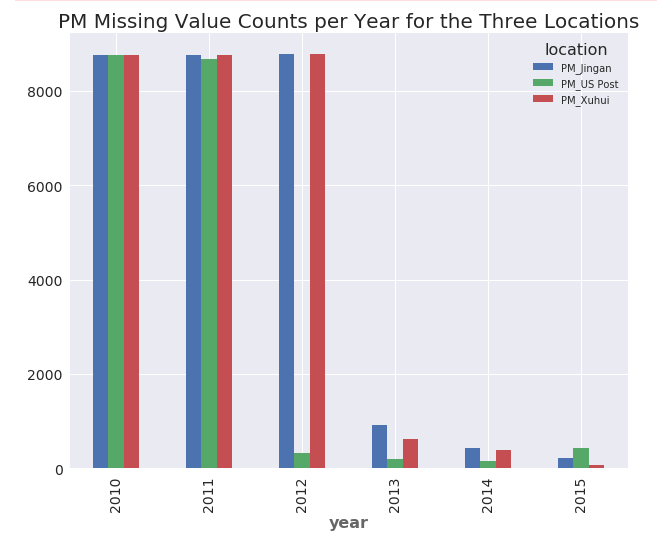
Project: Capstone Project - Data Wrangling

The Shanghai dataset used is between Jan 1st, 2010 to Dec 31st, 2015. Missing data are denoted as NA. The percentage of missing values for all attributes is as follows:

* No: 0.000000
* year: 0.000000
* month: 0.000000
* day: 0.000000
* hour: 0.000000
* season: 0.000000
* PM\_Jingan: 53.027537
* PM\_US Post: 35.267382
* PM\_Xuhui: 52.097596
* DEWP: 0.024722
* HUMI: 0.024722
* PRES: 0.053248
* TEMP: 0.024722
* cbwd: 0.000000
* Iws: 0.022821
* precipitation: 7.623992
* Iprec: 7.623992

PM2.5 (dependent variable) collected at the three locations namely PM\_Jingan, PM US Post and PM\_Xuhui have high percentage of missing values. After investigating the annual distribution of PM missing values for the three locations, we found that PM\_US Post started collecting data since 2012 while the other two locations started since 2013.



We didn’t find any special pattern for the missing values from the three locations related to season, month and weekday. We performed some statistics analysis and the results are as follows:

------------------------------------

|Statistics for location at An Hui:|

------------------------------------

Mean of PM2.5 is:57.67898685934336

Median of PM2.5 is:43.0

Standard Deviation of PM2.5 is:48.412620382306855

Skewness of PM2.5 is:2.67797650833

Kurt of PM2.5 is:13.3443464096

------------------------------------

|Statistics for location at Jing An:|

------------------------------------

Mean of PM2.5 is:57.395991902834005

Median of PM2.5 is:44.0

Standard Deviation of PM2.5 is:47.61617929355327

Skewness of PM2.5 is:2.59401623698

Kurt of PM2.5 is:12.6398761958

------------------------------------

|Statistics for location at US Post:|

------------------------------------

Mean of PM2.5 is:52.91054378800787

Median of PM2.5 is:41.0

Standard Deviation of PM2.5 is:42.576137669466284

Skewness of PM2.5 is:2.57515565047

Kurt of PM2.5 is:13.8007368705

From the statistics and other distribution analysis for the PM2.5 values from three locations, we came to the following conclusions:

* All three PM2.5 values had similar distributions that were right skewed with long tails.
* PM2.5 values from Xuhui and Jingan had higher median and standard deviation than those from US Post
* We cannot reject the null hypothesis that PM2.5 values from Xuhui and Jingan had identical average values.

Because higher PM2.5 values represent poorer air quality, we decide to choose the average PM2.5 values from Xuhui and Jingan and drop the records where both values are missing. (see the following code)

PM\_gov = PM\_gov[PM\_gov['PM\_Jingan'].notnull() | PM\_gov['PM\_Xuhui'].notnull()]

PM\_gov['PM25'] = PM\_gov[['PM\_Jingan','PM\_Xuhui']].mean(axis=1)

Since the missing value percentage for the independent variables are all less than 10%, we will use mean values to replace missing values respectively. (see the following code)

PM\_gov = PM\_gov.fillna(PM\_gov.mean())

Since the PM2.5 value is right skewed with outliers occur at the high values, we won’t remove the outliers as we’d like to have them represented in the model itself.