Data Stats final project

Covid Group #5 – Prafull Goel, Abhishek Jaiswal

Dataset Description

This dataset comes from the COVID Analytics Group which is collecting data and building models to track the disease and make policy decisions based on predictions from these models. Key is to account for the prevalence of asymptomatic patients and include only sufficiently representative studies.

It's largely derived from studies run in hospitals and nations affected with Covid-19, majority of the studies being from China as of now. The dataset includes 539 cases. Each row represents a study with a cohort of patients. Cohorts then belong to a paper (one or more cohorts may belong to the same paper).

Given a cohort, each column represents information about this cohort of patients, divided roughly in the following categories:

1. demographic information (e.g. number of patients in the cohort, aggregated age and gender statistics)
2. comorbidity information (e.g. prevalence of diabetes, hypertension, etc.)
3. symptoms (including fever, cough, sore throat, etc.)
4. treatments (including antibiotics, intubation, etc.)
5. standard labs (including lymphocyte count, platelets, etc.)
6. outcomes (including discharge, hospital length of stay, death, etc.)

The data is observational, and hence, making causal conclusions is not advisable.

Reference : https://covidanalytics.io/dataset_documentation

Problem Statement

Check how well can we build and fit a model to predict whether a person with heart, lunch, kidney or liver disease, along with diabetes, is more likely to die after contracting covid.

2

In the problem that we are exploring, we'll be looking at the following variables –

Response variable: Mortality

Explanatory variables:

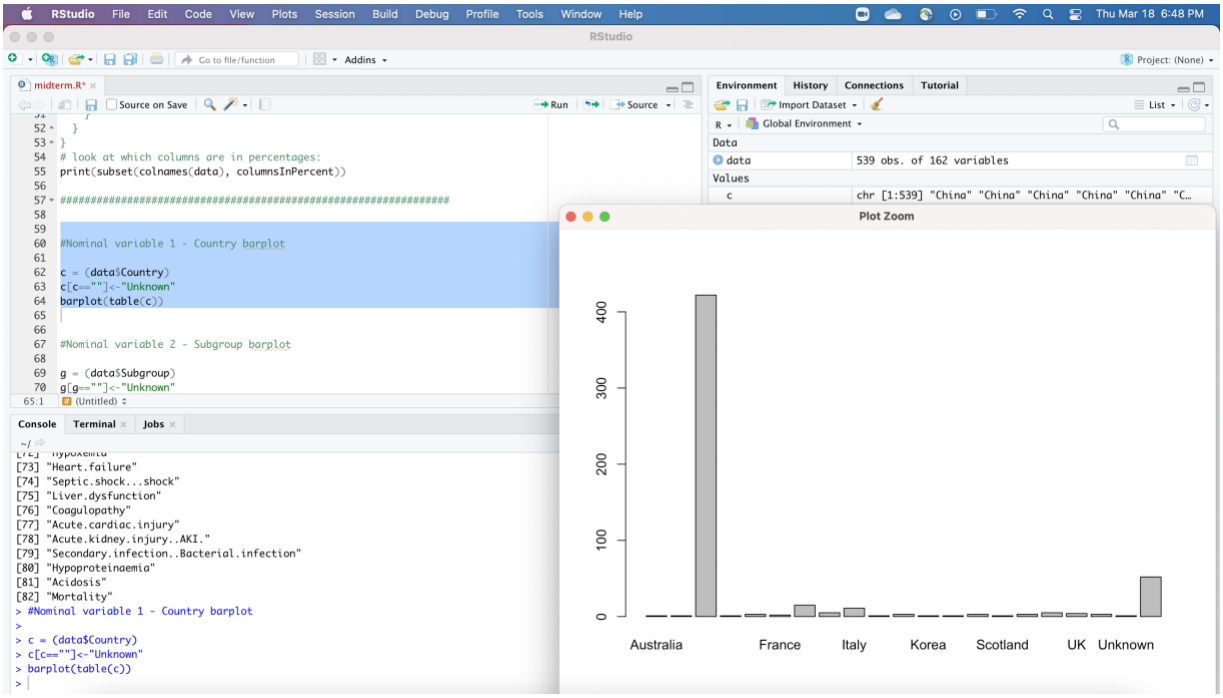
Diabetes
Cardiovascular Disease (incl. CAD)
Chronic obstructive lung (COPD)
Chronic kidney/renal disease

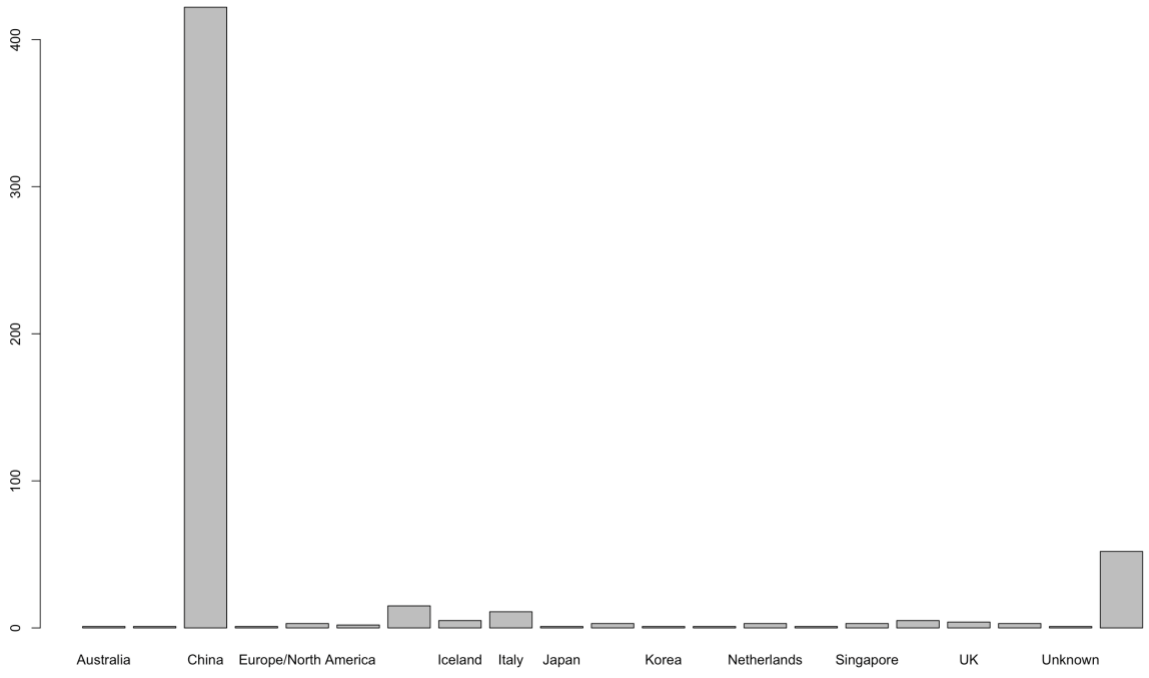
Part 1 and 2 – Summarization

PROMPT 1 – PLOTS

1. Categorical Nominal Variable 1 : Country barplot

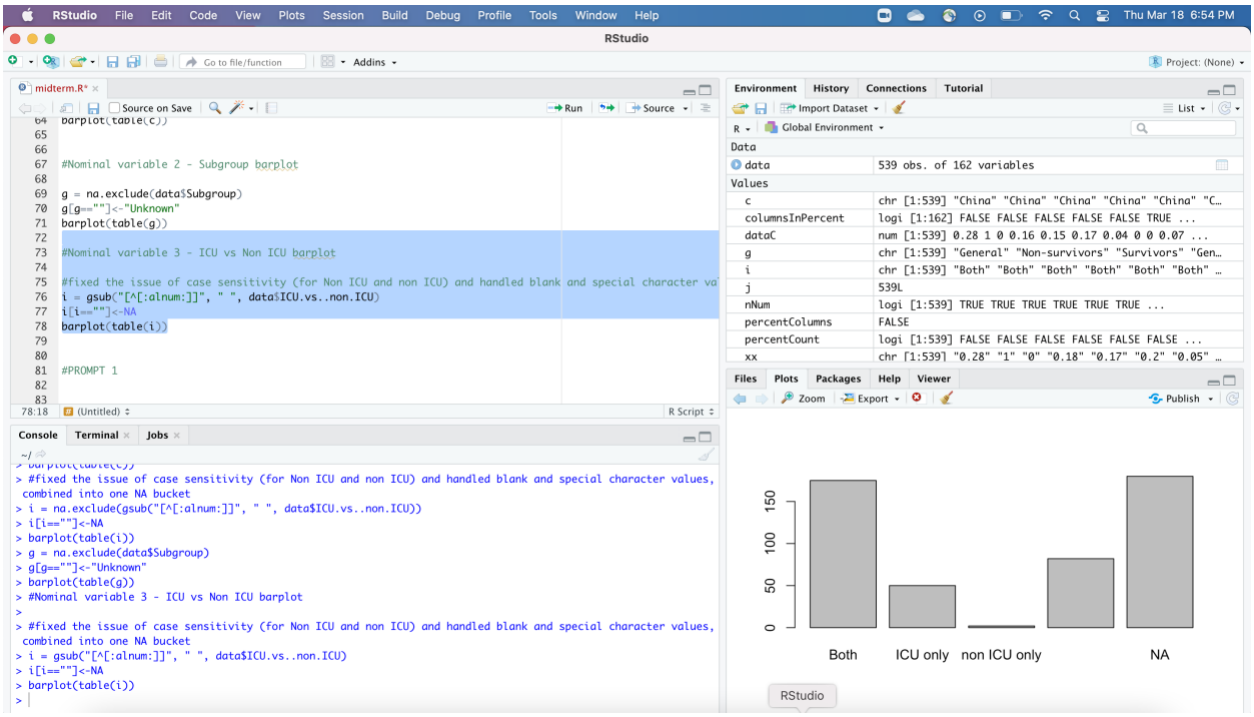
Used logic to replace blank with “Unknown”

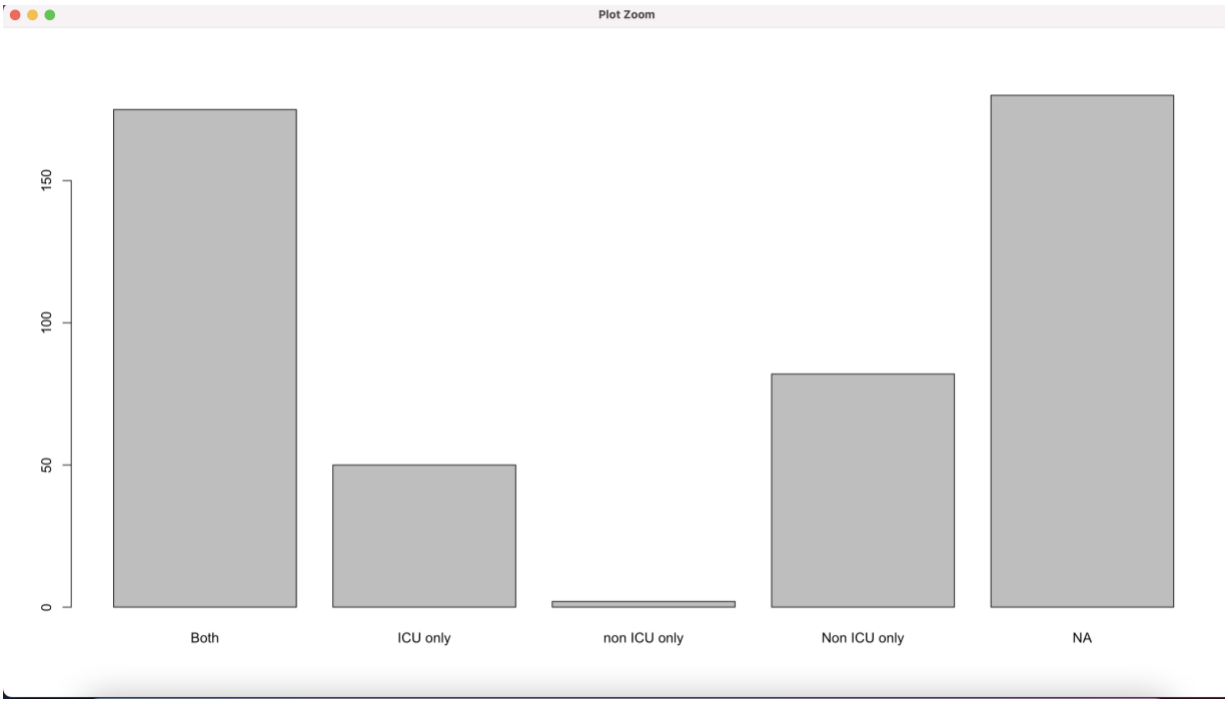




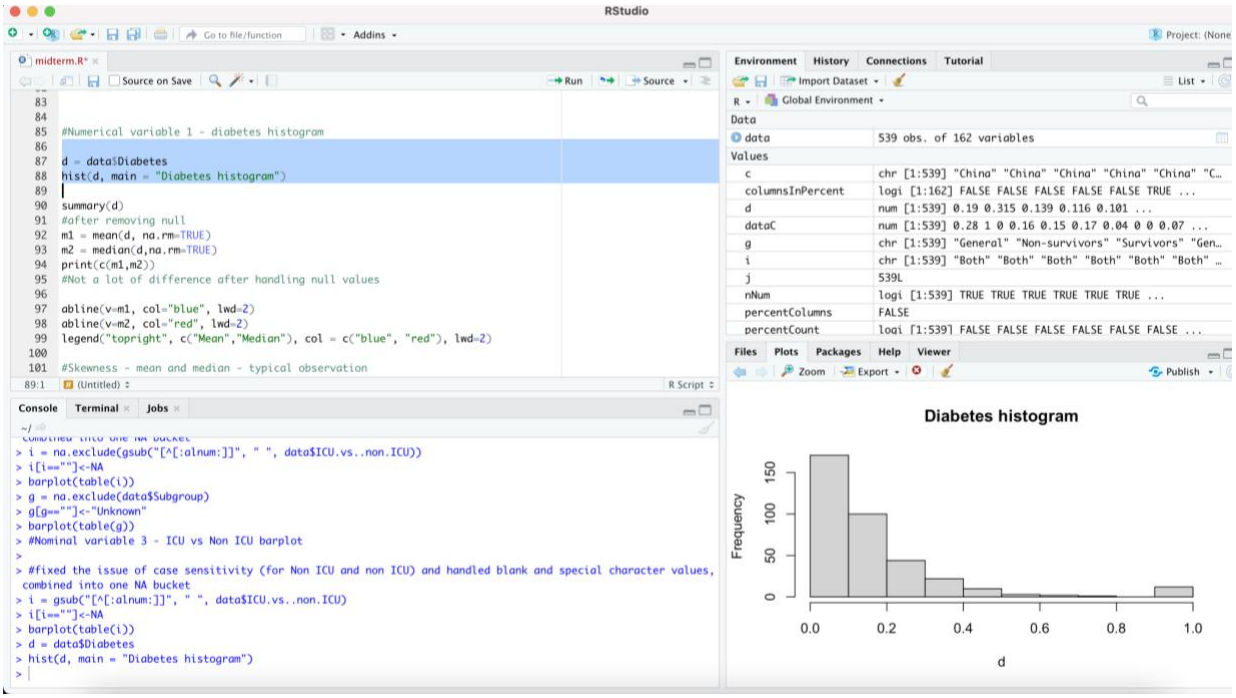
2. Categorical Nominal Variable 2 : ICU vs Non-ICU barplot

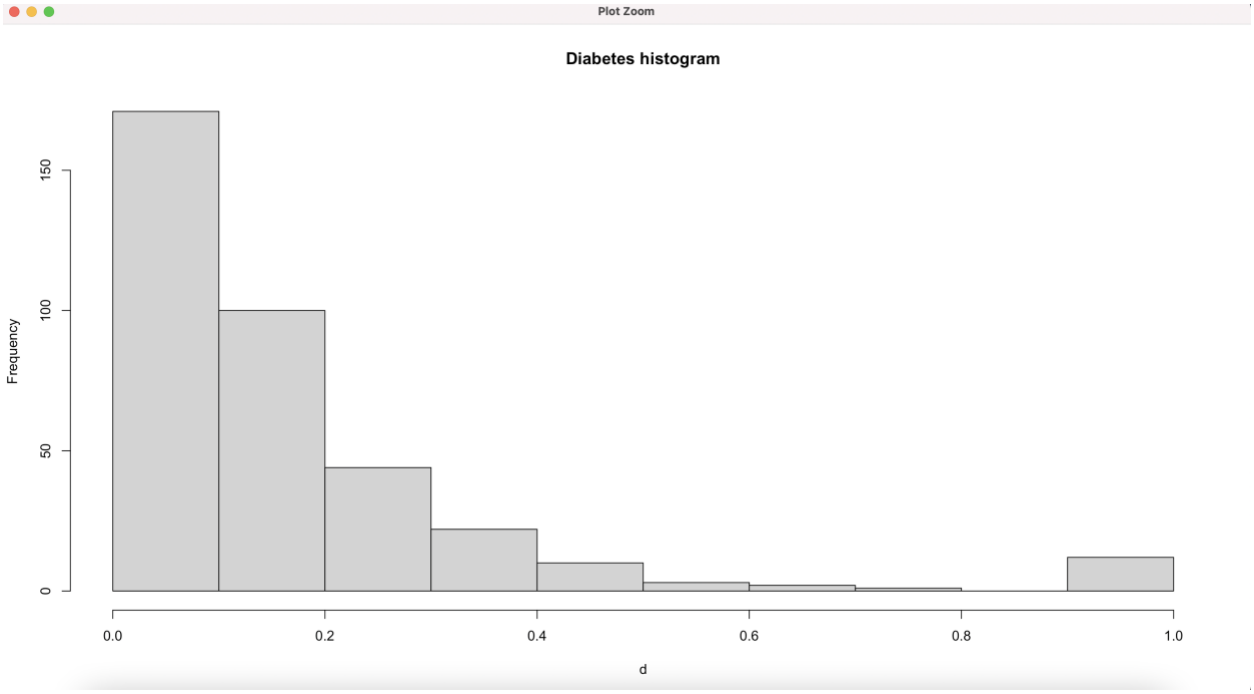
Fixed the issue of case sensitivity (for Non-ICU and non ICU) and handled blank and special character values, combined into one NA bucket. Learned about the gsub function.





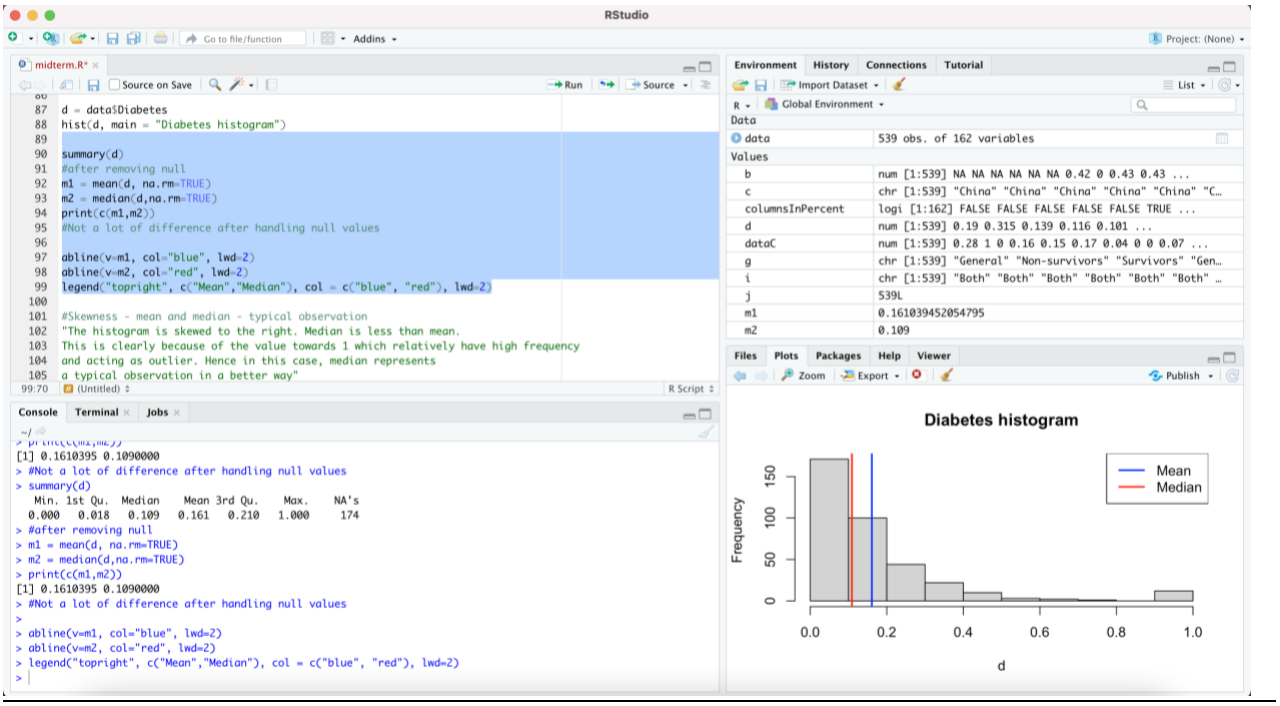
3. Numerical Variable 1 : Diabetes histogram

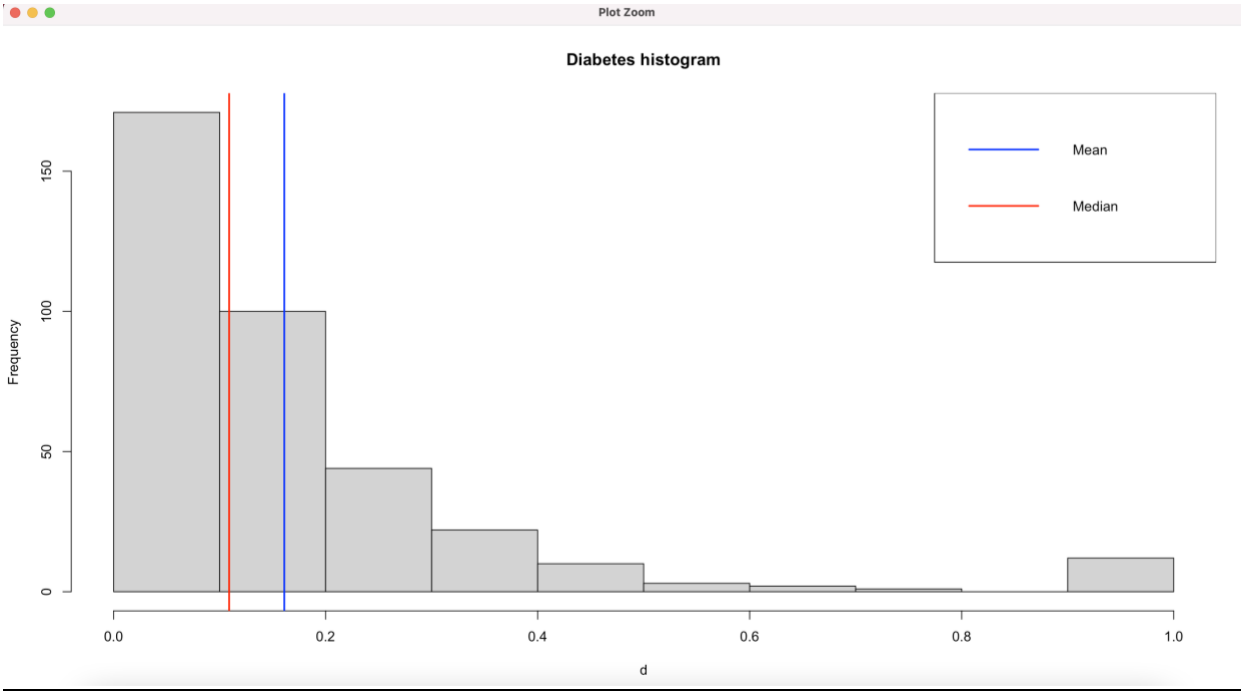




Mean, Median and Skew

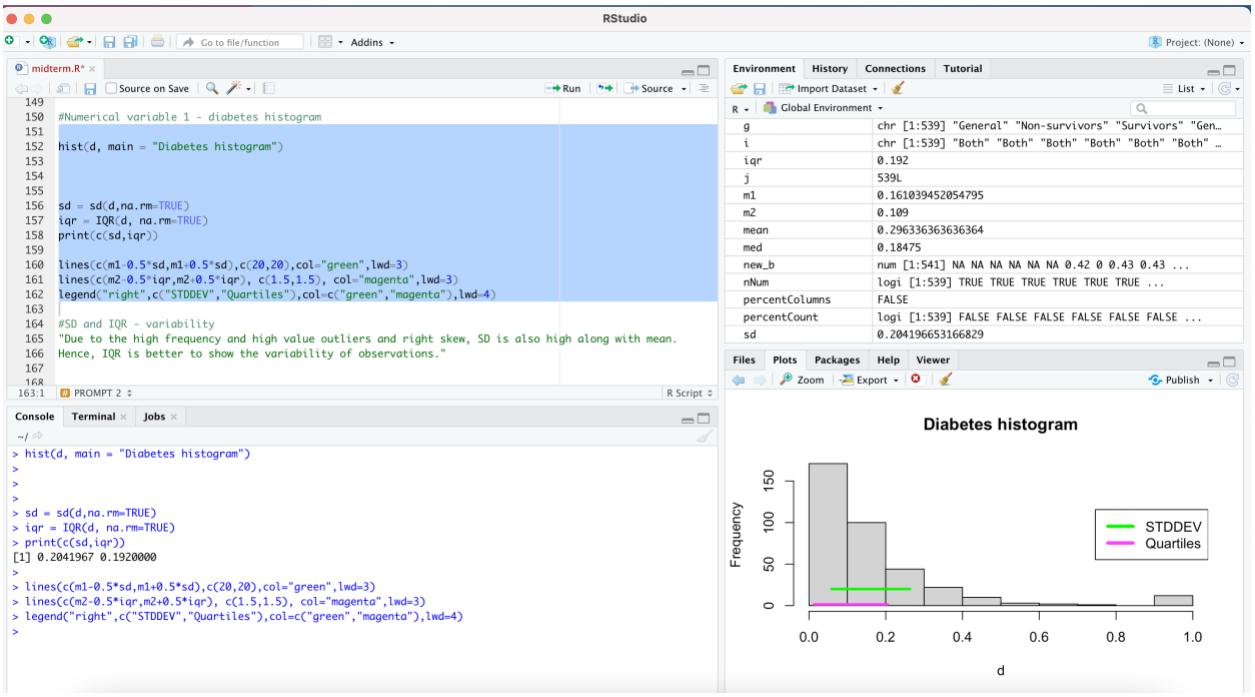
Note that there is no difference in mean and median when I take summary vs after removing na (na.rm = TRUE)

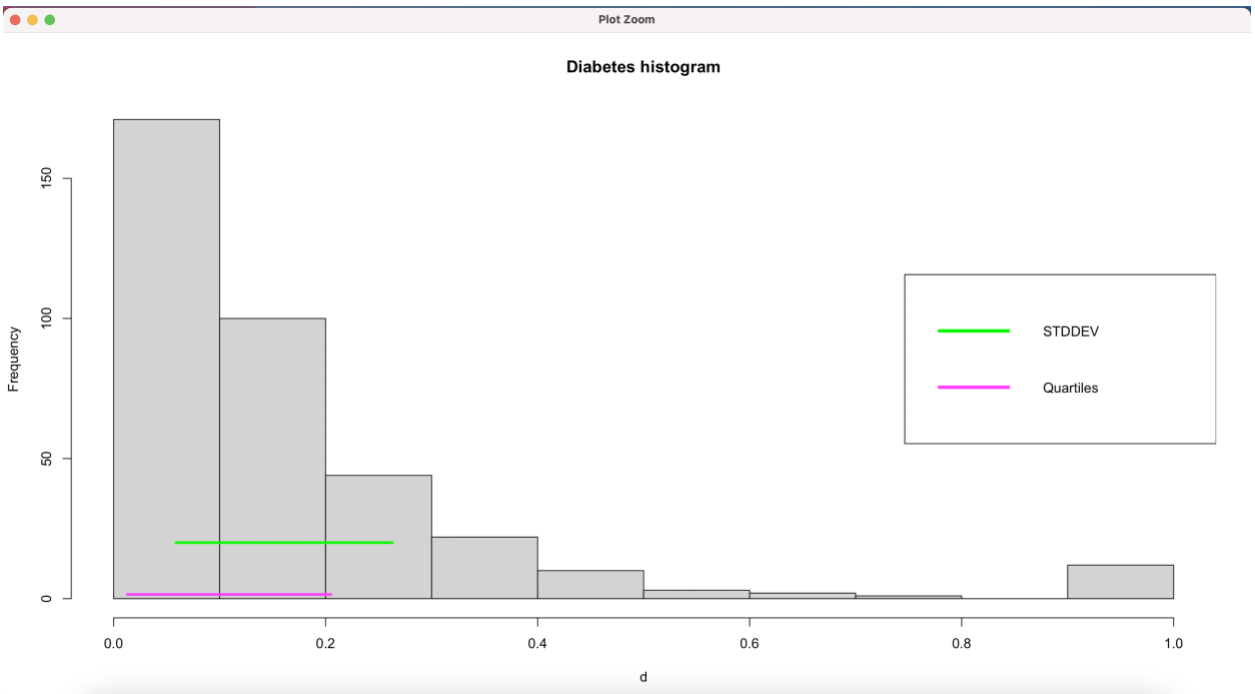




Comments –
The histogram is skewed to the right since Median is less than mean. (0.109 and 0.16 resp.) This is clearly because of the value towards 1 with relatively high frequency that is acting as outlier. Hence in this case, median represents a typical observation in a better way

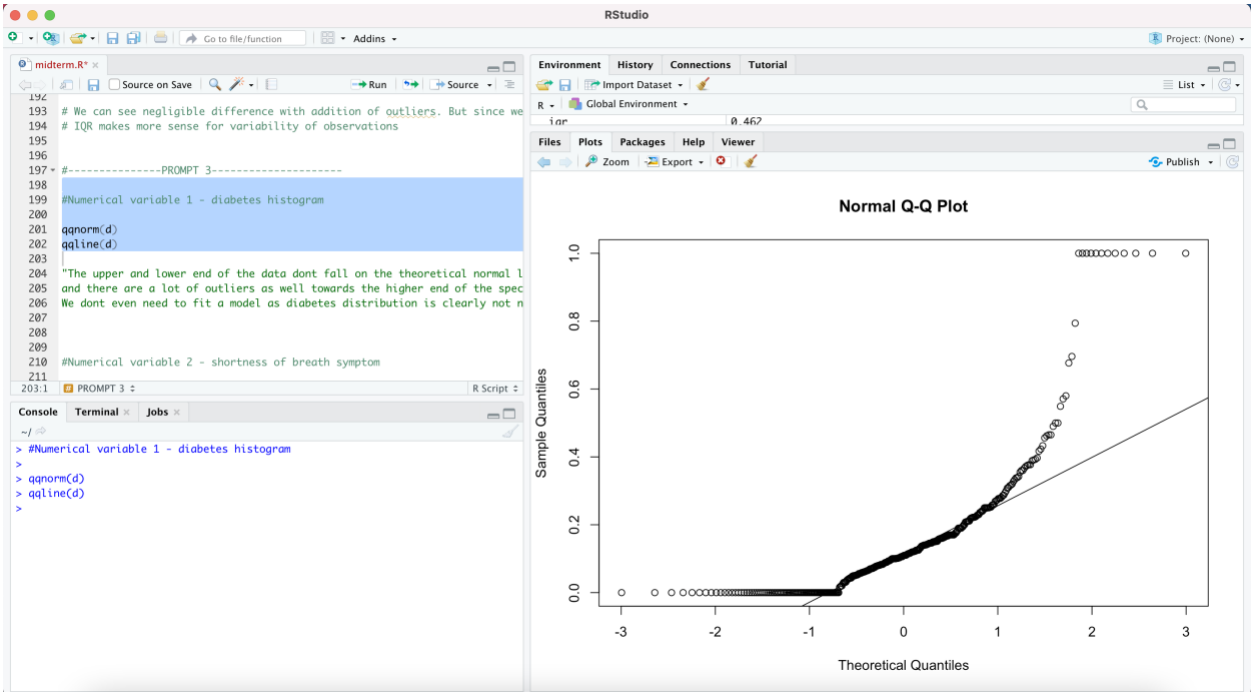
SD and IQR

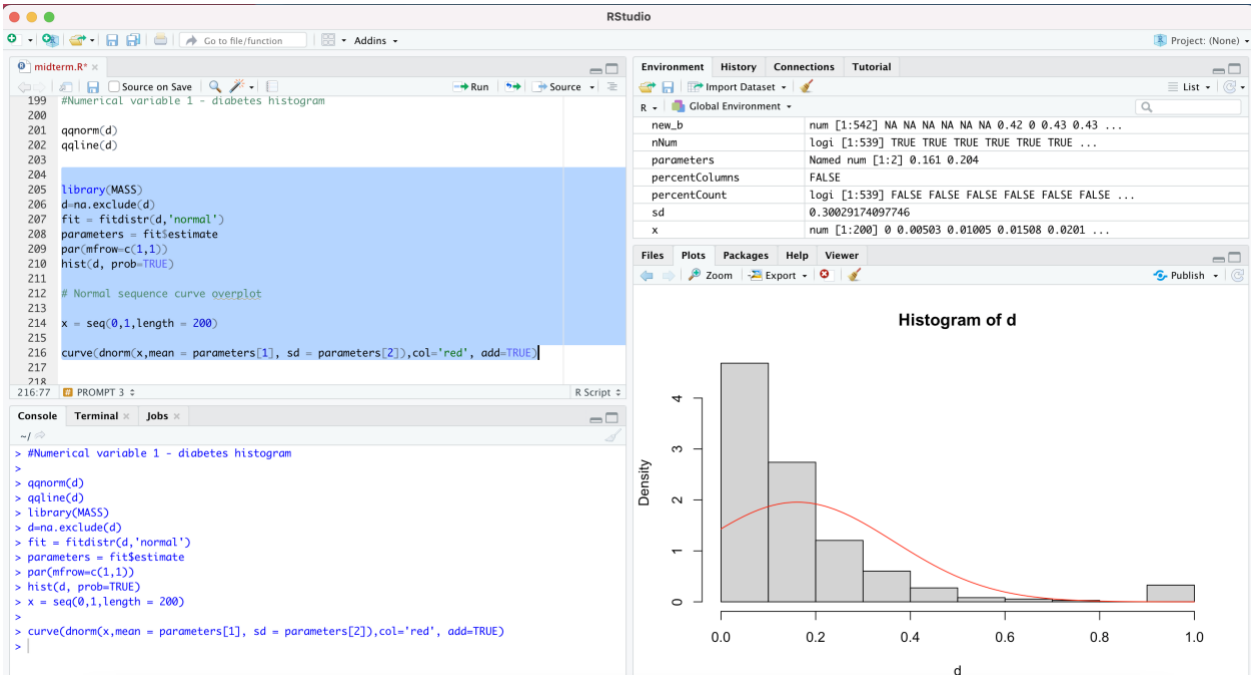


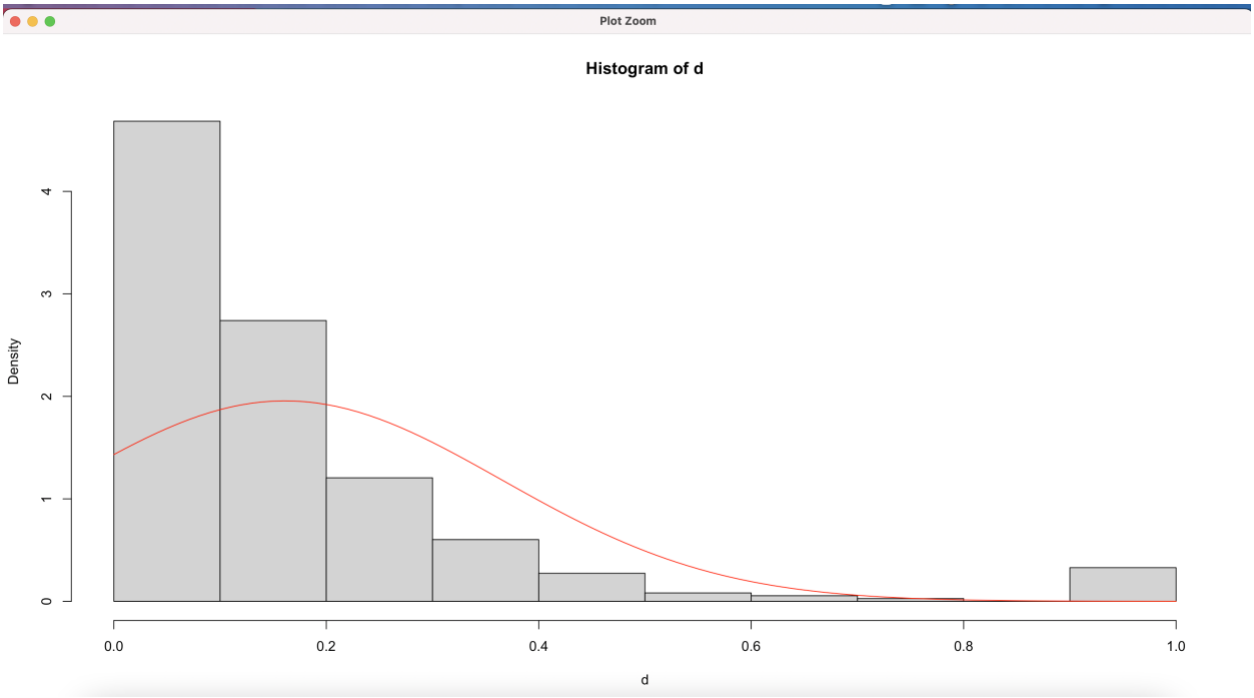


Comments - Due to the high frequency and high value outliers and right skew, SD is also high along with mean. (0.204). Hence, IQR is better to show the variability of observations. (0.192)

Check for Normality

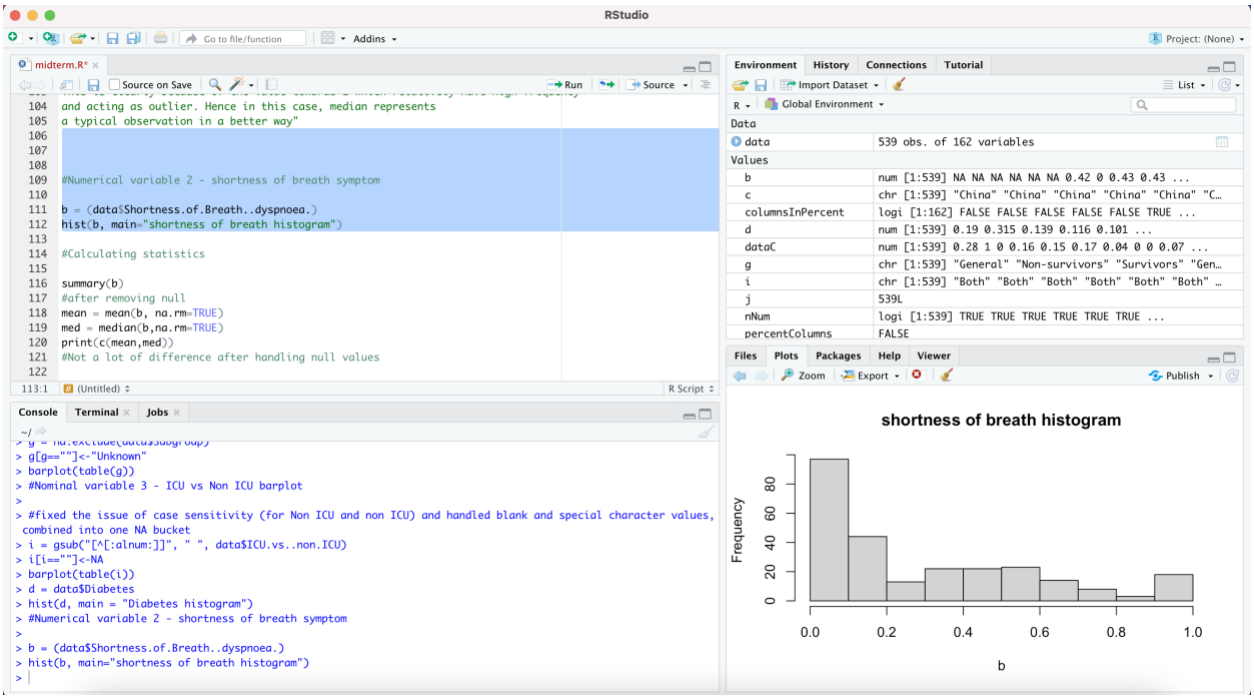


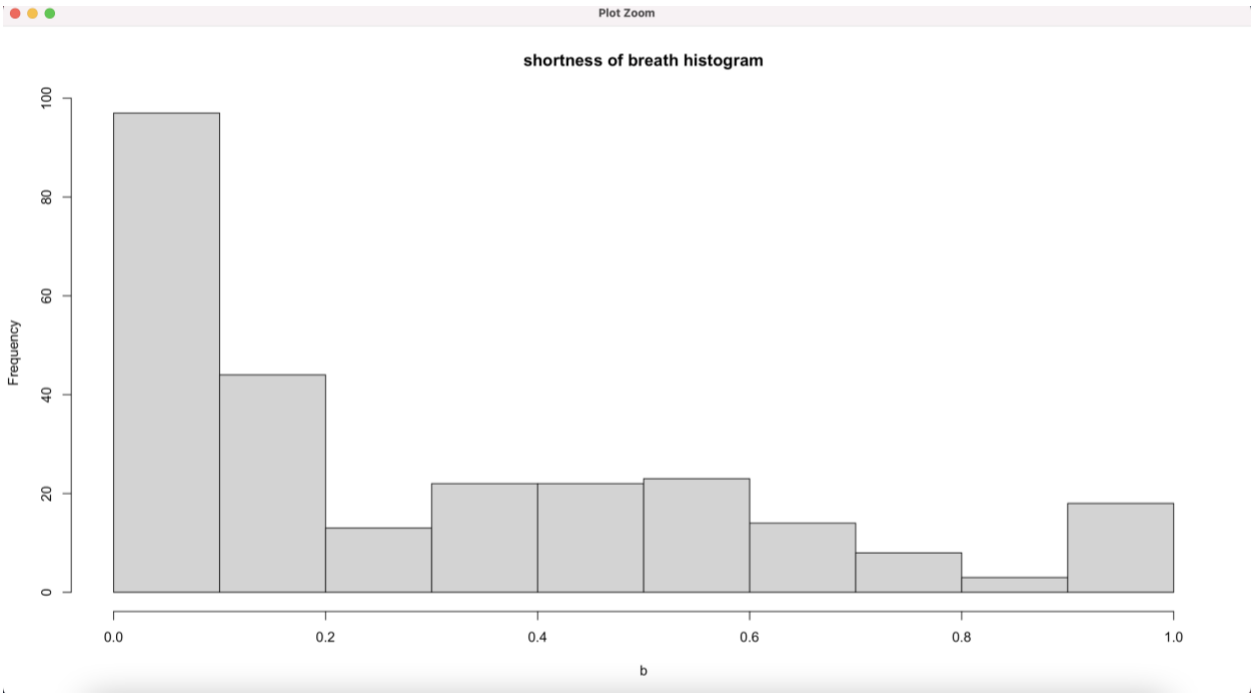




Comments - The upper and lower end of the data don't fall on the theoretical normal line and there are a lot of outliers as well towards the higher end of the spectrum.
We didn't even need to fit a model as diabetes distribution is clearly not normal. However, I used the **na.exclude** function and did fit a model over the distribution and we can see all the values that go over the bell curve.

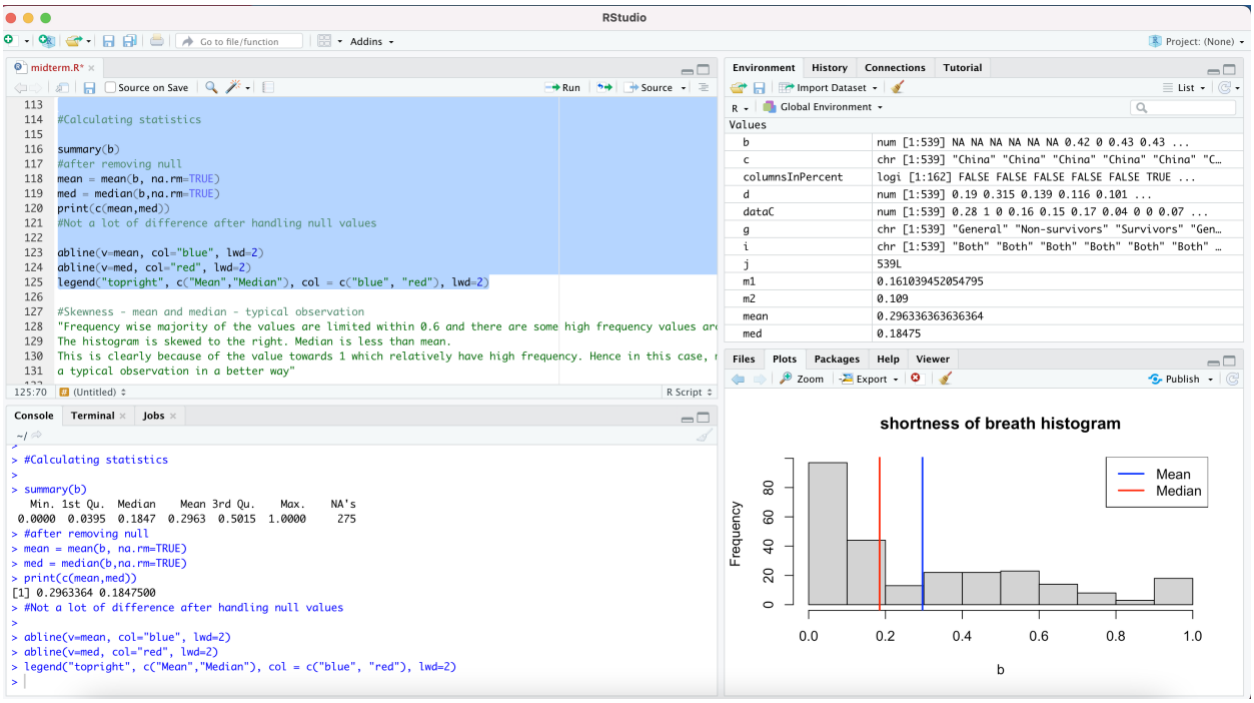
4. Numerical Variable 2 : Shortness of Breath histogram

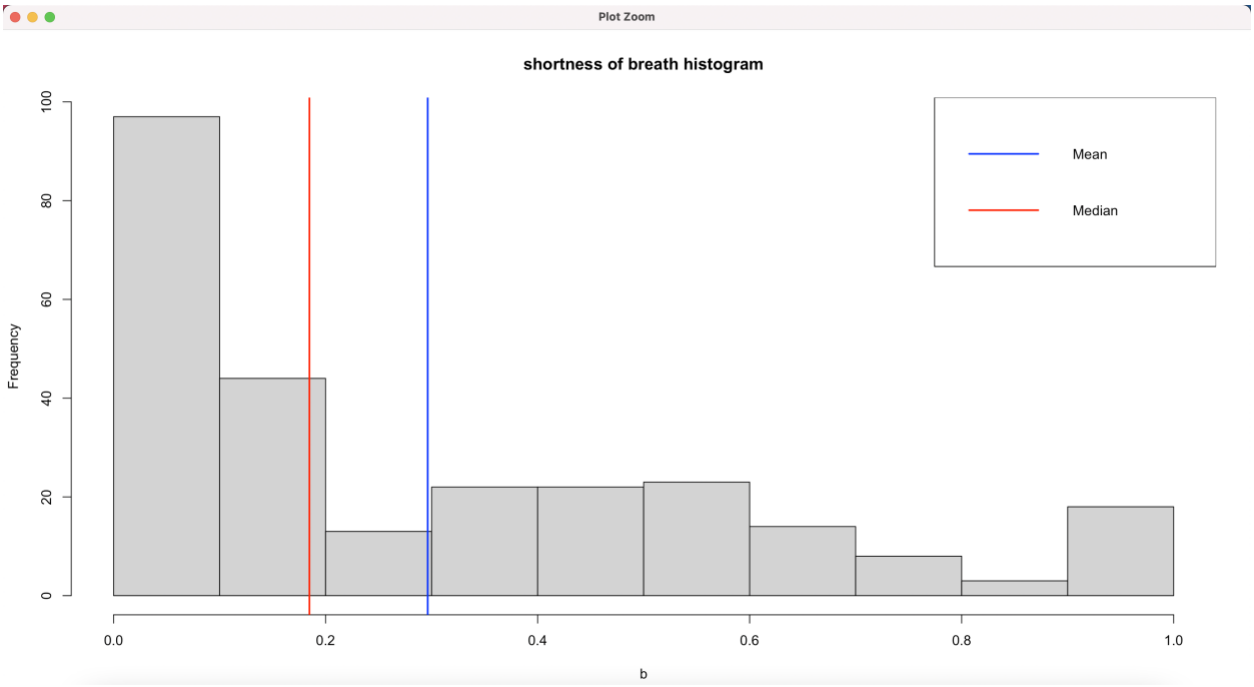




Mean, Median and Skew

Note that there is no difference in mean and median when I take summary vs after removing na (na.rm = TRUE)

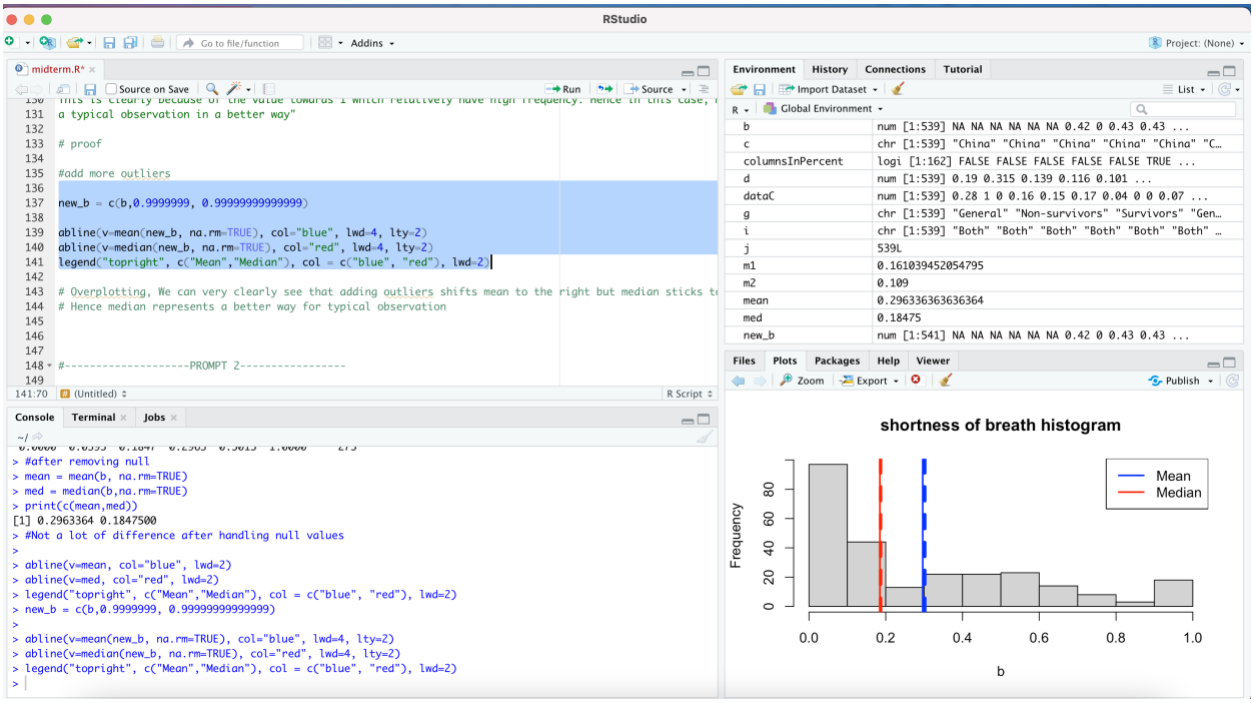


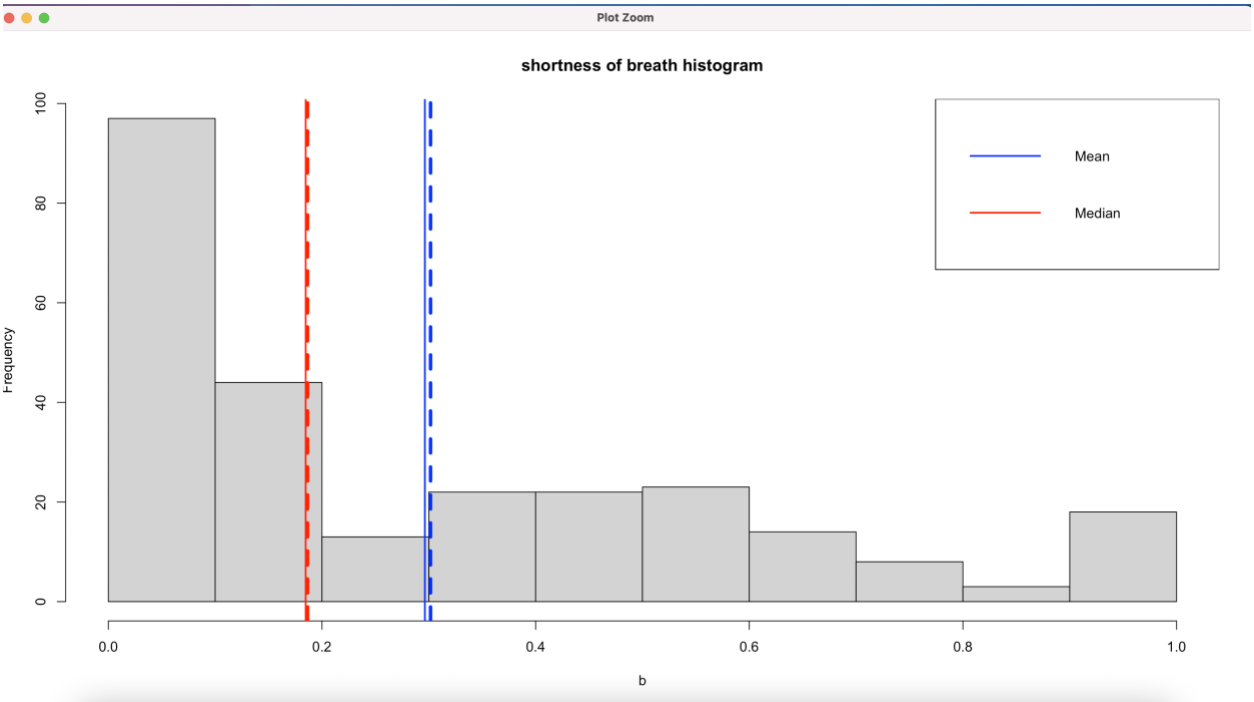


Comments –

Frequency wise majority of the values are limited within 0.6 and there are some high frequency values around 1. The histogram is skewed to the right since Median is less than mean.

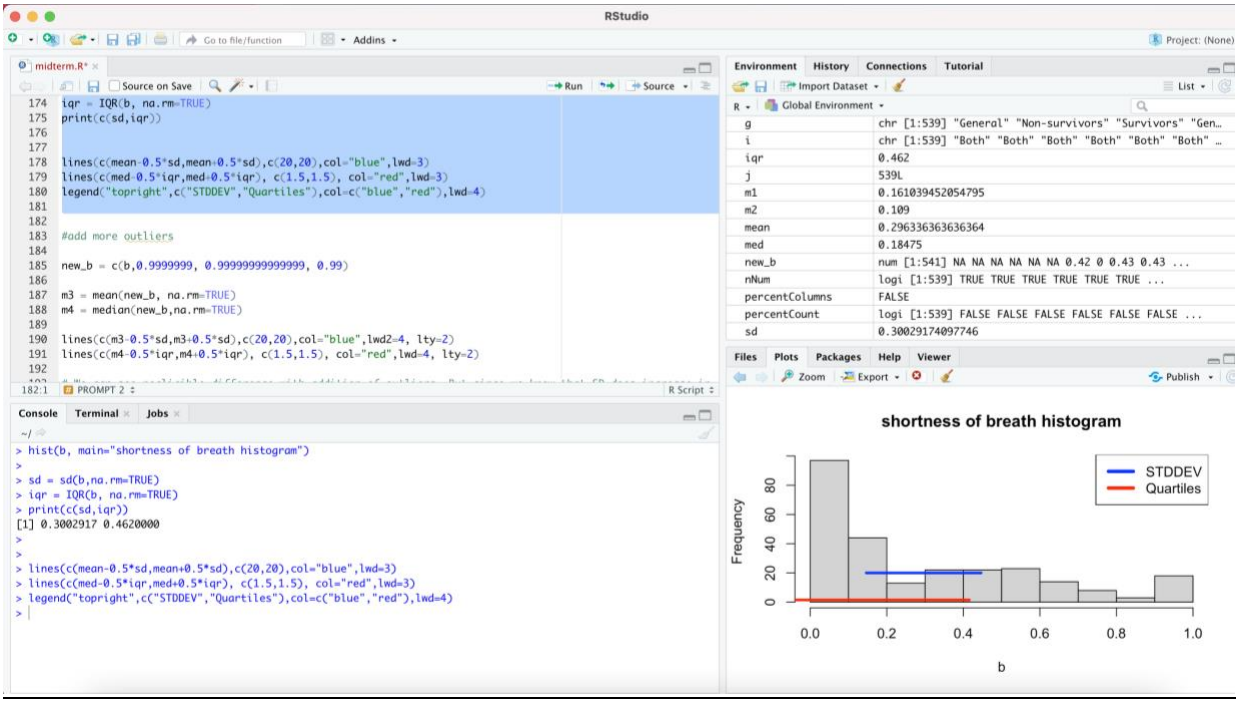
This is clearly because of the value towards 1 which relatively have high frequency. Hence in this case, median represents a typical observation in a better way. But we have done a more robust test for this variable by adding 2 more outliers and seeing the reaction of mean and median.

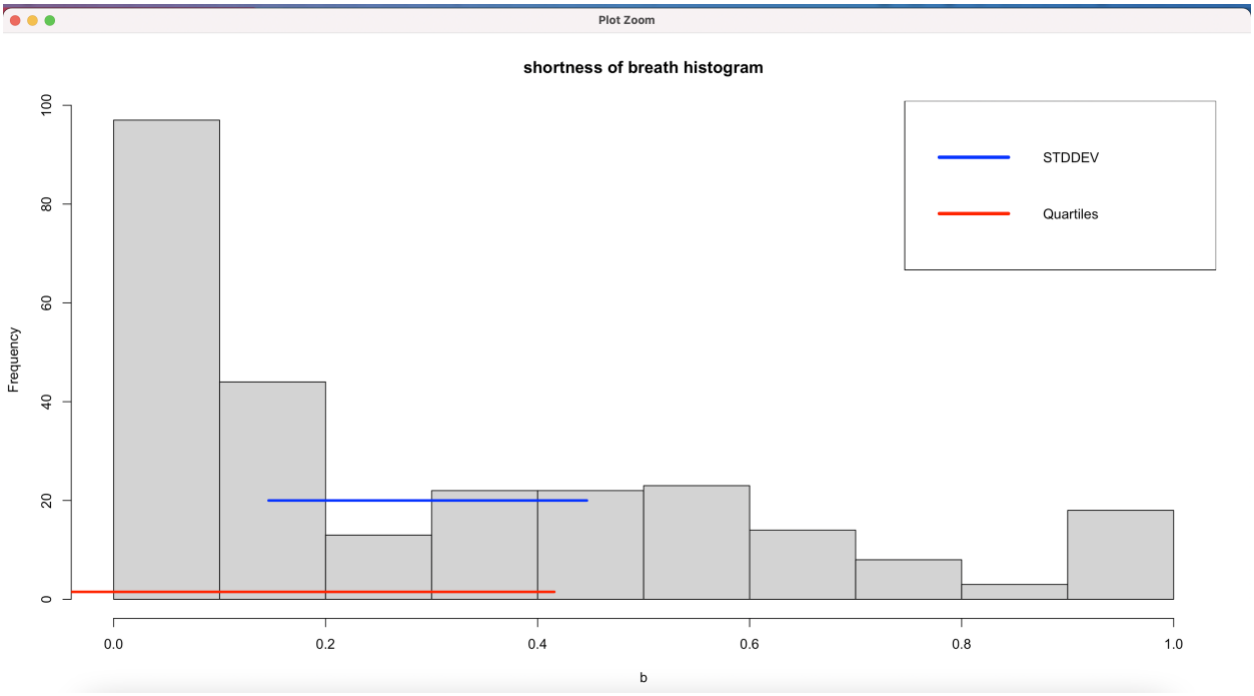




Overplotting, we can very clearly see that adding outliers shifts mean to the right but median sticks to the same.

SD and IQR





Comments - Due to the relatively high frequency numbers around 1 and the right skew, IQR is better to show the variability of observations.

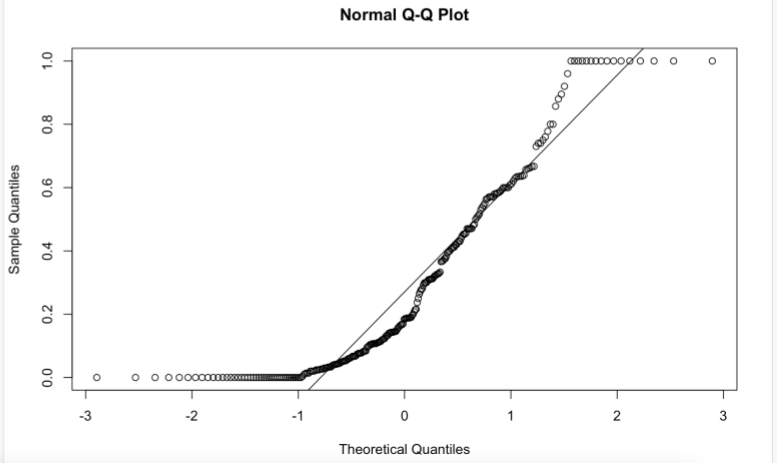
Check for Normality

RStudio

```
222 We dont even need to fit a model as diabetes distribution is clea
223
224
225
226 #Numerical variable 2 - shortness of breath symptom
227
228 qqnorm(b)
229 qqline(b)
230
231
232 library(MASS)
233 b=na.exclude(b)
234 fit = fitdistr(b,'normal')
235 parameters = fit$estimate
236 par(mfrow=c(1,1))
237 hist(b, prob=TRUE)
238
239 # Normal sequence curve overplot
240
```

```
> par(mfrow=c(1,1))
> hist(d, prob=TRUE)
> x = seq(0,1,length = 200)
>
> curve(dnorm(x,mean = parameters[1], sd = parameters[2]),col='red', add
=TRUE)
> qqnorm(b)
> qqline(b)
> qqnorm(d)
> qqline(d)
> qqnorm(b)
> qqline(b)
> b = (data$Shortness.of.Breath..dyspnoea.)
> qqnorm(b)
> qqline(b)
>
```

new_b	num [1:542] NA NA NA NA NA NA 0.42 0 0.43 0.43 ...
nNum	logi [1:539] TRUE TRUE TRUE TRUE TRUE TRUE ...

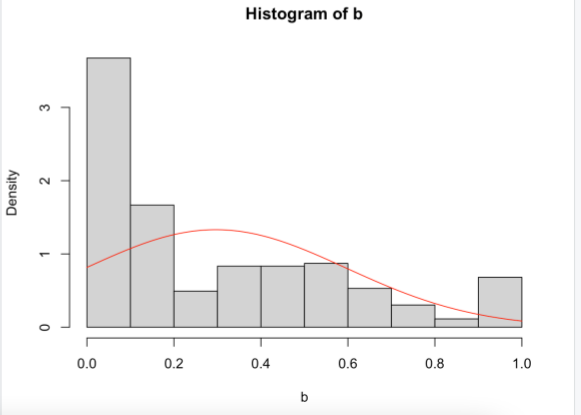


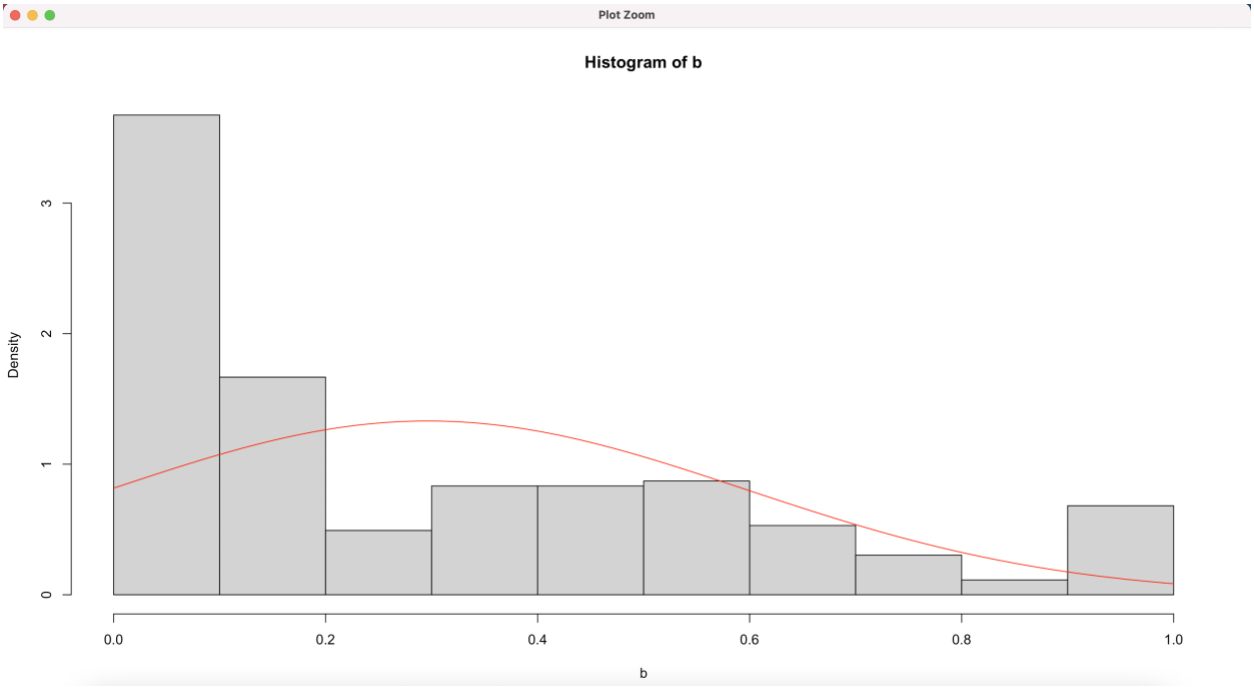
RStudio

```
229 qqline(b)
230
231
232 library(MASS)
233 b=na.exclude(b)
234 fit = fitdistr(b,'normal')
235 parameters2 = fit$estimate
236 par(mfrow=c(1,1))
237 hist(b, prob=TRUE)
238
239 # Normal sequence curve overplot
240
241 x = seq(0,1,length = 200)
242
243 curve(dnorm(x,mean = parameters2[1], sd = parameters2[2]),col='red', add=TRUE)
244
245 "Both the upper and lower end of the distribution has outliers that dont
246 fall on the theoretical normal line. Hence, shortness of breath distribution
247 is not normal"
```

```
> parameters2 = fit$estimate
> par(mfrow=c(1,1))
> hist(b, prob=TRUE)
> library(MASS)
> b=na.exclude(b)
> fit = fitdistr(b,'normal')
> parameters2 = fit$estimate
> par(mfrow=c(1,1))
> hist(b, prob=TRUE)
>
> # Normal sequence curve overplot
>
> x = seq(0,1,length = 200)
>
> curve(dnorm(x,mean = parameters2[1], sd = parameters2[2]),col='red', add=TRUE)
>
```

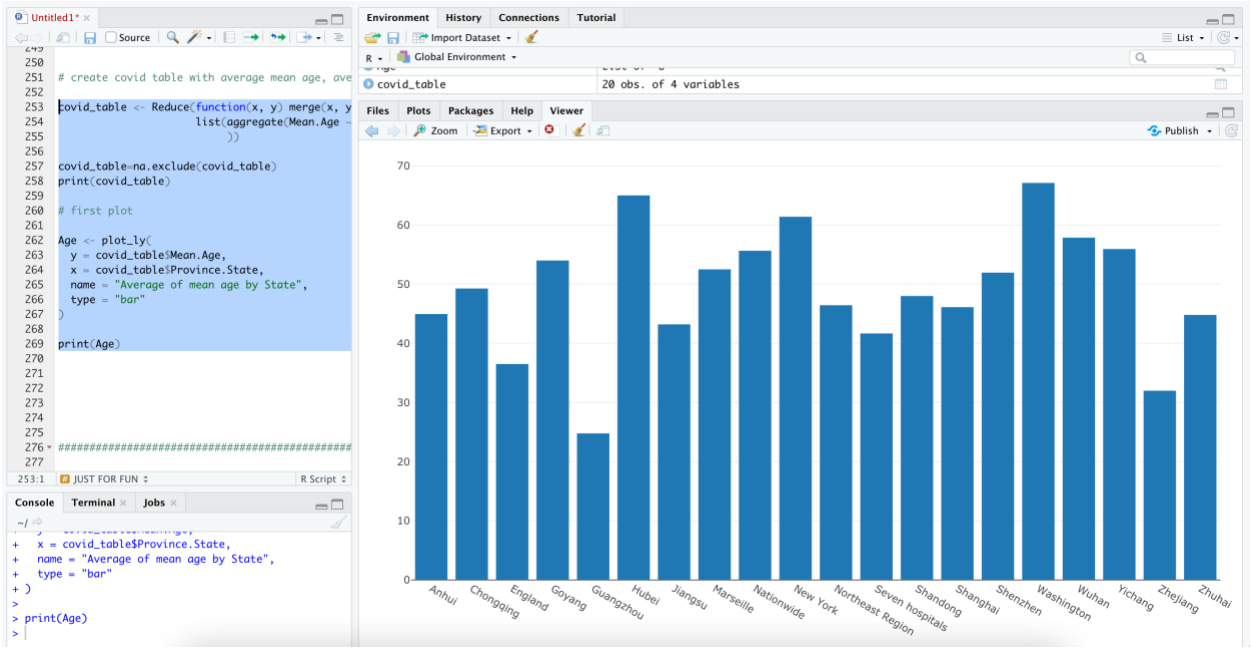
parameters2	Named num [1:2] 0.296 0.3
percentColumns	FALSE
percentCount	logi [1:539] FALSE FALSE FALSE FALSE FALSE FALSE ...
sd	0.30029174097746
x	num [1:200] 0 0.00503 0.01005 0.01508 0.0201 ...





Comments – Same as the diabetes variable, the upper and lower end of the data don't fall on the theoretical normal line. (see QQ chart)
After having used the **na.exclude** function and fitting a model over the distribution, we can clearly see all the values that are undercut by the bell curve.

 # Plotly plot – average age of population by state



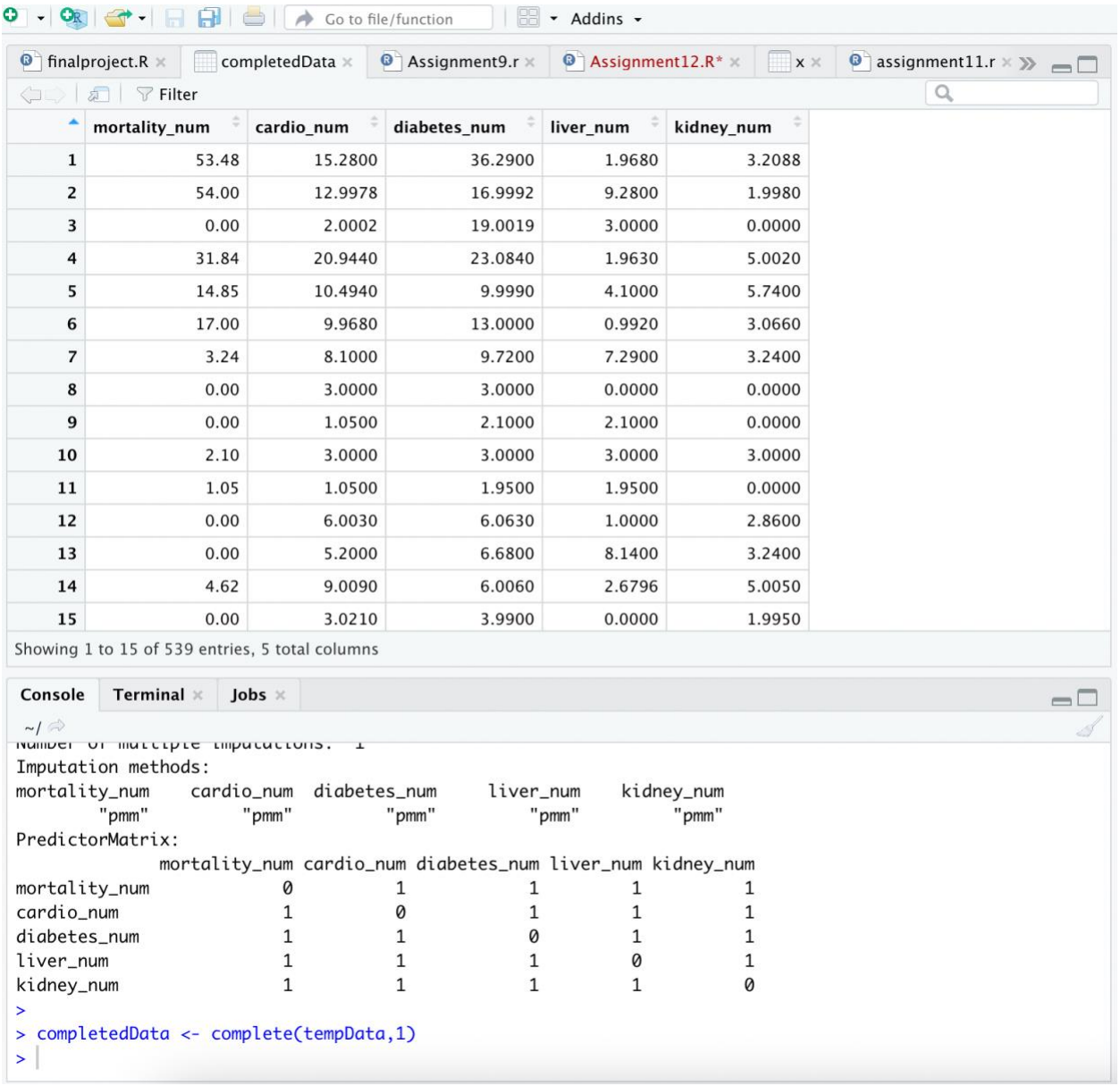
PART 3

For our problem statement, using a statistical model is very important to predict the mortality rate based on the % of diabetes, COPD, and CAD and this would answer a very important medical question and also help in prioritization if the statistical prediction is accurate. If modelled well, the applications are immensely useful. We opted for MLR because of multiple explanatory parameters.

We tried to incorporate the feedback by professor about proportional dependent variable and researched on it.
Per research paper https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf,

Beta regression uses logit to transform a mean of distribution assumed for data (beta distribution in this case) while linear regression with logit-transformed dependent variable transforms a data. So, in beta regression we have $logit(E(y))$ modelled while in linear regression with logit-transformed dependent variable we have $E(logit(y))$. These two are not the same.

Ultimately though, to tackle this problem, we have multiplied all the below percentage variables with pop size to take out the absolute number of mortal people, diabetic people, etc. (see image)



finalproject.R | completedData | Assignment9.r | Assignment12.R* | x | assignment11.r

Filter

	mortality_num	cardio_num	diabetes_num	liver_num	kidney_num
1	53.48	15.2800	36.2900	1.9680	3.2088
2	54.00	12.9978	16.9992	9.2800	1.9980
3	0.00	2.0002	19.0019	3.0000	0.0000
4	31.84	20.9440	23.0840	1.9630	5.0020
5	14.85	10.4940	9.9990	4.1000	5.7400
6	17.00	9.9680	13.0000	0.9920	3.0660
7	3.24	8.1000	9.7200	7.2900	3.2400
8	0.00	3.0000	3.0000	0.0000	0.0000
9	0.00	1.0500	2.1000	2.1000	0.0000
10	2.10	3.0000	3.0000	3.0000	3.0000
11	1.05	1.0500	1.9500	1.9500	0.0000
12	0.00	6.0030	6.0630	1.0000	2.8600
13	0.00	5.2000	6.6800	8.1400	3.2400
14	4.62	9.0090	6.0060	2.6796	5.0050
15	0.00	3.0210	3.9900	0.0000	1.9950

Showing 1 to 15 of 539 entries, 5 total columns

Console | Terminal | Jobs

```
Imputation methods:
mortality_num    cardio_num  diabetes_num     liver_num    kidney_num
        "pmm"         "pmm"         "pmm"         "pmm"         "pmm"
PredictorMatrix:
             mortality_num cardio_num diabetes_num liver_num kidney_num
mortality_num             0          1            1         1          1
cardio_num                1          0            1         1          1
diabetes_num              1          1            0         1          1
liver_num                 1          1            1         0          1
kidney_num                1          1            1         1          0
>
> completedData <- complete(tempData,1)
>
```

Having solved the proportion dependent variable problem, we'll continue fitting the MLR model using all 4 variables (as we check in Lasso later)

- Variables

  Response variable: Mortality

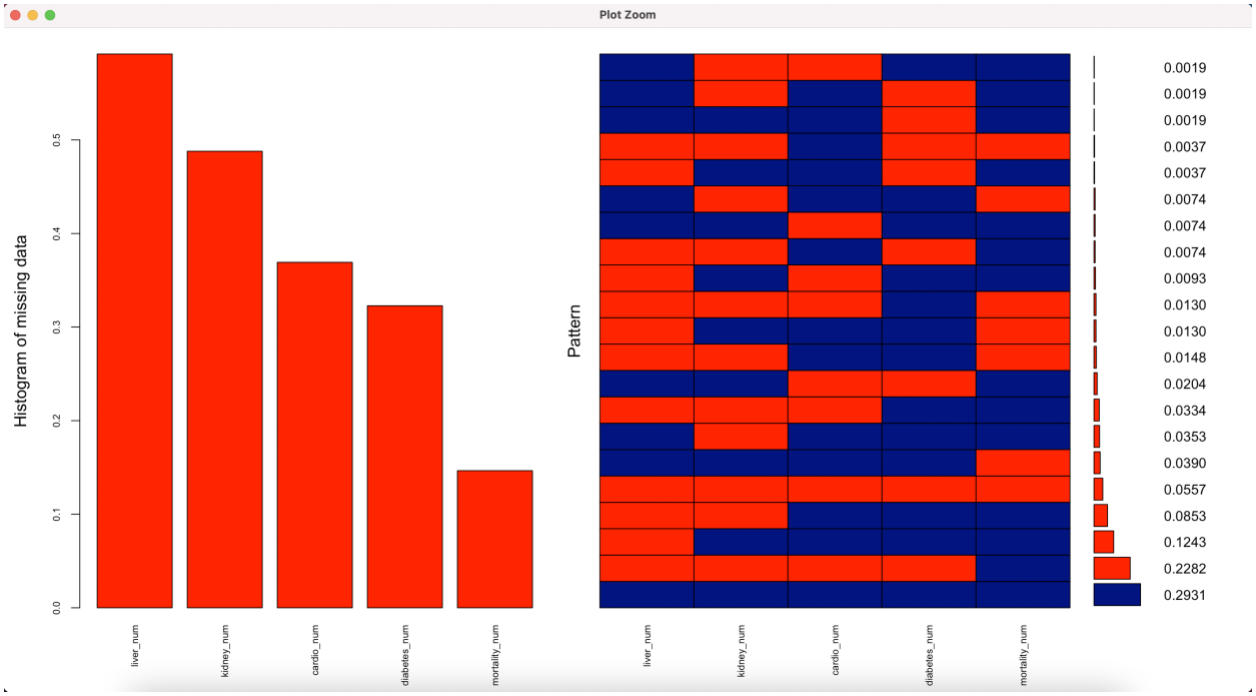
  Explanatory variables:
  Diabetes
  Cardiovascular Disease (incl. CAD)

Chronic obstructive lung (COPD)
Chronic kidney/renal disease

PART 4

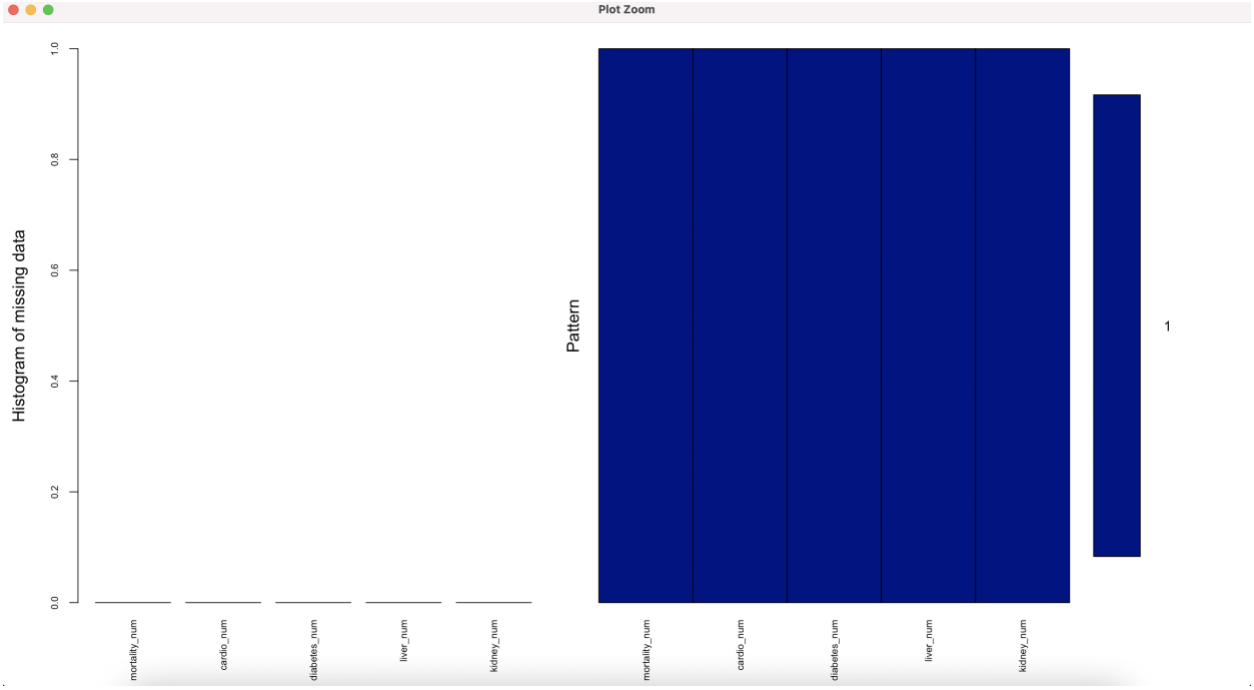
- **DATA CLEANING (Handling NA values)**

For the variables we have chosen in our model, below is the missing data pattern.

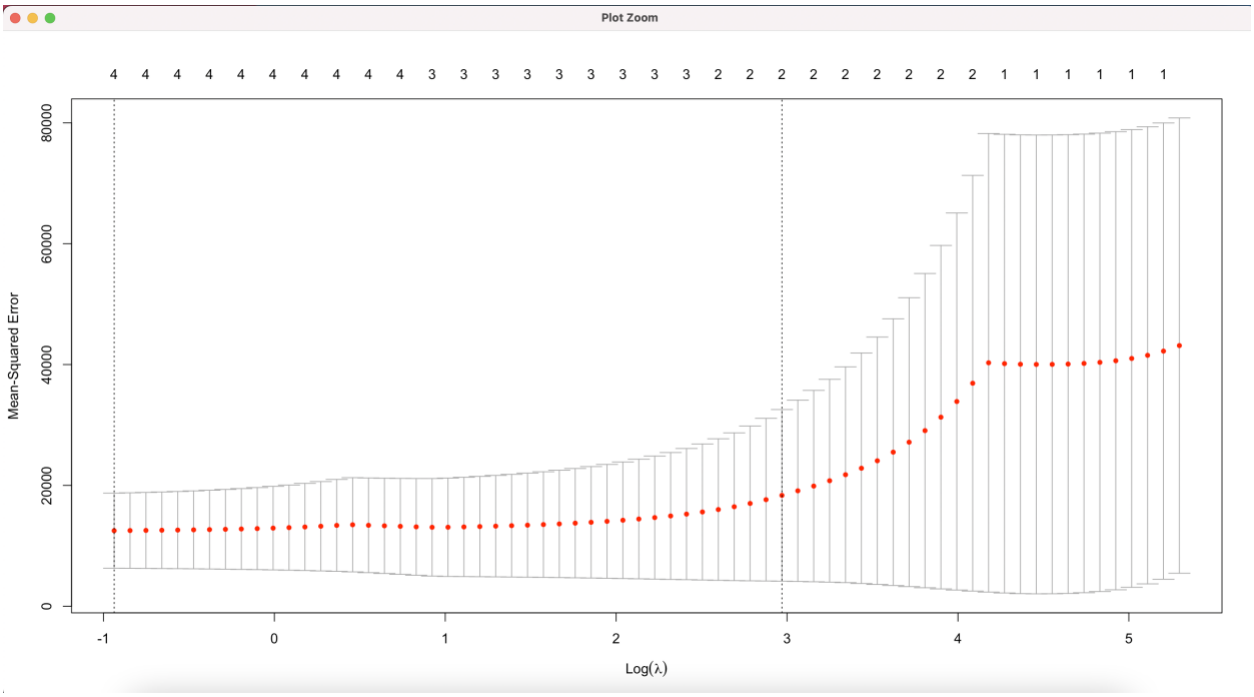


Per professor's feedback, we have utilized the mice package to impute the features and select **predictive mean matching** as the imputation method.

The new dataframe **completeData** has no missing values and this is where we'll fit our model on.



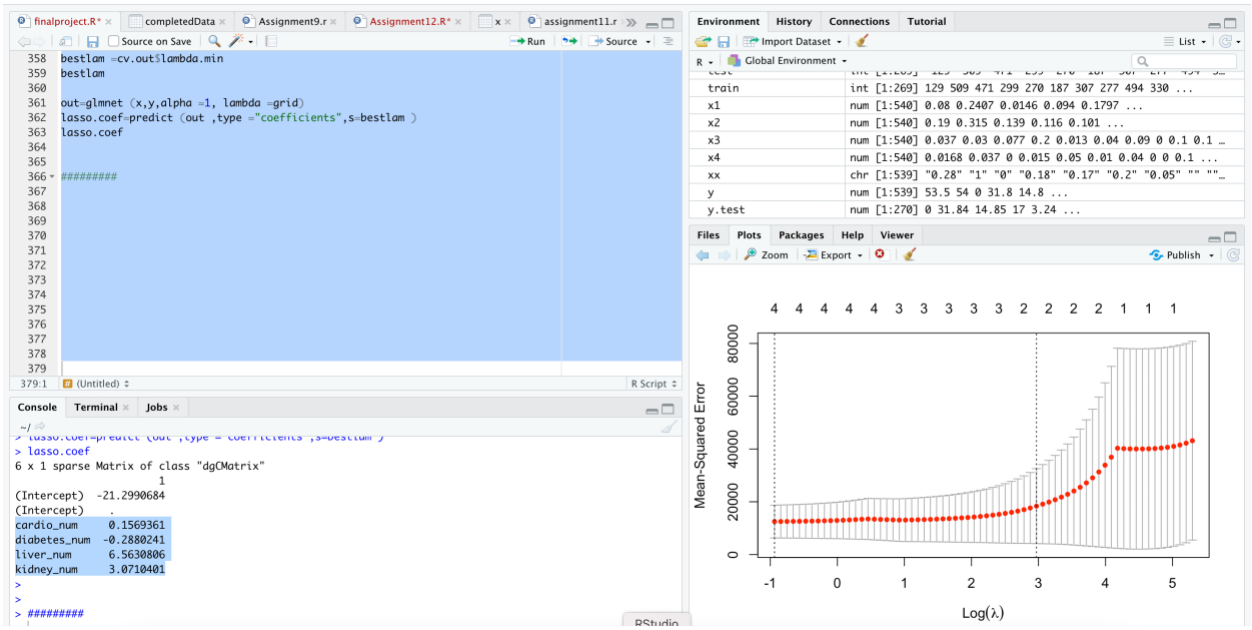
- **FINDING BEST PARAMETERS (using Lasso)**



We have applied Lasso regression to be able to regularize the collinearity of Mortality. Using best lambda value of 0.39158, we can see non-zero lasso coefficient values for all of the four explanatory variables. (see image below). (Diabetes_num has negative lasso coefficient)

cardio_num 0.1569361
diabetes_num -0.2880241
liver_num 6.5630806
kidney_num 3.0710401

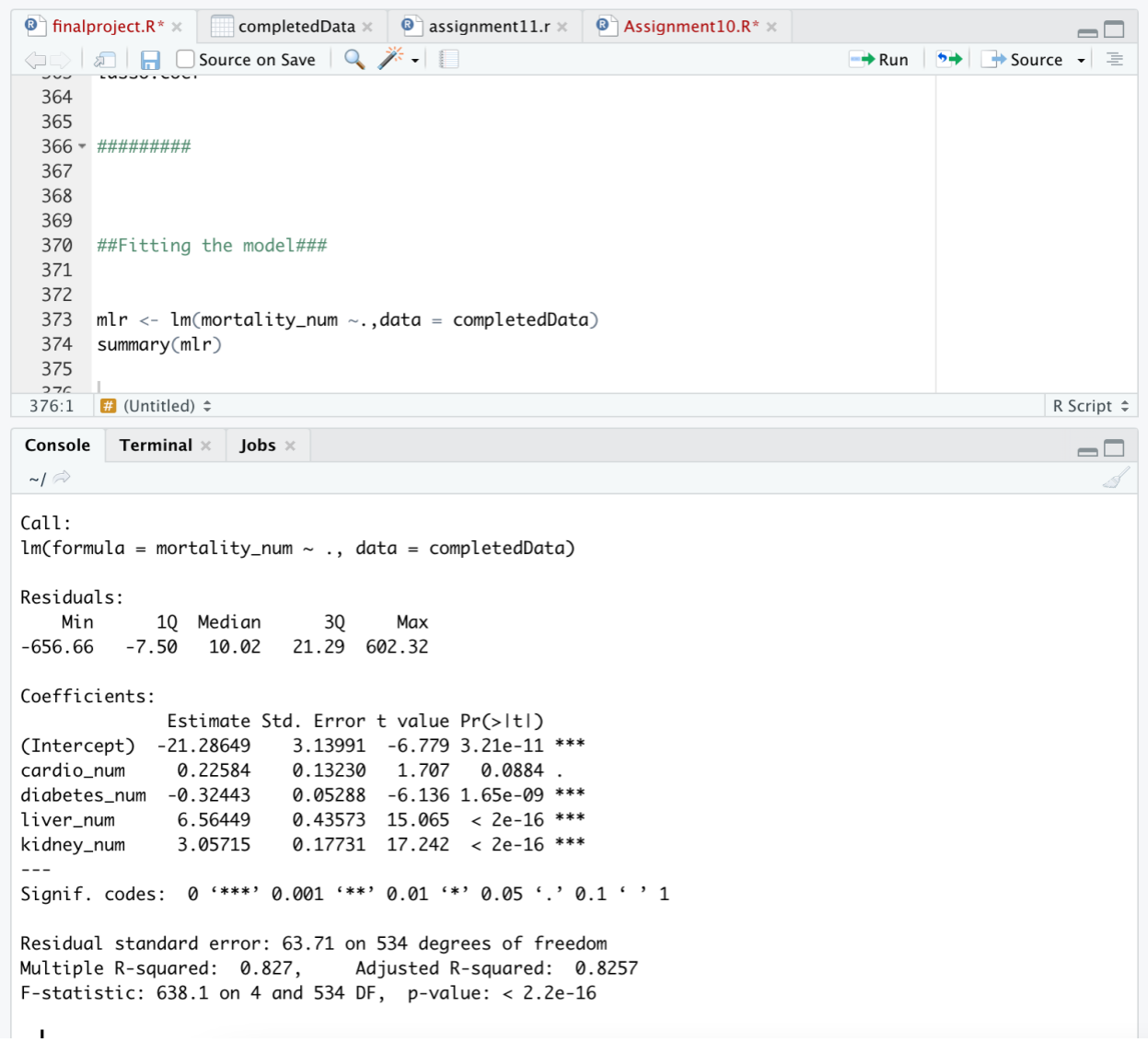
We can also conclude that the best parameters are ***liver_num*** and ***kidney_num***



- **SIGNIFICANCE OF RESULT**

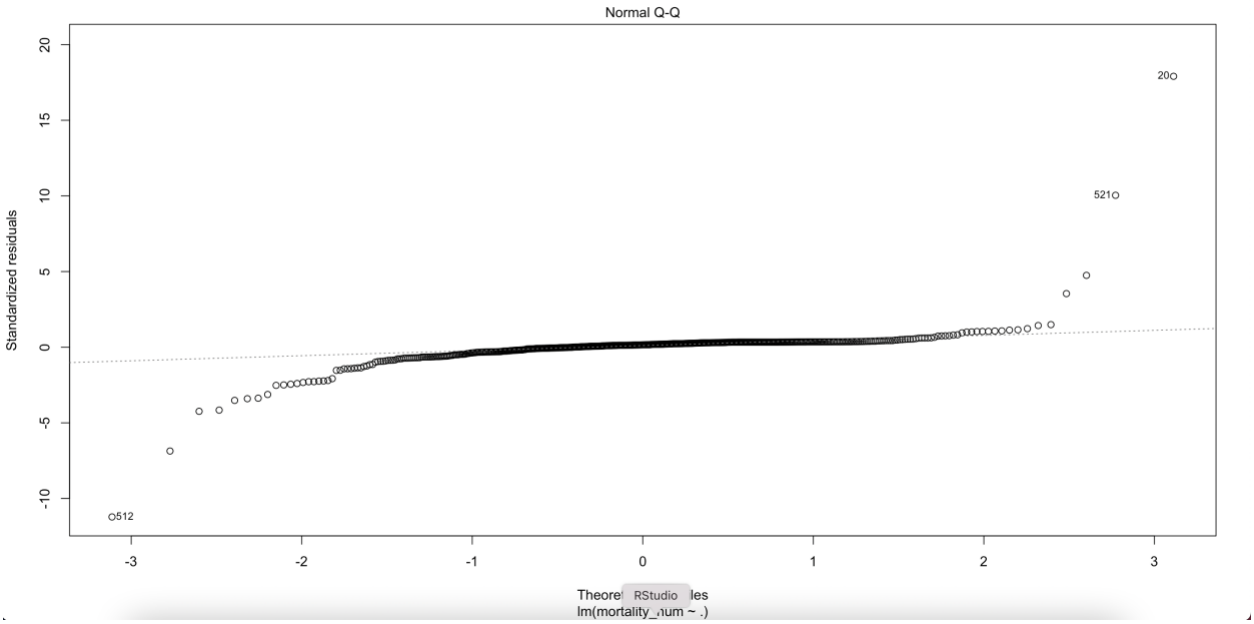
Fitting our MLR model (dependent variable mortality_num, explanatory variables diabetes_num, cardio_num, liver_num, kidney_num), we can see that the R-squared is almost **83%** which is good fit and the level of significance overall is also 2.2e-16, so it is statistically significant.

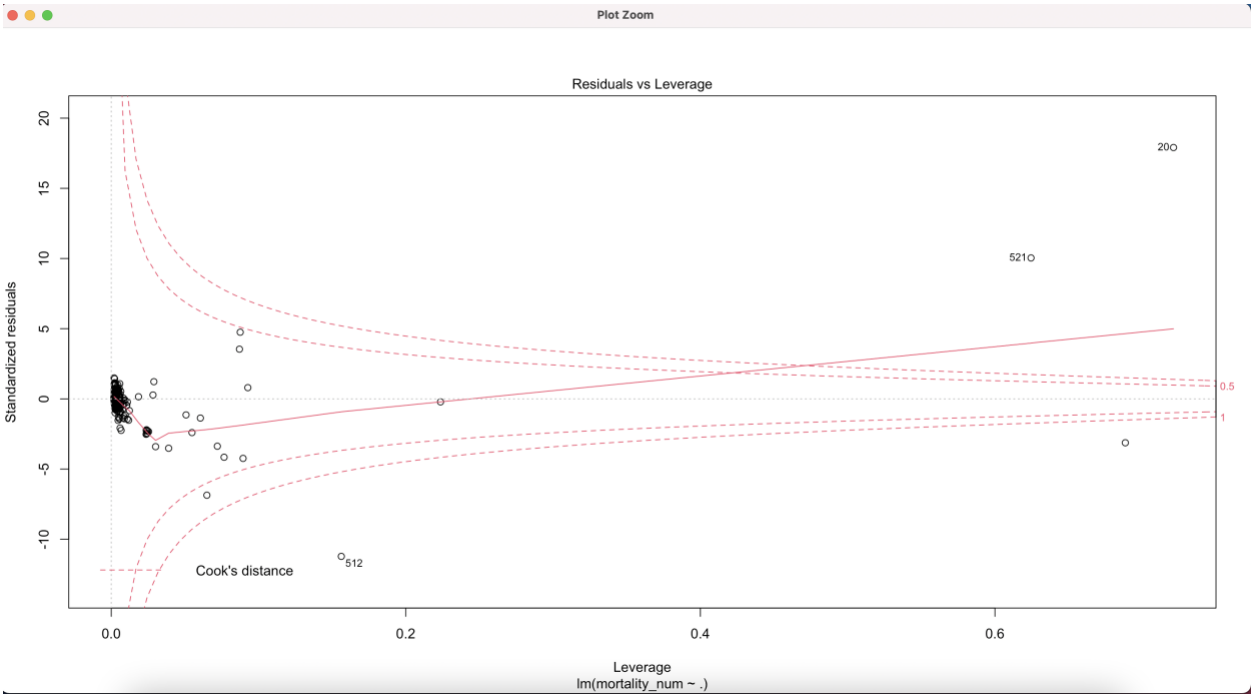
(cardio_num is not statistically significant though)



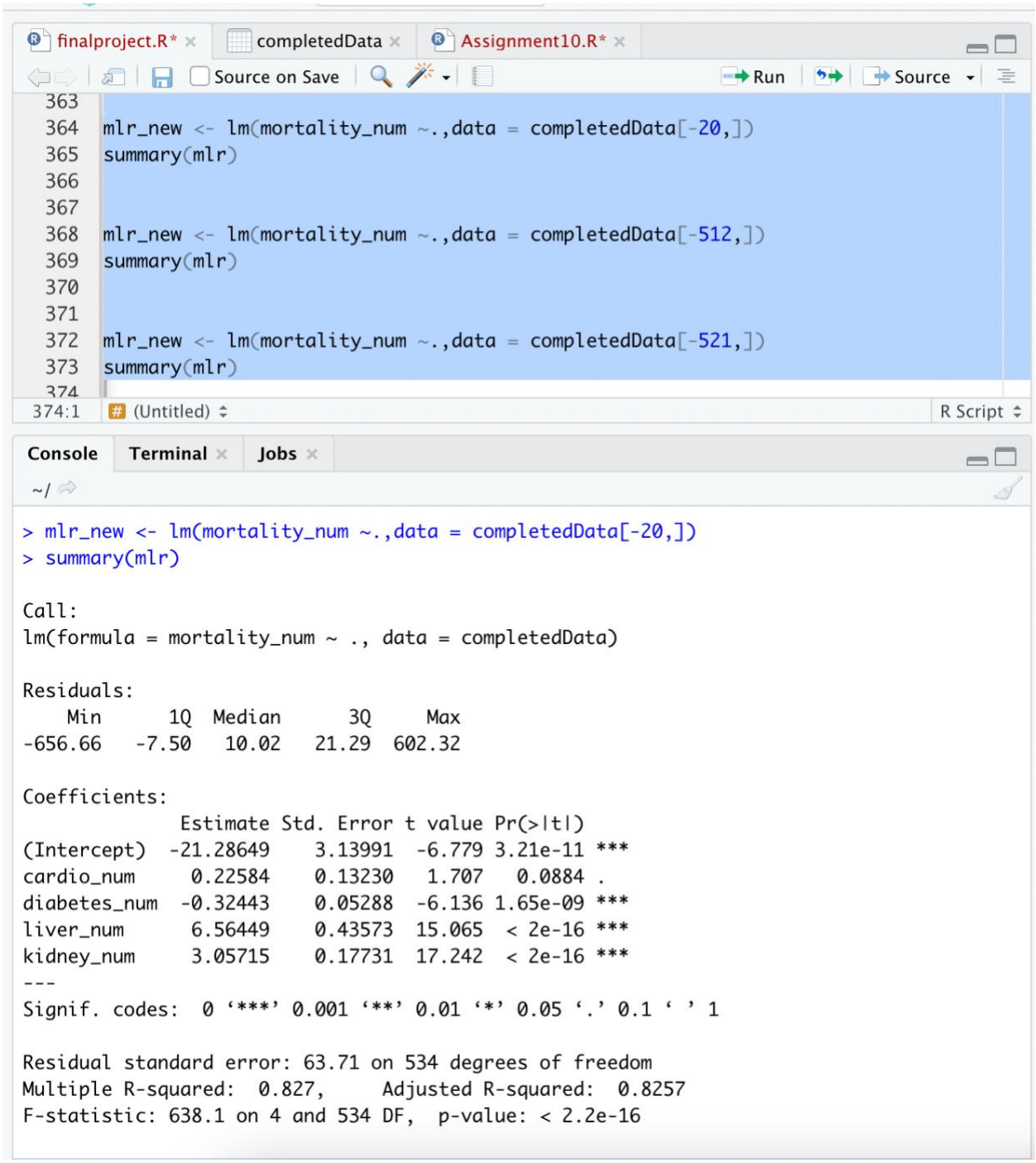
- **OUTLIERS**

We can see observations 512, 521 and 20 as outliers.

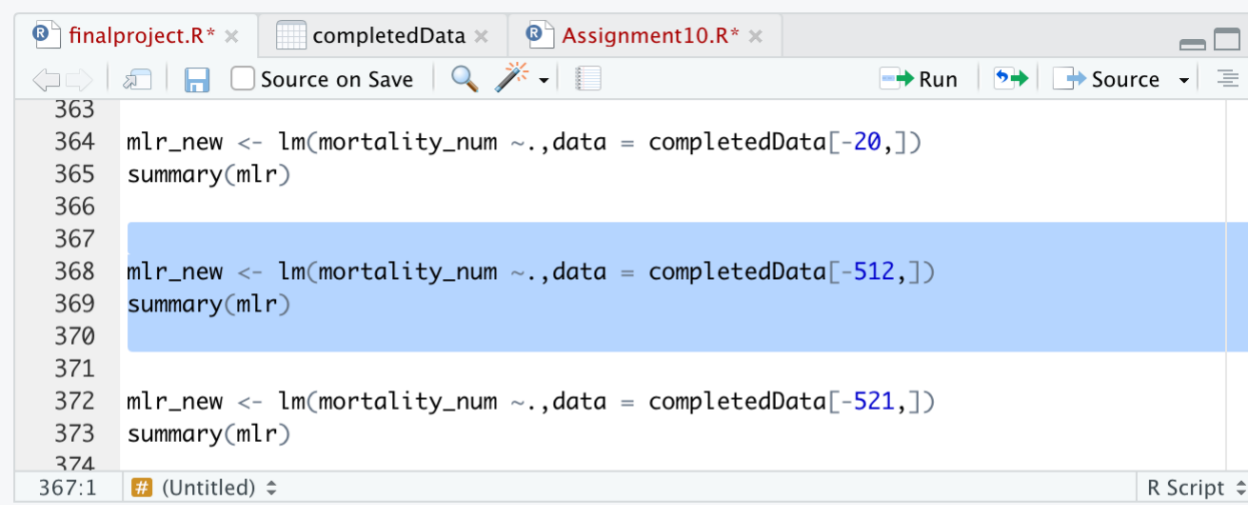


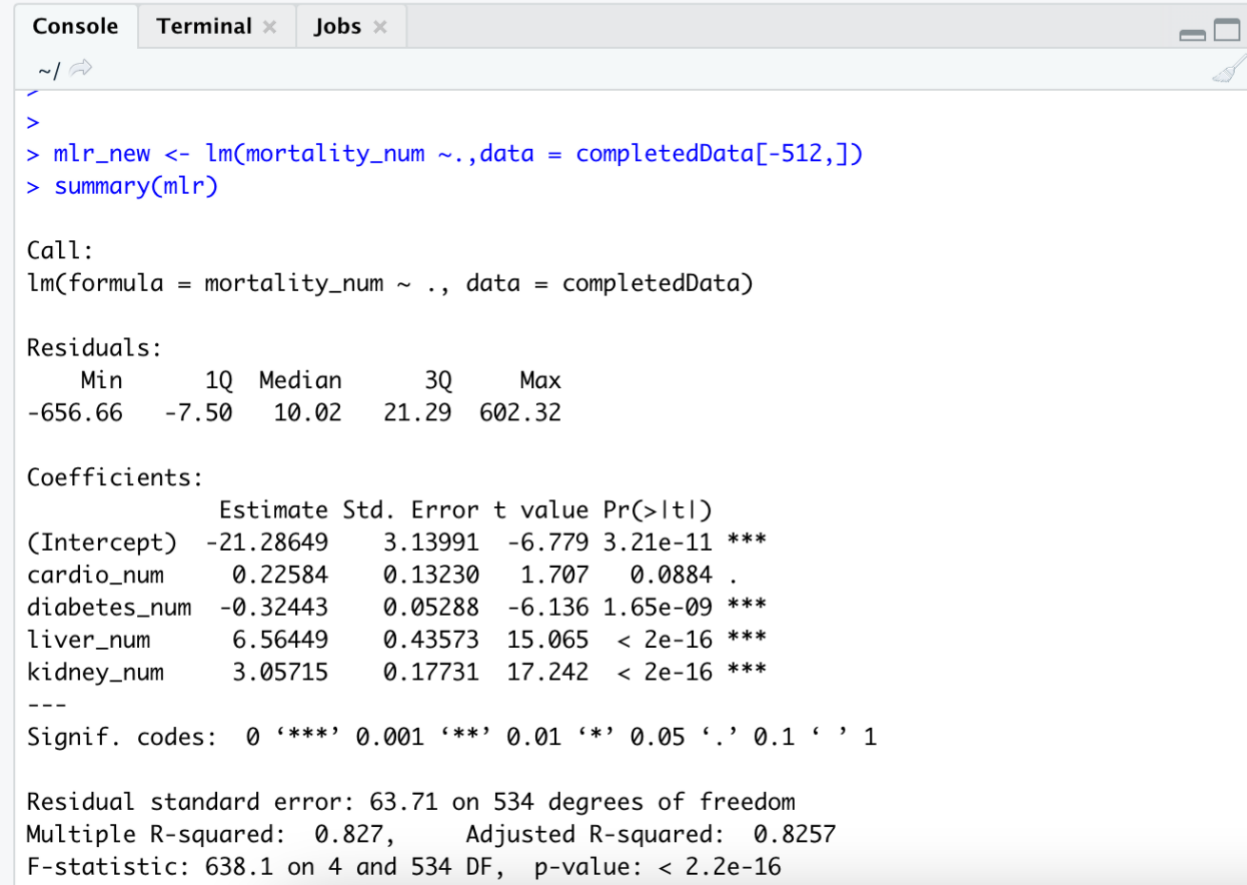


We remove the outliers one at a time and re-calculate R-squared after fitting the model.


finalproject.R*
completedData
Assignment10.R*
Source on Save
Run
Source
363
364 mlr_new <- lm(mortality_num ~.,data = completedData[-20,])
365 summary(mlr)
366
367
368 mlr_new <- lm(mortality_num ~.,data = completedData[-512,])
369 summary(mlr)
370
371
372 mlr_new <- lm(mortality_num ~.,data = completedData[-521,])
373 summary(mlr)
374
374:1
(Untitled)
R Script
Console
Terminal
Jobs
~/
> mlr_new <- lm(mortality_num ~.,data = completedData[-20,])
> summary(mlr)
Call:
lm(formula = mortality_num ~ ., data = completedData)
Residuals:
Min 1Q Median 3Q Max
-656.66 -7.50 10.02 21.29 602.32
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.28649 3.13991 -6.779 3.21e-11 ***
cardio_num 0.22584 0.13230 1.707 0.0884 .
diabetes_num -0.32443 0.05288 -6.136 1.65e-09 ***
liver_num 6.56449 0.43573 15.065 < 2e-16 ***
kidney_num 3.05715 0.17731 17.242 < 2e-16 ***

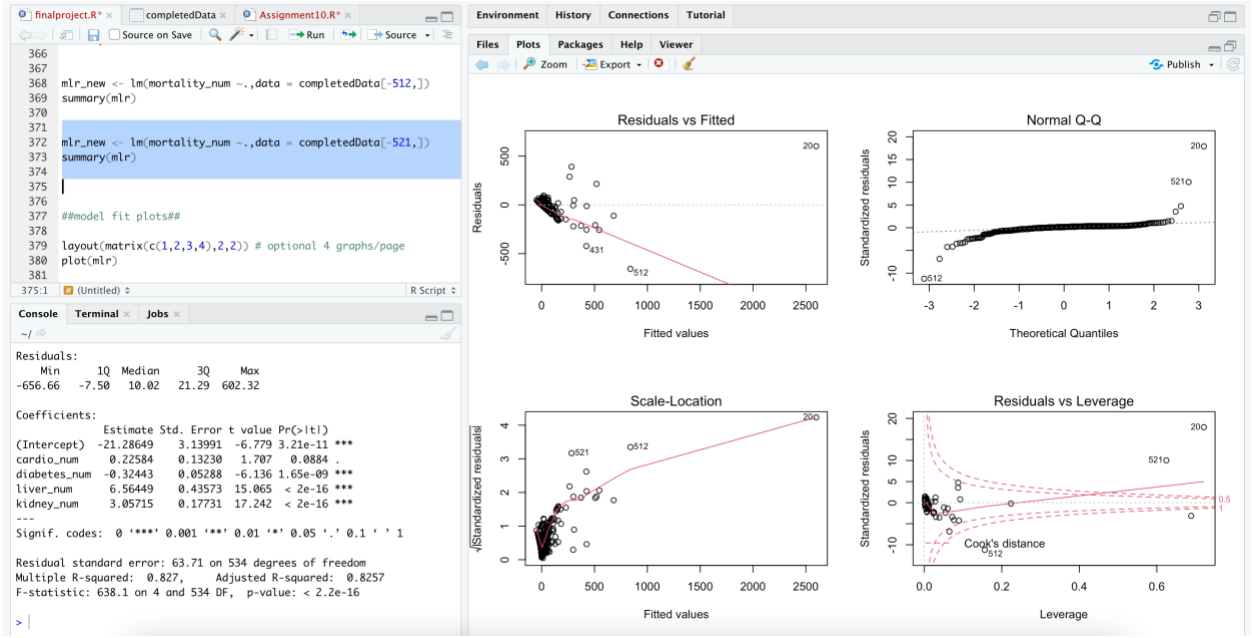
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 63.71 on 534 degrees of freedom
Multiple R-squared: 0.827, Adjusted R-squared: 0.8257
F-statistic: 638.1 on 4 and 534 DF, p-value: < 2.2e-16

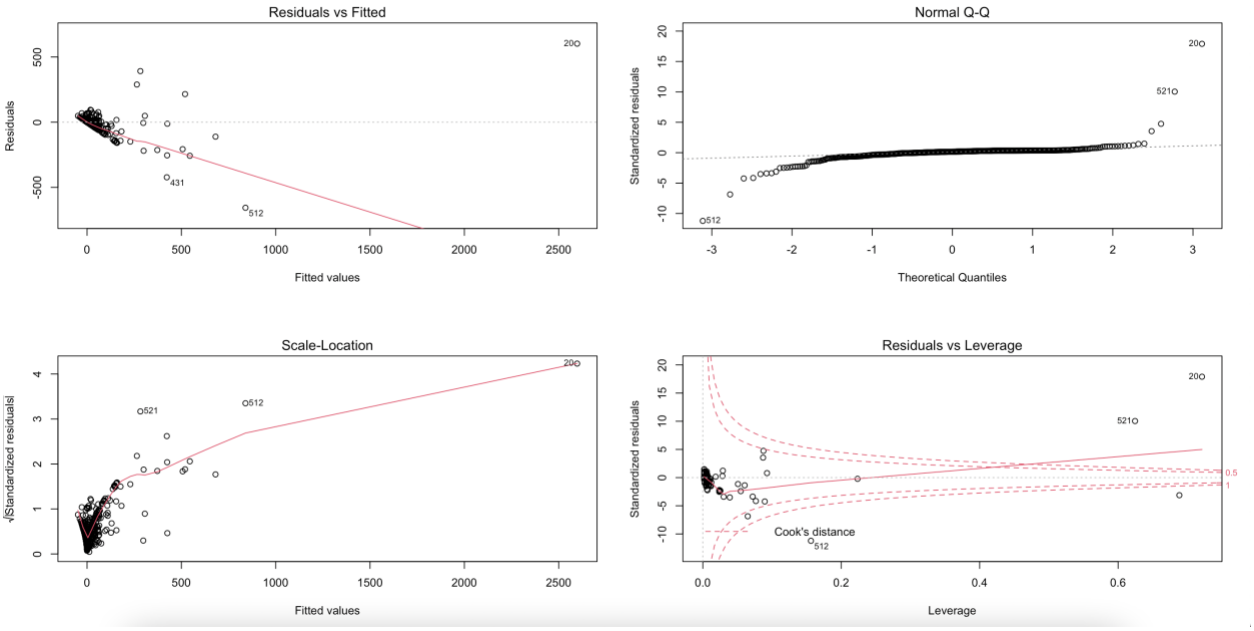




In all three cases, we see the R-squared values are the same as when we fit the model having included these observations. So, these values are not really acting as outliers in terms of affecting the model fit.

- **PLOTS SUMMARIZING MODEL FIT**





PART 5 – PREDICTION

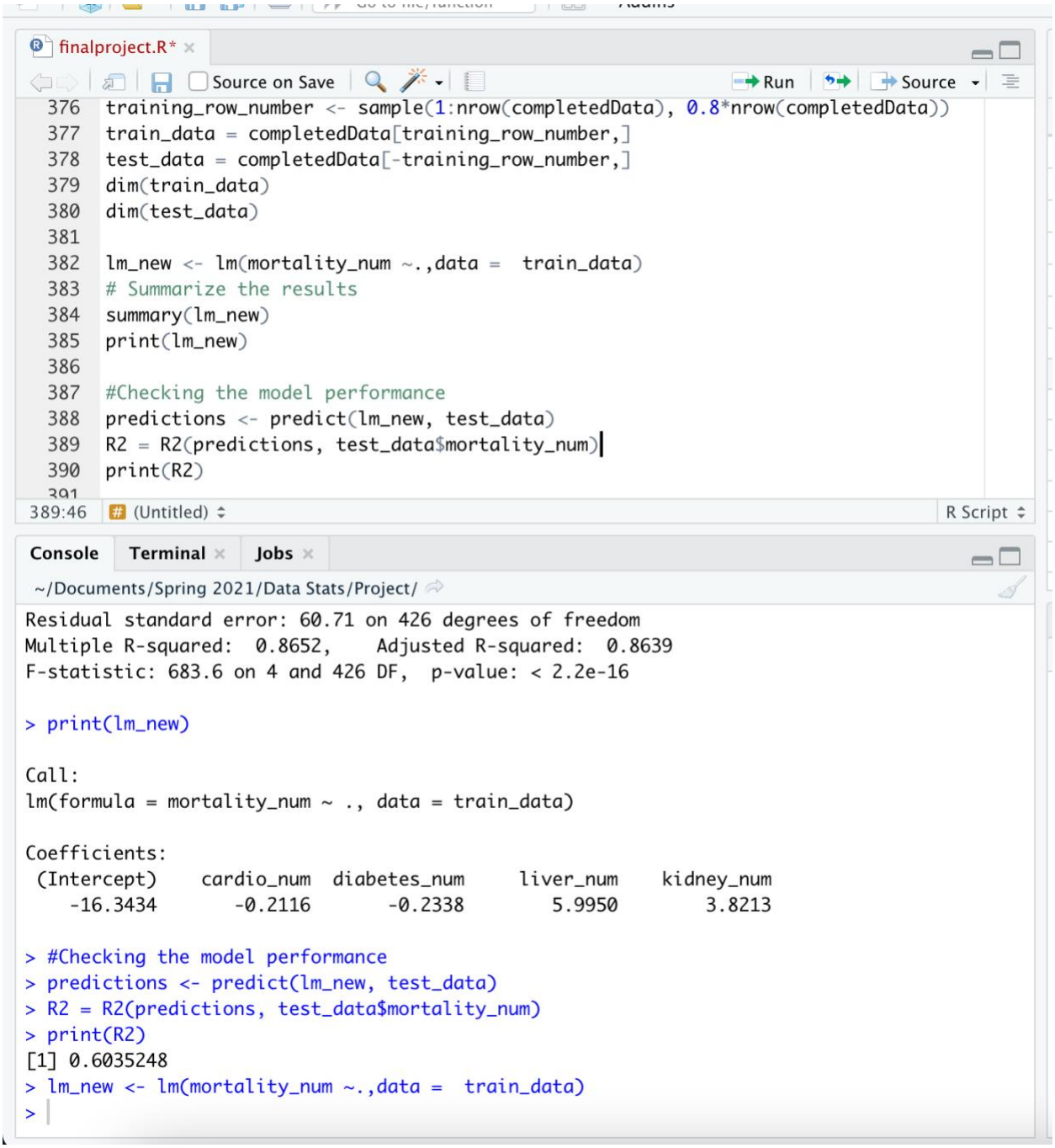
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -21.28649    3.13991  -6.779 3.21e-11 ***
cardio_num     0.22584    0.13230   1.707   0.0884 .
diabetes_num  -0.32443    0.05288  -6.136 1.65e-09 ***
liver_num      6.56449    0.43573  15.065  < 2e-16 ***
kidney_num     3.05715    0.17731  17.242  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.71 on 534 degrees of freedom
Multiple R-squared:  0.827,     Adjusted R-squared:  0.8257
F-statistic: 638.1 on 4 and 534 DF,  p-value: < 2.2e-16
```

From the model fit summary, we can give the following predictions and results -

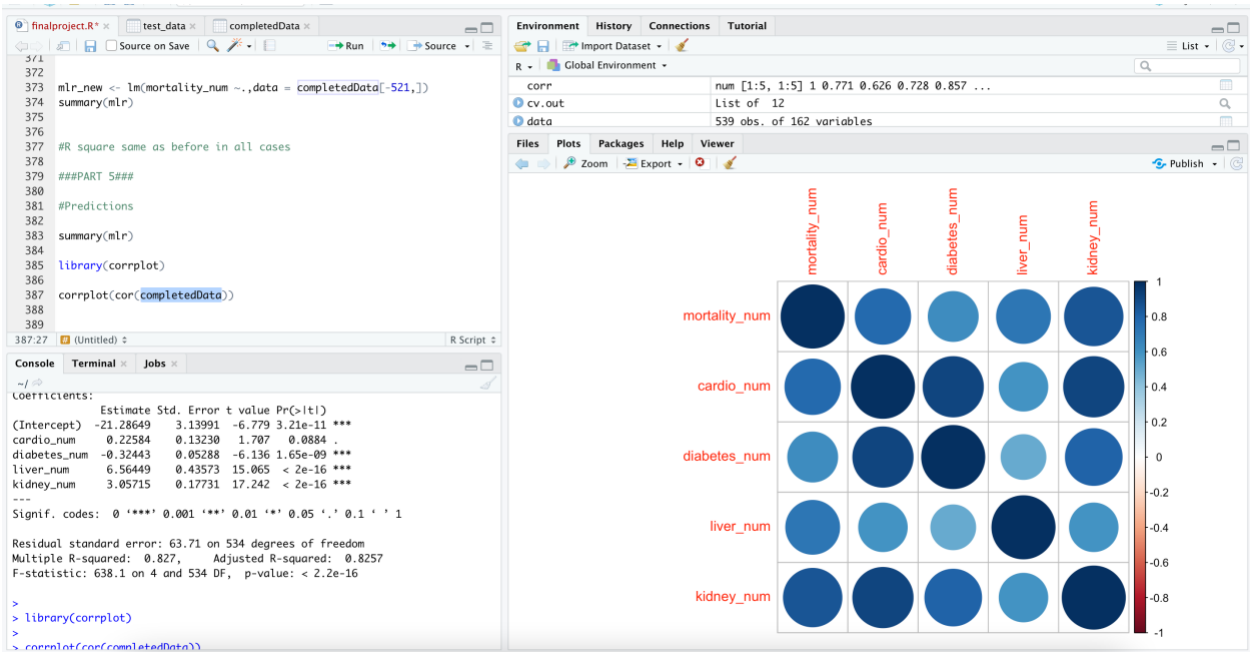
- Change in mortality_num is 6.5 with one unit change in liver_num
- Change in mortality_num is 3 with one unit change in kidney_num
- Liver diseases, kidney diseases and diabetes are statistically significant in correlation to mortality.
- Liver and kidney disease specifically have huge positive correlation to the mortality of covid affected patients.
- Contrary to common logic, the model suggests diabetes history has negative correlation to mortality (also statistically significant).

Predictions



The predicted values for liver_num is 5.995 (vs 6.56 actual) ; for kidney_num is 3.82 (vs 3.05)

The R-square of the prediction is 60%.



The corrplot shows linear relationships between variables, it also shows that kidney_num and liver_num has highest correlation to mortality.

PART 6 – CAVEATS and ERRORS

Measurement error:

- We don't know how accurate the studies are, as it's mentioned on the website that the data has not been subject to peer review.
- Our model wanted to look at all comorbid conditions before trying lasso regression over it. But the majority of the comorbidity columns had such high incompleteness in data that we had to settle with a subset of just 4.
- Papers are not consistent on their reporting of mortality or discharge rates. Some only report mortality; others report only discharges; many patients remain in the hospital at the conclusion of the study
- Studies in this dataset do not always have the same purpose, which affects data reporting. For instance, many papers from Italy seem to report data only on non-survivors. In addition, some studies focus on the disease's contagion profile, with little information on death, discharge, stay length. Data points from these studies may exhibit a high proportion of missing features.
- All our variables in the model had up to 50% missing values which is also measurement error. We fixed this by imputing by mean using mice package.

Sampling error:

- This data is only research data of people who have been critically ill so the mortality rate is possibly higher than what it should ideally be. To be able to make real life predictions, we are leaving out a major amount of population in asymptomatic patients, those who

don’t even get tested but do contain the virus. Hence, this sampling technique has a lot of bias. (https://covidanalytics.io/dataset_documentation)

Modelling error:

- Overall, the residual standard error is 63% which is relatively on the higher side, but the accuracy of the model is still good and it does produce statistically significant outcomes.

Modelling bias - What is the predictive power

For liver_num and kidney_num, the probability of their coefficient values (7.0321 and 3.2439 resp) lying outside 2 standard deviations is very low. (2e-16 each)