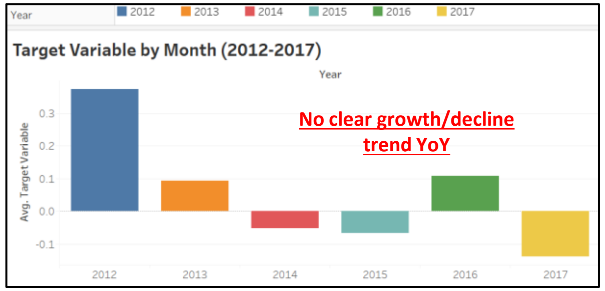
# Forecasting Challenge – Hamza Akhtar

**Based upon the data provided, I am confident that I have built a model that can predict the 2018 target values to the required degree of accuracy**

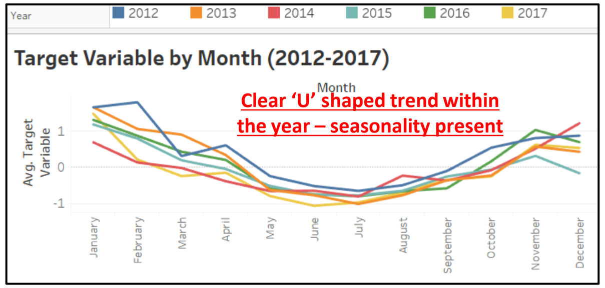
## 1

Looking at the data, we are working with a set that is defined by consecutive calendar dates. Dates come with properties such as months and 'days of the week' so we'll use these to gain some insight.
Initially we plot the data to see if there are any patterns that we see. This gives us an insight into the nature of the data we will work with.

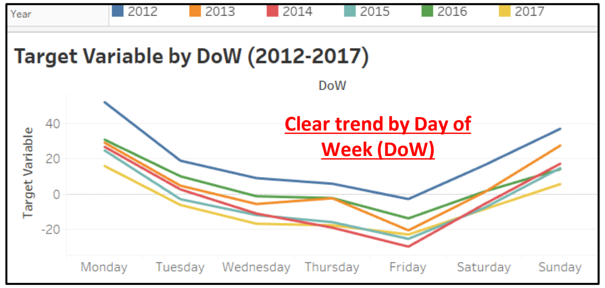
Below we see plots of the target variable by year, month and by DoW (day of the week). It's clear to see that there are trends at the latter two levels so we'll keep this in mind for when we begin to model the data in order to predict the 2018 'target variable' outputs. I will classify 'Month' and 'DoW' as variables in addition to the 5 input variables.


Target Variable by Month (2012-2017)

No clear growth/decline trend YoY


Target Variable by Month (2012-2017)

Clear 'U' shaped trend within the year – seasonality present


Target Variable by DoW (2012-2017)

Clear trend by Day of Week (DoW)

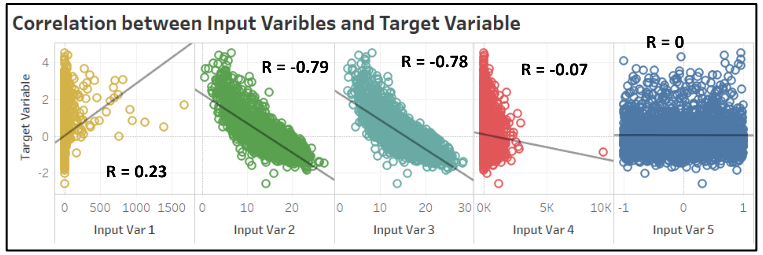
## 2

We plot the input variables against the target variable to see if there are any relationships that are substantial. We can see their correlation coefficients (R) on the plot. The closer R is to -1 or 1, the stronger the correlation. If R is 0, then there is no correlation:

- Input variable 1 and 4 have weak correlations
- Input variables 2 and 3 have fairly strong **negative** correlations
- Input variable 5 has no relationship

Given that we need to use the input variables to predict the 'target' variable for 2018, **using linear regression to model the data** we will aim to do expose the following:

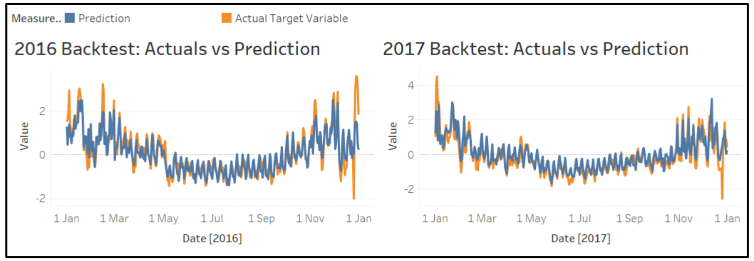
- Strength of the effect of input variables on the target variable
- Seasonality within the dataset


Correlation between Input Varibles and Target Variable

## 3

After running through linear regression models and testing the significance of variables, the conclusion reached is that in my model, only input variable 2 along with month and DoW are necessary to effectively forecast 2018 (details on reaching this conclusion are in the next slide). To support this claim, backtests have been ran over both 2016 and 2017, producing the following results:

**2016 produced an R-Squared value of 82%**
**2017 produced an R-Squared value of 78%**
**Given that the current best score for the 2018 prediction is 62.2%, the backtests give me confidence in the selected model.**


2016 Backtest: Actuals vs Prediction

2017 Backtest: Actuals vs Prediction

2016 Backtest: Date used to train the model spans 2012-2015

2017 Backtest: Date used to train the model spans 2012-2016

## 5

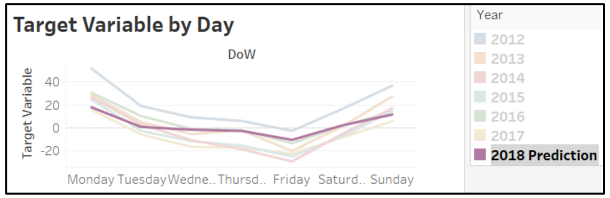
### RISKS

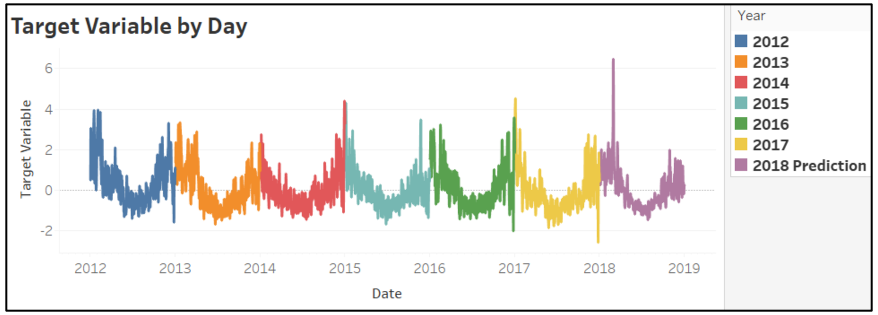
Given that the context of the data provided is unknown, this can pose a number of risks within the assumptions that we have made within the model selection and outputs:

- There are a number of extreme data points that have a big influence on the model. We can't remove these from the model as their nature is unknown at this time.
- When forecasting we can compute accuracy measures such as MAPE on the back tests, however given that a majority of our data lies between the number -1 and 1, MAPE is not a good measure to use. Again, knowing what the data represents would indicate the best accuracy measure to use.
- There are couple of extreme points in the prediction dataset. These could have a big effect on the R-Squared of the 2018 prediction however we have to go off of the data provided. If the context was known, there's a case for capping these extremes

## 4

In the visual on the right, we can see the target variable from 2012-2017 along with my prediction for 2018 based upon the input variables I was given. We can see that the shape of the 2018 prediction generally follows that seen in the prior years. The seasonality (monthly trends) are present in the prediction and below we can see that the 2018 prediction follows a similar trend at the DoW level:


Target Variable by Day


Target Variable by Day

# Forecasting Challenge – Hamza Akhtar

The accuracy of the model could be increased through exploring interaction terms. Due to the unknown nature of the dataset, it's difficult to determine the contextual requirement of variables and extreme data points.

As mentioned in the previous slide, we will use linear regression to model the data and produce a prediction for 2018. Specifically, **we will use the ordinary least squares (OLS) method**. All instances of regression have been initialised and ran in Python.

***Please note the following abbreviation: Input variable X = VarX***

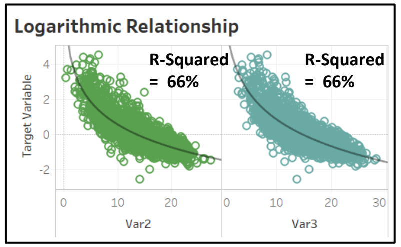
The model that includes maximum variables (from what we have explored already) is as follows: Target ~ Var1 + Var2 + Var3 + Var4 + Var5 + Month + DoW

This model may not be the best selection, especially seeing as we determined in the last slide that some variables had extremely weak correlations to the Target variable.

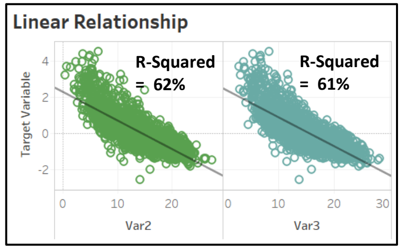
There are some further points to note:

## 1

When we plot Var2 and Var3 against the Target variable, the relationship seems to support a logarithmic pattern more than a linear pattern. Therefore within the models used, **the values represented by Var2 and Var3 will be log(Var2) and log(Var3) respectively.**



Logarithmic Relationship



Linear Relationship

When running regression models, outputs can be significantly skewed by the presence of collinearity/multicollinearity. This is the term given to the condition where 2 or more independent variables have an extremely high correlation. In the case of our dataset, Var2 and Var3 have collinearity present **therefore we will remove Var3 from the modelling process entirely**.

Correlation Coefficients

| | Var1 | Var2 | Var3 | Var4 | Var5 |
|---|---|---|---|---|---|
| Var1 | - | 0.28 | 0.29 | 0.03 | 0.00 |
| Var2 | 0.28 | - | 0.99 | 0.03 | 0.00 |
| Var3 | 0.29 | 0.99 | - | 0.02 | 0.00 |
| Var4 | 0.03 | 0.03 | 0.02 | - | 0.03 |
| Var5 | 0.00 | 0.00 | 0.00 | 0.03 | - |

## 2

### Model Selection

We need to undergo a process of model selection. This can be done using a variety of methods. We can use a method such as 'backward selection' or you can run all of the possible models (128 models for the 7 variables defined) and compare the diagnostics and outputs. I have tested two methods 'backward selection' and 'forward selection' using **OLS** and I arrive at the same final model. We will be using the 95% confidence interval (CI) as a bench mark for variable inclusion. This will give us a probability of 95% that our model contains the true value of the regression coefficients.

I will describe the process of backward selection below:

- Start with the following model: Target ~ Var1 + log(Var2) + Var4 + Var5 + Month + DoW (non-inclusion of Var3 due to collinearlity)
- Check the F-value is significantly high. This indicates that we should reject the null hypothesis that the regression coefficients are all 0 (hence an 'intercept-only'model)
- Remove the variable with the p-value that is greatest and furthest away from the critical value of 5%. Repeat until all remaining variables are "statistically significant" (p-value<=5%)

## 3

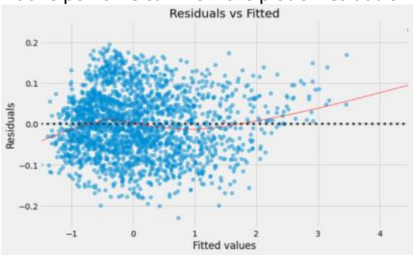
### Model Selection Continued

- In doing so we removed variables in the following order: Var5, Var4, Var1
- We were left with the following model: Target ~ log(Var2) + Month + DoW. All variables had a p-value less than 1%. This tells us that 'Month' and 'DoW' are both significant in predicting the target variable along with Var2 (albeit transformed as log(Var2) in this case).
- This model explained 66.4% of the variance seen within the dataset (adjusted R-Squared)

At this point, we then convert 'Month' and 'DoW' into categorical/dummy variables. As we saw in the previous slide that there were clear trends within the target variable by DoW and Month, we should therefore model at both of these levels.

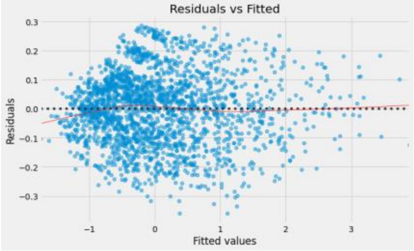
Using the model Target ~ log(Var2) + C(Month) + C(DoW) we report back a model that explains 70.5% of the observed variance.

At this point we can view the plot of residuals:



We can see that residuals are generally symmetrical and evenly distributed. This is a good fit for a vast majority of the data. However, we see a tail off for the extreme positive values. Although there are not many values that are extreme highs, it's clear to see that these point are very influential.

If we ran the same model but removed data points that were 3 standard deviations away the average by Month by DoW, we would see the following residual with an adjusted R-Squared of 80.4%.



This is telling us that **0.7% of the entire data set has an influence of 10.0% on the adjusted R-squared.**

Removing this small amount of data would produce a better fit however **we risk overfitting and we have no means to remove this data as we don't know what it represents due to the unknown context.**

## 4

### Model Improvements - Interaction Terms

As we've chosen to use regression analysis, a notable property worth exploring is interaction terms,

Interaction effects occur when the effect of one variable depends on the value of another variable.

Hence in our model we could classify 'Var4Var1' or 'Var2Var3Var5' as additional variables.

Given our 5 input variables, there are a possible 26 interaction variables (ignoring permutation of course). This means that we would have 31 variables along with our two dummy variables (month and DoW) – we would have 33 variables overall.

This would result in over 8 billion possible models which would require lots of processing power.

When looking at the correlation between the target variable and different interaction terms, immediately I noticed that **every interaction term in which Var5 is present has a correlation coefficient of 0**. Removing these would immediately bring the total number of variables down to 17, resulting in 131,071 models. For every model, you could export the diagnostics (such as AIC, BIC, R-squared and Log-Likelihood) along with backtest data. You could then choose the most suitable model based upon the defined criteria.

### Model Improvements – Backtest Training Data

When we conduct the backtests, we could try and adjust the date span of the training data (eg. Last 2 years, last 3 years). This could have an effect on the accuracy of predictions. In the 0.7% of data that had a 10% influence on the prediction model for 2018, half of the 0.7% data was from 2012.

## 5

### Conclusion

I am confident that the process of model selection has resulted in the a model that will predict the 2018 values to a respectable degree of accuracy.

I believe that in order to maximise the potential from the entire dataset, if we had the required processing power we could run all of the potential models and effectively 'leave no stone unturned'.