Data Analyst

Pei-Mei Lee

Based in Stuttgart, Germany

CITYWIDE MOBILITY SURVEY | insights & regression analysis

INSTACART | marketing strategy for an online grocery store

ROCKBUSTER | business insights for an online video rental service

INFLUENZA | preparing for flu season in the U.S.

GAMECO | analyzing global video game sales

All logos were downloaded from worldvectorlogo.com.

NEW YORK CITYWIDE MOBILITY SURVEY

Photo by Tim Hüfner on Unsplash

Project Intro

- The New York City Department of Transportation conducted an annual travel survey named the Citywide Mobility Survey between 2017 and 2020.
- It aims to assess the travel behavior, preferences, and attitudes of New York City residents.
- This project focuses solely on the survey dataset of 2019.

My Tasks

✓ Discover New York residents' transportation needs and preferences via exploratory visual analysis and make suggestions for improvement

✓ Examine the relationship between trip distance and trip duration via regression and clustering analysis

NEW YORK CITYWIDE MOBILITY SURVEY

DATA

- 5 CSV files and 1 GeoJSON file
- Survey data collected and preprocessed by a consulting company RSG
- Data source: nyc.gov

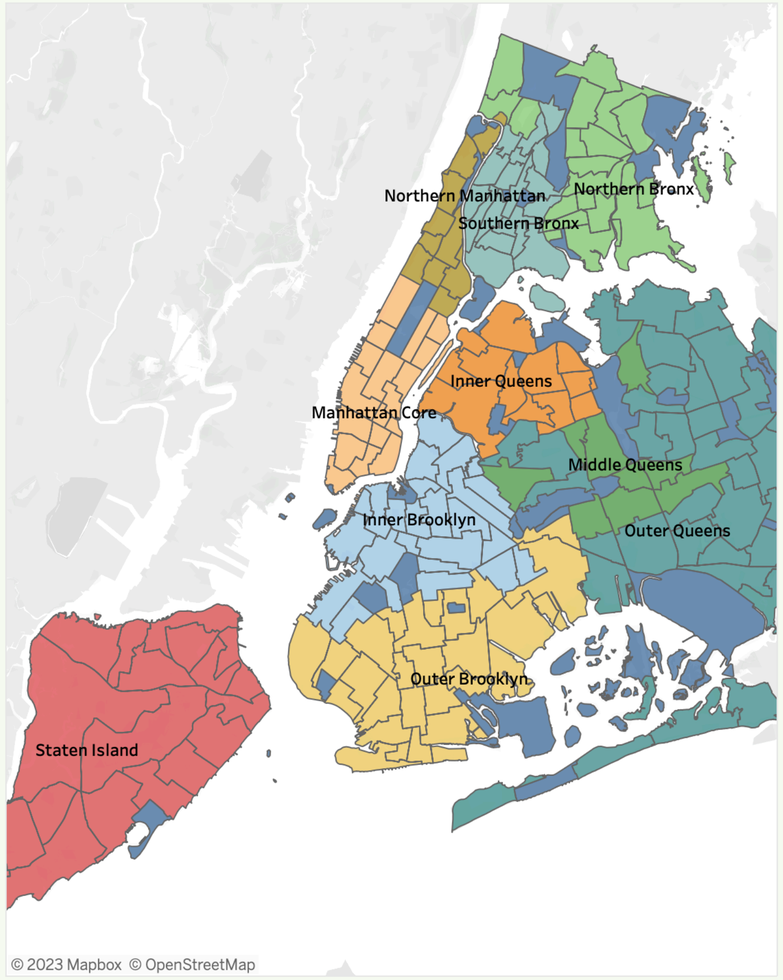
TOOLS

- Anaconda / Jupyter
- Python
  - ✓Pandas
  - ✓Numpy
  - ✓Seaborn
  - ✓Matplotlib
- Excel
- Tableau
- ChatGPT
- GitHub

KEY SKILLS

- Sourcing open data for own project
- Exploratory visual analysis
- Using Python to do the following:
  - ✓Data cleaning & wrangling
  - ✓Creating correlation heatmap
  - ✓Regression analysis
  - ✓Clustering analysis
- Tableau storyboard reporting

NEW YORK CITYWIDE MOBILITY SURVEY



About the Survey

- All participants are residents of New York City.

- 3.346 residents participated in the survey and are included in the orriginal dataset.

- 3.009 of the particiants are presented in this analysis as a result of data preprocessing.

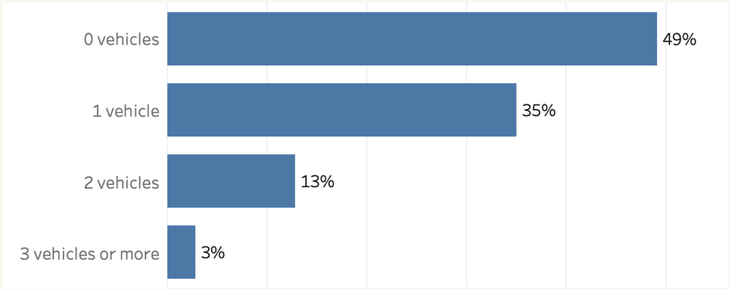
- 75% participated in the survey by smartphone, 20% participated online, and 5% participated through the survey call center.

- The survey took place between May 22 and June 30, 2019.

Survey Participants per Zone of Residence

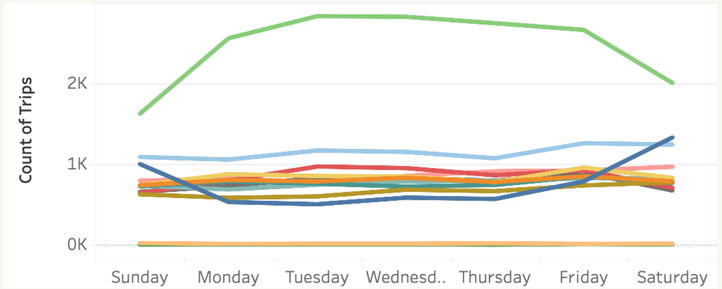
NEW YORK CITYWIDE MOBILITY SURVEY

Vehicle Ownership



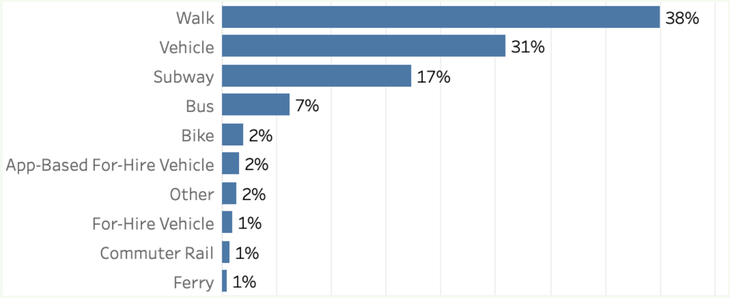
- Almost half of the participants don't have a car home.
- 16% of the participants have more than one car.

The Most Popular Destination



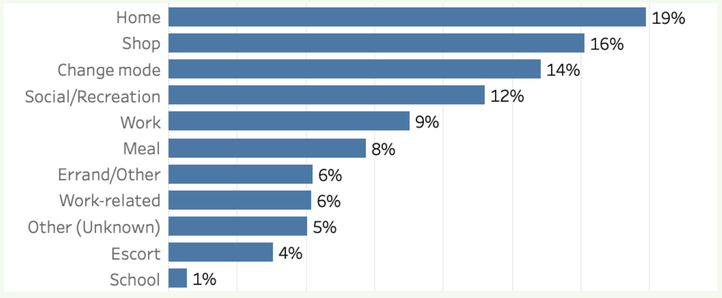
- Manhattan Core is the most popular destination throughout the week.
- Traffic leaving New York City increases during weekend.

Travel Mode Share



- Most popular choice of mode is walking, followed by driving own car or company car.
- 26% were by public transit (subway, bus, commuter rail & ferry).

Trip Purposes

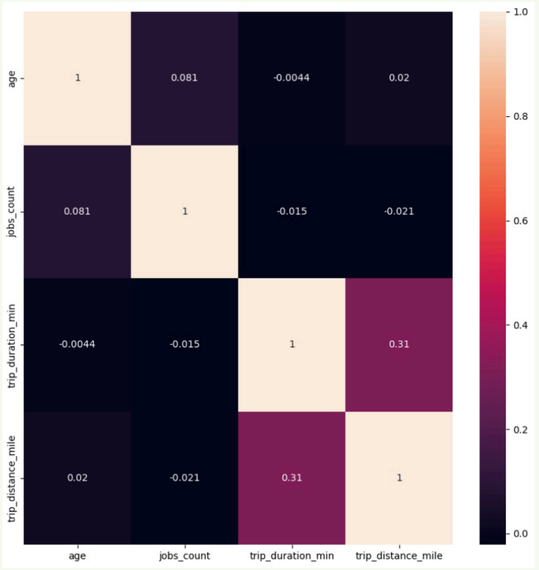


- 19% of trips were trips home
- 16% of trips were to go shopping
- 15% of trips were to go to work or were work-related.

NEW YORK CITYWIDE MOBILITY SURVEY

Summary of Regression Analysis

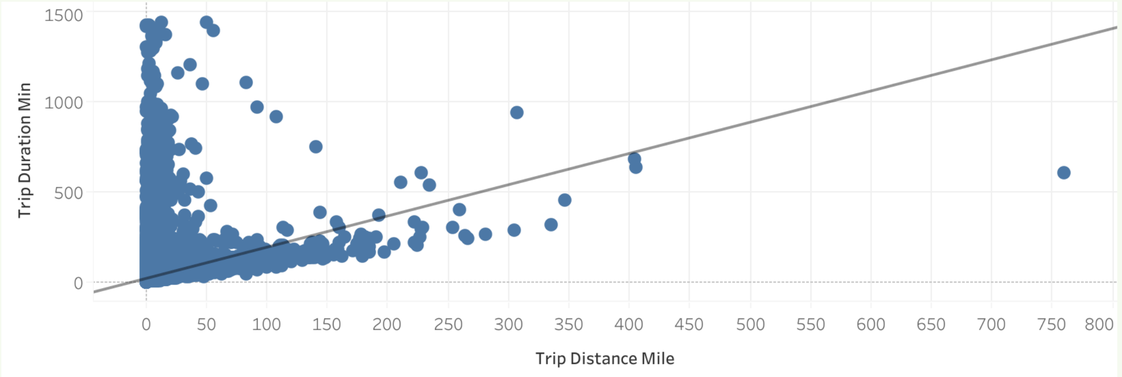
Continuing the exploratory analysis, a correlation heatmap was created to see if any numerical variables have a strong correlation with trip distance or duration.



Interpretation of Correlation Heatmap

- The trip distance and trip duration have a week-moderate correlation.
- All the rest are values close to zero, which indicates there is no correlation.

A regression analysis was then performed to access the relationship between trip distance and duration. **My hypothesis:** The longer the trip distance is, the longer the trip lasts in time.



Model Performance

- The MSE is at around 2738, which indicates that the regression line isn't an accurate representation of the data and can't accurately predict the influence of trip distance on the trip duration.
- The R2 score is around 0.09, which is low. This tells us that the model doesn't explain the variance in the data well. It is a poor fit.

Conclusion

Based on the statistical results, we can say the relationship between trip distance and duration doesn't follow a single, straight regression line. I can rule out my above stated hypothesis. To accurately represent the data, new hypotheses need to be formed to find the best fitting model.

NEW YORK CITYWIDE MOBILITY SURVEY

What affects trip distance and duration?

As it turns out, factors like **mode of transportation**, **where one lives** and **the day of week** all have some influence on the trip distance and duration.

Averagely speaking...

- The longest trip duration is by ferry (42 min.), the shortest is on foot (12 min.).
- If you live in MiddleQueens, your trip takes longer time than anybody else's.
- Weekday trips are shorter in distance than those during weekend.

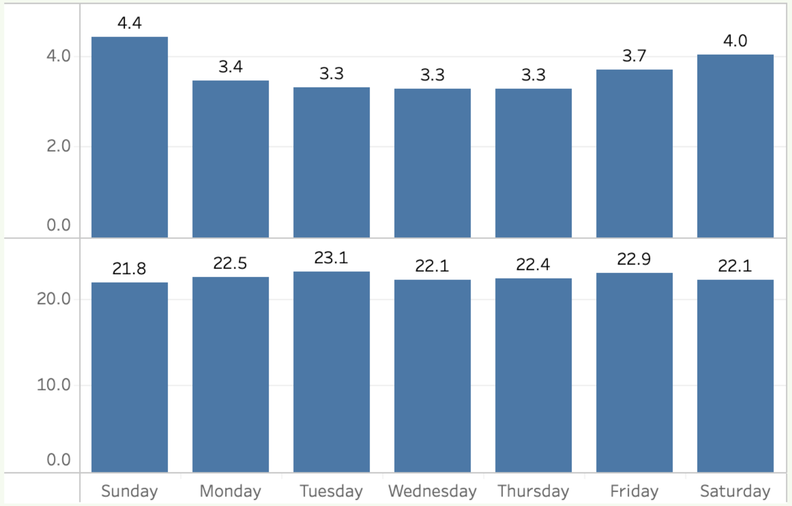
Use the dashboard below to gain further insights such as...

- Residents of Staten Island travel the longest distance by bus.
- Residents of Manhattan Core drive longest distance in own vehicles.
- Outer Queens travel the longest distance whereas Manhanttan Core travel the shortest by subway.

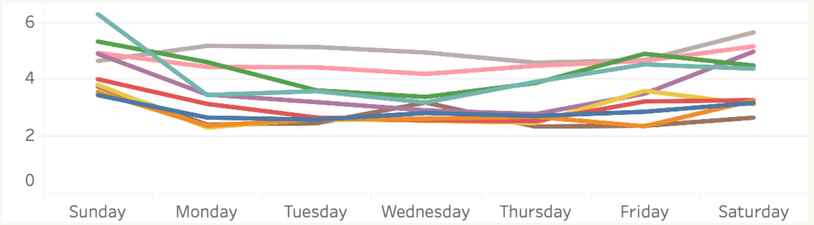
Inner Brooklyn
Inner Queens
Manhattan Core
Middle Queens
Northern Bronx
Northern Manhattan
Outer Brooklyn
Outer Queens
Southern Bronx
Staten Island

*** Click around to Generate Insights! ***

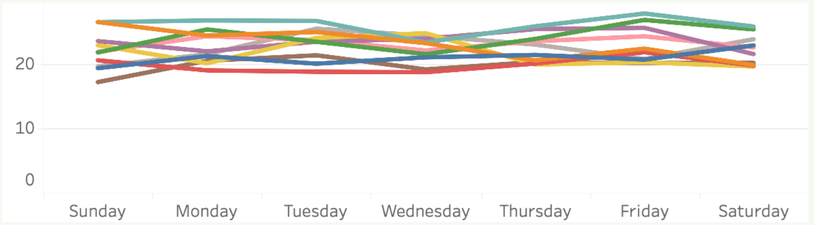
Citywide Average



Average Trip Distance (in Miles)



Average Trip Duration (in Minutes)



Home Zone

- [x] (All)
- [x] Inner Brooklyn
- [x] Inner Queens
- [x] Manhattan Core
- [x] Middle Queens
- [x] Northern Bronx
- [x] Northern Manhattan
- [x] Outer Brooklyn
- [x] Outer Queens

Mode Type

- [x] (All)
- [x] App-Based For-Hire Ve...
- [x] Bike
- [x] Bus
- [x] Commuter Rail
- [x] Ferry
- [x] For-Hire Vehicle
- [x] missing, skip logic
- [x] Other

NEW YORK CITYWIDE MOBILITY SURVEY

Photo by Tim Hüfner on Unsplash

Limitations of Survey Data

Data limitations need to be considered when interpreting the results.

Scope of survey participants: The survey exclusively includes New York City residents, rendering it unrepresentative of the diverse needs and behaviors of tourists in the city. (Approximately 66.6 million people visited New York in 2019.)

Seasonality: All trips took place in May or June. This limits our observations to only the participants' travel behavior and needs in Summer.

Bias: The data exhibits both sampling and geographical biases, as the demographic composition of participants deviates from the actual population distribution within the surveyed zones.

Conclusion

- The relationship between trip distance and trip duration cannot be explained solely by linear correlation.

- Factors such as mode of transportation, day of the week and where one lives all have an influence on how long the trip is, both in distance and in time.

Recommendations

- Conduct a dedicated survey focused on tourists' mobility preferences, given that there is a constant presence of many tourists in New York.

- Conduct the same survey in the other three seasons to achieve a comprehensive understanding of the year-round transportation needs and preferences.

- Prioritize efforts to improve conditions for pedestrians, as walking is the most frequently chosen mode.

- Allocate more budget for infrastructure maintenance in Manhattan Core, as it receives the most arrrivals among all zones.

INSTACART Grocery Sales Analysis

Photo by No Revisions on Unsplash

Project Intro

- Instacart is an online grocery store that operates through an app.
- Business is going well. But the management wants to know more about their customers such as their purchasing behaviors.

My Tasks

✓ Performing an initial data exploratory analysis of their data to derive insights

✓ Suggesting strategies for better segmentation based on the provided criteria.

INSTACART Analysis Overview

DATA

- 7 CSV files
- 34 columns in total
- Data Dictionary provided by jeremystan on GitHub
- The Instacart Online Grocery Shopping Dataset 2017 was accessed via Kaggle.

TOOLS

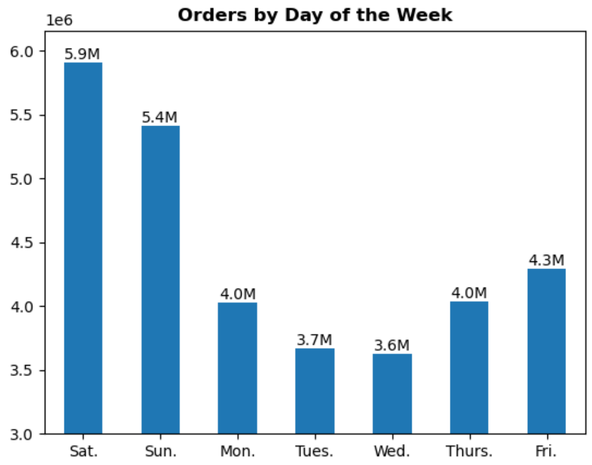
- Anaconda / Jupyter
- Python
  - ✓ Pandas
  - ✓ Numpy
  - ✓ Seaborn
  - ✓ Matplotlib
- Excel
- ChatGPT
- GitHub

KEY SKILLS

- Using Python to do the following:
  - ✓ Wrangling & merging data
  - ✓ Grouping & aggregating data
  - ✓ Visualizing data
- Creating a population flowchart
- Using Anaconda libraries manager & Jupyter Notebooks
- Creating repositories on GitHub

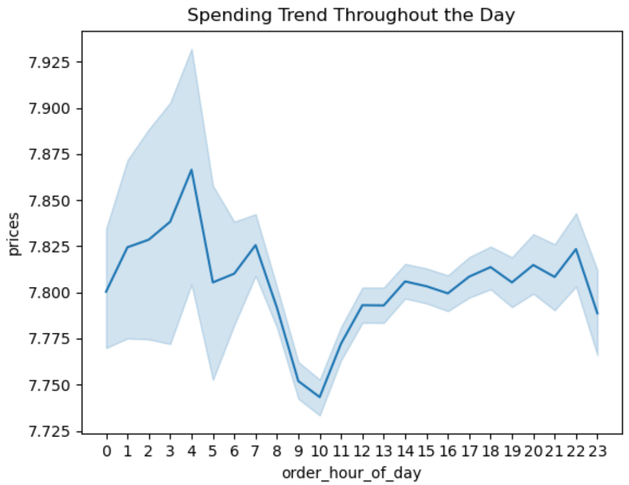
INSTACART Business Q&A

What are the busiest days of the week?



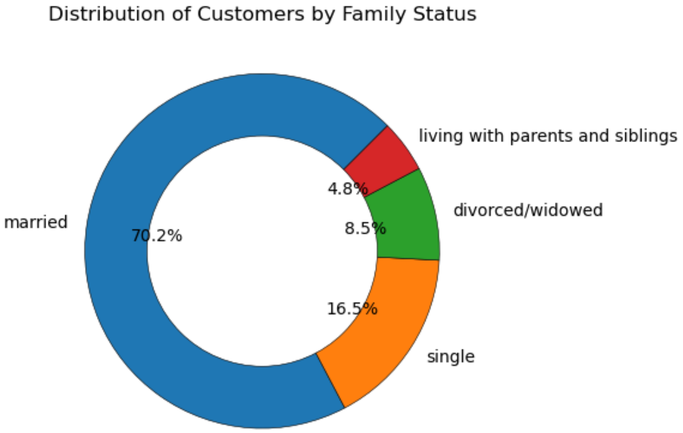
- Saturday and Sunday are the busiest days of the week.
- Wednesday is the least busy day of the week.

Are there particular times of the day when people spend the most money?



- Customers spend the most money in the late night hours, especially from 3am to 4am.
- This would be the ideal time to advertise pricier products.

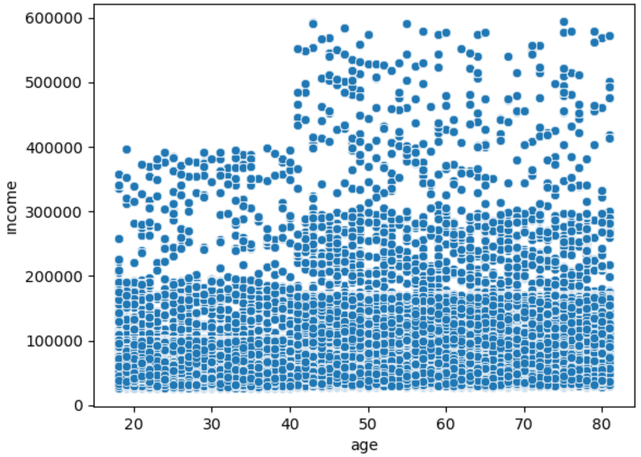
Are our customers single or married?



- The majority (70%) of our customers are married.

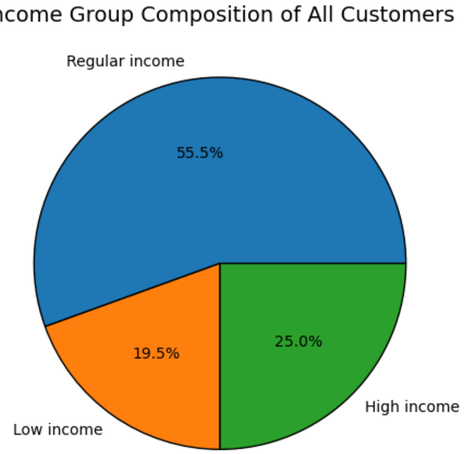
INSTACART Business Q&A

What's the relationship between a customer's income and age?



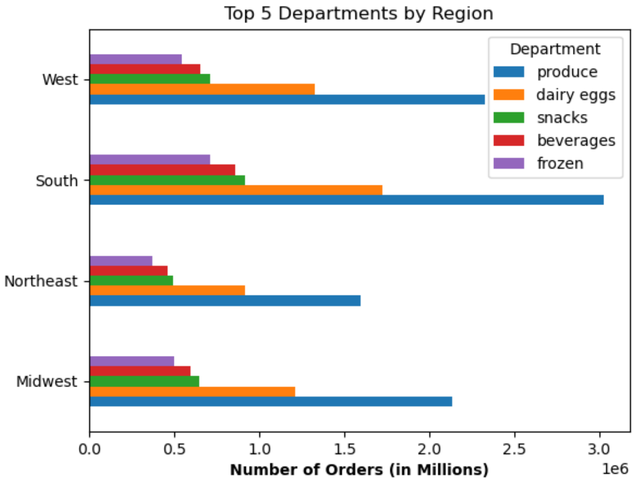
- The majority of customers have an annual income below $200K.
- A cluster of customers aged over 40 earn an income between $200K and $300K.
- Max. income for < 40 yr: ~ $400K
  Max. income for > 40 yr: ~ $600K

Customer segmentation: What kind of income do our customers earn?



- More than half of our customers are regular-income earners.
- One quarter of our customers are high-income earners.

What are the top 5 bestselling product departments in each region?



- Top 5 bestselling departments: Produce, dairy eggs, snacks, beverages and frozen
- No differences are found in terms of top 5 best-selling departments between regions.

ROCKBUSTER Online Movie Rentals

Photo by Glenn Carstens-Peters on Unsplash

Project Intro

- Rockbuster Stealth LLC is a movie rental company that used to have stores worldwide.
- Confronted by intense competition from streaming services, the management plans to use current movie licenses to introduce an online movie rental service to remain competitive.

My Tasks

✓ Use SQL to analyze the data and answer various business questions

✓ Compile analysis findings into an easily digestible format for executives

ROCKBUSTER Analysis Overview

DATA

- 16 tables
- 91 columns in total
- Data source via Postgresqltutorial

TOOLS

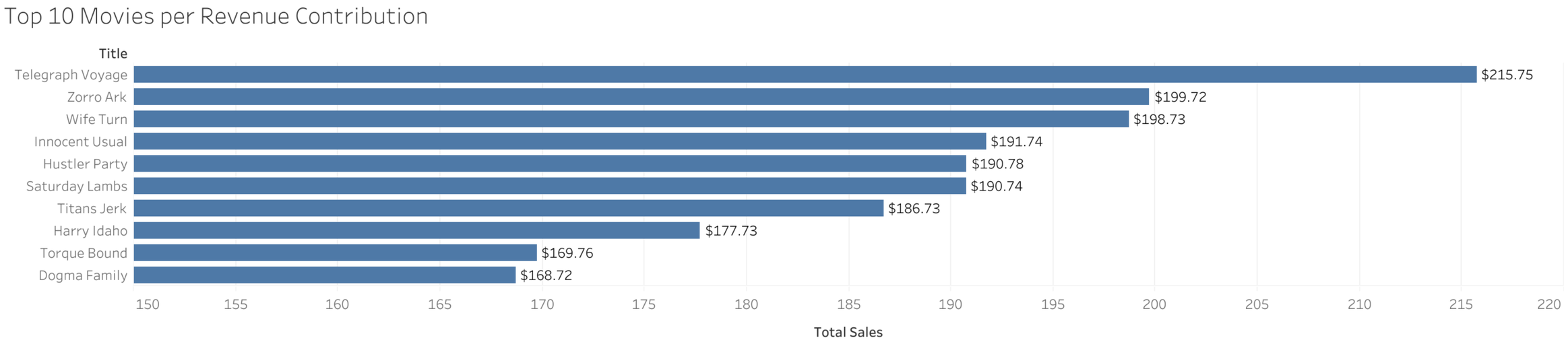
- PostgreSQL
- Excel
- Tableau
- GitHub

KEY SKILLS

- Using SQL to do the following:
  - ✓ Wrangling data
  - ✓ Joins
  - ✓ Subqueries
  - ✓ Common table expressions (CTEs)
- Data visualization in Tableau
- Creating a data dictionary
- Creating repositories on GitHub

ROCKBUSTER Business Q&A

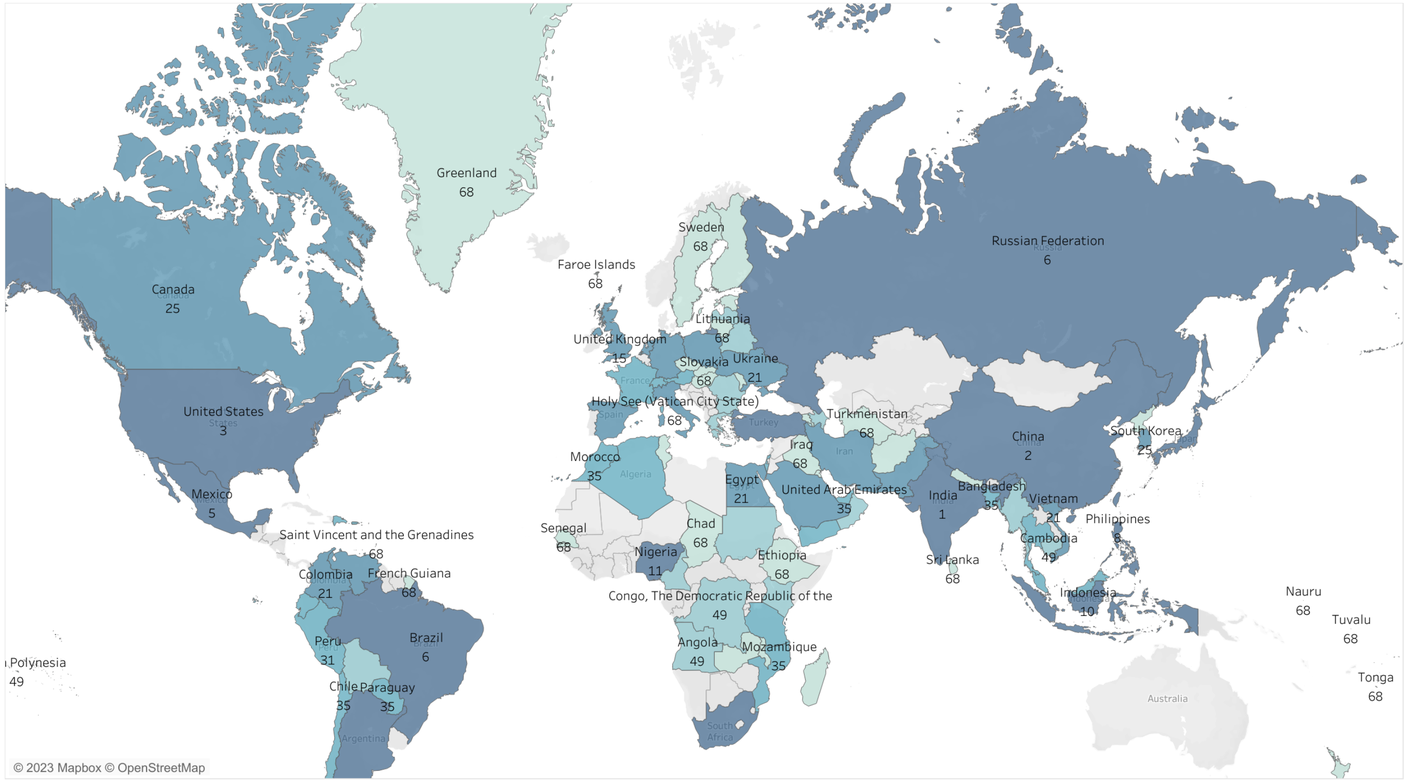
Which movies contributed the most/least to revenue gain?



- Highest-grossing movie: Telegraph Voyage ($215.75)
- Average total revenue per movie: $64

ROCKBUSTER Business Q&A

Which countries are customers based in?

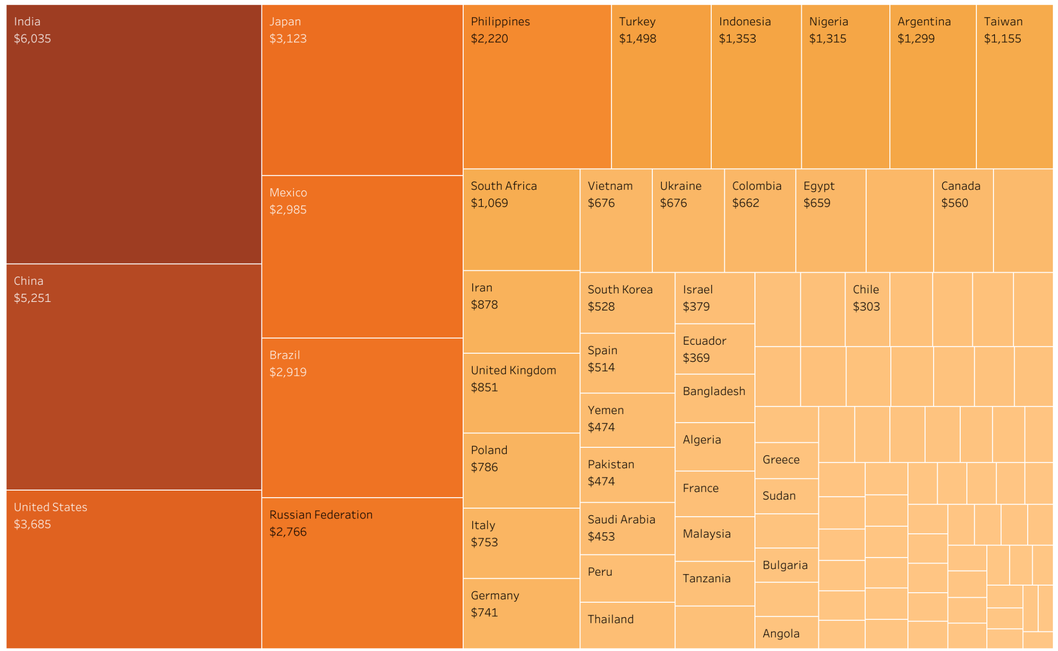


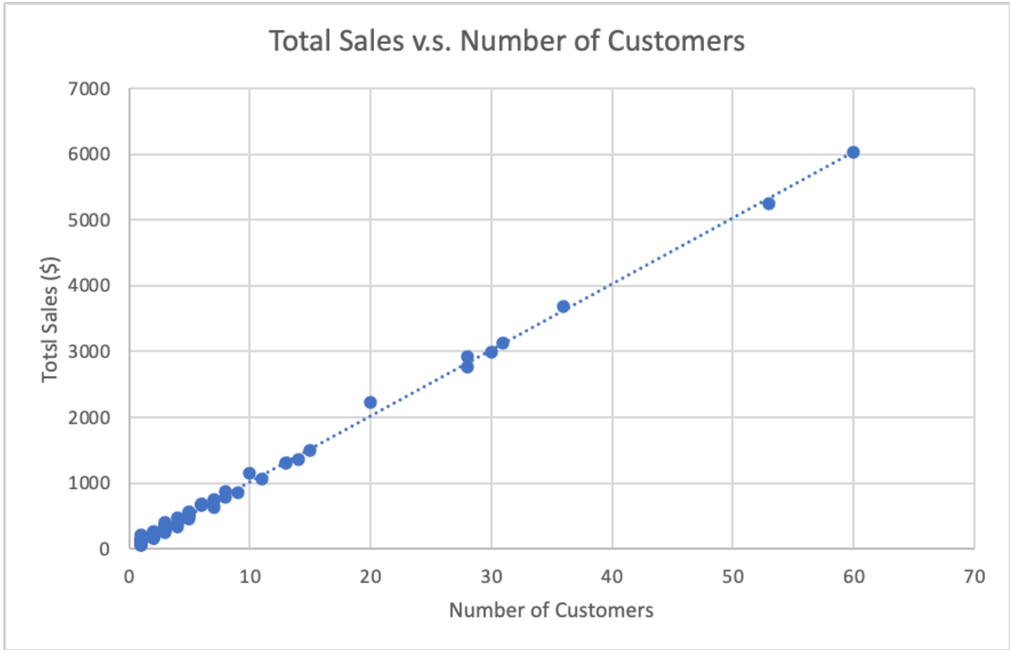
- In total: 599 customers based in 108 Countries
- Number of customers per country ranges from 1 to 60
- Top 5 countries with most customers:

- India (60)
- China (53)
- The U.S. (36)
- Japan (31)
- Mexico (30)

ROCKBUSTER Business Q&A

Do sales figures vary between geographic regions?





- Total sales per country ranges between $48 and $6,035.
- Countries with highest sales: India, China, The U.S., Japan, Mexico

- Total sales and number of customers have a positive correlation
- Increasing the customer base is paramount in all regions.

ROCKBUSTER Online Movie Rentals

Recommendations:

- Diversify movie offerings by adding movies of further languages (e.g. Hindi, Chinese, Japanese, Spanish)
- Replace low-grossing movies with new movies (a separate analysis for selection of new movies is needed.)
- Increase customer base worldwide (currently less than 10 customers in most countries!)
- Collect more recent data for analysis (All payment data are from 2007!)

Photo by Glenn Carstens-Peters on Unsplash

INFLUENZA Preparing for Flu Season

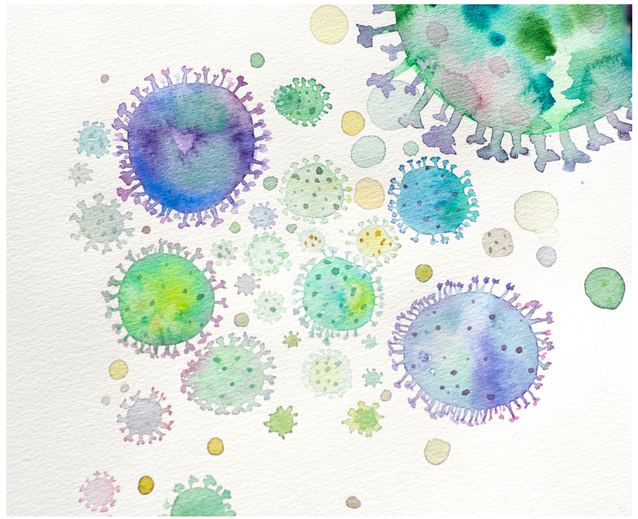


Photo by Elena Mozhvilo on Unsplash

Project Intro

- The United States has an influenza season where more people than usual suffer from the flu. A medical staffing agency provides temporary workers to clinics and hospitals on an as-needed basis.
- The agency wishes to examine trends in influenza and determine how they can plan for staffing needs across the country.

My Tasks

✓ Determine whether influenza occurs seasonally or throughout the entire year.

✓ Provide insights to support a staffing plan, detailing what data can help inform the timing and spatial distribution of medical personnel countrywide.

INFLUENZA Analysis Overview

DATA

- 5 Excel files
- Data source 1: Centers for Disease Control and Prevention
- Data source 2: US Census Bureau

TOOLS

- Excel
- Tableau

KEY SKILLS

- Descriptive statistics
- Statistical hypothesis testing
- Visual analysis
- Dealing with data limitations
- Forecasting in Excel
- Tableau storyboard reporting

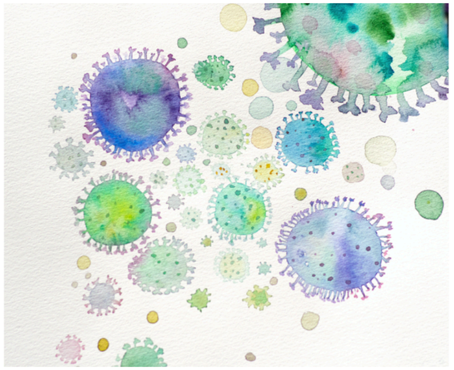
INFLUENZA Tableau Storyboard

Overview

Motivation:
- The United States has an influenza season where more people than usual suffer from the flu
- Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital
- Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff

Objective:
Determine when to send staff, and how many, to each state.

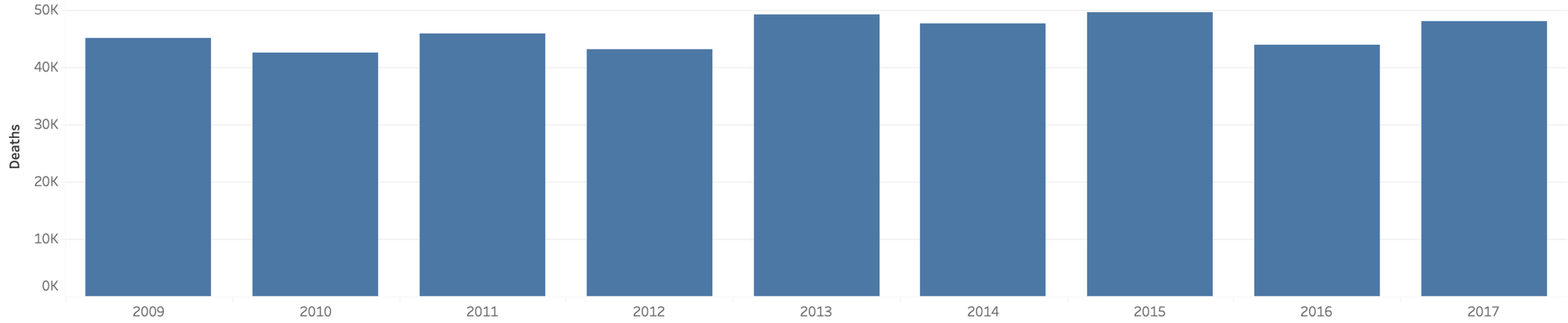
Scope:
The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.



Source: Image by Elena Mozhvilo on Unsplash

Influenza Deaths across the U.S. (2009 - 2017)

46.158 people died of influenza yearly on average.



INFLUENZA Tableau Storyboard

Vulnerable Populations

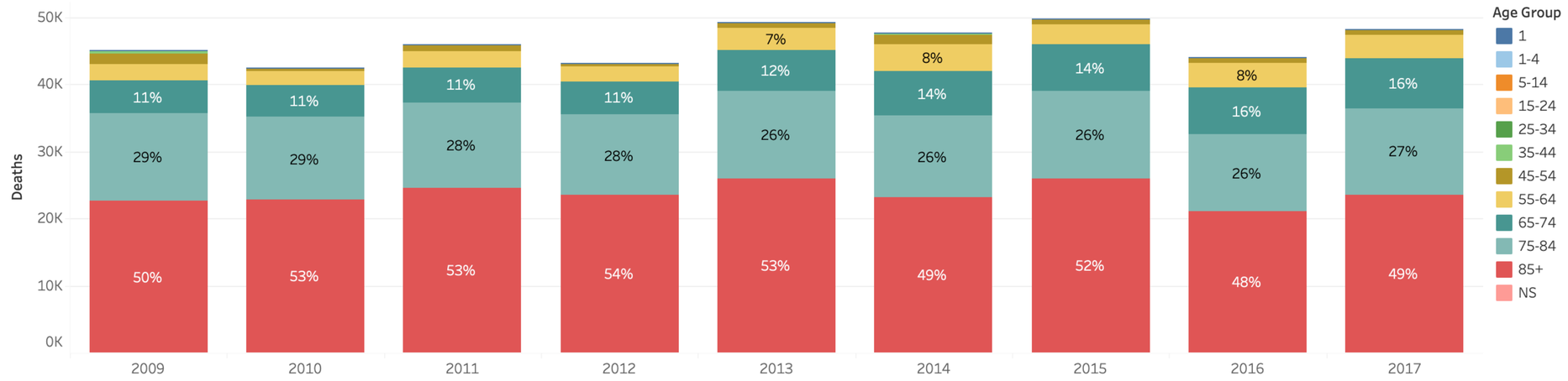
- Elderly population (aged 65 and older) has a higher risk of dying from the flu
- Additional resources are to be assigned to each state according to the size of its elderly population
- Only elderly population is discussed here among the vulnerable populations due to data limitations

Source: Image by Jacek Dylag on Unsplash

Who is vulnerable to influenza?
Vulnerable populations refer to patients likely to develop flu complications requiring additional care, as identified by the Centers for Disease Control and Prevention (CDC). These include adults over 65 years, children under 5 years, and pregnant women, as well as individuals with HIV/AIDs, cancer, heart disease, stroke, diabetes, asthma, and children with neurological disorders.

Influenza Deaths by Age Group (2009 - 2017)

Around 90% of the influenza deaths fall into the age groups of 65 years and older



INFLUENZA Tableau Storyboard

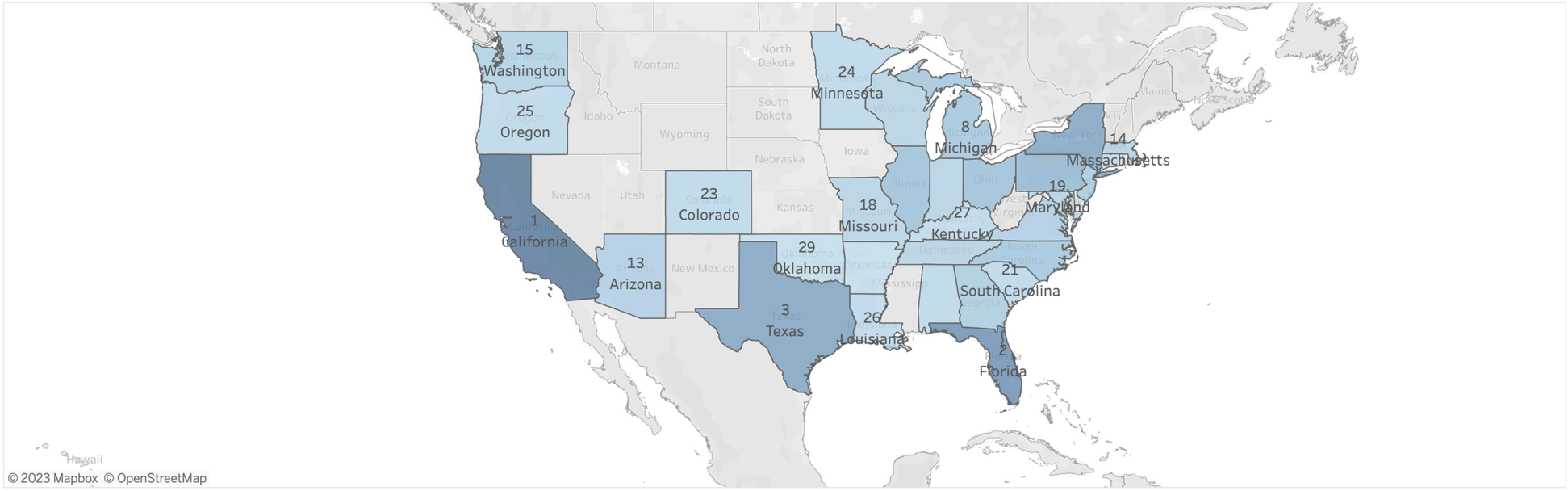
Where are the Elderly Populations?

- To define where to send additional medical staff, we need to identify where the elderly populations (65 and older) are located.
- In 2017, California, Florida, Texas, New York and Pennsylvania are the top 5 states with the most elderly population, ranging from 2.171.552 and 5.078.704.

Slide to See Which States Have the Largest Elderly Populations

1 30

Ranking in 2017: Top States with the Largest Elderly Population



INFLUENZA Tableau Storyboard

Seasonality of Influenza

Looking at the historical data from 2009 to 2017, we can establish the following:
- In most states, influenza occurs seasonally and its activity rises in September and peaks in January
- Hawaii and Florida, for instance, have a year-round flu activity
- This factor is to be reflected in the staff planning

Deaths
30 153

(All)
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire

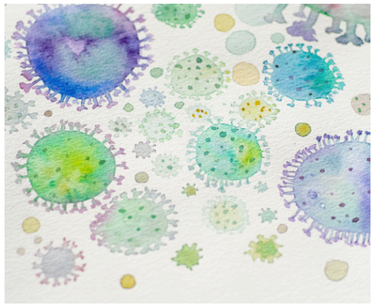
Choose a State to See Its Flu Activity! ->

Deaths by Month

	2009	2010	2011	2012	2013	2014	2015	2016	2017
Jan.	80	102	120	72	142	153	153	86	119
Feb.	53	63	80	54	80	105	127	93	92
Mar.	72	71	71	67	88	59	101	106	121
Apr.	46	57	36	61	88	52	99	59	72
May	53	63	66	71	50	59	60	74	80
Jun.	54	64	52	72	44	51	43	51	63
Jul.	61	45	59	37	32	51	58	74	80
Aug.	30	59	50	69	45	54	51	50	63
Sep.	54	46	61	40	40	41	51	55	83
Oct.	82	69	46	40	56	53	74	54	84
Nov.	84	73	64	59	83	75	75	75	72
Dec.	86	97	71	119	113	134	85	98	115

INFLUENZA Tableau Storyboard

Conclusion

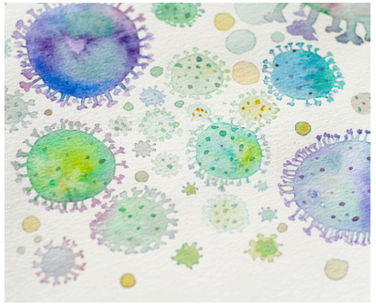


- Prioritize the elderly population (age 65 and older)

- Top 5 states with highest elderly population: California, Florida, Texas, New York and Pennsylvania

- Flu activity rises in September and peaks in January in most states; Active year-round in few states

Recommendations



- Distribute additional staff according to size of state’s elderly population

- Timing for dispatching staff: Ahead of September

- Consider sending additional staff twice a year to states with year-round flu activity

- Research and analysis on further vulnerable populations, such as children under 5, to be conducted

GAMECO Global Video Game Sales

Photo by Nikita Kachanovsky on Unsplash

Project Intro

- GameCo is a new video game company that wants to use data to inform the development of new games.
- Company executives are open to hearing any insights pulled from the data.

My Tasks

✓ Perform a descriptive analysis of historical data on global video game sales

✓ Derive business insights for allocating market budgets

GAMECO Analysis Overview

DATA

- 1 excel file
- 11 columns x 16.599 rows
- 37 years of sales data (1998 - 2016)
- Data source: VGChartz

TOOLS

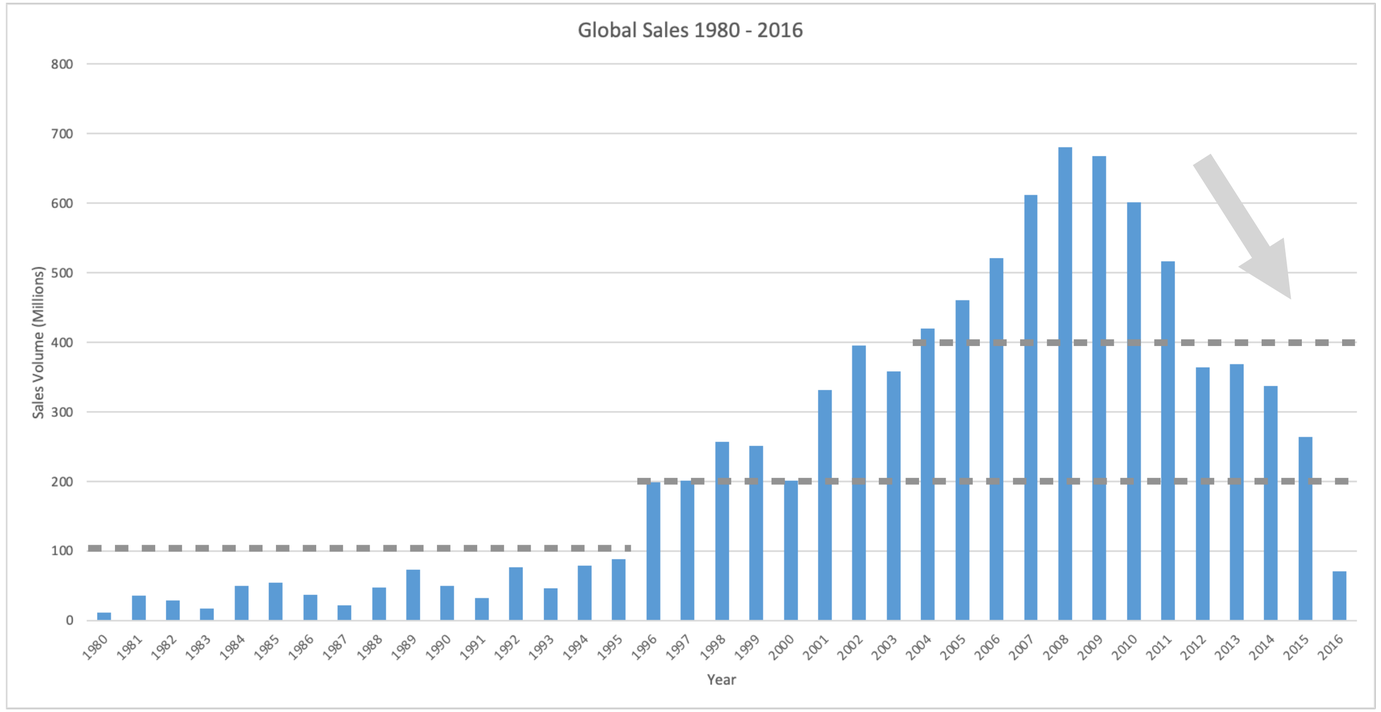
- Excel

KEY SKILLS

- Understanding business problems
- Using Excel to do the following:
  - ✓Summary stats
  - ✓Data cleaning
  - ✓PivotTables: Grouping & summarizing data
  - ✓Data Visualization
- Formulating hypotheses
- Using data to (dis)confirm hypotheses

GAMECO Business Insights

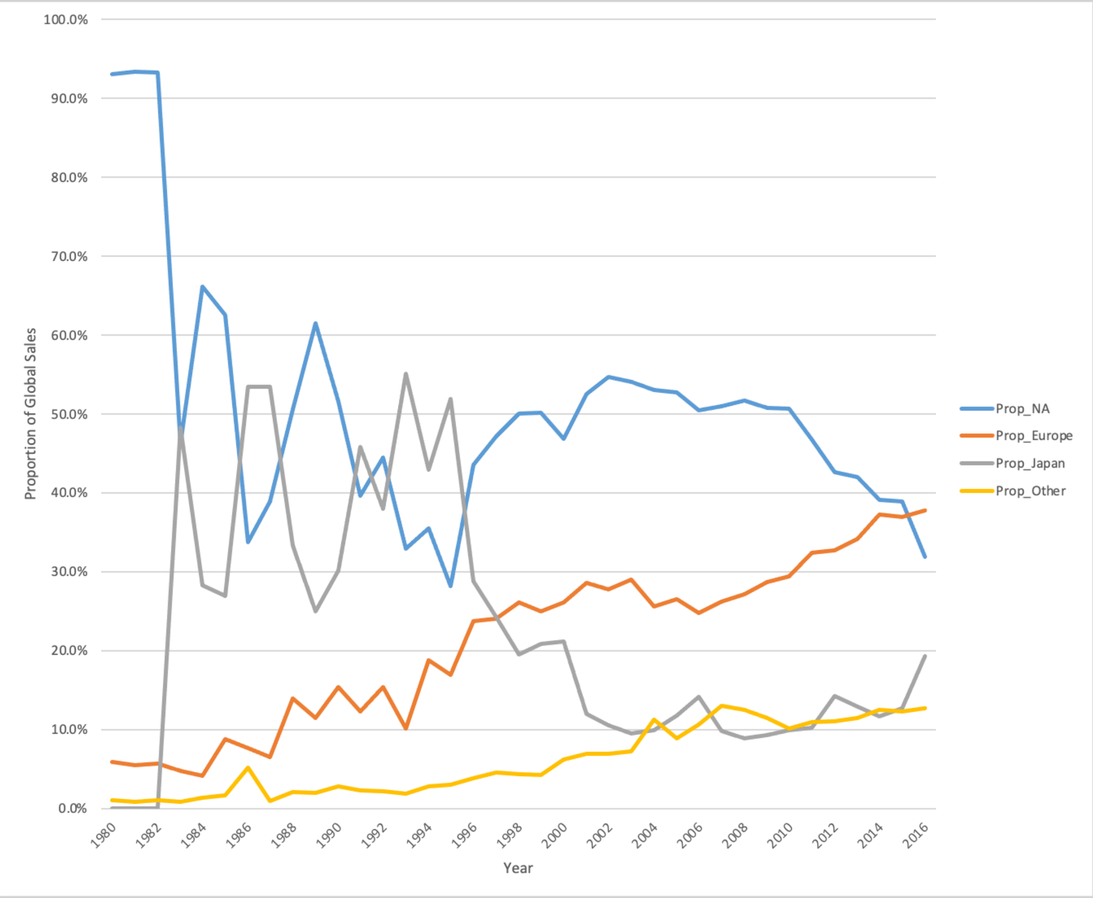
Global Sales Trends



- Total sales volume fluctuated and remained below 100 million units between 1980 and 1995.
- In 1996, total sales volume doubled and neared 200 million units. It remained above this level ever since up until 2016 (as of data cutoff).
- Since 200-million-mark, total sales fluctuated and doubled again in 2004. It continued to grow and peaked in 2008 at almost 680 million units.
- After the historical peak, total sales dropped below 400-million-mark in 2012 and still continues to fall ever since-except a mild climb in 2013.

GAMECO Business Insights

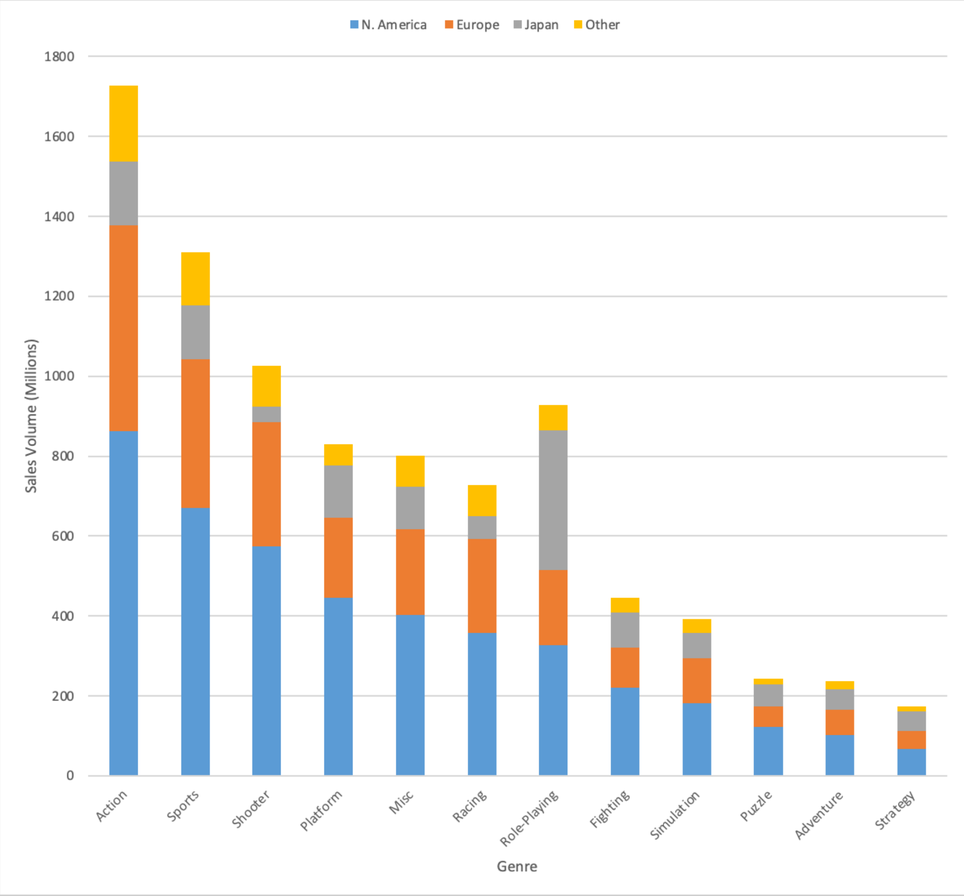
Global Sales Share by Region



- North America and Japan had a fierce competition in the 80s and early 90s, taking turns to be the leader in global sales.
- All the while Europe zigzagged its way up to the top, taking over Japan in 1998 and surpassing North America to claim the crown in global sales this year.
- Since mid 2000s, other regions combined make up around 10% of global sales, overall a promising uptrend worth further investigation alongside Europe.

GAMECO Business Insights

Genre Popularity by Sales Volume



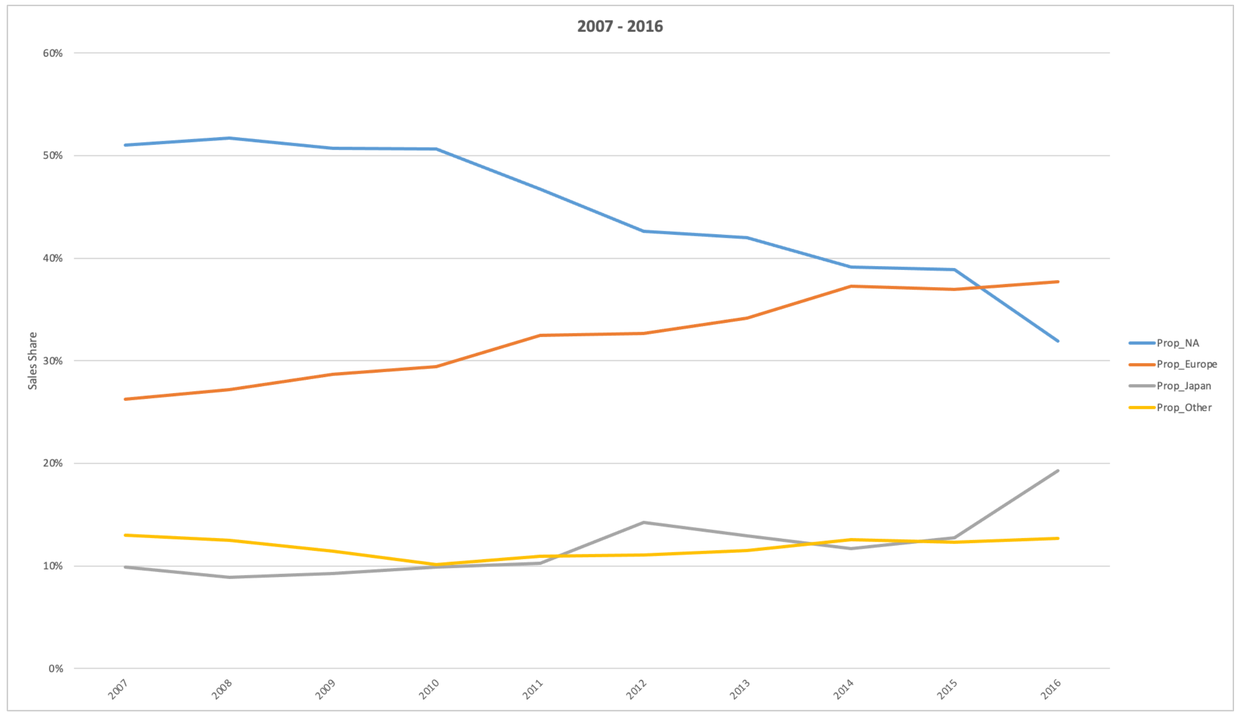
Insights:

- Top 3 genres are Action, Sports and Shooter worldwide.
- North America and Europe share very similar preferences (Action, Sports, Shooter)
- Japan shows a strong liking for Role-Playing, Action and Sports.

Recommendation: Focus our product development on Action, Sports and Shooter games as 1.) these are historically the most popular genres worldwide AND 2.) both Europe and other regions (our future sales driving forces) indicate preferences that match the global trends.

GAMECO Business Insights

Sales Trend in the Past Decade (2007 - 2016)



Insights:

- Europe will surpass the North America this year, assuming Q4 will continue the ongoing uptrend.
- Japan indicates an overall increasing contribution to the global sales in the past decade.
- Other regions combined have been accounted for more and more of the global sales since 1986.
- Its combined sales share is well above 10% in the past decade, signaling a quiet uptrend with lots of potentials.

Recommendation:

✓ Allocate more marketing resources to Europe and Japan.

✓ Invest to identify emerging markets.

Thank you.