AS404 Data Intergrity Management
and
Data Analysis

Assignment 1

2020 - 2021

Instructions

1. Create a folder and name it **AS404_SXX_XXX** by replacing **SXX_XXX** with your registration number.
2. Create an R project in the **AS404_SXX_XXX** folder you have created and name it - **AS404_A1**.
3. Download the crimeData.csv file from Google Classroom.
4. Create an R-markdown file and answer all the questions (a) - (h).
5. Knit your R-markdown file as a pdf and name the PDF file as **AS404_A1_SXX_XXX.pdf** by replacing **SXX_XXX** with your registration number.
6. Upload the pdf and R project to Google Classroom as a ZIP file.

Question

1. The crimeData data set contains 47 rows and 14 columns that describe the crime rate and some other relevant variables in different states of the US. The dependent variable "crime_rate" describes the average number of offences per million population in each state. Description of other variables is as follows;

    i Youth - number of young males aged 18-24 per 1000

    ii Southern - whether the state is a Southern or non-southern

    iii Education - average number of years schooling up to 25

    iv ExpenditureYear0- per capita expenditure on police the succeeding year

    v LabourForce- males employed aged 18-24 per 1000

    vi Males - number of males per 1000 females

vii MoreMales- whether more males are identified per 1000 females

viii StateSize- state size in hundred thousand

ix YouthUnemployment- Number of unemployed males aged 18-24 per 1000

x MatureUnemployment- number of unemployed males aged 35-39 per 1000

xi HighYouthUnemploy- whether there exist higher youth unemployment ($highifyouth\_unemployment > 3 \times mature\_unemployment$)

xii Wage- median weekly wage

xiii BelowWage- number of families below half of the median weekly wage

(a) Read the dataset into R, define the column types accordingly and rename the column names in 'snake case' format.

(b) Use an appropriate graph to identify the shape of the distribution of crime_rate

(c) Calculate the average crime rates for southern and non-southern states and discuss your findings.

(d) Using a suitable graph, compare the crime rates in southern and non-southern states and interpret the graph.

(e) Compare the crime rates according to the following variables using a single graph.

1. MoreMales
2. HighYouthUnemploy

(f) Visualise and compare the youth unemployment of the states based on gender composition.

(g) Identify the relationship between the crime rate and median weekly wage using an appropriate graph.

(h) Fit a simple linear regression model to describe the identified relationship between the crime rate and median weekly wage in part (g) and interpret the fitted model.