Project 2
Exploratory Data Analysis (EDA) of Two Data Sets.

CPS Graduate Analytics program, Northeastern University, Seattle

ALY 6000: Introduction to Data Analytics

Professor: SELCUK BARAN

Submitted by

Syed Tanveer

Introduction: A step in the data analysis process when various methodologies are utilized to better understand the dataset in use is called exploratory data analysis (EDA), sometimes known as data exploration.

Overview: Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate)

Key Findings:

1. Understanding the dataset' can refer to several things including but not limited to…Extracting important variables and leaving behind useless variables.

2. Identifying outliers, missing values, or human error

3. Understanding the relationship(s), or lack of, between variables.

4. Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process.

Explanation:

PART-1

#1-data_2015 <- read_csv("Desktop/Tanveer_Project2./2015 (1).csv")

>>> The data set file is read and stored in the variable "data_2015."

#2-names(data_2015)
>>>names of each column on data_2015 file is called and read.
#3-view(data_2015)
>>>The called data_2015 is opened in a different tab.
#4-glimpse(data_2015)

>>>When used this the rows and columns configuration are changed. Rows are changed as column and columns are changed as rows.

#5-p_load(janitor)
data_2015 <- clean_names(data_2015)
data_2015

>>> clean_names function of janitor package makes the names of the data file R friendly.

#6-happy_df <-
data_2015[c("country","region","happiness_score","freedom")]
happy_df

>>> happiness_score, country, region, freedom are selected and stored in happy_df variable.

#7-top_ten_df <- happy_df[1:10,]
top_ten_df

>>>First 10 rows of happy_df are sliced and stored in the variable stated above.

#8-no_freedom_df <- filter(happy_df,freedom<0.20)
no_freedom_df

>>> freedom values below 0.20 are filtered and stored on a new table.

#9-best_freedom_df <- happy_df[order(happy_df$freedom),]
best_freedom_df

>>>Values of freedom is stored in descending order.

#10-data_2015 <- mutate(data_2015, gff_stat=family+freedom+generosity)
data_2015

>>>values of family, freedom and generosity are added and created in new column and stored

#11-library(dplyr)
happy_summary <- happy_df %>%summarise(
mean_happiness=mean(happiness_score),
max_happiness=max(happiness_score),
mean_freedom=mean(freedom),
max_freedom=max(freedom))

>>>A happy summary is created where mean of happiness and freedom is consolidated and max of happiness and freedom is consolidated in the variable listed above.

#12-regional_stats_df <- happy_df%>%
group_by(region)%>%summarise(country_count=n(),
mean_happiness=mean(happiness_score), mean_freedom=mean(freedom))

>>>region wise data is collected and stored.

#13>>>The average gdp of the of sub-saharan and western Europe data is called and stored

#14>>the scatter plot of mean happiness vs. mean freedom

from smallest to largest values are obtained

PART-2

#1
baseball <- read.csv("Desktop/Tanveer_Project2./baseball.csv")
baseball

>>>The data downloaded is read which is stored in the directory.

#2 Spend time with the data using various exploration functions.

>>>The downloaded file was studied briefly to understand what the dataset says and what does it mean.

#3_baseball <- class(baseball)
baseball

>>>class function is used to get the type of the table that is data frame.

#4 >>The age of players are computed and reflected as the average of home runs, hit and runs scored.

#5>>>filteration is done and the players who scored zero while batting are removed from the table and stored in the variable.

#6 baseball <- read.csv("Desktop/Tanveer_Project2./baseball.csv")
baseball <- baseball %>%mutate(BA = H / AB)
baseball

>>>BA(Batting average) is mutated and stored that is number of hits/no of at bats.

#7 baseball <- baseball %>%

mutate(BA = round(H / AB, 3))

>>batting average value is rounded to 3 decimal points.

#8>> baseball <- baseball %>%
mutate(OBP = (H + BB) / (AB + BB))

On base percentage is calculated using this syntax.

#9 baseball <- baseball %>%
mutate(OBP = round((H + BB) / (AB + BB), 3))

>>>On base percentage is rounded to 3 decimal points and created as new column.

#10 strikeout_artists <- baseball %>%
arrange(desc(SO)) %>% head(10)

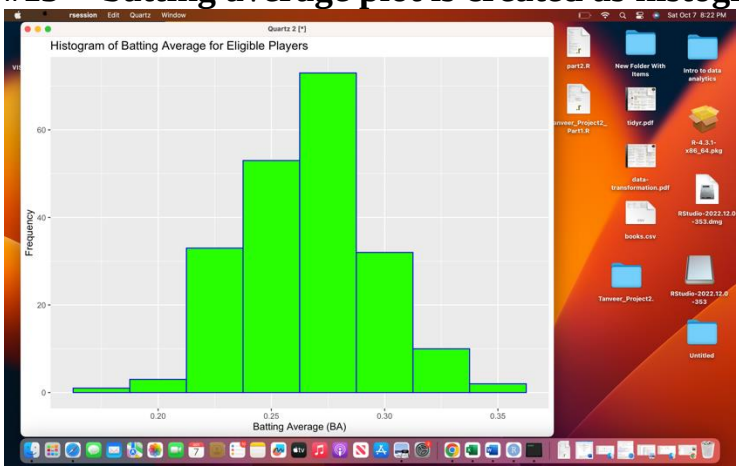
The striked out player that is the 10 players who stayed out of the season most is stored as a table.

#11>>>scatter plot is created.

#12 eligible_df <- baseball %>%filter(AB>=300,G>= 100)

>>>for the awards nominations the players who scored 100 or attended the 300 matches are collected**.**

#13>>batting average plot is created as histogram.



#14

eligible_df <- eligible_df |>
mutate(RankHR =rank(-1 * HR, ties.method = "min"))

>>players are ranked based on home runs.

#15 >>Players are ranked on RBI and OBP.

#16-eligible_df <- eligible_df %>%mutate(TotalRank = RankHR + RankRBI + RankOBP)

>>>Previous three ranks are called as total rank and the those who score 1st rank in all the three are ranked as 3.

#17-mvp_candidates <- eligible_df %>%
arrange(TotalRank) %>%
head(20)

>>>ascending order of total ranks are called and lowest 20 ranks are stored in the new variable.

#18-mvp_candidates_abbreviated <- mvp_candidates %>%
select(First, Last, RankHR, RankRBI, RankOBP)

>>>names rank of HR, RBI and OBP are consolidated in the table.

#19 Leagues most valuable player according to me is:

Mike schmith

Conclusion: Hence all the. Elements of the project is completed.

Works Cited :

https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

https://stackoverflow.com/questions/17031039/how-to-sort-a-character-vector-according-to-a-specific-order

https://www.statmethods.net/management/sorting.html#:~:text=To%20sort%20a%20data%20frame,sign%20to%20indicate%20DESCENDING%20order.

https://community.rstudio.com/t/rstudio-server-community-edition-hangs-on-lo