Nancy Kolaski

Data Analyst

nancykolaski@gmail.com
github.com/Nancy-Kolaski
https://www.linkedin.com/in/nancy-kolaski/
https://public.tableau.com/app/profile/nancy.kolaski/vizzes/
nancykolaski.com

My Background

I have worked the majority of my professional life as an Occupational Therapist trying to make my mark in the world by improving people's lives.

I have obtained a Data Analytics Certification through CareerFoundry. This self-paced online program has prepared me in my transition to a remote work, computer based setting. I am excited to shift my skills into a Data Analyst role, working with raw data as I have always had a passion for research (since research supports everything we do in our lives). Instead of impacting people through direct contact, I want to work from the data perspective to touch people's lives in a different way moving forward.

PORTFOLIO PROJECTS & TOOLS USED

- Gamco Financial Analysis
- Influenza Analysis
- Rockbuster Stealth Analysis
- Instacart Basket Analysis
- Pig E. Bank Analysis
- U.S.A. Real Estate Analysis
- ClimateWins Weather Predictions & Climate Change

GameCo Financial Analysis

Introduction:
GameCo is a video game company that wants to develop new games and establish better marketing strategies. Data analysis looked at different variables that impact sales such as genre/types of games, game platforms, publishers, sales across time (historical to 2016), and sales across geographical regions (North America, Europe, & Japan).

Goal:
Analyze regional and temporal sales trends to make more informed business decisions and develop improved marketing strategies.

Steps and Skills:
- Data Cleaning
- Descriptive Analysis (mean, mode, median)
- Pivot Tables
- Visualizations (line chart, bar chart, stacked bar chart)
- Interpret Results and Summarize findings/insights.

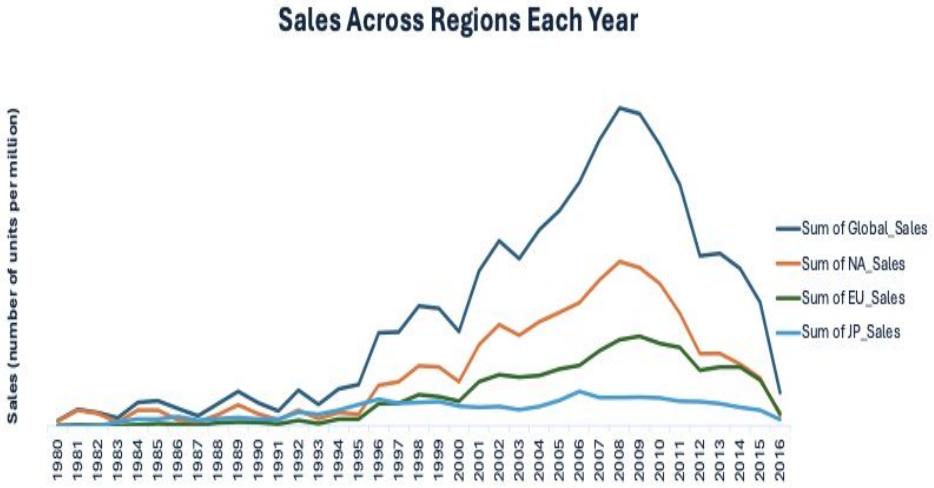
GameCo Data set

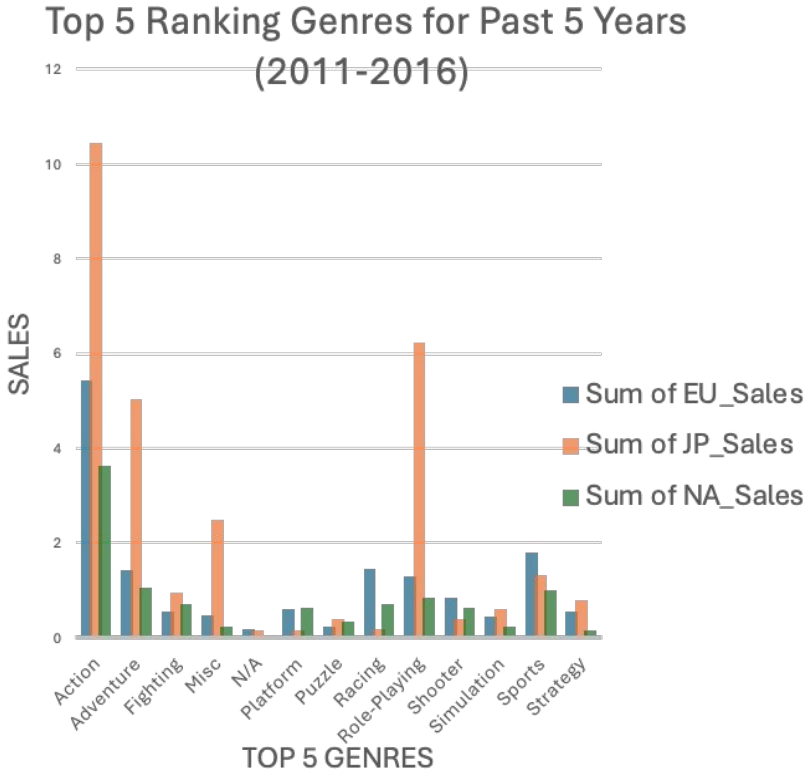
Tools Used:

GameCo Financial Analysis

North America consistently generates the most video game sales in history, even during fluctuations. This line graph demonstrates the positive correlation between North American sales and global sales.

Note that there is a steady and sudden decrease in sales after the peak from 2008-2010. Global sales dropped from a peak of $679.9 million to $70.93 million in 2016.

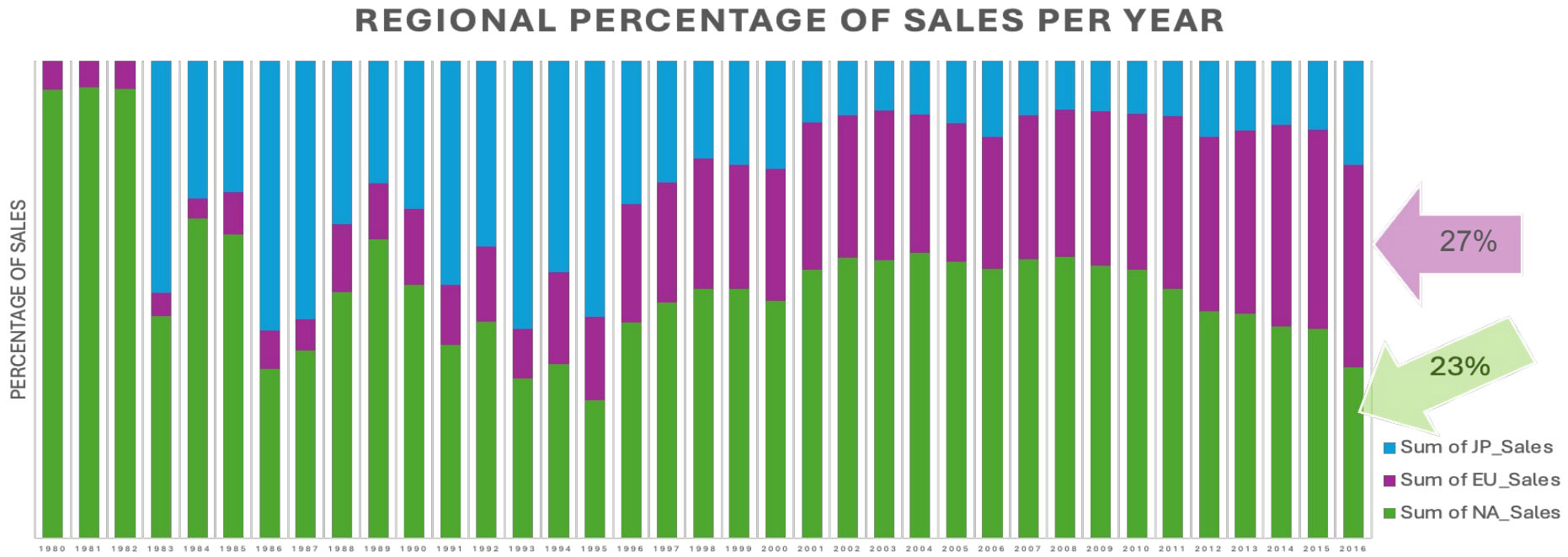




GameCo Financial Analysis

There has been a recent shift in sales trends,. European sales are on the rise.

- Look at the percentage of sales distributed for 2016. Europe is bringing in 27% of sales and North America is bringing in 23% of sales.
  - This is the first year in history that Europe has generated more sales than North America. This is an important insight!



GameCo Financial Analysis

Insights:

- Gameco's sales for video games are made primarily in the North American region, however all sales started dropping from 2010 to 2016.
- Europe has started to bring in more sales percentages than all regions including North America for the first time! It is likely that this trend can continue.
- Majority of video game sales sold globally are **action games.**
- In addition to this, North America generates more sales in **shooting** and **sports games**
- While Japan generates more sales in **role-playing games**
- Nintendo's PlayStation generated the most sales in 2016 for all regions.

Recommendations:

- **PlayStation Game Marketing**
  - Focus marketing strategies moving forward to target the latest release of action-based PlayStation games across all regions. This is important since Nintendo/PS4 is generating the main source of sales globally.
- **Marketing for top 3 genres:**
  - 1) action -> TOP PICK globally
  - 2) shooter -> in North America & Europe
  - 3) sports -> in North America & Europe
  - 4) role playing -> in Japan only
- **Marketing focus in Europe:**
  - More marketing in general should be concentrated in Europe as they have brought in more sales than North America for the first time ever in 2016, and could potentially continue to be a top consumer of GameCo video games in the future.

Influenza Analysis

Introduction:

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 hospitalized. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff. The purpose of this analysis was to determine when to send staff, and how many, to each states. The agency covers all hospitals in each of the 50 United States, and the project will plan for the upcoming influenza season.

Hypothesis:

If a person is 65+ years old, then they are more likely to die from flu.

Goal:

To help medical staffing agencies that provide temporary workers to clinics and hospitals on an as-needed basis. The analysis will help plan for influenza season, a time when additional staff are in high demand. The final results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

Steps and Skills:

- Data Cleaning
- Statistical hypothesis
- Hypothesis testing (t-testing)
- Temporal & Statistical Visualizations in Tableau (bar chart, scatter plot, choropleth map, line graph)
- Interpret Results and Summarize findings/insights.
- Presenting results

Data Sets:

Influenza Deaths
Vaccines
Lab Test Results
Survey of Flu Shot Rates

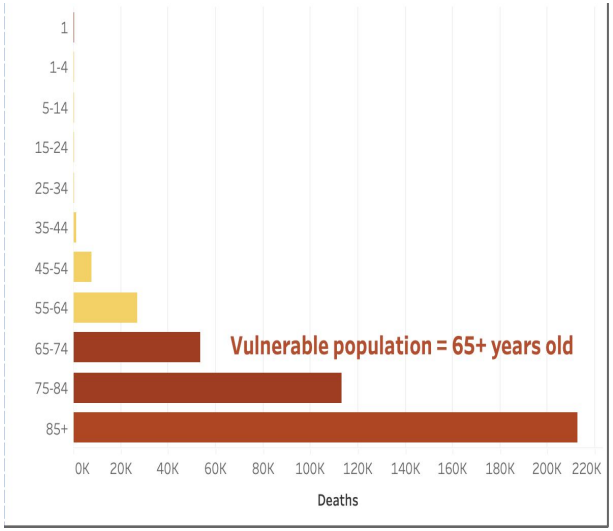
Link for Project Presentation:

Tools Used:

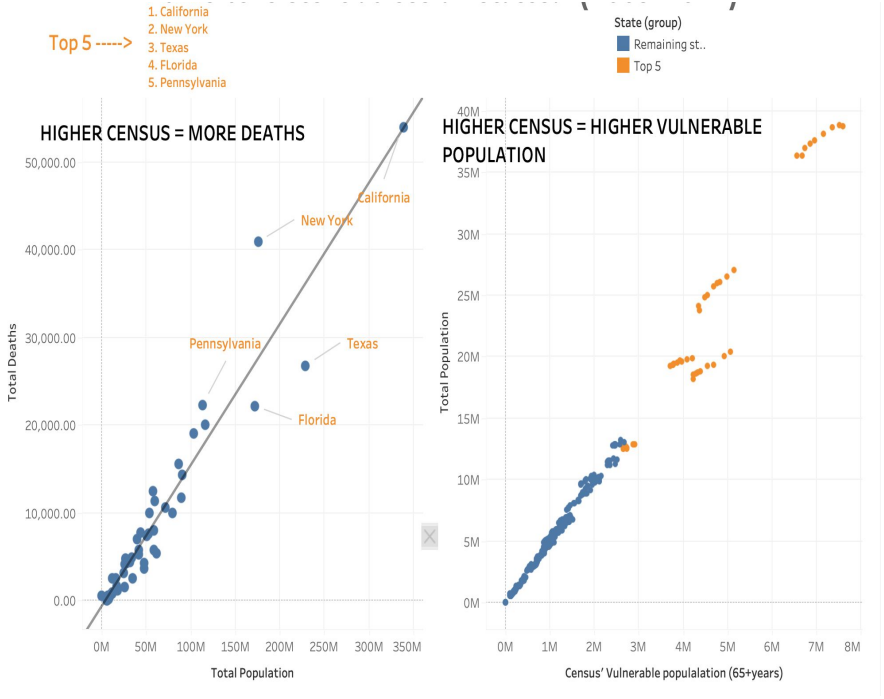
Influenza Analysis

Population/Age Groups

By grouping into 10-year age ranges, it was easier to see that the vulnerable population makes up more deaths.



Top 5 States With Flu-Related Deaths

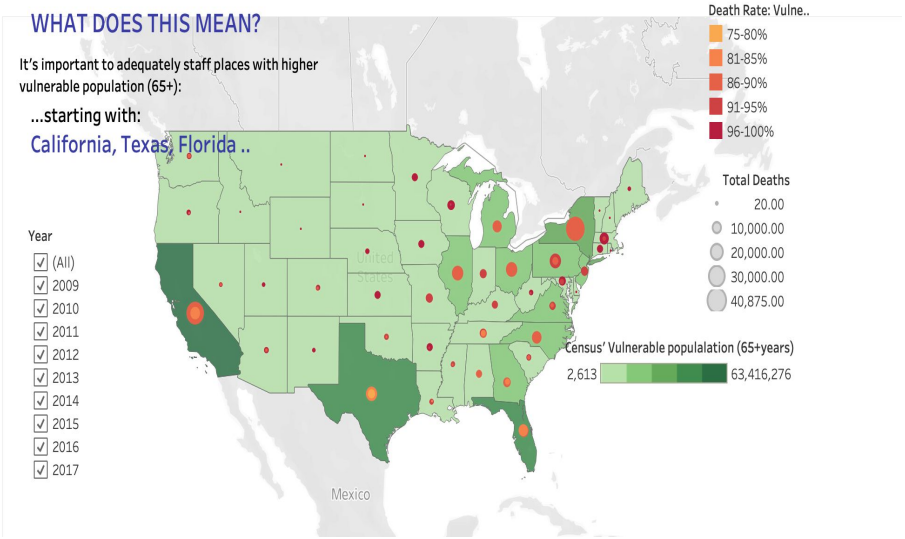


Influenza Analysis

Seasonality

United States census , vulnerable population death rate, and total death count (2009-2017)

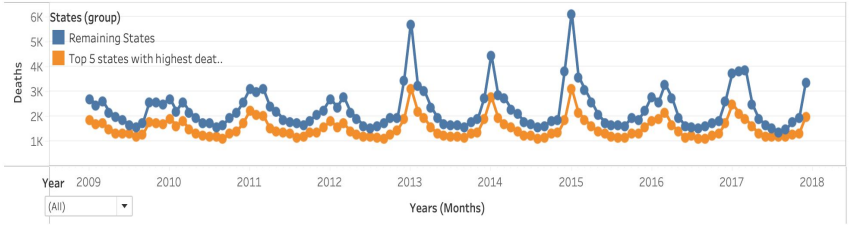
Death rates for vulnerable populations are HIGH for all states! (low of 75-80%!)



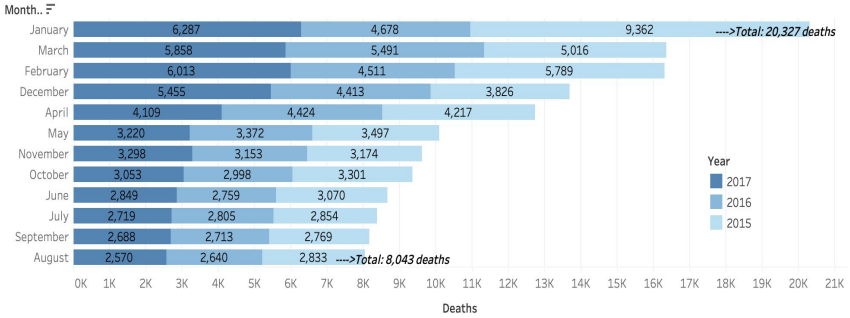
Seasonality of the flu: Watch out in January!

Winter months with highest death count shown at end of the year:

(December through beginning of the next year, January through March)



Monthly view of death count per year (last 3 years)



Influenza Analysis

Insights:

- Vulnerable populations (65+) make up most influenza deaths
  - 5 states with highest vulnerable population count: California, New York, Texas, Florida, Pennsylvania
- States with higher census bring higher death counts
- More deaths by flu occur in the winter months (December - March). January consistently has highest death counts.

Recommendations:

- **Top 5 High Risk States to prioritize staffing needs:**
  - California
  - New York
  - Texas
  - Florida
  - Pennsylvania
- **Seasonality**
  - Concentrated staffing needed during January mostly, along with winter months (Dec-March)
- **Further analysis needed on:**
  - Vaccination rates and their impact on death rates on the vulnerable and non-vulnerable populations (65+ vs below 65 years old)
  - Inclusion of other contributors to our vulnerable population, including those who are under 5 years old, are pregnant, or who have other comorbidities.

Conclusion:

- Based on the analysis of this data, recommendations include prioritizing the top 5 identified states (with the most influenza related deaths & highest population counts) with increased hospital/agency staffing during the months of January,, along with the winter months of December through March.

Rockbuster Stealth Analysis

Introduction:

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

Goal:

Assist the Rockbuster Stealth management board to find data-driven answers to key questions that can be utilized in their 2020 company strategy.

Steps and Skills:

- SQL/Postgres
- Relational Database (Lucidchart)
- Database querying, cleaning, filtering
- Data Descriptions
- Joining, Subqueries, CTEs (Common Table Expressions)
- Visualizations in Tableau (line chart, bar chart, stacked bar chart, choropleth map, bubble chart)
- Interpret Results and Summarize findings/insights.

Data Link

[Tableau Project Presentation]

Tools Used:

Rockbuster Stealth Analysis

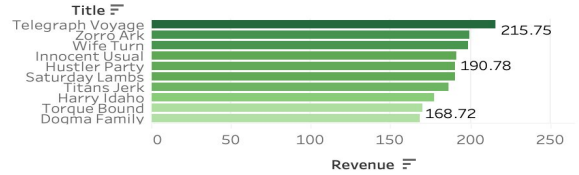
A look into the films' revenue gain

Data Description

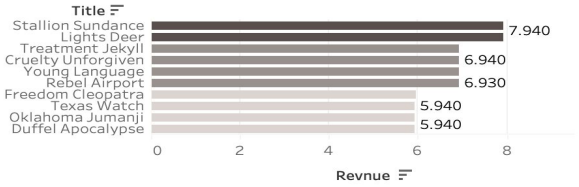
Descriptive Statistics: Film Table (Numeric Data)					
	rent_duration	rent_rate	length	replacement_cost	release_year
Minimum	3	$0.99	46	$9.99	1999
Average	4.985	$2.98	115.272	$19.98	1999
Maximum	7	$4.99	185	$29.99	1999

Film Table, Non-Numeric (Descriptive Statistics)				
	Title	rating	special features	last_update
Mode	Academy Dinosour	PG-13	{Trailers,Commentaries,"Behind the Scenes"}	2024-03-26

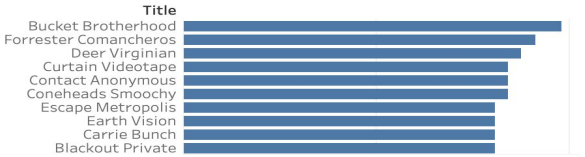
Customer Table, Non-Numeric (Descriptive Statistics)						
	first_name	last_name	address_id	activebool	create_date	active value
mode	Jamie	Abney	5	TRUE	2006-02-14	1



Top 10 movies - Highest Revenue



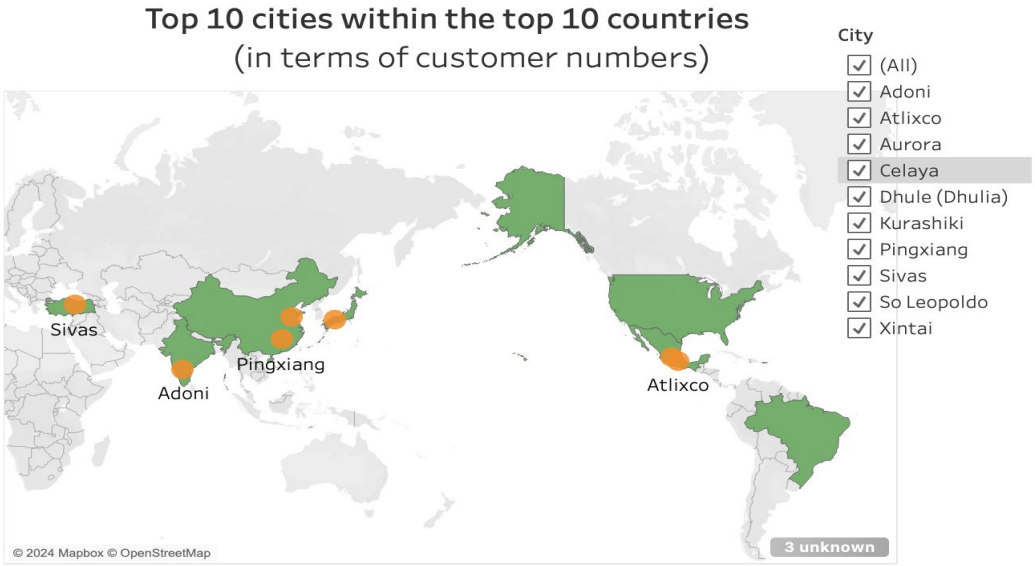
Bottom 10 Movies - Lowest Revenue

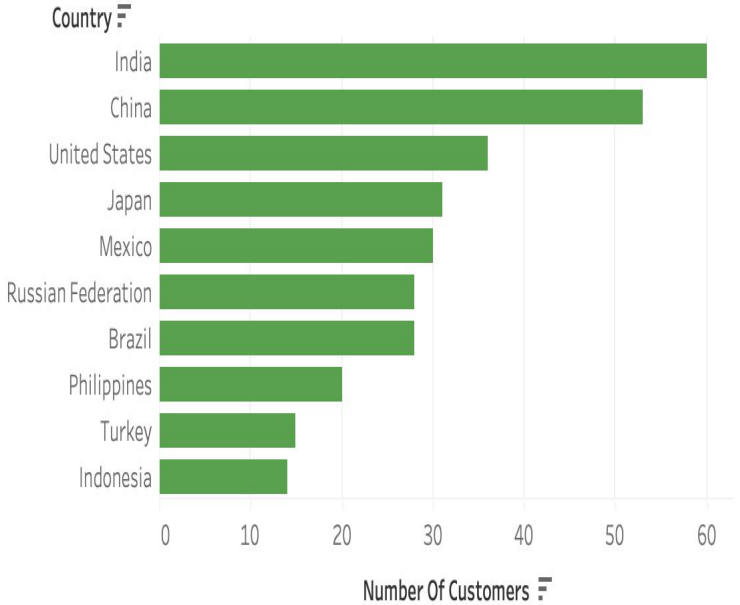


Movies rented the longest:
(no correlations to top 10 movies)

Rockbuster Stealth Analysis

Countries With the Most Customers

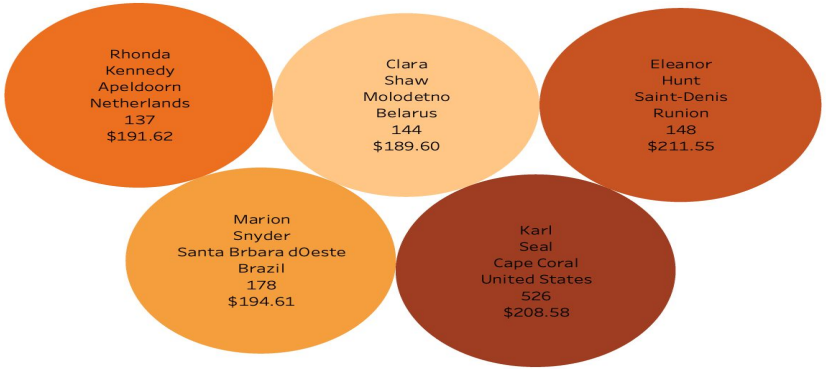




Rockbuster Stealth Analysis

Countries With the Most Revenue

Top 5 customers from top 10 cities, who've paid the highest amounts to Rockbuster



Rockbuster Stealth Analysis

Insights:

- The top 5 countries that bring the most revenue are:
  - India $6,000
  - China $5,200
  - U.S.A #3,600
  - Japan #3,100
  - Brazil & Mexico $2,900
- Geographical distribution is spread worldwide.
  - No trends or centralized locations (interesting insight since most movies are in English)
- Top sales per top 5 customers averages to $200 each
- Average rental duration is around 5 days for all videos
  - No correlation between longer movie rentals and top movies

Recommendations:

- Interview top customers using questionnaires to find out what they like/don't like about Rockbuster's rental service.
  - Since they are invested, they will likely know the process well and provide helpful feedback.
- Use AI to formulate algorithms for movies that are similar to the top 10 movie list in order to advertise more movies similar (and vice versa for bottom list).
  - Collect revenue lists 3 months following to monitor for any increases.
- Focus marketing for Rockbuster in higher populated regions since they produce more revenue.
- Introduce more movies with language dialects native to these top countries: India, China, Japan, Brazil, & Mexico.

Instacart Basket Analysis

Introduction:

Instacart is a well known and widely used online grocery shopping service looking to uncover more information on their sales patterns. This project from the CareerFoundry coursework was geared to introduce python to data analyst students.

Goal:

Assist Instacart stakeholders to understand the variety of customers in their database and their purchasing behaviors in order to develop targeted marketing strategies. The goal for this analysis is to ensure that Instacart targets the right customer profile with the appropriate products. They have specifically requested data regarding busiest times of day/days of the week, price range groupings on their products in order to simplify them, and departments/products that are most popular (most products sold).

Steps and Skills:

- Python/Anaconda
- Data Cleaning and data wrangling
- Deriving variables, merging
- Grouping datasets
- Aggregating data
- Population flows
- Visualizations in Python matplotlib, Seaborn, Scipy (line chart, bar chart, histogram, pie chart, scatterplot, stacked bar chart)
- Interpret Results and Summarize findings/insights.
- Reporting in Excel

instacart

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from www.instacart.com/datasets/grocery-shopping-2017 via Kaggle on [4/20/24].

Note: the customer data was fabricated for the learning purposes of this course.

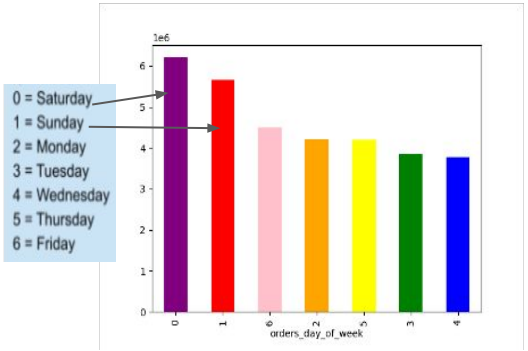
Tools Used:

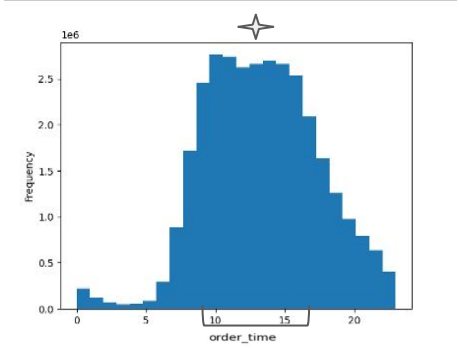
jupyter GitHub

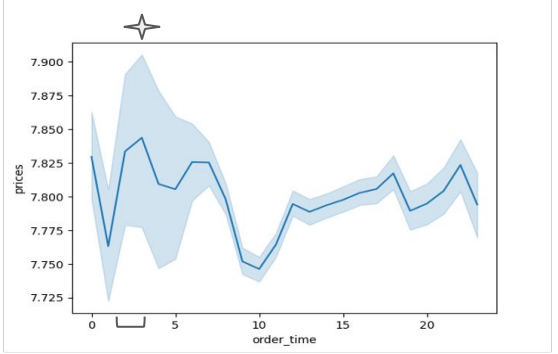
[Github Project Link]

Instacart Basket Analysis

The analysis shows that **Saturday and Sunday** are the busiest days of the week.
Peak hours are between **9am-5pm** with **10am,** the busiest hour of the day. People tend to spend most around **2-3am.**

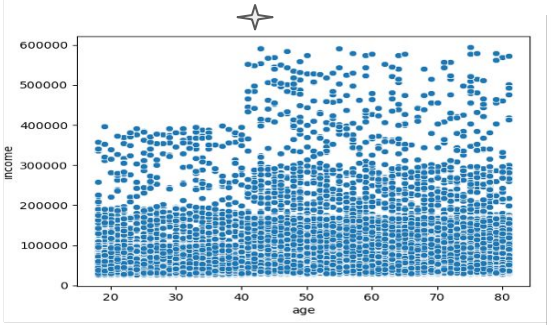


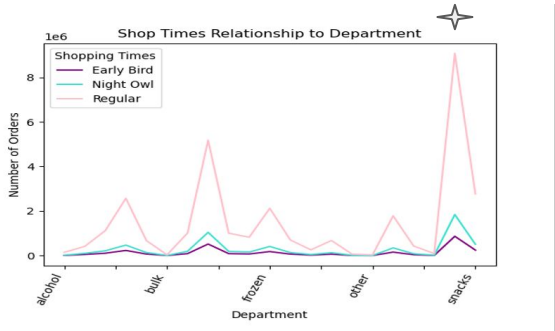




There is a correlation here between age and spending power represented in this scatterplot as there is a **definitive jump in income at age 40**, from a max of $400,000 to $600,000.

The line chart shows that **snack items are the most frequently purchased** for all shoppers, no matter the time. Recommend snack options to always be included in advertisements at **all hours of the day.**



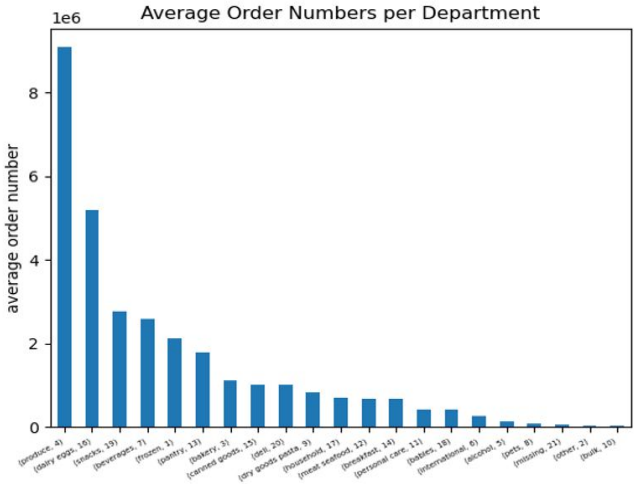


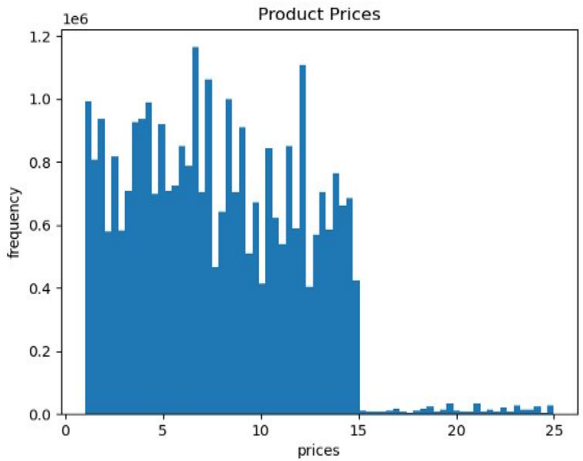
Instacart Basket Analysis

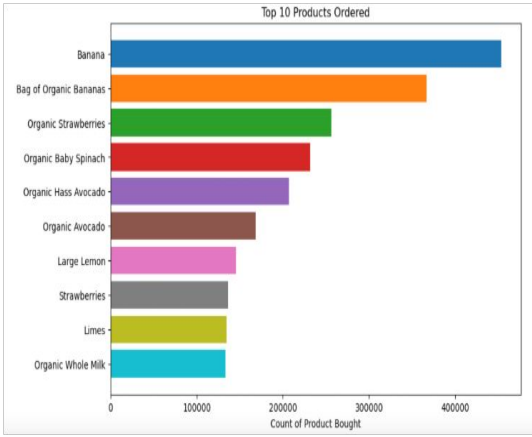
-**Produce, dairy/eggs, snacks**, **beverages**, **frozen**, and **pantry** departments are the **most popular departments** (with least popular departments being international, alcohol, and pets).

- **Most products cost** between **$1-$15**, while a few are priced higher in **$15-$25** range. This information will help to figure out appropriate and simpler price range groupings, as majority of items or lower priced. Aim for $5 increments for simplify price groupings.

-**The top 10 products:** ranked with regular bananas at the top, followed by organic bananas, organic strawberries, organic baby spinach, organic avocados, lemon, regular strawberries, limes, and organic whole milk (interesting to note the popularity of **organic** products).

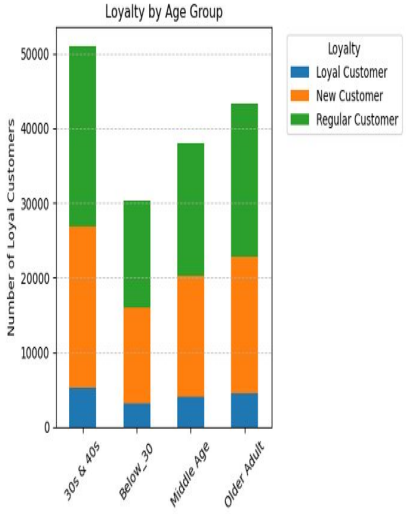


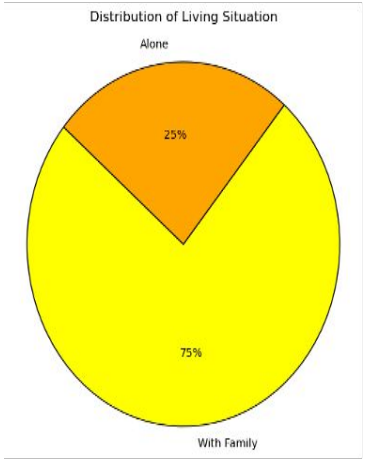


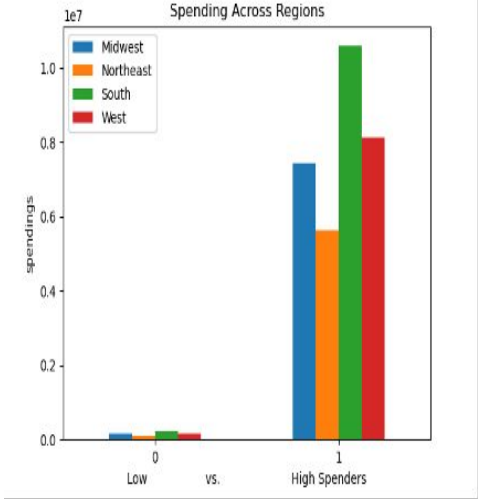


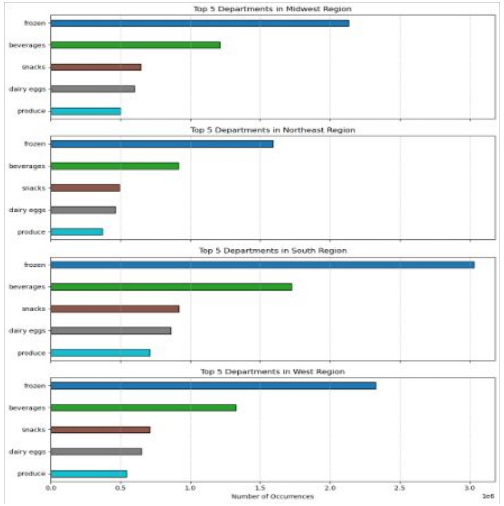
Instacart Basket Analysis

- The **30s/40s** age group tends to rank the highest in all aspects of customers, according to the first bar chart.
  - They make up: 1) the most **regular customers**, 2) majority of **new customers**, & 3) majority of **loyal customers**.
- Most customers **have families** (pie chart).
- The next bar chart shows that **Southern Regions** generate the most of Instacart's spendings. This is followed by Western Regions, then Midwest Regions, with Northeast spending the least. This pattern is the same for both groups of people (high and low spenders).
- These final horizontal bar charts are broken up by regions, showing the **top 5 departments for that region.** All of them show a similar distribution of department preference in the following order: **1) <u>Frozen</u>, 2) Beverages, 3) Snacks, 4) Dairy Eggs, 5) Produce.**









Instacart Basket Analysis

Recommendations:

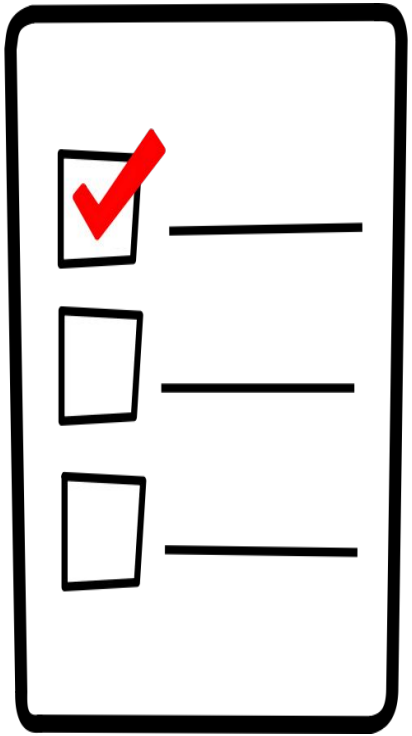
- Increase Ads during mid week (Tuesdays & Wednesdays).
- Schedule Ads between 12pm and 3pm, the busiest hours for placing orders.
- Schedule Ads for higher priced items in the early morning hours between 2-3am.
- Advertise snacks during all hours
- Place eye-catching poster Ads within the popular departments (Produce, dairy/eggs, snacks, beverages, frozen, and pantry).
- Use promotional Ads in popular sections for popular items to increase amount purchased, as well as Ads for lesser populated products to direct viewing towards less popular sections (international, pet, & alcohol).
- Keep $5 price range groupings as the majority of products are between $5-$15.
- Promotional Ads for organic items (placed in highly populated banana section)
- Keep organic foods well stocked and in good shape to maintain the integrity of these sales.
- Recommend targeting 30s-40s, young family groups in advertisements - busy lifestyles (prepared food, quick snacks, food on-the-go).
- Recommend increasing advertisements in southern regions to maintain integrity of their higher sales, as well as promotion of Instacart to the other regions to gain more customers (keeping in mind these variables and target populations in the types of Ads used.

Instacart Basket Analysis

Challenges:

- Creating flagged columns or column derivations did not always go as planned.
    - **Solution**: Since I carefully kept track of each step of the wrangling process, I could trace my footsteps and was able to restart the steps prior to my error. This taught me to record each manipulation of original datasets, no matter how small or insignificant it may have seemed. As a result, I have learned to be diligent with organization and record keeping throughout the entire wrangling process.
- I experienced difficulty in creating certain visualizations with python using this large dataset, particularly pie charts.
    - **Solution:** I transferred categorized data into a csv file and then uploaded into Tableau to create a pie chart there instead.

★ This project was extremely time-consuming for me, taking much longer than the CareerFoundry time frame expectation for task completion.
    - Because of the amount of time invested, I found that this was a great learning tool for me to fully understand Python and Jupyter Notebook. I grew a great appreciation for Python with hands on work throughout the tasks involved. As my knowledge grew, I found it fun to dissect and manipulate this large dataset to find the key insights that I was targeting. I spent time deep diving into this coursework since I knew that this process was going to be essential for my future work as a data analyst.



Pig E. Bank Analysis

Introduction:

Pig E. Bank is a well known global bank looking for analytical support for it's anti-money-laundering compliance department.

Goal:

Help Pig E. Bank assess client risk and transaction risk, as well as reporting on metrics. Build and optimize models that assist the bank in running their compliance program more efficiently.

Steps and Skills:

- Big Data
- Time-series analysis & Time-series forecasting
- Data bias impact on ethics
- Data mining
- GitHub

Pig E. Bank Data Set

GitHub Repository Link:

https://github.com/Nancy-Kolaski

Tools Used:

Pig E. Bank Analysis

Comparing Current & Former Customers

COUNTRIES

FORMER		CURRENT		COMBINED	
COUNTRIES		COUNTRIES		COUNTRIES	
France	37.75%	France	51.15%	France	48.38%
Germany	36.76%	Germany	23.16%	Germany	25.96%
Spain	25.49%	Spain	25.70%	Spain	25.66%
Grand Total	100.00%	Grand Total	100.00%	Grand Total	100.00%

★ Germany has a higher amount who leave (36% vs 24% who stay).
★ France makes up most of the clientele for Pig E. Bank, (less people leave).
★ Spain is split pretty equally between those who leave and those who stay (25-26%).

AGE RANGE

AGE RANGE		AGE RANGE		AGE RANGE	
18-24 years	0.99%	18-24 years	4.33%	18-24 years	3.64%
25-34 years	11.82%	25-34 years	35.75%	25-34 years	30.84%
35-44 years	35.47%	35-44 years	43.51%	35-44 years	41.86%
45-54 years	33.50%	45-54 years	9.67%	45-54 years	14.56%
55-64 years	15.76%	55-64 years	3.94%	55-64 years	6.37%
65 years and up	2.46%	65 years and up	2.80%	65 years and up	2.73%
Grand Total	100.00%	Grand Total	100.00%	Grand Total	100.00%

★ Most customers are between the ages of 25-44 years.
★ Those who leave the bank tend to be a little older (35-54 years).

GENDER

GENDER		GENDER		Row Labels	
Female	59.31%	Female	341	Female	46.71%
Male	40.69%	Male	444	Male	53.29%
Grand Total	100.00%	Grand Total	785	Grand Total	100.00%

★ More females leave (59%) .

TENURE

Tenure		TENURE		TENURE	
0	2.94%	0	4.07%	0	3.84%
1	14.71%	1	9.29%	1	10.40%
2	13.24%	2	10.56%	2	11.11%
3	9.80%	3	10.31%	3	10.20%
4	8.33%	4	8.78%	4	8.69%
5	9.80%	5	10.05%	5	10.00%
6	10.29%	6	9.29%	6	9.49%
7	7.35%	7	9.92%	7	9.39%
8	9.80%	8	11.20%	8	10.91%
9	10.29%	9	11.07%	9	10.91%
10	3.43%	10	5.47%	10	5.05%
Grand Total	100.00%	Grand Total	100.00%	Grand Total	100.00%

★ Most people left with 1 or 2 years of Tenure (28%.)

NUMBER OF PRODUCTS

NumOfProducts		NumOfProducts		NumOfProduts	
1	69.61%	1	46.82%	1	33.84%
2	15.69%	2	52.54%	2	59.06%
3	13.73%	3	0.64%	3	6.57%
4	0.98%	Grand Total	100.00%	4	0.53%
Grand Total	100.00%			Grand Total	100.00%

★ Majority of the clients who left only had one product. Of the current clients, 53% have 2 products.

IS ACTIVE MEMBER

(1=YES, 0=NO)

Is Active Member		Is Active Member		Is Active Member	
0	70.10%	0	43.77%	0	49.19%
1	29.90%	1	56.23%	1	50.81%
Grand Total	100.00%	Grand Total	100.00%	Grand Total	100.00%

★ 70% of those who left were not active members.
★ 56% of current clients are active members.
★ Combined groups (current & former), are equally divided for active and not active.
  - Perhaps, more member engagement/involvement could influence them to stay.

FORMER VS. CURRENT CUSTOMERS

Exited from bank?		Exited from bank?		Former vs. current	
1	204	0	786	0	79.39%
Grand Total	204	Grand Total	786	1	20.61%
				Grand Total	100.00%

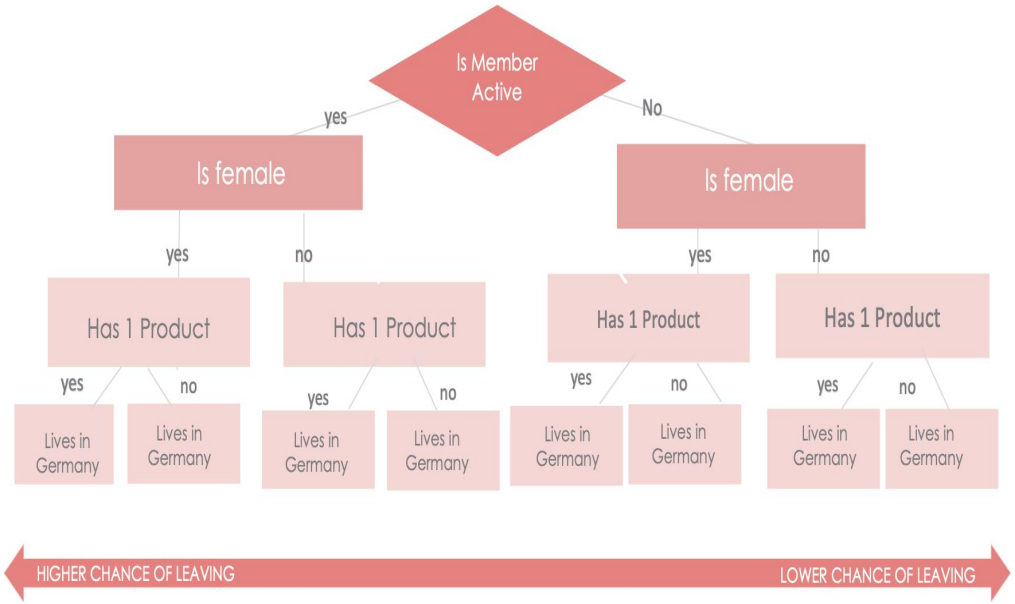
★ 79% (a good majority) have not left the bank. 21% of clients left.

Pig E. Bank Analysis

DESCRIPTIVE STATISTICS FOR CURRENT VS FORMER CLIENTS (AND COMBINED				
	Mean	Max	Min	Count
Age of Former	45	69	22	205
Age of Current	38	82	18	787
Age of Combined Total	39	82	18	992
Balance of Former	$90,239.22	$213,146.20	$0.00	205
Balance of Current	$74,807.56	$197,041.80	$0.00	787
Balance of Combined	$78,003.00	$213,146.20	$0.00	992
Tenure of Former	5	10	0	205
Tenure of Current	5	10	1	787
Tenure of Combined	5	10	1	992
Number of Products for Former	1	4	1	205
Number of Products for Current	2	3	1	787
Number of Products for Combined	2	4	1	992
Estimated Salary of Former	$97,155.20	$199,725.39	$417.41	205
Estimated Salary of Current	$98,984.61	$199,661.50	$371.05	787
Estimated Salary of Combined	$98,574.54	$199,725.39	$371.05	992
Credit Score of Former	637	850	376	205
Credit Score of Current	652	850	411	787
Credit Score of Combined	649	850	376	992

Decision Tree

21% of all Pig E. Bank's clients leave. Who is most likely to be in this group?



Pig E. Bank Analysis

Insights:

- 21% of Pig E Bank's client's have left.
- Females make up the 59% of those who leave.
- 70% of clients who left had only one product.
- 70% of those who left were not active members.
- Most people left with 1 or 2 years of Tenure, making up 28% of those who left.
- France makes up most of the clientele for Pig E. Bank, with smaller numbers that leave in comparison to their presence in the company. Germany tends to have the most departures as they average a higher percent of those who leave (36%), than those who are current (24%).
- The Majority of Pig E Bank's current customers fall between the ages of 25-44 years, however those who leave the bank tend to be a little older at 35-54 years. These age groups make up 69% of all who leave.

Recommendations:

- Target inactive customers.
- Keep customers informed through regular emails, newsletters, and company updates.
- Offer incentives to customers in Germany to see if departures from Pig E. Bank can be avoided. If this is successful, continue this in France.
- Conduct further research with geographical focus to better understand why more customers leave in Germany and/or why they stay in France.
- Create leveled flags for customers who fit into the "departure" categories, in order to offer more support and/or incentives to stay. The more categories they belong in, the bigger they are flagged.
    - Those with only one product, who are not active members, who have only 1-2 years of Tenure, are female, and are from Germany.

U.S.A. Real Estate Analysis

Introduction:

This case study looked at real estate across the U.S.A., specifically across the different regions and investigating whether certain variables influenced the market.

Goal:

Analyze the United States real estate market to see what factors or variables influence sales the most.

Hypothesis:

As house size increases, price increases.

Steps and Skills:

- Data Cleaning
- Exploratory Analysis
- Linear Regression
- Scatterplots
- Clusterplot Analysis
- Pairplots

Data Set

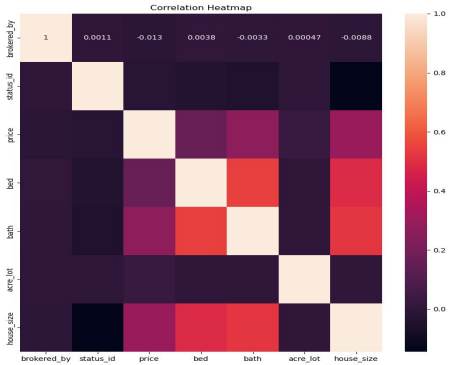
Tools Used:

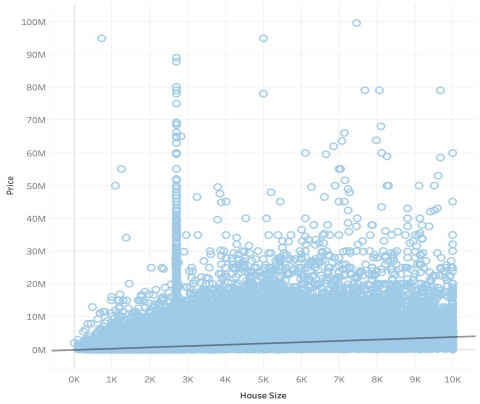
Tableau Presentation Link

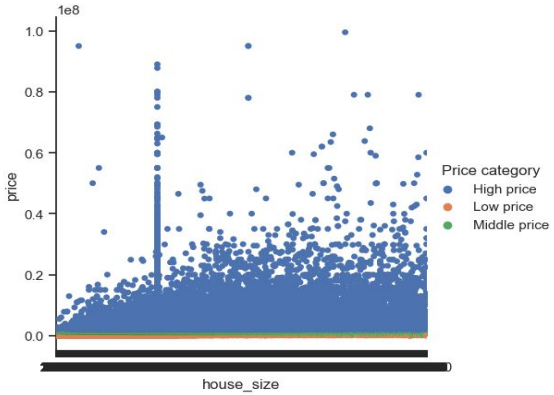
U.S.A. Real Estate Analysis

Exploratory Analysis:

By using this correlation heatmap (below), I was able to determine the strongest positive correlations. Since bed, bath, and house size all seem to be obvious in their correlation, it was determined to further explore the relationship between house size and price (noted by the lighter purple color).







***HYPOTHESIS:**

As the house size increases, so does the price.

To test this hypothesis, I conducted a linear regression.
There are many points that fall outside of the regression line. This isn't enough to draw a significant conclusion. I will try another approach -> Clusters.

In order to prove the hypothesis besides only using linear regression, I conducted a cluster analysis. This categorical plot groups data into 'clusters' in order to compare each group to uncover patterns.

-Cluster 0 represents low priced homes.
-Cluster 1 represents medium to high priced homes.
-Cluster 2 represents very high priced homes.

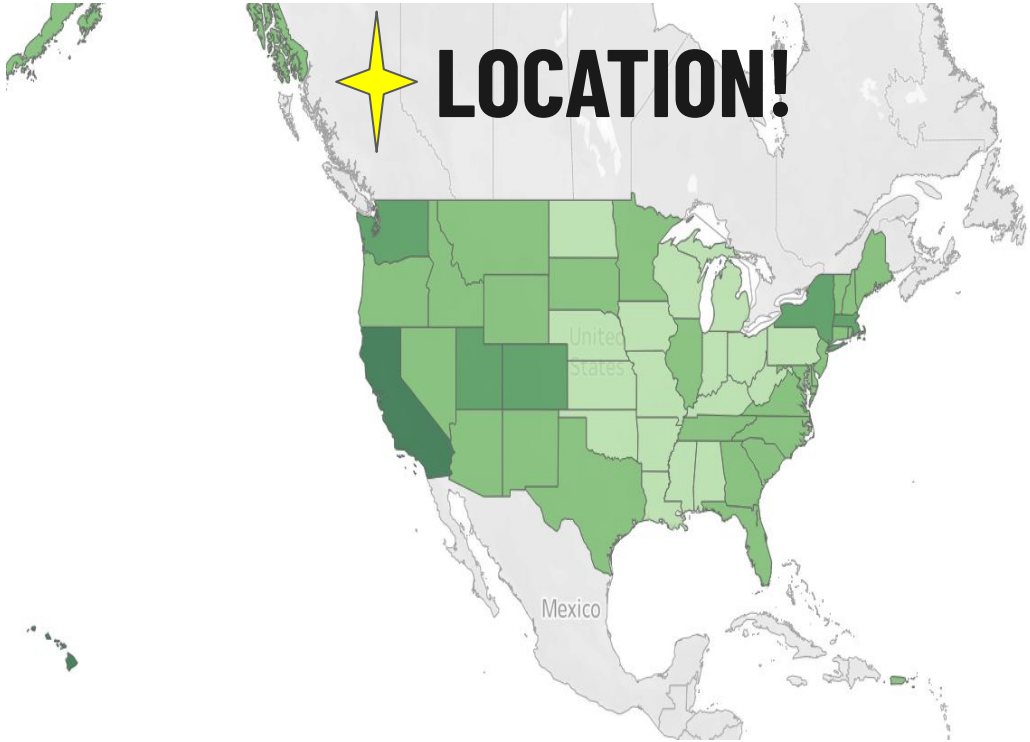
This plot shows that cluster 1 (med-to-high priced homes) have outliers for very high priced homes while the other two clusters stay within a certain range.

All three clusters share a large amount of homes at around 2700 square feet (as noted previously), meaning this is a common house size across all price ranges.

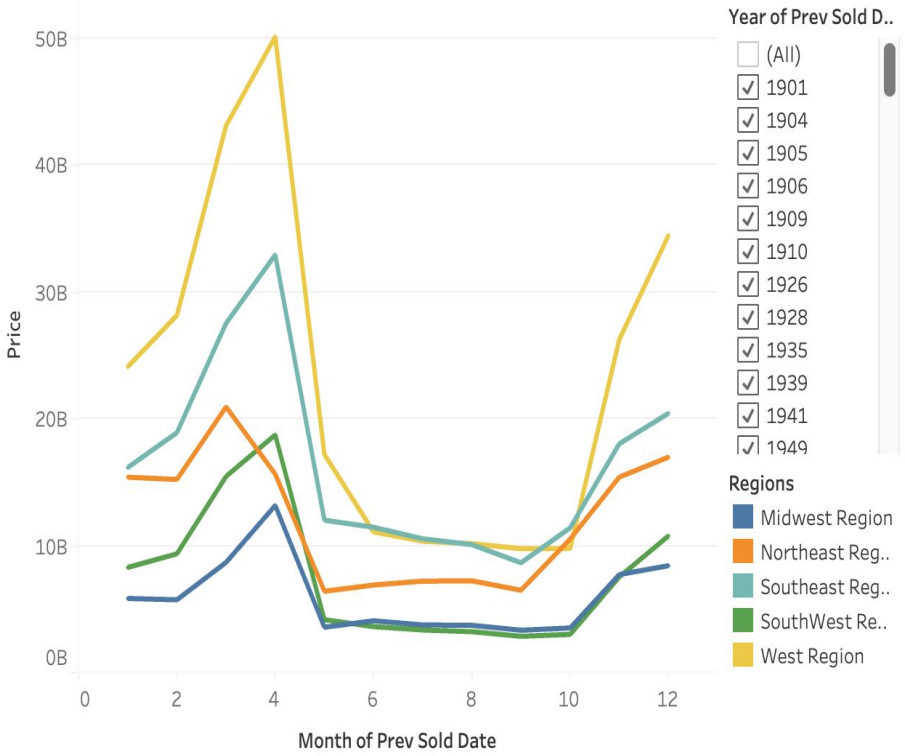
U.S.A. Real Estate Analysis

Geography plays an important role on the price. Western and Northeast Regions, indicated by the darkest green color, demonstrate the highest revenue generated from real estate within the U.S.A. These top states include **Hawaii, California, Montana, Colorado, Utah, New York, & Massachesetts.**

Lighter green states (midwest) show the least revenue generating states.



U.S.A. Real Estate Analysis



Seasonality (1901-2023)

Stays the same across all regions

Most homes are sold in March and April. This is the same across all regions and across time. Revenue is represented highest in total for West Regions, then Southeast, Northeast, Southwest, and Midwest marked as producing the lowest revenue. This is parallel to cost of living parameters.

Trends remain the same.

April is the hot month for selling/buying.

Note: 2024 was left out as homes without a previous sold date were marked for today's date (raising the month to current month of august 2024). By removing this year, we have eliminated the risk of tainting data to show more sales where they have not occurred.

U.S.A. Real Estate Analysis

Insights:

- PRICES and LOCATION:
  - Highest real estate prices exist within the west and northeast regions:
  - CA, HI, MT, CO, VT, NY, MA
  - Lowest real estate prices are in the midwest
- PRICES and HOUSE SIZE:
  - Strongest correlations exist between house size and price
  - The bigger the house size, the higher the price is, generally.
  - 2700 square feet tends to be a house size that is standard across ALL PRICES!
- SEASONALITY:
  - Most homes are sold in April (and March).

Recommendations:

- There was a large amount of data where previous sold dates were blank. It was assumed that these were newer builds without a previously sold date. These homes were left out of the seasonality. Next steps would be to clarify this.
- Acre lot sizes and house sizes had a HUGE variance in size, along with price range. It would have been insightful to have more details with these homes/listings to understand the context of these ranges (particularly acre lots and house sizes marked as '0'). This does not make sense as it is impossible to have a house or lot size equal to 0. Next steps would be to clarify the number markers, determine if these were missing values filled in, and eliminate or impute values.

ClimateWins Weather Predictions & Climate Change

Introduction:

ClimateWins is a fictional European nonprofit organization that is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world. It's concerned with extreme weather events, especially in the past 10-20 years. Through use of machine learning, it wants to see if weather conditions can be predicted by looking historically at the temperature highs and lows, and exploring whether conditions can be predicted to a specific given day and can prevent dangerous weather conditions.

Goal:

Utilize machine learning (both supervised and unsupervised) algorithms to make weather predictions and understand the consequences of climate change around Europe and, potentially the world.

Data Set

Hypothesis:

- ClimateWins can help predict climate change around Europe (and potentially, around the world).
- The weather climate across Europe will gradually increase over time.
- Supervised and unsupervised deep learning algorithms are optimal tools in predictive analysis necessary for weather forecasting.

Steps and Skills:

- Data Cleaning
- Gradient Descent Optimization
- Pairplots

- **Supervised Learning Algorithms**
  - K-Nearest Neighbor (KNN)
  - Artificial Neural Network (ANN)
  - Decision Tree

- **Unsupervised Learning Algorithms:**
  - Random Forest
  - K-Means Clustering
  - Dendrograms
  - Dimensionality Reduction
  - Convolution Neural Network (CNN)
  - Recurrent Neural Networks (RNN)
  - Generative Adversarial Networks (GANs)

Tools Used:

Github Project Link

ClimateWins Weather Predictions & Climate Change

***Optimization* lowers the risk of error and improves the accuracy of a model, often used to determine which algorithms to use.**
It helps understand valleys and peaks of the local/global landscape of the data.

- **Gradient Descent** (used in both linear and nonlinear data) was used in this study to determine the local minimums and maximums of the data points.
  - Three iterations performed, adjusting step lengths (alpha) in order to get a result as near to 0 as possible

Is Climate Increasing?

- Belgrade has freezing minimum temps getting colder over past 20 years. It has warmed up by about 5 degrees over past 60 years, when looking at the mean per year. The max mean raised 1 degree higher than 60 years ago).
- In general, all means, mins, and max temperatures have increased, with the exception of Valencia (where data is likely skewed due to permanency of 10.7 report).

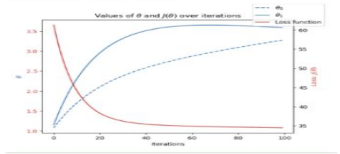
This chart shows data for approximately a 60 years span of temperatures in Madrid, Valentia, and Belgrade in the years: 1980, 2000, & 2018

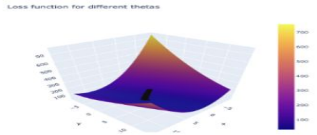
Weather Station	Year	Theta0		Theta1		Iterations	Step size	Mean Temp	Max Mean	Min Mean
		Start	End	Start	End					
MADRID	1980	1	0	1	0	100	.01	14.19 (57.5F)	30.4 (87F)	-.2 (31F)
MADRID	2000	1	-1	1	0	100	.01	15 (59F)	29.4 9 (84F)	.6 (33F)
MADRID	2018	1	0	1	0	100	.01	15.57 (60F)	32.9 (91F)	1.6 (35F)
VALENTIA	1980	1	0	1	0	100	.01	10.35 (50.6F)	17.9 (64F)	1.5 (34F)
VALENTIA	2000	1	0	1	0	100	.02	10.76 (51F)	19.5 (67F)	1.4 (34.5F)
VALENTIA	2018	1	0	1	0	100	.01	10.7 (51F)	10.7 (51F)	10.7 (51F)
BELGRADE	1980	1	0	1	0	100	.01	10.77 (51F)	27.4 (81F)	-9.3 (15F)
BELGRADE	2000	1	1	1	1	100	.03	14.19 (57F)	32.8 (91)	-9.9 (14F)
BELGRADE	2018	1	.5	1	1	100	.02	14.94 (57.5F)	28.4 (83F)	-5.8 (21.5F)

Belgrade:

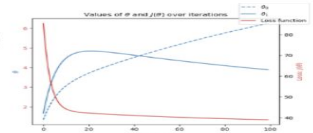
3 iterations showing *loss of function* and *loss profile*

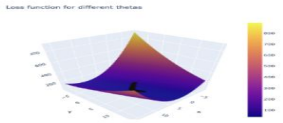
BELGRADE 1980



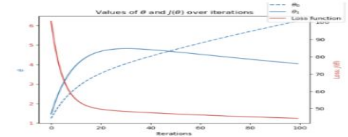


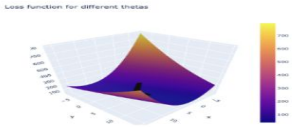
BELGRADE 2000





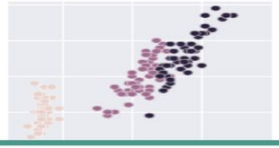
BELGRADE 2018

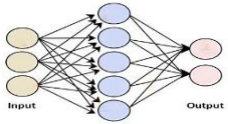


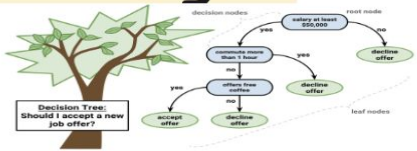


ClimateWins Weather Predictions & Climate Change

Supervised Machine Learning



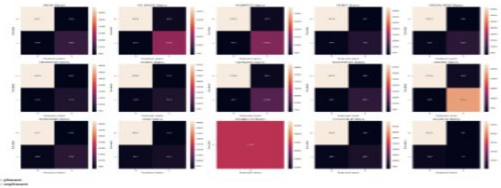




K-Nearest Neighbor (KNN):

classifies data on its proximity to its neighbors

Test Accuracy Scores = 88.46%



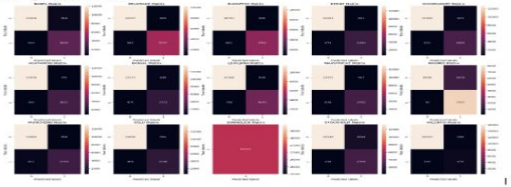
Weather Station	Accurate Predications 0 (nonplacement)	Accurate Predications 1 (placement)	False Positive	False Negative	Accuracy:
Basel	3917	961	421	439	4878/5738 – 85%
Belgrade	3252	1544	524	418	4796/5738 – 84%
Budapest	3424	1462	476	376	4886/5738 – 85%
Debilt	4320	723	317	378	5043/5738 – 88%
Dusseldorf	4164	810	343	421	4974/5738 – 87%
Heathrow	4138	744	432	424	4882/5738 – 85%
Kassel	4563	614	252	309	5177/5738 – 90%
Ljubljana	3740	1180	455	363	4920/5738 – 86%
Maastricht	4253	824	309	352	5077/5738 – 88%
Madrid	2750	2261	418	309	5011/5738 – 87%
Munchenb	4237	792	309	400	5029/5738 – 88%
Oslo	4637	512	242	347	5149/5738 – 90%
Sonnblick	5738	0	0	0	5738/5738 – 100%
Stockholm	4483	607	283	365	5090/5738 – 89%
Valentia	5404	74	58	202	5478/5738 – 95%

Artificial Neural Network (ANN):

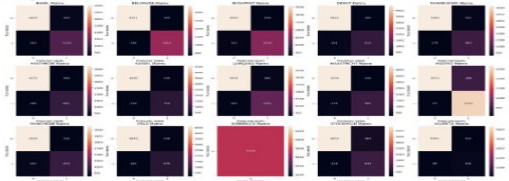
Replicates the human brain, consisting of input and output layers, along with hidden layers by adjusting weights to obtain outcomes

Train accuracy score = 52%
Test Accuracy Score = 49%

Training: 52% accuracy



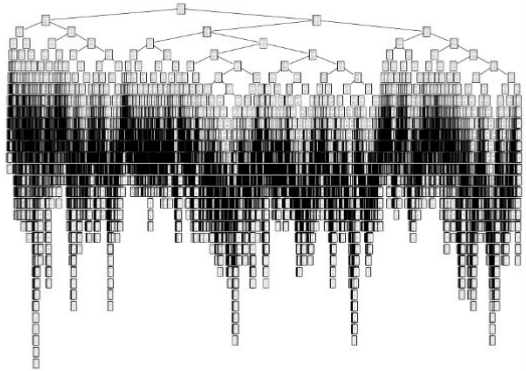
Testing: 49% accuracy



Decision Tree

Decision trees model these kinds of questions for many objects at once. They have roots, branches, and leaves. The *root* is the first question asked, with yes/no answer. Leading to another question/ *branch*. The answer is the *leaf*/stopping point.

Train accuracy score = 46%,
Test Accuracy Score = 47%



ClimateWins Weather Predictions & Climate Change

Algorithm Overview: Machine Learning
(Supervised/Unsupervised)

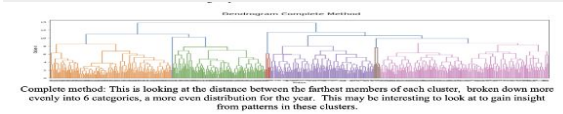
Random Forest:
(Clustering with Dendrograms)

Identify extreme weather conditions:

Compare all stations with Complete method (reduced data) to find trends, then narrow down specific locations to find the extreme weather condition patterns using Single method.

Combines multiple decision trees to make the most accurate predictions. Each decision tree trains on a random sample of the total data, with final prediction made by averaging the predictions of all trees.

Dendrogram Single Method can find outliers, interpreted as extreme weather conditions *when looking at specific weather stations.* It is not as useful when comparing all weather stations



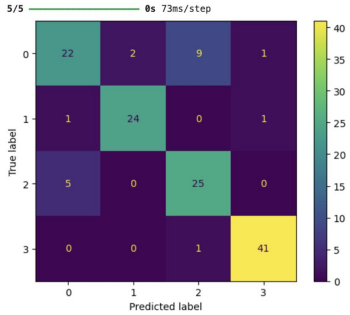
Dendrogram Complete Method gives clear, distinct clusters with less noise to find key patterns in the data (all weather stations per year).

Deep Learning with CNN & RNN:
(Convolutional Neural Network & Recurrent Neural Network)

Determine if unusual weather patterns are increasing:

Use **CNNs** and **GAN**s to analyze spatial data to find trends across these regions.

CNNs handle **images & numerical** data as they were inspired by visual cortex processes of the brain. RNNs handle temporal data such as **text, handwriting, and speech**. RNNs also use LSTM (long short-term memory), imitating a brain that forgets unimportant data – updating the significant data as needed).



CNN model Confusion Matrix
showing 4 classes: cloudy, rain, sunshine and sunrise
91% Accuracy

Incorrect prediction:
Class cloudy, predicted shine

GANs:
(Generative Adversarial Network)

Determine safest places to live in the future:

Use **Random Forest / Decision Trees (for 2010s)** to use predicted weather conditions to determine regional safety. Use the identified features within top locations to gauge future weather predictions.

Use two adversarial neural networks working against each other – generator (creates data) & discriminator (samples of real & artificial, discriminating which is real). Can create artificial data (text & images).

ClimateWins Weather Predictions & Climate Change

Insights:

- Supervised and Unsupervised Machine Learning models can make accurate weather predictions.
  - This study has shown slight temperature increase over time, and has potential to generalize to the rest of the world.
- **KNN (**K-Nearest Neighbor) **model was the best choice** for unsupervised method (88% accuracy, much higher when compared with the other two).
  - **ANN (Artificial Neural Network) has potential** to produce a more complete depiction of the data, presuming adjustments are made when eliciting this model as long as there is a good understanding of the model and how to manipulate it for best outcomes.
  - **Decision Tree** was not useful as the data was too deep and complex for a meaningful insight, nor was **RNN** due to its low accuracy rate and lack of insightful information. These algorithm may run better on sections of data.
- **Random Forest Algorithms** are highly accurate and provide useful breakdown of data to pinpoint weather data features and top locations.
  - Dendrograms are useful in finding outliers or extreme weather events (single) and for finding season trends year round (complete).

Recommendations:

- **Prune Decision Tree** data to reach a better conclusion and avoid overfitting.
- **Run more iterations on ANN model** to find more options for training the model to find more optimal variables.
  - Rule out if Sonnblick's 100% accuracy was due to overfitting or data error.
- Make adjustments within the weather data.
  - **Further clean/check for error**s: as in the case of Valencia showing no variability in min/max/mean temps for the year chosen in this analysis.
  - Incorporate more complete data: **include more descriptions features that encompass 'pleasant' vs. 'unpleasant'.**
  - **Separate locations into hot/cold regions** to eliminate cultural bias, particularly in perception of what is pleasant vs. unpleasant.
- **Combine supervised and unsupervised algorithms** for best results.
  - Use **Random Forest to narrow down** the data features (either by location or year). **Create subsets** of data on these important variables. Then, **run those through a CNN** to produce higher accuracy. Then, use final results to **plug into a GAN** to generate more realistic artificial results.
- **Increase sample size to other parts of the world** besides Europe and run the algorithms again to compare and see if these methods can be used as a generalization for world data.

Thank you for your time and consideration.

- nancykolaski@gmail.com
- github.com/Nancy-Kolaski
- https://www.linkedin.com/in/nancy-kolaski/
- https://public.tableau.com/app/profile/nancy.kolaski/vizzes/
- nancykolaski.com