Stellenbosch University : Economics Department

Data Science Practical Project

Practical Examination: Semester I

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

TOTAL MARKS: 100
TIME ALLOWED: 48 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/PracData25.zip**
4. Put all the 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.class@gmail.com**
7. Make sure about the email (I will not accept 'I sent to the wrong email') **nfkatzke.class@gmail.com**
8. Use the functional programming paradigm throughout.

Question 1: Baby Names

Summary You have been approached by a New York based kids' toy design agency that wants to do data analytics around baby naming trends in the US through the years. They are open to be guided by your expertise in data analysis to shed light on e.g.: the factors influencing the naming of children (e.g. popular movie character names, popular US presidential candidates, celebrities, billboard topping song or artist names, etc), and also the longevity of naming trends (whether some names have seen persistence in naming popularity, or whether some fads have completely faded).

They are hoping that better understanding naming trends will help them better predict which character names to use in naming their toys.

To do the analysis, you have been given a list of baby names, by year, for all the US states between 1910 - 2014. The toy design agency further advised that you can use whatever data source you want to supplement your analysis and to be creative as you possibly can.

The agency asked that you first show a time-series representation of the *rank-correlation* (TIP: use Spearman rank correlation - think carefully about how to do this as it effectively looks at the rank similarity of two tables) between each year's 25 most popular boys' and girls' names and that of the next 3 years - specifically to get a sense whether today's popular names persist into the future. See if you can confirm or deny their suspicion that since the 1990s, popular name trends have been slower to persist than in earlier decades.

Tips: also look at year-on-year surges in popularity by names - and do some research into what could have caused that. Use your discretion in looking for patterns. Your older generation supervisor remembers that in 1974 there was an odd spike for the name *Katina* - and mentioned that in 1974 it was a character on the popular TV show 'Where the Heart Is'. You overheard him saying to the client: "we can maybe look for similar interesting examples in showing how a TV show / characters / singers / celebrities through the years have caused baby name spikes. Putting this on a plot e.g. with Years or Decades on the Y-axis and most popular Names on the X (N being the size of the name bubble), while highlighting popular character names in adult or children series. . . ''

You proceeded to compile data on music and movies / series to to facilitate your analysis, and stumbled upon some interesting data sets that contain, e.g., the Top 100 Billboard songs for

each week since 1958. This should help in giving you some insight into whether people name their children after popular singers / songs (see e.g. the spike in *Whitney* names in the 1980s). There's also a list of HBO movie and series titles that you can work with - including their audience popularity scores under *tmdb_score*, as well as the actor credits under the Credits file corresponding to each movie / series' unique ID.

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

```
library(tidyverse)
Baby_Names <- read_rds("_Data/data/US_Baby_names/Baby_Names_By_US_State.rds")
Top_100_Billboard <- read_rds("_Data/data/US_Baby_names/charts.rds")
HBO_titles <- read_rds("_Data/data/US_Baby_names/HBO_titles.rds")
HBO_credits <- read_rds("_Data/data/US_Baby_names/HBO_credits.rds")
```

* **NOTE**: use your full discretion on how to make sense of the data and use it in your report (whether by state or nationally for the US, whether by looking at sports stars, music, politics book characters, etc.).

* Try and produce some interesting plots showing naming persistence as described above, distributions for some popular names, spikes in names contemporaneous with big events, etc.

* You've been asked to produce a report in PDF or HTML format to be given to the client - your results should be summarised concisely - avoid long paragraphs and instead supplement graphs / tables with bullet summaries.

Question 2: Music Taste

You were approached by Spotify to write a short report on the longevity and musical progression of some of the most famous bands over time.

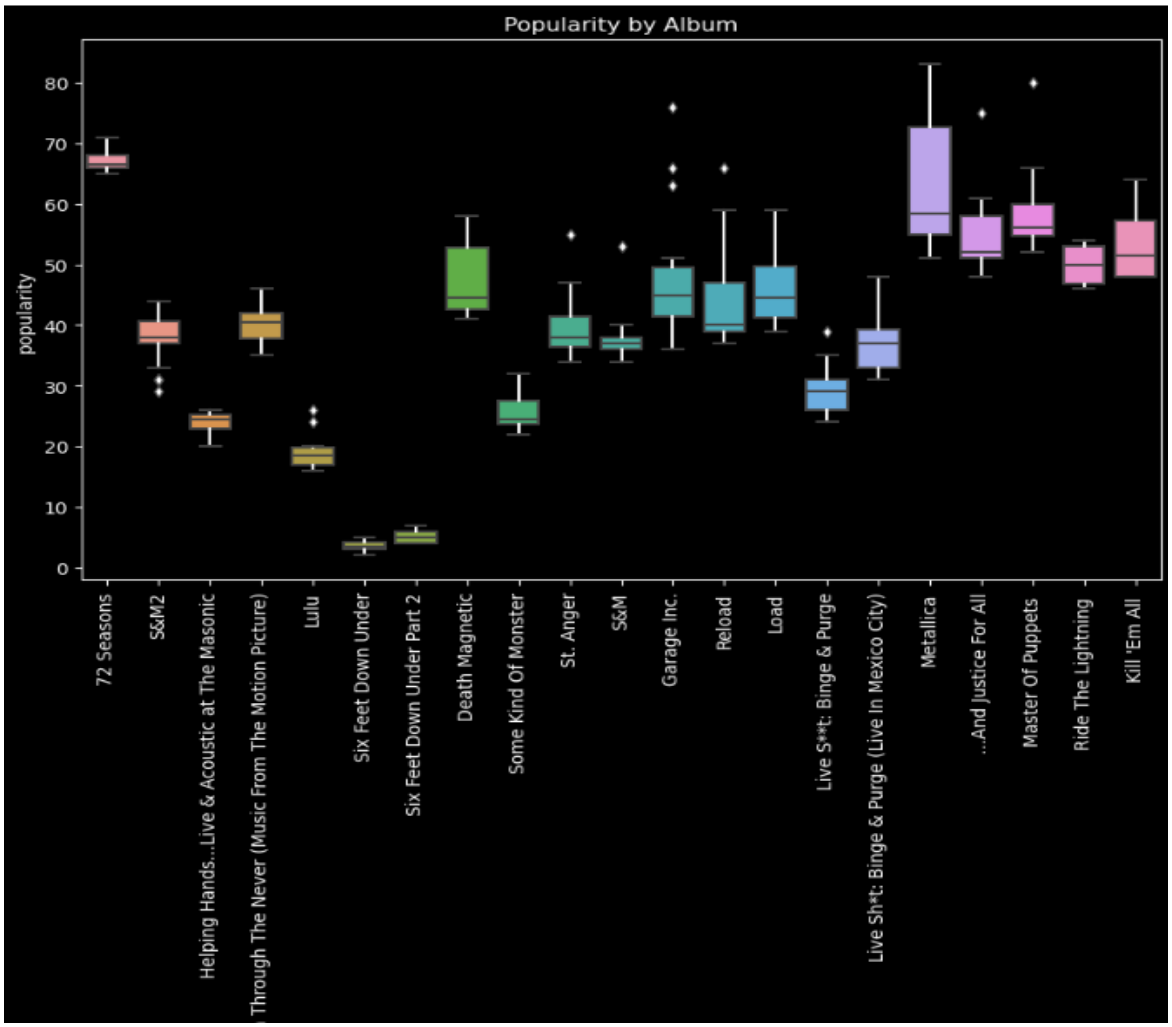
You compiled data from Spotify on Coldplay and Metallica (two bands that have more than 2 decades' worth of musical development). Compare these two famous bands using the data contained in Coldplay_vs_Metallica. Use Definitions.txt for column context. You are welcome to research further the meaning and relevance of these terms, and compare it to more modern music (should you want to add colour) - use your discretion.

You have been advised to make any direct comparisons like-for-like, by comparing studio recordings (thus filtering out Live performances - TIP: look for this in the song name).

Important: Consider that there are several iterations of the same songs (e.g. Metallica has songs with different names, e.g. being live, studio recorded, demo, etc.) This can be used for interesting analyses (e.g. the change in tempo of songs sung live, in studio or as a demo - you can be creative with this), but be careful to account for this when comparing bands directly.

Also note there is a file (albeit a bit outdated) that gives information on other songs played on Spotify - you can use this to give broader context of how both bands changed their styles through the years compared to broader trends within the music industry. You can use this, in combination with the Billboard Top 100 list (shown per week since the 1950s to supplement your analysis), to supplement your analysis with a broader music comparison.

Be creative with your analysis. Your colleague recommended including a graphic similar to the one below, as an example to get your creative juices flowing:



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

```
coldplay <- read_csv("Data/Coldplay_vs_Metallica/Coldplay.csv")
metallica <- read_csv("Data/Coldplay_vs_Metallica/metallica.csv")
spoitfy <- read_rds("Data/Coldplay_vs_Metallica/Broader_Spotify_Info.rds")
billboard_100 <- read_rds("Data/Coldplay_vs_Metallica/charts.rds")
```

Question 3: Netflix

With Netflix seeing a decline in users in recent months and a sharp decline in their share price, you've been asked to assist your superiors in doing some research into Netflix. Your superiors have ambitions of launching their own streaming service, and need a few nice figures and or tables to take to their investors to show what works and what does not work in streaming content.

Format You've been supplied with data on the titles of shows and movies available on Netflix up to 2023. You've been tasked by the team that collated the information to provide perspective on streaming content and preferences.

The source of the data provided is IMDb (an abbreviation of Internet Movie Database), an online database of information related to films, television series, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

Given that you're a respected quantitative analyst, you've been given full freedom by your superiors to explore the topic and supply insights from the data that you deem interesting.

To this end, you've been provided with Titles and Credits info, as well as info on Movies specifically. You've been asked to comment specifically on the types of movies listed on the platform from different countries (up to 2022 as provided). In particular, comment on the types of movies preferred (e.g. consider textual analysis on descriptions), ratings comparisons and length of movies.

You can use whatever format you like to present your findings (HTML, PDF, etc), but please note that you've been asked to be concise - so stick to key statistics, graphs or tables - with a short summary accompanying each. A few figures and tables (where applicable) should be sufficient.

- NOTE: You are also welcome to use the HBO data set from Question 1 to compare popularity of movies / series across genres too. There's no constraint on your creativity here...

```r
Titles <- read_rds("_Data/data/netflix/titles.rds")
Credits <- read_rds("_Data/data/netflix/credits.rds")
Movie_Info <- read_csv("_Data/data/netflix/netflix_movies.csv")
```

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

Question 4: Billionaires

Forbes recently surveyed a number of the wealthiest individuals in South Africa, and realised that they need a better understanding of the changing wealth patterns seen globally.

You've been approached by Forbes to test some of the claims made by one of the participants. Use the data in the Billionaires data folder to test the voracity of these claims. The data span 3 decades, namely the 90s, early 2000s and middle 2010s. If the analysis proves insightful, Forbes endevours to update this database in the future. Test the following claims:

- "In the US, you saw an increasing number of new billionaires emerge that had little to no familial ties to generational wealth. Other developed markets and emerging markets tend to have less entrepreneurial successes and tend to house mostly inherited wealth.''
- "Most new self-made millionaires are in software, compared to consumer services type industries in the 90s. This is related to different countries' GDP, of course, with richer countries providing more innovation in consumer services.''

Format You are encouraged to be creative and concise in providing your perspective on these statements, as well as any other interesting insights you can uncover in the data.

* NOTE: Use Info_file.xlsx to create a function to efficiently read in the Billions.csv file, by noting the column type and explicitly labelling each according to their type. Tip: see this thread.

```
Billions <- "_Data/data/Billions/billionaires.csv" %>% Bespoke_Read_Function()
```

Question 5: Health

As a follow up to your analysis on wealth, you've been asked to do a short report on health as well. The World Health Organization (WHO) funded a TV segment where you are tasked with providing insights into the different determinants of good health care for Channel 1 News.

The interviewer stressed that you only have a short time to present your findings on air, and so requested that you create a power point deck with a few charts / tables / regression analyses to provide health care insights. You are also required to do a short write-up explaining what you intend to say, in a summarised form, to accompany your Powerpoint slides.

You have been given freedom to explore the data at your disposal and to provide a practical narrative from the data stressing how health care can be improved. In a casual chat with the host in the past, you've made the point that sleeping is more important to your health than exercise, and living a stress free lifestyle has a major impact on your health as well. This resonated particularly well with the host - as she is an avid gamer that does not like exercise at all.

Format Provide analyses in the form of a few Powerpoint slides with an accompanying text document that explains what you want to say on air.

```
Health <- "_Data/data/Health/HealthCare.csv" %>% read_csv()
```

END OF PAPER