Milestone_2

Group_D

2024-10-08

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
CustData2 <- read.csv("~/R scripts/CustData2.csv", sep=";")
#View(CustData2)

library(tinytex)

#  Split the combined column into separate columns using commas as the
delimiter
split_data <-
strsplit(CustData2$year_of_birth.marital_status.street_address.postal_code.ci
ty.state_province.Country_id.phone_number.email.Education.Occupation.househol
d_size.yrs_residence, ",")

#  Convert the list of split values into a data frame
split_data_df <- do.call(rbind, split_data)

## Warning in (function (..., deparse.level = 1) : number of columns of
result is
## not a multiple of vector length (arg 1)

#  Convert the split data to a data frame with the correct column names
split_data_df <- as.data.frame(split_data_df, stringsAsFactors = FALSE)
colnames(split_data_df) <- c("year_of_birth", "marital_status",
"street_address", "postal_code", "city","state_province", "Country_id",
"phone_number", "email","Education", "Occupation", "household_size",
"yrs_residence")

#  Combine the split columns with the original dataset
# Make sure that the 'Annual.Salary' column is preserved and not overwritten
CustData2_cleaned <- cbind(CustData2, split_data_df)


# Remove commas from 'Annual.Salary' and 'yrs_residence' columns before
conversion
```

```r
CustData2_cleaned$Annual.Salary <- gsub(",", "",
CustData2_cleaned$Annual.Salary)
CustData2_cleaned$Gross.Pay.Last.Paycheck <- gsub(",", "",
CustData2_cleaned$Gross.Pay.Last.Paycheck)
CustData2_cleaned$Gross.Year.To.Date <- gsub(",", "",
CustData2_cleaned$Gross.Year.To.Date)
CustData2_cleaned$Gross.Year.To.Date...FRS.Contribution <- gsub(",", "",
CustData2_cleaned$Gross.Year.To.Date...FRS.Contribution)
CustData2_cleaned$yrs_residence <- gsub(",", "",
CustData2_cleaned$yrs_residence)

#  Convert chr columns to numeric
CustData2_cleaned$Annual.Salary <-
as.numeric(CustData2_cleaned$Annual.Salary)
CustData2_cleaned$Gross.Pay.Last.Paycheck <-
as.numeric(CustData2_cleaned$Gross.Pay.Last.Paycheck)
CustData2_cleaned$Gross.Year.To.Date <-
as.numeric(CustData2_cleaned$Gross.Year.To.Date)
CustData2_cleaned$Gross.Year.To.Date...FRS.Contribution <-
as.numeric(CustData2_cleaned$Gross.Year.To.Date...FRS.Contribution)
CustData2_cleaned$yrs_residence <-
as.numeric(CustData2_cleaned$yrs_residence)
CustData2_cleaned$household_size <-
as.numeric(CustData2_cleaned$household_size)

## Warning: NAs introduced by coercion

CustData2_cleaned$year_of_birth <-
as.numeric(CustData2_cleaned$year_of_birth)



#  Removing the last column from the dataset(added a column by mistake at the
end)
CustData2_cleaned <- CustData2_cleaned[, -ncol(CustData2_cleaned)]

#View(CustData2_cleaned)
head(CustData2_cleaned)

##   X Last.Name First.Name Middle.Initial                        Title
## 1 1    ALBERT    JESSICA              M         CORRECTIONAL OFFICER
## 2 2  ARGUELLO     ADRIAN              A               POLICE OFFICER
## 3 3    TUCKER      KEVIN              K         CORRECTIONAL OFFICER
## 4 4      DELL      JAMES              A         WASTE SCALE OPERATOR
## 5 5    THOMAS    MICHAEL              D RAIL VEHICLE ELECTRONIC TECH
## 6 6   QUINTAS      DAVID              F               POLICE SERGEANT
##                    Department.Name Annual.Salary Gross.Pay.Last.Paycheck
## 1    CORRECTIONS & REHABILITATION      54619.76                 2501.62
## 2                          POLICE      65250.38                 3467.63
## 3    CORRECTIONS & REHABILITATION      62393.76                 4513.71
```

```
## 4          SOLID WASTE MANAGEMENT      37735.10                 1561.67
## 5 TRANSPORTATION AND PUBLIC WORKS      64386.40                 6665.66
## 6                          POLICE      89621.22                 3802.71
##   Gross.Year.To.Date Gross.Year.To.Date...FRS.Contribution
## 1           48025.48                                46616.58
## 2           57932.07                                56222.79
## 3           49968.35                                48501.19
## 4           35469.59                                34432.85
## 5          132850.76                               128948.86
## 6           97945.90                                95047.65
##
year_of_birth.marital_status.street_address.postal_code.city.state_province.C
ountry_id.phone_number.email.Education.Occupation.household_size.yrs_residenc
e
## 1 1976,married,27 North Sagadahoc
Boulevard,60332,Ede,Gelderland,52770,519-236-
6123,Ruddy@company.com,Masters,Prof.,2,4
## 2 1964,,37 West Geneva Street,55406,Hoofddorp,Noord-Holland,52770,327-194-
5008,Ruddy@company.com,Masters,Prof.,2,4
## 3 1942,single,47 Toa Alta Road,34077,Schimmert,Limburg,52770,288-613-
9676,Ruddy@company.com,Masters,Prof.,2,4
## 4 1977,married,47 South Kanabec Road,72996,Scheveningen,Zuid-
Holland,52770,222-269-1259,Ruddy@company.com,Masters,Prof.,2,4
## 5 1949,,57 North 3rd Drive,67644,Joinville,Santa Catarina,52775,675-133-
2226,Ruddy@company.com,Masters,Prof.,2,4
## 6 1950,single,67 East Mcintosh Avenue,83786,Nagoya,Aichi,52782,183-207-
2933,Ruddy@company.com,Masters,Prof.,2,4
##   year_of_birth marital_status               street_address postal_code
## 1          1976        married 27 North Sagadahoc Boulevard       60332
## 2          1964                         37 West Geneva Street       55406
## 3          1942         single              47 Toa Alta Road       34077
## 4          1977        married        47 South Kanabec Road       72996
## 5          1949                           57 North 3rd Drive       67644
## 6          1950         single      67 East Mcintosh Avenue       83786
##           city state_province Country_id phone_number            email
## 1          Ede     Gelderland      52770 519-236-6123 Ruddy@company.com
## 2    Hoofddorp  Noord-Holland      52770 327-194-5008 Ruddy@company.com
## 3    Schimmert        Limburg      52770 288-613-9676 Ruddy@company.com
## 4 Scheveningen   Zuid-Holland      52770 222-269-1259 Ruddy@company.com
## 5    Joinville Santa Catarina      52775 675-133-2226 Ruddy@company.com
## 6       Nagoya          Aichi      52782 183-207-2933 Ruddy@company.com
##   Education Occupation household_size yrs_residence
## 1   Masters      Prof.              2             4
## 2   Masters      Prof.              2             4
## 3   Masters      Prof.              2             4
## 4   Masters      Prof.              2             4
## 5   Masters      Prof.              2             4
## 6   Masters      Prof.              2             4
```

```r
# Deleting a useless column
cust <- subset(CustData2_cleaned, select = -
year_of_birth.marital_status.street_address.postal_code.city.state_province.C
ountry_id.phone_number.email.Education.Occupation.household_size.yrs_residenc
e)

#View(cust)
head(cust)
```

```
##   X Last.Name First.Name Middle.Initial                        Title
## 1 1    ALBERT    JESSICA              M         CORRECTIONAL OFFICER
## 2 2  ARGUELLO     ADRIAN              A               POLICE OFFICER
## 3 3    TUCKER      KEVIN              K         CORRECTIONAL OFFICER
## 4 4      DELL      JAMES              A         WASTE SCALE OPERATOR
## 5 5    THOMAS    MICHAEL              D RAIL VEHICLE ELECTRONIC TECH
## 6 6   QUINTAS      DAVID              F               POLICE SERGEANT
##                   Department.Name Annual.Salary Gross.Pay.Last.Paycheck
## 1    CORRECTIONS & REHABILITATION      54619.76                 2501.62
## 2                          POLICE      65250.38                 3467.63
## 3    CORRECTIONS & REHABILITATION      62393.76                 4513.71
## 4          SOLID WASTE MANAGEMENT      37735.10                 1561.67
## 5 TRANSPORTATION AND PUBLIC WORKS      64386.40                 6665.66
## 6                          POLICE      89621.22                 3802.71
##   Gross.Year.To.Date Gross.Year.To.Date...FRS.Contribution year_of_birth
## 1           48025.48                              46616.58          1976
## 2           57932.07                              56222.79          1964
## 3           49968.35                              48501.19          1942
## 4           35469.59                              34432.85          1977
## 5          132850.76                             128948.86          1949
## 6           97945.90                              95047.65          1950
##   marital_status              street_address postal_code         city
## 1        married 27 North Sagadahoc Boulevard       60332          Ede
## 2                      37 West Geneva Street       55406    Hoofddorp
## 3         single            47 Toa Alta Road       34077    Schimmert
## 4        married       47 South Kanabec Road       72996 Scheveningen
## 5                         57 North 3rd Drive       67644    Joinville
## 6         single     67 East Mcintosh Avenue       83786       Nagoya
##   state_province Country_id phone_number            email Education
Occupation
## 1     Gelderland      52770 519-236-6123 Ruddy@company.com   Masters
Prof.
## 2  Noord-Holland      52770 327-194-5008 Ruddy@company.com   Masters
Prof.
## 3        Limburg      52770 288-613-9676 Ruddy@company.com   Masters
Prof.
## 4   Zuid-Holland      52770 222-269-1259 Ruddy@company.com   Masters
Prof.
## 5 Santa Catarina      52775 675-133-2226 Ruddy@company.com   Masters
Prof.
## 6          Aichi      52782 183-207-2933 Ruddy@company.com   Masters
```

```
Prof.
##   household_size yrs_residence
## 1              2             4
## 2              2             4
## 3              2             4
## 4              2             4
## 5              2             4
## 6              2             4
```

```
keeper <- c("year_of_birth", "marital_status", "Education", "Occupation",
                    "Annual.Salary", "Gross.Pay.Last.Paycheck",
"Gross.Year.To.Date",

"Gross.Year.To.Date...FRS.Contribution","household_size","yrs_residence")
cust <- cust[keeper]



names(cust)[6:7] <- c("Gross_Pay_Last_Paycheck","Gross_Year_To_Date")
names(cust)[8] <- "Gross_FRS_Contribution"

View(cust)

# Imputing the Marital Column
df <- cust

#table(df$marital_status)
#table(df$Education)
#table(df$Occupation)

# Cleaning the Marital status column
df$marital_status <- gsub("(?i)divorced|Divorc.|separ.","Divorced",
df$marital_status, perl = TRUE)
df$marital_status <- gsub("(?i)married|mabsent|mar-af", "Married",
df$marital_status, perl = TRUE)
df$marital_status <- gsub("(?i)widow(ed)|widow", "Widow",df$marital_status,
perl = TRUE)
df$marital_status <- gsub("(?i)NeverM|single","Single",df$marital_status,
perl = TRUE)
table(df$marital_status)
```

```
##
##          Divorced  Married   Single    Widow
##    60795     2697    56014    71184      633
```

```
#mode_m <- names(sort(table(df$marital_status), decreasing = TRUE))

df2 <- df

df2[df2 == ""] <- NA
```

```r
df2$marital_status[is.na(df2$marital_status)] <- "Unknown"
table(df2$marital_status)
```

```
##
## Divorced  Married   Single  Unknown    Widow
##     2697    56014    71184    60795      633
```

```r
#table(cust$marital_status)
#table(df2$marital_status)

# Imputing the Education column
df3 <- df2
df3$Education <- gsub("[[:alnum:]._%+-]+@[[:alnum:].-]+\\.[a-zA-Z]{2,}",
"",df3$Education)

# Finding the mode of the column Education
mode <- names(sort(table(df3$Education), decreasing = TRUE))[1]
mode
```

```
## [1] "Bach."
```

```r
df3[df3 == ""] <- NA
df3$Education[is.na(df3$Education)] <- mode

table(df3$Education)
```

```
##
##   Bach. HS-grad Masters
##   81045   55139   55139
```

```r
#write.csv(cust, "Splitted data.csv", row.names = TRUE)

# Imputing the Occupation column
df3$Occupation <- gsub("Bach.|HS-grad|Masters", "", df3$Occupation)

mode_O <- names(sort(table(df3$Occupation), decreasing = TRUE))[1]
mode_O
```

```
## [1] "Cleric."
```

```r
df3[df3 == ""] <- NA
sum(is.na(df3$Occupation))
```

```
## [1] 1228
```

```r
df3$Occupation[is.na(df3$Occupation)] <- mode_O
table(df3$Occupation)
```

```
##
## Cleric.   Exec.   Prof.   Sales
##   56367   24678   55139   55139
```

```r
# Dealing with missing numerical data and imputing it
df4 <- df3
table(df4$household_size)
```

```
##
##      2      3
## 165417  24678
```

```r
sum(is.na(df4$household_size))
```

```
## [1] 1228
```

```r
# Using the median to impute data
med_h <- median(df4$household_size, na.rm = TRUE)
med_h
```

```
## [1] 2
```

```r
df4$household_size[is.na(df4$household_size)] <- med_h
table(df4$household_size)
```

```
##
##      2      3
## 166645  24678
```

```r
# Checking if we still have missing values
missings <- colSums(is.na(df4))
missings
```

```
##           year_of_birth          marital_status               Education
##                       0                       0                       0
##              Occupation           Annual.Salary Gross_Pay_Last_Paycheck
##                       0                       6                       6
##      Gross_Year_To_Date  Gross_FRS_Contribution          household_size
##                       6                       6                       0
##           yrs_residence
##                       0
```

```r
# Imputing numerical data with the mean of each respective column
mean_a <- mean(df4$Annual.Salary, na.rm = TRUE)
mean_a
```

```
## [1] 63932.63
```

```r
mean_GPL <- mean(df4$Gross_Pay_Last_Paycheck, na.rm = TRUE)
mean_GPL
```

```
## [1] 2868.06
```

```r
mean_GYT <- mean(df4$Gross_Year_To_Date, na.rm = TRUE)
mean_GYT
```

```
## [1] 57923.08
```

```r
mean_FRS <- mean(df4$Gross_FRS_Contribution, na.rm = TRUE)
mean_FRS
```

```
## [1] 56379.05
```

```r
df4$Annual.Salary[is.na(df4$Annual.Salary)] <- mean_a
df4$Gross_Pay_Last_Paycheck[is.na(df4$Gross_Pay_Last_Paycheck)] <- mean_GPL
df4$Gross_Year_To_Date[is.na(df4$Gross_Year_To_Date)] <- mean_GYT
df4$Gross_FRS_Contribution[is.na(df4$Gross_FRS_Contribution)] <- mean_FRS

# Checking if there is still any missing data
last_check <- colSums(is.na(df4))
last_check
```

```
##           year_of_birth          marital_status               Education
##                       0                       0                       0
##              Occupation           Annual.Salary Gross_Pay_Last_Paycheck
##                       0                       0                       0
##      Gross_Year_To_Date  Gross_FRS_Contribution          household_size
##                       0                       0                       0
##           yrs_residence
##                       0
```

```r
# Dealing with Duplicates
df5 <- df4

# Viewing the duplicate data
dup <- df5[duplicated(df5), ]
table(duplicated(df5))
```

```
##
##  FALSE   TRUE
## 190252   1071
```

```r
#View(dup)

# Getting rid of the duplicated data
df6 <- df5[!duplicated(df5), ]

dup2 <- df6[duplicated(df6), ]
table(duplicated(df6))
```

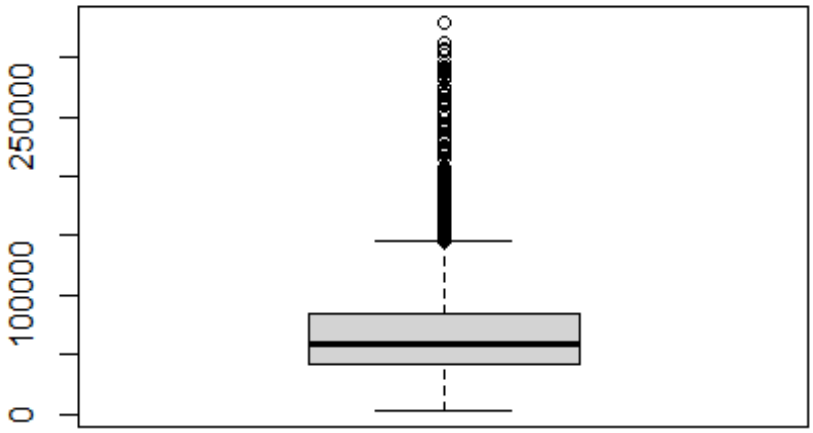
```
##
##  FALSE
## 190252
```

```r
# Deleting the Outliers
# Function to calculate and remove outliers using the IQR method

df7 <- df6
#summary(df6)
```

```r
#outliers <- boxplot.stats(df7$Annual.Salary)$out
#print(outliers)


#sum(is.na(df6$Annual.Salary))

#sum(is.na(df7$Annual.Salary))
# Remove outliers based on 1.5*IQR rule
Q1 <- quantile(df7$Annual.Salary, 0.30)
Q3 <- quantile(df7$Annual.Salary, 0.70)
IQR <- Q3 - Q1

# Define the lower and upper bounds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filter out outliers
df7_clean <- df7[df7$Annual.Salary >= lower_bound & df7$Annual.Salary <=
upper_bound, ]


# Count outliers for each column
summary(df7$Annual.Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2756   42537   58987   63940   83850  329680
```

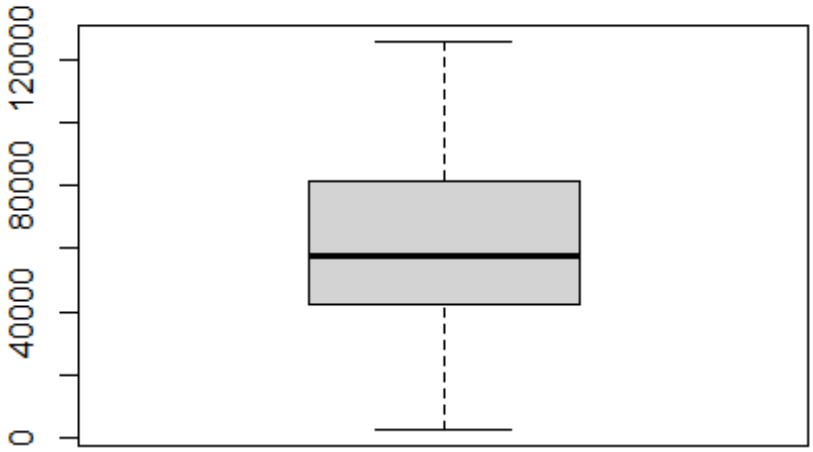
```r
boxplot(df7$Annual.Salary)
```



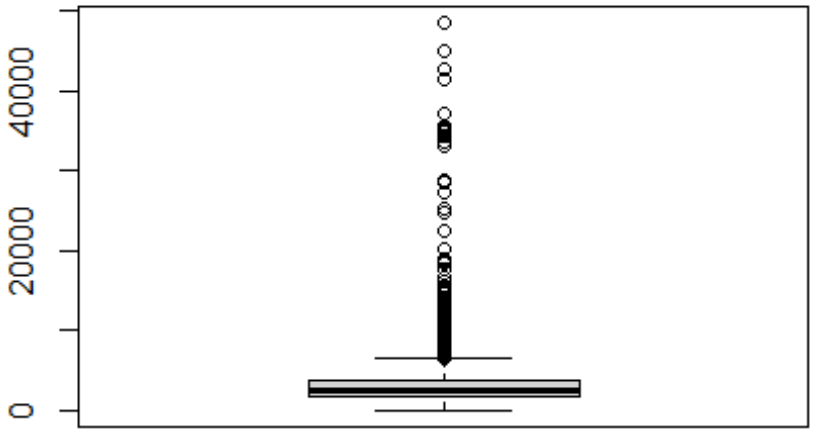

```r
summary(df7_clean$Annual.Salary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2756   42146   57767   61051   81335  125633
```

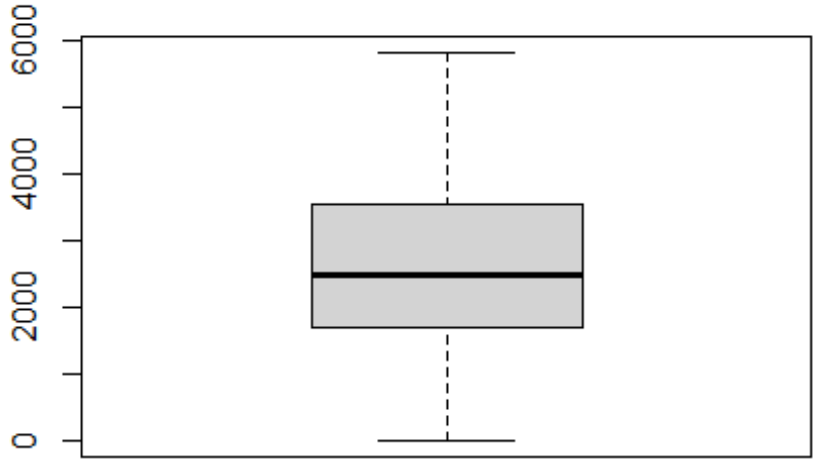
```r
boxplot(df7_clean$Annual.Salary)
```



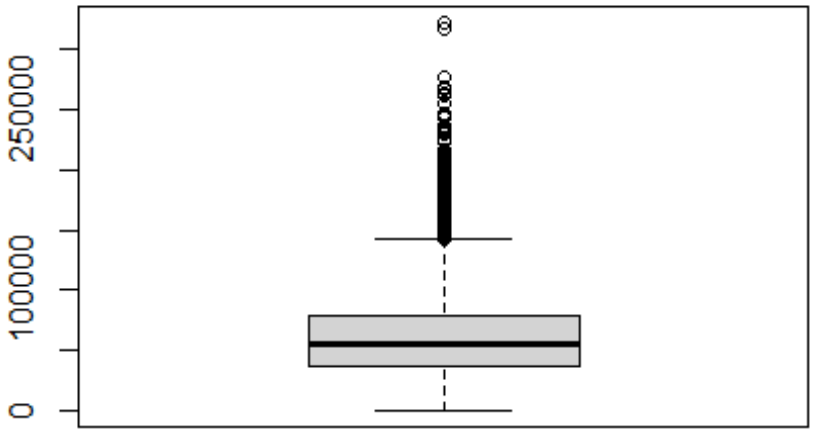

```r
Q1 <- quantile(df7$Gross_Pay_Last_Paycheck, 0.30)
Q3 <- quantile(df7$Gross_Pay_Last_Paycheck, 0.70)
IQR <- Q3 - Q1

# Define the lower and upper bounds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filter out outliers
df7_clean <- df7[df7$Gross_Pay_Last_Paycheck >= lower_bound &
df7$Gross_Pay_Last_Paycheck <= upper_bound, ]

summary(df7$Gross_Pay_Last_Paycheck)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   -11.33  1740.11  2583.10  2869.37  3683.72 48530.27

boxplot(df7$Gross_Pay_Last_Paycheck)
```




```
summary(df7_clean$Gross_Pay_Last_Paycheck)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -11.33 1698.70 2477.45 2603.56 3523.86 5810.12

boxplot(df7_clean$Gross_Pay_Last_Paycheck)
```



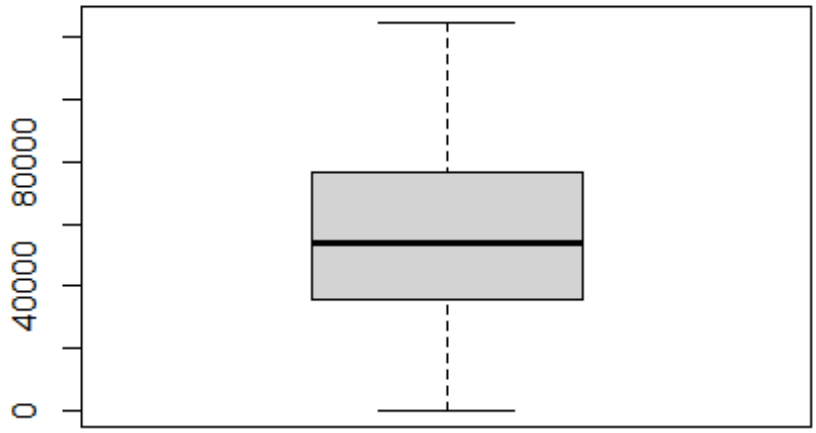

```r
Q1 <- quantile(df7$Gross_Year_To_Date, 0.30)
Q3 <- quantile(df7$Gross_Year_To_Date, 0.70)
IQR <- Q3 - Q1

# Define the lower and upper bounds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filter out outliers
df7_clean <- df7[df7$Gross_Year_To_Date >= lower_bound &
df7$Gross_Year_To_Date <= upper_bound, ]

summary(df7$Gross_Year_To_Date)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   36032   54717   57957   78600  322713

boxplot(df7$Gross_Year_To_Date)
```



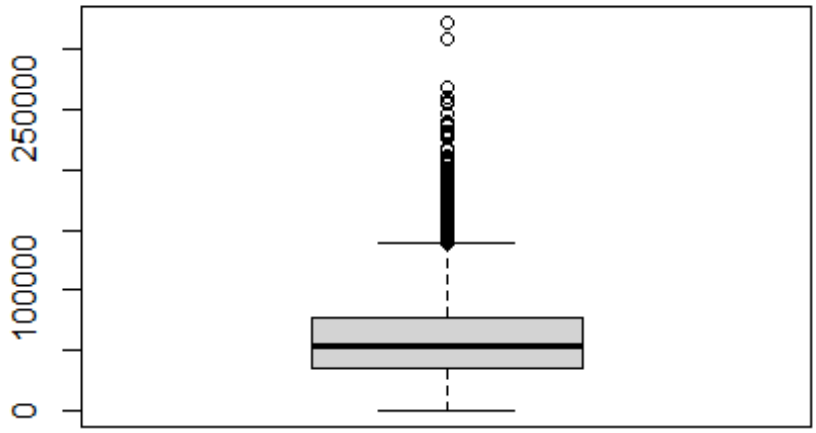

```r
summary(df7_clean$Gross_Year_To_Date)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   35491   53669   55496   76638  124830
```

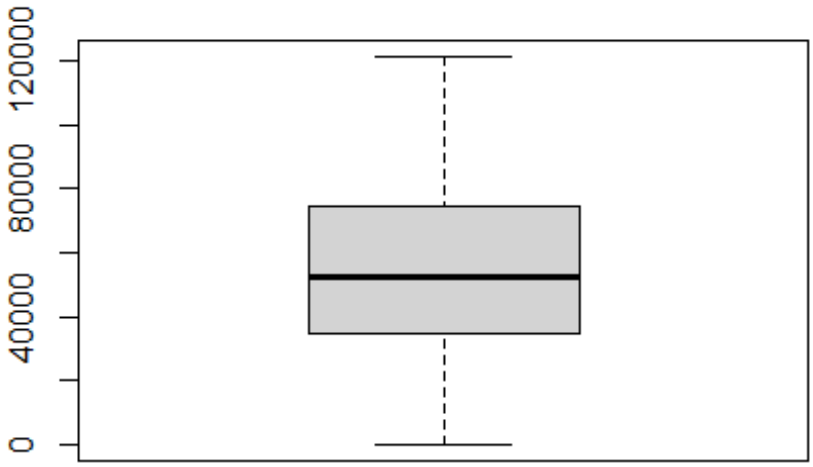
```r
boxplot(df7_clean$Gross_Year_To_Date)
```




```r
Q1 <- quantile(df7$Gross_FRS_Contribution, 0.30)
Q3 <- quantile(df7$Gross_FRS_Contribution, 0.70)
IQR <- Q3 - Q1

# Define the lower and upper bounds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR


# Filter out outliers
df7_clean <- df7[df7$Gross_FRS_Contribution >= lower_bound &
df7$Gross_FRS_Contribution <= upper_bound, ]

summary(df7$Gross_FRS_Contribution)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   35055   53195   56412   76463  322713

boxplot(df7$Gross_FRS_Contribution)
```




```
summary(df7_clean$Gross_FRS_Contribution)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   34488   52153   53962   74524  121488

boxplot(df7_clean$Gross_FRS_Contribution)
```




```
df8 <- df7_clean

current <- as.numeric(format(Sys.Date(), "%Y"))

df8$Age <- current - df8$year_of_birth

View(df8)

# Create a new column 'age_group' based on the age classification
df9 <- df8

#install.packages("dplyr")

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
```

```r
df9 <- df9 %>%
  mutate(age_group = case_when(
    Age >= 16 & Age <= 24 ~ "Youth",    # Youth: Age 16-24
    Age >= 25 & Age <= 64 ~ "Adult",    # Adult: Age 25-64
    Age >= 65             ~ "Senior"    # Senior: Age 65 and over
  ))

View(df9)

table(df9$age_group)
```

```
##
##  Adult Senior
##  79113 105943
```

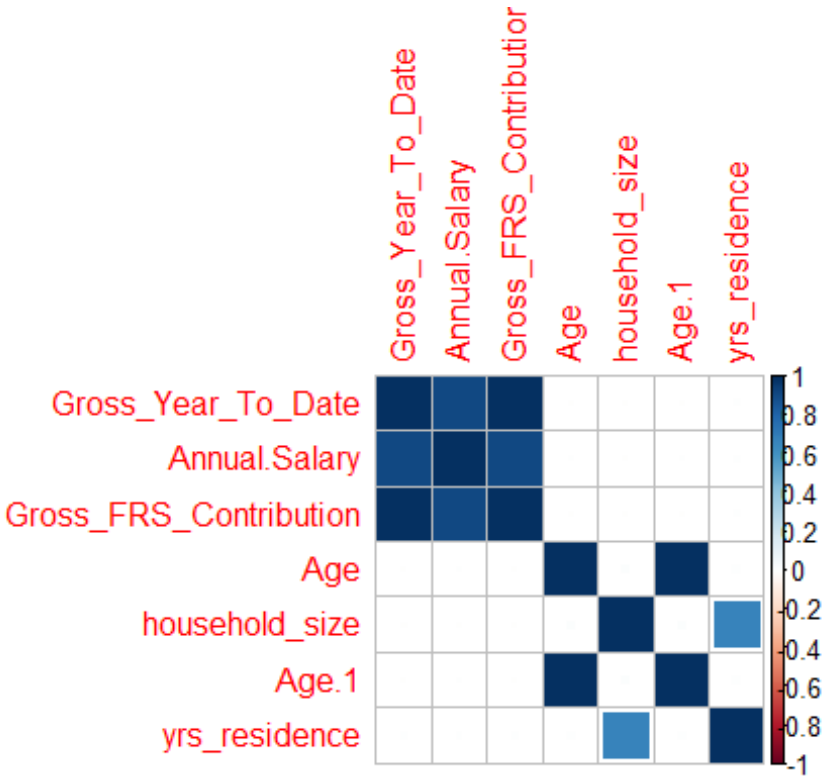
```r
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```r
numeric_data <-
c("Gross_Year_To_Date","Annual.Salary","Gross_FRS_Contribution","Age","household_size","Age","yrs_residence")

corr_data <- df9[numeric_data]

names(df9)
```

```
##  [1] "year_of_birth"           "marital_status"
##  [3] "Education"               "Occupation"
##  [5] "Annual.Salary"           "Gross_Pay_Last_Paycheck"
##  [7] "Gross_Year_To_Date"      "Gross_FRS_Contribution"
##  [9] "household_size"          "yrs_residence"
## [11] "Age"                     "age_group"
```

```r
# Create a correlation matrix
corr_matrix <- cor(corr_data)

# Plot the correlation matrix
corrplot(corr_matrix, method = "square")
```




```
#write.csv(df9, "Prepared_Data.csv", row.names = TRUE)
```