



DATA ANALYSIS OF CASE STUDIES

MARY SASSAQUI

Get in touch!



AGENDA

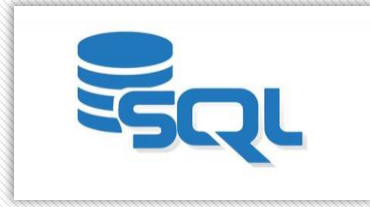
Game Co. - Market Research

Influenza Season

Rockbuster - Movie Rental

Instacart - Online Groceries

Olist – Online marketplace



Upcoming Deposits



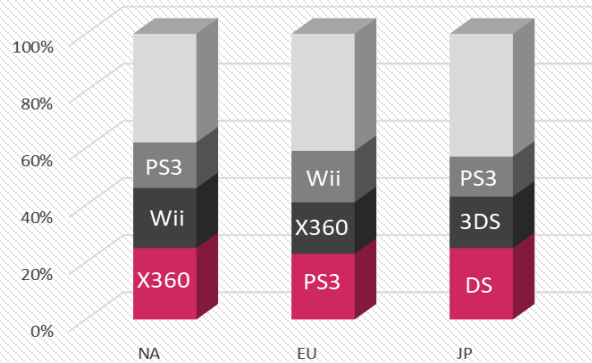
TOOLS

- These programs were used to analyse the data in the following case studies, accordingly to the data available.
- Python displays many advantages for being an open source and having a huge community ready to help. When I want to dig deeper in the coding world, I feel supported.
- SQL shows me that I need to have in mind in advance what I want to explore and insights are really valuable.
- Excel allows me to feel closer to the dataset, once I can check each cell if needed and I am more used to it.
- Tableau and Power Point allows me to better build informative, powerful and easy on the eyes visualizations. I have great pleasure to work with these tools.



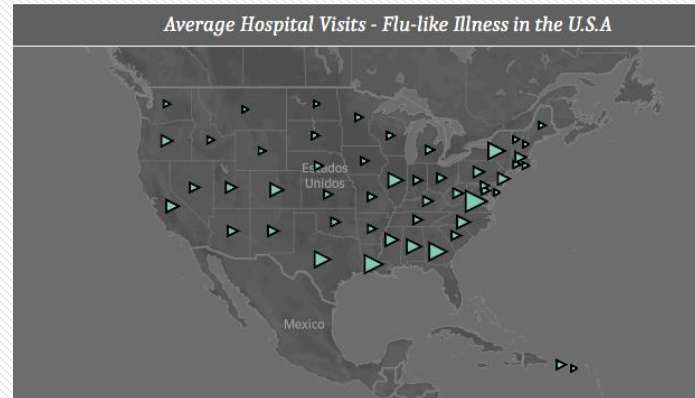
DATA VISUALIZATION

Top 3 Platforms per Region

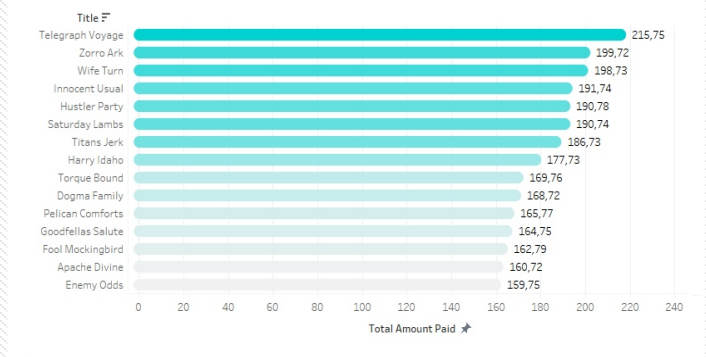


*Data from the last 10 years

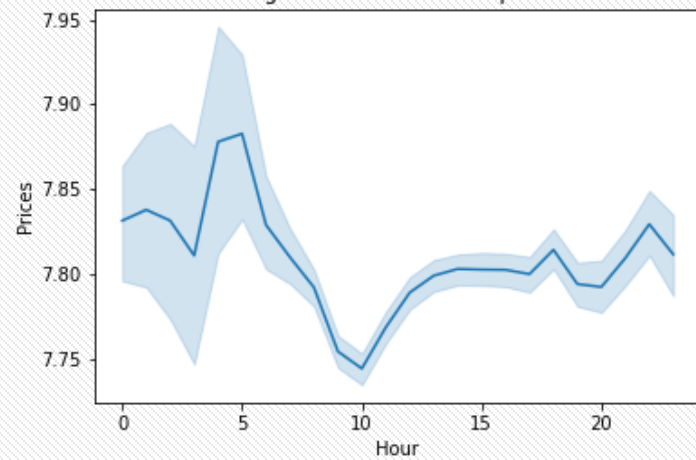
Average Hospital Visits - Flu-like Illness in the U.S.A



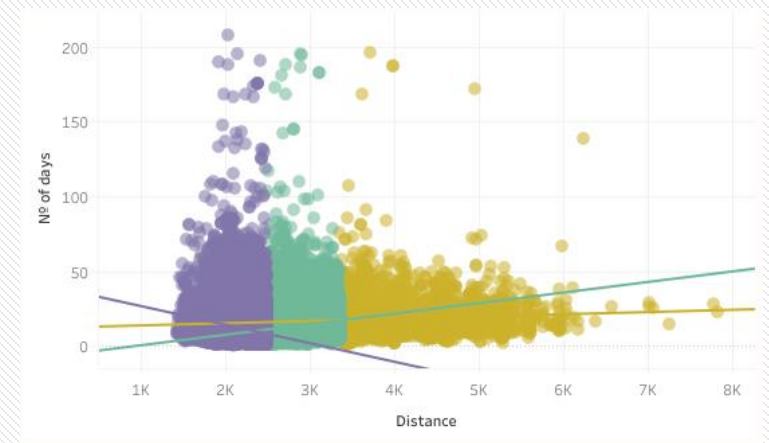
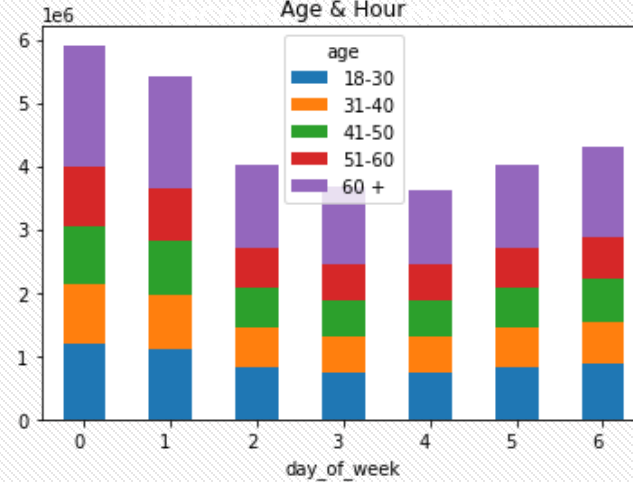
Most Popular Titles



Average Value of Purchase per Hour



Age & Hour



Clustering: to analyse the delivery of Olist

GAME CO.

Market Analysis

- Sales history and Market Share since 1980;
- Regional Analysis;
- Target: Regional top genres and platforms.

Data

- 16.5K titles;
- File: Excel - CSV;
- Regions: North America, Europe, Japan and others;
- Informations: title, platforms, year, genre, publisher;
- Pivot Tables and Charts;
- Data Source: [GVZCharts](#).

Challenges

- Data Cleaning, missing values, formatting and standardizing the dataset;
- Lack of access to current data;
- Decrease of sales do not reflect the real numbers because online channels are not included in the dataset.

REGIONAL ANALYSIS DASHBOARD: NORTH AMERICA



93%

1980 Market

Sales Units: 10,6M



32%

2016 Market

Sales Units: 22,6M



Shooter (52%*)

363M

Misc (51%)

288M

Sports (50%)

396M

Fighting (50%)

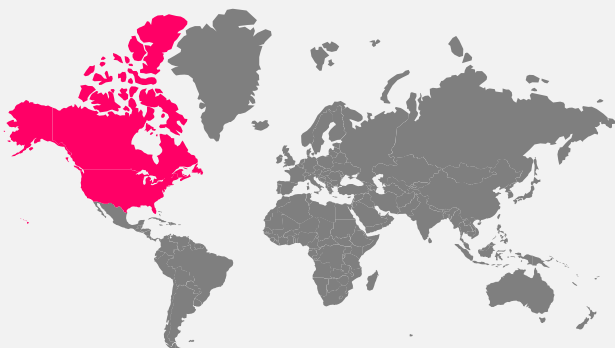
95M

Simulation (49%)

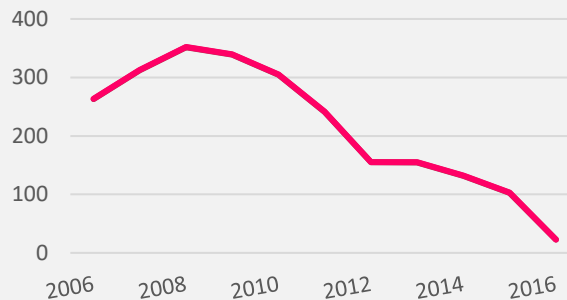
109M

* % from Global Sales

Market Location



Sales



*Data from the last 10 years



NA lost market share



Increase and decrease of sales

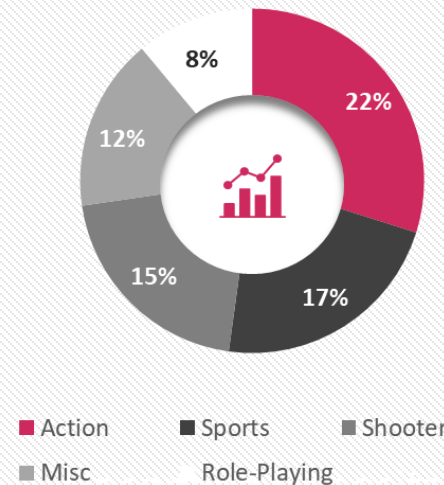


STABLISHING TARGETS: GENRES AND PLATFORMS

Year	(Multiple Items)
2006-2016	
Row Labels	Sum of NA_Sales
Action	529,05
Sports	395,53
Shooter	363,49
Misc	287,94
Role-Playing	196,06
Platform	135,65
Racing	132,9
Simulation	108,78
Fighting	94,54
Adventure	61,98
Puzzle	43,34
Strategy	28,65
N/A	0,03
Grand Total	2377,94

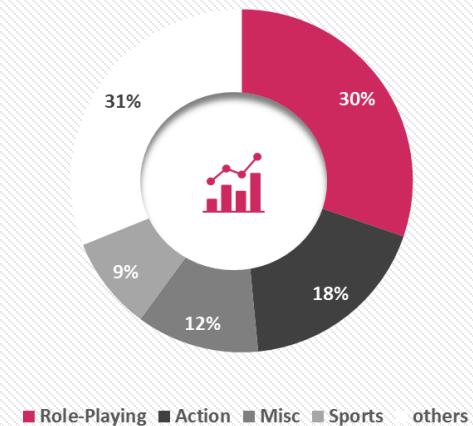
The differences between regions were well observed through Pivot Tables and Charts

NA and Eu % are similar:

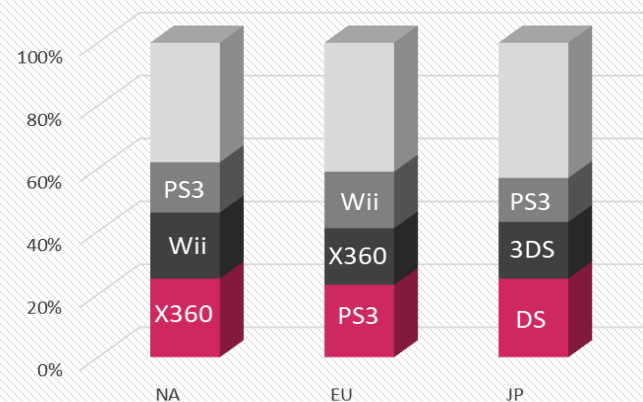


Japan has as different top seller:

Role-playing



Top 3 Platforms per Region



Year	(Multiple Items)
2006-2016	
Row Labels	Sum of NA_Sales
X360	587,44
Wii	497,28
PS3	388,9
DS	331,4
PS2	114,89

*Data from the last 10 years

INFLUENZA SEASON

Goals

- ❁ Work with a staff agency to relocate professionals during Influenza Season;
- ❁ Formulate and test hypothesis to background the relocation;

Data

- ❁ File: Excel - CSV;
- ❁ Informations: Hospital Visits 2010 - 2019, Mortality 2009 - 2017, Vaccination in children until 35 months in 2017, laboratory tests 2010 - 2015, Census of the population 2010 – 2019;
- ❁ Tableau: Data Visualization;
- ❁ Data Source: CDC Years 2009-2017 and US Census Bureau.

Challenges

- ❁ Data Cleaning, inconsistency, different levels of granularity;
- ❁ Detect the high risk groups, create the correlation of the variables, elaborate the hypothesis to support it and test it, was a nice challenge;
- ❁ Learn how to use Tableau and write a Report Interim;
- ❁ Lack of access to current data, including the pandemic Covid-19 and its impacts on the flu season numbers.

DESCRIPTIVE ANALYSIS

- 🦠 If the population from high-risk groups is higher in some states, then there are more hospitalizations flu illness related.
- 🦠 High risk groups: < 5 years and 65 + years.

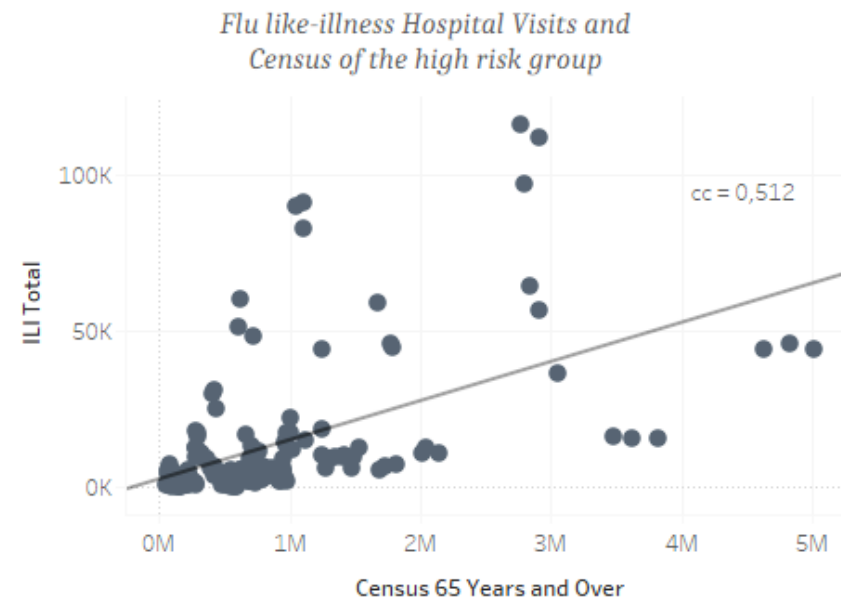
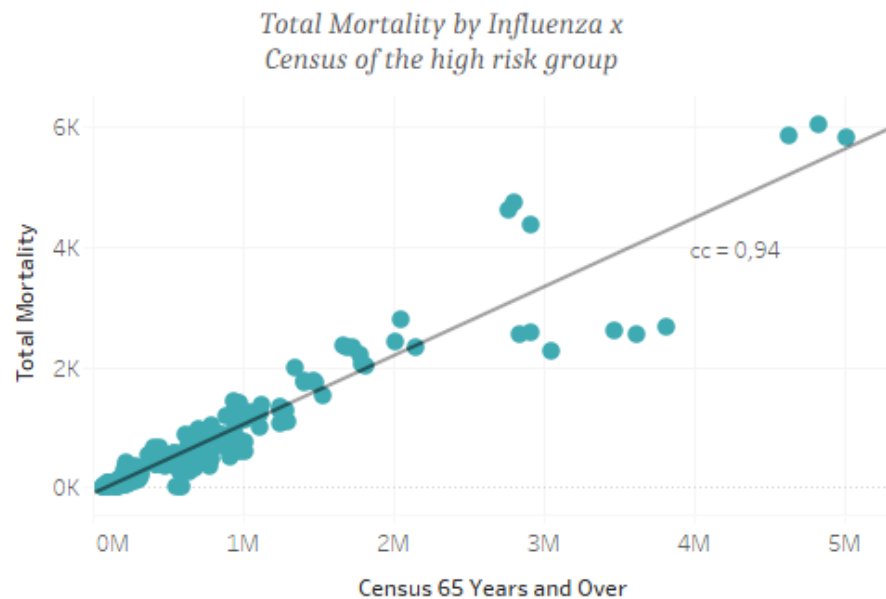
* But... how strongly are they correlated?



The **correlation coefficient** is measured between two variables to understand how strongly are they correlated or connected.
The values to measure the relationship are:

cc = 0: no relationship - 0.1-0.3: weak relationship - 0.3-0.5: moderate relationship - 0.5-1.0: strong relationship

There is a **strong** relationship among the three variables!



HYPOTHESIS TESTING

- Independent Variable: Number of vulnerable people (sum of population < 5 years and 65+)
- Dependent Variables: Mortality and Hospitalization by flu-like illness.

Test 1

Null Hypothesis

$$H_0: \mu_{\text{high}} = \mu_{\text{low}}$$

1. If States have higher number of people from high risk groups, then they have the same mortality rates from flu like illness.

Alternative Hypothesis

$$H_A: \mu_{\text{high}} \neq \mu_{\text{low}}$$

1. If States have higher number of people from high risk group, then they do not have the same mortality rates from flu like illness.

T test

two - tailed because it can be higher or lower

	States with higher Average of mortality per 100,000 habitants	States with lower Average of mortality per 100,000 habitants
Mean	0,129	0,017
Variance	0,010	0,000
Observations	26,000	26,000
Pooled Variance	0,005	
Hypothesized Mean Difference	0,000	
df	50,000	
t Stat	5,570	
P(T<=t) one-tail	0,000	
t Critical one-tail	1,676	
P(T<=t) two-tail	0,000	
t Critical two-tail	2,009	

Test 2

Null Hypothesis

$$H_0: \mu_{\text{high}} = \mu_{\text{low}}$$

2. If States have a higher number of people from high risk groups, then they have the same rates of hospitalized people with flu-like illness.

Alternative Hypothesis

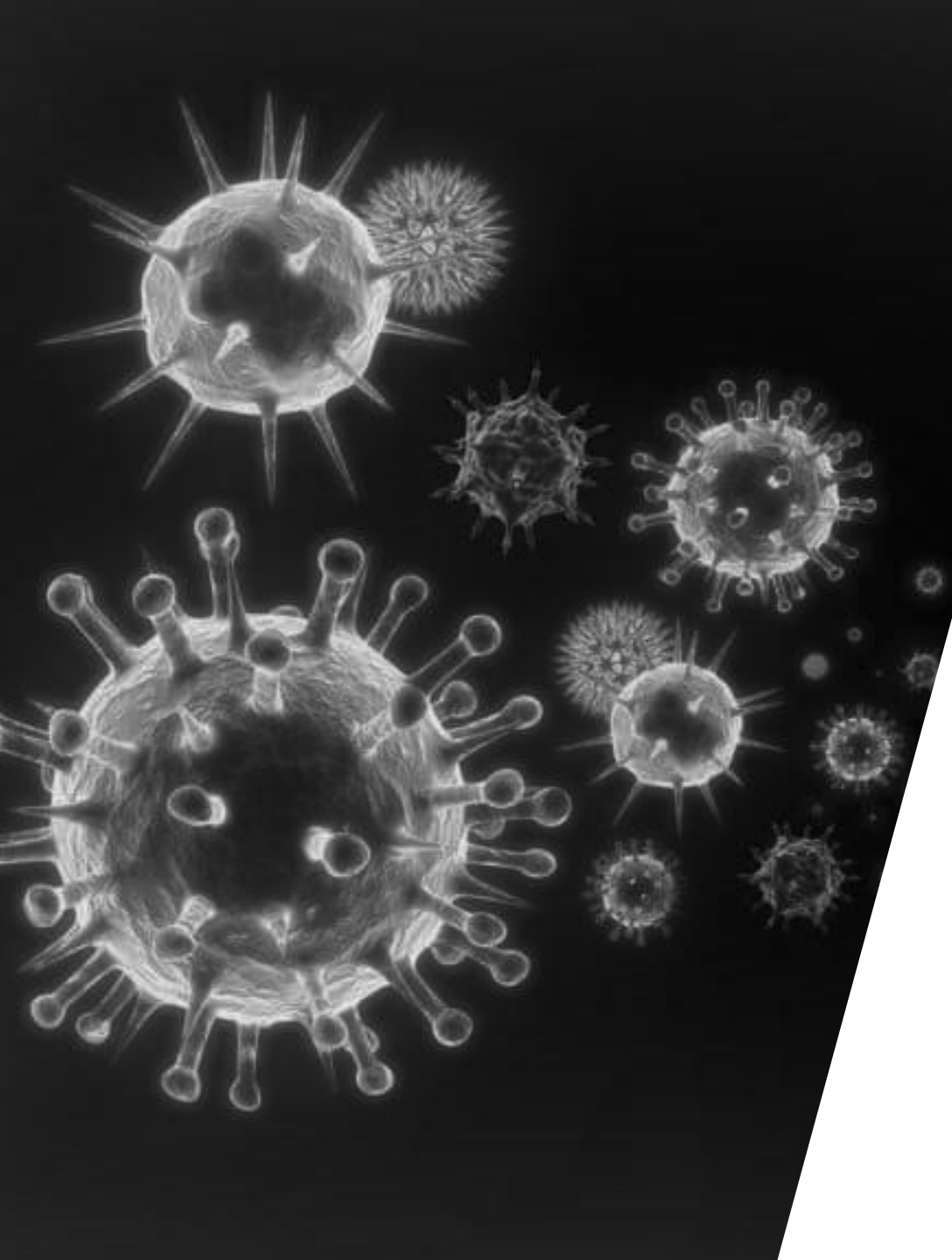
$$H_A: \mu_{\text{high}} \neq \mu_{\text{low}}$$

2. If States have a higher number of people from high risk groups, then they don't have the same rates of hospitalized people with flu-like illness.

T test

two - tailed because it can be higher or lower

	States with higher Average of flu-like illness like hospitalizations per 100,000 habitants	States with lower Average of flu-like illness hospitalizations per 100,000 habitants
Mean	1,763	0,196
Variance	3,367	0,014
Observations	26,000	26,000
Pooled Variance	1,690	
Hypothesized Mean Difference	0,000	
df	50,000	
t Stat	4,347	
P(T<=t) one-tail	0,000	
t Critical one-tail	1,676	
P(T<=t) two-tail	0,000	
t Critical two-tail	2,009	



For both analyses, the P was lower than 0,05 and indicates that both null hypotheses can be rejected.

Meaning that **States with higher number of high-risk people don't have the same rates of mortality and hospitalizations than other states.**

* *Final Considerations:*

- 🦠 *Each State has different needs;*
- 🦠 *States with higher number of vulnerable people, hospitalizations and deaths: should be treated as **priority**.*
- 🦠 *The best treatment is always prevention: Vaccinations, Information and Hygiene.*

My [Tableau](#) vizualizations for this project

Check out the [Interim Report](#)

ROCKBUSTER STEALTH

Goals

- Help the company understand its operations and revenue.
- Keep the company competitive in the market.

Data





- Database: SQL;
- Program: PostgreSQL, DB Visualizer, Tableau;
- Data Source: Rockbuster data.




Challenges

- Understand the complexity of the DB;
- Program in SQL to extract the information of the DB;
- Write a Data Dictionary;
- Visualization with Tableau;
- Presentation in Power Point.

DATA DICTIONARY

All the data connections were described here:

Payment			
Columns	Data type	Description	
 payment_id	SERIAL	Primary key of payment fact table, is the payment id number	
 customer_id	SMALLINT	Foreign key, is a customer id number and has a dimension table	
 staff_id	SMALLINT	Foreign key, is a staff id number and has a dimension table	
 rental_id	INTEGER	Foreign key, is a rental id number and has a dimension table	
amount	NUMERIC(5,2)	Numeric data, is the amount of payment	
payment_date	TIMESTAMP(6) WITHOUT TIME ZONE	Date and time registered at the moment of the last update	

Links to			
Table	Join	Description	
 rental	payment.inventory_id = inventory.inventory_id	Foreign Key constraint referencing inventory.inventory_id	
 customer	payment_customer_id = customer.customer_id	Foreign Key constraint referencing customer.customer_id	
 staff	payment.staff_id = staff.staff_id	Foreign Key constraint referencing staff.staff_id	



SQL CODING & OUTPUT

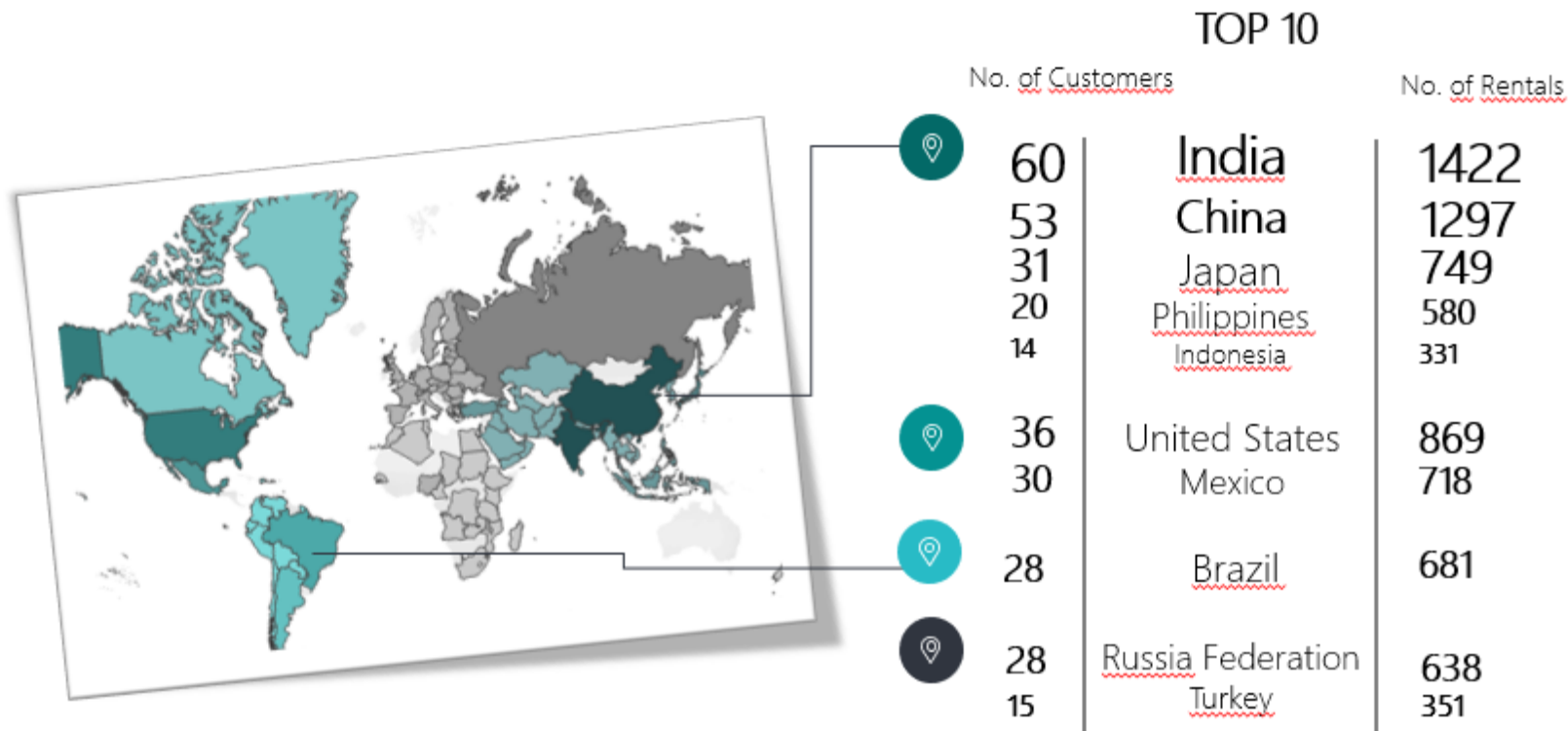
Coding in SQL was interesting and rewarding, once I could understand, I could expand the search for more information:

```
SELECT SUM(payment.amount) AS total_amount_paid,  
       COUNT(payment.payment_id) AS total_payments,  
       country.country  
FROM payment  
INNER JOIN rental ON payment.rental_id = rental.rental_id  
INNER JOIN inventory ON rental.inventory_id = inventory.inventory_id  
INNER JOIN film ON inventory.film_id = film.film_id  
INNER JOIN film_category ON film_category.film_id = film.film_id  
INNER JOIN category ON film_category.category_id = category.category_id  
INNER JOIN customer ON payment.customer_id = customer.customer_id  
INNER JOIN address ON customer.address_id = address.address_id  
INNER JOIN city ON address.city_id = city.city_id  
INNER JOIN country ON city.country_id = country.country_id  
GROUP BY country  
ORDER BY total_amount_paid DESC
```

total_amount_paid	total_payments	Country
6034,78	1422	India
5251,03	1297	China
3685,31	869	United States
3122,51	749	Japan
2984,82	718	Mexico
2919,19	681	Brazil
2765,62	638	Russian Federation
2219,7	530	Philippines
1498,49	351	Turkey
1352,69	331	Indonesia

PRESENTATION

Where are customers with a high lifetime value based?



Top 10 Sales Numbers

\$6.035	India
\$5.251	China
\$3.685	U.S.A
\$3.123	Japan
\$2.985	Mexico
\$2.919	Brazil
\$2.766	Russia
\$2.220	Philippines
\$1.498	Turkey
\$1.353	Indonesia

Check the whole Presentation in [Tableau!](#)

INSTACART

Goals

- Profile the customers and their behaviour;
- Target marketing actions accordingly.

Data

- Database: Open-source data and customer data set (fictitious);
- Program: Anaconda/Jupyter, Python, Matplot/Seaborn;
- Data Souce: [Instacart](#), [Dictionary](#), [Customer dataset](#)

Challenges

- Understand the complexity of the data;
- Program in Python to extract the information;
- Learn programming language and expand through Stackflow;
- Cleaning, Wrangling, Labeling, Creating new columns, Merging data sets;
- Visualization with Matplot/Seaborn;
- Create a Final Report with findings and observations.

```

In [20]: ords_prods_customers_new.head()
Out[20]:
  order_id  user_id  eval_set  order_number  day_of_week  hour  days_since_prior_order  product_id  add_to_cart_order  reordered  _merge  product_name  price
0    2539329      1    prior         1         2         8             0.0             196             1             0    both    Soda Water  1.50
1    2398795      1    prior         2         3         7             15.0             196             1             1    both    Soda Water  1.50
2    473747      1    prior         3         3        12             21.0             196             1             1    both    Soda Water  1.50
3    2254736      1    prior         4         4         7             29.0             196             1             1    both    Soda Water  1.50
4    431534      1    prior         5         4        15             28.0             196             1             1    both    Soda Water  1.50

In [27]: # Create a income flag
ords_prods_customers_new.loc[ords_prods_customers_new['income'] < 50000, 'income_level'] = 'Low-income'

C:\Users\Mary\anaconda3\lib\site-packages\pandas\core\indexing.py:1599: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-copy
    self.obj[key] = infer_fill_value(value)
C:\Users\Mary\anaconda3\lib\site-packages\pandas\core\indexing.py:1720: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-copy
    self._setitem_single_column(loc, value, pi)

In [28]: ords_prods_customers_new.loc[(ords_prods_customers_new['income'] <= 100000) & (ords_prods_customers_new['income'] > 50000), 'income_level'] = 'Mid-income'

C:\Users\Mary\anaconda3\lib\site-packages\pandas\core\indexing.py:1720: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-copy
    self._setitem_single_column(loc, value, pi)

In [29]: ords_prods_customers_new.loc[ords_prods_customers_new['income'] > 100000, 'income_level'] = 'High-income'

C:\Users\Mary\anaconda3\lib\site-packages\pandas\core\indexing.py:1720: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-copy
    self._setitem_single_column(loc, value, pi)

In [30]: ords_prods_customers_new['income_level'].value_counts(dropna = False)
Out[30]:
High-income    14207028
Mid-income     13394982
Low-income     3362554
Name: income_level, dtype: int64

In [31]: ords_prods_customers_new.shape
Out[31]: (30964564, 38)

In [32]: ords_prods_customers_new.head()
Out[32]:

```

HANDS ON DATA!

PYTHON = CHALLENGE

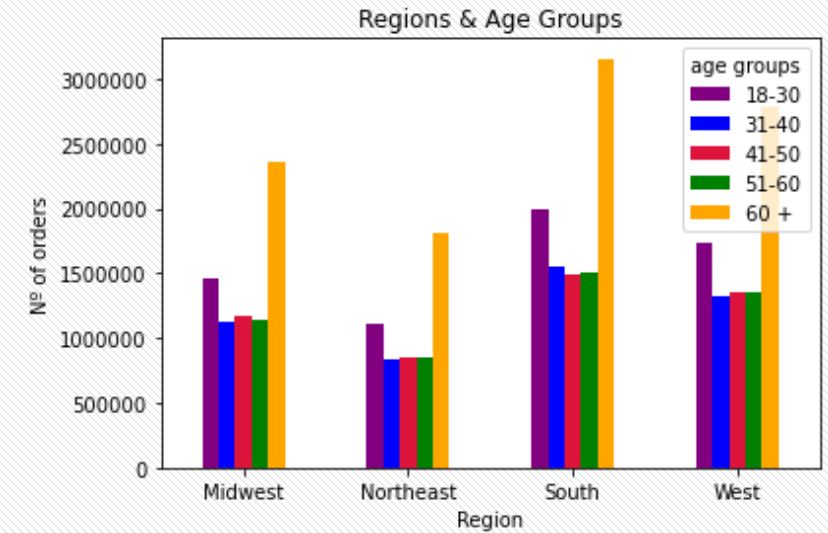
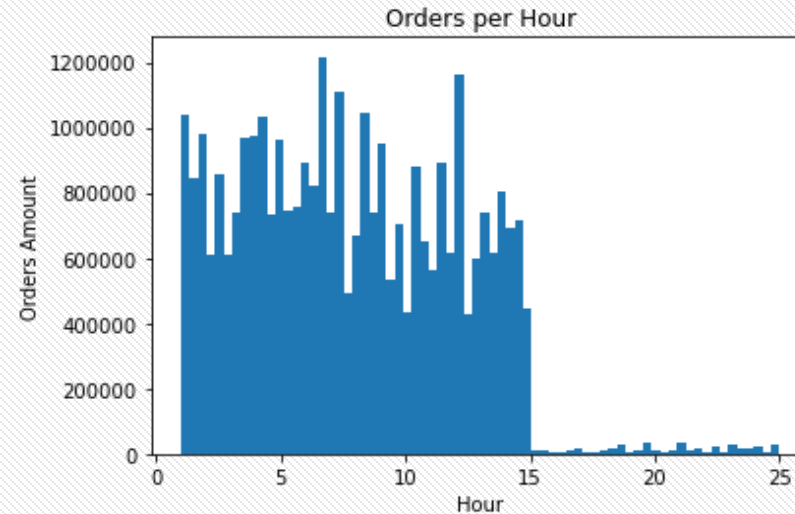
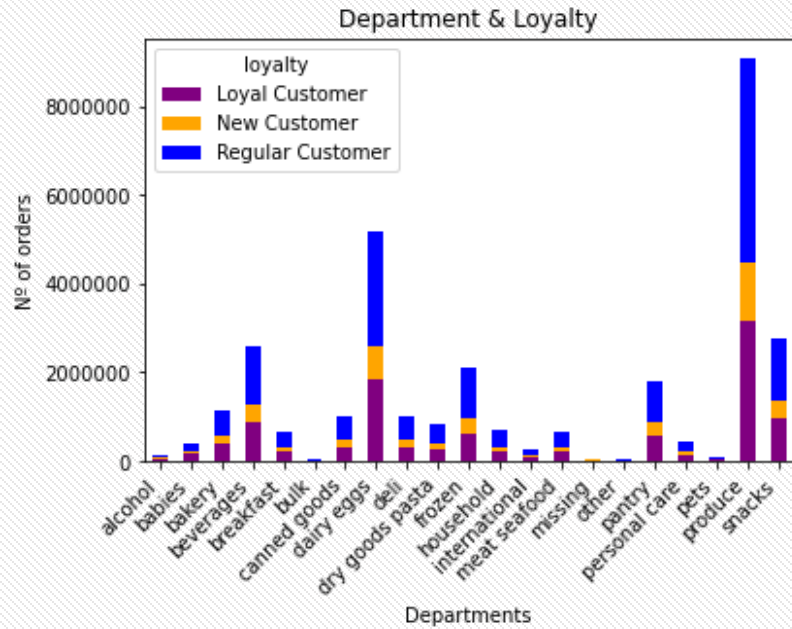
- Data Wrangling, Data Consistency, Merging, New Variables, Labeling, Grouping, Aggregating.



This was all new!

- A lot was taught and learned while dealing with this huge data set;
- I searched for codes in Stackflow as I wanted to improve in every step.

VISUALIZATIONS & RECOMMENDATIONS



Upcoming Deposits

BUSINESS PRIORITIES

- Loyal customers are responsible for 41% of the orders of Babies and Bulker Departments, special offers could be directed to this group;
- Email marketing could be sent in the afternoon to target morning and lunch-time purchases, also weekends, increasing avg ticket;
- South and West regions present more orders (31% and 28%) , also 60+ years old is the public that most buy in Instacart.

OLIST

Goals

- ❖ Perform Exploratory Analysis;
- ❖ Obtain insights about: customers, company's growth, reviews.

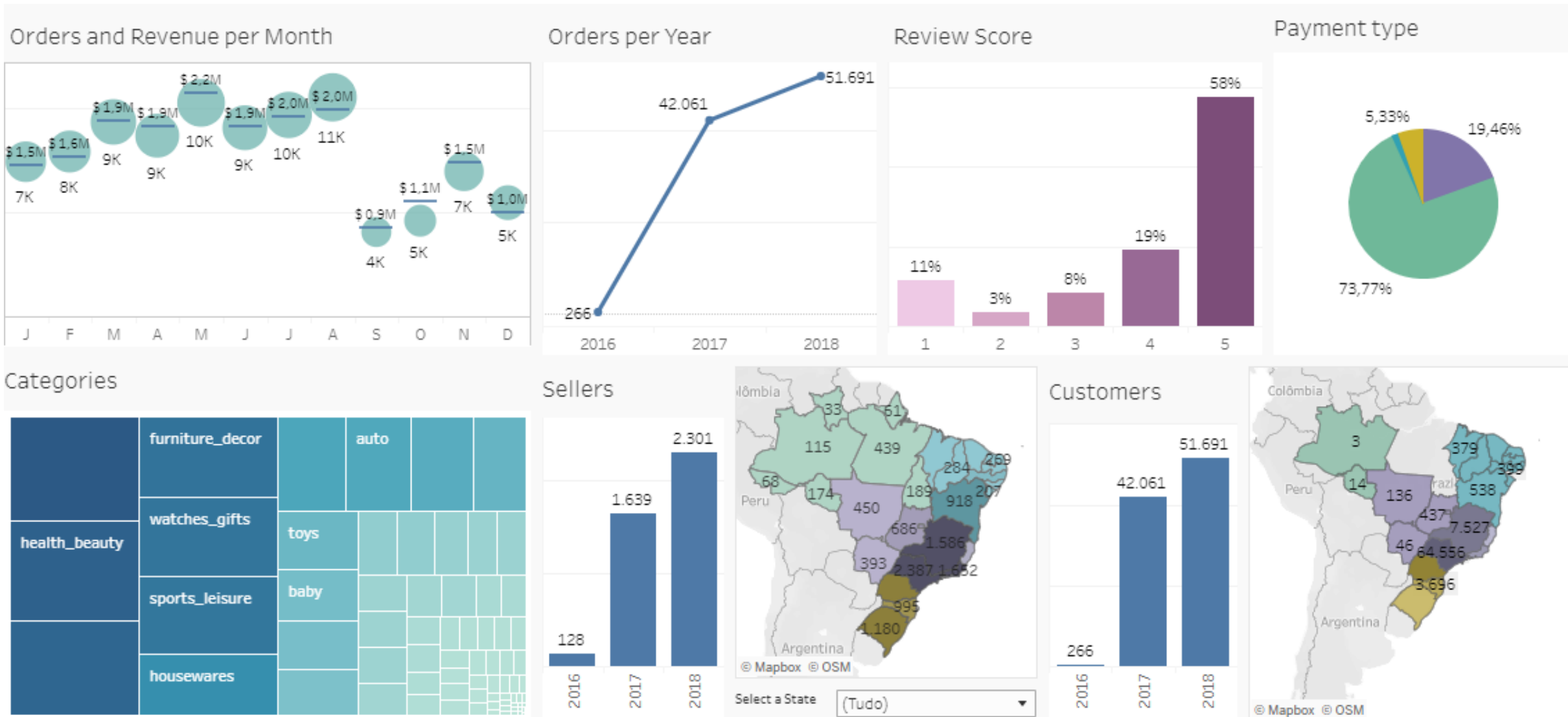
Data

- ❖ Database: Open-source data;
- ❖ Program: Anaconda/Jupyter, Python, Pandas, Matplot/Seaborn, Statsmodel, Tableau;
- ❖ Data Souce: [Kaggle](#)

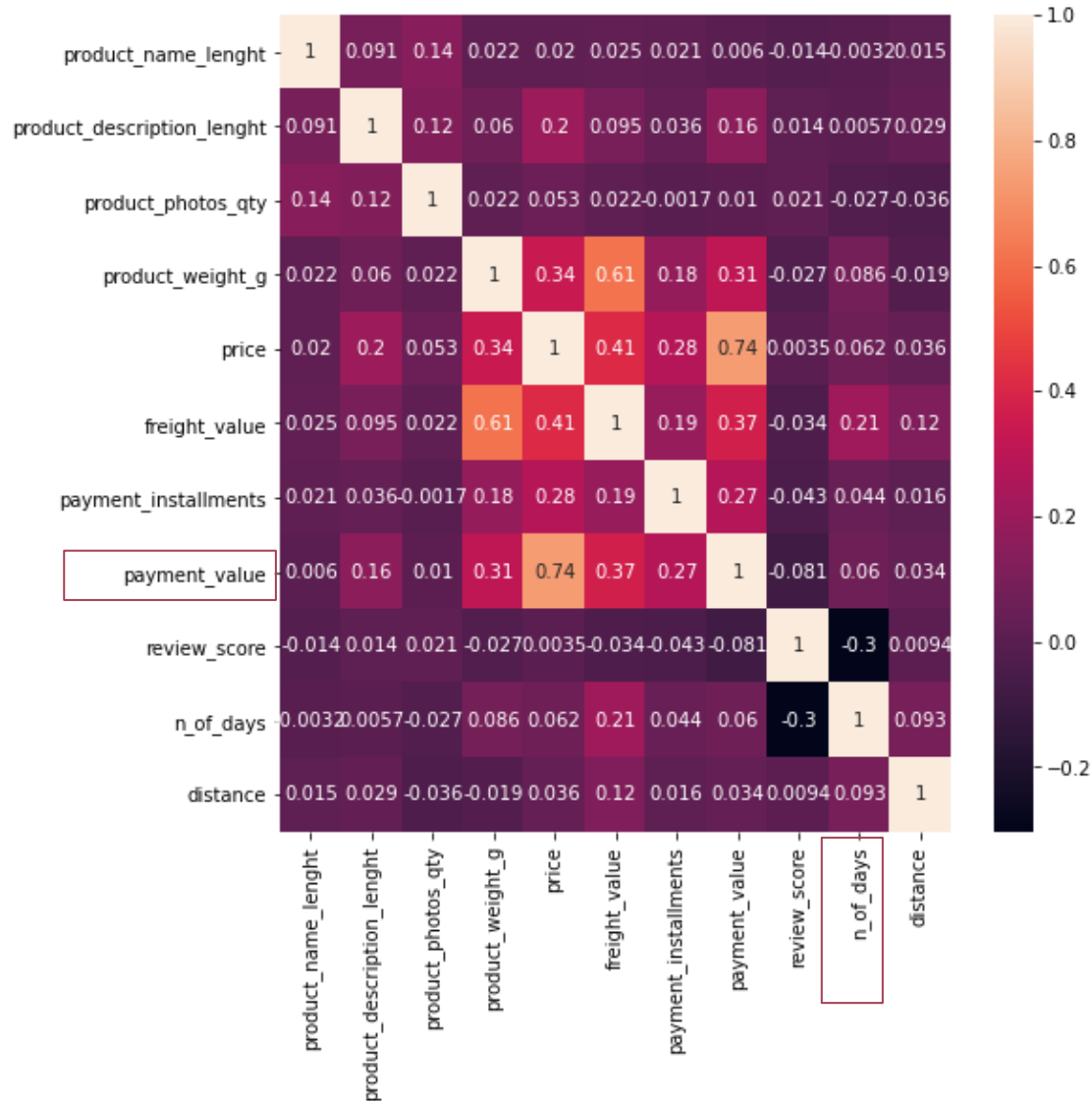
Challenges

- ❖ Understand the complexity of the data;
- ❖ Program in Python to extract the information;
- ❖ Expand knowledge through Stackflow and other blogs;
- ❖ Cleaning, Wrangling, Merging, Clustering, Statistics finding;
- ❖ Visualization with Matplot/Seaborn;
- ❖ Create an interective dashboard to obtain a business overview;
- ❖ Create a final presentation with findings and observations.

MY INTERACTIVE BUSINESS DASHBOARD IN TABLEAU



CORRELATION MATRIX AND CHANGE OF PLANS

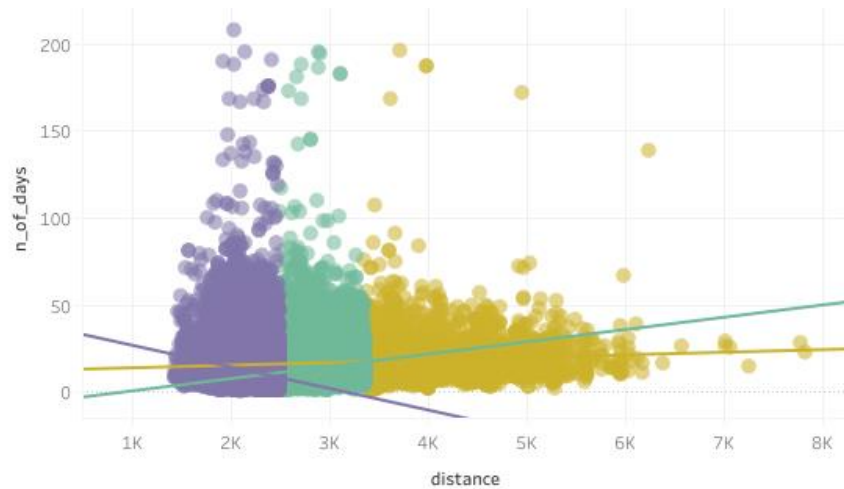


- I did a **correlation matrix** with the data available, performed a **linear regression** with categorical and numerical data and proceeded to **clustering** to try and find patterns.
- My linear regression with review_score and n_of_days to process the order didn't had good results due to the high mean squared error.
- Therefore, I had to change my focus on payment value and distance between seller and customer.
- Using these variables, I proceeded to clustering.

RESULTS

&

RECOMMENDATIONS



Cluster 1: the longer the distance, more expensive is the order.

Cluster 3: the shorter the distance, the cheaper is the order.

Cluster 2: the longer the distance, the cheaper is the order.

So, is it a matter of regional logistic costs or lower product's prices?

Cluster	1	2	3
avg_distance	3652	2788	2282
avg_payment	354	170	166
avg_n_of_days	18	12	9

Next steps:

- ❖ Check the reasons why the avg price is cheaper for cluster 2;
- ❖ Keep track of reasons for low review score;
- ❖ Check the products, categories and regions where the logistic costs are lower;
- ❖ Adopt another method for analyzing the relationship between review_score and n_of_days to process the order;
- ❖ Gather more data to trace seasonality;
- ❖ Check how many of the sellers and customers are active and not only registered.



Check my [Tableau](#) presentation!

THANK YOU



Get in touch!