INTRODUCTION

This report includes a diagnosis that shows how to use RStudio to fit, analyses, and assess a regression model. This assignment module's learning outcomes are as follows:

• Using standard functions and diagnostic procedures, fit, analyze, and assess chi square & ANOVA models.

• Address chi square independence and goodness of fit test, One way and Two way ANOVA.

What is One way ANOVA?

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test (Kent State University., 2022).

This test is also known as:

One-Factor ANOVA

One-Way Analysis of Variance

Between Subjects ANOVA

The variables used in this test are known as:

Dependent variable

Independent variable (also known as the grouping variable, or factor)

This variable divides cases into two or more mutually exclusive levels, or groups

What is Two-way ANOVA?

ANOVA stands for analysis of variance and tests for differences in the effects of independent variables on a dependent variable. A two-way ANOVA test is a statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable.

A two-way ANOVA tests the effect of two independent variables on a dependent variable. A two-way ANOVA test analyzes the effect of the independent variables on the expected outcome along with their relationship to the outcome itself. Random factors would be considered to have no statistical influence on a data set, while systematic factors would be considered to have statistical significance (Investopedia., 2021).

ANALYSIS SECTION

Libraries Used

```
library(psych)
library(readxl)
library(tidyverse)
library(dplyr)
library(RColorBrewer)
library(knitr)
library(ggplot2)
library(DT)
library(corrplot)
library(dlookr)
library(car)
library(leaps)
```

First Task Part One

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

Perform these steps:

a. State the hypotheses and identify the claim.

#Ho = Blood type distribution is same as the distribution of general population.

#Ha = Blood type distribution differs from the distribution of general population.

b. Find the critical value and test value. Then make the decision.

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 5.4714, df = 3, p-value = 0.1404
```

```
ifelse(result$p.value > alpha, "Fail to reject Ho", "Reject Ho")
```

```
## [1] "Fail to reject Ho"
```

c. Summarize the results

There is not enough evidence to reject the claim that blood type distribution is same as the distribution of general population.

First Task Part Two

According to the Bureau of Transportation Statistics, on-time performance by the airlines is described as follows:

Action - % of Time

On time - 70.8

National Aviation System delay - 8.2

Aircraft arriving late - 9.0

Other (because of weather and other conditions) - 12.0

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

a. State the hypotheses and identify the claim.

#Ho = Results are same as government statistics

#Ha = Results are different from government statistics

b. Find the critical value and test value. Then make the decision.

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 39.504, df = 3, p-value = 1.357e-08
```

```
ifelse(result$p.value > alpha, "Fail to reject Ho", "Reject Ho")
```

```
## [1] "Reject Ho"
```

c. Summarize the results

There is enough evidence to reject the claim that results are same as government statistics. Hence, the results are different from government statistics.

Second Task Part One

Are movie admissions related to ethnicity? A 2014 study indicated the following numbers of admissions (in thousands) for two different years. At the 0.05 level of significance, can it be concluded that movie attendance by year was dependent upon ethnicity?

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

a. State the hypotheses and identify the claim.

#Ho = Movie attendance by the year was independent upon ethnicity

#Ha = Movie attendance by the year was dependent upon ethnicity

b. Find the critical value and test value. Then make the decision.

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

```
ifelse(result$p.value > alpha, "Fail to reject Ho", "Reject Ho")
```

```
## [1] "Reject Ho"
```

c. Summarize the results

There is enough evidence to reject the claim that Movie attendance by the year was independent upon ethnicity. Hence, the Movie attendance by the year was dependent upon ethnicity.

Second Task Part Two

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

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

a. State the hypotheses and identify the claim.

#Ho = There is no relationship between rank and branch of Armed forces.

#Ha = There is a relationship between rank and branch of Armed forces.

b. Find the critical value and test value. Then make the decision.

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

```
ifelse(result$p.value > alpha, "Fail to reject Ho", "Reject Ho")
```

```
## [1] "Reject Ho"
```

c. Summarize the results

There is enough evidence to reject the claim that there is no relationship between rank and branch of Armed forces. Hence, there is a relationship between rank and branch of Armed forces.

Third Task

The amount of sodium (in milligrams) in one serving for a random sample of three different kinds of foods is listed. At the 0.05 level of significance, is there sufficient evidence to conclude that a difference in mean sodium amounts exists among condiments, cereals, and desserts?

Condiments	Cereals	Desserts
270	260	100
130	220	180
230	290	250
180	290	250
80	200	300
70	320	360
200	140	300
		160

a. State the hypotheses and identify the claim.

#Ho = all the means are equal

#Ha = all means are not equal (i.e., At least one of the means is different from other)

b. Find the critical value and test value. Then make the decision.

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

```
ifelse(Ftest_Val > cv, "Reject Ho", "Fail to reject Ho")
```

```
## [1] "Fail to reject Ho"
```

c. Summarize the results

There is not enough evidence to reject the claim that all the means are equal.

Fourth Task Part One

The sales in millions of dollars for a year of a sample of leading companies are shown. At α = 0.01, is there a significant difference in the means?

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

a. State the hypotheses and identify the claim.

#Ho = all the means are equal

#Ha = all means are not equal (i.e., At least one of the means is different from other)

b. Find the critical value and test value. Then make the decision.

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

```
ifelse(Ftest_Val > cv, "Reject Ho", "Fail to reject Ho")
```

```
## [1] "Fail to reject Ho"
```

```
## [1] "The decision is that mean1(cereal) is not significantly different from mean2(chocolate candy)"
```

```
## [1] "The decision is that mean1(cereal) is not significantly different from mean3(coffee)"
```

```
## [1] "The decision is that mean1(chocolate candy) is not significantly different from mean3(coffee)"
```

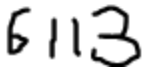
c. Summarize the results

There is not enough evidence to reject the claim that all the means are equal. We can notice that when we conduct Scheffe's test on each mean values to find out that they are not significantly different from another mean value.

Fourth Task Part Two

The expenditures (in dollars) per pupil for states in three sections of the country are listed. Using α = 0.05, can you conclude that there is a difference in means?

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



a. State the hypotheses and identify the claim.

#Ho = all the means are equal

#Ha = all means are not equal (i.e., At least one of the means is different from other)

b. Find the critical value and test value. Then make the decision.

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

```
ifelse(Ftest_Val > cv, "Reject Ho", "Fail to reject Ho")
```

```
## [1] "Fail to reject Ho"
```

```
## [1] "The decision is that mean1(eastern third) is not significantly different from mean2(middle third)"
```

```
## [1] "The decision is that mean1(eastern third) is not significantly different from mean3(western third)"
```

```
## [1] "The decision is that mean1(middle third) is not significantly different from mean3(western third)"
```

c. Summarize the results

There is not enough evidence to reject the claim that all the means are equal. We can notice that when we conduct Scheffe's test on each mean values to find out that they are not significantly different from another mean value.

Fifth Task Part One

A gardening company is testing new ways to improve plant growth. Twelve plants are randomly selected and exposed to a combination of two factors, a "Grow-light" in two different strengths and a plant food supplement with different mineral supplements. After a number of days, the plants are measured for growth, and the results (in inches) are put into the appropriate boxes.

	Grow-light 1	Grow-light 2
Plant food A	9.2, 9.4, 8.9	8.5, 9.2, 8.9
Plant food B	7.1, 7.2, 8.5	5.5, 5.8, 7.6

Can an interaction between the two factors be concluded? Is there a difference in mean growth with respect to light? With respect to plant food? Use α = 0.05

a. State the hypotheses and identify the claim.

#Ho = There is no difference in mean growth with respect to light

#Ha = There is a difference in mean growth with respect to light

#Ho = There is no difference in mean growth with respect to plant food

#Ha = There is a difference in mean growth with respect to plant food

#Ho = There is no interaction between plant food and light

#Ha = There is an interaction between plant food and light

b. Find the critical value and test value. Then make the decision.

```
##                                             Df Sum Sq Mean Sq F value
## plant_growth_df$Plant                        1 12.813  12.813  24.562
## plant_growth_df$Growlight                    1  1.920   1.920   3.681
## plant_growth_df$Plant:plant_growth_df$Growlight  1  0.750   0.750   1.438
## Residuals                                    8  4.173   0.522
##                                              Pr(>F)
## plant_growth_df$Plant                       0.00111 **
## plant_growth_df$Growlight                   0.09133 .
## plant_growth_df$Plant:plant_growth_df$Growlight 0.26482
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Summarize the results

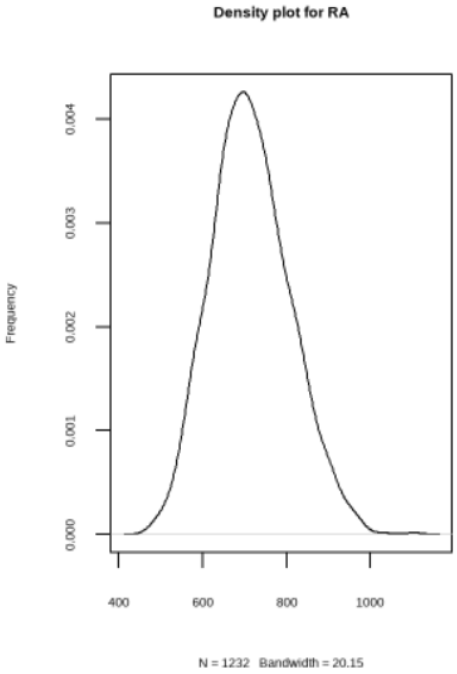
We have enough evidence to conclude that there is difference in mean growth with respect to light & mean growth with respect to plant food. For the interaction, we have don't have enough evidence to reject the claim that there is no interaction between plant food and light.

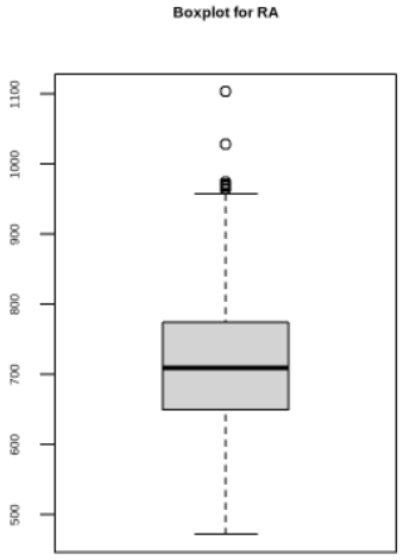
Fifth Task Part Two

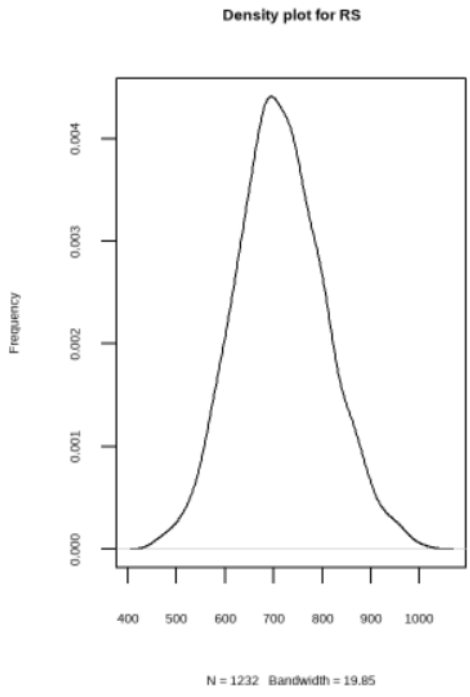
Present the baseball dataset and perform chi square test on it.

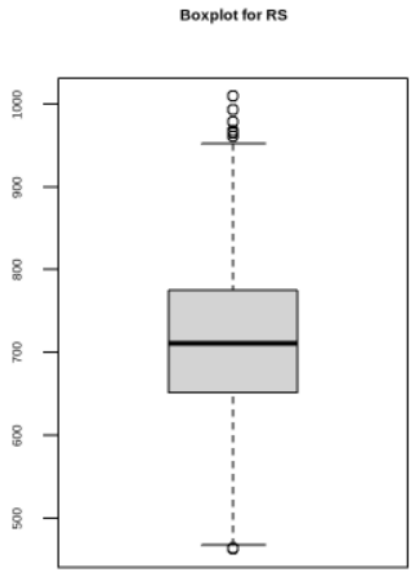
Performing EDA on the provided baseball dataset.

```
## spec_tbl_df [1,232 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Team        : chr [1:1232] "ARI" "ATL" "BAL" "BOS" ...
##  $ League      : chr [1:1232] "NL" "NL" "AL" "AL" ...
##  $ Year        : num [1:1232] 2012 2012 2012 2012 2012 ...
##  $ RS          : num [1:1232] 734 700 712 734 613 748 669 667 758 726 ...
##  $ RA          : num [1:1232] 688 600 705 806 759 676 588 845 890 670 ...
##  $ W           : num [1:1232] 81 94 93 69 61 85 97 68 64 88 ...
##  $ OBP         : num [1:1232] 0.328 0.32 0.311 0.315 0.302 0.318 0.315 0.324 0.33 0.335 ...
##  $ SLG         : num [1:1232] 0.418 0.389 0.417 0.415 0.378 0.422 0.411 0.381 0.436 0.422 ...
##  $ BA          : num [1:1232] 0.259 0.247 0.247 0.26 0.24 0.255 0.251 0.251 0.274 0.268 ...
##  $ Playoffs    : num [1:1232] 0 1 1 0 0 0 1 0 0 1 ...
##  $ RankSeason  : num [1:1232] NA 4 5 NA NA NA 2 NA NA 6 ...
##  $ RankPlayoffs: num [1:1232] NA 5 4 NA NA NA 4 NA NA 2 ...
##  $ G           : num [1:1232] 162 162 162 162 162 162 162 162 162 162 ...
##  $ OOBP        : num [1:1232] 0.317 0.306 0.315 0.331 0.335 0.319 0.305 0.336 0.357 0.314 ...
##  $ OSLG        : num [1:1232] 0.415 0.378 0.403 0.428 0.424 0.405 0.39 0.43 0.47 0.402 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Team = col_character(),
##   ..   League = col_character(),
##   ..   Year = col_double(),
##   ..   RS = col_double(),
##   ..   RA = col_double(),
##   ..   W = col_double(),
##   ..   OBP = col_double(),
##   ..   SLG = col_double(),
##   ..   BA = col_double(),
##   ..   Playoffs = col_double(),
##   ..   RankSeason = col_double(),
##   ..   RankPlayoffs = col_double(),
##   ..   G = col_double(),
##   ..   OOBP = col_double(),
##   ..   OSLG = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```









On the baseball data set, the str function was used to do basic descriptive statistics. We can see that the full data set has 1232 rows and 15 columns by utilizing that function. Some of the columns, such as RankSeason and RankPlayoff, have a lot of null values. We can see that the density map for RA (Run Average) is right skewed, implying that the mean will be bigger than the median. In addition, the boxplot shows that Run Average has a lot of outliers. The similar

scenario occurs with RS, or Run Support, where we can see that the boxplot includes a large number of outliers.

```
## # A tibble: 6 x 2
##   Decade  wins
##    <dbl> <dbl>
## 1   1960 13267
## 2   1970 17934
## 3   1980 18926
## 4   1990 17972
## 5   2000 24286
## 6   2010  7289
```

Decade was extracted from year and then using the grouped Decade data was summarized with wins per decade to create a table which will be used to calculate the Chi Square test

a. State the hypotheses and identify the claim.

#Ho = There is no difference in number of wins by decade.

#Ha = There is a difference in number of wins by decade.

b. Find the critical value and test value. Then make the decision.

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 9989.5, df = 5, p-value < 2.2e-16
```

```
ifelse(result$p.value > alpha, "Fail to reject Ho", "Reject Ho")
```

```
## [1] "Reject Ho"
```

c. Summarize the results

We have enough evidence to reject the claim that there is no difference in number of wins by decade. Hence, there is a difference in number of wins by decade.

Fifth Task Part Two

Import the crop dataset and perform two way a nova test on it.

a. State the hypotheses and identify the claim.

#Ho = There is no difference in crop with respect to fertilizer

#Ha = There is a difference in crop with respect to fertilizer

#Ho = There is no difference in crop with respect to density

#Ha = There is a difference in crop with respect to density

#Ho = There is no interaction between fertilizer and density

#Ha = There is an interaction between fertilizer and density

b. Find the critical value and test value. Then make the decision.

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

c. Summarize the results

We have enough evidence to conclude that there is difference in crop with respect to fertilizer & crop with respect to density. For the interaction, we have don't have enough evidence to reject the claim that there is no interaction between fertilizer and density.

CONCLUSION

Finally, while working on several problems, I improved my understanding of hypothesis testing by studying Chi-square goodness of fit and independence tests. I also learnt how to do one-way

and two-way ANOVA analyses using manually entered and imported data. I learnt how to utilize pipe commands and the tibble () function to summarize data for ANOVA testing as well.

BIBLIOGRAPHY

1) LibGuides: SPSS Tutorials: One-Way ANOVA. (2022). Kent State University. https://

libguides.library.kent.edu/spss/onewayanova#:%7E:text=One%2DWay%20ANO A%

20(%22analysis,One%2DFactor%20ANOVA

2) Two-Way ANOVA. (2021, April 30). Investopedia.

https://www.investopedia.com/terms/t/two-way

anova.asp#:%7E:text=A%20two%2Dway%20ANOVA%20test,variables%20on%20a%2

dependent%20variable.