Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6015: Intermediate Analytics**

**Assignment:**

Module 2 Assignment - Chi Square and ANOVA

**Submitted on:**

January 22, 2023

**Submitted to:**  **Submitted by:**

Professor: PAROMITA GUHA NIKSHITA RANGANATHAN

# **INTRODUCTION**

**Chi-Square test:**

The Chi-square test is an approach used to evaluate if there is a difference between projected and actual frequencies between various groups. It is used to evaluate the independence between two events or to see if there is a significant correlation between two variables.

The advantages of this test are that it is relatively easy to calculate and interpret and can be used to compare large numbers of variables. Additionally, it is a non-parametric test, meaning it does not require any assumptions about the distribution of the data.

The Chi-square test is utilized in numerous applications. In marketing research, it is used to compare the efficacy of multiple advertising campaigns and in medical research to evaluate the effectiveness of different treatments. Additionally, it is used in areas like survey analysis, quality assurance, and genetics. It is an effective tool for data analysis and decision-making across a wide range of industries.

**ANOVA test:**

An ANOVA test is employed to compare the group means It is used to determine whether the means of the various groups vary in any way. The basic idea behind ANOVA is to compare the variability amongst groups with the variability within groups.

Numerous areas, including psychology, economics, and medicine, use ANOVA. ANOVA can also be used to explore for variable interactions. It can be used, for instance, to find out whether age and gender have an influence on a certain activity. It can also be used to test for differences in variance between groups, such as differences in the variance of test scores between males and females.

There are two types of ANOVA tests.

One-way ANOVA: This type of ANOVA is used to compare the means of independent groups with respect to one factor.

Two-way ANOVA: This type of ANOVA is used to compare the means of independent groups on a range of factors.

**Steps involved in Chi-Square and ANOVA tests:**

1. State the alternative and null hypotheses.

2. Find the test statistic.

3. Compute the degrees of freedom.

4. Find the critical value.

5. Compare the test statistic to the critical value.

6. Make a decision about the null hypothesis.

7. Interpret the results.

**ANALYSIS & INTERPRETATION**

**Section 11-1**

**6. Blood Types**

**Null Hypothesis H0:** Type A=0.20, Type B=0.28, Type O=0.36, Type AB= 0.16

**Alternate Hypothesis H1:** The blood type distribution of hospital patients is not the same as the general population.

**Alpha value: 0.10**

**Table

Description automatically generated**

**Figure 1- Blood types matrix**

1. The critical value for a chi-square test in R can be calculated using the qchisq() function. The syntax for this function is qchisq(p, df), where p is the desired p-value and df is the degree of freedom.

**Critical Value: 6.251389**

**Chart

Description automatically generated with medium confidence**

**Figure 2 – Chi-square test of Blood types matrix**

**X-squared Test Value: 5.471429**

**Degree of Freedom: 3**

**P value: 0.1403575**

The p-value is greater than the alpha of 0.1

**Result: Fail to reject the null hypothesis**

Since the p-value exceeds the significance level(α) of 0.1, we are unable to reject the test's null hypothesis. There is insufficient evidence to claim that the blood type distribution of hospital patients is not the same as the general population.

**8. On-Time Performance by Airlines**

**Null Hypothesis H0:** On-Time = 0.708, National Aviation System Delay = 0.082,Aircraft Arriving Late = 0.09, Other (because of weather and other conditions) = 0.12

**Alternate Hypothesis H1:** The on-time performance distribution of airlines is not the same as the government's statistics.

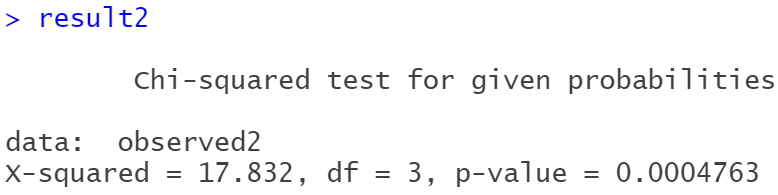
**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 3- Airline matrix**

**Critical Value: 7.814728**

1. ****

**Figure 4 – Chi-square test of Airline matrix**

**X-squared Test Value: 17.8325**

**Degree of Freedom: 3**

**P value: 0.0004762587**

The p-value is less than the alpha of 0.05

**Result: Reject the null hypothesis**

We reject the test's null hypothesis since the p-value is lower than the significance level(α) of 0.05. There is sufficient evidence to conclude that the airline on-time performance distribution of airlines is not the same as the government's official statistics.

**Section 11-2**

**8. Ethnicity and Movie Admissions**

**Null Hypothesis H0:** Movie admissions are independent to ethnicity

**Alternate Hypothesis H1:** Movie admissions are dependent to ethnicity

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 5- Movies matrix**

**Critical Value: 7.814728**

1. **Chart

   Description automatically generated with medium confidence**

**Figure 6 – Chi-square test of Movies matrix**

**X-squared Test Value: 60.14352**

**Degree of Freedom: 3**

**P value: 0.0000000000005477507**

The p-value is less than the alpha of 0.05

**Result: Reject the null hypothesis**

Because the p-value is less than the level of significance (α) of 0.05, we reject the null hypothesis of the test. There is sufficient evidence to claim that ethnicity influences movie attendance in the years 2013 and 2014.

**10. Women in the Military**

**Null Hypothesis H0:** Ranks and branches of Armed Forces are independent.

**Alternate Hypothesis H1:** Ranks and branches of Armed Forces are dependent.

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 7- Military matrix**.

**Critical Value: 7.814728**

1. **Text

   Description automatically generated with low confidence**

**Figure 8 – Chi-square test of Military matrix**

**X-squared Test Value: 654.2719**

**Degree of Freedom: 3**

**P value: 1.726418e-141**

The p-value is less than the alpha of 0.05

**Result: Reject the null hypothesis**

Since the p-value is less than the significance level of 0.05, we reject the test's null hypothesis. There is enough evidence to say that the different ranks and branches of the Armed Forces are interdependent.

**Section 12-1**

**8. Sodium Contents of Foods**

**Null Hypothesis H0:** μ\_Condiments = μ\_Cereals = μ\_Desserts

**Alternate Hypothesis H1:** There is a difference in mean sodium amounts among condiments, cereals, and desserts

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 9- Sodium dataframe**

**Critical Value: 3.521893***.*

1. Text

   Description automatically generated with medium confidence

**Figure 10 – Anova test of Sodium dataframe**

**ANOVA test F Value: 2.398538**

**Degree of Freedom: Numerator -2, Denominator -19**

**P value: 0.1178108**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

We fail to reject the test's null hypothesis since the p-value is greater than the significance level(α) of 0.05. There is insufficient evidence to prove that condiments, cereals, and sweets differ in their mean sodium contents.

**Section 12-2**

**10. Sales for Leading Companies**

**Null Hypothesis H0:** μ\_Cereals = μ\_Chocolatecandy = μ\_Coffee

**Alternate Hypothesis H1:** There is a difference in mean sales among cereals, chocolate candy and coffee

**Alpha value: 0.01**

**Table

Description automatically generated**

**Figure 11- Sales dataframe**

**Critical Value: 7.205713**

Text

Description automatically generated

**Figure 12 – Anova test of Sales dataframe**

**ANOVA test F Value: 2.171782**

**Degree of Freedom: Numerator -2, Denominator -11**

**P value: 0.1603487**

The p-value is greater than the alpha of 0.01

**Result: Fail to reject the null hypothesis**

We fail to reject the test's null hypothesis since the p-value is greater than the significance level(α) of 0.01. There is insufficient evidence to suggest that cereals, chocolate candy, and coffee have different mean sales.

The Tukey test is often used in ANOVA and other statistical tests to compare the means of different groups.

**Text

Description automatically generated**

**Figure 13 – Tukey test of Sales dataframe**

**12. Per-Pupil Expenditures**

**Null Hypothesis H0:** μ\_Eastern Third = μ\_Middle Third = μ\_Western Third

**Alternate Hypothesis H1:** There is a difference in mean expenditures of three sections of the country ( Eastern Third, Middle Third, and Western Third)

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 14- Expenditure dataframe**

**Critical Value: 4.102821**

Text

Description automatically generated with medium confidence

**Figure 15 – Anova test of Expenditure dataframe**

**ANOVA test F Value: 0.6488214**

**Degree of Freedom: Numerator -2, Denominator -10**

**P value: 0.5433264**

The p-value is greater than the alpha of 0.05

**Result: Fail to reject the null hypothesis**

We fail to reject the test's null hypothesis since the p-value is greater than the significance level(α) of 0.05. There is insufficient evidence to indicate that the mean expenditures of the three regions of the country differ from one another (Eastern Third, Middle Third, and Western Third).

**Section 12-3**

**10. Increasing Plant Growth**

**1st Pair: (Grow-light (Factor-1))**

**Null Hypothesis H0:** Meangrowth\_Grow-light 1 = Meangrowth\_Grow-light 2

**Alternate Hypothesis H1:** There is a difference in the mean growth of plants with respect to Grow-light 1 and Grow-light 2

**2nd Pair: (Plant-Food (Factor-2))**

**Null Hypothesis H0:** Meangrowth\_Plant-Food A = Meangrowth\_Plant-Food B

**Alternate Hypothesis H1:** There is a difference in mean growth with respect to Plant-Food A and Plant-Food B

**3rd Pair: (Factor1 & Factor2-Grow-light & Plant-Food)**

**Null Hypothesis H0:** There is no interaction between Plant-Food and Growth-Light.

**Alternate Hypothesis H1:** There is an interaction between the Plant-Food and Growth-Light.

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 16- Plants dataframe**

**Critical Value (Grow-light): 5.317655**

**Critical Value (Plant-Food): 5.317655**

1. **Critical Value (Grow-light & Plant-Food): 5.317655**

A picture containing table

Description automatically generated

**Figure 17 – Anova test of Plants dataframe**

**ANOVA test F Value (Grow-light): 3.680511**

**ANOVA test F Value (Plant-Food): 24.5623**

**ANOVA test F Value (Grow-light & Plant-Food): 1.4377**

**Degree of Freedom: Numerator (Grow-light): -1, Numerator (Plant-Food): -1, Numerator (Grow-light & Plant-Food): -1, Denominator -8**

**P value (Grow-light): 0.09133137**

**P value (Plant-Food): 0.001112418**

**P value (Grow-light & Plant-Food): 0.2648194**

The p-value **(Grow-light)** is greater than the alpha of 0.05

The p-value **(Plant-Food)** is less than the alpha of 0.05

The p-value **(Grow-light & Plant-Food)** is greater than the alpha of 0.05

**Result:**

**Failed to reject the null hypothesis**

**Reject the null hypothesis**

**Failed to reject the null hypothesis**

**(Grow-light):**

We are unable to reject the test's null hypothesis because the p-value exceeds the significance level of 0.05. There is insufficient evidence to imply that plants with Grow-light 1 and Grow-light 2 have different mean growths.

**(Plant-Food):**

Since the p-value is less than the significance level() of 0.05, we reject the test's null hypothesis. There is enough evidence to conclude that plants with Plant-Food A and Plant-Food B have different means growths.

**(Grow-light & Plant-Food):**

The p-value is greater than the significance level of 0.05, so we are unable to reject the null hypothesis of the test. There is not sufficient evidence to support the idea that Plant-Food and Growth-Light interact.

**Baseball**

With the help of str(), different datatypes can be noticed. This dataset consists of data types - characters, numbers, and integers. “baseball” consists of 1232 observations and 15 columns.

**A screenshot of a computer

Description automatically generated with medium confidence**

Text

Description automatically generated

**Figure 18 – str() and summary()**

glimpse() is a function that provides a quick summary of a data frame. It displays the number of rows and columns, the names of the variables, the type of each variable, and the first few rows of the data frame.

Text, letter

Description automatically generated

**Figure 19 – glimpse()**

**Descriptive statistics**

The skim() function in R provides a summary of the data including the number of observations, mean, standard deviation, quartiles, and other summary statistics. It is useful for exploring a dataset and understanding the distribution of values.

**Table

Description automatically generated with medium confidence**

**Figure 20 – skim()**

The mean of RS (Runs Scored) is 715.08, with a standard deviation of 91.53. 463 and 1009 runs, respectively, are the minimum and maximum values.

The mean of W (Wins) is 80.9, with a standard deviation of 11.46. 40 and 116 respectively are the minimum and maximum values.

Means of RS and RA (Runs scored and runs allowed) have mean values approximately equal to median values which indicates a normal distribution.

RS, RA, and W columns have skew values between -0.5 to 0.5 which signifies symmetric distribution.

Kurtosis of RS, RA, and W are less than 3 which indicates that their distribution is flat and less peaked (Platykurtic).

**Table

Description automatically generated with medium confidence**

**Figure 21 – describe()**

**Understanding the Correlation Matrix of baseball file**

Runs allowed and Wins are negatively correlated. Wins has a correlation coefficient of 0.5 with the Runs scored column which means it has a moderate positive correlation.

**Chart

Description automatically generated**

**Figure 22 – Correlation Matrix**

A correlogram is used to show how strongly two variables are related and can be used to spot trends or patterns in the data. Below graph gives the distribution of data variables.

Diagram

Description automatically generated

**Figure 23 – Correlogram**

Scatterplot signifies a relationship between two variables and here we can see both variables (runs scored and wins) are proportional to each other. This suggests that teams with higher runs scored are more likely to win games.

**Chart, scatter chart

Description automatically generated**

**Figure 24 – Scatterplot**

In Figure 25, we can see the fluctuations in the number of wins across the decades. During the 2000s, maximum wins were secured by the teams in both leagues.

**Chart, bar chart

Description automatically generated**

**Figure 25 – Stacked bar graph**

There is a significant increase in the number of years through the years 1960 – 2010.

**Chart, line chart

Description automatically generated**

**Figure 26 – Line chart**

**Null Hypothesis H0:** The number of wins by decade does not differ from one another.

**Alternate Hypothesis H1:** There is a difference in the number of wins by decade

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 27- Wins\_decade**

**Critical Value: 11.0705**

Text

Description automatically generated with medium confidence

**Figure 28 – Chi-square test of Wins\_decade**

**X-squared Test Value: 1558.5**

**Degree of Freedom: 5**

**P value: 0**

The p-value is less than the alpha of 0.05

**Result: Reject the null hypothesis**

Since the p-value falls below the significance level of 0.05, we reject the null hypothesis of the test. There is sufficient proof to conclude that the number of wins varies by decade.

**Crop Data**

Data was imported from a CSV file and datatypes were changed from integers to factors.

Yield – Dependent variable

Fertilizer and density – Independent variable

**1st Pair: (Fertilizer (Factor-1))**

**Null Hypothesis H0:** Mean Yield\_Fertilizer 1 = Mean Yield\_Fertilizer 2 = Mean Yield\_Fertilizer 3

**Alternate Hypothesis H1:** There is a difference in mean yields due to Fertilizer 1, Fertilizer 2 and Fertilizer 3

**2nd Pair: (Density (Factor-2))**

**Null Hypothesis H0:** Mean Yield\_Density 1 = Mean Yield\_Density 2

**Alternate Hypothesis H1:** There is a difference in mean yields due to Density 1 and Density 2

**3rd Pair: (Factor1 & Factor2-** **Fertilizer & Density)**

**Null Hypothesis H0:** There is no interaction between density and fertilizer.

**Alternate Hypothesis H1:** There is an interaction between density and fertilizer.

**Alpha value: 0.05**

**Table

Description automatically generated**

**Figure 29- Crop dataframe**

**Critical Value (Fertilizer): 3.097698**

**Critical Value (Density): 3.946876**

1. **Critical Value (Fertilizer & Density): 3.097698**

Text

Description automatically generated

**Figure 30 – Anova test of Crop dataframe**

**ANOVA test F Value (Fertilizer): 9.001052**

**ANOVA test F Value (Density): 15.19452**

**ANOVA test F Value (Fertilizer & Density): 0.6346053**

**Degree of Freedom: Numerator (Fertilizer): -2, Numerator (Density): -1, Numerator (Fertilizer & Density): -2, Denominator -90**

**P value (Fertilizer): 0.000273189**

**P value (Density): 0.0001864075**

**P value (Fertilizer & Density): 0.5325001**

The p-value **(Fertilizer)** is less than the alpha of 0.05

The p-value **(Density)** is less than the alpha of 0.05

The p-value **(Fertilizer & Density)** is greater than the alpha of 0.05

**Result:**

**Reject the null hypothesis**

**Reject the null hypothesis**

**Fail to reject the null hypothesis**

**(Fertilizer):**

We reject the null hypothesis of the test because the p-value falls below the level of significance(α) of 0.05. There is sufficient evidence to imply that Fertilizer 1, Fertilizer 2, and Fertilizer 3 have different mean yields.

**(Density):**

We reject the null hypothesis of the test because the p-value falls below the level of significance(α) of 0.05. There is sufficient evidence to imply that Density 1 and Density 2 have different mean yields.

**(Fertilizer & Density):**

Because the p-value is above the level of significance of 0.05, we are unable to reject the null hypothesis of the test. There is insufficient evidence to imply that there is an interaction between the Fertilizer and Density.

**CONCLUSION**

We used Chi-Square and Analysis of variance tests to analyze various datasets for this assignment.

The key findings are :

* ANOVA and Chi-Square tests are essential tools for data professionals to use in order to make valid inferences from their data.
* In the baseball question, we can notice that the number of wins by decade differs significantly. This could be due to various factors, such as the talent level of players, changes in the strategies used by teams, etc. The results of this test can also be used to understand the history of the game and how it has changed over time. Knowing the differences in wins by decade can help teams make better decisions and become more successful.
* For the crop dataset, we performed a two-way ANOVA test to test the effects of different fertilizers and densities on the yield of a crop. The results of this test suggest that the yield of the crop is affected by both the fertilizer and the density, but not by the interaction between the two.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

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

GeeksforGeeks. (2020, June 30). Chi-Square Test in R. https://www.geeksforgeeks.org/chi-square-test-in-r/

**APPENDIX: CODE**

#---------------------- Week\_2\_Module\_2 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Module 2 Assignment - Chi Square and ANOVA")

print("Course Name - ALY6015: Intermediate Analytics")

# Installing and loading the packages

library(dplyr)

library(psych)

library(tibble)

library(skimr)

library(corrgram)

library(GGally)

library(ggplot2)

library(hrbrthemes)

library(wesanderson)

# Section 11-1

# Q6. Blood Type

# State the hypotheses

# H0: Type A=0.20, Type B=0.28, Type O=0.36, Type AB= 0.16

# H1: The blood type distribution of hospital patients is not the same as the general population

# Creating matrix

Data\_bloodtype <- matrix(c(0.20, 0.28, 0.36, 0.16,12, 8, 24, 6),ncol=2,byrow=FALSE)

colnames(Data\_bloodtype) <- c("Expected","Observed")

rownames(Data\_bloodtype) <- c("Type A", "Type B","Type O","Type AB")

Data\_bloodtype

# Set significance level

alpha<-0.1

# Create a vector of the values

observed <- c(12, 8, 24, 6)

# Create a vector of the probabilities

p <- c(0.20, 0.28, 0.36, 0.16)

# Run the test and save the results in result variable

result <- chisq.test(x = observed, p = p)

# View the test statistic and p-value

result$statistic

result$p.value

result$parameter

result

# Critical Value

qchisq(alpha,result$parameter, lower.tail=FALSE)

# Compare the p-value to alpha and make a decision

ifelse(result$p.value > alpha,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q8. On-Time Performance by Airlines

# State the hypotheses

# H0: On-Time = 0.708, National Aviation System Delay = 0.082,Aircraft Arriving Late = 0.09, Other (because of weather and other conditions) = 0.12

# H1: The on-time performance distribution of airlines is not the same as government's statistics

# Creating matrix

Data\_airlines <- matrix(c(0.708, 0.082, 0.09, 0.12,125, 10, 25, 40),ncol=2,byrow=FALSE)

colnames(Data\_airlines) <- c("Expected","Observed")

rownames(Data\_airlines) <- c("On time", "National Aviation System delay","Aircraft arriving late","Other (because of weather and other conditions)")

Data\_airlines

# Set significance level

alpha2 = 0.05

# Create a vector of the values

observed2 <- c(125, 10, 25, 40)

# Create a vector of the probabilities

p2 <- c(0.708, 0.082, 0.09, 0.12)

# Run the test and save the results to result variable

result2 <- chisq.test(x = observed2, p = p2)

# View the test statistic and p-value

result2$statistic

result2$p.value

result2$parameter

result2

# Critical Value

qchisq(alpha2,result2$parameter, lower.tail=FALSE)

# Compare the p-value to alpha and make a decision

ifelse(result2$p.value > alpha2,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 11-2

# Q8. Ethnicity and Movie Admissions

# State the hypotheses

# H0: Movie admissions are independent to ethnicity

# H1: Movie admissions are dependent to ethnicity

# Set significance level

alpha3 = 0.05

# Create a vector for each row

year2013<-c(724, 335, 174, 107)

year2014<-c(370, 292, 152, 140)

# State the number of rows for the matrix

rows1=2

# Create a matrix from the rows

Movies=matrix(c(year2013,year2014),nrow=rows1,byrow=TRUE)

# Name the rows and columns of the matrix

rownames(Movies) <- c("2013", "2014")

colnames(Movies) <- c("Caucasian", "Hispanic", "African American", "Other")

# View the matrix

Movies

# Run the test and save the results to result variable

result3 <- chisq.test(Movies)

# View the test statistic and p-value

result3$statistic

result3$p.value

result3$parameter

result3

# Critical Value

qchisq(alpha3,result3$parameter, lower.tail=FALSE)

# Compare the p-value to alpha and make a decision

ifelse(result3$p.value > alpha3,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Q10. Women in the Military

# State the hypotheses

# H0: Ranks and branches of Armed Forces are independent

# H1: Ranks and branches of Armed Forces are dependent

# Set significance level

alpha4 = 0.05

# Create one vector for each row

army <- c(10791, 62491)

navy <- c(7816, 42750)

marine\_corps <- c(932, 9525)

air\_force <- c(11819, 54344)

# State the number of rows for the matrix

rows2=4

# Create a matrix from the rows

Military=matrix(c(army,navy,marine\_corps,air\_force),nrow=rows2,byrow=TRUE)

# Name the rows and columns of the matrix

rownames(Military) <- c("Army", "Navy", "Marine Corps", "Air Corps")

colnames(Military) <- c("Officers", "Enlisted")

# View the matrix

Military

# Run the test and save the results to result variable

result4 <- chisq.test(Military)

# View the test statistic and p-value

result4$statistic

result4$p.value

result4$parameter

result4

# Critical Value

qchisq(alpha4,result4$parameter, lower.tail=FALSE)

# Compare the p-value to alpha and make a decision

ifelse(result4$p.value > alpha4,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 12-1

# Q8. Sodium Contents of Foods

# State the hypotheses

# H0: μ\_Condiments = μ\_Cereals = μ\_Desserts

# H1: At least one mean is different from the others

# Set significance level

alpha5 = 0.05

# Create a dataframe for Condiments

Condiments <- data.frame('Sodium' = c(270, 130, 230, 180, 80, 70, 200), 'Food' = rep('condiments', 7),stringsAsFactors = FALSE)

# Create a data frame for the Cereals

Cereals <- data.frame('Sodium' = c(260, 220, 290, 290, 200, 320, 140), 'Food' = rep('cereals', 7), stringsAsFactors= FALSE)

# Create a data frame for the Desserts

Desserts <- data.frame('Sodium' = c(100, 180, 250, 250, 300, 360, 300, 160), 'Food' = rep('desserts', 8),stringsAsFactors = FALSE)

# Combine the data frames into one

Sodium <- rbind(Condiments, Cereals, Desserts)

Sodium$Food <- as.factor(Sodium$Food)

# Run the ANOVA test

anova <- aov(Sodium ~ Food, data = Sodium)

# View the model summary

summary(anova)

# Save the summary to an object

a.summary <- summary(anova)

a.summary

# Degrees of Freedom

# k - 1: between group variance - numerator

df.numerator <- a.summary[[1]][1, "Df"]

df.numerator

# N - k: within group variance - denominator

df.denominator <- a.summary[[1]][2, "Df"]

df.denominator

# Extract the F test value from the summary

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

f.value

# Extract the P-value value from the summary

p.value <- a.summary[[1]][[1, "Pr(>F)"]]

p.value

# Critical Value

qf(p=alpha5, df1=df.numerator, df2 = df.denominator, lower.tail=FALSE)

# Determine if we should reject the null hypothesis

ifelse(p.value>alpha5,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Section 12-2

# Q10. Sales for Leading Companies

# State the hypotheses

# H0: μ\_Cereals = μ\_Chocolatecandy = μ\_Coffee

# H1: At least one mean is different from the others

# Set Significance Level

alpha6 = 0.01

# Create a data frame for the cereals

Cereal <- data.frame('Sales' = c(578, 320, 264, 249, 237), 'Food' = rep('cereal', 5), stringsAsFactors = FALSE)

# Create a data frame for the chocolate Candy

Chocolatecandy <- data.frame('Sales' = c(311, 106, 109, 125, 173), 'Food' = rep('chocolate candy', 5),stringsAsFactors = FALSE)

# Create a data frame for the coffee

Coffee <- data.frame('Sales' = c(261, 185, 302, 689), 'Food' = rep('coffee', 4), stringsAsFactors = FALSE)

# Combine the data frames into one

Sales <- rbind(Cereal, Chocolatecandy, Coffee)

Sales$Food <- as.factor(Sales$Food)

# Run the ANOVA test

anova2 <- aov(Sales ~ Food, data = Sales)

# View the model summary and save it

a.summary2 <- summary(anova2)

a.summary2

# Degrees of Freedom

# k - 1: Between Group Variance - Numerator

df.numerator2 <- a.summary2[[1]][1, "Df"]

df.numerator2

# N - k: Within Group Variance - Denominator

df.denominator2 <- a.summary2[[1]][2, "Df"]

df.denominator2

# Extract the F test value from the summary

f.value2 <- a.summary2[[1]][[1, "F value"]]

f.value2

# Extract the P-value from the summary

p.value2 <- a.summary2[[1]][[1, "Pr(>F)"]]

p.value2

# Critical Value

qf(p=alpha6, df1=df.numerator2, df2 = df.denominator2, lower.tail=FALSE)

# Determine if we should reject the null hypothesis

ifelse(p.value2>alpha6,"Fail to reject the null hypothesis","Reject the null hypothesis")

# See differences

TukeyHSD(anova2)

# Q12. Per-Pupil Expenditures

# State the hypotheses

# H0: μ\_Eastern Third = μ\_Middle Third = μ\_Western Third

# H1: At least one mean is different from the others

# Set Significance Level

alpha7 = 0.05

# Create a data frame for the Eastern Third

Easternthird <- data.frame('Expenditure' = c(4946, 5953, 6202, 7243, 6113), 'States' = rep('Eastern Third', 5),stringsAsFactors = FALSE)

# Create a data frame for the Middle Third

Middlethird <- data.frame('Expenditure' = c(6149, 7451, 6000, 6479), 'States' = rep('Middle Third', 4),stringsAsFactors = FALSE)

# Create a data frame for the Western Third

Westernthird <- data.frame('Expenditure' = c(5282, 8605, 6528, 6911), 'States' = rep('Western Third', 4),stringsAsFactors = FALSE)

# Combine the data frames into one

Expenditure <- rbind(Easternthird, Middlethird, Westernthird)

Expenditure$States <- as.factor(Expenditure$States)

# Run the ANOVA test

anova3 <- aov(Expenditure ~ States, data = Expenditure)

# View the model summary and save it

a.summary3 <- summary(anova3)

a.summary3

# Degrees of Freedom

# k - 1: Between Group Variance - Numerator

df.numerator3 <- a.summary3[[1]][1, "Df"]

df.numerator3

# N - k: Within Group Variance - Denominator

df.denominator3 <- a.summary3[[1]][2, "Df"]

df.denominator3

# Extract the F test value from the summary

f.value3 <- a.summary3[[1]][[1, "F value"]]

f.value3

# Extract the P-value from the summary

p.value3 <- a.summary3[[1]][[1, "Pr(>F)"]]

p.value3

# Critical Value

qf(p=alpha7, df1=df.numerator3, df2 = df.denominator3, lower.tail=FALSE)

# Determine if we should reject the null hypothesis

ifelse(p.value3>alpha7,"Fail to reject the null hypothesis","Reject the null hypothesis")

# See differences

TukeyHSD(anova3)

# Section 12-3

# Q10. Increasing Plant Growth

# State the hypotheses

# 1. Hypotheses for Grow-light (Factor-1)

# H0: μ\_Grow-light 1 = μ\_Grow-light 2

# H1: There is a difference in means of Grow-light 1 and Grow-light 2

# 2. Hypotheses for Plant-Food (Factor-2)

# H0: μ\_Plant-Food A = μ\_Plant-Food B

# H1: There is a difference in means of Plant-Food A and Plant-Food B

# 3. Hypotheses for interaction (Factor1:Factor2-Grow-light:Plant-Food)

# H0: There is no interaction between the Growth-Light and Plant-Food

# H1: There is interaction between the Growth-Light and Plant-Food

# Set Significance Level

alpha8 = 0.05

# Create a data frame

Data\_plants <- data.frame('plant\_growth' = c(9.2, 9.4, 8.9, 8.5, 9.2, 8.9, 7.1, 7.2, 8.5, 5.5, 5.8, 7.6),'grow\_light' = c('1', '1', '1', '2', '2', '2', '1', '1', '1', '2', '2', '2'),'plant\_food' = c('A', 'A', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'B', 'B'),stringsAsFactors = TRUE)

# Run the ANOVA test

anova4 <- aov(plant\_growth ~ grow\_light + plant\_food + grow\_light:plant\_food, data = Data\_plants)

# View the model summary

summary(anova4)

# Save the summary to an object

a.summary4 <- summary(anova4)

a.summary4

# Degrees of Freedom

# k - 1: Between Group Variance - Numerator (Grow-light (Factor-1))

df.numerator\_growlight <- a.summary4[[1]][1, "Df"]

df.numerator\_growlight

# k - 1: Between Group Variance - Numerator (Plant-Food (Factor-2))

df.numerator\_plantfood <- a.summary4[[1]][2, "Df"]

df.numerator\_plantfood

# k - 1: Between Group Variance - Numerator (Factor1:Factor2-Grow-light:Plant-Food)

df.numerator\_growlight\_plantfood <- a.summary4[[1]][3, "Df"]

df.numerator\_growlight\_plantfood

# N - k: Within Group Variance - Denominator

df.denominator4 <- a.summary4[[1]][4, "Df"]

df.denominator4

# Extract the F test value (Grow-light (Factor-1))

f.value\_growlight <- a.summary4[[1]][[1, "F value"]]

f.value\_growlight

# Extract the F test value (Plant-Food (Factor-2))

f.value\_plantfood <- a.summary4[[1]][[2, "F value"]]

f.value\_plantfood

# Extract the F test value (Factor1:Factor2-Grow-light:Plant-Food)

f.value\_growlight\_plantfood <- a.summary4[[1]][[3, "F value"]]

f.value\_growlight\_plantfood

# Extract the P-value (Grow-light (Factor-1))

p.value\_growlight <- a.summary4[[1]][[1, "Pr(>F)"]]

p.value\_growlight

# Extract the P-value (Plant-Food (Factor-2))

p.value\_plantfood <- a.summary4[[1]][[2, "Pr(>F)"]]

p.value\_plantfood

# Extract the P-value (Factor1:Factor2-Grow-light:Plant-Food)

p.value\_growlight\_plantfood <- a.summary4[[1]][[3, "Pr(>F)"]]

p.value\_growlight\_plantfood

# Critical Value (Grow-light (Factor-1))

qf(p=alpha8, df1=df.numerator\_growlight, df2 = df.denominator4, lower.tail=FALSE)

# Critical Value (Plant-Food (Factor-2))

qf(p=alpha8, df1=df.numerator\_plantfood, df2 = df.denominator4, lower.tail=FALSE)

# Critical Value (Factor1:Factor2-Grow-light:Plant-Food)

qf(p=alpha8, df1=df.numerator\_growlight\_plantfood, df2 = df.denominator4, lower.tail=FALSE)

# Determine if we should reject the null hypothesis (Grow-light (Factor-1))

ifelse(p.value\_growlight>alpha7,"There is no difference in means of Grow-light 1 and Grow-light 2","There is a difference in means of Grow-light 1 and Grow-light 2")

# Determine if we should reject the null hypothesis (Plant-Food (Factor-2))

ifelse(p.value\_plantfood>alpha7,"There is no difference in means of Plant-Food A and Plant-Food B","There is a difference in means of Plant-Food A and Plant-Food B")

# Determine if we should reject the null hypothesis (Factor1:Factor2-Grow-light:Plant-Food)

ifelse(p.value\_growlight\_plantfood>alpha7,"There is no interaction between the Growth-Light and Plant-Food","There is interaction between the Growth-Light and Plant-Food")

# Baseball Question

# Importing the dataset

getwd()

baseball<-read.csv("baseball.csv")

# Descriptive statistics & EDA

str(baseball)

summary(baseball)

headTail(baseball)

dim(baseball)

skim(baseball)

describe(baseball,quant = c(0.25, 0.75),IQR = T)

glimpse(baseball)

baseball <- baseball %>% mutate(RankSeason = ifelse(is.na(RankSeason), 0,RankSeason)) %>%

mutate(RankPlayoffs = ifelse(is.na(RankPlayoffs), 0,RankPlayoffs))

# Correlation Matrix

corr <- select\_if(baseball, is.numeric)

ggcorr(corr,nbreaks = 6,label = TRUE,label\_size = 3,hjust = 0.75,size=3,color = "black")

corrgram(corr, lower.panel=panel.pts, upper.panel=panel.conf,diag.panel=panel.density)

# Graph 1 - Scatterplot

ggplot(baseball, aes(x=W, y=RS))+geom\_point(aes(color=factor(Playoffs)))+theme\_ipsum()+labs(colour = "Playoffs",y = "Runs Scored", x = "Wins")+ggtitle("Scatterplot between Runs scored and Wins")+ scale\_color\_manual(values= wes\_palette("Darjeeling1", n = 2))

baseball$decade <- baseball$Year - (baseball$Year %% 10)

Wins\_decade\_league<-baseball %>% group\_by(decade,League) %>% summarize(wins=sum(W))

Wins\_byyear<-baseball %>% group\_by(Year,League) %>% summarize(wins=sum(W))

Wins\_decade\_league %>% ggplot( aes(x=decade, y=wins,fill=League)) + geom\_bar(position="stack", stat="identity")+ scale\_fill\_brewer(palette="Accent")+ggtitle("Stacked Bar chart")+labs(x = "Decade", y = "Wins")

Wins\_byyear %>% ggplot( aes(x=Year, y=wins,colour=League)) + geom\_line(size=1) +scale\_color\_manual(values= wes\_palette("GrandBudapest1", n = 2))+ggtitle("Line Chart")+labs(x = "Year", y = "Wins")

Wins\_decade<-baseball %>% group\_by(decade) %>% summarize(wins=sum(W))%>% as.tibble()

# State the hypotheses

# H0: There is no difference in number of wins by decade

# H1: There is difference in number of wins by decade

# Set Significance Level

alpha9 = 0.05

# Run the test and save the results

result\_baseball <- chisq.test(x = Wins\_decade)

# View the test statistic and p-value

result\_baseball$statistic

result\_baseball$p.value

result\_baseball$parameter

result\_baseball

# Critical Value

qchisq(alpha9,result\_baseball$parameter, lower.tail=FALSE)

# Compare the p-value to alpha and make a decision

ifelse(result\_baseball$p.value > alpha9,"Fail to reject the null hypothesis","Reject the null hypothesis")

# Crop Data Question

# Importing the dataset

getwd()

Data\_crop<-read.csv("crop\_data.csv")

# Changing datatypes to factors

Data\_crop$density=as.factor(Data\_crop$density)

Data\_crop$fertilizer=as.factor(Data\_crop$fertilizer)

Data\_crop$block=as.factor(Data\_crop$block)

# Descriptive statistics

str(Data\_crop)

summary(Data\_crop)

headTail(Data\_crop)

dim(Data\_crop)

skim(Data\_crop)

describe(Data\_crop,quant = c(0.25, 0.75),IQR = T)

glimpse(Data\_crop)

# State the hypotheses

# 1. Hypotheses for Fertilizer (Factor-1)

# H0: μ\_Fertilizer 1 = μ\_Fertilizer 2 = μ\_Fertilizer 3

# H1: There is a difference in means of Fertilizer 1,Fertilizer 2 and Fertilizer 3

# 2. Hypotheses for Density (Factor-2)

# H0: μ\_Density 1 = μ\_Density 2

# H1: There is a difference in means of Density 1 and Density 2

# 3. Hypotheses for interaction (Factor1:Factor2-Fertilizer:Density)

# H0: There is no interaction between the Fertilizer and Density

# H1: There is interaction between the Fertilizer and Density

# Set Significance Level

alpha10 = 0.05

# Run the ANOVA test

anova5 <- aov(yield ~ fertilizer + density + fertilizer:density, data = Data\_crop)

# View the model summary

summary(anova5)

# Save the summary to an object

a.summary5 <- summary(anova5)

a.summary5

# Degrees of Freedom

# k - 1: Between Group Variance - Numerator (Fertilizer (Factor-1))

df.numerator\_fertilizer <- a.summary5[[1]][1, "Df"]

df.numerator\_fertilizer

# k - 1: Between Group Variance - Numerator (Density (Factor-2))

df.numerator\_density <- a.summary5[[1]][2, "Df"]

df.numerator\_density

# k - 1: Between Group Variance - Numerator (Factor1:Factor2-Fertilizer:Density)

df.numerator\_fertilizer\_density <- a.summary5[[1]][3, "Df"]

df.numerator\_fertilizer\_density

# N - k: Within Group Variance - Denominator

df.denominator5 <- a.summary5[[1]][4, "Df"]

df.denominator5

# Extract the F test value (Fertilizer (Factor-1))

f.value\_fertilizer <- a.summary5[[1]][[1, "F value"]]

f.value\_fertilizer

# Extract the F test value (Density (Factor-2))

f.value\_density <- a.summary5[[1]][[2, "F value"]]

f.value\_density

# Extract the F test value (Factor1:Factor2-Fertilizer:Density)

f.value\_fertilizer\_density <- a.summary5[[1]][[3, "F value"]]

f.value\_fertilizer\_density

# Extract the P-value (Fertilizer (Factor-1))

p.value\_fertilizer <- a.summary5[[1]][[1, "Pr(>F)"]]

p.value\_fertilizer

# Extract the P-value (Density (Factor-2))

p.value\_density <- a.summary5[[1]][[2, "Pr(>F)"]]

p.value\_density

# Extract the P-value (Factor1:Factor2-Fertilizer:Density)

p.value\_fertilizer\_density <- a.summary5[[1]][[3, "Pr(>F)"]]

p.value\_fertilizer\_density

# Critical Value (Fertilizer (Factor-1))

qf(p=alpha10, df1=df.numerator\_fertilizer, df2 = df.denominator5, lower.tail=FALSE)

# Critical Value (Density (Factor-2))

qf(p=alpha10, df1=df.numerator\_density, df2 = df.denominator5, lower.tail=FALSE)

# Critical Value (Factor1:Factor2-Fertilizer:Density)

qf(p=alpha10, df1=df.numerator\_fertilizer\_density, df2 = df.denominator5, lower.tail=FALSE)

# Determine if we should reject the null hypothesis (Fertilizer (Factor-1))

ifelse(p.value\_fertilizer>alpha10,"There is a no difference in means of Fertilizer 1,Fertilizer 2 and Fertilizer 3","There is a difference in mean yields for Fertilizer 1,Fertilizer 2 and Fertilizer 3")

# Determine if we should reject the null hypothesis (Density (Factor-2))

ifelse(p.value\_density>alpha10,"There is a no difference in means of Density 1 and Density 2","There is a difference in mean yields for Density 1 and Density 2")

# Determine if we should reject the null hypothesis (Factor1:Factor2-Fertilizer:Density)

ifelse(p.value\_fertilizer\_density>alpha10,"There is no interaction between the Fertilizer and Density","There is interaction between the Fertilizer and Density")