**Group 4**

**Data Analytics Interim Project Proposal**

**Overview**

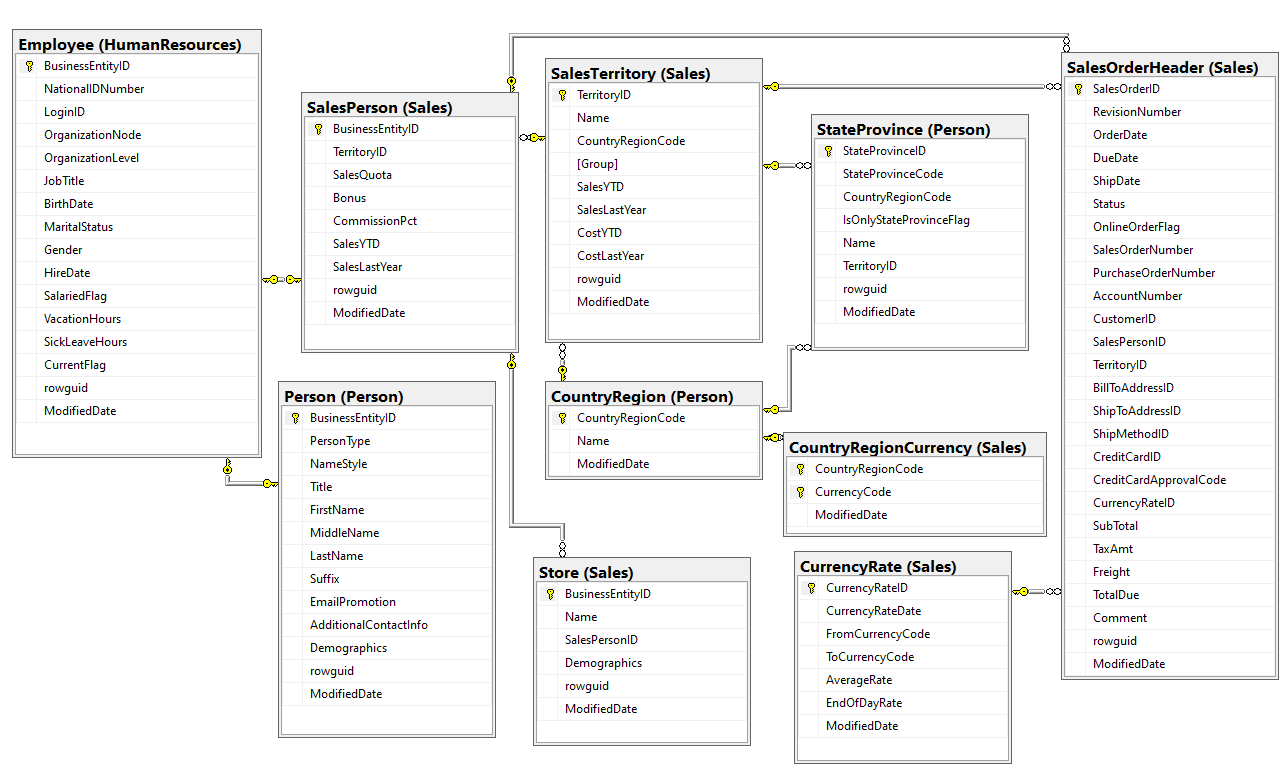
Our group looked into AdventureWork2019 database and investigated the relationship between multiple variables in the data in order to determine any issues, make meaningful insights and identify potential opportunities for the company. We used AdventureWork2019 database to create SQL queries to simplify and efficiently extract refined data in a CSV format and importing it to Python using Pandas.

SQL queries and Python calculations and visualisation highlighted a number of trends, patterns and relationships in AdventureWorks2019 in regard to its Stores, Sales, Revenue and Employees.

We presented the results using charts and graphs to get a better clarity of the relationships (if there is any) between the variables. The insights presented key results, highlighted any outliers, and identify areas for improvement.

|  |  |
| --- | --- |
| Proposed by: | Group 4  Sylwia Kisielewska  Ola Abdelrazek  Vina Suzette Matibag  Oluwaseyi Badero |
| Timeframe: | Completion by 02/04/2025 time 14:55.  Presentation on 02/04/2025 time 14:15. |

**Concept Schema:**



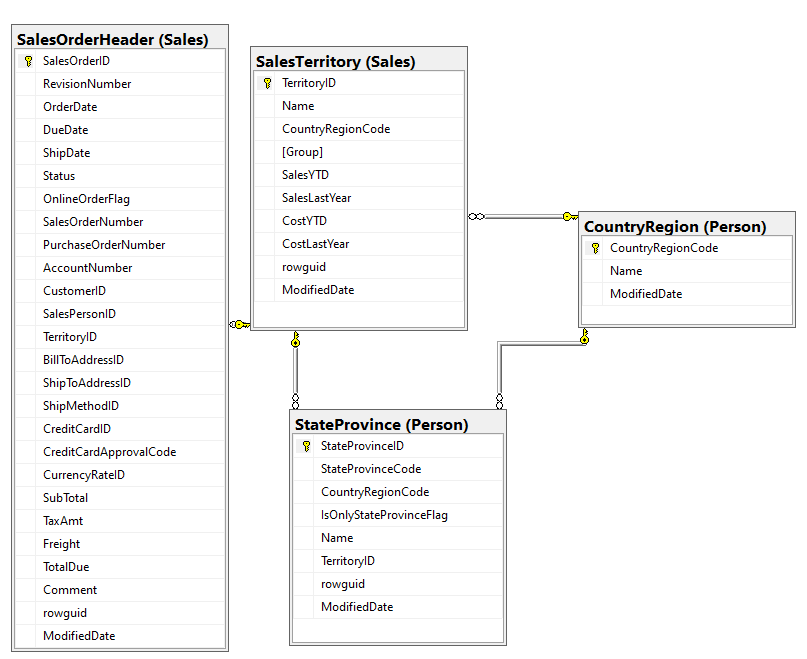
**Objectives and steps**

With the use of specific data found in Adventure works we were tasked to answer the following questions:

1. What are the regional sales in the best performing country?

**Introduction**

This report provides an analysis of total sales across different countries and sales performance in various regions of the highest-performing country. The data was extracted from a relational database using SQL queries and then processed in Python for visualization. The findings provide insights into the best-performing countries and regions, helping in strategic decision-making.

Schema of tables used

**Methodology**

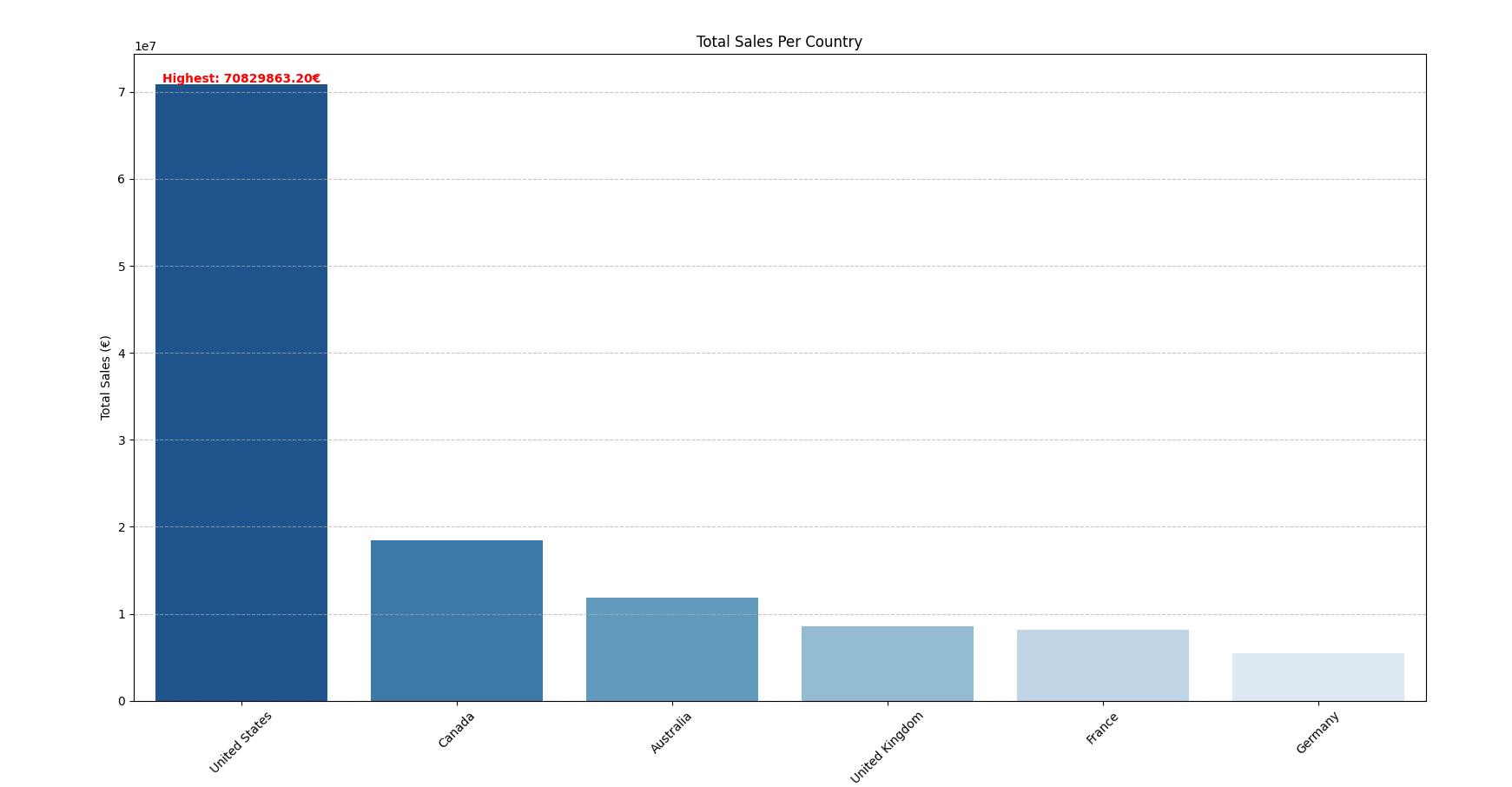
The analysis involved the following steps:

* Extract Data using SQL Queries
* Total sales per country were computed using SQL joins on SalesOrderHeader, SalesTerritory, StateProvince, and CountryRegion tables.
* The highest-performing country was identified from the query results.
* Regional sales data was retrieved for this top country.
* SQL query results were exported as CSV files.
* Python was used to load and process these files.
* Visualize Data using Matplotlib and Seaborn
* Two bar charts were plotted: one for total sales per country and another for sales per region in the highest-performing country.

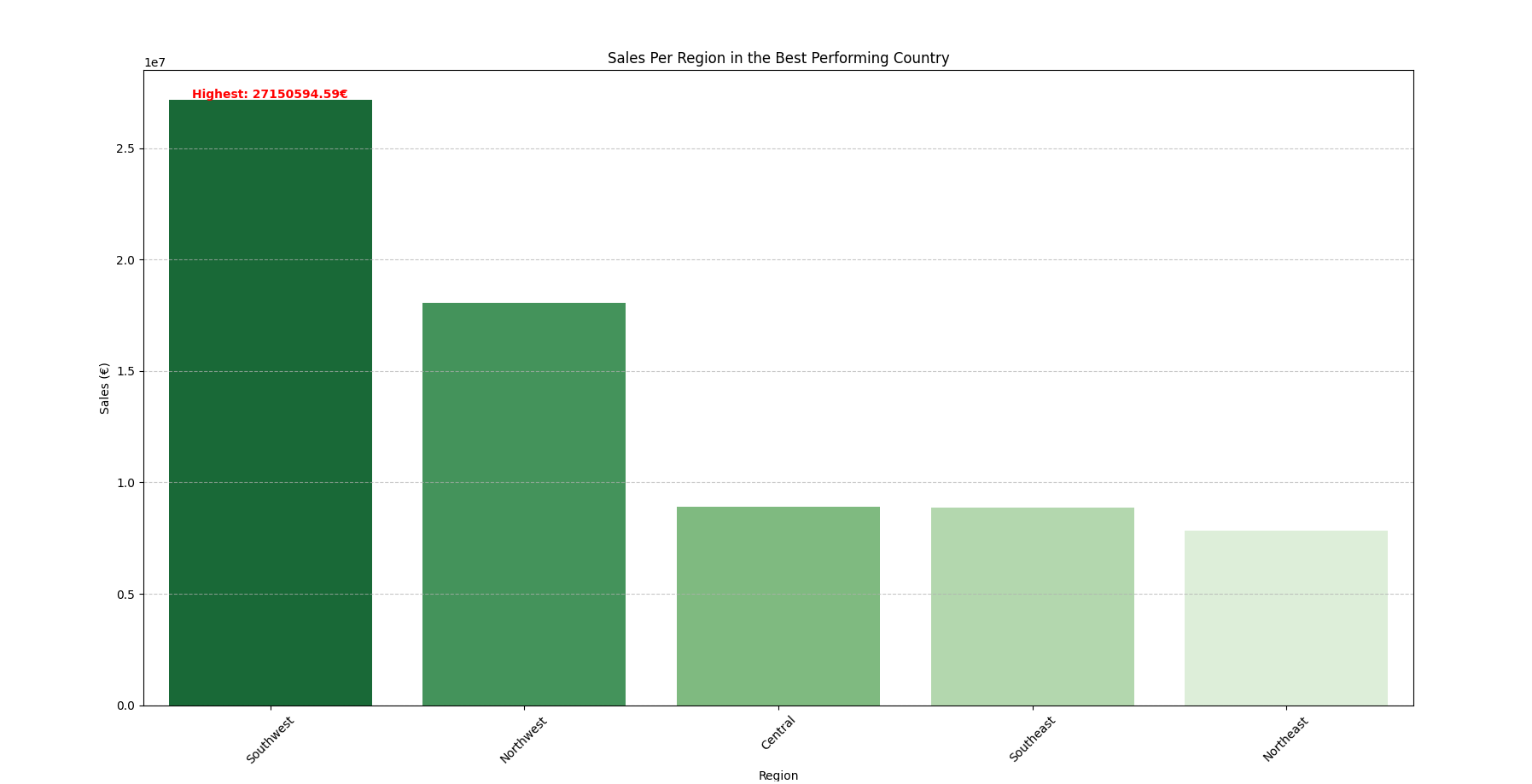
**Key Python Steps:**

* Load sales data from CSV files using pandas.
* Sort data by total sales for correct ranking.
* Use matplotlib and seaborn to create bar charts.
* Highlight the highest-performing country and region.

**Results & Insights**

*Country Sales Analysis*

* The highest-performing country is United States with €70.8M in total sales.
* Other top performers include Canada (€18.4M), Australia (€11.8M), and United Kingdom (€8.6M).
* France, ranks fifth with €8.1M.

*Regional Sales in the Best Country (USA)*

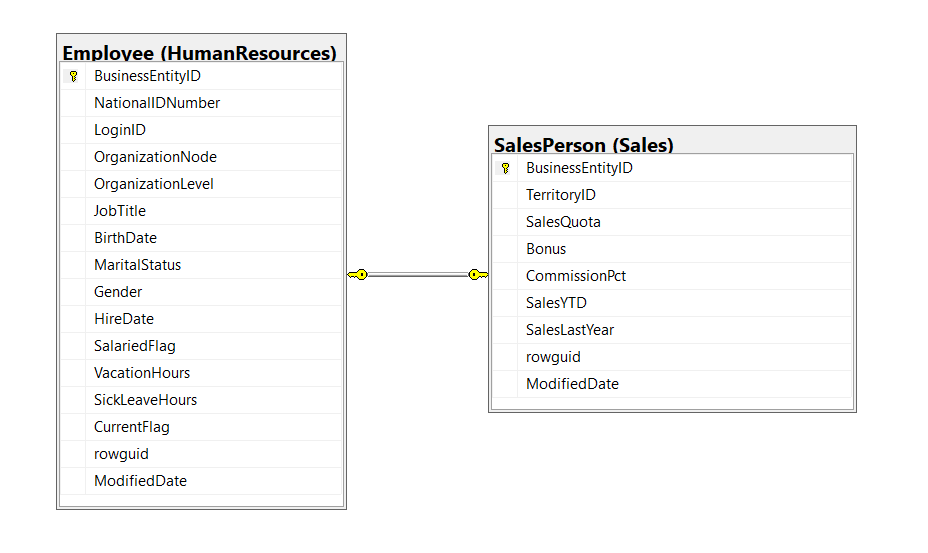
* The region with the highest sales within the USA is Southwest with 27150594.5893.
* Other regions show varying performance, and further analysis could determine key market influences.

**Conclusion**

This analysis successfully identified the top-performing country and its most successful regions. These insights can guide market strategy and regional focus for optimizing sales performance. Further exploration is made below in terms of relationships between store size, number of employees and revenue.

1. What is the relationship between annual leave taken and bonus?

**Introduction**  
This report provides an analysis if there is a relationship between annual leave taken and bonus. The data was extracted from a relational database using SQL queries and then processed in Python for further calculation and visualization. The findings provide insights into establishing if there is any relationship between bonus and annual leave taken.



Schema of tables used

**Methodology**

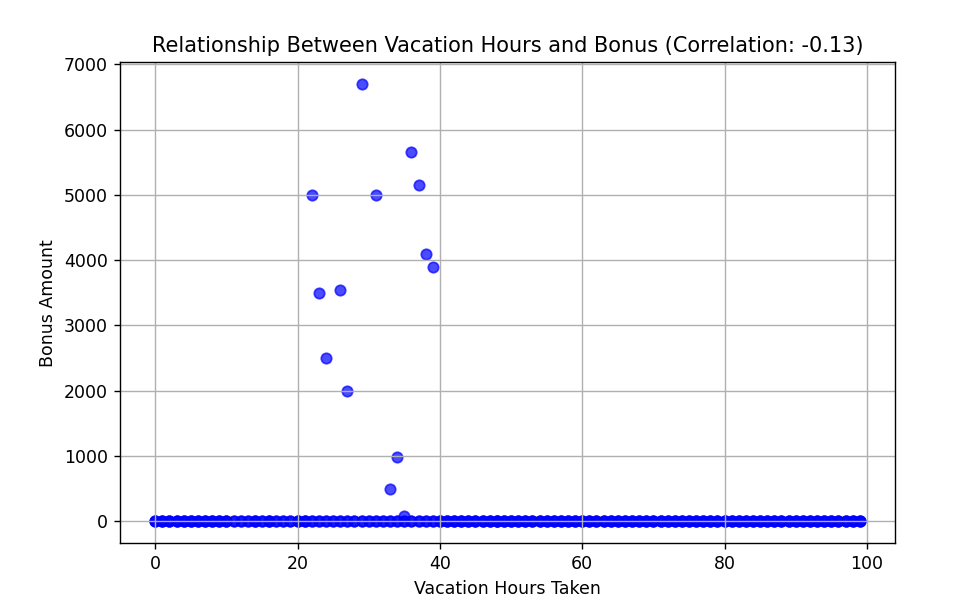
The analysis involved the following steps:

* Linking Employee table with SalesPerson on BusinessEntitityID.
* Extract Data using SQL Queries to retrieve a list of employees, their job titles, their vacation hours, and any bonuses they might have (or 0 if no bonus). The results are ordered by the amount of vacation hours, from the most to the least and joined together.
* Extracted the results into CSV file.
* Python was used to load, process the file using pandas and to perform a calculation.
* One scatterplot was created to visualize the results using matplotlib.

**Key Python steps**

* Load CSV file using pandas.
* Perform a check if there are any missing values.
* Calculate correlation between Vacation hours and bonus.
* Visualize the results using matplotlib scatterplot.

**Results & Insights**



*Correlation between Vacation Hours Taken and Bonus*

* The scatter plot shows the relationship between Vacation Hours and Bonus. There is no strong visible trend. This suggests that employees who take more vacation hours tend to have slightly lower bonuses, but the relationship is weak (weak negative correlation).
* We can also see that it may be a slight connection for more frequent bonuses awarded when annual leave hours were in range from 20 to 40 hours rather than others. Another possible factors could be job performance, role and type, gender, economic factors, bonus timing.

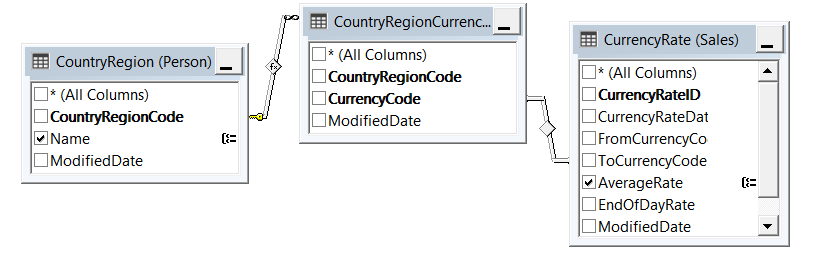
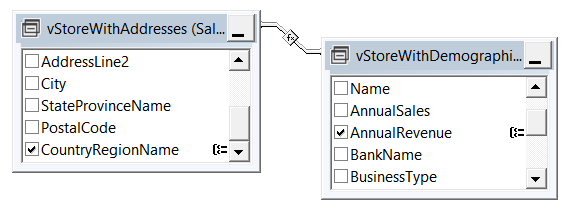
**Conclusion**

This analysis presents that there is a little relationship between Vacation Hours and Bonus but not significant enough to suggest a strong or meaningful connection. There could be other factors not considered for the bonuses rather than just annual leave such as job performance, role and type, gender, economic factors, bonus timing.

1. What is the relationship between Country and Revenue?

**Introduction**

This query analyses revenue across different countries using data extracted from a relational database via SQL. Python was then used for calculations and visualizations.



Schema of tables used

**Methodology**

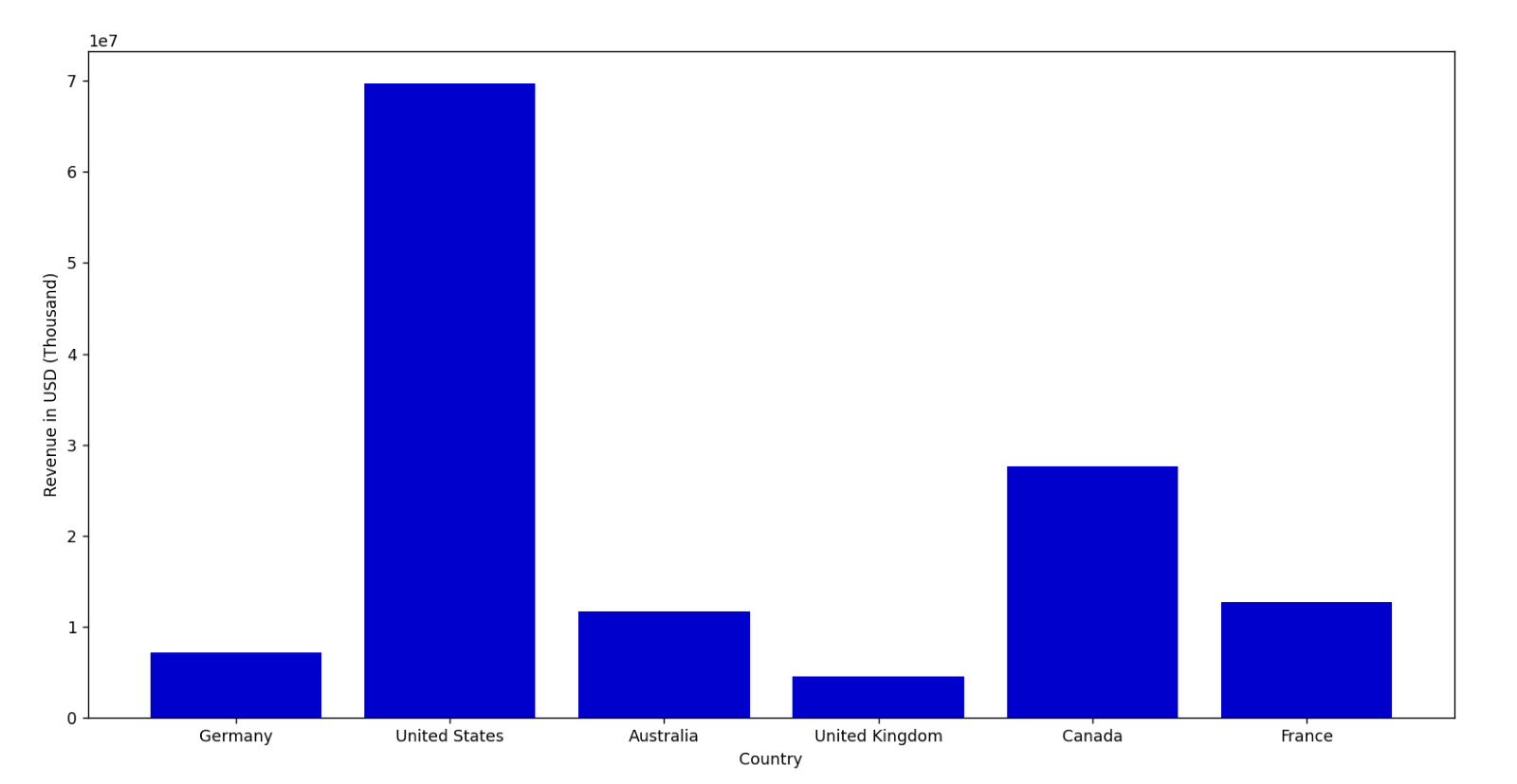
The analysis involved the following steps:

* Linking revenue data from the StoreWithDemographics view table with country data from the StoreWithAddresses view table to create a view table (CountryRevenue).
* Assuming revenues are recorded in different currencies, an additional relationship was established between three tables to retrieve the exchange rates data, to convert all revenue values to USD.
* The CountryRegionCurrency table links currency codes to country codes. Since the revenue table contains country names, the CountryRegion table was used to match country codes with their respective names.
* Next, the CurrencyRate table was related to the previous tables to retrieve currency exchange rate data, creating a view (CountryCurrencyRate).
* Finally, the CountryRevenue and CountryCurrencyRate views were combined to present country names, revenue, and currency rates in a single table, which was then exported as a CSV file for further analysis in Python.

**Key Python Steps**

* Read CSV file using Pandas package from Python library.
* Add a new column to the dataset that calculates revenue in USD by multiplying the original revenue by the corresponding exchange rate.
* Visualize and compare revenue across countries using Matplotlib.

**Results & Insights:**

*Country and Revenue Analysis*

* From the bar chart, the revenue for US is significantly higher than other countries. A key factor influencing this is the number of stores, with the US having the highest store count.
* When excluding US (with more than 400 stores) and Canda (with more than 100 stores), the remaining countries have almost a similar number of stores (about 40 store), However, their total revenues still vary, further analysis of other factors would help (analysis of some factors is demonstrated in the last two queries).

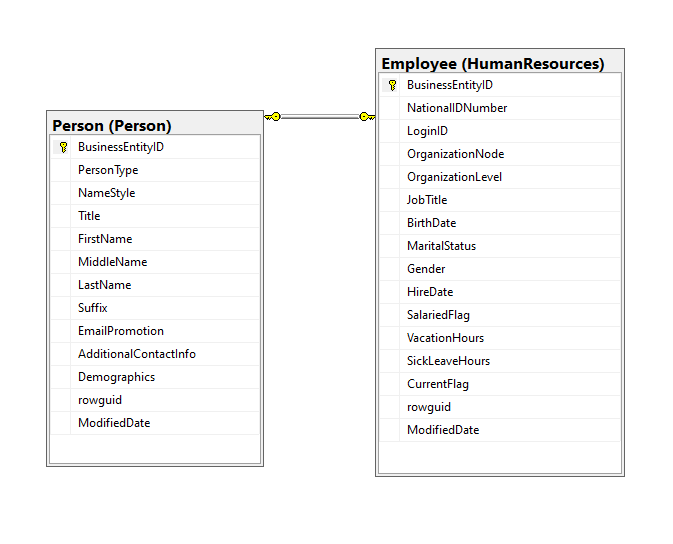
**Conclusion**

This analysis highlights the relationship between country and revenue, revealing that the store count strongly impacts revenue. However, among countries with similar store counts, revenue still varies, suggesting additional influencing factors. Further exploration of these factors could provide deeper insights into revenue. Some will be explored further down the report i.e. relationship between store size, number of employees and revenue.

1. What is the relationship between sick leave and Job Title?

**Introduction**

This report analyses the relationship between sick leave hours and job title/person type. The data was extracted from a relational database using SQL queries and then processed in Python for further calculation and visualization. The insights will provide the relationship between person type and sick leave/ job title and sick leave and presents which job titles have the highest sick leave entries. This analysis will also check the gender comparison for sick leave and job title as a result of further analysis.



Schema of tables used

**Methodology**

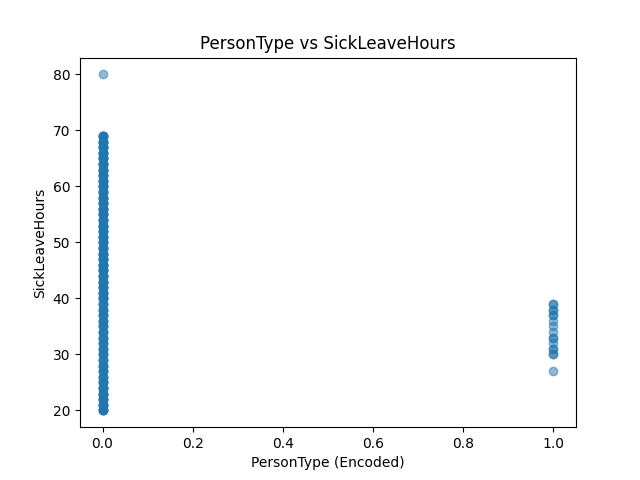
The analysis involved the following steps:

* Extract Data using SQL Queries
* Used SQL to join tables Person.PersonType and HumanResources.Employee to get the relevant information for calculations.
* In the query, I filtered the data to not include null values for SickLeaveHours to get a better representation of the data assuming that those who have null values for SickLeaveHours in HumanResources.Employee are not considered employees.
* Export SQL query results as CSV files and process it in Python.
* Python was used to load and process these files using Pandas.
* Visualize Data using Matplotlib and Seaborn
* Two scatter plots to check correlation between PersonType and SickLeaveHours AND JobTitle and SickLeave hours.
* Two bar charts to get insight on which JobTitle had the most SickLeaveHours and to further investigate other factors like Gender comparison.

**Key Python Steps**

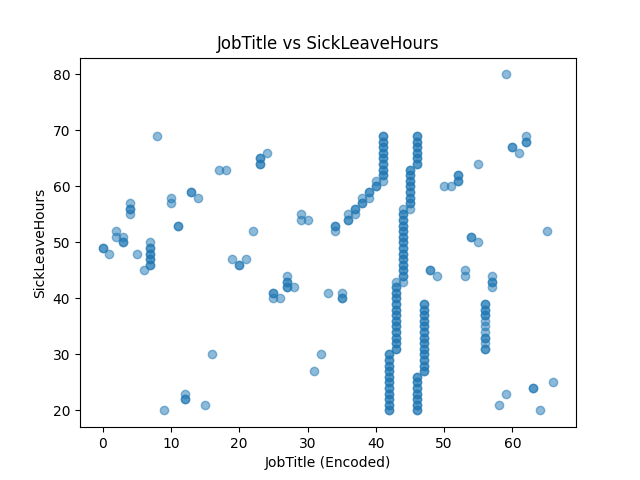
* Load data from CSV files using pandas.
* Using cat.codes to convert JobTitle and PersonType as integer so that it can be used for correlation calculations with SickLeaveHours.
* Check if there is a relationship by using Pearson’s correlation.
* Sort data using sort\_values for SickLeaveHours so that I get the values in order.
* Use matplotlib and seaborn to create bar charts and scatterplot.
* Highlight the highest sick leave entries with corresponding job title.

**Results & Insights**



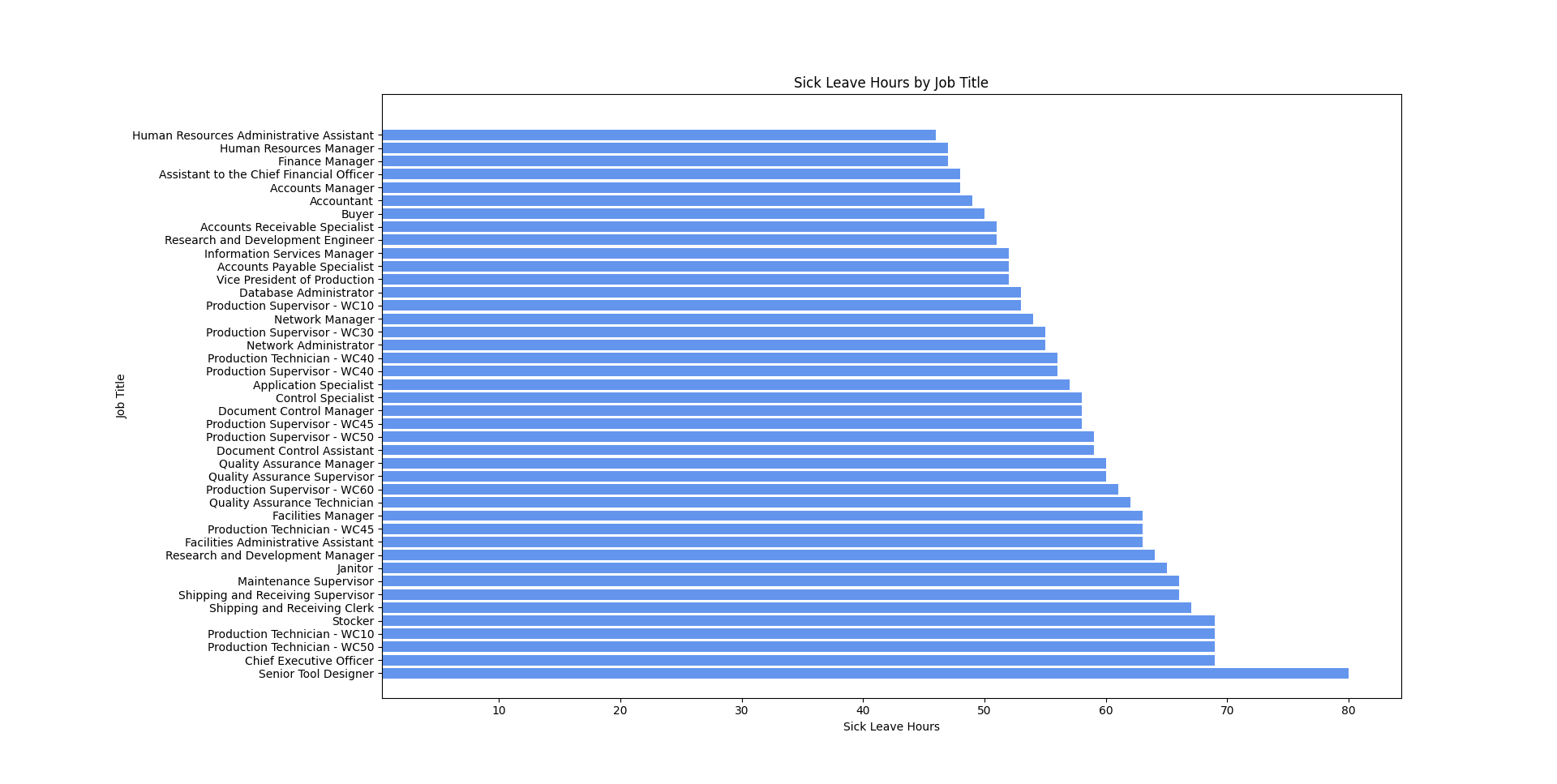
*Correlation between PersonType and SickLeaveHours*

* Out of 290 employees, there are 274 PersonType “EM” and 17 PersonType which are not “EM”
* The correlation between PersonType and SickLeaveHours is weak negative relationship (Correlation = -0.19236) as shown in the scatterplot.
* Which means that there could be other factors that can affect the relationship



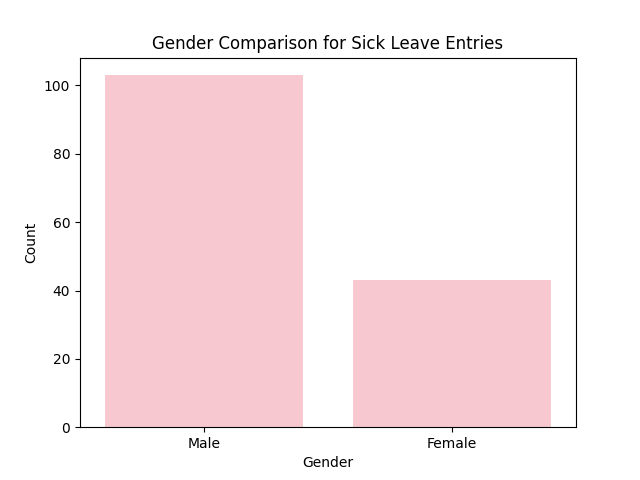
*Correlation between JobTitle and SickLeaveHours*

* Since there is no relationship between PersonType and SickLeaveHours, I investigated the relationship between JobTitle and SickLeaveHours instead.
* Analysis still shows weak negative relationship with correlation being -0.1289721.
* Which means that there could be other factors that needs to be further investigated i.e. Gender or OrganisationalLevel for each JobTitle



*Highest Sick Leave Entries Analysis*

* Highest entry for Sick Leave Hours is the Senior Tool Designer with 80 hours of Sick Leave
* This analysis shows those entries that went above the average/normal sick leave hours (45.3 hours) and grouped by their Job Title.
* There were 146 entries in total for those that went above the average sick leave hours.



*Gender Analysis for Sick Leave Entries Above Average*

* Just an additional investigation of other factors that can be highlighted on the relationship between JobTitle and SickLeaveHours.
* Insight shows that amongst 146 entries for above average sick leave, there were 103 males and 43 females.
* Results shows that Male Employees are twice more likely to take sick leave more often than the Female Employees.

**Conclusion**

Overall, this analysis presents that there is a negative relationship or the relationship between PersonType/JobTitle and SickLeaveHours is so weak that there is no significant connection. However, it identified the highest entries for sick leave and highlights job titles who had the highest entries.

These insights can support further investigations on sick leave and job titles and why others have more entries than the rest but also helps the business to realise other factors that may be related to sick leave hours i.e. workload of each department/employee, time management, employee’s wellbeing and more. All of which, can help the Business to provide better work-life balance, improve benefit and rewards policy, retain current employees, leverage traction and apply better work performance and productivity.

1. What is the relationship between store trading duration and revenue?

**Introduction**

In this analysis, we explore the relationship between store trading duration and revenue to understand how a store's age might influence its annual revenue. Using a dataset from AdventureWorks, we calculate the duration of each store from the year it was opened until the current year, then analyze whether older stores tend to generate higher revenues. We apply correlation analysis and visualization techniques to investigate this relationship and derive meaningful insights.

**Methodology**

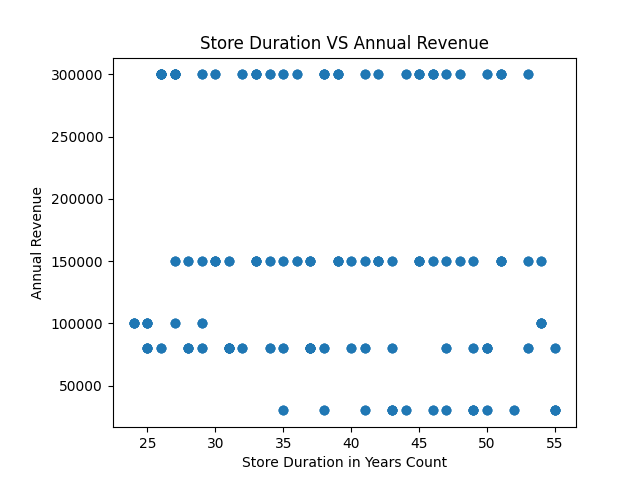
The analysis involved the following steps:

* Extract Data using SQL Queries
* The data was sourced from the StoreDemographicsTable.csv file, which contains information on the stores, including the year they were opened and their annual revenue.
* The dataset was loaded into Python using pandas.
* Created a new column, Store\_duration, to calculate the duration of each store by subtracting the store's opening year (YearOpened) from the current year.
* Used Pearson correlation to calculate the correlation between the Store\_duration and AnnualRevenue to determine if there is any linear relationship between the age of the store and its revenue.
* The store with the longest trading duration was identified to understand which store has the longest history in the dataset.
* The store with the highest annual revenue was also identified to compare how revenue relates to store duration.
* Grouped store duration to see pattern of the annual revenues.

**Key Python Steps**

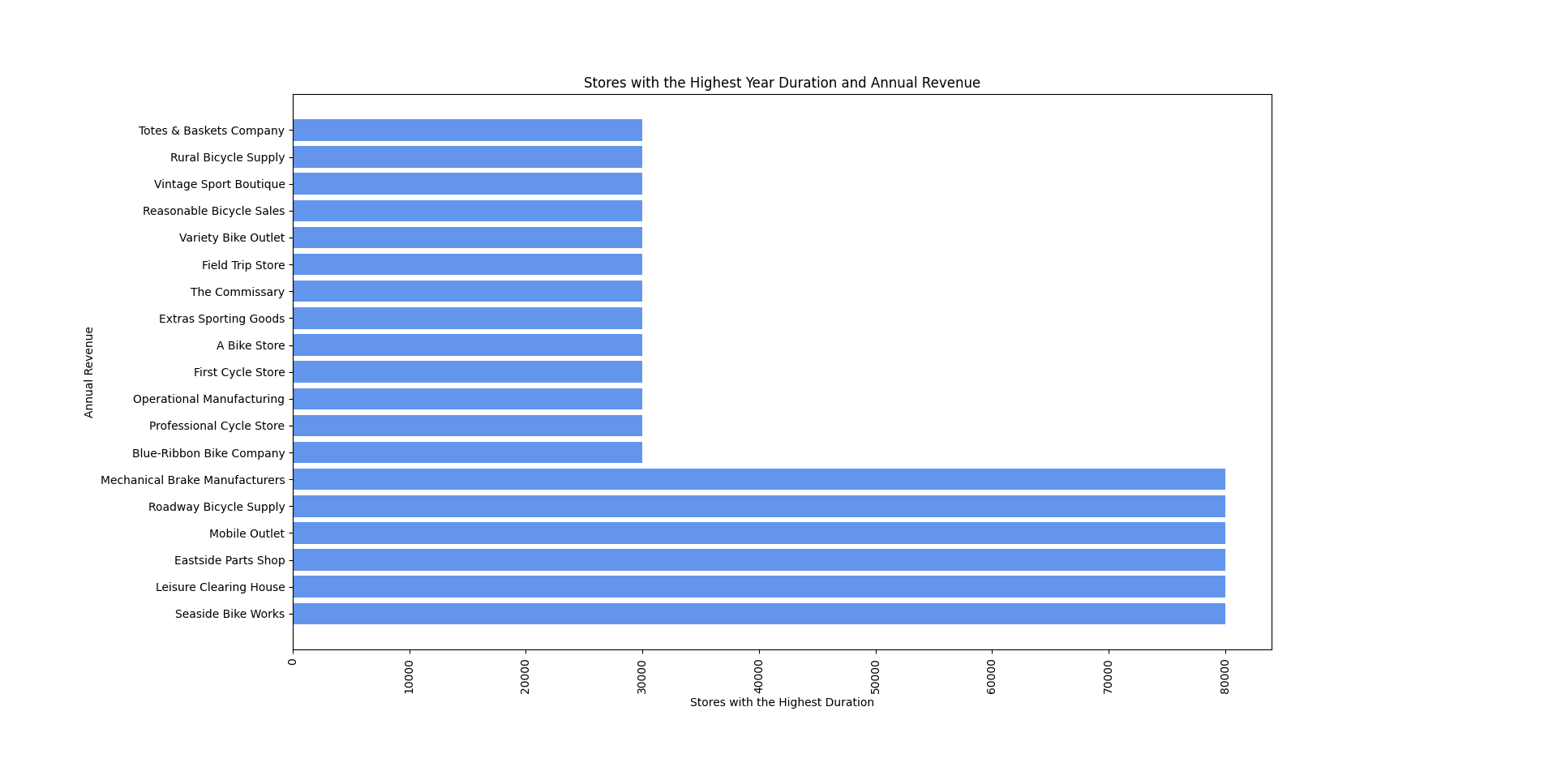
* Pearson correlation method to find any relationships
* Matplotlib and seaborn for visualisation
* Scatter plots were used to visually explore the relationship between store duration and annual revenue.
* A bar plot visualized the store with the longest duration and its annual revenue.
* A line plot was created to show the trends in annual revenue across different store durations.

**Results & Insights**

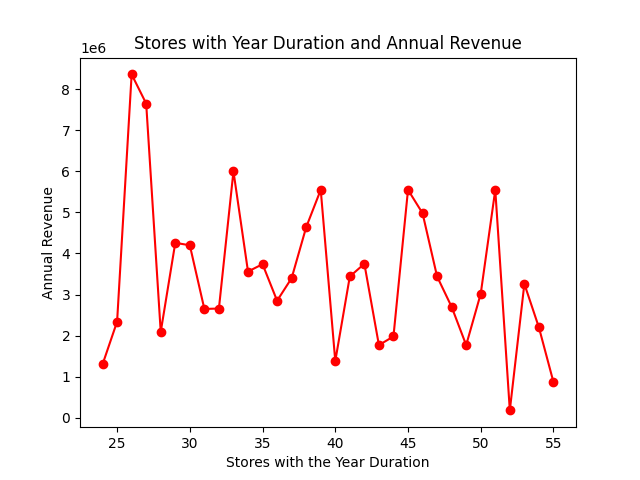


*Correlation between Store Duration and Annual Revenue*

* The relationship between the store duration and annual revenue is negative and weak, indicating that older stores tend to have lower revenues, at least based on this dataset with correlation of –0.133
* However, this might be due to factors not accounted for in this analysis i.e. store location, store size, number of employees which is later explored.

*Stores with the Highest Store Duration and their Annual Revenue*

* All these stores opened in 1970 with 55 years of trading duration (the highest trading duration) and its revenue was visualized using a bar plot for better clarity. Interestingly, some of the stores with longer durations do not necessarily show high revenue which can be further justified below with the line plot.



*Total Stores Year Duration and Annual Revenue*

* The line plot showed that stores with durations of 25-26 years appear to generate the highest revenues and the stores with durations of 50 years and above generate the lowest revenues.
* Additionally, there is a noticeable increase in revenue every 5-7 years, suggesting that stores might experience growth spurts at specific intervals so there could be other factors that can justify this i.e. they might have remodelled or modernised their stores every 5-7 years, but this will need further investigations.

**Conclusion**

In conclusion, the analysis of the relationship between store trading duration and revenue reveals that while there is a negative correlation, the relationship is not straightforward and suggests that other factors may need to be further explored. The revenue trends show that stores with durations of 25-26 years tend to have higher revenues, but this is not a guarantee for all stores.

This insight could be valuable for AdventureWorks to leverage the business decision-making and understand how store maturity impacts performance and planning for future business strategies. Further analysis with additional variables e.g. Store size, number of employees are explored below.

1. What is the relationship between the size of the stores, number of employees and revenue?

**Introduction**

This report provides an analysis if there is a relationship between the size of the stores, number of employees and revenue. The data was extracted from a relational database using SQL queries and then processed in Python for further calculation and visualization. The findings provide insights into establishing if there are any connections between.

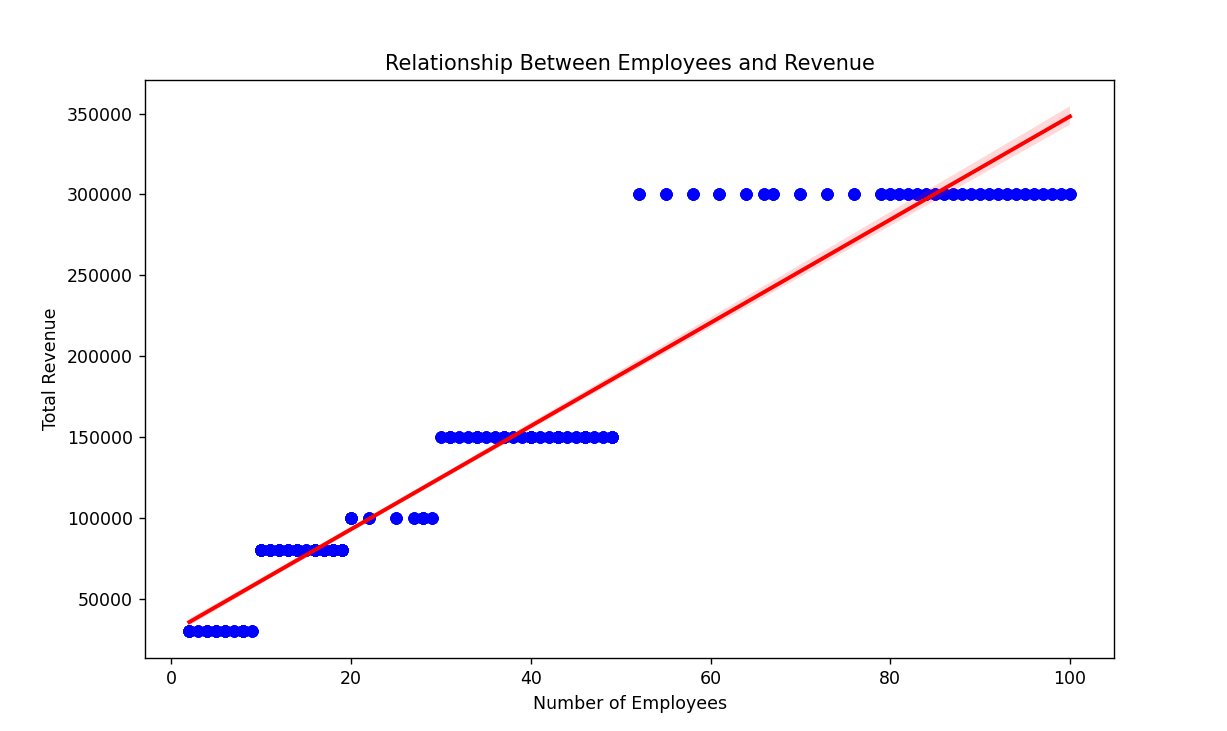
**Methodology**

The analysis involved the following steps:

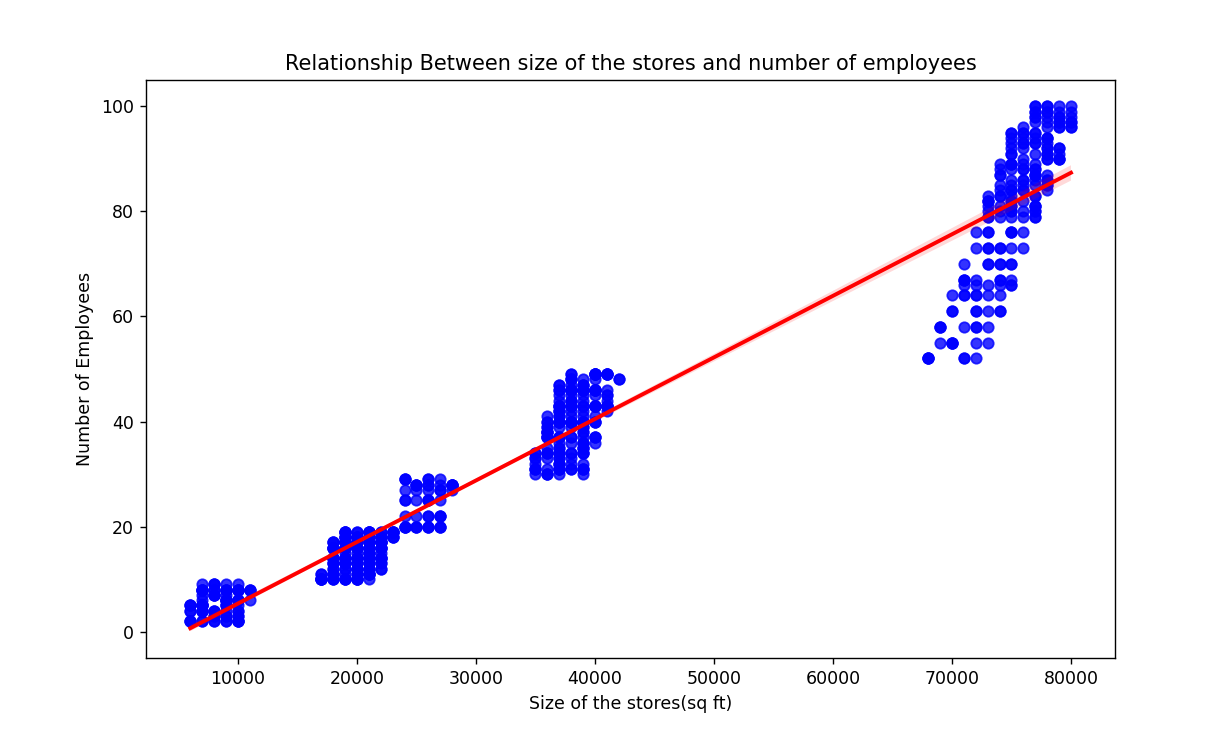
* Extract Data table StoreWithDemographics using SQL Queries to retrieve a table with size of stores (sq. ft), number of employees and revenue.
* Extracted the results into CSV file.
* Python was used to load, process the file using pandas and to perform a calculation.
* Three scatterplot was created to visualize the results using matplotlib and one pair plot for comparisons.

**Key Python Steps**

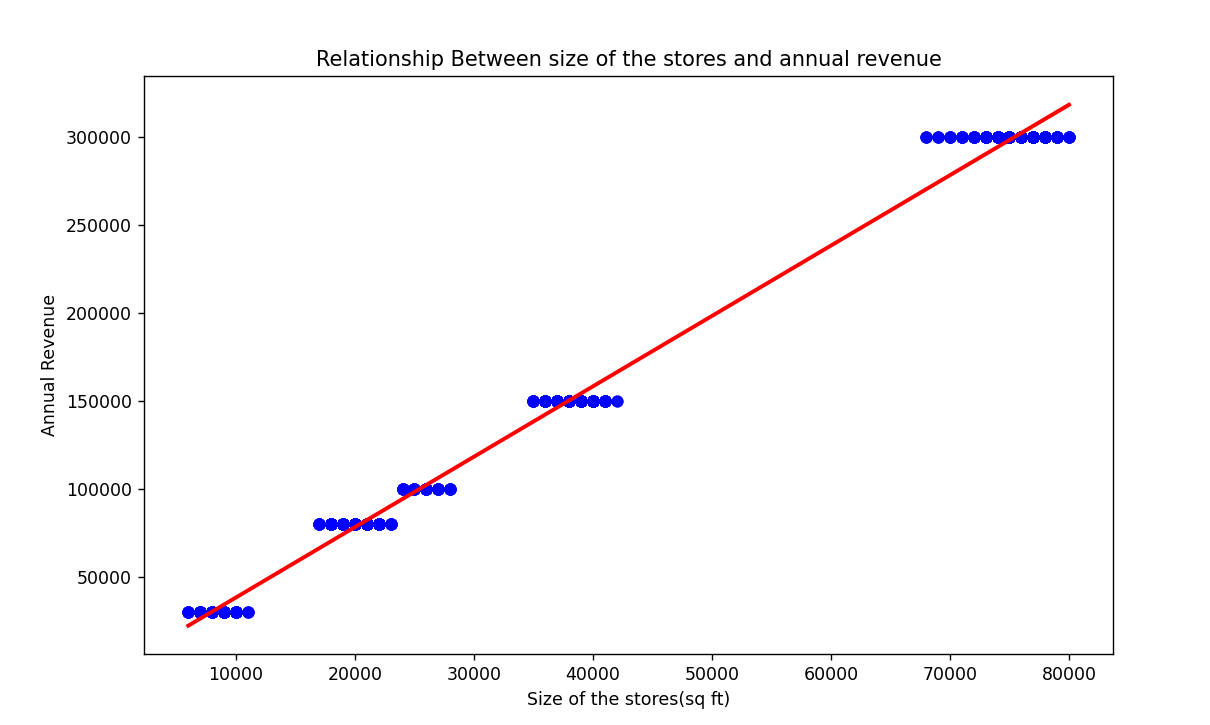
* Load CSV file using pandas.
* Print first rows.
* Calculate correlations between size of stores, number of employees.
* Visualize the results using matplotlib scatterplot and seaborn pairplot.

**Results & Insights***Relationship between Employees and Revenue*

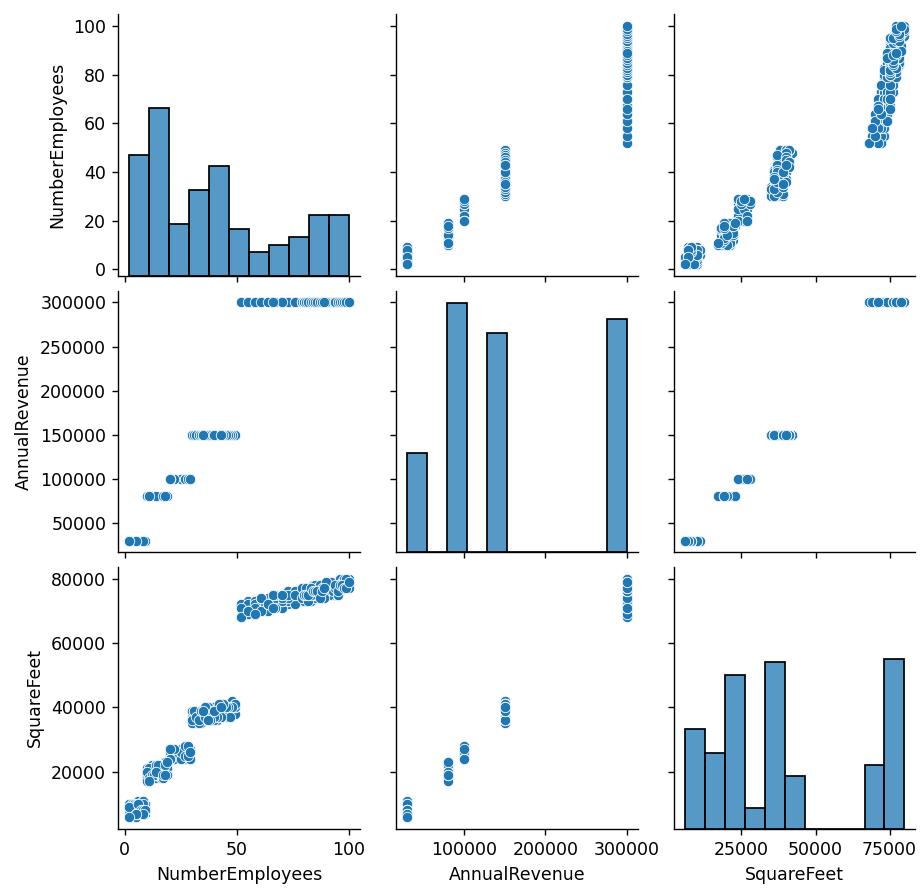
Correlation between number of employees and revenue: 0.9581073442505238 shows there is strong correlation between revenue and number of employees.

*Relationship Between size of the stores and number of employees*

Correlation between the size of the stores and number of employees: 0.9717822547114197 shows strong connection between size of the store and number of employees.

*Relationship between size of the stores and annual revenue*

Correlation between the size of the stores and revenue: 0.99653089194352 shows strong connection between revenue and size of the stores.

**Conclusion**

Based on the insights, we can conclude that there is a strong relationship between store size, number of employees, and revenue. The high correlation between the number of employees and revenue (0.96) suggests that as businesses expand their workforce, their revenue tends to increase. Similarly, the strong correlation between store size and the number of employees (0.97) indicates that larger stores require more staff. Finally, the exceptionally strong correlation between store size and revenue (0.99) highlights that store size is a factor influencing revenue. These findings suggest that expanding store size and workforce can significantly contribute to higher revenue growth.

**Overall Conclusion**

* Our analysis of the database provided valuable insights into the relationships between various business variables.
* We were able to uncover key trends, patterns, and potential areas for improvement.
* The visualizations helped to clarify these relationships and highlight opportunities that could drive better decision-making and business growth.
* The analysis can help further investigate other factors to optimize operations and boost performance within the company

**Extra Findings: Alternative way to access SQL query in Python using sqlalchemy for Q1**

A screen shot of a computer program

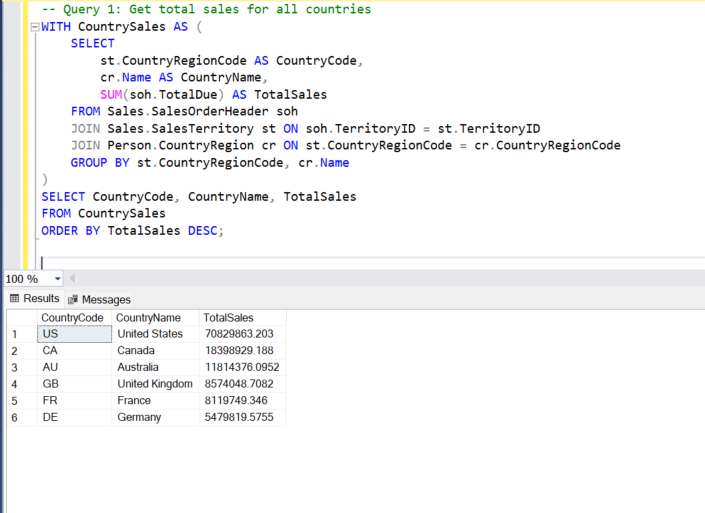
AI-generated content may be incorrect.A screenshot of a computer program

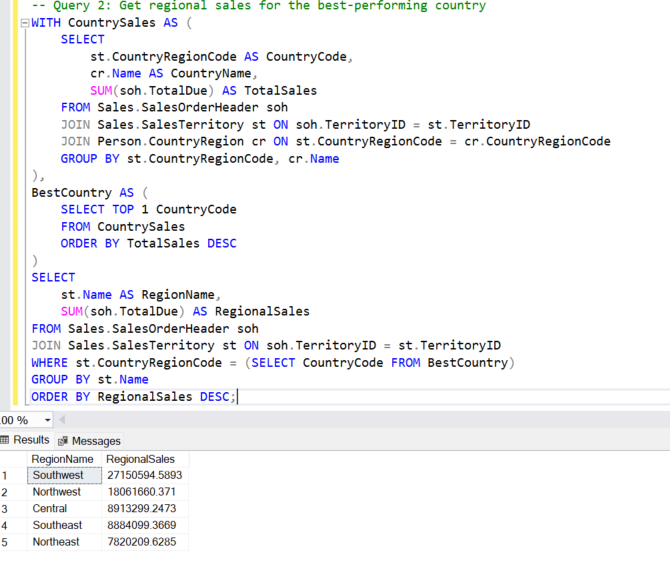
AI-generated content may be incorrect.A computer screen shot of a code

AI-generated content may be incorrect.

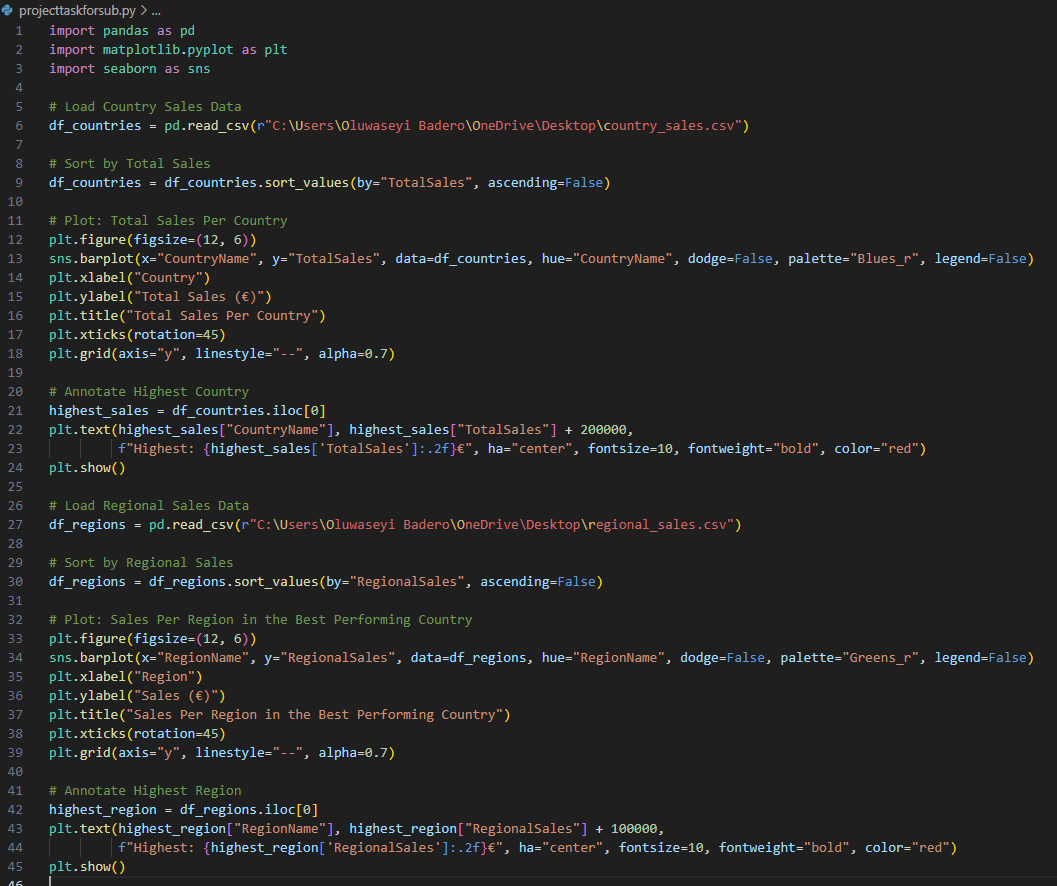
**Appendix:**

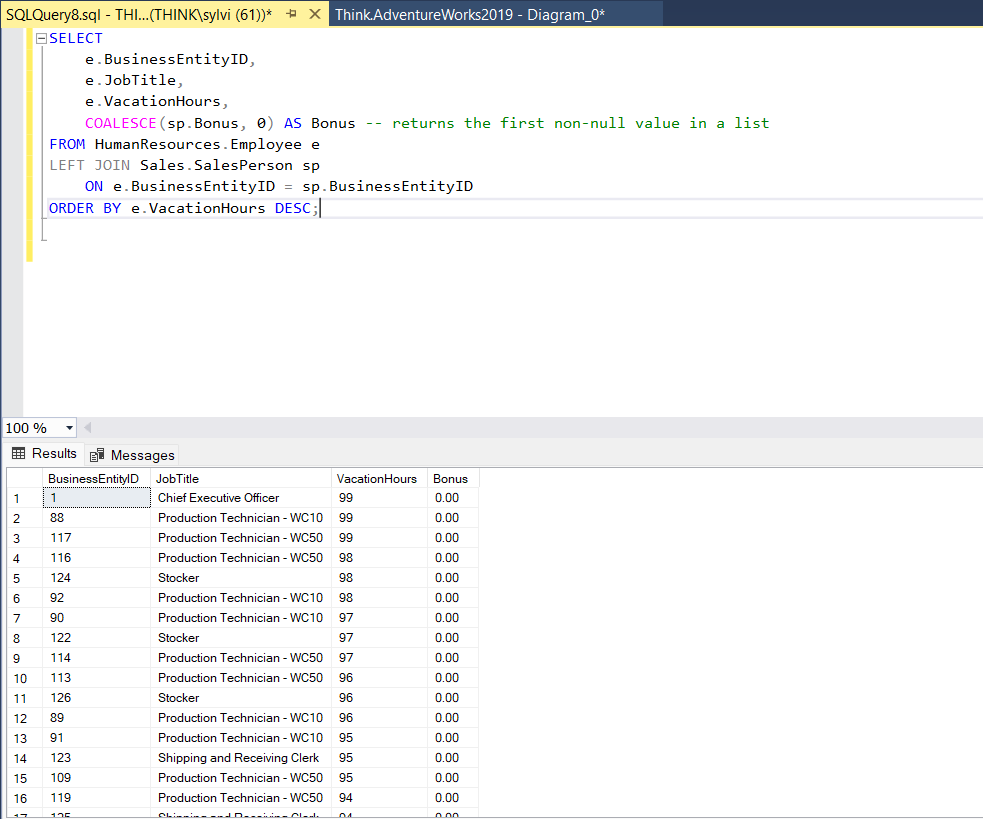
SQL CODE for Q1:

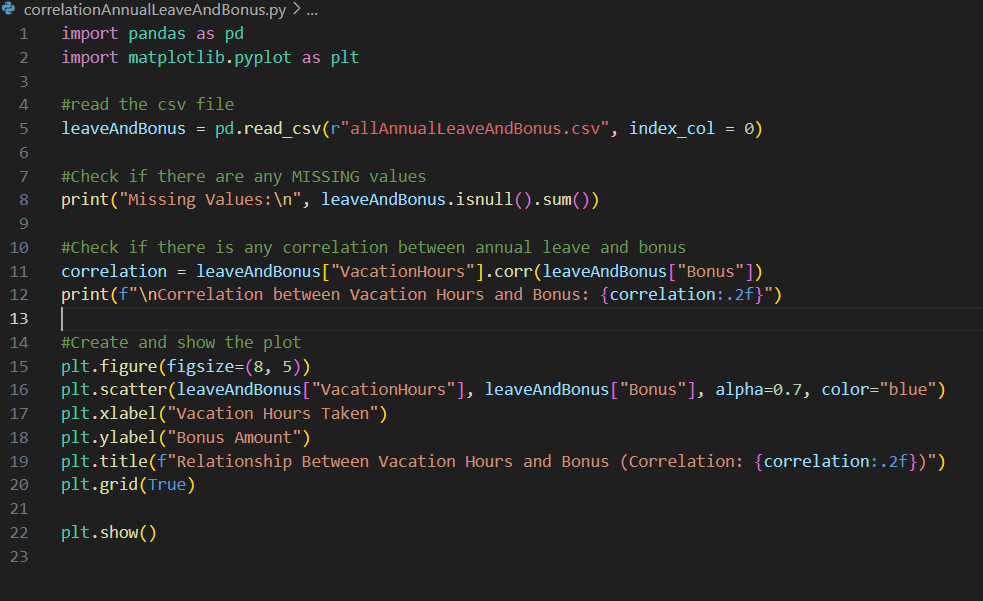




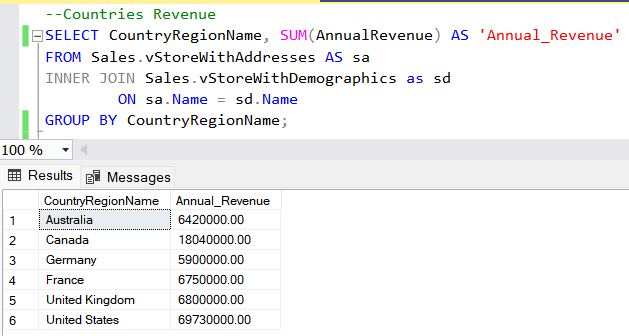
PYTHON CODE for Q1

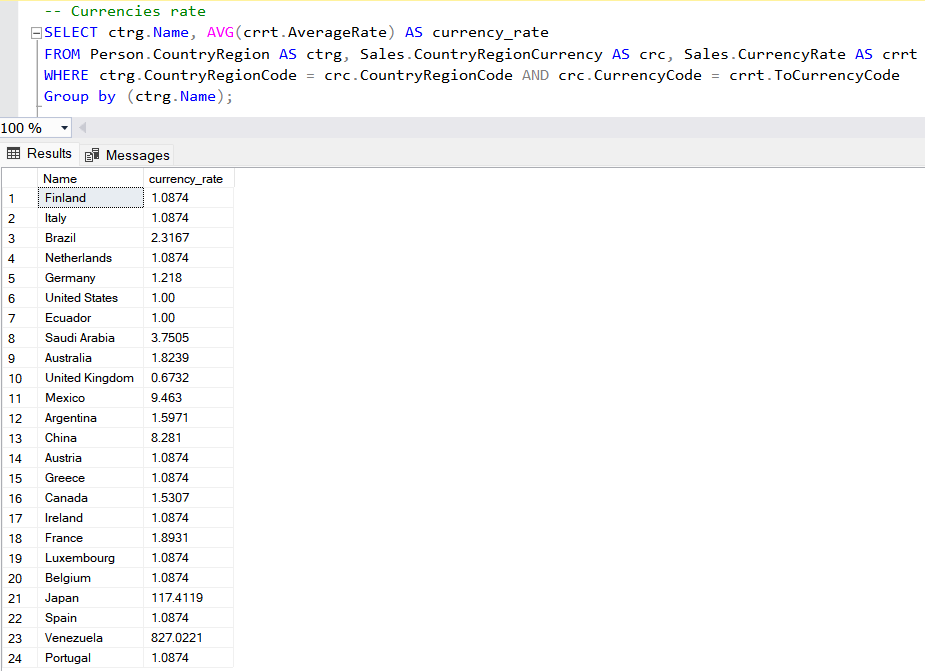


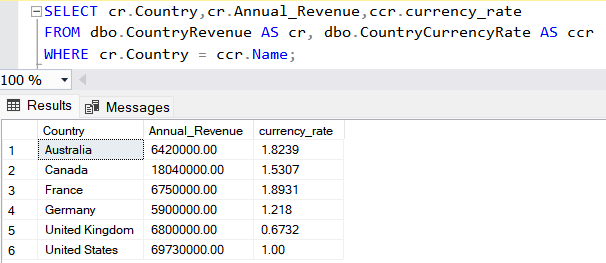
SQL CODE for Q2:  


PYTHON CODE for Q2:  


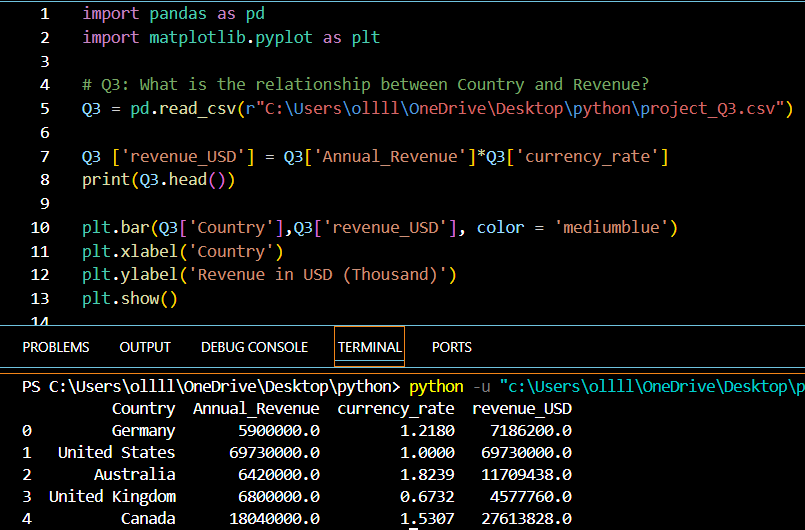
SQL CODE for Q3:



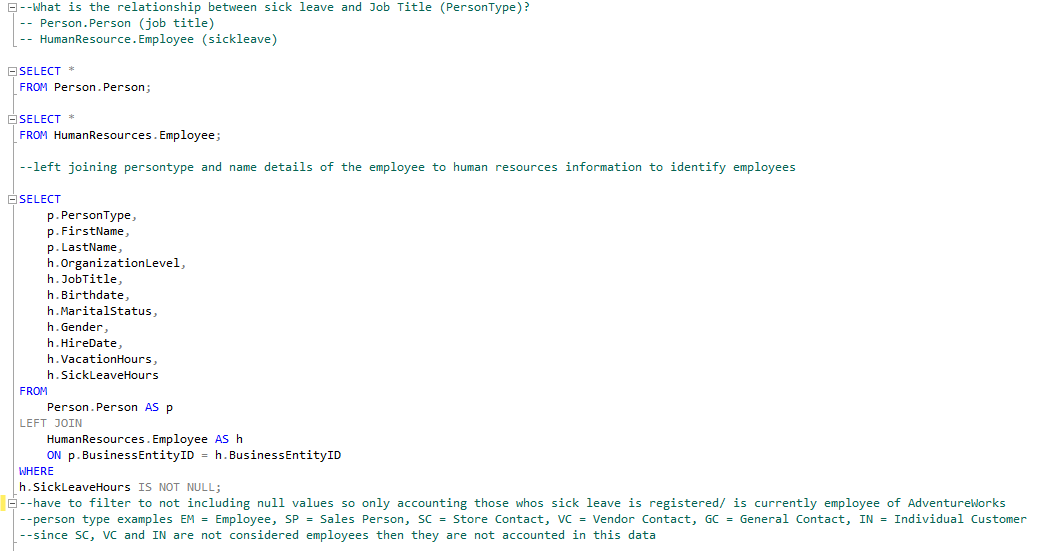




PYTHON CODE for Q3:



SQLCODE for Q4:



PYTHON CODE for Q4:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv(r"C:\Users\vinam\Desktop\Project Adventure Works 2019\SickLEave&JobTitleTable.csv")

print(df.head())

# I want to check the following:

# 1. does sick leave hours increase the higher the job title?

# 2. does males take more sick leave than females?

# 3. which jobtitles has more sick leave hours?

# 4. who has taken the most sick leave hours?

# 5. which persontype has the most sick leave?

#total employees using len() since its a string

total\_employees = len(df["PersonType"])

#total sick leave hours using sum() since its an int

sickleave\_total = df["SickLeaveHours"].sum()

#average sick leave amongst employees

avg\_sickleave = sickleave\_total/total\_employees

#display result

print("Average sick leave hours: ",avg\_sickleave)

print("Total Employees: ",total\_employees)

#displaying the number of employees and their person type

em\_employees = df[df["PersonType"]== "EM"]

other\_employees = df[df["PersonType"]!= "EM"]

print("PersonType that is EM totals: ",len(em\_employees))

print("PersonType that is not EM totals: ",len(other\_employees))

# x = df["SickLeaveHours"]

# y = df["PersonType"]

# pearson = x.corr(y)

# print(pearson)

# spearmans = x.corr(y, method="spearman")

# print(spearmans)

# kendalls = x.corr(y, method="kendall")

# print(kendalls)

#the above code does not work as person type is a string - needs to be converted to int/float so using cat.codes to asign an int to each person type

df["PersonTypeEncoded"] = df["PersonType"].astype("category").cat.codes

#calculate correlation between PersonType (encoded) and SickLeaveHours

correlation = df["PersonTypeEncoded"].corr(df["SickLeaveHours"])

#display the correlation result

print("Correlation Matrix:")

print(correlation)

#results shows that there is a weak negative relationship between the two variables as the results is close to zero

#there could be other factors or non-linear relationship but in terms of correlation, it's weak

#visualising the data points on a scatter plot

plt.scatter(df["PersonTypeEncoded"], df["SickLeaveHours"], alpha=0.5)

plt.title("PersonType vs SickLeaveHours")

plt.xlabel("PersonType (Encoded)")

plt.ylabel("SickLeaveHours")

plt.show()

#still shows no linear relationship and it showcases that there are only 2 types of persontype which is why there is no correlation

#checking how many entries are above the average with corresponding job title instead

#filtering to group it by job title and have sick leave hours equal or above average

count\_sickleave = df.groupby(["JobTitle"]) and df[df["SickLeaveHours"]>=avg\_sickleave]

print("Total sick leave entries above average is:")

#using len() to get total entries for sick leave hours above average for job title

print(len(count\_sickleave))

#sorting out the result in order with ascending values of sick leave hours with 80 being the highest

sickleave\_sorted = count\_sickleave.sort\_values(by= "SickLeaveHours", ascending=False)

#displaying result in a barchart for better readability

plt.barh(sickleave\_sorted["JobTitle"],sickleave\_sorted["SickLeaveHours"], color = "cornflowerblue")

plt.title('Sick Leave Hours by Job Title')

plt.ylabel('Job Title')

plt.xlabel('Sick Leave Hours')

plt.show()

print(sickleave\_sorted)

#using the same methods as persontype above, I want to see if there is a correlation between job title and sick leave

df["JobTitleEncoded"] = df["JobTitle"].astype("category").cat.codes

#calculate correlation between JobTitle (encoded) and SickLeaveHours

correlation\_j = df["JobTitleEncoded"].corr(df["SickLeaveHours"])

#display the correlation result

print("Correlation Matrix:")

print(correlation\_j)

#results shows that there is still a weak negative relationship between the two variables as the results is close to zero

#there could be other factors or non-linear relationship but in terms of correlation, it's weak

#visualising the data points on a scatter plot

plt.scatter(df["JobTitleEncoded"], df["SickLeaveHours"], alpha=0.5)

plt.title("JobTitle vs SickLeaveHours")

plt.xlabel("JobTitle (Encoded)")

plt.ylabel("SickLeaveHours")

plt.show()

#curious on the gender comparison for this as well

#gender analysis for above average sick leave entries

male\_count = count\_sickleave["Gender"].value\_counts()["M"]

female\_count = count\_sickleave["Gender"].value\_counts()["F"]

print("Male count for sick leave entries:",male\_count)

print("Female count for sick leave entries:",female\_count)

#display analysis as bar chart

#creating a new DataFrame for plotting

gender\_data = pd.DataFrame({

"Gender": ["Male", "Female"],

"Count": [male\_count, female\_count]

})

#create a bar plot using seaborn

sns.barplot(x="Gender", y="Count", data=gender\_data, color = "pink")

#add titles and labels

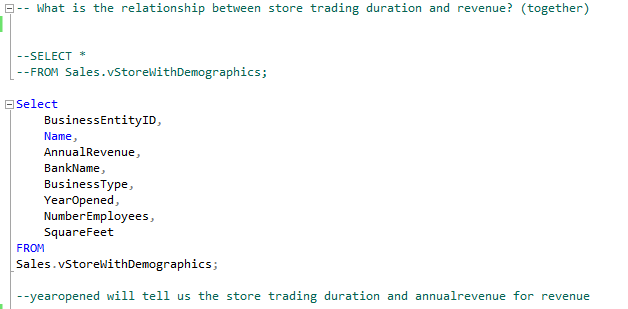
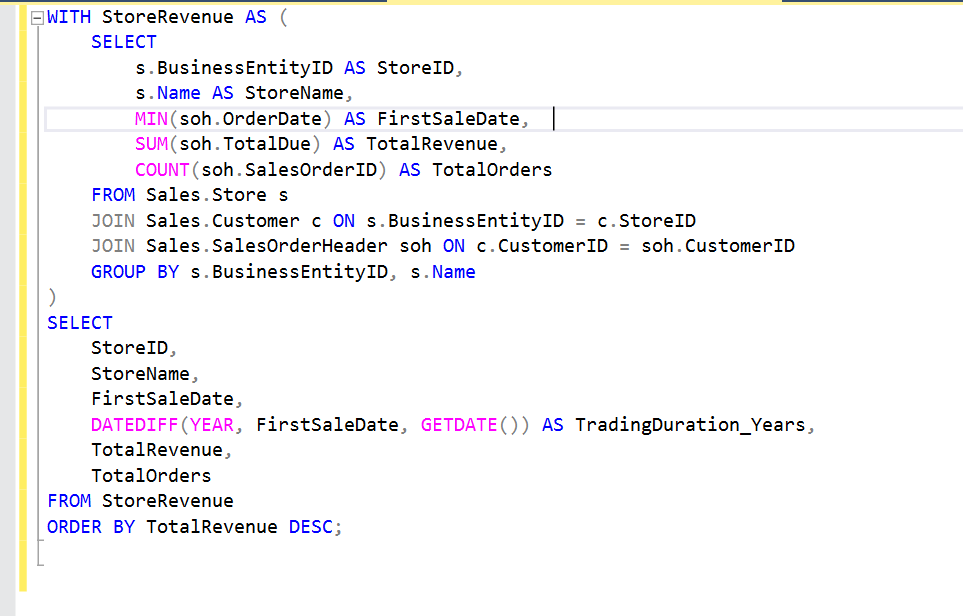
plt.title("Gender Comparison for Sick Leave Entries")

plt.xlabel("Gender")

plt.ylabel("Count")

plt.show()

SQL CODE for Q5: (two codes from Olu and Vina to check if there’s any missing information and check for comparisons – both gave similar answers)



PYTHON CODE for Q5:

#What is the relationship between store trading duration and revenue? (together)

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from datetime import datetime

df = pd.read\_csv(r"C:\Users\vinam\Desktop\Project Adventure Works 2019\StoreDemographicsTable.csv")

print(df.head())

# I want to check the following:

# 1. is there a relationship with store trading duration and revenue via pearson correlation?

# 2. which store has the longest trading duration

# 3. which store has the highest annual revenue

#calculate the store duration first i.e.current date - year opened

#function to calculate store duration

def year\_duration(year\_opened):

#get the current year

current\_year = datetime.now().year

#calculate the duration by subtracting the store opening year from the current year

duration = current\_year - year\_opened

return duration

#get the year in data

year\_opened = df["YearOpened"]

#replace with the actual year the store opened

print(f"Store duration: {year\_duration(year\_opened)} years")

#now adding result to a new column for better readability and easier calculations

df["Store\_duration"] = df["YearOpened"].apply(year\_duration)

#checking result

print(df)

#now I want to check whether high store duration equals to high annual revenue

#calculate correlation between Year Opened and Annual Revenue

correlation = df["Store\_duration"].corr(df["AnnualRevenue"])

#display the correlation result

print("Correlation Matrix:")

print(correlation)

#check if the correlation is positive or negative and analyze the relationship

if correlation > 0:

print("There is a positive correlation between Store Duration and Annual Revenue.")

elif correlation < 0:

print("There is a negative correlation between Store Duration and Annual Revenue.")

else:

print("There is no linear correlation between Store Duration and Annual Revenue.")

#results shows that there is a negative relationship between the two variables

#there could be other factors or non-linear relationship but in terms of correlation, it's negative

#visualising the data points on a scatter plot

plt.scatter(df["Store\_duration"], df["AnnualRevenue"], alpha=0.5)

plt.title("Store Duration VS Annual Revenue")

plt.xlabel("Store Duration in Years Count")

plt.ylabel("Annual Revenue")

plt.show()

#finding the longest store duration

longest\_duration = df["Store\_duration"].max()

#finding the row(s) with the longest store duration

longest\_store = df[df["Store\_duration"] == longest\_duration]

#showcasing other relevant details of the store with the longest duration

detail = longest\_store[["Name", "AnnualRevenue","Store\_duration","YearOpened"]]

sorted\_details = detail.sort\_values(by = "AnnualRevenue", ascending=False)

#displaying the results

print("Longest Duration:", longest\_duration)

#displaying results in a list

print("Details of the Store with Longest Duration:")

print(detail)

#displaying as a bar

plt.barh(sorted\_details["Name"],sorted\_details["AnnualRevenue"],color = "cornflowerblue")

plt.title("Stores with the Highest Year Duration and Annual Revenue")

plt.ylabel("Annual Revenue")

plt.xlabel("Stores with the Highest Duration")

plt.xticks(rotation=90)

plt.show()

#curious to see the overall graph for store duration and annual revenue

#using group by to group the years of duration and sum the annual revenue for each duration year

count\_store = df.groupby("Store\_duration")["AnnualRevenue"].sum()

#using line plot this time to get a clearer insight

#.index will show unique store duration values

#.values contains the summed annual revenue for each store duration

plt.plot(count\_store.index, count\_store.values, color = "red", marker = "o")

plt.title("Stores with Year Duration and Annual Revenue")

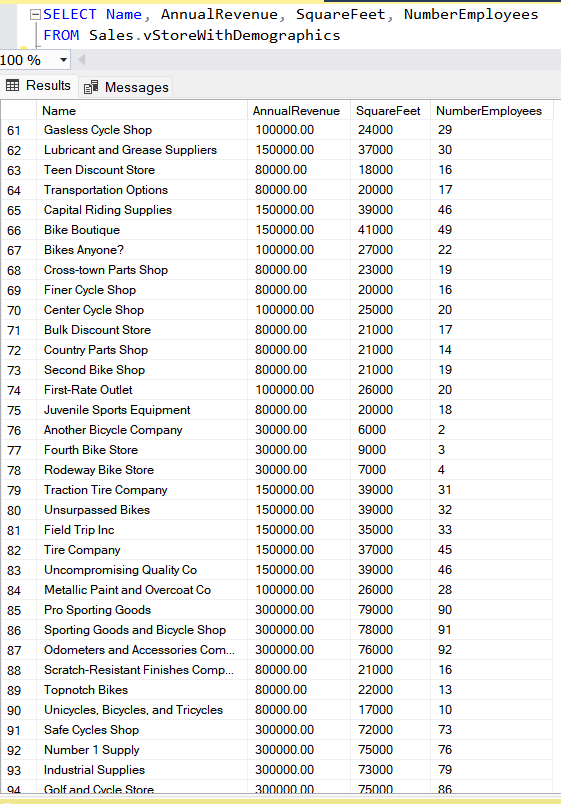
plt.ylabel("Annual Revenue")

plt.xlabel("Stores with the Year Duration")

plt.show()

#output shows that stores with store duration of 25/26 years have the highest revenues

#it also showed that every 5-7 years onwards, there is a bit of jump in revenues

SQL CODE for Q6:  


PYTHON CODE for Q6  
