Data Scientist

Heidi Hodges

Video Game Sales Analysis

Premise (from the briefing): A new video game company is looking to use data to inform the development of new games.

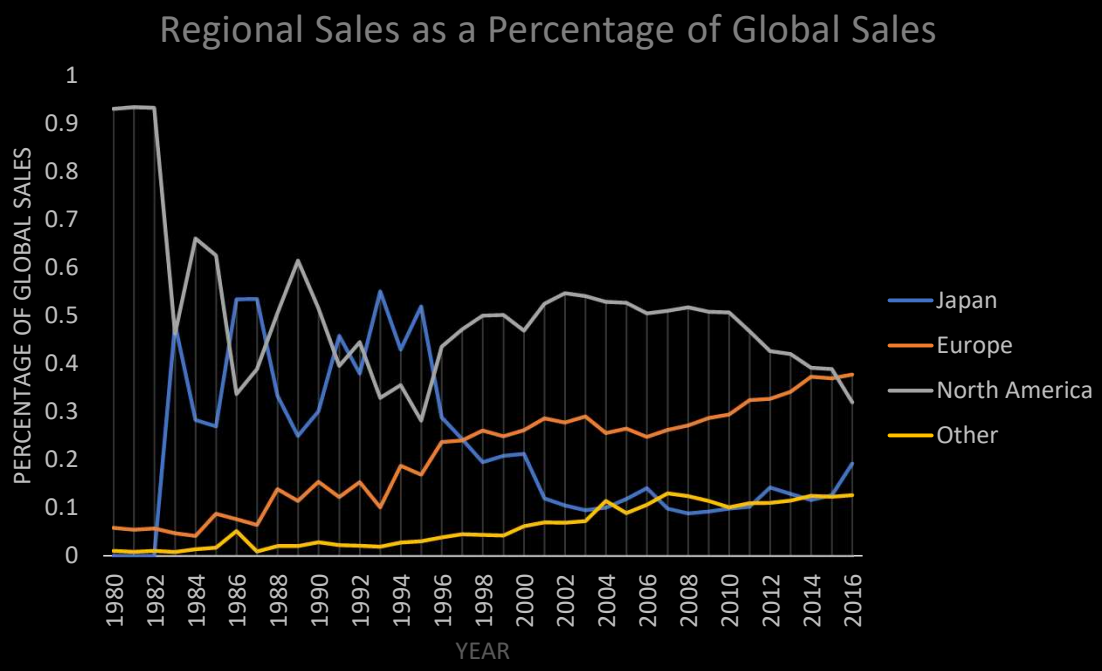
Key Questions

- 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: video game sales from https://www.vgchartz.com/ (data collection stops at 2016).
Tools and skills: Excel, Word, grouping and summarizing data, descriptive analyses, visualizations, presenting results.

Process: After cleaning the data, I answered broader questions so that I could then answer more concise questions

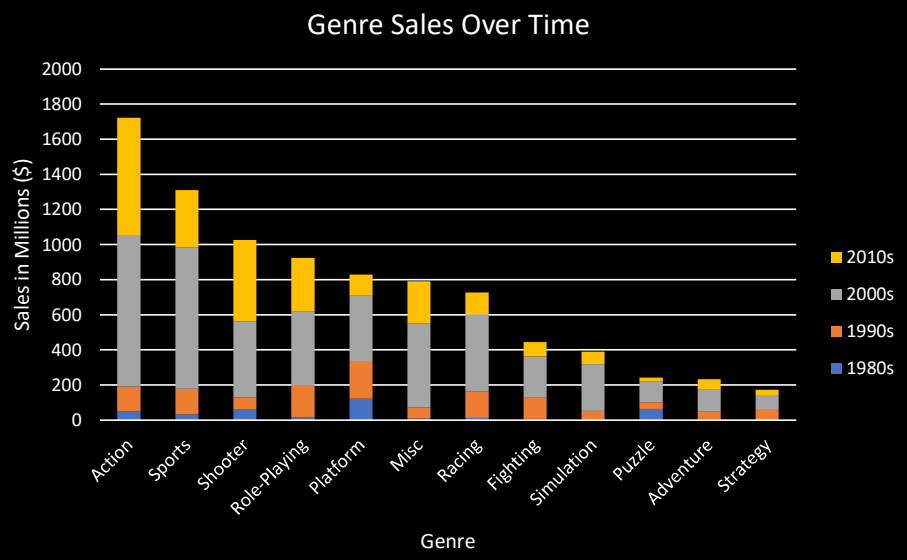
GameCo's current understanding is that video game sales remain steady across all sales regions over time



After charting sales by region over time, I determined sales are steadier as the gaming industry matures in the different regions. The European and Other sales have been increasing in percentage (and dollars) over time, while Japanese sales have been mostly steady since the mid 1990s.

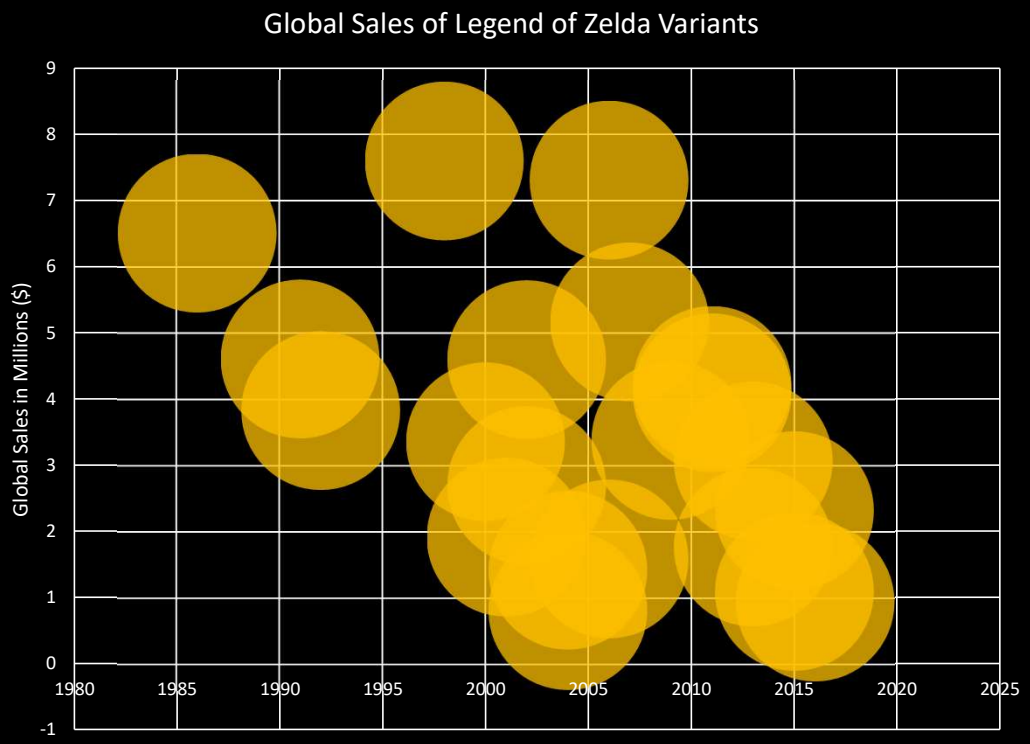
In accordance with these findings, GameCo should focus marketing efforts in Europe and North America to maximize ROI.

Next, I compared genre popularity over time



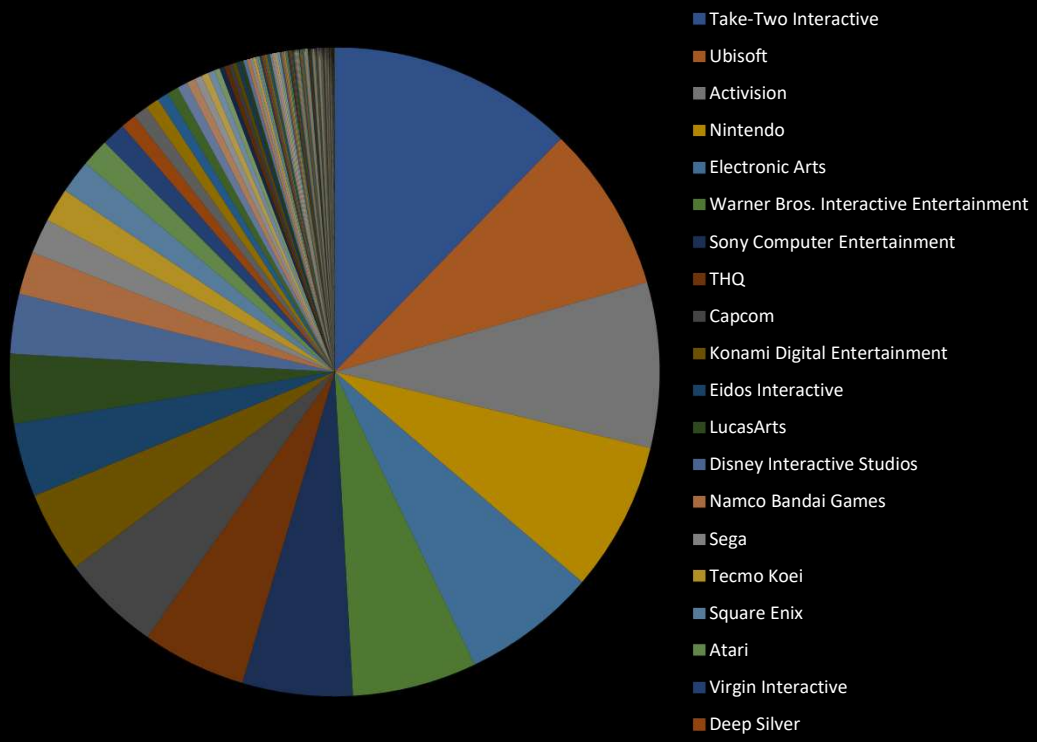
I used a Pivot table to determine that Action, Sports, and Shooter genres are the most popular over time, which I then used to inform the analysis for the next questions: individual game popularity and game staying power. I then used a line chart for each region to determine that action games were the most popular in most regions (second in Japan to Role-Playing) to further narrow down that Action games will have the largest Return on Investment.

Next, I compared game popularity over time: the three most popular games from Action games



I used bubble charts to determine that back-to-back releases and expansion packs accounted for a large portion of sales.

Finally, I compared publishers for Action games in all regions



I used pie charts to determine Take-Two Interactive, Electronic Arts, and Ubisoft were the main competitors in most regions, while Nintendo and Capcom were the main competitors in Japan.

In conclusion, sales by region have changed quite a bit since the inception of the video game industry in 1980. We should anticipate market share should continue to even out between the regions, with North American and European sales mirroring each other, while Japanese and Other sales continue to converge at a lower level. At this juncture, we recommend that GameCo focus its marketing and sales on the North American and European regions.

The top selling genres of games are Action, Shooter, Sports, and in Japanese markets, Role-Playing. Action games perform well in all regions and best in North American, European, and Other regions. We recommend that GameCo focus its efforts on developing an action game as it will have wider appeal in all regions.

Games (and their sequels and expansion packs) do vary in their popularity over time, but the best performing games have multiple releases within short periods of time with as minimal wait period between releases as possible. We recommend that GameCo have an aggressive release schedule with sequels and expansion packs developing concurrently to maximize sales and enthusiasm.

The publishers competing for sales within the Action genre vary by market but are primarily Take-Two Interactive, Ubisoft, Electronic Arts, and Nintendo.

Things I would improve: I would change using a pie chart for game publishers as there were too many options and a pie chart was not the best way to showcase the publishers. I would choose a treemap to better show the difference between bigger publishers and smaller publishers.

Flu Season Preparations in the United States

Premise (from the briefing): To help a medical staffing agency that provides temporary workers to clinics and hospitals on an as-needed basis for influenza season. Using CDC and Census information from 2011-2017 to inform future needs for additional influenza staffing.

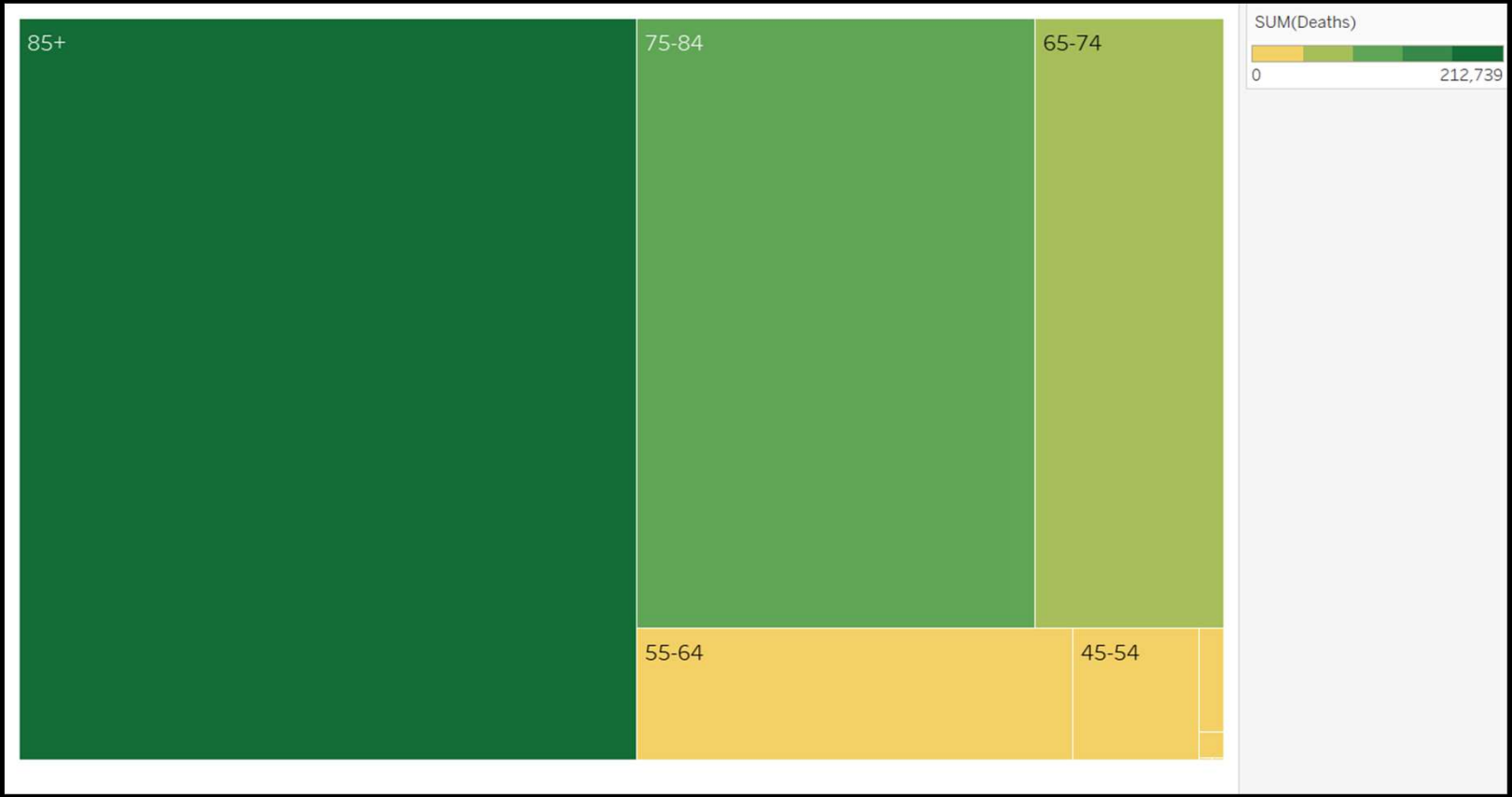
Key Questions

- Does influenza occur seasonally or throughout the entire year?
  - If seasonal, does it start and end at the same time in every state?
- Prioritize states with large vulnerable populations.
  - Vulnerable population defined as "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," does the data bear this out?
- Does poverty contribute to influenza deaths per state?

Data: data sets from US Census Bureau and CDC. Poverty and income information not broken out by age so cannot cross reference vulnerable populations with poverty. Influenza deaths have many suppressed records, especially for vulnerable persons under 5 years of age so we may not be getting all the true information.

Tools and skills: Excel, translating business requirements, data cleaning, integration, and transformation, statistical hypothetical testing, visual analysis, forecasting, storytelling in Tableau, presenting results to an audience.

I started by verifying that age contributed significantly to death rates for influenza as displayed in my Tableau presentation.



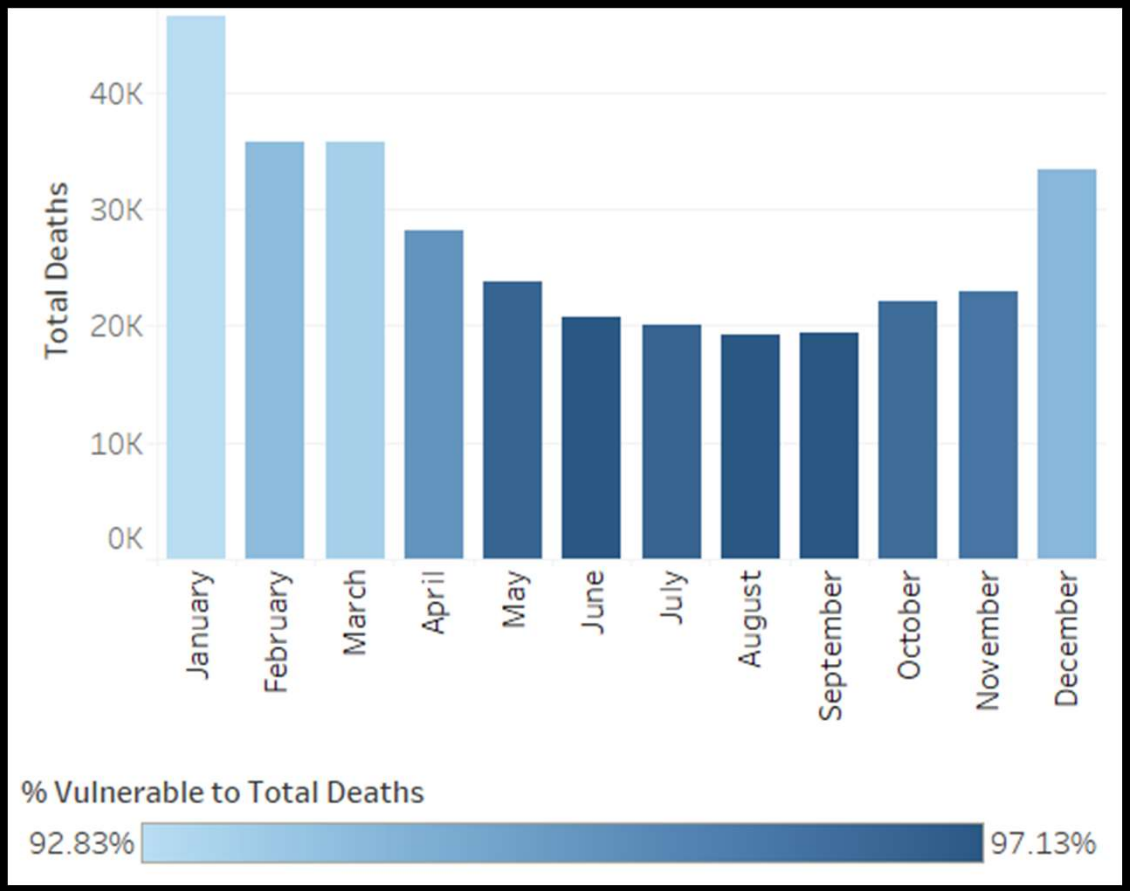
I used a treemap to show how influenza deaths double for every 10-year age increase.

I leveraged poverty data according to the US Census information against the deaths from influenza.

I used a bubble chart and used Tableau to insert the R-squared line to show that there doesn't appear to be any correlation between poverty and influenza deaths.

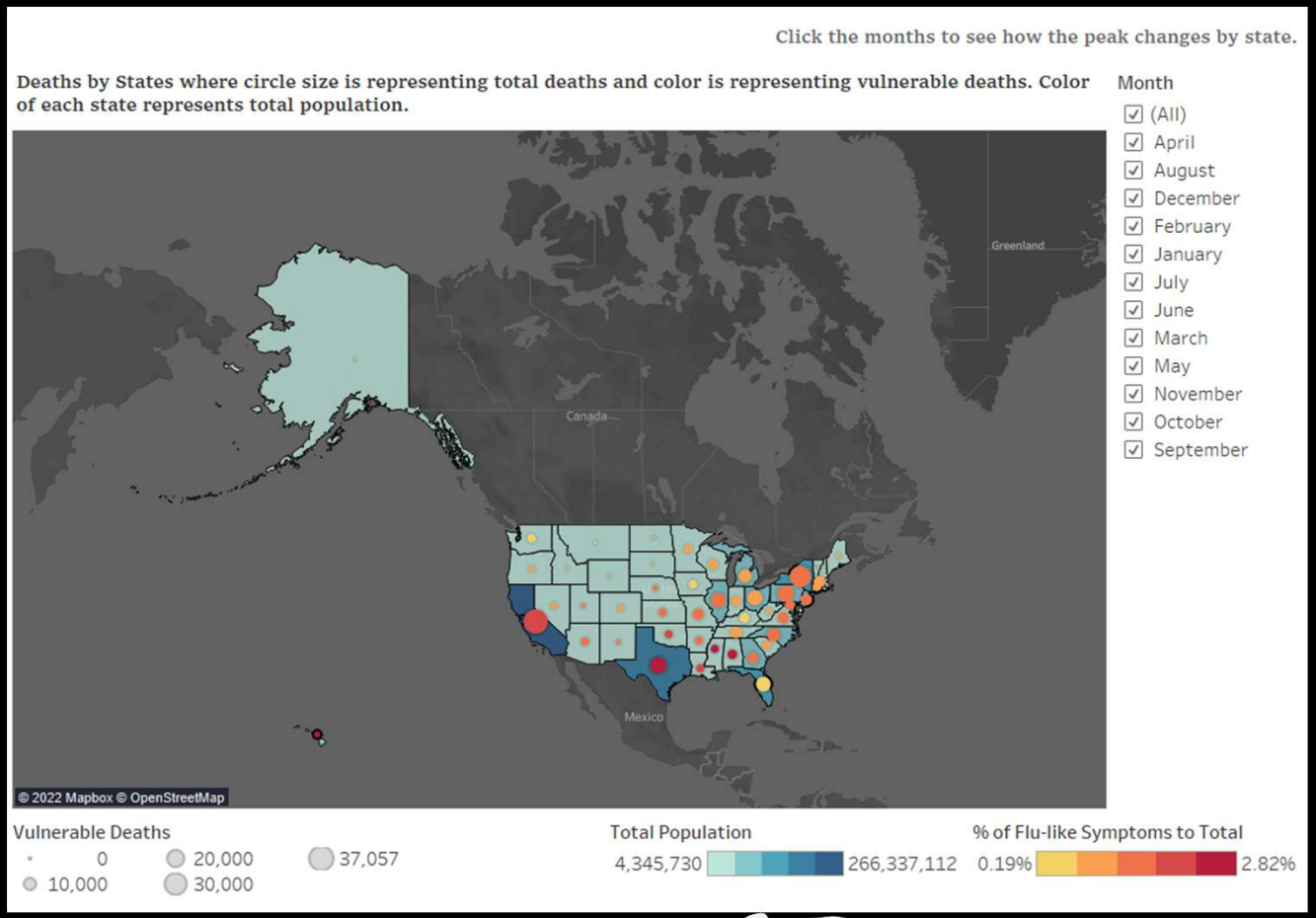
I have also utilized callouts to bring attention to the outliers: Hawaii with a lower poverty rate but higher mortality rate and New Mexico with a higher poverty rate but much lower mortality rate.

I used CDC data to determine which months have the most deaths and compared to which months have the highest percentage of influenza deaths of vulnerable patients.



This line chart shows that while the overall deaths are higher in the winter months, 97% of deaths in summer are vulnerable persons.

I created a map in Tableau using combined CDC and Census Bureau information to determine where the greatest need lay.



The map shows that California, New York, Texas, and Florida have the greatest need per the total and vulnerable persons population. Those states also have the highest death rates. The map also shows that Hawaii is disproportionate in its vulnerable population deaths to total population.

I also added a filter so the map is interactive and can show how the information changes by month. One can also use a mouse and rollover any object and it will highlight the object and give specific information per state.

I presented my project in video format using Vimeo

In conclusion, our next steps will be to send additional hospital and clinic staffing to California, Texas, New York, and Florida starting in December until April.

Monitor how the fatality rate is changed over the coming years.

Determine why Hawaii has such a high percentage of fatalities to its population. Likely due to its international tourism industry, but there may be other factors to consider.

There were limitations in the data. In the file of deaths from the CDC, there were a number of months and years for many age groups that were suppressed to protect PHI. We accounted for these as 0 but the CDC suppresses any number below 20 and it could actually be higher.

Data limitations: Florida did not submit any information on flu visits so we cannot determine if they have a higher percentage of patients with flu-like symptoms. It would be helpful to have this information so we could leverage if the number of visits correlates to the fatalities or patients recovering and not dying.

Rockbuster Video Rental Analysis

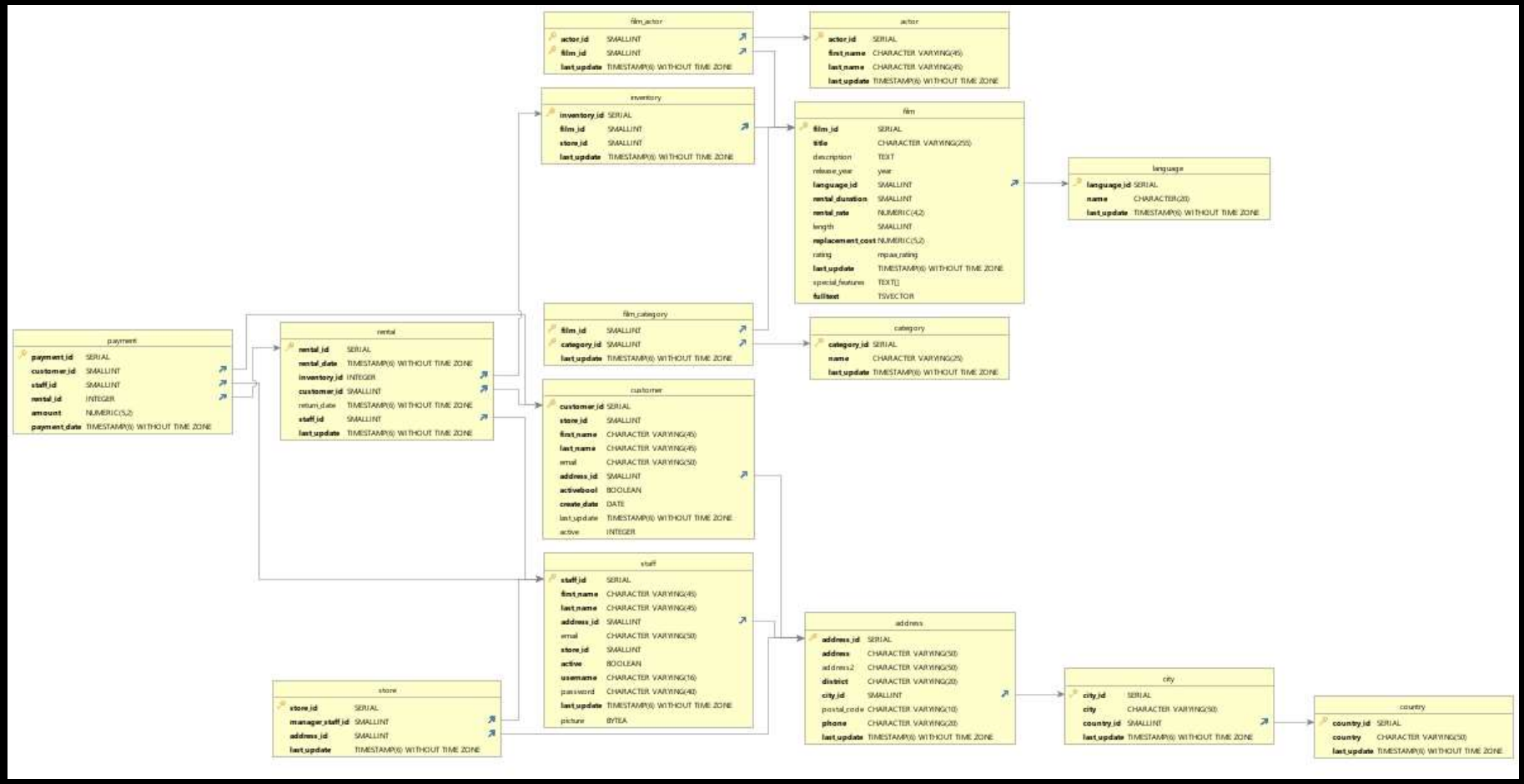
Premise (from the briefing): to provide insight on online video rentals so Rockbuster can compete with Netflix and Amazon Prime.

Key Questions

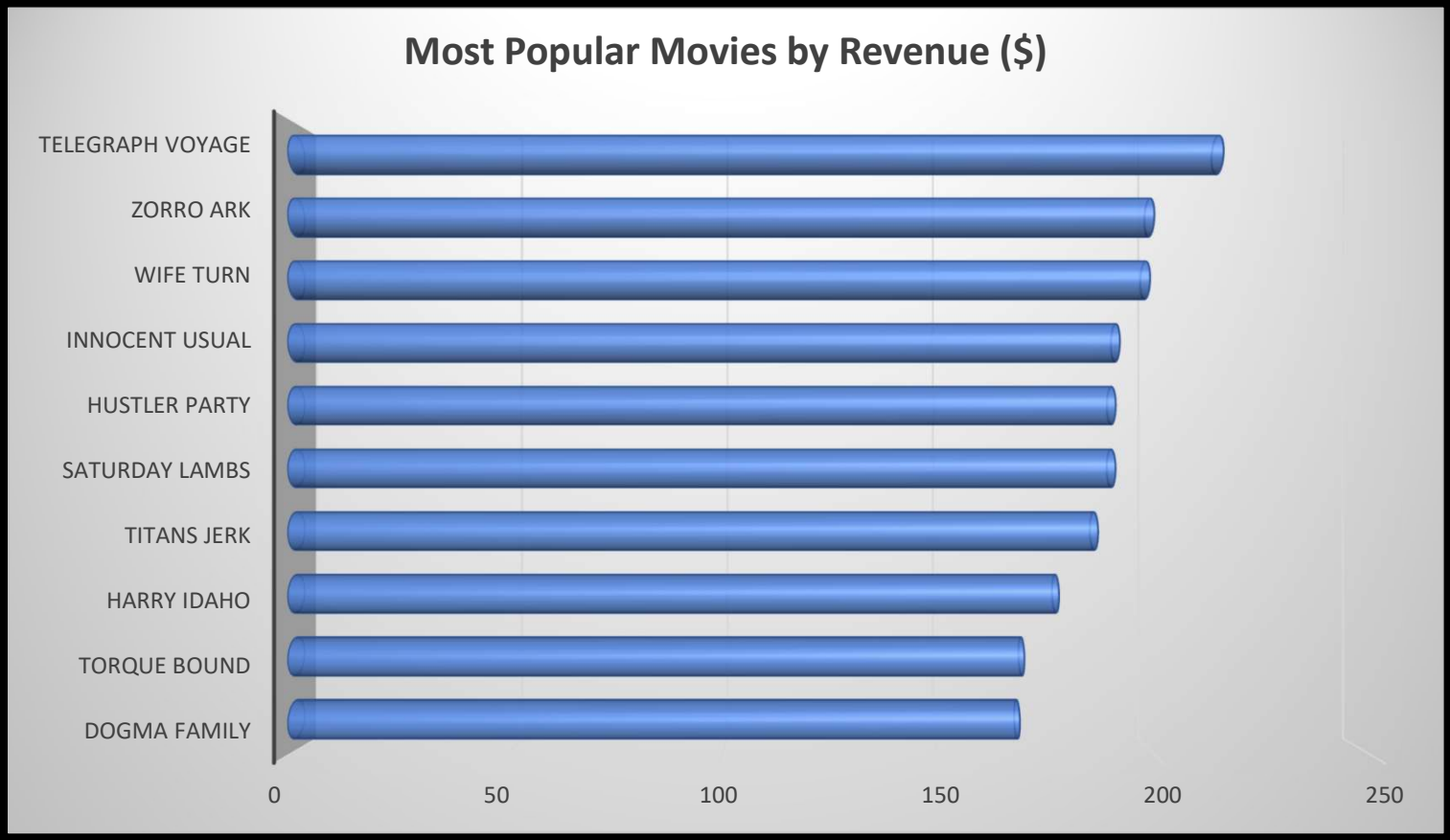
- Which movies contributed the most/least to revenue gain?
- What was the average rental duration for all videos?
- Which countries are Rockbuster customers based in?
- Where are customers with a high lifetime value based?
- Do sales figures vary between geographic regions?

Data: data sets from Rockbuster containing information about inventory, customers, payments, etc.
Tools and skills: SQL, PostgreSQL, PGAdmin4, relational databases, database querying, filtering, cleaning, and summarizing, joining tables, subqueries, common table expressions
Github: Rockbuster

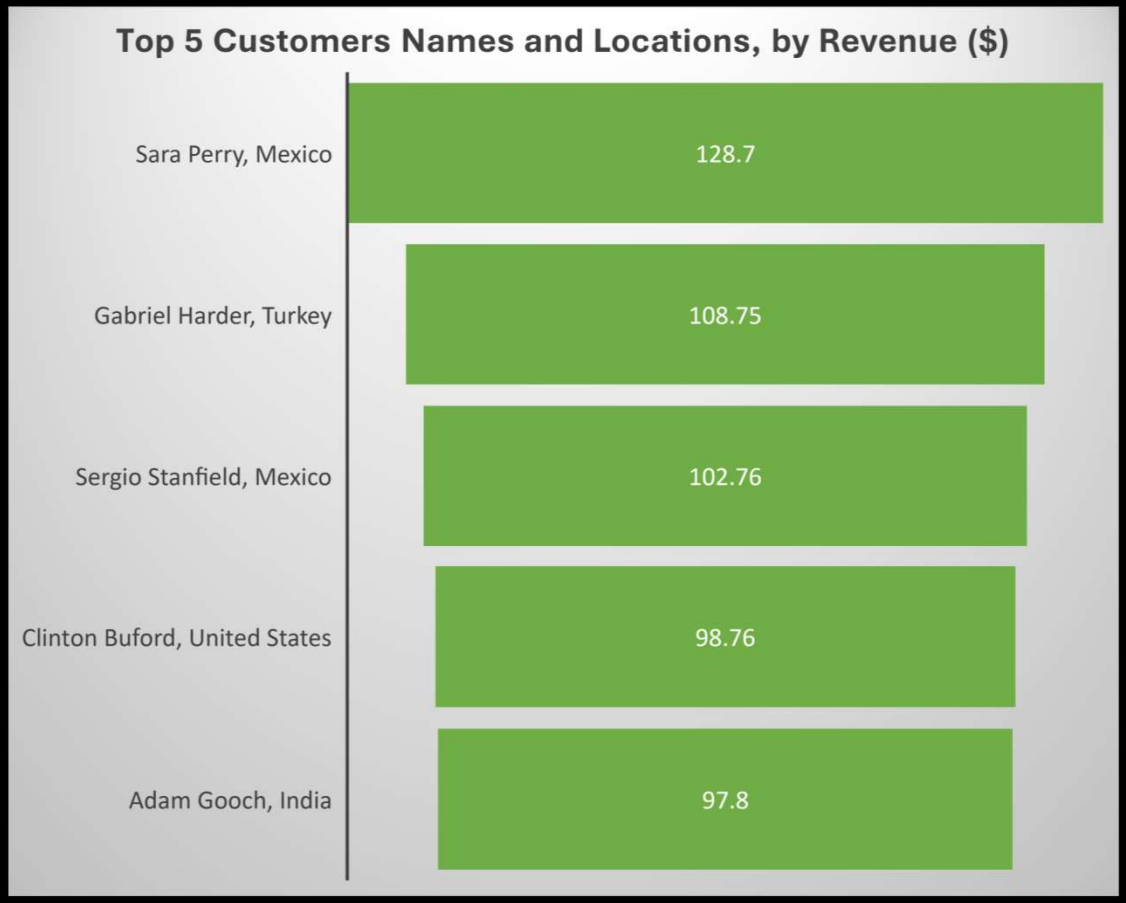
I created a data dictionary using DB Visualizer so I could track how to join my tables in SQL.



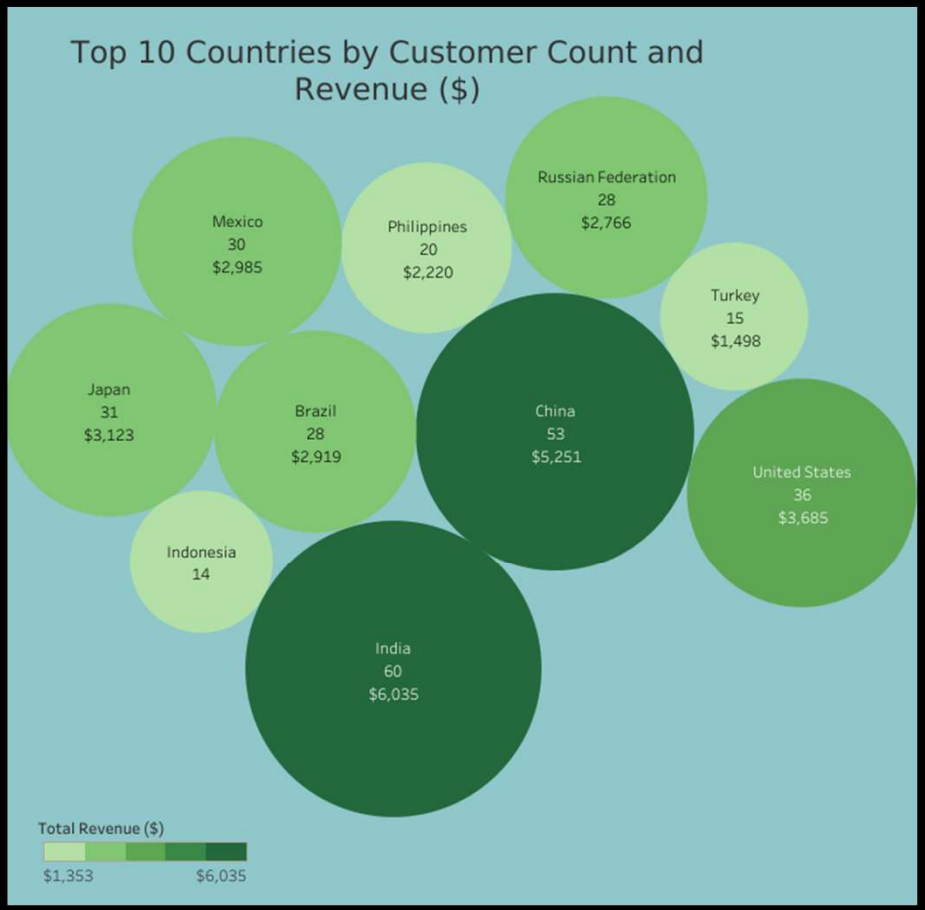
Using my data dictionary, I joined tables in SQL to determine revenue of movies in the Rockbuster database to determine the movies with the most and least revenue and used the exported tables in Excel to create line charts.



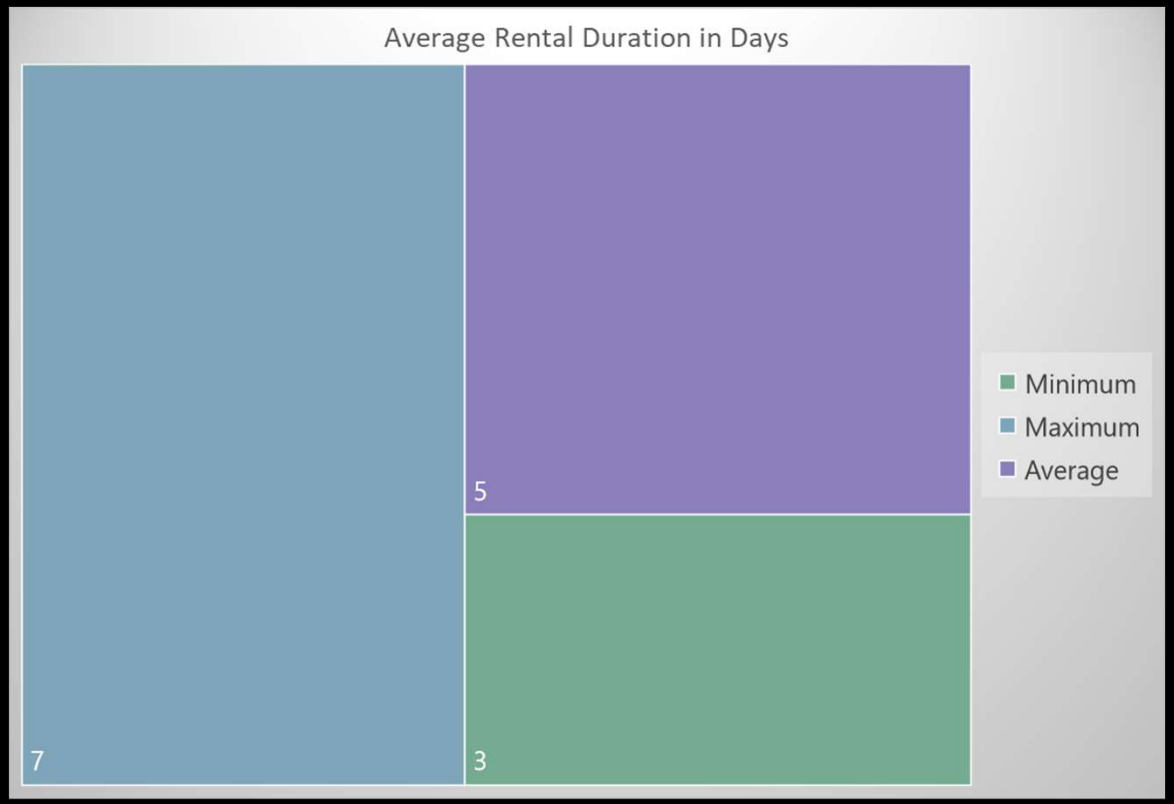
Using my data dictionary, I joined tables in SQL to determine where the top 5 customers by revenue live and presented as a funnel chart.



Using my data dictionary, I joined tables in SQL to determine which countries produced the most revenue and had the highest customer count, I presented this information in a bubble chart made in Tableau.



I then determined the average rental duration and presented as a treemap.



Rockbuster also wanted to know where their customers are based, I used SQL to join tables to determine where customers are based and revenue by country. I presented this as a map with a bubble overlay to show revenue.

In conclusion, I recommended that Rockbuster:

- Focus sales on the top 10 countries currently generating revenue
- Focus sales on countries that could generate more revenue such as England, Germany, France

Data Limitations:

- Database only has:
  - Movies from 2006
  - Rentals from February 14, 2007 to May 14, 2007
  - 1,000 movies

Things I would improve: I would present average rental days differently, as a treemap is confusing. A column chart would have been clear and concise.

Marketing Strategy Analysis for Online Grocery Store

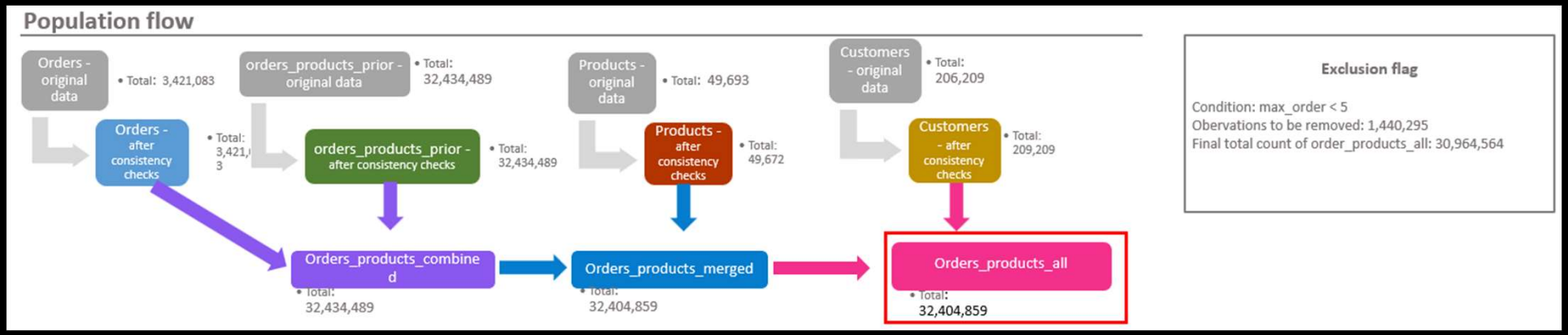
Premise (from the briefing): to provide insight on customers' shopping habits for Instacart, an online grocery store.

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.
- Instacart has a lot of products with different price tags. Marketing and sales want to use simpler price range groupings to help direct their efforts.
- Are there certain types of products that are more popular than others? The marketing and sales teams want to know which departments have the highest frequency of product orders.
- The marketing and sales teams are particularly interested in the different types of customers in their system and how their ordering behaviors differ.

Data: Instacart open-source data sets and CareerFoundry created data on customer information
Tools and skills: Python, data wrangling, data merging, deriving variables, grouping data, aggregating data, reporting in Excel, population flows.
Github: Instacart

I filled in a population flow with noted exclusion flag to keep track of how the data sets changed before and after cleaning in the Excel report provided by CareerFoundry for this purpose:



I kept track of consistency checks and what I did to fix/clean the data:

Consistency checks

Dataset	Missing values	Missing values treatment	Duplicates
orders	206,209	changed to 0, first time customer	none
products	16	remove from dataframe	none
orders_products_prior	none	none	none
customers	11259 (first names)	removed names from dataframe anyway	none

I wrangled the data and further kept track of any changes made:

Wrangling steps

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
	order_dow to orders_day_of_week		columns dropped due to not adding value to the analysis
first_name	order_hour_of_day to orders_hour		columns renamed to add clarity to their contents
last_name			no columns needed data type changed
Unnamed : 0			
Unnamed : 1			
_merge			

I derived several new variables and kept track of new columns made:

Column derivations and aggregations

Dataset	New column	Column/s it was derived from	Conditions
df_ords_prods_merged	price_range_loc	prices	low-range = price below 5 mid-range = price between 5 and below 15 high-range = price above 15
df_ords_prods_merged	max_order	order_number, user_id	groups user id by order number to determine max number of orders per user/customer
df_ords_prods_merged	busiest_day	orders_day_of_week	busiest day = busiest day of week according to value count least busy day = least busy day according to value count regularly busy = all other days
df_ords_prods_merged	busiest_days	orders_day_of_week	busiest days = 2 busiest day of week according to value count least busy days = 2 least busy day according to value count regularly busy = all other days
df_ords_prods_merged	busiest_period_of_da	order_hour_of_day	5 pm fewest orders = hours when least orders come in, between 12 am and 7 am
df_ords_prods_merged	loyalty_flag	max_order	loyal customer = max_orders over 40 regular customer = max_orders between 10 & 40 new customer = max_orders 10 and under
df_ords_prods_merged	spend_flag	avg_order	high spender = avg_order higher than 10 low spender = avg_order 10 and under
df_ords_prods_merged	freq_flag	order_freq	non-frequent customer = order_freq fewer than 20 days regular customer = order_freq between 10 and 20 days frequent customer = order_freq 10 days or fewer
orders_products_customers_merge	profile	depedents, fam_status, age	young parent = between 18 and 35 years of age, 1 or more dependents single adult = above 18, fam status anything other than married married adult = age 35 and older and fam status = married
orders_products_customers_merge	region	state	group states into regions based on wikipedia https://simple.wikipedia.org/wiki/List_of_regions_of_the_United_States
df_ords_prods_merged	avg_order	prices, user id	groups user id by prices to determine average prices per
df_ords_prods_merged	order_freq	user id, days since prior orde	groups user id by days since prior order to determine order frequency

I created several visualizations in Python to answer key questions:

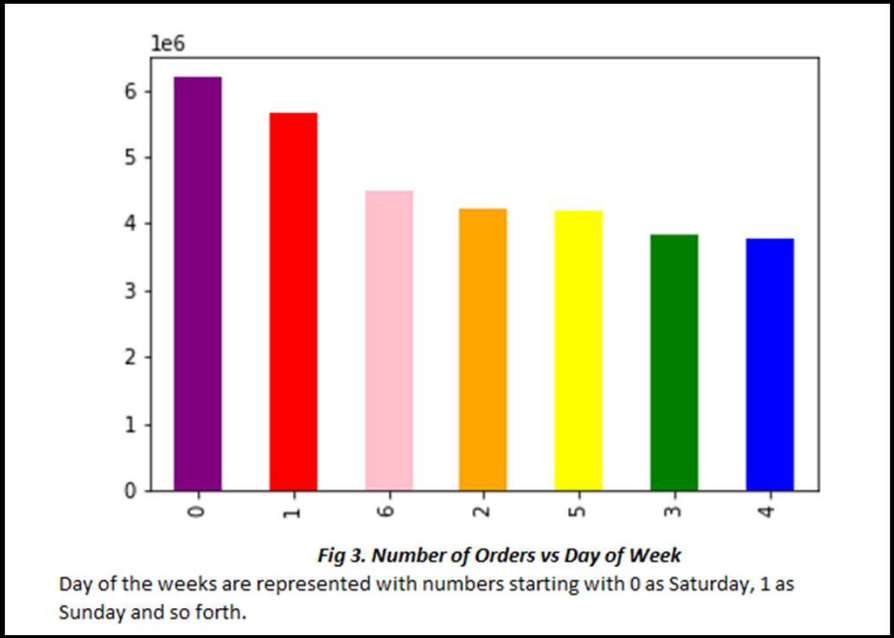


Fig 3. Number of Orders vs Day of Week

Day of the weeks are represented with numbers starting with 0 as Saturday, 1 as Sunday and so forth.

Figure 3 is a column chart created in Python with colors defined to determine busiest shopping day of the week.

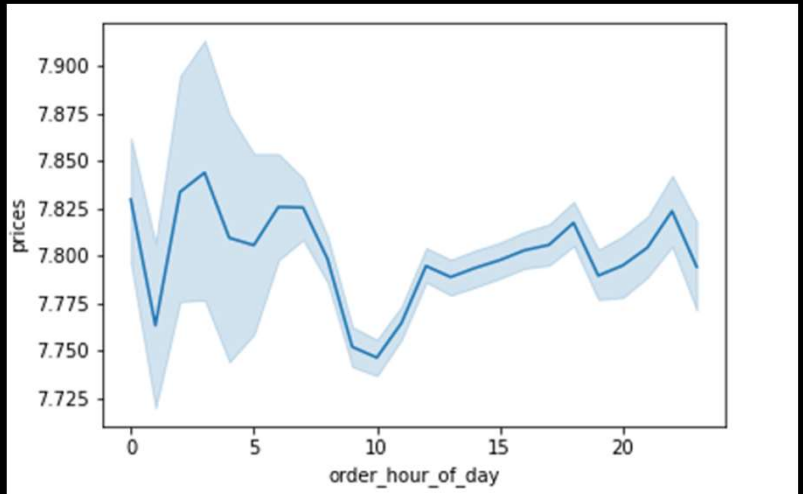


Fig 6. Prics vs Hour of Day

We can see that more expensive items are typically bought in the morning with a sharp drop around mid-morning but steadily growing again in the afternoon. Mid-morning might be the time to advertist more expensive items.

Figure 6 is a line chart created in Python to view the price of items bought versus the hour of day that they were ordered.

In conclusion, the distribution is twice as many regular customers as new customers, with loyal customers in the middle. More information about the differences in spending habits between regular and loyal customers would provide insight as to what makes a customer loyal vs regular. I would recommend looking at money spent overall and if coupons or rewards play a part.

There are differences based on loyalty and spending but they are not statistically significant. About 2% of all customers, regardless of loyalty designation, are high spenders.

There are more frequent regular customers than loyal customers. Loyalty was determined based on total orders while the frequency flag was based on order frequency. Instacart may want to include order frequency within their loyalty parameters and cross check against when the account was created.

Most customers are not going to be high spenders. South and West regions have more customers and therefore more numbers overall, but the percentages are not statistically different. However, we should also bear in mind that much of the Northeast is very rural, many folks are not going to be as likely to adapt to new things and the grocery stores themselves may not be set up for orders.

There are real differences in ordering habits dependent upon family status and the current needs of the families. Married people over 35 appear to buy more produce than single adults and young parents.

We can determine that age and family composition affect a customer's spending and shopping habits by which departments they shop. Younger parents will shop more in 'Baby' than those without babies. We can also see that Married adults shop much more frequently than Single adults or Young parents, most likely because life is more stable. Young parents could probably benefit more from the convenience of ordering from Instacart.

Married adults spend a lot more on product and dairy & eggs than their counterparts, likely because these are more expensive items and married couples are more stable. Likely, married couples, as they are older than 35, have more stability in their careers and the benefit of 2 incomes. Single adults only have the benefit of 1 income but also have fewer mouths to feed. Young parents likely have the least income and have to spend their money more wisely, as babies are expensive.

Married adults place orders more frequently than their counterparts and buy much more produce and dairy & eggs than their counterparts. Perhaps advertising of those departments should be more focused on single adults and young parents.

Data limitations: the data didn’t seem to account for single, divorced, or widowed parents.

Things I would improve: I would use better color palettes for the column charts, colors that are meant to go together.

Real Estate, Income, Graduate Analysis

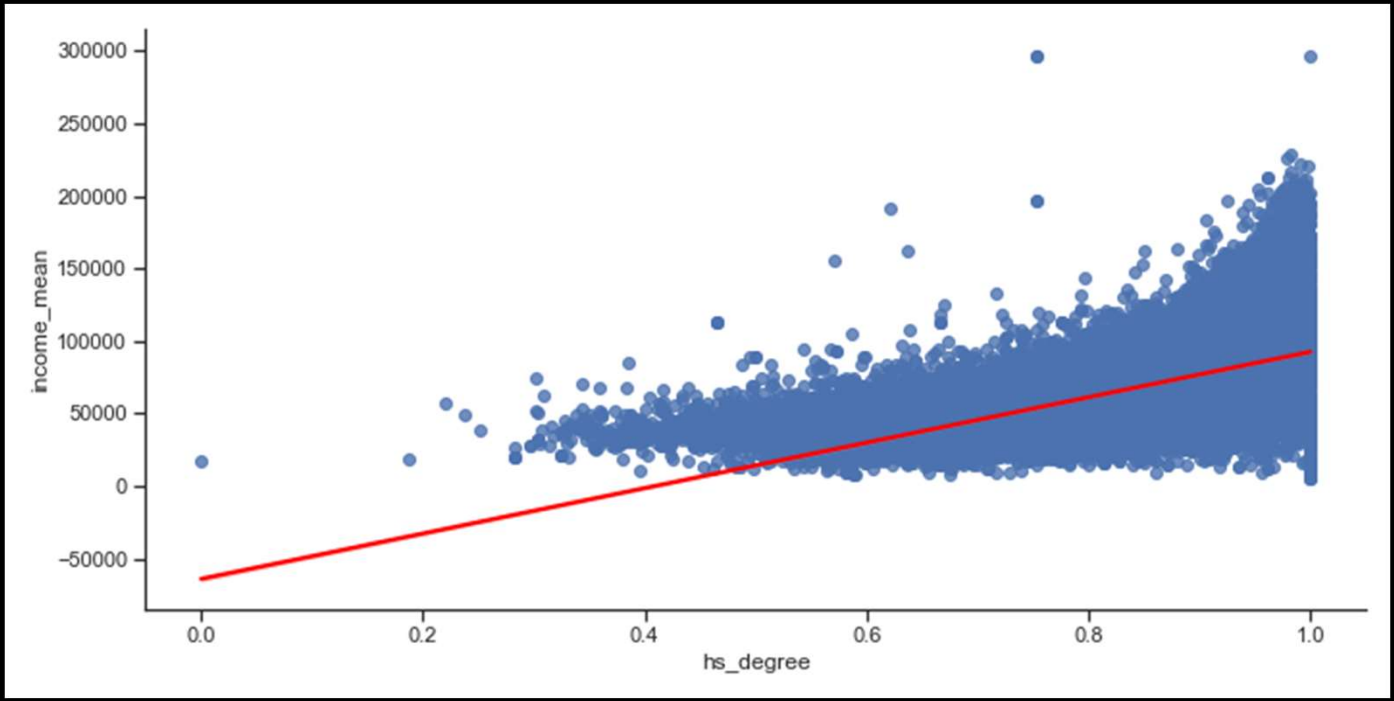
Premise: A government official for the United States wants to know the relationships between average income, population, real estate (mortgage and rent), and the percentage of high school graduates per state.

Key Questions

1. Does population play a role in percentage of home ownership per area?
2. Does high school diploma play a role in family income?
3. Does high school diploma play a role in mortgage and expenses?

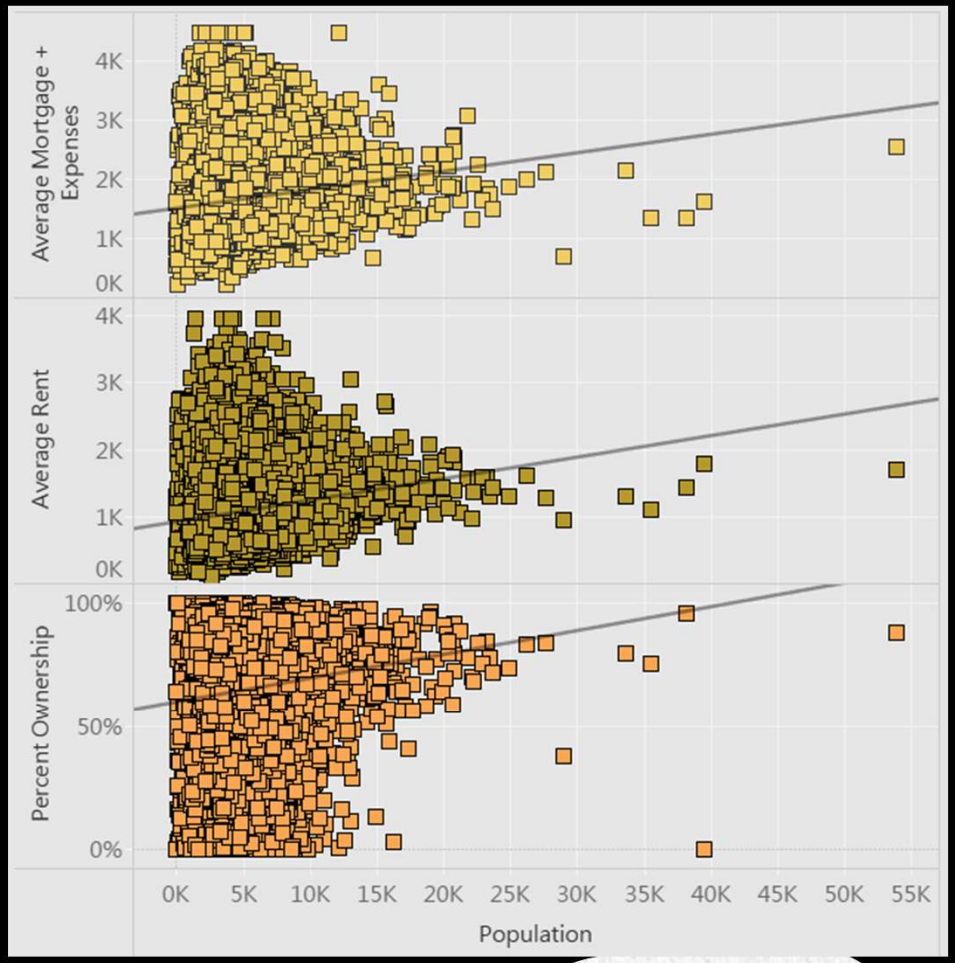
Data: https://www.kaggle.com/datasets/goldenoakresearch/us-acs-mortgage-equity-loans-rent-statistics
Tools and skills: Python, supervised machine learning, linear regression, k-means and clustering, presenting results in a dashboard
Github: https://github.com/HeidiHodges/Real-Estate-Income-Graduate-Analysis
Tableau: https://public.tableau.com/app/profile/h.hodges/viz/HHodgesCF6_7Analysis/Analyses

To start, I conducted some statistical analyses of percentage of high school degree per state against mortgage and income. I produced scatterplots in Python with a trend line to showcase the correlation.



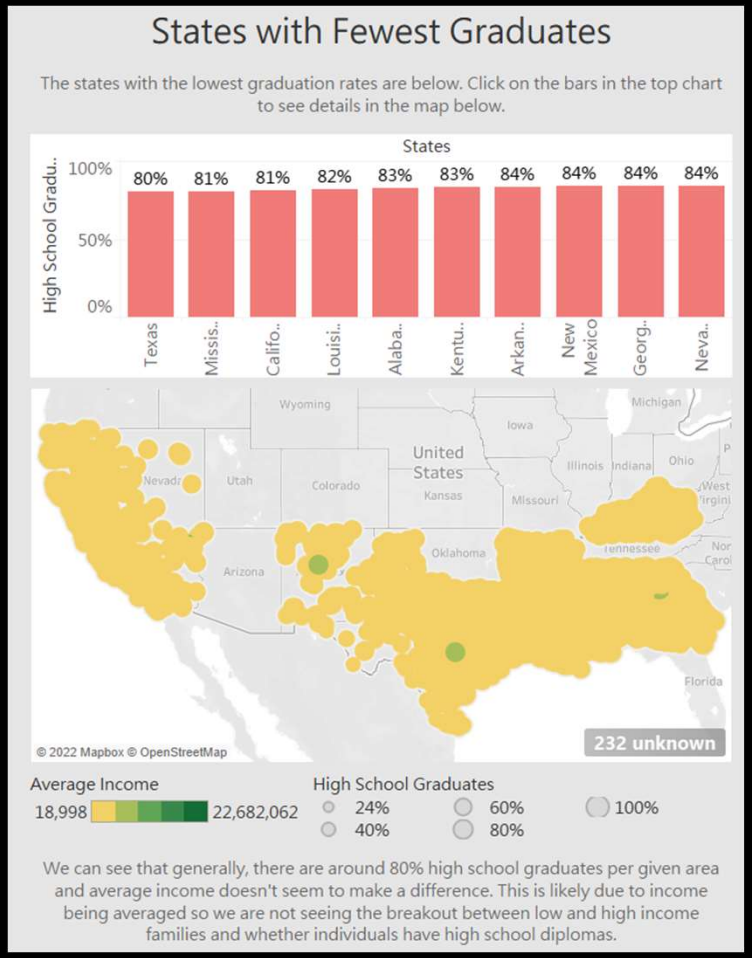
This shows that there is a correlation and trend between high school graduates and the average income of a state. The outliers above the line to the right of the graph are likely due to higher education. The outliers above the line and to the left of the graph could be from areas where there is little industry or need for education.

I then conducted an analysis to see if population was a determining factor in mortgage and expenses, rent, or percent ownership per location:



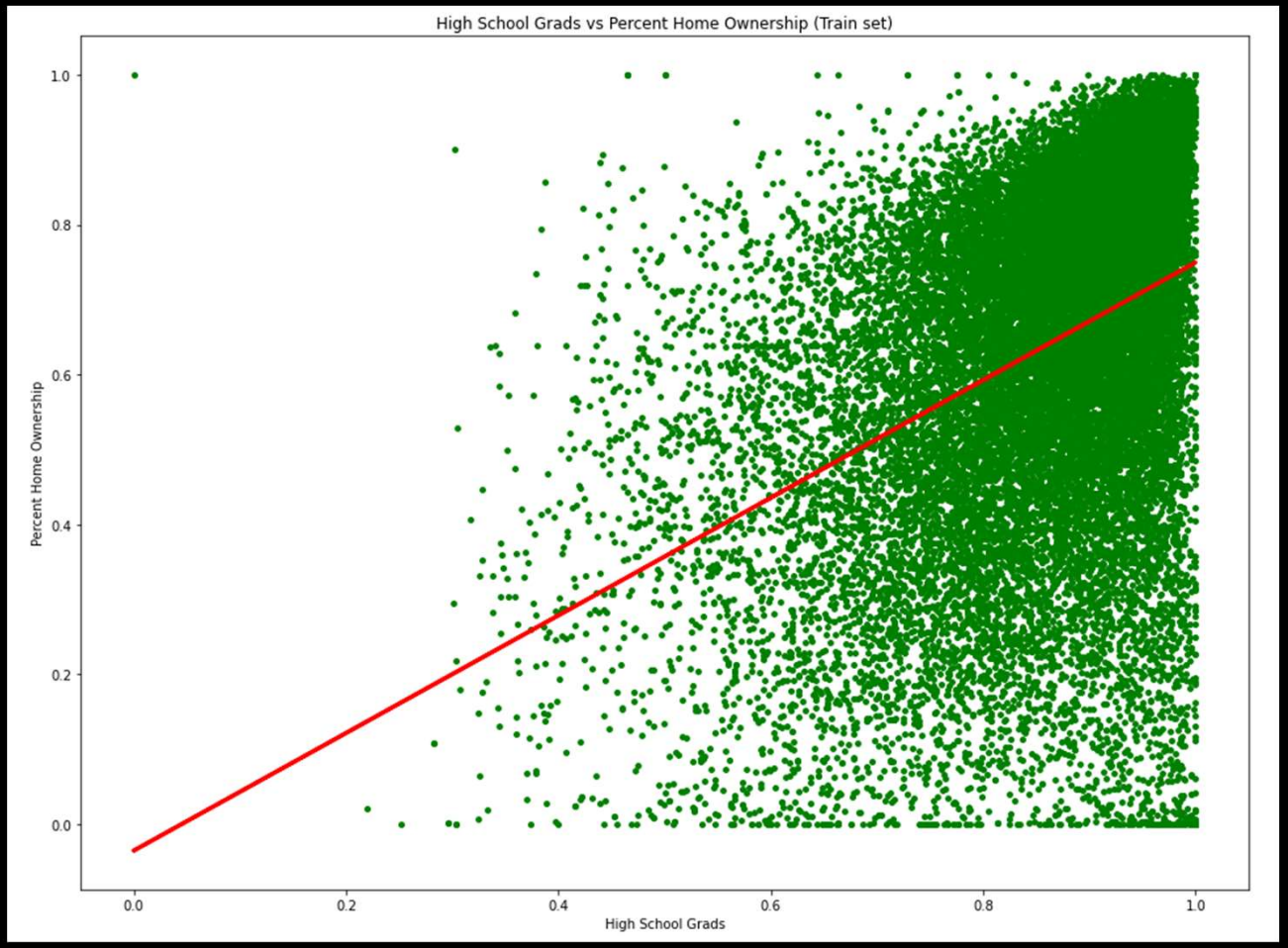
This shows that there is little to no correlation between population size and mortgage, rent, or percent ownership per area. Clearly having a high school diploma is a better factor to explore.

I then explored the states with the fewest graduates, and plotted income vs high school graduates:



This is an interactive dashboard I created in Tableau to showcase our bottom 10 performing states. You can see areas with high average income but most have around the same percentage of graduates. The lack of definition in the data set by averaging income and high school graduate percentages by city instead of individuals is likely leading to this lack of insight.

I also ran a linear regression using supervised machine learning:



This shows the correlation between home ownership and high school graduates. There are many outliers likely due to higher education, inheritances, etc.

In conclusion, while there is a cyclical correlation between income and having a high school diploma, there is more information needed about the individual cities with the lowest graduation rates.

Limitations: The data set was from 2017 and only 2017 so there is no opportunity to be able to see how this information changes over time. We do not have industry data within the dataset to determine the main industry in the low performing areas (Silicon Valley vs Mining Towns, for example).

Things I would improve: I would love to do an analysis on just the low performing states and go area by area to see how proximity to metropolitan centers affects any graduation rates.