Summary:

Using Python 3 with Pandas, Matplotlib, and SKLearn, two tiers of charges are discovered in the data. Some data (lower-tiers) are explained *very well* by the factors we have, and other data (higher-tiers) are *hardly* explained by the factors we have. This suggests that there are other factors affecting the charges which we don't have in our data. A missing factor might be the effect of different health plans on their charges.

Analysis

Beginning with some preliminary statistics, the code output gives:

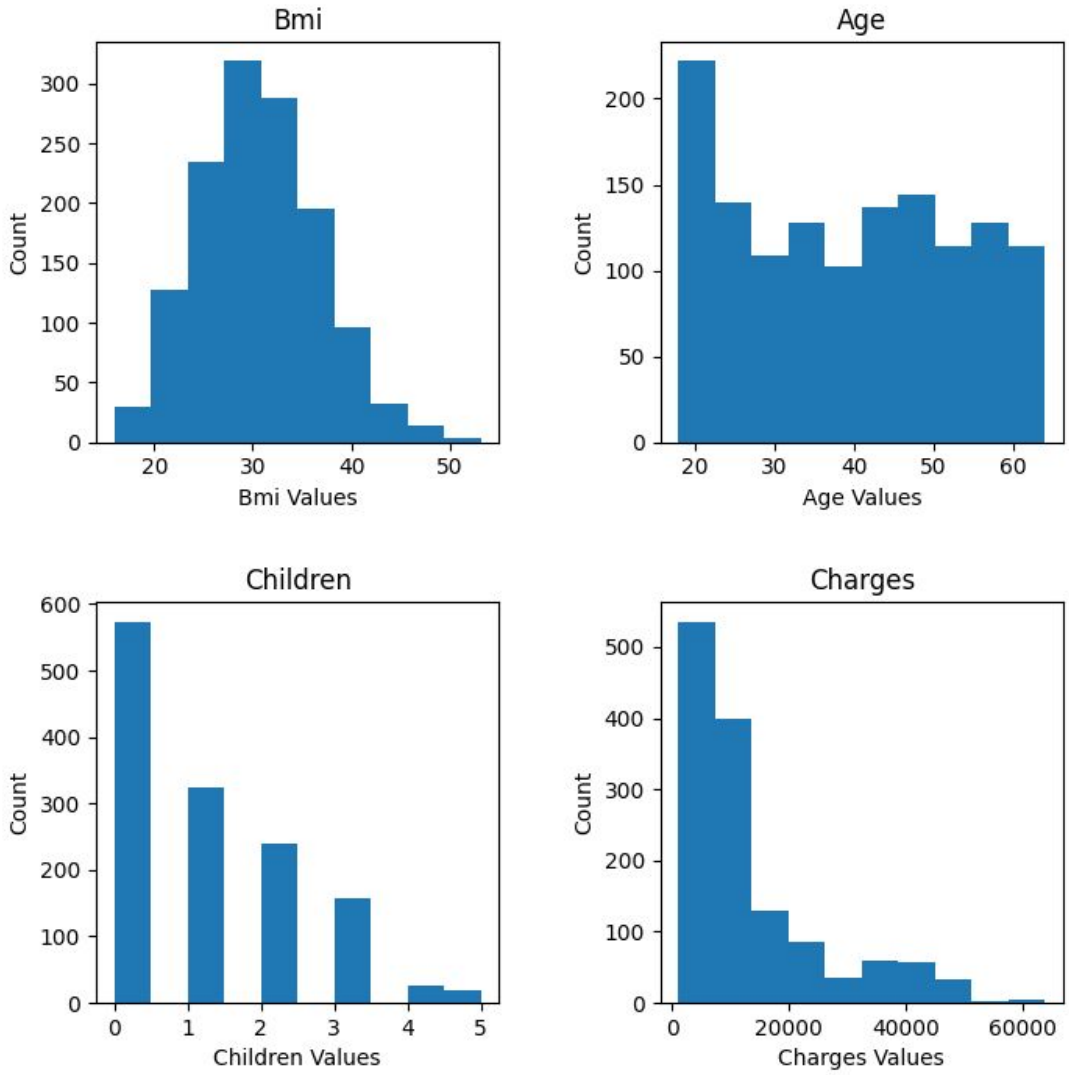
Stats:
bmi--Average: 30.66, Standard Deviation: 6.1
age--Average: 39.21, Standard Deviation: 14.05
children--Average: 1.09, Standard Deviation: 1.21
charges--Average: 13270.42, Standard Deviation: 12110.01

Proportions for sex
female: 49.48%
male: 50.52%

Proportions for smoker
yes: 20.48%
no: 79.52%

Proportions for region
southwest: 24.29%
southeast: 27.2%
northwest: 24.29%
northeast: 24.22%

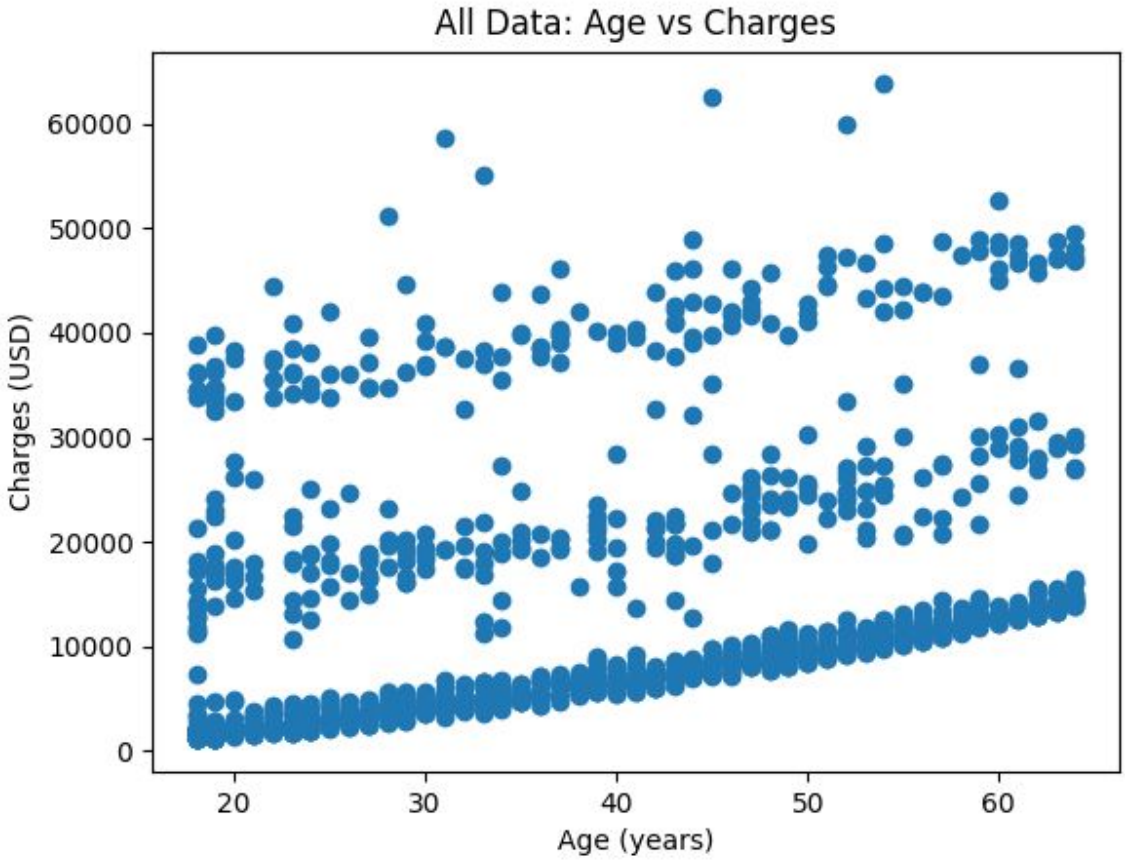
Wow, look at the standard deviation on the charges! It's almost the size of the mean. Let's get a visual of the shape of the data by making histograms:



It turns out the mean and standard deviation aren't all that helpful here because the ‘age,’ ‘children,’ and 'charges' data aren't normal distributions; they’re skewed left. The BMI is a normal distribution.

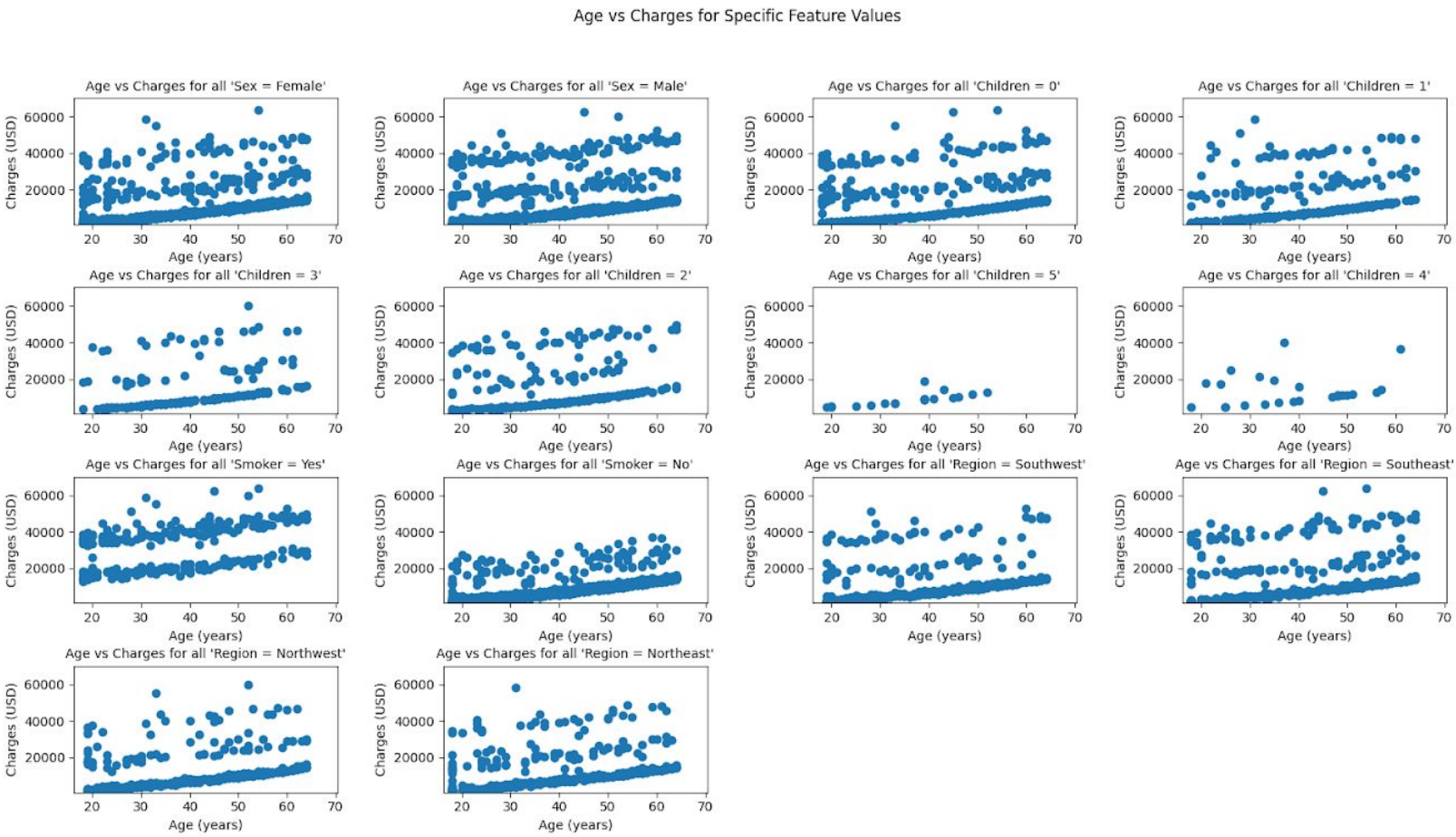
From now on, we’ll use the person’s age as the index, or the primary independent variable representing the person. Age was chosen because each person has one, and it’s a continuous, fairly static, non-arbitrary piece of information. Charges will be our dependent variable.

With that said, let's get an idea of what the data looks like as a whole:



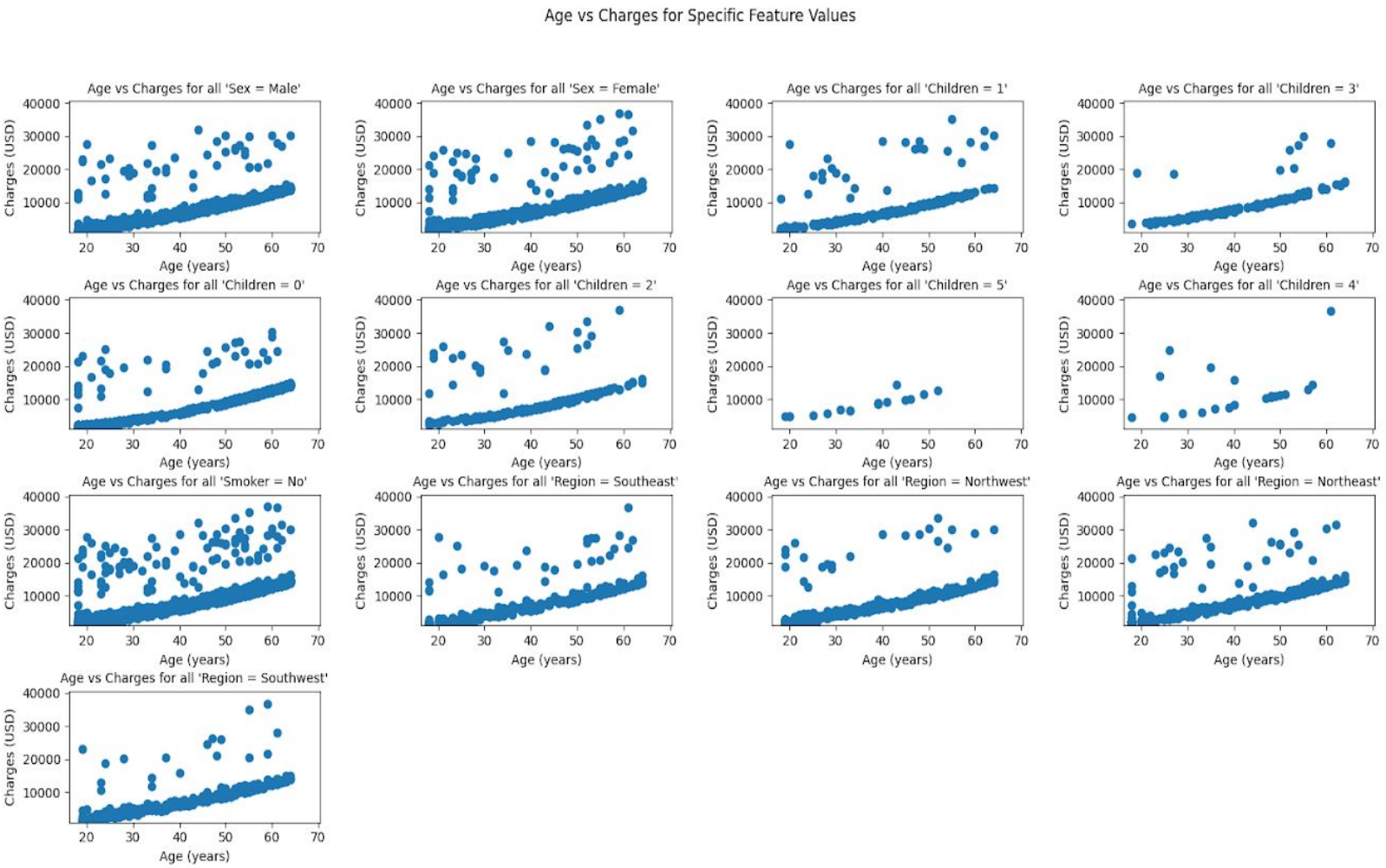
It looks like the charges fall generally into three tiers: High, middle, and low. Why would this be?

If we can look at the data for each unique feature by itself, we might get a picture of which feature(s) are accounting for these tiers. The code yields the following subplots:



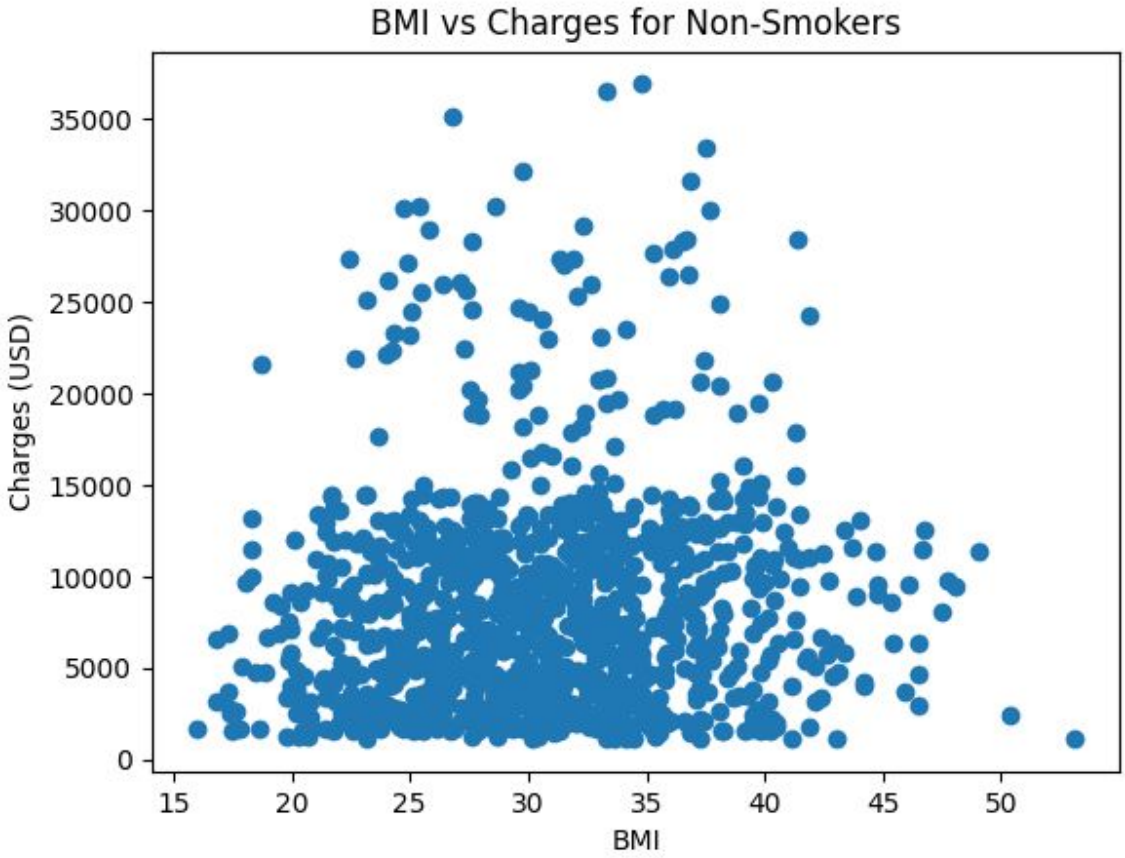
The smokers and non-smokers account for distinct sets of data! It looks like there aren't three tiers for all data, but rather two tiers of data each for smokers and non-smokers, where the lower-tier smoker data falls around the same charges values as the higher-tier non-smoker data. *All of the highest-tier charges are accounted for by smokers*. We were able to separate the smoker and non-smoker data to isolate a tier, but even these still contain tiers! That means there's still something else splitting them.

So let's see if we can isolate the features again, but only among non-smokers, to see if we can further pinpoint which feature accounts for the tiers. Here is the same method applied only to non-smokers:



Notice that the top tier is completely gone (evident visually, but also from comparing the limits of the y-axis)! But there are still two tiers. The low-charges tier looks dense and regular, with most charges (as our histogram showed) falling into a distinct line. Such regularity suggests that the features/independent variables are uniformly weighted in determining the charges; if there were some feature outsizing the rest to fling the charges into the higher tier, then the subplot for that feature should show only the higher-tier data. What we see, however, is that all the rest of the features share in this two-tier split (excluding the people with five children, who couldn't be our answer, since they're so few and *all* in the lower tier anyways). Since *all* of the data still has the split, there must be some other factor behind it all affecting it.

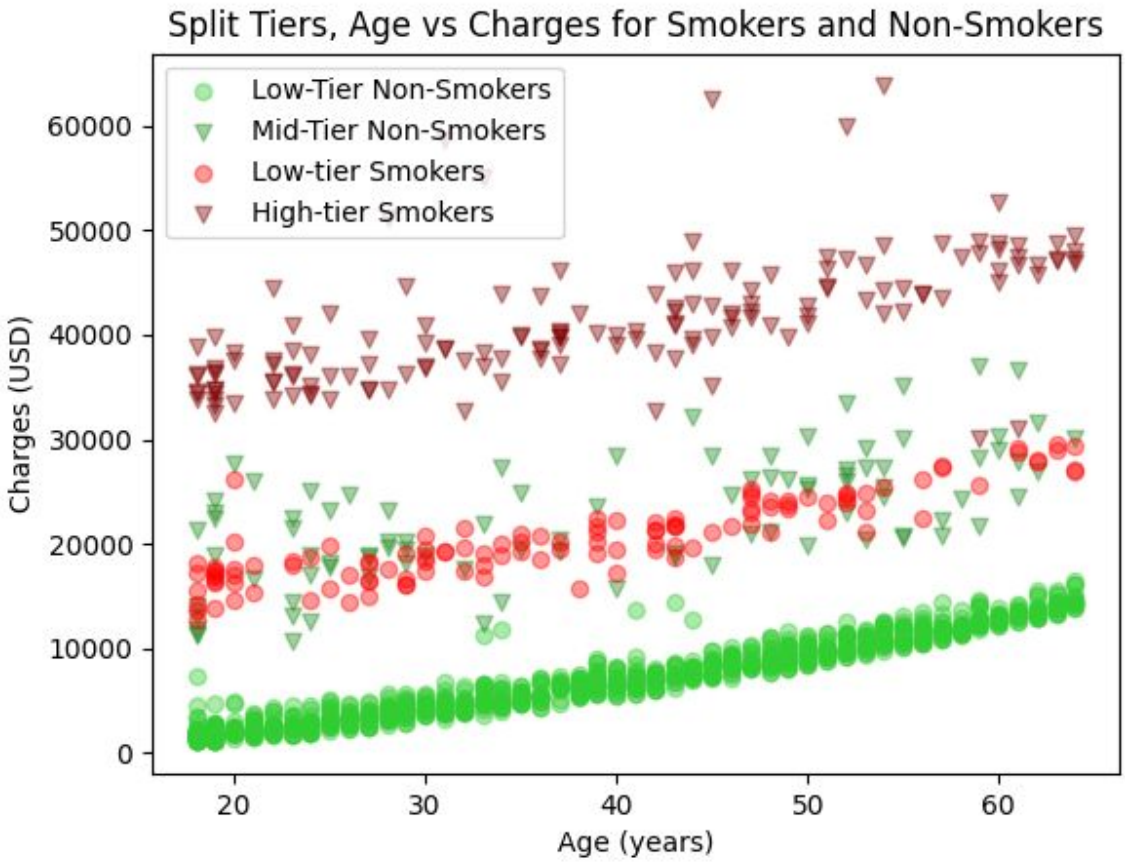
Could the hidden factor be BMI? Here is the BMI plotted against charges for non-smokers:



This data is all over the place! There are people with around 50 BMI paying the same as people with around 17 BMI, and everything in between. Just as with the other features, most BMIs have charges densely gathered under $15k in charges, with more scattered, irregular points up top. It seems the BMI is just as affected by some hidden factor as the rest.

As further evidence of a hidden factor, we could use machine learning to calculate coefficients which can show us the weight of each factor.

Maybe if we split the lower- and higher-tier data and perform multiple linear regression on them, we can compare coefficients to see which factors might account for the split. With (age, charges) being a data point, I used (18, 7500) and (65, 21000) to find a line that separates the tiers. The equation is: y = 287.23x + 2329.79
This will split the higher and lower tiers nicely. We'll include the smokers in this. Here's what our split up data looks like visually:



Now we can use multiple linear regression to analyze each line separately, then see if we get any insights from comparing coefficients. Here's the output:

Multiple Linear Regression Insights for All Data:
R2 value: 0.7599
age coef: 251.768
sex coef: 297.926
bmi coef: 322.375
children coef: 362.199
smoker coef: 24192.21
region_southwest coef: -315.44
region_southeast coef: -406.334
region_northwest coef: 115.005
region_northeast coef: 606.769

Multiple Linear Regression Insights for High-tier Smoker Data:
R2 value: 0.5008
age coef: 269.539
sex coef: 601.309
bmi coef: 611.184

children coef: -216.308
smoker coef: 0.0
region_southwest coef: -216.193
region_southeast coef: -647.111
region_northwest coef: 1123.909
region_northeast coef: -260.605

Multiple Linear Regression Insights for Low-tier Smoker Data:
R2 value: 0.9795
age coef: 246.423
sex coef: 890.034
bmi coef: 452.499
children coef: 284.075
smoker coef: 0.0
region_southwest coef: -74.546
region_southeast coef: -55.267
region_northwest coef: 82.706
region_northeast coef: 47.107

Multiple Linear Regression Insights for High-tier Non-smoker Data:
R2 value: 0.0616
age coef: 294.521
sex coef: 1336.802
bmi coef: -129.896
children coef: 1218.024
smoker coef: -0.0
region_southwest coef: -75.698
region_southeast coef: -1219.444
region_northwest coef: 1496.042
region_northeast coef: -200.901

Multiple Linear Regression Insights for Low-tier Non-smoker Data:
R2 value: 0.9657
age coef: 267.582
sex coef: 448.532
bmi coef: 3.96
children coef: 462.843
smoker coef: -0.0
region_southwest coef: -294.11
region_southeast coef: -219.587
region_northwest coef: 112.628
region_northeast coef: 401.068

Conclusions

The smoker and non-smoker data parallel one another, only the smoker data is shifted upward due to its effect on insurance charges. Although shifted up, both sets are parallel one another in their upper- and lower-tiers. There must be something else splitting them.

With R^2 values greater than 0.9 for the low-tier data, we can say that the features (independent variables) in our data account for 90%+ of the variance in their 'charges' data. This data is strongly linear, and so we expect the good R^2 value for a linear regression model.

The higher-tier data for both smokers and non-smokers, however, is so scattered that we expect a low R^2 value when fitting to a line. These lower R^2 values show that *the features we have in our data don't sufficiently account for the higher-tier charges data*.

So we're left with mystery. What accounts for the higher numbers and irregularity in the higher-tier charges? Why Is the lower tier so much more linear and predictable? Some data (lower-tiers) are explained *very well* by the factors we have, and other data (higher-tiers) are *hardly* explained by the factors we have. This leads me to suppose that there are other factors affecting the charges which we don't have in our data. Perhaps low- and high-deductible health plans?

For future research, I'd want to consult with a health insurance or benefits specialist for some insight into why this split might exist, and see if we could obtain data accordingly.

Concerning the more predictable lower tiers, we can say with certainty that being a smoker is a significant factor in the costs. Our low-tier data with good linear models show that sex is a fairly significant factor in determining charges. Age is consistent between the two low-tier models, but sex and BMI are much more significant factors for the smokers. Region is a more significant factor for non-smokers.