**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.

**Project Description**

Many travelers now have access to fantastic, straightforward, and practical lodging from Airbnb. Similar to that, it has provided several people with a means of generating additional income by marketing their homes as places for travelers to stay. To complete the project milestone, I have selected 3 data sources named as (**Listings**, **Calendar**, and **Reviews** ) to query and design the Airbnb API and websites to easily provide the data to end users and facilitate easy booking and querying from UI.

**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 and Consent:** When working with any datastes, it's important to ensure that you have the necessary permissions and consents to use the data. This is particularly critical when dealing with personal or sensitive information. Ethical concerns may arise if data is collected, shared, or used without the knowledge and consent of the individuals involved. Additionally managing sensitive or private data and adhering to privacy regulations (e.g., GDPR) is crucial. Anonymization and data protection techniques must be applied correctly
* **Data Quality:** Ensuring that the dataset is accurate, complete, and reliable is one of the primary challenges. Cleaning, validating, and curating data can be time-consuming
* **Data Security:** Security is a top concern for datasets including websites and APIs. Protecting against unauthorized access, data breaches, and attacks such as SQL injections or cross-site scripting requires robust authentication and authorization mechanisms. Data wrangling often involves data transfer, storage, and sharing. It's crucial to take measures to protect the data from unauthorized access, breaches, or leaks. Data security breaches can have serious ethical and legal implications.
* **Data Anonymization and De-identification:** If the dataset contains personal or sensitive information, it's essential to anonymize or de-identify the data to protect the privacy of individuals. Failure to do so could result in ethical and legal issues.
* **Bias and Fairness:** Data wrangling can introduce bias if not done carefully. For example, data cleaning or sampling methods may inadvertently favor certain groups or perspectives. It's crucial to be aware of and address bias to ensure fairness in data analysis and decision-making.