

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.

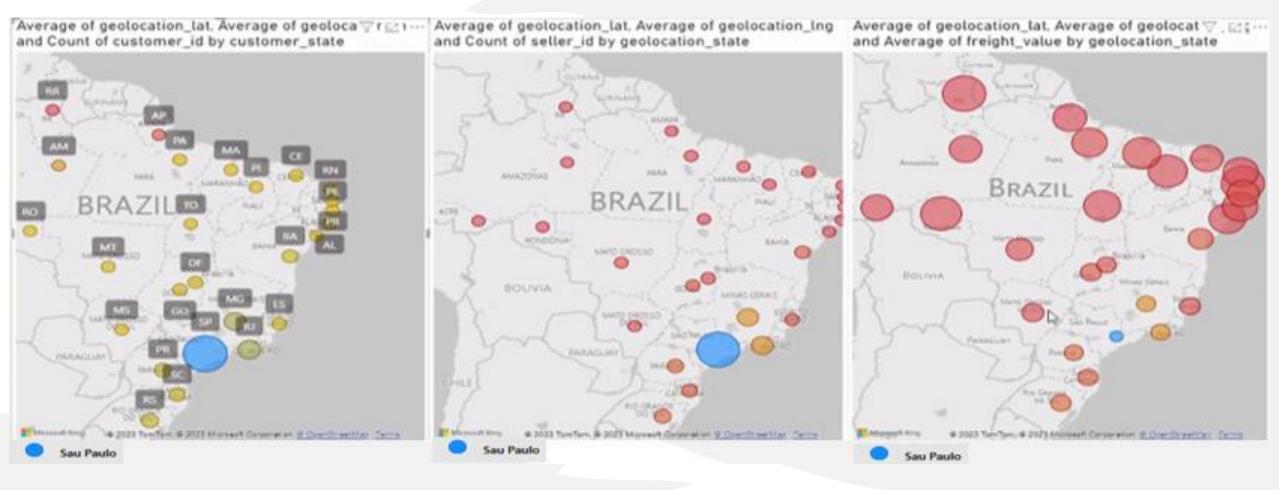
Exploratory data analysis done to understand data structure, identify key variables and modelling of database relationships.

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

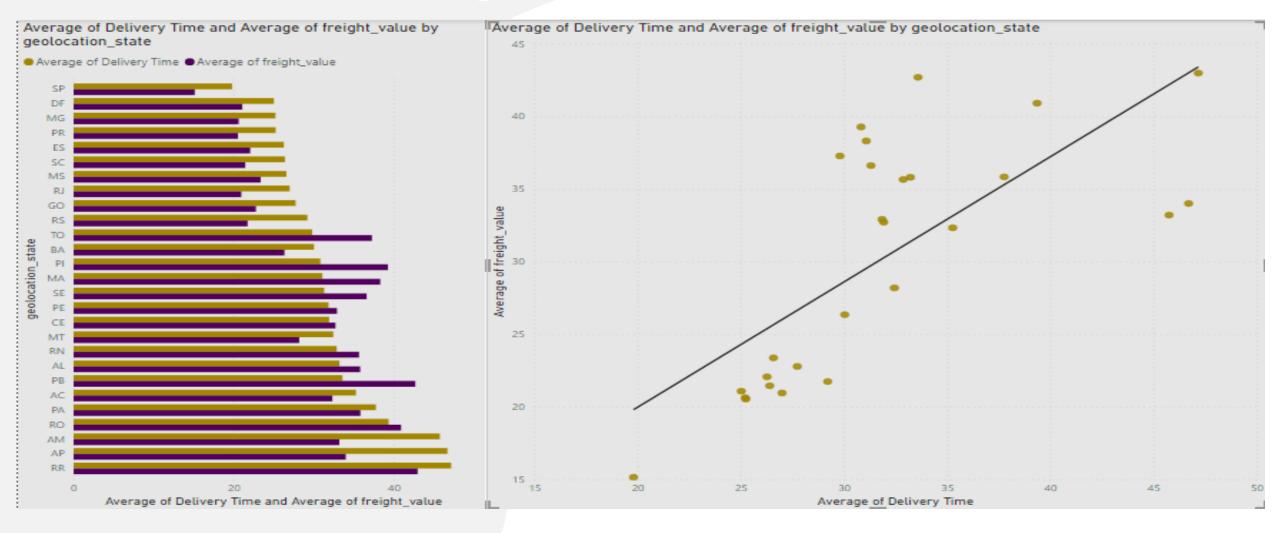


The Southeast region of Brazil, which includes states such as São Paulo, Rio de Janeiro, Minas Gerais, and Espírito Santo, is generally considered the most economically vibrant region in Brazil

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.

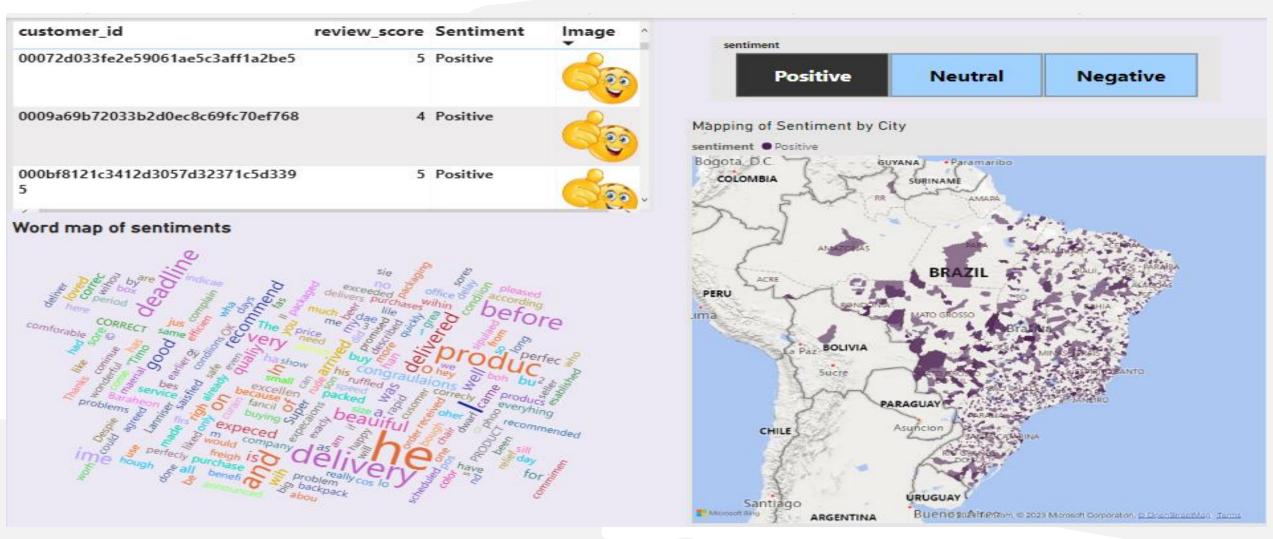
The bigger the size of the bubbles in the average freight value by geolocation chart, the higher delivery costs for sellers to deliver in those outer regions, whereas the bubbles near Sao Paulo are smaller. This be the reason why there are less orders from regions further out.

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

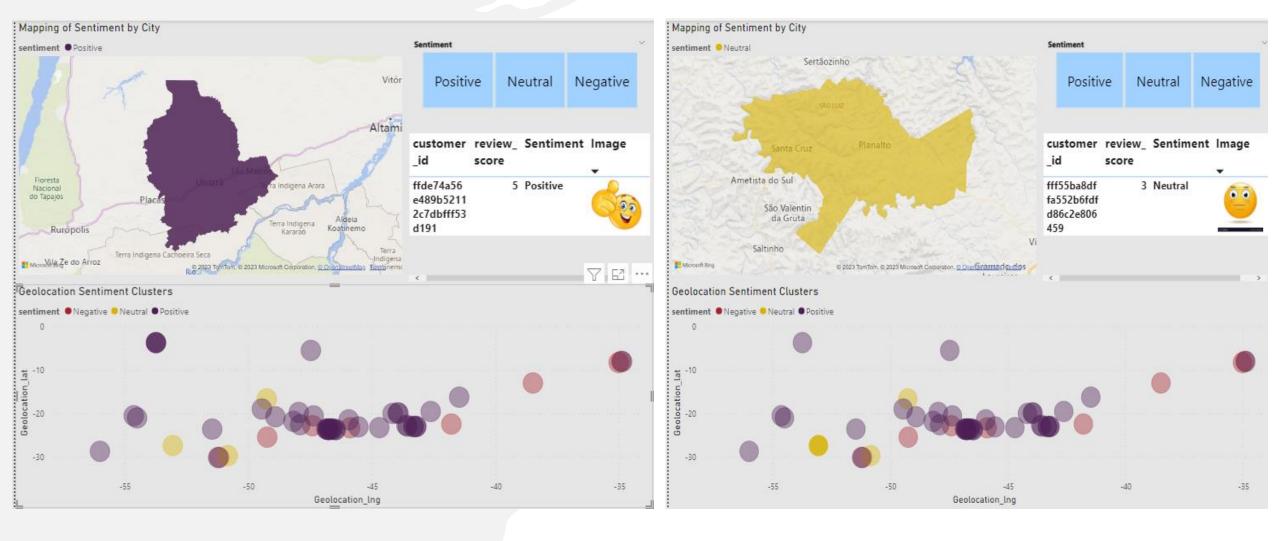
Analysis of sentiment spread across regions



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.

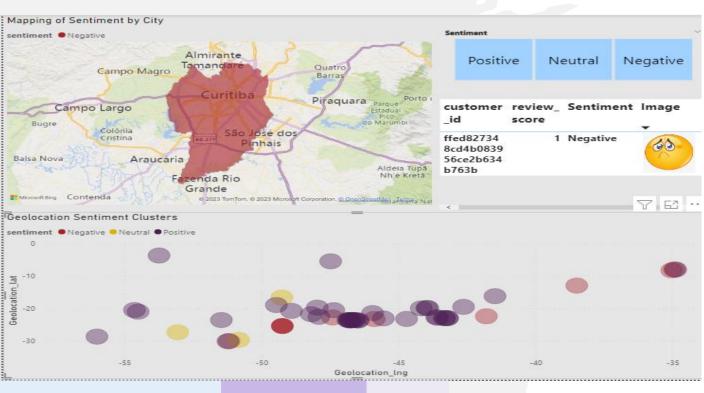
Clustering and sentiment patterns across regions

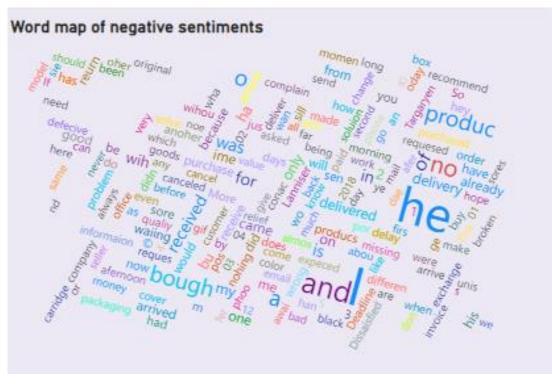


Clustering analysis showed into patterns of customer sentiments across the regions. Customers in the southern regions of — Sao Marcos and Santa Cruz had positive and neutral sentiments, respectively.

The word map gives insight into the sentiment from these customers

Clustering and sentiment patterns across regions





76.47K
Positive Sentiments

99.22K
Total Count of Customer Sentiments

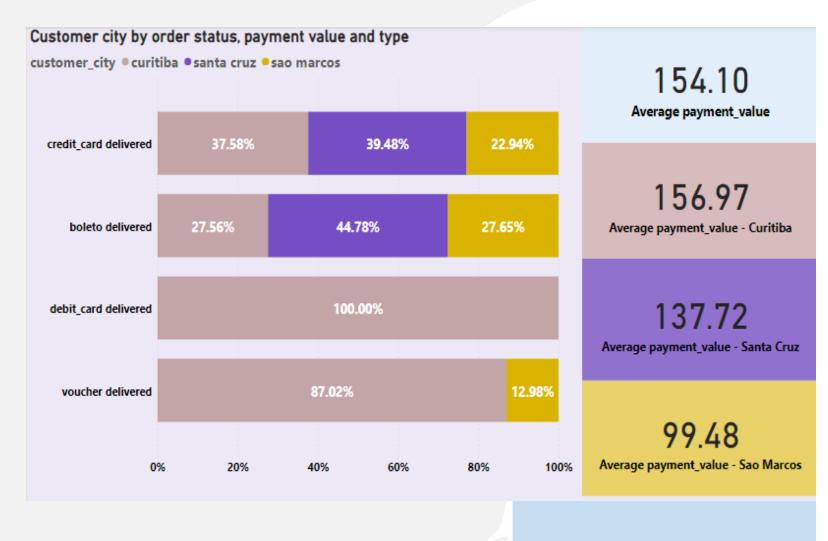
8179
Neutral Sentiments

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



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.

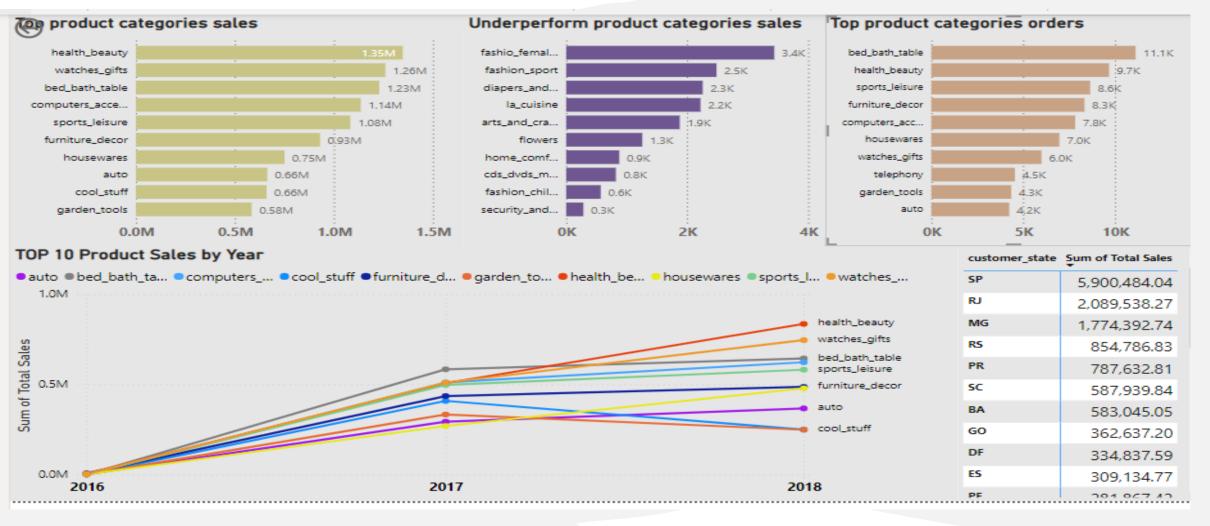
Sao Paulo in southeast region and home to major economic centres, had payment value below the average base rate but higher than the average for **Sao Marcos**, with a neutral customer sentiment.

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

135.83

Average payment_value - Sao Paulo

Analysis of product performance by state, category and year

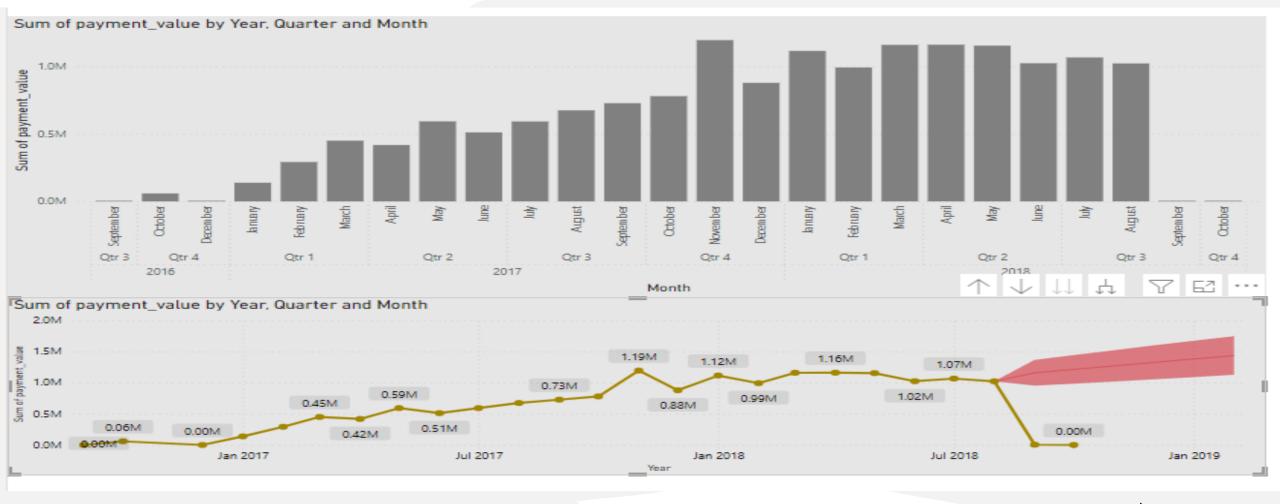


The state **'SP'** where Sao Paulo is located was top performing state with sales of over 5 million R\$, **'RS'** where Sao Marcos is located had sales of over 854,000 R\$ followed by **'PR'** where Curitiba is located at over 787,000 R\$

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

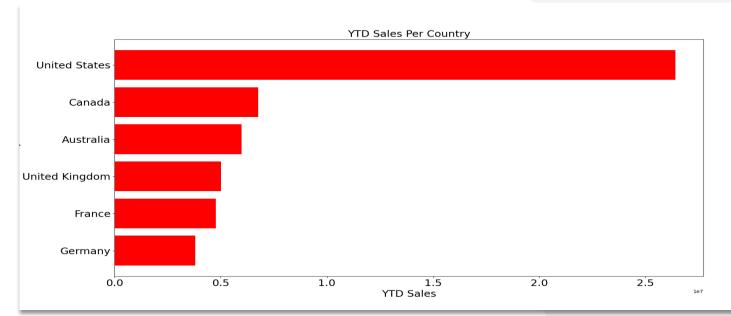
The focus was on exploring the relationship between annual sales and store size, number of employees, revenue and determining the regions with the best sales,

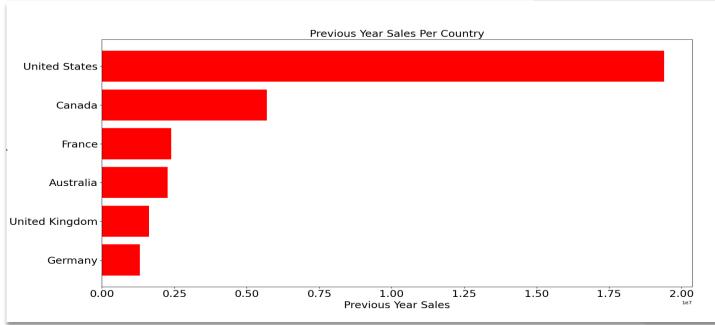
Python and SQL were used to analyze the data, including deciphering specific data relevant to the questions.

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.

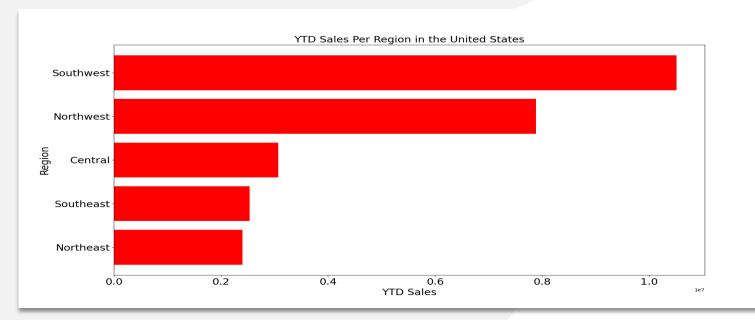
Sales in best performing country

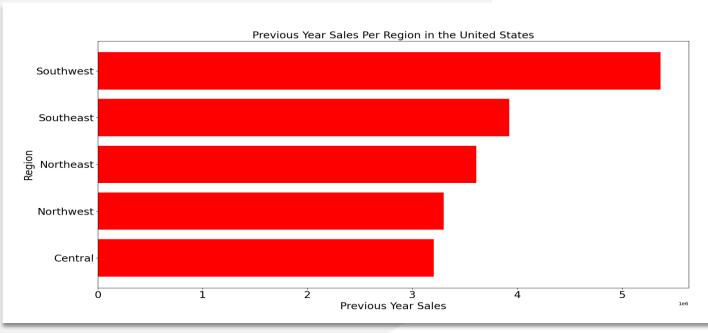




Which country has the best performance? United States is the best performing country Previous year's performance United States is the best performing country, both on a YTD and Previous Year basis France was in third place in previous year but in fifth place for current year

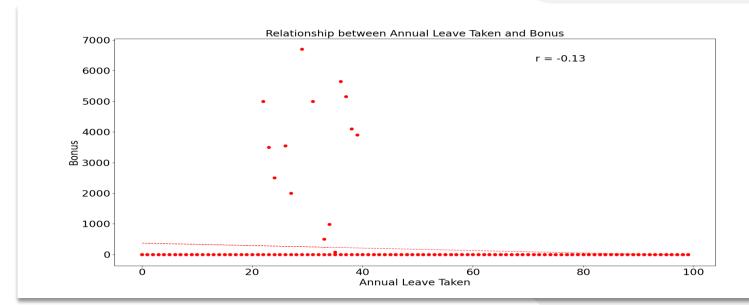
Regional sales in best performing country

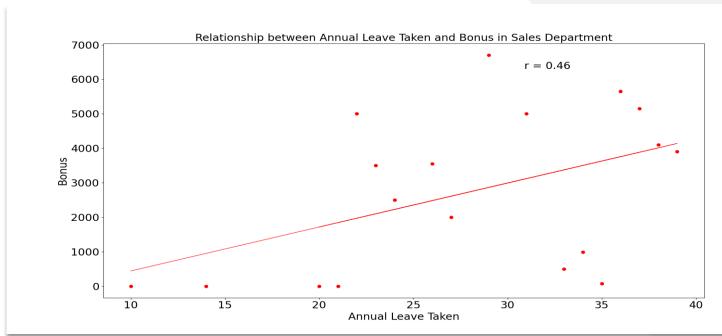




Revenue split by region in the best performing country YTD(year to date) regional sales Southwest region is the best performing region, followed by Northwest Previous Year regional sales Southwestregion is the best performing region

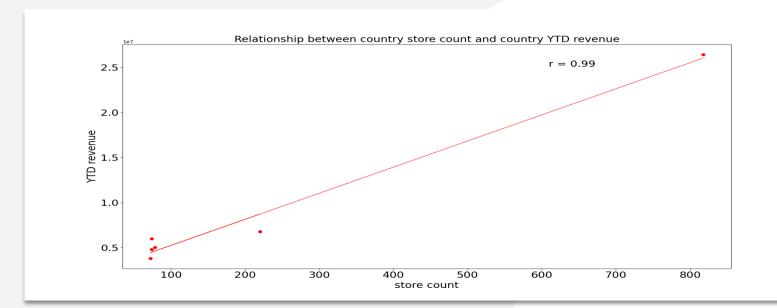
Relationship between annual leave taken and bonus

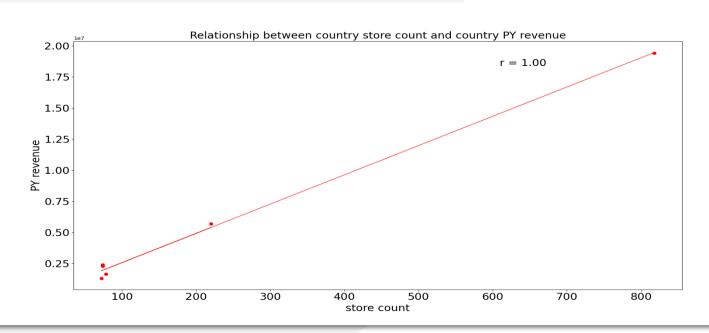


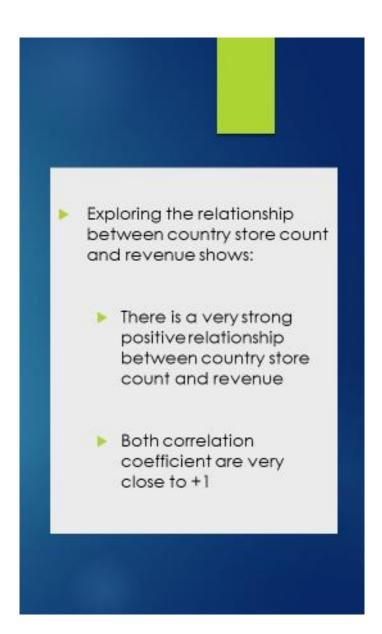


- The chart shows that a lot of employees do not receive a bonus, and in fact only the sales department employees received bonuses
- The correlation coefficient of -0.13 suggest a very weak or no relationship between annual leave and bonuses
- The correlation coefficient of 0.46 suggest a moderate positive relationship between annual leave taken and bonus
- However, moderate positive relationship does not mean there is a causal factor.

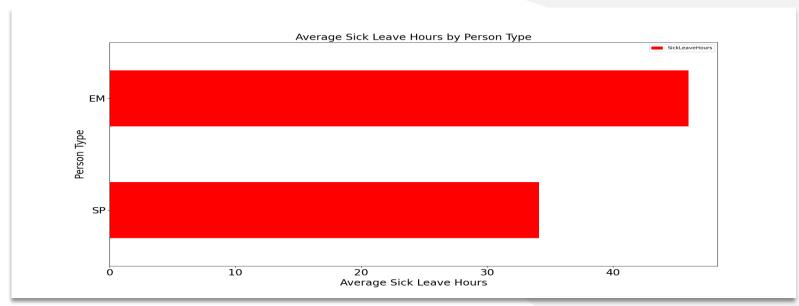
Relationship between Country store count and Revenue

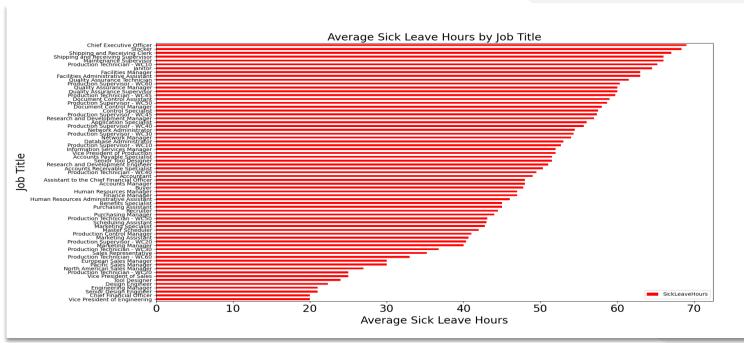






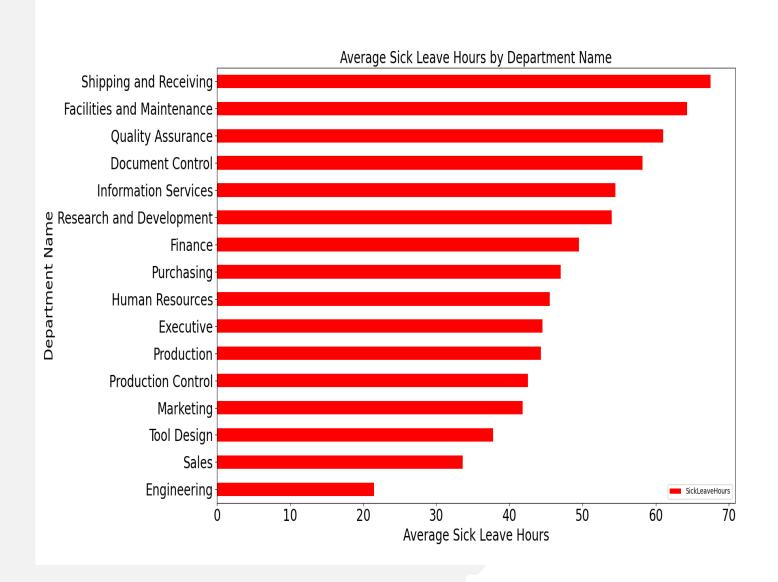
Relationship between sick leave and job title

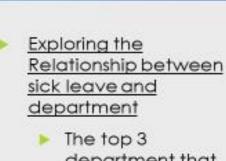




- Using average sick leave for comparison across different job type and title.
- Relationship between sick leave and Person Type
 - Non-sales employee had higher average sick leave than sales employee(SP – salesperson)
- Relationship between sick leave and Job Title
 - The top 3 job titles that had the most average sick leave were
 - 1) CEO
 - 2) Stocker
 - 3) Shipping and Receiving Clerk

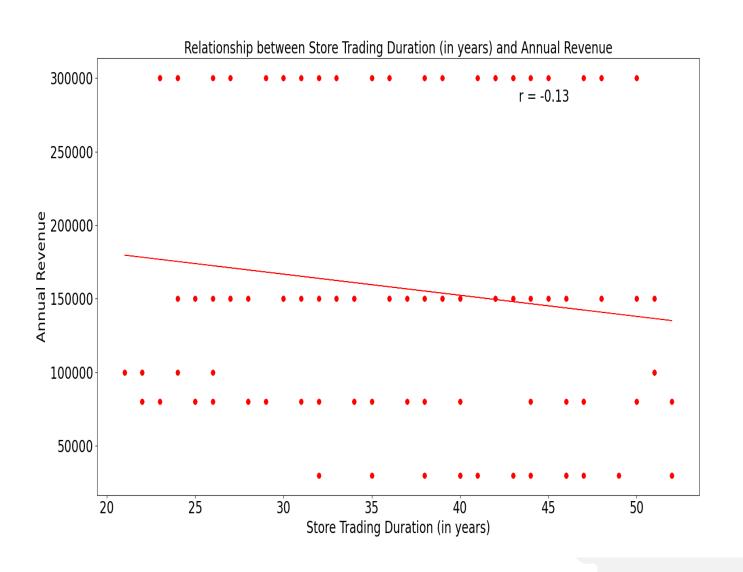
Relationship between sick leave and department





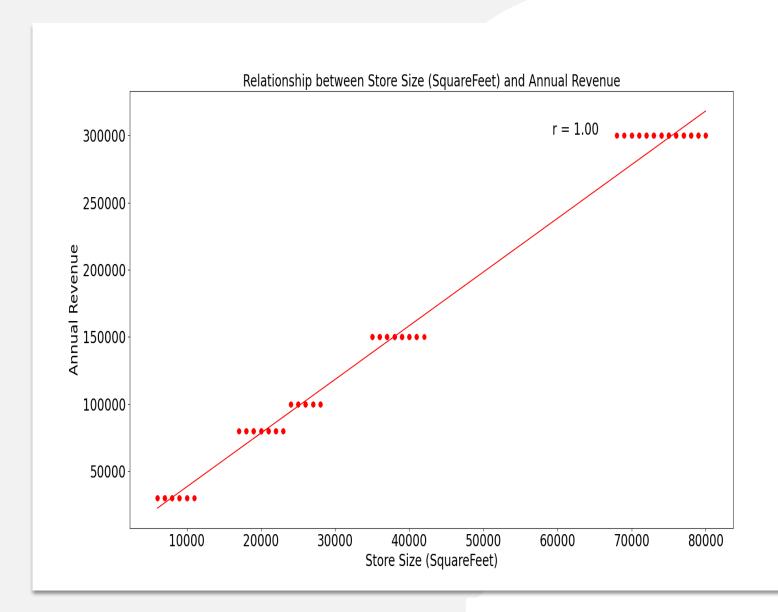
- department that had the most average sick leave were
- 1) Shipping and Receiving
- Facilities and Maintenance
- Quality Assurance

Relationship between store trading duration and revenue



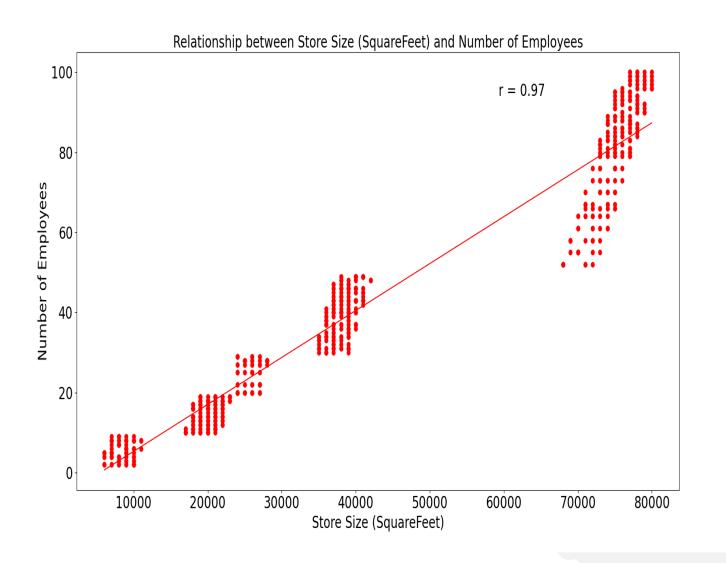


Relationship between size of store and revenue





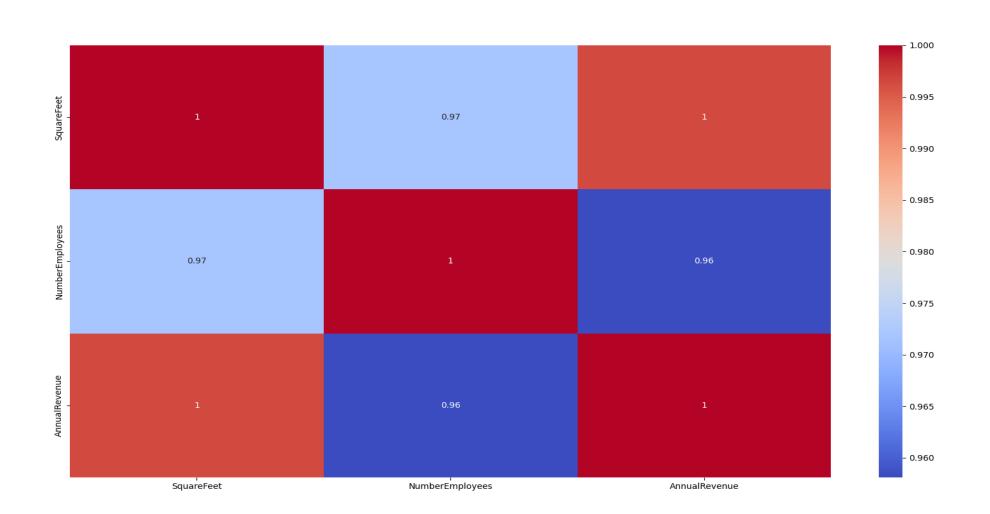
Relationship between size of store and number of employees





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

SQL CODE: Sales in best performing country

```
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

```
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'
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'

WHERE cr.[Name] = 'United States'

SQL CODE: Relationship between annual leave taken and bonus

Specify columns to be retrieved using SELECT statement

```
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'.
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')
WHERE sp.[Bonus] IS NOT NULL) AS adj table
ORDER BY Bonus DESC
```

PYTHON CODE: Sales in best performing country

```
# 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

```
# 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

```
# 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

```
# 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.

```
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

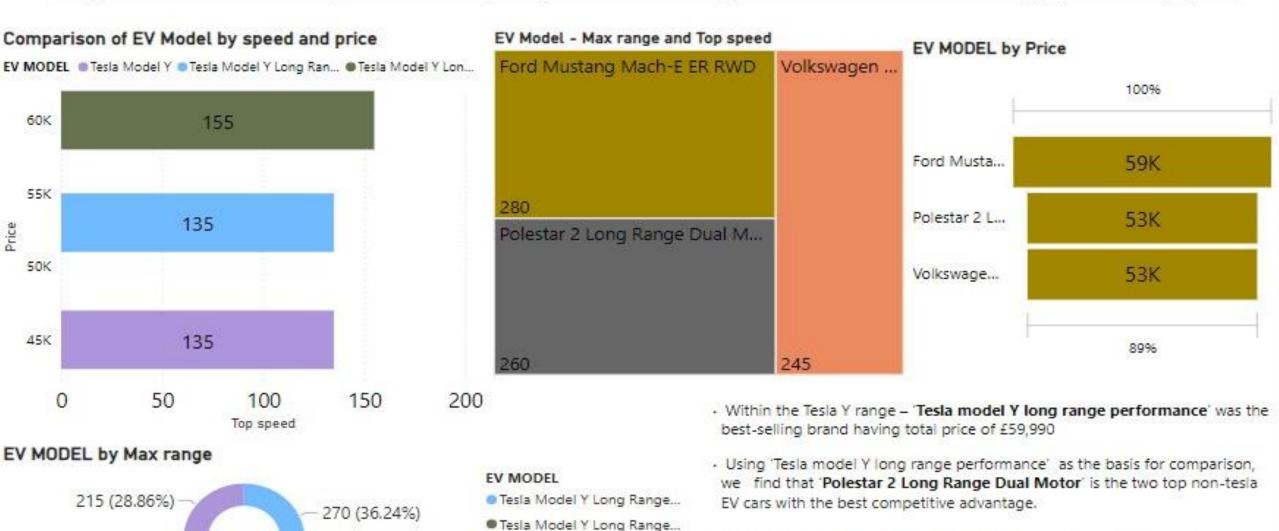
```
■ Specify columns to be retrieved using SELECT statement
SELECT *
FROM (
 SELECT Coin, Mkt_Cap,[_1h],[_24h],[_7d],
     Assign a rank to each row based on market capitalization value (Mkt_Cap)
    RANK() OVER (ORDER BY Mkt Cap DESC) AS Mkt Cap rank,
     Assign a rank to each row based on percentage 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;
```

Competitor Analysis - Tesla Electric Vehicle (EV) Dashboard.pbix



260 (34.9%)

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



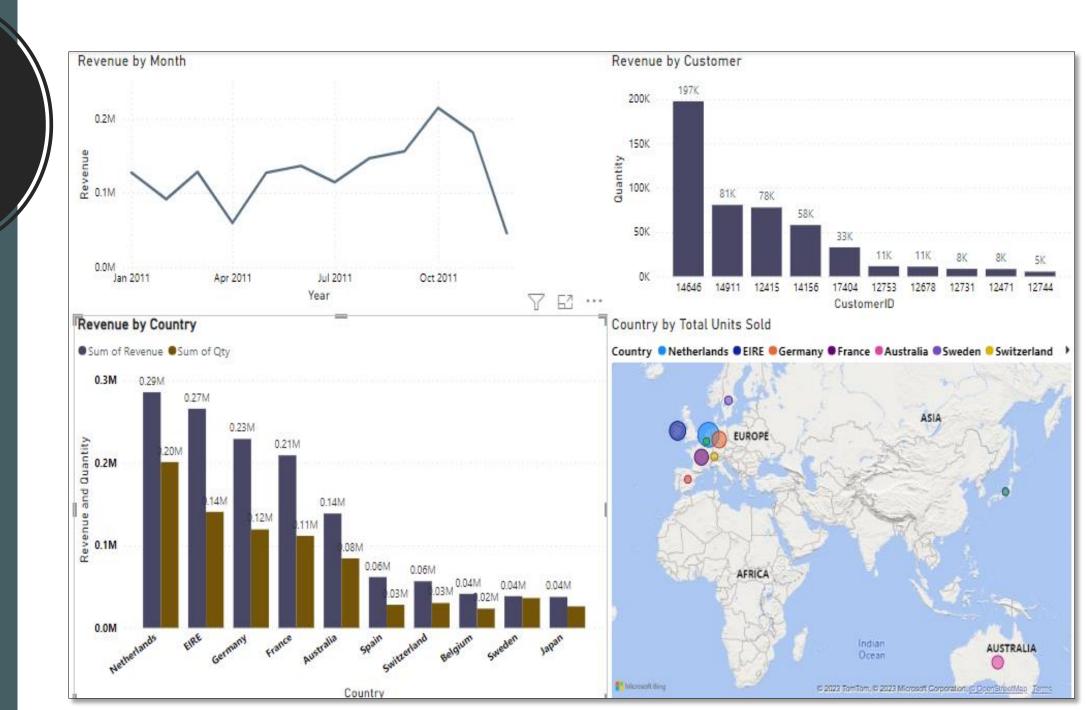
Tesla Model Y

Polestar 2 Long Range Dual Motor is the best EV car in terms of price

for money in terms of max range (number of miles per charge)

and speed, while 'Ford Mustang Mach - EER RWD' gives better value

<u>Data</u>
<u>Visualization</u>
<u>Dashboard –</u>
<u>Business</u>
<u>Insight</u>





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

Q datacaмр



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

Adatacamp

A

Jonathan Cornelissen, CEO



STATEMENT OF ACCOMPLISHMENT

#29,447,819

HAS BEEN AWARDED TO

Oge Ibezi

FOR SUCCESSFULLY COMPLETING

Understanding Data Engineering

LENGTH

2 HOURS

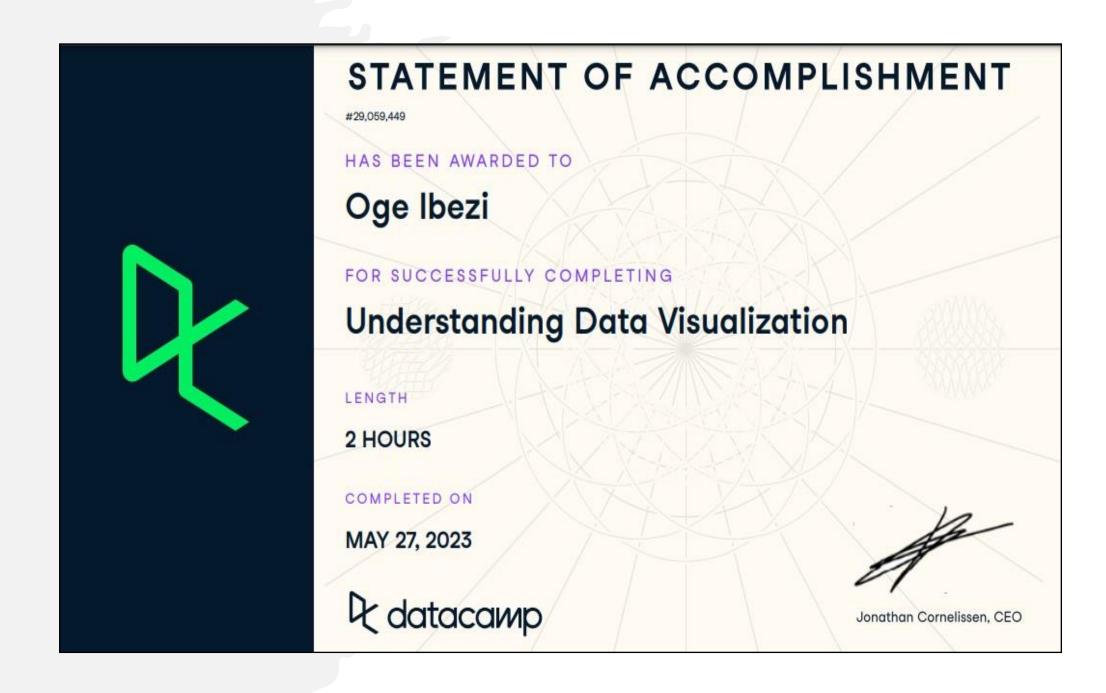
COMPLETED ON

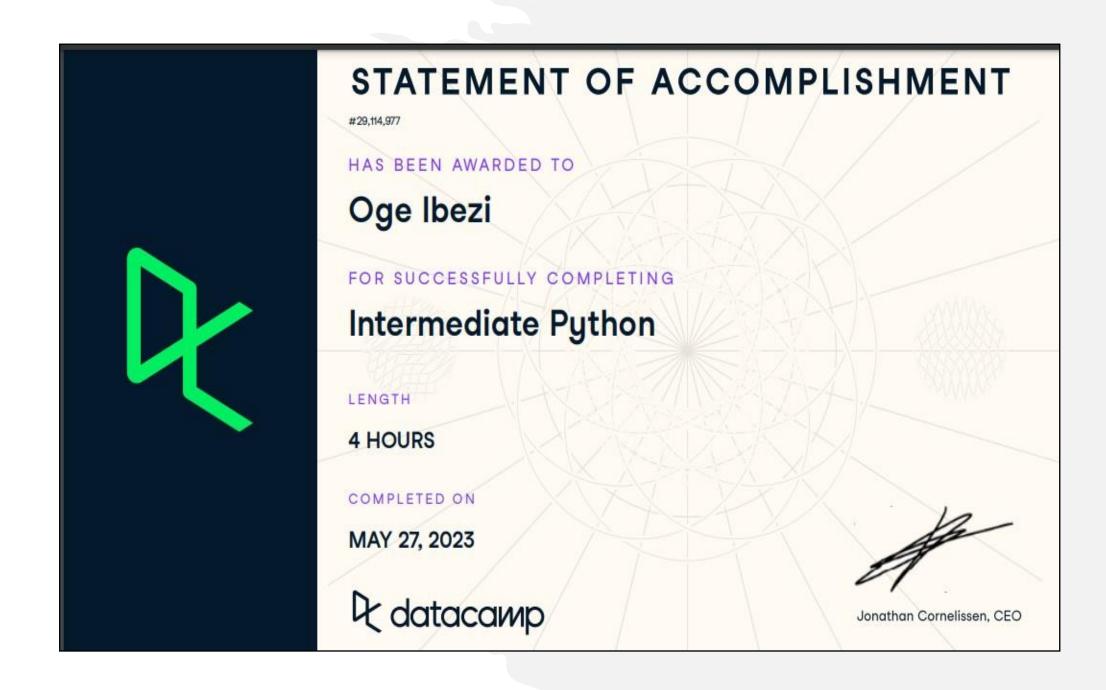
JUN 12, 2023

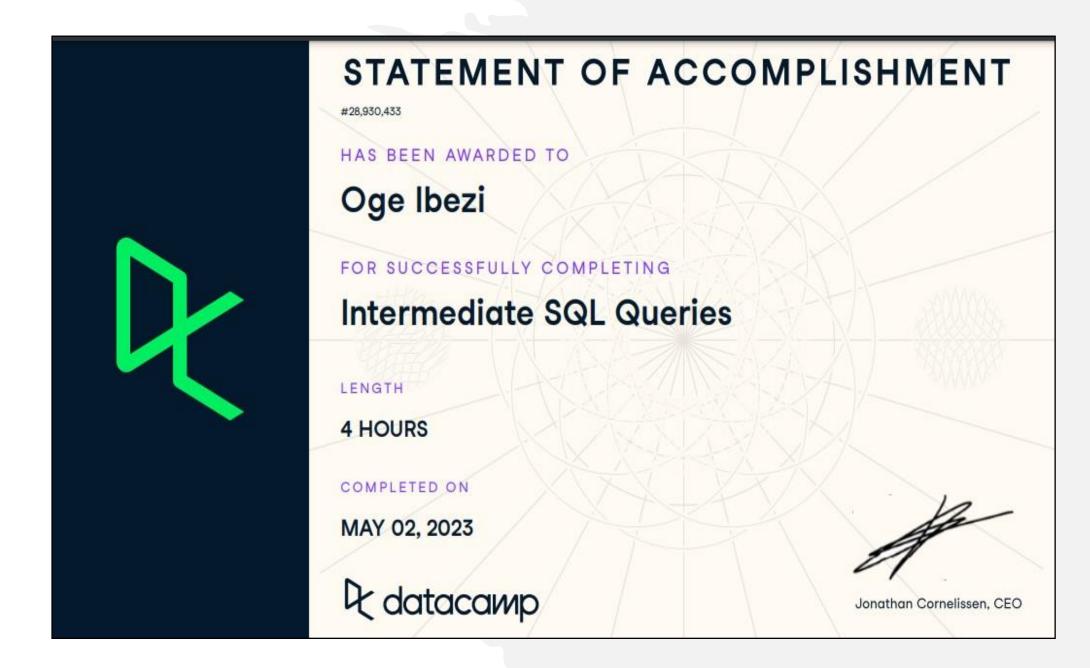
Q datacaмр

A

Jonathan Cornelissen, CEO







Many