📊

Data Science Report

Made by:

Khaled Abdelreheam - 20217004
Ahmed Khaled - 20216004
Yassin Saleh - 20216117
Malak Mahmoud - 20217010
Malak Gamal - 20217009

Under Supervision of:

Dr. Ayman El-Kilany
TA. Mohamed Ramadan

Introduction and Goals

As in any project, there must be clear goals defined. In this Data Science Project we are represented with sales data across multiple stores in the U.S. (This is clear since the temperatures are in Fahrenheit). In this Data Science we have already obtained the data, so we need to understand it and then gain value, our target with this data is to be able to predict weekly sales accurately.

Data Cleaning

1- Data Descriptions:

- Fuel Prices:
    - The Data Contains time series data about the fuel price on each week
- Sales:
    - The Data Contains time series data about the sales of each store in each category on a specific week
- Weather:
    - The Data Contains time series data about the temperature on each week

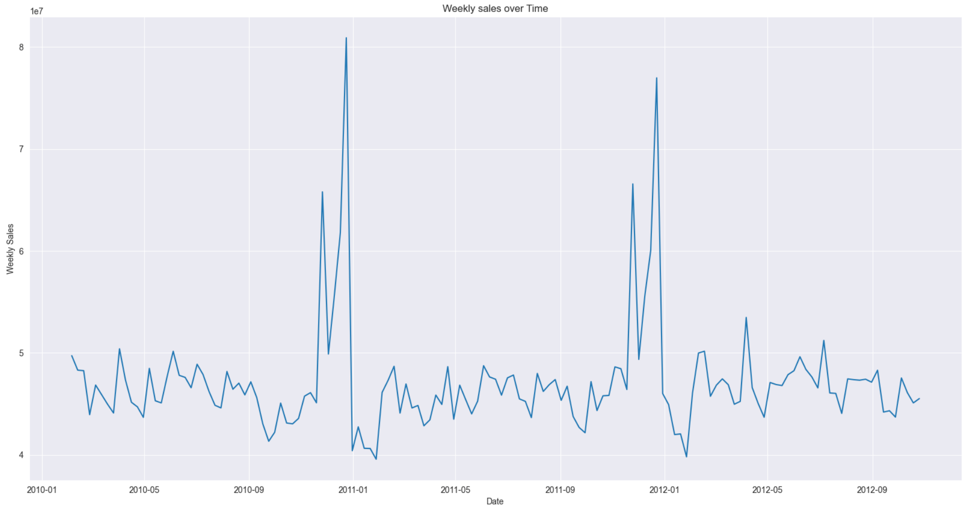
The CSV files don’t contain null values when read using Pandas, reports for statistical measures are visible in the notebook.

3- Data Preprocessing:

- Temperature values were rounded (converted to int) in order to make analysis easier.
- Negative Weekly Sales were dropped, since mean imputation will introduce bias towards the mean and their count as a percentage isn’t significant, Weekly Sales less than 300 (2% of the mean) were dropped as well since values near zero would affect the MAPE measure we use in Modeling significantly.
- Data was merged based on Date.

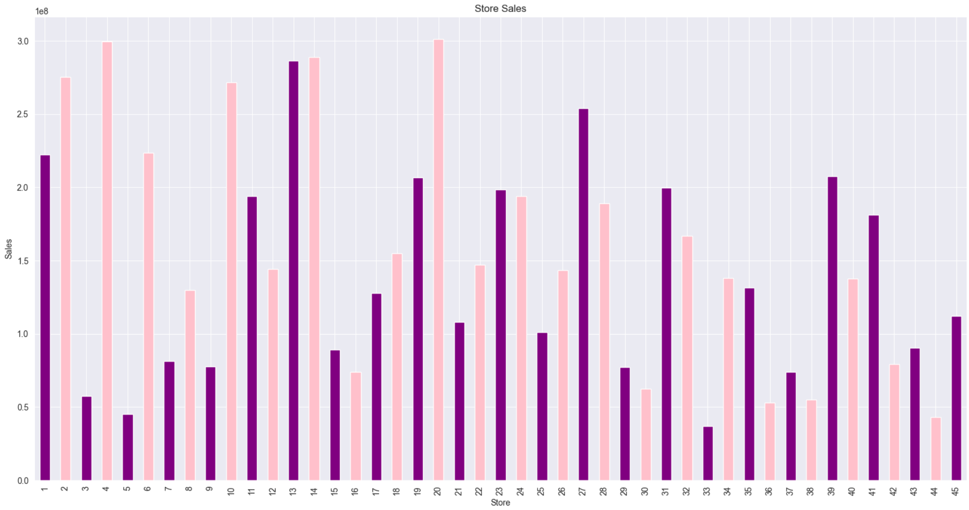
Data Visualization

1- Weekly Sales Over Time



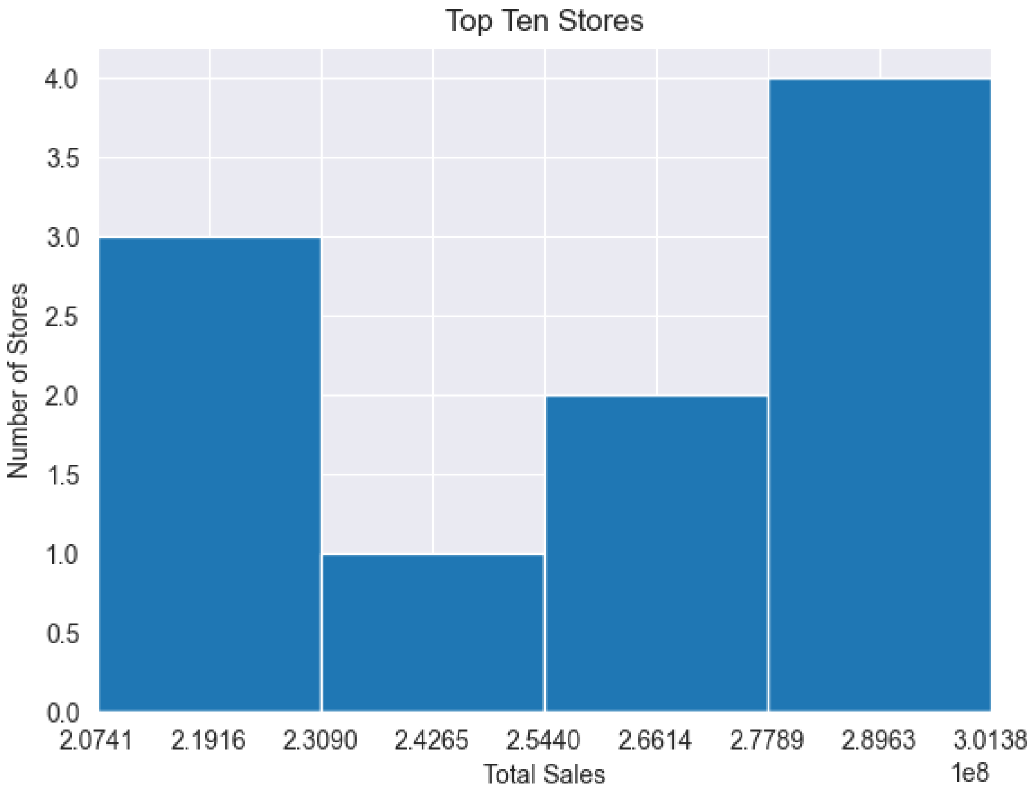
Based on the analysis of our line graph, it can be inferred that the sales of our stores reach their maximum levels just prior to the New Year. Subsequently, there is a fluctuation in sales both upward and downward after the New Year until they stabilize around the average sales figure.

2- Store Sales Bar Chart



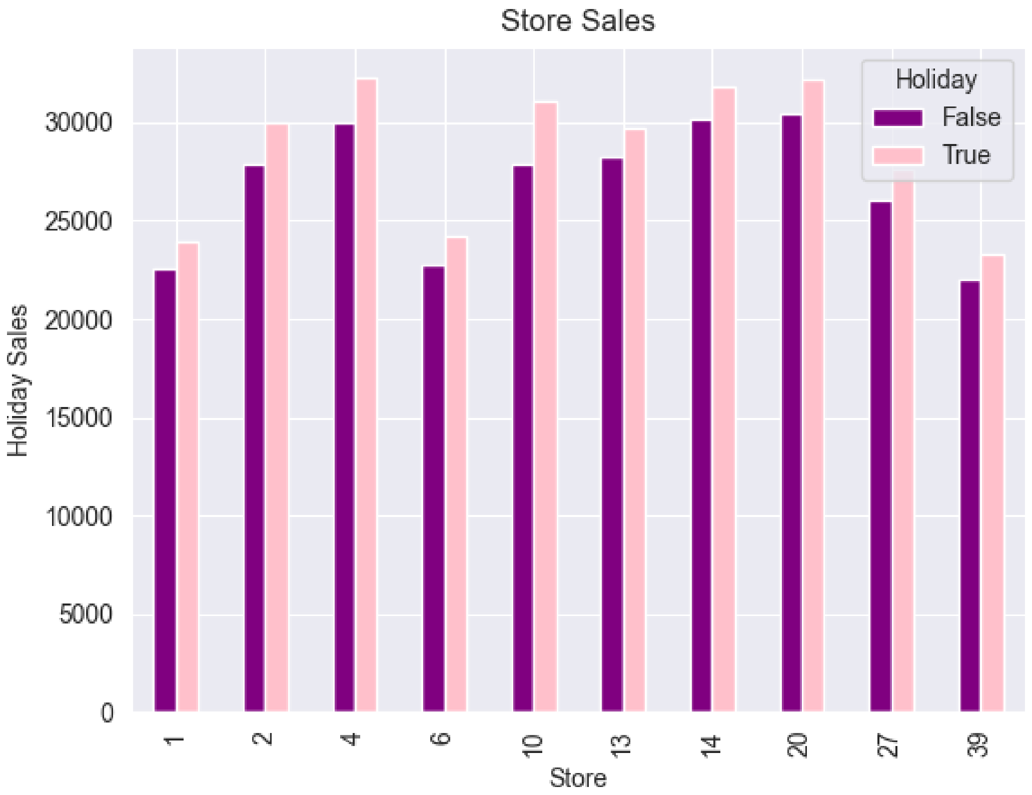
The bar graph clearly indicates that Store 20 boasts the highest weekly sales among all the stores.

4- Top 10 Stores



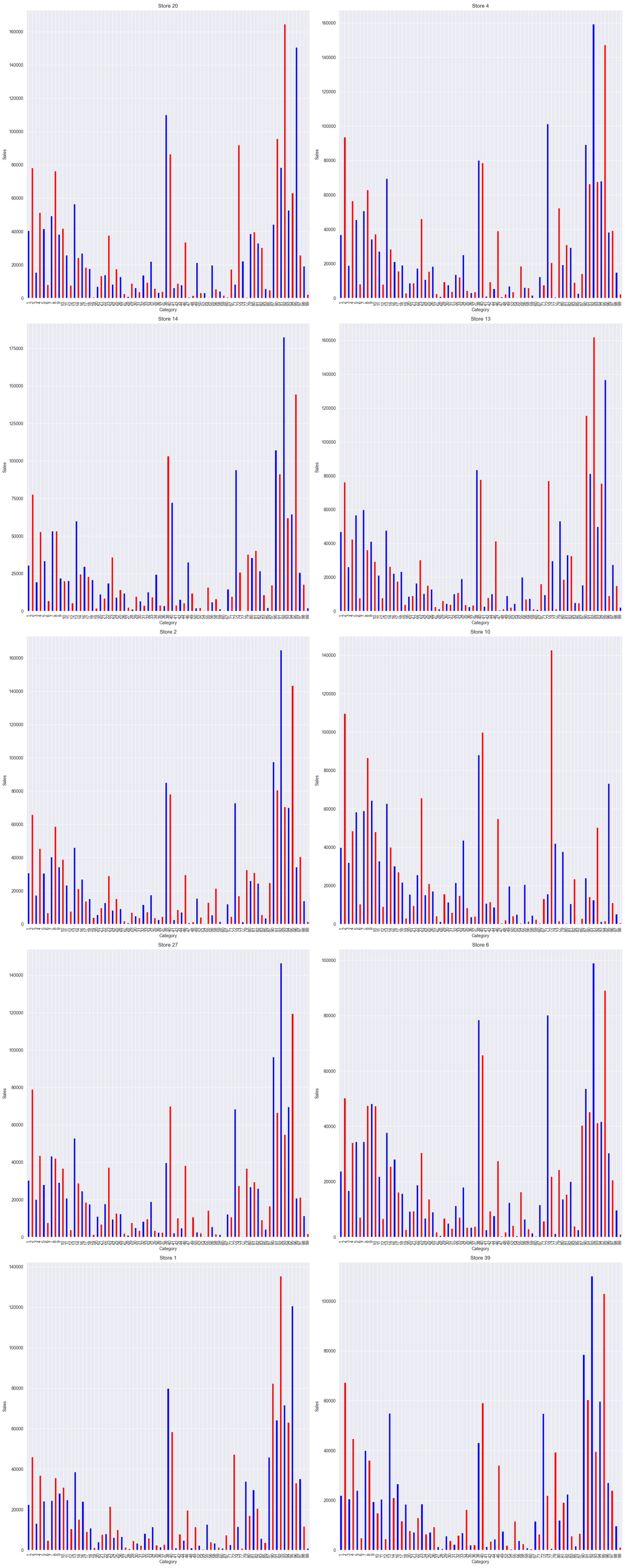
The histogram reveals that three stores exhibit sales of $207M to $230M, one store exhibit sales figures within the range of $231M to $254M, two stores exhibit sales in the range of $255M to $277M and four stores exhibit sales in the range of $278M to $301M.

5- Top Ten Store Sales (Holidays)



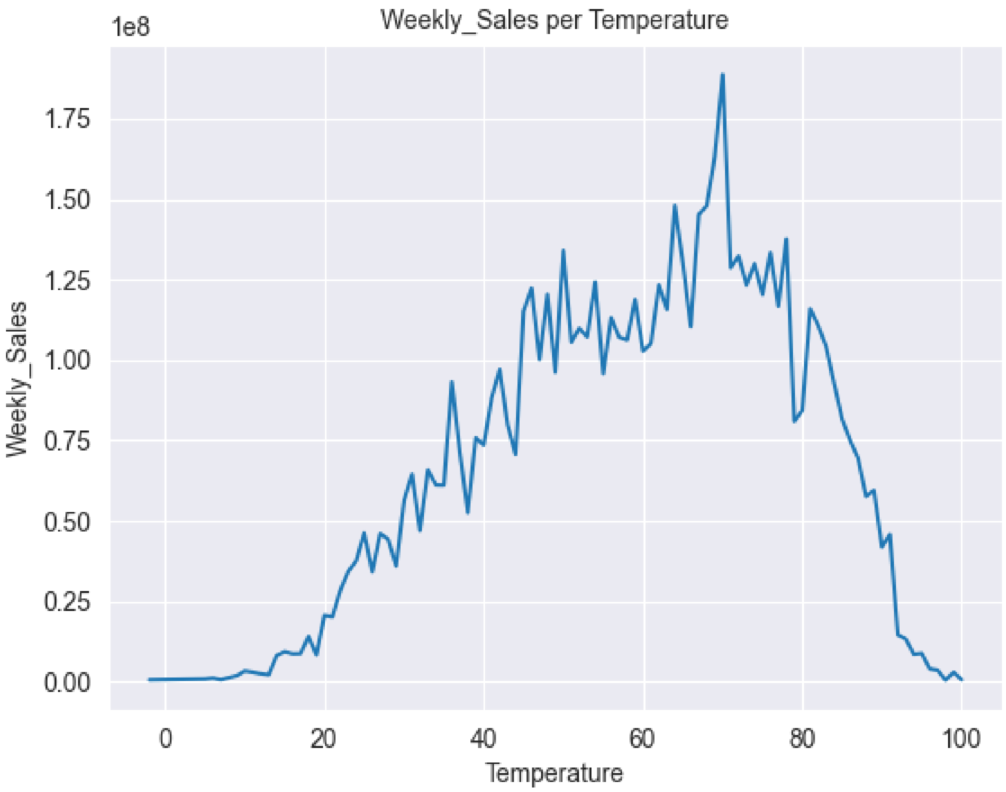
The corresponding bar chart depicting holiday sales illustrates that the top ten stores experience higher sales volumes during holidays as opposed to non-holiday periods. Store 20 emerges as the frontrunner in terms of sales during non-holiday periods, Stores 20 and 4 are tied for sales during holiday periods.

6- Categorical Sales in the Top 10 Stores


Store 20
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
160000
Store 4
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
160000
Store 14
Sales
Category
0
25000
50000
75000
100000
125000
150000
175000
Store 13
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
160000
Store 2
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
160000
Store 10
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
Store 27
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
Store 6
Sales
Category
0
20000
40000
60000
80000
100000
Store 1
Sales
Category
0
20000
40000
60000
80000
100000
120000
140000
Store 39
Sales
Category
0
20000
40000
60000
80000
100000

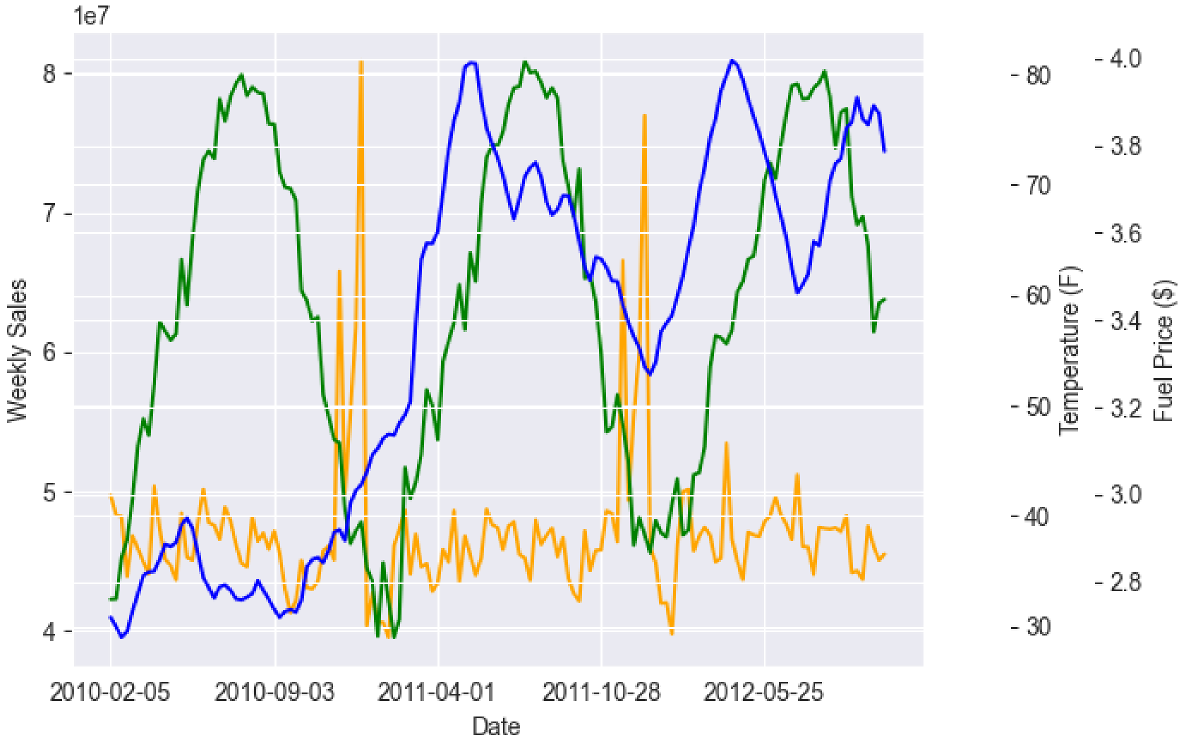
Based on the bar chart, the categories with the highest sales in all 10 stores are those with the lowest and highest numbers. Meanwhile, the categories with sales ranging from approximately 40 to 70 have the lowest sales in all 10 stores. (With some spikes in the median categories)

7- Sales per Temperature



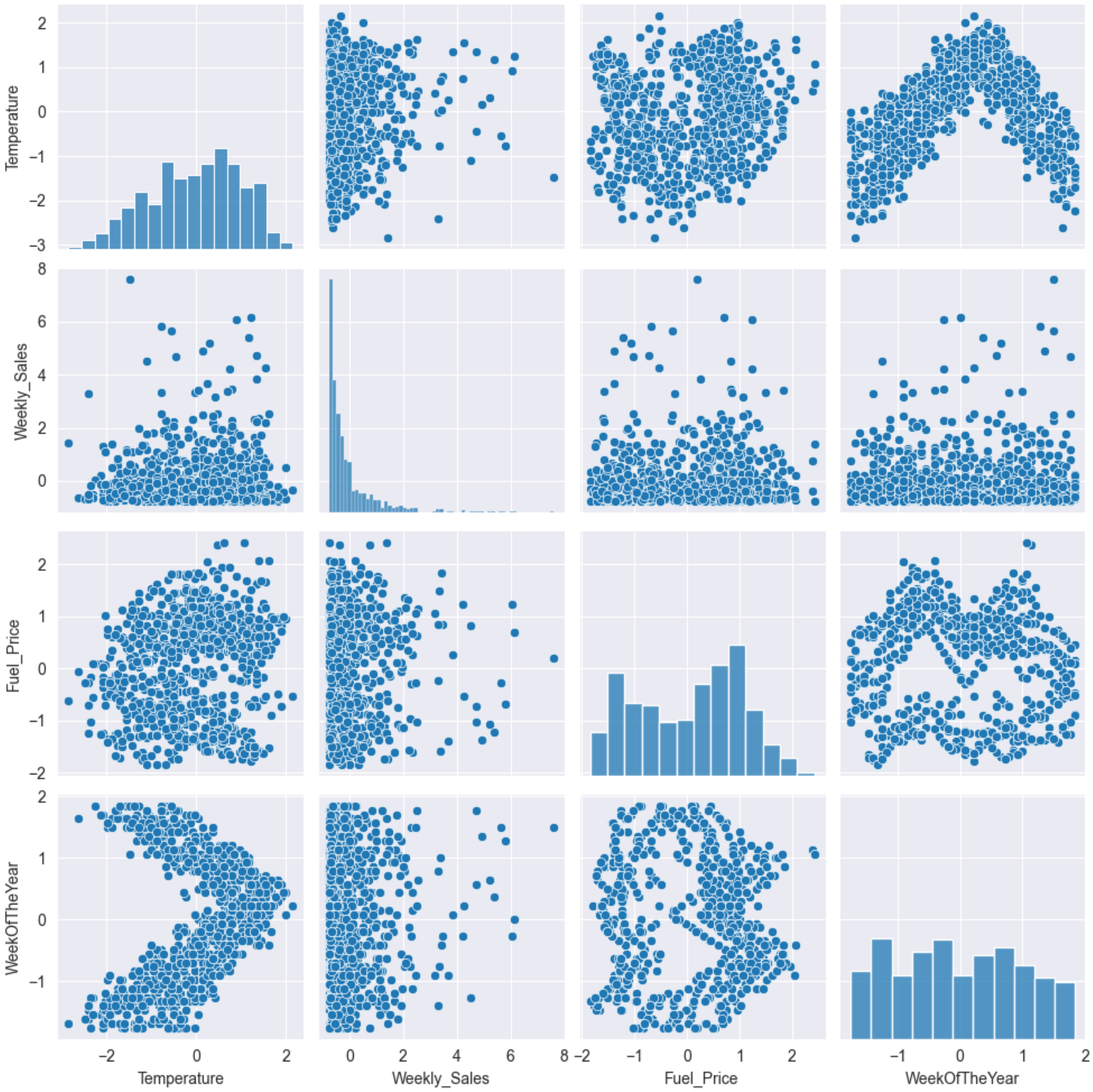
It is evident from the line graph titled "Weekly sales per temperature" that the weekly sales of our stores gradually increase until they reach a peak value. This peak occurs within a temperature range of 60 to 80 degrees. However, as the temperature surpasses 80 degrees and reaches 100 degrees, the sales show a declining trend. Furthermore, the lowest sales figures are observed at temperature extremes, namely around 0 and 100 degrees.

8- Weekly Sales, Temperature, Fuel Prices over Time



The chart indicates a negative correlation between weekly sales and fuel prices, suggesting that higher fuel prices tend to coincide with lower weekly sales. Although the relationship between temperature and weekly sales is less conclusive, there appears to be a weak positive correlation between the two variables. The general increase in Fuel Prices from the period of 2010 to 2011 is due to the increase of the price of crude oil during that period. Source: https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-bureau-economics-gasoline-price-changes-and-petroleum-industry-update/federal-trade-commission-bureau-economics-gasoline-price-changes-and-petroleum-industry.pdf . The Temperature over time follows the seasons (with peaks in August and lows in February)

9- Pairplot



The pair plot suggests that there is a strong correlation between the week of the year and temperature. This implies that when the temperature is low, it is most likely winter and when it is high, it is summer. Additionally, we can deduce that there is a weak correlation between weekly sales and all other attributes

Modeling

1- Preprocessing for ML

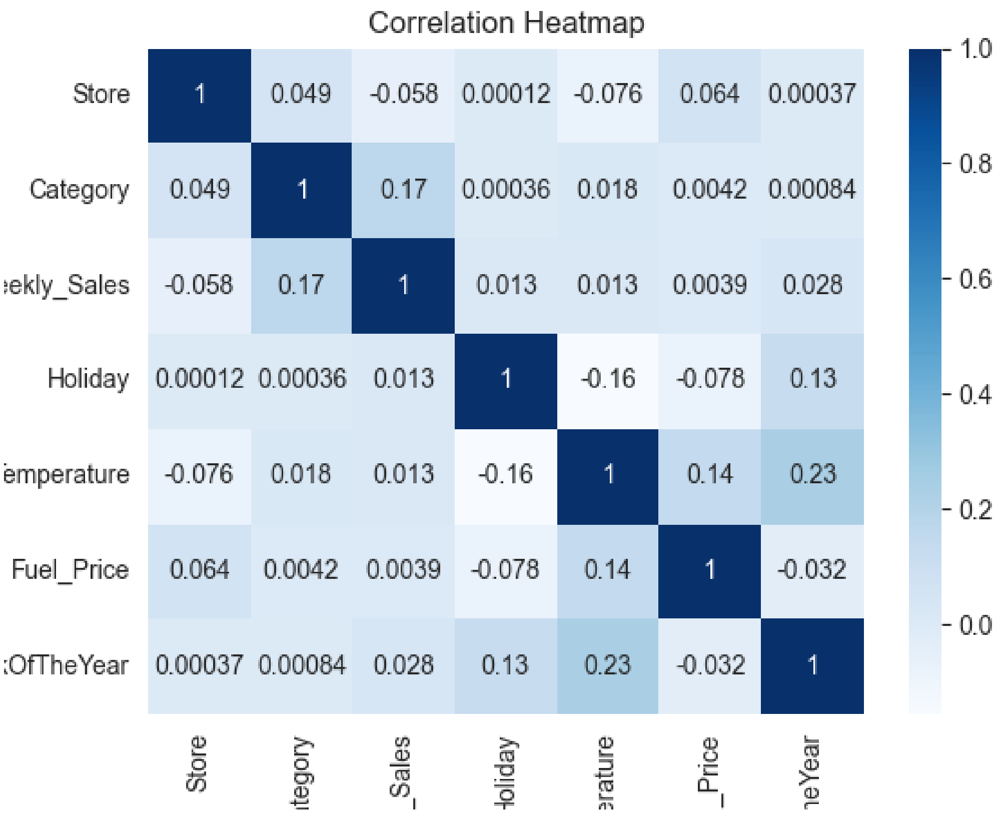
To be able to analyze the significance of the Date Attribute, we first converted it to a number signifying which Week of the Year it is in.

For the measures we choose the R^2 Score (R2 Score) and the Mean Absolute Percentage Error (MAPE), since MAPE is extremely sensitive to values around 0 we put in an offset by 100 for positive sales.

For KMeans we scaled the data since that improves the fitting prices and doesn’t over dramatized the measures.

2- Linear Regression

We first choose the features Temperatures, Fuel Price and Week of The Year since these features are what showed some correlation in the heatmap. Still, the correlation is weak and a linear model like linear regression wouldn’t give good results.



We used Train Test Split to Split the Data into 80% for training and 20% for testing. We then fit the model and calculated various measures:

```
Mean Absolute Error: 15632.170861346602
Mean Squared Error: 527408749.1680182
Root Mean Squared Error: 22965.381537610436
Mean Absolute Percentage Error: 4.6833322979876355 %
R2 score: 3.2610425373168495 %
```

The model doesn’t appear to show good performance, the MAPE measure is a bit acceptable but not great and the errors are just too big, the R2 score is very low as well.

3- Random Forest

For this non-linear model we choose to include all the features present, since a Random Forest can handle non-linear relationships.

We used Train Test Split to Split the Data into 80% for training and 20% for testing. We then fit the model and calculated various measures:

Mean Absolute Error: 1737.5696874359924
Mean Squared Error: 19167841.758979432
Root Mean Squared Error: 4378.1093817970595
Mean Absolute Percentage Error: 0.1684238845940851 %
R2 score: 96.48417469088547 %

The model performance improved immensely, the MAE and RMSE values are acceptable for the scale of our data, the MAPE value and R2 Score are near perfect.

4- XGBoost

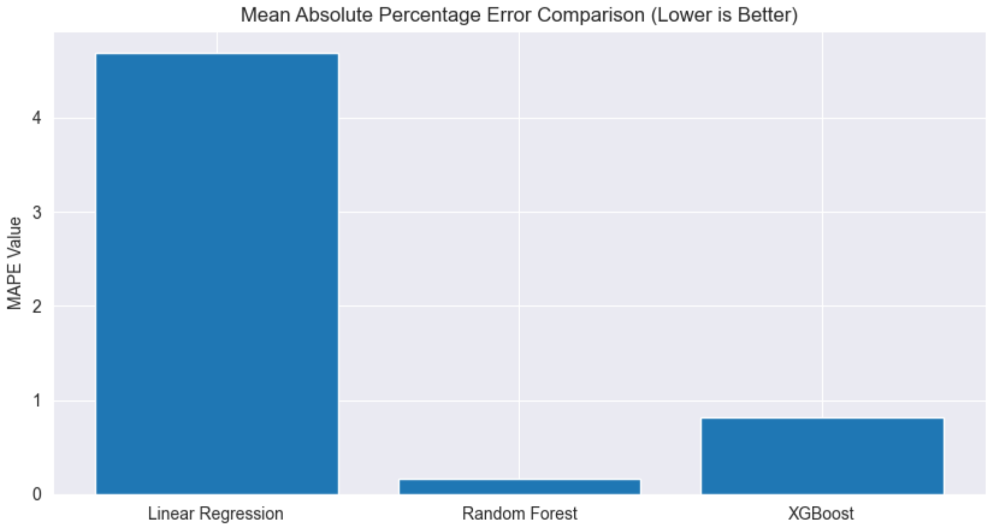
For this model we also choose to include all the features present, since an XGBoost Classifier can handle non-linear relationships.

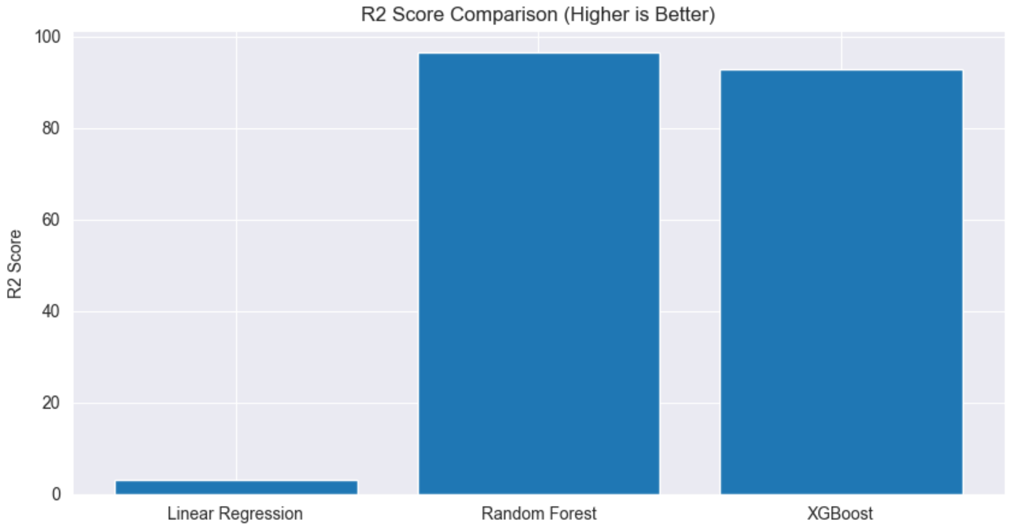
We used Train Test Split to Split the Data into 80% for training and 20% for testing. We then fit the model and calculated various measures:

Mean Absolute Error: 3751.3522344930307
Mean Squared Error: 39285028.71486795
Root Mean Squared Error: 6267.777015407292
Mean Absolute Percentage Error: 0.8121234276875631 %
R2 score: 92.79421752528191 %

The model is a clear improvement over Linear Regression, but it’s performance doesn’t live up to the Random Forest Classifier.

5- Model Evaluation





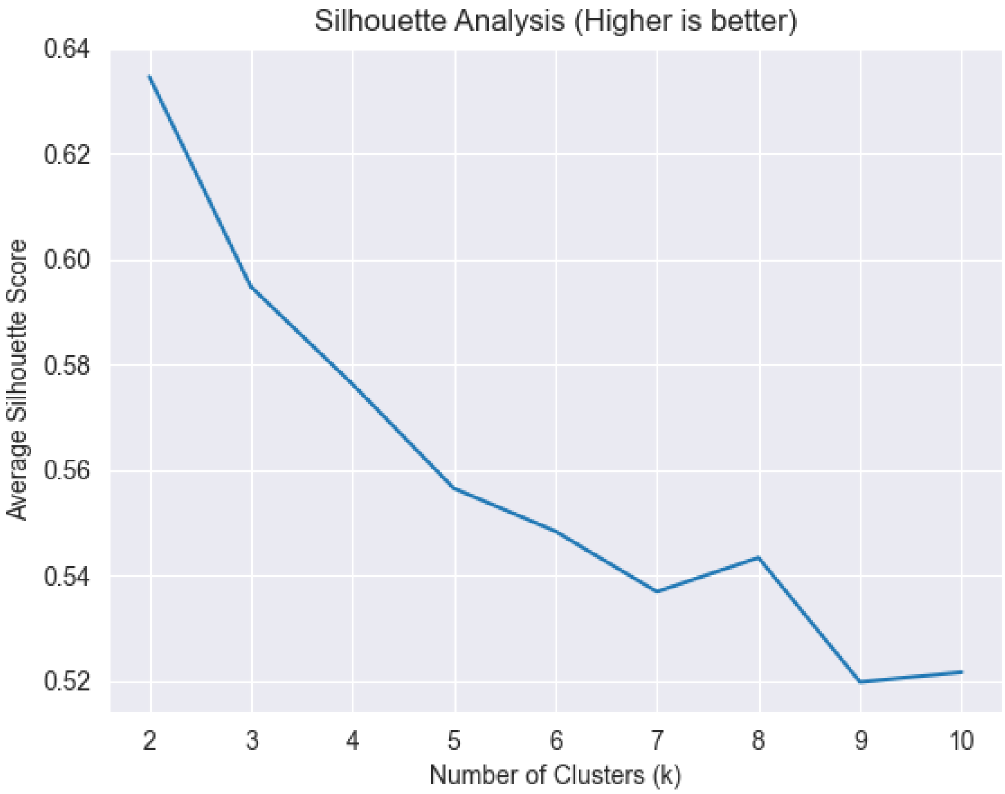
As we saw, the Random Forest Classifier consistently had the best performance, this can be seen in the following charts comparing R2 Score and MAPE (Which are the metrics we want to optimize for):

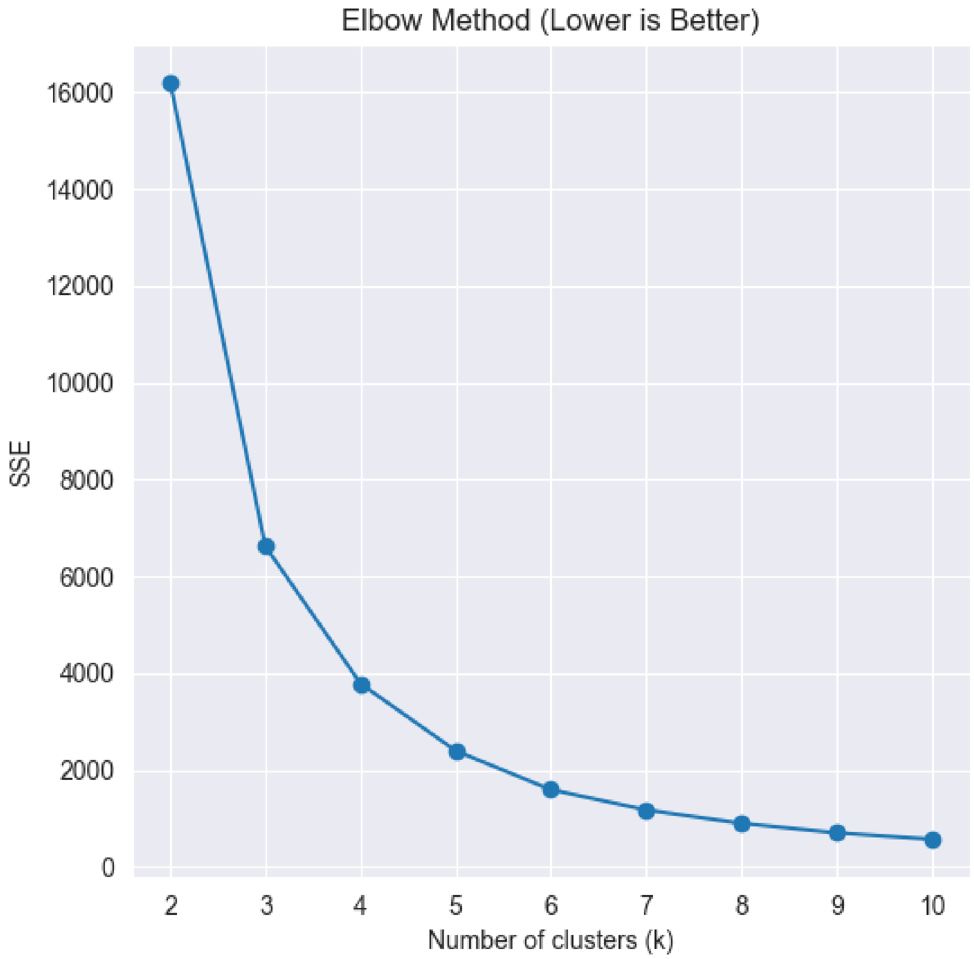
The results makes sense, since XGBoost and Random Forests can capture non-linear relationships in the Data.

6- KMeans

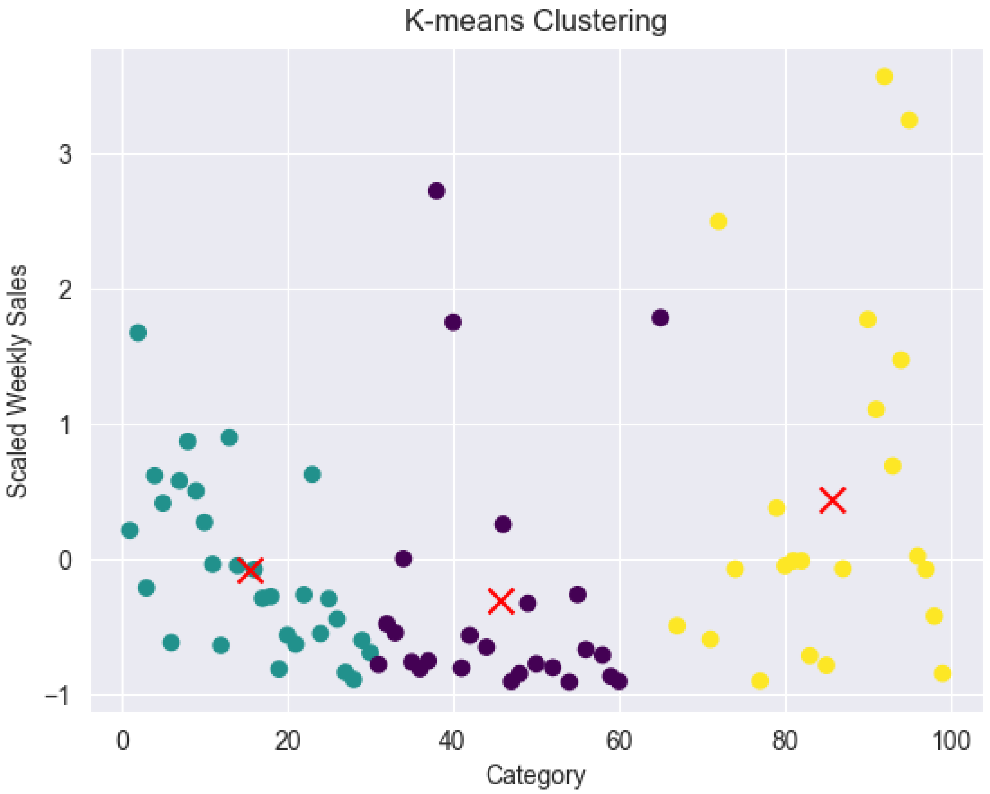
First we grouped data by category, summing sales in similar categories.

Then we employed 2 Measures to find the optimal number of Clusters, Sillhoutee Analysis (Higher is Better) and the Elbow Method (Lower is Better), we can see the Graphs below.


Silhouette Analysis (Higher is better)
Average Silhouette Score
0.64
0.62
0.60
0.58
0.56
0.54
0.52
2
3
4
5
6
7
8
9
10
Number of Clusters (k)



From the Graphs it appears that the Sillhoutee Analysis suggests the optimal number of Clusters to be 2 or 3, the Elbow Method Suggests a value in [3,4,5,6]. After Visualizing the clusters and given these methods we decided on 3 for the number of clusters, the clustered data can be visualized below:



The Clusters don’t appear to be cleanly seperated as we hope for, this implies that the data we have don’t allow us to seperate the categories as well as we hope for, despite that, the Clusters we have are linked to Visualization #6 in the Top 10 stores, where higher and lower categories tended to have better sales than middle ones. For Clusters over 3-Means it is clear that there isn’t that much of an improvement.

Results & Next Steps

From the Correlation map and Pair plot, we saw that the relationships in the data are mostly non-linear, this evident from the performance with the Linear Regression model performing weakly and the strong performance of non-linear models such as Random Forests and XGBoost Classifiers. The Random Forest classifier is the most suited to the Data available at hand. We then can see that for KMeans that there isn’t a clear seperation between different categories, like the visualization in number 6 (Visualization section) we can see that the upper and lower numbered categories tend to have higher sales than the lower numbered ones.

As for the next steps, we can collect data on other economic measures at the time to explain our results better, focus on the categories in the middle numbers to see why they are generally weaker in comparison to lower and higher numbered ones, we should collect further data on each category as well to try to Cluster our results better.