Herald College

Concepts and Technologies of AI

5CS037

Assignment-1

Statisitical Interpretation and Exploratory Data Analysis.

December 9, 2023

Contents

1 Assignment Details and Submission Guidelines

1.1 Assignment Details:

Due	Marks	Submission
December-19.	10	2-4 page report see details below

1.2 Plagiarism and AI Generated Content

Plagiarism of more than 20% and any AI generated content found in the report will be reported for academic misconduct. Thus we highly encourage you to submit your original work.

1.3 Submission Guidelines:

In this assignment you will work in Group you formed during your first assignment.

1. **In Group:**
   (a) You are allowed to write code in group but should submit individual report. Comments can not be same for two or more members of the group. For more details on report and report style please read report and overview section.
   (b) You are to make a 5 mins. presentation regarding your work.(One presentation per group; every member of the group must present something.)
2. **Individual:**
   - There would be small viva after your presentation, where individual be asked question.
   - You are supposed to submit a report in .pdf format in CANVAS.

The Final Date for submission is: Dec 19.

1.3.1 Naming Conventions:

You are supposed to follow naming conventions strictly any file not following the naming conventions will be marked "0".

```
File Name: WLVID_FullName(firstname+last).ipynb
Example: 00000_ABC Sharma.ipynb
```

1.3.2 What and Where to Submit?

- **What to submit?**
  You are expected to submit 2-4 page report based on the task and exercise asked from you along with the code base.
  Please check the section **Report Guidelines** for detailed guidelines on how to write a report.
  You are also expected to make a small presentation in group based the work you all did.
- **Where to Submit?**
  Designated Portal Opened at Canvas or as instructed by your instructor.

2 Assignment Overview

2.1 About Assignment:

In this assignment, you will perform a statistical interpretation and exploratory data analysis for a small dataset and provide a rigorous rationale for your choices. We will determine scores by judging both the soundness of your **design**, the quality of the **write-up(report)** and your ability to answer the question during **viva**. Here are examples of aspects that may lead to **point deductions:**

- Use of misleading, unnecessary, or unmotivated graphic elements.
- Missing chart title, axis labels, or data transformation description.
- Missing or incomplete design rationale in write-up.
- Ineffective encodings for your stated goal (e.g., distracting colors, improper data transformation).

Tools and Python Package which can be used for this assignments (listed but not limited to):

1. **Pandas library(pd)**
2. **Numpy library(np)**
3. **Matplotlib library(plt)**
4. **Seaborn library(sns)**

2.2 Learning Outcomes:

Learning outcomes can be following but not limited to:

1. Use Pandas as the primary tool to process structured data in Python with CSV files,
2. Use matplotlib and seaborn library to produce various plots for visualization,
3. Extract various information from a given dataset using statistical and visualizing techniques.

2.3 Data Selection:

1. Please feel free to pick any structured datasets in csv format that matches the task requirements. But please take pre-approval from your respected instructor and Module leader.
2. If you are not sure about which dataset to pick, select one from the options provided.
3. Note: No two groups in one Section can have same datasets.

The best source to find datasets are but not limited to:

1. Kaggle Datasets:
   Kaggle provides a high-quality dataset in different formats that we can easily find and download.
2. UCI Machine Learning Repository:
   This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms.

3 Tasks and Marks Division

For this assignment we will use the recommended workflow for EDA described in slide 5 of Workshop-03.

3.1 Choose; Load and Inspect your Data [2].

- **Pick a Domain and Dataset you are interested in:**
  For this assignment, you will need to find a dataset of your choosing (interest) and load into dataframe object with PANDAS library.

  Perform a initial observation regarding the dataset, while doing that try to answer following question:

  - Detailed description about the dataset:
    1. When and Who created the dataset?
    2. How did you get acess to the dataset?
    3. List out the attributes (columns) of a dataset.
  - Guess some probable question that dataset could answer.
  - Assess the basic fitness of the dataset.
- **Inspect your Data:**
  Before you begin exploring the data, write some code in chunks to preview and summarize the data frame using some of the method's we've used in class. You can consider following questions in your exploration:
  1. What is the total size of the data frame?
  2. What type of data is each variable (numeric, character, logical, date)?
  3. Do any variables have missing values? Why might that be?

3.2 Data Cleaning and Summary Statistics [3]

- **Data Cleaning:** Based on your previous inspection you can explore following options but are not limited to clean your data:
  1. If found any missing values use appropriate data imputations techniques discussed in your workshops to fill the missing values. Please justify why did you pick particular techniques.
  2. If found any duplicates remove those observations.
  3. Perform more data cleaning as you think is required for your particular dataset.Justify your action.
- **Summary Statistics:** The next step in EDA is to examine the measures of variability and central tendency for the variables in the data set. For this step you can do the following:
  Split the Dataset into numeric column and Categorical Column.
  - **For Numeric Column:**
    1. Use descriptive function to gather basic statistical information about the variables such as: min., max., range, mean, median, Variance, standard deviation.
    2. Observe the output and explain if anything interesting note it.
  - **For Categorical Column:**
    1. For character variables, what are the unique values in the variable?
    2. Figure out the mode of such variables.
    3. If any date variables, what time period do the observations in these data frames span.

3.3 Visualize the Data: Make Interpret and Save your Charts [5].

Now that you have a basic understanding of the dataset, make some charts/figures to explore the variables in the data and their potential relationships. You may use base pandas plotting functions or matplotlib or seaborn or any package to make your figures, but you must make at least figures of at least two different variables for each category below:

- **Univariate Analysis: [3]** Pick any of the two variables of your choice and perform the following:
  - **Make Chart:**For your selected two variables:
    1. Make Line Chart or Bar chart or Pie chart, whichever best fits your dataset.
    2. For your selected two variables plot a figure to analyze Skewness and Modality of Distributions present in dataset.
    3. For your selected two variables plot a figure to analyze Outliers present in dataset.
  - **Interpret a Chart:** For all the figure write a short description and interpretation of your chart. While doing that make sure you address at least the following questions:
    1. Describe what variables you are plotting and why.
    2. Describe the primary relationship or information reader will gain from your visualization.

- **Bi-variate Analysis: [2]**Split the dataset into numeric and categorical column, store the Numeric Column in different dataframe and for that dataframe:
  - **Make Chart:**
    1. Explore correlation visualization with scatter plot between the variables.

       {Hint: Explore the `seborn pairplot` function.}
    2. Explore the correlation visualization with the `heatmap`.
  - **Interpret Chart:** For all the figure write a short description and interpretation of your chart. While doing that make sure you address at least the following questions:
    1. Describe what variables you are plotting and why.
    2. Describe the primary relationship or information reader will gain from your visualization.
- **Save Chart:** All the chart/figure made above in Univariate and Bi-variate Analysis must be save in ".png" format and use in the report.

4 Conclusion and Reflect

Write a short reflection on what you've learned and any questions or points of confusion you have about what we've covered thus far. This can just few a few sentences related to this assignment or the contents covered thus far.