Rajit Sikka
Data Analytics Portfolio

Projects

GAMECO 

Analyzed global video game sales data and created a marketing plan

INFLUENZA SEASON 

Used historical Influenza data to prepare a medical staff deployment schedule

ROCKBUSTER 

Analyzed Rockbuster historical customer data to create an effective movie rental launch strategy

INSTACART 

Analyzed Instacart customer demographic and sales data to profile customers and their habits

PIG E BANK 

Analyzed Pig E Bank customer data to figure out the factors that lead to client loss.

WORLD HAPPINESS REPORT 

Analyzed World Happiness Reports (2015-2022) to learn more about the factors that influence a countries happiness.

GameCo Project Overview

Context

GameCo is a video game development company trying to understand how their new games will fare in the global market (primarily North America, Japan, and Europe).

Objective

Create a marketing plan for GameCo's new games going into 2017 based on historical video game sales data.

Tools

(Excel)

Techniques

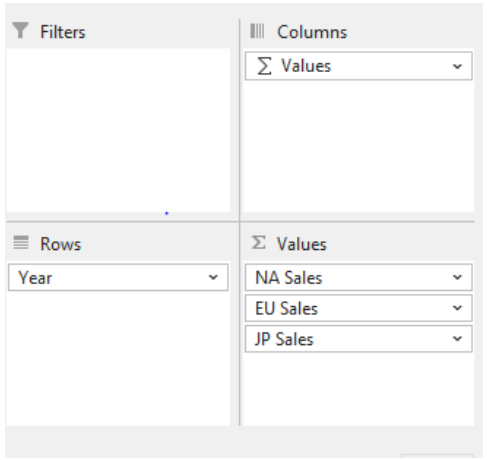
- Grouping Data
- Summarizing Data
- Descriptive Analysis
- Hypothesis Testing
- Visualizing Results in Excel
- Presenting Results

GameCo Process Overview

A look into the main process for this project, **testing a hypothesis and visualizing my findings using Excel.**

Hypothesis: Video game sales have **NOT** changed over time for each region.

1. Pivot Table Values



Step 1: Create a Pivot table from the main dataset, looking at video game sales for each major region.

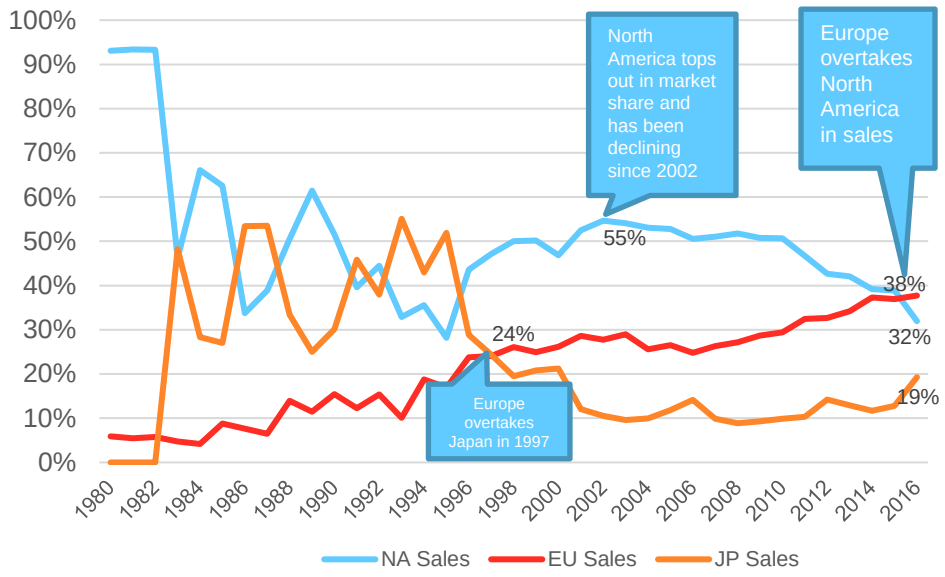
2. Pivot Table Data

Row Labels	NA Sales	EU Sales	JP Sales
1980	93%	6%	0%
1981	93%	5%	0%
1982	93%	6%	0%
1983	46%	5%	48%
1984	66%	4%	28%
1985	63%	9%	27%
1986	34%	8%	53%
1987	39%	6%	53%
1988	51%	14%	33%
1989	61%	11%	25%
1990	52%	15%	30%
1991	40%	12%	46%
1992	44%	15%	38%
1993	33%	10%	55%
1994	36%	19%	43%
1995	28%	17%	52%
1996	44%	24%	29%
1997	47%	24%	24%
1998	50%	26%	20%
1999	50%	25%	21%
2000	47%	26%	21%
2001	52%	29%	12%
2002	55%	28%	11%
2003	54%	29%	10%
2004	53%	26%	10%
2005	53%	27%	12%
2006	50%	25%	14%
2007	51%	26%	10%
2008	52%	27%	9%
2009	51%	29%	9%
2010	51%	29%	10%
2011	47%	32%	10%
2012	43%	33%	14%
2013	42%	34%	13%
2014	39%	37%	12%
2015	39%	37%	13%
2016	32%	38%	19%

Step 2: View table and ensure I have all the necessary values.

3. Line Chart

Figure 4: Proportional Global Sales over Time



Step 3: Visualize findings with a line chart, adding captions and labels for easy viewing.

Conclusion:

By creating a pivot table to gather the relevant variables, then viewing the pivot table that was created; ensuring its exactly what I want, then finally using the best type of chart to visualize the data (line chart) I was able to find that **video game sales have changed significantly over time for each region**

GameCo Project Insights

More analysis, highlighting some of the main questions within the project.

Key Questions

What are the most popular genres?

What % of global sales do the top 4 genres hold?

What are the most popular platforms?

Data

Figure 2: Global Sales by Genre (2012 – 2016)

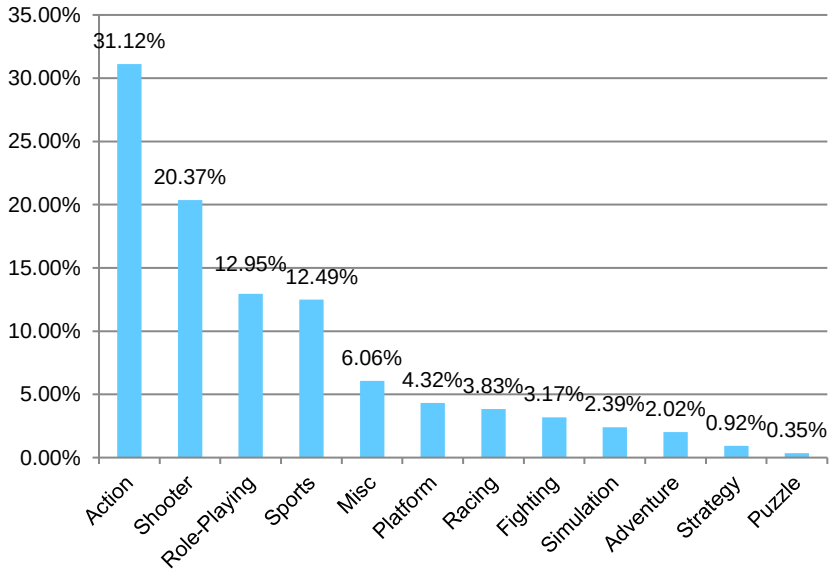


Figure 3: Proportion of Global Sales by Region for Top 4 Genres (2012-2016)

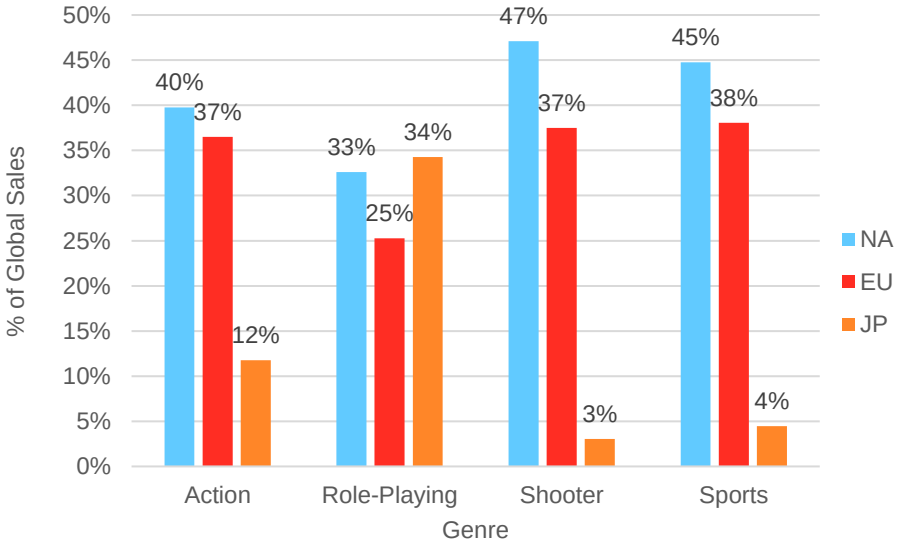
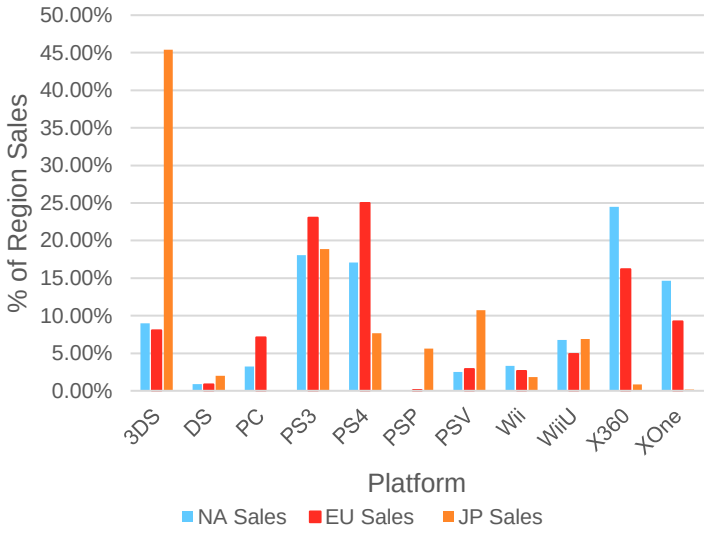


Figure 7: Regional Sales Distribution by Platform (2012 – 2016)



GameCo Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. Europe has emerged as the leader in global market share (38%). Prioritize Sports and Shooting games specifically on the PS4
2. North America's global market share (32%) has been declining over time, but still holds a strong position. Prioritize Shooting games on the PS4 and Sports games on both the PS4 and Xbox One
3. Japan has a smaller global market share and realistically Role-Playing games on the Nintendo 3DS are the only games that should be prioritized

Takeaways

1. It's important to understand a business's needs.
2. Pivot tables in Excel are useful for isolating data to do specific analysis.
3. Good presentation comes with being very detail oriented. Being specific to leave no confusion for the reader is also important.

Deliverables

Project Brief

Presentation

Influenza Season Project Overview

Context

A medical staffing agency needs to understand trends within influenza season, so they can proactively plan and staff temporary workers to clinics and hospitals across the country.

Objective

Create a medical staff allocation plan for each state in the U.S.A

Tools

(Tableau)

Techniques

- Translating business requirements
- Data Cleaning, Integration, and Transformation
- Forecasting
- Storytelling in Tableau

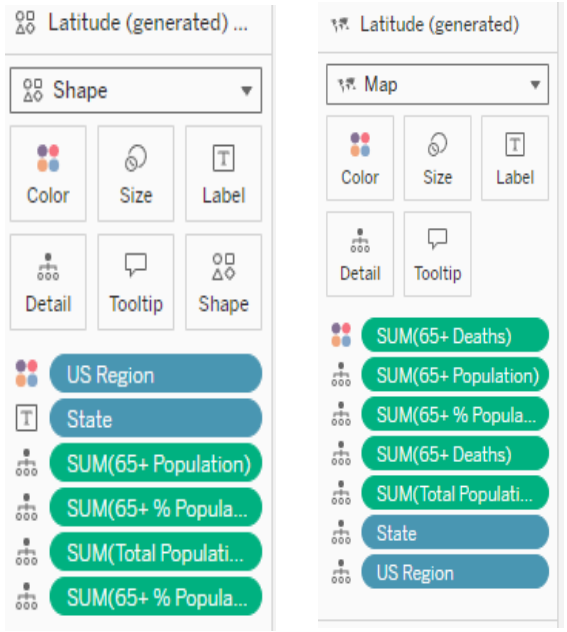
Influenza Process Overview

A look into one of the main tasks for this project, **creating a combination map inside of Tableau**.

1. New Sheet and Map Chart

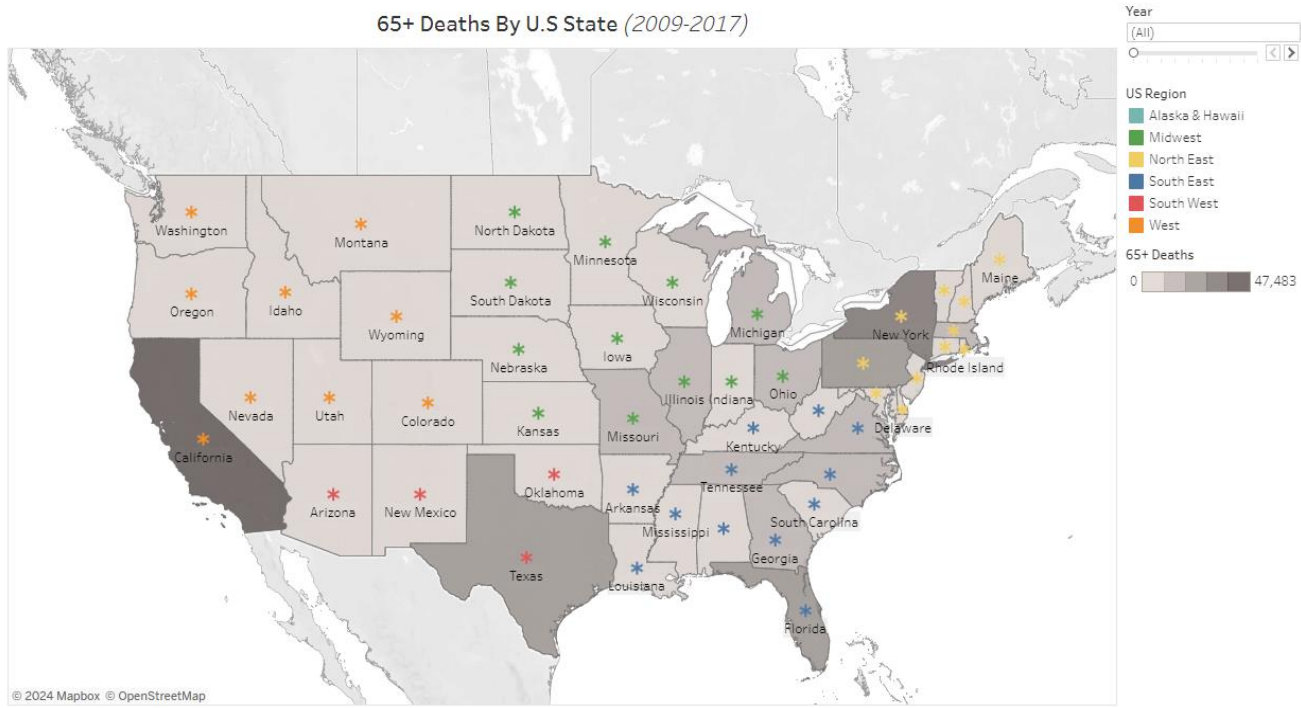
Step 1: After importing the Excel tables into Tableau. I created a new sheet and started with a latitude map. I duplicated the Latitude field so I could have a combination map.

2. Marks for Visualization



Step 2: Having two latitude fields for our map allows us to add different marks for each one. For the Map field I added our 65+ deaths and for the shape field I added US Region.

3.Combination Map



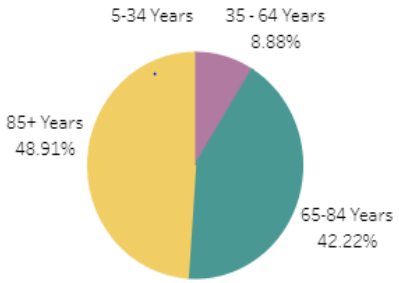
Result: Pictured above is the result of our previous steps. From our chart we can see how influenza deaths are congested in popular states (due to their higher population). Most of the West and Midwest have significantly less influenza deaths compared to New York, California, and Texas.

Influenza Project Insights

A brief look at some of the main questions answered in this project, alongside their charts

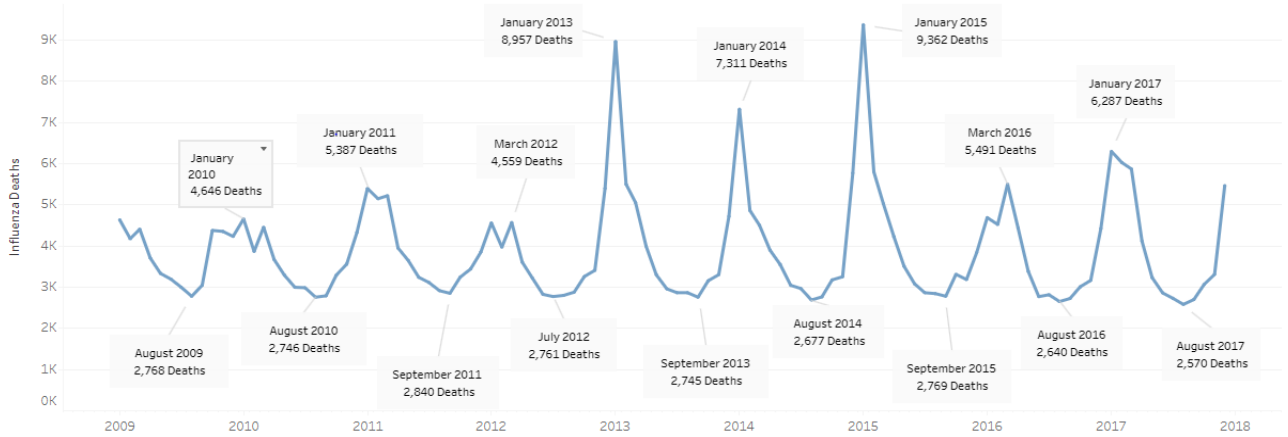
What Age Group is most vulnerable for influenza?

Total U.S Influenza Deaths by Age Group
(2009-2017)

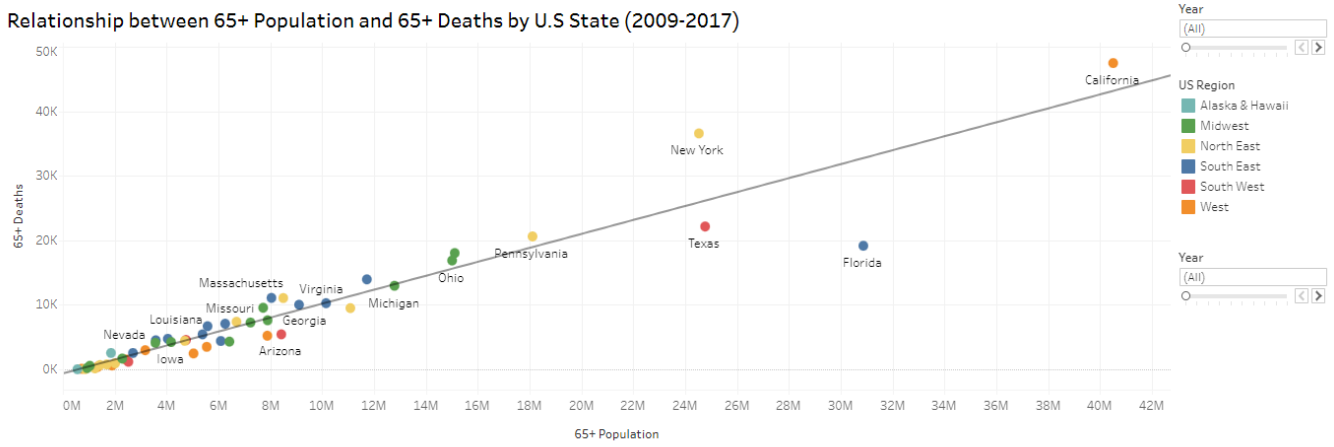


Age Group
- 5-34 Years
- 35 - 64 Years
- 65-84 Years
- 85+ Years

Total US Influenza Deaths over Time (2009 -2017)



Relationship between 65+ Population and 65+ Deaths by U.S State (2009-2017)



(Positive Correlation between 65+ Population and 65+ Deaths)

When is influenza season?

How does the number of 65+ U.S population impact Influenza Deaths?

Influenza Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. Influenza Season begins around September and tends to peak in January. Influenza Season is very volatile, quickly increasing and decreasing in cases during this period.
2. Ages 65+ are the most vulnerable age group as they made up about 90% of total U.S Influenza Deaths.
3. Population is the main driver for influenza deaths. As 65+ U.S population increases, 65+ U.S influenza deaths will increase as well.

Takeaways

1. Tableau Storyboards is a better PowerPoint when you are present lots of data together. Being able to highlight certain parts of a data (ex: U.S Region) is nice as it can work with both charts at the same time.
2. The Forecasting feature inside of Tableau is also very useful when trying to predict future trends (forecasting influenza season).
3. Being specific when labeling anything is important. 65+ Influenza Deaths vs. 65+ U.S. Influenza Deaths makes a big difference.

Deliverables

Influenza Season: Interim Report
Tableau Storyboard

Rockbuster Project Overview

Context

Rockbuster Stealth is a movie rental company that used to have stores around the world. Facing stiff competition from services such as Netflix and Amazon Prime, the Rockbuster Stealth management team is planning to use its existing licenses to launch an online video rental service to stay competitive.

Objective

Analyze Rockbuster's historical data and create and effective launch strategy

Tools

(PostgreSQL, Tableau, and Excel)

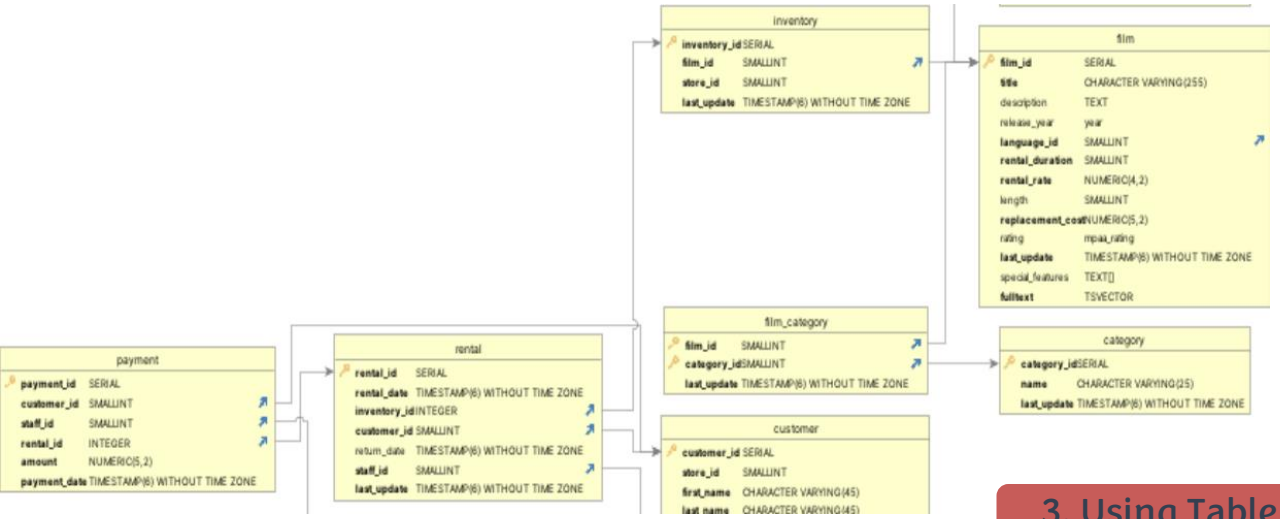
Techniques

- Working with relational databases
- Database querying
- Joining Tables
- Subqueries
- Common table Expressions

Rockbuster Process Overview

A look into the primary process for this project. **Navigating a database through PostgreSQL and turning it into a visualization.**

1. Navigating through SQL database using Data Dictionary



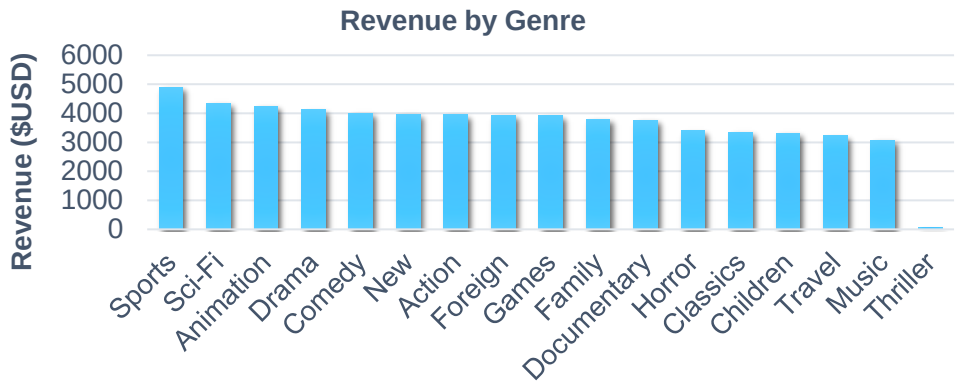
2. Creating appropriate SQL query and exporting table

Query | Query History

```sql
SELECT
F.name AS "Genre",
SUM(amount) AS "Revenue"

FROM Payment A
INNER JOIN rental B on A.rental_id = B.rental_id
INNER JOIN inventory C on B.inventory_id = C.inventory_id
INNER JOIN film D on C.film_id = D.film_id
INNER JOIN film_category E on D.film_id = E.film_id
INNER JOIN category F on E.category_id = F.category_id

GROUP BY F.name
ORDER BY SUM(amount) DESC
```

Genre	Revenue
Sports	4892.19
Sci-Fi	4336.01
Animatior	4245.31
Drama	4118.46
Comedy	4002.48
New	3966.38
Action	3951.84
Foreign	3934.47
Games	3922.18
Family	3782.26
Document	3749.65
Horror	3401.27
Classics	3353.38
Children	3309.39
Travel	3227.36
Music	3071.52
Thriller	47.89

3. Using Table to create a visualization inside of Excel.



Notes: In this example, our objective is to find **Revenue by Movie genre**. Since the data is not in the same table, we must use **INNER JOINS with SQL** to connect the tables we want. Once we have tables are properly connected, we can query for the data we need.

Rockbuster Project Insights

A brief look at some of the main questions answered in this project, alongside their charts

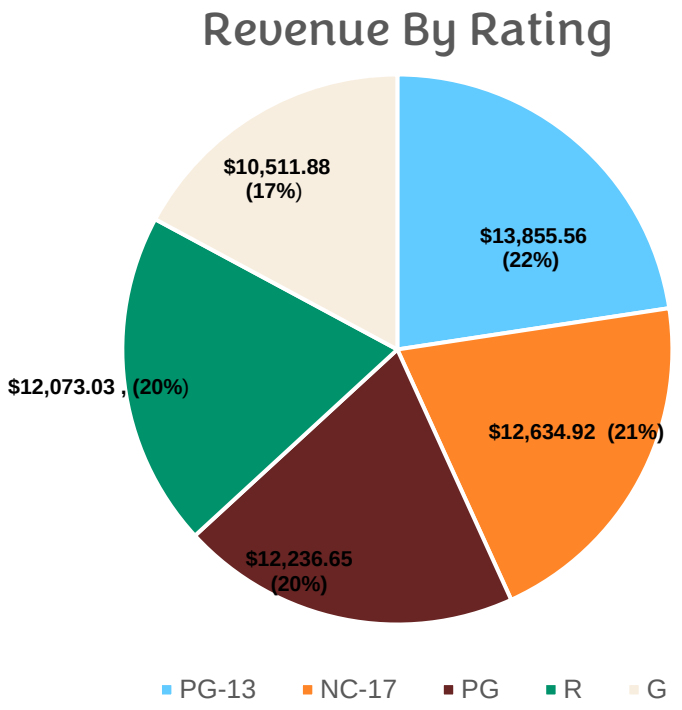
Rockbuster Movie Statistics

- Total Company Revenue: $61,312
- 1000 Films
- Average Replacement Cost: $19.98
- Average Customer Value: $104
- 17 Genres
- Average Rental Duration: 5 days
- Average Movie length: ~2 hours (115 min)
- 599 Customers (108 Countries)
- Average Rental Rate: $3/rental movie

What are the top 10 countries with the most customers?

Country	Total Customers
India	60
China	53
United States	36
Japan	31
Mexico	30
Brazil	28
Russian Federation	28
Philippines	20
Turkey	15
Indonesia	14

How is revenue distributed across each movie rating?



Rockbuster Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. Rockbuster has customers across the world. Creating a referral system (especially the top 10 countries listed in 13) could help expand Rockbusters customer base.
2. The Thriller genre should be dropped from Rockbuster's movie library as it had only $48 in revenue, the other 16 genres had a minimum of $3,000 of revenue.

Takeaways

1. An ERD (Entity Relationship Diagram) is absolutely necessary when trying to navigate around a database to get the queries you want.
2. When working with units of measurement, pick what's faster to understand if you're able to round to a higher unit. (ex: ~2 hours is quicker to understand than 115 minutes. So, putting ~2 hours first is better.)

Deliverables

Presentation
Excel SQL Workbook
Data Dictionary

Instacart Project Overview

Context

Instacart stakeholders are primarily interested in the variety of their customers and their purchasing behaviors. They want to be able to use more targeted marketing towards different types of customers.

Objective

Perform exploratory analysis on Instacart customers to create customer profiles that can be used for targeted marketing.

Tools

(Python and Excel)

Techniques

- Data Wrangling
- Data Merging
- Deriving variables
- Grouping and Aggregating Data
- Excel Reporting

Instacart Process Overview

A look into one of the main process for this project, **creating new columns and dataframes in Python to build a visualization (ex: stacked bar chart)**

1. Adding a Region Column to main dataframe using loc function

```
# Defining the list of states for each region
east_region = ['Maine','New Hampshire', 'Vermont', 'Massachusetts', 'Rhode Island', 'Connecticut', 'New York', 'Pennsylvania', 'New Jersey']
midwest_region = ['Wisconsin', 'Michigan', 'Illinois', 'Indiana', 'Ohio', 'North Dakota', 'South Dakota' , 'Nebraska', 'Kansas', 'Minnesota', 'Iowa', 'Missouri']
south_region = ['Delaware', 'Maryland', 'District of Columbia', 'Virginia', 'West Virginia', 'North Carolina', 'South Carolina', 'Georgia', 'Florida','Kentucky', 'Tennessee', 'Mississippi', 'Alabama','Oklahoma', 'Texas
                'Arkansas', 'Louisiana']
west_region = ['Idaho', 'Montana', 'Wyoming', 'Nevada', 'Utah', 'Colorado', 'Arizona', 'New Mexico', 'Alaska', 'Washington', 'Oregon', 'California', 'Hawaii']

#Creating Region Column
df_comC['Region'] = 'N/A'

#Using Loc function to define all East States
df_comC.loc[df_comC['State'].isin(east_region), 'Region'] = 'East'

#Using Loc function to define all Midwest States
df_comC.loc[df_comC['State'].isin(midwest_region), 'Region'] = 'Midwest'

#Using Loc function to define all West States
df_comC.loc[df_comC['State'].isin(west_region), 'Region'] = 'West'

#Using Loc function to define all South States
df_comC.loc[df_comC['State'].isin(south_region), 'Region'] = 'South'

#Checking Region FLag
df_comC['Region'].value_counts(dropna = False)
```

```
Region
South      10791885
West        8292913
Midwest     7597325
East        5722736
Name: count, dtype: int64
```

2. Adding an Age Range Column to dataframe

```
# Creating New Column "Age Range" and Defining Young Adults
df_5plus_com.loc[(df_5plus_com['Age'] > 17) & (df_5plus_com['Age'] <= 35 ), 'Age_Range'] = 'Young Adult'

# Defining Middle Adults for the column
df_5plus_com.loc[(df_5plus_com['Age'] > 35) & (df_5plus_com['Age'] <= 64 ), 'Age_Range'] = 'Middle Adult'

# Defining Late Adults for the column
df_5plus_com.loc[df_5plus_com['Age'] > 64, 'Age_Range'] = 'Late Adult'

# Viewing Age Distribution
df_5plus_com['Age_Range'].value_counts(dropna = False)
```

```
Age_Range
Middle Adult    14030215
Young Adult      8738805
Late Adult       8195544
Name: count, dtype: int64
```

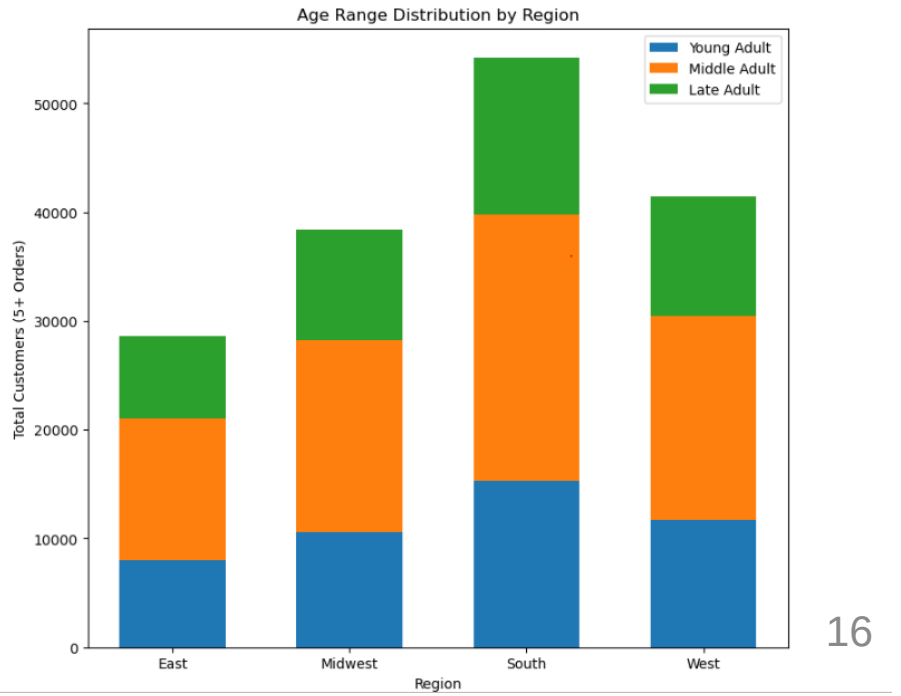
3. Creating new dataframes to group age range and region

```
#Creating Dataframe for only Young Adult customers grouped by region
young_adult_region = df_qual[df_qual['Age_Range'] == 'Young Adult'].groupby('Region').size()

#Creating Dataframe for only Middle Adult customers grouped by region
middle_adult_region = df_qual[df_qual['Age_Range'] == 'Middle Adult'].groupby('Region').size()

#Creating Dataframe for only Young Adult customers grouped by region
late_adult_region = df_qual[df_qual['Age_Range'] == 'Late Adult'].groupby('Region').size()

#Testing new dataframes for accuracy
late_adult_region
```

```
Region
East        7580
Midwest    10169
South      14433
West       11068
dtype: int64
```

4. Creating a Stacked bar chart

```
# Visualizing Distribution of Special Profiles
plt.figure(figsize = (9,8))
plt.bar(young_adult_region.index, young_adult_region.values, 0.6, label = "Young Adult")
plt.bar(middle_adult_region.index, middle_adult_region.values, 0.6, bottom = young_adult_region.values, label = "Middle Adult")
plt.bar(late_adult_region.index, late_adult_region.values, 0.6, bottom = (young_adult_region.values + middle_adult_region.values), label = "Late Adult")
plt.xlabel('Region')
plt.ylabel('Total Customers (5+ Orders)')
plt.title('Age Range Distribution by Region')
plt.legend()

#Exporting Chart
plt.savefig(os.path.join(path, '04 Analysis','Visualizations', 'age_range_region.png'))
```



Instacart Project Insights

A brief look at some of the main questions answered in this project, alongside their charts

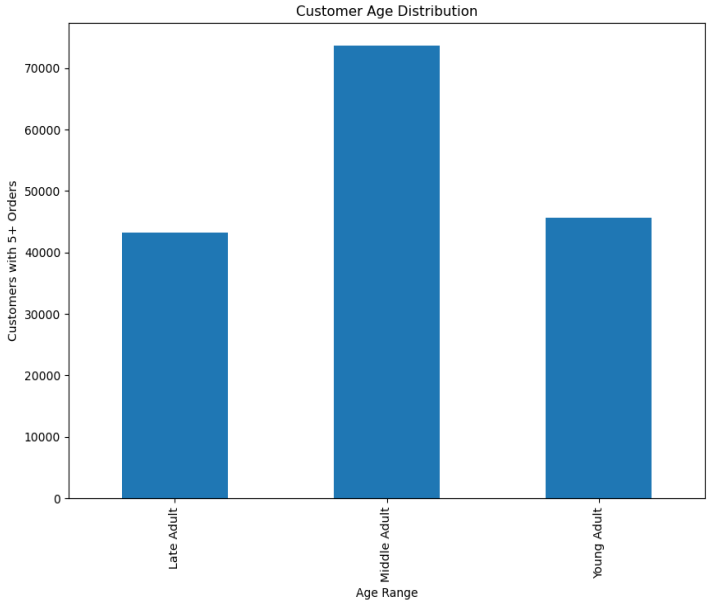
Key Questions

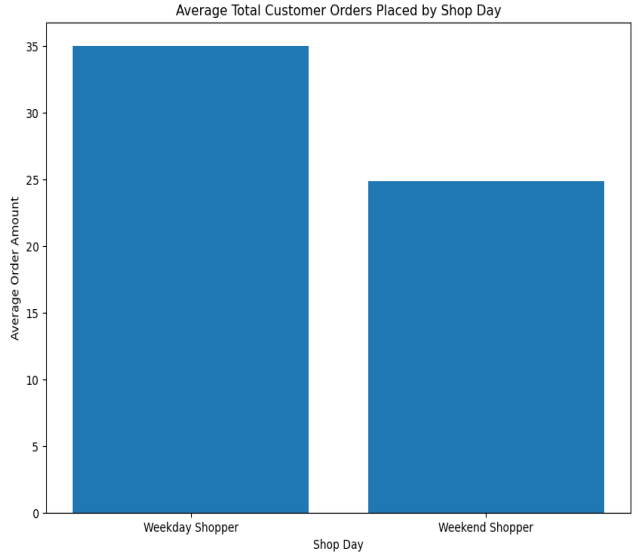
What is the average age of our customer base?

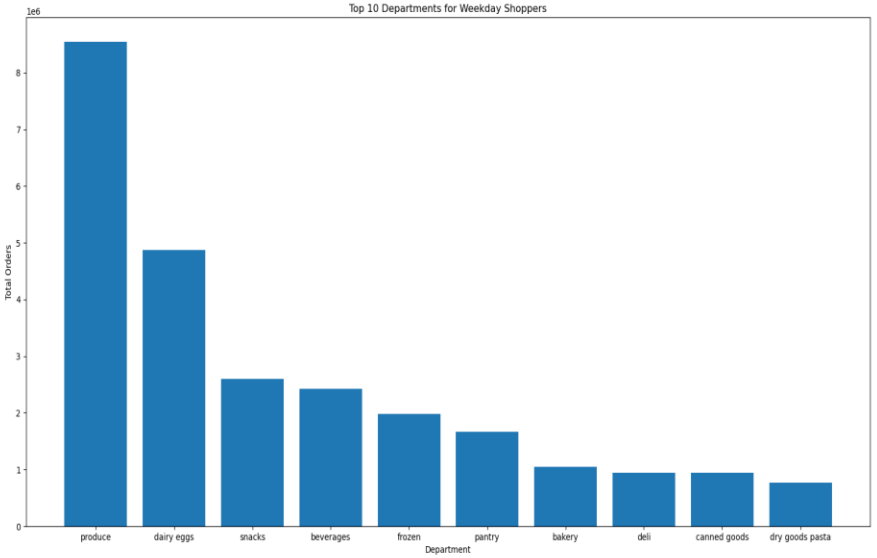
Do our customers shop more on the weekdays or weekends?

What departments are most popular?

Data







Context: Young Adult (18-35), Middle Adult (35-64), Late Adult (65+)

Context: Weekdays are considered Mon - Thu, while Weekends is Fri-Sun

Instacart Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. Customers shop throughout the week, but more on Weekdays

2. The top 5 departments are consistently produce, dairy/eggs, beverages, snacks, and frozen items. There are some deviations between income ranges, but these 5 remain at top.

3. Most of our customer base is middle/upper class.

Takeaways

1. Python is incredibly complex and versatile. I feel like I only scratched the surface when figuring out how to do stacked bar chart
2. Having good comments for your code is not only helpful for the other person reading the code, but also yourself to quickly understand what a code block did.
3. One struggle I had in Python was running out of disk space on my SSD. I found out you can use the command prompt to select what drive you want to use Python in and run jupyter-lab from there.

Deliverables

Excel Final Report

Pig E Bank Project Overview

Context

Pig E Bank sales team is trying identify the leading indicators that a customer will leave the bank.

Tools

(Excel)

Objective

Use the client attributes table to identify the top risk factors that contribute to client loss and model them in a decision tree.

Techniques

- Data Ethics
- Data Mining
- Predictive analysis
- Decision Trees

Pig E Bank Process Overview

A look into the main process for this project, **creating a decision tree for factors that impact client loss.**

1. Ensuring Clean Data (+documentation)

Changes Made
Checked for Duplicates, No Duplicates found
Changed abbreviations in country column to full names (FR -> France, ES -> Spain, DE -> Germany)
Deleted Row number 215 from data set as it had blanks
Replaced Rows with "F" with Female and "M" for Male inside of the gender column
Added "N/A" for columns with blank last names (as last name doesn't really matter for analysis)
Removed two rows that didn't include their credit score
Removed row because of null age
Removed row 22 because of blank salary
Removed Row number column because of redundancy
Replaced Rows with Age 2 with 20

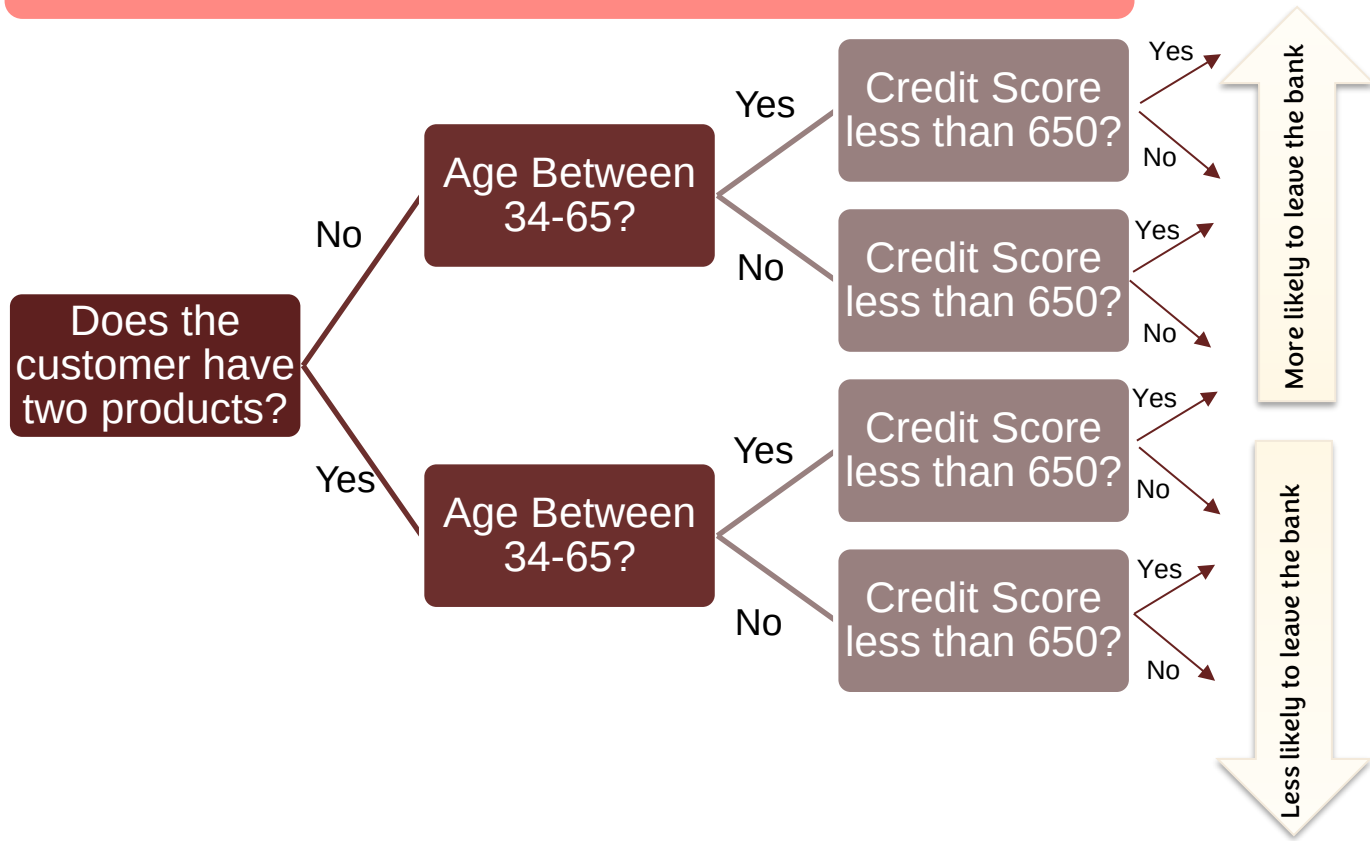
2. Using pivot tables to find the most impacting factors that lead to client loss.

Credit Score	With the Bank	Left the Bank
350-449	70.00%	30.00%
450-549	76.51%	23.49%
550-649	78.28%	21.72%
650-749	80.91%	19.09%
750-850	83.02%	16.98%

Age	With the Bank	Left the Bank
18-33	92.74%	7.26%
34-49	77.96%	22.04%
50-65	50.45%	49.55%
66-82	82.61%	17.39%

Number of Products	With the Bank	Left the Bank
1	72.39%	27.61%
2	92.79%	7.21%
3	15.15%	84.85%
4	0.00%	100.00%

3. Creating a Decision Tree based on insights



Pig E Bank Project Insights

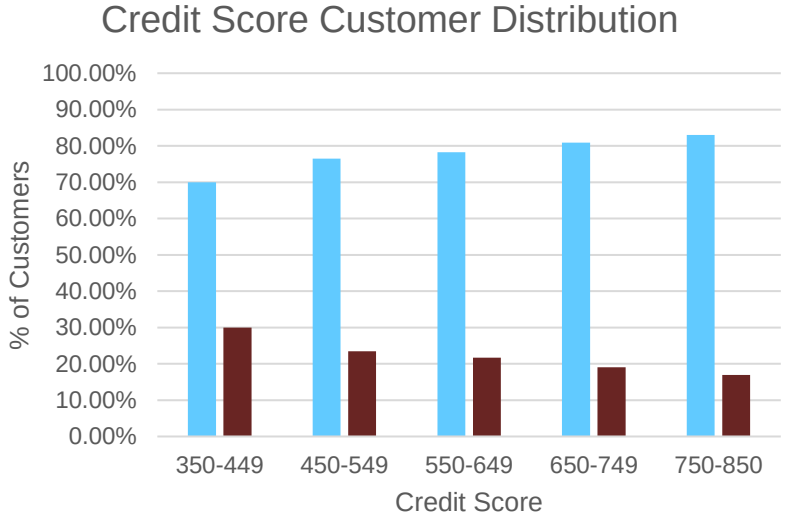
A brief look at some of the main factors that lead to client loss.

Legend:
Customers with the Bank
Customers that have left the bank

Factor 1: # of Products

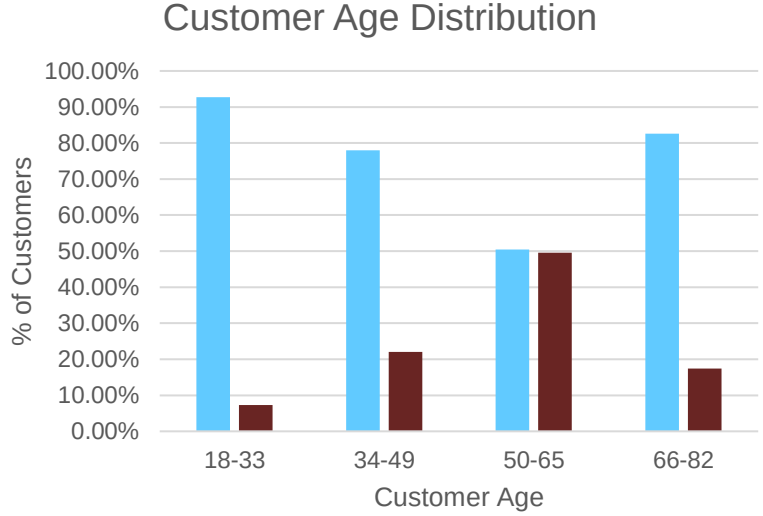
Customers that specifically have 2 products have the lowest chance of leaving the bank

Factor 2: Credit Score



The percentage of customers that leaves that bank decreases as credit score increases. Subsequently, as credit score increases the percentage of customers in the bank increase.

Factor 3: Customer Age



Customers specifically between 34-65 are more likely to leave the bank compared to customers in the other age groups.

Pig E Bank Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. The main factors that lead to client loss are Credit Score, Number of Products, and Age

2. People who leave the bank on average are also 8 years older, have a higher balance, and a slightly lower credit score.

Deliverables

Excel Pig E Bank Report

Takeaways

1. Documenting every step you do (especially cleaning) is very important if you want someone to repeat the process.
2. There are many forms of bias, so it's important to go through all the potential things (stuff like even cultural bias, and measurement bias have their own levels of depth) that you must think through properly before evaluating a dataset.
3. Decision trees are useful when trying to relate multiple factors that can impact a specific instance (client loss)

World Happiness Report Project Overview

Context

Personal project, conducting exploratory analysis on the World Happiness Report from 2015-2022

Tools

(Excel, Python, Tableau)

Objective

Analyze the World Happiness Report and discover insights related to Happiness across each country and the other variables that impact it.

Techniques

- Geographical Visualizations with Python
- Regression Analysis
- Cluster Analysis
- Creating Data Dashboards

World Happiness Report Process Overview

A look into the main tasks for this project, **conducting cluster analysis.**

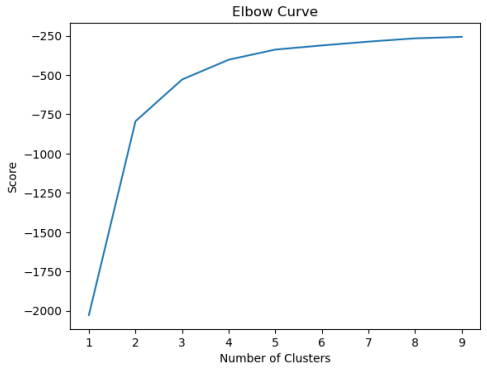
1. Determing the number of clusters needed by using the **Elbow Technique**

```
# Defines the range of potential clusters in the data.
num_cl = range(1, 10)

# Defines k-means clusters in the range assigned above.
kmeans = [KMeans(n_clusters=i) for i in num_cl]
```

```
# Creates a score that represents a rate of variation for the given cluster option.
score = [kmeans[i].fit(subM).score(subM) for i in range(len(kmeans))]

score
```

2. Plotting elbow curve in python.

```
# Plot the elbow curve using PyLab.

pl.plot(num_cl,score)
pl.xlabel('Number of Clusters')
pl.ylabel('Score')
pl.title('Elbow Curve')
pl.show()
```



2. Running k-means algorithm to do cluster analysis

```
# Create the k-means object.

kmeans = KMeans(n_clusters = 5)
```

```
# Fit the k-means object to the data.

kmeans.fit(subM)
```

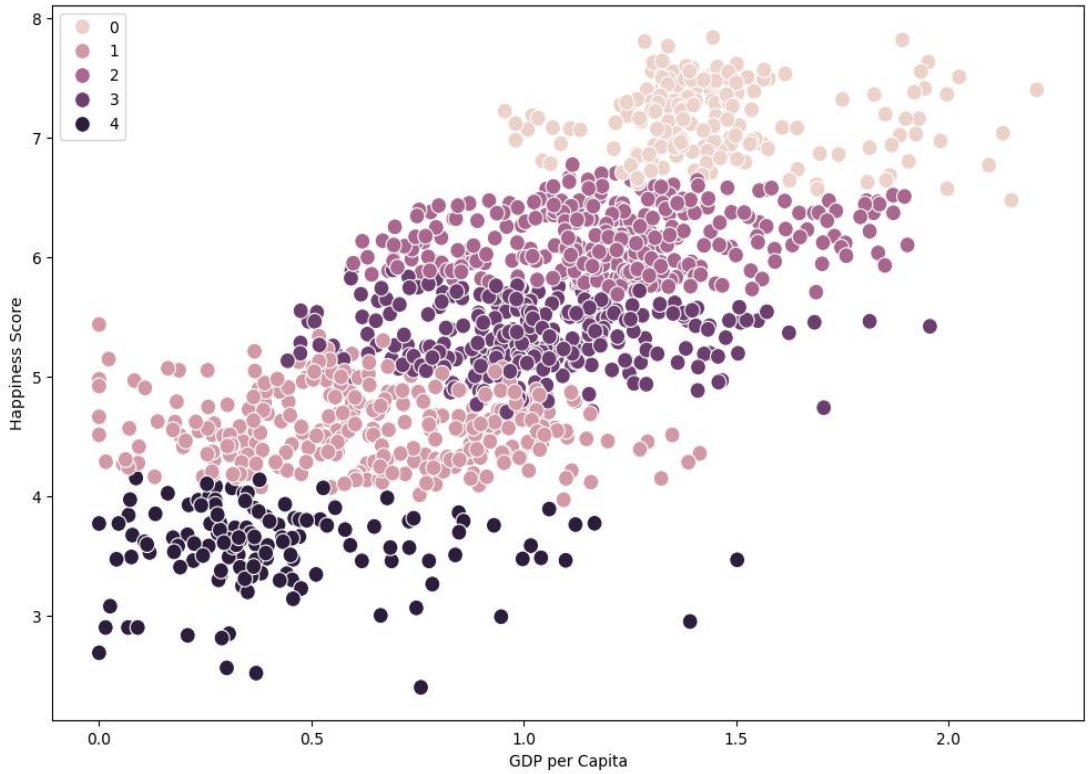
```
KMeans
KMeans(n_clusters=5)
```

```
#Create a new clusters column
subM['clusters'] = kmeans.fit_predict(subM)
```

4. Visualizing k-mean algorithm

```
# Plot the clusters for the "Happiness Score" and "GDP per Capita" variables.

plt.figure(figsize=(12,8))
ax = sns.scatterplot(x=subM['GDP per Capita'], y=subM['Happiness Score'], hue=kmeans.labels_, s=100)
# Here, you're subsetting `X` for the x and y arguments to avoid using their labels.
# `hue` takes the value of the attribute `kmeans.labels_`, which is the result of running the k-means algorithm.
# `s` represents the size of the points you want to see in the plot.

ax.grid(False) # This removes the grid from the background.
plt.xlabel('GDP per Capita') # Label x-axis.
plt.ylabel('Happiness Score') # Label y-axis.
plt.show()
```

5. **Result:** Cluster Analysis of GDP per Capita vs Happiness Score

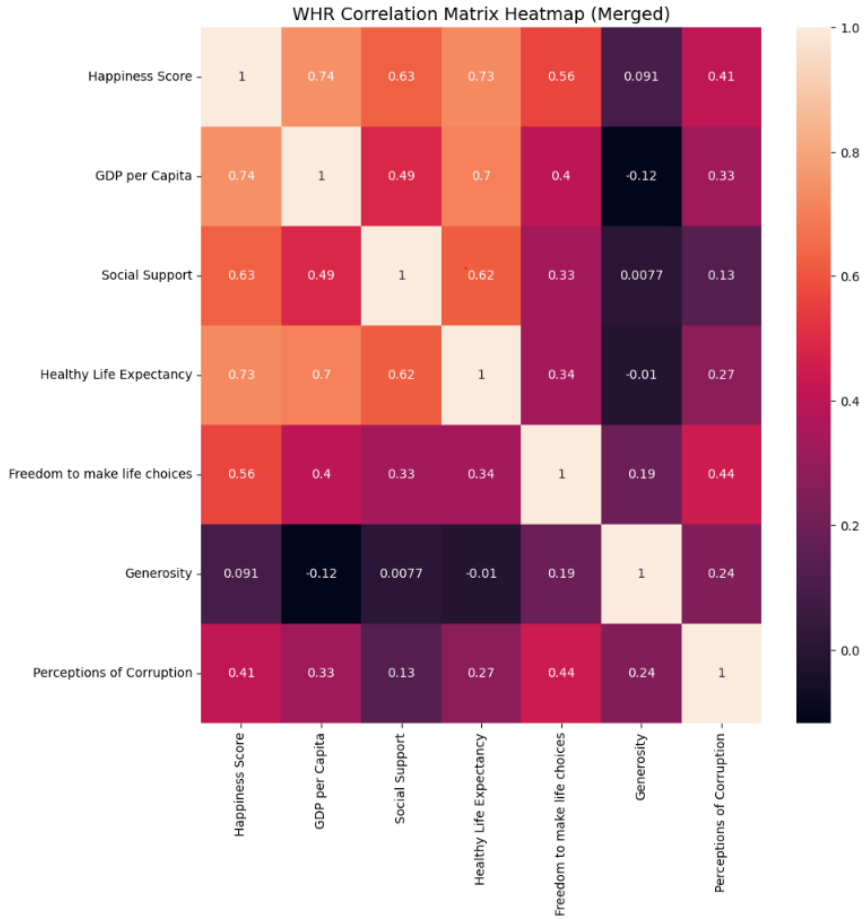


All code for this process (6.5) and others is under Scripts in the Github Repository

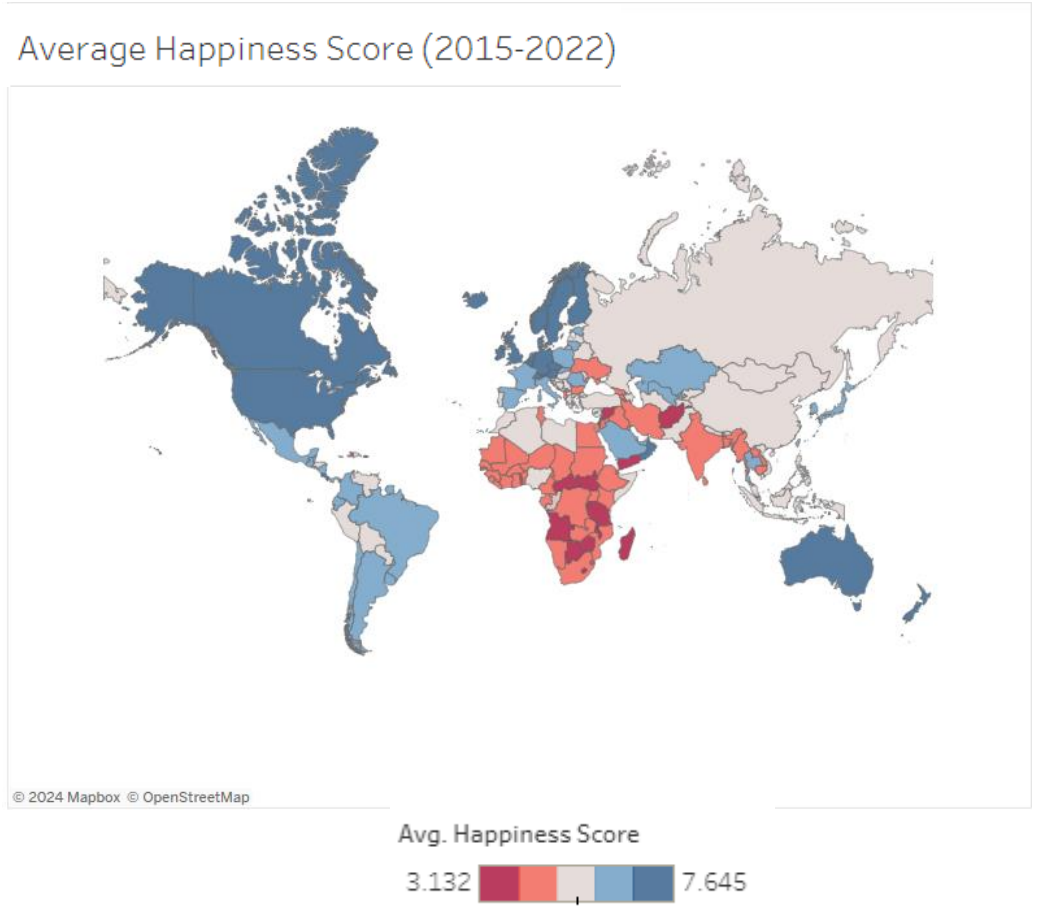
World Happiness Report Project Insights

A brief look at some of the main factors that lead to client loss.

How do each of the factors correlate to one another?



What does Happiness look like across the world?



World Happiness Report Project Conclusion

Key findings for the project, personal takeaways, and deliverables for project.

Project Findings

1. Happiness varies greatly across the world with North America and Europe yielding higher average levels of happiness to other countries.
2. GDP per Capita, Healthy Life expectancy and Social Support have a strong correlation with Happiness Score.
3. The average countries GDP per Capita and Freedom to make life choices has increased overtime, while generosity has decreased.

Takeaways

1. There is so much you can do in Tableau, more than my brain can process.
2. Correlation Matrix's are very useful when comparing many variables to find the correlations with each other.
3. Cluster analysis is also very helpful in finding relationships you wouldn't have been able to find before

Deliverables

Tableau Story Viz

Github Repository

Thank You!

Find me on….

Tableau

GitHub