EDA Vera

2024-05-14

```
load("Irish.RData")
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- as.data.frame(Irish[2])[,2:6]
```

```
summary(df)
```

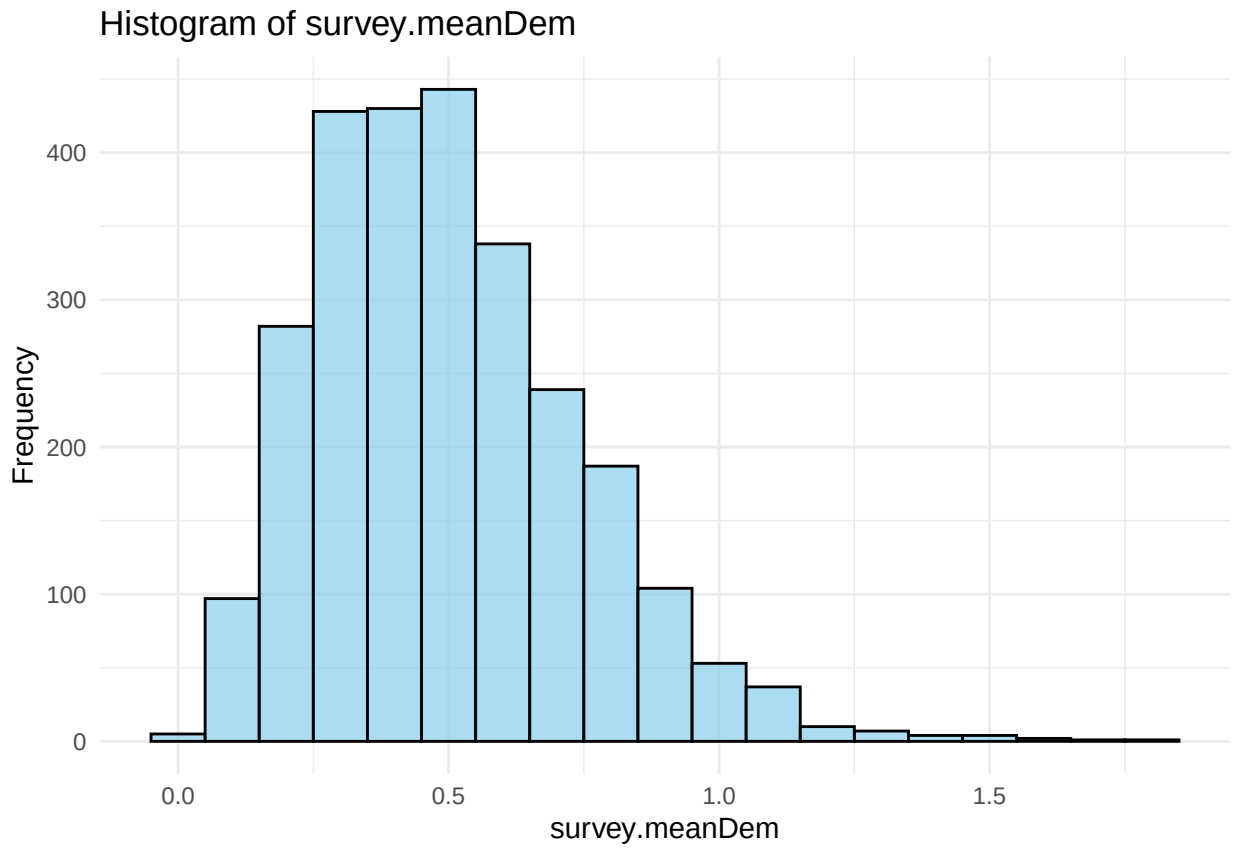
```
##  survey.meanDem    survey.SOCIALCLASS survey.OWNERSHIP   survey.BUILT.YEAR
##  Min.   :0.02032   AB: 410            Length:2672        Min.   :1674
##  1st Qu.:0.31820   C1: 730            Class :character   1st Qu.:1962
##  Median :0.46698   C2: 449            Mode  :character   Median :1979
##  Mean   :0.49938   DE:1018                               Mean   :1972
##  3rd Qu.:0.64220   F :  65                               3rd Qu.:1997
##  Max.   :1.75077                                         Max.   :2008
##  survey.HEAT.HOME
##  Length:2672
##  Class :character
##  Mode  :character
##
##
##
```

Null vals:

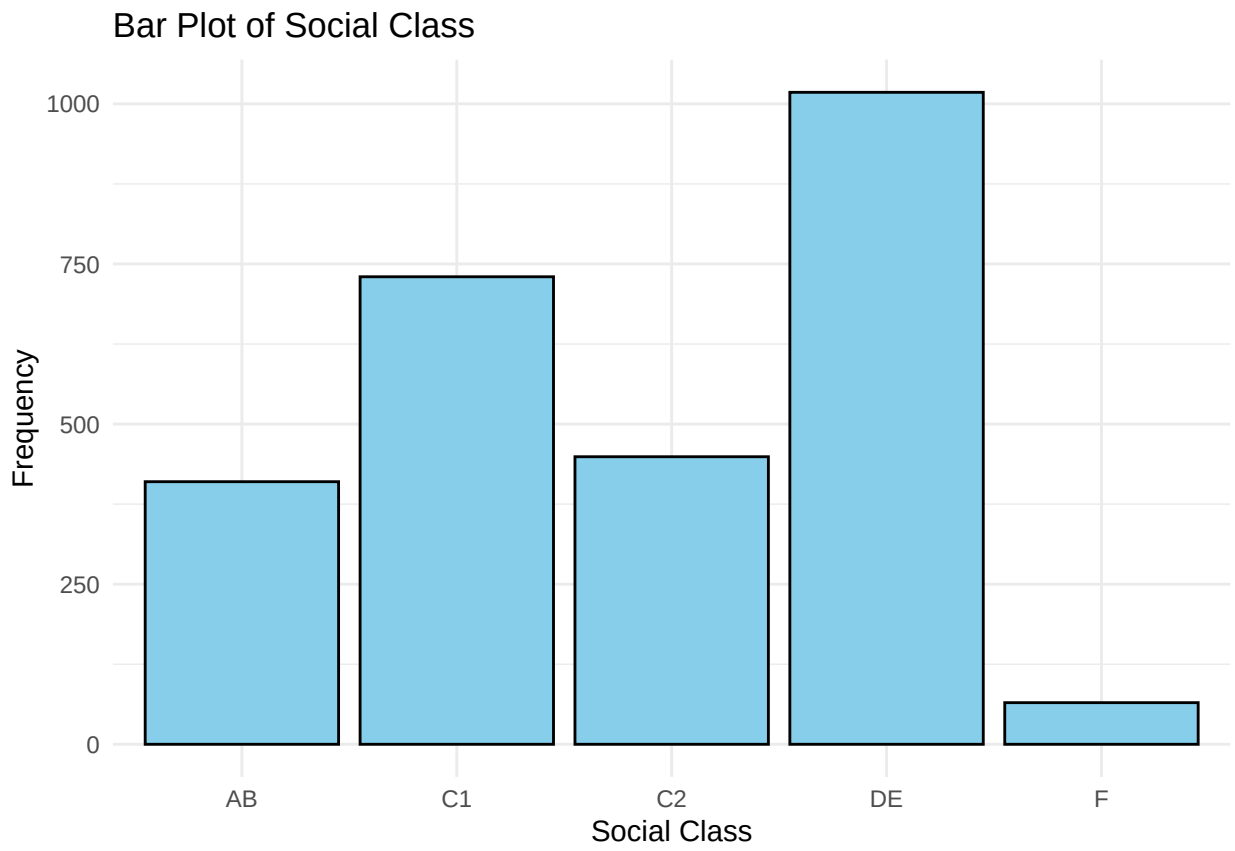
```
colSums(is.na(df))
```

```
##     survey.meanDem survey.SOCIALCLASS   survey.OWNERSHIP  survey.BUILT.YEAR
##                  0                  0                  0                  0
##   survey.HEAT.HOME
##                  0
```

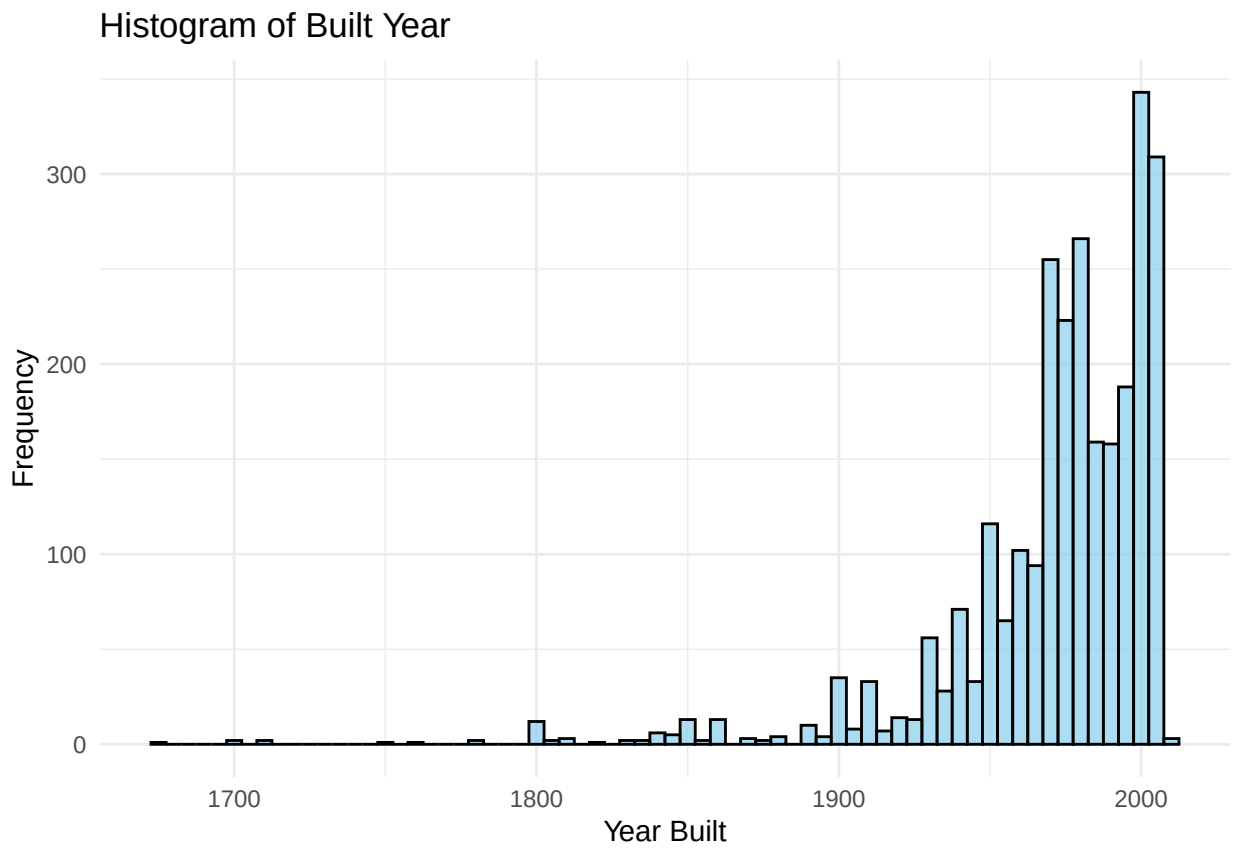
```r
library(ggplot2)

# Histogram of survey.meanDem
ggplot(df, aes(x=survey.meanDem)) +
  geom_histogram(binwidth=0.1, fill="skyblue", color="black", alpha=0.7) +
  theme_minimal() +
  labs(title="Histogram of survey.meanDem", x="survey.meanDem", y="Frequency")
```



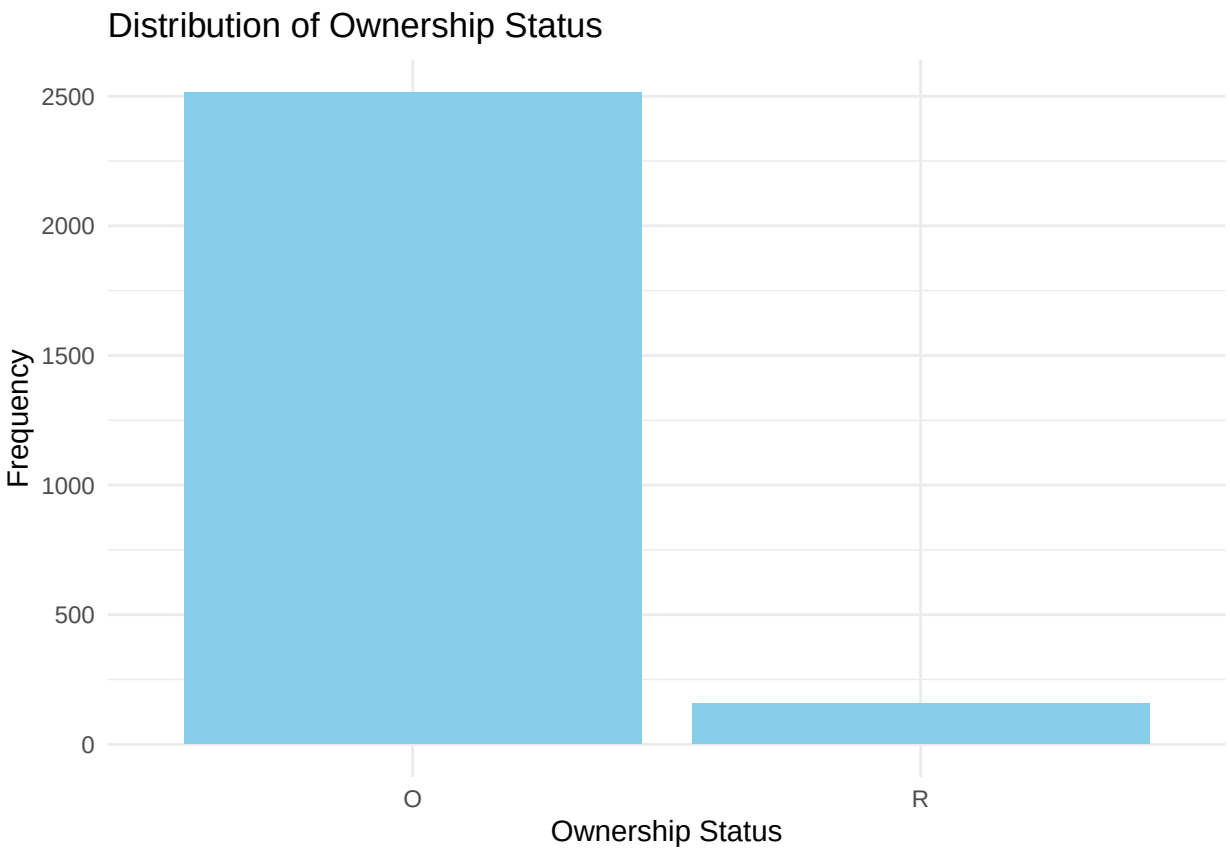

```r
# Bar plot for survey.SOCIALCLASS
ggplot(df, aes(x=survey.SOCIALCLASS)) +
  geom_bar(fill="skyblue", color="black") +
  theme_minimal() +
  labs(title="Bar Plot of Social Class", x="Social Class", y="Frequency")
```



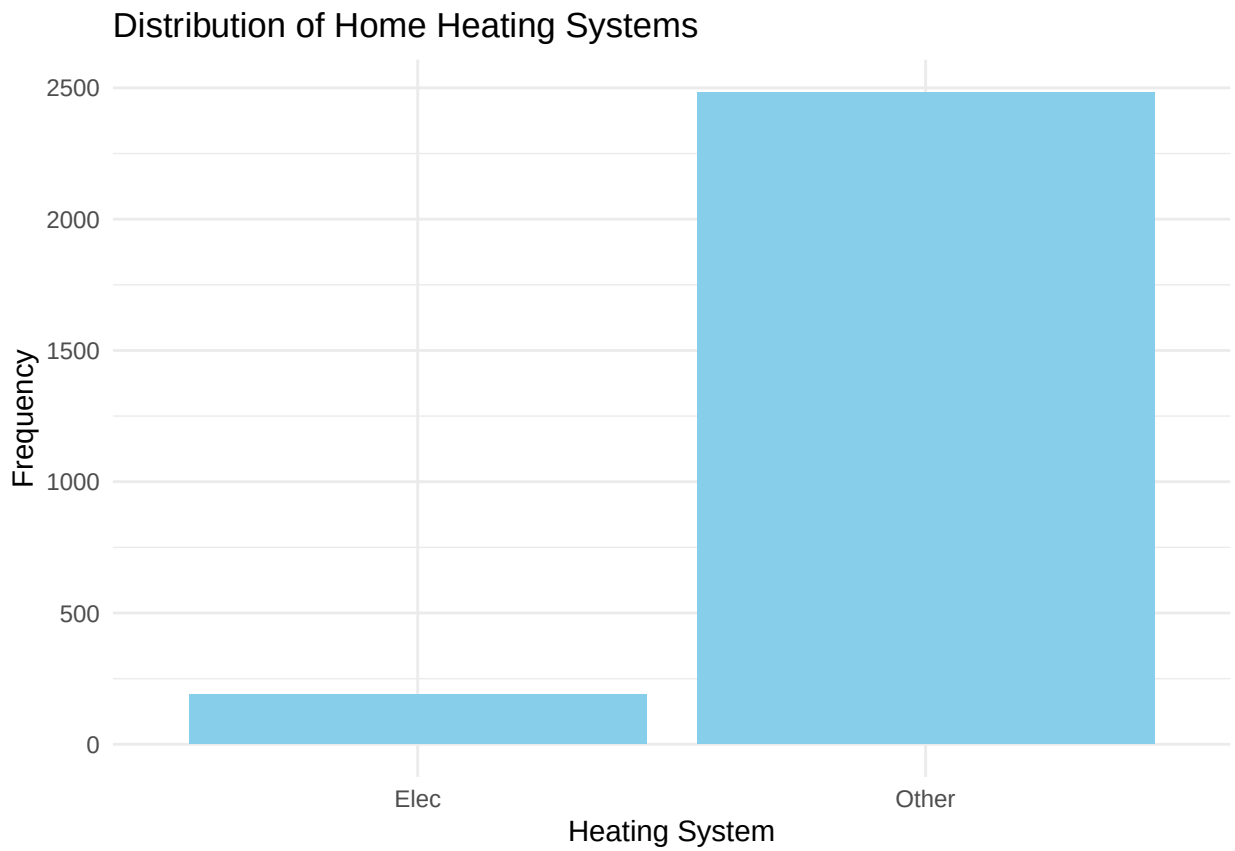

```
# Histogram of survey.BUILT.YEAR
ggplot(df, aes(x=survey.BUILT.YEAR)) +
  geom_histogram(binwidth=5, fill="skyblue", color="black", alpha=0.7) +
  theme_minimal() +
  labs(title="Histogram of Built Year", x="Year Built", y="Frequency")
```




```
# Bar plot for survey.OWNERSHIP
ggplot(data=df, aes(x=survey.OWNERSHIP)) +
  geom_bar(fill="skyblue") +
  theme_minimal() +
  labs(title="Distribution of Ownership Status", x="Ownership Status", y="Frequency")
```



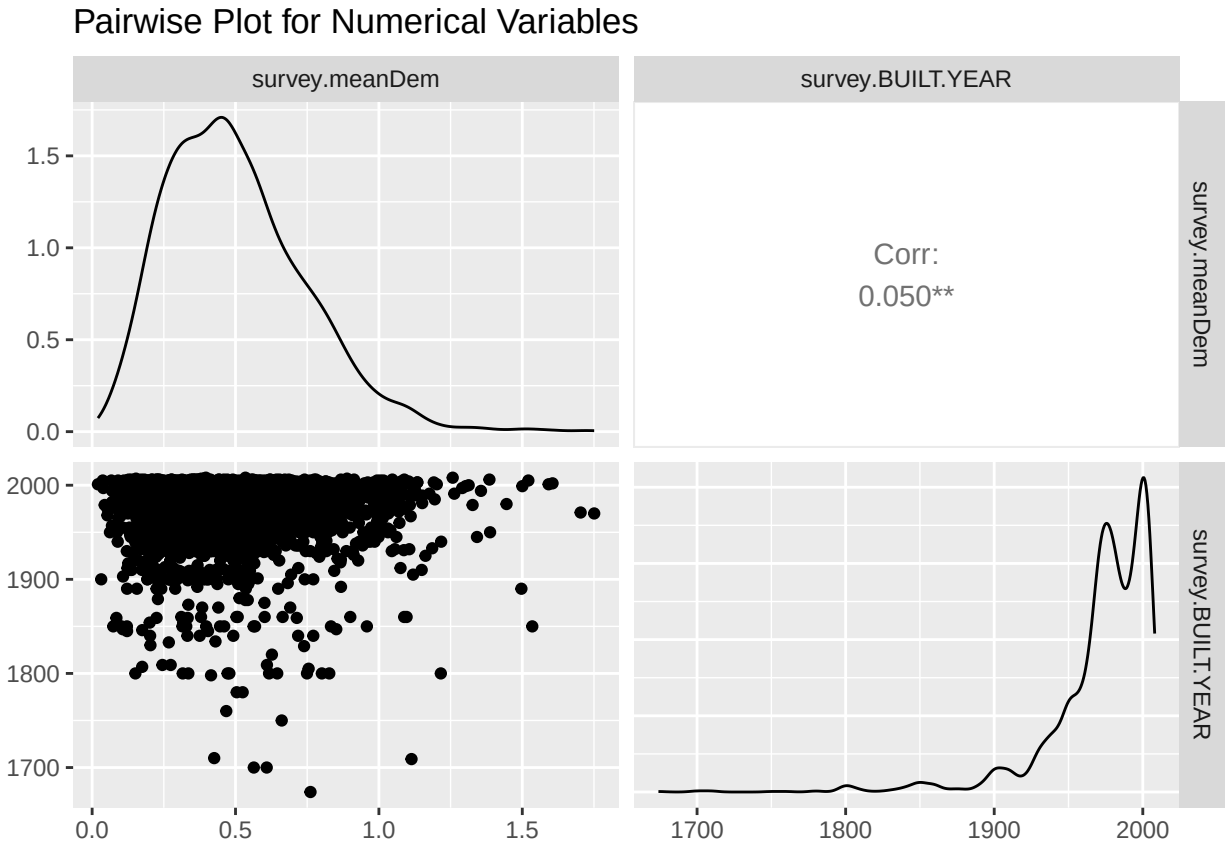

```
# Bar plot for survey.HEAT.HOME
ggplot(data=df, aes(x=survey.HEAT.HOME)) +
  geom_bar(fill="skyblue") +
  theme_minimal() +
  labs(title="Distribution of Home Heating Systems", x="Heating System", y="Frequency")
```



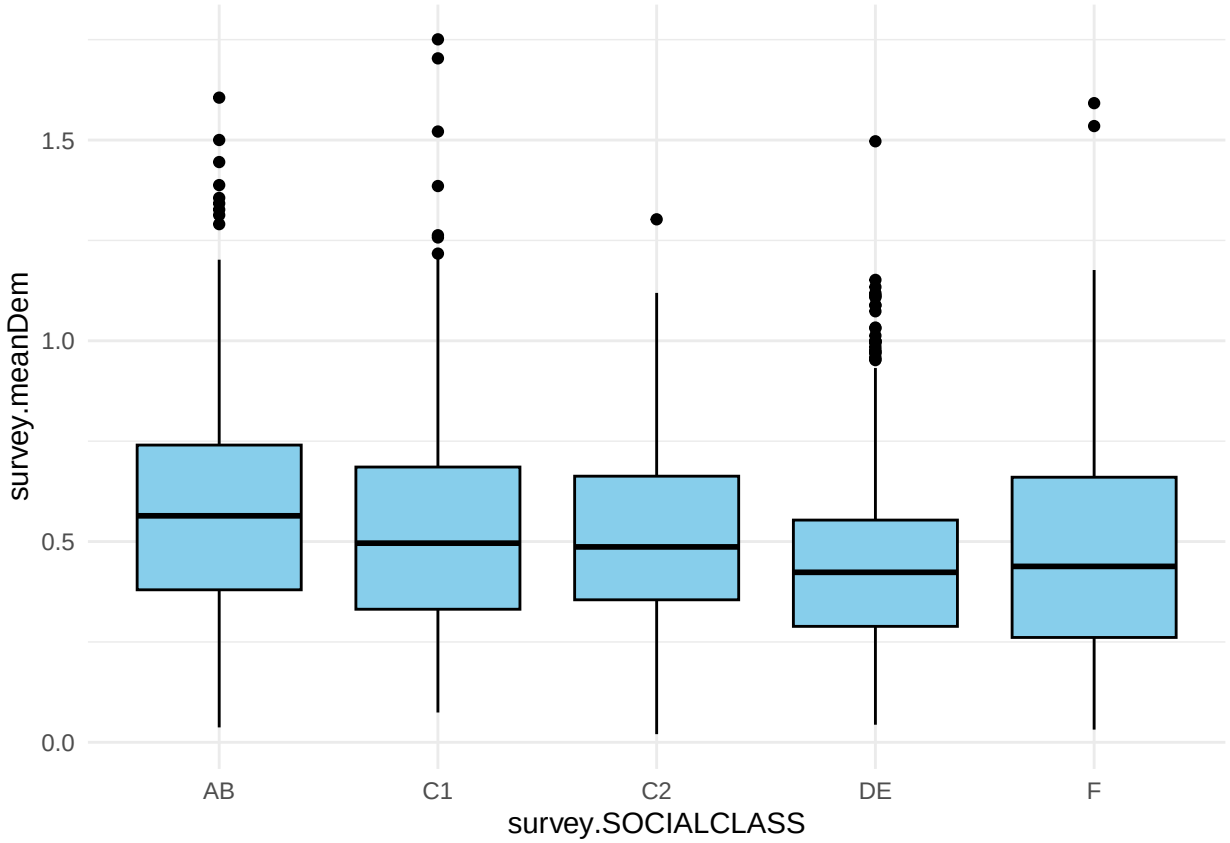

```
library(GGally)  # For pairwise plots
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

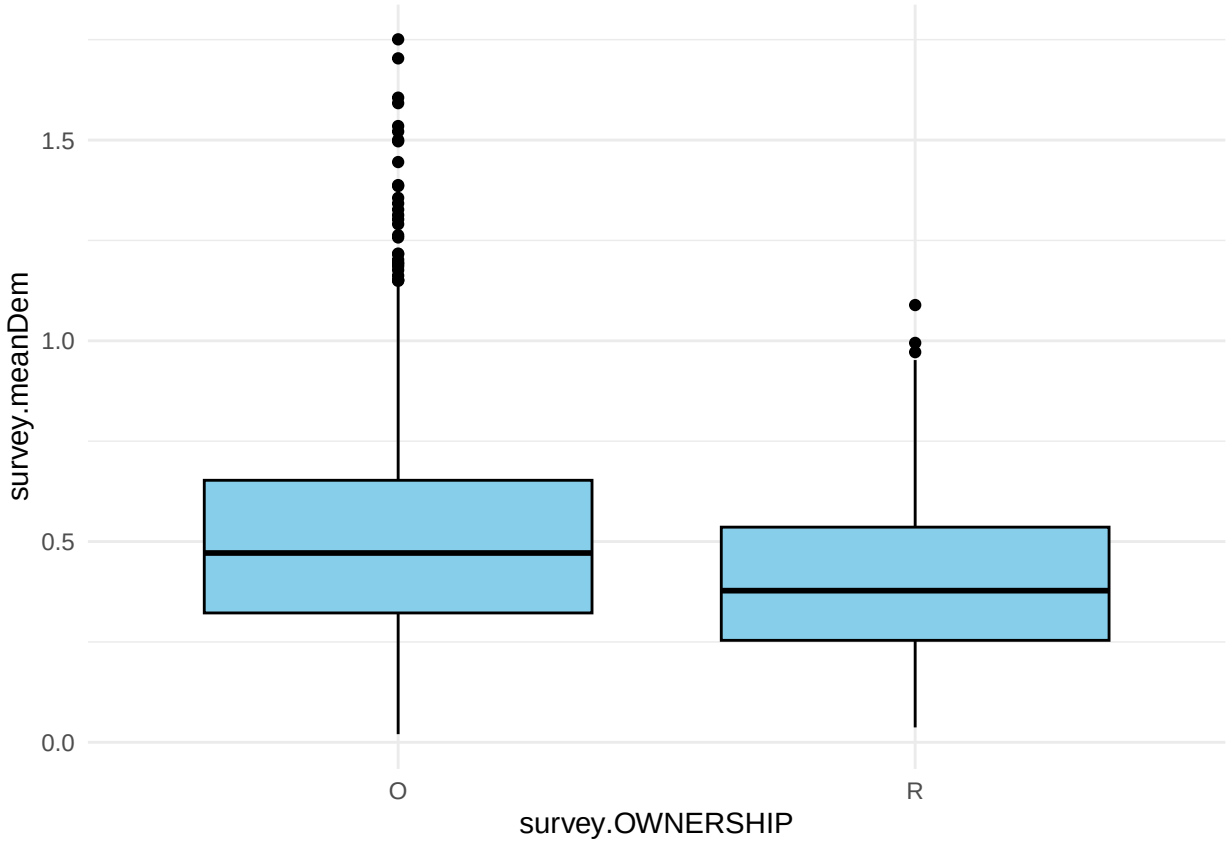
```
# Subset the dataframe to include only numerical variables
numerical_df <- df[, c("survey.meanDem", "survey.BUILT.YEAR")]

# Pairwise plot for numerical variables
ggpairs(numerical_df,
        title = "Pairwise Plot for Numerical Variables")
```



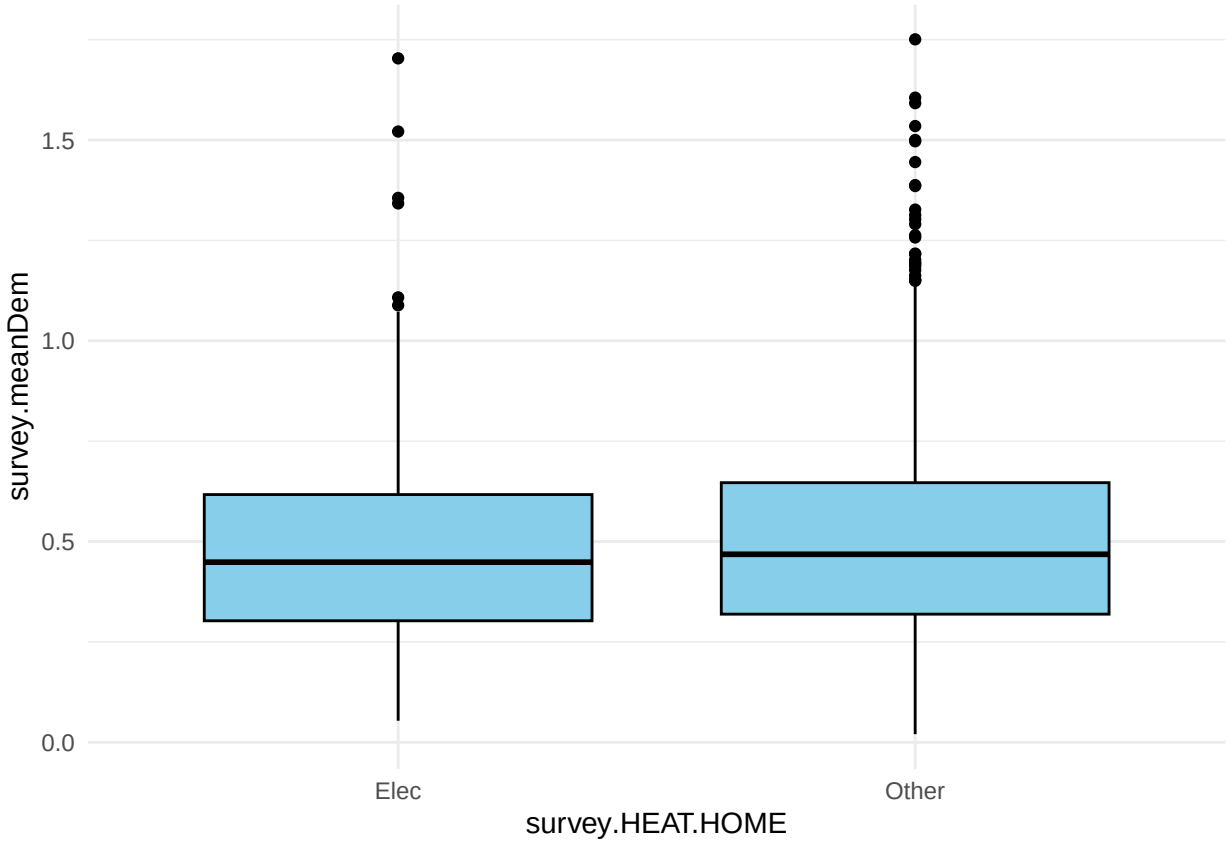

```
# Box plot for survey.meanDem vs survey.SOCIALCLASS
ggplot(df, aes(x=survey.SOCIALCLASS, y=survey.meanDem)) +
  geom_boxplot(fill="skyblue", color="black") +
  theme_minimal()
```




```
# Box plot for survey.meanDem vs survey.OWNERSHIP
ggplot(df, aes(x=survey.OWNERSHIP, y=survey.meanDem)) +
  geom_boxplot(fill="skyblue", color="black") +
  theme_minimal()
```




```
# Box plot for survey.meanDem vs survey.HEAT.HOME
ggplot(df, aes(x=survey.HEAT.HOME, y=survey.meanDem)) +
  geom_boxplot(fill="skyblue", color="black") +
  theme_minimal()
```


survey.meanDem
1.5
1.0
0.5
0.0
Elec
Other
survey.HEAT.HOME