**Mapeamento da Criminalidade em São Paulo**

**Projeto da Disciplina de BI**

Prof. Anderson Nascimento

prof.anderson@ica.ele.puc-rio.br

**Componentes do Projeto:**

Alvaro A. S. Junior – aasjr1000@gmail.com

Alexsander Lima – aleglima@gmail.com

Patrick Geller – patrickgeller@gmail.com

|  |
| --- |
| **Histórico de Versões** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Versão** | **Descrição** | **Autor** | **Aprovado por** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Sumário**

[1 Introdução 5](#_heading=h.gjdgxs)

[2 Estudo de Caso 6](#_heading=h.30j0zll)

[2.1 Descrição do Estudo de Caso 6](#_heading=h.1fob9te)

[3 Descrição da Base de Dados 7](#_heading=h.3znysh7)

[4 Proposta de Processo de BI 8](#_heading=h.2et92p0)

[5 Modelo Multidimensional 9](#_heading=h.tyjcwt)

[6 Elaboração do Data Warehouse 11](#_heading=h.1t3h5sf)

[7 Projeto de ETL 12](#_heading=h.2s8eyo1)

[7.1 Descrição do Projeto de ETL 12](#_heading=h.17dp8vu)

[7.2 Extração e Tratamento dos Dados 12](#_heading=h.3rdcrjn)

[7.3 Carregando as Dimensões do DW 14](#_heading=h.26in1rg)

[7.4 Carregando a Tabela Fato do DW 15](#_heading=h.lnxbz9)

[8 Dashboard 16](#_heading=h.35nkun2)

[8.1 Descrição da Elaboração 16](#_heading=h.1ksv4uv)

[8.2 Telas do Dashboard 16](#_heading=h.44sinio)

[9 Conclusão 18](#_heading=h.2jxsxqh)

# Introdução

Este documento tem por finalidade de mostrar como foi realizado o processo de B.I. para o estudo de caso “**Mapeamento da Criminalidade em São Paulo”**, desde a seleção e análise da base de dados, a modelagem do Data Warehouse, o tratamento e carga dos dados no processo de ETL, e por fim a criação dos *Dashboards* para visualização das informações extraídas.

# Estudo de Caso

## Descrição do Estudo de Caso

Um problema grave que aflige o brasileiro é a violência urbana, um problema que ataca a todos, principalmente os moradores das chamadas áreas de risco, onde a frequência de crimes é maior.

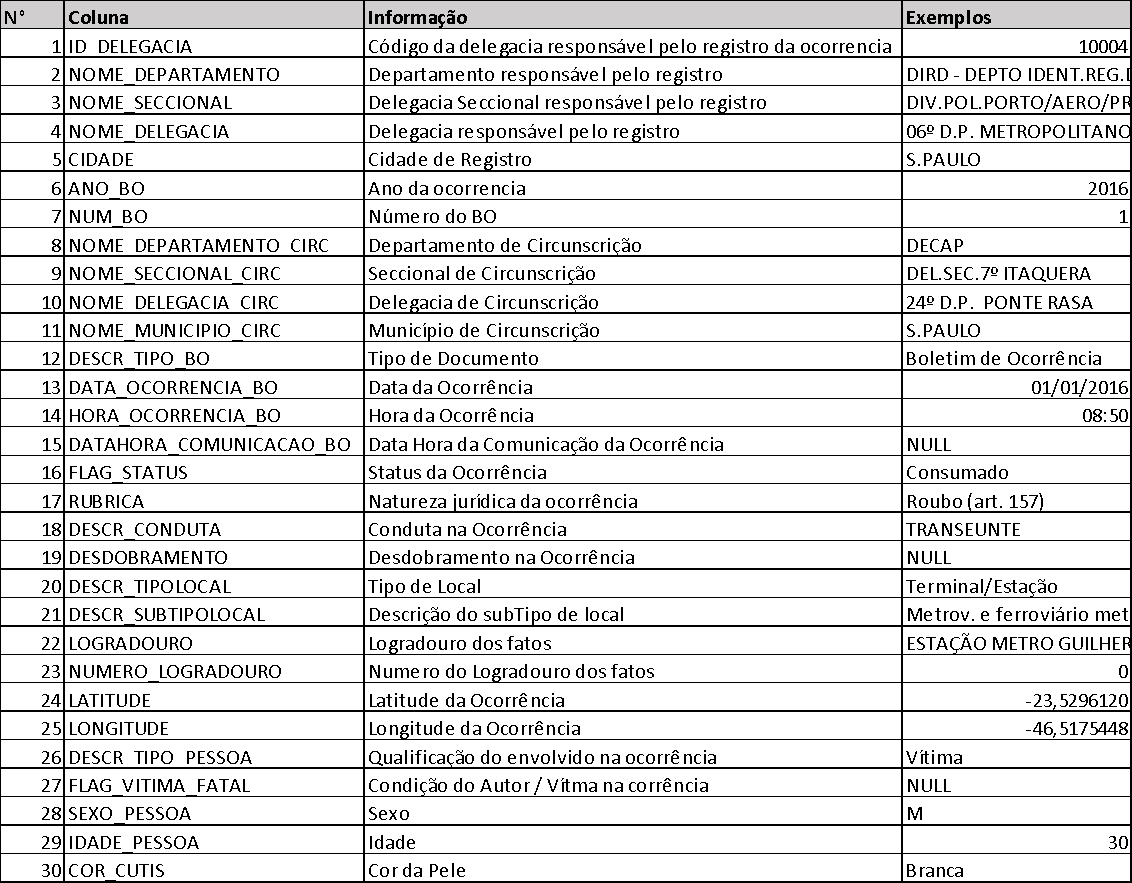
O objetivo do projeto é apresentar mapas de calor que indiquem a frequência da ocorrência de determinados tipos de crimes em uma determinada região, subsidiando as autoridades de informações para a melhor alocação de recursos referentes à segurança pública.

Nesse estudo de caso serão estudados os dados da cidade de São Paulo, mas a estrutura do Data Warehouse e do processo de B.I. é flexível o bastante para analisar dados de outras regiões, bastando adicionar a nova base de dados e ajustar o ETL.

# Descrição da Base de Dados

A base de dados utilizada nesse estudo de caso se chama “Crime Data in Brazil” foi obtida a partir do site Kaggle [1], ela é formada por arquivos no formato “.csv” que contém dados resumidos das ocorrências criminais registradas no estado de São Paulo entre os anos de 2007 e 2016.

No estudo será utilizada a parte mais recente da base de dados que contém os dados extraídos do sistema de Registro Digital de Ocorrências (R.D.O.), com os dados do ano de 2016 e apresenta as informações contidas na Tabela 1.



*Tabela 1 – Descrição das colunas dos arquivos RDO\_3.csv*

# Proposta de Processo de BI

O processo de BI proposto neste estudo começa pela seleção das informações relevantes da base de dados. A seleção dessas informações será explicada na próxima seção pela lógica dos pontos cardeais.

Em seguida no processo de ETL, as colunas referentes a cada informação relevante para o estudo têm seus dados extraídos da base de dados, tratados e por fim carregados e armazenados no Data Warehouse (DW).

Na última etapa os dados contidos no DW são capturados pelo player de BI onde os *dashboards* com gráficos de tendências e mapas são apresentados.

Para o processo de ETL será utilizada a ferramenta “Pentaho Data Integration (PDI)”, para o DW será utilizado o sistema de gerenciamento de banco de dados “PostgreSQL”, e o player de BI utilizado será o “Power BI”.

Na imagem 1 está exposto o passo-a-passo do processo de BI.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

*Figura 1 – Processo de BI*

# Modelo Multidimensional

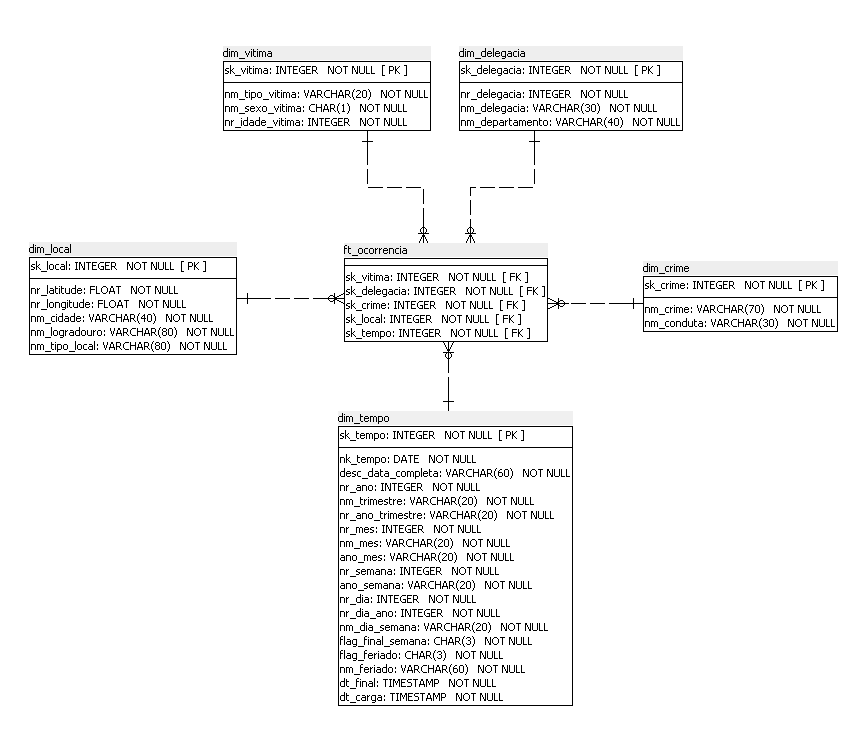
Para definição das dimensões do modelo estrela (star model) e as informações que serão contidas nelas, é utilizada a lógica dos pontos cardeais, onde se busca responder certas perguntas básicas com relação ao fato estudado, que nesse caso é uma ocorrência criminal.

As perguntas, relacionadas ao fato, que devem ser respondidas são: “O quê?", “Quem?”, “Quando?” e “Onde?”.

A dimensão crime (dim\_crime) responde à pergunta “O que?”, o que aconteceu, qual foi o crime cometido. As dimensões vítima (dim\_vitima) e delegacia (dim\_delegacia) respondem à pergunta “Quem?”, quem foi a vítima e que delegacia apurou a ocorrência. A dimensão local (dim\_local) responde à pergunta “Onde?”, onde ocorreu o crime. E a dimensão tempo (dim\_tempo) responde à pergunta “Quando?”, quando ocorreu o crime.

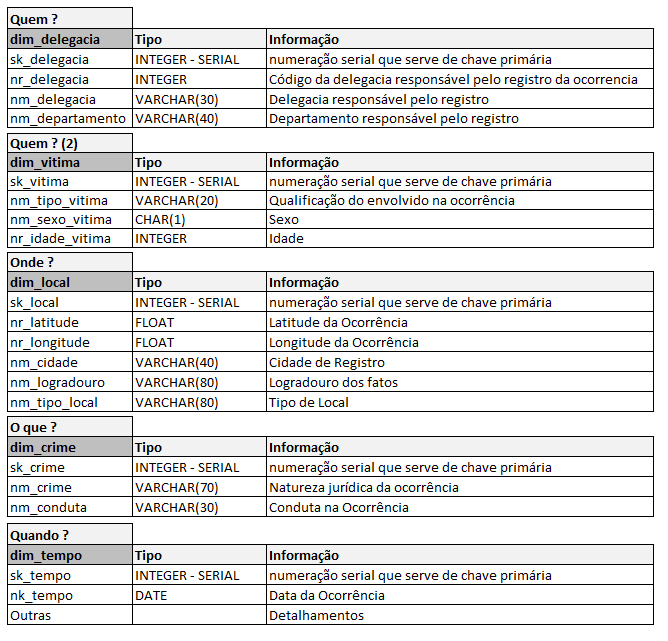
Conectando essas dimensões ao fato, o modelo multidimensional fica como é exibido na Figura 2.

A ferramenta utilizada para a montagem do modelo foi o “SQL Power Architect”, que gerou o arquivo com o modelo multidimensional “modelo\_dw\_crime\_sp” [2].



*Figura 2 – Modelo Estrela do DW*

Abaixo na Tabela 2 são descritos os componentes de cada dimensão, o seu tipo de dado em SQL e a informação contida neles. A descrição consolidada das dimensões está contida no arquivo “Descrição das dimensões.xlsx” [3].



*Tabela 2 – Descrição dos componentes das Dimensões.*

# Elaboração do Data Warehouse

Nessa seção será apresentada a estrutura do *Data Warehouse* (DW), no que se refere a arquitetura, tipo de abordagem, frequência de carga e estrutura física.

A construção lógica do DW está descrita no arquivo “sql\_contrucao\_dw.sql” [4], as dimensões foram construídas conforme o modelo multidimensional apresentado na seção anterior, na dimensão tempo foi utilizado o script cedido pelo Prof. Anderson, e para facilitar o processo de ETL foi criada uma tabela stage para recepção dos dados que vêm diretamente do CSV.

A estrutura do DW tem arquitetura independente, desenvolvido apenas para este estudo, é flexível para alteração da base de dados se o processo de ETL for ajustado, mas não é prevista outras conexões externas.

Quanto ao tipo de abordagem se trata de uma abordagem “Bottom-Up”, as bases de dados são carregadas no DW onde são armazenados de forma estruturada.

É prevista para este estudo uma carga incremental anual, e devido a isso uma arquitetura física local (*On-premises*) de baixo custo seria o suficiente.

# Projeto de ETL

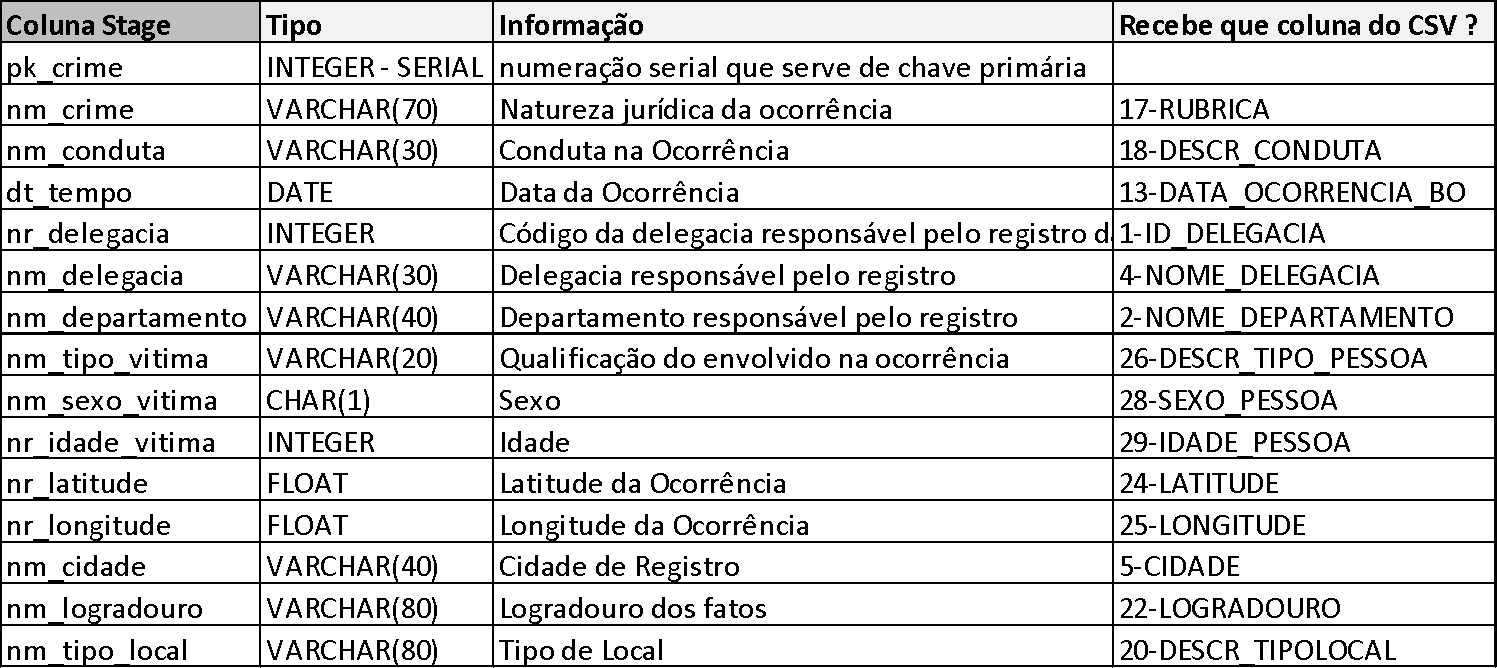
## Descrição do Projeto de ETL

O processo de ETL para esse estudo segue os seguintes passos: primeiro os dados contidos no arquivo “RDO3.csv” são tratados e carregados na tabela stage (st.st\_crime\_sp), o segundo passo é popular as dimensões do DW e o terceiro e último passo é popular a tabela fato do DW.

A script para construção da tabela stage e do DW está no arquivo “sql\_contrucao\_dw.sql” [4].

## Extração e Tratamento dos Dados

Os dados do arquivo “RDO3.csv” serão tratados e carregados na tabela *stage* “st.st\_crime\_sp”, abaixo na Tabela 3 estão descritas as informações contidas na tabela e que origem tem.



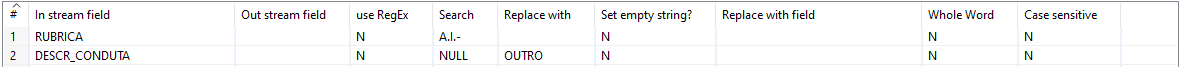
*Tabela 3 – Descrição das colunas da tabela Stage.*

No PDI serão tratados os dados a serem carregados, na Figura 3 está exibido o processo de tratamento, e no arquivo “Tratamento CSV.xlsx”[5] estão descritos os tratamentos realizados, e o arquivo das transformação é o “etl\_st\_crime\_sp.ktr”[6]:



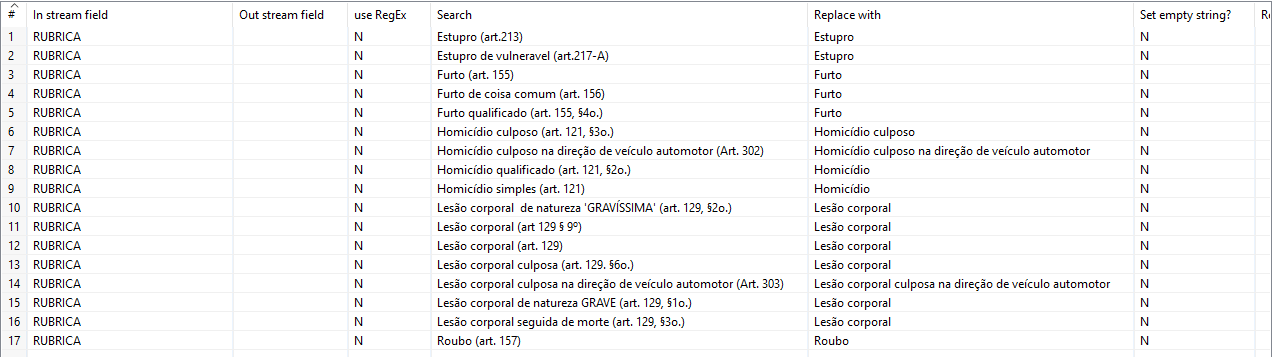
*Figura 3 – Tratamento dos dados do CSV e carga na tabela Stage*

O primeiro tratamento “Tratamento\_CRIME\_1”, consiste em utilizar a transformação Replace in string para substituir a string “NULL” da coluna DESCR\_CONDUTA pela string “OUTRO” e eliminar da coluna RUBRICA os prefixos “A.I.-“, conforme a Figura 4 :



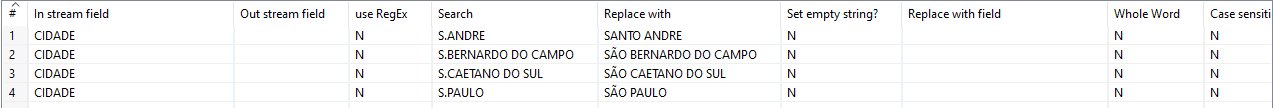
*Figura 4 – Tratamento\_CRIME\_1*

O segundo tratamento “Tratamento\_CRIME\_2”, também utiliza a transformação Replace in string, mas agora com o objetivo de simplificar a descrição dos crimes contidos na coluna RUBRICA do arquivo original da base de dados, conforme a Figura 5:



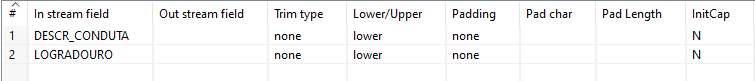
*Figura 5 – Tratamento\_CRIME\_2*

No terceiro tratamento “Tratamento\_CIDADE”, a transformação Replace in string é utilizada para retirar a abreviação do nome de algumas cidades para que player de BI consiga identificar a localização dos crimes adequadamente, na Figura 6 o tratamento é explicitado:



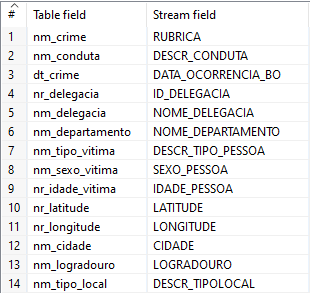
*Figura 6 – Tratamento\_CIDADE*

O quarto tratamento “Tratamento\_NOMES”, utiliza da transformação “String Operations”, e basicamente, passa os nomes das colunas “DESCR\_CONDUTA” e “LOGRADOURO” para letras minúsculas, o objetivo seria facilitar a concatenação com os nomes dos crimes localizados na coluna “RUBRICA”, caso seja necessário. Na Figura 7 a transformação é apresentada:



*Figura 7 – Tratamento\_Nomes*

Por último é feita a seleção das colunas e carregamento na tabela “st.st\_crime\_sp” através das transformações “Select Values” e “Table output”, conforme a Figura 8:



*Figura 8 – Seleção e carregamento das colunas*

## Carregando as Dimensões do DW

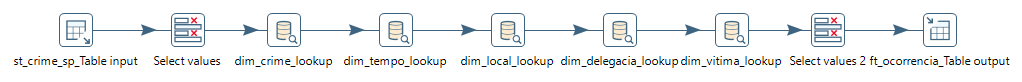
Após o carregamento dos dados na tabela “st.st\_crime\_sp” as dimensões serão populadas com a ajuda do comando *DISTINCT*, na Figura 9 é exibido um exemplo, o carregamento da dim\_crime, essa e as outras operações que estão contidas no arquivo “sql\_populando\_dim\_dw.sql”[7].



*Figura 9 – Carregamento da dim\_crime*

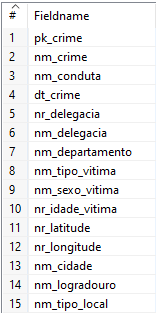
## Carregando a Tabela Fato do DW

Para o carregamento da tabela fato do DW a “dw.ft\_ocorrencia”, primeiro os dados da tabela stage “st.st\_crime\_sp” foram selecionados , após isso cada linha da tabela contendo os dados das ocorrências foram comparadas com as suas respectivas categorias localizadas nas dimensões, para a definição de quais surrogate keys seriam carregadas na tabela fato. A processo está contido no arquivo “dw\_carga\_fato.ktr” [8], e na Figura 10 está explicitado:



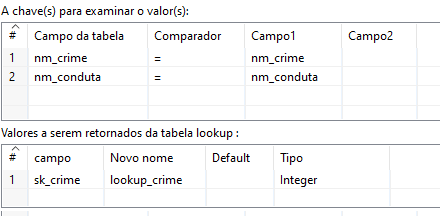
*Figura 10 – Carregamento da Tabela Fato do DW*

Primeiro ocorre a seleção dos valores da tabela st.st\_crime\_sp, basicamente todas as colunas são selecionadas, conforme está na Figura 10.



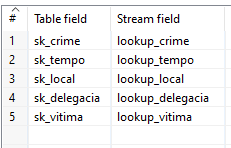
*Figura 11 – Seleção das colunas da tabela stage*

No segundo passo as colunas nm\_crime e nm\_conduta da tabela stage são comparadas com as suas respectivas colunas na dim\_crime, através da operação “Database lookup” e assim é selecionada a chave sk\_crime correspondente, conforme a Figura 12.



*Figura 12 – Definição da sk\_crime correspondente*

Esse processo é repetido para todas as outras dimensões (tempo, local, delegacia e vítima), e após a definição das suas respectivas surrogate keys, essas chaves são selecionadas e carregadas na tabela fato “dw.ft\_ocorrencia”, conforme a Figura 13.



*Figura 13 – Carregamento das SKs na Tabela Fato*

Após o carregamento da Tabela Fato o processo de ETL está concluído.

# Dashboard

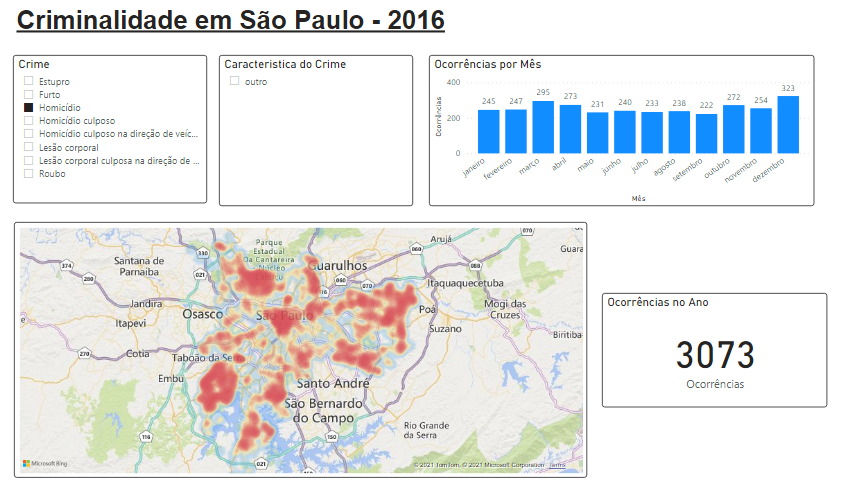
## Descrição da Elaboração

Para criação do Dashboard, dados do DW foram importados para o Power BI Desktop versão 2.98.1025.0 64-bit (October 2021), e estão no arquivo “BI\_Mapeamento da Criminalidade em São Paulo.pibx”[9]

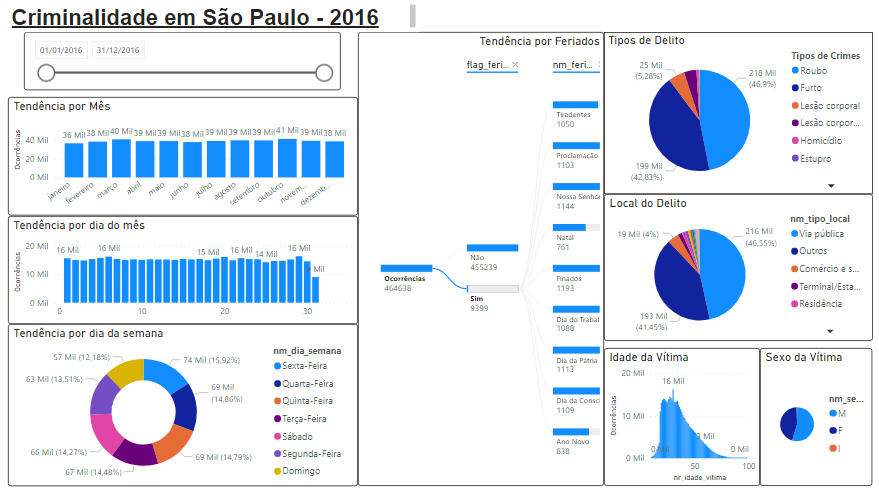
## Telas do Dashboard

Abaixo, reproduzimos as telas criadas no Dashboard para mostrar as estatísticas e tendências de criminalidade em São Paulo. Inserimos filtros para os estudos desejados, além de um *heatmap* de crimes plotados no mapa de São Paulo usando a latitude/longitude da ocorrência, presente na base de dados. Analisamos também as tendências por mês do ano, por dia do mês, por dia da semana e feriados, observando um comportamento bastante uniforme ao longo do tempo. Apresentamos também dados para caracterizar os principais tipos de delitos, incluindo tipo de crime, tipo de local e características das vítimas.

Obviamente a ferramenta é flexível o suficiente para criação de novas visualizações, com os dados existentes ou mesmo carregamento de novos dados para mais análises.



*Figura 14 – Capa e Mapa de Calor – Exemplo selecionado “Homicídio”*

**

*Figura 15 – Análise descritiva dos dados e tendências*

# Conclusão

Este trabalho permitiu praticar todos os conceitos aprendidos na disciplina de *Business Intelligence,* passando pela seleção e análise da base de dados, a modelagem do *Data Warehouse*, o tratamento e carga dos dados no processo de ETL, e por fim a criação dos *Dashboards* para visualização das informações extraídas.

Abordamos o tema da segurança pública, caracterizando o problema na cidade de São Paulo e usando análise de dados para analisar comportamento e tendências da violência urbana, com base em *Database* existente.

Além da análise específica criada, o modelo é flexível o bastante para extensão a novas análises e ampliação da base de dados, sendo bastante útil para verificação de tendências na violência urbana.

1. **Anexos**
2. **Arquivos**

[1] Base de Dados: <https://www.kaggle.com/inquisitivecrow/crime-data-in-brazil>

[2] Arquivo com o modelo multidimensional: modelo\_dw\_crime\_sp

[3] Arquivo com a descrição das dimensões do DW: Descrição das Dimensões.xlsx

[4] Arquivo para a construção do DW: sql\_contrucao\_dw.sql

[5] Arquivo com a descrição do tratamento do CSV: Tratamento CSV.xlsx

[6] Arquivo PDI de tratamento do CSV e carga na table Stage: etl\_st\_crime\_sp.ktr

[7] Arquivo de carregamento as dimensões:sql\_populando\_dim.sql

[8] Arquivo PDI de carregamento da Tabela Fato: etl\_carga\_fato.ktr

[9] Arquivo com os Dashbords: BI\_Mapeamento da Criminalidade em São Paulo.pibx