**Task 3**

STEPS PERFORMED BY ME:

1) After reading the dataset into the notebook, I concatenated the tables

1. "London\_historical\_aqi\_forecast\_stations\_20180331" and "London\_AirQuality\_Stations"
2. "London\_historical\_aqi\_other\_stations\_20180331" and "London\_AirQuality\_Stations"

Using the column “ station\_id” as we needed to interpolate other columns based on the “longitude” and “ latitude” which were only present in "London\_AirQuality\_Stations".

2) According to my understanding, interpolation means filling in the missing and null values. I tried this in the first merged table with griddata (linear, cubic and nearest). There were approximately 122000 missing values in pm2.5, no2 and pm10 but after performing interpolation I was still left with 10,000 missing cells.

3) I then used linear regression to fill in those missing values, which made no changes probably because pm2.5, no2 and pm10 are all dependent features.

4) So I filled these using the median, which does not seem the right thing to do.

5) I am not able to perform the 3 fold cross validation because of some error in the X\_train. I tried imputation as well, but couldn’t perform it successfully.