

# MVP DATA ENGINEERING/PIPELINE - 3rd Sprint of the Data Science and Analytics Post-Graduation Program of PUC-Rio

- Professors: Tatiana Escovedo, Fernanda Baião, Marcos Villas, Anthony Seabra, Silvio Alonso, and Victor Almeida

Student: Dr. Vagner Zeizer Carvalho Paes

## MVP's Title: Crimes's Investigations in Parana over 2018-2023 years

In this work, it is going to be created a **Data Pipeline** in DataBricks using the free service named **DataBricks Community Edition**. Here, a plain Table will be created and the correspondent *ETL* (Extract, Transform, Load) pipeline is going to be detailed and documented, as shown below. The pipeline was written in PySpark and the *Transformation* part was fully documented in order to get governance over the entire dataset.

## Definition of the Problem

The State of Paraná, in Brazil, has experienced significant variations in *crime rates* over the past years. While some municipalities have seen a decline in crime, others have seen an increase. This raises questions about the factors that contribute to crime rates in different areas of the state. By understanding these relationships, we aim to identify patterns, trends, and potential causal factors that could help in developing targeted crime prevention strategies and socioeconomic policies.

## Basic Research Questions:

- To what extent do socioeconomic factors, such as the *Municipal Human Development Index (MHDI)* and *Gross Domestic Product (GDP)* per capita, influence crime rates in the State of Paraná, Brazil?
- What are the *most dangerous cities* according to a specific kind of Crime in the State of Paraná?
- *Impact of Drug-Related Crimes*: How do drug-related crimes (trafficking or use/consumption) relate to other types of crimes? Is there a significant overlap or correlation?

## MVP's Goal

- From the collected data, the following questions about Crimes in the State of Paraná along the 2018-2023 years will be addressed:

1. What is the correlation between Crimes, such as Rapes and Thefts, and GDP or MHDÍ?
2. What are the Top 10 cities in State of Paraná with the highest number of Rape Crimes?
3. What are the Top 10 cities in State of Paraná with the highest number of Robbery Crimes?
4. What are the Top 10 cities in State of Paraná with the highest number of Drug Trafficking Occurrences?
5. What are the Top 10 cities in State of Paraná with the highest number of Vehicle Thefts?
6. What are the Top 10 cities in State of Paraná with the highest Disturbing of Piece/Tranquility?
7. What is the mean Drug Trafficking and Drug use Occurrences in Curitiba?
8. Concerning the 7-th question, what about specifically in year 2020?

## 1 - Data Search

The dataset collected to investigate in this work was obtained from this [website](#) from government of the State of Paraná, as it will be detailed in the next sections.

### 1.1 Data License

The license for this public, that data was made available by the Brazilian government, is the **Creative Commons Attribution 4.0 International (CC BY 4.0) license**. This license allows anyone to reuse, distribute, and modify the data for any purpose, including commercial purposes, as long as they give appropriate credit to the original source.

### 1.2 - Data Dictionary

Information about the dataset features:

- 1 - 'Ano': years investigated throughout the study;
- 2 - 'Localidade': Location where the Crimes happened;
- 3 - 'Índice de Desenvolvimento Humano Municipal (IDHM) ': Municipal Human Development Index (MHDÍ);
- 4 - 'Produto Interno Bruto (PIB) per Capita (R\$ 1,00)': Gross Domestic Product (GDP);
- 5 - 'Crimes de Ameaça ': Threat Crimes;
- 6 - 'Crimes de Estelionato ': Fraud crimes;
- 7 - 'Crimes de Estupro ': Rape Crimes;
- 8 - 'Crimes de Furto ': Theft Crimes;
- 9 - 'Crimes de Lesão Corporal ': Bodily Injury;
- 10 - 'Crimes de Roubo ': Robbery Crimes;
- 11 - 'Furtos de Veículos ': Vehicle Thefts;
- 12 - 'Ocorrências Envolvendo Tráfico de Drogas ': Drug

Trafficking Occurrences;  
 13 - 'Ocorrências Envolvendo Uso/Consumo de Drogas ': Drug Use Occurrences;  
 14 - 'Perturbação do Sossego/Tranquilidade ': Disturbing the Peace/Tranquility;  
 15 - 'Roubos de Veículos ': Vehicle Robberies.

## 2. Data Gathering

The dataset was collected from this [web page](#), and by clicking on *Base de Dados do Estado (BDEweb)*, being redirected to this [page](#). Now, it was selected the variables ("*Seleção de Variáveis*") present in the **Crimes** section and it was chosen the aforementioned features in the previous section (such as 'Ano', rape and theft crimes, and so on). In order to choose the Locations/Municipalities, it was clicked on *Seleção de Localidade*, and selected "State of Paraná" (*Estado do Paraná*) together with all the cities (*Todos os Municípios*) from State of Parana. This dataset comprehends all the cities within the State of Parana, along with the overall quantities for the whole State of Parana, over the 2018-2023 years.

## 3. Data Modeling

Once the data was collected, as shown in the previous Section, it was organized as a single table, and therefore the **flat table Model** was utilized. Despite Models like *star* or *snowflake* could be more suitable for querying and analyze structured relational databases, the dataset here is simple, lightweight and flat, showing a flexible structure. The dataset will be imported from the local computer, loaded in *PySpark* and transformed into a cleaned format, and saved as a Table in order to perform queries, answering the eight business questions listed in the beginning of this document.

### 3.1 Data Lineage/Data Catalog

The file *./DataCatalog\_MVPIII.xlsx* contains a description of the features, their domains, containing expected values for numerical data, and possible categories for categorical data. A printscreen of this Data Catalog is provided below.

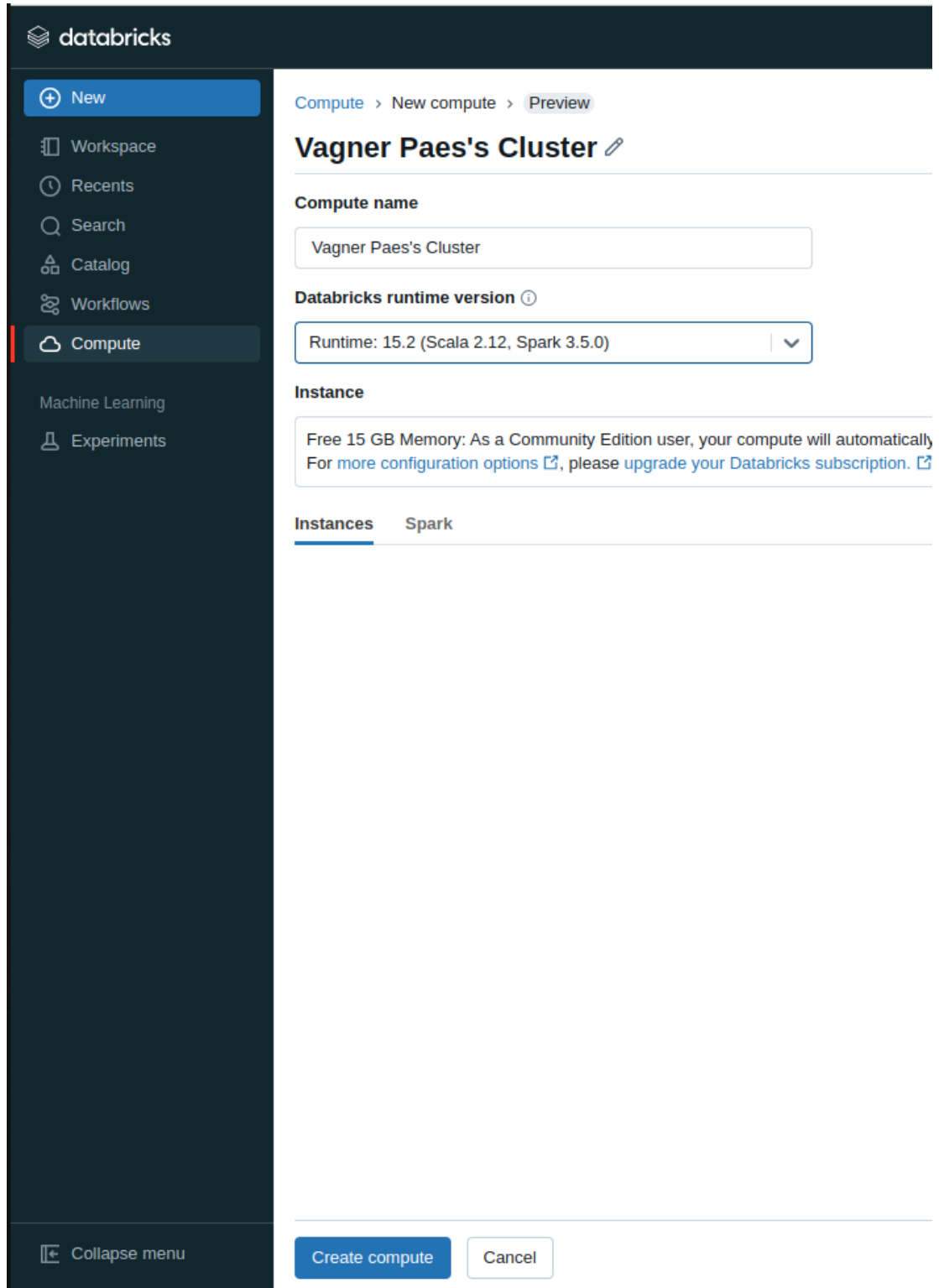
Column Name	Data Type	Description	Accepted Values (**)
Ano	Integer	years investigated throughout the study	[2018,2019,2020,2021,2022,2023]
Localidade	String	Location where the Crimes happened	Any City within State of Paraná
Índice de Desenvolvimento Humano Municipal (IDHM)	Float	Municipal Human Development Index (MHDI)	Any Positive Value between 0 and 1
Produto Interno Bruto (PIB) per Capita (R\$ 1,00)	Float	Gross Domestic Product (GDP)	Any Positive Value
Crimes de Ameaça	Integer	Threat Crimes	Any Positive Integer value
Crimes de Estelionato	Integer	Fraud Crimes	Any Positive Integer value
Crimes de Estupro	Integer	Rape Crimes	Any Positive Integer value
Crimes de Furto	Integer	Theft Crimes	Any Positive Integer value
Crimes de Lesão Corporal	Integer	Bodily Injury	Any Positive Integer value
Crimes de Roubo	Integer	Robbery Crimes	Any Positive Integer value
Furtos de Veículos	Integer	Vehicle Thefts	Any Positive Integer value
Ocorrências Envolvendo Tráfico de Drogas	Integer	Drug Trafficking Occurrences	Any Positive Integer value
Ocorrências Envolvendo Uso/Consumo de Drogas	Integer	Drug Use Occurrences	Any Positive Integer value
Perturbação do Sossego/Tranquilidade	Integer	Disturbing the Peace/Tranquility	Any Positive Integer value
Roubos de Veículos	Integer	Vehicle Robberies	Any Positive Integer value

## 4. Load

In this section, **Screenshots** will be pasted in order to prove that work was really done according to the MVP's criteria.

### 4.1 Cluster Creation

The figure below shows the DataBricks's cluster being created.



The screenshot displays the Databricks web interface for creating a new compute cluster. The left sidebar contains navigation links: New, Workspace, Recents, Search, Catalog, Workflows, Compute (highlighted), Machine Learning, and Experiments. The main content area shows the 'Compute > New compute > Preview' breadcrumb. The cluster name is 'Vagner Paes's Cluster'. The Databricks runtime version is set to 'Runtime: 15.2 (Scala 2.12, Spark 3.5.0)'. The Instance section notes 'Free 15 GB Memory: As a Community Edition user, your compute will automatically' and provides links for 'more configuration options' and 'upgrade your Databricks subscription'. At the bottom, there are 'Instances' and 'Spark' tabs, and 'Create compute' and 'Cancel' buttons.

databricks

+ New

Workspace

Recents

Search

Catalog

Workflows

Compute

Machine Learning

Experiments


Collapse menu


Compute > New compute > Preview

### Vagner Paes's Cluster

Compute name

Vagner Paes's Cluster

Databricks runtime version 

Runtime: 15.2 (Scala 2.12, Spark 3.5.0) 

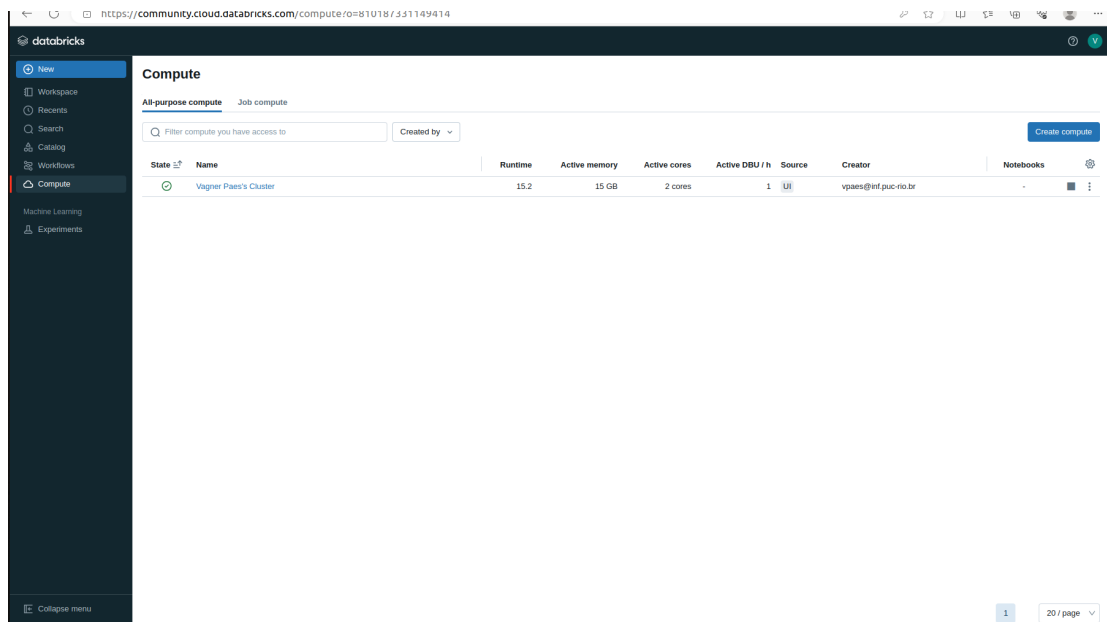
Instance

Free 15 GB Memory: As a Community Edition user, your compute will automatically  
For [more configuration options](#), please [upgrade your Databricks subscription](#).

Instances Spark

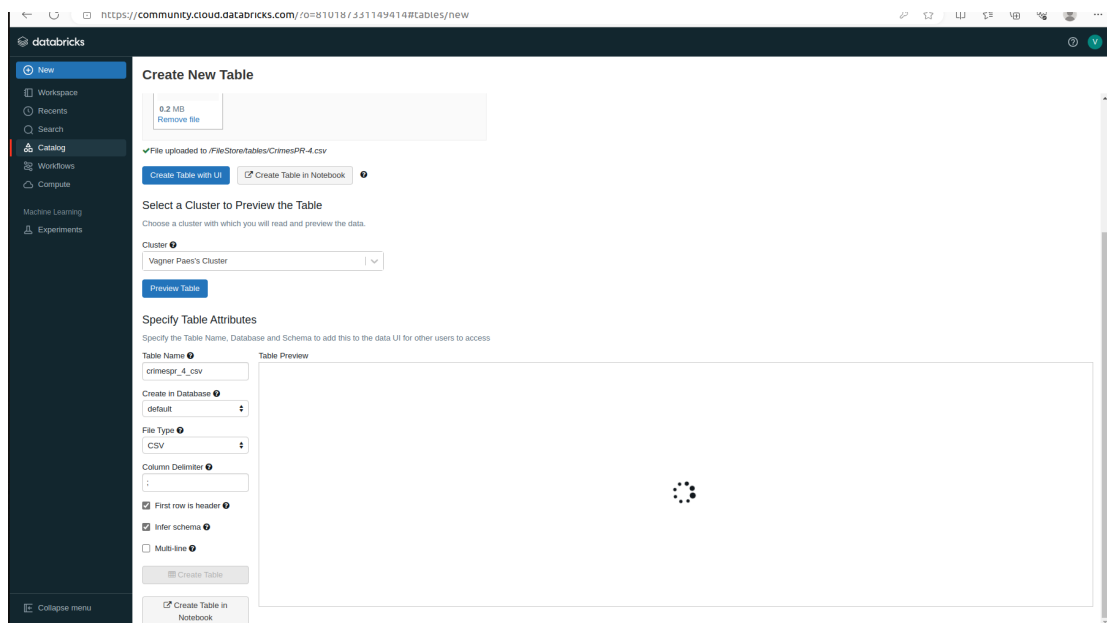
Create compute Cancel

The figure below shows that the DataBricks's cluster was successfully created.



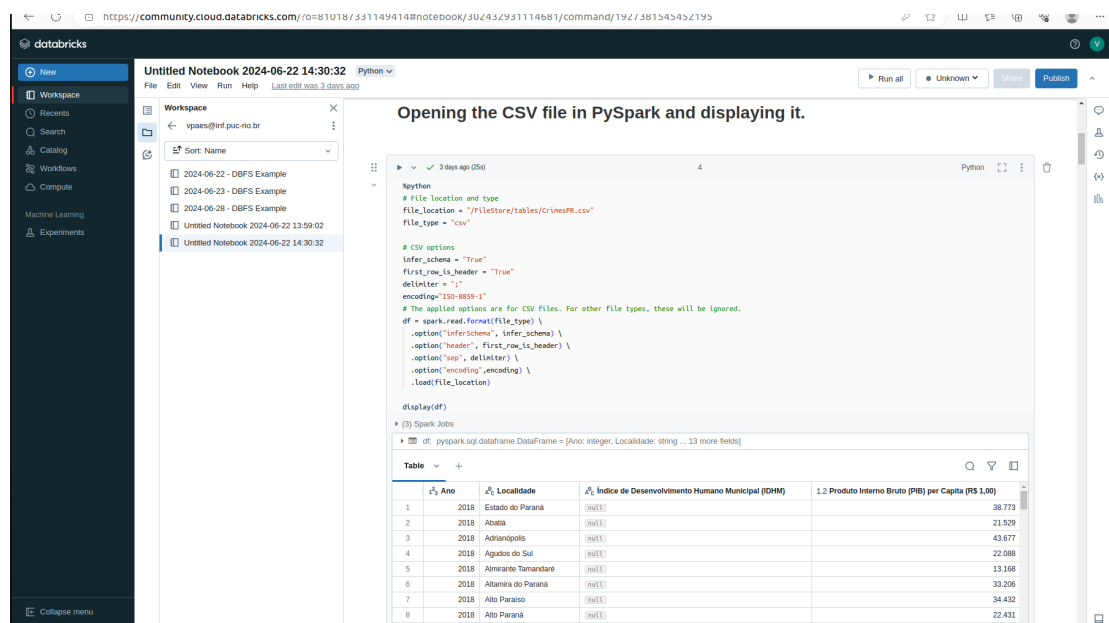
## 4.2 Table creation

The figure below shows the creation of a Table from the *Crimes Dataset*. The layer containing the dataset at this stage is going to be named as **bronze**, as usual.



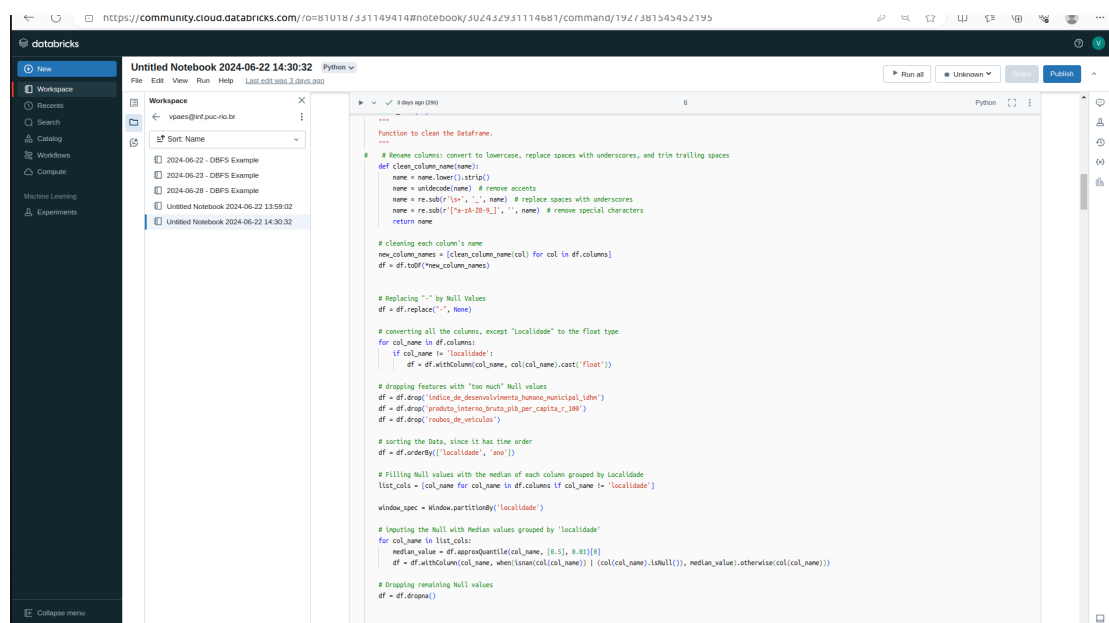
## 4.3 Loading the Dataset

The figure below shows the load of the *Crimes Dataset* in DataBricks by using PySpark.



## 4.4 Transforming the Dataset

The figure below shows the **Transformation** step of the *Crimes Dataset* in DataBricks by using PySpark programming Language.

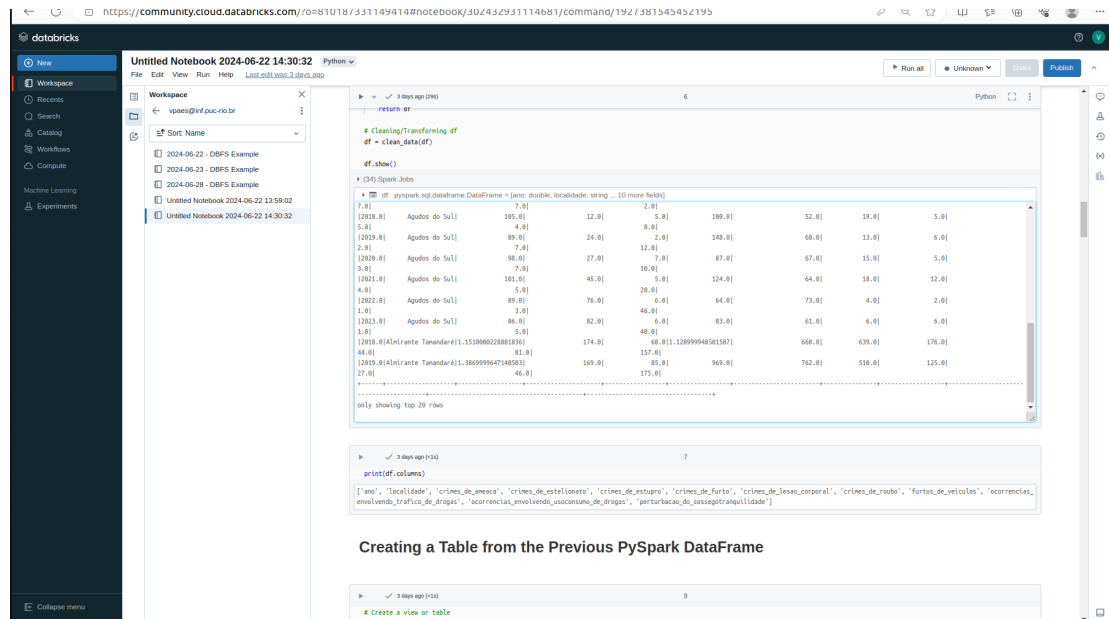


The following procedures have been performed in order to perform the data's transformation:

1. Column's names were standardized by putting them in lower case, stripping whitespaces, and removing hyphens, as well as special characters;
2. Next, all the columns were converted to *float* type, except 'localidade', which is a *string*;
3. Columns related to '*IDHM*', '*GDP*' and '*roubo de veiculos*' were dropped, since they have a very large number of null values;
4. The DataFrame/RDD was order by 'localidade', and 'ano', since this dataset has a *temporal order*;

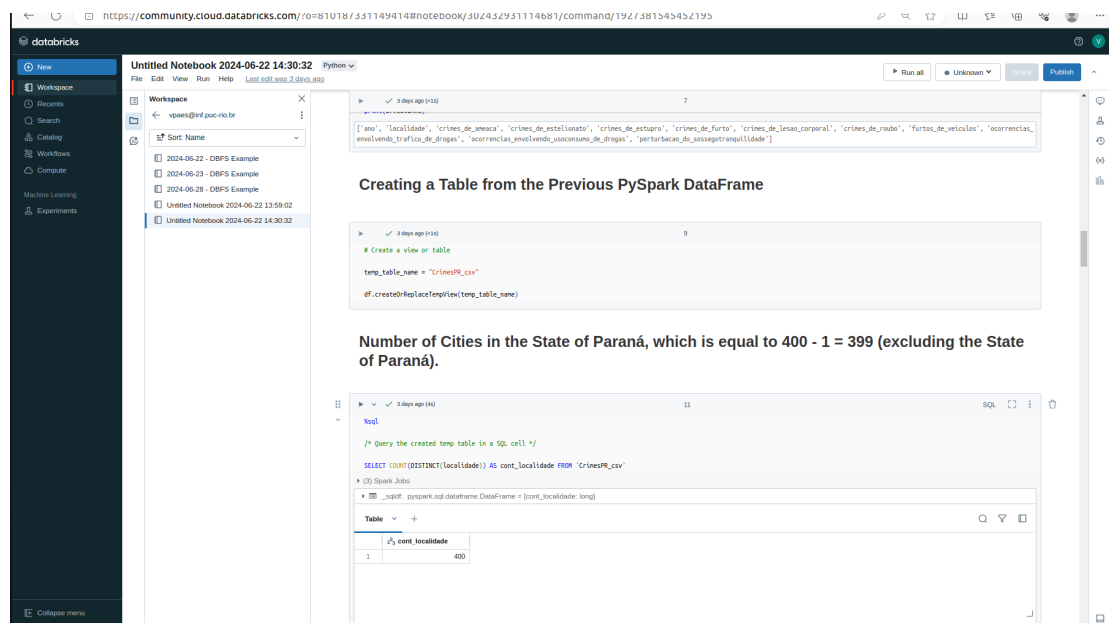
- Next, null values were *imputed by their median values* according to a given municipality ('localidade');
- Rows with null values were *dropped from the dataset*, since that if they were not imputed in the last step, they lacked all measurements over the 5 years (from 2018 up to 2023) and must be dropped.

The figure below shows a screenshot, evidencing that the cleaning procedure was successfully applied in this dataset. Now, there is a cleaned dataset ready-to-be-used in further data analysis (see next section). The layer containing this dataset is going to be named, as usually, **silver**.



## 4.5 Creating a Table (Temporary View)

The figure below shows a screenshot evidencing that the creation of a Table, or rather a *Temporary View*, was successfully performed in Databricks. In the next section, it is going to be performed data analysis on this transformed dataset by using both *SQL* and *Python*.



## 5. Analysis

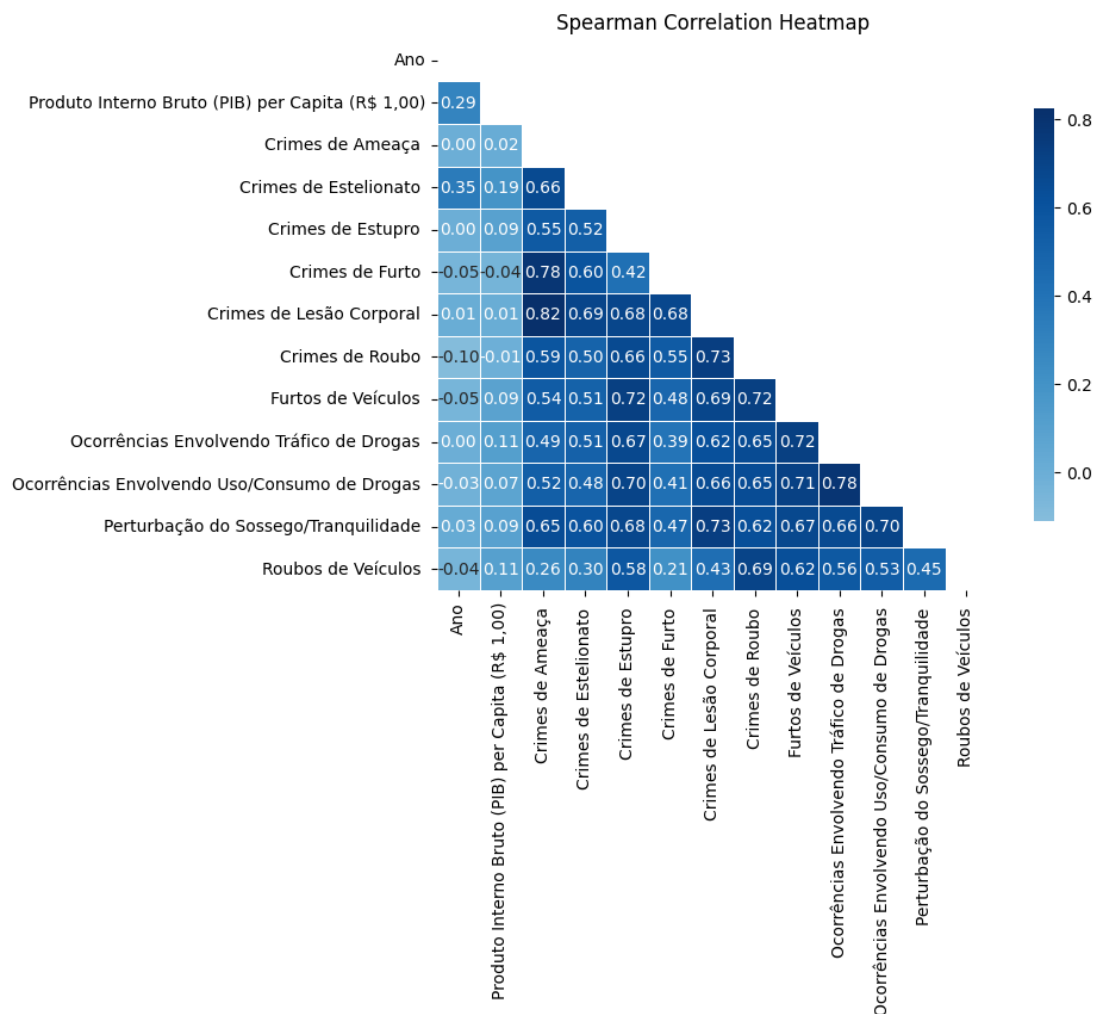
### 5.1 Data Analysis

- Below, it is going to be addressed the questions listed in the beginning of the MVP. Screenshots of the SQL queries will be displayed, evidencing the work has been done.

Q1. What is the correlation between Crimes, such as Rapes and Thefts, and GDP or MHDI?

- Correlation Analysis

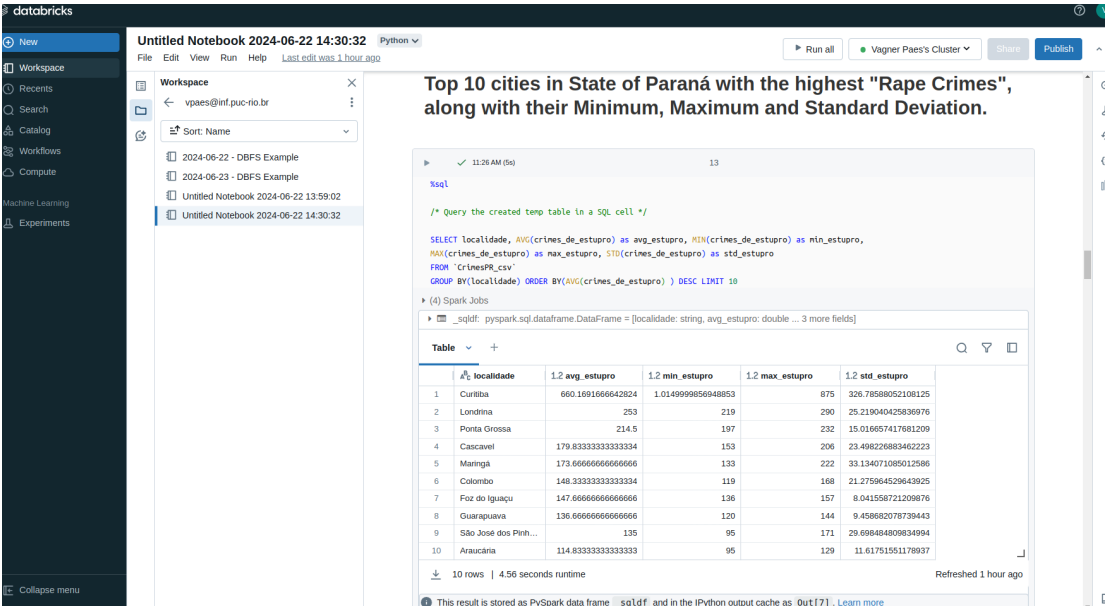
The figure below shows that, surprisingly, that crimes, in general, have a low correlation with the Cities's GDP, and it was not possible to answer this question concerning *MHDI* feature, because this specific feature was completely absent of data. Additionally, **Crimes** seem to have a medium or high-correlation between each other. It is noteworthy that the *Spearman's scorumlation method* was used here to create the correlation's heatmap plot (see the file `./notebooks/EDA_CrimesPR.ipynb`), since this data is not normally distributed.



Q2. What are the Top 10 cities in State of Paraná with the highest number of Rape Crimes?

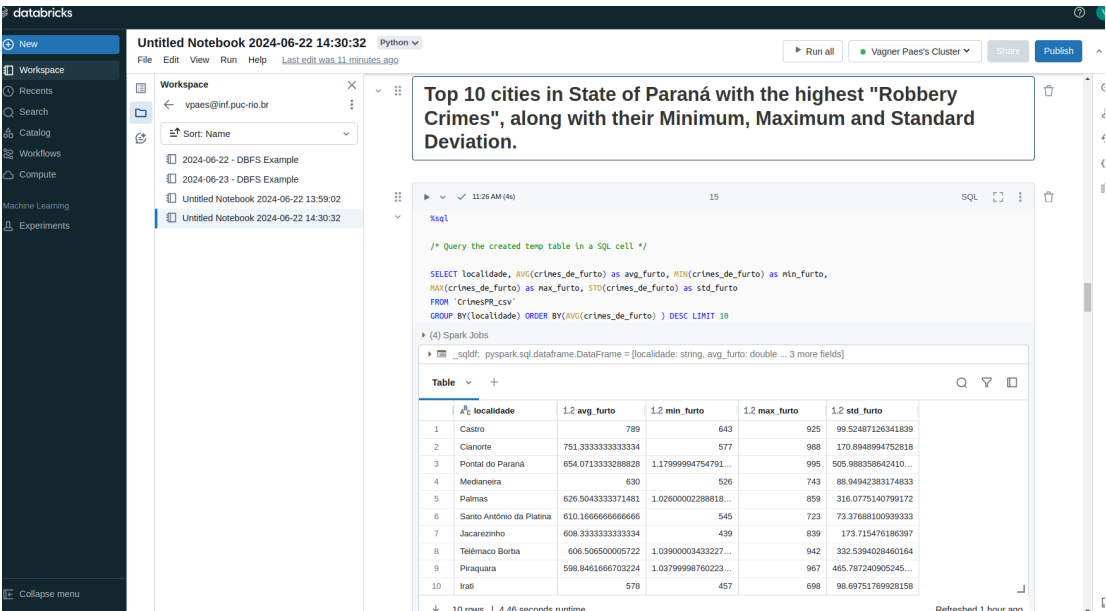


The top 10 cities are shown in the Figure below. The Top 3 are: *Curitiba, Londrina and Ponta Grossa.*



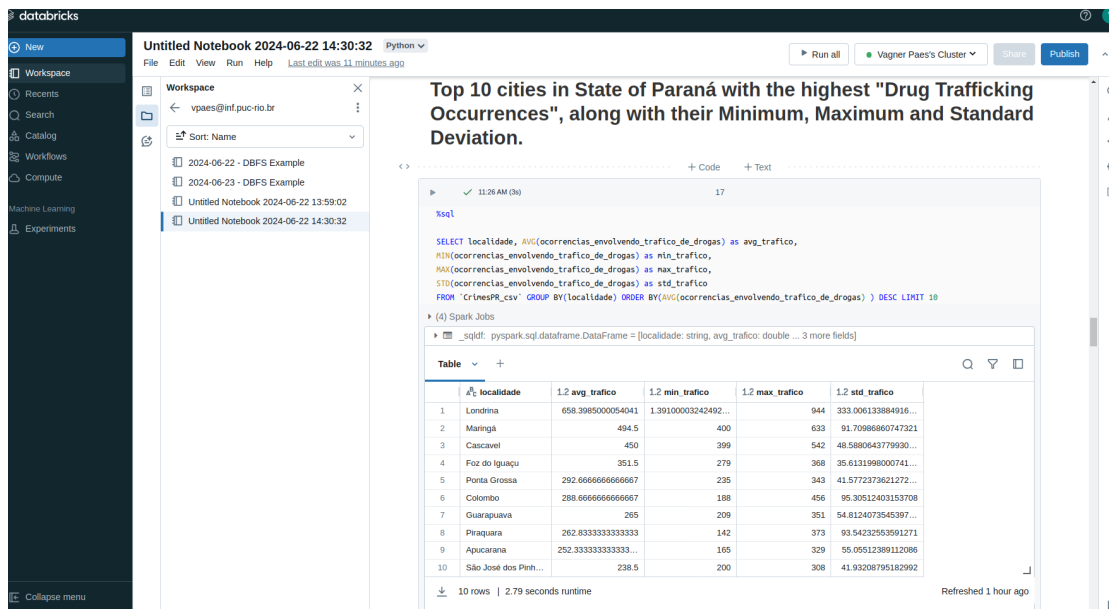
Q3. What are the Top 10 cities in State of Paraná with the highest number of Robbery Crimes?

The top 10 cities are shown in the Figure below. The Top 3 are: *Castro, Cianorte and Pontal do Paraná.*



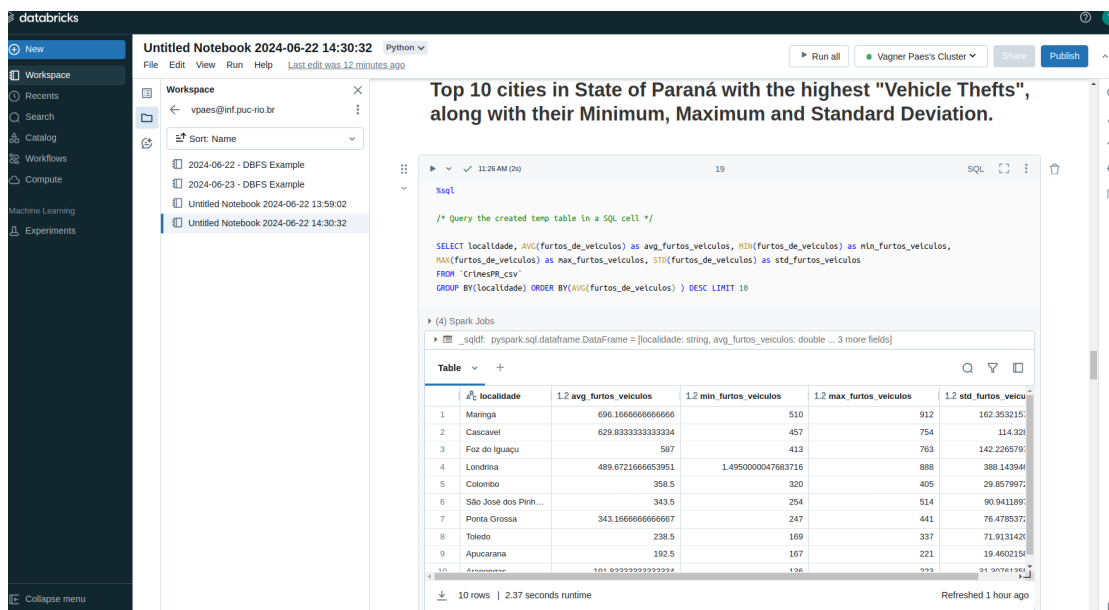
Q4. What are the Top 10 cities in State of Paraná with the highest number of Drug Trafficking Occurrences?

The top 10 cities are shown in the Figure below. The Top 3 are: *Londrina, Maringá and Cascavel.*



Q5. What are the Top 10 cities in State of Paraná with the highest number of Vehicle Thefts?

The top 10 cities are shown in the Figure below. The Top 3 are: *Maringá, Cascavel and Foz do Iguaçu.*



Q6. What are the Top 10 cities in State of Paraná with the highest Disturbing of Piece/Tranquility?

The top 10 cities are shown in the Figure below. The Top 3 are: *Cascavel, Francisco Beltrão and Londrina.*

Top 10 cities in State of Paraná with the highest "Disturbing the Piece/Tranquility", along with their Minimum, Maximum and Standard Deviation.

```

/* Query the created temp table in a SQL cell */

SELECT localidade, AVG(perturbacao_do_sossegotranquilidade) as avg_sossego_tranq,
MIN(perturbacao_do_sossegotranquilidade) as min_sossego_tranq,
MAX(perturbacao_do_sossegotranquilidade) as max_sossego_tranq,
STD(perturbacao_do_sossegotranquilidade) as std_sossego_tranq
FROM 'crimesPR_csv'
GROUP BY(localidade) ORDER BY(AVG(perturbacao_do_sossegotranquilidade)) DESC LIMIT 10

```

localidade	1.2 avg_sossego_tranq	1.2 min_sossego_tranq	1.2 max_sossego_tranq	1.2 std_sossego_tranq
1 Cascavel	706.359166618983	1.154999713897705	961	358.6130794177245
2 Francisco Beltrão	646.3333333333334	430	926	192.7284036572744
3 Londrina	558.7086666623751	1.09399981880188	950	445.063656704879
4 Maringá	528.6666666666666	340	688	145.7060867541102
5 Piraquara	508.5	192	734	179.6850215655175
6 Colombo	497.3653333385785	1.1920000314712534	868	328.99932280237385
7 Paranaguá	492.5	340	772	167.02784190048972
8 São José dos Pinh...	472.85199998909265	1.1119999885559082	998	344.84232719629915
9 Guaratuba	466	236	695	180.92318812136824
10 Pato Branco	456.3333333333333	304	694	172.7039856710512

Q7. What is the mean Drug Trafficking and Drug use Occurrences in Curitiba?

According to the query below, the mean Drug Trafficking and Drug use Occurrences in Curitiba are 1.7 and 3.1, respectively.

What is the mean Drug Trafficking and Drug use Occurrences in Curitiba?

```

SELECT localidade, AVG(ocorrencias_envolvendo_trafico_de_drogas) as avg_trafico_curitiba,
AVG(ocorrencias_envolvendo_usoconsumo_de_drogas) as avg_druguse_curitiba
FROM 'crimesPR_csv'
GROUP BY localidade
HAVING localidade = 'Curitiba'

```

localidade	1.2 avg_trafico_curitiba	1.2 avg_druguse_curitiba
1 Curitiba	1.734333336353302	3.14300004641215

Concerning the last question, what about specifically in year 2020?

Q8. Concerning the last question, what about specifically in year 2020?

According to the query below, the Drug Trafficking and Drug use Occurrences in Curitiba in year 2020 was 2.0 and 3.4, respectively.

The screenshot shows a Databricks workspace with a notebook titled "Untitled Notebook 2024-06-22 14:30:32". The notebook contains a SQL query that filters for crimes in Curitiba in the year 2020. The query is as follows:

```
SELECT localidade, ocorrencias_envolvendo_trafico_de_drogas as trafico_curitiba,
       ocorrencias_envolvendo_usoconsumo_de_drogas as druguse_curitiba
FROM 'crimesPR_csv'
WHERE localidade = 'Curitiba' AND ano = 2020
```

The query results are displayed in a table with 4 columns: localidade, trafico\_curitiba, and druguse\_curitiba. The table shows 1 row of data for Curitiba.

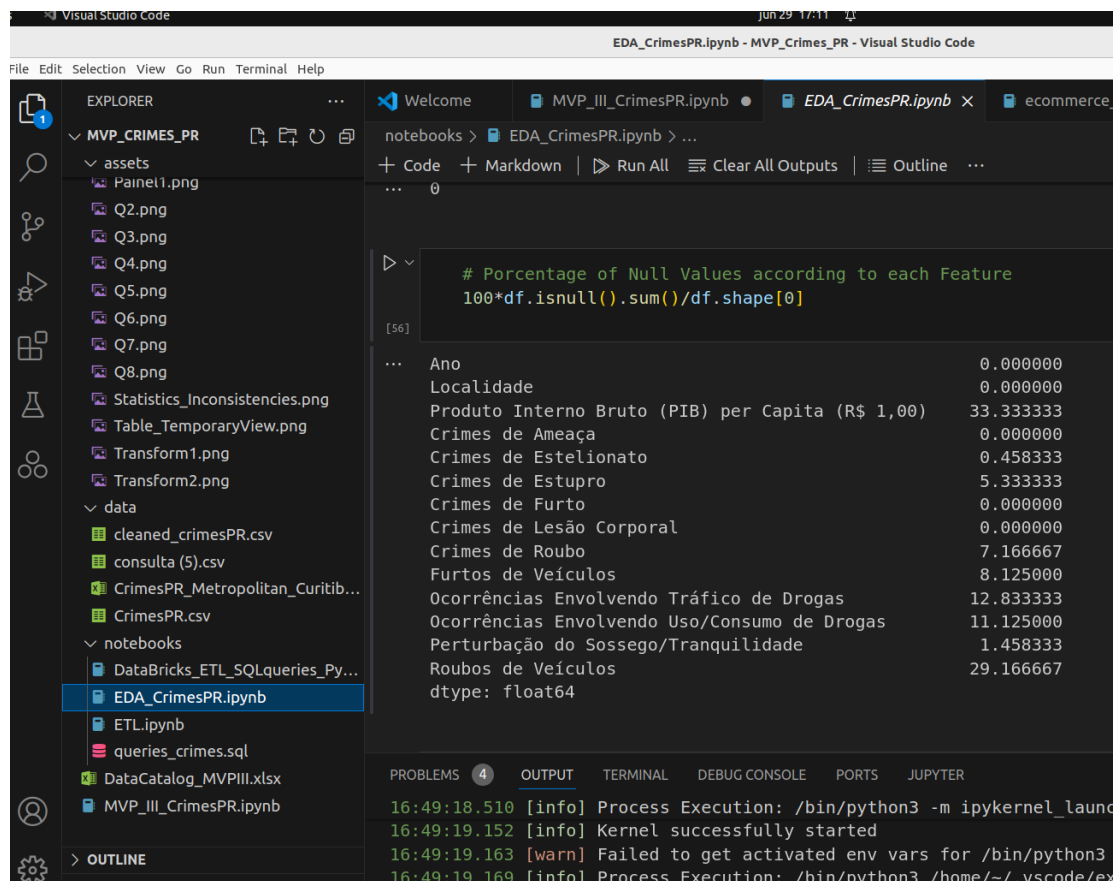
localidade	trafico_curitiba	druguse_curitiba
Curitiba	1.975000023841858	3.3800000114440918

Below the table, there is a section titled "Data Quality Issues Investigations" with a list of tasks. The first task is: "1. Next, it is going to be addressed descriptive statistics about Rape Crimes in Curitiba in order to find inconsistencies in the data."

It is worth emphasizing that in the top 10 cities related to a given crime, there are not cities in some of the last questions like *Curitiba*, and *São José dos Pinhais*, which might be related to a lack of standardization of the measurements in different cities across the State of Paraná, as it will be discussed in the next sections and subsections. So, these results may not be totally true.

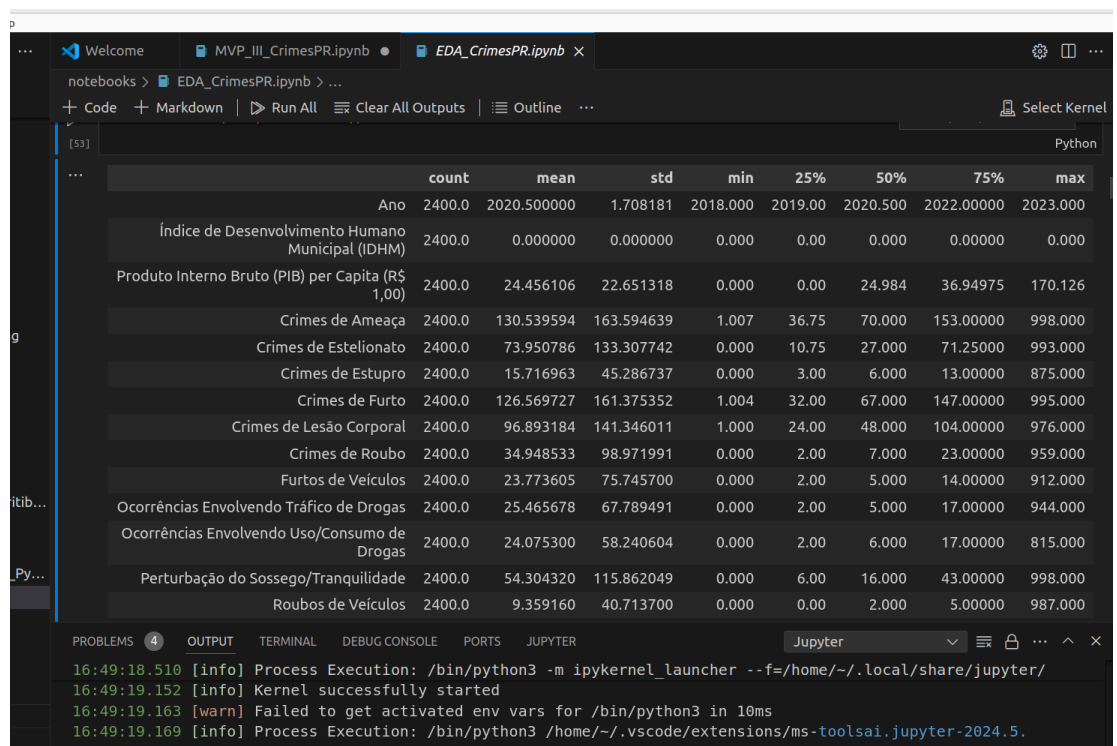
## 5.2 Data Quality

- There have been found many found *Null* values in the dataset, according to no values assigned or the "-" character, which in the **ETL procedure** have been imputed by the median values grouped by *Location*. Below, it is shown a screenshot showing the percentage of null values presented in the raw data, obtained from Exploratory Data Analysis, or *EDA*, in python, as presented in the file **EDA\_CrimesPR.ipynb**.



- Data Quality Issues Investigations

Before addressing specific quality issues, the figure below shows the descriptive statistics of the overall dataset obtained through Exploratory Data Analysis.



1. Next, it is going to be addressed descriptive statistics about Rape Crimes in **Curitiba** in order to find inconsistencies in the data. From this figure, it is noticed that the minimum value is of around 1.0 and the maximum value is of around 875, clearly pointing out to a data standardization issue.

databricks

Untitled Notebook 2024-06-22 14:30:32 Python

File Edit View Run Help Last edit was 56 minutes ago

Run all Vagner Paes's Cluster Share Publish

Workspace

Sort: Name

- 2024-06-22 - DBFS Example
- 2024-06-23 - DBFS Example
- Untitled Notebook 2024-06-22 13:59:02
- Untitled Notebook 2024-06-22 14:30:32

## Data Quality Issues Investigations

1. Next, it is going to be addressed descriptive statistics about Rape Crimes in **Curitiba** in order to find inconsistencies in the data.

```

%sql

SELECT localidade, AVG(crimes_de_estupro) as avg_estupro, MIN(crimes_de_estupro) as min_estupro_curitiba,
MAX(crimes_de_estupro) as max_estupro_curitiba
FROM 'CrimesPR_csv'
GROUP BY(localidade) HAVING localidade = 'Curitiba'

```

(4) Spark Jobs

\_sqlidf: pyspark.sql.dataframe.DataFrame = [localidade: string, avg\_estupro: double ... 2 more fields]

	localidade	1.2 avg_estupro	1.2 min_estupro_curitiba	1.2 max_estupro_curitiba
1	Curitiba	660.1691666642824	1.0149999856948853	875

1 row | 2.45 seconds runtime Refreshed 3 minutes ago

This result is stored as PySpark data frame \_sqlidf and in the IPython output cache as Out[22]. [Learn more](#)

2. Next, it is going to be addressed descriptive statistics about Drug Trafficking Occurrences in **Curitiba** in order to find inconsistencies in the data. Now, the dataset seems ok within Curitiba, but the overall maximum of this feature in the dataset is 944, clearly pointing out to inconsistency/lack of standardization.

databricks

Untitled Notebook 2024-06-22 14:30:32 Python

File Edit View Run Help Last edit was 58 minutes ago

Run all Vagner Paes's Cluster Share Publish

Workspace

Sort: Name

- 2024-06-22 - DBFS Example
- 2024-06-23 - DBFS Example
- Untitled Notebook 2024-06-22 13:59:02
- Untitled Notebook 2024-06-22 14:30:32

2. Next, it is going to be addressed descriptive statistics about Drug Trafficking Occurrences in **Curitiba** in order to find inconsistencies in the data.

```

%sql

SELECT localidade, AVG(ocorrencias_envolvendo_trafico_de_drogas) as avg_traffic_curitiba, MIN(
ocorrencias_envolvendo_trafico_de_drogas) as min_traffic_curitiba,
MAX(ocorrencias_envolvendo_trafico_de_drogas) as max_traffic_curitiba
FROM 'CrimesPR_csv'
GROUP BY(localidade) HAVING localidade = 'Curitiba'

```

(4) Spark Jobs

\_sqlidf: pyspark.sql.dataframe.DataFrame = [localidade: string, avg\_traffic\_curitiba: double ... 2 more fields]

	localidade	1.2 avg_traffic_curitiba	1.2 min_traffic_curitiba	1.2 max_traffic_curitiba
1	Curitiba	1.734333336353302	1.378000020980835	2.132999897003174

1 row | 1.84 seconds runtime Refreshed 5 minutes ago

This result is stored as PySpark data frame \_sqlidf and in the IPython output cache as Out[23]. [Learn more](#)

3. Next, it is going to be addressed descriptive statistics about Bodily Injuries in **Curitiba** in order to find inconsistencies in the data. The same as item 2., with the maximum possible value of around 976.

3. Next, it is going to be addressed descriptive statistics about Bodily Injuries in **Curitiba** in order to find inconsistencies in the data.

```

SELECT localidade, AVG(crimes_de_lesao_corporal) as avg_lesion_curitiba, MIN(crimes_de_lesao_corporal) as min_lesion_curitiba,
MAX(crimes_de_lesao_corporal) as max_lesion_curitiba
FROM "CrimesPR_csv"
GROUP BY(localidade) HAVING localidade = "Curitiba"

```

localidade	1.2 avg_lesion_curitiba	1.2 min_lesion_curitiba	1.2 max_lesion_curitiba
Curitiba	7.769500017166138	6.984000205993652	8.48900032043457

1 row | 1.87 seconds runtime | Refreshed 6 minutes ago

4. Next, it is going to be addressed descriptive statistics about Bodily Injuries in **São José dos Pinhais** in order to find inconsistencies in the data. The same as item 2., with the maximum possible value of around 976. This shows that are many inconsistencies between many cities concerning standardization in the dataset across different cities across the State of Paraná.

4. Next, it is going to be addressed descriptive statistics about Bodily Injuries in **São José dos Pinhais** in order to find inconsistencies in the data.

```

SELECT localidade, AVG(crimes_de_lesao_corporal) as avg_lesion_SJP, MIN(crimes_de_lesao_corporal) as min_lesion_SJP,
MAX(crimes_de_lesao_corporal) as max_lesion_SJP
FROM "CrimesPR_csv"
GROUP BY(localidade) HAVING localidade = "São José dos Pinhais"

```

localidade	1.2 avg_lesion_SJP	1.2 min_lesion_SJP	1.2 max_lesion_SJP
São José dos Pinh...	1.418999898274739	1.2300000190734863	1.6740000247955322

1 row | 1.48 seconds runtime | Refreshed 7 minutes ago

5. More inconsistencies in the dataset are shown in the file named **"./data/CrimesPR\_Metropolitan\_Curitiba\_statistics.xlsx"**. This table shows descriptive statistics of a simplified dataset, such as *minimum*, *maximum*, *median* and *mean values* of each Feature for all the main cities in the Metropolitan area of Curitiba, Paraná, Brazil. In order to demonstrate discrepant values, rows with inconsistencies were highlighted in yellow, while inconsistent values for specific features were highlighted in red.

File Edit View Insert Format Styles Sheet Data Tools Window Help											
Calibri 11pt B I U A E T Center Vertically											
C35:D35	f, Σ = 1,4										
	A	B	D	E	F	G	H	I	J	K	
1		Almirante Tamandare	Araucária	Campo Largo	Campo Magro	Colombo	Curitiba	Fazenda Rio Grande	Pinhais	Piraquara	São José dos Pinhais
2	Crimes de Ameaça_min	1,1	1	1,1	221	2	13,5	1,3	1,2	1	2,3
3	Crimes de Estelionato_min	1,1	1,7	1,2	47	1,3	8,7	1,6	1,2	1,1	1,5
4	Crimes de Estupro_min	53	95	55	16	119	7	77	51	60	95
5	Crimes de Furto_min	1	1,4	1,1	198	2,3	36,2	1,4	2	1	3,2
6	Crimes de Lesão Corporal_min	564	600	508	151	1,2	7	760	556	1	1,4
7	Crimes de Roubo_min	264	230	179	32	1,1	9,3	402	1,2	173	1
8	Furtos de Veículos_min	108	104	73	7	320	2,9	92	125	44	254
9	Ocorrências Envolvendo Tráfico de Drogas_min	27	46	56	6	188	1,4	89	57	142	200
10	Ocorrências Envolvendo Uso/Consumo de Drogas_min	25	61	40	2	86	2,2	57	73	122	139
11	Perturbação do Sossego/Tranquilidade_min	82	214	275	19	1,2	5,2	1,5	282	192	1,1
12	Crimes de Ameaça_max	924	1,4	969	299	3,1	17,3	1,5	998	993	3,2
13	Crimes de Estelionato_max	913	985	740	252	830	40,9	829	738	960	984
14	Crimes de Estupro_max	96	129	69	23	168	2,9	103	77	98	171
15	Crimes de Furto_max	969	1,9	1,6	274	3,2	53,2	1,9	2,5	967	4,7
16	Crimes de Lesão Corporal_max	762	767	842	220	1,8	8,5	888	677	957	1,1
17	Crimes de Roubo_max	639	923	671	69	867	25,4	942	773	565	811
18	Furtos de Veículos_max	176	173	138	21	405	4,9	198	227	70	514
19	Ocorrências Envolvendo Tráfico de Drogas_max	83	123	107	20	456	2,1	203	120	373	308
20	Ocorrências Envolvendo Uso/Consumo de Drogas_max	88	184	220	10	445	3,5	120	156	295	320
21	Perturbação do Sossego/Tranquilidade_max	338	345	568	71	868	10,9	807	489	734	998
22	Crimes de Ameaça_mean	155	1,2	317,1	273,2	2,4	15,1	1,4	167,4	166,5	2,8
23	Crimes de Estelionato_mean	279,7	309,3	276,4	150,2	229,7	25,8	260,9	194,8	325,4	167,1
24	Crimes de Estupro_mean	70,5	114,8	65,7	19,7	148,3	68,2	91	62	77,2	135
25	Crimes de Furto_mean	162,4	1,7	1,7	234,5	2,7	44	1,7	2,3	598,8	4,1
26	Crimes de Lesão Corporal_mean	658,7	694	635,8	172	1,4	7,8	817,3	607,7	679,8	1,1
27	Crimes de Roubo_mean	391,2	575,5	367,3	51,8	355,1	14,4	599,7	398,4	339	267,2
28	Furtos de Veículos_mean	127,5	132,5	101,5	15,5	358,5	4	127,8	190,7	58,2	343,5
29	Ocorrências Envolvendo Tráfico de Drogas_mean	45,7	76,8	80,8	11,5	288,7	1,7	123,3	81,3	262,8	238,5
30	Ocorrências Envolvendo Uso/Consumo de Drogas_mean	50,3	138,3	105,8	6,3	270,7	3,1	100,2	117	174,5	220,8
31	Perturbação do Sossego/Tranquilidade_mean	174,3	276,3	435,8	36,8	497,4	7,6	352,7	387,3	508,5	472,9
32	Crimes de Ameaça_median	1,2	1,2	1,5	275	2,3	15	1,4	1,3	1,2	2,8
33	Crimes de Estelionato_median	171,5	165,6	196,3	153	3	27,4	145,5	2,3	208	4,7
34	Crimes de Estupro_median	67,5	116,5	67,5	20	150,5	779,3	91	59	75	142,5
35	Crimes de Furto_median	1,2	1,8	1,7	236,5	2,6	42,9	1,7	2,3	858,5	4,2
36	Crimes de Lesão Corporal_median	639	708	564	168	1,4	1,8	816,5	601,5	784,5	1,1
37	Crimes de Roubo_median	331	532,5	291,5	56	288,5	11,2	538,5	471,5	314	2,5
38	Furtos de Veículos_median	121	132,5	91	18	351,5	3,9	118	199,5	57,5	323,5
39	Ocorrências Envolvendo Tráfico de Drogas_median	43,5	67	79	11	278,5	1,7	111,5	78,5	264	224
40	Ocorrências Envolvendo Uso/Consumo de Drogas_med	39,5	155	87,5	6,5	284,5	3,4	111	114,5	143	208,5
Sheet											

## 5.3 Data Quality's Issues Solution Suggestion

- In order to better deal with *Null* values, be represented as lack of values or the "-" character, a more robust data collection procedure is recommended;
- The data collected is a **data silo**, which means that there might be duplicated data, limited data access (this could explain null values), poor data quality (e.g., brutal discrepancies in the values of the same feature for different cities, as well as in the same city for different years, even taking into account population's size) because Features (e.g., *Threat Crimes* or *Rape Crimes*) could have been measured in different scales for different cities. This means that **this dataset lacks a standardization procedure** across the cities within the State of Paraná, leading to **inconsistencies in the data**, hardening further data interpretation.

Here are summarized several steps and strategies to address and heal data silos:

- 1. Identify and Understand the Silos Assessment:** Identify all data sources and silos within the organization. In the present dataset, there are many inconcistent values of the same feature for a given city and between different cities, even taking population's size into consideration. Analysis: Understand the type of data each silo contains, how it is stored, and who manages it.
- 2. Establish a Data Governance Framework** Policies and Procedures: Develop and enforce policies for data management, access, and sharing. Roles and Responsibilities: Assign clear roles and responsibilities for data stewardship and management.
- 3. Centralize Data Storage** Data could be centralized stored in a Data Warehouse or a Data Lake. *Data Warehouse:* Implement a data warehouse to centralize structured data from different silos. *Data Lake:* Use a data lake to centralize unstructured and semi-structured data.
- 4. Adopt Modern Data Platforms and Technologies** Cloud Solutions: Leverage cloud-based platforms like DataBricks, AWS, Azure, or Google Cloud for scalable and flexible data storage and integration. Data Virtualization: Use data



virtualization to create a virtual data layer that integrates data from multiple sources without moving it.

5. **Ensure Data Quality and Consistency** Data Cleaning: Regularly clean and standardize data to ensure accuracy and consistency, as well as search for local outliers and inconsistent values.
6. **Facilitate Cross-Departmental Collaboration** Communication: Foster a culture of communication and collaboration among departments (in this case, among cities). Shared Objectives: Align data management objectives with overall business goals to ensure all departments work towards a common purpose. In the present study, collect data aiming to understand the relationship between crimes and socioeconomic factors.
7. **Utilize Data Governance Tools and Platforms** Metadata Management: Use tools to manage metadata and ensure that data is well-documented and easily discoverable. *Data Catalogs*: Implement data catalogs to help users find and understand the data available across the organization (among different cities).

## Visualizations in Tableau

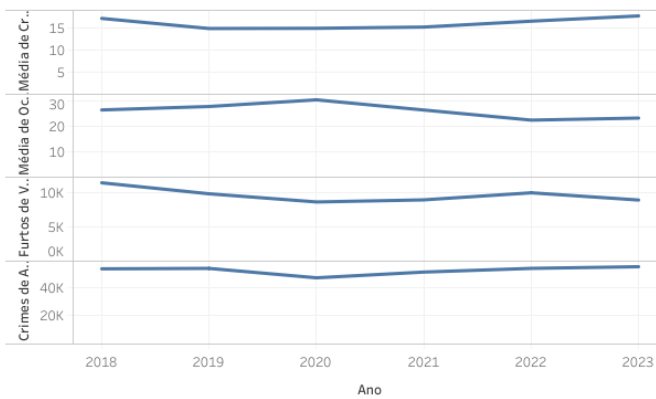
By using the cleaned DataFrame/RDD (file in `/data/cleaned_crimesPR.csv`), which was obtained after *EDA* (see file `./notebooks/EDA_CrimesPR.ipynb`) and it is sql-ready, insightful visualizations were performed in **Tableau**, as shown in the Dashboard below.

Below, there is a brief description of each Figure in the Dashboard:

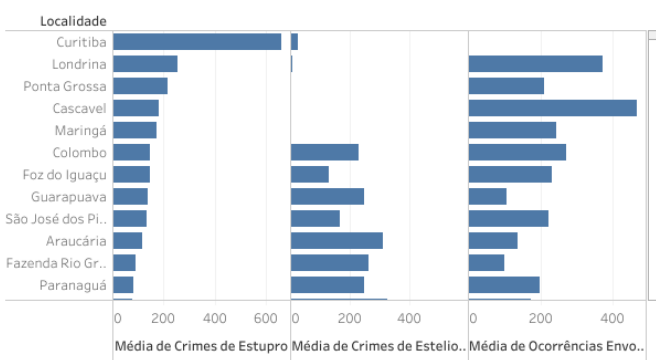
1. Top-left Figure: it shows the average Rape Crimes, Vehicles Theft, Drug Trafficking Occurrences, and Threat Crimes for different years;
2. Bottom-left Figure: it shows the average *Rape Crimes*, *Fraud Crimes*, and *Drug Use Occurrences* for each city within the State of Paraná;
3. Figure on the right corner: it shows in tabular form *Rape Crimes*, *Theft Crimes*, *Vehicles Theft*, and *Drug Trafficking Occurrences* for each city within the State of Paraná.

However, these visualizations could not be made publicly available through an *URL*, because **Public Tableau** was used.

Média de Crimes por Ano



Médias de Crimes por Cidade



Crimes Furtos e Ocorrência de Tráfico de Drogas

Localidade	Crimes ..	Crimes ..	Furtos ..	Ocorrên
Abatiá	23	398	36	2
Adrianópolis	33	151	12	
Agudos do Sul	30	606	36	1
Almirante Tamand..	423	975	765	27
Altamira do Paraná	19	98	8	
Alto Paraíso	13	136	13	9
Alto Paraná	62	970	79	8
Alto Piquiri	37	405	27	7
Altônia	62	800	44	6
Alvorada do Sul	29	597	30	3
Amaporã	17	344	9	2
Ampére	82	780	65	4
Anahy	12	118	13	
Andirá	73	1.125	77	42
Ângulo	13	225	24	1
Antonina	70	2.128	45	13
Antônio Olinto	24	204	11	1
Apucarana	367	14	1.155	1.51
Arapongas	255	8	1.151	85
Arapoti	126	2.022	135	13
Araruna	28	836	82	3
Araucária	689	10	795	46
Ariranha do Ivaí	9	102	12	
Assaí	45	873	46	24
Assis Chateaubria..	89	2.750	242	17
Astorga	80	1.554	163	10
Atalaia	6	196	16	
Balsa Nova	58	978	28	9

## 6. Self-Evaluation

The following questions were (reasonably) answered:

[Q1] - **"What is the correlation between Crimes, such as Rapes and Thefts, and GDP or MHDI?"** Ans.: This question could not be fully answered by using this data, since *GDP* and *MHDI* features presented a very large percentual number of *Null* values. *GDP* could be used, despite 30% of null values, in order to investigate its correlation with several types of Crimes, and, surprisingly, only low-level correlations were found. *MHDI* was removed from the dataset, since there is no data for these cities in the given range of years (2018-2023). Therefore, improving the data gathering/collection step would significantly enhance the data analysis and the conclusions/inferences obtained from data analyses;

[Q2-Q8] - These questions concerned statistics about different types of Crimes in the State of Paraná. It was noticed that "big cities" in population, like *Curitiba* or *São José dos Pinhais* were not present in the top 10 cities concerning a given crime, and this could be related to a lack of standardization of the measurements among different cities. This was confirmed in the investigations performed in the *Data Quality* section, which also pointed out that the data is a *data silos*, meaning that there is, among other things, lack of standardization and communication between different systems in the State of Paraná. Addressing and improving these issues would significantly increase the value obtained from this data, and help battle against Crimes for different cities within the State of Paraná.

It is possible to further extend this project by:

1. Adding more exogenous variables in the dataset, such as *Population*. However, the data silos presented in this dataset would harden creating features like "Crimes by 100K Habitants";
2. Creating more insight visualizations and making them publicly available. However, this would imply in costs in the credit card, since a paid service should be used;
3. Increasing the granularity of the "municipality" of this dataset, by addressing crimes by neighborhood in a given city. However, more socioeconomic variables/features would be necessary to be collected. Again, the data silos presented here would harden these investigations. Therefore, it is mandatory to address and solve these data silos's issues.