Capstone 2 Milestone Report

Problem:

Walmart is an American retail operation that operates a chain of grocery stores by the same name. Since its opening in 1962, it has expanded its ventures to many other countries, including Mexico, China, and Germany. As of 2019, Walmart is the world's largest company by both revenue and number of employees, at $514.405 billion per year and 2.2 million people, respectively.

Like all companies, Walmart's revenue is impacted by a variety of factors, some of which are included in the Walmart Recruiting dataset on Kaggle https://www.kaggle.com/bletchley/course-material-walmart-challenge#test.csv). The dataset contains information about the weekly revenue of 45 Walmart stores over 2.5 years, their promotion events, and other information about the area in which the stores operate, such as the average temperature and gas price. Using this dataset, I would like to fit a regression model to predict the weekly sales made given the factors.

Data Description and Wrangling:

The data comes in two .csv files, one for testing and one for training. The training dataset contains 282,000 rows and 16 columns, one of which is the weekly revenue column that's the object of our prediction. The other 15 features are:

- Store ID (number)
- Department ID (number)
- Date
- Whether the week contains a holiday (boolean)
- The temperature of the area in Fahrenheit (number)
- The gas price of the area in dollars (number)
- 5 markdown columns with information about promotional offers (number).
- The Consumer Price Index (number)
- The unemployment rate (number)
- The Type of the store (strings)
- The size of the store (number)

The Testing dataset has identical features with the training dataset but does not have the weekly revenue column, which we would have to predict ourselves.

To clean the data, I looked at the summary of the dataframe and found that 1) there were several null values in the MarkDown columns, 2) the date column is not in DateTime format, but in object, and 3) the type column is shown to be type object,

when type category would both consume less memory and be more convenient for OneHotEncoding later.

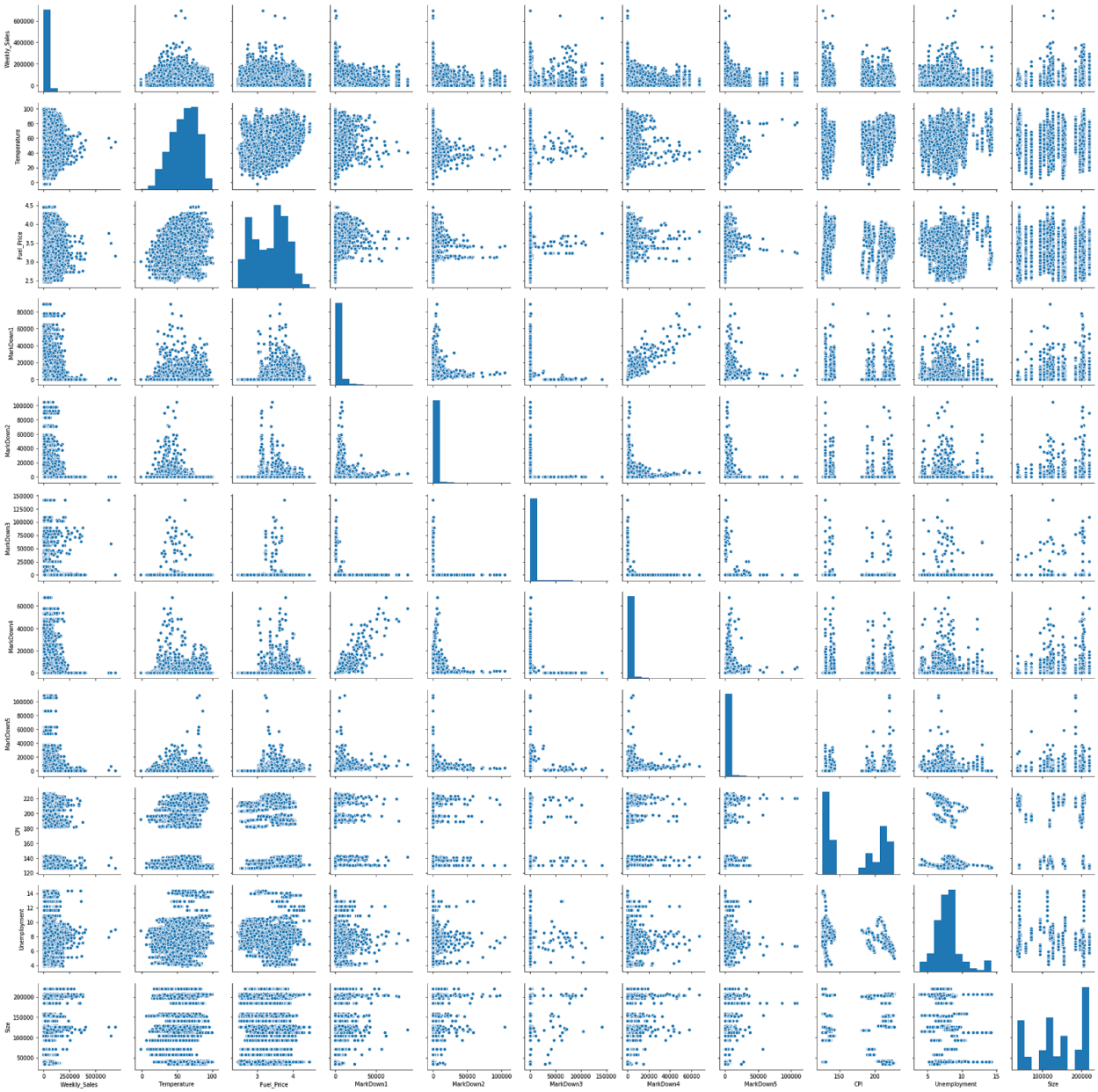
To make this dataset easier for analysis, I first filled in all the null values in the MarkDown columns with 0. The description of the dataset said that the MarkDown columns “contain information about promotional events”, so it makes sense to conclude that null values in MarkDown columns mean that there were no promotional events on that week in that particular store and department. Then, I changed the Date column to datetime format and the type column to category format for easier analysis.

Exploratory Data Analysis and Inferential Statistics -- Numerical:

The numerical variables are: Temperature, Fuel_Price, MarkDowns 1-5, CPI, Unemployment, and Size. I put all of these columns into a Seaborn Pairplot to look at their distributions, as well as visualize any relationship with other variables, especially with our dependent variable (Weekly Revenue) in Figure 1. It appears that both the weekly revenue and the markdown columns have an exponential distribution, with most of the values on the left side of the graph. Thus, I created a new column from the logarithm of the weekly sales for easier visualization. CPI and fuel prices are bimodal, which is related to the year in which the data was taken, and the size column is trimodal which corresponds with the three categories in the type column. The remaining columns, unemployment and temperature, are both approximately normal.

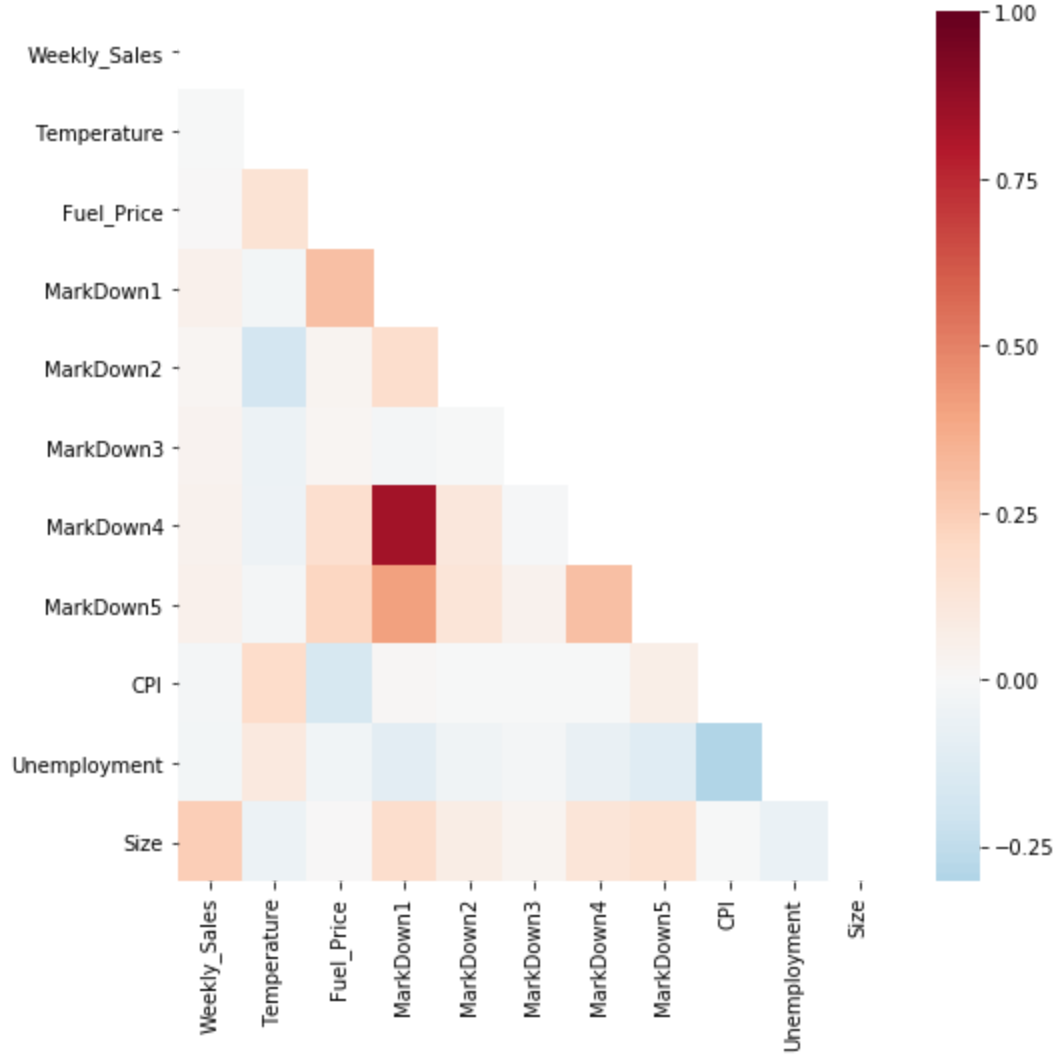
Figure 1. Pairplot of the numerical variables in the dataset.

(Note: The columns might be a bit unreadable on this report, but the columns are, from top to bottom and from left to right, Weekly_sales, Temperature, Fuel_Price, MarkDowns 1-5, CPI, Unemployment, and Size. Every intersection between two of the same columns show a histogram of the variable’s distribution, and every intersection between two different columns show a scatter plot with the left column on the y-axis and the right column on the x-axis.)



Judging from the Weekly Sales' interactions with the other numerical variables in the scatter plots, fuel price, the markdown columns, Unemployment, and CPI has a negative effect on the weekly revenue while the others have a positive effect. There are also some multicollinearity among the dependent variables (Most notably between the Markdown columns), which I looked at in more detail in a heatmap (Figure 2)

Figure 2. Heatmap of the numerical variables.



This heatmap confirms our previous observations: the markdown columns are intercorrelated, and the weekly sales is most heavily correlated with the size. To reduce the multicollinearity, I may condense all of the markdown columns into one and deleted some other columns. I may also not do anything about it, if the intercorrelation doesn't translate to worse results.

Next, I looked at the significance of correlations between the independent variables and the weekly revenue by fitting a model using the statsmodel ordinary least square linear regression model.

Figure 3. OLS report of numerical variables.

OLS Regression Results

Dep. Variable:	Weekly_Sales	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	226.1
Date:	Mon, 16 Sep 2019	Prob (F-statistic):	0.00
Time:	18:28:53	Log-Likelihood:	-3.2323e+06
No. Observations:	282451	AIC:	6.465e+06
Df Residuals:	282441	BIC:	6.465e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.606e+04	494.488	52.711	0.000	2.51e+04	2.7e+04
Temperature	23.9436	2.471	9.691	0.000	19.101	28.786
Fuel_Price	-1483.2012	102.857	-14.420	0.000	-1684.798	-1281.604
MarkDown1	0.1904	0.014	13.386	0.000	0.163	0.218
MarkDown2	0.0513	0.009	5.919	0.000	0.034	0.068
MarkDown3	0.1500	0.008	19.263	0.000	0.135	0.165
MarkDown4	-0.0727	0.021	-3.545	0.000	-0.113	-0.033
MarkDown5	0.1940	0.011	17.401	0.000	0.172	0.216
CPI	-25.0676	1.203	-20.834	0.000	-27.426	-22.709
Unemployment	-388.1754	24.458	-15.871	0.000	-436.113	-340.238

Omnibus:	205761.418	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5733316.107
Skew:	3.215	Prob(JB):	0.00
Kurtosis:	24.114	Cond. No.	9.45e+04

All of the correlations among the numerical variables are significant. It's mostly as we expected: Fuel price, CPI, unemployment all have a negative correlation with the weekly sales while temperature has a positive correlation. The actual impact might have been impacted by

To see if the categorical variables had a significant impact on the revenue, I performed a frequentist hypothesis test to see if the weekly sales are different for variables with two categories (IsHoliday and IsMarkDown). For the Type column, which contains 3 options, I did an ANOVA (Analysis of Variance) test to see if the type of store is relevant for the revenue prediction.

There are three requirements that must be met for frequentist hypothesis testing:

- First, the samples must be independent.

This assumption may be questionable, because the dataset has a time-series element to it and each week's revenue may have an impact on the revenue next week. For our purposes, though, we can assume that that each row is an independent observation based only on the features. If I have time, I can utilize a more Times Series based approach.

- Second, the samples must be random.

This is an assumption, but since the description of the dataset doesn't mention any biases in selection we can assume that there is none.

- Third, the distribution must be normal.

By "distribution" here, we are referring to the distribution of the mean of the data according to the central limit theorem, which means that the sample data should either be 1) normal, or 2) greater than 30. As both the test and training data have more than 30 entries, this condition is satisfied.

Holiday

There are 19819 rows with IsHoliday==True included out of the total 282450 rows. Next I looked at the unique values included to see if these dates are holidays, and if there are any holidays missing. Below are the holidays Walmart gave:

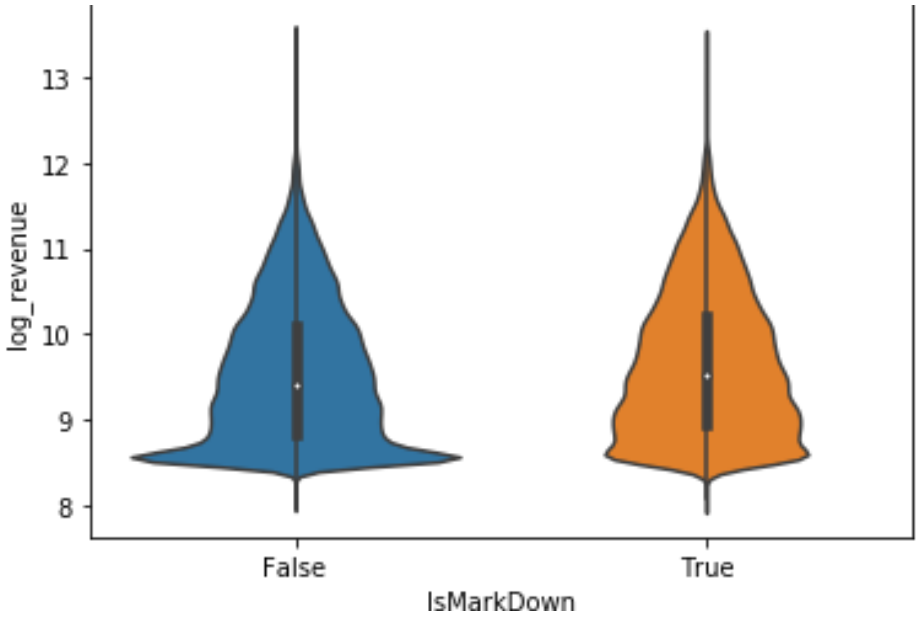
Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Given that the range of dates included go from 2010-02-05 to 2012-10-26, we should have all of the holidays except Thanksgiving and Christmas of the years 2012 and 2013. A look at the unique values say we do.

Next, I compared the distribution of the revenue for holiday and non-holiday weeks in Figure 3. As can be seen, holiday weeks generated more revenue compared to non-holiday weeks.

As the distribution is almost exponential, I took the log of the weekly revenue for clearer comparison

Figure 4. Distribution of log revenues for Holiday and Non-Holiday weeks

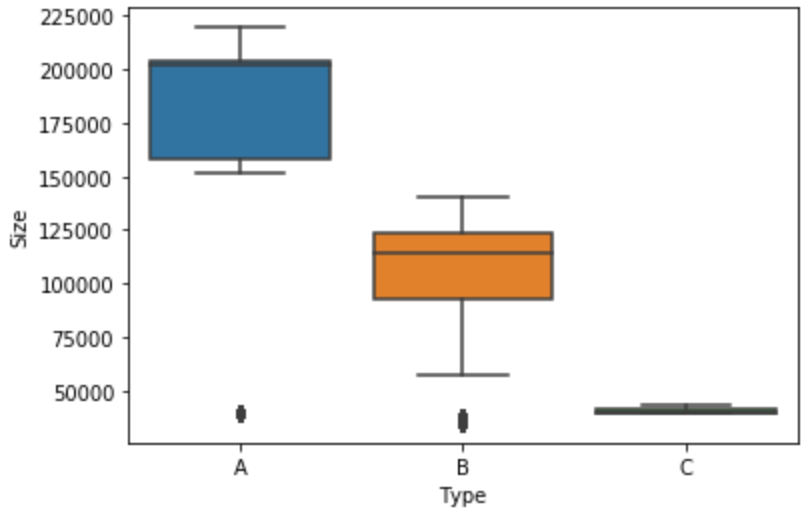


Descriptive statistics and a hypothesis test confirms what we see: the average revenue for holiday weeks is 16986.67 dollars with a standard deviation of 26955.16 dollars, while non-holiday weeks have a mean revenue of 15907.72 dollars with a standard deviation of 22301.76 dollars, and holiday weeks see significantly higher weekly revenue compared to non-holiday weeks.

Type

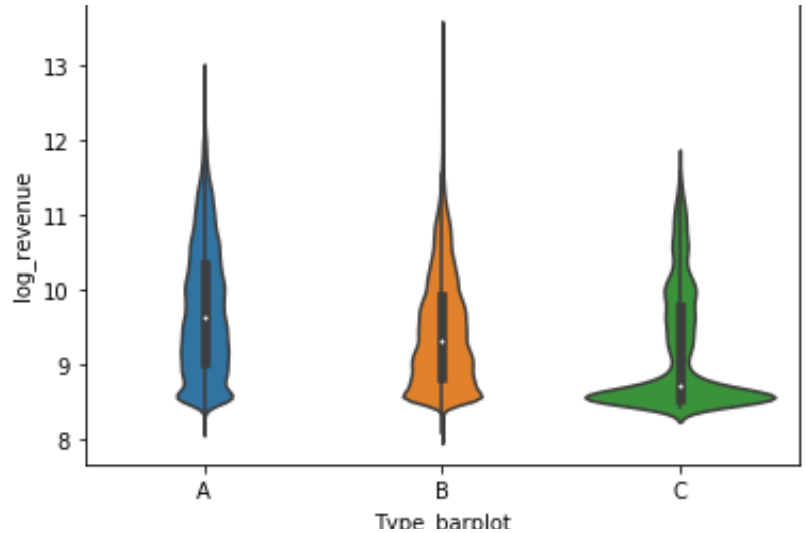
Browsing the data, I noticed that type A stores tend to be larger than type B stores, which are in turn larger than type C stores. A boxplot (Figure 7) confirms my observation

Figure 5. A box-and-whiskers plot of the sizes for each type of store.



Next we look at the weekly revenue. Surprisingly, the highest revenue was generated by a type B store (Figure 8a). After removing outliers, however, the distribution is as we expected with the largest type of store generating the most revenue (Figure 8b).

Figure 6. Types of Stores vs. Log of Weekly Sales

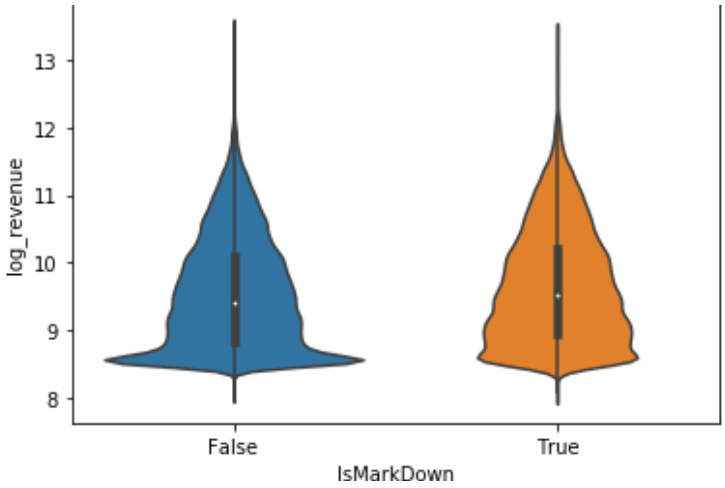


The ANOVA analysis yielded a F-value of 5290, which is larger than 3, or the threshold F-value for alpha = 0.05. This would say that the type of store significant affects the weekly revenue.

IsMarkDown

There were 64818 rows with markdowns and 217633 rows without markdowns. First, we looked at violin plots of weekly revenues made during weeks with markdowns vs. during weeks without markdowns (Figure 9a)

Figure 7. IsMarkDown vs. Log of Weekly_Sales



Surprisingly, the higher number of sales were made on a week without markdowns. Let's eliminate outliers and see if this is still true. However, the average revenue for markdown weeks is 19867.04 dollars with a standard deviation of 24729.34 dollars, while non-holiday weeks have a mean revenue of 15422.34 dollars with a standard deviation of 21976.45 dollars. A hypothesis test yielded a p-value of 1.95 * 10^-8, so it appears that revenues made during markdown weeks are significantly higher than non-markdown weeks.