Project Title: Healthcare Insurance

Here is the link to the project completed by WILLIAN OLIVEIRA GIBIN. He completed a very comprehensive analysis of the data.

https://www.kaggle.com/datasets/willianoliveiragibin/healthcare-insurance/data

He separates the work into 6 components:

1. Data review
2. Visualization
3. Feature engineering
4. Modelling: 5 models are used
    a. Linear regression
    b. Random vector regressor
    c. Support vector regressor
    d. Decision tree regressor
    e. Gaussian process regressor
5. Testing model
6. Conclusion

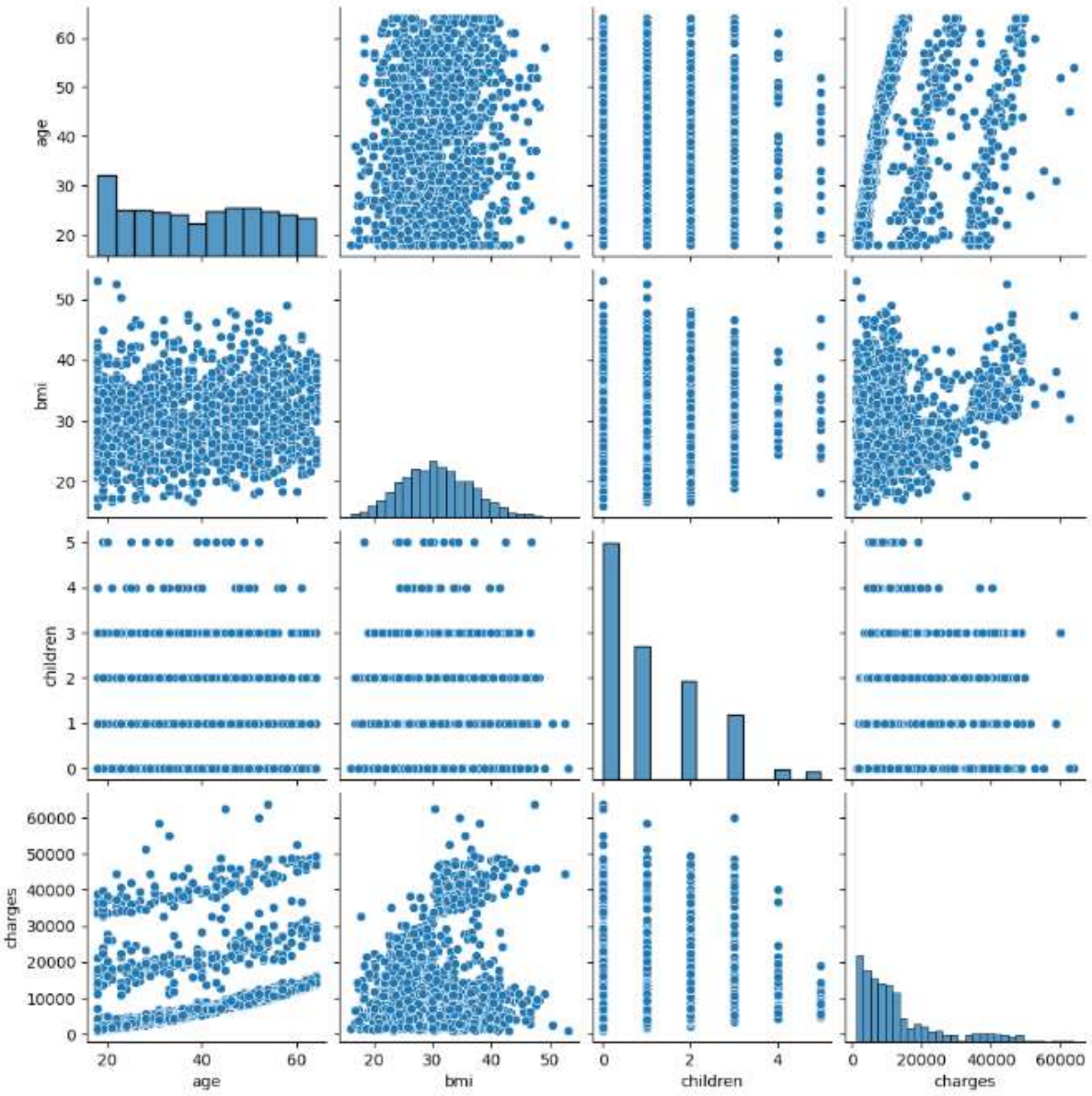
I think in the data review section, there should be some checking to be done, such as missing values and outliers.

In [4]:
```
#Exploring the Discrete Statistics of the Data Set
data.describe(include = 'all')
```

Out[4]:

	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.663397	1.094918	NaN	NaN	13270.422265
std	14.049960	NaN	6.098187	1.205493	NaN	NaN	12110.011237
min	18.000000	NaN	15.960000	0.000000	NaN	NaN	1121.873900
25%	27.000000	NaN	26.296250	0.000000	NaN	NaN	4740.287150
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.033000
75%	51.000000	NaN	34.693750	2.000000	NaN	NaN	16639.912515
max	64.000000	NaN	53.130000	5.000000	NaN	NaN	63770.428010

In the visualization section, he used the following code to visualize the data: sns.pairplot(data). It's very powerful and produces the charts of various interactions.



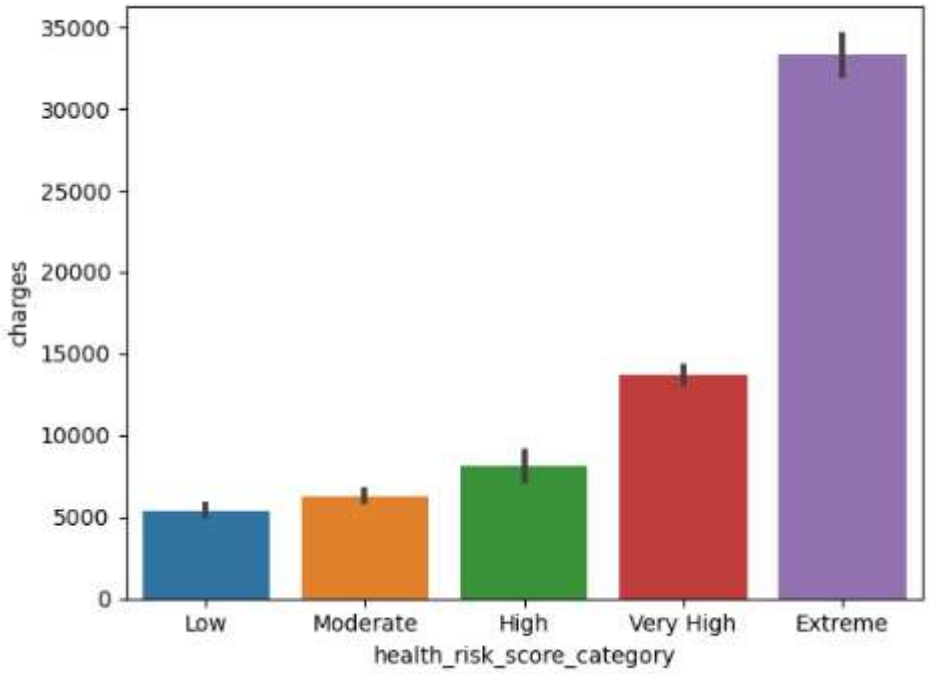
In the feature engineering section, he groups age, BMI and children into new categories. A new feature named health risk based on smoking status, BMI, and medical problem is created.

Another great thing I learned from the code is to create the heatmap of the correlation: sns.heatmap(cor, annot=True, cmap=plt.cm.Reds)

```python
def calculate_health_risk(row):
    score = 0

    # Assign scores based on bmi_NEW
    if row['bmi_NEW'] == 'Underweight':
        score += 0.3
    elif row['bmi_NEW'] == 'Normal':
        score += 0
    elif row['bmi_NEW'] == 'Overweight':
        score += 0.3
    elif row['bmi_NEW'] in ['Obesity I', 'Obesity II', 'Extreme Obesity']:
        score += 0.9

    # Assign scores based on medical_problem
    if row['medical_problem'] == 'Negligible':
        score += 0
    elif row['medical_problem'] == 'Mild':
        score += 0.4
    elif row['medical_problem'] == 'Moderate':
        score += 0.8
    elif row['medical_problem'] == 'Severe':
        score += 1.2

    # Assign scores based on smoker
    if row['smoker'] == 'yes':
        score += 0.8
    elif row['smoker'] == 'no':
        score += 0

    return score

data['health_risk_score'] = data.apply(calculate_health_risk, axis=1)

# Calculate maximum and minimum health risk scores
max_score = data['health_risk_score'].max()
min_score = data['health_risk_score'].min()

# Calculate percentage of health risk score
score_range = max_score - min_score
```

In the modelling section, he rans 5 different models using the categorized values and new features. Clearly, the health risk will increase the health charge.

One thing he can improve in the section is to compare the performance of 5 different values.



Another potential model that he can try is the polynomial regression. It runs the model on higher order of regression variables.

Finally, he mentions in the conclusion that “Feature engineering is an essential and very important part and plays a major role in increasing the rates.”. It’s a correct statement but it can be expanded, such as what’s the major driver of the health charge, how general population can reduce the charge, what models works best for the dataset.