Student name: Hoang Long Tran

Student ID: s223128143

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.

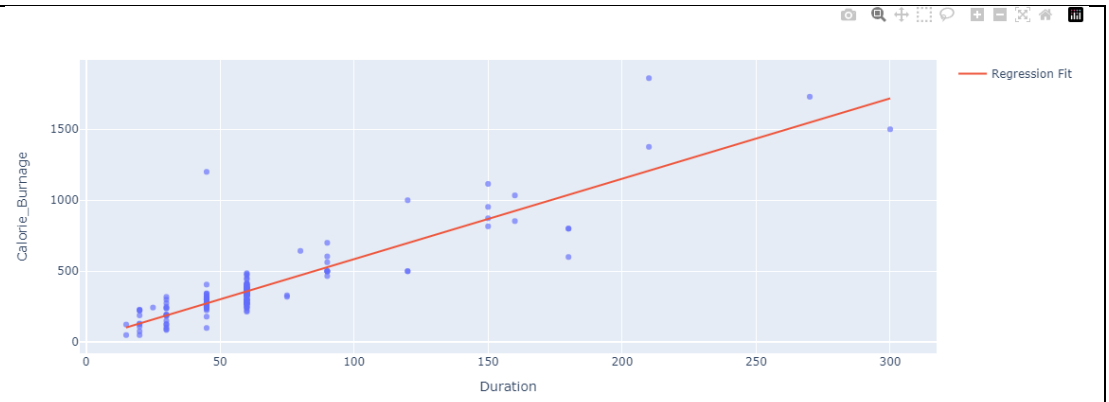
Software Required

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: ok
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: We can observe that 50% the `Average_Pulse` is in quantile 25 to 75 or 100 to 111 beats per minute. Using quantile can be a great way to identify outliers by setting the boundary for data points below Q1 - 1.5IQR or above Q3 + 1.5IQR.
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: Causation relationship means a cause-and-effect relationship. In our case, it means that the sale of ice cream is causing more drowning cases, which is in no way correct. The positive correction between them indicates that both values increase during the summer, and it might be due to a third factor. A third factor can be hot weather, which causes people to visit the beach more often and leads to higher ice cream sales and drowning accidents.
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:

	 Using `Duration` as predictor to predict ` Burnage` shows an overall better performance than using `Average_Pulse`. For ` Average_Pulse` the Mean Squared Error shows for 74716, so the square difference between actual values and predicted values is 74716, which is very large. Couple with an R^2, an indicator of goodness of fit, with only 0.0003, meaning that this model is useless. For the ` Duration` model, the mse is 15796.8, and R^2 is 0.79, which shows an acceptable performance.

Weekly task

Q2.

Up until step 7, my analysis is similar to the example given by the unit chair. I use the Temperature as the predictor and Humidity as the response variable. I plot the scatterplot, and the min/max temperature is similar to the example. However, I used IQR method to remove outliers, since if I just remove the 5 highest and lowest temperature, I will just be repeating myself. Also, after the outliers is removed, I noticed that our data might not have a linear relationship, and using linear model might not capture everything, so I use a non-linear one, which shows higher performance on the metrics, indicating my guesses are correct. To improve the predictions, I have also added new features deriving from Temperature and Humidity.

Q3.

https://www.youtube.com/watch?v=yvDXZyV92n8

Q4.

https://github.com/tomadonna1/SIT225_2024T2/tree/main/Pass%20Task%20Data%20analysis%20and%20interpretation