Stellenbosch University : Economics Department

Data Science Practical Project

Practical Examination: Semester I

Lecturer: NF Katzke
Internal Moderator: Prof. R. Burger
2024

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/PracData24.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. You were also given a database of US population by state and city to give you proportional insights at US state level.

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_csv("Data/US_Baby_names/Baby_Names_By_US_State.csv")
charts <- read_csv("Data/US_Baby_names/charts.csv")
Population_data <- read_csv("Data/US_Baby_names/Total_Population_By_City.csv")
HBO_Titles <- read_csv("Data/US_Baby_names/HBO_titles.csv")
HBO_Credits <- read_csv("Data/US_Baby_names/HBO_credits.csv")
```

Instruction: Use your Baby_Names folder in the Data folder to access the data.

* **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.

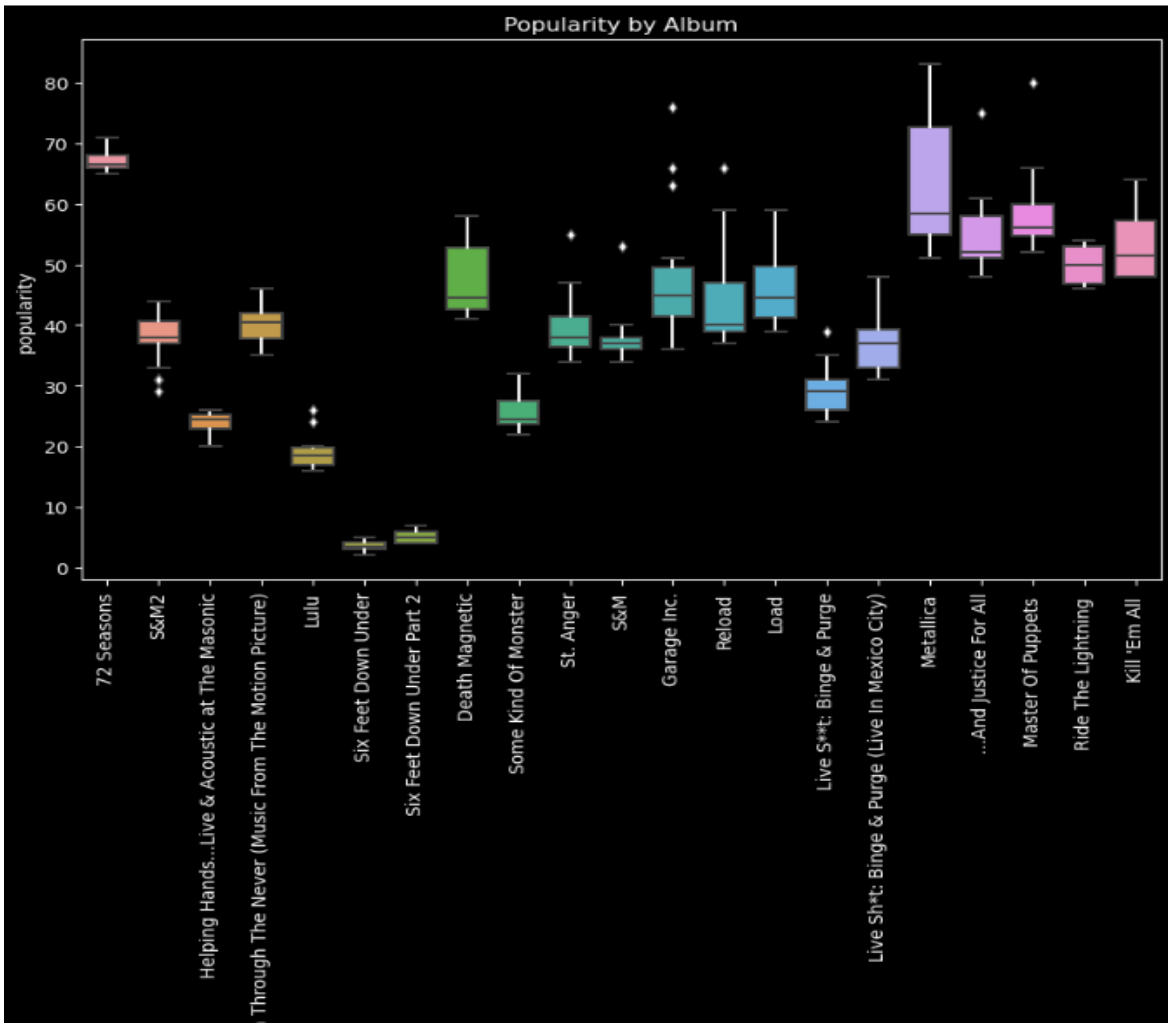
You compiled data from Spotify on Coldplay and Metallica (two bands that have more than 2 decades' worth of musical development). Compare these two famous bands using the data contained in Coldplay_vs_Metallica. Use Definitions.txt for column context. You are welcome to research further the meaning and 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_csv("Data/Coldplay_vs_Metallica/Broader_Spotify_Info.csv")
billboard_100 <- read_csv("Data/Coldplay_vs_Metallica/charts.csv")
```

Question 3: Russia - Ukraine Conflict

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.

Question 4: Olympics

Following your stellar performance on the news segment on Russia / Ukraine, another producer, this time from a TV station in New Delhi, India, gave you a call asking to produce some interesting perspectives on the upcoming Olympics. Seeing that the Olympics that is soon to be held in Paris is promising to be a riveting affair, the producer sent you the following information to consider:

- How has India faired in past summer Olympics compared to similarly sized economies, to other emerging market economies and also select South American countries?
- TIP: be careful when collating the data - the data shows medals per person, and some events (like hockey) show multiple winners (whole team) while the country receives only one medal. Be creative in your code and analysis and account for this as far as possible.
- Which countries have been most dominant in both Winter and Summer Olympics over time? Maybe show some time-series analyses of a few countries side-by-side.
- Which countries best punch above their weight when it comes to winning medals (however you define this)?
- Which is your personal favourite event at the Olympics? Show some interesting analyses on past winners / countries related to your chosen event.

In your research you came across some interesting files that may be useful in your analysis.

```
Summer <- read_csv("Data/olympics/winter.csv")
Winter <- read_csv("Data/olympics/summer.csv")
GDP <- read_csv("Data/olympics/GDP.csv")
```

Please share your figures / tables and a concise bullet point summary of what you want to share that is interesting to me in PDF form.

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

Question 5: Using SQL Queries (Dr. Odendaal's section)

By now, you're seen as a veteran news guest, and after your previous segment you've been approached by eNCA to share your views on local inflation. Your brief is to be creative in visualising the impact of inflation using a practical example that non-technical viewers would find easy to understand.

You've decided to construct a *Braaibroodjie* index. The *Braaibroodjie* (braai-bread) is synonymise with South African culture.

Your task is to analyse the inflationary impact on the *braaibroodjie* over a three-year period. You are provided with two detailed data sources: Stats SA and daily scraped prices from a major retailer. The data contains the essentials for making a braaibroodjie:

- White bread, cheddar, margarine, tomatoes, onions, salt and chutney

Connect to the PSQL prices database where the data is stored using the following credentials:

```
# Connect Suggestion:

library(dbbasic)
usethis::edit_r_environ()

gp_user="exam"
gp_pass="e4oSSMr6TXgASfkoysY5"
gp_host="102.222.21.138"
gp_port="5010"

# Database: datascience
# Remember to restart R for the above to take affect
```

A) Produce a few key graphs / tables where appropriate, and provide a summary of your findings to share with the producer ahead of time. You are required to be sufficiently concise in describing your results and clear with what your findings mean. While your

brief is to be creative, your superior suggested creating indexes for each ingredient (base: 100), showing it as time-series, calculating rolling correlations and producing informative summary tables. Be creative in making your visualizations clear and informative - you'll only be able to share a few on the segment so spend time making your plot as compelling as possible.

B) You were also asked by **Stats SA** to provide some more technical insight into the data that you've used to construct your index (this should include basic data summaries - and also compare the Stats SA data with Retailer data). Try and be as convincing as possible in having them publish the index - specifically being careful in cleaning your data and comparing your Braaibroodjie index vs the actual SA Inflation index (Tip: download CPI here - using CPS00000 CPI Headline).

END OF PAPER