MATH 208 Assignment 2

The assignment contains two questions, with 60 possible points. Your answers must be submitted in the form of a PDF and include both the answers to the questions, along with your R code and output used to generate your answers.

Question 1 (30 points)

The FiveThirtyEight website publishes many of the datasets that are used in their articles. A former McGill statistics undergrad (now an assistant professor at Smith College) created an R package, **fivethirtyeight** that allows easier access to the all but the largest data sets.

For this question, we will look at the **biopics** dataset used in the article: *Straight Outta Compton Is The Rare Biopic Not About White Dudes*. Note that the article defines "biopics" as a class of movies that "are dramatizations, loosely based on the real-life events of actual people. Biopics offer an interpretation of lives deemed important (and profitable) by Hollywood, and they often try to make a statement about their subjects' historical or cultural significance."

```
## install.packages("fivethirtyeight")
library(fivethirtyeight)
data(biopics)
```

(a) Using the plot of your choice, assess whether the total number of biopics released per year has increased over time based on the data collected from the IMDB movie database.

(b) Produce a stacked barplot *similar* to the barplot in the original article showing the relative numbers of male and female subjects over time (Note the figures will not exactly be the same as the data in the article figures is not the same as in the dataset).

(c) Produce a stacked barplot *similar* to the barplot in the original article showing the relative numbers of white subjects, subjects who are persons of color, and unknown race subjects over time. (Mote the figures will not exactly be the same as the data in the article figures is not the same as in the dataset).

(d) Based on a mosaic plot (collapsing over year of release), which sex / white-nonwhite-NA group is the most underrepresented in biopics based on number of subjets?

(e) Produce a summary table containing counts and proportions of biopic subjects per year for each sex/white-nonwehite-NA factor combination.

(f) Create (i) a line plot showing the counts of these groups over time and (ii) a line plot showing the relative proportions of subjects over time. Would you infer from these plots that the imbalance is improving over time or not? Explain your answer.

Question 2 (30 points)

For this question, we will examine a famous diabetes dataset analyzed by Reaven and Miller (1979). We can obtain this from the **heplots** R package. We won't use any of the functions in the package (for now), but we can access the data. Using the help (?) we can see the definition of the variables.

```
install.packages("heplots")
library(heplots)
data(Diabetes)
?Diabetes
```

(a) First, create a summary table that finds the mean and median for each of the six quantitive variables with a column for each group. (Hint: use summarise, pivot_longer, and pivot_wider). Which varible(s) seem to differentiate amongst the different types of diabetes?

(b) Create 3 scatterplots, comparing all possible pairs of the glucose test variable, the insulin test varible and the sspg variable. Which pair of variables seems to allow for the strongest distinction amongst the three groups?

(c) Using the pair of variables that you chose in part (b), make 2-d histograms and contour plots for each group separately. Do you find for this dataset that these plot provide useful summaries of the differences in distributions in the three groups? Feel free to adjust the amount of binning/smoothing and the number of levels from the defalut levels.