Module2 Chi Square and ANOVA Assignment

Naveenkumar Govindasamy
NUID-001542203
ALY6015-21591 SEC 10 Intermediate Analytics
Northeastern University
Instructor: Vladimir Shapiro
Date:23 January, 2022

Introduction

In this assignment we are going to use chi-square and Anova test to solve different types of problems. We are going to test distribution of a fit, test two variables for independence and test proportions for homogeneity using chi-square. And using one-way Anova technique to determine significance among three or more means. Using two-way Anova technique to see the significant difference in the main effects.In first task we are going to compare blood samples and bring out the results, in the second task we are going to see the performance results of the air crafts. In third task. And likewise, ethnicity and Movie admissions and women in military, these tasks share the same similarities and process. From task 5 we'll do Anova test to compare two variables, if the null hypothesis is rejected in Anova test we'll use Tukey test to see the significant difference in the pairs of means. Task 6 and task 7 share same similarities, when it comes to task 9 we have given two datasets for each subdivision so one is baseball and the other is crops. In baseball task we are asked to do exploratory data analysis, so i'm planning to EDA and some graphs and report for. And then we need to perform chi-square and goodness-of-test to determine if there is a difference in the number of wins by decade. In Crops task we need to do ANOVA test using yield as dependent variable and fertilizer as the independent variable

Task-1

- **A medical researcher wishes to see if hospital patients in a large hospital have the same blood type distribution as those in the general population. The distribution for the general population is as follows: type A, 20%; type B, 28%; type O, 36%; and type AB = 16%. He selects a random sample of 50 patients and finds the following: 12 have type A blood, 8 have type B, 24 have type O, and 6 have type AB blood. At alpha = 0.10, can it be concluded that the distribution is the same as that of the general population?**

library(dplyr)

```
#The observed probabilities are,
#Type A = 20%
#Type B = 28%
#Type C = 36%


#The observed blood types values of patients are,
#Type A = 12
#Type B = 8
#Type C = 24

#The significance level is 0.10

alpha1 <- 0.10

#Creating vector for the observed probability blood types

Probability_of_blood_types <- c(0.20, 0.28, 0.36)

#Creating vector using the observed blood types

Observed_values_of_blood_types <- c(12, 8, 24)

#Chi-Square Test

Result <- chisq.test(x = Observed_values_of_blood_types, p = Probability_of_blood_types,rescale.p = TRUE)

#Comparing the P-Value to alpha and make the decision.

ifelse(Result$p.value > alpha1, "Fail to reject the null hypothesis", "Reject the null hypothesis")
```

```
## [1] "Reject the null hypothesis"
```

```
Result$statistic
```

```
## X-squared
##  4.654545
```

```
Result$p.value
```

```
## [1] 0.09756146
```

```
Result$parameter
```

```
## df
##  2
```

```
Result
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Observed_values_of_blood_types
## X-squared = 4.6545, df = 2, p-value = 0.09756
```

Observation

In this task first we are collecting the observed blood types and the probability types and the significance level is given as 0.10. Then we are creating vectors for the observed and probability values. After creating vector we need to do chi-square test. The results from chi-square tests for the given probabilities are x-squared = 4.6545 and p-value = 0.09756. After that we are comparing the p.value to the given alpha if we are taking the null hypothesis. After comparing, the result p.value is lesser than the alpha and so we reject the null hypothesis here.

Task-2

- **According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows: Action % of Time On time 70.8 National Aviation System delay 8.2**

Aircraft arriving late 9.0 Other (because of weather and other conditions) 12.0 Records of 200 randomly selected flights for a major airline company showed that 125 planes were on time; 40 were delayed because of weather, 10 because of a National Aviation System delay, and the rest because of arriving late. At α = 0.05, do these results differ from the government's statistics?

```
#The observed Probabilities of on_time performance of the airlines,
#On-time = 70.8%
#National Aviation System Delay = 8.2%
#Aircraft arriving late = 9.0%

#The observed values of on-time performance of the airlines,
#On-time = 125
#National Aviation System Delay = 40
#Aircraft delay = 10

#The significance level is 0.05

alpha2 <- 0.05

#Creating vector for probability of 0n-time performance of the airlines,

Probability_of_performance_of_airlines <- c(0.70, 0.08, 0.09, 0.12)

#Creating vector for the observed On-time performance of the airline,

Probability_of_Observed_performance_of_airline <- c(125, 40, 10, 25)

#Chi-Square Test

Result2 <- chisq.test(x = Probability_of_Observed_performance_of_airline, p = Probability_of_per
formance_of_airlines,rescale.p = TRUE)

#Comparing the P-Value to alpha and make the decision.

ifelse(Result$p.value > alpha2, "Fail to reject the null hypothesis", "Reject the null hypothesi
s")
```

```
## [1] "Fail to reject the null hypothesis"
```

```
Result2$statistic
```

```
## X-squared
##  40.77232
```

```
Result2$p.value
```

```
## [1] 7.308203e-09
```

```
Result2$parameter
```

```
## df
##  3
```

```
Result2
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Probability_of_Observed_performance_of_airline
## X-squared = 40.772, df = 3, p-value = 7.308e-09
```

Observation

In this task the significance level is given as aplha = 0.05, and the vector for the probability is given for the performance by the airlines is(70.8, 8.2, 9.0, 12.0) and the observed values vector for the performance of the airlines is(125, 40, 10, 25).After creating vector we need to do chi-square test. The results from chi-square tests for the given probabilities are x-squared = 40.772 and p-value = 7.308 10*^8. After that, we are comparing the p.value to the given alpha. After comparing the result.The p.value is lesser than the alpha and so we reject the null hypothesis here.

Task-3

- **Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity? Caucasian Hispanic African American Other 2013 724 335 174 107 2014 370 292 152 140.**

library(ggplot2)

```
# Set significance level
alpha <- 0.05

#Creating vector for the rows
R1 <- c(724, 335, 174, 107)
R2 <- c(370, 292, 152, 140)

rows = 2

#Number of rows on the given table
Matrx = matrix(c(R1, R2), nrow = rows, byrow = TRUE)

result3 <- chisq.test(Matrx)
result3
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  Matrx
## X-squared = 60.144, df = 3, p-value = 5.478e-13
```

```
#test statistic and p-value
result3$statistic
```

```
## X-squared 
##  60.14352
```

```
result3$p.value
```

```
## [1] 5.477507e-13
```

```
result3$parameter
```

```
## df 
##  3
```

```
ifelse(result3$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesi
s")
```

```
## [1] "Reject the null hypothesis"
```

Observation

Ethnicity and Movie Admission: Significance level is given 0.05, Vector for the year is created with (724, 335, 174, 107) and vector for the year 2014 is given as (370, 292, 152, 140). After creating vector we need to do chi-square test. The results from chi-square tests for the given probabilities are x-squared = 60.14352 and p-value = 5.478*10^13. After that, we are comparing the p.value to the given alpha. After comparing the result.The p.value is lesser than the alpha and so we reject the null hypothesis here.

Task-4

- **This table lists the numbers of officers and enlisted personnel for women in the military. At α = 0.05, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces? Officers Enlisted Army 10,791 62,491 Navy 7,816 42,750 Marine Corps 932 9,525 Air Force 11,819 54,344**

```
# Set significance level
alpha <- 0.05

#Creating vector for the rows
r1 <- c(10791, 62491)
r2 <- c(7816, 42750)
r3 <- c(932, 9525)
r4 <- c(11819, 54344)

rows = 4

#Number of rows on the given table
Matrx = matrix(c(r1, r2, r3, r4), nrow = rows, byrow = TRUE)

result4 <- chisq.test(Matrx)
result4
```

```
##
##  Pearson's Chi-squared test
##
## data:  Matrx
## X-squared = 654.27, df = 3, p-value < 2.2e-16
```

```
#test statistic and p-value
result4$statistic
```

```
## X-squared
##  654.2719
```

```
result4$p.value
```

```
## [1] 1.726418e-141
```

```
result4$parameter
```

```
## df
##  3
```

```
ifelse(result4$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesi
s")
```

```
## [1] "Reject the null hypothesis"
```

Observation

Women in military: Significance level is given 0.05, Vector for the officers is created with (10791, 7816, 932, 11819) and enlisted is given as (62491, 42750, 9525, 54344). After creating vector we need to do chi-square test. The results from chi-square tests for the given probabilities are x-squared = 654.27 and p-value = 2.2*10^16. After that, we are comparing the p.value to the given alpha. After comparing the result.The p.value is lesser than the alpha and so we reject the null hypothesis here.

Task-5

- **The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts? Condiments Cereals Desserts 270 260 100 130 220 180 230 290 250 180 290 250 80 200 300 70 320 360 200 140 300**

  160

```
#Setting the significance level
alpha5 <-0.05

#Creating a data frame for the condiments
Condiments <- data.frame('sodium' = c(270, 130, 230, 180, 80, 70, 200), 'food' = rep('Condiments', 7), stringsAsFactors = FALSE)

#Creating a data frame for cereals
cereals <- data.frame('sodium' = c(260, 220, 290, 290, 200, 320, 140), 'food' = rep('cereals', 7), stringsAsFactors = FALSE)

#Creating a data frame for deserts
desserts <- data.frame('sodium' = c(100, 180, 250, 250, 300, 360, 300,160), 'food' = rep('desserts', 8), stringsAsFactors = FALSE)

#combining the dataframes into one
sodium <- rbind(Condiments, cereals, desserts)
sodium$food <- as.factor(sodium$food)

#ANOVA test
anova <- aov(sodium ~ food, data = sodium)

#summary of the model
a.summary <- summary(anova)

#Degrees of freedom
#k-1:between group variance - numerator
df.numerator <- a.summary[[1]][1, "DF"]
df.numerator
```

```
## NULL
```

```
# N - K: within group variance - denominator
df.denominator <- a.summary[[1]][2, "DF"]
df.denominator
```

```
## NULL
```

```
#Extracting the F-test value from the summary

F.value <- a.summary[[1]][[1, "F value"]]
F.value
```

```
## [1] 2.398538
```

```
#Extract the p-value from the summary
p.value <- a.summary[[1]][[1, "pr(>F)"]]
p.value
```

```
## NULL
```

```
#Determining if we should reject the null hypothesis
ifelse(p.value > alpha5, "fail to reject the null", "reject the null")
```

```
## logical(0)
```

```
#To see the difference
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sodium ~ food, data = sodium)
##
## $food
##                            diff        lwr       upr     p adj
## Condiments-cereals  -80.000000 -182.89588  22.89588 0.1456674
## desserts-cereals     -8.214286 -107.84279  91.41422 0.9761344
## desserts-Condiments  71.785714  -27.84279 171.41422 0.1866850
```

Observation

Amount of sodium in a food: Significance level is given 0.05, Vector for the Condiments is created with (270, 10, 230, 180, 80, 70, 200) and cereals is created with (260, 220, 290, 290, 200, 320, 140) and vector for desserts is created with (100, 180, 250, 250, 300, 360, 300). After creating vector we need to do chi-square test. After that, we are comparing the p.value to the given alpha. After comparing the result.The p.value is lesser than the alpha and so we reject the null hypothesis here.When the null hypothesis is rejected we have to consider tukeys test into consideration. In tukeys test lower value upper value and p adjusted

value will be considered. We can see the p adjusted value for the cereals. The adjusted p-value is used for multiple comparison. This is used to compare the significance difference between the mean levels and we don’t have any significance level level found here.

Task-6

- **Perform a complete one-way ANOVA. If the null hypothesis is rejected, use either the Scheffé or Tukey test to see if there is a significant difference in the pairs of means. Assume all assumptions are met. The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means? Cereal Chocolate Candy Coffee 578 311 261 320 106 185 264 109 302 249 125 689 237 173**

```
#Setting the significance level
alpha6 <-0.01

#Creating a data frame for the condiments
cereal <- data.frame('leadcomp' = c(578, 320, 264, 249, 237), 'food' = rep('cereal', 5), strings
AsFactors = FALSE)

#Creating a data frame for cereals
candy <- data.frame('leadcomp' = c(311, 106, 109, 125, 173), 'food' = rep('candy', 5), stringsAs
Factors = FALSE)

#Creating a data frame for deserts
coffee <- data.frame('leadcomp' = c(261, 185, 302, 689), 'food' = rep('coffee', 4), stringsAsFac
tors = FALSE)

#combining the data-frames into one
leadcomp <- rbind(cereal, candy, coffee)
leadcomp$food <- as.factor(leadcomp$food)

#ANOVA test
anova1 <- aov(leadcomp ~ food, data = leadcomp)

#summary of the model
a.summary1 <- summary(anova1)

#Degrees of freedom
#k-1:between group variance - numerator
df.numerator1 <- a.summary1[[1]][1, "DF"]
df.numerator1
```

```
## NULL
```

```
# N - K: within group variance - denominator
df.denominator1 <- a.summary1[[1]][2, "DF"]
df.denominator1
```

```
## NULL
```

```
#Extracting the F-test value from the summary

F.value <- a.summary1[[1]][[1, "F value"]]
F.value
```

```
## [1] 2.171782
```

```
#Extract the p-value from the summary
p.value <- a.summary1[[1]][[1, "pr(>F)"]]
p.value
```

```
## NULL
```

```
#Determining if we should reject the null hypothesis
ifelse(p.value > alpha6, "fail to reject the null", "reject the null")
```

```
## logical(0)
```

```
#To see the difference
TukeyHSD(anova1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = leadcomp ~ food, data = leadcomp)
##
## $food
##                  diff        lwr      upr     p adj
## cereal-candy  164.80  -99.22409 428.8241 0.2535458
## coffee-candy  194.45  -85.58983 474.4898 0.1916553
## coffee-cereal  29.65 -250.38983 309.6898 0.9561014
```

Observation

Significance level is given 0.05, Vector for the Cereals is created with (578, 320, 264, 249, 237) and cereals is created with (311, 106, 109, 125, 173) and vector for desserts is created with (261, 185, 302, 689). After creating vector we need to do chi-square test. After that, we are comparing the p.value to the given alpha. After comparing the result.The p.value is lesser than the alpha and so we reject the null hypothesis here.When the null hypothesis is rejected we have to consider tukeys test into consideration. In tukeys test lower value upper value and p adjusted value will be considered. We can see the p adjusted value for the coffee. The adjusted p-value is used for multiple comparison. This is used to compare the significance difference between the mean levels and we don’t have any significance level level found here, significance level is same as the original p-value

Task-7

- **Perform a complete one-way ANOVA. If the null hypothesis is rejected, use either the Scheffé or Tukey test to see if there is a significant difference in the pairs of means. Assume all assumptions are met.**

The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using α = 0.05, can you conclude that there is a difference in means? Eastern third Middle third Western third 4946 6149 5282 5953 7451 8605 6202 6000 6528 7243 6479 6911 6113

```
#Setting the significance level
alpha7 <-0.05

#Creating a data frame for the Eastern Third
Easternthird <- data.frame('exp' = c(4946, 5953, 6202, 7243, 6113), 'states' = rep('Easternthird', 5), stringsAsFactors = FALSE)

#Creating a data frame for Middle third
Middlethird <- data.frame('exp' = c(6149, 7451, 6000, 6479), 'states' = rep('Middlethird', 4), stringsAsFactors = FALSE)

#Creating a data frame for Western third
Westernthird <- data.frame('exp' = c(5282, 8605, 6528, 6911), 'states' = rep('Westernthird', 4), stringsAsFactors = FALSE)

#combining the data-frames into one
exp <- rbind(Easternthird, Middlethird, Westernthird)
exp$states <- as.factor(exp$states)

#ANOVA test
anova2 <- aov(exp ~ states, data = exp)

#summary of the model
a.summary2 <- summary(anova2)

#Degrees of freedom
#k-1:between group variance - numerator
df.numerator2 <- a.summary2[[1]][1, "DF"]
df.numerator2
```

```
## NULL
```

```
# N - K: within group variance - denominator
df.denominator2 <- a.summary2[[1]][2, "DF"]
df.denominator2
```

```
## NULL
```

```
#Extracting the F-test value from the summary

F.value <- a.summary2[[1]][[1, "F value"]]
F.value
```

```
## [1] 0.6488214
```

```
#Extract the p-value from the summary
p.value <- a.summary2[[1]][[1, "pr(>F)"]]
p.value
```

```
## NULL
```

```
#Determining if we should reject the null hypothesis
ifelse(p.value > alpha7, "fail to reject the null", "reject the null")
```

```
## logical(0)
```

```
#To see the difference
TukeyHSD(anova2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = exp ~ states, data = exp)
##
## $states
##                              diff       lwr      upr     p adj
## Middlethird-Easternthird  428.35 -1372.582 2229.282 0.7954670
## Westernthird-Easternthird 740.10 -1060.832 2541.032 0.5203918
## Westernthird-Middlethird  311.75 -1586.599 2210.099 0.8954324
```

Observation

The significance level is is given as 0.05 and the eastern-third dataframe is created, likewise middle-third and west-third is created. And we are combining the all three data-frames into one data frame using RBIND(). Then we are doing the Anova test with states and exp. Then degrees of freedom is calculated with numerator and denominator values. if we compare the p-value and the f.test value and if we fail to reject the null hypothesis we need to perform Tukeys test. Tukeys multiple comparison of means 95% family-wise confidence level. If we perform anova test we'll get the values like lower and upper and p-adjusted values. We can use the adjusted p values for all three variables.

Task-8

Assume that all variables are normally or approximately normally distributed, that the samples are independent, and that the population variances are equal. a. State the hypotheses. b. Find the critical value for each F test. c. Complete the summary table and find the test value. d. Make the decision. e. Summarize the results. (Draw a graph of the cell means if necessary.)

A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a “Grow-light” in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes. Grow-light 1 Grow-light 2 Plant food A 9.2, 9.4, 8.9 8.5, 9.2, 8.9 Plant food B 7.1, 7.2, 8.5 5.5, 5.8, 7.6 Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use α = 0.05

```
#Creating the row vectors

lit1 <- c(9.2, 9.4, 8.9, 7.1, 7.2, 8.5)
lit2 <- c(8.5, 9.2, 8.9, 5.5, 5.8, 7.6)
foodA <- c(9.2, 9.4, 8.9, 8.5, 9.2, 8.9)
foodB <- c(7.1, 7.2, 8.5, 5.5, 5.8, 7.6)
#Creating a data frame containing the vectors
vecdf <- data.frame(c(lit1, lit2),c(foodA, foodB))
#chi-square test
chitest8 <- chisq.test(vecdf)
chitest8
```

```
##
##  Pearson's Chi-squared test
##
## data:  vecdf
## X-squared = 0.75748, df = 11, p-value = 1
```

```
growth <- 0.05
lit_type <- 0.07
#To find the critical values
litmatrix <- matrix(c(rep("lit 1",6),rep("lit 2",6),lit1,lit2),ncol=2)
litdf <- data.frame(litmatrix)
names(litdf) <- c("lighttype", "Growth")
anova_lit <- aov(as.numeric(Growth) ~ lighttype,data=litdf)
summary(anova_lit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## lighttype    1   1.92   1.920   1.083  0.323
## Residuals   10  17.74   1.774
```

```
fd_mtrx <- matrix(c(rep("Food A",6), rep("Food B", 6),foodA, foodB),ncol=2)
fd_df <- data.frame(fd_mtrx)
names(fd_df) <- c("fd_type", "Growth")
anova_fd <- aov(as.numeric(Growth)~fd_type,data = fd_df)
summary(anova_fd)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## fd_type      1 12.813  12.813   18.72 0.0015 **
## Residuals   10  6.843   0.684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observation

we are creating vectors first with light 1 and light 2 and with food A and Food B that is given in the question, creating the data frame with light and food vectors. Performing chi-square test will give out the p-value which is 1 which is greater than significance level that is given and so we reject the null hypothesis. For the next task 8.2, we are giving the matrix with only light given and performing Anova test, which shows the result with sum, mean and F.value. Simultaneously we are doing for the foods which results in giving the values which will be helpfull in comparing the results.

Task-9

```
#Task-9(1). Importing the data-set
library(readr)
baseball <- read_csv("baseball.csv")
```

```
## Rows: 1232 Columns: 15
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): Team, League
## dbl (13): Year, RS, RA, W, OBP, SLG, BA, Playoffs, RankSeason, RankPlayoffs,...
```

```
## 
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(baseball)

#Task-9(2). Exploratory data analysis in the dataset.

library(dplyr)
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

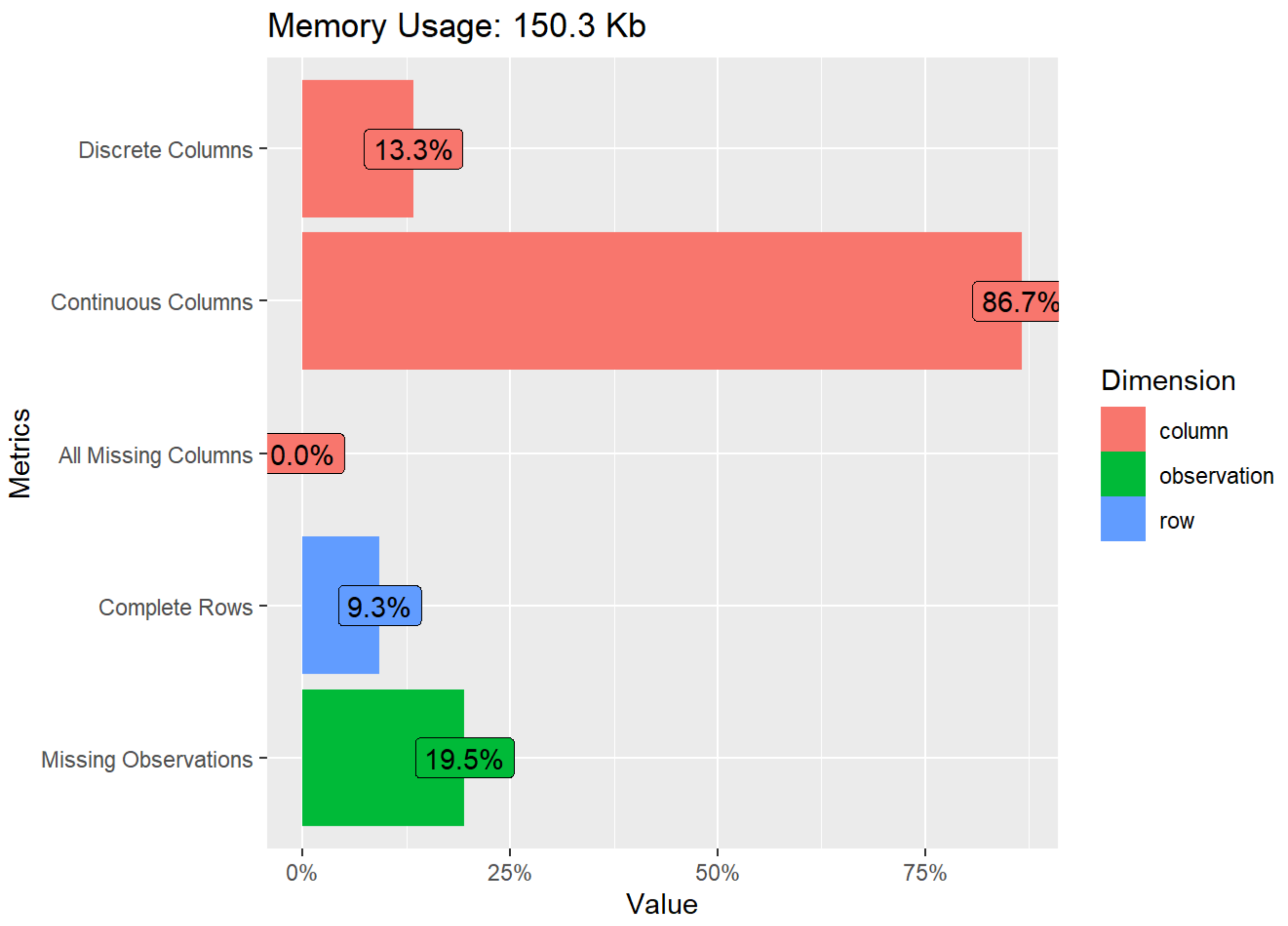
```
library(DataExplorer)
#install.packages("DataExplorer")

#Dataset name = baseballds

baseball %>% glimpse()
```

```
## Rows: 1,232
## Columns: 15
## $ Team         <chr> "ARI", "ATL", "BAL", "BOS", "CHC", "CHW", "CIN", "CLE", "~
## $ League       <chr> "NL", "NL", "AL", "AL", "NL", "AL", "NL", "AL", "NL", "AL~
## $ Year         <dbl> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 201~
## $ RS           <dbl> 734, 700, 712, 734, 613, 748, 669, 667, 758, 726, 583, 67~
## $ RA           <dbl> 688, 600, 705, 806, 759, 676, 588, 845, 890, 670, 794, 74~
## $ W            <dbl> 81, 94, 93, 69, 61, 85, 97, 68, 64, 88, 55, 72, 89, 86, 6~
## $ OBP          <dbl> 0.328, 0.320, 0.311, 0.315, 0.302, 0.318, 0.315, 0.324, 0~
## $ SLG          <dbl> 0.418, 0.389, 0.417, 0.415, 0.378, 0.422, 0.411, 0.381, 0~
## $ BA           <dbl> 0.259, 0.247, 0.247, 0.260, 0.240, 0.255, 0.251, 0.251, 0~
## $ Playoffs     <dbl> 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
## $ RankSeason   <dbl> NA, 4, 5, NA, NA, NA, 2, NA, NA, 6, NA, NA, NA, NA, NA, N~
## $ RankPlayoffs <dbl> NA, 5, 4, NA, NA, NA, 4, NA, NA, 2, NA, NA, NA, NA, NA, N~
## $ G            <dbl> 162, 162, 162, 162, 162, 162, 162, 162, 162, 162, 162, 16~
## $ OOBP         <dbl> 0.317, 0.306, 0.315, 0.331, 0.335, 0.319, 0.305, 0.336, 0~
## $ OSLG         <dbl> 0.415, 0.378, 0.403, 0.428, 0.424, 0.405, 0.390, 0.430, 0~
```

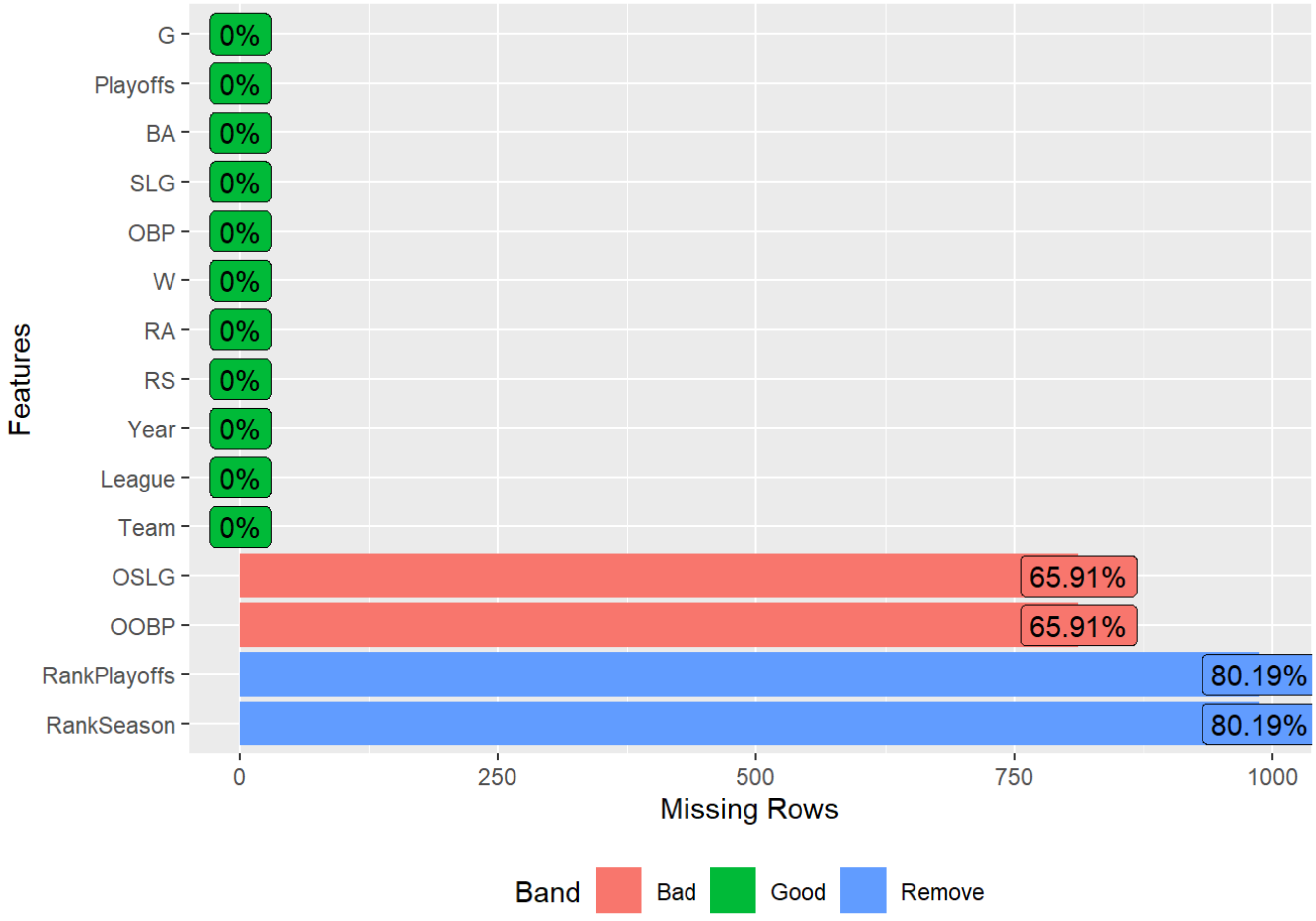
```
baseball %>% introduce()
```

```
## # A tibble: 1 x 9
##    rows columns discrete_columns continuous_columns all_missing_columns
##   <int>   <int>            <int>              <int>               <int>
## 1  1232      15                2                 13                   0
## # ... with 4 more variables: total_missing_values <int>, complete_rows <int>,
## #   total_observations <int>, memory_usage <dbl>
```

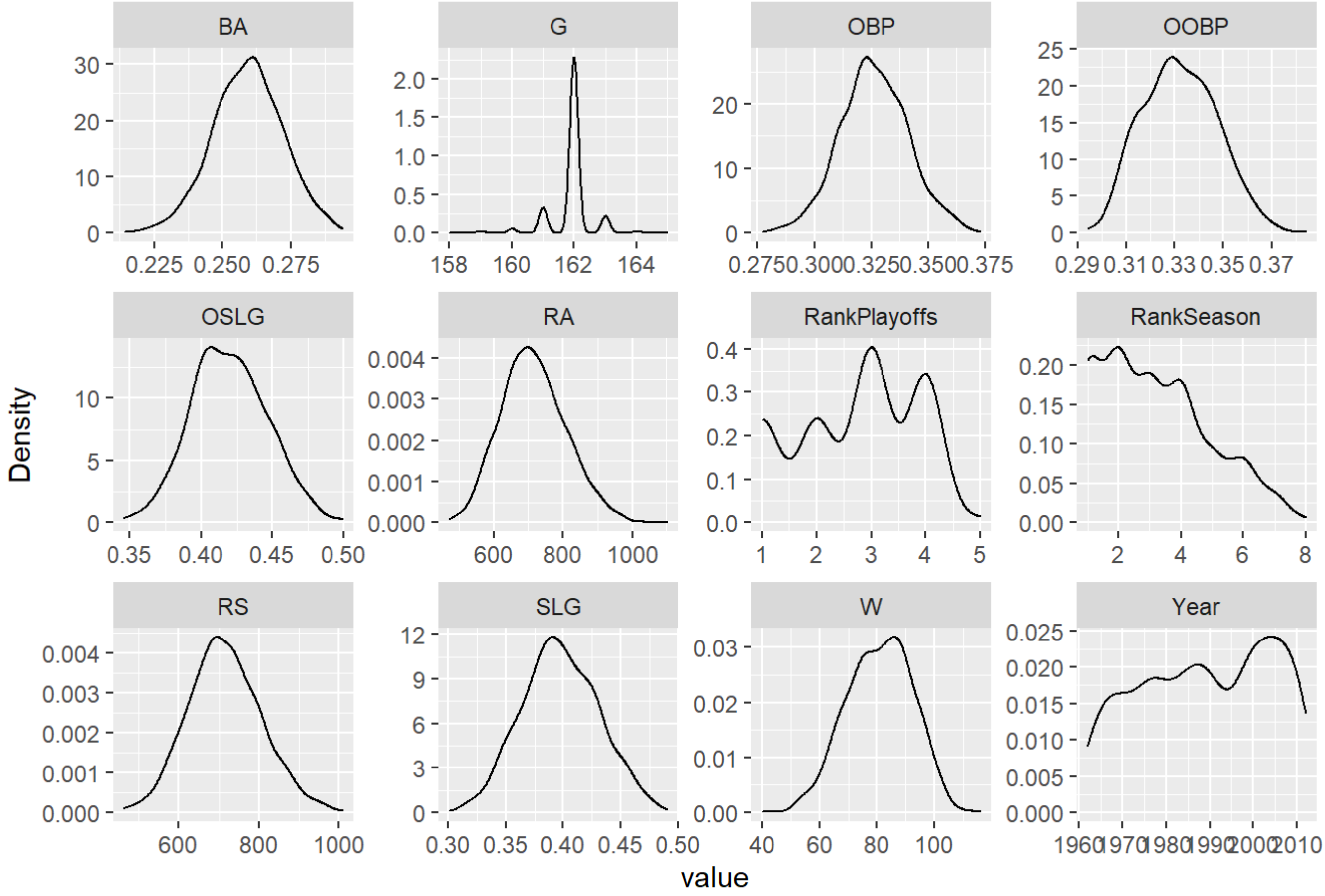
```
baseball %>% plot_intro()
```



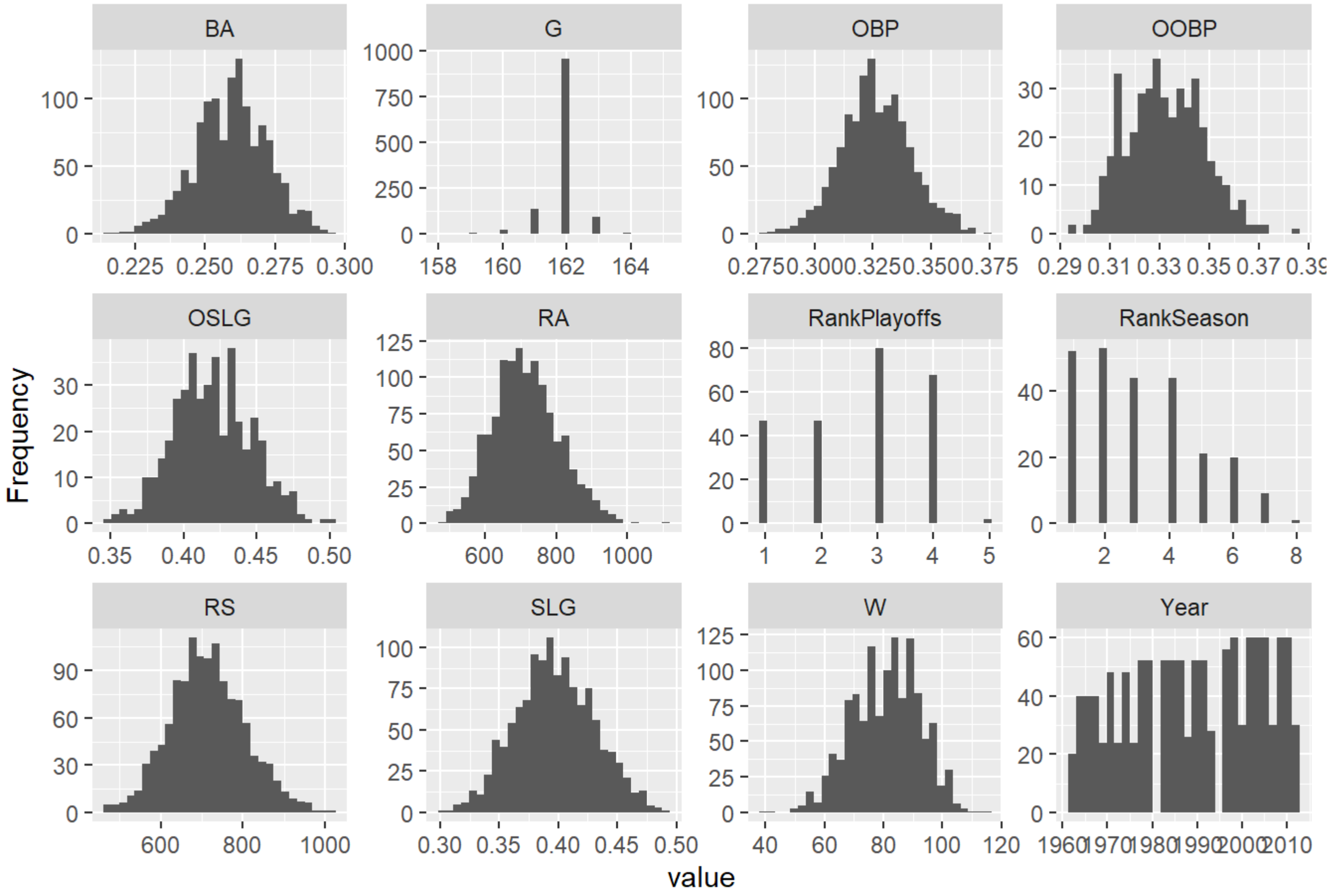
```
baseball %>% plot_missing()
```



```
baseball %>% plot_density()
```



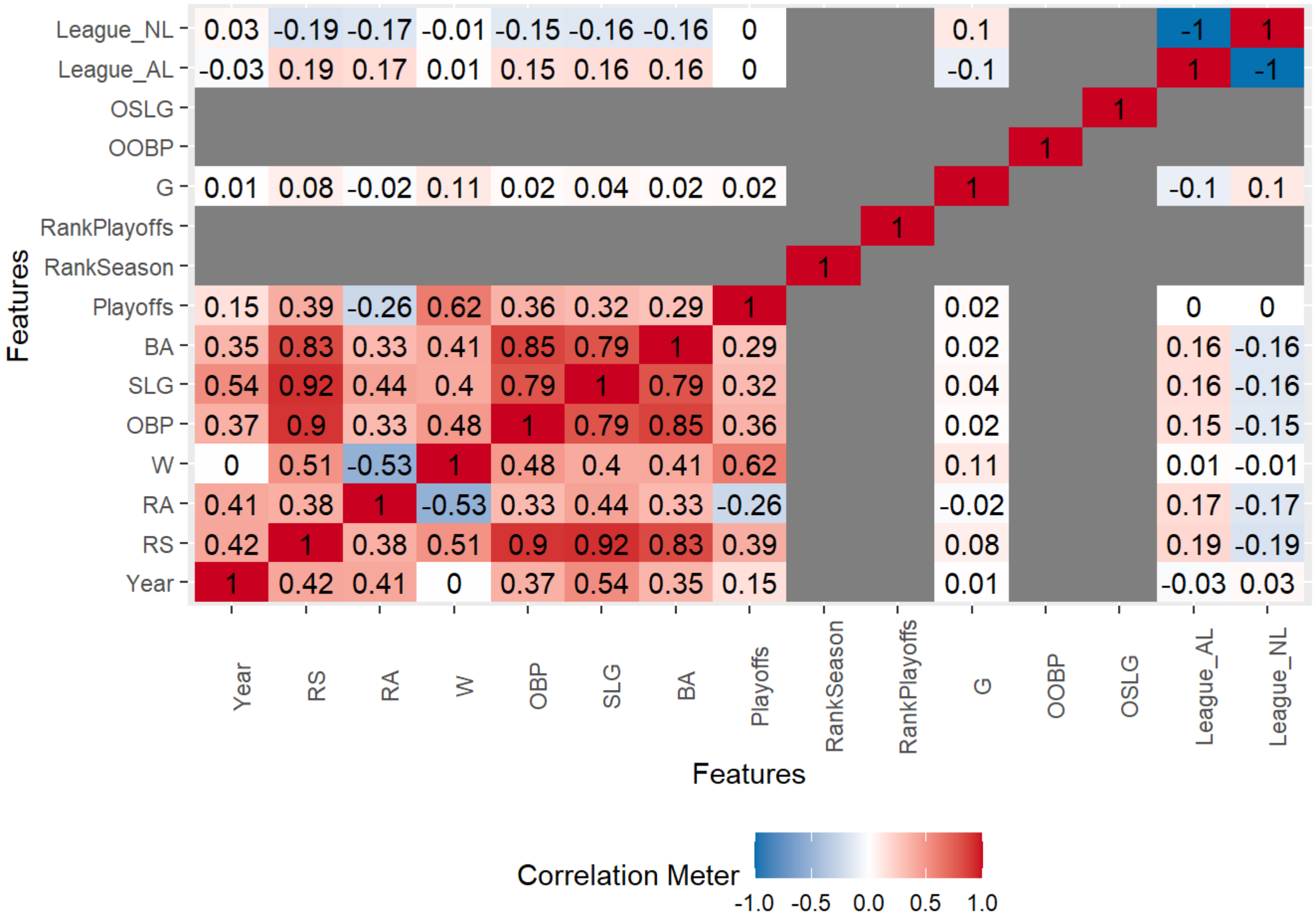
```
baseball %>% plot_histogram()
```



```
baseball %>% plot_correlation(maxcat = 4)
```

```
## 1 features with more than 4 categories ignored!
## Team: 39 categories
```

```
## Warning: Removed 100 rows containing missing values (geom_text).
```



```
#Chi-square goodness-of-Fit test to determine if there is a difference in number of wins by deca
de

#the significance level is given = 0.05
alpha9 = 0.05
#Chisquare test
result9 <- chisq.test(x = baseball$W, p = baseball$W, rescale.p = TRUE)

#result9

ifelse(result9$p.value > alpha9, "Fail to reject the null hypothesis", "Reject the null hypothes
is")
```

```
## [1] "Fail to reject the null hypothesis"
```

Observation

In this task we need to perform exploratory Data Analysis. I used glimpse function to do perform the EDA. We found out missing values using missing function as same like density of whole variables in dataset is found using density plot and likewise for the histograms, to check the frequency of the variable values. And finally correlation plot and this plot is used to compare the variables which has maximum relationship which will be shown in blue. So we can check the player with the year, if it has positive or negative relationships.

```
#Task-9(2)
#Importing crop data-set

library(readr)
crop_data <- read_csv("crop_data.csv")
```

```
## Rows: 96 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): density, block, fertilizer, yield
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

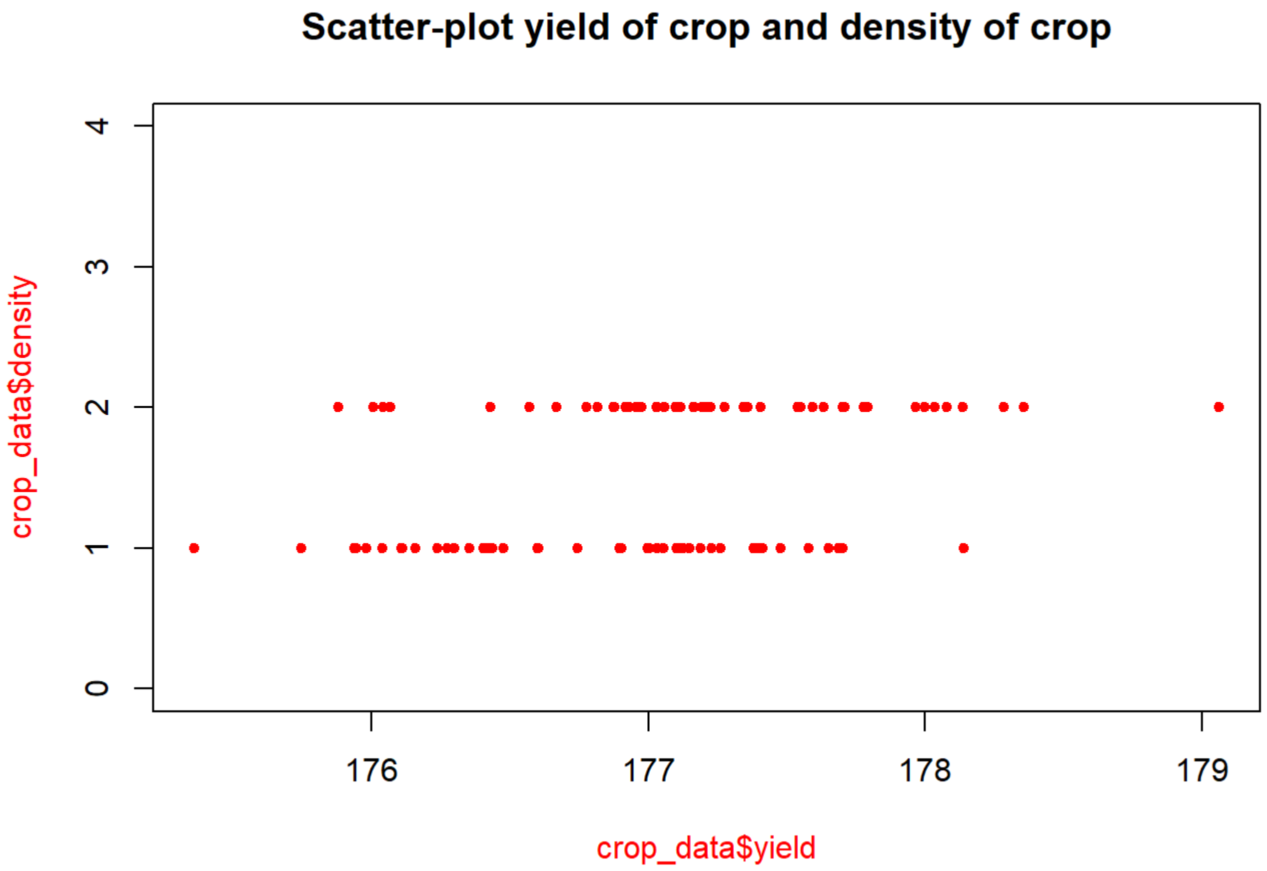
```
View(crop_data)

as.factor(crop_data$fertilizer)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2
## [39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
## [77] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## Levels: 1 2 3
```

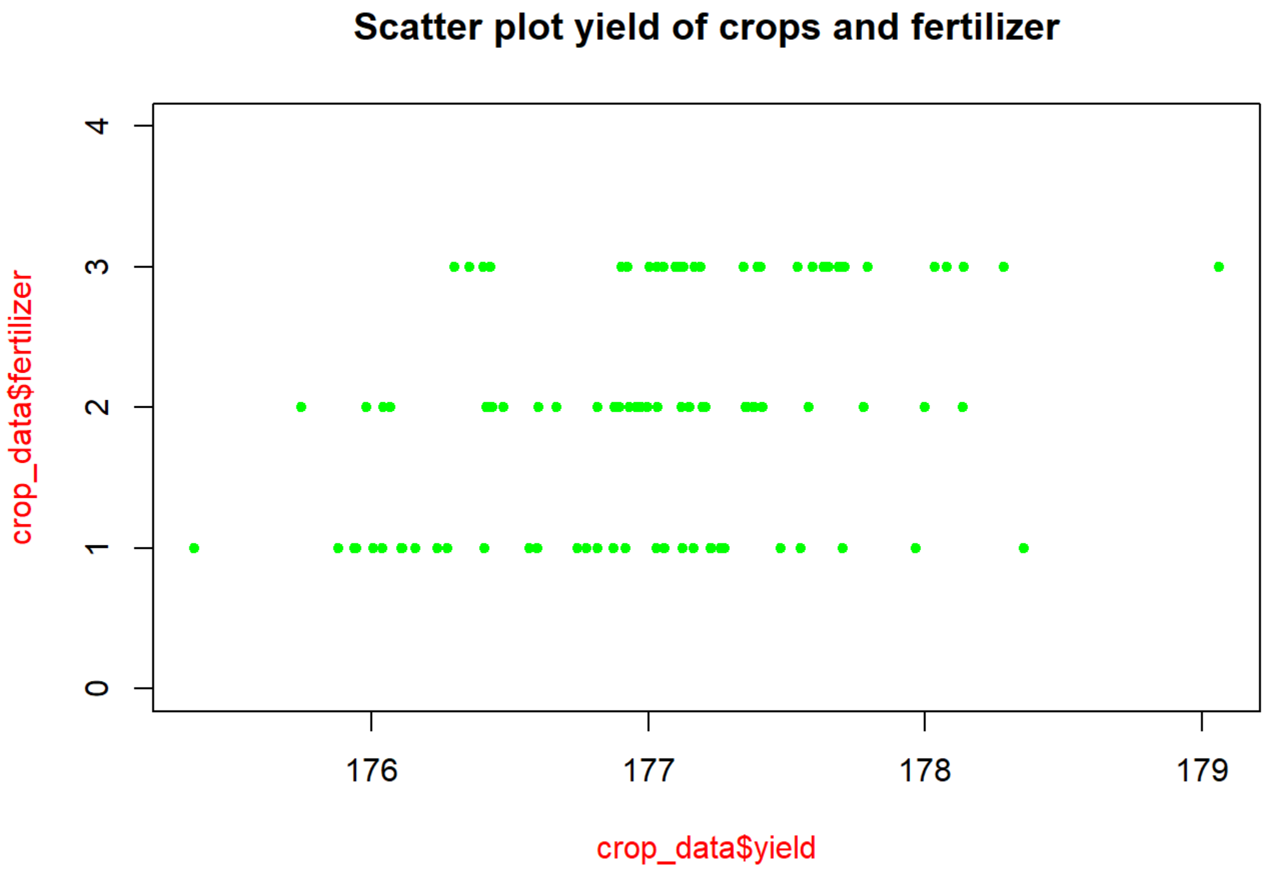
```
as.factor(crop_data$density)
```

```
##  [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## [39] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## [77] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
## Levels: 1 2
```

```
plot(crop_data$density~crop_data$yield,
     pch = 20, col="red", col.lab="red", bg= "#F504FC", ylim= c(0,4),
     main = "Scatter-plot yield of crop and density of crop")
```



```
plot(crop_data$fertilizer~crop_data$yield, pch = 20, col="green", col.lab="red", bg= "#F504FC",
 ylim= c(0,4),
     main = "Scatter plot yield of crops and fertilizer")
```



```
aplha = 0.05
cropsaov = aov(yield ~ density * fertilizer, data = crop_data)
summary(cropsaov)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)    
## density             1  5.122   5.122  15.230 0.000181 ***
## fertilizer          1  5.743   5.743  17.078  7.9e-05 ***
## density:fertilizer  1  0.150   0.150   0.447 0.505630    
## Residuals          92 30.939   0.336                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observation

In this task crop data is given with yield, density, block and fertilizer. The variables density and fertilizers are send to R factors. Using scatter-plot we can compare the relationship between crop density and crop yield. So if the density of the crops in the farm is 1 we are getting 176 to 177 yield in maximum. Likewise if we need to see the comparison between fertilizer and the crop yield we'll get the values.For two-way Anova test we can use yield as the dependent variable and fertilizer and density as the independent variable. From the Anova test we get the values sum, mean, favlue and the other comparitive values. That's how we use values to compare the best fit for the case.

Conclusion

In this module we used chi-square to test a distribution for goodness of fit, testing the two variables for independence and the proportions of homogeneity. We also performed one and two way Anova function.For the questions like blood types, on-time performance by airlines, ethnicity and movie admissions, women in military we used chi-square test and found out best fit and the values. And for the task like plant growth and crop we have used Anova test. The task performed are stating the hypothesis, finding the critical value, computing the test value, making a decision that the null hypothesis must be rejected or not. And finally comparing the critical value with the test value provided the same result as comparing the p-value from R with the significance level.

Reference

Two-Way ANOVA Test in R - Easy Guides - Wiki - STHDA. (2021). Sthda. http://www.sthda.com/english/wiki/two-way-anova-test-in-r (http://www.sthda.com/english/wiki/two-way-anova-test-in-r)

Team, D. (2021, August 25). Chi-Square Test in R | Explore the Examples and Essential concepts! DataFlair. https://data-flair.training/blogs/chi-square-test-in-r/ (https://data-flair.training/blogs/chi-square-test-in-r/)