IS 665 Midterm Spring 2018

Yegin Genc

3/8/2018

Instructions

- Please submit your answers are in .rmd format into Blackboard under Midterm.
- You will have only attempt so please submit once you finalize your work. Submission will close at 9.00 pm sharp.

Datasets

Oscars(oscars.csv) Best actor and actress Oscar winners from 1929 to 2012.

gender Gender of the oscar winner
oscar_no The number of the oscar. 1 means 1st oscar ever
name Name of the actor/actress
movie Name of the movie for which oscar was given
age Age of the winner at the time
birth_pl Birth place of the year
birth_mo Month the winner was born
birth_d Day the winner was born
birth_y Year the winner was born

Husbands and Wives Data (hw_clean.csv and hw_raw.csv). The Great Britain Office of Population Census and Surveys once collected data on a random sample of 199 married couples in Britain, recording the age (in years) and heights (in millimeters) of the husbands and wives.

Age_Husband Husband's current age
Ht_Husband Husband's height
Age_Wife Wife's age
Ht_Wife Wife's height
Age_Husb_at_Marriage Husband's age at the time of marriage
Years_Married Number of years since marriege
Age_Wife_At_Marriage Wife's marriage at the time of the marriage
Duration Whether the years married is less than or equal to 20 years (duration = 1) or more than 20 years (duration = 0)

Questions

1. (10 pts) Assuming you have a data set called `mydata` . Please explain what the following commands will do.

a. `names(mydata)`

b. `head(mydata[,3])`
c. `typeof(mydata)`
d. `df=data.frame(mydata)`
e. `cbind(mydata, 1:nrow(mydata))`

2. (20 pts) The first Oscar awards for best actor and best actress were given out in 1929. The `oscars` dataset includes information about all the oscar winning actors and actresses between 1929 and 2012.

a. Check if there are any technically correct variables but that don't make sense ? If there are any, explain why you think they don't make sense.

b. Show the movie(s) that Jack Nicholson won the oscar for.

c. What is the maximum number of oscars won by an actor or actress. Who has the max number of oscars? (Find it with R commands)

d. From the data, create a new variable that shows the year the oscar was given. (Hint: You will need to calculate it based on what is given in the data)

3. (25 pts) Answer the following questions based on the Husbands and Wives data sets. Please provide the R commands necessary for every answer.

a. At what stage of the data preparation process is 'hw_raw.csv'. Please conduct the next steps to get it ready for analaysis.

b. Using the clean data (hw_clean.csv), provide descriptive statistics on one of the numerical variables in the data.

c. What does the difference between mean and median of Husbands' age at the time of marriage suggest about the age distribution of husbands at the time of marriage?

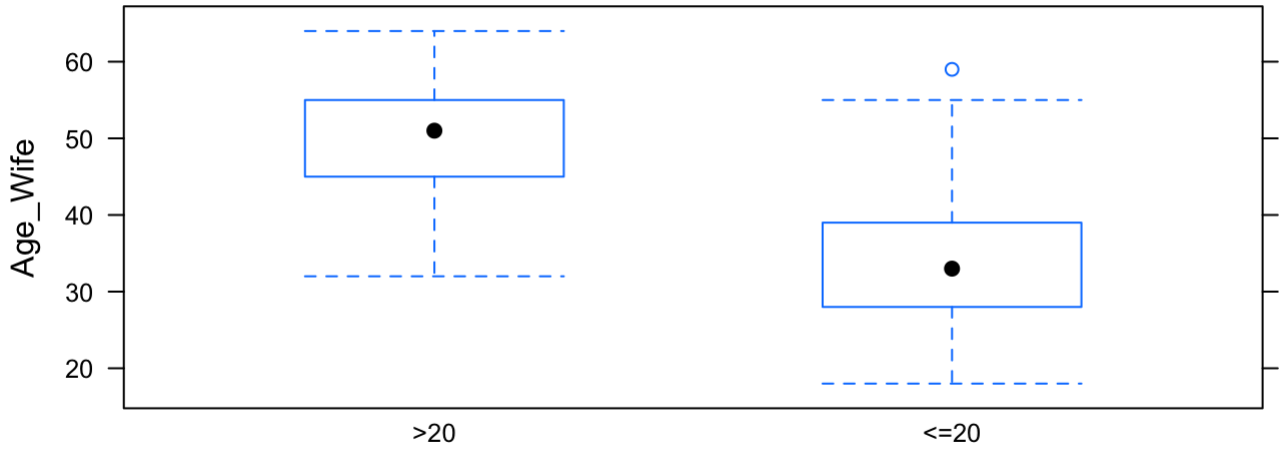
d. Using the clean data (hw_clean.csv), show the number of couples that are married less than 20 years and the number of couples that are married 20 or more years. The result should look like below.

```
| <20   | >=20 |
| 75    | 94   |
```

e. Plot two graphs that show the husbands' and wives' height distributions (you can create two separate graphs). What do the graphs suggest about the heights of husbands and wives in the sample set? Briefly explain.

4. (20 pts) Perform the following exploratory analyses on the Husbands and Wives data.

a. Using the clean data (hw_clean.csv), plot boxplot diagrams for Wives' ages for different Duration categories ('<=20' and '>20') in the dataset. The graph should look like below. What does the graph reveal about the differences for the wives' ages for the two duration values? What is a plausable explanation for what the graph reveals ?



b. Using the clean data (hw_clean.csv), create a scatterplot with variable of your choosing and try to explain the graph.

c. Create a histogram showing the age differences between husbands and wives in the data set. (Hint For this, you will need to create a new variable that shows the age difference between couples. We expect difference to be always positive).

d. What is average number of years married for couples such that both husband and wife were less than 20 years old when they got married.

5. (25pts) Answer the following questions in a few sentences (or less).

a. When do we need to transform a variable? Why ?

b. What percent of a distribution is expected to be between the first quartile (Q1) and the median?

c. What is a suitable graph for studying relationships between two numerical variables?

d. With an example, explain how exploratory analysis is different than hypothesis testing?

e. What is the difference between descriptive and predictive analytics?