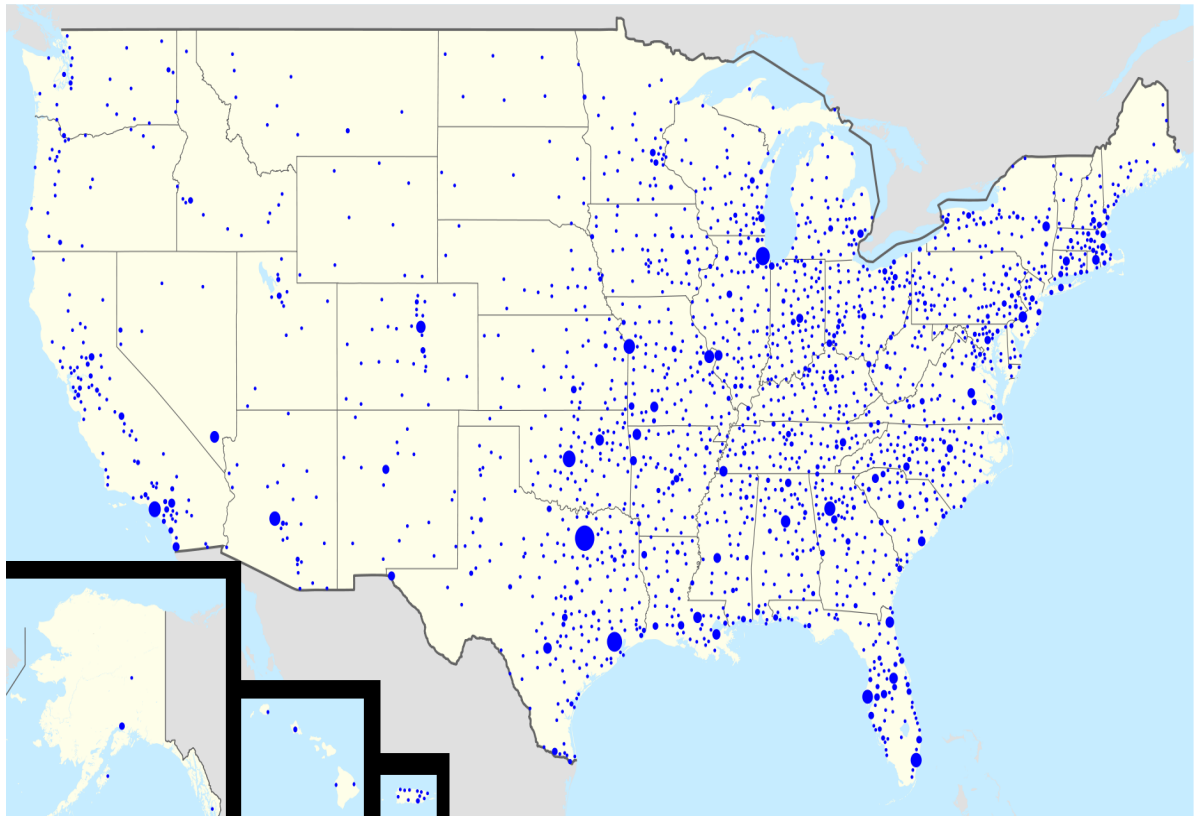
PREDICTING SALES MODEL

Presented by: Arnaldo Alonso

Problem Identification

- On this project, we made a model that predicts the amount of weekly sales a Walmart store could have as a function of variables such as store, and department number, unemployment rate of the area, CPI, size and type of store, and many others.

What is Walmart?

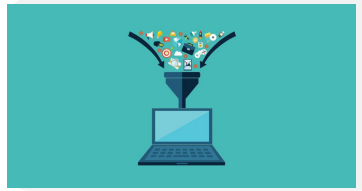


- Walmart is a multinational retail corporation characterized for the competitive prices on their products.
- Walmart is a corporation which has operations in all the states of the United States.
- Walmart is the world's largest company by revenue according to the Fortune Global 500 list.

Who might be interested?

- Walmart's CEO.
- Walmart's shareholders.
- Walmart's stakeholders

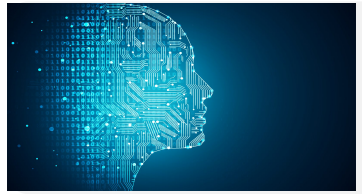
Steps



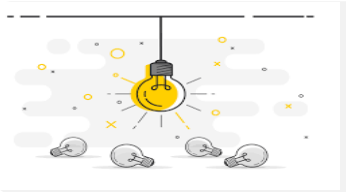
Data Collection

Data Wrangling

Data Visualization



Modeling



Conclusions

Data Collection

- Data was collected from Kaggle dataset and yahoo finance.
- Number of data tables: 4.
- Number of features: 17.
- Dimensions: 409,727 rows and 17 columns.

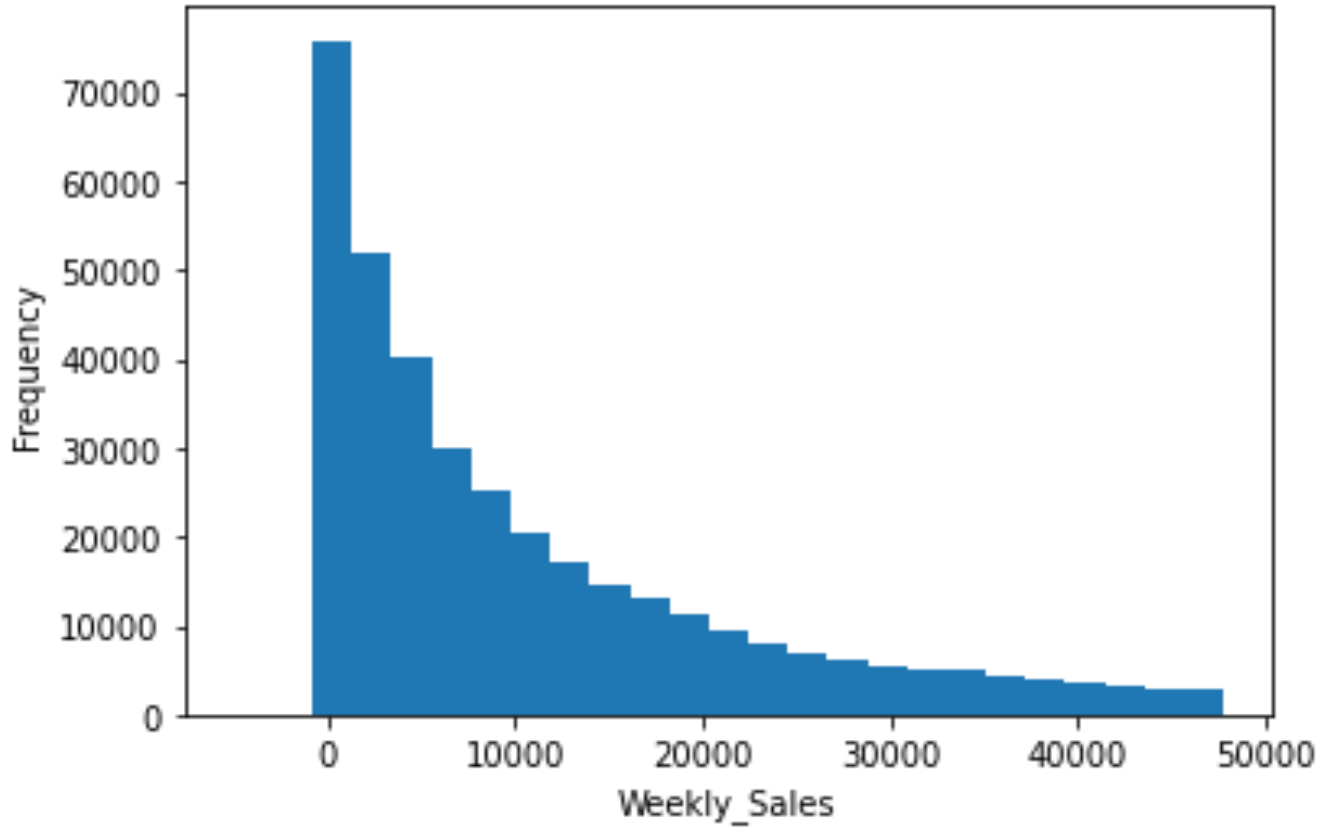
Data Wrangling

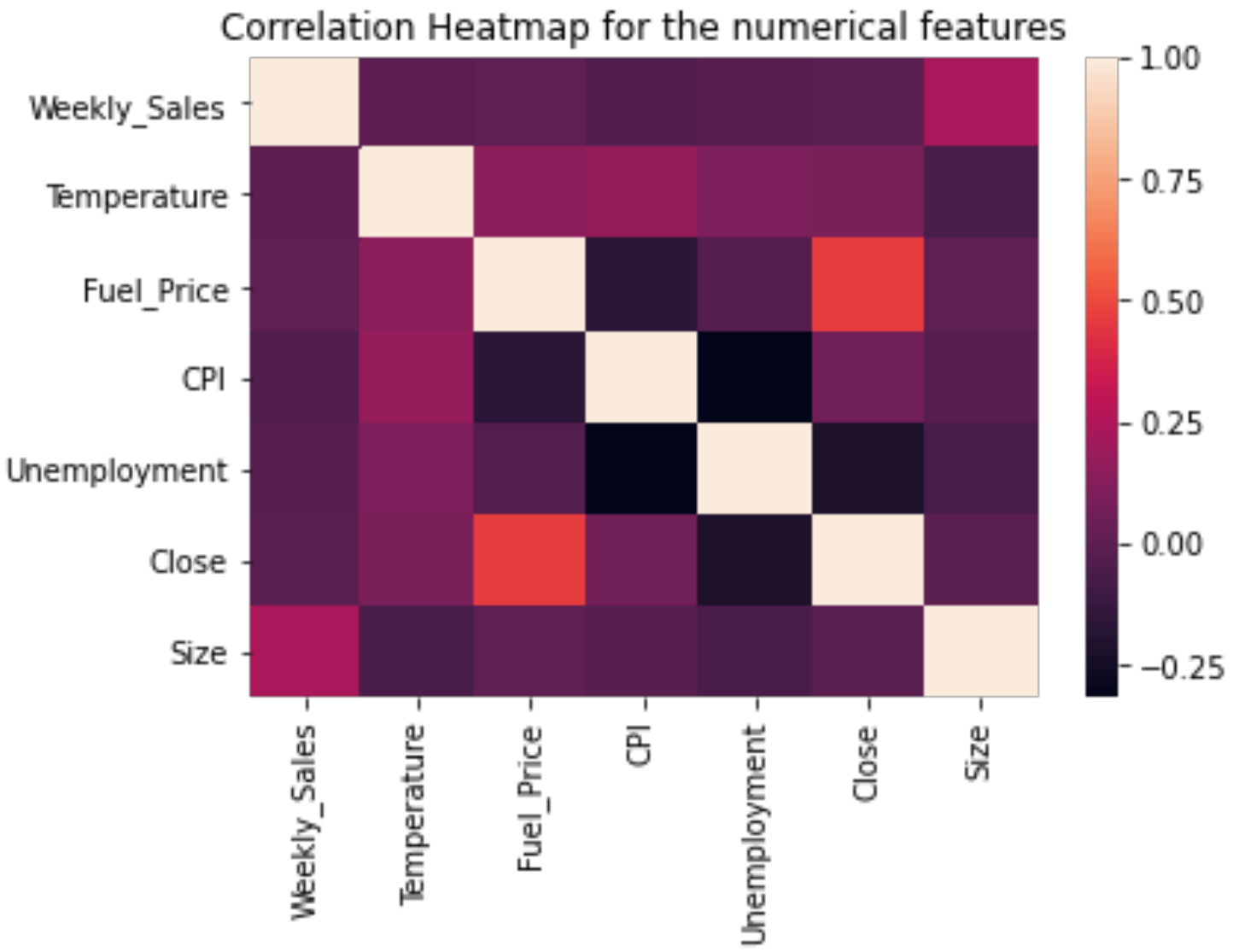
1. Pandas function: Inner Merge.

2. Handling null-values: pandas fillna function.

3. Transform to categorical features: Pandas function get dummies.

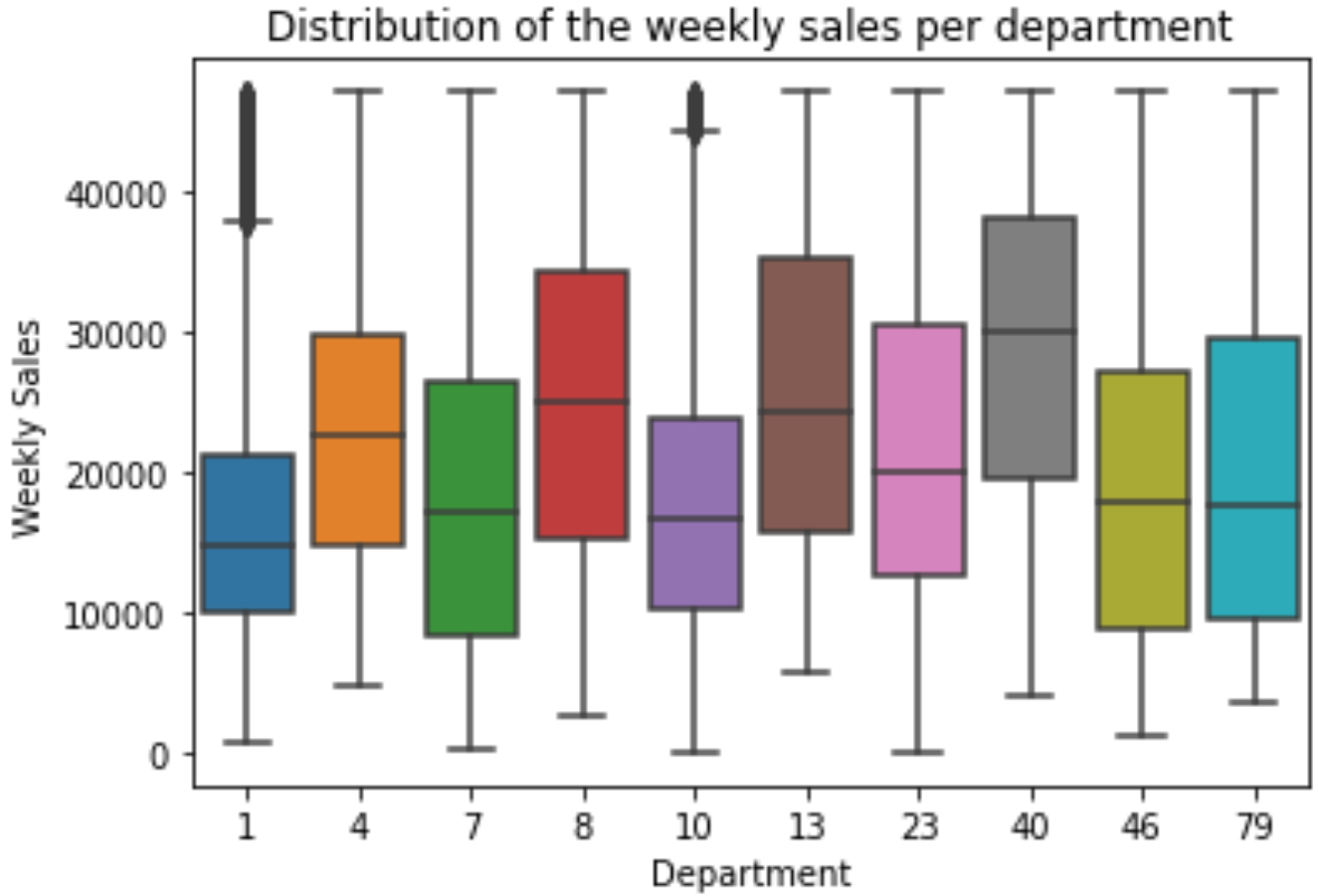
Data Visualization



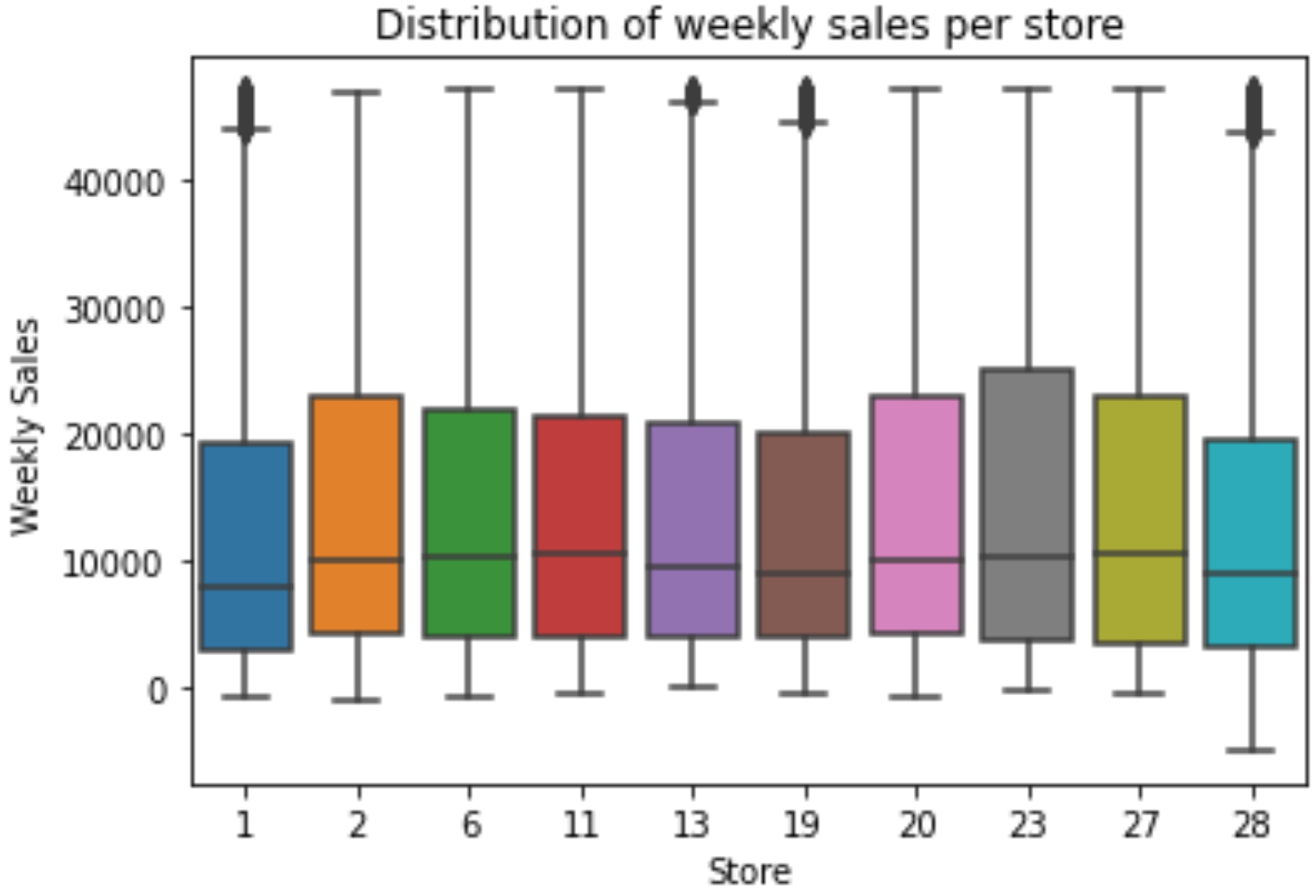


Correlation Heatmap

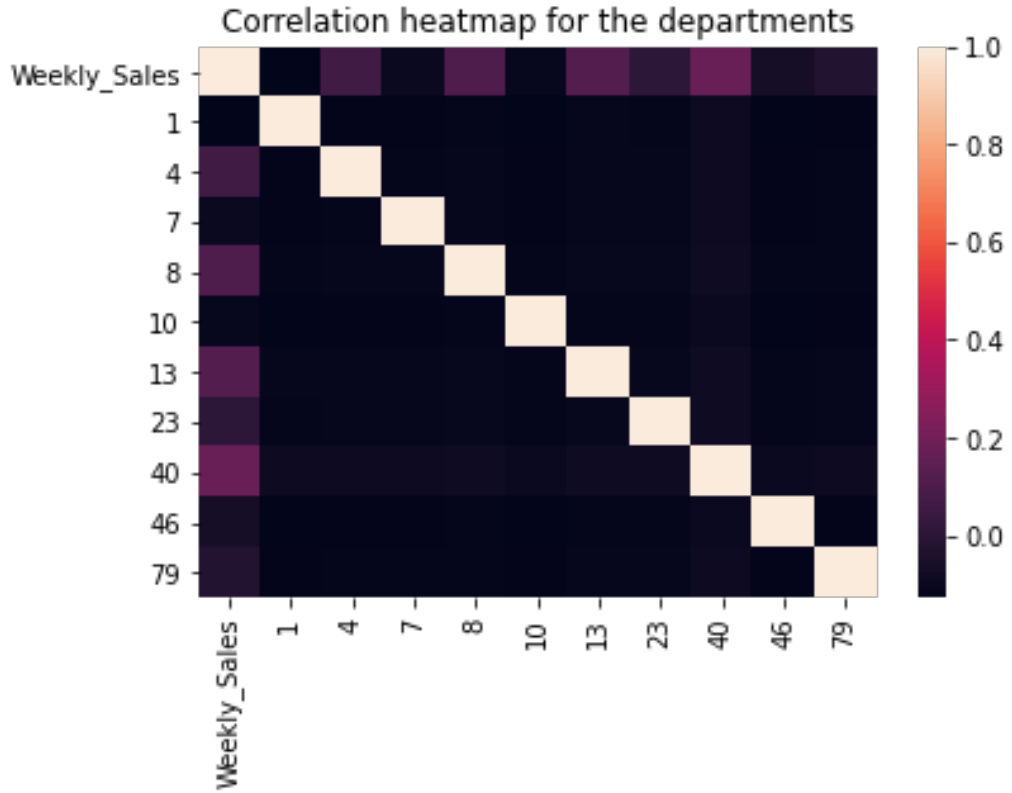
Distribution by department

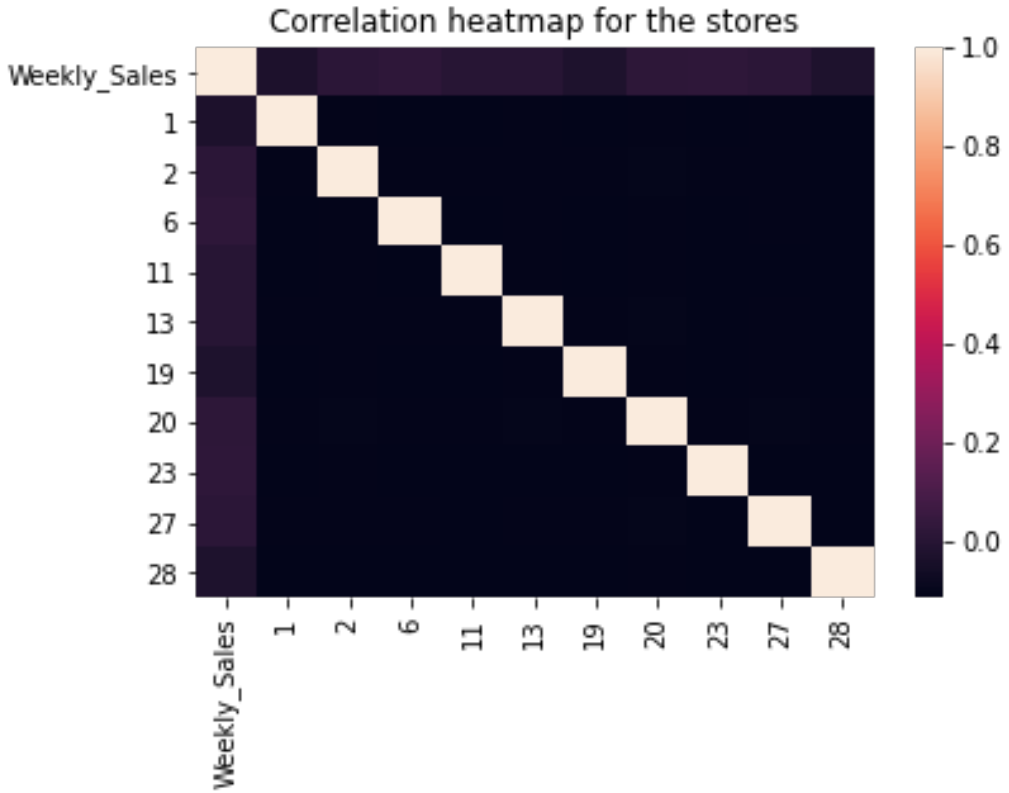


Distribution by store



Departments and Stores Correlation Heatmap





Hypothesis Testing

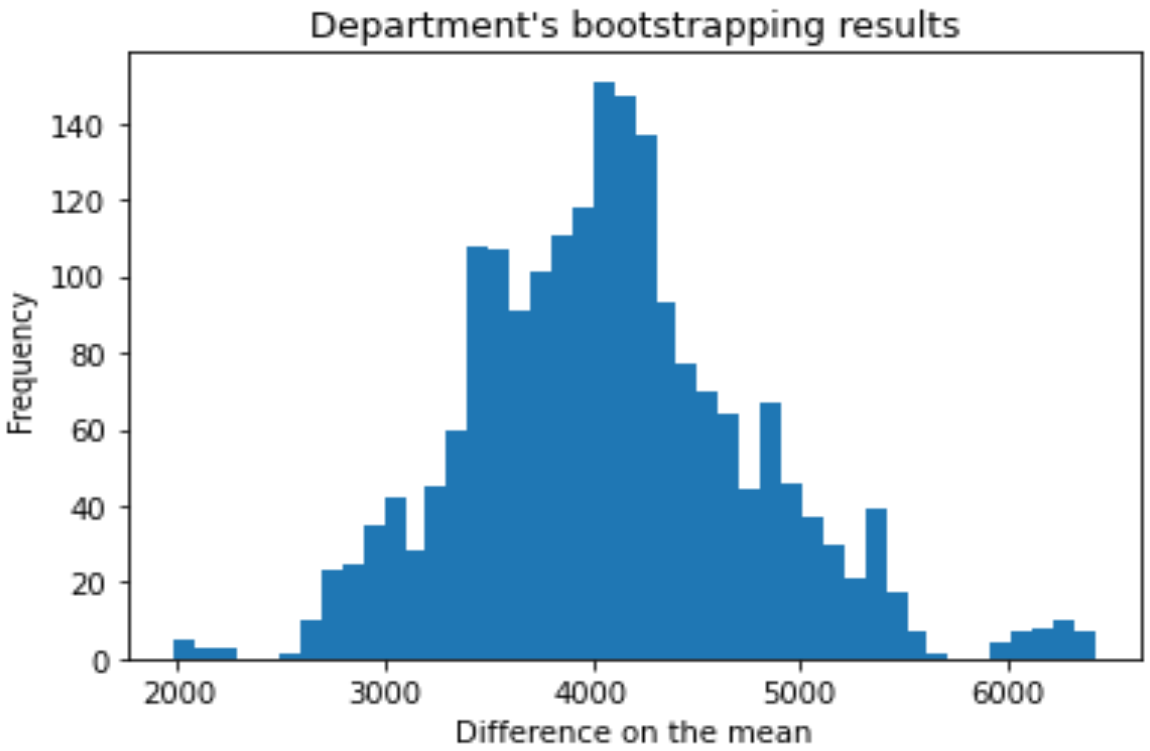
- For department

“The average of weekly sales for store 23 would tend to be higher than the average of the other 9 stores”.

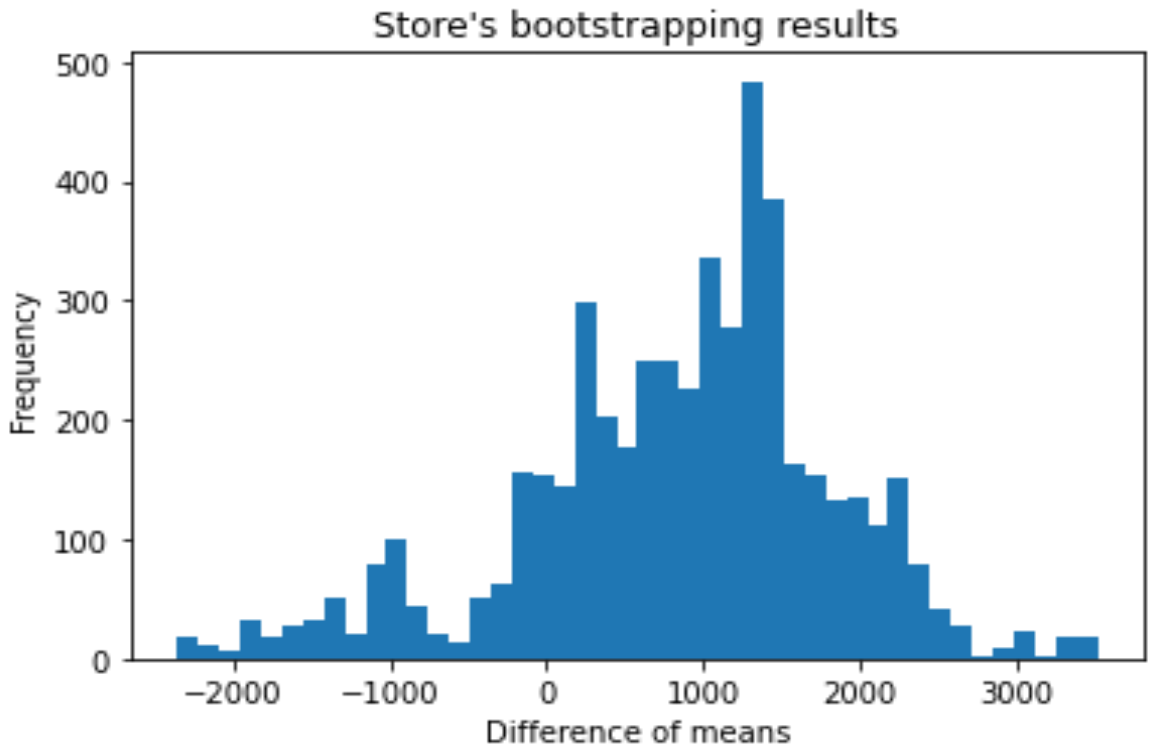
For Store

“The average of weekly sales for store 23 would tend to be higher than the average of the other 9 stores”.

Results



p-value = 0.0



p-value = 0.1718

Modeling

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- Gradient Boosting Regressor
- Decision Tree Regressor

Modeling Process

1. Reduce Dimensionality
2. Define independent variables and dependent variable (weekly sales).
3. Divide into train and test data
4. For some models we did grid search and for some we did not.
5. Define the model, fit based on the train data and make the predictions based on the test data.

Results

Model	Train Score	Test Score
Standard Linear Regression	0.7556	0.7660
Lasso Regression	0.7556	0.7660
Ridge Regression	0.7556	0.7660
Decision Tree Regressor	0.8481	0.8528
Random Forest Regressor	0.8491	0.8538
Gradient Boosting Regressor	0.9284	0.9283

Results without reducing dimensionality

Model	Train Score	Test Score
Random Forest Regressor	0.8560	0.8602
Gradient Boosting Regressor	0.9429	0.9340

Features Importance

Variable	Feature Importance
Size	0.22
Dept 54	0.1597
Dept 52	0.1377
Dept 60	0.1168
Dept 28	0.1036
Dept 59	0.0876
Dept 38	0.0204
Dept 95	0.0204
Dept 79	0.0161
Dept 23	0.01582

Variable	Feature Importance
Size	0.1799
Dept 54	0.1548
Dept 52	0.1328
Dept 60	0.1132
Dept 28	0.10
Dept 59	0.084
Dept 38	0.024
Dept 95	0.021
Dept 23	0.018
Dept 79	0.017

Final Insights

- I was able to create a good model to predict the weekly sales a Walmart store could have by using all the variables.
- I was able to know the most important features in our Gradient Boosting Regressor model. With this information, Walmart can know the features they need to focus on to improve the weekly sales.