The Cost of Medical Care

Can general health factors predict insurance billing?

Corey Quackenbush
CS 522 – Data Mining

Data summary

Summary statistics for project data

age	
Min.	18.0
1st Qu.	27.0
Median	39.0
Mean	39.2
3rd Qu.	51.0
Max.	64.0

charges	
Min.	1122
1st Qu.	4740
Median	9382
Mean	13270
3rd Qu.	16640
Max.	63770

bmi	
Min.	16.0
1st Qu.	26.3
Median	30.4
Mean	30.7
3rd Qu.	34.7
Max.	53.1

children	
0	574
1	324
2	240
3	157
4	25
5	18

sex	
female	662
male	676

region	
northeast	324
northwest	325
southeast	364
southwest	325

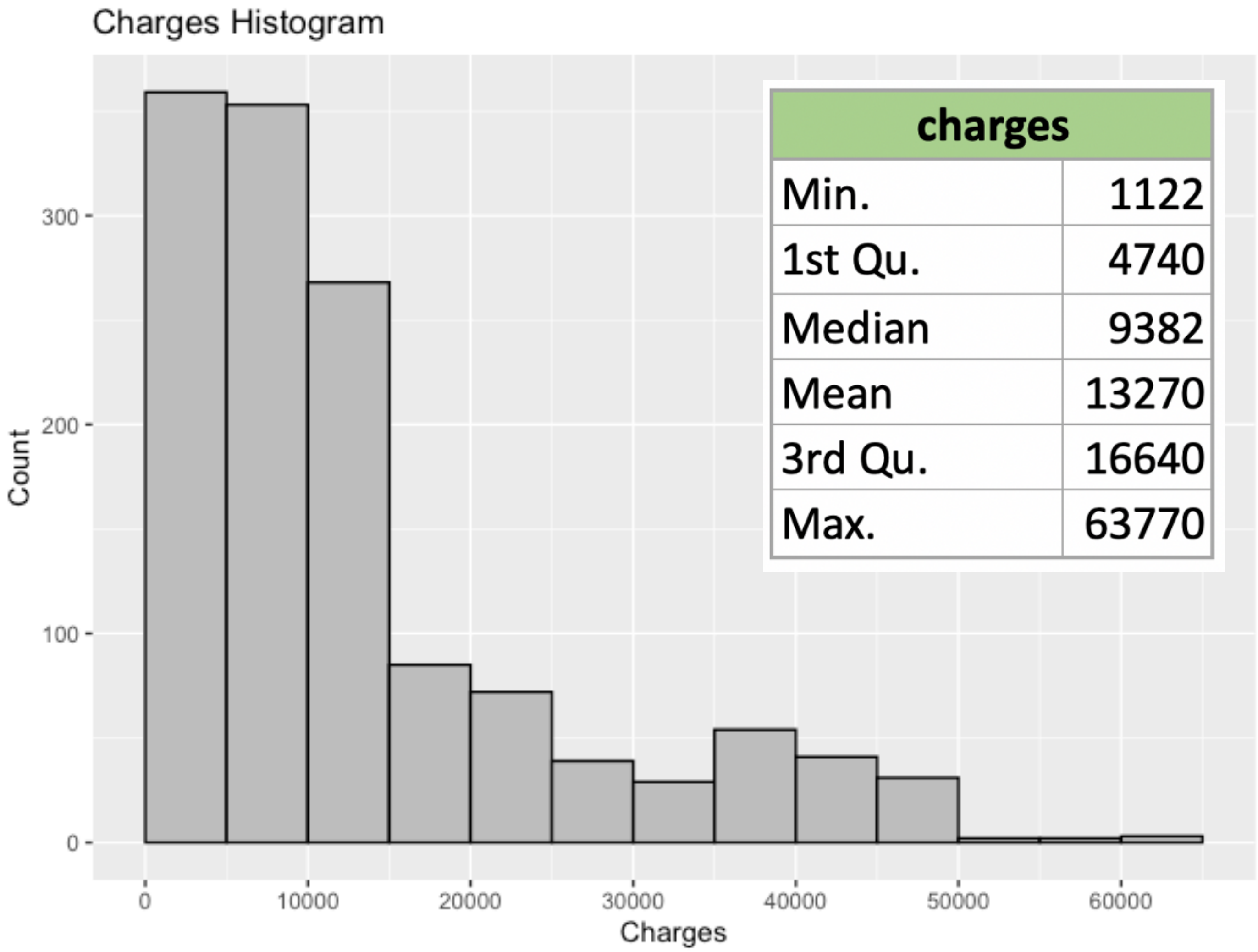
bmi_category	
Underweight	20
Normal	225
Overweight	386
Obese	707

smoker	
yes	274
no	1064

obese	
yes	707
no	631

Field of interest – yearly medical charges

- Can we predict these values based on some or all the other fields present?
- Difference between median and mean (Right-skew).
- Could be a problem because assumes data is normally distributed.



charges	
Min.	1122
1st Qu.	4740
Median	9382
Mean	13270
3rd Qu.	16640
Max.	63770

Partition data in to test and training

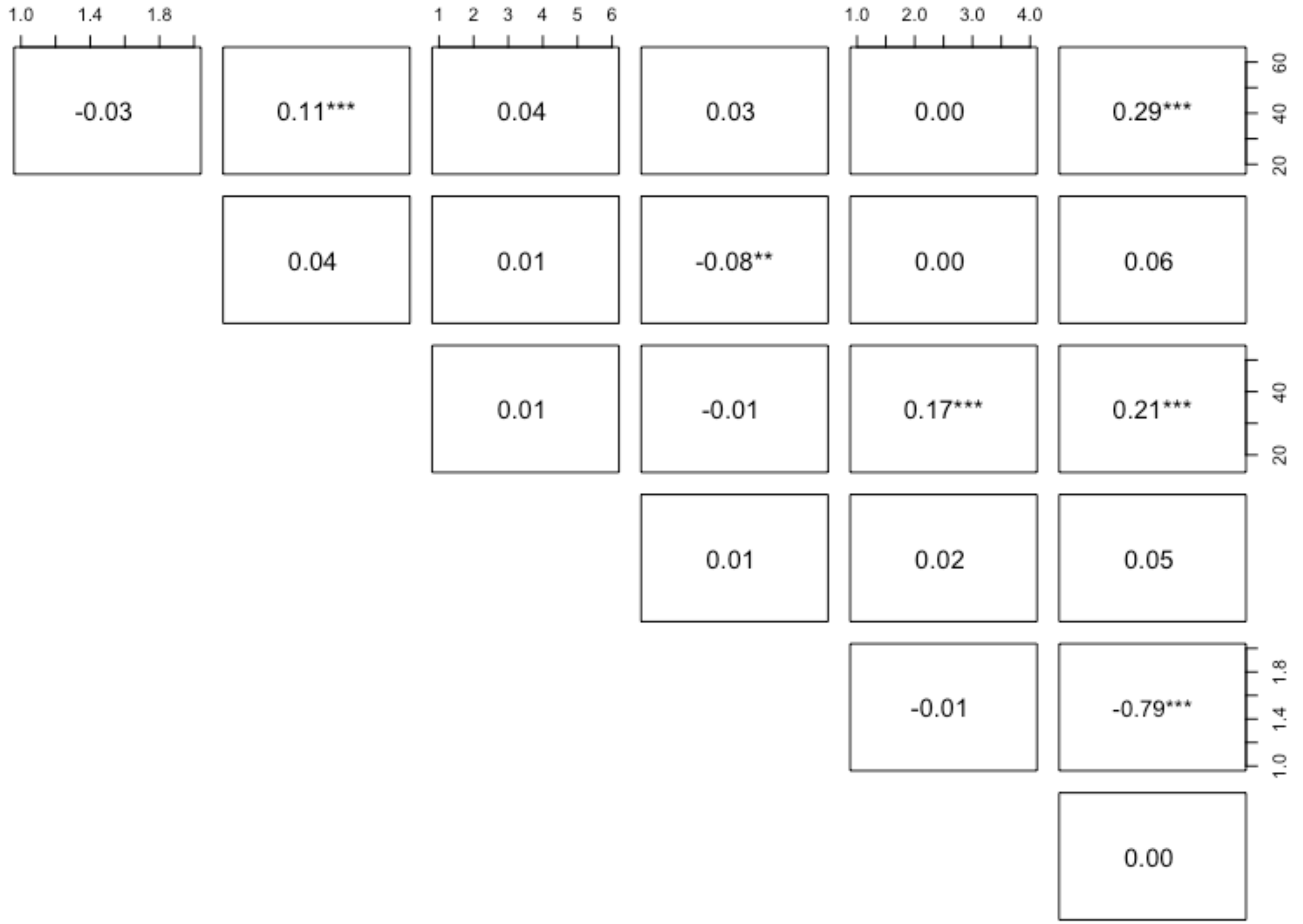
```r
# Set seed to make sample reproducible.
# Split data into testing and training sets. Using ~10% (134) from testing.
set.seed(123)
sampleID <- sample(nrow(insurance), 134, replace = F)

test_set     <- insurance[sampleID, ]
training_set <- insurance[-sampleID, ]
```

- Using 90% of the data for training – 1204 entries.
- Using 10% of the data for testing – 134 entries.
- Total – 1338 entries.

Check for correlation between data attributes

Scatter Plot Matrix

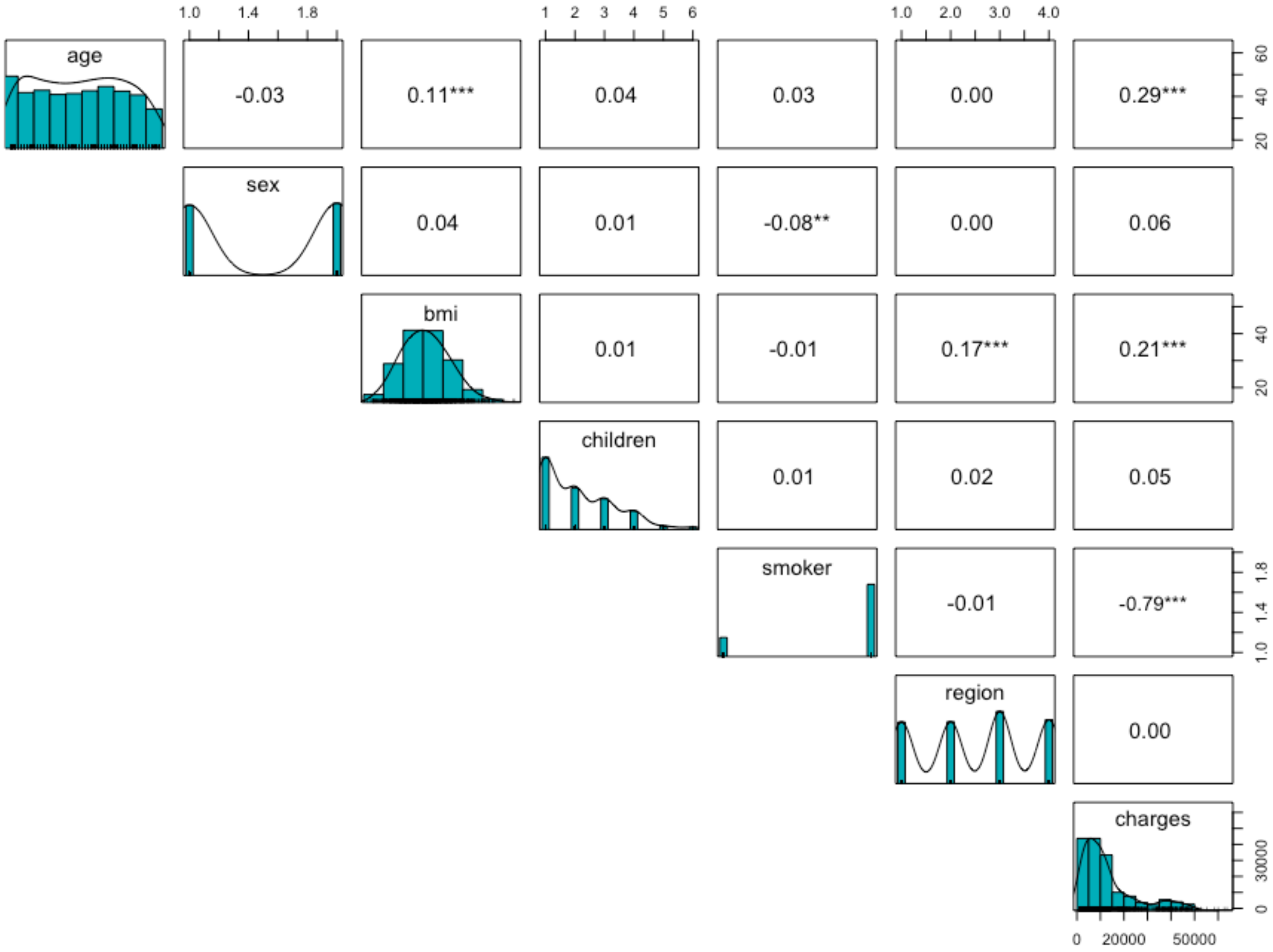
- Correlation in upper right



Check for correlation between data attributes

Scatter Plot Matrix

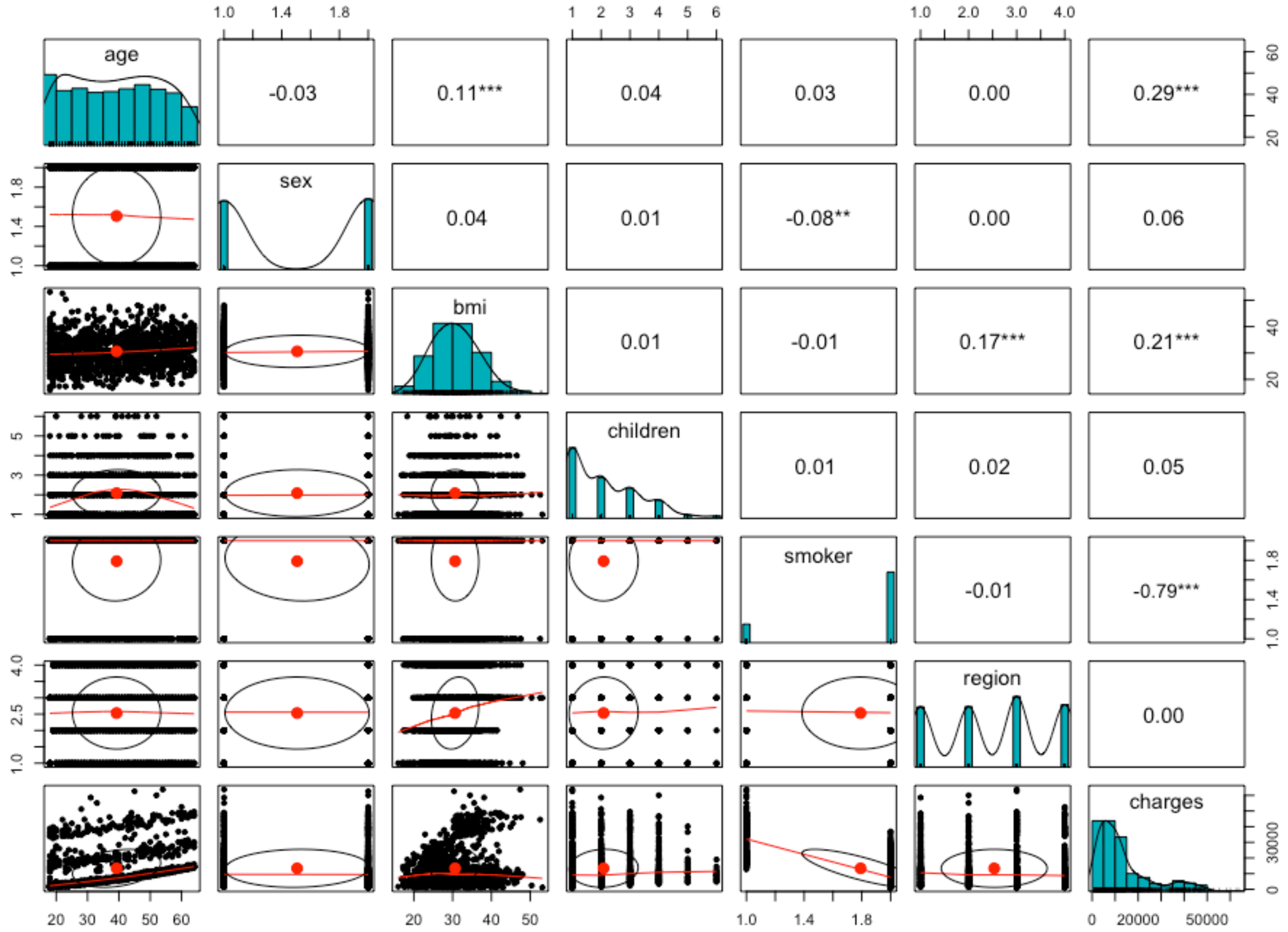
- Correlation in upper right
- Attribute histograms on diagonal



Check for correlation between data attributes

Scatter Plot Matrix

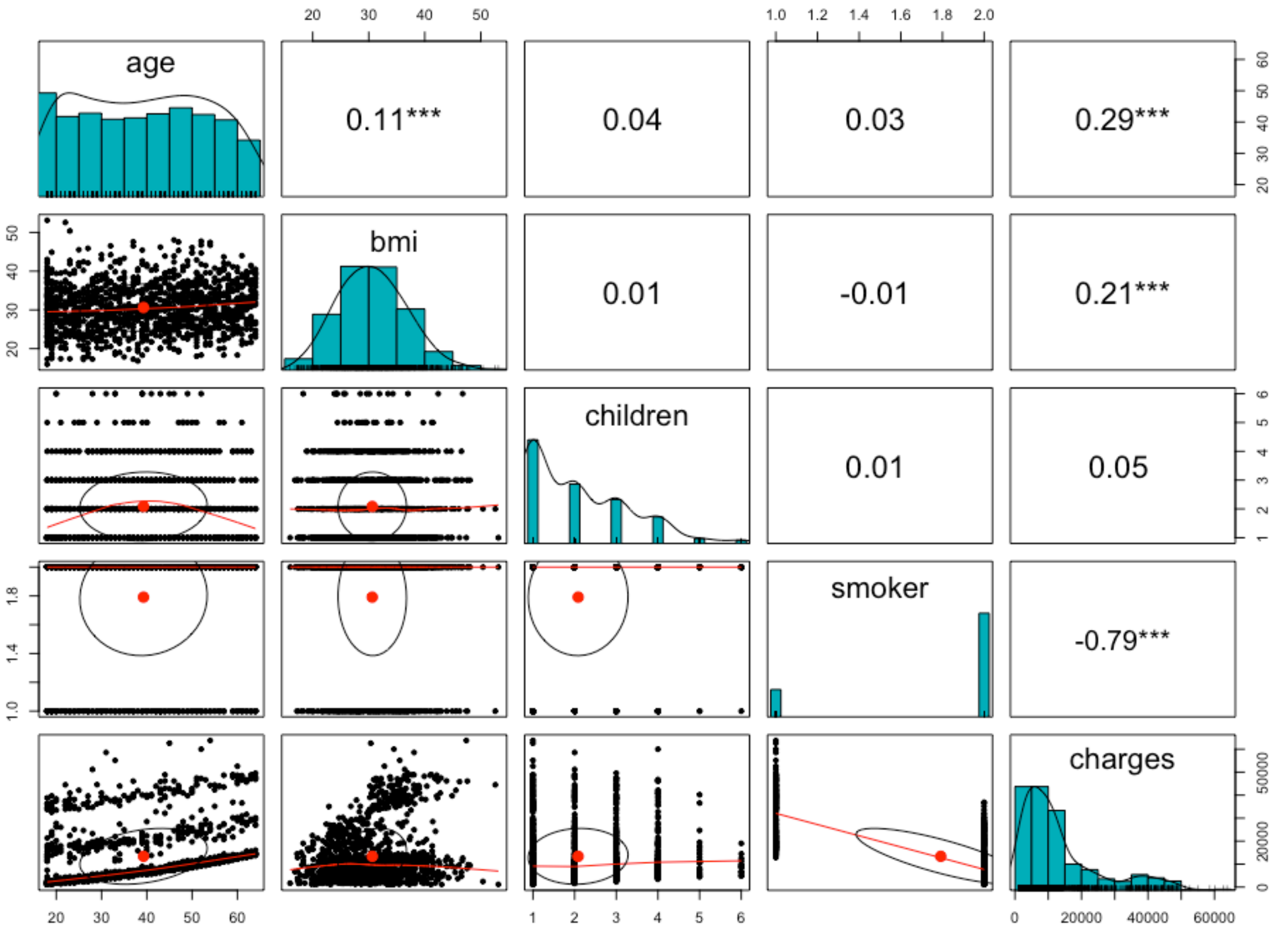
- Correlation in upper right
- Attribute histograms on diagonal
- Scatter plots between attributes in lower left
- Oval – correlation ellipse
  - Stretched: strong correlation
  - Round: little/no correlation
- Curve - loess smooth
  - General relationship between x/y



Choose attributes for model

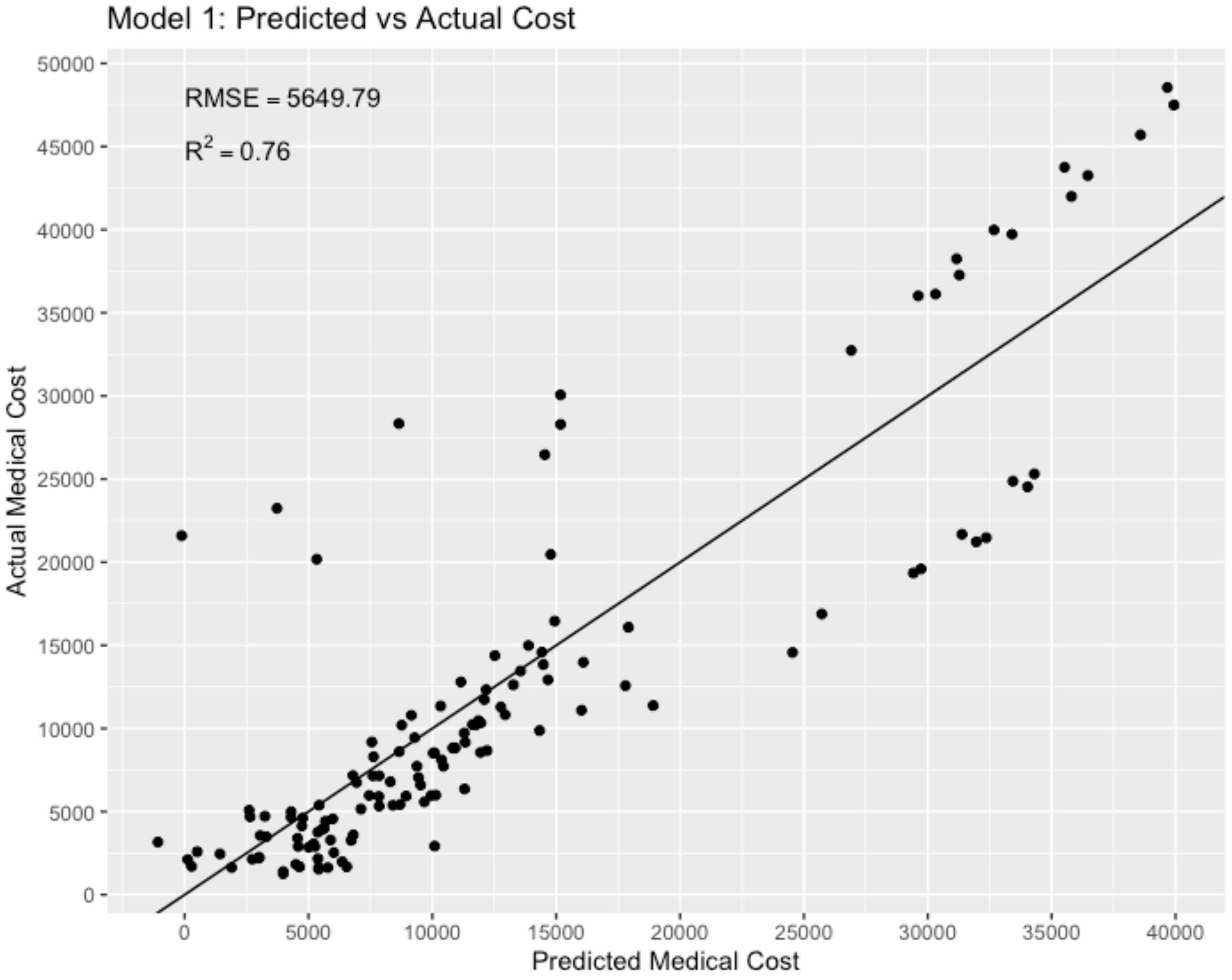
```
# Explore the effects of dropping variables from the model.
summary(lm(charges ~ age + sex + bmi + children + smoker + region, data = training_set))
summary(lm(charges ~ age + sex + bmi + children + smoker, data = training_set))
summary(lm(charges ~ age + bmi + children + smoker, data = training_set))
```

- Tested various models.
- Focused on including attributes correlated with the charges attribute.
- Used summary statistics of model to choose the best one.
- Scatter plot matrix of final attributes chosen for model.



Model generation and testing

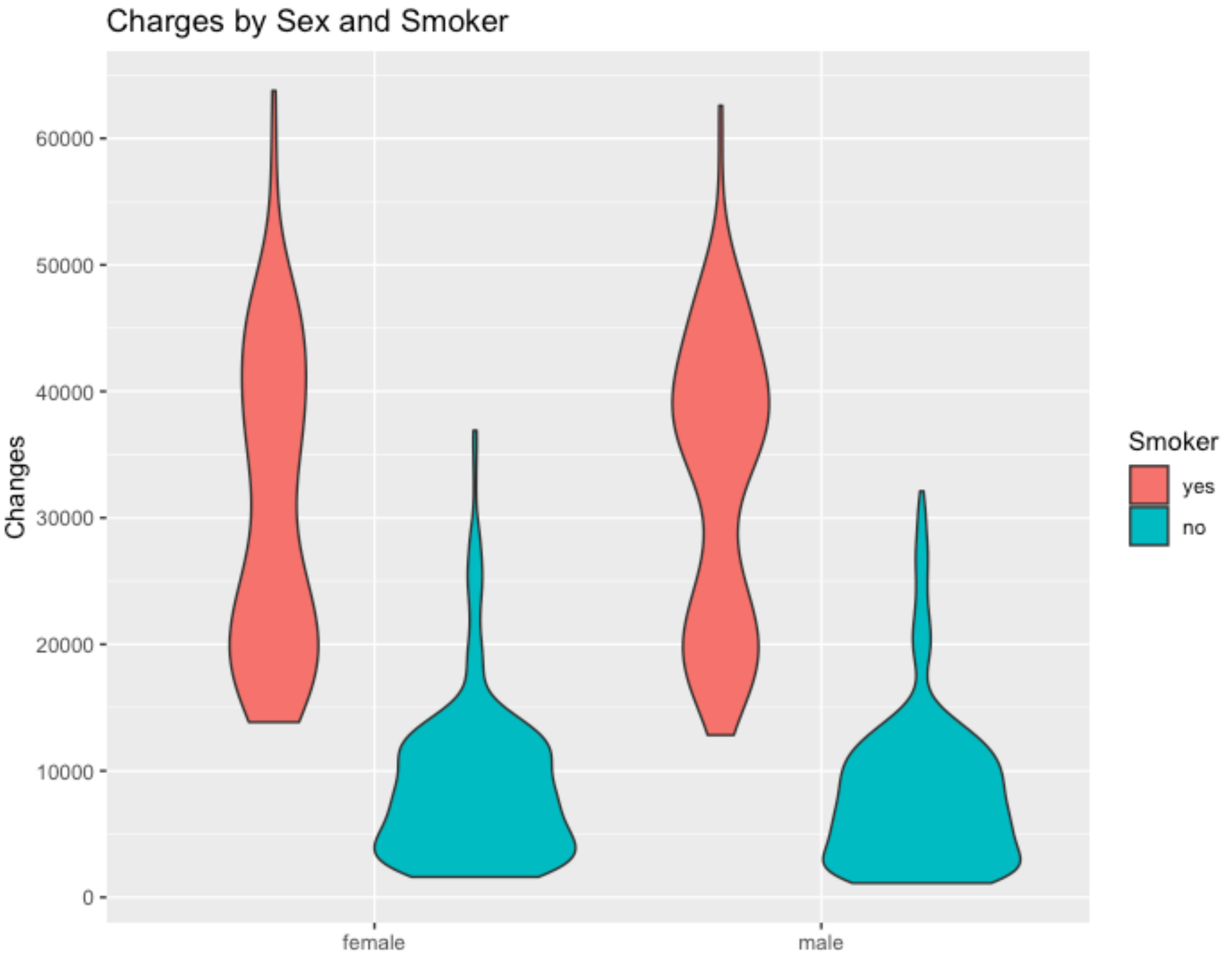
```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-11884.0  -3015.2   -868.3   1679.2  29418.9 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11582.90    1055.97  10.969   <2e-16 ***
age            255.10      12.63  20.197   <2e-16 ***
bmi            333.05      29.05  11.464   <2e-16 ***
children1      183.05     445.66   0.411   0.6813    
children2     1250.12     493.14   2.535   0.0114 *  
children3      802.76     597.51   1.344   0.1794    
children4     3247.49    1277.76   2.542   0.0112 *  
children5      489.65    1552.87   0.315   0.7526    
smokerno    -23804.78     435.30 -54.686   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6112 on 1195 degrees of freedom
Multiple R-squared:  0.7493,    Adjusted R-squared:  0.7476 
F-statistic: 446.4 on 8 and 1195 DF,  p-value: < 2.2e-16
```



Data Exploration

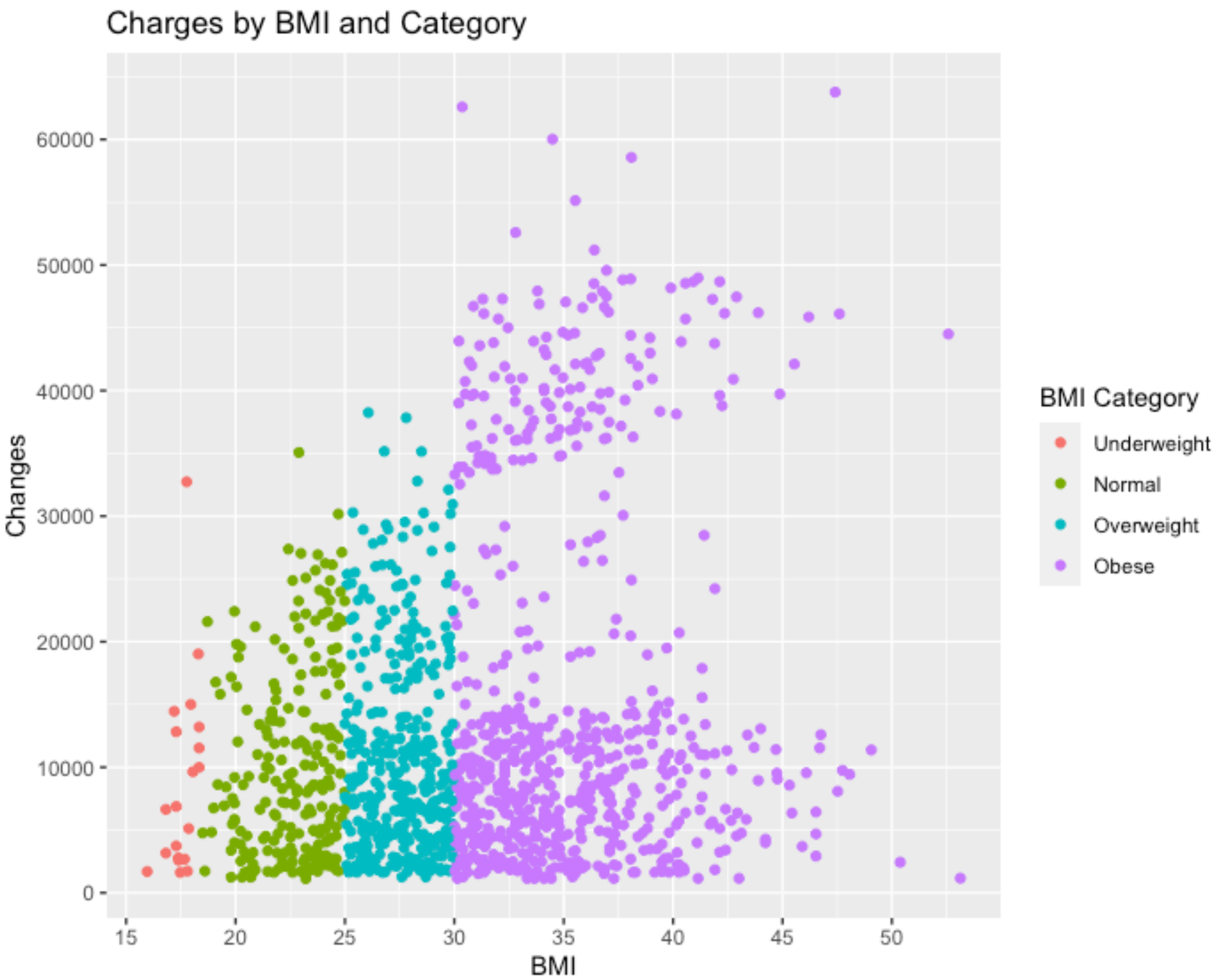
Data exploration

- Charges higher if you smoke.
- Not much difference between male/female smoker and non-smokers.
- Might be two groups within the smoker set.



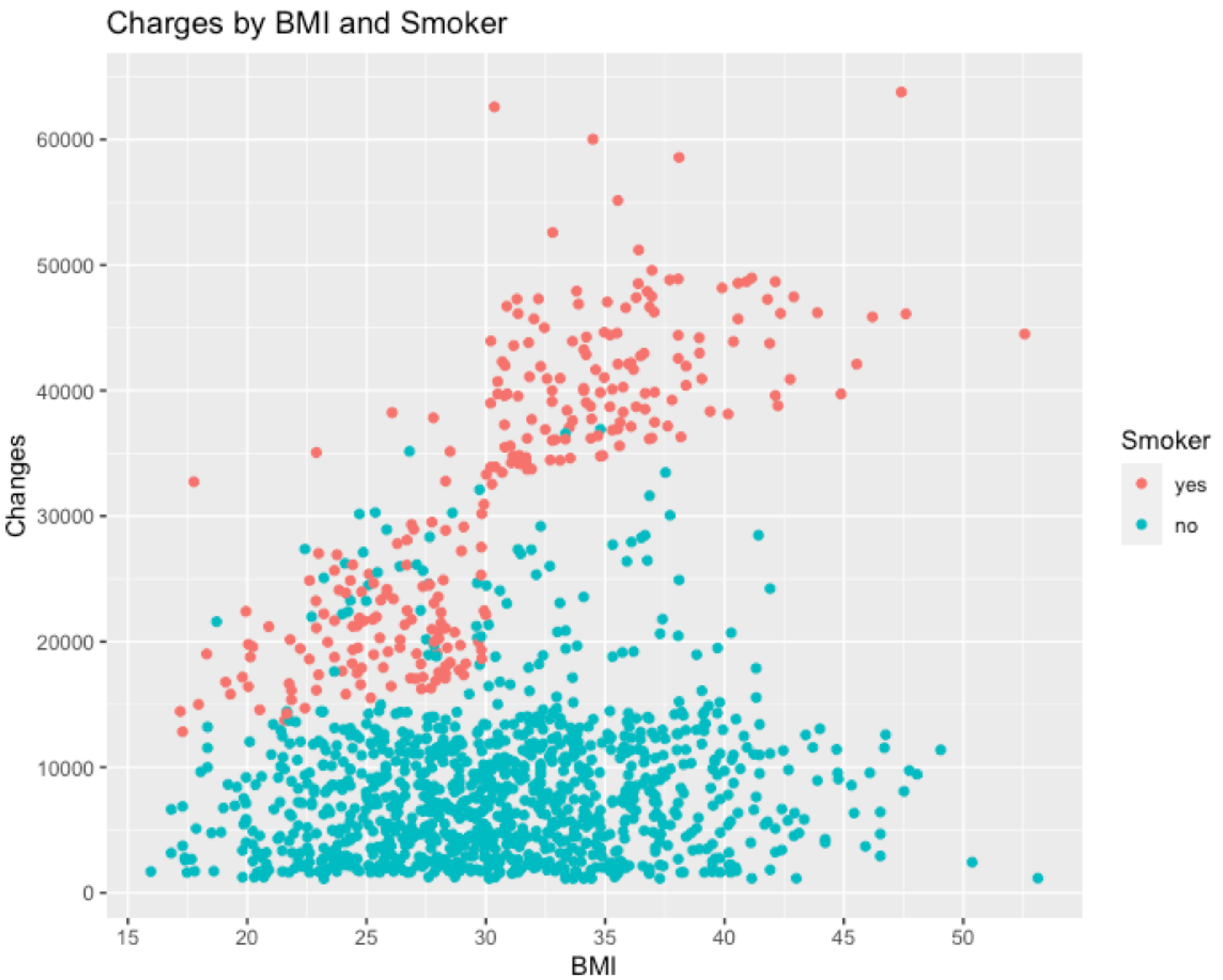
Data exploration

- Interesting second group in the obese category.
- Much higher medical charges in second obese group.



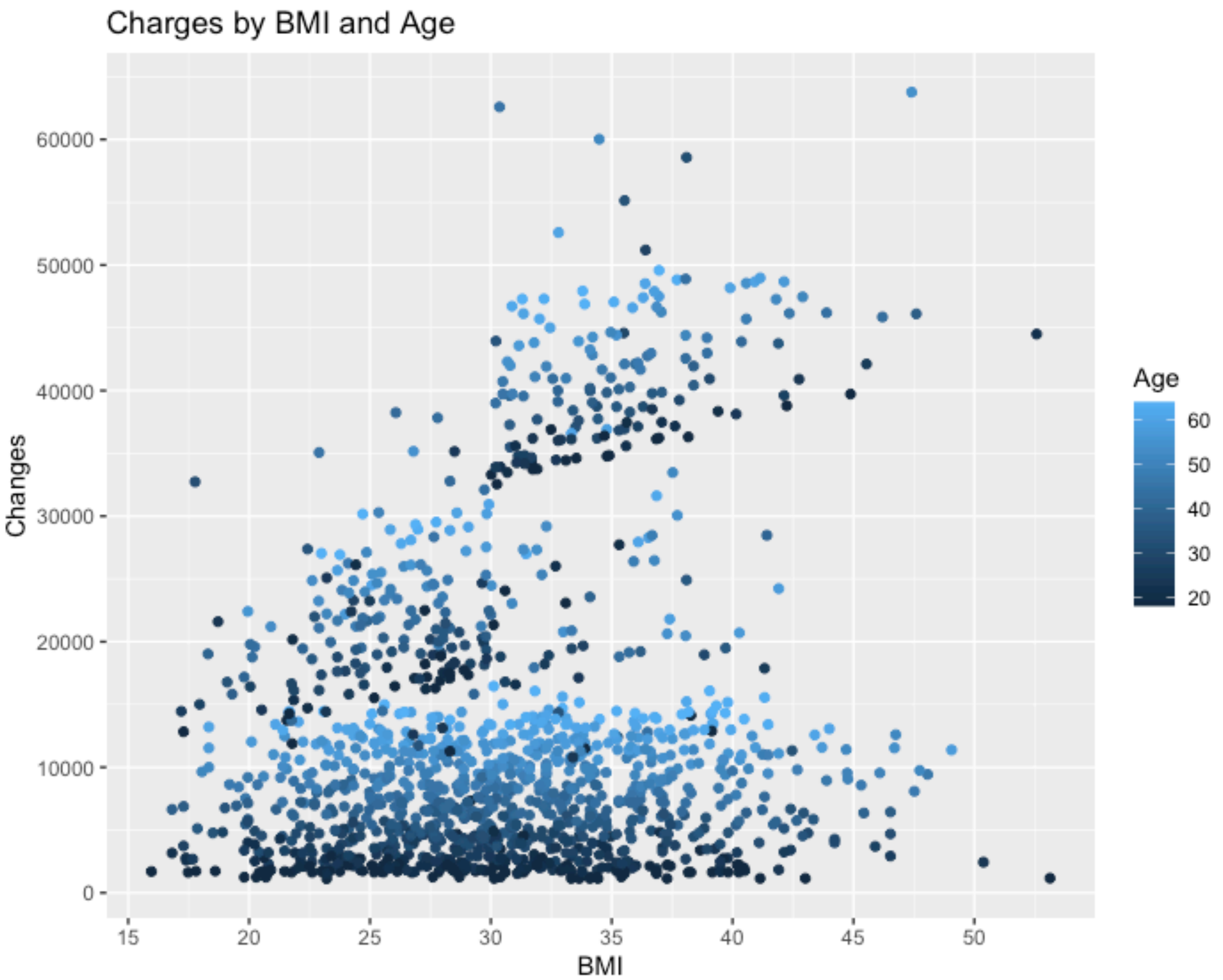
Data exploration

- Again, smokers pay more than non-smokers.
- Seems reasonable that the second group from previous plot could be obese smokers.



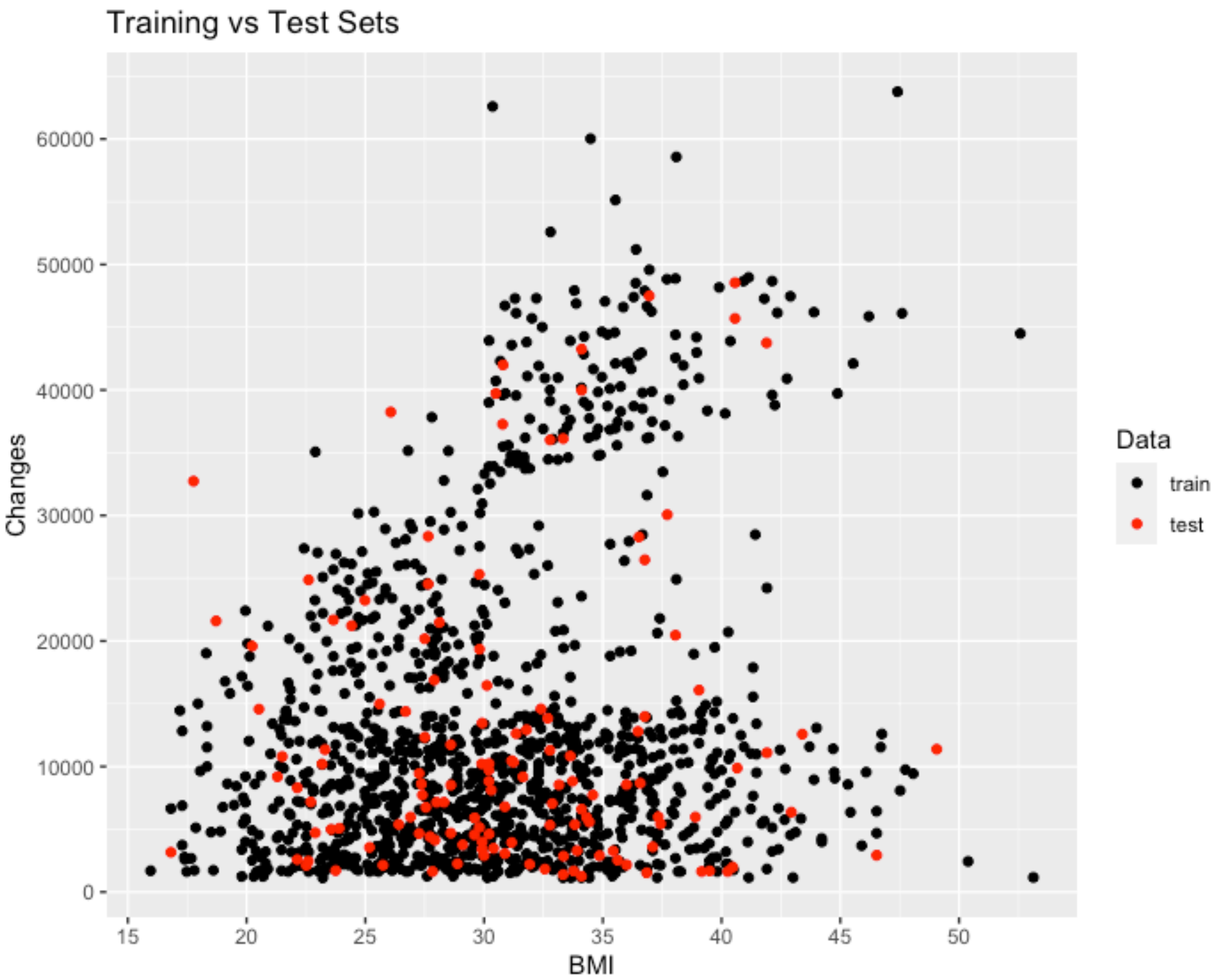
Data exploration

- Older people tend to pay more in medical expenses.
- The trend replicates well within groups.



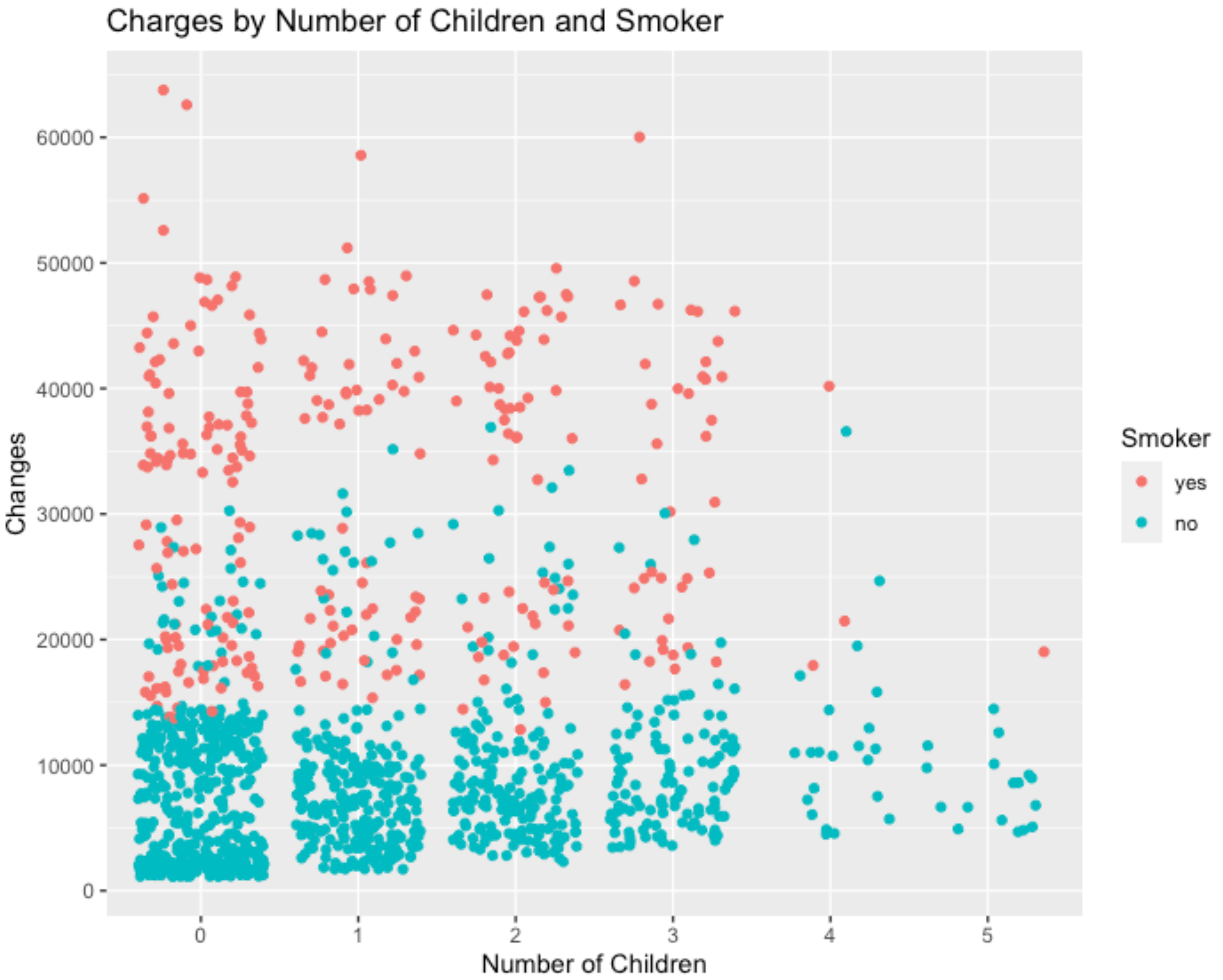
Data exploration

- Visualize the training and testing data.
- Nice random distribution.



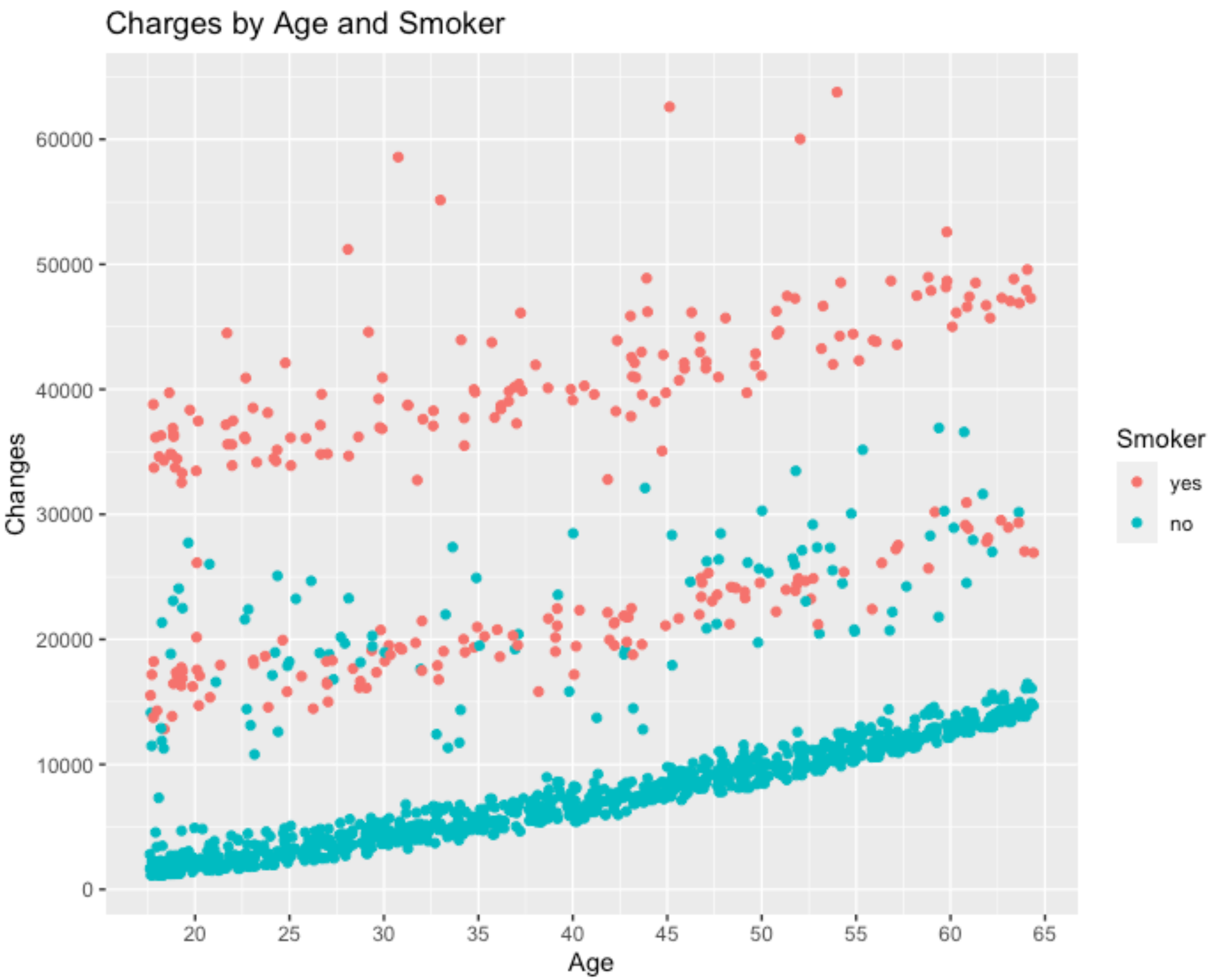
Data exploration

- Minimum amount paid in medical expenses increases with number of children.
- Whether or not you smoke dominates the relationship with charges.



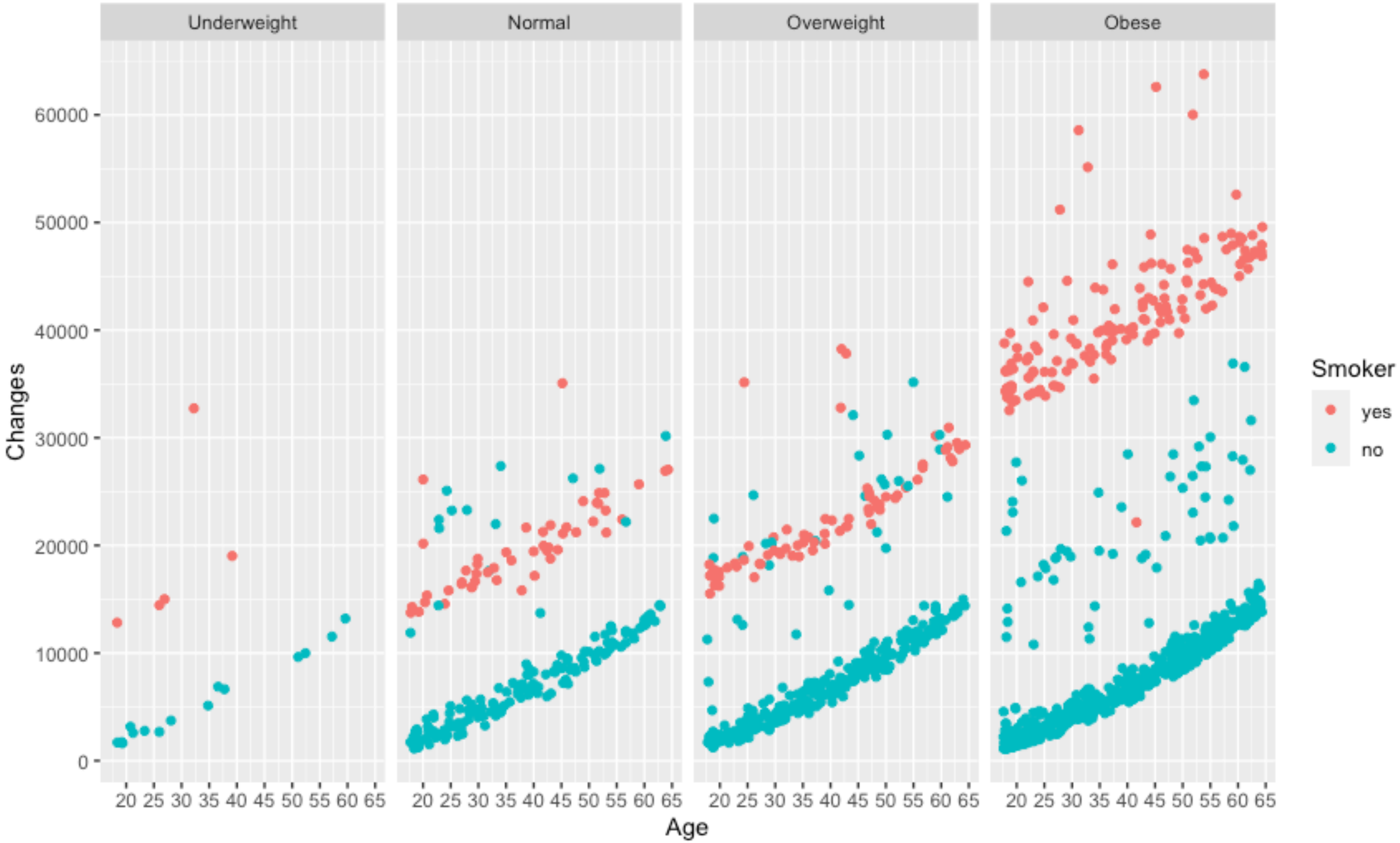
Data exploration

- Again, older people pay more in medical expenses.
- Again, you tend to pay more if you smoke.
- There is clearly a second group within the smokers



Data exploration

- There is a combined effect for being obese and smoking.

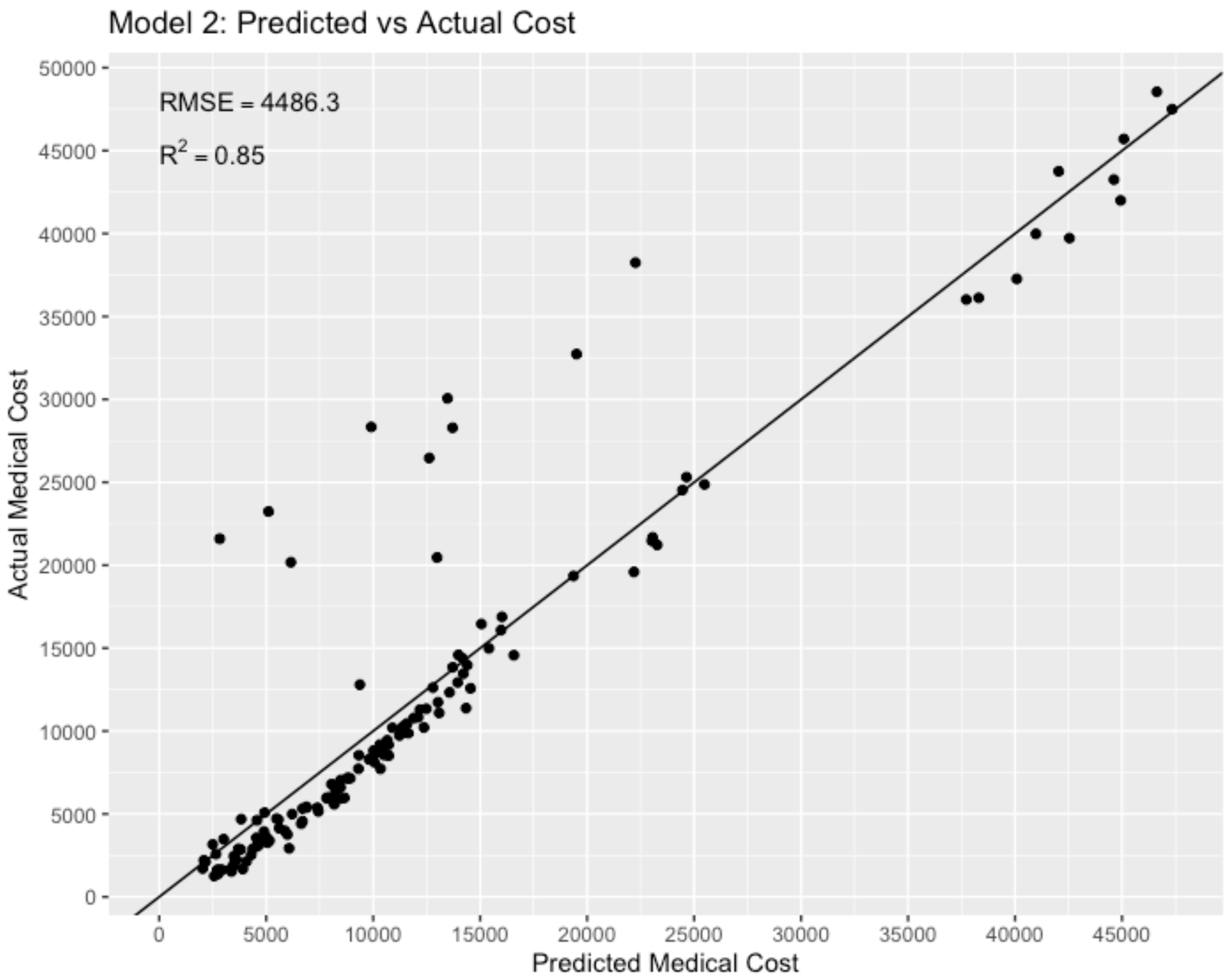


Data exploration

- Again, smoking results in higher medical expenses.
- There is a combined effect for being obese and smoking.
- We can use this observation to update our regression model.

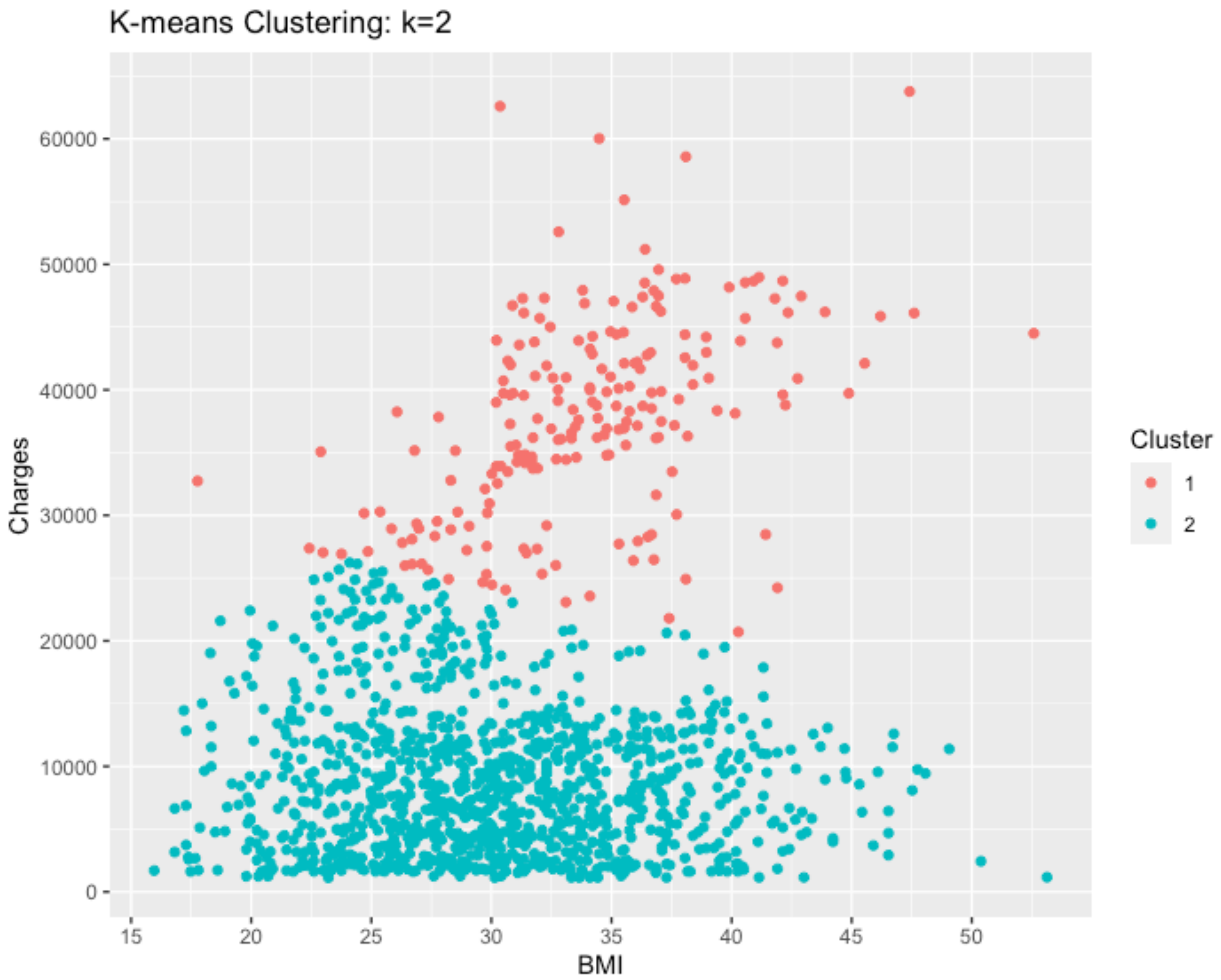
Update model: smoker – obese interaction

```
Call:
lm(formula = charges ~ age + bmi + children + smoker * obese,
    data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-19168.7  -1882.7  -1180.8   -396.4  24375.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       27215.68    1359.66  20.017  < 2e-16 ***
age                 262.54       9.32  28.170  < 2e-16 ***
bmi                 102.35      35.82   2.858  0.00434 **
children1           391.23     329.08   1.189  0.23473
children2          1112.69     363.85   3.058  0.00228 **
children3          1082.03     441.11   2.453  0.01431 *
children4          3572.56     943.35   3.787  0.00016 ***
children5          1040.89    1145.79   0.908  0.36382
smokerno         -33122.92     440.10 -75.262  < 2e-16 ***
obeseno          -19041.27     677.16 -28.119  < 2e-16 ***
smokerno:obeseno  19823.46     641.77  30.889  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4508 on 1193 degrees of freedom
Multiple R-squared:  0.8638,    Adjusted R-squared:  0.8627
F-statistic: 756.6 on 10 and 1193 DF,  p-value: < 2.2e-16
```

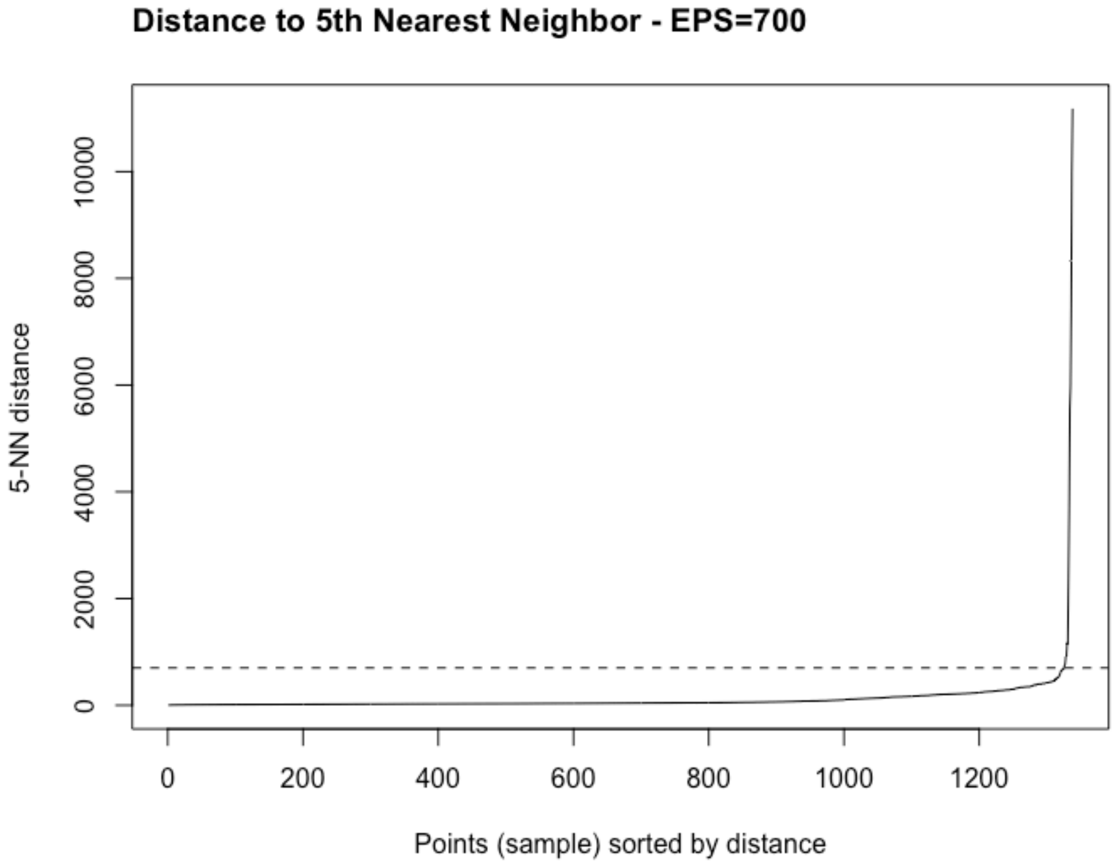


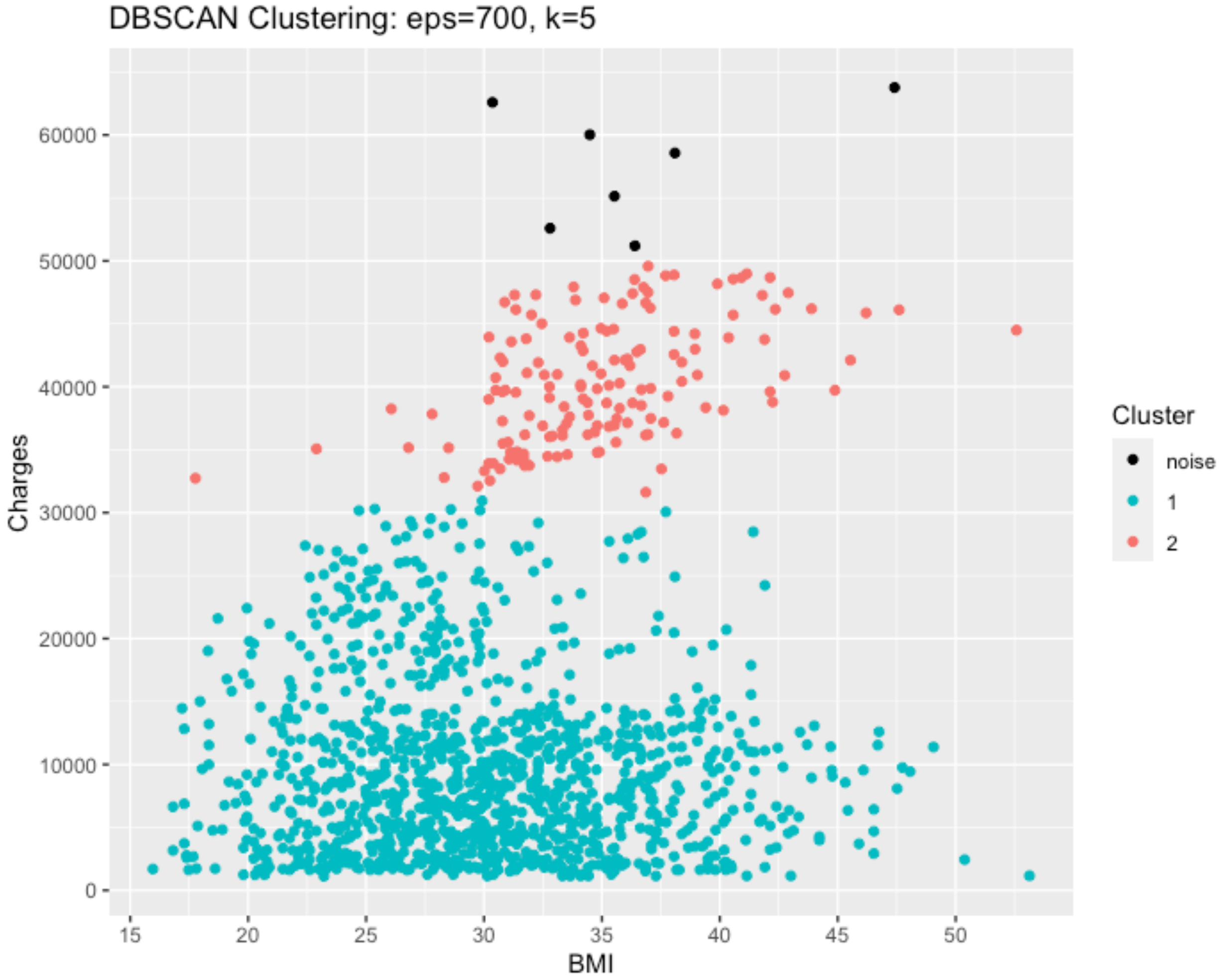
Clustering

K-means clustering

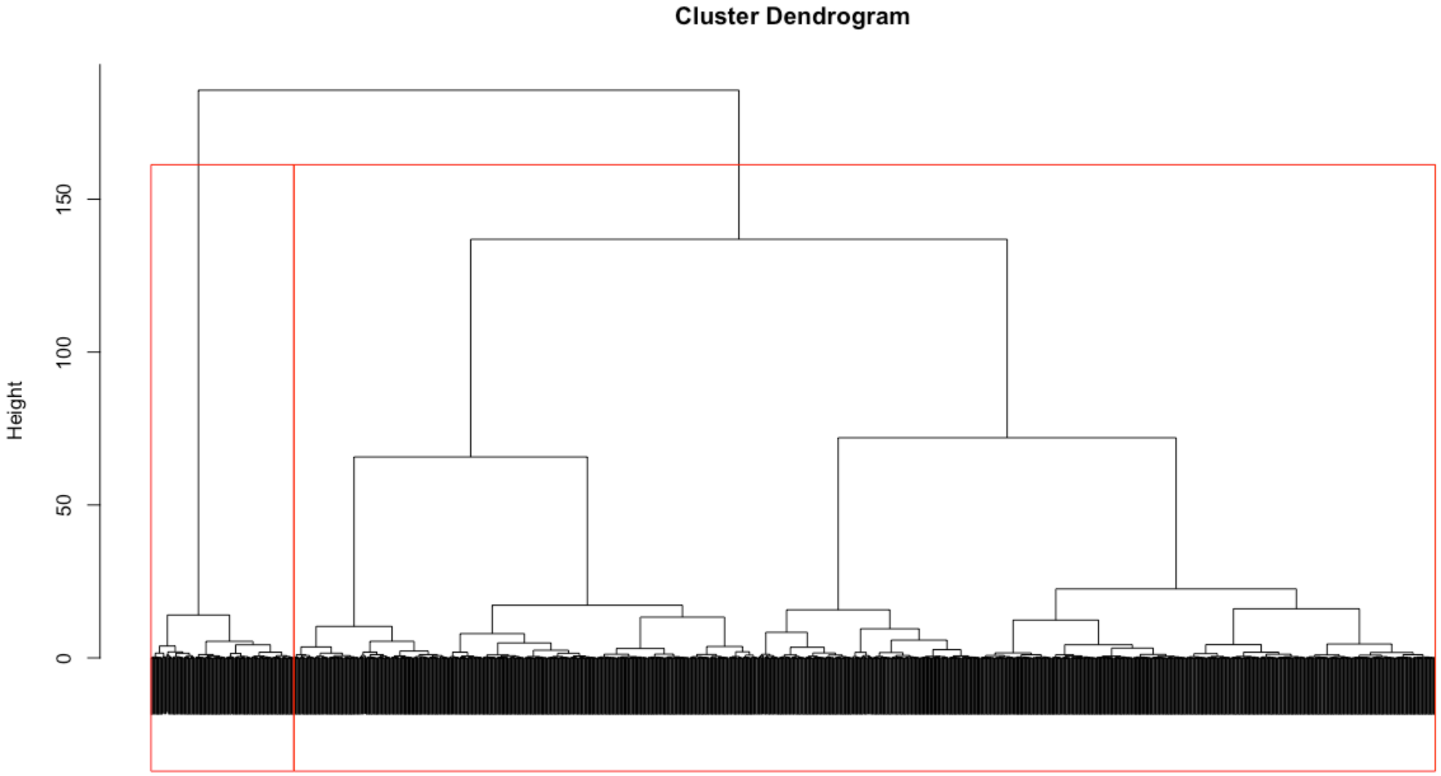


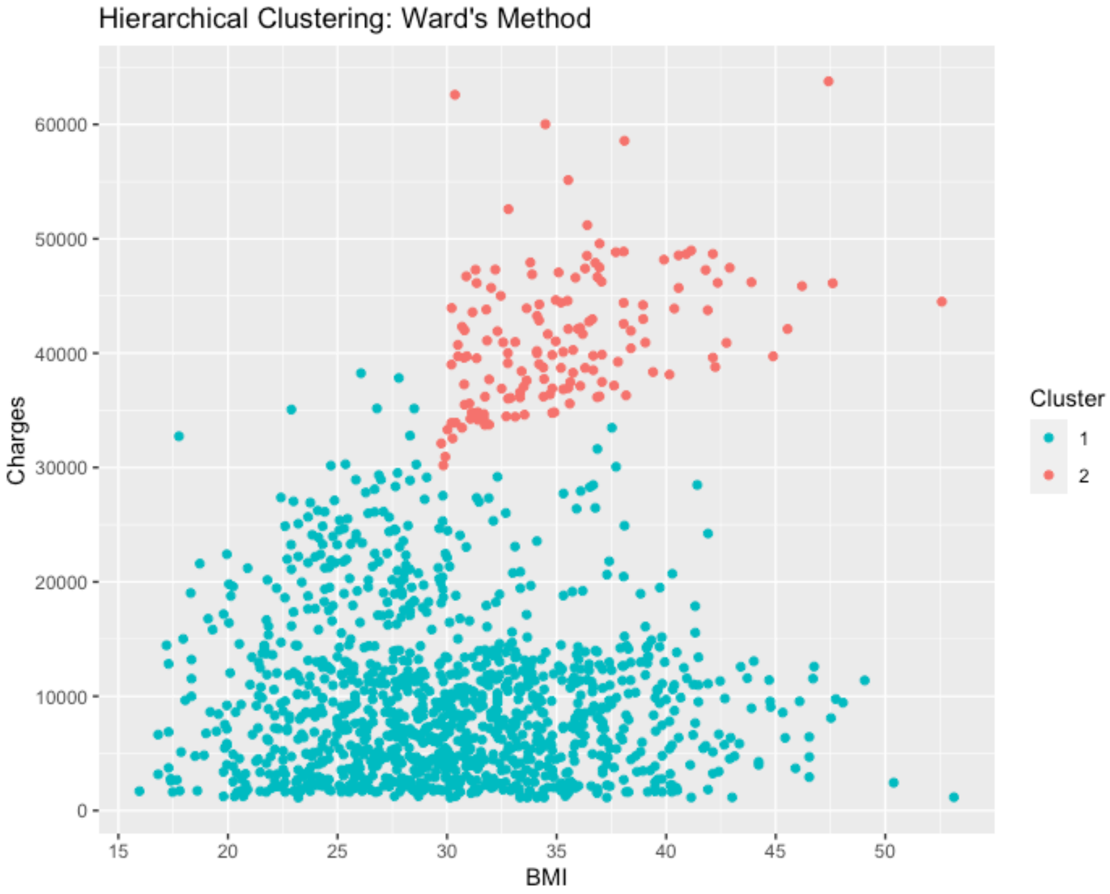
DBSCAN clustering



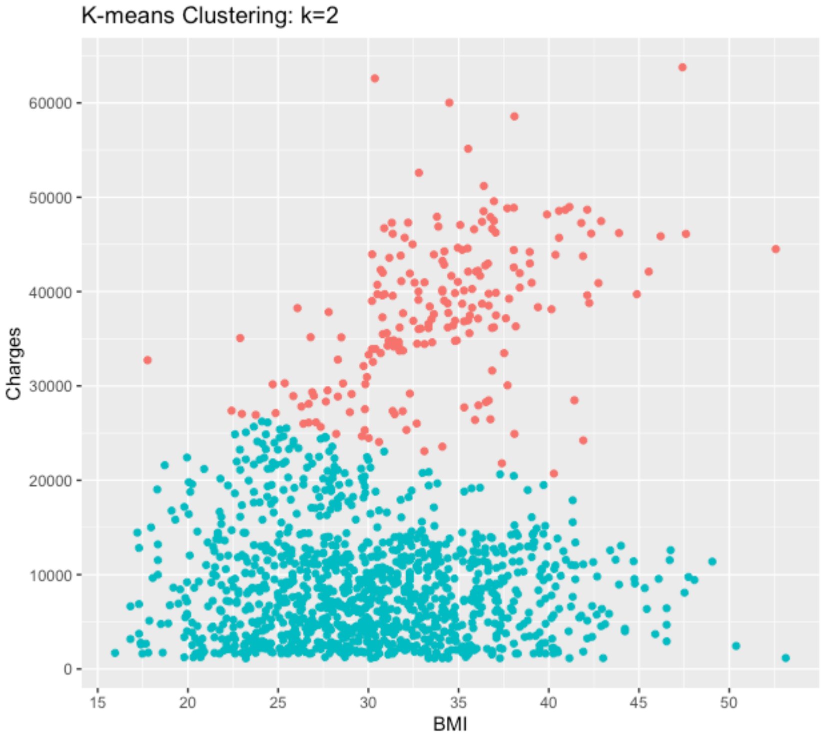


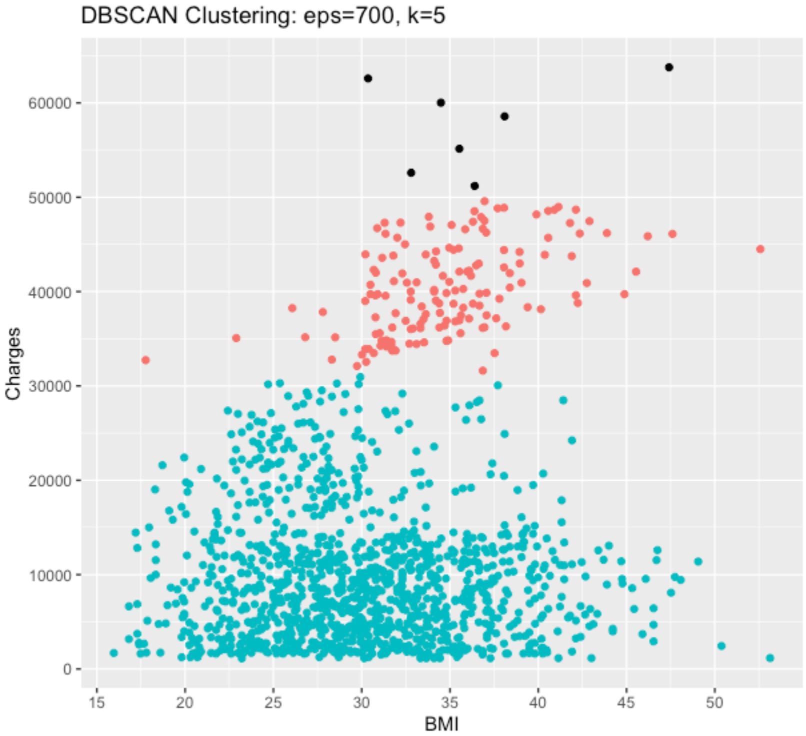
Hierarchical clustering: Ward’s method

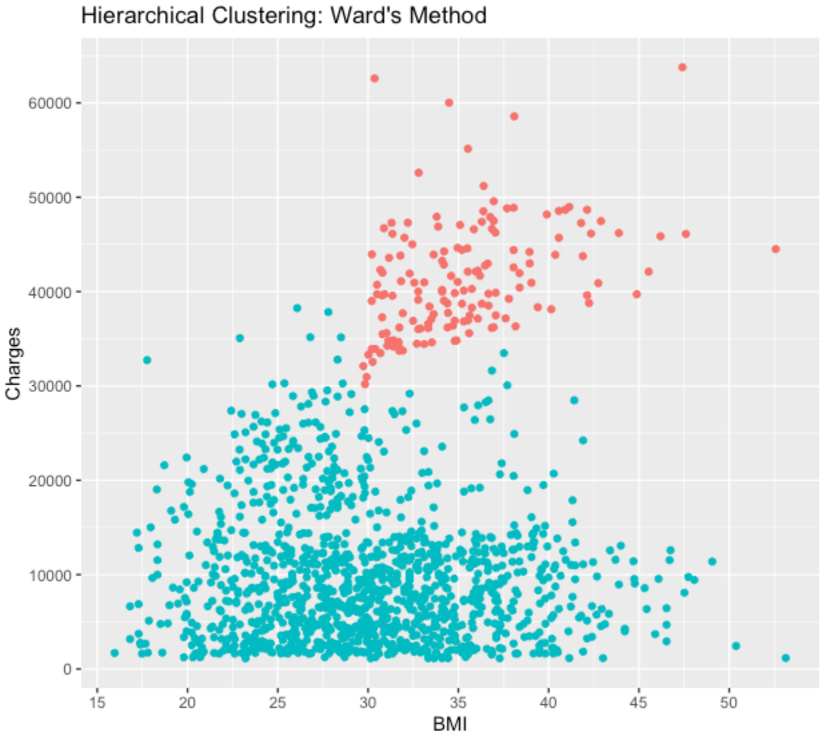




Clustering method comparison







Final Notes

- Can predict yearly medical costs given the following information:
  - Age, BMI, Children number, Smoker or not
- Model explains 86% of variation in medical costs.
- Majority of predictions are over predicted by only $400 to $1800.
  - Not bad considering the input information.
- Can group the data nicely with hierarchal clustering using Ward's method.
- Can use nearest neighbor to assign new data to a cluster.
  - Use groups to infer general information.
    - BMI/Charges → Smoker
    - BMI/Smoker → Charge range
    - Smoker/Charges → Obese / BMI range