Data Preparation Homework

TEAM NAME: ………………………..

Assignment 1: Understand COVID-19 Situation in Thailand

This assignment will let you explore the real data which is open data provided by the government and apply data preparation techniques necessary before showing in the dashboard. The sample data product could be seen in <https://covid19.workpointnews.com/>

Dataset: <https://raw.githubusercontent.com/jumpai/data_preapration_dataset/master/20200510_covid_case.csv>

1. Profiling: Understand Your Data
   1. Inspect data quality by showing numbers of missing values of each column
   2. Explore patients profile:
      1. What's the time period of data collected in this dataset?
      2. How many cases have been reported?
      3. Show the distribution of patients by ‘province of isolation’
      4. List nationalities and number of patients of top 5 non-Thai patients in Thailand
2. Apply data preparation techniques
   1. Create ‘age group’ column by grouping age in the range of 10  
      (E.g. 0-10, 11-20, 21-30, …. , 91-100) and show the distribution of this column
   2. To analyze by date range, we might need to do date parsing
      1. Extract month from “Announce Date” column and answer these questions  
         (In order to process date we can separate date column by “/” or parse the date to datetime type then selects and compares month)
         1. How many cases of COVID announced in April?
         2. How many cases of COVID announced during 23 - 30 March 2020?
         3. (Extra) How many cases of COVID announced over the weekend in April?  
            \* Note: Need to parse ‘Announce Date’ column to datetime type and then extract day (Mon - Sun) from that column
   3. Capture your dataset after applying data preparation techniques

Assignment 2: Explore your own dataset

1. Write a problem statement and explain your dataset
2. Provide your dataset sample url
3. Profiling: Understand Your Data
   1. Inspect data quality by showing numbers of missing values of each column
4. Apply data preparation techniques
   1. Identify data messy and suggest data preparation techniques
   2. Apply feature synthesis and data preprocessing techniques (if needed)
5. Provide URL of your dataset after applying data preparation techniques