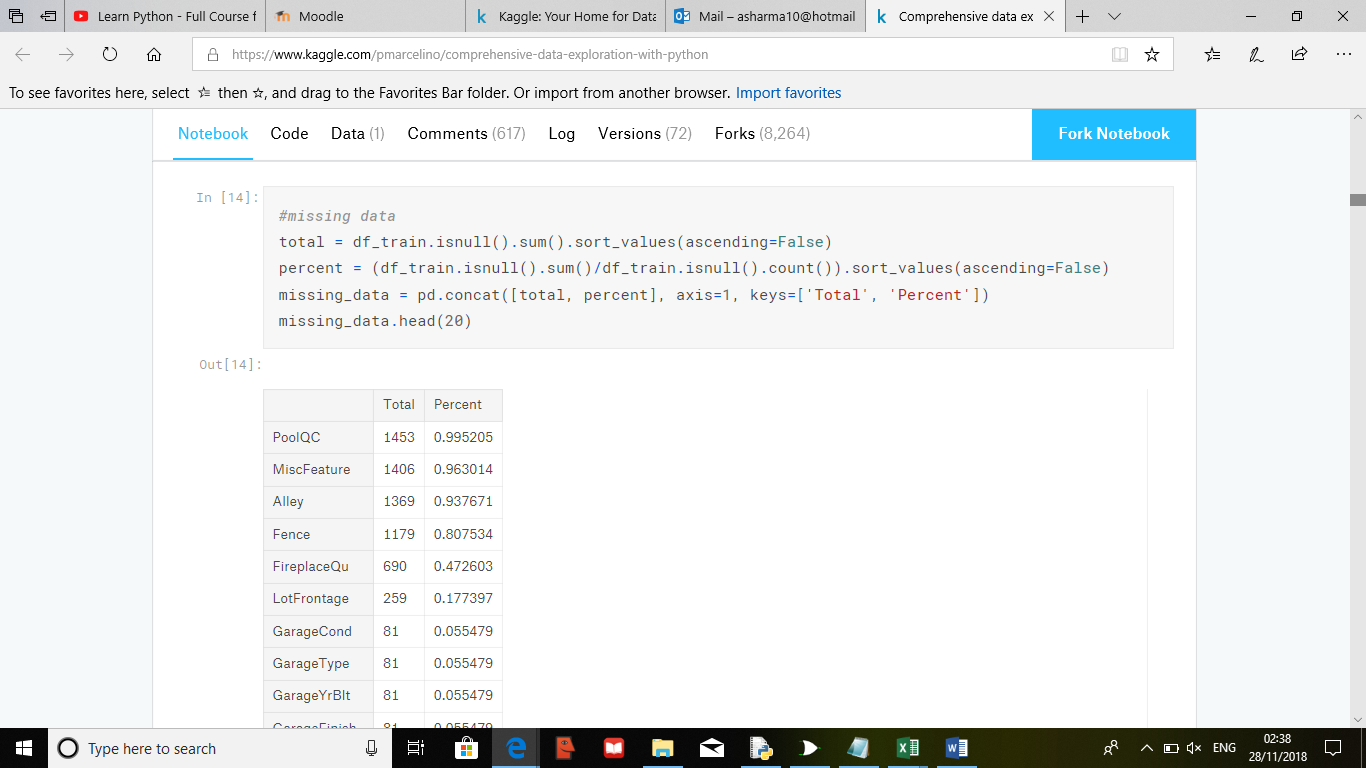
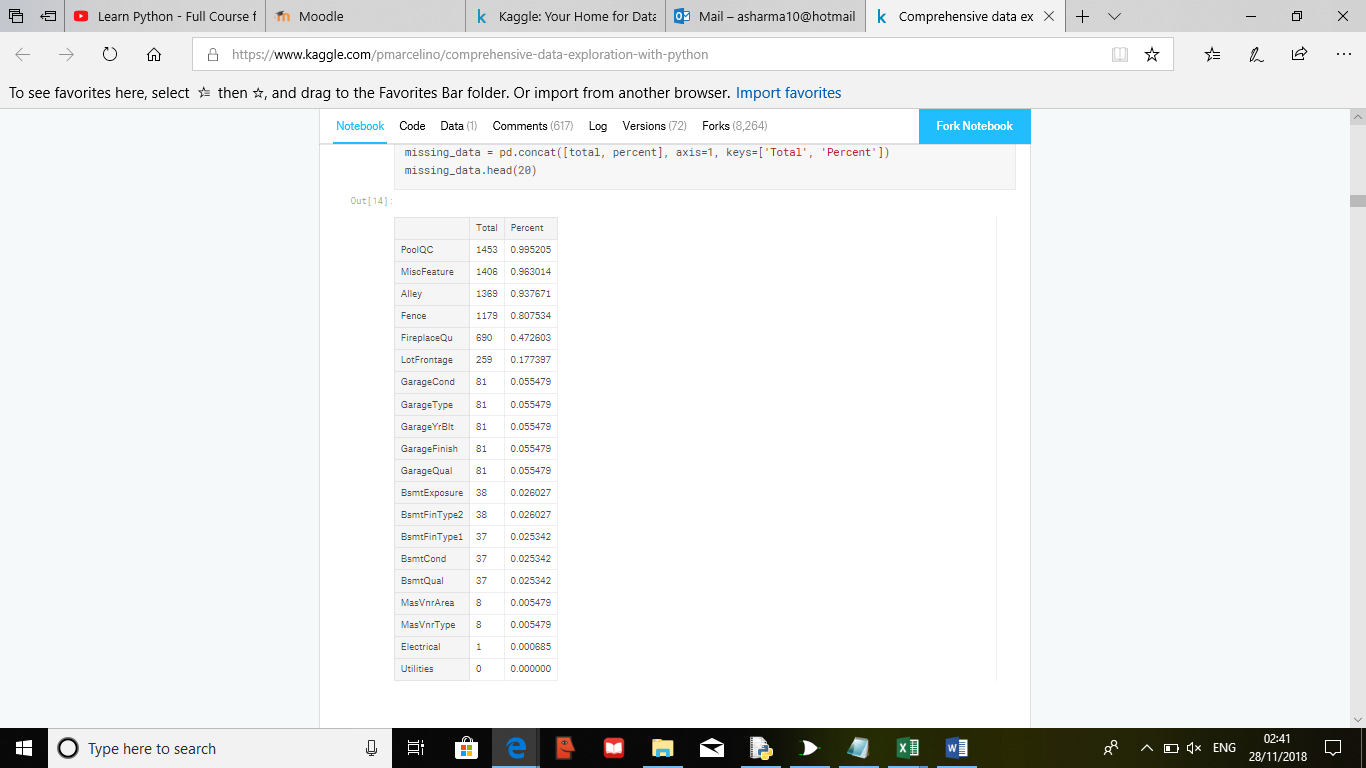
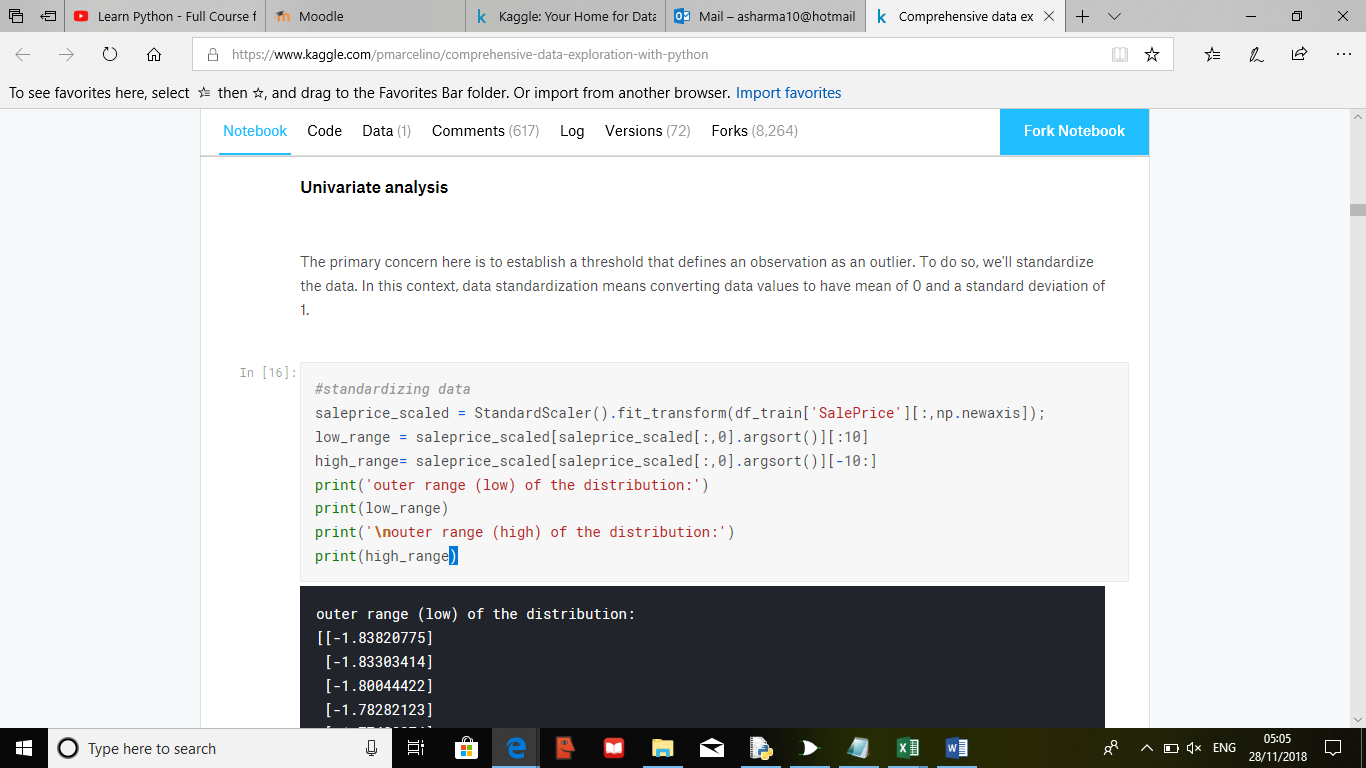
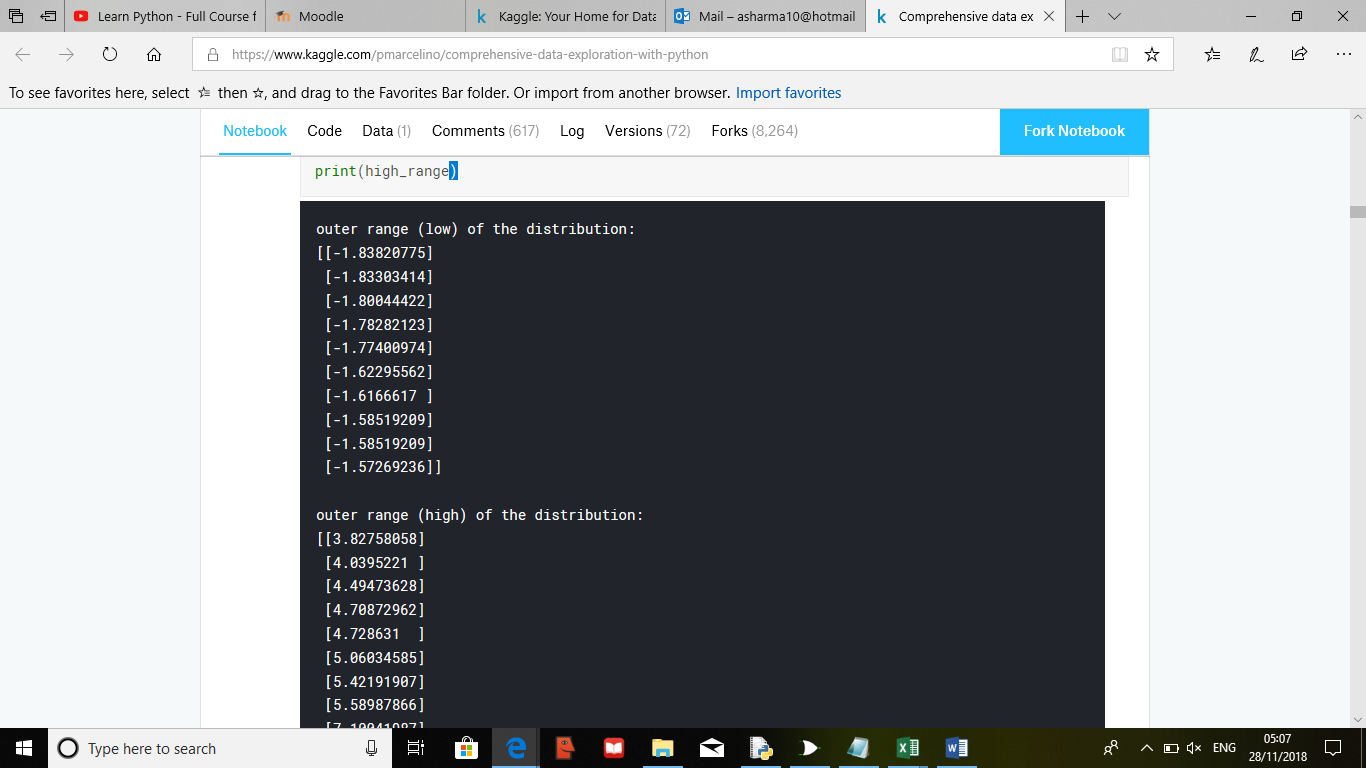
Missing Data Part 4

* Does missing data have a pattern or is it completely random??
* Missing data usually means that the sample size is reduced. So we need to be sure that the missing data isn’t bias
* When more than 15% of data is missing it is sometimes best to delete the variable that the missing data corresponds to, we don’t try and fill in the missing data
* 
* The table on the left represents different variables, we should for example delete, ‘PoolQC’. Deleting these variables doesn’t will not matter as there not factors which are usually looked at when buying a house and can even be considered as an outlier.
* Outliers are important can provide information about specific behaviours

Univariate analysis

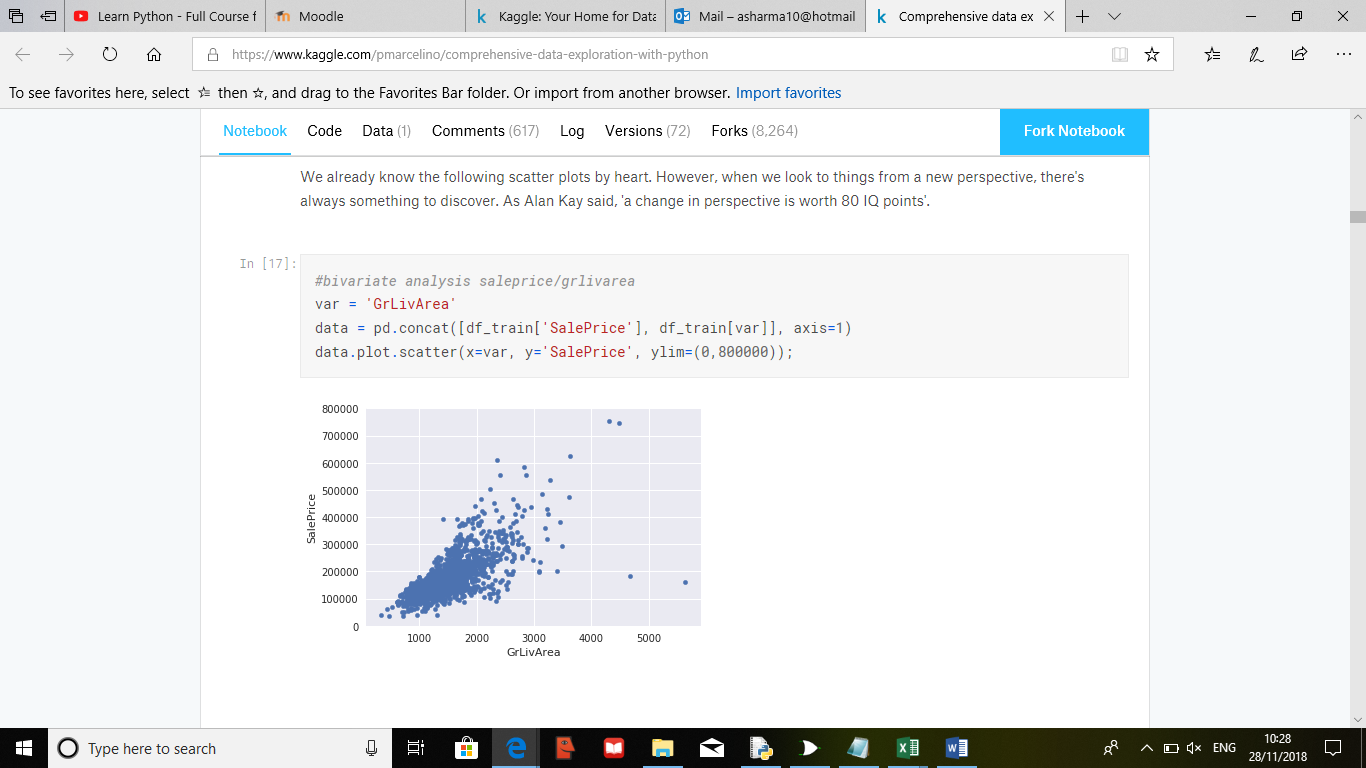
1. First we need to create a threshold that defines an observation as an outlier. To do this we standardize the results, this means converting data values to have a mean of 0 and standard deviation of 1.





* As you can see the outer range (low) of the distribution is close to 0 whereas the (high) is between 0-7 which is close to 1

Bivariate analysis – Another way of quantitative analysis, analysis of two variables to determine the relationship between them



* The scatter graph above contains outliers, and we would usually delete them (two with the biggest ‘GrLIVArea’ values). However in this case the two outliers seem to be following the trend and looks like two special cases. So we will keep them