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”
Data Used : https://www.kaggle.com/mirichoi0218/insurance

Agenda

Presentation consists of:

1. Data Explanation
2. Data visualization
3. Analyzing Data
4. Prediction Model

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

Data description :

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Insurance Data Explanation

What is the target of the data analysis?

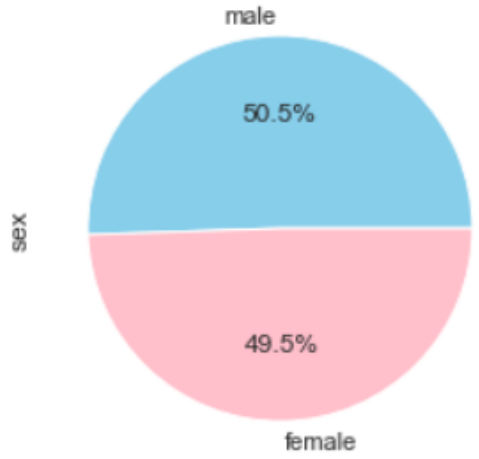
Our data analysis aims to predict the insurance charge with high accuracy given certain features of the user like his Bmi,age,if he is a smoker or not…etc.

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

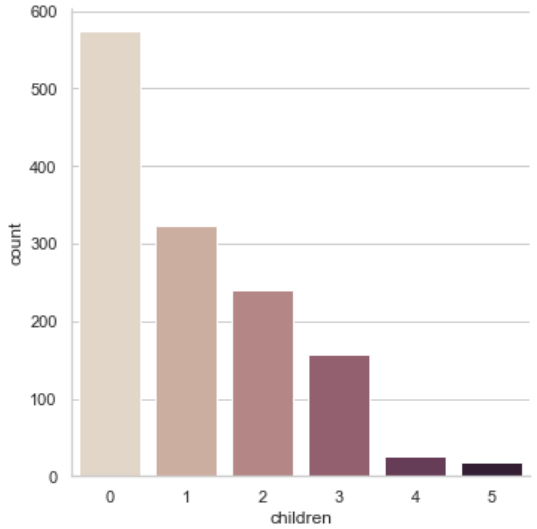
Insurance Data Visualization

Data Features :

As mentioned before our data set contains 7 features. Here are some graphs to visualize the data and give some information about it:



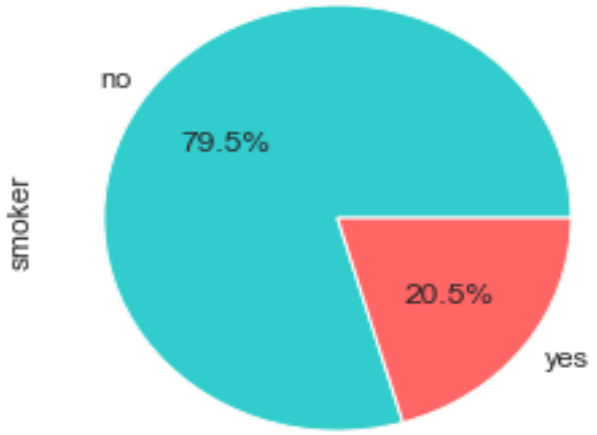
- Ratio between males to females



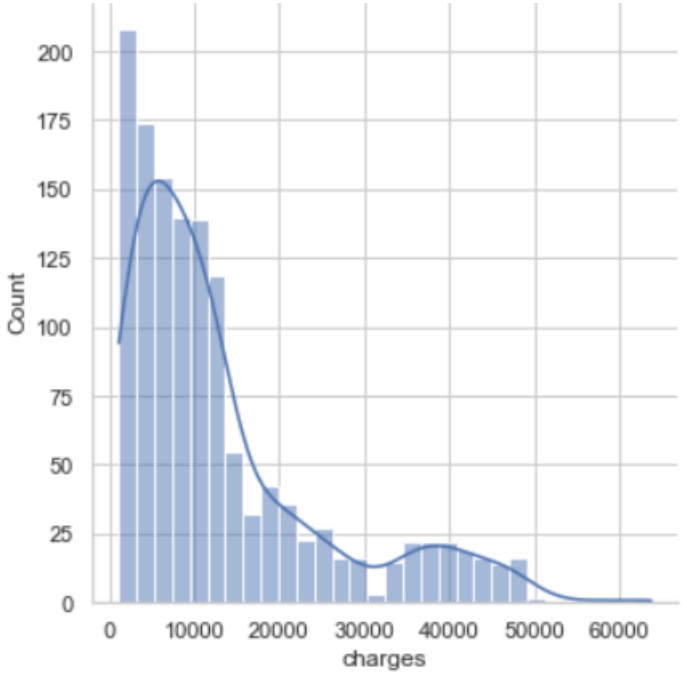
- Number of children a user has

Insurance Data Visualization

Data Features :



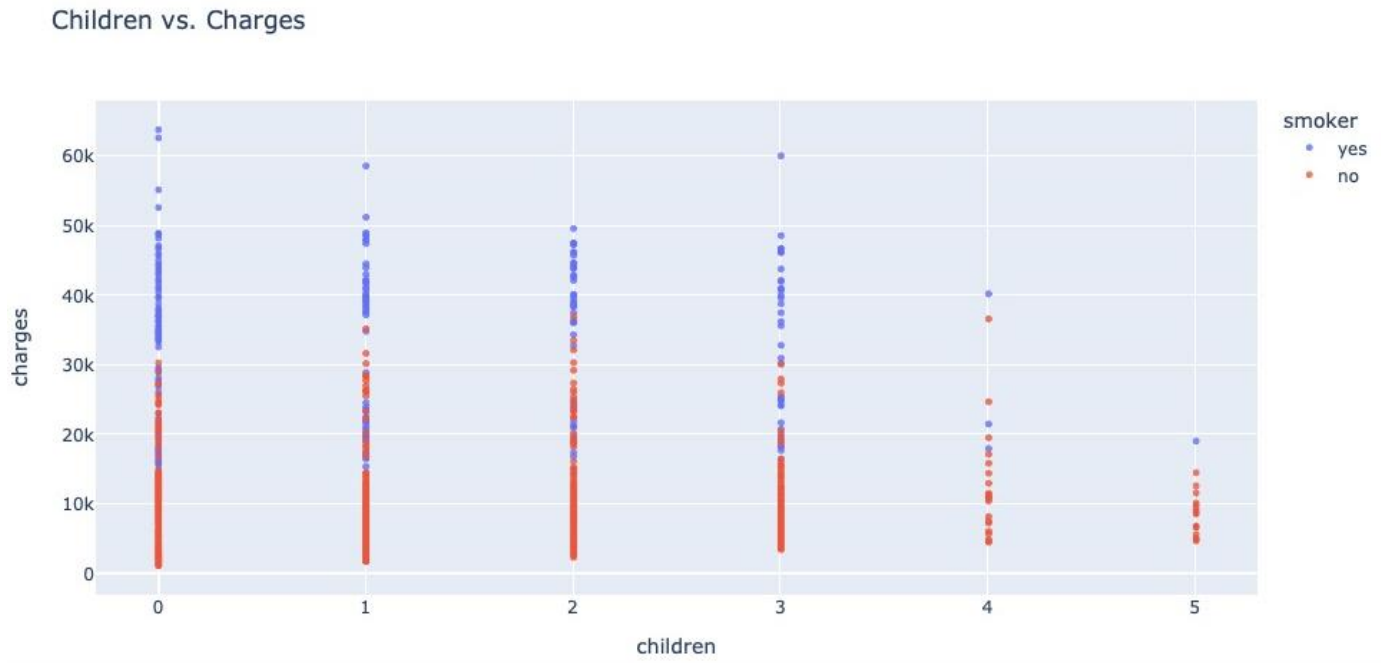
- Ratio between smokers and non-smokers



- Plot for number of users at each charge

Insurance Data Visualization

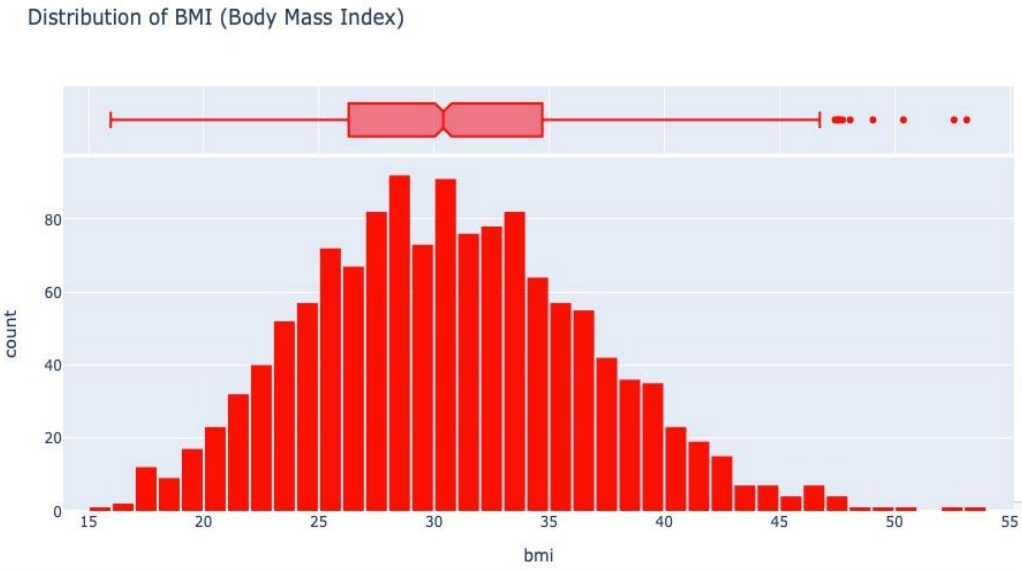
Data Features :

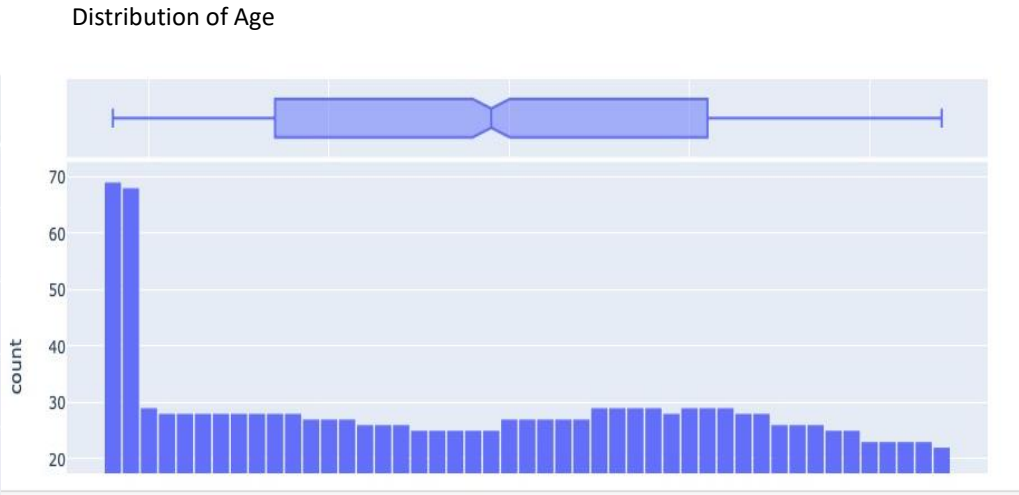


- Children effect on Charges
https://plotly.com/~h3mkader/14/

Insurance Data Visualization

Data Features :





https://plotly.com/~h3mkader/10/

https://plotly.com/~h3mkader/6/

Analyzing Insurance Data

Correlation matrix :

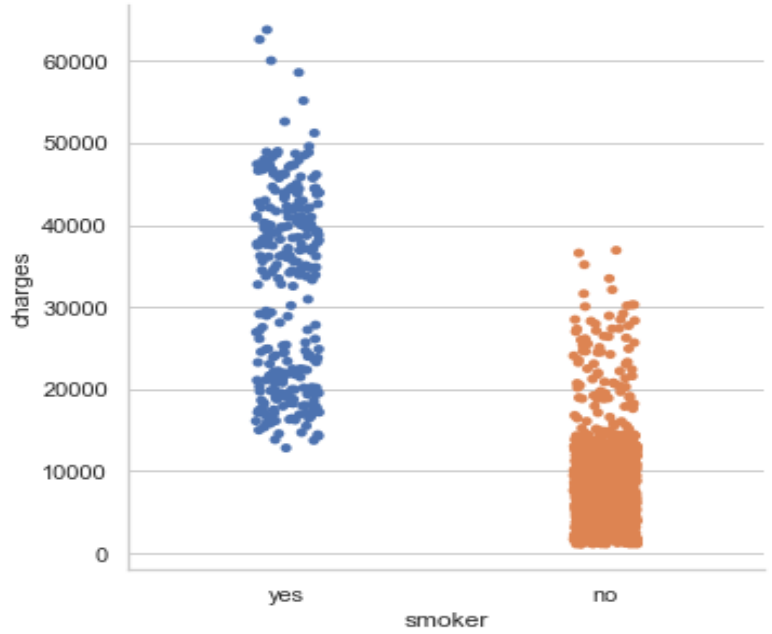
First step in analyzing the data set is getting the correlation matrix :

- Charges & Smoker has a strong relation.
- Charges & Age has a weak relation.
- Charges & BMI has a weak relation.
- Charges & Children has a very weak relation.
- Charges & Sex has a very weak relation.
- Charges & Region has nearly no relation.

Analyzing Insurance Data

Features affecting the charge:

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:

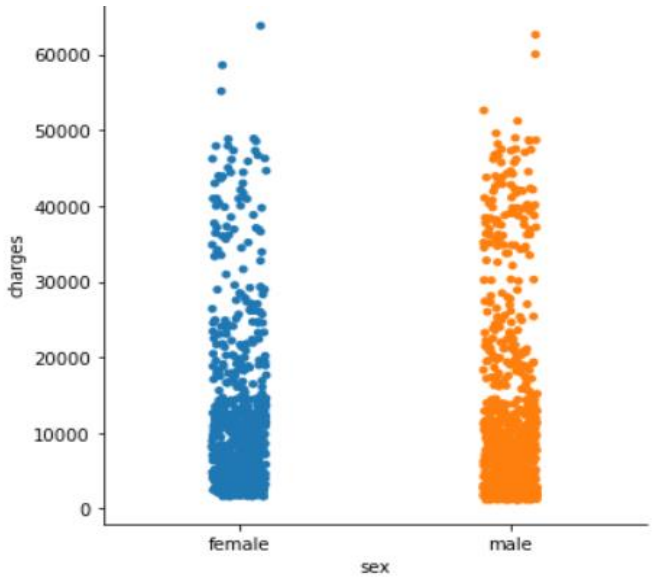


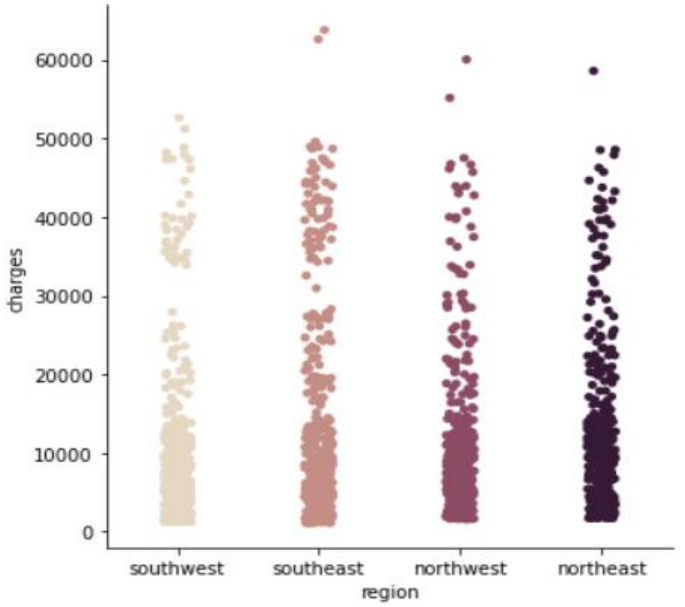
We notice that being a smoker increases the charge.

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:



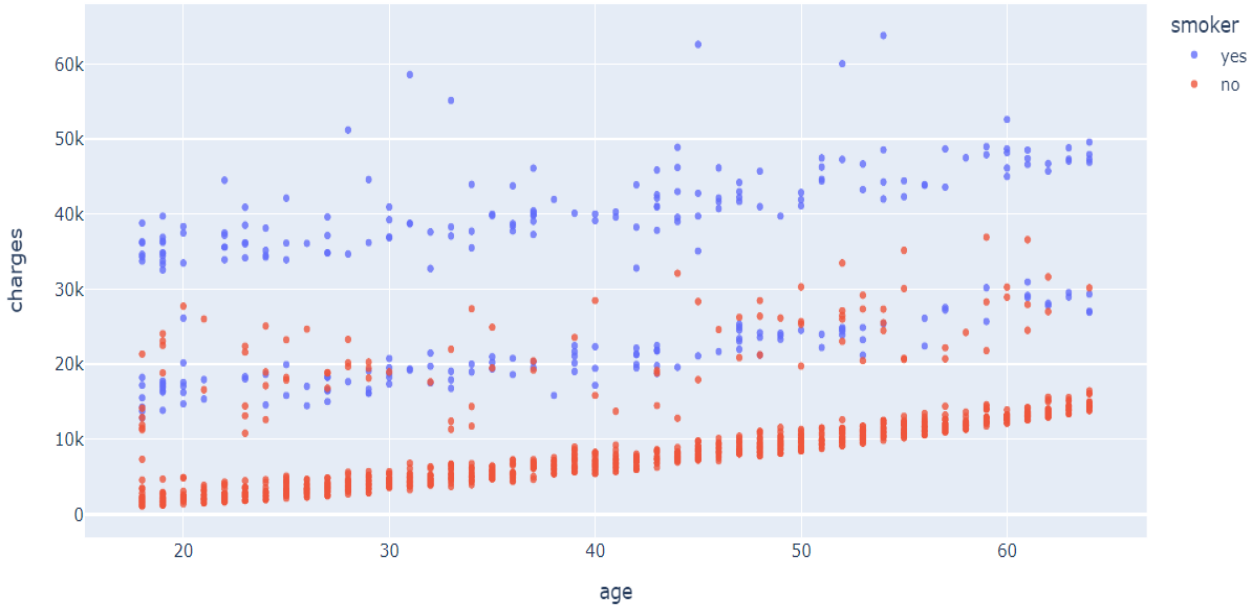


It is clear in the shown graphs that being a male or female has nearly no effect on the charge also the region of the user is non-effective as charge doesn't depend on neither of them.

Analyzing Insurance Data

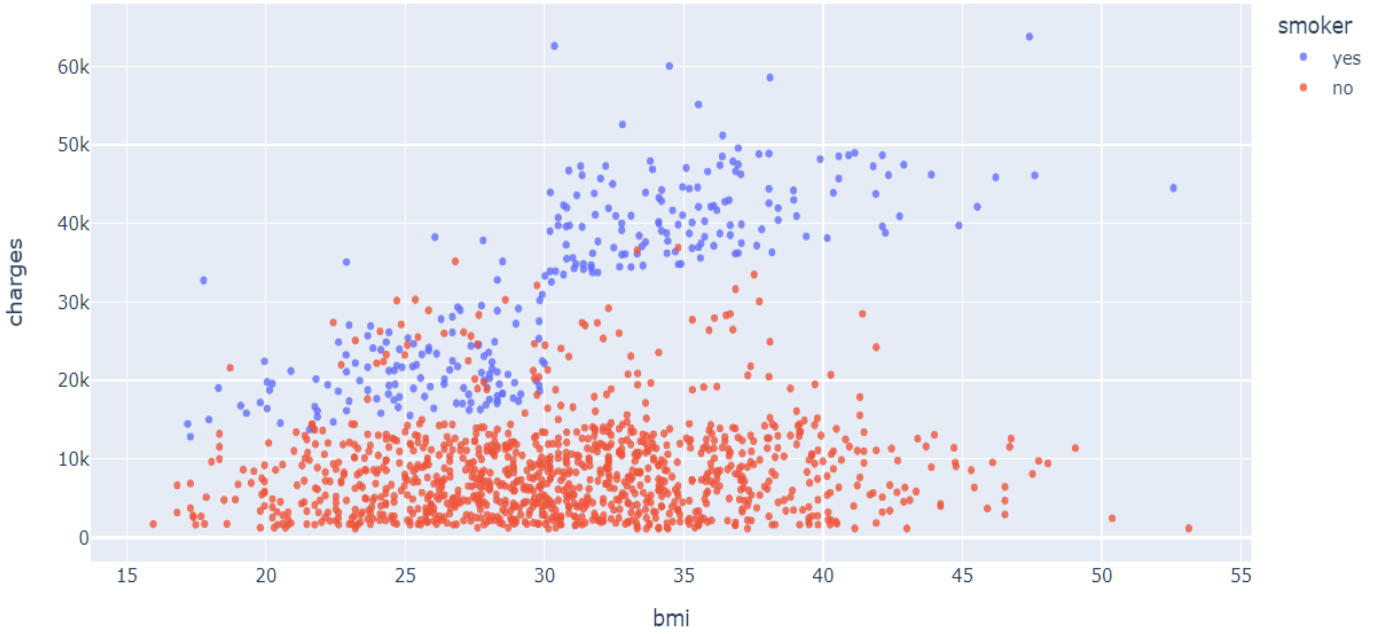
Features with minimum effect:

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:



https://plotly.com/~h3mkader/8/

Analyzing Insurance Data



https://plotly.com/~h3mkader/12/

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.

Prediction Model

After training the data and preprocessing it (label encoding) . We used multiple algorithms to get the best accuracy possible, and the accuracies where as follow :

- SVR algorithm:

  Train score : -10.311815697384263
  Test score : -13.319626486483415

- KNN algorithm:

  Train score : 20.094222013147743
  Test score : 15.470289852889863

- Linear Regression algorithm:

  Train score : 74.19903528835611
  Test score : 79.15641662953337

- Decision Tree algorithm:

  Train score : 87.05906089983986
  Test score : 87.92069936210474

Prediction Model

Random Forrest Algorithm:

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.

Project links :

Kaggle:

- Zyad Farag: https://www.kaggle.com/zyadelfakharany/notebook36c8d667fc
- Hadeer Emad: https://www.kaggle.com/hadeeremad/medical-cost
- Abdelrahman Ehab: https://www.kaggle.com/abdelrahmanehabb/insurance-data-set

GitHub:

- Zyad Farag: https://github.com/ZYAD-ELFAKHARANY2001/insurance-pr/upload/main
- Hadeer Emad: https://github.com/h3mkader/insurance-DataAnalysis/blob/c0189bcb3e77edb89f8bb9b26d1e3e9515dda25f/Insurancee.ipynb
- Abdelrahman Ehab: https://github.com/AbdelrahmanEhabb/Medical-insurance-

With our best regards.

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations