Data Science Journey Using R

Jaynal Abedin

PhD in Data Science

16-Mar-2025

Creating a Tidy Data from a Messy Data – Example from VEGIEHAT

A Crowd Surveillance for Monitoring Essential Food and Vegetable Prices in Bangladesh

VEGIEHAT Data – Current Structure

1	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

VEGIEHAT Data – Current Structure

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

- **What is the frequency of selecting Soybean Oil?**
- **What is the frequency of selecting Soybean Oil and Eggs?**

VEGIEHAT Data – Current Structure

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

- **What is the frequency of selecting Soybean Oil?**
- **What is the frequency of selecting Soybean Oil and Eggs?**

Is the current layout suitable to answer the above question?

Is it a tidy data?

VEGIEHAT Data – Current Structure

	Submission.ID	UserId	Rice	Soybean C	Sugar	Green Chi	Lentil	Flour	Salt	Eggs	Potato	Onion	Chicken	Eggplant
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee709	1	1	0	0	0	0	0	0	0	0	0	0
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fa	0	0	1	1	0	0	0	0	0	0	0	0
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76	1	1	0	0	0	0	0	0	0	0	0	0
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0	1	1	0	1	1	1	1	1	1	1	0	0
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	3c4da5d53	1	1	1	1	1	1	1	1	1	1	1	1

Is this a tidy data?

Import/Load Data in R Session

```
dfVEGIEHAT <- read.xlsx2(
  file = "./data/VEGIEHAT-Pilot-Database.xlsx",
  sheetName = "Sheet1",
  colClasses = c("character","POSIXct","character")
)
```

Use of Pipe Operator

```
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

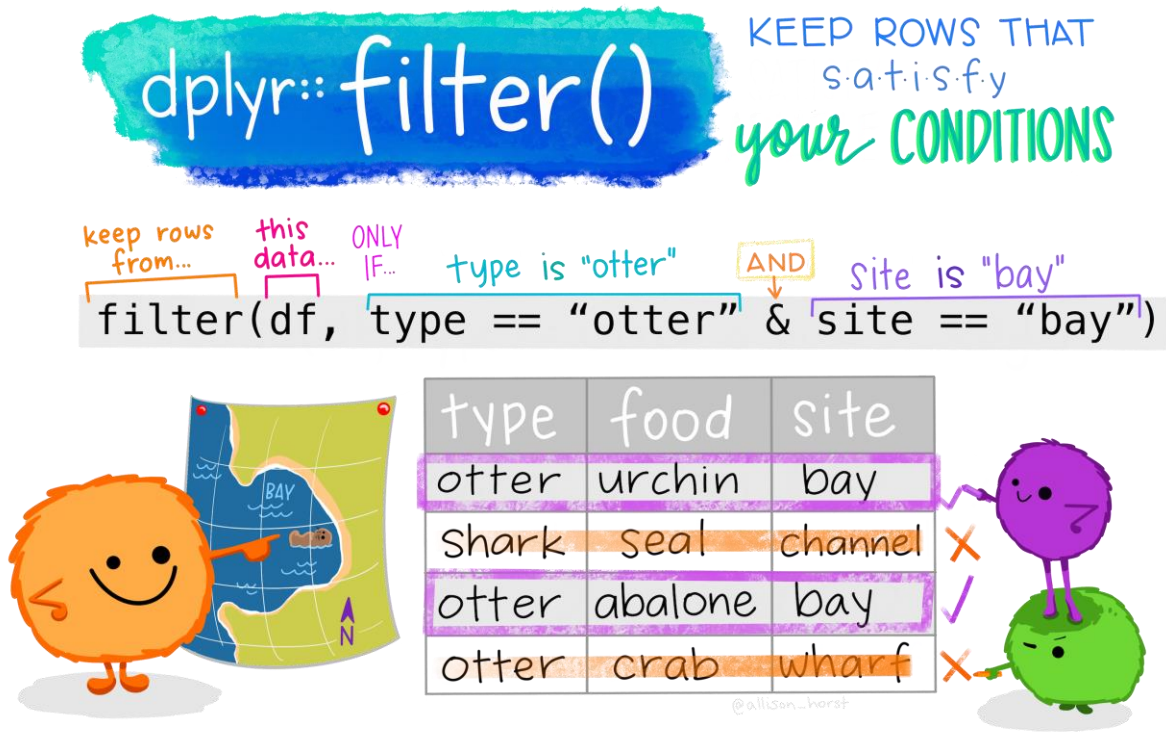
Use of Pipe operator %>% is like a fountain waterfall

Use of Pipe Operator

```
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

Subset Rows - Filtering

```
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```



Art: @allison_horst

Subset Columns - Selecting

```
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

It is better to use only a subset of columns to create a tidy data and then combine with the original data later

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

Create one row for each selected items

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

Trim white spaces if there is any from the “Item.to.Choose” column

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

Art: @allison_horst

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

- **Keep rows with unique combination of Submission ID, User Id and Chosen Items**
- **Create a new column with 1 in every rows**

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

	Submission.ID	UserId	Items.to.Choose	itemChosen
1	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Rice	1
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Rice	1
4	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
5	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Rice	1
6	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Soybean Oil	1
7	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Sugar	1
8	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Green Chilli	1
9	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Rice	1
10	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Soybean Oil	1
11	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Rice	1
12	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Lentil	1
13	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Flour	1
14	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Soybean Oil	1

```r
dfItemsChosen <- dfVEGIEHAT %>%
  filter(
    !is.na(Items.to.Choose),
    nchar(Items.to.Choose)>0,
    nchar(UserId)>0
  ) %>%
  select(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  separate_rows(
    Items.to.Choose, sep = ","
  ) %>%
  mutate(
    Items.to.Choose = str_trim(
      Items.to.Choose,
      side = "both"
    )
  ) %>%
  distinct(
    Submission.ID, UserId, Items.to.Choose
  ) %>%
  mutate(
    itemChosen = 1
  ) %>%
  pivot_wider(
    names_from = Items.to.Choose,
    values_from = itemChosen,
    values_fill = list(itemChosen=0)
  )
```

- **Convert Each Selected Item in a column**
- **The value will be 1 if the item was selected, otherwise it will be 0**

VEGIEHAT Data – Current Structure

	Submission.ID	UserId	Rice	Soybean C	Sugar	Green Chi	Lentil	Flour	Salt	Eggs	Potato	Onion	Chicken	Eggplant
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee709	1	1	0	0	0	0	0	0	0	0	0	0
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fa	0	0	1	1	0	0	0	0	0	0	0	0
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76	1	1	0	0	0	0	0	0	0	0	0	0
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0	1	1	0	1	1	1	1	1	1	1	0	0
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	3c4da5d53	1	1	1	1	1	1	1	1	1	1	1	1

Now we can use this data to merge with original data using submission Id and Users Ids

VEGIEHAT Data – Current Structure

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

VEGIEHAT Data – Current Structure

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

	Submission.ID	UserId	Items.to.Choose	itemChosen
1	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Rice	1
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Rice	1
4	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
5	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Rice	1
6	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Soybean Oil	1
7	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Sugar	1
8	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Green Chilli	1
9	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Rice	1
10	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Soybean Oil	1
11	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Rice	1
12	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Lentil	1
13	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Flour	1
14	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Soybean Oil	1

VEGIEHAT Data – Current Structure

	Submission ID	Submission time	UserId	DistrictName	UpazilaName	Items to Choose
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	2024-11-30 23:32:19	173dcb872759	Dhaka	Adabor	Rice,Soybean Oil
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	2024-11-30 23:48:36	173dcb872759	Dhaka	Dhanmondi	Rice,Soybean Oil
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	2024-12-01 4:32:29	0859ee7098b0	Dhaka	Badda	Rice,Soybean Oil
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	2024-12-02 6:36:04	0c29694fadfcb	Dhaka	Mohammadpur	Sugar,Green Chilli
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	2024-12-03 16:25:33	fae74bc76cf60	Dhaka	Pallabi	Rice,Soybean Oil
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	2024-12-02 12:06:00	24d4f26a0359	Dhaka	Mohammadpur	Rice,Lentil,Flour,Soybean Oil,Salt,Eggs,Potato,Onion,Green Chilli
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	2024-12-13 8:30:45	3c4da5d5300f	Dhaka	Mohammadpur	Rice,Flour,Lentil,Soybean Oil,Sugar,Salt,Eggs,Chicken,Potato,Eggplant,Green Chilli,Onion

	Submission.ID	UserId	Items.to.Choose	itemChosen
1	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Rice	1
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Rice	1
4	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872759ee4b05745b23a2e7ae34	Soybean Oil	1
5	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Rice	1
6	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee7098b0fcd674aeb822126f986a	Soybean Oil	1
7	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Sugar	1
8	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fadfcb564ebe238a66c3b99f8	Green Chilli	1
9	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Rice	1
10	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76cf6016774eb072b7ad97bb8	Soybean Oil	1
11	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Rice	1
12	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Lentil	1
13	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Flour	1
14	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0359322aa9de9b83d25d2875	Soybean Oil	1

	Submission.ID	UserId	Rice	Soybean C	Sugar	Green Chi	Lentil	Flour	Salt	Eggs	Potato	Onion	Chicken	Eggplant
2	6b7483a2-25c7-4fa9-864e-4b5545bb1783	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
3	0e3b61f1-8e21-4f05-aea0-b7226a0946e6	173dcb872	1	1	0	0	0	0	0	0	0	0	0	0
4	8d07e7ca-e22f-44be-b5c6-3eb9a8ec71c1	0859ee709	1	1	0	0	0	0	0	0	0	0	0	0
5	988acbef-9e63-4ab0-aea6-b3baa020c32f	0c29694fa	0	0	1	1	0	0	0	0	0	0	0	0
6	c66083e5-1cd2-4697-856b-31c3f1e3b00d	fae74bc76	1	1	0	0	0	0	0	0	0	0	0	0
7	228625d6-bc6c-4776-b20b-8ac43923c8d2	24d4f26a0	1	1	0	1	1	1	1	1	1	1	0	0
8	d7ed0df1-13ae-4691-a3a5-55d0dafaffc8	3c4da5d53	1	1	1	1	1	1	1	1	1	1	1	1

Its time to Practice

- Go to **DSJR** Repository in GitHub
- Browse to the folder "**Tasks**"
- Click on the file "`data-processing-group-work.md`"
- Participants in **Room 1, Room 4 and Room 7** will choose "`VEGIEHAT Pilot Database`"
- Participants in **Room 2, Room 5 and Room 8** will choose "`Air Quality Data`"
- Participants in **Room 3, Room 6, and Room 9** will choose "`English Premier League Player Stats`"
- Do the necessary processing and create the data as per the expected data layout
- Create a single R Script file from each Room and submit directly into GitHub Repository under "Tasks" folder

Hint: you will need to use various dplyr verbs e.g. `filter()`, `mutate()`, `pivot_longer()`, `pivot_wider()`