Edwin Ng

Weather-based Restaurant Sales Prediction Final Report:

Problem Statement:

Like any other business, understanding your customers is one of the keys to success. Even if you have a brilliant product, if you don't know when and how your customer wants to buy said product, it will never get sold. This also rings true for restaurants. Especially when profit margins are so thin and inventory will expire within a few days, as a restaurateur you need to understand when certain demand for your food will rise and fall. Whether it's a sporting event nearby or road construction blocking off access to your neighborhood, there are a variety of factors that can affect customer volume and product demand.

For restaurants, one of these factors is weather. While this might seem a bit arbitrary as to whether customers are hungry or not, "More than 90 percent of restaurant operators indicate that changes in local weather conditions affect their sales and customer counts." (Devincent-Reinbold). For example, hot weather might drive up sales of salads and cold/lighter foods and drive down sales of heavier, hotter meals such as soups and stews. Rain, for example, might make customers want to stay indoors and not go out to eat, resulting in less overall customer volume whereas in other

locations (ie. outdoor malls) it might drive customers indoors to seek shelter from the rain, in turn driving up sales. So, knowing that weather can affect our customers' moods, how can we leverage this information to create a model that predicts how this affects the sales of certain kinds of food at a restaurant?

Data Wrangling:

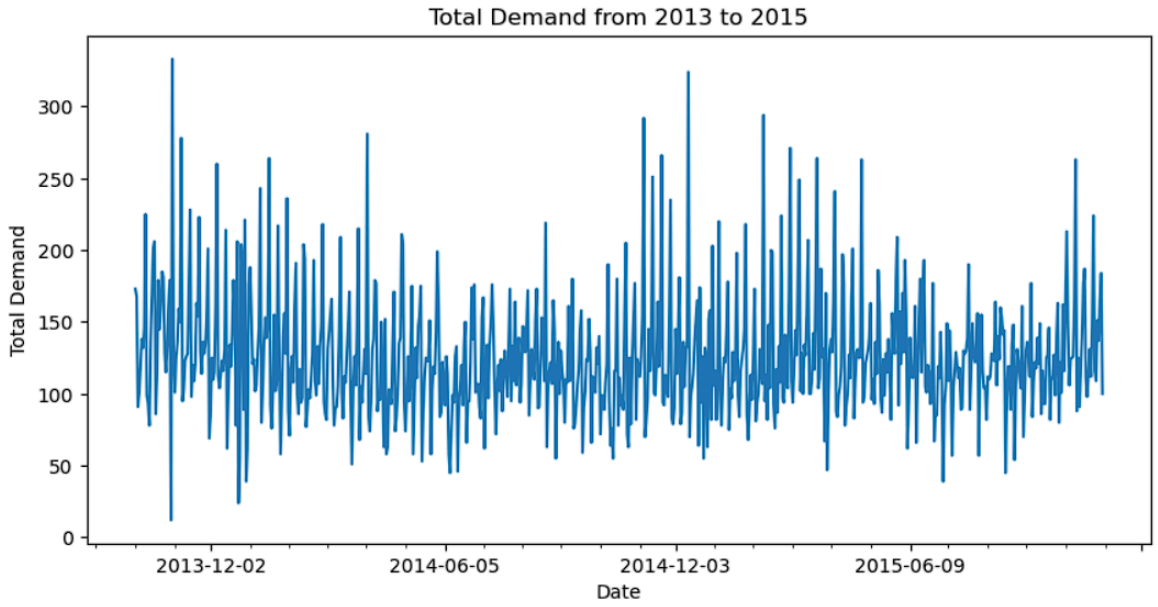
The restaurant sales data was acquired from Mendeley data and contained two years of daily sales data from a restaurant based in Stuttgart, Germany. The data originally contained 760 entries and 293 columns. After assessing which columns were necessary and useful to answering our question, we trimmed unnecessary data from the dataset. Columns such as those containing data from arbitrary and undescribed time periods were deemed unusable since there was no explanation as to the grouping of the time periods and how that affected the data collection. This resulted in a final dataset size of 760 entries and 40 columns.

After this, the next step to this phase was to ensure that all our columns contained the right kind of information. Most of our data in this case was already in the right type with the exception of the “Demand date” column - which we quickly converted to the datetime data type. After ensuring column data type integrity we made sure that there were no duplicate entries within the dataset before making sure that our dataset contained no null values. Finally, once the dataset was trimmed, casted, and inspected for null values, we exported the cleaned data to an output file for ease of access in the subsequent notebooks.

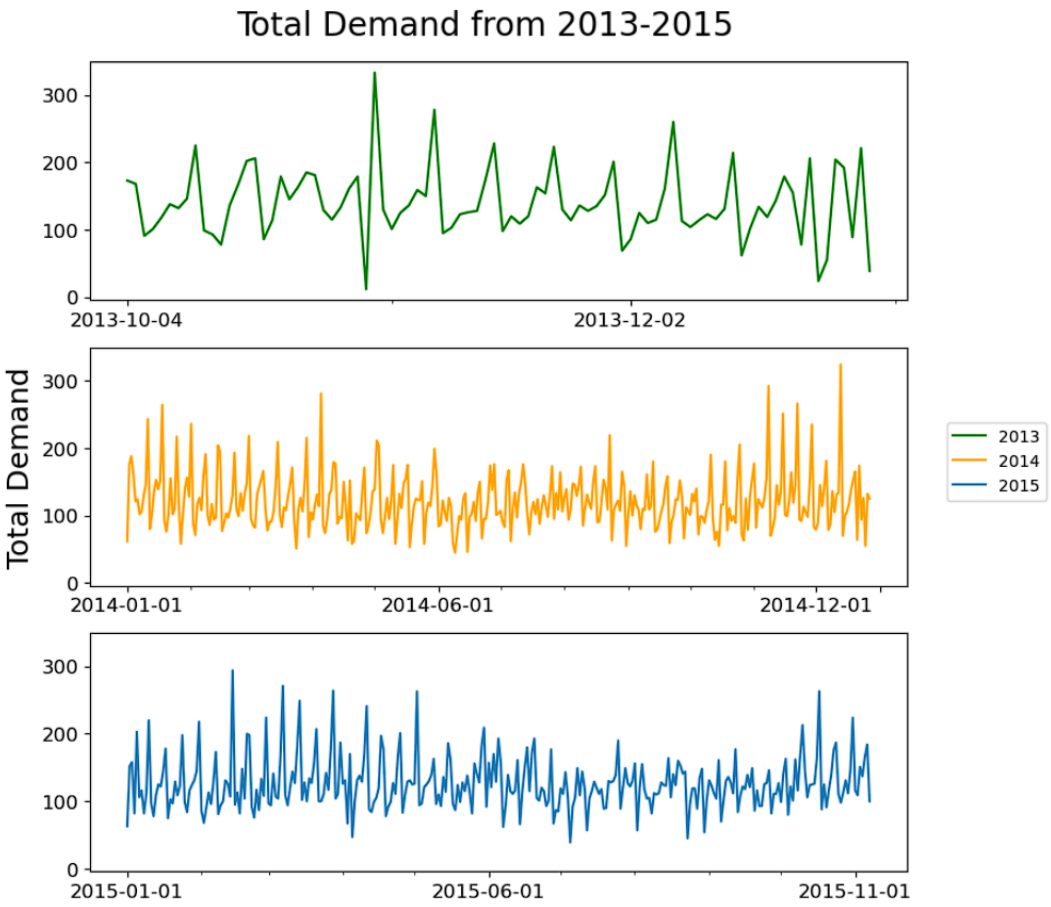
Exploratory Data Analysis:

We first took a look at our main focus of the project - total demand. In the world of restaurants, aside from the day to day operations, a large part of understanding your business is looking for long term cycles in demand. This can help with things such as supply levels, staffing rates, labor hours, etc. We first explore this metric by getting a sense of the data’s seasonality by graphing the total demand from 2013 to 2015.

Edwin Ng



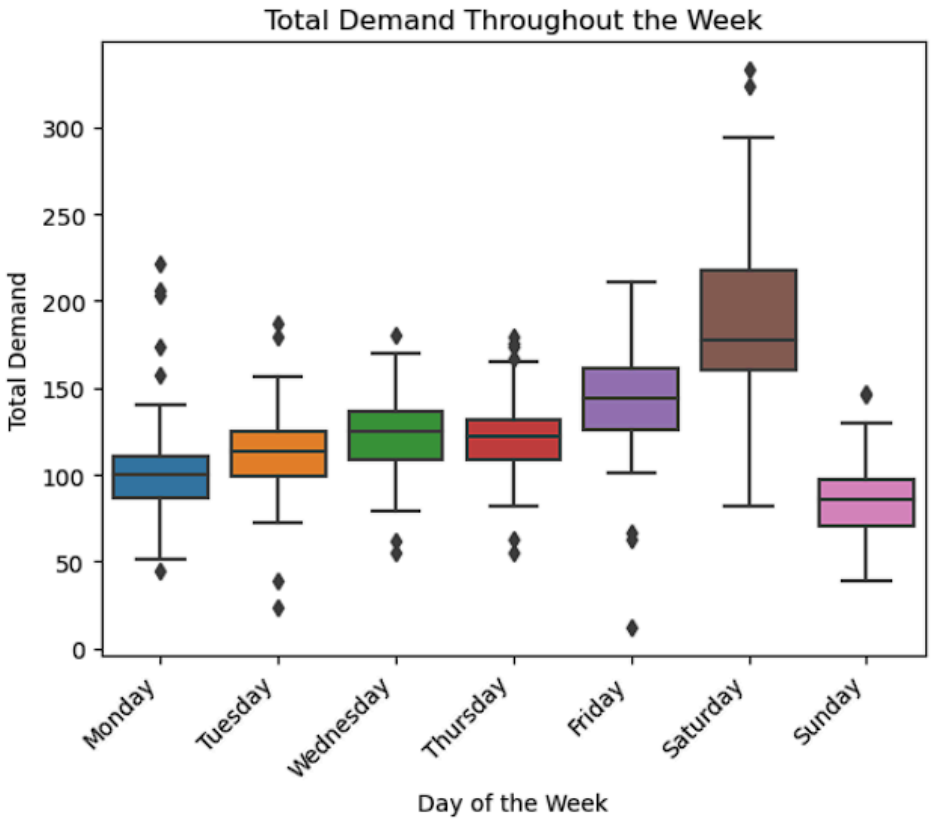
Total demand, as the name might suggest, takes into account all other factors and is a pure representation of how many customers the restaurant experiences. That being said, while there are some interesting peaks here and there within the graph, the graph is too packed for us to really get an idea of any true seasonality. From here, we split up the years and stack the graphs on top of each other to see if we can identify any repeating patterns.



Though not entirely super informative, we can already see that there is a slight amount of seasonality within our dataset. Granted we only have two full years of data, we can see that for both 2014 and 2015 there is, on average, a lower level of demand during the Summer months and into Fall - specifically from June to roughly October. From November through April/May there is an elevated level of demand compared to the rest of the year.

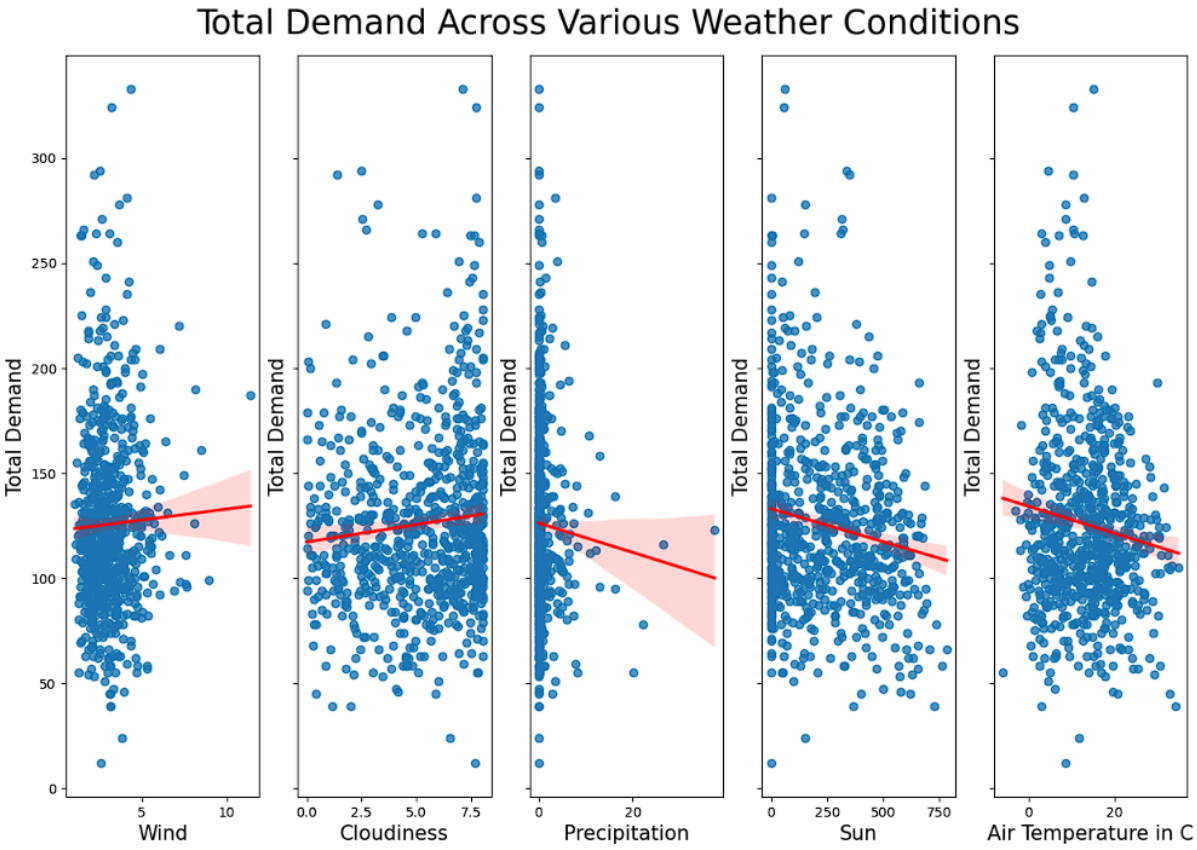
Now that we've taken a look at the demand on a yearly scale, we then move into analyzing the demand on a weekly scale to look for which days are the busiest. For this application, we plot a series of box-and-whisker plots to visualize the distribution of total demand levels across the days of the week. This way we can compare the levels and determine if there are significant differences between the amount of demand experienced across different days.

Edwin Ng



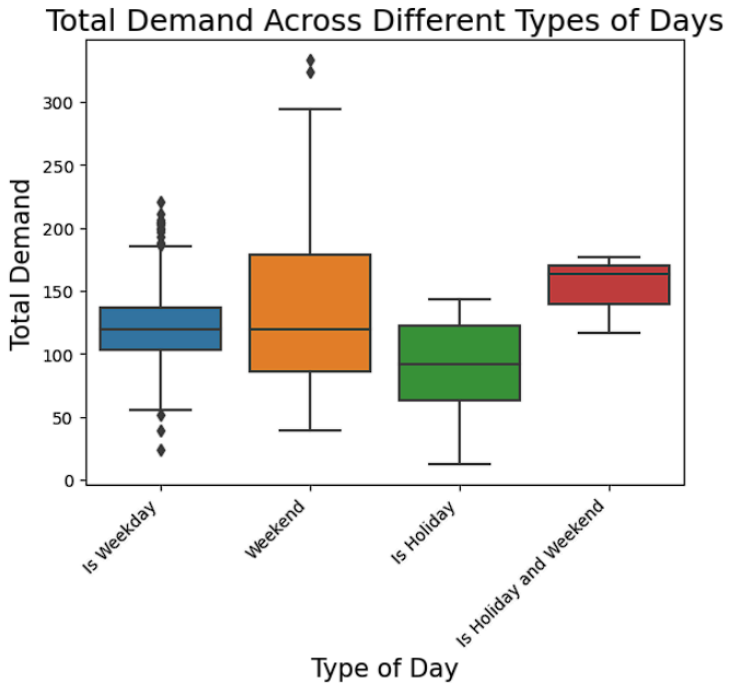
Now that we've explored the relationships between the days of the week and total demand as well as the time of year and total demand, it's time we turned our attention to the second important variable to our question - the weather. As mentioned in the introduction, weather is shown to have an impact on customers and their moods - thus hinting at possible impacts on restaurant demand. To investigate these relationships, we graphed total demand against different types of weather to see if we can elucidate any relationships between the two variables.

Edwin Ng



While from a first glance, there are not any particularly strong relationships between weather conditions and total demand, there are slight relationships to be seen. With increasing wind and cloudiness came slight increases in total demand, whereas increases in precipitation, sun, and air temperature exhibited lower total demand. This could be potentially rationalized by the sentiment that wind and cloudiness makes customers want to come inside for a warm meal whereas heavy precipitation might make customers not want to go out at all. Furthermore, increased sun and air temperature might make customers want to stay indoors or perhaps even venture outdoors and opt for other outdoor options as opposed to a stuffy indoor restaurant.

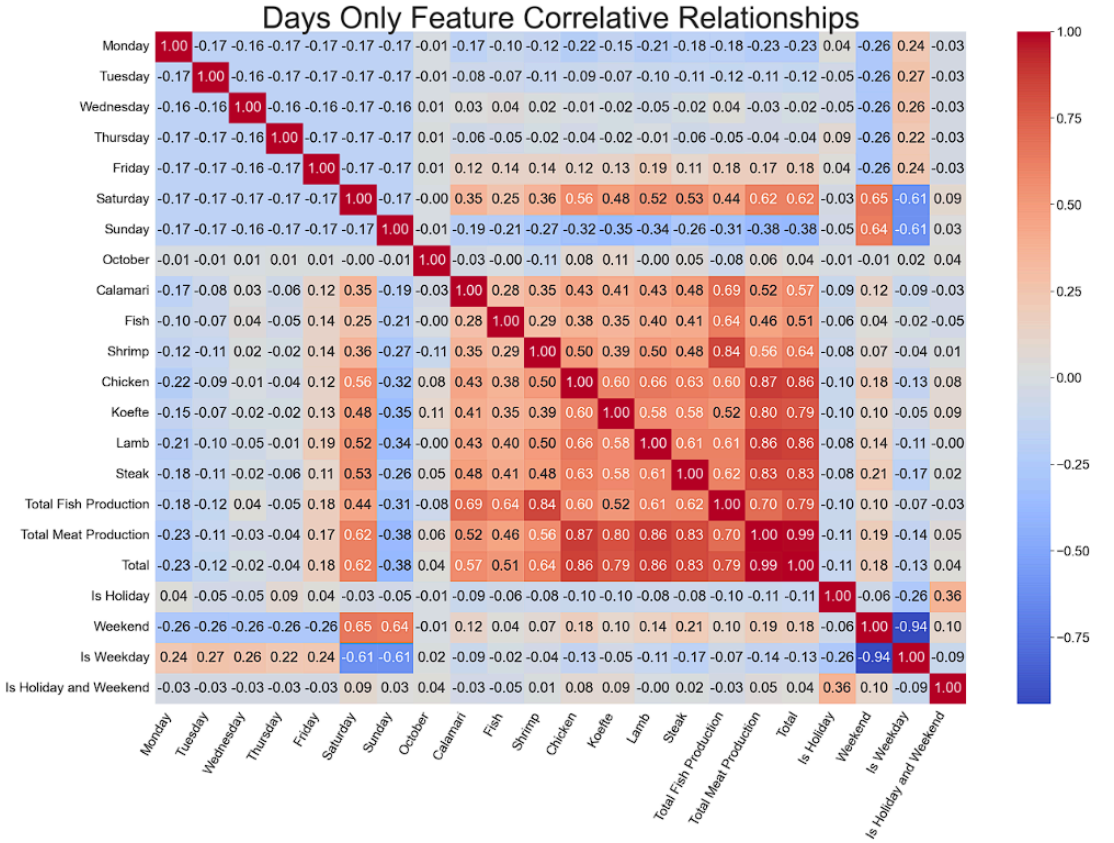
Before moving on to variable relationships, we took one final look at total demand across a variety of day-categories. We visualized relationships between total demand and weekdays, weekends, holidays and days that were both holidays AND weekends using box-and-whisker plots.

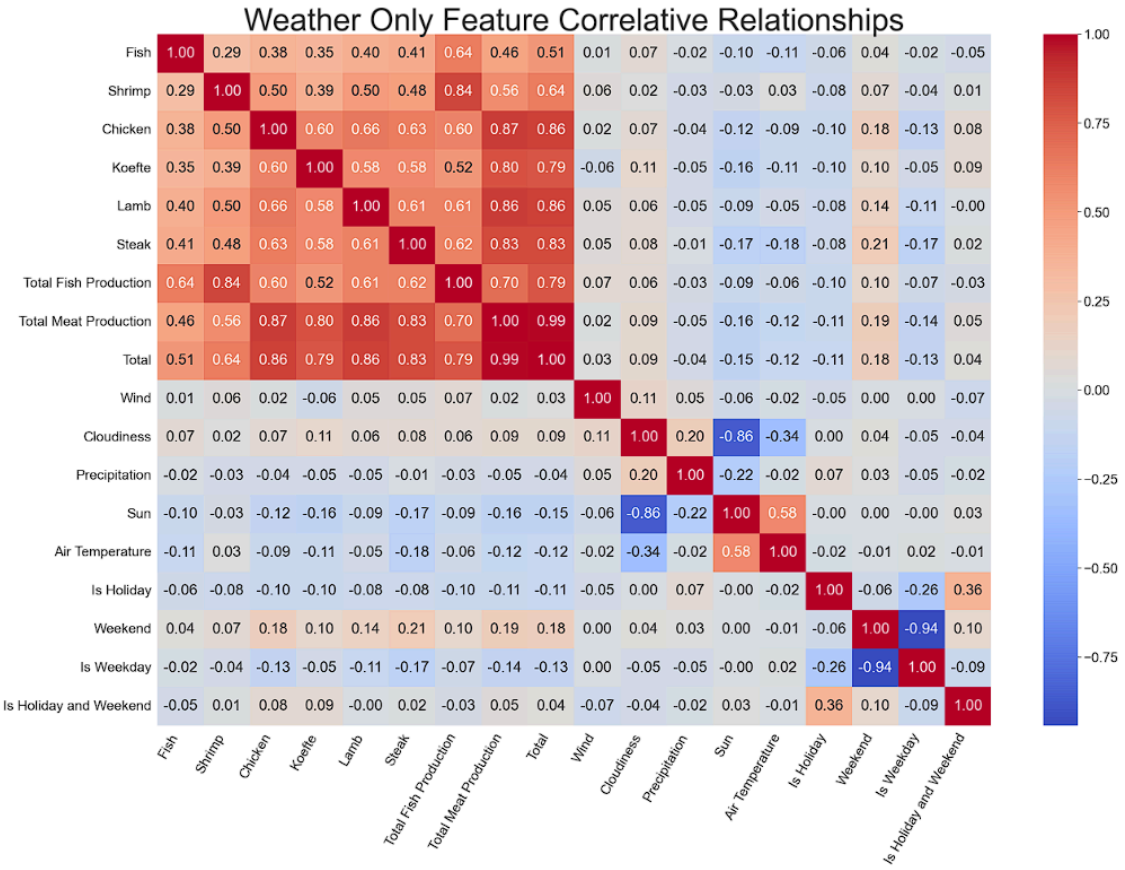


These results were a bit surprising. On average there wasn't a clear statistical difference between weekday and weekend as I had thought there would be. Even more surprising is that the median demand for holidays was lower than both weekdays and weekends. However, the median demand for days that are both holidays and weekends was the highest across the board as expected.

As for our heatmap describing feature relationships, we had to split the categories into a days-only heatmap and weather-only heatmap for readability.

Edwin Ng


Days Only Feature Correlative Relationships
Monday 1.00 -0.17 -0.16 -0.17 -0.17 -0.17 -0.17 -0.01 -0.17 -0.10 -0.12 -0.22 -0.15 -0.21 -0.18 -0.18 -0.23 -0.23 0.04 -0.26 0.24 -0.03
Tuesday -0.17 1.00 -0.16 -0.17 -0.17 -0.17 -0.17 -0.01 -0.08 -0.07 -0.11 -0.09 -0.07 -0.10 -0.11 -0.12 -0.11 -0.12 -0.05 -0.26 0.27 -0.03
Wednesday -0.16 -0.16 1.00 -0.16 -0.16 -0.17 -0.16 0.01 0.03 0.04 0.02 -0.01 -0.02 -0.05 -0.02 0.04 -0.03 -0.02 -0.05 -0.26 0.26 -0.03
Thursday -0.17 -0.17 -0.16 1.00 -0.17 -0.17 -0.17 0.01 -0.06 -0.05 -0.02 -0.04 -0.02 -0.01 -0.06 -0.05 -0.04 -0.04 0.09 -0.26 0.22 -0.03
Friday -0.17 -0.17 -0.16 -0.17 1.00 -0.17 -0.17 0.01 0.12 0.14 0.14 0.12 0.13 0.19 0.11 0.18 0.17 0.18 0.04 -0.26 0.24 -0.03
Saturday -0.17 -0.17 -0.17 -0.17 -0.17 1.00 -0.17 -0.00 0.35 0.25 0.36 0.56 0.48 0.52 0.53 0.44 0.62 0.62 -0.03 0.65 -0.61 0.09
Sunday -0.17 -0.17 -0.16 -0.17 -0.17 -0.17 1.00 -0.01 -0.19 -0.21 -0.27 -0.32 -0.35 -0.34 -0.26 -0.31 -0.38 -0.38 -0.05 0.64 -0.61 0.03
October -0.01 -0.01 0.01 0.01 0.01 -0.00 -0.01 1.00 -0.03 -0.00 -0.11 0.08 0.11 -0.00 0.05 -0.08 0.06 0.04 -0.01 -0.01 0.02 0.04
Calamari -0.17 -0.08 0.03 -0.06 0.12 0.35 -0.19 -0.03 1.00 0.28 0.35 0.43 0.41 0.43 0.48 0.69 0.52 0.57 -0.09 0.12 -0.09 -0.03
Fish -0.10 -0.07 0.04 -0.05 0.14 0.25 -0.21 -0.00 0.28 1.00 0.29 0.38 0.35 0.40 0.41 0.64 0.46 0.51 -0.06 0.04 -0.02 -0.05
Shrimp -0.12 -0.11 0.02 -0.02 0.14 0.36 -0.27 -0.11 0.35 0.29 1.00 0.50 0.39 0.50 0.48 0.84 0.56 0.64 -0.08 0.07 -0.04 0.01
Chicken -0.22 -0.09 -0.01 -0.04 0.12 0.56 -0.32 0.08 0.43 0.38 0.50 1.00 0.60 0.66 0.63 0.60 0.87 0.86 -0.10 0.18 -0.13 0.08
Koefte -0.15 -0.07 -0.02 -0.02 0.13 0.48 -0.35 0.11 0.41 0.35 0.39 0.60 1.00 0.58 0.58 0.52 0.80 0.79 -0.10 0.10 -0.05 0.09
Lamb -0.21 -0.10 -0.05 -0.01 0.19 0.52 -0.34 -0.00 0.43 0.40 0.50 0.66 0.58 1.00 0.61 0.61 0.86 0.86 -0.08 0.14 -0.11 -0.00
Steak -0.18 -0.11 -0.02 -0.06 0.11 0.53 -0.26 0.05 0.48 0.41 0.48 0.63 0.58 0.61 1.00 0.62 0.83 0.83 -0.08 0.21 -0.17 0.02
Total Fish Production -0.18 -0.12 0.04 -0.05 0.18 0.44 -0.31 -0.08 0.69 0.64 0.84 0.60 0.52 0.61 0.62 1.00 0.70 0.79 -0.10 0.10 -0.07 -0.03
Total Meat Production -0.23 -0.11 -0.03 -0.04 0.17 0.62 -0.38 0.06 0.52 0.46 0.56 0.87 0.80 0.86 0.83 0.70 1.00 0.99 -0.11 0.19 -0.14 0.05
Total -0.23 -0.12 -0.02 -0.04 0.18 0.62 -0.38 0.04 0.57 0.51 0.64 0.86 0.79 0.86 0.83 0.79 0.99 1.00 -0.11 0.18 -0.13 0.04
Is Holiday 0.04 -0.05 -0.05 0.09 0.04 -0.03 -0.05 -0.01 -0.09 -0.06 -0.08 -0.10 -0.10 -0.08 -0.08 -0.10 -0.11 -0.11 1.00 -0.06 -0.26 0.36
Weekend -0.26 -0.26 -0.26 -0.26 -0.26 0.65 0.64 -0.01 0.12 0.04 0.07 0.18 0.10 0.14 0.21 0.10 0.19 0.18 -0.06 1.00 -0.94 0.10
Is Weekday 0.24 0.27 0.26 0.22 0.24 -0.61 -0.61 0.02 -0.09 -0.02 -0.04 -0.13 -0.05 -0.11 -0.17 -0.07 -0.14 -0.13 -0.26 -0.94 1.00 -0.09
Is Holiday and Weekend -0.03 -0.03 -0.03 -0.03 -0.03 0.09 0.03 0.04 -0.03 -0.05 0.01 0.08 0.09 -0.00 0.02 -0.03 0.05 0.04 0.36 0.10 -0.09 1.00
Monday Tuesday Wednesday Thursday Friday Saturday Sunday October Calamari Fish Shrimp Chicken Koefte Lamb Steak Total Fish Production Total Meat Production Total Is Holiday Weekend Is Weekday Is Holiday and Weekend
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75


Weather Only Feature Correlative Relationships
Fish 1.00 0.29 0.38 0.35 0.40 0.41 0.64 0.46 0.51 0.01 0.07 -0.02 -0.10 -0.11 -0.06 0.04 -0.02 -0.05
Shrimp 0.29 1.00 0.50 0.39 0.50 0.48 0.84 0.56 0.64 0.06 0.02 -0.03 -0.03 0.03 -0.08 0.07 -0.04 0.01
Chicken 0.38 0.50 1.00 0.60 0.66 0.63 0.60 0.87 0.86 0.02 0.07 -0.04 -0.12 -0.09 -0.10 0.18 -0.13 0.08
Koefte 0.35 0.39 0.60 1.00 0.58 0.58 0.52 0.80 0.79 -0.06 0.11 -0.05 -0.16 -0.11 -0.10 0.10 -0.05 0.09
Lamb 0.40 0.50 0.66 0.58 1.00 0.61 0.61 0.86 0.86 0.05 0.06 -0.05 -0.09 -0.05 -0.08 0.14 -0.11 -0.00
Steak 0.41 0.48 0.63 0.58 0.61 1.00 0.62 0.83 0.83 0.05 0.08 -0.01 -0.17 -0.18 -0.08 0.21 -0.17 0.02
Total Fish Production 0.64 0.84 0.60 0.52 0.61 0.62 1.00 0.70 0.79 0.07 0.06 -0.03 -0.09 -0.06 -0.10 0.10 -0.07 -0.03
Total Meat Production 0.46 0.56 0.87 0.80 0.86 0.83 0.70 1.00 0.99 0.02 0.09 -0.05 -0.16 -0.12 -0.11 0.19 -0.14 0.05
Total 0.51 0.64 0.86 0.79 0.86 0.83 0.79 0.99 1.00 0.03 0.09 -0.04 -0.15 -0.12 -0.11 0.18 -0.13 0.04
Wind 0.01 0.06 0.02 -0.06 0.05 0.05 0.07 0.02 0.03 1.00 0.11 0.05 -0.06 -0.02 -0.05 0.00 0.00 -0.07
Cloudiness 0.07 0.02 0.07 0.11 0.06 0.08 0.06 0.09 0.09 0.11 1.00 0.20 -0.86 -0.34 0.00 0.04 -0.05 -0.04
Precipitation -0.02 -0.03 -0.04 -0.05 -0.05 -0.01 -0.03 -0.05 -0.04 0.05 0.20 1.00 -0.22 -0.02 0.07 0.03 -0.05 -0.02
Sun -0.10 -0.03 -0.12 -0.16 -0.09 -0.17 -0.09 -0.16 -0.15 -0.06 -0.86 -0.22 1.00 0.58 -0.00 0.00 -0.00 0.03
Air Temperature -0.11 0.03 -0.09 -0.11 -0.05 -0.18 -0.06 -0.12 -0.12 -0.02 -0.34 -0.02 0.58 1.00 -0.02 -0.01 0.02 -0.01
Is Holiday -0.06 -0.08 -0.10 -0.10 -0.08 -0.08 -0.10 -0.11 -0.11 -0.05 0.00 0.07 -0.00 -0.02 1.00 -0.06 -0.26 0.36
Weekend 0.04 0.07 0.18 0.10 0.14 0.21 0.10 0.19 0.18 0.00 0.04 0.03 0.00 -0.01 -0.06 1.00 -0.94 0.10
Is Weekday -0.02 -0.04 -0.13 -0.05 -0.11 -0.17 -0.07 -0.14 -0.13 0.00 -0.05 -0.05 -0.00 0.02 -0.26 -0.94 1.00 -0.09
Is Holiday and Weekend -0.05 0.01 0.08 0.09 -0.00 0.02 -0.03 0.05 0.04 -0.07 -0.04 -0.02 0.03 -0.01 0.36 0.10 -0.09 1.00
Fish Shrimp Chicken Koefte Lamb Steak Total Fish Production Total Meat Production Total Wind Cloudiness Precipitation Sun Air Temperature Is Holiday Weekend Is Weekday Is Holiday and Weekend
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

Exploratory Data Analysis Conclusions:

From what we could see from our exploratory data analysis, while there are strong relationships between time-based variables such as days of the week or even days of the month, there aren't super strong relationships between nebulous variables such as weather and total demand. This can be rationalized by the idea that while weather has been thought to have strong influences on us as humans - the exact effect is not concrete and might vary from person to person. However, this doesn't mean that there is not any effect either. While the evidence from our EDA does not indicate the presence of any strong correlative relationships between weather and demand, it also doesn't indicate that said relationships do not exist. Our EDA has indicated that while weak there might be some relationships between different types of weather and overall demand.

Modeling:

Standardizing Values:

Now from what we can see, this data set already came with the days of the week, months, and year turned into dummy variables. So we can exclude these from any preprocessing steps for now and then join them back into our dataset before we move onto modeling.

As for value standardization, because there are vastly different kinds of units being measured across the various columns it is a good idea for us to standardize them. This way no one variable dominates the analysis being done due to the scale of its measurements. For example, we can see above that wind is measured in vastly different units than Sun - with a range of 10 units whereas Sun has a range of 782. In this step, since the weather categories had vastly different units, the weather columns wind, cloudiness, precipitation, sun, and air temperature were the only columns that needed to be standardized.

Splitting the data:

From here, we split our data on an 80/20 train-test split - focusing on the various types of demand. This would prove to be a challenging issue in this project since instead of just focusing on overall demand like we did in our EDA, we are now choosing to predict multiple factors such as total meat production and total fish production. If successful, this would allow us to identify what weather patterns drive certain demands in meat items or in fish items and what weather patterns drive overall demand across the board. Moving onto our model selection, we chose the following three models:

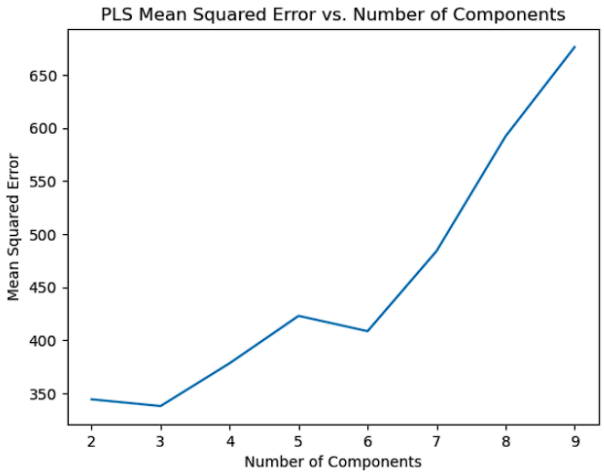
Edwin Ng

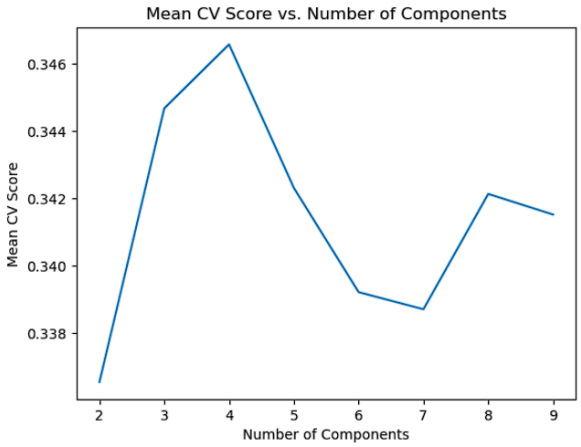
Linear Regression:

- The linear regression mean squared error: **1130.1110929801362**
- The linear regression model mean absolute percentage error: **325.9499954553569**
- **34.06%** accuracy with a standard deviation of **10.39**

Partial Least Squares Regression (PLS):

After iterating through various components to identify our best parameter, we found that the best number of components was 4.





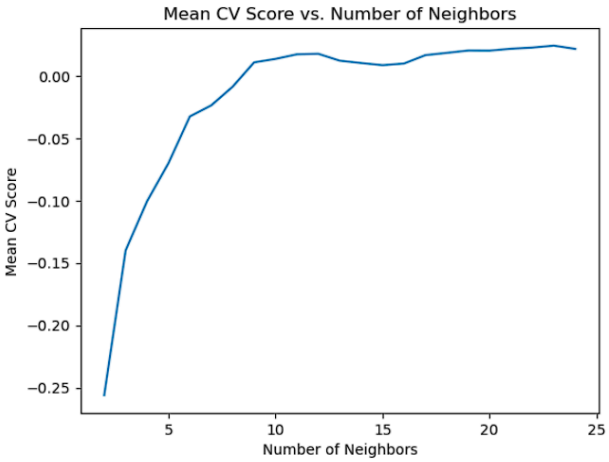
This yielded the results:

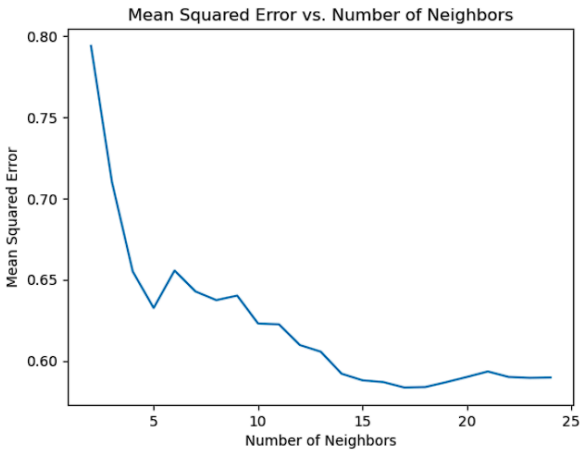
- The linear regression mean squared error: **378.3965955342321**
- The linear regression model mean absolute percentage error: **293.09247169633494**
- **34.66%** accuracy with a standard deviation of **10.62**

Edwin Ng

K Nearest Neighbors (KNN):

After iterating through various n neighbors to identify our best parameter, we found that the best number of neighbors was 23.





This yielded the results:

- The KNN mean squared error: **378.3965955342321**
- The KNN model mean absolute percentage error: **1.8644701453911603**
- **2.43%** accuracy with a standard deviation of **1.44**

Random Forest:

For our Random Forest regression model, we took the following approaches to optimizing the parameters of the model:

Randomized Grid Search with CV:

Again using GridSearchCV and the following parameter grid:

```
param_distributions={'bootstrap': [True, False],
                    'max_depth': [5, 6, 7, 8, 9, 10],
                    'max_features': [0.3, 0.4, 0.5],
                    'min_samples_leaf': [2, 3, 4, 5, 6],
                    'min_samples_split': [2, 4, 6, 8, 10,
                                          20, 40, 80],
```

```
'n_estimators': [20, 50, 100, 200, 400,
                 600, 800, 1000, 1200,
                 1400, 1600, 1800,
                 2000]}
```

We found that the best parameters for our model were:

RandomForestRegressor(max_depth=7, max_features=0.3, min_samples_leaf=2,
min_samples_split=6, n_estimators=800)

Earning a score of:

- The randomized search random forest regression mean squared error: **0.6041993547375255**
- The randomized search random forest regression regression model mean absolute percentage error: **3.1700858599681427**
- **42.39** percent accuracy with a standard deviation of **10.70**

Bayesian Optimization:

Using the results from our randomized grid search process, we constructed a new parameter grid for the Bayesian Optimization grid search:

```
params_new = {
    'bootstrap': [True],
    'criterion':['squared_error'],
    'min_samples_leaf': [2,3,4,5],
    'max_depth': [3,4,5,6,7,8],
    'min_samples_split': [5,6,7,8,9],
    'max_features': [0.2,0.25,0.3,0.35,0.4],
    'n_estimators': [1000,1100,1200,1300,1400,1500, 1600,1700,1800,1900,2000]
}
```

From here, we found that our best parameters were:

RandomForestRegressor(bootstrap=True, criterion='squared_error', max_depth=8,
max_features=0.3, min_samples_leaf=3, min_samples_split=8,
n_estimators=1400)

Earning a score of:

- The bayesian optimization random forest regression mean squared error: **0.5886198228920206**
- The bayesian optimization random forest regression regression model mean absolute percentage error: **3.0606415017178636**
- **42.32** percent accuracy with a standard deviation of **9.25**

Support Vector Regression (SVR):

For this approach, in an attempt at reducing the difficulty of the predictions - I chose to try a support vector regression model. This way, we could separately predict the responding variables - this made it so that the model could predict the responding variables one at a time instead of trying to predict all 3 all at once. This approach earned the following scores:

- The SVR model mean squared error is: **0.8531301195919251**
- The SVR Model mean absolute percentage error is: **3.299470230116503**
- **-4.66** percent accuracy with a standard deviation of **4.21**

In the end our model performances were ranked as follows:

1. Random Forest Regression Model with an accuracy of 42.39%
2. Partial Least Squares Model with an accuracy of 34.66%
3. Linear Regression Model with an accuracy of 34.06%
4. K Nearest Neighbors Model with an accuracy of 2.43%
5. SVR Model with an accuracy of -4.66%

Modeling Conclusions:

Among the many models that we tried, it is clear that the Random Forest model performed the best. Surprisingly enough the support vector regression model performed the worst. This might be because of the high amounts of collinearity between the responding variables - therefore making any differences between the predicted variables that much worse when the problem was split into 3 different single-variable regression problems. Finally, while it was disappointing to see, out of all the models that we tried, we were unable to create a model that performed well. This is mostly in part due to the difficult nature of the problem. To start, there were no strong relationships to base these models on. While there were strong relationships between variables such as overall demand and days of the week, the relationships between demand and weather

variables were slim to none. On top of this, this problem was multivariate in nature - adding that much more complexity to an already strenuous problem. However, as a challenge in working with multivariate regression problems - I believe I put my best foot forward.

Further Investigations:

Overall, while I was disappointed by the fact that I was not able to uncover any worthwhile information as to whether weather plays an impact on customer demand, this project allowed me to gain a deeper understanding of multivariate regression. It showed me that there are many factors at play when it comes to model predictive success and problem complexity. It also helped me to understand the importance of exploratory data analysis and highlighted the different warning signs of problem complexity and model prediction difficulties that I should pay attention to before moving on to modeling. However, while my project so far has enabled me to get a deeper understanding of what goes into shaping the outcomes of regression problems - it has also left me with more questions.

What kind of data would enable this kind of analysis to be performed successfully? What kinds of data was I missing and how else could I have created stronger models given the data I have? As for my understanding of how weather impacts customer decisions, it has also left me wondering about other scenarios. For example, are there weather-driven impacts on other consumer sectors such as retail, entertainment, event attendance, etc.? Also, what impacts does location have - with regards to culture? For example, our dataset here was based in Germany. One thing we noticed during our EDA was that restaurant demand was statistically significantly lower than the rest of the week on Sundays - does this also ring true for restaurants in North America, Asia, or the rest of Europe? How does culture impact consumer spending and consumer behaviors? There are still many frontiers to be explored from this project alone!