

**DSC421 FUNDAMENTALS OF DATA SCIENCE**

**PROJECT PROPOSAL**

**Submitted to: Prof. Ustun Ilyas**

**Submitted by: Salik Hussaini, Peter Gillespie, Nick Edington & Zoya Hafeez**

**Group 11**

**1 DOCUMENT PURPOSE:**

The project scope statement is a document that describes the major deliverables of a project including the key milestones, high-level requirements, assumptions, and risks. It also defines the boundaries of a given project and clarifies what deliverables are in and out of scope.

1. **PROJECT SCOPE:**

Following has been identified as the project scope, “Predictive analysis of healthcare cost to help reduce debt accumulation from medical providers”.

* 1. **PROBLEM STATEMENT:**

United is notorious for high healthcare costs, many politicians preach that the healthcare system is broken. With these ideas in mind, our group will focus on solving why is healthcare so expensive? What is actually going wrong? Our data from the US. Department of Health and Human Services provides common inpatient and outpatient services, all physician and other supplier’s procedures and services, and also Part d prescription. With this immense data source, we hope to apply several machine learning algorithms to combat high healthcare costs.

* 1. **IMPORTANCE:**

Healthcare is important as age increases. Disease burden exponentially takes a toll on an individual’s life. Visiting a healthcare provider also increases, so in result debt also increases. What we aim to do is reduce billing and unneeded ICD-10 codes applied to a patient’s visit. The benefit of our analysis is a reduction of the debt of a patient's hospital visit.

1. **DATA SOURCE:**

The Centers for Medicare & Medicaid Services (CMS), is a federal agency that provides standards for all healthcare-related programs (HIPPA & CHIPS). Our data is acquired from a single source that used is used by CMS (<https://www.cms.gov>).

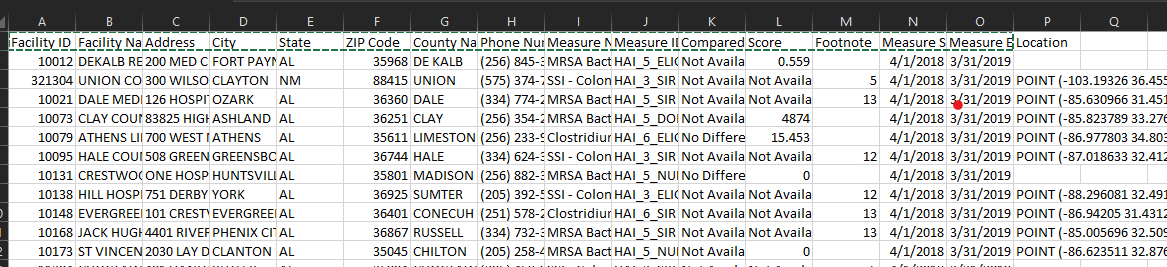
****

Fig 1: Snapshot of data set

**3.1 TOOLS:**

For data analysis, we have identified three tools, R, SQL and Excel. R is a powerful tool for performing the statistical analysis, we will be using it for a) data analysis and b) exploratory work (creating statistical models). SQL, is a SQL is a standard language that is used for accessing and manipulating databases. In our project, we will be using it for data reading, writing, updating and retrieving the data. Also, we will be using excel to merge the data.

1. **SCOPE FOR ANALYSIS:**

**4.1 RESEARCH FOCUS:**

The focus of our research will be around two outstanding questions;

1. If I were to go to a hospital which provider would bill me the least?
2. If I had a medical condition which city will bill me the least.
   1. **TECHNIQUES:**

Most of our data is in numerical, categorical, geographical and demographical form. Linear regression, PCA and Hypotheses testing have been identified as some of the techniques that we will use in our project.

* 1. **CHALLENGES:**
     + Having an R-squared that is low,
       - however, is still able to adequately predict future medical services.
     + Multicollinearity
     + Assumption not met for regression model.