EDA project: 80% of the overall grade

EDA project: 80% of the overall grade

In this assignment, you are required to design and implement an exploratory data analysis (EDA) project.

Datasets

You can choose one of three datasets for analysis:

(1) Loan applications: loanapp.csv . The dataset contains data on loan applications to a bank, including various types of information on the applicant and the purpose of the loan, along with the eventual loan decision (approve or reject - see the column loan_decision). A detailed description of the columns can be found in loanapp_desc.txt

(2) Major League Baseball: mlb.csv . Data on salaries and other information (such as race, position and performance information) on baseball players in MLB in 1993. A detailed description of the columns can be found in mlb_desc.txt

(3) Wages: wage.csv . Data on employees, such as their hourly wage, gender, race, marital status, etc. A detailed description of the columns can be found in wage_desc.txt .

The source for all the three datasets is the companion page for Woolridge J. (2013). Introductory Econometrics: A Modern Approach.

Format of submission

You are free to choose a programming language to use for the assignment- either Python or R. You can also use Excel, however, do note that many types of the expected analysis are not directly available in Excel and you will need to do quite a lot of manual manipulation of the data if you do use Excel. You can prepare your project in the form of a Jupyter notebook (Python), RMarkdown notebook (R) or an Excel file. With any of the three options, the final submission should be converted to a PDF file (when using Excel you may consider inserting relevant tables and plots into a MS Word document and before converting it to PDF).

The pages in the file should be in the portrait orientation.

The word limit is 1000 words +/-20%, excluding the code and any code output. To count the number of words in Markdown cells in a Jupyter notebook, you may find it useful to use this code: Word count in Jupyter.ipynb . To count the number of words in Markdown in an Rmd notebook, please see this code: Word count in Rmd.Rmd .

Steps of the EDA

The EDA project will need to include the following steps, presented in this sequence:

1. Display descriptive statistics on the dataset.

2. Check if any records in the data have any missing values; handle the missing data as appropriate.

3. Build a graph visualizing the distribution of one or more individual continuous variables of the dataset

4. Build a graph visualizing the relationship in a pair of continuous variables. Determine the correlation between them.

5. Display unique values of a categorical variable.

6. Build a contingency table of two potentially related categorical variables. Conduct a statistical test of the independence between them.

7. Retrieve one or more subset of rows based on two or more criteria and present descriptive statistics on the subset(s).

8. Conduct a statistical test of the significance of the difference between the means of two subsets of the data.

9. Create one or more tables that group the data by a certain categorical variable and displays summarized information for each group (e.g. the mean or sum within the group).

10. Implement a linear regression model and interpret its output.

Each step of the analysis should be documented with comments, describing what the step is meant to achieve, and interpreting the result of the step. If the result of the step is a graph, interpret the graph. The project should conclude with a conclusion section summarizing key findings.

Before you start to work on this assignment, please familiarise yourself with the detailed evaluation criteria for this assignment by studying the assessment brief (see above).