

EDA

Ken Ye, Ejay Lin, Gorden Gao

2023-09-11

Load Libraries & Data

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

```
library(ggplot2)
library(dplyr)
```

```
# make sure you are in the EMS Stations Project directory
# alternatively, click the emsData.RData file to load it into your global environment
load("./Data/emsData.RData")
```

```
head(x)
```

```
##      REF.GRID DISPATCH.PRIORITY.NAME REF.GPS.LAT REF.GPS.LON BASE.NAME VEH.GRID
## 1 3 South Emergency 36.3085 -78.4563 Company 9 Medic 5
## 2 2 Central Emergency 36.3306 -78.4040 Company 9 Medic 6
## 3 2 Central Emergency 36.3335 -78.4399 Company 9 Medic 1
## 4 2 Central Emergency 36.3351 -78.4410 Company 9 Medic 5
## 5 2 Central Non Emergency 36.3401 -78.4017 Company 9 Medic 6
## 6 2 Central Emergency 36.3315 -78.3929 Company 9 Medic 1
##      VEHCGPS DT.DISP DT.ENROUTE DT.ARRIVE
## 1 36.345, -78.3905 2024-01-01 06:46:00 2024-01-01 06:46:00 2024-01-01 06:52:00
## 2 36.345, -78.3905 2024-01-01 08:30:00 2024-01-01 08:30:00 2024-01-01 08:34:00
## 3 36.345, -78.3905 2024-01-01 10:22:00 2024-01-01 10:22:00 2024-01-01 10:27:00
## 4 36.345, -78.3905 2024-01-01 11:38:00 2024-01-01 11:38:00 2024-01-01 11:44:00
## 5 36.345, -78.3905 2024-01-01 12:33:00 2024-01-01 12:33:00 2024-01-01 12:37:00
## 6 36.345, -78.3905 2024-01-01 14:18:00 2024-01-01 14:18:00 2024-01-01 14:22:00
##      DT.LVREF DT.ARVREC DT.AVAILABLE
## 1 2024-01-01 07:07:00 2024-01-01 07:13:00 2024-01-01 07:32:00
## 2 2024-01-01 08:39:00 2024-01-01 08:46:00 2024-01-01 09:00:00
## 3 2024-01-01 10:36:00 2024-01-01 10:39:00 2024-01-01 10:54:00
## 4 <NA> <NA> 2024-01-01 12:08:00
## 5 2024-01-01 12:38:00 2024-01-01 12:45:00 2024-01-01 12:52:00
## 6 2024-01-01 14:38:00 2024-01-01 14:47:00 2024-01-01 15:11:00
##      REC.NAME REC.LON REC.LAT observedTT onSceneDur toHospitalTT
## 1 Maria Parham Hospital -78.44931 36.33089 360 secs 900 secs 360 secs
## 2 Maria Parham Hospital -78.44931 36.33089 240 secs 300 secs 420 secs
## 3 Maria Parham Hospital -78.44931 36.33089 300 secs 540 secs 180 secs
## 4 NA NA 360 secs NA secs NA secs
```

```

## 5 Maria Parham Hospital -78.44931 36.33089 240 secs 60 secs 420 secs
## 6 Maria Parham Hospital -78.44931 36.33089 240 secs 960 secs 540 secs
## atHospitalDur arriveToClearTime Dist.So Dist.Ce Dist.NN Dist.FN eTT.GL.So
## 1 1140 secs 2400 secs 9258 8434 17426 25709 561
## 2 840 secs 1560 secs 7048 2422 12212 20495 578
## 3 900 secs 1620 secs 10969 5301 12540 20823 759
## 4 NA secs 1440 secs 8781 5068 12307 20590 734
## 5 420 secs 900 secs 8967 1516 12228 20511 770
## 6 1440 secs 2940 secs 7800 2298 13267 21550 696
## eTT.GL.Ce eTT.GL.NN eTT.GL.FN eTT.Pe.So eTT.Pe.Ce eTT.Pe.NN eTT.Pe.FN
## 1 411 827 1198 616 440 859 1267
## 2 234 635 1007 650 251 689 1076
## 3 366 752 1124 918 384 796 1217
## 4 298 685 1056 1026 363 784 1193
## 5 191 656 1027 965 220 725 1123
## 6 245 795 1166 890 250 963 1358
## eTT.BG.So eTT.BG.Ce eTT.BG.NN eTT.BG.FN eTT.Op.So eTT.Op.Ce eTT.Op.NN
## 1 539 406 805 1173 507 372 758
## 2 549 220 633 993 508 210 587
## 3 752 346 743 1127 699 343 728
## 4 770 306 712 1085 679 283 671
## 5 766 181 670 1036 728 187 642
## 6 722 235 815 1186 668 243 753
## eTT.Op.FN hosp.Dist hosp.GL eTT.Pe.Hosp eTT.BG.Hosp eTT.Op.Hosp
## 1 1097 3151 309 300 271 267
## 2 933 6060 443 489 404 393
## 3 1090 1372 219 222 178 208
## 4 1032 NA NA NA NA NA
## 5 994 6076 447 557 438 414
## 6 1124 8416 557 661 553 531

```

```
colnames(x)
```

```

## [1] "REF.GRID" "DISPATCH.PRIORITY.NAME" "REF.GPS.LAT"
## [4] "REF.GPS.LON" "BASE.NAME" "VEH.GRID"
## [7] "VEHCGPS" "DT.DISP" "DT.ENROUTE"
## [10] "DT.ARRIVE" "DT.LVREF" "DT.ARVREC"
## [13] "DT.AVAILABLE" "REC.NAME" "REC.LON"
## [16] "REC.LAT" "observedTT" "onSceneDur"
## [19] "toHospitalTT" "atHospitalDur" "arriveToClearTime"
## [22] "Dist.So" "Dist.Ce" "Dist.NN"
## [25] "Dist.FN" "eTT.GL.So" "eTT.GL.Ce"
## [28] "eTT.GL.NN" "eTT.GL.FN" "eTT.Pe.So"
## [31] "eTT.Pe.Ce" "eTT.Pe.NN" "eTT.Pe.FN"
## [34] "eTT.BG.So" "eTT.BG.Ce" "eTT.BG.NN"
## [37] "eTT.BG.FN" "eTT.Op.So" "eTT.Op.Ce"
## [40] "eTT.Op.NN" "eTT.Op.FN" "hosp.Dist"
## [43] "hosp.GL" "eTT.Pe.Hosp" "eTT.BG.Hosp"
## [46] "eTT.Op.Hosp"

```

EDA

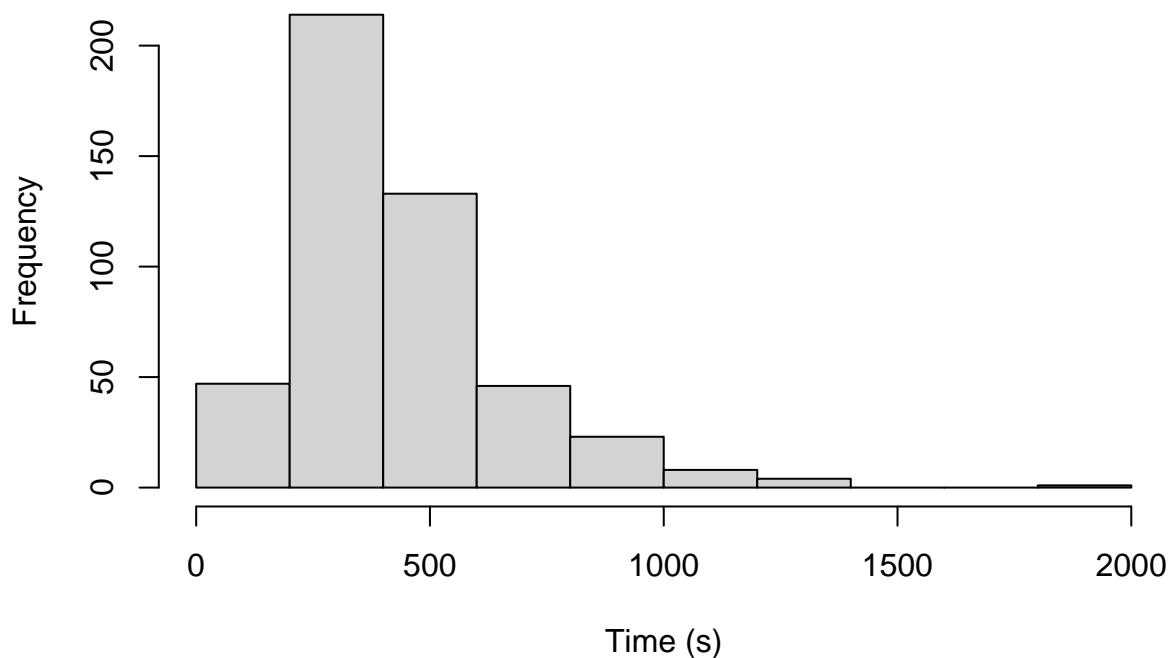
Observed Response Time

```
# observed response time distribution
x$observedTT_numeric <- as.numeric(x$observedTT)
# remove rows with NA and 0 values in the observedTT_numeric column
x <- x[!is.na(x$observedTT_numeric) & x$observedTT_numeric != 0, ]
summary(x$observedTT_numeric)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       60     300     360     432     540    1980
```

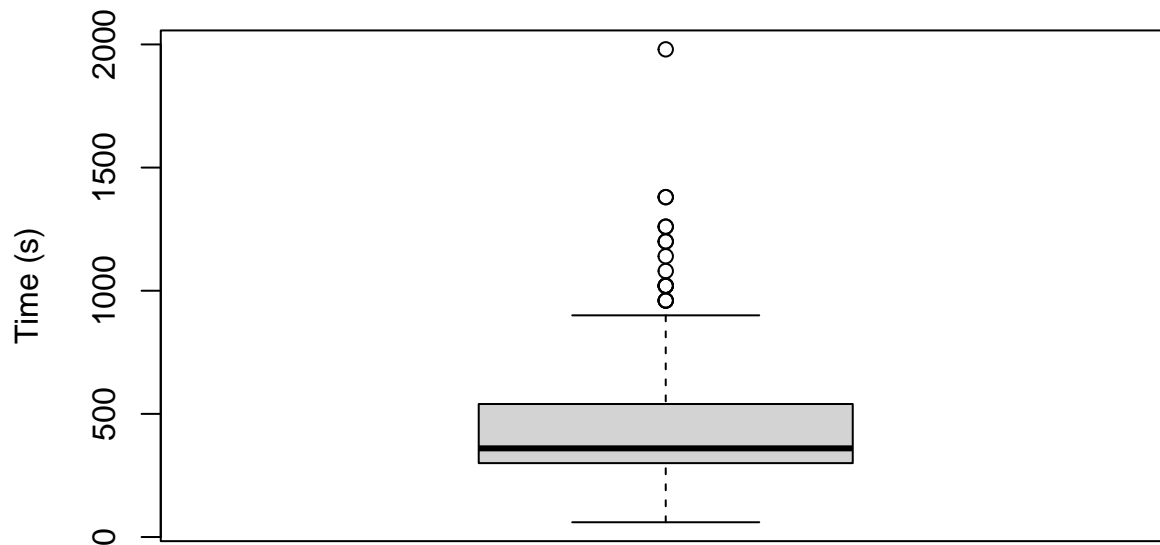
```
# observed response time histogram
hist(x$observedTT_numeric,
     main = "Distribution of Observed Response Time",
     xlab = "Time (s)",
     ylab = "Frequency")
```

Distribution of Observed Response Time



```
# observed response time boxplot
boxplot(x$observedTT_numeric,
     main = "Distribution of Observed Response Time",
     ylab = "Time (s)")
```

Distribution of Observed Response Time



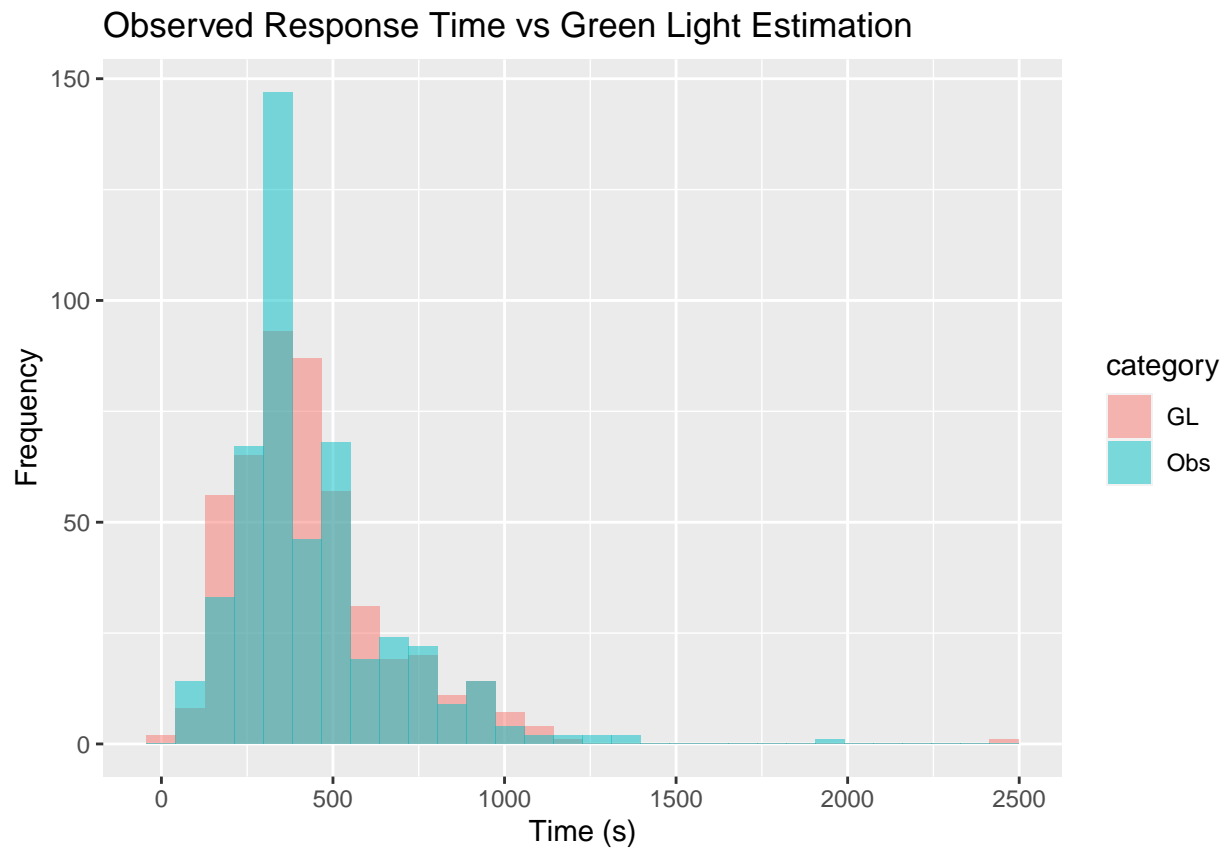
Comparison with Estimated Response Time

Visualization

```
# for each record, extract Google map API estimation based on observed base and destination
x$eTT.GL <- ifelse(x$BASE.NAME == "Company 1", x$eTT.GL.So, x$eTT.GL.Ce)
x$eTT.Pe <- ifelse(x$BASE.NAME == "Company 1", x$eTT.Pe.So, x$eTT.Pe.Ce)
x$eTT.BG <- ifelse(x$BASE.NAME == "Company 1", x$eTT.BG.So, x$eTT.BG.Ce)
x$eTT.Op <- ifelse(x$BASE.NAME == "Company 1", x$eTT.Op.So, x$eTT.Op.Ce)

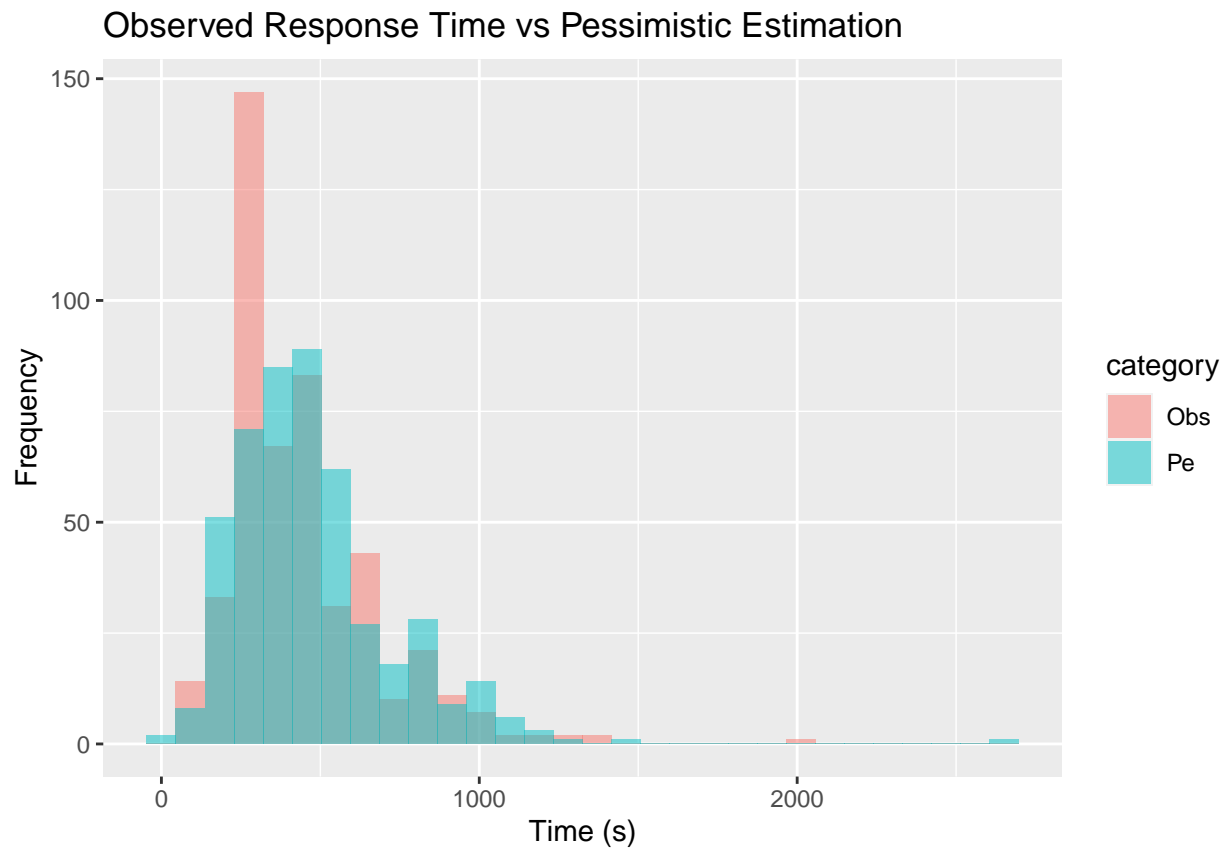
# compare observed with green light
df1 <- data.frame(category = rep(c("Obs", "GL"), each = length(x$observedTT_numeric)),
                  value = c(x$observedTT_numeric, x$eTT.GL))

ggplot(df1, aes(x = value, fill = category)) +
  geom_histogram(position = "identity", alpha = 0.5) +
  labs(title = "Observed Response Time vs Green Light Estimation",
       x = "Time (s)",
       y = "Frequency")
```



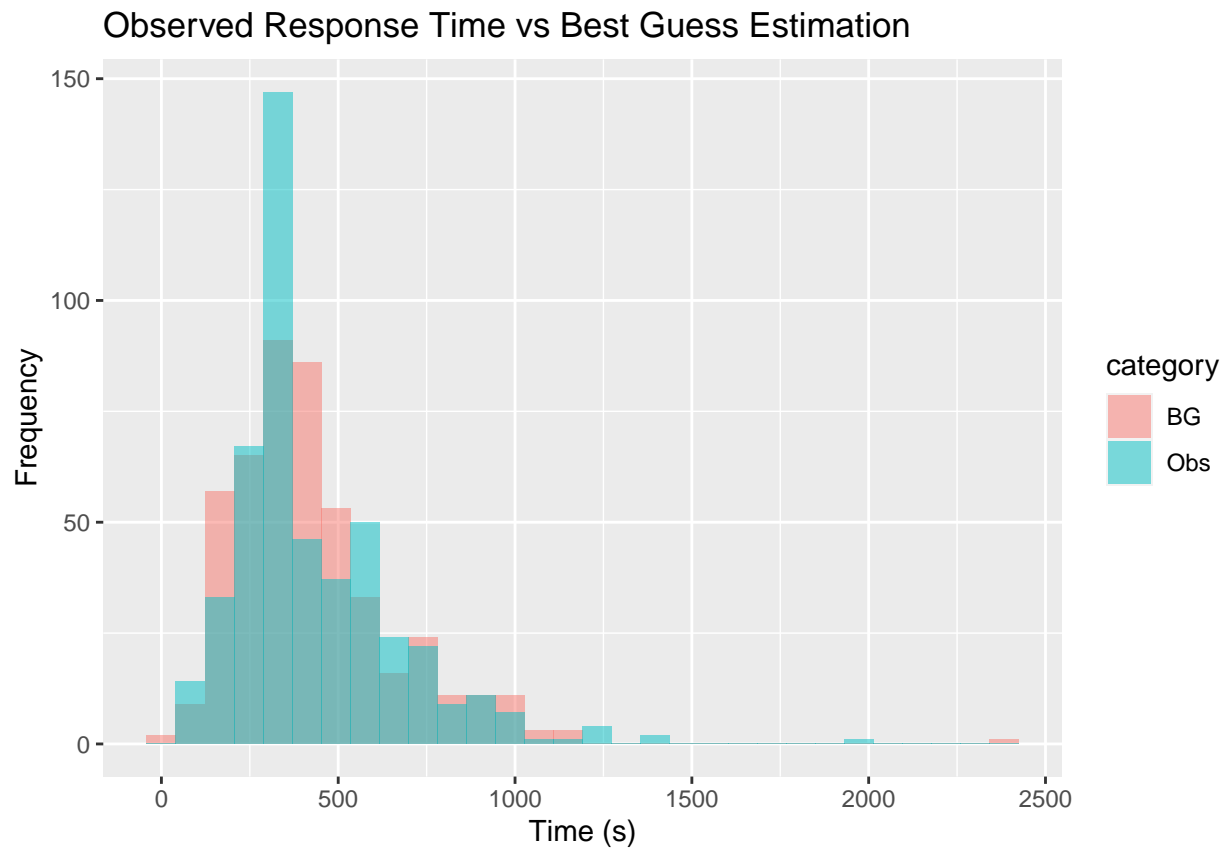
```
# compare observed with pessimistic
df2 <- data.frame(category = rep(c("Obs", "Pe"), each = length(x$observedTT_numeric)),
                  value = c(x$observedTT_numeric, x$eTT.Pe))

ggplot(df2, aes(x = value, fill = category)) +
  geom_histogram(position = "identity", alpha = 0.5) +
  labs(title = "Observed Response Time vs Pessimistic Estimation",
       x = "Time (s)",
       y = "Frequency")
```



```
# compare observed with best guess
df3 <- data.frame(category = rep(c("Obs", "BG"), each = length(x$observedTT_numeric)),
                  value = c(x$observedTT_numeric, x$eTT.BG))

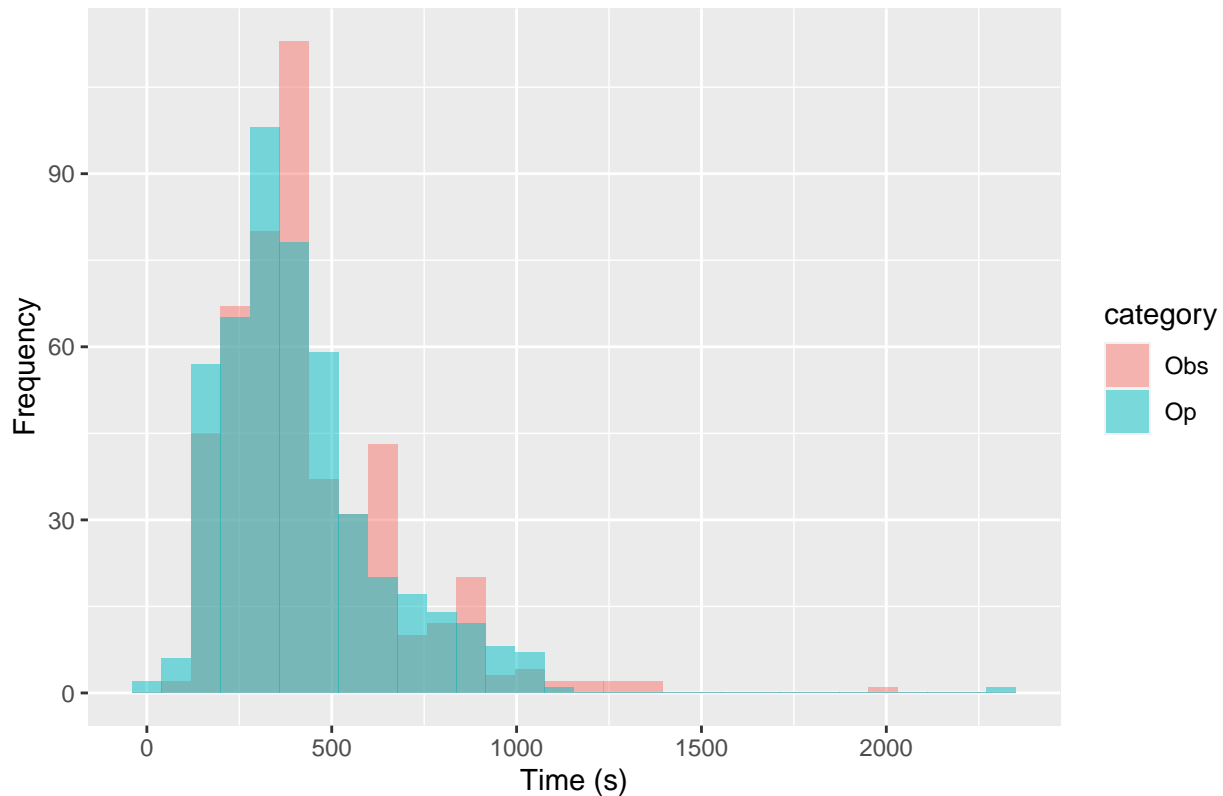
ggplot(df3, aes(x = value, fill = category)) +
  geom_histogram(position = "identity", alpha = 0.5) +
  labs(title = "Observed Response Time vs Best Guess Estimation",
       x = "Time (s)",
       y = "Frequency")
```



```
# compare observed with optimistic
df4 <- data.frame(category = rep(c("Obs", "Op"), each = length(x$observedTT_numeric)),
                  value = c(x$observedTT_numeric, x$eTT.Op))

ggplot(df4, aes(x = value, fill = category)) +
  geom_histogram(position = "identity", alpha = 0.5) +
  labs(title = "Observed Response Time vs Optimistic Estimation",
       x = "Time (s)",
       y = "Frequency")
```

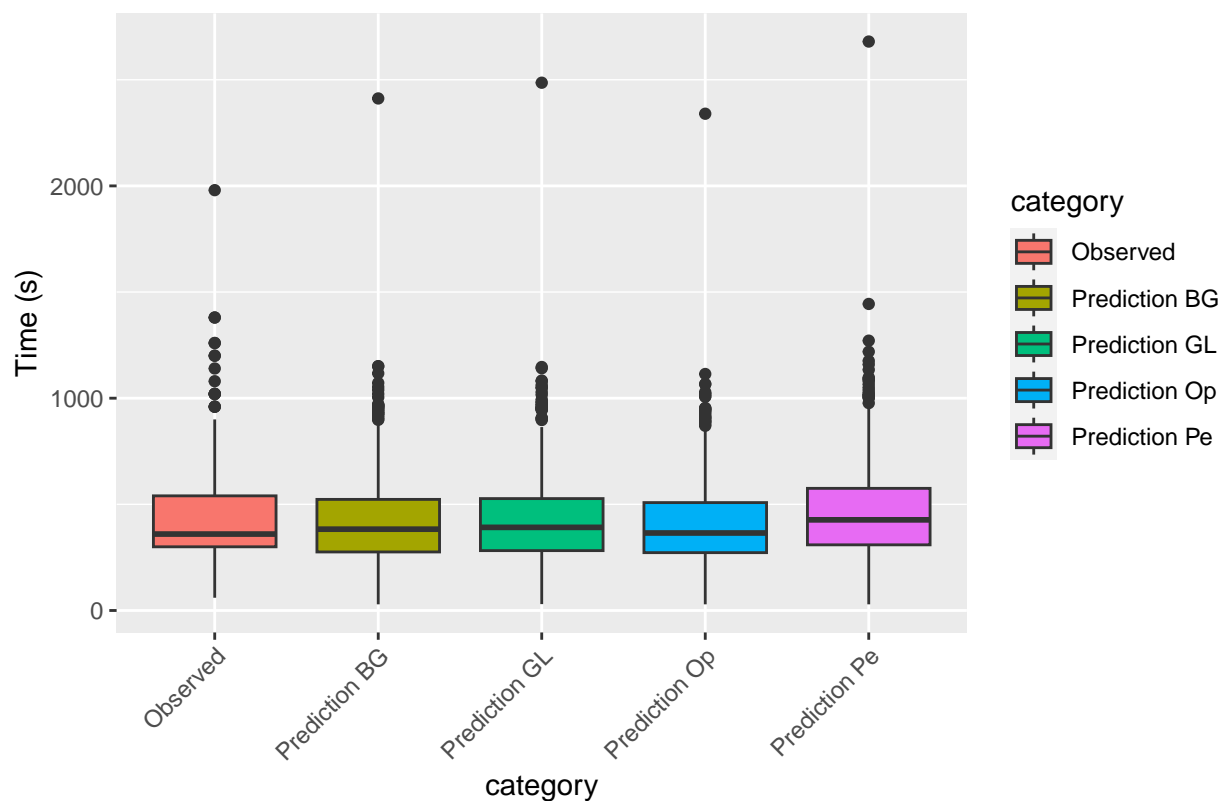
Observed Response Time vs Optimistic Estimation



```
# boxplot to compare five response time distributions
observed_df <- data.frame(category = "Observed", value = x$observedTT_numeric)
pred1_df <- data.frame(category = "Prediction GL", value = x$eTT.GL)
pred2_df <- data.frame(category = "Prediction Pe", value = x$eTT.Pe)
pred3_df <- data.frame(category = "Prediction BG", value = x$eTT.BG)
pred4_df <- data.frame(category = "Prediction Op", value = x$eTT.Op)
combined_df <- bind_rows(observed_df, pred1_df, pred2_df, pred3_df, pred4_df)

ggplot(combined_df, aes(x = category, y = value, fill = category)) +
  geom_boxplot() +
  labs(title = "Observed Response Time vs Different Estimations",
       y = "Time (s)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Observed Response Time vs Different Estimations



RMSE

```
rmse1 <- sqrt(mean((observed_df$value - pred1_df$value)^2))
rmse2 <- sqrt(mean((observed_df$value - pred2_df$value)^2))
rmse3 <- sqrt(mean((observed_df$value - pred3_df$value)^2))
rmse4 <- sqrt(mean((observed_df$value - pred4_df$value)^2))
```

```
print(rmse1)
```

```
## [1] 189.8238
```

```
print(rmse2)
```

```
## [1] 212.1679
```

```
print(rmse3)
```

```
## [1] 188.9788
```

```
print(rmse4)
```

```
## [1] 183.6995
```

The optimistic Google Map API has the best estimation among the four, based on the RMSE criteria.

Linear Regression (Obs vs Op)

```
# perform the linear regression
lm_model <- lm(observed_df$value ~ pred4_df$value)
summary(lm_model)

##
## Call:
## lm(formula = observed_df$value ~ pred4_df$value)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1369.36   -86.97   -20.68    56.54   1440.51
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  136.99639   16.41909   8.344 7.92e-16 ***
## pred4_df$value    0.70614    0.03454  20.443 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170.9 on 474 degrees of freedom
## Multiple R-squared:  0.4686, Adjusted R-squared:  0.4674
## F-statistic: 417.9 on 1 and 474 DF,  p-value: < 2.2e-16

# scatter plot
ggplot(data = NULL, aes(x = pred4_df$value, y = observed_df$value)) +
  geom_point() +
  geom_abline(intercept = coef(lm_model)[1],
              slope = coef(lm_model)[2],
              color = "red") +
  labs(title = "Observed Response Time vs Optimistic Estimation",
       x = "Optimistic Estimation",
       y = "Observed Response Time")
```

