**DSC 540**

**DATA PREPARATION**

**BELLEVUE UNIVERSITY**

**WEEK 11 & 12: PROJECT MILESTONE [5]**

**Milestone 5 Objective:** Now that you have cleaned and transformed your 3 datasets, you need to load them into a database. You can choose what kind of database (SQLLite or MySQL, Postgre SQL are all free options). You will want to load each dataset into SQL Lite as an individual table and then you must join the datasets together in Python into 1 dataset. Once all the data is merged together in your database, create 5 visualizations that demonstrate the data you have cleansed. You should have at least 2 visualizations that have data from more than one source (meaning, if you have 3 tables, you must have visualizations that span across 2 of the tables – you are also welcome to use your consolidated dataset that you created in the previous step, if you do that, you have met this requirement).

For the visualization portion of the project, you are welcome to use a python library like Matplotlib, Seaborn, or an R package ggPlot2, Plotly, or Tableau/PowerBI.

PowerBI is a free tool that could be used – Tableau only has a free web author. If your use Tableau/PowerBI you need to submit a PDF with your assignment vs the Tableau/PowerBI file.

**Some great Learnings from (Milestone - 5) exercise.**

The Data Wrangling course provided me with valuable learning experiences. I gained a deeper understanding of the significance of data preparation, which is often the most time-consuming aspect of any Data Science project. This course has significantly boosted my confidence in handling data clean-up and formulation tasks. While I had some prior experience with pandas and Python language at work and in the DSC530 (Exploratory Data Analysis) course, this learning opportunity allowed me to dive much deeper into pandas, as well as numpy and matplotlib.

The course has covered essential topics such as Fuzzy Matching, Hierarchical Index, Time Series Data analysis, and web scraping, which proved to be valuable tools in my skill set. The textbooks associated with the course were particularly helpful. Additionally, the weekly posts on Microsoft Teams channel from various colleagues played a crucial role in keeping the learning process engaging and dynamic. I made a conscious effort to take notes on intriguing topics, further enhancing my understanding of the material. Regarding project, I chose public datasets. Using file datasets and API datasets was easy part. I found web scraping little tricky as it relies on html DOM (elements and stylesheet names). I felt chances of error are more and hence need careful look at websites to find any discrepancy. Data cleaning, adding column names, standardizing names etc. tasks were really interesting but easy.

The visualizations created with Seaborn and Matplotlib libraries were exceptionally well-crafted, providing me with valuable insights and learning opportunities. The process of selecting the appropriate visualization has always posed a challenge for me, but I am confident that with experience, this skill will improve. Customizing visualizations requires a substantial amount of learning, and I found it to be a significant aspect of the overall learning experience.

One specific challenge I faced was positioning the labels correctly on a Pie chart, which proved to be a bit challenging. Nevertheless, relying on Google as a constant resource proved to be immensely helpful, reinforcing the notion that it remains a reliable companion in navigating and troubleshooting various aspects of data visualization.

**Ethical implications of data wrangling specific to the data source and the steps completed (Milestone - 5)**

While choosing the datasets we should be mindful of not promoting any hate speech, discrimination, or any form of harmful or offensive content. So, it is very important to create a website that contribute positively to the online community and promote inclusivity, diversity, and respect. Depending on the circumstances and situation, the data source, and the particular methodology taken, data wrangling may have different ethical ramifications. The type of data and the analysis's objectives determine the particular ethical ramifications of data wrangling. Maintaining the confidence and respect of the people and communities from whom the data is derived is just as important as adhering to rules and regulations when it comes to ethical data wrangling. Responsible data science includes ethical data management and analysis as a fundamental component. The following are few challenges related to data wrangling:

* **Data Privacy:** When engaging with datasets, it is crucial to verify that we possess the requisite permissions and consent to utilize the datasets. This becomes especially pivotal when handling personal or sensitive information. Ethical considerations come into play if data is acquired, disseminated, or employed without the awareness and approval of the individuals concerned. Furthermore, the responsible management of sensitive or private data, along with adherence to privacy regulations such as GDPR, is of utmost importance. Implementing proper anonymization and data protection techniques is imperative in maintaining ethical standards and legal compliance.
* **Data Quality:** One of the main challenges lies in confirming the accuracy, completeness, and reliability of the dataset. The processes of cleaning, validating, and curating data can be labor-intensive and time-consuming.
* **Data Security:** Security ranks as a paramount consideration for datasets, encompassing websites and APIs. Safeguarding against unauthorized access, data breaches, and potential threats like SQL injections or cross-site scripting demands robust authentication and authorization mechanisms. Given that data wrangling frequently involves tasks such as data transfer, storage, and sharing, it becomes imperative to implement measures ensuring protection against unauthorized access, breaches, or leaks. The repercussions of data security breaches extend to serious ethical and legal implications.
* **Data Anonymization:** If the dataset comprises personal or sensitive information, it is imperative to anonymize or de-identify the data to safeguard the privacy of individuals. Neglecting this precautionary step could lead to ethical and legal ramifications.
* **Bias and Fairness:** Data wrangling can also introduce bias, where actions such as data cleaning or sampling methods may unintentionally favor specific groups or perspectives. It is essential to recognize and rectify bias to guarantee fairness in data analysis and decision-making.