23084103

Linda

2025-06-19

Purpose

This folder details the steps I followed to solve the exam questions. The code and individual plots can be found in the readme files within each solution folder, which are updated separately on GitHub. ## Question 1

I analyzed U.S. baby naming trends from 1910 to 2014 using Social Security data, enriched with Billboard, movie, and TV datasets to assess cultural influences. I examined the stability of popular names over time using Spearman rank correlations and identified growing volatility in name persistence, especially post-1990. I highlighted historical surges in names like John and Mary and linked modern spikes (e.g., Kanye, Arya) to pop culture events. Time-series and bubble plots were used to visualize these trends. The analysis shows a shift from traditional to media-influenced naming.

Question 2

To explore the musical progression of Coldplay and Metallica, Spotify data focusing solely on studio recordings (excluding live tracks) was analyzed. Key musical attributes such as tempo, danceability, and energy were extracted for each band. A boxplot comparing their tempos revealed that Coldplay tends to favor moderate tempos, while Metallica leans toward faster, more energetic rhythms. A scatter plot examining tempo versus danceability further showed that Metallica generally achieves higher danceability scores. These visual tools emphasize the contrasting musical styles and audience engagement approaches of the two bands.

Question 3

In this question, I began by reading the dataset into R using the read.csv() function to load the movies infor data set and readRDS to load the tittles and credits data sets. I then used the dfSummary() function from the summarytools package to inspect the structure of each variable, identify missing values, and understand the frequency of categories. This helped confirm that the datasets iare mostly clean and ready for analysis. Additionally, I created bar plots and boxplots to visualize the distribution of key variables. I also generated frequency tables to identify the most popular actors and directors based on the number of appearances.

Question 4

In this question I created a bespoke R function to efficiently load the billionaire dataset by leveraging the detailed column type information provided in a separate Excel file. This function reads the Excel file to understand the data types for each column, then uses this metadata to explicitly specify column types while reading the CSV file. By doing so, this ensured that each column is correctly interpreted (e.g., integers, strings, dates), preventing common data import issues. After defining the function, I called it with the appropriate file paths to import the dataset into R. Finally, I verified the successful import by inspecting

the dataset's column names and structure, confirming that the data matched the expected format.I first visualised the data to se interesting trends about the billionaires across the world then went on to do the analysis using different plots.

Question 5

This slide pack explores key lifestyle factors influencing good health, based on participant-level data. We examine how stress levels, sleep quality, age, and physical activity relate to overall well-being. Using visual analysis and regression, we highlight consistent patterns between health outcomes and daily habits. While some statistical limitations exist, the trends offer valuable insights. Our goal is to inform simple, actionable steps for improving public health.

Conclusion

I genuinely tried my best, I know it does not look like it but I struggled in this exam. But I want to say thank you for the challenge this exam actually taught me a lot and the things you taught in class now finally made sense. Thank you.