Dashboard para Dados sobre a Mobilidade Urbana em Salvador

Luis Borges Filho
Departamento de Computação
Instituto Federal da Bahia
Salvador, Brasil
E-mail: luisborges.ssa@gmail.com

Pablo Vieira Florentino (Orientador)
Departamento de Computação
Instituto Federal da Bahia
Salvador, Brasil
E-mail: pablovf@ifba.edu.br

Resumo – O transporte público é um dos principais temas nos grandes centros urbanos, onde pesquisadores buscam equacionar o aumento recorrente do número de veículos particulares, crescimento populacional, mobilidade e qualidade de vida. Tradicionalmente em nosso país as políticas públicas têm privilegiado segmentos que contribuem para este cenário. O presente trabalho tem como objetivo investigar as possibilidades de uso da Linguagem R e do pacote Shiny para a criação de um dashboard a partir da análise exploratória dos dados e informações existentes nos Anuários de Transportes Urbano mais recentes produzidos pela Secretaria de Mobilidade da Prefeitura Municipal de Salvador, de forma buscar um maior entendimento sobre o panorama da mobilidade na cidade de Salvador, suas tendências, possibilidades e oportunidades.

Palavras chaves – mobilidade urbana, transporte público, transporte público em Salvador, dados abertos, análise de dados, visualização de dados, dashboard, storytelling, Linguagem R, Shiny.

I. INTRODUÇÃO

A formação do cenário atual de mobilidade urbana em nosso país desenvolveu-se ao longo de décadas com diversas variações, embora apresentando um elemento recorrente: a diminuição progressiva de investimentos públicos em transporte de massa, por um lado e, por outro, um aumento da motorização individual [1]. Esse conjunto de decisões resultou em um impacto nas relações sociais e na qualidade de vida, notadamente em grandes centros urbanos onde a desigualdade se evidencia nas condições de mobilidade em função da renda [2].

Em 2016 na Conferência das Nações Unidas sobre Habitação e Desenvolvimento Urbano Sustentável [2], um dos temas de destaque foi a mobilidade urbana segura, saudável e sustentável, onde observa-se no item 100:

"*Apoiaremos a oferta de redes bem projetadas de ruas e espaços públicos seguros, inclusivos a todos, acessíveis, verdes e de qualidade, livres de crime e violência, incluindo o assédio sexual e a violência de gênero, considerando a escala humana, bem como a adoção de medidas que permitamo melhor uso comercial possível dos andares no nível da rua, impulsionando o comércio e mercados locais, tanto formais como informais e iniciativas comunitárias sem fins lucrativos, trazendo as pessoas para os espaços públicos e promovendo a mobilidade de pedestres e ciclistas com o objetivo de melhorar a saúde e o bem-estar*."

Embora instituições da sociedade civil como a Organização das Nações Unidas, entre outras, reconheçam que uma mobilidade urbana de qualidade e acessível seja de suma importância para o bem-estar de uma sociedade, o lobby da indústria automobilística tem sido um fator de influência na formação de políticas públicas de transporte, com impactos em toda a população, embora, de forma mais contundente entre as camadas de baixa renda [3].

Diante deste cenário é importante que a população de um município disponha de informações organizadas e integradas sobre as condições de mobilidade urbana em sua cidade, seja para a compreensão de uma realidade, suas alternativas e possibilidades, seja para demandar políticas públicas que atendam suas necessidades, estimulando a participação do cidadão nos processos decisórios da gestão pública. Assim, este trabalho tem como objetivo apresentar um conjunto de dados e informações sobre a situação de mobilidade urbana na cidade de Salvador nos últimos anos, através de um painel digital interativo – um *dashboard*.

A coleta de informações e o desenvolvimento da base de dados utilizada neste projeto, envolveu a pesquisa em fontes de dados abertos governamentais, fontes criadas por iniciativas da sociedade civil, assim como fontes obtidas via Lei de Acesso à Informação. O tratamento e estruturação desses dados foram efetuados a partir de conceitos e recursos de Engenharia de Dados e, para a organização visual dos gráficos produzidos a partir dessas fontes, foram considerados conceitos de *storytelling* – técnicas para comunicação de dados, informações, análises ou ideias através de recursos multimídia –, e de princípios da psicologia Gestalt [4], que aborda a forma como nosso cérebro lida com informações visuais.

II. REFERENCIAL BIBLIOGRÁFICO

Cidade de Salvador – Cenário atual

A cidade de Salvador possui uma extensão territorial de 693,453Km², uma população estimada de 2.900.319 habitantes [5], tendo mais de 3.700 Km de pistas de rolamento, sendo que, de acordo com o Plano de Mobilidade (PlanMob) [6], 78Km dessas pistas tem situação típica de congestionamento em horários de pico, ao passo que 175Km estão em situação próxima à saturação.

Foram identificadas, ainda de acordo com o PlanMob – SSA, 33 regiões/localidades que constituem "gargalos de trânsito" das quais 13 necessitam de intervenções de adequação diversas, tais como pontilhões, viadutos, etc, além de 22 áreas que demandam revisão no esquema de circulação, sinalização de tráfego e/ou coordenação de semáforos, além de pequenas obras de natureza viária.

Publicado pela Prefeitura de Salvador em 2018, o PlanMob – SSA apresenta um conjunto importante de metas, contemplando os diversos modais de transporte,

incluindo o transporte ativo – meios de transporte por propulsão humana, tais como bicicletas, patinetes, patins, entre outros. Entretanto, no quesito “Dados Abertos”, embora conste no escopo de uma ação prevista, transcorridos quatro anos desde sua publicação, ainda não dispomos de tais dados para consulta.

Observa-se, contudo, através das publicações existentes no site da Secretaria de Mobilidade da Cidade de Salvador (SEMOB) [7], que existem estudos e pesquisas, cujos resultados encontram-se no formato PDF, porém, tais dados não se encontram organizados e tampouco estruturados em formatos mais acessíveis, tais como JSON ou CSV, o que dificulta o seu acesso e utilização, em inobservância às boas práticas de políticas de abertura de dados, bem como à Cartilha Técnica para Publicação de Dados Abertos no Brasil [8].

DADOS ABERTOS

Segundo a Open Knowledge Foundation – OKF, uma organização sem fins lucrativos de promoção do conhecimento livre com enfoque em Dados Abertos, com representação no Brasil através da Open Knowledge Brasil – OKBR [9], “*dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribuí-los, estando sujeito a, no máximo, a exigência de creditar a sua autoria e compartilhar pela mesma licença*”.

Em 11 de maio de 2016 foi publicado o decreto federal nº 8.777 que instituiu a Política de Dados Abertos do Poder Executivo federal. Embora destinado ao Poder Executivo, este decreto assinala conceitos importantes, pois estabelecem orientações essenciais ao processo de abertura de dados, assim, destacam-se no Artigo 2:

I – dado – sequência de símbolos ou valores, representados em qualquer meio, produzidos como resultado de um processo natural ou artificial;

II – dado acessível ao público – qualquer dado gerado ou acumulado pelo Governo que não esteja sob sigilo ou sob restrição de acesso nos termos da Lei nº 12.527, de 18 de novembro de 2011;

III – Dados Abertos – dados acessíveis ao público, representados em meio digital, estruturados em formato aberto, processáveis por máquina, referenciados na internet e disponibilizados sob licença aberta que permita sua livre utilização, consumo ou cruzamento, limitando-se a creditar a autoria ou a fonte;

IV – Formato Aberto – formato de arquivo não proprietário, cuja especificação esteja documentada publicamente e seja de livre conhecimento e implementação, livre de patentes ou qualquer outra restrição legal quanto à sua utilização;

V – Plano de Dados Abertos – documento orientador para as ações de implementação e promoção de abertura de dados de cada órgão ou entidade da administração pública federal, obedecidos aos padrões mínimos de qualidade, de forma a facilitar o entendimento e a reutilização das informações.

O Tribunal de Contas da União, por sua vez, destaca 5 motivos para a abertura de dados na administração pública [10]:

1. A sociedade exige mais transparência na gestão pública;
2. A própria sociedade pode contribuir com serviços inovadores ao cidadão;
3. Ajuda a aprimorar a qualidade dos dados governamentais;
4. Viabiliza novos negócios;
5. É obrigatório por Lei.

Apesar de termos uma política de abertura de dados no Brasil em crescente desenvolvimento, não é incomum encontramos portais da transparência de órgãos públicos onde são veiculados somente dados de natureza financeira. A transparência deve englobar, naturalmente, as demais áreas da gestão pública, desde que os dados aí contidos não estejam sob sigilo ou restrição de acesso.

Outro aspecto que chama a atenção é que, não raro, os dados divulgados não se encontram em formatos abertos, muitas vezes estando em PDF ou mesmo sob a forma de imagens cujo acesso editável ao conteúdo, tendo em vista um processamento automatizado, só seria possível mediante uso de tecnologias como o Reconhecimento Ótico de Caracteres ou OCR (Optical Character Recognition), o que dificulta o acesso ao seu conteúdo pelos pesquisadores e sociedade civil em geral.

Em 2018 a Diretoria de Análise de Políticas Públicas da Fundação Getúlio Vargas e a Open Knowledge Brasil publicaram o Índice de Dados Abertos para as Cidades [11], onde foram avaliadas 136 bases de Dados Abertos, distribuídas em 17 dimensões, das cidades de Belo Horizonte-MG, Brasília-DF, Natal-RN, Porto Alegre-RS, Rio de Janeiro-RJ, Salvador-BA, São Paulo-SP e Uberlândia-MG. Os dados avaliados foram aqueles considerados de utilidade para o público em geral.

Segundo esse estudo, apenas 25% das bases de Dados Abertos estão 100% de acordo com a definição de Dados Abertos, sendo que 62% apresentam problemas de usabilidade e 38% de processo. Ainda de acordo com estudo, os problemas mais frequentes são “*a dificuldade de trabalhar dados (incluindo os metadados insuficientes), indisponibilidade de download da base de dados completa, dataset incompleto e ausência da informação em formato aberto*”.

A pontuação de cada cidade se baseou em dois aspectos: o “Escore”, determinado pelo alinhamento dos dados aos critérios internacionais de transparência e o "%Open", que

estabelece o percentual do total de dados que atende a todos os critérios da metodologia.

No quesito “Escore”, a cidade de Salvador ficou em 6º lugar com 55% - uma diferença de 29 pontos percentuais da primeira colocada, a cidade de São Paulo, com 84% –, ao passo que no "%Open", ficou na última colocação, com apenas 5%, neste caso, a diferença é de 42 pontos percentuais da primeira colocação, também São Paulo, com 47% (figura 1).

Cidades	Escore	Cidades	%Open
São Paulo	84%	São Paulo	47%
Rio de Janeiro	75%	Belo Horizonte	35%
Belo Horizonte	73%	Rio de Janeiro	29%
Porto Alegre	68%	Brasília	29%
Brasília	68%	Porto Alegre	23%
Salvador	55%	Uberlândia	17%
Uberlândia	53%	Natal	11%
Natal	43%	Salvador	5%

Figura 1: Fonte e elaboração FVG/DAPP e OKBr [11].

FORMAS DE DISTRIBUIÇÃO DE DADOS

A abertura de dados deve resultar em mecanismos que facilitem a sua utilização. Dentre esses mecanismos os mais comuns são a viabilização de arquivos em formato aberto para download e as APIs. Uma API – Application Programming Interface –, é um “conjunto de regras definidas que explica como computadores ou aplicações se comunicam uns com os outros” [12]. Na figura 2, vemos o funcionamento de uma API. Basicamente uma API funciona em quatro etapas [12]:

1. Um aplicativo cliente inicia uma chamada de API para recuperar informações, também conhecida como solicitação. Essa solicitação é processada de um aplicativo para o servidor da Web por meio do URI (Uniform Resource Identifier – URIs são identificadores do nome e local de um arquivo ou recurso em um formato uniforme) da API e inclui um “*request verb*”, cabeçalhos e, às vezes, um corpo da solicitação;

2. Após receber uma solicitação válida, a API faz uma chamada para o programa externo ou servidor web;

3. O servidor envia uma resposta à API com as informações solicitadas;

4. A API transfere os dados para o aplicativo solicitante inicial.

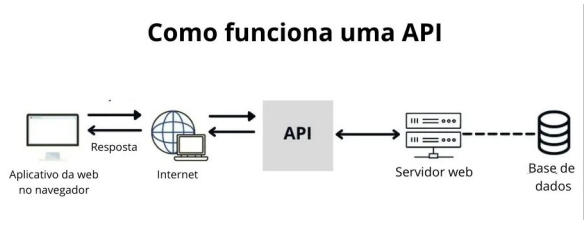

Figura 2: Funcionamento de uma API [13].

Devido a essas características, uma API constitui um recurso inestimável para o processo de abertura de dados, conferindo agilidade no consumo de dados e no desenvolvimento de soluções e serviços. Atualmente existem diversos protocolos que são utilizados no desenvolvimento de APIs, dentre os quais se destacam [12]:

- SOAP (Simple Object Access Protocol) é um protocolo de API construído com XML, permitindo que os usuários enviem e recebam dados através de SMTP e HTTP. Com APIs SOAP, é mais fácil compartilhar informações entre aplicativos ou componentes de software que estão sendo executados em ambientes diferentes ou escritos em linguagens diferentes;

- XML-RPC é um protocolo que se baseia em um formato específico de XML para transferir dados, enquanto o SOAP usa um formato XML proprietário. O XML-RPC é mais antigo que o SOAP, mas muito mais simples e relativamente leve, pois usa largura de banda mínima;

- JSON-RPC é um protocolo semelhante ao XML-RPC, pois ambos são chamadas de procedimento remoto (RPCs), mas este usa JSON em vez do formato XML para transferir dados. Ambos os protocolos são simples. Embora as chamadas possam conter vários parâmetros, elas esperam apenas um resultado;

- REST (Representational State Transfer) é um conjunto de princípios de arquitetura de API web, o que significa que não existem padrões oficiais (ao contrário daqueles com protocolo). Para ser uma API REST (também conhecida como API RESTful), a interface deve obedecer a certas restrições arquitetônicas. É possível construir APIs RESTful com protocolos SOAP, mas os dois padrões geralmente são vistos como especificações concorrentes.

Como suporte a uma API está a engenharia de dados que transformará os dados brutos em diversos formatos e oriundos de variadas fontes, em formatos passíveis de serem consumidos via APIs.

DADOS ABERTOS CONECTADOS LINKED OPEN DATA – LOD

A Tecnologia da Informação evolui de uma forma constante, criando produtos cada vez mais avançados. Computadores e dispositivos inteligentes mais rápidos e com mais memória têm se tornado cada vez mais acessíveis ao grande público, assim como internet mais rápida e de melhor qualidade.

Segundo dados do Internet World Stats [14], site especializado em estatísticas sobre a internet, somente na América Latina e Caribe, o número de pessoas acessando a *web* aumentou 2.907% entre os anos 2000 e 2022. Nesse

contexto, uma quantidade crescente de dados e informações é produzida e compartilhada pela internet, seja pelas organizações, seja pelo indivíduo.

USO INTERNET MUNDIAL E ESTATÍSTICAS POPULACIONAIS
Estimativas do ano de 2022

Mundo Regiões	População (2022 Est.)	População % do mundo	Usuários da Internet 30 de junho de 2022	Penetração Taxa (% Pop.)	Crescimento 2000-2022	Internet % Mundial %
África	1.394.588.547	17,6 %	652.865.628	46,8 %	14.362 %	11,9 %
Ásia	4.352.169.960	54,9 %	2.934.186.678	67,4 %	2.467 %	53,6 %
Europa	837.472.045	10,6 %	750.045.495	89,6 %	614 %	13,7 %
América Latina / Caribe.	664.099.841	8,4 %	543.396.621	81,8 %	2.907 %	9,9 %
Norte América	374.226.482	4,7 %	349.572.583	93,4 %	223 %	6,4 %
Oriente Médio	268.302.801	3,4 %	211.796.760	78,9 %	6.378 %	3,9 %
Oceania / Austrália	43.602.955	0,5 %	31.191.971	71,5 %	309 %	0,6 %
TOTAL MUNDIAL	7.934.462.631	100,0 %	5.473.055.736	69,0 %	1.416 %	100,0 %

NOTAS: (1) As estimativas de uso da Internet e estatísticas da população mundial são para 31 de julho de 2022. (2) CLIQUE em cada nome de região mundial para obter informações detalhadas sobre o uso regional. Os números (3) Demográfico (População) são baseados em dados do Divisão de População das Nações Unidas. As informações de uso da Internet (4) vêm de dados publicados por Nielsen Online, pelo Internacional União das Telecomunicações, por GfK, pelos reguladores locais de TIC e outras fontes confiáveis. (5) Para definições, ajuda na navegação e isenções de responsabilidade, por favor consulte o Guia de surf no site. (6) As informações deste site podem ser citadas, dando o devido crédito a www.internetworldstats.com.

Figura 3: Crescimento do uso da internet de 2000 a 2022 [14].

A *web* acaba por se tornar um vasto repositório de dados e informações. Assim, é muito comum que, antes de tomarmos uma decisão, procuremos informações e o primeiro lugar onde pensamos em buscar é a internet e, da mesma forma que os aparelhos citados, a internet também se encontra em constante evolução.

Em 2001, o físico inglês Tim Bernes-Lee pensou em uma série de princípios cujo objetivo era buscar maneiras mais simples de representação do conhecimento, de forma que o processo de entendimento e manipulação pelos computadores fosse facilitado. Esses princípios ficariam conhecidos como *Web* Semântica. A internet, até então, se baseava no inter-relacionamento entre documentos ou links, criando conexões que referenciavam outras conexões resultando em uma vasta rede de conteúdo [15].

Segundo Rauntemberg [15] "*Com a Web Semântica, em vez de termos links para documentos, são usados links para dados. De certa forma, esses links entre os dados garantem uma semântica de relacionamento. E se estes mesmos links forem baseados em vocabulários bem formados, tem-se o fundamento da web de dados. Se os dados forem certificados e mantidos por organizações ou pessoas confiáveis, forma-se uma imensa teia de dados inter-relacionados que podem ser explorados nos mais variados contextos*".

Ainda segundo Rautenberg [15]: "*Nesse sentido um novo significado dos dados é adicionado à web tradicional, permitindo a troca de informações entre seres humanos e máquinas por meio de vocabulários compartilhados e de uso comum. Entretanto, para que isso seja possível, é necessário disponibilizar os dados com o uso de formatos padronizados, acessíveis e gerenciáveis por ferramentas da web de dados*".

O consórcio W3C (World Wide Web Consortium) [16], criado em 1994 por Bernes-Lee e outros, é a principal organização de padronização da World Wide Web. Composto atualmente por 464 membros, dentre os quais podemos destacar: Adobe, Amazon, Apache, Apple, AT&T, Autodesk, IBM, Microsoft, entre outras importantes empresas de tecnologia, apresentou um conjunto de diretrizes para a publicação e consumo de dados na web, conhecidas como as Melhores Práticas [17], essas diretrizes são de propósito geral e independem do modelo de aplicação. A adoção destas práticas resultaria em um conjunto de benefícios [12]:

- **Compreensão** – habilita os seres humanos a melhor entender a estrutura e o significado dos dados, dos metadados e da natureza dos conjuntos de dados;
- **Processabilidade** – habilita máquinas para processar automaticamente e manipular os dados dentro de um conjunto de dados;
- **Descoberta** – habilita máquinas para descobrirem automaticamente dados ou conjuntos de dados;
- **Reúso** – aumenta as chances do reúso de um conjunto de dados por diferentes consumidores de dados;
- **Confiança** – melhora a confiança dos consumidores no conjunto de dados;
- **Ligação** – possibilita a criação de ligações (links) entre recursos de dados (conjuntos de dados e itens de dados);
- **Acesso** – permite aos humanos e máquinas acessar dados atualizados em uma variedade de formatos;
- **Interoperabilidade** – facilita a obtenção de um consenso entre publicadores e consumidores de dados;

A seguir, as Melhores Práticas (MPs) [17]recomendadas pelo consórcio W3C:

- **Metadados – MP:** "*Forneça metadados para usuários humanos e aplicativos de computador*". Os metadados contêm informações que permitem aos consumidores de dados conhecer sua estrutura, significado, licença de uso, organização criadora, qualidade dos dados, e periodicidade de atualização. Formatos indicados: Turtle, JSON, ou incorporados na página HTML através de [HTML-RDFA] ou [JSON-LD]. Exemplos legíveis para humanos e máquinas podem ser encontrados aqui [18] e [19], respectivamente.
- **Licenças de Dados – MP:** "*Forneça um link ou uma cópia do contrato de licença que controla o uso dos dados*". Licenças informam ao consumidor de dados o que é permitido e/ou proibido. É sempre importante que dados publicados na internet tenham uma licença de uso, tendo em vista a proteção legal dos dados e o controle do seu uso.
- **Proveniência dos Dados – MP:** "*Forneça informações completas sobre as origens dos dados e quaisquer alterações feitas*". Informar a

proveniência permite ao consumidor de dados confiar na integridade e credibilidade dos dados.

- **Qualidade dos Dados – MP:** "*Forneça informações sobre a qualidade dos dados e adequação para fins específicos*". A avaliação da qualidade de um conjunto de dado envolve dimensões diversas da qualidade, representando aspectos que são importantes para cada público, seja editor ou consumidor de dados. Existe um Vocabulário de Qualidade de Dados – Data Catalog Vocabulary (DCAT), criado para facilitar a interoperabilidade entre catálogos de dados publicados na *web* [20].

- **Versionamento de Dados – MP: "***Atribua e indique um número de versão ou data para cada conjunto de dados*". Dados Abertos podem ser atualizados de forma programada ou à medida em que estão sendo criados, assim, uma periodicidade estabelecida permite ao consumidor de dados organizar também o seu cronograma de consumo, atualização e divulgação.

- **Identificadores de Dados – MP:** "*Usar URIs persistentes como identificadores de conjuntos de dados*".

- **Formatos de Dados – MP:** "*Disponibilize os dados em um formato de dados padronizado e legível por máquina que seja adequado ao seu uso pretendido ou potencial*". Exemplos de formatos recomendados: CSV, XML, HDF5, JSON, RDF, RDF/XML, JSON-LD ou Turtle.

- **Vocabulários de Dados – MP:** "*Use termos de vocabulários compartilhados, preferencialmente padronizados, para codificar dados e metadados*".

- **Acesso a Dados – MP:** "*Permita que os consumidores recuperem o conjunto de dados completo com uma única solicitação*".

- **APIs de acesso a dados – MP:** "*Ofereça uma API para fornecer dados se você tiver os recursos para isso*".

- **Preservação de Dados – MP:** "*Ao remover dados da Web, preserve o identificador e forneça informações sobre o recurso arquivado*".

- **Comentários (*feedback*) – MP:** "*Forneça um meio facilmente detectável para os consumidores oferecerem feedback*".

- **Enriquecimento de Dados – MP:** "*Enriqueça seus dados gerando novos dados quando isso aumentar seu valor*".

- **Republicação – MP:** "*Informe ao editor original quando você estiver reutilizando os dados dele. Se você encontrar um erro ou tiver sugestões ou elogios, informe-os*".

Na figura 4 pode-se ver como os diferentes ambientes de dados – Dados na Web, Dados Abertos, Dados Conectados e Dados Abertos Conectados se relacionam.

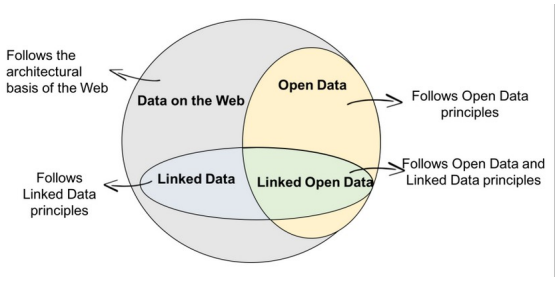

Figura 4: Dados na *Web* x Dados Abertos x Dados Conectados [21]

Uma evolução do conceito de Dados Abertos Conectados é o "*LOD Cloud*" ou Nuvem de Dados Abertos Conectados – estruturada como um grafo de conhecimento, constituindo uma Web Semântica de Dados Conectados – figura 5.

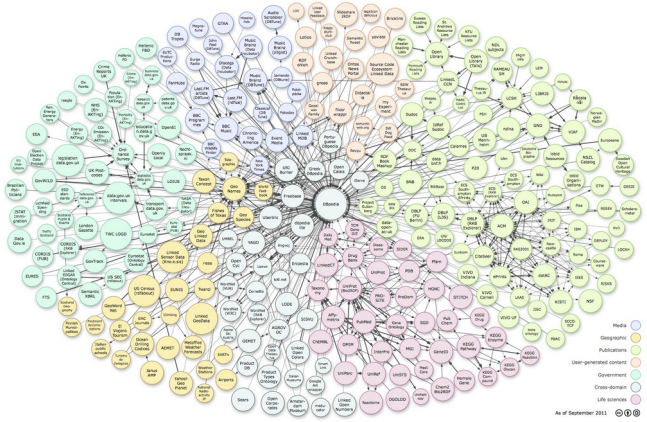

Figura 5: Nuvem de Dados Abertos Conectados estabelecido pelo projeto *LOD Cloud* [22].

A Computação em Nuvem, ou *Cloud Computing*, é a oferta de serviços de computação em *Data Centers*, sob demanda, por meio da internet, onde o contratante paga somente pelos recursos que são consumidos, resultando em economia, flexibilidade e também escalabilidade, quando necessária.

Diferente do que ocorria anteriormente, onde uma aplicação *web* precisava de uma equipe para cuidar da manutenção da disponibilidade do serviço online, com o conceito de *Cloud* empresas viabilizam toda a estrutura de hospedagem, disponibilidade e segurança, permitindo que um projeto *web* tenha sua equipe focada no processo de desenvolvimento.

Com a evolução da adoção destas práticas por parte das organizações, a realização de um projeto de Ciência de Dados fica muito mais ágil e objetiva.

Em novembro de 2007, foi aberto o escritório da W3C Brasil [23], onde é possível encontrar bastante conteúdo, além de cursos sobre Dados Abertos, Web Semântica,

Ontologia, manuais, guias de referência, etc, o que pode colaborar para o fortalecimento de uma cultura alinhada a essas boas práticas.

ENGENHARIA DE DADOS

A tendência crescente na quantidade e diferentes tipos de dados resultou no desenvolvimento de tecnologias para lidar com este cenário. O conceito de *Big Data* desenvolveu-se a partir dessa necessidade, possibilitando análises em grandes volumes de dados, nas mais diversas áreas e situações.

A *Encyclopedia of Big Data Technologies* [24], lista dezenas de diferentes tipos de cenários e aplicações de *Big Data*, tais como: saúde, cidades inteligentes, herança cultural, bem-estar social, sequenciamento de DNA, supercomputação, cibersegurança, processamento de dados semânticos na área de ciências da vida e na ciência dos materiais, entre outros.

Inicialmente o conceito de *Big Data* era caracterizado pelos, então chamados, 3 ‘Vs’:

- Volume – é possível processar grandes volumes de dados, superiores à ordem de Terabytes;
- Variedade – processamento de dados estruturados, semiestruturados e não estruturados;
- Velocidade – capacidade de processamento em tempo real.

Posteriormente, outros ‘Vs’ foram sendo adicionados a essa lista, por exemplo:

- Veracidade – se referindo à qualidade dos dados;
- Valor – relacionado à capacidade de transformar dados em valor a um projeto ou organização.

Essas características demandam um conjunto de ações que possibilitem uma estruturação e apresentação dos dados de forma que possam ser utilizados de acordo com o projeto ou modelo de negócio de uma organização. Chamamos esse conjunto de ações de Engenharia de Dados.

A engenharia de dados faz usos de técnicas para prover dados aos setores e serviços de uma organização ou projeto. Uma das técnicas principais chama-se ETL – formada pelas iniciais das palavras em inglês *Extract*, *Transform* e *Load*, Extrair, Transformar e Carregar, respectivamente, designam as operações a seguir:

- Extrair – consiste na coletas dos dados armazenados em uma ou mais fontes e em diferentes formatos;
- Transformar – aplicar técnicas que estruturem os dados de acordo a necessidade.
- Carregar – prover os dados transformados a um armazém de dados (em inglês, *Data Warehouse*) onde ficarão disponíveis para utilização.

A figura 6 ilustra essas etapas:

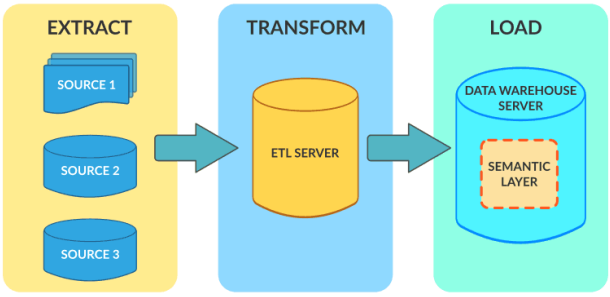

Figura 6: ETL – *Extract*, *Transform* e *Load* [25].

A etapa de transformação pode incluir diversos subprocessos:

- Limpeza – tratamento das inconsistências e valores ausentes;
- Padronização – os dados são padronizados de acordo com critérios pré estabelecidos;
- Deduplicação – remoção de dados redundantes;
- Verificação – os dados que não podem ser utilizados serão removidos, sinalizando essas anomalias;
- Classificação – os dados são organizados de acordo com suas características;
- Outras tarefas — aplicação de regras adicionais, quando necessário, visando a melhoria da qualidade dos dados.

Na etapa de carga, os dados são disponibilizados em um *Data Warehouse*. Os processos de carga podem ser:

- Carga Completa – todos os dados anteriores são apagados e substituídos pelos novos;
- Carga Incremental ou Diferencial – inclui somente aqueles dados que não existiam anteriormente.

Existem dois tipos de cargas incrementais, de acordo com o volume de dados a ser carregado: carga incremental em *streaming*, que ocorre geralmente em lotes menores e em tempo real, e a carga incremental em lote ou *batch*, na qual os dados são armazenados até o momento em que serão carregados todos de uma vez.

Entre as principais vantagens da carga completa estão: É simples de implementar, pois, consiste em apagar os dados anteriores e substituir pelos novos; baixa necessidade de manutenção, pois, não é necessário efetuar o gerenciamento de versões, ou verificar se os dados foram ou não

atualizados; seu design simples – processo de fácil configuração, basicamente, o envio ou *upload* dos arquivos atuais, substituindo os anteriores, se necessidade de controle de versão.

Entre as desvantagens podemos citar: Quando somente parte dos dados foi atualizado em relação à versão anterior, será preciso atualizar todo o conteúdo, o que acarreta uma demora na finalização do processo; desempenho lento – à medida que a quantidade de dados aumenta, também aumentará o tempo necessário à conclusão da operação por completo.

Na carga incremental ou diferencial, as principais vantagens são: maior rapidez na conclusão do processo se comparado à carga completa, pois, somente os dados que mudaram serão carregados; histórico de armazenamento. Segundo Mitchell [26], a carga incremental é o padrão de projeto ideal para a maioria das operações de ETL, entretanto, apresenta alguns desafios, tais como: desenvolvedor deve estabelecer a lógica de carga incremental para localizar os dados que foram alterados e os novos; nem sempre existe uma forma clara de identificar os dados que precisarão entrar na atualização [26] o que exigirá a utilização de ferramentas adicionais de controle.

Segundo Reis [27], a Engenharia de Dados está a montante em relação à Ciência de Dados, que se encontra a jusante daquela. Em outras palavras, é a Engenharia de Dados que provê as entradas, os dados que serão utilizados pelos cientistas e analistas de dados (figura 7).

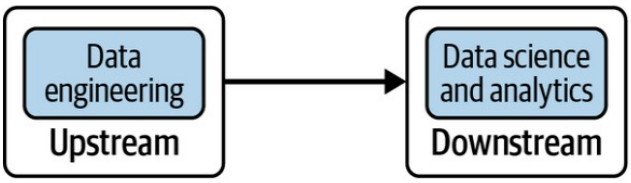

Figura 7: Engenharia de Dados subsidia a Ciência a Análise de Dados [27].

A ciência de dados é uma área do conhecimento multidisciplinar que combina programação, matemática, em especial estatística, análise avançada, inteligência artificial e aprendizado de máquina para descobrir padrões, relações, *insights* nos dados de uma organização.

DADO X INFORMAÇÃO

Segundo Boscarioli [28], "*O dado é um fato, um valor documentado ou um resultado de medição. Quando um sentido semântico ou um significado é atribuído aos dados, gera-se informação. Quando estes significados se tornam familiares, ou seja, quando um agente os aprende, este se torna consciente e capaz de tomar decisões a partir deles, e surge o conhecimento*".

Podemos entender dados como valores brutos, não contextualizados; a partir do momento em que uma contextualização é feita chegamos à informação. Além disso, um outro aspecto que diferencia dado e informação é que um dado pode ser lido, ao passo que uma informação pode ser interpretada. Para melhor compreensão, um exemplo na figura 8.

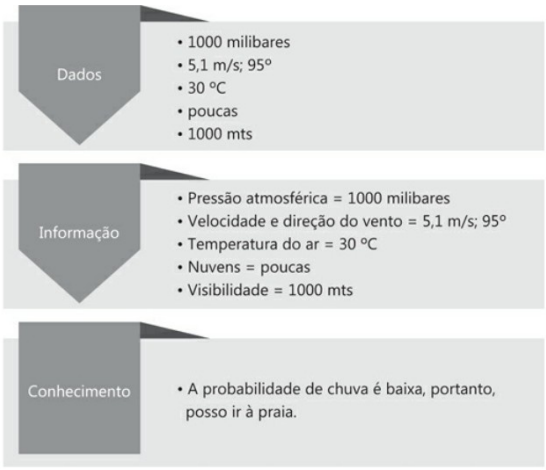

Figura 8: Dados, informação e conhecimento [29].

A partir da leitura e interpretação, chegamos ao Conhecimento, onde é possível fazer projeções ou previsões a partir do histórico de fatos anteriores. Na sequência podemos chegar à Sabedoria, onde teremos condições de decidir sobre o que fazer. Esta sequência de passos é conhecida como "Pirâmide do Conhecimento" (figura 9), e também pela sigla DIKW, formada pelas palavras em inglês: Data, Information, Knowledge e Wisdom.

Figura 9: A pirâmide do conhecimento [30].

Também encontramos uma variação dessa pirâmide onde após o Conhecimento chegamos à Ideia ou *Insight* – aquele lampejo que nos ocorre quando uma situação ou problema é bem compreendido e, assim, solucionado.

NOMENCLATURA E TIPOS DE DADOS

Um conjunto de dados pode apresentar uma ou mais variedades de acordo com a sua estruturação [29]:

- Dados estruturados – Os dados estão organizados em campos fixos, como planilha, tabela ou banco de dados;

- Semiestruturados – Não apresentam uma estruturação completa, mas, parcial, geralmente utilizando marcadores que identificam parte dos dados; arquivos HTML, que têm uma hierarquia construída a partir de elementos semânticos – as

tags; arquivos XML que estabelecem regras para a formatação de documentos; e-mails, cujos campos ‘remetente, ‘destinatário’, ‘assunto’, e outros dados do e-mail, que podem ser utilizados como campos fixos;

- Não estruturados – Não apresentam uma estrutura ou arquitetura definida, por exemplo, textos, imagens, áudio, vídeo e outros.

Segundo Pickell [31], as principais diferenças entre dados estruturados e não estruturados (tabela 1):

Dados Estruturados	Dados não Estruturados
Dados estruturados são dados quantitativos que consistem em números e valores.	Dados não estruturados são dados qualitativos que consistem em áudio, vídeo, sensores, descrições e muito mais.
Os dados estruturados são usados no aprendizado de máquina e acionam algoritmos de aprendizado de máquina.	Os dados não estruturados são usados no processamento de linguagem natural e na mineração de texto.
Os dados estruturados são armazenados em formatos tabulares, como planilhas do Excel ou bancos de dados SQL.	Armazenados como arquivos de áudio, arquivos de vídeos ou bancos de dados NoSQL
Os dados estruturados têm um modelo de dados predefinido.	Os dados não estruturados não possuem um modelo de dados predefinido.
Os dados estruturados são provenientes de formulários on-line, sensores de GPS, logs de rede, logs de servidor da Web, sistemas OLTP e similares.	Os dados não estruturados são provenientes de mensagens de e-mail, documentos de processamento de texto, arquivos PDF e assim por diante.
Os dados estruturados são armazenados em data warehouses	Dados não estruturados são armazenados em data lakes
Os dados estruturados requerem menos espaço de armazenamento e são altamente escaláveis.	Dados não estruturados requerem mais espaço de armazenamento e são difíceis de dimensionar.

Tabela 1: Dados estruturados x dados não estruturados

DADOS – ATRIBUTOS

Quanto aos atributos, os dados podem ser classificados em “Numéricos” e “Categóricos” (figura 10), que são subdivididos em:

Numéricos:

- Discreto – São valores inteiros, positivos e negativos;
- Contínuos – São valores em números reais – números decimais, dizimas periódicas;
- Razão – Representam valores que podem inciar a partir do zero, não podendo ser negativos, por exemplo, medidas de altura, velocidade.

Categóricos:

- Binários – Podem assumir os valores verdadeiro ou falso – em programação podem ser encontrados como TRUE, True, true, FALSE, False ou false;
- Nominal – Informam nomes, rótulos ou categorias;
- Ordinal – Informam valores que fazem parte de uma sequência ou gradação de uma medida, dados que podem ser classificados segundo uma ordem, como por exemplo: “Bom”, “Ótimo”, “Excelente”.

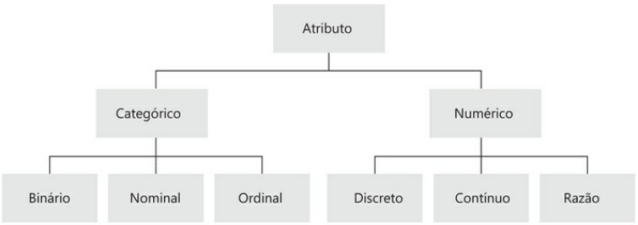

Figura 10: Atributos e suas subdivisões [29].

Os dados podem, ainda, ser divididos em:

- Qualitativos – expressam valores não numéricos, como por exemplo: nome, gênero, etnia, estado civil;
- Quantitativos – expressam valores numéricos contínuos ou discretos.

TIPOS DE ANÁLISES DE DADOS

A natureza das análises de dados efetuadas através da Ciência de Dados pode ser divididas em [28],[32]:

- Análise descritiva – investiga em dados históricos e atuais se existem padrões que possam ser mapeados; procura responder “o que aconteceu” ou “o que está acontecendo”.
- Análise diagnóstica – procura respostas para entender “o porquê” de uma situação ter ocorrido; utiliza técnicas como descoberta de dados, mineração de dados e correlações;
- Análise preditiva – investiga em dados históricos e atuais para prever resultados e tendências futuras; utiliza técnicas como aprendizado de máquina, previsão, correspondência de padrões e modelagem preditiva;

- Análise prescritiva – procura responder qual o melhor curso de ação para uma determinada situação, com recomendações para possíveis situações futuras com base em dados.

No âmbito de negócios, existe também o conceito de "cenarização" [33], também conhecido com "descoberta de cenários de negócio", onde são feitas análises multivariadas (relação entre três ou mais variáveis) que visam a descoberta de combinações sistêmicas entre variáveis que possam levar a um resultado específico.

O modelo de análise será escolhido, portanto, de acordo com a natureza do problema ou situação para a qual se busca uma resposta.

CICLO DE VIDA DA ANÁLISE DE DADOS

O ciclo de vida da análise de dados estabelece um conjunto de processos que partem dos dados até a fase de tomada de decisão. Este ciclo é dividido em:

- Perguntar: É preciso definir qual a pergunta a ser respondida. Uma pergunta bem formulada não deve conter qualquer tipo de enviesamento, caso contrário, os resultados serão comprometidos;
- Preparar: Todo o processo envolvendo os dados desde sua criação à dinâmica de gerenciamento;
- Processar: As ações que visam a preparação dos dados para sua utilização nos processos de análise;
- Analisar: As ações desde a análise exploratória, passando pela análise em si, até a visualização;
- Compartilhar: Comunicar que foi descoberto;
- Agir: Utilização do que foi aprendido para responder a pergunta inicial. Operacionalização.

O valor de uma base ou conjunto de dados está relacionado à capacidade que temos de extrair conhecimentos que permitam a tomada inteligente de decisões. Uma das formas de avaliar os resultados das análises desenvolvidas pelos diferentes tipos de tarefas é através da análise visual de dados. Dentre as diversas opções, destacamos os painéis interativos, também conhecidos como *dashboards*, que permitem aplicar a análise visual dos dados trabalhados e extrair uma série de informações e conhecimentos.

VISUALIZAÇÃO DE DADOS – DASHBOARDS

Segundo Pandey [34], pesquisas experimentais a visualização de dados possui a capacidade de tornar a transmissão de uma informação mais persuasiva. De acordo com esse estudos, uma mensagem tem um poder maior de influência, sendo mais facilmente assimilada, quando organizada de forma visual.

Entretanto, para que uma mensagem seja persuasiva, é necessária a existência de no mínimo três fatores [34]:

1. Informação contextual – é importante fornecer uma introdução às evidências apresentadas por meio de dados e estatísticas;
2. A própria evidência – normalmente composta por números e tendências;
3. Os dados fornecidos em apoio à evidência.

Ainda segundo Pandey [34], destacam-se os diferentes caminhos que podem ocorrer no processo de persuasão e contato com a informação:

"*Quando a elaboração é alta (e, portanto, ocorre através da "rota central"), a persuasão depende das características centrais do argumento, principalmente sua qualidade e força. Quando a elaboração é baixa (e, portanto, ocorre por meio da "rota periférica"), o receptor reverte para heurísticas cognitivas, que se baseiam em características periféricas da mensagem, como credibilidade da fonte e fatores estéticos. A quantidade de elaboração e, portanto, se o processo persuasivo ocorre mais pela via central ou periférica, depende, por sua vez, da motivação e capacidade do receptor para processar a mensagem. A motivação depende em grande parte da relevância pessoal do tópico, e a habilidade depende se o receptor é capaz de processar a mensagem sem esforço cognitivo excessivo*".

Percebe-se que apresentação de uma informação visualmente estruturada e organizada constitui um dos aspectos a serem considerados quando da elaboração de uma solução em visualização de dados, porém, não menos importante estão a contextualização, a conscientização da importância da informação, o que, por sua vez, pode conduzir à motivação para sua utilização, além da adequação de formato ao público-alvo, facilitando o esforço cognitivo necessário à sua compreensão.

O estatístico Edward R. Tufte em sua obra "The Visual Display of Quantitative Information" [35], apresenta o princípio da "excelência gráfica" cujo objetivo é comunicar o maior número de ideias, no menor tempo, usando um mínimo de tinta e de espaço. Esse conceito subtende a otimização de recursos tendo em vista uma maximização do processo de apresentação de um conteúdo informacional.

Ainda segundo Tufte [35], uma apresentação gráfica deveria:

- Mostrar os dados;
- Induzir o espectador a pensar na substância em vez da metodologia, design gráfico, tecnologia de produção gráfica, ou qualquer outra coisa;
- Evitar distorcer o que os dados têm a dizer;

- Apresentar muitos números em um pequeno espaço;
- Tornar grandes conjuntos de dados consistentes;
- Incentivar o olho a comparar diferentes dados;
- Revelar os dados em vários níveis de detalhes, desde uma ampla visão geral até a estrutura fina;
- Servir a um propósito razoavelmente claro: descrição, exploração, tabulação ou decoração;
- Estar intimamente integrado com as descrições estatísticas e verbais de um conjunto de dados.

Assim, o processo de desenvolvimento de um *dashboard* deve ser bem planejado tendo em vista uma utilização inteligente de recursos e sua distribuição, além do aspecto estético, o que nos leva à observação de mais um importante conjunto de conceitos.

A PSICOLOGIA GESTALT

Segundo a Associação Brasileira de Gestalt-Terapia [36] "*Gestalt é uma palavra alemã, traduzida aproximadamente por "configuração", totalidade, um todo que é uma realidade em si, diferente da soma de suas partes. O sentido da experiência está em como os elementos da mesma estão configurados em um todo significativo*".

Esta definição acena com diversas possibilidades e, neste projeto, procuraremos apresentar algumas aplicações de princípios que foram estabelecidos pela Gestalt, direcionados ao design gráfico.

Na obra "Storytelling com Dados" Nussbaumer [37] destaca os princípios da Gestalt:

Proximidade – ao vermos objetos próximos uns dos outros, tendemos a pensar que fazem parte de um grupo. Na figura 11, o que separadamente seriam apenas pontos, quando colocados próximos ou organizados formam colunas, linhas e quadrados.

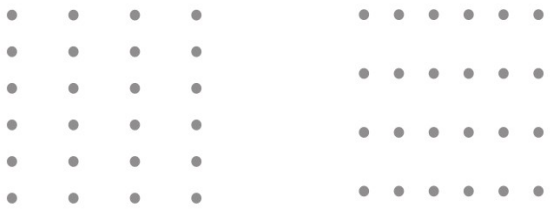

Figura 11: Proximidade – colunas e linhas [37].

Similaridade – tendemos a pensar que objetos de uma mesma cor, forma, tamanho ou orientação são pertencente a um mesmo grupo. Na figura 12 imediatamente identificamos as figuras semelhantes.

Figura 12: Similaridade – mesma cor, forma e tamanho [37].

Acercamento – objetos que estejam fisicamente delimitados são vistos como um grupo. Na figura 13 a delimitação sugere grupos.

Figura 13: Acercamento – delimitação criando grupos [37].

Fechamento – tendemos a seguir construções já existentes em nossa mente, dando sentido ao que percebemos de acordo com essas construções. Nas figuras 14 e 15 dois exemplos da abordagem baseada no conceito de fechamento.

Figura 14: Fechamento – conjunto de arcos se tornam um círculo [37].

Figura 15: Fechamento – manchas formam um dálmata [38].

Continuidade – princípio próximo ao do fechamento, tendemos a criar uma continuidade do que é visto mesmo onde não é aparente, de uma forma explícita (figura 16) – o alinhamento em sequência das barras substitui o eixo y.

Figura 16: Continuidade – o alinhamento em sequência das barras à esquerda substitui o eixo y [37].

Conexão – objetos fisicamente conectados são visto como parte de um grupo – mesmo quando outros princípios da Gestalt estão presentes, a conexão tem um apelo difícil de ignorar. Na figura 17 exemplos de conexão.

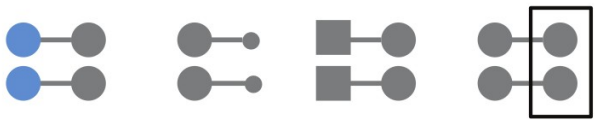

Figura 17: Conexão – mesmo diante de outros princípios da Gestalt, a conexão tem um forte apelo [37].

Através da utilização destes princípios em um projeto de visualização de dados, podemos transmitir informações de uma forma mais objetiva, concentrando nos elementos que são essenciais à comunicação de um conteúdo, evitando, assim, uma carga cognitiva desnecessária.

TRABALHOS CORRELATOS

MOBILIDADOS

Segundo a página da União de Ciclistas do Brasil [39], o projeto a Mobilidados [40] é "*uma base de dados que pretende contribuir com processos de elaboração, participação e controle social das políticas de mobilidade. O objetivo principal desta iniciativa é a colaboração para a construção de cidades mais equitativas e alinhadas aos princípios de Desenvolvimento Orientado ao Transporte Sustentável*". Criado em 2017, o projeto apresenta indicadores gerais das capitais, com maior detalhamento de dados de oito capitais e suas regiões metropolitanas.

Contempla indicadores sócio-econômicos, infraestrutura de transporte, ambientais e óbitos. Tecnologia utilizada no desenvolvimento: 100% linguagem R. Fonte dos dados: Não informado. Natureza da iniciativa: Voluntária da sociedade civil. (Figura 18).

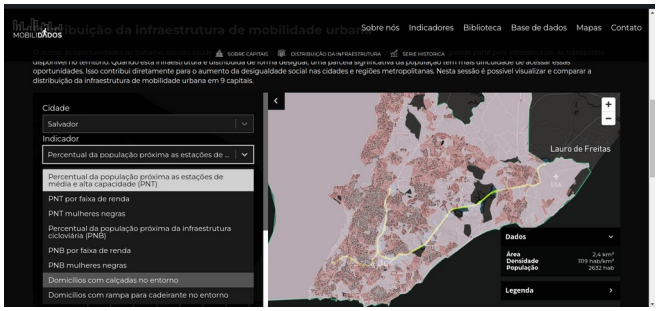

Figura 18: Mobilidados e a cidade de Salvador [40].

CICLOMAPA

Seguindo a página do projeto, o CicloMapa [41] "*surgiu do esforço conjunto da União dos Ciclistas do Brasil (UCB) e do Instituto de Políticas de Transporte e Desenvolvimento (ITDP Brasil) para apresentar mapas cicloviários padronizados das cidades brasileiras. O banco de dados escolhido foi o OpenStreetMap (OSM)* [42] *que é considerado o maior banco de dados abertos de mapas do mundo, e todo o projeto do CicloMapa também é open source, tendo seu código disponível no Github*". Voltado para o público ciclista, o projeto é desenvolvido de forma voluntária e colaborativa, realizando reuniões periódicas abertas à comunidade, aceitando a participação de novos colaboradores voluntários.

Contempla indicadores socioeconômicos, ciclovias, ciclofaixas, ciclorrotas e infraestrutura de apoio ao ciclista. Tecnologias utilizadas no desenvolvimento: 90.5% Javascript, 5.8% CSS, 2.3% Less e 1.4% HTML. Fonte dos dados: Dados abertos do OpenStreetMap. Natureza da iniciativa: Voluntária da sociedade civil. (Figura 19)

Figura 19: Ciclomapa e a cidade de Salvador [41].

SIMU

Iniciativa do Ministério do Desenvolvimento Regional, o SIMU – Infraestrutura De Mobilidade Urbana [43] apresenta inúmeros indicadores de infraestrutura, frotas de veiculos leves e pesados particulares, dados econômicos, ambientais, números de acidentes, com abrangência nacional e municipal, embora nem todos os municípios estejam contemplados no filtro da pesquisa. Salvador, por exemplo, não consta nesse filtro (Figura 20), embora esteja presente em parte dos gráficos. Não possuía informações sobre as fontes dos dados ou a tecnologia utilizada no desenvolvimento do projeto.

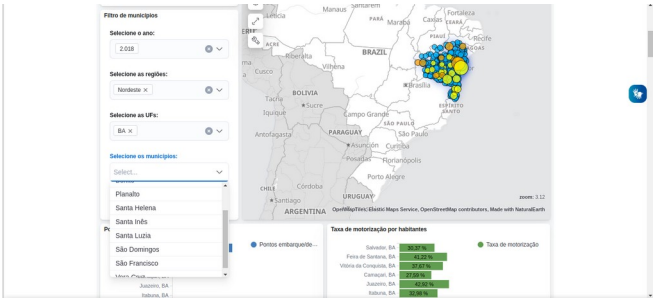

Figura 20: SIMU e o filtro de municípios do estado da Bahia [43].

O DESENVOLVIMENTO DA SOLUÇÃO

A seguir, descreveremos as etapas do processo de desenvolvimento da solução, as dificuldades que foram encontradas, motivos que nortearam nossas escolhas, cuidados e recomendações.

A FONTE DOS DADOS – SECRETARIA MUNICIPAL DE MOBILIDADE URBANA (SEMOB)

A SEMOB [7] vem, há alguns, anos produzindo o Anuário de Transportes Urbanos, entretanto, a localização desses documentos não é possível a partir da página da própria Secretaria. De fato, somente mediante busca via Google, encontramos os anuários de 2017 e 2018, sendo que os de 2019 a 2021 só foram obtidos via Lei de Acesso à Informação. Sobre o anuário de 2016, foram encontradas partes isoladas, além da capa, igualmente localizados via pesquisa no Google.

Lamentavelmente não existe, disponível ao público, uma base de dados com os dados dos Anuários que possa ser baixada e utilizada, nem tampouco uma API. Assim, foi preciso trabalhar a partir dos dados disponíveis nos arquivos em PDF.

TECNOLOGIA UTILIZADA – A LINGUAGEM R

A tecnologia escolhida para a realização deste projeto foi a Linguagem R, conhecida pelas suas bibliotecas para análises de dados e criação de gráficos. Trata-se de uma linguagem desenvolvida em 1993 por professores do Departamento de Estatística da Universidade de Auckland, Nova Zelândia, de código aberto e licença de uso gratuita, executável em sistemas UNIX, Windows e MacOS, multi-paradigma, incluindo a orientação a objetos e o paradigma funcional, com uma vasta coleção de pacotes e bibliotecas – atualmente mais de 18.000 pacotes [44] para computação estatística (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, clustering, entre outros), além de bibliotecas para visualização de dados. Neste último quesito, um dos destaques é o pacote Shiny para a criação de *dashboards* – que será utilizado neste projeto.

TECNOLOGIA DO DASHBOARD – PACOTE SHINY R

O Shiny é um pacote da linguagem R e um sistema de desenvolvimento de aplicações web interativas, criado por Joe Cheng, Diretor de Tecnologia do RStudio – Posit [45]. O Shiny tem, ainda, uma versão servidor, com opções gratuitas e pagas.

Dentre os diversos recursos que este pacote oferece, estão:

- Criação de painéis interativos;
- Relatórios em R Markdown;
- Plotagem de gráficos;
- Aplicativos Python interativos;
- Criação de APIs;
- Encapsulamento e empacotamento de uma aplicação sob a forma de um pacote R;
- Empacotamento da aplicação em contêiners;
- Execução em *localhost* e também em servidores;
- Inclusão de código ou biblioteca Javascript;
- Inclusão de HTML e uso de CSS.

A possibilidade de encapsular uma aplicação Shiny permite a sua distribuição sob a forma de um pacote. Da mesma forma, ao colocá-la em um contêiner, facilita a sua distribuição, também possibilitando a colocação em produção.

O PROCESSO DE EXTRAÇÃO DE DADOS

Como vimos, devido à indisponibilidade de recursos amigáveis para o uso facilitado dos dados da mobilidade urbana produzidos pela SEMOB, utilizamos os dados encontrados nos Anuários em PDF.

Para a leitura e extração dos dados nas tabelas existentes nos arquivos em PDF, utilizamos o pacote "*pdftools*" [46] [47]. Para as operações de transformação e visualização de dados utilizamos o pacote "tidyverse" [48].

A biblioteca "*pdftools*" faz parte do CRAN, Comprehensive R Archive Network, o repositório padrão para pacotes R. Segundo a página do CRAN, esta biblioteca apresenta "*Utilitários baseados em 'libpoppler' para extrair texto, fontes, anexos e metadados de um arquivo PDF. Também oferece suporte à renderização de alta qualidade de documentos PDF em formato PNG, JPEG, TIFF ou em vetores bitmap brutos para processamento posterior em R*" [47].

O pacote “tidyverse” [48] possui todo um conjunto de ferramentas necessárias ao fluxo de dados (figura 21):

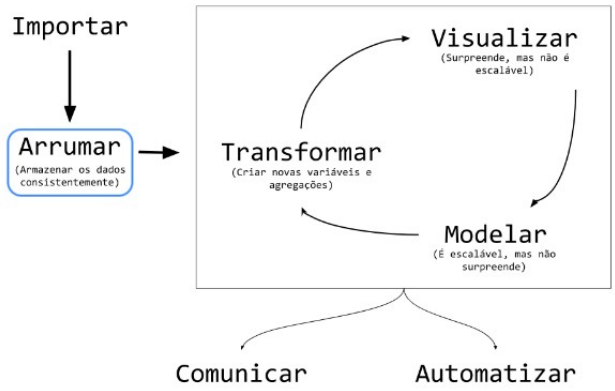

Figura 21: O fluxo de processamento de dados [49].

Um exemplo de extração da dados da tabela “Linhas Regulares”, existente na página 11 do Anuário 2020 pode ser visto na figura 22.

Ano: 2020

LINHAS REGULARES	CSN	OTT	PLAT	SUB TOTAL	POOL	TOTAL
JANEIRO	110	103	102	315	0	**315**
FEVEREIRO				0	0	-
MARÇO				0	0	-
ABRIL				0	0	-
MAIO				0	0	-
JUNHO				0	0	-
JULHO				0	0	-
AGOSTO				0	0	-
SETEMBRO				0	0	-
OUTUBRO				0	0	-
NOVEMBRO				0	0	-
DEZEMBRO				0	0	-

Figura 22 Linhas Regulares, página 11 do Anuário de 2020 da SEMOB

Inicialmente precisamos localizar a página com a tabela que queremos extrair, no caso, a página 11. Em seguida obtemos o nome da primeira coluna, ou seja, "LINHAS REGULARES", e o conteúdo da última linha desta mesma coluna, "DEZEMBRO". Esses dados permitirão a seleção da página e a extração da tabela através dos recursos da biblioteca *pdftools*.

Após a extração é sempre importante verificarmos o seu conteúdo, assim transformamos em um *tibble*, um tipo de *dataframe* que exibe os seus 10 primeiros itens, assim como os seus tipos, neste exemplo temos três colunas do tipo “chr”, ou seja “caractere” e uma do tipo “dbl”, “double”, como mostrado na figura 23.

O tipo de “num_linhas” precisa ser convertido para “inteiro”, para que possamos efetuar os cálculos e análises. A partir deste *dataframe* criamos um arquivo CSV, que será utilizado para as operações de criação de nossa base dados em conjunto com os dados dos demais anos que serão utilizados para a geração e alimentação dos gráficos.

Através dos recursos do *pdftools* foi possível extrair uma parte dos dados a serem utilizados no dashboard. Umas das limitações desta biblioteca está em páginas com mais de uma tabela com os mesmos nomes na primeira coluna, nesse caso, somente a primeira tabela seria detectada.

```
> tb <- as_tibble(tabela)
> tb
# A tibble: 36 × 4
   mes   operadora num_linhas   ano
   <chr> <chr>     <chr>      <dbl>
 1 jan   CSN       110         2020
 2 fev   CSN       0           2020
 3 mar   CSN       0           2020
 4 abr   CSN       0           2020
 5 mai   CSN       0           2020
 6 jun   CSN       0           2020
 7 jul   CSN       0           2020
 8 ago   CSN       0           2020
 9 set   CSN       0           2020
10 out   CSN       0           2020
# … with 26 more rows
# i Use `print(n = ...)` to see more rows
```

Figura 23: Conteúdo extraído da tabela "Linhas Regulares".

Uma alternativa seria dividir a página digitalmente para então proceder a extração, em todo caso, isso reforça a necessidade de uma política de abertura de dados em formatos de fato abertos, pois, no caso citado, seriam necessários diversos procedimentos para o acesso digital aos dados e informações contidos nestas páginas.

Tabelas com agrupamento de mais de uma variável também tem sua extração dificultada. Ainda existem casos onde uma variável é desdobrada naqueles que seriam os seus valores, resultando na necessidade de uma manipulação de forma a melhor estruturar os dados.

Uma situação encontrada, ainda nesse processo de extração a partir de um arquivo PDF, foi a falha de codificação do próprio PDF, onde ao selecionarmos e copiarmos um texto, os caracteres vêm truncados, impossibilitando a leitura de seu conteúdo – isso ocorreu com o Anuário da SEMOB de 2021, onde foi preciso efetuar a coleta de forma manual dos dados ali existentes. Neste caso, os dados foram inseridos em uma planilha do LibreOffice e em seguida salvos no formato CSV – nesta etapa é importante atentar para o formato dos dados numéricos – se a casa decimal separada por ponto ou por vírgula –, de forma a evitar perdas de dados, assim como o elemento separador dos dados – espaço, tabulação, vírgula ou ponto e vírgula.

O mesmo cuidado se dá no processo de importação de arquivos CSV, onde se deve indicar qual dos elementos citados é o separador. Assim, como um fator a mais de segurança, tendo em vista manter a integridade dos dados, sempre que importamos um arquivo CSV é preciso conferir se o resultado da importação está de acordo com o desejado, o que pode ser feito mediante uso do *tibble*.

A ARQUITETURA DA SOLUÇÃO

A arquitetura básica de uma aplicação Shiny é bastante enxuta. Neste projeto a estrutura é composta pela interface de usuário – arquivo ui.R –, o arquivo server.R e um diretório onde estão os arquivos CSV (figura 24).

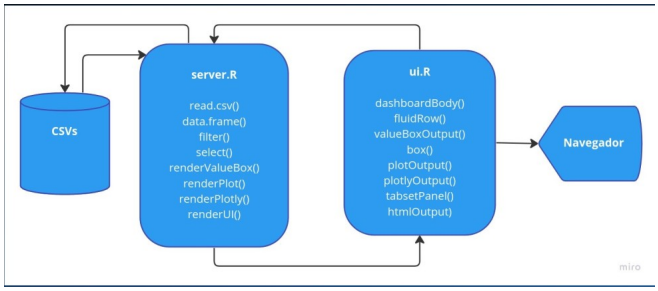

Figura 24: A arquitetura da aplicação (fonte: do autor).

Ao executarmos a aplicação, o servidor busca os arquivos CSV, efetua o seu carregamento, processamento, apresentando os resultados na interface de usuário.

Para a composição do layout do *dashboard*, adotamos o componente *fluidRow* que cria linhas que são divididas em colunas. Cada linha tem um comprimento de 12 unidades que pode ser dividido de várias formas, por exemplo três boxes de 4 unidades, dois boxes de 6, entre outras opções, sendo o único requisito que a soma não ultrapasse 12 unidades.

O *fluidRow* também permite a responsividade da página, dimensionando e organizando automaticamente, em tempo real, a disposição dos gráficos de acordo com o tamanho ou formato da tela ou janela do navegador.

A interface de usuário está organizada da seguinte forma:

- valueBoxOutput("cidade")
- valueBoxOutput("linhas")
- valueBoxOutput("operadoras")
- valueBoxOutput("passageiros")
- valueBoxOutput("viagens")
- valueBoxOutput("quilometragem")
- valueBoxOutput("cicloviaria")
- valueBoxOutput("pnb")
- valueBoxOutput("infra_apoio_ciclista")

O componente *valueBoxOutput*, como seu nome sugere, criam caixas nas quais são apresentadas valores (informações) (Figura 25), inclusive com a possibilidade de atualização automática em tempo real, permitindo a visualização de dados à medida que estes estão sendo recebidos pela aplicação, embora este recurso não tenha sido implementado neste modelo, pois as informações são atualizadas anualmente. A escolha do parâmetro "cidade" teve por objetivo permitir a escalabilidade do sistema para outras cidades.

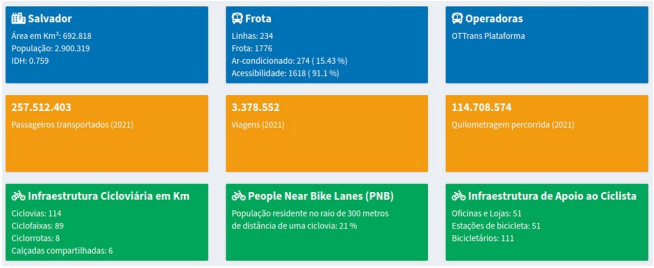

Figura 25: Apresentação de dados via *valueBoxOutput*

Em seguida temos nove componentes do tipo *box* com sete saídas interativas do tipo plotlyOutput, e uma estática do tipo *plotOutput*. e mais um *box* com saída em texto html. Uma das saídas interativas foi estruturada via componente *tabsetPanel*, que envolve seis componentes *tabPanel*. Todos os gráficos são estruturados via pacote *ggplot2* e renderizados no servidor pela função *renderPlot e renderPlotly* – estático e interativos, respectivamente –, em seguida exibidos via interface de usuário ui.R no navegador.

O primeiro é um *plotOutput* é responsável pela exibição de um gráfico no modelo pizza (figura 26) com os valores absolutos e percentuais da frota (média mensal) de cada operadora.

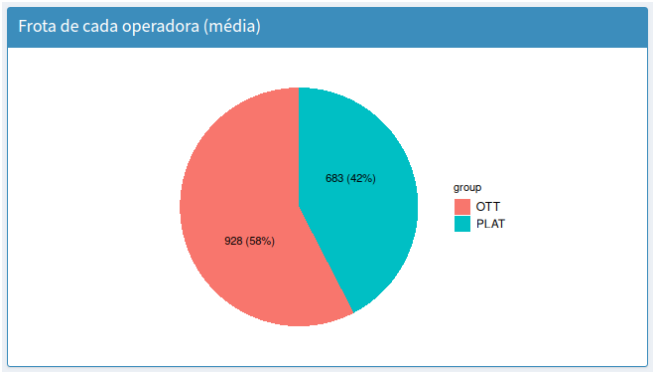

Figura 26: Número de veículos de cada operadora

Um plotlyOutput exibe um gráfico com o número médio das linhas por operadoras, de 2018 a 2021 (figura 27).

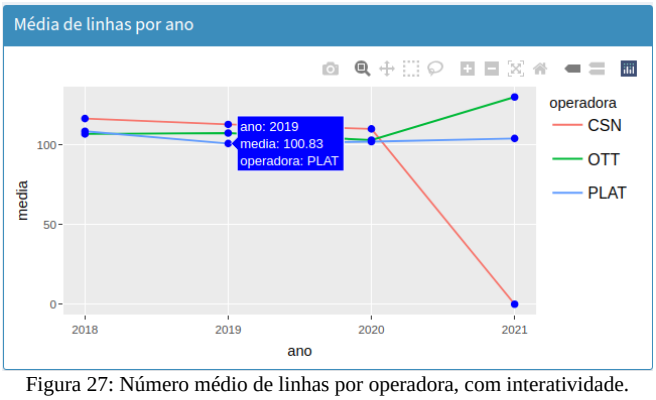

Figura 27: Número médio de linhas por operadora, com interatividade.

O próximo é um gráfico de barras (figura 28) mostrando a média de idade da frota de 2018 a 2021.

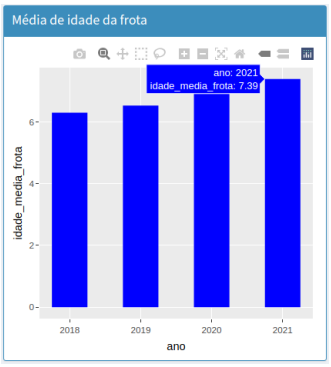

Figura 28: Média de idade da frota de ônibus em Salvador.

Em seguida temos um gráfico de barras horizontais com os principais corredores rodoviários, de acordo com a SEMOB – são 23 ao todo –, agrupados por ano e divididos em seis abas. O indicador apresentado neste gráfico é a quantidade de ônibus por hora (figura 29), nos últimos quatro anos.

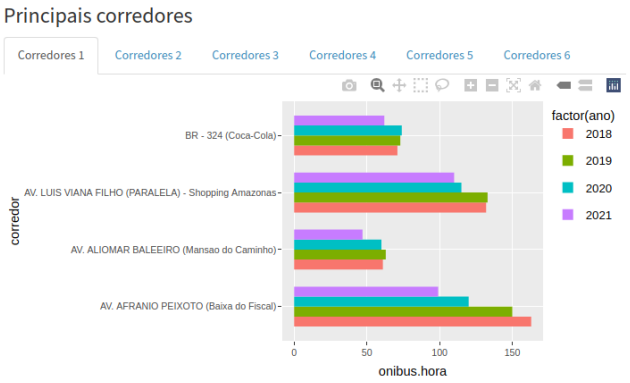

Figura 29: Quantidade de ônibus por hora nos principais corredores.

O número total de corredores foi dividido em seis abas, de forma a evitar uma saturação de informações.

A escolha pela posição horizontal das barras foi motivada pela necessidade de mostrar os nomes dos corredores sem abreviações. O agrupamento por corredor com diferenciação dos anos por cores permitiu uma economia de espaço ao mesmo tempo em que possibilita, logo à primeira vista, identificar as variações deste indicador ao longo dos anos.

O próximo gráfico apresenta a variação da frota operante ao longo de nove anos (figura 30).

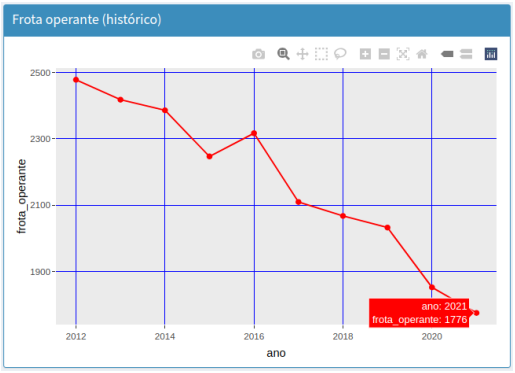

Figura 30: Variação da frota total operante em nove anos.

As aquisições de ônibus novos nos últimos vinte anos podem ser vistas na figura 31.

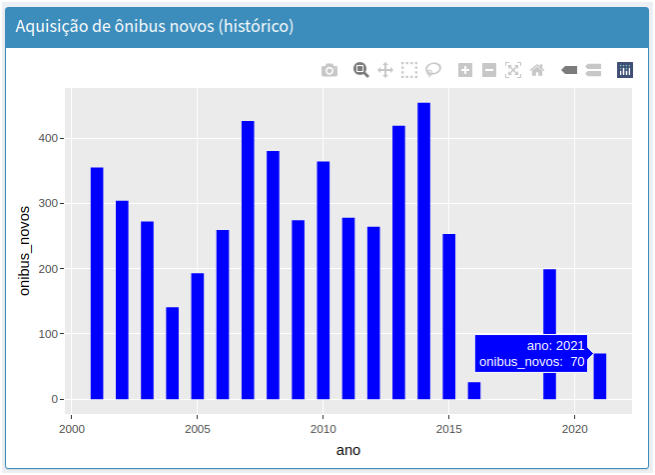

Figura 31: Variação da aquisição de ônibus novos em 21 anos.

A variação da quantidade de passagens de ônibus que podem ser compradas com um salário-mínimo ao longo dos últimos dez anos, está representada na figura 32.

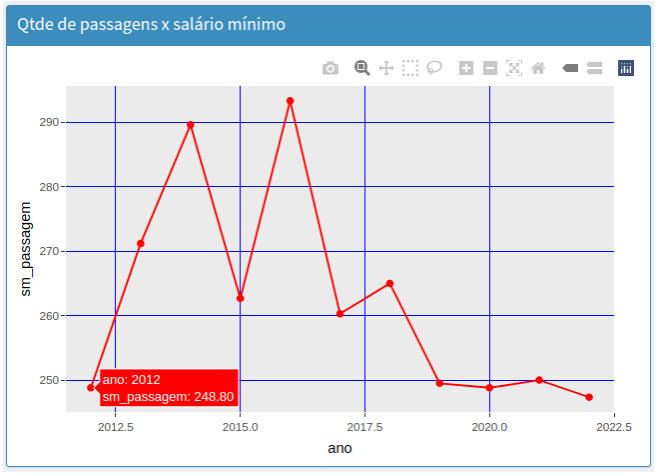

Figura 32: Variação da quantidade de passagens compradas com um salário-mínimo.

Em seguida, a quantidade de passagens de ônibus que podem ser compradas com a renda mensal domiciliar *per-capita* do Estado da Bahia, na figura 33.

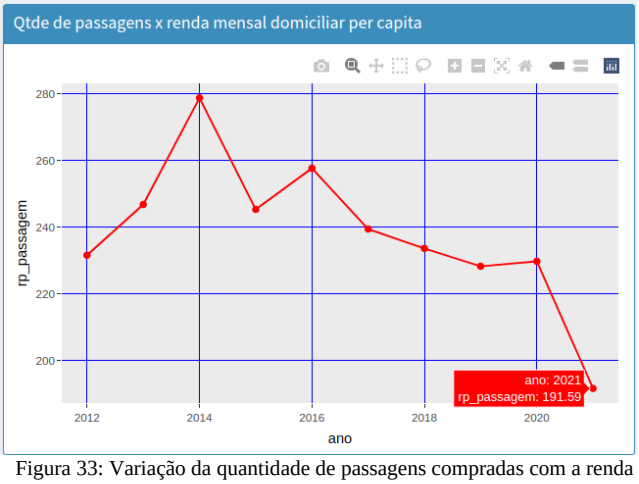

Figura 33: Variação da quantidade de passagens compradas com a renda domiciliar *per-capita* no Estado na Bahia.

A terminologia, os créditos e as fontes dos dados e informações utilizados para a elaboração deste *dashboard* podem ser vistos na figura 34.

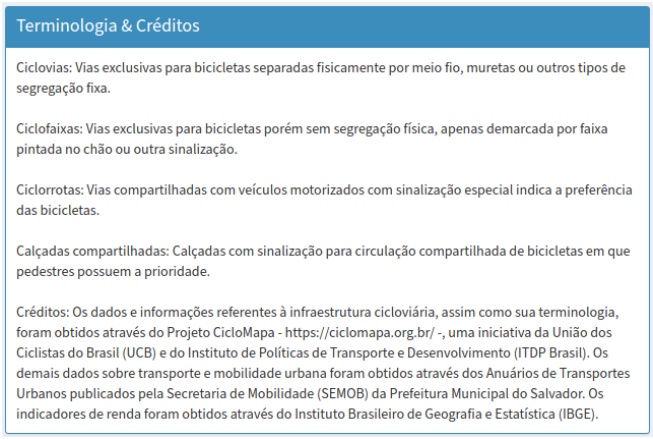

Figura 34: Terminologia, fontes e créditos.

DIFICULDADES ENCONTRADAS

Neste projeto, no quesito “Acesso a Dados”, foi preciso acionarmos a Lei de Acesso à Informação [50], tendo que aguardar não apenas o prazo legal inicial de 20 dias, mas, também o de 10 dias da prorrogação para que as solicitações que fizemos fossem atendidas.

Vale observar que se tratam de **anuários**, documentos criados a cada ano e de interesse público, não contendo nenhuma informação sensível ou sigilosa, podendo, desta forma, serem publicado livremente sem quaisquer problemas legais.

Na etapa de Preparação, identificamos dados faltantes na tabela “Linhas Regulares” dos anuários de 2020 e 2021. Até 2019 constavam nestas tabelas a quantidade de linhas a cada mês do ano. Como não havia tempo hábil para uma nova solicitação via Lei de Acesso à Informação – no pior dos casos levaria ao menos 30 dias –, optamos por trabalhar com a média dos valores de 2018 e 2019.

Ainda nesta etapa, identificamos que o anuário de 2021 estava com a sua codificação corrompida, o que impossibilitou a extração de dados de forma automatizada. Foi utilizada uma planilha do LibreOffice para gerar os arquivos CSVs.

CONSIDERAÇÕES

Como vimos, os dados encontrados referentes ao cenário de mobilidade urbana em Salvador, fornecidos pela SEMOB, eram de natureza não estruturada, se encontrando no formato PDF. Em sua grande maioria, compostos por variáveis quantitativas do tipo discreto, seguidas por quantitativas contínuas e, por fim, qualitativas nominais.

A partir destes dados foi possível efetuar análises descritivas – o que aconteceu e está acontecendo –, referentes ao panorama de mobilidade urbana na capital baiana. Também pudemos, em menor escalar e de forma parcial, entender o porquê (análise diagnóstica) da situação atual.

Uma análise preditiva, a princípio, não seria possível neste momento, pois, envolvem diversos aspectos, notadamente no campo da macroeconomia e da política, que podem passar por mudanças imprevisíveis.

Os conceitos da Gestalt, por sua vez, possibilitaram a organização dos dados e informações de forma a tornar a experiência de visualização ágil, evitando uma sobrecarga cognitiva. Sob essa perspectiva, utilizamos os conceitos listados a seguir nos gráficos:

- Figura 26: proximidade, similaridade e acercamento;
- Figura 27: proximidade e acercamento;
- Figura 28: similaridade e conexão;
- Figura 29: similaridade;
- Figura 30: proximidade;
- Figura 31: similaridade e conexão;
- Figura 32: similaridade;
- Figura 33: similaridade e conexão;
- Figura 34: similaridade e conexão;
- Figura 35: acercamento;

RESULTADOS ALCANÇADOS

Com este trabalho foi possível produzir uma visão geral do quadro do transporte público por ônibus em Salvador. Conseguimos organizar, agregar e visualizar dados para obter uma compreensão da dinâmica do sistema de transporte público em Salvador, mas, novas perguntas surgiram durantes esse processo.

Quem se locomove em Salvador através de transporte público tem notado um maior tempo de espera nos pontos de ônibus ao longo dos anos, o que ficou demonstrado pela redução seguida dos valores ônibus/hora nos principais corredores nestes últimos 4 anos, conforme apresentado no gráfico “Principais corredores”.

Entretanto, identificamos sete exceções no ano de 2019, nos corredores:

- Av. Luis Viana Filho (Paralela) Shopping Amazonas;
- Av. Aliomar Baleeiro (Mansão do Caminho);
- A. S. Marcos AV. S. Rafael (Hospital São Rafael);

- Av. Mario Leal Ferreira (Bonocô) Escola da Bíblia;
- Av. Centenário (Hospital Santo Amaro);
- Rua Carlos Gomes (Igreja Universal);
- Av. Joana Angelica (Colégio Central);

Nestes corredores o fluxo ônibus/hora aumentou. É importante lembrar que neste ano foram adquiridos 199 ônibus novos, assim, essas duas variáveis podem estar relacionadas. Nos anos seguintes, a redução se deu possivelmente em razão das medidas de isolamento e da crise do transporte público que já se fazia presente, mesmo antes da pandemia de Covid-19.

Essa diminuição do número de ônibus/hora resulta, obviamente, em um maior tempo de espera nos pontos de ônibus, aumentando o tempo de deslocamento como um todo, maior exposição ao calor, além do fator segurança, em especial durante a noite, impactando na qualidade de vida da população que utiliza o transporte público.

Outro aspecto que chamou a atenção foi a média de idade da frota em 2021 que está 7.39 anos, sendo que a vida útil dos ônibus, segundo informação contida nos próprios anuários é de, no máximo, 7 anos. No ano de 2020, 52,31 % da frota tinha mais de 7 anos de uso [52], em 2021 este percentual subiu para 69,98% [53], um aumento de 33,78%. Veículos nessas condições podem apresentar uma redução do nível de conforto, diminuindo a qualidade da experiência de uso do transporte público, afetando também a qualidade de vida dos usuários.

É importante acrescentar que, de acordo com o Termo de Compromisso de Ajustamento de Conduta Nº 35/2009 – obtido via Lei de Acesso à Informação –, do Grupo de Atuação Especial de Defesa do Patrimônio Público e da Moralidade Administrativa (GEPAM) do Ministério Público do Estado da Bahia, encontramos na página 9:

II — A idade média da frota prevista contratualmente será objeto, nas próximas revisões tarifárias, de definição pelo Poder Concedente, no intervalo entre 3,5 anos e 5 anos, em conformidade com o projeto de investimentos que vier a ser considerado na revisão tarifária, sendo utilizados no modelo de remuneração do contrato os parâmetros fixados nos estudos realizados, referidos no ANEXO 01, mediante as seguintes medidas provisórias:

Seguimos no aguardo de uma definição quanto essa questão. No mesmo documento, ainda na página 9, encontramos:

a) até dezembro de 2022, os Contratos de Concessão remunerarão veículos com até 10 (dez) anos de uso;

Se por um lado essa cláusula determina um fator condicionante, fixando um limite para a remuneração de veículos em função de uma idade máxima, o que é muito positivo, por outro lado essa idade máxima de 10 anos se encontra 42,85% acima do fator vida útil de 7 anos.

Ainda no mesmo documento, sobre a aquisição de ônibus climatizados:

b) as Concessionárias adquirirão, a cada ano, no mínimo 250 novos ônibus (0 km) com ar-condicionado, totalizando, até dezembro de 2022, ao menos 1.000 novos ônibus com ar-condicionado;

c) a partir do exercício de 2019, todos os veículos novos, O km, que passarem a operar no STCO (Sistema de Transporte Coletivo por Ônibus de Salvador), deverão ser equipados com ar-condicionado;

Entretanto, segundo o Anuário de Transportes Urbanos de Salvador referente ao ano de 2022, o total de ônibus climatizados é de 383, correspondendo a somente 21,4% da frota, ou seja, um aumento de 38,6% em relação ao ano anterior. É importante lembrarmos que a pandemia da COVID 19 certamente impactou no cumprimento do que fora acordado, entretanto, sigamos acompanhando o processo de aquisição de novos veículos.

Também notamos nos anuários a ausência de indicadores sobre:

- Ciclovias;
- Mobilidade ativa;
- Passarelas;
- Segurança;
- Velocidade das vias;
- Legendas ou explicações sobre os termos e siglas.

Embora parte desses itens, a exemplo das ciclovias e mobilidade ativa, constem da proposta do PlanMob Salvador, divulgado pela Prefeitura de Salvador em 2018.

PERSPECTIVAS PARA ESTE PROJETO

Nosso objetivo, tanto sob a ótica da pesquisa quanto de aprendizado, é continuar acompanhando os anuários e aperfeiçoando este projeto. Atualmente consideramos a possibilidade de implementar as funcionalidades:

- Mostrar o percentual de ônibus novos;
- O percentual de redução ônibus/hora;
- A velocidade das vias;
- Acidentes no trânsito – tipologia, localização, data e hora;
- Reclamações;
- Segurança;

- Geração de relatórios;
- *Download* dos *datasets;*
- Criação de uma API

CONCLUSÃO

Os dados e informações aqui apresentados, como dissemos, foram obtidos a partir de arquivos em formato não estruturados, o que nos levou a efetuar um trabalho de engenharia de dados, nem sempre possível de forma automatizada. Esses dados por serem públicos e de interesse também público deveriam estar disponíveis na página da SEMOB, em formatos abertos, o que permitiria seu livre uso por pesquisadores de uma forma bem mais ágil e amigável.

Vimos com a iniciativa do consórcio W3C, um conjunto de melhores práticas que poderiam ser utilizadas pelos gestores públicos quando do processo de abertura de dados, e seus diversos benefícios, assim como tivemos conhecimento, através do Tribunal de Contas da União, dos cinco motivos para a abertura de dados governamentais, algo que, por si só, resumem a importância dessa política. Entretanto, como apresentado pela pequisa feita pela Fundação Getúlio Vargas e a OpenKnowledge Brasil, a cultura de dados abertos governamentais em nosso país se encontra um tanto quanto incipiente, em especial em nossa cidade.

A mudança dessa condição passa não apenas pela conscientização e compromisso do gestor público, mas, também, em função de uma cobrança da sociedade como um todo e, principalmente, por parte dos pesquisadores e da academia.

É importante lembrar que o trabalho de pesquisa pode inclusive enriquecer esses dados, oferecendo análises e *insights* igualmente passíveis de serem utilizados pelos gestores públicos, no desenvolvimento de programas sociais, projetos, campanhas e políticas públicas, dispondo, assim, de mais subsídios técnicos para o planejamento estratégico e a tomada inteligente de decisões. De fato, é precisamente isso que as empresas privadas vêm fazendo, aumentando, assim, a sua competitividade.

Se uma empresa pública, ao menos internamente, não se encontra numa situação de competição, isso não impede que o processo decisório ocorra de forma a agregar valor aos seus produtos e serviços – inclusive de forma contínua. Lembremos que um dos setores responsáveis pelo aumento da competitividade de uma empresa é justamente o de Pesquisa e Desenvolvimento. Assim, dispor de apoio sob a forma de estudos feitos por pesquisadores e centros de pesquisa acadêmica, pode constituir um importante diferencial na qualidade de um processo de gestão pública, resultando em benefícios tanto na dinâmica processual de planejamento e decisão, bem como para a sociedade como um todo.

Esperamos que este trabalho seja uma amostra do que a academia pode realizar; uma modesta contribuição para o entendimento de nossa realidade, de forma a buscarmos e propormos mudanças tendo em vista a construção de uma sociedade mais participativa, mais democrática e mais cidadã.

AGRADECIMENTOS

Nossos mais sinceros agradecimentos a todos e todas que contribuíram, direta ou indiretamente, para realização deste trabalho. Minha esposa Alessandra e meus filhos João Vitor, Rafaella e Pedro, os amigos e colegas Márcia Teles, Christiane Veigga, Rafaela Fernandes, Marília Libório, Neri, Damián Meneses. Aos professores Cláudio Amorim (UNEB) pela amizade, conselhos e troca de ideias, Sandro Andrade (IFBA), Romilson Sampaio (IFBA), e por fim ao mestre e orientador Pablo Florentino pelo apoio, compreensão, conselhos e orientações sempre precisas e pertinentes. Um forte abraço a todos e todas.

REFERÊNCIAS

[1] J. M. Rodrigues, “Qual o estado da mobilidade urbana no Brasil?”, Mobilidade Urbana no Brasil: Desafios e Alternativas, https://br.boell.org/sites/default/files/mobilidade_urbana_boll_brasil_web_.pdf (acessado Março 25, 2022).

[2] *Instituto de Pesquisa Econômica Aplicada, “Relatório brasileiro para a habitat III,” in Relatório Brasileiro para A Habitat III. IPEA,* https://habitat3.org/wp-content/uploads/National-Report-LAC-Brazil-Portuguese.pdf , *2016.*

[3] Bartelt, Dawid. Paula, Marilene de. “Qual o estado da mobilidade urbana no Brasil?” Mobilidade Urbana no Brasil: Desafios e Alternativas, 2016 (acessado Março 25, 2022).

[4] Princípios da Gestalt em UI Design. https://grandecircular.medium.com/princ%C3%Adpios-da-gestalt-em-ui-design-38148814e5ee. (acessado Março 25, 2022).

[5] IBGE, “Panorana,” https://cidades.ibge.gov.br/ , 2022 (acessado Março 25, 2022).

[6] Prefeitura Municipal de Salvador, “Relatório técnico rt14: Plano de mobilidade urbana sustentável de salvador tomo i,” http://www.mobilidade.salvador.ba.gov.br/documentos/RT_14-PlanMob_SSA-TOMO_I.pdf , 2018 (acessado Março 25, 2022).

[7] Secretaria de Mobilidade de Salvador. http://www.mobilidade.salvador.ba.gov.br/, (acessado Novembro 10, 2022)

[8] Cartilha Técnica para Publicação de Dados Abertos no Brasil v1.0 Secretaria de Logística e Tecnologia da Informação - SLTI Ministério do Planejamento Orçamento e Gestão - MP https://wiki-dados-h.cgu.gov.br/GetFile.aspx?Page=Tecnologia&File=Cartilha%20T%c3%a9cnica%20para%20Publica%c3%a7%c3%a3o%20de%20Dados%20Abertos%20no%20Brasil%20v1.pdf. 2012. (acessado Março 25, 2022).

[9] OpenKnowledge Fundation Brasil. https://ok.org.br/. (acessado Março 25, 2022).

[10] T. de Contas da União, “5 motivos para a abertura de dados na administração pública,” https://portal.tcu.gov.br/5-motivos-para-a-abertura-de-dados-na-administracao-publica.htm , 2015 (acessado Março 25, 2022)

[11] F. G. V. e Open Knowledge Brasil, “Índice de Dados Abertos para cidades,” https://ok.org.br/wp-content/uploads/2020/04/WEB-Índice-de-dados-abertos-1.pdf , 2018 (acessado Março 25, 2022).

[12] I. C. Education. Application programming interface (api). [Online].Available: https://www.ibm.com/cloud/learn/api (acessado Outubro 25, 2022)

[13] O que é SDK? Qual a diferença entre SDK e API? - https://programadoresdepre.com.br/o-que-e-sdk-qual-a-diferenca-entre-sdk-e-api/ (acessado Novembro 12, 2022)

[14] Internet World Stats – https://www.internetworldstats.com/stats.htm , 2022. (acessado Novembro 10, 2022)

[15] Rautenberg, Sandro. Guia prático para publicação de dados abertos conectados. 2018.

[16] World Wide Web Consortium (W3C), https://www.w3.org/ , 2017. (acessado Novembro 10, 2022)

[17] Data on the Web Best Practices – https://www.w3.org/TR/dwbp/#bestPractices, 2017.(acessado Novembro 10, 2022)

[18] Bus stops of MyCity – https://www.w3.org/TR/dwbp/dwbp-example.html , 2015.(acessado Novembro 10, 2022)

[19] Examples – https://www.w3.org/TR/dwbp/dwbp-example.ttl , 2015. (acessado Novembro 10, 2022)
[20] Data Catalog Vocabulary (DCAT) - Version 2 – https://www.w3.org/TR/vocab-dcat/ , 2020. (acessado Novembro 10, 2022).
[21] Lóscio, Bernadette. Burle, Caroline. Calegari, Newton. Data on the Web Best Practices: Challenges and Benefits. https://acervo.ceweb.br/acervos/conteudo/7294e1c6-813d-4d8e-b84f-76d2332373ab, Open Data Research Symposium, 2016. (acessado Novembro 10, 2022)
[22] Nuvem de Dados Abertos Conectados desenvolvida pelo projeto LOD Cloud Fonte: "The Linking Open Data Cloud Diagram", disponível em https://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.svg. (acessado Novembro 10, 2022).
[23] W3C Brasi, https://www.w3c.br/ , 2022. (acessado Novembro 10, 2022).
[24] Sakr, Sherif. Zomaya, Albert. Encyclopedia of Big Data Technologies. Springer, Switzerland, 2019. (acessado Novembro 10, 2022)
[25] O que é ETL – Extract Transform Load? https://diogonvidal.wixsite.com/powercenter/post/o-que-%C3%A9-etl-extract-transform-load , 2018. Acessado Novembro 10, 2022)
[26] Mitchell, Tm. The What, Why, When, and How of Incremental Loads – https://www.timmitchell.net/post/2020/07/23/incremental-loads , Acesso em: novembro, 2022.
[27] Reis, Joe. Housley , Matt. Fundamentals of Data Engineering - Plan and Build Robust Data Systems, CA, USA, 2022.
[28] Boscarioli, Clodis. da Silva, Leandro Augusto. Peres, Sarajane Peres. Introdução à Mineração de Dados com Aplicações em R. 2016, Rio de Janeiro
[29] Castro, Leandro Nunes de, et al. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo, Saraiva, 2016.
[30] Massena, Cristiane. Segurança da Informação e Governança de Dados: parceria benéfica para todos. https://abracd.org/seguranca-da-informacao-e-governanca-de-dados-parceria-benefica-para-todos/. (acessado Novembro 10, 2022).
[31] Pickell, Devin. https://www.g2.com/articles/structured-vs-unstructured-data, (acessado Novembro 10, 2022).
[32] Neto, José Antônio Ribeiro. Análise de Dados em Big Data. https://medium.com/xnewdata-portugal/aula-11-an%C3%A1lise-de-dados-em-big-data-a71cfdba290, (acessado Novembro 10, 2022).
[33] Hoppen, Joni. Santos, Marco. Análise descritiva, preditiva, prescritiva e cenarização: Como gerar valor nos negócios. https://www.aquare.la/analise-descritiva-preditiva-prescritiva-e-cenarizacao/, (acessado Novembro 10, 2022).
[34] A. e. N. O. e. S. M. e. B. E. Pandey, Anshul Vikram e Manivannan, “The persuasive power of data visualization,” in IEEE Transactions on Visualization and Computer Graphics. IEEE, 2014.
[35] E. R. Tufte, The Visual Display of Quantitative Information, ser. International series of monographs on physics. Graphics Press, 2001.
[36] A. B. de Gestalt-Terapia, “O que é gestalt terapia,” https://gestalt.com.br/institucional/gestalt-terapia/, 2022. (acessado Novembro 10, 2022).
[37] C. N. Knaflic, Storytelling com Dados. Graphics Press, 2019. (acessado Novembro 10, 2022).
[38] Os 7 Princípios de Gestalt e Como Utilizá-los em Projetos de UI Design. https://aelaschool.com/designdeinteracao/os-7principios-de-gestalt-e-como-utiliza-los-em-projetos-de-ui-design/,(acessado Novembro 10, 2022).
[39] União de Ciclistas do Brasil – https://uniaodeciclistas.org.br/atuacao/ciclomapa/, (acessado Novembro 10, 2022).
[40] Mobilidados – https://mobilidados.org.br/, (acessado Novembro 10, 2022).
[41] Ciclomapa – https://ciclomapa.org.br/, (acessado Novembro 10, 2022).
[42] OpenStreetMap. https://www.openstreetmap.org/ (acessado Novembro 10, 2022).
[43] SIMU: Infraestrutura de Mobilidade Urbana – https://simu.mdr.gov.br/?page_id=43, (acessado Novembro 10, 2022).
[44] R Packages (2e) https://r-pkgs.org/Introduction.html, 2022. (acessado Novembro 10, 2022)
[45] RStudio – Posit. https://posit.co/, (acessado Janeiro 17, 2023)
[46] Package ‘pdftools’ – https://cran.r-project.org/web/packages/pdftools/pdftools.pdf, 2022. (acessado Novembro 10, 2022)
[47] pdftools: Text Extraction, Rendering and Converting of pdf Documents – https://cran.r-project.org/web/packages/pdftools/index.html, 2022. (acessado Novembro 10, 2022)
[48] tidyverse: Easily Install and Load the 'Tidyverse' – https://cran.r-project.org/web/packages/tidyverse/index.html, 2022. (acessado Novembro 10, 2022)
[49] Ciência de Dados em R, capítulo 7 – Curso-R. https://livro.curso-r.com/7-manipulacao.html, (acessado Novembro 11, 2022)
[50] LEI Nº 12.527, DE 18 DE NOVEMBRO DE 2011. https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. (acessado Novembro 10, 2022).
[51] Transferência de linhas da CSN para outras operadoras será concluída nesta quinta (30), diz prefeitura de Salvador, https://diariodotransporte.com.br/2021/09/29/transferencia-de-linhas-da-csn-para-outras-operadoras-sera-concluida-nesta-quinta-30-diz-prefeitura-de-salvador/ (acessado Dezembro, 2022)
[52] Anuário de Transportes Urbanos, pág 25. Secretaria Municipal de Mobilidade (SEMOB), Prefeitura Municipal de Salvador, 2020.
[53] Anuário de Transportes Urbanos, pág 32. Secretaria Municipal de Mobilidade (SEMOB), Prefeitura Municipal de Salvador, 2021.
[54] Wickham, Hadley. R para Data Science. Alta Books, Rio de Janeiro. 2019.
[55] Schmuller, Joseph. Projetos em R para Leigos. Alta Books, Rio de Janeiro. 2019.
[56] Alcoforado, Luciane F. Utilizando a Linguagem R. Alta Books, 2021.
[57] Isotani, Seiji. Dados abertos conectados. São Paulo. Novatec Editora, 2015.
[58] Wijaya, Adi. Data Engineering with Google Cloud Platform – A practical guide to operationalizing scalable data analytics systems on GCP. Birmingham. Packt Publishing. 2021.
[59] EMC Education Services. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Indianapolis. Wiley. 2015.
[60] Kroese, Dirk P. et Al. Data Science and Machine Learning - Mathematical and Statistical Methods. 2022