STAT230 Project Starting Point

Group D: NAMES

Date

[DELETE & DON'T INCLUDE the following instructions in your proposal]

This template is provided to help you get started in finding the data you want to use for your projects and to develop a proposal.

- Initial proposal: The proposal part should be no more than 3 pages (in the knitted PDF), including at least the following:
  - Work Team , all members' names, and pre-Project Title (if there is one).
  - What is the question of interest? WHY is this question interesting or important to your team?
  - Which dataset (pick one) & variables are you planning to use? Which one is your response variable?

I would recommend each team to choose at least 6 variables from the ACS data, but no more than 10 variables (including the response variable) for the project.

- Perform an EDA (including graphs, numerical summaries, AND verbal descriptions) on the related variables that will be used in your modeling. Use appropriate graphical displays to justify that MLR is a reasonable choice for modeling (i.e. the CHOOSE step in 4-step modeling). You may find the document **Basic R you may Remember** under *Week 1 Tile* in Moodle particularly helpful. Additional Notes -

This document should also include all the code used to *wrangle* data, as well as those used to create the *EDA* output (plots, tables, numerical summaries, etc). Recall that you've learned various `R` functions for data wrangling back in **R Activity 2** (Part 3), as well as from **R Tutorial 2** (under *Useful Resources*). Don't hesitate to ask for help from me or the TA on this front.

Proposal

Import Data & Wrangling

On RStudio, be sure to save the dataset file, `2019ACS_pums.csv` or `2019ACS_housing.csv` in *the SAME folder* you saved this RMD file. I would suggest that you create a specific folder on RStudio for the Project. When you wrangle your data, it's always safer to save the mutated/filtered dataset with a new dataset name.

```
#getwd() in console to find the directory where the csv file is stored on your computer
#For your computer you need to change the path so you can reed the file
#alldata <- read.csv("/home/class26/FIRSTPARTOFEMAIL/ConcussionInjuries2012-2014.csv")

#For Donna's computer
```

```
alldata <- read.csv("C:/Users/donna/OneDrive/Documents/STAT230/Stats230FinalProject/ConcussionInjuries20

#For Kaitlyn's computer
#alldata <- read.csv("/home/class27/khuang27/STAT230FinalProject/ConcussionInjuries2012-2014.csv")

glimpse(alldata)
```

```
## Rows: 392
## Columns: 18
## $ ID                              <chr> "Aldrick Robinson - Washington Redskins~
## $ Player                          <chr> "Aldrick Robinson", "D.J. Fluker", "Mar~
## $ Team                            <chr> "Washington Redskins", "San Diego Charg~
## $ Game                            <chr> "Washington Redskins vs. Tampa Bay Bucc~
## $ Date                            <chr> "30/09/2012", "22/09/2013", "28/09/2014~
## $ Opposing.Team                   <chr> "Tampa Bay Buccaneers", "Tennessee Tita~
## $ Position                        <chr> "Wide Receiver", "Offensive Tackle", "W~
## $ Pre.Season.Injury.              <chr> "No", "No", "No", "No", "Yes", "Yes", "~
## $ Winning.Team.                   <chr> "Yes", "No", "No", "Yes", "Yes", "No", ~
## $ Week.of.Injury                  <int> 4, 3, 4, 6, 1, 1, 7, 9, 1, 1, 1, 1, 1, ~
## $ Season                          <chr> "2012/2013", "2013/2014", "2014/2015", ~
## $ Weeks.Injured                   <int> 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Games.Missed                    <int> 1, 1, 1, 1, NA, NA, NA, NA, NA, NA, NA,~
## $ Unknown.Injury.                 <chr> "No", "No", "No", "No", "No", "No", "Ye~
## $ Reported.Injury.Type            <chr> "Head", "Concussion", "Concussion", "He~
## $ Total.Snaps                     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Play.Time.After.Injury          <chr> "14 downs", "78 downs", "25 downs", "82~
## $ Average.Playtime.Before.Injury <chr> "37.00 downs", "73.50 downs", "17.50 do~
```

```
names(alldata)
```

```
##  [1] "ID"                             "Player"
##  [3] "Team"                           "Game"
##  [5] "Date"                           "Opposing.Team"
##  [7] "Position"                       "Pre.Season.Injury."
##  [9] "Winning.Team."                  "Week.of.Injury"
## [11] "Season"                         "Weeks.Injured"
## [13] "Games.Missed"                   "Unknown.Injury."
## [15] "Reported.Injury.Type"           "Total.Snaps"
## [17] "Play.Time.After.Injury"         "Average.Playtime.Before.Injury"
```

```
acsp <- read.csv("https://pmatheson.people.amherst.edu/stat230/2019ACS_pums.csv")         ##read in per
acsh <- read.csv("https://pmatheson.people.amherst.edu/stat230/2019ACS_housing.csv")      ##read in hou
#dim(acsp)         ##size of the dataset
#names(acsp)      ##variable names
#glimpse(acsp)
```

As you create new subsets of data and/or mutate variables make sure to save them under a new datafile name.

```
## wrangle!!
```

```
studydata <- rename(alldata, OppTeam = Opposing.Team, PostInjuryPlayTime = Play.Time.After.Injury, Pres
#names(studydata)
```

Sample code to filter or create categories with labels

#Take out high values of HINCP (over 3 million) all_data <- filter(all_data, HINCP < 30000)

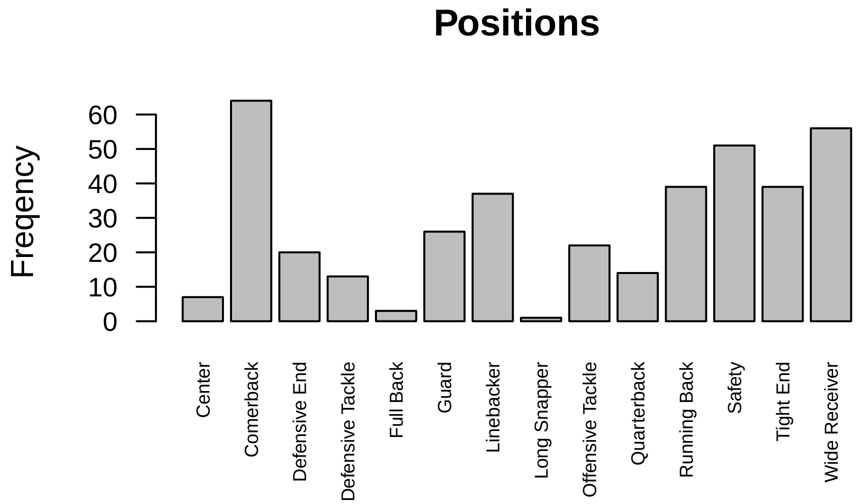
#Mutate SCHL with fewer categories all_data <- mutate(all_data, SCHL = cut(SCHL, breaks = c(0, 15.5, 17.5, 20.5, 21.5, 24.5), labels = c("no hs", "hs", "some college", "BA", "beyond"), include.lowest = TRUE))

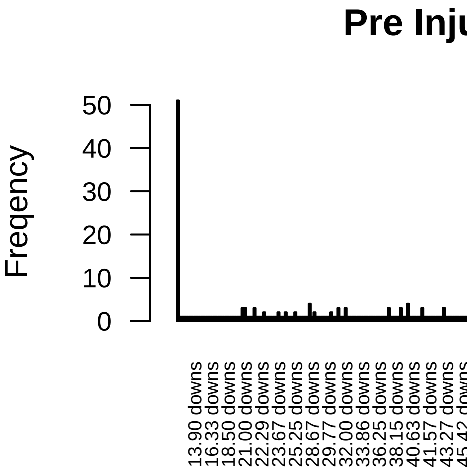
#Mutate JWTRNS with fewer categories all_data <- mutate(all_data, JWTRNS = cut(JWTRNS, breaks = c(0, 1.5, 8.5, 11.5), labels = c("private vehicle", "public", "bike/walk"), include.lowest = TRUE)) tally(~JWTRNS, data = all_data) #only two motorcycles, so we put it in public

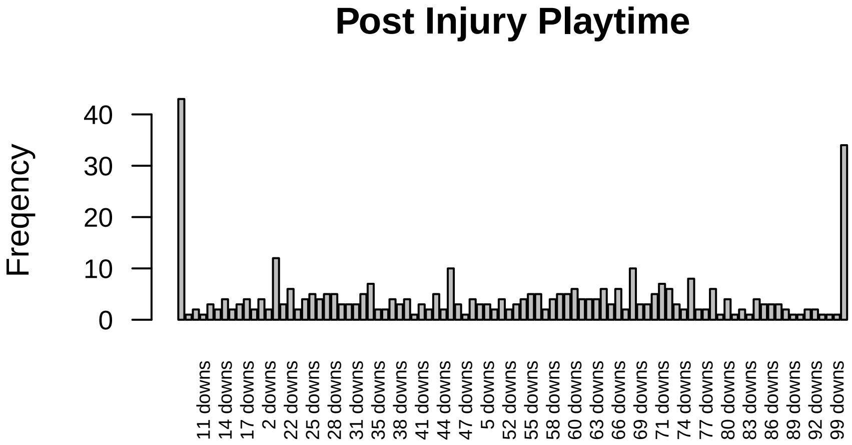
#Mutate HHT fewer categories all_data <- mutate(all_data, HHT = cut(HHT, breaks = c(0, 1.5, 3.5, 7.5), labels = c("married couple", "single parent", "nonfamily"), include.lowest = TRUE)) tally(~HHT, data = all_data)

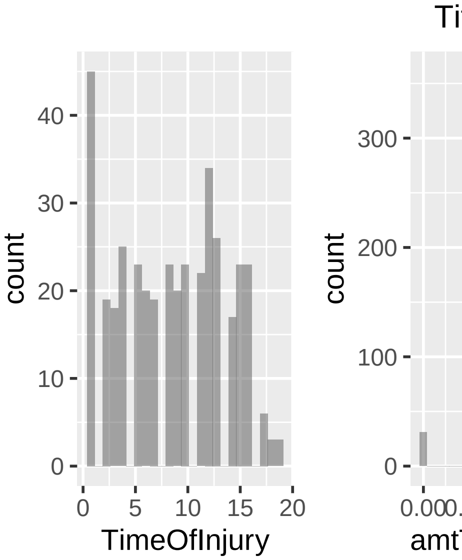
#HHT, JWTRNS, SCHL all_data <- all_data %>% mutate(HHT = as.factor(HHT), JWTRNS = as.factor(JWTRNS))

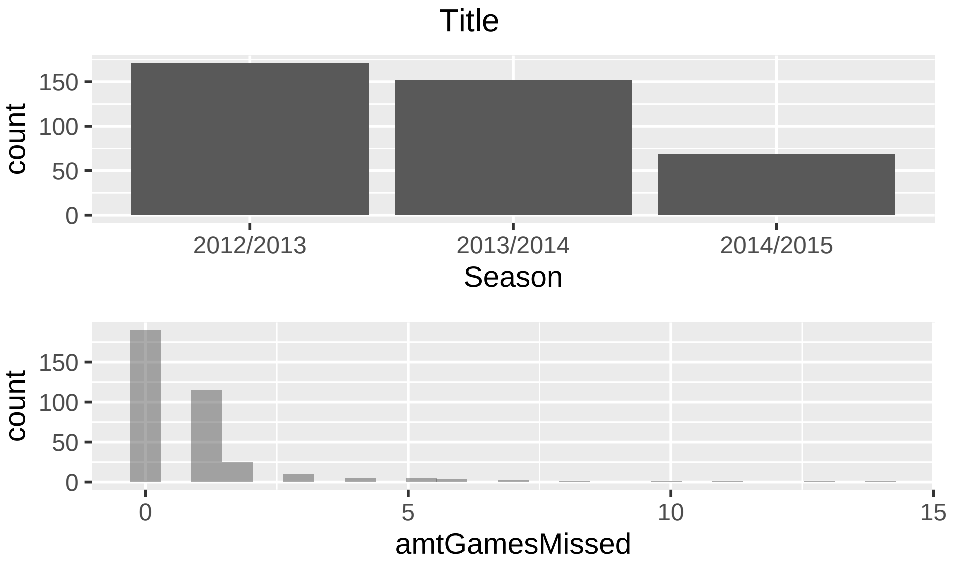
Exploratory Data Analysis (EDA)











While you don’t have to do all of the exploratory data analysis for the proposal, you can get started here by looking at univariate descriptives and graphs for the variables your team is considering.

Explore distributions and associations graphically and numerically.

Help with plotting more than one at a time using grid.arrange()

add to library section at top of RMD

library(gridExtra) library(grid)

Sample code to make plots together. These are not your variables but you can see how it’s done.“‘{r, warning = FALSE, echo = FALSE ##univariate analysis of response variable - describes shape, center and spread #favstats(~GASP, data = acs1) #gf_histogram(~GASP, bins = 10, data = acs1)

#univariate EDA on response variable (density plots of each variable individually; allows for determination of shape, center, and spread for each individual variable) m1 <- gf_dens(~ FULP, data = acs1, color = “blue”)
m2 <- gf_dens(~ ELEP, data = acs1, color = “deepskyblue”) m3 <- gf_dens(~ RNTP, data = acs1, color = “pink”) m4 <- gf_dens(~ BDSP, data = acs1, color = “purple”)

#bivariate EDA between response and qualitative predictors m5 <- gf_boxplot(GASP ~ REGION, data = acs1, color = “firebrick2”) #4 categories: Northeast, Midwest, South, West m6 <- gf_boxplot(GASP ~ HFL, data = acs1, color = “chartreuse1”) #9 categories: Gas from underground pipes, #Gas: bottled, tank, or LP, Electricity, Fuel oil, etc., coal or coke, wood, solar energy, other, None grid.arrange(m1, m2, m3, m4, m5, m6, ncol=3)

Help with rotating labels - use las=2

barplot(mytable,main=“Car makes”,ylab=“Freqency”,xlab=“make”,las=2)