

DATA ANALYTICS PORTFOLIO

SHRAVANI IYTHA SUBRAMANYAM

Student Data Analyst Portfolio



About Me



Hi, I'm Shravani,

My journey started in finance and auditing, where I developed a knack for uncovering patterns and solving problems. Over time, my curiosity about data grew, leading me to transition into data analysis—a field where I combine my analytical mindset with a love for storytelling through numbers .

I enjoy diving into complex datasets, whether it's to uncover trends, solve puzzles, or create visuals that bring insights to life. With tools like Python, SQL, Tableau, and Power BI, I craft solutions that make a difference. When I'm not working, you'll find me exploring new cuisines, traveling, or brushing up on languages . Feel free to explore my portfolio to learn more about my work and connect with me for potential collaborations.

PROJECTS

01

GameCo.

Global market analysis of video game sales.

02

Preparing for Influenza Season

Staff deployment planning for influenza season.

03

Rockbuster

Launching Rockbuster Stealth online movie service.

04

Instacart

Market segmentation analysis to uncover sales.

05

Pig E.Bank

Analysing customer attrition.

06

Citibike Analysis

Analyzing the usage Trends and Station dynamics.

07

ClimateWins

choose an appropriate machine learning algorithm to predict climate change

TOOLS



Microsoft Excel
Tableau,
PowerPoint



Microsoft Excel
Tableau



PostgreSQL
Tableau
GitHub



Python
GitHub



Microsoft Excel



Python
Tableau
GitHub



Python
Tableau
GitHub



01 GAMECO MARKET ANALYSIS

Analyzing global video game sales



Company Overview:

Game Co. is a leading global gaming company with a strong presence in key markets, including North America, Europe, Japan, and other regions. The company offers a wide variety of games across multiple genres, available for both purchase and rental, catering to a diverse audience of gamers worldwide.

Project Objective:

In October 2016, I was tasked by Game Co's executive board to analyze historical game sales and rental data for the 2017 marketing budget, with an assumption of stable sales across regions. The goal was to identify key trends and patterns to guide strategic decisions, potentially redistributing the marketing budget for maximum ROI. Although I had limited data expertise at the time, I was responsible for presenting actionable insights effectively to help optimize the timing and positioning of future game releases in various markets. This project focused on developing predictive models to forecast market reception and improve overall marketing efficiency.

GAMECO MARKET ANALYSIS

1

Are certain types of games more popular than others?

2

What other publishers will likely be the main competitors in certain markets?

3

How have their sales figures varied between geographic regions over time?

4

Have any games decreased or increased in popularity over time?



DATA

Data source: [VGChartz](#). ([Link](#))

File: Excel - CSV

Period of data: 1980-2016

Regions: North America, Europe, Japan, and others

Information: title, platforms, year, genre, publisher



SKILLS APPLIED

For this project, I used Excel. For specific visualizations, I used Tableau.

- Improving data quality
- Data grouping and summarizing
- Descriptive analysis
- Pivot table
- Visualization results in Excel and Tableau
- Presenting results ([link](#))

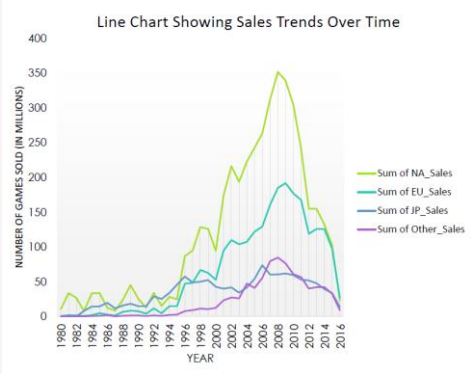


CHALLENGES

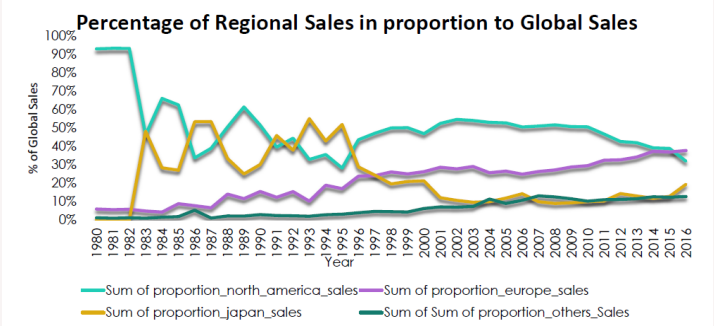
In my first project, I embraced the opportunity to learn and translate business requirements into actionable analysis. Despite everything being new, I quickly adapted and successfully aligned the analysis with business goals.

ANALYSIS

Historical Trends 1980-2016



Regional Sales Trends by Percentage of Global Sales



There is a general upward trend in global sales, with notable peaks in the mid-2000s. and then gradually decline in recent years. With notably high sales include 2006, 2008, and 2009. These peak years might be influenced by factors such as the release of popular gaming consoles, blockbuster game titles, or other industry trends

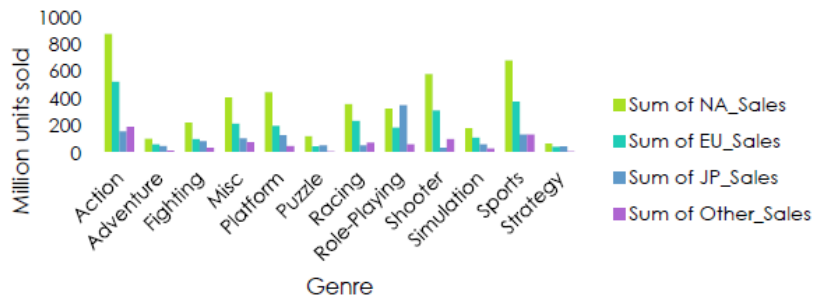
This trend can be seen across all regions.
70.93 million games were sold in 2016, **90%** less than in 2008.

Based on the above chart, it is clear that the sales over years have not been the same throughout the regions.
In the view of marketing budget allocation ,It is proposed for GameCo to focus more on EU's growing market .EU has steadily increased over the years ,overtaking JP in 1997 and taking over NA in 2015.

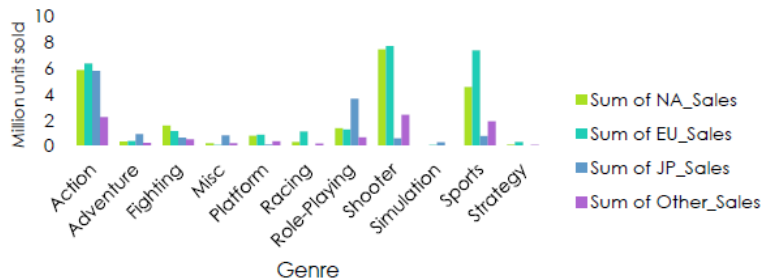
ANALYSIS

Analysis of the Most Popular Game Genres (1980-2016)

Most popular genre in each region(1980-2016)



Most popular genre in each region 2016



Top 10 Selling Games From 1980-2016:



All these games were published by Nintendo.

- 5 of these games were for the Wii platform, but no Wii games were purchased in 2016.
- The most popular platform in 2016 is the PS4, making up 55% of all video game sales this year.
- Recently, Sports and Shooter games became the highest selling genre in North America and Europe.
- Action games, however, still make up the most sales overall. In 2016 Action games made up 28% of all sales.

ANALYSIS

New Publisher Trends Emerging in 2016

Top 10 Selling Games of 2016

1. FIFA 17 (PS4)
2. Uncharted 4: A Thief's End (PS4)
3. Tom Clancy's The Division (PS4)
4. Far Cry Primal (PS4)
5. Tom Clancy's The Division (XOne)
6. Overwatch (PS4)
7. No Man's Sky (PS4)
8. Dark Souls III (PS4)
9. FIFA 17 (XOne)
10. Doom 2016 (PS4)

Top 5 Selling Publishers Percent of Sales by Popular Genres

Publishers	Action	Fighting	Role-Playing	Shooter	Sports
Electronic Arts	--	--	--	8%	88%
Ubisoft	39%	--	--	61%	--
Sony Computer Entertainment	18%	--	--	64%	--
Namco Bandai Games	12%	42%	40%	--	--
Activision	6%	--	8%	86%	--

Tom Clancy's The Division and FIFA 17 both show up twice, selling on different platforms

Most of the sales from top publishers were in popular genres, but there are some gaps as well.

- ❑ Electronic Arts didn't have any Action game sales.
- ❑ Ubisoft, Sony Computer Entertainment, Namco Bandai Games, and Activision didn't have any Sports game sales.
- ❑ Namco is unique for having most sales from Fighting and Role-Playing genre games.

INSIGHTS & RECOMMENDATIONS

KEY LEARNINGS

1

Despite a general sales decline since 2010, Europe surpassed North America in market share post-2015, with Japan also experiencing a notable sales spike.

2

While Action, Shooter, and Sports dominate in North America and Europe, Japan has seen a shift towards Action genres, although Role-Playing games remain the second most preferred genre.

3

Platform popularity has undergone notable shifts; PS4 and Xbox One have emerged as leaders in the West, replacing past favorites, while in Japan, the enduring preference for the 3DS highlights distinct and lasting regional differences.

4

The marked post-2015 divergence in regional preferences highlights the need for tailored strategies in game development and marketing.

RECOMMENDATIONS

Global Marketing Strategy

- Adjust marketing strategies to match regional sales trends and shifts in market preferences post-2015.
- Craft advertising campaigns that align with regional tastes and current trends to boost engagement and market share.

Genre

- **North America and Europe:** Increase focus on Shooters and Sports genres, which continue to perform strongly.
- **Japan:** Capitalize on the growing interest in Action genres while also prioritizing Role-Playing Games, which remain highly popular,

Platform

- **North America and Europe:** Leverage the established console base, particularly the popularity of PS4 and Xbox One.
- **Japan:** Continue to capitalize on the strong preference for handheld platforms like the 3DS.
- **General:** Maintain vigilance for emerging platforms to quickly adapt to shifts in platform popularity.

LINKS &
DELIVERABLES



Final Presentation

Reflection

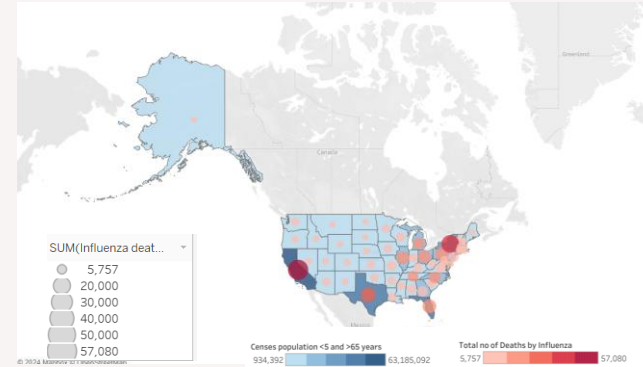
*Please click links above to view relevant project work

02 Preparing for Influenza Season

Preparing for flu season in the U.S..

OBJECTIVE

In the U.S., when flu season ramps up, hospitals require extra help, especially for vulnerable individuals facing complications. As their data analyst, I'm here to forecast the optimal timing and staffing numbers for each state, ensuring a well-coordinated response to provide the necessary care.



Goal: Analyze trends for a medical staffing agency during influenza season, ensuring proactive national staffing planning for increased demand.

Preparing for Influenza Season

KEY REQUIREMENTS

1

Support staffing plan with data on medical personnel distribution in the U.S.

2

Investigate seasonality of influenza across states.

3

Prioritize states based on vulnerable population size and categorize as low-, medium-, or high-need.

4

Identify data limitations that hinder analysis.



DATA

1. Census Population Dataset

Source: US Census Bureau

Contents: Population information from the US by country, time, age and gender for 2009-2017.

2. Influenza Deaths Dataset

Source: CDC

Contents: Information about influenza deaths by age groups in the US by state and time for between 2009-2017.



SKILLS APPLIED

- Translating business requirements into analytical questions
- Sourcing relevant datasets
- Data integration and cleaning
- Statistical hypothesis testing
- Visual analysis in Tableau
- Forecasting
- Storytelling in Tableau
- Presenting results to an audience



CHALLENGES

A key challenge for this project was effectively integrating diverse data sources to ensure cohesive analysis.

Another significant challenge was managing 'suppressed' variables, initially posing uncertainties in how to effectively utilize them in the analysis.

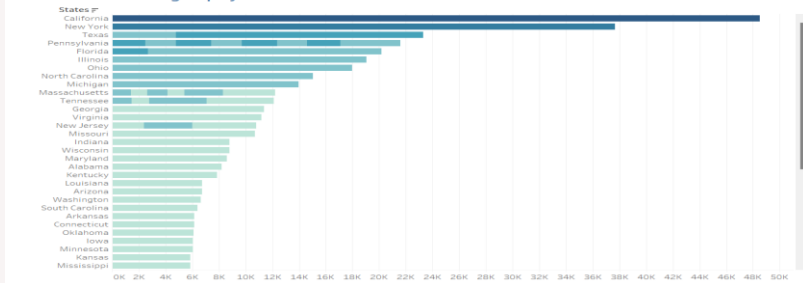
Preparing for Influenza Season

ANALYSIS

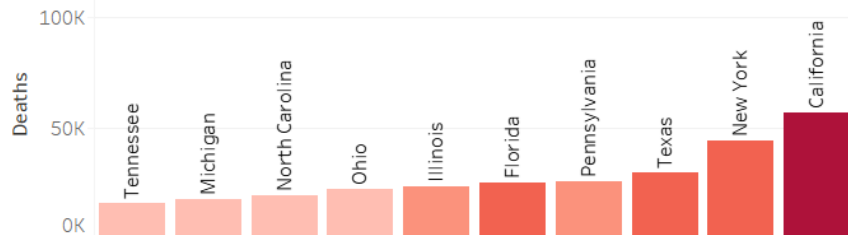
Which States experience the highest impact from the flu?

States with highest need for additional medical staff are States(California , New York , Texas , Pennsylvania and Florida) with highest population of people from Vulnerable group and at the same time with highest count of influenza caused deaths

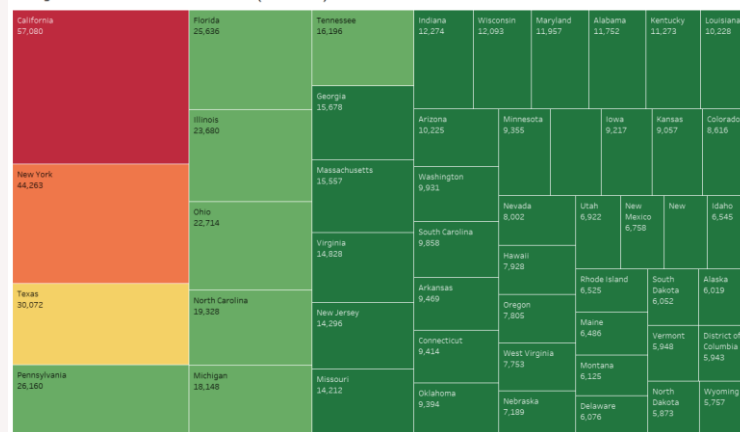
Deaths in Vulnerable group by State



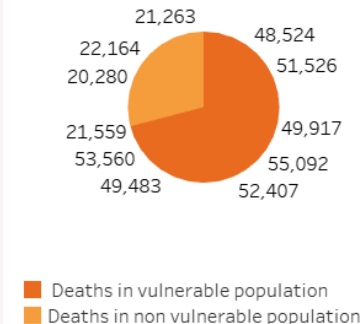
Top 10 States with most Influenza Death Rates



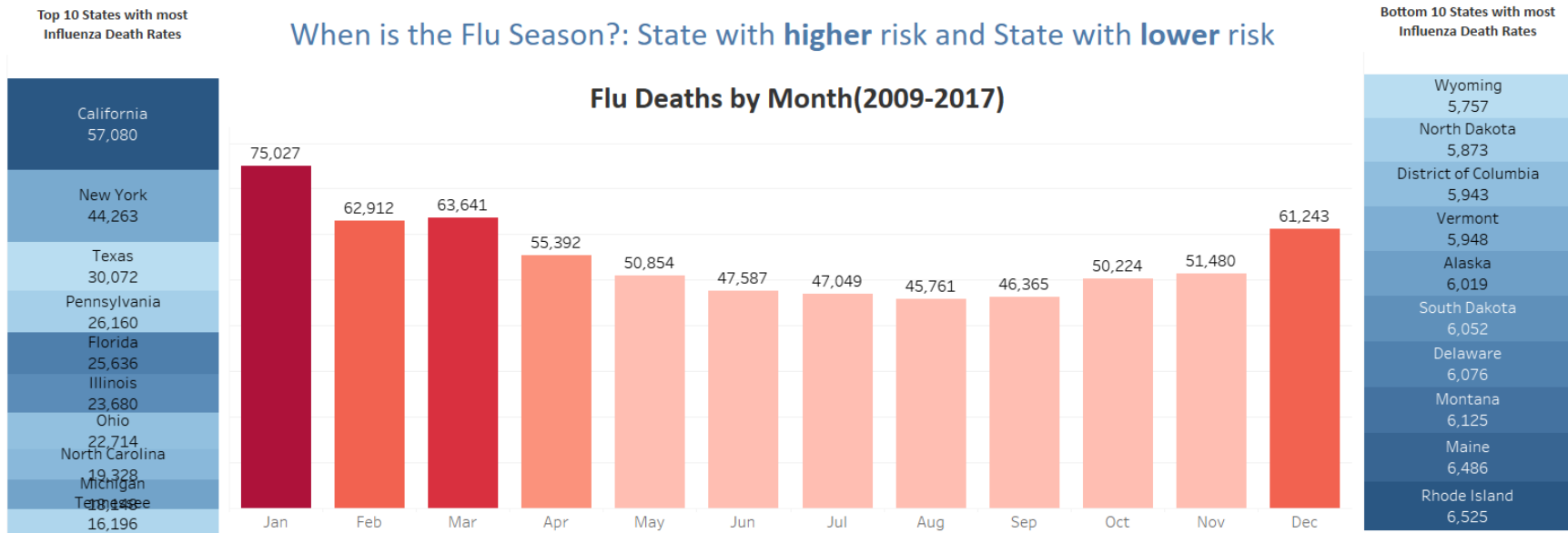
Average State-Wise Influenza Total Deaths(2009-2017)



Vulnerable group
Elderly >65 years & children <5 years



When is the peak season for influenza?



- In the States with higher risk, the flu season typically begins in mid-fall, peaks in **January**, and then gradually declines. While in the States with the lower risk, the number of influenza deaths is more dispersed throughout the year.
- Most states coincide that **January** is the month when they reach the **peak** of influenza deaths.

INSIGHTS & RECOMMENDATIONS

KEY LEARNINGS

1

States with a larger population of individuals aged 65 and older experience more influenza-related deaths.

2

The flu season starts in mid-fall, peaking in January, especially in higher-risk states. While lower-risk states see more dispersed influenza deaths, most states experience a consistent pattern: deaths rise in December, peak in January, and stay high through March.

3

The virus's impact on vulnerable age groups highlights the need for targeted interventions and preventative measures.

RECOMMENDATIONS

Priority

States with larger populations over 65 years of age should increase their medical staffing during the influenza season.

Very high- and high-need states:
California, New York, Texas, Pennsylvania.

Seasonality

To prepare for the influenza season, which spans from **December to March**, medical staff should be allocated to each state based on priority.

Further Analysis

States with smaller populations or fewer individuals over 65 but higher influenza death rates warrant closer examination to identify **additional contributing factors**. Future analysis should **explore influenza's impact on all vulnerable groups** such as pregnant women, individuals with HIV/AIDS, and other diseases.

LINKS & DELIVERABLES



[Tableau Story Board](#)



[Interim Report](#)



[Project Management Plan](#)

*Please click links above to view relevant project work



03 Rockbuster Stealth LLC:

Transition to Online Video Rental Service

Rockbuster Stealth LLC, a global movie rental company with traditional physical stores, faces increasing competition from digital streaming platforms such as Netflix and Amazon Prime. In response, Rockbuster is considering a transition to an online video rental model.

OBJECTIVE

Rockbuster Stealth LLC is a video rental company tasked with launching an online video service to stay competitive against streaming giants. As their data analyst, my responsibilities include loading data into RDBMS and utilizing SQL for insightful analysis, supporting various departments with ad-hoc queries.

Goal: The Rockbuster Stealth Management Board has asked a series of business questions and they expect data-driven answers that they can use for their 2020 company strategy.

KEY 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

Data Source: Postgres Tutorial Data Set

SKILLS APPLIED

- Relational databases
- SQL
- Creating a data dictionary
- Database querying
- Data filtering
- Data cleaning and summarizing
- Joining tables
- Subqueries
- Common table expressions
- Presentation

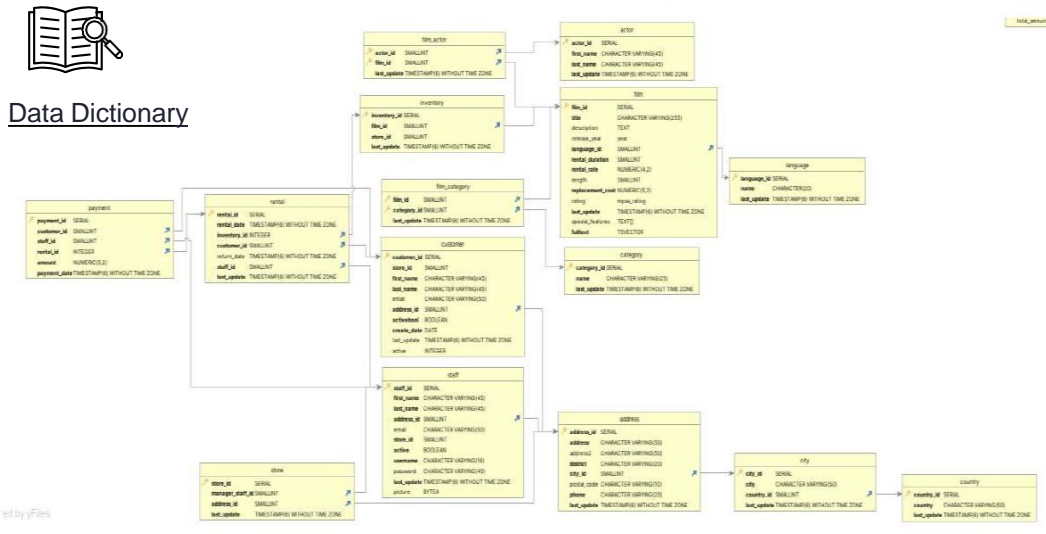


CHALLENGES

Challenges included managing complex queries and maintaining data accuracy.

Rockbuster Stealth ERD and overview of statistics

Rockbuster Entity Relationship Diagram (ERD)



Data Dictionary

Calculate descriptive statistics for numerical columns

SELECT

```
MIN(rental_duration) AS min_rent_duration,
MAX(rental_duration) AS max_rent_duration,
round(AVG(rental_duration),2) AS avg_rent_duration,
COUNT(rental_duration) AS count_rental_duration,
MIN(rental_rate) AS min_rent_rate,
MAX(rental_rate) AS max_rent_rate,
round(AVG(rental_rate),2) AS avg_rent_rate,
COUNT(rental_rate) AS count_rental_rate,
MIN(length) AS min_length,
MAX(length) AS max_length,
round(AVG(length), 2) AS avg_length,
COUNT(length) AS count_length,
MIN(replacement_cost) AS min_replace_cost,
MAX(replacement_cost) AS max_replace_cost,
round(AVG(replacement_cost),2) AS avg_replace_cost,
COUNT(replacement_cost) AS count_replace_cost,
COUNT(*) AS count_rows
```

FROM film;

This Entity-Relationship Diagram (ERD) snowflake schema was extracted using DbVisualizer for the purpose of visual representation used in my SQL data analysis to illustrate the relationships between entities (tables) of Rockbuster Stealth in PostgreSQL.

•A screenshot taken from my Excel workbook displays one SQL query used in my exploratory data analysis for descriptive statistics.

ANALYSIS

Which regions and customers stand out as top performers for Rockbuster Stealth?

Top 10 customers from the top 10 cities who've paid the highest total amounts to Rockbuster. The customer team would like to reward them for their loyalty

First Name	Last Name	City	Country	
Arlene	Harvey	Ambattur	India	\$111.76
Kyle	Spurlock	Shanwei	China	\$109.71
Marlene	Welch	Iwaki	Japan	\$106.77
Glen	Talbert	Acua	Mexico	\$100.77
Clinton	Buford	Aurora	United States	\$98.76
Betty	White	Citrus Heights	United States	\$96.77
Francisco	Skidmore	So Leopoldo	Brazil	\$93.79
Dora	Medina	Tianjin	China	\$88.81
Norman	Currier	Cianjur	Indonesia	\$73.76
Juan	Fraley	Teboksary	Russian Federation	\$63.79

Top 10 countries with highest number of customers

India 6,035 60	United States 3,685 36	Brazil 2,919 28	Russian Federation 2,766 28
China 5,251 53	Japan 3,123 31	Philippines 2,220	
	Mexico 2,985 30	Turkey 1,498	

Online rental service reaches the global market, servicing 108 countries. Asia is the highest sales revenue at 45% with 273 customer s.

Top 10 Countries by customers and Revenue are: India, China, US, Japan, Mexico, Russia, Brazil, Philippines, Turkey and Indonesia.

These countries provide more than 50% of the total revenue. India, China and United States are the leaders, making them..

TOP 10 REVENUE GENERATING COUNTRIES

Country	Total no of Custom..	
India	60	\$6,034.78
China	53	\$5,251.03
United S..	36	\$3,685.31
Japan	31	\$3,122.51
Mexico	30	\$2,984.82
Brazil	28	\$2,919.19
Russian ..	28	\$2,765.63
Philippin..	20	\$2,219.70
Turkey	15	\$1,498.49
Indonesia	14	\$1,352.69



[Tableau link of this graph.](#)

ANALYSIS

DATA OVERVIEW

 1000 Movies	 Film length	 Rental Rate	 Replacement Cost	 Rental Duration
MAXIMUM	185 MIN	4.99\$	29.99\$	7
MINIMUM	46 MIN	0.99\$	9.99\$	3
AVERAGE	115.7 MIN	2.98\$	19.984\$	4.98
COUNT	1000	1000	1000	1000

 109 COUNTRIES 600 CITIES 16044 RENTALS 599 CUSTOMERS 2 STORES AVAILABLE MOVIE
LANGUAGE: ENGLISH 1000 FILMS 6132.04 \$ REVENUE PG 13 RATING IS THE HIGHEST

The top highest-earning 20 films contributed 6% of the revenue share while the bottom 20 fell under 0.5%.

20 Highest Earning
Films

Telegraph Voyage	\$215.75
Zorro Ark	\$199.72
Wife Turn	\$198.73
Innocent Usual	\$191.74
Hustler Party	\$190.78
Saturday Lambs	\$190.74
Titans Jerk	\$186.73
Harry Idaho	\$177.73
Torque Bound	\$169.76
Dogma Family	\$168.72
Pelican Comforts	\$165.77
Goodfellas Salute	\$164.75

20 Lowest Earning
Films

Duffel Apocalypse	\$5.94
Oklahoma Jumanji	\$5.94
Texas Watch	\$5.94
Freedom Cleopatra	\$5.95
Rebel Airport	\$6.93
Young Language	\$6.93
Cruelty Unforgiven	\$6.94
Treatment Jekyll	\$6.94
Lights Deer	\$7.93
Japanese Run	\$7.94
Stallion Sundance	\$7.94
Ghostbusters Elf	\$8.93

Rockbuster Stealth reached
\$61,312 in total revenue with
958 films across the regions.

KEY LEARNINGS

1

Telegraph Voyage contributed the most revenue gain while Duffel Apocalypse contributed the least.

2

The average rental duration across all videos is 5 days, indicating consistent consumer engagement.

3

Sports, sci-fi, and animation genres are the most popular. However, different regions have different preferences.

4

Top customers are coming from all around the globe.

5

India, China, and the United States lead in both customer numbers and revenue share, driving significant market impact.

RECOMMENDATIONS

Region/Country Specific

Increase marketing efforts and tailor promotional campaigns specifically for **high-impact markets: the broader Asian region, and countries like India, China, and the USA.**

Genre Specific

Capitalize on regional differences in genre preferences by promoting **Animation and Sports heavily in Asia, and Sci-Fi in the USA and Europe.**

Customer Specific

Develop a global loyalty program tailored to **reward top customers. Implement flexible rental options** to cater to the average 5-day rental duration.

LINKS & DELIVERABLES

[GitHub Repository](#)[Final Presentation](#)[Data Dictionary](#)

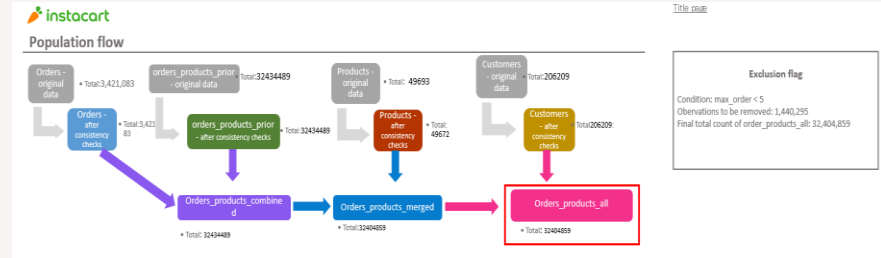
*Please click links above to view relevant project work



Marketing Strategy for an Online Grocery Store

View Project in [GitHub](#)

Instacart, a leading online grocery store, enables customers to order groceries through an app. The company aimed to boost sales by adopting a more strategic approach to targeting customers and improving segmentation through the analysis of historical data.



Expectation

As being Instacart's data analyst, I am tasked with enhancing sales insights through initial data and exploratory analysis. Focusing on customer segmentation for targeted marketing strategies, while ensuring personalized campaigns align with customer profiles and boost product sales.

Goal: Analyze customer purchasing behaviors to create a customer segmented classification model for targeted marketing strategies and boosting sales revenue.

KEY QUESTIONS

1

What are the busiest days of the week and hours of the day?

2

At what times of the day do people tend to spend the most money?

3

Is there a connection between age and family status in terms of ordering habits?

4

Which types of products are most popular?

5

What are the characteristics and spending habits of different customer profiles

DATA

Data Sets:


- Customers: Analyzed for purchasing patterns and loyalty.
- Orders: Studied to determine busy times and spending habits.
- Products: Categorized to understand popularity and sales impact.
- Departments: Analyzed for sales volume per department.

Data Citations:

- "[The Instacart Online Grocery Shopping Dataset 2017](#)", Accessed via Kaggle.
- "[Customers Data Set](#)", Provided by CareerFoundry.

SKILLS APPLIED

- Data wrangling
- Data merging
- Deriving variables
- Grouping data
- Aggregating data
- Reporting in Excel
- Population flows
- Visualising results using Python



Access to full
Python code

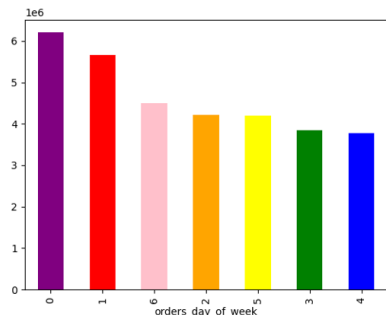
CHALLENGES

Handling the dataset's 30 million records posed significant challenges in cleaning, standardizing, and formatting for analysis. Issues such as memory shortages were overcome through efficient coding practices to ensure meaningful data transformation.

ANALYSIS

ORDERING PATTERNS

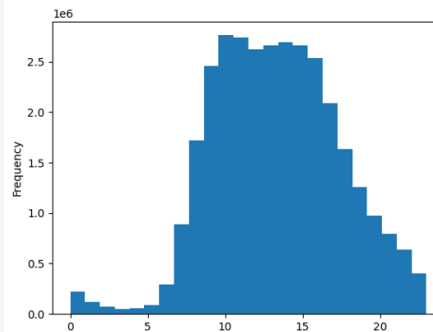
Number of Orders by Day of Week



#	Day
0	Saturday
1	Sunday
2	Monday
3	Tuesday
4	Wednesday
5	Thursday
6	Friday

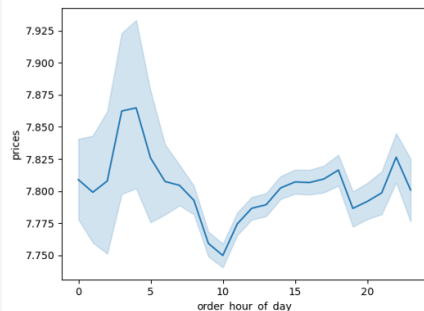
The busiest days of the week are Saturday and Sunday between the hours of 8 am to 4 pm. The least busy day is Wednesday between the hours of 2 am to 5 am.

Prices by hour of day



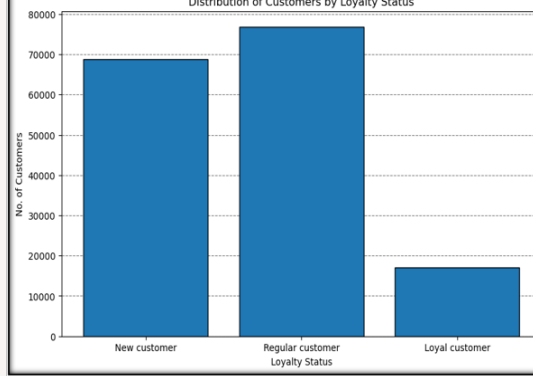
- There is a clear peak in order activity during the mid-morning to early afternoon hours (9 AM to 3 PM), with the highest number of orders occurring at 10 AM.
 - Order activity starts to increase significantly from 7 AM, reaching its peak between 9 AM and 3 PM.
 - After 3 PM, the number of orders gradually decreases, with a notable decline from 5 PM onwards.
 - The lowest order volumes are observed in the early morning hours (midnight to 6 AM), with the least activity around 3 AM.
- This distribution suggests that Instacart's customers are most active during typical waking hours.

Prices by Order hour of day placed



People spend the most money during the hours of 8am to 4pm. 9 am is when people spend the most money. Therefore, Instacart should utilize those times to advertise specific products.

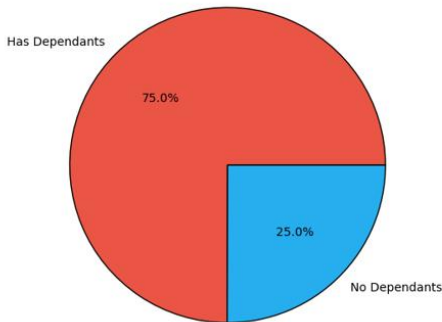
Distribution of Customers by Loyalty Status



In terms of distribution, there are far more new customers than there are regular or loyal customers. This either indicates recent growth or many one-time customers.

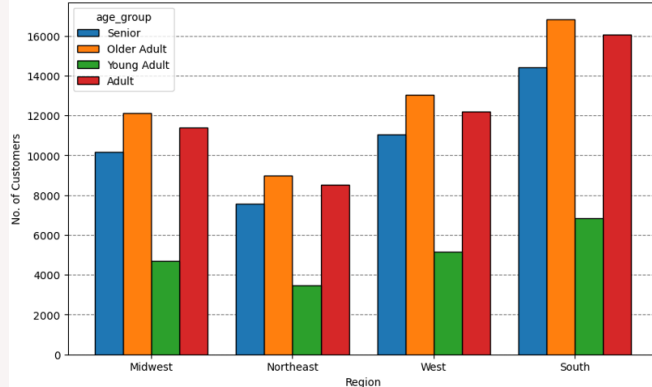
Most of Instacart frequent, non-frequent, and regular customers are married. It is followed by single people, divorced/widowed people and lastly people who live with their parents and siblings.

Instacart Customers with Dependents vs without Dependents



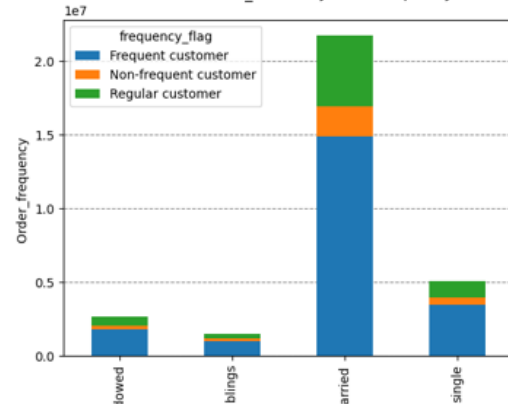
Dependents: This category is based on whether a customer has a dependent(s). Most customers have a dependent(s) (75%).

Distribution of Customer Age Groups by Region

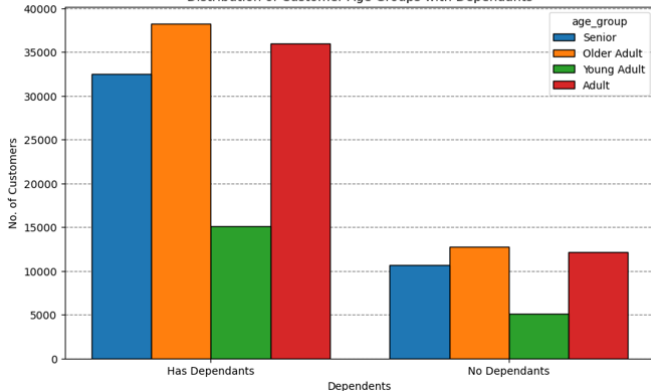


In every region, Instacart customers are mostly Older adults from the graph on the left. Also, the South has the most customers in Adult and older adult category followed by the West, then the Midwest and lastly the Northeast.

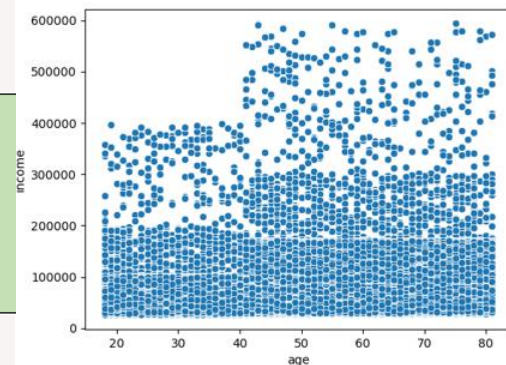
Distribution of Marital_status by order frequency chart



Distribution of Customer Age Groups with Dependents



There is a positive correlation between age and income. As the older you become, the more income you'll earn. The scatterplot showing the relationship between age and spending power (income) of instacart customers. Based on the scatterplot, there is a large concentration of customers across all ages that earns up to 200,000 (USD) and customers who are 40 years and older are earning up to 600,000 (USD). So, customers who are 40 years and older might spend more because of their higher income.



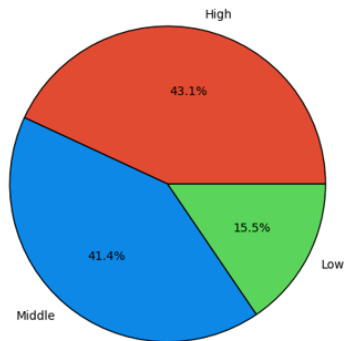
Income: The income groups were based on distributions of household incomes in the U.S. during 2022 . Most of instacart's customers fit into the high and middle income categories.

Order Amount:

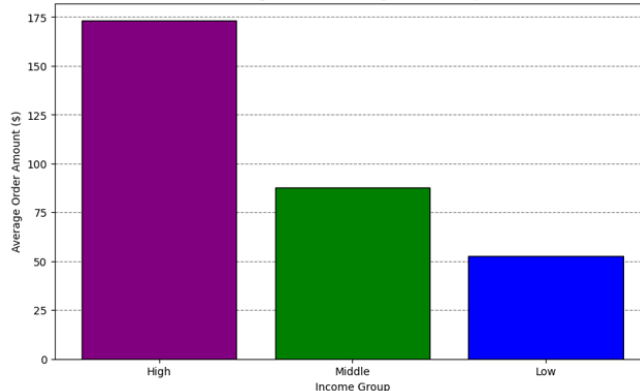
High-income earners spend significantly more per order compared to middle and low-income earners. There is a substantial drop in average order amount from high-income to middle-income and again from middle-income to low-income groups.

High-income and middle-income groups have similar high-frequency ordering habits. Low-income group tends to have a higher percentage of medium and low-frequency orders compared to high-income and middle-income groups.

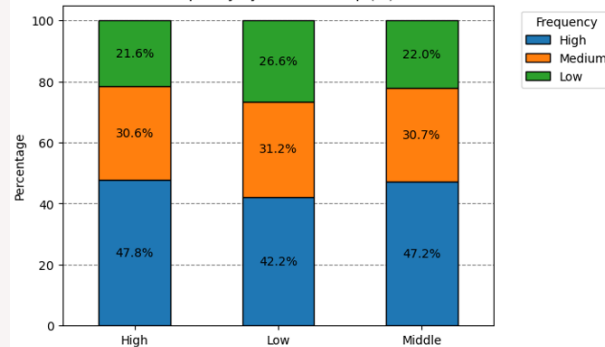
Instacart Distribution of Customers by Income Group



Average Order Amount by Income Group



Frequency by Income Group (%)



If income ≤ 50000 then low . If income ≥ 50000 and income ≤ 100000 then middle . If income ≥ 100000 then high .

Product Preferences:

Produce is the most purchased category across all income groups, indicating a common preference for fresh products.

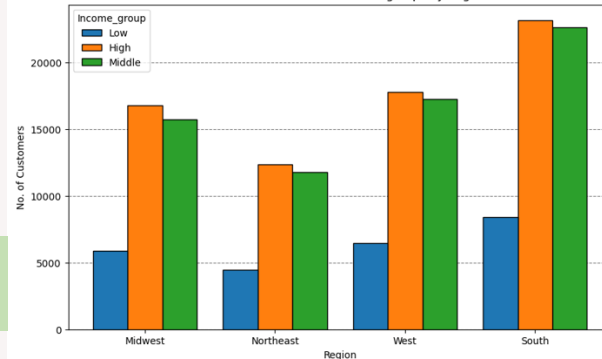
Dairy & Eggs is the second most popular among high and middle earners, while it is third for low earners.

Snacks are more popular among low earners compared to beverages.

Beverages and **Frozen** goods are consistently in the top five for all groups.

The chart on the right shows that the high Income group in all the four region contribute more when compared to low in come group .

Distribution of Customer Income groups by Region



KEY LEARNINGS

1

The busiest days of the week are Saturday and Sunday between the hours of 8 am to 4 pm. The least busy day is Wednesday between the hours of 2 am to 5 am.

2

People spend the most money during the hours of 8am to 4pm. 9 am is when people spend the most money

3

The majority of the products people order are low-range (<5 USD) and mid-range (5-10 USD) products.

4

Most popular products are fresh food products (produce, meat and sea food, dairy and eggs, deli, bakery).

5

The most significant variation in shopping behavior among customer groups relates to income, particularly in the average price of orders.

RECOMMENDATIONS

Marketing and Sales

Schedule targeted ads during lower order volumes in the early evening hours on weekdays, capitalizing on increased social media usage.

Focus on advertising high-range products early in the morning and late at night, using time-limited promotions to generate urgency.

Customer Profiles

Utilize loyalty data to tailor campaigns, particularly promoting premium products to high-income customers like Gen X and Baby Boomers.

Implement referral programs that reward loyal customers for bringing in new shoppers, focusing on bulk purchase promotions for families.

Further Analysis

Explore variations in the busiest shopping hours and days, and analyze purchasing trends for different product types during specific times.

Examine underperforming high-range products to understand customer purchasing behaviors and preferences better.

LINKS & DELIVERABLES



[GitHub Repository](#)



[Project Report](#)

*Please click links above to view relevant project work



Data analyst portfolio

05 Global Bank

Pig E. Bank is a global bank dedicated to providing exceptional financial services.

Predicting consumer churn rate with a classification model.

OBJECTIVE

Assuming a new role in sales analytics at Pig E. Bank, I'm leading a customer retention project. Using client attributes like age and estimated salary, I'll pinpoint key risk factors leading to client loss, modeling them in a decision tree.

Goal: Use a predictive model to identify and segment banking members with a high likelihood of either exiting the bank or remaining as active or non-active members.

KEY QUESTION

1

What are the key risk-factors in identifying customers who are most likely to churn?

Global Bank

DATA

Data source: Career Foundry

SKILLS APPLIED

- Big data
- Data ethics
- Data mining
- Predictive analysis
- Time series analysis and forecasting
- Using GitHub

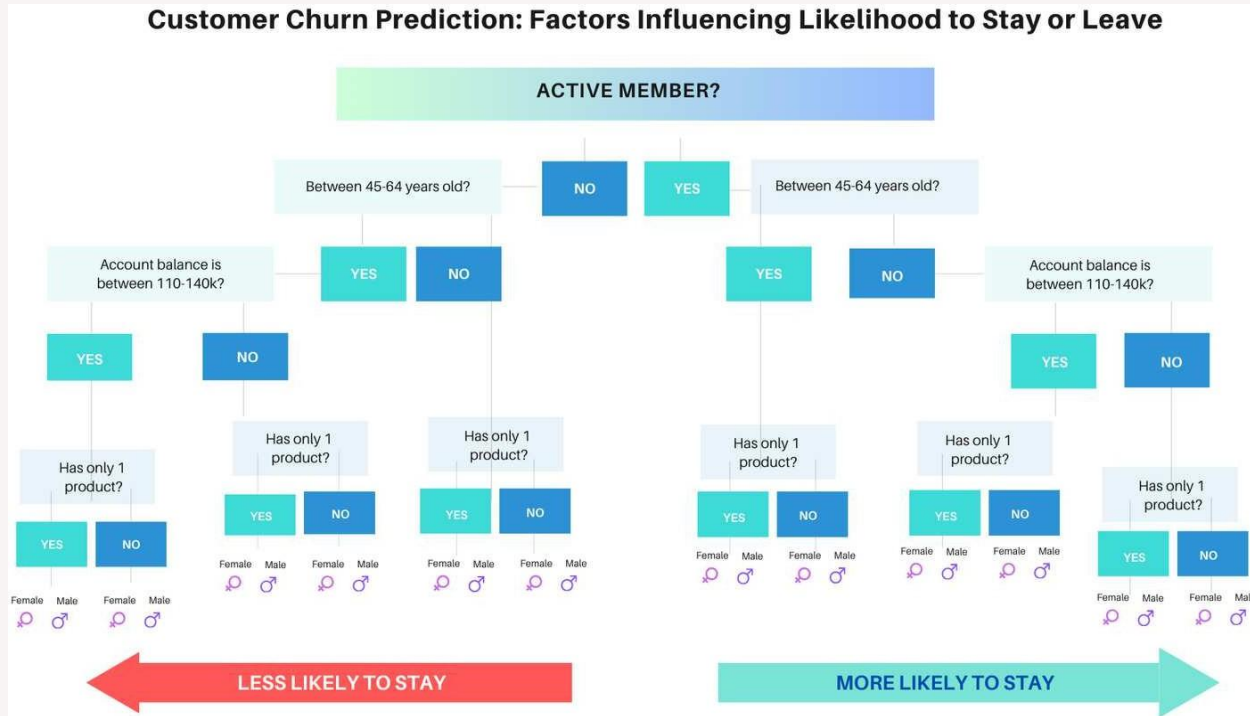
	Current Customers			Former Customers		
	Mean	Max	Min	Mean	Max	Min
Credit Score	652	850	411	637	850	376
Age	38	82	18	45	69	22
Tenure	5	10	0	5	10	0
Balance	\$74 921,00	\$197,04	\$0,00	\$90 239,22	\$213 146,20	\$0,00
Number of Products	2	3	1	1	4	1
Estimated Salary	\$98 837,57	\$199 661,50	\$371,00	\$97 155,20	\$199 725,39	\$417,41

		All Customers	Current Customers	Former Customers
Age Group	18-24	3.63%	4.32%	0.98%
	25-34	30.78%	35.71%	11.76%
	35-44	41.98%	43.58%	35.78%
	45-54	14.53%	9.66%	33.33%
	55-64	6.36%	3.94%	15.69%
	65-74	2.22%	2.16%	2.45%
	75+	0.50%	0.64%	
Gender	Female	46.62%	43.33%	59.31%
	Male	53.28%	56.54%	40.69%
	Unknown	0.10%	0.13%	
Account Balance	0	35.22%	37.23%	27.45%
	30,001-40,000	0.20%	0.25%	1.47%
	40,001-50,000	0.71%	0.51%	1.47%
	50,001-60,000	1.21%	1.40%	0.49%
	60,001-70,000	1.82%	1.91%	1.47%
	70,001-80,000	2.93%	3.30%	1.47%
	80,001-90,000	3.83%	4.32%	1.96%
	90,001-100,000	4.94%	4.83%	5.39%
	100,001-110,000	7.16%	6.99%	7.84%
	110,001-120,000	8.78%	7.50%	13.73%
	120,001-130,000	9.38%	9.02%	10.78%
	130,001-140,000	8.07%	7.37%	10.78%
Number of Products	140,001-150,000	4.84%	4.96%	4.41%
	150,000+	10.90%	10.42%	12.75%
	1	51.46%	46.76%	69.61%
	2	45.01%	52.60%	15.69%
	3	3.33%	0.64%	13.73%
Is Active Member	4	0.20%		0.98%
	No	49.24%	43.84%	70.10%
	Yes	50.76%	56.16%	29.90%



ANALYSIS

Data Acquisition and Cleanup Client Segmentation: Divided the data into two groups for analysis: former clients and current clients. Trend Analysis: Conducted a detailed examination of the former clients' data to detect prevalent patterns of behavior. Model Development: Constructed a decision tree to systematically understand and predict the determinants of customer churn.



KEY INSIGHTS

1

Being an inactive member seems to be a major contributing factor in leaving the bank.

2

A higher proportion of the people who left the bank are above age 45.

3

A larger proportion of the former customers have higher account balance (between 100k-140k and also balances more than 150k).

4

Majority of the former customers held only one product.

5

More females are among former customers while current customers are predominantly male.

RECOMMENDATIONS

Activity

Increase customer engagement through loyalty programs, personalized offers, and regular communication.

Age

Develop tailored financial products and services that cater specifically to the needs of people above 45.

Account Balance

Enhance personalized financial advisory services tailored for the financial needs of people with higher balances.

Products

Encourage product diversification among customers.

Gender

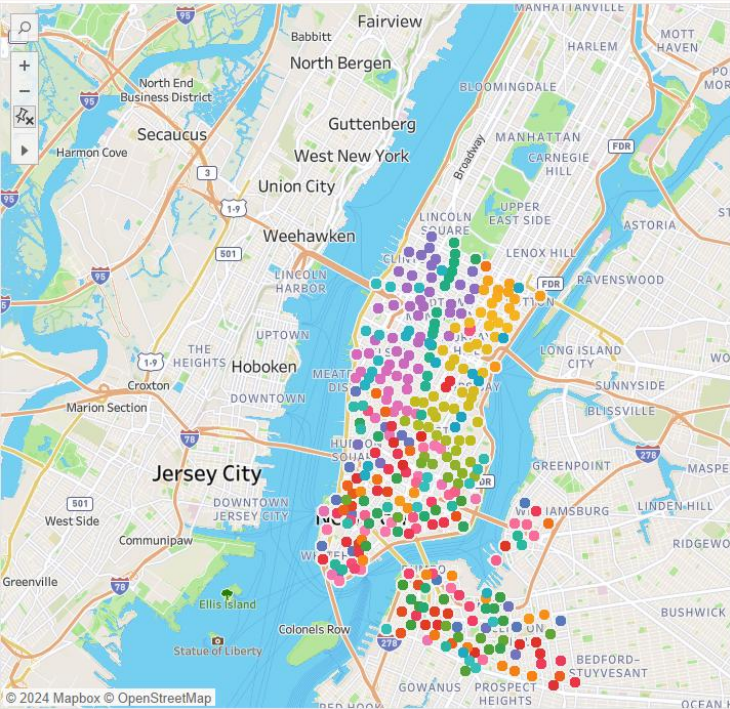
Understand the specific needs and preferences of female customers.

LINKS & DELIVERABLES

[Project Report](#)

*Please click links above to view relevant project work

06 New York Citibike Analysis



"Data-Driven Optimization: Analyzing User Behavior and Station Demand for Strategic Insights"

Objective: As a Citi Bike data analyst, my role is to examine station locations, customer demographics, and user behavior patterns to uncover insights that drive station demand. By analyzing trends such as peak usage times, geographic preferences, and customer needs, I aim to provide actionable recommendations to optimize station placement, improve service efficiency, and enhance the overall user experience. My work ensures that Citi Bike remains a convenient, data-driven solution aligned with the needs of its diverse customer base.

New York Citibike Analysis

KEY QUESTIONS

1

What day is the busiest day for Citi Bike?

2

What time of day do people use Citi Bike the most?

3

Does the day of the week change this outcome?

4

What age group uses Citi Bike the most/least?

5


What are riders' habits? Which stations are most busy? least busy?

6

Does trip duration increase or decrease based on various factors? Time of day? What day of the week? Age?

DATA

Data Source: [New York Citibike Dats Set](#)



**Access to
full Python
code**

SKILLS APPLIED

- Data Ethics
- Big Data
- Data mining
- Predictive Analysis
- Linear Regression
- Cluster Analysis Time Series and Forecasting
- Designing and building a Dashboard
- Relationships and patterns spotting

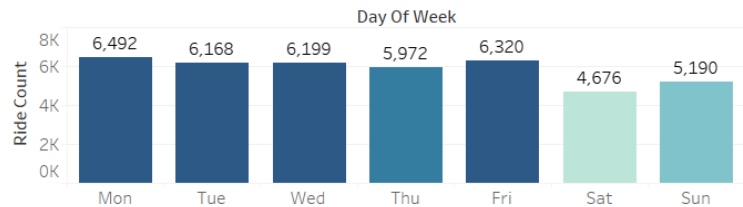
CHALLENGES

Identifying meaningful patterns in user habits was challenging, requiring careful analysis to separate daily fluctuations from long-term trends. Performing cluster analysis on non-numeric data, like hourly usage, required creative solutions to make time-based insights useful. Finally, designing a clear and effective Tableau dashboard was essential for balanced communication of key findings..

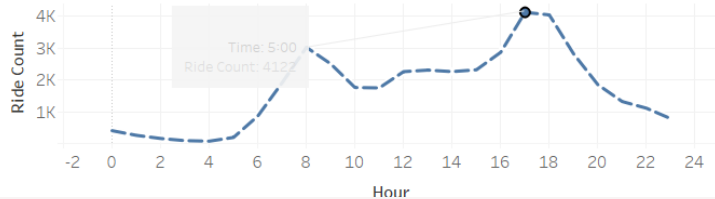
ANALYSIS

Exploratory Data Analysis

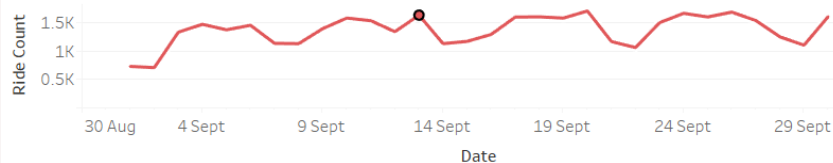
Busiest Day of the Week



Busiest Hour of Day



Ride Counts in September



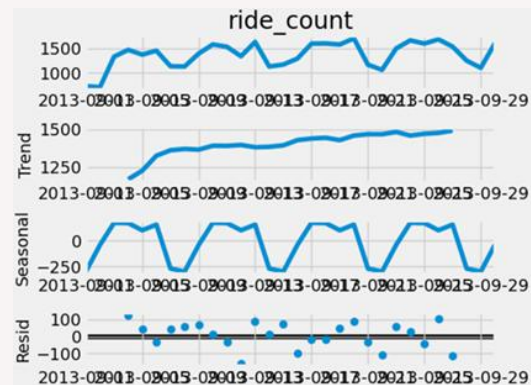
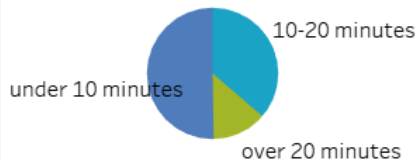
Busiest time: Rides peak at 5:00 PM, coinciding with the evening commute, followed by a secondary peak at 8:00 AM, reflecting morning commuter usage.

Busiest day: Monday sees the highest ride count, followed closely by Friday, indicating strong weekday usage.

Least busy: Saturday and Sunday have the fewest rides, suggesting Citi Bike is less popular for leisure or non-commuting purposes.

Weekday trend: Ride counts are consistently high on weekdays, especially during commuting hours, underscoring the system's role as a transportation option for daily work commutes.

Monthly pattern: Throughout September, ride counts show a steady increase on weekdays, highlighting a regular commuter habit.

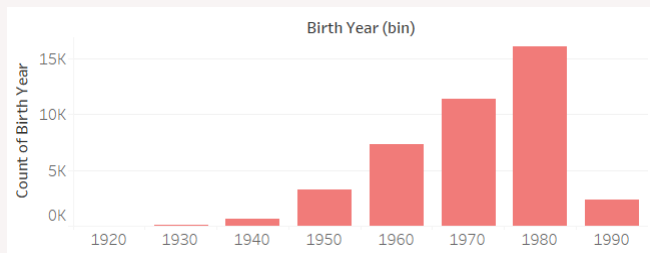


New York Citibike Analysis

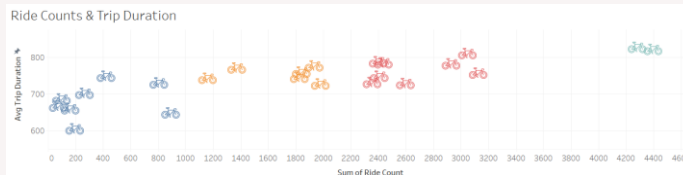
ANALYSIS

Access to
visualizations

Relational Analysis



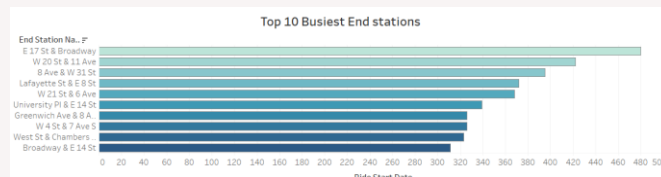
Cluster and Spatial analysis



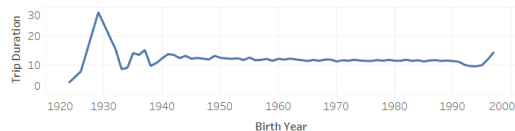
Where do most Citi Bike Rides take place?

Busiest Start Station is *Pershing Square N.* with 451 rides in September
Busiest End Station is *E 17th St & Broadway* with 480 rides in September

The busiest station on Monday, Tuesday, Wednesday & Friday is Pershing Square N.
Thursday is Lafayette St
Saturday is E 17th & Broadway
Sunday is E 8 St 111 8 Ave & W 31 St



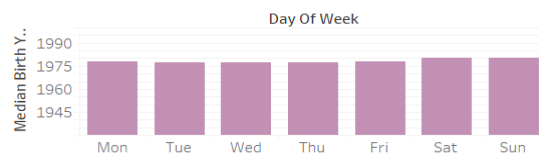
Average Trip duration and Age of Rider



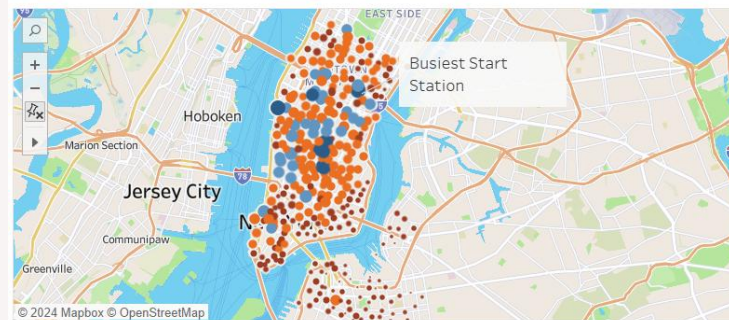
How Does Age Play a Role in Citi Bike Usage ?

Most Citi Bike users are middle-aged, born in the 1980s. Age minimally impacts trip duration, as younger riders show only a slight, statistically insignificant decrease in ride length. However, age influences ride timing—younger riders favor weekends, especially Saturdays, likely for leisure, and are more active during early morning and late hours. In contrast, middle-aged users exhibit steadier patterns, reflecting commuting habit

Median Birth Year of Riders



Start station



KEY INSIGHTS

1

Citi Bike's busiest days are Monday and Friday
Citi Bike's least busy days are Saturday and Sunday.

2

Rides peak at 5:00 PM during the evening commute, with a secondary peak at 8:00 AM for morning commuters.

3

Most riders are born in the 1980s, indicating that the largest user group consists of middle-aged individuals, roughly 30-50 years old.

4

Younger riders show a slight decrease in trip duration, prefer weekends, especially Saturdays, and are more active early morning and late at night.

5

Our Tableau maps show Pershing Square N. as the busiest start station with 451 rides, and E 17th St & Broadway as the busiest end station with 480 rides, varying by day.

RECOMMENDATIONS

Encourage Off-Peak Commuting

Target the commuting demographic with incentives for off-peak riding, such as a 10% discount for six-day riders and free rides for exceeding a set duration.

Local Partnerships

Collaborate with local businesses to provide discounts for riders who show their Citi Bike receipt, promoting both biking and local commerce.

Group Discounts

Encourage group rides by offering discounts for friends or family riding together, especially on weekends or during events.


Optimize Bike Distribution

Ensure ample bikes at busy stations and minimize unused bikes at less busy ones to enhance service efficiency and user satisfaction.

Promote Weekend Usage

Create targeted incentives for longer weekend rides to boost participation and revenue..

LINKS & DELIVERABLES

 [Tableau Story Board](#)



[GitHub Repository](#)

*Please click links above to view relevant project work



07 ClimateWins Weather Data

Help ClimateWins choose an appropriate machine learning algorithm to predict climate change

Objective: As a data analyst for ClimateWins, a European nonprofit organization dedicated to combating climate change, I'll lead the charge in integrating machine learning to forecast climate consequences, empowering ClimateWins to address extreme weather events with cutting-edge algorithms to derive a data-driven strategy.

Goal: Utilize machine learning algorithms with Python to educate ClimateWins on choosing the most optimized algorithm to predict European extreme weather conditions.

KEY QUESTIONS

1

Can machine learning predict significant temperature increases in Europe over the next decade?

2

Which machine learning models are most effective for forecasting extreme weather events like heatwaves and storms?

3

What patterns or correlations can machine learning identify between heatwaves and variables like air quality, urbanization, or economic impact?

4

What age group uses Citi Bike the most/least?

5

What are riders' habits? Which stations are most busy? least busy?

6

Does trip duration increase or decrease based on various factors? Time of day? What day of the week? Age?

DATA

Data Source:

This data was collected by the [European Climate Assessment & Data Set Project](#)

[Data Set Link.](#)

DATA BIAS

- **Collection Bias.**-Data is from 18 weather stations across Europe, while over 26,321 stations exist
- **Temporal Bias** - Data spans from the late 1800s to 2022. Older records may no longer represent current conditions, potentially misleading machine learning models.

Access to
full Python
code

SKILLS APPLIED

- History and tools of ML
- Ethics & direction of ML programs
- Optimization in relation to problem solving.
- Supervised ML algorithms.
- Presenting ML results.

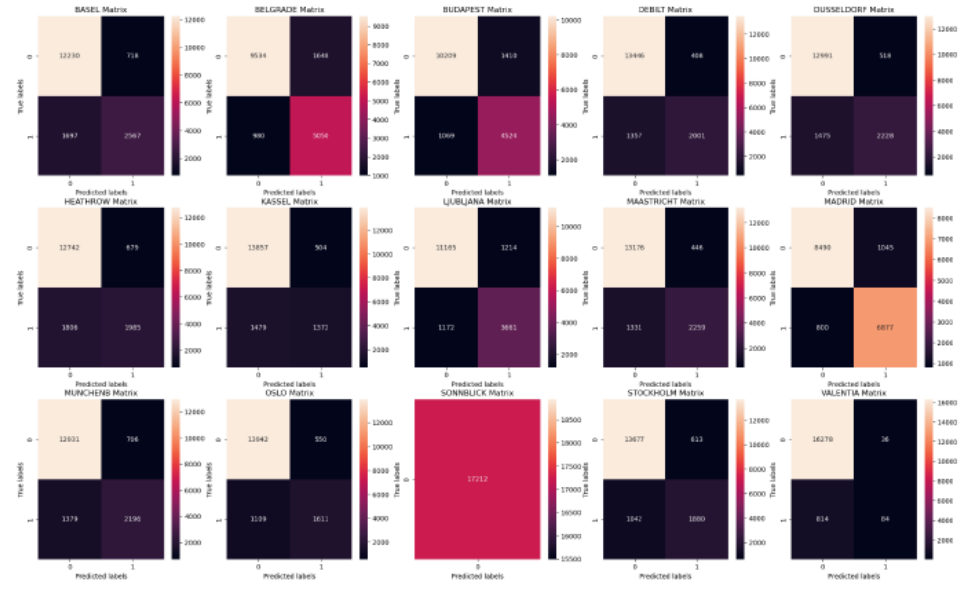
ANALYSIS

What methods were utilized?

- **Decision Tree:** Accuracy not assessed due to the need for pruning.
 - **ANN:**
 - **Test Accuracy:** 46% - 49%
 - **Training Accuracy:** 45% - 51%
 - Moderate performance, with a slight risk of overfitting.
 - **KNN:**
 - **Test Accuracy:** 88%
 - Best performing model, indicating it is well-suited for this dataset.
- Conclusion:**
- **KNN** is the top performer with 88% accuracy.
 - **ANN** shows stable but lower accuracy (46% - 49%).
 - **Decision Tree** requires further adjustments before accuracy can be determined.

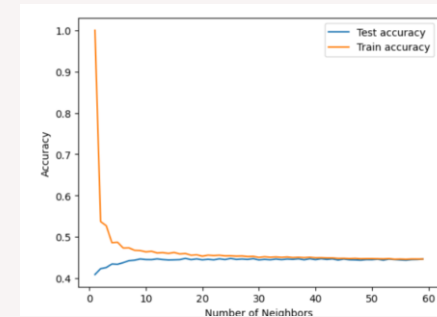
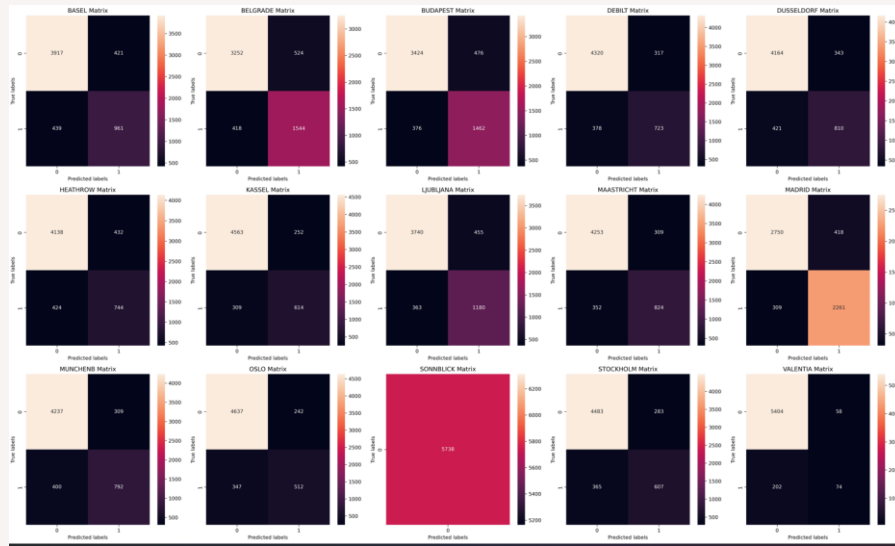
Displayed on the right, a confusion matrix scores chart that was used to score The accuracy for all of our optimization & predictive models (gradient descent, KNN, decision tree, ANN). Of all models, VALENTIA had the highest accuracy scores.

Training Accuracy:



Weather Station	Accurate Predictions	False Positive	False Negative	Accuracy Rate
Basel	3917	421	439	85%
Belgrade	3252	1544	524	84%
Budapest	3424	1462	476	85%
Debilt	4320	723	317	88%
Dusseldorf	4164	810	421	87%
Heathrow	4138	744	432	85%
Kassel	4563	614	252	90%
Ljubljana	3740	1180	455	86%
Maastricht	4253	824	309	88%
Madrid	2750	2261	418	87%
Munchenb	4237	792	309	88%
Oslo	4637	512	242	90%
Sonnblick	5738	0	0	100%
Stockholm	4483	607	283	89%
Valentia	5404	74	58	96%
			Average	88%

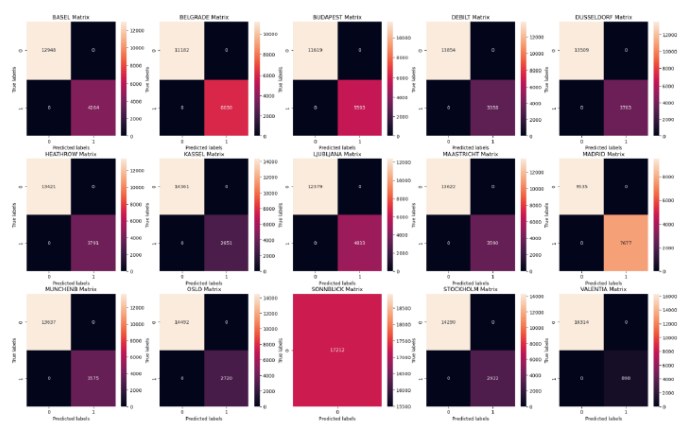
- Promising Weather Stations: Basel, Belgrade, Budapest, Ljubljana, and Madrid show promising weather predictions, as accurate predictions are higher than false positives.
- Sonnblick: Achieved 100% accuracy for unpleasant weather, inflating the overall average to 88%. This may suggest overfitting due to imbalanced data or noise in the model.
- Valentia: Delivered the most reliable accuracy rate of 96%, exceeding the 88% average.
- Belgrade & Budapest: Recorded the lowest accuracy (84% and 85%), falling below the average accuracy of 88%.



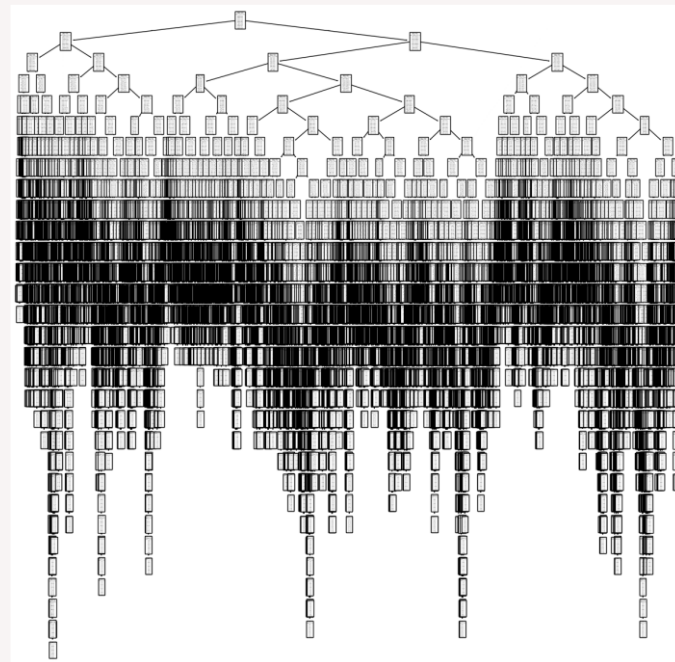
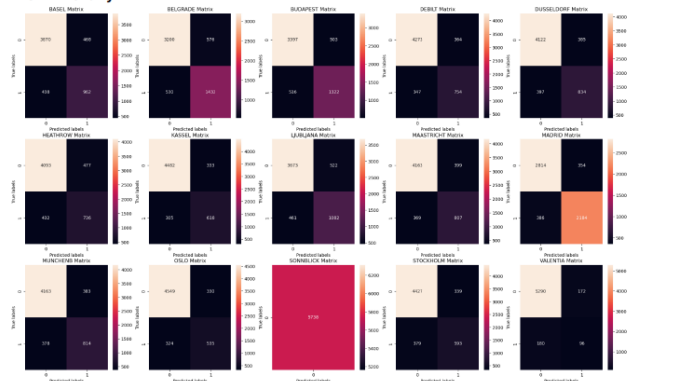
Decision Tree

ANALYSIS

Training Accuracy:



Test Accuracy:



```
#What is the testing accuracy score? Using the cross-validation method
y_pred = weather_dt.predict(X_test)
print('Test accuracy score: ', accuracy_score(y_test, y_pred))
multilabel_confusion_matrix(y_test, y_pred)
```

Test accuracy score: 0.47385848727779717

ANALYSIS

What are the overall scores for each model? Which models performed at the highest accuracy?

```
#Create the ANN
#hidden_layer_sizes has up to three layers, each with a number of nodes. So (5, 5)
#and (100, 50, 25) is three hidden layers with 100, 50, and 25 nodes.
mlp = MLPClassifier(hidden_layer_sizes=(5, 5), max_iter=500, tol=0.0001)
#Fit the data to the model
mlp.fit(X_train, y_train)
```

```
MLPClassifier(hidden_layer_sizes=(5, 5), max_iter=500)
```

```
y_pred = mlp.predict(X_train)
print(accuracy_score(y_pred, y_train))
y_pred_test = mlp.predict(X_test)
print(accuracy_score(y_pred_test, y_test))
```

```
0.4654891935858703
```

```
0.4696758452422447
```

Displayed above, is the Artificial Neural Network parameters and accuracy score, by which, yielded the highest accuracy at 45% but still relatively low accuracy.

```
#What is the training accuracy score? Using the cross validation method
y_pred_train = weather_dt.predict(X_train)
print('Train accuracy score: ', cross_val_score(weather_dt, X_train, y_train,
multilabel_confusion_matrix(y_train, y_pred_train))
```

```
Train accuracy score: 0.46153827226214633
```

```
array([[12948, 0],
       [ 0, 4264]],

       [[11182, 0],
       [ 0, 6030]],

       [[11619, 0],
       [ 0, 5593]],

       [[13854, 0],
       [ 0, 3358]],

       [[13509, 0],
       [ 0, 3703]],
```

In the screenshot above, the decision tree tested at an overall accuracy score of 40% for the sample of 15 weather stations and their mean temperatures while the individual scores for each weather station performed at 82-95%, suggesting overfitting.

KEY INSIGHTS

1

The current data was better predicted by the KNN algorithm, considering that it had an average of 88% in the test set.

2

The decision tree needs to be prune for better accuracy.

3

As for the ANN, besides the fact that is unpredictable, it has shown an accuracy of around 46% for the test set, which is a lot lower than the KNN algorithm that.

4

since this project it is still incomplete, I'd consider the KNN algorithm the best option, as this one had 88% of accuracy predicting the climate temperature.

5

Based on the European Climate Assessment dataset:
Minimum Temperature: -34.3°C at Sonnblick on January 13, 1968.
Maximum Temperature: 43.6°C at Belgrade on July 24, 2007.

RECOMMENDATIONS

Model refinement

Further analysis and model refinement are necessary to address overfitting and improve generalization performance..

Decision tree

Further prune the decision tree for better accuracy

Future analysis

Search for other options, such as new algorithms, or a combination of algorithms that were already used, to potentially lead us to patterns that were not defined before

Combined analysis

Combine both supervised and unsupervised methods to create a complete climate model that predicts both specific weather events and wider climate trends

Analyse Valentia

Conduct feature importance analysis to leverage Valentia's strengths..

LINKS & DELIVERABLES



[GitHub Repository](#)



[Project Report](#)



[Proposal-strategy](#)

*Please click links above to view relevant project work

THANK YOU

Shravani Iytha Subramanyam

Click to Connect

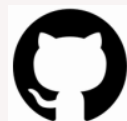
via LinkedIn



via e-mail



via GitHub



via Tableau



via Website

