Assignment: ASSIGNMENT 4.2 Exercise

Name: Praghuwanshi, Prashant

Date: 2021-07-03

Check your current working directory using `getwd()`

getwd()

List the contents of the working directory with the `dir()` function

dir()

If the current directory does not contain the `data` directory, set the

working directory to project root folder (the folder should contain the `data` directory

Use `setwd()` if needed

setwd(“D:/MS_DataScience/DSC 520-Datastatiics.pdf/dsc520”)

Load the file `data/Scores.csv` to `score_df1` using `read.csv`

Examine the structure of `score_df1` using `str()`

score_df1 <- read.csv(file = “data/Scores.csv”, sep = “,”, header=TRUE)

##A professor has recently taught two sections of the same course with only one difference between the sections. ##In one section, he used only examples taken from sports applications, and in the other section, ##he used examples taken from a variety of application areas. The sports themed section was advertised as such; ##so students knew which type of section they were enrolling in. The professor has asked you to compare student performance in the two sections using course grades and ##total points earned in the course. You will need to import the Scores.csv dataset that has been provided for you. ##Use the appropriate R functions to answer the following questions ##What are the observational units in this study?

score_df1 str(score_df1)

Answer: we have 38 observations with three varibales, Score & Counts are observation varibales which is having measurements and Section is observation unit

##Identify the variables mentioned in the narrative paragraph and ##determine which are categorical and quantitative? ##Answer: quantitative variables : Score & Counts categorical:Section

##Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section

install.packages(“dplyr”) library(dplyr) sports_Section <- select(filter(score_df1, Section == ‘Sports’),c(‘Count’,‘Score’,‘Section
Regular_Section <- select(filter(score_df1, Section == ‘Regular’),c(‘Count’,‘Score’,‘Section’)) Regular_Section

##Use the Plot function to plot each Sections scores and the number of students achieving that score. vsports_Section_cnt <- dplyr::pull(sports_Section, 1) vsports_Section_cnt vsports_Section_score <- dplyr::pull(sports_Section, 2) vsports_Section_score vRegular_Section_cnt <- dplyr::pull(Regular_Section, 1) vRegular_Section_cnt vRegular_Section_score <- dplyr::pull(Regular_Section, 2) vRegular_Section_score

sports_Section_plot <- plot(vsports_Section_score, vsports_Section_cnt) Regular_Section_plot <- plot(vRegular_Section_score, vRegular_Section_cnt)

##Use additional Plot Arguments to label the graph and give each axis an appropriate label.

sports_Section_plot <- plot(vsports_Section_score, vsports_Section_cnt, main="Sport Section Students Score Vs Count",ylab="Student count", xlab="Student score",col="blue")

Regular_Section_plot <- plot(vRegular_Section_score, vRegular_Section_cnt, main="Regular Section Students Score Vs Count",ylab="Student count", xlab="Student score", col="red")

##Once you have produced your Plots answer the following questions ##Comparing and contrasting the point distributions between the two section, ##looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer

plot(vsports_Section_score, vsports_Section_cnt, main="Overlaying Graphs", ylab="", type="l", col="blue") lines(vRegular_Section_score,vRegular_Section_cnt, col="red") legend("topleft",c("vsports_Section_cnt","vReg

ANSWER :by using dot plots & Line distribution seesm the tendaecny and consistency of both section is differnet and seesm sport section tends to score more as compare to regular section

##Did every student in one section score more points than every student in the other section? ##If not, explain what a statistical tendency means in this context.

boxplot(vRegular_Section_score,vsports_Section_score, main = "Multiple boxplots for comparision", at = c(1,2), names = c("Regulal", "sports"), las = 2, col = c("orange","red"), border = "brown", horizontal = TRUE, notch = TRUE) ##ANSWER if we compare the mean score of each section by using boxplots then will come to know that in regular section the mean score of students is more than the mean score of sport section

##What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

i believe we need have one more variable called lecture no (like first lecture, second , third lecture)

##We interact with a few datasets in this course, one you are already familiar with, ##the 2014 American Community Survey and the second is a Housing dataset, ##that provides real estate transactions recorded from 1964 to 2016. For this exercise, ##you need to start practicing some data transformation steps - ##which will carry into next week, as you learn some additional methods.
##For this week, using either dataset (or one of your own - although I will let you know ahead of time that ##the Housing dataset is used for a later assignment, so not a bad idea for you to get more comfortable with now!),

##perform the following data transformations: ##Use the apply function on a variable in your dataset

install.packages("readxl") library(readxl) housing_df1 <- read_excel("data/week-7-housing.xlsx", sheet = "Sheet2") housing_df1 housing_mat1 <- as.matrix(housing_df1) housing_mat <- as.matrix(housing_df1['Sale_Price']) apply(housing_mat,2,sum) ##toltal sales price (sum of all sales amount in sales price column)

##Use the aggregate function on a variable in your dataset

aggregate(housing_df1$square_feet_total_living\ housing_df$1year_built, housing_df1, mean)

##Use the plyr function on a variable in your dataset - more specifically, ##I want to see you split some data, perform a modification to the data, and then bring it back together housing_df2 <- head(housing_df1, 200) housing_df2 split_test <- ddply(housing_df2, ‘bedrooms’, identity) #splitting the datafram into list , group by sale_instrument variable

##Check distributions of the data split_test unsplit_test <- unsplit(split_test, bedrooms) unsplit_test

##Identify if there are any outliers

sales=c(1,2,50,45,67,200,230,55,56,49) boxplot(sales)

#find outliers values OutVals = boxplot(sales)$out

#print outlier OutVals

#find outlier index position in vector which(sales %in% OutVals)

##Create at least 2 new variables

library(“tidyverse”) library(dplyr)

df <- data.frame(player = c(‘a’, ‘b’, ‘c’, ‘d’, ‘e’), position = c(‘G’, ‘F’, ‘F’, ‘G’, ‘G’), points = c(12, 15, 19, 22, 32), rebounds = c(5, 7, 7, 12, 11)) df

#define new variable ‘scorer’ & type using mutate() and case_when()

df %>% mutate(scorer = case_when(points < 15 ~ ‘low’, points < 25 ~ ‘med’, points < 35 ~ ‘high’),type = case_when(player == ‘a’ | player == ‘b’ ~ ‘starter’, player == ‘c’ | player == ‘d’ ~ ‘backup’, position == ‘G’ ~ ‘reserve’))