**Group OHLC**

**Introduction- Stock Market:**

I will be looking at stock data and applying regression modeling to derive insights, relationships, and predictions between variables of the stock dataset. This is both interesting and important, as the Stock market is the biggest money machine and people lose/gain billions every day. This is not only an opportunity to test out data science skills but also to analyze if we can *actually* predict and guide our investment strategy better with historical data and machine learning.

**Aim of the model– Predictions of closing price depending on the type of day (Holidays/Weekends) and volatility index.**We will run the model at 4 PM to predict the closing price for that day. Depending on this predicted price we will either -

Questions answered by prediction-

1. **Do we Hold/Sell/Buy more shares based on the predicted closing price before and after WEEKENDS and HOLIDAYS?**
2. **Can we use the Volatility index (VIX) to improve our previous model for predicting the closing price of a stock?**

**Dataset –**

Stock market datasets are mostly focused on prices, volume, and trading behavior. In our dataset, we would be using the open and close price of a/few stocks along with their trading volume. This dataset is the basis of our experiment and has been used by all the team members in one way or other in this project. Hence not going into details here

However, in addition to the base dataset of High,Low,Close,Volume etc. I have also added dummy variables to mark weekends and holidays derived from the date variable.

Lastly, regardless of variables or Model1/2, our dependent variable would be the stock’s closing price when markets closed.i.e at 5:00 PM.

**Model outcome**

Predicting closing price is the main aim here to guide the investor to either:

* + **HOLD/BUY Shares** - In case the predicted closing price is higher than the current price. We will make a profit as prices will increase and hence, we can sell at a higher price later.   
                        **OR**
  + **SELL Shares** - In case the predicted closing price is lower than the current price. We need to sell as the prices will go down and hence, we need to avoid a loss.

**Feature Selection:**

We need to ensure we only select relevant features and avoid overfitting/underfitting. Hence we look to:

* **Removing Redundant Features -** We need to select the most relevant features before building a model. A good practice is to remove redundant features from the data and we noticed that we could choose between Close and Adj Close as they were very similar.
* **Measuring collinearity –** We measure the extent of interdependence between the Explanatory variables of the dataset by checking the **Pearson Correlation** and/or **VIF** scores**.** We noticed that all the Explanatory independent variables of the Stock dataset are highly correlated and has VIF exceeding 10 hence we remove some variables. Multicollinearity will likely exist when using these independent variables together so we need to address this before fitting the model

**The above steps would be repeated once we are done with further feature engineering.**

**Feature Engineering:  
  
For Model 1 – Predicting Closing Price by Type of Day (Weekend/Holiday/Normal)**

After **Feature Selection** we performed Feature Engineering and added **Dummy** variables in the Model. Variable **Day** was created by marking days before and after Weekends and Holidays, days on which no trading is done. Variable Day was plotted as a function of ‘date’ and converted into the given categorical variables-

* 1. **BW** - All Fridays (the day before the weekend). The two levels of BW are:

0 – Not Friday   
1 – Friday

* 1. **AW** – After Weekend – Every Monday, the first day after the weekend. The 2 levels of AH are:

0 – Not Monday   
1 – Monday

* 1. **AH** – After Holiday - First Market Open day after a holiday (market close). The 2 levels of AH are:

0 – Regular Day

1 – After Holiday

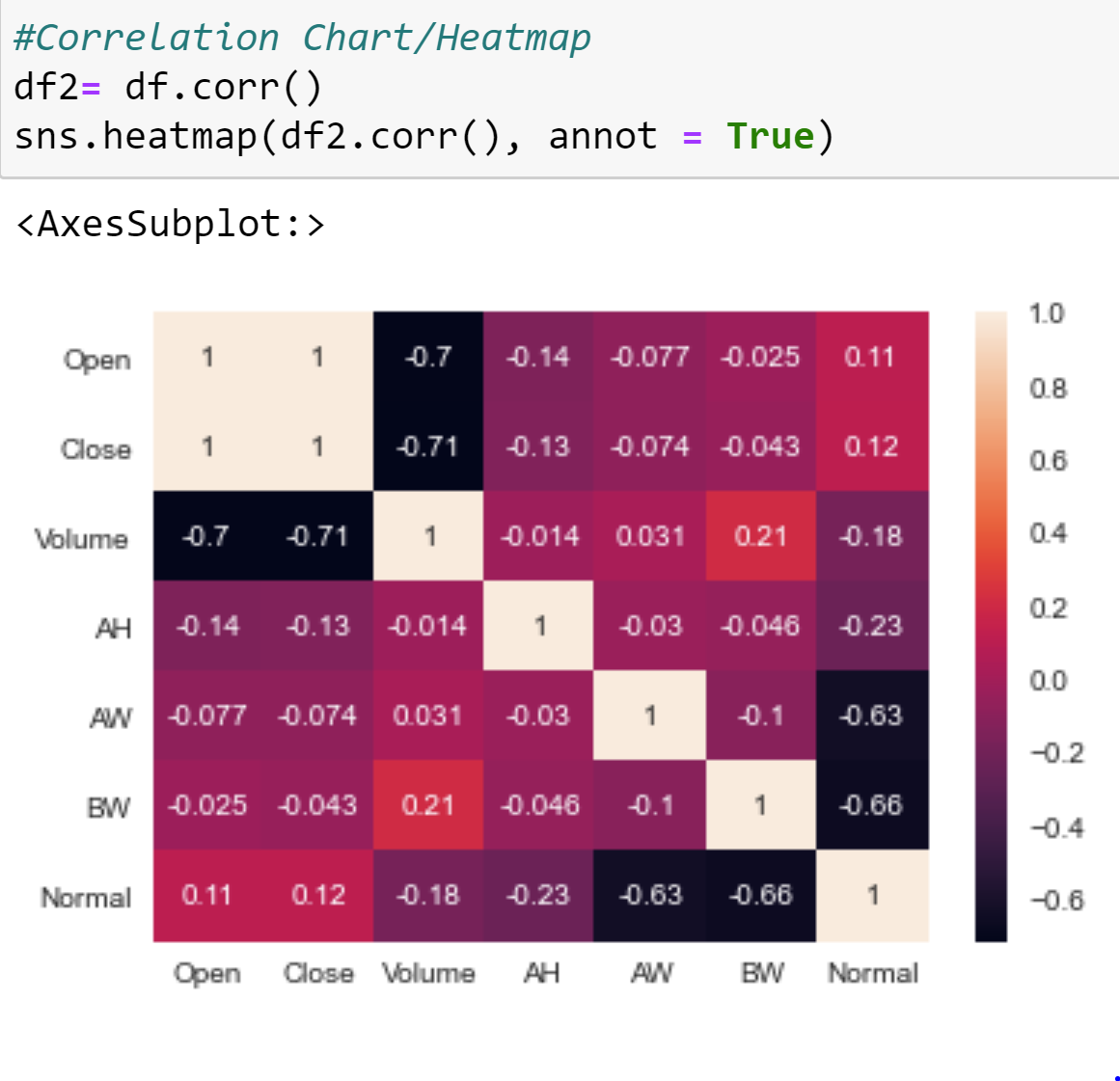
* 1. **Normal** –Market days that did not fall under any of the three categories above (Tuesday, Wed, Thurs) – This is our base level in the categorical variable Type\_of\_Day.

**For Model 2 – Predicting Closing Price with the help of the VIX price with Type of Day**

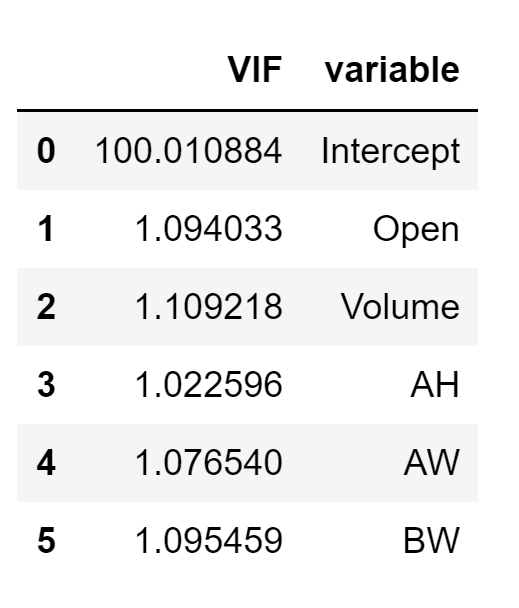
**Volatility Index (VIX)** - Checks how Volatile the market is. It is used to measure the level of risk, fear, or stress in the market. The source of data for VIX is the daily VIX data set at the same timeframe from- <https://finance.yahoo.com/quote/%5EVIX/history/>

**Correlations between pairs of FINAL independent variables in the model:**

1. To measure the extent of interdependence between variables we performed a **Pearson Correlation**.



1. Another tool used to measure the severity of multicollinearity in regression analysis is the **Variance Inflation Factor (VIF).** It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. We can see the VIF scores below –



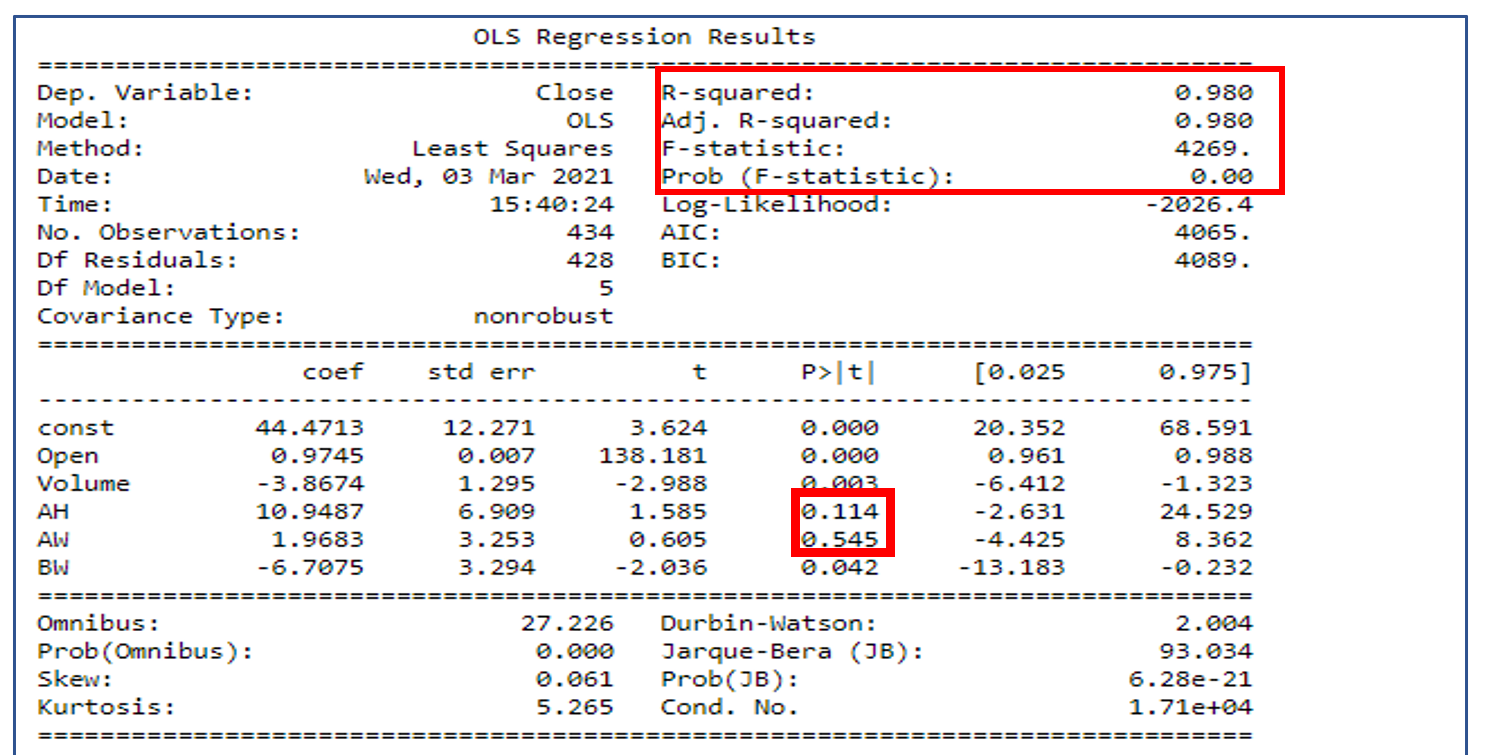
The Correlation between variables looks good and VIFs is less than 10, we can say that there are **no signs of multicollinearity.**

**Multiple Regression Analysis-**

**First Order Model**

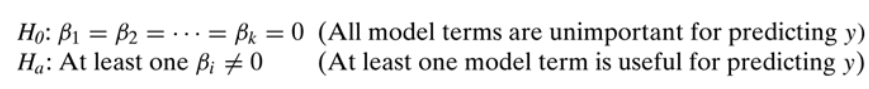
We run the first order model with our dummy variables and see how the base first order model performs. The multiple linear regression equation-**The First-order model relating y** to each of the five independent variables is:

**y(Close) = β0 + β1Open + β2Volume+ β3AH+ β4AW + β5BW**



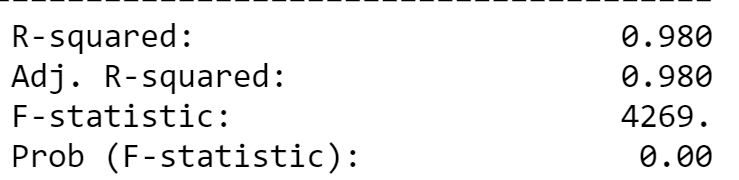
* R square (R2) is 0.98. It means that our model/the predictors (Xi ) explain 98% of the variance of Y (*Close*).
* R2 is between 0 and 1, and the closer it is to 1, the better the model.
* We see that AH and AW have high p-value and appear to be insignificant to the model.

**Model Utility Test:**

**Hypothesis:**

**Assumptions**:

* Everything else /Other factors are constant.
* The standard regression assumptions about the random error component.
* Standard Errors assume that the covariance matrix of the errors is correctly specified.



* **Testing at α = 0.05 for 95% significance.**

**P-value** = P(F > F0)=0.000

**Reject H0 since P-value< α (0.00 < 0.05)**

At a significance level of 0.05, the data provide sufficient evidence that that the model is

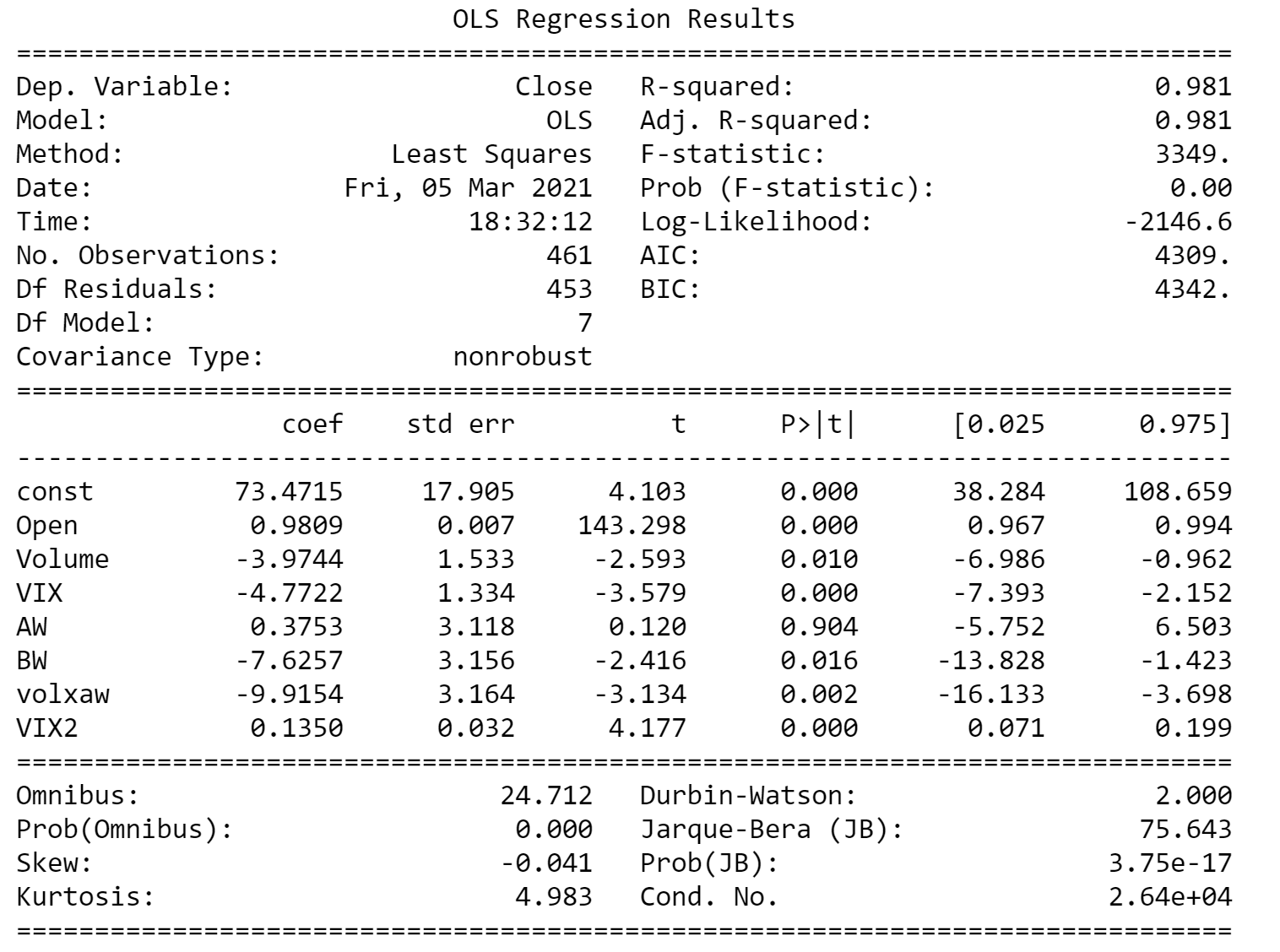
useful in explaining the closing price of Amazon stocks by using the 5 independent variables. So, **our linear regression Model, provides a better fit than the model which does not contain independent variables** AH, BW, AW**.**

**Complete Second-Order Model with Interaction Terms:**

After feature selection and multiple iterations with the above variable, we get to this final complete Second-order model. We have included VIX along with the significant predictors from the last first order model. We **keep AW regardless of its p-value as its interaction term is significant**. Multicollinearity does not exist and VIF all variables is less than 10.

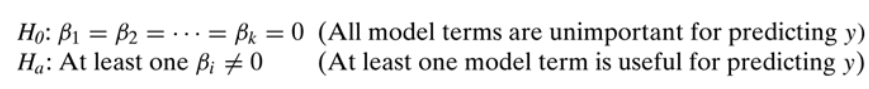
**Equation-**

**y(Close) = β0 + β1Open + β2Volume+ β3VIX + β4AW+ β5BW + β6VolxAW + β7VIX2**

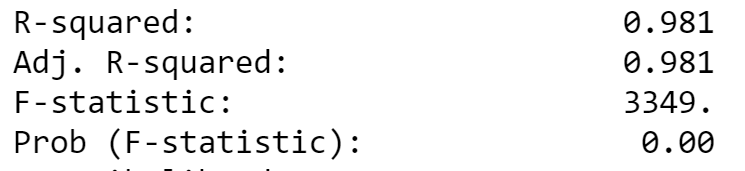


* R square (R2) is 0.98 (same as the First Order Model). It means that our model/the predictors (Xi) explain 98% of the variance of Y (*Close*).
* R2 is between 0 and 1, and the closer it is to 1, the better the model.

1. **Model Utility Test:**
2. **Hypotheses:**



**F-Test Statistic from output: 5.292 and P(F-statistic) = 0.00743**

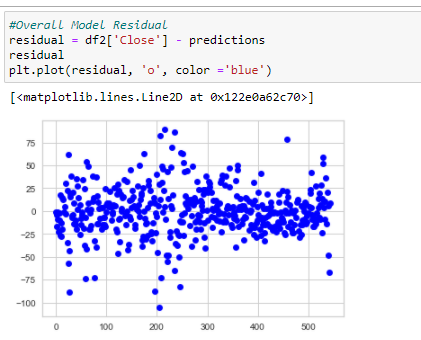


**Testing at α = 0.05 for 95% significance.**

1. **P-value** = P (F > F0) = 0.00
2. **Reject H0 since P-value< α (0.00 < 0.05)**

At a significance level of 0.05, the data provide sufficient evidence that the model is useful in explaining the closing price of Amazon stocks by using these 7 independent variables.

**Final Model Residuals:**

****

**Residuals do not have a trend so we can say that the model is a good fit. Predictors are significant as well.**

**CONCLUSION:**

We recommend using the **Holiday/Weekend Model** in practice because of the following reasons:

1. **R2 equals 0.98.** It means that by taking into the first-order effect of the predictors (Xi), the interaction between them (XiYi), and the second-order effect (Xi2), we can explain only 98% of the variance of Y(Close Price). In other words, 98% of the variation in predicting the Closing Price of Amazon’s stock has been explained by the **Complete Second-Order Model.**
2. **Model Utility**– We rejected the Null Hypothesis above which means that the model is useful in explaining the Closing Price by using the 7 independent variables, and hence our Model is a Good Fit.
3. **Residual Spread –** No trends seen in the residual spread hence we can say that the model is a fit.