ISCG 8025 Introduction to Data Science
Assignment B

Due: 12 November 2017

Introduction

This is Part B of the ISCG 8025 course assignment and is worth 30% of your final grade (total Mark is 30). For this assignment, you need to create a R Script file that interacts with the dataset accompanying this assignment. The file containing the data can be downloaded from Moodle "`data set 6.csv`". The data set showing the monthly power consumption of 500 residential houses contains the following variables.

1. "`Area`": The area of the house in sqm
2. "`City`": The city at which the house is placed
3. "`P.Winter`": The average monthly power consumptions of the house in winter in kW.h
4. "`P.Summer`": The average monthly power consumptions of the house in summer in kW.h

Part 1 – Data Cleaning and Transformation

A) Write a data cleaning function that makes the data set ready for further analysis. This function may perform various data cleaning tasks including but not limited to
   - Correcting possible typos
   - Removing irrelevant data (only houses in Auckland and Wellington are considered)
   - Removing outliers, e.g. negative area, negative power consumptions, very high areas, very high power consumptions

   Note: You should not clean the data set manually. All the data cleaning tasks should be carried out by the data cleaning function automatically.

B) Write a function that calculates the annual average power consumption given "`P.Winter`" and "`P.Summer`". (you just need to add "`P.Winter`" and "`P.Summer`" and divide the result by two). By using this function, create a new variable named "`P.Annual`" and add it to the dataset.

Part 2 – Univariate Analysis

A) Write R codes that calculate the mean and standard deviation of the annual, winter and summer power consumption. Show the results in your report by using a table.

B) Write R codes that plots the density function of the annual, winter and summer power consumption. Use appropriate labels for the plots. Use same scale for the plots. Add the plots to your report.

C) Write R codes that creates the boxplots for the annual, winter and summer power consumption. Use appropriate labels for the plots. Use same scale for the plots. Add the plots to your report.

D) Write R codes that divide the data set into two subsets based on the values of "`City`" variable.

E) Write R codes that repeat tasks A, B, C for the two subsets.

F) Compare the results obtained from the above tasks and make comments on the power consumptions of Auckland and Wellington residential houses during winter and summer.

Part 3 – Bivariate Analysis

A) Write R codes that create a scatterplot from "P.Annual" and "Area" variables. Use appropriate labels for the plots. Use same scale for the plots. Add the plots to your report.
B) Write R codes that calculate a linear regression model for "P.Annual" and "Area" variables. Show the linear model in the scatterplot. Calculate the MSE (mean square error) of the .
C) Write R codes that calculate a second order polynomial regression model for "P.Annual" and "Area" variables. Show the linear model in the scatterplot. Calculate the MSE of the model.
D) Write R codes that calculate a third order polynomial regression model for "P.Annual" and "Area" variables. Show the linear model in the scatterplot. Calculate the MSE of the model.
E) Make comments on the three MSE values obtained in the previous tasks. Which regression model has the highest accuracy?
F) Repeat tasks A-E for "P.Winter" and "P.Summer".
G) Repeat task A-F for Auckland and Wellington sub data sets.

Marking scheme

This assignment will be marked according to the following table.

	Results	Interview	Total
Part 1	2	3	5
Part 2	3	7	10
Part 3	5	10	15
			30

- Marks allocated to "Results" will be awarded if the R Script file results in correct operations, calculations, and plots.
- Marks allocated to "Interview" will be awarded if the student is able to describe the operation of the R Script correctly and to modify the codes as requested by the tutor.