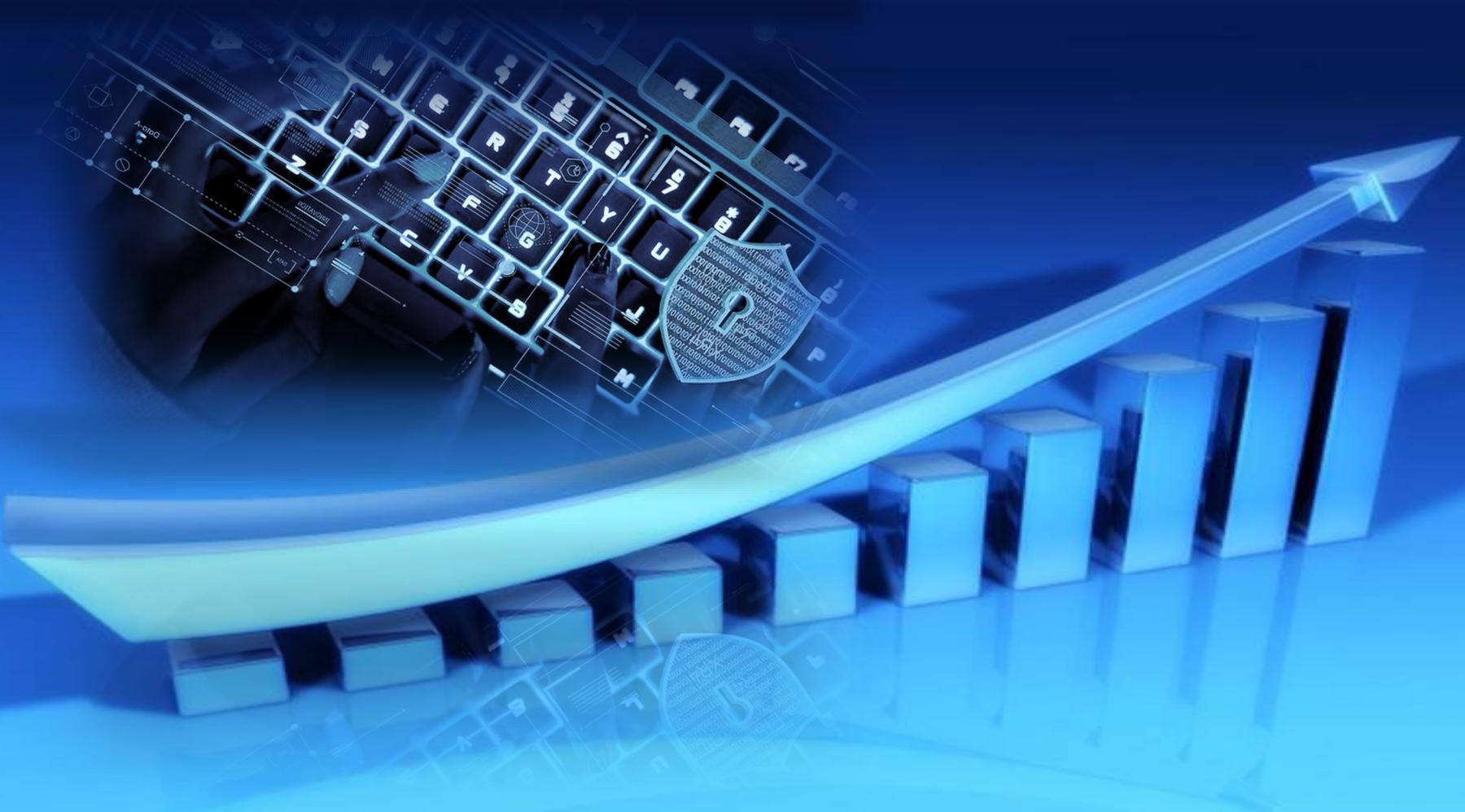


Jay Haygood

Data Analytics Portfolio



About Me

I have an IT background in Platform Validation and Software Quality Assurance. My recent skillset includes knowledge of Excel, PowerPoint, Tableau, SQL, and Python as tools related to the data analysis projects featured in this portfolio.

Professionally, I've been a business owner and performance artist. My personal interests include music, fitness, outdoor recreation, and astronomy.



Featured Projects

GameCo, Inc.

Descriptive analysis of how new games may sell in the current market

Instacart

Exploratory analysis and customer profiling, using insights to optimize marketing campaigns

Influenza Season

Analysis examining influenza trends in the US, for use in forecasting seasonal hospital staffing

Pig E. Bank

Predictive analysis on customer retention, and decision tree

Rockbuster Stealth

Geographic analysis of movie rentals, with intent to launch online video streaming service

UFO Sightings

Geospatial analysis and exploratory passion project, using open source data on 80,000+ events.

GameCo, Inc. Overview



OBJECTIVE

Perform an analysis of video game sales data to optimize GameCo marketing plan and ROI.



ASSUMPTION

Game sales for the various geographic regions have stayed consistent over time.



DATA

[Video Game Sales](#)
Source: [VGChartz](#)



TOOLS

Excel
PowerPoint

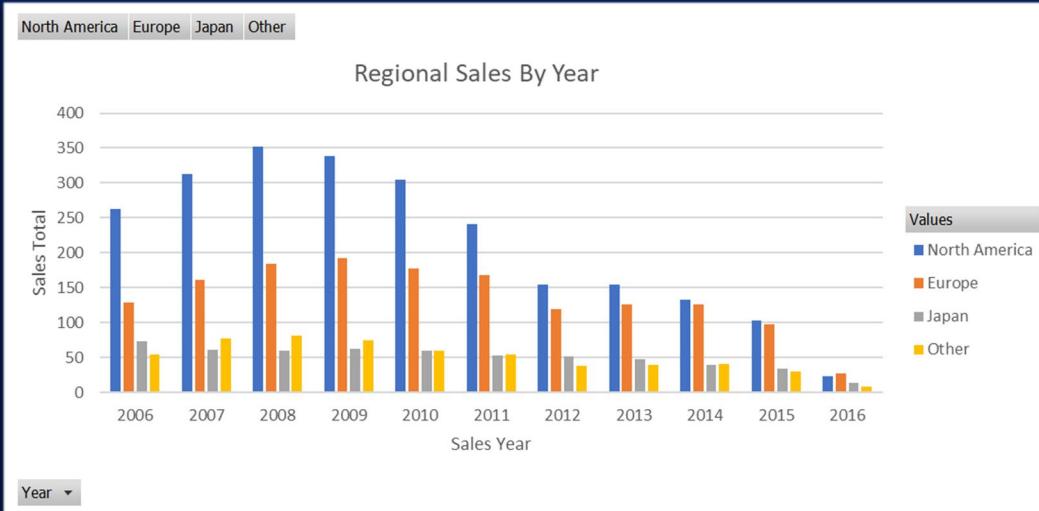


SKILLS

Grouping and Summarizing Data
Descriptive Analysis
Visualizing Insights
Presenting Findings



GameCo, Inc. Analysis

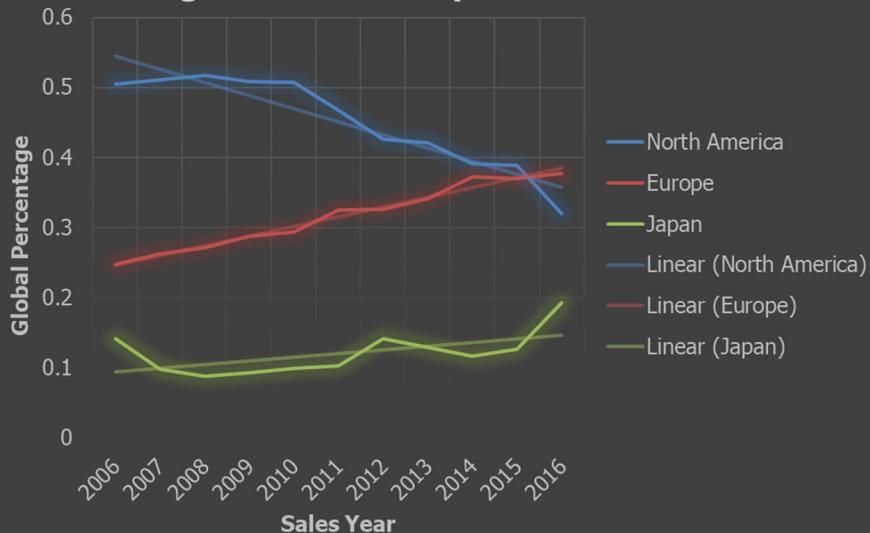


- Shows a steady decline in overall global sales since 2008, most noticeable with drop in North American sales.
- Comparing the various regional totals relative to each other in recent years, they appear generally consistent and unchanged, which would concur with GameCo's current assumptions.

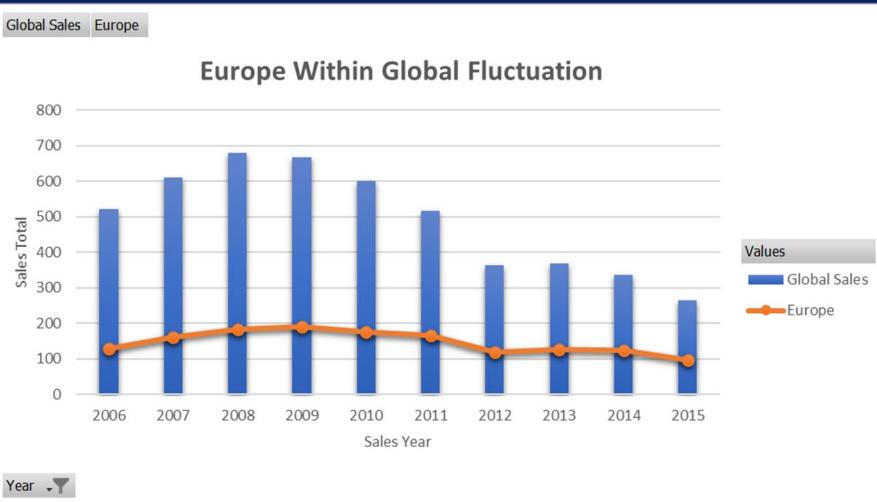
- Regional sales for Europe have eclipsed that of North America for the first time, as of Q3 2016.
- While we would typically not view this as indication of a trend, being the first such occurrence, it is actually a snapshot of a bigger picture.



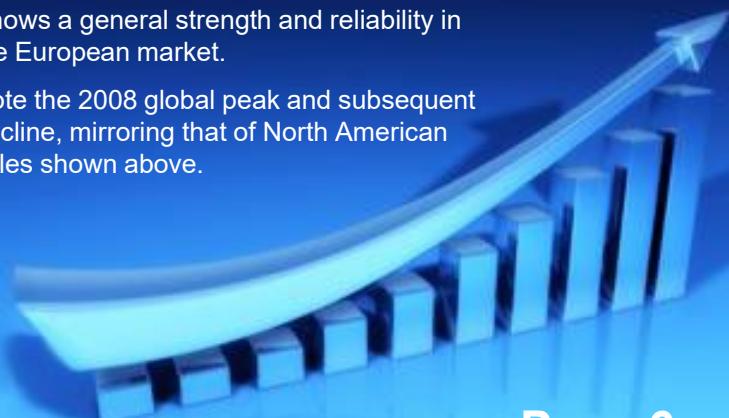
Regional Sales Proportion Trends



- Steady and consistent increase in Europe over the last 10 years.
- The Europe increase is proportional to a decline in North America.
- Despite a sharp increase in Japan for 2016, it is too recent to affect the trendline. Still an upward trend, it is also worth consideration.



- While somewhat parallel to global sales variations, Europe still maintains an overall consistency through this 10-year cycle.
- Shows a general strength and reliability in the European market.
- Note the 2008 global peak and subsequent decline, mirroring that of North American sales shown above.



Summary and Proposals

FINDINGS

- ❖ GameCo's assumption that regional sales have remained mostly consistent is not accurate, when considering how they figure into a steady decline in global sales.
- ❖ While European sales have eclipsed that of North America for the first time this year, their percentage of global sales has steadily increased, in direct contrast to a decline in the North American proportion.
- ❖ Japan has shown a steady increase in global proportion as well, but to a lesser degree. This is worthy of consideration in the future.

RECOMMENDED ACTION

Focus more marketing resources on Europe, in an effort to capitalize on the existing increase in percentage of global sales for this region. As this has been a consistent trend, it would likely be a safe approach for optimal return on marketing investment.

FINAL REPORT



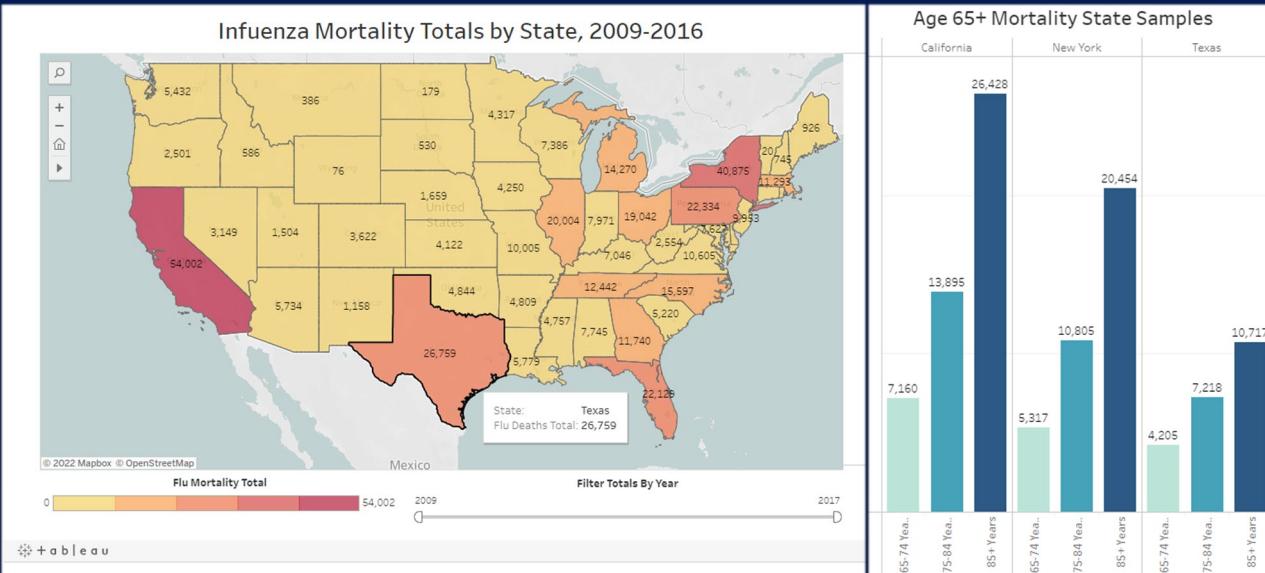
Influenza Season Overview

				
OBJECTIVE	ASSUMPTION	DATA	SKILLS	TOOLS
Analysis will help plan for influenza season, when additional hospital staff are needed. Final results will examine trends to proactively plan for staffing demands across the country.	Vulnerable populations suffer the most severe impacts from the flu, and are the most likely to be hospitalized.	Influenza Deaths Source: CDC Population Data Source: US Census Bureau Influenza Visits Lab Test Data Source: CDC (Fluvview) Flu Shot Survey Source: CDC	Data Profiling, Cleaning, Integration, and Transformation Statistical Hypothesis Testing Visual Analysis and Forecasting Storytelling with Data	Excel Tableau

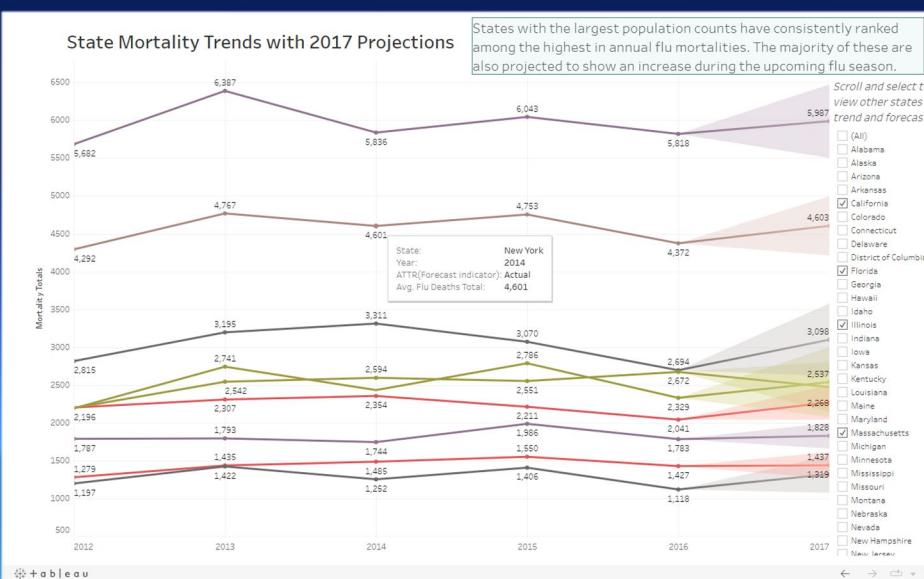
INTERIM REPORT



Influenza Season Analysis

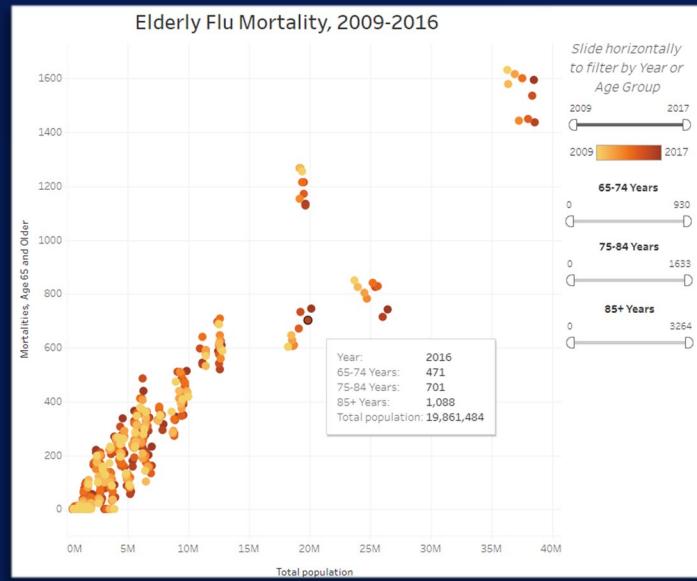


➤ Densely populated states tend to show much higher flu-related mortalities overall. These cases are notably highest among the vulnerable population, age 65 and older. This should justify increased hospital staffing to these regions for the upcoming flu season.

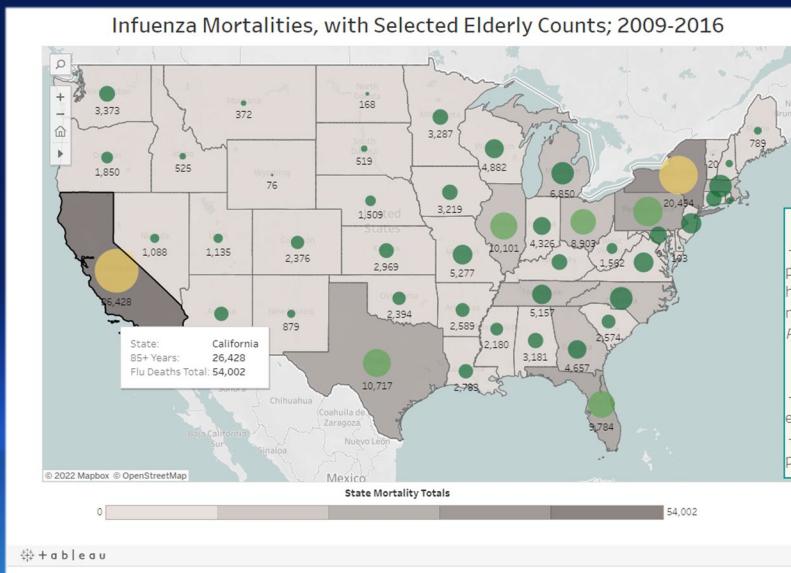


➤ States with the largest population counts have consistently ranked among the highest in annual flu mortalities. The majority of these are also projected to show an increase during the upcoming flu season.





- The plot shows a distinct correlation with increased influenza mortalities among all vulnerable age groups in the elderly population, relative to the total population, through the previous seven flu seasons.



- We clearly see a correlation in the US between densely populated states and increasing influenza mortality cases, especially among the elderly population.

Summary and Proposals

FINDINGS

There is a direct correlation in the United States between increasing influenza mortality cases and the densely populated states, particularly among the elderly population.

RECOMMENDED ACTION

- ❖ Increase hospital staffing beyond that of previous flu seasons, among the most heavily populated states.
- ❖ Deploy staff members to these areas earlier where feasible, as budget allows.
- ❖ Place an emphasis on earlier vaccinations, particularly among the elderly population.

FINAL REPORT



Rockbuster Stealth Overview

				
OBJECTIVE	QUESTIONS	DATA	TOOLS	SKILLS
Apply a relational database analysis to a series of business questions with data-driven answers for use in the Rockbuster 2020 company strategy.	<ul style="list-style-type: none">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?	Rockbuster Data Set PostgreSQL Database	PostgreSQL Tableau	Relational Databases Database Querying Filtering, Cleaning and Summarizing Data Table Joins and Subqueries in SQL Common Table Expressions

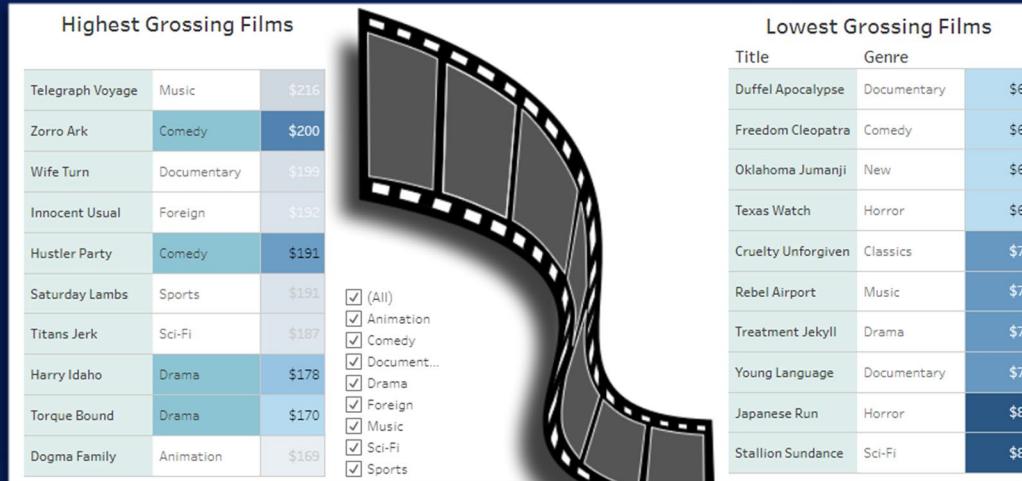


Rockbuster Stealth Analysis

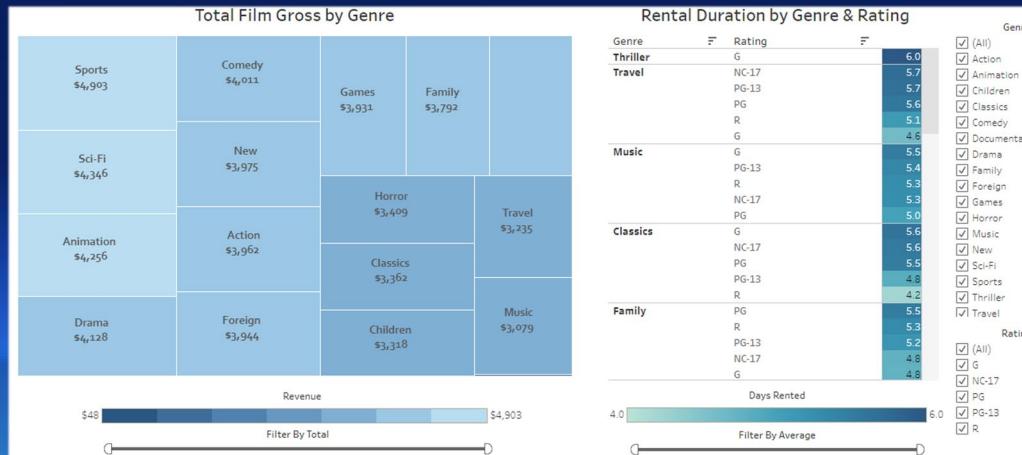
- Which movies contributed the most and least to revenue gain?

The movie which generated the most revenue was Telegraph Voyage, an outlier hit from the Music genre.

Note two appearances each from Comedy and Drama among the top grossing films. ==>



- Which film genres generated the highest revenue?
- What was the average rental duration for all videos?

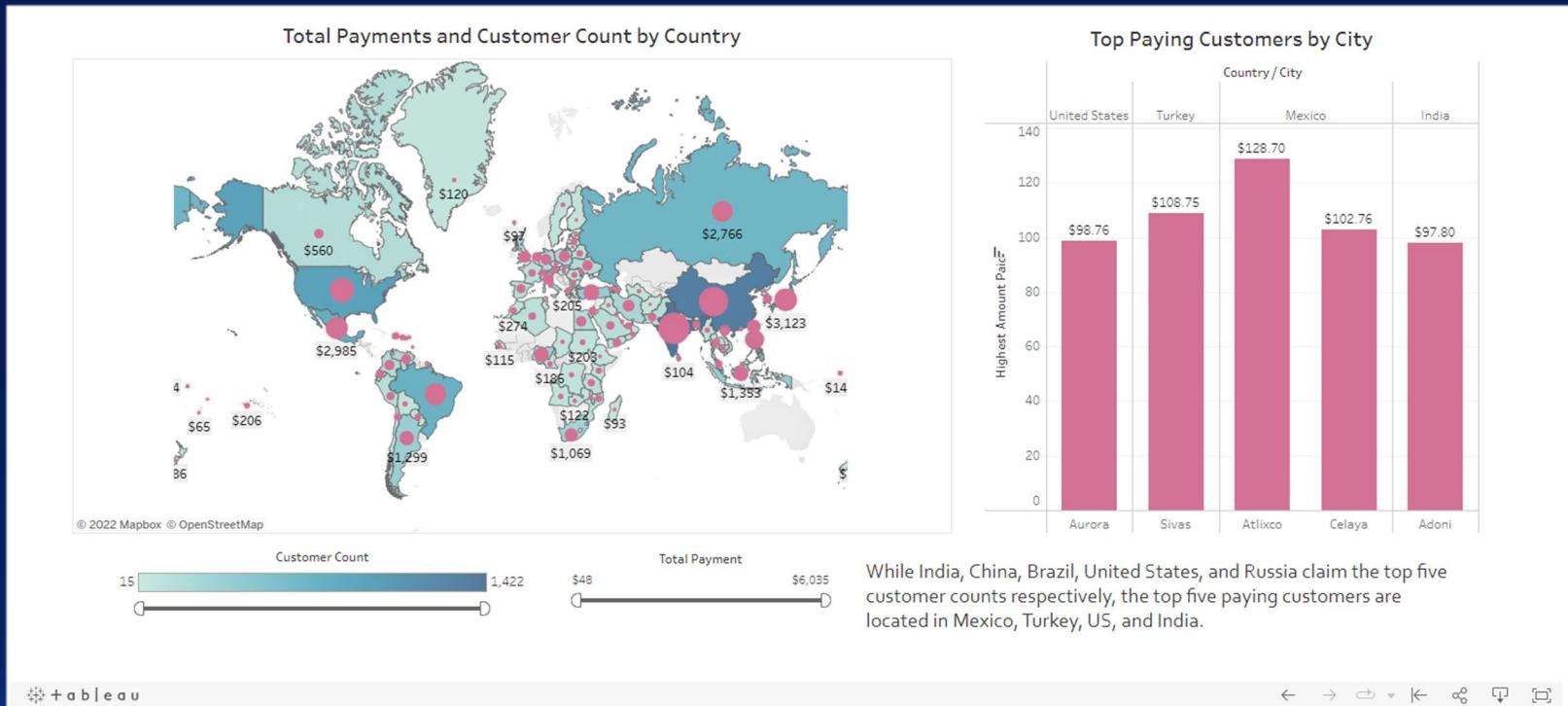


➤ Several movies of various genres were among the lowest grossing.

➤ Overall average rental duration is 5 days. While longer durations will generate greater revenue per title, there is no direct correlation with the higher grossing genres.

Sports, Sci-fi, Animation, Drama, and Comedy are clearly the top grossing categories. Therefore, expanding the Rockbuster Stealth catalog of films from these genres should lead to an increase in revenue.

- Which countries are Rockbuster Stealth customers based in?
- Do sales figures vary between various geographic regions?
- Where are customers with a high lifetime value based?



While there is no distinct correlation between high value customers and countries with high customer counts, both India and the US are home to one of the top five customers, and rank in the top five for customer count as well.

Summary and Proposals

FINDINGS

- ❖ Rockbuster Stealth has established a diverse International customer base, among the strongest being India, China, Brazil, United States, and Russia. These regions also rank the highest in overall revenue, and claim two of the top paying customers.
- ❖ Globally popular genres such as Sci-Fi, Drama, and Comedy are also among our highest grossing movie rental categories. This is ideal for the high-profile streaming market.
- ❖ These factors put Rockbuster Stealth in an ideal position to launch a highly successful online platform.



RECOMMENDED ACTION

- ❖ In preparation to launch the online platform, we should concentrate our marketing strategy towards targeting the top five countries, in an effort to capitalize on their existing customer base.
- ❖ With Sports, Sci-fi, Animation, Drama, and Comedy being the most popular movie rental genres, increasing our film roster accordingly will optimize our strength in the online streaming market.
- ❖ Consider test marketing to the strongest International markets prior to a global rollout, while also considering any potential government restrictions regarding internet content in these markets.

FINAL REPORT



Instacart Overview



OBJECTIVE
Instacart wants to target customers with applicable marketing to gauge sales of their products. Perform an exploratory analysis of Instacart data to derive insights and suggest strategies based on provided criteria.



QUESTIONS

- What are the busiest days of the week and hours of the day?
- Are there particular times of the day when people spend the most money?
- 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?
- What are the different types of customers and how do their ordering behaviors differ?



DATA

[Customers Data Set](#)

[Data Dictionary](#)

Citation (required in final report)

Source: [Instacart](#)



TOOLS

Jupyter Notebooks

Python

Excel



SKILLS

Data Wrangling and Consistency Checks

Combining and Exporting Data

Deriving New Variables

Grouping and Aggregating Data

Visualizing with Python

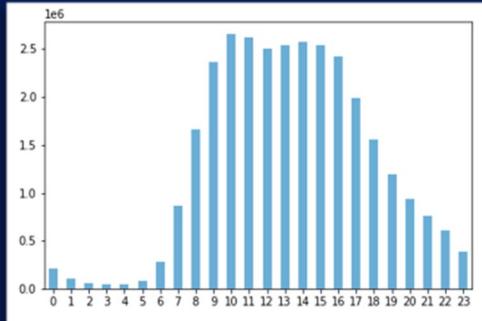
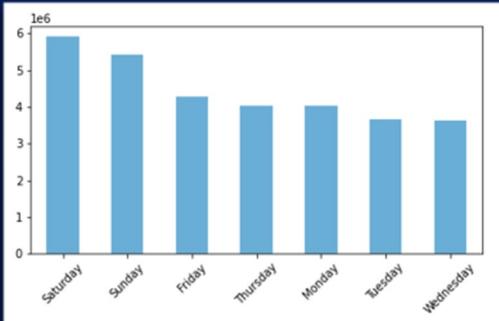
Reporting with Excel



instacart

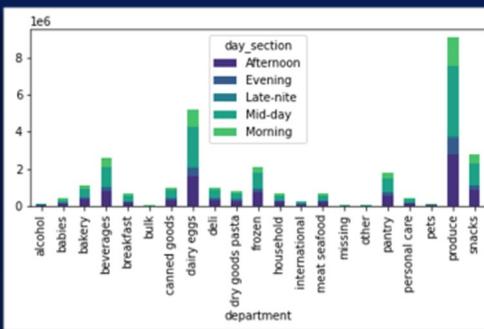
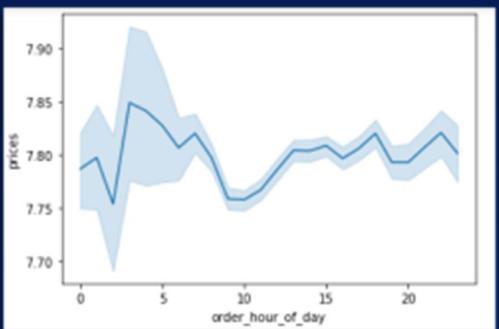
Instacart Analysis

- The sales team needs to know what the busiest days of the week and hours of the day are in order to schedule ads at times when there are fewer orders.



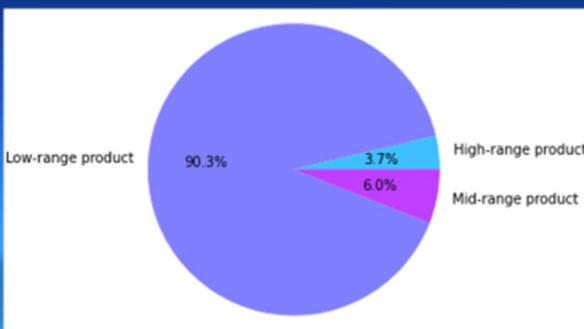
➤ The busiest days of the week are Saturday and Sunday, as shown in the bar chart to the left. The busiest hours are 10am to 4pm, shown as 10-16 respectively, in the chart to the right.

- 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.



➤ Interestingly, we can see in the graph to the left a very slight peak in average pricing occur during the least busiest hours of the day, between 3 and 4am. While the chart on the right clearly shows a majority of Mid-day and Afternoon orders, among all of the major product departments.

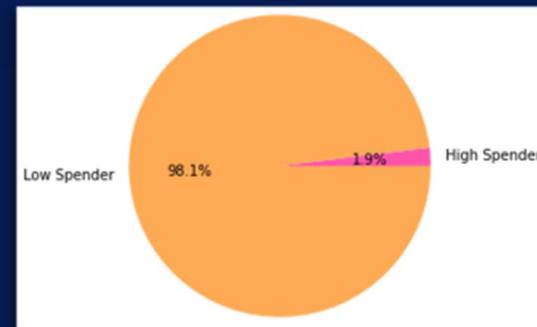
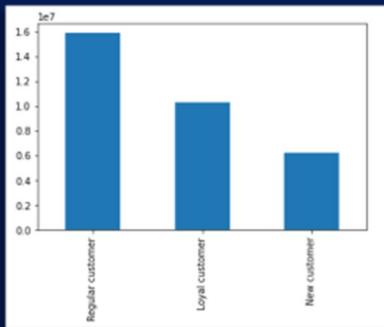
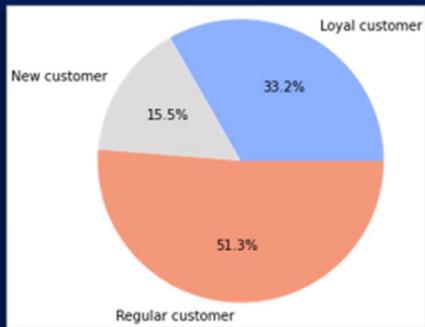
- Marketing and sales want to use simpler price range groupings to help direct their efforts.



➤ Clearly, the Instacart inventory is dominated by lower priced items. Both the Mid- and High-range products combined only encompass 9.7% of the entire product line.

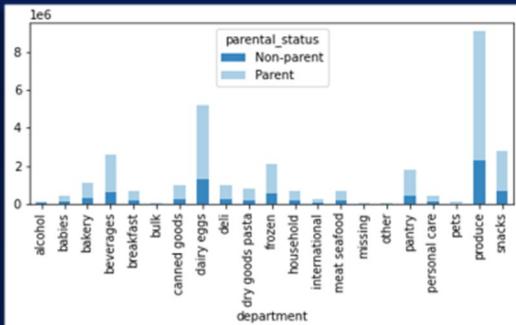
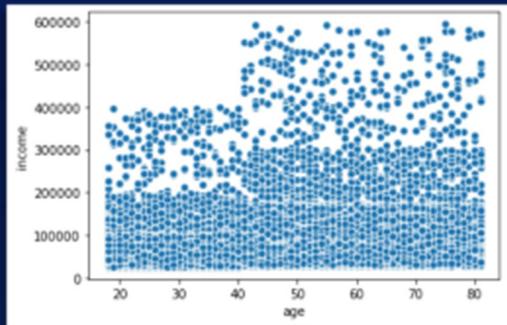


- What's the distribution among users in regards to their brand loyalty (i.e., how often do they return to Instacart)? Are there differences in ordering habits based on a customer's loyalty status?



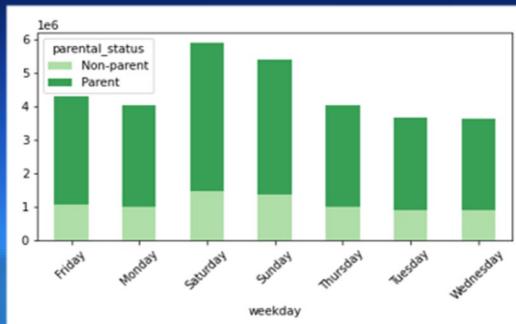
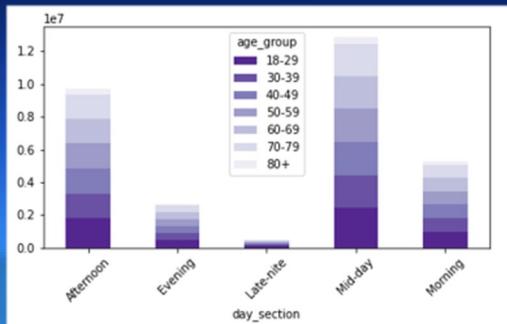
➤ As the pie chart to the left displays the data from the bar chart in clearer detail, the most returning or loyal customers make up a third of the overall customer base, at 33.2%, while regular customers are slightly over half the customer base, at 51.3%.

- What different classifications does the demographic information suggest? Age? Income? Income? Certain types of goods? Family status?



➤ We can see in the scatterplot to the left, a distinct increase in middle and higher income ranges after the age of 40. The chart on the right clearly shows a minority of non-parent customers across virtually every product department.

- What differences in ordering habits are seen among different customer profiles?



➤ While ordering habits among the middle and older age groups show an even distribution across each section of the day, we can see a slight dominance with the youngest 18-29 group, particularly during the busiest times of day. Again there is also an obvious majority of parents, by contrast to non-parents, distinctly shown across every day of the week.

Summary and Proposals

FINDINGS

- ❖ The busiest days of the week are Saturday and Sunday. The busiest hours are 10am to 4pm.
- ❖ Orders are most frequent during Mid-day and Afternoon, among all of the major product departments.
- ❖ Instacart inventory is dominated by lower priced items.
- ❖ Most returning or loyal customers make up a third of the overall customer base, while regular customers are slightly over half the customer base.
- ❖ There is an increase in middle and higher income ranges after the age of 40, and a minority of non-parent customers across virtually every product department.
- ❖ Ordering habits among the middle and older age groups show an even distribution across each section of the day, with a slight dominance among the youngest 18-29 group, particularly during the busiest times of day. There is also an obvious majority of parents, by contrast to non-parents, distinctly shown across every day of the week.

RECOMMENDED ACTION

- ❖ Run ad campaigns during weekdays, specifically Monday through Thursday, focused on increasing orders in the late-night and evening hours.
- ❖ With Produce, Dairy, Beverages, and Snacks claiming the majority of sales, increased advertising for higher priced perishables such as Meat & Seafood, as well as higher profit margin products such as Personal Care items, could potentially increase overall profits.
- ❖ It may be more appropriate to combine the Mid- and High-range products, designating these as High-range, while dividing the Low-range products somewhat proportionately, to create more equal ranges of Low- and Mid-range products.
- ❖ While there is always a marketing push to increase the count of new customers, increased incentives such as discounts and/or product promotions, would keep the regular customers returning, in effect converting them to loyal customers.
- ❖ Consider an increase in marketing towards the older age groups, as they would likely benefit highly from the Instacart business model, as well as non-parents.

FINAL REPORT



Pig E. Bank Overview



OBJECTIVE

To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank.



TASKS

- Use predictive analysis to identify top client loss risk factors
- Model the findings in a decision tree



DATA

[Pig E. Bank Client Data Set](#)



TOOLS

Excel
PowerPoint
GitHub



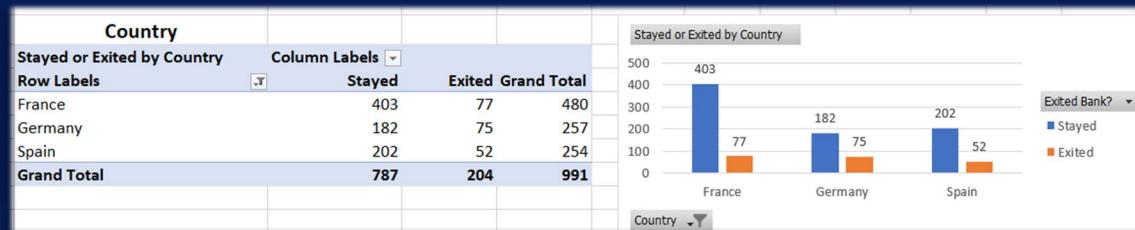
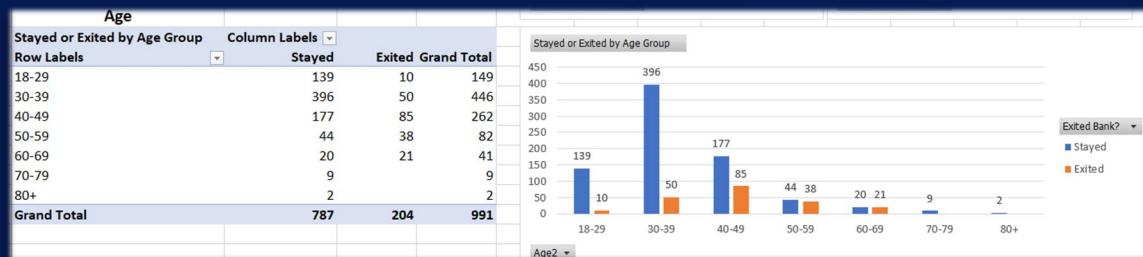
SKILLS

Data Ethics
Data Mining and Cleaning
Predictive Analysis
Time-Series Analysis and Forecasting

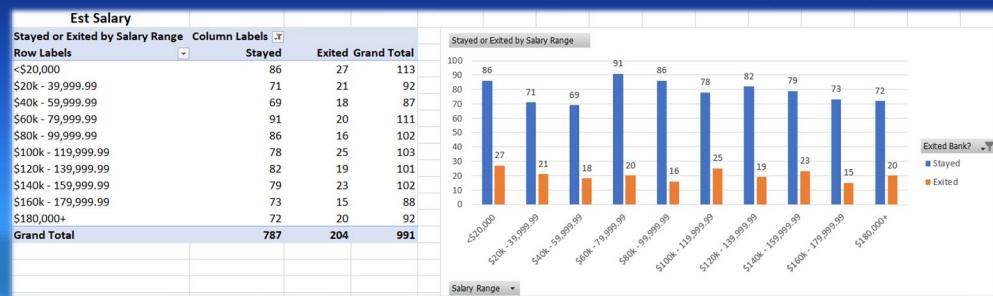
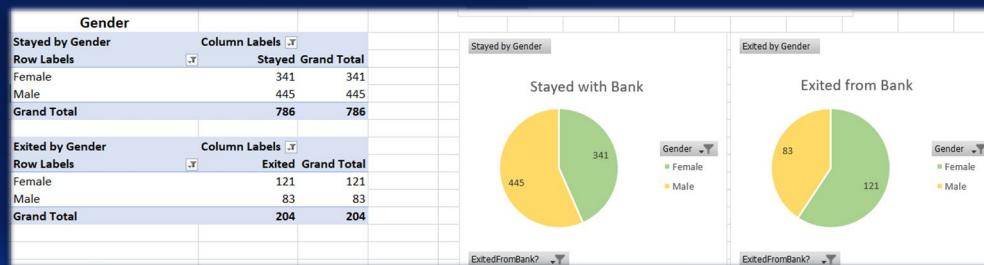


Pig E. Bank Analysis

Nearly half the customers age 50-59 and 60-69 have exited. The likelihood of exiting the bank clearly starts to increase within the 40-49 age group, indicating a high exit risk. ======>

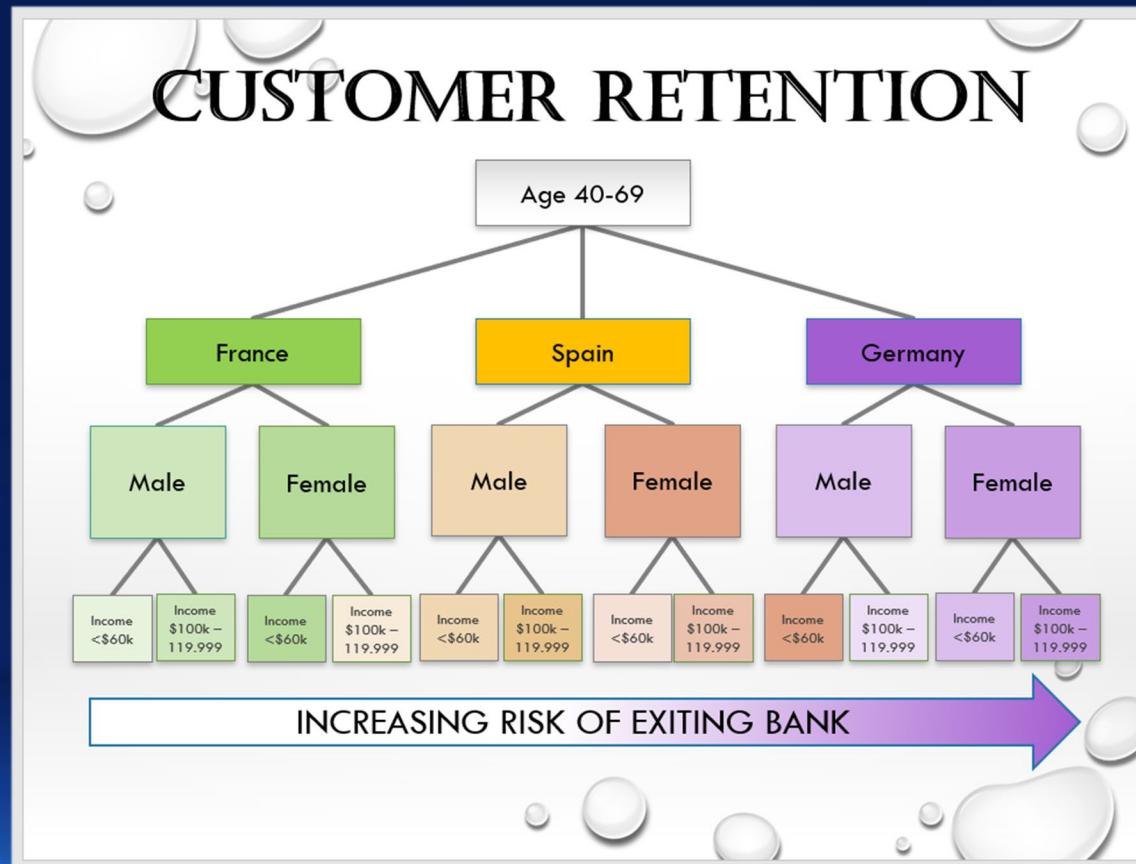


- The highest ratio of Germany's total population has exited, followed by Spain and France, respectively.



- We can generalize this data for the purpose of the analysis, stating the \$100k bracket as the highest exit risk, followed by income less than \$60k.

As indicated by the chart findings above, customers aged 40-69 show the overall highest likelihood of exiting, followed by customers based in Germany. All three countries are shown in the **decision tree**, by order of risk, in effect including the entire 40-69 population. These higher risk factors are further expanded by gender, and finally according to the two higher risk income brackets. These show higher exit by a narrow margin, although still significant.



Statistics and Conclusion

Age

Exited		Stayed	
Mean	45	Mean	38
Median	45	Median	36
Max	69	Max	82
Min	22	Min	18

Notably higher mean and median age indicated among those who exited.

Balance

Exited		Stayed	
Mean	\$90,239.22	Mean	\$74,830.87
Median	\$112,433.97	Median	\$93,147.00
Max	\$213,146.20	Max	\$197,041.80
Min	\$0.00	Min	\$0.00

Mean, median, and max balance are somewhat higher with exited customers.

Credit Score

Exited		Stayed	
Mean	637	Mean	652
Median	644	Median	657
Max	850	Max	850
Min	376	Min	411

Slightly lower mean, median and min credit scores found with exited customers.

- The highest ratio of Germany's total population has exited, followed by Spain and France, respectively.
- Clearly a higher percentage of exited customers are female, concurring with the higher male percentage among those who stayed.
- Close to half the total population of customers age 50-59 and 60-69 have exited. While the total of these age groups is notably smaller by comparison, the likelihood of exiting the bank clearly starts to increase within the 40-49 age group. This indicates a high exit risk.
- There appears to be no correlation with credit card status, among customers who stayed or exited, as the ratio of credit card holders is nearly identical among both customers who stayed or exited.
- While most stayed to exited ratios are comparable among the various income brackets, the highest exit percentages are as follows: \$100k - 24.2%; <\$20k - 23.2%; \$20k - 22.8%; \$40k - 20.6%; We can generalize this data for the purpose of the analysis, stating the \$100k bracket as the highest exit risk, followed by income less than \$60k.

Salary

Exited		Stayed	
Mean	\$97,155.20	Mean	\$98,942.45
Median	\$100,375.40	Median	\$98,368.24
Max	\$199,725.39	Max	\$199,661.50
Min	\$417.41	Min	\$371.05

Only slight salary variations indicated between those who stayed or exited.

FINAL REPORT



UFO Sightings Overview



OBJECTIVE

To build an interactive dashboard that will visually showcase well-curated results of an advanced exploratory analysis conducted in Python.

Research Hypothesis:
If specific geographic areas have many recorded sightings, then a majority of these will have common descriptive attributes, such as shape, motion, or sound.



QUESTIONS

- Is there an identifiable pattern to the sightings, such as certain days of the week? If people are on their way home from a night on the town, are they more likely to see little green men?
- Are certain shapes or descriptive factors of UFO sightings, commonly reported in various geographic regions?
- Can the duration of the UFO event be predicted from the available data?
- Can correlations be identified, such as proximity to a military base, or celestial event?



DATA

[UFO Sightings Around the World](#)

Compiled into Kaggle from:
[NUFORC](#)

DATA PROFILE



TOOLS

Jupyter Notebooks

Python

Excel

Tableau



SKILLS

Sourcing Open Data

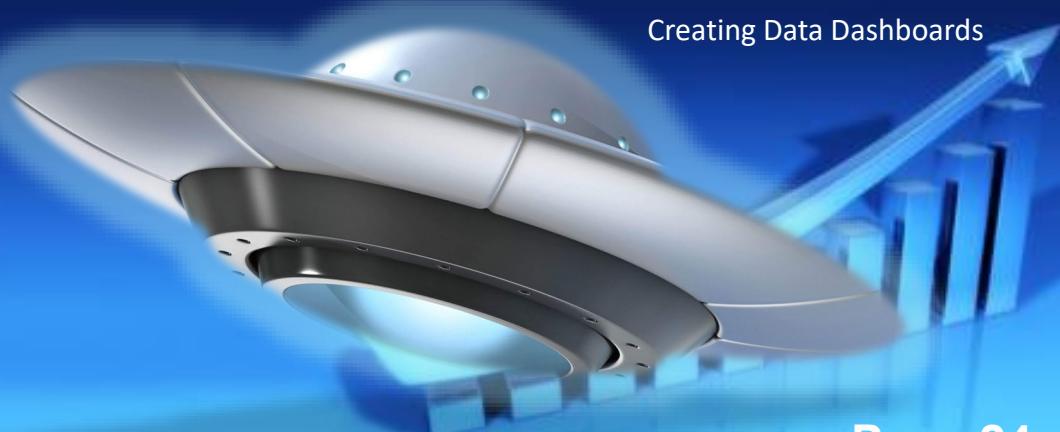
Geographical Visualizations

Filtering, Cleaning and Summarizing Data

Unsupervised Machine Learning: Clustering

Analyzing Time Series Data

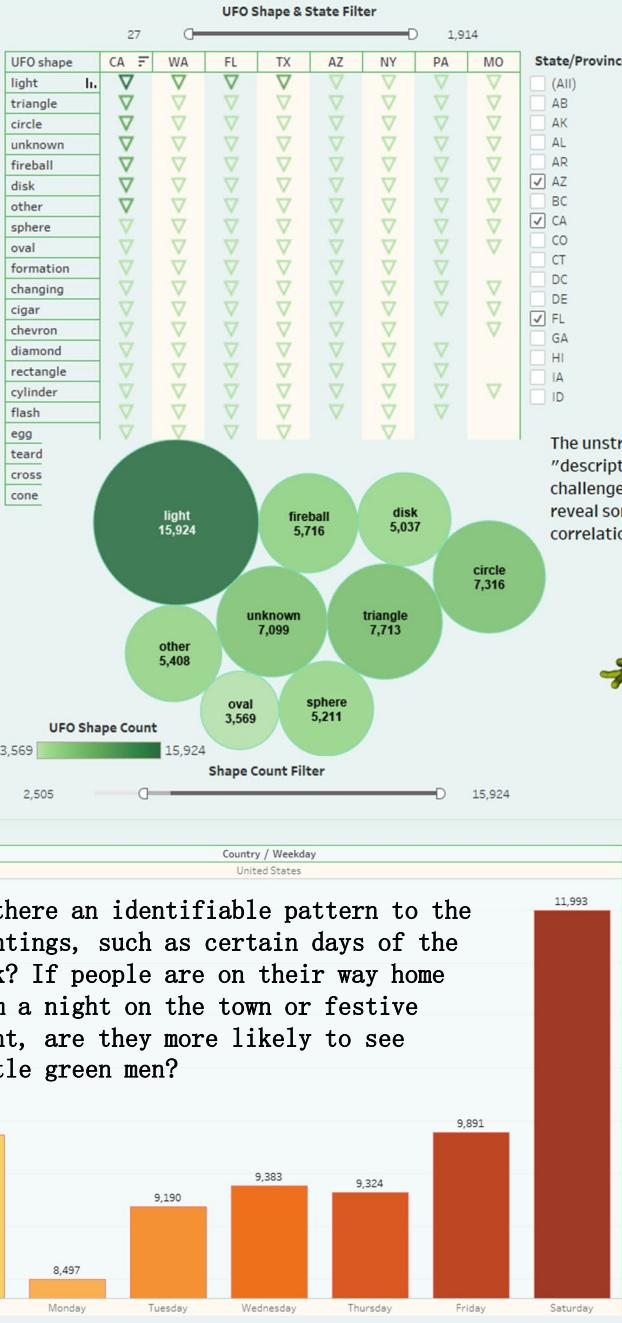
Creating Data Dashboards



UFO Sightings Analysis



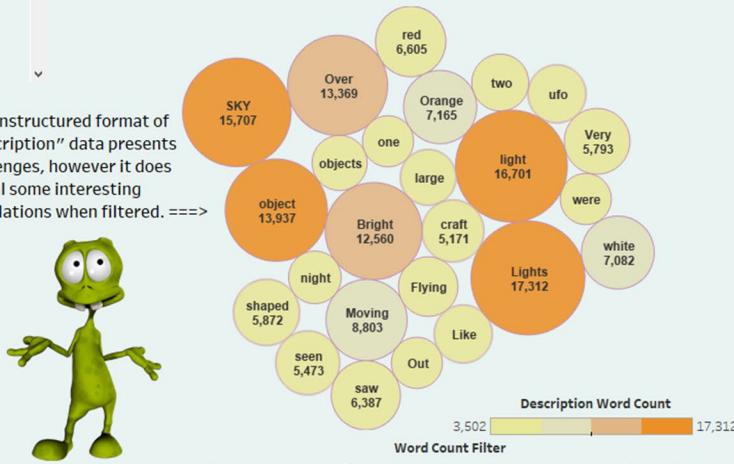
- * A fascination with UFO phenomena, including all of the theories, speculation, and possible explanations therein, continues to grow within the public counterculture of UFO enthusiasts.
- * Ranging from plausible to extravagant, often reading like modern science fiction, these events have been well documented and heavily investigated, over the last century in particular.



Are certain shapes, or other descriptive factors of UFO sightings, more frequently reported in various geographical regions?

<== With the top reported states selected by default, we see that many lights, triangles, circles, disks, & fireballs have been spotted in their skies.

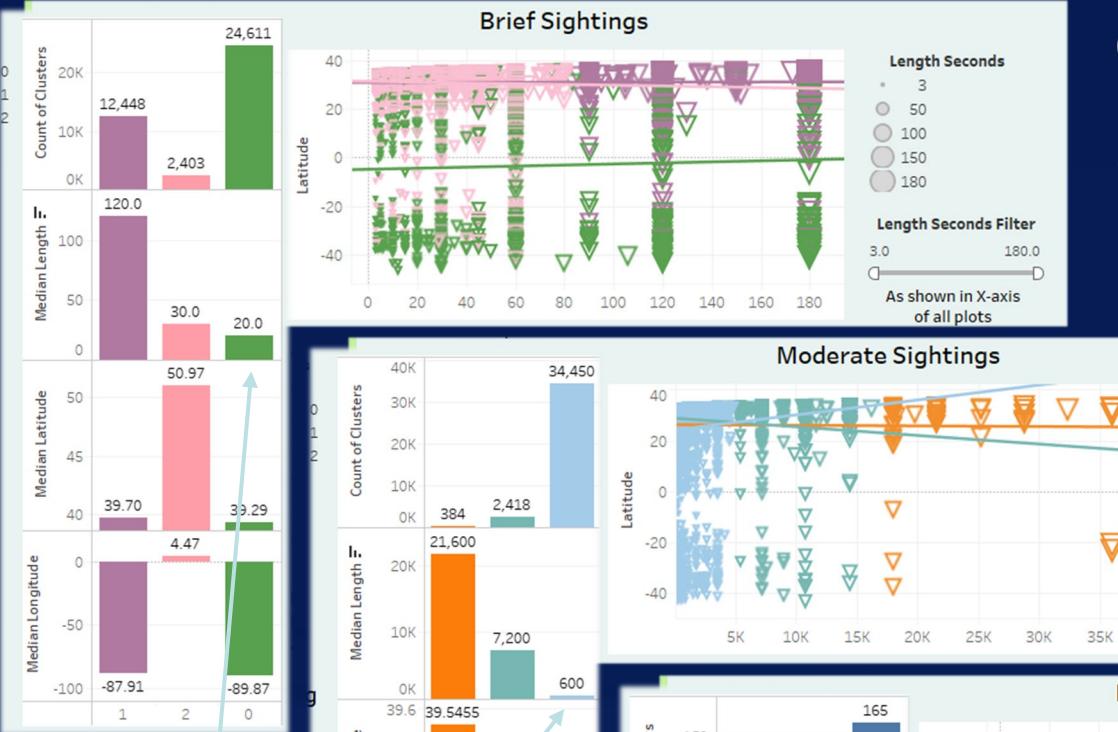
Also noteworthy are the frequency of less typical shapes and descriptors, such as cigar, chevron, cylinder, orange, & red.



Is there an identifiable pattern to the sightings, such as certain days of the week? If people are on their way home from a night on the town or festive event, are they more likely to see little green men?

➤ Viewing individual countries with the highest number of sightings, we see a clear increase on weekends.

This would indicate more encounters on late-night Fridays and Saturdays, often occurring after midnight.



The shortest sighting duration for each range is actually the common median value in the largest cluster:

~ Brief: 20 seconds

~ Moderate: 10 minutes

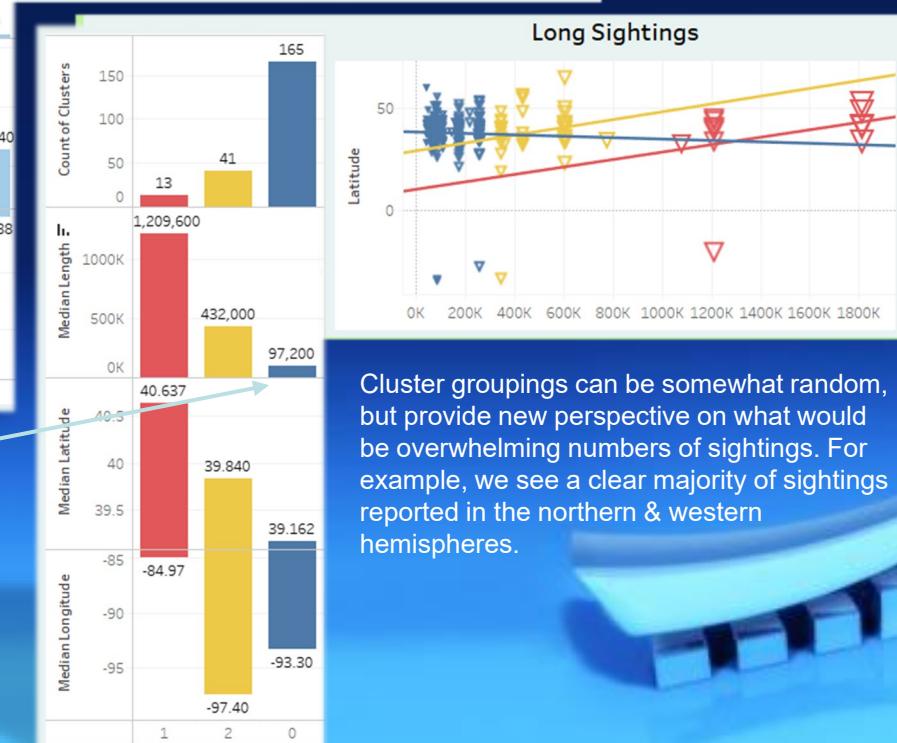
~ Long: 27 hours

Can the duration of the UFO event be predicted from the other available data?

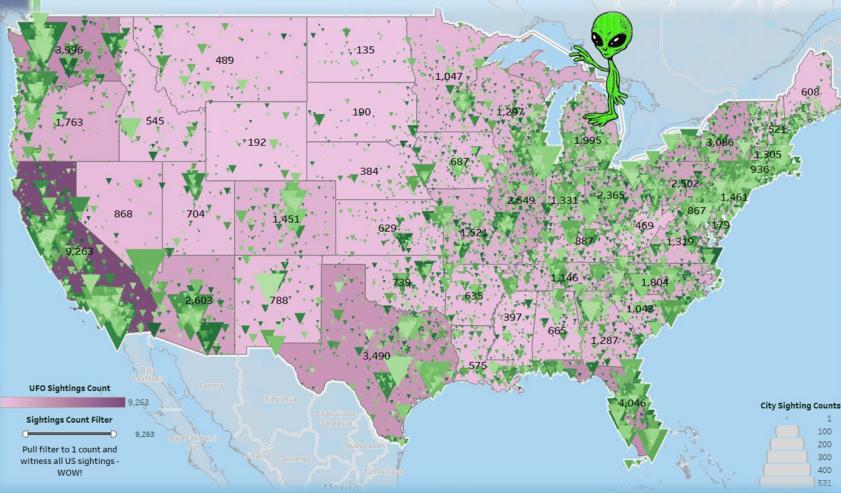
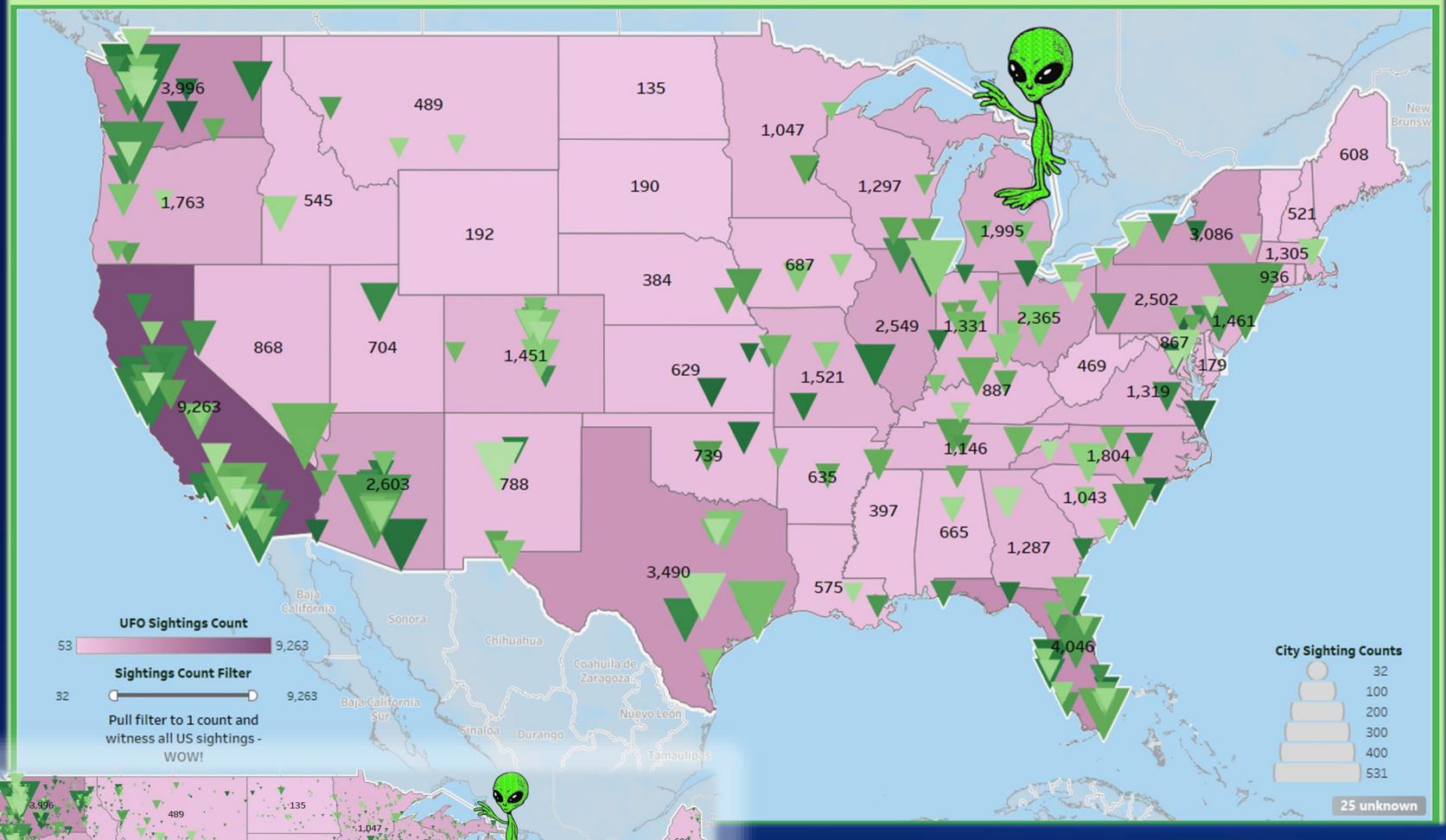
Due to over 80,000 reported observations in the raw data, subsets of duration times were used for cluster analyses, as follows:

- ~ **Brief Sightings:** Less than or equal to 3 minutes
- ~ **Moderate Sightings:** Between 3 minutes & 12 hours
- ~ **Long Sightings:** Greater than or equal to 12 hours

The cluster trend lines will display various contrasting skews, depending on how filtering is applied, which can reveal how sightings of any given duration may be more frequent in certain latitude or longitude.



Cluster groupings can be somewhat random, but provide new perspective on what would be overwhelming numbers of sightings. For example, we see a clear majority of sightings reported in the northern & western hemispheres.



With over 68,000 reported sightings, the United States is the clear leader in UFO phenomena, claiming over 85% of documented events worldwide.

Summary and Proposals

FINDINGS

- ❖ For the eternal UFO enthusiasts, California, Florida, Washington, Texas, and New York are the hot-spot states for sightings. Significantly less than the US, Canada and the United Kingdom have also reported frequent sightings by comparison to other countries. While they have occurred consistently every day of the week, your odds of an encounter will increase Friday through Sunday.
- ❖ A noticeable absence of reported sightings in the southern Nevada desert. This region includes the US air test ranges of Nellis AFB, and Groom Lake, previously known as Area 51. Considering much activity here would be classified by the US government, any reported sightings were likely suppressed by the [National UFO Reporting Center](#).
- ❖ The sightings have been known to take many forms with various descriptions, however you should look to the skies for bright orange, white, and red lights, triangles, circles, and disks.
- ❖ Finally, while there have been many documented sightings of extended duration, look quickly and have the camera ready, as the vast majority are brief, typically lasting 20 seconds or less.

FURTHER ANALYSIS

- ❖ Compare and contrast state populations with sighting counts, to identify regions with highest sightings per capita.
- ❖ As additional recent and future sightings are documented, a more refined and focused analysis could further identify correlations with these events, and possibly help determine which if any, are actually of extraterrestrial origin.

FINAL REPORT





Thank You

Jay Haygood

jay.haygood@outlook.com

