</talentlabs>

Chapter 3 - Data Collection Mini-project

Instructions

1. Please submit the assignment through TalentLabs Learning System.

In this exercise you will work through the data collection process. An example solution is discussed in the final lesson of chapter 2. By the end of the exercise, you will submit a zip file that will contain three files – the data file, the documentation file, and **this word document** with your answers to previous questions.

Your challenge is to pick a topic of your choosing (something that interests you) and then **ask a question** that can be answered with data. Next, you need to state:

- The **aim** of your project
- The **type of data** that you require.
- Whether primary or secondary data collection is better for the task and **why**
- With what **method** will the data be collected and **why?**

All this information should be documented into the documentation file.

Next you should collect the data. Once the data is collected create a folder and add the data to it. In the same folder add this word document with your answers to previous questions in Part 1. To the folder, also add the documentation file and document your project within that file. When documenting, state the aim, type of data, method of data collection etc. Also explain the reasons behind your decisions and include anything else you think should be documented.

Finally, create a zip version of your folder that has the three files and submit the zip file to Talent Labs.

(9 marks)

(A) Aim of project:

1. Which location in Kuala Lumpur's area has the lowest monthly rental?
2. Which location in Kuala Lumpur's area has the lowest house price?

(B) Type of data required:

Variable	Data type	Description
property_id	integer	Unique identifier for the properties, serving as primary key

transaction_type	string - categorical	Indicates whether the property is listed for sale or rent
Property_category	string - categorical	The general category to which the property belongs, such as "House" or "Apartment/Condominium."
property_type	string - categorical	The specific type of property, providing more detailed information about its structure or style, e.g., "2-storey Terraced House" or "Condominium."
location_city	string – categorical	The city where the property is situated, in this case, Kuala Lumpur.
Location_subarea	string - categorical	A more specific area inside the city, for example, both Ampang and Cheras which are in KL city.
bedrooms	Integer	The number of bedrooms in the property
bathrooms	Integer	The number of bathrooms in the property
property_size_sqft	Integer	The size of the property in square feet
property_price	float	The price at which the property is listed for sale. This column may be empty for properties listed for rental.
monthly_rent	float	The monthly rental cost if the property is listed for rent. This column may be empty for properties listed for sale.

(C) Whether primary or secondary data collection is better for the task and why

- Primary data is the superior choice for this task.
- Primary data is usually more quality, reliable, and understanding to analyse.
- With getting up-to-date data of housing price and monthly rental cost, it will be easier for its users to take fast action (e.g., finding best affordable place to rent or buy)

(D) What does method will the data be collected and why?

Web scraping via python requests from mudah.my website (see code at Kaggle: https://www.kaggle.com/code/tanyongsheng/learn-scrape-mudah-my)

The amount of data is large (around 10k rows) for manual collecting. Thus, writing a python script could save the manual work and we could run the script from time to time if we need to data again in future.