EDA(S670)-Mini-Project-1-Report

Syam Sundar Herle,Rahul Rahagate,Sidharth,Saheli Saha

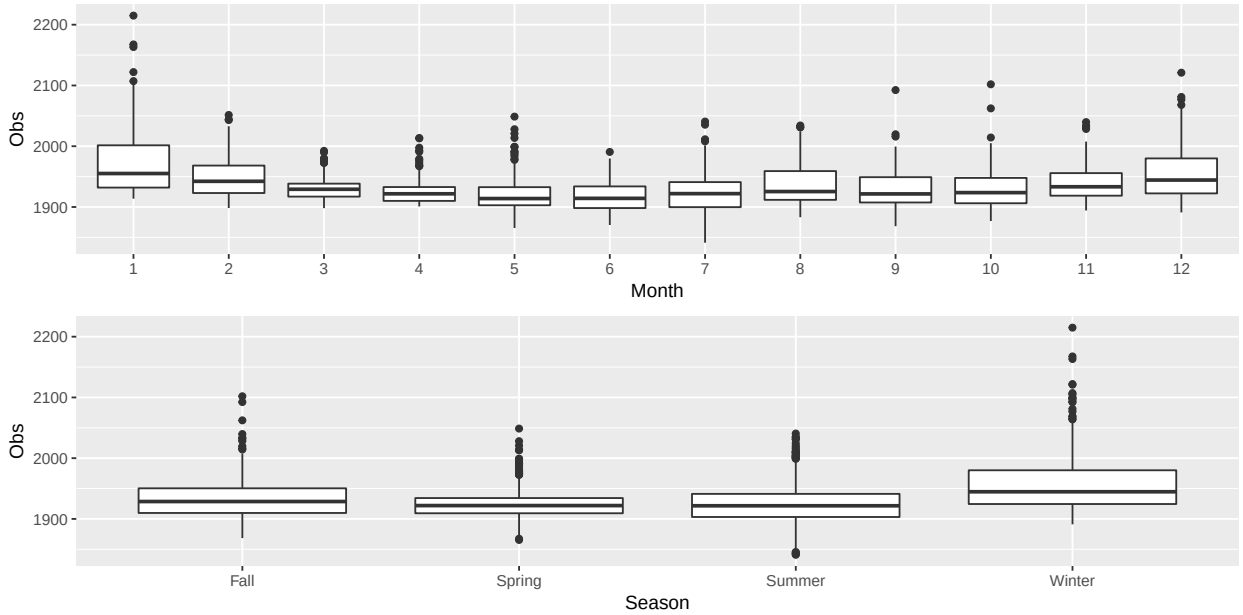
October 6, 2017

Solution 1 a)

For this problem we will do the following steps,

- We will read the COP_CH4_Obs_Mod_Bg_Sep2012-Aug2013.csv which gives us the observation of CH4 from September 2012 to August 2013.
- We created season column, and we assigned the data as per below categories based on Month values.
  - Months value of 9,10,11 = 'Fall'
  - Months value of 3,4,5 = 'Spring'
  - Months value of 6,7,8 = 'Summer'
  - Months value of 12,1,2 = 'Winter'
- Subseting the data for 16-21 UTC
  - Hr =[16-21]
- We replaced the N/A values in the data by the mean of the data at the same Hour.

Now we have the data required for this problem, the data is formatted as season wise, and month wise. Let us try to create a box plot to visualise the distribution of the data season wise and month wise,



Let us have the summary statistic for the season wise observation of CH4,

```
FALSE # A tibble: 4 x 3
FALSE   Season     mean       sd
FALSE    <chr>    <dbl>    <dbl>
FALSE 1   Fall 1932.878 30.77271
FALSE 2 Spring 1925.381 23.94319
```

```
FALSE 3 Summer 1925.150 34.09360
FALSE 4 Winter 1959.179 46.73533
```

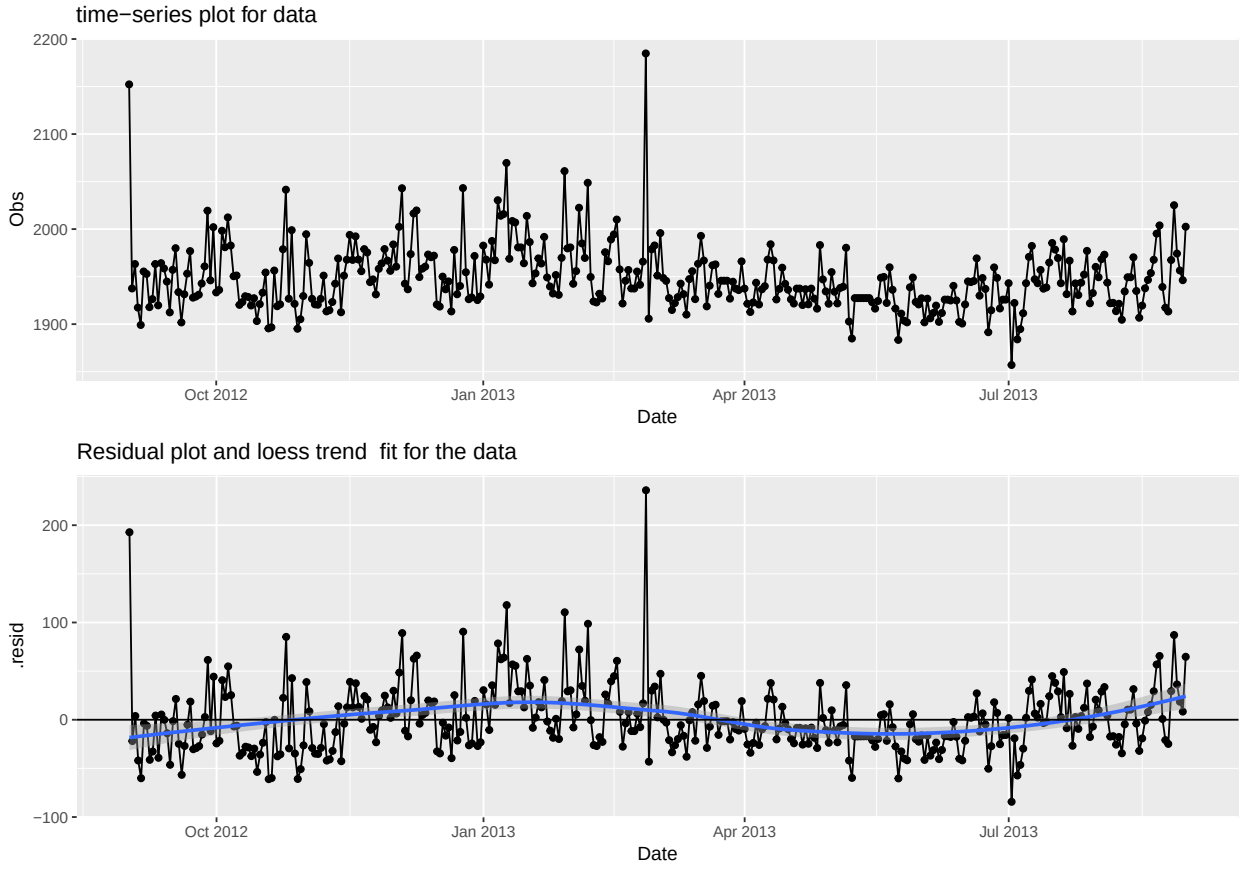
From the above plot of distribution of observation month wise we have the following findings,

- The observation of CH4 in the time frame (16-21) for the month of December,January,February is higher than the observation of same in other months
- The Mean of the observation of CH4 for the month of 'March','April',and 'August' is almost same.

Findings from the season-wise distribution,

- All follows non-normal right skewed distribution. It is visible that there are a number of leverage points in each season.
- The observation of CH4 is higher during 'winter' than any other season.
- The mean of Observation of CH4 is simillar for the 'Summer' and 'Spring' period.
- The Variance and Standard deviation of CH4 observation for season-wise is different.
- There are more outliers for 'Fall' and 'Winter' season compared to others. Among the 'Fall' and 'Winter', Winter data has more outliers resulting in high variance.

Solution 1b)



There is no particular trend from the loess curve, though there appears that there might be seasonality in the data. There are less oscillation(1cycle) due to only one year data. There are two peaks which look like outliers in the data. Most of the variation is accounted by the residuals concluding that time series model

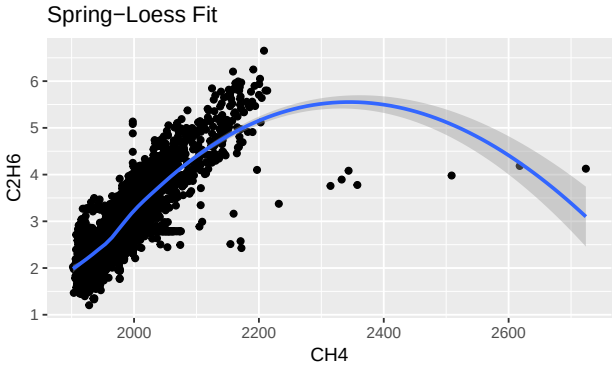
performs very poorly and probably there are other factors that are needed to explain the variation in the given data.

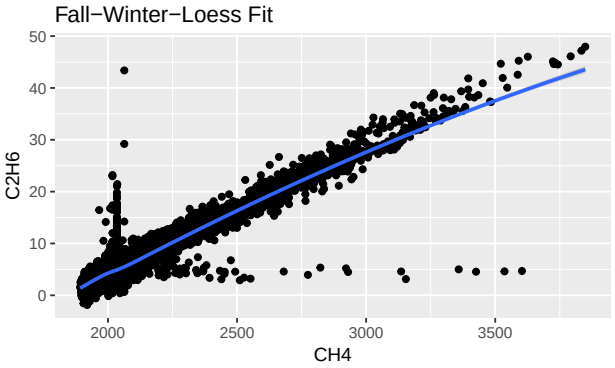
Solution 2)

For the problem of fitting linear fit in Spring and fall-winter observation of CH4,C2H6 we will consider the following data,

- BU C2H6 CH4 5min May-Jun2014.csv, for the Spring season observation of CH4 and C2H6.
- BU C2H6 CH4 5min Oct2012-Jan2013.csv, for the Fall-Winter season observation of CH4 and C2H6.
- The data values which has N/A are replaced by the mean of the data for the Month of the 'N/A' value.

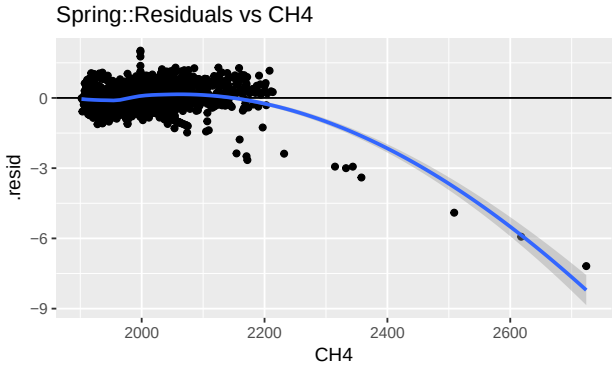
Lets explore scatterplot for both spring and fall-winter of CH4 and C2H6 observations using loess smoothing.

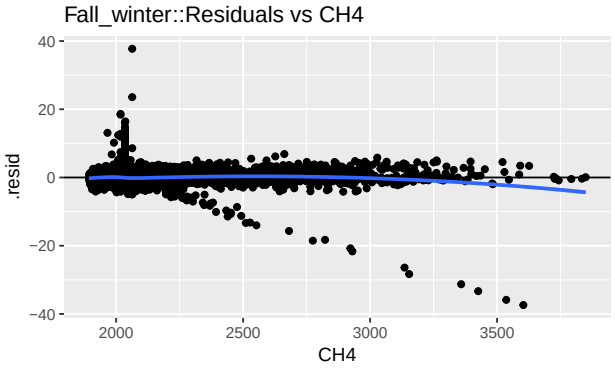




From the above plot we get to know the following details,

- Although it seems Linear Model may fits very well for Fall-Winter, there are too many points which clearly don't fit, maybe outliers. For Spring outliers are weighing too much and we can clearly see the curve.
- They cannot be simply removed as we don't know about their correctness/validity.





The following are the observation from the above plot,

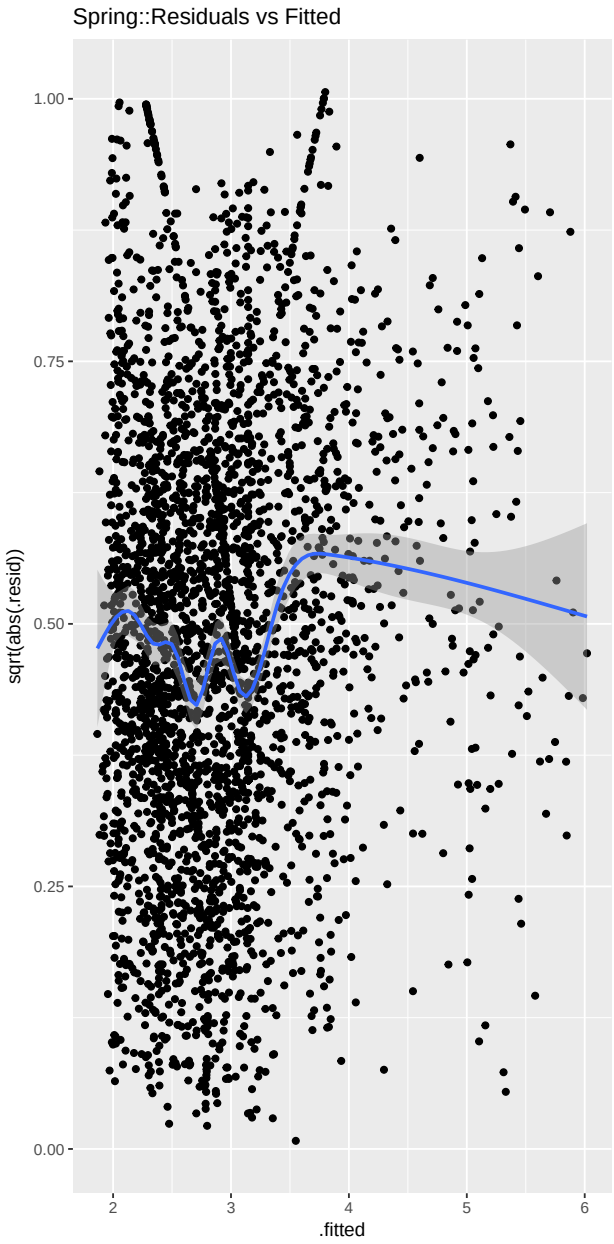
- There is clear curve in the residuals for spring and linear model is not good fit,whereas for fall-winter data, the linear model may fit well if the outlier are taken care of.

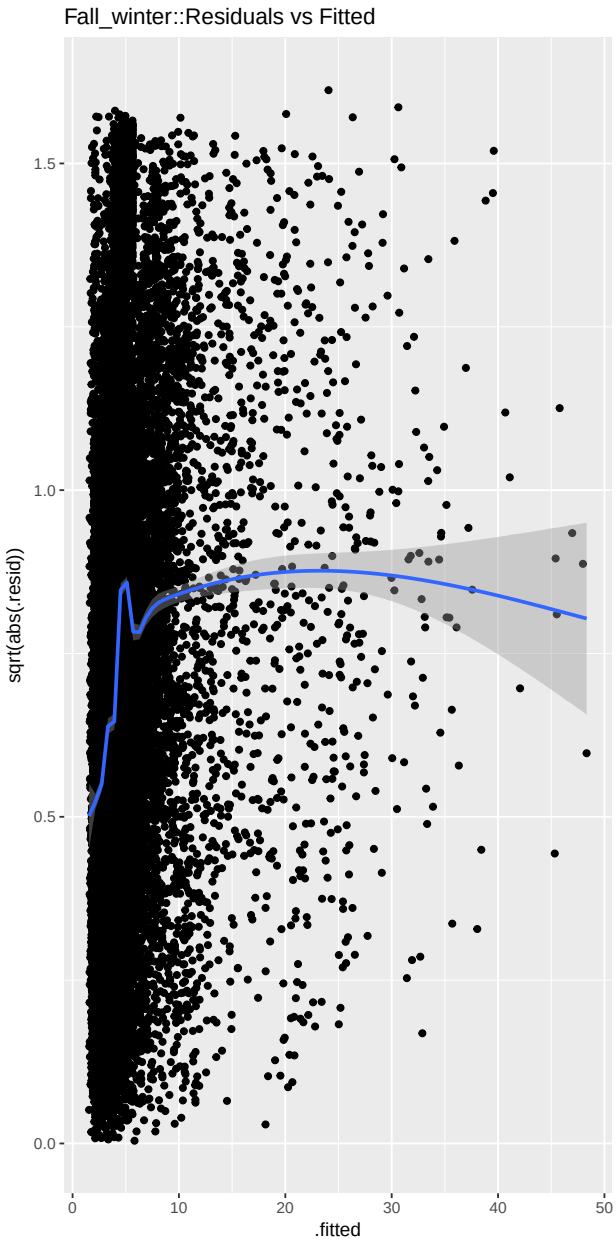
Lets remove the outlier,

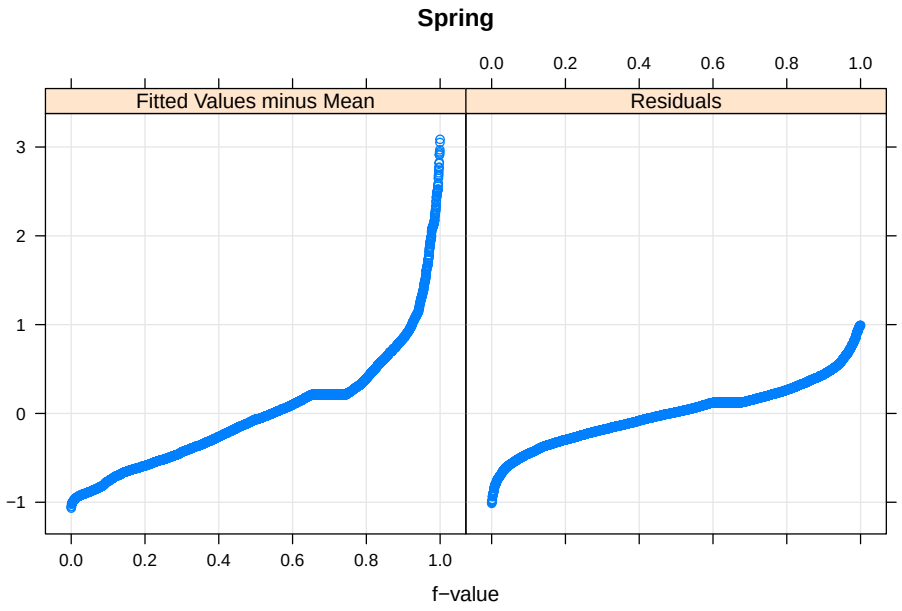
In both datasets we removed the extreme points(outliers) as they were very stronlgy affecting the loess curve.

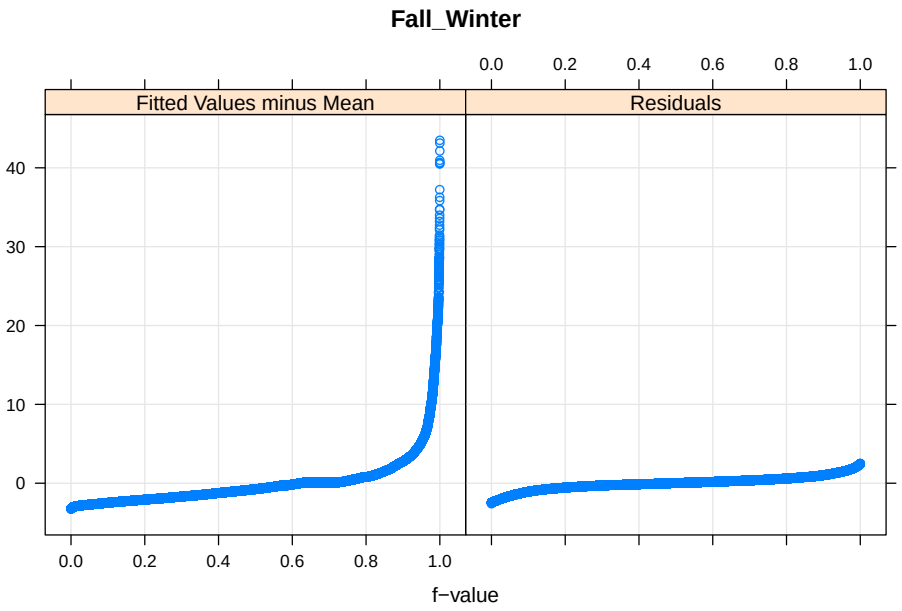
The scatterplot for new data and qq_normal plots for checking normality of residuals shows clear linear-fit for the Fall-Winter Data compared to little curve in Spring Data and qq-plots of the residuals in both the cases reasonably follow a normal distribution though not as perfect. Also we do not need any variable transformations.

Consider the residual vs fitted plots shown below, In case of fall-winter, fiited value dispersion is higher against residuals and indicates better explanation of variance by CH4 compared to Spring.


Spring::Residuals vs Fitted
sqrt(abs(.resid))
1.00
0.75
0.50
0.25
0.00
2
3
4
5
6
.fitted


Fall_winter::Residuals vs Fitted
sqrt(abs(.resid))
1.5
1.0
0.5
0.0
0
10
20
30
40
50
.fitted





Above plots indicates that in the case of fall-winter data the linear fit seems to be better. Model for fall-winter seems to be a very tight fit. The model for spring does not show as much variation explanation as the winter model. But Still the above models are build based on assumption of large number of outlier which requires to be verified from authors of paper.