MODULE ONE PROJECT

CHI-SQUARE AND ANOVA ASSESSMENT

Submitted By: **HARSHIT GAUR**
MASTER OF PROFESSIONAL STUDIES IN ANALYTICS
ALY 6015 : INTERMEDIATE ANALYTICS
CRN : 21454
MARCH 5, 2022
WINTER 2022

Submitted To: **PROF. ROY WADA**

INTRODUCTION

A **chi-square (χ2) statistic** is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a fair coin meet these criteria.

Chi-square tests are often used in hypothesis testing. The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = Degrees of freedom

O = Observed value(s)

E = Expected value(s)

Figure 1.1: Formula to calculate Chi-Square Statistic

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

There are two main types of ANOVA: one-way (or unidirectional) and two-way. One-way or two-way refers to the number of independent variables in your analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents.

$$\mathrm{F} = \frac{\mathrm{MST}}{\mathrm{MSE}}$$

where:

F = ANOVA coefficient

MST = Mean sum of squares due to treatment

MSE = Mean sum of squares due to error

Figure 1.2: Formula to calculate ANOVA Statistic

ANALYSIS

With the medium of this project, we will understand the concepts of Chi-Square test and ANOVA test by performing these tests on various data sets which are either formed using data in the questions or are available to us.

Before moving forward to conduct/perform these tests, let's look at the general steps to follow for these tests.

1. *State the hypotheses and identify the claim.*
2. *Find the critical value.*
3. *Compute the test value.*
4. *Make the decision.*
5. *Summarize the conclusion/results.*

PROBLEM 1 - BLOOD TYPES

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.

Blood Types Table

	Expected	Observed
Type-A	10	12
Type-B	14	8
Type-O	18	24
Type-AB	8	6

Table 1.1: Blood Types Table

At $\alpha = 0.10$, can it be concluded that the distribution is the same as that of the general population?

Alpha value :
0.10

Percentages of the Blood-Type in general population, from which "Expected" values have been calculated, are :
Type A - 20%; Type B - 28%; Type O - 36%; and Type AB - 16%

Null Hypothesis -
H_0 ***: Type-A = 0.20, Type-B = 0.28, Type-O = 0.36, Type-AB = 0.16***

Alternate Hypothesis -

H_1 : The blood type distribution is not the same in the hospital's patient population as stated in the null hypothesis

Chi-Square Test Value :

5.47142857142857

Chi-Square P-Value :

0.140357527293565

Degree of Freedom :

3

Critical Value :

6.2514

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qchisq()** function.

Result :

Failed to reject the null hypothesis as the P-value of 0.1404 is greater than the alpha value of 0.1. The results are not significant. Therefore, we do not have sufficient evidences to claim that the blood type distribution is not the same in the hospital's patient population as stated in the null hypothesis

PROBLEM 2 - ON-TIME PERFORMANCE BY AIRLINES

According to the Bureau of Transportation Statistics, on- time performance by the airlines is described and we need to check if the results from our calculated statistics differ from the government's statistics.

On-Time Performance by Airlines Table

	Expected	Observed
On-Time	141	125
National Aviation System Delay	16	10
Aircraft Arriving Late	18	25
Other (because of weather and other conditions)	24	40

Table 1.1: On-Time Performances by Airlines Table

Alpha value :

0.05

Percentages of the on-time performances by airlines, from which "Expected" values have been calculated, are :
On Time - 70.8%; National Aviation System Delay- 8.2%;
Aircraft Arriving Late - 9%; and Other (because of weather and conditions - 12%

Null Hypothesis -
H_0 : 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 -
H_1 : The on-time performance distribution of airlines is not the same as stated in the null hypothesis

Chi-Square Test Value :
17.8324950622388

Chi-Square P-Value :
0.00047625874475707

Degree of Freedom :
3

Critical Value :
7.8147

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qchisq()** function.

Result :
Reject the null hypothesis as the P-value of 0.0005 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that the on-time performance distribution of airlines is not the same as stated in the null hypothesis

PROBLEM 3 - ETHNICITY AND MOVIE ADMISSIONS

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?

Ethnicity and Movie Admissions Table

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

Table 1.3: Ethnicity and Movie Admissions

Alpha value :
0.05

Null Hypothesis -
H_0 : Movie admissions are independent of ethnicity

Alternate Hypothesis -
H_1 : Movie admissions are dependent on ethnicity

Chi-Square Test Value :
60.1435247416858

Chi-Square P-Value :
0.0000000000005477507

Degree of Freedom :
3

Critical Value :
7.8147

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qchisq()** function.

Result :
Reject the null hypothesis as the P-value of 0.000000000000548 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that the movie admissions are dependent on ethnicity.

PROBLEM 4 - WOMEN IN THE MILITARY

The table lists the numbers of officers and enlisted personnel for women in the military. At $\alpha = 0.05$, is there sufficient evidence to conclude that a relationship exists between rank and branch of the Armed Forces?

Women in the Military Table

	Officers	Enlisted
Army	10791	62491
Navy	7816	42750
Marine Corps	932	9525
Air Corps	11819	54344

Table 1.4: Women in the Military

Alpha value :
0.05

Null Hypothesis -
H_0 : Ranks of women in Armed Forces are independent of their branches

Alternate Hypothesis -
H_1 : Ranks of women in Armed Forces are dependent on their branches

Chi-Square Test Value :
654.271888875628

Chi-Square P-Value :
$1.72641801107315e^{-141}$

Degree of Freedom :
3

Critical Value :
7.8147

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qchisq()** function.

Result :
Reject the null hypothesis as the P-value of almost 0 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that the *Ranks of women in Armed Forces are dependent on their branches.*

ANOVA Tests (One-Way and Two-Way ANOVA Tests)

Now, we will conduct/perform One-way or Two-way ANOVA tests on the questions asked in the assignment. We will formulize the Null and Alternate Hypotheses and find out whether the results are significant or not to claim the alternate hypothesis.

PROBLEM 5 - SODIUM CONTENTS OF FOODS

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?

Sodium Contents of Food Table

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

Table 1.5: Sodium content of Foods

Alpha value :
0.05

Null Hypothesis -
H_0 : μ-Condiments = μ-Cereals = μ-Desserts

Alternate Hypothesis -
H_1 : At least one mean is different from the others in the null hypothesis.

Degree of Freedom -

k - 1: Between Group Variance - Numerator : **2**
N - k: Within Group Variance - Denominator : **19**

ANOVA test F-Value :
2.398538

ANOVA test P-Value :
0.1178108

Critical Value :
3.5219

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qf()** function.

Result :

Failed to reject the null hypothesis as the P-value of 0.1178108 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that *At least one mean is different from the others in the null hypothesis.*

PROBLEM 6 - SALES FOR LEADING COMPANIES

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?

Sales for Leading Companies Table

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

Table 1.6: Sales for Leading Companies

Alpha value :

0.01

Null Hypothesis -

H_0 : μ-Cereals = μ-`Chocolate Candy` = μ-Coffee

Alternate Hypothesis -

H_1 : At least one mean is different from the others in the null hypothesis.

Degree of Freedom -

k - 1: Between Group Variance - Numerator : **2**

N - k: Within Group Variance - Denominator : **11**

ANOVA test F-Value :

2.171782

ANOVA test P-Value :

0.1603487

Critical Value :

7.2057

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qf()** function.

Result :
Failed to reject the null hypothesis as the P-value of 0.1603487 is greater than the alpha value of 0.01. The results are not significant. Therefore, we do not have sufficient evidences to claim that *At least one mean is different from the others in the null hypothesis.*

PROBLEM 7 - PER-PUPIL EXPENDITURES

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?

Per-Pupil Expenditures Table

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

Table 1.7: Per Pupil Expenditure in 3 sections of a country

Alpha value :
0.05

Null Hypothesis -
H_0 : μ-Eastern Third = μ-Middle Third = μ-Western Third

Alternate Hypothesis -
H_1 : At least one mean is different from the others in the null hypothesis.

Degree of Freedom -
k - 1: Between Group Variance - Numerator : **2**
N - k: Within Group Variance - Denominator : **10**

ANOVA test F-Value :
0.6488214

ANOVA test P-Value :
0.5433264

Critical Value :
4.1028

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qf()** function.

Result :
Failed to reject the null hypothesis as the P-value of 0.5433264 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that *At least one mean is different from the others in the null hypothesis.*

PROBLEM 8 - INCREASING PLANT GROWTH

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. 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 $\alpha = 0.05$

Increasing Plant Growth Table

GROWTH	GROW LIGHT	PLANT FOOD
9.2	1	A
9.4	1	A
8.9	1	A
8.5	2	A
9.2	2	A
8.9	2	A
7.1	1	B
7.2	1	B
8.5	1	B
5.5	2	B
5.8	2	B
7.6	2	B

Table 1.8: Increasing Plant Growth

Alpha value :
0.05

Since, this test is a Two-Way ANOVA test and because we need null and alternative hypotheses for the effect on both categorical factors, and the effect of the categorical factors on each other, the null and alternative hypothesis pairs may be expressed as follows.

(3 Pairs in 2-Way ANOVA)
1st Pair:

Null Hypothesis -
H_0 : The means of all Plant-Food Supplement groups are same

Alternate Hypothesis -
H_1 : The means of all Plant-Food Supplement groups are different

2nd Pair:

Null Hypothesis -
H_0 : The means of all Growth-Light groups are same

Alternate Hypothesis -
H_1 : The means of all Growth-Light groups are different

3rd Pair:

Null Hypothesis -
H_0 : There is no interaction between the Growth-Light and Plant-Food Supplement

Alternate Hypothesis -
H_1 : There is interaction between the Growth-Light and Plant-Food Supplement

```
# Run the ANOVA test
anova <- aov(growth ~ growth_light + plant_food + growth_light:plant_food, data = plantsGrowth)
```

Figure 1.3: Two-Way ANOVA Test script code.

Increasing Plant Growth ANOVA Test Summary Table

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
growth_light	1	1.920000	1.9200000	3.680511	0.091331368
plant_food	1	12.813333	12.8133333	24.562300	0.001112418
growth_light : plant_food	1	0.750000	0.7500000	1.437700	0.264819413
Residuals	8	4.173333	0.5216667		

Table 1.9: Increasing Plant Growth ANOVA Test Summary

Degree of Freedom -

k - 1: Between Group Variance - Numerator (Growth Light) : **1**
k - 1: Between Group Variance - Numerator (Plant Food) : **1**
k - 1: Between Group Variance - Numerator (Growth Light : Plant Food) : **1**
N - k: Within Group Variance - Denominator : **8**

ANOVA test **F-Value** :

Growth Light	Plant Food	Growth Light : Plant Food
3.680511	**24.5623**	**1.4377**

ANOVA test **P-Value** :

Growth Light	Plant Food	Growth Light : Plant Food
0.09133137	**0.001112418**	**0.2648194**

Critical Value :

Growth Light	Plant Food	Growth Light : Plant Food
5.3177	**5.3177**	**5.3177**

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qf()** function.

Results :

Failed to reject the null hypothesis that means of all Growth-Light groups are same as the P-value of 0.09133137 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that *the means of all Growth-Light groups are different.*

Reject the null hypothesis that means of all Plant-Food Supplement groups are same as the P-value of 0.001112418 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that *the means of all Plant-Food Supplement groups are different.*

Failed to reject the null hypothesis that there is no interaction between the Growth-Light and Plant-Food Supplement as the P-value of 0.2648194 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that *there is interaction between the Growth-Light and Plant-Food Supplement.*

PROBLEM 9 - BASEBALL

Perform a Chi-Square Goodness-of-Fit test to determine if there is a difference in the number of wins by decade.

The first 6 data points of the baseball data set has been presented below for a glimpse of the data set and what all variables are included in the data set.

Baseball Dataset

Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
ARI	NL	2,012	734	688	81	0.328	0.418	0.259	0			162	0.317	0.415
ATL	NL	2,012	700	600	94	0.320	0.389	0.247	1	4	5	162	0.306	0.378
BAL	AL	2,012	712	705	93	0.311	0.417	0.247	1	5	4	162	0.315	0.403
BOS	AL	2,012	734	806	69	0.315	0.415	0.260	0			162	0.331	0.428
CHC	NL	2,012	613	759	61	0.302	0.378	0.240	0			162	0.335	0.424
CHW	AL	2,012	748	676	85	0.318	0.422	0.255	0			162	0.319	0.405

Table 1.10: Baseball Dataset

Exploratory Data Analysis -

We will analyse the data set to find out some insights and investigate the variables.

Descriptive Statistics of Baseball Dataset

e	n	mean	sd	median	min	max	range	skew	kurtosis
Team*	1,232	18.93	10.61	20.00	1.00	39.00	38.00	0.06	-1.25
League*	1,232	1.50	0.50	1.50	1.00	2.00	1.00	0.00	-2.00
Year	1,232	1,988.96	14.82	1,989.00	1,962.00	2,012.00	50.00	-0.15	-1.21
RS	1,232	715.08	91.53	711.00	463.00	1,009.00	546.00	0.17	-0.03
RA	1,232	715.08	93.08	709.00	472.00	1,103.00	631.00	0.30	-0.02
W	1,232	80.90	11.46	81.00	40.00	116.00	76.00	-0.18	-0.31
OBP	1,232	0.33	0.02	0.33	0.28	0.37	0.10	0.02	0.06
SLG	1,232	0.40	0.03	0.40	0.30	0.49	0.19	0.05	-0.33
BA	1,232	0.26	0.01	0.26	0.21	0.29	0.08	-0.11	0.00
Playoffs	1,232	0.20	0.40	0.00	0.00	1.00	1.00	1.51	0.29
RankSeason	244	3.12	1.74	3.00	1.00	8.00	7.00	0.56	-0.58
RankPlayoffs	244	2.72	1.10	3.00	1.00	5.00	4.00	-0.27	-1.12
G	1,232	161.92	0.62	162.00	158.00	165.00	7.00	-1.04	6.97
OOBP	420	0.33	0.02	0.33	0.29	0.38	0.09	0.19	-0.37
OSLG	420	0.42	0.03	0.42	0.35	0.50	0.15	0.12	-0.21

Table 1.11: Baseball Dataset Descriptive Statistics Summary

1. The data set contains 1,232 data points with 16 features.
2. The mean of WINS (W) 80.90 which is almost close to the median of 81. This shows that the distribution of wins might be a normal distribution. We can check the normality using Q-Q Plot and Shapiro Wilks test as well. The distribution of wins is a normal distribution from the tests results.

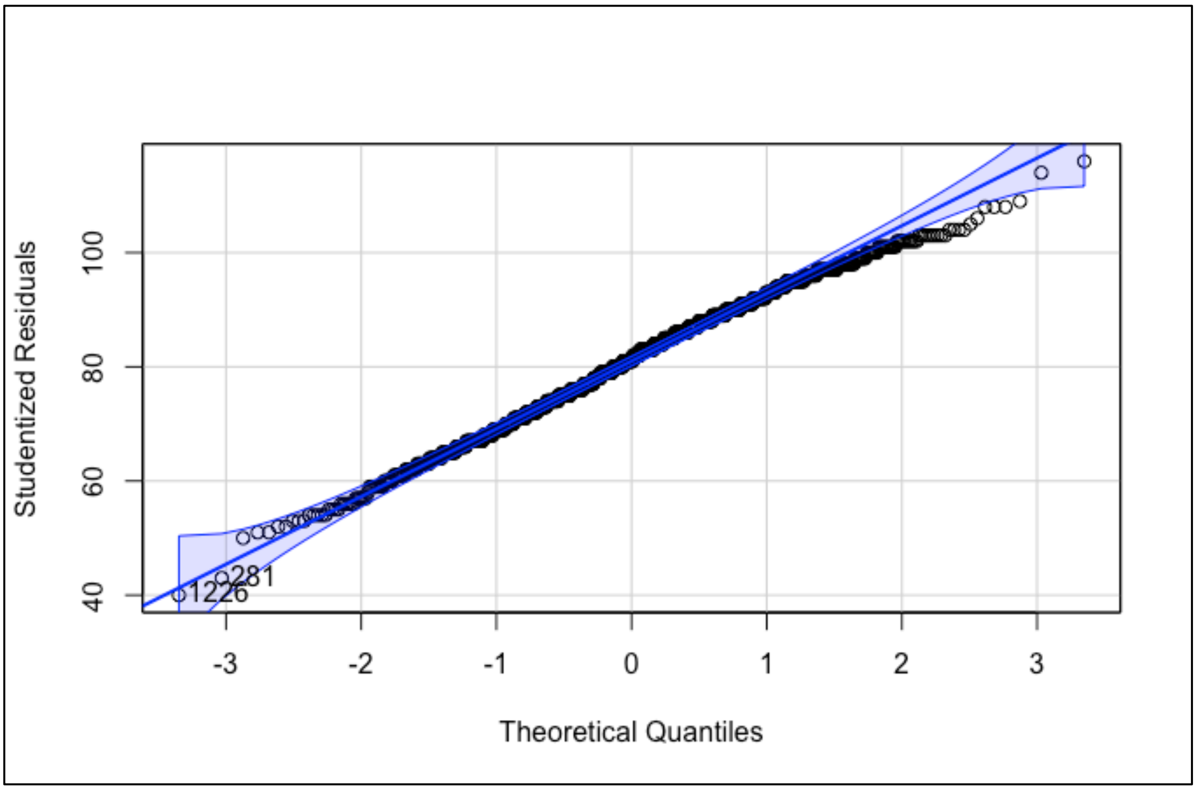

Figure 1.4: Quantile-Quantile Plot for Normality

```
        Shapiro-Wilk normality test

data:  baseball$W
W = 0.99442, p-value = 0.0001479
```

Figure 1.5: Shapiro Wilks Test for Normality

3. The minimum wins registered by a team in a season is 40 and the maximum is 116.
4. The standard deviation of the wins distribution is 11.46 which means that the data is dispersed.
5. The wins distribution is slightly negatively skewed as suggested by the value of -0.18 of skewness.

6. The RUNS SUPPORT (RS) and RUNS AGAINST (RA) variables also have their means almost equal to their respective medians. This also suggests that there is a possibility of these distributions to be normal in nature.

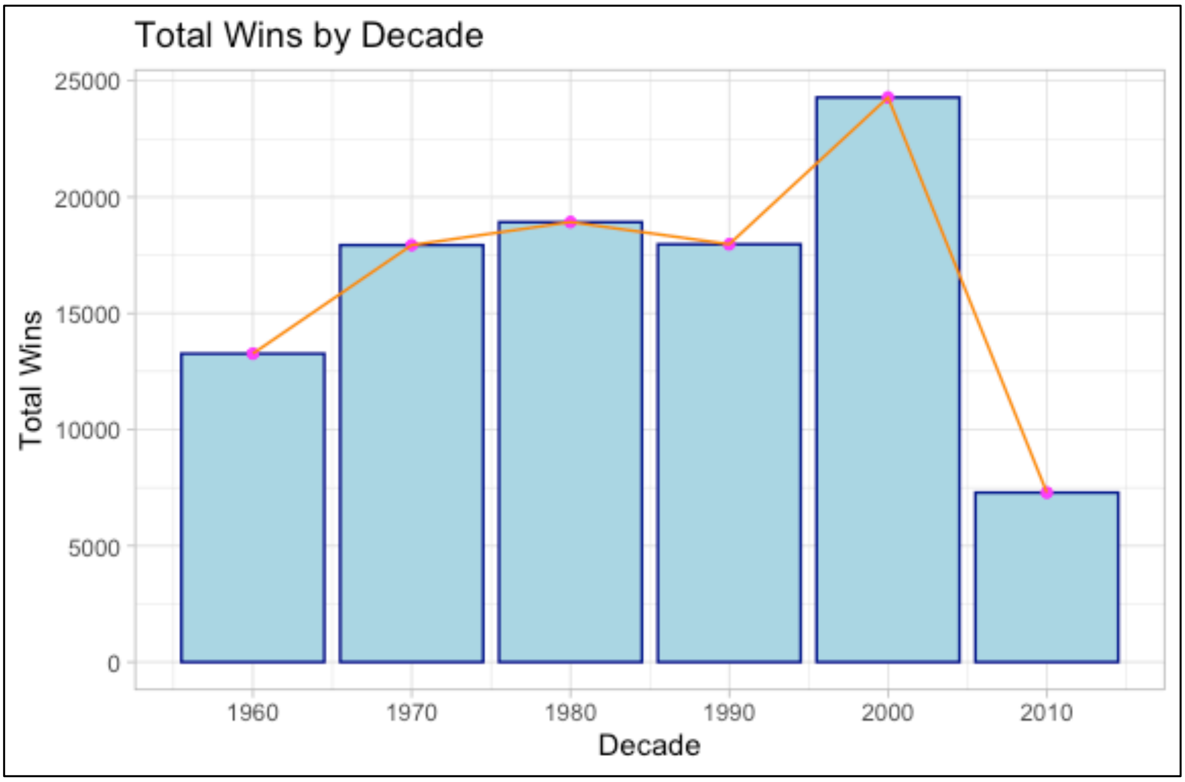

Figure 1.6: Total Wins by Decade

1. The Total wins by decade plot shows that there is a general increasing pattern from 1960 to 2000.
2. We cannot say anything about 2010 decade as the data is not available after 2012 year.
3. There is a decrease in the total wins in 1990 decade but increased further up.

Data Imputation -

We need to extract decade from the variable 'Year' and create a table containing information about the **wins per decade**.

```
# Extract Decade from Year
baseball$decade <- baseball$Year - (baseball$Year %% 10)

# Create a wins table by summing the wins by decade
baseballDecadeWins <- baseball %>%
  group_by(decade) %>%
  summarise(wins = sum(W)) %>%
  as.tibble()
```

Figure 1.5: Data Imputation in the Crop dataset

Alpha value :
0.05

Null Hypothesis -
H_0 : There is no difference in number of wins by decade

Alternate Hypothesis -
H_1 : There is difference in number of wins by decade

Chi-Square Test Value :
30

Chi-Square P-Value :
0.224289004834403

Degree of Freedom :
25

Critical Value :
37.6525

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qchisq()** function.

Result :
Failed to reject the null hypothesis as the P-value of 0.2243 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that t*here is difference in number of wins by decade.*

PROBLEM 10 - CROP DATA

Perform a Two-way ANOVA test using *yield* as the dependent variable and *fertilizer* and *density* as the independent variables. Explain the results of the test. Is there reason to believe that fertilizer and density have an impact on yield?

The first 6 data points are presented below of the 'Crop' dataset to showcase the variables and information.

Crop Dataset

density	block	fertilizer	yield
1	1	1	177.2287
2	2	1	177.5500
1	3	1	176.4085
2	4	1	177.7036
1	1	1	177.1255
2	2	1	176.7783

Table 1.11: Top 6 data points of the Crop Dataset

Data Imputation -

We need to factorize the integer type/domain variables like *density, block, fertilizer.*

```
# Convert variables in factors
cropData <- cropData %>%
  mutate(
    density = as.factor(density),
    block = as.factor(block),
    fertilizer = as.factor(fertilizer)
  )
```

Figure 1.4: Data Imputation in the Crop dataset

Crop Data for Fertilizer and Density ANOVA Test Summary Table

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
fertilizer	2	6.0680466	3.0340233	9.0010522	0.0002731890
density	1	5.1216812	5.1216812	15.1945174	0.0001864075
fertilizer:density	2	0.4278183	0.2139091	0.6346053	0.5325000914
Residuals	90	30.3366866	0.3370743		

Table 1.12: Crop Data for Fertilizer and Density ANOVA Test Summary

Alpha value :
0.05

Since, this test is a Two-Way ANOVA test and because we need null and alternative hypotheses for the effect on both categorical factors, and the effect of the categorical factors on each other, the null and alternative hypothesis pairs may be expressed as follows.

(3 Pairs in 2-Way ANOVA)
1st Pair:

Null Hypothesis -
H_0 : The means of all Fertilizer groups are same

Alternate Hypothesis -
H_1 : The means of all Fertilizer groups are different

2nd Pair:

Null Hypothesis -
H_0 : The means of all Density groups are same

Alternate Hypothesis -
H_1 : The means of all Density groups are different

3rd Pair:

Null Hypothesis -

H_0 : There is no interaction between the Fertilizer and Density

Alternate Hypothesis -

H_1 : There is interaction between the Fertilizer and Density

```
# Run the ANOVA test
anova <- aov(yield ~ fertilizer + density + fertilizer:density, data = cropData)
```

Figure 1.5: Two-Way ANOVA Test script code.

Degree of Freedom -

k - 1: Between Group Variance - Numerator (Fertilizer) : **2**
k - 1: Between Group Variance - Numerator (Density) : **1**
k - 1: Between Group Variance - Numerator (Fertilizer: Density) : **2**
N - k: Within Group Variance - Denominator : **90**

ANOVA test **F-Value** :

Fertilizer	Density	Fertilizer : Density
9.001052	**15.19452**	**0.6346053**

ANOVA test **P-Value** :

Fertilizer	Density	Fertilizer : Density
0.000273189	**0.0001864075**	**0.5325001**

Critical Value :

Fertilizer	Density	Fertilizer : Density
3.0977	**3.9469**	**3.0977**

The chi-square test is performed using the **chisq.test()** function in R and its critical value is computed using the **qf()** function.

Results :

Reject the null hypothesis that means of all Fertilizer groups are same as the P-value of 0.000273189 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that *the means of all Fertilizer groups are different.*

Reject the null hypothesis that means of all Density groups are same as the P-value of 0.0001864075 is smaller than the alpha value of 0.05. The results are significant. Therefore, we have sufficient evidences to claim that *the means of all Density groups are different.*

Failed to reject the null hypothesis that there is no interaction between the Fertilizer and Density as the P-value of 0.5325001 is greater than the alpha value of 0.05. The results are not significant. Therefore, we do not have sufficient evidences to claim that *there is interaction between the Fertilizers and Densities.*

CONCLUSION

We have conducted/performed the *Chi-Square Test and ANOVA Test* on various assignment questions and found some insights of the data sets used in them.

We used Chi-Square Test to analyse the Goodness-of-Fit and the Independence of dependent variables on independent variables. We also used ANOVA test to figure out the equality of three or more population means by analysing the sample variances to determine whether a relationship exists between them or not. There are two methods present in ANOVA test. One-way ANOVA test and Two-way ANOVA test have been used to analyse the questions present in the assignment.

In various questions, we conducted the Chi-Square or ANOVA test and were able to reject the Null Hypothesis which helped us to gather sufficient evidences to claim their respective Alternate Hypotheses. In some questions, we failed to reject the Null Hypothesis and therefore, couldn't gather sufficient evidences to claim their respective Alternate Hypotheses.

The dataset of baseball was analysed and we found that we were not able to reject the Null Hypothesis that there is no difference in number of wins by decade. Therefore, we do not have sufficient evidences to claim the alternate hypothesis that there is difference in numbers of wins by decade.

The crop dataset was investigated from which we can conclude that we have sufficient evidences to claim that the means of all Fertilizer groups are different and also that the means of all Density groups are different. But, we failed to reject the null hypothesis that there is no interaction between the Fertilizer and Density. Therefore, we cannot say that there is interaction between the Fertilizer and Density and that they have effect on the yield variable on the dataset.

We can conclude that the Chi Square Test of Association Method and ANOVA Test of Hypothesis Testing allow businesses to test theories regarding the relationship of one or more data points to another data point to determine possible influencing factors for product purchases, or other outcomes.

BIBLIOGRAPHY

1. *Home - RDocumentation*. (2021). Functions in R - Documentation. https://www.rdocumentation.org/

2. ALY 6015 - Prof Roy Wada - *Lesson 2-1 — Chi-Square Goodness-of-Fit* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-2-1-chi-square-goodness-of-fit?module_item_id=6646955

3. ALY 6015 - Prof Roy Wada - *Lesson 2-3 — Chi-Square Independence Test* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-2-2-chi-square-independence-test?module_item_id=6646958

4. ALY 6015 - Prof Roy Wada - *Lesson 2-4 — Analysis of Variance* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-2-3-analysis-of-variance-anova?module_item_id=6646960

5. ALY 6015 - Prof Roy Wada - *Lesson 1-6 — Feature/Variable Selection* (2022, February), https://northeastern.instructure.com/courses/98028/assignments/1207970?module_item_id=6646970

6. *Chi-Square (χ2) Statistic Definition*. (2021, September 20). Investopedia. https://www.investopedia.com/terms/c/chi-square-statistic.asp

APPENDIX

```
#-------- ALY6015_M2_ChiSquare&ANOVA_HarshitGaur --------#

print("Author : Harshit Gaur")
print("ALY 6015 Week 2 Assignment - Chi-Square and ANOVA")

# Declaring the names of packages to be imported
packageList <- c("tidyverse", "vtable", "RColorBrewer", "psych", "flextable")

for (package in packageList) {
  if (!package %in% rownames(installed.packages()))
  { install.packages(package) }

  # Import the package
  library(package, character.only = TRUE)
}


# Steps required to properly perform/conduct the Chi-Square or ANOVA test.
# Step 1 - State the hypotheses and identify the claim.
# Step 2 - Find the critical value.
# Step 3 - Compute the test value.
# Step 4 - Make the decision.
# Step 5 - Summarize the conclusion/results.


####################################################################
# Section 11-1
####################################################################

###### Question 6. Blood Type ######

# State the Hypothesis
# H0: Type-A = 0.20, Type-B = 0.28, Type-O = 0.36, Type-AB = 0.16
# H1: The blood type distribution is not the same in the hospital's patient population
#     as stated in the null hypothesis

# Set Significance Level
alpha = 0.10

# Create a vector of the values
observed <- c(12, 8, 24, 6)

# Create a vector of the probabilities
prob <- c(0.20, 0.28, 0.36, 0.16)

# Create a matrix from the rows
matrix_obj <- matrix(c(c("Type-A", "Type-B", "Type-O", "Type-AB"), sum(observed) * prob, observed), nrow =
length(observed), byrow = FALSE,
                    dimnames = list(c(), c("", "Expected", "Observed")))

# Save 3-Line Table
save_as_docx('Blood Types Table' = flextable(data = as.data.frame(matrix_obj)),
         path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/11-1-
Blood.docx')

# Run the test and save the results
result <- chisq.test(x = observed, p = prob)
```

```
# View the test statistic and p-value
paste("Chi-Square Test Value :", result$statistic)
paste("Chi-Square P-Value :", result$p.value)
paste("Degree of Freedom :", result$parameter)

# Critical Value
paste("Critical Value :", round(qchisq(p = alpha, df = result$parameter, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))


###### Question 8. On-Time Performance by Airlines ######

# State the 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
# H1: The on-time performance distribution of airlines is not the same
#     as stated in the null hypothesis

# Set Significance Level
alpha = 0.05

# Create a vector of the values
observed <- c(125, 10, 25, 40)

# Create a vector of the probabilities
prob <- c(0.708, 0.082, 0.09, 0.12)

# Run the test and save the results
result <- chisq.test(x = observed, p = prob)

# Create a matrix from the rows
matrix_obj <- matrix(c(c("On-Time", "National Aviation System Delay", "Aircraft Arriving Late", "Other
(because of weather and other conditions)"), sum(observed) * prob, observed), nrow = length(observed),
byrow = FALSE,
                  dimnames = list(c(), c("", "Expected", "Observed")))

# Save 3-Line Table
save_as_docx('On-Time Performance by Airlines Table' = flextable(data = as.data.frame(matrix_obj)),
         path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/11-1-
Airlines.docx')

# View the test statistic and p-value
paste("Chi-Square Test Value :", result$statistic)
paste("Chi-Square P-Value :", result$p.value)
paste("Degree of Freedom :", result$parameter)

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))

# Critical Value
paste("Critical Value :", round(qchisq(p = alpha, df = result$parameter, lower.tail = FALSE), 4))
```

```
###############################################################
# Section 11-2
###############################################################

###### Question 8. Ethnicity and Movie Admissions ######

# State the Hypothesis
# H0: Movie admissions are independent of ethnicity
# H1: Movie admissions are dependent on ethnicity

# Set Significance Level
alpha = 0.05

# Create one vector for each row
row_2013 <- c(724, 335, 174, 107)
row_2014 <- c(370, 292, 152, 140)

# State the number of rows for the matrix
rows <- 2

# Create a matrix from the rows
matrix_obj <- matrix(c(row_2013, row_2014), nrow = rows, byrow = TRUE)

# Name the rows and columns of the matrix
rownames(matrix_obj) <- c("2013", "2014")
colnames(matrix_obj) <- c("Caucasian", "Hispanic", "African American", "Other")

# Verify the matrix
matrix_obj

# Save 3-Line Table
save_as_docx('Ethnicity and Movie Admissions Table' = flextable(data = as.data.frame(cbind(c("2013", "2014"),
matrix_obj))),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/11-2-
Ethnicity.docx')

# Run the test and save the results
result <- chisq.test(matrix_obj)

# View the test statistic and p-value
paste("Chi-Square Test Value :", result$statistic)
paste("Chi-Square P-Value :", format(result$p.value, scientific = FALSE))
paste("Degree of Freedom :", result$parameter)

# Critical Value
paste("Critical Value :", round(qchisq(p = alpha, df = result$parameter, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
     paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 15), scientific =
FALSE), "is greater than the alpha value of", alpha),
     paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 15), scientific = FALSE),
"is smaller than the alpha value of", alpha))


###### Question 8. Women in the Military ######

# State the Hypothesis
# H0: Ranks of women in Armed Forces are independent of their branches
# H1: Ranks of women in Armed Forces are dependent on their branches
```

```
# Set Significance Level
alpha = 0.05

# Create one vector for each row
row_army <- c(10791, 62491)
row_navy <- c(7816, 42750)
row_marine <- c(932, 9525)
row_air <- c(11819, 54344)

# State the number of rows for the matrix
rows <- 4

# Create a matrix from the rows
matrix_obj <- matrix(c(row_army, row_navy, row_marine, row_air), nrow = rows, byrow = TRUE)

# Name the rows and columns of the matrix
rownames(matrix_obj) <- c("Army", "Navy", "Marine Corps", "Air Corps")
colnames(matrix_obj) <- c("Officers", "Enlisted")

# Verify the matrix
matrix_obj

# Save 3-Line Table
save_as_docx('Women in the Military Table' = flextable(data = as.data.frame(cbind(c("Army", "Navy", "Marine
Corps", "Air Corps"), matrix_obj))),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/11-2-
Women.docx')

# Run the test and save the results
result <- chisq.test(matrix_obj)

# View the test statistic and p-value
paste("Chi-Square Test Value :", result$statistic)
paste("Chi-Square P-Value :", format(result$p.value, scientific = FALSE))
paste("Degree of Freedom :", result$parameter)

# Critical Value
paste("Critical Value :", round(qchisq(p = alpha, df = result$parameter, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 10), scientific =
FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 10), scientific = FALSE),
"is smaller than the alpha value of", alpha))



######################################################################
# Section 12-1
######################################################################

###### Question 8. Sodium Contents of Foods ######

# State the Hypothesis
# H0: µ-Condiments = µ-Cereals = µ-Desserts
# H1: At least one mean is different from the others in the null hypothesis.

# Set Significance Level
alpha = 0.05
```

```
# Create a data frame for the 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)

# Create a matrix from the rows
matrix_obj <- matrix(c(
            c(270, 130, 230, 180, 80, 70, 200, ''),
            c(260, 220, 290, 290, 200, 320, 140, ''),
            c(100, 180, 250, 250, 300, 360, 300, 160)
           ), nrow = 8, byrow = FALSE,
           dimnames = list(c(), c("Condiments", "Cereals", "Desserts")))

# Save 3-Line Table
save_as_docx('Sodium Contents of Food Table' = flextable(data = as.data.frame(matrix_obj)),
         path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/12-1-
Sodium.docx')

# 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 and save it
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
F.value <- a.summary[[1]][[1, "F value"]]
F.value

# Extract the P-value
P.value <- a.summary[[1]][[1, "Pr(>F)"]]
P.value

# Critical Value
paste("Critical Value :", round(qf(p = alpha, df1 = df.numerator, df2 = df.denominator, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(P.value > alpha,
    paste("Failed to reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE),
"is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE), "is
smaller than the alpha value of", alpha))
```

```
##################################################################
# Section 12-2
##################################################################

###### Question 10. Sales for Leading Companies ######

# State the Hypothesis
# H0: µ-Cereals = µ-Chocolate Candy = µ-Coffee
# H1: At least one mean is different from the others in the null hypothesis.

# Set Significance Level
alpha = 0.01

# Create a data frame for the Cereals
cereals <- data.frame('sales' = c(578, 320, 264, 249, 237), 'food' = rep('cereals', 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)

# Create a matrix from the rows
matrix_obj <- matrix(c(
  c(578, 320, 264, 249, 237),
  c(311, 106, 109, 125, 173),
  c(261, 185, 302, 689, '')
), nrow = 5, byrow = FALSE,
dimnames = list(c(), c("Cereals", "Chocolate Candy", "Coffee")))

# Save 3-Line Table
save_as_docx('Sales for Leading Companies Table' = flextable(data = as.data.frame(matrix_obj)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/12-2-
Sales.docx')

# Combine the data frames into one
sales <- rbind(cereals, chocolateCandy, coffee)
sales$food <- as.factor(sales$food)

# Run the ANOVA test
anova <- aov(sales ~ food, data = sales)

# View the model summary and save it
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
F.value <- a.summary[[1]][[1, "F value"]]
F.value
```

```
# Extract the P-value
P.value <- a.summary[[1]][[1, "Pr(>F)"]]
P.value

# Critical Value
paste("Critical Value :", round(qf(p = alpha, df1 = df.numerator, df2 = df.denominator, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(P.value > alpha,
      paste("Failed to reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE),
"is greater than the alpha value of", alpha),
      paste("Reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE), "is
smaller than the alpha value of", alpha))


###### Question 12. Per-Pupil Expenditures ######

# State the Hypothesis
# H0: μ-Eastern Third = μ-Middle Third = μ-Western Third
# H1: At least one mean is different from the others in the null hypothesis.

# Set Significance Level
alpha = 0.05

# Create a data frame for the Eastern Third
easternThird <- data.frame('expenditure' = c(4946, 5953, 6202, 7243, 6113), 'state' = rep('Eastern Third', 5),
stringsAsFactors = FALSE)

# Create a data frame for the Middle Third
middleThird <- data.frame('expenditure' = c(6149, 7451, 6000, 6479), 'state' = rep('Middle Third', 4),
stringsAsFactors = FALSE)

# Create a data frame for the Western Third
westernThird <- data.frame('expenditure' = c(5282, 8605, 6528, 6911), 'state' = rep('Western Third', 4),
stringsAsFactors = FALSE)

# Create a matrix from the rows
matrix_obj <- matrix(c(
  c(4946, 5953, 6202, 7243, 6113),
  c(6149, 7451, 6000, 6479, ''),
  c(5282, 8605, 6528, 6911, '')
), nrow = 5, byrow = FALSE,
dimnames = list(c(), c("Eastern Third", "Middle Third", "Western Third")))

# Save 3-Line Table
save_as_docx('Per-Pupil Expenditures Table' = flextable(data = as.data.frame(matrix_obj)),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/12-2-
Pupil.docx')


# Combine the data frames into one
expenditure <- rbind(easternThird, middleThird, westernThird)
expenditure$state <- as.factor(expenditure$state)

# Run the ANOVA test
anova <- aov(expenditure ~ state, data = expenditure)

# View the model summary and save it
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
F.value <- a.summary[[1]][[1, "F value"]]
F.value

# Extract the P-value
P.value <- a.summary[[1]][[1, "Pr(>F)"]]
P.value

# Critical Value
paste("Critical Value :", round(qf(p = alpha, df1 = df.numerator, df2 = df.denominator, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(P.value > alpha,
     paste("Failed to reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE),
"is greater than the alpha value of", alpha),
     paste("Reject the null hypothesis as the P-value of", format(round(P.value, 10), scientific = FALSE), "is
smaller than the alpha value of", alpha))



##################################################################
# Section 12-3
##################################################################

###### Question 10. Increasing Plant Growth ######

# State the Hypothesis (3 Pairs in 2-Way ANOVA)
# H0: The means of all Plant-Food Supplement groups are same
# H1: The means of all Plant-Food Supplement groups are different

# H0: The means of all Growth-Light groups are same
# H1: The means of all Growth-Light groups are different

# H0: There is no interaction between the Growth-Light and Plant-Food Supplement
# H1: There is interaction between the Growth-Light and Plant-Food Supplement

# Set Significance Level
alpha = 0.05

# Create a data frame
plantsGrowth <- data.frame('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),
                    'growth_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)

# Save 3-Line Table
save_as_docx('Increasing Plant Growth Table' = flextable(data = plantsGrowth),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/12-3-
PlantGrowth.docx')


# Run the ANOVA test
anova <- aov(growth ~ growth_light + plant_food + growth_light:plant_food, data = plantsGrowth)
```

```
# View the model summary and save it
a.summary <- summary(anova)
a.summary

# Save 3-Line Table
df <- data.frame(unclass(a.summary), stringsAsFactors = FALSE, check.rows = TRUE)
save_as_docx('Increasing Plant Growth ANOVA Test Summary Table' = flextable(data =
cbind(trimws(rownames(df)), df)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class 2/Assignment/Tables/12-3-
PlantGrowth_Summary.docx')

# Degrees of Freedom
# k - 1: Between Group Variance - Numerator (Growth Light)
df.numerator_growthLight <- a.summary[[1]][1, "Df"]
df.numerator_growthLight

# k - 1: Between Group Variance - Numerator (Plant Food)
df.numerator_plantFood <- a.summary[[1]][2, "Df"]
df.numerator_plantFood

# k - 1: Between Group Variance - Numerator (Growth Light : Plant Food)
df.numerator_growthLight_plantFood <- a.summary[[1]][3, "Df"]
df.numerator_growthLight_plantFood

# N - k: Within Group Variance - Denominator
df.denominator <- a.summary[[1]][4, "Df"]
df.denominator

# Extract the F test value (Growth Light)
F.value_growthLight <- a.summary[[1]][[1, "F value"]]
F.value_growthLight

# Extract the F test value (Plant Food)
F.value_plantFood <- a.summary[[1]][[2, "F value"]]
F.value_plantFood

# Extract the F test value (Growth Light : Plant Food)
F.value_growthLight_plantFood <- a.summary[[1]][[3, "F value"]]
F.value_growthLight_plantFood

# Extract the P-value (Growth Light)
P.value_growthLight <- a.summary[[1]][[1, "Pr(>F)"]]
P.value_growthLight

# Extract the P-value (Plant Food)
P.value_plantFood <- a.summary[[1]][[2, "Pr(>F)"]]
P.value_plantFood

# Extract the P-value (Growth Light : Plant Food)
P.value_growthLight_plantFood <- a.summary[[1]][[3, "Pr(>F)"]]
P.value_growthLight_plantFood

# Critical Value (Growth Light)
paste("Critical Value of Growth Light :", round(qf(p = alpha, df1 = df.numerator_growthLight, df2 =
df.denominator, lower.tail = FALSE), 4))

# Critical Value (Plant Food)
paste("Critical Value of Plant Food :", round(qf(p = alpha, df1 = df.numerator_plantFood, df2 =
df.denominator, lower.tail = FALSE), 4))

# Critical Value (Growth Light : Plant Food)
```

```
paste("Critical Value of Growth Light:Plant Food =", round(qf(p = alpha, df1 =
df.numerator_growthLight_plantFood, df2 = df.denominator, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result (Growth Light)
ifelse(P.value_growthLight > alpha,
      paste("Failed to reject the null hypothesis that means of all Growth-Light groups are same as the P-value
of", format(round(P.value_growthLight, 10), scientific = FALSE), "is greater than the alpha value of", alpha),
      paste("Reject the null hypothesis that means of all Growth-Light groups are same as the P-value of",
format(round(P.value_growthLight, 10), scientific = FALSE), "is smaller than the alpha value of", alpha))

# Compare the p-value and alpha to decide the result (Plant Food)
ifelse(P.value_plantFood > alpha,
      paste("Failed to reject the null hypothesis that means of all Plant-Food Supplement groups are same as the
P-value of", format(round(P.value_plantFood, 10), scientific = FALSE), "is greater than the alpha value of",
alpha),
      paste("Reject the null hypothesis that means of all Plant-Food Supplement groups are same as the P-value
of", format(round(P.value_plantFood, 10), scientific = FALSE), "is smaller than the alpha value of", alpha))

# Compare the p-value and alpha to decide the result (Growth Light : Plant Food)
ifelse(P.value_growthLight_plantFood > alpha,
      paste("Failed to reject the null hypothesis that there is no interaction between the Growth-Light and Plant-
Food Supplement as the P-value of", format(round(P.value_growthLight_plantFood, 10), scientific = FALSE), "is
greater than the alpha value of", alpha),
      paste("Reject the null hypothesis that there is no interaction between the Growth-Light and Plant-Food
Supplement as the P-value of", format(round(P.value_growthLight_plantFood, 10), scientific = FALSE), "is
smaller than the alpha value of", alpha))



#################################################################
# Baseball.CSV
#################################################################

# Import the data set
baseball <- read.csv('Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/baseball.csv', header = TRUE)


save_as_docx('Baseball Dataset' = flextable(data = head(baseball)),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/Tables/Baseball_Data_Table.docx')

# Get the glimpse of data set
glimpse(baseball)

describeFlexball <- baseball %>%
  psych::describe(quant = c(.25, .75), IQR = TRUE) %>%
  select(n, mean, sd, median, min, max, range, skew, kurtosis)

describeFlexball <- round(describeFlexball, 2)
describeFlexball <- cbind(e = rownames(describeFlexball), describeFlexball)
save_as_docx('Descriptive Statistics of Baseball Dataset' = flextable(data = describeFlexball),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/Tables/Baseball_Desc_Stats_Table_main.docx')

# Normality Check for 'Wins' using Q-Q Plot and Shapiro-Wilks Test.
qqPlot(baseball$W, ylab = "Studentized Residuals", xlab = "Theoretical Quantiles")
shapiro.test(baseball$W)


# Extract Decade from Year
baseball$decade <- baseball$Year - (baseball$Year %% 10)
```

```
# Create a wins table by summing the wins by decade
baseballDecadeWins <- baseball %>%
  group_by(decade) %>%
  summarise(wins = sum(W)) %>%
  as.tibble()

# Plot to investigate the trend of Wins segregated by Decade.
ggplot(baseballDecadeWins, mapping = aes(x= decade, y= wins)) +
  geom_bar(stat = "identity", fill = "LIGHTBLUE", colour = "DARKBLUE") +
  geom_point(colour = "MAGENTA") +
  geom_line(colour = "DARKORANGE") +
  labs(title = "Total Wins by Decade", x = "Decade", y = "Total Wins") +
  scale_x_continuous(breaks = scales::pretty_breaks(n=7)) +
  theme_light()


# State the Hypothesis
# H0: There is no difference in number of wins by decade
# H1: There is difference in number of wins by decade

# Set Significance Level
alpha = 0.05

# Run the test and save the results
result <- chisq.test(x = baseballDecadeWins$decade, y = baseballDecadeWins$wins)

# View the test statistic and p-value
paste("Chi-Square Test Value :", result$statistic)
paste("Chi-Square P-Value :", result$p.value)
paste("Degree of Freedom :", result$parameter)

# Critical Value
paste("Critical Value :", round(qchisq(p = alpha, df = result$parameter, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result
ifelse(result$p.value > alpha,
       paste("Failed to reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific =
FALSE), "is greater than the alpha value of", alpha),
       paste("Reject the null hypothesis as the P-value of", format(round(result$p.value, 4), scientific = FALSE), "is
smaller than the alpha value of", alpha))




####################################################################
# Crop Data.CSV
####################################################################


# Import the data set
cropData <- read.csv('Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/crop_data.csv', header = TRUE)

save_as_docx('Crop Dataset' = flextable(data = head(cropData)),
             path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/Tables/Crop_Data_Table.docx')

# Get the glimpse of data set
glimpse(cropData)

# Convert variables in factors
```

```
cropData <- cropData %>%
  mutate(
    density = as.factor(density),
    block = as.factor(block),
    fertilizer = as.factor(fertilizer)
  )

# State the Hypothesis (3 Pairs in 2-Way ANOVA)
# H0: The means of all Fertilizer groups are same
# H1: The means of all Fertilizer groups are different

# H0: The means of all Density groups are same
# H1: The means of all Density groups are different

# H0: There is no interaction between the Fertilizer and Density
# H1: There is interaction between the Fertilizer and Density

# Set Significance Level
alpha = 0.05

# Run the ANOVA test
anova <- aov(yield ~ fertilizer + density + fertilizer:density, data = cropData)

# View the model summary and save it
a.summary <- summary(anova)
a.summary

# Save 3-Line Table
df <- data.frame(unclass(a.summary), stringsAsFactors = FALSE, check.rows = TRUE)
save_as_docx('Crop Data for Fertilizer and Density ANOVA Test Summary Table' = flextable(data =
cbind(trimws(rownames(df)), df)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
2/Assignment/Tables/Crop_Data_Summary.docx')

# Degrees of Freedom
# k - 1: Between Group Variance - Numerator (Fertilizer)
df.numerator_fertilizer <- a.summary[[1]][1, "Df"]
df.numerator_fertilizer

# k - 1: Between Group Variance - Numerator (Density)
df.numerator_density <- a.summary[[1]][2, "Df"]
df.numerator_density

# k - 1: Between Group Variance - Numerator (Fertilizer : Density)
df.numerator_fertilizer_density <- a.summary[[1]][3, "Df"]
df.numerator_fertilizer_density

# N - k: Within Group Variance - Denominator
df.denominator <- a.summary[[1]][4, "Df"]
df.denominator

# Extract the F test value (Fertilizer)
F.value_fertilizer <- a.summary[[1]][[1, "F value"]]
F.value_fertilizer

# Extract the F test value (Density)
F.value_density <- a.summary[[1]][[2, "F value"]]
F.value_density

# Extract the F test value (Fertilizer : Density)
F.value_fertilizer_density <- a.summary[[1]][[3, "F value"]]
F.value_fertilizer_density
```

```
# Extract the P-value (Fertilizer)
P.value_fertilizer <- a.summary[[1]][[1, "Pr(>F)"]]
P.value_fertilizer

# Extract the P-value (Density)
P.value_density <- a.summary[[1]][[2, "Pr(>F)"]]
P.value_density

# Extract the P-value (Fertilizer : Density)
P.value_fertilizer_density <- a.summary[[1]][[3, "Pr(>F)"]]
P.value_fertilizer_density

# Critical Value (Fertilizer)
paste("Critical Value of Fertilizer :", round(qf(p = alpha, df1 = df.numerator_fertilizer, df2 = df.denominator,
lower.tail = FALSE), 4))

# Critical Value (Density)
paste("Critical Value of Density :", round(qf(p = alpha, df1 = df.numerator_density, df2 = df.denominator,
lower.tail = FALSE), 4))

# Critical Value (Fertilizer : Density)
paste("Critical Value of Fertilizer:Density =", round(qf(p = alpha, df1 = df.numerator_fertilizer_density, df2 =
df.denominator, lower.tail = FALSE), 4))

# Compare the p-value and alpha to decide the result (Fertilizer)
ifelse(P.value_fertilizer > alpha,
    paste("Failed to reject the null hypothesis that means of all Fertilizer groups are same as the P-value of",
format(round(P.value_fertilizer, 10), scientific = FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis that means of all Fertilizer groups are same as the P-value of",
format(round(P.value_fertilizer, 10), scientific = FALSE), "is smaller than the alpha value of", alpha))

# Compare the p-value and alpha to decide the result (Density)
ifelse(P.value_density > alpha,
    paste("Failed to reject the null hypothesis that means of all Density groups are same as the P-value of",
format(round(P.value_density, 10), scientific = FALSE), "is greater than the alpha value of", alpha),
    paste("Reject the null hypothesis that means of all Density groups are same as the P-value of",
format(round(P.value_density, 10), scientific = FALSE), "is smaller than the alpha value of", alpha))

# Compare the p-value and alpha to decide the result (Fertilizer : Density)
ifelse(P.value_fertilizer_density > alpha,
    paste("Failed to reject the null hypothesis that there is no interaction between the Fertilizer and Density as
the P-value of", format(round(P.value_fertilizer_density, 10), scientific = FALSE), "is greater than the alpha
value of", alpha),
    paste("Reject the null hypothesis that there is no interaction between the Fertilizer and Density as the P-
value of", format(round(P.value_fertilizer_density, 10), scientific = FALSE), "is smaller than the alpha value of",
alpha))




#-------- END --------#
```