Project II

Requirement:

- Each group selects one topic and follows the instructions.
- In each report, there must student names and IDs on the cover page, the table of content, and the questions.
- R-Studio must be used to analyze the data set and the codes must be inside framed environments. Detailed explanations must be provided to receive full credit.

Project II - Topic 1

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Attribute Information:

- *price* - Price of each home sold
- *sqft_living* - Square footage of the apartments interior living space
- *floors* - Number of floors
- *condition* - An index from 1 to 5 on the condition of the apartment,
- *sqft_above* - The square footage of the interior housing space that is above ground level
- *sqft_living15* - The square footage of interior housing living space for the nearest 15 neighbors

Steps:

1. Import data: **house_price.csv**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Transformation
   (b) Descriptive statistics for each of the variables
   (c) Graphs: hist, boxplot, pairs.
4. Fitting linear regression models: We want to explore what factors may affect home prices in King County.
5. Predictions:
   - Case 1: sqft_living15 = mean(sqft_living15), sqft_above = mean(sqft_above), sqft_living = mean(sqft_living), floor = 2, condition = 3
   - Case 2: sqft_living15 = max(sqft_living15), sqft_above = max(sqft_above), sqft_living = max(sqft_living), floor = 2, condition = 3.

Project II - Topic 2

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.

Attribute Information:

- *sex* - student's sex (binary: 'F' - female or 'M' - male)
- *age* - student's age (numeric: from 15 to 22)
- *studytime* - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- *failures* - number of past class failures (numeric: n if 1<=n<3, else 4)
- *higher* - wants to take higher education (binary: yes or no)
- *absences* - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:

- *G1* - first period grade (numeric: from 0 to 20)
- *G2* - second period grade (numeric: from 0 to 20)
- *G3* - final grade (numeric: from 0 to 20, output target)

Steps:

1. Import data: **grade.csv**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Transformation
   (b) Descriptive statistics for each of the variables
   (c) Graphs: hist, boxplot, pairs.
4. Fitting linear regression models: We want to explore what factors may affect the final grade.
5. Predictions:

Project II - Topic 3

This data set contains information on 78 people using one of three diets (The University of Sheffield).

Attribute Information:

- *Person*: Participant - number
- *gender*: Gender (1 = male, 0 = female) - Binary
- *Age*: Age (years) - Scale
- *Height*: Height (cm) - Scale
- *preweight*: Weight before the diet (kg) - Scale
- *Diet*: Diet - Binary
- *weight10weeks*: Weight after 10 weeks (kg) - Scale
- *weightLOST*: Weight lost after 10 weeks (kg) - Scale

Steps:

1. Import data: **Diet.csv**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Descriptive statistics for each of the variables
   (b) Graphs: boxplot.
4. t.test: between *pre.weight* and *weight6weeks*
5. One way ANOVA: What is the best diet for weight loss?
6. Two way ANOVA: How do *Diet* and *gender* affect *weightLOST*?

Project II - Topic 4

This data set contains information about all flights that departed from the two major airports of the Pacific Northwest (PNW), SEA in Seattle and PDX in Portland, in 2014: 162049 flights in total.

Attribute Information:

- *year*, *month*, *day*: date of departure.
- *carrier*: carrier
- *origin*: departure airport
- *dest*: destination airport
- *dep_time*: estimated time departure
- *arr_time*: estimated arrival departure
- *dep_delay*: departure delay
- *arr_delay*: arrival delay
- *distance*: distance between two airports (in miles)

Steps:

1. Import data: **flights.rda**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Descriptive statistics for each of the variables
   (b) Graphs: boxplot - *dep_delay* for each *carrier*. Remove outliers.
4. One way ANOVA: Is there a difference in average delayed departure times among airlines for flights departing from Portland in 2014?
5. Generalize linear model: How do *dep_delay* and *carrier* affect *arr_delay*?

Project II - Topic 5

Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement.

Attribute Information:

- *weigth*: a numeric variable giving the chick weight.
- *feed*: a factor giving the feed type.

Steps:

1. Import data: **chicken_feed.csv**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Descriptive statistics for each of the variables
   (b) Graphs: boxplot - *weight* for each *feed* type.
4. One way ANOVA.
5. Multiple comparison.
6. Kruskal- Wallis test.