Business Optimisation through Analysis of Brazilian e-commerce data

Oge Ibezi

Introduction

The analysis of e-commerce dataset is carried out to give insight into the business' operational dynamics. This analysis was done using Power BI.

Exploratory data analysis done to understand data structure, identify key variables and modelling of database relationships.

Modelling of database enabled the linkage of entities and generation of complex insights.

Vendor location, customer demographics, clustering, sentiment analysis, product categories and sales trends were explored, with a view to draw out factors that could impact effective strategic development and business optimization.

Utilized Power BI functionalities for ETL processes such as importing data from different sources, use of power query for data cleaning and transformation, DAX for expanding the dataset and charts and dashboard using Power BI desktop and service.

Comparing customers by state, sellers by state and average freight cost by state


The Southeast region of Brazil, which includes states such as São Paulo, Rio de Janeiro, Minas Gerais, and Espírito Santo, is generally considered the most economically vibrant region in Brazil

We can see that most customers and sellers are based around Sao Paulo region and the further away we go from here the less the size of the bubbles.

The bigger the size of the bubbles in the average freight value by geolocation chart, the higher delivery costs for sellers to deliver in those outer regions, whereas the bubbles near Sao Paulo are smaller. This be the reason why there are less orders from regions further out.

# Analysis of delivery time by freight value and spatial distribution across states
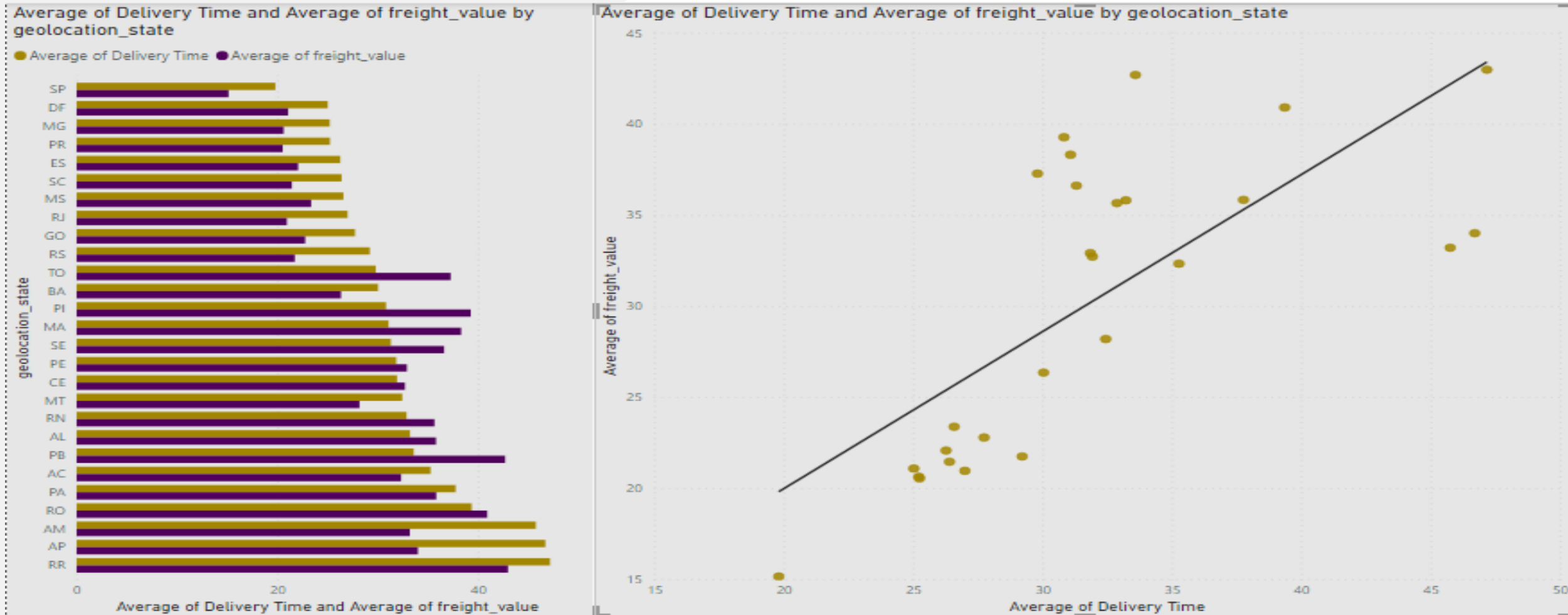
There is a correlation between the   time of delivery and cost of shipping products
The higher the delivery time the higher the shipping cost for sellers/vendors.
Also, the further or more remote the region with lower economic activity, the higher the shipping cost

Analysis of sentiment spread across regions

customer_id	review_score	Sentiment	Image
00072d033fe2e59061ae5c3aff1a2be5	5	Positive	
0009a69b72033b2d0ec8c69fc70ef768	4	Positive	
000bf8121c3412d3057d32371c5d3395	5	Positive	

Word map of sentiments

sentiment

Positive	Neutral	Negative

Mapping of Sentiment by City

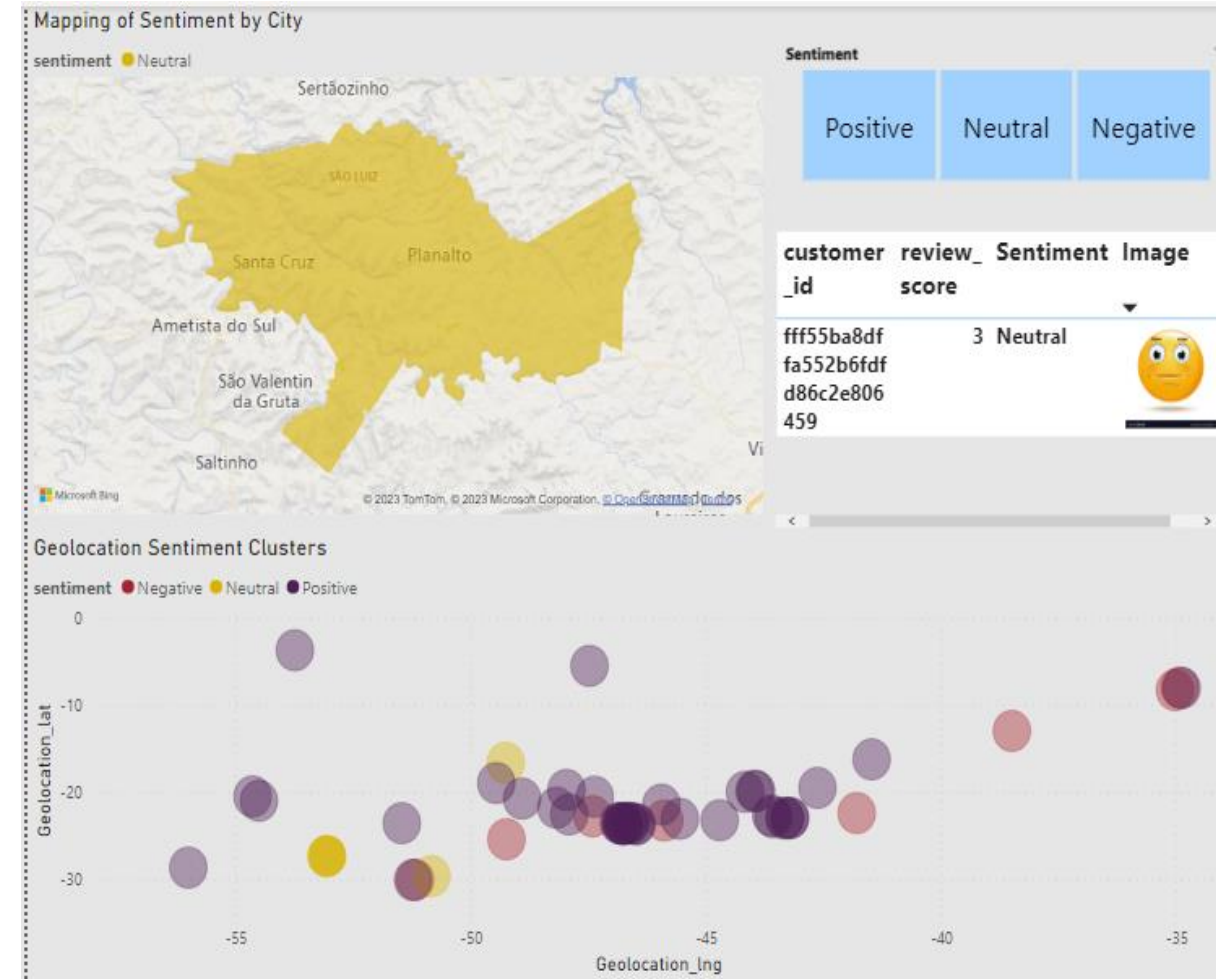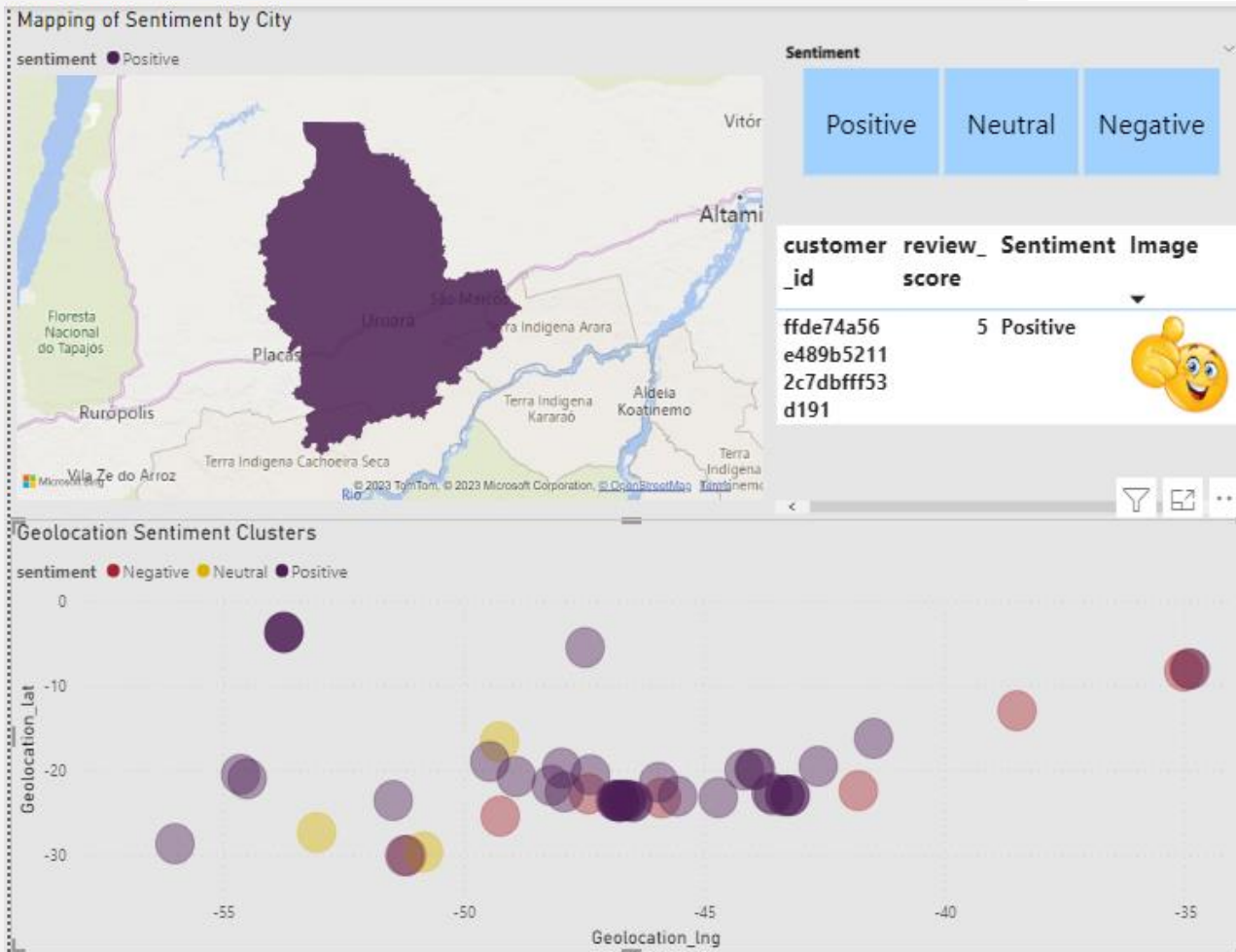sentiment ● Positive
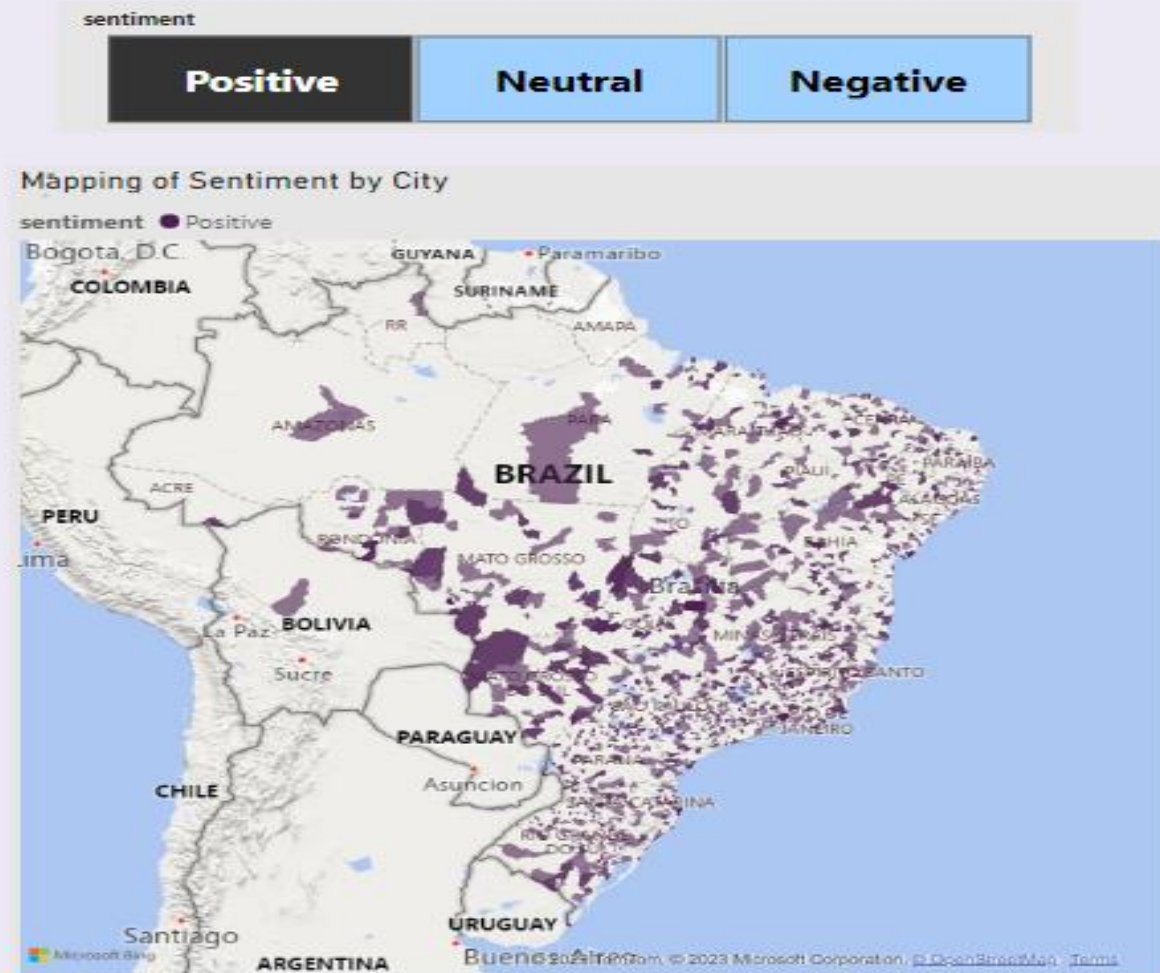
Sentiment analysis of customer reviews were categorised into positive, negative and neutral feedback.

The word map gives insight into the sentiment landscape. Words like delivery, arrived, earlier, deadline, before all point to delivery time as a factor for positive sentiment.
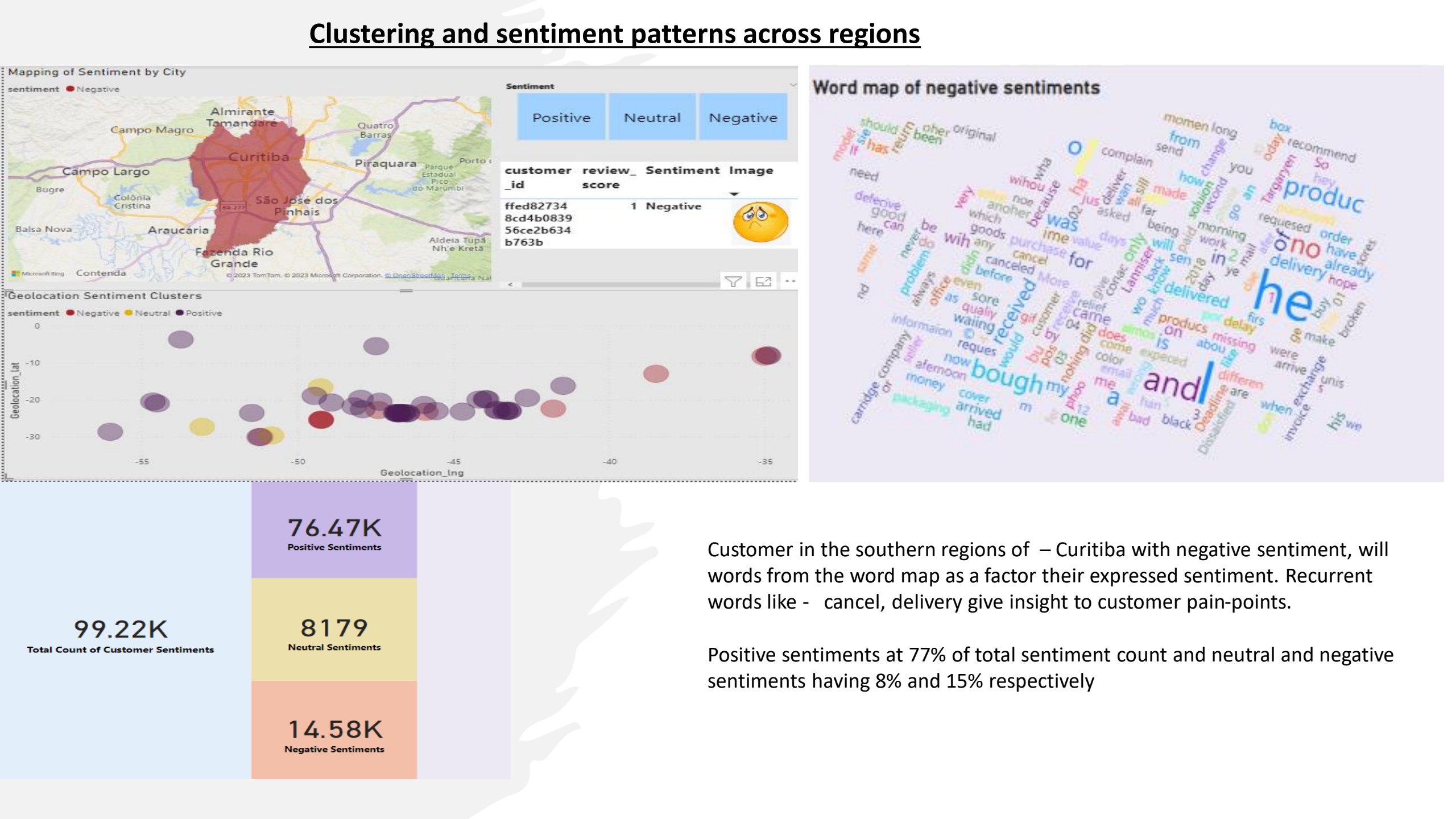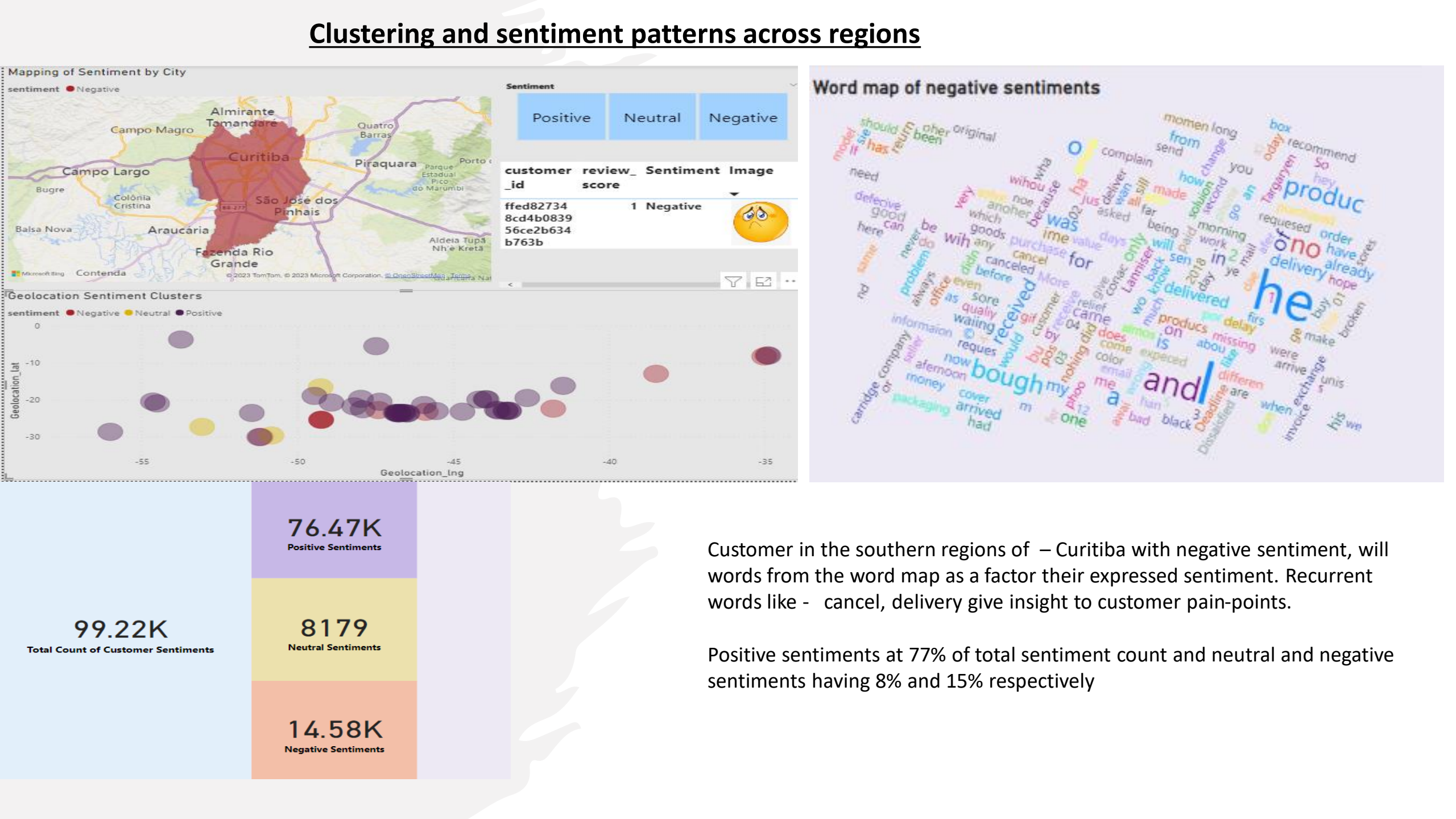
Clustering and sentiment patterns across regions

Clustering analysis showed into patterns of customer sentiments across the regions. Customers in the southern regions of – Sao Marcos and Santa Cruz had positive and neutral sentiments, respectively.

The word map gives insight into the sentiment from these customers

Clustering and sentiment patterns across regions

Mapping of Sentiment by City

Sentiment: Positive | Neutral | Negative

customer_id	review_score	Sentiment	Image
ffed827348cd4b083956ce2b634b763b	1	Negative	🥺

Geolocation Sentiment Clusters
sentiment ● Negative ● Neutral ● Positive

Word map of negative sentiments

76.47K
Positive Sentiments

99.22K
Total Count of Customer Sentiments

8179
Neutral Sentiments

14.58K
Negative Sentiments

Customer in the southern regions of – Curitiba with negative sentiment, will words from the word map as a factor their expressed sentiment. Recurrent words like - cancel, delivery give insight to customer pain-points.

Positive sentiments at 77% of total sentiment count and neutral and negative sentiments having 8% and 15% respectively

Comparison of payment value and type across customer cities

Customer city by order status, payment value and type

customer_city ● curitiba ● santa cruz ● sao marcos

	curitiba	santa cruz	sao marcos
credit_card delivered	37.58%	39.48%	22.94%
boleto delivered	27.56%	44.78%	27.65%
debit_card delivered	100.00%		
voucher delivered	87.02%		12.98%

154.10
Average payment_value

156.97
Average payment_value - Curitiba

137.72
Average payment_value - Santa Cruz

99.48
Average payment_value - Sao Marcos

135.83
Average payment_value - Sao Paulo

Analysis of overall average payment type values for orders delivered in customer city, with all order IDs having a status of delivered.

Curibita had all 4 payment types - credit card @ 100% , boleto, debit card and voucher @ 87%.
It was the only city that used debit card as a payment system of all three cities.
It had payment value above the average base rate of 154.10

Sao Marcos had 3 payment types - credit card, boleto and voucher. Had lowest payment value below the average base rate.

Santa Cruz had 2 payment types- credit card and boleto. It had the highest number of payments for both payment types across the three cities. Had payment value below the average base rate.
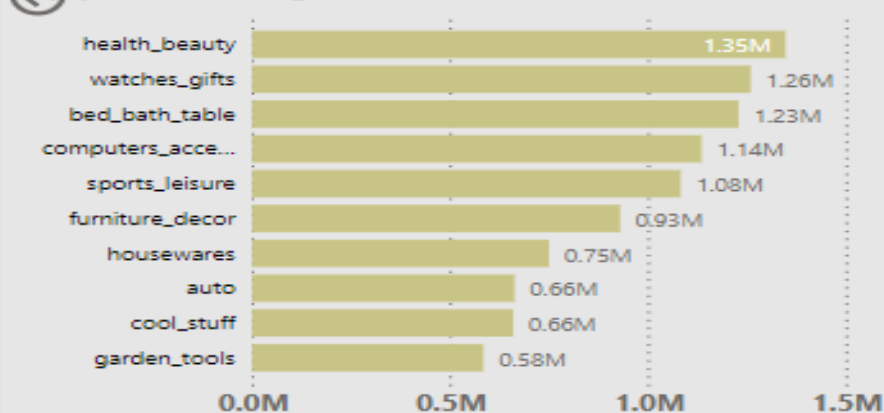
Sao Paulo in southeast region and home to major economic centres, had payment value below the average base rate but higher than the average for **Sao Marcos,** with a neutral customer sentiment.

This could indicate that not there could be other factors for neutral sentiments other than a monetary factor.
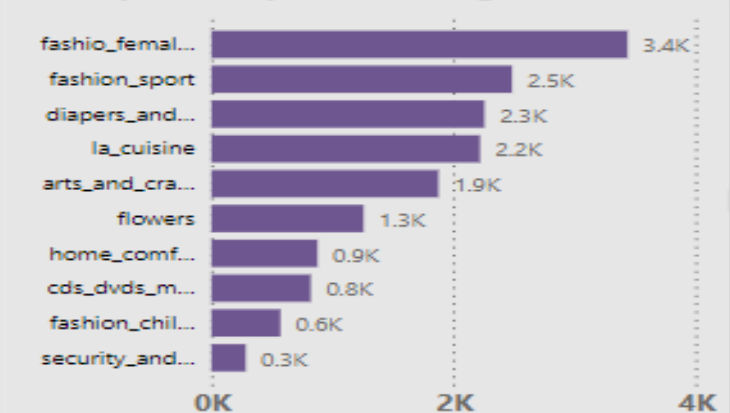
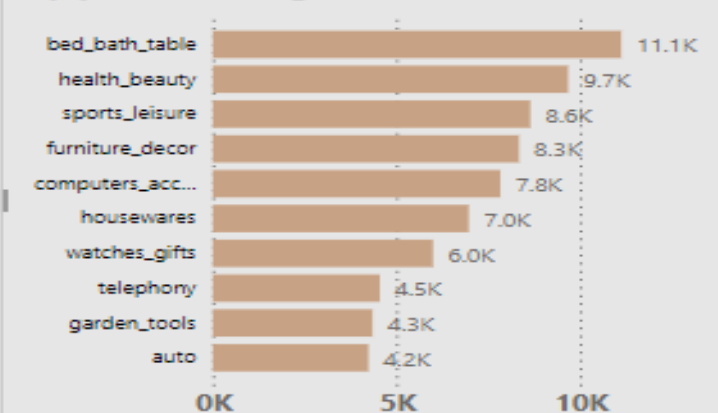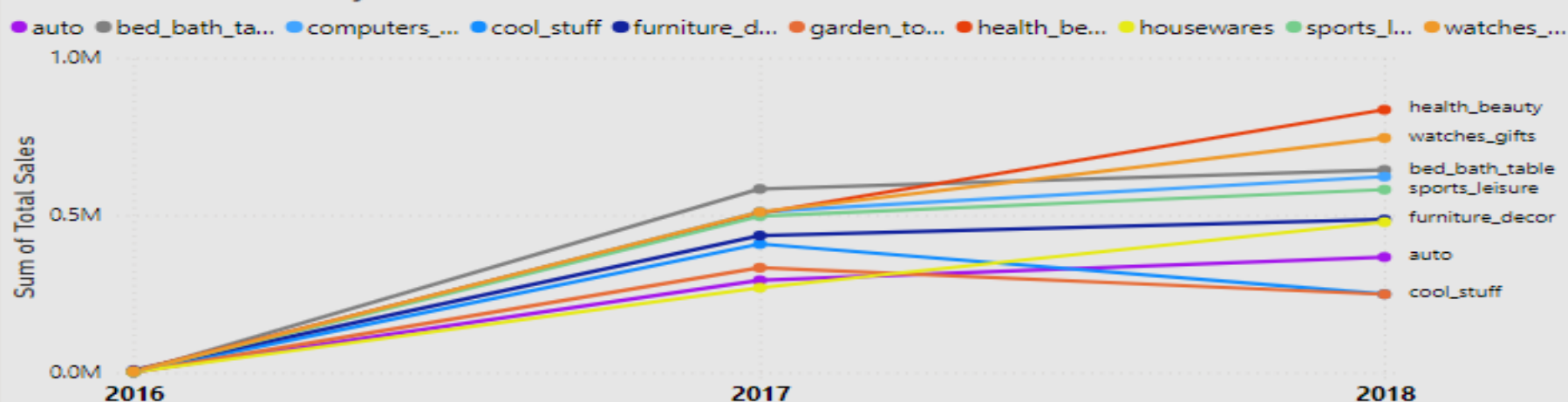# Analysis of product performance by state, category and year

The state **'SP'** where Sao Paulo is located was top performing state with sales of over 5 million R$, **'RS'** where Sao Marcos is located had sales of over 854,000 R$ followed by **'PR'** where Curitiba is located at over 787,000 R$

Top performing product category was bed bath table, least performing was security and services.

Health beauty was the best product category for year 2018

Sales Trends Over the Years and Forecasting Sales for 2019
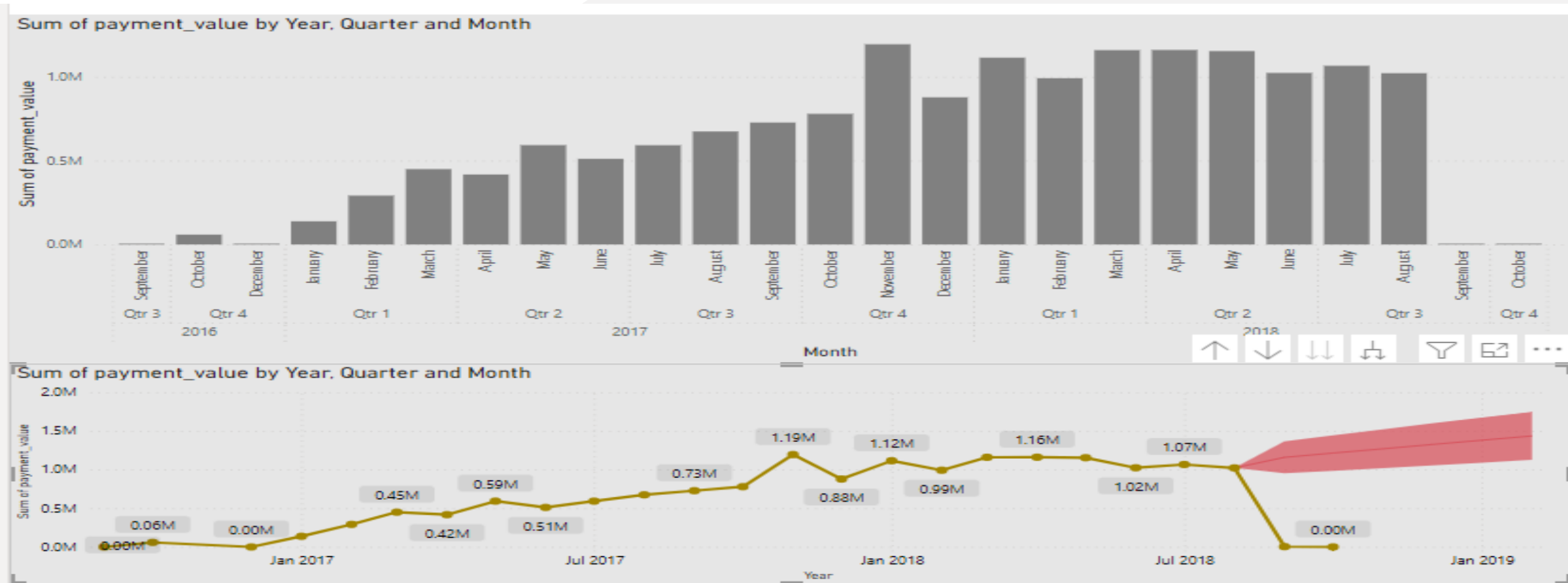
The month November had the highest volume in fourth quarter of year 2017 at over 1 million R$ in sales.

The last quarter of year 2017 outperformed the first three quarters of that year whereas for year 2018 the second quarter was better than first and third quarters but only by a slim margin.

Forecast shows an upward trend for the last quarter of year 2018 and year 2019.

Recommendations

This e-commerce business demonstrates promising performance with a positive sales trajectory, a high level of customer satisfaction, and a wide range of product offerings.

Minimize delivery costs and shorten the duration of shipments in remote/low concentrated areas through promotional offers (including free delivery option) that enable customers purchase from local vendors.

To encourage growth and expansion of business, it can upskill independent business owners in the low concentrated regions or where products underperform, to help improve sales.

Expand the criteria for collection of reviews including options to indicate reasons for review i.e., vendor service, product or speed of delivery to improve customer experience.

Further analysis on high-value and low-value customers, can show spread of sentiments of these customers across different regions if there is additional data to show frequency of orders and segmentation by volume of purchase. Important for targeted marketing.

Curitiba is the capital city of the state of Paraná (PR) in Brazil and is considered economically vibrant. Its high utilization of vouchers requires further analysis because high dependency on vouchers usually suggests an informal economy and one where there are potential barriers to economic growth.

Analyzing a bike enterprise

Oge Ibezi

Introduction

This analysis involved exploring key metrics on a global database of bike enterprise selling bikes, clothing, and accessories.

Data selection was done using sub-queries and joins where applicable

The database included different types of data, such as data warehouse and OLTP

The focus was on exploring the relationship between annual sales and store size, number of employees, revenue and determining the regions with the best sales,

Python and SQL were used to analyze the data, including deciphering specific data relevant to the questions.

Data validation and readability were ensured by cross-referencing the data in excel and converting it to a csv file for analysis in python.

Matplotlib was utilized in python to create graphs, such as scatter plots and bar charts, to visually present the findings, while anomalies in the data were identified and removed from the database.

Sales in best performing country

YTD Sales Per Country

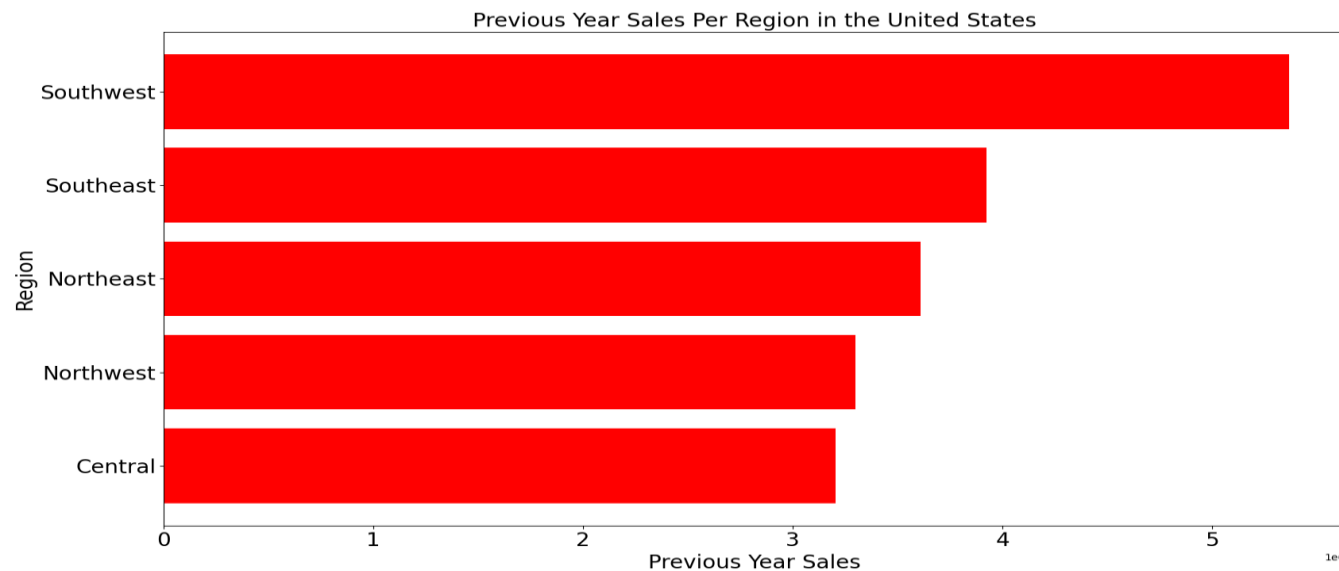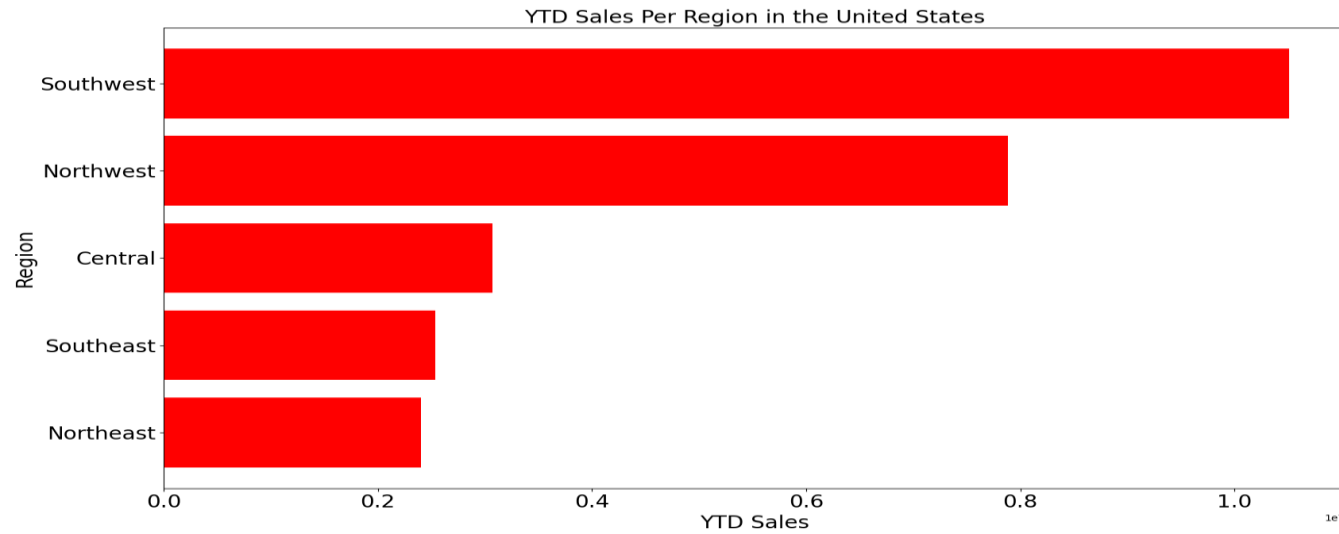Previous Year Sales Per Country
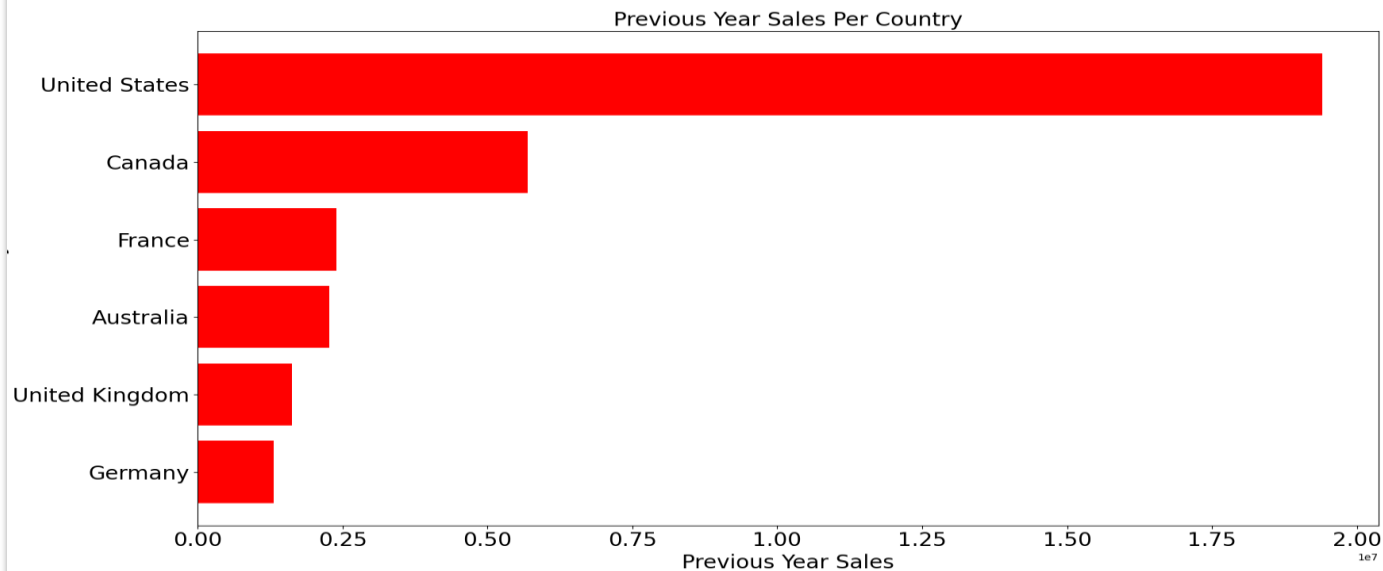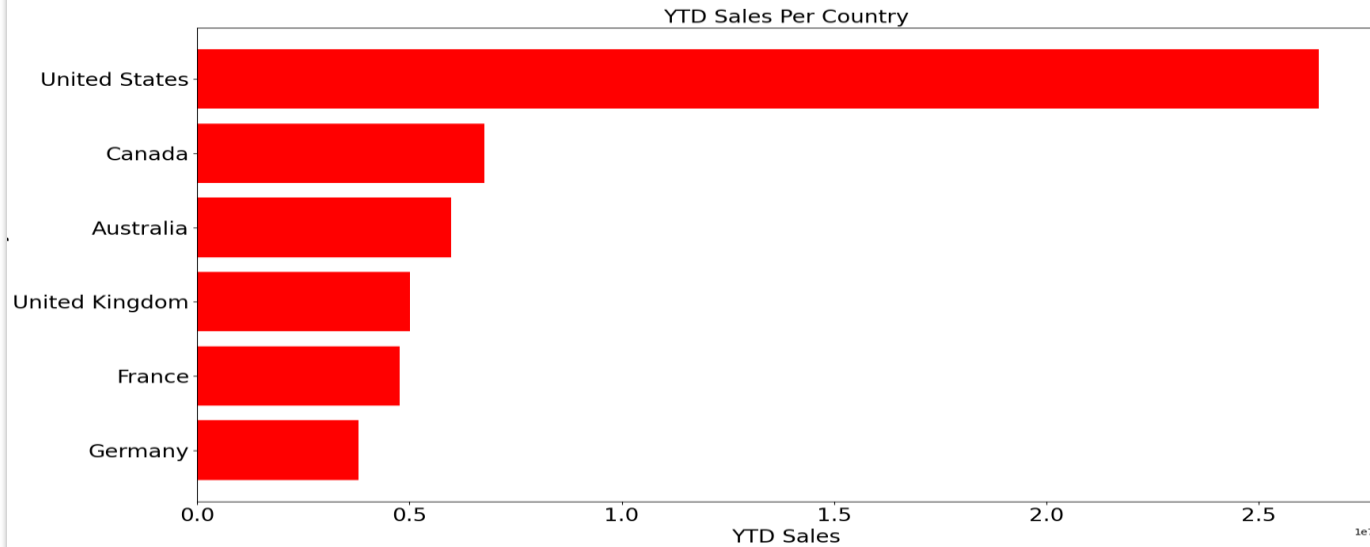
- Which country has the best performance?
  - United States is the best performing country

- Previous year's performance
  - United States is the best performing country, both on a YTD and Previous Year basis
  - France was in third place in previous year but in fifth place for current year

Regional sales in best performing country

YTD Sales Per Region in the United States

Previous Year Sales Per Region in the United States

- Revenue split by region in the best performing country

- YTD(year to date) regional sales
  - Southwest region is the best performing region, followed by Northwest

- Previous Year regional sales
  - Southwest region is the best performing region

Relationship between annual leave taken and bonus

Relationship between Annual Leave Taken and Bonus

r = -0.13

Relationship between Annual Leave Taken and Bonus in Sales Department

r = 0.46
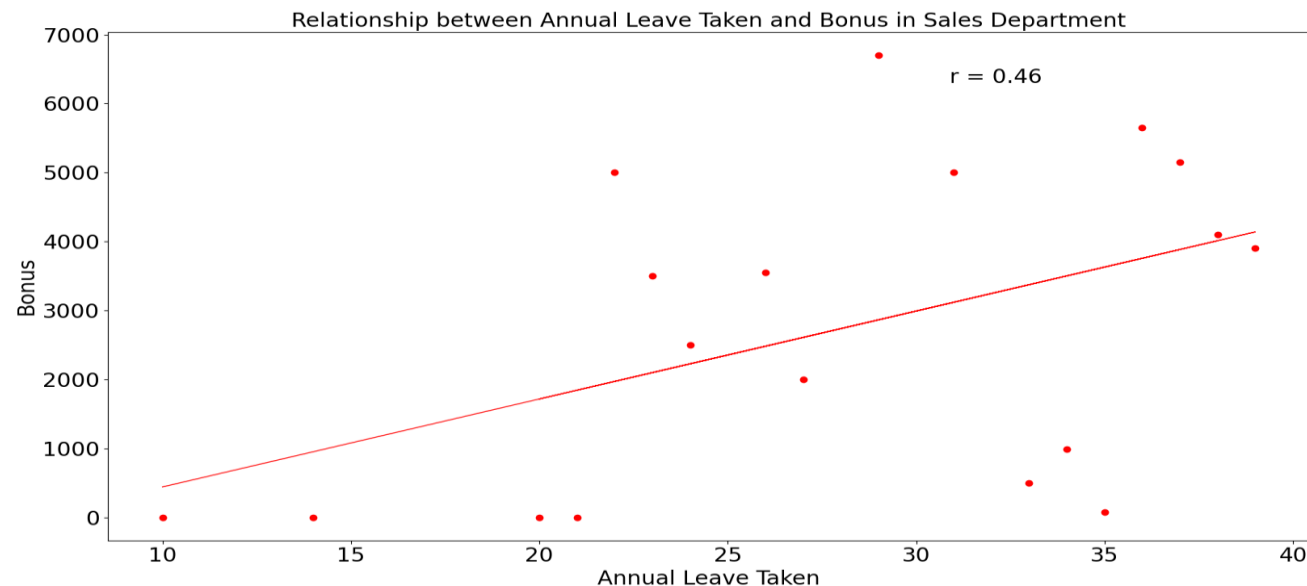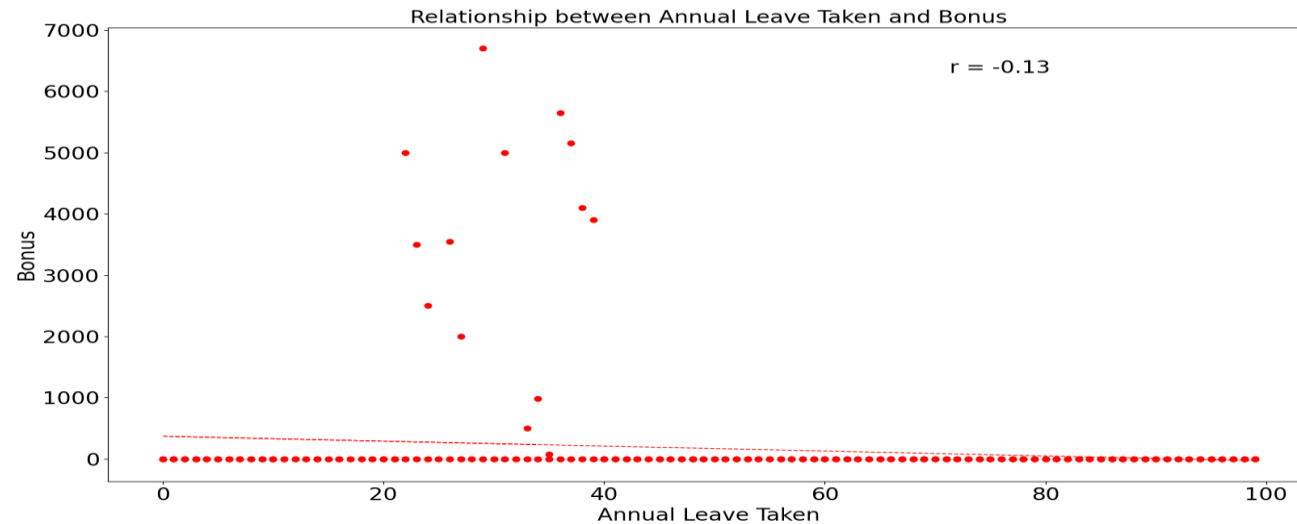
- ▶ The chart shows that a lot of employees do not receive a bonus, and in fact only the sales department employees received bonuses

- ▶ The correlation coefficient of -0.13 suggest a very weak or no relationship between annual leave and bonuses

- ▶ The correlation coefficient of 0.46 suggest a moderate positive relationship between annual leave taken and bonus

- ▶ However, moderate positive relationship does not mean there is a causal factor.

Relationship between Country store count and Revenue

Relationship between country store count and country YTD revenue

r = 0.99

Relationship between country store count and country PY revenue

r = 1.00
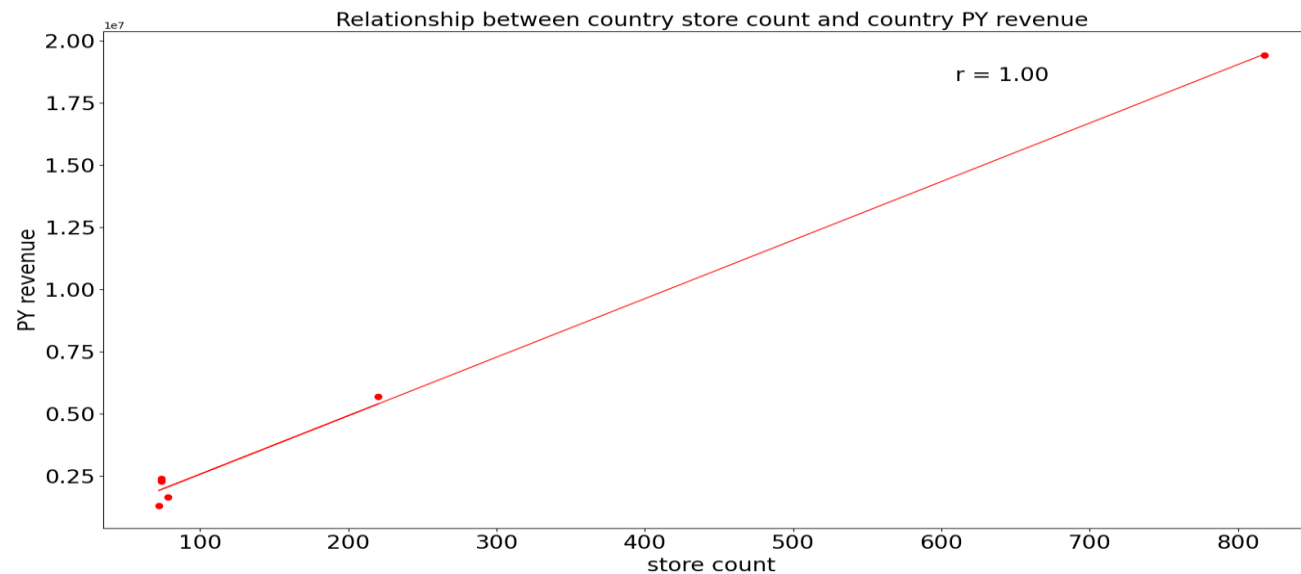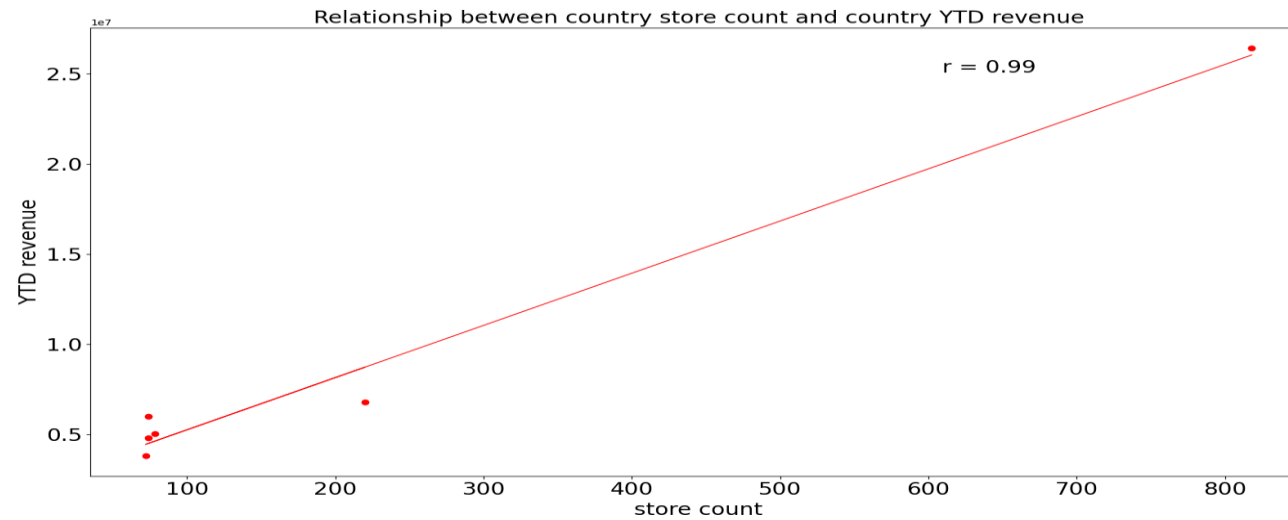
▶ Exploring the relationship between country store count and revenue shows:

    ▶ There is a very strong positive relationship between country store count and revenue

    ▶ Both correlation coefficient are very close to +1

Relationship between sick leave and job title

Average Sick Leave Hours by Person Type

Average Sick Leave Hours by Job Title
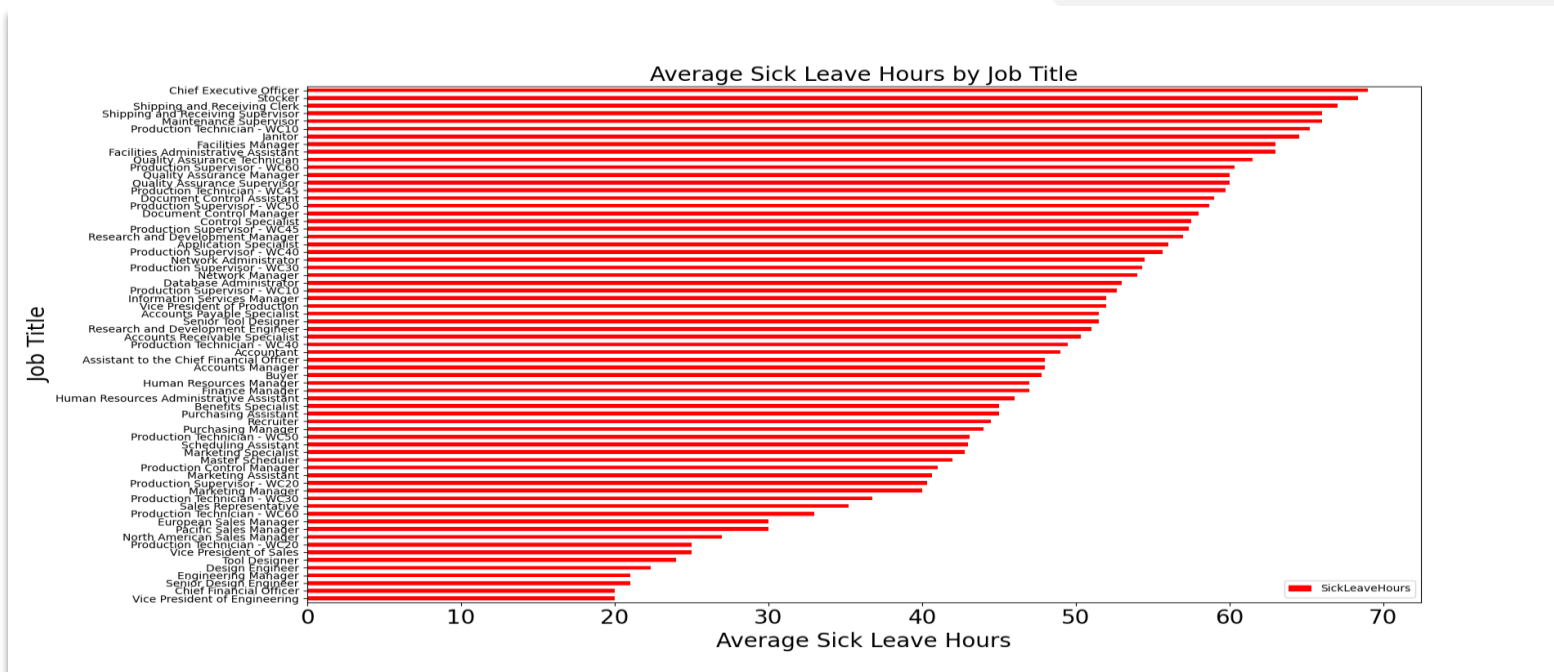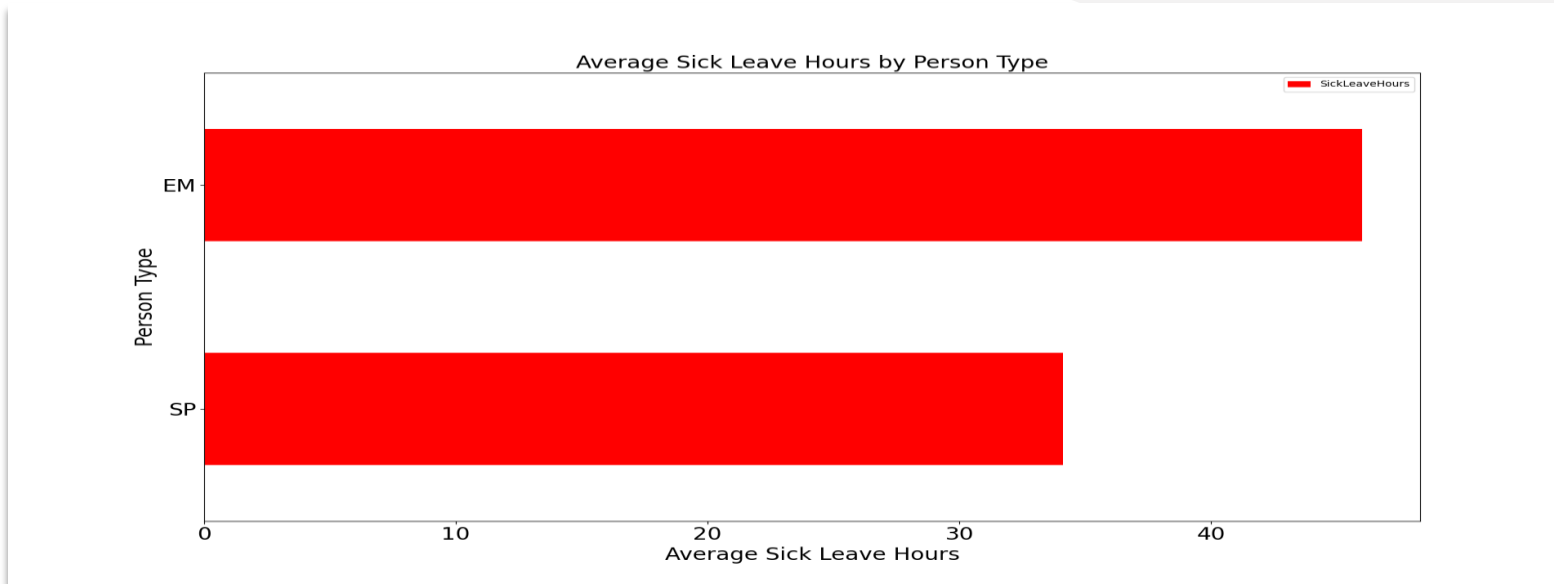
- Using average sick leave for comparison across different job type and title.

- Relationship between sick leave and Person Type

  - Non-sales employee had higher average sick leave than sales employee(SP – salesperson)

- Relationship between sick leave and Job Title

  - The top 3 job titles that had the most average sick leave were

    - 1) CEO

    - 2) Stocker

    - 3) Shipping and Receiving Clerk

Relationship between sick leave and department

Average Sick Leave Hours by Department Name

Relationship between store trading duration and revenue

Relationship between Store Trading Duration (in years) and Annual Revenue

r = -0.13

- ▶ Here we define store trading duration as the age of the store ("Year 2022 minus the year the store opened")

- ▶ The correlation coefficient of -0.13 suggest there is a very weak or no relationship between store trading duration and revenue

Relationship between size of store and revenue

Relationship between Store Size (SquareFeet) and Annual Revenue

r = 1.00
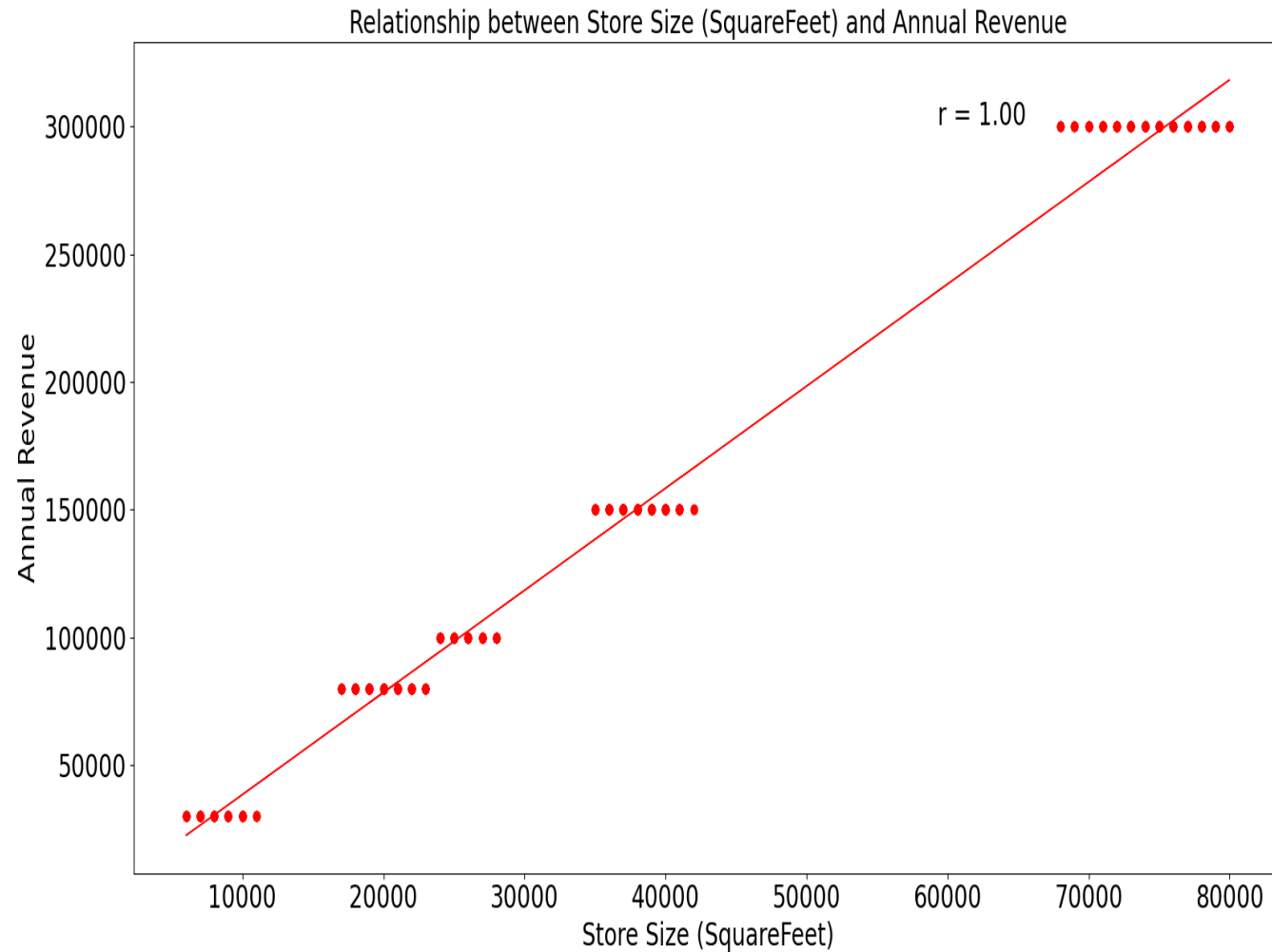
- <u>Relationship between the size of the stores and revenue</u>
    - The correlation coefficient of +1.00 suggest there is a very strong relationship between size of the stores and revenue

Relationship between size of store and number of employees

Relationship between Store Size (SquareFeet) and Number of Employees

r = 0.97
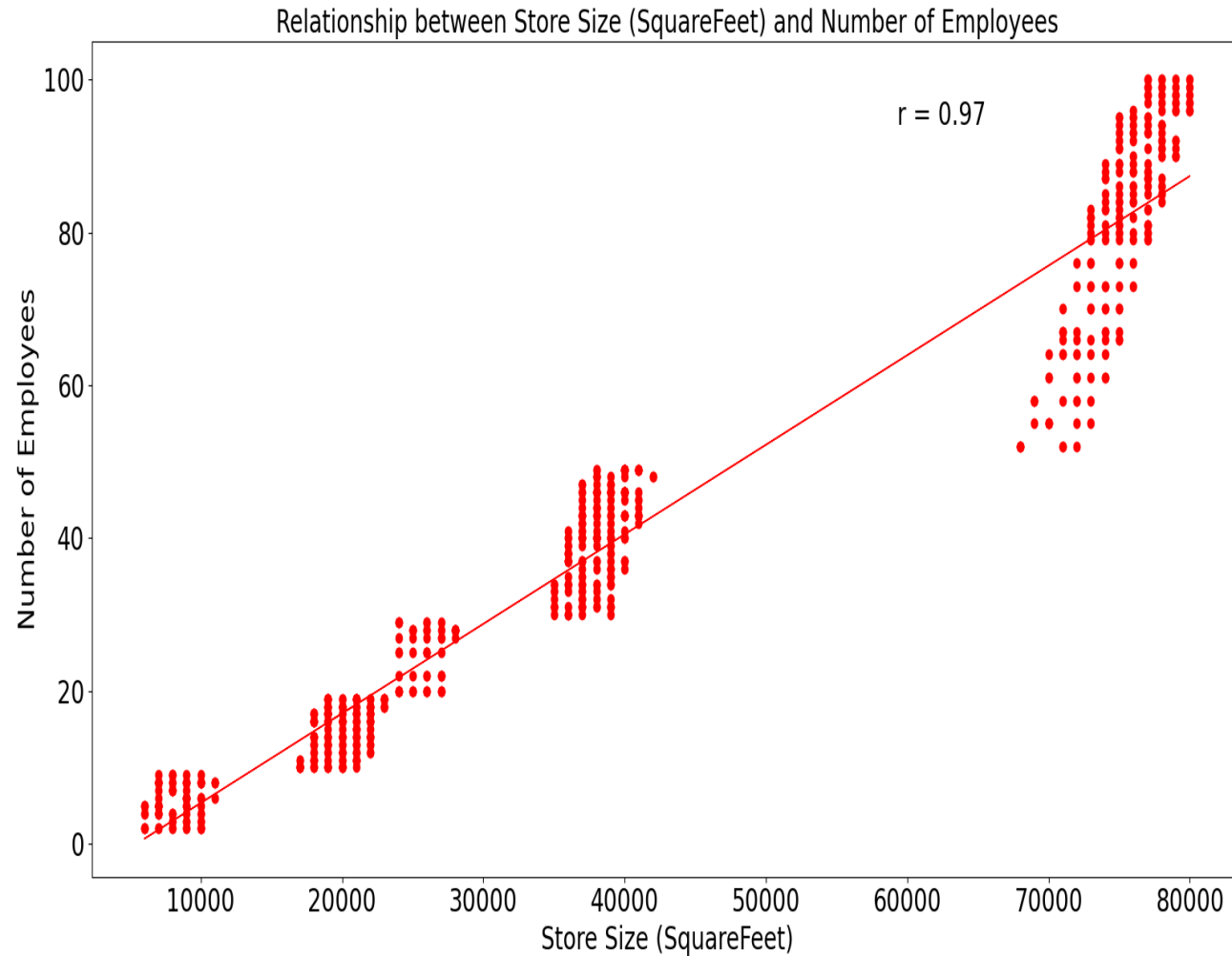
- Relationship between the size of stores and number of employees
  - The correlation coefficient of +0.97 suggest there is a very strong relationship between the size of the stores and number of employees

CORRELATION MATRIX
(SquareFeet, NumberofEmployees and AnnualRevenue)
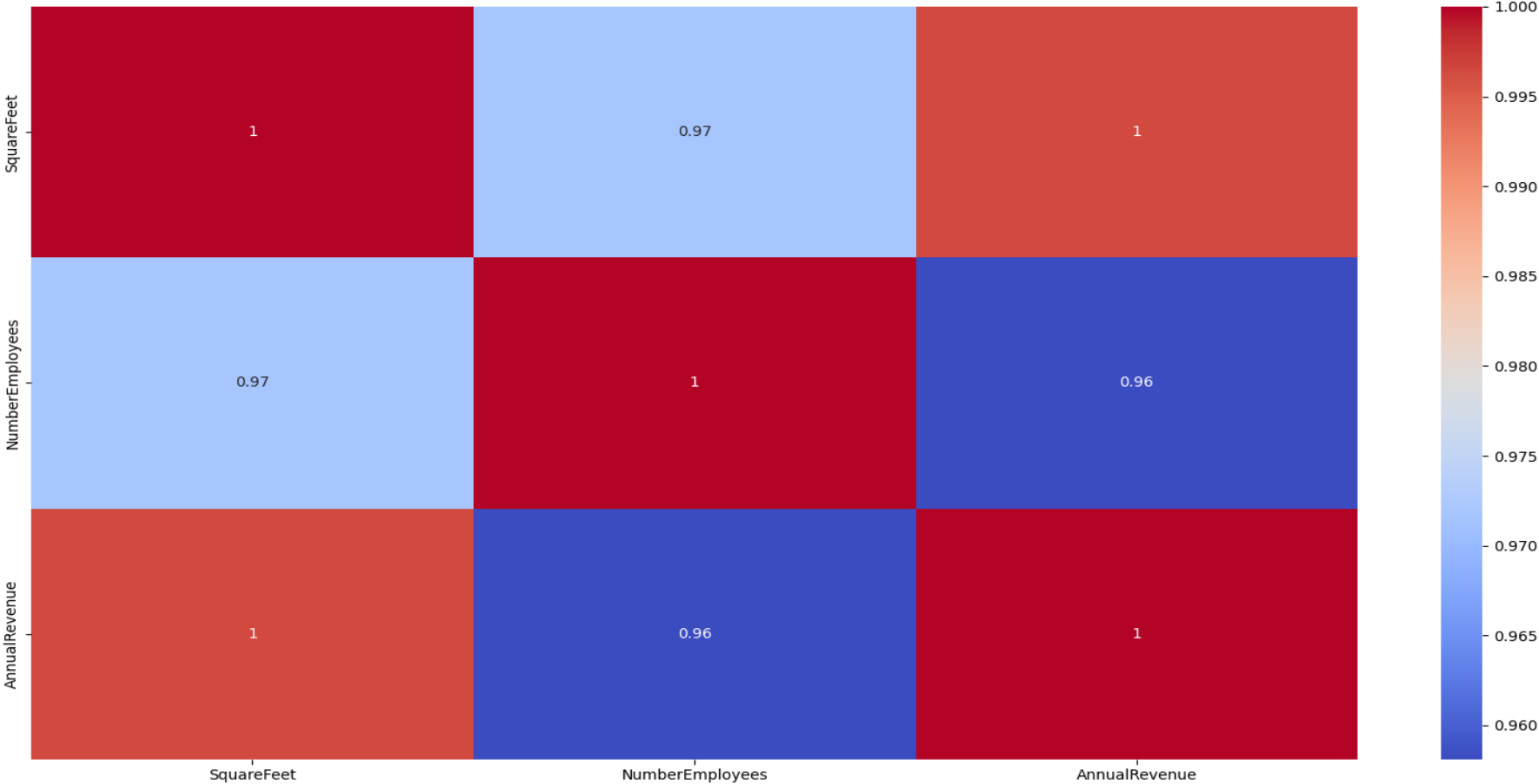
CONCLUSION

- The best performing country for sales revenue generated was United States
- South-west region of United States was the best performing region for sales
- The sales department staff were the only employees to receive bonuses in the organisation
- Analysis showed a moderate positive correlation between annual lave taken and bonuses
- There is a strong positive correlation between country store count and revenue generated
- The job titles with most average sick leave hours were: CEO, Stocker and Shipping and receiving clerk
- Top three departments with most average sick leave hours were: Shipping and receiving, Facilities and maintenance and Quality Assurance
- There is little or no correlation between when business store started and revenue generated
- Relationship between size of store and revenue showed a strong positive correlation, meaning that the bigger the store the higher the revenue generated
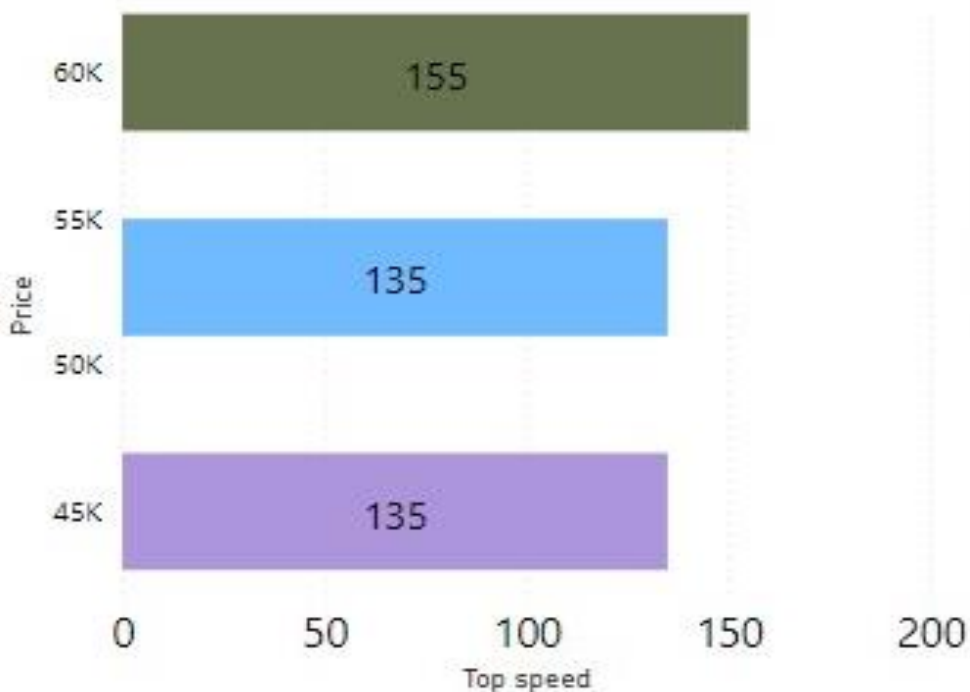
Competitor Analysis - Tesla Electric Vehicle (EV) Dashboard.pbix

Analysis of Non-Tesla EV Models, to determine top competitor for best-selling Tesla Y EV Model in terms of range, price and top speed

Comparison of EV Model by speed and price

EV MODEL ● Tesla Model Y ● Tesla Model Y Long Ran... ● Tesla Model Y Lon...

Price	Top speed
60K	155
55K	135
45K	135

Top speed: 0, 50, 100, 150, 200

EV Model - Max range and Top speed

Ford Mustang Mach-E ER RWD — 280

Volkswagen ... — 245

Polestar 2 Long Range Dual M... — 260

EV MODEL by Price

100%

Ford Musta...	59K
Polestar 2 L...	53K
Volkswage...	53K

89%
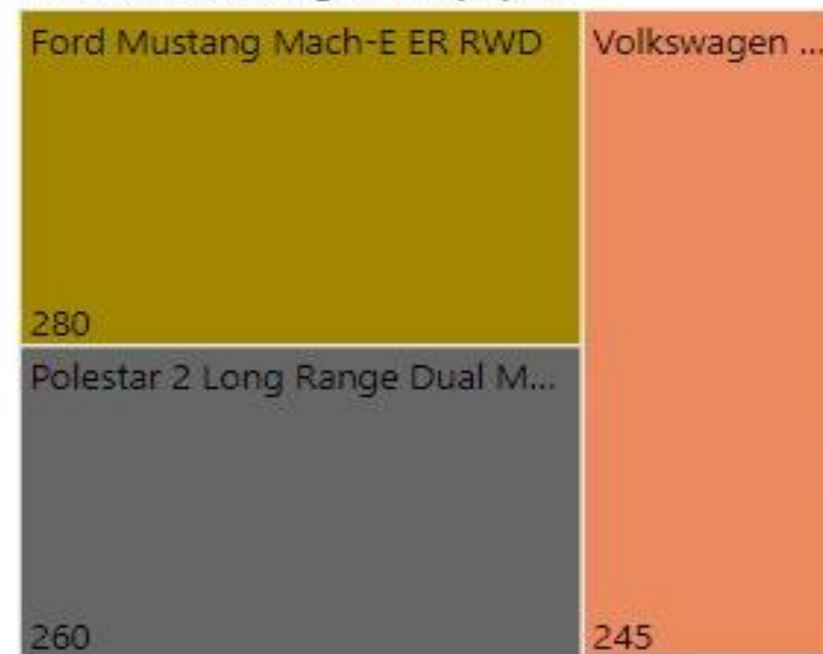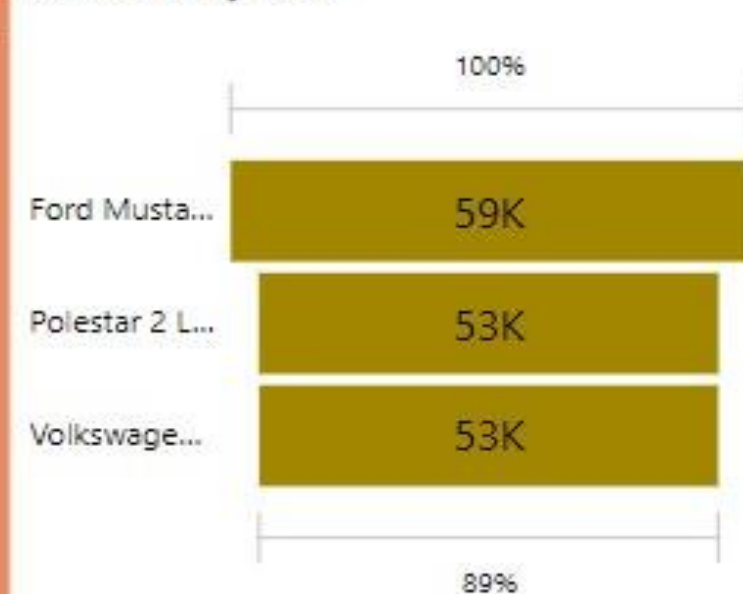
- Within the Tesla Y range – **Tesla model Y long range performance** was the best-selling brand having total price of £59,990

- Using 'Tesla model Y long range performance' as the basis for comparison, we find that **Polestar 2 Long Range Dual Motor** is the two top non-tesla EV cars with the best competitive advantage.

- Polestar 2 Long Range Dual Motor is the best EV car in terms of price and speed, while **Ford Mustang Mach - E ER RWD** gives better value for money in terms of max range (number of miles per charge)

EV MODEL by Max range

215 (28.86%)

270 (36.24%)

260 (34.9%)

EV MODEL
● Tesla Model Y Long Range...
● Tesla Model Y Long Range...
● Tesla Model Y

SQL CODE: Sales in best performing country

```sql
Best performing country - United States

■   Specify columns to be retrieved using SELECT statement
SELECT CountryName,
                SUM(SalesYTD) AS SalesYTD    -- SUM(SalesLastYear) AS SalesLastYear

■   Use subquery to combine data from 'SalesTerritory' and 'CountryRegion' tables using LEFT JOIN and matching/primary key
    'CountryRegionCode'
FROM(
SELECT st.[TerritoryID]
     ,st.[Name] AS RegionName
     ,st.[CountryRegionCode]
     ,st.[Group]
     ,st.[SalesYTD]
     ,st.[SalesLastYear]
         ,cr.[Name] AS CountryName
FROM [Sales].[SalesTerritory] AS st
LEFT JOIN [Person].[CountryRegion] AS cr
        ON st.[CountryRegionCode] = cr.[CountryRegionCode]) AS mergedTable

■   Group data with same country name

GROUP BY CountryName

■   Organise the results using SalesYTD column in descending order

ORDER by SalesYTD DESC
```

SQL CODE: Regional sales in best performing country

■   Specify columns to be retrieved using SELECT statement

```sql
SELECT st.[TerritoryID]
      ,st.[Name] as RegionName
      ,st.[CountryRegionCode]
      ,st.[Group]
      ,st.[SalesYTD]
      ,st.[SalesLastYear]
      ,cr.[Name] as CountryName
```

■   Combine data from 'SalesTerritory' and 'CountryRegion' tables using LEFT JOIN and matching/primary key 'CountryRegionCode'

```sql
FROM [Sales].[SalesTerritory] AS st
LEFT JOIN [Person].[CountryRegion] AS cr
ON st.[CountryRegionCode] = cr.[CountryRegionCode]
```

■   Filter to show results for rows with country name 'United States'

```sql
WHERE cr.[Name] = 'United States'
```

SQL CODE: Relationship between annual leave taken and bonus

- Specify columns to be retrieved using SELECT statement

```sql
SELECT
    [BusinessEntityID]
    ,[JobTitle]
    ,[VacationHours]
    ,[Bonus]
    ,[Department_Name]
```

- Use subquery to combine data from 'Employee', 'SalesPerson', 'EmployeeDepartmentHistory' and 'Department' tables using LEFT JOIN and matching/primary key 'BusinessEntityID'.

```sql
FROM(
SELECT emp.[BusinessEntityID]
        ,emp.[JobTitle]
        ,emp.[VacationHours]
        ,sp.[Bonus]
        ,edh.[DepartmentID]
        ,edh.[EndDate] AS Department_End_Date
        ,dp.[Name] AS Department_Name
        ,dp.[GroupName] AS Department_GroupName

FROM [HumanResources].[Employee] AS emp
LEFT JOIN [Sales].[SalesPerson] AS sp
        ON emp.[BusinessEntityID] = sp.[BusinessEntityID]
LEFT JOIN [HumanResources].[EmployeeDepartmentHistory]  AS edh
        ON emp.[BusinessEntityID] = edh.[BusinessEntityID]
LEFT JOIN [HumanResources].[Department]  AS dp
        ON edh.[DepartmentID] = dp.[DepartmentID]
```

- Filter to show records only current department ('Sales')

```sql
WHERE sp.[Bonus] IS NOT NULL)  AS adj_table
ORDER BY Bonus DESC
```

PYTHON CODE: Sales in best performing country

```python
# Step 1: Create a table to show which country has the highest sales
table_1a_YTD = pd.pivot_table(Q1_data, values='SalesYTD', index=['CountryName'], columns=[], aggfunc=np.sum)
table_1a_YTD = table_1a_YTD.sort_values(by='SalesYTD', ascending=True)

table_1a_PY = pd.pivot_table(Q1_data, values='SalesLastYear', index=['CountryName'], columns=[], aggfunc=np.sum)
table_1a_PY = table_1a_PY.sort_values(by='SalesLastYear', ascending=True)

# Step 2: Create a horizontal bar chart to show which country has the highest sales
# YTD Sales bar chart
plt.barh(table_1a_YTD.index, table_1a_YTD['SalesYTD'], data=table_1a_YTD, color="red") # plot the bar chart
plt.title("YTD Sales Per Country", fontsize=20) # add a title
plt.xlabel("YTD Sales", fontsize=20) # add a label to the x-axis
plt.ylabel("Country", fontsize=20) # add a label to the y-axis
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot

# PY Sales bar chart
plt.barh(table_1a_PY.index, table_1a_PY['SalesLastYear'], data=table_1a_PY, color="red") # plot the bar chart
plt.title("Previous Year Sales Per Country", fontsize=20) # add a title
plt.xlabel("Previous Year Sales", fontsize=20) # add a label to the x-axis
plt.ylabel("Country", fontsize=20) # add a label to the y-axis
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot
```

PYTHON CODE: Regional sales in best performing country

```python
# Step 3: Create a table to show the regional sales in the best performing country (i.e. United States)
US_data = Q1_data[Q1_data['CountryName'] == 'United States']
table_1b_YTD = pd.pivot_table(US_data, values='SalesYTD', index=['RegionName'], columns=[], aggfunc=np.sum)
table_1b_YTD = table_1b_YTD.sort_values(by='SalesYTD', ascending=True)

table_1b_PY = pd.pivot_table(US_data, values='SalesLastYear', index=['RegionName'], columns=[], aggfunc=np.sum)
table_1b_PY = table_1b_PY.sort_values(by='SalesLastYear', ascending=True)

# Step 4: Create a horizontal bar chart to show the regional sales in the best performing country (i.e. United States)
# YTD Sales bar chart
plt.barh(table_1b_YTD.index, table_1b_YTD['SalesYTD'], data=table_1b_YTD, color="red") # plot the bar chart
plt.title("YTD Sales Per Region in the United States", fontsize=20) # add a title
plt.xlabel("YTD Sales", fontsize=20) # add a label to the x-axis
plt.ylabel("Region", fontsize=20) # add a label to the y-axis
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot


# PY Sales bar chart
plt.barh(table_1b_PY.index, table_1b_PY['SalesLastYear'], data=table_1b_PY, color="red") # plot the bar chart
plt.title("Previous Year Sales Per Region in the United States", fontsize=20) # add a title
plt.xlabel("Previous Year Sales", fontsize=20) # add a label to the x-axis
plt.ylabel("Region", fontsize=20) # add a label to the y-axis
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot
```

PYTHON CODE: Relationship between annual leave taken and bonus

```python
# Step 1: Create a scatter chart to show annual leave taken and bonus
plt.scatter(Q2_data['VacationHours'], Q2_data['Bonus_adj'], color="red") # plot the scatter chart
plt.title("Relationship between Annual Leave Taken and Bonus", fontsize=20) # add a title
plt.xlabel("Annual Leave Taken", fontsize=20) # add a label to the x-axis
plt.ylabel("Bonus", fontsize=20) # add a label to the y-axis
# adding  the correlation coefficient
r, p = stats.pearsonr(Q2_data['VacationHours'], Q2_data['Bonus_adj'])
plt.annotate('r = {:.2f}'.format(r), xy=(0.7, 0.9), xycoords='axes fraction', fontsize=20)
# adding best fit line
m, b = np.polyfit(Q2_data['VacationHours'], Q2_data['Bonus_adj'], 1)
plt.plot(Q2_data['VacationHours'], m*Q2_data['VacationHours'] + b, linestyle='--', linewidth=1, color="red")
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot


# Step 2: Create a table to show the relationship between department and bonus
table_2 = pd.pivot_table(Q2_data, values='Bonus_adj', index=['Department_Name'], columns=[], aggfunc=np.sum)
table_2 = table_2.sort_values(by='Bonus_adj', ascending=True)
print("-----------------------------------")
print("Create a table to show the relationship between department and bonus:")
print(table_2)
print("From here we can see that only Sales department receive bonus, bt not other departments")
```

PYTHON CODE: Relationship between annual leave taken and bonus

```python
# Step 3: Create a scatter chart to show annual leave taken and bonus in Sales department
Sales_data = Q2_data[Q2_data['Department_Name'] == 'Sales']
plt.scatter(Sales_data['VacationHours'], Sales_data['Bonus_adj'], color="red") # plot the scatter chart
plt.title("Relationship between Annual Leave Taken and Bonus in Sales Department", fontsize=20) # add a title
plt.xlabel("Annual Leave Taken", fontsize=20) # add a label to the x-axis
plt.ylabel("Bonus", fontsize=20) # add a label to the y-axis
# adding  the correlation coefficient
r, p = stats.pearsonr(Sales_data['VacationHours'], Sales_data['Bonus_adj'])
plt.annotate('r = {:.2f}'.format(r), xy=(0.7, 0.9), xycoords='axes fraction', fontsize=20)
# adding best fit line
m, b = np.polyfit(Sales_data['VacationHours'], Sales_data['Bonus_adj'], 1)
plt.plot(Sales_data['VacationHours'], m*Sales_data['VacationHours'] + b, linestyle='solid', linewidth=1, color="red")
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
plt.show() # show the plot
```

SQL CODE: Analysis of Electric Vehicle (EV) Stock by Country

Task : List the top 5 countries with the highest number of EV cars currently.

```sql
-- Specify columns to be retrieved using SELECT statement
SELECT TOP(5) [region]
        ,SUM(value) AS Number_of_Evs

FROM AdventureWorks2017.[dbo].[IEA-EV-dataEV stockCarsHistorical2]

-- Filter to show records for regions that are not 'World', 'Europe' or 'EU27'
WHERE [region] <> 'World' AND [region] <> 'Europe' AND [region] <> 'EU27'

-- Group data with same country name
GROUP BY [region]

-- Organise the results using Number_of_EVs columns in descending order
ORDER BY Number_of_EVs DESC;
```

Task :   Analysis to determine the ranking of 'Mkt_Cap' within top 25, '24h volume' within top 50 and '7 day%' increase highest and positive

```sql
    ■  Specify columns to be retrieved using SELECT statement
SELECT *
FROM (
  SELECT Coin, Mkt_Cap,[_1h],[_24h],[_7d],
    ■  Assign a rank to each row based on market capitalization value (Mkt_Cap)
    RANK() OVER (ORDER BY Mkt_Cap DESC) AS Mkt_Cap_rank,
    ■  Assign a rank to each row based on percentage change in the last 24hrs
    RANK() OVER (ORDER BY _24h DESC) AS Vol_rank

  FROM [dbo].[Crypto_SQL_data_csv]


    ■  Filter to show records to meet specified condition within the subquery
  WHERE Mkt_Cap >= 2754035669
    ) AS subquery
■  Organise the results using Mkt_Cap column in descending order
 ORDER BY Mkt_Cap DESC;
```

STATEMENT OF ACCOMPLISHMENT

#29,120,932

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Introduction to Data Science in Python

LENGTH

4 HOURS

COMPLETED ON

MAY 23, 2023

datacamp

Jonathan Cornelissen, CEO

STATEMENT OF ACCOMPLISHMENT

#29,071,617

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Database Design

LENGTH

4 HOURS

COMPLETED ON

MAY 18, 2023

datacamp

Jonathan Cornelissen, CEO

STATEMENT OF ACCOMPLISHMENT

#29,447,819

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Understanding Data Engineering

LENGTH

2 HOURS

COMPLETED ON

JUN 12, 2023

datacamp

Jonathan Cornelissen, CEO

STATEMENT OF ACCOMPLISHMENT

#29,059,449

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Understanding Data Visualization

LENGTH

2 HOURS

COMPLETED ON

MAY 27, 2023

datacamp

Jonathan Cornelissen, CEO

STATEMENT OF ACCOMPLISHMENT

#28,930,433

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Intermediate SQL Queries

LENGTH

4 HOURS

COMPLETED ON

MAY 02, 2023

datacamp

Jonathan Cornelissen, CEO