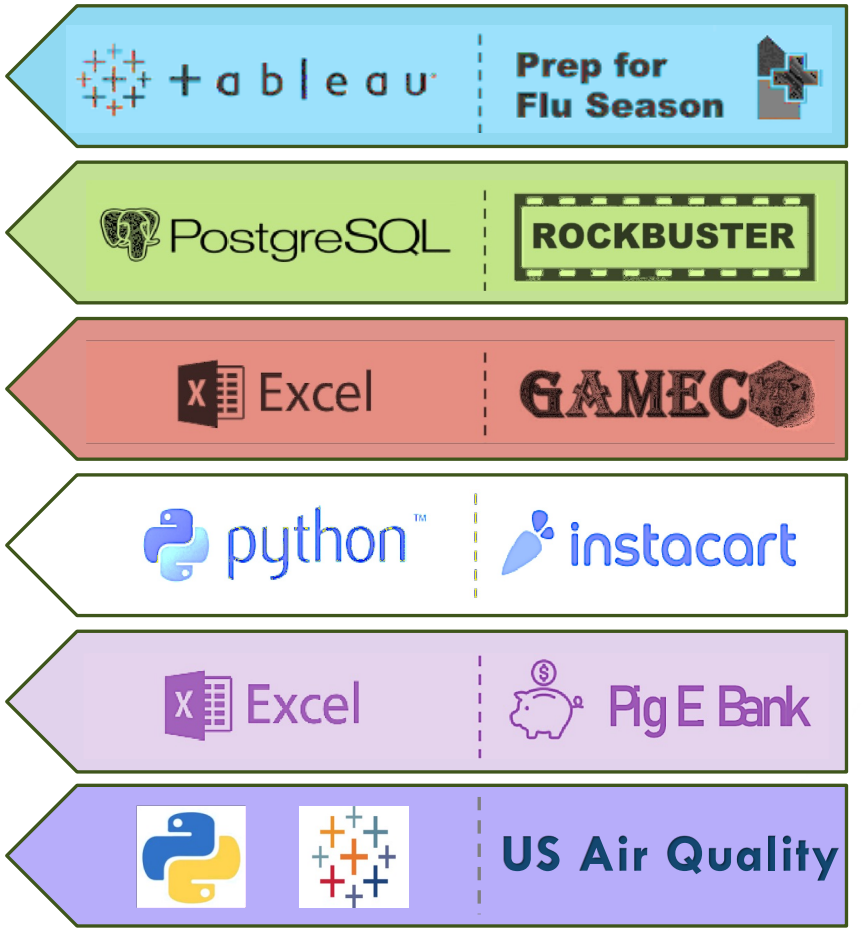
Shaili Oza

Data Analyst Portfolio



About Me

I am a multilingual data analyst with a strong background in medicine. My passion lies at the intersection of data and healthcare, where I utilize my diverse skill set to uncover valuable insights that can drive meaningful improvements in patient care and overall healthcare systems.

I am adept at extracting, analysing, and communicating complex data from various sources, enabling data-driven decision-making. Through my analytical prowess, belief in continuous learning and medical expertise, I aim to contribute to the advancement of healthcare practices and create a positive impact on people's lives.

Through this portfolio, I humbly aim to share my efforts in harnessing the power of data, using multiple analytics tools and ultimately, the well-being of individuals worldwide.

Case Study:

Preparing for Influenza Season

Company

The Medical Staffing Agency provides temporary medical staff to hospitals and clinics around the United States to aid in the fight against influenza.

Context

There have been an average of 63,229 deaths in the US due to Influenza from 2009-2017. The goal is to help the agency determine where and how much staff to allocate for the upcoming influenza season.

Goals

- What is the vulnerable age group showing highest mortality rate?
- Which region are in highest need of assistance?
- Which months have peak influenza season?

Disclaimer

All data in this project is fictional. Insights about sales and popularity of films should not be extrapolated in real world.

The Process

Preparation

- Translated business requirements into a project plan.
- Cleaned data for accuracy, consistency and clarity.
- Transformed data (grouping, filtering, sorting, transposing, etc.).
- Merged data from multiple sources to get desired outcomes.
- Derived new variables from existing data.

Tools used: MS Word, Excel

Analysis

- Performed visual analysis using histograms, box and whiskers charts, choropleth maps and scatterplots.
- Checked for correlation between variables using linear regression.
- Performed statistical hypothesis testing (two sample t-test).
- Forecasted where and when extra staffing will be needed.

Tools used: Excel, Tableau

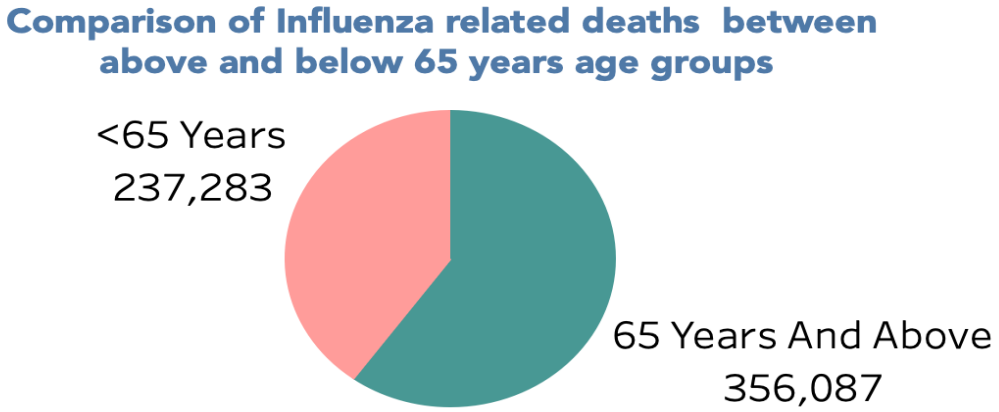
Visualization

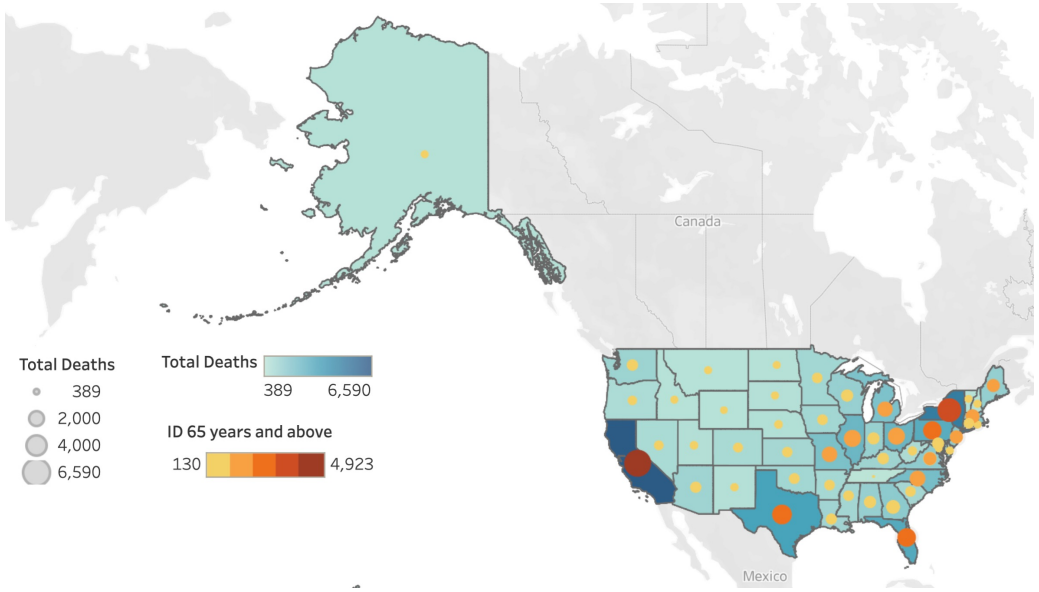
- Used data storytelling principles to design visualizations.
- Created an interactive Tableau dashboard that allowed stakeholders to filter and explore data by state or year.
- Presented findings and recommendations via recorded video presentation.

Tools used: Tableau, PowerPoint

The Results

- Statistical hypothesis testing confirmed that those aged 65+ are significantly more likely to die of the flu than those under 65 (p-value= 5.7x10-157).
- The density of elderly population was also the highest in states with highest number of cases. We narrowed maximum cases to top five states with highest mortality.
- Maximum deaths were observed during the winter months from December to March.
- These factors, along with each state's population size, were combined to create a model to predict the number of deaths, by state, for the next flu season.





Our model predicts which states will require maximum assistance and at which time of the year in order to help the staffing agency deploy healthcare workers.

Links for Further Exploration

- Tableau Interactive Dashboard
- Video Presentation of entire project
- Project Management Plan
- Interim Report
- Original Dataset Link

Case Study:

Rockbuster Stealth Data Analysis

Company

Rockbuster Stealth is a movie rental company with stores and customers around the world.

Disclaimer

All data in this project is fictional. Insights about sales and popularity of films should not be extrapolated in real world.

Context

The management team is planning to use its existing movie licenses to launch an online video rental service in order to stay competitive market of services such as Netflix and Amazon Prime.

Project Goals

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

The Process

Preparation

- Loaded data into a relational database management system.
- Cleaned data as necessary.
- Extracted entity relationship diagram (ERD).
- Created a data dictionary documenting the relationships between tables, keys, data types and other details.

Tools used: DbAnalyzer, PostgreSQL, Word

Analysis

Wrote queries that

- Joined tables
- Filtered, grouped, and ordered data
- Applied aggregation functions

Revised queries to more efficiently use

- Subqueries
- Common Table Expressions (CTE)

Tools used: PostgreSQL, PGAdmin

Visualization

- Used results of queries to design appropriate visualizations for each key business question.
- Supplied a PowerPoint presentation summarizing key findings.

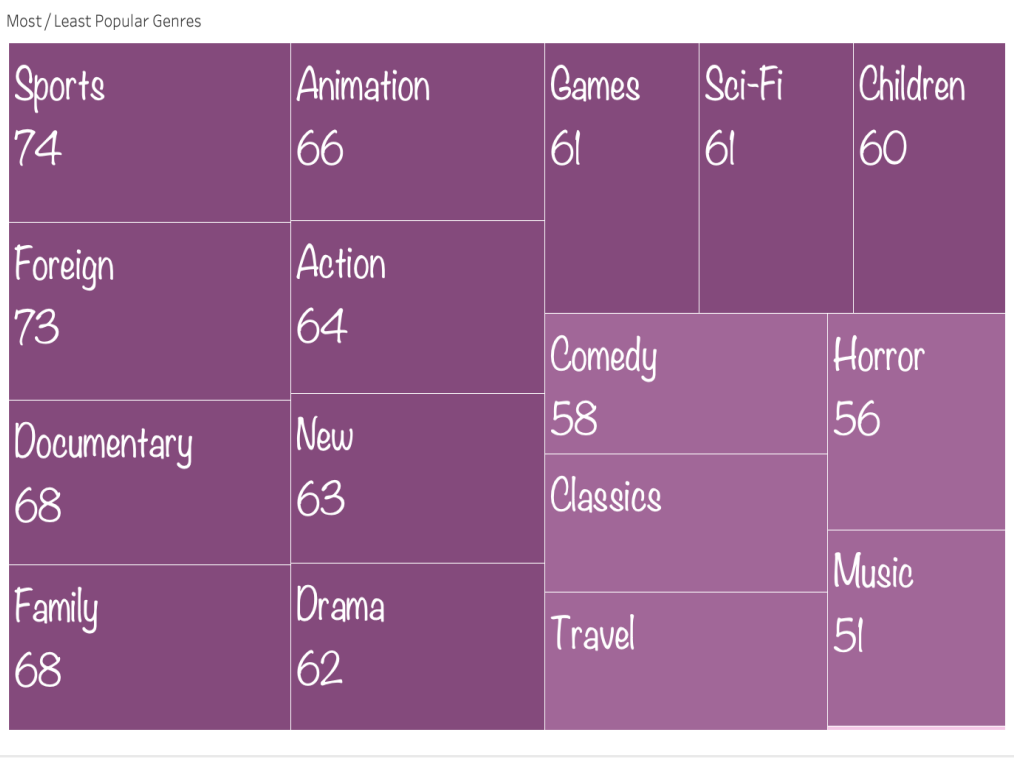
Tools used: Tableau, PowerPoint

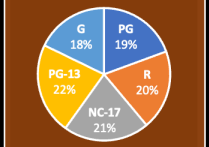
The Results

- India and China are two major contributors to the company's revenue.
- Top genres are sports, foreign, documentary, family and animation
- Analysis also shows which customers have high lifetime value.

Top 10 countries by customer count and revenue generated

Number of Customers	Revenue Generated	Country
1422	6035	India
1297	5251	China
869	3685	United States
749	3123	Japan
718	2985	Mexico
681	2919	Brazil
638	2766	Russian Federation
530	2220	Philippines
351	1498	Turkey
331	1353	Indonesia



# of Movies in our Catalog: **1000**	Release Year: **2006**	Language(s): **English**	Breakdown of Ratings: 	Longest Film(s) **185 min.** Average Film(s) **115 min.** Shortest Film(s) **46 min.**
# of Active Customers: **584**	Countries: **109**	# of stores: **2**	Rental Rate Maximum: $4.99/day Average: $2.98/day Minimum: $0.99/day	Rental Duration: Maximum: **7 Days** Average: **5 Days** Minimum: **3 Days**

Links for Further Exploration

- Rockbuster PowerPoint Presentation
- Rockbuster Stealth LLC Data Dictionary
- SQL Queries and Documentation in GitHub
- Rockbuster Original Dataset

Case Study:

GameCo Market Analysis

Company

GameCo sells a wide variety of video games around the world. In 2008, they created marketing teams to target different channels for purchasing games – for example, purchasing packaged games at a brick-and-mortar store versus buying via digital download or online subscription.

Context

It's October 2016, and GameCo's executives are looking for current insights about the sales of packaged video games in order to better plan their marketing strategy for 2017.

Goals

- Understand overall global and regional trends in sales.
- Detect regional markets with quickest growth.
- Look at top genre, publishers, and platforms by region.
- Develop insight on the current market and give recommendation.

Disclaimer

All data in this project is fictional. Insights about sales and popularity of films should not be extrapolated in real world.

The Process

Preparation

- Cleaned data for accuracy, consistency and clarity.
- Documented cleaning process and how inconsistent data had been addressed.
- Created new variables to aid in finding insights.

Tools used: Excel, Word

Analysis

- Performed descriptive analysis of all key variables, including exploring the shape and spread of data via histograms and scatterplots.
- Used pivot tables to group, filter, and sort data to find insights.

Tools used: Excel

Visualization

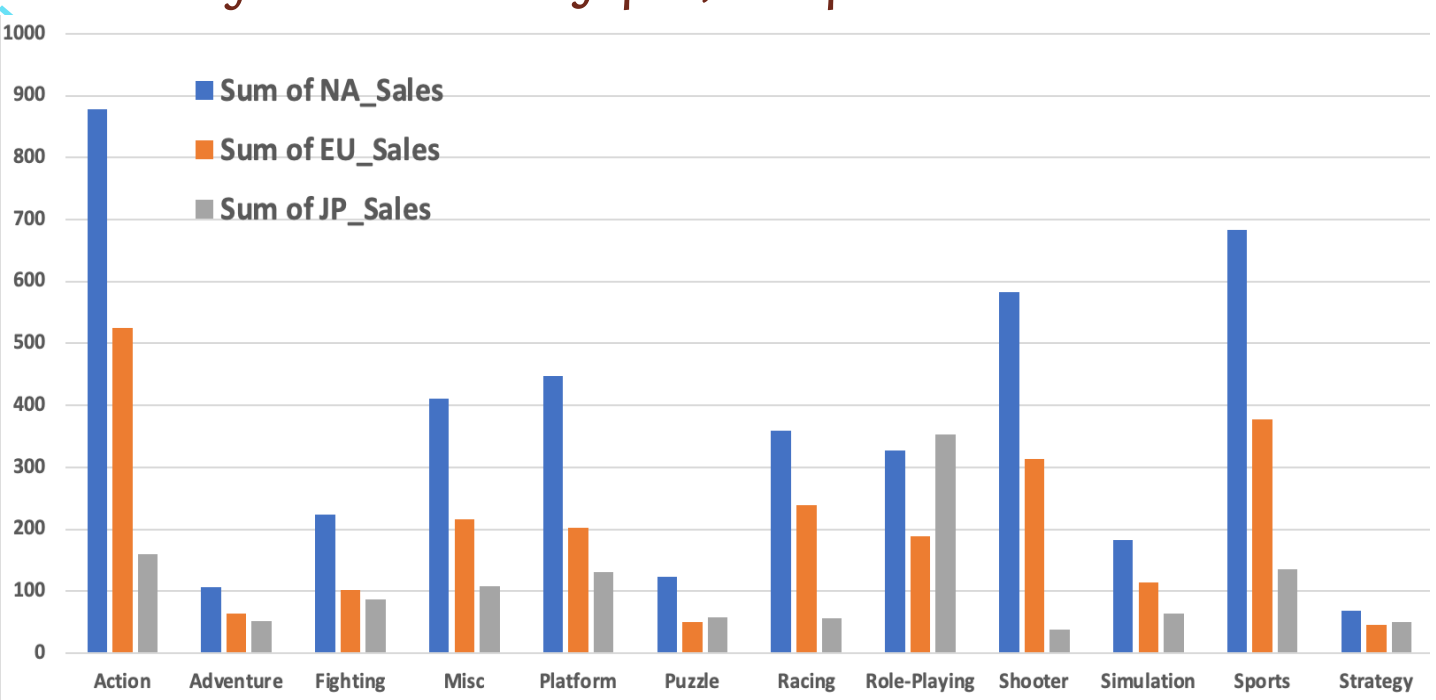
- Designed visualizations that would clearly and intuitively deliver information to executives.
- Created a presentation that not only answered key questions but framed insights within our company's narrative to make their significance easier to understand.

Tools used: Excel, PowerPoint

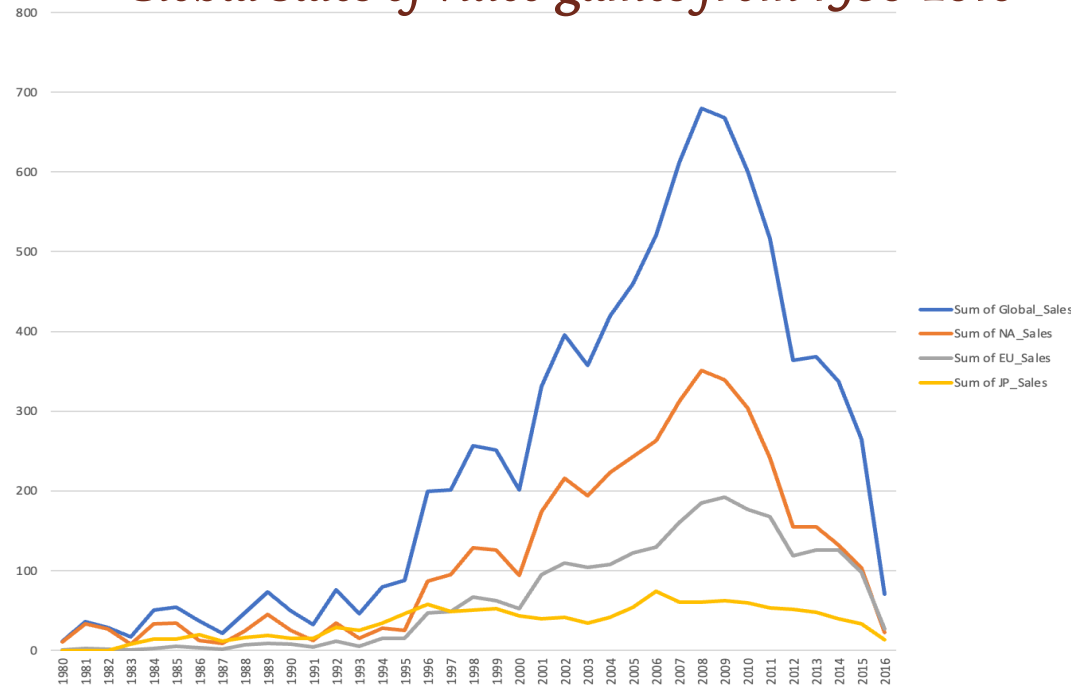
The Results

- Most popular genres are action, sports, shooter, and platform video games.
- North America has been the biggest contributor in the Global Sales of video games among the three regions.
- Japan sales have been consistently low compared to the other regions and yet, after 2009, they have reduced even further.
- Sudden crash in overall sales in the year 2016 may be due to affecting factors like economy, global crisis, supply chain issues; which need further evaluation.

Preferred Genres in Japan, Europe and North America



Global Sales of video games from 1980-2016



Links for Further Exploration

- GameCo PowerPoint Presentation (PDF format)
- GameCo Original Dataset

Case Study:

Instacart Grocery Basket Analysis

Company

Instacart is an online delivery company that operates a grocery delivery and pick-up service in the United States and Canada. The users can order groceries via the Instacart app and have those groceries delivered directly to their home.

Context

Instacart executives want to explore the data they have about their customers and sales patterns to find insights that will allow them to create targeted marketing campaigns for specific segments of their user base as well as promotions to increase usage during non-peak hours.

Goals

- When are our busiest times of day/week?
- Which types of products bring in the most revenue?
- How does ordering behaviour vary based on brand loyalty?
- What profile groups should get targeted marketing?

The Process

Preparation

- . Cleaned data for accuracy, consistency and clarity.
- Wrangled data (adjusting datatypes, transposing, etc.)
- Merged data frames.
- Derived new variables from existing data.
- Documented changes and population flows.

Tools used: Python

Analysis

- Filtered, sorted, grouped, and aggregated data to answer key questions.
- Created new profiles/flags to enable marketing to specific segments within the customer base.
- Performed exploratory analysis on each profile/flag to uncover additional insights.

Tools used: Python

Visualization

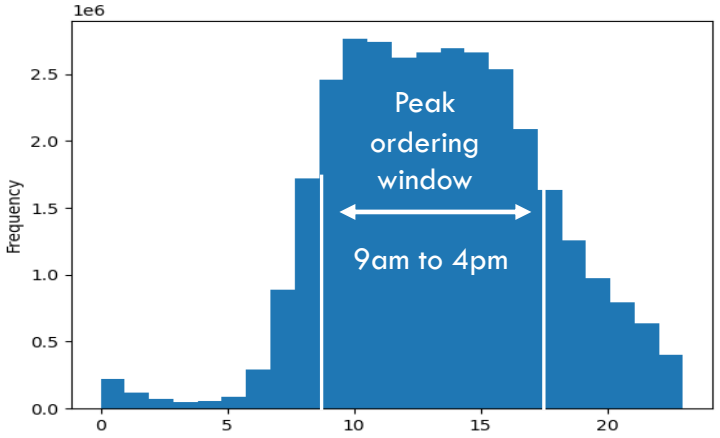
- Designed visualizations that would clearly and intuitively deliver information to executives.
- Created a presentation that not only answered key questions but framed insights within our company's narrative to make their significance easier to understand.

Tools used: Python

Python coding was done in Jupyter Notebooks using Pandas, Numpy, Matplotlib and Seaborn.

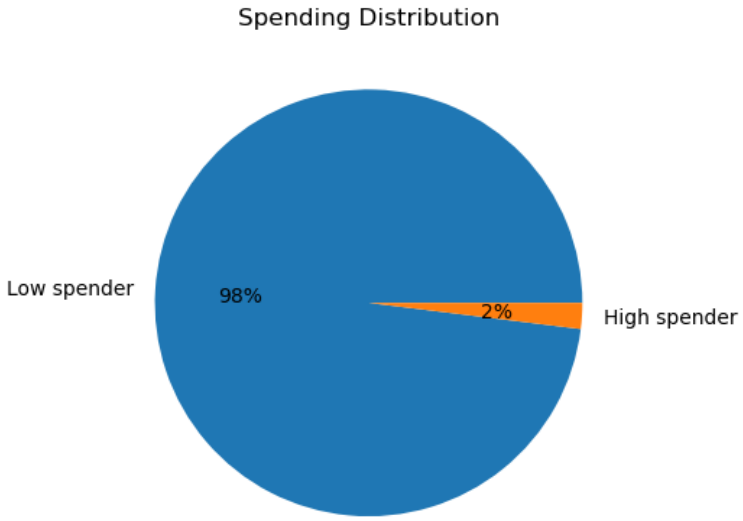
The Results

- The regular customers appear to be the most frequent consumers. The loyal customers are typically also frequent customers.
- Since majority customers are low spenders, marketing team must target them for advertisements.
- Marketing team must focus efforts toward families as they spend most compared to other people.
- Advertise those departments which show potential to increase sales, such as alcohol, personal care and pets.

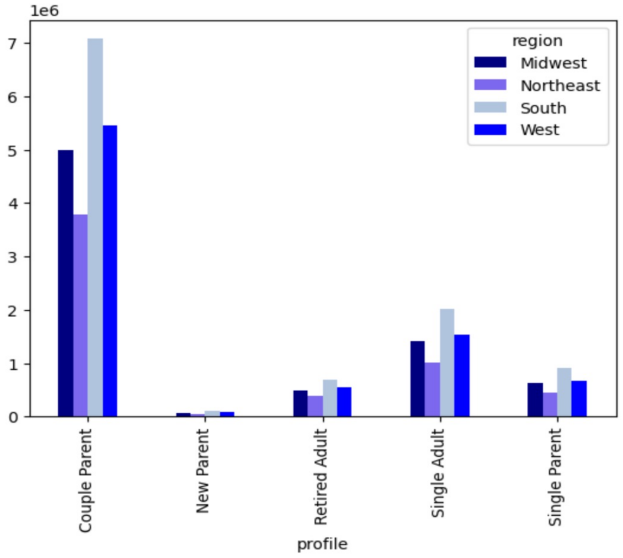


Peak ordering window is 9am to 4pm

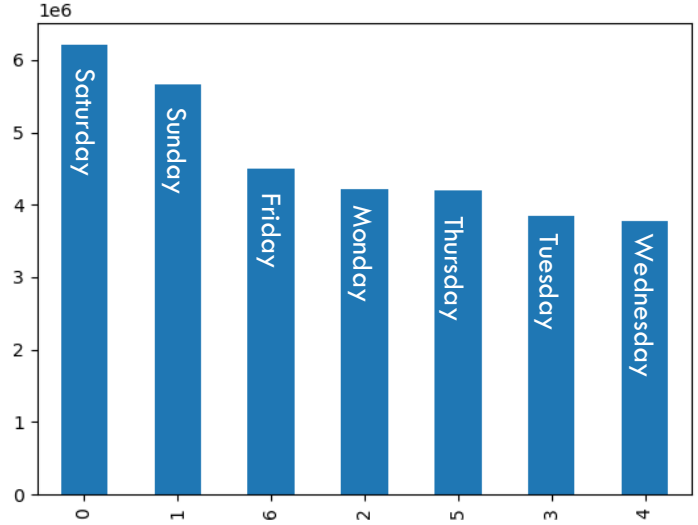
98% customers are low spenders



Majority of customers are coupled parents and maximum sales happen in southern US



Most sales occur on Saturday and Sunday



Links for Further Exploration

- GitHub Repository with full project
- Jupyter Notebooks with Python Code
- Final Report in Excel
- Original Datasets

Excel Pig E Bank

Case Study:

Indicators Why A Client May Leave

Company

Pig E. Bank is a global bank.

Company

All data in this project is fictional. Insights should not be extrapolated in the real world.

Context

The Customer Retention team wants to be able to identify when a client is considering leaving the bank before they actual exit the bank community. This would allow personal banking representatives an opportunity to check in with those clients and repair/retain the relationship.

Goals

- Identify leading indicators that a client is considering leaving the bank.
- Create a decision tree that can be used to assign a risk level to each client based on the probability of them exiting the bank community.

The Process

Preparation

- Extracted only needed data from the data source, preventing personally identifiable data from being included in the analysis.
- Cleaned data for accuracy, consistency and clarity.
- Use pivot tables to explore data and identify potential leading indicators.

Tools used: Excel

Analysis

- Calculated descriptive statistics on each variable for both current and former clients.
- Performed two-sample t- tests to check the significance of different means with quantitative variables.

Tools used: Python

Visualization

- Created a decision tree model with leading indicators sequenced from most influential to least.
- Provided documentation for reasoning behind each variables inclusion (or exclusion) from the model.

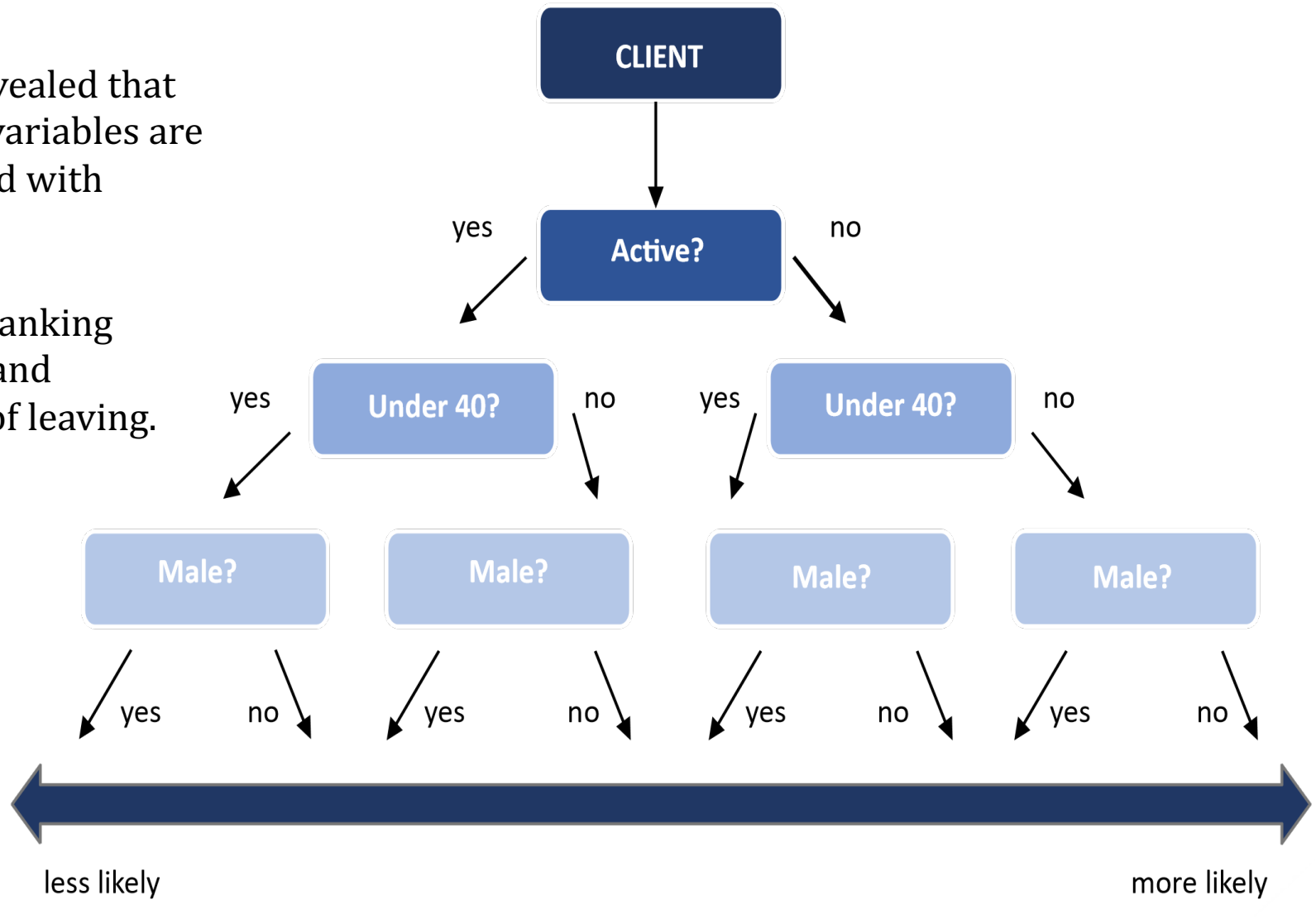
Tools used: Python

Excel Pig E Bank

The Results

Analysis of client data revealed that age, activity and gender variables are more commonly observed with clients exiting the bank.

The decision tree helps banking representatives identify and prioritize clients at-risk of leaving.



Excel Pig E Bank

Links for Further Exploration

Analysis of leading indicators why customers leave the bank

Case Study:

US Air Pollution Historical Data Analysis

Company

The US Environmental Protection Agency monitors air quality index in country and provides necessary instructions and precautions when necessary.

Context

Air Pollution is a major contributor to the Greenhouse Effect and is the cause of various respiratory illnesses in the country.

Goals

- What are the causes of specific air pollutants?
- Which regions have highest concentrations of air pollutants?
- Do these levels show seasonality or trend based on current events?

Disclaimer

All data in this project is real. Insights about pollutant concentrations and index can be extrapolated in real world.

The Process

Preparation

- Sourced data from US EPA.
- Documented cleaning process and how inconsistent data had been addressed.
- Created new variables to aid in finding insights.
- Removed unnecessary columns irrelevant to analysis.

Tools used: Python

Analysis

- Created new categories to analyse current air quality scenario.
- Created geographic visualizations.
- Performed linear regression, clustering.
- Analysed time series data.

Tools used: Python, Tableau

Visualization

- Designed visualizations that would clearly and intuitively deliver information to executives.
- Created a dashboard story that not only answers key questions but frames insights within the company's narrative to make their significance easier to understand.

Tools used: Python, Tableau

The Process, Explained

1. Data Cleaning and Preparation (Python): Cleaned and pre-processed the data using Python libraries such as Pandas and NumPy. Handled missing values, removed duplicates, and formatted variables for further analysis.

2. Feature Engineering (Python): Created additional features from the air pollution dataset, such as aggregating pollutant concentrations over specific time periods or extracting temporal features like day of the week or month.

3. Machine Learning Model Selection (Python): Chose appropriate machine learning models for air pollution analysis, such as regression (e.g., Linear Regression, clustering) or time series models.

4. Model Training and Evaluation (Python): Split the data into training and testing sets. Trained the selected machine learning models on the training set and evaluated their performance on the testing set using suitable metrics (e.g., Mean Squared Error, R-squared).

5. Model Tuning and Validation (Python): Fine-tuned the chosen models by adjusting hyperparameters and validated their performance using techniques like cross-validation or time series cross-validation to ensure robustness.

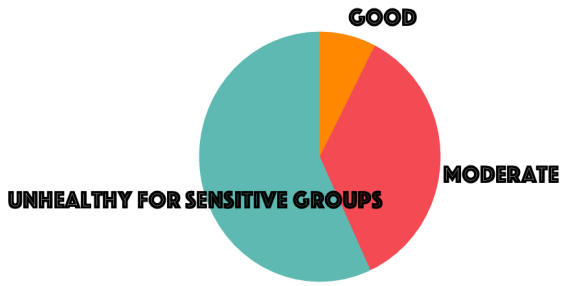
6. Prediction and Forecasting (Tableau): Utilized the trained models to make predictions and forecast future air pollution levels based on the historical data.

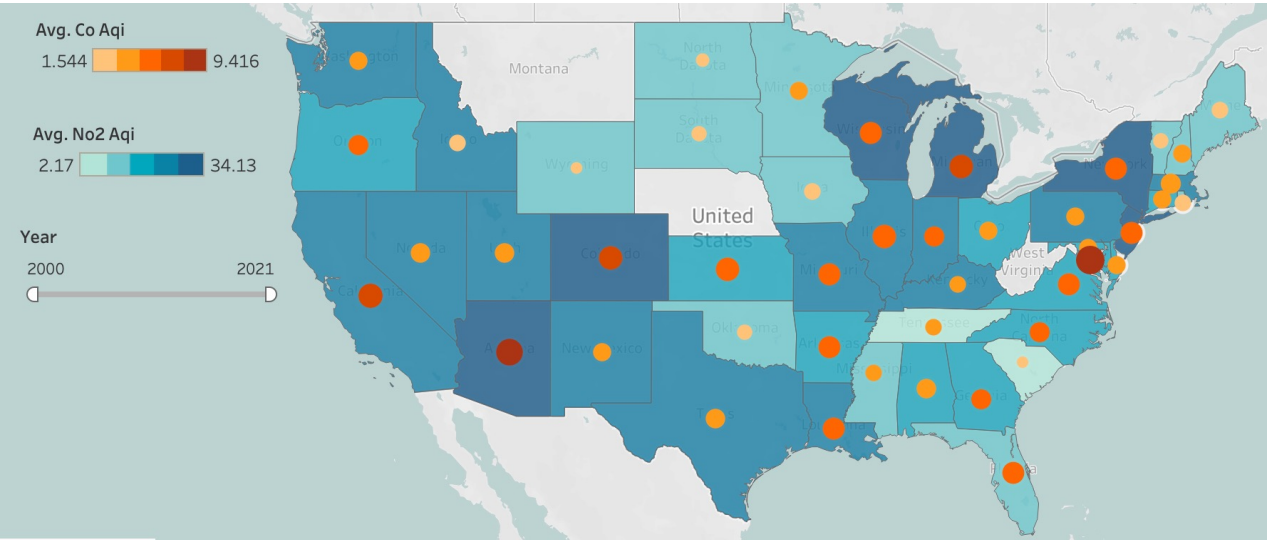
7. Visualization (Python, Tableau): Created visualizations of the predicted and actual air pollution levels using Python libraries such as Matplotlib or Seaborn to illustrate the model's performance and provide insights.

The Results

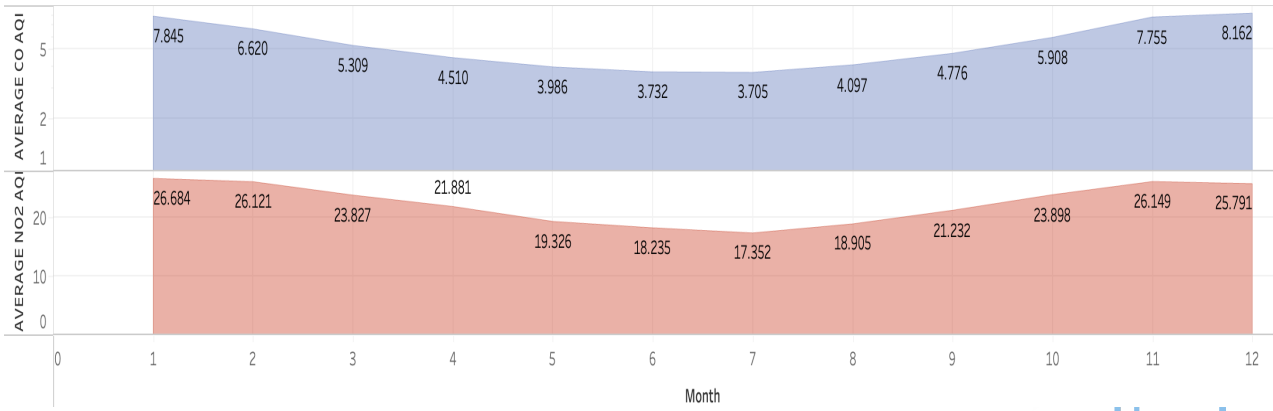
- Top states with high concentrations of pollutants coincide.
- States with maximum pollution are: Arizona, Michigan, Colorado and Alaska.
- Pollutants show significant decrease over the course of two decades.
- There does not seem to be any change in specific gas AQIs during the COVID- 19 pandemic.
- CO and NO2 show decrease in summer months, while they increase in winter months
- Sulphur dioxide show opposite trend. Increase in summer and decrease in winter (probably due to low temperatures).
- Data on deaths due to respiratory diseases is required to analyse effect of pollutants on health.

NO2 AQI Category over 2 Decades





CO and NO2 Seasonal Trend



Links for Further Exploration

- Github Repository
- Tableau Story
- Data Source

Let's connect!

shaili1994@gmail.com

https://www.linkedin.com/in/shaili-oza-690916120/

https://github.com/Shaili612

https://public.tableau.com/app/profile/shaili.oza

Excel

- Pivot Tables
- Filtering, Sorting, Grouping
- Vlookup
- Formulas
- Hypothesis Testing
- Visualizations
- Mail Merge w/ Word

PostgreSQL

- Filtering, Grouping, Ordering
- Joins
- Common Table Expressions

python

- Data Cleaning & Wrangling
- Merging
- Filtering, Sorting, Grouping, and Aggregating.
- Deriving new variables
- Visualizations

tableau

- Filtering, Sorting, Grouping
- Deriving new variables
- Data Storytelling
- Dashboards

GitHub

Photoshop Elements & Premiere Elements

Adobe

Notion

Office

jupyter matplotlib pandas NumPy

G Suite