# VEGIEHAT Data Analysis Markdown

Shihab Sarker

2025-03-16

## 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:`
- `UserId:` User ID
- `ItemsToChoose:` Food/Vegetable Item - Separated in each column

## Objectives

After completing data cleaning and necessary processing determine followings: - What is the frequency of selecting `Soybean Oil` ? - What is the frequency of selecting `Soybean Oil` and `Eggs` ?

## Data Cleaning and Processing

```
# Load necessary Packages
library(readxl)
library(dplyr)
library(tidyr)
library(stringr)

# Import data
dfVEGIEHAT <- read_xlsx("./Data/VEGIEHAT-Pilot-Database.xlsx", sheet = "Sheet1",
                        range = cell_cols("A:F"),
                        col_types = c("text","date","text","text","text","text" ))
show(dfVEGIEHAT)
```

```
## # A tibble: 185 × 6
##    SubmissionId              SubmissionTime      UserId DistrictName UpazilaName
##    <chr>                     <dttm>              <chr>  <chr>        <chr>
##  1 6b7483a2-25c7-4fa9-864e-… 2024-11-30 23:32:19 173dc… Dhaka        Adabor
##  2 0e3b61f1-8e21-4f05-aea0-… 2024-11-30 23:48:36 173dc… Dhaka        Dhanmondi
##  3 8d07e7ca-e22f-44be-b5c6-… 2024-12-01 04:32:29 0859e… Dhaka        Badda
##  4 988acbef-9e63-4ab0-aea6-… 2024-12-02 06:36:04 0c296… Dhaka        Mohammadpur
##  5 c66083e5-1cd2-4697-856b-… 2024-12-03 16:25:33 fae74… Dhaka        Pallabi
##  6 228625d6-bc6c-4776-b20b-… 2024-12-02 12:06:00 24d4f… Dhaka        Mohammadpur
##  7 d7ed0df1-13ae-4691-a3a5-… 2024-12-13 08:30:45 3c4da… Dhaka        Mohammadpur
##  8 70f25470-7afe-4eca-a9ac-… 2024-12-13 10:25:14 def25… Naogaon      Mahadebpur
##  9 96ac0237-9540-4781-b6a0-… 2024-12-13 10:34:37 2121c… Naogaon      Dhamoirhat
## 10 b52f7f5c-4bde-44d5-a339-… 2024-12-13 10:35:37 2121c… Naogaon      Dhamoirhat
## # ℹ 175 more rows
## # ℹ 1 more variable: ItemsToChoose <chr>
```

```
# Filter rows with non-empty values
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(ItemsToChoose),            # Remove NA in ItemsToChoose
    nchar(ItemsToChoose) > 0,         # Remove empty ItemsToChoose
    nchar(UserId) > 0                 # Remove empty UserId
  ) %>%
  select(
    SubmissionId, UserId, ItemsToChoose  # Select relevant columns
  ) %>%
  separate_rows(
    ItemsToChoose, sep = ","            # Split comma-separated items
  ) %>%
  mutate(
    ItemsToChoose = str_trim(
      ItemsToChoose,                    # Trim leading/trailing spaces
      side = "both"
    )
  ) %>%
  distinct(
    SubmissionId, UserId, ItemsToChoose  # Remove duplicates
  ) %>%
  mutate(
    ItemChosen = 1                      # Mark item as chosen
  ) %>%
  pivot_wider(
    names_from = ItemsToChoose,           # Create columns for each item
    values_from = ItemChosen,             # Set value to 1 for chosen items
    values_fill = list(ItemChosen = 0)    # Fill missing values with 0
  )
show(dfItemsChosen)
```

```
## # A tibble: 183 × 14
##    SubmissionId     UserId  Rice `Soybean Oil` Sugar `Green Chilli` Lentil Flour
##    <chr>            <chr>  <dbl>         <dbl> <dbl>          <dbl>  <dbl> <dbl>
##  1 6b7483a2-25c7-4… 173dc…     1             1     0              0      0     0
##  2 0e3b61f1-8e21-4… 173dc…     1             1     0              0      0     0
##  3 8d07e7ca-e22f-4… 0859e…     1             1     0              0      0     0
##  4 988acbef-9e63-4… 0c296…     0             0     1              1      0     0
##  5 c66083e5-1cd2-4… fae74…     1             1     0              0      0     0
##  6 228625d6-bc6c-4… 24d4f…     1             1     0              1      1     1
##  7 d7ed0df1-13ae-4… 3c4da…     1             1     1              1      1     1
##  8 70f25470-7afe-4… def25…     1             1     0              0      0     0
##  9 96ac0237-9540-4… 2121c…     0             0     0              0      0     0
## 10 b52f7f5c-4bde-4… 2121c…     0             0     0              0      0     0
## # ℹ 173 more rows
## # ℹ 6 more variables: Salt <dbl>, Eggs <dbl>, Potato <dbl>, Onion <dbl>,
## #   Chicken <dbl>, Eggplant <dbl>
```

```
# Find the frequency of selecting 'Soybean oil'
frequency_soybean_oil <- dfItemsChosen %>%
  count(`Soybean Oil`)
print(frequency_soybean_oil)
```

```
## # A tibble: 2 × 2
##   `Soybean Oil`     n
##           <dbl> <int>
## 1             0   106
## 2             1    77
```

The frequency of selecting Soybean Oil is 77.

```
# Find the frequency of selecting 'Soybean oil'
frequency_soybean_oil_eggs <- dfItemsChosen %>%
  count(`Soybean Oil`, Eggs)
print(frequency_soybean_oil_eggs)
```

```
## # A tibble: 4 × 3
##   `Soybean Oil`  Eggs     n
##           <dbl> <dbl> <int>
## 1             0     0    71
## 2             0     1    35
## 3             1     0    24
## 4             1     1    53
```

The frequency of selecting Soybean Oil and Eggs is 53