EDA

Xin

2024-05-14

```
library(electBook)
```

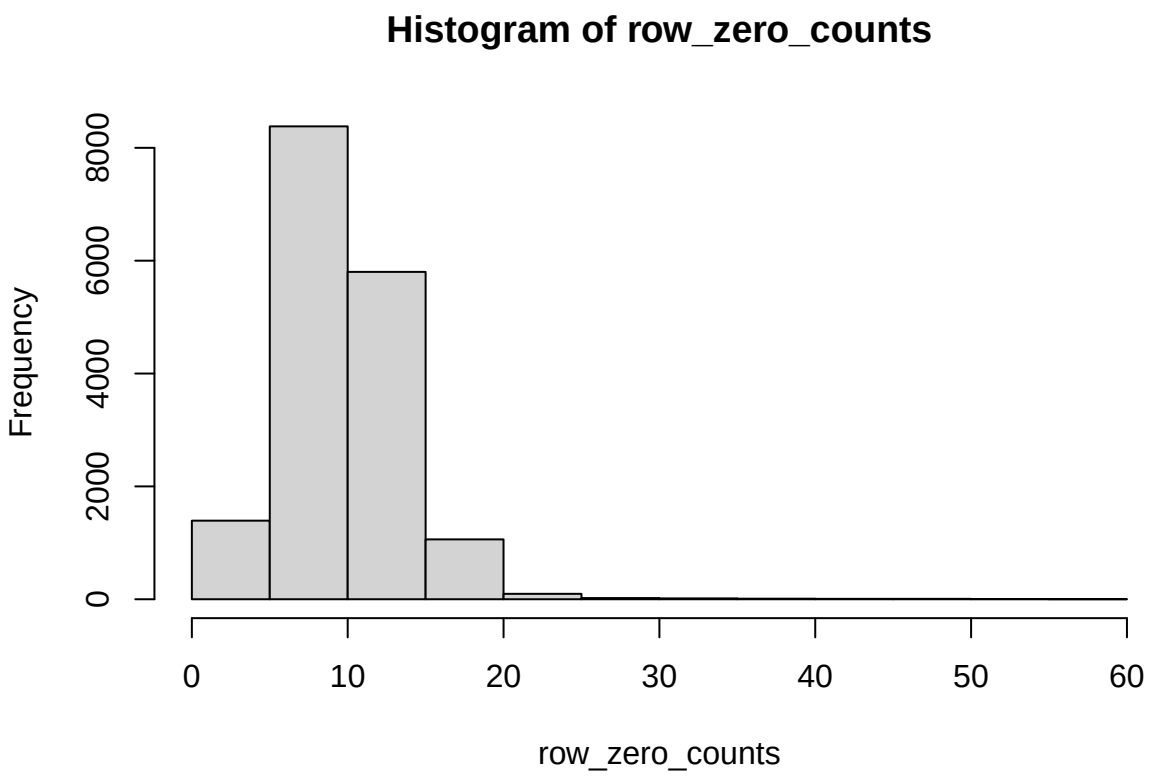
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

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

```
head(Irish$indCons[,1:10])
```

```
##      I1002 I1003 I1004 I1005 I1013 I1015 I1018 I1020 I1022 I1024
## 8114 0.022 0.593 2.002 0.755 0.035 0.398 0.547 0.376 0.229 1.030
## 8115 0.133 0.707 1.602 0.898 0.112 0.689 0.603 0.275 0.198 0.807
## 8116 0.094 0.684 1.525 0.736 0.046 0.407 0.511 0.259 0.201 0.859
## 8117 0.023 0.563 1.393 0.738 0.036 0.223 0.593 0.249 0.212 0.210
## 8118 0.133 0.489 1.221 0.849 0.065 0.132 0.570 0.241 0.121 0.056
## 8119 0.090 0.521 1.032 0.695 0.093 0.117 0.481 0.122 0.127 0.169
```

```
# Count zeros in each row
row_zero_counts <- rowSums(Irish$indCons == 0)

hist(row_zero_counts)
```



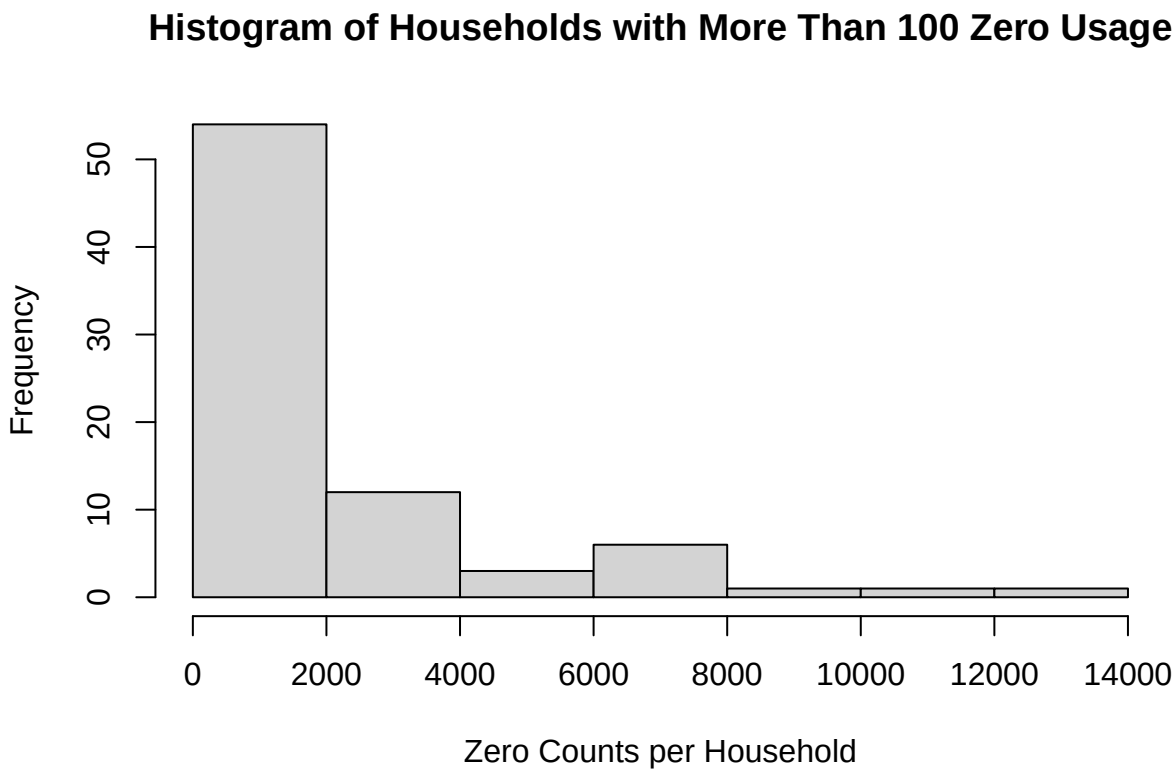

```
sum(apply(Irish$indCons, 2, function(x) sum(x == 0)))
```

```
## [1] 170228
```

```
#There are many zeros in the demand data frame
print(16799*2674)
```

```
## [1] 44920526
```

```
# Count zeros in each column
col_zero_counts <- colSums(Irish$indCons == 0)

# Histogram of columns with more than 100 zeros
hist(col_zero_counts[col_zero_counts > 100], main="Histogram of Households with More Than 100 Zero Usag
```




```
sum(col_zero_counts > 30*48)
```

```
## [1] 27
```

```
cols_to_remove <- which(col_zero_counts > 30*48)
```

```
df0 <- Irish$indCons[,-cols_to_remove]

df0$time_mean_dem <- rowSums(Irish$indCons)/ncol(Irish$indCons)
```

```
head(Irish$extra)
```

```
##   time       toy dow  holy tod temp            dateTime
## 1    1 0.9863014 Wed FALSE   0    4 2009-12-29 23:00:00
## 2    2 0.9863014 Wed FALSE   1    4 2009-12-29 23:30:00
## 3    3 0.9863014 Wed FALSE   2    4 2009-12-30 00:00:00
## 4    4 0.9863014 Wed FALSE   3    4 2009-12-30 00:30:00
## 5    5 0.9863014 Wed FALSE   4    4 2009-12-30 01:00:00
## 6    6 0.9863014 Wed FALSE   5    4 2009-12-30 01:30:00
```

```
df <- cbind(df0[,"time_mean_dem"],Irish$extra)
colnames(df) <- c("time_mean_demand", colnames(Irish$extra))
```

```r
head(df)
```

```
##   time_mean_demand time       toy dow  holy tod temp            dateTime
## 1        0.6266460    1 0.9863014 Wed FALSE   0    4 2009-12-29 23:00:00
## 2        0.5256755    2 0.9863014 Wed FALSE   1    4 2009-12-29 23:30:00
## 3        0.4419034    3 0.9863014 Wed FALSE   2    4 2009-12-30 00:00:00
## 4        0.3827193    4 0.9863014 Wed FALSE   3    4 2009-12-30 00:30:00
## 5        0.3282253    5 0.9863014 Wed FALSE   4    4 2009-12-30 01:00:00
## 6        0.2903952    6 0.9863014 Wed FALSE   5    4 2009-12-30 01:30:00
```

Visualizing main characteristics

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

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

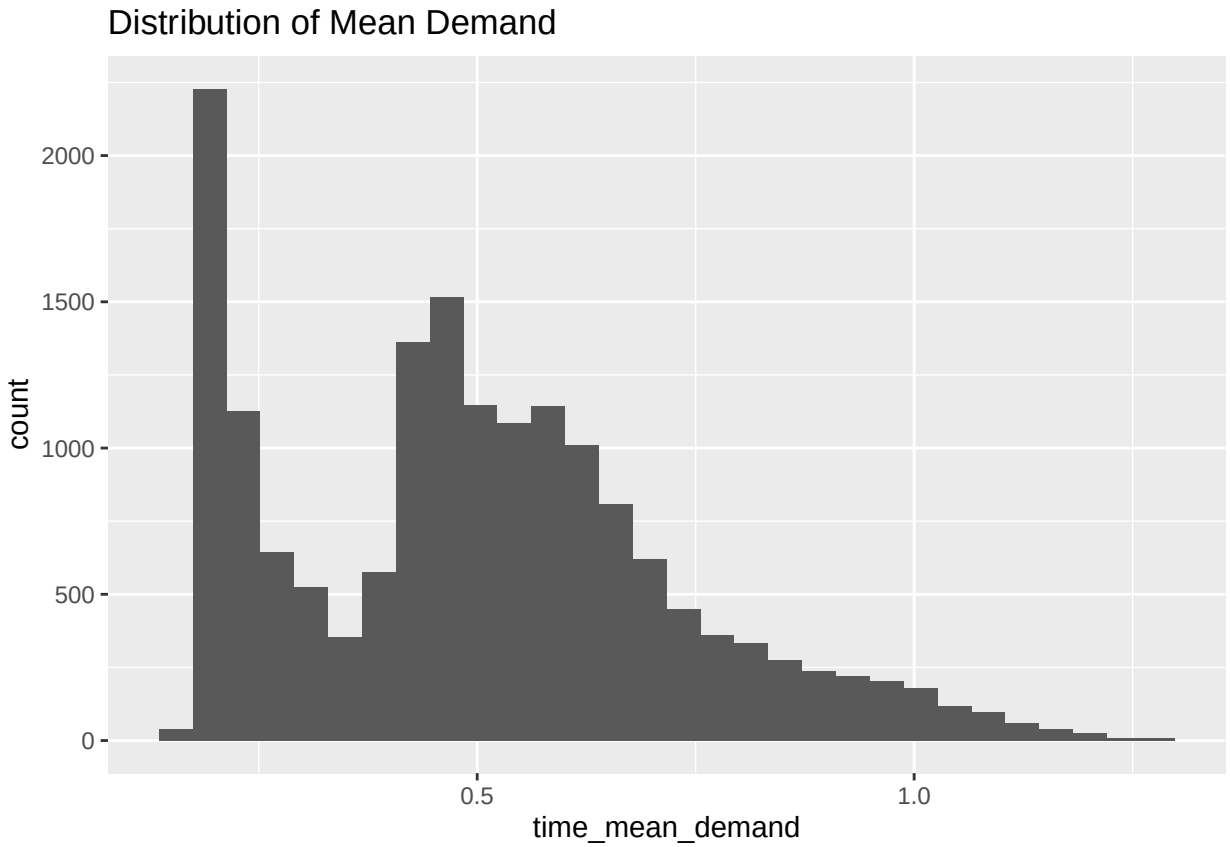
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

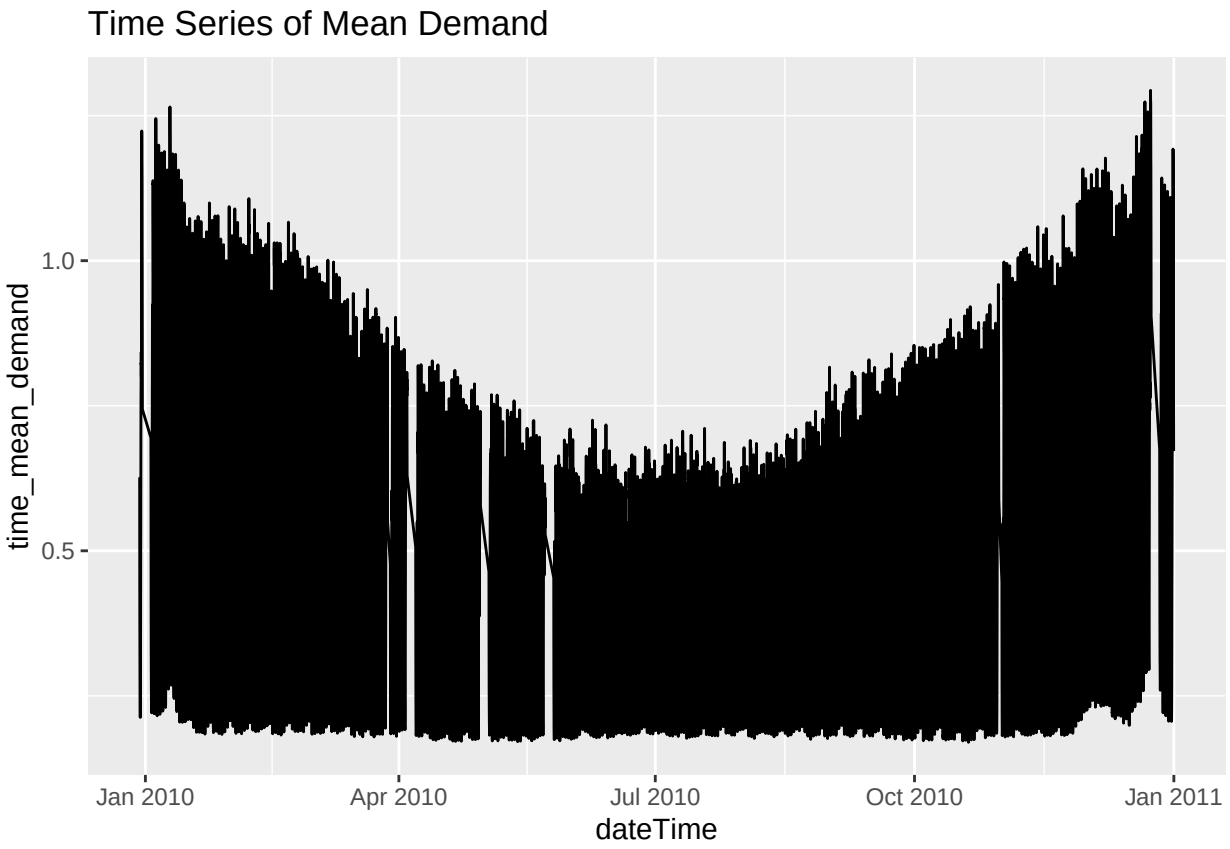
```r
# Basic summary of each column
summary(df)
```

```
##  time_mean_demand      time            toy           dow          holy
##  Min.   :0.1699   Min.   :    1   Min.   :0.0000   Sun:2208   Mode :logical
##  1st Qu.:0.3005   1st Qu.: 4200   1st Qu.:0.2411   Thu:2496   FALSE:16799
##  Median :0.4854   Median : 8400   Median :0.5041   Mon:2400
##  Mean   :0.4994   Mean   : 8400   Mean   :0.4975   Tue:2400
##  3rd Qu.:0.6320   3rd Qu.:12600   3rd Qu.:0.7452   Wed:2544
##  Max.   :1.2936   Max.   :16799   Max.   :0.9918   Sat:2352
##                                                    Fri:2399
##       tod             temp           dateTime
##  Min.   : 0.0   Min.   :-10.000   Min.   :2009-12-29 23:00:00.00
##  1st Qu.:12.0   1st Qu.:  4.000   1st Qu.:2010-03-31 10:45:00.00
##  Median :24.0   Median :  9.000   Median :2010-07-05 22:30:00.00
##  Mean   :23.5   Mean   :  8.616   Mean   :2010-07-03 00:08:03.46
##  3rd Qu.:35.5   3rd Qu.: 14.000   3rd Qu.:2010-10-01 10:15:00.00
##  Max.   :47.0   Max.   : 24.000   Max.   :2010-12-31 22:30:00.00
##
```

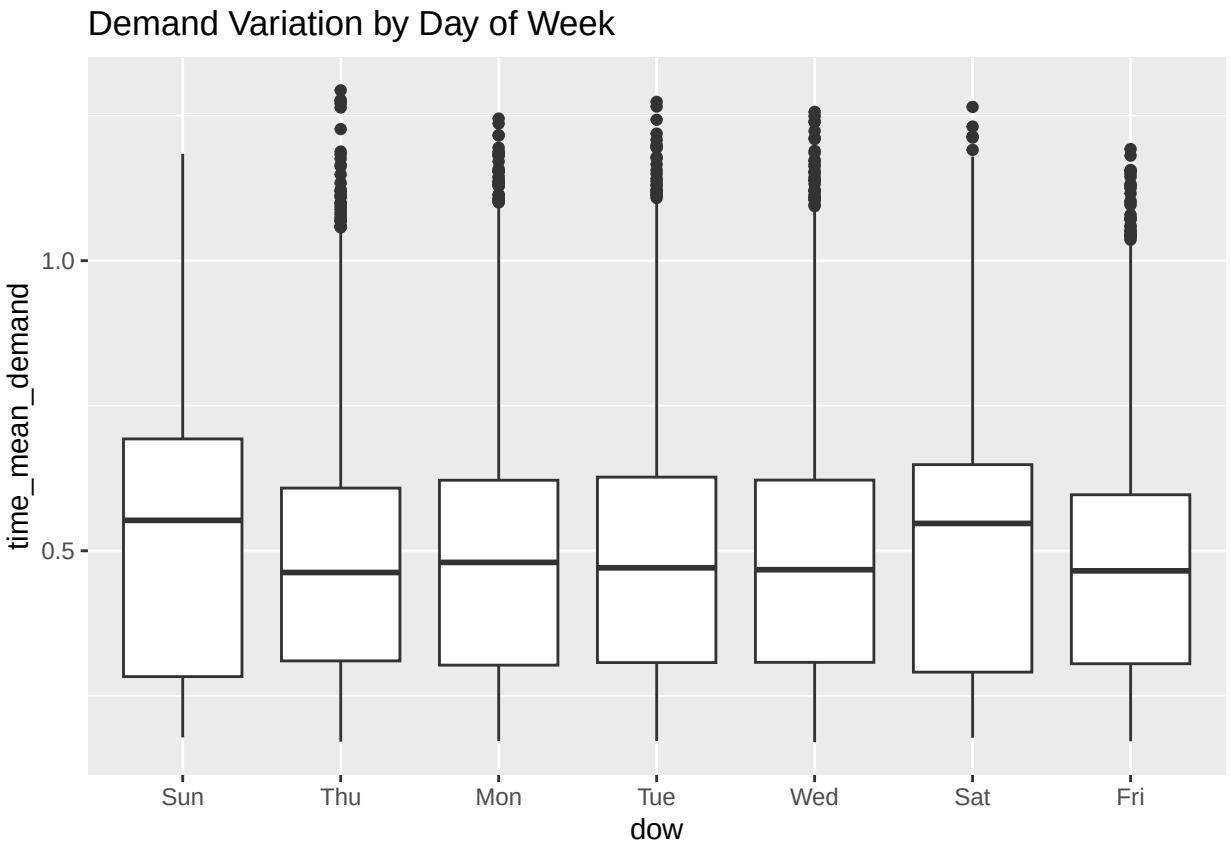
```r
#`holy` is all FALSE

# Visualizing distribution of time_mean_demand
ggplot(df, aes(x=time_mean_demand)) + geom_histogram(bins=30) + ggtitle("Distribution of Mean Demand")
```




```r
# Time series plot of time_mean_demand
ggplot(df, aes(x=dateTime, y=time_mean_demand)) + geom_line() + ggtitle("Time Series of Mean Demand")
```

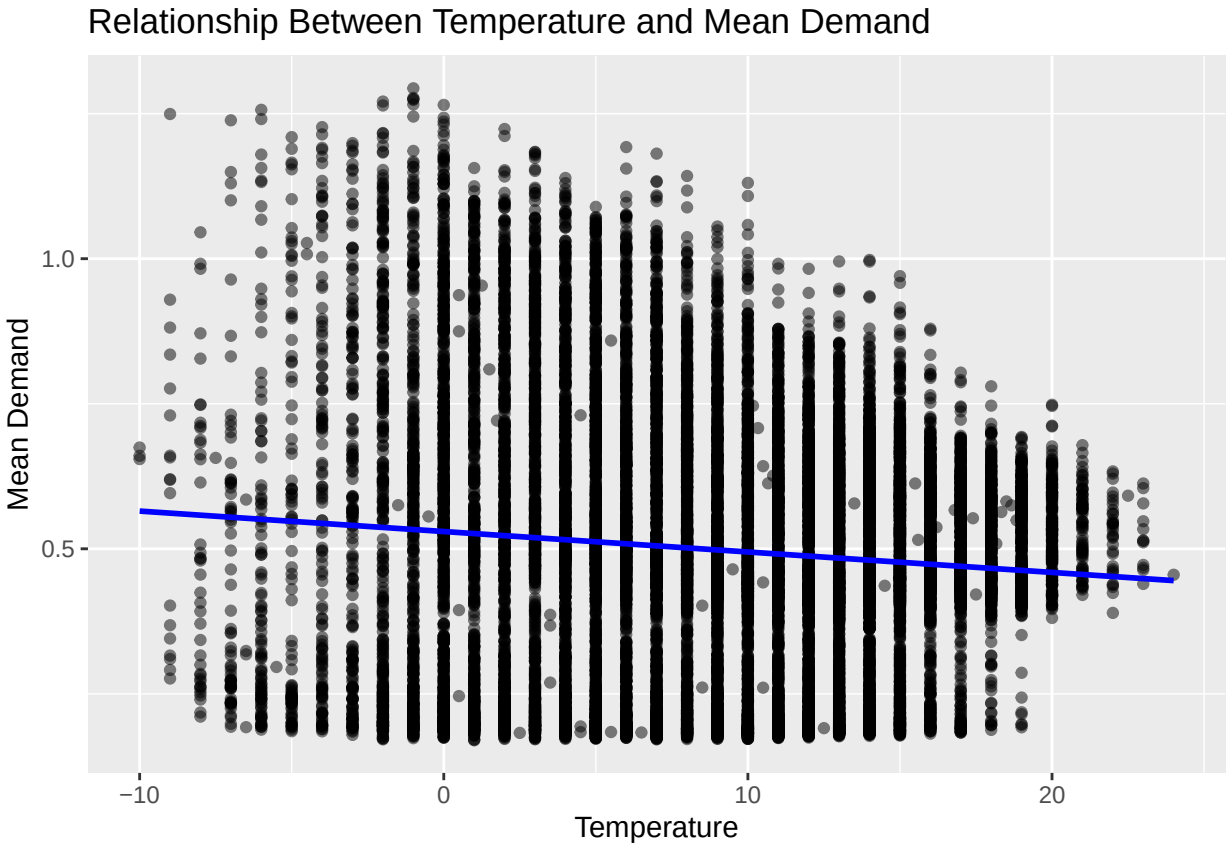



```
# Boxplots to check variation of time_mean_demand across days of the week
ggplot(df, aes(x=dow, y=time_mean_demand)) + geom_boxplot() + ggtitle("Demand Variation by Day of Week"
```



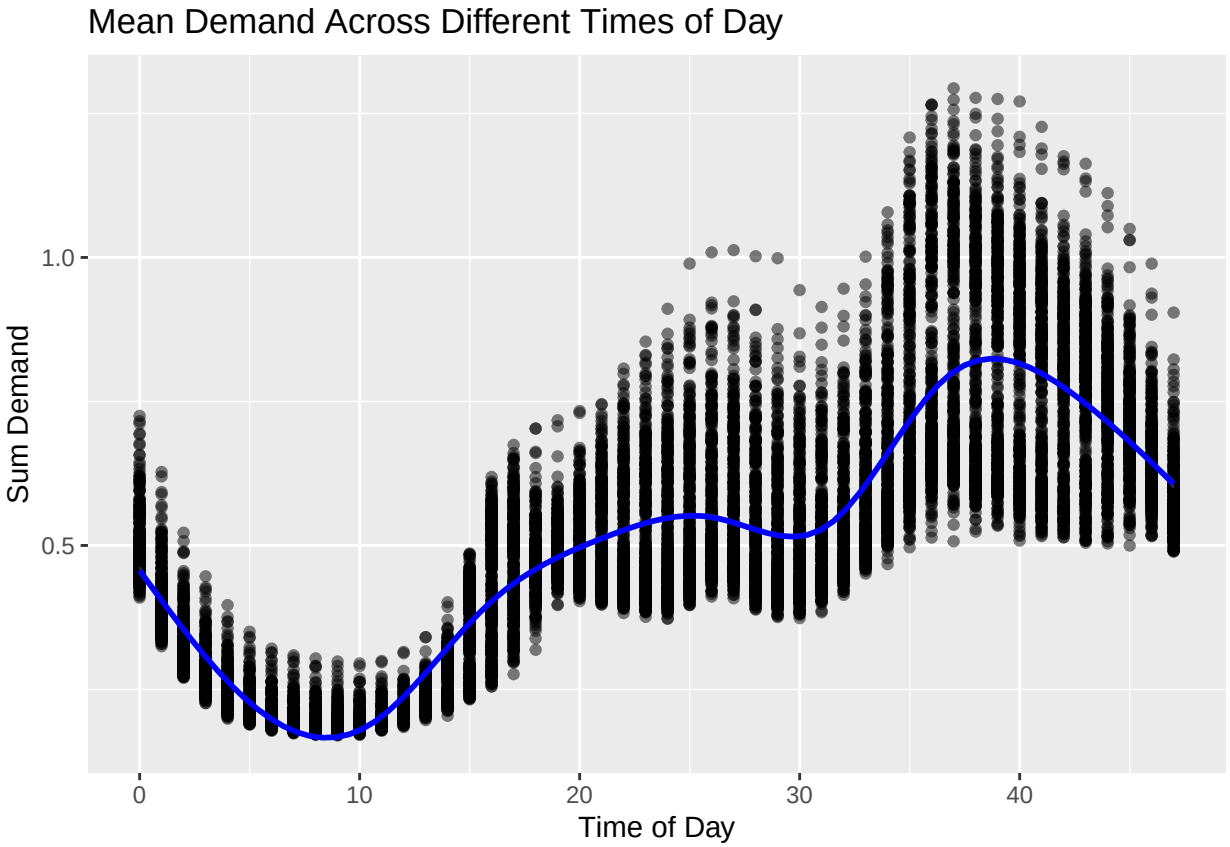

```
# Scatter plot of time_mean_demand vs. temperature
ggplot(df, aes(x=temp, y=time_mean_demand)) +
    geom_point(alpha=0.5) +
    geom_smooth(method="lm", se=FALSE, color="blue") +
    labs(x="Temperature", y="Mean Demand", title="Relationship Between Temperature and Mean Demand")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Line plot for time_mean_demand across different times of day
ggplot(df, aes(x=tod, y=time_mean_demand, group=1)) +
    geom_point(alpha=0.5) +
    geom_smooth(color="blue") +
    labs(x="Time of Day", y="Sum Demand", title="Mean Demand Across Different Times of Day")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```


Mean Demand Across Different Times of Day
Sum Demand
1.0
0.5
0
10
20
30
40
Time of Day