VEGIEHAT Data Processing

Shihab Sarker

2025-03-17

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

Data Processing

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

# Import data
dfVEGIEHAT <- read_xlsx("./Data/VEGIEHAT-Pilot-Database.xlsx", sheet = "Sheet1")

# Filter rows with non-empty values
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(ItemsToChoose),
    nchar(ItemsToChoose) > 0,
    nchar(UserId) > 0
  ) %>%
  select(
    SubmissionId, SubmissionTime, UserId, DistrictName,
    UpazilaName, ItemsToChoose, `Value - Rice`, `Value - Flour`, `Value - Lentil`,
    `Value - Soybean Oil`, `Value - Salt`, `Value - Sugar`, `Value - Eggs`,
    `Value - Chicken`, `Value - Potato`, `Value - Eggplant`, `Value - Onion`,
    `Value - Green Chilli`
  ) %>%
  mutate(
    ItemsToChoose = str_trim(ItemsToChoose, side = "both"),
    SubmissionDateOnly = format(as.Date(SubmissionTime), "%d/%m/%Y"),
    SubmissionTimeOnly = format(SubmissionTime, "%H:%M:%S")
  ) %>%
  separate_rows(
    ItemsToChoose, sep = ","
  ) %>%
  mutate(
    ItemsToChoose = str_trim(
      ItemsToChoose,
      side = "both"
    )
  ) %>%
  mutate(
    PurchaseUnit = case_when(
      ItemsToChoose %in% c("Rice", "Flour", "Lentil", "Salt", "Sugar", "Potato",
                           "Eggplant", "Onion", "Green Chilli", "Chicken") ~ "1 kg",
      ItemsToChoose == "Soybean Oil" ~ "1L",
      ItemsToChoose == "Eggs" ~ "1 Hali",
      TRUE ~ NA_character_
    )
    )%>%
  mutate(
    Price = case_when(
      ItemsToChoose == "Rice" ~ `Value - Rice`,
      ItemsToChoose == "Flour" ~ `Value - Flour`,
      ItemsToChoose == "Lentil" ~ `Value - Lentil`,
      ItemsToChoose == "Soybean Oil" ~ `Value - Soybean Oil`,
      ItemsToChoose == "Salt" ~ `Value - Salt`,
      ItemsToChoose == "Sugar" ~ `Value - Sugar`,
      ItemsToChoose == "Eggs" ~ `Value - Eggs`,
      ItemsToChoose == "Chicken" ~ `Value - Chicken`,
      ItemsToChoose == "Potato" ~ `Value - Potato`,
      ItemsToChoose == "Eggplant" ~ `Value - Eggplant`,
      ItemsToChoose == "Onion" ~ `Value - Onion`,
      ItemsToChoose == "Green Chilli" ~ `Value - Green Chilli`,
      TRUE ~ NA_real_
    )
  ) %>%
  select(
  SubmissionId, SubmissionDateOnly, SubmissionTimeOnly, UserId, DistrictName,
  UpazilaName, ItemsToChoose, PurchaseUnit, Price
  )

show(dfItemsChosen)
```

```
## # A tibble: 824 × 9
##    SubmissionId        SubmissionDateOnly SubmissionTimeOnly UserId DistrictName
##    <chr>               <chr>              <chr>              <chr>  <chr>
##  1 6b7483a2-25c7-4fa9… 30/11/2024         23:32:19           173dc… Dhaka
##  2 6b7483a2-25c7-4fa9… 30/11/2024         23:32:19           173dc… Dhaka
##  3 0e3b61f1-8e21-4f05… 30/11/2024         23:48:36           173dc… Dhaka
##  4 0e3b61f1-8e21-4f05… 30/11/2024         23:48:36           173dc… Dhaka
##  5 8d07e7ca-e22f-44be… 01/12/2024         04:32:29           0859e… Dhaka
##  6 8d07e7ca-e22f-44be… 01/12/2024         04:32:29           0859e… Dhaka
##  7 988acbef-9e63-4ab0… 02/12/2024         06:36:04           0c296… Dhaka
##  8 988acbef-9e63-4ab0… 02/12/2024         06:36:04           0c296… Dhaka
##  9 c66083e5-1cd2-4697… 03/12/2024         16:25:33           fae74… Dhaka
## 10 c66083e5-1cd2-4697… 03/12/2024         16:25:33           fae74… Dhaka
## # ℹ 814 more rows
## # ℹ 4 more variables: UpazilaName <chr>, ItemsToChoose <chr>,
## #   PurchaseUnit <chr>, Price <dbl>
```