Assignment 2

In this assignment, you will design and build an RShiny Application. The purpose of the shiny app is to visualize and provide analytical output for a dataset of your choice (see below), using appropriate visualization and analysis methods covered in this course. The RShiny Application should be a stand alone application. That is, the user should be able to understand it's purpose and outcomes from just the application (so without requiring any additional explanation outside of the application).

For this assignment, you will use one of the six datasets provided, each of which are outlined in brief below. Each of these datasets has a large number of variables, and it is up to you which variables you will use/not use within your visualization and analysis process.

The deadline for handing in Assignment 2 is on Thursday **June 24th before your lab meeting**. The assignment should be emailed to your lab instructor as a zip file, which contains the following:

- a .Rproj file;
- the raw shiny files (ui.R/server.R);
- if appropriate, a seperate data preprocessing file (.Rmd)

Step 1:

Select one of the provided datasets, and as a group select which variables you wish to use within your RShiny Application. You are welcome to use as many of the provided variables in the dataset as you like, however it is not expected for you to use all variables, so select those which you feel are best able to use.

Step 2:

Set up your Shiny application, which includes a title and one or two lines containing the purpose of your application. Your application will use **at least three** and at maximum six interactive components (such as Checkboxes, Sliders, Select Boxes, etc…).

Step 3

Produce an interactive visualization of your data. In total, use a maximum of two plots to visualize your data **and** analysis results (see step 4).

Step 4:

Based on the dataset you have selected, intergrate at least one of the statistical methods you have been taught in this course (linear regression, logisitic regression, K-Nearest Neighbours, etc) into your application. Make sure that the application has an added value in terms of the analysis done. For example:

- use the application to compare the performance of different statististical methods on predicting an outcome in your dataset
- use the application to show the process of tuning a parameter of a statististical method

- ...

In addition, make sure the statistical method applied includes an interactive componenet. Vizualize (part of) your analysis results as part of your first interactive visualization or as a second (interactive) visualization.

Step 5:

At the bottom of the application, include an interpretation of the analysis results that includes a reactive component. This can either consist of a reactive text, or a fixed line of text combined with a reactive table providing the exact analysis results. Note: a reactive comonent is not an interactive component, and thus does not count for the min. 3 and max. 6 interactive components.

Step 6:

Present your application, as both a completed application (ui.R/server.R) in addition to well documented support files, so that it can be understood the steps you have taken both inside and outside (e.g., data preparation) of the RShiny Application to create and present this application.

Notes:

The following should be noted:

- Attempting to use too much information in a plot is as bad as not using enough information, so use your judgement to decide how to use the Shiny functions to your advantage to display the data in the most appropriate way.
- You are welcome to provide a seperate data preprocessing file (.Rmd), if you would prefer.
- Since you will be using *real-world* datasets, there is likely to be some missing data, it is up to you how you would like to manage this.
- Making the assignment alone is not allowed. Students have to stay in their assigned groups.

Grading:

Your grade will be determined by:

- Your preprocessing of the data, including the handling of missing values, and selection of variables.
- Quality and appropriateness of the data visualizations created within the RShiny Application
- Your understanding and appropriate use of different RShiny Interactive Components
- Quality, appropriateness and presentation of statistical techniques used relating to your dataset
- Quality and appropriateness of the interpretation of the analysis results and it's reactive component
- The quality of your R code (i.e., clear structure and in accordance with Hadley Wickhams Guidance)
- Overall quality of the Application (i.e., lay-out, title and description, stand-alone) and its usability

Datasets:

All datasets will be provided as .csv files via Microsoft Teams, a brief discription of each can be found below:

1. World Bank Indictors (WDB.csv)

This dataset of different global indicators from the World Bank Open Data, which includes data from over 200 countries from the 1960s - 2019. This contains the following variables (Variable name in (I):

- Country Name (*Country Name*)
- Country Code (*Country Code*)
- Continent (*Continent*)
- Year (*Year*)
- Population (*Pop*)
- Female Population (*Pop.fe*)
- Male Population (*Pop.ma*)
- Birth Rate, crude per 1000 people (*birthrate*)
- Death Rate, crude per 1000 people (*deathrate*)
- Life Expetency at Birth in years (*lifeexp*)
- Female Life Expetency at Birth in years (*lifeexp.fe*)
- Male Life Expetency at Birth in years (*lifeexp.ma*)
- Educational Spending, percetage of GDP (*ed.spend*)
- Compulsory Education Duration in Years (*ed.years*)
- Labour Force Total (*labour*)
- Literature Rate in adults, percentage % (*lit.rate.per*)
- CO2 Emissions, kt (*co2*)
- Gross Domestic product, $ (*gdp*)
- Unemployment, percentage of total labour force (*unemp*)
- Female Unemployment, percentage of total labour force (*unemp.fe*)
- Male Unemployment, percentage of total labour force (*unemp.ma*)
- Health Expenditure per capita, $ (*health.exp*)
- Hospital Beds per 1000 people (*medbeds*)
- Number of Surgical Procedures per 1000 people (*surg.pro*)
- Number of Nurses & Midwives per 1000 people (*nurse.midwi*)

2. College Basketball Dataset (colbaskdat.csv)

This is a dataset from the 2015-2020 Division I college basketball (USA), provided by Kaggle (https://www.kaggle.com/andrewsundberg/college-basketball-dataset). This contains the following variables (Variable name in (*I*):

- Ranking (*RK*): The ranking of the team at the end of the regular season according to barttorvik
- Team (*TEAM*): The Division I college basketball school
- Athletic Conference (*CONF*): The league the school participates in (A10 = Atlantic 10, ACC = Atlantic Coast Conference, AE = America East, Amer = American, ASun = ASUN, B10 = Big Ten, B12 = Big 12, BE = Big East, BSky = Big Sky, BSth = Big South, BW = Big West, CAA = Colonial Athletic Association, CUSA = Conference USA, Horz = Horizon League, Ivy = Ivy League, MAAC = Metro Atlantic Athletic Conference, MAC = Mid-American Conference, MEAC = Mid-Eastern Athletic Conference, MVC = Missouri Valley Conference, MWC = Mountain West, NEC = Northeast Conference, OVC = Ohio Valley Conference, P12 = Pac-12, Pat = Patriot League, SB = Sun Belt, SC = Southern Conference, SEC = South Eastern Conference, Slnd = Southland Conference, Sum = Summit League, SWAC = Southwestern Athletic Conference, WAC = Western Athletic Conference, WCC = West Coast Conference)
- Number of Games Played (*G*)
- Number of Games Won (*W*)
- Adjusted Offensive Efficiency (*ADJOE*): An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense
- Adjusted Defensive Efficiency (*ADJDE*): An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense
- Power Rating (*BARTHAG*) Chance of beating an average Division I team
- Effective Field Goal Percentage Shot (*EFG_0*)
- Effective Field Goal Percentage Allowed (*EFG_D*)
- Turnover Percentage Allowed (*TOR*): Turnover Rate
- Turnover Percentage Committed (*TORD*): Steal Rate
- Offensive Rebound Percentage (*ORB*)
- Defensive Rebound Percentage (*DRB*)
- Free Throw Rate (*FTR*): How often the given team shoots Free Throws
- Free Throw Rate Allowed (*FTRD*): Free Throw Rate Allowed
- Two-Point Shooting Percentage (*2P_O*)
- Two-Point Shooting Percentage Allowed (*2P_D*)
- Three-Point Shooting Percentage (*3P_O*)
- Three-Point Shooting Percentage Allowed (*3P_D*)
- Adjusted Tempo (*ADJ_T*): An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo
- Wins Above Bubble (*WAB*): The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it
- Post Season (*POSTSEASON*) Round where the given team was eliminated or where their season ended (R68 = First Four, R64 = Round of 64, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness Tournament for that given year)
- Seed in the NCAA (*SEED*): Seed in the NCAA March Madness Tournament
- Season/Year (*Year*)

3. Spotify - Top 2000 (Spotify-2000.csv)

This is a dataset which contains the audio statistics from the top 2000 tracks on Spotify, provided by Kaggle (https://www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset). This contains the following variables (Variable name in (I):

- *Index*: ID
- *Title*: Name of the Track
- *Artist*: Name of the Artist
- *Top Genre*: Genre of the track
- *Year*: Release Year of the track
- *Beats per Minute (BPM)*: The tempo of the song
- *Energy*: The energy of a song - the higher the value, the more energtic. song
- *Danceability*: The higher the value, the easier it is to dance to this song.
- *Loudness (dB)*: The higher the value, the louder the song.
- *Valence*: The higher the value, the more positive mood for the song.
- *Length (Duration)*: The duration of the song.
- *Acoustic*: The higher the value the more acoustic the song is.
- *Speechiness*: The higher the value the more spoken words the song contains
- *Popularity*: The higher the value the more popular the song is.

4. Housing Sales in King County, USA (2014-2015); (kc_house_data.csv)

This is a dataset which contains Housing sales in King County, USA, provided by Kaggle (https://www.kaggle.com/harlfoxem/housesalesprediction?select=kc_house_data.csv). This contains the following variables (Variable name in (I):

- ID (*ID*)
- Date (*Date*)
- Price of House (*Price*)
- Number of Bedrooms (*Bedrooms*)
- Number of Bathrooms (*Bathrooms*)
- Size of Living space, measured in sqft (*sqft_living*)
- Total Size of Sold Space, measured in sqft (*sqft_lot*)
- Number of Floors (*floors*)
- Is the Property on the Waterfront (*waterfront*)
- View Quality (*view*)
- House Condition (*condition*)
- House Grade (*grade*)
- Size of Floors above groundfloor (*sqft_above*)
- Size of Floors below groundfloor (*sqft_basements*)
- Year Built (*yr_built*)
- Year Renovated (*yr_renovated*)
- Zipcode (*zipcode*)
- Latitude (*lat*)
- Longitude (*long*)
- Size of Living space in 2015, measured in sqft (*sqft_living15*)
- Total Size of Sold Space in 2015, measured in sqft (*sqft_lot15*)

5. Coffee Quality from Coffee Quality Institute (CQI) (coffee.sort.csv)

This is a dataset which contains data relating to the quality of coffee, provided by Kaggle (https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi?select=merged_data_cleaned.csv). This contains the following variables (Variable name in (I):

- Coffee Bean Species (*Species*)
- Country of Origin (*Country.of.Origin*)
- Region (*Region*)
- Name of the Farm (*Farm.Name*)
- Farm Owner (*Owner*)
- Farm Company (*Company*)
- Coffee Bean Certification Body (*Certification.Body*)
- Measurement unit (*unit_of_measurement*)
- Farm Altitude (*Altitude*)
- Highest Altitude Point (*altitude_high_meters*)
- Lowest Altitude Point (*altitude_low_meters*)
- Number of Bags Produced (*Number.of.Bags*)
- Weight of Bags Produced (*Bag.Weight*)
- Year of Harvest (*Harvest.Year*)
- Date of Grading (*Grading.Date*)
- Date of Expiration (*Expiration*)
- Bean Variety (*Variety*)
- Method of Bean Processing (*Processing.Method*)
- Aroma (*Aroma*)
- Flavour (*Flavor*)
- Aftertaste (*Aftertaste*)
- Acidity (*Acidity*)
- Body (*Body*)
- Balance (*Balance*)
- Bean Uniformity (*Uniformity*)
- Clean Cup (*Clean.Cup*)
- Sweetness (*Sweetness*)
- Cupper Points (*Cupper.Points*)
- Total Cupper Points (*Total.Cup.Points*)
- Moisture (*Moisture*)
- Category One Defects (*Category.One.Defects*)
- Category Two Defects (*Category.Two.Defects*)
- Quakers (*Quakers*)
- Bean Colour (*Color*)

6. Pokemon, Gens 1-7 (pokemon.sort.csv)

This is a dataset which contains pokemon statistics, for all pokemon generations 1-7, provided by Kaggle (https://www.kaggle.com/rounakbanik/pokemon). This contains the following variables (Variable name in (*I*):

- The English name of the Pokemon (*name*)
- The Original Japanese name of the Pokemon (*japanese_name*)
- The entry number of the Pokemon in the National Pokedex (*pokedex_number*)
- The numbered generation which the Pokemon was first introduced (*generation*)
- A stringified list of abilities that the Pokemon is capable of having (*abilities*)
- The Primary Type of the Pokemon (*type1*)
- The Secondary Type of the Pokemon (*type2*)
- Denotes if the Pokemon is legendary. (*is_legendary*)
- The percentage of the species that are male. Blank if the Pokemon is genderless. (*percentage_male*)
- Height of the Pokemon in metres (*height_m*)
- The Weight of the Pokemon in kilograms (*weight_kg*)
- The Classification of the Pokemon as described by the Sun and Moon Pokedex (*classification*)
- The Experience Growth of the Pokemon (*experience_growth*)
- Capture Rate of the Pokemon (*capture_rate*)
- The Base total of the Pokemon (*base_total*)
- The Base Attack of the Pokemon (*attack*)
- The Base Defense of the Pokemon (*defense*)
- The Base HP of the Pokemon (*hp*)
- The Base Special Attack of the Pokemon (*sp_attack*)
- The Base Special Defense of the Pokemon (*sp_defense*)
- The Base Speed of the Pokemon (*speed*)
- The number of steps required to hatch an egg of the Pokemon (*baseeggsteps*)
- Base Happiness of the Pokemon (*base_happiness*)
- Eighteen features that denote the amount of damage taken against an attack of a particular type (*against_?*)