

Projeto Integrado

Análise de Acidentes no Estado de São Paulo

HISTÓRICO DE REVISÃO

Versão	Descrição da Mudança	Autor	Data
1.0	Criação da estrutura do documento e formatação.	Tatiane Conceição Alves	17/ 08 / 2021
2.0	Definição de parâmetros	Tatiane Conceição Alves	00 / 08 / 2021
3.0	Inserção das etapas de parte do processo	Tatiane Conceição Alves	03 / 09 / 2021
4.0	Cada participante incluiu imagens e referências	Tatiane Conceição Alves	08 / 09 / 2021
5.0	Finalização de formatação	Tatiane Conceição Alves	10 / 09 / 2021

IDENTIFICAÇÃO DOS ENVOLVIDOS

Papel	Nome	E-mail
Líder	Daniela Vieira Ferreira Domingues	daniela.domingues@blueshift.com.br
Trainee	Antony Cassiano Bueno	antony.bueno@blueshift.com.br
Trainee	Flávio de Salles Caló	flavio.calo@blueshift.com.br
Trainee	Tatiane Conceição Alves	tatiane.alves@blueshift.com.br
Trainee	Gilberto Kiyoshi Abe	gilberto.abe@blueshift.com.br
Trainee	Elias Isaac Salas Contreras	elias.contreras@blueshift.com.br

SUMÁRIO

1	VISÃO GERAL DO PROJETO	4
2	FONTE DE DADOS - DATASETS	4
Resp	peito a vida	4
Kag	gle	5
3	FERRAMENTAS	5
4	ARQUITETURA DE DADOS	7
5	PROCESSO ETL	7
6	DATA FACTORY	7
Pip :	Master	8
Pip_	Web_Scrapping	9
Pip '	Tratamento	10
7	DATABRICKS	11
Lib	- Pasta de Bibliotecas	12
Mou	ib - Pasta de Bibliotecas	
Web	Scraping	12
Trat	amento de dados	13
8	SQL SERVER	13
9	POWER BI	19
Grá	fico	19
10	CONCLUSÃO	21
11	REFERÊNCIA – BIBLIOGRAFIA	22

1 VISÃO GERAL DO PROJETO

Por conta da pandemia seguido do lockdown o comportramento dos motoristas em meio ao trânsito foi alterado bruscamente. E pra que isto se desse alguns fatos foram cruciais como as compras por delivery, colocando em pauta os vários motociclistas que foram mortos, em segundo lugar os pedestres e em terceiro lugar os passageiros. Em contrapartida os acidentes em si foram menores em relação aos anos anteriores.

As pesquisas apontam que esta mudança se deu pois, os motoristas sentiram que poderiam flexibilizar as regras de trânsito o que acabou causando este aumento e mudança do padrão das vítimas em seu contexto geral.

O nosso projeto tem como objetivo avaliar os últimos três anos 2019,2020 e 2021 no Estado de São Paulo com relação ao impacto, comportamentos e padrões nos acidentes de trânsito

2 FONTE DE DADOS - DATASETS

Respeito a vida

A primeira fonte que utilizaremos será do site 'Respeito a Vida', uma das principais iniciativas do Governo do Estado de São Paulo no qual faz a elaboração de um banco de dados abertos com informações de acidentes de trânsito do Estado.

As informações são atualizadas mensalmente, no dia dezenove ou próximo dia útil, para todos os seiscentos e quarenta e cinco municípios e tem como foco a redução de acidentes de trânsito.

Com ele, é possível traçar um diagnóstico mais preciso das características dos acidentes e, assim, planejar e estabelecer políticas públicas mais eficazes de prevenção em benefício de toda a sociedade.

Link para download: http://www.respeitoavida.sp.gov.br/relatorios/

- BASE COMPLETA DE ÓBITOS : obitos publico.csv
- BASE COMPLETA DE ACIDENTES FATAIS: acidentes fatais.xlsx
- BASE COMPLETA DE ACIDENTES NÃO FATAIS: acidentes naofatais.csv.

Kaggle

Nossa segunda fonte será o site do Kaggle, uma comunidade online de cientistas de dados e profissionais de aprendizado de máquina. O Kaggle permite que os usuários encontrem e publiquem conjuntos de dados, explorem e construam modelos em um ambiente de ciência de dados baseado na web, trabalhem com outros cientistas de dados e engenheiros de aprendizado de máquina e participem de competições para resolver desafios de ciência de dados.

Link para download:

https://www.kaggle.com/mcamera/brazil-highway-traffic-accidents datasets:

- por ocorrencia: datatran2019, dataran2020 e datatran2021.
- Por pessoa: acidentes2019.csv.zip, acidentes2020.csv.zip e acidentes2021.csv.zip.

3 FERRAMENTAS

Dataset

É um conjunto de dados que geralmente é tabulado, ele é finito e tem uma função e característica específica. usado para tratar dados que vieram do banco de dados através de alguma consulta, muitas vezes através de um comando SQL, então vem as colunas e e linhas das tabelas que você determinar. Um dataset pode ser montado manualmente ou através de um framework.

Web Scrapping

Permite a coleta de dados públicos disponibilizados em sites, rede sociais e outros serviços online. Acela a extração de dados e consulta de informações e esta rapidez se deve a linguagem de programação scripts e copia de dados em grande escala. No projeto o web scrappinp foi utilizado junto com python. A coleta de dados feita por esta ferramenta em nosso projeto está dentro da legalidade pois utilizados dados de base pública.

Blob Storage Azure

É uma ferramenta dentro do Azure para armazenamento na nuvem, é otimizado para uma grande quantidade de dados não estruturados. Ele foi criado para armazenar arquivos de acesso distribuidos, fornecimento de imagens ou documentos direto a um navegador, transmissão por streaming de audio

e vídeo, gravar arquivos de log, armazenamento de backups, restaurações, recuperações de desastres e arquivamento, armazenamento para análise por um serviço local ou hospedado no Azure.

Databricks

É uma plataforma de análise de dados aperfeiçoados, com processamento escalável, para fluxos de trabalho simplificado, workspace interativo que permite a colaboração entre os seus usuários. Unifica dados, análise e IA em uma plataforma comum para todos os casos de uso de dados, agrupa o ecossistema de dados com código aberto, padrões e formatos, colaboração com todos os dados e fluxos de IA e datafactory.

Data Factory

É o serviço de integração de dados e ETL baseado em nuvem que lhe permite criar fluxos de trabalho orientados a dados para orquestrar a movimentação e a transformação de dados em escala. Possível criar e agendar pipelines que podem ingerir dados de diferentes repositórios de dados. Cria-se processos de ETL complexos que transformam os dados visualmente com fluxos de dados ou usando serviços de computação.

Sql Azure

É criado com base no já conhecido mecanismo do SQL Server, para que você possa migrar aplicativos com facilidade e continuar usando as ferramentas, as linguagens e os recursos que você conhece. Suas habilidades e sua experiência são transferidas para a nuvem. Assim, você pode fazer ainda mais com o que você já tem.

Power BI

É uma coleção de serviços de software, aplicativos e conectores que trabalham juntos para transformar suas fontes de dados não relacionadas em informações coerentes, visualmente envolventes e interativas. Os dados podem estar em uma planilha do Excel ou em uma coleção de data warehouses híbridos locais ou baseados na nuvem. Com o Power BI, você pode se conectar facilmente a fontes de dados, visualizar e descobrir conteúdo importante e compartilhá-lo com todas as pessoas que quiser. O Power BI consiste em vários elementos que trabalham juntos, começando com estes três elementos básicos: um aplicativo de desktop do Windows chamado Power BI Desktop, um serviço SaaS (software como serviço) online chamado de serviço do Power BI e um aplicativos móveis do Power BI para dispositivos Windows, iOS e Android.

4 ARQUITETURA DE DADOS

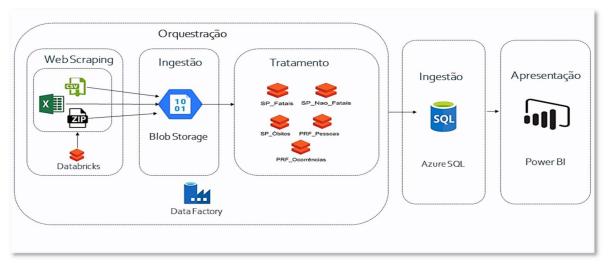


Imagem 01- Arquitetura do Projeto

A arquitetura na imagem 01 demonstra como o projeto foi estruturado e a utilização de cada uma das soluções da Microsoft.

5 PROCESSO ETL

Utilizamos o processo ETL no qual o Data Factory é orquestrador de toda a extração dos dados através de um notebook pelo databricks passando o link do download, e armazenando no Blob Storage. Todo o tratamento foi realizado também nos notebooks pelo databricks utilizando o pyspark e dataframes para as devidas tratativas e ao final envia para o banco SQL em formato de tabelas. Após as tabelas serem enviadas para o banco SQL Server, realizamos as consultas e criamos as views necessárias para a apresentação dos relatórios no Power BI.

6 DATA FACTORY

Este projeto foi elaborado através da solução de ferramentas ETL/ELT da Microsoft Data Factory considerada uma das melhores do mercado atual, totalmente gerenciada, robusta e de fácil utilização com conceitos intuitivos de arrasta e solta no qual é utilizada principalmente para ingestão de dados em bash (lotes)Pensando em como é o funcionamento de uma orquestra, o ADF neste

projeto tem a função de maestro que orquestra a execução desde o web scrapping que realiza o download dos arquivos cvs e xlsx até a ingestão dos dados no SQL devidamente tratados para a apresentação em relatórios.

O data factory possui apenas duas atividades de execução de pipeline. Veja:

Pipeline: Uma fábrica de dados pode ter um ou mais pipelines. Pipeline é um agrupamento lógico de atividades que realiza uma unidade de trabalho. Juntas, as atividades em um pipeline executam uma tarefa.

Tarefa: Um pipeline pode ter uma ou mais atividades. As Atividades definem as

Pip Master

O Pipeline "Executar o pipeline" é responsável por fazer as chamadas de outros pipelines e o principal gerenciador de todo o processo.

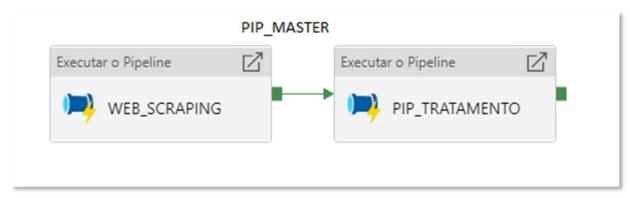


Imagem 02- PIP Master

Para a elaboração do Pipeline da imagem 03, realizamos a seguinte configuração:

PIP MASTER:

Atividade: executar pipeline.

Aba configurações:

- Pipeline invocado: PIP WEB SCRAPING

Pip_Web_Scrapping

O Pipeline na imagem 03 acima "WEB_SCRAPING", realiza a chamada do

"PIP WEB SCRAPING" no qual possui três Notebooks em Databricks conforme a imagem 04.



Imagem 03- PIP_WEB_SCRAPING

Para a elaboração do Pipeline da imagem 04, realizamos a seguinte configuração:

PIP WEB SCRAPING à atividade: Notebook Aba Geral:

- Notebook: PRF_OCORRENCIAS
- Aba Azure Databricks:
 - -nome: Is Databricks
 - -Conectar-se via runtime de integração: AutoResolveIntegrationRuntime
 - -Método de seleção de conta: Inserir Manualmente
 - -URL do Workspace do Databricks: link Databricks

Exemplo: "https://adb-00000000000.0.azuredatabricks.net"

-Tipo de Autenticação: Accss Token

Obs.: Em todos os notebooks é necessário configurar a conexão na aba Azure Databricks e na aba configurações conforme a descrição abaixo:

Notebook: PRF OCORRENCIAS

- Aba configurações:

-Caminho do notebook: /WEB SCRAPING/PRF/prf_ocorrencias

Notebook: PRF_PESSOAS

- Aba configurações:

-Caminho do notebook: /WEB SCRAPING/PRF/prf pessoas

Notebook: SP ESTADUAL

- Aba configurações:

- Caminho do notebook: /WEB SCRAPING/ESTADUAL/sp geral.

Cada notebook realiza uma solicitação utilizando a biblioteca em python "urllib", no qual é criado um dicionário que contém uma lista com a URL de download e o nome do arquivo que será baixado. Após o download, estes arquivos serão enviados para blob Storage conforme a imagem 04. Sua estrutura foi configurada em pastas conforme a imagem abaixo:

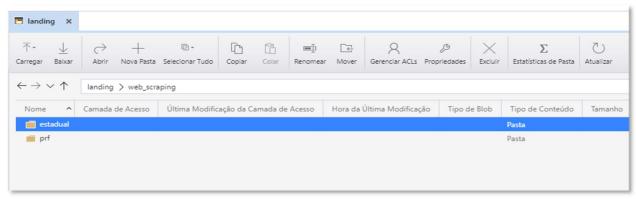


Imagem 04- Blob Storage

Para que o Databricks consiga realizar a leitura dos arquivos armazenados, é necessário a montagem do blob storage no primeiro momento. Este notebook "mount" foi criado para esta função conforme anexado na pasta do projeto Azure Data Factory.

Pip Tratamento

O pipeline de tratamento possui 5 notebooks, no qual cada um é responsável por um arquivo baixado que será transformado em uma tabela e enviado para o SQL pelo próprio notebook. Todo o código de tratamento foi realizado em pyspark.

Conforme o artigo "Funcionamento do PySpark" no site médium "O Pyspark é construído em cima da API Java, ele é apenas um fina camada de software Python que repassa as chamadas de funções para o core Java, é bastante interessante e acredito que este seja um dos motivos de sucesso

da plataforma que o código do PySpark seja tão simplório, se der uma conferida no repositório oficial do spark você pode ver por exemplo que o código de um dataframe em Python que é a principal abstração do Apache Spark tem apenas ~2000 linhas e diria que pelo menos 50% são comentários para geração de documentação."

O desenvolvimento dos códigos nos notebooks foi utilizado em formato de dataframes que nada mais é do que um conjunto de dados. Após todo o processo de tratamentos nos notebooks, como seleção de campos, linhas com informações nulas e erros de ortografias, é realizado nas 2 últimas células a conexão com o banco de dados SQL Serve que realiza a ingestão de dados em formato de tabela final.

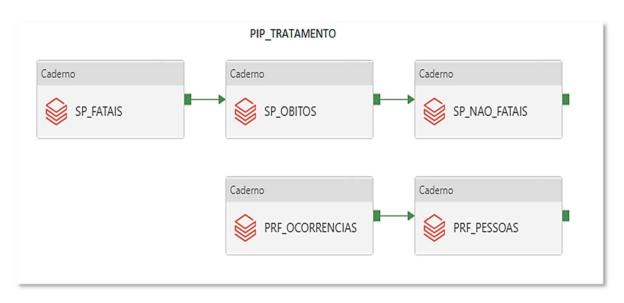


Imagem 05- PIP_TRATAMENTO

7 DATABRICKS

Configurações

Compute: Foi criado um cluster nome "projeto grupob"

Especificação: Contendo dois nós

Runtime: 9.0 (includes Apache Spark 3.1.2, Scala2.12)

Driver: Standart DS4_v2

Workspace contém 4 estrutura: Lib, Mount, Tratamento e WebScraping.

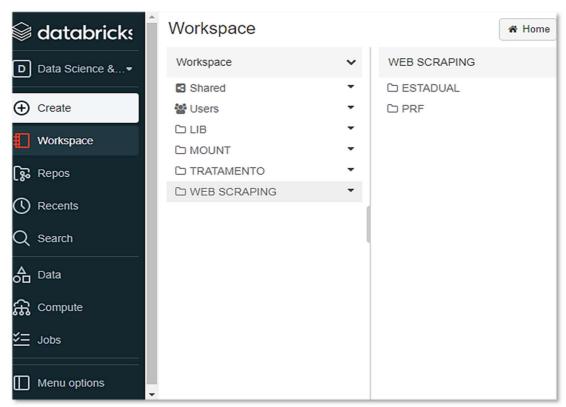


Imagem 06 - Estrutura do Workspace Databricks

Lib - Pasta de Bibliotecas

Duas bibliotecas instaladas: openpyxl e xlrd. São bibliotecas em Python, criadas para ler e escrever em arquivos do Excel 2010, cujas extensões são xlsx/xlsm/xltx/xltm.

Mount

Notebook Criado para fazer a conexão do Blob Storage com o databricks no qual serão armazenados todos os arquivos datasets a serem utilizados.

Web Scraping

Notebook responsável pelo download dos arquivos datasets. Utiliza a biblioteca "urllib", no qual é criado um dicionário contendo as urls de download e o nome do arquivo que será baixado.

Em seguida e realizado uma request, recomendado para uma interface de cliente HTTP informando o local onde será armazenado no Blo Storage.

Alguns arquivos foram baixados em formato zip, neste mesmo notebook os mesmos foram descompactados no storage e em seguida excluidos.

Tratamento de dados

Cinco notebooks (Fatais, Não Fatais, óbitos, prf_Ocorrências, prf_Pessoas). Notebooks criados para fazer a leitura do arquivo no Blob <u>Storage</u>, realiza todo o tratamento necessário com DataFrames e Pyspark que na finalização do mesmo envia para o banco SQL em formato de tabela.

8 SQL SERVER

Após a ingestão dos dados em formato de tabelas e através dos notebooks com os dataframes finais, decidimos fazer a apresentação dos relatórios em formato de views ou visões, que são consultas sql visando a performance no Power Bi. Devido a grande quantidade de dados disponíveis no banco não seria viável importar para o power bi já que o mesmo suporta o limite de apenas 1 Gb por conjunto de dados. Uma das vantagens da criação de views é que porque podemos filtrar exatamente as colunas que precisamos para o projeto na apresentação dos relatórios não sendo necessário carregar todos os dados facilitando a disponibilidade e performance.

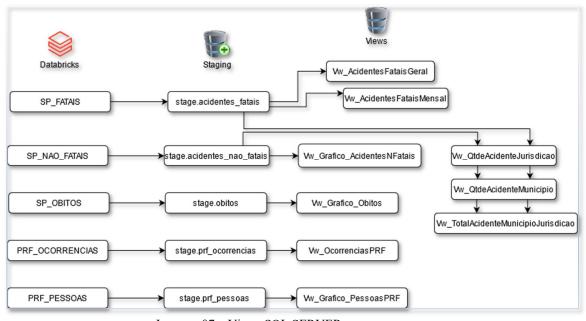


Imagem 07 – Views SQL SERVER

Vw_AcidenteFataisGeral

 Representa uma visão geral de acidentes fatais apenas com informações necessárias para o consumo do dashboard.

• Vw AcidentesFataisMensal

O Quantidade total de acidentes fatais por mês.

• Vw Grafico AcidentesNFatais

 Representa uma visão geral de acidentes não fatais apenas com informações necessárias para o consumo do dashboard.

• Vw QtdeAcidenteJurisdicao

Soma a quantidade de acidentes (fatais e não fatais) por Jurisdição buscando na view
 (Vw TotalAcidenteMunicipioJurisdicao)

• Vw QtdeAcidenteMunicipio

Soma a quantidade de acidentes (fatais e não fatais) por Município buscando na view
 (Vw TotalAcidenteMunicipioJurisdicao)

• Vw TotalAcidenteMunicipioJurisdicao

O Quantidade total de acidentes, (fatais e não fatais), por Município e Jurisdição.

• Vw Grafico Obitos

 Representa uma visão geral de óbitos apenas com informações necessárias para o consumo do dashboard.

• Vw OcorrenciasPRF

 Representa uma visão geral de ocorrências da PRF, apenas com informações necessárias para o consumo do dashboard.

Vw_Grafico PessoasPRF

 Representa uma visão geral de pessoas da PRF, apenas com informações necessárias para o consumo do dashboard.

0

Script construção das Views:

• Vw_AcidentesFataisGeral

```
ALTER VIEW [dbo].[Vw_AcidentesFataisGeral]
AS

SELECT
    logradouro
    ,convert(int,quantidade_vitimas) quantidade_vitimas
FROM
    [stage].[acidentes_fatais]
GO
```

Imagem 08 - View Acidentes Fatais Geral

• Vw_AcidentesFataisMensal

```
ALTER VIEW [dbo].[Vw_AcidentesFataisMensal]
as

select
    ano_acidente
    ,count(*) QtdeAcidente,
        convert(date, data_acidente) as data_acidente
from
    [stage].[acidentes_fatais]
group by
    ano_acidente
    ,convert(date, data_acidente)
```

Imagem 09 – View Acidentes Fatais Mensal

• Vw Grafico AcidentesNFatais

```
□ALTER VIEW [dbo].[Vw_Grafico_AcidentesNFatais]
 SELECT
       municipio
      ,jurisdicao
      ,ano_acidente
      ,CONVERT(int,pes_env_grave)
                                                         pes_env_grave
      ,CONVERT(int,pes_env_ileso)
                                                         pes_env_ileso
                                                         pes_env_leve
      ,CONVERT(int,pes_env_leve)
      , {\color{red} \texttt{CONVERT}} ( \textbf{int}, \texttt{tipo}\_ a cidente\_ a tropelamento) \\ \phantom{\textbf{convert}} \texttt{tipo}\_ a cidente\_ a tropelamento \\
      ,CONVERT(int,tipo_acidente_choque)
                                                          tipo_acidente_choque
      ,CONVERT(int,tipo_acidente_colisao)
                                                          tipo_acidente_colisao
                                                    tipo_acidente_outros
      ,CONVERT(int,tipo_acidente_outros)
 FROM
      [stage].[acidentes_nao_fatais]
 GO
```

Imagem 10 – View Acidentes Não Fatais

• Vw_QtdeAcidenteJurisdicao

```
ALTER VIEW [dbo].[Vw_QtdeAcidenteJurisdicao]

AS

SELECT
SUM(QdteAcidente) as QdteAcidente,
jurisdicao
FROM
Vw_TotalAcidenteMunicipioJurisdicao
GROUP BY
jurisdicao
GO
```

Imagem 11 - View Quantidade Acidentes Jurisdição

• Vw QtdeAcidenteMunicipio

```
AS

SELECT top 10
Municipio,
SUM(QdteAcidente) as QdteAcidente
FROM
Vw_TotalAcidenteMunicipioJurisdicao
GROUP BY
Municipio
ORDER BY QdteAcidente DESC
```

Imagem 12 – View Quantidade Acidentes Município

• Vw_TotalAcidenteMunicipioJurisdicao

```
□ ALTER VIEW [dbo].[Vw_TotalAcidenteMunicipioJurisdicao]
 AS
 SELECT
      municipio
      ,jurisdicao
      ,count(*) QdteAcidente
      [stage].[acidentes_fatais]
 GROUP BY
      municipio
      , jurisdicao
 UNION all
 SELECT
      municipio
     ,jurisdicao
      ,count(*) QdteAcidente
 FROM
      [stage].[acidentes_nao_fatais]
 GROUP BY
      municipio
      , jurisdicao
 GO
```

Imagem 13 – View Total Acidente Municipio Jurisdicao

• Vw_Grafico_Obitos

```
AS

SELECT

convert(numeric(18),id_delegacia)
,convert(date,data_obito)
,tipo_veiculo_vitima

FROM

[stage].[obitos]

GO

AS

SELECT

convert(date,data_obito)
data_obito

data_obito
```

Imagem 14 – View Gráfico Óbitos

• Vw_OcorrenciasPRF

```
AS

SELECT

convert(numeric(18),id)
,causa_acidente
FROM
[stage].[prf_ocorrencias]
GO
```

Imagem 15 – View Ocorrências PRF

• Vw_Grafico_PessoasPRF

Imagem 16 – View Gráfico Pessoas PRF

9 POWER BI

Inicialmente utilizamos a conexão por DirectQuery, onde a mesma obtém os dados com o Power Bi diretamente com o banco de dados SQL. Por um lado, significa que abrir ou atualizar um relatório ou um dashboard sempre mostra os dados mais recentes na fonte.

No entanto, volume de dados muito grande também pode indicar que o desempenho das consultas nessa fonte subjacente é muito lento ao utilizar o DirectQuery, observamos também que a conexão direta não permite a transformação de dados como a importação e as datas não apresentavam a hierarquia para filtros.

Nossa solução foi selecionar as views com os devidos campos necessários para a apresentação dos relatórios e realizar a importação no qual a performance é muito melhor.

Gráficos

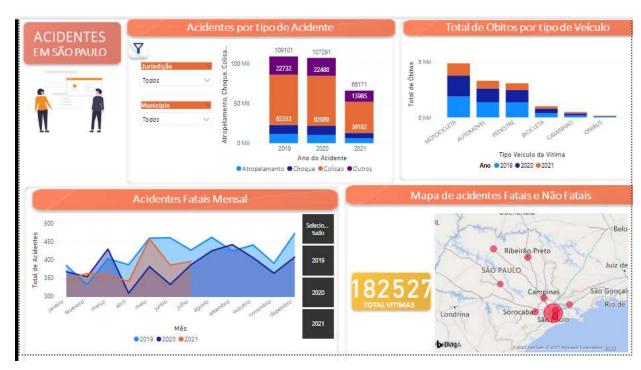


Imagem 17 – Gráfico Acidentes em São Paulo

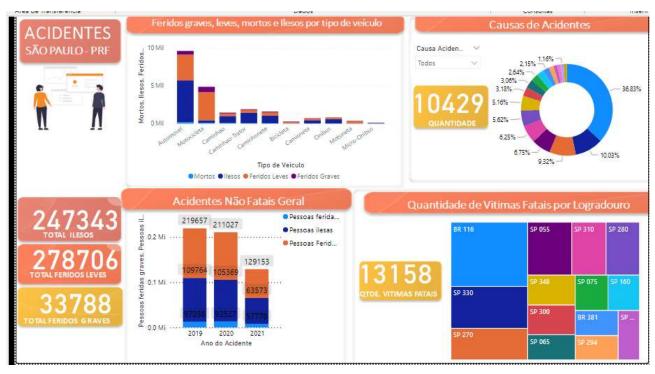


Imagem 18 - Gráfico Acidentes São Paulo - PRF

10 CONCLUSÃO

Após a elaboração dos relatórios, podemos observar que os acidentes não diminuiram como esperado no período da pandemia.

Houve uma alteração no padrão de acidentes e isto se dá por conta da flexibilização as regras de trânsito, diminuição de carros em circulação e aumento nos deliverys com motocicletas. Por consequência os motociclistas foram os mais atingidos, vindo após os pedestres e os passageiros.

Podemos ressaltar também que uma das maiores causas de acidentes é a falta de atenção com 36,86% e que a BR 116 é considerada a rodovia que mais mata no País com 492 vítimas fatais.

O Estado de São Paulo é o que mais possui índices de acidentes de trânsito no Brasil. Segundo a Organização Mundial da Saúde (OMS), o Brasil ocupa a quarta posição no ranking mundial de mortes por acidente de trânsito. De acordo com a Seguradora Líder, no primeiro semestre de 2020, o trânsito brasileiro matou mais do que crimes violentos.

Estudos como estes podem ser a solução e uma das melhores estratégias de decisão e alerta para diminução de acidentes não somente no estado, mas em todo o país, identificando formas de solucionar e identificar pontualmente algumas soluções para circulação de veículos nas principais vias que possuem altos índices

11 REFERÊNCIA – BIBLIOGRAFIA

Documentação Databricks - Disponível em: < https://docs.databricks.com/ >. Acesso em: 24, Agosto de 2021.

Como buscar recursos da Internet usando o pacoteurlib. Disponível em: https://docs.python.org/3howto/urlib2.html>. Acesso em 16, Agosto de 2021.

brazil-highway-traffic-accidents - Disponível em:highway-traffic-accidents. Acesso em: 18, Agosto de 2021.

Relatórios e Base de Dados. Disponível em: http://www.respeitoavida.sp.gov.br/relatorios/>. Acesso em: 16, Agosto de 2021.

Mesmo com pandemia, acidentes de trânsito em SP aumentam 4% em janeiro de 2021 na comparação com ano anterior. Disponível em: https://g1.globo.com/sp/sao-paulo/noticia/2021/02/19/mesmo-com-pandemia-acidentes-de-transito-em-sp-aumentam-4percent-em-janeiro-de-2021-na-comparação-com-ano-anterior.ghtml>. Acesso em: 01, Setembro de 2021.

Estado de são paulo tem queda de 3,8% nas fatalidades de trânsito nos quatro primeiros meses de 2021. Disponível em: http://www.cetsp.com.br/noticias/2021/05/27/cai-numero-total-de-sinistros-de-transito-na-cidade-de-sao-paulo-em-2020.aspx. Acesso em: 06, Setembro de 2021.

Cidade de SP registra queda de 8% das mortes no trânsito. Disponível em: https://noticias.r7.com/sao-paulo/cidade-de-sp-registra-queda-de-8-das-mortes-no-transito-18062021>. Acesso em: 01, Setembro de 2021.

Estado mantém estabilidade nas fatalidades de trânsito nos cinco primeiros meses de 2021. Disponível em :http://www.respeitoavida.sp.gov.br/estado-mantem-estabilidade-nas-fatalidades-de-transito-nos-cinco-primeiros-meses-de-2021/>. Acesso em: 01, Setembro de 2021