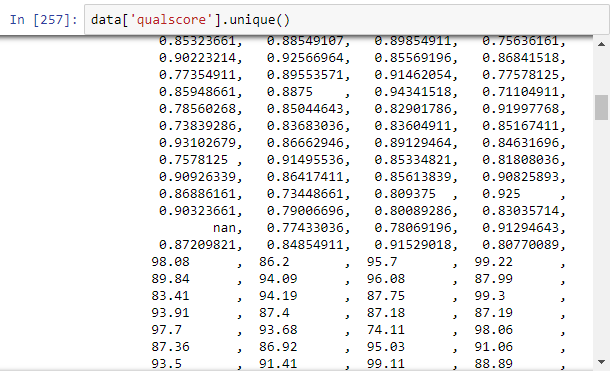
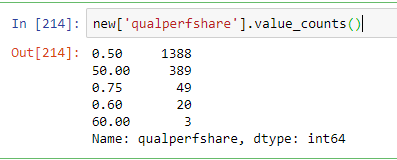
1. **Improving/Preparing the Data for Analysis**

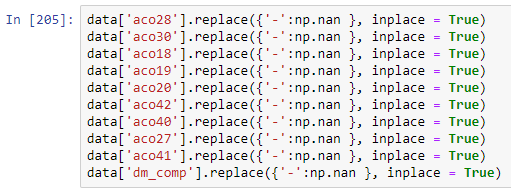
* **Erroneous Values:** Some columns (‘qualscore’, ‘qualperfshare’) had some mixed values of percentages and fractions. For example:

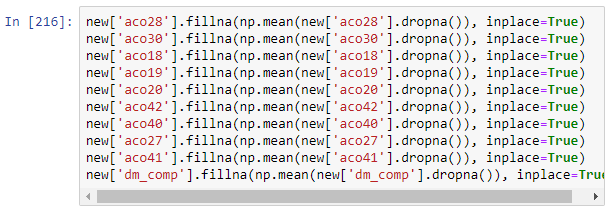




So, these types of values were replaced with new values so that they all are on same scale.

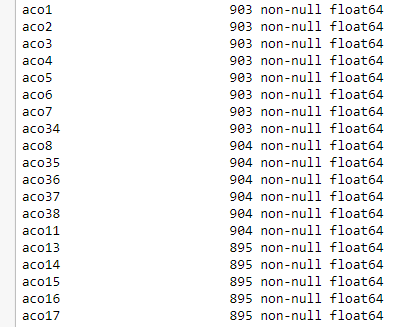
* **Unidentified values:** Some other columns (variables) like ‘aco28’, ‘aco18’, ‘aco30’, ‘ac027’, ‘dm\_comp’ etc had unidentified values like ‘-‘ which needed to be replaced.





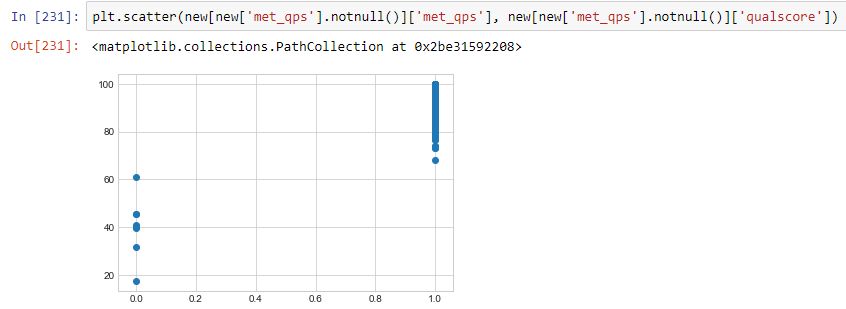
So, replaced these values first with NaN and then imputed the missing values with Mean of the variables (Could also be replaced by Median).

* **Null Values:** Variables like ‘aco1’, ‘aco2’ ……. ‘aco37’ , ‘aco38’ had null values which I filled with the mean of the corresponding variables:



For Variables like ‘n\_ben\_race\_asian’ , ‘n\_ben\_race\_hisp’ , ‘n\_ben\_race\_native’ , null values were filled with the Median values of the corresponding variable.

For ‘met\_qps’ , filled null values with the help of the other variable ‘qualscore’ because they both were correlated and interdependent.





Filled the remaining missing values in the variables ‘aim’ , ‘qualscore’ , ‘met\_qps’ by -1.0 as now these values were seemed as MCAR (missing completely at random), so -1.0 equivalent to ‘missing’.

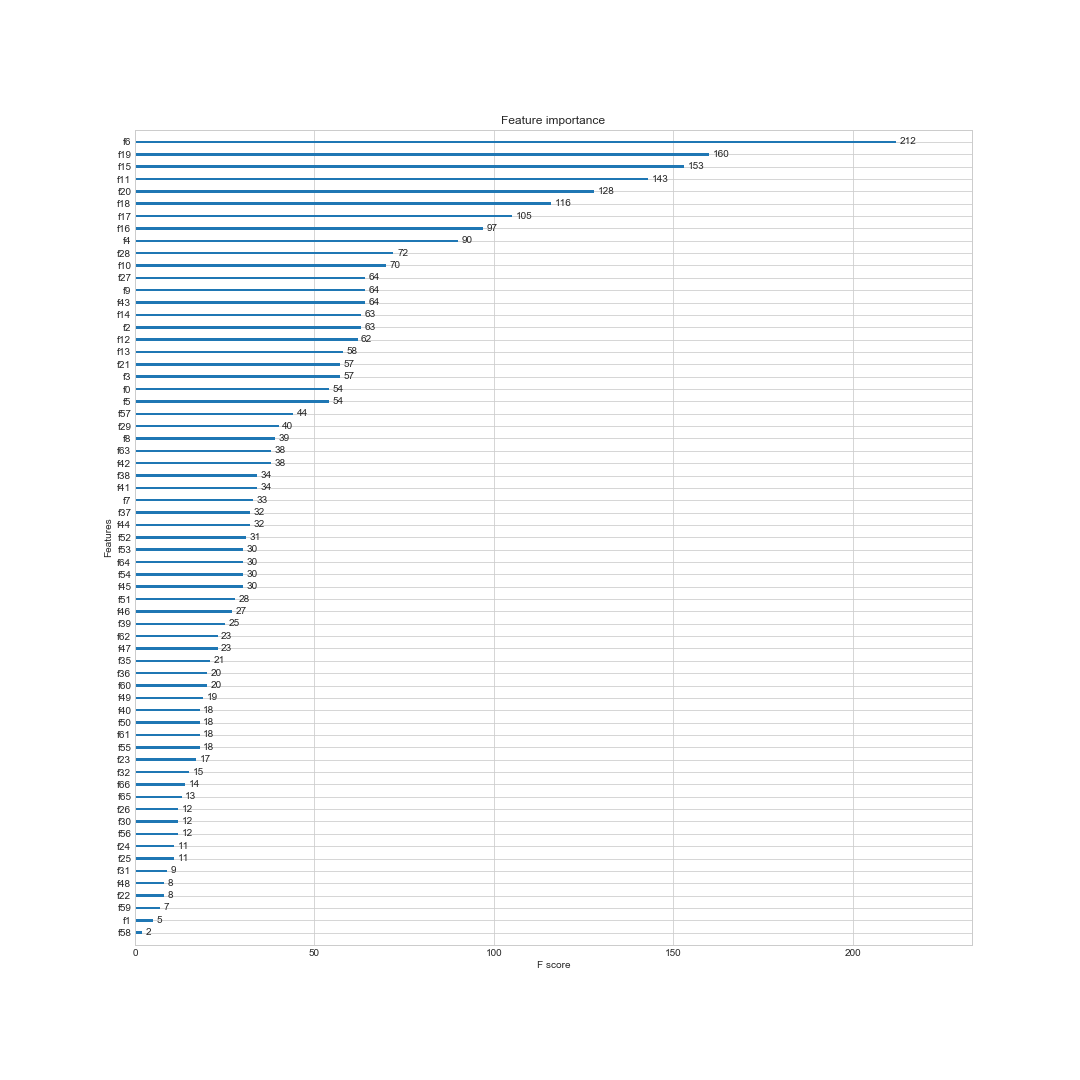
**Then, I saved this data in csv format for my convenience as the name ‘15118012\_Amarpreet\_Singh\_Data’.**

1. **Key Drivers of ‘per\_capita\_exp\_total\_py’ : By Feature Selection Techniques and Regression Analysis:**

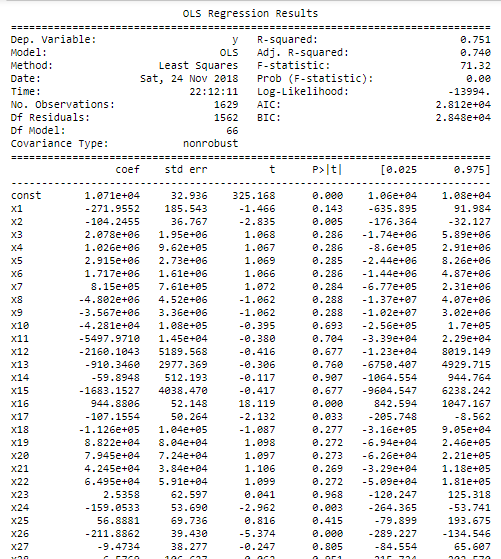
* **Correlation based Feature Selection/Importance:** Here Feature Selection is done by watching the correlation (pearsonr) between the target variable (‘per\_capita\_exp\_total\_py’) and other input variables. I put the threshold at 0.3 .



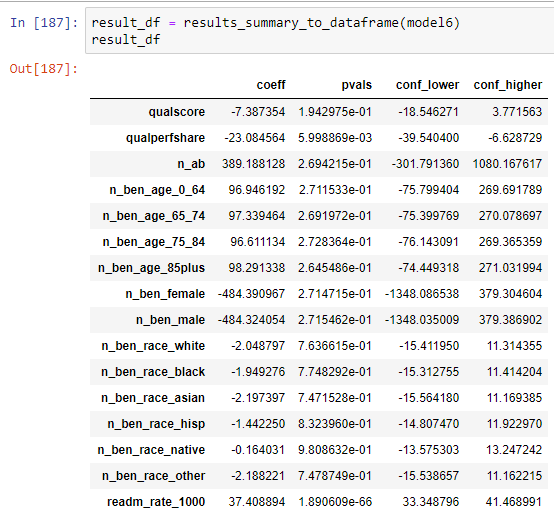
* **Feature Importance by Tree-Based Models:** Tree – Based models like XGBoost and RandomForest have an excellent function. We can check Feature Importance on the basis of F-Score (Variance explained) by the independent (input variables).



* **Estimating Regression Coefficients:** Larger the coefficients of the input variables in the Regression Models, more important is the corresponding feature. This also needs standardizing the Data first.



* **Calculating R-squared when each variable is added to a model that already contains some of the other variables:** The associated change in R-squared represents the improvement in the goodness-of-fit that is due solely to that last variable. Here I have used the Forward Selection Technique.



**THANKS**