SCHOOL OF COMPUTING (SOC)

CA2 Specification

DIPLOMA IN APPLIED AI & ANALYTICS

ST1510
Programming for Data Analytics

2022/2023 Semester-2

Assignment rubrics

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

Table of Contents

Contents

Section 1
Instructions and Guidelines

1. This is an **INDIVIDUAL** assignment which requires the student to write Python code that retrieves data from files (like CSV, text or json) and perform basic data manipulation operations such ascleansing, 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 **Monday 6 Feb 2023 (8am)**.

4. Submissions should 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
   "YourClass-YourStudentID-YourName.zip"
   e.g. "DAAA1B04-2228883-StevenLee.zip"

6. Zip file should include the following items:
   - One or more Jupyter notebooks (.ipynb) or Python source code files (.py) that accomplishes the given tasks using the Python programming language
   - A set of Powerpoint slides that summarizes the extraction, transformation and loading of the datasets you have used.
   - For completeness, you can include the final visualisations that were created from your cleansed datasets.

7. As part of the assignment requirements, you will need to give a short (not more than 10 minutes) presentation / interview to your module tutor using the Powerpoint slides you have prepared. Your module tutor may ask you questions related to the Python code during this interview / presentation session.

8. Subsequent to the submission of your codes and slides, your Module Lecturer will arrange assignment interviews with you separately. Please take note that the dates of the interviews you arrange with your lecturer do not affect our records of the date that you submitted your assignment.

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

10. No marks will be awarded, if the work is copied or you have allowed others to copy your work.

Section 2
Scope of the assignment

In this individual assignment, you are required to produce a data analysis presentation for datasets relating to **the Environment** based on the requirements as stated below.

Basic Requirements

1. You must use **at least three** datasets, including at least one from Data.gov.sg
   https://data.gov.sg/search?groups=environment

2. For each dataset you use, you must write Python code that uses the **Pandas** and/or **NumPy** package to extract useful statistical or summary information about the data. The code should read from multiple datasets, merge the data, identify and handle missing values, identify and remove outliers, if any.

   A sample of the expected output of this requirement is given in Section 4 of this document.

3. For each dataset you use, you will write Python code to produce useful data visualizations that explain the data. Your submission requirement for PDAS is the final cleansed dataset and the python code you have used to extract, transform and load the input datasets to the final datasets used for charting.

   Your code should include:

   - Reading data from multiple sources
   - Identify and handle missing values
   - Identify and handle outliers
   - Writing the final cleansed dataset to a file
   - Use the cleansed data to produce meaningful insights

4. Your Python codes should help you to gain deeper insights into the chosen datasets such that you are able to produce an interesting data analysis on it.

   Compile your findings into a deck of **Powerpoint slides**

   Your Powerpoint slides should include the following sections:

   - 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 briefly explain the **nature of that dataset** (i.e. what is in that dataset) or any peculiarities about it you wish to highlight

- For each dataset, one slide or more to explain the **process** you went through to extract, transform, load and analyse that dataset. Where possible, you should specifically mention how you used the Pandas functions to achieve a certain outcome e.g. to transform the data or identify points of interest (missing or outlier)
- For each dataset, the **insights** you have gained from analysing the data and any conclusions or recommendations you want to make as a result of the analysis.

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.
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. If you are guilty of plagiarism, you may fail all modules in the semester, or even be liable for expulsion.

Section 3
Marking Scheme

Marks will be awarded to each student based on the following rubrics:

Component	Weightage
Assignment requirements are met • Use of at least 3 different datasets, including at least one from data.gov.sg to form environment related insights • Python codes that extract useful insights from the datasets • Presence of code to check and handle data inconsistencies eg missing data, outliers • Python codes that produce useful data visualizations to identify points of interest (missing data, outliers or any special characteristics of the datasets) • A deck of Powerpoint slides that explain the datasets, what was done to process these datasets and summarizes the insights gained from the analysis of the data	50%
Quality of application • Technical complexity • Code quality • User-friendliness (of the graphs) • Aesthetics • Creativity	30%
Data analysis • Completeness in the analysis of data • Quality of analysis and presentation	20%

Section 4
Sample Output expected ✓

This section contains sample screenshots of some sample outputs.

Do not 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

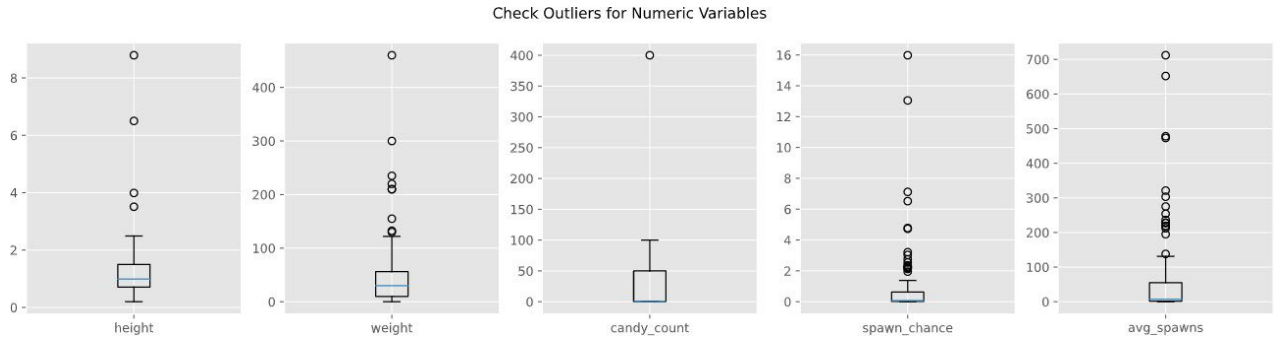
```
# Extract Pokemon gaming data
import requests
import pandas as pd
url = : https://github.com/Biuni/PokemonGO-Pokedex/blob/master/pokedex.json
r = requests.get(url)
df = pd.DataFrame(data=r.json()['pokemon'])
```

	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.

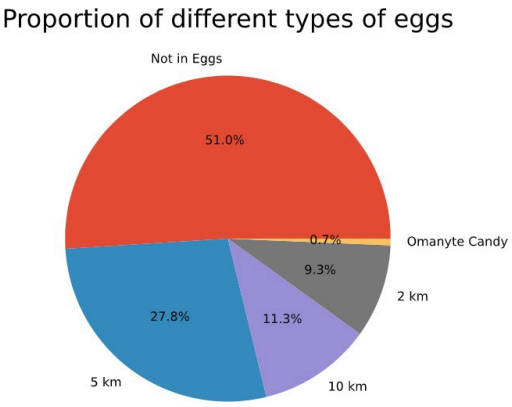
```
Check Missing Values in the Pokemon dataset
--------------------
id                0
name              0
type              0
height            0
weight            0
candy             0
candy_count      81
egg               0
spawn_chance      0
avg_spawns        0
spawn_time        0
multipliers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



Example 3
Extract useful insights for data analysis

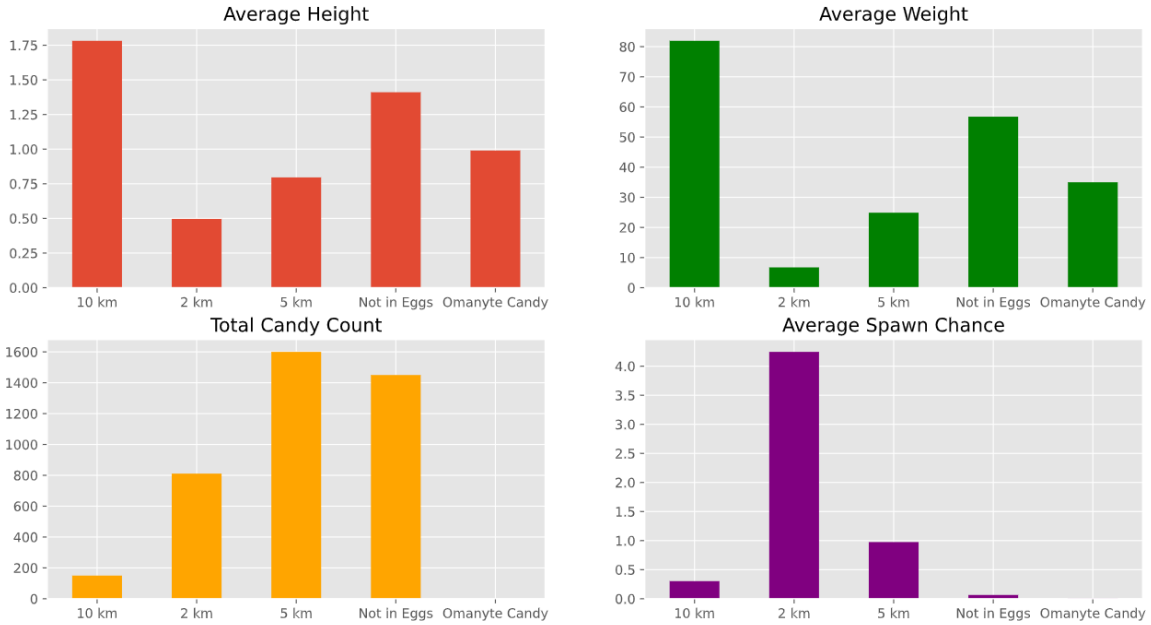
These simple outputs using pandas function to extract some insights, which can be used for data analysis, insight generation and data visualization.

```
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 --