CS521 C1 Information Structures with Python (Summer 1 2022)
Name: Rong Jiang (Hanna)

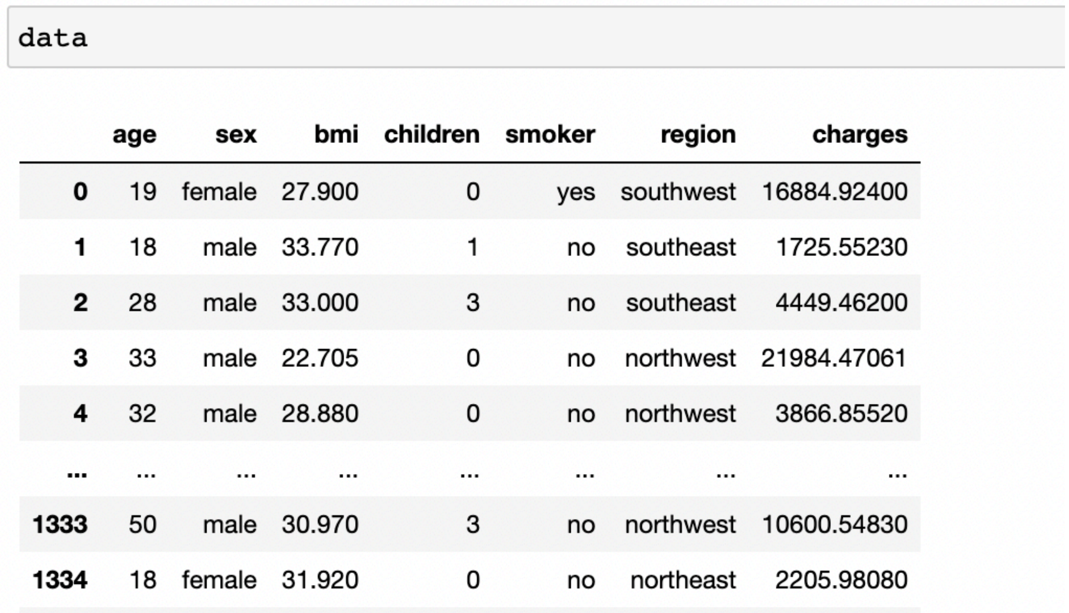
Final Project:
Predicting medical expenses using linear regression

Introduction

In order for a health insurance company to be profitable, it needs to collect appropriate yearly premiums and more than it spends on medical care to its beneficiaries. Therefore, insurers invest a great deal of time and money in developing models that accurately forecast medical expenses for the insured population.

Medical expenses are difficult to estimate because the most costly conditions are infrequent and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical care expenses for population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs. I found Medical Cost Personal Datasets. I am curious how these variables affect the charges. Since we are



```
data
```

	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
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080

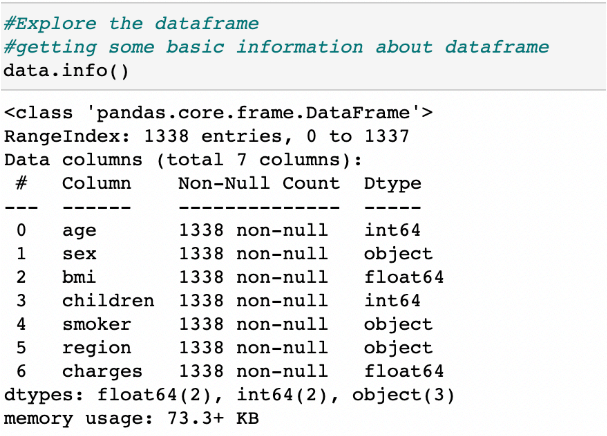
interested in medical expenses, we choose the expenses column as the dependent variable and explore it more. I used multiple linear regression method in this project.

The Dataset

The dataset consists of 7 variables.

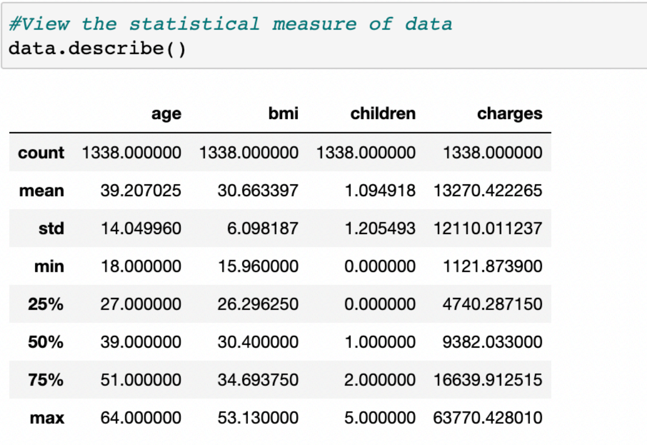
Dependent variables I used in project: charges(numerical variables)
Independent variables: age, sex(male/female), bmi, smoker(yes/no), children, region(southeast, southwest, northeast, northwest).
Among these independent variables: age, bmi, children, charges are numerical variables while smoker, region and sex are categorical variables.

It is important to give some thought to how these variables may be related to billed medical expenses. For instance, we might expect that older people and smokers are at higher risk of large medical expenses.



```
#Explore the dataframe
#getting some basic information about dataframe
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
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)
memory usage: 73.3+ KB
```

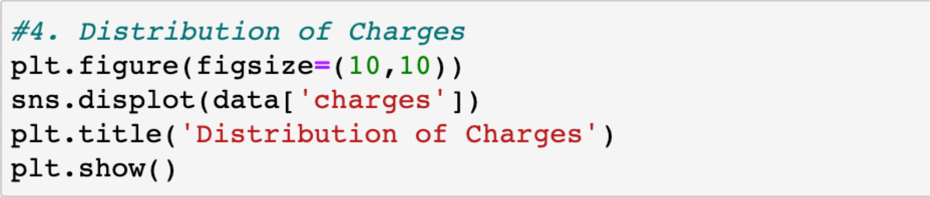


```
#View the statistical measure of data
data.describe()
```

	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

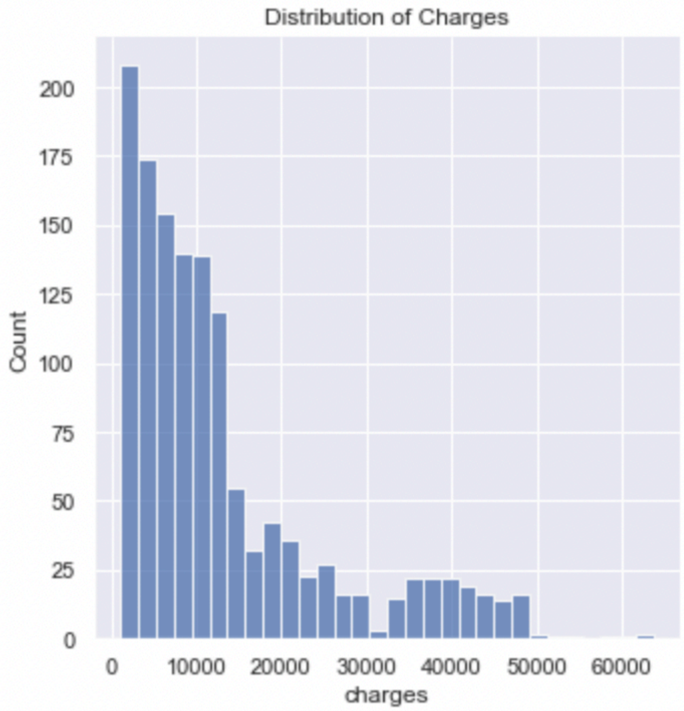
Data Exploration

Our model's dependent variable is expenses, which measures the medical costs each person charged to the insurance plan for the year. Prior to building a regression model, it is often helpful to check for normality. Although linear regression does not strictly require a normally distributed dependent variable, the model often fits better when this is true. Let's take a look at the summary statistics:
Because the mean value is greater than the median, this implies that the distribution of insurance expenses is right-skewed. We can confirm this visually using a histogram and the output is shown as follows:



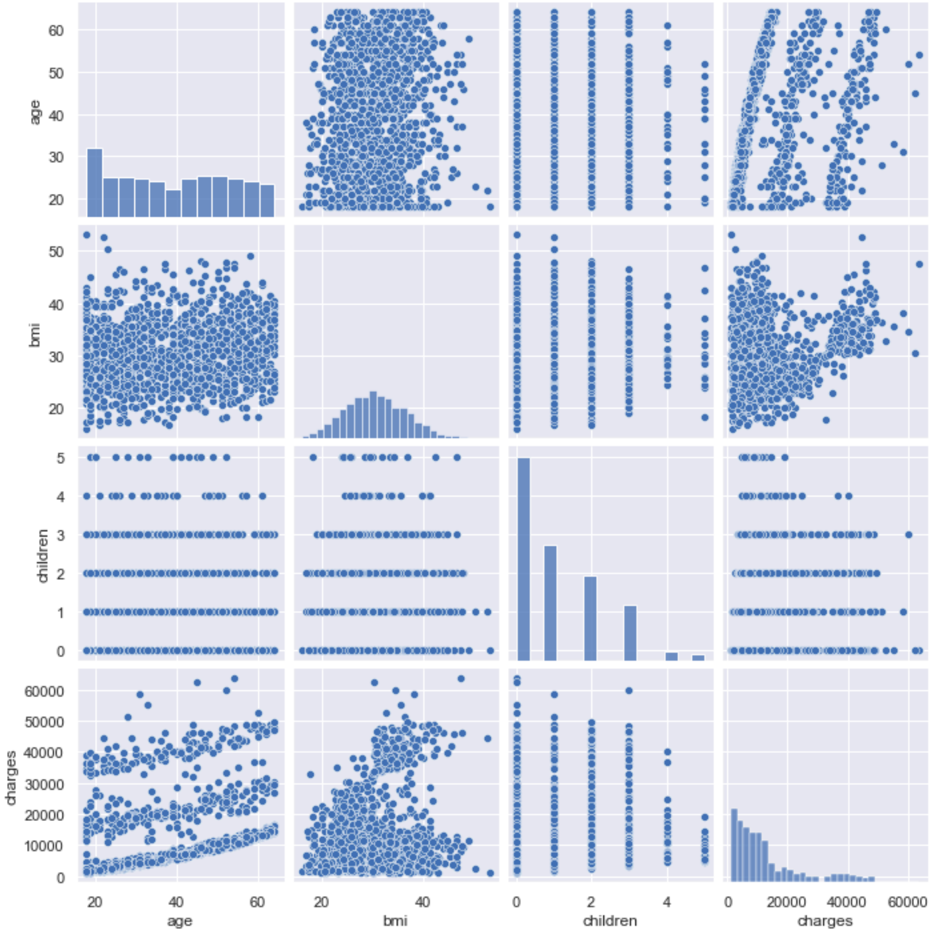
```
#4. Distribution of Charges
plt.figure(figsize=(10,10))
sns.displot(data['charges'])
plt.title('Distribution of Charges')
plt.show()
```

```
<Figure size 720x720 with 0 Axes>
```



Visualizing relationships among features

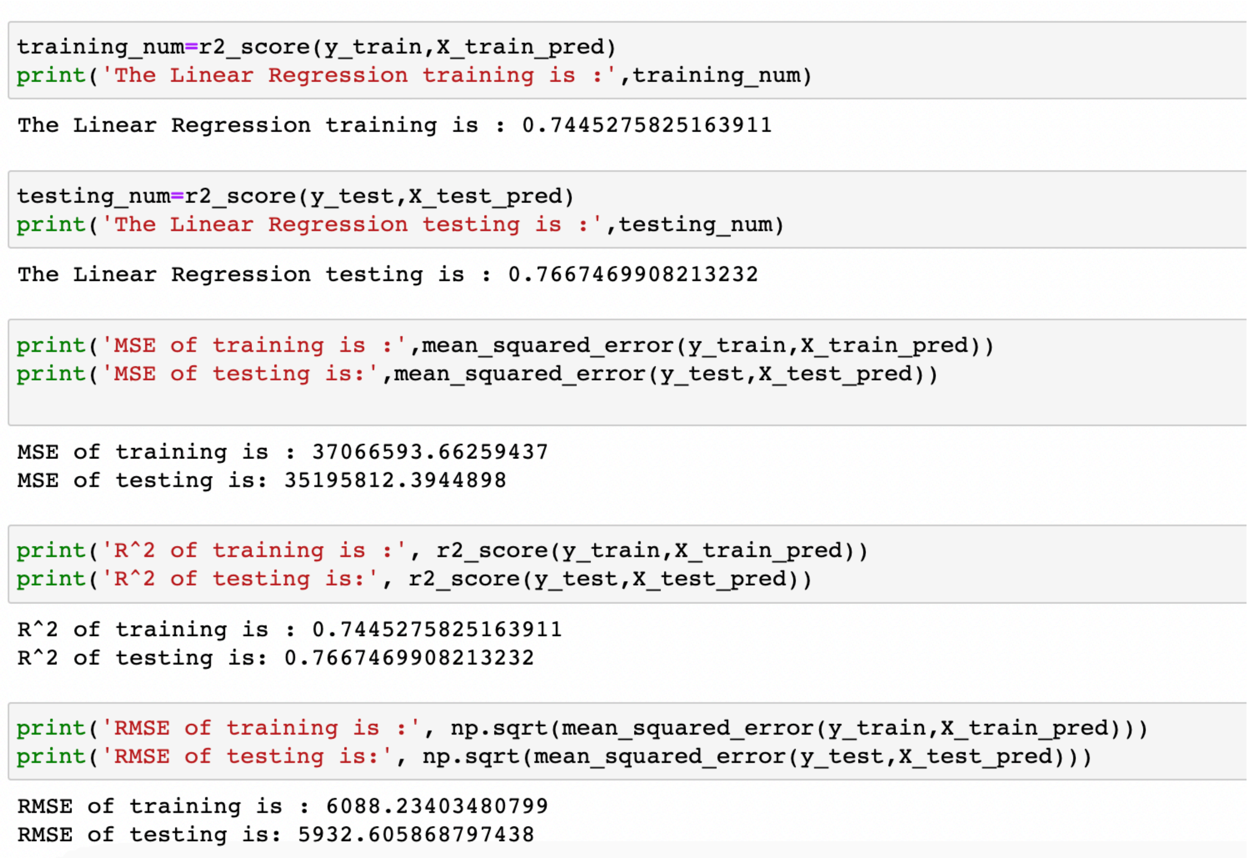
It can also be helpful to visualize the relationships among numeric features by using a scatterplot. Although we could create a scatterplot for each possible relationship, doing so for a large number of features might become tedious.



Split the data and train a model on the data, evaluate model performance

Dummy coding allows a nominal feature to be treated as numeric by creating a binary variable, often called a dummy variable, for each category of the feature.
The dummy variable is set to 1 if the observation falls into the specified category or 0 otherwise. For instance, the sex feature has two categories: male and female.
The same coding applies to variables with three or more categories. For example, four-category feature region into four dummy variables: regionnorthwest, regionsoutheast, regionsouthwest, and regionnortheast.

The results of the linear regression model make logical sense: old age, smoking, and obesity tend to be linked to additional health issues, while additional family member dependents may result in an increase in physician visits and preventive care such as vaccinations and yearly physical exams. However, we currently have no sense of how well the model is fitting the data.



```
training_num=r2_score(y_train,X_train_pred)
print('The Linear Regression training is :',training_num)
```

```
The Linear Regression training is : 0.7445275825163911
```

```
testing_num=r2_score(y_test,X_test_pred)
print('The Linear Regression testing is :',testing_num)
```

```
The Linear Regression testing is : 0.7667469908213232
```

```
print('MSE of training is :',mean_squared_error(y_train,X_train_pred))
print('MSE of testing is:',mean_squared_error(y_test,X_test_pred))
```

```
MSE of training is : 37066593.66259437
MSE of testing is: 35195812.3944898
```

```
print('R^2 of training is :', r2_score(y_train,X_train_pred))
print('R^2 of testing is:', r2_score(y_test,X_test_pred))
```

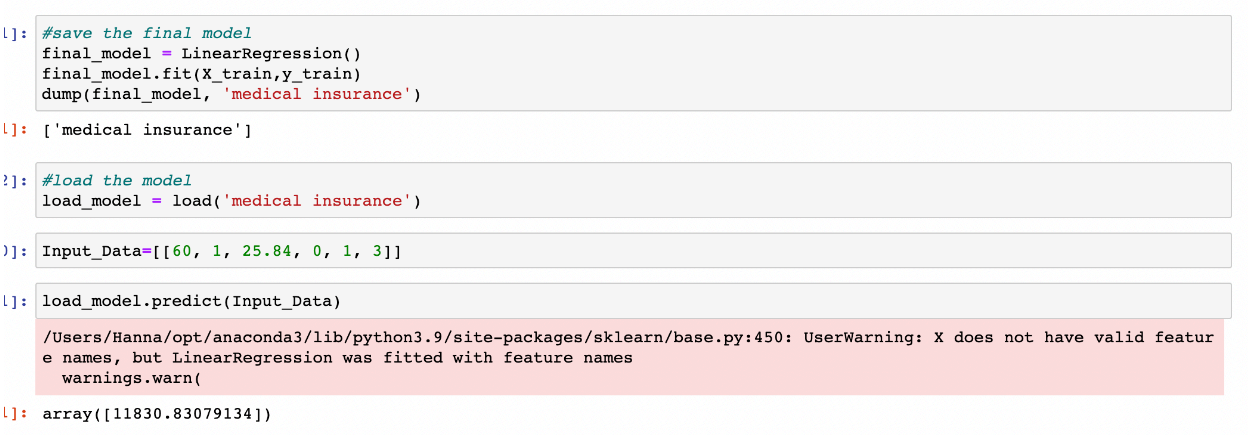
```
R^2 of training is : 0.7445275825163911
R^2 of testing is: 0.7667469908213232
```

```
print('RMSE of training is :', np.sqrt(mean_squared_error(y_train,X_train_pred)))
print('RMSE of testing is:', np.sqrt(mean_squared_error(y_test,X_test_pred)))
```

```
RMSE of training is : 6088.23403480799
RMSE of testing is: 5932.605868797438
```

Build a predictive model

For example, if we have a patient whose age is 60, gender is female, BMI is 25.84, 0 children, is not a smoker and lives in northeast area. This patient's expected insurance cost is 11830.8307913.



```
l]: #save the final model
    final_model = LinearRegression()
    final_model.fit(X_train,y_train)
    dump(final_model, 'medical insurance')

l]: ['medical insurance']

2]: #load the model
    load_model = load('medical insurance')

)]: Input_Data=[[60, 1, 25.84, 0, 1, 3]]

l]: load_model.predict(Input_Data)

    /Users/Hanna/opt/anaconda3/lib/python3.9/site-packages/sklearn/base.py:450: UserWarning: X does not have valid featur
    e names, but LinearRegression was fitted with feature names
      warnings.warn(

l]: array([11830.83079134])
```

Conclusion:

Smoking is the biggest indicator of increased charges billed by medical insurance. Also, if BMI is high, it also increases the charges.

We also see that age is an important indicator for higher charges. Because as we get older, our physical condition would decline naturally. However, we can quit smoking and lower the BMI to the healthy level.

Not only we can use multiple linear regression model, but also I found it's also helpful if I use random forest regression model. I will practice and apply this model to the dataset when I have more time.