PORTFOLIO 2024

Meghan McGrath

Data Analysis Portfolio

Projects

01	GameCo	Video game market analysis to advise strategic planning for future business initiatives
02	Influenza	Influenza data analysis to inform staffing needs to upcoming flu seasons
03	Rockbuster	Movie rental service analysis to guide new online video strategy
04	Instacart	Customer spending habit analysis to inform business strategies
05	Pig E. Bank	Customer retention analysis for global banking provider
06	Airbnb	Geographical & machine learning analysis for NYC rentals
07	ClimateWins	Machine learning analysis to predict climate change conditions




01 GameCo

GameCo, a new video game company, wants to use data to inform the development of new games.

Goal: Perform a descriptive analysis of a video game data set to foster a better understanding of how GameCo's new games might fare in the market.

01 Overview

Key Business Questions

1. Are certain types of games more popular than others?
2. What other publishers will likely be the main competitors in certain markets?
3. Have any games decreased or increased in popularity over time?
4. How have their sales figures varied between geographic regions over time?

Data & Tools

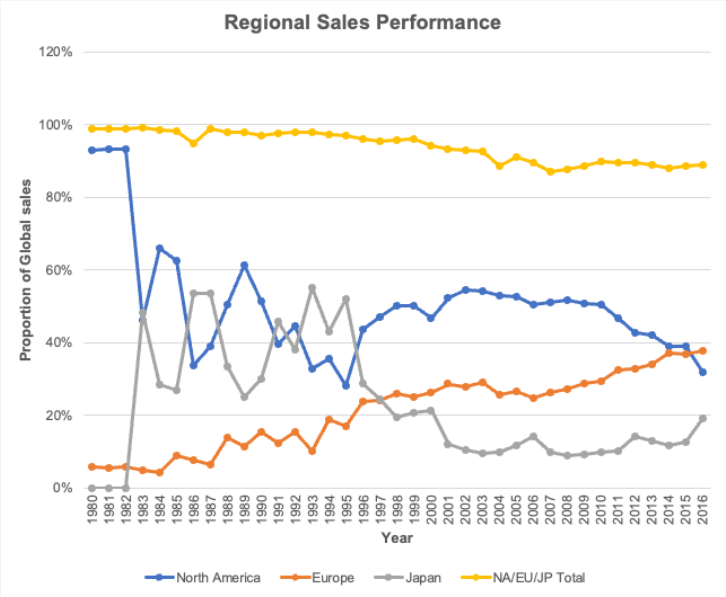
Data:
- Video Game Sales
- Source: VGChartz Data Set

Tools/Skills:
- Excel
- Cleaning data
- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing data insights in Excel
- Storytelling with data

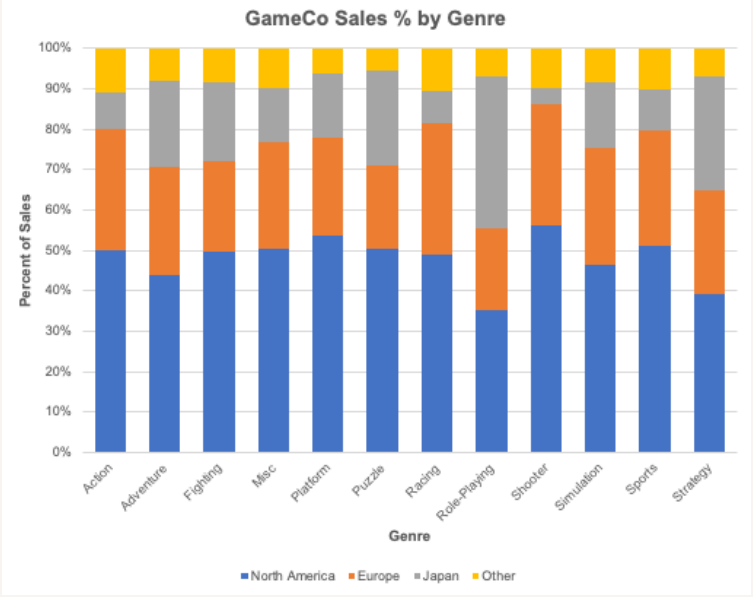
Challenges & Limitations

1. **Data Completeness and Accuracy:**
   - Includes data from 1980-2016 - may lack recent trends/newer game releases that could impact current market analysis.
   - Sales figures include units sold, not revenue - difficult to fully assess financial performance or profitability
2. **Genre Classification:** Genres evolve over time and new genres emerge which may not be accurately represented in historical data

Findings



Sales in Europe have gradually increased while North American and Japanese sales are volatile and in decline.



Shooter and sports games continue to increase in share while there is a declining interest in platform and puzzle games. There is a consistent low interest in adventure and racing games.

Recommendations

New Insights		
Trouble in North America & success in Europe	**Declining markets still have growing subset categories**	**Opportunity within Emerging Markets**
• North America is a problem region – with the market share decrease in NA, it's left room for other regions to undercut NA and grow their share, with one of the main regions being Europe	• Consumer interests are changing North America (and around the world) • North American sales show that in a declining market, there are still genres that continue to perform well	• New markets are cannibalizing the existing regions' market shares – there is potential to continue their growth with the right strategy.
Recommendations		
Focus on high performing genres in NA	**Shift Resources to Europe**	**Focus on Emerging Markets**
• Allocate more resources and budget to promoting shooter and sports games & pull back on platform and puzzle games due to loss in interest • Alter marketing budgets, reorganize store layouts, and reallocate stock	• The more streamlined resourcing for North America should free up more resources that can be allocated in Europe • With the European population topping NA by 200 million, there is a larger target market aka more potential consumers to capture	• With a new 10% of market share coming from secondary geographical regions, there are entire new markets that can be explored • Source more granular data to identify specifically where these spikes are occurring and what is contributing to them, both internally and externally.

02 Influenza

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

Goal: Help a medical staffing agency that provides temporary workers to clinics/hospitals to plan for influenza season when additional staff are in high demand. The results will examine trends in influenza and how they can be used to proactively plan for staffing needs across the country.

02 Overview

Key Business Questions

1. What time(s) of year is influenza's impact so great that ancillary staff will be required to tackle additional patient load?
2. What areas of the USA are most likely to demand this extra support?
3. How do vulnerable populations factor into this matter?

Data & Tools

Data:

- Influenza deaths by geography: CDC Data Set
- Population data by geography, time, age, and gender: US Census Bureau Data Set

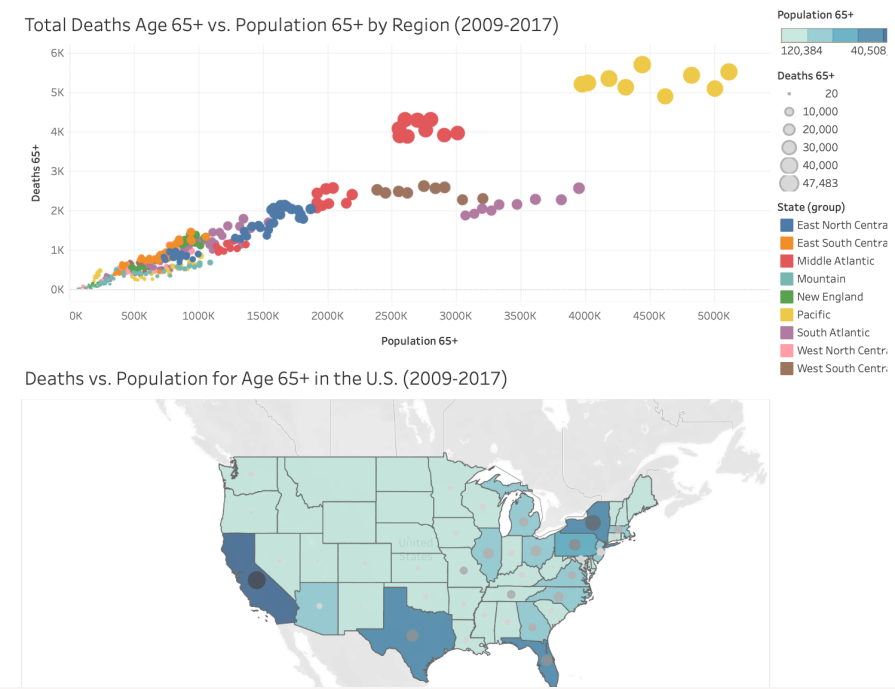
Tools/Skills:

- Excel & Tableau
- Data profiling and integrity
- Data transformation & integration
- Statistical hypothesis testing
- Data visualization Spatial & textual analysis
- Forecasting

Challenges & Limitations

1. **Data Quality:** Variations in reporting practices and healthcare infrastructure across states/years may cause inconsistencies in death/laboratory test data.
2. **Resource Constraints:** Finite number of staff available with no additional hiring permitted, risking under or overstaffing to address all patient needs.
3. **Predictive Accuracy:** Flu trends can be unpredictable i.e. new flu strains, public health initiatives, or changes in public behavior can significantly impact the severity of flu season.

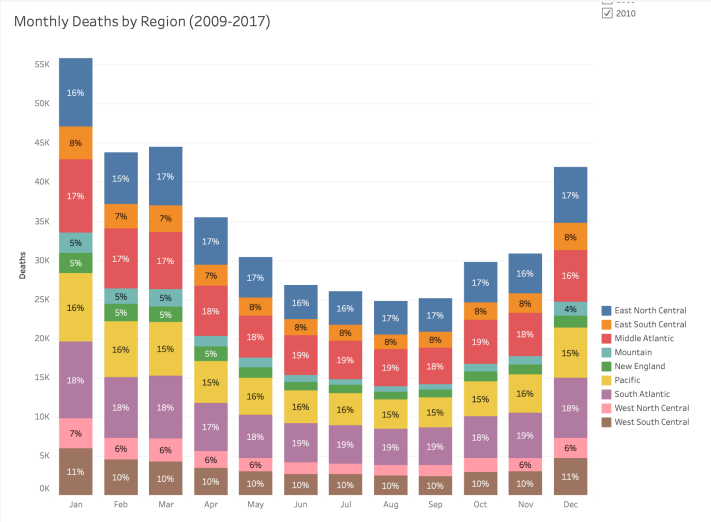
Findings



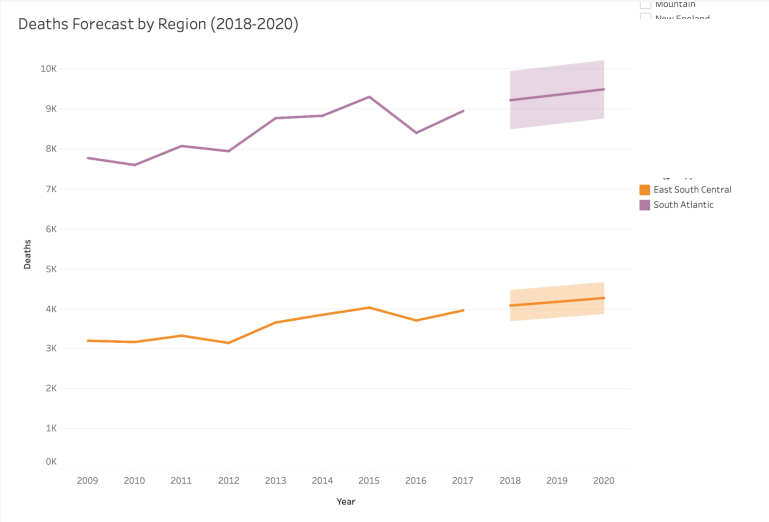
Relationship between population and deaths among 65+ age group:

- The trendline indicates a positive correlation, meaning that as the 65s+ population increases, so do their deaths.

Findings

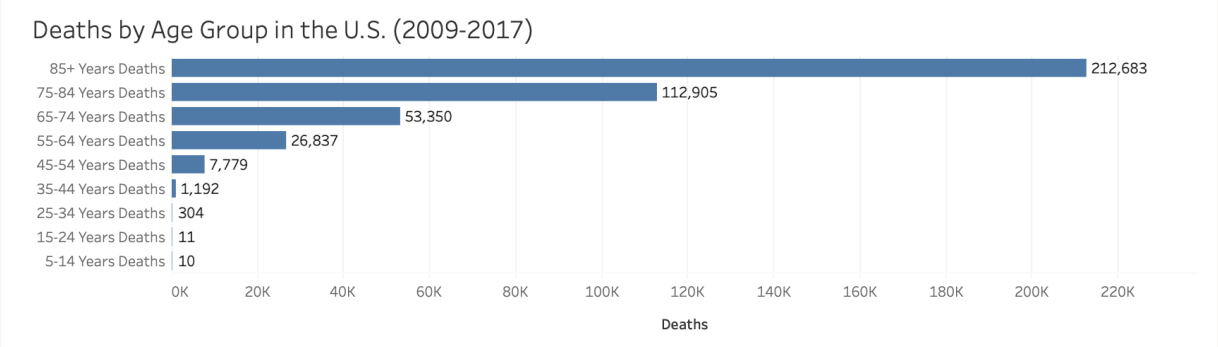


Influenza is seasonal – the highest mortality rates occur during the winter months of November through April.



Priority regions for influenza staffing include South Atlantic and East South Central states. The deaths for these regions are forecasted to increase by the greatest rates, by 6% and 7%, respectively.

Recommendations



WHERE to send staff: Deaths for populations aged 65 and older make up over 90% of total deaths. The staffing agency will need to prepare to send more medical personally to the following states with the largest vulnerable populations/deaths:

- California, Texas, Florida, North Carolina, Tennessee, Pennsylvania, Illinois, New York, Maine, Michigan

WHEN to send staff: The staffing agency will need to prepare to send more medical personnel during the winter/spring months, November-April as influenza is highly seasonality.

Tableau: found here

03 Rockbuster

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 service to stay competitive.

Goal: Analyze the performance of the current catalogue and develop a launch strategy.

03 Overview

Key Business Questions

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

Data & Tools

Data:

- Career Foundry Data Set: Rockbuster Data Set

Tools/Skills:

- Relational databases SQL
- Database querying
- Filtering data
- Summarizing & cleaning data
- Joining tables
- Performing subqueries
- Common table expressions
- Presenting SQL results

Challenges & Limitations

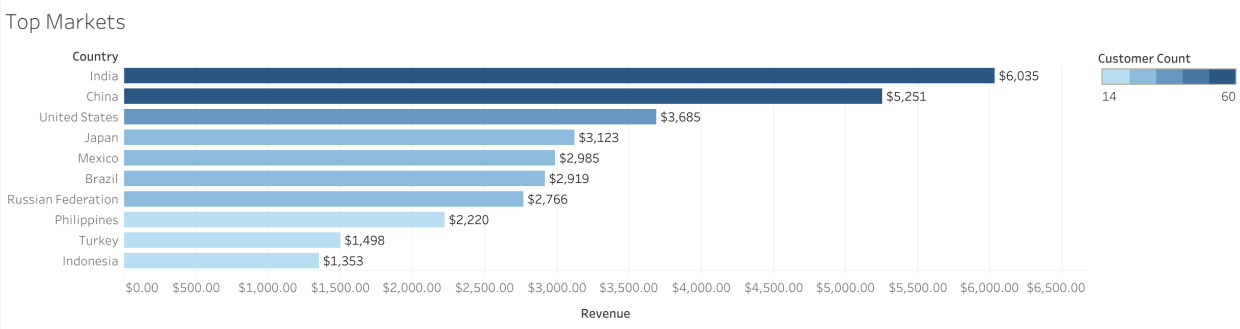
1. **Data Limitations:** Fabricated sales performance data set with limited information i.e. no rental costs or rental period included – some questions may require more granular data than what is available.
2. **Historical Context/Trends:** Data may not reflect recent trends in the evolving online video streaming market making it challenging to predict future performance due to market dynamics and changing consumer behavior.

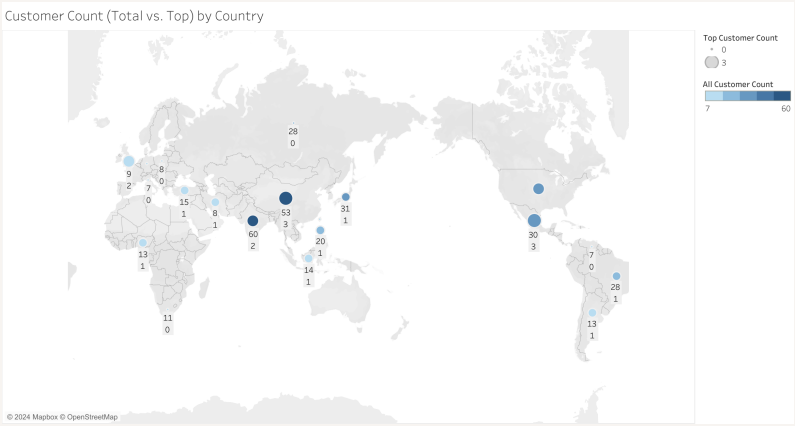
Findings

Unique Titles	Unique Genres	Average Rental Duration	Average Rental Rate	Average Customer Revenue	Languages
1001	17	4.98 days Min: 3 days Max: 7 days	$2.98 Min: 0.99 Max: 4.99	$102	1

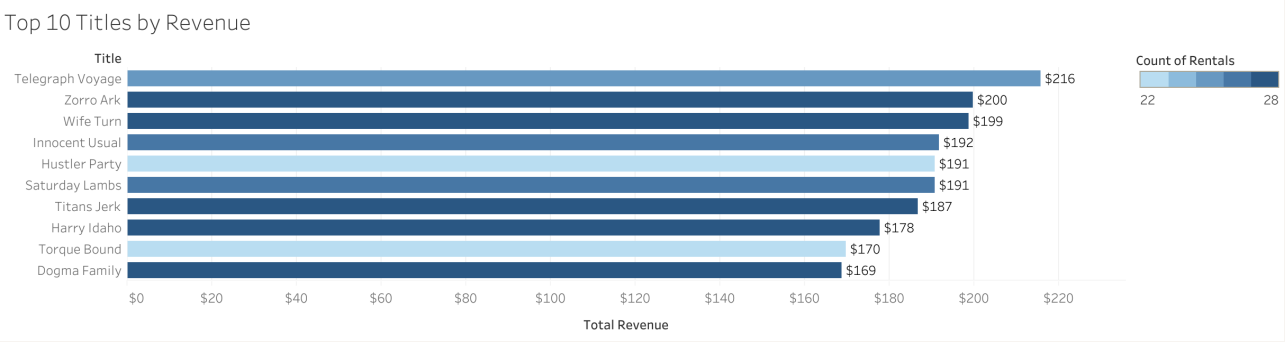
Findings

Top Markets: India, China, U.S & Japan



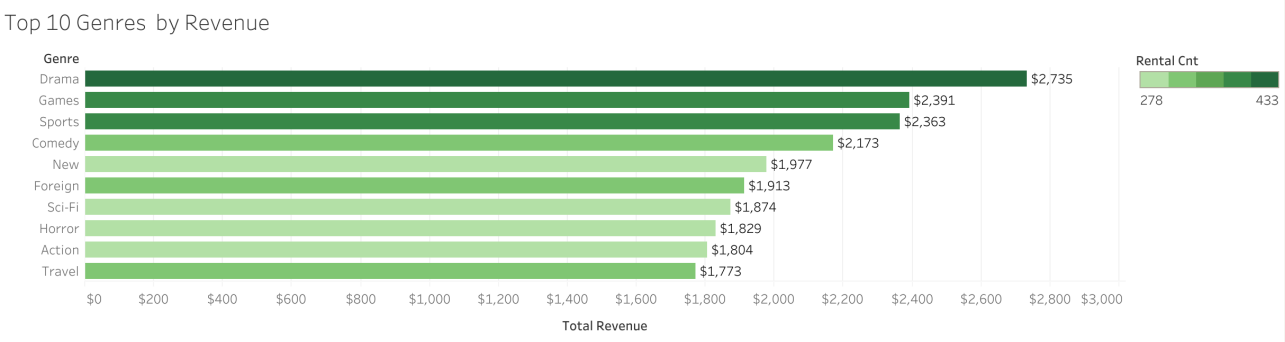


Findings



Top Titles:

- Telegraph Voyage
- Zorro Ark
- Wife Turn



Top Genres:

- Drama
- Games
- Sports

Recommendations

Diversify

- Diversify language offering
- Focus on high performing markets: India, China and the U.S.
- Diversify offerings to include titles in Hindi, Chinese and Japanese

Expand

- Expand library for popular genres
- Increase rental fees for popular genres
- Conduct further analysis to determine title additions to the drama, games and sports categories

Reimagine

- Adjust budget to support locations that are struggling and boost those that are thriving
- Consider introducing penalties for extended rental duration

Data Dictionary: found here
Full Analysis: found here
GitHub: found here

04 Instacart

Instacart, an online grocery store that operates through an app, wants to uncover more information about their sales patterns. The Instacart stakeholders are most interested in the variety of customers in their database along with their purchasing behaviors to implement a targeted marketing strategy.

Goal: Perform a data and exploratory analysis to derive insights and suggest strategies for better segmentation based on the provided criteria. Inform what this strategy might look like to ensure Instacart targets the right customer profiles with the appropriate products.

04 Overview

Key Business Questions

1. When do users spend the most money?
2. What are the most popular products?
3. What's the distribution among users in terms of brand loyalty? Are there differences in ordering habits based on a customer's loyalty status?
4. Are there differences in ordering habits based on a user's region?
5. Is there a connection between age and family status in terms of ordering habits?
6. What different classifications does the demographic information suggest?
7. What differences are there in the ordering habits of different customer profiles?

Data & Tools

Data:

- **CareerFoundry Data Sets:** Customers Data Set
- **Instacart:** Grocery Shopping Data Set
  Data Dictionary

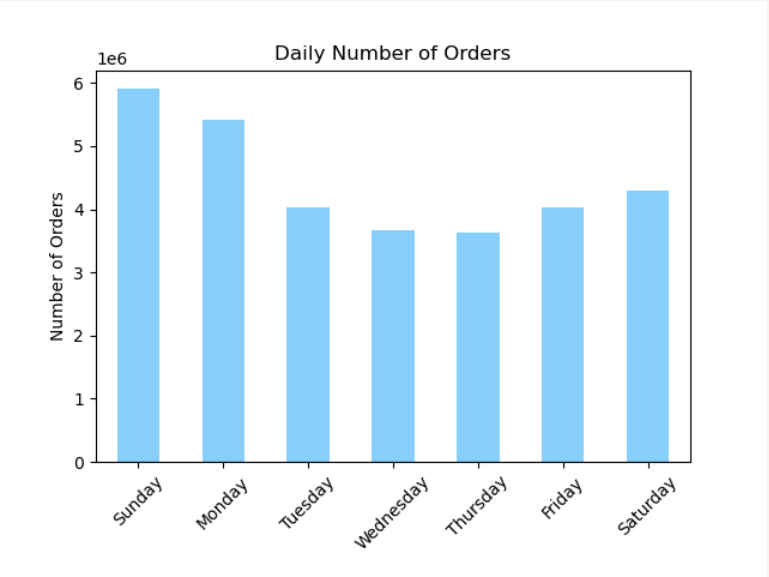
Tools/Skills:

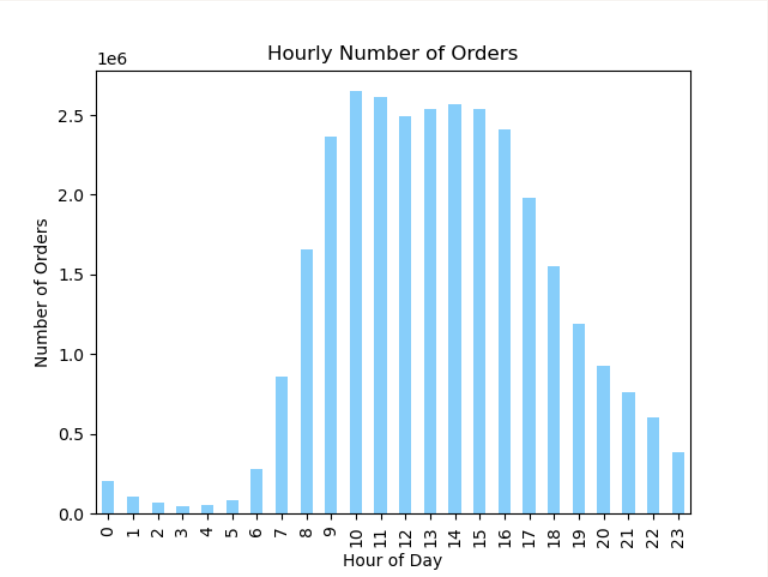
- Python
- Data wrangling & subsetting
- Data consistency
- Data combining & exporting
- Deriving new variables
- Grouping data
- Aggregating variables
- Data visualization with Python
- Reporting in Excel

Challenges & Limitations

1. **Data Limitations:** Fabricated customer data set on pricing with no information on the quantities of ordered items.
2. **Customer Degmentation:** Basic segmentation of current data may not capture the diverse and nuanced behaviors of different customer groups, which could lead to ineffective marketing strategies..
3. **Ethical Considerations/Data Privacy:** Maintaining ethical standards and customer data privacy is most important - strict data privacy regulations may limit access to PII data, limiting the scope of the analysis.

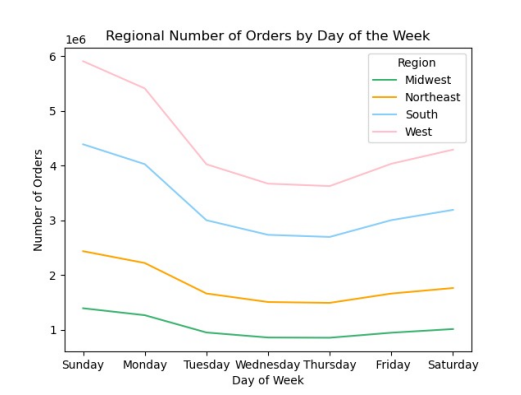
Findings

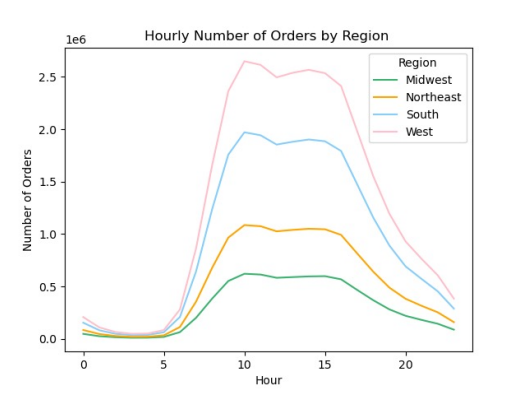




Busiest hours of the day: 9 AM to 5PM
Busiest days of the week: Saturday - Monday

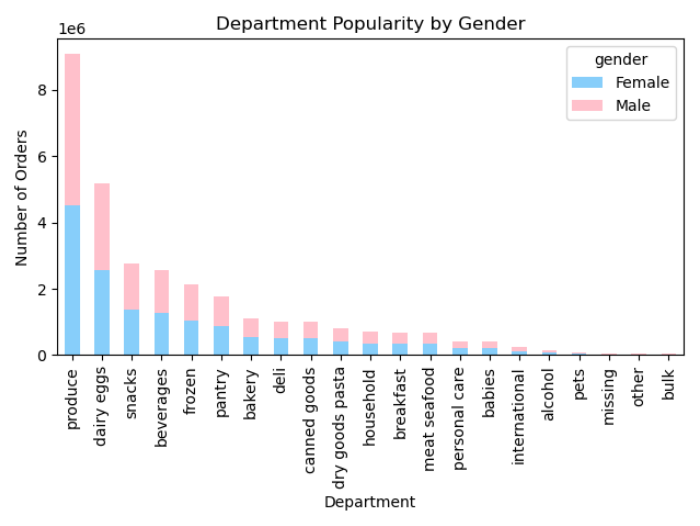
Findings - Geographic

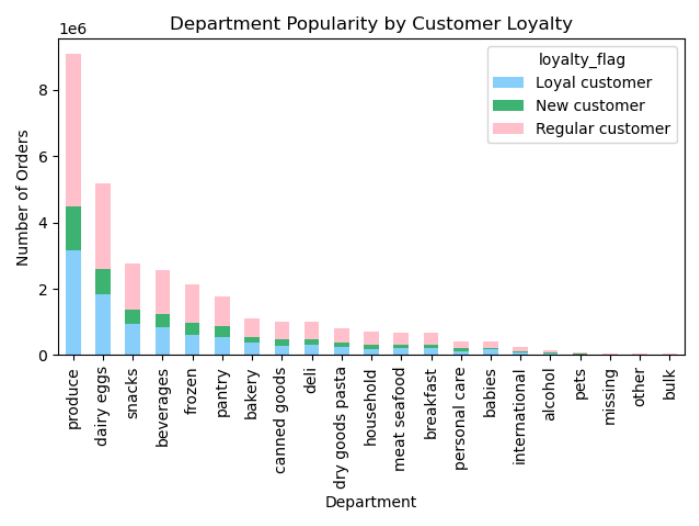




There are **no notable differences in purchasing behavior between the 4 regions** - for all regions, the average price, orders by day and orders by hour remain the same. Instacart does not need to segment its strategy by region.

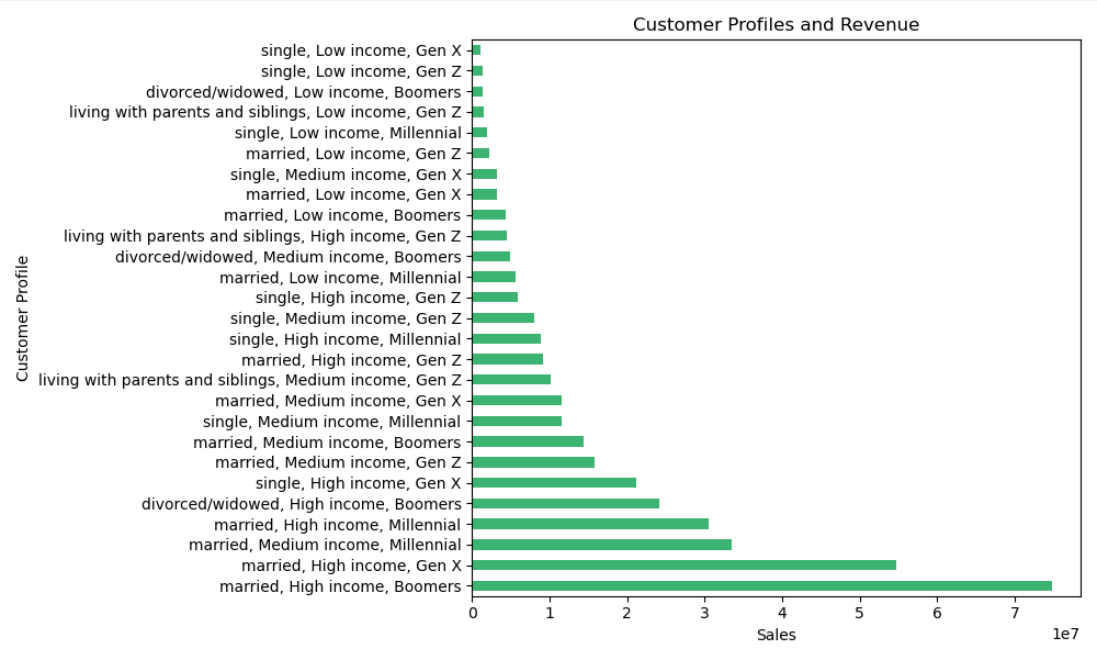
Findings – Customer Profiles





While there are segments that are higher revenue drivers, there is **no significant difference** in what items are being purchased between the different gender and loyalty segments alone.

Findings – Customer Profiles



Highest spenders:

- high income, married, Boomers
- high income, married, Gen Xers

Lowest spenders:

- single, low income, Gen X
- Single, low income, Gen Z

Recommendations

- **Schedule ads during the busiest hours of the day**, primarily 9 AM to 5PM, and refrain from advertising during the evening/early morning hours when consumers are asleep. Focus on advertising during the weekend into the new week (Saturday - Monday) as consumers are most active.

- **Focus on upselling regular customers to become loyal customers and maintaining loyal customers through extensive retention marketing** (email, push notifications, re-targeting through PPC, SMS, organic social, etc.).

- **Differentiate marketing strategy among customer profiles:**
  Older, wealthier, family-oriented – advertise via mail/print vs. online, suggest pricier or larger serving add-ons before checkout
  Younger, lower-income singles - push notifications for sales on previously bought or similar items, offer deals on produce with a closer sell-by-date or minor defects, suggest low-priced shelf-stable staples

- **Capitalize on popular departments,** produce and dairy/eggs: these items need to be bought frequently as they have a short shelf life. With these users being on the site more frequently, Instacart should target them more often and continue to collect data to improve their business.

Full Analysis: found here
GitHub: found here

05
Pig E. Bank

Pig E. Bank is a well-known global bank looking for anti-money laundering and customer retention support. The Pig E. Bank stakeholders are most interested in assessments of client risk and transaction risk, as well as reporting on metrics.

Goal: Highlight risk and help build/optimize models that assist the bank in running their compliance and retention programs more efficiently.

05 Overview

Key Business Questions

1. What are the leading factors leading to client loss?
2. Do certain customers have a higher risk of leaving the bank than others?
3. Are there any data privacy, data security or data bias issues?

Data & Tools

Data:

- Career Foundry Datasets: Pig E. Bank's customer data

Tools/Skills:

- Big data
- Data ethics – data bias and security & privacy
- Data mining
- Predictive analysis
- Time series analysis & forecasting
- GitHub

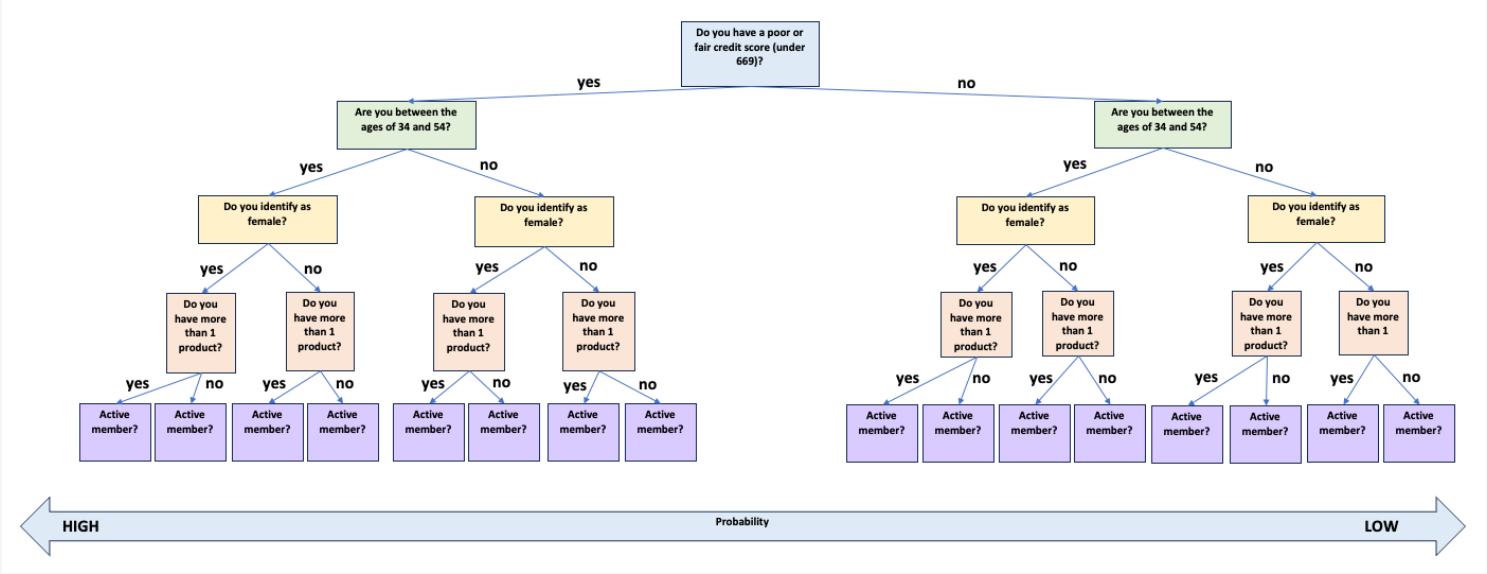
Challenges & Limitations

1. **Risk of Bias:**

- **Sample:** Dataset includes U.S. bank deposits within 100 miles of Mexico border & ATM withdrawals in Mexico. Possibility that this random sample is not representative of the total.
- **Collection & Measurement:** No information on analysts scoring ATM transactions – lack of understanding on who these analysts are, how they have been trained, what materials they have access to, and the accuracy of their classifications based on their interpretations.
- **Exclusion:** No details on how data has been collected or "cleaned" by the team leader.

Findings

This predictive model determines the **probability of customers leaving the bank**



The analysis determined the **highest risk factors** as:

- credit score, age, number of products, member activity status

Recommendations

Promote financial health

Higher credit scores decrease exit rates – provide financial advice and implement customer-centric features (ex. fee-free overdrafts, early paycheck access)

Incentivize activity

Inactive customers are more at risk of leaving – encourage customer activity by expanding number of products or offering promotions (ex. referral bonuses)

Customer Feedback

Better understand customer banking needs through regular surveys and interviews

Biases that may be present in the exercise:

- **Collection** – no explanation of how data has been collected as the team leader for Pig E. Bank has provided stale data from 2017
- **Sample** - the sample provided lacks detail on which ATM transactions are included in the dataset meaning we cannot be sure this is a random sample representative of the total
- **Exclusion** – no explanation of how data provided has been cleaned
- **Measurement** – lack of standardization in identifying transactions as “positive” or “negative”

06 Airbnb

Since 2008, Airbnb has revolutionized travel by offering guests and hosts the opportunity to explore unique and personalized experiences around the world. This dataset provides insights into listing activity and metrics within New York City.

Goal: Make predictions and draw conclusions to further enhance understanding of the Airbnb market in this dynamic urban landscape.

06 Overview

Key Business Questions

1. Is there evidence of spatial autocorrelation in listing prices or booking patterns? Are nearby listings more like each other in terms of pricing and popularity than listings farther apart?
2. Are there any specific features that correlate with higher ratings or booking frequency?
3. Can we forecast future demand or pricing trends based on historical data using time series analysis techniques?

Data & Tools

Data:

- Open-Sourced Airbnb Data: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Tools/Skills:

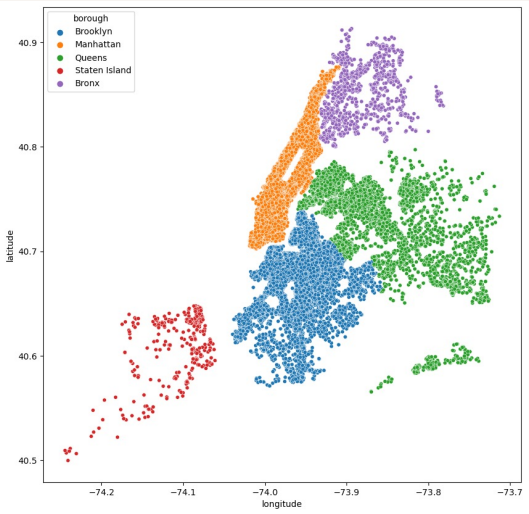
- Geographical visualizations with Python
- Supervised machine learning – regression
- Unsupervised machine learning – clustering
- Sourcing and analyzing time series data
- Creating data dashboards
- GitHub

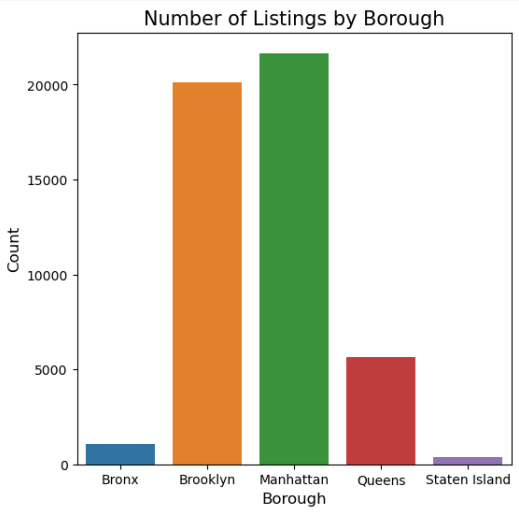
Challenges & Limitations

1. **Data Limitations:** Dataset spans 2011-2019 i.e. it is not inclusive of the past 5 years – the pandemic may have led new trends/patterns in rentals that would be more representative of the current state.
2. **Ethics:**

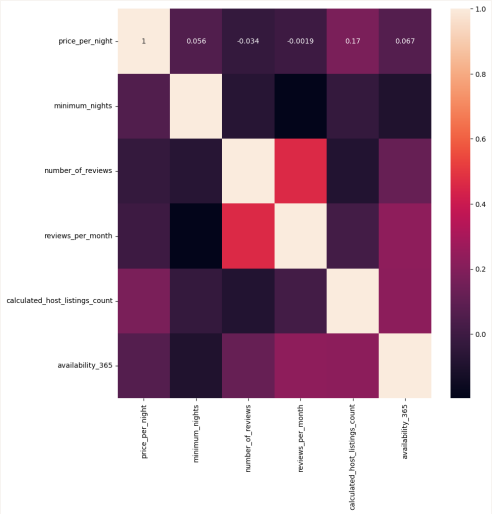
- No date of booking/stay available in this data set – date of last review has been used for time series.
- PII present in this dataset including host name, host ID and Airbnb ID –these fields are not relevant to analysis and have been removed.

Preliminary Analysis



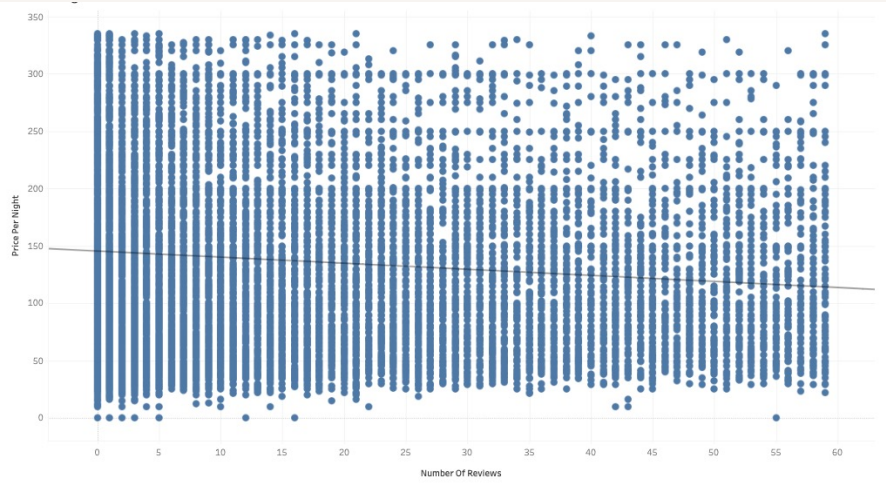


Geospatial Analysis:
As expected, Manhattan (the borough with the most tourist/visitor attractions) has the highest number of Airbnb listings with Brooklyn as a close follower.



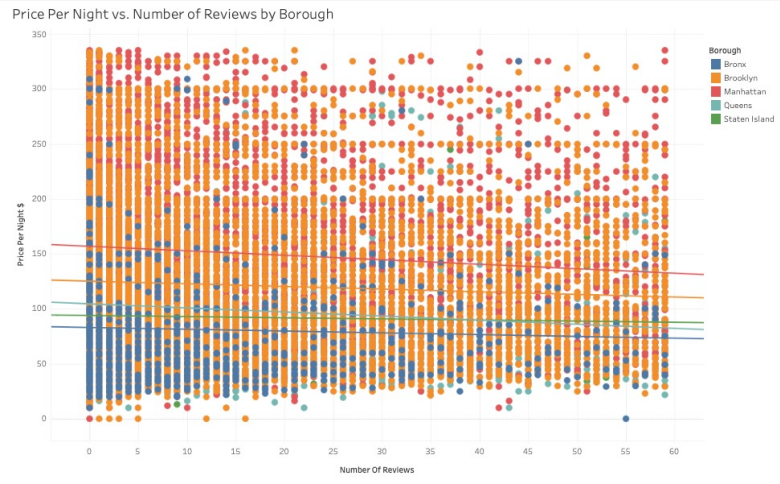
Exploratory Analysis:
Minimal correlation between any of the variables

Findings



Hypothesis: The higher the number of reviews, the higher the price per night. To test this hypothesis, a linear regression was conducted.

With an r-squared value of 0.01, we can deduce that this model is not a good fit.
With a p-value of near 0, we can reject our null hypothesis - there is no relationship between the number of reviews and the price per night of an Airbnb.



We see the same lack of fit when looking at each borough individually.

However, there are 2 boroughs where we **cannot** reject the null hypothesis due to the p-value: Staten Island (.66) and the Bronx (0.23).

Findings

Hypothesis tested via linear regressions: The higher the minimum nights required to book, the higher the year-round availability.

With an r-squared value of 0.01, we can deduce that this model is not a good fit. With a p-value of near 0, we can reject our null hypothesis - there is no relationship between the number of reviews and the price per night of an Airbnb when looking at all boroughs.

However, looking into individual boroughs, there are 3 boroughs where we **cannot** reject the null hypothesis due to the p-value: Staten Island (0.32), the Bronx (0.87) and Queens (0.22).

The dark purple cluster has the highest availability (i.e. days the Airbnb is available to book online) and the highest number of total Airbnb listings per host - both significantly larger than the pink and purple clusters.

Recommendations

Hypothesis testing: Both were disproved by a linear regression analysis where the r-squared and p-values were both insignificant.

1. The higher the number of reviews, the higher price per night
2. The higher the minimum nights required to book, the higher the year-round availability

Next Steps:

- Analyze the individual boroughs where we could NOT reject the null hypothesis:
  - Hypothesis 1: Staten Island and the Bronx
  - Hypothesis 2: Staten Island, the Bronx and Queens

07 ClimateWins

CimateWins, a European nonprofit organization, is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

Goal: Leverage machine learning techniques to analyze European climate data, identifying patterns and trends that could provide critical insights into how climate change is impacting different regions. Pinpoint which variables are the most relevant in forecasting future climate conditions.

07 Overview

Key Business Questions

1. Can machine learning be used to predict whether weather conditions will be favorable on a certain day? (If so, it could also be possible to predict danger.)
2. Historically, what have the maximums and minimums in temperature been?
3. Are there any ethical concerns surrounding machine learning and AI for this project?

Data & Tools

- Data set based on weather observations from 18 different European weather stations from the late 1800s to 2022.
- Includes daily recordings for temperature, wind speed, snow, global radiation, and more.
- Data set collected by: European Climate Assessment & Data Set Project

Challenges & Limitations

1. Risk of Bias:

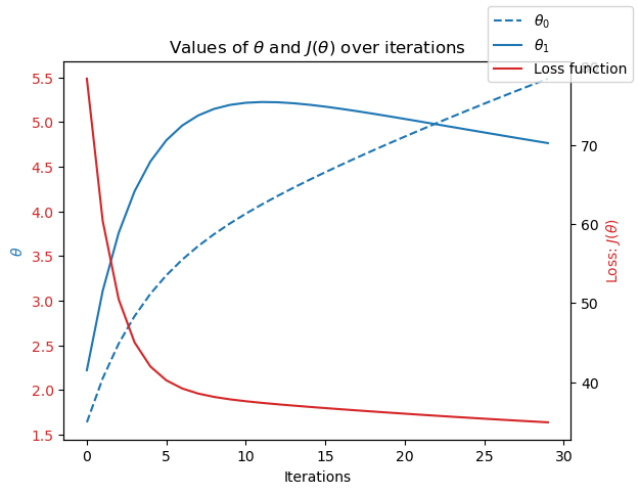
- **Measurement bias** – no confirmation of measurement standardization across each station (i.e. varying equipment) & more accuracy in current vs. past data as technology has improved
- **Human bias** – stations interpreting data in different ways
- **Selection bias** - are the 18 stations a good representation of Europe/the world?
- **Cultural bias** – cultural beliefs/values can shape how people perceive and respond to climate change. A machine learning algorithm that reflects certain European cultural perspectives may not be a good representation of the rest of the world with different cultural norms.

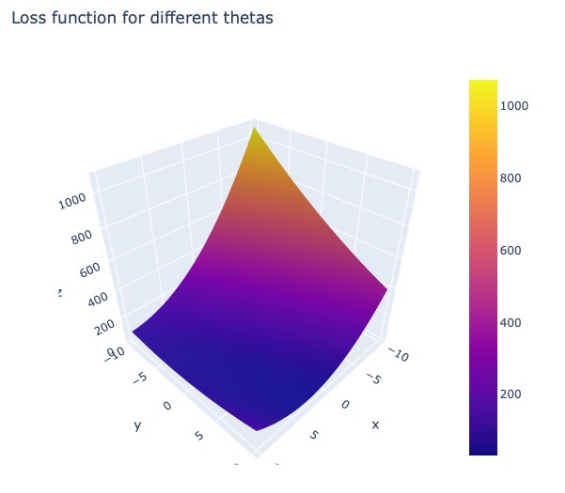
Hypotheses

- Machine learning models can accurately predict future weather patterns in Europe by analyzing historical weather data from the past century.
    - Training models on historical data to identify patterns/trends which allows for better forecasting of future conditions.
- Machine learning can identify significant changes in weather patterns, correlating these changes with rising global temperatures & other indicators of climate change.
    - Comparing recent data with historical trends so machine learning can highlight deviations and provide insights into the impact of climate change.
- Machine learning models can accurately predict daily weather conditions, including temperature, wind speed, and precipitation
    - Training on historical weather data so machine learning can develop models to provide reliable daily forecasts.
- ClimateWins can use existing weather data and machine learning tools to make significant advancements in climate prediction and fight climate change.

Gradient Descent Optimizations

Next step: Run optimizations (for 3 stations across 3 different years) to get the maximum and minimums for each station and measure how well a model's predictions match the actual data.





Station: Madrid
Year: 2018

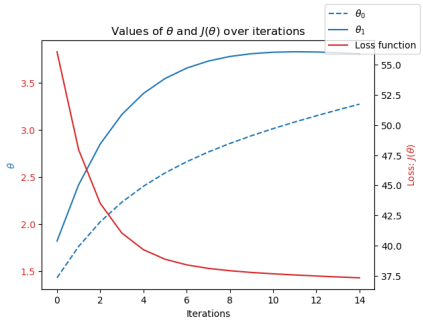
1st Optimization:
num_iterations=30
theta_init=np.array([[1],[1]])
alpha=0.05

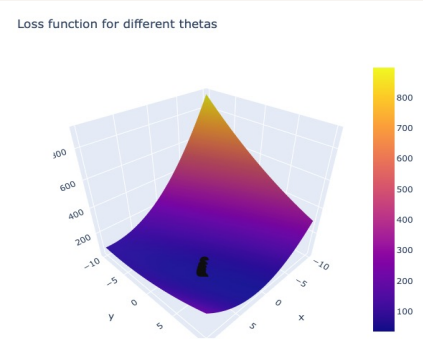
2nd Optimization:
num_iterations=30
theta_init=np.array([[4],[6]])
alpha= 0.05

Goal: For the loss profile, move the red line to trend toward 0 to get the smallest possible error and improve accuracy. The 3D "canyon" shows the path to the lowest loss per the optimization inputs.

Gradient Descent Optimizations

Station: Budapest
Year: 2017



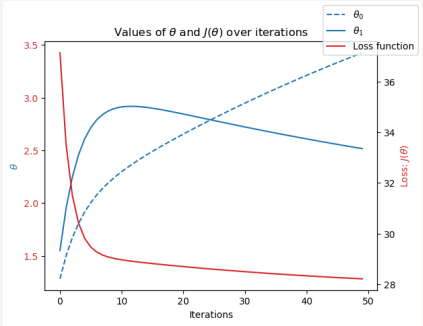


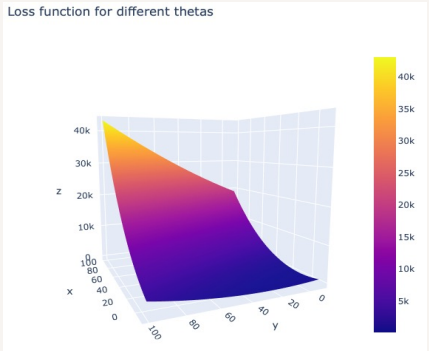
1st Optimization:
num_iterations=15
theta_init=np.array([[1],[1]])
alpha=0.05

2nd Optimization:
num_iterations=50
theta_init=np.array([[-1],[-1]])
alpha= 0.05

3rd Optimization:
num_iterations=50
theta_init=np.array([[-1],[-1]]
alpha= 0.01

Station: Budapest
Year: 1997





1st Optimization:
num_iterations=50
theta_init=np.array([[1],[1]])
alpha=0.05

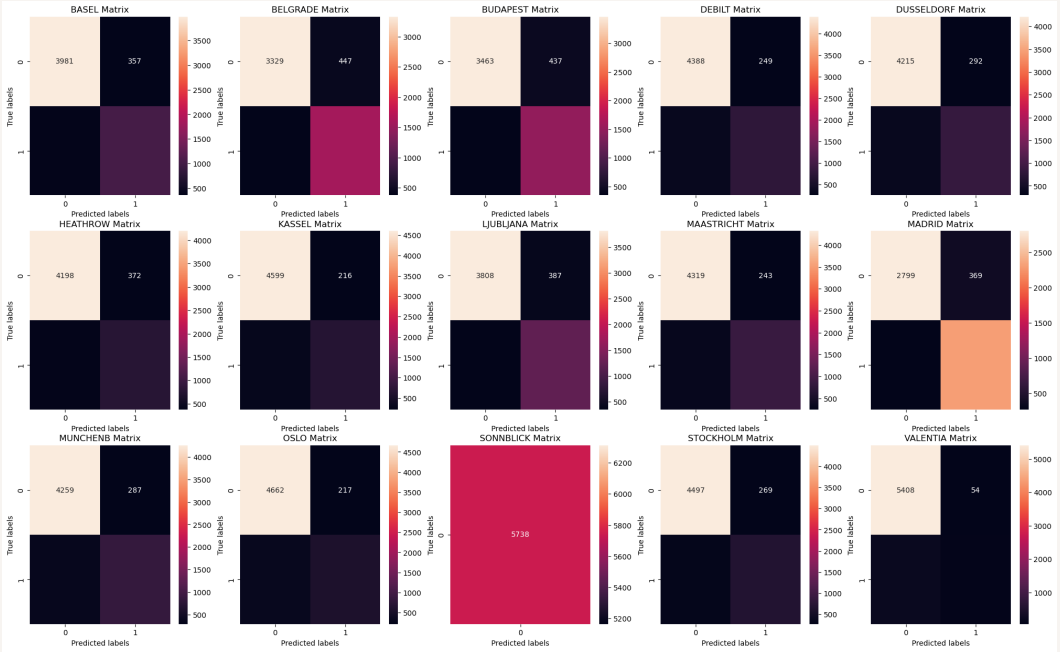
2nd Optimization:
num_iterations=50
theta_init=np.array([[0],[0]])
alpha= 0.05

KNN Model – Confusion Matrix

Accuracy score: 52%

What is a confusion matrix?

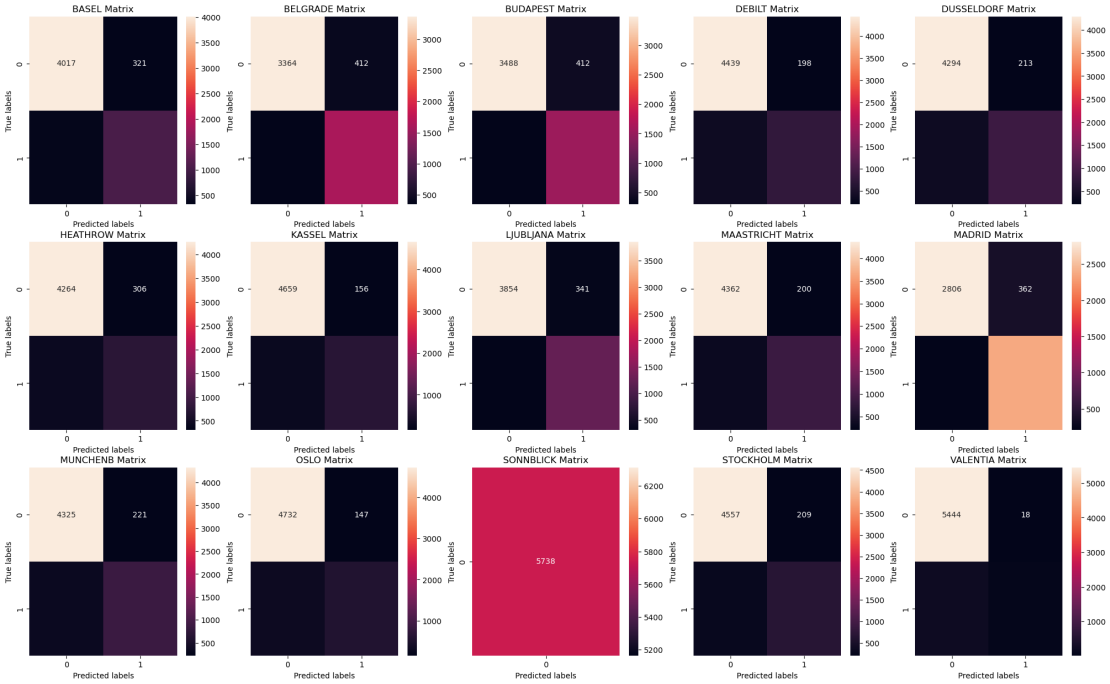
A confusion matrix is a tool used in machine learning to represent the accuracy of a model – is this example, the accuracy of the algorithm for each station.. It shows the **actual vs. predicted values** and how many of the variables have been classified or misclassified by the model in the four quadrants (i.,e. true positives, true negatives, false positives, and false negatives).



#Run the model with neighbors equal to 1 to 4, test the accuracy
k_range = np.arange(1,4)

KNN Model – Confusion Matrix

Accuracy score: 58%



#Run the model with neighbors equal to 1 to 10, test the accuracy
k_range = np.arange(1,10)

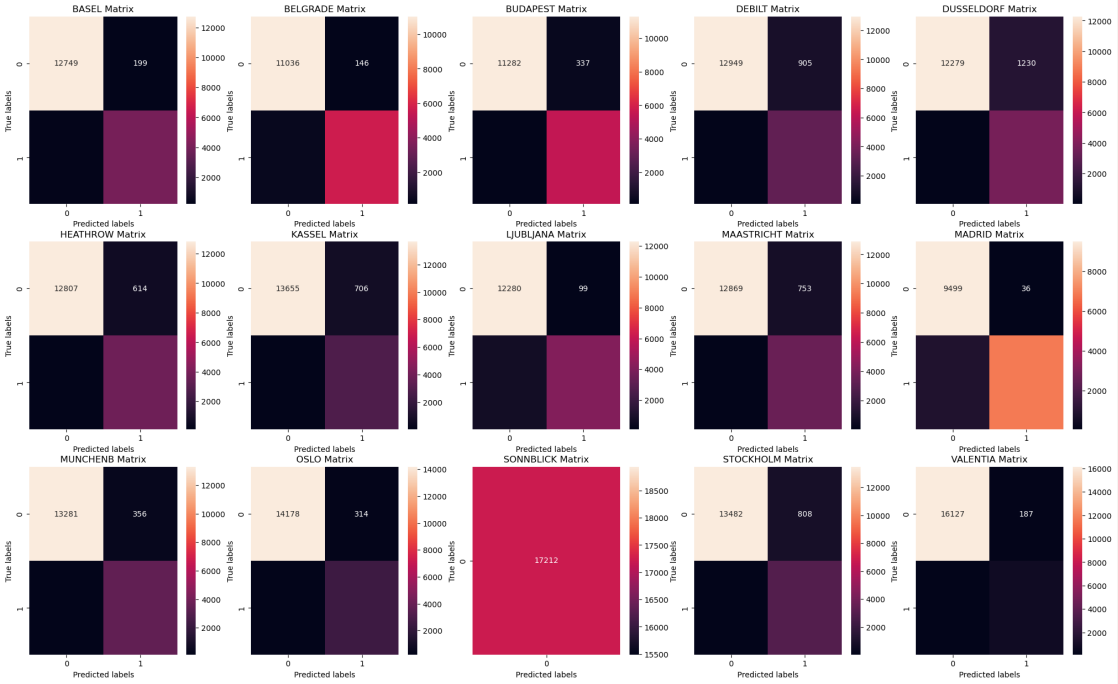
The KNN model is **not** an accurate predictor as most quadrants in the confusion matrix show **no relationship.**
For each station, only the top left-hand corner (the variables 0 and 0) shows the occurrences when the variables were labeled correctly.

Features that may contribute to overall inaccuracy:

- Overfitting – the algorithm may conform to one aspect and not represent the entire data set.
- Some weather stations may have more unique/volatile weather conditions that could impact the “outliers”.
- Stations more exposed to environmental factors = more susceptible to inaccurately collected data (ex. Stockholm is situated on islands & Oslo is surrounded by forests/mountains)

ANN Model – Confusion Matrix

Accuracy score: 82%

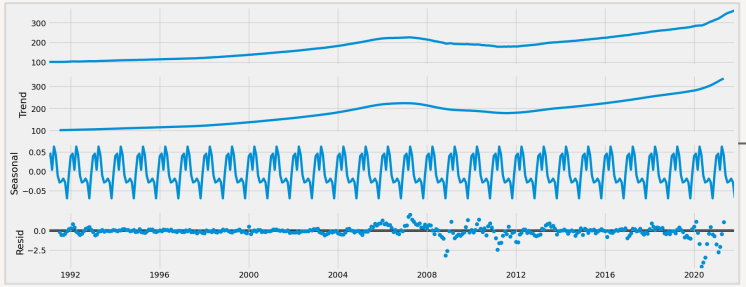


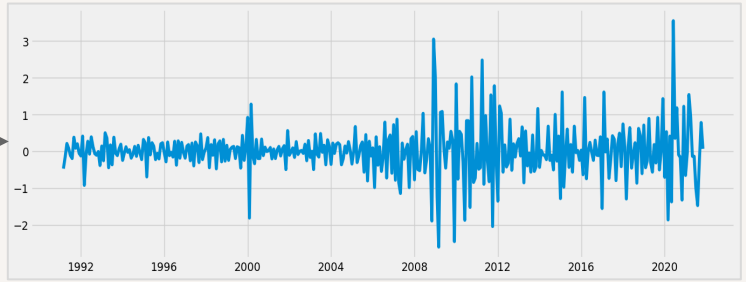
The ANN model is the better predictor in comparison to the KNN model. The number of correct labels is much higher across the stations.

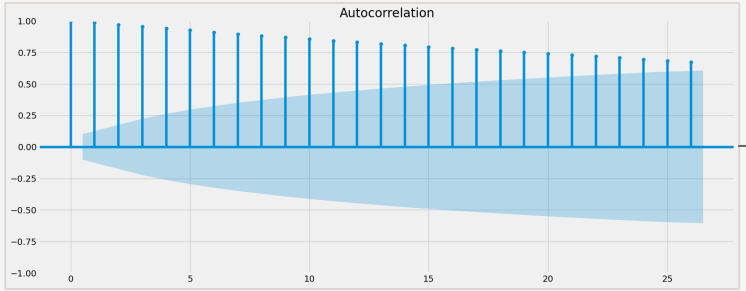
Key Takeaways

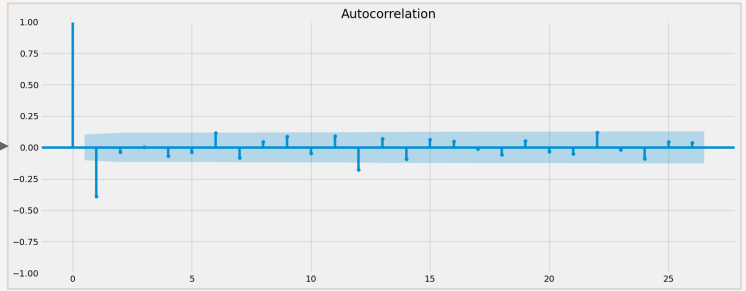
- Neither the confusion matrixes for the KNN or ANN model are fully accurate by station.
    - Example of overfitting in Sonnblick - overfitting may be caused by the model learning not only the underlying patterns in the training data but also the noise/random fluctuations. Sonnblick is a mountainous area in Austria that most likely incurs large weather fluctuations.
- Accuracy of a model is influenced by the quality (accuracy and completeness) and relevance of the datasets. Relevant features are going to provide stronger predictive power and reduce noise. It's important to use features that are strongly correlated with the target variable of the model – in this example, we have access to each stations cloud cover, wind speed, humidity and precipitation to name a few.
    - Next steps: Trim down these features, as making the model simpler could lower the risk of overfitting. It's important to use relevant features so the model can learn the underlying patterns rather than false correlations, leading to greater accuracy in real-world applications.
- Overall, the **ANN model should be recommended** over the KNN model to ClimateWins as it is more successful in predicting weather station data.

Bonus – Time Series









Conducted a **Dickey-Fuller test** on real estate data to check for stationarity & performed a round of differencing to check the data's autocorrelations.

Thank you!

Any questions?

meghanmcgrath97@gmail.com
(973) 309 - 5186