CLL 798

Term Project

Statistical Analysis

Of

Different Composition Glass Properties

Shivanshu Verma (2015CH70186)
Anurag Vashistha (2015CH70194)
Saransh Maheshwari(2015CH70183)

DATA DESCRIPTION

The data we will be working upon is purchased and available for IIT Delhi students to perform various statistical analysis and enhance statistical learning. The data contains over 10 lakh observations for more than 900 properties of different compositions of glasses. This is a huge data set and we will only be performing statistical analysis on a very small part of it which is training set. Therefore we cropped the dataset and the training dataset, contains 45000 observations and 30 variables.
The variables are:

- **Glass ID :** Glass ID is unique for every observation and not used in any analysis whereas just helped in the identification of particular glass.
- **Compositions :** Numbered as 1, 2, 3……. , these IDs tells about the composition of glass. Every number represents different oxides present in the glass and the corresponding value represents it's concentration in the particular glass. Therefore, the row sum must be equal to 100% for a valid composition of glass.

  Original Data Contained almost 900 oxides and their compositions, whereas we have taken 30 oxides, which had significant amount of impact in the density. The following are the oxides we have taken into analysis :

  - 001 ----> SiO_2
  - 002 ----> B_2O_3
  - 003 ----> Al_2O_3
  - 007 ----> Li_2O
  - 008 ----> Na_2O

  Other IDs included in the dataset are not taken for statistical analysis.

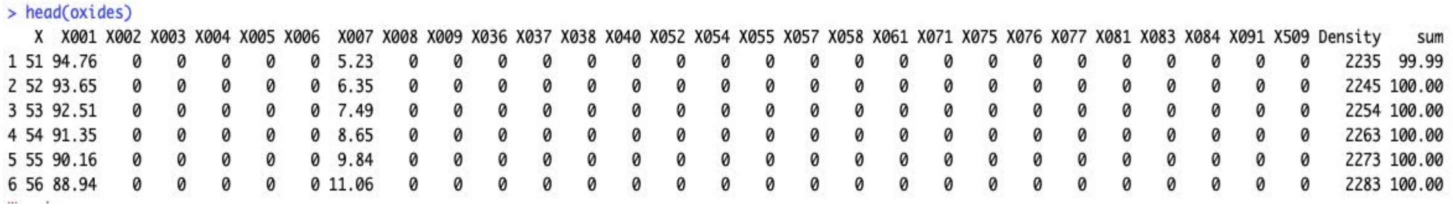
- **Density :** The last column contained density of the glass corresponding to the different compositions of different oxides present within the glass.

Density will be the response here and other variables as predictor.

DATA CLEANING

As this data was not proper for directly performing statistical analysis on it, we required to clean the data in the desirable way. Firstly, we deleted all the zero and non integer values of density from the data set. After that, we added the another column named as sum, which represents the sum of all the oxides composition of glass and deleted all the data where the sum of compositions was greater than 100 and less than 100, as for a valid glass, the sum of its compositions must be 100. After performing these tasks in the data, the data is now fit for analysing.

There are now no other abnormalities present in the dataset. The result of data cleaning is a dataset with 35351 observations and 31 variables. A head() function output (Figure 01) can give an idea how the data structures after cleaning.



```
> head(oxides)
   X  X001 X002 X003 X004 X005 X006  X007 X008 X009 X036 X037 X038 X040 X052 X054 X055 X057 X058 X061 X071 X075 X076 X077 X081 X083 X084 X091 X509 Density    sum
1 51 94.76    0    0    0    0    0  5.23    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2235  99.99
2 52 93.65    0    0    0    0    0  6.35    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2245 100.00
3 53 92.51    0    0    0    0    0  7.49    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2254 100.00
4 54 91.35    0    0    0    0    0  8.65    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2263 100.00
5 55 90.16    0    0    0    0    0  9.84    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2273 100.00
6 56 88.94    0    0    0    0    0 11.06    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    2283 100.00
```

Figure 1 : Representing head() function output on the cleaned dataset

EXPLORATORY DATA ANALYSIS

Now to visualise the effect of different compositions on the density of Glass, we will be plotting different boxplots. Considering the following dataset, we found that glass is mostly consist of only 2 oxides whereas the composition of the rest oxides is zero within the particular glass. So, what we do is partition the following dataset into small partitions containing only two oxides composition and these small partitions are named as following :

- Comp12 : Contains dataset where the glass is made up of only oxide 001 and oxide 002.
- Comp13 : Contains dataset where the glass is made up of only oxide 001 and oxide 003.
- Comp17 : Contains dataset where the glass is made up of only oxide 001 and oxide 007.
- Comp178: Contains dataset where the glass is made up of oxides 001, 007 and 008.

The reason behind taking the same oxide 001 with three others is to get the feel of how different oxides mixed with oxide 001 affects the density. Another reason for that is that in all 45 K observations, all the glasses consisted of oxide 001 with different compositions , therefore taking oxide 001 as the base helps to understand better the effect of other oxides’ compositions on the density of the glass.

Note: All the plots are plotted between composition of Oxide 001 vs Density in different partitions including Comp12, Comp13 and Comp17 for being on the same basis.

Now, as the dataset is made up of only two oxides, we can perform analysis between the composition of oxide 001 and density. Similarly for other partitions too.

For visualization, following is the boxplot of density vs composition for different partitions shown in Figure 2:

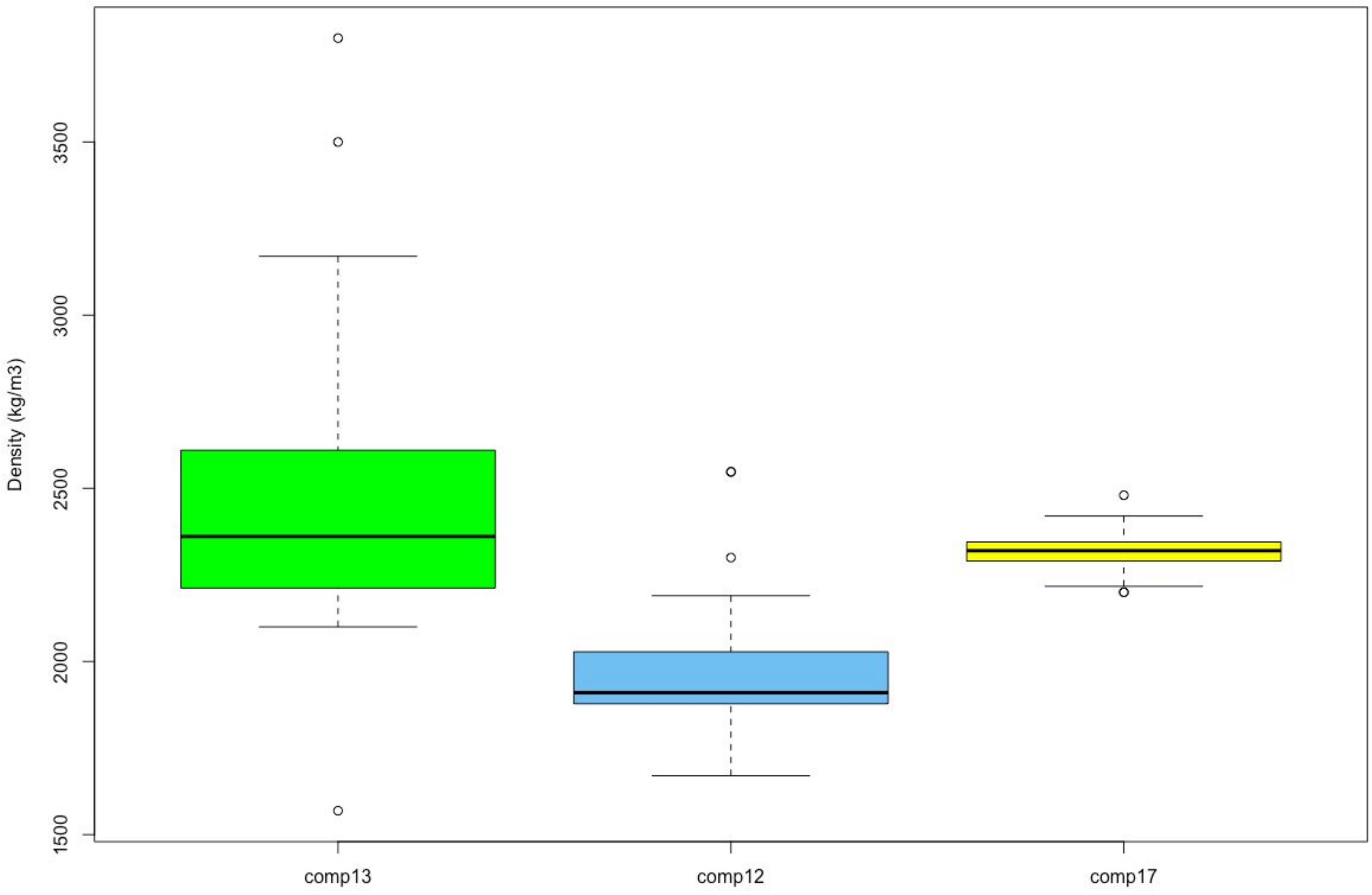


Figure 2 : Different boxplots of compositions (001,002), (001,003) and (001 , 007)

Inferences : The following boxplots are plotted on the same graph to compare how the different oxides are affecting density of the glass. The boxplot of different compositions against density ensures the composition trend with the density. Density is generally low in the glass made up of oxides (001 and 002) and it peaks in (001 and 007). Different oxides compositions are therefore one of the determining factors that affects glass density .

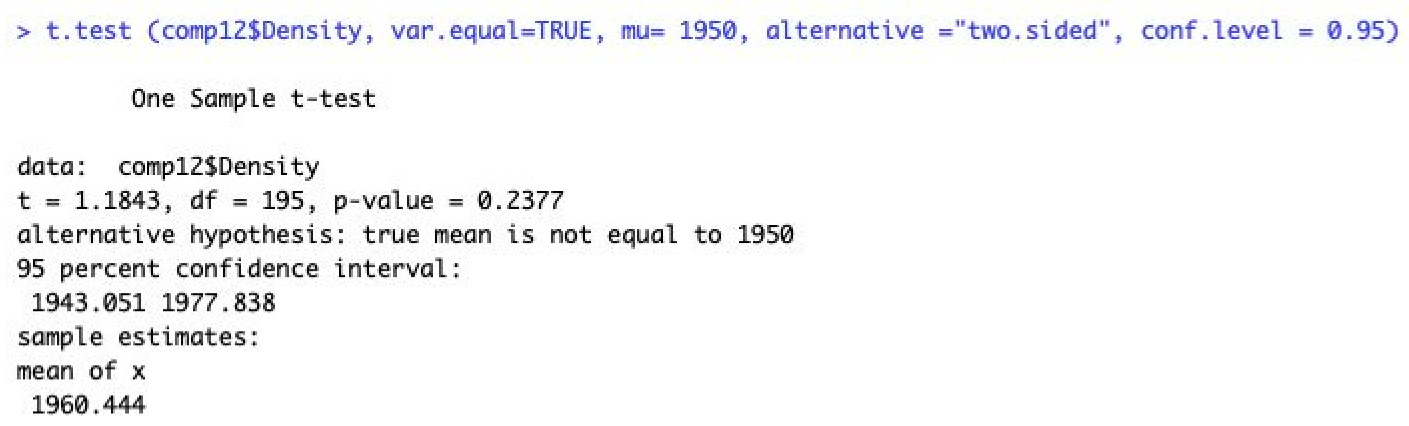
HYPOTHESIS TESTING

From the boxplot given above, we roughly estimated mu value for all the partitions and then performed t test on them to check the hypothesis. T value is simply the calculated difference represented in units of standard error. The greater the magnitude of T (it can be either positive or negative), the greater the evidence against the null hypothesis that there is no significant difference. The closer T is to 0, the more likely there isn't a significant difference.

The assumed mu values for the different partitions are given below:

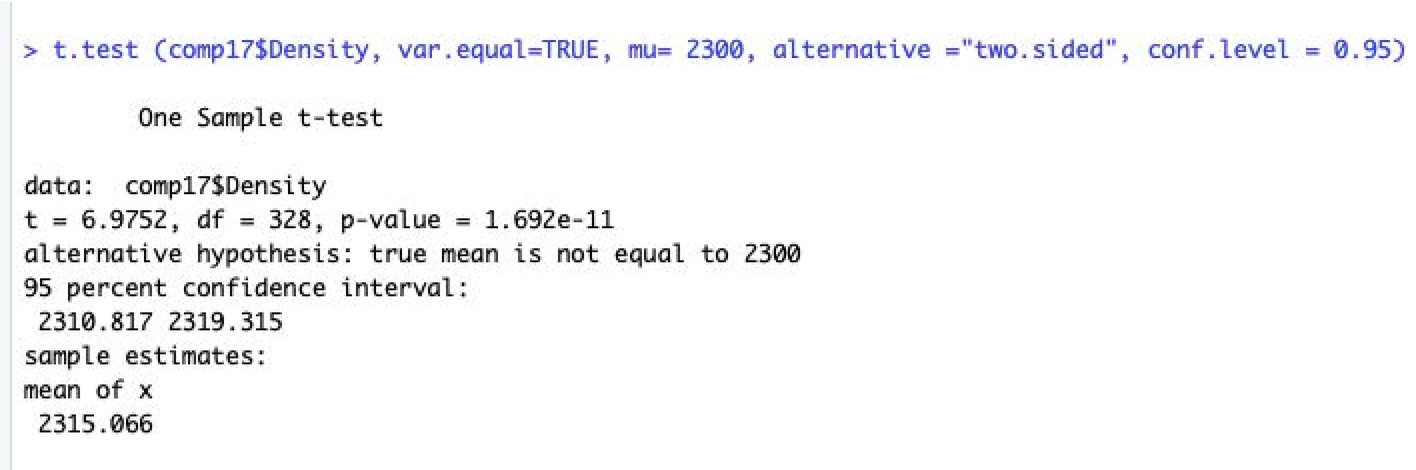
- Comp12 -------- 1950
- Comp13 -------- 2450
- Comp17 -------- 2300

The results for the following were as follows:



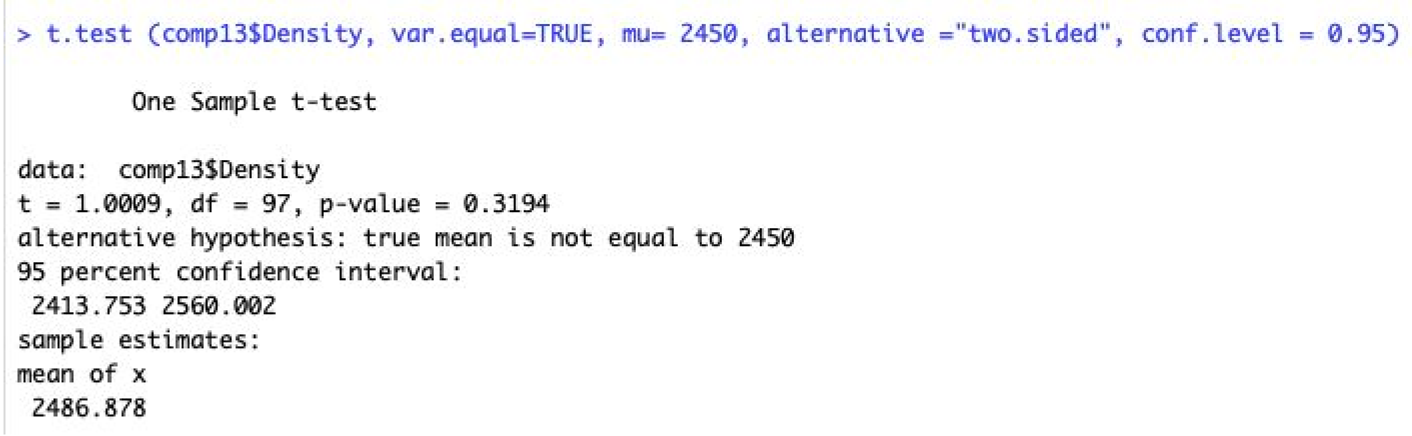
```
> t.test (comp12$Density, var.equal=TRUE, mu= 1950, alternative ="two.sided", conf.level = 0.95)

        One Sample t-test

data:  comp12$Density
t = 1.1843, df = 195, p-value = 0.2377
alternative hypothesis: true mean is not equal to 1950
95 percent confidence interval:
 1943.051 1977.838
sample estimates:
mean of x
 1960.444
```

Figure 3 : t test for comp12 at mu = 1950



```
> t.test (comp17$Density, var.equal=TRUE, mu= 2300, alternative ="two.sided", conf.level = 0.95)

        One Sample t-test

data:  comp17$Density
t = 6.9752, df = 328, p-value = 1.692e-11
alternative hypothesis: true mean is not equal to 2300
95 percent confidence interval:
 2310.817 2319.315
sample estimates:
mean of x
 2315.066
```

Figure 4 : t test for comp17 at mu = 2300



```
> t.test (comp13$Density, var.equal=TRUE, mu= 2450, alternative ="two.sided", conf.level = 0.95)

        One Sample t-test

data:  comp13$Density
t = 1.0009, df = 97, p-value = 0.3194
alternative hypothesis: true mean is not equal to 2450
95 percent confidence interval:
 2413.753 2560.002
sample estimates:
mean of x
 2486.878
```

Figure 5 : t test for comp13 at mu = 2450

Inferences : From the above tests, we inferred that, in Figure 3, assumed mu value was 1950, with a confidence interval of 95%, which gave a low t value and p value greater than .05, which passes the null hypothesis. Similarly in Figure 5, t value is very less (near to 1) and p value is greater than .05, which also provides Null Hypothesis true. Whereas in Figure 2, assumed mu value was 2300, which gave the t value larger and also p value very small, proving alternative hypothesis true.

SIMPLE LINEAR REGRESSION

Now, we would like to see how the density varies with the composition as linear single variable regression. By applying the simple linear regression model on different partitions, we get different parameters and figures. For e.g., figure 6 is the plot of linear regression between composition of oxide 001 and density in the partition of Comp17, which consists of oxide 001 and 007 only.

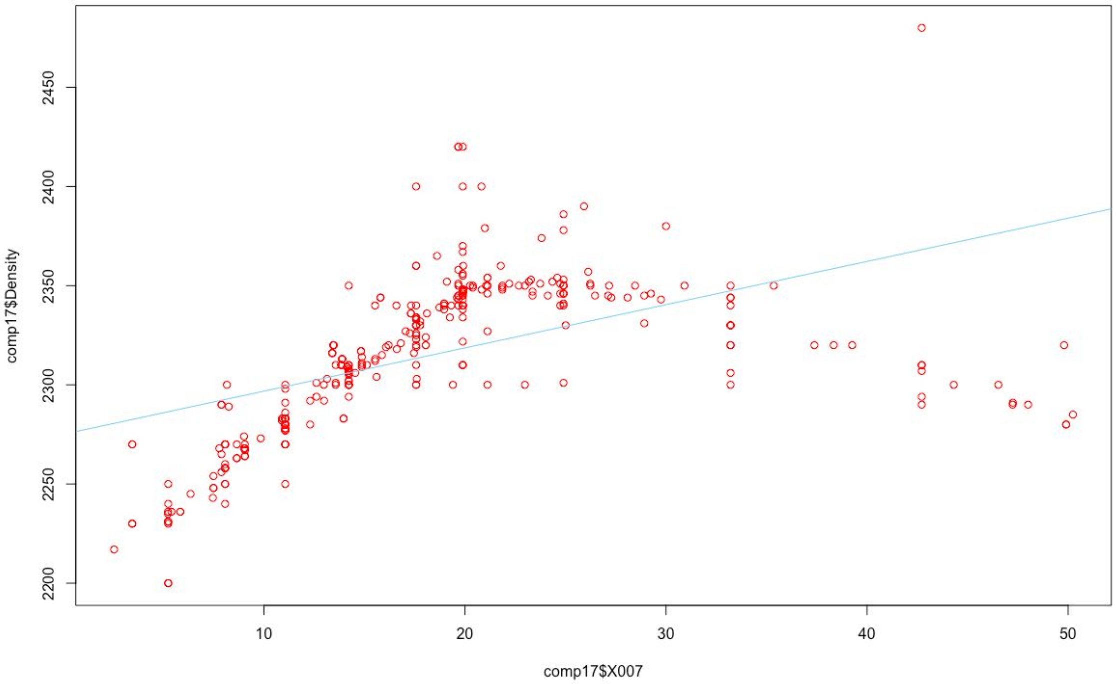


Figure 6 : Linear Regression on Comp17

Whereas, the Single variable Linear Regression in Comp13 gives better results (See figure 7)

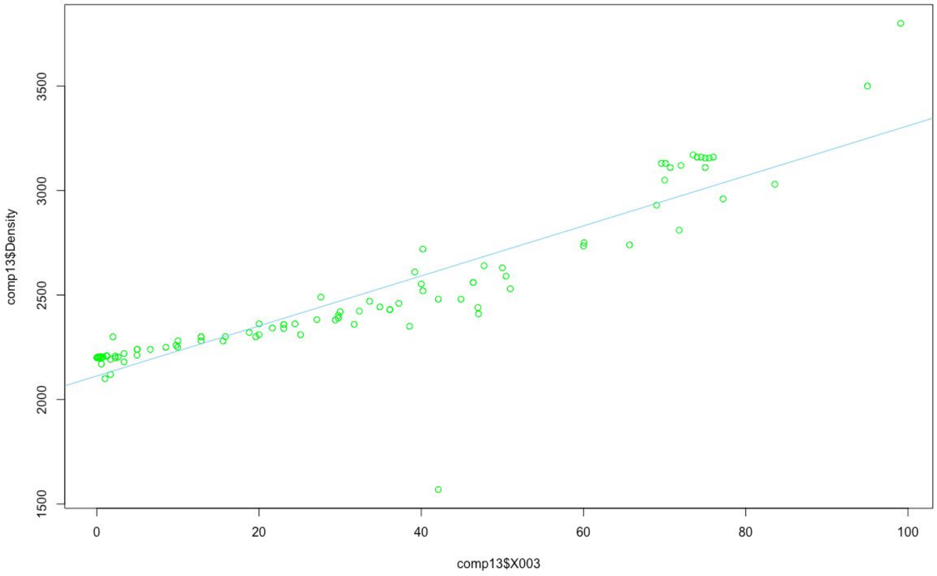
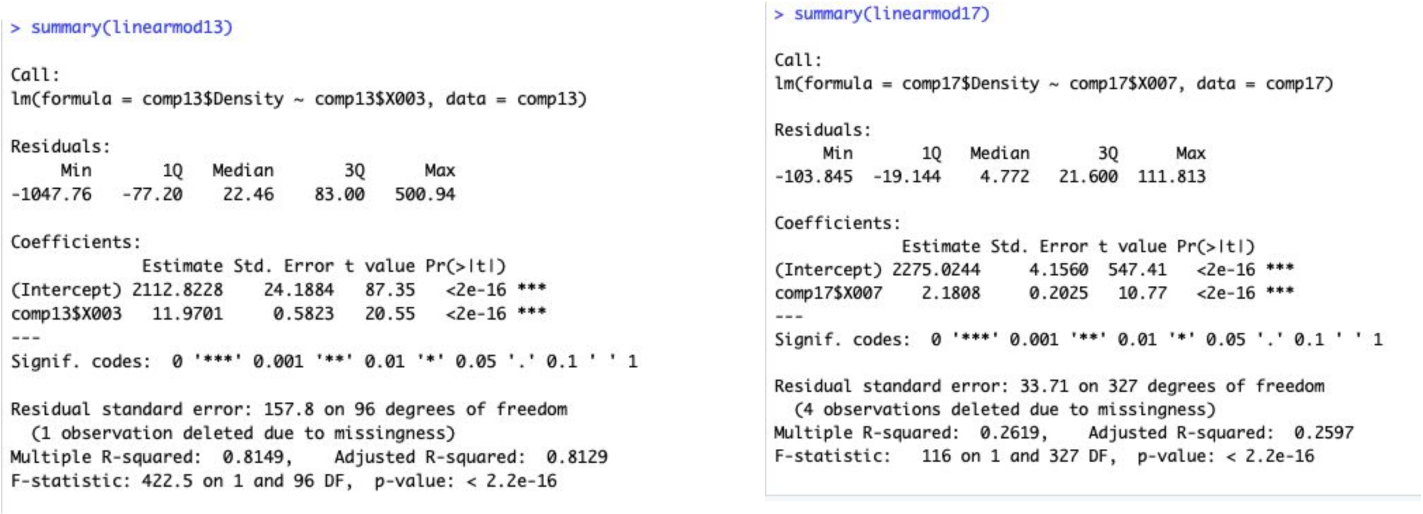


Figure 7 : Linear Regression on Comp13

The reason behind this is also given by the statistical values provided by the summary function of both the linear regressions(See Figure 8).



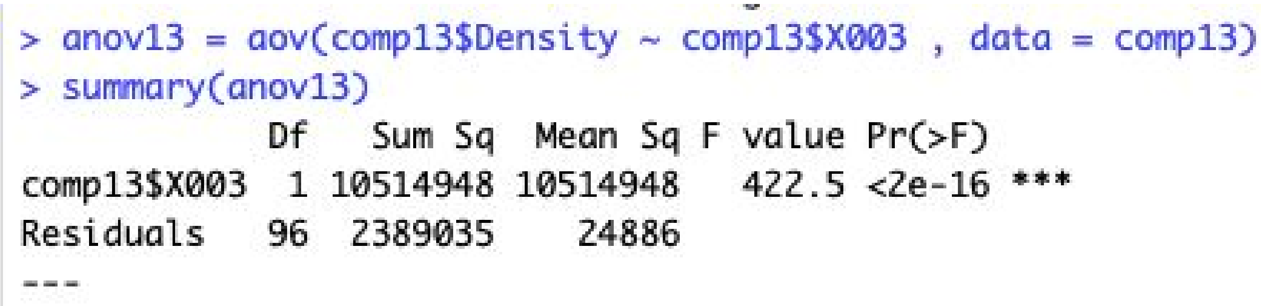
```
> summary(linearmod13)

Call:
lm(formula = comp13$Density ~ comp13$X003, data = comp13)

Residuals:
     Min       1Q   Median       3Q      Max 
-1047.76   -77.20    22.46    83.00   500.94 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2112.8228    24.1884   87.35   <2e-16 ***
comp13$X003   11.9701     0.5823   20.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 157.8 on 96 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8149,    Adjusted R-squared:  0.8129 
F-statistic: 422.5 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
> summary(linearmod17)

Call:
lm(formula = comp17$Density ~ comp17$X007, data = comp17)

Residuals:
     Min       1Q   Median       3Q      Max 
-103.845  -19.144    4.772   21.600  111.813 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2275.0244     4.1560  547.41   <2e-16 ***
comp17$X007    2.1808     0.2025   10.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.71 on 327 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.2619,    Adjusted R-squared:  0.2597 
F-statistic:   116 on 1 and 327 DF,  p-value: < 2.2e-16
```

Figure 8 : Comparing Statistical Values between Comp13 and Comp17

The R value for comp13 is 0.812 whereas for comp17 is 0.26, which can also be observed by looking at the figure 7 and figure 6. The more the value of R near 1, the more data fits in the linear model.

ANOVA Test :



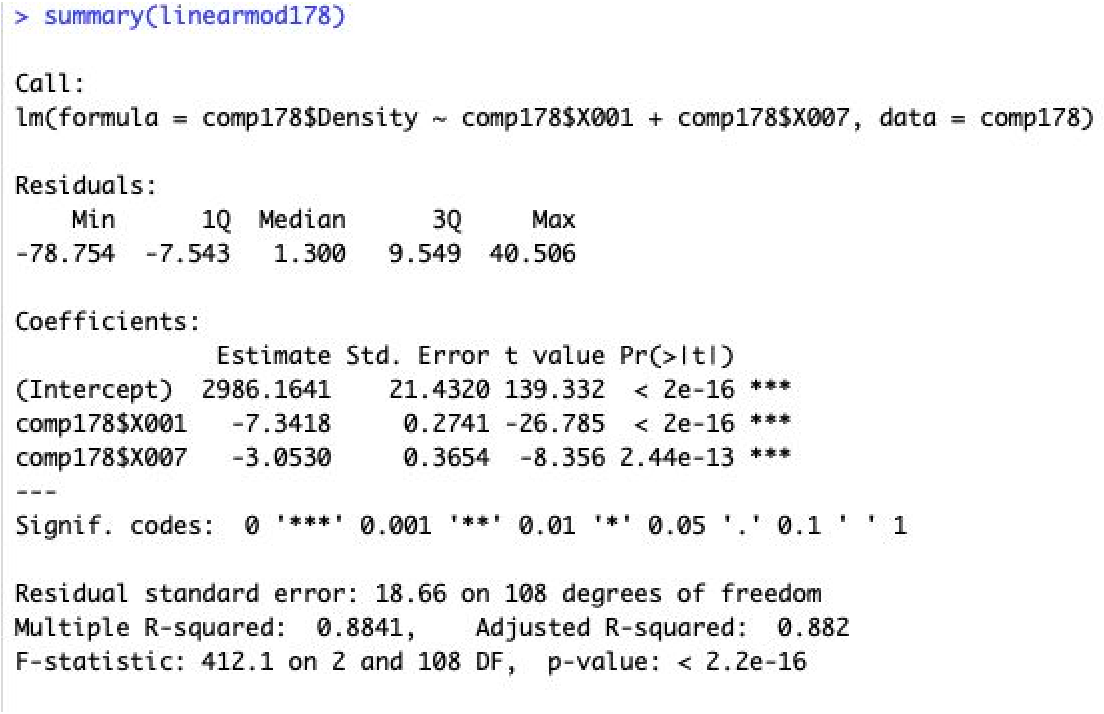
```
> anov13 = aov(comp13$Density ~ comp13$X003 , data = comp13)
> summary(anov13)
            Df   Sum Sq  Mean Sq F value Pr(>F)    
comp13$X003  1 10514948 10514948   422.5 <2e-16 ***
Residuals   96  2389035    24886                   
---
```

Figure 9 : ANOVA test for Comp13

Here, my F value is 422.5, and p-value is very low too. In other words, the variation of X003 means among density and our p-value is less than 0.05 (as suggested by normal scientific standard). Hence we can conclude that for our confidence interval we accept the alternative hypothesis H1 that there is a significant relationship between density and X003.

MULTIVARIATE LINEAR REGRESSION

For Multivariate Linear Regression, we have made a different partition which is comp178, consisting of glasses composed of only oxide 001, 007 and 008. In this , we will find the more than one parameters, whereas in R, the code for this almost same as that for Simple Linear Regression.

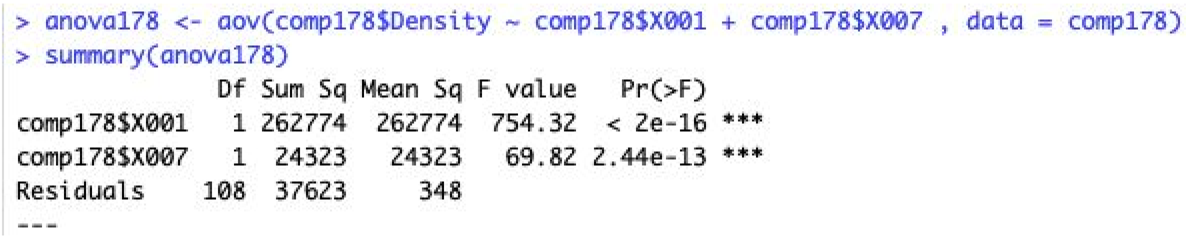


```
> summary(linearmod178)

Call:
lm(formula = comp178$Density ~ comp178$X001 + comp178$X007, data = comp178)

Residuals:
    Min      1Q  Median      3Q     Max 
-78.754  -7.543   1.300   9.549  40.506 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2986.1641    21.4320 139.332  < 2e-16 ***
comp178$X001   -7.3418     0.2741 -26.785  < 2e-16 ***
comp178$X007   -3.0530     0.3654  -8.356 2.44e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.66 on 108 degrees of freedom
Multiple R-squared:  0.8841,	Adjusted R-squared:  0.882 
F-statistic: 412.1 on 2 and 108 DF,  p-value: < 2.2e-16
```

Figure 10 : Multivariate Linear Regression Summary

From the summary of multivariate Linear Regression , we can see that the R value is .884 and adjusted R value is .882, which means that the model fits good. Now, look at other information through anova test (shown in figure 11):

ANOVA Test :



```
> anova178 <- aov(comp178$Density ~ comp178$X001 + comp178$X007 , data = comp178)
> summary(anova178)
              Df Sum Sq Mean Sq F value   Pr(>F)    
comp178$X001   1 262774  262774  754.32  < 2e-16 ***
comp178$X007   1  24323   24323   69.82 2.44e-13 ***
Residuals    108  37623     348                     
---
```

Figure 11 : Multivariate Linear Regression ANOVA Test

Here also, the F value is large and p-value is very low too. I Hence we can conclude that for our confidence interval we accept the alternative hypothesis H1 that there is a significant relationship between density , X001 and X007.

RIDGE REGRESSION

Ridge regression puts a penalty on the size of the coefficient. We will calculate the optimal λ (the complexity parameter that controls the amount of shrinkage) for the multivariate regression model where density is the predicted value and components 1, 7, and 8 are predictors. We have used the library *glmnet* for this.

Gives various regression models in a matrix form

```
> x <- model.matrix(comp178$Density ~ comp178$X1 + comp178$X7, data=comp178)
```

Inputs the density in a vector form

```
> y <- comp178$Density
```

Specifying the values of lambda

```
> lambdas <- 10^seq(3, -2, by = -.1)
```

Fitting the ridge regression model

```
> fit <- glmnet(x, y, alpha = 0, lambda=lambdas)
> summary (fit)
#          Length Class     Mode
a0             51           -none- numeric
beta           153          dgCMatrix S4
df             51           -none- numeric
dim            2            -none- numeric
lambda         51           -none- numeric
dev.ratio      51           -none- numeric
nulldev        1            -none- numeric
npasses        1            -none- numeric
jerr           1            -none- numeric
offset                 1            -none- logical
call           5            -none- call
nobs           1            -none- numeric
```

To calculate optimal lambda

```
cv_fit <- cv.glmnet(x, y, alpha = 0, lambda = lambdas)
plot (cv_fit)
opt_lambda <- cv_fit$lambda.min
opt_lambda
[1] 0.01
```

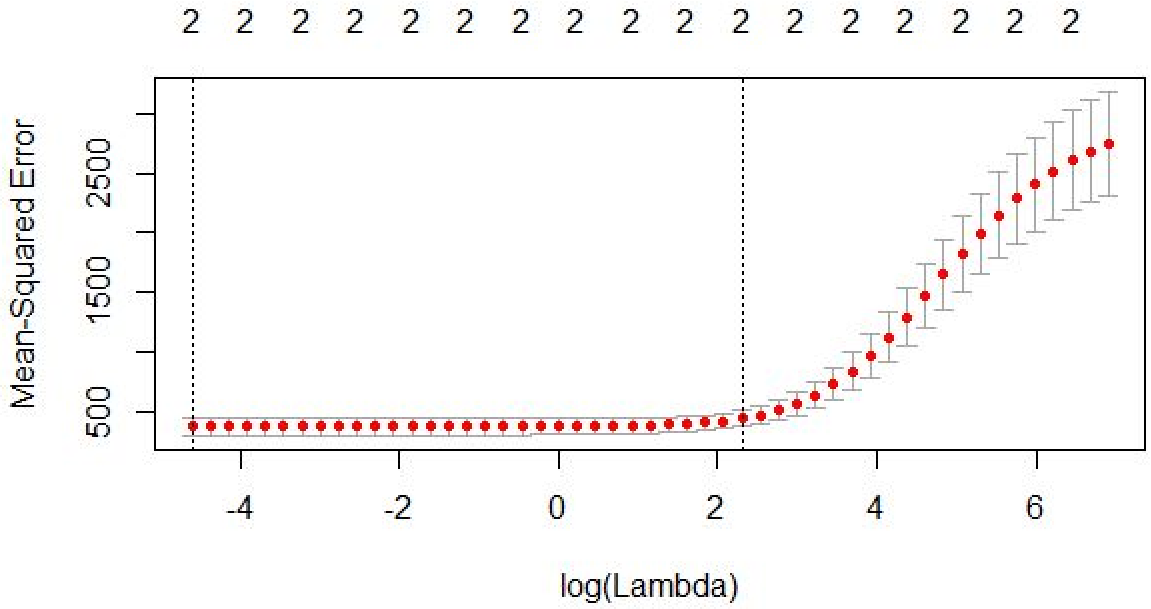


Figure 12 - The lowest value gives the optimal lambda

We obtain the value of optimal lambda to be 0.01. This low value implies a very low shrinkage in the regression coefficients. Thus doing ridge regression and imposing a penalty had very little effect on the values of our regression coefficients.

Thank You