**Central Information System – Zekelman School of IT**

**(Data Analytics for Business)**

**Dataset**

The purpose of this dataset is to gather information related to the experiences, needs, and issues of students, and it was collected via an online survey. However, since the data is self-reported, there is a potential for response bias. The survey was divided into three parts: demographics, the convenience of living, and employment and scholastics. By analyzing this dataset, it is possible to gain valuable insights into the student experience, particularly in terms of their needs and issues.

However, it is important to be aware of potential biases and limitations in the survey method. Additionally, careful data cleaning and processing will be necessary to ensure the accuracy and usefulness of the data for analysis.

**Dataset assessment**

Initial data is pushed into a Raw\_data table which acts as a landing zone for Survey data sheets, next the data is pushed into a Base\_Survey\_data table where column names are given along with correct Data types. The dataset is further broken down into seven different tables, each with a unique ID to identify and join them using Google big query.

* **Academic** has columns for intake and current semester to track student progress.
* **Demographics** has 10 columns for gathering demographic information to understand the student population.
* **Accommodation** has three columns related to student housing situation for potential areas of improvement.
* **Accommodation feedback** has five columns for scoring living and service quality.
* **Academic feedback** has five columns for students to rate their course experience.
* **Services** focuses on college services, gathering information on awareness, usage, and satisfaction for identifying areas of improvement.
* **Feedback** consists of data in which students can give their feedback related to their experience" means that there is a specific table within the dataset that is designed to gather feedback from students about their experiences. This feedback can then be analyzed and used to identify areas for improvement and to make data-driven decisions about how to improve the student experience.

**Data Collection and Pre-Processing**

**Data Preprocessing**

Raw Survey data collected is stored in an Excel online file. Using Microsoft Power Automate, the data is connected to Google Drive and then transferred to Google Cloud (Views in Big Query). Data is then segregated and stored into multiple tables using queries. These tables can be used for data preprocessing which involves transforming the raw data into a usable format for analysis. The processed data is fetched from Database to Google Colaboratory for analysis.

**Data Cleaning:** The following steps can be taken for data cleaning:

* Handling missing values
* Outlier detection and treatment
* Shortening/Transforming long values.
* Separating semi-colon separated values.
* Cleaning text field data
* Encoding categorical variables

**Data Integration:** Our data is partitioned into multiple tables, and there is a need for integration of relevant data from multiple tables to conduct bivariate analysis for observing variable correlations and machine learning for satisfaction score prediction.

**Data Transformation:** To analyze data accurately, it must be transformed into a format that is suitable for analysis. This involves techniques such as normalization, standardization, and scaling. Normalization ensures that all data values fall within a specific range, while standardization scales data to have a mean of zero and a standard deviation of one.

**Data Reduction**: As our dataset is not large enough then data reduction may not be necessary or may be less important. However, if the data contains many features or variables, then data reduction techniques can still be useful for simplifying analysis and improving efficiency. Some data reduction techniques that can be used in this case include:

* Feature selection
* Feature engineering
* Clustering

Overall, data reduction techniques can be useful even for smaller datasets, particularly if the data contains many features or variables. However, the specific technique used will depend on the nature of the data and the goals of the analysis.

**Data Encoding:** To analyze our data, we need to transform categorical variables into numerical format using encoding techniques.

**Ethical principles**

There are several ethical concerns related to collecting and analyzing student experience survey data. Some of these concerns include:

**Informed consent:** It is important to obtain informed consent from participants before collecting their data. This includes providing clear information about the purpose of the survey, how the data will be used, and any potential risks or benefits of participating.

**Confidentiality:** Participants have the right to privacy and confidentiality. It is important to ensure that their personal information is kept secure and that their responses are anonymized.

**Bias:** Surveys can be biased in their design or administration, which can result in inaccurate or misleading data. It is important to ensure that the survey is designed in a way that is neutral and unbiased, and that the questions are clear and unambiguous.

**Data usage:** The data collected should only be used for the purpose for which it was collected, and any additional usage should be clearly communicated to participants and be within the bounds of their original consent.

**Data ownership:** Participants should be informed about who will own the data that is collected and how it will be used.

**Data security:** Surveys can collect sensitive information such as demographic information or personal experiences. It is important to ensure that the data is stored securely and that appropriate measures are in place to prevent unauthorized access or disclosure.

**Method and Solution**

Google Big Query will act as the Database (views) for the data to be stored, there will be connections between Google Drive (act as passage for data), Google Colab (for Python and ML work) and Tableau Visualization. Flow of data is single or both ways depending upon the use and phase in progress.

A picture containing chart

Description automatically generated

Diagram

Description automatically generated

**Future data needs**

**Second phase of Survey**: More elaborated data related to College Services and Student’s issues. This data will be needed for further analysis on the issues and services that can be related.

**Code**

* <https://github.com/JayrajRadadiya/Capstone-Project/tree/main/Code>
* <https://github.com/JayrajRadadiya/Capstone-Project/blob/main/DesignDocuments/LatestDb_Design.xlsx>

**References**

* **Tableau**: [Tableau Public](file:///E:\04%20-%20Semester4\01%20-%20capstone\Github\Capstone-Project\Assessment%201\Tableau%20Public), <https://www.tableau.com/learn>, [Datacamp](https://app.datacamp.com/learn/skill-tracks/tableau-fundamentals)
* **Python**: [Data science life hacks (linkedin.com)](file:///E:\04%20-%20Semester4\01%20-%20capstone\Github\Capstone-Project\Assessment%201\Data%20science%20life%20hacks%20(linkedin.com)), <https://keras.io>, <https://numpy.org>, [NLP - Datacamp](https://www.datacamp.com/tracks/natural-language-processing-in-python?)
* **Google Big Query:** <https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv>
* **SQL: W3 schools**
* **Power Automate:** <https://learn.microsoft.com/en-us/power-apps/>