SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Samsung Innovation Campus

Artificial Intelligence Course

SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Medical Cost Personal.

Index :

Team :

- Rawan Elframawy.
- Yasmin Osama.
- Ziad Omar Yousef.

❖ Members of **Eng. Maryam Ashraf**'s team

About :

- **The dataset is for a group of residents in the United States.**
- **It contains some information about patients including health insurance charges.**
- **We want to predict health insurance charges.**

Data Features :

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Content :

- **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

Content :

Category	BMI range - kg/m^2
Severe Thinness	< 16
Moderate Thinness	16 - 17
Mild Thinness	17 - 18.5
Normal	18.5 - 25
Overweight	25 - 30
Obese Class I	30 - 35
Obese Class II	35 - 40
Obese Class III	> 40

Content :

- **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.

•Data info :

- **The data :**
- **1338** instances.
- **7** attributes.
- **2** integer type.
- **2** float type.
- **3** object type (Strings in the column).

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
```

In [7]:
```
# The number of rows and columns in this data frame
df.shape
```

Out[7]:
```
(1338, 7)
```

- **Findings:**
- There are four numerical variables:
  - Continuous : Age - BMI - Charges
  - Discrete : children
- There are three categorical variables: - Sex - Smoker - Region

Describing data:

- **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

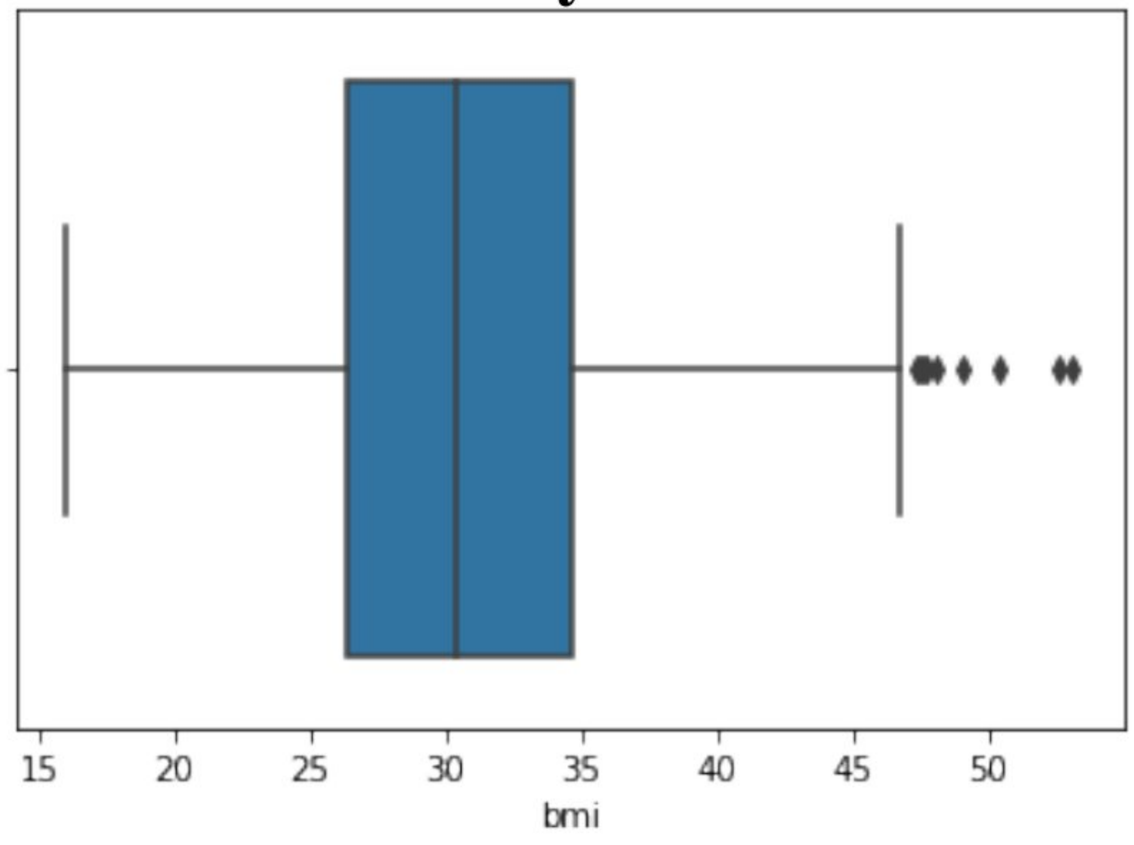
Is there missing values:

- **No missing values.**

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Detecting outliers :

Does it indicate that the data is really stable? Check the outliers...



I got them 😈

Detecting Outliers :

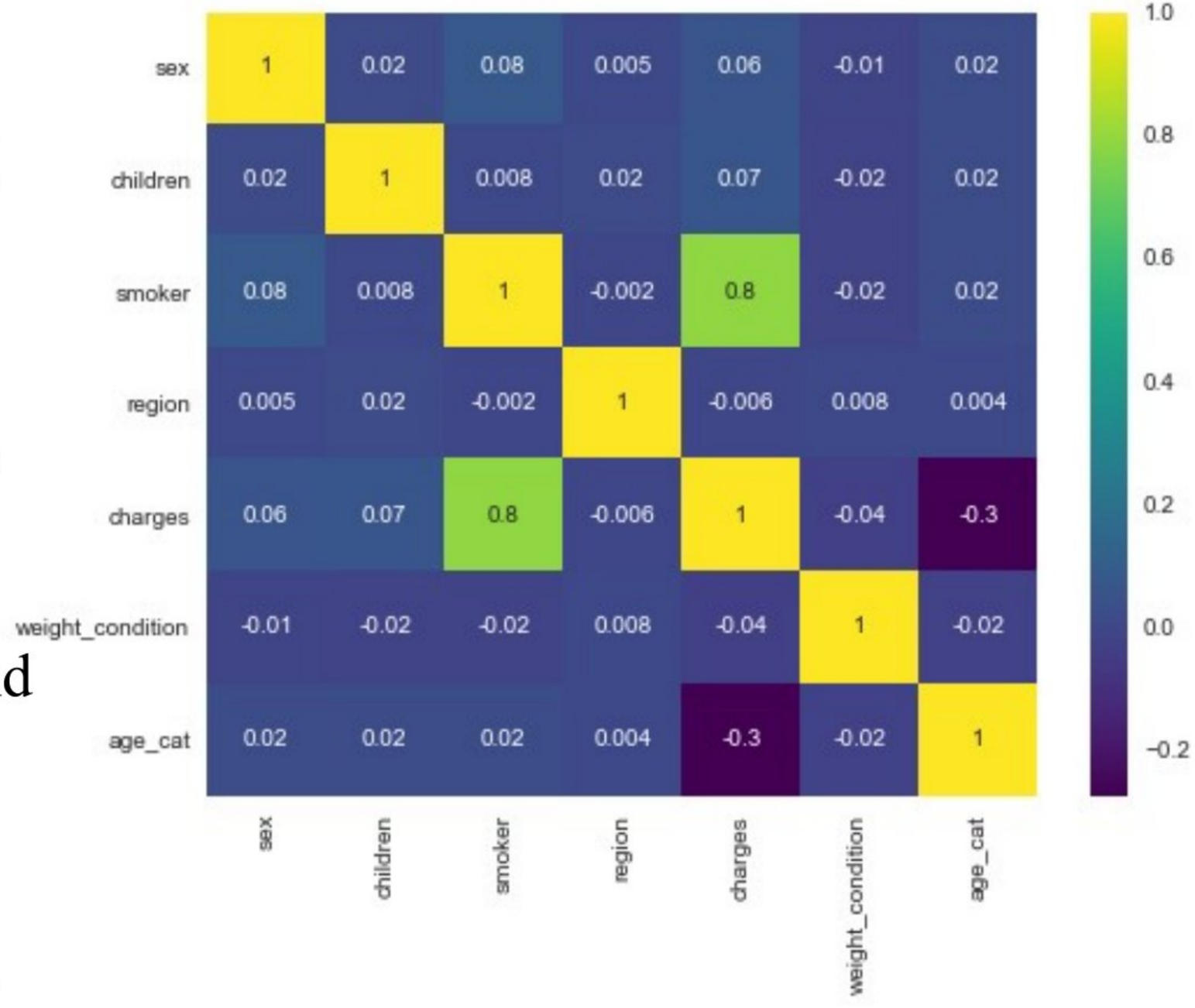
Let's find out which this data is..

	age	sex	bmi	children	smoker	region	charges
116	58	male	49.06	0	no	southeast	11381.32540
286	46	female	48.07	2	no	northeast	9432.92530
401	47	male	47.52	1	no	southeast	8083.91980
543	54	female	47.41	0	yes	southeast	63770.42801
847	23	male	50.38	1	no	southeast	2438.05520
860	37	female	47.60	2	yes	southwest	46113.51100
1047	22	male	52.58	1	yes	southeast	44501.39820
1088	52	male	47.74	1	no	southeast	9748.91060
1317	18	male	53.13	0	no	southeast	1163.46270

only 9 records! we can get red of them safely

Correlation :

- **Is there is a relationship between the features and each other?**



- The correlation between the data is positive but weak.
- The highest correlation is between medical charges and age (0.8), big .

Label Encoding :

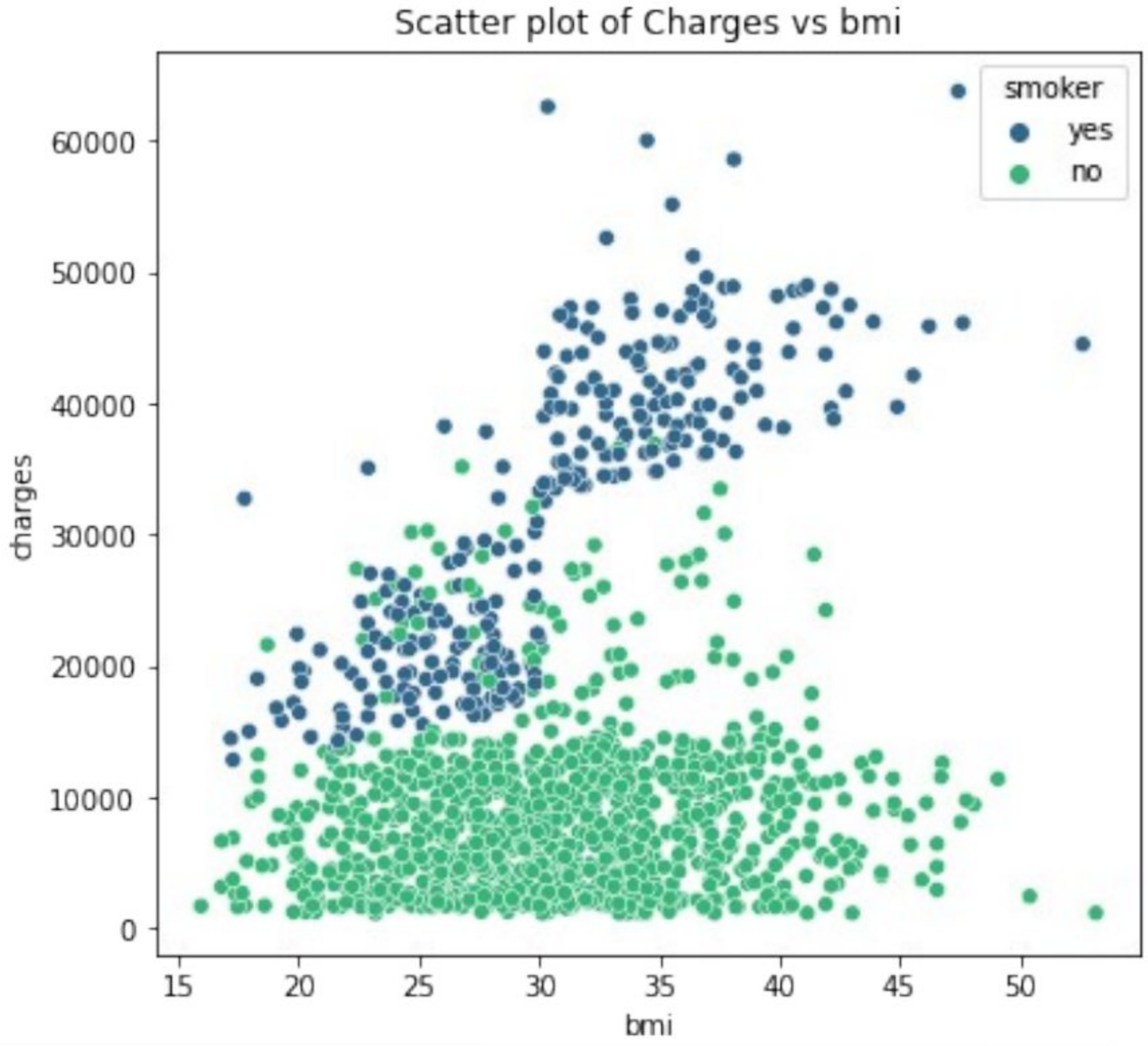
- Converting the labels (Categorical Data) into numeric form. Such as (Sex , Smoker , Region)

region
southwest
southeast
southeast
northwest
northwest

region
3
2
2
1
1

EDA:

take a look at the features and its effect in the target "charges" ...



insights:

- Most of the people who smoke are obese people
- Even though a person is not obese but he/she smokes, his/her charges tends to be higher.

EDA:

I see that It will be clearer if we classify people into categories according to their "bmi" 💡

- bmi < 18.5 , be underweight
- bmi > or = 18.5 & bmi <24.5 , be normal
- bmi > or = 24.5 & bmi < 30 , be overweight
- bmi > 30 , be obese

bmi	c
27.900	
33.770	
33.000	
22.705	
28.880	

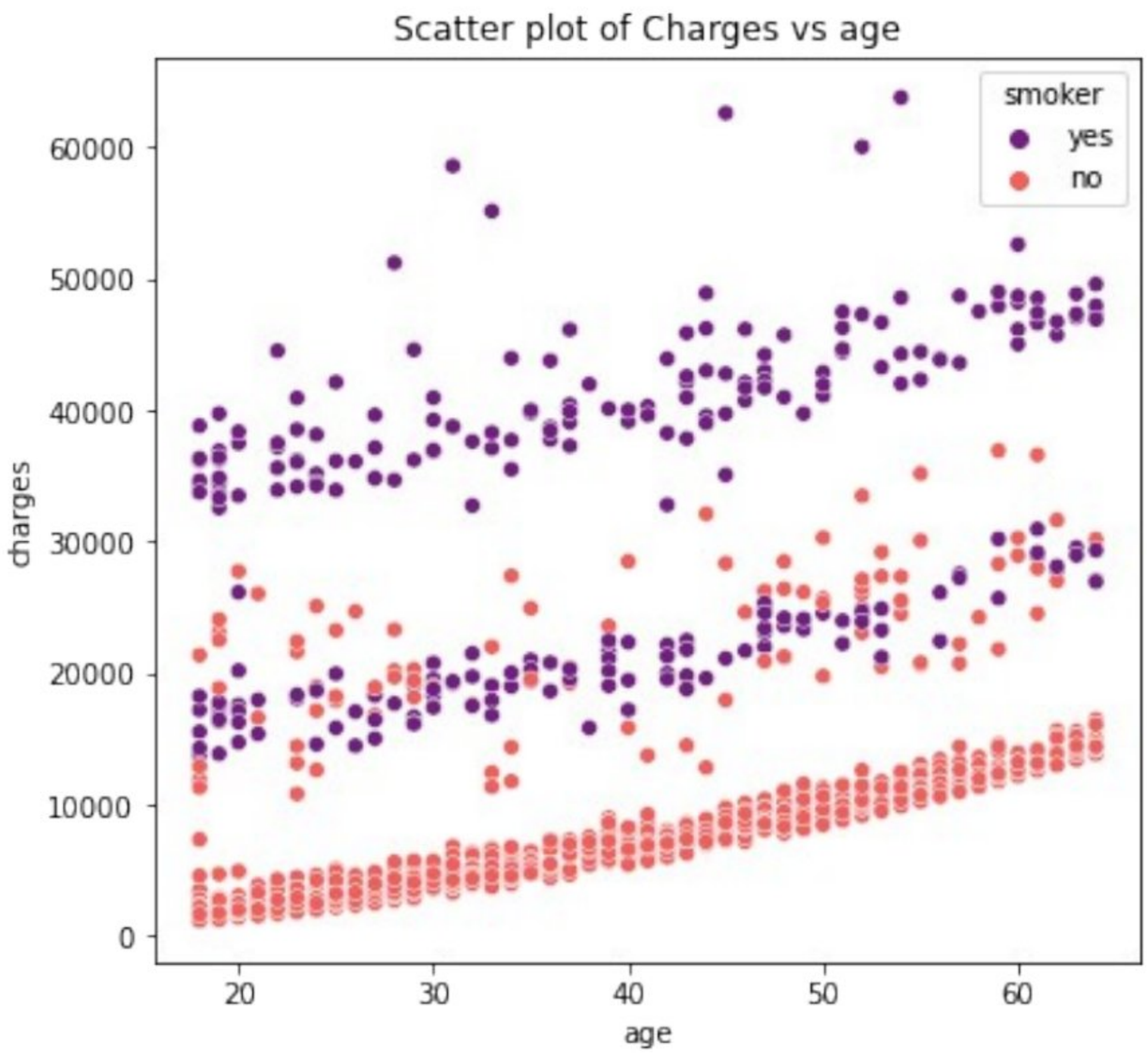
weight_condition
Overweight
Obese
Obese
Normal
Overweight

EDA:

age VS charges

Insights

- Charges increases as the age of the person increases
- Smokers and obese people have the highest charges regardless of their age.



EDA:

I'll do the same thing with "age" so that the data to be follow the same flo 👍

- Age <= 35 year , be Young Adult
- Age <=50 year , be Senior Adult
- Age > 50 year , be Elder

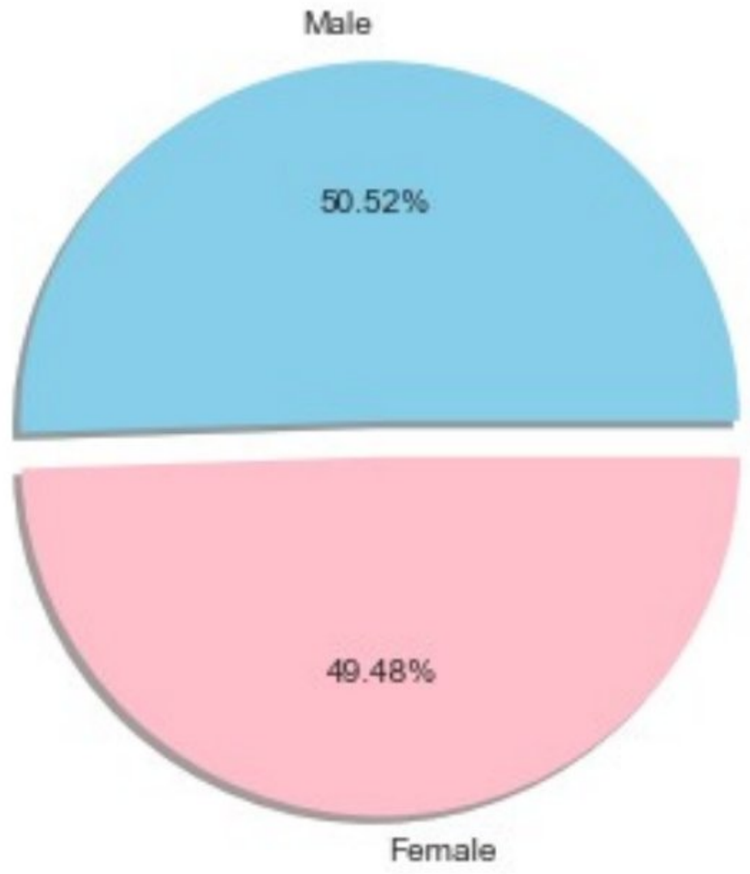
age
19
18
28
33
32

age_cat
Young Adult
Young Adult
Young Adult
Young Adult
Young Adult

EDA:

Relation between sex and charges:

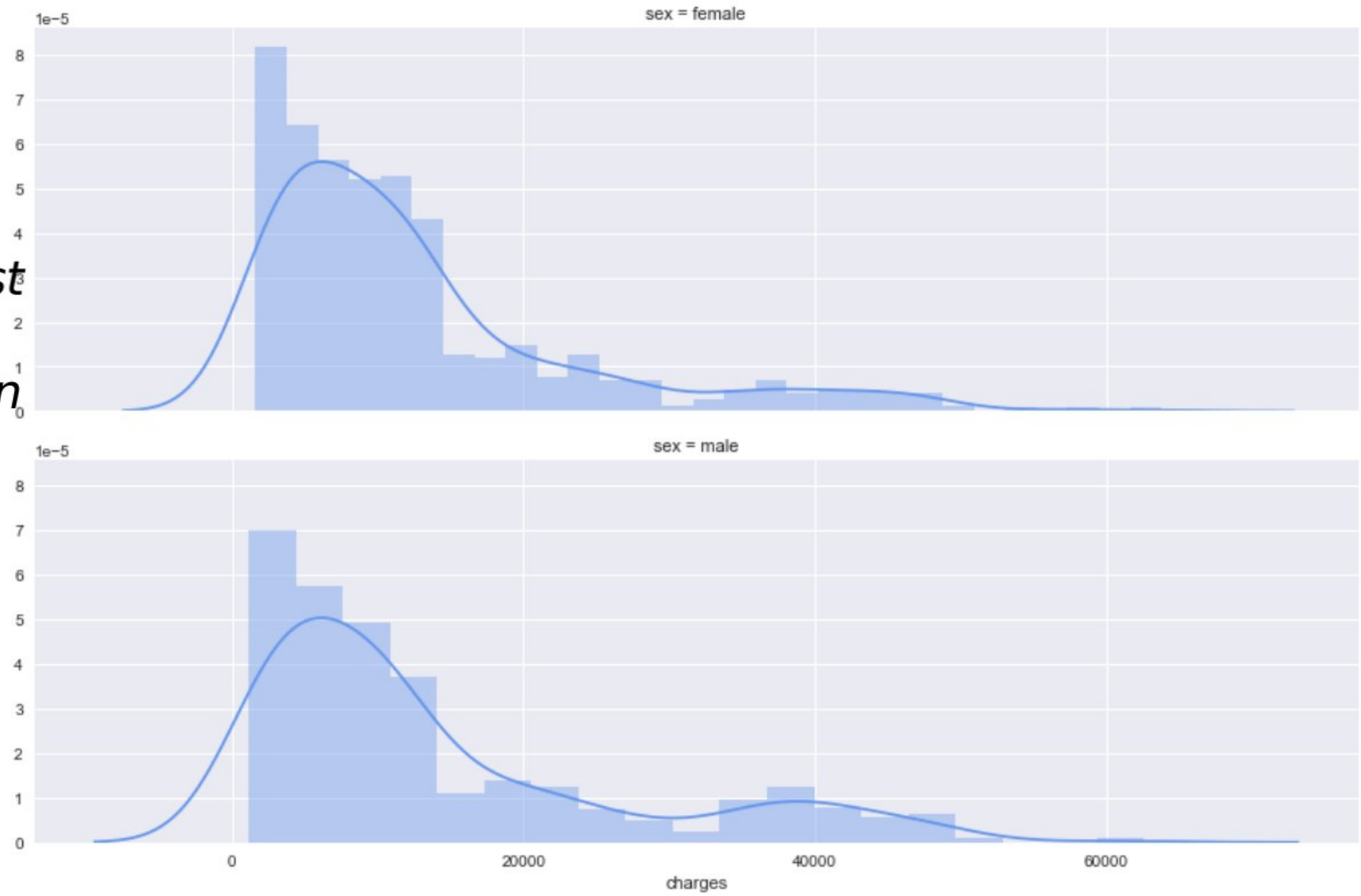
Almost the same proportion...



EDA:

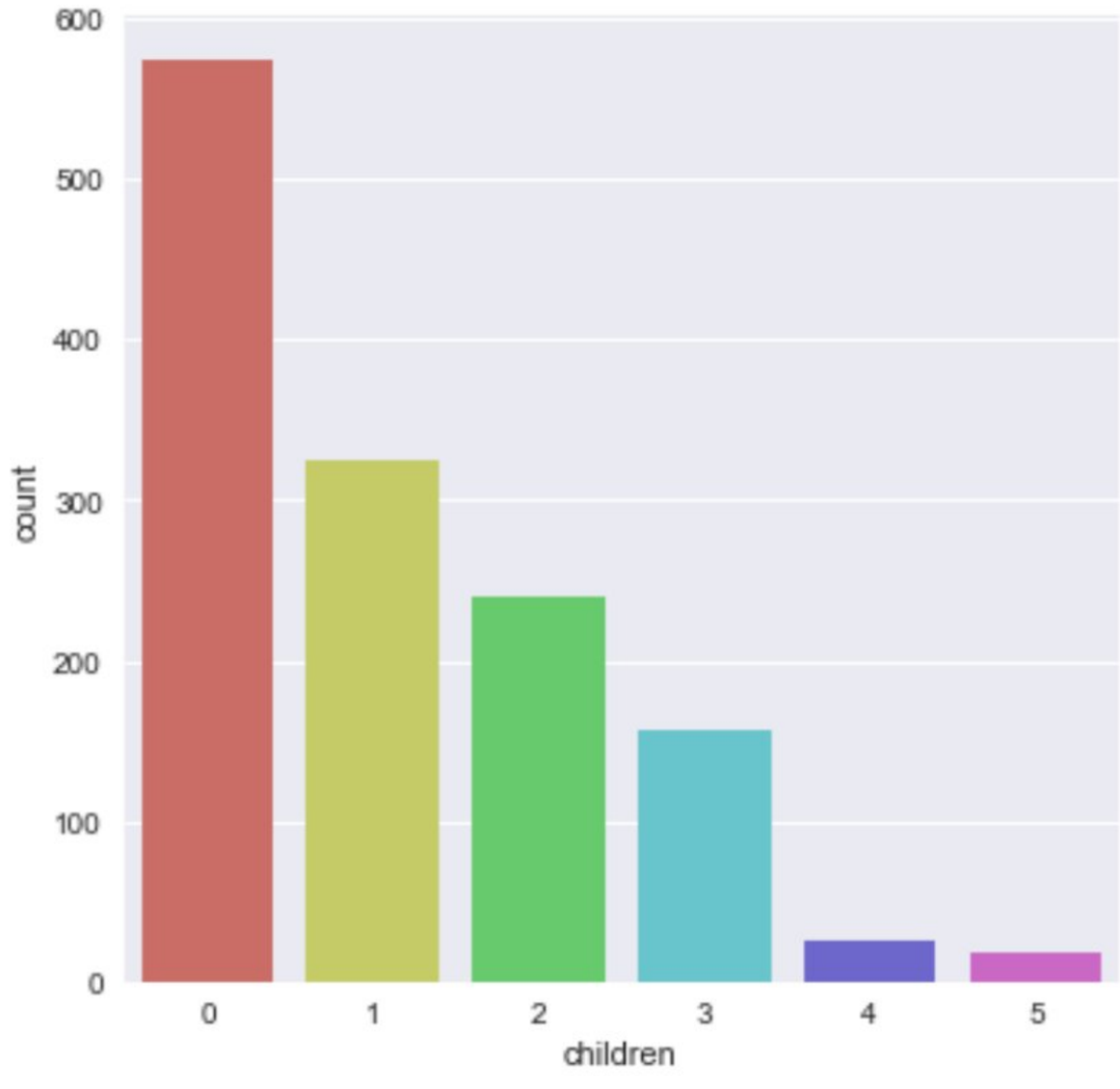
Checking the charges distributions for males and females

As we can see the two distributions are almost the same for both women/men, so we can affirm that there is no influence on the medical charges when it comes to the sex variable.



EDA:

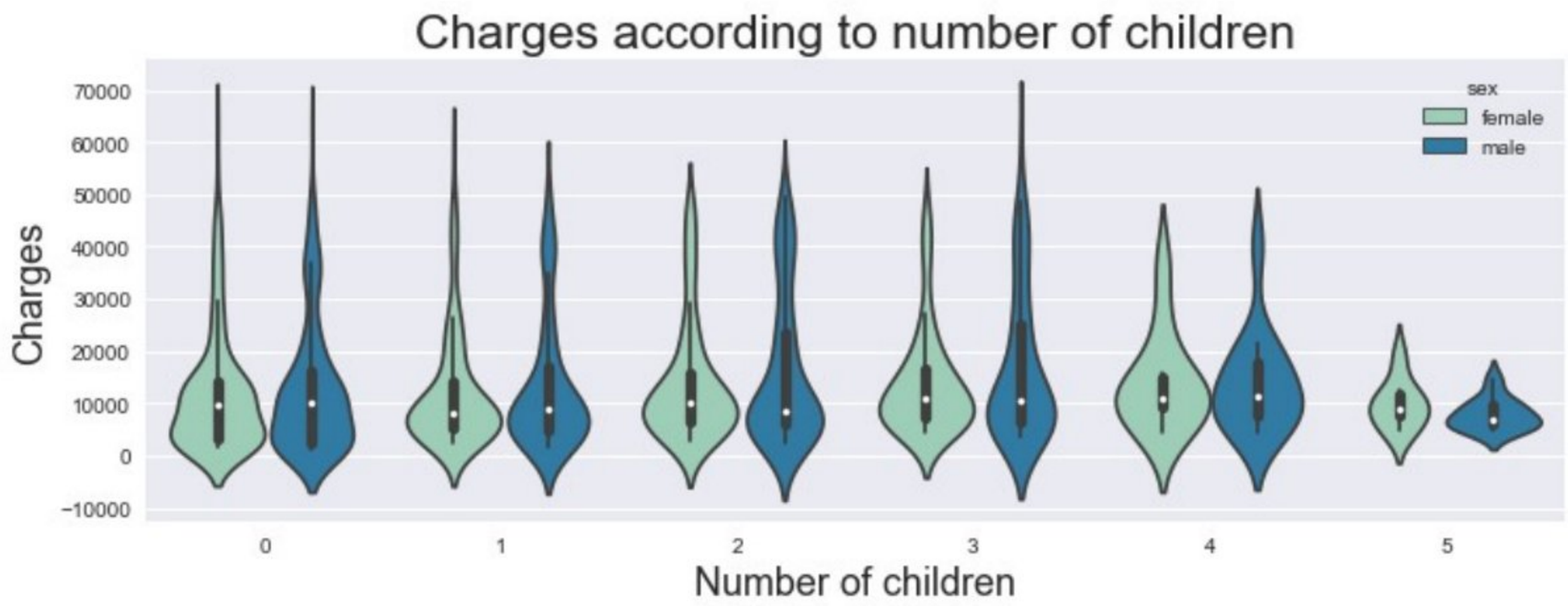
let's see how many children our patients have.



Most patients do not have children

EDA:

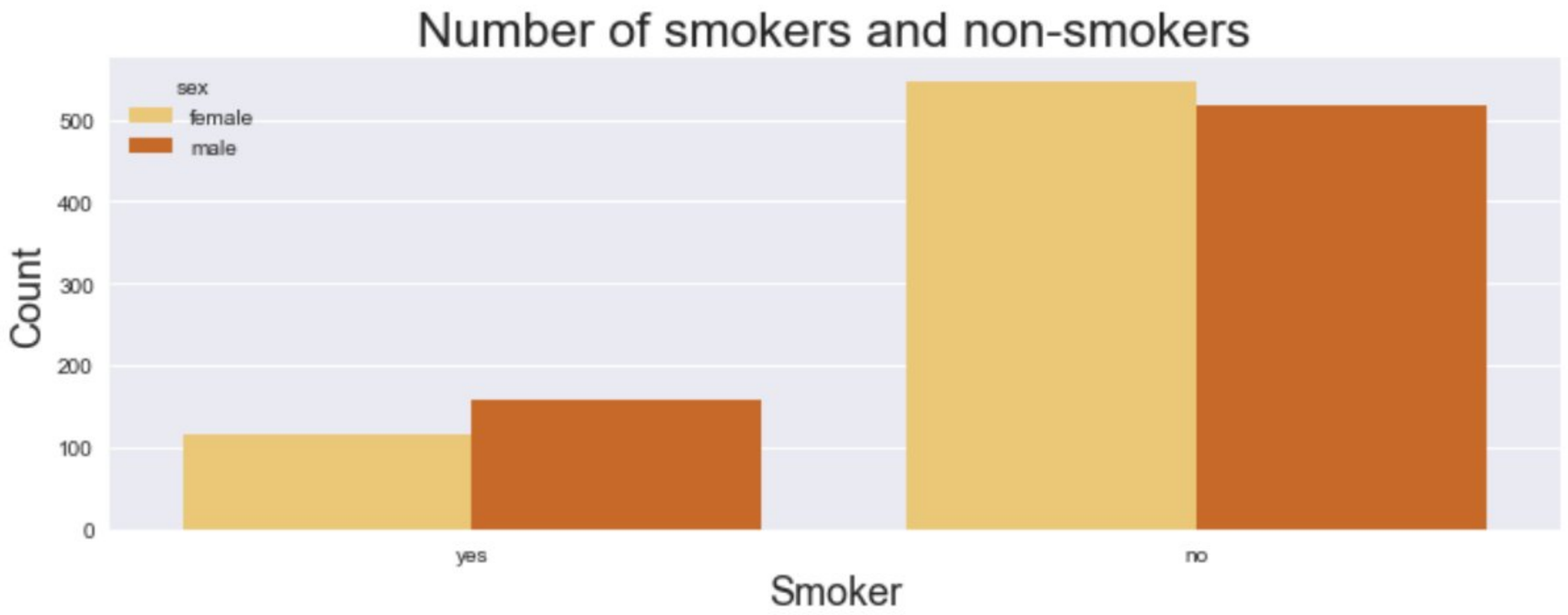
Creating a violin plot for each category



As we can see, almost all categories have the same range and mean of costs also the distributions are very similar, except for the people who have 5 children. This might be because of the small size of the sample of this kind of people!

EDA

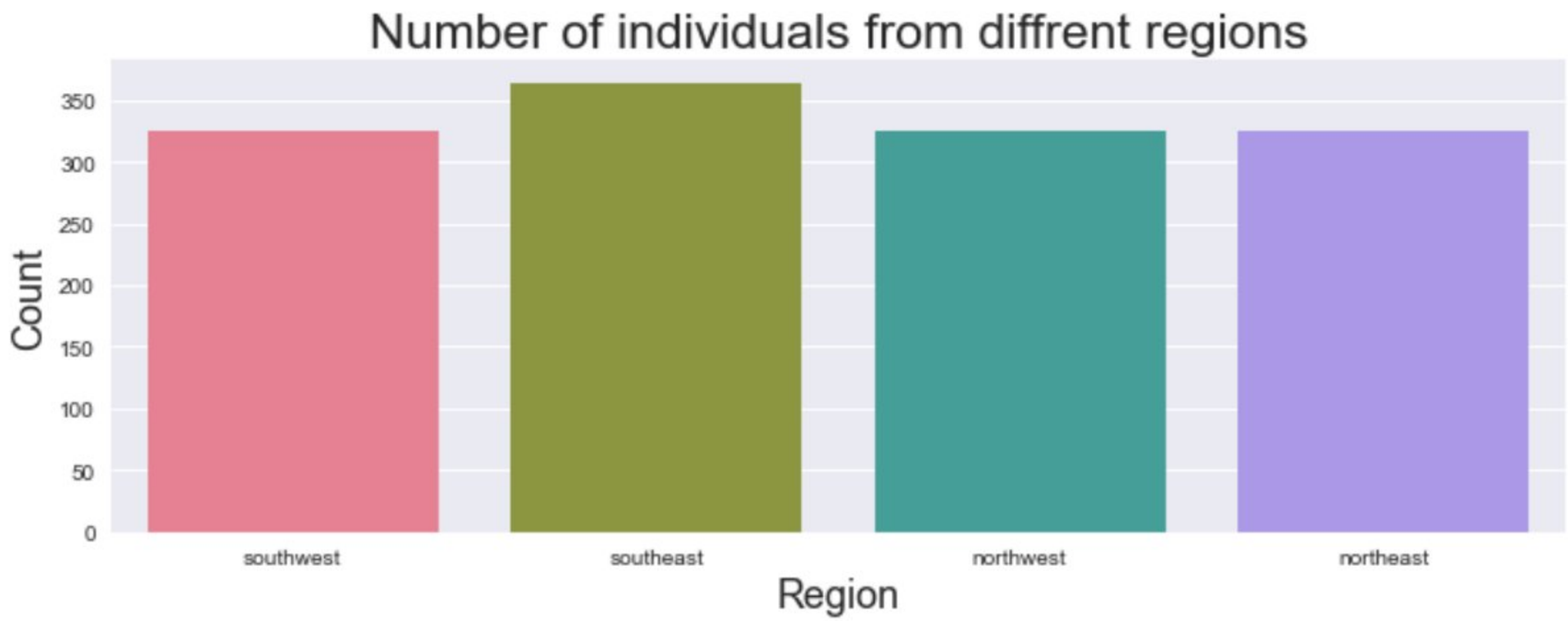
Relation between smoking and charges:



As we can see charges for smokers are much higher than charges for non-smokers. Sex doesn't have any effect on charges when you are a smoker

EDA:

Relation between regions and charges:



Almost all the regions have equal number of individuals.

Conclusion:

We have found out that region and gender does not bring significant difference on charges among its groups. Age, BMI, number of children and smoking are the once that drives the charges.

As:

- age has an impact on the charges, when a person is older the health costs are larger.
- if you are a smoker, you must expect some huge medical charges compared to non-smokers. Especially for people who have high BMI values (>35) it will result very serious health care charges.
- no matter where you live, this won't have any impact on your medical insurance bills.
- the number of children doesn't affect the medical costs billed by health insurance.
- it doesn't matter if you are a men or a women your health bills won't change.

Modeling

From the data analyzes we found that it is possible to drop [sex & regions]
Then we reprocessed the data like label Encoding and others..
Also, we split the data based on the features and the target, and train & test
and then we started the algorithm tests to choose from.

We tested eight algorithms :

1. Linear Regression
2. KNeighborsRegressor
3. DecisionTreeRegressor
4. RandomForestRegressor
5. AdaBoostRegressor
6. GradientBoostingRegressor
7. XGBRegressor
8. CatBoostRegressor

Modeling

We have set parameters for each algorithm based on the best result.. We got them after a lot of searching

```
lr  = LinearRegression()
knn = KNeighborsRegressor(n_neighbors=10)
dt  = DecisionTreeRegressor(max_depth = 3)
rf  = RandomForestRegressor(max_depth = 3, n_estimators=500)
ada = AdaBoostRegressor( n_estimators=50, learning_rate =.01)
gbr = GradientBoostingRegressor(max_depth=2, n_estimators=100, learning_rate =.2)
xgb = XGBRegressor(max_depth = 3, n_estimators=50, learning_rate =.2)
cb  = CatBoostRegressor(learning_rate =.01, max_depth =5, verbose = 0)
```

Modeling

What about parameters !

1-

```
knn = KNeighborsRegressor(n_neighbors=10)
```

2-

```
dt  = DecisionTreeRegressor(max_depth = 3)
rf  = RandomForestRegressor(max_depth = 3, n_estimators=500)
```

Modeling

3-

```
gbr = GradientBoostingRegressor(max_depth=2, n_estimators=100, learning_rate =.2)
xgb = XGBRegressor(max_depth = 3, n_estimators=50, learning_rate =.2)
cb  = CatBoostRegressor(learning_rate =.01, max_depth =5, verbose = 0)
```

Modeling

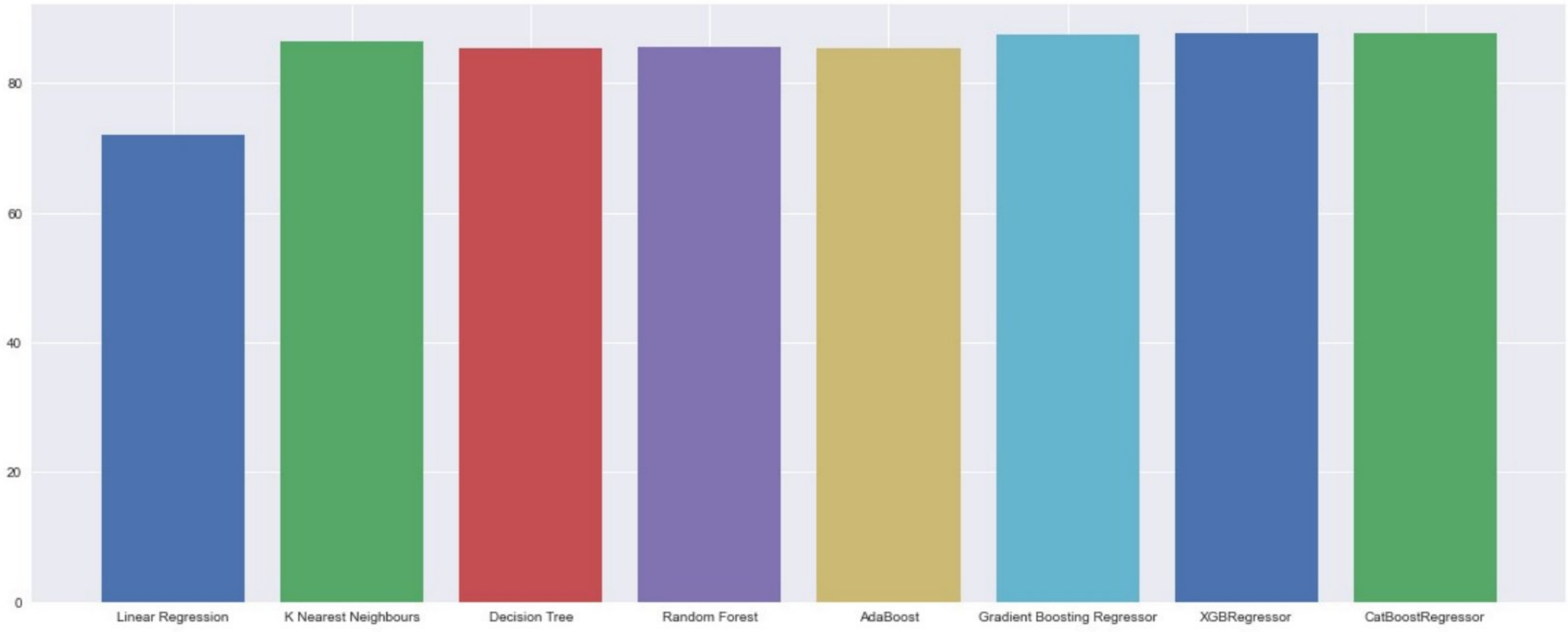
Why didn't we use Gridsearch ?

We found …

- It takes too much time to make the selection process
- In most algorithms it does not choose the best case
- When we searched for the best values of parameters to put.. the results of our search were better than Gridsearch

Modeling

Results :



Modeling

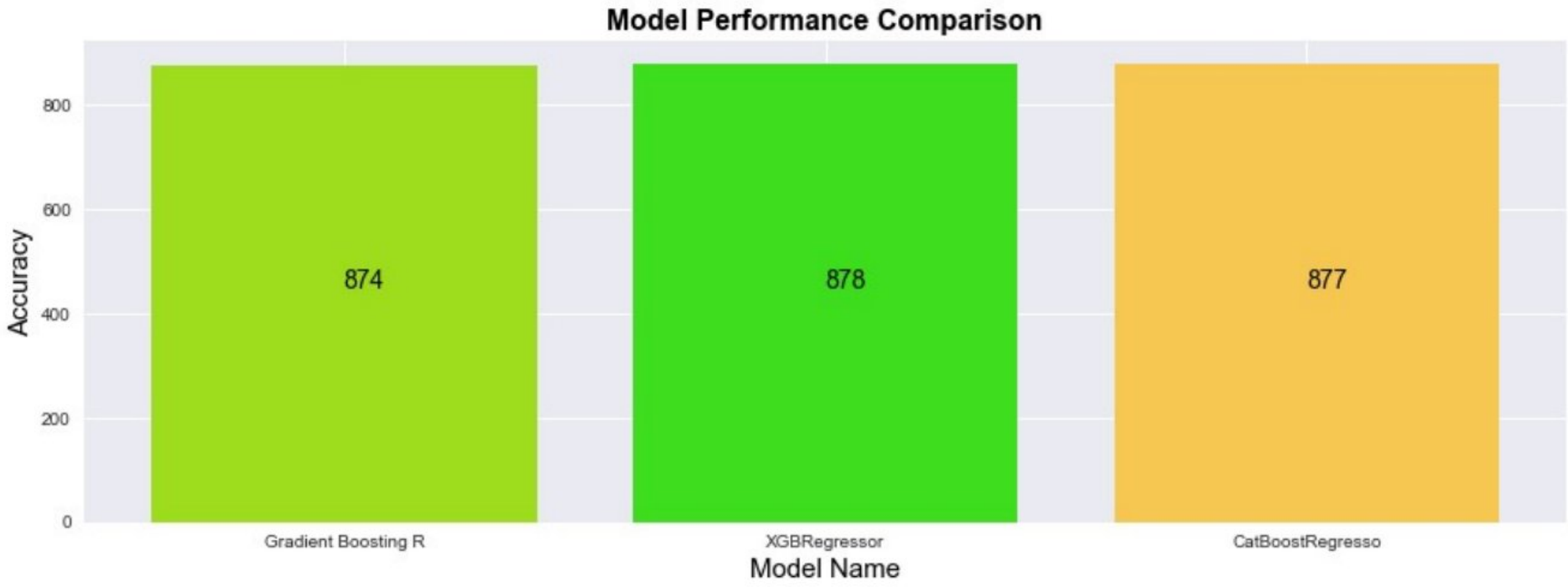
- Linear Regression : 72 %
- K Nearest Neighbours : 86 %
- Decision Tree : 85 %
- Random Forest : 86 %
- AdaBoost : 85 %
- Gradient Boosting Regressor : 88 %
- XGBRegressor : 88 %
- CatBoostRegressor : 88 %

Best Models

1. Gradient BoostingRegressor : 88 %
2. CatBoostRegressor : 88 %
3. XGBRegressor : 88 %

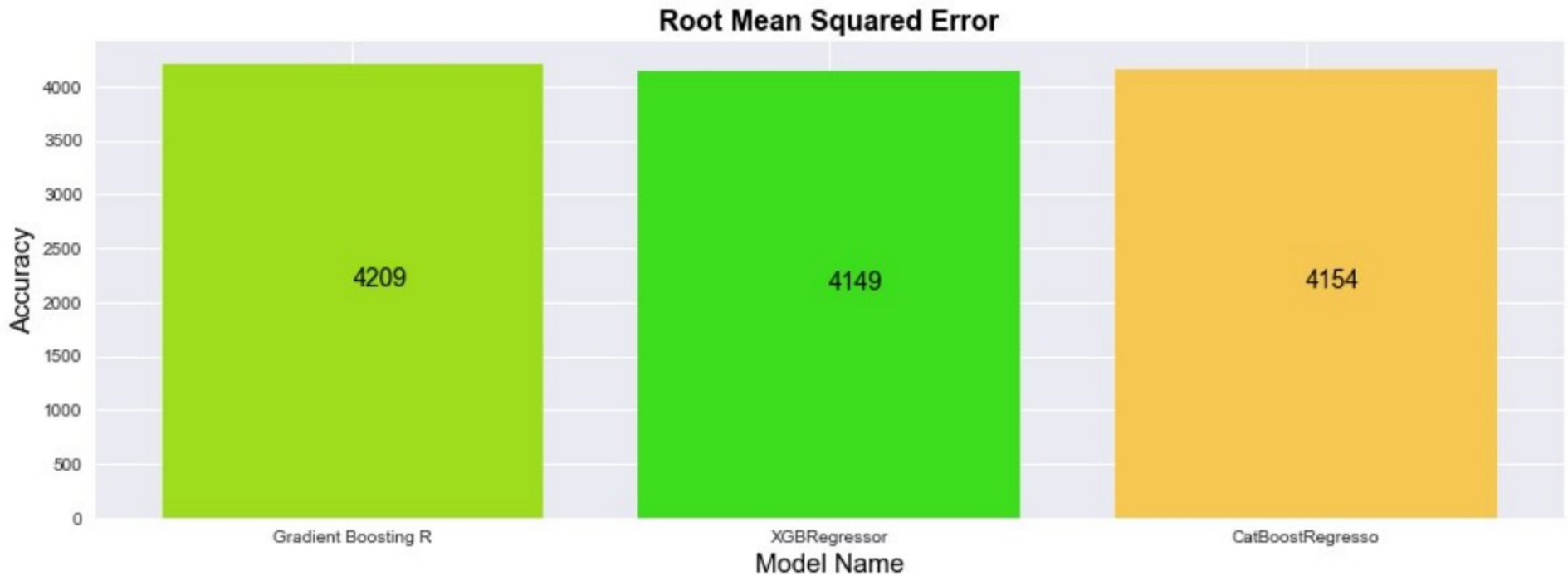
Modeling

R square (r2) Score & Root Mean Squared Error(RMSE) :



Modeling

Root Mean Squared Error(RMSE):
It is the squared root of the mean of the difference between actual and the predicted values.



Modeling

What are the advantages of this algorithm?

- It is designed to handle missing data with its in-build features.
- It Works well in small to medium dataset
- One nice feature of Trees (XGBoost, Random Forest, etc.) is you don’t have to normalize your features as you should with SVM, etc
- work well if the data is nonlinear, non-monotonic, or when segregated clusters.

Disadvantages

can over-fit the data, especially if the trees are too deep with noisy data.

Modeling

Cross Validation

It is a resampling procedure which is used to evaluate the machine learning models on limited data samples. Its goal is to predict new data that is that is not tested before.

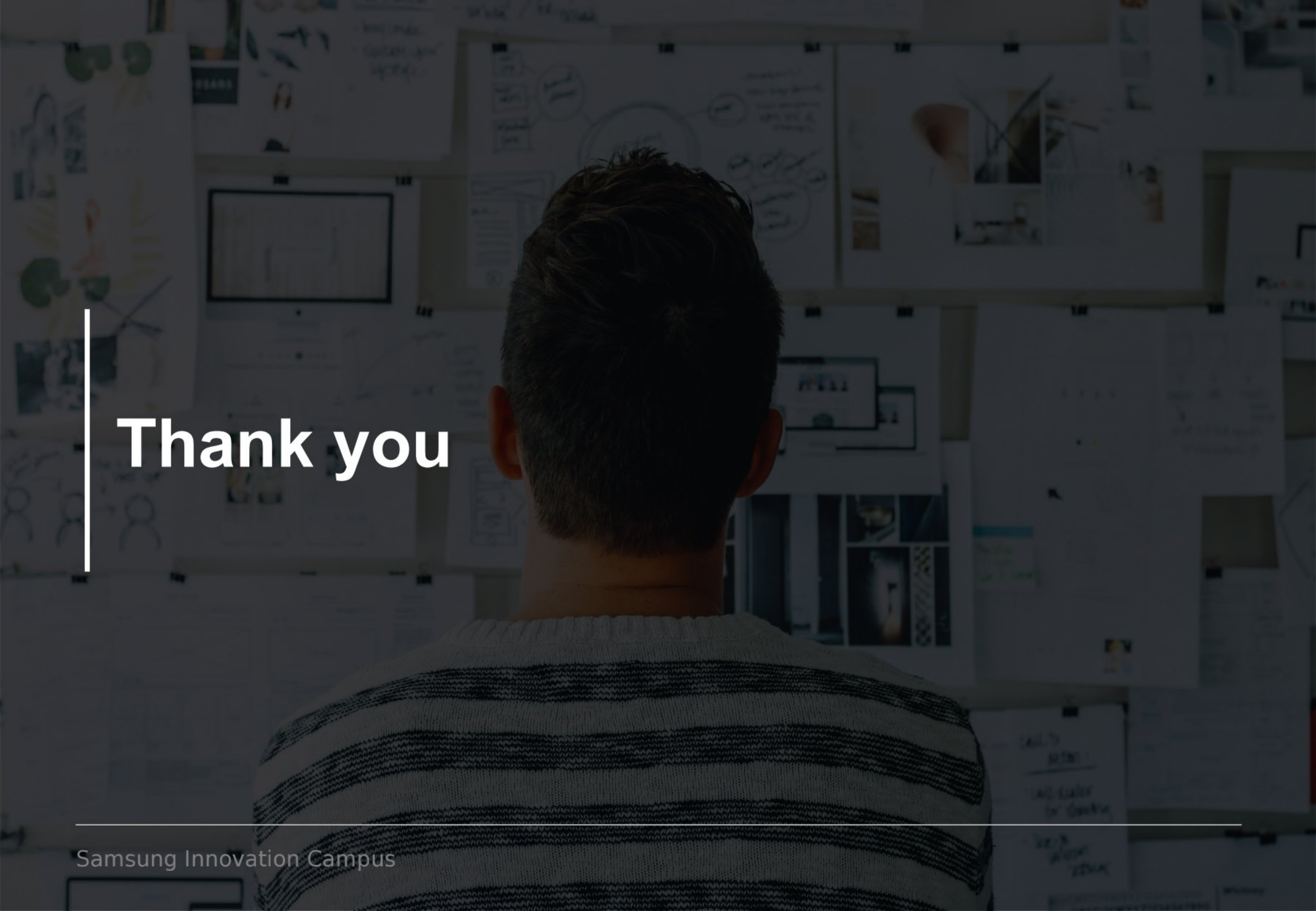
```
score = cross_val_score(xgb, x, y, cv=5)
print(score)
```

[0.85890086,
0.79149981,
0.85040421,
0.83031141,
0.86551417]

Where is the code

GitHub : https://github.com/Ziad-o-Yusef/MCP
Kaggle: https://www.kaggle.com/rawanelframawy/medical-cost-personal

Questions?



Thank you

Samsung Innovation Campus

SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations