ALY 6015 Intermediate Analytics

Descriptive Statistics and Regression Analysis in R

Nalini Macharla

College of Professional Studies, Northeastern University

**INTRODUCTION:**

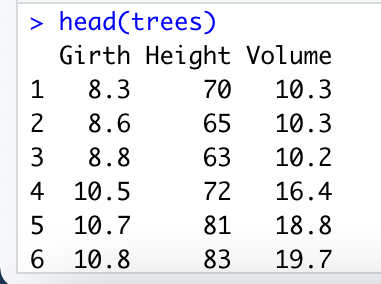
Here in the module 1, we will be using R programming and R studio in order to perform the descriptive and regression data analysis on various datasets. There are so many built in datasets in the R studio and these datasets are really useful to perform test and for practicing data analysis. With the help of data() function, once can see every list of data sets that are available in R. In return with the help of those data sets we can perform and test data analysis functions and other tools as well. We all know that Descriptive Statistics has different things such as standard deviation, mean, variance, median, skewness of graph, minimum and maximum values, kurtosis, quartiles and many more. Descriptive statistical analysis is used to analyze and present the data in a meaningful manner and is used to study and it is used to give a clarity about already existing data in order to summarize the sample. Whereas, regression analysis is used to predict the value of any dependent variable on the basis of the value of variables that are independent.

In this assignment we will be using descriptive statistical analysis and regression analysis in order to perform the given tasks.

**PART A:**

Firstly, we used a built-in function called as ‘trees’ to get the summary regarding the columns present in the dataset.

head(trees)



summary(trees)

Text

Description automatically generated with medium confidence

Here we used summary(trees) function to get the information about min, 1st Quartile, Median, Mean, 3rd Quartile, Max.

cor(trees)

Text

Description automatically generated

Trees is the name which we gave to our dataset. Further we used $Gir, $Vol and $Height to set a parameter to find the output for that particular column.

Below is the plot to find the girth and volume of the tress and we used the following function:

plot(

x = trees$Gir,

y = trees$Vol,

xlab = "Girth of trees",

ylab = "Volume of trees",

main = "Girth and Volume of trees",

col = "red"

)

Chart, scatter chart

Description automatically generated

abline(lm(trees$Vol ~ trees$Gir), col = "black")

Chart, scatter chart

Description automatically generated

Going further we have plotted histograms and density plots for the columns in the dataset ‘trees’.

We can plot a histogram using the function hist().

hist(

trees$Gir,

main = "Histogram for the girth of the trees",

xlab = "Girth",

ylab = "Freq of Girth",

col = "blue"

)

Chart, histogram

Description automatically generated

Above is the histogram that we plotted for the Girth of the trees.

hist(

trees$Height,

main = "Histogram for the Height of the trees",

xlab = "Height",

ylab = "Freq of Height.",

col = "yellow",

)

Chart, histogram

Description automatically generated

Above is the histogram that we plotted for the Height of the trees.

hist(

trees$Vol,

main = "Histogram for the Volume of the trees",

xlab = "Volume",

ylab = "Freq of Vol.",

col = "green"

)

Chart, histogram

Description automatically generated

Above is the histogram that we plotted for the Volume of the trees.

Now, we use the function density() to plot the density plots for the columns of the trees.

plot(

density(trees$Gir),

main = "Density Plot for the Girth of the trees",

xlab = "Girth",

ylab = "Prob of Girth",

col = "red"

)Chart, line chart, histogram

Description automatically generated

Above is the density plot that we plotted for the Girth of the trees.

plot(

density(trees$Height),

main = "Density Plot for the Height of the trees",

xlab = "Height",

ylab = "Prob of Height",

col = "pink"

)

Chart, line chart, histogram

Description automatically generated

Above is the density plot that we plotted for the Height of the trees.

plot(

density(trees$Vol),

main = "Density Plot for the Volume of the trees",

xlab = "Volume",

ylab = "Prob of Volume",

col = "blue"

)

Chart, histogram

Description automatically generated

Above is the density plot that we plotted for the Volume of the trees.

Going further we can plot the boxplots for the columns of the trees dataset and the function that we use is boxplot()

boxplot(

trees$Gir,

main = "Box Plot for Girth of trees",

col = "orange"

)

Chart, box and whisker chart

Description automatically generated

Above is the boxplot for the girth of the trees. Here the boxplot is displaying the min-max values, median and also 1st quartile plus third quartile values. The min value is around 8 and the maximum is around 20. You can notice that the median is between 12 and 14 values. The 1st quartile value here is at 11 and the 3rd quartile value is at 15.

boxplot(

trees$Height,

main = "Box Plot for Height of trees",

col = "green"

)

Chart, box and whisker chart

Description automatically generated

Above is the boxplot for the height of the trees. Median is around 76. Min value is 62 and Max value is 87. 1ST quartile lies somewhere near 72 and 3rd quartile lies around 79.

boxplot(

trees$Vol,

main = "Box Plot for Volume of trees",

col = "red"

)

Chart, box and whisker chart

Description automatically generated

Above is the boxplot for the volume of the trees. Median is around 23. Min value is 10 and Max value is 58. 1ST quartile lies somewhere near 20 and 3rd quartile lies around 35.

Using the function qqnorm() we can plot the normal plot for the columns of the trees dataset.

qqnorm(trees$Gir,

main = "Normal plot for Girth of the trees",

col = "purple"

)

Chart, scatter chart

Description automatically generated

The above is the normal plot for the girth of the trees.

qqnorm(trees$Height,

main = "Normal plot for the Height of the trees",

col = "purple"

)

Chart, scatter chart

Description automatically generated

The above is the normal plot for the height of the trees.

qqnorm(trees$Vol,

main = "Normal plot for the Volume of the trees",

col = "purple"

)

Chart, scatter chart

Description automatically generated

The above is the normal plot for the Volume of the trees.

> skewness(trees$Gir)

[1] 0.5263163

> kurtosis(trees$Gir)

[1] 2.444206

> skewness(trees$Height)

[1] -0.374869

> kurtosis(trees$Height)

[1] 2.430937

> skewness(trees$Vol)

[1] 1.064357

> kurtosis(trees$Vol)

[1] 3.466049

**PART B:**

head(Rubber)

Table

Description automatically generated with medium confidence

summary(Rubber)

Text

Description automatically generated

Using the summary(Rubber) function we found the information about min., 1ST quartile, median, mean, 3rd quartile and max. parameters columns for the dataset ‘rubber’.

head(oddbooks)

Table

Description automatically generated

summary(oddbooks)

Text

Description automatically generated with low confidence

Using the summary(oddbooks) function we found the information about min., 1ST quartile, median, mean, 3rd quartile and max. parameters columns for the dataset ‘oddbooks’.

Regression analysis is used to predict the value of any dependent variable on the basis of the value of variables that are independent. Correlation analysis helps you to find and learn the connection between the variables.

Below, we used regression and correlation functions to determine the coefficients of the dataset’s ‘rubber’ and ‘oddbooks’.

regrRubber <- lm(loss ~ hard + tens, data = Rubber)

regrRubber

Text

Description automatically generated

regrOddbooks <- lm(weight ~ thick + height + breadth, data = oddbooks)

regrOddbooks

Text

Description automatically generated with medium confidence

corrRubber <- cor(Rubber)

corrRubber

Text, letter

Description automatically generated

corrOddbooks <- cor(oddbooks)

corrOddbooks

Graphical user interface, text

Description automatically generated

ggcorrplot is used for the visualization of correlation matrix using the package ggplot2. ggcorrplot() is used for obtaining a graphical representation such as below output. Here we used ggcorrplot() for datasets ‘rubber’ and ‘oddbooks’ as you can see below.

ggcorrplot(

correlationRubber,

colors = c("red", "blue", "yellow"),

method = "square"

)

Chart

Description automatically generated

ggcorrplot(

corrOddbooks,

colors = c("pink", "blue", "orange"),

method = "square"

)

Chart

Description automatically generated

ggpairs() is a form of ggmatrix() which produces a pairwise type of comparison of different sets of data. Below we used ggpairs() for the datasets ‘rubber’ and ‘oddbooks’ to get the output.

ggpairs(

Rubber,

diag = list(continuous = "densityDiag"),

columns = 1:ncol(Rubber),

upper = list(continuous = "cor"),

axisLabels = c("show", "internal", "none"),

lower = list(continuous = "points")

)

Chart, scatter chart

Description automatically generated

ggpairs(

oddbooks,

diag = list(continuous = "densityDiag"),

columns = 1:ncol(Rubber),

upper = list(continuous = "cor"),

axisLabels = c("show", "internal", "none"),

lower = list(continuous = "points")

)

Chart, line chart

Description automatically generated

**CONCLUSION:**

R is a tool which we use to perform different types analysis under different domains. In this assignment we have performed various analytical operations on a wide range of data using R. Here we performed descriptive statistical analysis and regression analysis on the data which is a the very first step before moving onto performing the inferential data analysis in order to retrieve insights from the raw data. We have created various histograms and graphical representations which can help others to understand and to interpret data through these visualizations.

**REFERENCES:**

[1] Surbhi, S. (2019, December 12). Difference between Descriptive and Inferential Statistics. Retrieved January 24, 2021 from <https://keydifferences.com/difference-between-descriptive-and-inferential-statistics.html#:~:text=Descriptive%20Statistics%20is%20a%20discipline%20which%20is%20concerned,analyzes%20and%20presents%20data%20in%20a%20meaningful%20way>.

[2] (2019, May 19). Package’ggcorrplot’. Retrieved January 24, 2021 from <https://cran.r-project.org/web/packages/ggcorrplot/ggcorrplot.pdf>

[3] Scholrke, B. (2015, October 19). ggpairs(): Pairwiseplot matrix. Retrieved January 24, 2021 from <http://ggobi.github.io/ggally/articles/ggpairs.html>

[4] Maindonald, J, H. (2008, January 19). Using R for Data Analysis and Graphics. Retrieved January 24, 2021 from <https://cran.r-project.org/doc/contrib/usingR.pdf>

[5] R Tutorial. Retrieved January 24, 2021 from <https://www.statmethods.net/r-tutorial/index.html>