Deakin University Trimester 2, 2022
School of IT Unit Assessment Task 2
Team: SIG742

SIG742: Modern Data Science

Group Work in group of up to 3 members. Group sign-up on **Great Learning** by 8:00pm on 04/10/2022 (Week 07 Sunday).

Extension Request Students with difficulty in meeting the deadline for proper reasons such as illness, etc. must apply for an assignment extension with supporting evidence, one day prior to the assessment due date. Apply via '**Great Learning**'.

Academic Integrity All assignment will be checked for plagiarism, and any academic misconduct will be reported to unit chair and university.

Instructions

Assessment Task 2 Questions

There are **2** parts in this assessment task:

Part 1 The first part will focus on the data manipulation skills which includes the `data wrangling`, the `EDA`, from **M04**.

Part 2 The second part is for those who are aiming to achieve '`High Distinction`' (HD) for this assessment task, and it will focus on more advanced data analysis for data science. This part will require the knowledge covered in **M05** and also **M06**.

What to Submit?

You (group) are required to submit the following completed files to the corresponding *Assignment* folder in Great Learning:

`SIG742Task2.ipynb` The completed notebook (one for each group) with all the run-able code on all requirements.

In general, you need to complete, save the results of running, download and submit your **notebook** from Python platform such as `Google Colab` to Great Learning. You need to clearly list the answer for each question, with sufficient coding comments, and the expected format from your notebook will be like in Figure 1. Also you (group) need to do the team work and distribute the work appropriately among all group members.

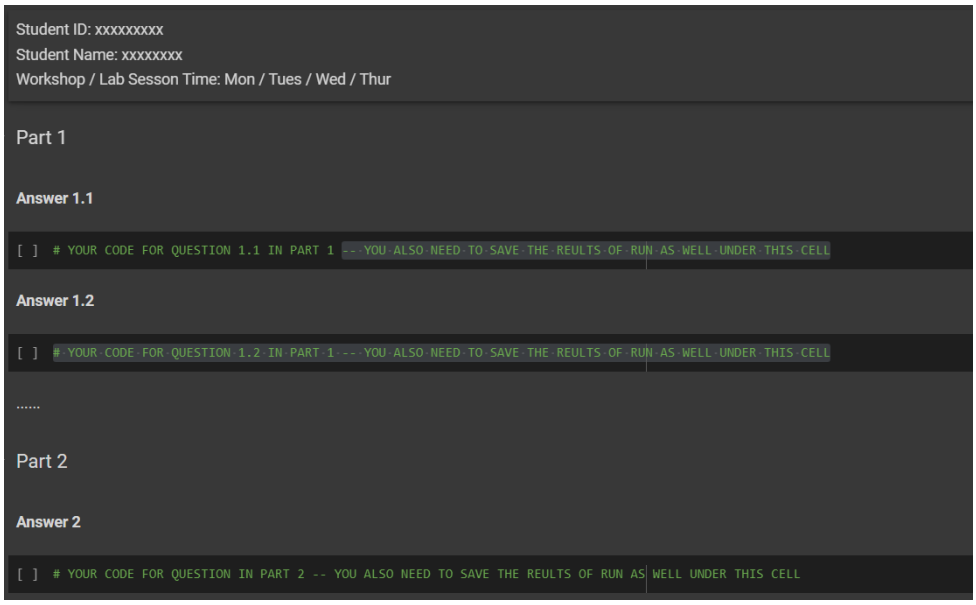

Figure 1: Notebook Format

`SIG742Task2Report.pdf` You (group) are also required to put your answer (code) and running results from `SIT742Task2.ipynb` into a `pdf` as the report for your task2 assignment (copy the code and paste into the report, the code format such as Indentation should be same in the ipynb notebook).

In this report (one for each group), you will need to include the questions for the assignment for both `Part 1` and `Part 2`. Also you will need to provide a clear explanation on your logic for solving each question. In the explanation, you will need to cover below parts: 1). why you decide to choose your solution; 2). are there any other solutions that could solve the question; 3). whether your solution is the optimal or not? why? The length of the explanation part for each question is limited below 100 words.

`SIG742Task2Report.avi` If you (group) are aiming to achieve '`High Distinction`' (HD) and choose to work on *Part 2* of this assessment task, one important submission is a **short video** in which *You* orally present the solutions that you provide in the notebook and illustrate the running of code line by line. You (group) are required to submit a video demonstration (one for each group) between 5 and 10 minutes for your *Part 2* only , and the file format can be other common ones, such as '`MKV`', '`WMV`', '`MOV`' etc. All members in you group need to be involved in the video demonstration.

Part I

Data Manipulation

There are **8** questions in this part for **80** marks, and each question is **10** marks.
You are required to use `Google Colab` to finish all the coding in the *code block cell*, provide sufficient coding comments, and also save the result of running as well. Also you need to put the code, running results and the explanation into pdf report as well into your `SIG742Task2Report.pdf`.

Question 1.1

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. The meaning of the column is in **assignment2data.pdf**

- print the shape of the csv dataframe and find how many rows are duplicated (use `pandas`);

- remove the duplicated rows and then print the new shape of the dataframe (use `pandas`);

Question 1.2

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. Removing the duplicated rows from dataframe and save as the new dataframe.

- define a function `missingdf(df)` with input argument `df`, which could print out all the column in dataframe `df` and also the missing value rate for each column. For example, with total 1000 rows, if column1 has missing value in 200 rows and its missing value rate will be 0.2 or 20%. The result of the function `missingdf(df)` will print the new dataframe which has two columns: the `column_name` and the `percent_missing`

Question 1.3

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. Removing the duplicated rows from dataframe and save as the new dataframe.

- define a function `column_list(df)` which could return a list which only contains the numerical column names and another list which only contains the categorical column names.
- Use `describe()` function from `pandas` to print out the statistics for all numerical columns.

Question 1.4

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. Removing the duplicated rows from dataframe and save as the new dataframe. The meaning of the column is in **assignment2data.pdf**

- look into the dataframe, there are two columns which are corrupted. Could you find it out and explain the reason? (you will need to draw some visualizations or check the statistics, also you may need to look into the data to understand the meaning the columns.)
- For the two corrupted columns, could you provide the solution to correct them?

Question 1.5

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. Removing the duplicated rows from dataframe and save as the new dataframe. The meaning of the column is in **assignment2data.pdf**

- Write code and return the results by using `pandas` package to find out "What percentage of customers who have purchased female items have paid by credit card?"
- Write code and return the results by using `pandas` package to find out "What was the total revenue to the nearest dollar for customers who have paid by credit card?"

Question 1.6

Open the **assignment2data.json** file and convert it to csv format as dataframe in `pandas`. Removing the duplicated rows from dataframe and save as the new dataframe. The meaning of the column is in **assignment2data.pdf**

- Write a code to change the value of 'Y' from column `is_newsletter_subscriber` to 1 and 'N'

to 0. (1 and 0 is "int" type)

- Print out the value count for column `is_newsletter_subscriber`.

Question 1.7

Open the **assignment2data.json** file and convert it to `csv` format as `dataframe` in `pandas`. Removing the duplicated rows from `dataframe` and save as the new `dataframe`. The meaning of the column is in **assignment2data.pdf**

Create some new features for the `dataframe` by using below code:

```
df['female_item_rate'] = df['female_items']/df['items']
df['male_item_rate'] = df['male_items']/df['items']
df['unisex_items_rate'] = df['unisex_items']/df['items']
```

- Write a code find out how many rows (customers) could have the value `female_item_rate` $== 1$ and the value `male_item_rate` $== 1$ and the value `orders` > 4.11

Question 1.8

Open the **assignment2data.json** file and convert it to `csv` format as `dataframe` in `pandas`. Removing the duplicated rows from `dataframe` and save as the new `dataframe`. The meaning of the column is in **assignment2data.pdf**

In this question, you will use the original format of the data to group data on the value of column `is_newsletter_subscriber` to show the average order value, the max order value, the median order value.

Part II

Advanced Data Analytics for Data Science

This part is for students (groups) who are aiming to achieve '`High Distinction`' (HD) for this assessment task.

There are **2** versions of Question 2 in this part for **20** marks: **10** marks for coding, and **10** marks for the explanation (as in '**What to Submit**'). You (group) should only work on one version based on your team members' own enrolment details as in below section, and working on the wrong one will result in zero for this question.

For your question, you (group) are required to use `Google Colab` to finish all the coding in the *code block cell*, provide sufficient coding comments, and also save the result of running. Also you need to put the code, running results and the explanation into pdf report as well for your `SIG742Task2Report.pdf`. In addition, a short video demonstration from all the members in group is required for submission.

Which version of Question 2 for you?

The code of determining your (group) Q2 version is provided:

```
def sum_digits(n):
    r = 0
    while n:
        r, n = r + n % 10, n // 10
    return r
```

```
def check_studentid(studentid_list):
    studentid = sum(studentid_list)
    x = sum_digits(studentid)
    if x % 2 == 0:
        print('version II')
    else:
        print('version I')
check_studentid([9876543210,9876543211,...])
#replace the value by your student ID list in your group
```

You need to copy this code to your notebook and run the function with your (group) student ID list. You will also need to print/save the result of the code running and put it in your `SIG742Task2Report.pdf`.

Question 2 (Version-I

Time Series Forecasting

In this part, we will use the data from **HK2012-2018 data**. This data is a multi-variate time series data with its granularity on month of the year. The details of the data and column explanation is on **HK2012-2018**

Question 2.1 Future value of the arrivals depends on the average of its k previous values. Therefore, we will use the **moving average** ($\hat{y}_t = \frac{1}{k}\sum_{n=1}^{k} y_{t-n}$) to forecast the future values. In this question, we use the data from `2012-01` to `2017-12` as training data and the data from `2018-01` to `2018-06` as testing data

- Define the function `moving_avg()` as below:

```
# define the moving average function by numpy
def moving_average(series, n):
    """
    Calculate average of last n observations
    """
    return xxxx
```

- You will need to run the STL decomposition to find out the seasonality pattern and also the trend pattern. Explain what you have find and then run the function `moving_avg()` on the trend component from the STL decomposition to forecast the trend value from `2018-01` to `2018-06`. You will need to report the forecasting error with `RMSE` (you will just forecast one step ahead which means you will know the value of arrival when your forecasting is moving ahead).

Question 2.2

The problem is how to forecast the future arrivals on given time series, in normal forecasting scenario, the types of forecasting are usually concluded as: `one-step forecasting` and `multi-steps forecasting`. In here, we will focus on `multi-steps forecasting`.

- Could you run `ARIMA` model to forecast the the arrivals from `2018-01` to `2018-06` by performing the `multi-steps forecasting`. (your model will need to only train on the training data and forecast the future on multi-steps at once.) You will also need to report the `RMSE` on your forecasting.

- Could you show the best p,d and q parameters with the evaluation metrics on `RMSE`? (you will need to do grid search on the three parameters, assume the range of the all three parameter is same from 1 to 3).
- Could you write down any other good models to do the `multi-steps forecasting` here rather than the `ARIMA`?

Question 2 (Version-II)

Transaction Data Analysis

In this part, we will do the analysis on the customer transaction data. The data is from **customer-transaction**. The row of the data represents the item transaction from customer (one item from a transaction for that customer). The product is represented as the `product_id` and the `commodity`. There is also a column `basket_id` to help group the transaction together into basket level (check out basket).

Question 2.1

You will need to group the `customer_id` and `basket_id` to find out the product `commodity` in each basket. Then you will need to answer:

- How many transactions based on basket level? what is the average basket size?
- What is the most popular product commodity (based on the frequency of the purchase)?
- What is the average of the total transaction price (average basket total price) for each customer?
- You will need to transform the data into a format of: the row represent the basket, the column will be all product commodity, the value of the column should indicate whether the basket contains particular product commodity. Name this new dataframe as `transaction_product`
- You will need to transform the data into a format of: the row represent the unique customer, the column will be all product commodity, the value of the column should be the frequency of the purchase on the particular commodity cross entire data. Name this new dataframe as `customer_product_freq`
- Using the `customer_product_freq` to find the top 5 similar customers for each customer. (Check out the **KNN**)

Question 2.2

Using the dataframe `transaction_product` to conduct association rule analysis (you are recommended to use `mlxtend` package). You will need to find out:

- The itemsets(basket) having length more than 1 and minimum support of 5%
- The association rules with minimum support of 2% and having lift more than 1.

The definition of the support and lift is in M05E, lecture slides and also Association rule learning.