Student name:

Student ID:

SIT225: Data Capture Technologies

Activity 7.1: Data analysis and interpretation

Data analysis is a broad term that covers a wide range of techniques that enable you to reveal any insights and relationships that may exist within raw data. As you might expect, Python lends itself readily to data analysis. Once Python has analyzed your data, you can then use your findings to make good business decisions, improve procedures, and even make informed predictions based on what you've discovered.

You have done data wrangling using Python Pandas module already in activity 5.2. In this activity, you will learn Data science statistics and linear regression models.

Hardware Required

No hardware is required.

Software Required

Python 3
Python packages including Pandas, Numpy, Scikit-learn, seaborn, plotly

Steps:

Step	Action
1	A Jupyter Notebook is provided for Data Science exploration here (https://github.com/deakin-deep-dreamer/sit225/tree/main/week_7). You will need to fill in your student ID and name and run all the cells to observe the output. Convert the Notebook into PDF and merge with this activity sheet which needs to be combined with this week's task for OnTrack submission.

	Question: There are sections in the Notebook. After running the cells and observing the outputs, provide your reflection in brief on the topic items for each section of the Notebook. Answer: <Your answer>
2	Question: In the 1.1 Percentile subsection of **Descriptive statistics** section in the Notebook, you have calculated 10%, 25%, 50% and 75% percentiles for *Max_Pulse*. Compare these percentiles with *Average_Pulse* percentiles for any trend, if exists. Answer: I noticed that the 25%, 50%, and 75% percentiles for Max_Pulse are consistently higher than those for Average_Pulse. This reflects that Max_Pulse values are generally higher, which makes sense since the Max_Pulse represents the highest heart rate during a session, while Average_Pulse represents the average. The trend suggests that while Average_Pulse fluctuates moderately, Max_Pulse has a higher range, indicating more variation in the maximum effort exerted during exercise sessions.
3	Question: In the “Correlation Does not imply Causality” section answer the question regarding the increase of ice cream sale in your own understanding. Answer: No, an increase in ice cream sales does not directly cause an increase in drowning accidents. This is an example of a spurious correlation, where two variables appear to be related, I think the correlation exists because of a common third factor (like summer season), not because of any direct causal relationship.
4	Question: In the **1.7 Linear Regression** section in the Notebook, a linear regression model was used to predict Calorie_Burnage from attributes such as Average_Pulse. The Duration value was predicted from the model for all the value range of Average_Pulse and a regression line was drawn. You will need to answer the follow up question next to 1.7 section where it is required to generate a linear regression model for Duration instead of Average_Pulse to predict the Calorie_Burnage. Take a screenshot of the regression line and paste it here. Also, comment on both the regression lines. Answer: **Calorie_Burnage vs. Average_Pulse**: The regression line shows a positive relationship, meaning as the Average_Pulse increases, the Calorie_Burnage tends to increase. This is expected because a higher pulse rate generally corresponds to higher physical exertion, leading to more calories burned.

	Calorie_Burnage vs. Duration: Similarly, the regression line for Calorie_Burnage vs. Duration shows a positive relationship. The longer the exercise session (Duration), the more calories are burned. This relationship is intuitive, as more time spent exercising results in greater energy expenditure.