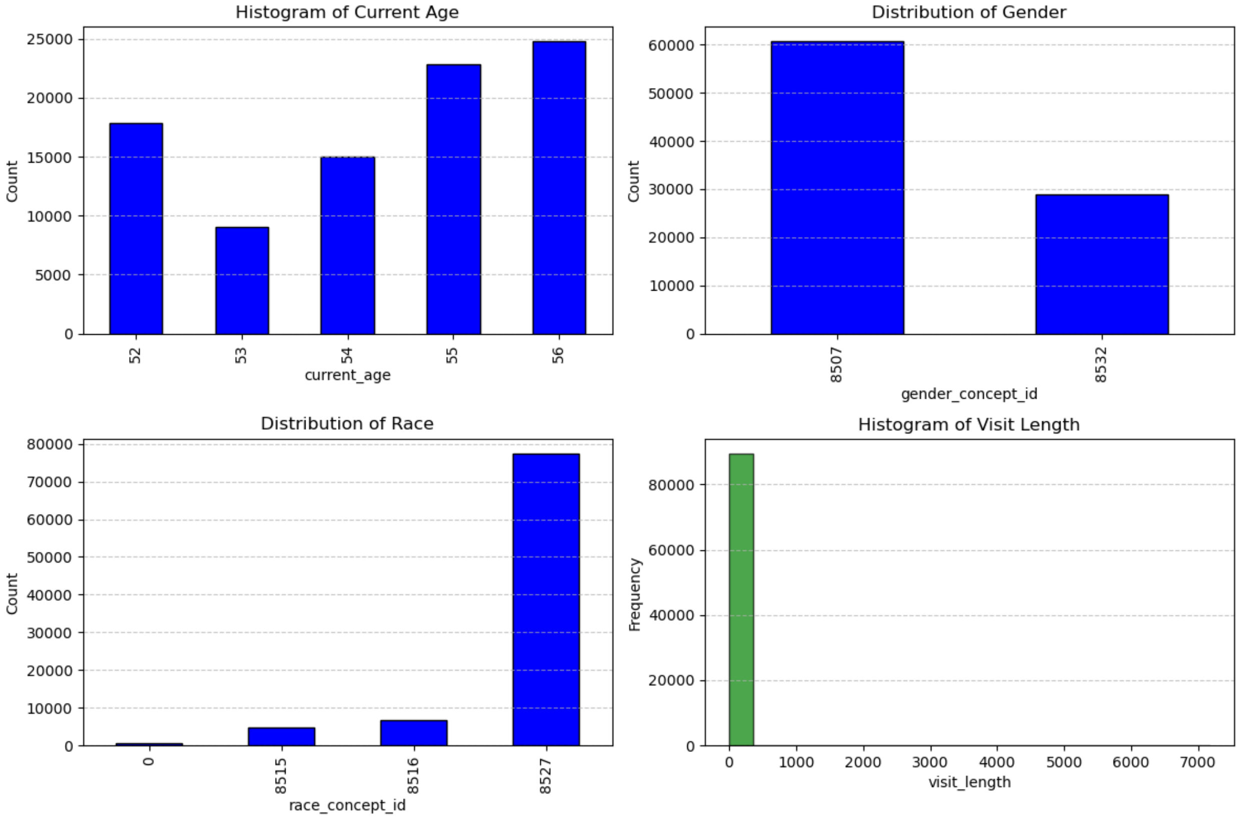
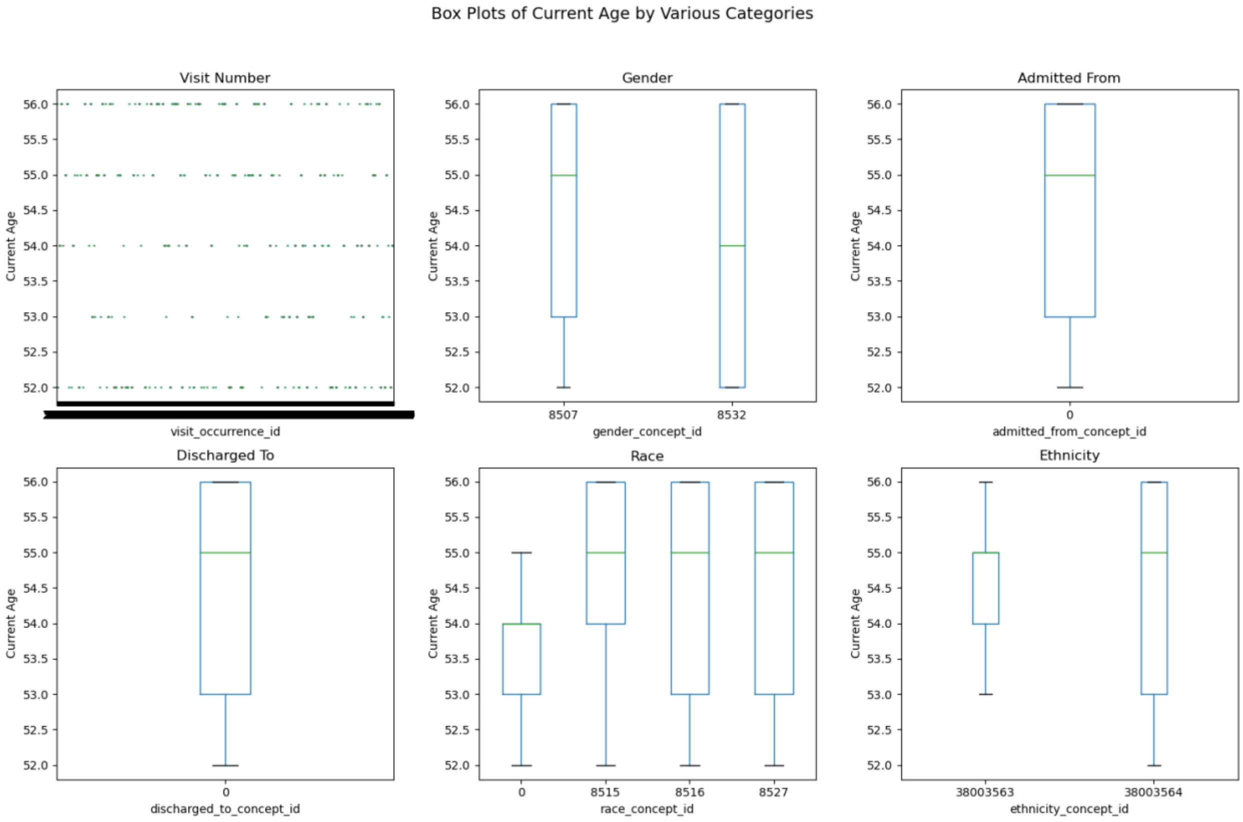
Programming Assignment 1

Descriptive Statistics Visualization

1. Bar plot and histogram
   From the age distribution, we can observe that the numbers of different ages are relatively balanced. From the gender distribution, we can observe an apparent difference between the two genders. The number of males is greatly larger than females. The majority of patients belong to the race concept ID 8527, while other racial categories have significantly lower representation. This indicates an imbalance in racial diversity in the dataset. The histogram for visit length suggests that most visits are of short duration, but there may be some outliers with exceptionally long visits.



2. Box plots
   The box plot for visit number shows age values spread evenly across visits, but the visualization is dense, suggesting high visit frequency for individual patients. From the age-gender plot, we observe that both genders shares a similar age distribution. Also, patients admitted from and discharged to the same place (ID = 0) have a wider spread in ages, meaning both younger and older individuals experience similar hospitalization patterns. Lastly, the racical and ethnic groups show the similar age distribution.



Question Answering

1. The concept table in the OMOP schema serves as a standardized vocabulary that maps medical terms across datasets using unique `concept_id`s, which are referenced in multiple tables, including `measurement` and `condition_occurrence`. The person table contains demographic details such as `person_id`, `gender_concept_id`, `race_concept_id`, and `ethnicity_concept_id`, acting as the central entity that links all other tables through `person_id`. The measurement table records clinical and laboratory results, storing values such as `measurement_concept_id`, `value_as_number`, and `unit_concept_id`, with each measurement tied to a specific patient via `person_id` and linked to a specific visit through `visit_occurrence_id`.

2. A feasible question we can explore is: Can we predict which patients are at a higher risk of prolonged hospital stays based on their demographics and biomarker measurements? This problem can be framed as a binary classification task, where the target variable is whether a patient's visit length exceeds a predefined threshold (e.g. this may be a hyperparameter or a threshold that fulfill a certain criteria). The features we can use are such as age, gender, race, ethnicity, hypertension/diabetes status, and various biomarkers. And for this binary classification task, some traditional ML methods will be easy-to-implement and efficient for example ramdom forest classfier.