Capstone Project 2 Proposal

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, such as weather and unemployment rates. Some of these factors are included in the Walmart Recruiting dataset on Kaggle (https://www.kaggle.com/bletchley/course-material-walmart-challenge#test.csv), which contains information about the weekly revenue of 45 Walmart stores over 2.5 years, promotion events, and other information about the area in which the stores operate, such as the average temperature and gas price.

Target Audience:

Understanding factors that contribute to revenue is key to the success of some stores and failure of others. The predictions and insights gained in this capstone project can be used to drive business decisions, and I target this proposal to an executive who could then use this information to gain a better understanding of what's affecting the company's revenue as well as projection numbers for the future.

Data Description:

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. Figure 1 shows a description of the numerical factors in the Training and Testing datasets.

Figure 1. Distribution of numerical values in training (left) and testing (right) datasets:

	Store	Dept	Weekly_Sales	Temperature
count	282451.000000	282451.000000	282451.000000	282451.000000
mean	22.193166	44.286138	15983.429692	60.113640
std	12.782138	30.503641	22661.092494	18.446485
min	1.000000	1.000000	-4988.940000	-2.060000
25%	11.000000	18.000000	2079.330000	46.780000
50%	22.000000	38.000000	7616.550000	62.150000
75%	33.000000	74.000000	20245.745000	74.290000
max	45.000000	99.000000	693099.360000	100.140000

	Fuel_Price	MarkDown1	MarkDown2	MarkDown3
count	282451.000000	100520.000000	74232.000000	91521.000000
mean	3.360300	7246.077559	3318.408122	1417.397841
std	0.458602	8254.606267	9485.575898	9547.858949
min	2.472000	0.270000	-265.760000	-29.100000
25%	2.932000	2241.190000	40.960000	5.060000
50%	3.452000	5363.520000	191.820000	24.340000
75%	3.737000	9235.590000	1919.790000	103.130000
max	4.468000	88646.760000	104519.540000	141630.610000

	MarkDown4	MarkDown5	CPI	Unemployment
count	90031.000000	101029.000000	282451.000000	282451.000000
mean	3379.591745	4639.476021	171.207802	7.968098
std	6269.428446	6060.459590	39.160808	1.868070
min	0.220000	135.160000	126.064000	3.879000
25%	508.100000	1877.810000	132.022667	6.891000
50%	1482.030000	3364.410000	182.350989	7.866000
75%	3607.570000	5563.800000	212.464799	8.572000
max	67474.850000	108519.280000	227.232807	14.313000

	Size
count	282451.000000
mean	136730.073220
std	61002.319363
min	34875.000000
25%	93638.000000
50%	140167.000000
75%	202505.000000
max	219622.000000

	Store	Dept	Temperature	Fuel_Price
count	139119.000000	139119.000000	139119.000000	139119.000000
mean	22.215528	44.207894	60.042182	3.362502
std	12.791743	30.468558	18.450840	0.458334
min	1.000000	1.000000	-2.060000	2.472000
25%	11.000000	18.000000	46.630000	2.935000
50%	22.000000	37.000000	62.010000	3.459000
75%	33.000000	72.000000	74.250000	3.738000
max	45.000000	99.000000	100.140000	4.468000

	MarkDown1	MarkDown2	MarkDown3	MarkDown4
count	50161.000000	37016.000000	45570.000000	44936.000000
mean	7247.106823	3367.157260	1483.652604	3390.333933
std	8364.196211	9454.875819	9772.352647	6338.190217
min	0.270000	-265.760000	-29.100000	0.220000
25%	2229.520000	42.160000	5.200000	500.910000
50%	5307.810000	193.900000	24.940000	1479.910000
75%	9160.250000	1958.520000	105.090000	3578.400000
max	88646.760000	104519.540000	141630.610000	67474.850000

	MarkDown5	CPI	Unemployment	Size
count	50403.000000	139119.000000	139119.000000	139119.000000
mean	4607.926737	171.190058	7.944434	136723.535441
std	5762.340423	39.156302	1.853472	60936.648098
min	135.160000	126.064000	3.879000	34875.000000
25%	1880.310000	132.022667	6.891000	93638.000000
50%	3332.990000	182.318780	7.866000	140167.000000
75%	5556.150000	212.403576	8.567000	202505.000000
max	108519.280000	227.232807	14.313000	219622.000000

Method:

After cleaning both datasets, I'm going to use all of the training dataset to create a regression model and leave the testing dataset alone until the model is finished, after which I'll use it to see how well the model performs on unseen data.

As seen in figure 1, the data contains some outlier values where the maximum is up to 10x the 75% quantile, and it's likely that our regression model will mispredict these outliers as they deviate significantly from the norm, so our evaluation metric needs to take that into account and not penalize deviations too much. One such metric is the Mean Absolute Error (MAE) which takes the absolute value of the error between the actual and predicted value, instead of squaring it, and is therefore more tolerant of errors. It's also easier to interpret.

I intend to fit several regression models and compare their outcomes. The data was originally used in a Kaggle competition, and the highest scorer had a weighted mean absolute error of 2301.

Deliverables:

I will have powerpoint slides, a report in PDF, and a jupyter notebook along with interim reports and projects on GitHub.