SUSAN WANG

Data Analytics Portfolio

PROJECTS

SPOTIFY ANALYSIS	*Python*
ROCKBUSTER	*SQL*
CDC INFLUENZA	*Excel & Tableau*
INSTACART	*Python*
CLIMATEWINS	*Machine Learning*

Spotify Music Analysis

Exploratory and Predictive Analytics

Overview

Context:

Spotify is one of the most popular music streaming apps today. There is ample data collected on artists and songs on its platform. An exploration of the top hits over two decades aims to discover trends and patterns of the most popular songs.

Key Questions:

- Where do most popular songs come from?
- Which audio features define a top hit?
- How has the music changed over the years?
- Can we predict how popular music sounds in the future?

Data:	Spotify Top Hit Playlist 2000 – 2023, Music Artists Popularity Data Set
Skills:	Data cleaning and wrangling, Exploratory analysis, Machine learning models, Dashboard creation
Tools:	Python, Tableau

Process:

- Preparing the data – cleaning and wrangling
- Exploratory visual analysis – finding correlations
- Regression analysis – testing a hypothesis
- Time series analysis – testing for stationarity
- Geospatial analysis – visual insights through mapping

Sourcing and Preparing Data

Primary Data Set:

- Sourced from Kaggle
- Top 100 hits per year on Spotify from 2000 – 2023
- 23 variables, including audio features such as *danceability, energy, key, mode, loudness, duration, tempo, valence, acousticness and danceability*
- Collected through Spotify API

Secondary Data Set:

- Sourced from Kaggle
- Data on more than 1.4 million artists
- Variables on artists, including *name, country, tags, and popularity*
- Collected from the MusicBrainz database and webscraping last.fm

Merged Data Set:

1. Data checked for missing values and duplicates
2. Wrangling procedure to prepare for merge
3. Merge on Artists' Name
4. Missing values in 'country' variable filled using assistance from ChatGPT

Exploratory Visual Analysis

1

Relevant variables were placed in a **correlation matrix heatmap** with the following results:

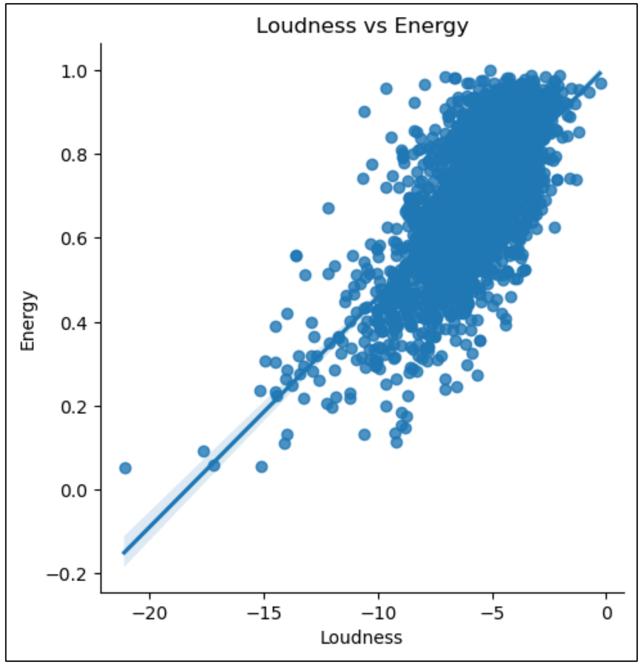
- Strong positive correlation (0.69) between energy and loudness
- Strong negative correlation (-0.55) between acousticness and energy

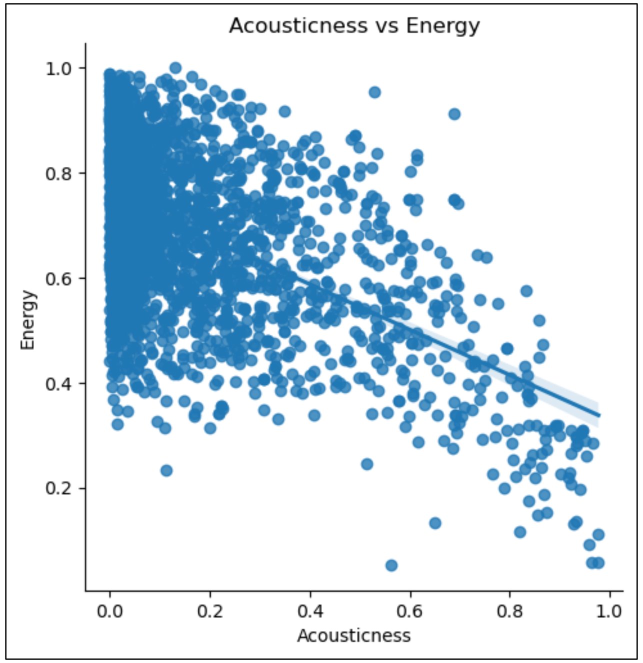
2

The correlations were visualized through scatterplots. The relationship between energy and acousticness was chosen for further exploration.

3

Hypothesis:
The more acoustic a song is, the less energy (or calmer) it feels.

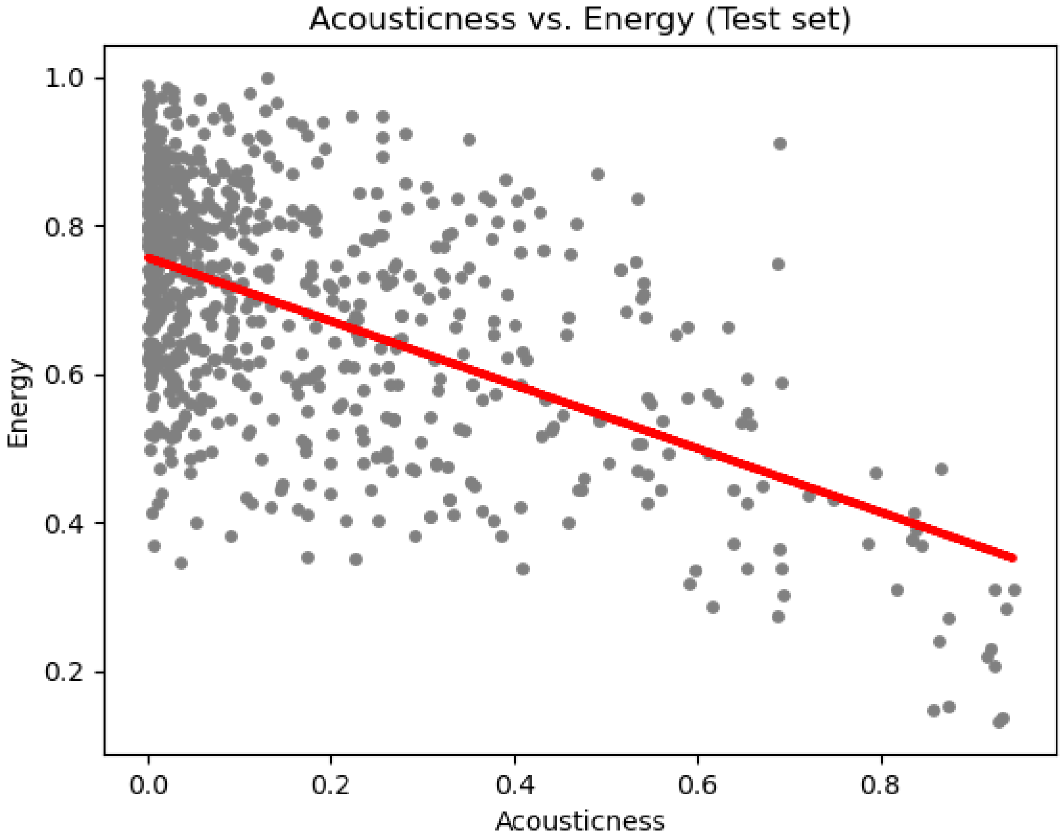




Linear Regression

The data was split into a training set and test set (70/30).
A **linear regression model** was applied to the training set based on the following hypothesis:

The more acoustic a song is, the less energy (or calmer) it feels.



The chart displays the results of the model on the test data set. The **model performance statistics** are as follows:

Slope: -0.42
MSE: -0.019
R2 Score: 0.309

- The negative slope confirms the negative correlation between acousticness and energy.
- The low MSE shows that the model's predictions are close to the actual values.
- The low R2 score indicates that this model **cannot** explain almost 70% of the variances in the data: the model is a **poor fit**.

Interpretation of results:

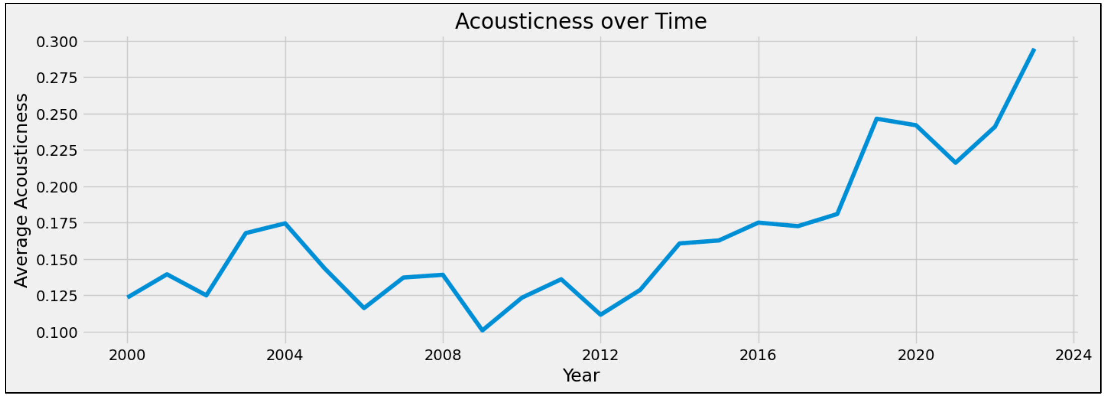
Acousticness alone does not explain energy level of a song. For the complexity of this case, a more advanced multiple linear regression model may prove a better fit.

Time Series Analysis

Acousticness over time:

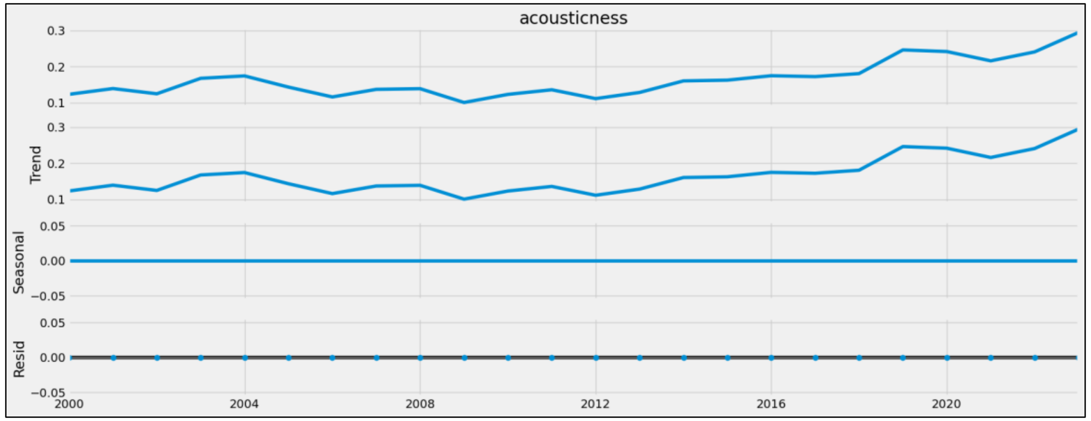
The introduction of electronics to music came only in about the last half century and is a relatively young genre in the long history of music.

Has the use of acoustic instrumentation continued to decrease since the beginning of this millenium?

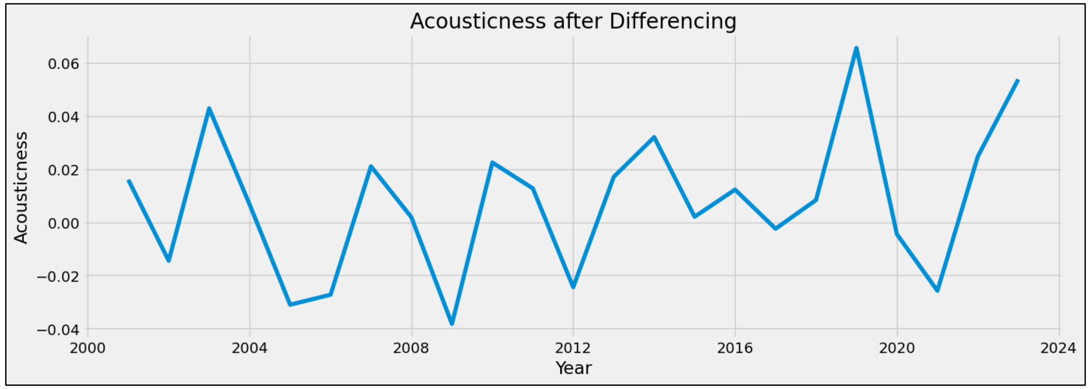


Surprisingly, average acousticness of the top hits since 2000 show a **rise in the last decade**. It seems that the electronic age may have peaked in the early 2000's and acoustic instrumentation is now making a comeback.

Stationarizing the time series:



The decomposition of the time series show a slight upwards trend, with no seasonality nor residuals.



Even after differencing, this time series could not be stationarized.

Conclusion: This time series is not suitable for forecasting.

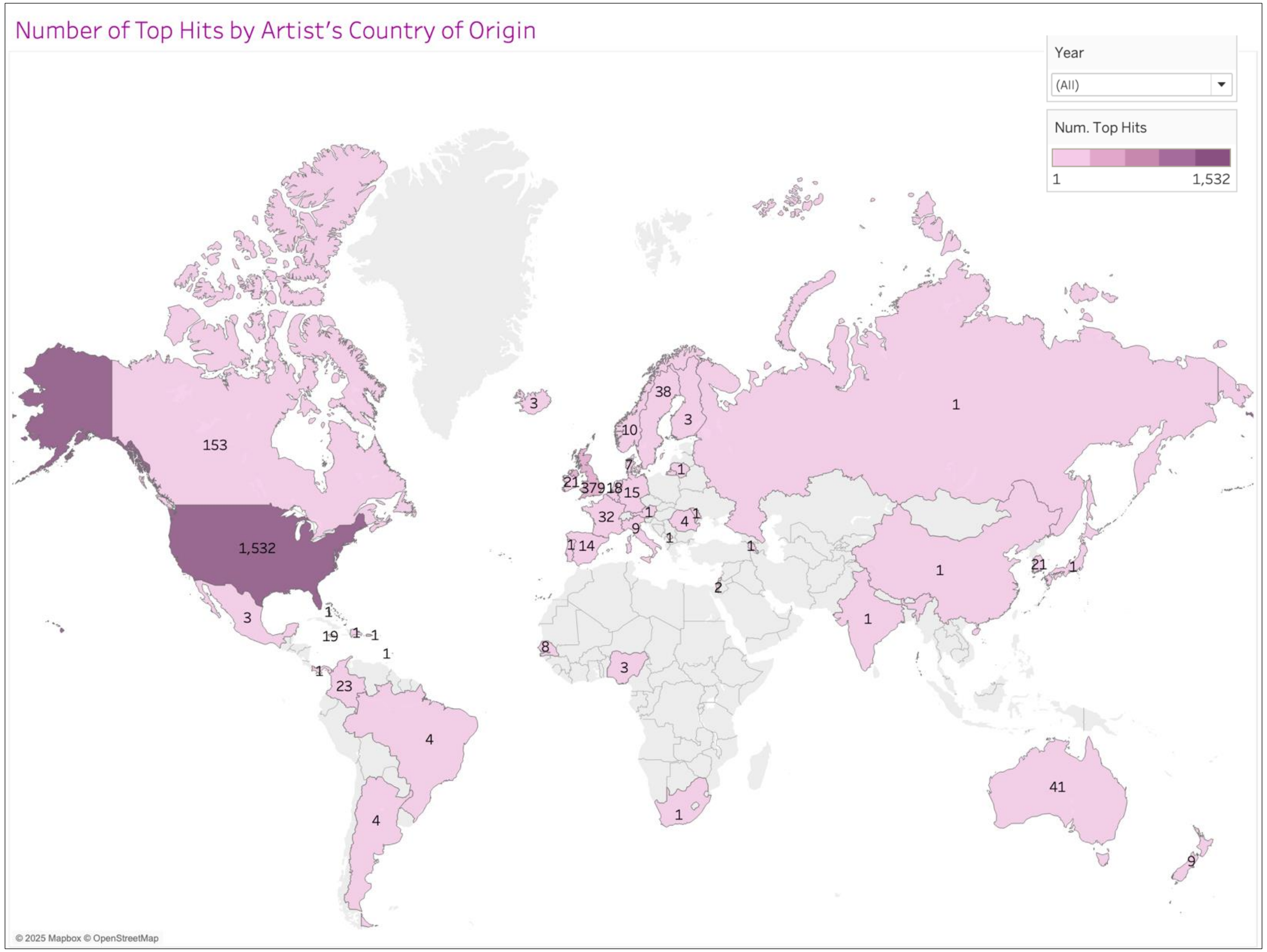
Geospatial Analysis

Where do the top hits come from?

Key takeaways from the geospatial analysis:

- The **USA** is home to the highest number of top hit artists.
- **English-speaking countries** (USA, UK, Canada and Australia) dominate in total number of hits over the last two decades.
- **South Korea** starts to gain significant numbers of top hits after 2019 and currently produces the highest number of popular artists in Asia.

Please click here to visit the interactive version of this map on Tableau.



Conclusions

Key Insights:

- American artists continue to top the charts in popular music.
- The last two decades have been a flourishing time for electronic (non-acoustic) music that are high in energy.
- Acoustic instrumentation has gradually returned over the last few years.
- The data doesn't allow for forecasting, therefore we cannot predict how popular music will evolve in the future.

Project assessment:

The accuracy and reliability of the data on audio features are highly dependent on the specifications of Spotify's algorithms.
Ultimately, the data proved unsuitable for various models, such as linear regression, cluster analysis and time series forecasting.
Although these limitations created challenges in producing the insights to popular music I was hoping for, it reminded me that music is an art, and numbers cannot always capture the essence of art.

Next steps for further analysis:

- Perform an advanced multiple linear regression model on multiple audio features.
- Observe and compare audio features of top hits by various artist's nationalities.

The Tableau Storyboard and GitHub Repository are available to view:

Rockbuster Analysis Project

Launch Strategy for an Online Movie Rental Service

Overview

Context:

Rockbuster Stealth LLC* is a movie rental company looking to launch an online video rental service in order to stay competitive with online streaming services such as Amazon Prime and Netflix.

Objectives:

The Rockbuster management would like to have a better understanding of their customer base. They have business questions and expect data-driven answers to use for their launch strategy.

Role:

Data analyst for Rockbuster’s business intelligence department, tasked to help with the launch strategy for the new online video service.

Data:	Rockbuster data set*
Skills:	data modeling, data cleaning and summarizing, database querying, data visualizations, reporting
Tools:	PostgreSQL, Excel, and Tableau

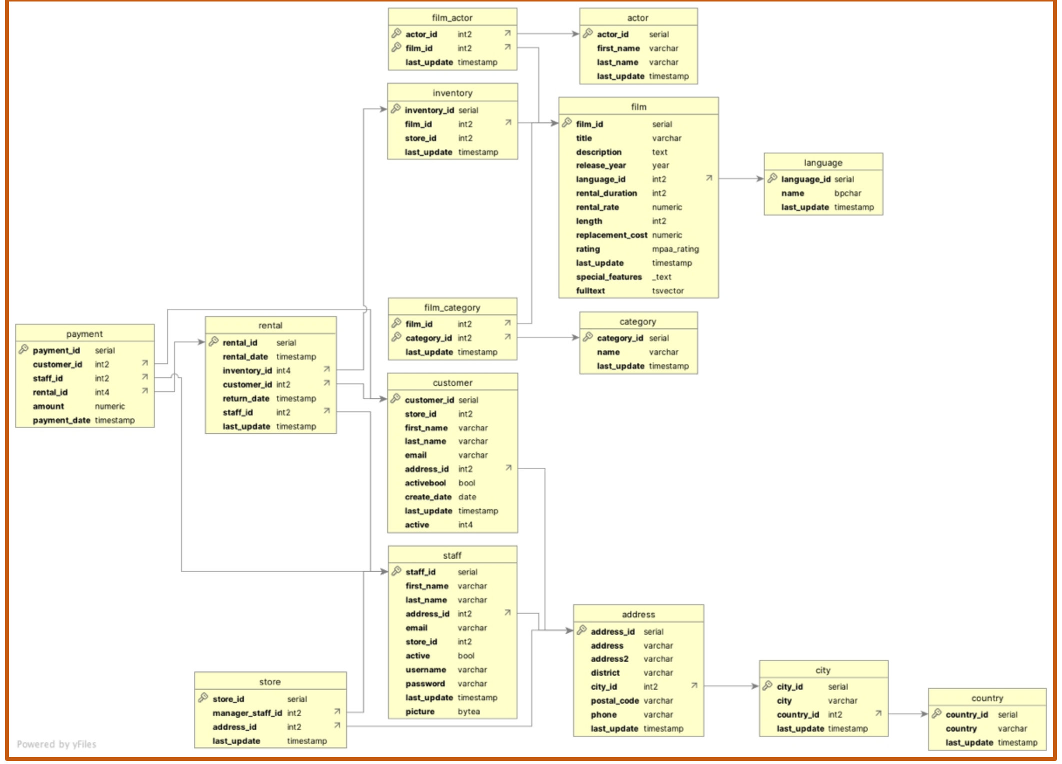
Process:

- Understanding the data structure
- Exploring and preparing the data
- Answer business questions through queries
- Report findings and offer recommendations

**This project and data set were created as part of the CareerFoundry data analytics course.*

Understanding the Data

The Entity Relationship Diagram of the Rockbuster database shows the relationships between the various tables and how they are linked.
Creating a data dictionary helps to clarify the complex relationships and define the data within each table.
The data dictionary serves as a reference document for the data analyst when accessing and querying the database.



Entity Relationship Diagram, extracted using DbVisualizer

2.1 Rental Table

Fact table with information on movie rental transactions.

COLUMNS

Name	Data Type	Description
rental_id	serial	Primary key for rental transaction.
rental_date	timestamp	Date rented out.
inventory_id	int4	Unique ID number for inventory record. Foreign key to Inventory table.
customer_id	int2	Unique ID number for customer. Foreign key to customer table
return_date	timestamp	Date of rental return.

Example from the data dictionary created for this project.

Exploring the Data

1. Data Cleaning: the data was checked for duplicates, missing values or inconsistencies.

2. Exploration: summary statistics were made on various tables:

- Stores
- Customers
- Rentals
- Payments
- Inventory
- Films

Statistics on customer data:	
Number of customers:	599
Most customers registered at store in:	Lethbridge, Canada
Number of active customers:	584
All customers’ accts created:	2006-02-14

Statistics on film data:	
Number of films:	1000
Average rental duration:	5 days
Average rental rate:	$ 2.98
Most frequent film rating:	PG-13
All films released in the year:	2006
All films in language:	English

Answering Business Questions

More complex queries use filtering, joining tables, or nesting subqueries to answer some of the management's key questions.

Who are our most valued paying customers?

First Name	Country	Total Rev
Eleanor	Runion	$ 211.55
Karl	United States	$ 208.58
Marion	Brazil	$ 194.61
Rhonda	Netherlands	$ 191.62
Clara	Belarus	$ 189.60

Which movies bring in the most revenue?

Query Query History

```sql
SELECT D.title, D.rating, SUM(A.amount
FROM payment A
INNER JOIN rental B ON A.rental_id = B
INNER JOIN inventory C ON B.inventory_
INNER JOIN film D ON C.film_id = D.fil
GROUP BY title, rating
ORDER BY SUM(A.amount) DESC
LIMIT 10
```

Snapshot of the SQL query input

Movie Title	Rating	Total Rev
Telegraph Voyage	PG	$ 215.75
Zorro Ark	NC-17	$ 199.72
Wife Turn	NC-17	$ 198.73
Innocent Usual	PG-13	$ 191.74
Hustler Party	NC-17	$ 190.78

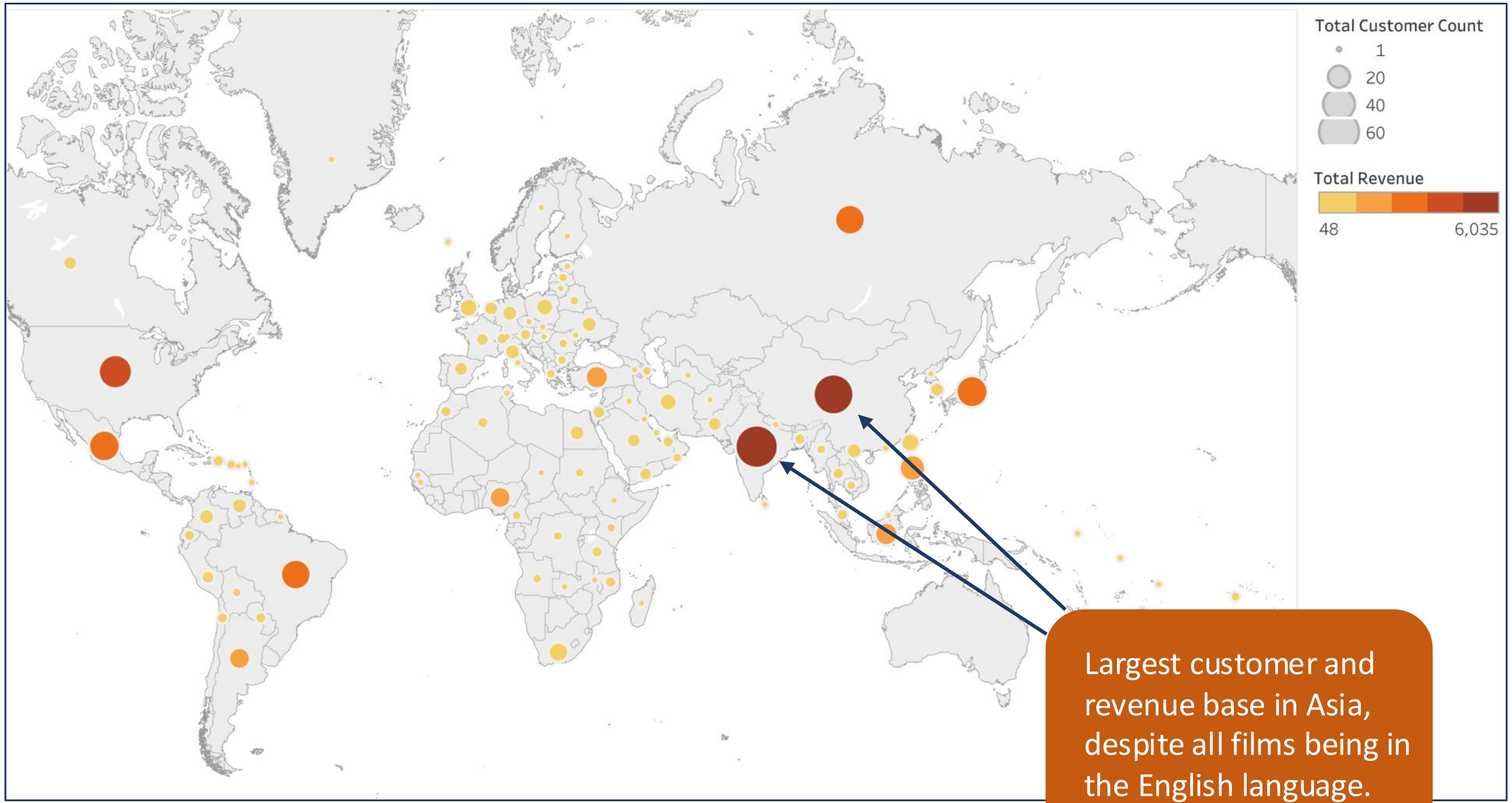
Please get in touch if you are interested in viewing the Excel Workbook with all SQL queries and results.

Data Visualization

Which countries have the most customers?
Which countries bring the most revenue?

Visualizing the result of a query can illuminate a pattern that may otherwise go unnoticed!

Number of Customers and Total Revenue by Country



Largest customer and revenue base in Asia, despite all films being in the English language.

Click to view the interactive map on Tableau.

Conclusions

Recommendations for Launch:

Following my exploratory analysis, these are some recommendations I would make to the Rockbuster launch team:

- Invest more advertising for the new platform in countries with the highest number of customers (India, China, US).
- Feature top revenue films on the front page of the online site.
- Expand and update the Rockbuster inventory to include films in more languages and release years.
- Retain valued customers when transitioning to the online platform by offering top payers a reward or discount.

Project assessment:

Through this project, I learned about the **challenges** in accessing data from a relational database. It is very important to understand the structure of the database, as it can be quite complex.

The various business questions gave me an opportunity to refine my SQL programming skills, especially in the task of **joining tables** and performing **subqueries**.

The frame of the project allowed me to answer the business questions presented, but I would have liked to further **explore** follow-up questions such as:

- *Which films are most popular in the top countries?*
- *Are there film ratings that produce more revenue?*

The final project presentation and GitHub Repository are available to view:

CDC INFLUENZA

Preparing for the Influenza Season in the U.S.

Case Study

The United States has an influenza season where more people suffer from the flu, particularly those in **vulnerable** populations.

The medical staffing agency provides temporary workers to hospitals and clinics to adequately treat patients during these times.

This analysis examines trends in influenza to advise the medical staffing agency on how to plan for the influenza season and optimize staff distribution across the country.

Data Sets	
CDC Influenza Deaths	**US Census**
2009 – 2017	2009 – 2017
Death counts by State and Age	Population counts by Region, Gender and Age

Data Cleaning
Data Integration
Data Transformation

Statistical Analysis

1

Research hypothesis:
If a state has a higher number of senior citizens over 65, then it will also have a higher proportion of severe cases or deaths from influenza.

Null hypothesis:
The flu mortality rate for people 65 years and older is the same or less than the flu mortality rate for the population under 65.

Alternative hypothesis:
The flu mortality rate for people 65 and older is higher than the flu mortality rate for people under 65.

2

Statistical Hypothesis Testing:
Conduct a one-tailed two-sample t-test on the null hypothesis.

t-Test: Two-Sample Assuming Unequal Variances

	Under 65 % deaths	65+ % deaths
Mean	0.000269161	0.001316014
Variance	7.59954E-08	2.7273E-07
Observations	459	459
Hypothesized Mean Difference	0	
df	695	
t Stat	-37.97962212	
P(T<=t) one-tail	5.1953E-172	
t Critical one-tail	1.647049044	
P(T<=t) two-tail	1.0391E-171	
t Critical two-tail	1.963383175	

Because the observed p-value is 5.1953E-172, which is significantly smaller than the alpha of 0.05, we can **reject the null hypothesis.**

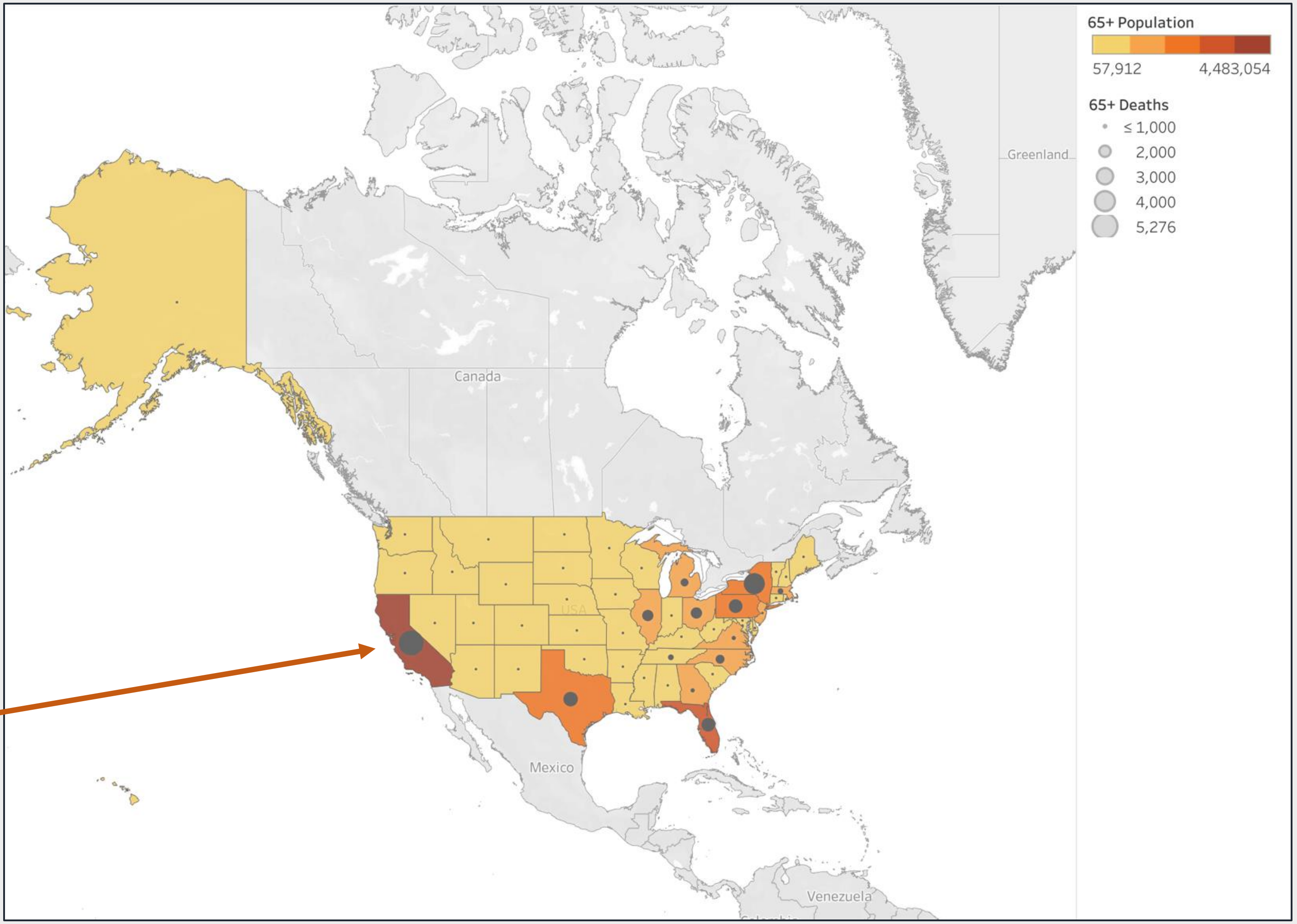
The staffing agency could optimize the distribution of their medical staff by focusing on areas with a higher proportion of citizens aged 65 and older.

Spatial Analysis

Building on the results of the hypothesis testing, the next step is to determine **which states** have a higher proportion of vulnerable elderly population.

This Tableau map shows the states with the highest population of citizens 65 years and above, as well as the states with the highest mortality rates in that age group.

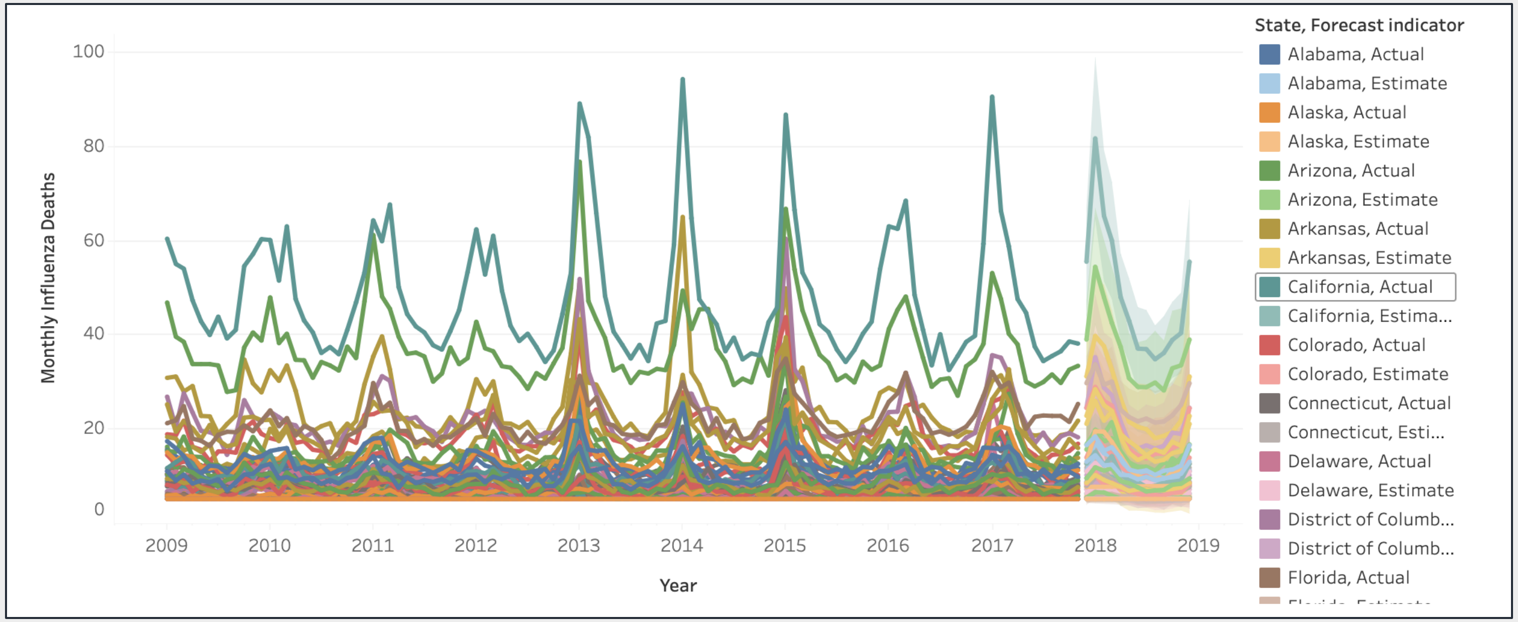
It is clear to see on this visualization that **California** leads in both highest population and highest mortality rate for seniors over the age of 65.



Results and Recommendations

This temporal forecast shows a clear seasonal trend in influenza cases every year. The flu season starts in the winter, peaking in January, and ends in spring.

With the combination of the results of the previous spatial analysis and this temporal chart, the medical agency can forecast the coming influenza season and distribute staff to the states with most need, at the time of most need.



On Tableau Public for this project, there is an **interactive tree map** that shows each state, its influenza statistics and its respective seasonal forecast chart.

On Vimeo, there is a video available on the presentation of this project.

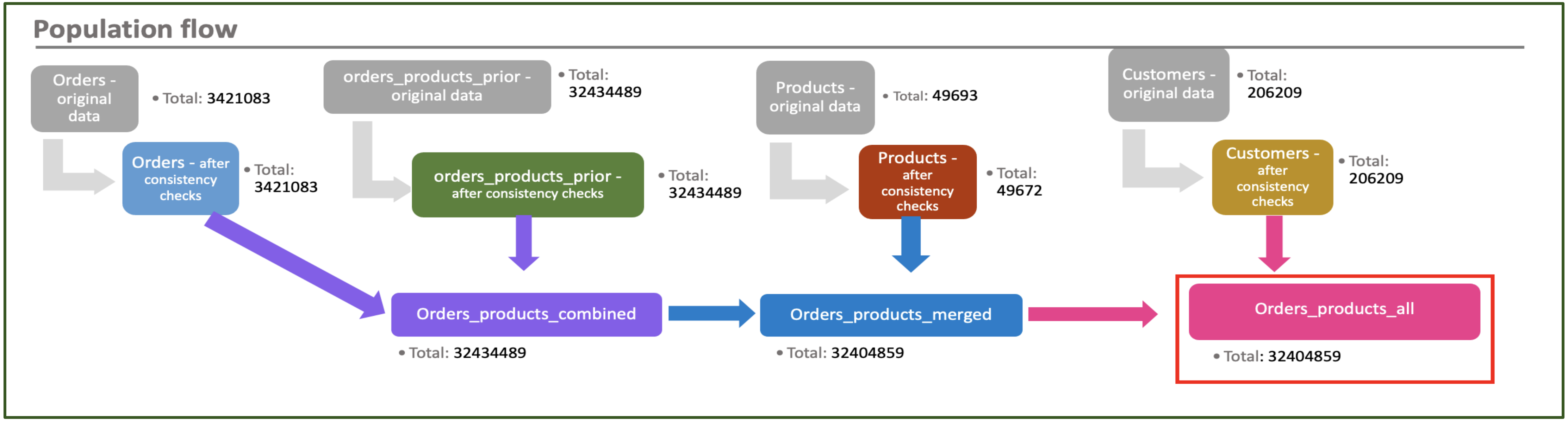
INSTACART

Marketing Strategy for an Online Grocery Store

Case Study

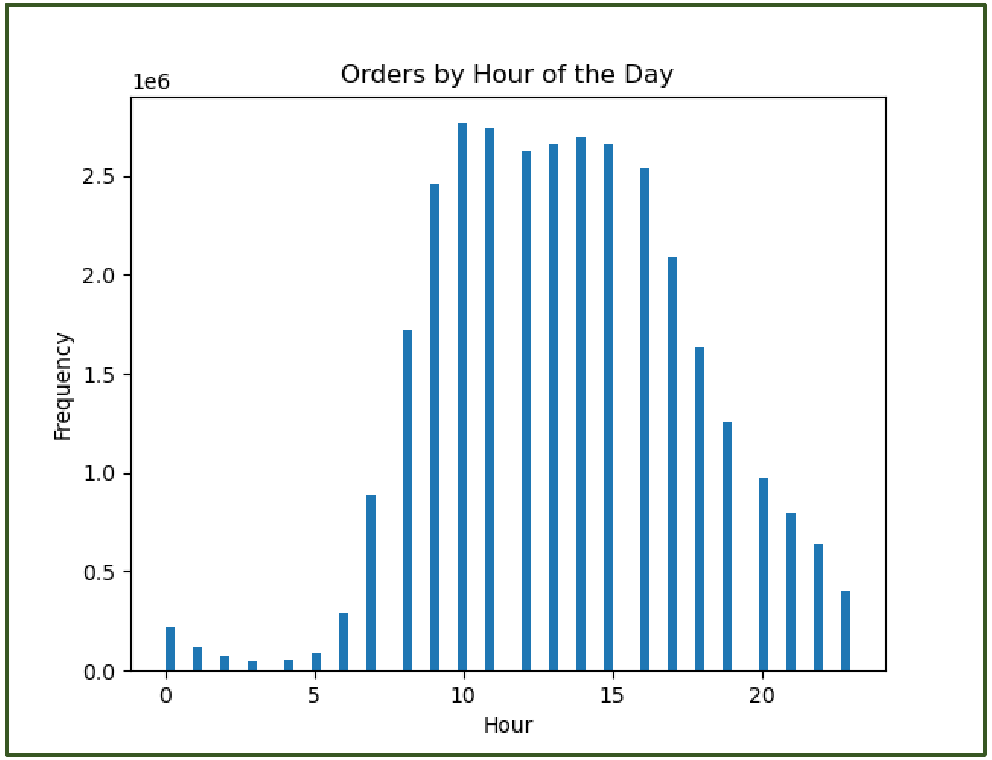
The Instacart stakeholders are interested in the variety of customers in their database. They are considering a targeted marketing strategy and would like to know more about sales patterns as well as their customers' purchasing behaviors. The large size of the data set requires the use of the powerful Python pandas library.

The Instacart data sets from Kaggle must go through **wrangling** and **consistency checks** before being **merged** into one useful data set for the purpose of this analysis. This population flow chart shows the transformation of the data sets through the various stages of consistency checks and merges.

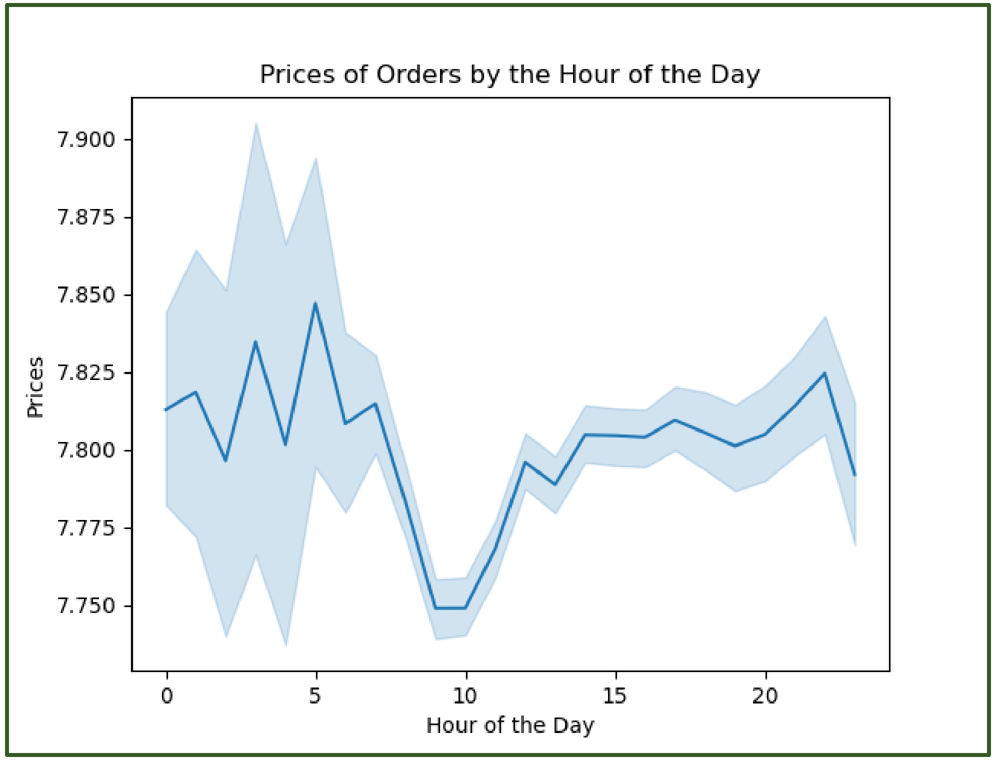


Exploratory Analysis I

Customer Purchasing Behavior by Time of Day



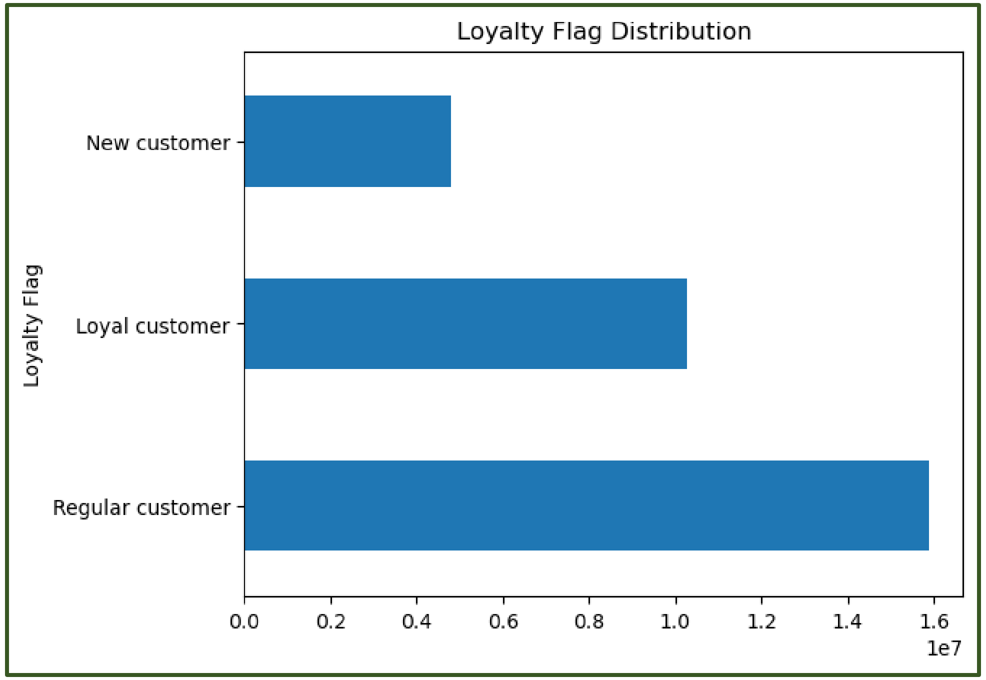
Viewing the data by the frequency of orders per hour shows the midday hours have the highest volume of orders.



However, an analysis of the average prices of orders show customers are purchasing lower priced items during those midday orders.

Exploratory Analysis II

Customer Ordering Habits and Demographics



A loyalty flag was created based on the number of orders each customer made. Loyal customers made over 40 orders while new customers made under 10 orders. Regular customers fall into the middle.

It is concerning to see that the smallest group are the new customers: if we want to increase the customer base, there should be efforts to increase as well as keep the new customers.

In order to further analyze customer ordering habits by demographics, customer profile tags were created on classifications of age and number of dependents.

There seems to be little difference in the average prices each group is paying.

Customer Profile	Mean Num Orders	Mean Prices
Older family	34.20	7.79
Single adult	**34.94**	7.79
Single youth	34.21	7.78
Young family	34.60	7.79

Additionally, we can see the breakdown of these customer types by regions in the US:

Customer Profile	Midwest	Northeast	South	West
Older family	48.12%	**48.42%**	47.55%	48.12%
Single adult	16.20%	15.83%	15.84%	**16.21%**
Single youth	8.90%	8.96%	**9.20%**	8.78%
Young family	26.78%	26.79%	**27.41%**	26.90%

Results and Recommendations

The marketing team can schedule ads in the early morning and times when they are less orders. They can choose to advertise pricier products, as customers seem more willing to pay for more during these off-hours.

The distribution of loyalty flags show that the smallest group of customers are new customers. Marketing may want to focus on attracting more customers as well as retaining the current ones in order to continue to expand the customer base.

Single adults over 40 order most frequently and also pay for pricier items.
Stores in various regions can target ads toward customer profile groups that are most prominent in their area.

Link to GitHub Repository with full Analysis and Jupyter Notebooks

ClimateWins

Implementing Machine Learning for Weather Prediction

Case Study

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

Role:
As a data analyst for ClimateWins, I will assess the tools available to categorize and predict the weather in Europe.

Tools: Python, Jupyter

Data set:
- Observations from 18 weather stations across Europe, dating from late 1800's to 2022
- Values such as temperature, wind speed, snow, precipitation, global radiation and more
- Collected by the European Climate Assessment & Data Set project

Hypothesis:
Machine-learning algorithms can be applied to the data to predict weather conditions.

Process for exploring hypothesis:
- Clean and prepare weather data
- Run optimization algorithm
- Apply various machine learning models, both supervised and unsupervised
- Evaluate accuracy and usefulness of different models on the weather data set
- Propose an effective method for machine learning application for weather prediction

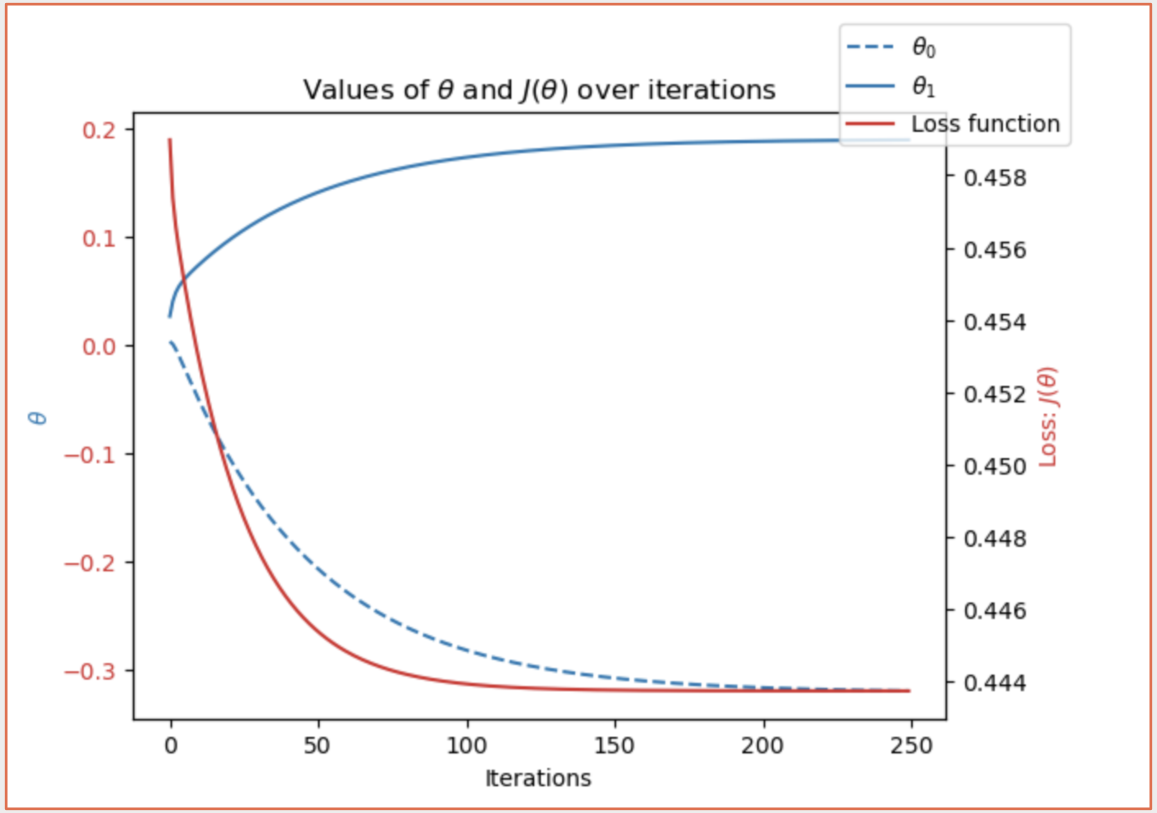
Optimization

To better understand the structure of our data and optimize it for machine learning, we can run an optimization algorithm such as **gradient descent** on a feature of the data.

Here we are fitting a regression model by running gradient descent on daily mean temperatures of various stations for a chosen year.

Gradient Descent on Kassel Station, 1960:

- Iteration: 200, Step size: 0.1
- Successfully converges
- Minimum achievable loss at 0.43



Kassel 1960 Loss Function

Supervised Learning

Preparation:

- An **answers** data set is provided to train the model on predicting if a certain day is pleasant or not.
- The data set is **scaled** to prevent the machine learning model attributing more weight to higher values.
- The data is split, **70%** for training, **30%** for testing.
- For this project, we applied the following models:
    - **K-Nearest Neighbor (KNN)**
    - **Decision Tree**
    - **Artificial Neural Networks (ANN)**

Challenges:

- **Parameters:** For each model, we must experiment on the parameters to produce the highest accuracy.
- **Overfitting:** The models are overfitting to one weather station (Sonnblick), therefore negatively affecting the overall accuracy.

Results on Supervised Models:

Model	Test Accuracy (Confusion Matrix)
KNN	90%
Decision Tree	95%
ANN	95%

Why the **ANN model** works best in predicting current data:

- Most accurate predictions on both training and testing sets.
- Less overfitting than the decision tree model.
- Room for improvement and experimentation with the parameters.

Unsupervised Learning

Neural Network Models

The success of the ANN model in the supervised learning exercise leads to an exploration of the **Recurrent Neural Network (RNN)** model in unsupervised learning. RNN's strength lies in handling **temporal data**.

The **Long Short-Term Memory (LSTM)** model is an improved version of RNN. The model is trained to predict the same weather conditions: pleasant or unpleasant.
Here is the final Keras model layout for running LSTM on the weather data.

```
[52]: epochs = 30
      batch_size =64
      n_hidden = 32

      timesteps = len(X_train[0])
      input_dim = len(X_train[0][0])
      n_classes = len(y_train[0])

      model = Sequential([
          Input(shape=(timesteps, input_dim)),
          LSTM(n_hidden),
          Dropout(0.5),
          Dense(n_classes, activation='sigmoid') ])

[53]: model.compile(loss='categorical_crossentropy',
                    optimizer='rmsprop',
                    metrics=['accuracy'])

[54]: model.fit(X_train,
                y_train,
                batch_size=batch_size,
                validation_data=(X_test, y_test),
                epochs=epochs)
```

Challenge:
It is time-consuming and difficult to experiment with the endless variations of hyperparameters in neural network models. How do we efficiently find the optimal values to improve the performance of deep learning models?

Bayesian Optimization:
Applying a Bayesian search on the hyperparameters produced a set of optimal values for running the model. It improved the accuracy of the pre-optimized model by 5%.

Proposal for ClimateWins

The following proposal results from thought experiments on the possibilities of implementing machine learning to predicting data and the study of various machine learning models throughout this project.

Thought Experiment	Action	Data Required	ML Models
Can machine learning models accurately identify future extreme weather events based on weather data from the last 60 years?	Optimize a model to accurately identify extreme weather conditions.	○ ClimateWins dataset ○ 'Answers' dataset of past extreme weather events	ANN
If we determine that extreme weather events are increasing, can we predict when and how frequent these events will occur?	Train model to predict the time and frequency of weather events.	○ ClimateWins dataset ○ 'Answers' dataset of past extreme weather events	LSTM
If we can predict when and how often these disasters will happen, will we be better prepared? Will there be enough time to make changes?	Analyze risk by location and population to determine the optimal time frame for disaster preparation.	○ ClimateWins dataset ○ Population dataset ○ Location dataset of disaster predictions	LSTM

Conclusion

Key Insights:

- Machine learning models are able to help predict weather conditions.
- Neural network models work best for the ClimateWins dataset.
- Complex deep learning models such as LSTM will need optimization algorithms to produce higher quality results.
- Thought experiments can assist in creating a viable method for applying machine learning for extreme weather prediction.

Project assessment:

Working with machine learning algorithms is a challenging yet exciting venture. This project has given me an insightful first look at the complex inner workings of various machine learning models. I've gained an understanding of which models are useful in different scenarios, or even that there are possibilities out there still to be explored.

As powerful as machine learning models are, I have also realized through this project how vital the quality and preparation of the data is to running a successful and useful model. In future projects, I would aim to always thoroughly clean and transform the data before modeling.

The project is available on GitHub:

CONTACT

If you have questions, or are interested in working with me, please get in touch!