CS-2704 Data Analytics with Python

Project Proposal (March-29th)

1. **Title: The Interdependence of Income with Education Levels**

2. **Group members:**
   1) Abdelrahman Abdelsadek - 3764220
   2) Darin Thomson   - 3776723

3. **Dataset Selection**
   a. *Mention the dataset name:*
      - World Bank or Kaggle income/education datasets.
   b. *Provide the source URL or description.*
      - https://www.kaggle.com/datasets/theworldbank/education-statistics
      - https://datatopics.worldbank.org/education/
   c. *Explain why this dataset is relevant to your hypothesis.*
      - These datasets are directly relevant to our project hypothesis as it gives us ample amount of data regarding income levels, education statistics and the interdependence between education levels and income.

4. **GitHub Repository**
   https://github.com/Darin07/CS-2704-Project-AD-
   a. Dataset
   b. Code
   c. Proposal
   d. Report

5. **Hypothesis**
   - The hypothesis being tested is as follows:

* *Hypothesis: Higher education levels lead to or yields higher salaries.*

6. Plan for Testing the Hypothesis

* *Data cleaning steps*

    - We will first, gather all necessary data from the respective datasets regarding the 2 topics: income and education level.

    - We would then go ahead and clean the data by selecting only a limited amount of data (ex: data from 2016 to 2024, etc.)

    - Once our data is cleaned, we will start implementing it into our model and work towards testing our hypothesis.

* *Descriptive statistics (mean, median, correlation, etc.)*

    - A descriptive statistic for our hypothesis would be to find the direct correlation between the 2 metrics: income and education level.

* *Data visualization techniques (scatter plots, heatmaps, etc.)*

    - Scatter plot (from the matplotlib library)
    - Box plot
    - Histograms
    - Bar chart
    - Heatmaps

* *Machine learning model (if applicable)*

    - Didn't think of implementing this into our model yet.

7. Expected Outcomes

* *What do you expect to find?*

    - We expect to find that there is a direct correlation between income and education levels, however this is only till a certain limit and not above that.

    - Ex: A graduate from a PhD would earn higher than an undergraduate. But a person who has multiple PhDs is not guaranteed to earn higher than a person with one PhD degree.

* *What conclusions might support or reject your hypothesis?*

- On observation of data, we expect to support the hypothesis as to higher the education level would result in a higher income.