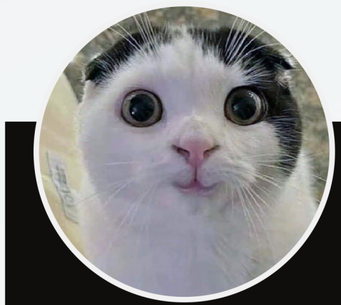
Medical Insurance

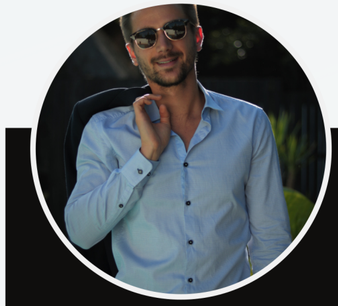
Instructors: Dr. Phauk Sokkey (Course)
Mr. Nhim Malai (TD)

CONTENT

OUR TEAM: GROUP9

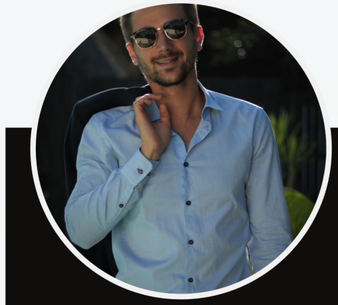


NANG
Sreynich
e20200447



BUTH
khemra
e20201690

VANNAK
Vireakyuth
e20200170



SENG
Panharith
e20200639

LONG
Channleap
e20200386

HOK
Ratanak
e20201106

PROJECT SCOPING

Medical expenses is one of the major expenses in a human life. It's a common that one life style and various physical parameters dictates diseases. According various studies, major factors that contribute to higher expenses in personal medical care include smoking, age, BMI, children, region, charges, sex.

The goal of this project, to help insurance companies to set insurance premiums that are fair and profitable. By understanding the trends in the population segments, insurance companies can better estimate the average medical care expenses for each segment. This information can then be used to set premiums that are fair to all segments and that are also profitable for the insurance company.

Data Preparation & EDA

- Define the Categorical & Numerical Data
- Statistical-Description
- Data-Visualization

Define the Categorical and Numerical 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

```
<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
```

Type of variable :

- Categorical variables : sex, smoker, region
- Quantitative variables : age, bmi, charges, children. Here children is a discrete variable where as age, bmi, and charges are continuous variables.
- There are no missing values

Statistical -Description of Data

	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

***Observation*:**

- Average age of the primary beneficiary is 39.2 and maximum age is 64
- Average BMI is 30.66, that is out of normal BMI range, Maximum BMI is 53.13
- Average medical costs billed to health insurance is 13270 $, median is 9382 $ and maximum is 63770 $
- Customer on an average has 1 child.
- For Age, BMI, children , mean is almost equal to median , suggesting data is normally distributed

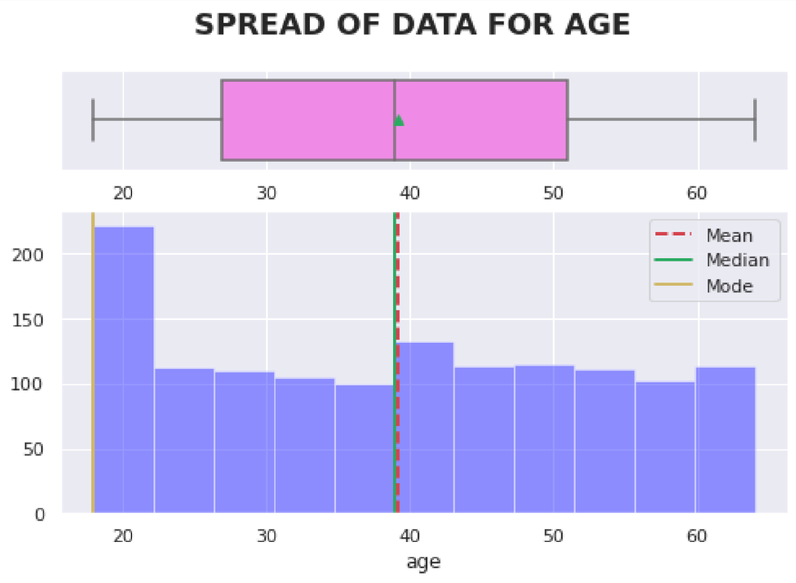
Data Preparation

```
male       676
female     662
Name: sex, dtype: int64
no      1064
yes      274
Name: smoker, dtype: int64
southeast    364
northwest    325
southwest    325
northeast    324
Name: region, dtype: int64
```

Observations

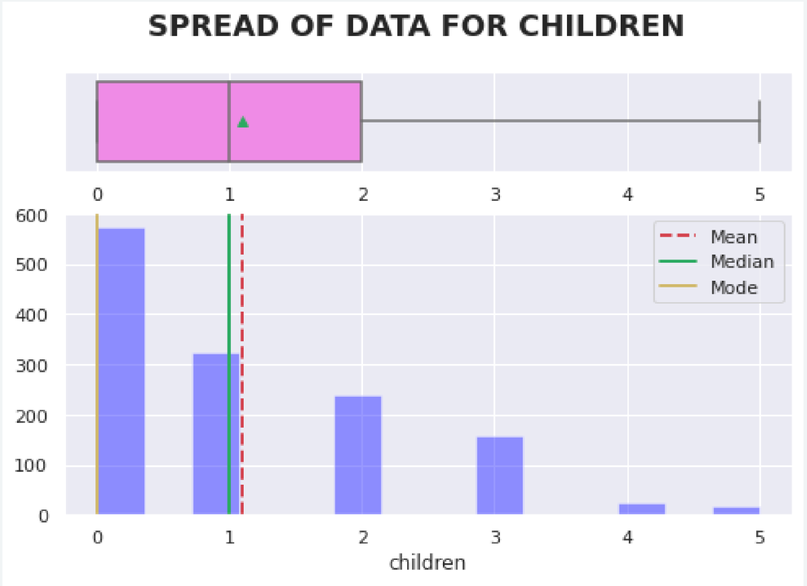
- 676 male and 662 female, indicated sample has slightly more males than females.
- 1064 nonsmoker and 274 smoker, indicating sample has more nonsmokers.
- Number of claims from customer who reside in southwest region is more compared to other regions

Data-Visualization



Age Distribution

The age of the primary beneficiary ranges from approximately 20 to 65 years old. The average age is about 40 years old. The majority of the customers are in their 18s and 20s.

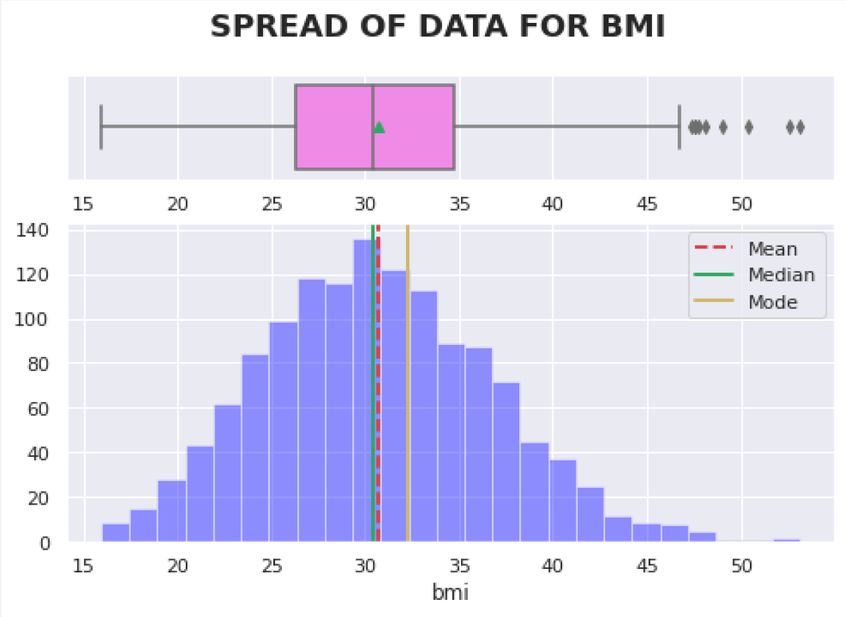


Number of Children Distribution

Most of the beneficiaries do not have children.

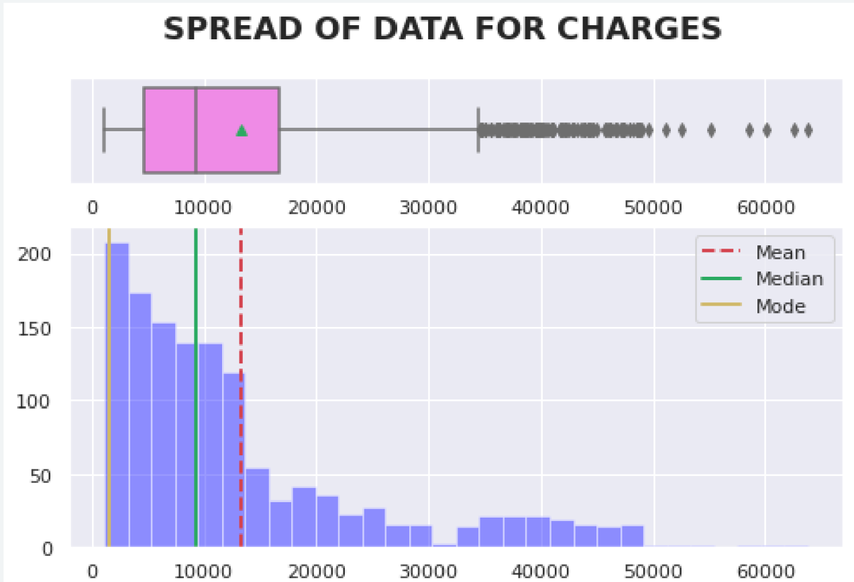
Data-Visualization

Numerical



BMI Distribution

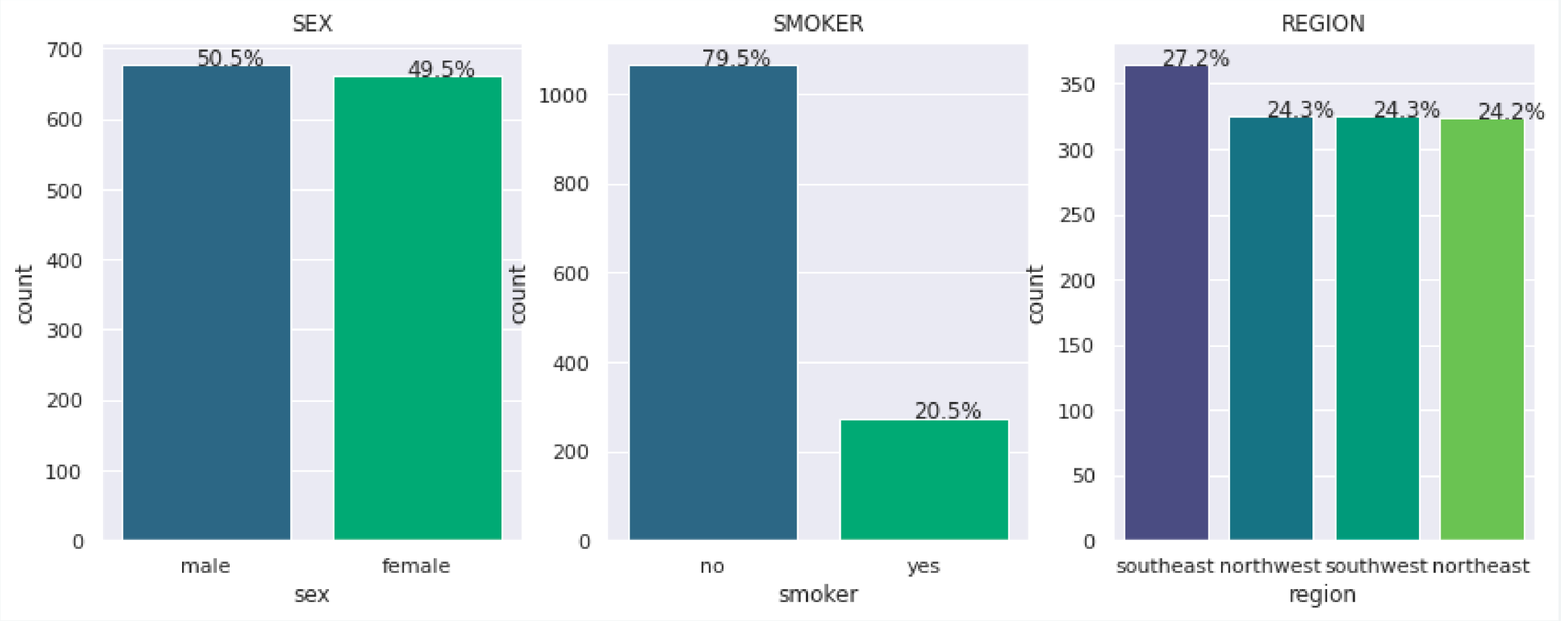
The BMIs of the beneficiaries are normally distributed, with an average BMI of 30. This BMI is outside the normal range of BMIs. There are a lot of outliers at the upper end.



Charge Distribution

The distribution of charges is unimodal and right-skewed. The average cost incurred to the insurance is approximately $130,000. The highest charge is $63,770. There are a lot of outliers at the upper end.

Data-Visualization
Categorical



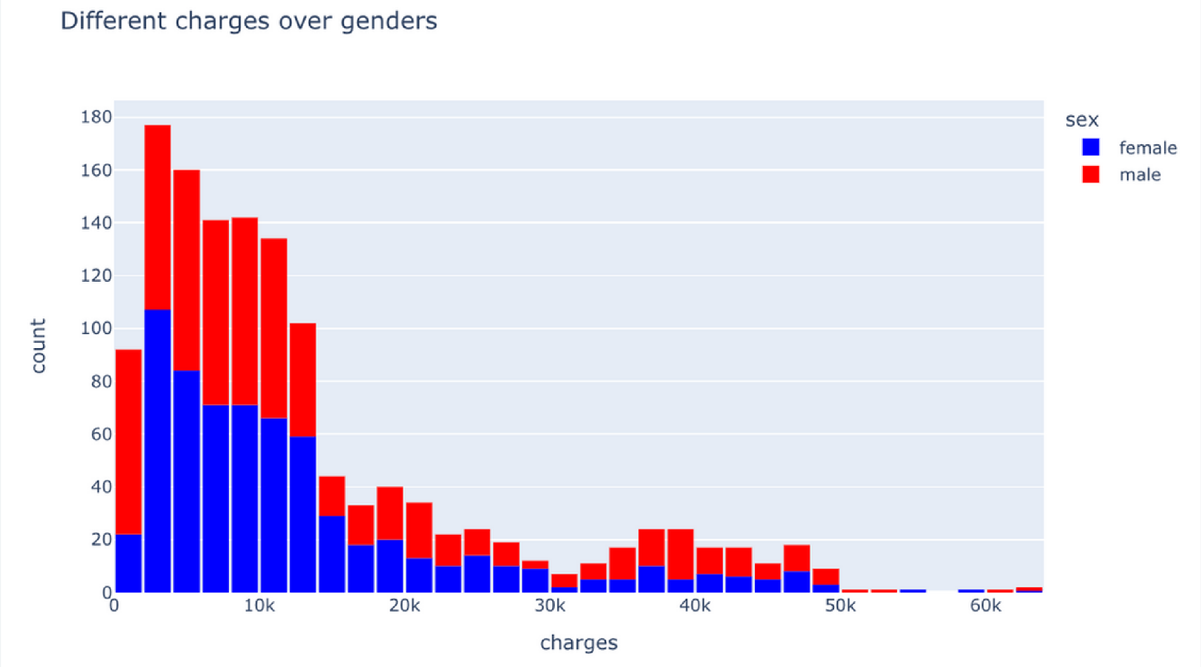
Observations

- 50.5% of beneficiaries are male and 49.5 % are female. Approximately same number of male and female beneficiary

- 20.5% of beneficiaries are smokers.

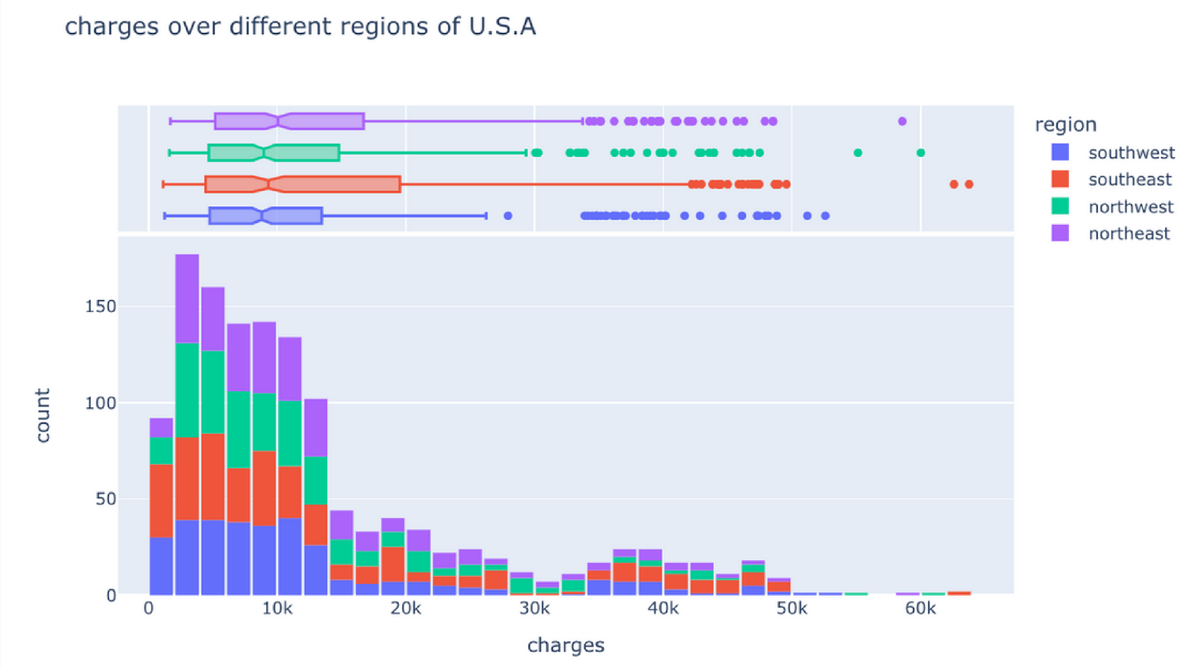
- Beneficiaries are evenly distributed across regions with South East being the most populous one (~27%) with the rest of regions each containing around ~24%

Bivariate & Multivariate Analysis



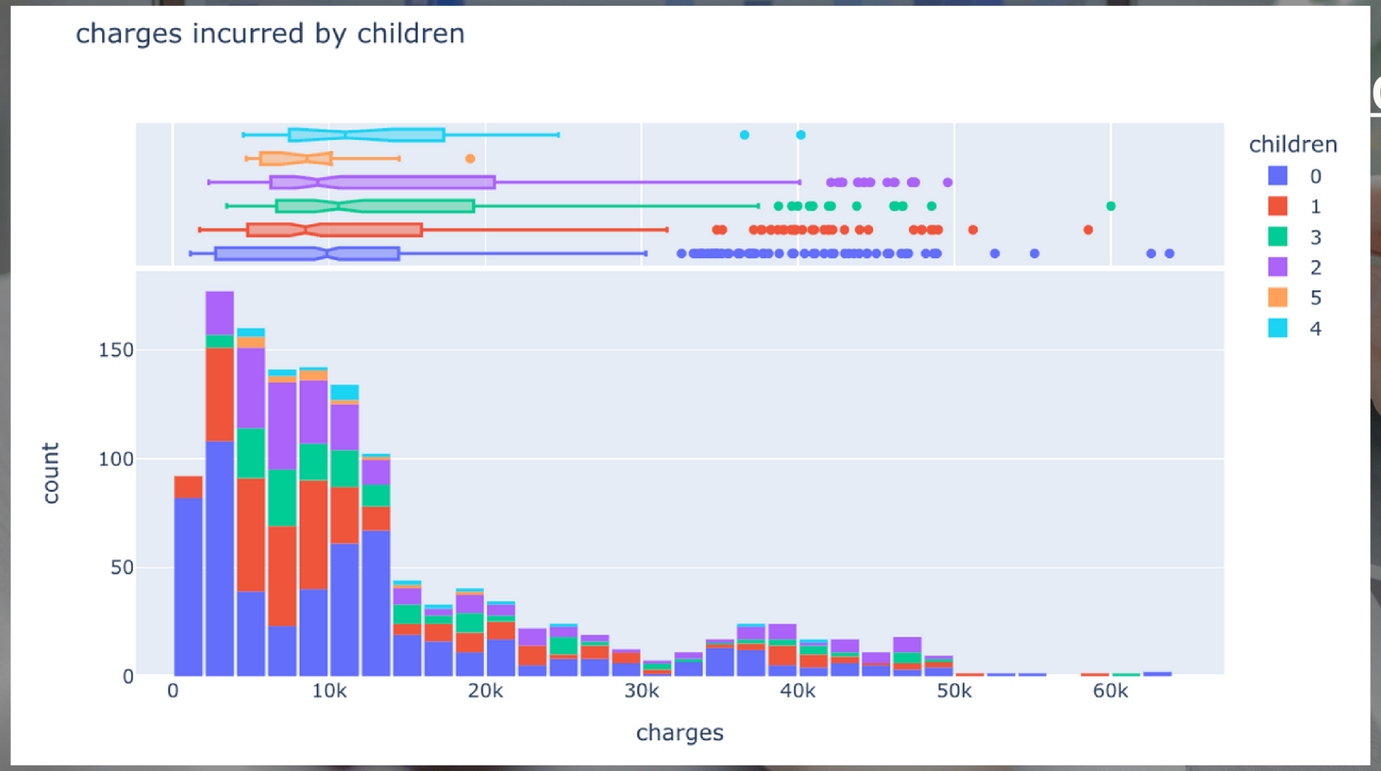
- **Distribution of charges over gender**: Males are charged substantially more than females for health insurance.
- **Subconscious behavio**r: Males are more likely to take risks than females, which may be due to subconscious behavioral patterns.

Bivariate & Multivariate Analysis



Location: The southeastern rigion has the highest average charges.

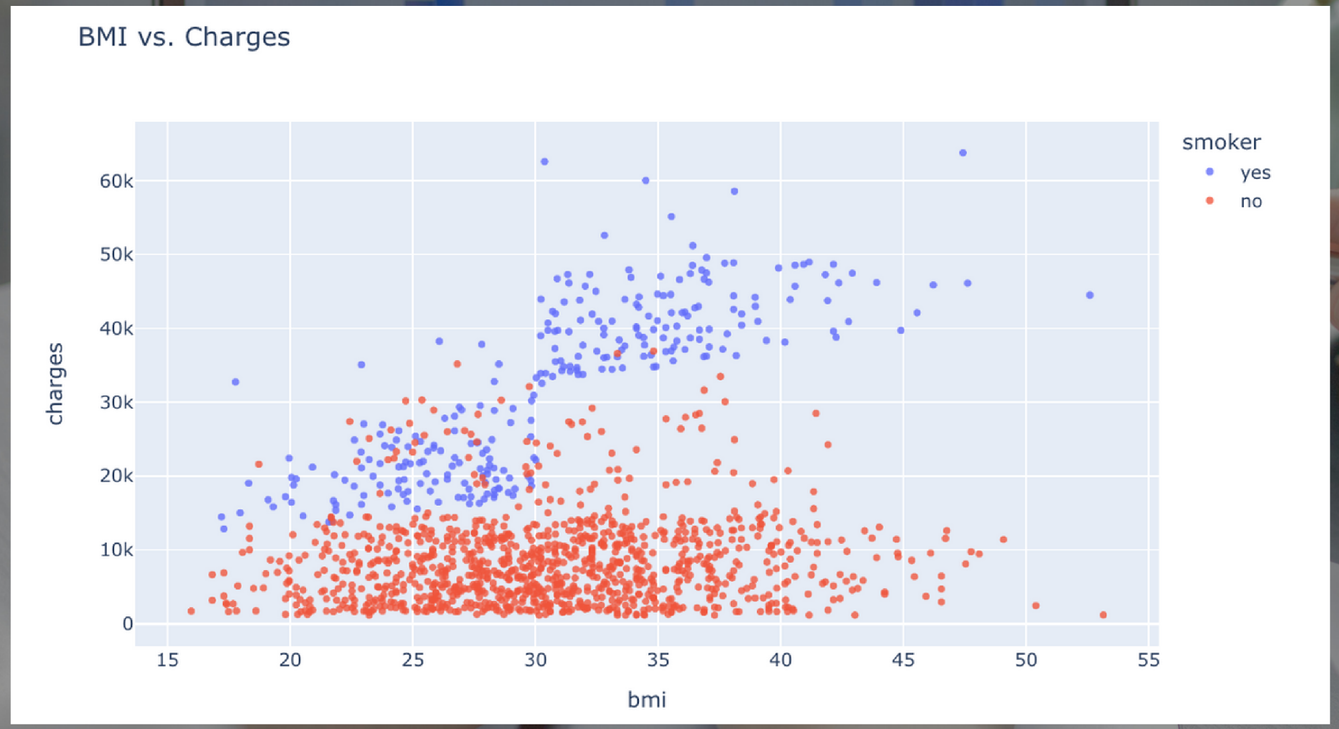
Bivariate & Multivariate Analysis



Observation:

- **Customers with 0 or 1 child**: The majority of customers have 0 or 1 child.
- **Median charges**: The median charges vary between 8.5k and 11k dollars. This suggests that most customers have relatively moderate medical expenses.

Exploratory Data Analysis
Bivariate & Multivariate Analysis



Observation:

- **BMI and medical charges for non-smokers:** For non-smokers, there is no clear relationship between BMI and medical charges. This suggests that BMI is not a significant predictor of medical charges for non-smokers.
- **BMI and medical charges for smokers**: For smokers, there is a clear relationship between BMI and medical charges. Smokers with a BMI greater than 30 have significantly higher medical charges than smokers with a BMI below 30.

Conclusion based on EDA

- As expected , as the age of the beneficiary increases ,the cost to insurance increases.
- Males who smoke have most claims and have higher bills.
- Female who are nonsmoker also have more claims to nonsmoker males this may be because of child birth , need to explore claims type to understand better.
- Customer with BMI >30 are on higher side of obesity, have more health issues and have higher claims.
- Females with BMI more than 45 have billed higher to insurance.
- Age, BMI and Smoking are important attributes which can cost insurance company more.

1st Fitting with Multiple Linear Regression

- By RFR (Random Forest Regression)
- Model Fitting
- Result

Multiple Linear Regression with 1st Method

```
from sklearn.model_selection import train_test_split as holdout
from sklearn.linear_model import LinearRegression
from sklearn import metrics
x = df.drop(['charges'], axis = 1)
y = df['charges']
x_train, x_test, y_train, y_test = holdout(x, y, test_size=0.2, random_state
=0)
Lin_reg = LinearRegression()
Lin_reg.fit(x_train, y_train)
print(Lin_reg.intercept_)
print(Lin_reg.coef_)
print(Lin_reg.score(x_test, y_test))
```

Result:

```
-11661.983908824413
[  253.99185244   -24.32455098   328.40261701   443.72929547
 23568.87948381  -288.50857254]
0.7998747145449959
```

The result we got is good enough, but we can try to improve it a bit by reducing unimportant features later.

Random Forest Regression with 1 Method

```
from sklearn.ensemble import RandomForestRegressor as rfr
x = df.drop(['charges'], axis=1)
y = df.charges
Rfr = rfr(n_estimators = 100, criterion = 'mse',
                              random_state = 1,
                              n_jobs = -1)
Rfr.fit(x_train,y_train)
x_train_pred = Rfr.predict(x_train)
x_test_pred = Rfr.predict(x_test)

print('MSE train data: %.3f, MSE test data: %.3f' %
      (metrics.mean_squared_error(x_train_pred, y_train),
       metrics.mean_squared_error(x_test_pred, y_test)))
print('R2 train data: %.3f, R2 test data: %.3f' %
      (metrics.r2_score(y_train,x_train_pred, y_train),
       metrics.r2_score(y_test,x_test_pred, y_test)))
```

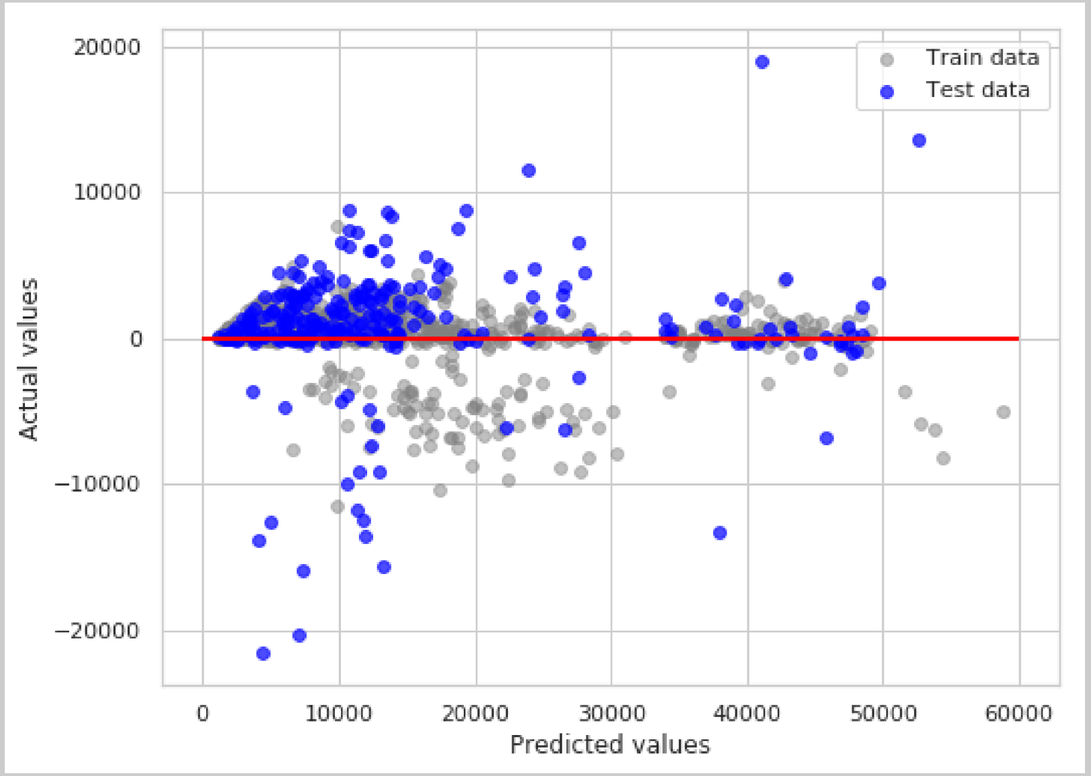
Result:

```
MSE train data: 3630549.354, MSE test data: 19737210.132
R2 train data: 0.971, R2 test data: 0.877
```

Random Forest Regression

```
from sklearn.ensemble import RandomForestRegressor as rfr
x = df.drop(['charges'], axis=1)
y = df.charges
Rfr = rfr(n_estimators = 100, criterion = 'mse',
                              random_state = 1,
                              n_jobs = -1)
Rfr.fit(x_train,y_train)
x_train_pred = Rfr.predict(x_train)
x_test_pred = Rfr.predict(x_test)

print('MSE train data: %.3f, MSE test data: %.3f' %
      (metrics.mean_squared_error(x_train_pred, y_train),
       metrics.mean_squared_error(x_test_pred, y_test)))
print('R2 train data: %.3f, R2 test data: %.3f' %
      (metrics.r2_score(y_train,x_train_pred, y_train),
       metrics.r2_score(y_test,x_test_pred, y_test)))
```



Result:

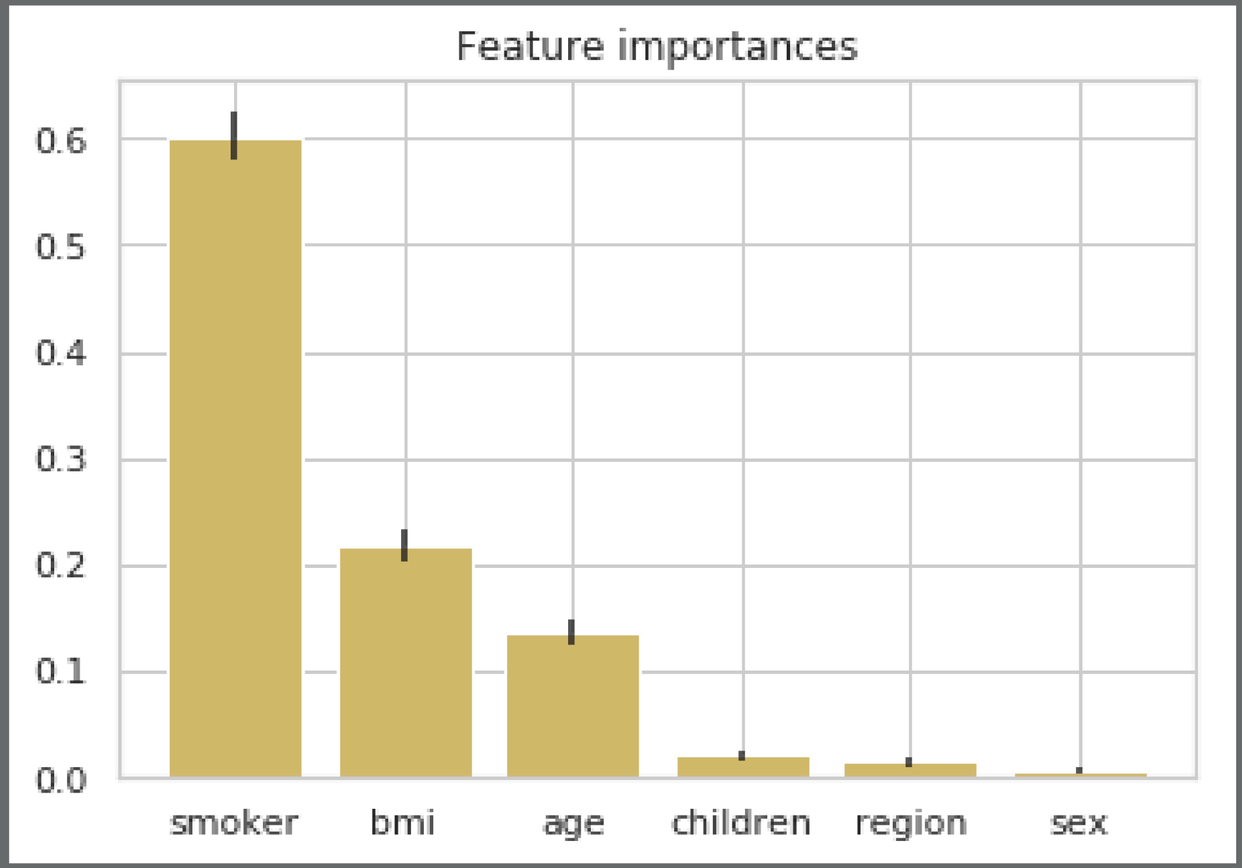
```
MSE train data: 3630549.354, MSE test data: 19737210.132
R2 train data: 0.971, R2 test data: 0.877
```

By Using RFR

1st Method

Feature importance ranking

1.smoker(0.601678)
2.bmi(0.218751)
3.age(0.136942)
4.children(0.021042)
5.region(0.015209)
6.sex(0.006377)



Polynomial Regression

```
from sklearn.preprocessing import PolynomialFeatures
x = df.drop(['charges', 'sex', 'region'], axis = 1)
y = df.charges
pol = PolynomialFeatures (degree = 2)
x_pol = pol.fit_transform(x)
x_train, x_test, y_train, y_test = holdout(x_pol, y, test_size=0.2, random_s
tate=0)
Pol_reg = LinearRegression()
Pol_reg.fit(x_train, y_train)
y_train_pred = Pol_reg.predict(x_train)
y_test_pred = Pol_reg.predict(x_test)
print(Pol_reg.intercept_)
print(Pol_reg.coef_)
print(Pol_reg.score(x_test, y_test))
```

```
-5325.881705252252
[ 0.00000000e+00 -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]
0.8812595703345225
```

Polynomial Regression

```
##Evaluating the performance of the algorithm
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_test_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_test_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_test_pred)))
```

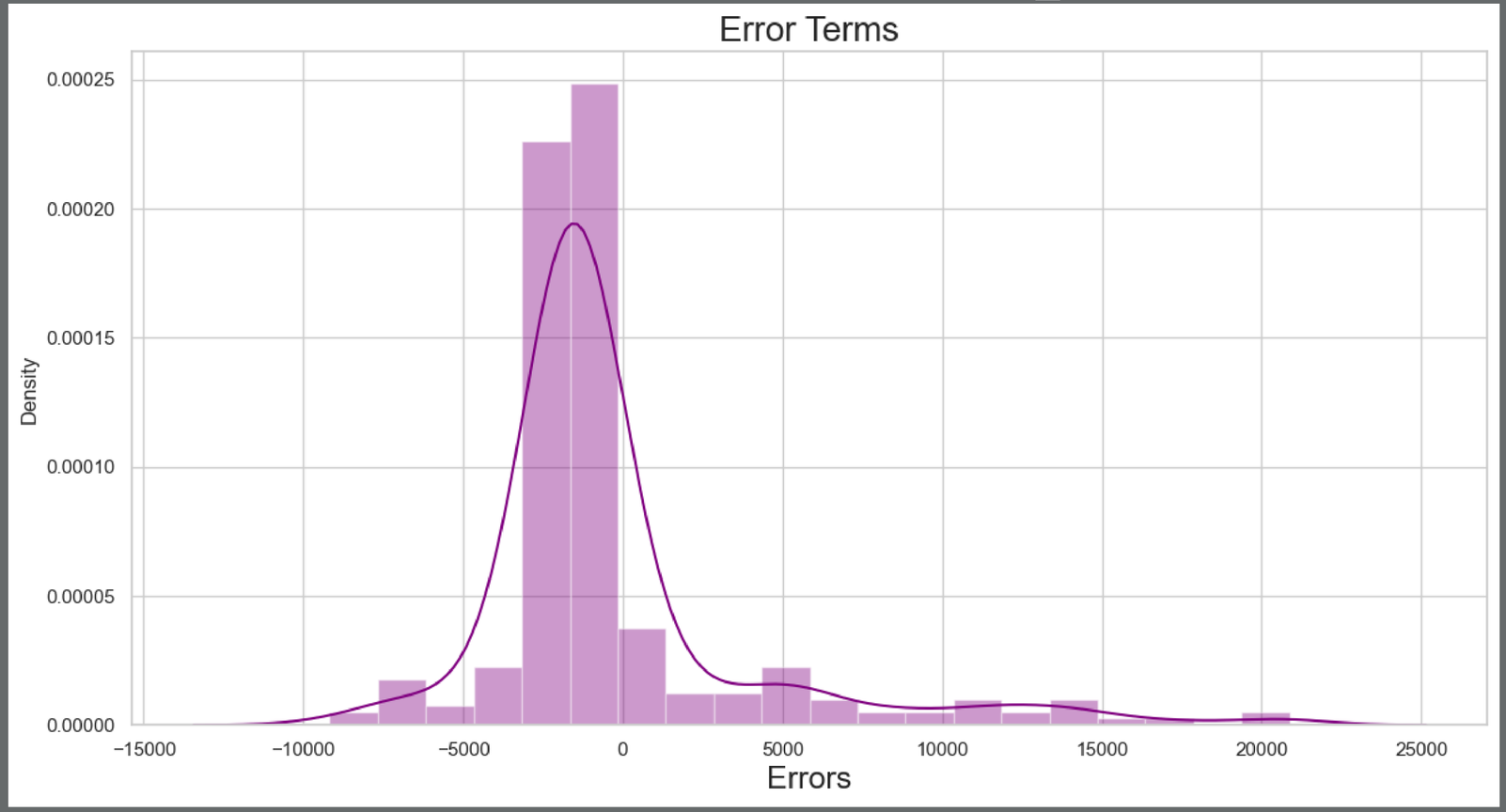
```
Mean Absolute Error: 2824.4950454776595
Mean Squared Error: 18895160.09878044
Root Mean Squared Error: 4346.856346692451
```

- **Polynomial Regression turned out to be the best model**

	Actual	Predicted
578	9724.530000	12101.156323
610	8547.691300	10440.782266
569	45702.022350	48541.022951
1034	12950.071200	14140.067522
198	9644.252500	8636.235727
981	4500.339250	5072.787029
31	2198.189850	3090.494817
1256	11436.738150	13171.361938
1219	7537.163900	9187.612192
1320	5425.023350	7496.320857
613	6753.038000	6653.904925
1107	10493.945800	11893.766490
1263	7337.748000	9291.317273
406	4185.097900	5326.271479
795	18310.742000	25726.734553
970	10702.642400	12643.360147
824	12523.604800	13099.011032
141	3490.549100	5336.149644
1173	6457.843400	8680.680007
1042	33475.817150	27696.731856
966	23967.383050	27202.661385
467	12643.377800	14490.242822
1098	23045.566160	10998.132160
757	23065.420700	29566.723915
1097	1674.632300	3611.941947
319	4667.607650	6215.673009
1286	3732.625100	3077.366908
459	7682.670000	9772.813467
5	3756.621600	5085.856146
517	8413.463050	10696.238443
535	6067.126750	7852.517542
853	11729.679500	12500.224437
1014	5383.536000	6780.484833
1186	37465.343750	35020.456915
215	7371.772000	9353.329096
1046	7325.048200	7794.117226
986	8410.046850	10584.144595
489	10461.979400	12489.567570
968	3279.868550	4999.229874
1160	7727.253200	9297.870472
792	2731.912200	3299.083436
1224	6858.479600	8156.022661
465	19521.968200	26671.512279
251	47305.305000	42890.820207
1017	3987.926000	5950.457712
1239	3238.435700	3930.573984
427	7323.734819	3192.996621
295	1704.568100	2778.118484
820	7445.918000	9919.906707
1335	1629.833500	2709.177041
884	4877.981050	6174.548189
326	3561.888900	5019.195968
1109	8605.361500	9777.556588
783	24520.264000	30818.987409
668	45710.207850	40988.027774
1084	15019.760050	16712.196281
726	6664.685950	8654.565461
1132	20709.020340	12372.050609
725	40932.429500	41465.617268
963	9500.573050	10941.780705

Checking the Error Term of Multiple Linear Regression

Versionı of Multiple



2nd Fitting with Multiple Linear Regression

- By RFE (Recursive Feature Elimination)
- Model Fitting
- Result

2nd Method

RFE (Recursive Feature Elimination)

Recursive Feature Elimination-RFE

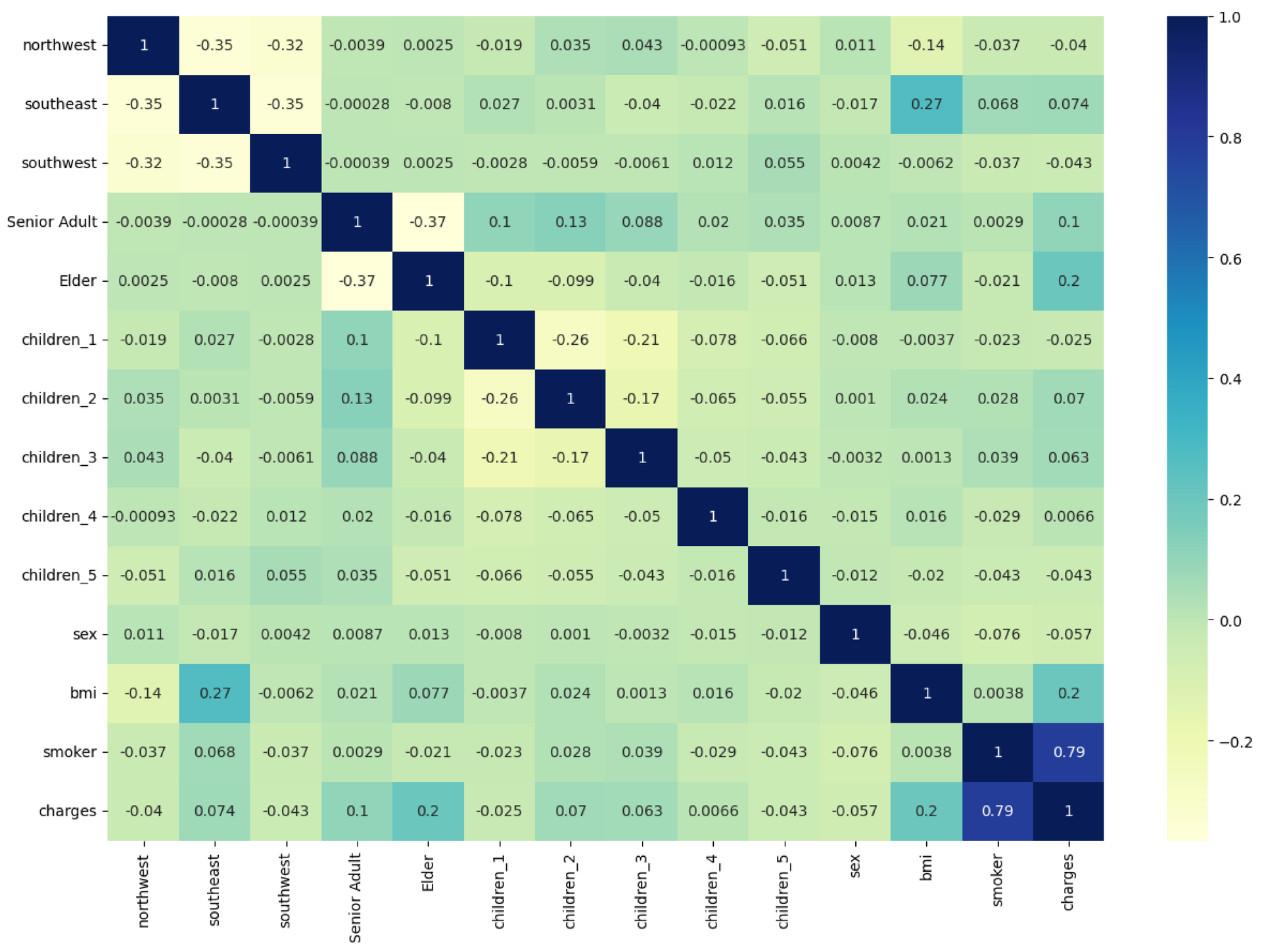
```
lm = LinearRegression()
lm.fit(X_train, y_train)

# rfe = RFE(lm, 8)             # running RFE
rfe = RFE(estimator=lm, n_features_to_select=8)
rfe = rfe.fit(X_train, y_train)
```

```
#List of variables selected
list(zip(X_train.columns,rfe.support_,rfe.ranking_))
```

```
[('northwest', False, 4),
 ('southeast', False, 3),
 ('southwest', False, 2),
 ('Senior Adult', True, 1),
 ('Elder', True, 1),
 ('children_1', False, 5),
 ('children_2', True, 1),
 ('children_3', True, 1),
 ('children_4', True, 1),
 ('children_5', True, 1),
 ('sex', False, 6),
 ('bmi', True, 1),
 ('smoker', True, 1)]
```

By RFE, we will choose these 8 variables to fit the model such as: ***'Senior Adult', 'Elder', 'children_2', 'children_3', 'children_4', 'children_5', 'bmi', 'smoker'***


northwest 1 -0.35 -0.32 -0.0039 0.0025 -0.019 0.035 0.043 -0.00093 -0.051 0.011 -0.14 -0.037 -0.04
southeast -0.35 1 -0.35 -0.00028 -0.008 0.027 0.0031 -0.04 -0.022 0.016 -0.017 0.27 0.068 0.074
southwest -0.32 -0.35 1 -0.00039 0.0025 -0.0028 -0.0059 -0.0061 0.012 0.055 0.0042 -0.0062 -0.037 -0.043
Senior Adult -0.0039 -0.00028 -0.00039 1 -0.37 0.1 0.13 0.088 0.02 0.035 0.0087 0.021 0.0029 0.1
Elder 0.0025 -0.008 0.0025 -0.37 1 -0.1 -0.099 -0.04 -0.016 -0.051 0.013 0.077 -0.021 0.2
children_1 -0.019 0.027 -0.0028 0.1 -0.1 1 -0.26 -0.21 -0.078 -0.066 -0.008 -0.0037 -0.023 -0.025
children_2 0.035 0.0031 -0.0059 0.13 -0.099 -0.26 1 -0.17 -0.065 -0.055 0.001 0.024 0.028 0.07
children_3 0.043 -0.04 -0.0061 0.088 -0.04 -0.21 -0.17 1 -0.05 -0.043 -0.0032 0.0013 0.039 0.063
children_4 -0.00093 -0.022 0.012 0.02 -0.016 -0.078 -0.065 -0.05 1 -0.016 -0.015 0.016 -0.029 0.0066
children_5 -0.051 0.016 0.055 0.035 -0.051 -0.066 -0.055 -0.043 -0.016 1 -0.012 -0.02 -0.043 -0.043
sex 0.011 -0.017 0.0042 0.0087 0.013 -0.008 0.001 -0.0032 -0.015 -0.012 1 -0.046 -0.076 -0.057
bmi -0.14 0.27 -0.0062 0.021 0.077 -0.0037 0.024 0.0013 0.016 -0.02 -0.046 1 0.0038 0.2
smoker -0.037 0.068 -0.037 0.0029 -0.021 -0.023 0.028 0.039 -0.029 -0.043 -0.076 0.0038 1 0.79
charges -0.04 0.074 -0.043 0.1 0.2 -0.025 0.07 0.063 0.0066 -0.043 -0.057 0.2 0.79 1
northwest southeast southwest Senior Adult Elder children_1 children_2 children_3 children_4 children_5 sex bmi smoker charges
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

MULTIPLE LINEAR REGRESSION WITH 2ND METHOD

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.723
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	485.6
Date:	Sun, 14 May 2023	Prob (F-statistic):	2.45e-256
Time:	23:26:23	Log-Likelihood:	809.68
No. Observations:	936	AIC:	-1607.
Df Residuals:	930	BIC:	-1578.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0126	0.009	-1.342	0.180	-0.031	0.006
Senior Adult	0.0779	0.007	10.592	0.000	0.063	0.092
Elder	0.1495	0.010	15.136	0.000	0.130	0.169
children_2	0.0245	0.009	2.810	0.005	0.007	0.042
bmi	0.1760	0.020	8.750	0.000	0.136	0.215
smoker	0.3826	0.008	45.290	0.000	0.366	0.399

Omnibus:	215.860	Durbin-Watson:	2.049
Prob(Omnibus):	0.000	Jarque-Bera (JB):	498.375
Skew:	1.241	Prob(JB):	6.02e-109
Kurtosis:	5.572	Cond. No.	7.84

	Features	VIF
3	bmi	2.34
0	Senior Adult	1.81
1	Elder	1.31
2	children_2	1.24
4	smoker	1.21

Now our model is good with p-values and VIF under the acceptable range

```
#Evaluate R-square for test
from sklearn.metrics import r2_score
r2_score(y_test,y_pred)
```

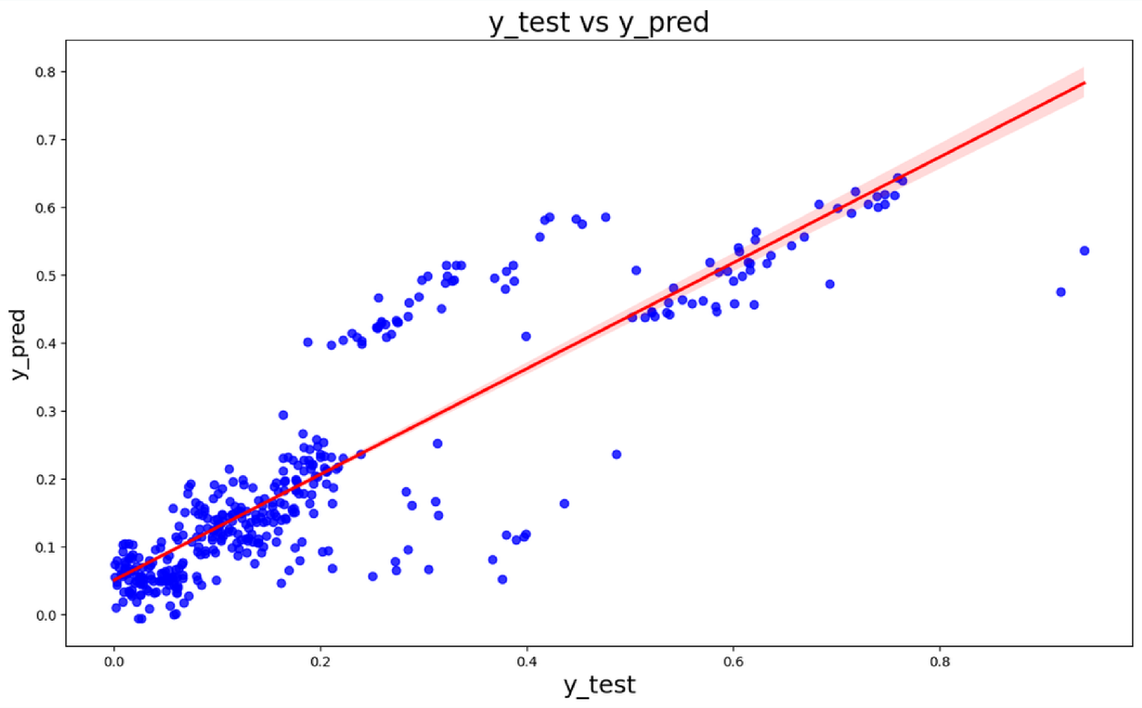
```
0.7628855670251862
```

```
#n =sample size , p = number of independent variables
n = X_test.shape[0]
p = X_test.shape[1]

Adj_r2=1-(1-0.75783003115855)*(n-1)/(n-p-1)
print(Adj_r2)
```

```
0.7497160889035529
```

RESULT MULTIPLE LINEAR REGRESSION WITH 2ND METHOD



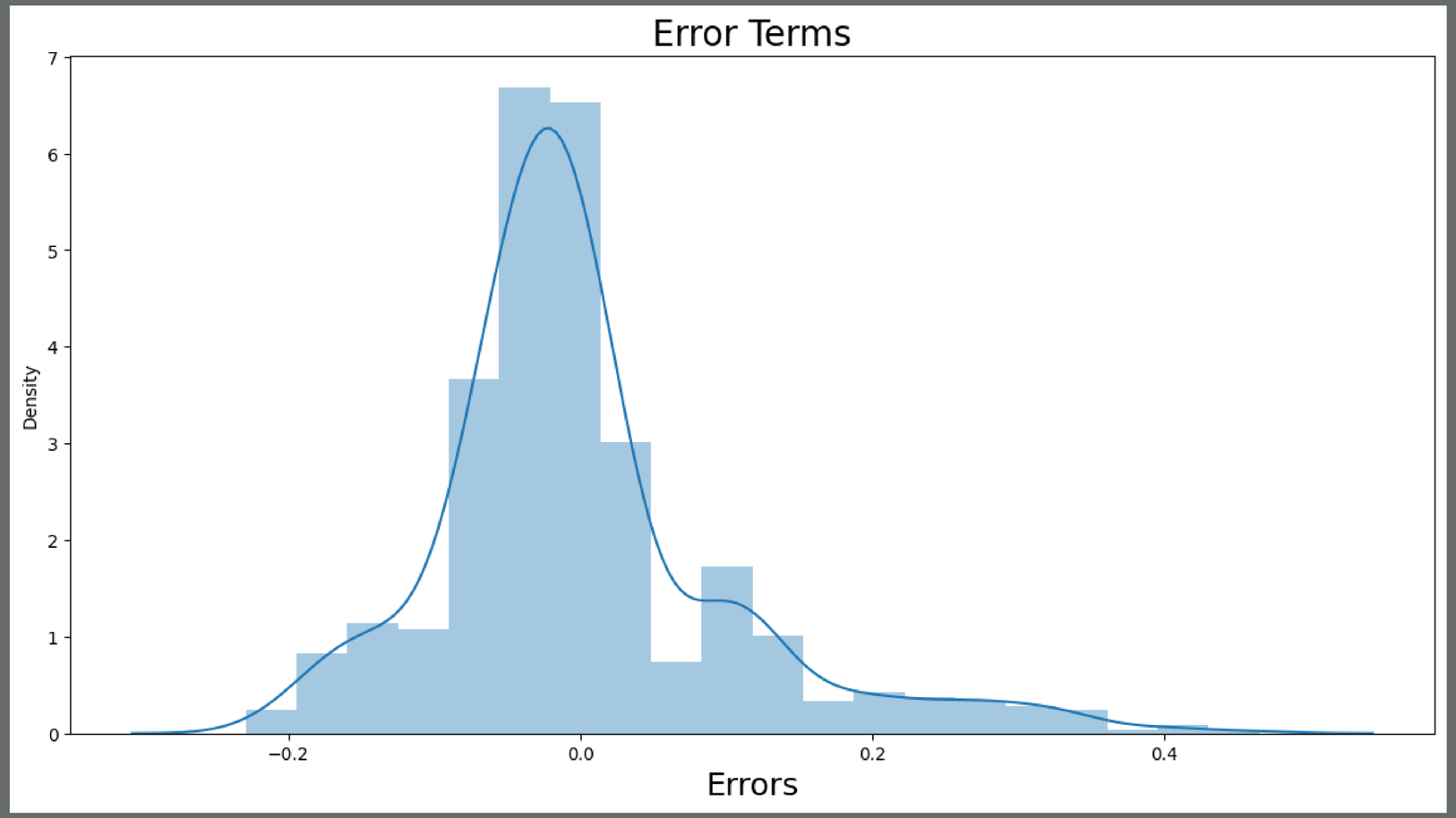
- From the regression analysis, we find that region and ***sex*** do not bring significant difference on charges.
- ***Age, BMI, number of children and smoking*** are the ones that drive the charges
- ***Smoking*** seems to have the most influence on the medical charges

We can see that the equation of our best fitted line is:

Charges = 0.3826 smoker + 0.077 Senior Adult + 0.149 Elder + 0.176 BMI + 0.024 Children_2

Checking the Error Term of Multiple Linear Regression

Version2 of Multiple



3rd Fitting with Multiple Linear Regression

- By Backward & Forward Selection
- Model Fitting
- Result

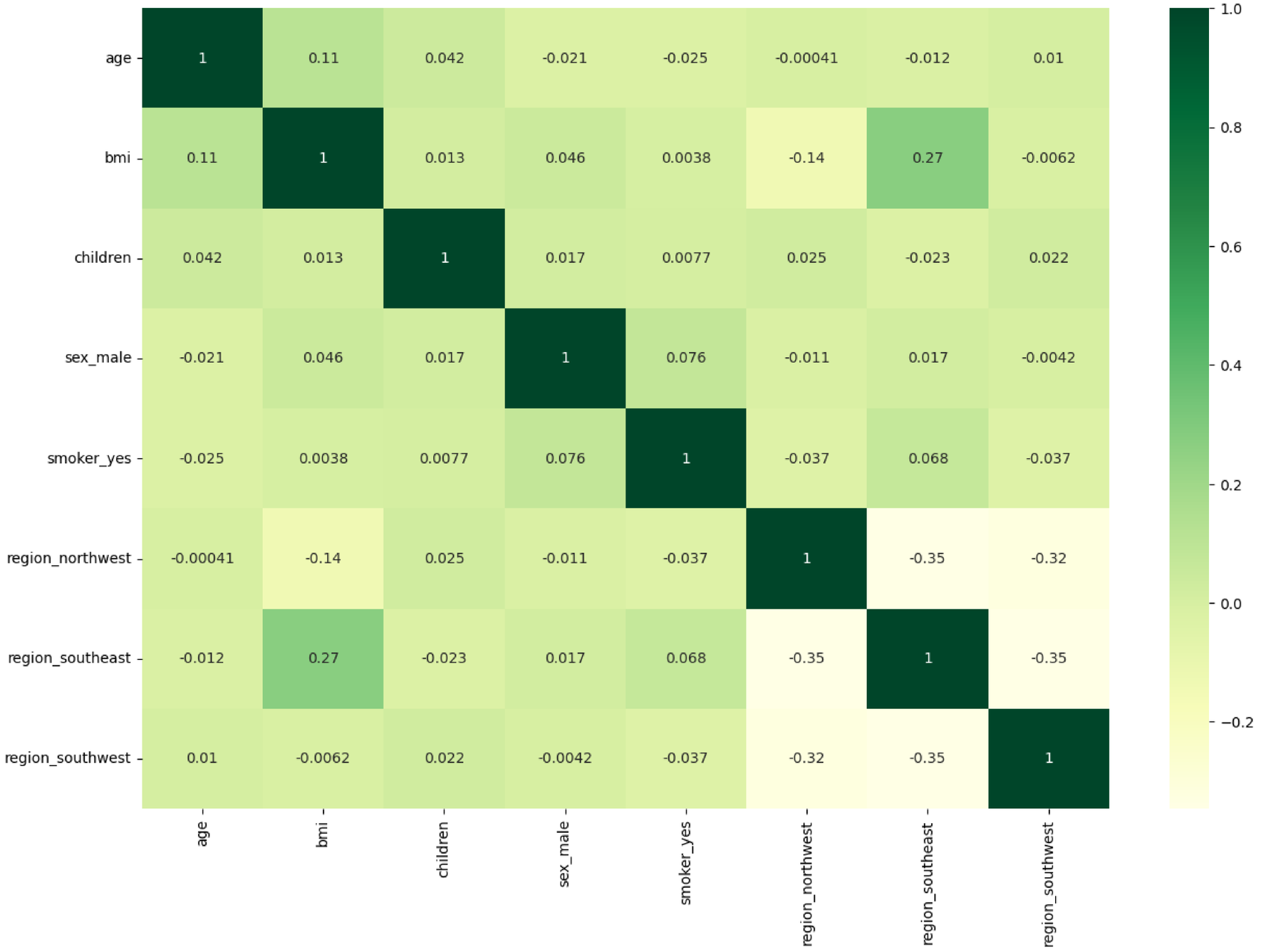
3rd Method

Backward Selection
&
Forward Selection

```
Out[15]: ['smoker_yes',
          'age',
          'bmi',
          'children',
          'region_southeast',
          'region_southwest']
```

```
In [16]: 1 set(predictor_back) == set(predictor_forward)

Out[16]: True
```


age 1 0.11 0.042 -0.021 -0.025 -0.00041 -0.012 0.01
bmi 0.11 1 0.013 0.046 0.0038 -0.14 0.27 -0.0062
children 0.042 0.013 1 0.017 0.0077 0.025 -0.023 0.022
sex_male -0.021 0.046 0.017 1 0.076 -0.011 0.017 -0.0042
smoker_yes -0.025 0.0038 0.0077 0.076 1 -0.037 0.068 -0.037
region_northwest -0.00041 -0.14 0.025 -0.011 -0.037 1 -0.35 -0.32
region_southeast -0.012 0.27 -0.023 0.017 0.068 -0.35 1 -0.35
region_southwest 0.01 -0.0062 0.022 -0.0042 -0.037 -0.32 -0.35 1
age
bmi
children
sex_male
smoker_yes
region_northwest
region_southeast
region_southwest
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

MULTIPLE LINEAR REGRESSION WITH 3RD METHOD

OLS Regression Results

Dep. Variable:	y	R-squared:	0.740
Model:	OLS	Adj. R-squared:	0.738
Method:	Least Squares	F-statistic:	378.2
Date:	Sat, 17 Jun 2023	Prob (F-statistic):	2.00e-304
Time:	19:17:58	Log-Likelihood:	-10846.
No. Observations:	1070	AIC:	2.171e+04
Df Residuals:	1061	BIC:	2.175e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.224e+04	1112.140	-11.010	0.000	-1.44e+04	-1.01e+04
age	267.4967	13.392	19.974	0.000	241.218	293.775
bmi	336.8498	32.103	10.493	0.000	273.857	399.843
children	406.3633	155.040	2.621	0.009	102.143	710.584
sex_male	-219.7955	376.724	-0.583	0.560	-959.004	519.413
smoker_yes	2.369e+04	475.879	49.777	0.000	2.28e+04	2.46e+04
region_northwest	-80.1959	532.812	-0.151	0.880	-1125.681	965.289
region_southeast	-985.3338	546.166	-1.804	0.072	-2057.022	86.354
region_southwest	-842.4895	538.704	-1.564	0.118	-1899.536	214.557

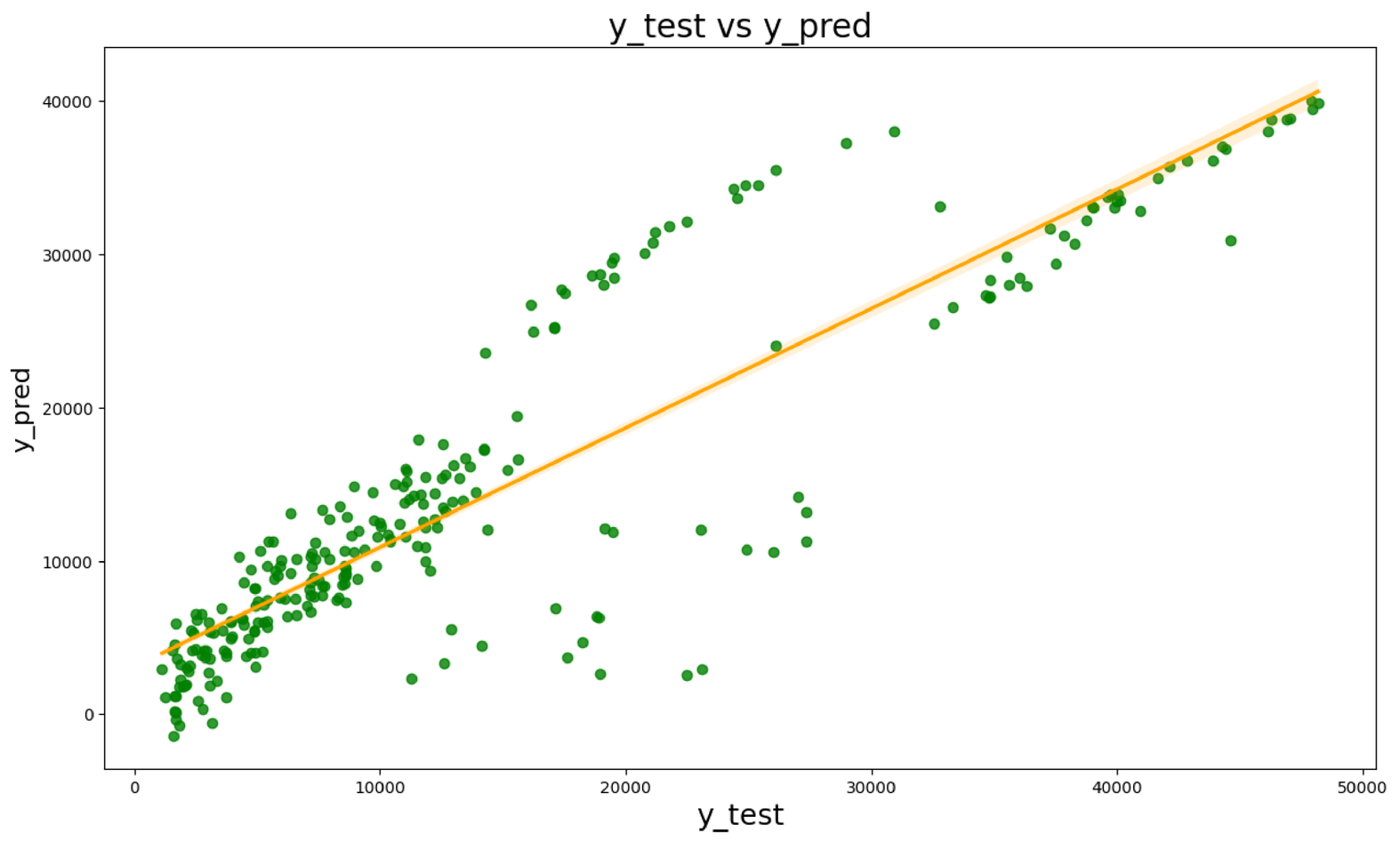
Omnibus:	273.543	Durbin-Watson:	2.012
Prob(Omnibus):	0.000	Jarque-Bera (JB):	708.949
Skew:	1.329	Prob(JB):	1.13e-154
Kurtosis:	5.973	Cond. No.	310.

	Variable	VIF
0	age	1.039
1	bmi	1.183
2	children	1.003
3	smoker_yes	1.090
4	region_northwest	1.499
5	region_southeast	1.624
6	region_southwest	1.541

```
1 #Evaluate R-square for test
2 from sklearn.metrics import r2_score
3 r2_score(y_test,y_pred)
```

```
0.7864665068297299
```

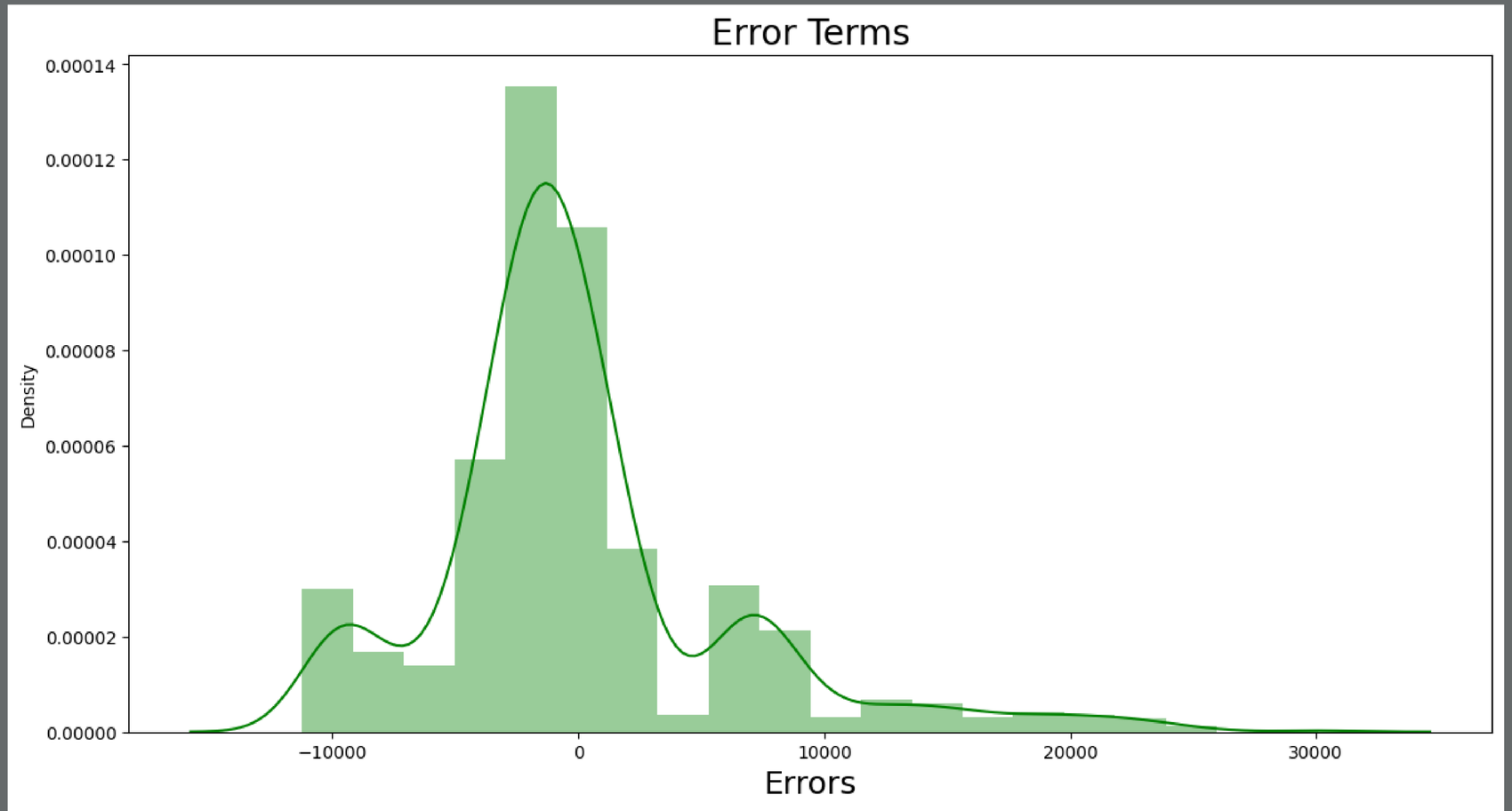
RESULT MULTIPLE LINEAR REGRESSION WITH 3RD METHOD



From the Multiple Linear Regression model, we know that being a smoker will cause a significant increase in beneficiary's charges than those who aren't.

Checking the Error Term of Multiple Linear Regression

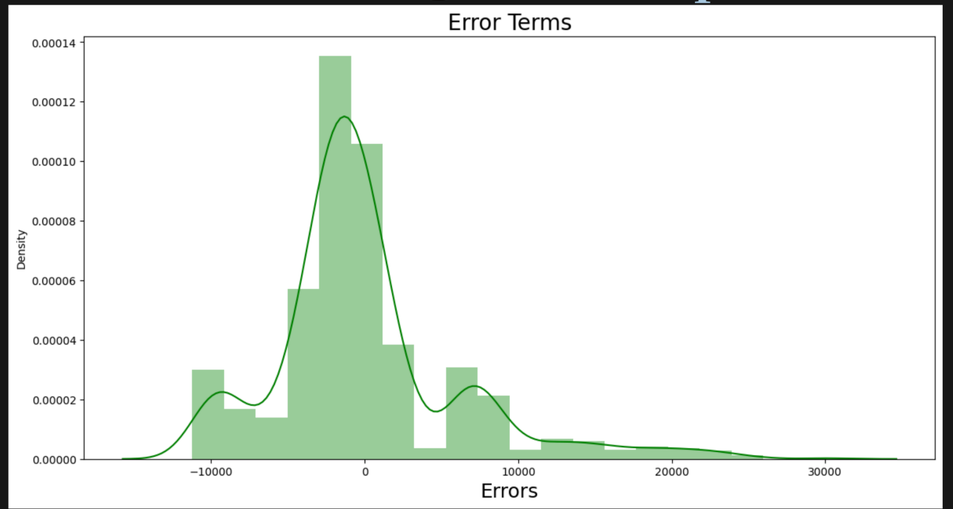
Version3 of Multiple



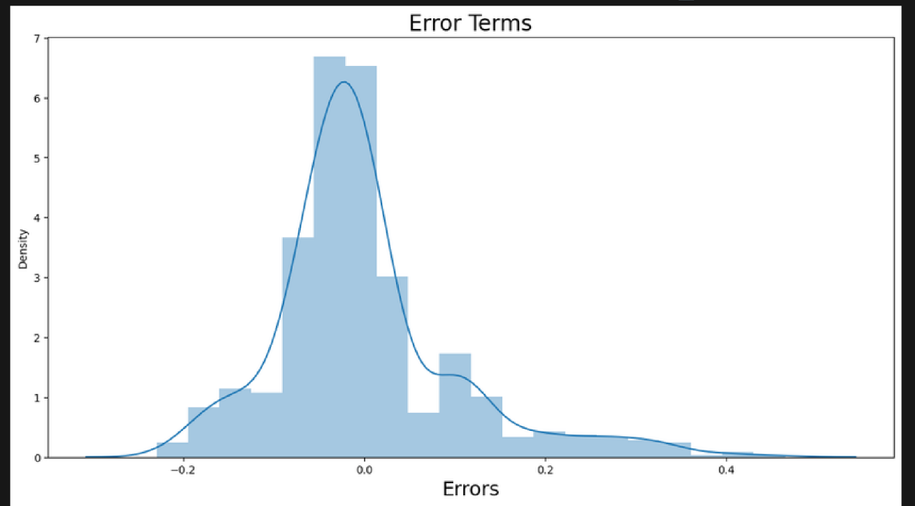
Compare the Error Term of Multiple Linear Regression

Version1 of Multiple

Version3 of Multiple



Version2 of Multiple



Recommendation

- Based on EDA and statistical evidence it can be seen that customer who smoke or have higher BMI have more higher claims. We can encourage customers to quit smoking by providing them incentive points for talking to life coach, get help for improving lifestyle habits, Quit Tobacco- 28 day program. Give gift cards when customer accumulates specific number of points.

- We can have Active wellness programs which can help up reduce claims related to BMI.

- High BMI is primarily because of unhealthy life choices. We can provide customers with Diet plans and wellness health coaches which can help them to make right choices.

- Provide discount coupons for Gym or fitness devices encouraging customers to exercis

Thank you for joining us today!