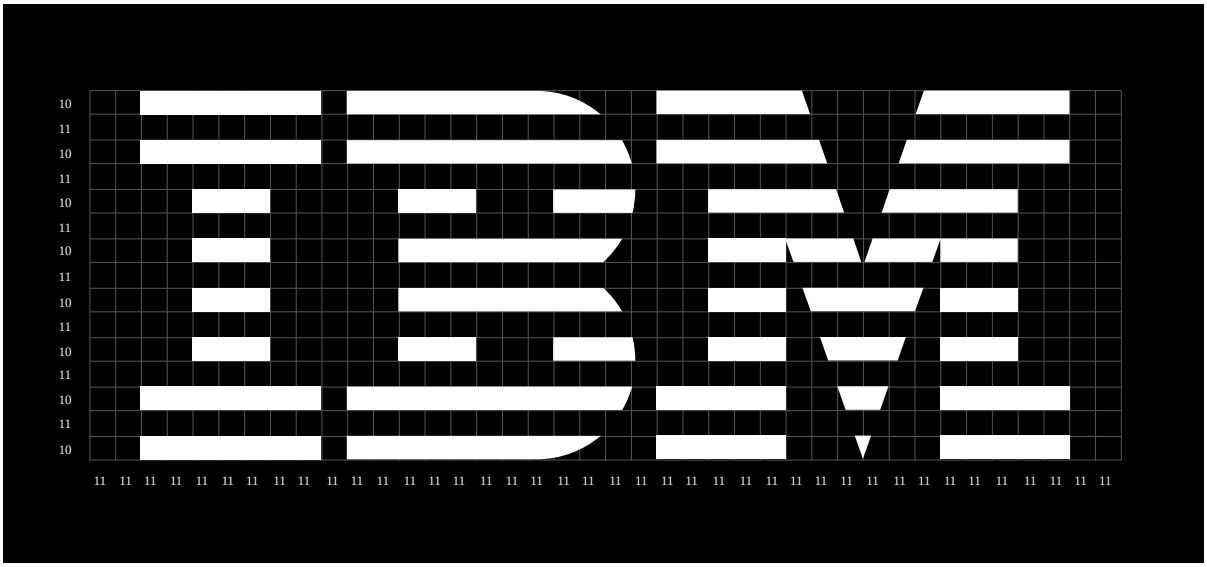


# Supervised Machine Learning Course Final Project

## Data Description

The name of the dataset is ‘Medical Cost Personal Datasets’.

| Name | Description |
|---|---|
| age | age of the primary beneficiary |
| sex | insurance contractor gender, female, male |
| BMI | Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height., |
| 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, and northwest |
| charges | individual medical costs billed by health insurance |

## Analytic Goal

This dataset's analysis's primary goal is to predict the insurance cost accurately.

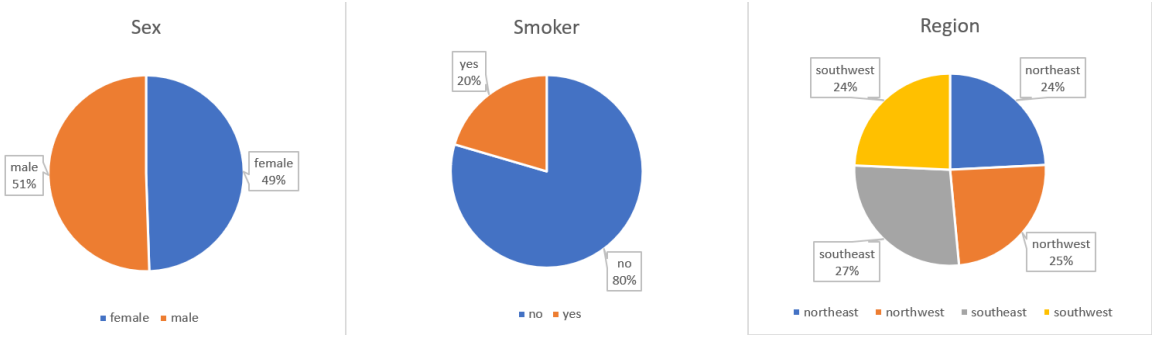
# EDA (Exploratory Data Analytics)

- Checking basic stats of the data frame

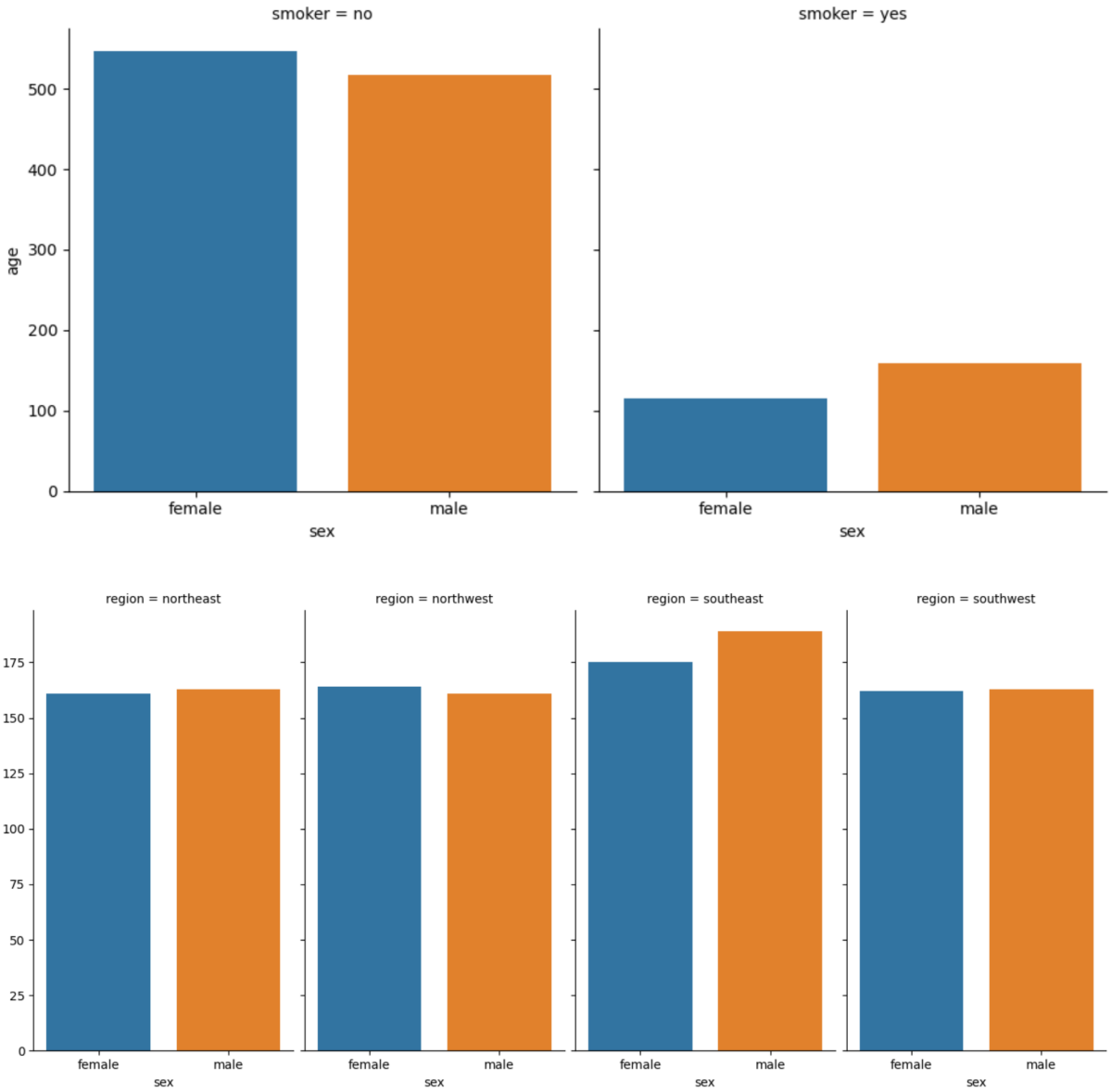
| | age | bmi | children | charges |
|---|---|---|---|---|
| count | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| mean | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| std | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| min | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| 25% | 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 |

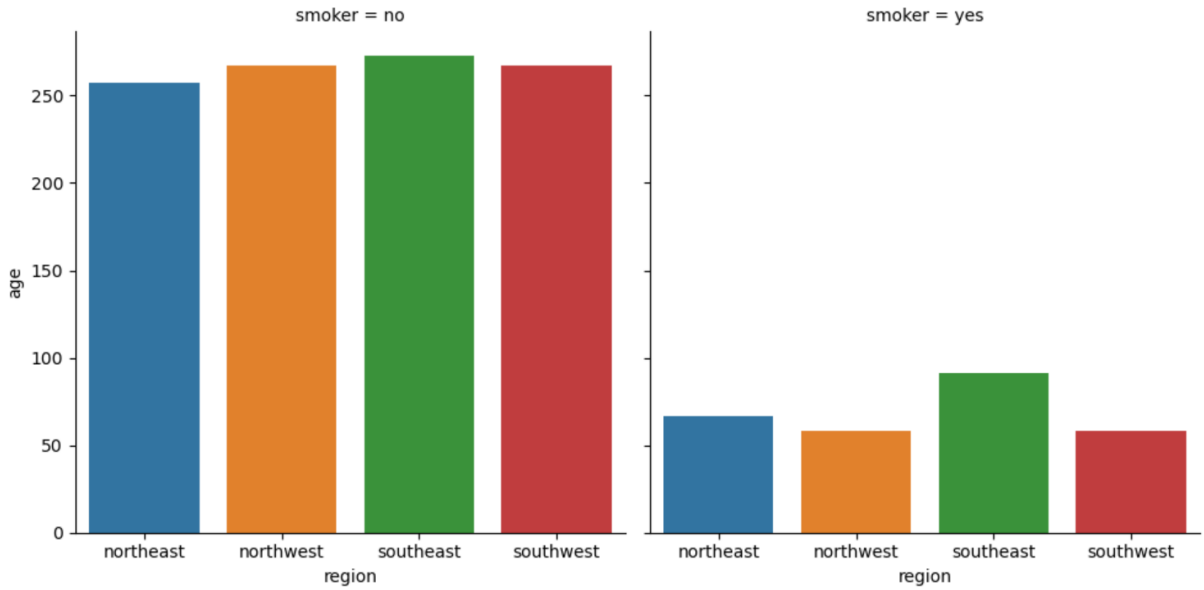
```
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
```

As you can see, there are 3 object variables: ‘sex’, ‘smoker’, and ‘region’.




The proportion of each instance is similar except for the ‘smoker’.  The proportion of non-smokers is more overwhelming than the number of smokers.






# Data Preprocessing

As I said before we have 3 object values and I'm going to split these features through One Hot Encoding.

| | age | bmi | children | ... | region_northwest | region_southeast | region_southwest |
|---|---|---|---|---|---|---|---|
| 0 | 19 | 27.900 | 0 | ... | 0 | 0 | 1 |
| 1 | 18 | 33.770 | 1 | ... | 0 | 1 | 0 |
| 2 | 28 | 33.000 | 3 | ... | 0 | 1 | 0 |
| 3 | 33 | 22.705 | 0 | ... | 1 | 0 | 0 |
| 4 | 32 | 28.880 | 0 | ... | 1 | 0 | 0 |

After this, I evaluate the model with many linear regressions.

# Model Evaluation

- Simple Linear Regression
  R2 Score: 0.71
- Polynomial

| | | |
|---|---|---|
| Polynomial | degree=2 | 0.77 |
| | degree=3 | 0.82 |
| | degree=4 | 0.66 |

- Linear Regression + LASSO Regularization

| | | |
|---|---|---|
| Linear + LASSO | alpha=0.1 | 0.71 |
| | alpha=0.05 | 0.71 |
| | alpha=0.001 | 0.71 |

- Linear Regression + Ridge Regularization

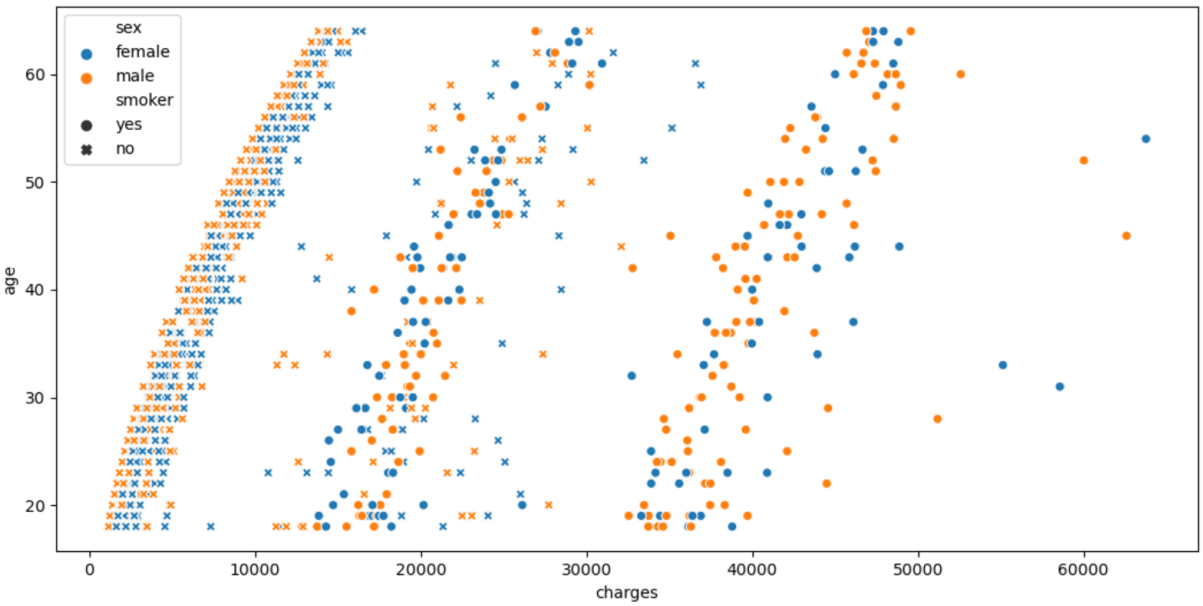
| | | |
|---|---|---|
| Linear + Ridge | alpha=0.1 | 0.71 |
| | alpha=0.05 | 0.71 |
| | alpha=0.001 | 0.71 |

- Linear Regression + Elastic Net

| | | alpha | | | |
|---|---|---|---|---|---|
| | | 0.1 | 0.05 | 0.001 | 0.0001 |
| l1_ratio | 0.1 | 0.64 | 0.69 | 0.716 | 0.716 |
| | 0.3 | 0.665 | 0.69 | 0.716 | 0.716 |
| | 0.5 | 0.686 | 0.707 | 0.716 | 0.716 |
| | 0.7 | 0.704 | 0.713 | 0.716 | 0.716 |
| | 0.9 | 0.715 | 0.716 | 0.716 | 0.716 |

As you can see I evaluated many models with different alphas and l1_ratios. The total evaluation concludes that the Polynomial Regression with the degree of 3 is the best-trained model and shows 82% accuracy in test data prediction.

# Key Findings and Insights.




The above figure is the scatter plot between medical charges/smoker/sex. As you can see the charges rarely depend on the sexual difference. We cannot find any correlation between sex and charges. However, when you see the dots and Xs in the graph you can easily find the correlation between smoking and medical charges. As you go to the left, the cost decreases, and if you look closely, you can see that most of them are non-smokers. On the contrary, the proportion of smokers is overwhelmingly higher as you go to the right. This is being shown regardless of age.

# Suggestions for the next step.

As I said earlier, it was confirmed that the relationship between smoking and medical charges is quite high. Next, I'm, planning to look at the factors that determine medical costs in more detail by comparing them with BMI, smoking, and age.