# **Group L2C5:**

# **Project Proposal: Predicting Insurance Charges**

1. The source of the data being used for the project:

[Prediction of Insurance Charges](https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender)

Wakefield, B. (2020). *Insurance* [Dataset]. <https://data.world/bob-wakefield/insurance>

This dataset contains detailed information about insurance customers. By analyzing this dataset, we can gain insights into customer behavior and the factors that contribute to insurance charges. This information can inform the development of insurance plans and marketing campaigns targeted at specific demographics. Additionally, we can explore solutions for addressing high insurance charges among certain groups to increase affordability.

2. The variables of this dataset on insurance customers is shown in the following table:

| **Variables** | **Units** | **Description** |
| --- | --- | --- |
| **Age** | years | The age of the customer. |
| **Sex** | male/ female | The gender of the customer. |
| **Bmi** | kg/m^2 | The body mass index of the customer. |
| **Children** | N/A | The number of children the customer has. |
| **Smoker** | false/ true | Whether or not the customer is a smoker. |
| **Region** | southeast/ southwest/ northwest/ northeast | The region the customer lives in. |
| **Charges** | dollar | The insurance charges for the customer. |

3. The research question, or other motivation, behind the analysis of the data.

The rising cost of healthcare has made insurance a crucial component of safeguarding individuals' financial and personal well-being. According to a study by the National Bureau of Economic Research, BMI, smoking status, age, and children are the primary factors contributing to the cost of healthcare premiums. Our project aims to predict insurance charges based on input variables such as age, gender, BMI, number of children, smoking status, and region of residence. Our goal is to develop a prediction model that accurately estimates insurance charges, enabling policyholders and insurers to manage risk and remain financially stable. We plan to use the forward selection method to compare the full model and reduced variables selected model and identify the "best" model based on the root mean squared error (RMSE). Accurate prediction of insurance charges based on these variables allows individuals to budget, make informed decisions about their insurance needs, and avoid financial hardship in the event of an unexpected illness, injury, or accident. Our study will contribute to this effort and help policyholders and insurance companies achieve financial stability.

Reference:<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9265373/>

4. An overview of "who will do what" on your project across your team members. This does not need to be in fine detail, but should broadly outline the responsibilities undertaken by each team member.

Haonan Su, Yitong Zhang:

Data collection and preprocessing:

* Importing the data set, cleaning, and splitting the data into training and testing sets.
* Checking linearity and other assumptions, such as normality (with Q-Q Plot), multicollinearity (with VIF), equal error variances (with a plot of residuals vs. fitted values), and the existence of outliers, etc.

Reporting:

* Writing the reports and references.

Project management:

* Coordinating team activities, ensuring timely completion of tasks, and facilitating communication among team members.

Xinru Jiang, Jun Ma:

Model development:

* Developing additive Linear models to predict insurance charges based on variables on the training set.
* Using forward selection to build the reduced model to compare with the full additive model to identify significant factors influencing insurance charges.

Model evaluation:

* Verifying the training model based on the testing set according to the RMSE and other statistics.
* Assessing and optimizing model performance for predicting insurance charges.

Visualization:

* Creating visual representations of data and analysis results to facilitate understanding and communication of insights.