Data Science Final Project

Theodor Clark & John Ferguson

12/12/2021

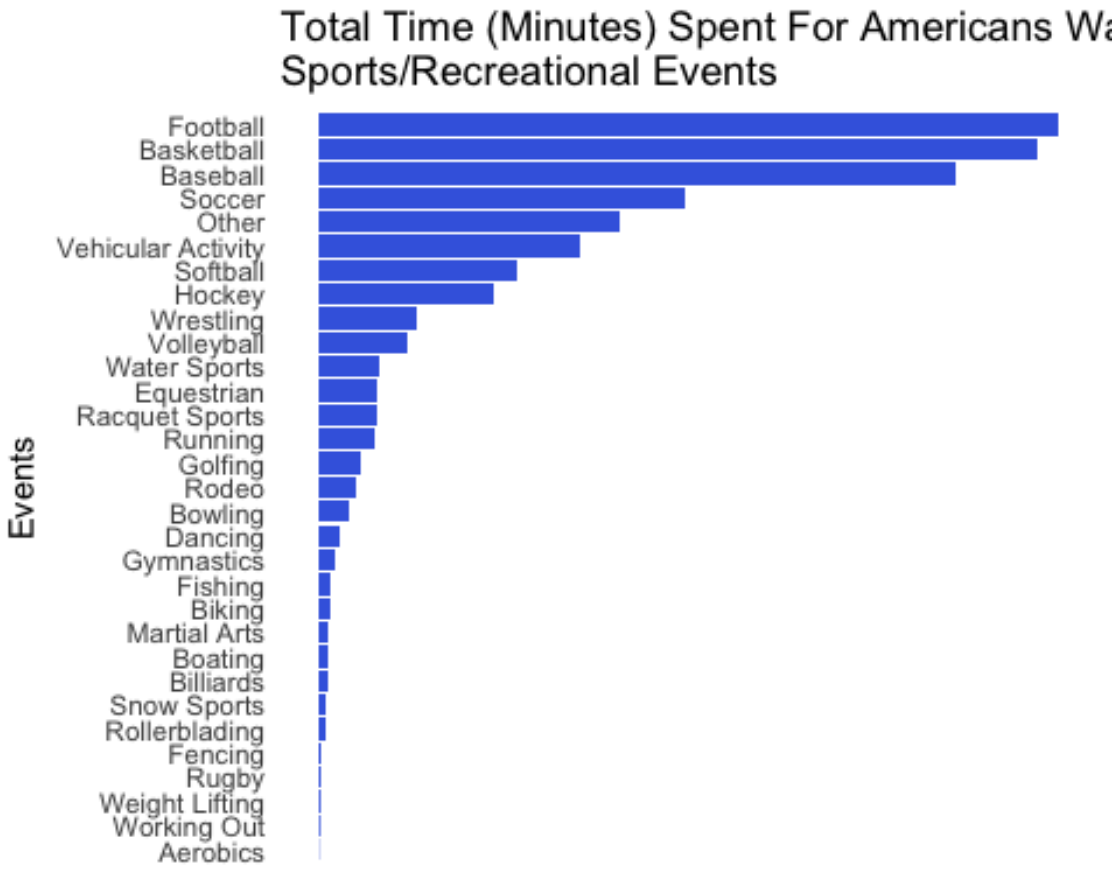
```
# Reading in the activity and roster data sets from 2003-20 ATUS
ATUS_data <- read.csv("/Users/theoclark/Desktop/Fall 2021/atusact-0320/atusact_0320.dat")
ATUS_roster <- read.csv("/Users/theoclark/Desktop/Fall 2021/atusrost-0320/atusrost_0320.dat")

# Loading required packages
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

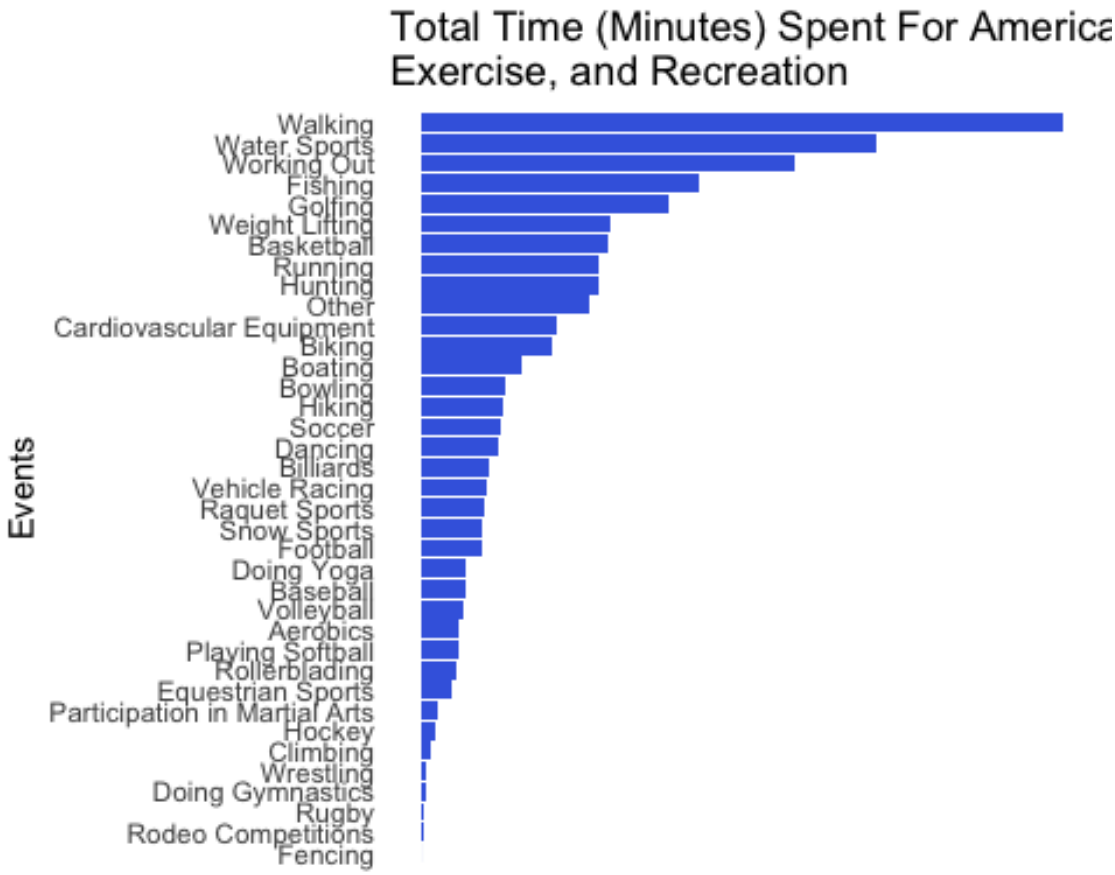
```
library(ggplot2)

# Removing all undesired variables from the activity dataset
ATUS_data <- ATUS_data %>%
  select(TRCODEP, TUACTDUR, TUCASEID)

ATUS <- ATUS_roster %>%
  left_join(ATUS_data, by="TUCASEID")
# Filtering the data to only include the time use for sports watching and participating
sports_data <- filter(ATUS, TRCODEP >= 130201 & TRCODEP <= 130299)
sports_play <- filter(ATUS, TRCODEP >= 130101 & TRCODEP <= 130199)

Events <- c("Aerobics", "Baseball", "Basketball","Biking", "Billiards", "Boating", "Bowling", "Dancing" ,"Equestrian", "Fencing", "Fishing", "Football" ," Golfing", "Gymnastics", "Hockey", "Martial Arts", "Racquet Sports", "Rodeo" , "Rollerblading", "Rugby","Running", "Snow Sports", "Soccer", "Softball"," Vehicular Activity","Volleyball","Water Sports","Weight Lifting","Working Out","Wrestling","Other")
```

```r
Events1 <- c("Aerobics", "Baseball", "Basketball","Biking", "Billiards", "Boa
ting", "Bowling", "Climbing", "Dancing" ,"Equestrian Sports", "Fencing", "Fis
hing", "Football" ,"Golfing", "Doing Gymnastics","Hiking", "Hockey", "Hunting
", "Participation in Martial Arts", "Raquet Sports" ,"Rodeo Competitions" ,"R
ollerblading", "Rugby","Running", "Snow Sports", "Soccer", "Playing Softball"
,"Cardiovascular Equipment", "Vehicle Racing","Volleyball","Walking", "Water
Sports","Weight Lifting","Working Out","Wrestling","Doing Yoga", "Other")

# Doing the proper wrangling in order to find the top sports for both watchin
g and participation
watch_total <- sports_data %>%
  group_by(TRCODEP) %>%
  summarize(
    Total_Time_Watched = sum(TUACTDUR)
  )
watch_total$Sporting_Events = Events
watch_total %>%
  select(-TRCODEP) %>%
  arrange(desc(Total_Time_Watched))

## # A tibble: 31 × 2
##    Total_Time_Watched Sporting_Events
##                 <int> <chr>
##  1             266041 "Football"
##  2             259004 "Basketball"
##  3             229585 "Baseball"
##  4             132161 "Soccer"
##  5             108740 "Other"
##  6              94184 " Vehicular Activity"
##  7              71711 "Softball"
##  8              63138 "Hockey"
##  9              35311 "Wrestling"
## 10              32095 "Volleyball"
## # … with 21 more rows

play_total <- sports_play %>%
  group_by(TRCODEP) %>%
  summarize(
    Total_Participation_Time = sum(TUACTDUR)
  )
play_total$Sporting_Events = Events1
play_total %>%
  select(-TRCODEP) %>%
  arrange(desc(Total_Participation_Time))

## # A tibble: 37 × 2
##    Total_Participation_Time Sporting_Events
##                       <int> <chr>
##  1                  1705422 Walking
##  2                  1206692 Water Sports
```

```
##  3                   990135 Working Out
##  4                   735643 Fishing
##  5                   655248 Golfing
##  6                   499979 Weight Lifting
##  7                   492745 Basketball
##  8                   466708 Running
##  9                   466670 Hunting
## 10                   447322 Other
## # … with 27 more rows

# Creating two plots for participation and watching sports time spent for eac
h
ggplot(data = watch_total, mapping = aes(x = factor(reorder(Events,Total_Time
_Watched)), y = Total_Time_Watched)) +
  geom_bar(stat = 'identity', fill = "#4169E1") +
  labs(title = "Total Time (Minutes) Spent For Americans Watching \nSports/Re
creational Events",
       x = "Events", y=NULL) +
   theme_bw() +
  theme(plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
  axis.ticks = element_blank(),
  axis.text.x = element_blank()) +
        coord_flip()
```




```
ggplot(data = play_total, mapping = aes(x = factor(reorder(Events1,Total_Part
icipation_Time)), y = Total_Participation_Time)) +
  geom_bar(stat = 'identity', fill = "#4169E1") +
  labs(x = "Events", title="Total Time (Minutes) Spent For Americans Particip
ating in Sports, \nExercise, and Recreation", y=NULL) +
   theme_bw() +
  theme(plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
  axis.ticks = element_blank(),
  axis.text.x = element_blank()) +
        coord_flip()
```




```r
# Creating subsets each of the top 3 from each graph into datasets
Walking <- ATUS %>%
  group_by(TRCODEP) %>%
  filter(TRCODEP == "130131")

Working_out <- ATUS %>%
  group_by(TRCODEP) %>%
  filter(TRCODEP == "130134")

Water_sports <- ATUS %>%
  group_by(TRCODEP) %>%
  filter(TRCODEP == "130132")

Football <- ATUS %>%
  group_by(TRCODEP) %>%
  filter(TRCODEP == "130213")

Baseball <- ATUS %>%
  group_by(TRCODEP) %>%
  filter(TRCODEP == "130202")

Basketball <- ATUS %>%
  group_by(TRCODEP) %>%
```

```
  filter(TRCODEP == "130203")

# Creating a matrix to save the summary stats of each 6 activitiesto display
options(digits=4)
matrix_a = data.frame(matrix(ncol = 3, nrow = 4))
colnames(matrix_a)=c("Watching Football", "Watching Baseball", "Watching Bask
etball")
rownames(matrix_a)=c("Minimum", "Maximum", "Mean", "Standard Deviation")


  matrix_a[1,1] = min(Football$TUACTDUR)
  matrix_a[1,2] = min(Baseball$TUACTDUR)
  matrix_a[1,3] = min(Basketball$TUACTDUR)
  matrix_a[2,1] = max(Football$TUACTDUR)
  matrix_a[2,2] = max(Baseball$TUACTDUR)
  matrix_a[2,3] = max(Basketball$TUACTDUR)
  matrix_a[3,1] = mean(Football$TUACTDUR)
  matrix_a[3,2] = mean(Baseball$TUACTDUR)
  matrix_a[3,3] = mean(Basketball$TUACTDUR)
  matrix_a[4,1] = sd(Football$TUACTDUR)
  matrix_a[4,2] = sd(Baseball$TUACTDUR)
  matrix_a[4,3] = sd(Basketball$TUACTDUR)

matrix_a

##                    Watching Football Watching Baseball Watching Basketball
## Minimum                         5.00              5.00                1.00
## Maximum                       495.00            565.00              530.00
## Mean                          163.12            158.77              134.69
## Standard Deviation             84.95             90.37               83.35

matrix_b = data.frame(matrix(ncol = 3, nrow = 4))
colnames(matrix_b)=c("Walking", "Working out", "Participating in Water Sports
")
rownames(matrix_b)=c("Minimum", "Maximum", "Mean", "Standard Deviation")


  matrix_b[1,1] = min(Walking$TUACTDUR)
  matrix_b[1,2] = min(Working_out$TUACTDUR)
  matrix_b[1,3] = min(Water_sports$TUACTDUR)
  matrix_b[2,1] = max(Walking$TUACTDUR)
  matrix_b[2,2] = max(Working_out$TUACTDUR)
  matrix_b[2,3] = max(Water_sports$TUACTDUR)
  matrix_b[3,1] = mean(Walking$TUACTDUR)
  matrix_b[3,2] = mean(Working_out$TUACTDUR)
  matrix_b[3,3] = mean(Water_sports$TUACTDUR)
  matrix_b[4,1] = sd(Walking$TUACTDUR)
  matrix_b[4,2] = sd(Working_out$TUACTDUR)
  matrix_b[4,3] = sd(Water_sports$TUACTDUR)
```

```
matrix_b
```

```
##                    Walking Working out Participating in Water Sports
## Minimum               1.00        1.00                           1.00
## Maximum             840.00     1400.00                         777.00
## Mean                 49.75       51.16                         103.27
## Standard Deviation   38.27       37.80                          76.95
```

```
# Linear model for sports watching time using duration as the response variab
le and age and sex as predictor variables
model.lm <- lm(TUACTDUR ~ TEAGE + factor(TESEX), data = sports_data)
summary(model.lm)
```

```
##
## Call:
## lm(formula = TUACTDUR ~ TEAGE + factor(TESEX), data = sports_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -151.1  -69.3  -25.2   45.2  879.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.4002     2.1115   67.91   <2e-16 ***
## TEAGE            0.1709     0.0522    3.28   0.0011 **
## factor(TESEX)2  -0.4840     2.0408   -0.24   0.8125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101 on 9861 degrees of freedom
## Multiple R-squared:  0.00109,    Adjusted R-squared:  0.000888
## F-statistic: 5.38 on 2 and 9861 DF,  p-value: 0.00462
```

```
# Linear model for sports participation time using duration as the response v
ariable and age and sex as predictor variables
modelplay.lm <- lm(TUACTDUR ~ TEAGE + factor(TESEX) , data = sports_play)
summary(modelplay.lm)
```

```
##
## Call:
## lm(formula = TUACTDUR ~ TEAGE + factor(TESEX), data = sports_play)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -86.0  -46.9  -22.1   17.1 1327.4
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    87.99485    0.43697   201.4   <2e-16 ***
## TEAGE          -0.23632    0.00942   -25.1   <2e-16 ***
```

```
## factor(TESEX)2 -4.23453    0.41539   -10.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.7 on 143594 degrees of freedom
## Multiple R-squared:  0.00519,    Adjusted R-squared:  0.00518
## F-statistic:  375 on 2 and 143594 DF,  p-value: <2e-16

cor(sports_data$TEAGE, sports_data$TUACTDUR)

## [1] 0.03293

cor(sports_play$TEAGE, sports_play$TUACTDUR)

## [1] -0.06688
```

Group Info:

```
Group Members: John Ferguson (fivethirtyeight dataset: classic_rock_song_list
) & Theodor Clark (fivethirtyeight dataset: drug_use)

Contributions: We coded most of the project together with about 50/50 contrib
ution, besides the regression model which Jack took most of the responsibilie
s for and the summary statistics table, which Theo took most of the responsib
ilies for. For the analysis, Theo did the introduction along with the analysi
s of the summary statistics table and the regression analysis. Jack then did
the analysis of the graphs for time use, along with the conclusion of the ana
lysis and where we could go from there if we had the proper time and resource
s.
```

Data Analysis Report:

```
    In order to retrieve, explore, and analyze useful data, we used the data
provided by the American Time Use Survey (ATUS) from 2003-2020. ATUS covers a
ll residents living in households in the United States that are at least 15 y
ears of age, with the exception of active military personnel and people resid
ing in institutions such as nursing homes and prisons. The ATUS sample is com
posed of the civilian, noninstitutional population residing in occupied house
holds in the United States. From this sample, the CPS selects approximately 5
9,000 eligible households every month. For goals of analysis, we were interes
ted in which sports were most/least popular for Americans to spend time watch
ing and which were most/least popular to participate in. Additionally, we wan
ted to see the trends in demographics (with a focus on sex and age) relating
to the amount of time dedicated to watching sports and participating in sport
s. In order to have demographics and activity time use in the same table, we
joined the 2003-2020 roster file with the activity file. This was able to giv
e us demographics along with activity durations for sports watching and parti
cipation, for further analysis of the time use in these activities.
    One of the important factors in the analysis of the data that we wanted t
o look at was the total amount of time spent participating and watching parti
cular sports or physical activities. We wrangled the data around to create a
```

couple of data sets that could reflect the amounts given to us by the ATUS. From these data sets, we learned that the most popular sports to watch in America are football, basketball, baseball, and soccer in terms of the total time watched by the survey participants. For participating in sports and activities the most popular were walking, water sports, working out, and fishing.

There seems to be an interesting distinction between what the survey participants enjoy watching, compared to what they enjoy to participate in. Very intense and physically demanding sports such as football are the most popular sports to consume as a viewer. This contrasts with the most participated in activities which tend to be much less physically demanding such as walking and fishing. Another significant difference between the two data sets, is that the highest total watch times are dominated by team sports, which is contrary to the individual sports and activities that dominate the highest participation total times.

After finding the most popular sports in America for both time spent watching and participating in, we show a summary statistic table for the top three sports from each. The mean for spending time watching sports was higher in each of the top three sports than time spent participating in each of the top three sports. The maximum values, however, report much higher values in the sports participation time than in the time spent watching sports. These mean and maximum results likely relate to the controlled time of a sporting event one watches, as this time is relatively constant across an individual sport. Conversely, individuals can control the amount of time they spend participating in their own form of sport or recreational activity.

Finally, we created a linear regression model using age and sex (as a numerical variable instead of categorical) to predict the activity duration for time spent watching sports and time spent participating in sports. We found, that each variable was significant in predicting durations, with the exception of using the female sex to predict time spend watching sports. This signifies a trend that male Americans tend to spend more time watching sports than females. Additionally, we reviewed the correlation between time age and activity duration for spent watching sports and time spent playing sports. There was nearly zero correlation found here, which signifies as one gets older (regardless of gender), there is no increase or decrease in the amount of time spent watching or participating in sports. These results suggest that as one gets older and stops participating in the more physically demanding sports, they fill the time spent with less physically demanding sports/activities, for example walking or golfing. For watching sports, it makes sense logically that age has no impact on the amount of time spent watching or attending sporting events, as there is no physical factor involved.

Through our analysis and findings, we discovered some rather interesting information ranging from the most popular sports to watch and participate in, to a regression that showed that there was no real correlation between age and participation. Although we are happy with our findings, we do have a couple areas that we think we could further into analysis with some extended time. One aspect that stood out to us when looking at the data, is that team sports are much more highly watched than individual sports. In the future we could look into a demographic difference perhaps, or just trying to determine why that is the case. From our data, we now understand the tendencies of Americans and their sport consumption through various means. From winter sports, to foo

tball, and to even spelunking, Americans like to participate and watch a vari
ety of different physical activities.