Stellenbosch University : Economics Department

Data Science Practical Test

Practical 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. **datsci.nfkatzke.com/Practical/Practical.zip**
4. Put all the data in your data folder, and do not commit the data folder on github.
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.

1 Question 1: 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 (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 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.

2 Question 2: Billionaires

> Note: when creating a map_df function to read in all the csv's by simply using read_csv, you will notice an error.
>
> Read the error carefully, and reason how you would solve it. Tip: you cannot simply use read_csv, but need to create a reading function to solve the error.

You have been tasked by Fin24 to create two figures that bring out some interesting facts from the recently released Forbes Billionaires list. Give a short description of both of your figures using a non-formal markdown template (i.e. not Texevier).

In your figures, you should at least have some plot displaying net-worth numbers (this might require some wrangling to make the column numeric). This is your short column - marks will be allocated for creativity in your plots. You are free to use your discretion to complete this assignment.

3 Question 3: Tweets

Over the past decade, social media has become an integral part of how we interact with others as well as how we consume information. People are today more inclined to consume media through platforms like Twitter.

You have been tasked to write a short formal report (using Texevier) to discuss how different media outlets use Twitter. To this end, you have been supplied with all the tweets over the past decade from the BBC, CNN and The Economist.

In your report, give some clarity on the following:

- Compare the amount of photo and video content (percentage) used by the different news outlets through time (for the video column - 0 contains no video and 1 contains video, while the photo column contains an image if it is not an empty "[]"). On your plot

- Compare how many different hashtags each news outlet used through time.
- Compare how readers interact to each news outlet's tweets using any of the measures that you find useful in conveying this information.
- Track how Donald Trump held the spotlight, through time, for each of the different news outlets (using whatever column and / or transformation(s) you deem necessary).
- Using the Country_list.csv file - show how often the news outlets reference countries in their tweets. See if you can share some interesting insights using the country list sheet and the available tweets.

General Tips answering this question:

- Aggregate numbers to monthly where applicable.
- For a time-series plot, you need a date column. You might need to do a join with a representative date for each month...

END OF PAPER