Stellenbosch University : Economics Department

Data Science Practical Test

Mock Examination: Semester I

Lecturer: NF Katzke
Internal Moderator: Prof. R. Burger
2025

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. **datsci.nfkatzke.com/Mock_Prac.zip**
4. Put all the data in Mock_Prac_data in your 'data' folder, and do not commit the data folder on github. Start every question in its own folder, with an accompanying code folder
5. Provide information as to how you approached your questions in your README in the root of your folder.
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.
9. You can hand in your project any time today, but no changes may be made after midnight tonight.
10. Any changes made to your project after midnight tonight will deem your project failed.

Question 1: World Happiness Report

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

You've been assigned to write a short formal pdf 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 (using any format of your choosing - either Texevier's pdf html output or just normal html), 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.

Question 3: Profitable Movies

After a heated discussion with one of your school friends at a braai, you decided to email her addressing some of the claims she made around movie critics being near perfect predictors of films' popularity and profitability amongst audiences. Your contention was that, while you were studying together in the mid 2000s - this certainly was not the case.

After some laughter, she suggested that you prove your point.

Luckily, a friend of yours working at Mr Video between 2007 - 2012 supplied you with movie critic and grossing data in the *Movies* folder you downloaded.

> Note: the profitability column is a ratio of gross profit relative to operating expenses; Rotten Tomatoes is an aggregator of critics' movie scores, where 0% is terrible, 100% is amazing. Same for Audience Scores, where it is a polled statistic.

Write an informal letter (use Texevier's HTML pdf output in Rstudio, or create a new Rmarkdown document and simply select to build a Pdf Document), in which you test her theories, and address the points she made:

- "I firmly remember that Rotten Tomatoes was always a great review platform - and if a movie had a rating of more than 80% on Rotten Tomatoes, audiences would rate it above 85% every time.''
- "Disney films may not have the highest grossing numbers, but they've always been the most profitable of all the leading studios.''
- "Audiences are always drawn to the highest grossing films. In fact, I bet the correlation between the world wide grossing numbers and audience scores would be near 80%.

As she is a visual arts major, your preference should be in showing figures to make your points. Try to pack your figures with information, and describe properly what is going on in your plots. Remember - your reputation among your friends is on the line.

Question 4: Tennis

Recently you read a post on Reddit singing the praises of Novak Djokovic as being the greatest tennis player of all time.

Being a staunch Nadal fan, this prompted you to do some of your own work looking into the performance of tennis players and tournament results through time. You are curious as to how tennis has evolved over time, and what type of tennis players have been winners in the past - and what the persistence of performance has been historically. You reached out to your good friend, Jeff Sackmann, who supplied you with some data to do some analytics.

Your superior at Sports.com suggested you do a short write-up citing some interesting facts and factors about past tournaments, attributes of players doing well, how likely players are to repeat good results and more.

See if you can get some interesting insights from the tennis data at your disposal - you can use your discretion as to what would be interesting to your readers. Consider things like winning percentages, ace and breakpoints saved percentages, time to beat oponents, etc. See if you can provide these answers by adding stratification layers (e.g. per surface, tall players, top 10 players, etc.). Try be creative and produce at least 3 plots in sharing your insights.

Important In your folder, create a function for reading in multiple sheets. You need not use all the data - it is up to you what you want to consider.

The file, matches_data_dictionary.txt, gives more information on the definition of the columns.

Also see the README.md file, which you can open directly in RStudio.

DIY Section Question 4: Russian Invasion of Ukraine

You've been asked to be a panelist on an Australian news desk to share insights into the Russia-Ukraine war, and specifically whether countries inside the EU has done enough to stem the tide of the war. The producer of the segment "From Russia With No Love" asked that you

summarise a few key bullets to discuss around the topic of country aid (in any PDF format to be sent to them) - providing viewers with intuitive and interesting insights into which countries are giving to the Ukrainian cause and which aren't.

He asked that you be concise and not spend much time on this. Format required is any PDF / HTML output, with the results clear and concise. Provide a short summary of how you intend to interpret the results on air.

```
alloc <- read_csv("Data/Ukraine_Aid/Financial Allocations.csv")
commit <- read_csv("Data/Ukraine_Aid/Financial Commitments.csv")
```

- Instruction: Use your **Ukraine_Aid** folder in the Data folder to access the data.

END OF PAPER