Andy Cen

Data Analyst Student

Case Study Portfolio

Projects

01

GameCo:

Analyzing global video game sales

02

Influenza Season:

Staffing necessities for flu season in the U.S.

03

Rockbuster Stealth:

Business questions for an online video rental company

04

Instacart:

Marketing strategy for an online grocery store

05

Salary:

Predicting the salary of different job titles in different countries

06

ClimateWins:

Predicting the consequences of climate change in Europe

07

ClimateWins:

Predicting extreme weather in Europe

Productivity Tools

Excel

PowerPoint

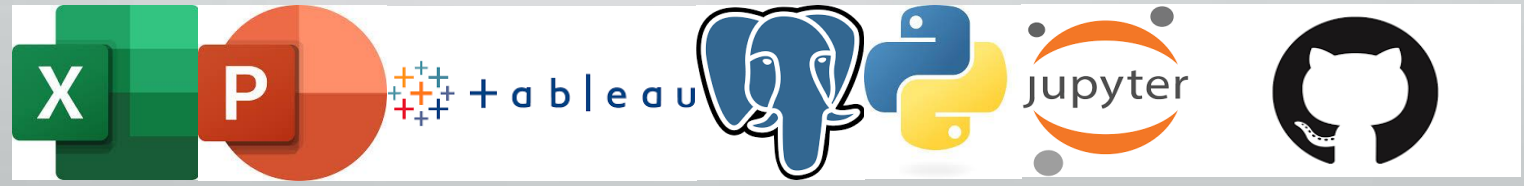
Tableau

PostgreSQL

Python

Jupyter Notebook

Github



Project Brief
Video Game Sales Dataset

GameCo

Context:

- GameCo is a new video game company that wants to use data to inform the development of new games. They want a better understanding oh how GameCo's new games might fare in the market.

Key Skills Applied:

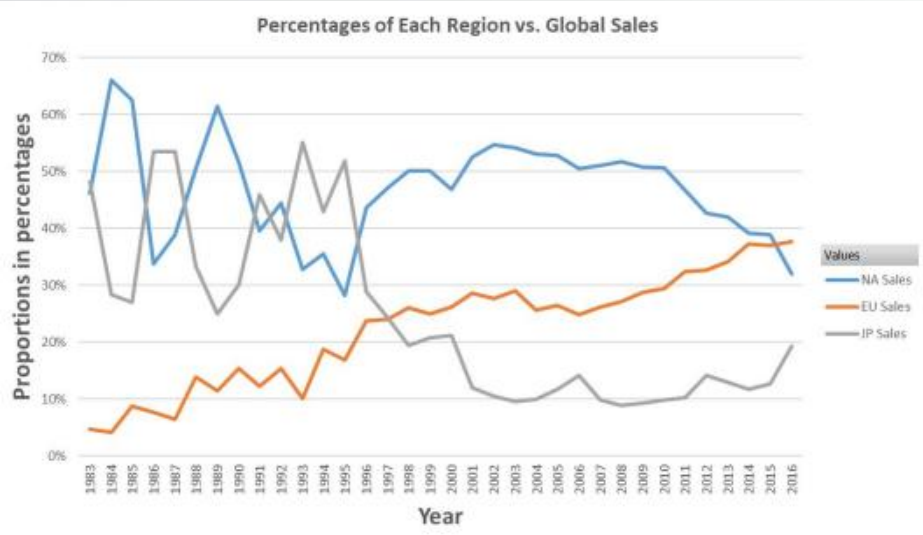
- Cleaning, filtering, sorting, summarizing and grouping data
- Calculated Fields
- Data Visualization in Excel
- Storytelling with Data
- Descriptive Analysis

Objectives:

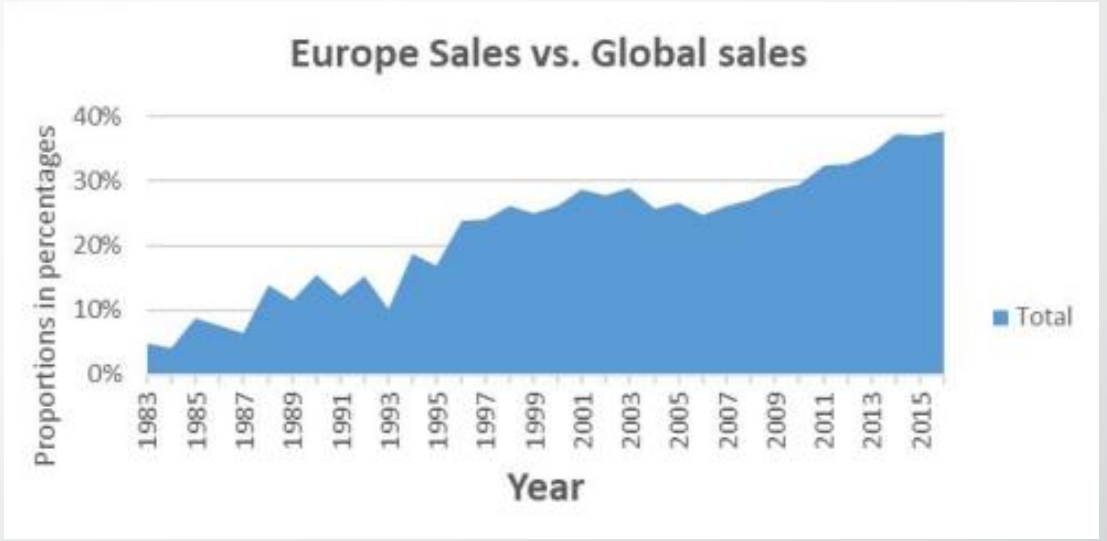
- Comparison of GameCo's popularity of games
- Identify GameCo's competitors
- Evaluate game's change in popularity over time
- Analyze sales over geographic regions over time

Productivity Tools Used:

Global Analysis

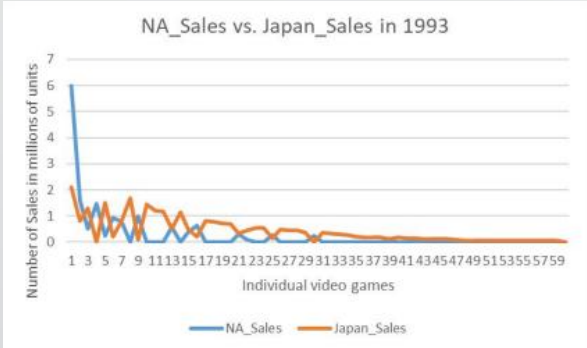


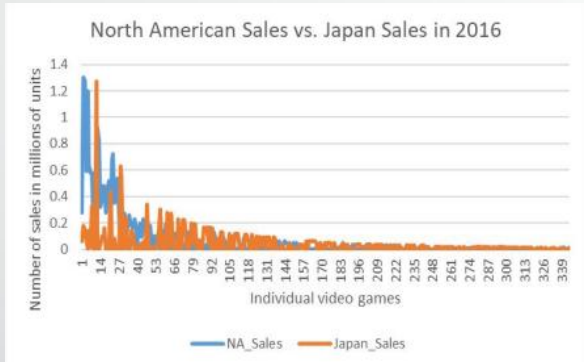
Europe has shown consistent growth, while North America and Japan display an inverse sales pattern.



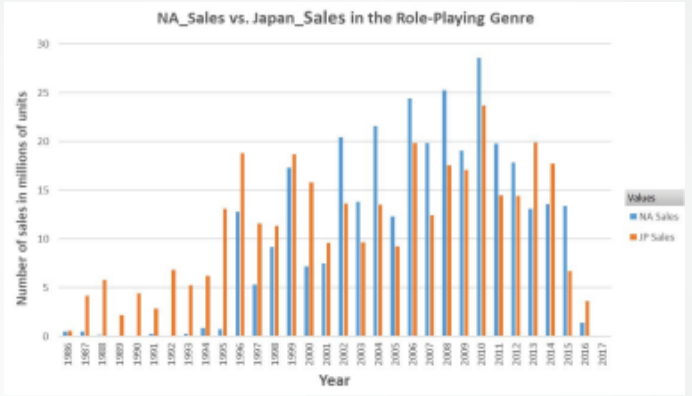
The graph shows European sales following a steady positive trend, surpassing North American sales in 2015-2016, after North America had previously led.

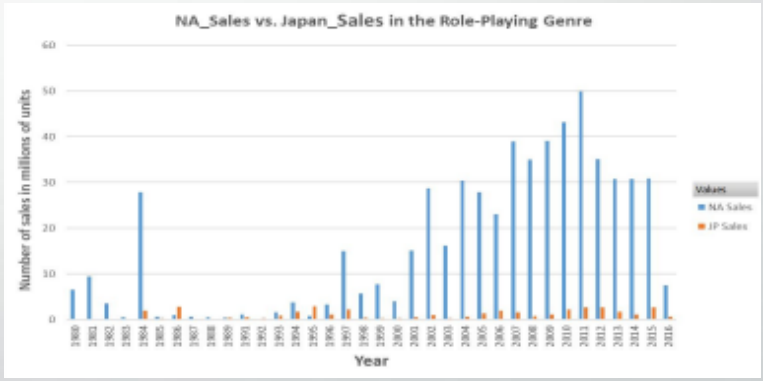
North American vs. Japan Analysis





The 1993 graph shows a mirroring sales pattern between North America and Japan in 1993, while the 2016 graph confirms the consistent alternating sales pattern.





Shooter games dominate North America, while Role-Playing games slightly lead there. Shooter releases boost North America but underperform in Japan, where Role-Playing games drive closer competition.

Insights & Recommendations

PowerPoint Presentation

01

Europe has demonstrated **a positive steady growth** and now **leads** in sales.

02

North America and Japan have experienced **fluctuating sales trends,** with North America holding up better than Japan.

03

GameCo should make efforts in bolstering Japan's low sales and **allocate targeted funding** for North America.

04

Prioritize additional resources in Europe to **maximize** on its **steady growth**.

05

Adjust its sales strategies accordingly **to regional trends** for **optimal performance**.

Influenza Season in the U.S.

[Flu Shot Rates Survey]
[Lab Tests Dataset]
[Influenza Visits Dataset]
[CDC (Fluview)]
[Population Dataset]
[CDC Influenza Deaths Dataset]

Context:

- The United States experiences an influenza season where more people than usual suffer from the flu, leading to more hospitalizations, especially among vulnerable populations. During this time, medical staffing agencies provide temporary staff to meet the increased demand.

Key Skills Applied:

- Data cleaning, Data Integrity, & Data Profiling
- Translating Business Requirements
- Forecasting
- Data Visualization with Tableau
- Storytelling with Data
- Statistical Hypothesis testing

Objective:

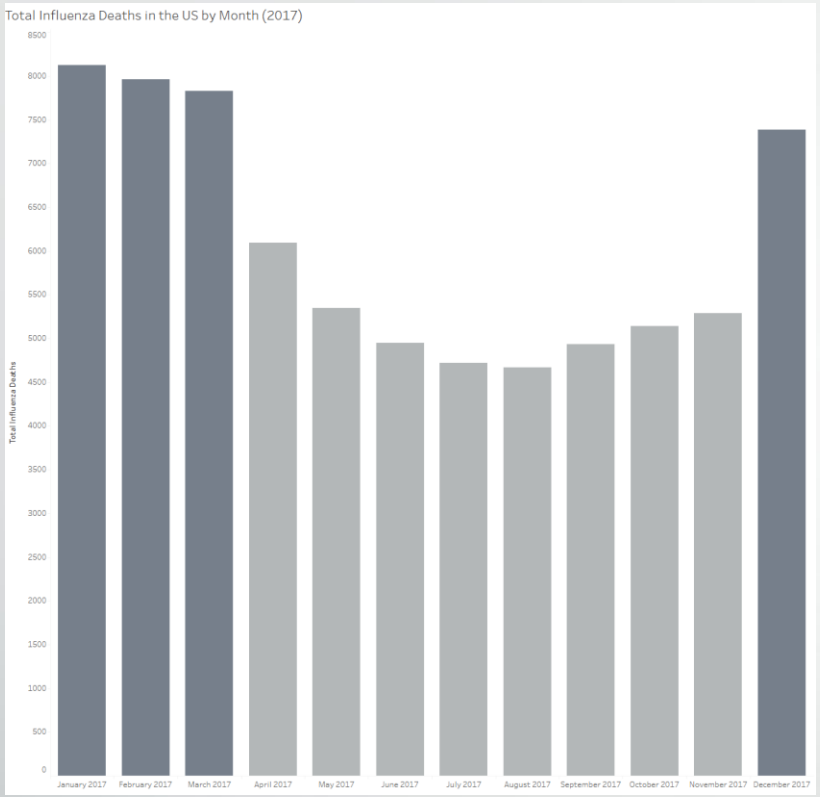
- To evaluate staffing requirements for influenza across the United States by comparing the variation influenza death counts between vulnerable (under 5 years and over 65 years) and non-vulnerable populations (between 5 and 65 years old).

Productivity Tools Used:

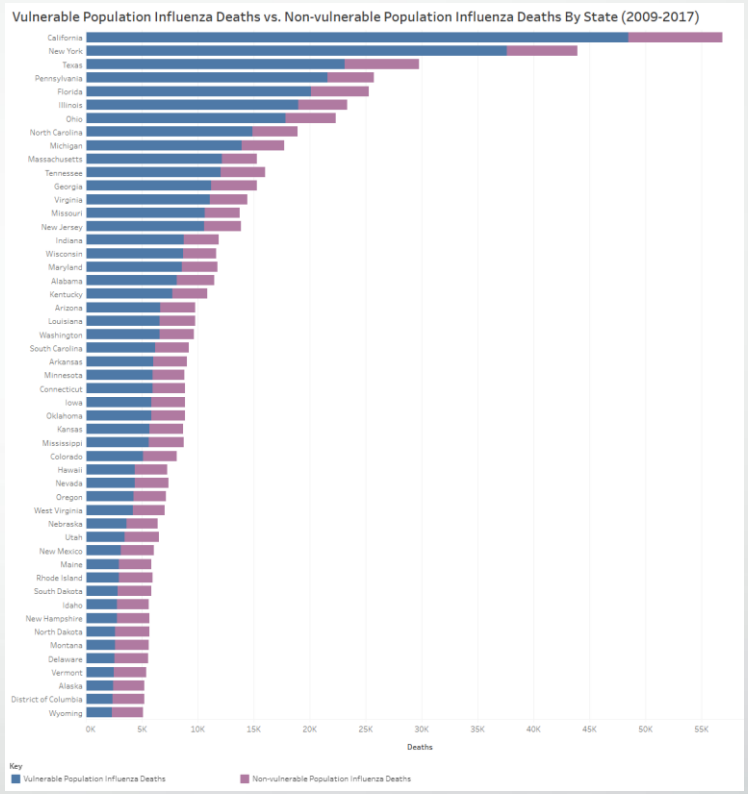
[Project Brief]

[Tableau Link]

Analysis



December, January, February, and March are the peak months of influenza. Winter has the highest death count of influenza and summer has the lowest death count of influenza.



California, New York, Texas, and Pennsylvania has the highest number of influenza deaths. We can distinguish the regional disparities and take insights for further action.

Insights & Recommendation

01

Resource Allocation: Since vulnerable populations account for the majority of influenza deaths, directing resources to areas with large vulnerable groups can help reduce mortality rates.

02

Targeted Focus; Prioritize regions and states (**Wyoming, Vermont, and District of Columbia**) experiencing high influenza death rates, with particular attention to vulnerable populations.

03

Staffing During Peak Seasons: Allocate additional staff and resources during the winter months (**December, January, February, March**) when influenza deaths are at the peak.

04

Enhance Public Awareness: Increase efforts to promote public awareness and preventative measures to mitigate influenza impacts.

Rockbuster Stealth

Context:

- 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 in order to stay competitive.

Key Skills Applied:

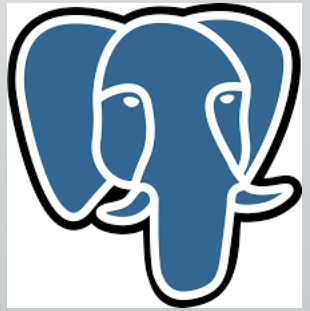
- Creating a Data Dictionary
- Relational Databases
- SQL:
  - PostgreSQL
  - Data Cleaning
  - Filtering
  - Joining Tables
  - Subqueries
  - CTEs
- Data Visualization with Tableau

Objective:

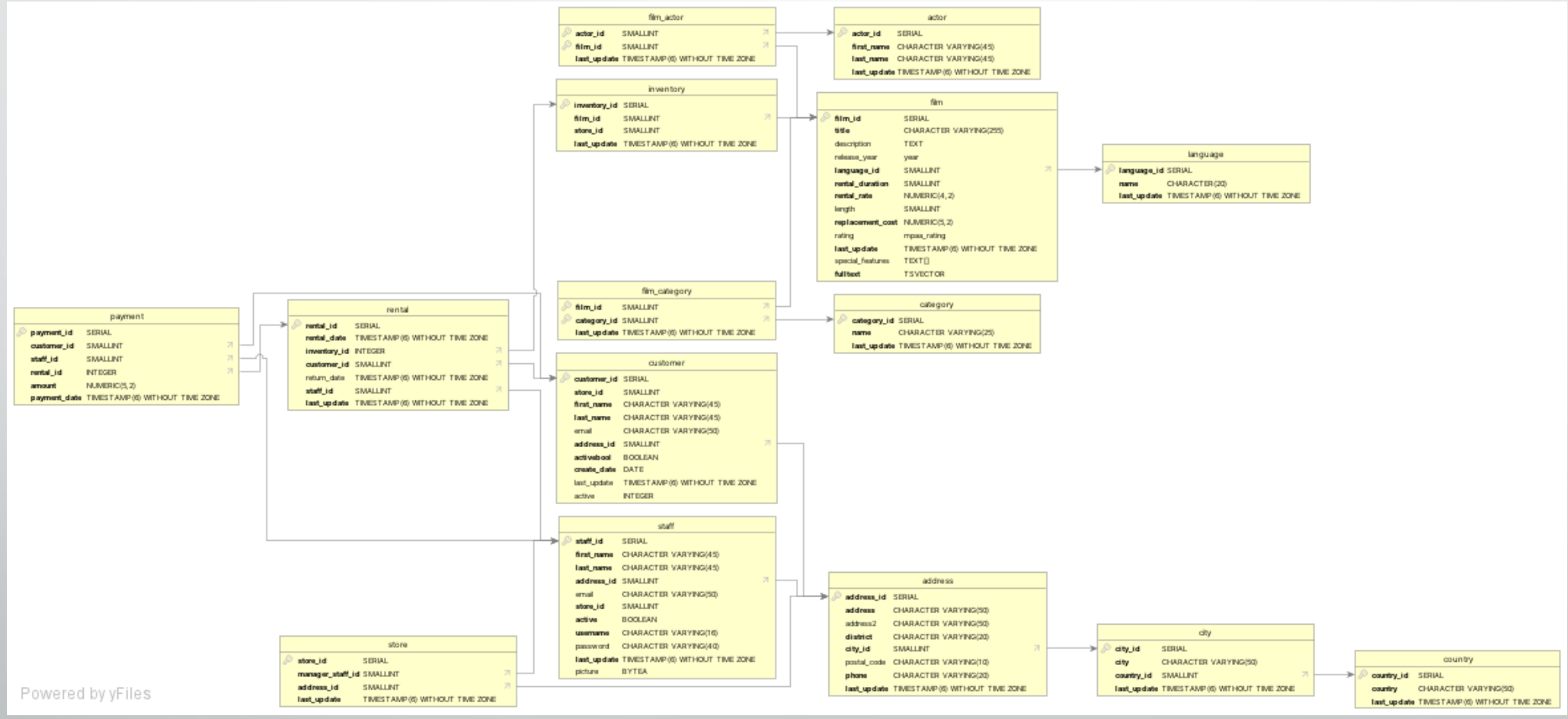
- Conduct an exploratory analysis of customer rental history to identify patterns by revenue contribution, customer geographic, sales by region, etc. to aid their new online video service.

Project Brief

Productivity Tools Used:

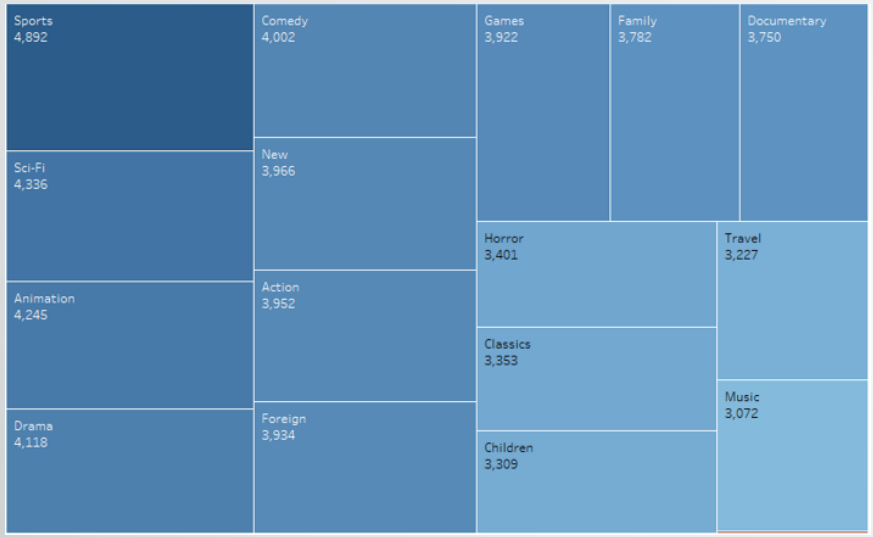


Entity Relationship Diagram



Analysis

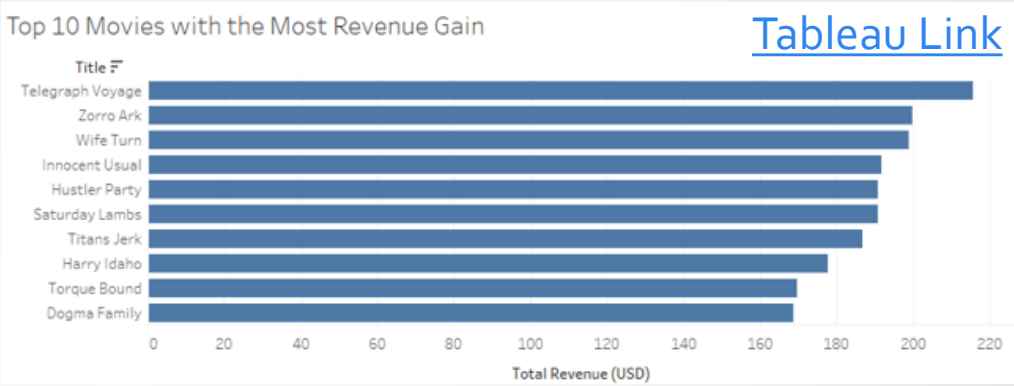
Tableau Link



The top movie categories or genres are Sports, Sci-Fi Animation, Drama, and Comedy.

Top 10 Movies with the Most Revenue Gain

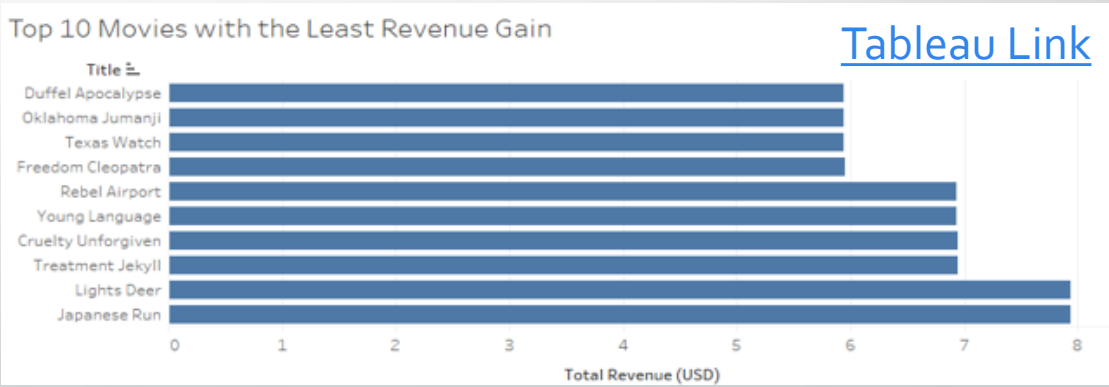
Tableau Link



Telegraph Voyage, Zorro Ark, and Wife Turn are the top movies with the most revenue gain.

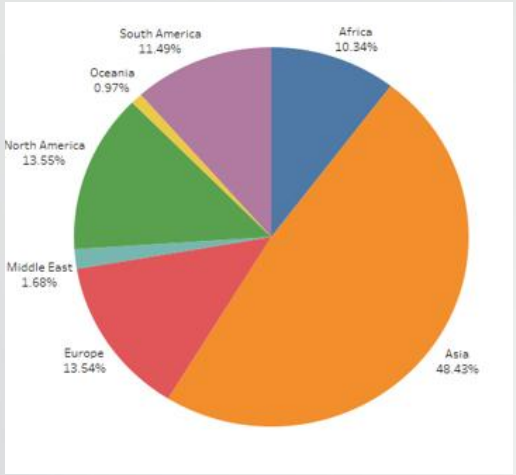
Top 10 Movies with the Least Revenue Gain

Tableau Link



Duffel Apocalypse, Oklahoma Jumanji, Texas Watch, and Freedom Cleopatra are the movies with the least revenue gain.

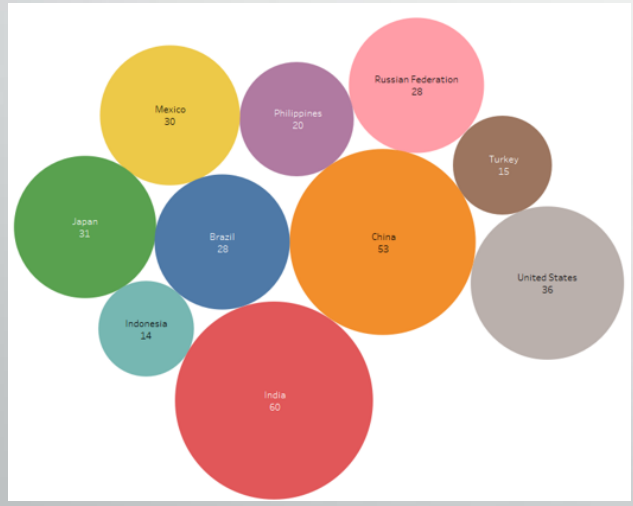
Revenue by Geographic Regions



Asia is leading in sales figures with over 48% of sales.

Tableau Link

Top 10 Countries by Number of Customers



India has the greatest number of Customers.

Tableau Link

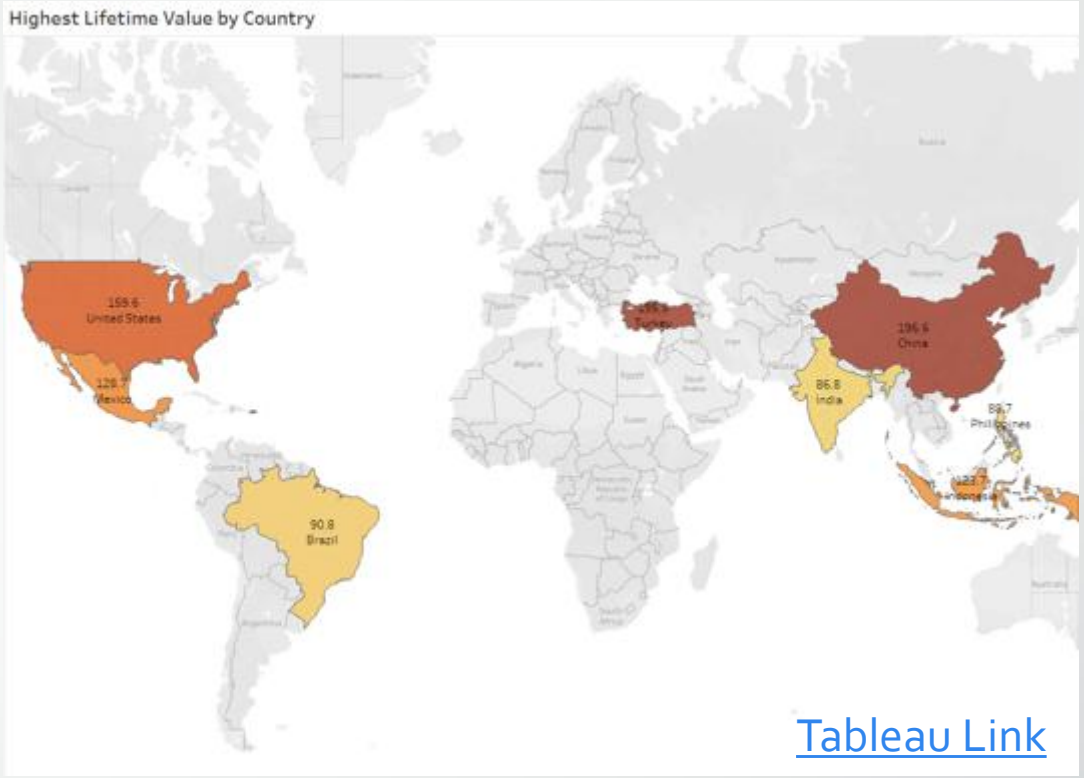


Tableau Link

China and Turkey has the highest lifetime value of customers.

United States, Mexico, Indonesia follow behind.

Brazil, India, and the Philippines are last.

Insights

PowerPoint Presentation

01

INDIA HAS THE GREATEST NUMBER OF CUSTOMERS WITH **60 CUSTOMERS**.

02

ASIA IS LEADING IN **SALES FIGURES** WITH **48.43%** OF SALES.

03

THE TOP MOVIE GENRES ARE **SPORTS**, **SCI-FI ANIMATION**, **DRAMA**, AND **COMEDY**.

04

TELEGRAPH VOYAGE, **ZORRO ARK**, AND **WIFE** TURN ARE THE TOP MOVIES WITH THE **MOST REVENUE GAIN**.

Recommendations

Global Market Expansion:

- Asia has over **48%** of Rockbuster's sales and revenue. Exerting **more resources** and engagement into Asia's market may better **improve profits**.
- Consider the lesser popular regions like **Oceania** and the **Middle East** to expand and explore **potential markets**.

Loyalty Programs:

- Offer customers with **high contributions** with great support and customer service. Include special offers and events that provide special privileges like **early access, reduced prices,** or **rewards**.

Top Performing Categories & Genres:

- Top movie categories or genres are **Sports, Sci-Fi Animation, Drama,** and **Comedy**. Rockbuster should consider promoting these genres more since they generate **high revenue**.

Capitalize on Top Revenue Generating Movies:

- Top revenue generating movies like **Telegraph Voyage** and **Zorro Ark** should be advertised and promoted more to yield more **profits**.

CareerFoundry Dataset
Instacart Dataset
Project Brief

Instacart

Context:

- Instacart is the leading grocery delivery platform in the US. Instacart stakeholders aim to understand customer diversity and purchasing behaviors to develop a targeted marketing strategy.

Key Skills Applied:

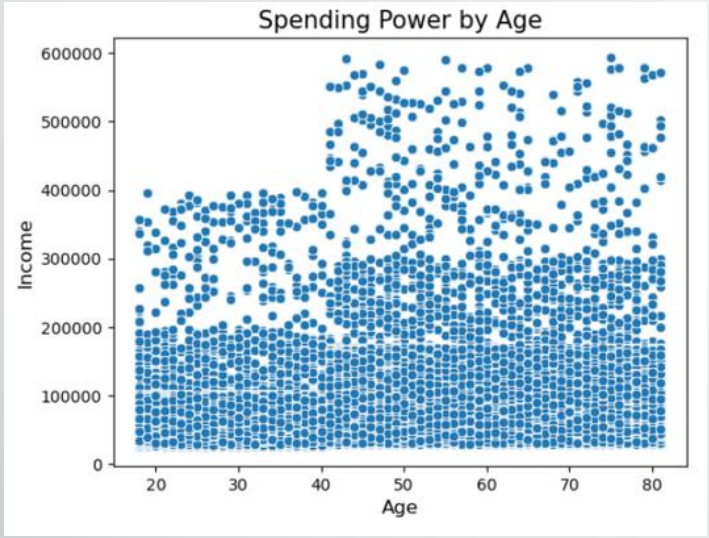
- Python(Jupyter Notebook)
- Data Cleaning
- Data Grouping, Wrangling, and Subsetting
- Merging Dataframes
- Deriving new variables
- Data Aggregation & Grouping
- Data Visualization

Objective:

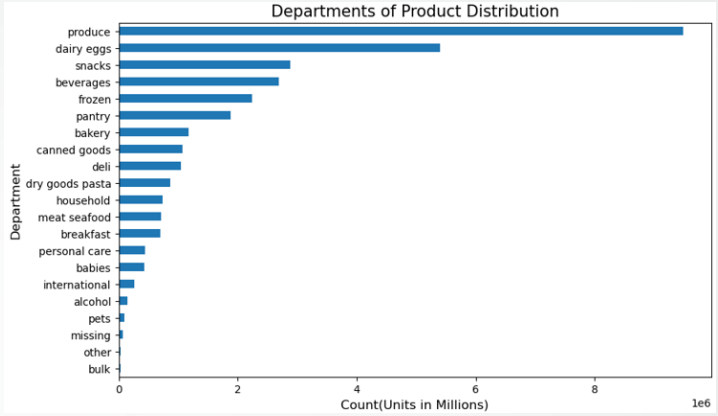
- Perform an initial data and exploratory analysis to derive insights and strategies for better segmentation.

Productivity Tools Used:

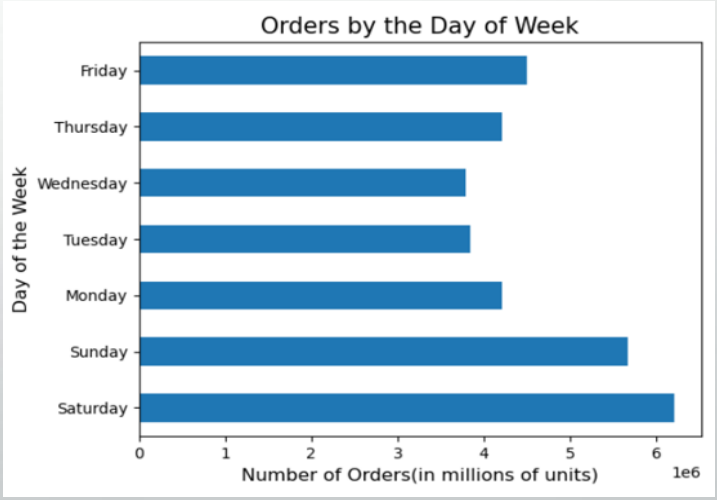
Analysis



People over the age of 40 have a higher range of income compared to those under 40.



Produce, dairy eggs, and snacks are the most popular products.



Saturday and Sunday seem to have the most orders of the week. Wednesday and Tuesday have the least orders of the week.

Insights & Recommendation

Excel Workbook

01

Saturday and **Sunday** seem to have the **most orders of the week**. Wednesday and Tuesday have the least orders of the week. To **increase sales** in the weekdays, consider having **discounted** or **promotions** on items **exclusively from Monday to Friday**.

02

Produce, dairy eggs, and **snacks** are the **most popular products**. Customers most likely use Instacart to get their produce, dairy eggs, snacks and everyday food items in general. Consider having **promotions** or **discounts** on the less popular items such as **pet items, alcohol,** and **international products**.

03

People **over the age of 40** have a higher range of income compared to those **under 40**. Consider promoting for potential customers over 40 due to their **high spending power**.

[Kaggle Dataset]
[Zillow Dataset]
[Project Brief]

Salary

Context:

- The "Salary by Job Title and Country" dataset contains salary data such as salary, job title, country, and race. The dataset can be used to analyze salary differences based on job roles, geographic locations, and experience level. It provides insights into global salary trends and can be useful for research on income disparities, workforce diversity, and regional economic conditions.

Key Skills Applied:

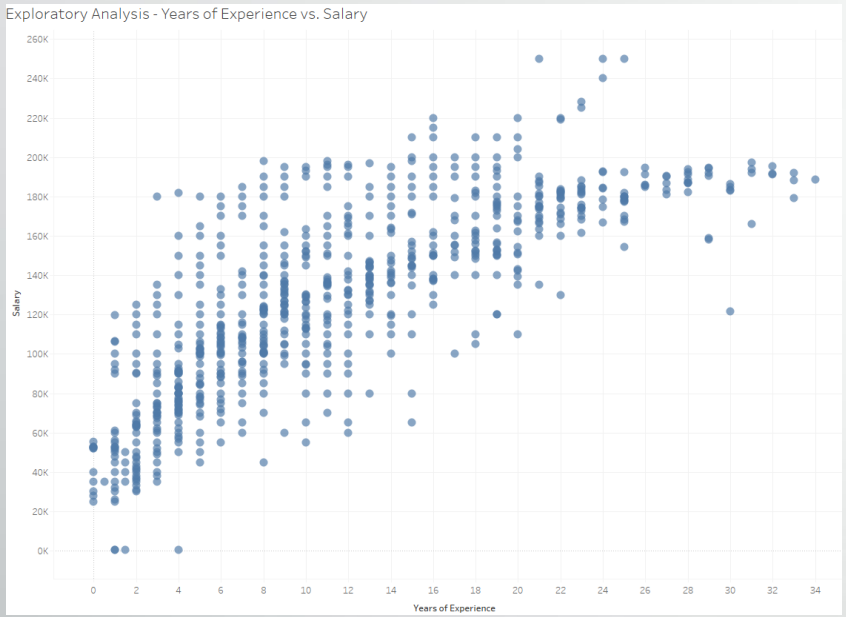
- Python(Jupyter Notebook)
- Tableau
- Sourcing Open Data
- Data Cleaning
- Exploratory Analysis
- Geospatial Analysis with a shapefile
- Regression Analysis
- Cluster analysis
- Sourcing & Analyzing Time Series Data
- Data Dashboards and Visualizations

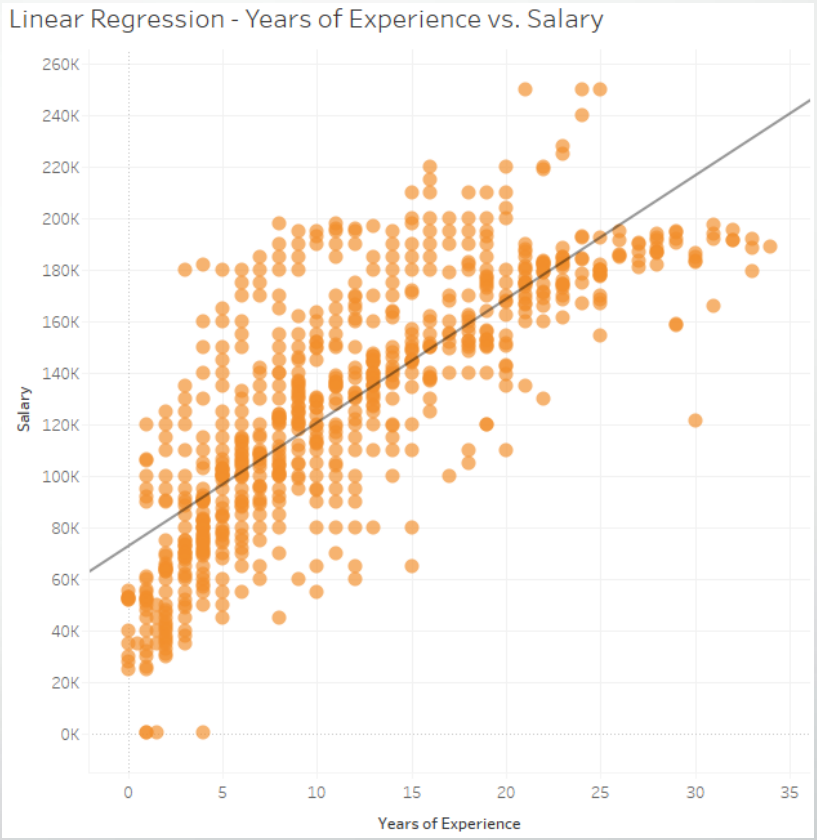
Objective:

- Conduct analysis to explore how geographic location, age, and job role affect salary differences, aiming to identify trends, inequities, and insights related to workforce diversity, income inequality, and regional economic conditions.

Productivity Tools Used:

Analysis

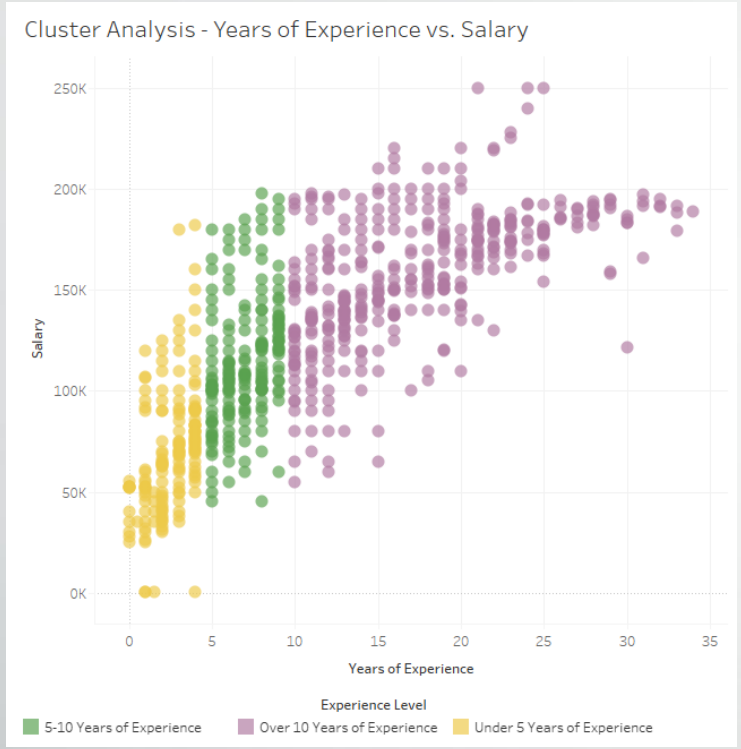




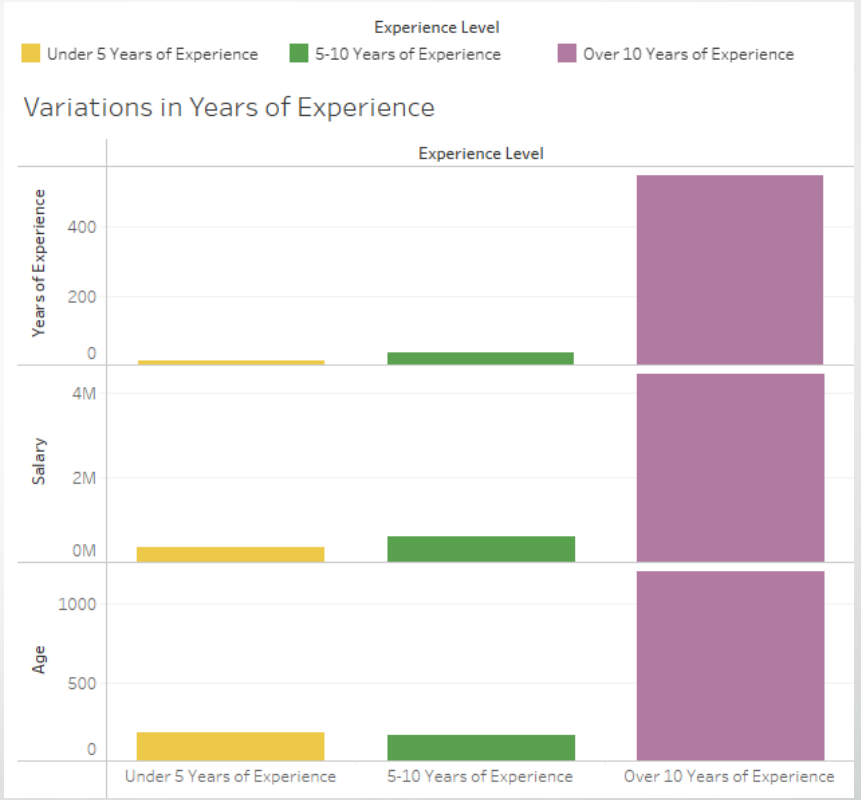
As the number of years of experience **increases,** the amount of salary compensated **rises**.

The results showed that the years of experience contributes to about **60% of the trend** in the data. The relationship between the two variables is **not entirely linear**.

Analysis



Our cluster analysis revealed three distinct experience levels, represented by the three different colors on the scatterplot: yellow, green, and purple.



The three clusters—"Under 5 Years of Experience," "5-10 Years of Experience," and "Over 10 Years of Experience"—show a clear correlation, with salary, age, and experience increasing with tenure.

Insights

GitHub Repo

Tableau Link

01

Years of Experience: This is a significant determinant, as more experience typically correlates with higher salary expectations.

02

Age: Age can also play a role, as it often reflects accumulated experience in a particular field.

03

Geographic Location: Salaries vary across countries due to differences in cost of living, local economic conditions, and government policies. Further analysis is required to fully understand these variations.

Next Steps

Sample Size:

- **Expand** the *sample size* for each job title to ensure more robust insights.

Geographic Expansion:

- Include *additional countries* to **enhance** the global perspective of salary data.

Algorithm Implementation:

- Implement a *classification algorithm* to **predict** average salaries based on available data.

Further Analysis:

- Explore additional variables, such as *race, gender, education level, etc.*, to uncover **deeper insights** into salary compensation across different countries.

Weather Datasets
Project Brief

ClimateWins

Context:

- ClimateWins is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world. ClimateWins has been sorting through hurricane predictions from The National Oceanic and Atmospheric Administration (NOAA) in the U.S., typhoon data from The Japan Meteorological Agency (JMA) in Japan, world temperatures, and a great deal of other data.

Key Skills Applied:

- Python(Jupyter Notebook)
- Data Ethics and Biases
- Hypotheses
- Data Optimization
- Supervised and Unsupervised Learning Algorithms
- Data Dashboards and Visualizations

Objective:

- Conduct analysis to explore how geographic location, age, and job role affect salary differences, aiming to identify trends, inequities, and insights related to workforce diversity, income inequality, and regional economic conditions.

Productivity Tools Used:

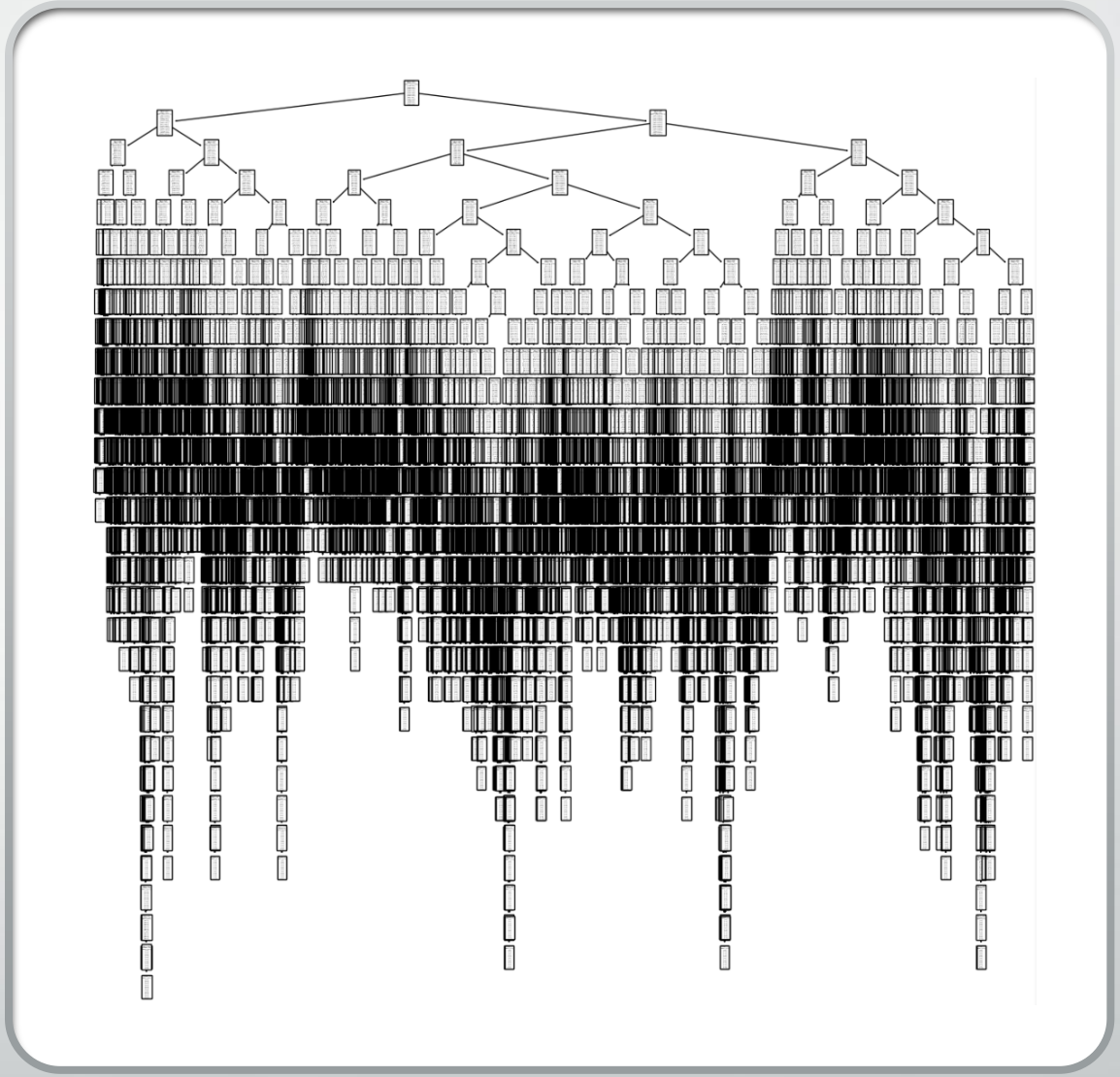
Machine Learning Algorithm: K-Nearest Neighbors

- KNN classifies data based on the majority class of its nearest neighbors.
- Used to predict pleasant weather in 15 European stations.
- Achieved 88.2% average accuracy.

Weather Station	Accurate Predictions		False Positive	False Negative	Accuracy Rate
BASEL	3907	935	431	465	84.8%
BELGRADE	3238	1502	538	460	82.6%
BUDAPEST	3416	1432	484	406	84.5%
DEBILT	4346	732	291	369	88.6%
DUSSELDORF	4167	800	340	431	86.6%
HEATHROW	4161	754	409	414	85.7%
KASSEL	4563	607	252	316	90.1%
LJUBLJANA	3726	1133	469	410	84.7%
MAASTRICHT	4249	819	313	357	88.3%
MADRID	2735	2257	433	313	87.0%
MUNCHENB	4222	766	324	426	86.9%
OSLO	4624	507	255	352	89.4%
SONNBLICK	5738	0	0	0	100%
STOCKHOLM	4449	588	317	384	87.8%
VALENTIA	5391	108	71	168	95.8%
				AVERAGE %	88.2%

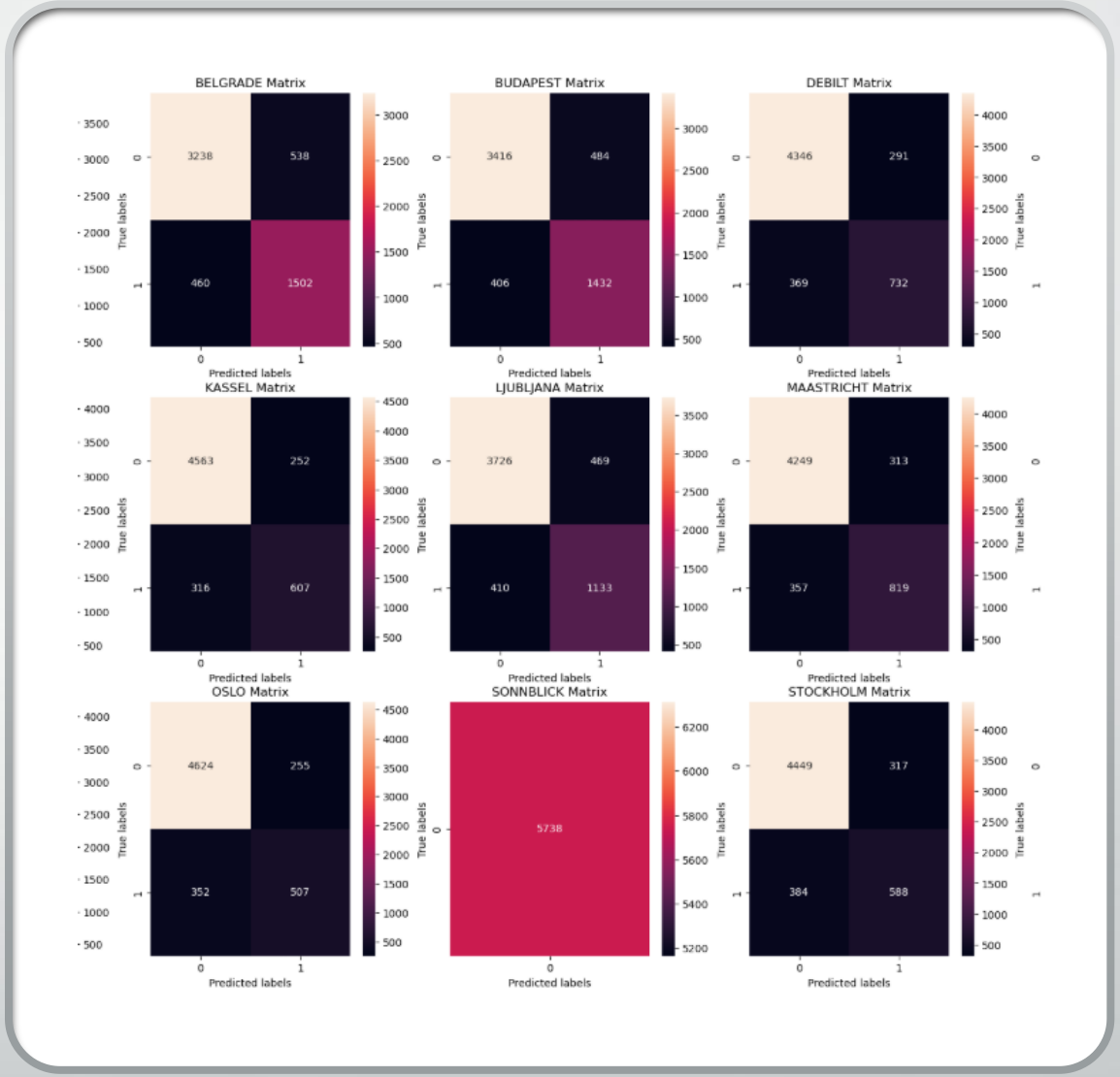
Machine Learning Algorithms: Decision Tree

- The decision tree algorithm classifies data by branching based on features, ending at a classification leaf.
- Used to predict pleasant weather in 15 European stations with 46.1% train accuracy.
- Likely overfitting; pruning may improve accuracy.



Machine Learning Algorithms: Artificial Neural Network (ANN)

- ANN mimics the brain using layers of interconnected neurons to learn patterns.
- Learns by adjusting neuron connections; optimized via parameters like iterations, layer size, and learning rate.
- Achieved a highest test accuracy of 67%.



Insights

- The decision tree model requires additional pruning to improve prediction accuracy.
- The K-Nearest Neighbors (KNN) algorithm achieved the highest accuracy at 88.2% for climate temperature prediction.
- The Artificial Neural Network (ANN) reached a test accuracy of 67% after three trials.
- The ANN model offers greater complexity in analysis compared to the simpler KNN approach.
- With more testing and the inclusion of additional variables, the ANN has the potential to outperform the other models.

PowerPoint Presentation

Next Steps

DIVERSIFY VARIABLES OTHER THAN TEMPERATURE TO FIND MORE INSIGHTS AND TRENDS.

INCORPORATE MORE WEATHER STATIONS TO EXPAND MACHINE LEARNING CAPABILITIES.

CONTINUE TO RUN FURTHER TESTS AND MAKE ADJUSTMENTS AND IMPROVEMENTS TO BOLSTER MODEL ACCURACIES.

Weather Datasets
Project Brief

ClimateWins Part II

Context:

- As climate change advances, weather patterns around the world are becoming increasingly unpredictable. This shift has resulted in a surge of extreme weather events, which pose growing risks to the safety of communities.

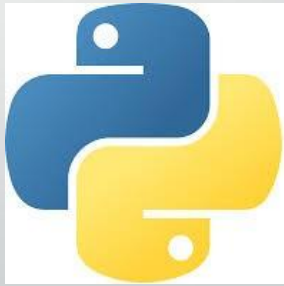
Key Skills Applied:

- Python(Jupyter Notebook)
- Evaluating Hyperparameters
- Complex Machine Learning Algorithms and Keras
- Unsupervised Learning Algorithms
- Visual Applications of Machine Learning

Objective:

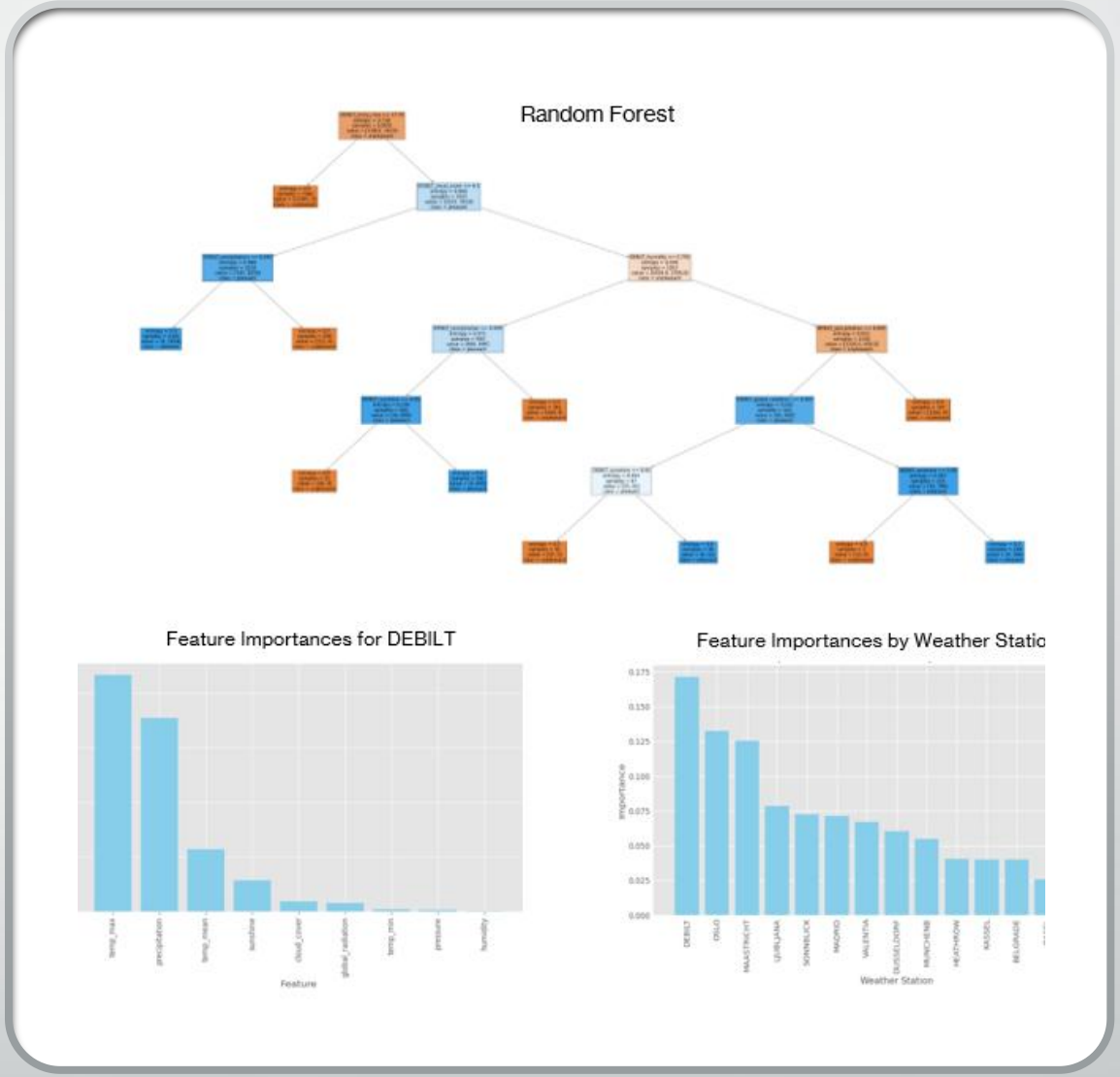
- Analyzing weather changes over the past 60 years to identify unusual patterns in Europe, assess whether such anomalies are becoming more frequent, project future climate conditions over the next 25 to 50 years, and determine the safest regions for habitation based on these trends.

Productivity Tools Used:



Thought Experiment #1: Classification of Unusual Weather

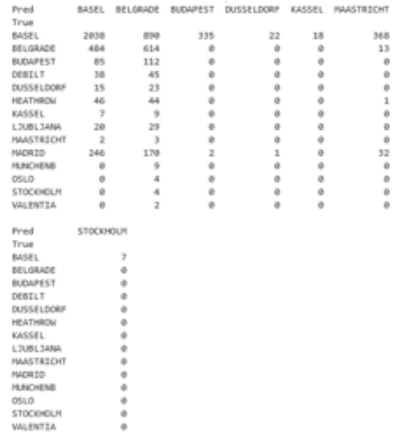
- **Hypothesis:** Random Forest can detect unusual weather trends using historical data.
- **Method:** Used RandomForestClassifier with temperature and precipitation data; optimized with RandomizedSearchCV.
- **Result:** Accuracy improved from *99.6%* to *100%*; effectively identified regions with increasing climate anomalies.



Thought Experiment #2: Deep Learning for Weather using Images

- **Hypothesis:** CNNs can enhance the interpretation of radar and satellite imagery leading to more accurate weather trend predictions.
- **Model Used**: Convolutional Neural Network (CNN) with Bayesian optimization.
- **Task**: Classified radar images into weather conditions (e.g., cloudy, rainy, sunny).
- **Method**: Tuned hyperparameters (neurons, batch size, learning rate) using Bayesian optimization.
- **Results**:
  - Accuracy improved from *47.83%* to *92.4%* after optimization.
  - Confusion matrix showed better distinction between similar classes (e.g., 'shine' vs. 'sunrise').
  - Demonstrated strong potential for analyzing complex weather image data.

CNN before Optimization

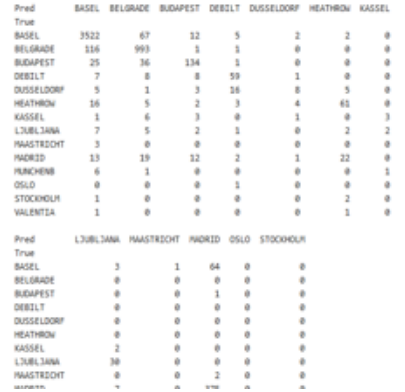


Pred / True	BASEL	BELGRADE	BUDAPEST	DUSSELDORF	KASSEL	MAASTRICHT
BASEL	2038	890	335	22	18	368
BELGRADE	484	614	0	0	0	13
BUDAPEST	85	112	0	0	0	0
DEBILT	38	45	0	0	0	0
DUSSELDORF	15	23	0	0	0	0
HEATHROW	46	44	0	0	0	1
KASSEL	7	9	0	0	0	0
LJUBLJANA	20	29	0	0	0	0
MAASTRICHT	2	3	0	0	0	0
MADRID	246	170	2	1	0	32
MUNCHENB	0	9	0	0	0	0
OSLO	0	4	0	0	0	0
STOCKHOLM	0	4	0	0	0	0
VALENTIA	0	2	0	0	0	0

Pred / True	STOCKHOLM
BASEL	7
BELGRADE	0
BUDAPEST	0
DEBILT	0
DUSSELDORF	0
HEATHROW	0
KASSEL	0
LJUBLJANA	0
MAASTRICHT	0
MADRID	0
MUNCHENB	0
OSLO	0
STOCKHOLM	0
VALENTIA	0

```
38/38 - 1s - 38ms/step - accuracy: 0.9240 - loss: 0.2204
```

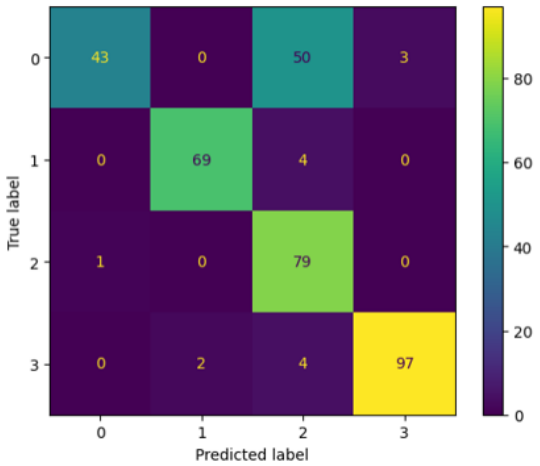
CNN after Optimization



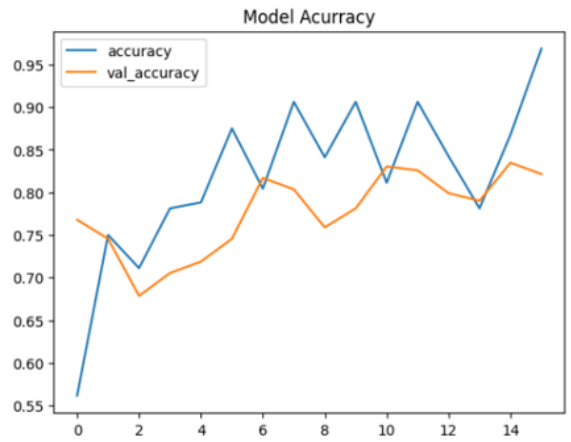
Pred / True	BASEL	BELGRADE	BUDAPEST	DEBILT	DUSSELDORF	HEATHROW	KASSEL
BASEL	3522	67	12	5	2	2	0
BELGRADE	116	993	1	1	0	0	0
BUDAPEST	25	36	134	1	0	0	0
DEBILT	7	8	8	59	1	0	0
DUSSELDORF	5	1	3	16	8	5	0
HEATHROW	16	5	2	3	4	61	0
KASSEL	1	6	3	0	1	0	3
LJUBLJANA	7	5	2	1	0	2	2
MAASTRICHT	3	0	0	0	0	0	0
MADRID	13	19	12	2	1	22	0
MUNCHENB	6	1	0	0	0	0	1
OSLO	0	0	0	1	0	0	0
STOCKHOLM	1	0	0	0	0	2	0
VALENTIA	1	0	0	0	0	1	0

Pred / True	LJUBLJANA	MAASTRICHT	MADRID	OSLO	STOCKHOLM
BASEL	3	1	64	0	0
BELGRADE	0	0	0	0	0
BUDAPEST	0	0	1	0	0
DEBILT	0	0	0	0	0
DUSSELDORF	0	0	0	0	0
HEATHROW	0	0	0	0	0
KASSEL	2	0	0	0	0
LJUBLJANA	30	0	0	0	0
MAASTRICHT	0	0	2	0	0

Thought Experiment # 2: CNN Model Performance



Accuracy: 0.96875, Val_Accuracy: 0.8214285969734192
Loss: 0.026281258091330528, Val_Loss: 0.0785900354385376



Confusion Matrix:

Illustrates the model's performance in classifying weather types. While most predictions are correct, some categories are occasionally misclassified.

Model Accuracy:

Monitors accuracy across epochs. Higher values reflect better performance, while fluctuations indicate changes in model stability.

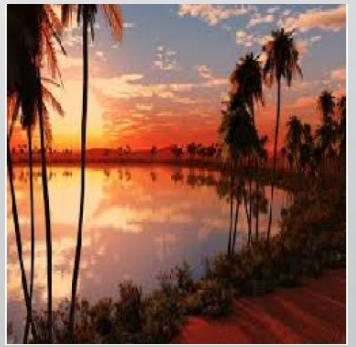
Model Loss:

Shows training and validation loss. Lower loss indicates a better fit, while spikes highlight areas that need improvement.

Thought Experiment #3: Synthetic Weather Projections using GANs to Improve Predictions

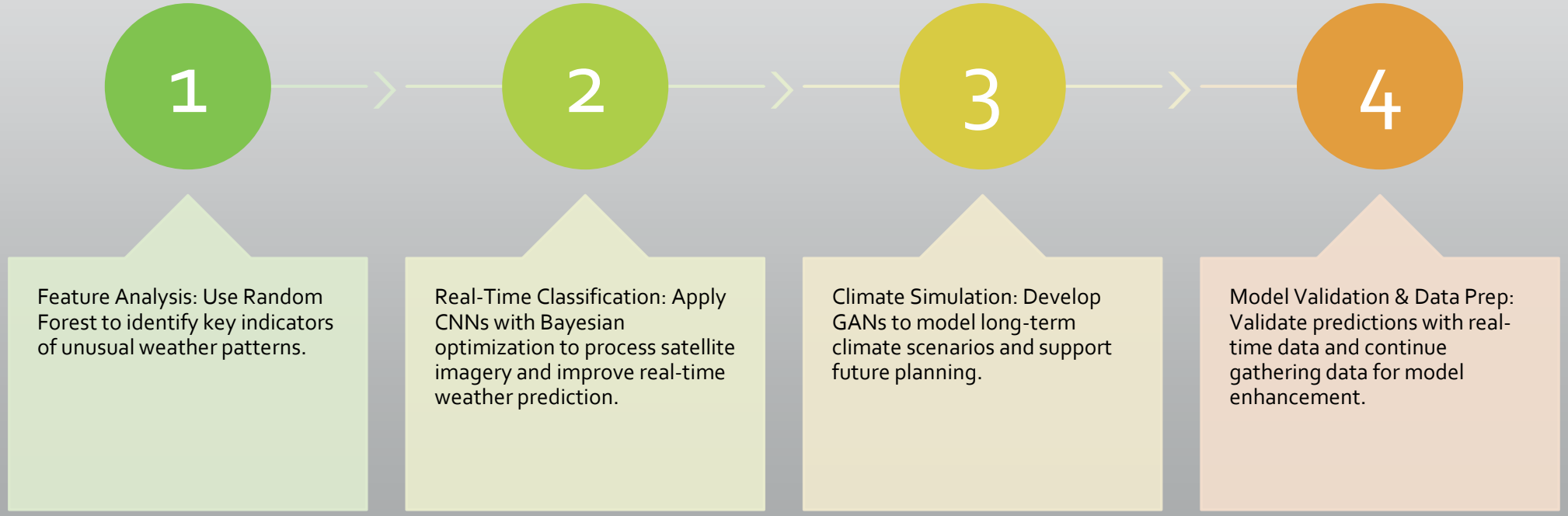
- Hypothesis:
  - GANs can generate realistic weather scenarios based on current trends, offering a range of possible future outcomes.
- Approach:
  - Used GANs to simulate temperature and precipitation changes.
  - Generated synthetic weather maps for long-term projections (25–50 years).
- Results:
  - Produced visualizations of potential future climate conditions.
  - Helps stakeholders plan for and adapt to extreme weather events.

Incorrect Prediction - class: Cloudy - predicted: Shine



Correct Prediction - class: Sunrise - predicted: Sunrise

Insights & Next Steps



Thank you!

Connect with me:

andycen7@gmail.com

LinkedIn

GitHub

(415) 806-3581