# Group Work Task: Data Processing

Welcome to the group work session! In this session, you will be utilizing not only the techniques discussed during the session but also your own prior skills. The ultimate goal is to create a data that is self informative and ready to do statistical analysis.

---

# Data

There are three different data covering different application domain. You will be assigned to one of the following data during the session. Your task is to create a tidy data.

1. **VEGIEHAT Pilot Database**
2. **Air Quality Data**
3. **English Premier League Player Stats**

# Data Dictionary

In this section you will get description of the variables in each of the data:

# VEGIEHAT Pilot Database

This is a crowd sourced data collected through VEGIEHAT (https://vegiehat.org/) platform. The objective was to create a database containing essential food and vegetable prices information in Bangladesh. The consumers can directly enter price information based on their own purchase. In this task, you are given a part of the data to work in a group. The following are the variables in the data.

- `SubmissionId:` Unique submission ID, this is column should be unique throughout the data
- `SubmissionTime:` The submission date and time (in GMT)
- `UserId:` User ID
- `DistrictName:` Name of Administrative Districts
- `UpazilaName:` Name of Administrative Upazila/Thana
- `ItemsToChoose:` Food/Vegetable Item to choose (Multiple allowed)
- `Value - Rice:` Price of 1 kg Rice in Bangladeshi Taka (BDT)
- `Value - Flour:` Price of 1 kg Flour in BDT
- `Value - Lentil:` Price of 1 kg Lentil in BDT
- `Value - Soybean Oil:` Price of 1L Soybean Oil in BDT
- `Value - Salt:` Price of 1 kg salt in BDT
- `Value - Sugar:` Price of 1 kg Sugar in BDT
- `Value - Eggs:` Price of 4 eggs in BDT
- `Value - Chicken:` Price of Farm Chicken per kg in BDT
- `Value - Potato:` Price of 1 kg potato in BDT
- `Value - Eggplant:` Price of 1 kg Eggplant in BDT
- `Value - Onion:` Price of 1 kg Onion in BDT
- `Value - Green Chilli:` Price of 1kg green chilli in BDT
- `Comments:` Any comments written by submitting users

## Expected Output from VEGIEHAT Data

The following is the expected output data after doing cleaning and necessary processing:

- `SubmissionId:`
- `SubmissionDateOnly:` Date part from the date time variable in the original data
- `SubmissionTimeOnly:` Time (24 hour) from the data time variable in the original data
- `UserId:` User ID
- `DistrictName:` Name of Administrative Districts
- `UpazilaName:` Name of Administrative Upazila/Thana
- `ItemsToChoose:` Food/Vegetable Item - One row should contain only one item
- `PurchaseUnit:` Unit of the purchase option (Either 1 kg / 1L as appropriate)
- `Price:` Price Per Unit
- `Comments:` Any comments written by submitting users

# Air Quality Data

A data of ambient air quality PM2.5 measured in every hours in major cities in Bangladesh covering 2012 to 2021. Your task is to create a tidy data so that subsequent analysis can be done easily. The data is in an Excel file. One worksheet contains one city. The following are the variables in the data:

- `Date:` Date in Datetime format
- `Time:` Time of the day in 24 hours format
- `PM2.5:` Concentration of PM2.5 measured in $\mu g/m^3$
- `Temperature:` in Degree Celsius

## Expected Output from Air Quality Data

The following is the expected output data after doing cleaning and necessary processing:

- `Date:` Only date part from the date time variable
- `Time:` Time of the day in 24 hours format
- `CityName:` Name of the city from the raw data (Worksheet name of the Excel file)
- `PM2.5:` Concentration of PM2.5 measured in $\mu g/m^3$
- `Temperature:` in Degree Celsius

# English Premier League Player Stats

This data contains English Premier League (https://www.premierleague.com/stats) Player's stats accessed from Premier League official website. The data contains all player up to Sep 24, 2020. One row contains one player's data. The following are the variables in the data.

- `Name`
- `Jersey Number`
- `Club`
- `Position`
- `Nationality`
- `Age`
- `Appearances`
- `Wins`
- `Losses`
- `Goals`
- `Goals per match`
- `Headed goals`
- `Goals with right foot`
- `Goals with left foot`
- `Penalties scored`
- `Freekicks scored`
- `Shots`
- `Shots on target`
- `Shooting accuracy %`
- `Hit woodwork`
- `Big chances missed`
- `Clean sheets`
- `Goals conceded`
- `Tackles`
- `Tackle success %`
- `Last man tackles`
- `Blocked shots`
- `Interceptions`
- `Clearances`
- `Headed Clearance`
- `Clearances off line`
- `Recoveries`
- `Duels won`
- `Duels lost`
- `Successful 50/50s`
- `Aerial battles won`
- `Aerial battles lost`
- `Own goals`

- Errors leading to goal
- Assists
- Passes
- Passes per match
- Big chances created
- Crosses
- Cross accuracy %
- Through balls
- Accurate long balls
- Saves
- Penalties saved
- Punches
- High Claims
- Catches
- Sweeper clearances
- Throw outs
- Goal Kicks
- Yellow cards
- Red cards
- Fouls
- Offsides

# Expected Output Data from EPL Player Stats

The following is the expected output data after doing cleaning and necessary processing:

- Name
- Jersey Number
- Club
- Position
- Nationality
- Age
- StatsName
- Statsvalue

It will be a long format data. Some of the stats are not relevant for certain position’s player.