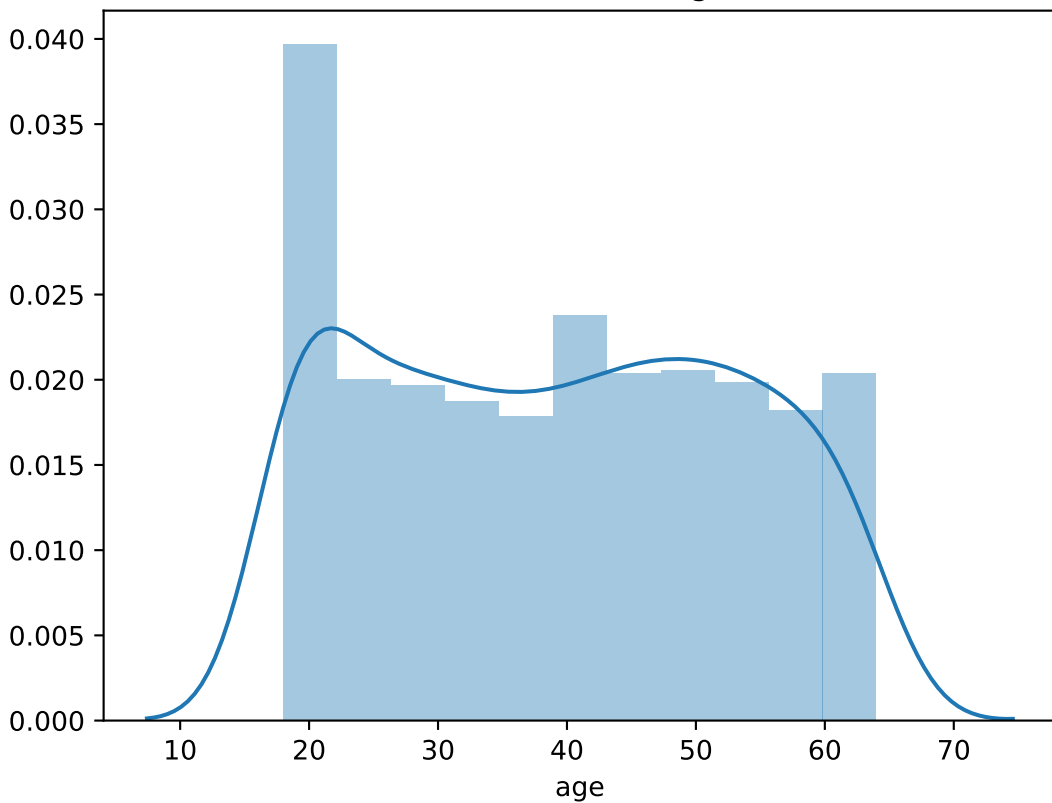
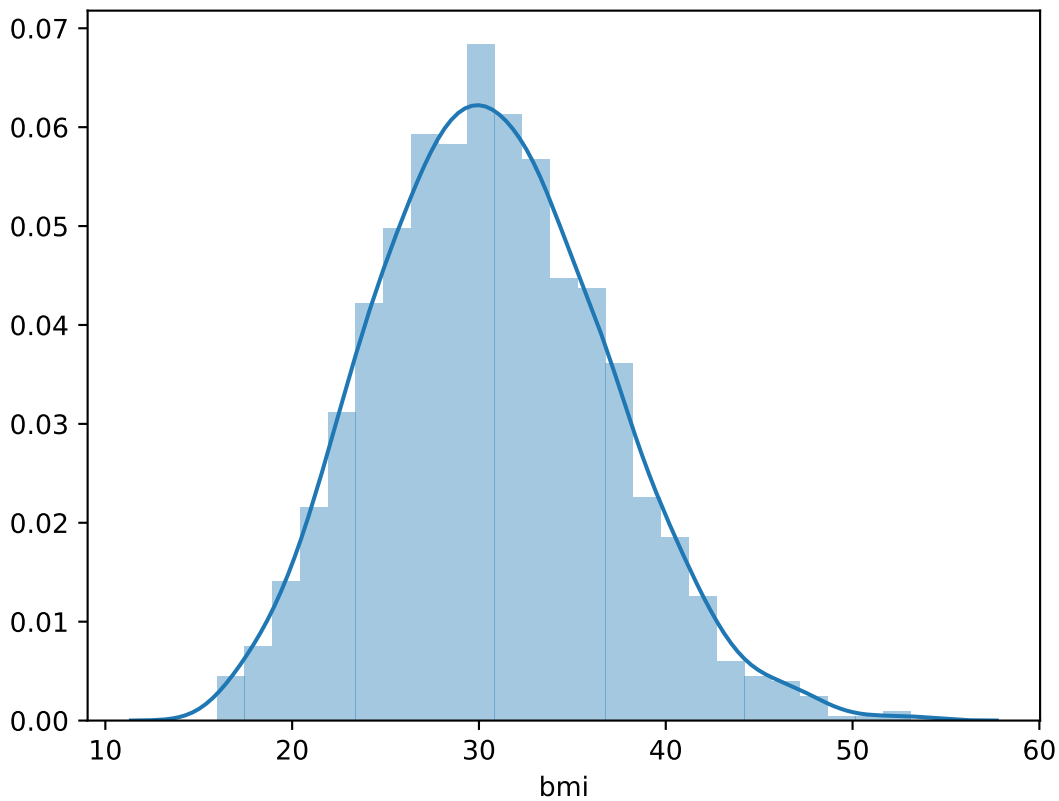


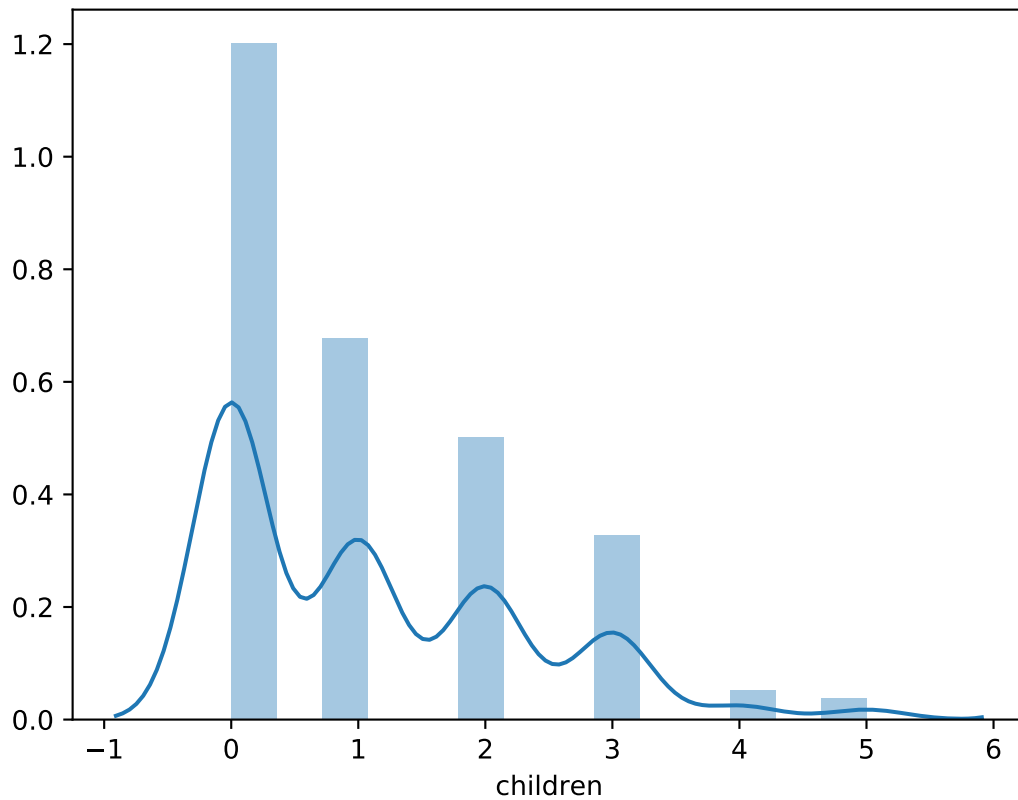
Distribution of Age

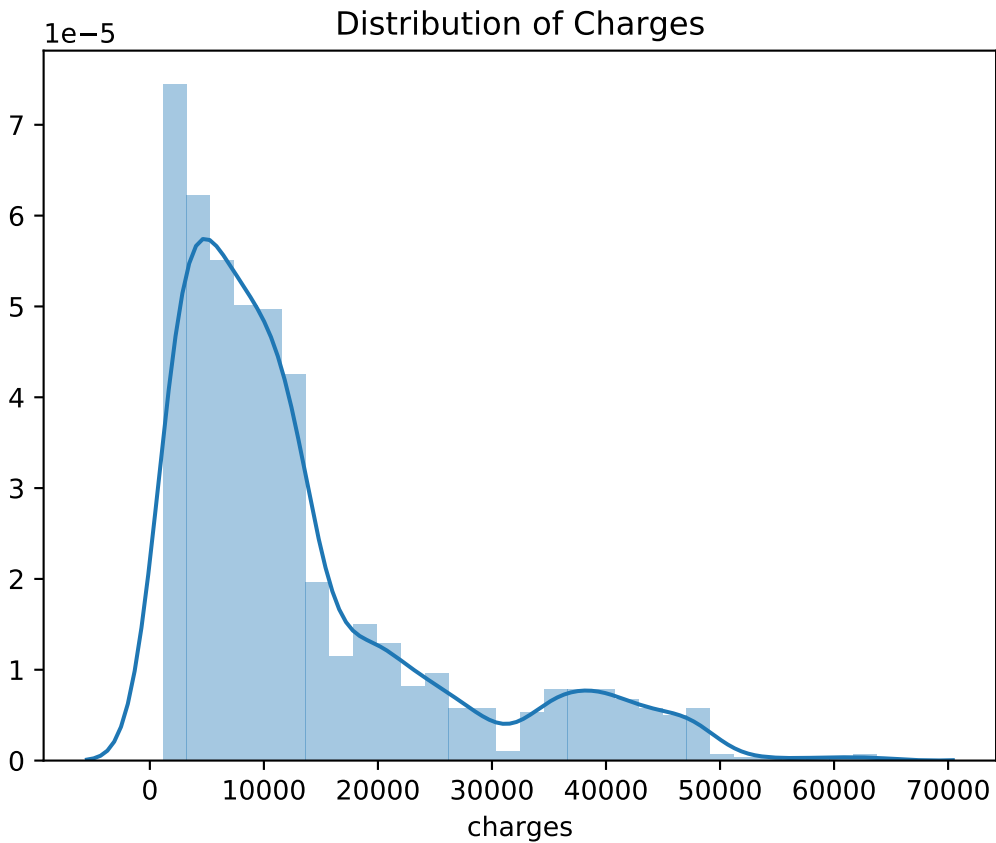


Distribution of BMI

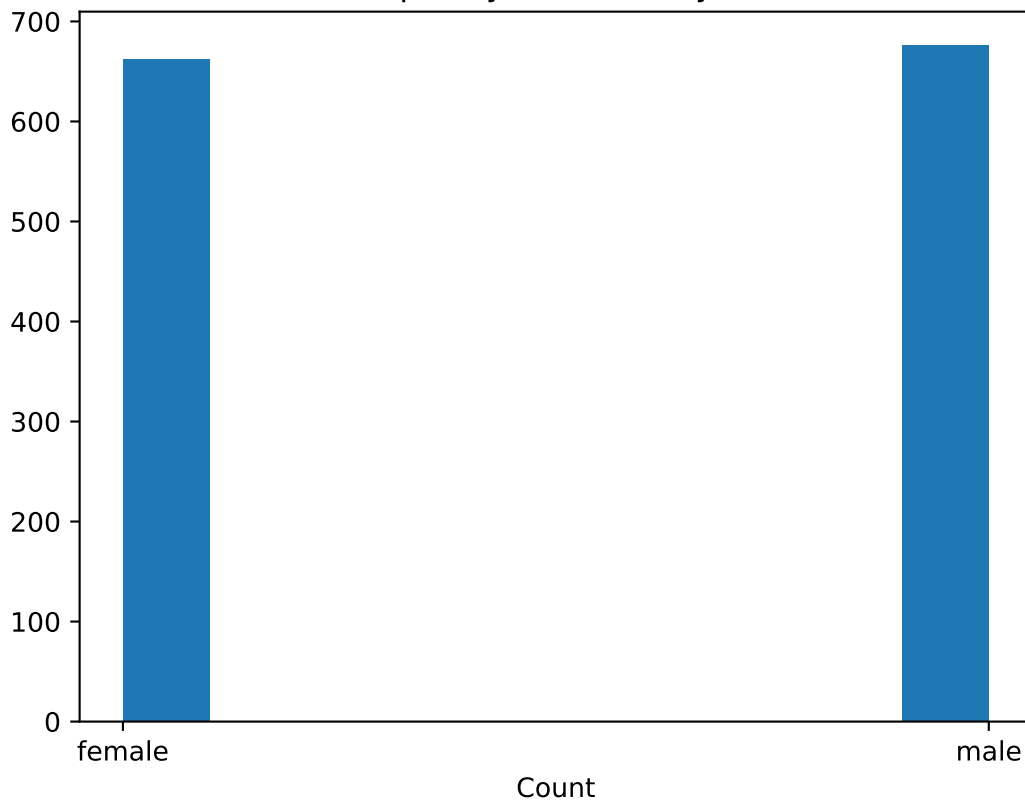


Distribution of # of Children

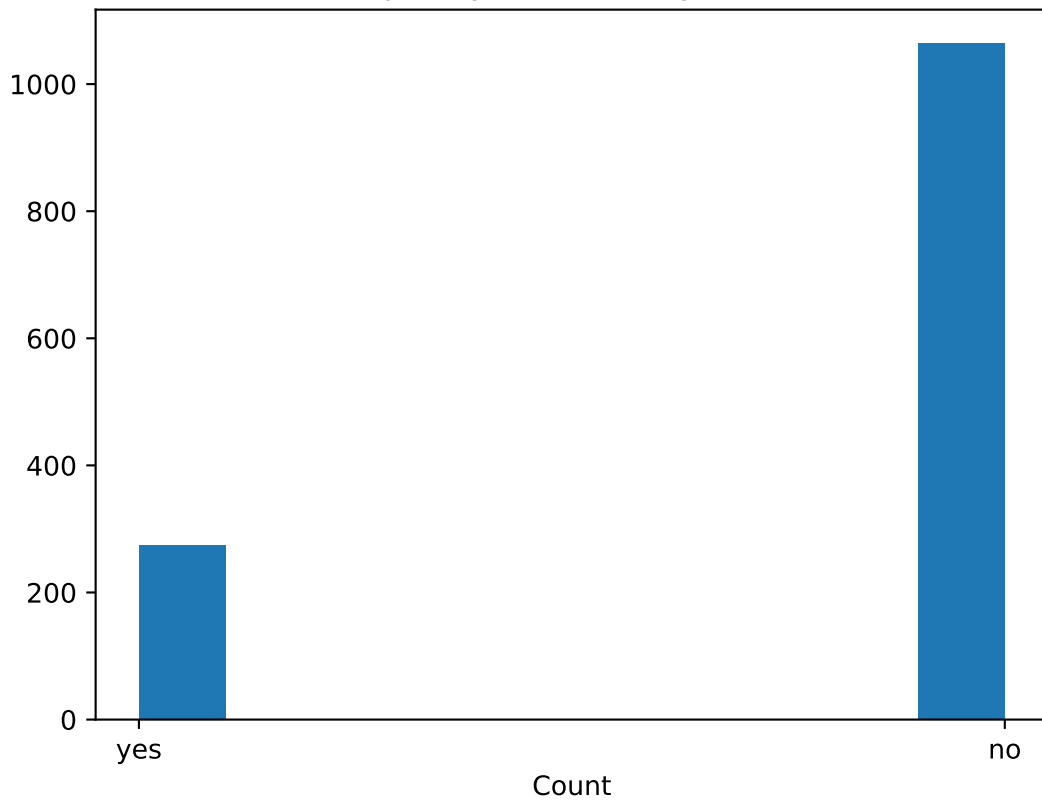




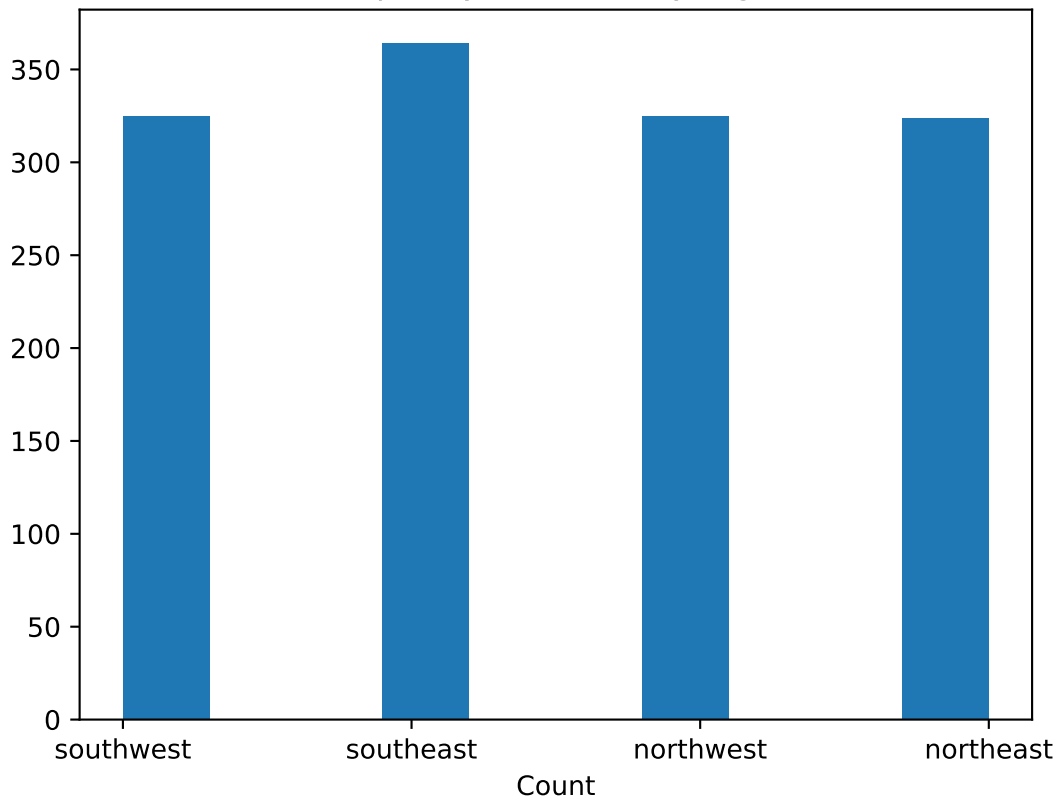
Frequency of Person by sex



Frequency of Person by smoker

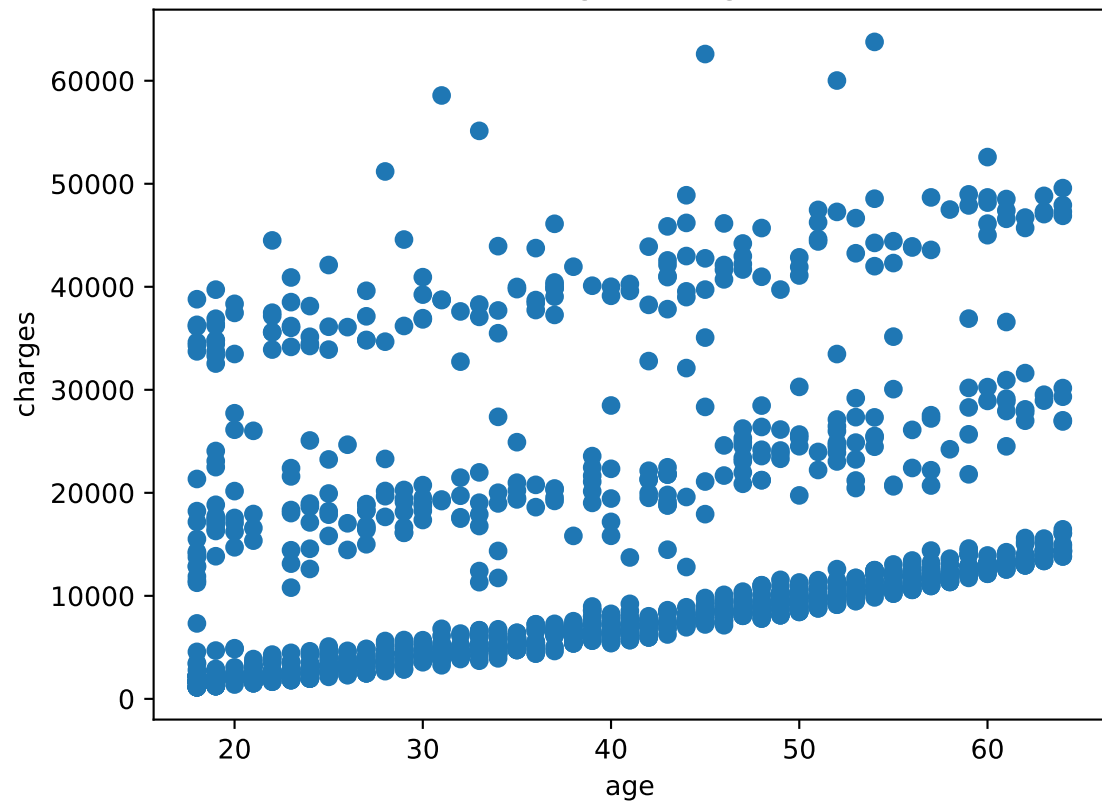


Frequency of Person by region

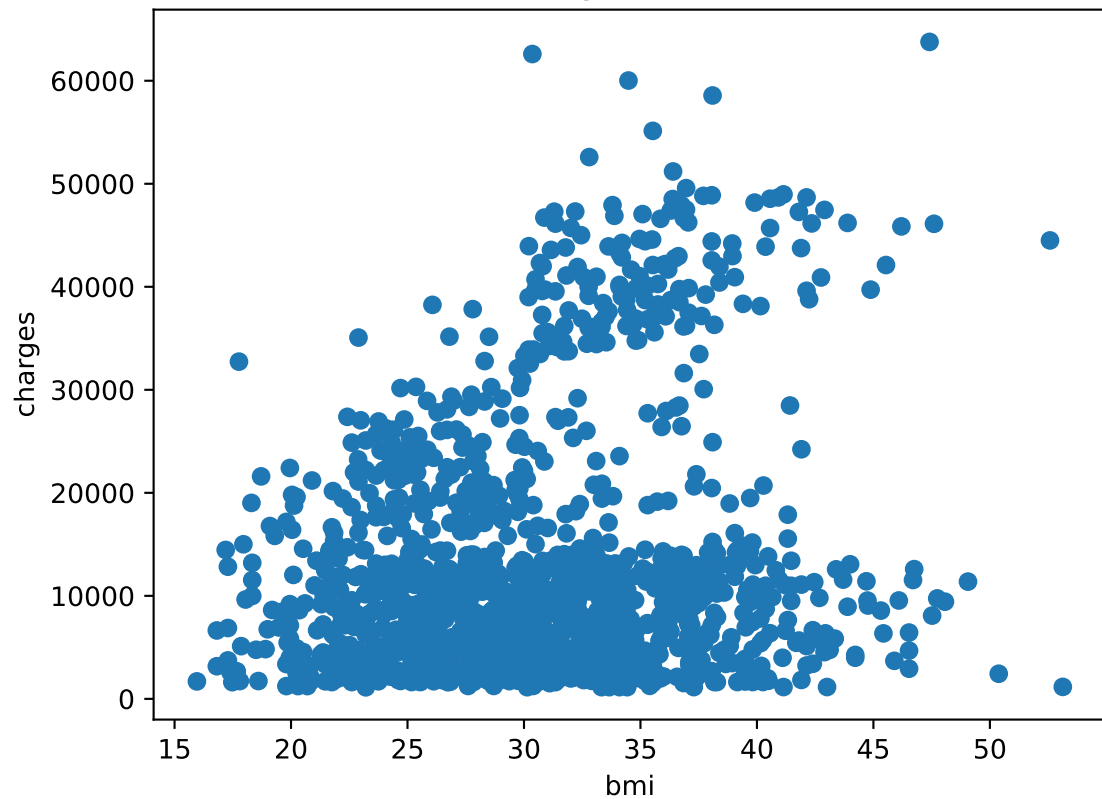


Now that we understand the data in the dataset, lets look at charges vs. each data point to see which have an effect

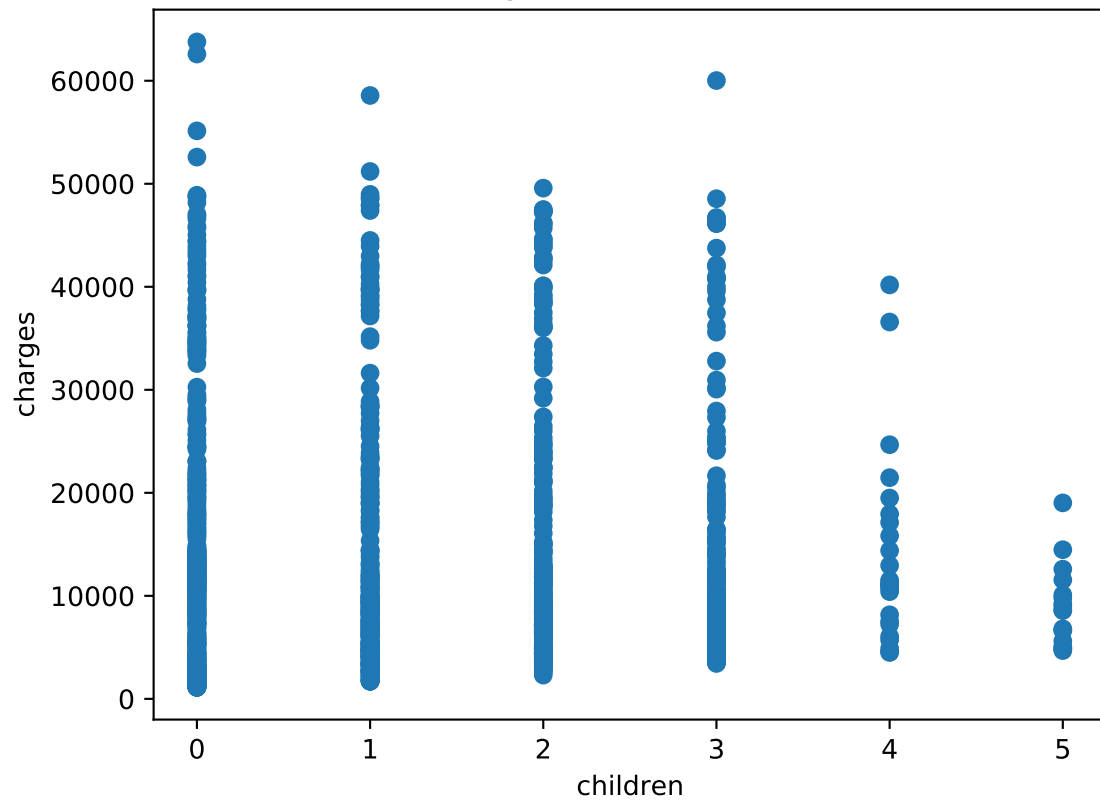
Charges vs. Age

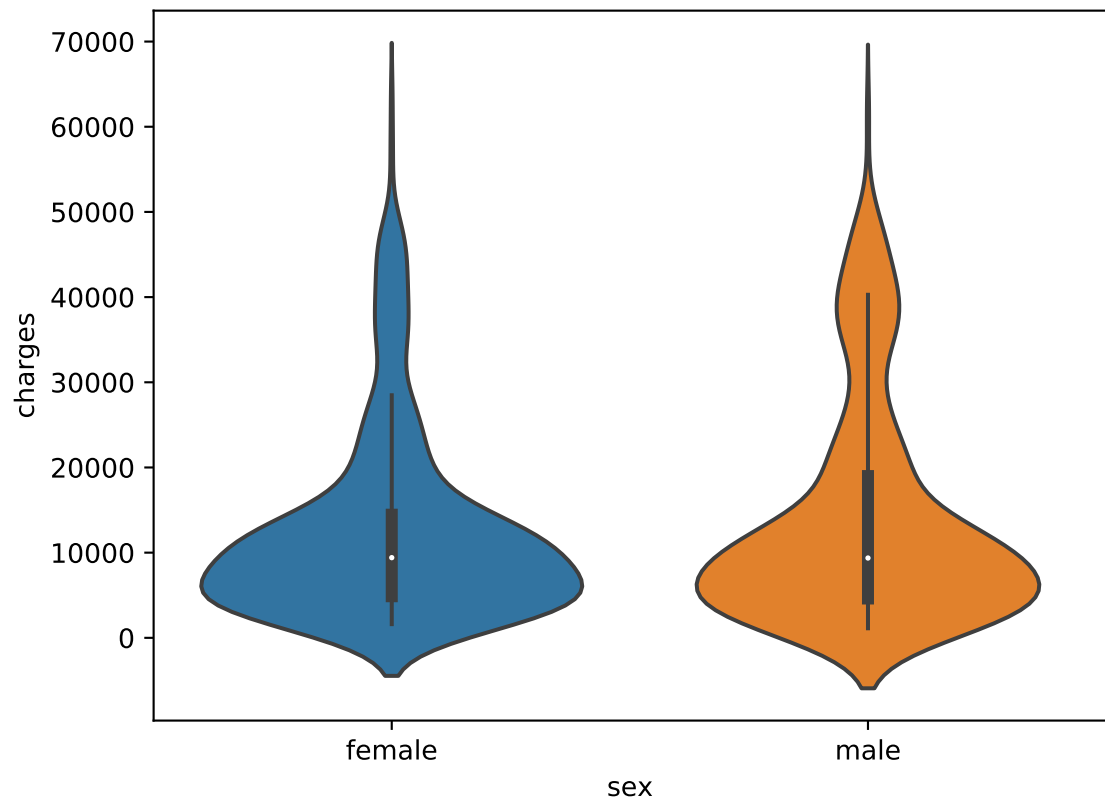


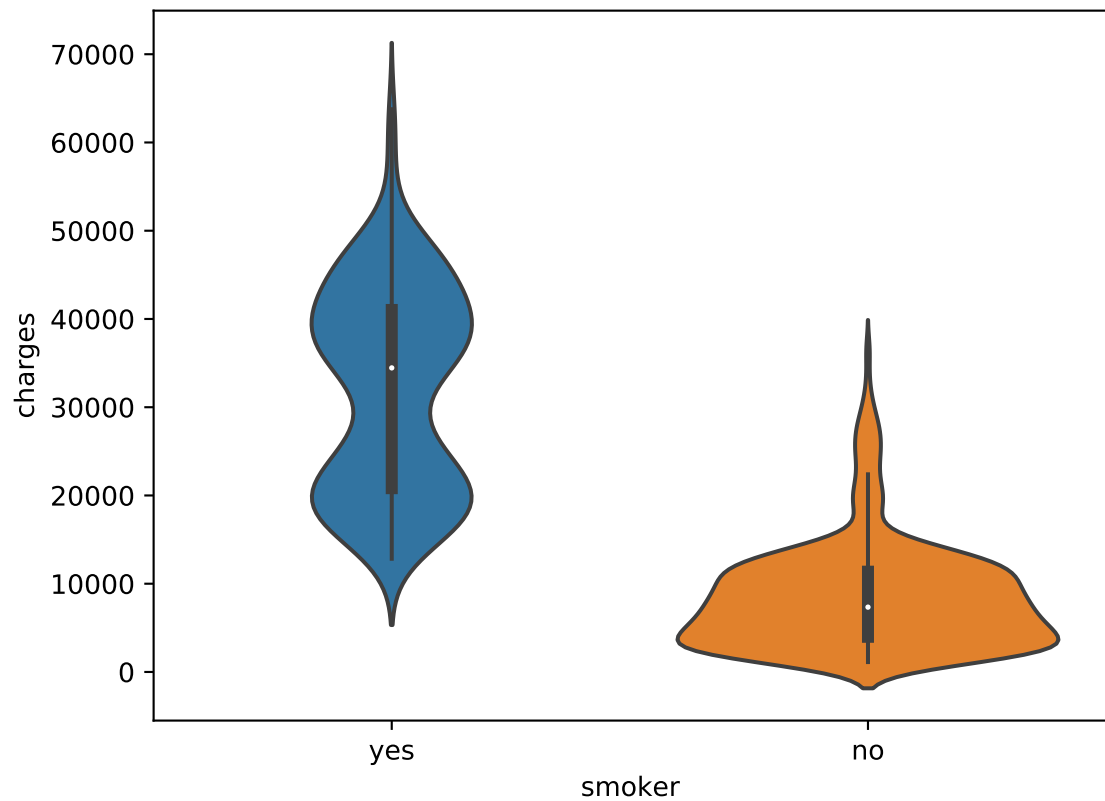
Charges vs. BMI

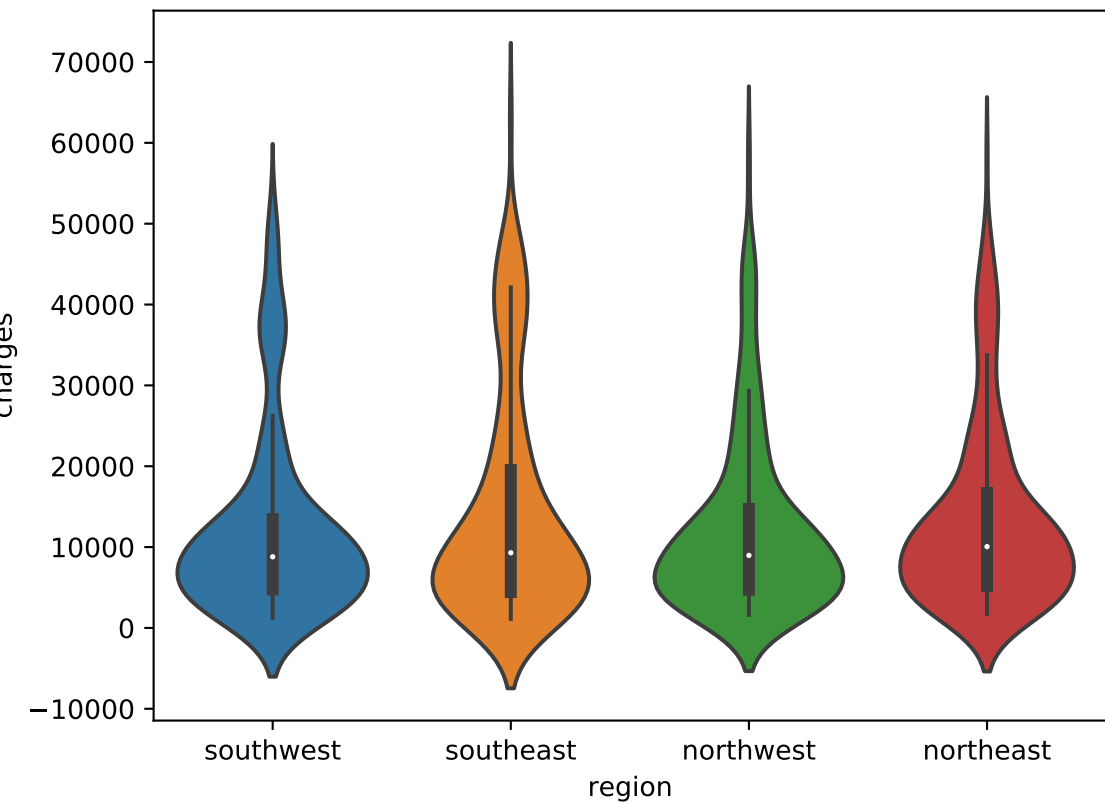


Charges vs. # of Children



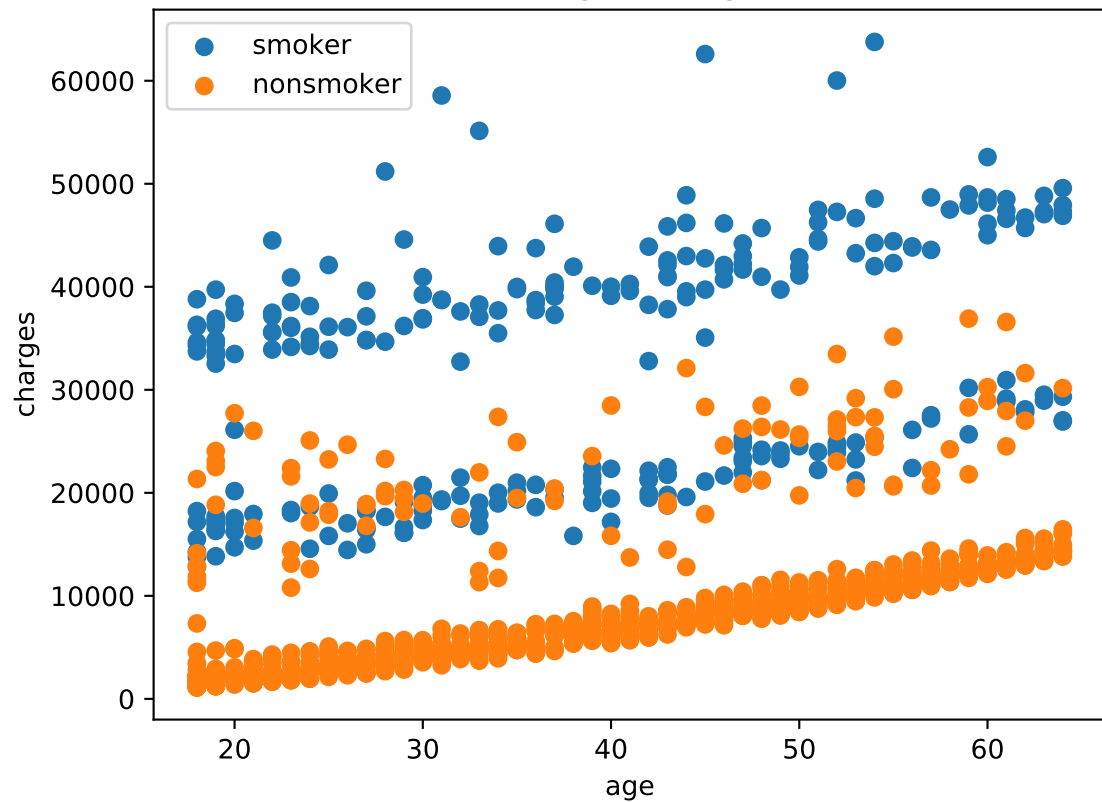




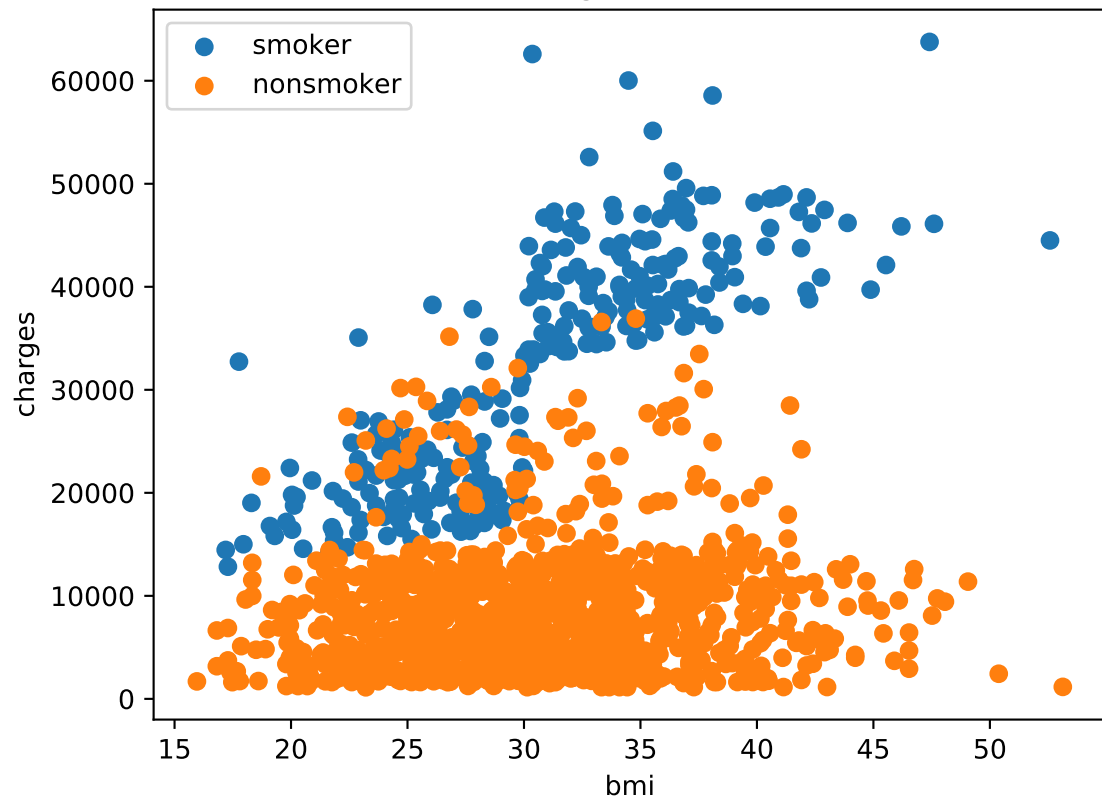


It is clear from these plots that smokers, people with BMI over 30, and old people have higher individual medical costs. Let us now look at the other plots with these data points highlighted to see what their effect is on charges

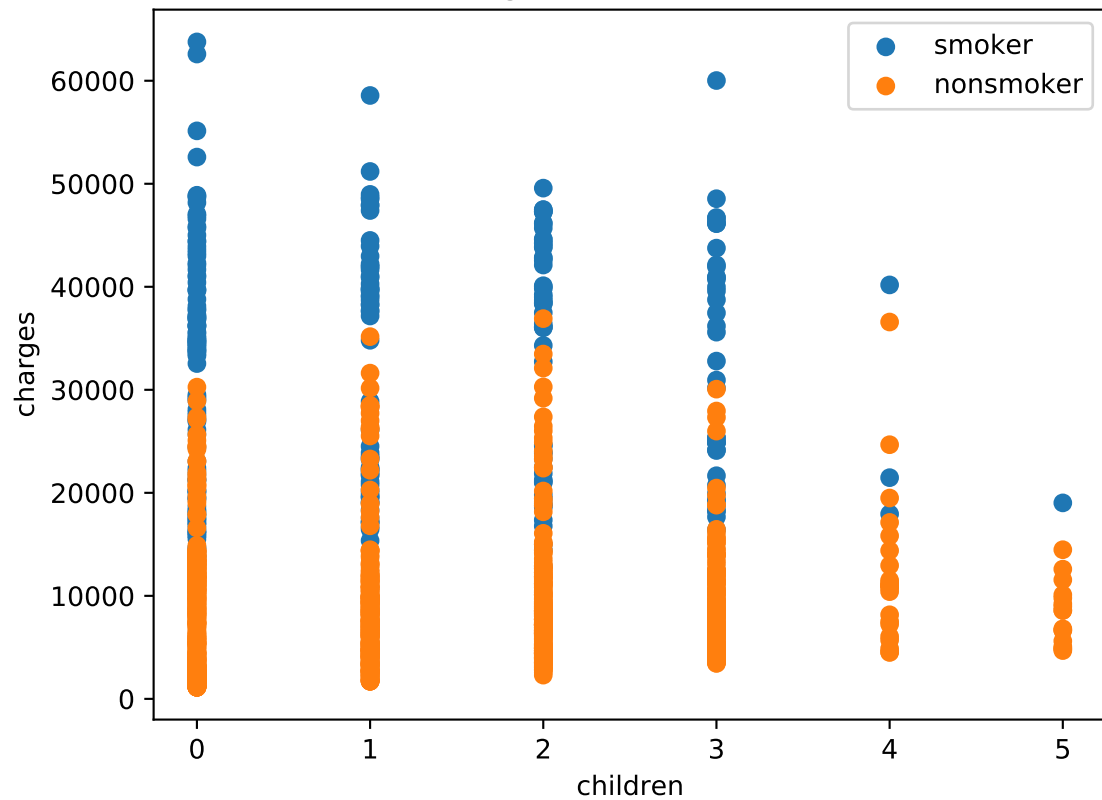
Charges vs. Age

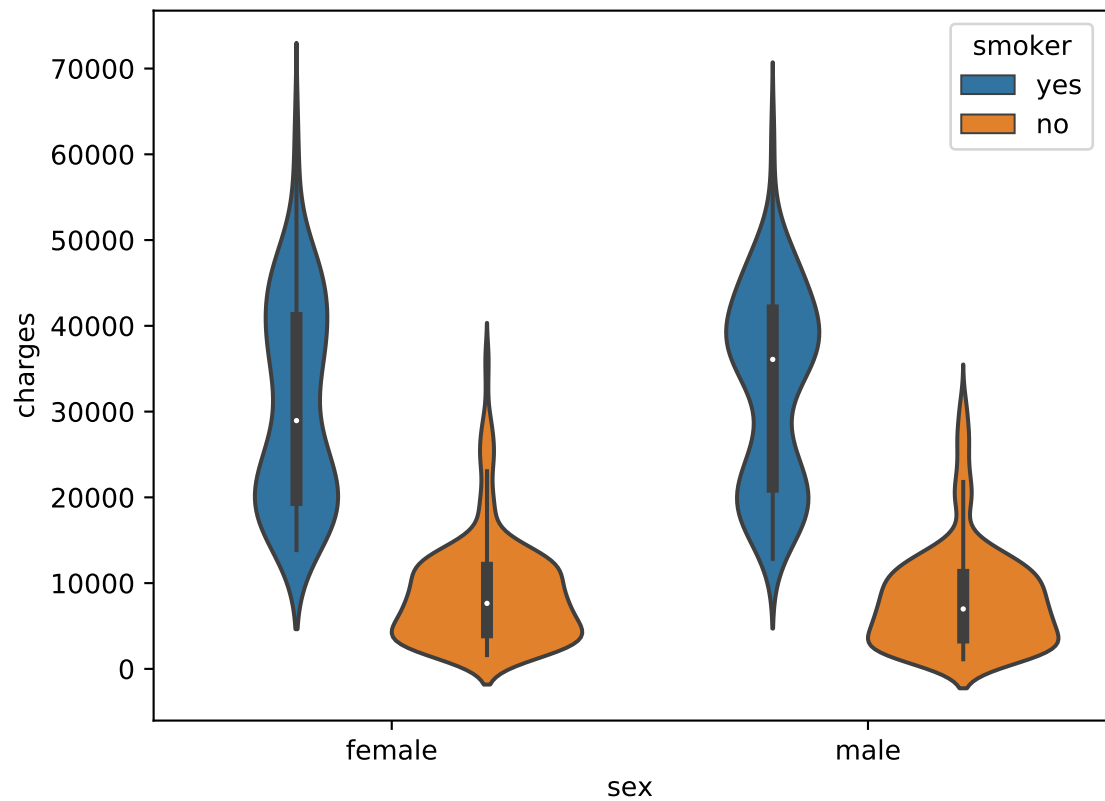


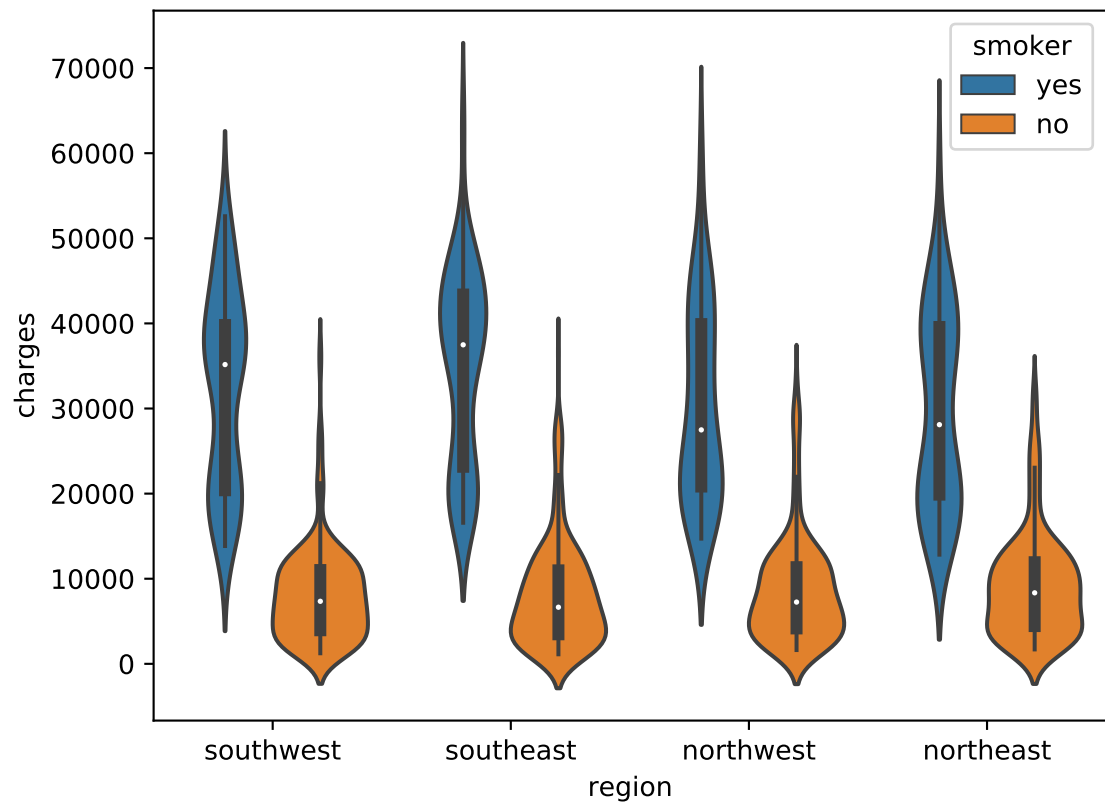
Charges vs. BMI



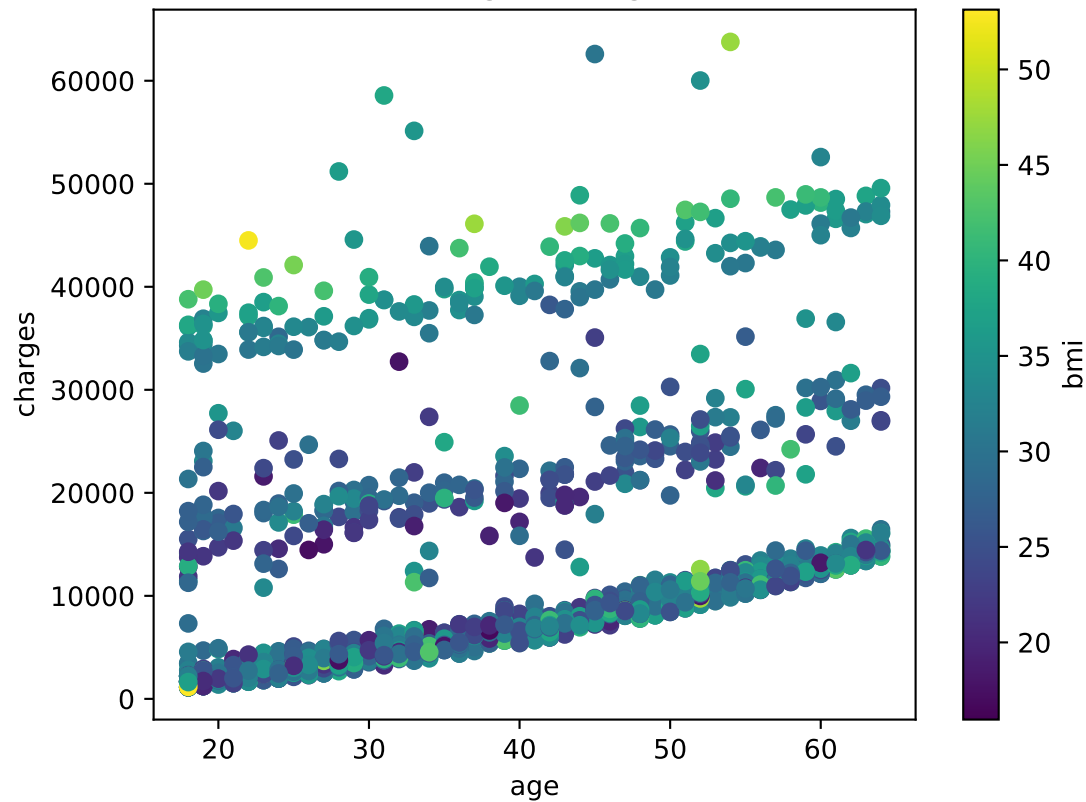
Charges vs. # of Children



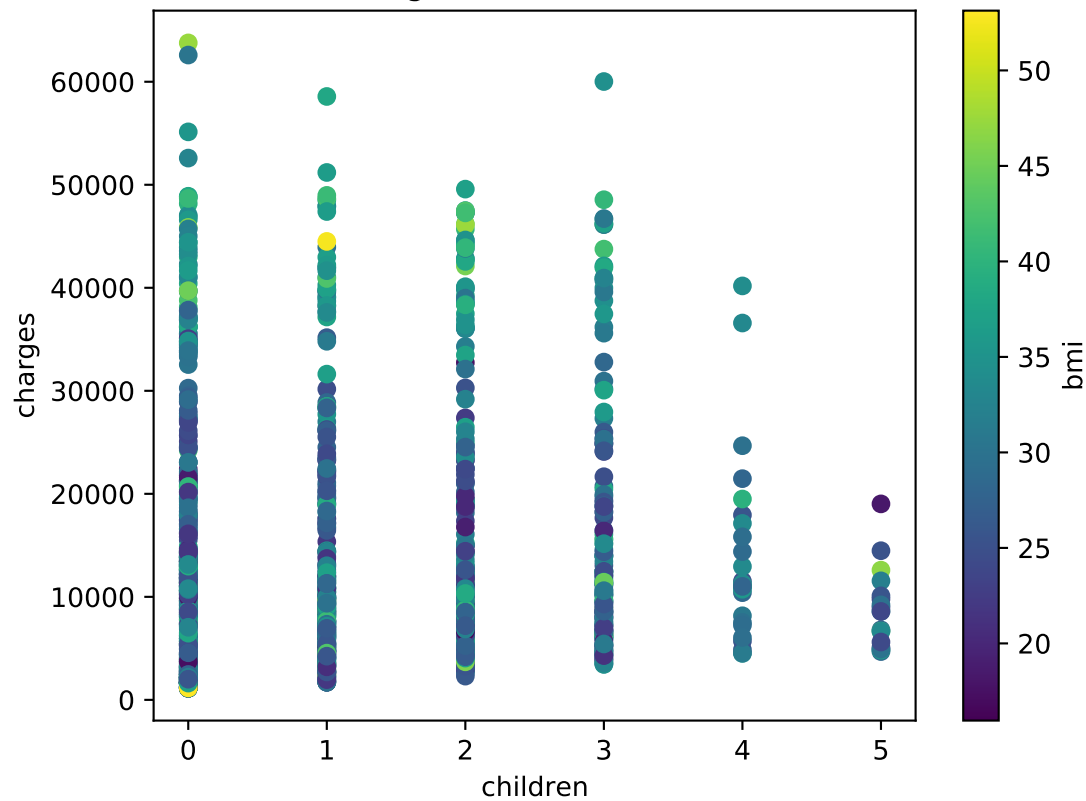




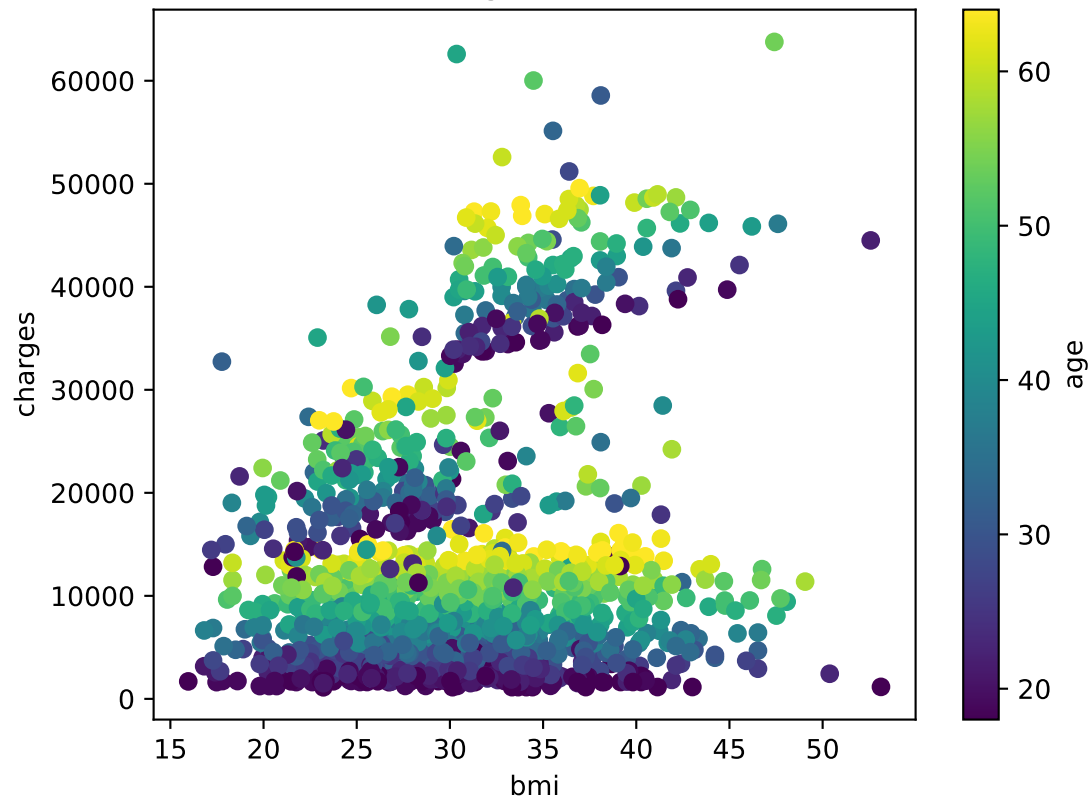
Charges vs. Age



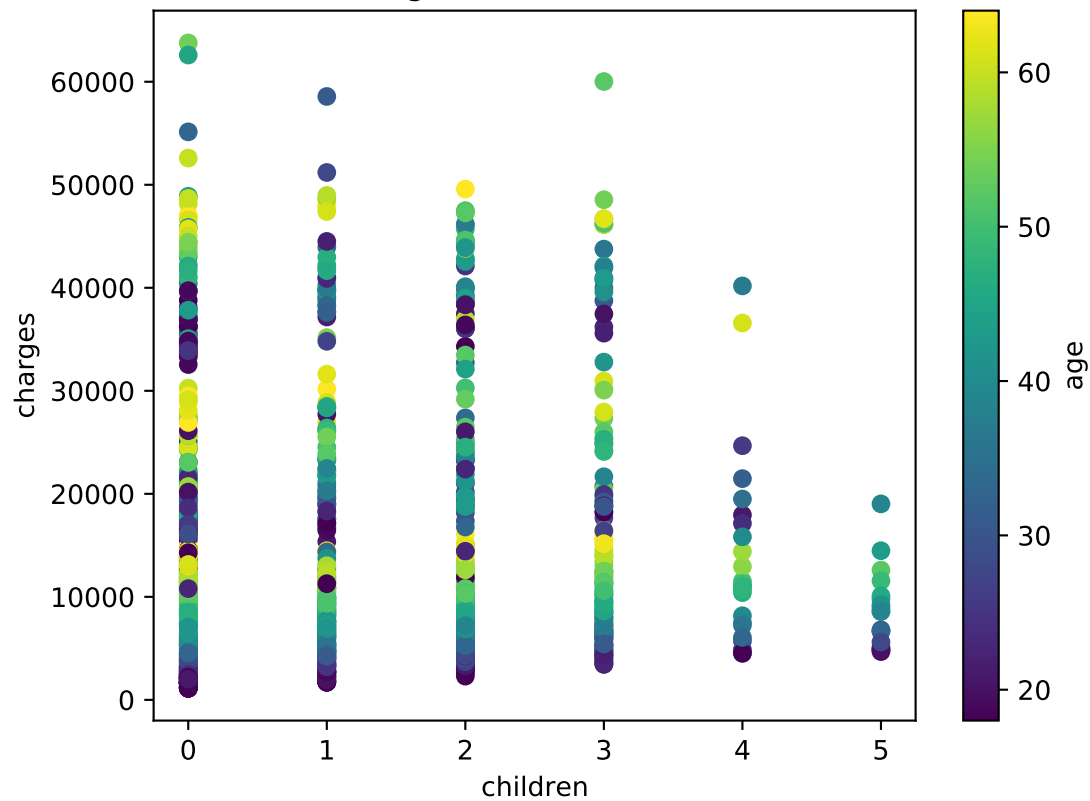
Charges vs. # of Children



Charges vs. BMI



Charges vs. # of Children



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:                   16:32:01  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

```

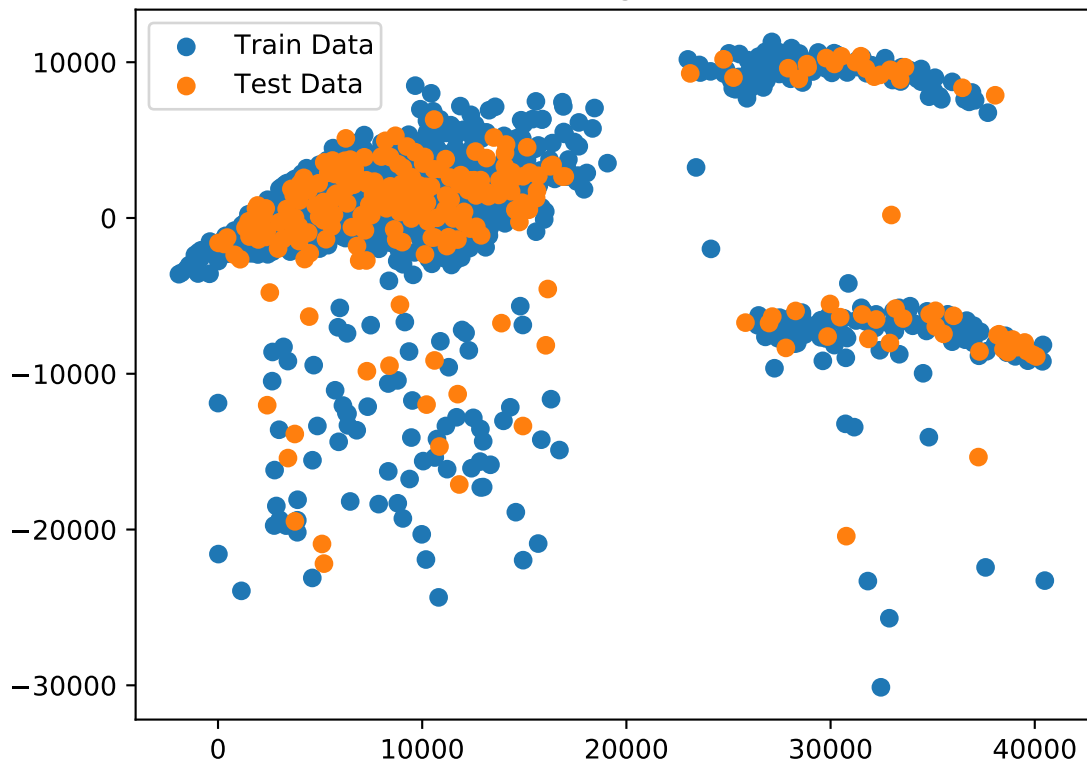
=====
Omnibus:                256.825  Durbin-Watson:          1.994
Prob(Omnibus):           0.000  Jarque-Bera (JB):        620.044
Skew:                    1.279  Prob(JB):                2.29e-135
Kurtosis:                5.715  Cond. No.                 9.60
=====

```

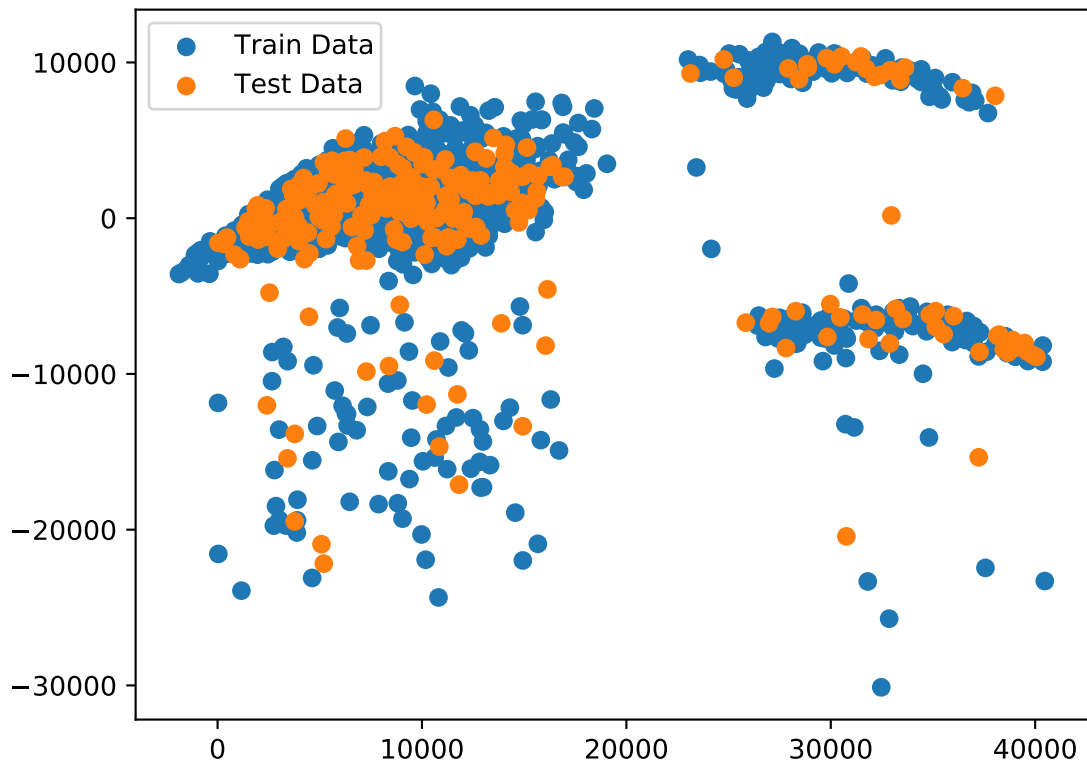
Warnings:

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

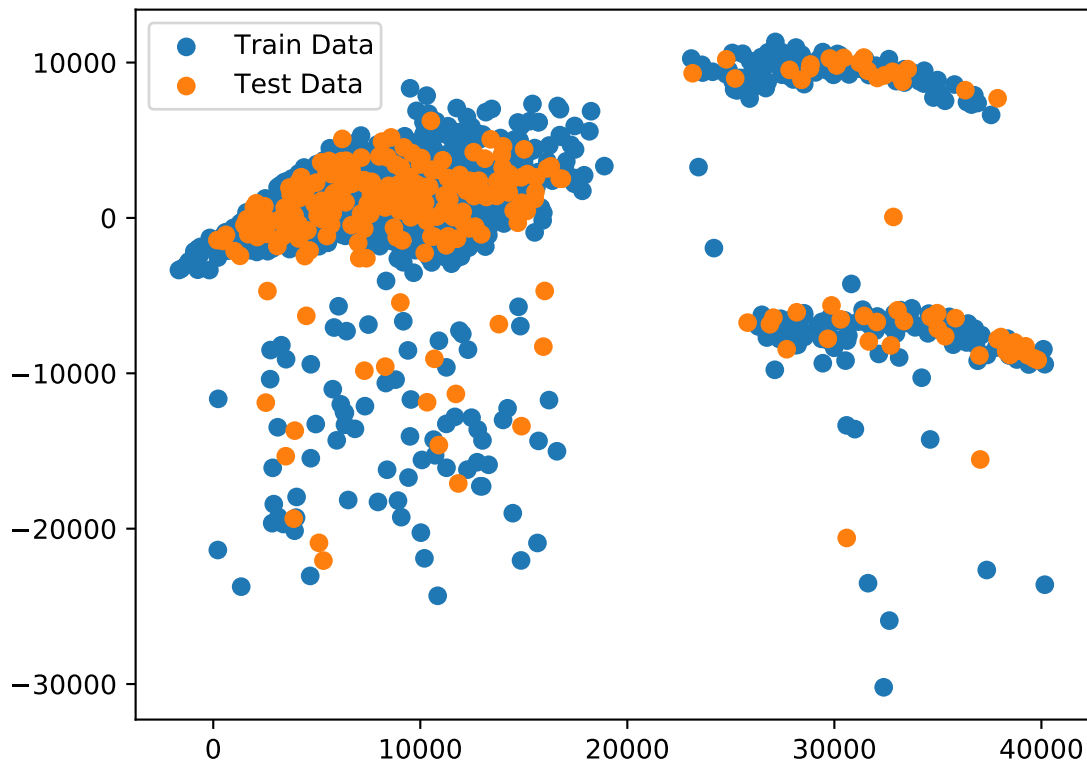
LinearRegression



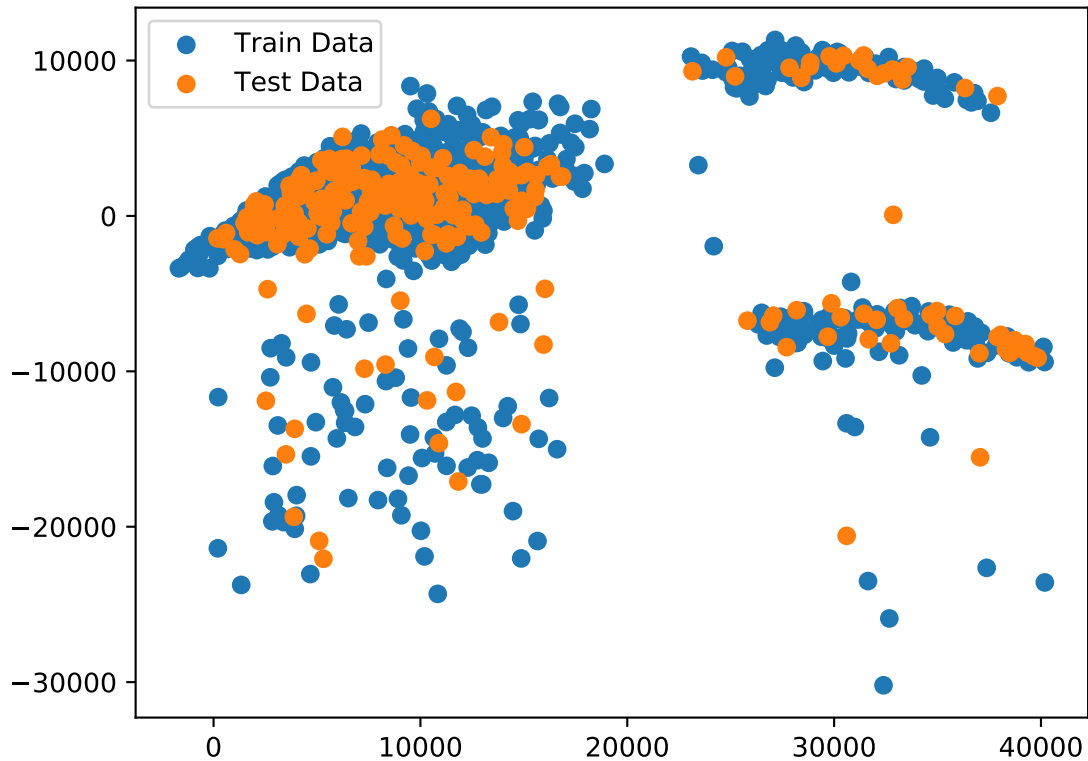
Lasso



ElasticNet



Ridge



RandomForestRegressor

