# Section A

## Part 1: Research Question

The dataset associated with the project contains observations about patient demographics, how long they stayed in the hospital, and the charges they received for their stay. As the projects data analyst, it is my job to answer the following research question: What demographic factors influence the additional charges that are billed during a patient’s hospital stay?

This information could be of practical use in order to allocate government funds to areas that display higher proportions of the features that drive additional costs.

## Part 2: Goals

The goals of the project are to determine which factors included in the dataset contribute to the additional charges that a patient receives, and to quantify the relationship. In order to meet these goals, the dataset will be prepared for statistical analysis. In particular, linear regression models will be used to analyze the data.

# Section B

## Part 1: Assumptions of Linear Regression

Four assumptions to be aware of when using a linear regression model are:

* Predictive purpose: Aside from the case of time series forecasting, linear regression models should not be used to extrapolate outside of the range of data.
* Multicollinearity: Multicollinearity can lead to dubious regression coefficients. As such, predictor variables should be removed from the model until multicollinearity is gone.
* Outliers: Outliers can distort the results of regression modeling. For this reason, predictor variables should have normal distributions.
* Heteroskedasticity: Heteroskedasticity (the lack of constant residual variance across the range of predicted values) may suggest that the model is incomplete and require more predictor variables (Bruce, Bruce, & Gedeck, 2019).

## Part 2: Python for Data Analysis

Python is a robust, widely-used programming language that has several libraries and packages for data analytics. For linear regression in particular, Python is useful because 1) the availability of packages used for data preprocessing and exploratory data analysis (pandas and scipy) and 2) the availability of packages used for building linear regression models (scikit-learn and statsmodels).

## Part 3: Justification of Methods

Linear regression quantifies the nature of the relationship between one or more predictor variables and the target variable (Bruce, Bruce, & Gedeck, 2019). As defined in Section A, Part 1 the research question seeks to understand what patient demographic factors (the predictor variables) influence additional charges (the target variable). Based on this, we can see that linear regression is an appropriate technique for analyzing the research question.