Project II- MT2013

Thi Huong PHAN

September 22, 2021

Requirements:

- Each group works on our assigned topic and follows the instructions.
- The report has a maximum of 30 pages.
- The table of content, the questions, the references must be included in the report.
- R/R-Studio must be used to analyze the data and the codes must be inside framed environments. Detailed explanations must be provided to receive full credit.

Bonuses:

- Students can use extended models which are not provided in the course.
- Students can show their points of view to give significant comments in your report.
- Students use novel clinical/experimental datasets which are closely relative to their majors.

Project II - Topic 1

Activity 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 (if it is necessary).
   (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.

Activity 2:

You must find a dataset subjected to your studying specification and then analyze the data. You are encouraged to be active in this activity that is

- to be free in finding the data, including sources from your collected experiments, from the Internet, or the reference source.
- to be free in using theoretical methods for the analysis.

Project II - Topic 2

Activity 1

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 (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 $1 \leq n < 3$, 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 (if it is necessary)
   (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

Activity 2:

You must find a dataset subjected to your studying specification and then analyze the data. You are encouraged to be active in this activity that is

- to be free in finding the data, including sources from your collected experiments, from the Internet, or the reference source.
- to be free in using theoretical methods for the analysis.

Project II - Topic 3

Activity 1

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
- *weight6weeks*: Weight after 6 weeks (kg) - Scale
- *weightLOST*: Weight lost after 6 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*?

Activity 2:

You must find a dataset subjected to your studying specification and then analyze the data. You are encouraged to be active in this activity that is

- to be free in finding the data, including sources from your collected experiments, from the Internet, or the reference source.
- to be free in using theoretical methods for the analysis.

Project II - Topic 4

Activity 1

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: Use suitable regression models to explore significant factors which affect the arrival delay.

Activity 2:

You must find a dataset subjected to your studying specification and then analyze the data. You are encouraged to be active in this activity that is

- to be free in finding the data, including sources from your collected experiments, from the Internet, or the reference source.
- to be free in using theoretical methods for the analysis.

Project II - Topic 5

Activity 1

The dataset was collected to assess the heating load and cooling load requirements of buildings (that is, energy efficiency) as a function of building parameters. The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

Specifically:

- *X1* - Relative Compactness
- *X2* - Surface Area
- *X3* - Wall Area
- *X4* - Roof Area
- *X5* - Overall Height
- *X6* - Orientation
- *X7* - Glazing Area
- *X8* - Glazing Area Distribution
- *y1* - Heating Load
- *y2* - Cooling Load

Steps:

1. Import data: **heat_data.csv**
2. Data cleaning: NA (Not available)
3. Data visualization
   (a) Transformation (if it is necessary).
   (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 Heating Load.
5. Propose a suitable test to compare the average Heating Load and the average Cooling Load.

Activity 2:

You must find a dataset subjected to your studying specification and then analyze the data. You are encouraged to be active in this activity that is

- to be free in finding the data, including sources from your collected experiments, from the Internet, or the reference source.
- to be free in using theoretical methods for the analysis.