KEANU GOMES

STUDENT DATA ANALYST PORTFOLIO - 2024

MY SKILLSET

- Detail-oriented data analyst skilled in strategy optimization and organization for informed decision-making.
- Managed and led driver teams, ensuring high-quality service delivery in the Miami-Fort Lauderdale area.
- Managed logs, incidents, and scheduling using MS Excel, fostering effective communication within diverse teams.
- Passionate about applying data ethics to enhance experiences, drive outcomes, and strengthen cybersecurity.
- Holds an IMDb credit for work on a Netflix movie, showcasing creative contributions in addition to analytical skills.
- Proficient in Excel, SQL, Tableau, Python, and Machine Learning, seeking a role emphasizing critical thinking, collaboration, innovation, and social responsibility.

Visit my Github repositories or Tableau storyboards

PROJECTS LIST

1	GAMECO MARKETING ANALYSIS 2017	Analyzing global video game sales.
2	PREPARING FOR INFLUENZA SEASON	Preparing for flu season in the U.S.
3	ROCKBUSTER STEALTH DATA ANALYSIS	Answering business questions for an online video rental company.
4	INSTACART GROCERY BASKET ANALYSIS	Marketing strategy for an online grocery store.
5	PIG E. BANK FINANCIAL SERVICES	Anti-money laundering projects at global bank.
6	YACHT AND BOAT WEBSITE VIEWS ANALYSIS	Utilizing supervised and unsupervised machine learning with python.
7	CLIMATEWINS WEATHER DATA - ML	Leading the charge in integrating machine learning to forecast climate consequences for ClimateWins.

View Project data set citation

Keanu Gomes

01 GAMECO MARKETING ANALYSIS

Analyzing global video game sales

Expectation

It's October 2016, GameCo's executive board is planning the 2017 marketing budget, assuming stable sales across regions. They've tasked me to analyze data, potentially redistributing the budget for maximum ROI. With limited data expertise, they rely on me to guide them through the results effectively.

Skills

- Grouping data
- Summarizing data
- Descriptive analysis
- Visualizing results in Excel
- Presenting results

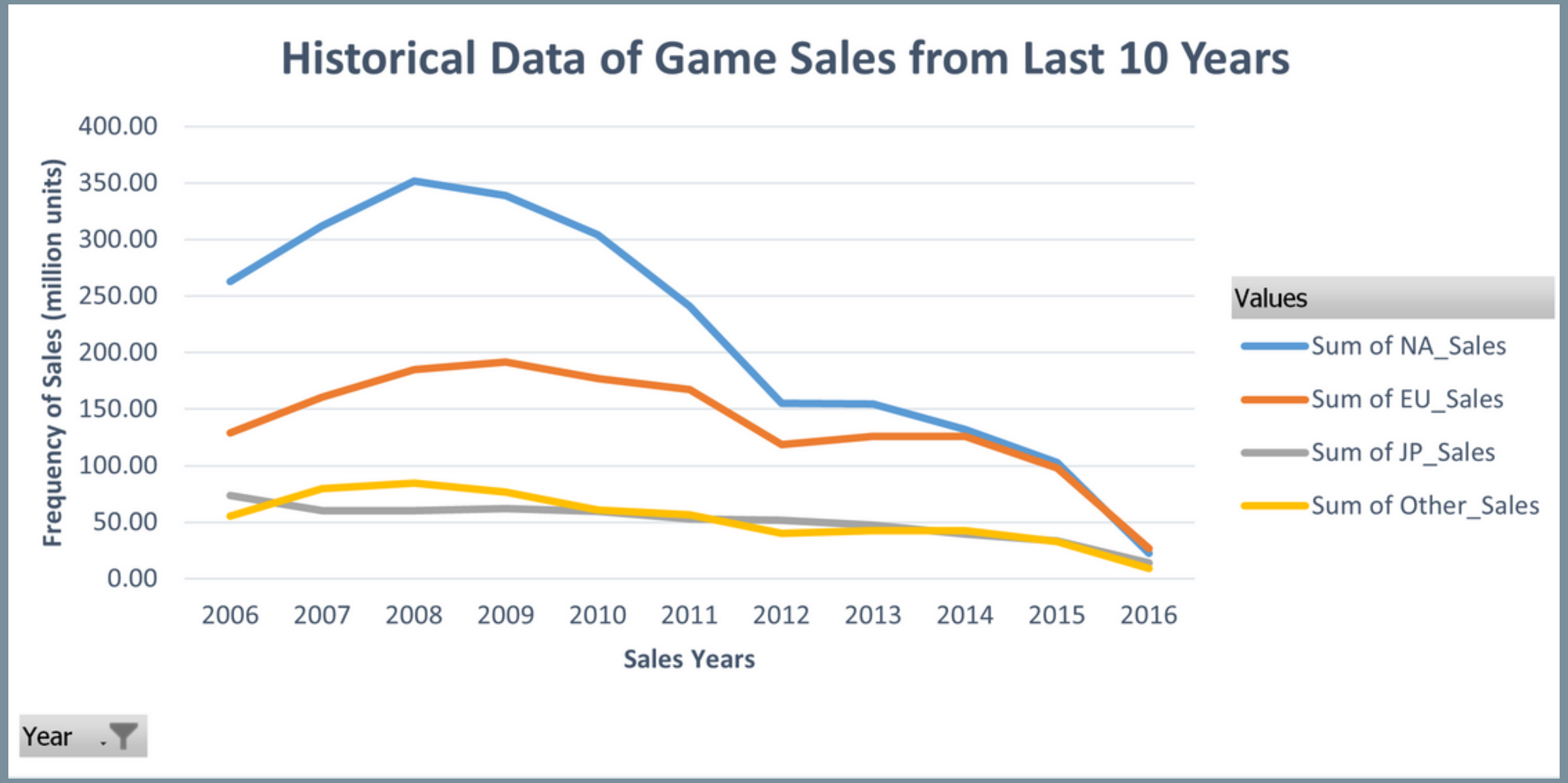
Tools

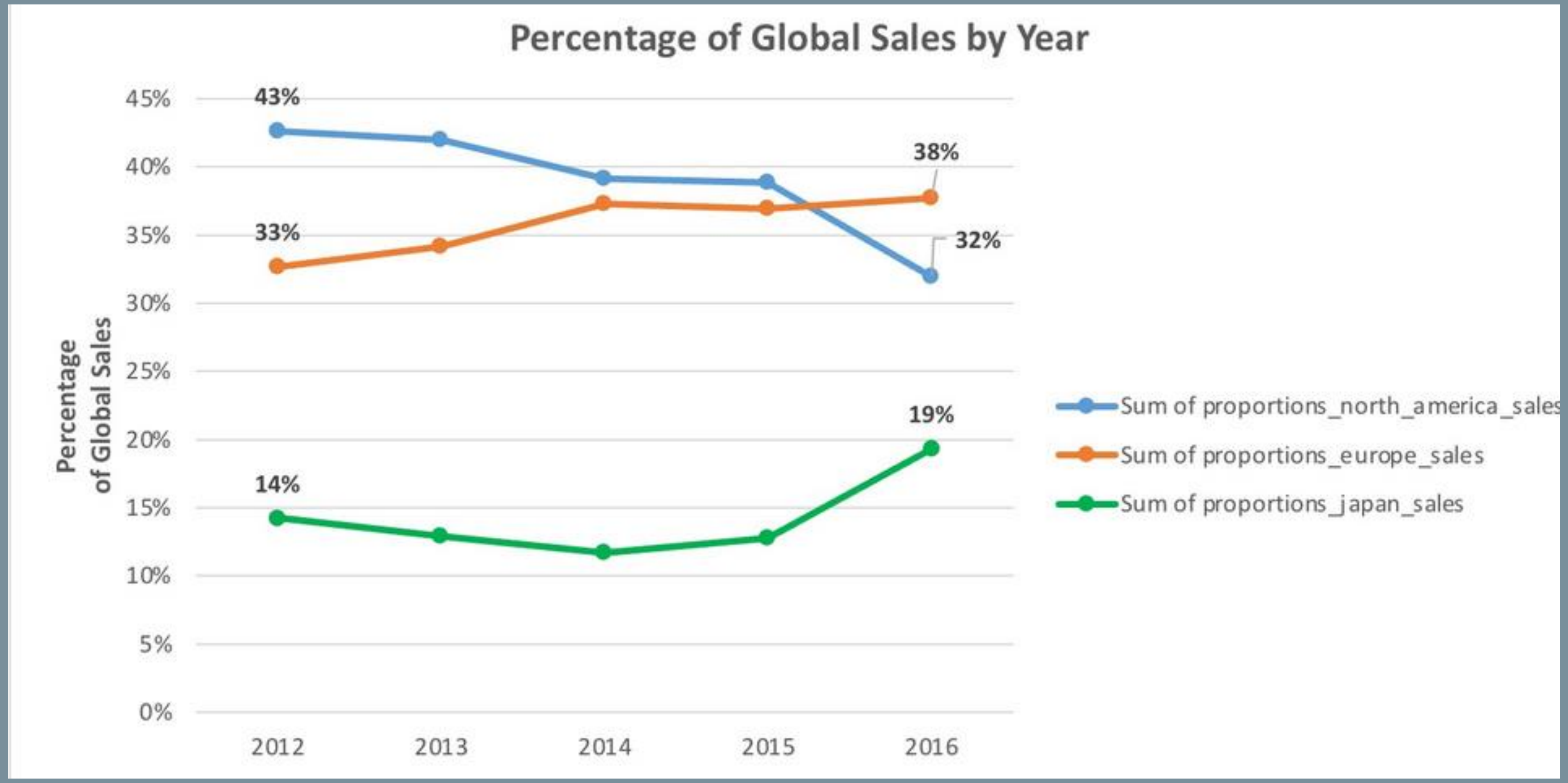
- Microsoft Excel
- Microsoft Powerpoint

Goal: Optimize 2017 marketing budget for maximum ROI. Guide the executive board through results, and facilitate informed budget redistribution decisions.

01 ANALYSIS

What does GameCo's historical data and regional market share value tell us?



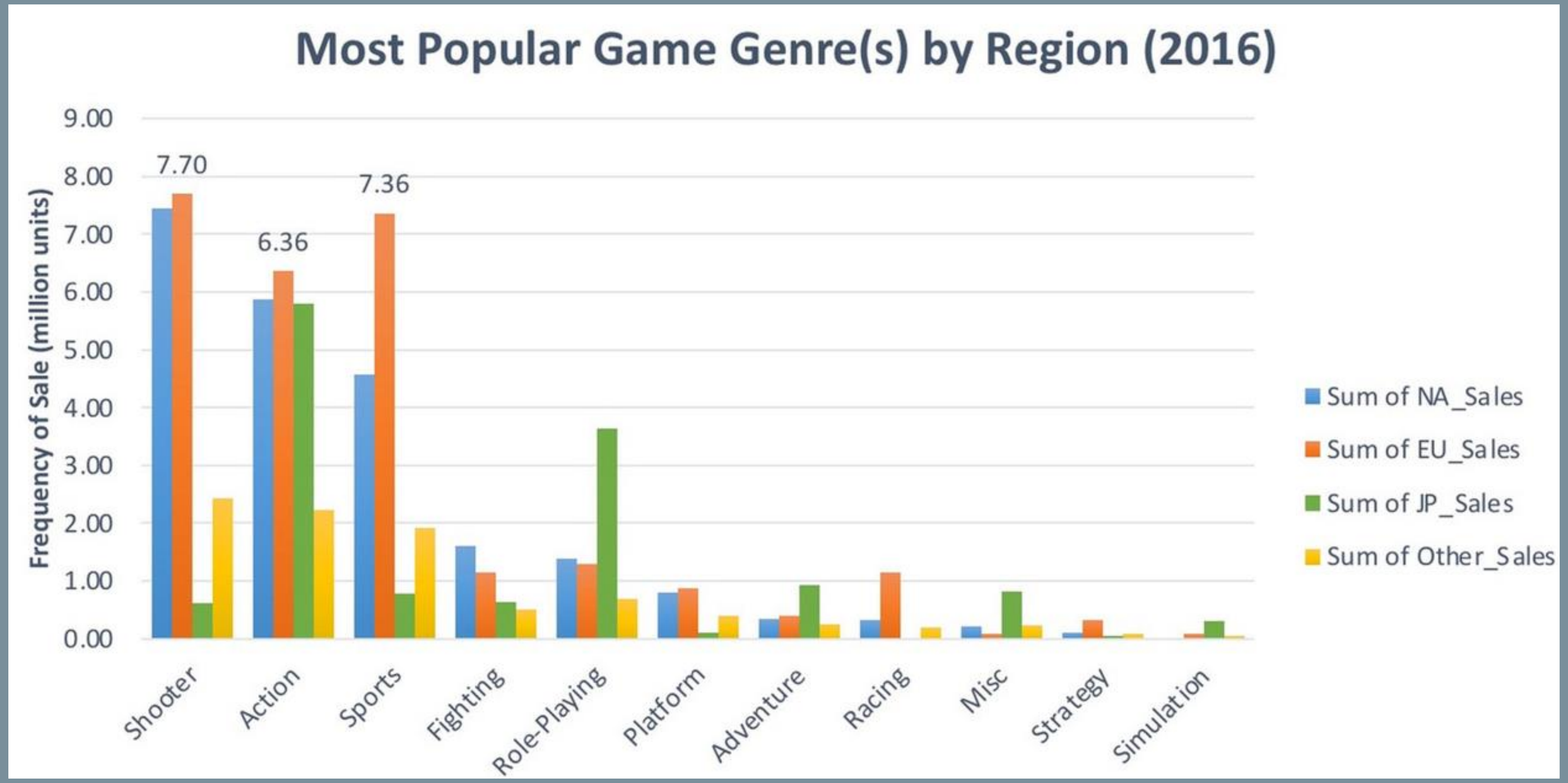


- As historical data reveals, GameCo's current anticipated outlook is being questioned due to the observed negative correlation over the past decade. There is a possibility that sales might dip below their baseline in 2017. Let's delve deeper into the data.

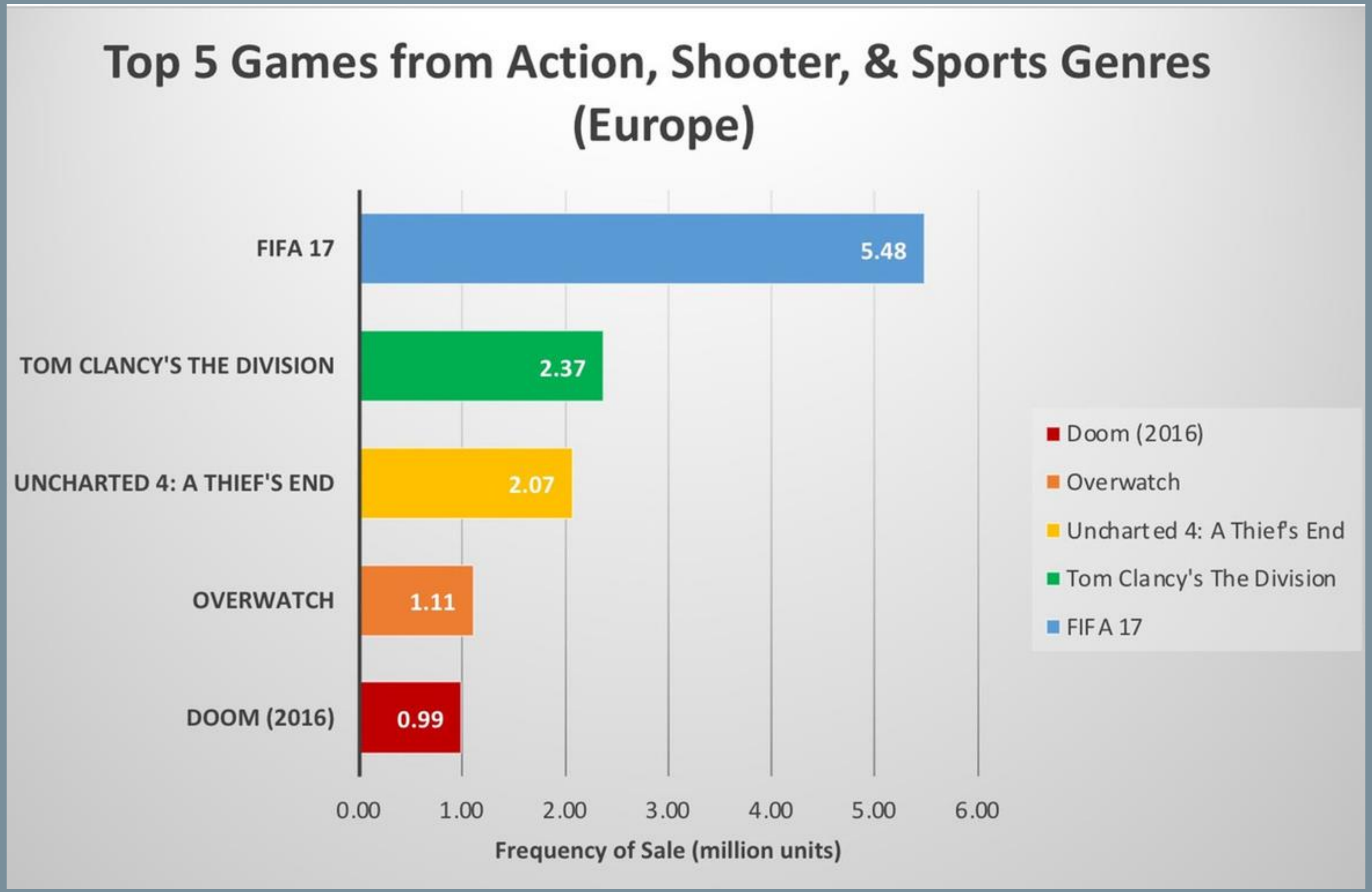
- Upon closer inspection, Japan and Europe exhibit upward trends from 2015-2016, while North America experiences a downward trend. Holding the highest market share value, Japan gained approximately 6%, suggesting a strong performance in GameCo sales for 2017. Europe follows with a 1% gain, and North America concludes with a 7% decline.

01 INSIGHTS

Who are GameCo’s top performers by their region?



- The grouped bar chart illustrates the top-performing game genres of 2016 from GameCo. Across all regions, Shooter, Action, and Sports games dominate, particularly in Europe and North America. However, Japan exhibits a preference for Action and Role-Playing genres.

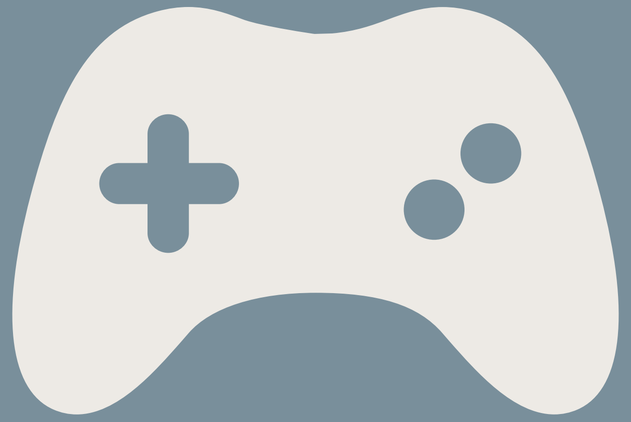


- Given that Europe stands as GameCo's most profitable region, the analysis will reveal that FIFA 17 was the highest-purchased game of 2016.

01 RECOMMENDATIONS

What do these insights tell us?

- Focus on picking up North American sales, while scaling Europe, and Japan’s customer retention.
- Emphasize action, shooter, and sports genres in Europe and North America; prioritize action, role-playing, and adventure genres in Japan.
- Allocate a significant marketing budget to Japan for potential growth in 2017.
- Tailor marketing campaigns based on regional preferences revealed in the data.
- Reassess campaigns from 2012-2016 for insights applicable to 2017 strategies.
- Prioritize showcasing top-performing games, genres, and console platforms to attract a larger customer base and optimize budget allocation.



02 PREPARING FOR INFLUENZA SEASON

Preparing for flu season in the U.S.

View Tableau Storyboard

Expectation

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.

Skills

- Translating business requirements
- Data cleaning, integration, and transformation
- Statistical hypothesis testing
- Visual analysis
- Forecasting
- Storytelling in Tableau
- Presenting results

Tools

- Microsoft Excel
- Tableau

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

02 ANALYSIS

How was the influenza-data prepared for this analysis?

Data Mapping			
CDC_Influenza_Deaths	**Example**	**US_Census_POP**	**Example**
State Code	AL		
Month	January		
Month Code	JAN		
10-year Age Groups	75-84 years		
10-year Age Groups Code	75-84 years		
Deaths	261		
State	Alabama	County/State	Autauga County, Baldwin County, Barbour County, Bibb County, Blount County, Bullock Co County, Calhoun County, Chambers County, Cherokee County, Chilton County, Choctaw C County, Clay County, Cleburne County, Coffee County, Colbert County, Conecuh County, Covington County, Crenshaw County, Cullman County, Dale County, Dallas County, DeKalb Elmore County, Escambia County, Etowah County, Fayette County, Franklin County, Genev Greene County, Hale County, Henry County, Houston County, Jackson County, Jefferson C County, Lauderdale, County, Lawrence County, Lee County, Limestone County, Lowndes C County, Madison County, Marengo County, Marion County, Marshall County, Mobile Coun County, Montgomery County, Morgan County, Perry County, Pickens County, Pike County, County, Russell County, St. Clair County, Shelby County, Sumter County, Talladega County County, Tuscaloosa County, Walker County, Washington County, Wilcox County, Winston
Year	2009	**Year**	2009
		Total Population	4713550
		75-84 years Population	217121
		etc.	

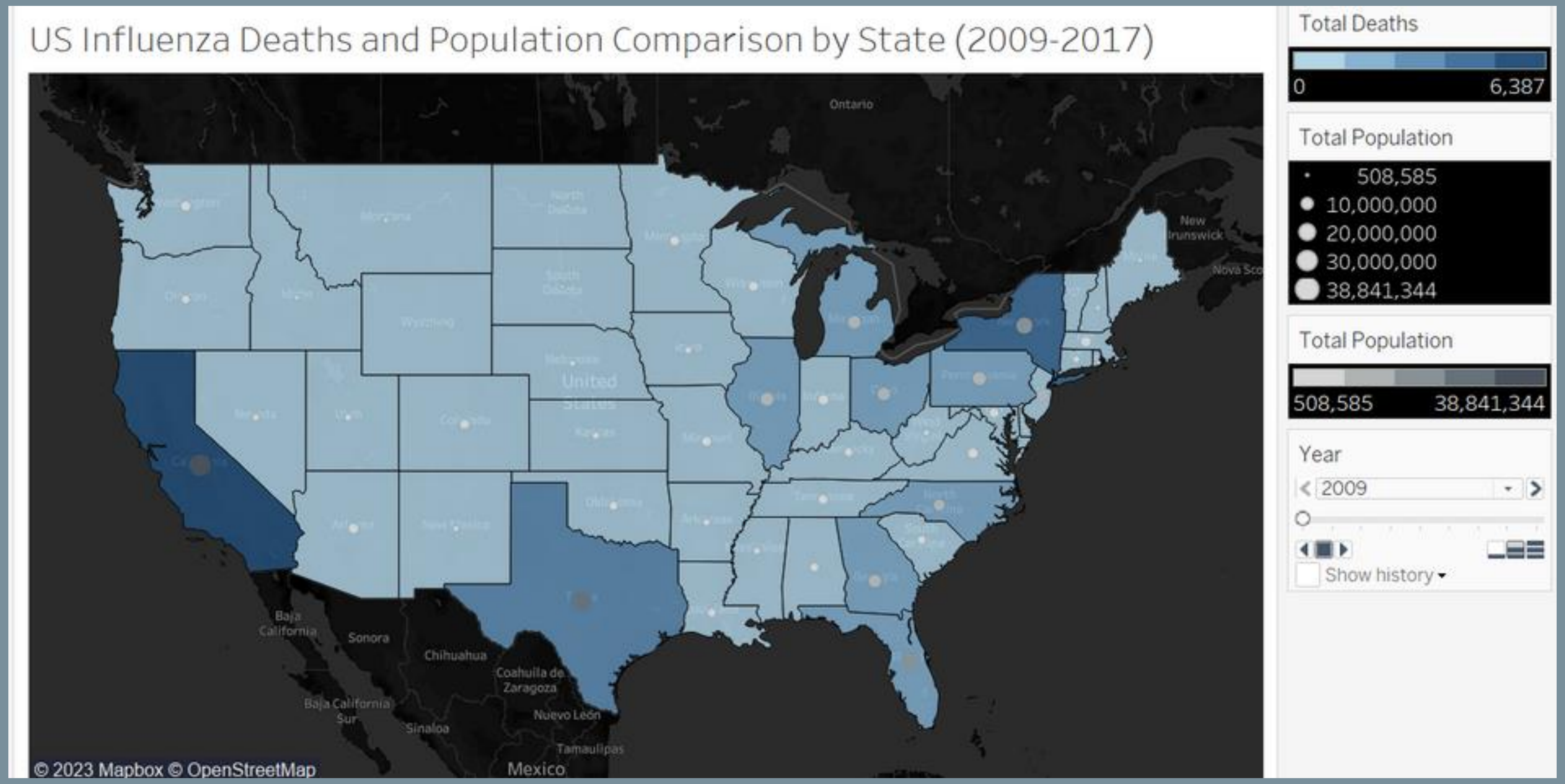
Key Variables/State/Year			US Census Population by 10-year Age Gro	
Combined Key	**State**	**Year**	**ars Population 35-44 years**	**Population 45-54 years**
Montana, 2015	Mc=VLOOKUP(A243, US_Census_POP_Pivot!A245:N712, 8, FALSE)			
Montana, 2016	Montana	2016	117866	132924
Montana, 2017	Montana	2017	107395	114763
Nebraska, 2009	Nebraska	2009	225027	249708
Nebraska, 2010	Nebraska	2010	225907	257586
Nebraska, 2011	Nebraska	2011	226436	259917
Nebraska, 2012	Nebraska	2012	218363	248312
Nebraska, 2013	Nebraska	2013	219688	248599
Nebraska, 2014	Nebraska	2014	223420	251812
Nebraska, 2015	Nebraska	2015	234160	252790
Nebraska, 2016	Nebraska	2016	233898	246750
Nebraska, 2017	Nebraska	2017	223639	226855
Nevada, 2009	Nevada	2009	370813	346271
Nevada, 2010	Nevada	2010	385294	365176
Nevada, 2011	Nevada	2011	386022	369463
Nevada, 2012	Nevada	2012	381116	370640

- The Key Variables for this dataset integration are State & Year.
- The US Census data set will be transformed to align with State level records, hence why all counties are included in the data map

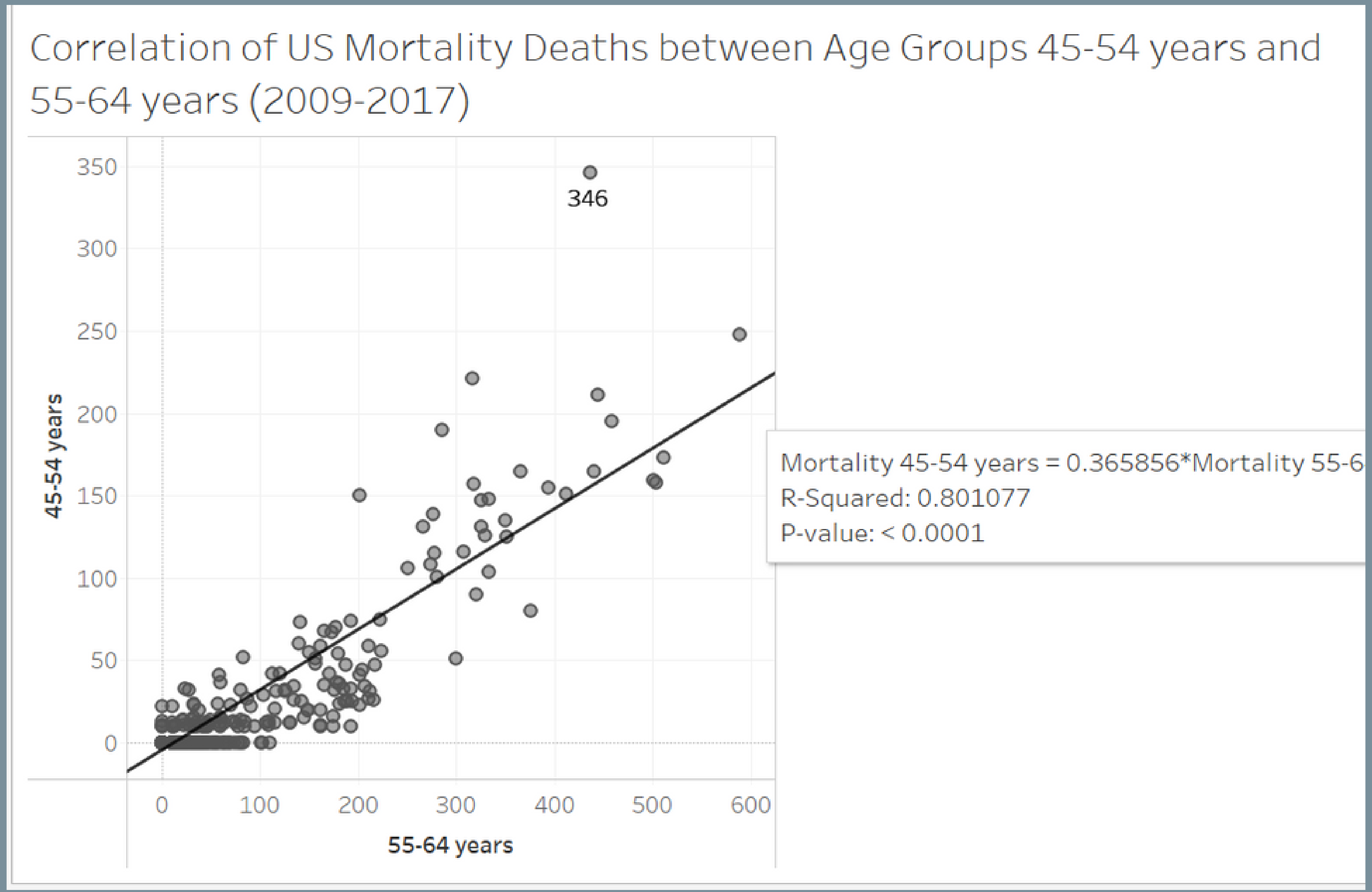
- Added new 'Combined Key' column using concatenate formula (e.g. =C2&", "&B2) in order to use in Pivot table.
- US_Census_Population_PivotTable formatted by 10-year Age Groups. (Tabular form w. CombinedKey)
- These pivot table variables were indexed according to VLOOKUP column order with "combined key" as column 1.

02 INSIGHTS

Who is at risk most to influenza sickness in the U.S. and is Age a factoring influence toward health decline?



- This combo heatmap illustrates that higher density populations tend to have higher frequencies of death due to many influencing factors such as living in closer proximity when compared to rural areas.



- This scatterplot illustrates a hypothesis that advancing age is likely a factor to influencing influenza deaths across the U.S. An R-Squared value close to 1 indicates a strong-positive correlation trending in the CDC data.

02 RECOMMENDATIONS

What do these insights tell us?

- California, New York, Texas, Pennsylvania, Florida are of the highest level of need for medical staffing in 2018 while District of Columbia, Alaska, Vermont, Wyoming, and Delaware are of the lowest need for medical staffing in preparation for 2018.

To note:

1. Flu outbreaks tend to vary in severity and timing across different geographic locations and demographic groups.
2. The severity of flu outbreaks can vary from year to year, and different states may experience. higher or lower levels of flu activity in any given season.
3. The number of medical staff needed during flu seasons can depend on several factors. Healthcare organizations and local health departments typically plan for flu seasons by considering historical data and projected demand.

Keanu Gomes

View Project in Tableau/GitHub

03 ROCKBUSTER STEALTH DATA ANALYSIS

Answering business questions for an online video rental company

Expectation

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.

Skills

- Relational databases
- Database querying Filtering
- Cleaning and summarizing
- Joining tables
- Subqueries
- CTEs

Tools

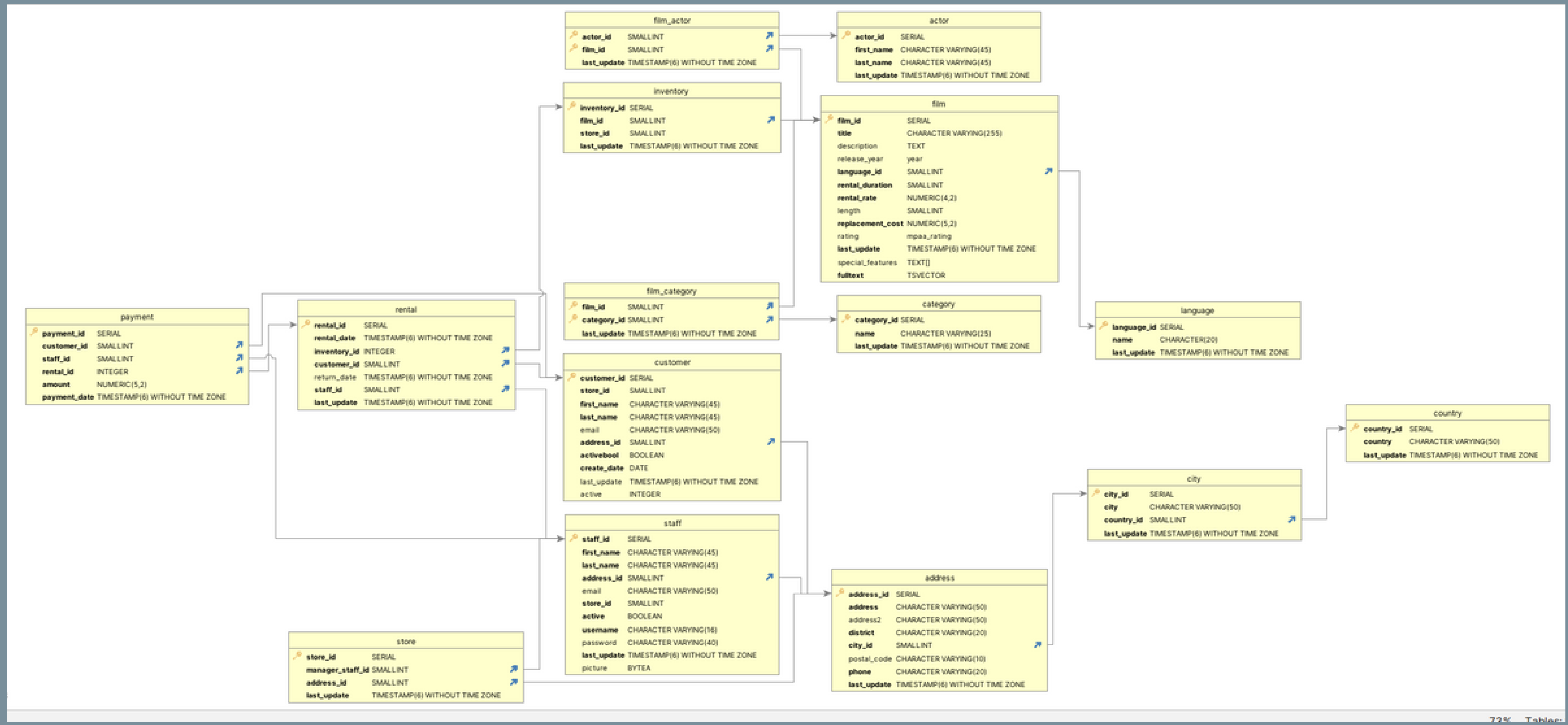
- PostgreSQL
- Microsoft Powerpoint
- Tableau

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.

Data analyst portfolio

03 ANALYSIS

Rockbuster Stealth ERD and overview of statistics



- 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.

View [Data Dictionary]

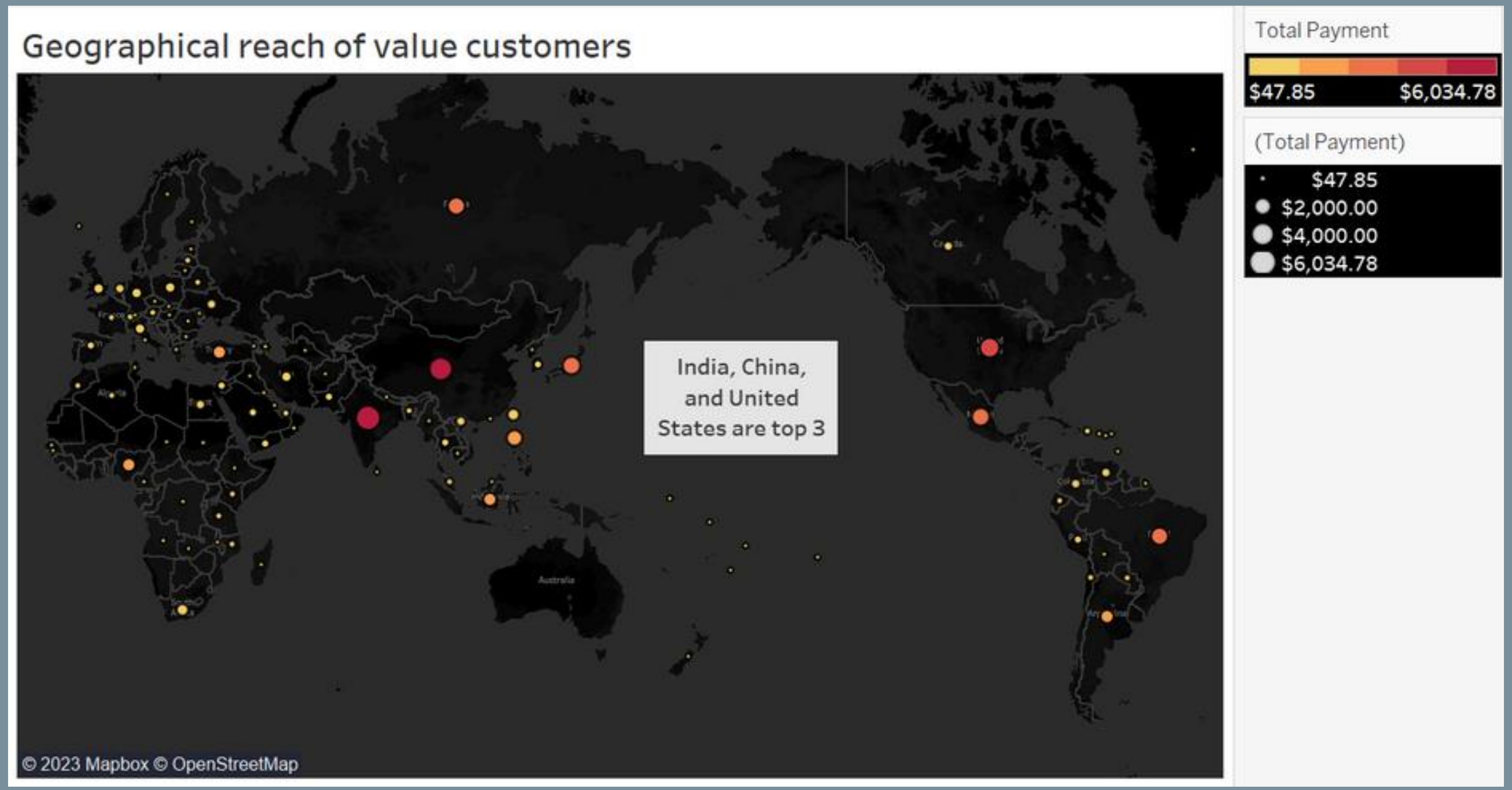
Calculate descriptive statistics for numerical columns

```
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;
```

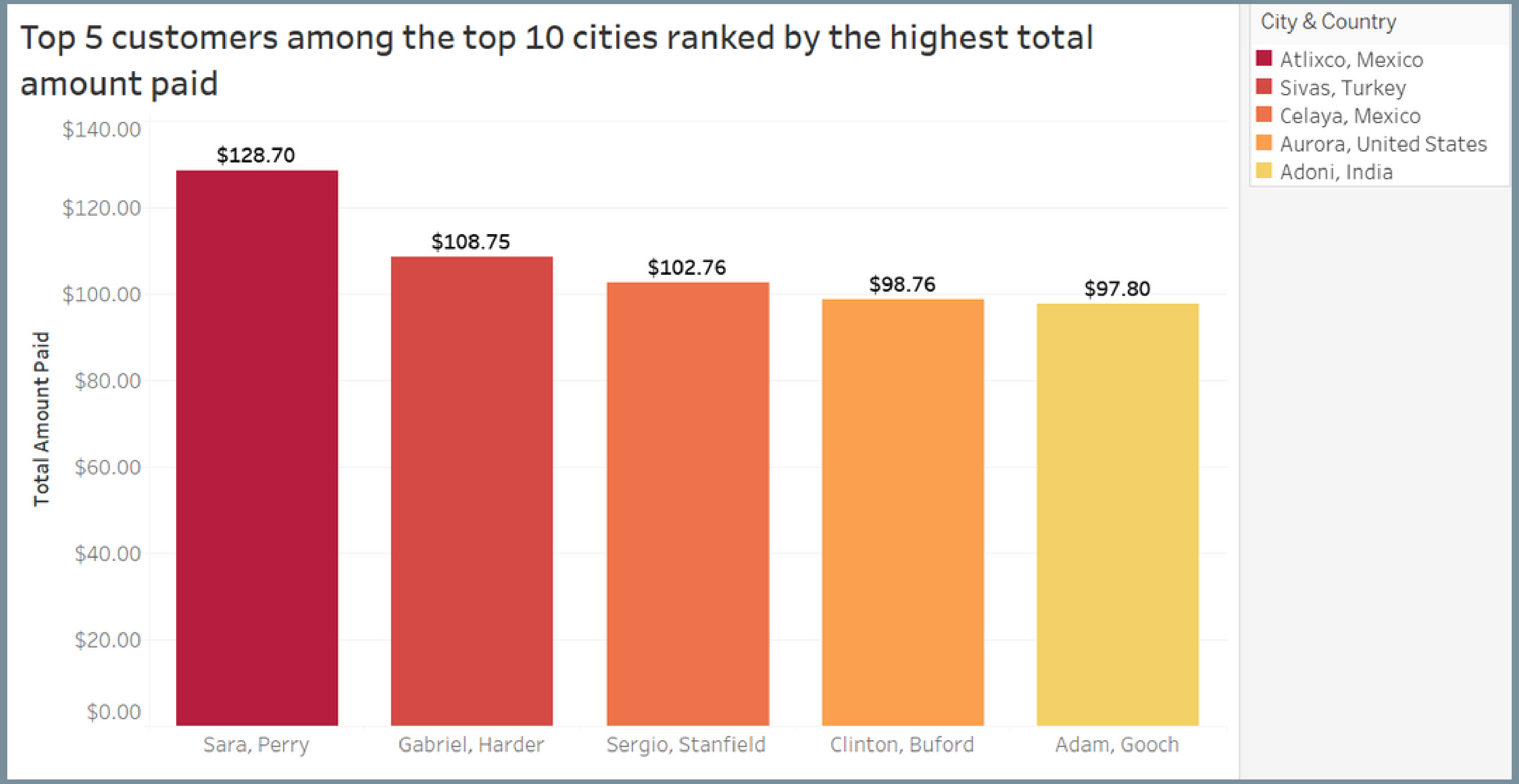
- A screenshot taken from my Excel workbook displays one SQL query used in my exploratory data analysis for descriptive statistics.

03 INSIGHTS

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



- Concluding from the heatmap created in Tableau, we can determine that India, China, and the United States are the top 3 regions in terms of total revenue.



- This bar chart illustrates the top 5 highest-paying customers among the top 10 highest-ranked cities. The top 10 cities were derived based on the highest customer count from both city and country. Here, we showcase the highest-paying customers from Mexico, Turkey, the United States, and India.

03 RECOMMENDATIONS

What do these insights tell us?

- Identify any possibilities of system errors, or reasons why (3) movie genres: Crime, Romance, and War haven't been included in the Rockbuster database payment system yet. Do this to ensure its validity.

- Focus marketing campaigns on top selling movies and genres and away from the less-contributing sellers.

- Develop a plan to give back to high value customers in order to help retain their commitment to the Rockbuster Stealth business.

04 INSTACART GROCERY BASKET ANALYSIS

Marketing strategy for an online grocery store

View Project in GitHub

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.

Skills

- Data wrangling
- Data merging
- Deriving variables
- Grouping data
- Aggregating data
- Reporting in Excel
- Population flows

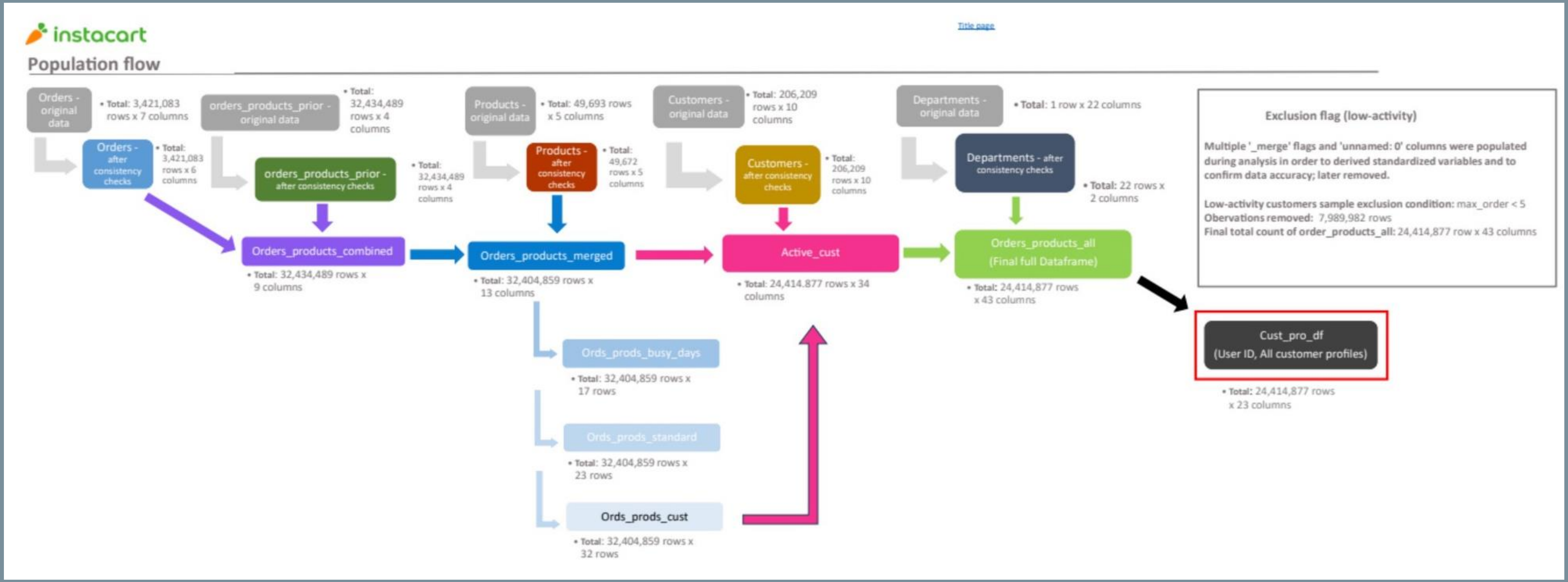
Tools

Microsoft Excel, Anaconda, Jupyter Notebook, Python

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

04 ANALYSIS

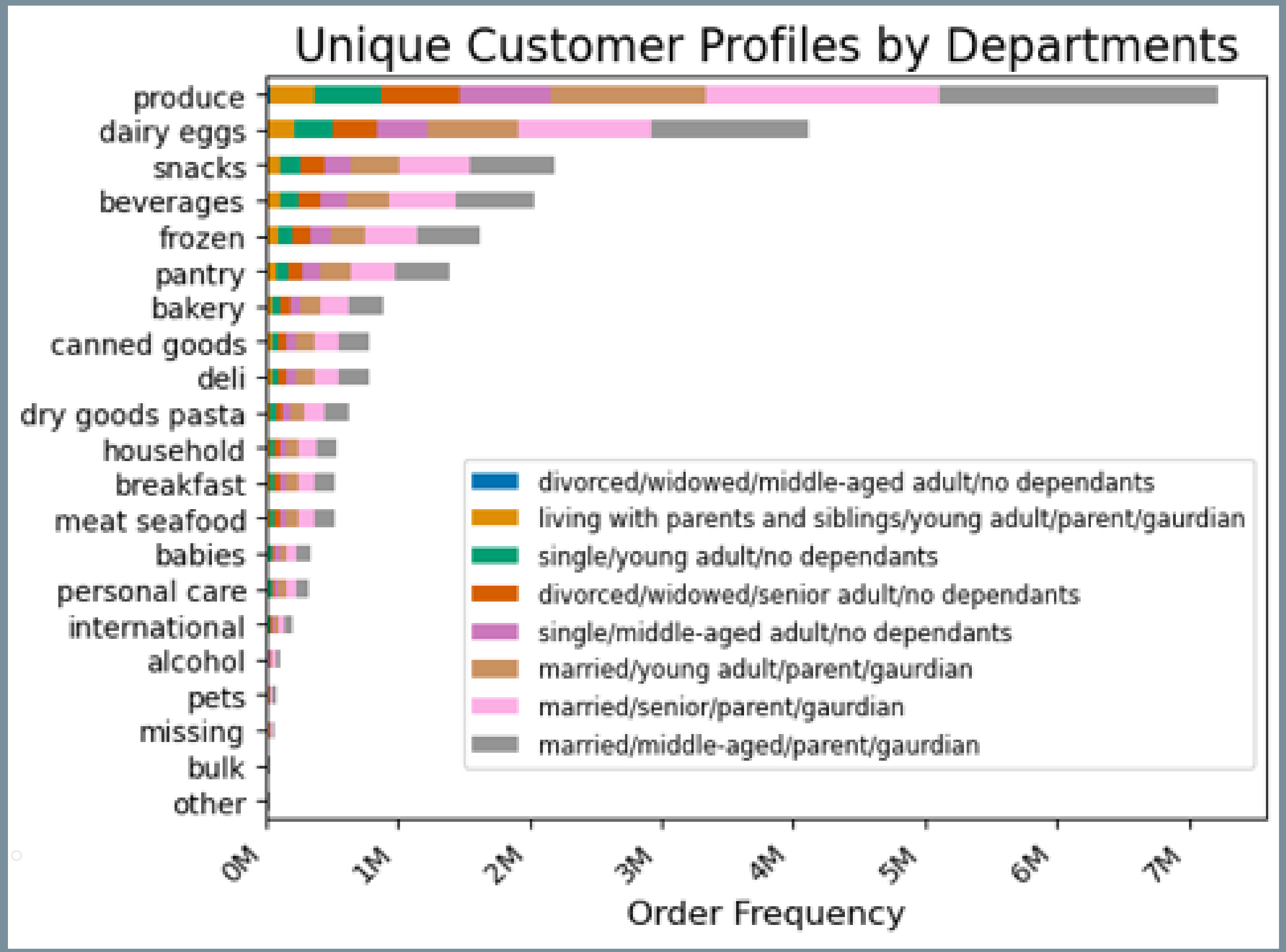
How was the Instacart Basket Analysis conducted?



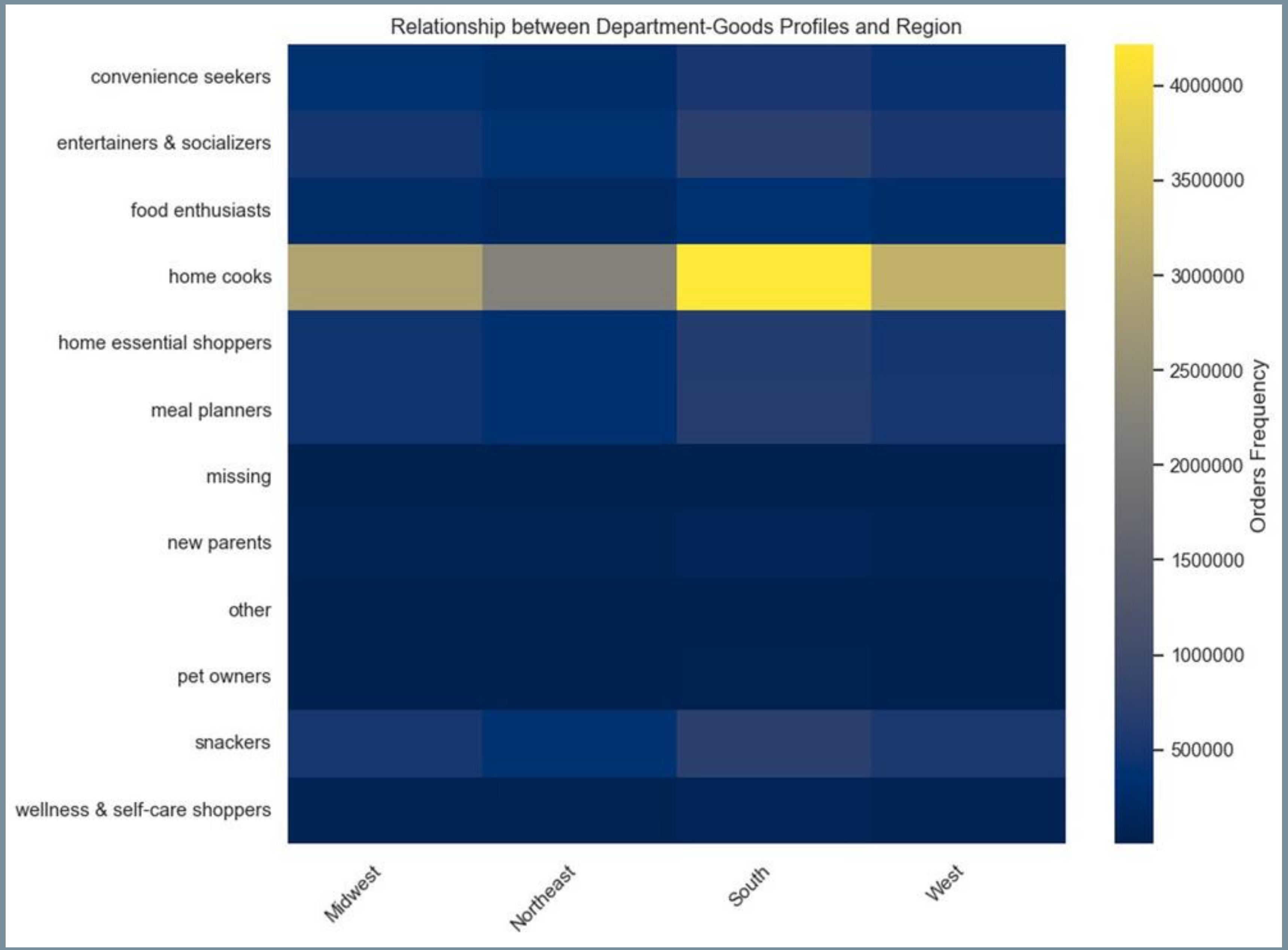
• Basket analysis population flow chart illustrated from my Excel workbook

04 INSIGHTS

Instacart's unique customer profiling classifications



- All customer profiles by department suggest the same ordering habits across all regions only varying by total order amounts.



- The relationship bewteen department goods profiles and region suggest a classifcation of (e.g. Home Cooks from the South region of the US)

04

RECOMMENDATIONS

What do these insights tell us?

- Run campaign ads Tues-Wed or Mon-Thurs (slowest weekdays).

- Additional review of loyalty program in order to attract new customers and maintain the attention/trust of current loyal customers.

- Create a discount campaign in order promote the loyalty program of instacart to its largest base of consumers (regular customer).

- Create/Promote campaigns targeted to customer profile demographics by region of the US:
  (e.g.)
    - (High-income/Male/Married/Middle-Aged/Parent/Gaurdian/Home Cooks)
    - (High-income/Female/Married/Middle-Aged/Parent/Gaurdian/Home Cooks)
    - (High-income/Male/Married/Senior/Parent/Gaurdian/Home Cooks)
    - (High-income/Female/Married/Senior/Parent/Gaurdian/Home Cooks)

- Early-morning hours around 3-6am suggest consumer habits willing to pay for higher priced items considering the time of day and food accessibility.

Keanu Gomes

05 PIG E. BANK FINANCIAL SERVICES

Predicting consumer churn rate with a classification model

Expectation

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.

Skills

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

Tools

- Microsoft Excel
- Microsoft PowerPoint

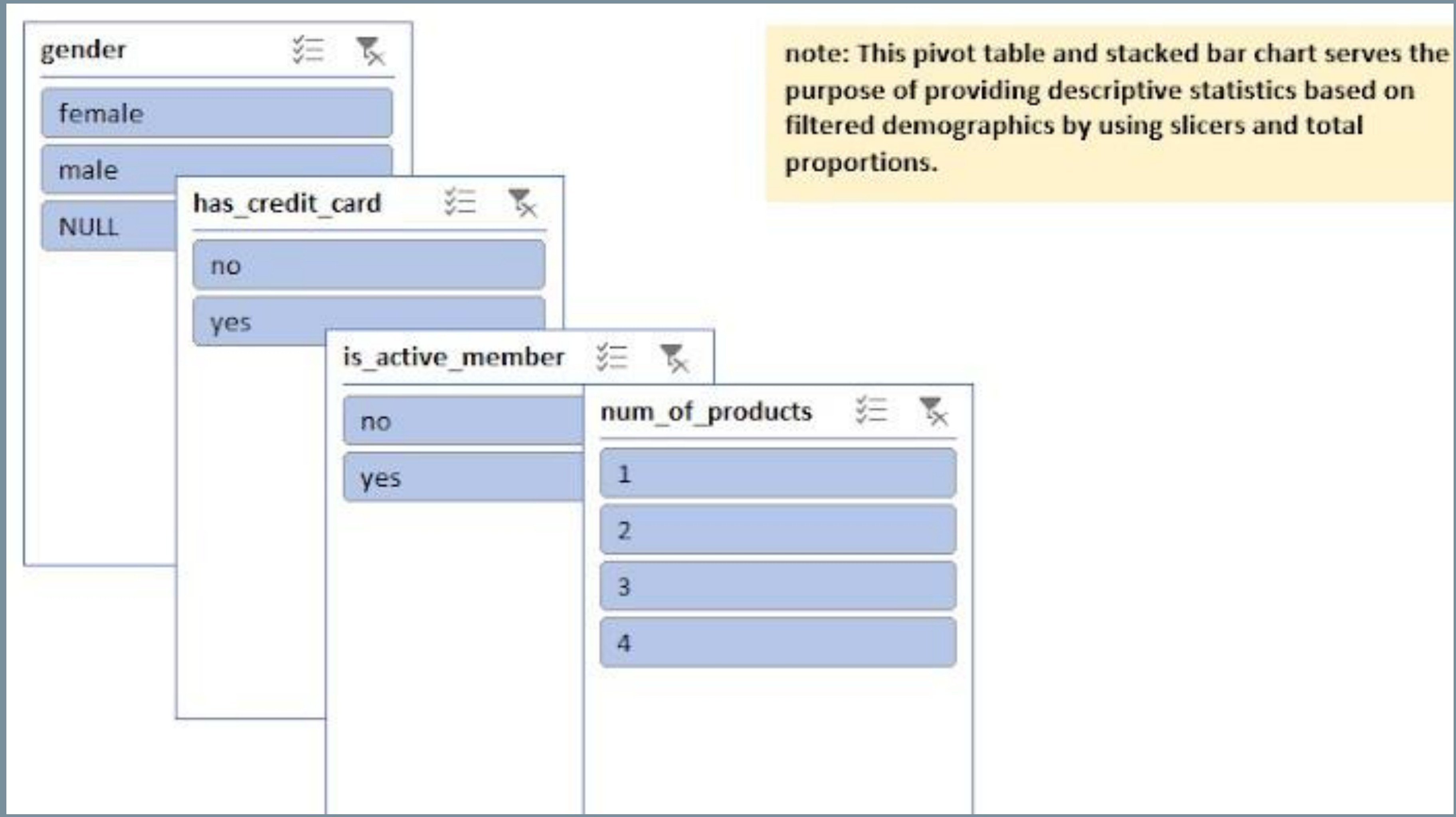
Goal: Use a predictive model to identify and segment banking members with a high likelihood of either exiting the bank or remaining as active or non-active members.

05 ANALYSIS

How was this financial services analysis conducted? What methods were used?

Columns	Missing values	Missing values treatment
last_name	1 NULL value, customer_id: 15752047	*PII Security Risk, entire column to be removed.*
credit_score	3 blank values, customer_id: 15627801, 15785542, and 15570060	*left-as-is.*
gender	1 NULL value, customer_id: 15737173	*left-as-is.*
age	1 NULL value, customer_id: 15699309	*left-as-is.*
est_salary	2 blank values, customer_id: 15597945, and 15785542	*left-as-is.*

Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
Row_Number			*Column is irrlevant to analysis.*
Last_Name			*PII Security Risk, column removed from data set analysis.*
	Tenure		*Tenure = the duration of the customer's relationship with the bank.*
	{Customer_ID:customer_id}, {Credit Score: credit_score}, {Country:country}, {Gender:gender}, {Age:age}, {Balance:balance}, {NumOfProducts:num_of_products}, {HasCrCard?:has_credit_card}, {IsActiveMember:is_active_member}, {Estimated Salary:est_salary}, {ExitedFromBank:exited_from_bank}		*Lowercasing and dashing implemented for smoother analysis.*
		balance	*inconsistent float numbers, change data type to float decimals at .00. Reformatted to account number format since these are balances of bank accounts.*

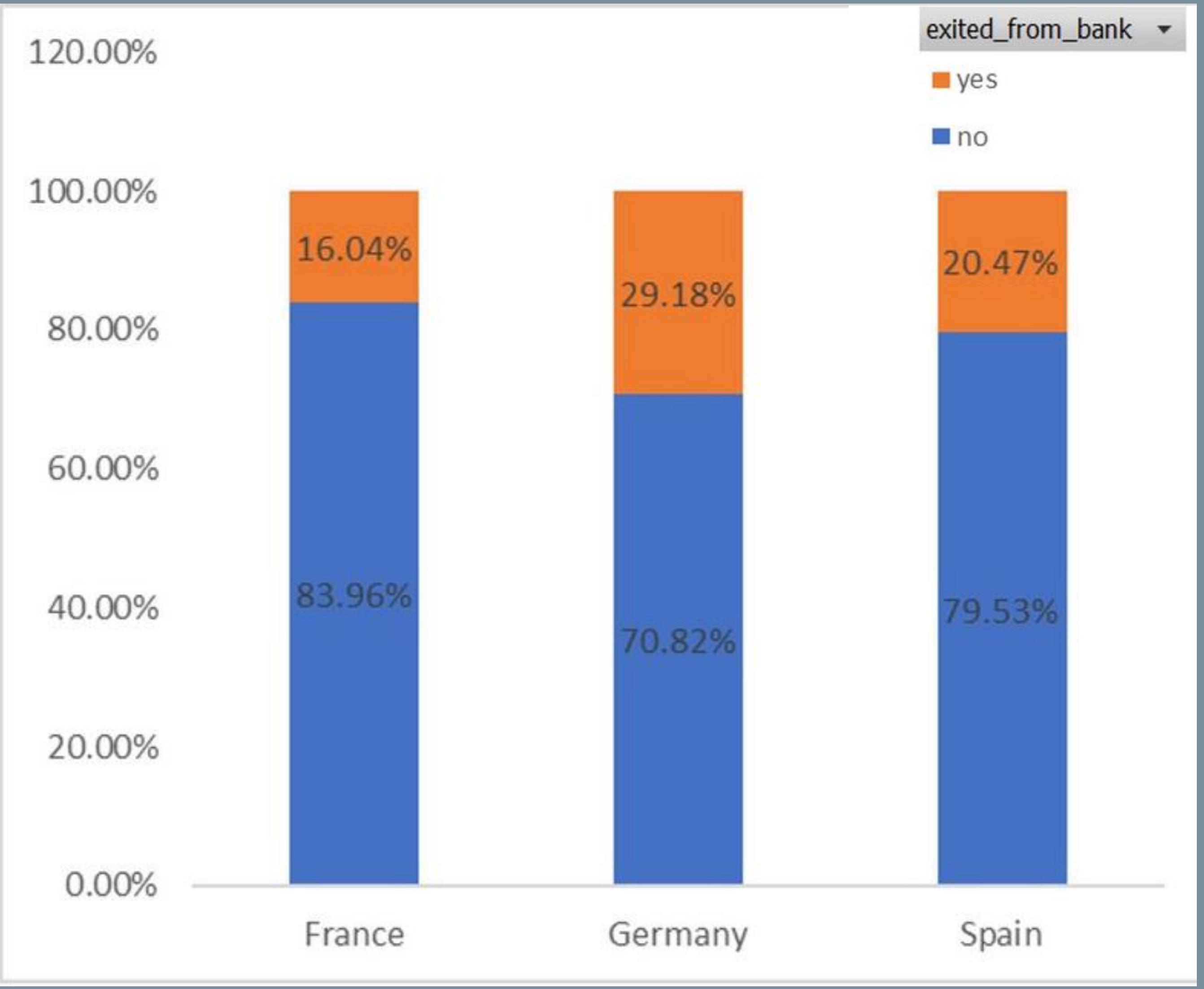


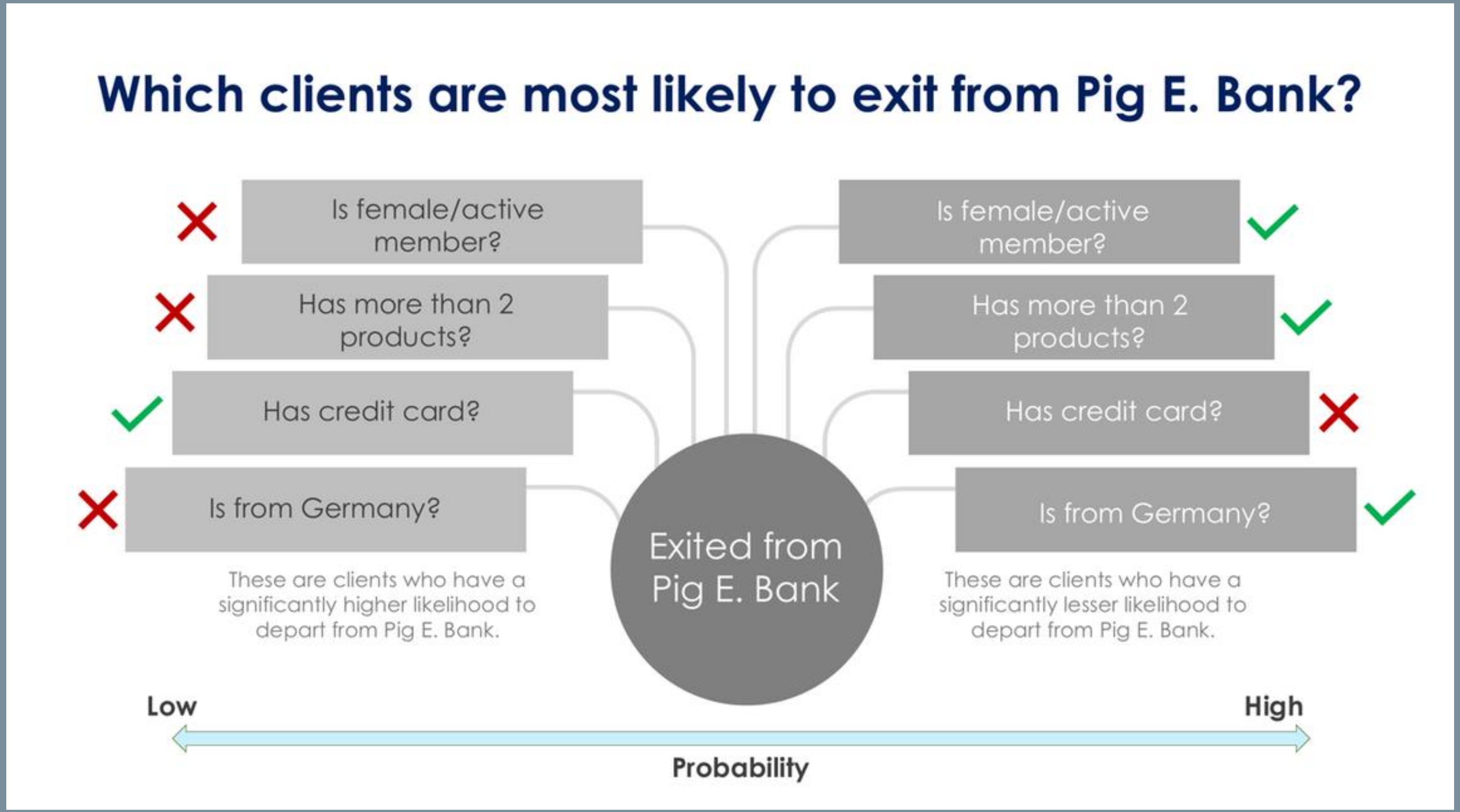
- The work illustrated in this analysis utilizes the CRISP-DM methodology. Above, displayed from Excel, I have showcased parts of the data understanding phase of my analysis.

- Here, in my data preparation phase after cleaning the data, I utilized a combination of slicers, pivot tables, and stacked bar charts to create a dynamic and user-friendly view. This view aims to analyze Pig E. Bank's banking clients for characteristics influencing members to either stay or exit the bank.

05 INSIGHTS

Which country had the highest exit rate? (logistic classification model)





- Germany had the highest exit rate from the bank at 29%, while France had the lowest exit rate at 16%.

- This simple decision-tree model illustrates that among female non-active members with a credit card, Germany had the highest exit rate from the bank at 39%. A more diverse model can be scaled by derived more variables, such as the region of Spain, which has the lowest exit rate at 28%, or the male gender.

05 RECOMMENDATIONS

What do these insights tell us?

- Germany had the highest exit rate from the bank at 29%, while France had the lowest exit rate at 16%.
- Among active members Germany had the highest exit rate from the bank at 20%, while France had the lowest exit rate with 9%.
- Among non-active members, Germany exhibits the highest exit rate from the bank at 39%, whereas France had the lowest exit rate at 23%.
- Among female non-active members without a credit card, Spain had the highest exit rate from the bank at 53%, while France had the lowest exit rate at 28%.
- Among female non-active members with a credit card, Germany had the highest exit rate from the bank at 39%, while Spain had the lowest exit rate at 28%.
- Among male non-active members with a credit card, Germany had the highest exit rate from the bank at 37%, while Spain has the lowest exit rate at 19%.

Keanu Gomes

06 BOAT WEBSITE VIEWS ANALYSIS

Yacht and Boat Website Views Analysis for Marketable Trends

VIEW PROJECT SCRIPTS IN GITHUB   VIEW TABLEAU DASHBOARD

Expectation

As a data analyst for a yacht and boat sales website, I've been tasked by the marketing team to analyze recent pricing and viewing data for their weekly newsletter. We're aiming to help sellers boost views and stay informed on market trends.

Skills

- Sourcing open data
- Exploring relationships
- Geograhical Visualizations
- Supervised ML
- Unsupervised ML
- Analyzing times series
- Creating data dashboards

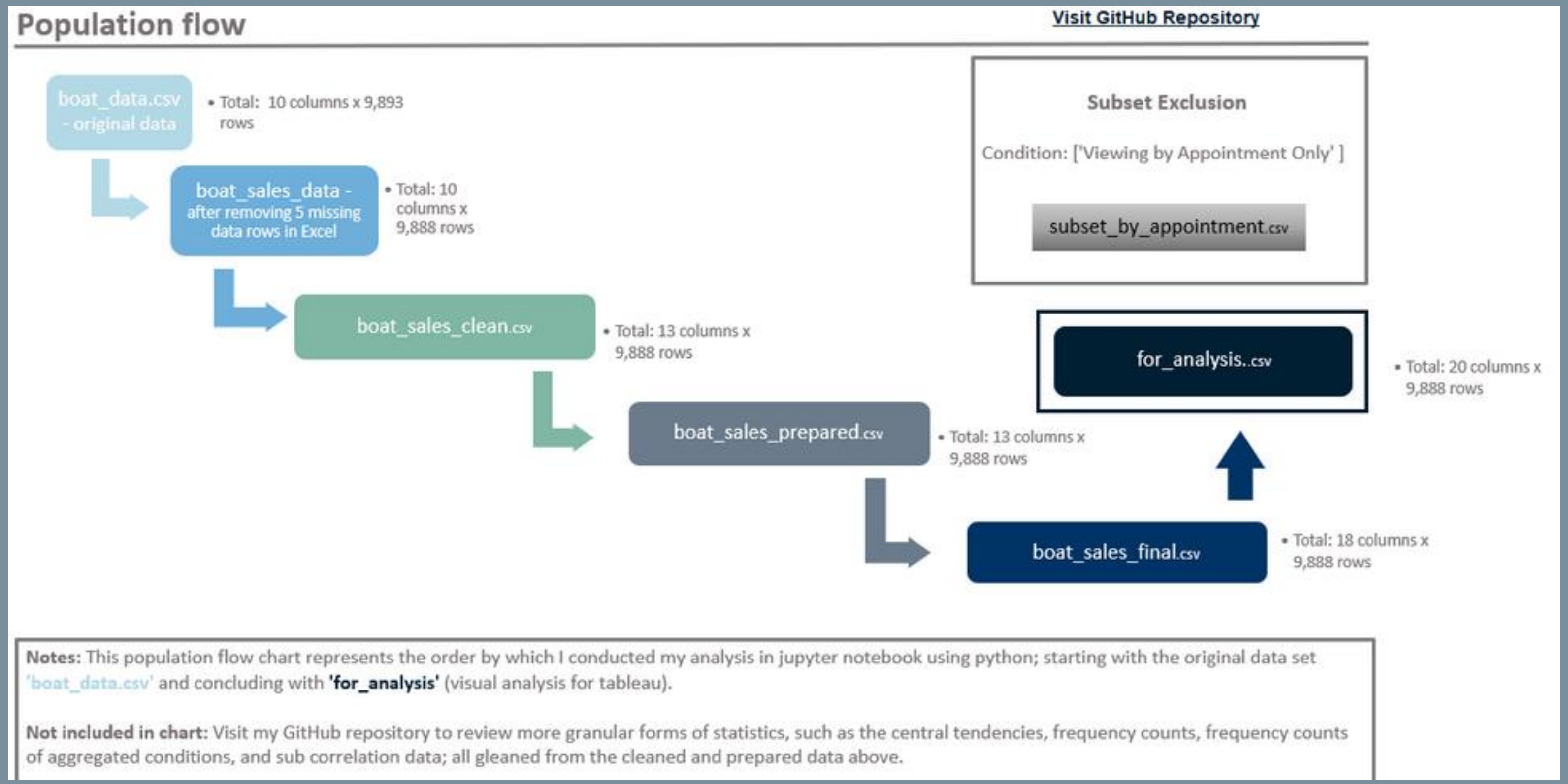
Tools

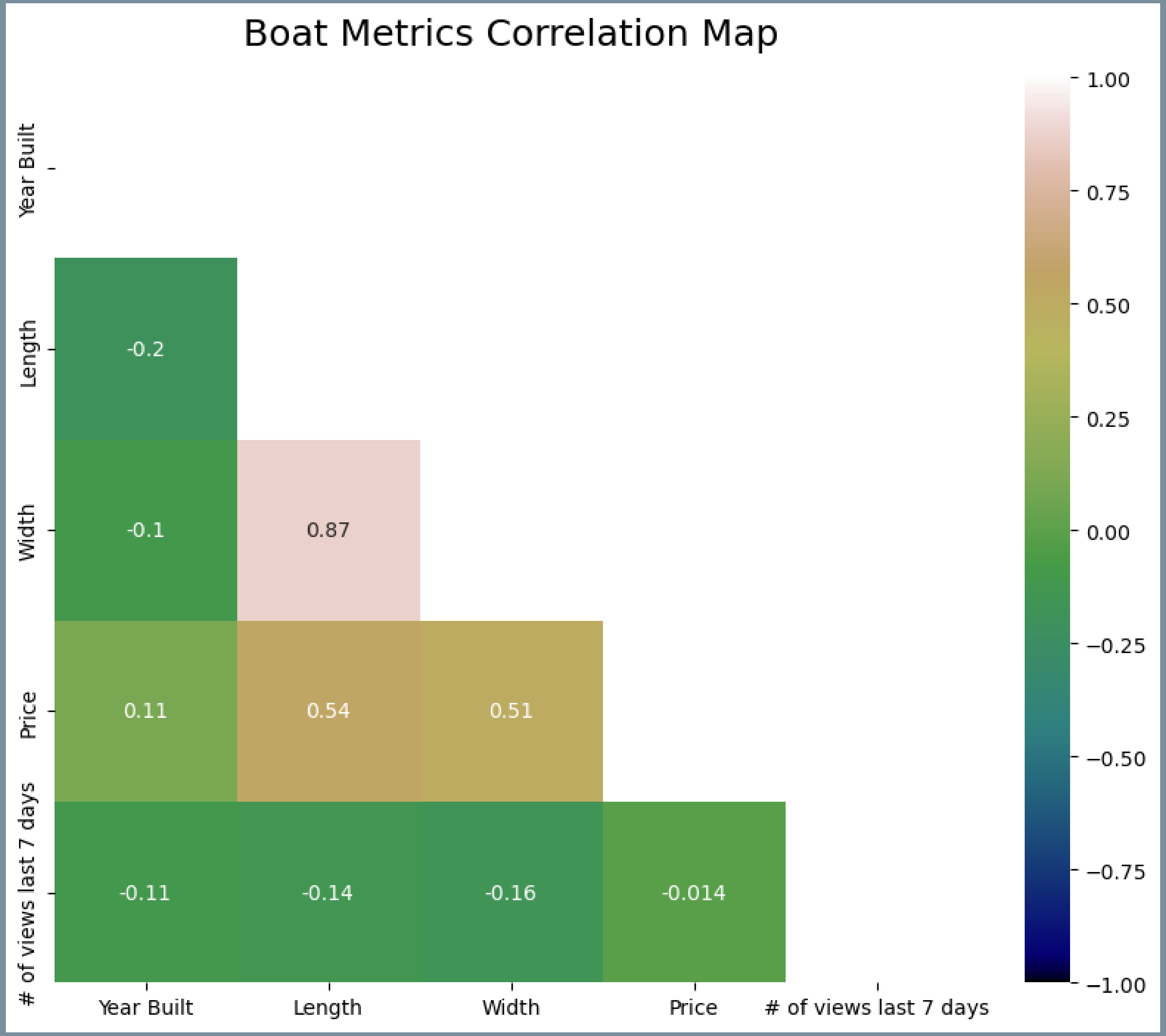
Microsoft Excel, Anaconda, Jupyter Notebook, Python

Goal: Utilize Python and machine learning (ML) to analyze views of yacht and boat listings on an online seller platform, aiming to discover the top characteristics of the most viewed boat listings.

06 ANALYSIS

How was the yacht and boat analysis conducted? What methods were utilized?

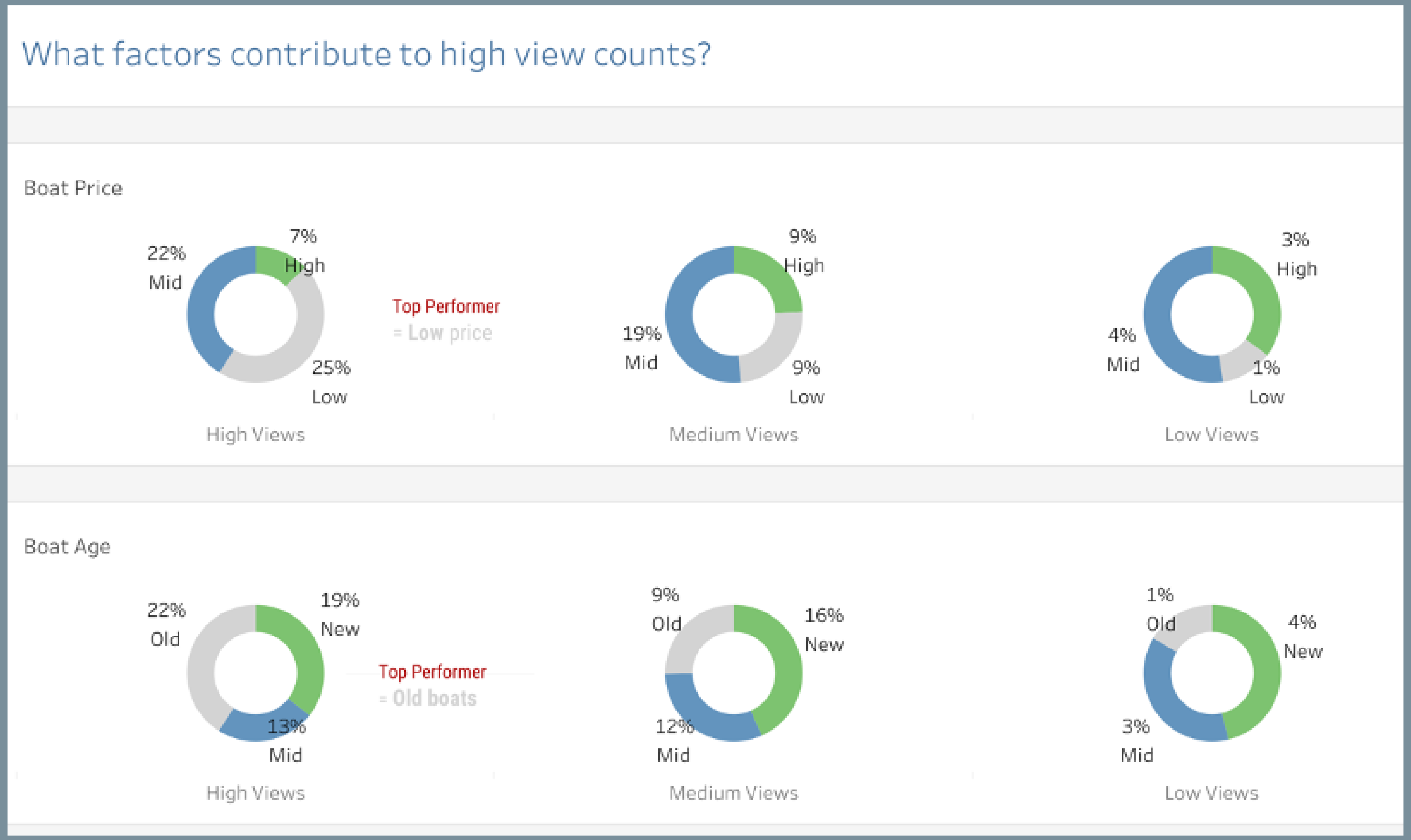


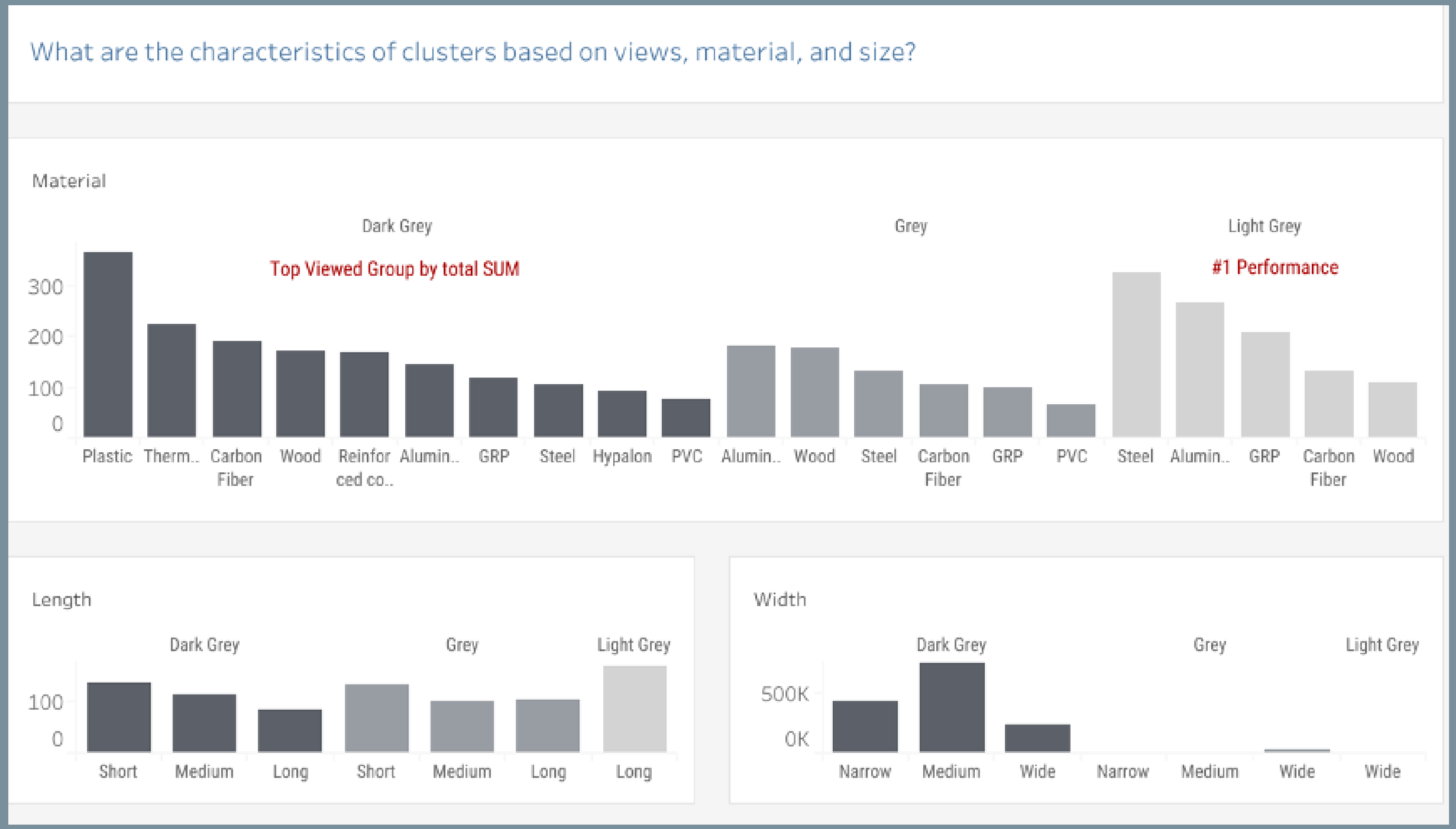


- Displayed above, the population flow chart represents the order by which my data cleaning processes were conducted in Jupyter Notebook utilizing Python.
- Displayed on the right, a correlation heat map was used to discover relationship strength between our numeric variables. A very weak positive correlation between top influencing characteristic Price by #Views was discovered.

06 INSIGHTS

What are the most common features among the top-viewed boats?





- Displayed above, I have compounded the two most correlated features into interactive business KPIs for easier interpretation. This was accomplished in my data preparation stage by aggregating new columns of data from its continuous variables into a hierarchy of 3 conditions.

- In this visualization above, I have created an interactive dashboard to compare analysis of the most featured characteristics of top viewed boats in the last 7 days. From these insights, I can now tailor marketable insights to the business owners and boat sellers.

06 RECOMMENDATIONS

What do these insights tell us?

- Highlight key attributes (age, price, condition) for better visibility & search engine optimization (SEO).
- Segment boats across price ranges and type to cater to a wider audience.
- Feature popular keywords (diesel, materials, brands) in listings to enhance SEO and attract more views.
- Focus marketing efforts on countries with high-viewed listings (Switzerland, Germany, Italy) to increase regional stability.
- Share market trends to help sellers optimize listings & improve search rank.

Main characteristics of top viewed:

- Boat Age: Old (<= 2000)
- Boat Price: Low-price (<= 44,000)
- Boat Condition: Used
- Boat Type: Motor yacht, Sport boat, Cabin boat
- Fuel Type: Diesel
- Boat Material: Plastic, Steel, Aluminum, GRP
- Manufacturers: Sunseeker, Beneteau, Jeanneau
- Boat Size: Long(> 13m), Short(< 8m) length & Medium(2-4m) width
- From Countries: Switzerland, Germany, Italy
- Most Used Currency: EU

Keanu Gomes

07 CLIMATEWINS WEATHER DATA

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

VIEW PROJECT SCRIPTS IN GITHUB   VIEW TABLEAU DASHBOARD

Expectation

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.

Skills

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

Tools

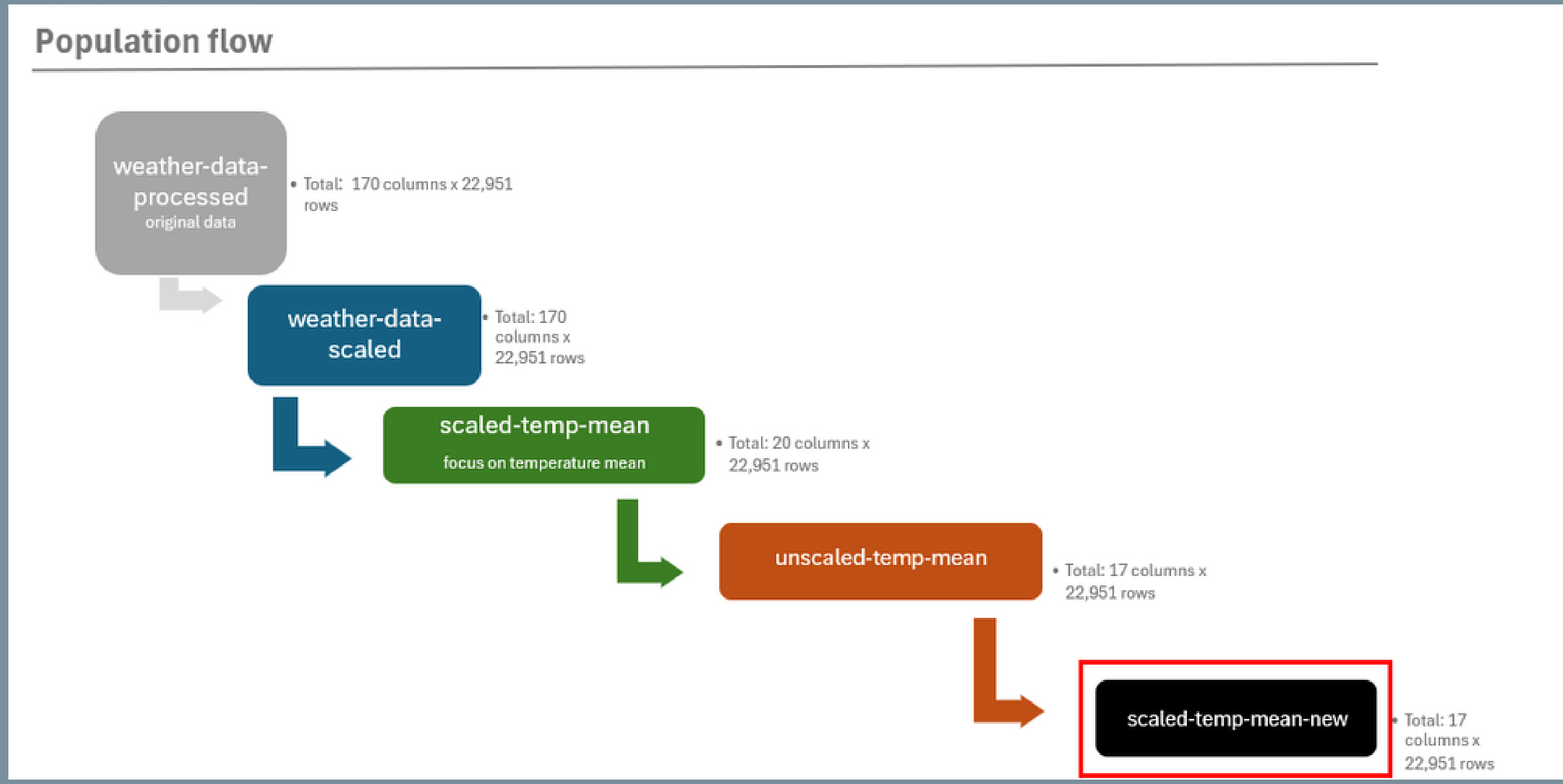
Microsoft Excel, Anaconda, Jupyter Notebook, Python

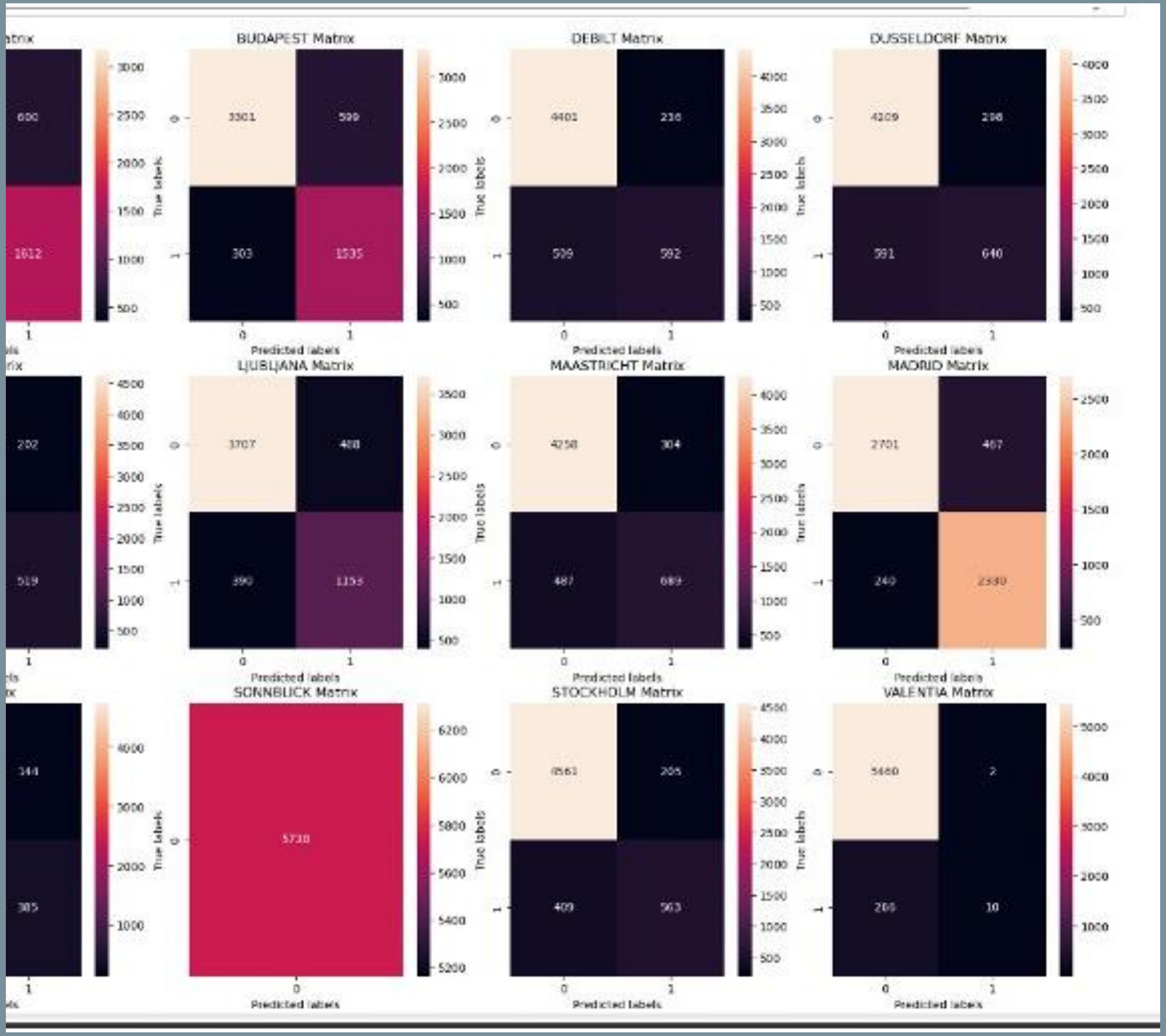
Goal: Utilize machine learning algorithms with Python to educate ClimateWins on choosing the most optimized algorithm to predict European extreme weather conditions.

Data analyst portfolio

07 ANALYSIS

How was this analysis conducted? What methods were utilized?





- Displayed on the right, is the population flow of my analysis through the (already) processed & cleaned data received from the weather stations.
- 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.

07 INSIGHTS

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

```
In [31]:
# Create the ANN
mlp = MLPClassifier(hidden_layer_sizes=(20, 10, 1(
#Fit the data to the model
mlp.fit(X_train, y_train)

Out[31]:
MLPClassifier
MLPClassifier(hidden_layer_sizes=(20, 10, 10), max_it(

In [32]:
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))

0.45044155240529865
0.4527710003485535
```

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

```
In [12]:
#What is the testing accuracy score? Usi
y_pred = weather_dt.predict(X_test)
print('Test accuracy score: ', accuracy_
multilabel_confusion_matrix(y_test, y_pr

Test accuracy score:  0.4051934471941443

Out[12]: array([[[3735,  603],
         [ 555,  845]],

        [[3143,  633],
         [ 622, 1340]],

        [[3339,  561],
         [ 578, 1260]],
```

- 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.

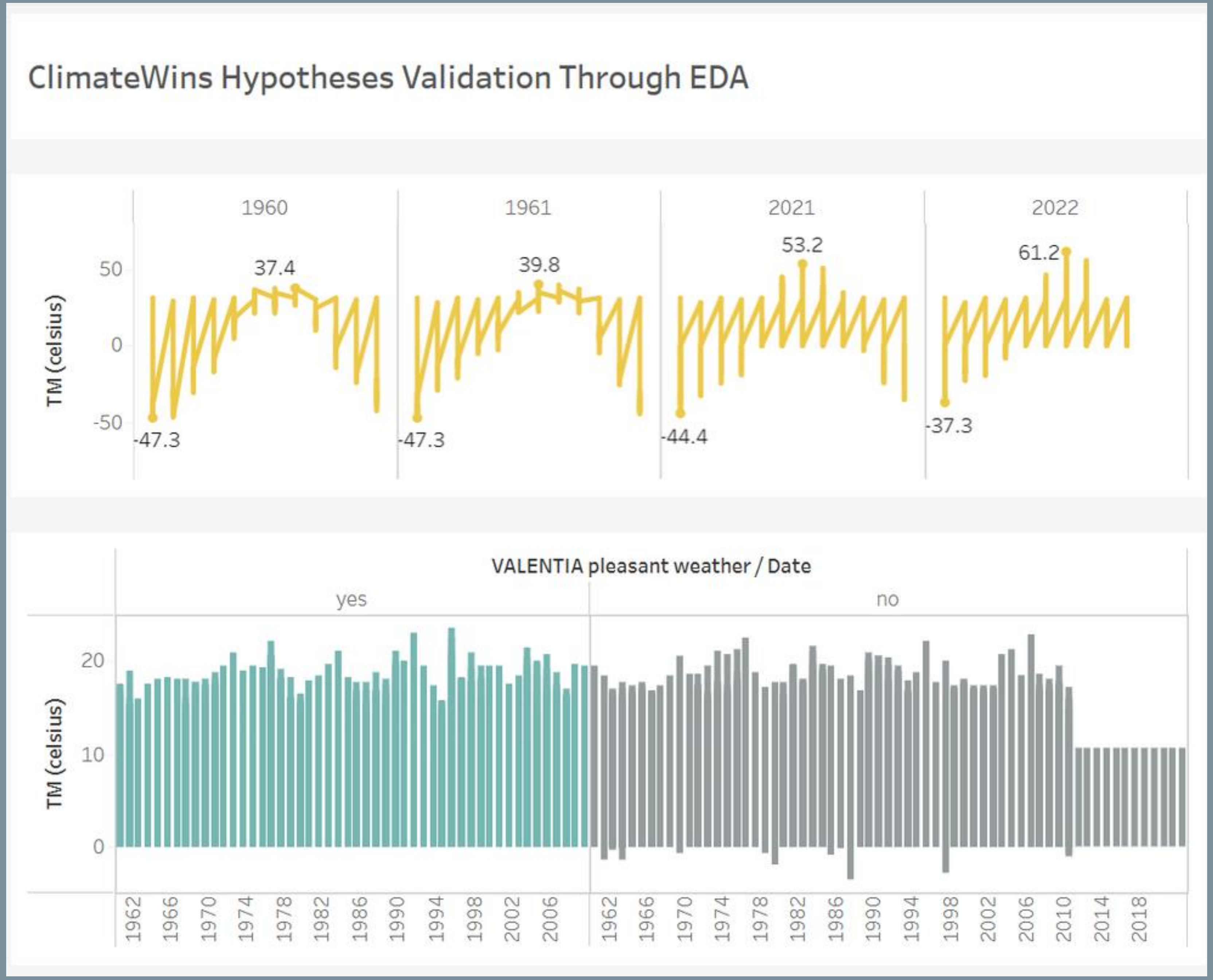
07

RECOMMENDATIONS

What do these insights tell us?

DISPLAYED BELOW - VIEW EDA RESULTS OF TABLEAU DASHBOARD

- All models show potential overfitting with lower overall accuracy compared to individual station scores.
- Further analysis and model refinement are necessary to address overfitting and improve generalization performance.
- Conduct feature importance analysis to leverage Valentia's strengths.
- Investigate data quality and potential biases.
- In summary, all three models (KNN, ANN, and decision tree) exhibit a discrepancy between their overall accuracy and the accuracy scores at the individual station level, indicating potential overfitting. The models may be fitting too closely to the training data and failing to generalize well to new or unseen data. Further analysis and model refinement are warranted to address overfitting and improve the models' generalization performance.



This concludes my portfolio, thank you for your time.

CONTACT ME

keanudatatech@gmail.com

Visit my Github repositories or Tableau storyboards