Data Analysis and Imputation of Survey Dataset

Table of contents

1 Task 1: Summarize the Data Structure

Explaining Variable Types:

Variable	Type
Case_id	Numerical (discrete)
123	Numerical (continuous)
1	Categorical (nominal)
Mstatus	Categorical (nominal)
Totper.	Numerical (discrete)
Adults.	Numerical (discrete)
Parent.	Categorical (nominal).
Age.	Numerical (discrete).
Education	Categorical (ordinal).
Income.	Numerical (continuous)
Hispanic	Categorical (nominal).

Variable	Type
Race.	Categorical (nominal).
Partyln.	Categorical (nominal).
Polview.	Categorical (ordinal).
Sex.	Categorical (nominal).
Religion	Categorical (nominal).
Date.	Numerical (discrete).
Q1.	Categorical (nominal).
Q2	Categorical (ordinal)
Q3a	Categorical (nominal).
Q3b	Categorical (nominal).
Q4	Categorical (ordinal).
Q5a	Categorical (nominal)
Q5b	Categorical (nominal).
Q5c	Categorical (nominal).
Q5d	Categorical (nominal).
Q5e	Categorical (nominal).
Q5f	Categorical (nominal).
Q6	Categorical (nominal).
Country2	Categorical (nominal).
servey.	Numerical (discrete).

Code to find the count of 'refused' and 'NA' values in each column

```
library(readxl)
df <- read_excel("Mini_Group_Project_1.xlsx")

for (i in 1: ncol(df)){
  a = as.integer(table(df[i])["Refused"])
  b = as.integer(table(df[i])["NA"])

  print(c(i, a, b))
}
```

```
[1]  1 NA NA
[1]  2 NA NA
[1]  3 NA NA
[1]  4  6 NA
[1]  5 11 NA
[1]  6 11 NA
[1]   7   1 796
```

```
[1]  8 23 NA
[1]  9  5 NA
[1] 10 65 NA
[1] 11  9 NA
[1] 12 21 NA
[1]  13  NA 634
[1] 14 19 NA
[1] 15 NA NA
[1] 16 22 NA
[1] 17 NA NA
[1] 18 NA NA
[1] 19 NA NA
[1] 20 NA NA
[1] 21 NA NA
[1] 22 NA NA
[1] 23 NA NA
[1] 24 NA NA
[1] 25 NA NA
[1] 26 NA NA
[1] 27 NA NA
[1] 28 NA NA
[1] 29 NA NA
[1] 30 NA NA
[1] 31 NA NA
```

Missing Values:

Variable	Refused	NA
Case_id	0	0
123	0	0
1	0	0
Mstatus	6	0
Totper.	11	0
Adults.	11	0
Parent.	1	796
Age.	23	0
Education	5	0
Income.	65	0
Hispanic	9	0
Race.	21	0
Partyln.	0	634
Polview.	19	0

Variable	Refused	NA
Sex.	0	0
Religion	22	0
Date.	0	0
Q1.	0	0
Q2	0	0
Q3a	0	0
Q3b	0	0
Q4	0	0
Q5a	0	0
Q5b	0	0
Q5c	0	0
Q5d	0	0
Q5e	0	0
Q5f	0	0
Q6	0	0
Country2	0	0
servey.	0	0

2 Task 2: Data Imputation

Code to replace 'Refused' and 'NA' values with mode of the column for categorical variables

```
df <- read_excel("Mini_Group_Project_1.xlsx")

calc_mode <- function(x){

  distinct_values <- unique(x)

  distinct_tabulate <- tabulate(match(x, distinct_values))

  distinct_values[which.max(distinct_tabulate)]
}

for(i in list(3, 4, 7, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,
  dt <- df[, i]
  dt_vec <- unlist(dt)
  df[,i]<-replace(df[,i], (df[,i] == 'Refused' | df[, i] == "NA" | df[, i]=="DK/Refused"),
}
```

Code to replace 'Refused' and 'NA' values with mode of the column for numerical variables
Finding the standard statistical properties of numerical columns like mean, median, etc.

```
# Statistics of numerical variable "Weight"
df$weight <- as.numeric(df$weight)
min(df$weight)
```

```
[1] 0.25
```

```
max(df$weight)
```

```
[1] 3.9598
```

```
mean(df$weight)
```

```
[1] 1
```

```
median(df$weight)
```

```
[1] 0.82
```

```
sd(df$weight)
```

```
[1] 0.6364925
```

```
IQR(df$weight)
```

```
[1] 0.72455
```

```
# Statistics of numerical variable "age"
df$age <- as.numeric(df$age)
min(df$age)
```

```
[1] 18
```

```r
max(df$age)
```

```
[1] 96
```

```r
mean(df$age)
```

```
[1] 49.83019
```

```r
median(df$age)
```

```
[1] 49
```

```r
sd(df$age)
```

```
[1] 16.78972
```

```r
IQR(df$age)
```

```
[1] 26
```

Statistical Properties of numerical columns

Variable	Min.	Max	Mean	Median	Standard Deviation	Interquartile range
Weight	0.25	3.9598	1	0.82	0.6264925	0.72455
Age	18	96	49.83019	49	16.78972	26

```r
for(i in list(3, 4, 7, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,
  a = names(which.min(table(df[, i])))
  b = names(which.max(table(df[, i])))

  print(c(i, a, b))
}
```

```
[1] "3"  "AK" "CA"
[1] "4"         "Separated" "Married"
[1] "7"  "No" "NA"
[1] "9"
[2] "Don't know"
[3] "Four year college or university degree/Bachelor.s degree (e.g., BS, BA, AB)"
[1] "11"         "Don't Know" "No"
[1] "12"                 "Other Race"         "White Non-Hispanic"
[1] "13"                         "Neither/Other (DO NOT READ)"
[3] "NA"
[1] "14"         "Don't know" "Moderate"
[1] "15"     "Female" "Male"
[1] "16"                       "Don't know"
[3] "Catholic, Roman Catholic"
[1] "18"              "Australia"      "United Kingdom"
[1] "19"            "Very bad"      "Somewhat good"
[1] "20"
[2] "VOL: Neither"
[3] "Having a close relationship to Germany"
[1] "21"
[2] "VOL: Neither"
[3] "Having a close relationship to Germany"
[1] "22"              "Very unlikely"   "Somewhat likely"
[1] "23"                "No, not a partner" "Yes, as a partner"
[1] "24"                "No, not a partner" "Yes, as a partner"
[1] "25"                "No, not a partner" "Yes, as a partner"
[1] "26"                "No, not a partner" "Yes, as a partner"
[1] "27"                "No, not a partner" "Yes, as a partner"
[1] "28"                "No, not a partner" "Yes, as a partner"
[1] "29"
[2] "Countries will cooperate more with other countries"
[3] "Everything will be the same as before the crisis"
[1] "30"            "United States" "United States"
```

Variable	State	Mstatus	Parent	Educ	Hispanic	Race	Partyln	Polview	sex	religion
Min	AK	Separated	No	Don't Know	Don't Know	Other Race	Neither	Don't Know	Female	Don't Know
Max	CA	Married	NA	4 Years...	No	White Non-Hispanic	NA	Moderate	Male	Catholic

3 Task 3: Exploratory Data Analysis

3.1 Part 1:

```
library(ggplot2)

df <- read_excel("Mini_Group_Project_1.xlsx")

df_new<-df[df[, 14]=="Moderate" & df[, 15] == "Female" &
         df$Q3b != "DK/Refused" & df$Q3b != "Both relationships are equally important"
       & df$Q3b != "VOL: Neither",]

df_new$age <- as.numeric(df_new$age)
```

```
Warning: NAs introduced by coercion
```

```
dt1 <- df_new[df_new[, 8] < 30,]
dt2 <- df_new[(df_new[, 8] < 50 & df_new[, 8] > 29), ]
dt3 <- df_new[(df_new[, 8] < 65 & df_new[, 8] > 49), ]
dt4 <- df_new[df_new[, 8] > 64 ,]

table(dt1$Q3b)
```

```
Having a close relationship to China Having a close relationship to Germany
                                   9                                      10
```

```
table(dt2$Q3b)
```

```
Having a close relationship to China Having a close relationship to Germany
                                  36                                      30
```

```
table(dt3$Q3b)
```

```
Having a close relationship to China Having a close relationship to Germany
                                  12                                      20
```

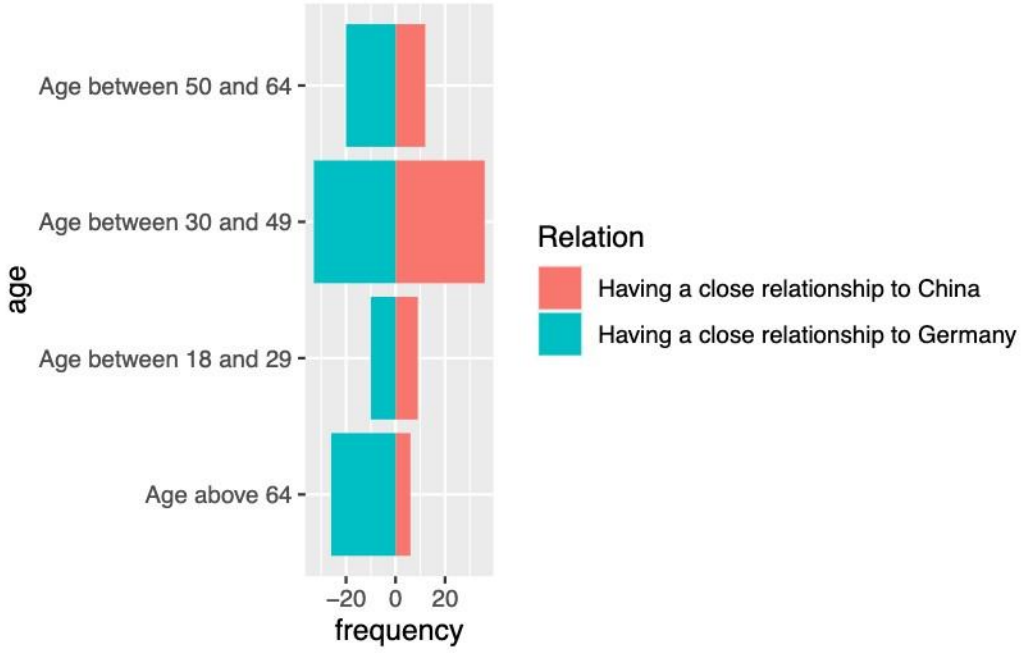
```r
table(dt4$Q3b)
```

```
Having a close relationship to China Having a close relationship to Germany
                                   6                                      26
```

```r
Relation = c("Having a close relationship to China", "Having a close relationship to Germa
frequency = c(9, -10, 36, -33, 12, -20, 6, -26)
age = c("Age between 18 and 29", "Age between 18 and 29", "Age between 30 and 49", "Age be

d <- data.frame(Relation, frequency, age)

ggplot(d, aes(x = age, y = frequency, fill = Relation))+
  geom_bar(stat = "identity")+
  coord_flip()
```



3.2 Part 2:

From the graph, we can infer that as older females prefer having a closer relationship with Germany over China, while in younger females, the ratio is more even. So, we can see a similar pattern as in the graph of Younger Americans.

4 Task 4: Visual Analysis

4.1 Part 1:

```
q1 = df[df$Q5a == 'Yes, as a partner',]
q2 = df[df$Q5b == 'Yes, as a partner',]
q3 = df[df$Q5c == 'Yes, as a partner',]
q4 = df[df$Q5d == 'Yes, as a partner',]
q5 = df[df$Q5e == 'Yes, as a partner',]
q6 = df[df$Q5f == 'Yes, as a partner',]
tot_f = nrow(df[df$sex == 'Female',])
tot_m = nrow(df[df$sex == 'Male',])
table(q1$sex)
```

```
Female   Male
   376    401
```

```
table(q2$sex)
```

```
Female   Male
   292    312
```

```
table(q3$sex)
```

```
Female   Male
   270    303
```

```
table(q4$sex)
```

```
Female   Male
   385    410
```

```
table(q5$sex)
```

```
Female   Male
   396    414
```

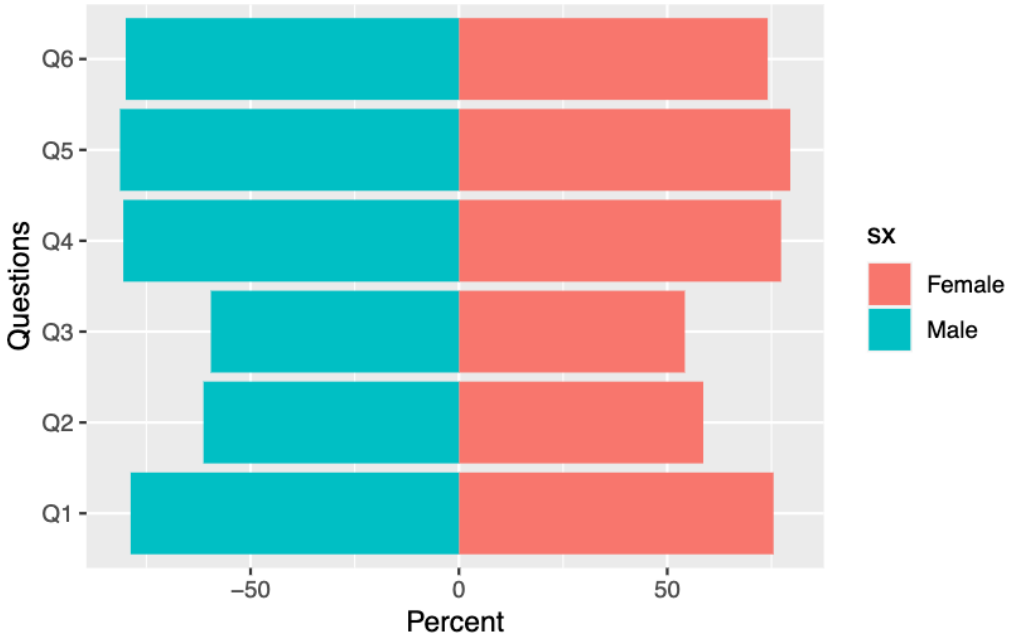
```r
table(q6$sex)
```

```
Female   Male
   369    407
```

```r
Questions = c('Q1', 'Q1', 'Q2','Q2', 'Q3','Q3', 'Q4', 'Q4', 'Q5', 'Q5', 'Q6', 'Q6')
Percent = c(c(as.numeric(table(q1$sex)['Female']/tot_f*100),
              -as.numeric(table(q1$sex)['Male']/tot_m*100)),
            c(as.numeric(table(q2$sex)['Female']/tot_f*100),
              -as.numeric(table(q2$sex)['Male']/tot_m*100)),
            c(as.numeric(table(q3$sex)['Female']/tot_f*100),
              -as.numeric(table(q3$sex)['Male']/tot_m*100)),
            c(as.numeric(table(q4$sex)['Female']/tot_f*100),
              -as.numeric(table(q4$sex)['Male']/tot_m*100)),
            c(as.numeric(table(q5$sex)['Female']/tot_f*100),
              -as.numeric(table(q5$sex)['Male']/tot_m*100)),
            c(as.numeric(table(q6$sex)['Female']/tot_f*100),
                -as.numeric(table(q6$sex)['Male']/tot_m*100))
)
sx = c('Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female', '

d <- data.frame(Questions, Percent, sx)

ggplot(d, aes(x = Questions, y = Percent, fill = sx))+
  geom_bar(stat = "identity")+
  coord_flip()
```



4.2 Part 2:

Looking at the graph, we can see that percent of males who think Germany as a partner on key issues is more than female percent in all the six questions, although they differ by a very small amount. So we can conclude that sex isn't a major factor in affecting the data.