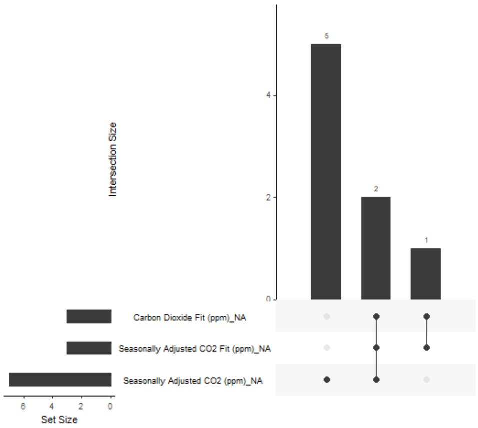
Stage 1:

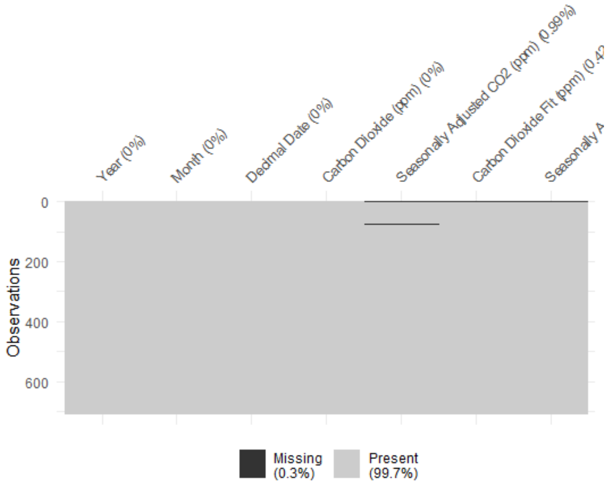
In this documentation I will review a dataset named “Carbon Dioxide Levels in Atmosphere”. This dataset consists of numerical values and is publicly available by following link https://www.kaggle.com/ucsandiego/carbon-dioxide. About the dataset itself: it consists of several columns: Year, Months, Decimal date, Carbon Dioxide (ppm), Carbon Dioxide Fit (ppm), Seasonally Adjusted CO2 (ppm), Seasonally Adjusted CO2 Fit (ppm). Here are descriptions of columns starting from 4th one:

- Column 4: Monthly CO2 concentrations in parts per million (ppm) measured on the 08A calibration scale and collected at 24:00 hours on the fifteenth of each month.
- Column 5: The fifth column provides the same data after a seasonal adjustment, which involves subtracting from the data a 4-harmonic fit with a linear gain factor to remove the seasonal cycle from carbon dioxide measurements
- Column 6: The sixth column provides the data with noise removed, generated from a stiff cubic spline function plus 4-harmonic functions with linear gain
- Column 7: The seventh column is the same data with the seasonal cycle removed.

There is an interesting aspect regarding these columns, in case of their correlations and other observations, which of them I will describe about in the next stages.

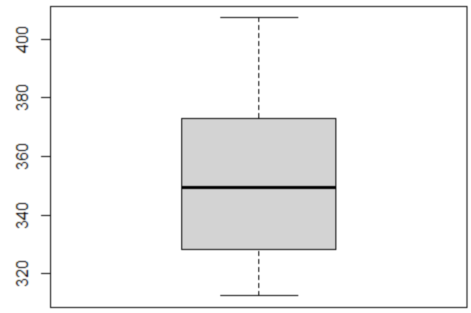
In overall, my training set has 701 rows of data in it, while test set has 421 rows. First of all, I have decided to look out for missing values in my data:



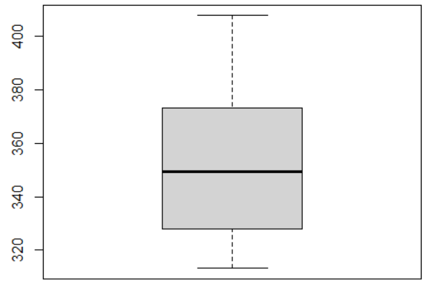


By using *vis_miss(archive) and gg_miss_upset(archive)* commands, I was able to visualize where my missing values are located so I can successfully erase them from my dataset.

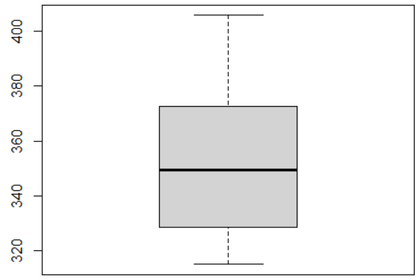
Another important thing is to check the dataset for the outliers or where they may occur, that is why I need help of boxplot function:



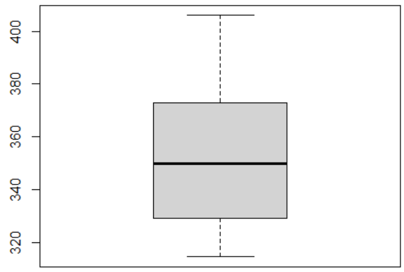
Boxplot for Carbon Dioxide Fit (ppm)



Boxplot for Carbon Dioxide Fit (ppm)



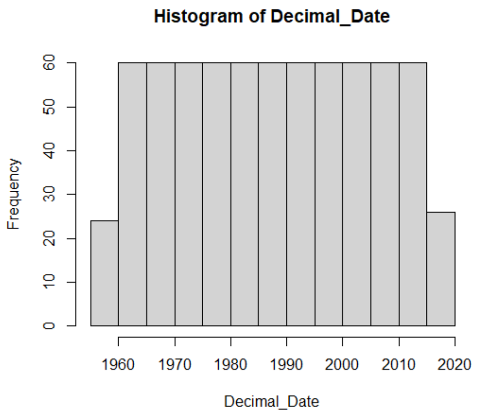
Boxplot for Seasonally Adjusted CO2 (ppm)

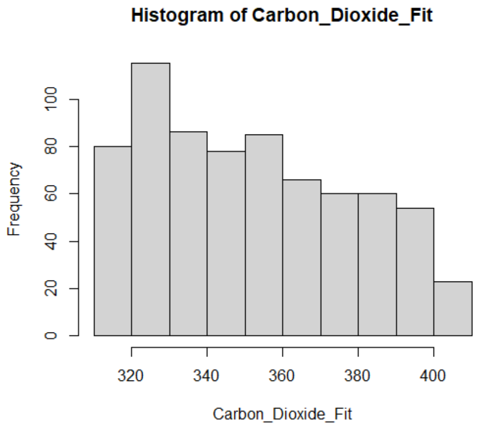


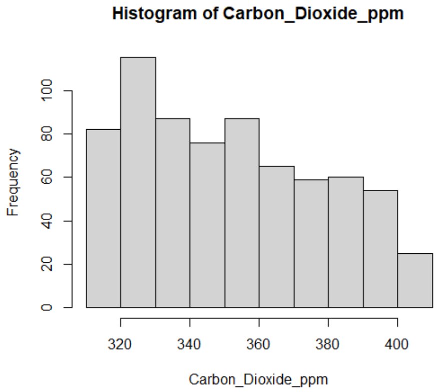
Boxplot for Seasonally Adjusted CO2 fit (ppm)

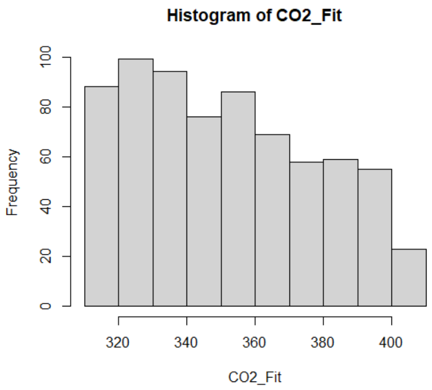
Thus, we can see that no outliers were detected.

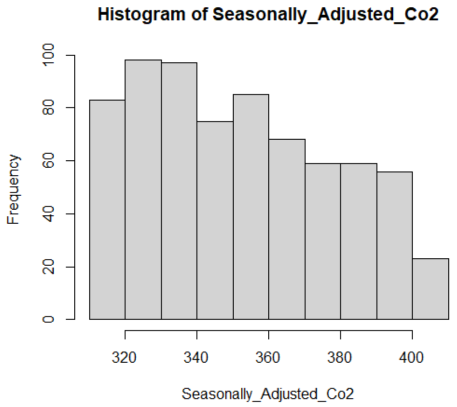
After getting rid of missing values and verifying the existence of outliers, it is time to look into the data itself, for this purpose I will use histograms:





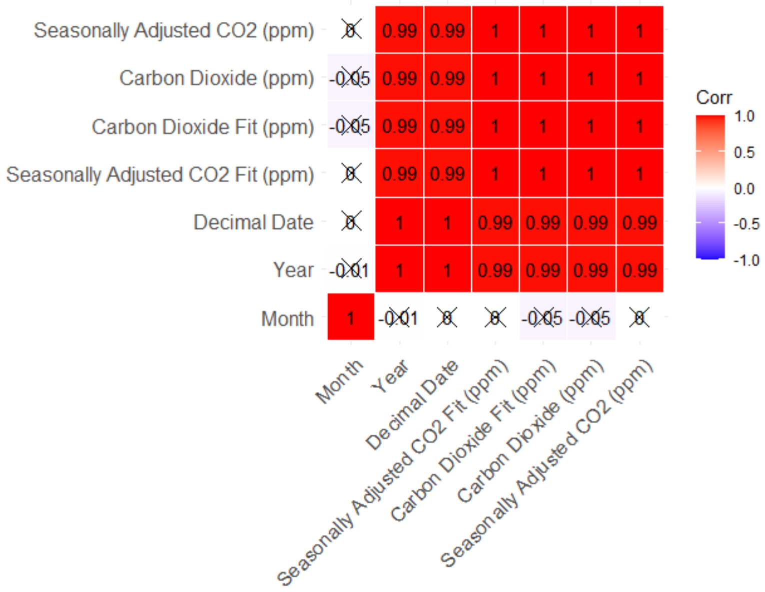






Here are the histograms of Date, Carbon Dioxide Fit (ppm), Carbon Dioxide (ppm), Seasonally Adjusted CO2 (ppm), Seasonally Adjusted CO2 Fit (ppm), where we can clearly see with which frequency their values are encountered. As it can be mentioned, most of the data was collected from 1965- 2015 years, while values of both Carbon dioxide and Seasonally Adjusted values have higher frequency in range 320-340 (ppm) or 320-360 (ppm).

Stage 2:



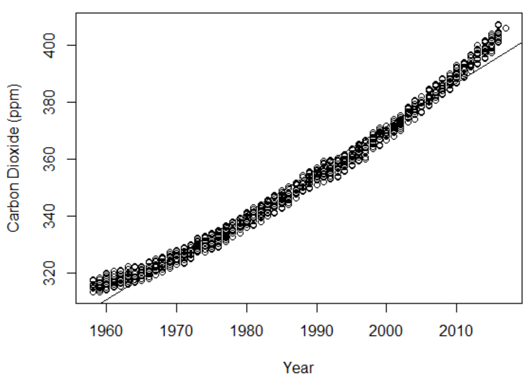
Here is depiction of correlation matrix of the dataset. As I have mentioned before, this dataset has incredible correlation values in it. After I have read an explanation of the dataset columns from 4-7 and looking upon correlation matrix above, I have decided mostly to use 2 main connections of data: Year – Carbon Dioxide (ppm) ; Year – Seasonally Adjusted CO2 (ppm), in order not to have almost identical values in other connections like Seasonally Adjusted CO2 (ppm) - Carbon Dioxide (ppm) and etc. where correlation equals to 1.

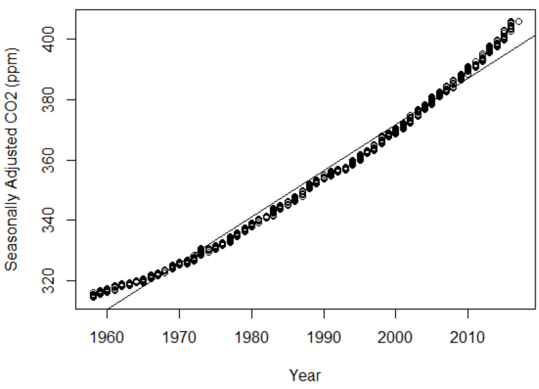
This table describes mean, median and mode values for columns 4-7 including Year column at the end. Table below represents Range, Interquartile Range, Variance and variance for the same columns. If we make reference to histograms depicted above, it is possible to see that values presented in the table have a tendency to be the truth.

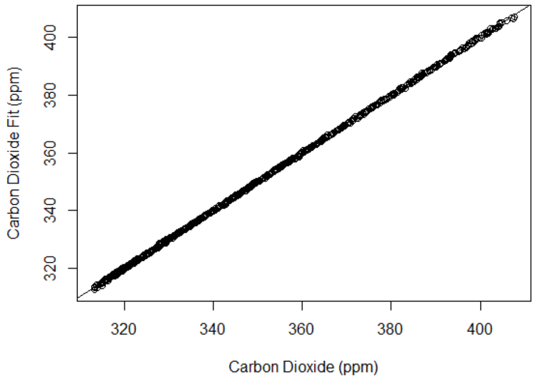
	MEAN	*MEDIAN*	*MODE*
Carbon Dioxide	352.2983	349.675	357.16
Carbon Dioxide Fit	352.2934	349.875	315.71
Seasonally Adjusted Co2	352.299	349.74	319.42
Seasonally Adjusted CO2 Fit	352.2941	349.825	314.89
Year	1987.306	1987	1959

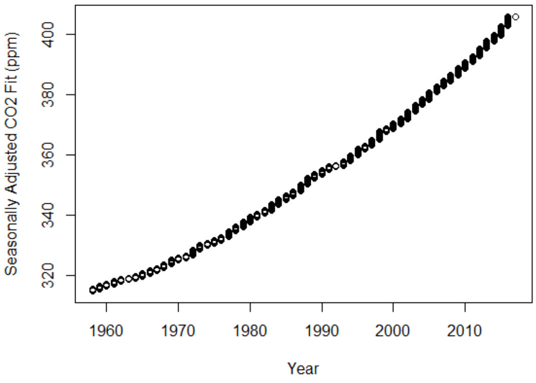
	Range	*Interquartile Range*	*Variance*	*Standard Deviation*
Carbon Dioxide	313.21 - 407.65	44.5075	685.4122	26.18038
Carbon Dioxide Fit	312.48 - 407.28	44.6825	685.3308	26.17882
Seasonally Adjusted Co2	314.42 406.04	43.82	681.8381	26.11203
Seasonally Adjusted CO2 Fit	314.89 - 405.83	43.6975	681.7428	26.11021
Year	1958-2017	29	287.023	16.94175

At this point, it is needed to depict the Dependent and Independent variables with the help of plot functions with line:



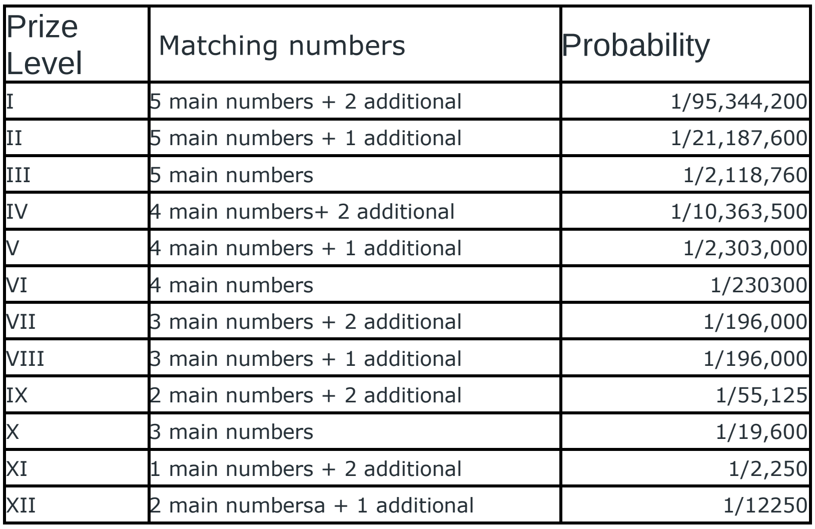






After analyzing these plots, I have decided to make “Year” an independent and “Carbon Dioxide (ppm), Seasonally Adjusted CO2 (ppm) dependent variables and exclude columns with “Fit”, because the difference between them will be too miniscule in order to have the difference even visually, by looking at future plots with predictions.

Stage 4:



Prize Level	Matching numbers	Probability
I	5 main numbers + 2 additional	1/95,344,200
II	5 main numbers + 1 additional	1/21,187,600
III	5 main numbers	1/2,118,760
IV	4 main numbers+ 2 additional	1/10,363,500
V	4 main numbers + 1 additional	1/2,303,000
VI	4 main numbers	1/230300
VII	3 main numbers + 2 additional	1/196,000
VIII	3 main numbers + 1 additional	1/196,000
IX	2 main numbers + 2 additional	1/55,125
X	3 main numbers	1/19,600
XI	1 main numbers + 2 additional	1/2,250
XII	2 main numbersa + 1 additional	1/12250

For this probability theory tasks, I have chosen EuroJackpot Loto, where there are 50 main and 10 additional numbers.

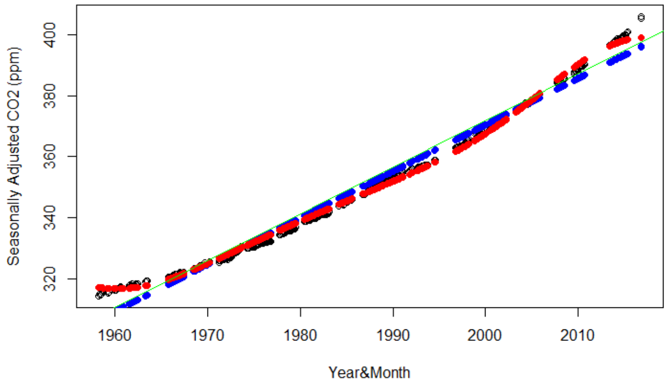
In order to calculate probability in case of only main numbers it is enough to use this formula, where n-(50) and k-(number of matched numbers)

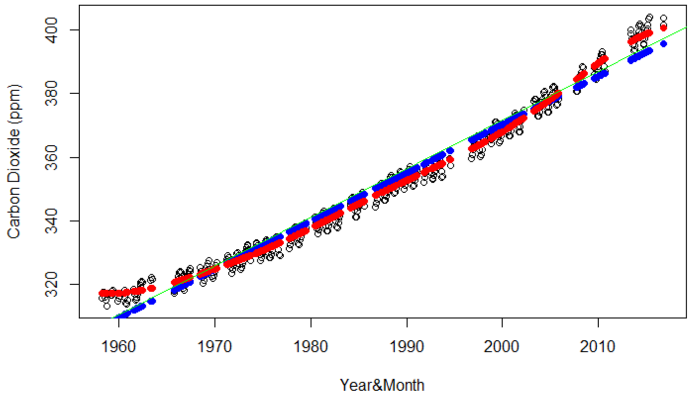
$$C_n^k = \frac{n!}{(n-k)!\,k!}$$

In cases where additional numbers are included, the same formula is used, where n-(10) and k-(1 or 2) and then we multiple the result we have with result we gained from calculating main matched numbers.

Stage 5:

As it was mentioned at the beginning, testing set consist of 421 rows of data and I have represented prediction of linear and Support Vector Machine predictions, with relationships of: Year-Carbon Dioxide (ppm) and Year-Seasonally Adjusted CO2 (ppm), where red line represents LM prediction, blue – SVM prediction, green line – regression line of training set and dots show us the representation of testing set values. Moreover, value “Year” does represent not the “Year” column but rather “Decimal Date” columns value, because it represents both Year and Month simultaneously in one value, which makes an observation more logical.At the end, using the MAPE function I have calculated the mean difference between original and predicted values of testing set, which is fairly to say, quite small after all.





	Year&Carbon Dioxide (ppm)	*Year&Seasonally Adjusted CO2 (ppm)*
lm	0.84%	0.70%
svm	053%	0.28%