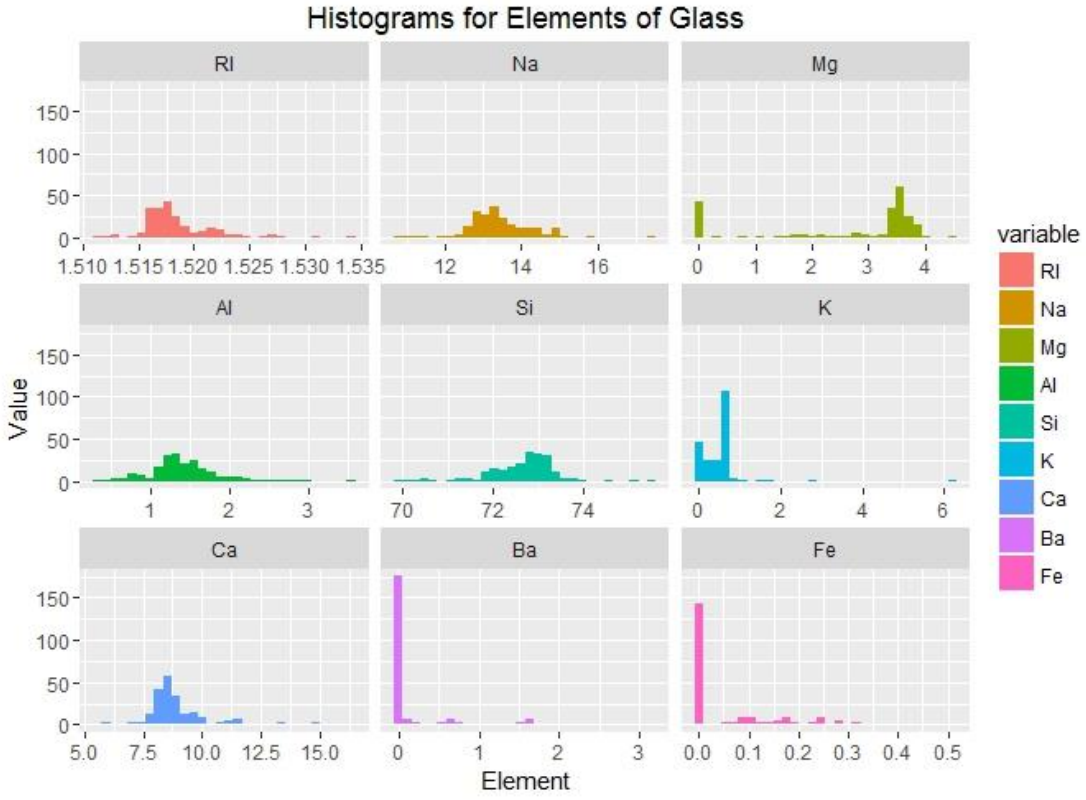
DSA/ISE 5103 Intelligent Data Analytics
Assignment #3
Name: Jiwon Jeon
Date: 10/07/2016

1 Glass Identification

(a) The Glass data can be seen with histogram to explore their distribution;



The skewness of each predictor (RI and elements) is tabulated as below;

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
skewness	1.62543	0.45418	-1.15255	0.90728	-0.73044	6.55164	2.04705	3.41642	1.75432

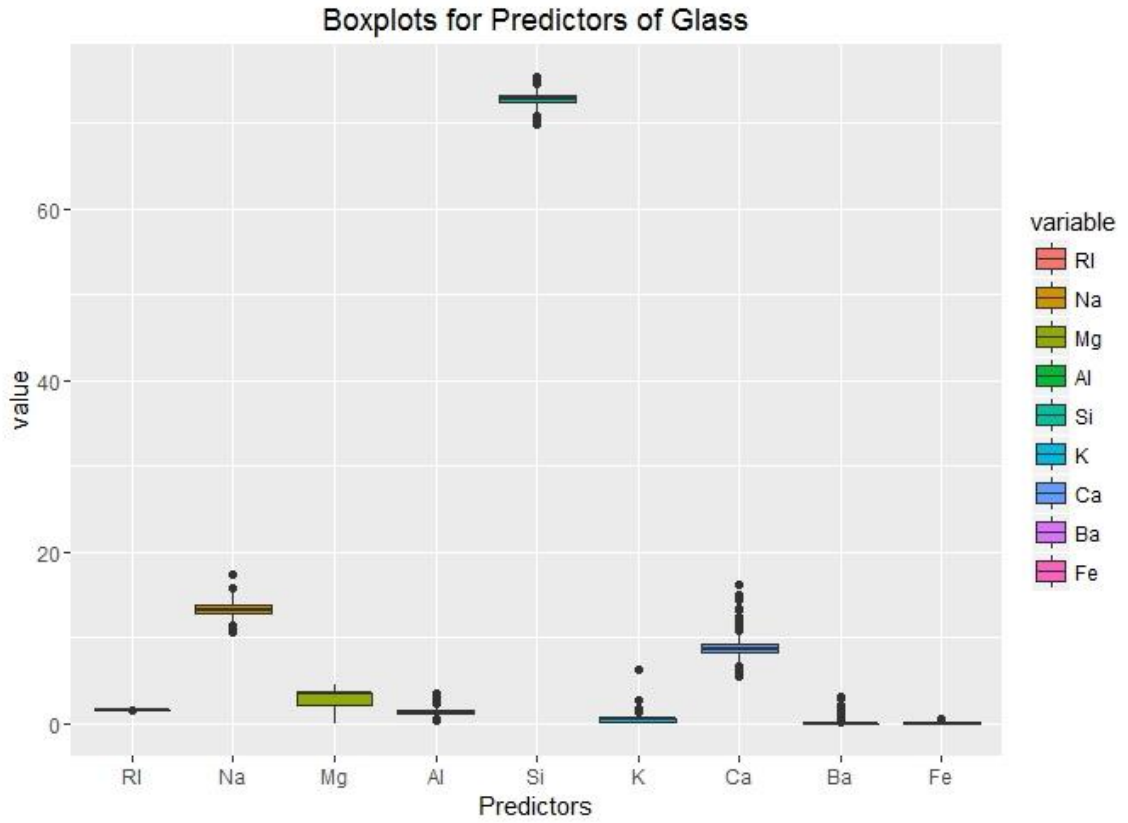
From the histograms and the skewness values, it can be said that Na and Si are fairly normal with small positive and negative skewness, respectively. RI, Al and Ca are slightly skewed to the positive direction. Mg shows somewhat low value in skewness, however, it does not show a normal distribution even though the dataset looks symmetric. Compared to others, K, Ba and Fe show highly skewed distribution.

In order to detect outliers in each predictor, boxplots are drawn in the next page. The number of outliers in each predictor is as below;

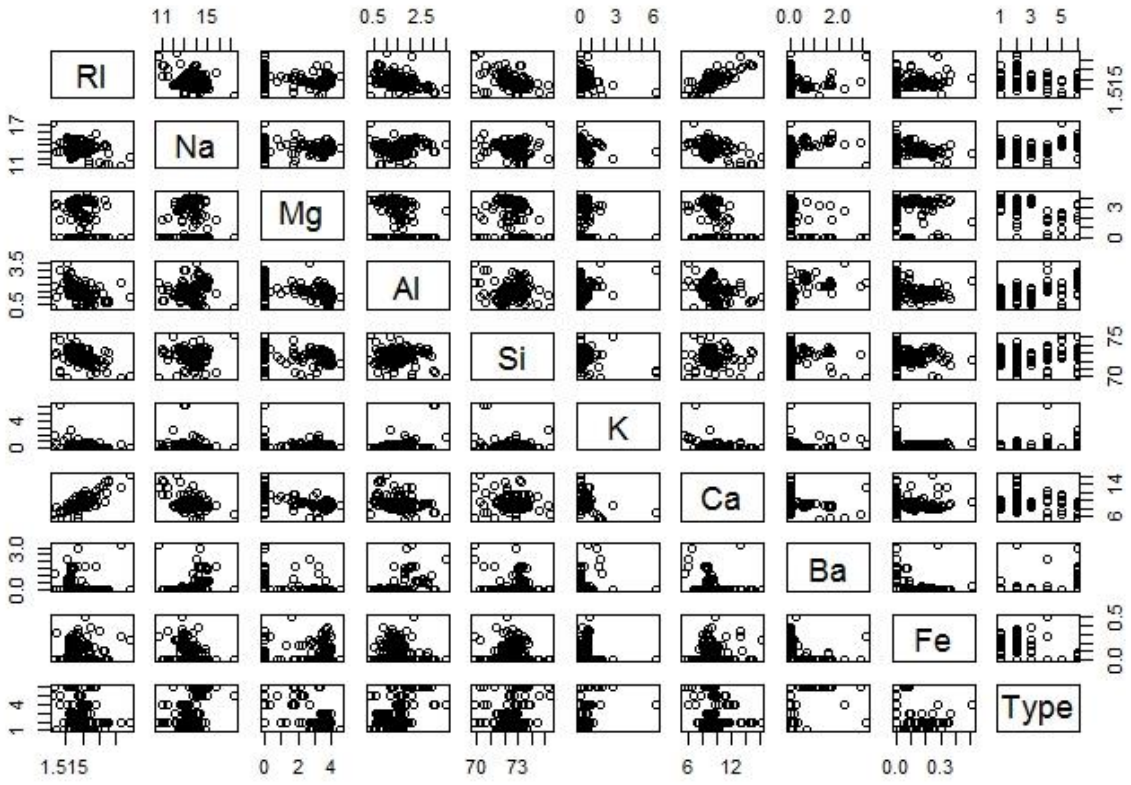
	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
number	17	7	0	18	12	7	26	38	12

Some of the outliers are tabulated as below;

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
outliers	1.52667	11.45		0.29	70.57	1.68	11.64	0.09	0.26
	1.5232	10.73		0.47	69.81	6.21	10.79	0.11	0.3
	1.51215	11.23		0.47	70.16	6.21	13.24	0.69	0.31
	1.52725	11.02		0.51	74.45	1.76	13.3	0.14	0.32
	1.5241	11.03		3.5	69.89	1.46	16.19	0.11	0.34
	…	…	…	…	…	…	…	…	…



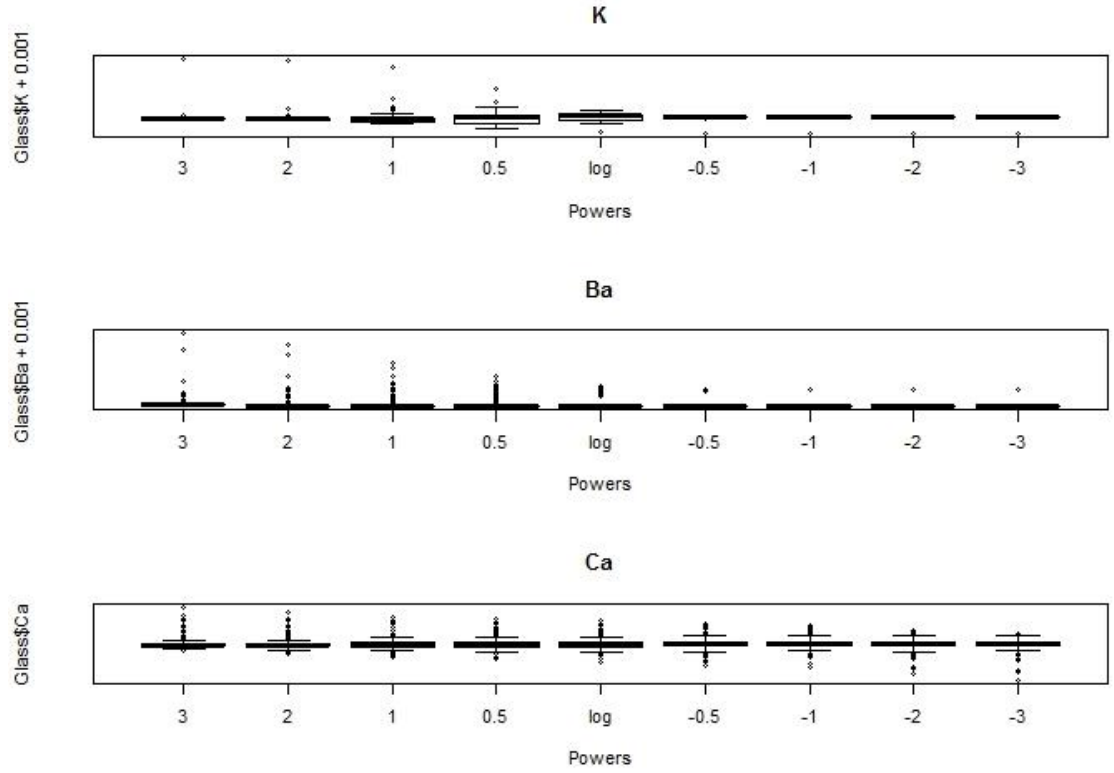
The relationships between between the predictors can be described by scatter plots. From the plots below, we can see that RI and Ca shows a great linear relationship while the others show quite random relations to each other;



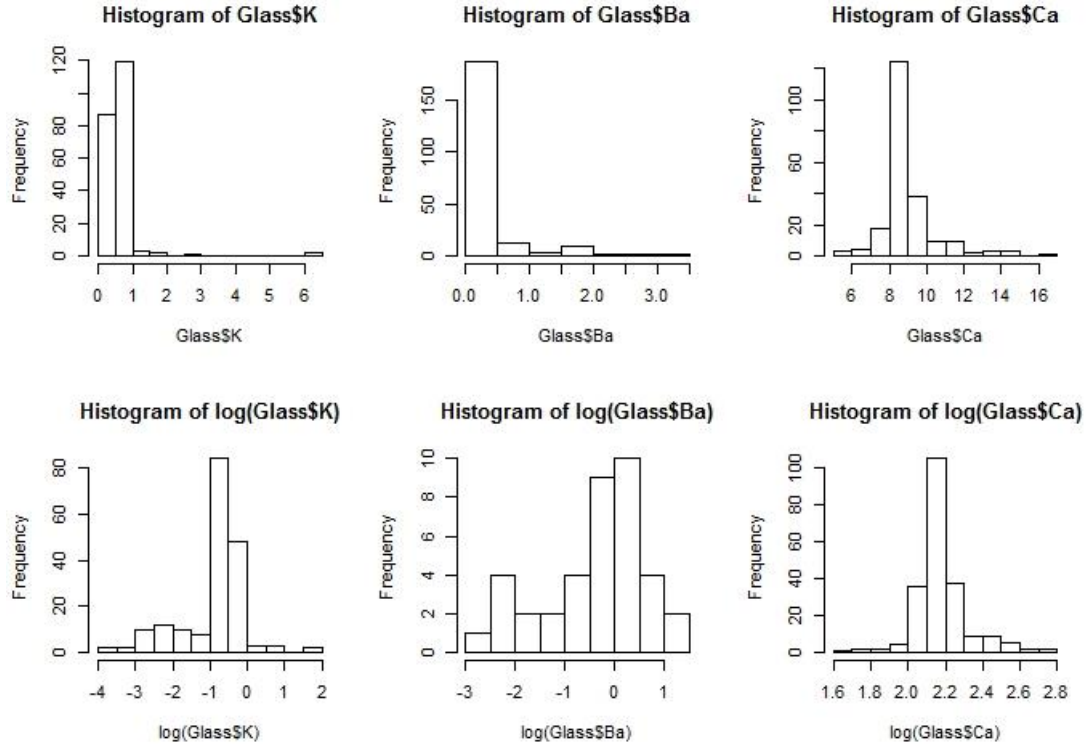
(b) From the skewness calculation, three elements **K, Ba and Ca** are selected for a skew transformation.

i. symbox function

The result of symbox with lambda = -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3, for each element is as below;

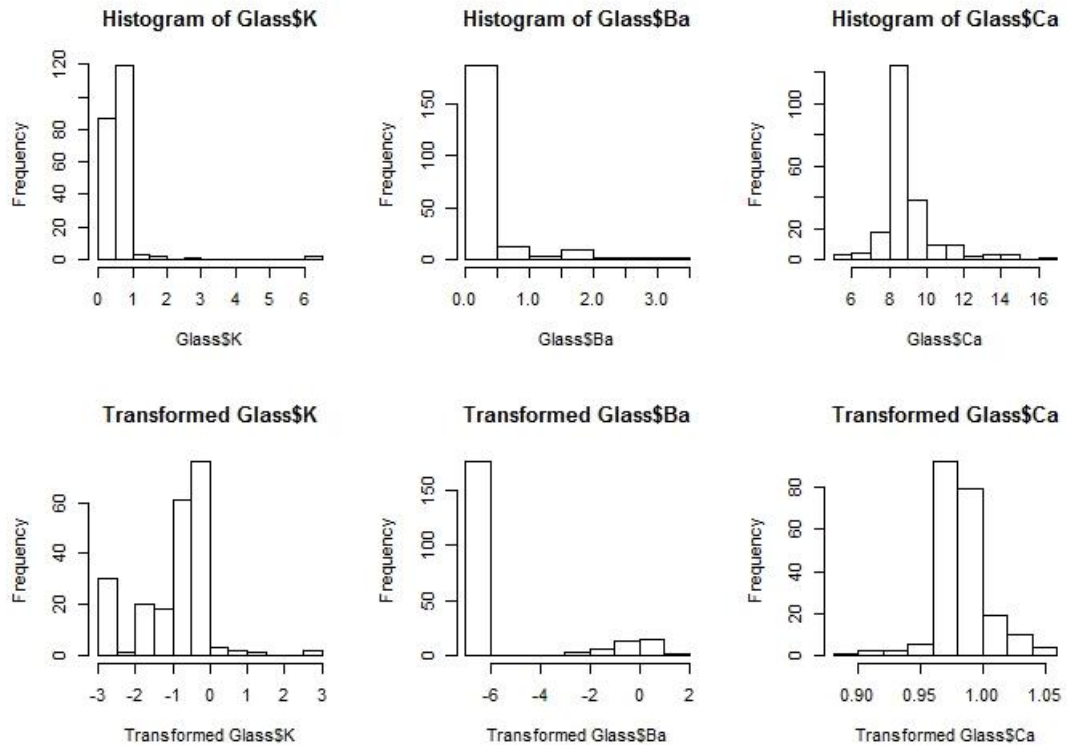


For each predictor, the log transformation is chosen to display the transformed distribution using histogram as it converted the original data to normal distribution most. Ca data actually showed the least skewness value when the lambda = -0.5, however, only log transformed distribution is shown here. (See R script for the transformed distribution with lambda = -0.5.)



ii. boxcox function

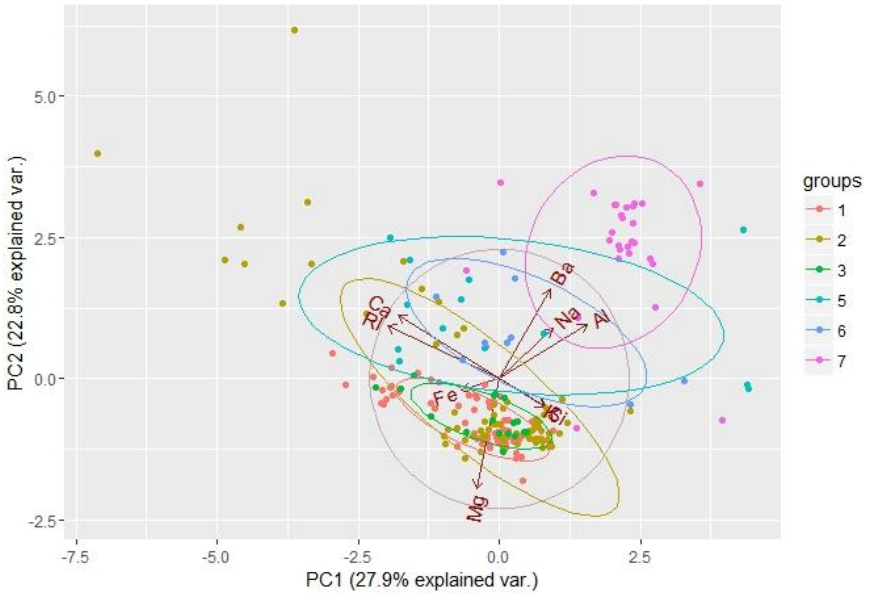
Using boxcox command, the lambda values were calculated 0.374133, 0.1563629, and -0.8593591 for K, Ba, and Ca, respectively. ;



Compared to K and Ca, Ba does not show a good transformation to normal distribution using boxcox function.

(c) The components of PCA object are as below;

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
RI	-0.5451766	0.28568318	-0.0869108293	-0.14738099	0.073542700	-0.11528772	-0.08186724	-0.75221590	-0.02573194
Na	0.2581256	0.27035007	0.3849196197	-0.49124204	-0.153683304	0.55811757	-0.14858006	-0.12769315	0.31193718
Mg	-0.1108810	-0.59355826	-0.0084179590	-0.37878577	-0.123509124	-0.30818598	0.20604537	-0.07689061	0.57727335
Al	0.4287086	0.29521154	-0.3292371183	0.13750592	-0.014108879	0.01885731	0.69923557	-0.27444105	0.19222686
Si	0.2288364	-0.15509891	0.4587088382	0.65253771	-0.008500117	-0.08609797	-0.21606658	-0.37992298	0.29807321
K	0.2193440	-0.15397013	-0.6625741197	0.03853544	0.307039842	0.24363237	-0.50412141	-0.10981168	0.26050863
Ca	-0.4923061	0.34537980	0.0009847321	0.27644322	0.188187742	0.14866937	0.09913463	0.39870468	0.57932321
Ba	0.2503751	0.48470218	-0.0740547309	-0.13317545	-0.251334261	-0.65721884	-0.35178255	0.14493235	0.19822820
Fe	-0.1858415	-0.06203879	-0.2844505524	0.23049202	-0.873264047	0.24304431	-0.07372136	-0.01627141	0.01466944



Principal component 1 (PC1) is highly affected by RI and Ca followed by Al as shown in the rotation matrix and ggbiplot. Meanwhile, principal component 2 (PC2) is mostly affected by Mg and Ba. PC1 and PC2 cover 27.9% and 22.8% of the total variability (variance) in the data, respectively.
The PCA distinguishes the data points of Type 7 fairly well from the other types, however, it is hard to recognize the differences in Type 1, 2 and 3 on PC1-PC2 panel using their projections as they are somewhat spread.

(d) The coefficients of LDA are calculated as below;

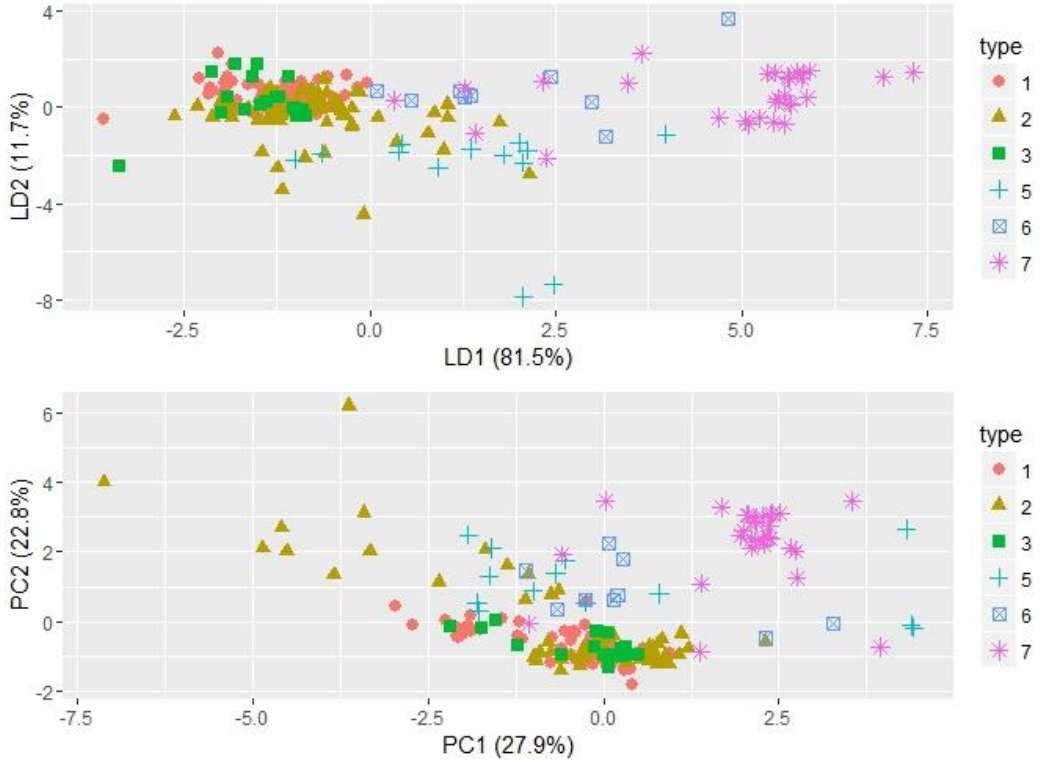
	LD1	LD2	LD3	LD4	LD5
RI	311.6912516	29.3910394	356.0188308	246.85720802	-804.6553938
Na	2.3812158	3.1650800	0.4596785	6.92435141	2.3987509
Mg	0.7403818	2.9858720	1.5728838	6.84983896	2.8002951
Al	3.3377416	1.7247396	2.2024668	6.41923638	0.9371345
Si	2.4516520	3.0063507	1.7026191	7.54220302	0.9562989
K	1.5714954	1.8620159	1.2861127	8.07611300	2.8209927
Ca	1.0063101	2.3729126	0.6475200	6.69663574	3.7110859
Ba	2.3140953	3.4431987	2.5964981	6.43849270	4.4077058
Fe	-0.5114573	0.2166388	1.2026071	-0.04474935	-1.3029207

The proportion of trace is as below;

LD1	LD2	LD3	LD4	LD5
0.8145	0.1169	0.0413	0.0163	0.0111

Linear discriminant 1 (LD1) explains 81.45% of the between-group variance in the data and LD2 covers 11.69% of them.

The comparison between LDA and PCA;



Both PCA and LDA reduce the dimensions of the data, which is useful for visualization. There is a difference between these two techniques when it comes to classification; PCA is used as an unsupervised learning technique where no class information is given while LDA is a supervised technique where pre-defined classes are referred to. Therefore LDA would provide better data classification than PCA. As shown in the comparison plots below, the LD1-LD2 panel (upper panel) let the data to be more clustered per types, especially Type 1, 2 and 3, compared to the PC1-PC2 panel (lower panel). This is expected since PCA tried to retain most of the variability in the data while LDA tries to retain most of the between-class variance in the data resulting in more powerful classification.

2 Missing Data

(a) When the listwise deletion method is used for the regression analysis, the multiple R-squared value is 0.9311. The regression model fits the data very well after the listwise deletion method, however, this high fitness means low predictivity.

(b) When the mean imputation method is used for the regression analysis, the multiple R-squared value is 0.6379. The regression model using the mean imputation method does not fit the imputed data as good as the listwise deletion method does, however the predictivity will be better in this case.

(c) For the regression analysis after using multiple imputation with MICE, 'sample' method is selected to impute values into missing data in the mice function (plus, m = 5 and maximum iteration = 10). After the regression analysis and recombination, the multiple R-squared values are 0.4214, 0.3957, 0.4882, 0.4229 and 0.439 for each imputation. These values might change whenever the method is run in MICE because MICE has a randomness when handling data. The fitness of the regression model using multiple imputation with MICE is worse than those of listwise deletion and mean imputation, however, the predictivity will be better.

(d) In order to see the result when single imputations specific to each attribute were applied, randomly selected methods for each variable are deployed in mice function. In this solution, the methods are 'sample', 'rf', 'sample', 'sample', 'rf', 'rf', 'cart', 'cart', 'sample' and 'rf' for year, country, tariff, polity, pop, gdp.pc, intresmi, signed, fiveop and usheg, respectively. The multiple R-squared values are 0.4155, 0.4798, 0.4429, 0.4395 and 0.4571. These values might change whenever the method is run in MICE. The values are similar with those from multiple imputation. The data seem to be barely improved.

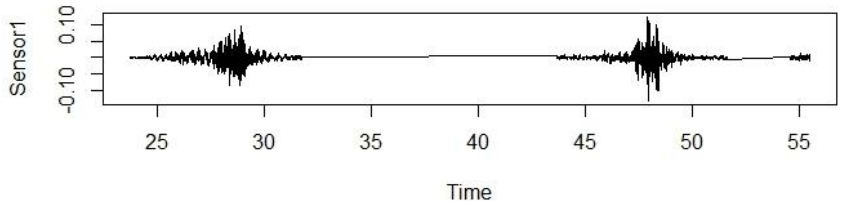
(e) The regression coefficients from (a) to (d) are as below;

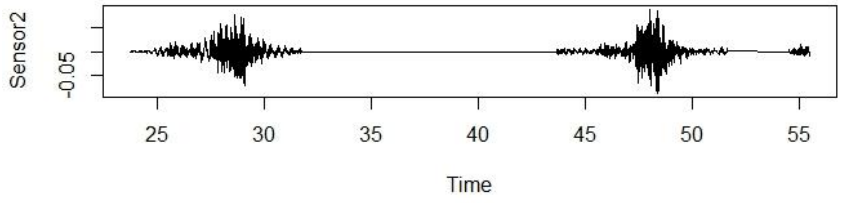
	Listwise Deletion	Mean Imputation	MI (pooled est)	SI's (pooled est)
(Intercept)	-2.650433e+02	1.633387e+03	1.764092e+03	1.740888e+03
year	3.580765e-01	-7.938926e-01	-8.710583e-01	-8.599067e-01
countryIndonesia	-1.900660e+02	-4.620179e+01	-4.720197e+01	-4.183327e+01
countryKorea	-2.254931e+02	-5.937894e+01	-6.248873e+01	-5.479903e+01
countryMalaysia	-2.318437e+02	-5.531281e+01	-5.784582e+01	-5.084796e+01
countryNepal	-2.270878e+02	-4.631776e+01	-4.739938e+01	-3.208214e+01
countryPakistan	-1.616933e+02	-1.440892e+01	-1.860809e+01	-6.192419e+00
countryPhilippines	-2.103454e+02	-5.033981e+01	-5.361710e+01	-4.140893e+01
countrySriLanka	-2.168838e+02	-4.536998e+01	-4.911930e+01	-3.471606e+01
countryThailand	-2.014832e+02	-4.141807e+01	-4.495145e+01	-3.284044e+01
polity	-1.902494e-01	-2.111236e-01	-1.043669e-01	-2.480084e-01
pop	-2.111286e-07	-2.628999e-08	-3.039428e-08	-1.403812e-08
gdp.pc	2.910265e-04	5.922484e-04	6.262182e-04	1.552850e-03
intresmi	2.929493e-01	-6.674644e-01	-1.050764e+00	-6.765417e-01
signed	-1.288913e+00	2.872480e+00	3.172251e+00	4.802864e+00
fiveop	-1.579368e+01	2.254838e+00	4.337617e+00	2.808252e+00
usheg	9.582074e+00	-1.988981e+01	-1.836046e+01	4.274891e-01

(f) From the values of R-squared for each approach, we can conclude that the predictivity of the listwise deletion and mean imputation will be significantly worse than multiple imputation with MICE and single imputations with MICE. The R-squared values for MI with MICE and SI with MICE are hard to compare, however, I would recommend the multiple imputation approach as it shows slightly less value in R-squared in this case. If different methods for each variable in single imputation approach are selected, the R-squared will be changed and might provide different interpretation.

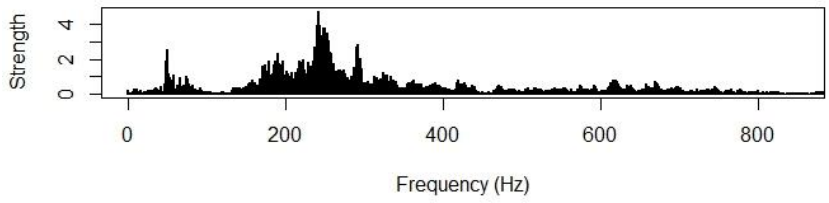
3 Truck - Bridge Sensor Data

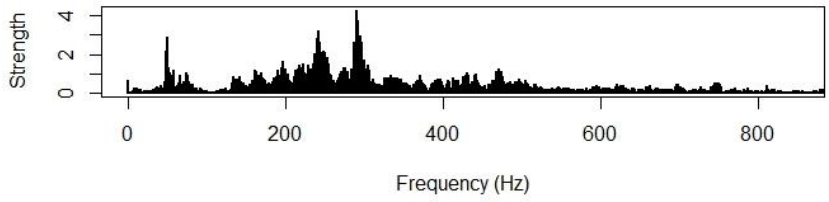
(a) By applying the fast fourier transform (FFT), the time domain can be converted into the frequency domain. The given time vs. sensor data distributions for truck type 1 (sensor 1) and type 2 (sensor 2) are as below;



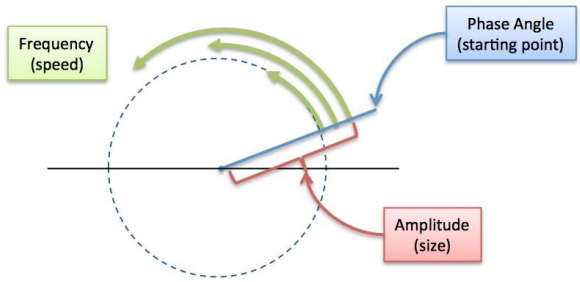


The transformed frequency vs. strength data distributions are as below;





In order to build a classification model to identify each type of trucks from the transformed frequency data (i.e. fft data), we can consider several candidate features based on the assumptions of fourier transform – every trajectory (time signal, signal) is a set of circular motions. That is, the fourier transform breaks the signal into a set of related cycles, and each cycle has a strength, a delay and a speed as shown in the diagram below;



(Source: https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/)

Therefore, the following variables can be considered as candidate features;
Cycle – a circular path of an event (a data point) containing the information
Strength – the amplitude or size of a cycle (a data point)
Delay – the starting point to start a cycle
Speed (frequency) – the vector of velocity to complete a cycle

Based on these features, the relative magnitude of mechanical information of trucks such as weight and length can also be found, which will allow us to classify the type of trucks.

(b) The following table shows the result of the feature construction;

FFT1 (Sensor 1)				FFT2 (Sensor 2)			
Cycle	Freq (speed)	Strength (size)	Delay (starting point)	Cycle	Freq (speed)	Strength (size)	Delay (starting point)
0	0.0000000	1.34e-04	0.000	0	0.0000000	4.00e-04	0.00
1	0.0005885	2.44e-05	-10.600	1	0.0005885	1.36e-05	-12.40
2	0.0011771	2.82e-05	-33.800	2	0.0011771	7.23e-06	-81.60
3	0.0017657	3.14e-05	-6.670	3	0.0017657	1.93e-05	-1.34
4	0.0023543	5.30e-05	31.100	4	0.0023543	2.35e-05	43.30
5	0.0029429	3.03e-05	-157.000	5	0.0029429	2.54e-05	-163.00

(c) It is usual to have data set not enough to construct and extract proper features to build a classification model or to analyze the nature of data. The given data set for this problem was not sufficient to find mechanical information of the vehicles passing by, therefore visualizing and identifying the differences between the types are difficult. In such case, we might be able to find any mathematical properties that characterize the differences in terms of ‘number’. As explained in Problem 3(a), we will also be able to find a proper feature extraction method to transform the data and analyze them.