BUSS6002 Assignment 1 2023 S1

Data Science in Business (University of Sydney)

BUSS6002 Assignment 1

Semester 1, 2023

Due Date

- Due: at 12:00pm (noon) on Monday, 17 April 2023 (start of week 8).
- A late penalty of 5% per day applies if you submit your assignment late without a successful special consideration or simple extension. See the Unit Outline for more details.
- You may submit multiple times but only your last submission will be marked.

Instructions

- You must submit a Jupyter Notebook (`.ipynb`) file with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_A1_STUDENTID.ipynb`.
- There is a limit of 1000 words for your submission (excluding code, tables, and captions).
- Do not include any more Python output than necessary and include only concise discussions.
- Each task must be clearly labelled with the corresponding question (and sub-question) number so that the marker can spot your solution easily.
- The submitted `.ipynb` file must be free of any errors, and the results must be reproducible.
- All figures must be appropriately sized (by setting `figsize`) and have readable axis labels and legends (where applicable).
- Use `plt.show()` instead of `plt.savefig('plot.png')` to display each figure.

Rubric

This assignment is worth 20% of the unit's marks. The assessment is designed to test your technical ability and statistical knowledge in performing important basic tasks associated with an exploratory data analysis (or EDA) of a real-world dataset.

Assessment Item	Goal	Marks
Question 1	Overall summary of the dataset	10
Question 2	Univariate analysis	10
Question 3	Multivariate analysis	19
Jupyter Notebook	Logical and clear presentation	1
Total		40

Table 1: Assessment Items and Mark Allocation

Overview

The Australian federal government is building a website to provide households with a tool to determine if installing a rooftop solar panel system is right for them. On the website, users will be able to enter information about their house and the website will provide an estimate of the possible solar power generation.

The government collected a random sample of existing households with solar panels, including information about the households, solar panel installation and the associated solar power generation. The generation from the solar panels was collected from 1/1/2022 to 31/12/2022.

The data has been assembled from a multiple sources including the customer energy retailer, energy distributors and solar installers. Sampling is limited to installations with:

- a single solar panel array or multiple arrays that are oriented identically,
- rooftop installation only.

As a data-scientist-in-training, you will assist the project by completing a variety of EDA tasks.

Data Files

The following files are available on Canvas.

File	Description
SolarSurvey.csv	Data file with 3000 observations.
DataDictionary.txt	Data dictionary containing description of each variable
BUSS6002 A1 STUDENTID.ipynb	A Jupyter Notebook template for getting you started

Table 2: Files Provided

Question 1 (10 Marks)

a) (4 marks) Write some code to automatically print out the column names of the variables with missing values, as well as the number of missing observations associated with each of those variables. The output should be sorted by the number of missing observations from most to least.

b) (4 marks) Write some code to cross-check the data against the data dictionary and identify discrepancies.

c) (2 marks) Briefly discuss your findings from a) and b).

Question 2 (10 Marks)

a) (4 marks) Graphically summarise the distributions of the variables `Generation` and `Panel_Capacity`, one at a time, and briefly discuss the distributional characteristics of the two variables. Your discussion should also connect the distributional characteristics to the domain-specific context of these variables.

b) (2 marks) Graphically summarise the distribution of the variable `Roof_Azimuth` and briefly discuss the distributional characteristics of the variable. Your discussion should also connect the distributional characteristics to the domain-specific context.

c) (3 marks) Given that there may be an association with `Roof_Azimuth` and generation, transform `Roof_Azimuth` so that equal angles either side of North are treated the same.

d) (1 mark) Visualise your new version of `Roof_Azimuth`.

Question 3 (19 Marks)

a) (2 marks) Generate a visualisation of the correlation coefficient between `Generation` and all variables. Briefly discuss your findings from in the context of predicting `Generation`.

b) (3 marks) Construct an appropriate plot to visualise the relationship between `Generation` and `Panel_Capacity`. Briefly discuss your findings.

c) (3 marks) Construct an appropriate plot to visualise the relationship between `Generation` and `Latitude`. Briefly discuss your findings.

d) (5 marks) For each city, compile a table that shows the mean generation for combinations of `Roof_Azimuth` and `Roof_Pitch`. Bin values of `Roof_Azimuth` into 45° groups. Briefly discuss your findings.

e) (4 marks) Pick one city, bin households based on `Panel_Capacity` and `Shading`. For each bin display the mean `Generation`. Display the results using an appropriate visualisation. Discuss your findings.

f) (4 marks) Pick one city, bin households based on `Panel_Capacity` and `Year`. For each bin display the mean generation. Display the results using an appropriate visualisation. Discuss your findings.