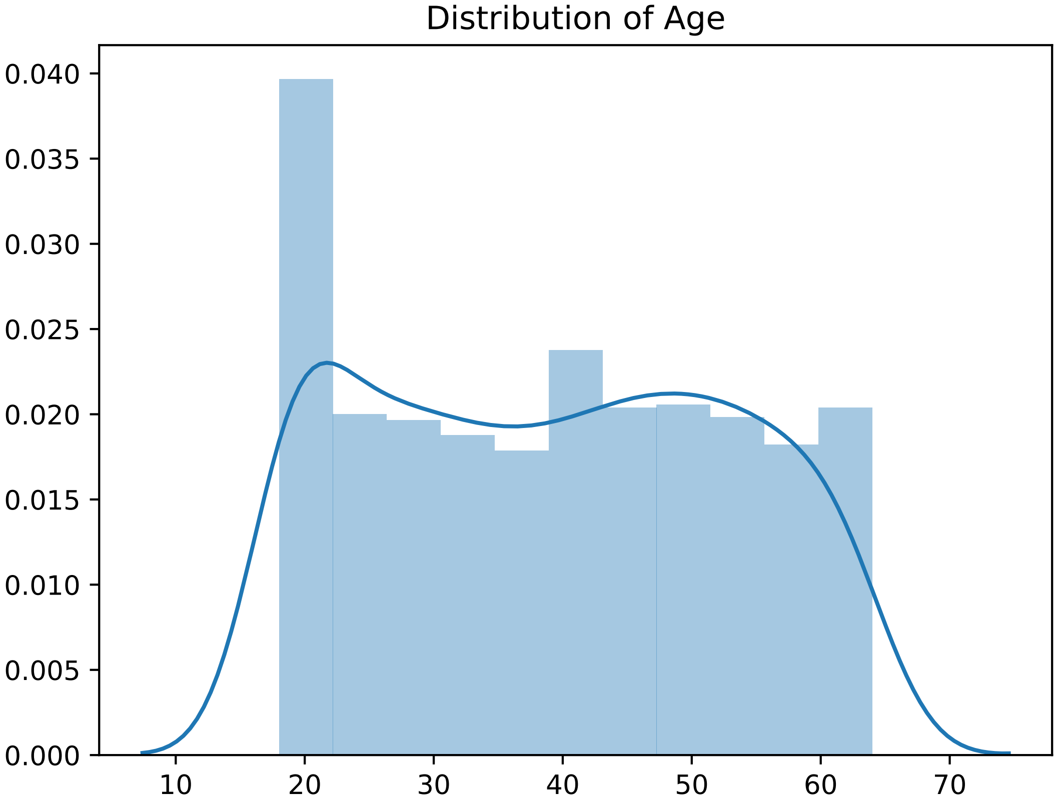
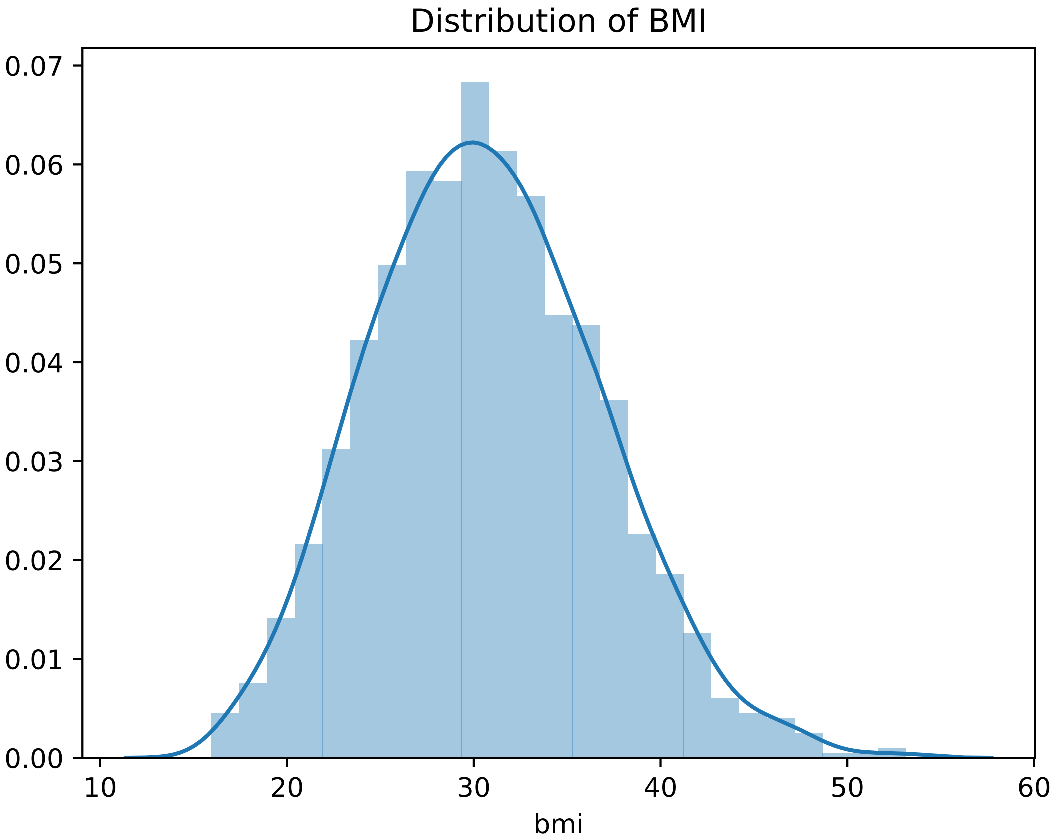
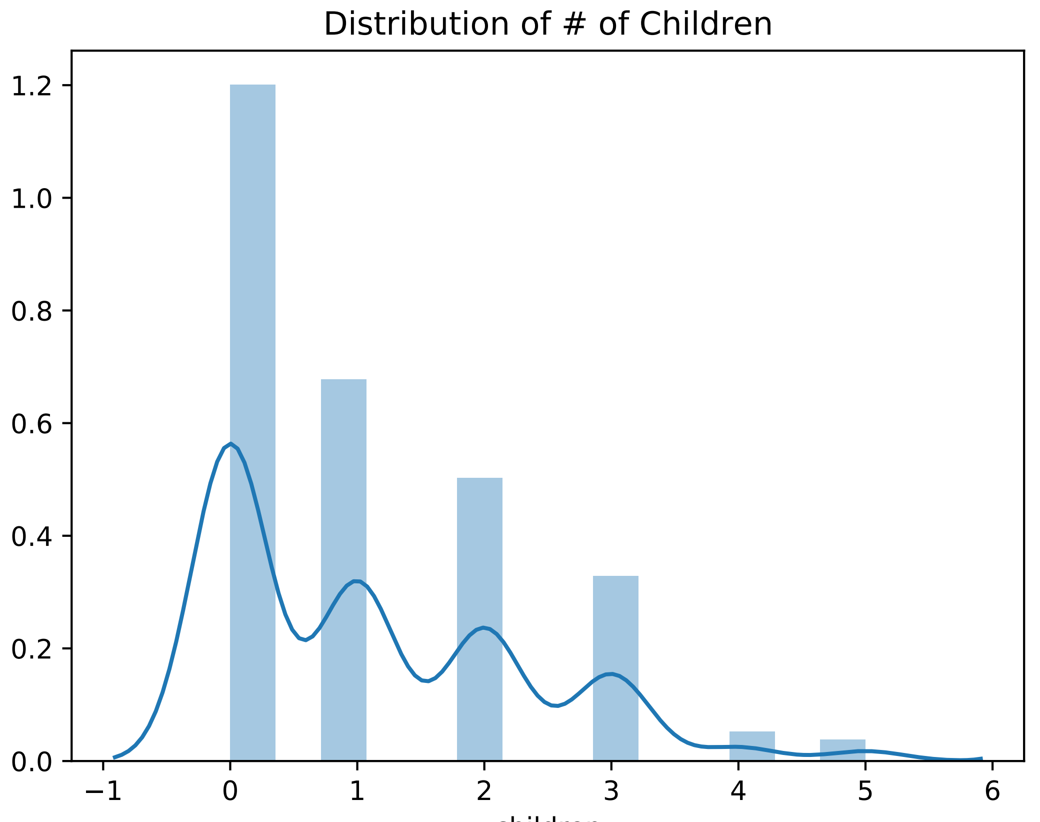
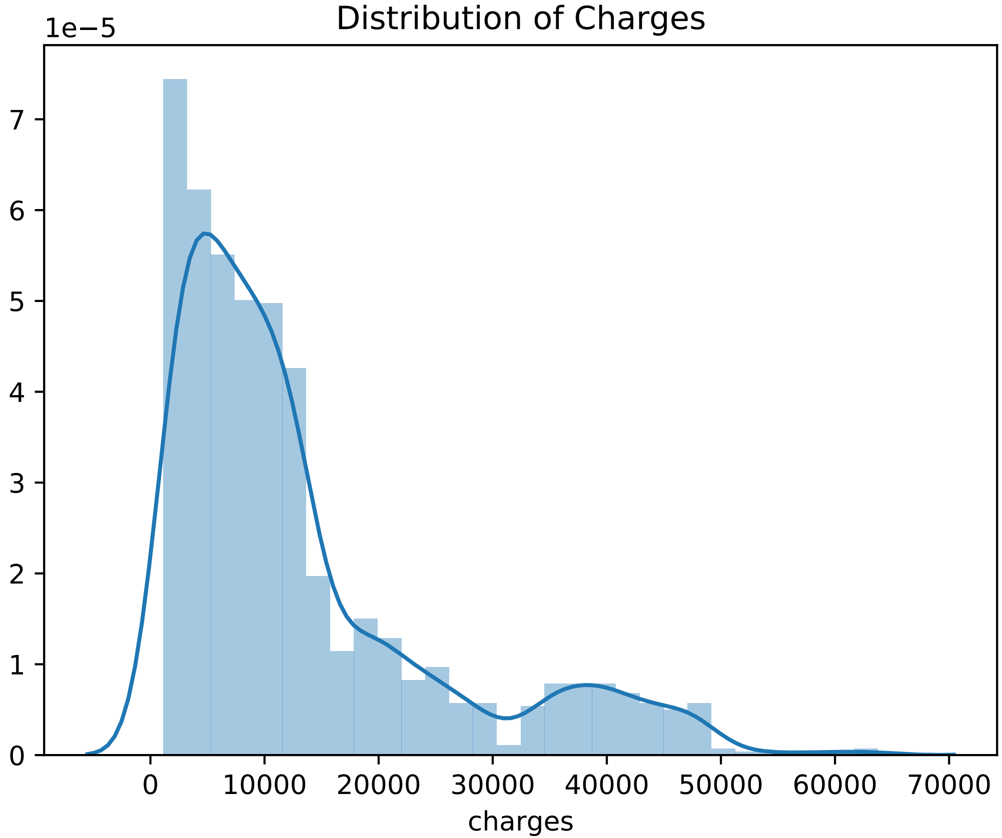
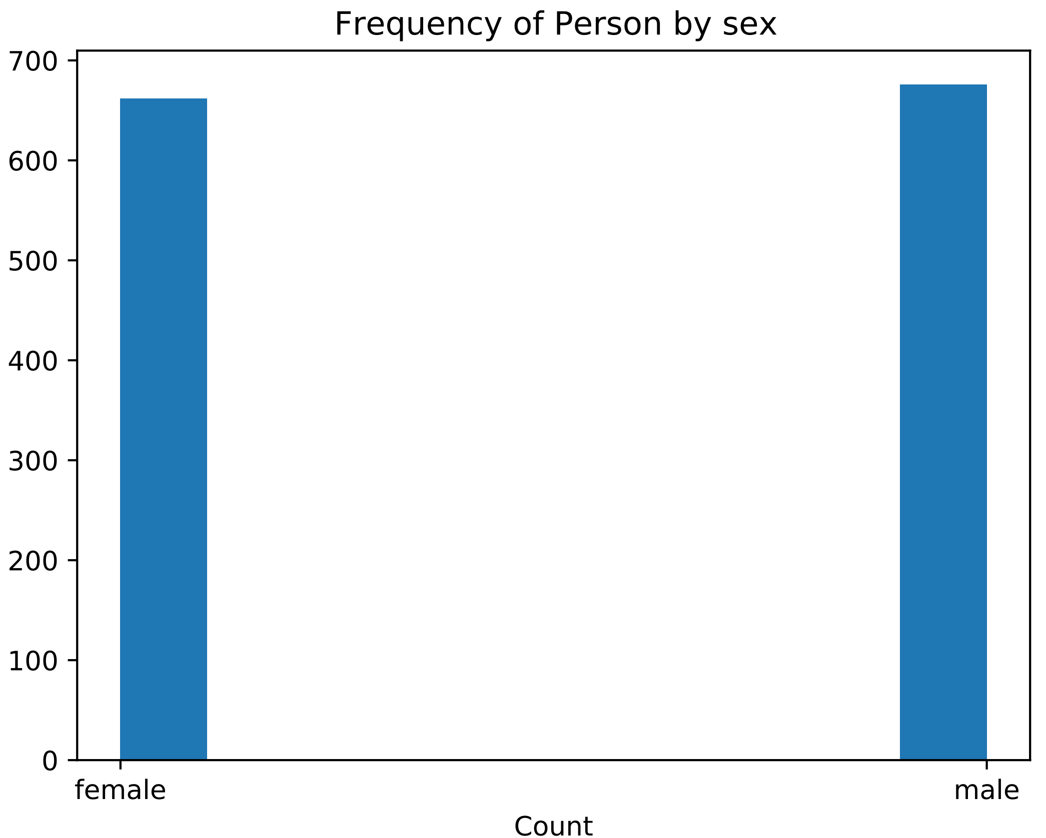

Distribution of Age
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
age

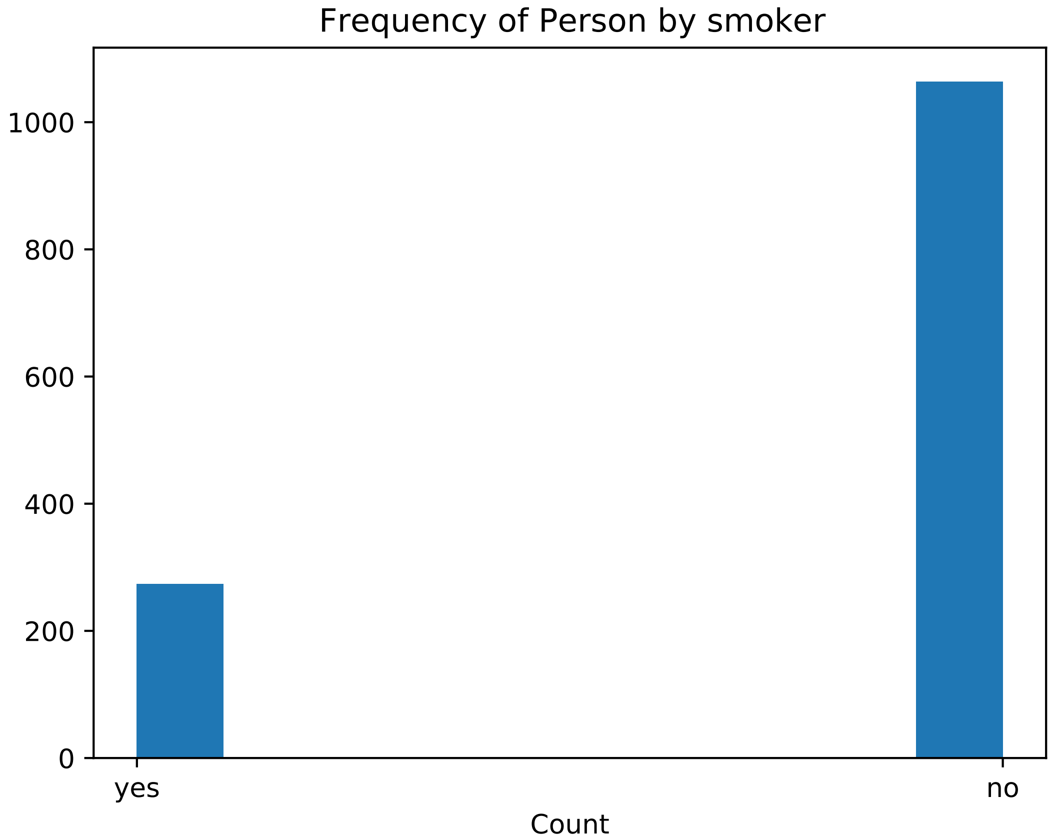

Distribution of BMI
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
10
20
30
40
50
60
bmi

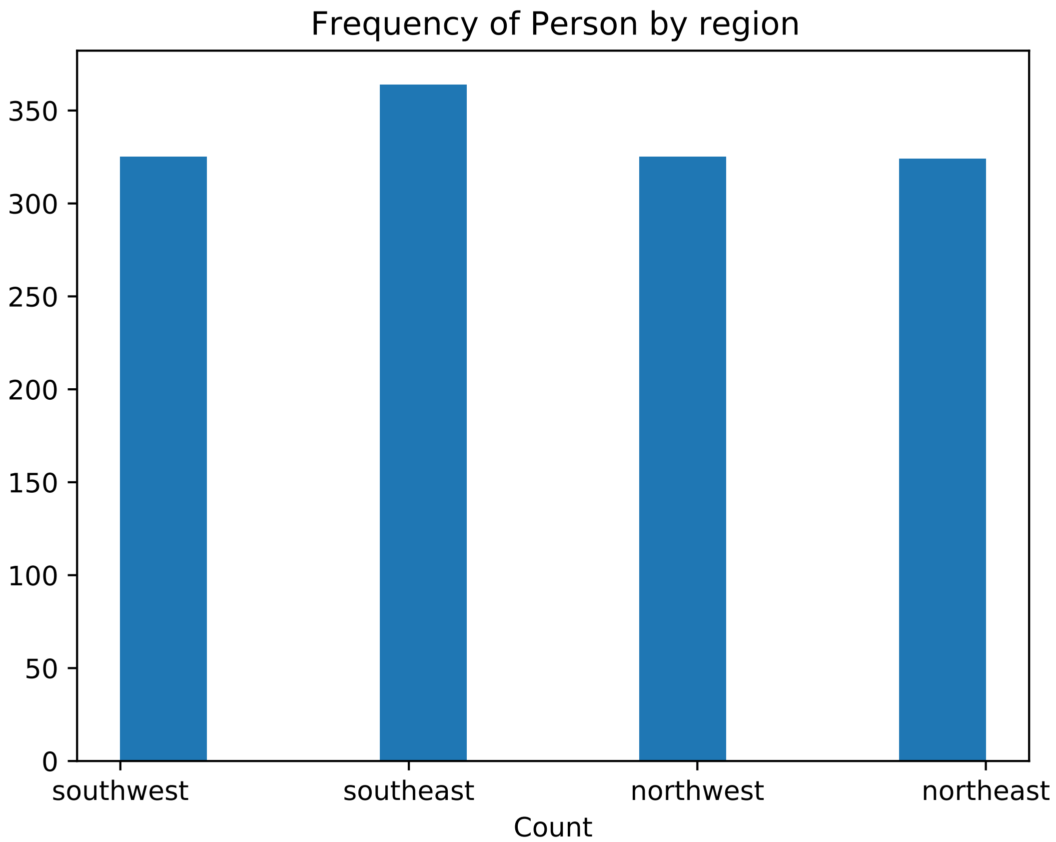


Distribution of # of Children
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
6
children

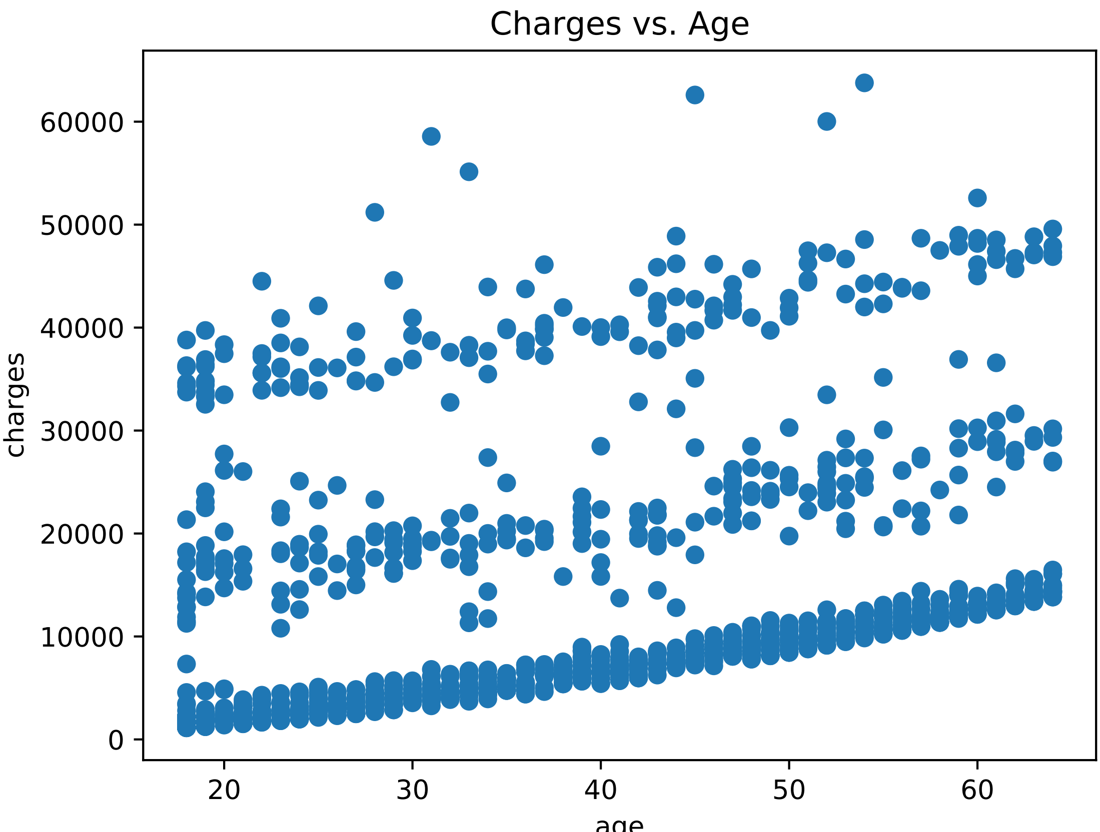

Distribution of Charges
1e−5
7
6
5
4
3
2
1
0
0
10000
20000
30000
40000
50000
60000
70000
charges

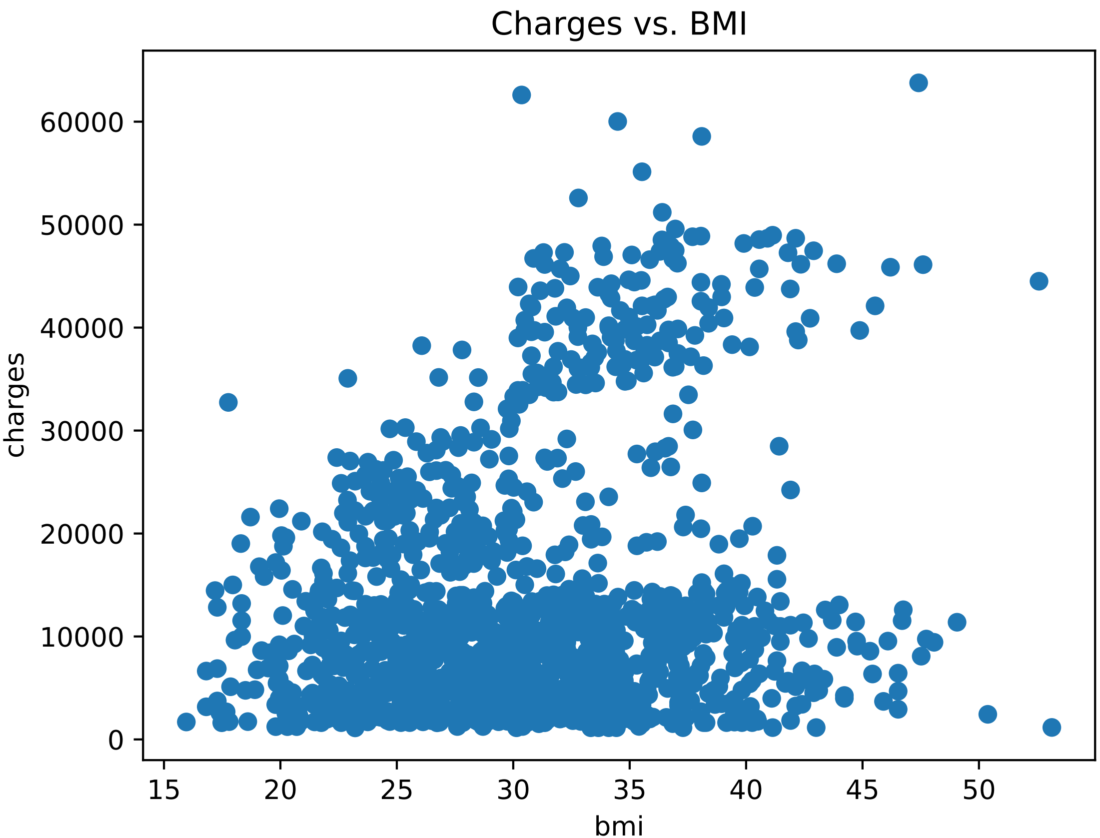

Frequency of Person by sex
700
600
500
400
300
200
100
0
female
male
Count

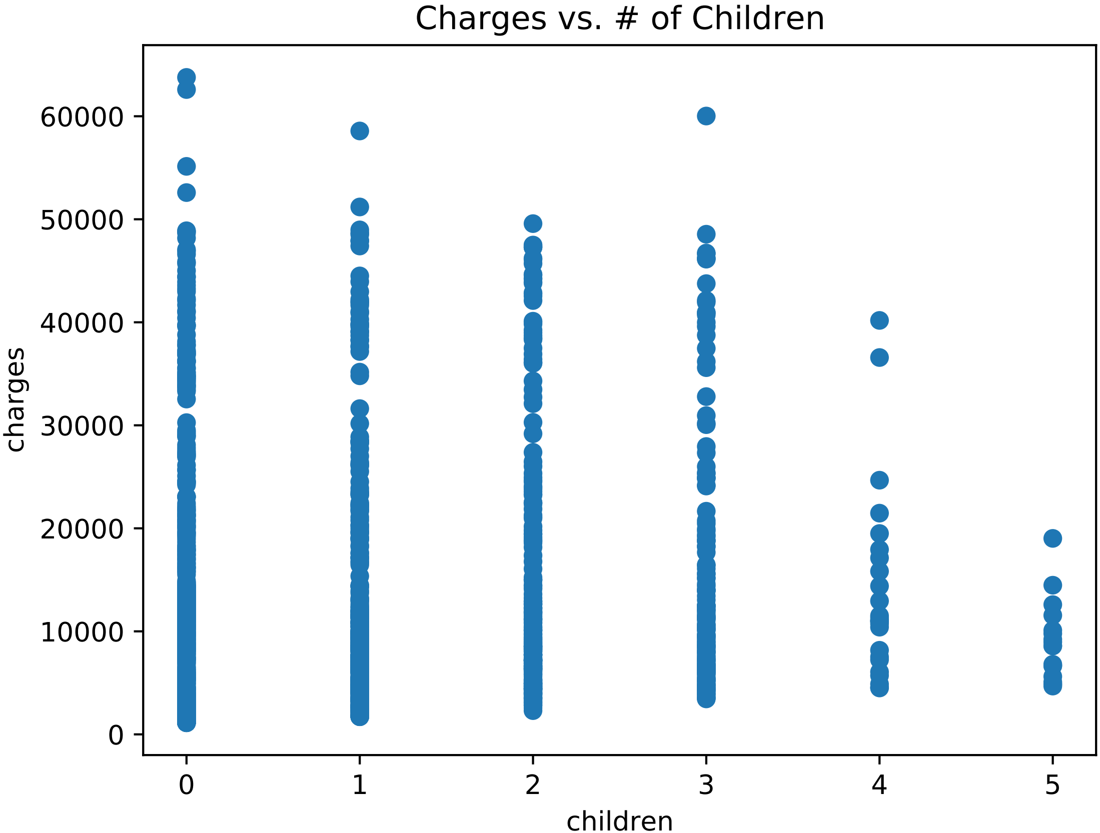

Frequency of Person by smoker
1000
800
600
400
200
0
yes
no
Count

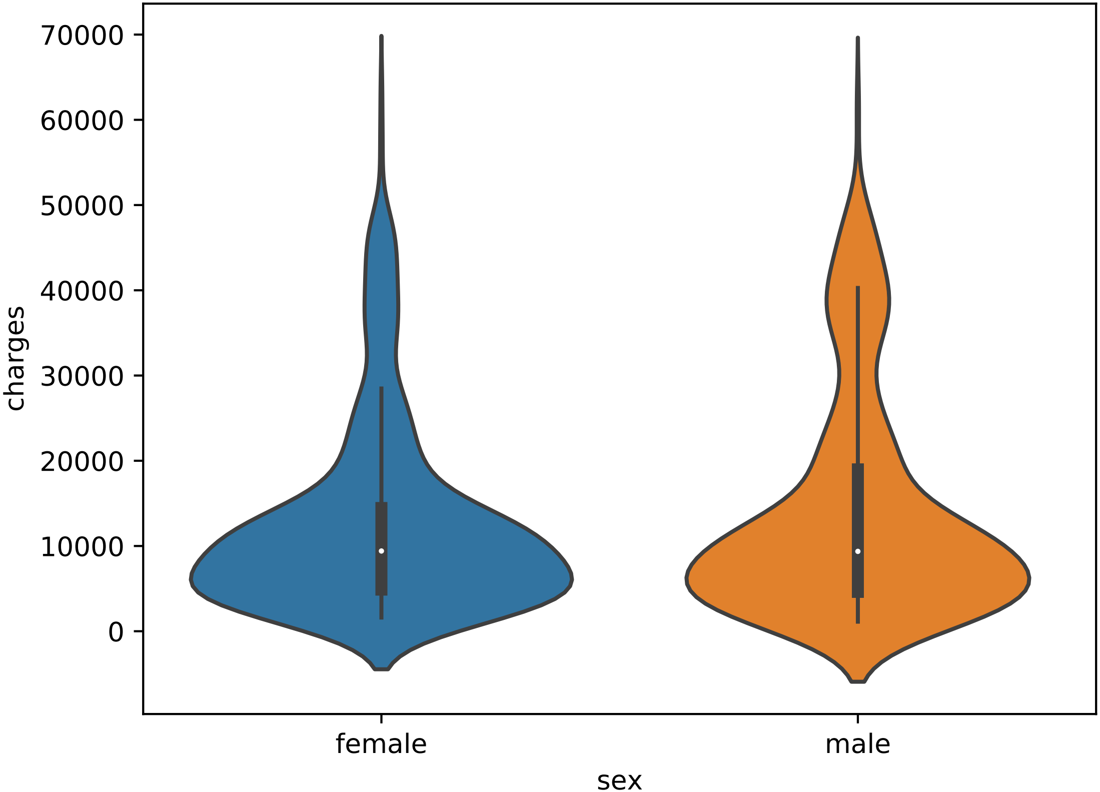

Frequency of Person by region
350
300
250
200
150
100
50
0
southwest
southeast
northwest
northeast
Count

Now that we understand the data in the dataset, lets look at charges vs. each data point to see which have an effect

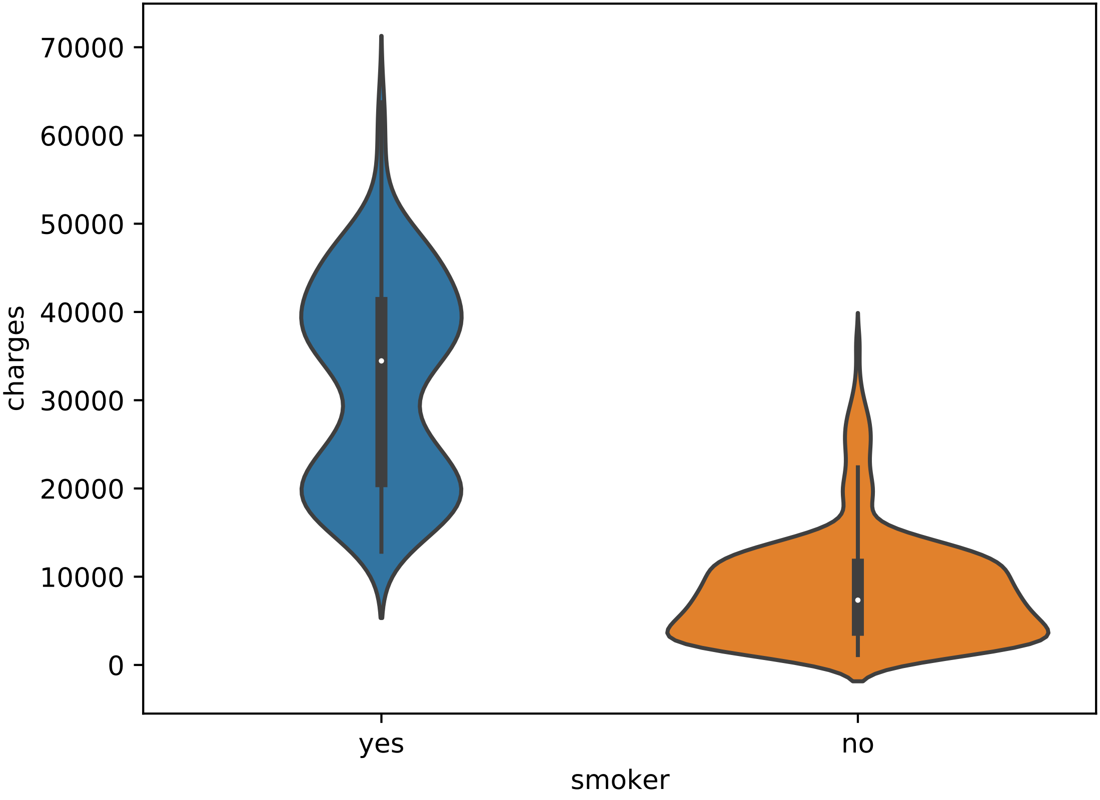

Charges vs. Age
60000
50000
40000
30000
20000
10000
0
charges
20
30
40
50
60
age


Charges vs. BMI
charges
bmi
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50

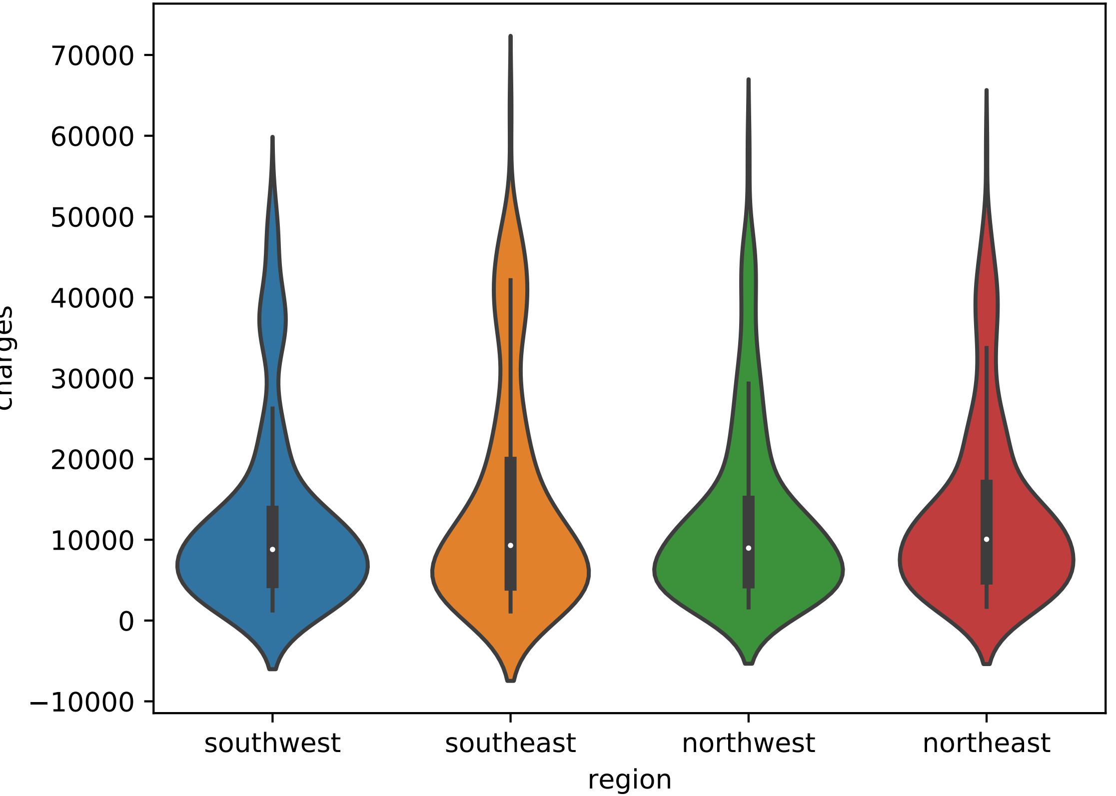

Charges vs. # of Children
charges
children
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5



70000
60000
50000
40000
30000
20000
10000
0
charges
female
male
sex

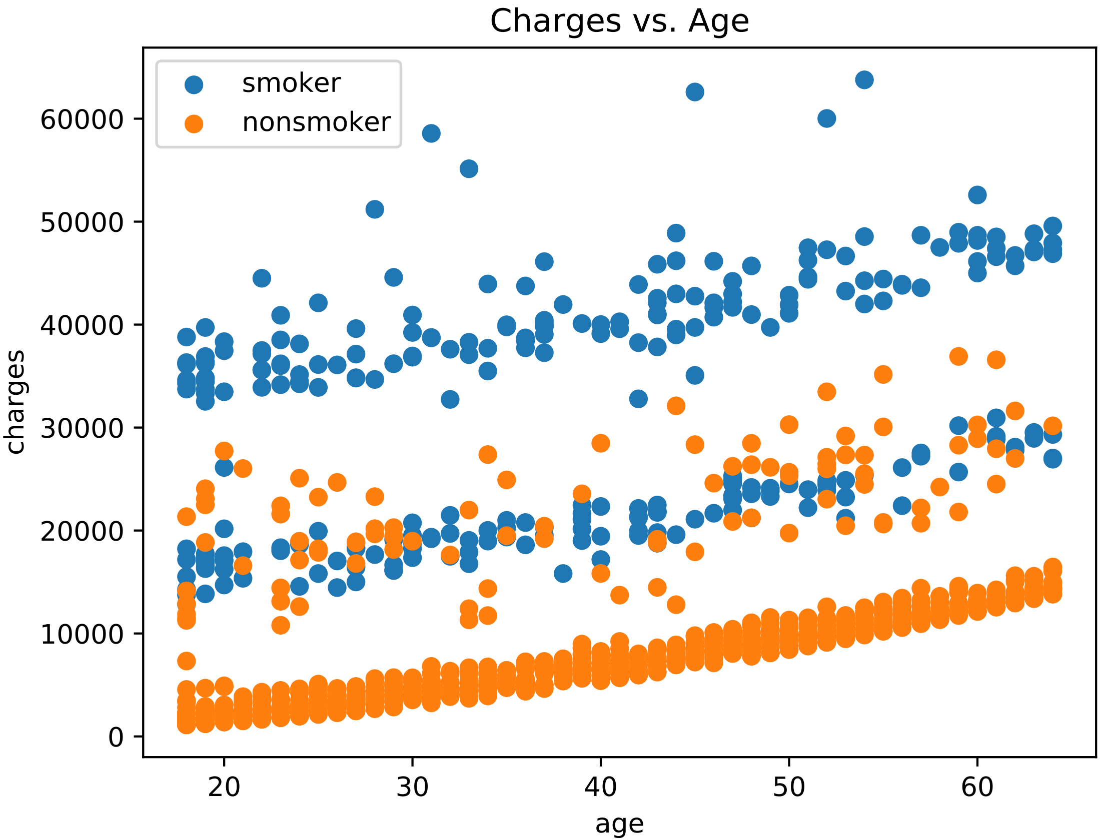


70000
60000
50000
40000
30000
20000
10000
0
charges
yes
no
smoker

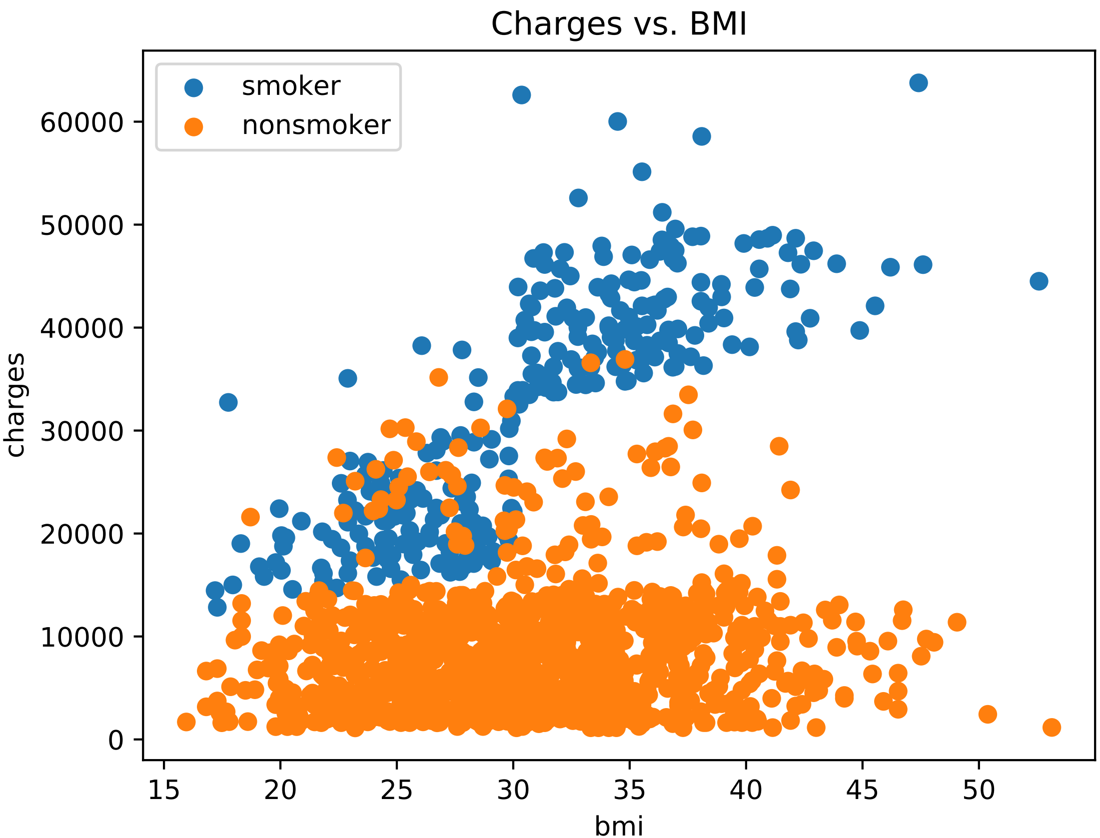


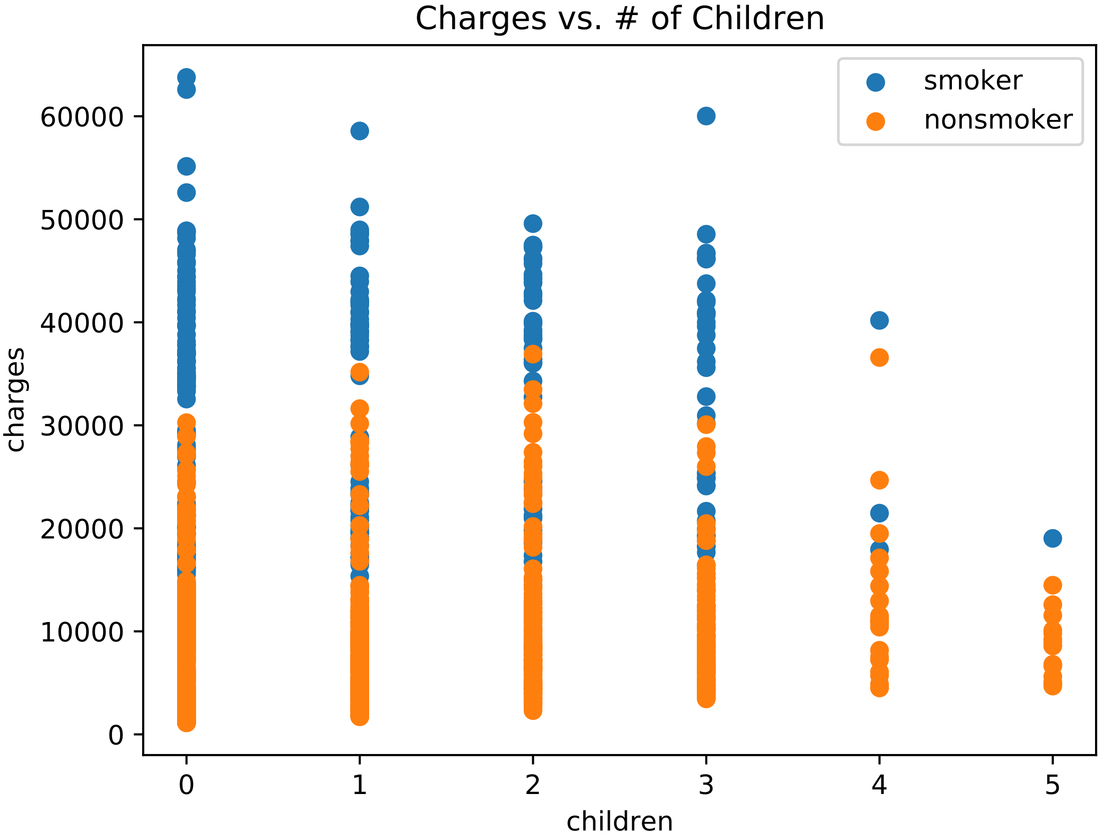
70000
60000
50000
40000
30000
20000
10000
0
−10000
charges
southwest
southeast
northwest
northeast
region

It is clear from these plots that smokers, people with BMI over 30, and old people have higher individual medical costs. Let us now look at the other plots with these data points highlighted to see what their effect is on charges

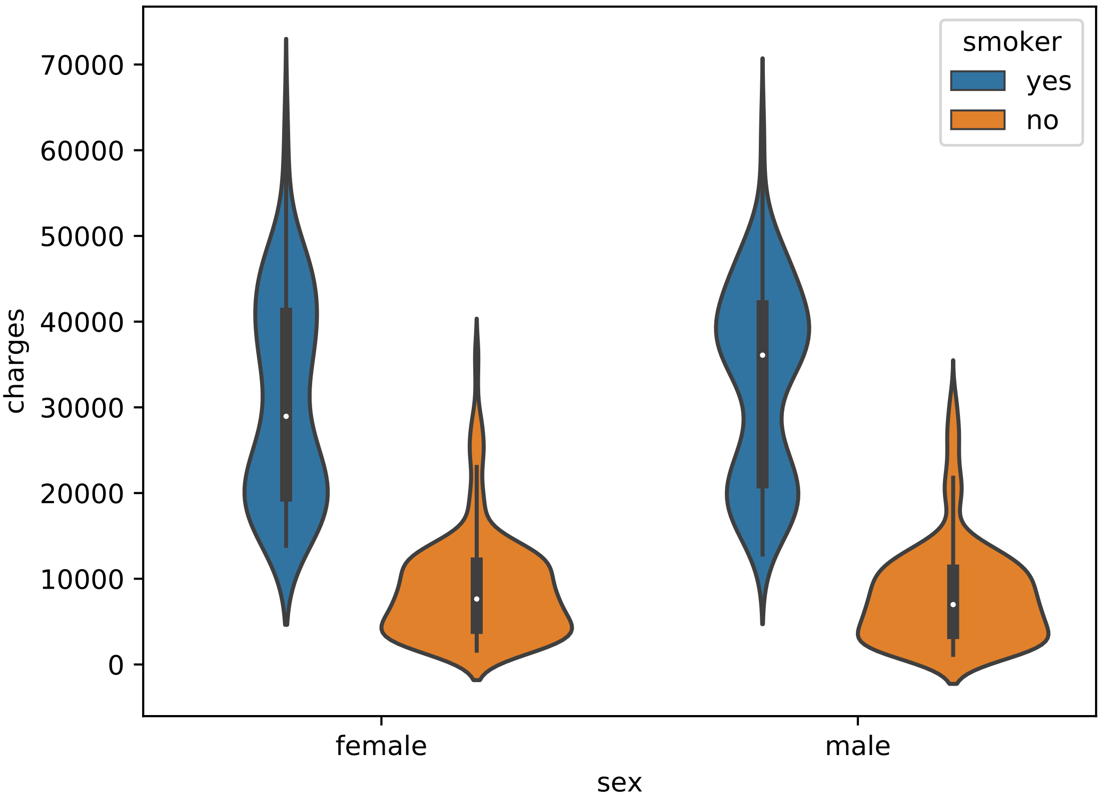


Charges vs. Age
smoker
nonsmoker
60000
50000
40000
30000
20000
10000
0
charges
20
30
40
50
60
age

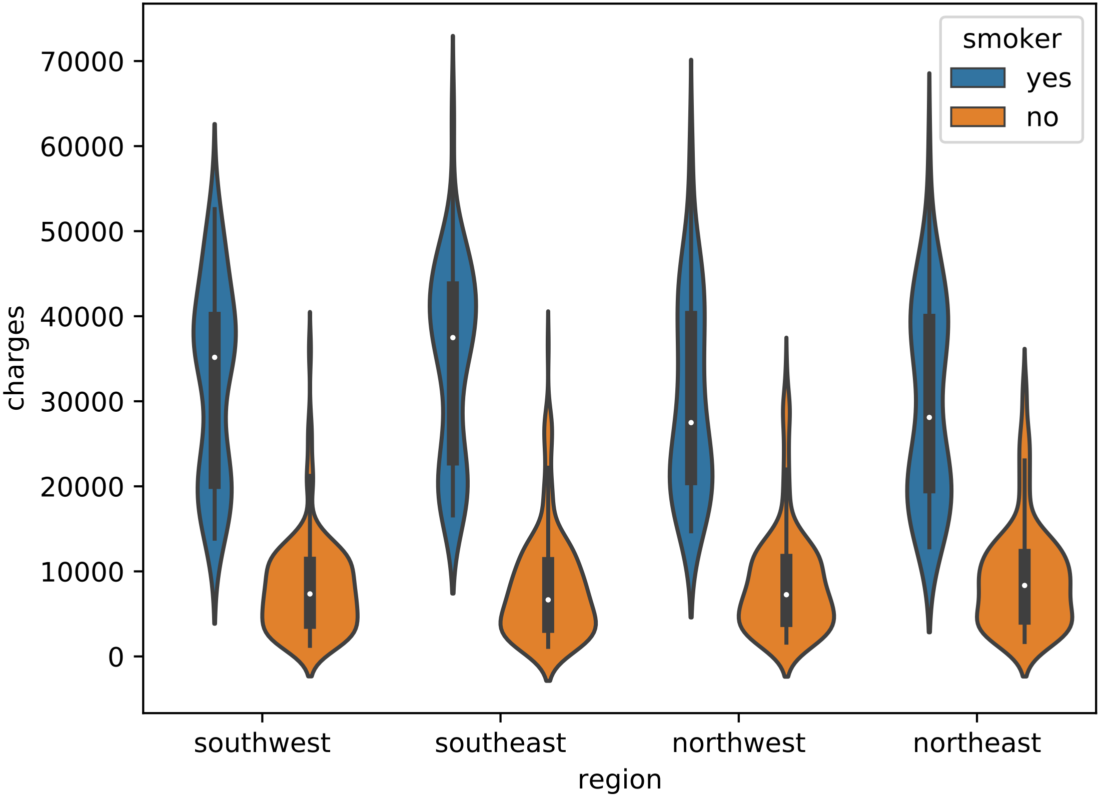

Charges vs. BMI
smoker
nonsmoker
charges
bmi
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50



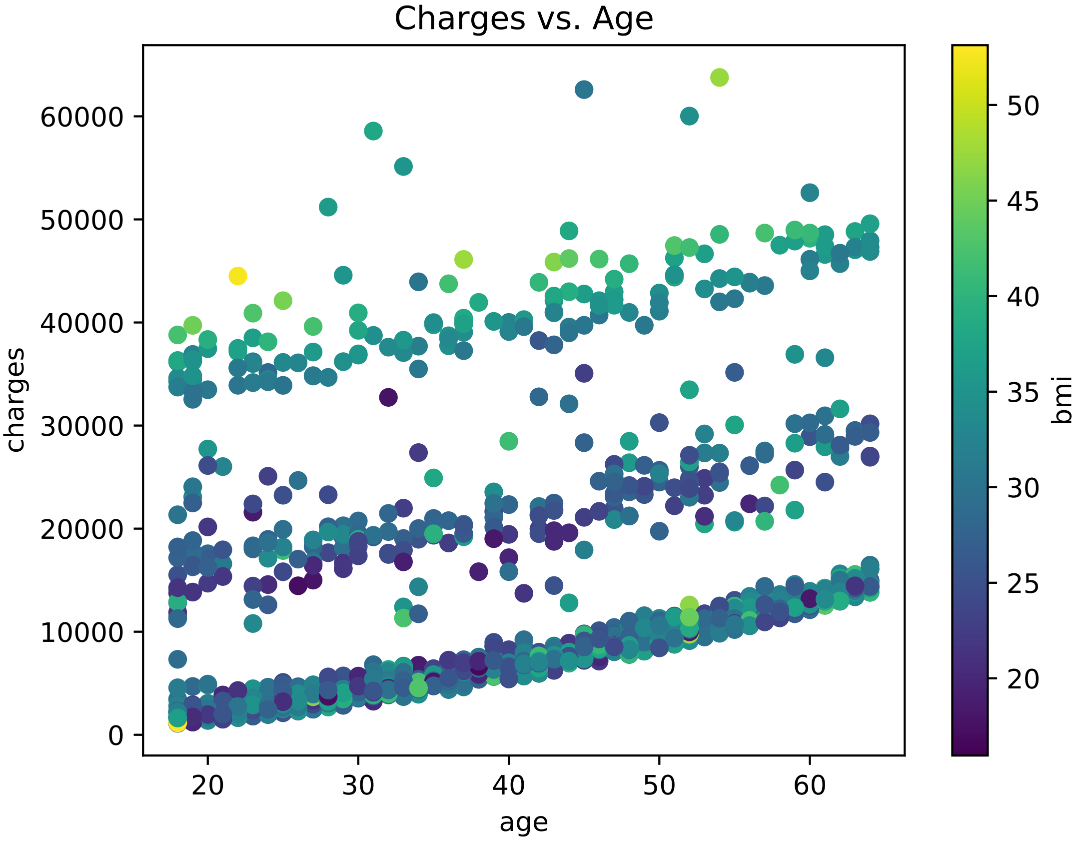
Charges vs. # of Children
smoker
nonsmoker
60000
50000
40000
30000
20000
10000
0
charges
0
1
2
3
4
5
children

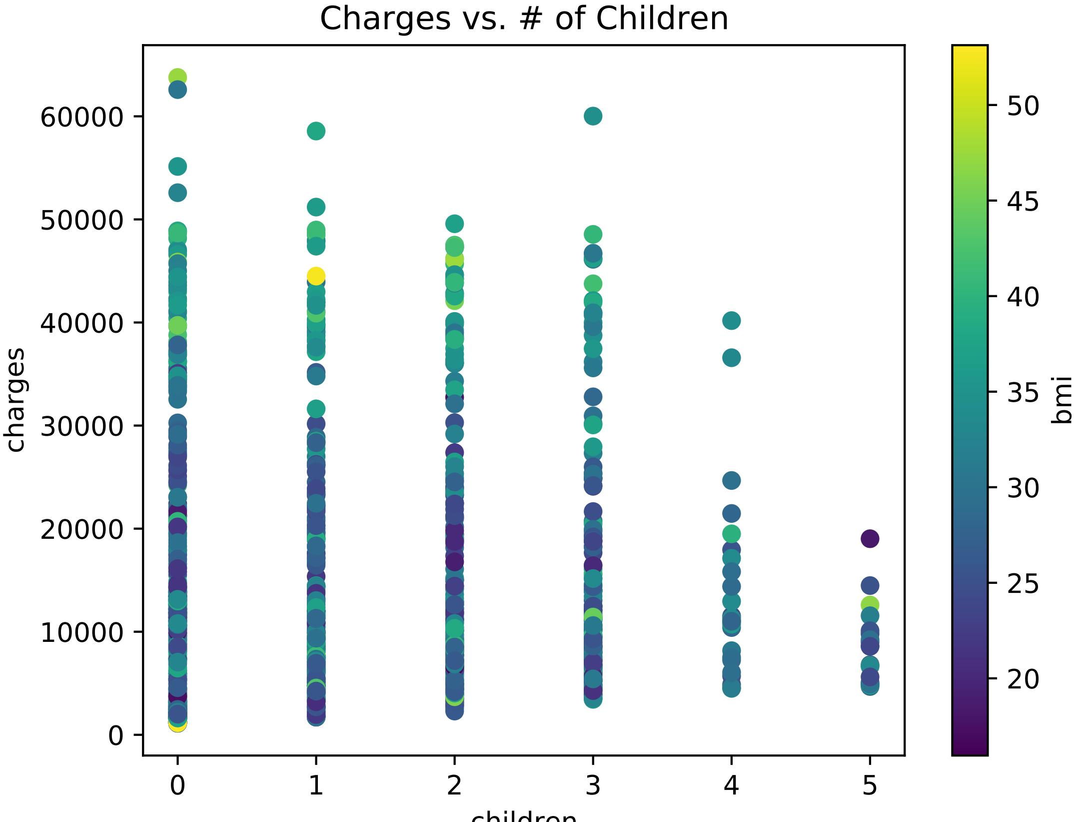


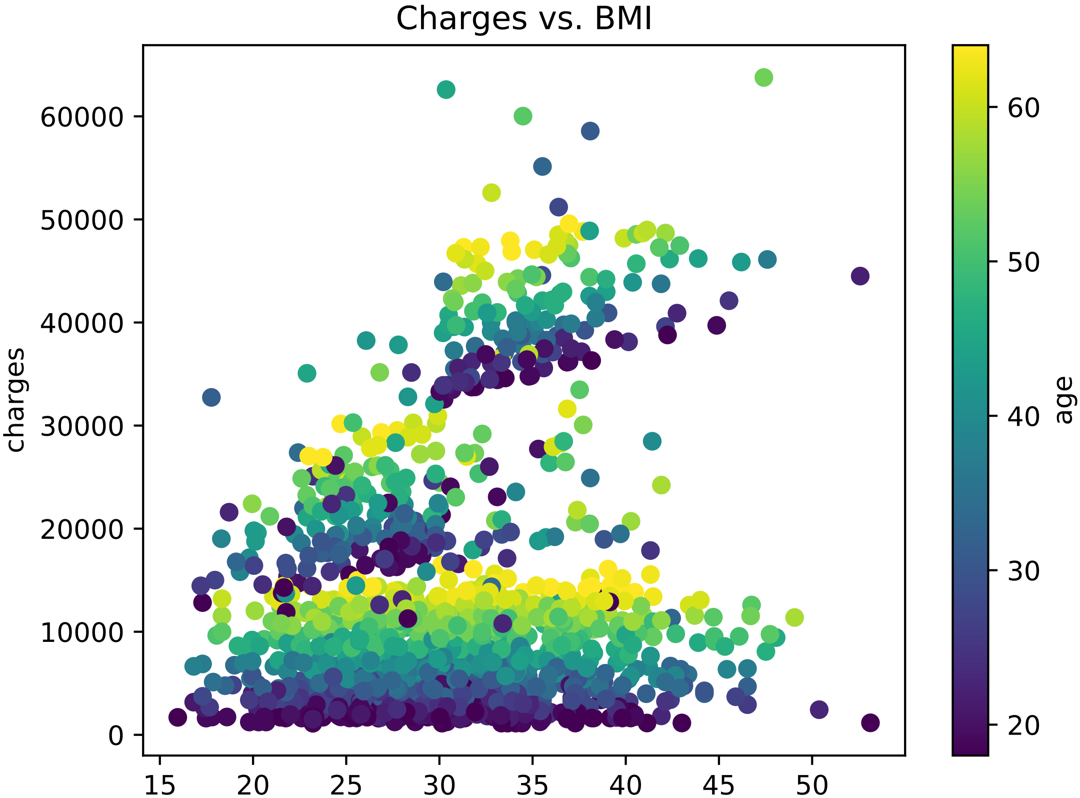
smoker
yes
no
70000
60000
50000
40000
30000
20000
10000
0
charges
female
male
sex

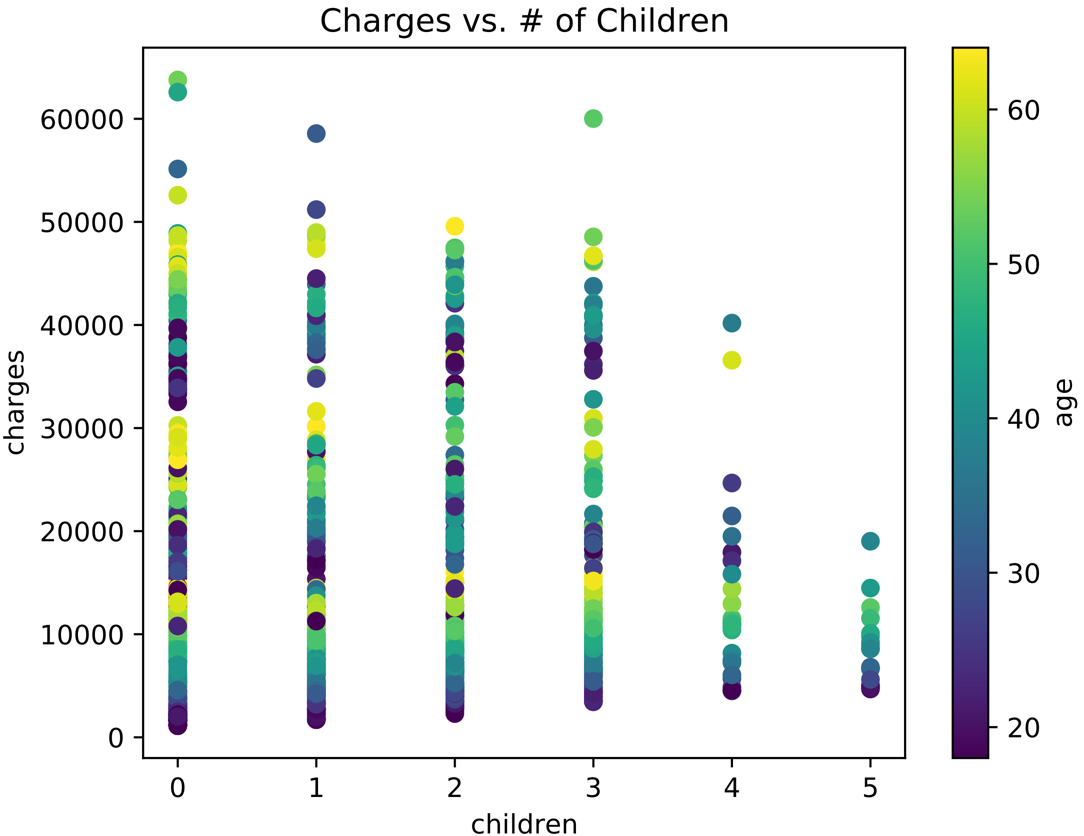


smoker
yes
no
charges
70000
60000
50000
40000
30000
20000
10000
0
southwest
southeast
northwest
northeast
region


Charges vs. Age
charges
age
bmi
60000
50000
40000
30000
20000
10000
0
20
30
40
50
60
50
45
40
35
30
25
20


Charges vs. # of Children
charges
children
bmi
0
10000
20000
30000
40000
50000
60000
0
1
2
3
4
5
20
25
30
35
40
45
50

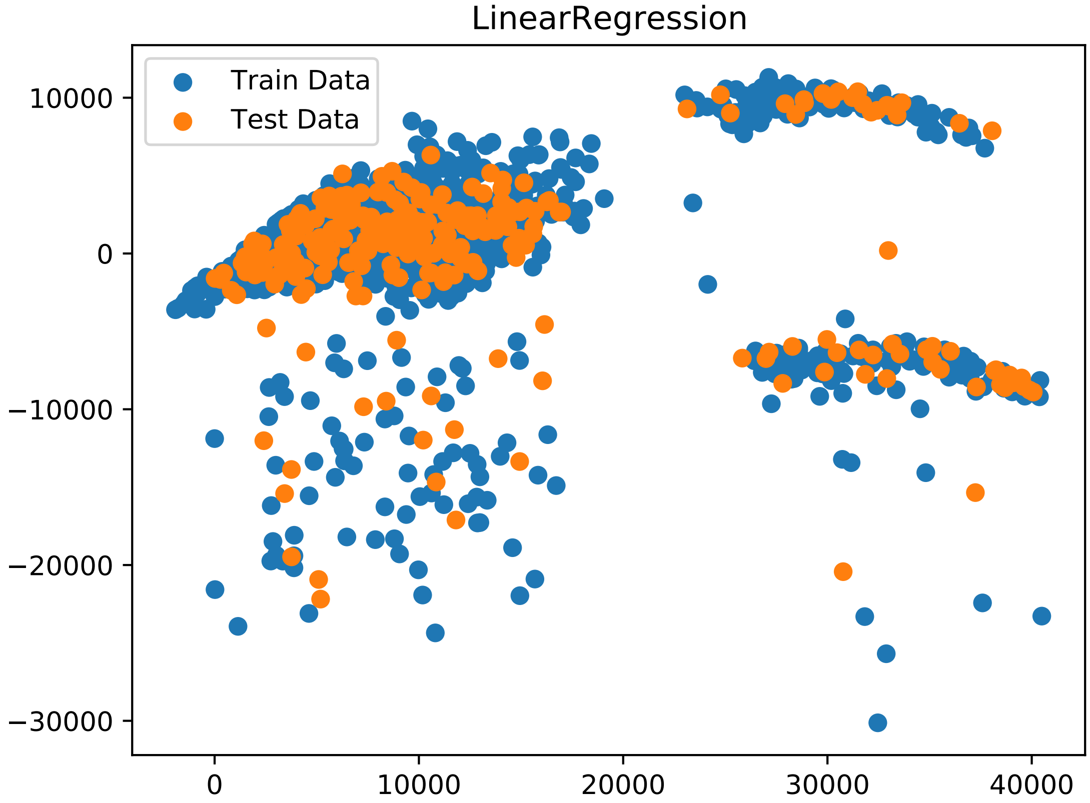

Charges vs. BMI
charges
60000
50000
40000
30000
20000
10000
0
15
20
25
30
35
40
45
50
bmi
age
60
50
40
30
20

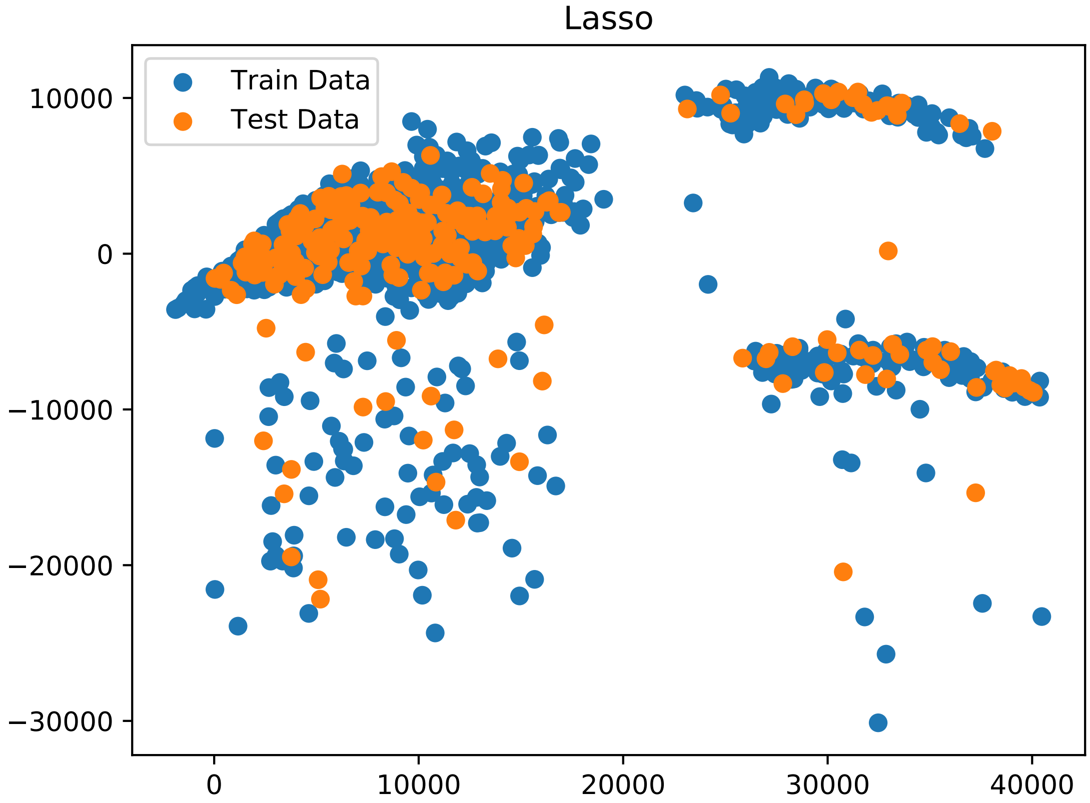


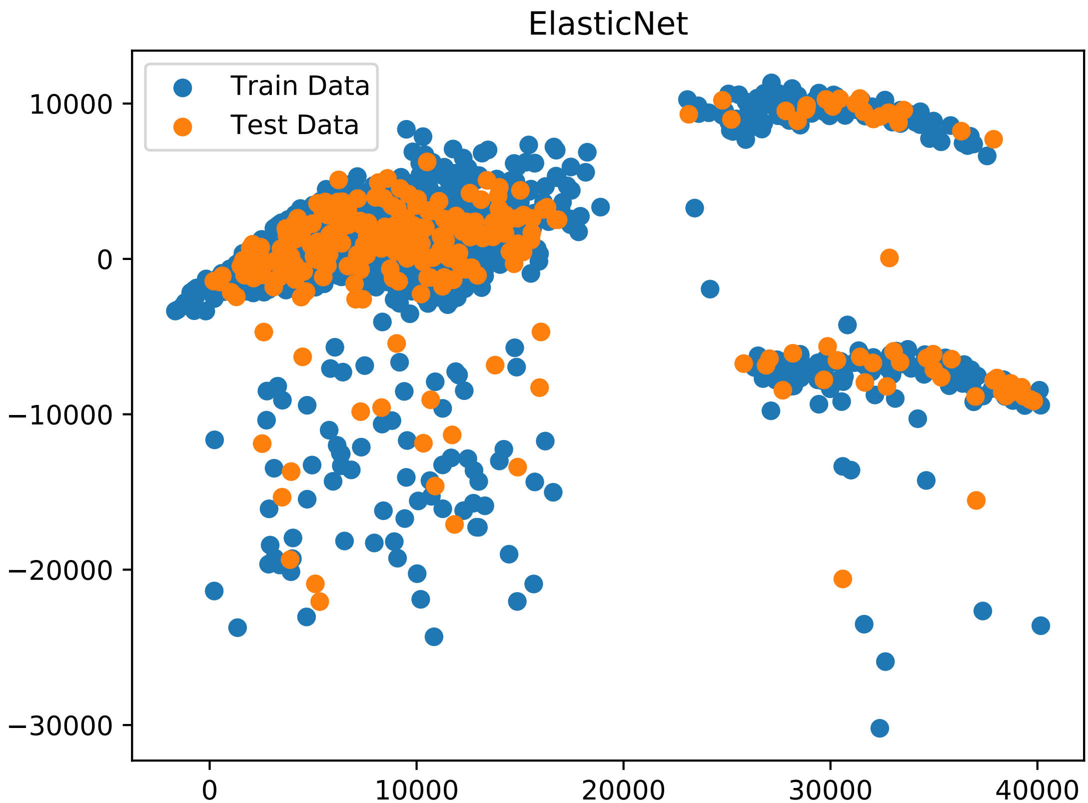
Charges vs. # of Children
charges
children
age
0
10000
20000
30000
40000
50000
60000
0
1
2
3
4
5
20
30
40
50
60

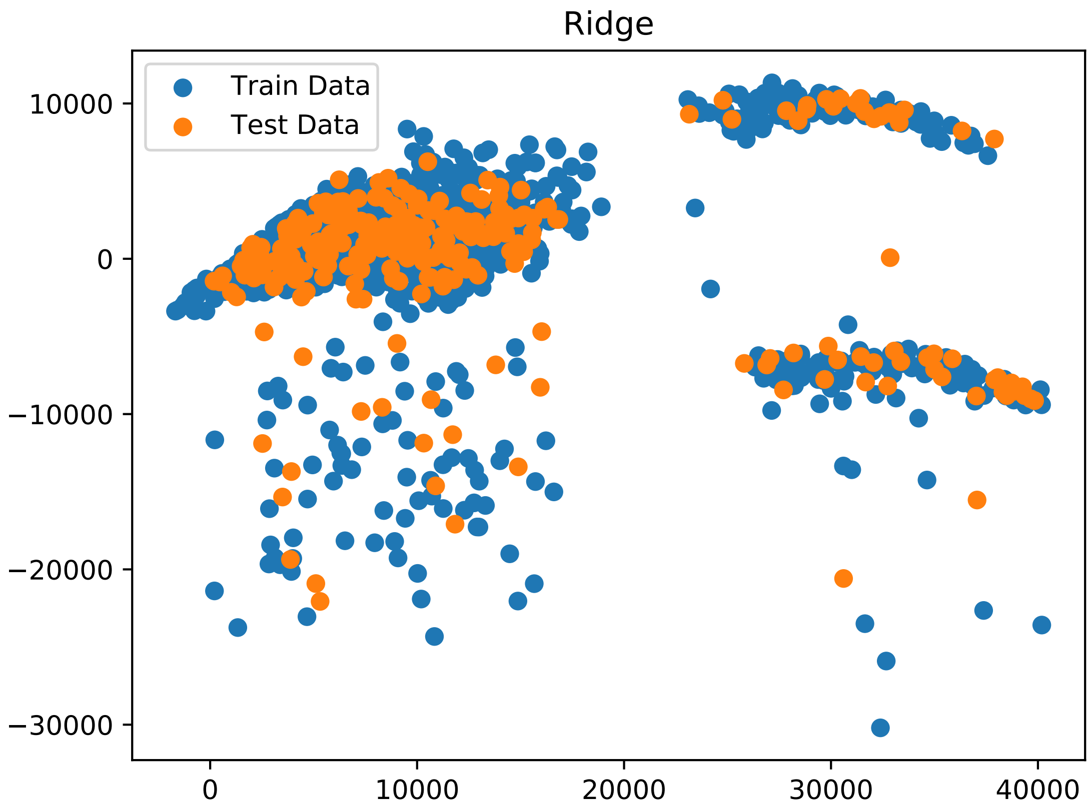
```
                          OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.737
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     371.7
Date:                Mon, 27 Apr 2020   Prob (F-statistic):          1.85e-301
Time:                        16:57:48   Log-Likelihood:                -10851.
No. Observations:                1070   AIC:                         2.172e+04
Df Residuals:                    1061   BIC:                         2.177e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            -2661.4467    685.554     -3.882      0.000   -4006.643   -1316.250
age               1.167e+04    622.384     18.751      0.000    1.04e+04    1.29e+04
bmi               1.249e+04   1197.929     10.424      0.000    1.01e+04    1.48e+04
children          2184.5506    782.920      2.790      0.005     648.304    3720.797
sex_male           -15.4637    378.193     -0.041      0.967    -757.555     726.627
smoker_yes        2.361e+04    470.606     50.159      0.000    2.27e+04    2.45e+04
region_northeast   761.9487    543.309      1.402      0.161    -304.134    1828.031
region_northwest   501.8160    540.283      0.929      0.353    -558.329    1561.961
region_southeast  -151.3301    531.148     -0.285      0.776   -1193.549     890.889
==============================================================================
Omnibus:                      256.825   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              620.044
Skew:                           1.279   Prob(JB):                    2.29e-135
Kurtosis:                       5.715   Cond. No.                         9.60
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```


LinearRegression
Train Data
Test Data
10000
0
−10000
−20000
−30000
0
10000
20000
30000
40000


Lasso
Train Data
Test Data
10000
0
−10000
−20000
−30000
0
10000
20000
30000
40000


ElasticNet
Train Data
Test Data
10000
0
−10000
−20000
−30000
0
10000
20000
30000
40000



Ridge
Train Data
Test Data
10000
0
−10000
−20000
−30000
0
10000
20000
30000
40000


RandomForestRegressor
Train Data
Test Data
15000
10000
5000
0
−5000
−10000
−15000
−20000
0
10000
20000
30000
40000
50000
60000

{"rmse": {"model": "LinearRegression", "train": 6140.157418880165, "test": 5641.626558850189}}
{"r2_scores": {"model": "LinearRegression", "train": 0.7370262574551634, "test": 0.7999876970680434}}
{"rmse": {"model": "Lasso", "train": 6140.165390774906, "test": 5642.508431546825}}
{"r2_scores": {"model": "Lasso", "train": 0.737025574606105, "test": 0.7999251622040369}}
{"rmse": {"model": "ElasticNet", "train": 6141.020521734997, "test": 5650.841822172486}}
{"r2_scores": {"model": "ElasticNet", "train": 0.7369523214548549, "test": 0.7993337468743069}}
{"rmse": {"model": "Ridge", "train": 6140.912604661044, "test": 5650.1775043410025}}
{"r2_scores": {"model": "Ridge", "train": 0.7369615665261751, "test": 0.7993809250960761}}
{"rmse": {"model": "RandomForestRegressor", "train": 1885.9619268363351, "test": 4270.773952245565}}
{"r2_scores": {"model": "RandomForestRegressor", "train": 0.9751904310801135, "test": 0.8853797871562548}}

The Random Forest Regression is best suited for this dataset by far. Linear regression is slightly favored over the other regression models. This makes sense becuase the other models are more robust the more features there are