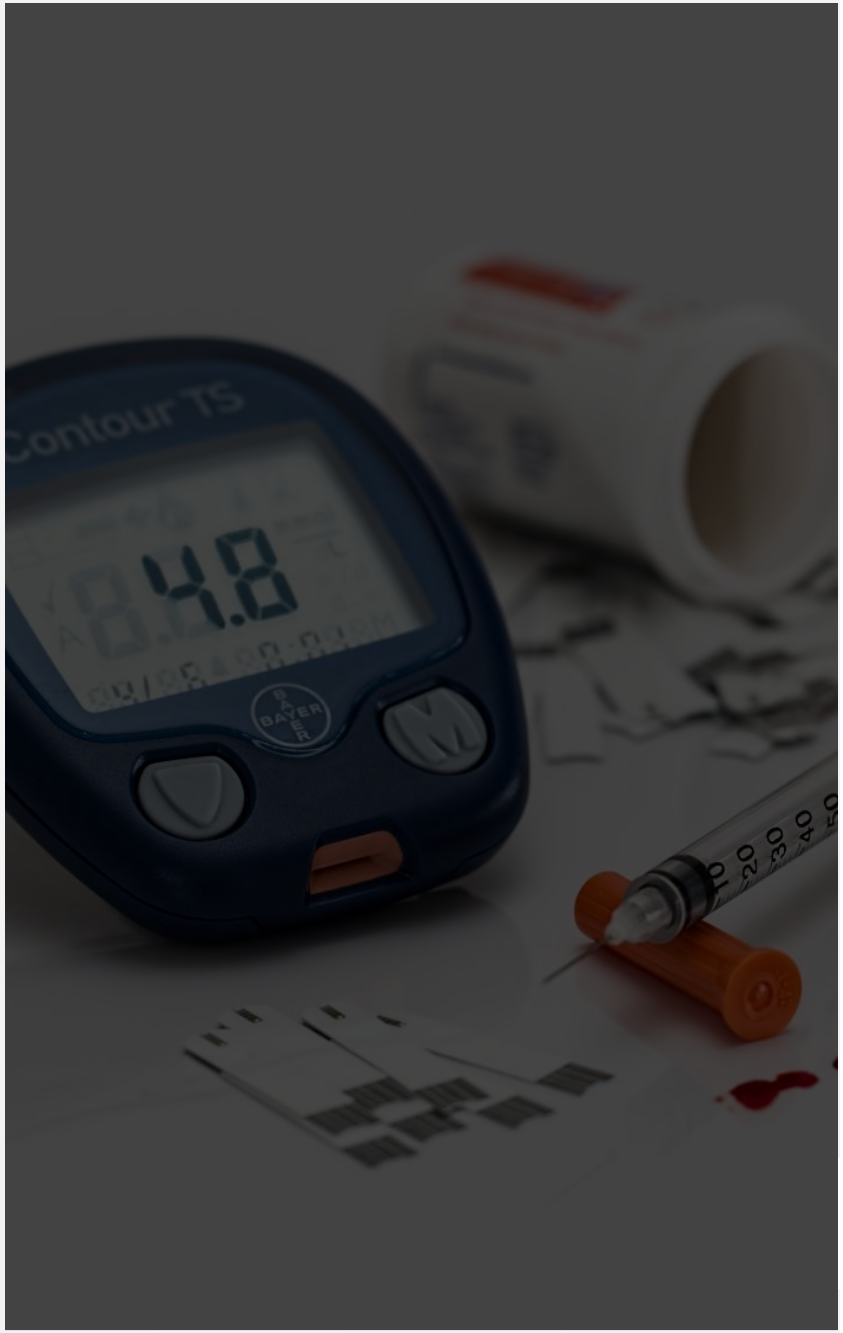
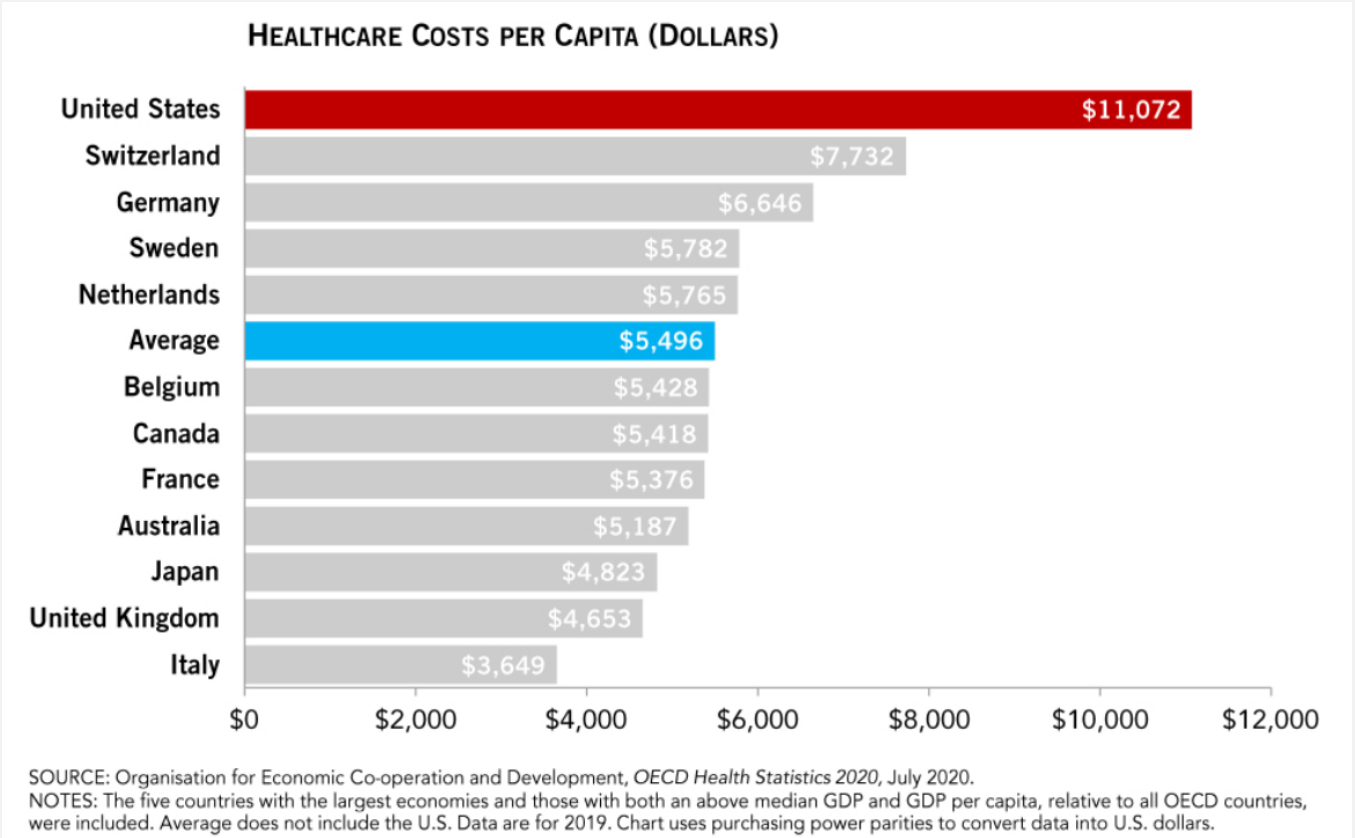
Prediction of Medical Charges for U.S. Citizens

By Eric Wheeler

The U.S. has the highest healthcare costs per capita of all the OECD nations and spends roughly twice as much as the average amongst other wealthy countries. The goal is to uncover insights into which factors contribute most to these high medical expenses.





Can we predict U.S. medical charges based on these 6 variables?

Smoking

- Yes
- No

Gender

- Male
- Female

Age

- 18-25
- 26-33
- 34-41
- 42-49
- 50-57
- 58-65

Children

- 0 - 5

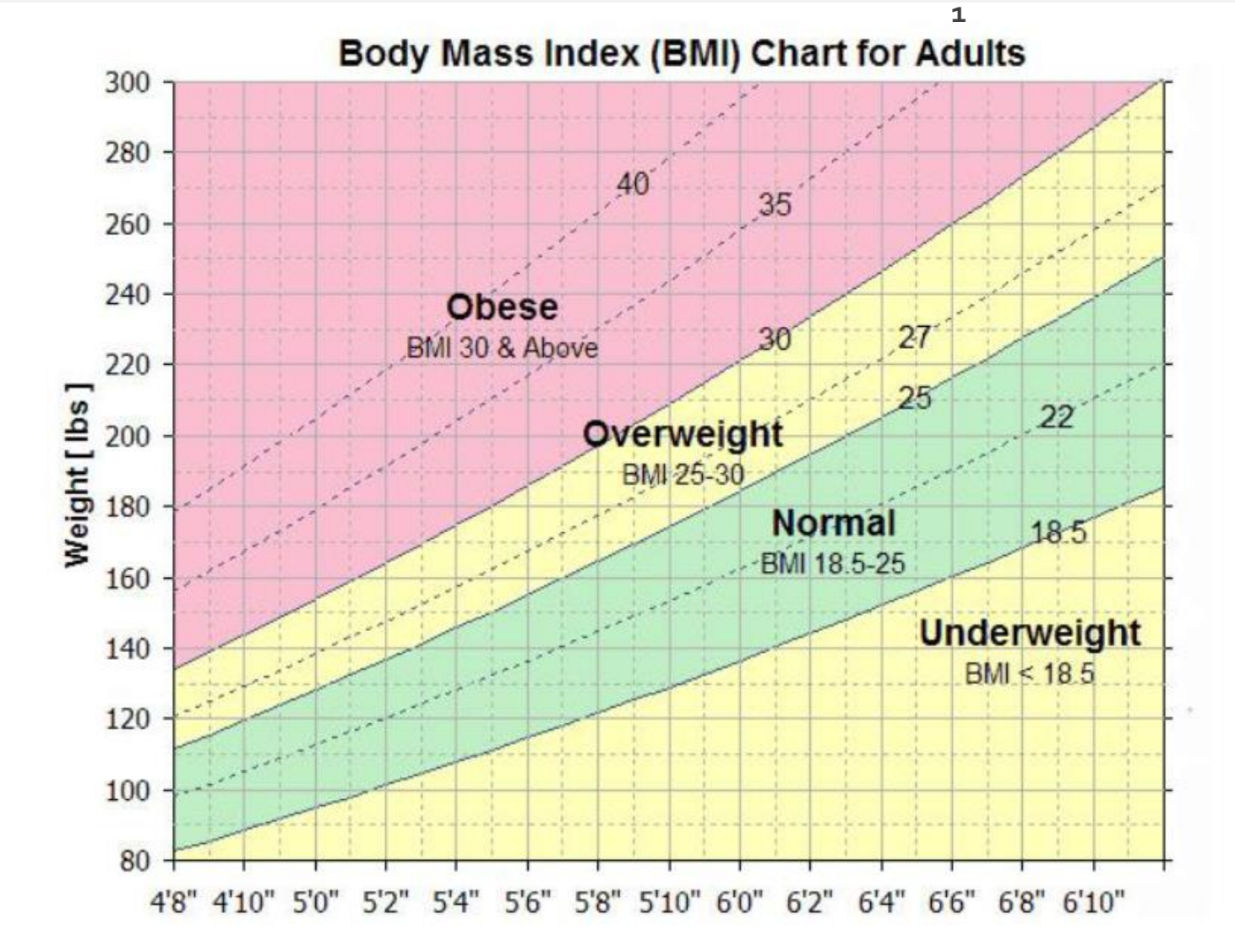
Region

- Northeast
- Southeast
- Northwest
- Southwest

BMI

- Underweight
- Normal
- Overweight
- Obese

Determining Weight Status



Underweight – BMI < 18.5

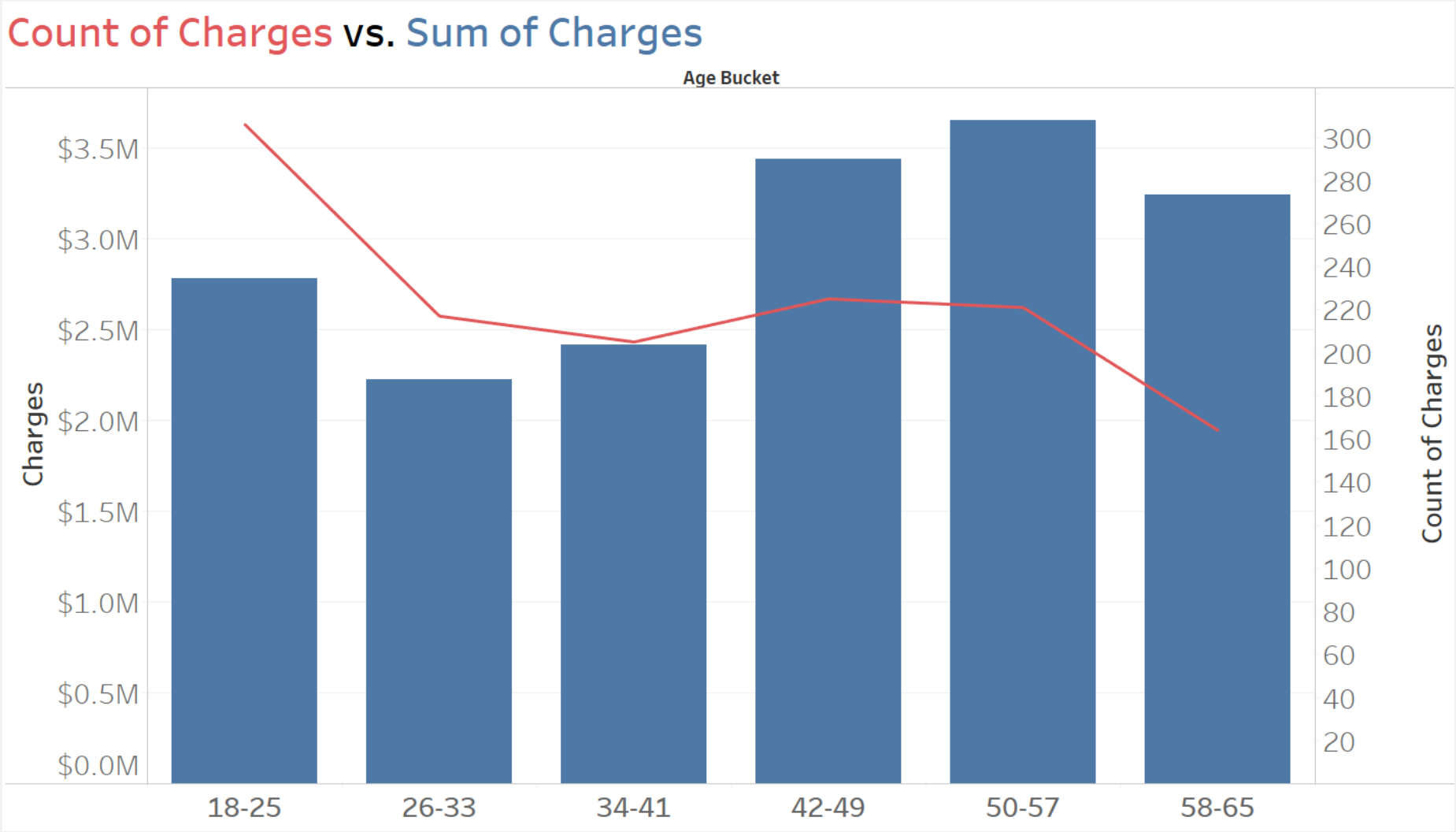
Normal – BMI 18.5-25

Overweight – BMI 25-30

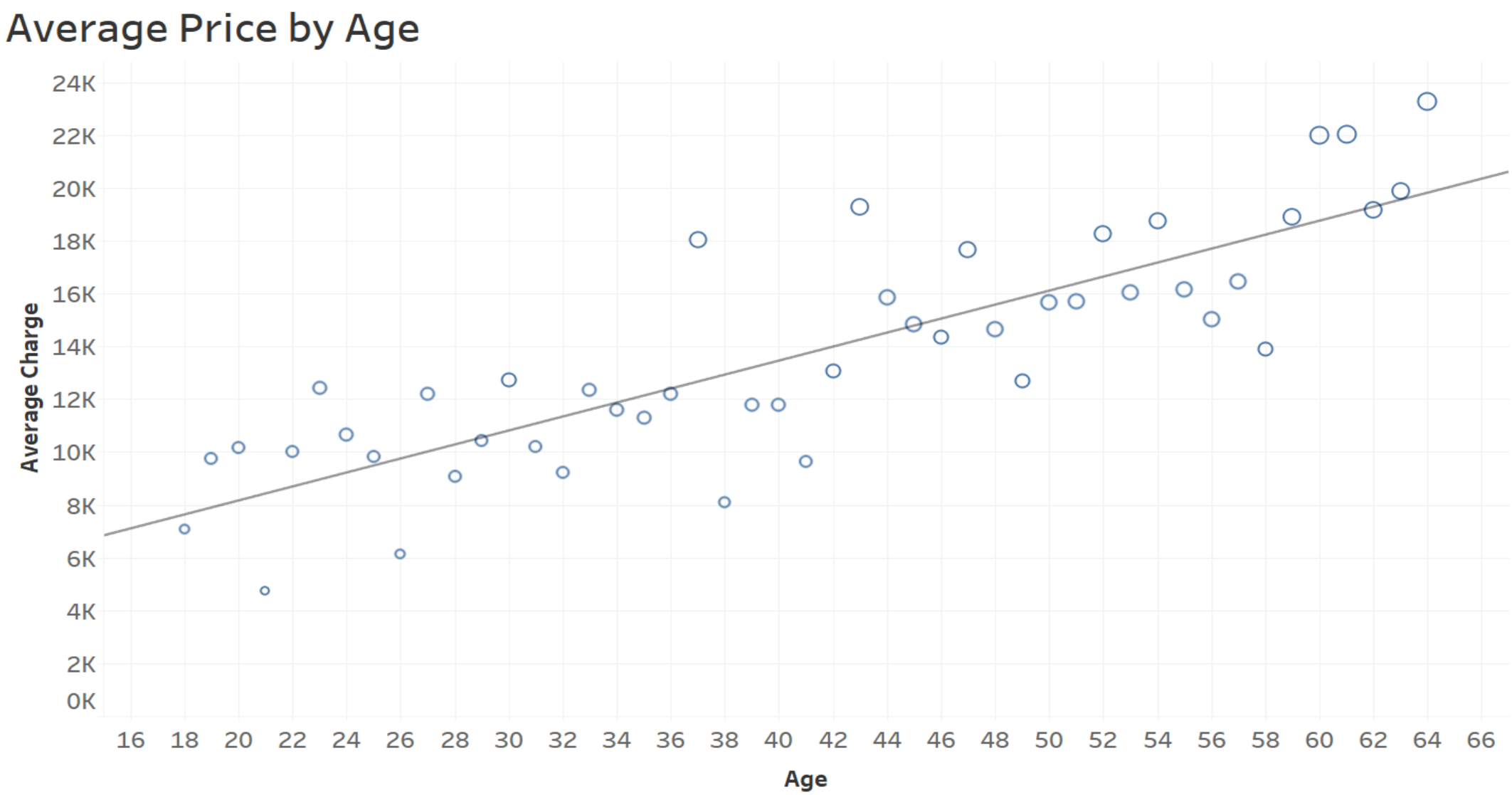
Obese – BMI > 30

[1] Jon Wittwer. BMI Chart (Body Mass Index) (2009). https://www.vertex42.com/ExcelTemplates/bmi-chart.html

Age distribution skews towards the youngest age group of 18-25-year-olds. Despite this, the older age groups of 42-49, 50-57, and 58-65 have higher overall charges.

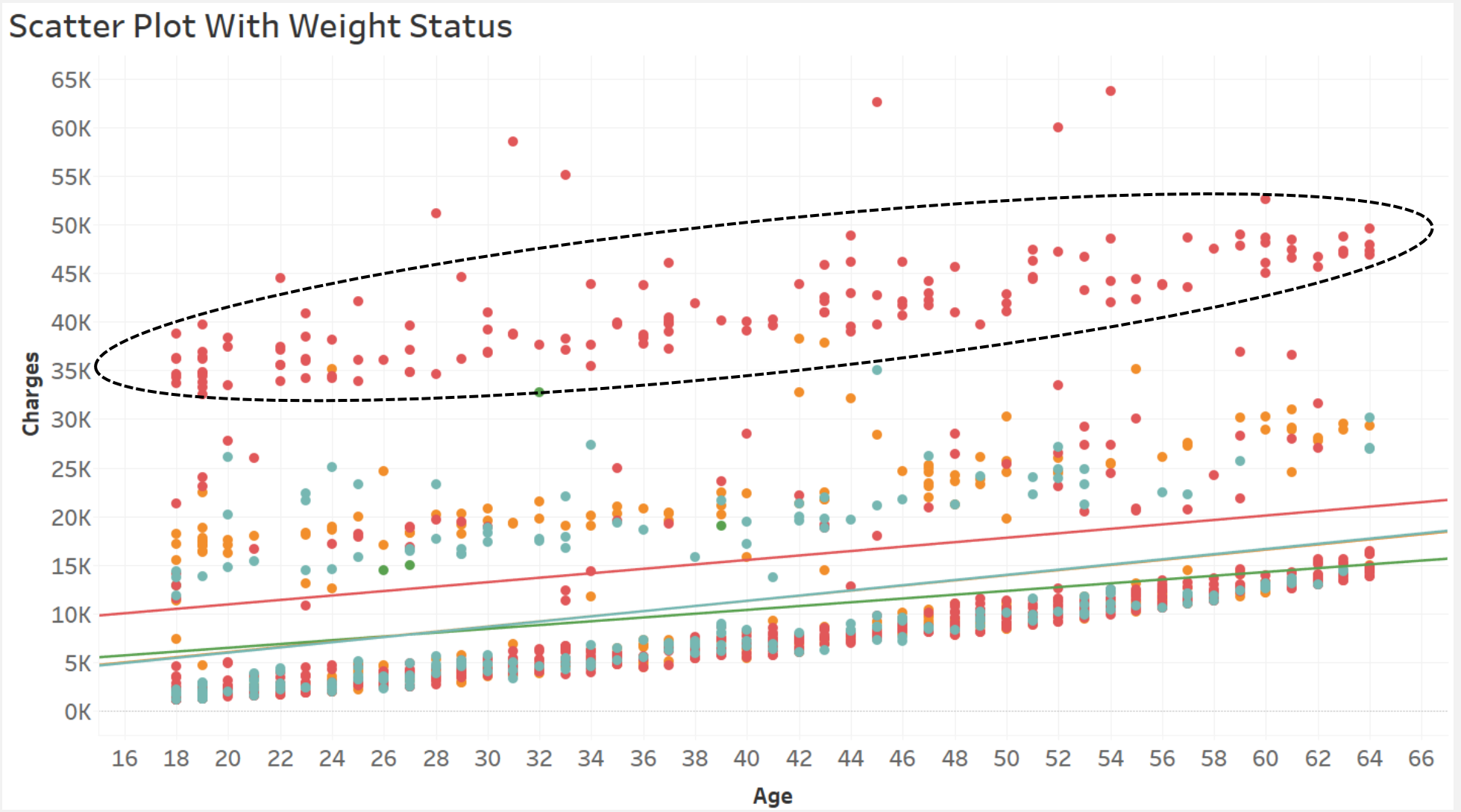


The average charge per age increases as people get older.

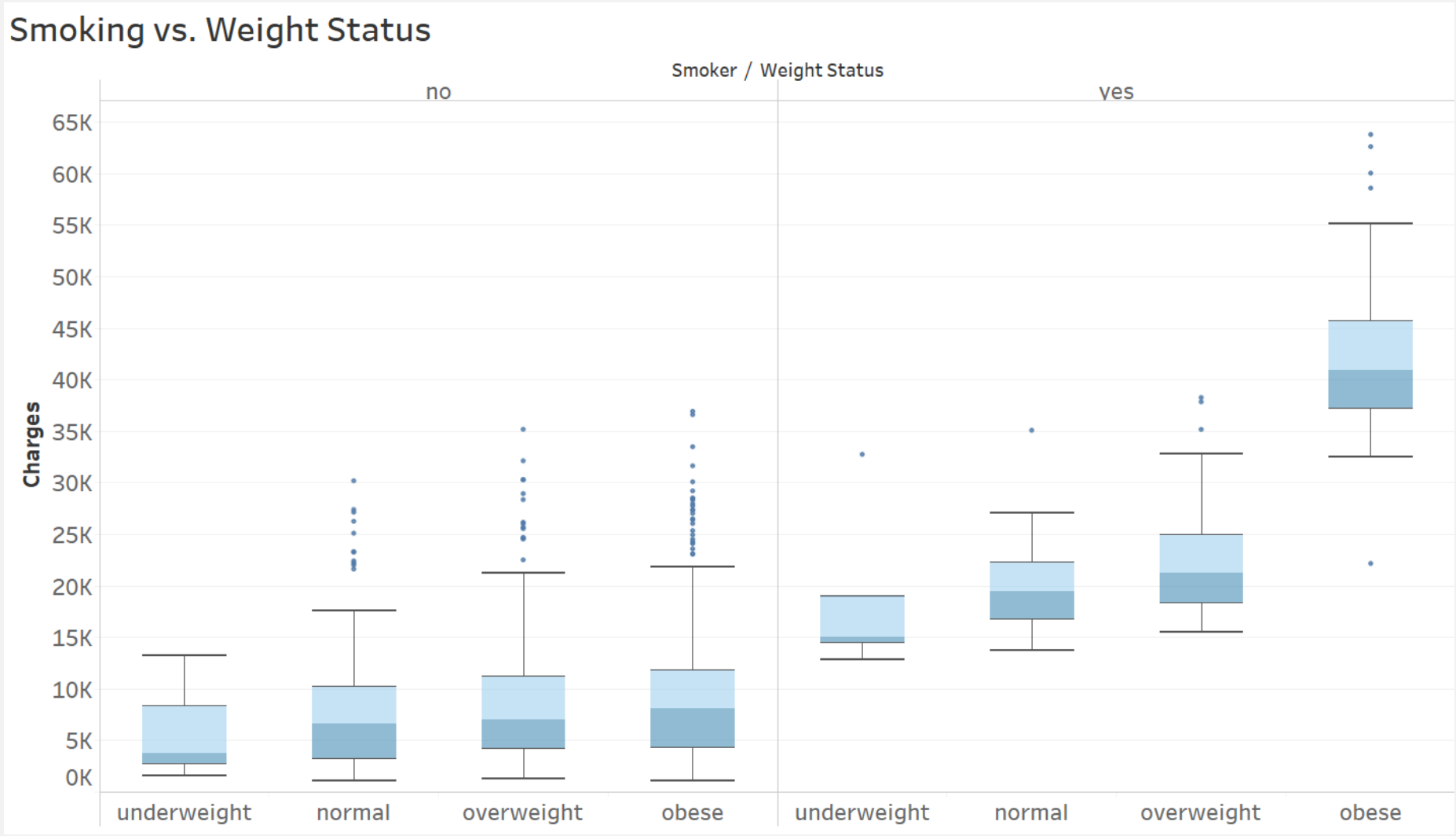


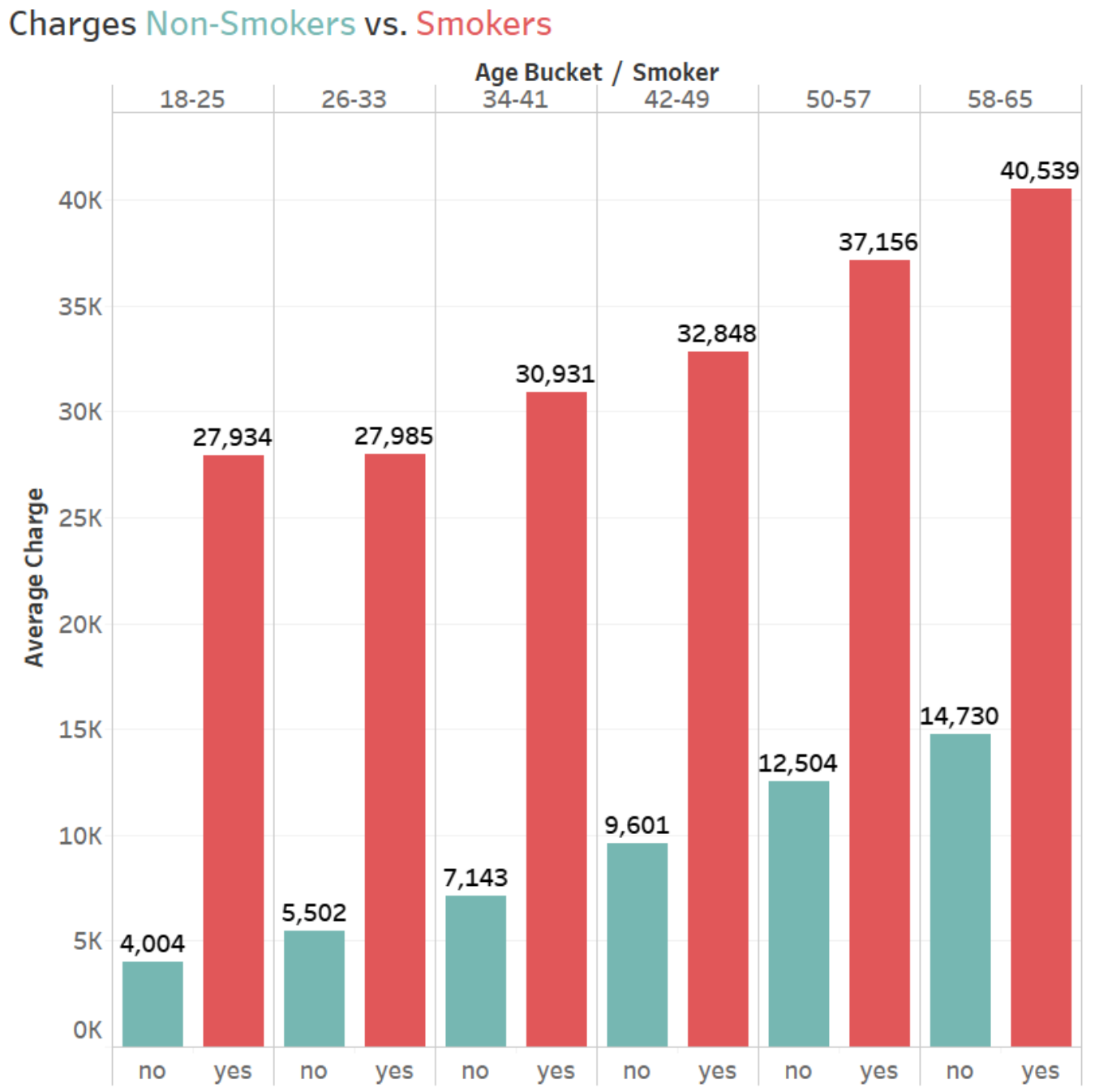
Disaggregating the data shows that there are data points that have particularly high charges. When grouping the data by smokers and non-smokers it is revealed that the highest charges are attributed to smokers.

Grouping the data by weight status reveals that these same large charges belong almost exclusively to people who fall under the obese category.

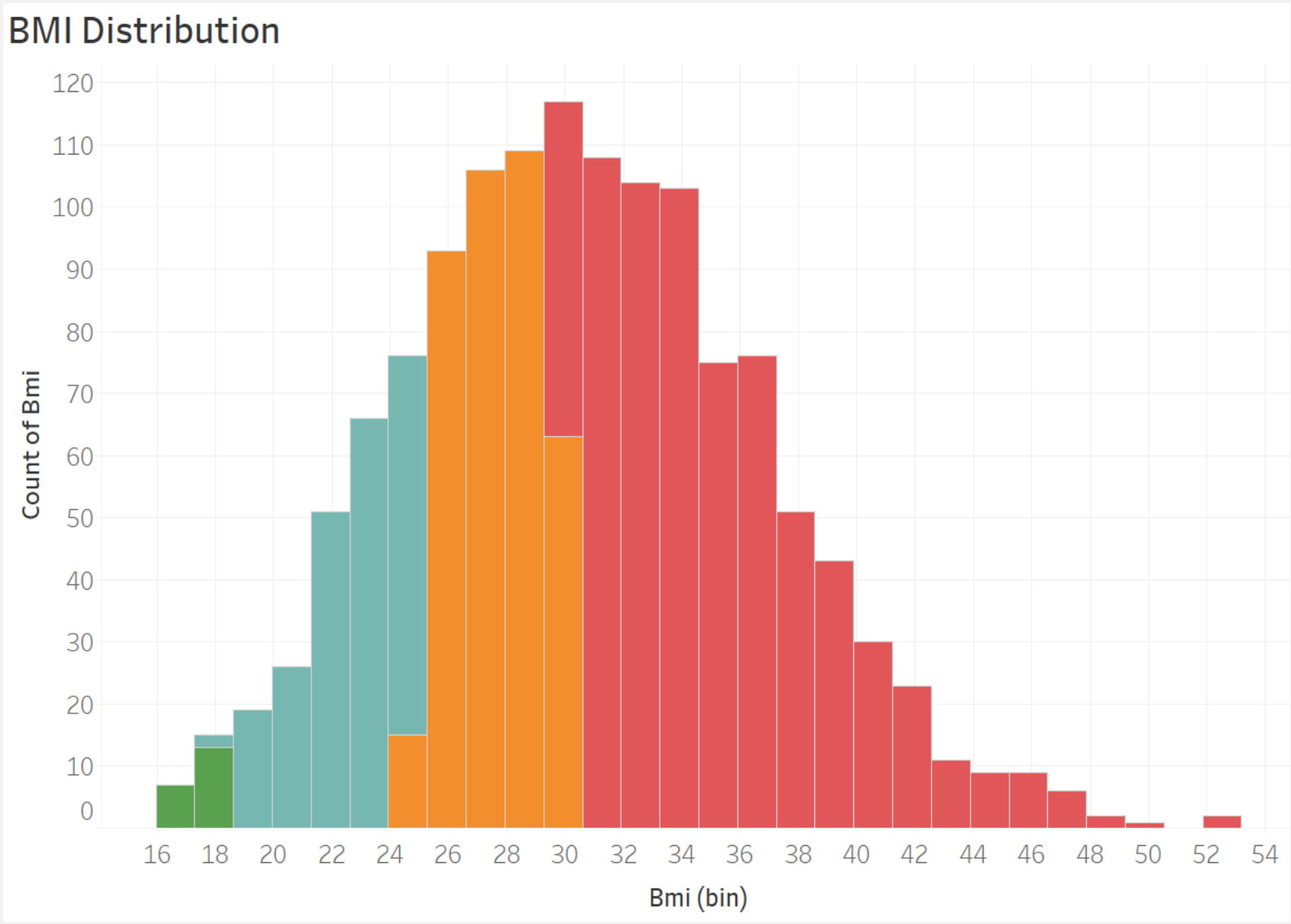


Smoking and Obesity together have a much larger affect on charges than either does alone.



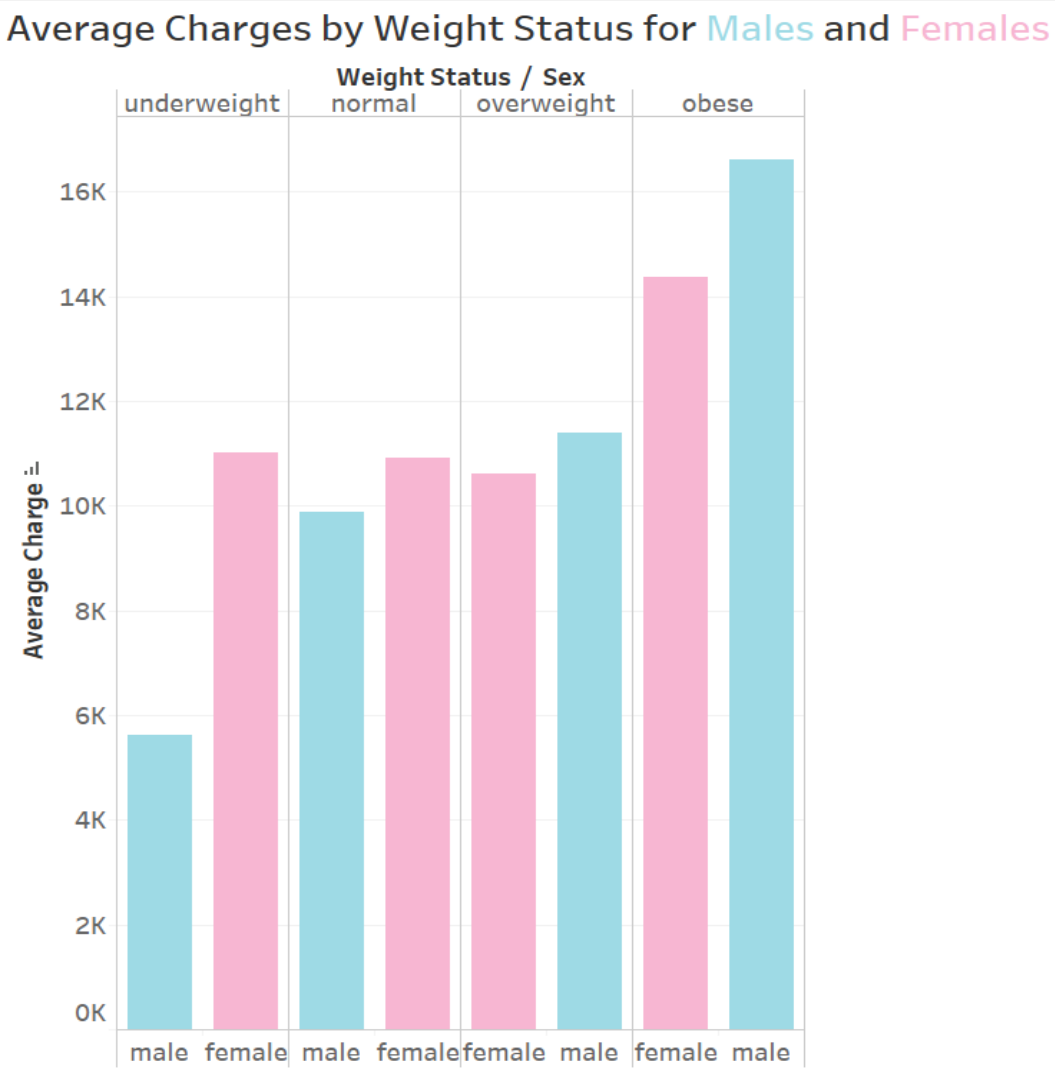


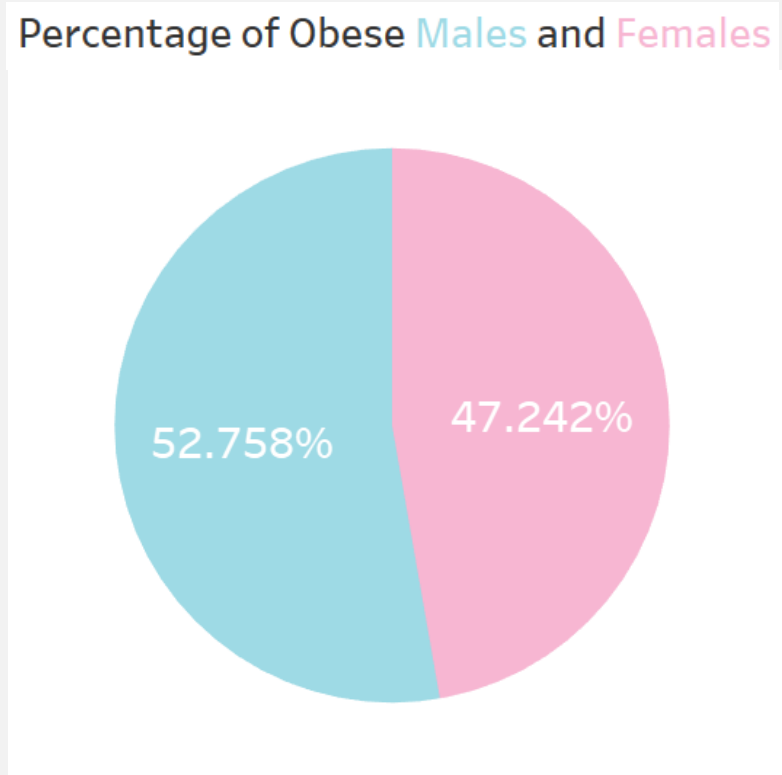
When taken alone, smokers experience average medical charges that are up to 7x more than their non-smoking peers!



The distribution of BMI shows that median BMI of people in the sample is considered obese and more than half of the samples in the study fall under the obese category.

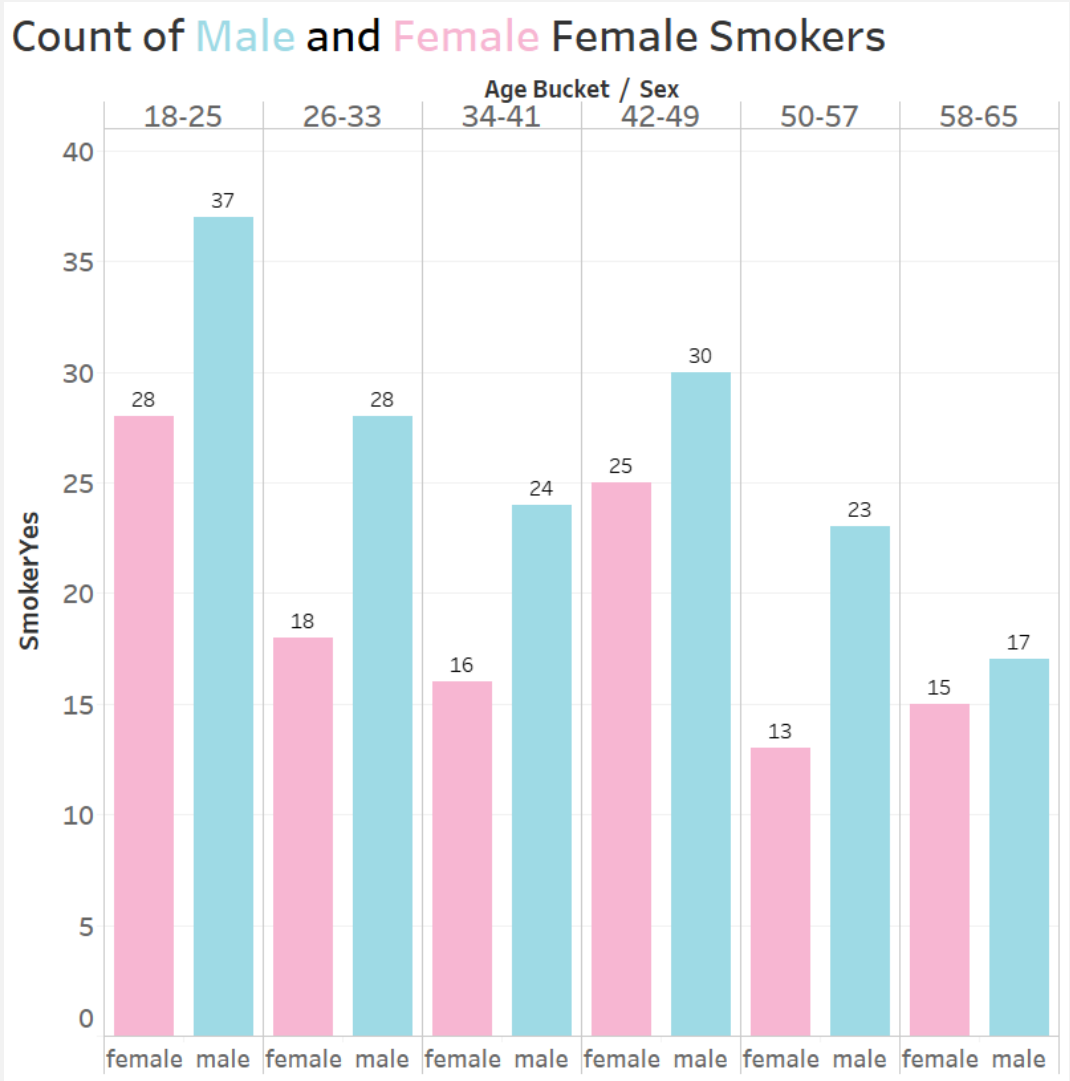
Women have steady medical charges regardless of weight until they hit a BMI that is obese. Men on the other hand experience an increase in medical charge as they move up in weight status. Ultimately obese men end up spending more on average when compared to obese women.

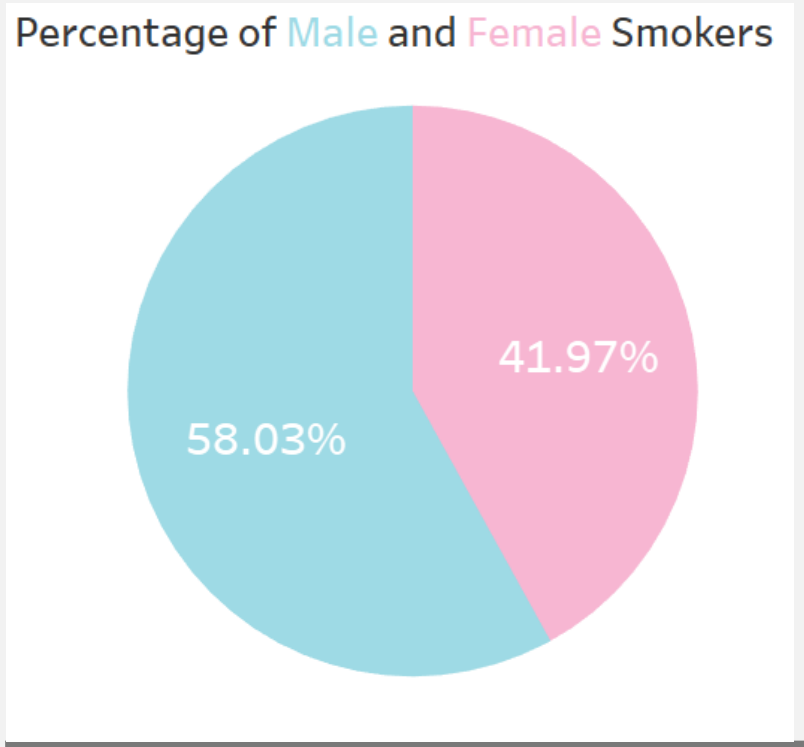




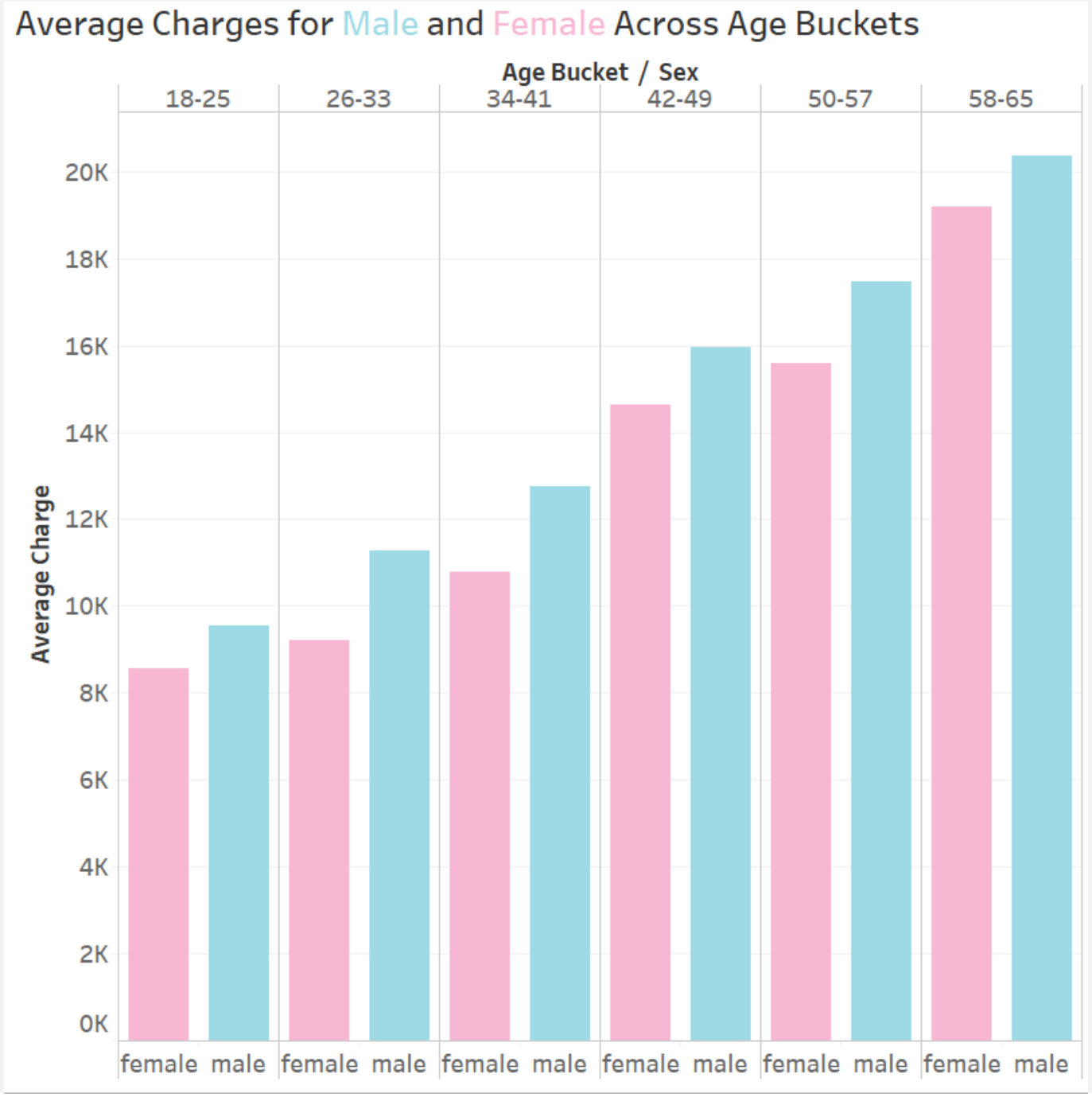
Men account for 53% of obese cases while women account for 47%.

Men are more likely to smoke than women regardless of age group.

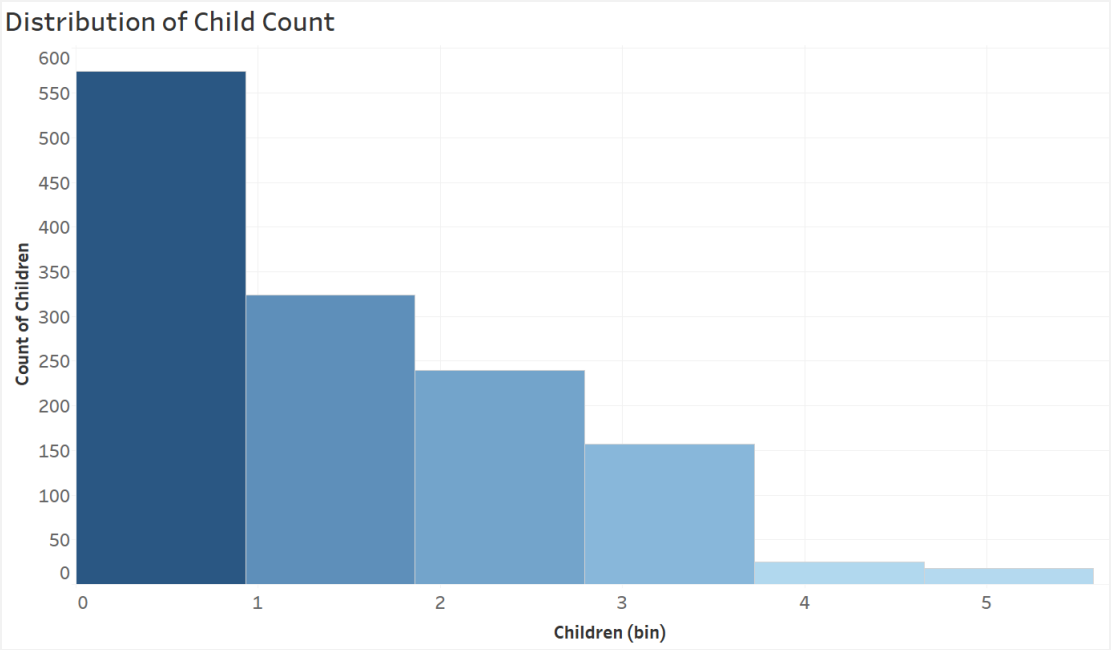




Men account for 58% of all smokers while women only account for 42%.

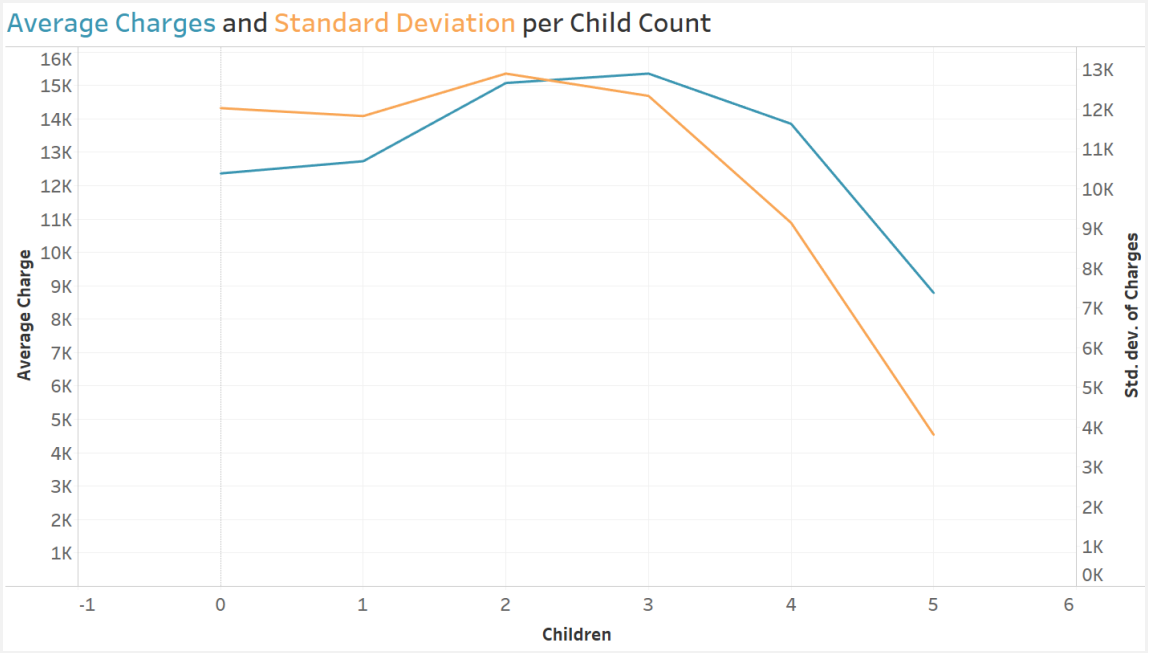


It appears that the higher rate of obesity and smoking amongst men could contribute to the higher average charges they experience across all age groups.

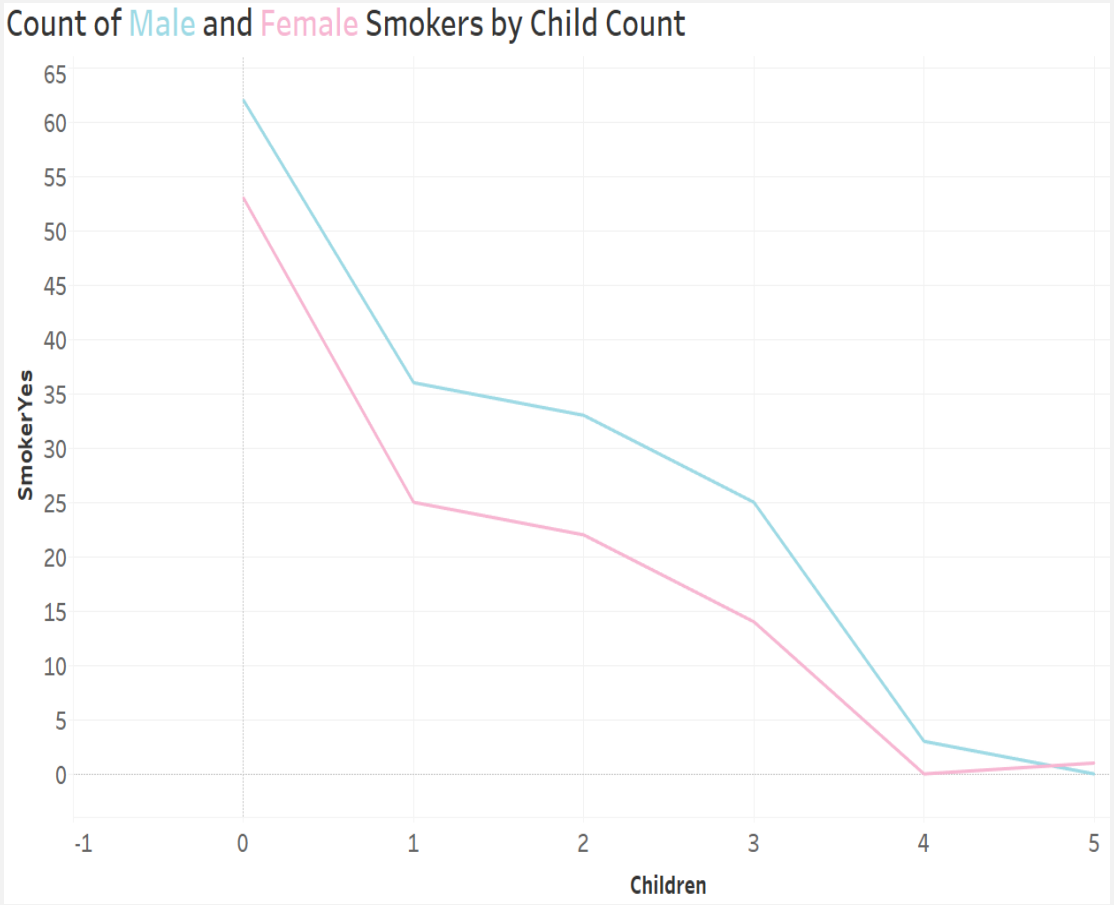


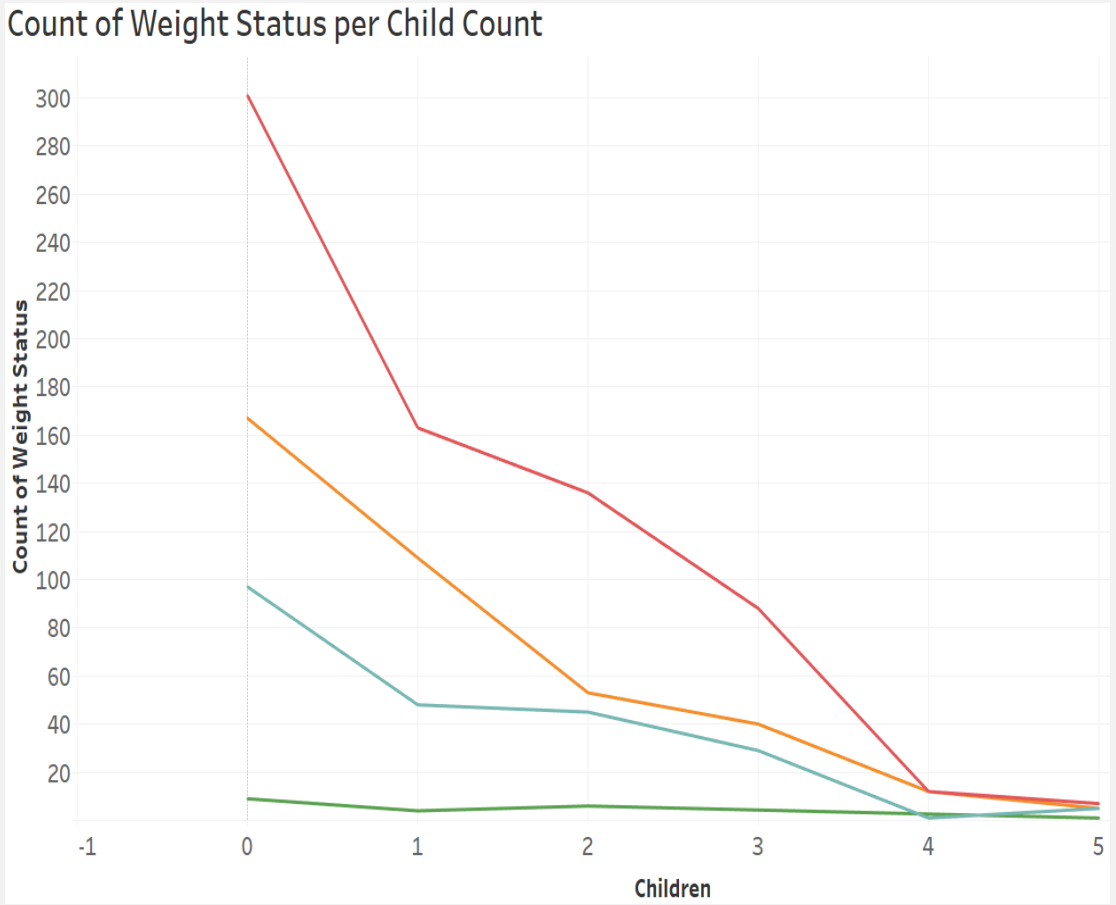
The distribution of child count is skewed heavily towards fewer children with zero children accounting for most of the samples.

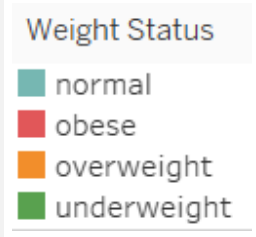
Average charges and standard deviation follow a similar path. Charges seem to increase with child count until reaching a count of 3. Charges and standard deviation begin to drop when child count reaches 4 and 5.



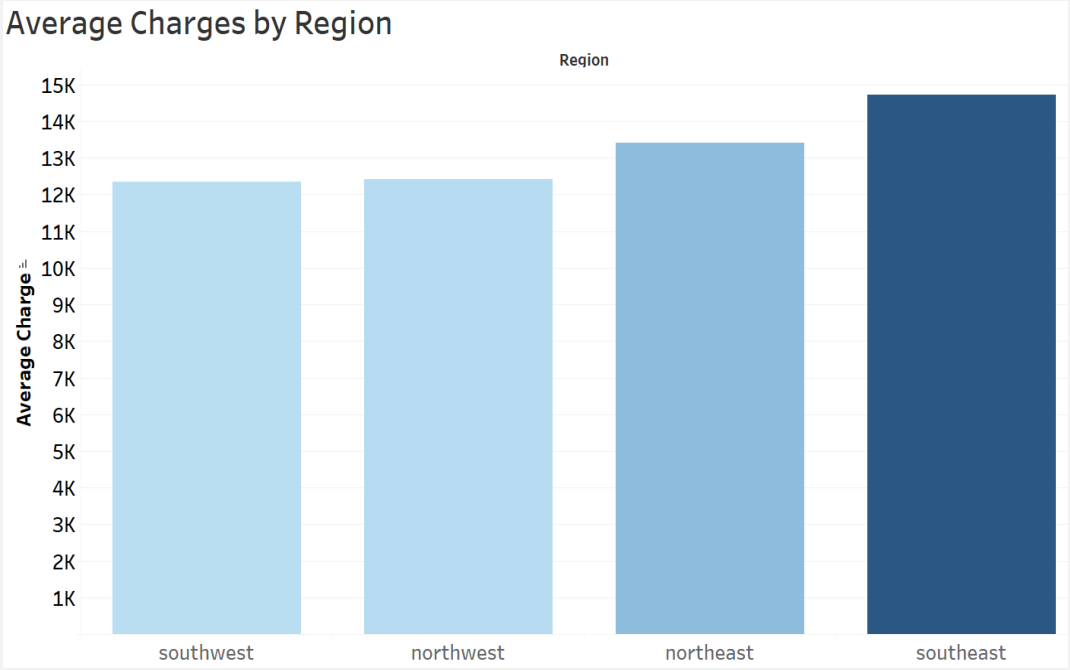
Exploration of the smoker and weight status variables in relation to child count reveals that there are far fewer cases of smoking and obesity amongst individuals who have four or five children.

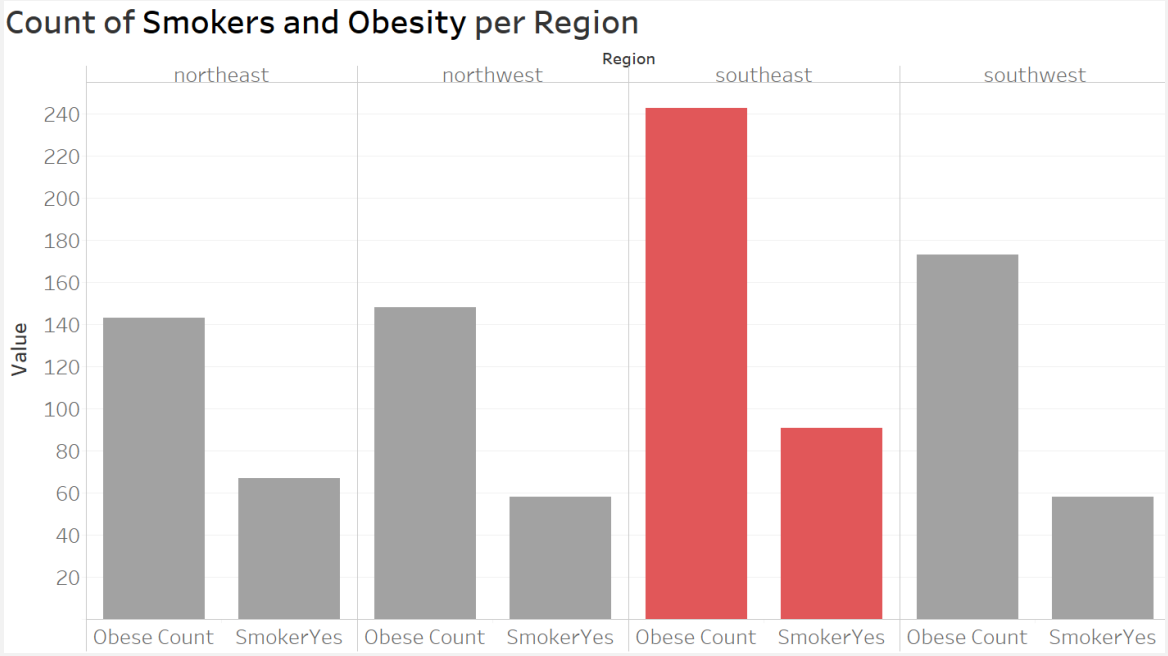






Region did not show to have a significant affect on charges. However, it is worth noting that the region with the highest average charges also has the highest count of obesity and smoking of all the areas in the study.





Correlation Matrix for the Independent Variables

The correlation matrix shows that smoking, age, and BMI are the variables most strongly correlated with medical charges.

Correlation	
Smoker_Yes	0.787251
Age	0.299008
BMI	0.198341
Region_SE	0.073982
Children	0.067998
Sex_Male	0.057292
Region_NE	0.006349
Region_SW	-0.04321

	Age	*Sex_Male*	*BMI*	*Children*	*Smoker_Yes*	*Region_SE*	*Region_SW*	*Region_NE*	*Charges*
Age	1								
Sex_Male	-0.02086	1							
BMI	0.109272	0.046371	1						
Children	0.042469	0.017163	0.012758901	1					
Smoker_Yes	-0.02502	0.076185	0.003750426	0.007673	1				
Region_SE	-0.01164	0.017117	0.270024649	-0.02307	0.06849841	1			
Region_SW	0.010016	-0.00418	-0.00620518	0.021914	-0.0369455	-0.346265	1		
Region_NE	0.002475	-0.00243	-0.13815622	-0.02281	0.00281113	-0.345561	-0.320177	1	
Charges	0.299008	0.057292	0.198340969	0.067998	0.78725143	0.0739816	-0.04321	0.0063488	1

Regression Using All Independent Variables

Absolute Percent Error (APE[1]) Using All Variables

Observation	Predicted Charges	Residuals	Actual charges	APE
1	$25,661.86	-$8,776.93	$16,884.92	51.98%
2	$3,818.78	-$2,093.23	$1,725.55	121.31%
3	$7,096.73	-$2,647.27	$4,449.46	59.50%
4	$3,643.43	$18,341.04	$21,984.47	83.43%
5	$5,376.30	-$1,509.44	$3,866.86	39.04%
6	$4,234.98	-$478.36	$3,756.62	12.73%
7	$11,057.62	-$2,817.03	$8,240.59	34.18%
8	$7,849.35	-$567.84	$7,281.51	7.80%
9	$7,920.04	-$1,513.63	$6,406.41	23.63%
10	$11,741.53	$17,181.61	$28,923.14	59.40%
11	$2,714.66	$6.66	$2,721.32	0.24%
12	$36,225.45	-$8,416.73	$27,808.73	30.27%
13	$4,836.13	-$3,009.29	$1,826.84	164.73%
14	$15,217.24	-$4,126.52	$11,090.72	37.21%
15	$32,182.34	$7,429.42	$39,611.76	18.76%
16	$1,120.43	$716.80	$1,837.24	39.02%
17	$11,746.54	-$949.20	$10,797.34	8.79%
18	$1,433.58	$961.59	$2,395.17	40.15%
19	$15,243.34	-$4,640.95	$10,602.39	43.77%
20	$30,753.80	$6,083.67	$36,837.47	16.51%
...				

Performing regression with all the independent values determines that the significant variables are age, BMI, children, and smoking.

P-Value	
Age	7.78E-89
BMI	6.5E-31
Children	0.000577
Smoker_Yes	0
Region_SE	0.154669
Region_SW	0.203533
Region_NE	0.458769
Sex_Male	0.693348

MAPE[2]
40.28%

With a MAPE of 40% some data manipulation was necessary to lower the chance of error.

[1] The Absolute Percent Error measures the size of the error of a predicted charge in relation to the actual charge.

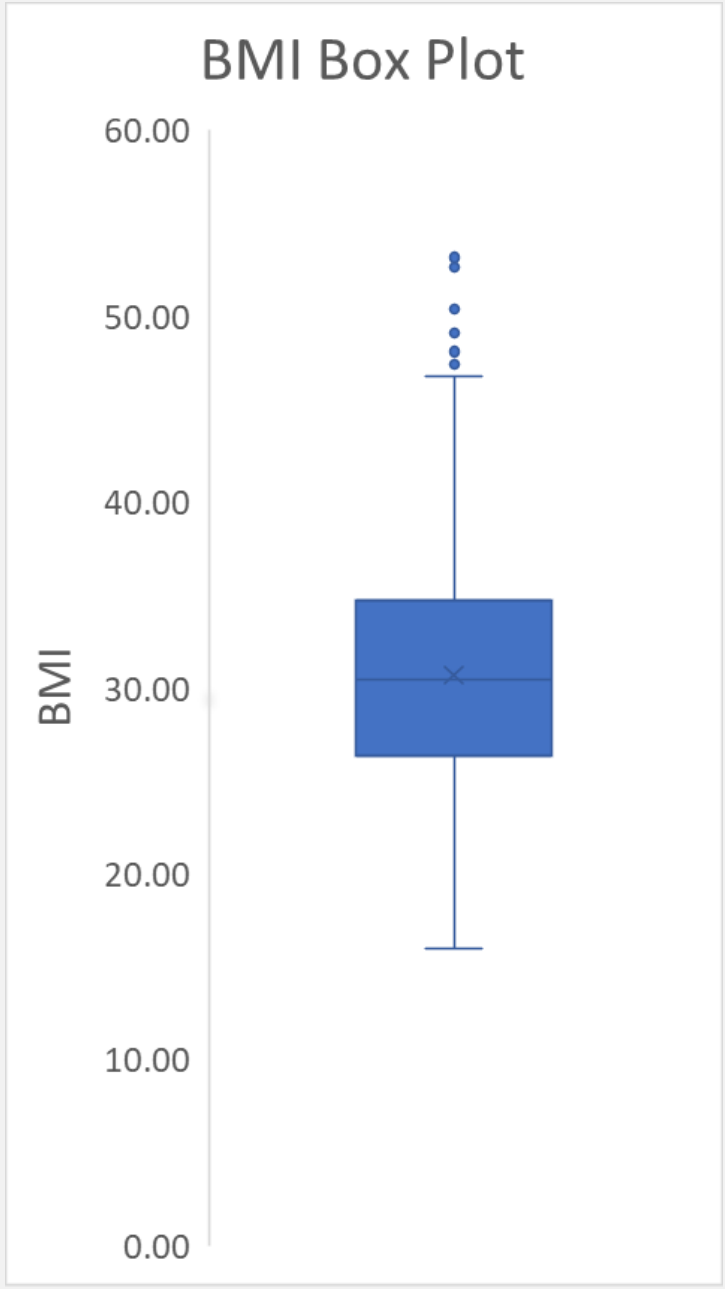
[2] The Mean Absolute Percent Error calculates the average of all the APE values to determine the prediction accuracy of a forecasting model

Adjusting the BMI to Remove Outliers

Outliers in the BMI variable were identified and removed to limit the affect of these exceptionally high figures.

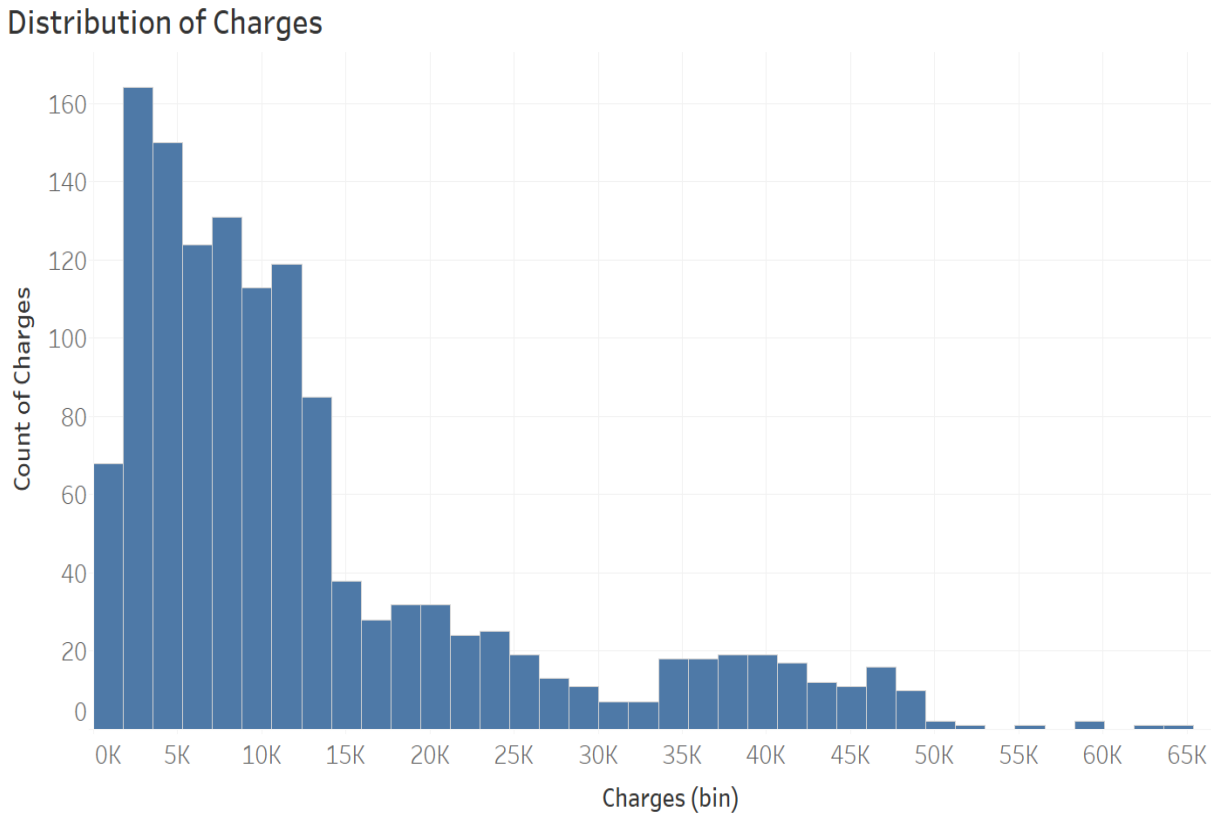
BMI Average	Quartile 1	Quartile 3	IQR	Upper Bound	Lower Bound
30.66	26.30	34.69	8.40	47.29	22.10

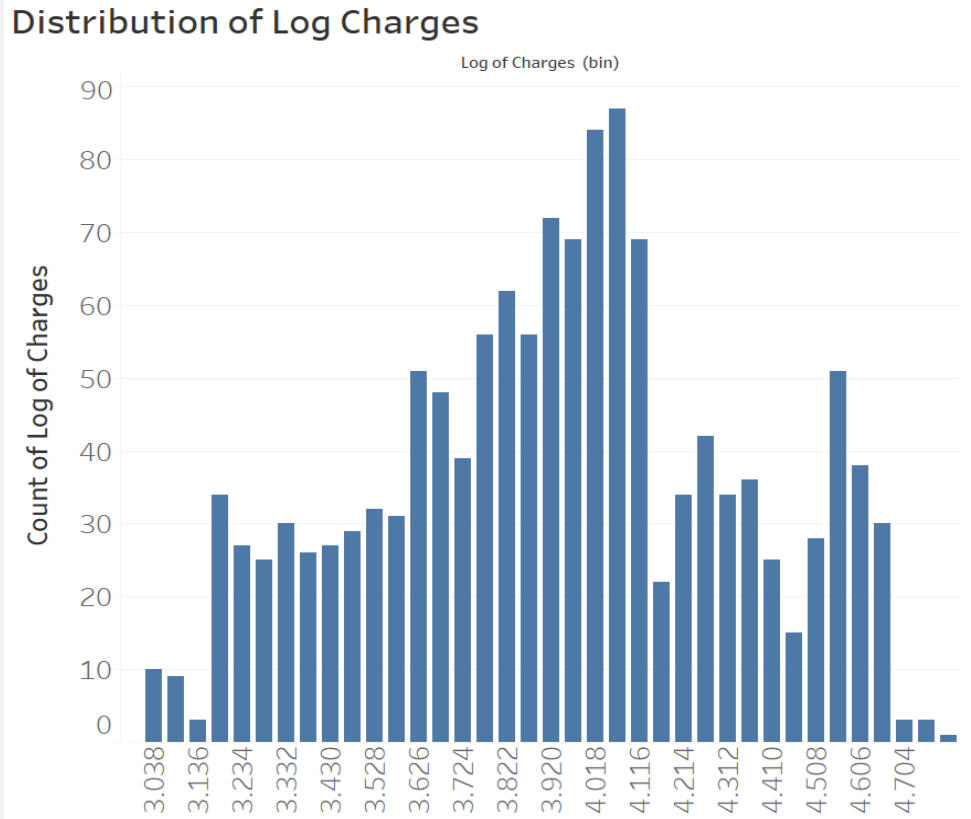
Upper Outliers	Lower Outliers
9	0



Taking the Log of Charges

The distribution of charges is right skewed in the dataset. In order to adjust this to a more normal distribution I took the logarithm of charges.





Regression Using Cleaned Data

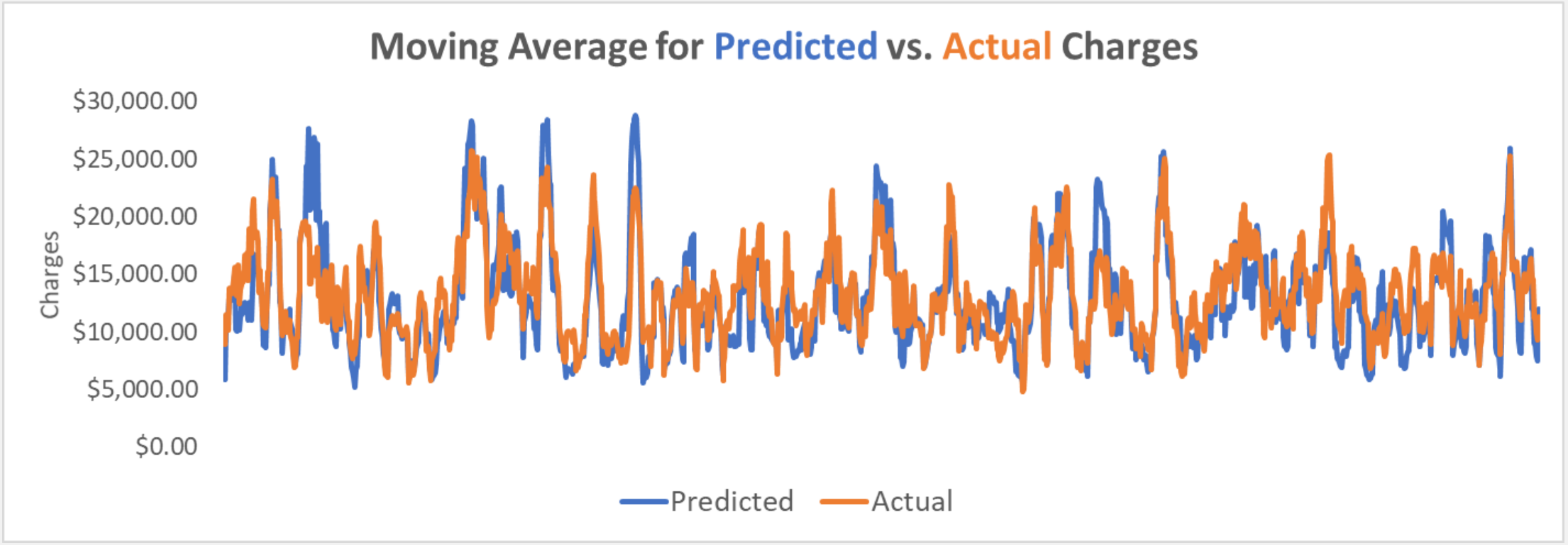
Observation	Predicted Log Charges	Residuals	Predicted Charges	Actual Charges	APE
1	9.42	0.31	$12,344.90	$16,884.92	26.89%
2	7.98	-0.53	$2,928.45	$1,725.55	69.71%
3	8.61	-0.20	$5,461.49	$4,449.46	22.74%
4	8.42	1.58	$4,520.30	$21,984.47	79.44%
5	8.44	-0.18	$4,647.99	$3,866.86	20.20%
6	8.37	-0.14	$4,327.85	$3,756.62	15.21%
7	9.07	-0.05	$8,650.10	$8,240.59	4.97%
8	8.89	0.00	$7,246.82	$7,281.51	0.48%
9	8.82	-0.05	$6,762.32	$6,406.41	5.56%
10	9.31	0.97	$11,001.82	$28,923.14	61.96%
11	8.13	-0.23	$3,411.80	$2,721.32	25.37%
12	10.91	-0.67	$54,604.37	$27,808.73	96.36%
13	8.13	-0.62	$3,409.27	$1,826.84	86.62%
14	9.34	-0.03	$11,401.11	$11,090.72	2.80%
15	9.93	0.66	$20,534.95	$39,611.76	48.16%
16	7.93	-0.42	$2,793.01	$1,837.24	52.02%
17	9.22	0.07	$10,103.22	$10,797.34	6.43%
18	8.02	-0.24	$3,048.78	$2,395.17	27.29%
19	9.35	-0.08	$11,459.21	$10,602.39	8.08%
20	9.98	0.54	$21,543.32	$36,837.47	41.52%
...					

Along with adjusting the BMI, removing non-significant variables, and taking the log of charges, I made one final adjustment by transforming the age variable using its square root. These adjustments allowed me to reduce the MAPE to a level of just under 27%.

MAPE
26.73%

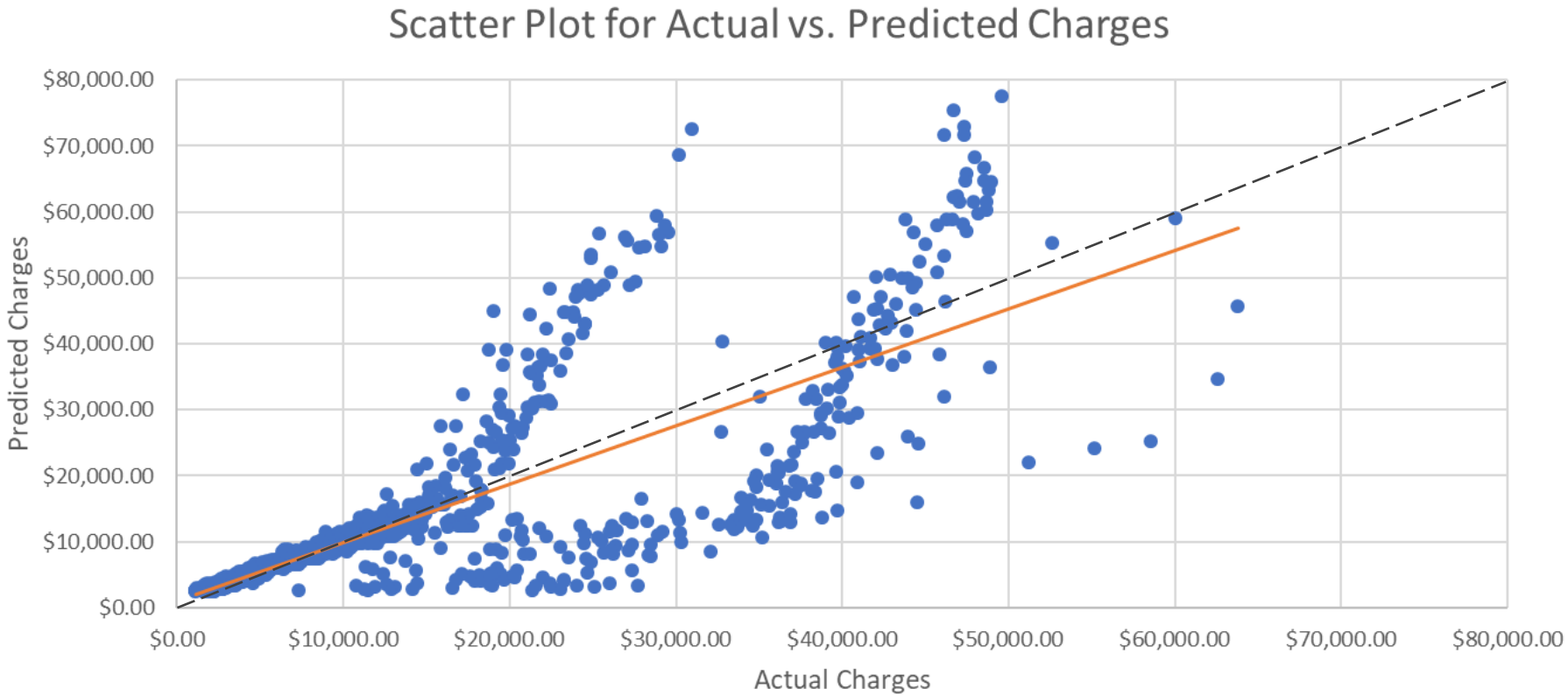
Predicted vs. Actual Charges

Using the moving average to smooth out the results, I was able to predict the medical charges of the sample population with 73% accuracy.



Predicted vs. Actual Charges

The scatter plot of actual vs. predicted charges shows that charges that are less than $15,000 annually can be predicted with much higher accuracy than charges that exceed this level.



Thank You

Eric Wheeler
917-562-3434
Ericwheeler.90@gmail.com