SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Samsung Innovation Campus

Artificial Intelligence Course

Insurance Data

Project presented by :

- Zyad Farag
- Hadeer Emad
- Abdelrahman Ehab

Team : Hunters “Eng.\Shaimaa Osman”
Data Used : https://www.kaggle.com/mirichoi0218/insurance

Insurance Data Explanation

What is the data about?

The Insurance data consists of 7 columns which are:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- Bmi: 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: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

Insurance Data Explanation

What is the target of the data analysis?

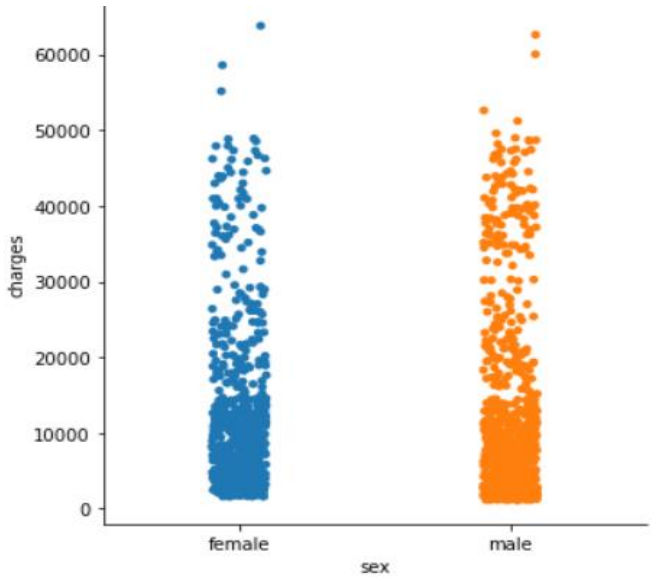
Our data analysis aims to minimize the insurance charge and predict the cost given certain features of the user like his Bmi,age,if he is a smoker or not…etc.

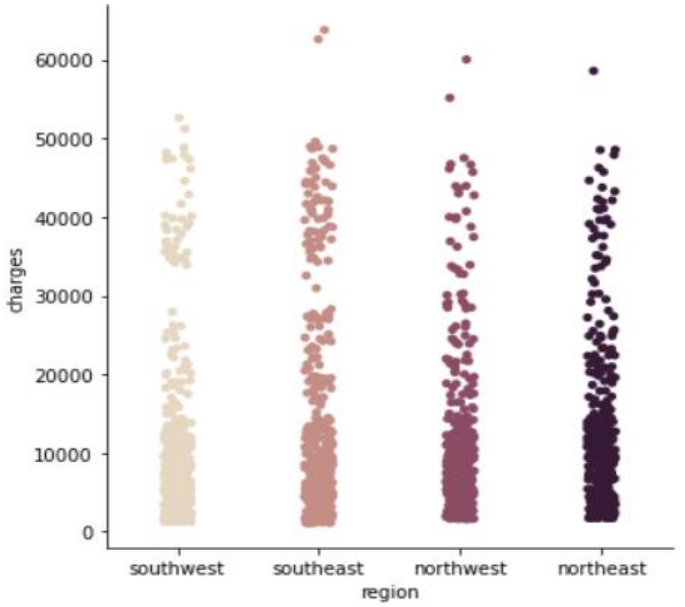
Those features are fairly effective over the medical insurance of the user as we will explain along the presentation.

Analyzing Insurance Data

Features not affecting the charge:

Not every feature in the data set can affect the cost of the insurance. For example, sex and region is irrelevant to the charge as shown in the following graphs:



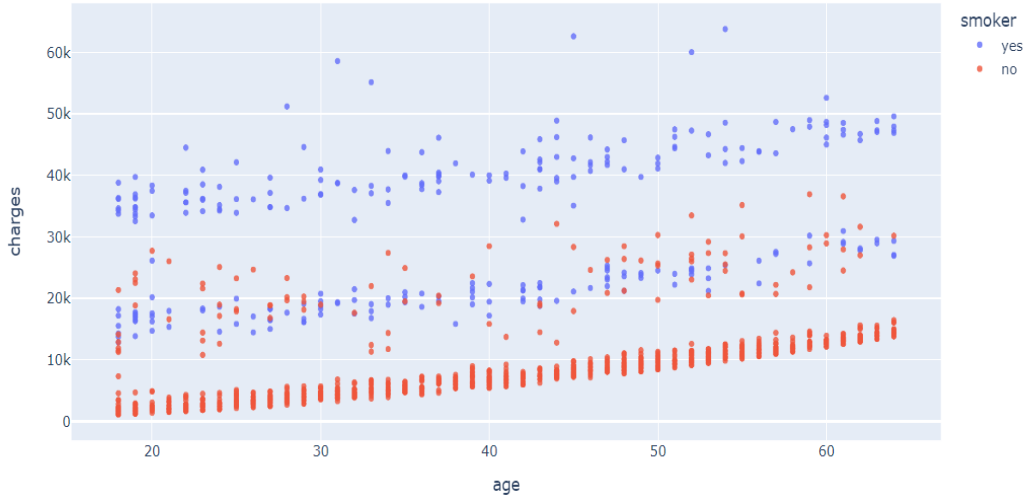


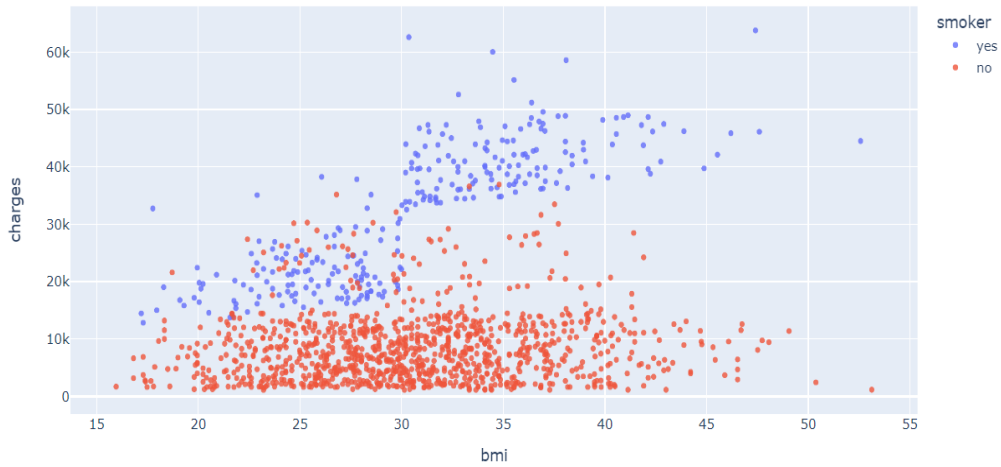
It is clear in the shown graphs that being a male or female has nearly no effect on the charge also the region of the user is non-effective as charge doesn't depend on neither of them.

Analyzing Insurance Data

Features with minimum effect:

Some other features has a little to none effect over the charge like age and body mass index of users as shown in the graphs below:





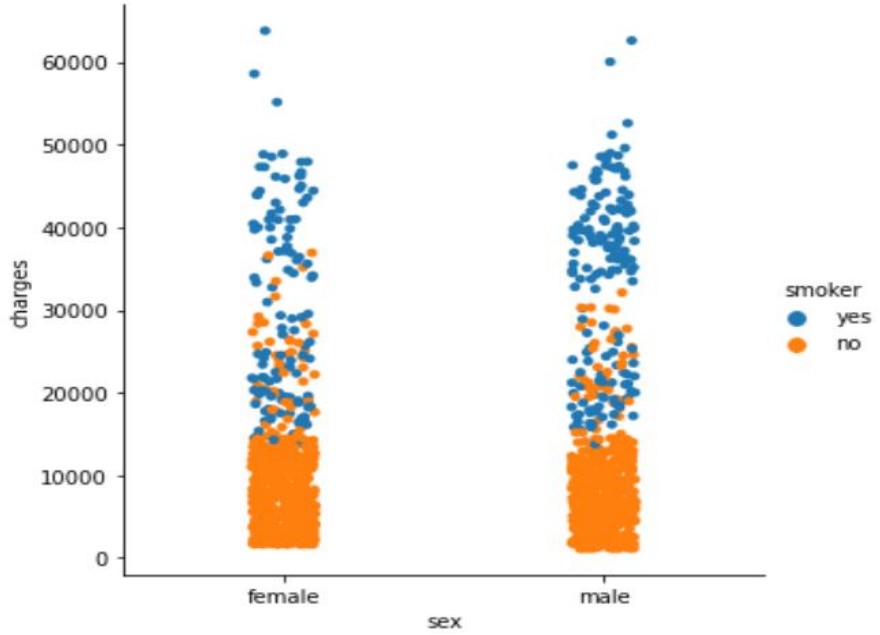
We notice that both features has a very low effect as older people have higher charge and BMI has a slight effect. But we can neglect both features in our analysis.

Analyzing Insurance Data

Features affecting the charge:

On the other hand, some features have great effect over the charge of the insurance like being a smoker or not

The following graph shows how being a smoker can change the charge:



We notice that being a smoker increases the charge either for males or females.

Minimizing the charge

After studying and analyzing the data, the best way to maintain optimum charge is by advising users not to smoke

Prediction Model

After training the data and preprocessing it . Using randomforrest algorithm we got the best possible accuracy for predicting charge of insurance for any new users as shown in the following code :

```
In [274]: from sklearn.ensemble import RandomForestRegressor
          rf = RandomForestRegressor(n_estimators=80,max_depth=6,max_features=5,random_state=123)

In [275]: rf.fit(X_train,Y_train)

Out[275]: RandomForestRegressor(max_depth=6, max_features=5, n_estimators=80,
                                random_state=123)

In [276]: rf.score(X_train,Y_train)

Out[276]: 0.9016105223789729

In [277]: rf.score(X_test,Y_test)

Out[277]: 0.8922595061637579
```

We can notice we have an accuracy (>90) with no overfitting or underfitting.

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations