Data Analysis Portfolio

Saher Aziz

Introduction

- This portfolio contains study cases from the Data Analytics program enrolled in Career Foundry educational platform for 2021.
- It consists of 5 projects through which I explored various techniques to analyze and visualize data using Excel, Tableau, SQL, and Python.
- Each project utilizes one of the tools listed.

GameCo

Analysis of Video Game Sales using Excel

- GameCo is a fictional video game seller present globally with its strongest markets in North America, Europe, and Japan. The goal of the analysis was to uncover insights from historical sales data to recommend changes in the marketing budget.
- I learned to conduct a complete analysis of the data set using the industry standard tool, Excel.
- The data set contains 16,600 observations with sales numbers for each game's title from 1980-2020, as well as the game's genre, platform, publisher, and publishing year.

GameCo

Exploring the data included:

- Understanding the data using filtering, sorting, grouping and summarizing functions, pivot tables, and charts to uncover first insights on the marketability of the games throughout the 40 years. Finding dirty data and cleaning helped with further analysis.
- Data Limitations that may have been caused by collection methods and bias.

Row Labels	Sum of Global_Sales
Action-Adventure	1989.62
Fighting	448.91
Misc	809.96
Platform	831.37
Puzzle	244.95
Racing	732.04
Role-Playing	927.37
Shooter	1037.37
Simulation	392.2
Sports	1330.93
Strategy	175.12
(blank)	0.36
Grand Total	8920.2

Fig. 1 Pivot table of Genre and Global Sales.

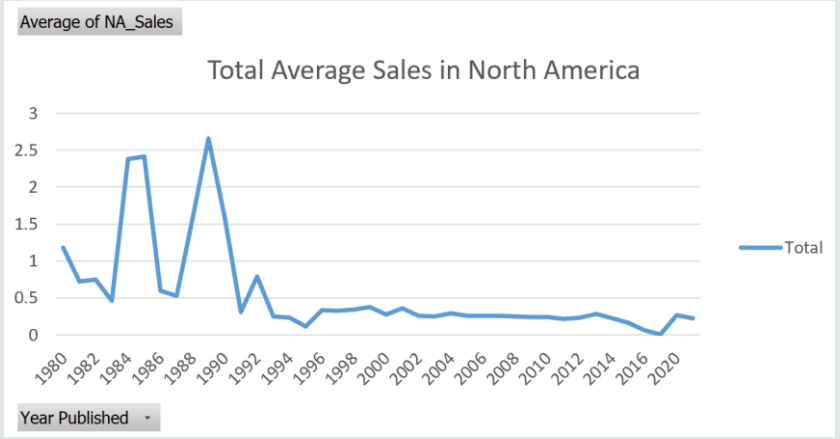


Fig. 2 Chart of Year and Average NA sales.

GameCo

- In the introduction to descriptive statistics, I learned the concepts such as measures of central tendency, distribution, spread, quartiles, or outliers that led me through the basics of exploratory data analysis.
- To read the results, I learned to prepare visualizations in form of histograms, box and whisker plots, and scatterplots. They are a great way to spot any unusual values and are easy to communicate to stakeholders during the process of data analysis.
- By exploring the sales in North America, I learned that the data is right-skewed caused by the low number of high sales which drives the average up and leaving the median to be lower.
- The relationship between the NA Sales and Global Sales have proved to be quite the discovery as North America has the highest shares in Global sales.

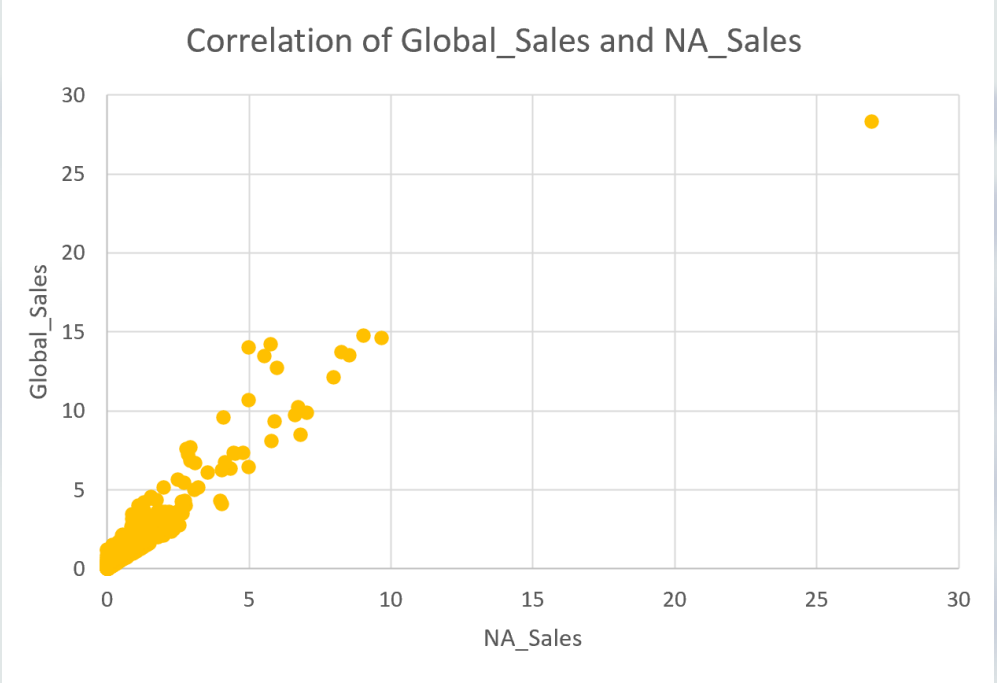


Fig. 3 Positive correlation between Global Sales and NA sales

GameCo

- In the final presentation, I presented my findings to the stakeholders and put them in the context of the goals and objectives of the analysis. According to the company, sales for the various geographic regions have stayed the same over time. I challenged the company's views and recommended changes in the marketing budget.
- I learned to combine the important descriptive statistics with functions and charts available in Excel. All of it assisted in presenting valuable insights for GameCo.
- View the entire presentation here.

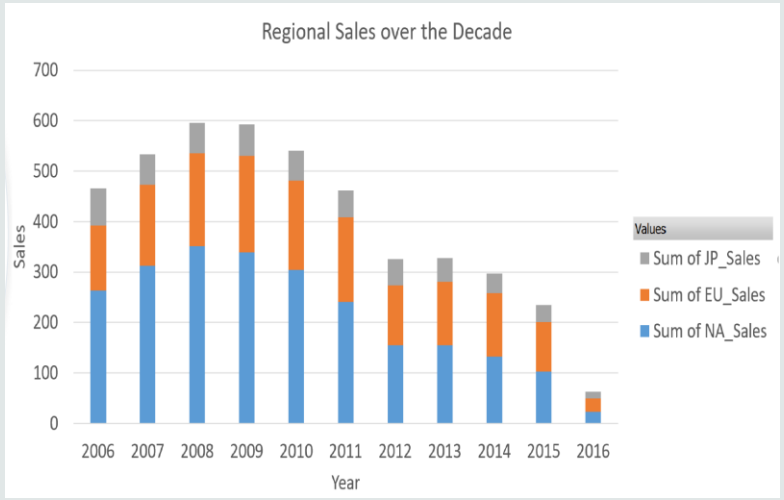


Fig. 4 Stacked Bar chart of regional sales over the Decade

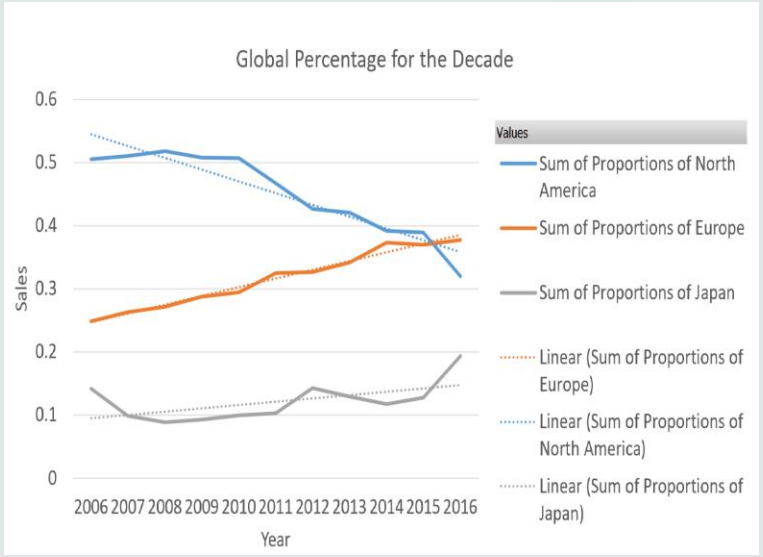
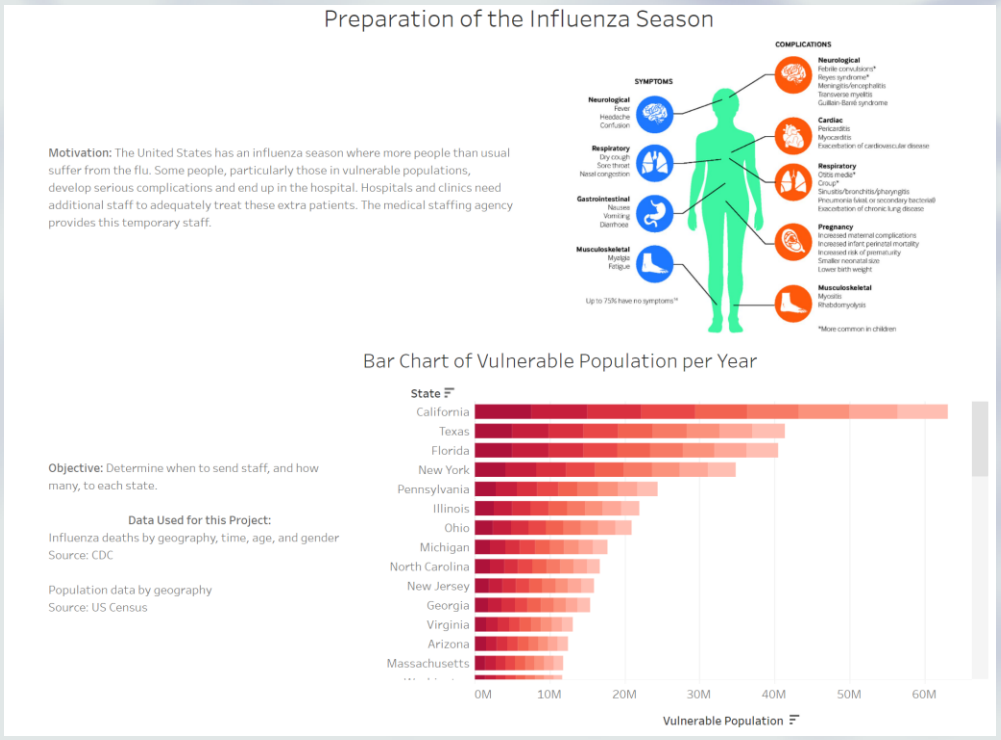


Fig. 5 Line chart with trend lines of the sales per region

Preparing for Influenza Season

Analysis of the Flu Season using Excel and Tableau

- The project was motivated by the fact that the United States experiences flu seasons when more people than usual suffer from the illness. It required to examine differences in staffing needs across each of the individual states.
- Expanded on analytical functions in Excel and work on visualizations using Tableau Public.
- The data about demography, flu deaths, flu shots, doctor's visits, and test results were collected in 5 separate sets.
- Sources of the data come from the CDC and US Census Bureau.



Preparing for Influenza Season

- Upon exploring the required documents, I learned to interpret the business requirements from a data perspective and translated the into questions that guided through the analysis.
- I constructed a project management plan and developed the hypothesis: If a state that a has a high population of vulnerable people, then influenza death rates increase.
- With this project, I worked closely with the Census and Flu Deaths sets in which I profiled the data and checked for accuracy and consistency.
- I needed to transform the data since there were two sets of data and integrate them to look for a correlation between selected variables.
- Finally, using the research hypothesis, I transformed it into a statistical hypothesis by creating a null hypothesis, calculating significance levels, and interpreting the p-value. In this study case, the null hypothesis was rejected and confirmed the research hypothesis.

Preparing for Influenza Season

- The second part of the project concentrated on the visualizations of the results in Tableau.
- I created comparison, temporal, spatial, and textual charts in which the scope of the project required looking for insights to prepare for the next season.
- I reported my findings using Tableau's dashboard, along with recommendations in a video presentation.

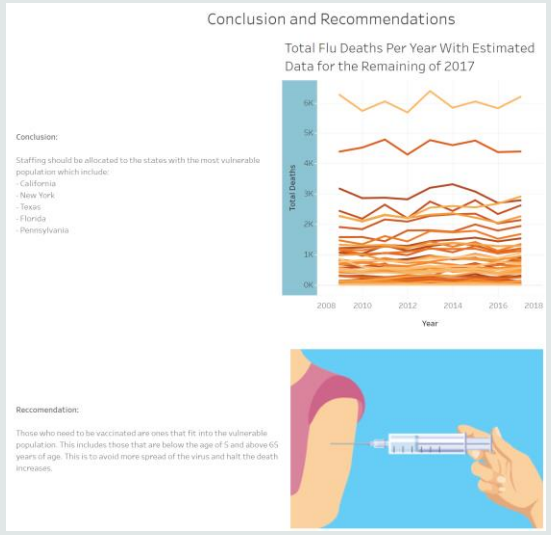


Fig. 6 Conclusions and recommendations slide in story on Tableau

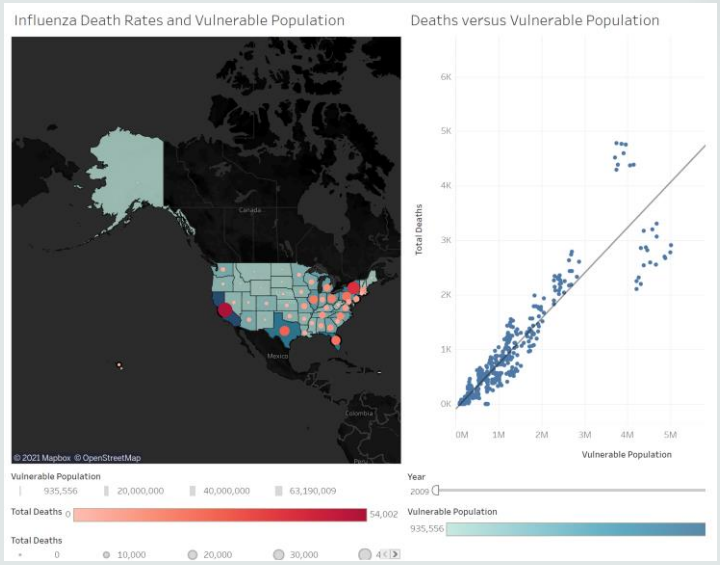
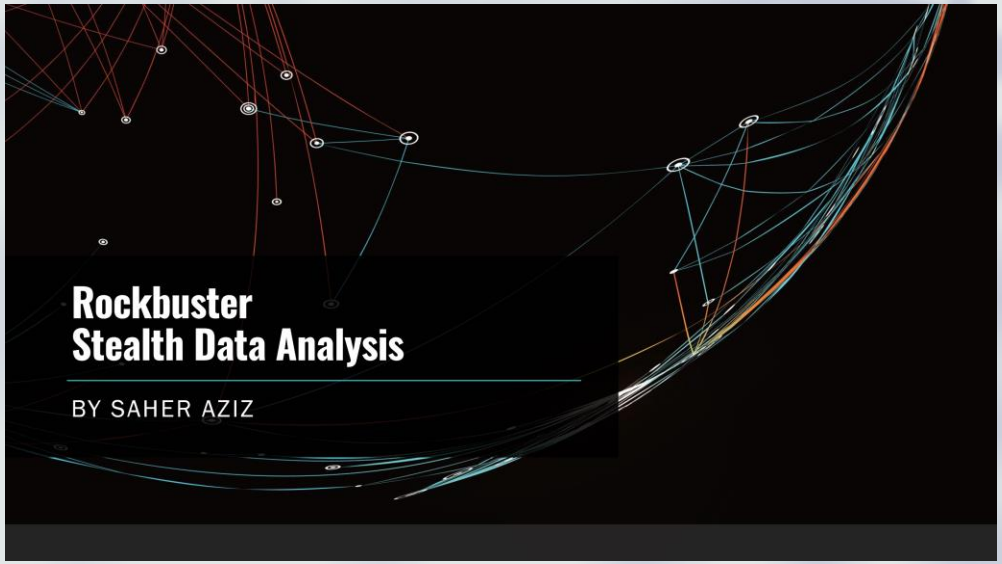


Fig.7 Map and Correlation of the vulnerable population and deaths

You can find the Tableau Presentation here and the video presentation here.

Rockbuster Stealth Summary of Inventory and Revenue using PostgreSQL and Tableau



- RockbusterStealth LLC is a fictional brick-and-mortar movie rental company with stores around the world. The management plans to use existing movie licenses to launch an online video rental service. To make informed decisions, they need to know what is in the store and how the sales performed. Additionally, the marketing department wants learn the company's most loyal customers.
- I learned to utilize SQL commands to present inventory and revenue details of the online video rental store.
- The data set is around 3MB and contains several files with film inventory, customers, and payments, among other variables.
- No source for the data was found

Rockbuster Stealth

- The goal of the project was achieved through exploration of the inventory and revenues of the stores by using the relational database management system pgAdmin4 within PostgreSQL.
- It was important to understand the relationships that exist within the database and get an overview of the tables. Using DBVisualizer, were able to extract the ERD of the database and discovered that it was a snowflake shape. Them, listing all the tables was needed to be presented in a data dictionary.
- I used the cleaning process to search for missing values, duplicates, and inconsistencies. Queries were used to calculate the descriptive statistics for selected columns.

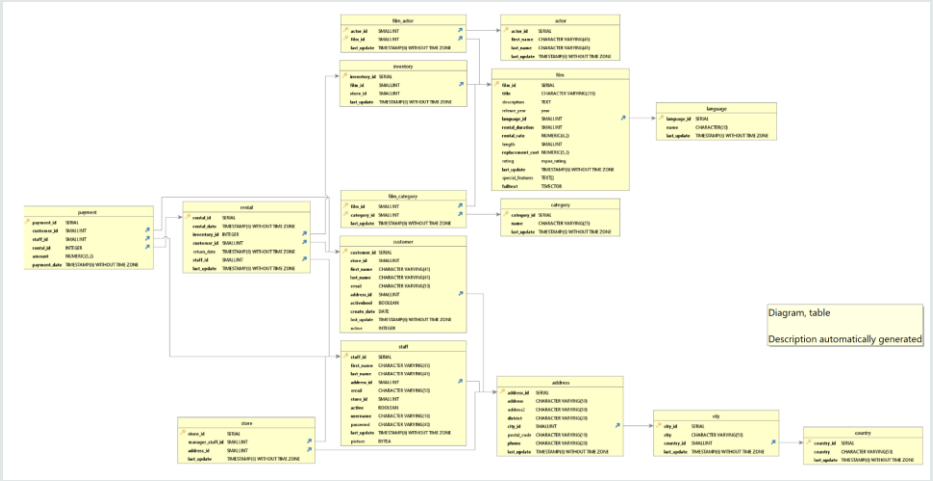


Fig. 8 ERD of the Rockbuster database

Rockbuster Stealth

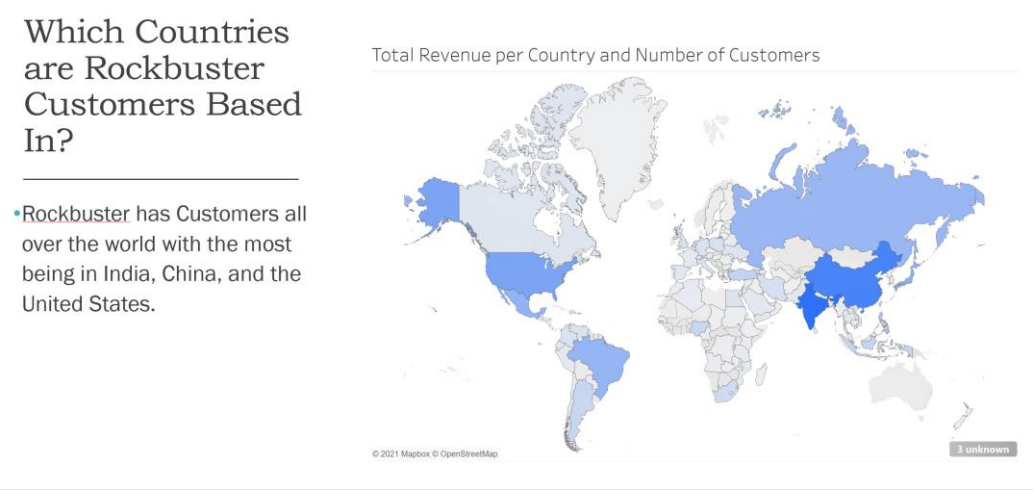
- By writing subqueries and joins between tables, I was able to find the answers to the advanced business questions. As the company was looking for the loyal customers to award them, I needed to write queries leading to the Top 5 customers.

```sql
SELECT E.customer_id,
A.first_name, A.last_name, C.city, D.country,
SUM(amount) AS total_spent
FROM customer A
INNER JOIN address B ON A.address_id=B.address_id
INNER JOIN city C ON B.city_id=C.city_id
INNER JOIN country D ON C.country_id=D.country_id
INNER JOIN payment E ON A.customer_id=E.customer_id
WHERE D.country IN ('India', 'China', 'United States', 'Japan', 'Mexico', 'Brazil', 'Russian Federation',
AND C.city IN ('Aurora', 'Pingxiang', 'Sivas', 'Dhule (Dhulia)', 'Kurashiki', 'Xintai', 'Adoni', 'Celaya',
GROUP BY E.customer_id, A.first_name, A.last_name, C.city, D.country
ORDER BY total_spent DESC
LIMIT 5
```

Fig. 9 Joins to find top 5 Customers.

Rockbuster Stealth

- My findings for the Rockbuster management board about the film inventory and revenues were presented using Tableau's storyboard.
- The full report is available through here.



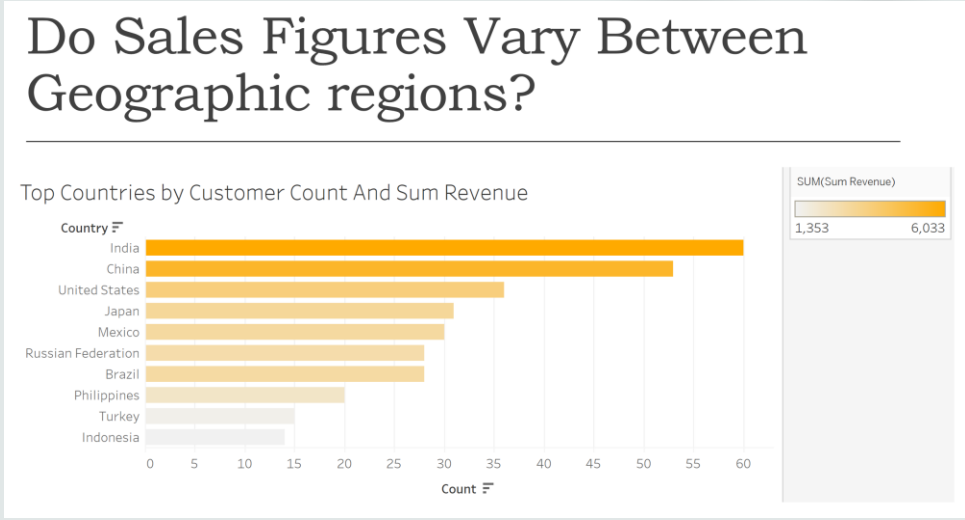


Fig. 10 & 11 Visualizations presented through Tableau.

Instacart

Analysis of Customer Profiles using Jupyter and Python

- Instacart is an online grocery store that operates through a mobile application. The stakeholders are most interested in the variety of customers in their database along with their purchasing behaviors to implement a targeted marketing strategy.
- The data analysis in this project was performed using Jupyter notebook in the Anaconda environment. I mostly used Pandas and NumPy libraries to conduct data analysis. Later on, I also added other libraries like Matplotlib, Seaborn, and SciPy to plot and visualize the results of my analysis
- The consumer data and the prices of the products were both fabricated for learning purposes. Some of the datasets contain over 32M observations.
- The datasets used for this project contain open-source data from 2017 made available online by Instacart.

Instacart

- For explanatory analysis, I used the basic functions such as .head(), .tail(), .shape, .dtypes, or .info to get a summary of the data.
- For descriptive analysis, the .describe function assisted in finding irregularities in the data.
- In the data wrangling process, changing dtypes, renaming, transposing data, and subsetting were applied.
- In the next steps, I performed consistency checks for missing values, mixed-data, and duplicates.

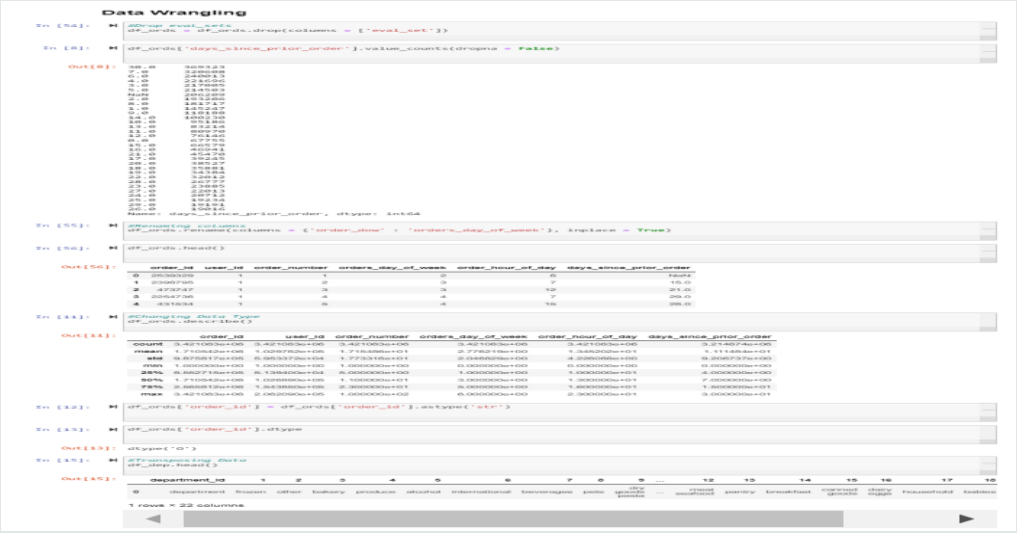


Fig. 12 Data Wrangling process

Instacart

- Large-scale manipulations include: combining data using merge, concatenate, or append. To extract further insights, used functions like .loc(), if-statements, and for-loops as well as aggregate, and grouping were applied.
- Derived new columns such as busiest days, age, income, Generations, and so on to create a customer profile.

Comparing customer behavior in different geographic areas

In [7]:
```
#Create Column that divides states by region
region =[]

for state in df['state']:
    if (state == 'Wisconsin') or (state == 'Michigan') or (state == 'Illinois') or (state == 'Indiana') or (state == 'Ohio')
        region.append('Midwest')
    elif (state == 'Maine') or (state == 'New Hampshire') or (state == 'Vermont') or (state == 'Massachusetts') or (state ==
        region.append('Northeast')
    elif (state == 'Deleware') or (state == 'Maryland') or (state == 'District of Columbia') or (state == 'Virginia') or (sta
        region.append('South')
    else:
        region.append('West')
```

In [8]:
```
#Create the region column
df['region'] = region
```

In [9]:
```
df['region'].value_counts()
```

Out[9]:
```
South        10229198
West          8994703
Midwest       7652535
Northeast     5764832
Name: region, dtype: int64
```

In [38]:
```
#creating age groups based on Generations (Source: https://www.beresfordresearch.com/age-range-by-generation/)
df_high.loc[df_high['age'] < 24, 'Generation'] = 'Gen_Z'
df_high.loc[(df_high['age'] >= 25) & (df_high['age'] <= 40), 'Generation'] = 'Millennials'
df_high.loc[(df_high['age'] >= 41) & (df_high['age'] <= 56), 'Generation'] = 'Gen_X'
df_high.loc[(df_high['age'] >= 57) & (df_high['age'] <= 66), 'Generation'] = 'Boomers_II'
df_high.loc[(df_high['age'] >= 67) & (df_high['age'] <= 75), 'Generation'] = 'Boomers_I'
df_high.loc[(df_high['age'] >= 76) & (df_high['age'] <= 93), 'Generation'] = 'Post_War'
df_high.loc[df_high['age'] >= 94, 'Generation'] = 'WW_II'
```

Fig. 13 & 14 Using for-loops and .loc() to create new columns for data profiling.

Instacart

- All operations allowed me to create visualizations to understand the insights with trends for Instacart's marketing department: For example, Gen_X and Millennials make up the most customers demographics that use Instacart.
- Charts like scatterplots let me explore not only the distribution of data points but also look for relations or spot outliers.

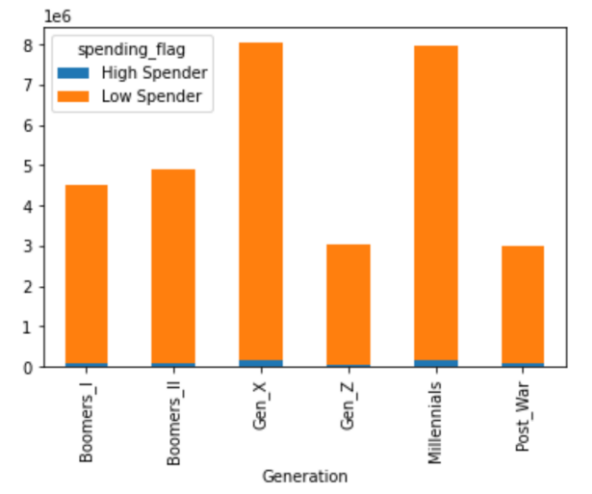


Fig. 15 Spending flags between generations.

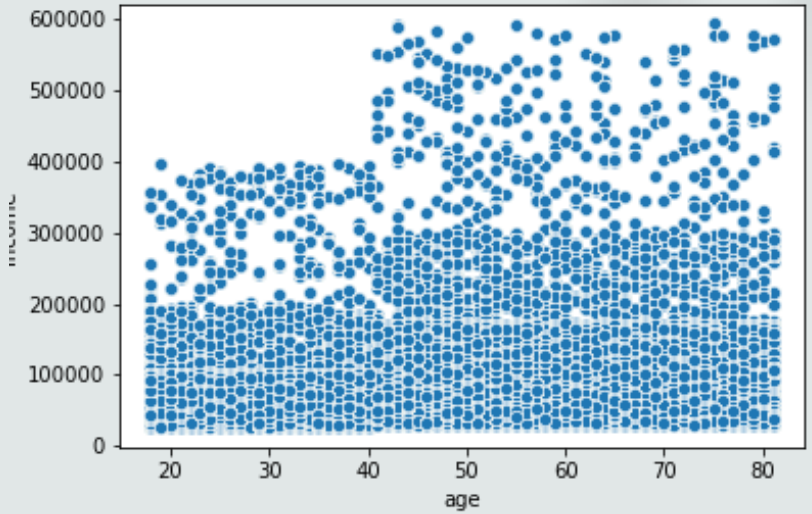


Fig. 16 Age and Income scatterplot.

Instacart

- The entire analysis process was put together using Excel Reporting where the population flow and recommendations are made for new marketing strategies.
- All files can be viewed here.

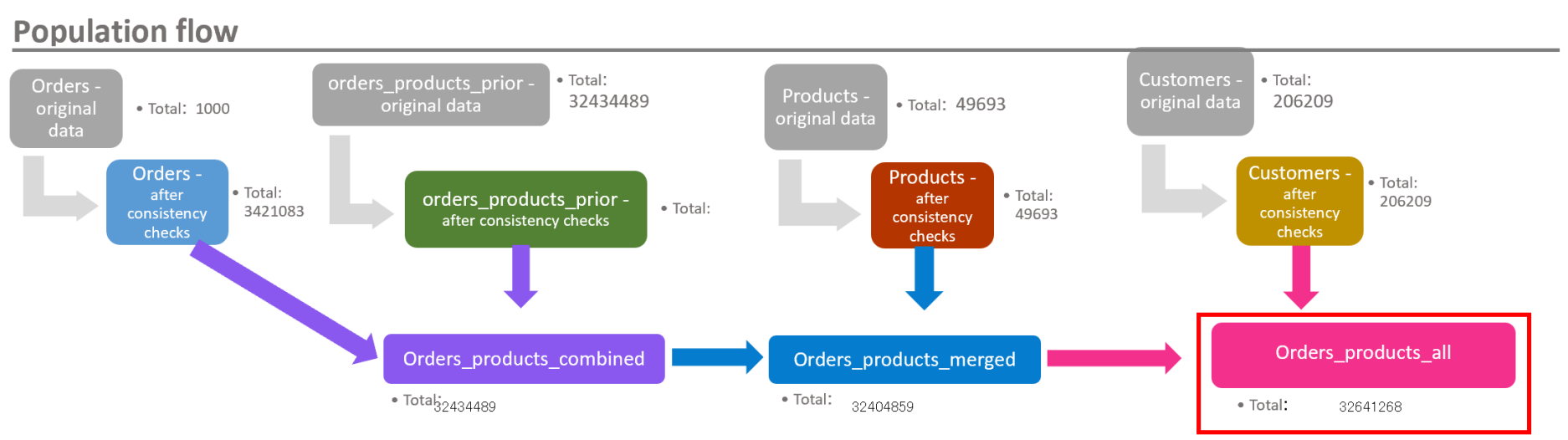


Fig. 17 Population flow for Instacart dataset

US Pollution

Data Analysis of Pollution Using Python and Tableau

- Pollution has affected many parts of the world and it increases more with time. The US has shown to have been having an increase in pollution. Because of this data has been collected to view the trends of pollution in the US.
- The data analysis in this project was performed using Jupyter notebook in the Anaconda environment. I added other libraries like Matplotlib, Seaborn, and SciPy to plot and visualize the results of my analysis. It was later presented using Tableau Storyboard.
- I conducted an advanced explanatory analysis in Python.
- The open-source dataset comes from the EPA that sources the data from 2000-2016. For the geographical visulatization, I used the us-states.json.

US Pollution

- The data set contained the information of the four main gases that contribute to the testing of pollution. These were separated by Mean, 1st Max Value, and AQI.
- For majority of the analysis, the Mean was used.
- After sourcing and cleaning the data, I learned to create a correlation diagram and interpreted the relationship between the variables.
- In the set, NO2 and CO stood out the most with the strongest relationship.

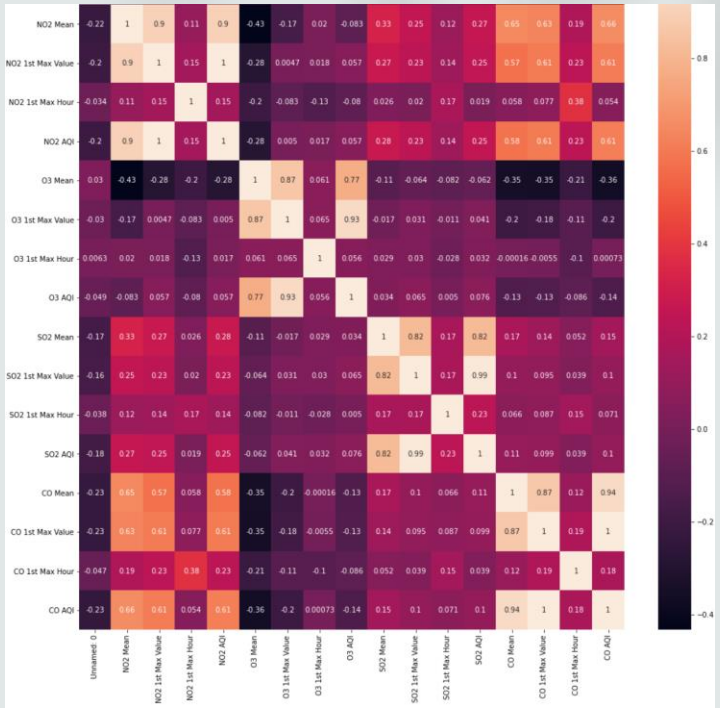


Fig. 18 Correlation Diagram

US Pollution

- Using Linear Regression, I explored the unique relationship between AQI and 1st Max Value to show the linear prediction.

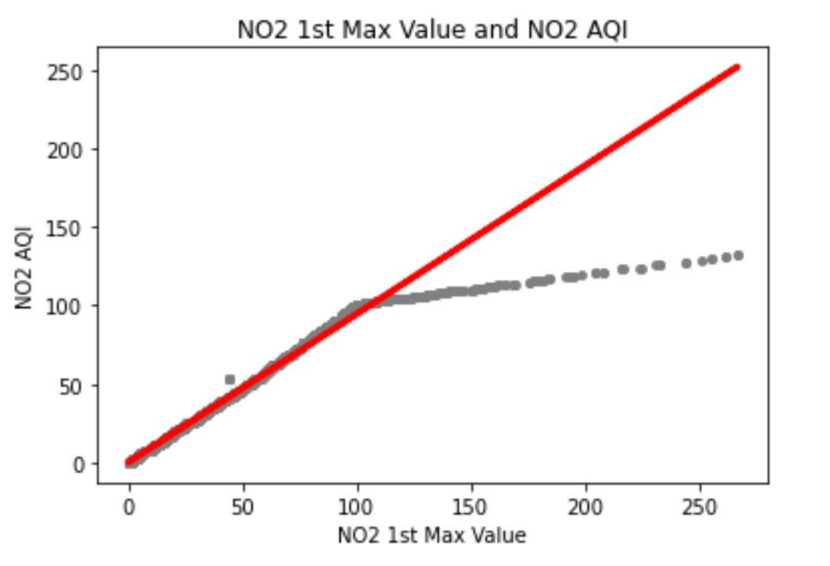


Fig. 19 Linear Regression Line

- In the unsupervised machine learning, I used clusters to confirm the linear relationship between NO2 and CO.

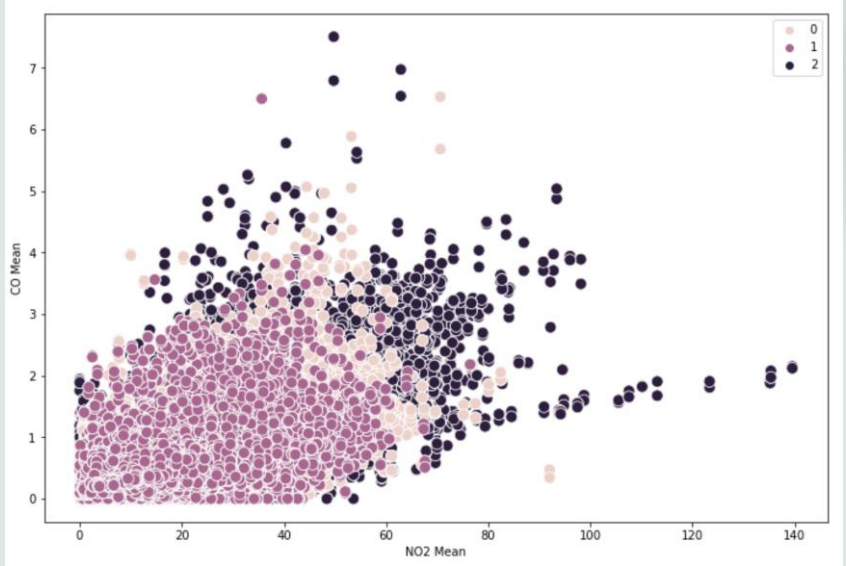


Fig. 20 Clusters Diagram

US Pollution

- Using the us-states JSON file, I created a map that represents the states with that have the most NO2 and CO gas emissions. It was interesting to uncover that the West had a higher density of NO2 and CO emissions when combined.

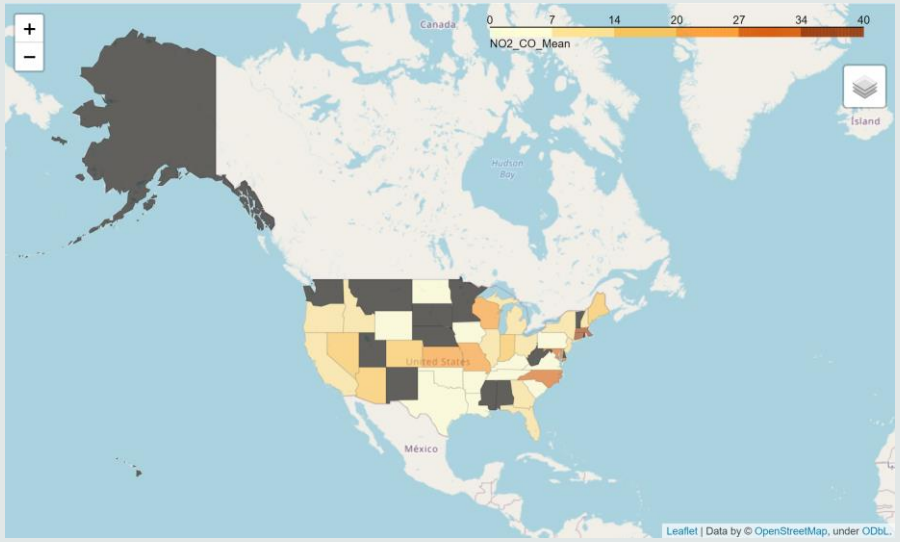


Fig. 21 Geographical Visualization of NO2 and CO Means

US Pollution

- I performed a time-series analysis to showcase the Western region effects of pollution. This used the data that is sourced from Quandl showing the natural gas purchases in the West.
- The Dickey-Fuller Test was performed to test the data's stationary to check whether the data is ready for predictions. After a few differencing operations were performed, the data was stationary.

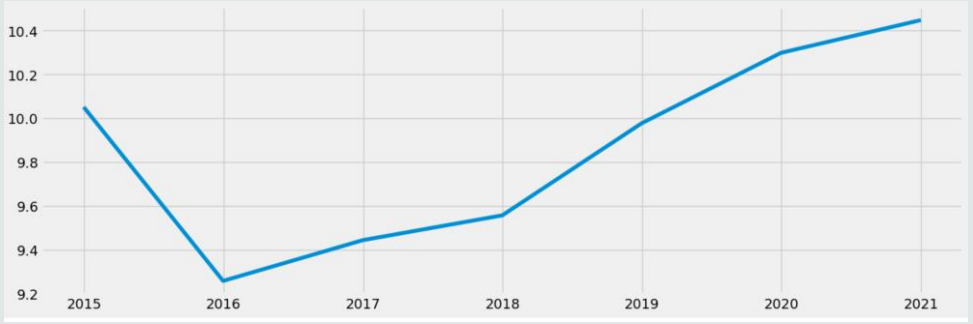


Fig. 22 Time Series

```
from statsmodels.tsa.stattools import adfuller

def dickey_fuller(timeseries): # Define the function
    # Perform the Dickey-Fuller test:
    print ('Dickey-Fuller Stationarity test:')
    test = adfuller(timeseries, autolag='AIC')
    result = pd.Series(test[0:4], index=['Test Statistic','p-value','Number of Lags Used','Number of Observations Used'])
    for key,value in test[4].items():
       result['Critical Value (%s)'%key] = value
    print (result)

dickey_fuller(data_diff['Value'])
```

```
Dickey-Fuller Stationarity test:
Test Statistic                -1.025397e+01
p-value                        4.410509e-18
Number of Lags Used            0.000000e+00
Number of Observations Used    2.500000e+01
Critical Value (1%)           -3.723863e+00
Critical Value (5%)           -2.986489e+00
Critical Value (10%)          -2.632800e+00
dtype: float64
```

Fig. 23 The Dickey-Fuller Test

US Pollution

- There are many factors that result in pollution. The four gas emissions were only the few factors that contribute to the problem.
- Between that set, a strong relationship was found between NO2 and CO based on the correlations.
- This resulted in the findings that the West had the most emissions of these gasses, resulting in more pollution in those regions.
- The Jupyter scripts are available here while the Tableau presentation is available here.

Summary

- Working on all five projects, I utilized the tools such as Excel, Tableau, SQL, and Python to explore what it takes to be a Data Analyst.
- Data Analytics is a fantastic field to work in and my goal is to expand my knowledge and skills for future projects.