MODULE 2 R PRACTICE

Week 2

Abstract

In this assignment, I will use everything that I learned so far to analyze data to complete a data rich and visually appealing report

Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

Introduction

For this assignment, I will use the data from Kaggle (Visser, 2015). The woodbine_horses.csv dataset and use the APA system for the bibliography. Also, the R code that I used to process the data is available on my GitHub account, which address is mentioned in the bibliography part. I also put my code in the appendix part.

The objective is to learn how produce several descriptive statistics tables. Secondly, to make various appealing plots.

Content and context of the dataset

Woodbine Racetrack is a Thoroughbred horse racing track located in the Etobicoke neighborhood of Toronto. The Queen's Plate, Canada's most famous race, is owned and managed by Woodbine. The track, which features a one-mile oval dirt track and a seven-eighths turf course, first opened in 1956. (Wikipedia Contributors, 2022)

This dataset represent a valuable data about of races in Woodbine Racetrack from the period of 06/2015 - 04/2017.

Methodology

This assignment is divided into three parts, each dealing with one of the aspects of the project.

Part 1: the variables of interest

This dataset has 27 variables, and it is a vast dataset; therefore, I choose only name , weight , age , sex (Colt, Gelding, Mare, Filly & Stallion) , speed rating ,

```
 > names(horse)

[1] "name"   "weight" "age"    "sex"    "speed"
```

Part 2: data cleaning

For this part, I corrected data formats and also deleted duplicated rows.

```
 #data cleaning part

str (horse)

horse$sex = as.factor(horse$sex)

#deleting duplicated datas

horse <- horse[!duplicated(horse), ]
```

Part 3: creating tables and graphs

For this part, at first I grouped data by gender and created a table from them

```
 # I group my data by gender so I can make better plots

gnames <- horse %>%

  group_by(sex)%>%

  dplyr::summarise_if(is.numeric,mean,na.rm = TRUE)



gnames <- na.omit(gnames)

print (gnames)
```

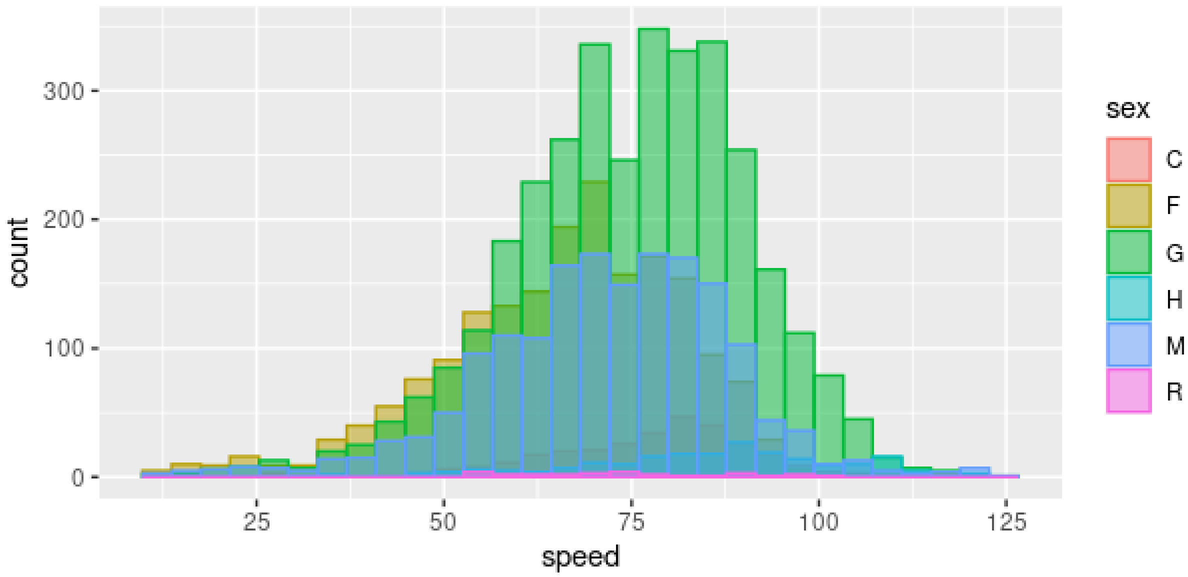
the resulting table is

```
 > print (gnames)

# A tibble: 6 × 4

  sex   weight   age speed

  <fct>  <dbl> <dbl> <dbl>

1 C       119.  2.56  76.8

2 F       118.  2.73  65.1

3 G       120.  4.33  73.9

4 H       121.  4.88  85.1

5 M       120.  4.78  70.6

6 R       119.  5     74.7
```

As it can be seen Colt horses were youngest but not the fastest. the fastest gender is stallion (H) with average speed rating of 85.1

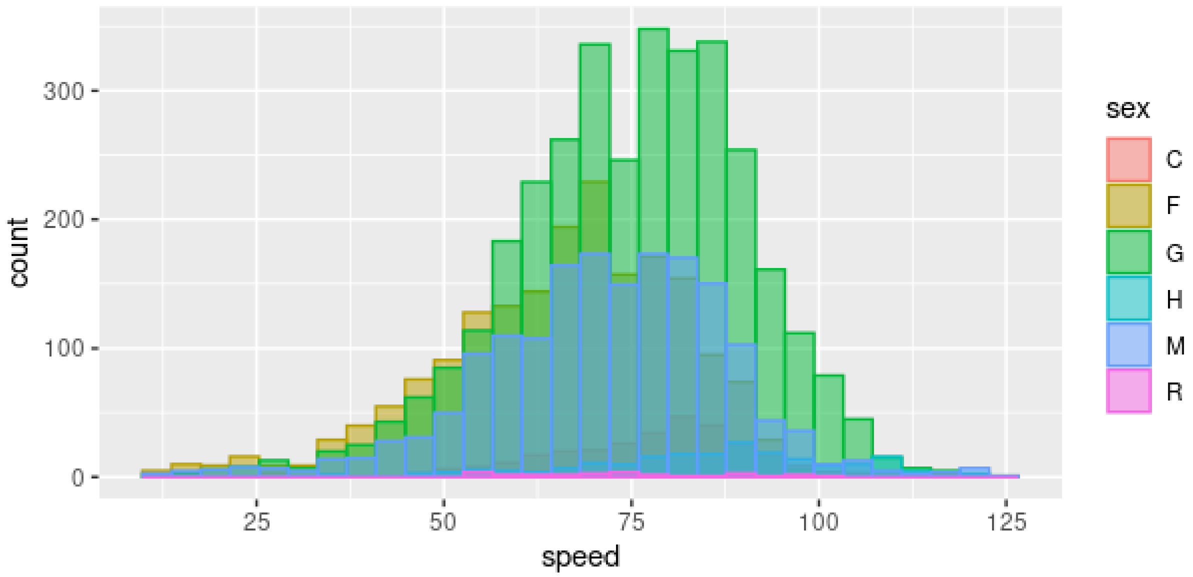
Also I created Histogram based on gender

```
 ggplot(horse, aes(x=speed, color=sex, fill=sex)) +

  geom_histogram(position="identity", alpha=0.5)
```

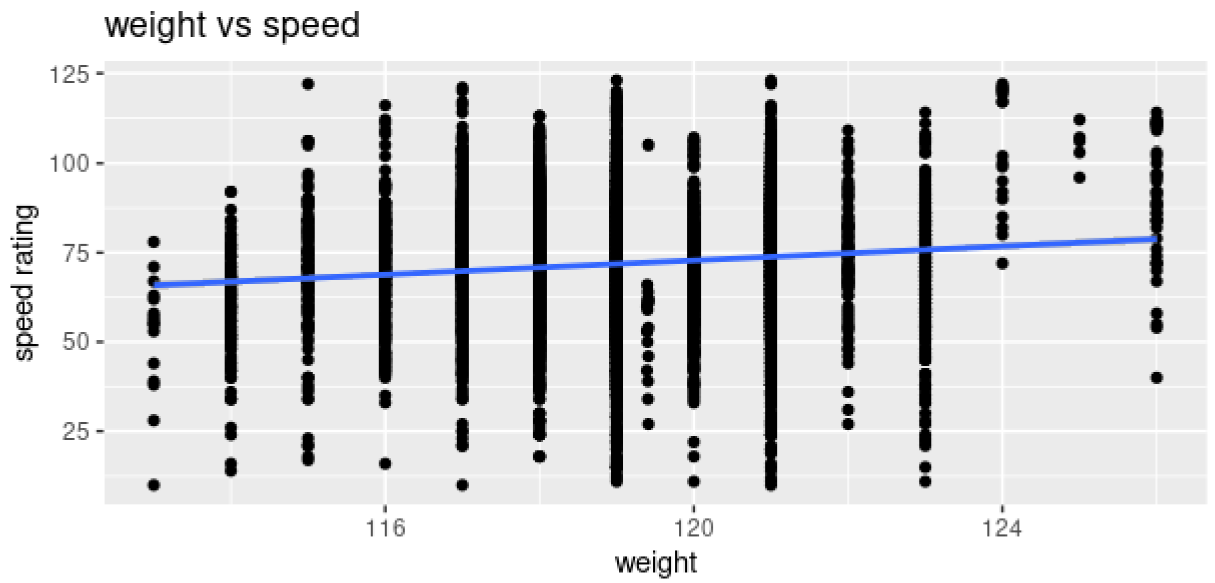


I also created a density plot too

```
ggplot(horse, aes(x=speed))+

 geom_density(data=subset(horse,sex == 'C'),fill = "red", alpha = 0.2) +

 geom_density(data=subset(horse,sex == 'F'),fill = "green", alpha = 0.2)+

 geom_density(data=subset(horse,sex == 'G'),fill = "yellow", alpha = 0.2) +

 geom_density(data=subset(horse,sex == 'H'),fill = "orange", alpha = 0.2)+

 geom_density(data=subset(horse,sex == 'M'),fill = "pink", alpha = 0.2) +

 geom_density(data=subset(horse,sex == 'R'),fill = "purple", alpha = 0.2)

 labs(title ="distibution of speed by genders", x="", y = "Density")
```

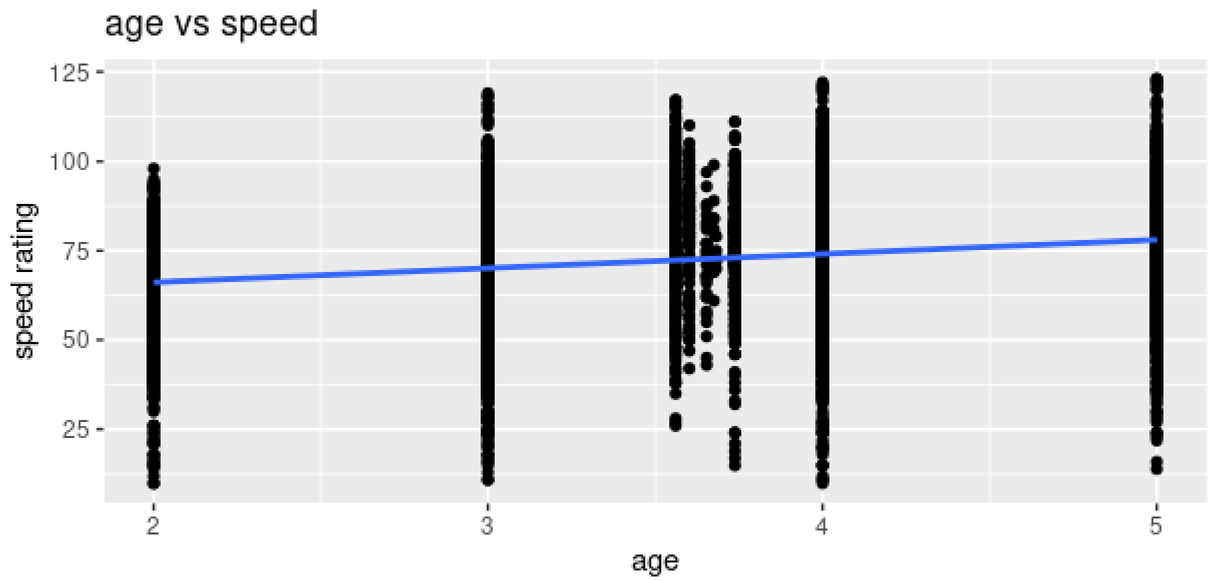


```
 # scatter plot

ggplot(horse, aes(x=weight, y=speed )) + geom_point()+

  labs(title ="weight vs speed", x="weight", y = "speed rating")+

  geom_smooth(method=lm)
```



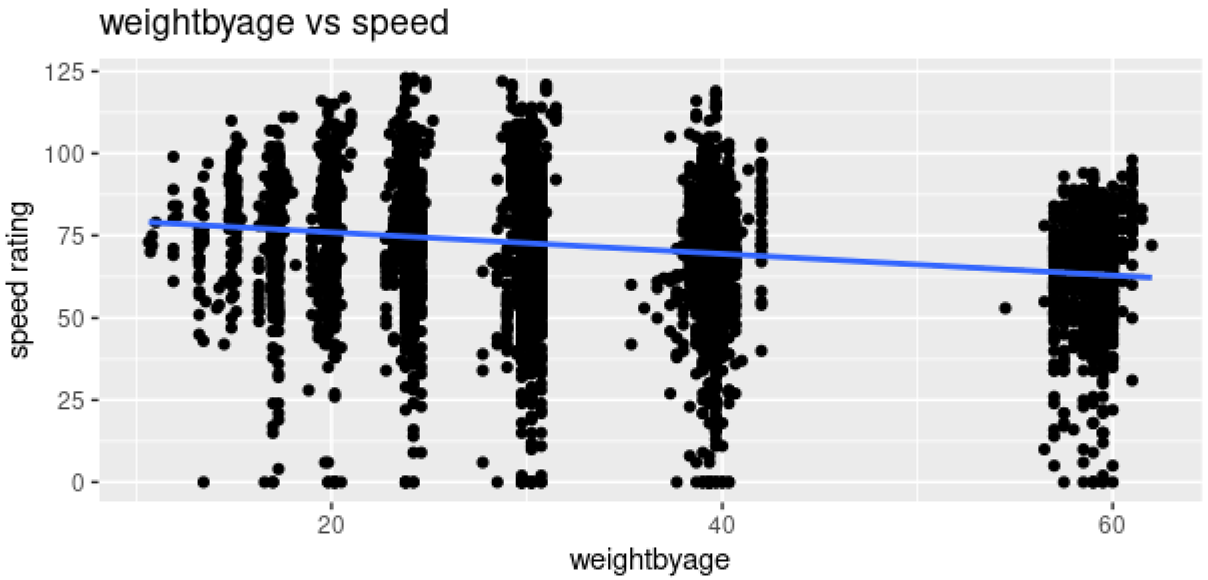
as we can see there is no relation between weight and speed rating

```
 # scatter plot

ggplot(horse, aes(x=age, y=speed )) + geom_point()+

  labs(title ="age vs speed", x="age", y = "speed rating")+

  geom_smooth(method=lm)
```



The same is ture for age too

```
 horse$weightbyage <- horse$weight/horse$age

# scatter plot

ggplot(horse, aes(x=weightbyage, y=speed )) + geom_point()+

  labs(title ="weightbyage vs speed", x="weightbyage", y = "speed rating")+

  geom_smooth(method=lm)
```



The scatter plot of the weight by age shows there are some cluster of data.

This database is very large so making jitter plot doesn't make sense here (Tidyverse.org, 2022)

Bibliography

Tidyverse.org. (2022). Jittered points — geom_jitter. [online] Available at: https://ggplot2.tidyverse.org/reference/geom_jitter.html [Accessed 9 Mar. 2022].

Horsey Hooves. (2020). Horse Genders: What is a Colt, Gelding, Mare, Filly & Stallion? [online] Available at: https://horseyhooves.com/horse-genders/#:~:text=When%20talking%20about%20a%20horse%27s,young%20male%20and%20female%20horses. [Accessed 9 Mar. 2022].

Sthda.com. (2020). ggplot2 - Essentials - Easy Guides - Wiki - STHDA. [online] Available at: http://www.sthda.com/english/wiki/ggplot2-essentials [Accessed 2 Mar. 2022].

Visser, B. (2015). Woodbine Horse Racing Results. [online] Kaggle.com. Available at: https://www.kaggle.com/noqcks/woodbine-races [Accessed 9 Mar. 2022].

momova97 (2022). momova97/ALY6010_Movahedi: This is the place that I will keep my projects R code. [online] GitHub. Available at: https://github.com/momova97/ALY6010_Movahedi [Accessed 8 Mar. 2022].

Appendix

```
 print("Mohammad Hossein Movahedi")

Print("week2")

#reading the dataset

data <- read.csv("woodbine_horses.csv")

#installing Important libraries

install.packages('FSA')

install.packages('FSAdata')

install.packages('magrittr')

install.packages('dplyr')

install.packages('tidyr')

install.packages('plyr')

install.packages('tidyverse')

install.packages('outliers')

install.packages('ggplot2')

library(ggplot2)

library(outliers)

library(FSA)

library(FSAdata)

library(magrittr)

library(dplyr)

library(tidyr)

library(plyr)

library(tidyverse)

library(scales)

# cheaking dataset structure

str(data)

names (data)
```

```
#selecting varibale of intrest

horse <- data.frame(name = data$name,weight = data$weight,age = data$age,speed =
data$speed_rating,sex =data$sex)

names(horse)

#data cleaning part

str (horse)

horse$sex = as.factor(horse$sex)



#deleting duplicated datas

horse <- horse[!duplicated(horse), ]

# I group my data by gender so I can make better plots

gnames <- horse %>%

  group_by(sex)%>%

  dplyr::summarise_if(is.numeric,mean,na.rm = TRUE)



gnames <- na.omit(gnames)

print (gnames)

#density plot

ggplot(horse, aes(x=speed))+

  geom_density(data=subset(horse,sex == 'C'),fill = "red", alpha = 0.2) +

  geom_density(data=subset(horse,sex == 'F'),fill = "green", alpha = 0.2)+

  geom_density(data=subset(horse,sex == 'G'),fill = "yellow", alpha = 0.2) +

  geom_density(data=subset(horse,sex == 'H'),fill = "orange", alpha = 0.2)+

  geom_density(data=subset(horse,sex == 'M'),fill = "pink", alpha = 0.2) +

  geom_density(data=subset(horse,sex == 'R'),fill = "purple", alpha = 0.2)

  labs(title ="distibution of speed by genders", x="", y = "Density")



#histogram

ggplot(horse, aes(x=speed, color=sex, fill=sex)) +
  geom_histogram(position="identity", alpha=0.5)




horse$weightbyage <- horse$weight/horse$age

# scatter plot

ggplot(horse, aes(x=weightbyage, y=speed )) +geom_jitter()

  labs(title ="weightbyage vs speed", x="weightbyage", y = "speed rating")+

  geom_smooth(method=lm)
```