STT 301 Quiz 9 Solutions
October 22, 2018

Name: _____ **PID:** _____

Directions: Answer two of the three questions below. Write "Do Not Grade" in the blank space of the question you do not want me to grade. If you answer all three questions, I will only grade the first two questions. Each question is worth 1 point. The > before any variable assignment is the prompt as you would see if you were typing this code in the RStudio console.

The `Lahman` library contains many data frames related to baseball. One data frame, called `Pitching` is about pitchers. Here are the names of the variables included in the data frame.

```
names(Pitching)
```

```
 [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"
 [6] "W"        "L"        "G"        "GS"       "CG"
[11] "SHO"      "SV"       "IPouts"   "H"        "ER"
[16] "HR"       "BB"       "SO"       "BAOpp"    "ERA"
[21] "IBB"      "WP"       "HBP"      "BK"       "BFP"
[26] "GF"       "R"        "SH"       "SF"       "GIDP"
```

1. Which of the following would output a data frame containing only players who played either for the New York Yankees (NYA) or the Boston Red Sox (BOS)? The variable containing team names is `teamID`.

```
filter(Pitching, teamID == "BOS" | teamID == "NYA")
```

2. Which of the following would compute the mean number of walks (BB) separately for each team? The output is below.

```
# A tibble: 149 x 2
  teamID mean_walks
  <fct>       <dbl>
1 ALT          10.4
2 ANA          29.1
3 ARI          21.4
4 ATL          25.6
5 BAL          28.8
# ... with 144 more rows
```

```
Pitching %>%  group_by(teamID) %>%
   summarize(mean_walks = mean(BB, na.rm = TRUE))
```

3. Which of the following would compute the mean ERA separately for each year, but for pitchers with less than 10 losses? The output is below.

```
# A tibble: 146 x 2
  yearID mean_era
   <int>    <dbl>
1   1871     6.98
2   1872     4.92
3   1873     4.88
4   1874     7.30
5   1875     4.92
# ... with 141 more rows
```

```
Pitching %>% filter(L < 10) %>% group_by(yearID) %>%
     summarize(mean_era = mean(ERA, na.rm = TRUE))
```