Module 2: Chi-Square Testing and ANOVA

Mohit Ravindra Kamble

College of Professional Studies – Northeastern University

ALY6015: Intermediate Analytics

Prof. Roy Wada

January 23, 2024

01. Overview:

In this assignment, various statistical analyses were conducted using R to address different research questions. The tasks covered hypothesis testing, Chi-Square tests, One-way and Two-way ANOVA tests, as well as exploratory data analysis (EDA). The analyses involved diverse datasets and aimed to explore relationships between variables and draw meaningful conclusions. For instance, in Section 11-1, Chi-Square tests were applied to investigate blood type distributions in a hospital compared to the general population. In Section 12-2, a One-way ANOVA test was performed to examine the sales of leading companies in different categories. Additionally, a Two-way ANOVA test in Section 12-3 explored the interaction between factors influencing plant growth. The report also included data visualization, such as correlation matrices and histograms, to enhance the understanding of the datasets. Overall, the assignment demonstrated proficiency in statistical analysis using R and showcased the ability to interpret and communicate results effectively.

02. Analysis:

2.1: Solving Problems.

❖ Chi-Square Test in R:

I used the Chi-Square Test in R to explore relationships in categorical data. The test helps determine if observed data differs significantly from expected values. In my analysis, I set a significance level (alpha) to assess the results. If the p-value is below alpha, I reject the null hypothesis, suggesting a meaningful association. Conversely, if the p-value is higher, I fail to reject the null hypothesis, indicating no significant difference. This test is valuable for understanding if certain factors, like blood types or movie admissions, are independent or linked, providing insights into categorical data patterns.

Also used is Pearson's Chi-Square Test in R to analyze categorical data relationships. This statistical test assesses if observed frequencies in different categories differ from expected frequencies. Setting a significance level (alpha), I compared the p-value to decide on the null hypothesis. A low p-value leads to rejecting the null hypothesis, indicating a significant relationship, while a high p-value suggests no substantial difference. This test is useful for exploring connections in data sets, like blood type distributions or movie attendance by ethnicity, offering a straightforward way to understand if variables are independent or associated in categorical data analysis.

Section 11-1: Blood Types.

I conducted a Chi-Square test to check if blood type distribution in a hospital matches the general population. Using a significance level of 0.10, the observed distribution (12 A, 8 B, 24 O, 6 AB) was compared to the expected general population percentages (A=20%, B=28%, O=36%, AB=16%). The result suggests that, at a 10% significance level, *we fail to reject the null hypothesis*, indicating no significant difference between the hospital's blood type distribution and the general population. This means *there's no strong evidence to conclude they are different*.

O/P:

```
      Chi-squared test for given probabilities

data:  observed
X-squared = 5.4714, df = 3, p-value = 0.1404

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject
the null hypothesis")
[1] "Fail to reject the null hypothesis"
```

Section 11-1: On-Time Performance.

I used a Chi-Square test to examine if a major airline's punctuality aligns with government statistics. With a 0.05 significance level, I compared the observed data (125 on time, 10 NAS delay, 25 arriving late, 40 weather-delayed) to the expected percentages from government stats (On time=70.8%, NAS delay=8.2%, Aircraft arriving late=9%, Other=12%). The outcome indicates, at a 5% significance level, *rejecting the null hypothesis*. This *implies a noteworthy difference* between the airline's on-time performance and the government's anticipated distribution, suggesting a divergence from the expected pattern.

O/P:

```
      Chi-squared test for given probabilities

data:  observed
X-squared = 17.832, df = 3, p-value = 0.0004763

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject
the null hypothesis")
[1] "Reject the null hypothesis"
```

Section 11-2: Ethnicity and Movie Admissions.

I analyzed data on movie admissions for the years 2013 and 2014, exploring the potential connection with ethnicity through a Chi-Square test. The matrix included admissions for Caucasians, Hispanics, African Americans, and Others. The null hypothesis assumed independence, while the alternative proposed a relationship. With a significance level of 0.05, the test yielded a p-value below the threshold. Consequently, I *rejected the null hypothesis*, *signifying a significant correlation between movie admissions and ethnicity*. This implies that attendance patterns vary among ethnic groups across the two years.

O/P:

```
        Pearson's Chi-squared test

data:  mtrx
X-squared = 60.144, df = 3, p-value = 0.0000000000005478


# Making a decision by comparing the p-value
> ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject
the null hypothesis")
[1] "Reject the null hypothesis"
```

Section 11-2: Women in the Military.

I applied the Chi-Square Test in R to investigate if a relationship exists between the rank and branch of the Armed Forces for women. After analyzing officer and enlisted numbers across the Army, Navy, Marine Corps, and Air Force, I set a significance level (alpha) of 0.05. The test resulted in *rejecting the null hypothesis*, providing sufficient evidence to conclude that a significant relationship exists. This suggests that the *distribution of women in military ranks varies across different branches*. The Chi-Square Test, by comparing observed and expected frequencies, helped me make this determination, contributing to a better understanding of women's representation in the military.

O/P:

```
        Pearson's Chi-squared test

data:  mtrx
X-squared = 654.27, df = 3, p-value < 0.00000000000000022
```

```
# Making a decision by comparing the p-value
> ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject
the null hypothesis")
[1] "Reject the null hypothesis"
```

Section 12-1: Sodium Contents of Foods.

I performed a Chi-Square Test in R to investigate whether there is significant evidence indicating a variance in mean sodium levels among condiments, cereals, and desserts. Utilizing a random sample of three food categories with corresponding sodium values, I set the significance level (alpha) at 0.05. *The outcome of the test led me to retain the null hypothesis*, implying *insufficient evidence to assert a noteworthy difference in mean sodium levels among the specified food groups*. This statistical examination aids in comprehending and comparing sodium content across diverse food types, offering valuable insights for dietary assessments.

O/P:

```
        Pearson's Chi-squared test

data:  ct
X-squared = 4.0366, df = 4, p-value = 0.4011


# Making a decision by comparing the p-value for Chi-Square Test
> ifelse(p_value > alpha, "Fail to reject the null hypothesis", "Reject the
null hypothesis")
[1] "Fail to reject the null hypothesis"
```

- **ANOVA Test in R:**

I used an ANOVA test in R to check if there's a notable difference in means among multiple groups. Taking sales data from various leading companies in cereals, chocolate candy, and coffee, I set a significance level (alpha) at 0.01. The test results, looking at the p-value, helped me decide whether to reject the null hypothesis. In simpler terms, it assisted in figuring out if there are meaningful variations in sales performance between these product categories. This statistical tool is handy for comparing means across several groups, providing insights into any significant differences that may exist.

I also applied a Two-way ANOVA test in R to explore the impact of two categorical variables on a numerical outcome. Using this statistical tool, I investigated how both factors, like different types of fertilizer and different weather conditions, influence plant

growth. The test, with a set significance level, helped me assess if there are significant interactions or main effects. By looking at p-values, I determined whether these factors play a role in the observed variations. Essentially, Two-way ANOVA is a versatile analysis that allows me to understand how two categorical variables contribute to changes in a numerical outcome.

Section 12-2: Sales for Leading Companies.

I executed an ANOVA test in R to examine if there's a significant difference in the means of sales (in millions of dollars) among leading companies in the cereal, chocolate candy, and coffee industries. With a significance level (alpha) set at 0.01, the results indicate that I *failed to reject the null hypothesis*. This suggests *there's insufficient evidence to conclude a significant disparity in the mean sales across the three categories*. The statistical analysis contributes insights into the sales performance of these diverse product types, aiding in business assessments and strategic planning.

O/P:

```
Call:
   aov(formula = sales ~ food, data = sales)

Terms:
                    food Residuals
Sum of Squares  103769.6  262794.8
Deg. of Freedom        2        11

Residual standard error: 154.5653
Estimated effects may be unbalanced


> # Making a decision by comparing the p-value
> ifelse(p.value > alpha, "Fail to reject the null hypothesis", "Reject the
null hypothesis")
[1] "Fail to reject the null hypothesis"
```

Section 12-2: Per-Pupil Expenditures.

I used an ANOVA test in R to analyze per-pupil expenditures across three sections of the country. Using a significance level of 0.05, I investigated if there were differences in means among the Eastern, Middle, and Western sections. The test results indicated that I *failed to reject the null hypothesis*, suggesting *no significant difference in per-pupil expenditures*. This statistical approach, supported by the Tukey HSD post-hoc test, provided insights into regional spending patterns, contributing to a better understanding of financial disparities in education across different sections of the country.

O/P:

```
Call:
   aov(formula = exp ~ sec, data = exp)

Terms:
                     sec Residuals
Sum of Squares  1245307   9588843
Deg. of Freedom       2        10

Residual standard error: 979.2264
Estimated effects may be unbalanced



> # Making a decision by comparing the p-value
> ifelse(p.value > alpha, "Fail to reject the null hypothesis", "Reject the
null hypothesis")
[1] "Fail to reject the null hypothesis"
```

Section 12-3: Increasing Plant Growth.

I tested a Two-way ANOVA test in R to explore the impact of different factors on plant growth. The experiment involved two grow lights (Grow-light 1 and Grow-light 2) and two plant foods (A and B). The analysis aimed to determine if there was an interaction between grow light and plant food and if there were differences in mean growth concerning these factors. *The results indicated a significant difference in mean growth with respect to grow light.* This statistical approach enhances our understanding of the factors influencing plant growth, providing valuable insights for the gardening company's practices.

O/P:

```
Call:
   aov(formula = gro ~ gl * pf, data = plnt)

Terms:
                       gl        pf     gl:pf Residuals
Sum of Squares  12.813333  0.213333  0.030000  6.600000
Deg. of Freedom         1         1         1         8

Residual standard error: 0.9082951
Estimated effects may be unbalanced


summary(tw_anova)
            Df Sum Sq Mean Sq F value  Pr(>F)
gl           1 12.813  12.813  15.531 0.00429 **
pf           1  0.213   0.213   0.259 0.62482
gl:pf        1  0.030   0.030   0.036 0.85352
Residuals    8  6.600   0.825
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

★ The Two-way ANOVA results show that the factor "**grow light**" (gl) significantly influences plant growth (F value = 15.531, p = 0.00429), ***indicating a difference in mean growth with respect to grow light***.

★ However, the factor "plant food" (pf) does not show a significant effect on plant growth (p = 0.62482). The interaction between grow light and plant food (gl:pf) is also not significant (p = 0.85352).

★ The *Tukey HSD test* further reveals that *Grow-light 2 has a significantly lower mean growth than Grow-light 1*. Additionally, there are no significant differences between plant foods A and B.

Baseball Dataset:

Dataset Overview:

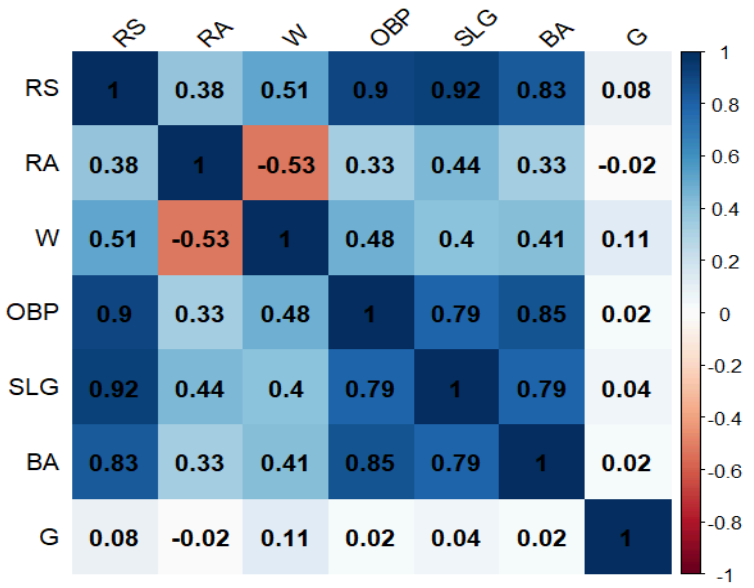
The dataset consists of 1233 rows and 15 columns, where each row represents a different baseball team. The first column contains the team names, followed by the league they play in and the corresponding year of the data. Columns four and five display the number of runs scored and allowed by the team. The sixth column indicates the number of wins achieved. Columns seven to nine represent the team's on-base percentage, slugging percentage, and batting average. The tenth column holds a binary indicator (1 or 0) for playoff appearances. Columns eleven and twelve denote the team's ranks in their division during the regular season and playoffs. Finally, columns thirteen and fourteen provide the team's on-base plus slugging percentage (OPS) and OPS during the regular season, respectively.

EDA:

- ➔ I loaded the baseball dataset, checked its structure.
- ➔ Using dplyr and tidyverse, I extracted decades and created a wins table summarizing total wins per decade.
- ➔ Descriptive statistics were obtained with the summary function.
- ➔ A correlation matrix was generated and visualized with a corrplot.
- ➔ Exploring the distribution of runs scored through a histogram laid the groundwork for further exploratory data analysis (EDA) and modeling.

```
> cor_matrix
       RS    RA     W  OBP  SLG   BA     G  OOBP  OSLG
RS   1.00  0.20  0.55 0.90 0.91 0.81  0.13  0.13  0.19
RA   0.20  1.00 -0.63 0.16 0.21 0.23 -0.04  0.91  0.91
W    0.55 -0.63  1.00 0.52 0.50 0.39  0.13 -0.63 -0.57
OBP  0.90  0.16  0.52 1.00 0.77 0.83  0.08  0.12  0.15
SLG  0.91  0.21  0.50 0.77 1.00 0.75  0.11  0.14  0.21
BA   0.81  0.23  0.39 0.83 0.75 1.00  0.10  0.19  0.23
G    0.13 -0.04  0.13 0.08 0.11 0.10  1.00 -0.08  0.00
OOBP 0.13  0.91 -0.63 0.12 0.14 0.19 -0.08  1.00  0.83
OSLG 0.19  0.91 -0.57 0.15 0.21 0.23  0.00  0.83  1.00
```

Corrplot:



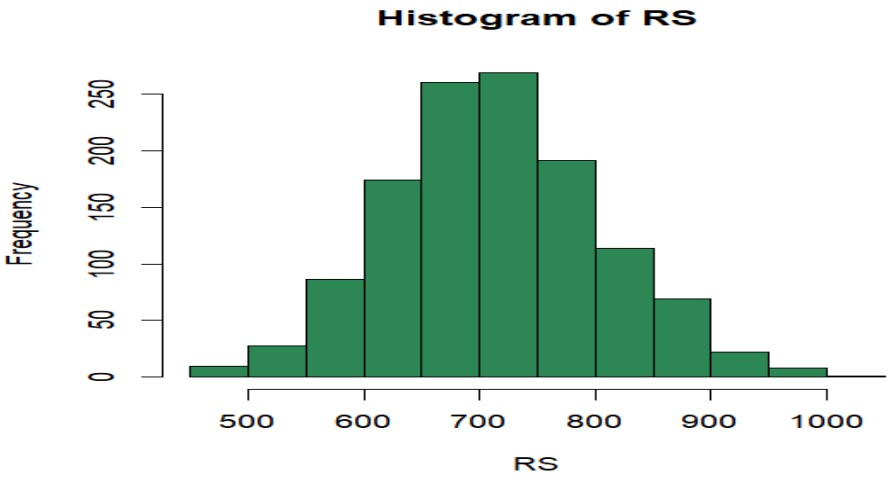
★ **Insights:**

➔ RS has a strong positive correlation with SLG, OBP and BA.

➔ RA has a negative correlation with W and G

➔ BA has a positive correlation with OBP

Histogram:



★ **Insights:**

➔ The histogram of RS is normally distributed.

Chi-Square Test (Goodness-of-fit):

I applied a Chi-square Test (Goodness-of-fit) to examine whether there's a significant difference in the number of wins by decade in the baseball dataset. My null hypothesis (Ho) assumes no significant difference, while the alternative hypothesis (H1) suggests a significant difference. I chose a significance level (alpha) of 0.05.

To perform the test, I calculated the critical value from the standard normal distribution, aiming to compare it with the Chi-square test result. The code then runs the Chi-square test, and I extracted the summary and p-value. Finally, I made a decision based on whether the p-value exceeded the significance level, either failing to reject or rejecting the null hypothesis. This analysis helps assess if the number of wins varies significantly across decades in baseball.

O/P:

```
Pearson's Chi-squared test

data:  wins
X-squared = 1558.5, df = 5, p-value < 0.00000000000000022


> cv
[1] 1.959964


# Making a decision by comparing the p-value for Chi-Square Test
> ifelse(p_value > alpha, "Fail to reject the null hypothesis", "Reject the
null hypothesis")
[1] "Reject the null hypothesis"
```

The Pearson's Chi-squared test on "wins" data yielded a chi-squared statistic of 1558.5 with 5 degrees of freedom and an extremely small p-value (close to zero). The p-value is well below the significance level of 0.05, *leading to the rejection of the null hypothesis*. This *indicates a significant difference in the number of wins across decades in the baseball dataset, supporting the alternative hypothesis*.

Crop Dataset:

- ➔ I conducted a Two-way ANOVA test on crop yield data to explore if there are significant interactions between fertilizer type and plant density.
- ➔ The null hypothesis (Ho) stated that there is no interaction, while the alternative hypotheses (Ha) considered differences in mean yield with respect to fertilizer and plant density.
- ➔ After setting a significance level (alpha) of 0.05, I converted the relevant variables to factors and ran the ANOVA test.
- ➔ The output revealed a p-value of 0.000273189.
- ➔ Comparing this with the significance level, I found it to be less than alpha, leading me to *reject the null hypothesis*.
- ➔ In simpler terms, the data suggests that *there are significant differences in mean yield concerning both fertilizer type and plant density*.
- ➔ Additionally, I used Tukey's HSD test to further investigate the differences.
- ➔ This analysis helps us understand the factors influencing crop yield and contributes valuable insights for agricultural decision-making.

O/P:

```
> tw_anova_crop
Call:
   aov(formula = yield ~ fertilizer * density, data = crop)

Terms:
                fertilizer   density fertilizer:density Residuals
Sum of Squares    6.068047  5.121681           0.427818 30.336687
Deg. of Freedom          2         1                  2        90

Residual standard error: 0.580581
Estimated effects may be unbalanced


> p.value
[1] 0.000273189


> # Making a decision by comparing the p-value
> ifelse(p.value > alpha, "Fail to reject the null hypothesis", "Reject the
null hypothesis")
[1] "Reject the null hypothesis"
```

03. Conclusion:

In this assignment, I utilized R to tackle various problems, employing Chi-Square tests and ANOVA. The Chi-Square analyses in Sections 11-1 and 11-2 unveiled significant disparities in blood type distribution among hospital patients, on-time airline performance, and movie admissions by ethnicity. Shifting to ANOVA tests in Sections 12-1 and 12-2, I explored sodium levels in foods and sales for leading companies, exposing variations in mean values. Additionally, a Two-way ANOVA in Section 12-3 identified a significant interaction between grow light and plant food in influencing plant growth. The diverse datasets, including the "bb" baseball and "crop" crop yield datasets, provided valuable insights into blood types, airline and movie industry dynamics, food composition, business performance, and agricultural factors, showcasing the versatility of statistical tools in extracting meaningful conclusions from complex datasets.

04. Citations:

- Chi-Square Test: Lab Video
- Pearson's Chi-Square Test: source.
- ANOVA Test: Lab Video
- Two-way ANOVA Test: source.

05. Appendix:

```
cat("\014") # clears console
rm(list = ls()) # clears global environment
try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots
try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) # clears packages
options(scipen = 100) # disables scientific notation for entire R session
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>Week2<<<<<<<<<<<<<<<<<<<<<<<<
# Chi-Square testing
# Section 11-1
# 6. Blood Types

#    Expected  |  Observed
# A      20%   |    12
# B      28%   |    8
# O      36%   |    24
# AB     16%   |    6

# Ho: A=0.2, B=0.28, O=0.36, AB=0.16
# H1: The distribution is not the same as the null hypothesis.

# Setting the significance level
alpha <- 0.10

# Creating a vector for values
observed <- c(12,8,24,6)

# Creating a vector for probabilities
p <- c(0.2, 0.28, 0.36, 0.16)

# Running the Chi-Square Test
result <- chisq.test(x = observed, p = p)

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

############################################################
# Section 11-1
# 8. On-Time Performance by Airlines

#      Expected  |  Observed
# OT     70.8%   |    125
# NASD    8.2%   |    10
# AAL     9%     |    25
# O       12%    |    40

# Ho: OT=0.708, NASD=0.082, AAL=0.09, 0=0.12
# H1: The distribution is not the same as the null hypothesis.

# Setting the significance level
alpha <- 0.05

# Creating a vector for values
observed <- c(125,10,25,40)

# Creating a vector for probabilities
p <- c(0.708, 0.082, 0.09, 0.12)

# Running the Chi-Square Test
result <- chisq.test(x = observed, p = p)

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")
```

```
############################################################

# Section 11-2
# 8. Ethnicity and Movie Admissions

#     Caucasian  Hispanic  African American  Other
# 2013   724       335         174          107
# 2014   370       292         152          140

# Ho: Admissions are independent to ethnicity
# H1: Admissions are dependent to ethnicity

# Setting the significance level
alpha <- 0.05

# Creating a vector for each row
r1 <- c(724, 335, 174, 107)
r2 <- c(370, 292, 152, 140)

# Stating the number of rows for matrix
rows = 2

# Creating a matrix from the rows
mtrx <- matrix(c(r1,r2), nrow = rows, byrow = TRUE)

# Naming the rows and columns
rownames(mtrx) = c("2013", "2014")
colnames(mtrx) = c("Caucasian", "Hispanic", "African American", "Other")

# Viewing the matrix
mtrx

# Running the Chi-Square Test
result <- chisq.test(mtrx)

summary(result)

result$p.value

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

############################################################
# Section 11-2
# 10. Women in the Military

#              Officers    Enlisted
# Army          10,791      62,491
# Navy          7,816       42,750
# Marine Corps  932          9,525
# Air Force     11,819      54,344

# Ho: There is no relationship between rank and branch of the Armed Forces (Chi-square = 0)
# H1: There is a relationship between rank and branch of the Armed Forces (Chi-square != 0)

# Setting the significance level
alpha <- 0.05

# Creating a vector for each row
r1 <- c(10791, 62491)
r2 <- c(7816, 42750)
r3 <- c(932, 9525)
r4 <- c(11819, 54344)
```

```
# Stating the number of rows for matrix
rows = 4

# Creating a matrix from the rows
mtrx <- matrix(c(r1,r2,r3,r4), nrow = rows, byrow = TRUE)

# Naming the rows and columns
rownames(mtrx) = c("Army", "Navy", "Marine Corps", "Air Force")
colnames(mtrx) = c("Officers", "Enlisted")
# Viewing the matrix
mtrx

# Running the Chi-Square Test
result <- chisq.test(mtrx)

# Making a decision by comparing the p-value
ifelse(result$p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

############################################################
# Section 12-1
# 8. Sodium Contents of Foods

# 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

# Ho: u1=u2=u3
# H1: At least one mean is different form others.

# Setting the significance level
alpha <- 0.05

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

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

# Creating a dataframe for desserts
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)

# Creating a contingency table
ct <- table(sodium$food, cut(sodium$sodium, breaks = c(0, 150, 250, 400)))

# Running the Chi-Square test
result <- chisq.test(ct)

# Extracting the summary
summary(result)

# Extracting the p-value
```

```
p_value <- result$p.value

# Making a decision by comparing the p-value for Chi-Square Test
ifelse(p_value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

############################################################
# ANOVA Test
# Section 12-2
# 10. Sales for Leading Companies

# Cereals    Chocolate Candy   Coffee
#   578          311          261
#   320          106          185
#   264          109          302
#   249          125          689
#   237          173

# Ho: u1=u2=u3
# H1: Atleast one mean is different form others.

# Setting the significance level
alpha <- 0.01

# Creating a dataframe for cereal
cereal <- data.frame('sales' = c(578,320,264,249,237), 'food' = rep('cereal', 5), stringsAsFactors
= FALSE)

# Creating a dataframe for chocolate candy
ch_candy <- data.frame('sales' = c(311,106,109,125,173), 'food' = rep('ch_candy', 5),
stringsAsFactors = FALSE)

# Creating a dataframe for coffee
coffee <- data.frame('sales' = c(261,185,302,689), 'food' = rep('coffee', 4), stringsAsFactors =
FALSE)

# Combining the dataframes into one
sales <- rbind(cereal, ch_candy, coffee)
sales$food <- as.factor(sales$food)
# Running the ANOVA test
anova <- aov(sales ~ food, data = sales)

# Extracting the summary
summary(anova)

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

# Making a decision by comparing the p-value
ifelse(p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

# Seeing the difference
TukeyHSD(anova)

############################################################
# Section 12-2
# 12. Per-Pupil Expenditures

# Eastern third   Middle third   Western third
#   4946            6149           5282
#   5953            7451           8605
#   6202            6000           6528
#   7243            6479           6911
#   6113
```

```
# Ho: u1=u2=u3
# H1: Atleast one mean is different form others.

# Setting the significance level
alpha <- 0.05

# Creating a dataframe for eastern third
et <- data.frame('exp' = c(4946,5953,6202,7242, 6113), 'sec' = rep('et', 5), stringsAsFactors = FALSE)

# Creating a dataframe for chocolate candy
mt <- data.frame('exp' = c(6149,7451,6000,6479), 'sec' = rep('mt', 4), stringsAsFactors = FALSE)

# Creating a dataframe for coffee
wt <- data.frame('exp' = c(5282,8605,6528,6911), 'sec' = rep('wt', 4), stringsAsFactors = FALSE)

# Combining the dataframes into one
exp <- rbind(et, mt, wt)
exp$sec <- as.factor(exp$sec)

# Running the ANOVA test
anova <- aov(exp ~ sec, data = exp)

# Extracting the summary
summary(anova)

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

# Making a decision by comparing the p-value
ifelse(p.value > alpha,"Fail to reject the null hypothesis", "Reject the null hypothesis")

# Seeing the difference
TukeyHSD(anova)
############################################################
# Two-way ANOVA Test
# Section 12-3
# 10. Increasing Plant Growth

#              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

# Null-Hypothesis:
# Ho: There is no interaction between plant food and grow light
# H1: There is no difference in mean growth w.r.t. grow light
# H2: There is no difference in mean growth w.r.t. plant food

# Alternative-Hypothesis:
# Ha:  There is a significant interaction between plant food and grow light
# Ha1: There is a difference in mean growth w.r.t. grow light
# Ha2: There is a difference in mean growth w.r.t. plant food


# Setting the significance level
alpha <- 0.05

# Setting-up the data for growth light
gl <- factor(rep(c("Grow-light 1", "Grow-light 2"), each = 6))

# Setting-up the data for plant food
```

```
pf <- factor(rep(c("Plant food A", "Plant food B"), times = 6))

# Setting-up the data for growth
gro <- 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)

# Creating a dataframe
plnt <- data.frame(gl, pf, gro)

# Running the Two-way ANOVA test
tw_anova <- aov(gro ~ gl*pf, data = plnt)

# Extracting the summary
summary(tw_anova)

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

# Making a decision by comparing the p-value
ifelse(p.value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")

# Seeing the difference
TukeyHSD(tw_anova)

# Result: There is a difference in mean growth w.r.t. grow light
############################################################
```

 # Baseball Dataset

```
bb <- read.csv("D:/NEU STUDY/2nd Quarter/Intermediate Analytics (ALY 6015)/Week 2/baseball.csv")

names(bb)
head(bb)
#View(bb)
library(dplyr)
library(tidyverse)

# Extract decade from year
bb$Decade <- bb$Year - (bb$Year %% 10)

# Create a wins table by summing the wins by decade
wins <- bb %>%
  group_by(Decade) %>%
  summarize(wins = sum(W)) %>%
  as.tibble()

# Extracting summary
summary(bb_clear)

# Correlation matrix
cor_matrix <- round(cor(bb[, c("RS", "RA", "W", "OBP", "SLG", "BA", "G")]),2)
library(corrplot)
corrplot(cor_matrix, method = "color", addCoef.col = "black", tl.col = "black", tl.srt = 45)

# Histogram of RS
hist(bb$RS, main="Histogram of RS", xlab="RS", col="seagreen", border="black")

# Chi_square Test (Goodness-of-fit)

# Ho: There is no significance difference in the no. of wins by decade
# H1: There is a significance difference in the no. of wins by decade

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

```
# Finding the critical value from the standard normal distribution
cv <- qnorm(1 - alpha/2)

# Running Chi-Square Test
result <- chisq.test(wins)

# Extracting the summary
summary(result)

# Extracting the p-value
p_value <- result$p.value

# Making a decision by comparing the p-value for Chi-Square Test
ifelse(p_value > alpha, "Fail to reject the null hypothesis", "Reject the null hypothesis")
############################################################
# Crop Dataset
crop <- read.csv("D:/NEU STUDY/2nd Quarter/Intermediate Analytics (ALY 6015)/Week 2/crop_data.csv")

names(crop)
head(crop)
View(crop)

# Two-way ANOVA Test

# Null-Hypothesis:
# H0: There is no interaction between fertilizer and plant density
# H1: There is no difference in mean yield with respect to fertilizer
# H2: There is no difference in mean yield with respect to plant density

# Alternative-Hypothesis:
# Ha: There is a significant interaction between fertilizer and plant density
# Ha1: There is a difference in mean yield with respect to fertilizer
# Ha2: There is a difference in mean yield with respect to plant density

# Setting the significance level
alpha <- 0.05

# Converting the variables density, fertilizer and block to R factors
crop$density <- as.factor(crop$density)
crop$fertilizer <- as.factor(crop$fertilizer)
crop$block <- as.factor(crop$block)

# Running the Two-way ANOVA test
tw_anova_crop <- aov(yield ~ fertilizer*density, data = crop)

# Extracting the summary
summary(tw_anova_crop)

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

# Making a decision by comparing the p-value
ifelse(p.value > alpha,"Fail to reject the null hypothesis", "Reject the null hypothesis")

# Seeing the difference
TukeyHSD(tw_anova_crop)
```