**DSC 540**

**DATA PREPARATION**

**BELLEVUE UNIVERSITY**

**WEEK 9 & 10: PROJECT MILESTONE 4**

**Project Objective:** Perform at least 5 data transformation and/or cleansing steps to your API data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, they will make completely different transformations. You could do the same transformation multiple times if you needed to clean your data. The goal is to clean a dataset at the end of the milestone.

* Replace Headers
* Format data into a more readable format
* Identify outliers and bad data
* Find duplicates
* Fix casing or inconsistent values
* Conduct Fuzzy Matching

Make sure you clearly label each transformation (Step #1, Step #2, etc.) in your code and

describe what it is doing in 1-2 sentences

**Project Description**

**Ethical implications of data wrangling specific to the data source and the steps completed (Milestone - 4)**

While designing an API and working with data we should be mindful of the data we work. The Json data should not contain any hate speech, discrimination, or any form of harmful or offensive wordings. It is very important to use API 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 data, 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 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.
* Scalability: APIs should handle increasing loads and user traffic efficiently. Designing for scalability often involves load balancing, caching, and considering potential bottlenecks in the system.