Bachelor in AI

Module
PG2104 Data Collection and Analysis

Due date for submission
(see Wiseflow)

Module leader and e-mail
Noha El-Ganainy |Noha.El-Ganainy@kristiania.no

Teacher and e-mail
Arvind Keprate| arvindke@oslomet.no

Learning outcomes

After successfully completing the course the student:

Knowledge

- can understand key principles of Python programming.
- can display knowledge of largely used Python libraries for data science, like NumPy, pandas, Pyplot and Scikit-learn.
- can understand how to effectively collect data and clean it.
- can understand and assess how data influences algorithms, and being able to counter these influences with different methods or changes in data collection.

Skills

- can exercise good programming skills in Python.
- can select and apply an efficient method for data collection.
- can select and apply efficient methods for data preparation and cleaning.

General competence

- can discuss theoretical and practical aspects of data collection.
- can discuss methods and theoretical approaches for data cleaning.
- can discuss and provide knowledge about which data to collect and how to collect it for new data science problems.

Assignment specification

1. Group Size=Only 1 (Individual Submission)
2. Please use comments and/or markdown cell wherever necessary explanation is required. Additional marks will be given for clean Jupyter notebook and understandable code.
3. Referencing: Wherever necessary use any acceptable academic style.
4. Submit the Jupyter Notebook in .ipynb and pdf format.

Please address the following questions in your submission.

Problem 1: Conceptual Questions (10 points)

Use Markdown cell in your Jupyter Notebook and explain the following:

1. What are escape sequences? Explain any 3 escape sequences.? **(2 points)**
2. What are data types? What are Python's built-in core data types. **(2 points)**
3. What do you mean by positive correlation and negative correlation? Give one example of each. **(2 points)**
4. What do you mean by five-point summary of data? Which plot is used to reflect five-point summary of data? **(2 points)**
5. What is difference between univariate, bivariate and multivariate analysis? Give one example each of the plots used for these analysis. **(2 points)**

Problem 2: Programming Problem (20 points)

1. Write a Python program to print the following pattern **(10 points):**

```
#######
#     #
#     #
#     #
#     #
#     #
#######
```

2. Write a Python program to input marks in 3 subjects; compute average and then calculate grade as per following guidelines **(10 points):**

Grade	Marks
A	Greater than or equal to 90
B	89 to 80
C	79 to 60
D	59 to 50
E	49 to 40
F	Less than or equal to 39

Problem 3: Data Wrangling and Plotting Problem (50 points)

For this exercise, we will use Automobile Dataset which has been provided to you in the file **auto_data.csv.** Perform the following tasks:

1. Read the csv file in Pandas and create a DataFrame named Auto_df. What is the shape of Auto_df**.** Print first 7 and last 7 rows of Auto_df**. (2 points)**
2. Do you find anything unusual in the dataset? If yes, then replace these unusual values with NaN values. **(3 points)**
3. Which columns have NaN values and how many? Finally, handle these null values using any strategy taught in the class. **(5 points)**
4. Name the columns that have datatype as "**object**". **(2 points)**
5. What is the mean and median value for the length and width column? **(3 points)**
6. The fuel consumption of the vehicles on the highway is given in miles per gallon (mpg) in the column "highway-mpg". Transform the mpg to L/100km (using formula L/100km = 235/mpg) in the column of "highway-mpg" and change the name of column to "highway-L/100km". **(5 points)**

7. Make a box plot for the curb-weight showing the 5-point summary. **(2 points)**
8. For the column “horsepower”, discretize the data into 5 bins and draw a histogram. **(3 points)**
9. Find the Spearman correlation between the following columns: bore, stroke, compression-ratio and horsepower. **(2 points)**
10. Given the correlation results between "price" and "stroke", do you expect a linear relationship? Verify your results using the function "regplot()" **(3 points)**
11. Make a box plot between “body-style” and “price”. Which body-style has the maximum outliers in price? **(5 points)**
12. Into how many categories can vehicles be categorized based on drive-wheels? How many vehicles are four-wheel drive (i.e. 4wd)? **(3 points)**
13. How many vehicles have rear engine location? **(2 points)**
14. What is the average price of the forward wheel drive (fwd) and rear wheel drive (rwd) category of vehicles? **(5 points)**
15. What is the average price of rwd-sedan and 4wd-hatchback? Which one is larger? **(5 points)**

Problem 4: Web Scraping Problem (20 points)

In this task you will webscrap the following website https://en.wikipedia.org/wiki/World_population. Perform the following tasks:

1. Scrap the table titled: “10 most populous countries” shown below and create a DataFrame named **DF_Pop (15 points)**.

10 most populous countries

Rank	Country / Dependency	Population	Percentage of the world	Date	Source (official or from the United Nations)
1	China	1,412,600,000	17.7%	31 Dec 2021	National annual estimate[94]
2	India	1,373,761,000	17.2%	1 Mar 2022	Annual national estimate[95]
3	United States	333,367,536	4.17%	25 Nov 2022	National population clock[96]
4	Indonesia	275,773,800	3.45%	1 Jul 2022	National annual estimate[97]
5	Pakistan	229,488,994	2.87%	1 Jul 2022	UN projection[98]
6	Nigeria	216,746,934	2.71%	1 Jul 2022	UN projection[98]
7	Brazil	215,436,807	2.69%	25 Nov 2022	National population clock[99]
8	Bangladesh	168,220,000	2.10%	1 Jul 2020	Annual Population Estimate[100]
9	Russia	147,190,000	1.84%	1 Oct 2021	2021 preliminary census results[101]
10	Mexico	128,271,248	1.60%	31 Mar 2022	National quarterly estimate[102]

2. Create a Bar Chart for the **DF_Pop**, depicting the country name on X-axis and population on the y-axis. **(5 points)**

Assignment criteria*

Grade	Learning Outcome 1: Knowledge	Learning Outcome 2: Skills	Learning Outcome 3: Competence
A Excellent	Excellent and comprehensive understanding of concepts	Demonstrates excellent analytical, technical and writing skills	Outstanding degree of judgment and independent critical thinking
B Very good	Very good understanding of concepts	Demonstrates very good analytical, technical and writing skills	Sound degree of judgment and independent critical thinking
C Good	Good understanding of theory in most important areas	Demonstrates good analytical, technical and writing skills	Reasonable degree of judgment and independent critical thinking
D Satisfactory	Satisfactory understanding of theory, but with significant shortcomings	Demonstrates limited analytical, technical and writing skills	Limited degree of judgment and independent critical thinking
E Sufficient	Meets the minimum understanding of concepts	Demonstrates sufficient analytical, technical and writing skills	Very limited degree of judgment and independent critical thinking
F Fail	Fail to meet the minimum academic criteria.	No demonstration of analytical, technical and writing skills	Absence of judgment and independent critical thinking

*Adapted from The Norwegian Association of Higher Education Institutions