# Project Proposal: Predicting healthcare insurance costs in the United States


| Name | Email | Contributions |
|---|---|---|
| Aditi Goyal | adigoyal@ucdavis.edu | Exploratory data analysis ,data visualization, data transformation,and web app. |
| Brandon Hom | bwhom@ucdavis.edu | Regression analysis, model tuning, interpretation of regression results,and web app. |
| Tammie Tam | tastam@ucdavis.edu | Regression diagnostics (i.e residual plots), conclusion, web app, and interpretation of regression diagnostics |


## Introduction/Background

I suck at writing this part; someone carry me.

## Dataset Description

The data set on healthcare insurance costs in the United States was obtained from kaggle: https://www.kaggle.com/mirichoi0218/insurance. The dimensions of the data set are 1338 by 7, meaning that we have 7 features within our data set. Six of the features may be of potential use to predict the target feature *charges*. Number of children, age and BMI are numerical variables. Region, smoker and sex are categorical variables.

## Key questions

- What numerical variables have the most impact on health insurance costs? For example, if increasing age contributes the most to health insurance costs, individuals should be wary about spending more on health insurance as they grow older.
- Does adding categorical variables to our model also influence the insurance costs? More specifically, is there a significant difference in insurance costs between males and females? Does being a smoker lead to higher insurance charges?

## Methodologies

The model that seems appropriate for this dataset is a linear regression model. We suspect that as age and BMI increase, the cost of insurance will also increase in a relatively linear fashion. To test this, we plan on using a multiple linear regression model, where the coefficients of the model will indicate the significance of that variable. For example, if we the coefficient for age and BMI is and 6 respectively, this informs us that BMI has a greater impact than age. Similarly, we plan on including the categorical variables sex and smoker to see whether those lead to significant increases in insurance cost.

To assess the appropriateness of a linear regression model, we will also perform diagnostics. Plots of the residuals vs fitted values and residuals vs predictor variables will be made to check for constant variance.

The normality assumption will be checked with a QQplot and histogram of the residuals.

Finally, there may also be multicollinearity present within our data. To address this problem, we plan on using forward and backward stepwise regression to tune our model by removing predictor variables that do not any significant change to the insurance costs.

# References