DATASCI 306, Fall 2024, Final Group Project

Your group number and each team member names

Throughout this course, you've dedicated yourself to refining your analytical abilities using R programming language. These skills are highly coveted in today's job market!

Now, for the semester project, you'll apply your learning to craft a compelling `Data Story` that can enrich your portfolio and impress prospective employers. Collaborating with a team (up to 5 members of your choosing), you'll construct a Data Story akin to the example provided here: https://ourworldindata.org/un-population-2024-revision

Data is already in the `data` folder. This data is downloaded from: https://population.un.org/wpp/Download/Standard/MostUsed/

You'll conduct Exploratory Data Analysis (EDA) on the provided data. The provided article already includes 6 diagrams. Show either the line or the map option for these 6 charts. You may ignore the table view. I'm also interested in seeing how each team will expand upon the initial analysis and generate additional 12 insightful charts that includes US and any other region or country that the author did not show. For e.g., one question you may want to answer is; US population is expected to increase to 421 million by 2100. You may want to show how the fertility rate and migration may be contributing to this increase in population.

Deliverable

1. **Requirement-1 (2 pt)** Import the data given in the .xlxs file into two separate dataframes;

   - one dataframe to show data from the `Estimates` tab
   - one dataframe to show data from the `Medium variant` tab

Hint: Some of the steps you may take while importing include:

- skip the first several comment lines in the spread sheet
- Importing the data as text first and then converting the relevant columns to different datatypes in step 2 below.

```
library(readxl)
library(tidyr)

estimates_table <- read_excel("data/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT.xlsx", sheet = "Esti
medium_variant_table <- read_excel("data/WPP2024_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT.xlsx", sheet =
```

2. **Requirement-2 (5 pt)**

You should show at least 5 steps you adopt to clean and/or transform the data. Your cleaning should include:

- Renaming column names to make it more readable; removing space, making it lowercase or completely giving a different short name; all are acceptable.
- Removing rows that are irrelevant; look at rows that have Type value as 'Label/Separator'; are those rows required?
- Removing columns that are redundant; For e.g., variant column
- Converting text values to numeric on the columns that need this transformation

You could also remove the countries/regions that you are not interested in exploring in this step and re-save a smaller file in the same `data` folder, with a different name so that working with it becomes easier going

forward.

Explain your reasoning for each clean up step.

3. Requirement-3 (3 pt) Replicate the 6 diagrams shown in the article. Show only the '2024' projection values where ever you have both '2022' and '2024' displayed. Show only the diagrams that are shown on the webpage with default options.

4. Requirement-4 (12 pt)

Select United States related data, and any other country or region(s) of your choosing to perform EDA. Chart at least 12 additional diagrams that may show relationships like correlations, frequencies, trend charts, between various variables with plots of at least 3 different types (line, heatmap, pie, etc.). Every plot should have a title and the x/y axis should have legible labels without any label overlaps for full credit.

Summarize your interpretations after each chart.

5. Requirement-5 (2 pt) Having developed a strong understanding of your data, you'll now create a machine learning (ML) model to predict a specific metric. This involves selecting the most relevant variables from your dataset.

The UN's World Population Prospects provides a range of projected scenarios of population change. These rely on different assumptions in fertility, mortality and/or migration patterns to explore different demographic futures. Check this link for more info: https://population.un.org/wpp/DefinitionOfProjectionScenarios

You can choose to predict the same metric the UN provides (e.g., future population using fertility, mortality, and migration data). Compare your model's predictions to the UN's.

How significantly do your population projections diverge from those of the United Nations? Provide a comparison of the two. If you choose a different projection for which there is no UN data to compare with, then this comparison is not required.

6. Requirement-5 (1 pt)

Conclusion

Your analysis should conclude with a summary of key findings. I'm especially interested in any novel insights you uncover that go beyond the article's original conclusions.

7. Extra Credit (1 pt) Develop an interactive Shiny app to visualize your machine learning model's projections. The app must include at least one interactive widget (e.g., dropdown, radio buttons, text input) allowing users to select a variable value (such as country/region) and view the corresponding projections.

Submission

- You will upload the zip file containing finals.Rmd file and its PDF as a deliverable to Canvas. If you created a shiny app for predictions, you will add those files also to your zip file.
- You will present your findings by creating a video of a maximum 15 minutes duration, explaining the code and the workings of your project; all team members should explain their part in the project to receive credit. You will share the URL of the video on Canvas for us to evaluate. An ideal way to create this video would be to start a Zoom meeting, start recording, and then every member share their screen and explain their contribution.

It is not necessary to prepare slides (if you do it doesn't hurt) for the presentation. You may speak by showing the diagrams and/or code from your Posit project. Every team member should explain their part in the project along with the insights they derived by explaining the charts and summaries for full credit to each member.

Your project will be evaluated for clean code, meaningful/insightful EDA and predictions.

Note:

- Each plot must be accompanied by a summary that clarifies the rationale behind its creation and what insights the plot unveils. Every diagram should possess standalone significance, revealing something more compelling than the other charts
- After the deadline, instructors will select the top three outstanding analytics projects. The teams responsible for these exceptional analyses will have their video shared with the class

We will not accept submissions after the deadline; December 10th 4 pm