crime and weather dataset cleaning,creating a merged dataset,data modelling: Regression and Lasso

###CRIME Cleaning

```
crime_data<-
read.csv('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/CrimesPres
ent4years.csv',header=TRUE,stringsAsFactors=FALSE)
dailyboard_data <-
read.csv('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/CTA_-
_Ridership_-
_Daily_Boarding_Totals_20240417.csv',header=TRUE,stringsAsFactors=FALSE)
print(names(crime_data))
```

```
##  [1] "ID"                          "Case.Number"
##  [3] "Date"                        "Time"
##  [5] "AM.PM"                       "Block"
##  [7] "IUCR"                        "Primary.Type"
##  [9] "Description"                 "Location.Description"
## [11] "Arrest"                      "Domestic"
## [13] "Beat"                        "District"
## [15] "Ward"                        "Community.Area"
## [17] "FBI.Code"                    "Year"
## [19] "Updated.On"                  "Latitude"
## [21] "Longitude"                   "Location"
## [23] "Historical.Wards.2003.2015" "Zip.Codes"
## [25] "Community.Areas"             "Census.Tracts"
## [27] "Wards"                       "Boundaries...ZIP.Codes"
## [29] "Police.Districts"            "Police.Beats"
## [31] "Wards.2023."
```

```
missing_values <- colSums(is.na(crime_data))
print(missing_values)
```

```
##                        ID               Case.Number
##                         0                         0
##                      Date                      Time
##                         0                         0
##                     AM.PM                     Block
##                         0                         0
##                     IUCR              Primary.Type
##                         0                         0
##              Description      Location.Description
##                         0                         0
##                   Arrest                  Domestic
##                         0                         0
##                     Beat                  District
```

```
##                           0                           0
##                        Ward              Community.Area
##                           2                           0
##                    FBI.Code                        Year
##                           0                           0
##                  Updated.On                    Latitude
##                           0                         122
##                   Longitude                    Location
##                         122                           0
## Historical.Wards.2003.2015                   Zip.Codes
##                         163                           0
##             Community.Areas               Census.Tracts
##                         162                         190
##                       Wards     Boundaries...ZIP.Codes
##                         162                         162
##            Police.Districts                Police.Beats
##                         163                         163
##                 Wards.2023.
##                         162
```

```
# Check for duplicate rows
duplicate_rows <- crime_data[duplicated(crime_data),]
print(duplicate_rows)
```

```
##  [1] ID                          Case.Number
##  [3] Date                        Time
##  [5] AM.PM                       Block
##  [7] IUCR                        Primary.Type
##  [9] Description                 Location.Description
## [11] Arrest                      Domestic
## [13] Beat                        District
## [15] Ward                        Community.Area
## [17] FBI.Code                    Year
## [19] Updated.On                  Latitude
## [21] Longitude                   Location
## [23] Historical.Wards.2003.2015 Zip.Codes
## [25] Community.Areas             Census.Tracts
## [27] Wards                       Boundaries...ZIP.Codes
## [29] Police.Districts            Police.Beats
## [31] Wards.2023.
## <0 rows> (or 0-length row.names)
```

```
str(crime_data)
```

```
## 'data.frame':    16444 obs. of  31 variables:
##  $ ID                          : int  13425318 13424224 13425209 13424232
13424188 13424155 13423997 13423705 13424120 13423655 ...
##  $ Case.Number                 : chr  "JH220414" "JH219164" "JH219155"
"JH219000" ...
##  $ Date                        : chr  "4/10/2024" "4/9/2024" "4/9/2024"
"4/9/2024" ...
```

```
##  $ Time                      : chr  "12:00:00" "11:30:00" "11:25:00"
"8:35:00" ...
##  $ AM.PM                     : chr  "AM" "PM" "PM" "PM" ...
##  $ Block                     : chr  "0000X W 87TH ST" "009XX W BELMONT
AVE" "109XX S MICHIGAN AVE" "012XX W LOYOLA AVE" ...
##  $ IUCR                      : chr  "460" "1350" "1506" "460" ...
##  $ Primary.Type              : chr  "BATTERY" "CRIMINAL TRESPASS"
"PROSTITUTION" "BATTERY" ...
##  $ Description               : chr  "SIMPLE" "TO STATE SUP LAND" "SOLICIT
ON PUBLIC WAY" "SIMPLE" ...
##  $ Location.Description      : chr  "CTA BUS STOP" "CTA TRAIN" "CTA BUS
STOP" "CTA TRAIN" ...
##  $ Arrest                    : logi  FALSE TRUE TRUE FALSE FALSE TRUE ...
##  $ Domestic                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE
...
##  $ Beat                      : int  634 1933 513 2432 815 1522 113 121 331
113 ...
##  $ District                  : int  6 19 5 24 8 15 1 1 3 1 ...
##  $ Ward                      : int  21 44 9 49 23 29 34 34 5 34 ...
##  $ Community.Area            : int  44 6 49 1 62 25 32 28 42 32 ...
##  $ FBI.Code                  : chr  "08B" "26" "16" "08B" ...
##  $ Year                      : int  2024 2024 2024 2024 2024 2024 2024
2024 2024 2024 ...
##  $ Updated.On                : chr  "4/17/2024 15:41" "4/17/2024 15:41"
"4/17/2024 15:41" "4/17/2024 15:41" ...
##  $ Latitude                  : num  41.7 41.9 41.7 42 41.8 ...
##  $ Longitude                 : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ Location                  : chr  "(41.736277427, -87.62510451)"
"(41.93991842, -87.653288878)" "(41.695555151, -87.62077676)" "(42.001064985,
-87.661254473)" ...
##  $ Historical.Wards.2003.2015: int  18 38 30 3 35 52 22 48 32 22 ...
##  $ Zip.Codes                 : chr  "21,554" "4,449" "21,861" "21,853" ...
##  $ Community.Areas           : int  40 57 45 10 60 26 38 29 9 38 ...
##  $ Census.Tracts             : int  1 681 237 48 793 67 92 97 134 92 ...
##  $ Wards                     : int  13 25 43 5 6 7 36 26 33 36 ...
##  $ Boundaries...ZIP.Codes    : int  59 22 19 9 56 32 29 44 24 29 ...
##  $ Police.Districts          : int  20 5 10 11 13 25 22 22 18 22 ...
##  $ Police.Beats              : int  242 31 260 41 105 99 126 91 262 126
...
##  $ Wards.2023.               : int  21 44 11 49 23 29 34 28 3 34 ...
```

```r
# Convert the "Date" column to Date type
crime_data$Date <- as.Date(crime_data$Date, format = "%m/%d/%Y")
# Check the structure of the dataframe after splitting
str(crime_data)
```

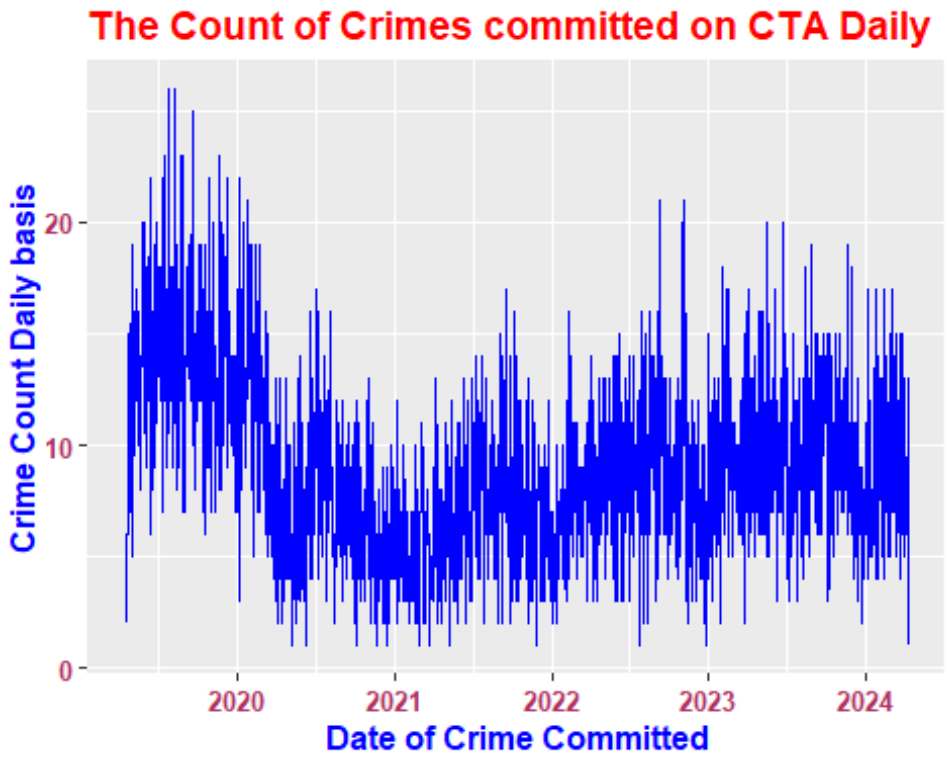
```
## 'data.frame':    16444 obs. of  31 variables:
##  $ ID                        : int  13425318 13424224 13425209 13424232
13424188 13424155 13423997 13423705 13424120 13423655 ...
##  $ Case.Number               : chr  "JH220414" "JH219164" "JH219155"
```

```
"JH219000" ...
##  $ Date                      : Date, format: "2024-04-10" "2024-04-09" ...
##  $ Time                      : chr  "12:00:00" "11:30:00" "11:25:00"
"8:35:00" ...
##  $ AM.PM                     : chr  "AM" "PM" "PM" "PM" ...
##  $ Block                     : chr  "0000X W 87TH ST" "009XX W BELMONT
AVE" "109XX S MICHIGAN AVE" "012XX W LOYOLA AVE" ...
##  $ IUCR                      : chr  "460" "1350" "1506" "460" ...
##  $ Primary.Type              : chr  "BATTERY" "CRIMINAL TRESPASS"
"PROSTITUTION" "BATTERY" ...
##  $ Description               : chr  "SIMPLE" "TO STATE SUP LAND" "SOLICIT
ON PUBLIC WAY" "SIMPLE" ...
##  $ Location.Description      : chr  "CTA BUS STOP" "CTA TRAIN" "CTA BUS
STOP" "CTA TRAIN" ...
##  $ Arrest                    : logi  FALSE TRUE TRUE FALSE FALSE TRUE ...
##  $ Domestic                  : logi  FALSE FALSE FALSE FALSE FALSE FALSE
...
##  $ Beat                      : int  634 1933 513 2432 815 1522 113 121 331
113 ...
##  $ District                  : int  6 19 5 24 8 15 1 1 3 1 ...
##  $ Ward                      : int  21 44 9 49 23 29 34 34 5 34 ...
##  $ Community.Area            : int  44 6 49 1 62 25 32 28 42 32 ...
##  $ FBI.Code                  : chr  "08B" "26" "16" "08B" ...
##  $ Year                      : int  2024 2024 2024 2024 2024 2024 2024
2024 2024 2024 ...
##  $ Updated.On                : chr  "4/17/2024 15:41" "4/17/2024 15:41"
"4/17/2024 15:41" "4/17/2024 15:41" ...
##  $ Latitude                  : num  41.7 41.9 41.7 42 41.8 ...
##  $ Longitude                 : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ Location                  : chr  "(41.736277427, -87.62510451)"
"(41.93991842, -87.653288878)" "(41.695555151, -87.62077676)" "(42.001064985,
-87.661254473)" ...
##  $ Historical.Wards.2003.2015: int  18 38 30 3 35 52 22 48 32 22 ...
##  $ Zip.Codes                 : chr  "21,554" "4,449" "21,861" "21,853" ...
##  $ Community.Areas           : int  40 57 45 10 60 26 38 29 9 38 ...
##  $ Census.Tracts             : int  1 681 237 48 793 67 92 97 134 92 ...
##  $ Wards                     : int  13 25 43 5 6 7 36 26 33 36 ...
##  $ Boundaries...ZIP.Codes    : int  59 22 19 9 56 32 29 44 24 29 ...
##  $ Police.Districts          : int  20 5 10 11 13 25 22 22 18 22 ...
##  $ Police.Beats              : int  242 31 260 41 105 99 126 91 262 126
...
##  $ Wards.2023.               : int  21 44 11 49 23 29 34 28 3 34 ...
```

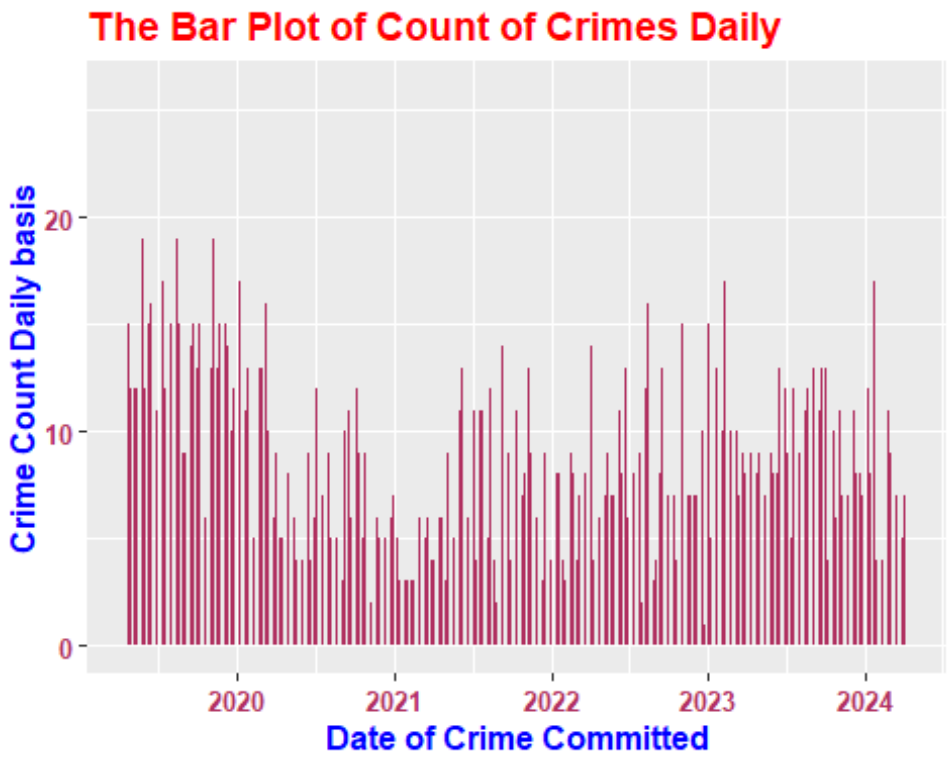
```r
crime_data$Date <- as.Date(crime_data$Date,"%m/%d/%y %I:%M:%S %p")
crime_data$day_of_week<-weekdays(crime_data$Date)
countdailycrimes<-
aggregate(crime_data$ID,by=list(crime_data$Date),FUN=length)
colnames(countdailycrimes)<- c("CrimeDate","countCrime")
#print(countdailycrimes)
#print(countdailycrimes,colnames())
```

```r
library(ggplot2)

ggplot(countdailycrimes, aes(x = CrimeDate, y = countCrime)) +
  geom_line(color = "blue") +
  labs(x = "Date of Crime Committed", y = "Crime Count Daily basis", title =
"The Count of Crimes committed on CTA Daily") +
  theme(
    text = element_text( size = 10, color = "black", face = "bold"),  # For
general text settings
    plot.title = element_text(color = "red", size = 14),  # For the plot
title
    axis.title = element_text(color = "blue", size = 12),  # For axis titles
    axis.text = element_text(color = "maroon", size = 10)  # For axis labels
  )
```




```r
ggplot(countdailycrimes, aes(x = CrimeDate, y = countCrime)) +
  geom_bar(stat="identity",width=0.5,fill = "maroon") +
  labs(x = "Date of Crime Committed", y = "Crime Count Daily basis", title =
"The Bar Plot of Count of Crimes Daily") +
  theme(
    text = element_text( size = 10, color = "black", face = "bold"),  # For
general text settings
    plot.title = element_text(color = "red", size = 14),  # For the plot
title
    axis.title = element_text(color = "blue", size = 12),  # For axis titles
    axis.text = element_text(color = "maroon", size = 10)  # For axis labels
  )
```



###END OF CRIME DATA FRAME

Preparation of Weather Data

```
weather_origData =
read.csv(file.path('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/
weather orig.csv'))
summary(weather_origData)
```

```
##      Date           TAVG..orig.        TAVG.Calc.F.
##  Length:24093       Length:24093       Length:24093
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit. PRCP..Inches.
##  Length:24093              Length:24093              Length:24093
##  Class :character          Class :character          Class :character
##  Mode  :character          Mode  :character          Mode  :character
##  SNOW..Inches.      SNWD..Inches.        location
##  Length:24093       Length:24093       Length:24093
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character

str(weather_origData)

## 'data.frame':    24093 obs. of  9 variables:
##  $ Date                     : chr  "5/1/1997" "5/2/1997" "5/3/1997"
"5/4/1997" ...
##  $ TAVG..orig.              : chr  "" "" "" "" ...
```

```
##  $ TAVG.Calc.F.             : chr  "48.00" "49.00" "49.50" "52.00" ...
##  $ TMAX..Degrees.Fahrenheit.: chr  "60.00" "53.00" "57.00" "64.00" ...
##  $ TMIN..Degrees.Fahrenheit.: chr  "36.00" "45.00" "42.00" "40.00" ...
##  $ PRCP..Inches.            : chr  "" "" "" "" ...
##  $ SNOW..Inches.            : chr  "" "" "" "" ...
##  $ SNWD..Inches.            : chr  "" "" "" "" ...
##  $ location                 : chr  "CHICAGO MIDWAY AIRPORT, IL US
(USW00014819)" "CHICAGO MIDWAY AIRPORT, IL US (USW00014819)" "CHICAGO MIDWAY
AIRPORT, IL US (USW00014819)" "CHICAGO MIDWAY AIRPORT, IL US (USW00014819)"
...
```

```
# Replace empty strings with 0
#weatherData[weatherData == ""] <- "0"
weather_origData[is.na(weather_origData) | weather_origData == ""] <- 0.0

# Printing the modified data
#print(weather_origData)

#check for missing rows
miss_vals <- colSums(is.na(weather_origData))
print(miss_vals)
```

```
##                      Date               TAVG..orig.
TAVG.Calc.F.
##                         0                         0
0
## TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
PRCP..Inches.
##                         0                         0
0
##             SNOW..Inches.             SNWD..Inches.
location
##                         0                         0
0
```

```
# Check for duplicate rows
duplicates <- weather_origData[duplicated(weather_origData),]
print(duplicates)
```

```
## [1] Date                      TAVG..orig.
## [3] TAVG.Calc.F.              TMAX..Degrees.Fahrenheit.
## [5] TMIN..Degrees.Fahrenheit. PRCP..Inches.
## [7] SNOW..Inches.             SNWD..Inches.
## [9] location
## <0 rows> (or 0-length row.names)
```

```
New_weather =
cbind(weather_origData[1],weather_origData[3],weather_origData[4],weather_ori
gData[5],weather_origData[6],weather_origData[7],weather_origData[8])
str(New_weather)
```

```
## 'data.frame':    24093 obs. of  7 variables:
##  $ Date                       : chr  "5/1/1997" "5/2/1997" "5/3/1997"
"5/4/1997" ...
##  $ TAVG.Calc.F.               : chr  "48.00" "49.00" "49.50" "52.00" ...
##  $ TMAX..Degrees.Fahrenheit.: chr  "60.00" "53.00" "57.00" "64.00" ...
##  $ TMIN..Degrees.Fahrenheit.: chr  "36.00" "45.00" "42.00" "40.00" ...
##  $ PRCP..Inches.              : chr  "0" "0" "0" "0" ...
##  $ SNOW..Inches.              : chr  "0" "0" "0" "0" ...
##  $ SNWD..Inches.              : chr  "0" "0" "0" "0" ...
```

```
# convert the date column to a Date object
New_weather$Date <- as.Date(New_weather$Date, format = "%m/%d/%Y")
# format the date column to the desired format
New_weather$Date <- format(New_weather$Date, "%Y-%m-%d")
str(New_weather)
```

```
## 'data.frame':    24093 obs. of  7 variables:
##  $ Date                       : chr  "1997-05-01" "1997-05-02" "1997-05-03"
"1997-05-04" ...
##  $ TAVG.Calc.F.               : chr  "48.00" "49.00" "49.50" "52.00" ...
##  $ TMAX..Degrees.Fahrenheit.: chr  "60.00" "53.00" "57.00" "64.00" ...
##  $ TMIN..Degrees.Fahrenheit.: chr  "36.00" "45.00" "42.00" "40.00" ...
##  $ PRCP..Inches.              : chr  "0" "0" "0" "0" ...
##  $ SNOW..Inches.              : chr  "0" "0" "0" "0" ...
##  $ SNWD..Inches.              : chr  "0" "0" "0" "0" ...
```

```
New_weather$Date <- as.Date(New_weather$Date)
summary(New_weather)
```

```
##       Date            TAVG.Calc.F.       TMAX..Degrees.Fahrenheit.
##  Min.   :1997-05-01   Length:24093       Length:24093
##  1st Qu.:2009-04-10   Class :character   Class :character
##  Median :2014-10-09   Mode  :character   Mode  :character
##  Mean   :2014-02-05
##  3rd Qu.:2020-03-01
##  Max.   :2024-04-15
##  NA's   :1
##  TMIN..Degrees.Fahrenheit. PRCP..Inches.      SNOW..Inches.
##  Length:24093              Length:24093       Length:24093
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
##
##
##
##
##  SNWD..Inches.
##  Length:24093
##  Class :character
##  Mode  :character
##
##
```

```
##
##

# create a subset of weatherData2 dataframe containing only the rows
corresponding to the year 2023.
subData <- subset(New_weather, Date >= "2023-01-01" & Date <= "2023-12-31")
str(subData)

## 'data.frame':    1460 obs. of  7 variables:
##  $ Date                      : Date, format: "2023-01-01" "2023-01-01" ...
##  $ TAVG.Calc.F.              : chr  "0.00" "0.00" "0.00" "41.00" ...
##  $ TMAX..Degrees.Fahrenheit.: chr  "0" "0" "0" "45.00" ...
##  $ TMIN..Degrees.Fahrenheit.: chr  "0" "0" "0" "37.00" ...
##  $ PRCP..Inches.             : chr  "0.12" "0.14" "0.19" "0.14" ...
##  $ SNOW..Inches.             : chr  "0" "0" "0" "0" ...
##  $ SNWD..Inches.             : chr  "0" "0" "0" "0.00" ...

#search for patterns at the beginning of each date that match two digits
(\\d{2}) and replace with "20". This converts two-digit years into four-digit
years.
subData$Date <- gsub("^\\d{2}", "20", subData$Date)
summary(subData)

##      Date           TAVG.Calc.F.       TMAX..Degrees.Fahrenheit.
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  TMIN..Degrees.Fahrenheit. PRCP..Inches.      SNOW..Inches.
##  Length:1460               Length:1460        Length:1460
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
##  SNWD..Inches.
##  Length:1460
##  Class :character
##  Mode  :character

#converting the datatypes of other columns as suited
subData$Date <- as.Date(subData$Date)
summary(subData)

##       Date            TAVG.Calc.F.       TMAX..Degrees.Fahrenheit.
##  Min.   :2023-01-01   Length:1460        Length:1460
##  1st Qu.:2023-04-02   Class :character   Class :character
##  Median :2023-07-02   Mode  :character   Mode  :character
##  Mean   :2023-07-02
##  3rd Qu.:2023-10-01
##  Max.   :2023-12-31
##  TMIN..Degrees.Fahrenheit. PRCP..Inches.      SNOW..Inches.
##  Length:1460               Length:1460        Length:1460
##  Class :character          Class :character   Class :character
##  Mode  :character          Mode  :character   Mode  :character
```

```
##
##
##
##  SNWD..Inches.
##  Length:1460
##  Class :character
##  Mode  :character
##
##
##
```

```
str(subData)
```

```
## 'data.frame':    1460 obs. of  7 variables:
##  $ Date                       : Date, format: "2023-01-01" "2023-01-01" ...
##  $ TAVG.Calc.F.               : chr  "0.00" "0.00" "0.00" "41.00" ...
##  $ TMAX..Degrees.Fahrenheit.: chr  "0" "0" "0" "45.00" ...
##  $ TMIN..Degrees.Fahrenheit.: chr  "0" "0" "0" "37.00" ...
##  $ PRCP..Inches.              : chr  "0.12" "0.14" "0.19" "0.14" ...
##  $ SNOW..Inches.              : chr  "0" "0" "0" "0" ...
##  $ SNWD..Inches.              : chr  "0" "0" "0" "0.00" ...
```

```
#converting the datatypes of other columns as suited
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
```

```
# Convert Date column to Date type
subData$Date <- as.Date(subData$Date)

# Convert other columns to appropriate data types
subData$TMAX..Degrees.Fahrenheit. <-
as.numeric(subData$TMAX..Degrees.Fahrenheit.)
subData$TMIN..Degrees.Fahrenheit. <-
as.numeric(subData$TMIN..Degrees.Fahrenheit.)
subData$PRCP..Inches. <- as.numeric(subData$PRCP..Inches.)
subData$SNOW..Inches. <- as.numeric(subData$SNOW..Inches.)
subData$SNWD..Inches. <- as.numeric(subData$SNWD..Inches.)
subData$SNOW <- as.numeric(subData$SNOW)

# Check the structure of the dataframe after data type conversion
str(subData)
```

```
## 'data.frame':    1460 obs. of  8 variables:
##  $ Date                       : Date, format: "2023-01-01" "2023-01-01" ...
##  $ TAVG.Calc.F.               : chr  "0.00" "0.00" "0.00" "41.00" ...
##  $ TMAX..Degrees.Fahrenheit.: num  0 0 0 45 0 0 0 44 0 0 ...
##  $ TMIN..Degrees.Fahrenheit.: num  0 0 0 37 0 0 0 32 0 0 ...
##  $ PRCP..Inches.              : num  0.12 0.14 0.19 0.14 0 0 0 0 0.82 0.87
...
##  $ SNOW..Inches.              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SNWD..Inches.              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SNOW                       : num  0 0 0 0 0 0 0 0 0 0 ...

#dataset with modified columns and data displayed correctly
summary(subData)

##       Date            TAVG.Calc.F.       TMAX..Degrees.Fahrenheit.
##  Min.   :2023-01-01   Length:1460        Min.   :  0.00
##  1st Qu.:2023-04-02   Class :character   1st Qu.:  0.00
##  Median :2023-07-02   Mode  :character   Median :  0.00
##  Mean   :2023-07-02                      Mean   : 15.66
##  3rd Qu.:2023-10-01                      3rd Qu.:  3.00
##  Max.   :2023-12-31                      Max.   :100.00
##  TMIN..Degrees.Fahrenheit. PRCP..Inches.   SNOW..Inches.
SNWD..Inches.
##  Min.   : 0.00             Min.   :0.000   Min.   :0.000000   Min.
:0.00000
##  1st Qu.: 0.00             1st Qu.:0.000   1st Qu.:0.000000   1st
Qu.:0.00000
##  Median : 0.00             Median :0.000   Median :0.000000   Median
:0.00000
##  Mean   :11.61             Mean   :0.106   Mean   :0.008151   Mean
:0.03418
##  3rd Qu.: 0.25             3rd Qu.:0.050   3rd Qu.:0.000000   3rd
Qu.:0.00000
##  Max.   :76.00             Max.   :4.680   Max.   :2.300000   Max.
:2.50000
##       SNOW
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.008151
##  3rd Qu.:0.000000
##  Max.   :2.300000

#return the indices of the elements in the "TMAX" column of newData where the
value is NA.
which(is.na(subData$TMAX))

## integer(0)
```

```r
#find the indices of missing values (NA) in the "TMIN" column of the newData dataframe.
which(is.na(subData$TMIN))

## integer(0)
```

#Since there are no empty values for Tmax and Tmin, we proceed with further analyiss.Or else we would have to remove the NA values individually.

```r
#replace NA values in the "SNOW" column of the subData dataframe with 0.
subData$SNOW <- ifelse(is.na(subData$SNOW), 0, subData$SNOW)
summary(subData)

##       Date            TAVG.Calc.F.       TMAX..Degrees.Fahrenheit.
##  Min.   :2023-01-01   Length:1460        Min.   :  0.00
##  1st Qu.:2023-04-02   Class :character   1st Qu.:  0.00
##  Median :2023-07-02   Mode  :character   Median :  0.00
##  Mean   :2023-07-02                      Mean   : 15.66
##  3rd Qu.:2023-10-01                      3rd Qu.:  3.00
##  Max.   :2023-12-31                      Max.   :100.00
##  TMIN..Degrees.Fahrenheit. PRCP..Inches.   SNOW..Inches.
SNWD..Inches.
##  Min.   : 0.00             Min.   :0.000   Min.   :0.000000   Min.
:0.00000
##  1st Qu.: 0.00             1st Qu.:0.000   1st Qu.:0.000000   1st
Qu.:0.00000
##  Median : 0.00             Median :0.000   Median :0.000000   Median
:0.00000
##  Mean   :11.61             Mean   :0.106   Mean   :0.008151   Mean
:0.03418
##  3rd Qu.: 0.25             3rd Qu.:0.050   3rd Qu.:0.000000   3rd
Qu.:0.00000
##  Max.   :76.00             Max.   :4.680   Max.   :2.300000   Max.
:2.50000
##       SNOW
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.008151
##  3rd Qu.:0.000000
##  Max.   :2.300000
```

SNOW

Merge Data

```r
lstation_data=
read.csv('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/CTA-RidershipL_Station_EntriesMonthly.csv')
summary(lstation_data)
```

```
##    station_id     stationame        month_beginning    avg_weekday_rides
##  Min.   :40010   Length:39053       Length:39053       Length:39053
##  1st Qu.:40370   Class :character   Class :character   Class :character
##  Median :40760   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40767
##  3rd Qu.:41160
##  Max.   :41700
##  avg_saturday_rides avg_sunday.holiday_rides  monthtotal
##  Length:39053       Length:39053             Length:39053
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
```

```
#changing the datatype
lstation_data$avg_weekday_rides <- as.numeric(gsub(",", "",
lstation_data$avg_weekday_rides))
lstation_data$avg_saturday_rides <- as.numeric(gsub(",", "",
lstation_data$avg_saturday_rides))
lstation_data$avg_sunday.holiday_rides <- as.numeric(gsub(",", "",
lstation_data$avg_sunday.holiday_rides))
lstation_data$monthtotal <- as.numeric(gsub(",", "",
lstation_data$monthtotal))
str(lstation_data)
```

```
## 'data.frame':    39053 obs. of  7 variables:
##  $ station_id             : int  40900 41190 40100 41300 40760 40880
41380 40340 41200 40770 ...
##  $ stationame             : chr  "Howard" "Jarvis" "Morse" "Loyola" ...
##  $ month_beginning        : chr  "01/01/2001" "01/01/2001" "01/01/2001"
"01/01/2001" ...
##  $ avg_weekday_rides      : num  6234 1489 4412 4664 3110 ...
##  $ avg_saturday_rides     : num  3814 1054 3064 3156 2126 ...
##  $ avg_sunday.holiday_rides: num  2409 718 2088 1953 1454 ...
##  $ monthtotal             : num  164447 40567 119772 125008 84189 ...
```

```
summary(lstation_data)
```

```
##    station_id     stationame        month_beginning    avg_weekday_rides
##  Min.   :40010   Length:39053       Length:39053       Min.   :    0
##  1st Qu.:40370   Class :character   Class :character   1st Qu.: 1258
##  Median :40760   Mode  :character   Mode  :character   Median : 2601
##  Mean   :40767                                         Mean   : 3541
##  3rd Qu.:41160                                         3rd Qu.: 4621
##  Max.   :41700                                         Max.   :24041
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal
##  Min.   :    0.0    Min.   :    0.0          Min.   :     0
##  1st Qu.:  720.8    1st Qu.:  499.0          1st Qu.: 32769
##  Median : 1324.5    Median :  924.8          Median : 65816
##  Mean   : 2093.4    Mean   : 1533.9          Mean   : 91761
```

```
##  3rd Qu.: 2693.5    3rd Qu.: 1918.0          3rd Qu.:120845
##  Max.   :19171.3    Max.   :15982.0          Max.   :670496

# Convert month_beginning to Date format
lstation_data$month_beginning <- as.Date(lstation_data$month_beginning,
format = "%m/%d/%Y")

# Filter the data for the year 2023
lstation_2023 <- subset(lstation_data, format(month_beginning, "%Y") ==
"2023")

# Print the filtered data
str(lstation_2023)

## 'data.frame':    1431 obs. of  7 variables:
##  $ station_id              : int  40900 41190 40100 41300 40760 40880
41380 40340 41200 40770 ...
##  $ stationame              : chr  "Howard" "Jarvis" "Morse" "Loyola" ...
##  $ month_beginning         : Date, format: "2023-01-01" "2023-01-01" ...
##  $ avg_weekday_rides       : num  2592 784 1995 2361 1785 ...
##  $ avg_saturday_rides      : num  1854 669 1511 2021 1558 ...
##  $ avg_sunday.holiday_rides: num  1505 486 1191 1376 1048 ...
##  $ monthtotal              : num  70868 22057 55090 65929 50015 ...

#Create new column called no_of_trips by adding values of other 3 columns
lstation_2023$no_of_trips <- lstation_2023$avg_weekday_rides +
lstation_2023$avg_saturday_rides + lstation_2023$avg_sunday.holiday_rides
str(lstation_2023)

## 'data.frame':    1431 obs. of  8 variables:
##  $ station_id              : int  40900 41190 40100 41300 40760 40880
41380 40340 41200 40770 ...
##  $ stationame              : chr  "Howard" "Jarvis" "Morse" "Loyola" ...
##  $ month_beginning         : Date, format: "2023-01-01" "2023-01-01" ...
##  $ avg_weekday_rides       : num  2592 784 1995 2361 1785 ...
##  $ avg_saturday_rides      : num  1854 669 1511 2021 1558 ...
##  $ avg_sunday.holiday_rides: num  1505 486 1191 1376 1048 ...
##  $ monthtotal              : num  70868 22057 55090 65929 50015 ...
##  $ no_of_trips             : num  5950 1939 4697 5759 4392 ...

# Check the columns of lstation_2023
print(colnames(lstation_2023))

## [1] "station_id"               "stationame"
## [3] "month_beginning"          "avg_weekday_rides"
## [5] "avg_saturday_rides"       "avg_sunday.holiday_rides"
## [7] "monthtotal"               "no_of_trips"

# Check the columns of subData
str(colnames(subData))
```

```
##  chr [1:8] "Date" "TAVG.Calc.F." "TMAX..Degrees.Fahrenheit." ...

# Merge the datasets on the common column
merged_data <- merge(lstation_2023, subData, by.x = "month_beginning", by.y =
"Date", all = TRUE)

# Check the structure of the merged dataset
str(merged_data)

## 'data.frame':    7144 obs. of  15 variables:
##  $ month_beginning          : Date, format: "2023-01-01" "2023-01-01" ...
##  $ station_id               : int  40900 40900 40900 40900 41190 41190
41190 41190 40100 40100 ...
##  $ stationame               : chr  "Howard" "Howard" "Howard" "Howard" ...
##  $ avg_weekday_rides        : num  2592 2592 2592 2592 784 ...
##  $ avg_saturday_rides       : num  1854 1854 1854 1854 669 ...
##  $ avg_sunday.holiday_rides : num  1505 1505 1505 1505 486 ...
##  $ monthtotal               : num  70868 70868 70868 70868 22057 ...
##  $ no_of_trips              : num  5950 5950 5950 5950 1939 ...
##  $ TAVG.Calc.F.             : chr  "0.00" "41.00" "0.00" "0.00" ...
##  $ TMAX..Degrees.Fahrenheit.: num  0 45 0 0 0 45 0 0 0 45 ...
##  $ TMIN..Degrees.Fahrenheit.: num  0 37 0 0 0 37 0 0 0 37 ...
##  $ PRCP..Inches.            : num  0.12 0.14 0.14 0.19 0.12 0.14 0.14 0.19
0.12 0.14 ...
##  $ SNOW..Inches.            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SNWD..Inches.            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SNOW                     : num  0 0 0 0 0 0 0 0 0 0 ...

summary(merged_data)

##  month_beginning        station_id     stationame        avg_weekday_rides
##  Min.   :2023-01-01   Min.   :40010   Length:7144        Min.   :    0
##  1st Qu.:2023-03-01   1st Qu.:40380   Class :character   1st Qu.:  764
##  Median :2023-06-01   Median :40780   Mode  :character   Median : 1600
##  Mean   :2023-05-26   Mean   :40786                      Mean   : 2164
##  3rd Qu.:2023-08-01   3rd Qu.:41180                      3rd Qu.: 2862
##  Max.   :2023-12-31   Max.   :41700                      Max.   :10725
##                       NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
```

```
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##  TAVG.Calc.F.       TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Length:7144        Min.   :  0.00            Min.   : 0.00
##  Class :character   1st Qu.:  0.00            1st Qu.: 0.00
##  Mode  :character   Median :  0.00            Median : 0.00
##                     Mean   : 16.33            Mean   :12.12
##                     3rd Qu.:  3.00            3rd Qu.: 0.25
##                     Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.          SNOW
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0.000000
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053   Mean   :0.001666
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000   Max.   :2.300000
##
```

```
#downloaded the merged file
#write.csv(merged_data, "mergeddata_weather+rider.csv", row.names = FALSE)

colnames(merged_data)
```

```
##  [1] "month_beginning"           "station_id"
##  [3] "stationame"                "avg_weekday_rides"
##  [5] "avg_saturday_rides"        "avg_sunday.holiday_rides"
##  [7] "monthtotal"                "no_of_trips"
##  [9] "TAVG.Calc.F."              "TMAX..Degrees.Fahrenheit."
## [11] "TMIN..Degrees.Fahrenheit." "PRCP..Inches."
## [13] "SNOW..Inches."             "SNWD..Inches."
## [15] "SNOW"
```

```
#taking the required columns only
merged_data =
cbind(merged_data[1],merged_data[2],merged_data[8],merged_data[13])
summary(merged_data)
```

```
##  month_beginning        station_id     no_of_trips    SNOW..Inches.
##  Min.   :2023-01-01   Min.   :40010   Min.   :    0   Min.   :0.000000
##  1st Qu.:2023-03-01   1st Qu.:40380   1st Qu.: 1655   1st Qu.:0.000000
##  Median :2023-06-01   Median :40780   Median : 3378   Median :0.000000
##  Mean   :2023-05-26   Mean   :40786   Mean   : 4927   Mean   :0.001666
##  3rd Qu.:2023-08-01   3rd Qu.:41180   3rd Qu.: 6299   3rd Qu.:0.000000
##  Max.   :2023-12-31   Max.   :41700   Max.   :27057   Max.   :2.300000
##                       NA's   :1420    NA's   :1420
```

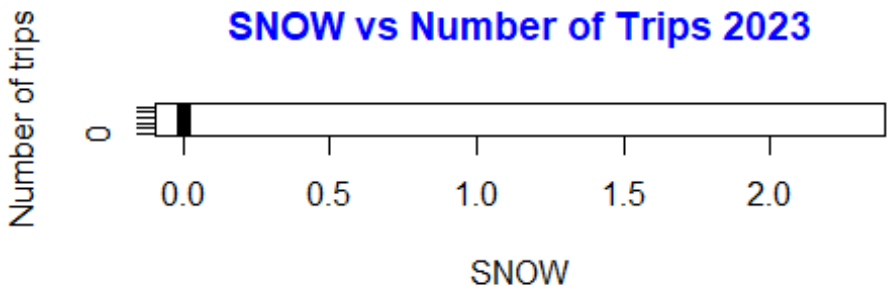
```
#assign 0 to NA
merged_data$SNOW <- ifelse(is.na(merged_data$SNOW), 0, merged_data$SNOW)
summary(merged_data)
```

```
##  month_beginning        station_id     no_of_trips    SNOW..Inches.
##  Min.   :2023-01-01   Min.   :40010   Min.   :    0   Min.   :0.000000
##  1st Qu.:2023-03-01   1st Qu.:40380   1st Qu.: 1655   1st Qu.:0.000000
##  Median :2023-06-01   Median :40780   Median : 3378   Median :0.000000
##  Mean   :2023-05-26   Mean   :40786   Mean   : 4927   Mean   :0.001666
##  3rd Qu.:2023-08-01   3rd Qu.:41180   3rd Qu.: 6299   3rd Qu.:0.000000
##  Max.   :2023-12-31   Max.   :41700   Max.   :27057   Max.   :2.300000
##                       NA's   :1420    NA's   :1420
##       SNOW
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.001666
##  3rd Qu.:0.000000
##  Max.   :2.300000
##
```

```
merged_data$Date <- gsub("^\\d{2}", "20", merged_data$month_beginning)
merged_data$Date <- as.Date(merged_data$month_beginning)
merged_data$Date <- gsub("^\\d{2}", "20", merged_data$month_beginning)
merged_data$Date <- as.Date(merged_data$month_beginning)
summary(merged_data)
```

```
##  month_beginning        station_id     no_of_trips    SNOW..Inches.
##  Min.   :2023-01-01   Min.   :40010   Min.   :    0   Min.   :0.000000
##  1st Qu.:2023-03-01   1st Qu.:40380   1st Qu.: 1655   1st Qu.:0.000000
##  Median :2023-06-01   Median :40780   Median : 3378   Median :0.000000
##  Mean   :2023-05-26   Mean   :40786   Mean   : 4927   Mean   :0.001666
##  3rd Qu.:2023-08-01   3rd Qu.:41180   3rd Qu.: 6299   3rd Qu.:0.000000
##  Max.   :2023-12-31   Max.   :41700   Max.   :27057   Max.   :2.300000
##                       NA's   :1420    NA's   :1420
##       SNOW              Date
##  Min.   :0.000000   Min.   :2023-01-01
##  1st Qu.:0.000000   1st Qu.:2023-03-01
##  Median :0.000000   Median :2023-06-01
##  Mean   :0.001666   Mean   :2023-05-26
##  3rd Qu.:0.000000   3rd Qu.:2023-08-01
##  Max.   :2.300000   Max.   :2023-12-31
##
```

```
#scatter plot 2*1
par(mfrow = c(2, 1))
plot(x=merged_data$SNOW,y=merged_data$no_of_trips,xlab = "SNOW" ,ylab =
"Number of trips",main = "SNOW vs Number of Trips 2023",col.main="Blue")
```



##TEMPERATURE

```
tempdata_merge =
read.csv(file.path('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/
mergeddata_weather+rider.csv'))
summary(tempdata_merge)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
```

```
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000
##
```

```
colnames(tempdata_merge)
```

```
##  [1] "month_beginning"           "station_id"
##  [3] "stationame"                "avg_weekday_rides"
##  [5] "avg_saturday_rides"        "avg_sunday.holiday_rides"
##  [7] "monthtotal"                "no_of_trips"
##  [9] "TAVG.Calc.F."              "TMAX..Degrees.Fahrenheit."
## [11] "TMIN..Degrees.Fahrenheit." "PRCP..Inches."
## [13] "SNOW..Inches."             "SNWD..Inches."
```

```
shortdata = cbind(tempdata_merge[8],tempdata_merge[9])
summary(shortdata)
```

```
##   no_of_trips     TAVG.Calc.F.
##  Min.   :    0   Min.   : 0.000
##  1st Qu.: 1655   1st Qu.: 0.000
##  Median : 3378   Median : 0.000
##  Mean   : 4927   Mean   :14.228
##  3rd Qu.: 6299   3rd Qu.: 1.625
##  Max.   :27057   Max.   :87.000
##  NA's   :1420
```

```
str(shortdata)
```

```
## 'data.frame':    7144 obs. of  2 variables:
##  $ no_of_trips : num  5950 5950 5950 5950 1939 ...
##  $ TAVG.Calc.F.: num  0 41 0 0 0 41 0 0 0 41 ...
```

```
# Calculate the mean value for the "Avg_temp" column
mean_Avg_temp <- mean(shortdata$TAVG.Calc.F., na.rm = TRUE)

# Check if there are missing values in the "Avg_temp" column
if (any(is.na(shortdata$TAVG.Calc.F.))) {
  # Replace missing values with the mean value
```

```r
  shortdata$TAVG.Calc.F.[is.na(shortdata$TAVG.Calc.F.)] <- mean_Avg_temp
} else {
  print("No missing values found in Avg_temp column.")
}
```

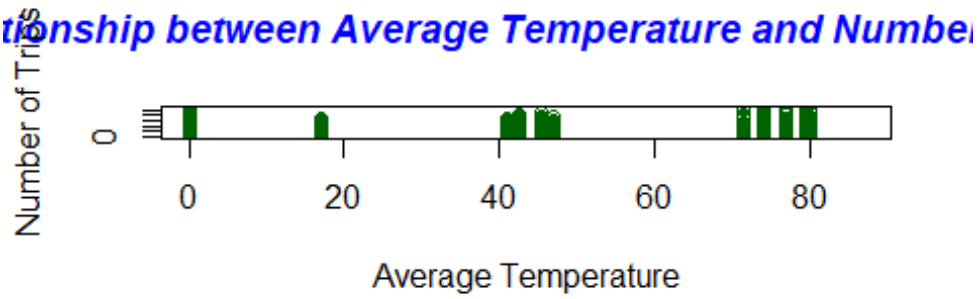
```
## [1] "No missing values found in Avg_temp column."
```

```r
# Display a summary of the updated dataframe
summary(shortdata)
```

```
##   no_of_trips     TAVG.Calc.F.
##  Min.   :    0   Min.   : 0.000
##  1st Qu.: 1655   1st Qu.: 0.000
##  Median : 3378   Median : 0.000
##  Mean   : 4927   Mean   :14.228
##  3rd Qu.: 6299   3rd Qu.: 1.625
##  Max.   :27057   Max.   :87.000
##  NA's   :1420
```

```r
# Set up a 2x1 plot layout
par(mfrow = c(2, 1))

# Create a scatter plot with customizations
plot(x = shortdata$TAVG.Calc.F.,
     y = shortdata$no_of_trips,
     xlab = "Average Temperature",
     ylab = "Number of Trips",
     main = "Relationship between Average Temperature and Number of Trips",
     col = "darkgreen",       # Set color of data points
         col.main = "blue",  # Set main title color
     font.main = 4       # Set main title font style (bold)
)
```



###PRECIPITATION

```
tempdata_merge =
read.csv(file.path('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/
mergeddata_weather+rider.csv'))
summary(tempdata_merge)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
```

```
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000
##
```

```
precipData = cbind(tempdata_merge[8],tempdata_merge[12])
summary(precipData)
```

```
##   no_of_trips    PRCP..Inches.
##  Min.   :    0   Min.   :0.0000
##  1st Qu.: 1655   1st Qu.:0.0000
##  Median : 3378   Median :0.0000
##  Mean   : 4927   Mean   :0.0624
##  3rd Qu.: 6299   3rd Qu.:0.0900
##  Max.   :27057   Max.   :4.6800
##  NA's   :1420
```

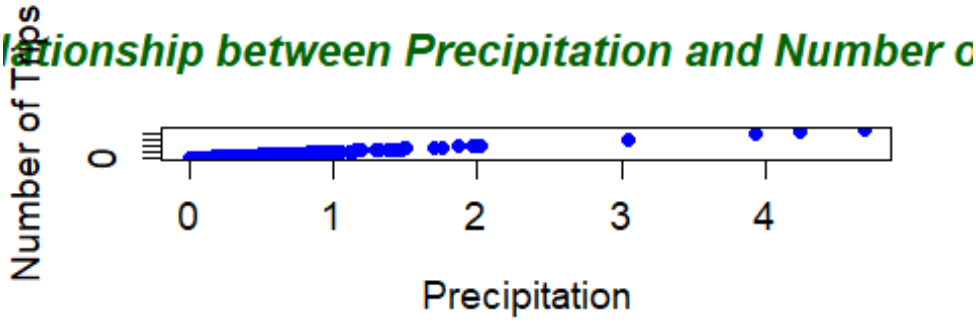
```
mean <- mean(precipData$PRCP, na.rm = TRUE)
precipData$PRCP <- ifelse(is.na(precipData$PRCP), mean, precipData$PRCP)
summary(precipData)
```

```
##   no_of_trips    PRCP..Inches.         PRCP
##  Min.   :    0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 1655   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3378   Median :0.0000   Median :0.0000
##  Mean   : 4927   Mean   :0.0624   Mean   :0.0624
##  3rd Qu.: 6299   3rd Qu.:0.0900   3rd Qu.:0.0900
##  Max.   :27057   Max.   :4.6800   Max.   :4.6800
##  NA's   :1420
```

```
# Set up a 2x1 plot layout
par(mfrow = c(2, 1))

# Create a scatter plot with customizations
plot(x = precipData$PRCP,
     y = precipData$PRCP,
     xlab = "Precipitation",
     ylab = "Number of Trips",
```

```
     main = "Relationship between Precipitation and Number of Trips",
     col = "blue",          # Set color of data points
     pch = 16,              # Set symbol shape (filled circle)
     cex = 1,             # Set symbol size
     cex.axis = 1.2,        # Set axis label size
     cex.lab = 1.2,         # Set axis title size
     cex.main = 1.4,        # Set main title size
     col.main = "darkgreen",# Set main title color
     font.main = 4          # Set main title font style (bold)
)
```



###MODEL

```
MODELdata =
read.csv(file.path('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/
mergeddata_weather+rider.csv'))
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
```

```
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000
##
```

```
MODELdata = cbind( MODELdata[8], MODELdata[9], MODELdata[12], MODELdata[13])
summary(MODELdata)
```

```
##   no_of_trips     TAVG.Calc.F.    PRCP..Inches.    SNOW..Inches.
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.: 1655   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median : 3378   Median : 0.000   Median :0.0000   Median :0.000000
##  Mean   : 4927   Mean   :14.228   Mean   :0.0624   Mean   :0.001666
##  3rd Qu.: 6299   3rd Qu.: 1.625   3rd Qu.:0.0900   3rd Qu.:0.000000
##  Max.   :27057   Max.   :87.000   Max.   :4.6800   Max.   :2.300000
##  NA's   :1420
```

```
mean <- mean(MODELdata$no_of_trips, na.rm = TRUE)
MODELdata$no_of_trips <- ifelse(is.na(MODELdata$no_of_trips), mean,
MODELdata$no_of_trips)
summary(MODELdata)
```

```
##   no_of_trips     TAVG.Calc.F.    PRCP..Inches.    SNOW..Inches.
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.: 2034   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median : 4828   Median : 0.000   Median :0.0000   Median :0.000000
```

```
##  Mean   : 4927   Mean   :14.228   Mean   :0.0624   Mean   :0.001666
##  3rd Qu.: 5481   3rd Qu.: 1.625   3rd Qu.:0.0900   3rd Qu.:0.000000
##  Max.   :27057   Max.   :87.000   Max.   :4.6800   Max.   :2.300000
```

```
mean <- mean(MODELdata$TAVG.Calc.F., na.rm = TRUE)
MODELdata$Avgtemp <- ifelse(is.na(MODELdata$TAVG.Calc.F.), mean,
MODELdata$TAVG.Calc.F.)
summary(MODELdata)
```

```
##   no_of_trips     TAVG.Calc.F.    PRCP..Inches.    SNOW..Inches.
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.: 2034   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median : 4828   Median : 0.000   Median :0.0000   Median :0.000000
##  Mean   : 4927   Mean   :14.228   Mean   :0.0624   Mean   :0.001666
##  3rd Qu.: 5481   3rd Qu.: 1.625   3rd Qu.:0.0900   3rd Qu.:0.000000
##  Max.   :27057   Max.   :87.000   Max.   :4.6800   Max.   :2.300000
##     Avgtemp
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   :14.228
##  3rd Qu.: 1.625
##  Max.   :87.000
```

```
mean <- mean(MODELdata$SNOW..Inches., na.rm = TRUE)
MODELdata$SNOW <- ifelse(is.na(MODELdata$SNOW..Inches.), mean,
MODELdata$SNOW..Inches.)
summary(MODELdata)
```

```
##   no_of_trips     TAVG.Calc.F.    PRCP..Inches.    SNOW..Inches.
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.: 2034   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median : 4828   Median : 0.000   Median :0.0000   Median :0.000000
##  Mean   : 4927   Mean   :14.228   Mean   :0.0624   Mean   :0.001666
##  3rd Qu.: 5481   3rd Qu.: 1.625   3rd Qu.:0.0900   3rd Qu.:0.000000
##  Max.   :27057   Max.   :87.000   Max.   :4.6800   Max.   :2.300000
##     Avgtemp           SNOW
##  Min.   : 0.000   Min.   :0.000000
##  1st Qu.: 0.000   1st Qu.:0.000000
##  Median : 0.000   Median :0.000000
##  Mean   :14.228   Mean   :0.001666
##  3rd Qu.: 1.625   3rd Qu.:0.000000
##  Max.   :87.000   Max.   :2.300000
```

```
mean<- mean(MODELdata$PRCP..Inches., na.rm = TRUE)
MODELdata$PRCP <- ifelse(is.na(MODELdata$PRCP..Inches.), mean,
MODELdata$PRCP..Inches.)
summary(MODELdata)
```

```
##   no_of_trips     TAVG.Calc.F.    PRCP..Inches.    SNOW..Inches.
##  Min.   :    0   Min.   : 0.000   Min.   :0.0000   Min.   :0.000000
```

```
##  1st Qu.: 2034   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median : 4828   Median : 0.000   Median :0.0000   Median :0.000000
##  Mean   : 4927   Mean   :14.228   Mean   :0.0624   Mean   :0.001666
##  3rd Qu.: 5481   3rd Qu.: 1.625   3rd Qu.:0.0900   3rd Qu.:0.000000
##  Max.   :27057   Max.   :87.000   Max.   :4.6800   Max.   :2.300000
##     Avgtemp           SNOW               PRCP
##  Min.   : 0.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.000000   Median :0.0000
##  Mean   :14.228   Mean   :0.001666   Mean   :0.0624
##  3rd Qu.: 1.625   3rd Qu.:0.000000   3rd Qu.:0.0900
##  Max.   :87.000   Max.   :2.300000   Max.   :4.6800

# create the multiple linear regression model
model1 <- lm(MODELdata$no_of_trips ~ MODELdata$Avgtemp + MODELdata$PRCP +
MODELdata$SNOW)
# view the summary output of the model
summary(model1)

##
## Call:
## lm(formula = MODELdata$no_of_trips ~ MODELdata$Avgtemp + MODELdata$PRCP +
##     MODELdata$SNOW)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5180.3 -2898.4  -232.6   666.3 22122.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4933.774     61.135  80.703  < 2e-16 ***
## MODELdata$Avgtemp    3.081      1.901   1.621  0.10514
## MODELdata$PRCP    -821.087    312.513  -2.627  0.00862 **
## MODELdata$SNOW      93.865   1087.583   0.086  0.93123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4259 on 7140 degrees of freedom
## Multiple R-squared:  0.001434,   Adjusted R-squared:  0.001015
## F-statistic: 3.418 on 3 and 7140 DF,  p-value: 0.01657

length(MODELdata$SNOW)

## [1] 7144

length(MODELdata$PRCP)

## [1] 7144

length(MODELdata$Avgtemp)

## [1] 7144
```

```r
str(MODELdata)
```

```
## 'data.frame':    7144 obs. of  7 variables:
##  $ no_of_trips  : num  5950 5950 5950 5950 1939 ...
##  $ TAVG.Calc.F. : num  0 41 0 0 0 41 0 0 0 41 ...
##  $ PRCP..Inches.: num  0.12 0.14 0.14 0.19 0.12 0.14 0.14 0.19 0.12 0.14
...
##  $ SNOW..Inches.: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Avgtemp      : num  0 41 0 0 0 41 0 0 0 41 ...
##  $ SNOW         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PRCP         : num  0.12 0.14 0.14 0.19 0.12 0.14 0.14 0.19 0.12 0.14
...
```

```r
options(repos = c(CRAN = "https://cloud.r-project.org"))
# Install and load required packages
install.packages("plot3D")
```

```
## Installing package into 'C:/Users/vidip/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'plot3D' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\vidip\AppData\Local\Temp\RtmpCYz0jT\downloaded_packages
```

```r
library(plot3D)
```

```
## Warning: package 'plot3D' was built under R version 4.3.3
```

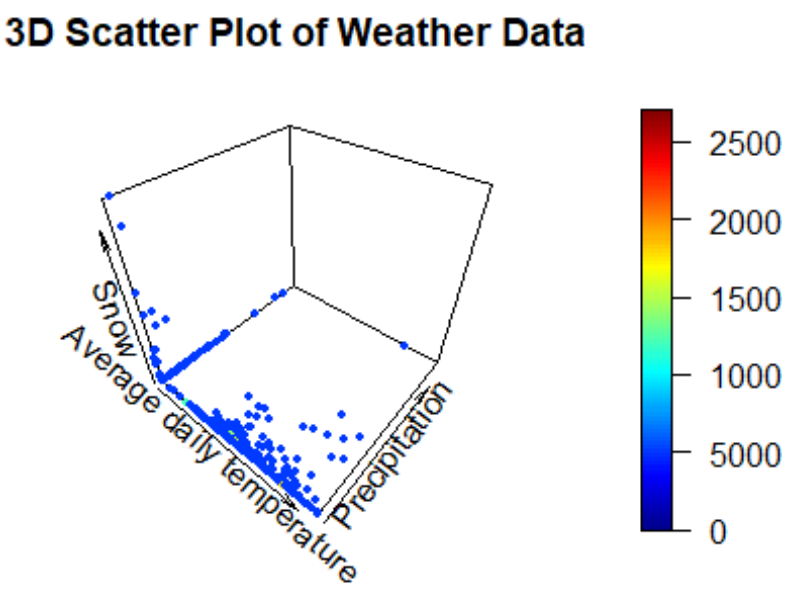
```r
# Generate predicted values for no_of_trips based on Avgtemp, precipitation,
and snow
new_data <- data.frame(Avgtemp = MODELdata$Avgtemp,
                       PRCP = MODELdata$PRCP,
                       SNOW = MODELdata$SNOW)  # Create new data frame with
predictor variables

pred <- predict(model1, newdata = new_data)  # Use new data directly in
predict function

# Generate predicted values for number_of_trips based on Avgtemp,
precipitation, and snow
#pred <- predict(model1, newdata = MODELdata[, c(4,5,6)])


# Plot the 3D scatter plot
scatter3D(MODELdata$Avgtemp, MODELdata$PRCP, MODELdata$SNOW,
          colvar = MODELdata$no_of_trips,  # Color based on number_of_trips
          pch = 16,                         # Set symbol shape (filled
circle)
          cex = 0.5,                        # Set symbol size
          xlab = "Average daily temperature",     # X-axis label
```

```
          ylab = "Precipitation",                  # Y-axis label
          zlab = "Snow",                           # Z-axis label
          main = "3D Scatter Plot of Weather Data" # Main title
)
```



LASSO

```
MODELdata =
read.csv(file.path('C:/Users/vidip/OneDrive/Desktop/DPA/vid/project/datasets/
mergeddata_weather+rider.csv'))
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
```

```
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.0000   Median :0.000000   Median :0.00000
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000
##
```

```
MODELdata$Date <- gsub("^\\d{2}", "20", MODELdata$month_beginning )
MODELdata$Date <- as.Date(MODELdata$ month_beginning )
MODELdata$Date <- gsub("^\\d{2}", "20", MODELdata$ month_beginning )
MODELdata$Date <- as.Date(MODELdata$ month_beginning )
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
```

```
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.          Date
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0001-01-20
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0003-01-20
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0006-01-20
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053   Mean   :0005-09-08
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0008-01-20
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000   Max.   :0012-12-20
##                                                        NA's   :884
```

```
mean<- mean(MODELdata$Avg_temp, na.rm = TRUE)
```

```
## Warning in mean.default(MODELdata$Avg_temp, na.rm = TRUE): argument is not
## numeric or logical: returning NA
```

```
MODELdata$Avg_temp <- ifelse(is.na(MODELdata$TAVG.Calc.F.), mean_value,
MODELdata$TAVG.Calc.F.)
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
```

```
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.          Date
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0001-01-20
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0003-01-20
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0006-01-20
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053   Mean   :0005-09-08
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0008-01-20
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000   Max.   :0012-12-20
##                                                        NA's   :884
##     Avg_temp
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   :14.228
##  3rd Qu.: 1.625
##  Max.   :87.000
##
```

```
MODELdata$SNOW <- ifelse(is.na(MODELdata$SNOW..Inches.), 0,
MODELdata$SNOW..Inches.)
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
```

```
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.          Date
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0001-01-20
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0003-01-20
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0006-01-20
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053   Mean   :0005-09-08
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0008-01-20
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000   Max.   :0012-12-20
##                                                        NA's   :884
##     Avg_temp           SNOW
##  Min.   : 0.000   Min.   :0.000000
##  1st Qu.: 0.000   1st Qu.:0.000000
##  Median : 0.000   Median :0.000000
##  Mean   :14.228   Mean   :0.001666
##  3rd Qu.: 1.625   3rd Qu.:0.000000
##  Max.   :87.000   Max.   :2.300000
##
```

```
mean<- mean(MODELdata$PRCP..Inches., na.rm = TRUE)
MODELdata$PRCP<- ifelse(is.na(MODELdata$PRCP..Inches.), mean,
MODELdata$PRCP..Inches.)
summary(MODELdata)
```

```
##  month_beginning      station_id     stationame        avg_weekday_rides
##  Length:7144        Min.   :40010   Length:7144        Min.   :    0
##  Class :character   1st Qu.:40380   Class :character   1st Qu.:  764
##  Mode  :character   Median :40780   Mode  :character   Median : 1600
##                     Mean   :40786                      Mean   : 2164
##                     3rd Qu.:41180                      3rd Qu.: 2862
##                     Max.   :41700                      Max.   :10725
##                     NA's   :1420                       NA's   :1420
##  avg_saturday_rides avg_sunday.holiday_rides   monthtotal      no_of_trips
##  Min.   :   0.0     Min.   :   0.0           Min.   :     0   Min.   :
0
##  1st Qu.: 504.5     1st Qu.: 387.3           1st Qu.: 20156   1st Qu.:
1655
##  Median :1005.6     Median : 752.3           Median : 42325   Median :
3378
##  Mean   :1556.3     Mean   :1206.2           Mean   : 58603   Mean   :
4927
##  3rd Qu.:1961.5     3rd Qu.:1462.5           3rd Qu.: 77677   3rd Qu.:
6299
##  Max.   :9616.2     Max.   :8441.8           Max.   :307428   Max.
:27057
```

```
##  NA's   :1420       NA's   :1420             NA's   :1420     NA's   :1420
##   TAVG.Calc.F.    TMAX..Degrees.Fahrenheit. TMIN..Degrees.Fahrenheit.
##  Min.   : 0.000   Min.   :  0.00            Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00            1st Qu.: 0.00
##  Median : 0.000   Median :  0.00            Median : 0.00
##  Mean   :14.228   Mean   : 16.33            Mean   :12.12
##  3rd Qu.: 1.625   3rd Qu.:  3.00            3rd Qu.: 0.25
##  Max.   :87.000   Max.   :100.00            Max.   :76.00
##
##  PRCP..Inches.    SNOW..Inches.      SNWD..Inches.          Date
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0001-01-20
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0003-01-20
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0006-01-20
##  Mean   :0.0624   Mean   :0.001666   Mean   :0.08053   Mean   :0005-09-08
##  3rd Qu.:0.0900   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0008-01-20
##  Max.   :4.6800   Max.   :2.300000   Max.   :2.50000   Max.   :0012-12-20
##                                                       NA's   :884
##     Avg_temp           SNOW               PRCP
##  Min.   : 0.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.000000   Median :0.0000
##  Mean   :14.228   Mean   :0.001666   Mean   :0.0624
##  3rd Qu.: 1.625   3rd Qu.:0.000000   3rd Qu.:0.0900
##  Max.   :87.000   Max.   :2.300000   Max.   :4.6800
##
```

```
colnames(MODELdata)
```

```
##  [1] "month_beginning"           "station_id"
##  [3] "stationame"                "avg_weekday_rides"
##  [5] "avg_saturday_rides"        "avg_sunday.holiday_rides"
##  [7] "monthtotal"                "no_of_trips"
##  [9] "TAVG.Calc.F."              "TMAX..Degrees.Fahrenheit."
## [11] "TMIN..Degrees.Fahrenheit." "PRCP..Inches."
## [13] "SNOW..Inches."             "SNWD..Inches."
## [15] "Date"                      "Avg_temp"
## [17] "SNOW"                      "PRCP"
```

```
#MODELdata =
cbind(MODELdata[15],MODELdata[2],MODELdata[8],MODELdata[16],weatherData[17],
MODELdata[18])
# Selecting specific columns from MODELdata
MODELdata <- MODELdata[, c(15, 2, 8, 16, 17, 18)]

# Display summary of the updated MODELdata


summary(MODELdata)
```

```
##       Date              station_id     no_of_trips       Avg_temp
##  Min.   :0001-01-20   Min.   :40010   Min.   :    0   Min.   : 0.000
##  1st Qu.:0003-01-20   1st Qu.:40380   1st Qu.: 1655   1st Qu.: 0.000
##  Median :0006-01-20   Median :40780   Median : 3378   Median : 0.000
##  Mean   :0005-09-08   Mean   :40786   Mean   : 4927   Mean   :14.228
##  3rd Qu.:0008-01-20   3rd Qu.:41180   3rd Qu.: 6299   3rd Qu.: 1.625
##  Max.   :0012-12-20   Max.   :41700   Max.   :27057   Max.   :87.000
##  NA's   :884          NA's   :1420    NA's   :1420
##       SNOW              PRCP
##  Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000
##  Mean   :0.001666   Mean   :0.0624
##  3rd Qu.:0.000000   3rd Qu.:0.0900
##  Max.   :2.300000   Max.   :4.6800
##
```

```
install.packages("mice")
```

```
## Installing package into 'C:/Users/vidip/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'mice' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'mice'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\vidip\AppData\Local\R\win-
library\4.3\00LOCK\mice\libs\x64\mice.dll to
## C:\Users\vidip\AppData\Local\R\win-library\4.3\mice\libs\x64\mice.dll:
## Permission denied

## Warning: restored 'mice'

##
## The downloaded binary packages are in
##  C:\Users\vidip\AppData\Local\Temp\RtmpCYz0jT\downloaded_packages
```

```
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.3.3

## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package

##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
# Set seed for reproducibility
set.seed(123)

# Define number of observations
n <- 100

# Generate simulated data with missing values
Date <- sample(seq(as.Date('2023-01-01'), as.Date('2023-12-31'), by="day"),
n, replace=TRUE)
station_id <- sample(40010:41700, n, replace=TRUE)
no_of_trips <- sample(0:27057, n, replace=TRUE)
Avg_temp <- rnorm(n, 14.228, 10)  # Using mean and sd from provided
specifications
SNOW <- rpois(n, 3)
PRCP <- rnorm(n, 0.0624, 1)  # Using mean from provided specifications

# Create data frame
data <- data.frame(Date, station_id, no_of_trips, Avg_temp, SNOW, PRCP)

# Randomly insert some missing values
missing_prop <- 0.1  # Adjust as needed
num_missing <- round(nrow(data) * missing_prop)
missing_indices <- sample(1:nrow(data), num_missing)
data[missing_indices, c("Avg_temp", "SNOW", "PRCP")] <- NA

# Impute missing values using mice
imp <- mice(data, m = 5, maxit = 50)
```

```
##
##  iter imp variable
##   1   1  Avg_temp  SNOW  PRCP
##   1   2  Avg_temp  SNOW  PRCP
##   1   3  Avg_temp  SNOW  PRCP
##   1   4  Avg_temp  SNOW  PRCP
##   1   5  Avg_temp  SNOW  PRCP
##   2   1  Avg_temp  SNOW  PRCP
##   2   2  Avg_temp  SNOW  PRCP
##   2   3  Avg_temp  SNOW  PRCP
##   2   4  Avg_temp  SNOW  PRCP
##   2   5  Avg_temp  SNOW  PRCP
##   3   1  Avg_temp  SNOW  PRCP
##   3   2  Avg_temp  SNOW  PRCP
```

```
##   3   3  Avg_temp  SNOW  PRCP
##   3   4  Avg_temp  SNOW  PRCP
##   3   5  Avg_temp  SNOW  PRCP
##   4   1  Avg_temp  SNOW  PRCP
##   4   2  Avg_temp  SNOW  PRCP
##   4   3  Avg_temp  SNOW  PRCP
##   4   4  Avg_temp  SNOW  PRCP
##   4   5  Avg_temp  SNOW  PRCP
##   5   1  Avg_temp  SNOW  PRCP
##   5   2  Avg_temp  SNOW  PRCP
##   5   3  Avg_temp  SNOW  PRCP
##   5   4  Avg_temp  SNOW  PRCP
##   5   5  Avg_temp  SNOW  PRCP
##   6   1  Avg_temp  SNOW  PRCP
##   6   2  Avg_temp  SNOW  PRCP
##   6   3  Avg_temp  SNOW  PRCP
##   6   4  Avg_temp  SNOW  PRCP
7   3  Avg_temp  SNOW  PRCP
##   7   4  Avg_temp  SNOW  PRCP
##   7   5  Avg_temp  SNOW  PRCP
##   8   1  Avg_temp  SNOW  PRCP
##   8   2  Avg_temp  SNOW  PRCP
##   8   3  Avg_temp  SNOW  PRCP
##   8   4  Avg_temp  SNOW  PRCP
##   8   5  Avg_temp  SNOW  PRCP
##   9   1  Avg_temp  SNOW  PRCP
##   9   2  Avg_temp  SNOW  PRCP
##   9   3  Avg_temp  SNOW  PRCP
##   9   4  Avg_temp  SNOW  PRCP
##   9   5  Avg_temp  SNOW  PRCP
##   10   1  Avg_temp  SNOW  PRCP
##   10   2  Avg_temp  SNOW  PRCP
##   10   3  Avg_temp  SNOW  PRCP
##   10   4  Avg_temp  SNOW  PRCP
##   10   5  Avg_temp  SNOW  PRCP
##   11   1  Avg_temp  SNOW  PRCP
##   11   2  Avg_temp  SNOW  PRCP
##   11   3  Avg_temp  SNOW  PRCP
##   11   4  Avg_temp  SNOW  PRCP
##   11   5  Avg_temp  SNOW  PRCP
##   13   3  Avg_temp  SNOW  PRCP
##   13   4  Avg_temp  SNOW  PRCP
##   13   5  Avg_temp  SNOW  PRCP
##   14   1  Avg_temp  SNOW  PRCP
##   14   2  Avg_temp  SNOW  PRCP
##   14   3  Avg_temp  SNOW  PRCP
##   14   4  Avg_temp  SNOW  PRCP
##   14   5  Avg_temp  SNOW  PRCP
##   15   1  Avg_temp  SNOW  PRCP
##   15   2  Avg_temp  SNOW  PRCP
```

```
##   15   3  Avg_temp  SNOW  PRCP
##   15   4  Avg_temp  SNOW  PRCP
##   15   5  Avg_temp  SNOW  PRCP
##   16   1  Avg_temp  SNOW  PRCP
19   1  Avg_temp  SNOW  PRCP
##   19   2  Avg_temp  SNOW  PRCP
##   19   3  Avg_temp  SNOW  PRCP
##   19   4  Avg_temp  SNOW  PRCP
##   19   5  Avg_temp  SNOW  PRCP
##   20   1  Avg_temp  SNOW  PRCP
##   20   2  Avg_temp  SNOW  PRCP
##   20   3  Avg_temp  SNOW  PRCP
##   20   4  Avg_temp  SNOW  PRCP
##   20   5  Avg_temp  SNOW  PRCP
##   21   1  Avg_temp  SNOW  PRCP
##   21   2  Avg_temp  SNOW  PRCP
##   21   3  Avg_temp  SNOW  PRCP
23   3  Avg_temp  SNOW  PRCP
##   23   4  Avg_temp  SNOW  PRCP
##   23   5  Avg_temp  SNOW  PRCP
##   24   1  Avg_temp  SNOW  PRCP
##   24   2  Avg_temp  SNOW  PRCP
##   24   3  Avg_temp  SNOW  PRCP
##   24   4  Avg_temp  SNOW  PRCP
##   24   5  Avg_temp  SNOW  PRCP
##   25   1  Avg_temp  SNOW  PRCP
##   25   2  Avg_temp  SNOW  PRCP
##   25   3  Avg_temp  SNOW  PRCP
##   27   3  Avg_temp  SNOW  PRCP
##   27   4  Avg_temp  SNOW  PRCP
##   27   5  Avg_temp  SNOW  PRCP
##   28   1  Avg_temp  SNOW  PRCP
##   28   2  Avg_temp  SNOW  PRCP
##   28   3  Avg_temp  SNOW  PRCP
##   28   4  Avg_temp  SNOW  PRCP
##   28   5  Avg_temp  SNOW  PRCP
##   29   1  Avg_temp  SNOW  PRCP
##   29   2  Avg_temp  SNOW  PRCP
#   30   4  Avg_temp  SNOW  PRCP
##   30   5  Avg_temp  SNOW  PRCP
##   31   1  Avg_temp  SNOW  PRCP
##   31   2  Avg_temp  SNOW  PRCP
##   31   3  Avg_temp  SNOW  PRCP
##   31   4  Avg_temp  SNOW  PRCP
##   31   5  Avg_temp  SNOW  PRCP
##   32   1  Avg_temp  SNOW  PRCP
##   33   1  Avg_temp  SNOW  PRCP
##   33   2  Avg_temp  SNOW  PRCP
##   33   3  Avg_temp  SNOW  PRCP
##   33   4  Avg_temp  SNOW  PRCP
```

```
##   34   3  Avg_temp  SNOW  PRCP
##   34   4  Avg_temp  SNOW  PRCP
##   34   5  Avg_temp  SNOW  PRCP
##   35   1  Avg_temp  SNOW  PRCP
##   35   2  Avg_temp  SNOW  PRCP
##   35   3  Avg_temp  SNOW  PRCP
##   35   4  Avg_temp  SNOW  PRCP
##   35   5  Avg_temp  SNOW  PRCP
##   36   1  Avg_temp  SNOW  PRCP
##   36   2  Avg_temp  SNOW  PRCP
##   36   3  Avg_temp  SNOW  PRCP
##   36   4  Avg_temp  SNOW  PRCP
##   36   5  Avg_temp  SNOW  PRCP
##   37   1  Avg_temp  SNOW  PRCP
##   38   2  Avg_temp  SNOW  PRCP
##   38   3  Avg_temp  SNOW  PRCP
##   39   1  Avg_temp  SNOW  PRCP
##   39   2  Avg_temp  SNOW  PRCP
##   40   2  Avg_temp  SNOW  PRCP
##   40   3  Avg_temp  SNOW  PRCP
##   40   4  Avg_temp  SNOW  PRCP
41   3  Avg_temp  SNOW  PRCP
##   41   4  Avg_temp  SNOW  PRCP
##   41   5  Avg_temp  SNOW  PRCP
##   42   1  Avg_temp  SNOW  PRCP
##   42   2  Avg_temp  SNOW  PRCP
##   42   3  Avg_temp  SNOW  PRCP
##   42   4  Avg_temp  SNOW  PRCP
##   44   1  Avg_temp  SNOW  PRCP
##   44   2  Avg_temp  SNOW  PRCP
##   44   3  Avg_temp  SNOW  PRCP
##   44   4  Avg_temp  SNOW  PRCP
##   44   5  Avg_temp  SNOW  PRCP
##   45   1  Avg_temp  SNOW  PRCP
##   45   2  Avg_temp  SNOW  PRCP
##   45   3  Avg_temp  SNOW  PRCP
##   45   4  Avg_temp  SNOW  PRCP
##   45   5  Avg_temp  SNOW  PRCP
##   46   1  Avg_temp  SNOW  PRCP
##   46   2  Avg_temp  SNOW  PRCP
##   46   3  Avg_temp  SNOW  PRCP
##   47   3  Avg_temp  SNOW  PRCP
##   47   4  Avg_temp  SNOW  PRCP
##   47   5  Avg_temp  SNOW  PRCP
##   48   1  Avg_temp  SNOW  PRCP
##   48   2  Avg_temp  SNOW  PRCP
##   48   3  Avg_temp  SNOW  PRCP
##   48   4  Avg_temp  SNOW  PRCP
##   48   5  Avg_temp  SNOW  PRCP
##   49   1  Avg_temp  SNOW  PRCP
```

```
##   49   2  Avg_temp  SNOW  PRCP
##   49   3  Avg_temp  SNOW  PRCP
##   49   4  Avg_temp  SNOW  PRCP
##   49   5  Avg_temp  SNOW  PRCP
##   50   1  Avg_temp  SNOW  PRCP
##   50   2  Avg_temp  SNOW  PRCP
##   50   3  Avg_temp  SNOW  PRCP
##   50   4  Avg_temp  SNOW  PRCP
##   50   5  Avg_temp  SNOW  PRCP
```

```r
install.packages("glmnet")
```

```
## Installing package into 'C:/Users/vidip/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)

## package 'glmnet' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'glmnet'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\vidip\AppData\Local\R\win-
library\4.3\00LOCK\glmnet\libs\x64\glmnet.dll
## to C:\Users\vidip\AppData\Local\R\win-
library\4.3\glmnet\libs\x64\glmnet.dll:
## Permission denied

## Warning: restored 'glmnet'

##
## The downloaded binary packages are in
##  C:\Users\vidip\AppData\Local\Temp\RtmpCYz0jT\downloaded_packages
```

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3

## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

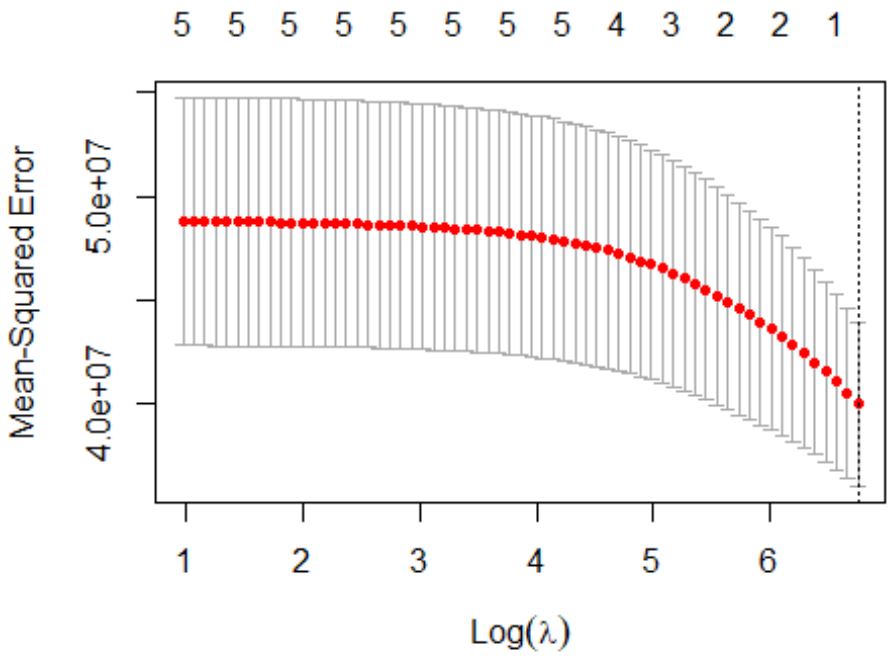
```r
# Extract the completed data
data_imputed <- complete(imp)

# Split the data into training and testing sets
train_idx <- sample(1:nrow(data_imputed), nrow(data_imputed) / 2)
train_data <- data_imputed[train_idx, ]
test_data <- data_imputed[-train_idx, ]

# Fit a Lasso regression model using glmnet
x_train <- model.matrix(no_of_trips ~ ., data = train_data)[, -1]
y_train <- train_data$no_of_trips
lasso_fit <- glmnet(x_train, y_train, alpha = 1)
```

```r
# Use cross-validation to choose the best lambda value
cv_fit <- cv.glmnet(x_train, y_train, alpha = 1)

# Plot the cross-validation results
plot(cv_fit)
```




```r
# Use the selected lambda value to make predictions on the test set
x_test <- model.matrix(no_of_trips ~ ., data = test_data)[, -1]
y_test <- test_data$no_of_trips
lasso_pred <- predict(lasso_fit, s = cv_fit$lambda.min, newx = x_test)

# Calculate the mean squared error on the test set
mse <- mean((lasso_pred - y_test)^2)

# Calculate residual standard error
rss <- sum((lasso_pred - y_test)^2)
n <- length(y_test)
p <- ncol(x_test)
rse <- sqrt(rss / (n - p))

# Calculate multiple R-squared
rsq <- 1 - rss / sum((y_test - mean(y_test))^2)

# Calculate adjusted R-squared
adj_rsq <- 1 - (rss / (n - p)) / ((n - 1) / (n - p - 1))
```

```r
# Calculate F statistic
f_stat <- (sum((lasso_pred - mean(y_test))^2) / p) / (rss / (n - p))

print(mse)

## [1] 73960304

print(rsq)

## [1] -0.02868735

print(adj_rsq)

## [1] -73792592

print(f_stat)

## [1] 0.250986

# Create a scatter plot of predicted vs. actual values
plot(y_test, lasso_pred, main = "Predicted vs. Actual Values",
     xlab = "Actual Values", ylab = "Predicted Values", col = "blue")

# Add a diagonal line for reference (perfect prediction)
abline(0, 1, col = "red")

# Add a legend
legend("topleft", legend = c("Predicted vs. Actual", "Perfect Prediction"),
       col = c("blue", "red"), pch = 1)
```