DSO 545: Statistical Computing and Data Visualization

Final Exam

Spring 2021

Instructions

- This is an open book/notes exam. You are NOT allowed to use the Internet as a resource except for downloading the files from blackboard, as well as for inspecting a website for scrapping the data for one of the questions below.
- You are NOT allowed to communicate with ANY PERSON in or outside the class during the exam period.
- If you are asked to create an exact copy of some graph, make sure to replicate the graph. Pay attention to axis names, legend name/ position, order of bars etc.
- Answer all questions below.
- Don't change the data file names and save the objects as instructed in the question.
- Submit your files to blackboard. You must submit either an R script and PDF or RMD file and PDF.

"I hereby certify that I have adhered to the university policies regarding ethical behavior in preparing for and completing this midterm exam. I will not discuss the exam questions and solutions with anyone in the classroom or outside the classroom via any means."

Name: ———————————————

Signature: ———————————————

1. (1 point) When scraping using the rvest package, what type of data structure is scraped?
   a) integer
   b) character
   c) vector
   d) list
   e) data frame
2. (1 point) In the following code, what is incorrect in the shinyApp code?

```
ui <- fluidPage(
        sliderInput("obs", "Number of observations", 0, 1000, 500),
        plotOutput("distPlot") )
server <- function(input, output) {
        output$distPlot <- renderPlot({
        dist <- isolate(rnorm(“input$obs”))
        hist(dist)
shinyApp(ui, server) }
```

   a) output$distPlot
   b) “input$obs”
   c) renderPlot({
   d) plotOutput(“distPlot”)
   e) None of the above
3. (1 point) Where should you insert the inputId?
   a) renderPlot function
   b) ui
   c) server
   d) shinyApp function
   e) None of the above
4. (1 point) When working with the stringr package which function should you use to identify a character, resulting in a logical vector?
   a) str_sub
   b) str_locate
   c) str_split
   d) str_detect
   e) str_replace

5. (5 points) Create an object called **pizzas** containing the following toppings; cheese, pepperoni, sausage, and green peppers. Then identify which orders contain the pattern "pepper" using the appropriate function from the *stringr* package.

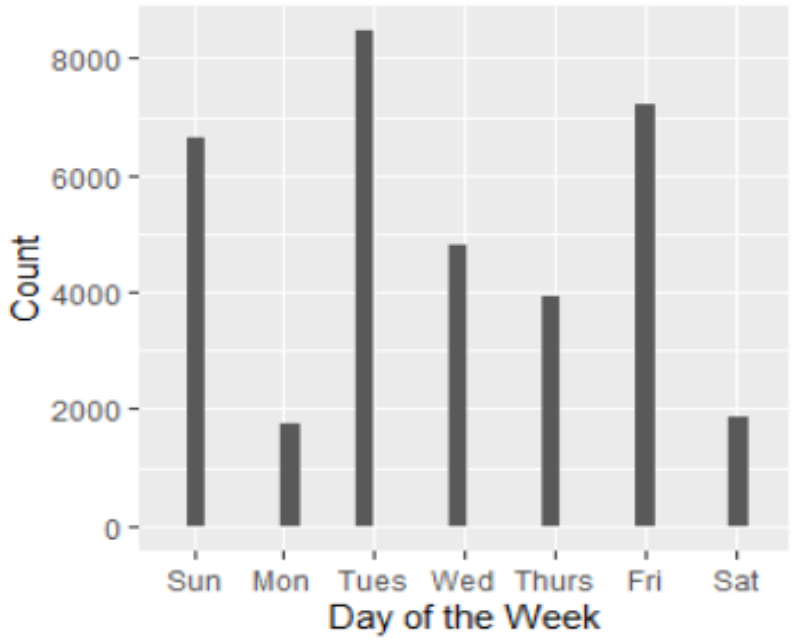
6. (5 points) Create an object called **currentvalue_cryptocurrency** containing the following current values of certain cryptocurrencies but combined with a comma in between for instance here is an example: 7.64,XTZ then using the appropriate function make the output become 7.64 XTZ (hint: *stringr* package and notice there is no longer a comma and now they is a space in between).

   Now do this for the following five using the appropriate function from the stringr package*:*

   55600,BIT 3496,ETH .58,DOGE 16.38,SUSHI 152,ETC

7. (6 points) The **lakers data set** contains play by play statistics of every major league basketball game played by the Los Angeles Lakers during the 2008-2009 season. This data is from http://www.basketballgeek.com/downloads/ (Parker 2010) and comes with the **lubridate package**. Using the str() command, we see that R recognizes the dates as integers. Before we can work with them as dates, we must parse them into R as date-time objects. (show your code, head of output, and show the structure). Convert the variable to a Date.

8. (6 points) Next, examine Lakers games throughout the week. Extract the day of the week and display the frequency of basketball games varies throughout the week. Create the graph below.

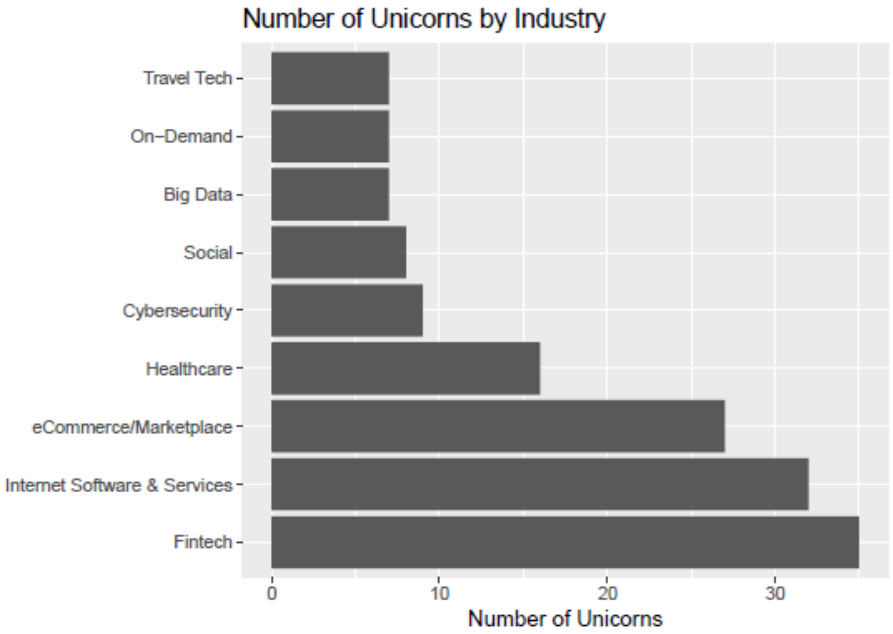


9. (6 points) Which is the statistical model that should be implemented if you were interested in gametype differences between games played home and away on number of points? Now run it and report whether there were statistically significant differences?

10. (8 points) A unicorn startup or unicorn company is a private company with a valuation over $1 billion. There are more than 300 unicorns around the world. In this exercise, we will study these unicorns, their valuations, industry, as top investors. Please download the unicorn dataset from blackboard and import into Rstudio.

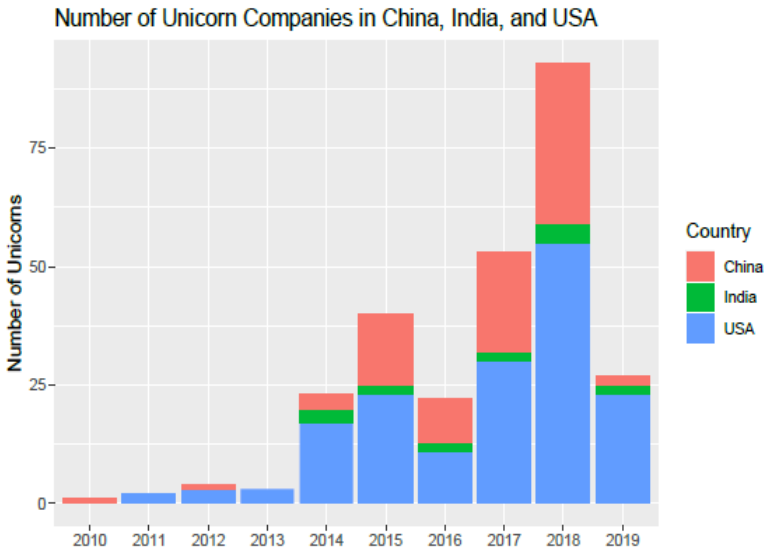
Variables	Description
Company	The name of the unicorn company
Valuation	The valuation (in billions) of the unicorn comp
DateJoined	Date in which the company became a unicorn
Country	Location of the Company
Industry	To which industry does it belong?
SelectedInvestors	A list of top investors in the unicorn

Create a bar chart like below that shows the distribution of the unicorns among different industries. Which industry has the greatest number of unicorns? Show your code.

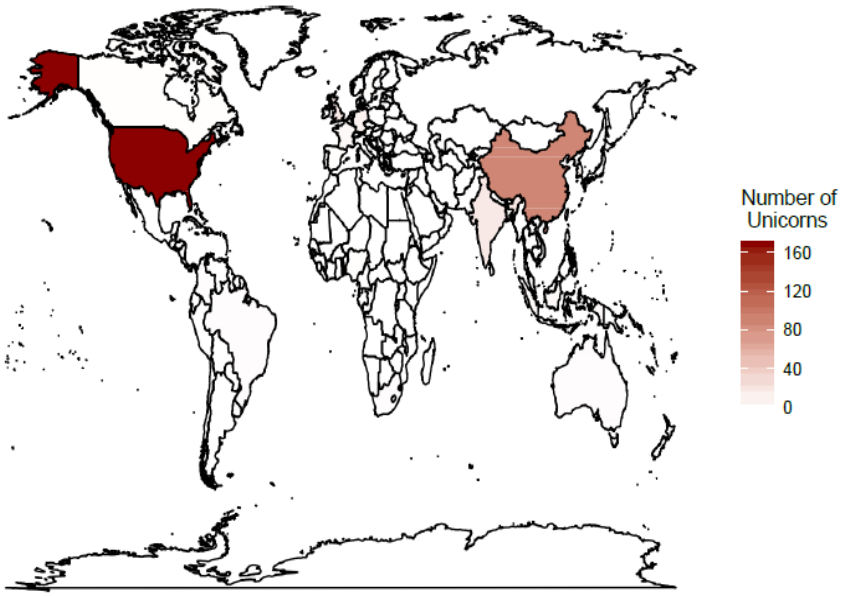


11. (5 points) What is the total valuation for the unicorns in the Fintech industry? Show your code.

12. (5 points) Use a stacked bar graph to show the total number of unicorns for India, China, and USA. Show your code.



13. (5 points) Create a choropleth map to show the number of unicorns in different countries. Show your code.



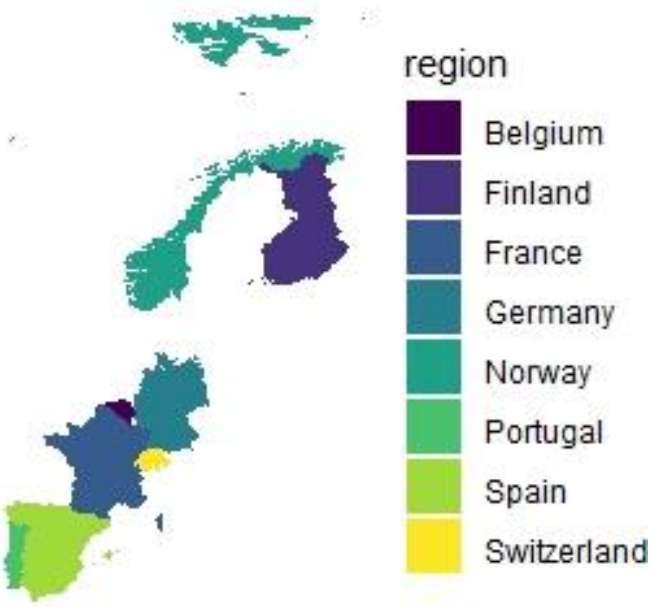
14. (5 points) Create an object called **some.eu.countries** that consists of the following countries as elements; Portugal, Spain, France, Switzerland, Germany, Belgium, Norway, and Finland.

15. (3 points) Then go ahead and retrieve the map data for some.eu.countries (which provides data for many countries in Europe) and save in an object called **some.eu.maps**.

16. (6 points) Finally, please visualize the coordinates for each of the countries' names by region and obtain a summary of the average of those coordinates, then save it in an object called **region.lab.data**.

17. (8 points) R comes with a dataset called **rock**. Run the appropriate model if we were interested in the effect's peri on area from the rock dataset (provide the R code, the output and the name of the model you chose). Is there a statistically significant finding? And what model would you use instead if you were interested in the effects of peri, shape, and perm variables in the rock dataset on area (provide the R code, the output, and the name of the model)?

18. (8 points) Now create the spatial map below and use a function scale_fill_viridis_d() + theme_void() to obtain the same color scheme.



19. (10 points) Now working with state data, please obtain state data and save it in an object called **states** and visualize the first six rows. Now that you have a states object, filter for Arizona and save it as an object **az** and display the first six rows with the appropriate function. Obtain the data at the county level and save it in an object called **counties**. Then narrow it to obtain the counties for the state of Arizona and save it in an object called **az_county**.

20. (5 points) Create a plot that shows the boundaries of the state of Arizona in the color blue.