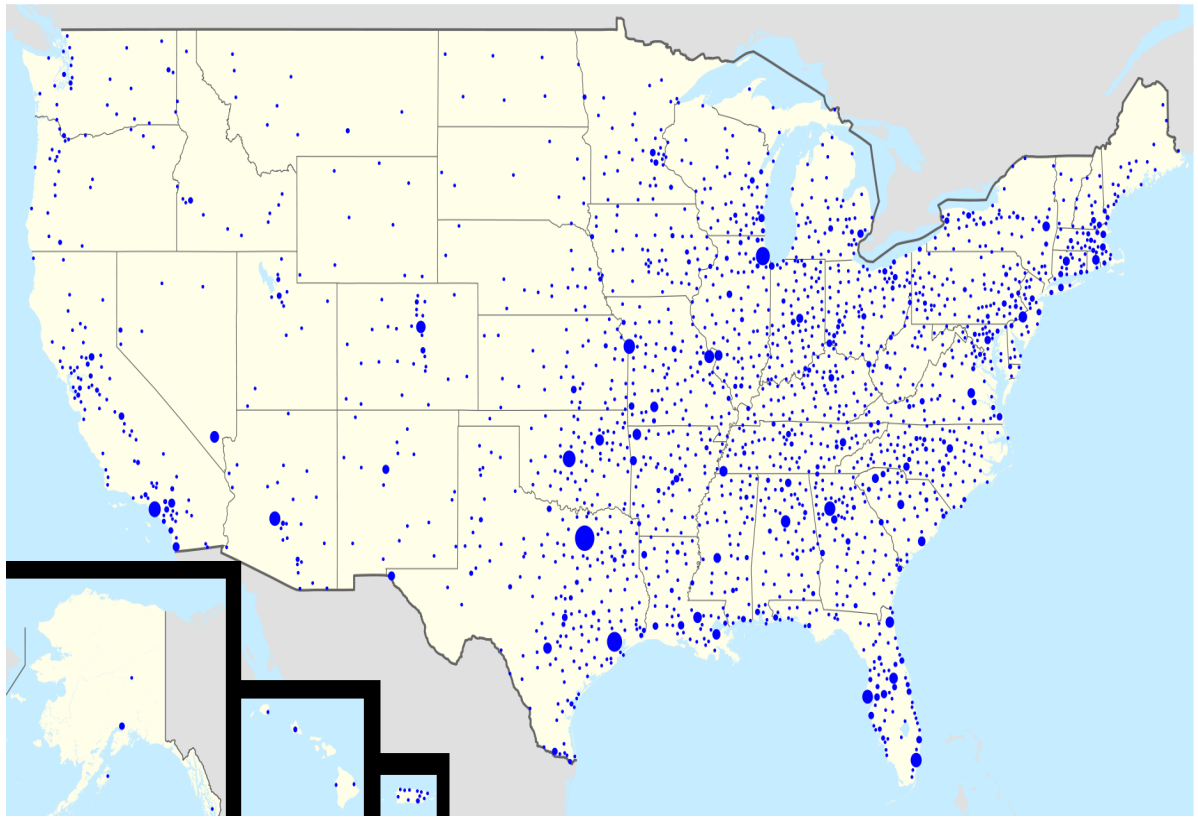
PREDICTING SALES MODEL

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?

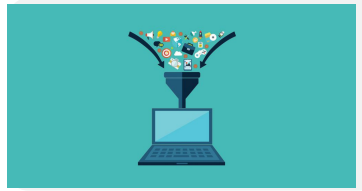


- Walmart is a multinational retail corporation characterized for the competitive prices on their products.
- Walmart is a corporation which has operations in all the states of the United States.
- Walmart is the world's largest company by revenue according to the Fortune Global 500 list.

Who might be interested?

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

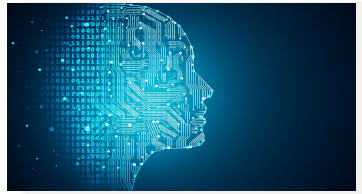
Steps



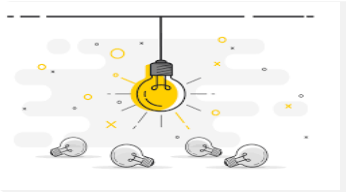
Data Collection

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.

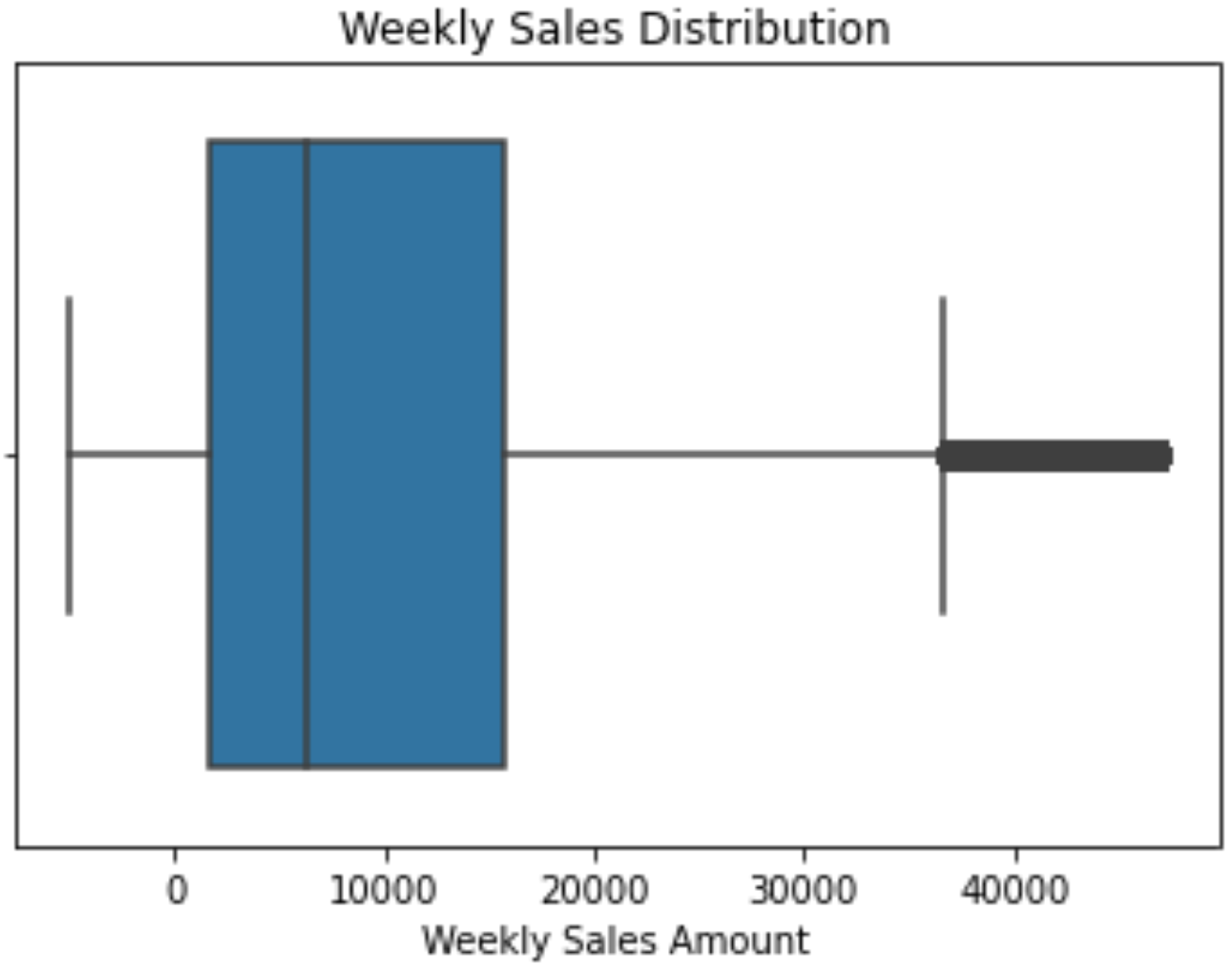
Data Wrangling

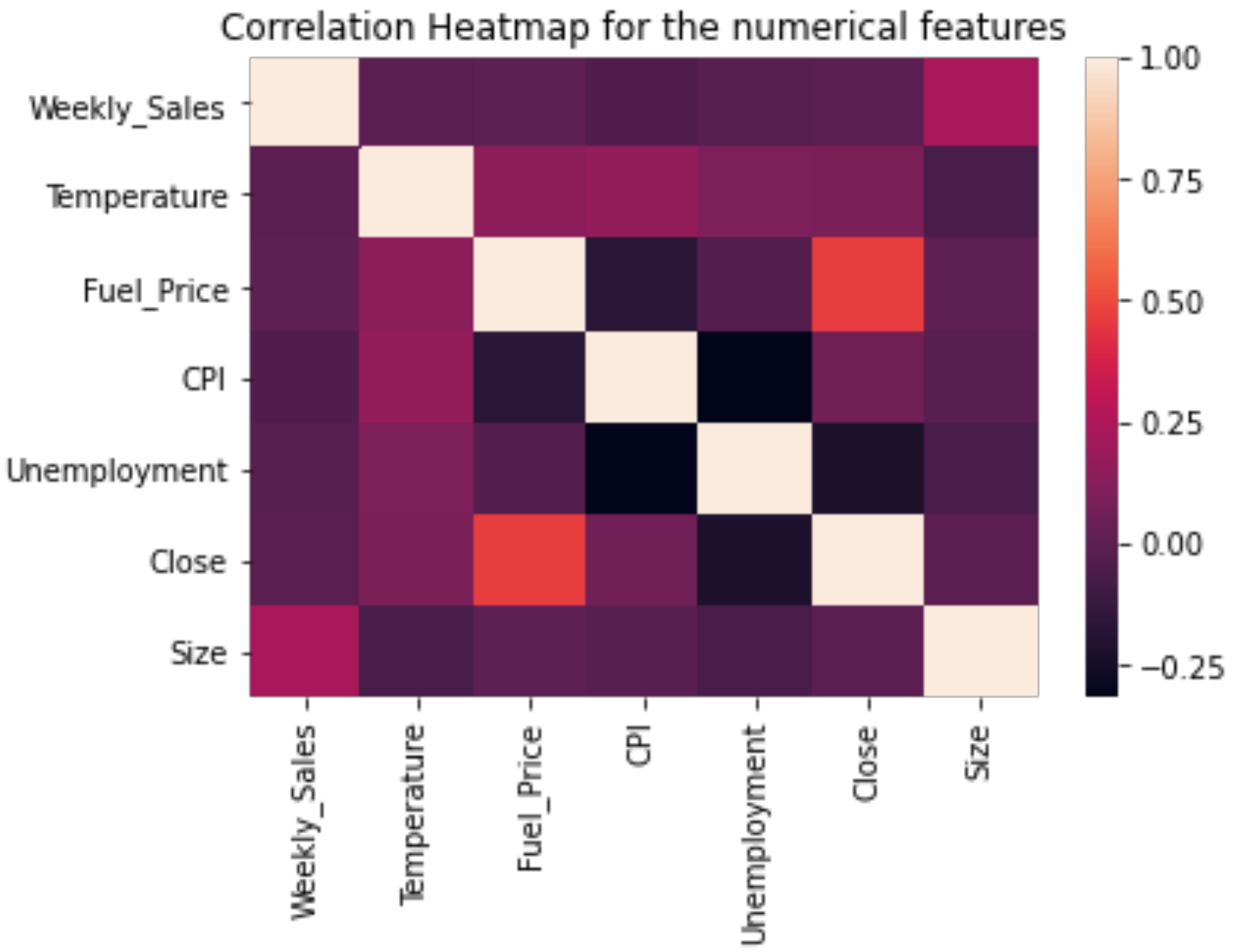
1. Pandas function: Inner Merge.

2. Handling null-values: pandas fillna function.

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

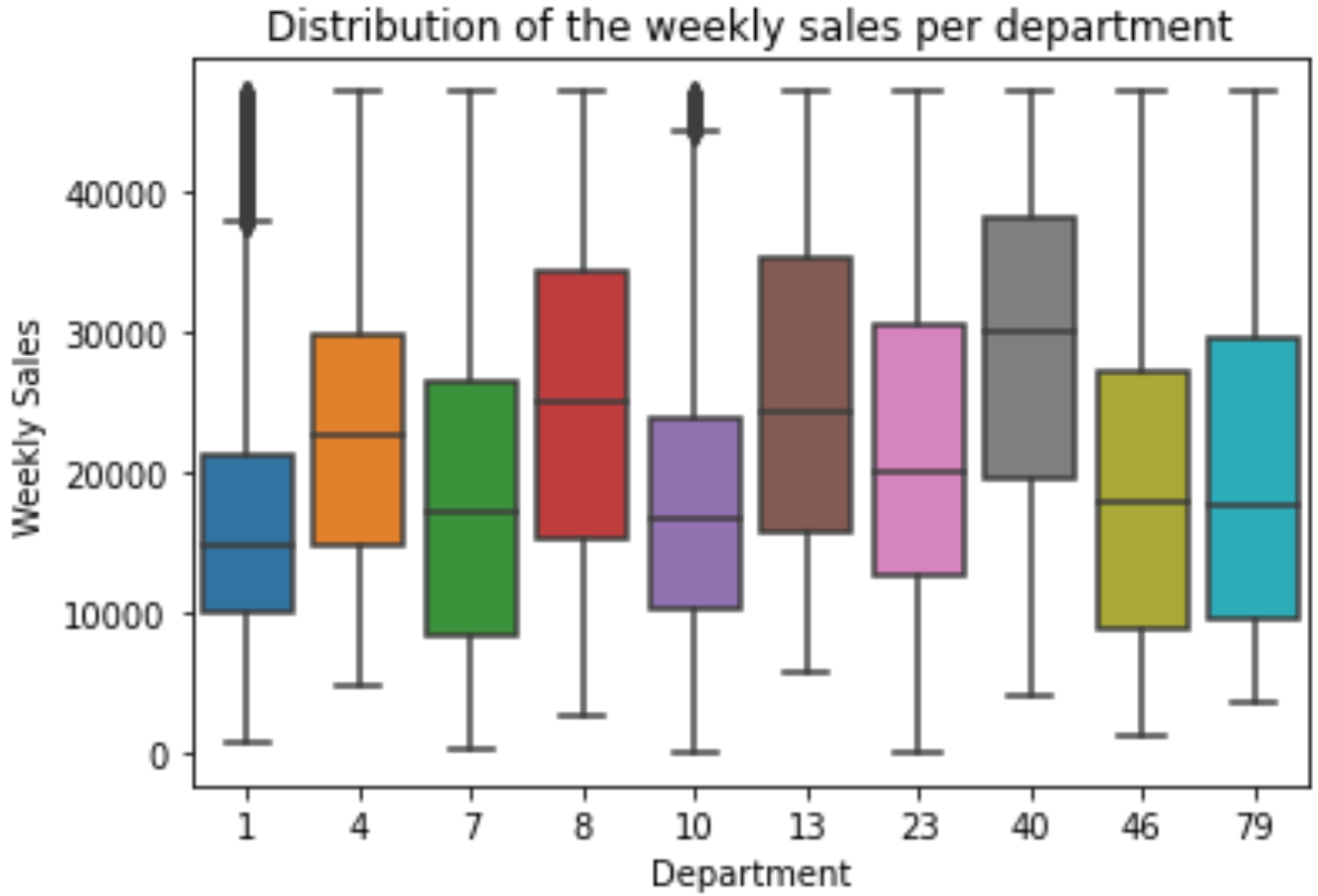
Data Visualization



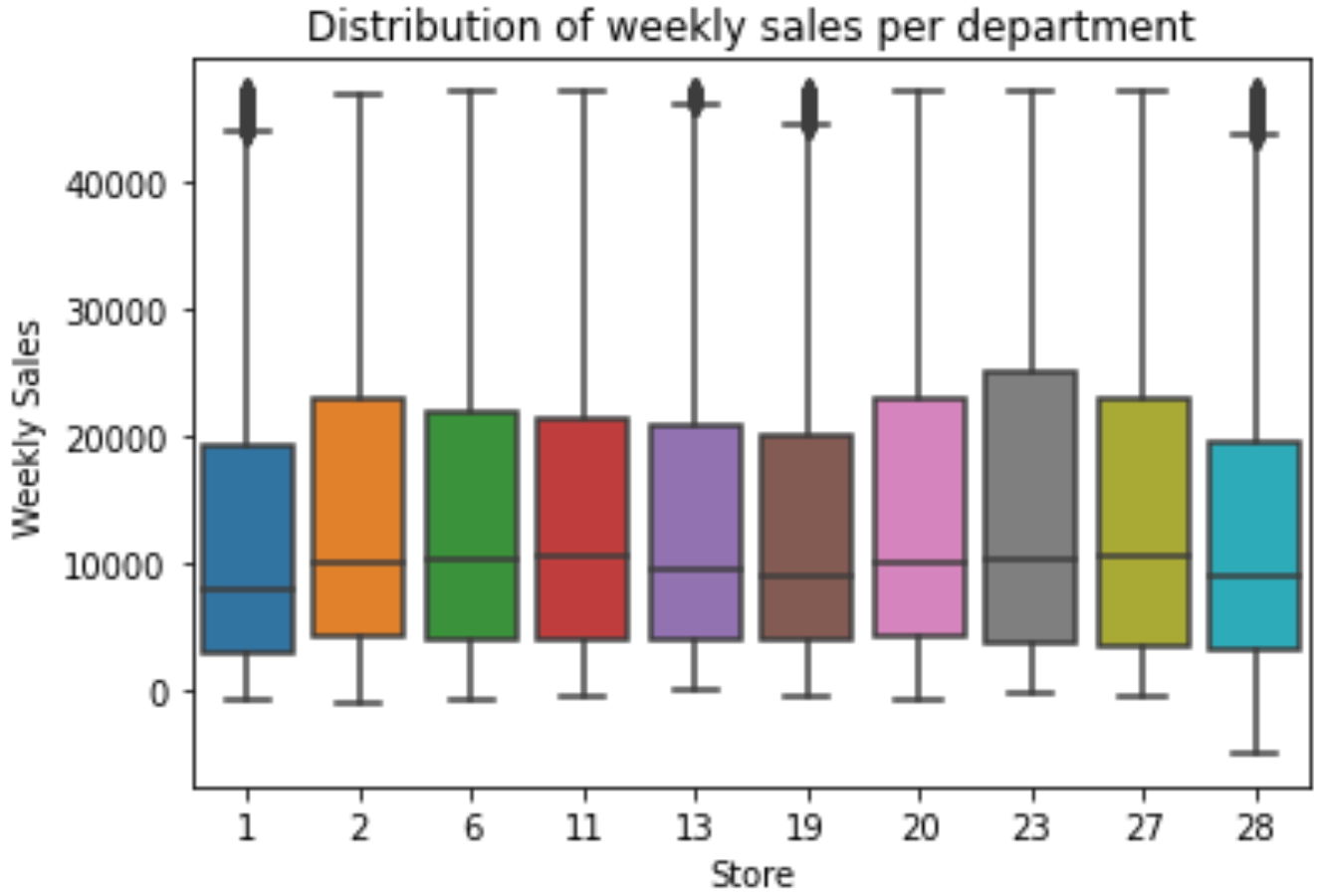


Correlation Heatmap

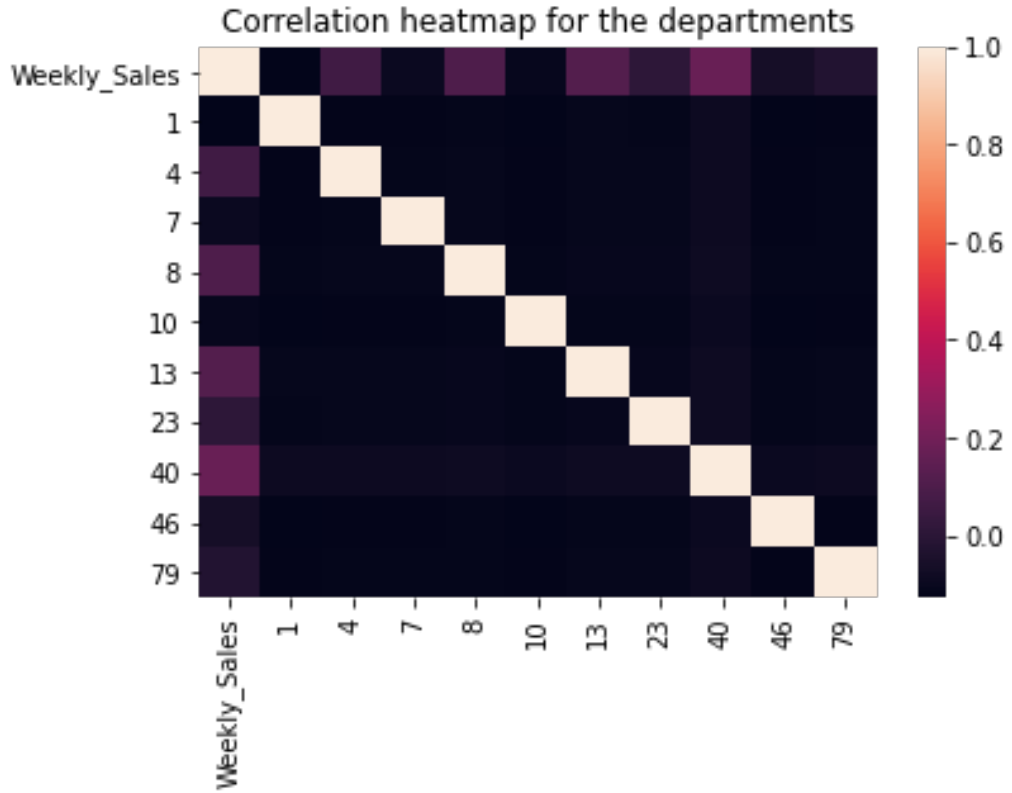
Distribution by department

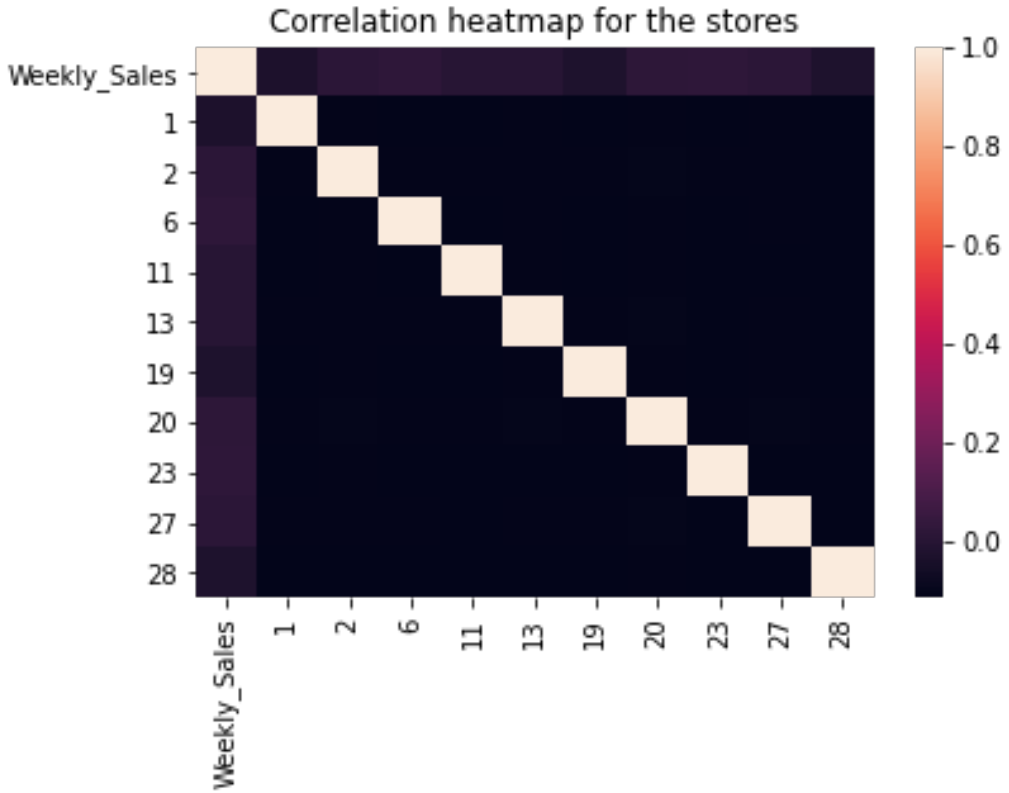


Distribution by store



Departments and Stores Correlation Heatmap





Hypothesis Testing

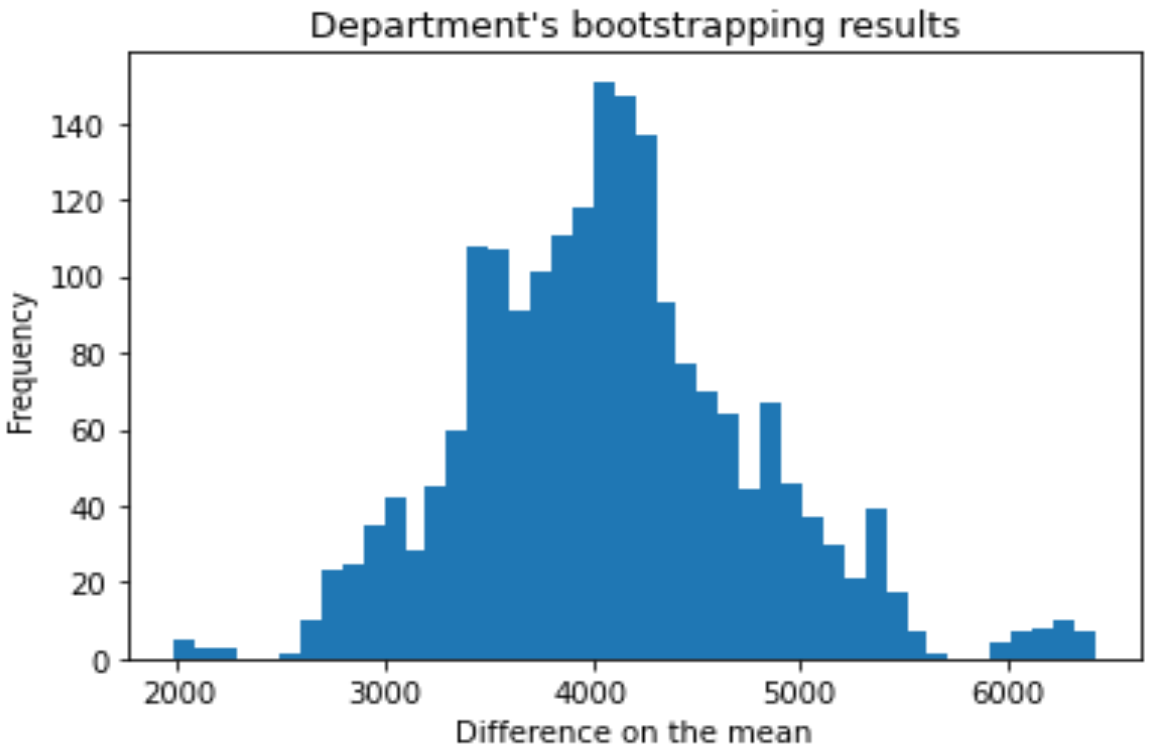
- For department

Hypothesis: Department 8 would tend to higher mean on sales than the other nine departments.

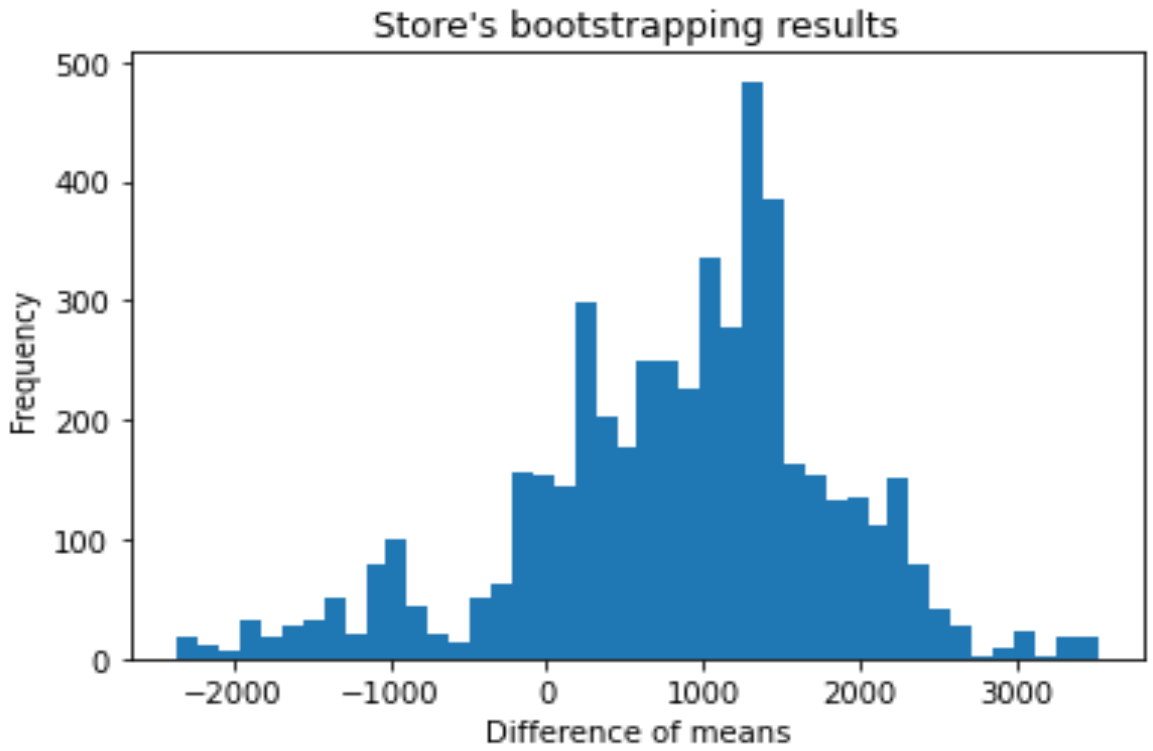
- For Store

Hypothesis : Store 23 would have in overall higher mean than the mean 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	0.7842	0.6237
Lasso Regression	0.7842	0.6238
Ridge Regression	0.7942	0.6237
Decision Tree Regressor	0.8529	0.7539
Random Forest Regressor	0.8564	0.76
Gradient Boosting Regressor	0.93507	0.83718

Results without reducing dimensionality

Model	Train Score	Test Score
Random Forest Regressor	0.8597	0.7617
Gradient Boosting Regressor	0.9463	0.835

Features coefficients

Variable	Coefficient
Dept 38	18,354.74
Dept 95	15,176.39
Dept 40	14,367.16
Dept 72	11,955.30
Dept 2	10,618.36
Dept 13	7,957.89
Dept 92	7,696.27
Dept 8	6,676.25
Dept 4	4,543.32
Type B	2,316.18

Intercept = 5077.49

Variable	Coefficient
Dept 38	17,836.78
Dept 95	16,211.45
Dept 40	15,323.88
Dept 72	12,693.05
Dep 2	11,439.24
Store 12	8,783.32
Dept 13	7,867.23
Store 23	7,723.57
Store 10	6,984.88
Store 35	6,611.97

Intercept = -8,230.66

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 more important variables by checking the coefficients. With the coefficients and intercept we could calculate the weekly sales for the store.