Data Analytics Portfolio

Samir Thomas

LinkedIn | GitHub | Tableau

Hello there! I’m Samir.

About Me:

Hi there! My name is Samir Thomas, and I am a Data Analyst based in the DMV (DC-Maryland-Virginia) area. I am experienced handling and **processing sensitive data** efficiently and ethically. My background includes a generous mix of **public service** and data processing. My previous experiences have led me to developing my skills in **data analytics** in order to better understand the meeting point of human decision making and technology, and how we can use technology to better the lives of those we affect. Clear and effective communication, **creative critical thinking**, & technical prowess are my core competencies.

Key skills:

- Data Analysis, Marketing Analysis, Business Intelligence
- Data Science, Machine Learning (supervised and unsupervised)
- Forecasting, Statistical Hypothesis Testing, Time Series Analysis
- Data Visualization, Data Sourcing and Reporting, Data Mining, Data Cleaning, Data Blending

Tools:

- Microsoft Excel
- Tableau
- SQL
- Python
- Power BI
- Microsoft Office Suite and Google Suite

PROJECTS

1. GameCo: Video Game Popularity (Excel)
   - Analysis of global video game sales using advanced Excel techniques
2. Preparing for Influenza Season (Excel & Tableau)
   - Analysis of population, vaccination, and flu mortality data using Excel to create Tableau visualizations
3. Rockbuster Stealth (PostgreSQL & Tableau)
   - Analysis of global movie rental database using PostgreSQL and presenting insights using Tableau visualizations
4. Instacart Co. (Python)
   - Analysis of Instacart online grocery shopping database using Python techniques and visualizations
5. World University Rankings (Python)
   - Analysis of the top ranked world universities using Python and Tableau visualizations

Project #1: GameCo.

Overview

Performed descriptive analysis of a video game data set to foster a better understanding of how GameCo's new games might fare in the market

Key Questions

- Are certain types of games more popular than others?
- What other publishers will likely be the main competitors in certain markets?
- Have any games decreased or increased in popularity over time?
- How have their sales figures varied between geographic regions over time?

Data

VGChartz historical video games sales data from 1980 to 2016

Key Skills

Microsoft Excel

Cleaning and Grouping Data

Summarizing Data

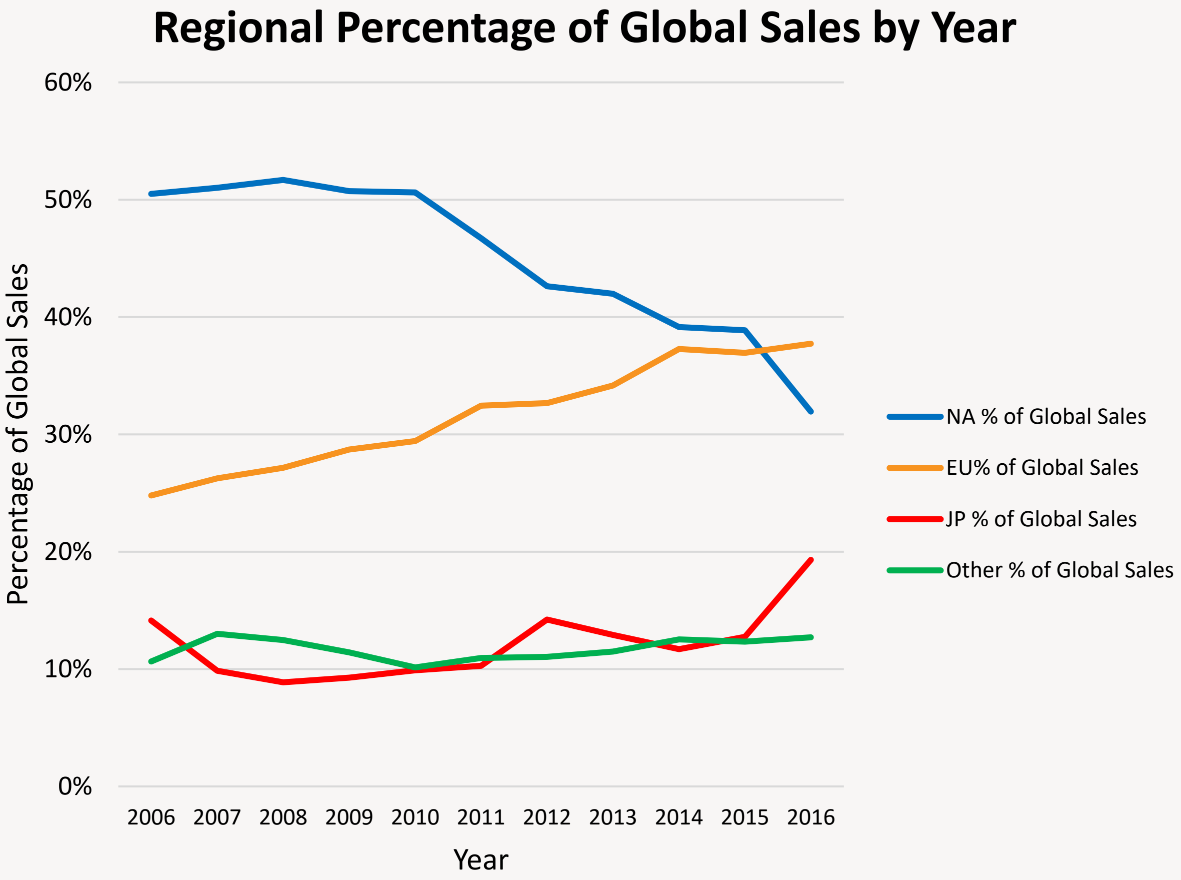
Pivot Tables

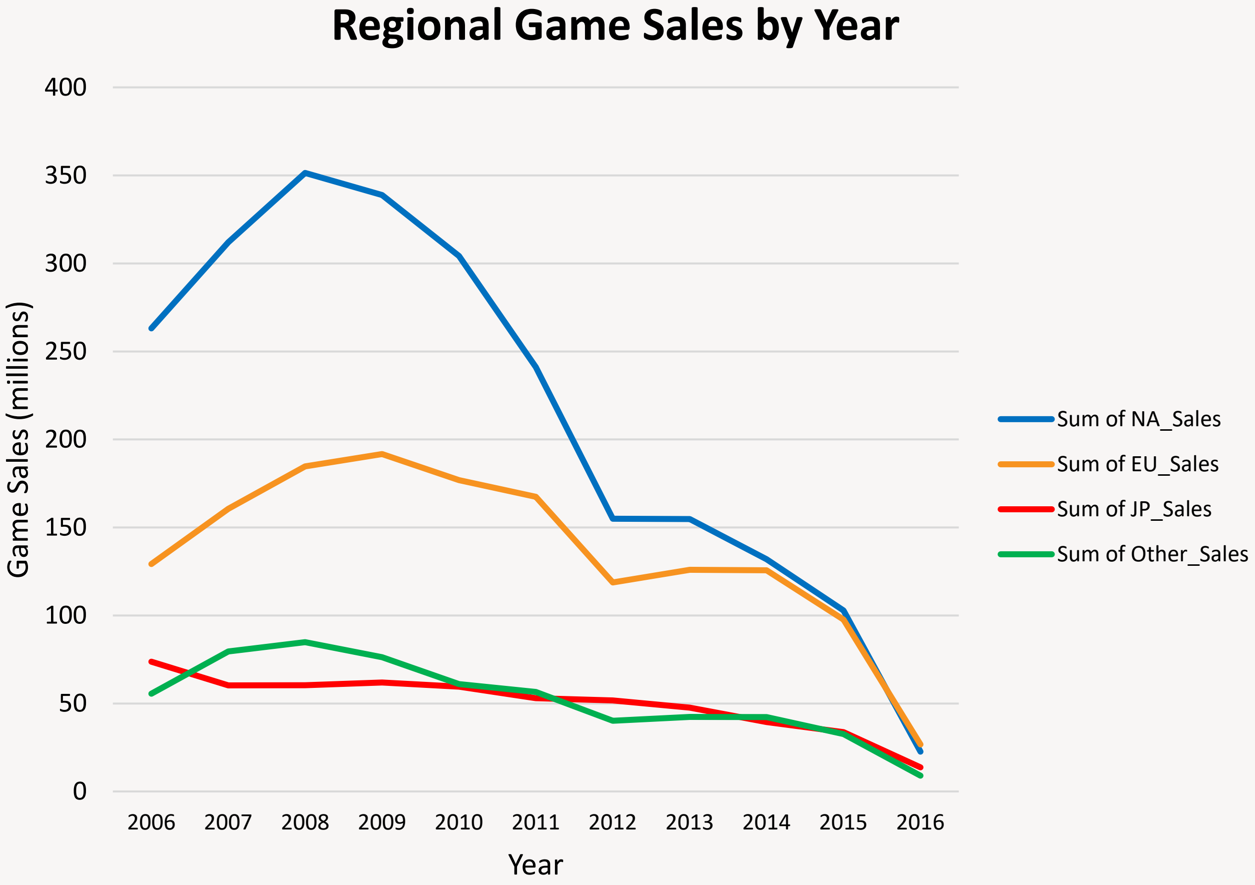
Descriptive Analysis

Visualizing results in Excel

Presenting Excel Results

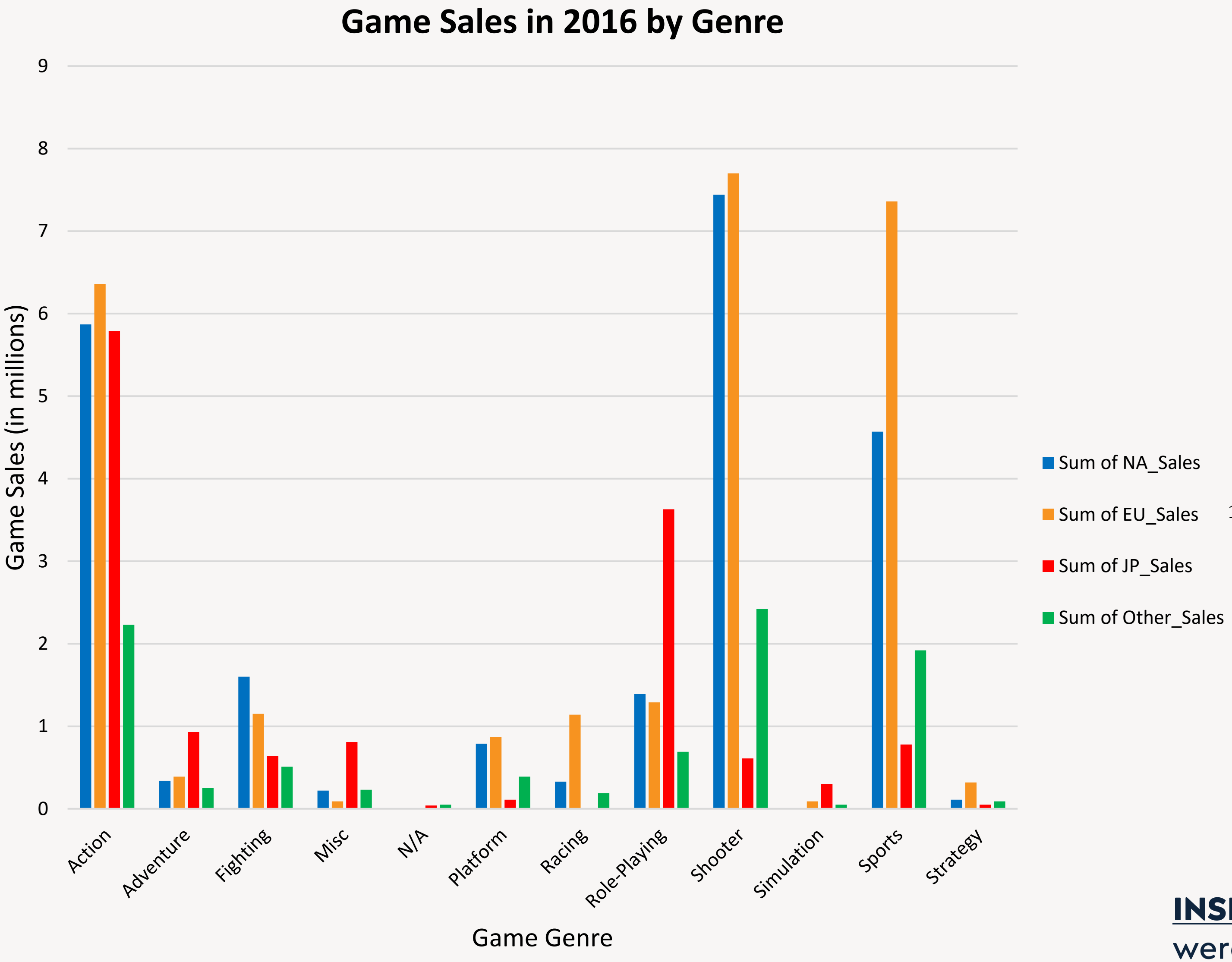
Regional Analysis

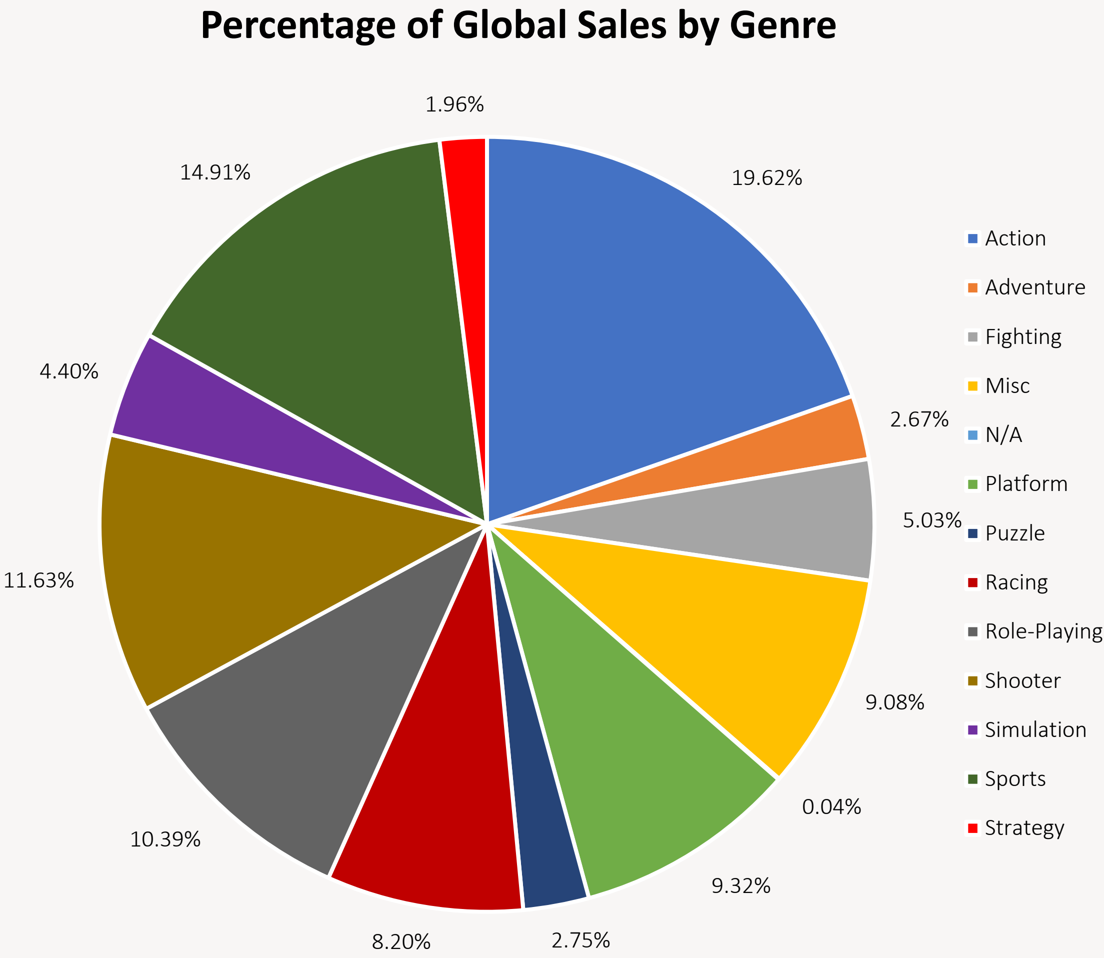




Details: After performing data cleaning procedures (missing, null, duplicate, formatting inconsistencies), pivot tables were used to analyze the sales trends by regions. By charting the data over the entire course of the data set we were able to identify trends in global sales percentages by region. We recognized that while global sales are decreasing, the percentage of global sales in the European and Japanese regions were increasing, signaling areas of interest.

Game Industry Analysis





INSIGHT: Overall, the highest performing game genres in 2016 were shooter, sports, and action.

GameCo Project Wrap Up

Recommendation #1

Revise our original assumption that regional sales stayed the same over time. Our new current understanding is that overall game sales are decreasing steadily both regionally and globally.

Recommendations #2

Adjust our marketing budget to reflect the regions that are showing growth for game sales. Focusing our marketing on the EU and JP region as they have both increased their proportion of global game sales over the last year showing that they are areas of opportunity.

Recommendation #3

Adjusting our attention to the game genre's that are popular in the EU and JP regions. Since the EU region has overtaken the NA region as the region that contributes the most to global sales, and JP global sales percentage is increasing, the genre's that are popular within those regions should be highly considered in marketing decision.

Project #2: Preparing for Influenza Season

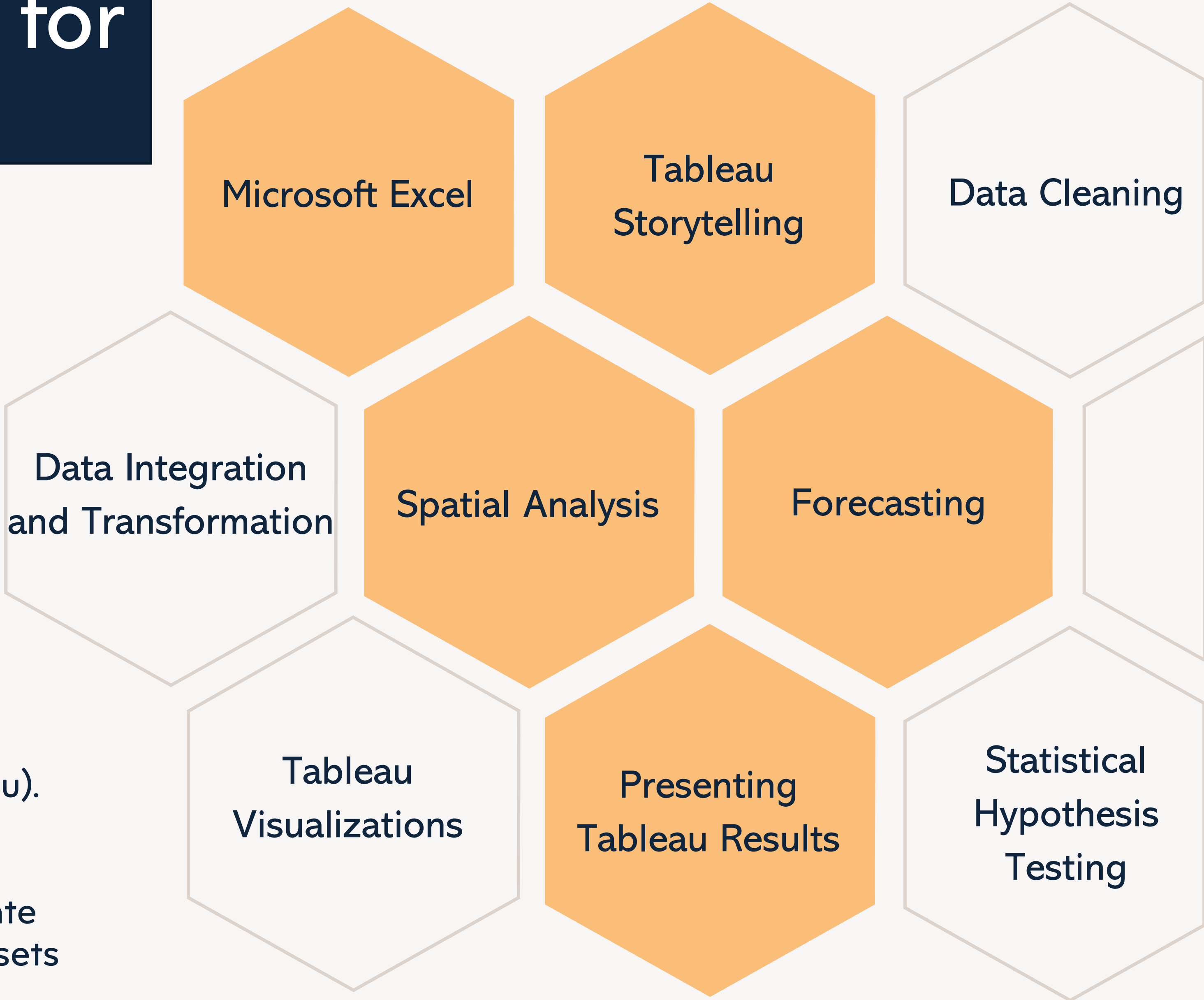
Overview

Analyze historical census and flu data to inform medical staffing agencies with data backed recommendations on when to send staff and how many to each state in preparation of the upcoming influenza season.

Data

- Influenza Deaths by geography, time, age, and gender (CDC). Data set
- Population data by geography (US Census Bureau). Data set
- Counts of Influenza laboratory test results by state (survey) (CDC). Influenza Visits & Lab Test Data sets
- Survey of Flu shot rates in children (CDC). Data set

Key Skills



Microsoft Excel

Tableau Storytelling

Data Cleaning

Data Integration and Transformation

Spatial Analysis

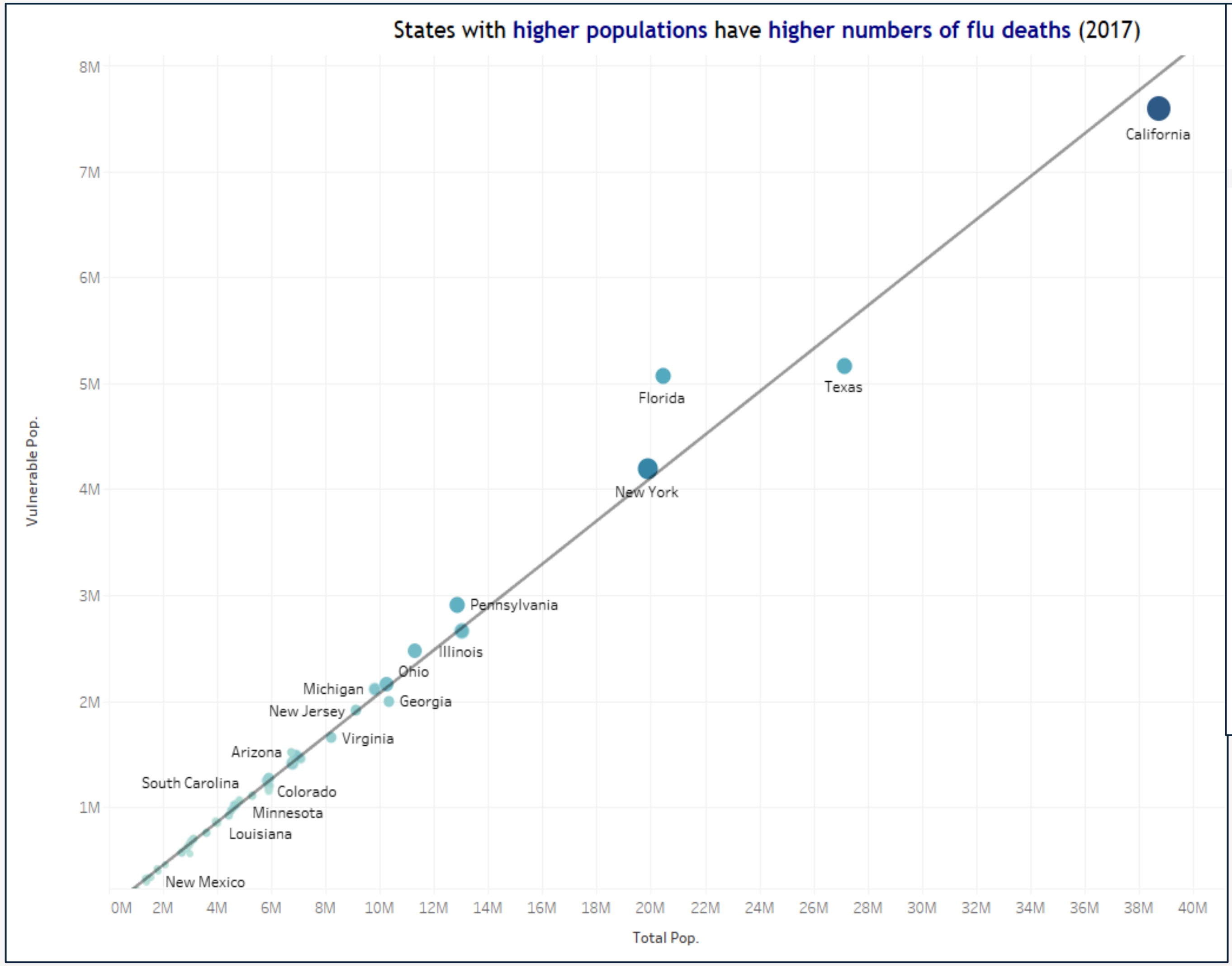
Forecasting

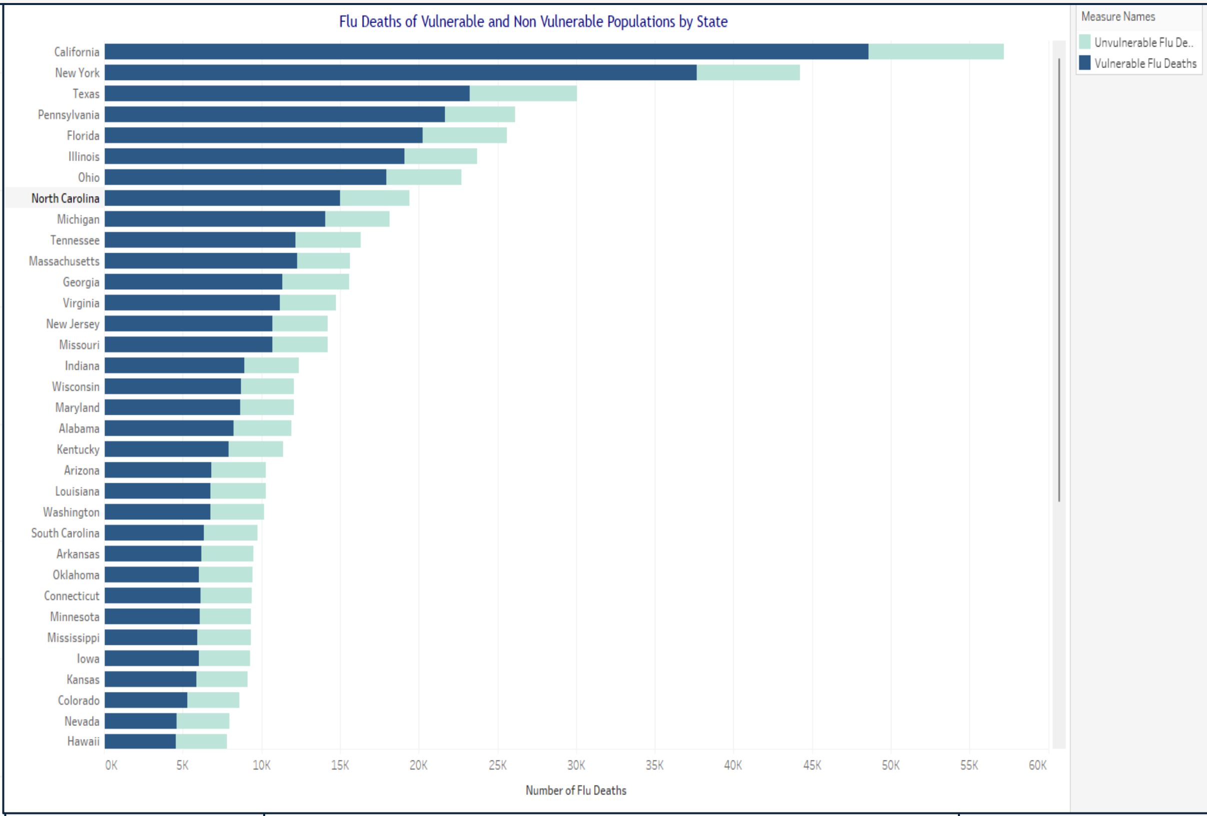
Tableau Visualizations

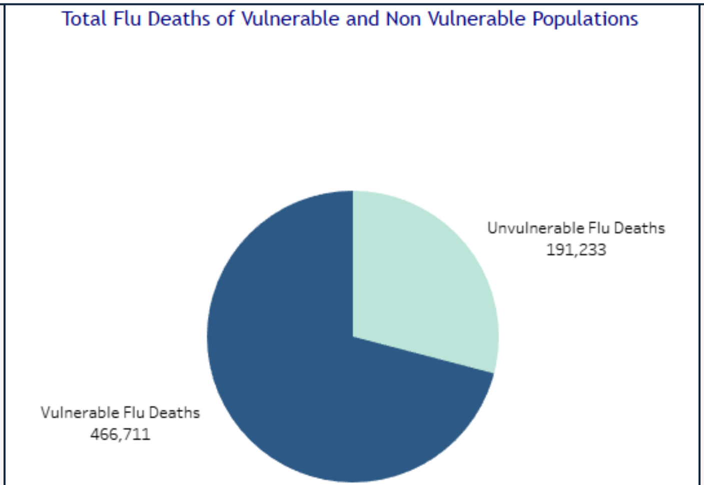
Presenting Tableau Results

Statistical Hypothesis Testing

Data Analysis

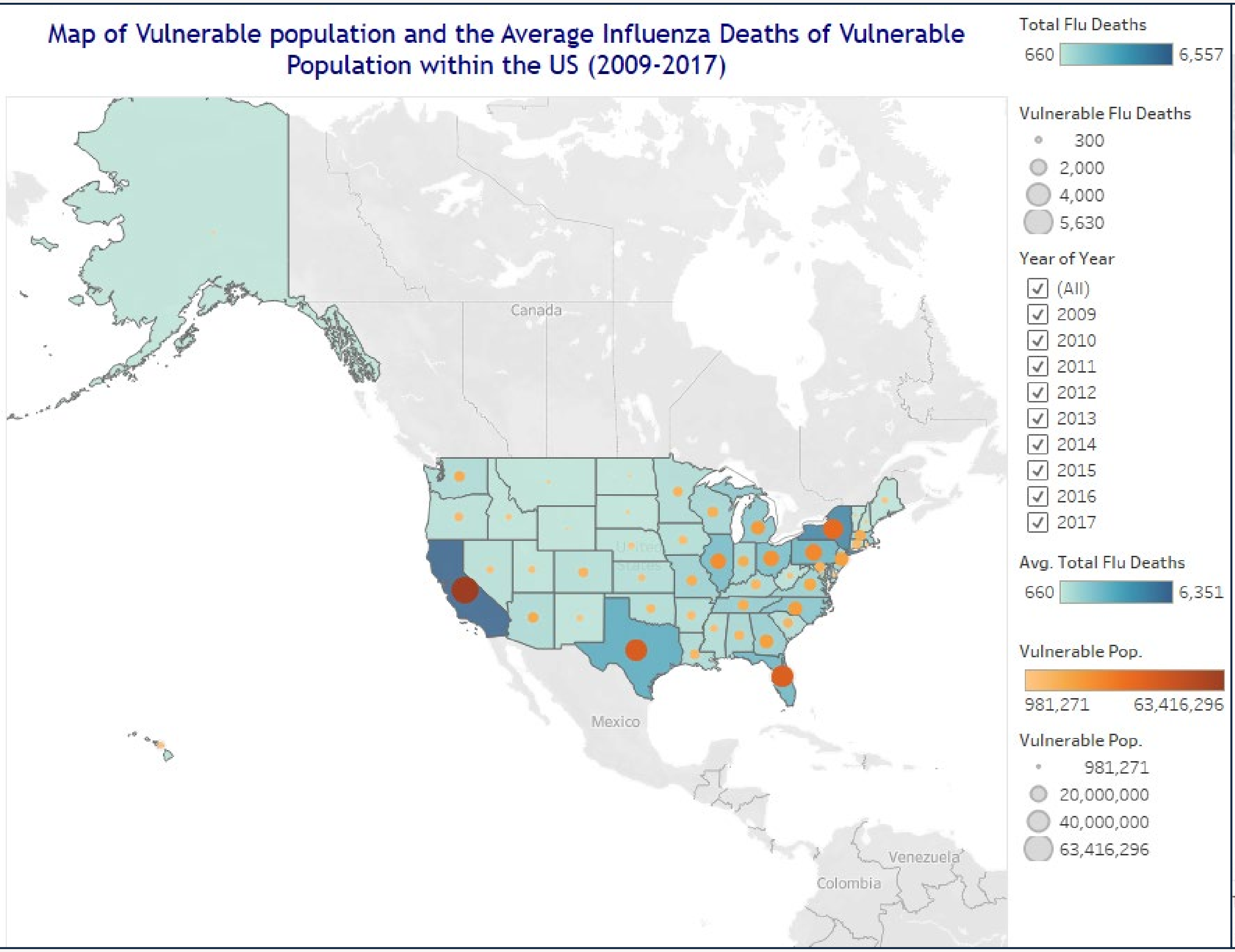




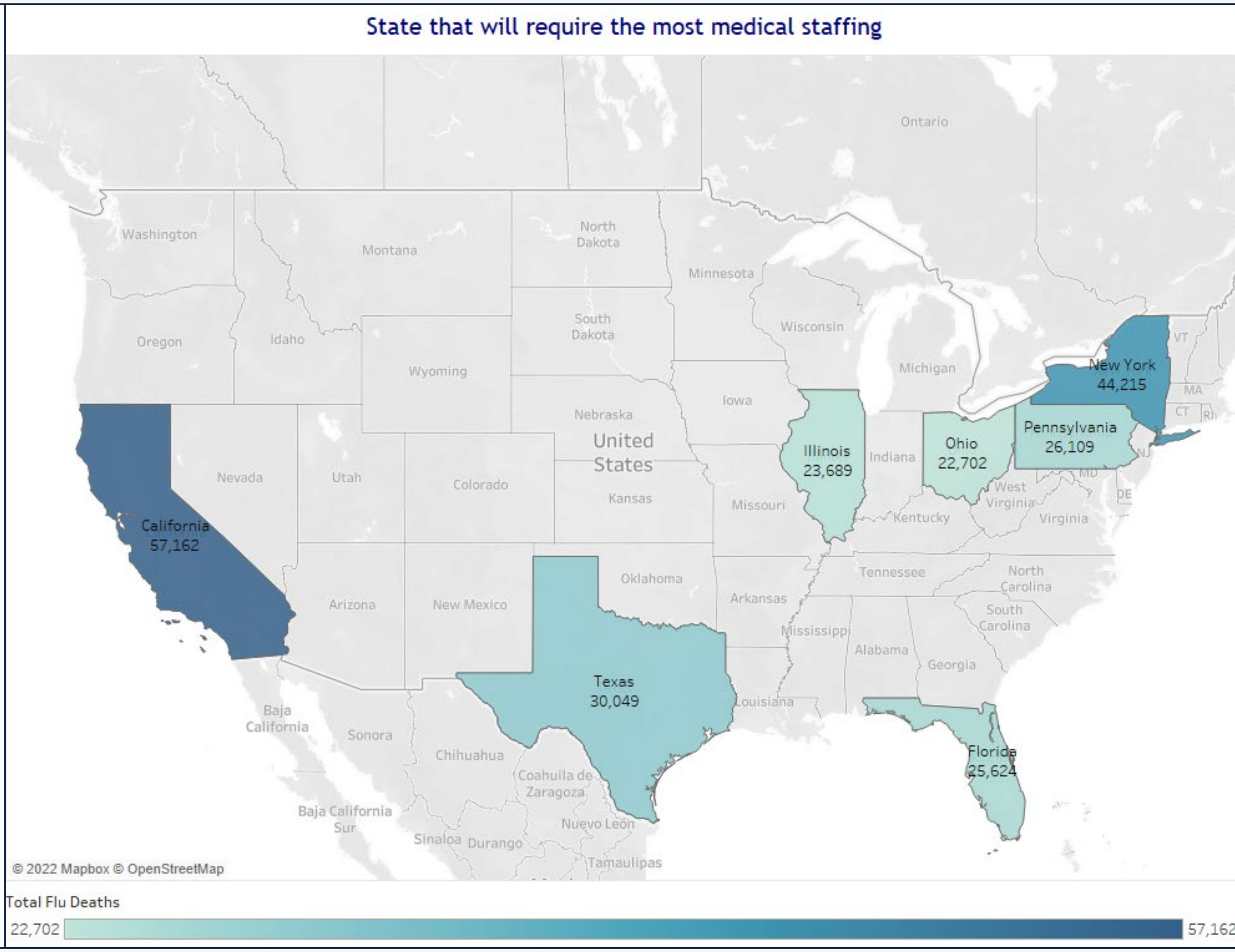


Details: The data sets were integrated into a singular dataset that included flu mortality numbers in addition to population census data. This data was then used in Tableau to analyze for insights on vulnerable populations and flu seasonality trends.

Data Visualizations & Storytelling



Map of Vulnerable population and the Average Influenza Deaths of Vulnerable Population within the US (2009-2017)



State that will require the most medical staffing

<u>Details:</u> Spatial analysis was conducted to highlight states of interest who would be impacted the most by the Flu season. States with the highest populations and the highest population of vulnerable people (0-5 yrs. old, and 65-over) were isolated as areas of interest for medical staffing.

Influenza Project Wrap Up

Conclusion

It has been determined that there is a strong positive correlation between the number of vulnerable people present within a population and the number of flu related deaths. Therefore, states with the highest number of vulnerable populations can be easily identified for medical staffing purposes.

Recommendations

- During the flu season, it would be of the greatest benefit to send additional medical staffing to states who have the largest population of vulnerable people - California, Texas, New York, Pennsylvania, Florida, Illinois, and Ohio.
- These additional medical staff should be deployed around November (at the earliest) and should be expected to provide medical assistance until at least March of the following year.

Deliverables

Interim Report

Tableau Dashboard

Stakeholder Presentation

Project #3: Rockbuster Stealth

Overview

Analyze historical data in order to better inform the Rockbuster Stealth Management Boards 2020 Company Strategy.

Key Questions

- Which movies contributed the most/least to revenue gain?
- What was the average rental duration for all videos?
- Which countries are Rockbuster customers based in?
- Where are customers with a high lifetime value based?
- Do sales figures vary between geographic regions?

Data

- Rockbuster Stealth data set

Key Skills

Relational Databases

PostgreSQL

Data Cleaning

Joining Tables

Database Querying

Filtering

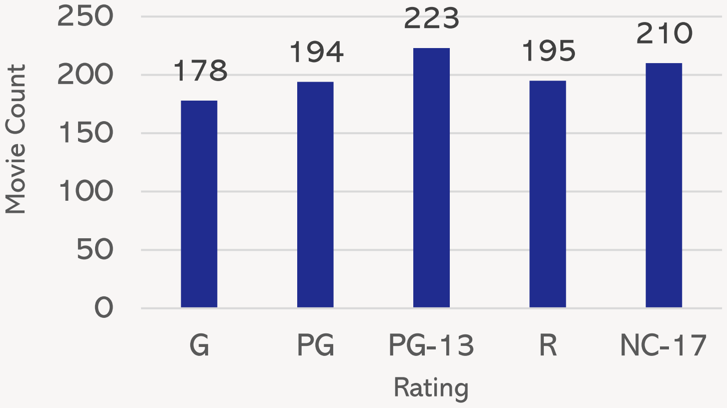
Common Table Expressions

Subqueries

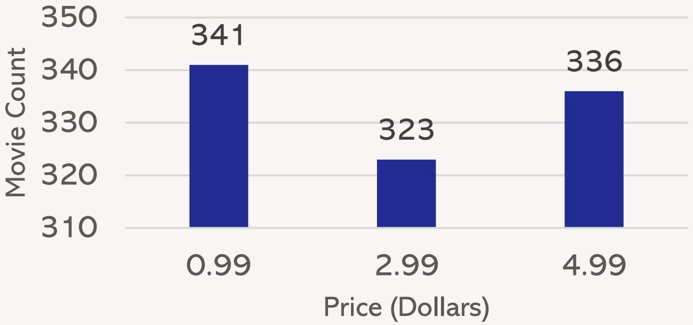
Cleaning and Summarizing

Rockbuster Overview

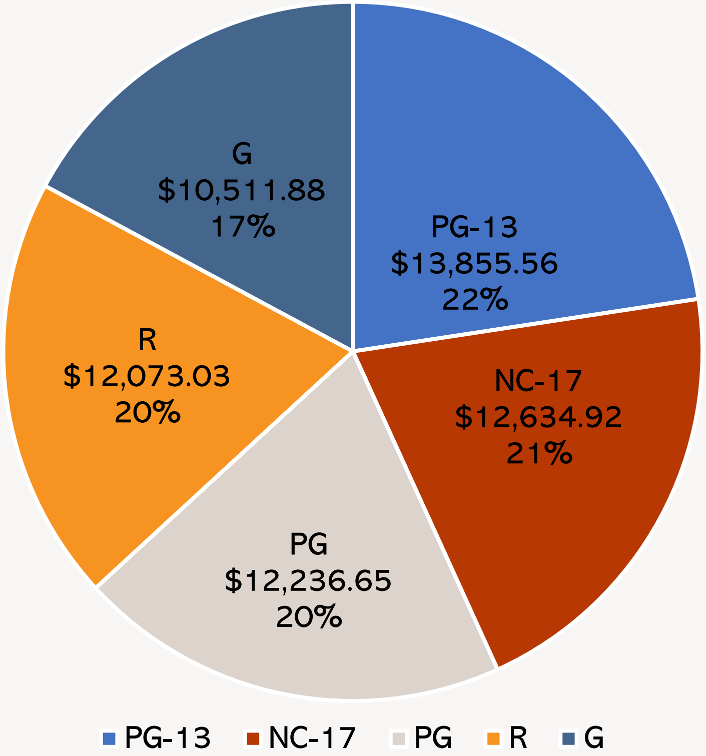
Number of Movies by Rating



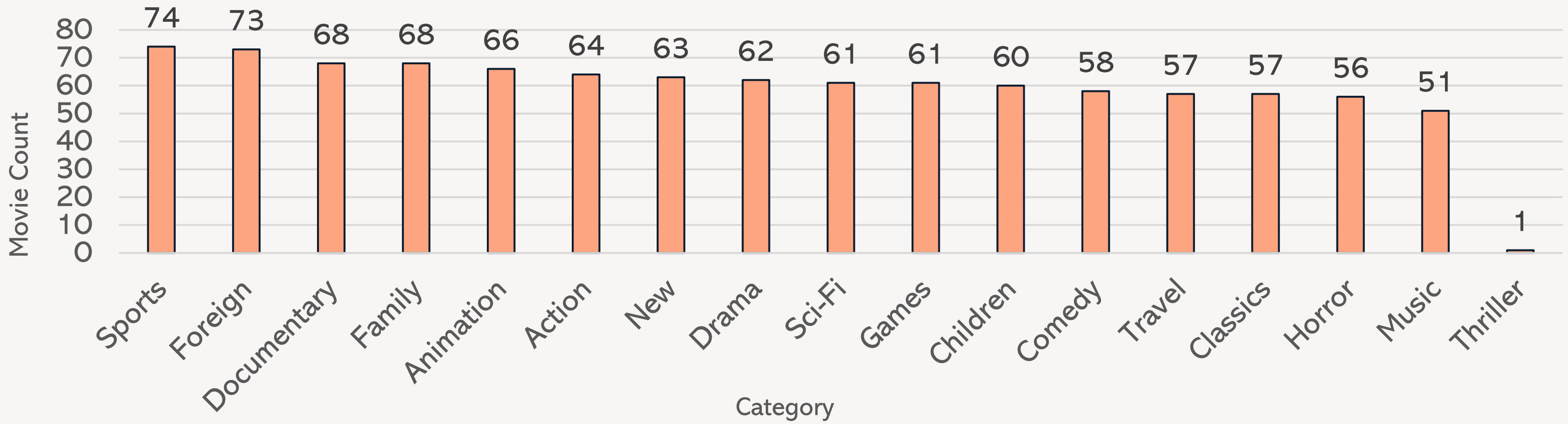
Number of Movies by Rental Rate



Movie Revenue by Rating

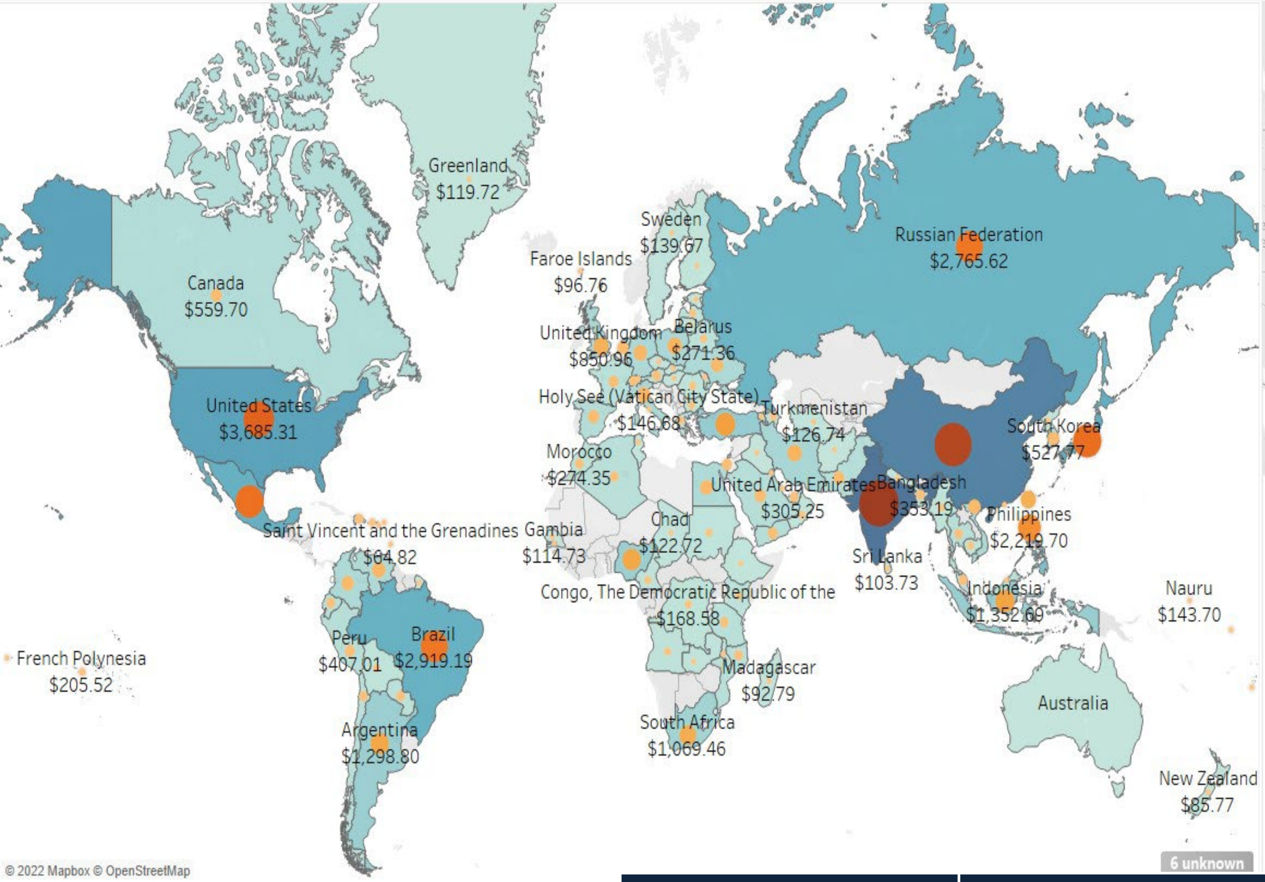


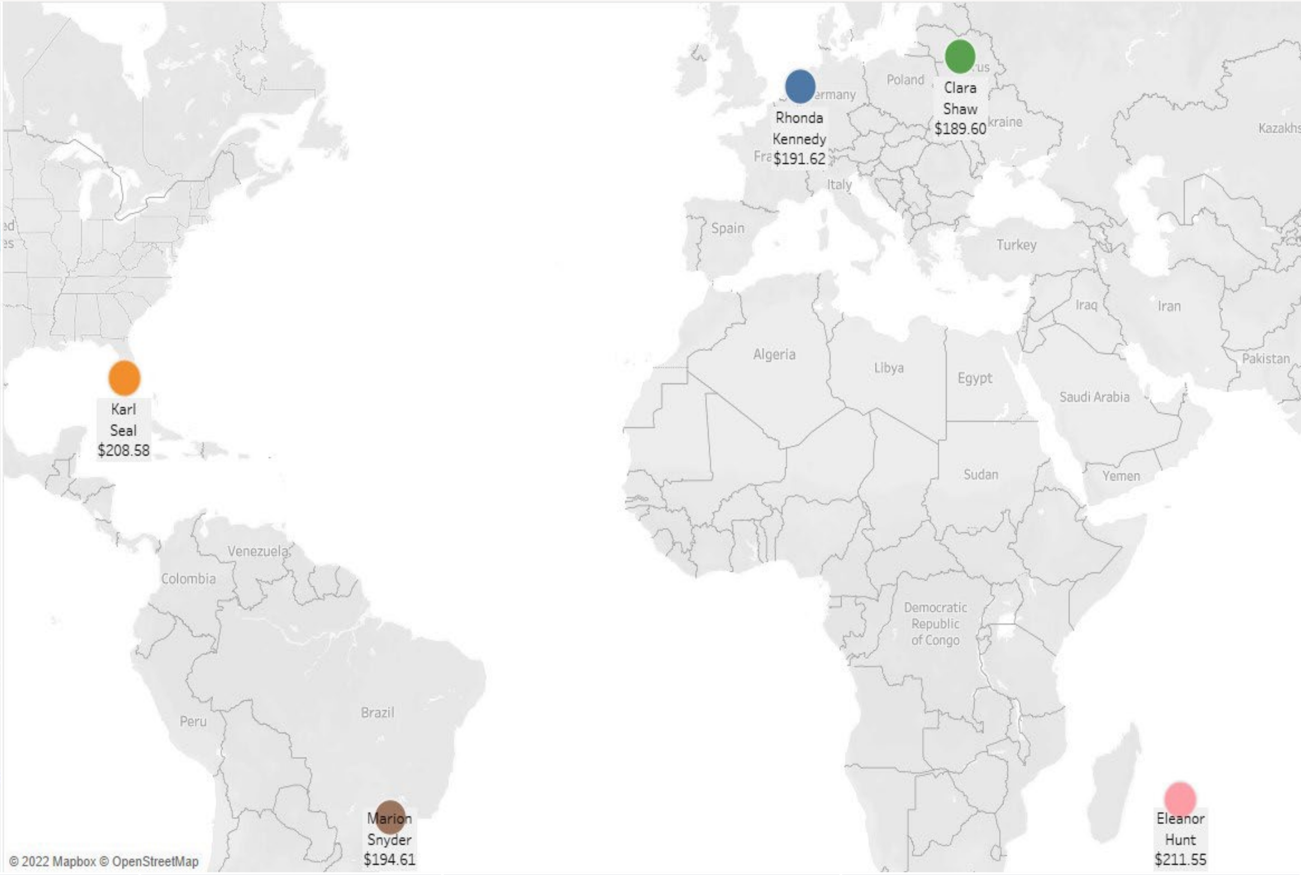
Number of Movies by Category



Details: Utilized CRUD commands to clean and analyze the dataset. Queries for filtering, grouping, joining, and CTE's were performed to isolate specific data points to reference within this project.

Customer Profile Analysis





Details: Regional analysis was also performed within SQL to find out the top performing countries as well as the top 5 customer.

Customer Name	City	Country	Number of Rentals	Total Revenue (dollars)
Eleanor Hunt	Saint-Denis	Reunion	46	$211.55
Karl Seal	Cape Coral	United States	45	$208.58
Marion Snyder	Santa Barbara d'Oeste	Brazil	42	$194.61
Rhonda Kennedy	Apeldoorn	Netherlands	42	$191.62
Clara Shaw	Molodechno	Belarus	41	$189.60

Rockbuster Project Wrap Up

Recommendation – Revamp Inventory

Focus on acquiring movies that are more likely to be rented out more frequently and for longer durations.

- Sports
- Science Fiction
- Animation
- Drama
- Comedy

Recommendation – Targeted Marketing

Hone in on location based research to expand sales within the top 5 performing countries

- India
- China
- United States
- Japan
- Mexico

Deliverables

Data Dictionary

Presenting SQL Results Presentation

Tableau Visualizations

SQL Query References

GitHub Repository

Project #4: Instacart Grocery Basket Analysis

Overview

Performed an initial data and exploratory analysis of some of their data in order to derive insights and suggest strategies for better segmentation based on the provided criteria.

Some Key Questions

- The sales team needs to know what the busiest days of the week and hours of the day are (i.e., the days and times with the most orders) in order to schedule ads at times when there are fewer orders.
- They also want to know whether there are particular times of the day when people spend the most money, as this might inform the type of products they advertise at these times.

Data

- Open-Sourced data sets provided by Instacart. Data Set
- Fictional data set provided by CareerFoundry. Data Set

Key Skills

Python

Data Wrangling

Data Cleaning

Deriving Variables

Data Merging

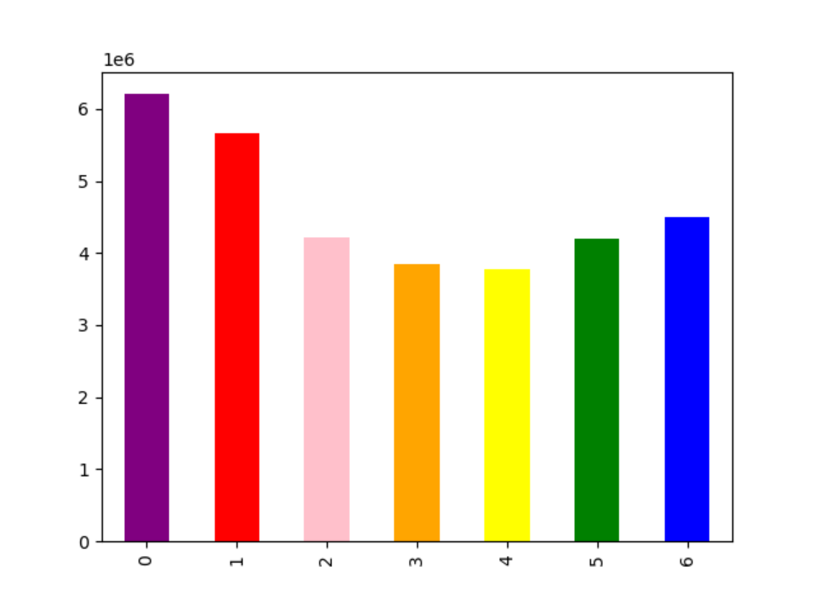
Grouping Data

Aggregating Data

Reporting in Excel

Population Flows

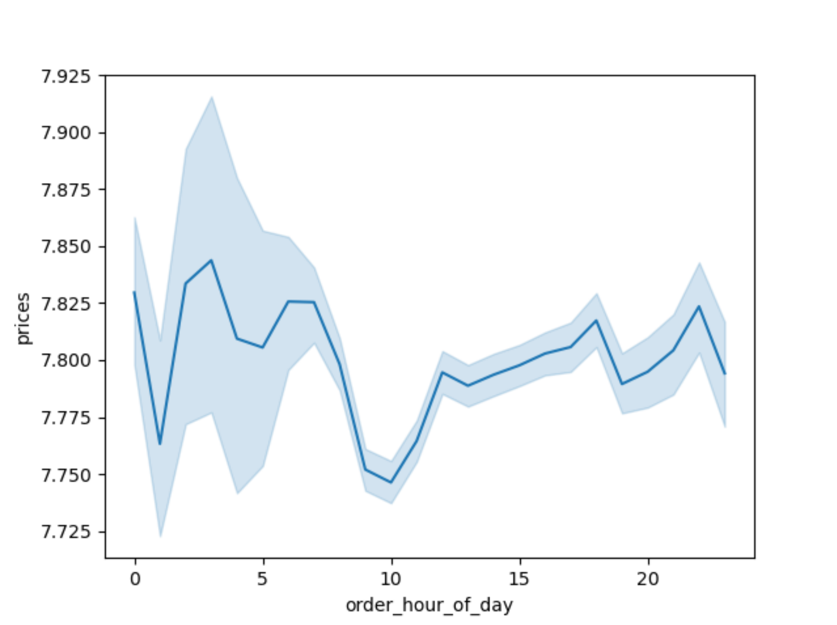
Exploratory data analysis

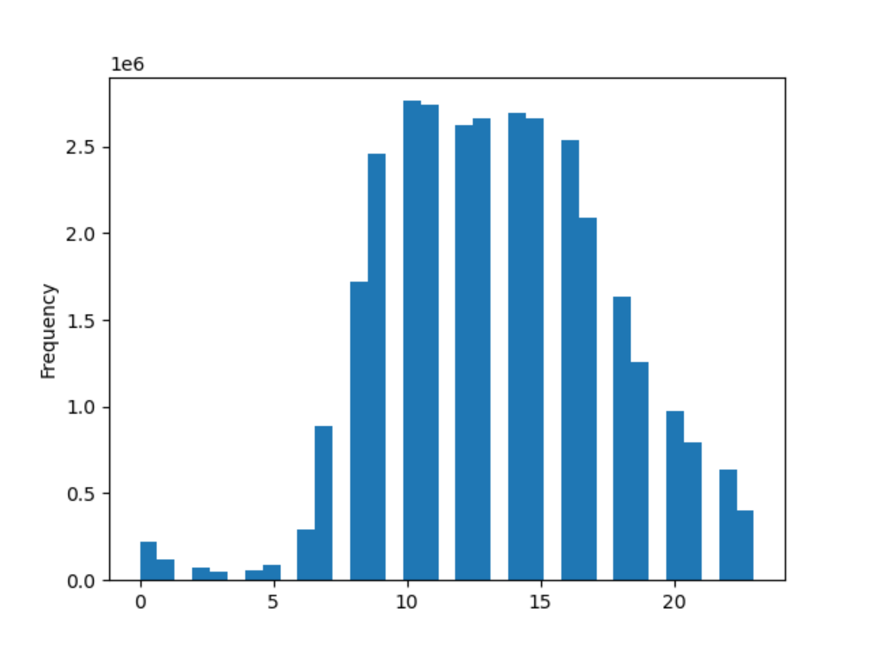


Legend
0 Saturday
1 Sunday
2 Monday
3 Tuesday
4 Wednesday
5 Thursday
6 Friday

This graph visualizes the busiest days of the week. Which is Saturday and Sunday.

This graph shows the average price of products purchased throughout the day.



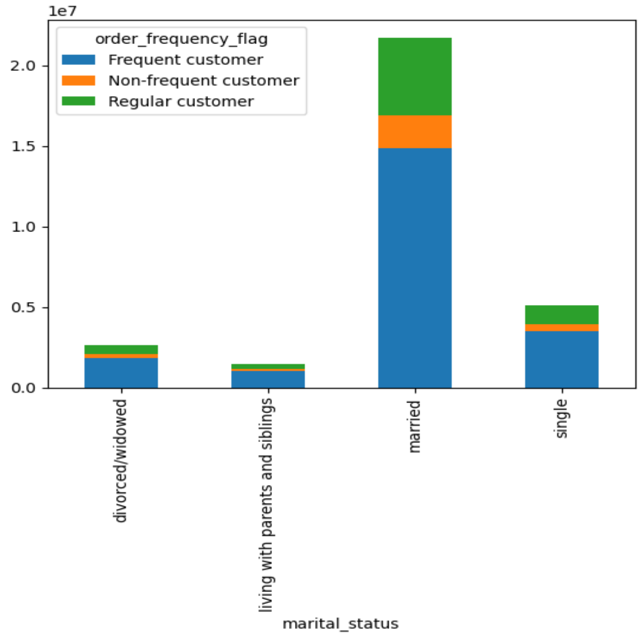


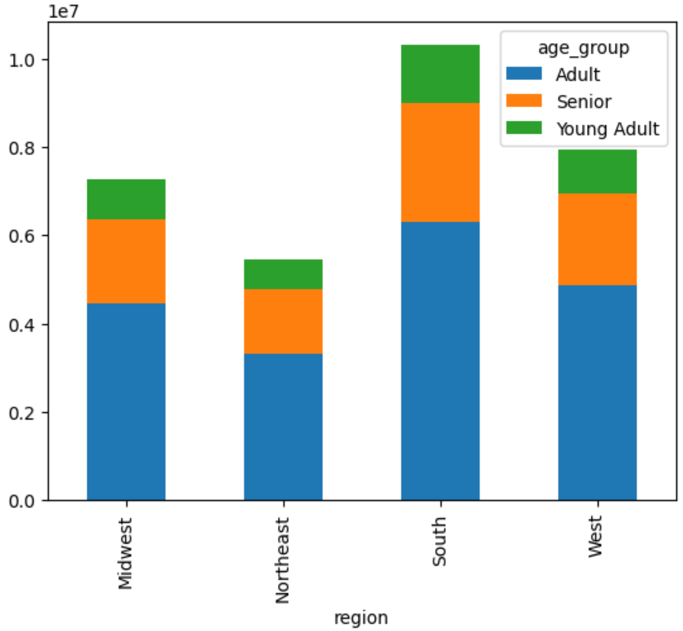
This graph shows the busiest hours of the day.

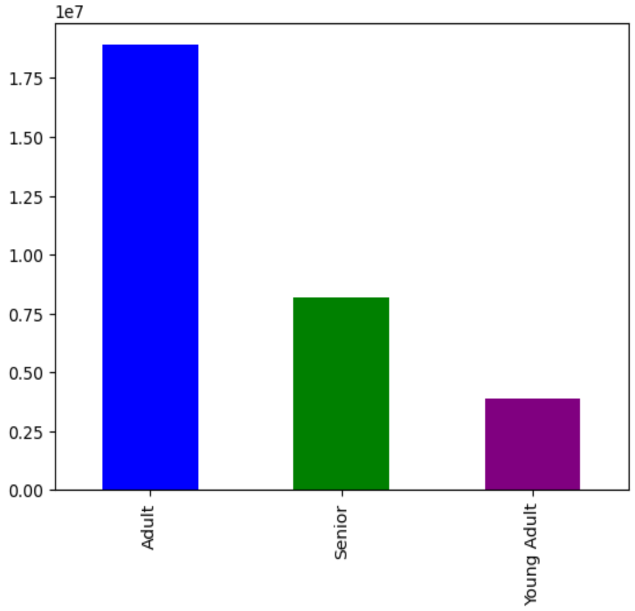
Details: Visualizations were created within Python after the data set was cleaned, wrangles, formatted, merged, and checked for inconsistencies.

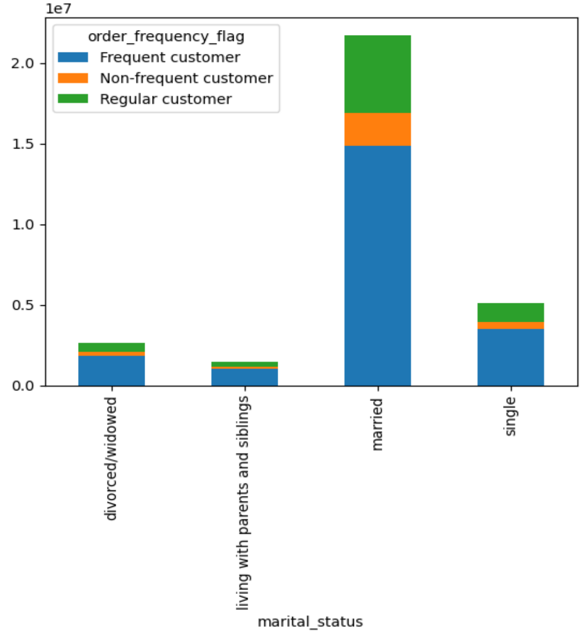
Customer Profile Analysis

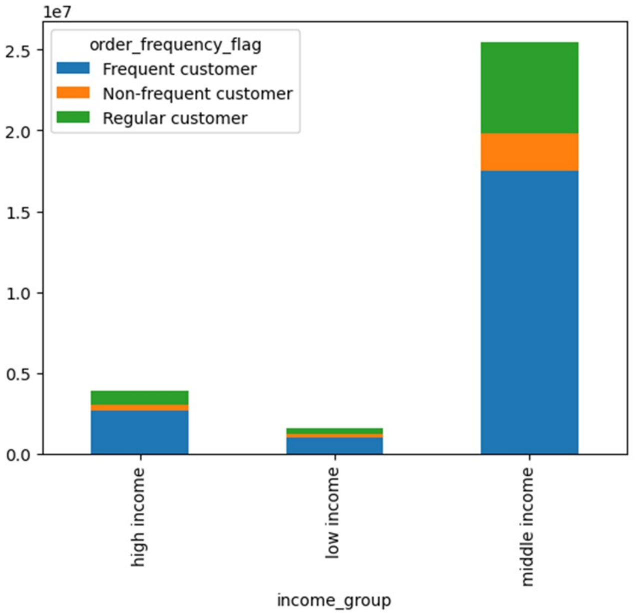
Details: Customers were group together by various criteria in order to tease out insights regarding Instacart's customer base. Some of the area’s looked at include marital status, income, and residential location.

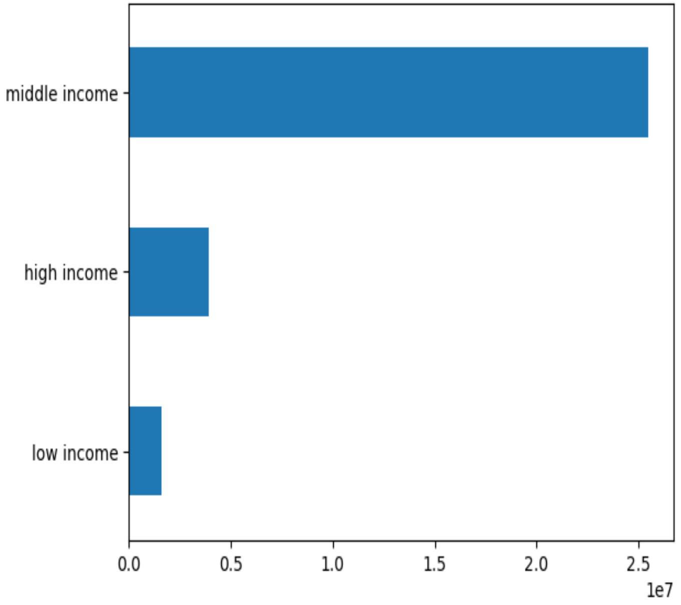


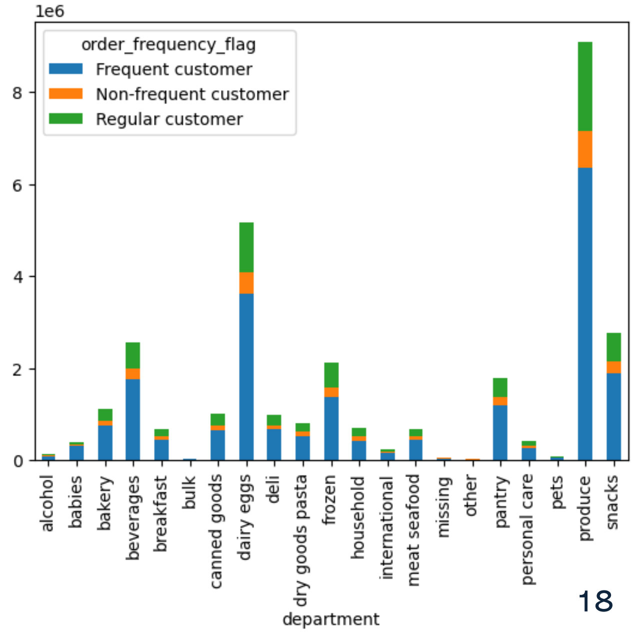












InstaCart Project Wrap Up

Derived Variables

- Customer Loyalty Flag: if order > 40 then loyal customer
- Spending Flag: if price < 10 then low spender; if >= 10 then high spender
- Order Frequency Flag: if < 10 then frequent customer
- Economic status: based on income thresholds

Recommendations

Special attention should be paid attention to the following in order to capitalize on Instacart's growth and resiliency

- Southern & Western Region
- People between the ages of 26-40 (Adults)
- People who are married
- People who make between $40,000 and $150,000 (middle-income)

Deliverables

Instacart Final Report

GitHub Repository (which includes scripts and visualizations)

Project #5: World University Rankings

Overview

Universities worldwide are the corner stone of intellect and advancement as they have a pivotal role within society to raise up the next generation of critical thinkers and promote the innovation within society.

While all schools have the mission of teaching and research, there are select few who excel in their efforts globally.

This project will be an analysis of the rankings of the world’s university in an effort to truly understand what makes the best universities so exceptional.

Data

- Open-sourced dataset from Kaggle provided by the Times Higher Education World University Ranking. Data Set

Key Skills

Python

Tableau

Correlation Analysis

Unsupervised Machine Learning

Regression & Clustering

Supervised Machine Learning

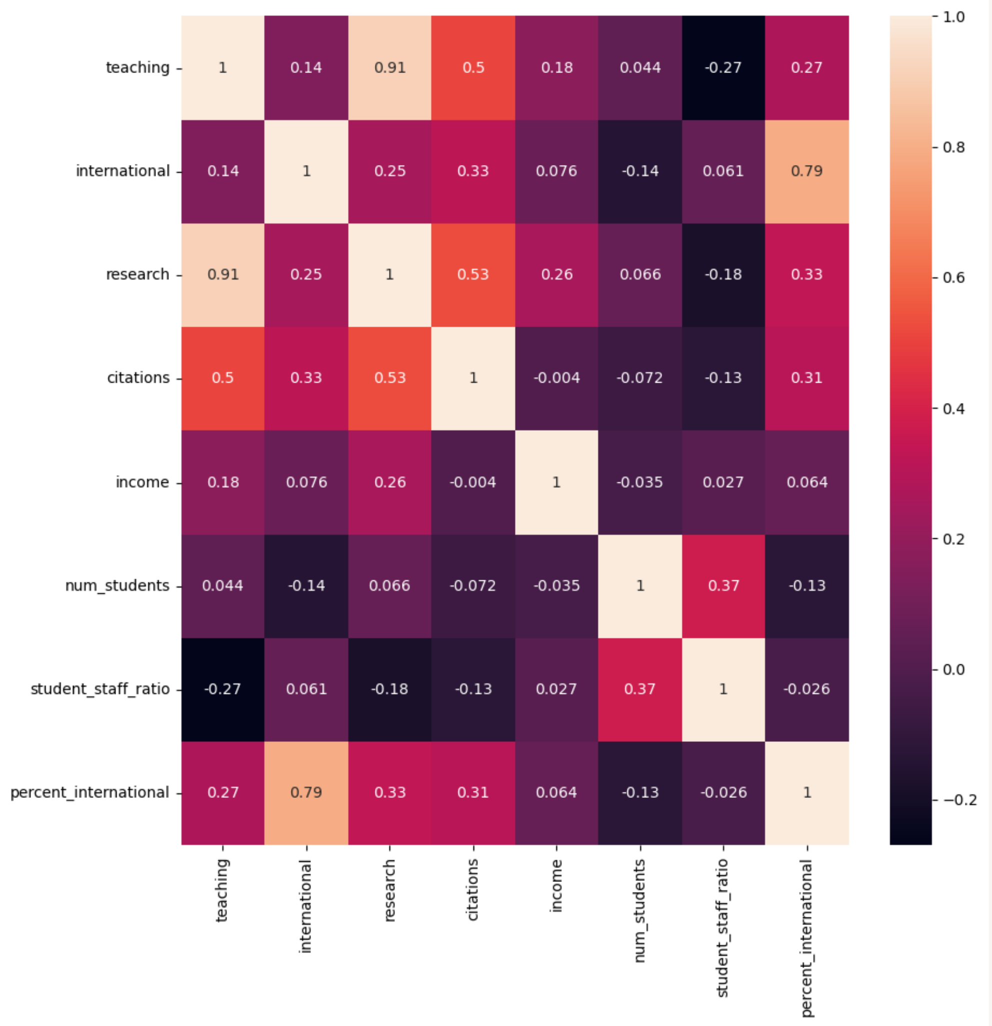
Data Wrangling and Cleaning

Data visualization

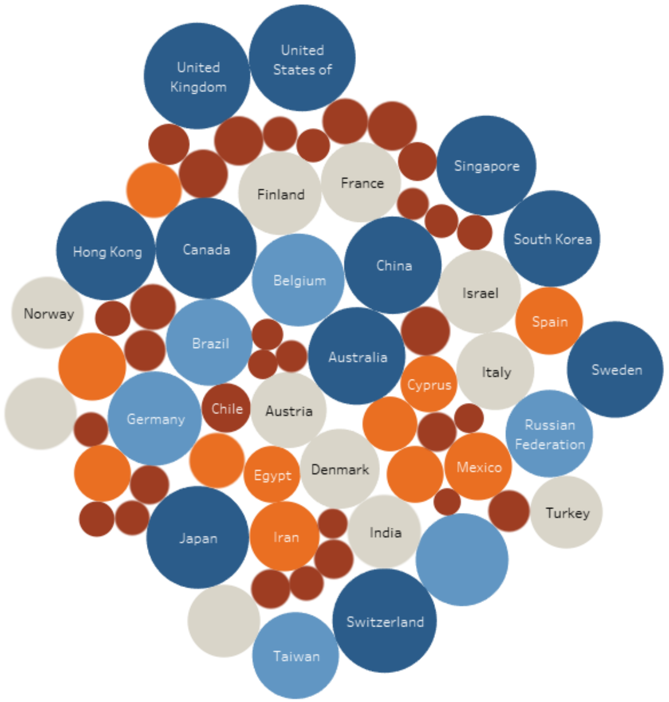
Data Sourcing

Data analysis

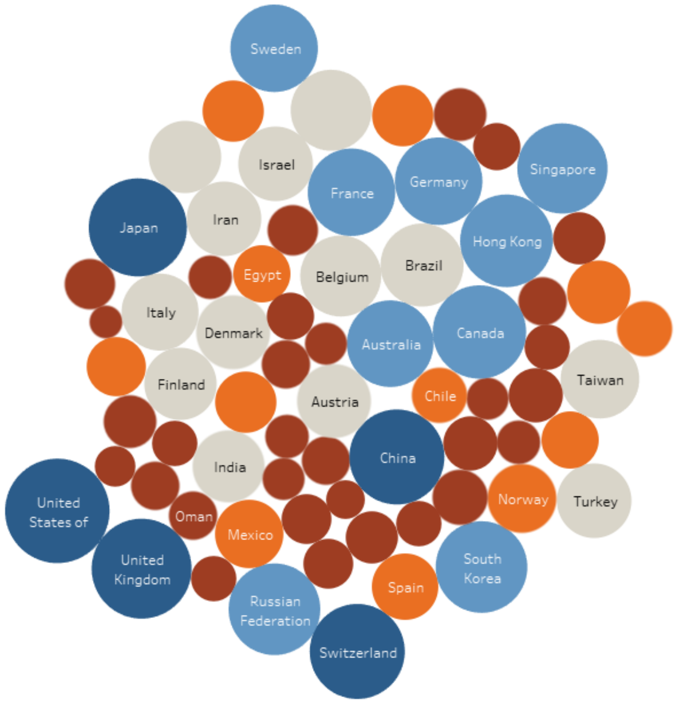
Exploratory Analysis begun by creating a correlation heatmap using seaborn within Python in order to visually see how the variables within the data correlate to each other. It was determined that the research score and teaching score share a strong positive correlation. As a result, these two variables were highlighted and later showcased within the analysis hypothesis which was, if a university has a high teaching score, then it will have a high research score.

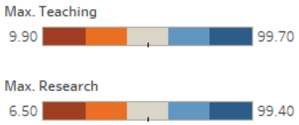


Country Max Research Score

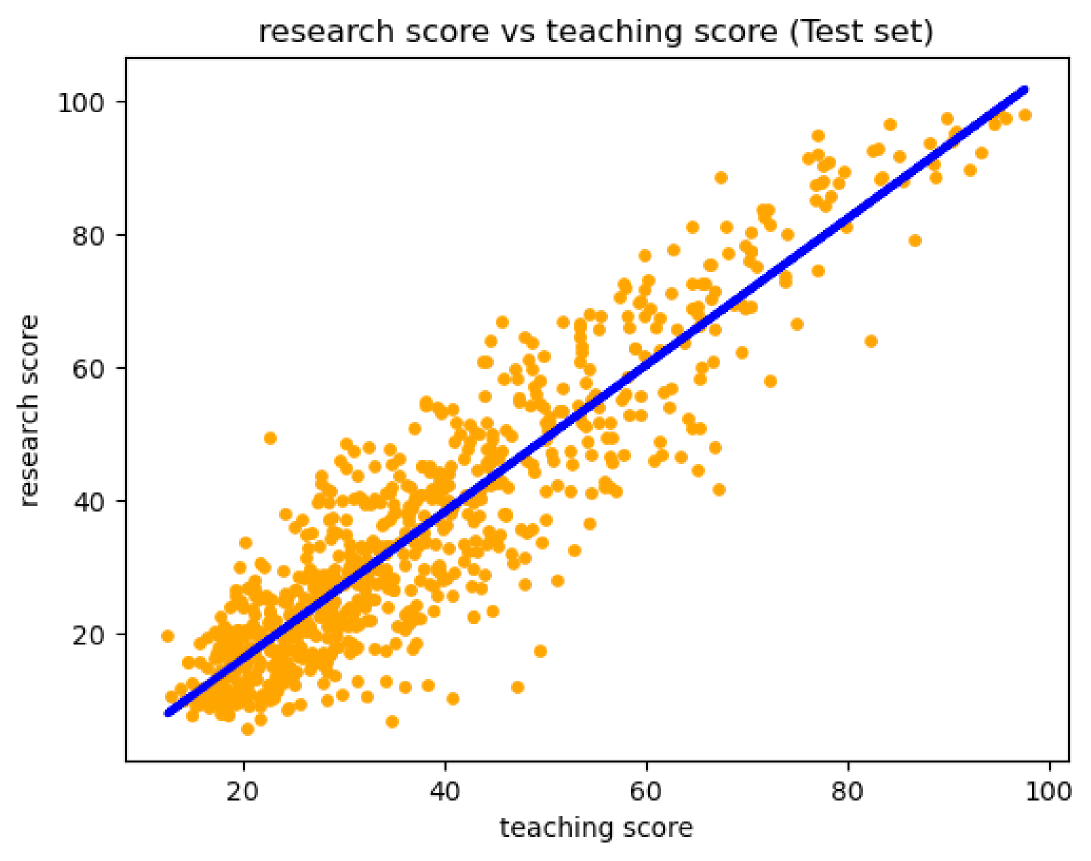


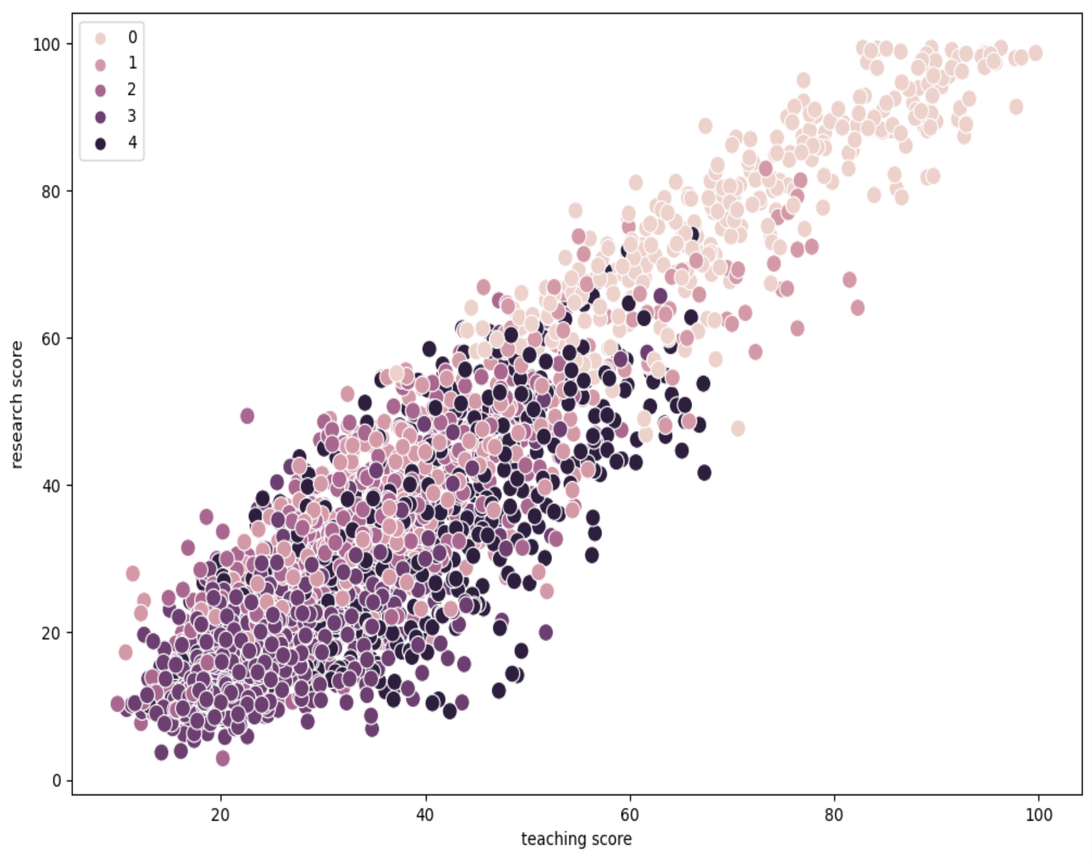
Country Max Teaching Score





Data Analysis pt.2





Conducting line regressions and cluster analysis, we are able to dive deeper into the analysis in order to illuminate trends and relationships between variables that are not easily determined.. This regression line shows that the teaching score contributes to approximately **83%** of the trend of the research score.

Achievement 6 Project Wrap Up

Results

Results: The hypothesis, if a university has a high teaching score, then it will have a high research score, has been proven to be correct. By finding the conducting various tests, linear regression and cluster analysis, in order to explore the correlation between teaching score and research score.

Recommendations

Universities who wish to increase their prestige and standing amongst the universities of the world can do audits on their personal teaching and research scores in order to raise their overall world rank.

Deliverables

Tableau Story Board

GitHub Repository (which includes scripts and visualizations)

Thank you

Samir Thomas

samirjthomas@gmail.com

LinkedIn | GitHub | Tableau