



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.

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 strategy 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



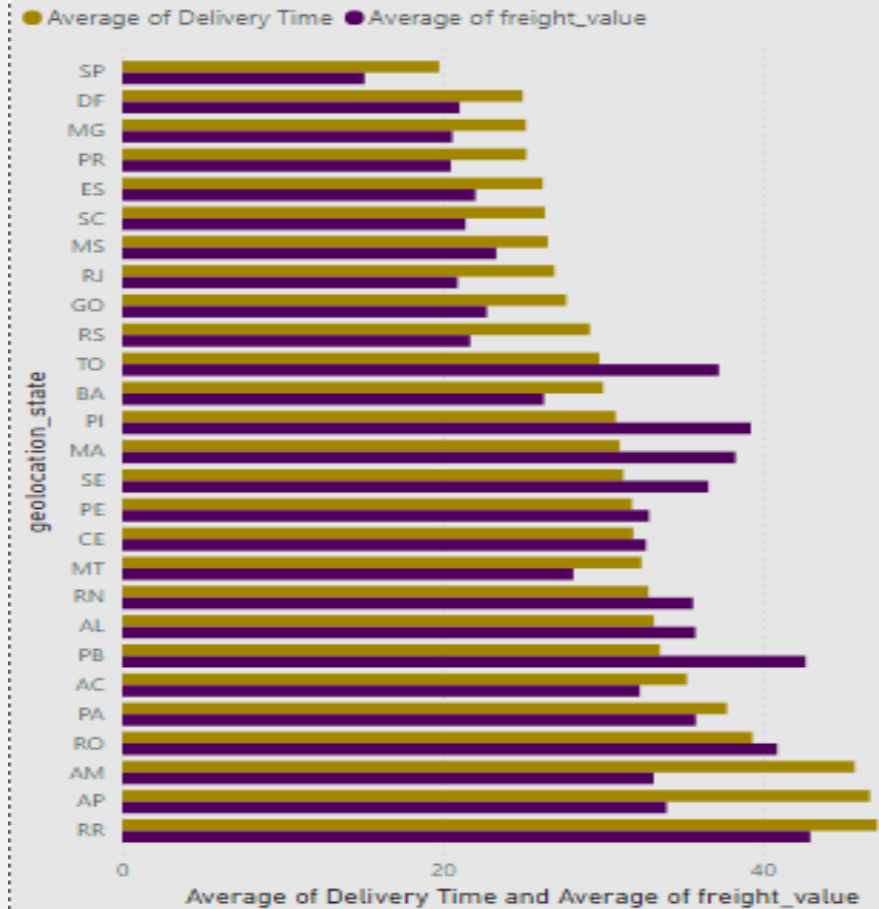
The Southeast region of Brazil, which includes states such as São Paulo, Rio de Janeiro, Minas Gerais, and Espírito Santo, is generally considered the most economically vibrant region in Brazil

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.

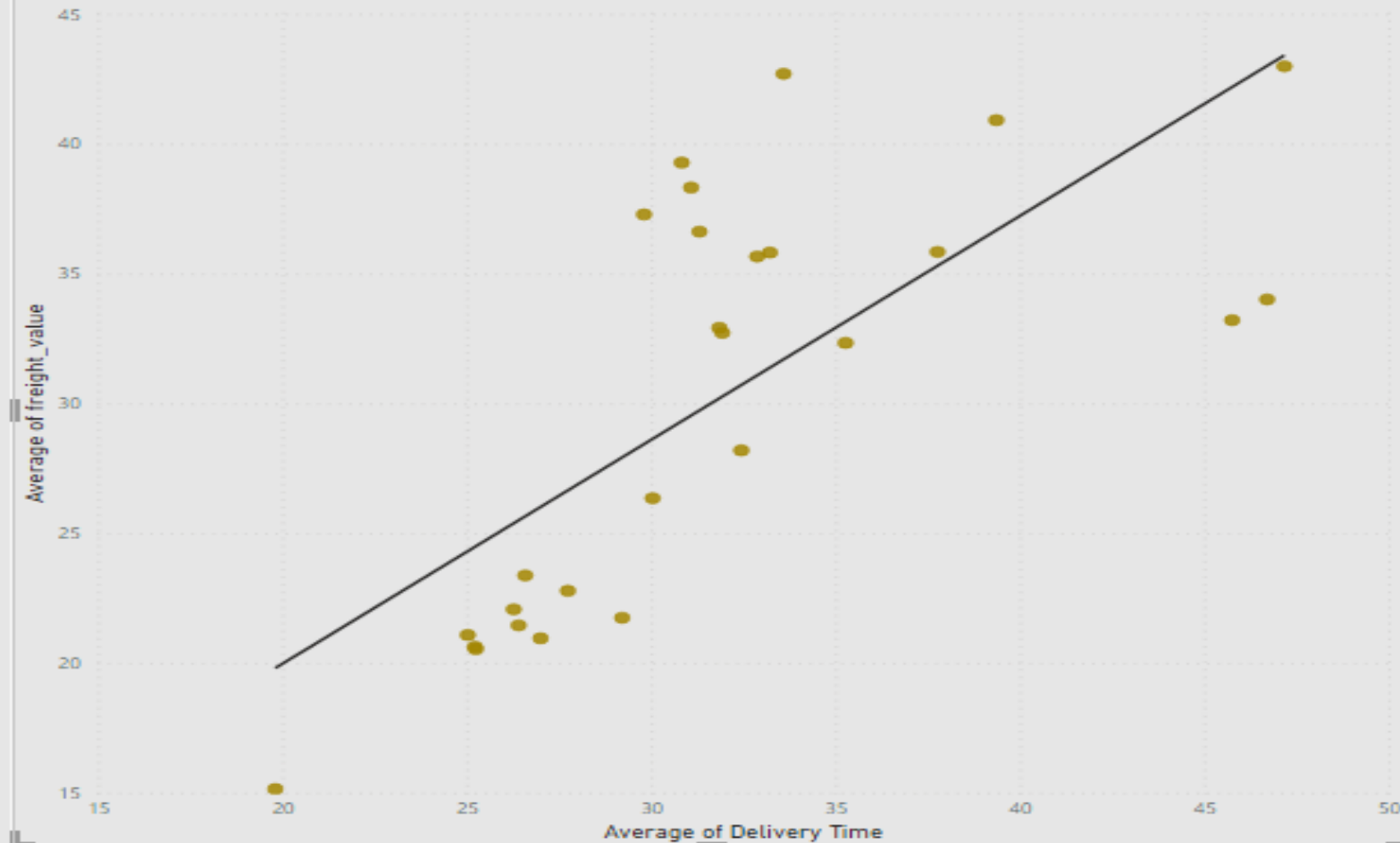
The bigger the size of the bubbles in the average freight value by geolocation chart, the higher delivery costs for sellers to deliver in those outer regions, whereas the bubbles near Sao Paulo are smaller. This be the reason why there are less orders from regions further out.

Analysis of delivery time by freight value and spatial distribution across states

Average of Delivery Time and Average of freight_value by geolocation_state






Average of Delivery Time and Average of freight_value by geolocation_state

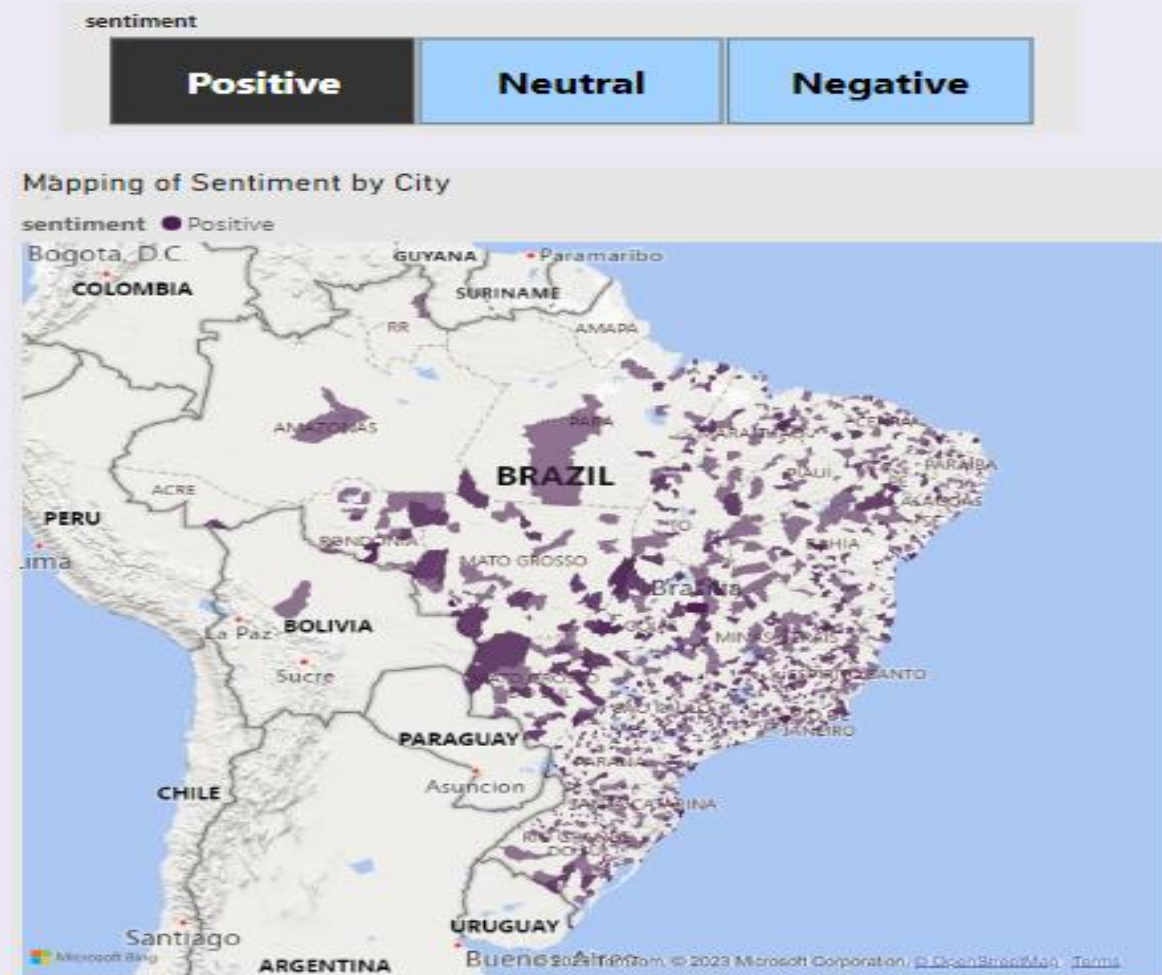


There is a correlation between the time of delivery and cost of shipping products. The higher the delivery time, the higher the shipping cost for sellers/vendors. Also, the further or more remote the region with lower economic activity, the higher the shipping cost.

Analysis of sentiment spread across regions

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

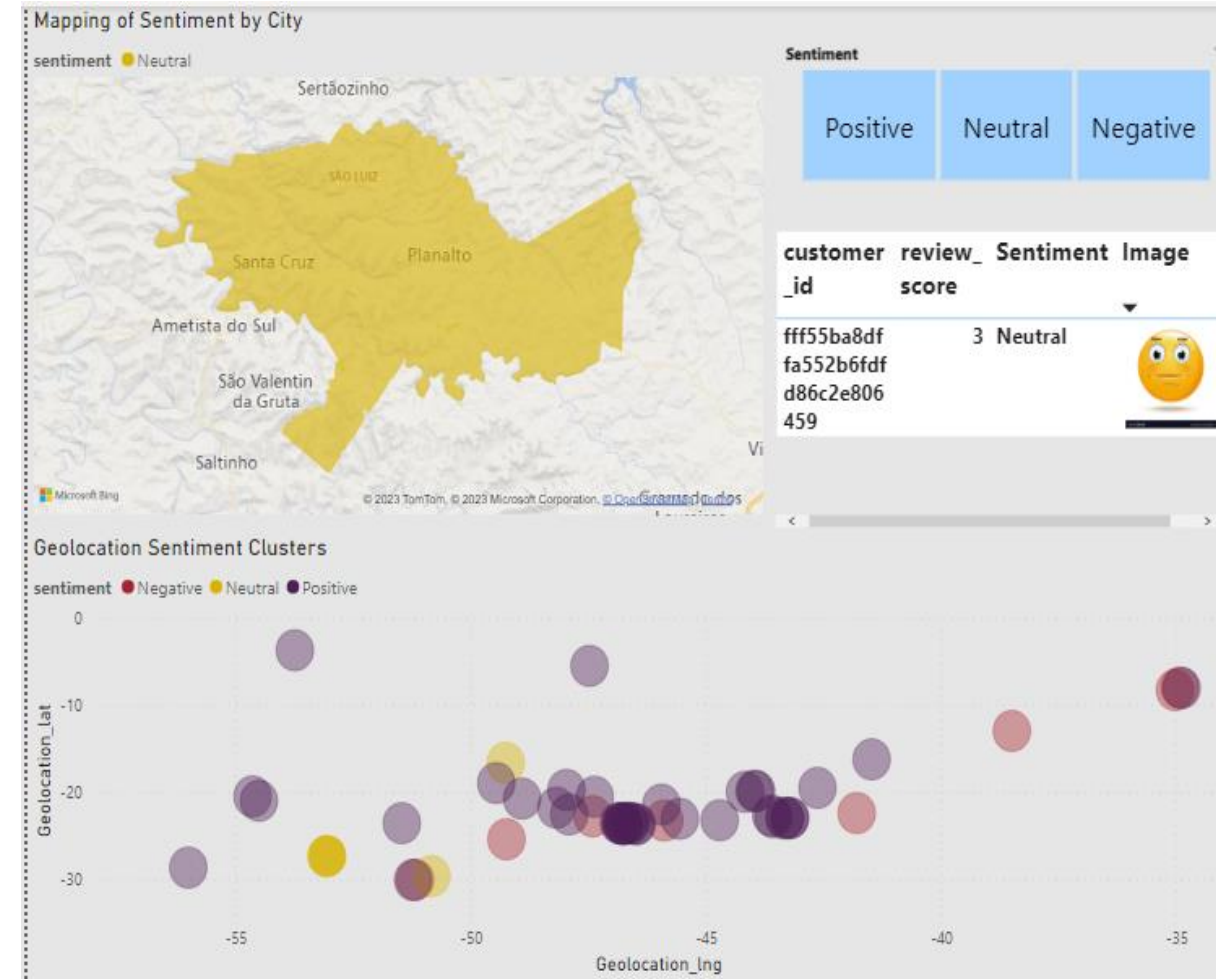
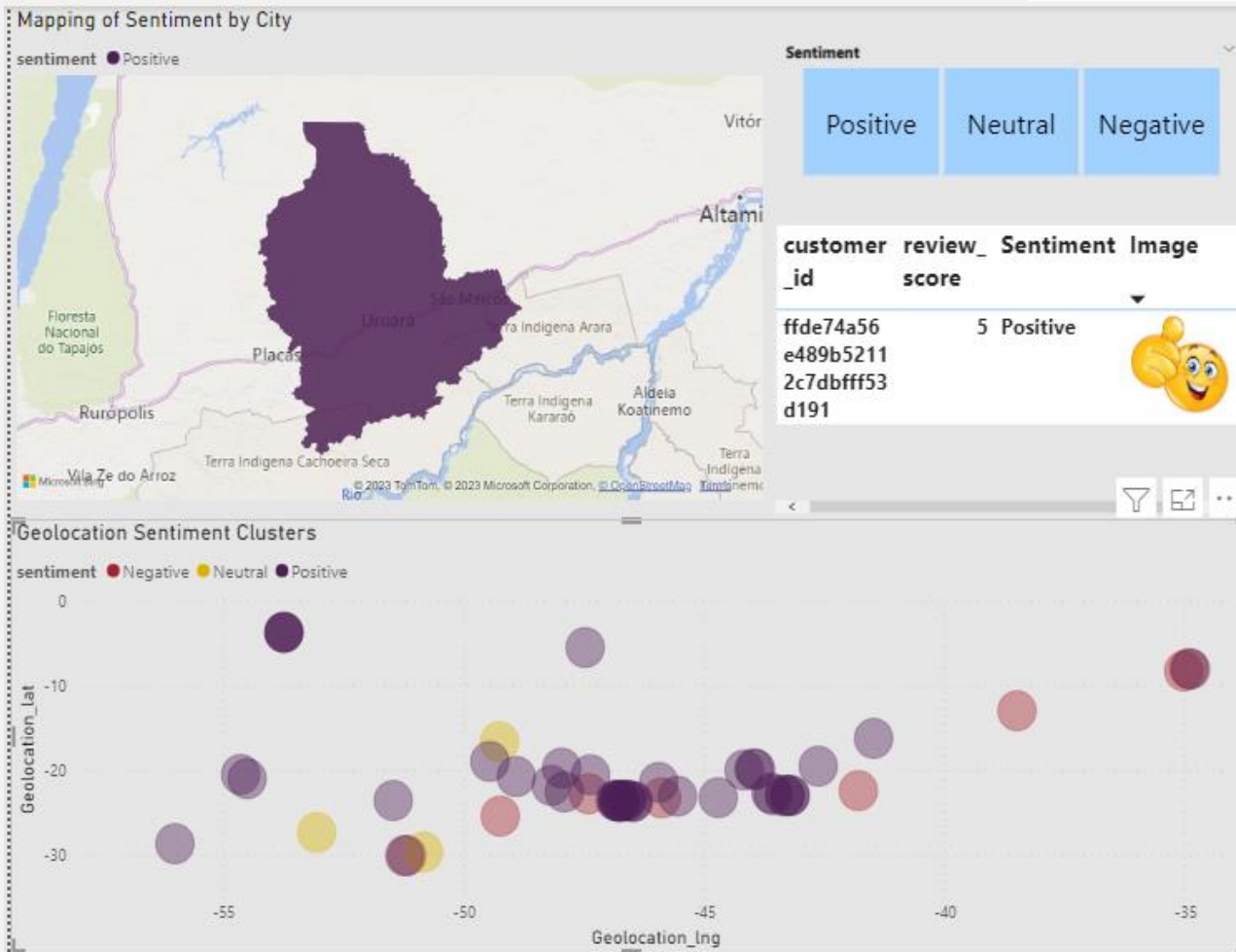
Word map of sentiments



Sentiment 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.

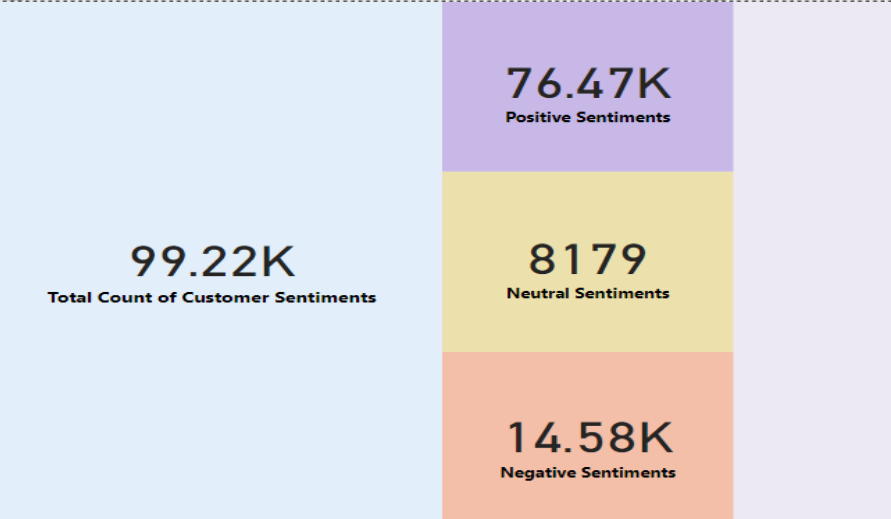
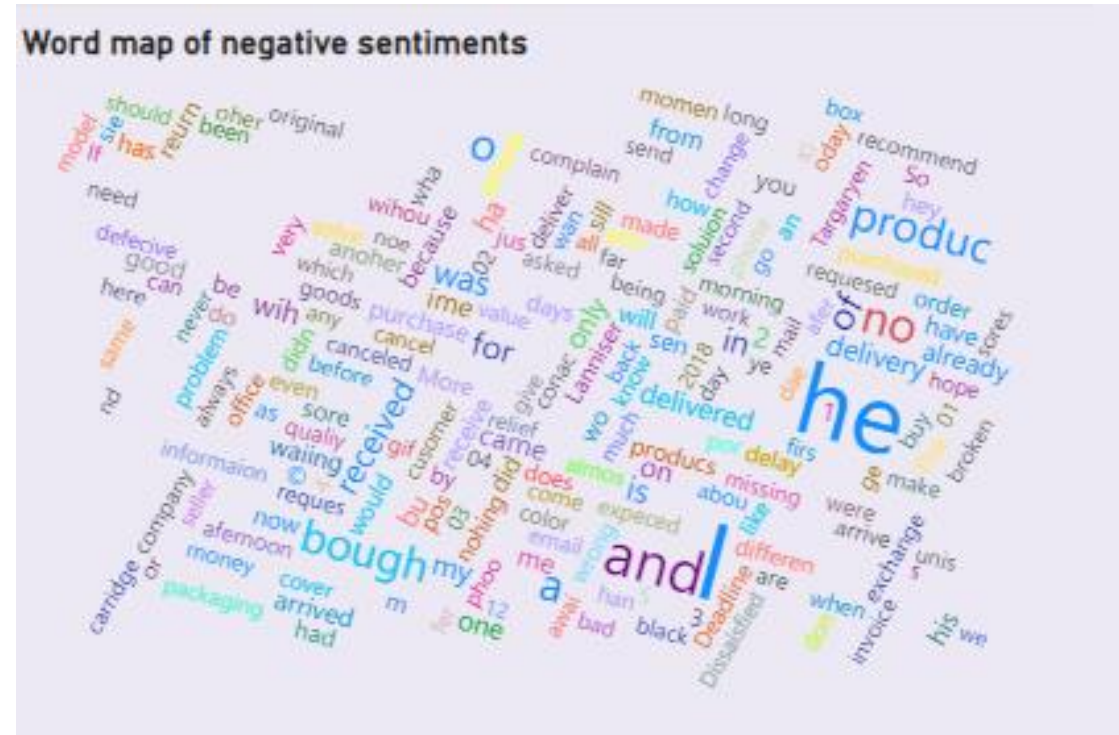
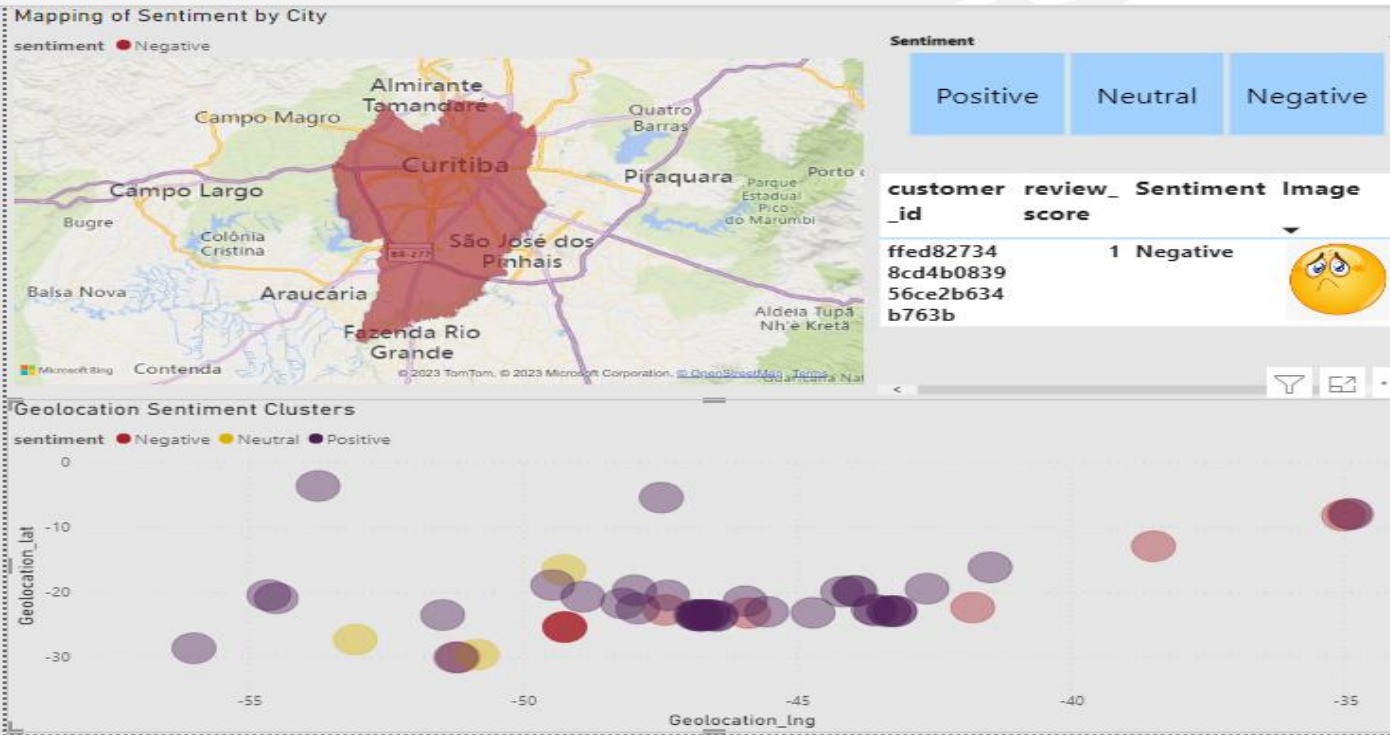
Clustering and sentiment patterns across regions



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



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

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%

0%

20%

40%

60%

80%

100%

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.

Curitiba 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.

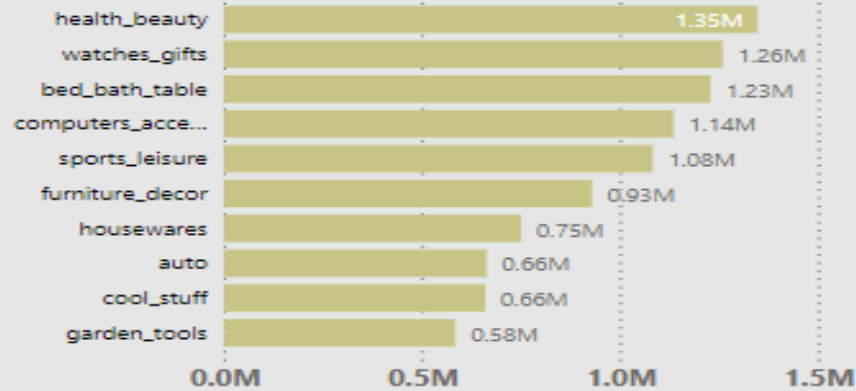
Santa Cruz had 2 payment types- credit card and boleto. It had the highest number of payments for both payment types across the three cities. Had payment value below the average base rate.

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.

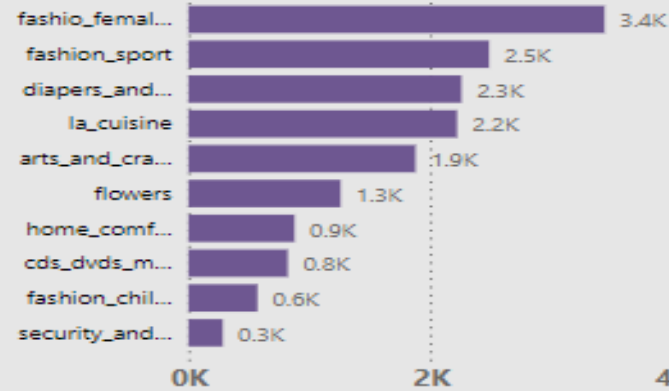
This could indicate that not there could be other factors for neutral sentiments other than a monetary factor.

Analysis of product performance by state, category and year

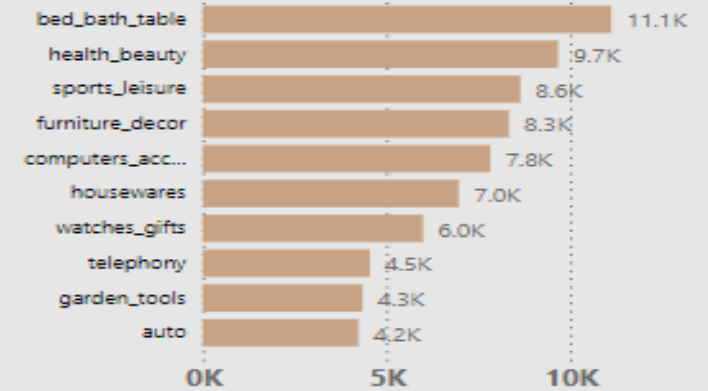
Top product categories sales



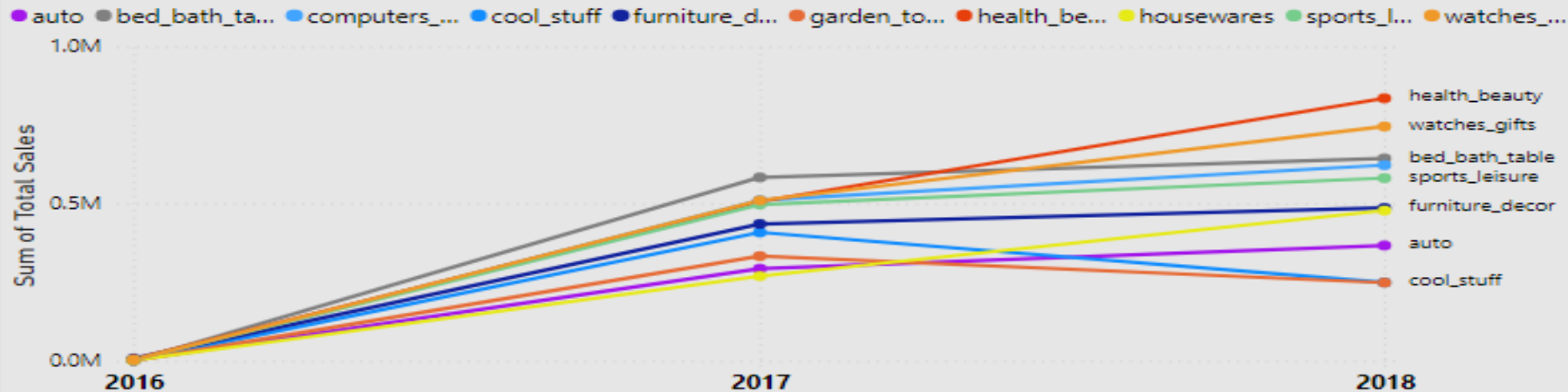
Underperform product categories sales



Top product categories orders



TOP 10 Product Sales by Year



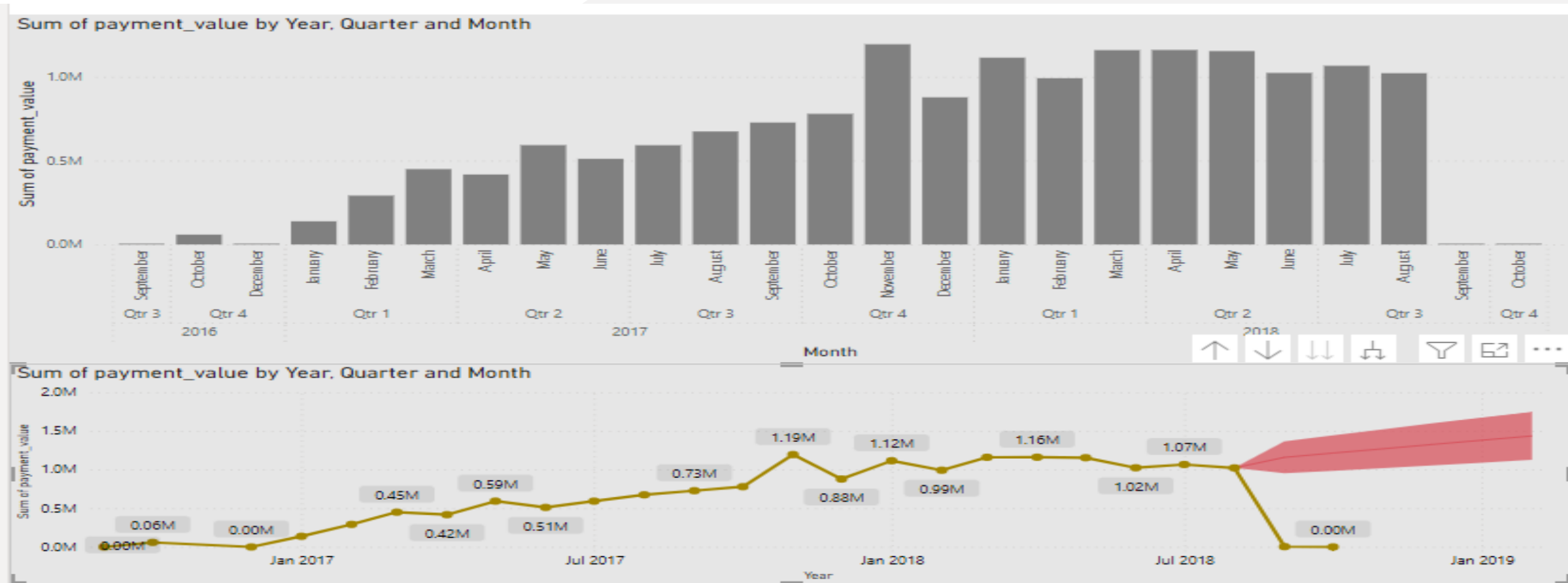
customer_state	Sum of Total Sales
SP	5,900,484.04
RJ	2,089,538.27
MG	1,774,392.74
RS	854,786.83
PR	787,632.81
SC	587,939.84
BA	583,045.05
GO	362,637.20
DF	334,837.59
ES	309,134.77
PE	281,867.42

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

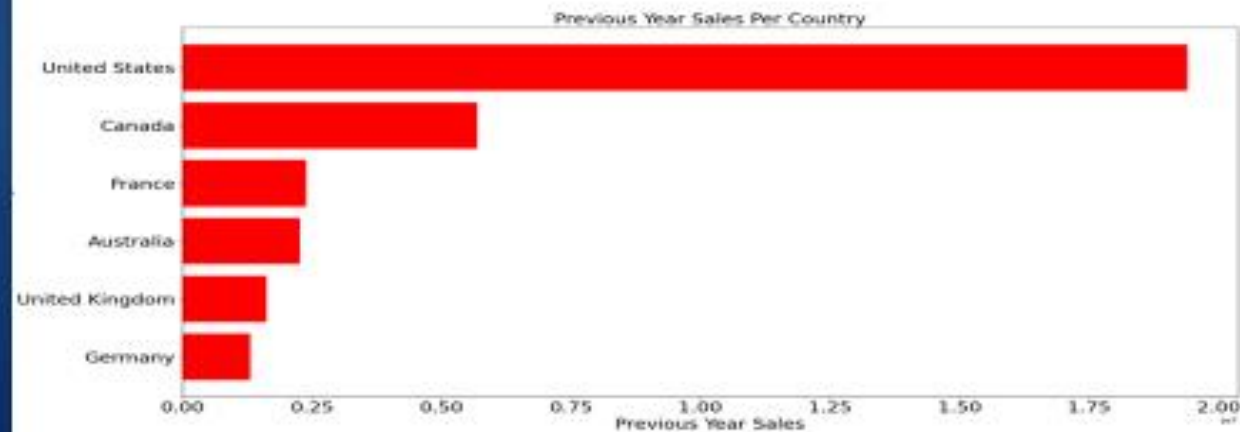
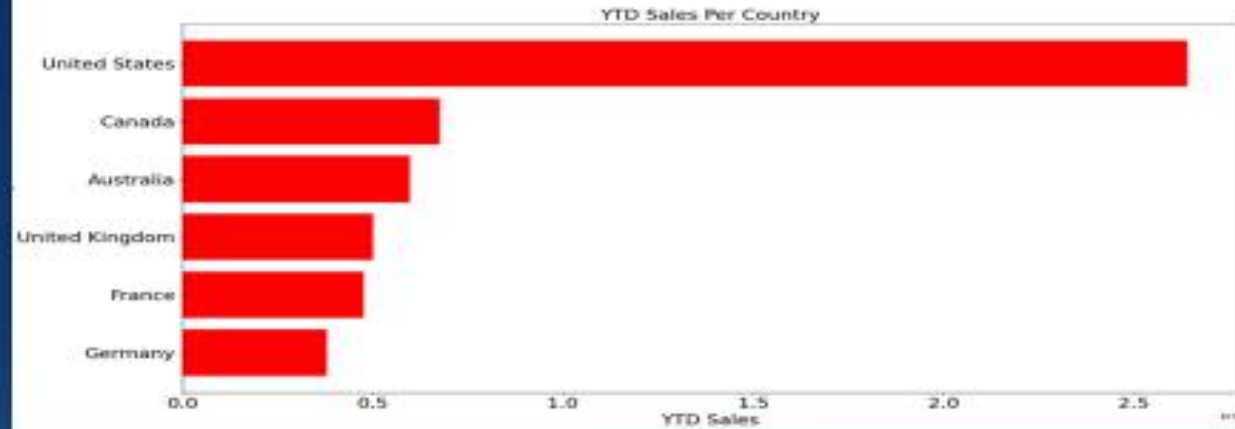
The focus was on exploring the relationship between annual sales and store size, number of employees, revenue and determining the regions with the best sales,

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 the best performing country



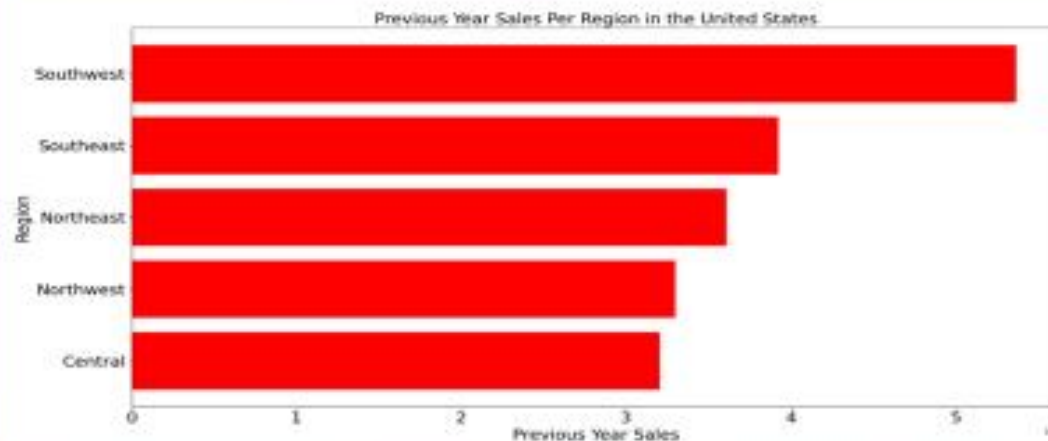
▶ Which country has the best performance?

- ▶ United States is the best performing country

▶ Previous year's performance

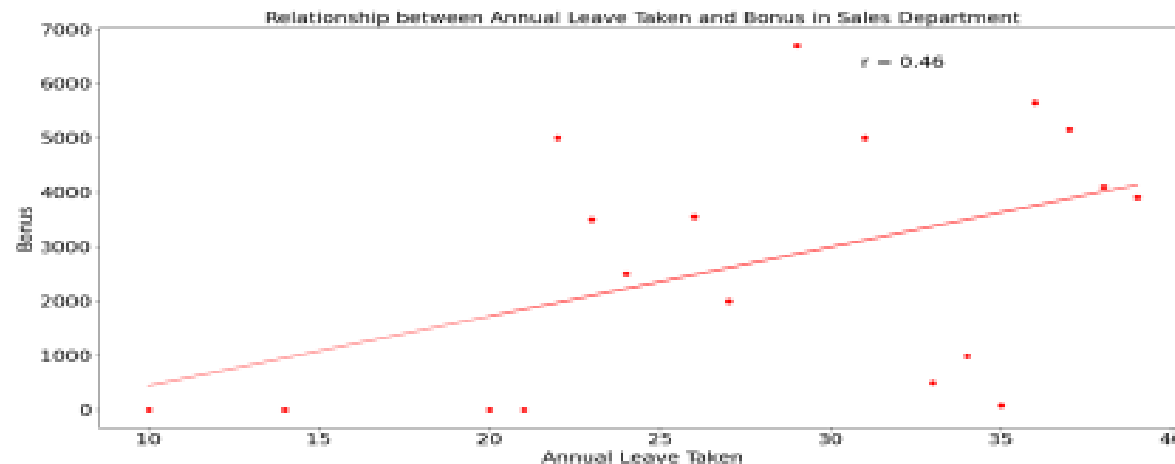
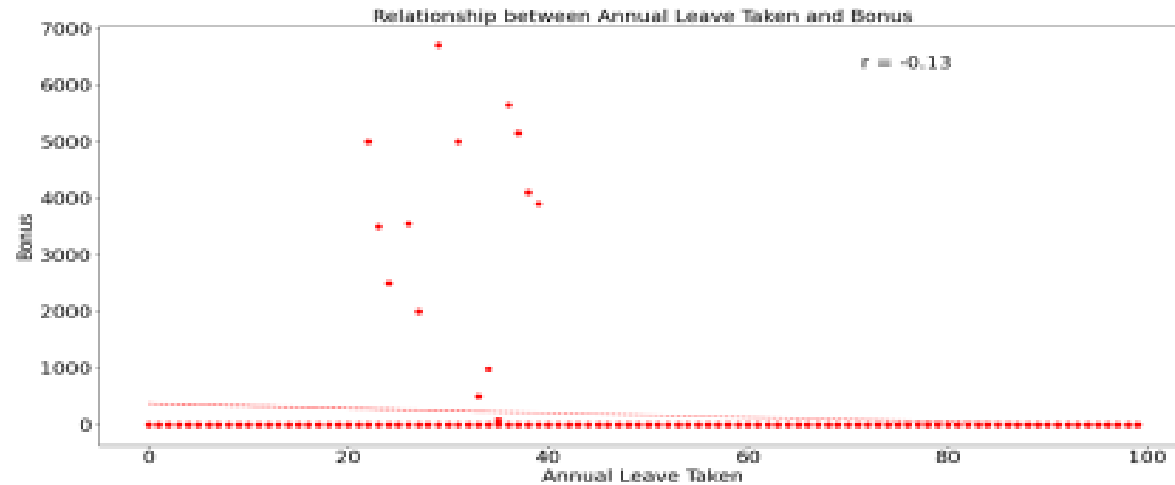
- ▶ United States is the best performing country, both on a YTD and Previous Year basis
- ▶ France was in third place in previous year but in fifth place for current year

Regional sales in the best performing country



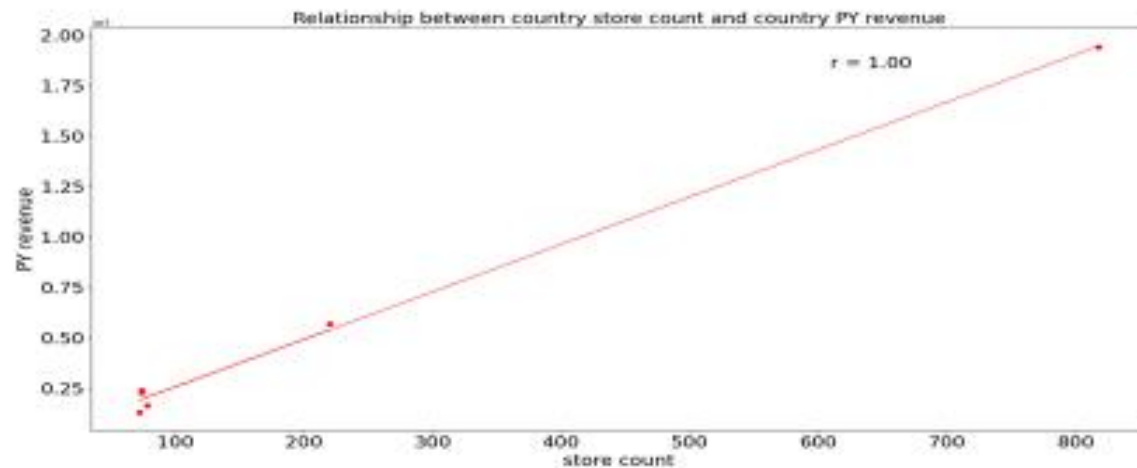
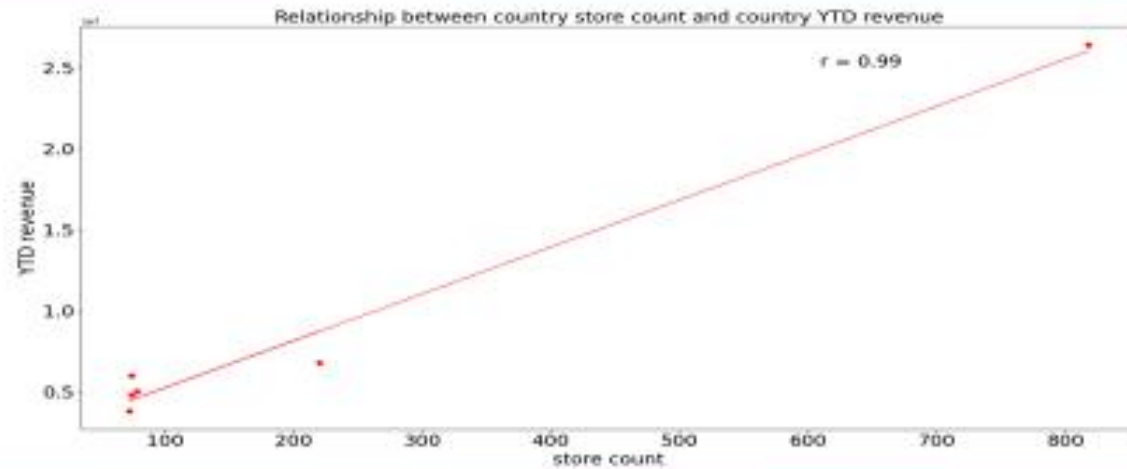
- ▶ Revenue split by region in the best performing country
- ▶ YTD(year to date) regional sales
 - ▶ Southwest region is the best performing region, followed by Northwest
- ▶ Previous Year regional sales
 - ▶ Southwest region is the best performing region

Relationship between annual leave taken and bonus



- ▶ 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

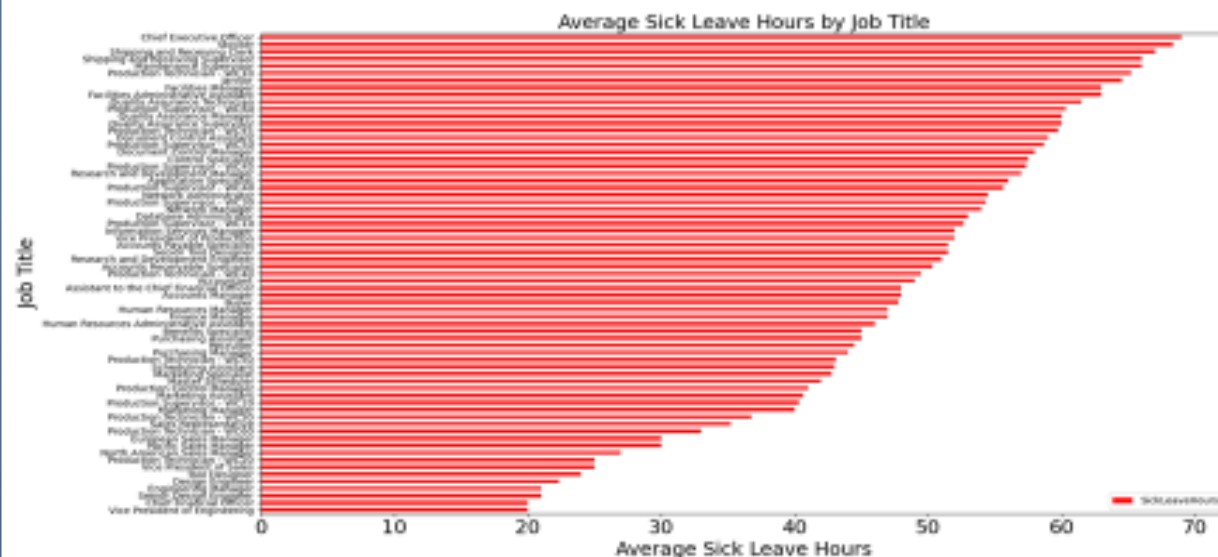
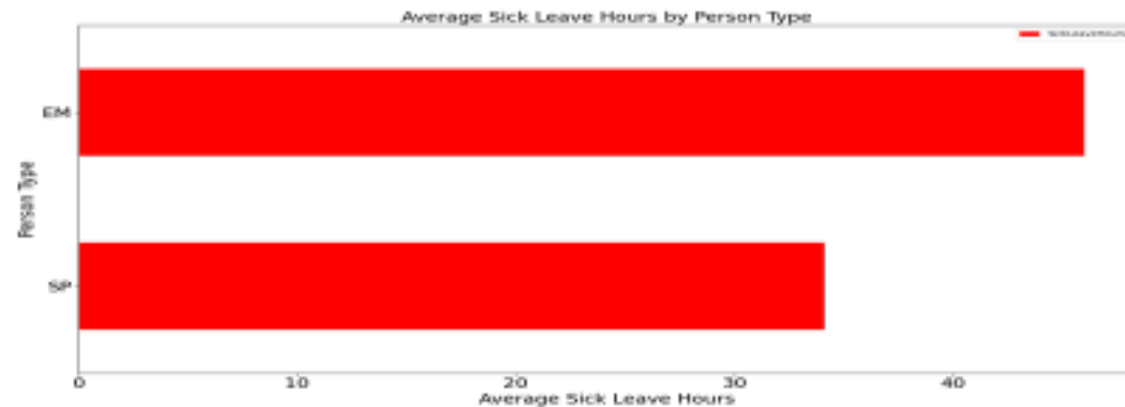


► Exploring the relationship between country store count and revenue shows:

► 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 (Person Type)



- ▶ Using average sick leave for comparison across different job type and title.

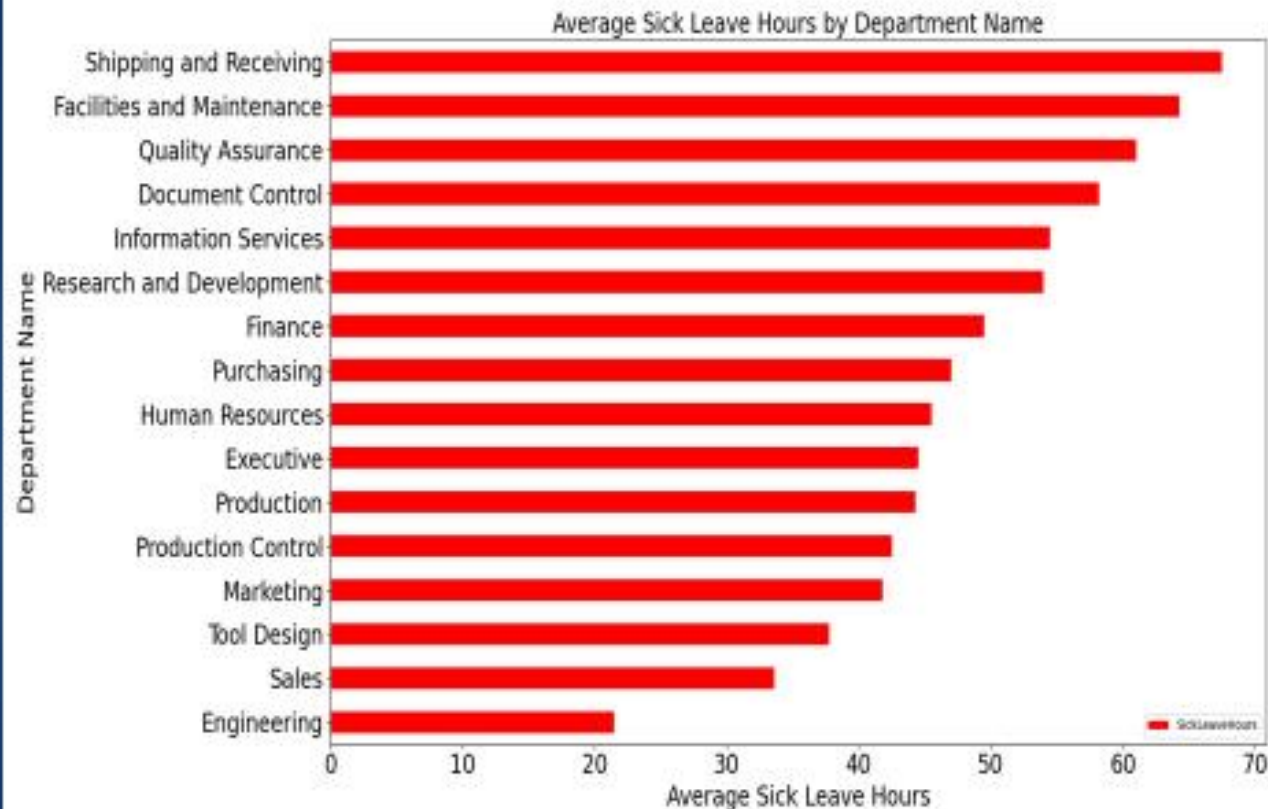
▶ Relationship between sick leave and Person Type

- ▶ 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

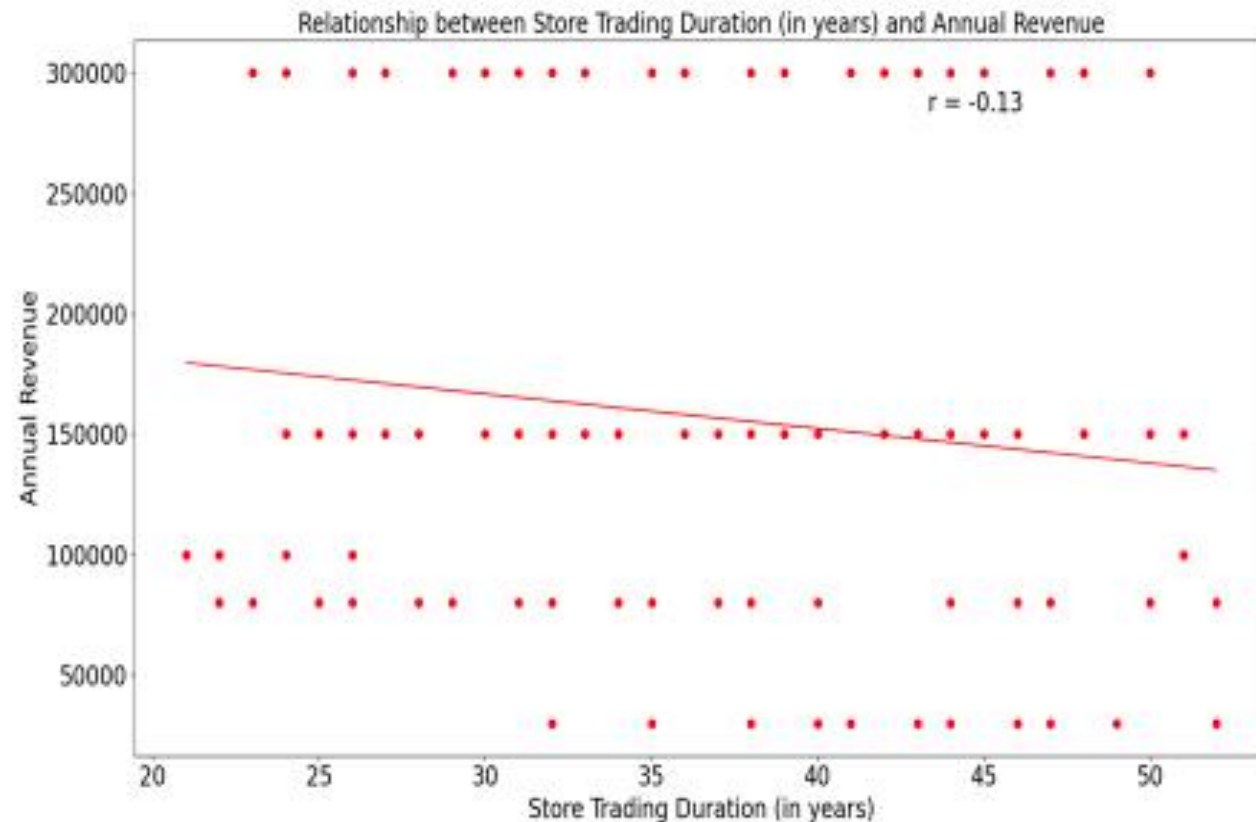
Relationship between sick leave and department



Exploring the Relationship between sick leave and department

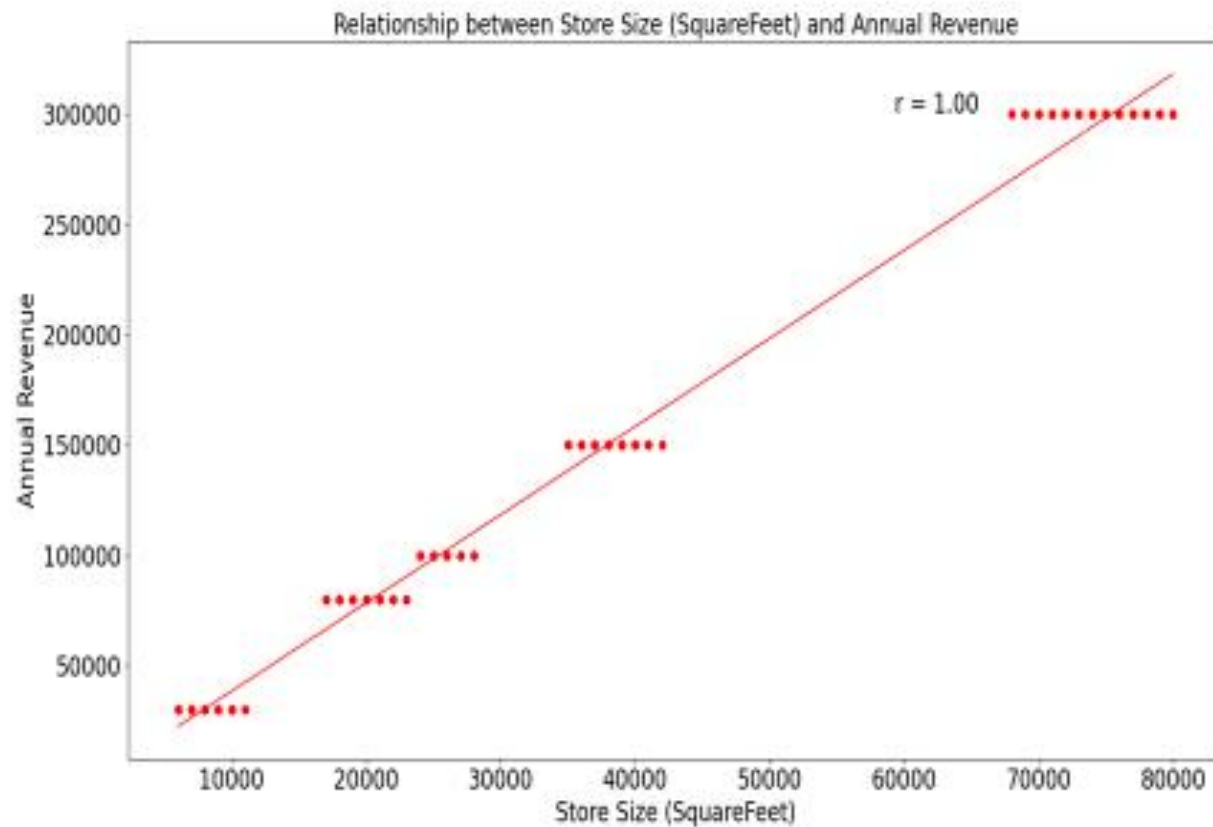
- ▶ The top 3 department that had the most average sick leave were
 - 1) Shipping and Receiving
 - 2) Facilities and Maintenance
 - 3) Quality Assurance

Relationship between store trading duration and revenue



- ▶ 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

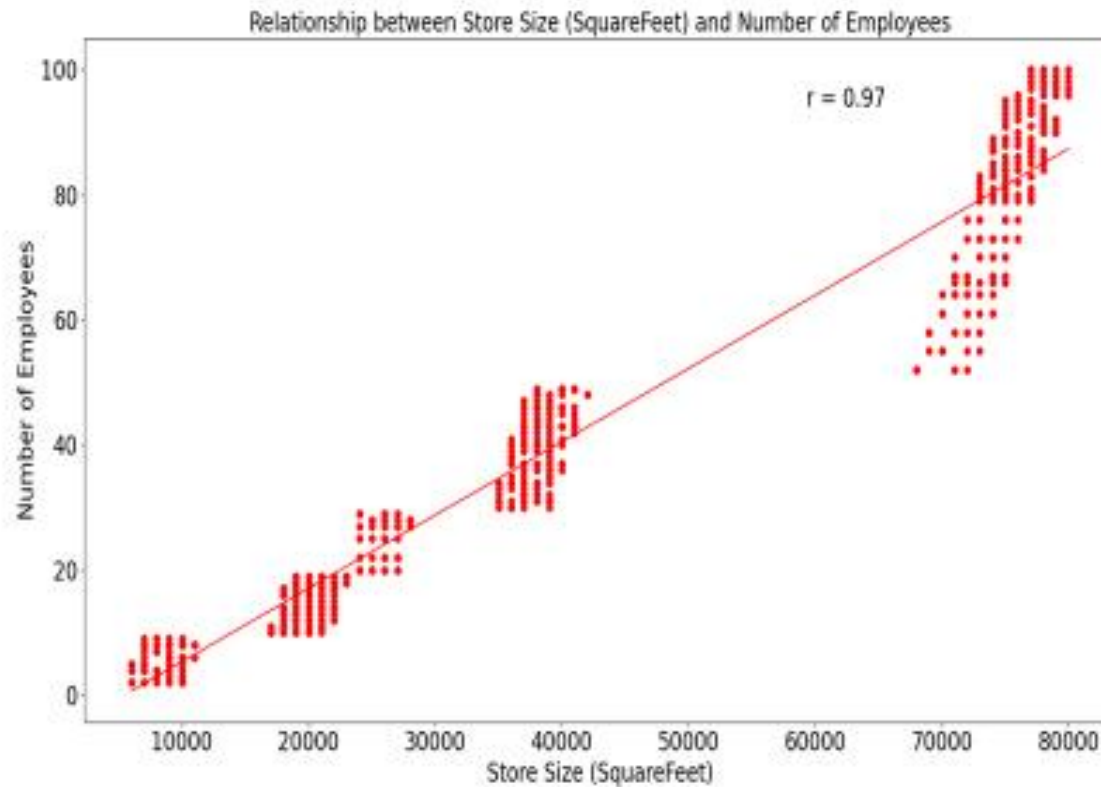
Relationship between the size of the stores and revenue



► Relationship between the size of the stores and revenue

- The correlation coefficient of +1.00 suggest there is a very strong relationship between size of the stores and revenue

Relationship between the size of stores and number of employees



- ▶ 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)

