SCHOOL OF COMPUTING (SOC)

CA2 Specification

DIPLOMA IN APPLIED AI & ANALYTICS

ST1510
Programming for Data Analytics

2023/2024 Semester 2

Assignment rubrics

1. Demonstrate basic competency in writing Python programs.
2. Demonstrate basic competency in using the Pandas and Statsmodel packages for data analysis and data visualization.
3. Demonstrate basic competency in applying the insights gained from the outputs of your Python programs to deliver a useful data analysis presentation.

Section 1
Instructions and Guidelines

1. This is an **INDIVIDUAL** assignment which requires the student to write Python code that retrieves data from data files and perform basic data manipulation operations such as cleansing, transformation and visualization on the data.

2. The requirements of this assignment are outlined in Section 2 of this document.

3. The deadline of this assignment is on Week 17, **5 Feb (Mon) 2024 on 2359Hr**.

4. Submissions must be made via the **BrightSpace CA2 Assignment Submission link** by the stated deadline.

5. Deliverable should be a zip file with the following file-naming convention.
   "PDASCA2_YourClass-YourStudentID-YourName.zip"
   e.g. "PDASCA2_1B04-2388888-StevenLee.zip"

6. The Zip file should include the following items:
   - One or more Jupyter Notebook (.ipynb) files that accomplishes the given tasks using the Python
   - A set of PowerPoint slides that summarize the data insights that you have gained through your Python code
   - All datasets used in your Jupter Notebook files
   - One Declaration of Academic Integrity

7. A compulsory presentation/interview will be conducted. During the session, you must present your work using the submitted PowerPoint slides and the Jupyter Notebook. Your module tutor will ask questions related to the submission and ask you to reproduce certain parts of your code during the session.

8. This assignment will account for **40%** of the **module grade**.

9. 50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter. Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the module tutor as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

10. No marks will be awarded, if the work is copied or you have allowed/enabled others to copy your work. Plagiarism is a serious offence, and if you are found to have committed, aided, and/or abetted the offence of plagiarism, disciplinary action will be taken against you.

    Warning: Plagiarism means passing off as one's own the ideas, works, writings, etc., which belong to another person. In accordance with this definition, you are committing plagiarism if you copy the work of another person and turning it in as your own, even if you would have the permission of that person.

Section 2
Assignment Requirements

1. You must use **at least three** datasets, including **at least one** datasets from data.gov.sg.
   The datasets should be related to either employment or environment. If you choose the environment topic, you can use datasets related to waste management, recycling or pollution.

   Example of datasets related to employment:
   - Graduate Employment Survey - NTU, NUS, SIT, SMU, SUSS & SUTD (2013 to 2021), https://beta.data.gov.sg/collections/415/view
   - Labour Force in Singapore 2022, https://stats.mom.gov.sg/Pages/Labour-Force-In-Singapore-2022.aspx (download the statistics)

   Example of datasets related to environment:
   - Recycling Rate By Waste Type, https://beta.data.gov.sg/collections/1405/view
   - Our World in Data, Per Capita plastic waste vs GDP per capital https://ourworldindata.org/grapher/per-capita-plastic-waste-vs-gdp-per-capita (download the statistics)

   Do NOT reuse more than one dataset from CA1. Do NOT use dataset from training website such as Kaggle.

2. You are encouraged to select interrelated datasets that align with a central theme of investigation. Clearly define an analysis question or problem statement and provide the answer through your data analysis. **Document the URLs of all datasets used** in both the Jupyter Notebook and PowerPoint slides.

3. For each dataset, your task is to write Python code to explain the data and produce useful data visualizations that explain the data.

   Your code should include:
   - Identify and handle missing values
   - Identify and handle outliers
   - Apply regression analysis using Statsmodels
   - Plot at least four charts to explain your observations
     The charts can be produced using visualization packages such as matplotlib, seaborn, pandas and statsmodels.

4. While you can submit more than four charts, keep it within a maximum of nine.

5. Your Python codes should extract meaningful insights from the chosen datasets, crafting an engaging and interesting data analysis.

6. Compile your findings into a deck of PowerPoint slides. The PowerPoint slides should include the following sections and to be presented in 6 minutes.

    - A cover page that lists your name and the title of your data analysis
    - A slide that lists the URLs of all the datasets you have used
    - For each dataset, one slide or more to explain the **process** you went through to analyse that dataset. Where possible, you should specifically mention **how you used the Pandas and Statsmodel** to achieve certain outcome, for example, to identify and handling of missing values and outliners.
    - For each dataset, the **insights** you have gained from analysing the data and any conclusions or recommendations of the analysis.

Section 3
Marking Scheme

Marks will be awarded based on the following rubrics:

Component	Weightage
Basic Requirements • Use of at least three different datasets that met the requirement outlined in Section 2 • At least four informative charts.	25%
Quality of application Application Coverage (20%) ▪ Identify and analysis of Missing Values and Outliners ▪ Regression Analysis using statsmodels) • Code Quality and Clarify (20%) ▪ Demonstration of mastery of the topics ▪ Reusability ▪ Efficiency ▪ Code clarity and documentation	40%
Data analysis and Presentation • Quality of the analysis • Organization and Quality of the presentation and the slides • Question and Answer • Ability to re-produce the submitted code during the presentation/interview **No mark** can be awarded if the PowerPoint slide is not submitted.	35%

Section 4
Sample Output

This section contains sample screenshots of some sample outputs.

Do note that these are simple outputs only, and you are highly encouraged to enhance your own version with more complex features or functionalities than what is shown here.

Example 1
Extracting Pokémon data from online JSON file

This output uses the **requests** library to retrieve the online json file, and uses **pandas** library to convert the json file into dataframe. Similarly, you can also extract other online csv files directly using pandas library, without downloading the csv files.

URL: https://github.com/Biuni/PokemonGO-Pokedex/blob/master/pokedex.json

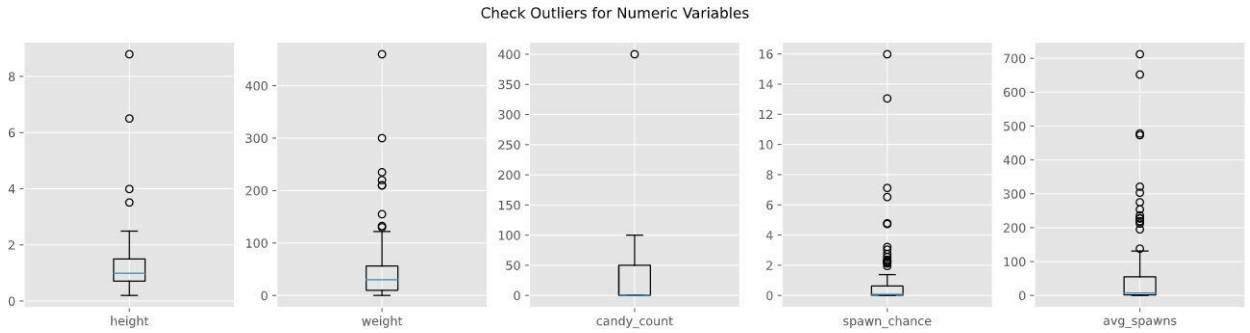
```
# Extract Pokemon gaming data
import requests
import pandas as pd
url = : https://github.com/Biuni/PokemonGO-Pokedex/blob/master/pokedex.json
r = requests.get(url)
df = pd.DataFrame(data=r.json()['pokemon'])
```

	id	name	type	height	weight	candy	candy_count	egg	spawn_chance	avg_spawns	spawn_time	multipliers	weaknesses
0	1	Bulbasaur	[Grass, Poison]	0.71 m	6.9 kg	Bulbasaur Candy	25.0	2 km	0.690	69.0	20:00	[1.58]	[Fire, Ice, Flying, Psychic]
1	2	Ivysaur	[Grass, Poison]	0.99 m	13.0 kg	Bulbasaur Candy	100.0	Not in Eggs	0.042	4.2	07:00	[1.2, 1.6]	[Fire, Ice, Flying, Psychic]
2	3	Venusaur	[Grass, Poison]	2.01 m	100.0 kg	Bulbasaur Candy	NaN	Not in Eggs	0.017	1.7	11:30	None	[Fire, Ice, Flying, Psychic]
3	4	Charmander	[Fire]	0.61 m	8.5 kg	Charmander Candy	25.0	2 km	0.253	25.3	08:45	[1.65]	[Water, Ground, Rock]
4	5	Charmeleon	[Fire]	1.09 m	19.0 kg	Charmander Candy	100.0	Not in Eggs	0.012	1.2	19:00	[1.79]	[Water, Ground, Rock]

Example 2
Data wrangling

It is a common practice to check any missing values, duplicate values or outliers before further processing and analyzing the dataset. You can use pandas function to perform the data wrangling task, or use some simple visualization techniques to do so.

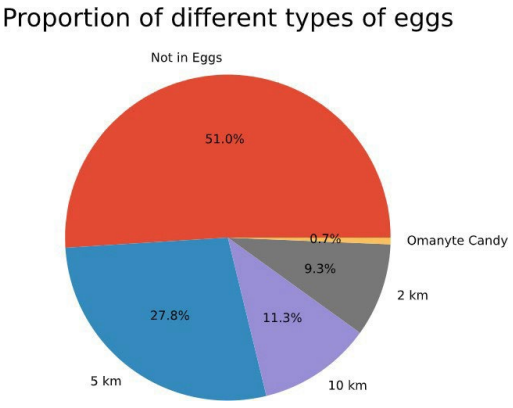
```
Check Missing Values in the Pokemon dataset
--------------------
id                0
name              0
type              0
height            0
weight            0
candy             0
candy_count      81
egg               0
spawn_chance      0
avg_spawns        0
spawn_time        0
multipliers      81
weaknesses        0
dtype: int64
```



Example 3
Extract useful insights for data analysis

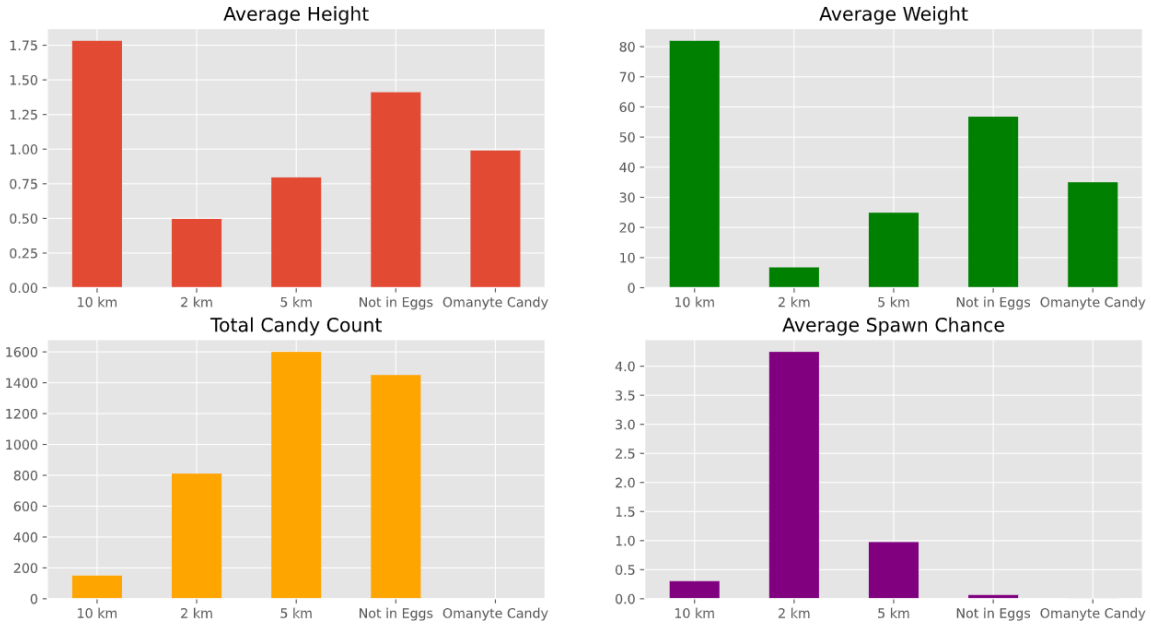
These simple outputs using pandas function to extract some insights, which can be used for data analysis, insight generation and data visualization.

```
Count the number of egg types
------------------------------
Not in Eggs      77
5 km             42
10 km            17
2 km             14
Omanyte Candy     1
Name: egg, dtype: int64
```



In this example, we calculate the number of different egg types in the dataset, and use it to plot a pie chart, in order to show the proportion of different egg types. As we can see from the pie chart, among the 151 Pokémons in the dataset, over half of them are not in eggs, and 27.8% of them belong to the 5km Egg.

	height			weight			candy_count	spawn_chance
	max	mean	min	max	mean	min	sum	mean
egg								
10 km	8.79	1.781765	0.30	460.0	81.941176	3.3	150.0	0.302476
2 km	0.89	0.496429	0.30	20.0	6.735714	1.8	811.0	4.246071
5 km	2.21	0.795952	0.20	115.0	24.895238	0.1	1600.0	0.974829
Not in Eggs	6.50	1.410519	0.30	300.0	56.763636	0.1	1450.0	0.064855
Omanyte Candy	0.99	0.990000	0.99	35.0	35.000000	35.0	0.0	0.006100



In this example, we group the data, summarize and extract the statistics, and visualize those using bar charts. From the statistics, we can obtain the information that 10km Egg Pokémon have bigger size in terms of height and weight, 5km Egg Pokémon have more candies in total, while 2km Egg Pokémon have higher spawn chance on average.

~~ End of Assignment Specifications ~~