Assignment 2

#For this assignment, you'll be working with the county level dataset to predict a home ownership rates using conditional means. You'll need to select the county-level characteristics that you think might be related to home ownership rates. Please complete the following steps:

```
## Clear environment
rm(list=ls())
```

#Load required datasets

```
load("pd.Rdata")
##view(pd)
```

The following section loads the codebook for this dataset. It is stored as another dataset, `labels_explain`. The first column in this dataset is variable names, the second column is a full explanation of that variable.

```
## Full explanation of data in codebook
load("pd_lab_explain.Rdata")

#or use View
##View(lab_explain)
```

#Load required libraries

#1. Calculate the mean of the outcome -> The section below calculates the mean homeownership rate.

```
##Unconditional Average
pd%>%summarize(mean_homeownership_rate=mean(homeown_rate,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   mean_homeownership_rate
##                     <dbl>
## 1                    72.7
```

#2. Use your mean as a prediction: Create a new variable that consists of the mean of the outcome -> The following section creates a new variable called mean_homeownership rate as a predictor of homeownership rate by county. It sets all means equal to the overal mean of 72.7

```
##Unconditional Average as a Predictor
pd<-pd%>%mutate(mean_homeownership_rate=mean(homeown_rate,na.rm=TRUE))
```

#3. Calculate a summary measure of the errors for each observation-the difference between your prediction and the outcome. -> The following section calculates the root mean square error (RMSE) that calculates the difference between the actual observation and the mean calculated in #1 above.

```
## RMSE

rmse_uncond_mean<-rmse(pd$homeown_rate,pd$mean_homeownership_rate)
rmse_uncond_mean
```

```
## [1] 7.653637
```

#4. Calculate the mean of the outcome at levels of a predictor variable. -> The following section divides all counties into quartiles based on percentage of black residents.

```
pd<-pd%>%mutate(black_pc_level=ntile(black_pc,4))
table(pd$black_pc_level)
```

```
##
##   1   2   3   4
## 772 772 772 772
```

#5. Use these conditional means as a prediction: for every county, use the conditional mean to provide a "best guess" as to that county's level of the outcome. -> The following section uses the quartiles created above to predict home ownership rate by percentage of black residents.

```
pd<-pd%>%group_by(black_pc_level)%>% ## Group by predictor
  ##Calculate mean at each level of predictor
  mutate(pred_black_homeown_rate=mean(homeown_rate))

pd%>%group_by(black_pc_level)%>% ## Group by predictor
  ##Calculate mean at each level of predictor
  summarise(pred_black_homeown_rate=mean(homeown_rate))
```

```
## # A tibble: 4 x 2
##   black_pc_level pred_black_homeown_rate
##            <int>                   <dbl>
## 1              1                    76.2
## 2              2                    73.4
## 3              3                    71.9
## 4              4                    69.5
```

Creating a rank variable to rank homeownership by black percentage rate in each county

```
pd<-pd%>%ungroup()%>%
  #Rank by prediction, with ties sorted randomly
  mutate(pred_black_homeown_rank=rank(pred_black_homeown_rate,ties.method="random"))
```

#6. Calculate a summary measureof the error in your predictions.

```
rmse_cond_mean_one<-rmse(pd$homeown_rate,pd$pred_black_homeown_rate)
rmse_cond_mean_one
```

```
## [1] 7.26581
```