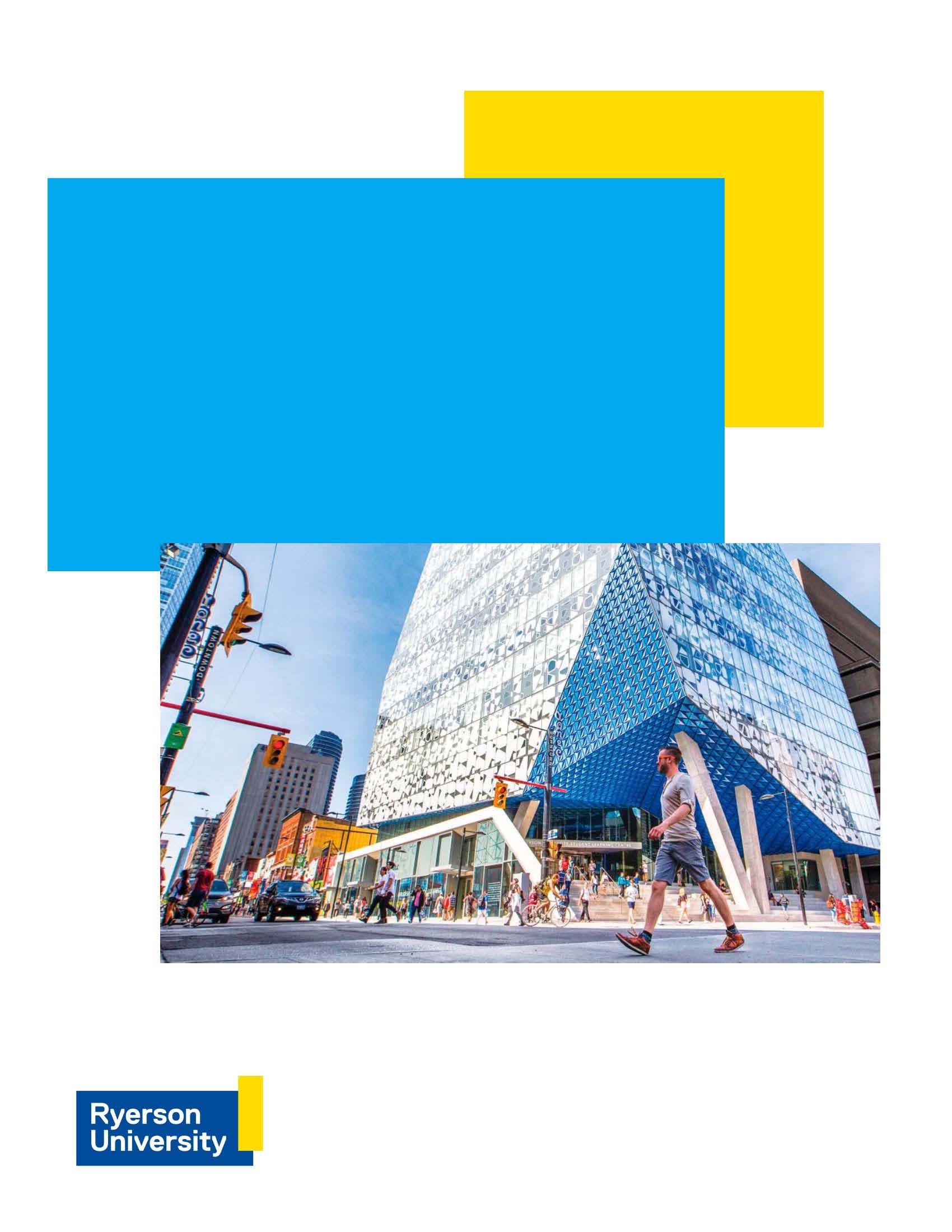
Data Mining

in Retail Industry

Name: Ammrit Preet Kaur Tuteja

Student Number: 501037031

Supervisor: Dr. Tamer

**Table**

[**Abstract**](#_heading=h.14a41ksbo3bk) **3**

[**Dataset**](#_heading=h.14a41ksbo3bk) **3**

[**Initial Analysis**](#_heading=h.7q48xwr35ls8) **4**

[**Exploratory Data Analysis**](#_heading=h.44yf9wrmkajp) **5**

[**References**](#_heading=h.gknyp65jzo7c) **7**

# Abstract:

# 

# The theme for the project is data mining.

# 

# Retail industry has to make adjustments on the prices of the items based on the various factors and how the change in those factors affected the sales in the past. We have historical data available for 45 stores in different regions; including the holiday and non-holiday season. Some of the factors which are uncontrollable; however they still affect the sales of the store such as environmental weather conditions, cost of operating (in the form of fuel prices). Other factors include markdowns and holiday season, which can hugely affect the sales.

# 

# Dataset:

# Here is the dataset we are using:

# 

# <https://www.kaggle.com/manjeetsingh/retaildataset>

# 

# It includes information about: i) Stores - 45 stores indicating the type and size of the store.

# ii) Features - Temperature, fuel prices, markdowns, employment status and consumer price index.

# iii) Weekly sales for a particular store and department and whether it is a holiday season week or not.

Problems to be tackled in this case are:

1. Whether costing factors like fuel price affect the markdowns offered by the stores?
2. Given the changes in the physical environmental conditions, does infusing markdowns have an effect on sales?

Tools used should be - Python (Jupyter Notebook), Scikit library, Numpy, Pandas and Matplotlib, Seaborn.

# Initial Analysis

**Features table:**

This table has market analysis for 45 stores belonging to the Retail industry. Each row represents the weekly data for one of the 45 stores. In this dataset:

The variables - store and date are qualitative and numeric discrete variables.

Temperature, fuel price, markdowns, CPI , unemployment are quantitative numeric continuous variable.

IsHoliday is an ordinal variable where the options for the answer are : True or False.

*Store* - Store number which will be 1 out of the 45 stores in the region. It is an independent variable.

*Date* - the date when the week starts for the given row of data

temperature - Average temperature for the given time period

*Fuel price* - cost of fuel in the region for the given time period

*Markdowns 1-5*: this is the value of the promotional markdowns for the week for the given store. There were no markdowns for the year 2010 and 2011; so the entire data has been marked as N/A in that case. For the years 2012 and later; not every store has markdowns running every week, hence some of them are marked as N/A too. We shall not fix any missing values in this case as it will hamper the accuracy of our model.

*CPI* - Consumer Price index is the weighted average change in prices over time that consumers pay for a group of goods or services.

*Unemployment* - rate of unemployment; some of the rows have missing data too. But as above, we shall not insert random values.

*IsHoliday* - does that week fall under the category of holiday season or not? it only has 2 possible values: True or False - boolean data. it will be considered as categorical data.

**Sales Dataset:**

This table has the sales information on a weekly basis between 2010 and 2013. Each of the stores has roughly about 98 departments. These departments are marked with numbers starting from 1.

All of these variables - Store, Department and date are qualitative in nature.

isHoliday is an ordinal variable

*Store*- It represents the store number - one of the 45 stores; qualitative data.

*Department* - it is basically just a number used to depict and differentiate between the departments. This is qualitative data.

*date* - it is the date of the start of the week, of which sales are represented.

weekly sales - this is the weekly sales statistics

*isHoliday* - Boolean variable representing if this week is a part of holiday season or not.

**Stores dataset:**

This table shows the store ID along with its type (one of the three values - A, B or C) and the size of the store. We are assuming it is the number of people that can be accommodated in that store.

Store ID and type are the qualitative numeric discrete variables; however the size of the store is the quantitative discrete numeric variable.

# Exploratory Data Analysis

Mainly the data that could give us answers lies in the file: Sales dataset. I have been analyzing what we have available as the initial data.

Here is the link to detailed analysis on this dataset:

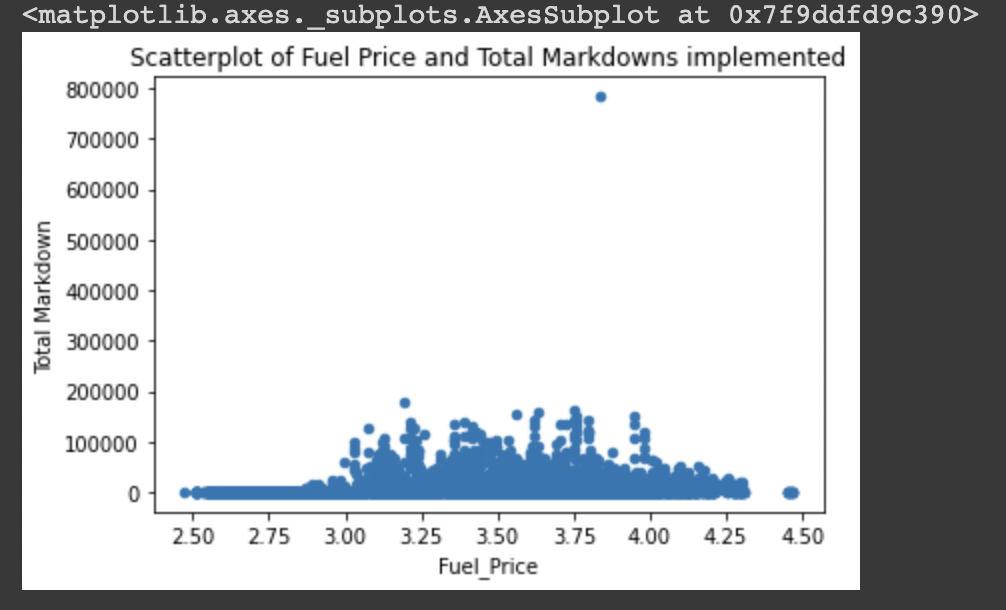
<https://github.com/aktuteja/DataMining_RetailIndustry>

Based on the analysis as performed above, the missing values were present in the columns - Markdown1, Markdown2, Markdown3, Markdown4, Markdown5, CPI and Unemployment.

All of the Markdown columns’ missing values have been replaced with 0 as there were no markdowns done in that time period.

In order to tackle the first problem which is to find if infusing Markdowns affects our Sales or not, we have used Linear Regression Model. That is so because we have one Dependent Variable and an Independent Variable. Regression estimates are used to describe data and to explain the relationship.

We imported the Features dataset in our code using a Pandas Dataframe named as “***features\_df”***. We created a scatter plot using pyplot from the Matplotlib library for Fuel Prices vs Total Markdowns. As per the plot, we do see a consistency in the Markdowns despite what the fuel prices are. There has been a time when the fuel prices were lying between 3.0 to 4.0 and that the markdowns were higher than the previous recorded values. However, they were still consistent. There is an outlier where the fuel price lies around 3.80 and the total markdown is above 700,000.



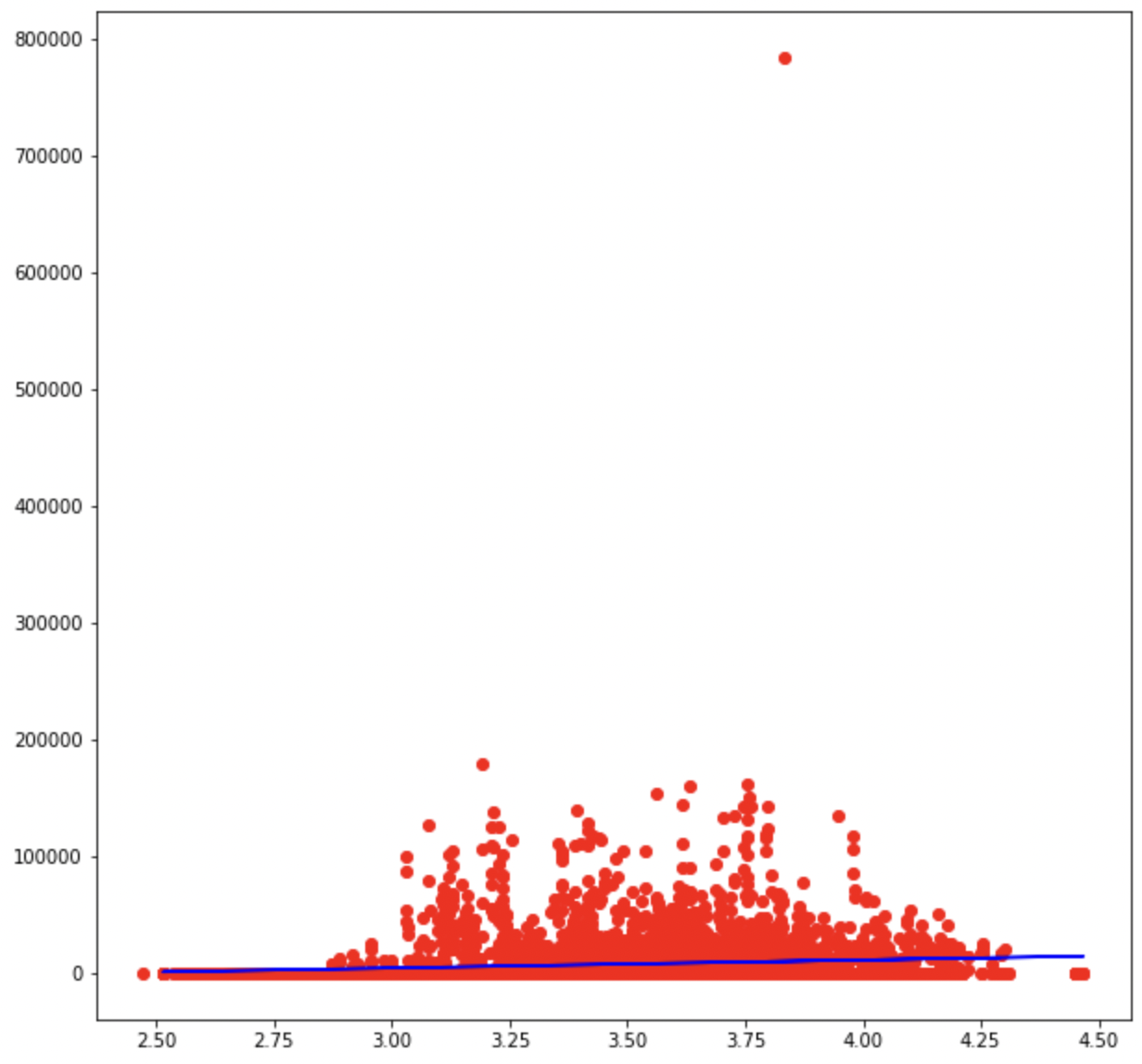
Then we split the data to be used for training and testing in the ratio of 75:25. So, now we implemented the LinearRegression() method on our training data to create our model and thereby testing the above said model on the testing data split. Here are the results for the Predictions:

Intercept: [-15503.88881997]

Co-efficient: [[6706.62370654]]

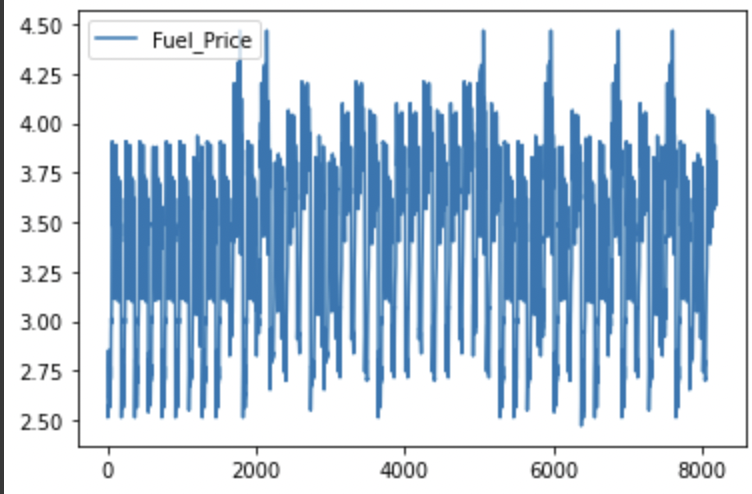
Regression Score: 0.0341075023022539

Although the Regression Score is not as simple to analyze, we do know that the Regression Score should lie between 0 and 100 in order to give positive results. However, in this case, the results are slightly above 0; still they are not negative so the two variables are not completely Irrelevant. However, we do not see the Markdowns to be highly dependent on the Fuel Price.



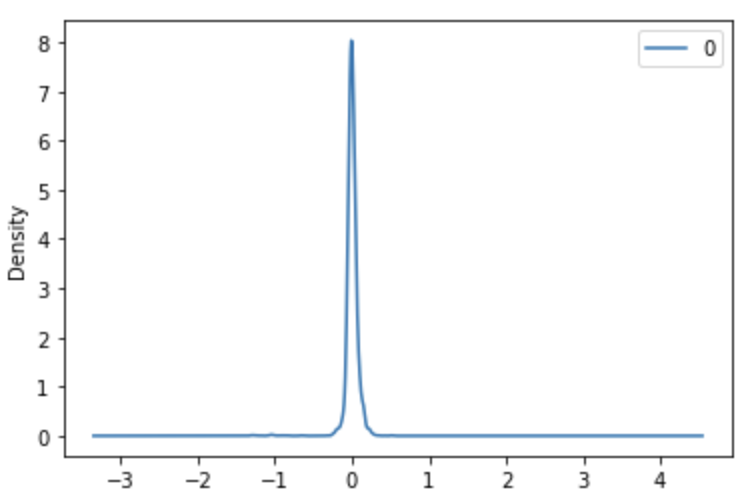
We also applied a time-based algorithm - ARIMA (Auto Regressive Integrated Moving Average) which is a forecasting algorithm that uses the information stored in the past over a longer period of time and is used to predict the future values that can form the trends and patterns.

We read the csv file and load it into a dataframe. Then we create a new dataframe to extract and use only the two variables that we need, for ease of coding the algorithm. then we plot the given two variables to see if there is a trend showing the co-variance of the two variables.



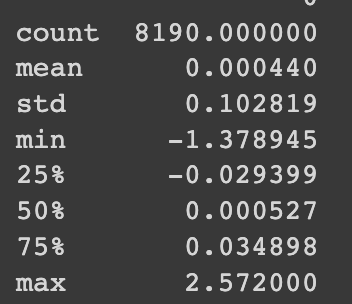
As seen in the screenshot, there is no set pattern going over here. The null hypothesis of the ADF test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) which in this case it is, we have rejected the null hypothesis and it infers that the time series is indeed stationary.

We also got a density plot of the residual error values, suggesting the errors are Gaussian, but may not be centered on zero.



We have also calculated the root mean squared error score (RMSE) for the predictions.

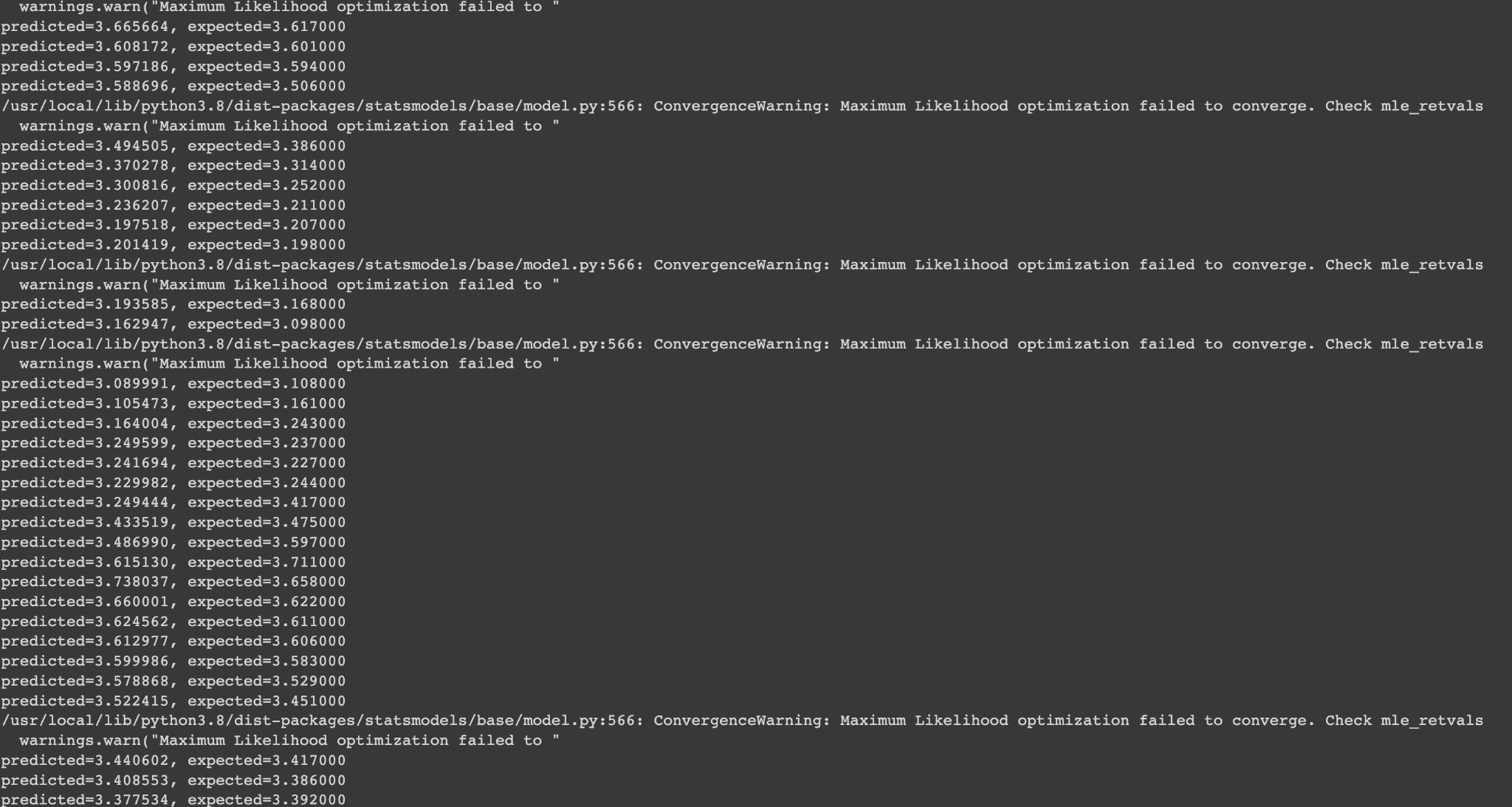
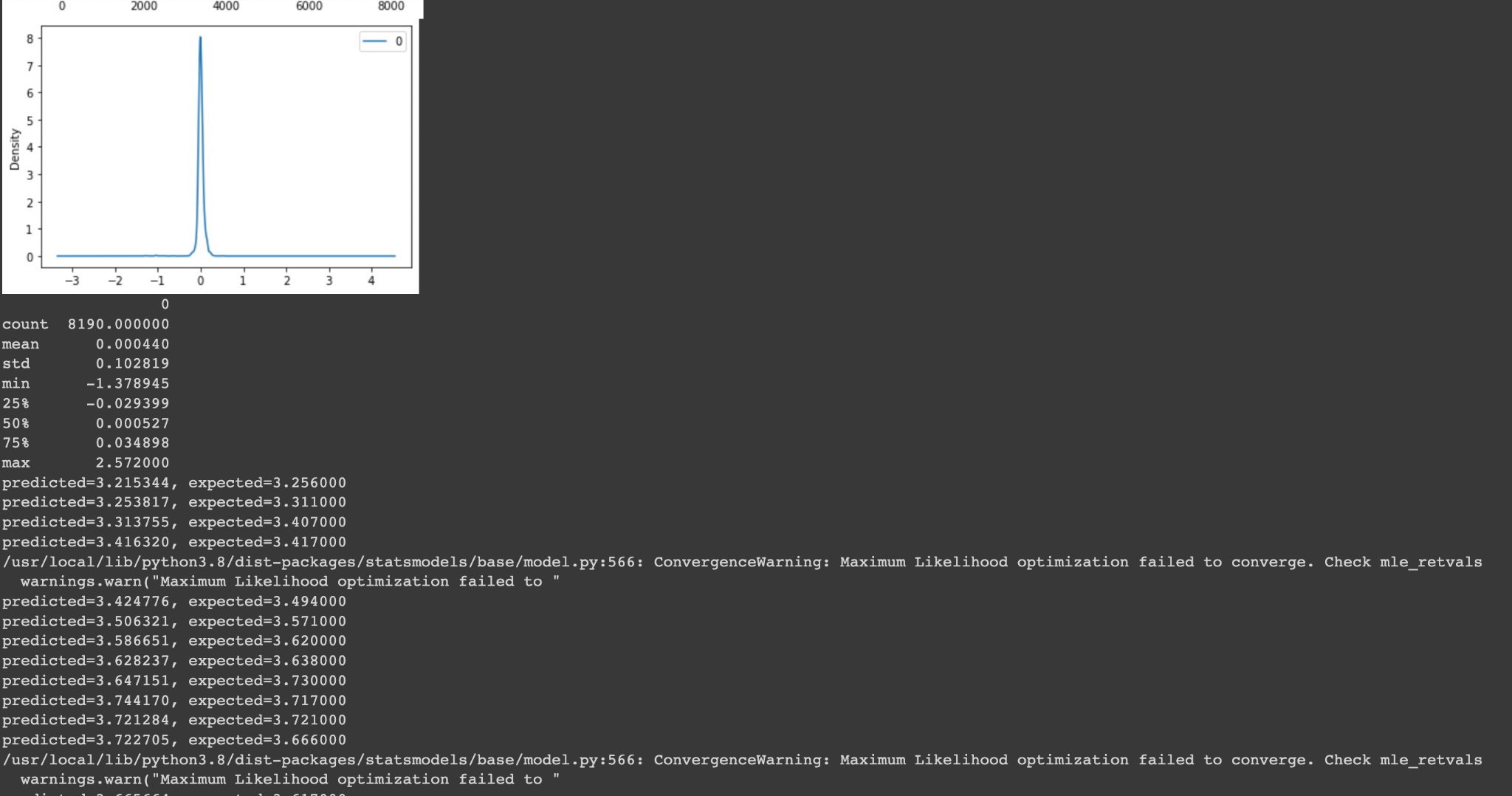
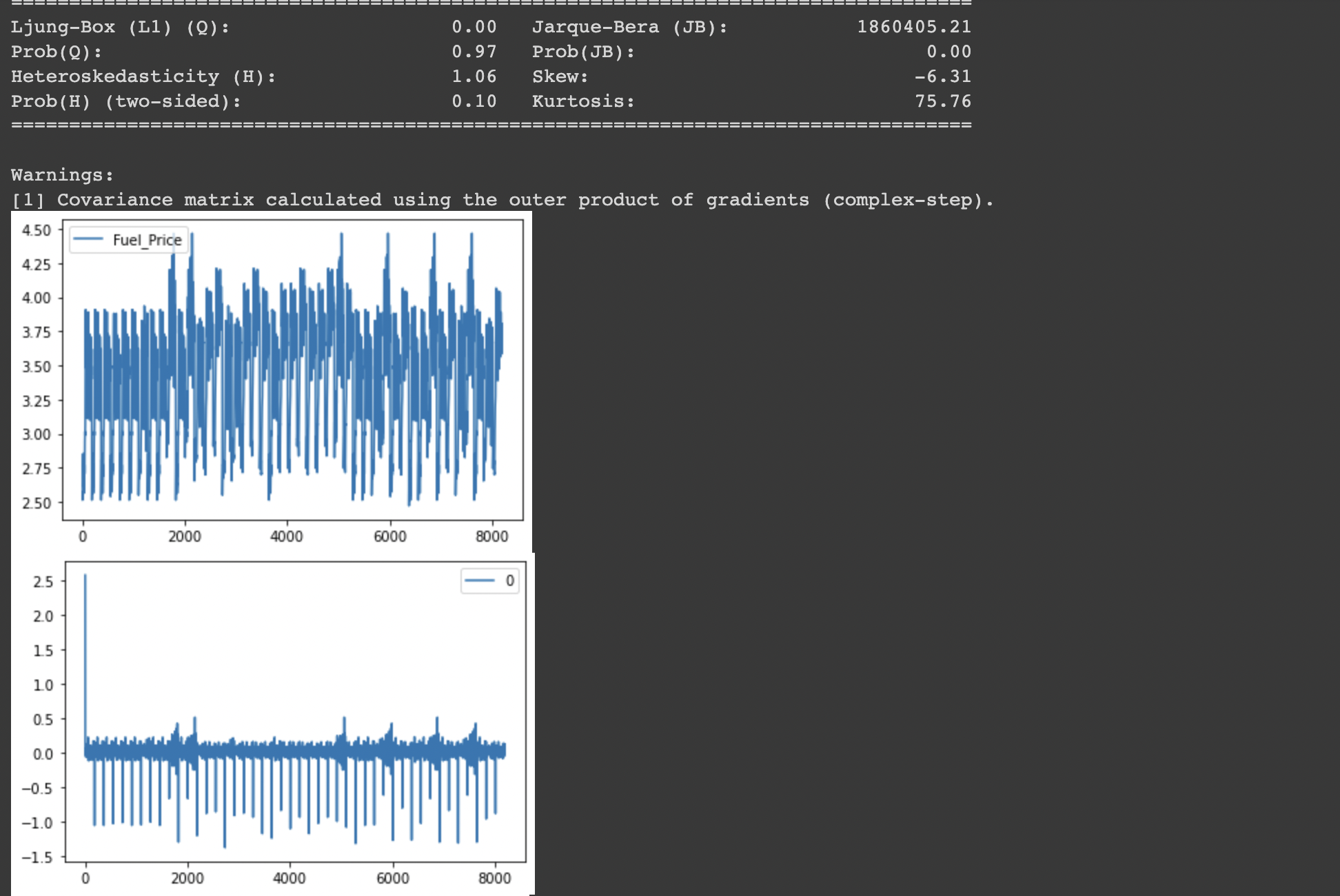
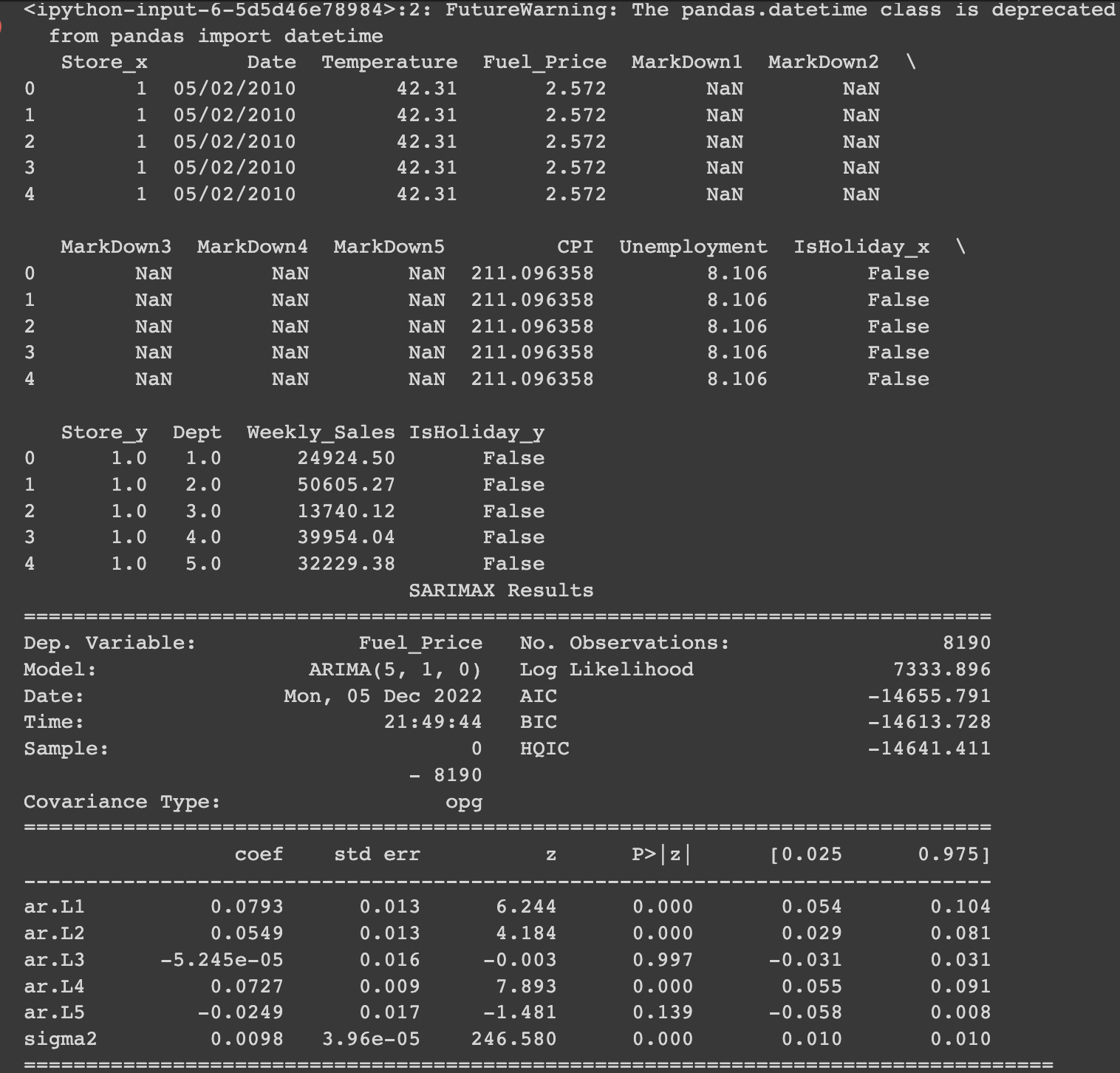
The distribution of the residual errors is displayed. The results show that indeed there is in fact almost no bias in the prediction (a zero mean in the residuals). The mean avg standard is around 0.05 which shows that there is very minor to no impact of the Fuel Prices on the markdowns infused on the sales.



For the second problem, we need to identify if the implementation of the additional Markdowns in the stores affect the Weekly Sales of the store in any manner - either positively or adversely.

We have used the ARIMA Algorithm in a similar manner and came across the point that these Weekly Markdowns are not affecting the sales in the given dataset.

We begin by creating a new data frame to merge the two CSV files which separately have Weekly Sales information and the Total markdowns infused in a given week. Then we split the dataset into training and testing modules. We can do multiple iterations of this to test the best fitting by trying various splits of training and testing modules. We also ran a density plot to see which suggests the normal distribution is being tested with mean sitting at zero. We then run the same model on the testing module to compare the expected value vs the predicted value (from the model we have created).



# References

<https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214817-eng.htm>

<https://serokell.io/blog/guide-to-data-mining-techniques>

<https://www.softwaretestinghelp.com/data-mining-techniques/>

<https://www.statisticshowto.com/probability-and-statistics/regression-analysis/>