A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Strategic Thinking |
| *Evaluation Title:* | CA 1 – Assignment of the final project proposal. |
| *Assessment Due Date:* | 29/03/2024. |
| *Presentation Date* | 29/03/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

[Task 1: Q1 (Student Exam Scores) Dataset Analysis 2](#_Toc167054719)

[Introduction 2](#_Toc167054720)

[Data Set Introduction 2](#_Toc167054721)

[Methodology 2](#_Toc167054722)

[Summary Statistics 3](#_Toc167054723)

[Hypothesis Testing Results 3](#_Toc167054724)

[Manual Calculation 3](#_Toc167054725)

[Visualization 4](#_Toc167054726)

[Histogram of Exam Scores 4](#_Toc167054727)

[Boxplot of Exam Scores and National Average 5](#_Toc167054728)

[Results and Discussion 5](#_Toc167054729)

[Task 2: Diamonds Dataset Analysis 5](#_Toc167054730)

[Introduction 5](#_Toc167054731)

[Dataset Introduction 5](#_Toc167054732)

[Methodology 6](#_Toc167054733)

[Summary Statistics 6](#_Toc167054734)

[Hypothesis Testing Results 7](#_Toc167054735)

[Confidence Intervals 7](#_Toc167054736)

[Visualization 8](#_Toc167054737)

[Results and Discussion 8](#_Toc167054738)

[Task 3: Plant Growth Analysis Report 9](#_Toc167054739)

[Introduction 9](#_Toc167054740)

[Dataset Introduction 9](#_Toc167054741)

[Methodology 9](#_Toc167054742)

[Summary Statistics 9](#_Toc167054743)

[Box Plot 10](#_Toc167054744)

[Violin Chart 10](#_Toc167054745)

[Bar Chart 11](#_Toc167054746)

[ANOVA Test Results 11](#_Toc167054747)

[Tukey HSD Test Results 11](#_Toc167054748)

[Visualization 12](#_Toc167054749)

[Histogram and KDE Graph of Plant Weights According to Treatment Group 12](#_Toc167054750)

[Result and Discussion 13](#_Toc167054751)

[Task 4: Trees Dataset Analysis 13](#_Toc167054752)

[Introduction 13](#_Toc167054753)

[Dataset Introduction 13](#_Toc167054754)

[Methodology 13](#_Toc167054755)

[Summary Statistics 14](#_Toc167054756)

[Correlation Matrix 14](#_Toc167054757)

[Correlation Analysis and Hypothesis Tests 14](#_Toc167054758)

[Regression Models 14](#_Toc167054759)

[Visualization 16](#_Toc167054760)

[Correlation Matrix Heat Map 16](#_Toc167054761)

[Relationship Between Girth and Volume 17](#_Toc167054762)

[Relationship Between Height and Volume 17](#_Toc167054763)

[Results and Discussion 18](#_Toc167054764)

[Github Link 18](#_Toc167054765)

# Task 1: Q1 (Student Exam Scores) Dataset Analysis

## Introduction

This report was prepared to test a teacher's concerns that students' performance was below the national average. The analysis included exam scores of 50 students from a secondary school in Dublin.

## Data Set Introduction

This dataset contains the exam scores of 50 students from a secondary school in Dublin. The dataset contains a column called “exam\_score”, which shows students exam scores as a percentage.

## Methodology

The following steps were followed to analyze the data:

1. Summary statistics were calculated.

2. A one-sample t-test was applied to test the teacher's concerns (α = 0.05).

3. Analysis results were interpreted and supported with visuals.

4. The results were verified by manual calculation.

## Summary Statistics

A screenshot of a computer screen

Description automatically generated

## Hypothesis Testing Results

A one-sample t-test was used to test the teacher's concern that students' performance was lower than the national average (national average = 70).

- t-Statistic: -0.9169

- P-Value: 0.181

- Sample Size: 50

Since the P-Value is greater than 0.05, there is no statistically significant evidence that students average performance is lower than the national average.

A black background with white text

Description automatically generated

## Manual Calculation

Calculations have been manually verified as follows:

- Sample mean: 68.74

- Population mean: 70

- Sample size: 50

- Sample standard deviation: 9.75

- Calculated t-Statistic: -0.9138

- Calculated P-Value: 0.1826

These values ​​are quite close to those obtained by our calculator and support the accuracy of the results.

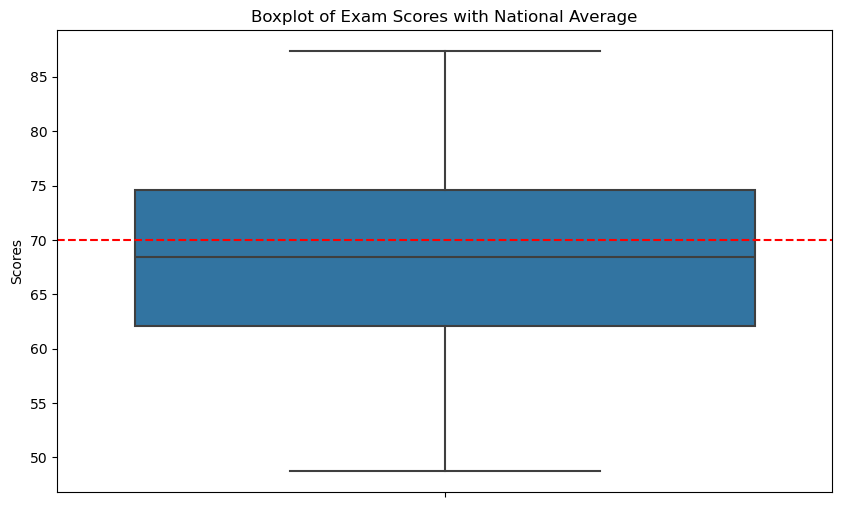
## Visualization

### Histogram of Exam Scores

A graph with a blue line

Description automatically generated

### Boxplot of Exam Scores and National Average



## Results and Discussion

According to the analysis results, the teacher's concerns that students' performance is below the national average are not statistically confirmed. The average performance of students is close to the national average and no significant decrease was observed.

# Task 2: Diamonds Dataset Analysis

## Introduction

This report aims to analyze the diamond dataset. The analysis involves creating a new binary variable, performing hypothesis testing to determine whether there is a relationship between diamond clarity and color, and calculating 90% confidence intervals for the average prices of colorless and non-colorless diamonds.

## Dataset Introduction

This dataset includes various characteristics of diamonds such as price, carat, clarity and color. The data set contains the following columns:

- price: Price of the diamond (in USD)

- carat: Weight of the diamond (in carats)

- clarity: The clarity of the diamond

- color: Color of the diamond

## Methodology

The following steps were followed to analyze the data:

1. Created a new binary variable called 'colourless' based on the 'colour' variable.

2. Chi-square test of independence was applied to test the relationship between the lightness status of a diamond and whether it is colorless.

3. 90% confidence intervals were calculated for the average prices of colorless and non-colorless diamonds.

4. Analysis results were interpreted and supported with visuals.

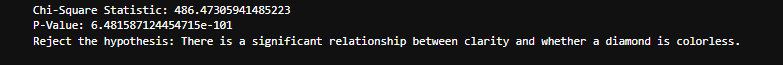
## Summary Statistics

The data set includes various characteristics of diamonds, such as price, carat weight, cut, colour, clarity and size.





## Hypothesis Testing Results



Since the P-Value is much smaller than 0.01, we can say that there is a significant relationship between clarity and whether a diamond is colorless or not.

## Confidence Intervals

- 90% Confidence Interval for Colorless Diamonds: (3301.86, 3373.66)

- 90% Confidence Interval for Non-Colorless Diamonds: (4448.78, 4533.68)

## Visualization

A graph of blue and orange bars

Description automatically generatedA diagram of different colored squares

Description automatically generated

## Results and Discussion



According to the analysis results, there is a significant relationship between the clarity status of a diamond and whether it is colorless or not. Also, the average price of colorless diamonds is lower than that of non-colorless diamonds. These findings show that the clarity and color characteristics of diamonds have a significant impact on price.

# Task 3: Plant Growth Analysis Report

## Introduction

This report provides a detailed analysis of the PlantGrowth data set, containing the results of a small study comparing plant yields under control and two different treatment conditions.

## Dataset Introduction

The PlantGrowth dataset consists of three groups: a control group and two treatment groups. The data set contains the following columns:

weight: weight of the plant

group: treatment group (ctrl, trt1, trt2)

## Methodology

The following steps were followed to analyze the data:

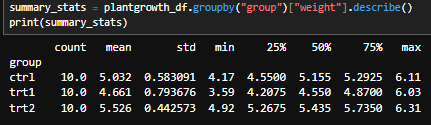
1. Summary statistics were calculated for each group.

2. Visualizations were created to understand the distribution of plant weights in each group.

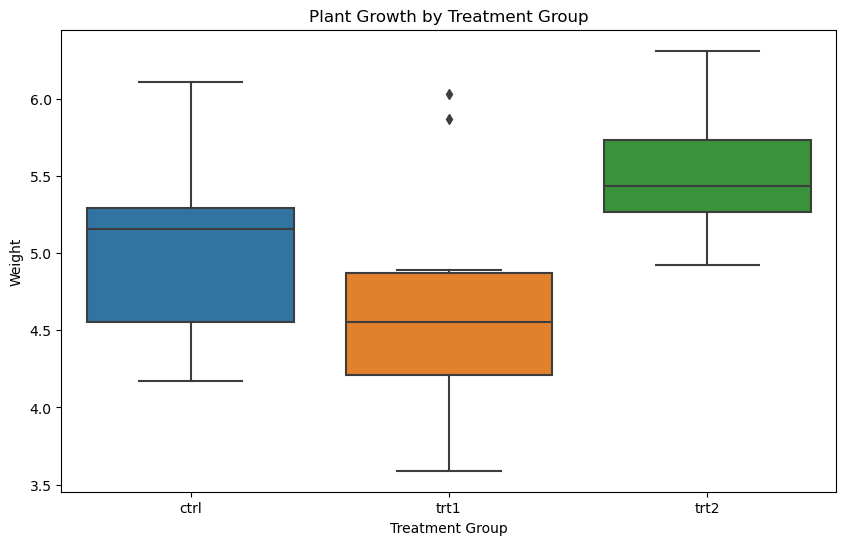
3. ANOVA test was performed to determine whether there were significant differences between groups.

4. If the ANOVA test was significant, Tukey HSD test was performed to determine where the differences were.

## Summary Statistics



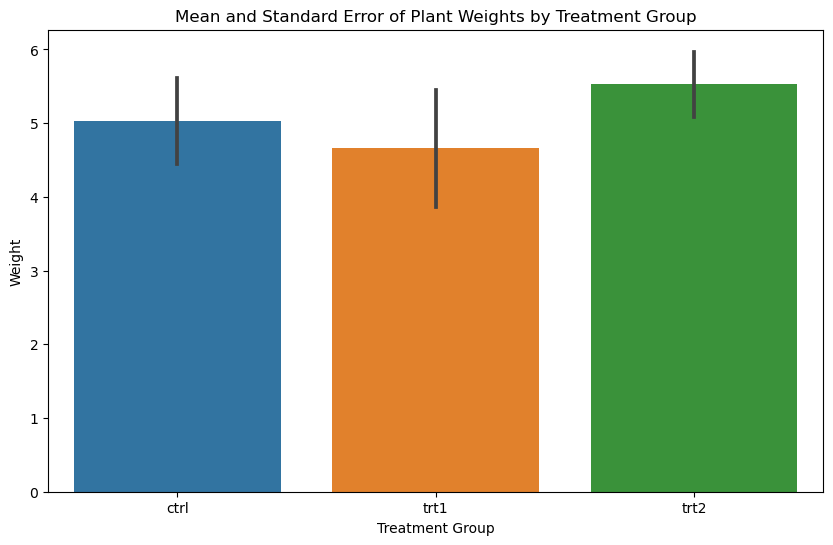
## Box Plot



## Violin Chart



## Bar Chart

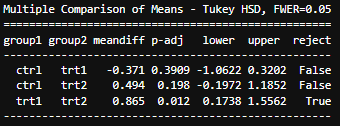


## ANOVA Test Results

ANOVA F-statistic: 4.846087862380136

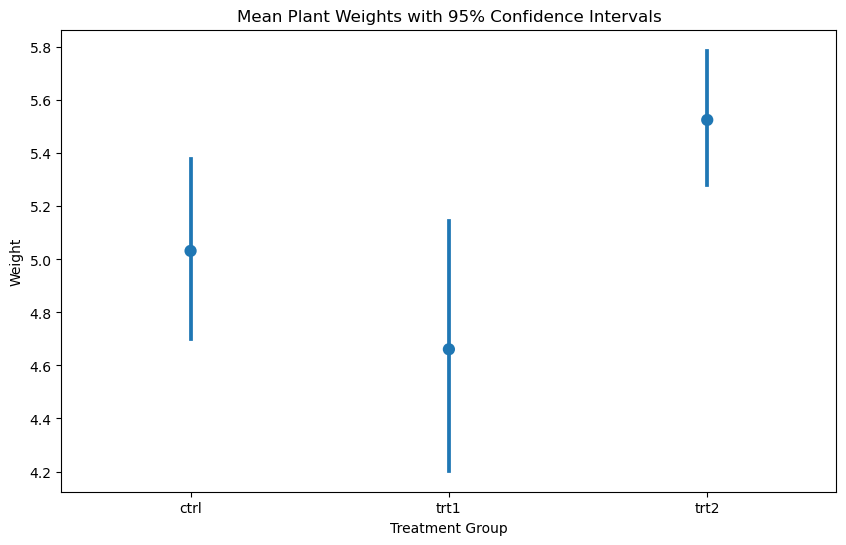
ANOVA p-value: 0.0159099583256229

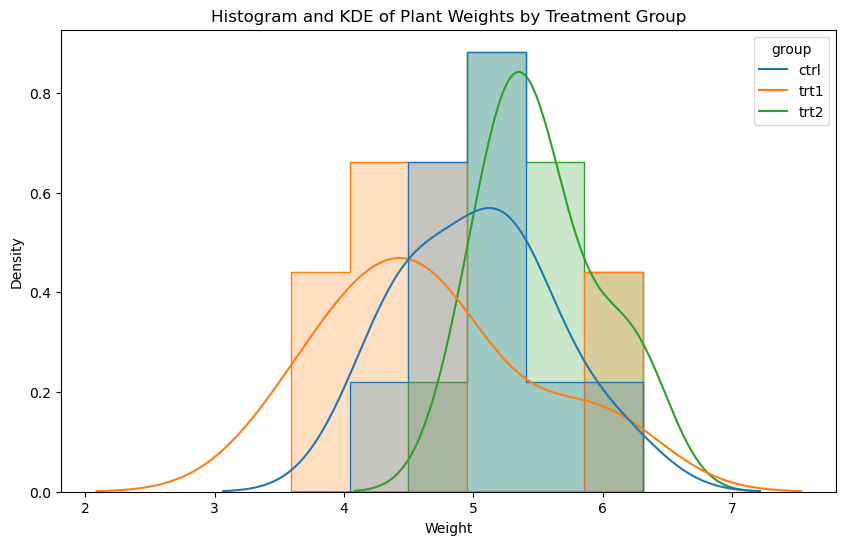
## Tukey HSD Test Results



Mean Plant Weights and 95% Confidence Intervals

## Visualization



Histogram and KDE Graph of Plant Weights According to Treatment Group

## Result and Discussion

The analysis shows significant differences between treatment groups. ANOVA test indicated that the means of the groups were not the same. Tukey HSD test revealed a significant difference between the first treatment group (trt1) and the second treatment group (trt2). There were no significant differences between the control group and any treatment group.

# Task 4: Trees Dataset Analysis

## Introduction

This report was prepared to analyze the relationships between the girth, height and volume of black cherry trees. Regression analysis was performed to determine the best model to use to predict the size of trees.

## Dataset Introduction

This dataset contains diameter, height and volume measurements of 31 felled black cherry trees. The data set contains the following columns:

- Girth: Diameter of the tree (in inches)

- Height: Height of the tree (in feet)

- Volume: Tree volume (in cubic feet)

## Methodology

The following steps were followed to analyze the data:

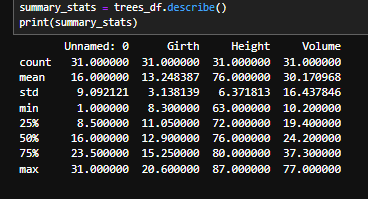
1. Summary statistics were calculated for each variable.

2. Correlations between variables were determined and evaluated with hypothesis tests.

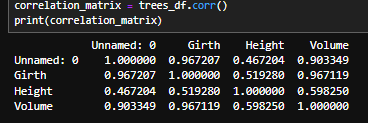
3. Linear regression models were created to predict the volume of trees.

4. It was determined which model gave the best results and the findings were interpreted.

## Summary Statistics



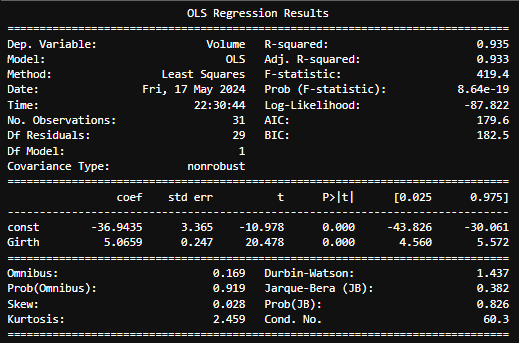
## Correlation Matrix



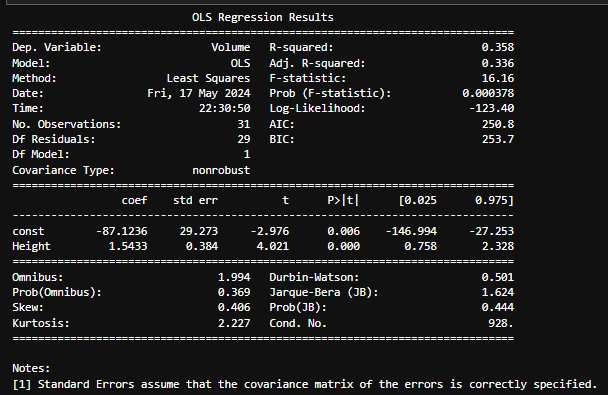
## Correlation Analysis and Hypothesis Tests



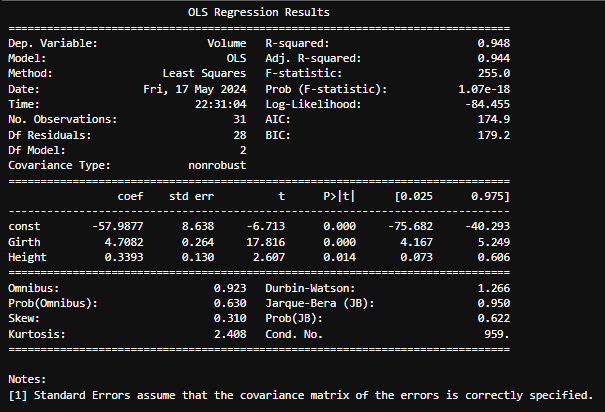
## Regression Models

Girth vs Volume

Height vs Volume

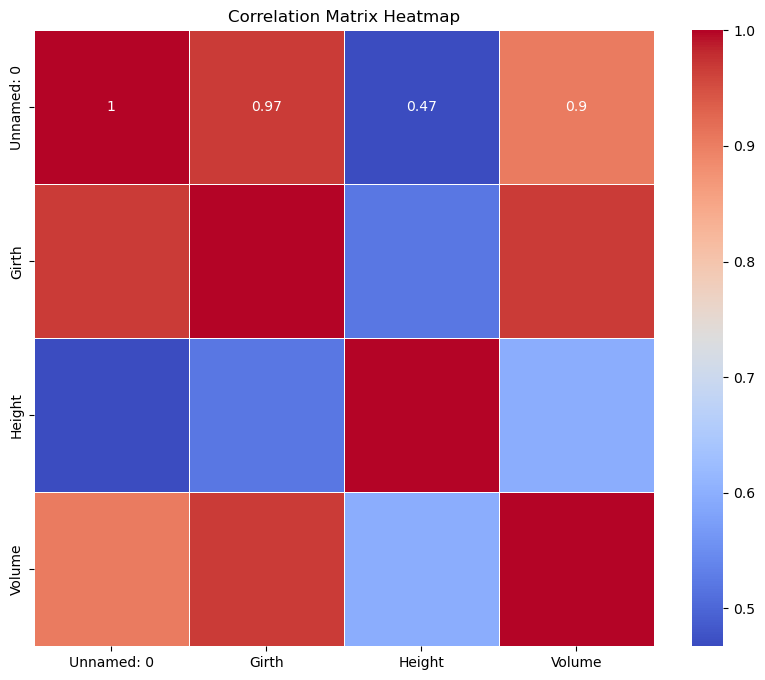


Girth and Height vs Volume

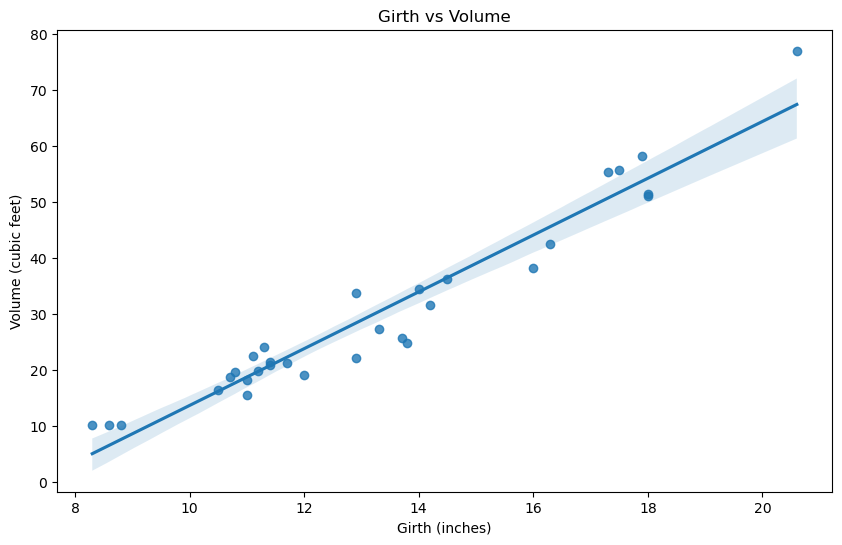


## Visualization

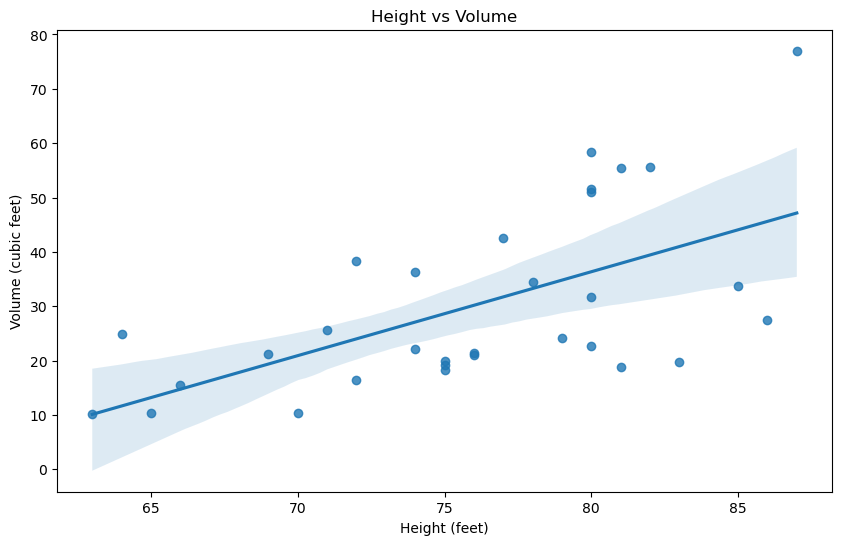
### Correlation Matrix Heat Map



### Relationship Between Girth and Volume



### Relationship Between Height and Volume



## Results and Discussion

According to the analysis results, there is a very strong positive relationship between girth and volume. The relationship between height and volume is weaker. Regression analyzes also support these findings. The best predictive model is a multiple linear regression model using diameter and height together, as this model has the highest R^2 value.

# Github Link

https://github.com/TaylanOzgur96/Statistical\_Techniques\_for\_Data\_Analysis-ca2