# **Project Proposal**

### **Motivation and problem statement:**

The problem statement for the project is a very simple question **Which Factors Influence the Price of Health Insurance?** The reason for the problem is that almost every person in the United States has health insurance and based on their selected plan, they get charged for it. With this problem, I want to give them a data-oriented review of how these different factors can increase or decrease their charges. If people know the reasons, they can optimize their routines to save money. Also, with exploratory analysis, I want to check whether any kind of bias exists in the different categories of features in data.

The project is a way for me to prepare myself to explore unknown and faulty datasets. It will be great to learn about statistical techniques that work with different kinds of variables, conditions where techniques might not work, and how to deal with them.

### **Data Selected for Analysis:**

For the analysis, I will be using a publicly available dataset from [Kaggle](https://www.kaggle.com/mirichoi0218/insurance). The dataset consists of individual records and what were they billed by health insurance companies. The dataset does not contain any PII data. Each individual is described by the following features:

1. **age**: age of the primary beneficiary
2. **sex**: insurance contractor gender, female, male
3. **bmi**: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, objective index of body weight using the ratio of height to weight, ideally 18.5 to 24.9
4. **children**: Number of children covered by health insurance / Number of dependents
5. **smoker**: Smoking (Yes/No)
6. **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
7. **charges**: Individual medical costs billed by health insurance

These features represent some of the most common details every customer is required to fill while applying for medical insurance. Providing an analysis using these features allows an easier interpretation for everyone (especially people who do not know about the small details of insurances). The description of data on Kaggle reveals that we have a similar number of male and female individuals, an almost equal number of individuals for each age (18-20 are slightly more than the rest) and there are more individual with 0 or 1 children when compared to individuals with greater than 2 children.

License:

[Database: Open Database, Contents: Database Contents](http://opendatacommons.org/licenses/dbcl/1.0/)

### **Unknowns and dependencies:**

* Although in the real world these features are not enough to adequately describe the health insurance bills, we can start with these limited columns and then gradually improve our analysis as we get new data
* The data is a small random subset of the whole population and inferences from this data cannot be generalized to the whole population