Hafid Pradipta
4496585
Final Project Stat-520

Factor Analysis on Ozone Level Detection

Introduction

Ozone day is the day when weather condition mix with the pollution of high-level ozone close to the ground. Ground level ozone(O3) is not created directly but formed as the result of complex chemical reaction called Volatile Organic Compounds. Short-term exposure to elevated ambient ozone not only can cause some health problems for human but also harmful to corpses (Fan & Wei, 2008).

Texas Commission on Environmental Quality is working on how to build accurate alarm forecasting model for the Houston area (Fan & Wei, 2008). However, there are numerous challenges to construct the model such as sparse and evolving dataset, limited data size, large irrelevant features, and the true model is stochastic as a function of measurable factors. The focus of this paper is to choose the most reliable factors that determine the future of the ozone level days with unknown parameter and different distribution. 15 statistical techniques applied in a study (Schlink, et al., 2003) and the results are none of them perform better from one to others in all aspects.

The author of this analysis realizes that the model that will be conducted in the future cannot fully predict ground-level ozone accurately due to limited knowledge of the author with the highly skewed dataset. However, the author will try to predict what are the main factors that contribute to the higher probability of the ozone level.

Dataset Explanation

The dataset is taken from https://archive.ics.uci.edu/ml/datasets/ozone+level+detection. This analysis will use one-hour peak dataset with the file name *ohehr.data*. There are ten categories in this dataset, and they are measured in a different level. The total variables in this dataset are 72 variables. Here are the table describing the categories and variables.

Category	**Label**	**Information**	**Measured at**
Temperature	T		22 times and three levels
Wind speed rate	Wsr		22 times and three levels
Relative humidity	Rh	percentage of partial pressure water vapor to the equilibrium vapor pressure	3 levels
East-West direction wind	U		3 levels
North-South direction Wind	V		3 levels
Geopotential height	Ht	vertical coordinate referenced to Earth's mean sea level	3 levels
K Index	Ki	quantification of the disturbances in the horizontal component of earth’s magnetic field. >5 is geomagnetic storm	

T Total	Tt	Index to assess storm strength. <44 less likely and >56 scattered severed storms	
Sea level Pressure	Slp	the pressure within the atmosphere of earth	current and previous day
Attribute	attribute	0 when normal day and 1 when ozone day	

The three levels in the dataset reflect the difference in atmospheric pressure which is 500,700 and 850 Hectopascal or roughly 5500, 3100 and 1500 meters height respectively

This analysis will start by assessing univariate and multivariate normality, conducting principal component analysis, factor analysis, clustering and two mean sample hoteling T^2 test. Since There are 1848 observations and 72 variables in this dataset, the author decided only to include 25 variables and omit 47 variables which are temperature and windspeed rate at 23 different times. The dependent variables for this dataset are attribute which is categorical data that denote ozone day as 1 and non-ozone day as 0. The rest of variables are independent variables.

METHODOLIGAL ANALYSIS

Descriptive Statistics

Based on the boxplot of temperature and wind speed rate (Figure 1 and Figure 2) at 22 different times, the only information that can be extracted is that as the temperature increase, the windspeed will also increase. Based on the boxplot of five variables (rh,u,v, ht,t) on Figure 3, temperature, relative humidity and north-south direction wind(v) are negatively correlated with the atmospheric meanwhile east-west wind direction (u), and geopotential height is positively correlated with the atmospheric pressure.

Assessment of Normality

Based on Shapiro Wilk test of each variable (Figure 5), there is only one variable that contains partial univariate normality which is north-south direction wind at 850 hpa with p-value 0.3. Therefore, the author uses boxcox transformation to find the optimum lambda to make the distribution more normal. In result, most of the lambda doesn't give even partial normality (Figure 4) to the variable. Peak wind speed rate will have cube square root transformation because lambda is close to 0.333

The dataset does not pass univariate normality, and it will result that the dataset is not multivariate normally distributed. The result of Mardia test, Chi-square plot to the squared Mahalanobis distance (Figure 6, Figure 7, and Figure 8) and the proportion of the contour give the same result. Therefore, the dataset is not normally distributed.

Correlation Matrix

The correlation matrix is quite hard to interpret but several takeouts from this standardized correlation matrix are geopotential height in general positively correlated with temperature meanwhile temperature, and geopotential height are negatively correlated with east-west wind direction. Relative humidity is positively correlated with the index to storm strength. However, correlation does not mean causation. Therefore, the author will not conclude anything based on this analysis.

Principal Component Analysis

Based on the scree plot of principal component analysis(Figure 10), the first component can explain 32.4% total variation in the dataset followed by 19.6%, 9%, and 7% for second, third and fourth factor respectively. Total variation explained by four components is 68%. Because adding more variable does not add significant total variance, this analysis will only use four principal components for further analysis. Because the sum of squared eigenvectors in each dimension is one. It means that squared eigenvector of each variable is the contribution of given variable to given component.

Based on Figure 11, Temperature and geopotential height are the most critical component for component one and those variables will be the contrast of east-west wind direction in the same component. The second component will be dominantly by North-South wind direction and contrast with sea level pressure. This analysis is in general form, and factor analysis will have a deeper analysis of the similarity and dissimilarity among variables.

Factor Analysis

Figure 12 is the bar chart of factor loadings of each variables using varimax rotation. The red bar reflects positive loadings, and blue bar reflects negative loadings. The noticeable variables that contribute in shaping factor one are the contrast between temperature and geopotential height with east wind direction. This factor will be called temperature and geopotential height. Furthermore, component two are mostly windspeed rate and north-south direction. Factor two will be called windspeed to north-south.

The author creates the group of the variables based on the pressure level (500,700, and 850 Hpa) and turns out that the higher the pressure, the more this group contributes to Factor 1 and Factor 2 (Figure 13). It seems that there is more information that can be extracted at higher pressure level. Furthermore, temperature plays an important role in both factors and based on the journal (Fan & Wei, 2008), more ozone days occurred in high temperature compared to low temperature. However, this information is not enough to explain why ozone occurred.

Based on the squared factor's loading, the first two variables that contribute the most to the Factor 1 are K-index and t-total followed by temperature and sea level pressure.

Two sample Hotelling T^2 test

Based on the Box M of equal variances, the group with attribute 0 and one do not have the equal variance-covariance matrix. The Chi-square approximation is 598,36 with 300 degrees of freedom (Figure 17). The two sample T^2 test shows the test statistic of 794 and follows the chi-square distribution with 24 degrees of freedom (Figure 16). Therefore, the multivariate mean group for ozone and non-ozone day is not equal to zero.

An individual test of the ANOVA (Figure 17) shows that relative humidity, k-index, t-total, sea level pressure, and precipitation are not different between the group ozone day and non-ozone day. It seems that the windspeed, temperature, and direction of the wind are different among the group. The authors conduct a test of MANOVA on three different level of pressure and turn out that altogether (v,u,t,rh, and ht) are different between ozone day and non-ozone day.

Discriminant Analysis

Based on the plot of the two principal components and the occurrence of the attribute, there is no pattern in the occurrence of the ozone day. It scattered across the dimension. The author split the dataset as “model” for the 80% of the data and “validation” for 20% of the data randomly. Since the variance-covariance matrix between first group and second group are not equal, linear discriminant analysis can not be used in here. The original probability for ozone and non-ozone day are 3% against 97%. The important thing in this dataset is that the occurrence of the ozone dataset is far worse than the non-occurrence. We can’t just assume that most of the time it doesn’t happen. Based on the table below, discriminant analysis is quite accurate. However, based on the numerical table, the model under predict by eight days (17%). The real consequences of this ozone day are quite grave if the model is not accurate. However, if it is compared with non-ozone day, the result will be 8/1848 (0.4%). Therefore, the accuracy of this model is dominated by the non-occurrence of ozone day.

DA	Covariance Matrix	Priors	Accuracy Method	Success	Failure	Overall
Quadratic	Unequal	Proportional	Cross Validation	0.9767	0.1351	0.9486
			Validation	0.9819	0.05	0.9568

Table of probability for Ozone day occurrence

Accuracy Method	Success	Failure
Original	57	1791
Expected Model	45	1433
Expected Validation	12	358
Model	37	1071
Validation	20	720

Table of numerical the occurrence of Ozone day

Conclusion

This dataset is not multivariate normal therefore any assessment that requires the assumption of multivariate normality will not give a good result in here. In the principal component analysis and factor analysis, temperature and geopotential heights are the key factors to generate the other number in this dataset. Furthermore, MANOVA and Two sample Hotelling T^2 test show that there is a difference of temperature and geopotential height during ozone and non-ozone day. Discriminant analysis shows high accuracy in the number, but the occurrence of the ozone day is far worse than the non-occurrence. Therefore, the author will conclude that this model is not accurate to predict the ozone level despite the high number of accuracy. A new dataset taken in the days with hot temperature and high geopotential height will give better accuracy in predicting ozone day using discriminant analysis.

References

Fan, K., & Wei, Z. ·. (2008). Forecasting skewed biased stochastic ozone days:. *Knowl Inf Syst 14*, 299–326.

Schlink, U., Dorling, S., Nunnari, E. P., Cawley, G., Junninen, H., Greig, A., & Foxal, R. (2003). A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment*, 3237-3253.

Appendix

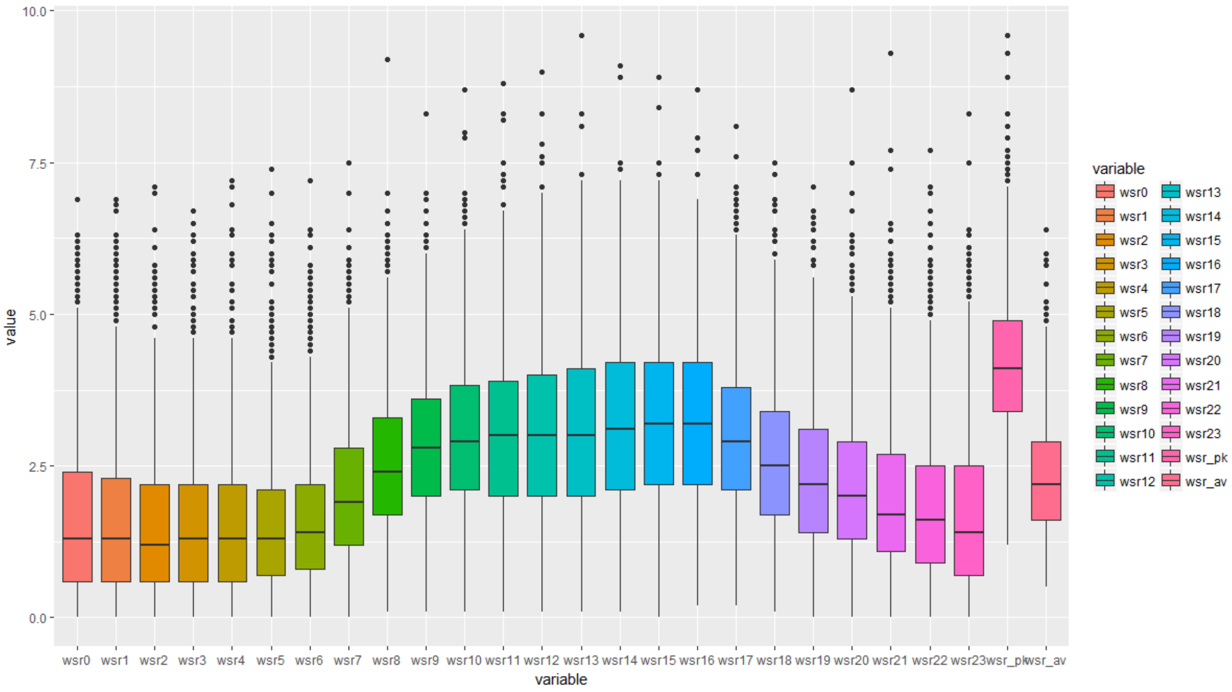


Figure 1 Boxplot of temperature at 22 separate times

Hafid Pradipta
4496585
Final Project Stat-520

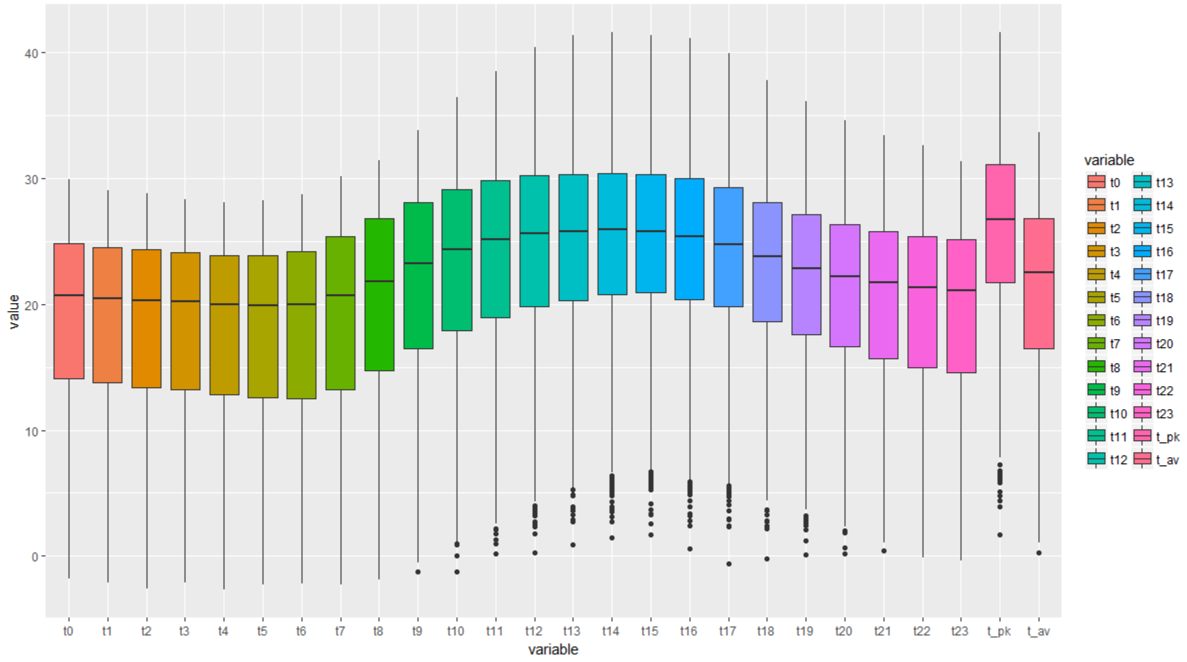


Figure 2 Boxplot of temperature at 22 separate times.

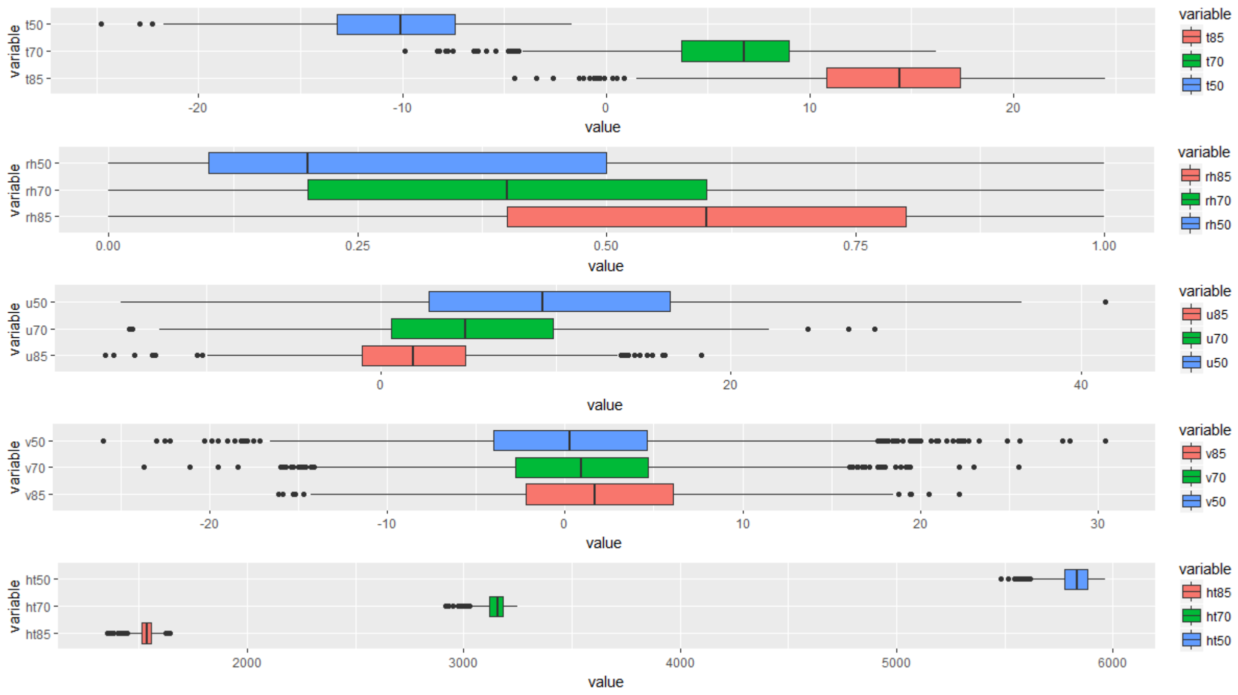


Figure 3 Boxplot of 5 variables (t,rh,u,v,ht) at three different levels.

Hafid Pradipta
4496585
Final Project Stat-520

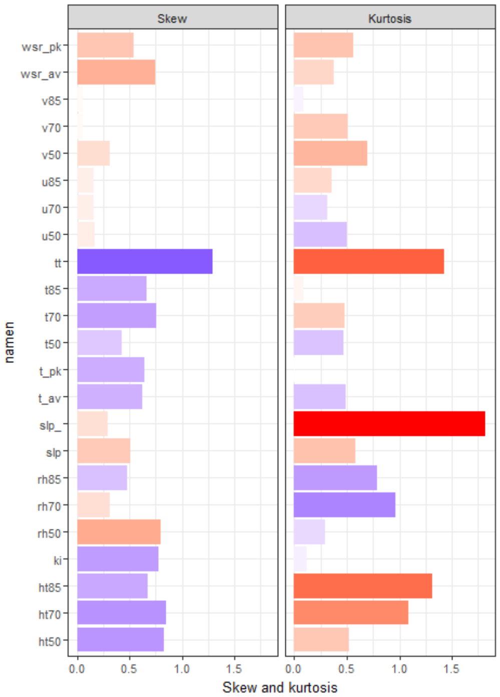
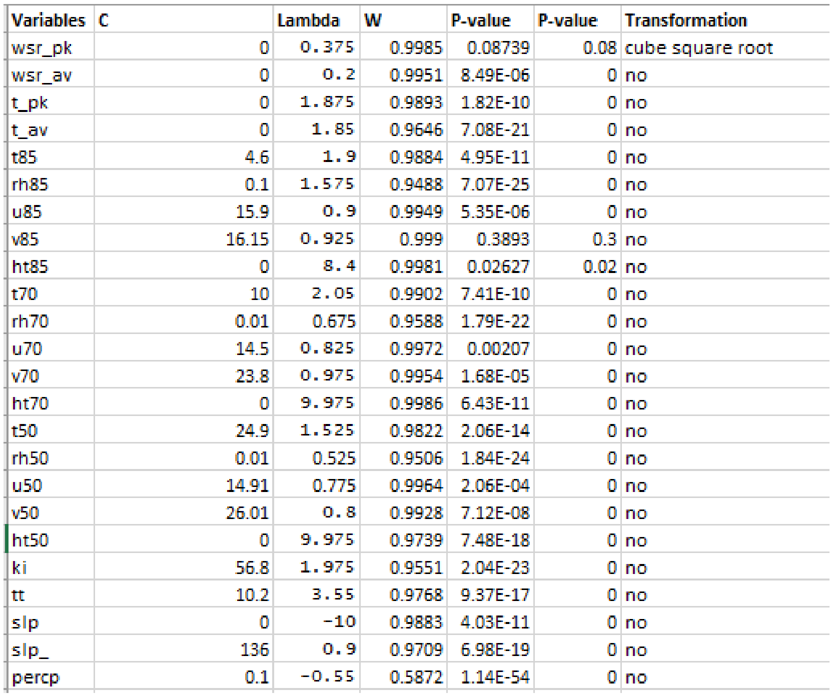
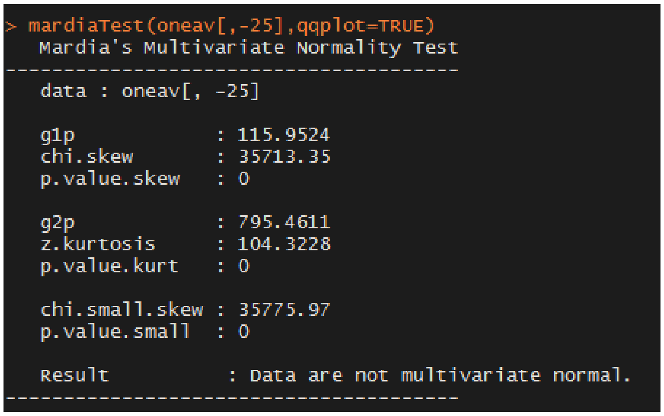


Figure 4 Bar chart of skewness and kurtosis



Variables	C	Lambda	W	P-value	P-value	Transformation
wsr_pk	0	0.375	0.9985	0.08739	0.08	cube square root
wsr_av	0	0.2	0.9951	8.49E-06	0	no
t_pk	0	1.875	0.9893	1.82E-10	0	no
t_av	0	1.85	0.9646	7.08E-21	0	no
t85	4.6	1.9	0.9884	4.95E-11	0	no
rh85	0.1	1.575	0.9488	7.07E-25	0	no
u85	15.9	0.9	0.9949	5.35E-06	0	no
v85	16.15	0.925	0.999	0.3893	0.3	no
ht85	0	8.4	0.9981	0.02627	0.02	no
t70	10	2.05	0.9902	7.41E-10	0	no
rh70	0.01	0.675	0.9588	1.79E-22	0	no
u70	14.5	0.825	0.9972	0.00207	0	no
v70	23.8	0.975	0.9954	1.68E-05	0	no
ht70	0	9.975	0.9986	6.43E-11	0	no
t50	24.9	1.525	0.9822	2.06E-14	0	no
rh50	0.01	0.525	0.9506	1.84E-24	0	no
u50	14.91	0.775	0.9964	2.06E-04	0	no
v50	26.01	0.8	0.9928	7.12E-08	0	no
ht50	0	9.975	0.9739	7.48E-18	0	no
ki	56.8	1.975	0.9551	2.04E-23	0	no
tt	10.2	3.55	0.9768	9.37E-17	0	no
slp	0	-10	0.9883	4.03E-11	0	no
slp_	136	0.9	0.9709	6.98E-19	0	no
percp	0.1	-0.55	0.5872	1.14E-54	0	no

Figure 5 Table of box cox transformation



```
> mardiaTest(oneav[,-25],qqplot=TRUE)
   Mardia's Multivariate Normality Test
---------------------------------------
   data : oneav[, -25]

   g1p            : 115.9524
   chi.skew       : 35713.35
   p.value.skew   : 0

   g2p            : 795.4611
   z.kurtosis     : 104.3228
   p.value.kurt   : 0

   chi.small.skew : 35775.97
   p.value.small  : 0

   Result          : Data are not multivariate normal.
---------------------------------------
```

Figure 6 Output of Mardia Test

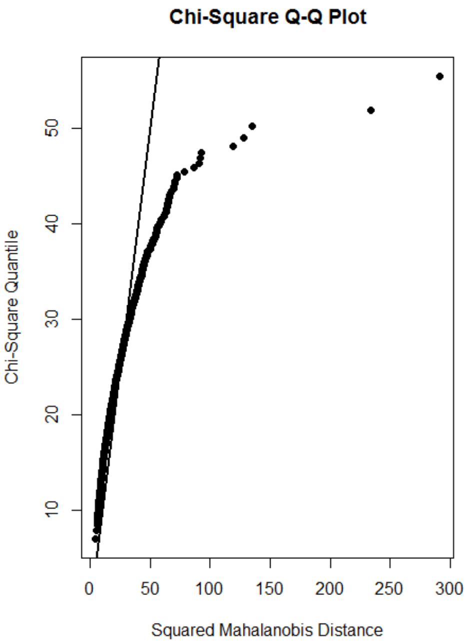


Figure 7 Plot of Chi Square quantile and Squared Mahalanobis Distance

Percentage of observation in the contour		
Expected in%	Observed in%	Difference
80	75.9	4.1
90	82.52	7.48
95	86.9	8.1

Figure 8 Multivariate normality expected and observed proportion

Hafid Pradipta
4496585
Final Project Stat-520

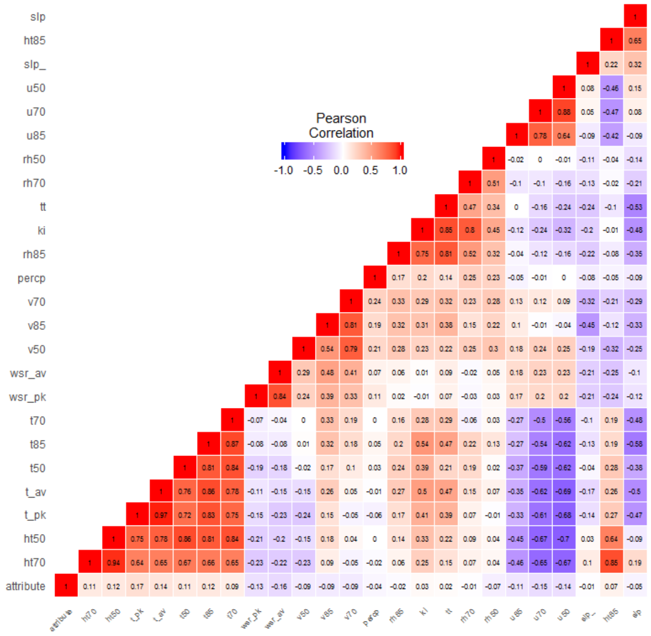


Figure 9 Correlation Matrix

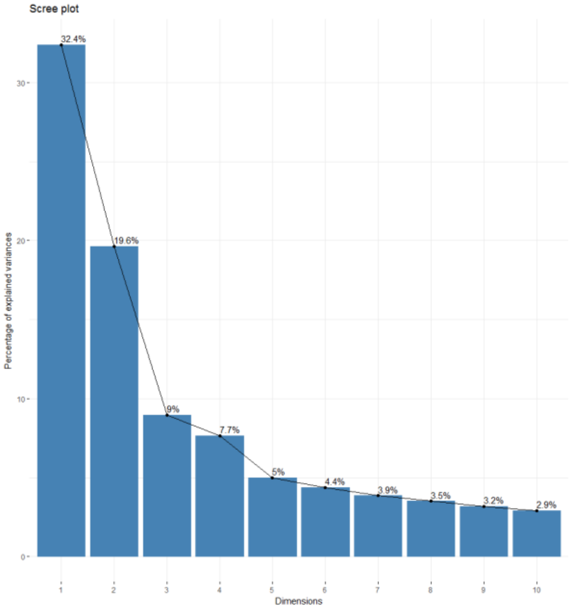


Figure 10 Scree Plot of Principal Component Analysis

Hafid Pradipta
4496585
Final Project Stat-520

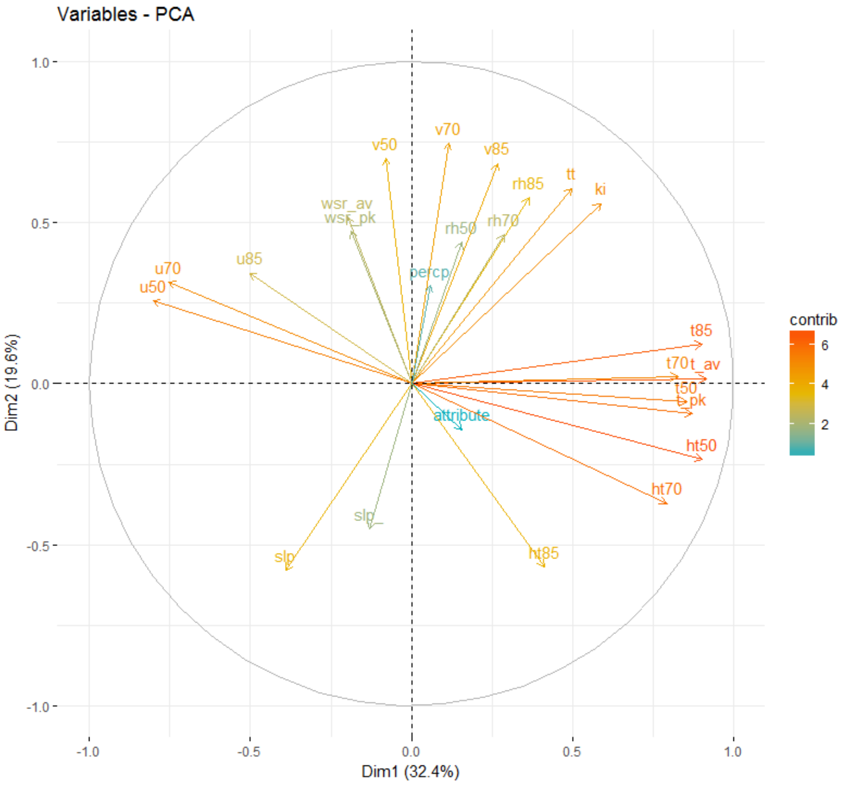


Figure 11 Direction and coordinate of each variable in the first two principal components

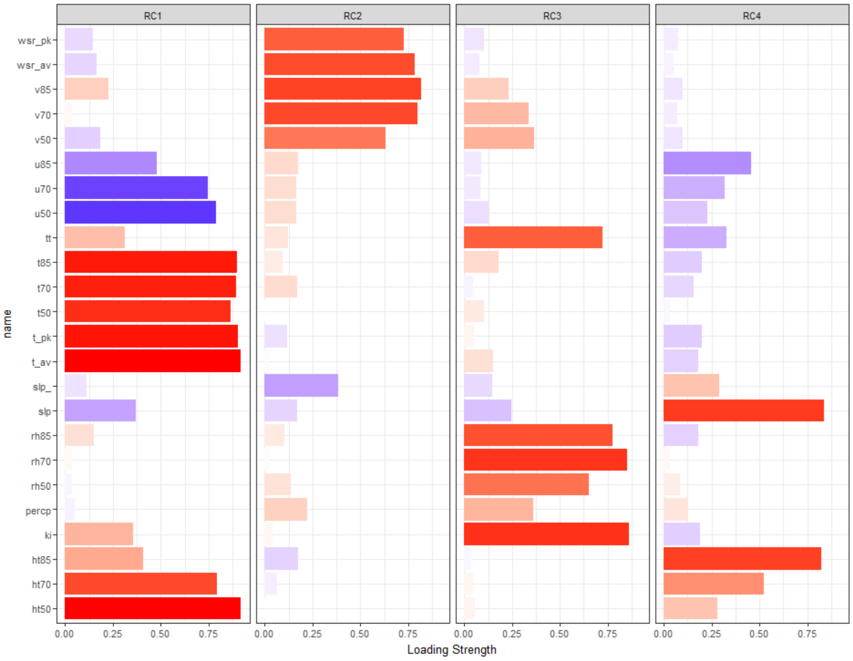


Figure 12 factor loadings using varimax rotation

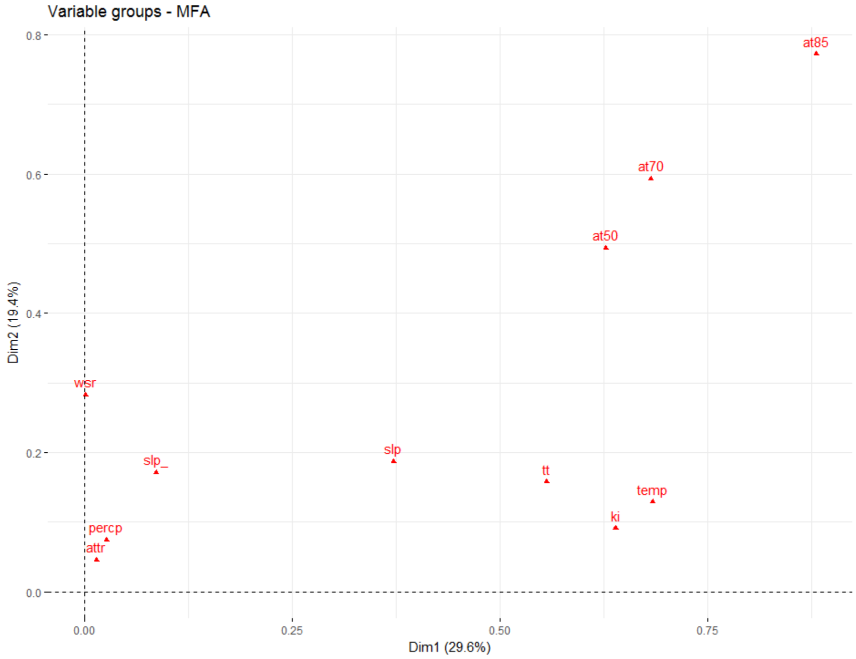


Figure 13 Plot of grouping at different level of pressure

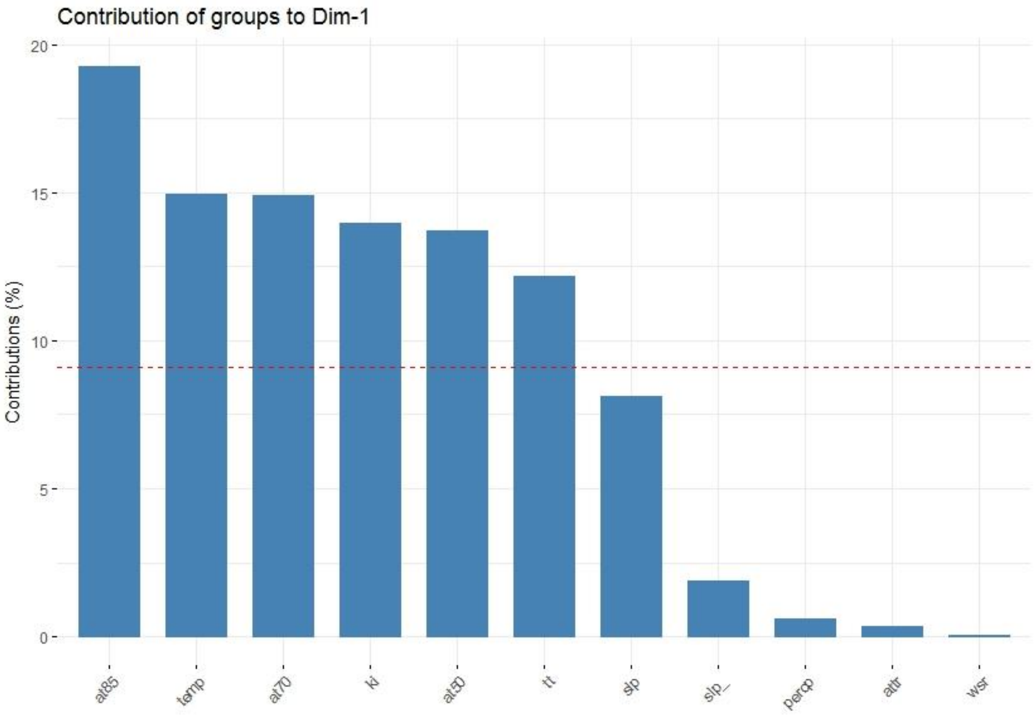


Figure 14 Contribution of the group to dimension 1

Hafid Pradipta
4496585
Final Project Stat-520

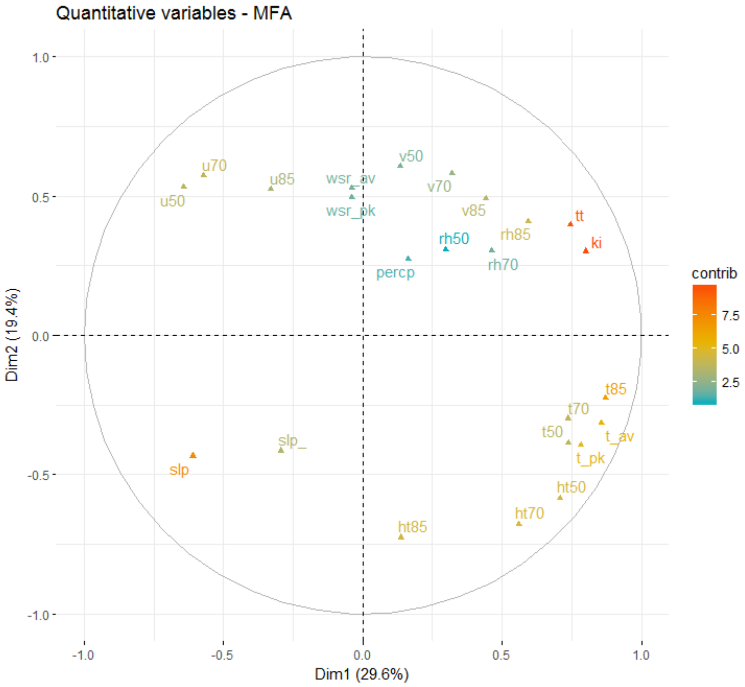
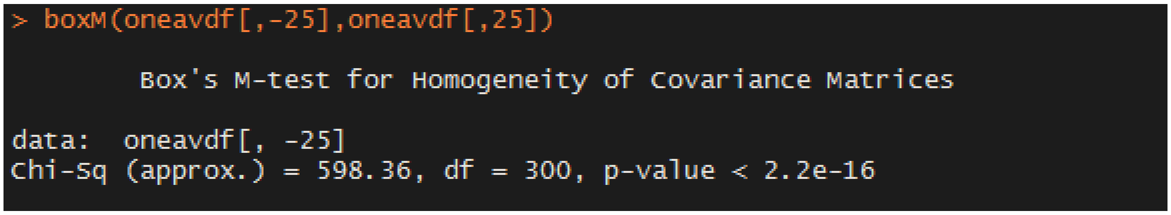


Figure 15 Biplot of Factor Analysis



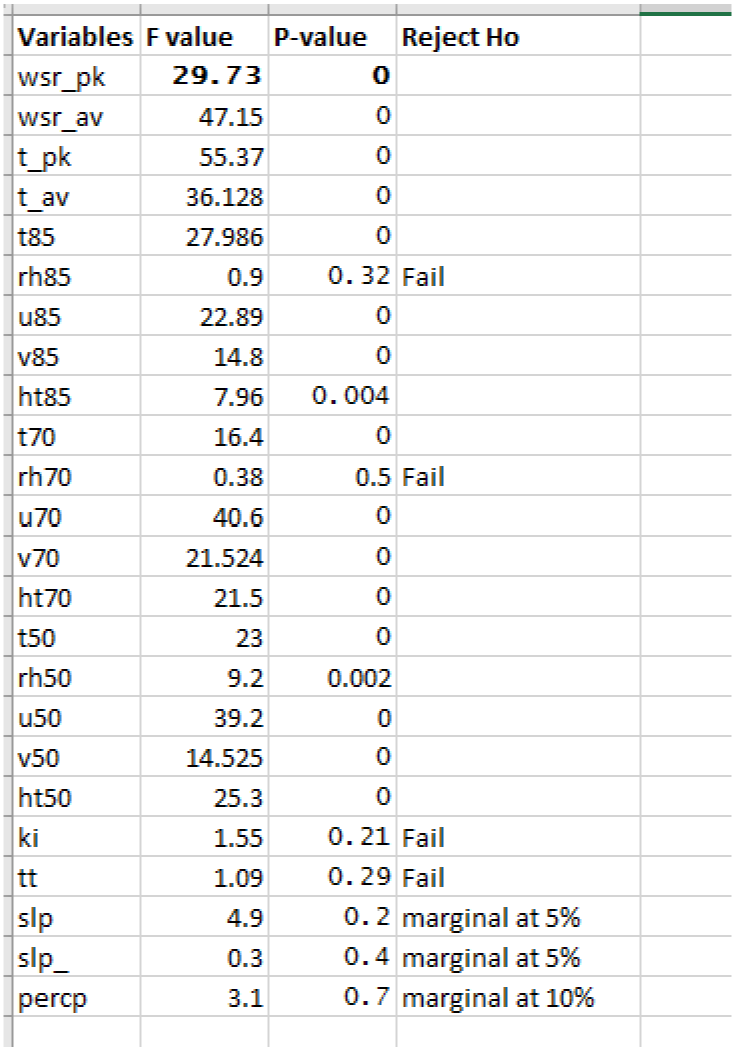
```
> boxM(oneavdf[,-25],oneavdf[,25])

        Box's M-test for Homogeneity of Covariance Matrices

data:  oneavdf[, -25]
Chi-Sq (approx.) = 598.36, df = 300, p-value < 2.2e-16
```

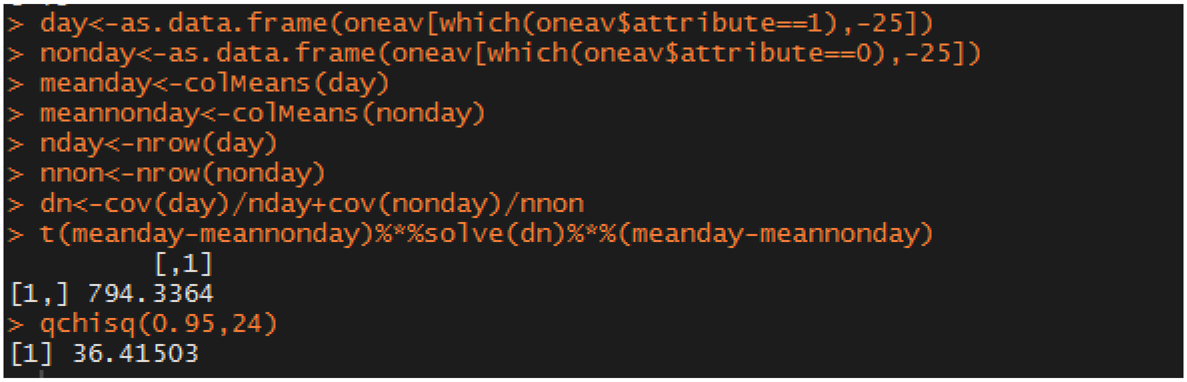
Figure 16 Box M test variance covariances equality

Hafid Pradipta
4496585
Final Project Stat-520



Variables	F value	P-value	Reject Ho
wsr_pk	29.73	0	
wsr_av	47.15	0	
t_pk	55.37	0	
t_av	36.128	0	
t85	27.986	0	
rh85	0.9	0.32	Fail
u85	22.89	0	
v85	14.8	0	
ht85	7.96	0.004	
t70	16.4	0	
rh70	0.38	0.5	Fail
u70	40.6	0	
v70	21.524	0	
ht70	21.5	0	
t50	23	0	
rh50	9.2	0.002	
u50	39.2	0	
v50	14.525	0	
ht50	25.3	0	
ki	1.55	0.21	Fail
tt	1.09	0.29	Fail
slp	4.9	0.2	marginal at 5%
slp_	0.3	0.4	marginal at 5%
percp	3.1	0.7	marginal at 10%

Figure 17 Individual ANOVA Test



```
> day<-as.data.frame(oneav[which(oneav$attribute==1),-25])
> nonday<-as.data.frame(oneav[which(oneav$attribute==0),-25])
> meanday<-colMeans(day)
> meannonday<-colMeans(nonday)
> nday<-nrow(day)
> nnon<-nrow(nonday)
> dn<-cov(day)/nday+cov(nonday)/nnon
> t(meanday-meannonday)%*%solve(dn)%*%(meanday-meannonday)
         [,1]
[1,] 794.3364
> qchisq(0.95,24)
[1] 36.41503
```

Figure 1815 Two sample Hotelling T2 test

Hafid Pradipta
4496585
Final Project Stat-520

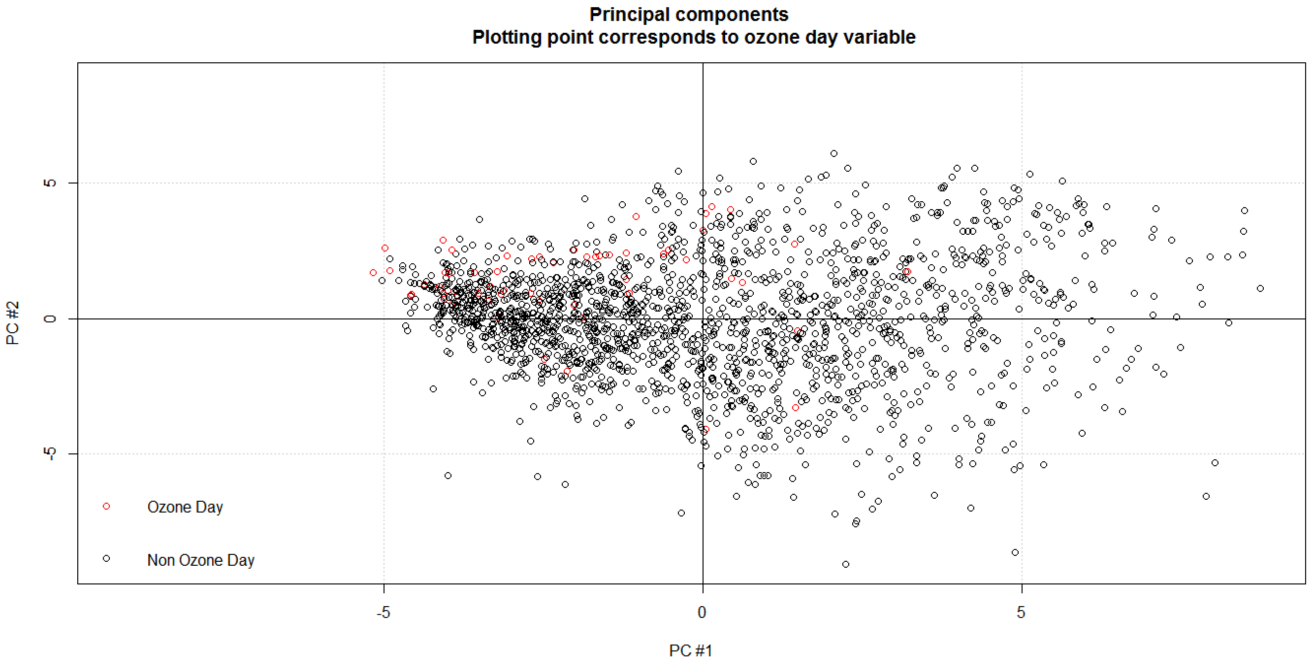


Figure 19 PC scores and ozone day variable