# PHASE-1 REPORT – Harsh Jivani

# Business Understanding

My data analytics project topic is *Predicting* *Personal Healthcare Costs influenced by various health factors*. We are a startup healthcare solutions provider where we audit healthcare insurance companies across the United States to make sure they are not over charging their customers so that customers will not leave the insurance company.

My motivation in selecting this project is to tell a story on how and why a person should be in healthy shape to avoid increase of medical costs. Within past decade, I have seen many companies investing in smart health technologies to help patients stay healthy and avoid high medical premiums.

The target variable for this dataset is *charges*. The target variable, charges, is a continuous number and the data type is numerical (float). The variable *charges* represent money in this dataset and all data for this variable is calculated in USD or United States Dollar amount.

## Business Problem

The business problem is that my company is in charge to audit a healthcare insurance companies who thinks they are losing customers at rapid rate and the business would like to find out if they are leaving because they overcharged them for their medical needs. My company will accurately predict medical costs for current and future members with similar characteristics and find out if the healthcare insurance company is under/over-charging members based on healthy/unhealthy lifestyle. I believe factors such as BMI, age, and someone with smoking history will affect whether they will pay higher medical costs than someone who meets the standard healthy lifestyle will pay lower medical costs.

## Dataset

The dataset I will be using has 7 attributes including 1 target attribute. The dataset comes from Kaggle.com and the data is derived from demographic statistics from US Census Bureau. Below, I have provided with the first few rows of the sample dataset for this project. The original dataset contains 1,338 rows; therefore, I have not linked it here.

The dataset attributes are described as:

* Age: The age of the person.
* Sex: The gender (male or female) of the person.
* BMI: Body Mass Index, is a measure of body size. It uses person’s weight and height to calculate the BMI. The result we get tells us whether the person is underweight if result is < 18.5, healthy weight if result is 18.5 – 24.9, overweight if result is 25.0 – 29.9, or obesity if result is 30 or higher.
* Children: Number of dependents that this person has.
* Smoker: If person smokes then yes or does not smoke then no.
* Region: This is residential area where the person resides. For example: Northwest/Southeast/Southwest, etc.
* Charges: This is the medical costs that is billed by the insurance providers.

First 4 instances:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| age | sex | bmi | children | smoker | region | charges |
| 19 | female | 27.9 | 0 | yes | southwest | 16884.924 |
| 18 | male | 33.77 | 1 | no | southeast | 1725.5523 |
| 28 | male | 33 | 3 | no | southeast | 4449.462 |
| 33 | male | 22.705 | 0 | no | northwest | 21984.4706 |
| 32 | male | 28.88 | 0 | no | northwest | 3866.8552 |

Source to dataset: https://www.kaggle.com/mirichoi0218/insurance

## Proposed Analytics Solution

The following is a potential proposed analytics solution:

* Medical Costs prediction
  + Predict and model medical costs based on age, bmi, smoking.
  + Use the model created to test the results with my family members to see the past and future charges make sense.
  + By using the models, we can use the predicted medical costs and find out how much money the patient would have saved.