Walmart dataset analysis report

Xiaoshu Luo

1.Introduction:

This dataset contains weekly sales from 99 departments belonging to 45 different stores.
Our aim is to forecast weekly sales from a particular department.

The objective of this case study is to forecast weekly retail store sales based on historical data.

The data contains holidays and promotional markdowns offered by various stores and several departments throughout the year.

Markdowns are crucial to promote sales especially before key events such as Super Bowl, Christmas and Thanksgiving.

Developing an accurate model will enable us to make informed decisions and make recommendations to improve business processes in the future.

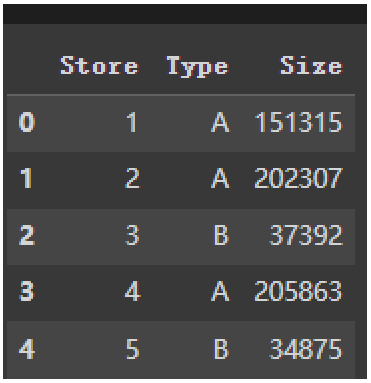
The data consists of three sheets:Stores, Features and Sales.

Data Source is from Kaggle:: https://www.kaggle.com/manjeetsingh/retaildataset

2. Data exploration and preprocessing:

The basic information about these three datasets include:

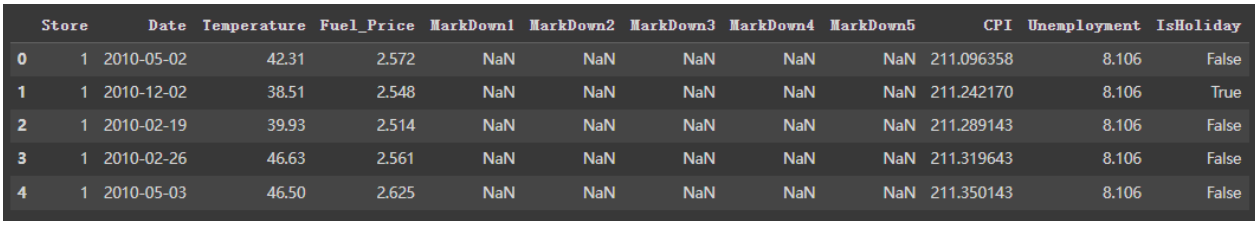
Stores:



	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

Anonymized information about the 45 stores, indicating the type and size of store

Features:

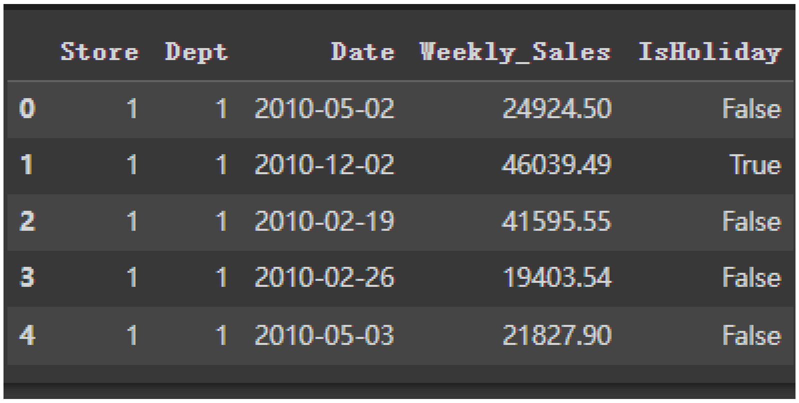


	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-05-02	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-12-02	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-05-03	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

Contains additional data related to the store, department, and regional activity for the given dates.

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

Sales:



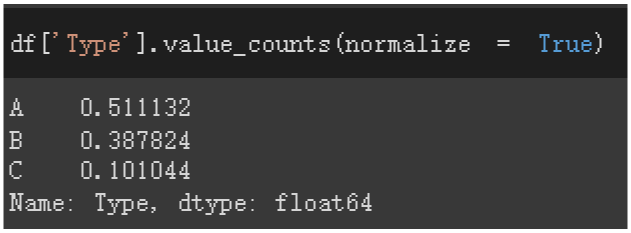
	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-05-02	24924.50	False
1	1	1	2010-12-02	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-05-03	21827.90	False

Historical sales data, which converts to 2010-02-05 to 2012-11-01.

- Store - the store number
- Dept - the department number
- Date - the week
- Weekly_Sales - sales for the given department in the given store
- IsHoliday - whether the week is a special holiday week

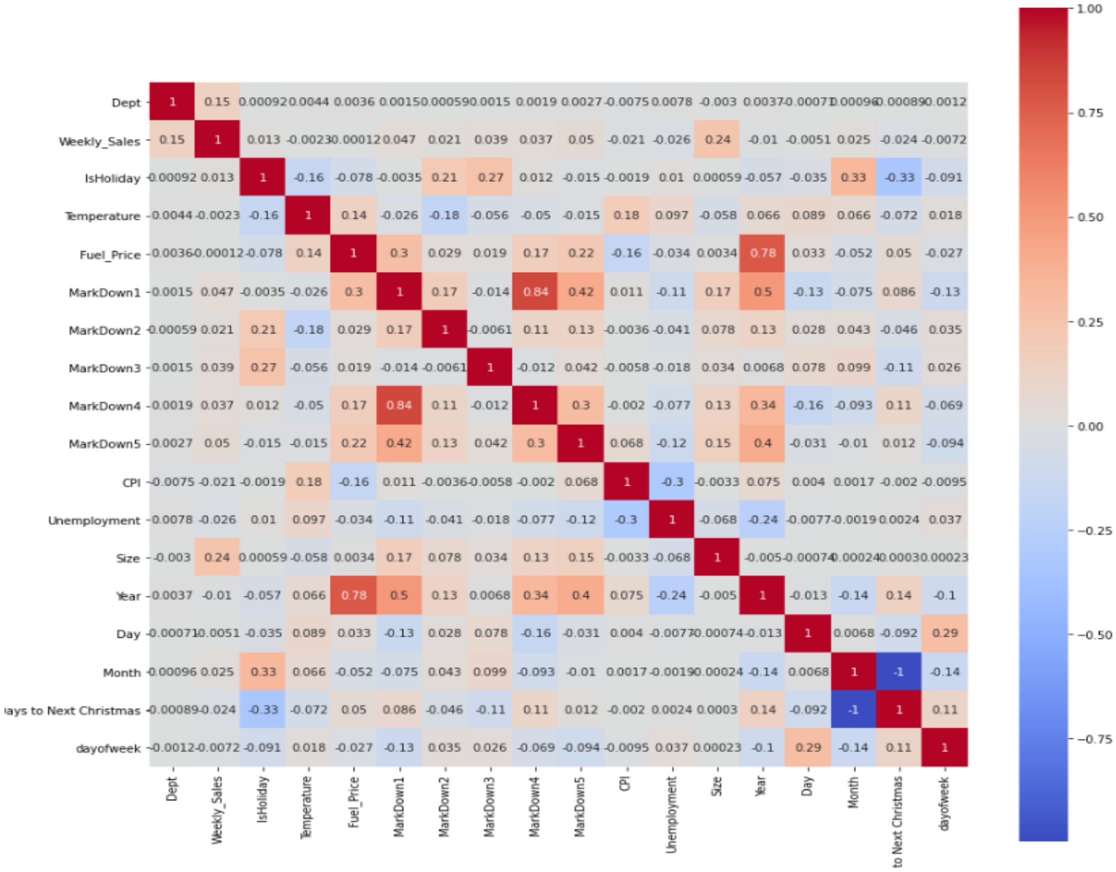
Then the date information was converted to pandas format for easier access and null values, all of which are in the markdown columns which represent no promotion, have been filled with zeroes.

Distribution of stores by type:

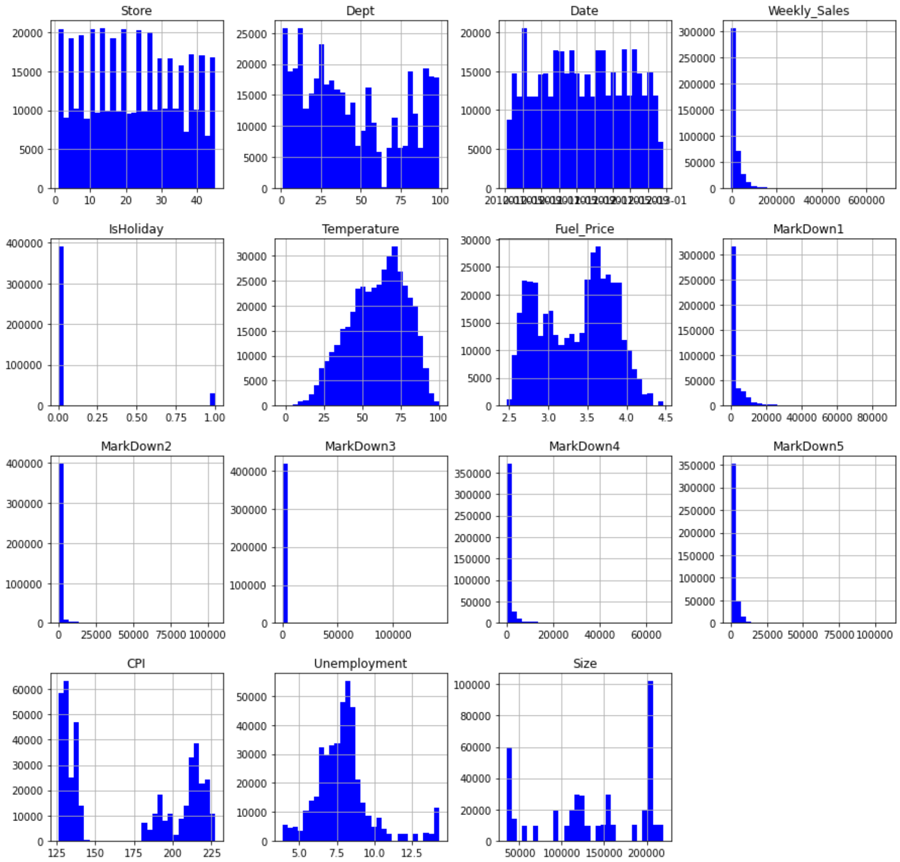


```
df['Type'].value_counts(normalize = True)
A    0.511132
B    0.387824
C    0.101044
Name: Type, dtype: float64
```

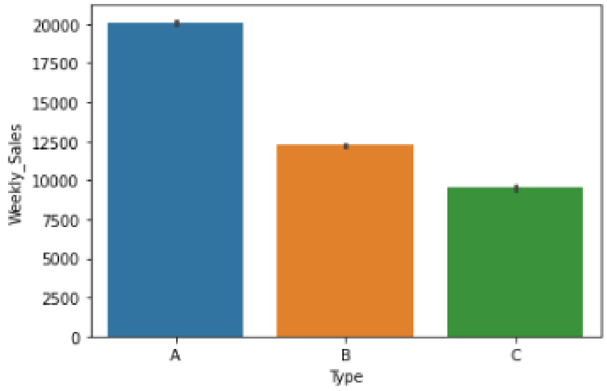
Correlation between variables:



Distribution of variables:



Weekly sales by Store Type:



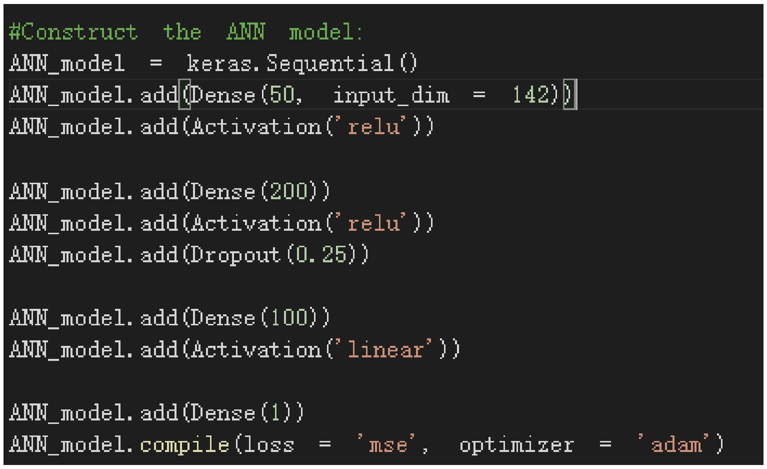
3. Building up machine learning models:

The weekly sales is used as the response variable and all the other variables are predictors. All the numerical variables are scaled, and the dataset is divided into a training set and a testing set by a ratio of 8:2.

All the three categorical variables are converted into one-hot encoding format.

Linear Regression with degree =1 as baseline:
Training R-squared: 0.659
Testing R-squared: 0.656

3-layer Deep learning:
Model structure:



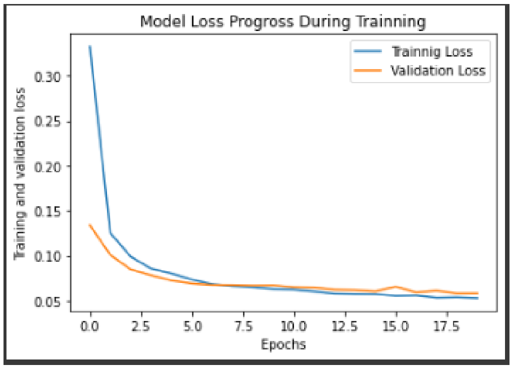
```
#Construct the ANN model:
ANN_model = keras.Sequential()
ANN_model.add(Dense(50, input_dim = 142))
ANN_model.add(Activation('relu'))

ANN_model.add(Dense(200))
ANN_model.add(Activation('relu'))
ANN_model.add(Dropout(0.25))

ANN_model.add(Dense(100))
ANN_model.add(Activation('linear'))

ANN_model.add(Dense(1))
ANN_model.compile(loss = 'mse', optimizer = 'adam')
```

Training R-squared: 0.960
Testing R-squared: 0.947
Training loss:



XG-Boost Tree (Grid Search):
Best parameters: {'gamma': 0.5, 'max_depth': 6}

Training R-squared: 0.854
Testing R-squared: 0.848

4. Conclusion:

By comparing the models, ANN has achieved the highest training and testing accuracy.