MEDICAL INSURANCE

Done by:
Nawaf Almutairi
Surayyi Alqahtani

Instructor :
Dr. Mejdal Alqahtani

OUTLINES OF PRESENTATION

1. INTRODUCTION
2. PROJECT GOAL
3. WEB SCRIBING
4. EXPLORATORY DATA ANALYSIS
5. REGRESSION MODELS
6. CONCLUSION

INTORDICATION

Health insurance is a type of insurance that covers medical expenses that arise due to an illness.

These expenses could be related to hospitalization costs, cost of medicines or doctor consultation fees.

INSURANCE

PROJECT GOAL

Problem statement/ question :
How can the insurance companies predict the cost for individual patients?

This project is aimed at giving the insurance companies a proximal prediction of costs for every individual.

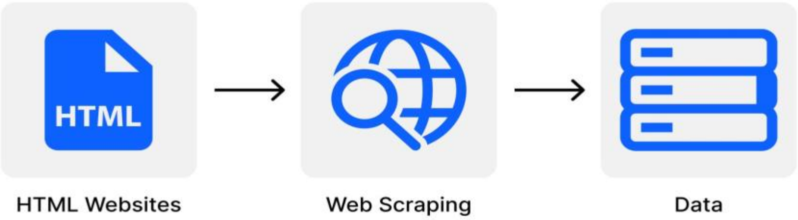
WEB SCRIBING:

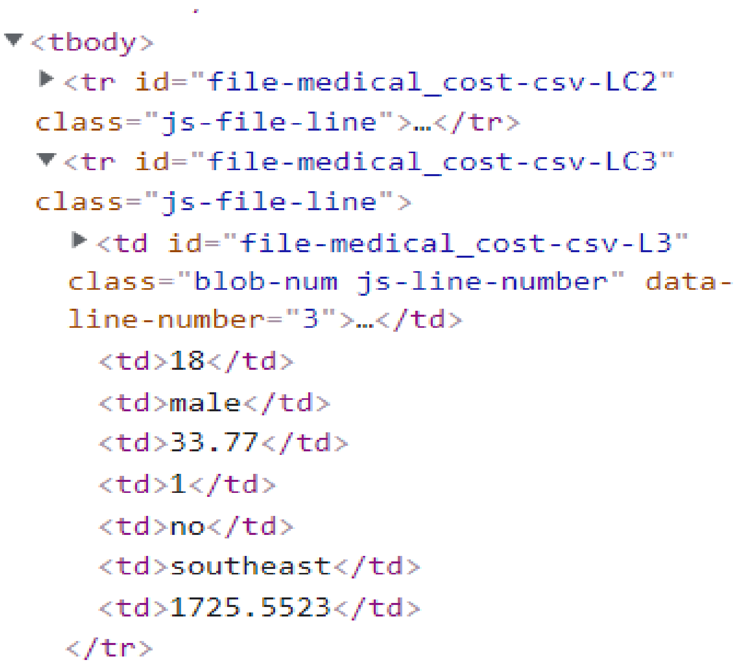
- **Web Scribing** is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format.

- The data was collated by using Web scribing on GitHub web page.

- Using Beautiful Soup and requests libraries.

- Problem that we face.





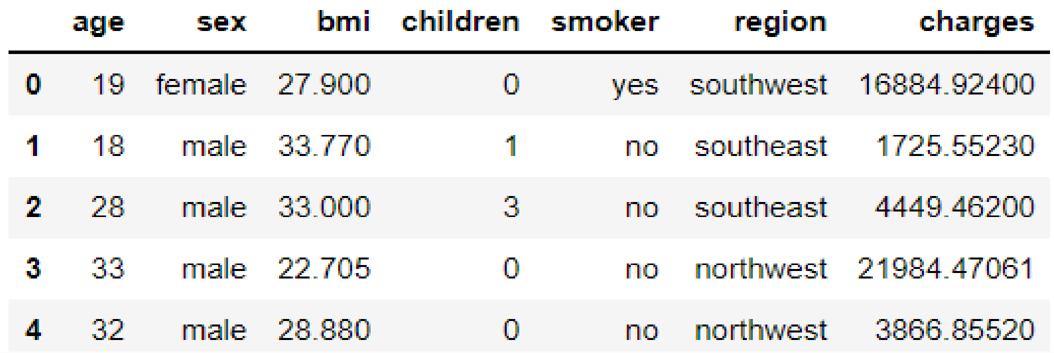
```
▼<tbody>
  ▶<tr id="file-medical_cost-csv-LC2"
  class="js-file-line">…</tr>
  ▼<tr id="file-medical_cost-csv-LC3"
  class="js-file-line">
    ▶<td id="file-medical_cost-csv-L3"
    class="blob-num js-line-number" data-
    line-number="3">…</td>
      <td>18</td>
      <td>male</td>
      <td>33.77</td>
      <td>1</td>
      <td>no</td>
      <td>southeast</td>
      <td>1725.5523</td>
    </tr>
```

EXPLORATORY DATA ANALYSIS

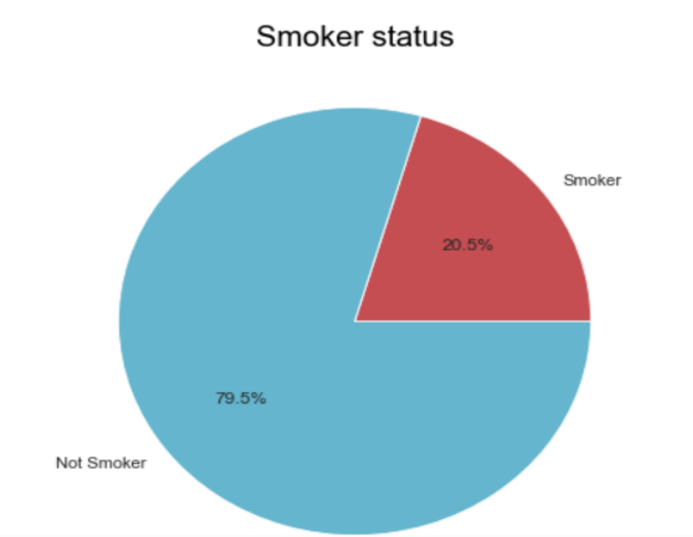
Data Structure:

It is consisted of 7 columns and 1338 rows.

df.head( )



	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

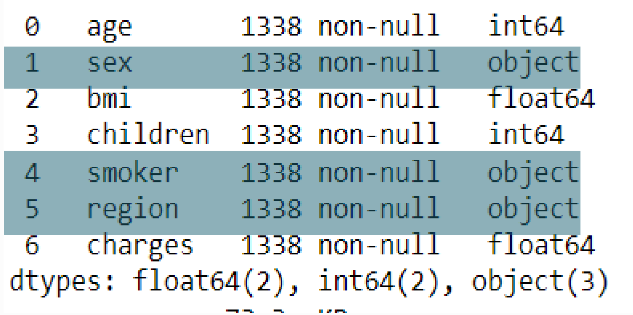


EXPLORATORY DATA ANALYSIS

Dummy Variables :

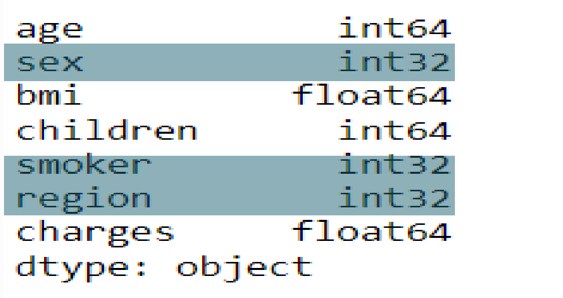
By using the label encoder from sklearn.preprocessing.

Before:



```
 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)
```

After:



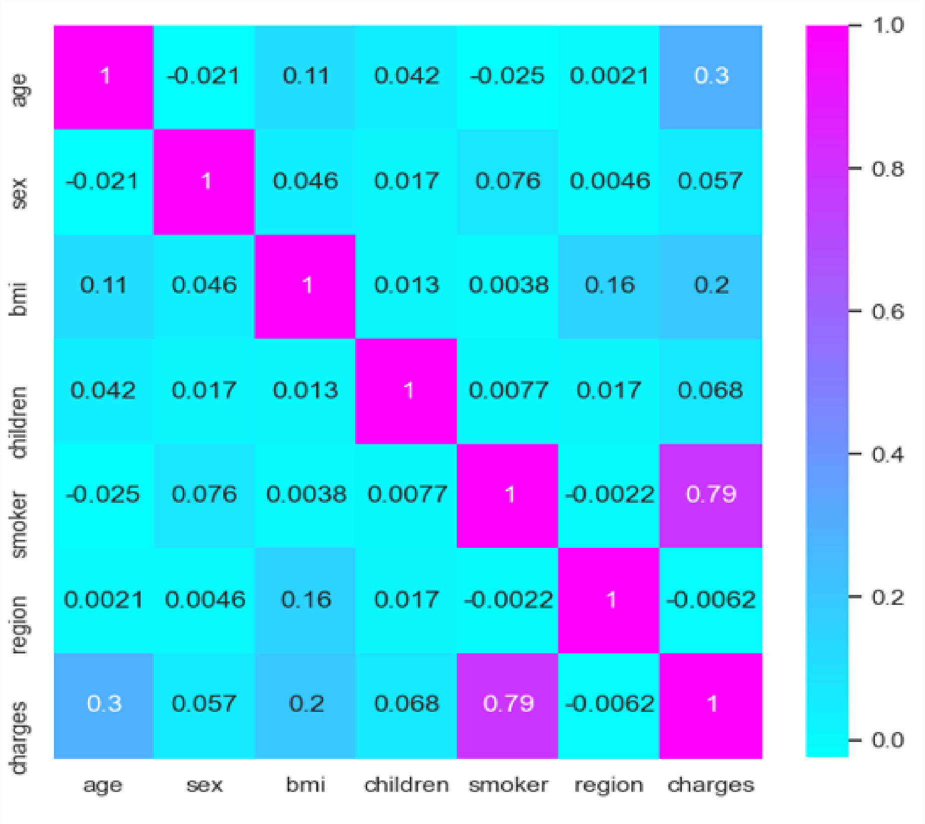
```
age             int64
sex             int32
bmi           float64
children        int64
smoker          int32
region          int32
charges       float64
dtype: object
```

EXPLORATORY DATA ANALYSIS

Heat Map :

The highest impact on the charges.

- Smoking.
- Age.
- Body Mass Index (BMI).



	age	sex	bmi	children	smoker	region	charges
age	1	-0.021	0.11	0.042	-0.025	0.0021	0.3
sex	-0.021	1	0.046	0.017	0.076	0.0046	0.057
bmi	0.11	0.046	1	0.013	0.0038	0.16	0.2
children	0.042	0.017	0.013	1	0.0077	0.017	0.068
smoker	-0.025	0.076	0.0038	0.0077	1	-0.0022	0.79
region	0.0021	0.0046	0.16	0.017	-0.0022	1	-0.0062
charges	0.3	0.057	0.2	0.068	0.79	-0.0062	1

REGRESSION MODELS

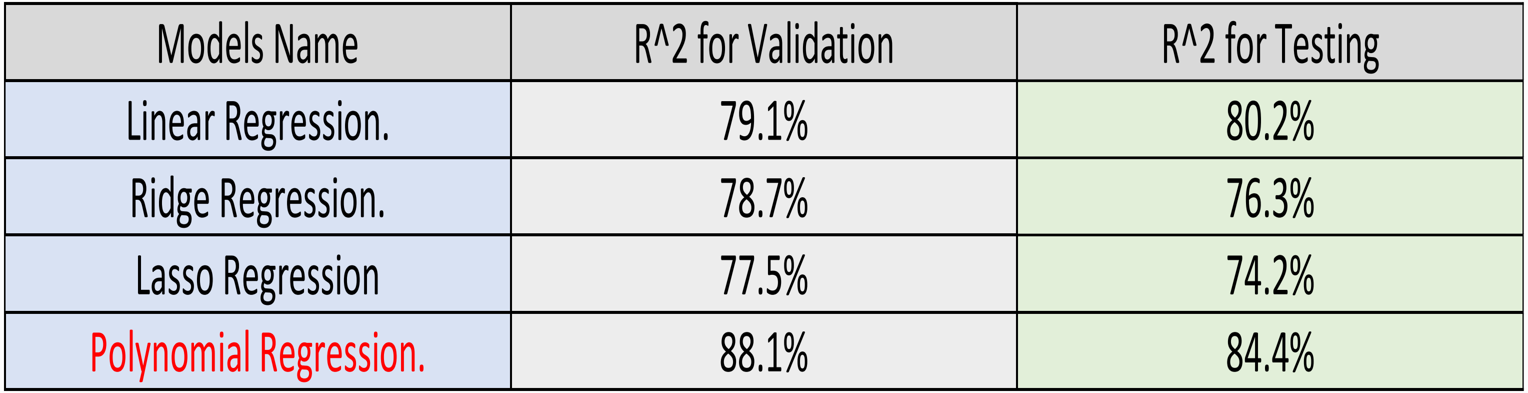
Regression is the relationships between a dependent variable (y) and one or more independent variables (x).

Different types of regression :

- ✓ Linear Regression.
- ✓ Ridge Regression.
- ✓ Lasso Regression.
- ✓ Polynomial Regression.

REGRESSION MODELS

R-squared (R2) for each models :



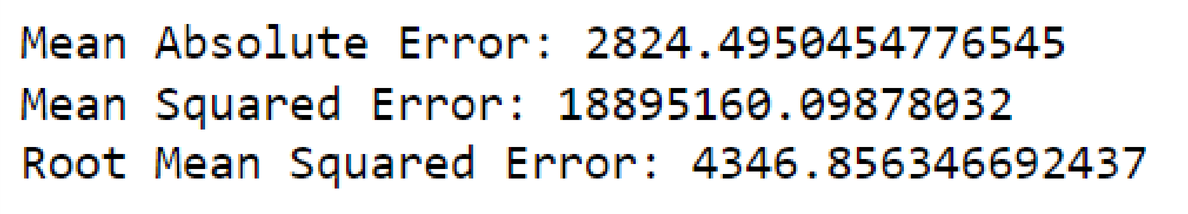
Models Name	R^2 for Validation	R^2 for Testing
Linear Regression.	79.1%	80.2%
Ridge Regression.	78.7%	76.3%
Lasso Regression	77.5%	74.2%
Polynomial Regression.	88.1%	84.4%

POLYNOMIAL REGRESSION

Polynomial is the best model:

- We drop some columns such as region and gender, so we have high focus on important features to increase R^2.

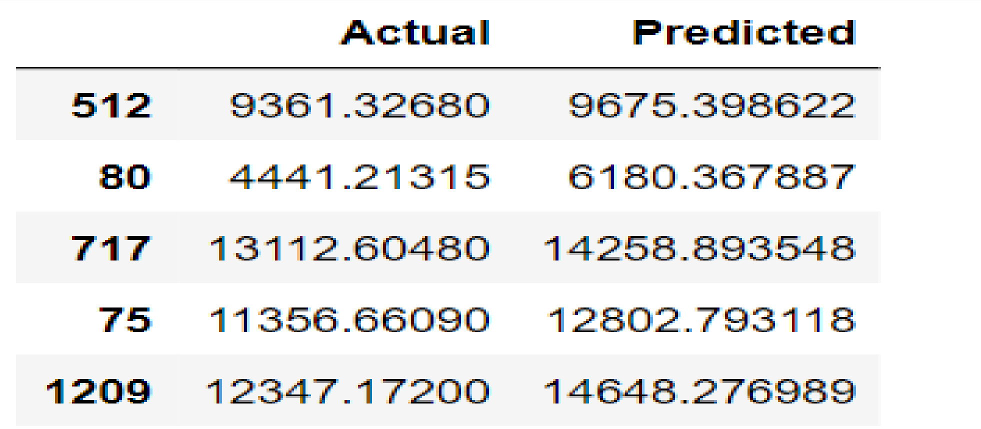
- The degree is (2).



```
Mean Absolute Error: 2824.4950454776545
Mean Squared Error: 18895160.09878032
Root Mean Squared Error: 4346.856346692437
```

THE PREDICATION

The different between actual and predicted values:



	Actual	Predicted
512	9361.32680	9675.398622
80	4441.21315	6180.367887
717	13112.60480	14258.893548
75	11356.66090	12802.793118
1209	12347.17200	14648.276989

The equation: $y = mx + b$

y = -5325.88 + [-4.01606591e+01 5.23702019e+02 8.52025026e+02-9.52698471e+03 3.04430186e+00 1.84508369e+00 6.01720286e+00 4.20849790e+00 -9.38983382e+00 3.81612289e+00 1.40840670e+03 -1.45982790e+02 -4.46151855e+02 -9.52698471e+03]

CONCLUSION

- **Health insurance** is a type of insurance that covers medical expenses that arise due to an illness.
- Smoking has the highest impact on medical costs, even though the costs are growing with age, bmi and children.
- We use some models to find the best R^2 and the Polynomial Regression turned out to be the best model.
- In the future we will try to improve the R^2 and minimize the error.

THENK YOU FOR LISINING