Stellenbosch University : Economics Department

Data Science Practical Test

Mock Examination: Semester I

Lecturer: NF Katzke
Internal Moderator: Prof. R. Burger
2021

TOTAL MARKS: 100
TIME ALLOWED: 3 HOURS

INSTRUCTIONS TO CANDIDATES

1. Start a new project (name it your student number), with a README and relevant code and data folders.
2. Download and unzip the following folder from the link:
3. **www.datsci.nfkatzke.com/Mock_Practical_Data.zip**
4. Put all the data in your data folder.
5. Answer all the questions in this paper in your README.
6. EMAIL me the link to your project at **nfkatzke@gmail.com**
7. Make sure about the email (I will not accept 'I sent to the wrong email') **nfkatzke@gmail.com**
8. Use the functional programming paradigm throughout.

Question 1: World Happiness Report

The old saying goes: *Money cannot buy you happiness.*

You've been assigned to write a short article by your employer that seeks to uncover *what really buys happiness.*

Name the report: YOURSTUDENTNUMBER_Happy.pdf

Use Texevier to write this report.

A colleague of yours sent you the data sets in the folder **data/Happy**, from what you've downloaded, which is used in the The World Happiness Report. Write a function to collate this data into a single data frame. Your function should also suppress the ugly read_csv messages. Use the collated data then to construct a written narrative along the following lines:

a) Plot, per region, the Ladder Score, upperwhisker and lowerwhiskers using ggplot. Also, add directly above each region's plot the average Healthy Life Expectancy. Your plot should also be arranged by Average Life Expectancy.

- Use your discretion in how to best visualize this information.
- TIP: use geom_errorbar.
- Create a function that arranges a data frame by a given input (factor); click **here** to see a gist to help you figure out how to do this.

b) Create a barplot that shows the breakdown of Ladder scores per region. Arrange the regions as they appear in the plot by Ladder score. Also, add South Africa's ladder to this plot (make SA the first bar).

- Tip: The Ladder Score is the sum of everything that starts with *Explained by*, as well as Dystopia + residual (think of Dystopia as happiness default, if you like)

Question 2: Wine Whine Wine

After watching the documentary, Somm, you decided to become a wine studying data scientist. After joining FruityWineClubSA.com, you were tasked to write a short piece on the preferences of Sommeliers globally, and also show which wines and regions are preferred in South Africa.

You then proceeded to scrape data from the Wine Enthusiast website (Credit: Zack Thoutt). See README in folder for column explanations (tip - load txt files using readr::read_table).

Write an informal report (i.e. in Rstudio - create a new Rmarkdown document and simply select to build a Pdf Document, don't use the Texevier template for this as she is neither a professor nor a journal editor), in which you give attention to the following:.

- Plot how many ratings each country received as a barplot, with the median score placed vertically above each bar.
- Create a table of the frequency with which Sommeliers use the words: Tannins ; smoke, smokey or ash ; wooded, wooden or woody.
- Create a plot of the most referenced fruits in Sommelier's descriptions for the countries listed below. Each country should have its own plot - arranged by the 5 most referenced fruit as a sum of the percentages (use the below country order as well):
  - South Africa, Italy, France, US and Spain
  - Tip: use purrr::map to map your function across the fruits list.
- Focusing on the local wine industry - plot the 5 most preferred wineries (using median points) above $20 for the tasters Laurne Buzzeo and Susan Kostrzewa. Use your own discretion in how you want to display this.
- Add two sentences discussing the two main tasters in SA's points awarded Spearman Rank correlation with price. Show your calculation.
- Run a regression using the following formula: lm("points ~ price + province + variety")
  - Plot the fitted vs actual values for both Lauren and Susan, in order to compare the impact that known factors have on their scores (such as price, province and variety)

- we'd prefer this to be low, implying the tasters truly only value the wine tasted, not other factors.

– Tip: use my answer to a stack question by **clicking here** to answer this question. Work through the example to understand the nuances of what I try to achieve here. and tailor it to the question at hand.

END OF PAPER