Project Proposal

It is important to protect yourself and loved ones and one way is being covered by health insurance. Everyone needs health insurance, regardless age, gender, region, or even if you are smoker or not. Health insurance provides some benefits if you become sick, but also, depending on some factors. it can be very expensive.

A research and analysis about some factors that determine the cost of health insurance policy will be performed. A dataset with insurance prices will be used to perform this analysis.

There are 1338 rows and 7 columns. The columns include 4 categorical variables, 2 continuous variables and 1 discrete variable.

The dataset used for this project can be found in this link: [insurance.csv](https://drive.google.com/file/d/1MFPdCXda9EV1yRDlkknJoYKabcNhioZj/view?usp=sharing)

|  |  |  |
| --- | --- | --- |
| Name | Data Type | Description |
| age | Nominal | age of primary beneficiary |
| sex | Nominal | insurance contractor gender, female, male |
| bmi | Continuous | Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 |
| children | Discrete | Number of children covered by health insurance / Number of dependents |
| smoker | Nominal | Smoking |
| region | Nominal | the beneficiary's residential area in the US, northeast, southeast, southwest, northwest |
| charges | Continuous | Individual medical costs billed by health insurance |

A dataset will be used to analyze and for comparing smokers and no smokers’ prices of health insurance. Also, for comparing health insurance prices between female and male genders.

The analysis will help to answer the following questions:

* Are smokers’ insurance prices higher than no smoker’s insurance prices?
* Are insurance prices higher for male gender than female gender?

Two hypotheses have been defined for the analysis:

1.- Null hypothesis: *H*0​: *u1-u2=0*  
There is not difference in insurance prices between non-smokers and smokers

Alternative hypothesis: *Ha*​: *μ* ≠ *μ*0

There is difference in insurance prices between non-smokers and smokers

2.- Null hypothesis: *H*0​: *u1-u2=0*

There is not difference in insurance prices for the male gender and insurance prices for the female gender

Alternative hypothesis: *Ha*​: *μ* ≠ *μ*0

There is difference in insurance prices for the male gender and insurance prices for the female gender

The smoker column, which is a categorical independent variable and the charges column, which is the dependent variable, will be used to perform a t-test to determine if insurance prices of smokers are higher than prices of non-smokers.

The sex column, which is an independent variable, and the charges variable will be used to perform a t-test to demonstrate if insurance prices for the male gender are higher than insurance prices for the female gender.

This analysis is important for insurance companies. It would be valuable information that can be used to project future prices for specific groups of people or regions. Also, it is valuable for individuals that want to buy a health insurance plan.