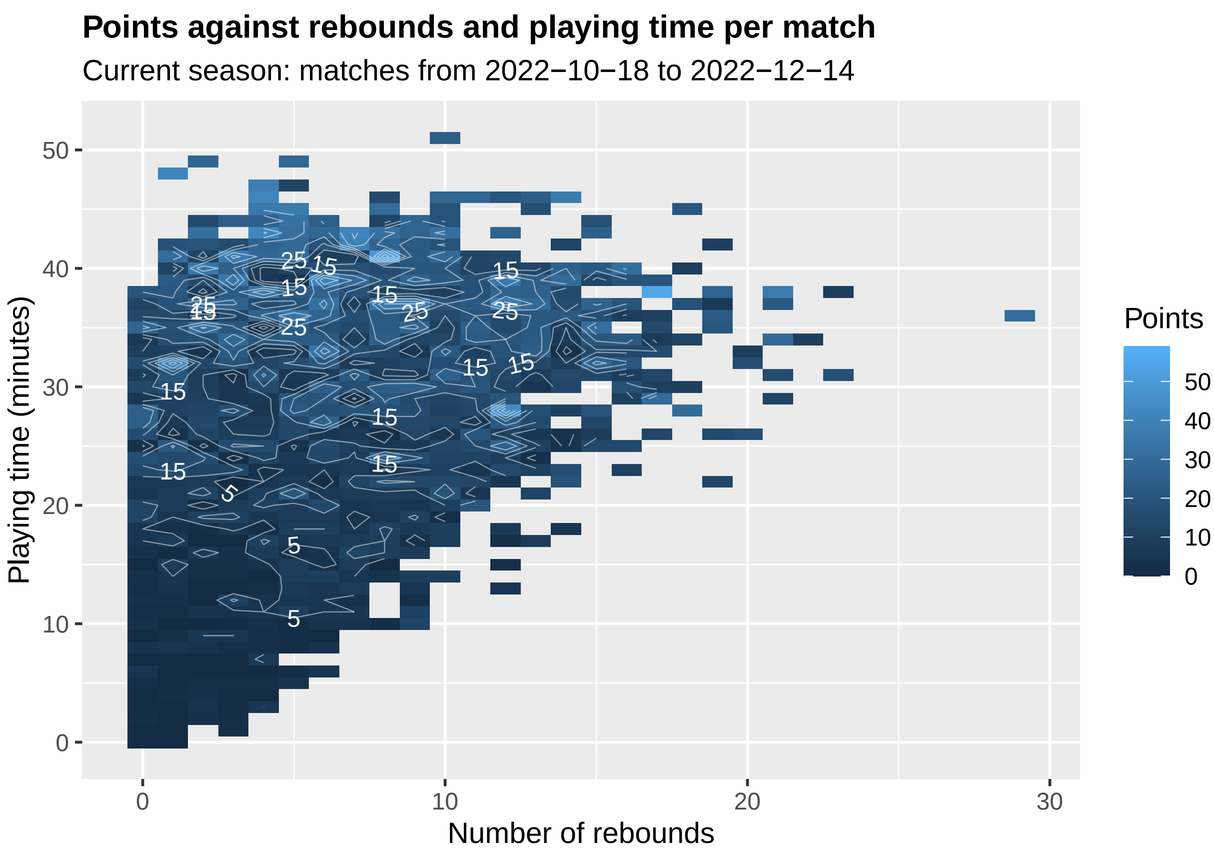
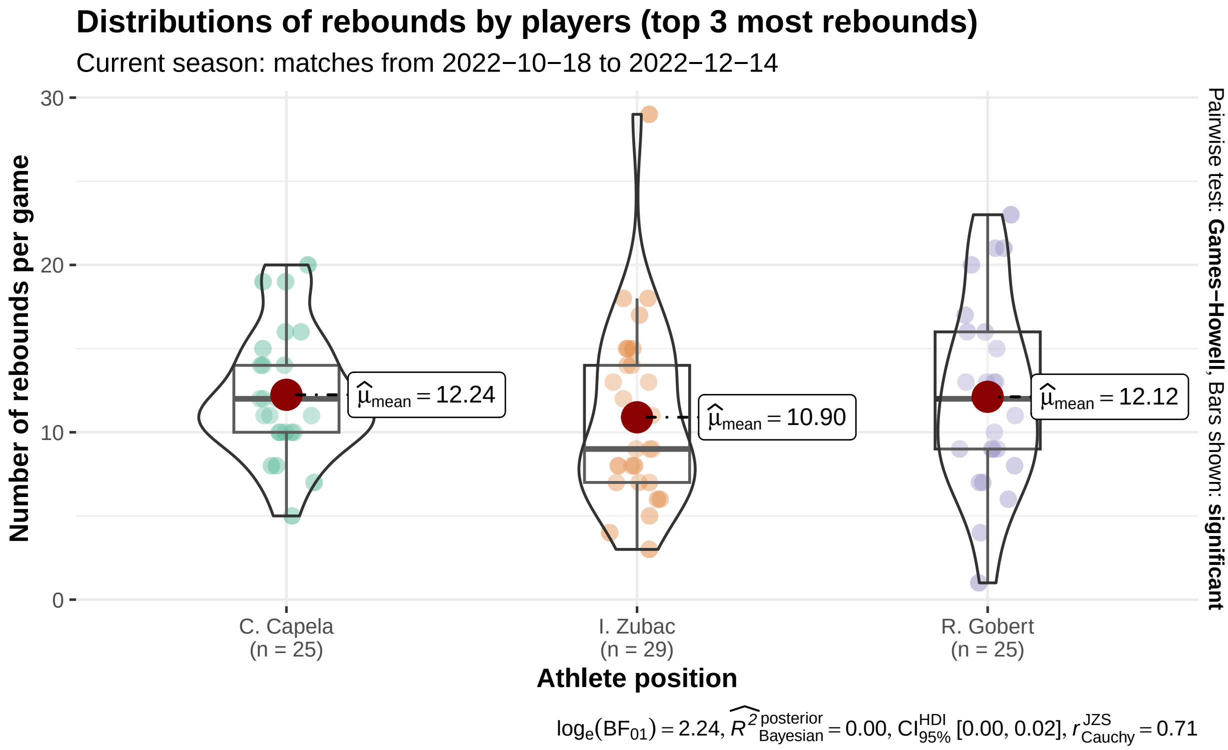
STAT 442: Final Project

The data used in the final project is the current season NBA player boxscore data consisting of matches from 2022-10-18 to 2022-12-14 imported from the `hoopr` package. The code used to generate the plots are attached at the end of the report.

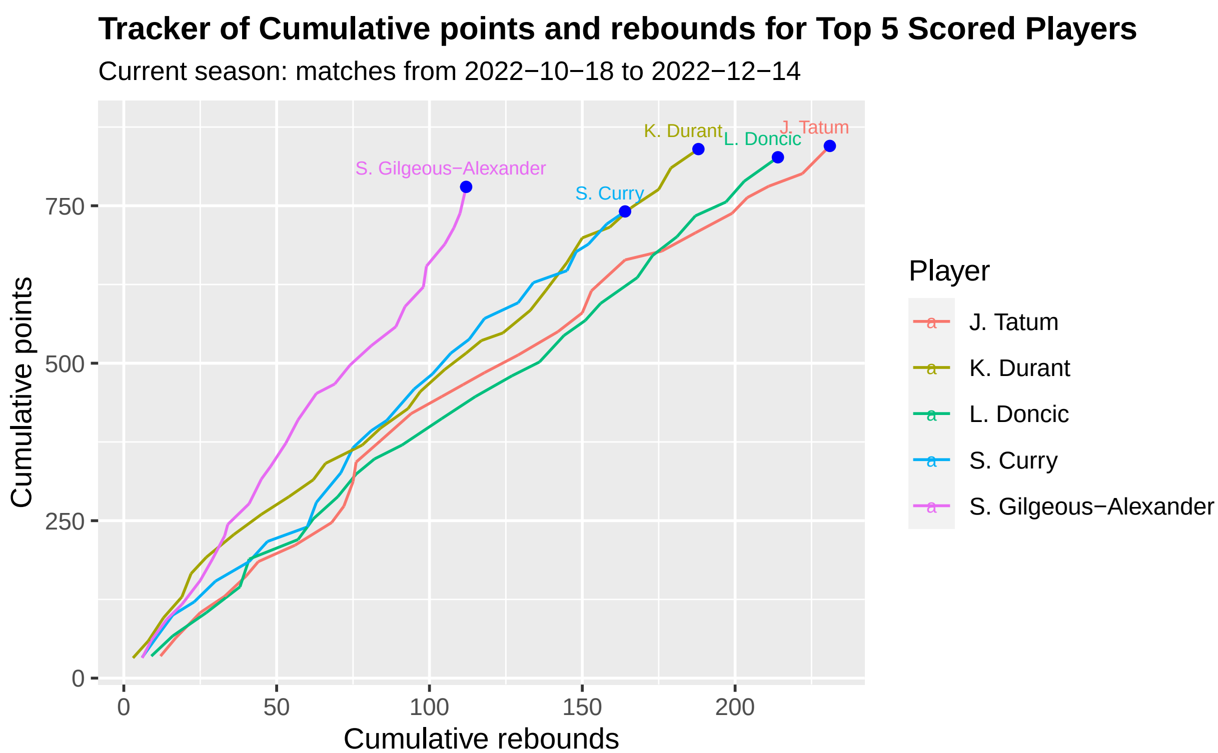
The first plot below shows the relationship between the points scored and number of rebounds and playing time per match.



The second plot below shows the distribution of the number of rebounds for players. The top three players with most number of rebounds are shown.



The third plot below tracks the total points and rebounds for top players up to date in the current season.



The last table shows the field goal (FG) percentage leader board for the current season for players who appeared in more than 10 games.

Table 1: Top 20 field goal percentage among players with more than 10 games played (Current season: matches from 2022-10-18 to 2022-12-14)

Name	Position	Team	Games Played	FGM	FGA	FG%
D. Jordan	C	Nuggets	21	46	60	76.7%
J. Sims	C	Knicks	19	35	46	76.1%
D. Gafford	C	Wizards	28	68	91	74.7%
N. Claxton	C	Nets	26	137	187	73.3%
D. Eubanks	PF	Trail Blazers	25	60	82	73.2%
W. Kessler	C	Jazz	27	73	100	73%
M. Robinson	C	Knicks	19	62	85	72.9%
C. Bassey	C	Spurs	17	43	61	70.5%
L. Kornet	C	Celtics	24	47	67	70.1%
T. Bryant	C	Lakers	13	41	59	69.5%
D. Powell	C	Mavericks	24	54	78	69.2%
R. Holmes	C	Kings	14	20	29	69%
K. Looney	C	Warriors	28	76	115	66.1%
R. Lopez	C	Cavaliers	19	27	41	65.9%
R. Gobert	C	Timberwolves	25	134	205	65.4%
M. Brown	C	Clippers	21	32	49	65.3%
L. Nance Jr.	PF	Pelicans	25	99	153	64.7%
J. Poeltl	C	Spurs	20	114	178	64%
K. Birch	C	Raptors	15	16	25	64%
B. Clarke	PF	Grizzlies	27	111	174	63.8%

Code:

```r
# import library
library(hoopR)
library(tidyr)
library(dplyr)
library(ggplot2)
library(ggstatsplot)
library(metR)
# import NBA player statistics for the most recent season
df <- load_nba_player_box(seasons = most_recent_nba_season())
# convert char to numeric for continuous variables in the data
df$pts <- as.numeric(df$pts)
df$reb <- as.numeric(df$reb)
df$oreb <- as.numeric(df$oreb)
df$min <- as.numeric(df$min)
```

```r
# Visualization 1 (2D): Contour plot for points against ofensive rebounds and playing time
ggplot(df, aes(x = reb, y = min)) +
    geom_raster(aes(fill = pts)) +
    geom_contour(aes(z = pts), colour = "white", size = 0.2, alpha = 0.5) +
    geom_text_contour(aes(z = pts),  colour = "white", size = 3) +
    labs(x = "Number of rebounds", y = "Playing time (minutes)",
         fill = "Points",
         title = "Points against rebounds and playing time per match",
         subtitle = "Current season: matches from 2022-10-18 to 2022-12-14") +
    theme(plot.title = element_text(size = 12, face = "bold"))
```

```r
# Visualization 2 (categorical): violin plot of distribution of rebounds by top athlete
# with the most rebounds.
# Order athlete by number of rebounds
pos.most.reb <- df %>%
    group_by(athlete_id) %>%
    summarise(n=sum(reb)) %>%
    arrange(desc(n))
# filter data to players with most rebounds
df.reb <- df %>%
    filter(athlete_id %in% pos.most.reb$athlete_id[1:3])
# create violin plot
plt <- ggbetweenstats(data = df.reb, x = athlete_short_name, y = reb) +
    labs(x = "Athlete position", y = "Number of rebounds per game",
        title = "Distributions of rebounds by players (top 3 most rebounds)",
        subtitle = "Current season: matches from 2022-10-18 to 2022-12-14")
plt
```

```r
# Visualization 3 (homebrew): tracking cumulative points/rebounds for top players
# find players with most points
player.most.pts <- df %>%
    group_by(athlete_id) %>%
    summarise(n = sum(pts)) %>%
    arrange(desc(n))
# filter top 5 players and extract their points/rebounds for each match
play.pts <- df %>%
```

```
    filter(athlete_id %in% player.most.pts$athlete_id[1:5]) %>%
    arrange(game_date) %>%
    select(athlete_short_name, athlete_id, reb, pts, game_date) %>%
    group_by(athlete_id) %>%
    mutate(rank = order(game_date), cum.pts = cumsum(pts), cum.reb = cumsum(reb))
# prepare data for line segment by left join game data to itself
df.combined <- play.pts %>%
    mutate(rank = rank + 1) %>%
    left_join(select(play.pts, c(athlete_id, cum.reb, cum.pts, rank)),
              by = c('athlete_id', 'rank'))
# extract line segment and last match
df.segment <- df.combined %>%
    filter(!is.na(cum.pts.y))
df.last <- df.combined %>%
    filter(is.na(cum.pts.y))
# create plot
ggplot(df.segment) +
    aes(x = cum.reb.x, y = cum.pts.x, xend = cum.reb.y, yend = cum.pts.y,
        group = athlete_short_name, colour = athlete_short_name) +
    geom_segment(size = .5) +
    geom_point(data = df.last, aes(x = cum.reb.x, y = cum.pts.x), col = 'blue') +
    geom_text(data = df.last, aes(x = cum.reb.x, y = cum.pts.x, label = athlete_short_name),
              size = 2.5, nudge_y = 30, nudge_x = -5) +
    labs(title = "Tracker of Cumulative points and rebounds for Top 5 Scored Players",
         subtitle = "Current season: matches from 2022-10-18 to 2022-12-14",
         x = "Cumulative rebounds", y = "Cumulative points",
         colour = "Player") +
    theme(plot.title = element_text(size = 12, face = "bold"),
          plot.subtitle = element_text(size = 10))
```

```
# Visualization 4 (wild card): Field Goal Percentage Leaders
# compute summary statistics: average pts, average rebounds, accuracy
df.tbl <- df %>%
    separate(fg, c('fg.a','fg.t')) %>%
    mutate(fg.a = as.numeric(fg.a), fg.t = as.numeric(fg.t)) %>%
    group_by(athlete_id) %>%
    summarise(pts = sum(pts), reb = sum(reb), n = n_distinct(game_id),
              team = min(team_name), name = min(athlete_short_name),
              pos = min(athlete_position_abbreviation),
              fg.a = sum(fg.a), fg.t = sum(fg.t)) %>%
    mutate(avg.pts = round(pts / n,1),
           avg.reb = round(reb / n, 1),
           acc1 = fg.a / fg.t,
           acc = paste(round(fg.a / fg.t,3)*100, '%', sep = '')) %>%
    arrange(desc(acc1)) %>%
    filter(n > 10) %>%
    select(name, pos, team, n, fg.a, fg.t, acc)
# display table
knitr::kable(df.tbl[1:20,], col.names =  c("Name", "Position", "Team", "Games Played",
                                   "FGM", "FGA","FG%"),
             caption = "Top 20 field goal percentage among players with more than 10 games played
             (Current season: matches from 2022-10-18 to 2022-12-14)")
```