Introduction to Data Science Project report

Data, source and collection method

The data is about the Linguistic diversity index (LDI), which is used to measure the diversity of languages spoken in a country. The data is quantitative and continuous, ranging from 0 to 1.

The data was taken from the Wikipedia page https://en.wikipedia.org/wiki/Linguistic_diversity_index using web scraping.

The data from the 'SIL International (2017)' table in the Wikipedia page was saved into a .csv file.

As there were 232 countries listed in the dataset, I decided to only produce visualisations showing the top and bottom ten countries for linguistic diversity in bar graph form.

Ethical statement

Data

Representation and biases

Wikipedia is authored by volunteers, instead of trained experts, so there may be more issues with accuracy and reliability than expert-authored articles. The article could be edited by anyone who has access to Wikipedia.

There are potential biases in the data from SIL International as they are a Christian religious organisation whose research in languages is conducted to encourage engagement with religious scripture. The rankings in SIL International's data did not match UNESCO data displayed on the same Wikipedia page and there were more countries listed than UNESCO's findings on linguistic diversity. UNESCO, as a United Nations specialised agency may have higher credibility and less biased approaches and objectives than SIL International data. While this could partly be due to changes in linguistic diversity between 2009, when the UNESCO data was dated, and 2017, when the SIL International Data was dated; this difference could raise concerns about reliability of the SIL International data. Additionally, the stated date of the data on the Wikipedia page is 2017, meaning it may be outdated if there have been changes in the languages used in that location in the last 8 years.

Potential uses of the project

The project could be used in conversations about diversity, culture and conservation. It could be used as a starting point to learn more about the languages spoken in different countries around the world. With further details about the types of languages spoken, and how this relates to cultural diversity, the data presented could be used to raise awareness about what linguistic diversity is and why it is important.

Challenges

One challenge was getting a 404 response code when using the code "print(response.status_code)" to make a request to the web page. I approached this by looking up why this error would occur during web scrapping. I found a way to address this on Stack Overflow, by accepting headers:

```
url = 'https://en.wikipedia.org/wiki/Linguistic_diversity_index'
headers = {'Accept': 'text/html'}
response = requests.get(url, headers=headers)
print(response.status_code)
```

There were issues producing graphs from the .csv file because the Python Interpreter interpreted the '/' from the 'Country / region' column table as division. I looked up ways to resolve this. After attempting a few methods unsuccessfully, I found the below as a solution to rename the column to "Country _ region." This was saved in a new .csv file.

```
for col in df.columns:
    if '/' in col:
        new_col_name = col.replace('/', '_')
        df = df.rename(columns={col: new_col_name})
        break
```

Despite me not aiming to create .tsv file, the table was being displayed with '\t' being added to column names. I found this to make reading the table more difficult, so I decided to recreate the Jupyter notebook, as I could not find a solution for the columns to be displayed without the '\t' in the initial attempt.

Learnings and what I would do differently

I learnt what a .tsv file is and how it is displayed using Python. I also learnt how to use '.sort_values,' 'ascending' and '.head' to identify the top and bottom 10 countries for linguistic diversity.

What I would do differently:

- I would explore different types of visualisation and different types of analysis. For example, presenting the data on a map.
- I could present the Top and Bottom 10 countries on the same graph.
- I would potentially removed the 'Unnamed: 0' column in the .csv file, as it does not add value to understanding the data, when displaying the data in table.
- I would potentially look and comparing Linguistic Diversity Index data from different sources, such as UNESCO.

What went well

- I did not encounter issues with retrieving the table data from the webpage.
- I was able to calculate the top and bottom ten countries for linguistic diversity and display the results in graphs without issues.
- I could adjust the graphs to improve legibility, for example, rotating the x-axis labels and increasing the distance between parts of the graph and text.
- Customising the bar graph colours, axes names and titles of the bar graphs.