CASE STUDY

By,
D.vincy

About data:

This dataset contains 1338 rows of insured data, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker and Region

Variable description:

Variables	description
Age	Age of primary beneficiary
Sex	Insurance contractor gender, female / male
BMI	Body mass index, providing an understanding of body, weights that are relatively high or low relative to
Children	Number of children covered by health insurance / Number of dependents
Smoker	Smoker / Non - smoker
Region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest
charges	Individual medical costs billed by health insurance

PROBLEM STATEMENT:

Predicting insurance premiums using linear regression

- Health insurance is one of the most marketed products offered by leading insurance firms. The bottom line in this industry is driven by the simple fact that the capital spent by the insurance company in response to beneficiary claims should not exceed customer premium. Higher the difference between settled claims and total premium received, higher are the profits.
- In the following analysis, we try to analyse the correlations between various customer attributes and develop a predictive model that would help the company charge adequete and appropriate premium to the clients.
- This dataset has 7 variables and 1338 records.

DATA PREPROCESSING:

Missing data:

There is no missing data in the dataset. so we may proceed without worrying about having to impute any data.

Summary of the dataset:

age	sex	bmi	children	smoker	region	charges
Min. :18.00	female:662	Min. :15.96	Min. :0.000	no :1064	northeast:324	Min. : 1122
1st Qu.:27.00	male :676	1st Qu.:26.30	1st Qu.:0.000	yes: 274	northwest:325	1st Qu.: 4740
Median :39.00		Median :30.40	Median :1.000		southeast:364	Median : 9382
Mean :39.21		Mean :30.66	Mean :1.095		southwest:325	Mean :13270
3rd Qu.:51.00		3rd Qu.:34.69	3rd Qu.:2.000			3rd Qu.:16640
Max. :64.00		Max. :53.13	Max. :5.000			Max. :63770

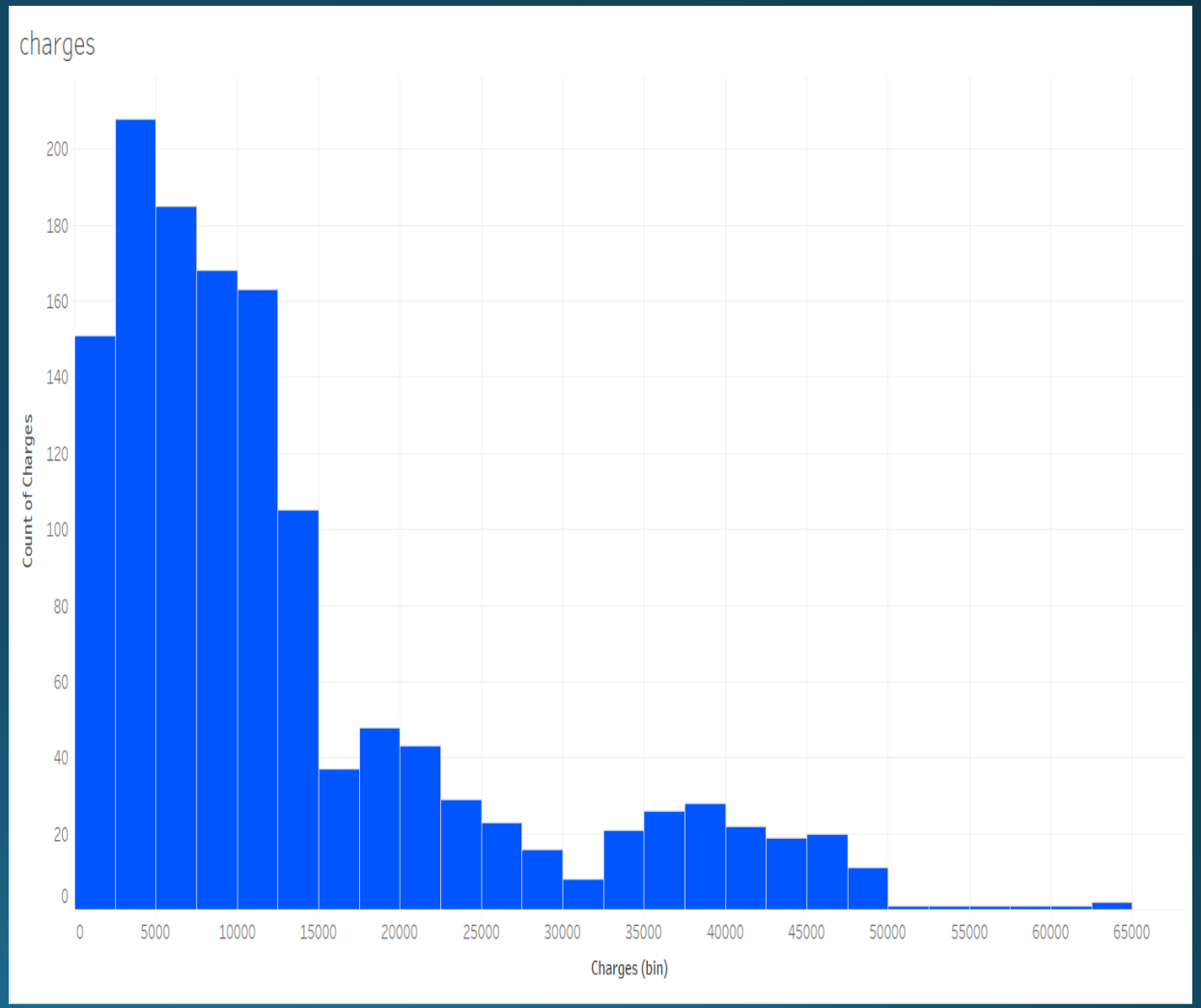
From the summary function,

- Looking at the response variable, the minimum value is 1122 while the maximum value is 63770. Most points cluster between 4740 and 16640. This large variance in the response variable indicates that there are potential outliers. The other quantitative variables are reasonably varied.
- smoker is unbalanced. more people are non-smokers.
- charges seems to be right skewed (median < mean).
- age, bmi, children seem to be normally distributed.
- sex, region, seems to be balanced.
- The Children variable seems to be a categorical variable because there are only a few values in the variable (1-5).

EDA(using Tableau):

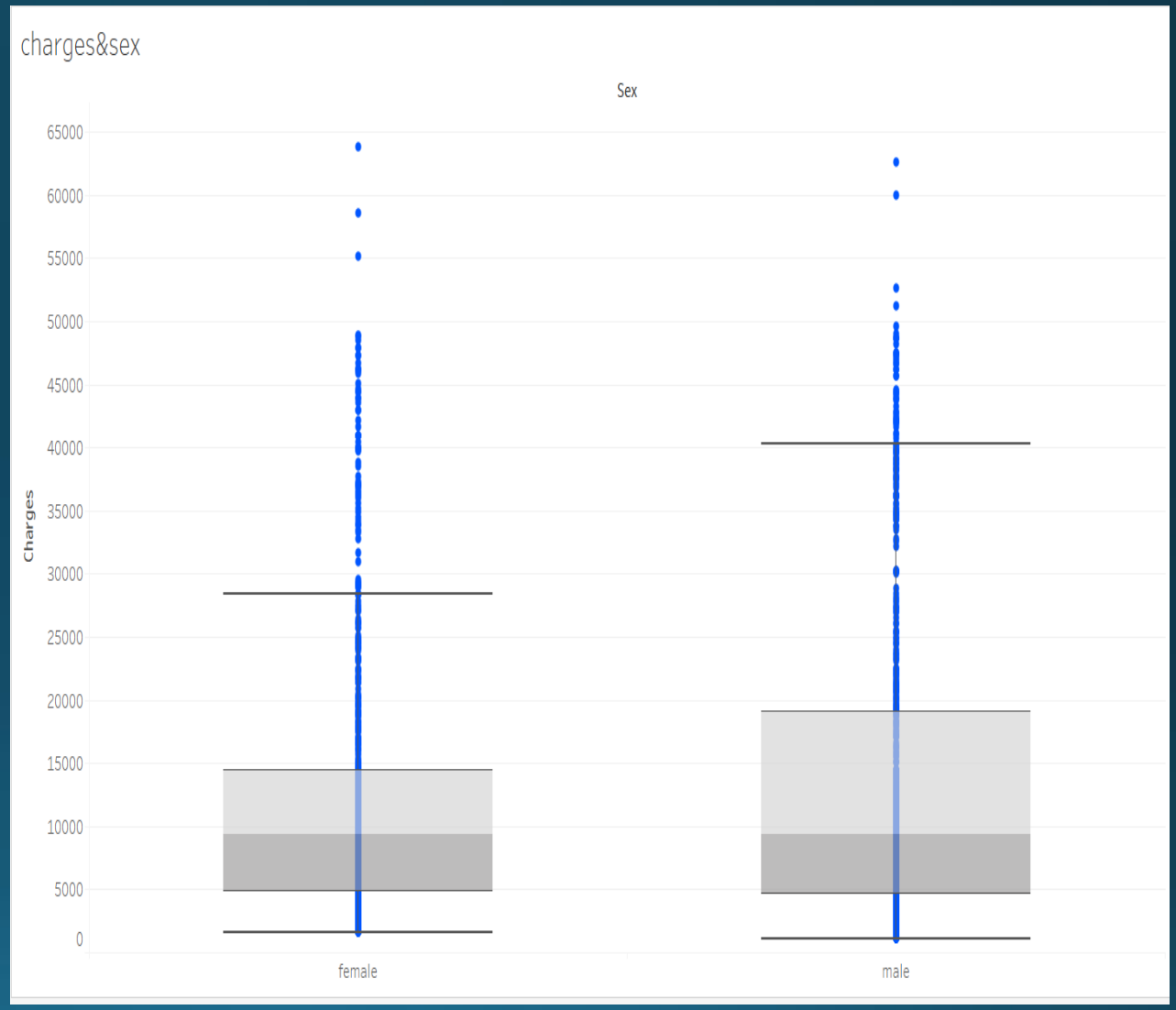
Distribution of the dependent variable: Charges

- Insurance charge (expense) is the dependent variable based upon other variables (predictors).
- From the summary function , we can see that, Median value is quite less than mean value implying right skew in distribution of insurance expenses.
- From the plot ,we see that most clients pay insurance charges of less than USD 25000 while very few pay beyond USD 50000.



Sex:

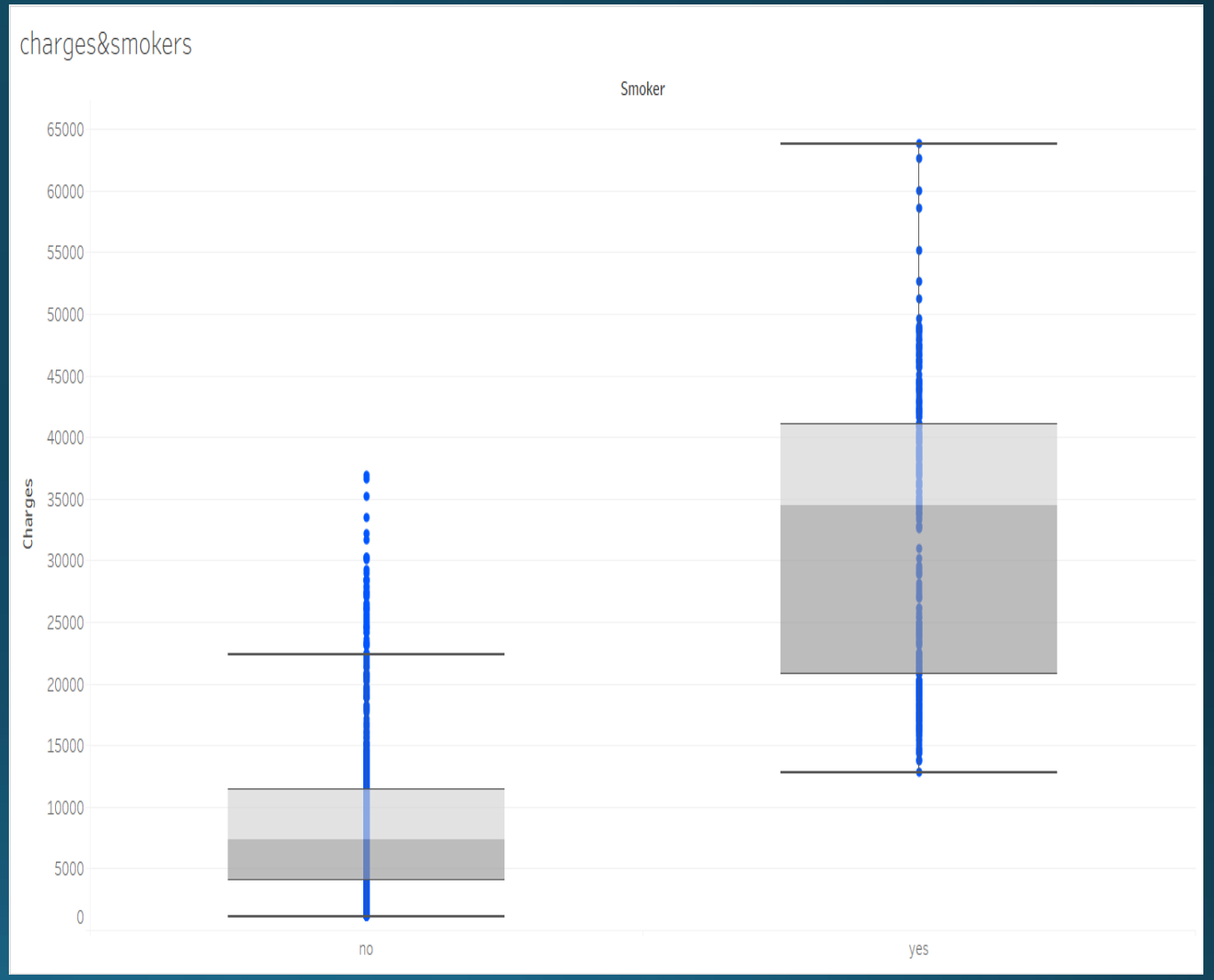
- The variable sex ahs two types. Male-676,female-662.
- The plot shows the boxplot of variable sex for insurance costs. The median costs for both sexes are pretty equal though there is more variance in insurance costs for male.
- Sex do not seem to have much impact on the target.



Smokers:

- Smoker is a categorical variable has yes for smokers and no for non-smokers.

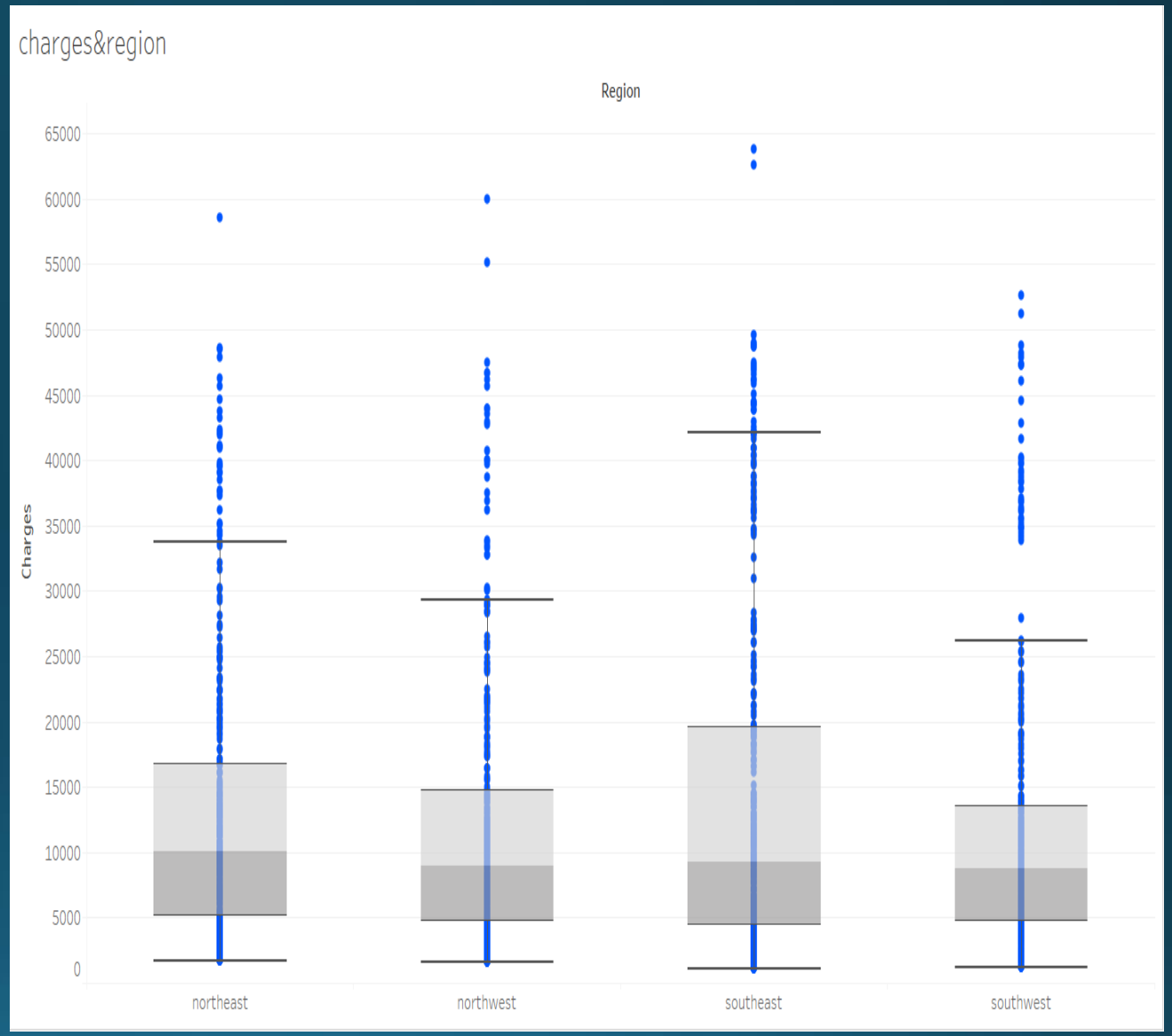
- There’s a clear trend in the plot. Smokers have a much higher median insurance costs in comparison with non-smokers.



Region:

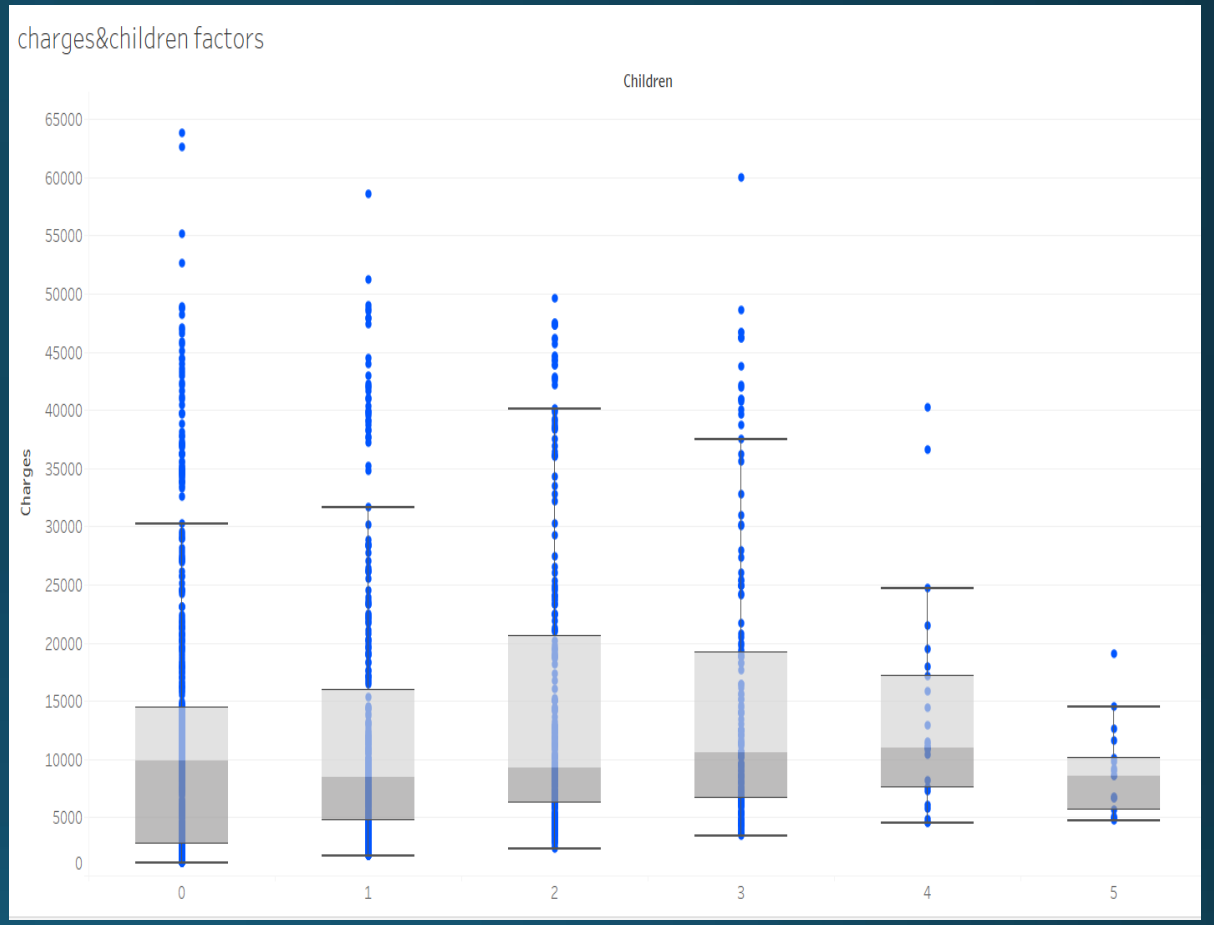
- Region is the categorical variable has 4types.
- Northeast-324
- Northwest-325
- Southeast-324
- Southwest-325

- There’s not a clear trend for variable region in relation with insurance costs. The insurance costs decreases slightly from east to west, however.



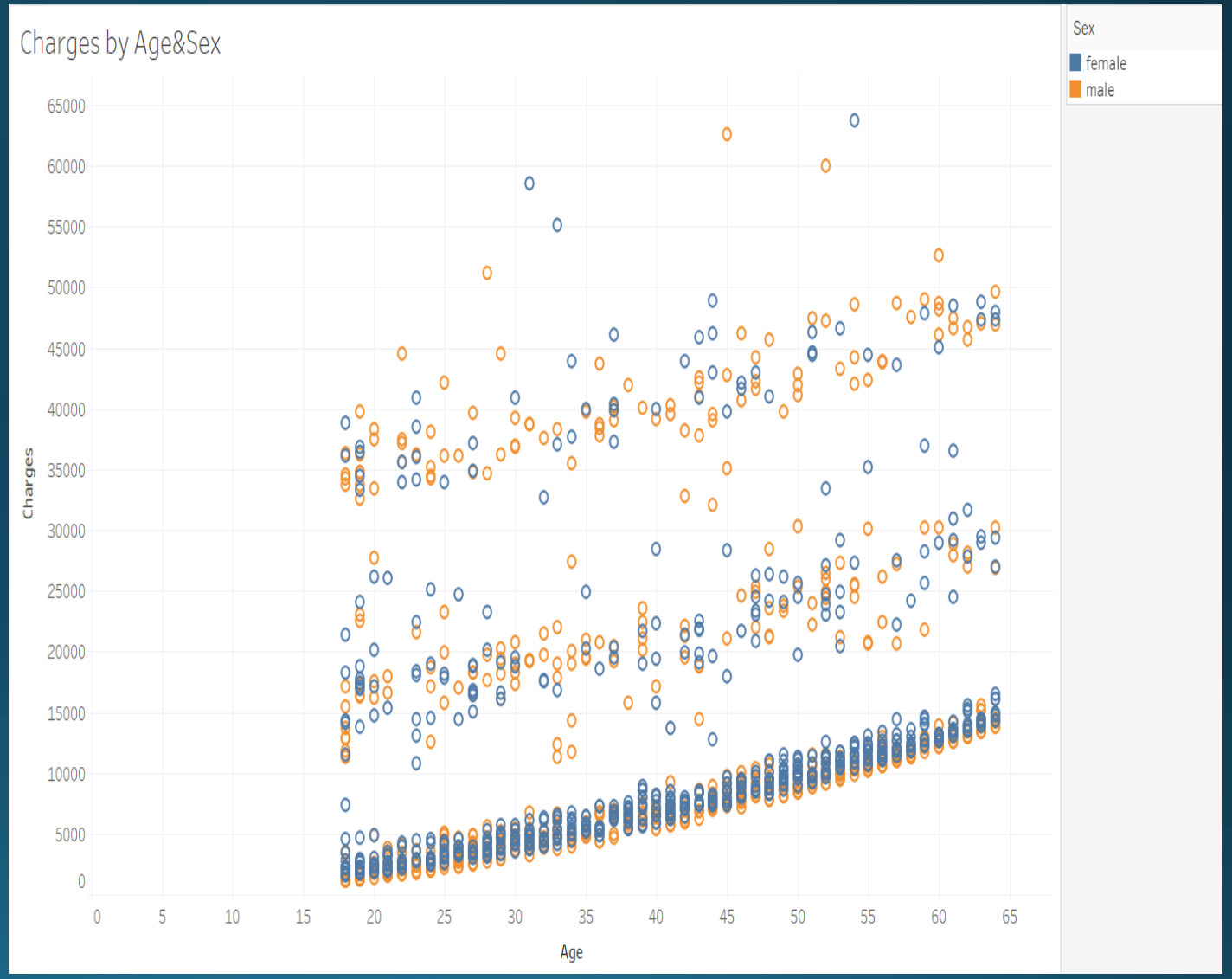
Children:

- The median insurance costs start high for contractors with zero children then goes down for 1 children contractors. The median costs keep increasing but then decreases when a contractor has 5 children. This could be due to the insurance companies policy to start with a high default cost. They give discount for contractors with children at a small rate then give really high discount for contractors with more than 5 children. One thing to note is that the boxplots show there are many outliers.



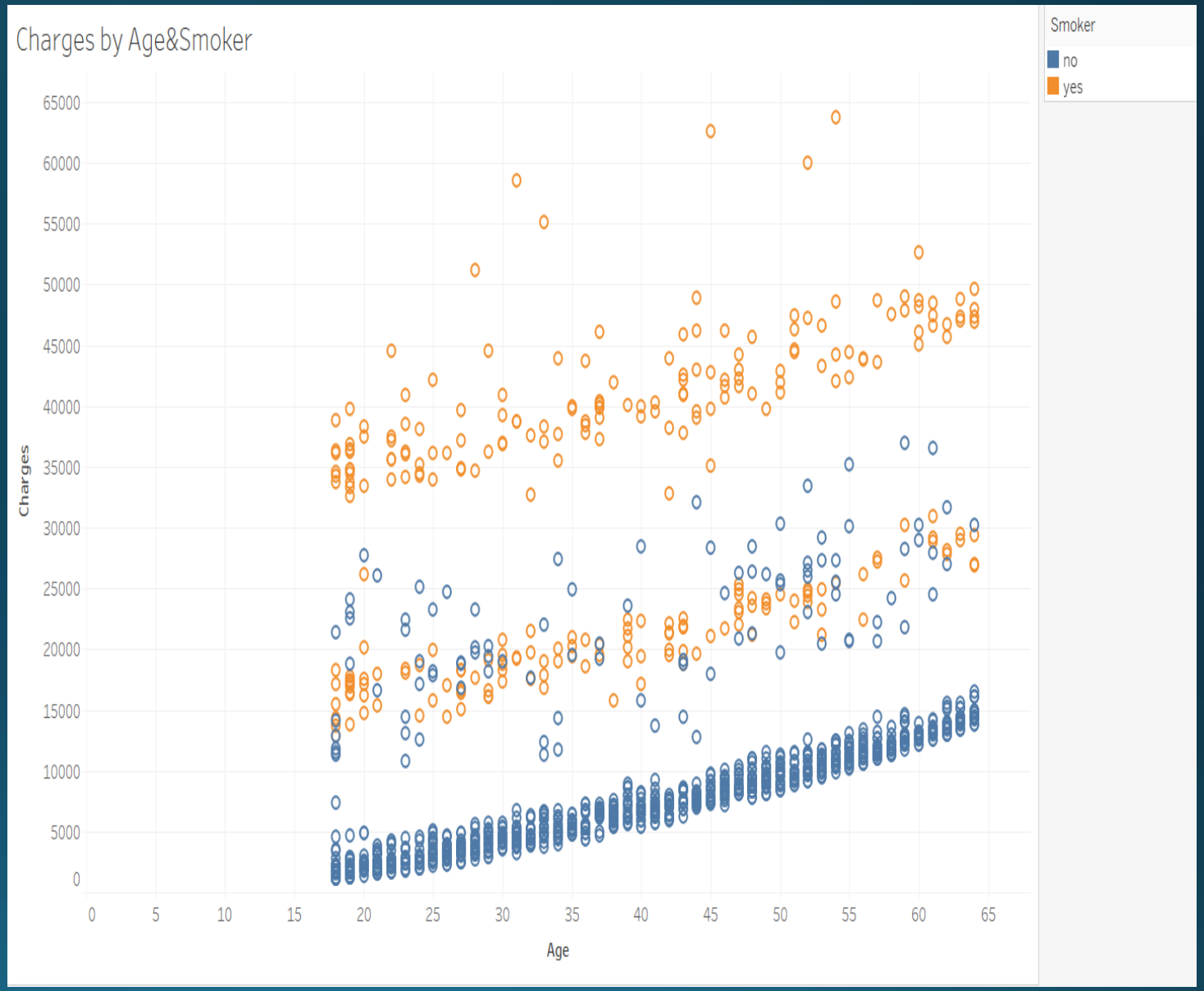
Age:

- The charges definitely do increase with respect to age.
- There is no clear difference in charges for male vs female.



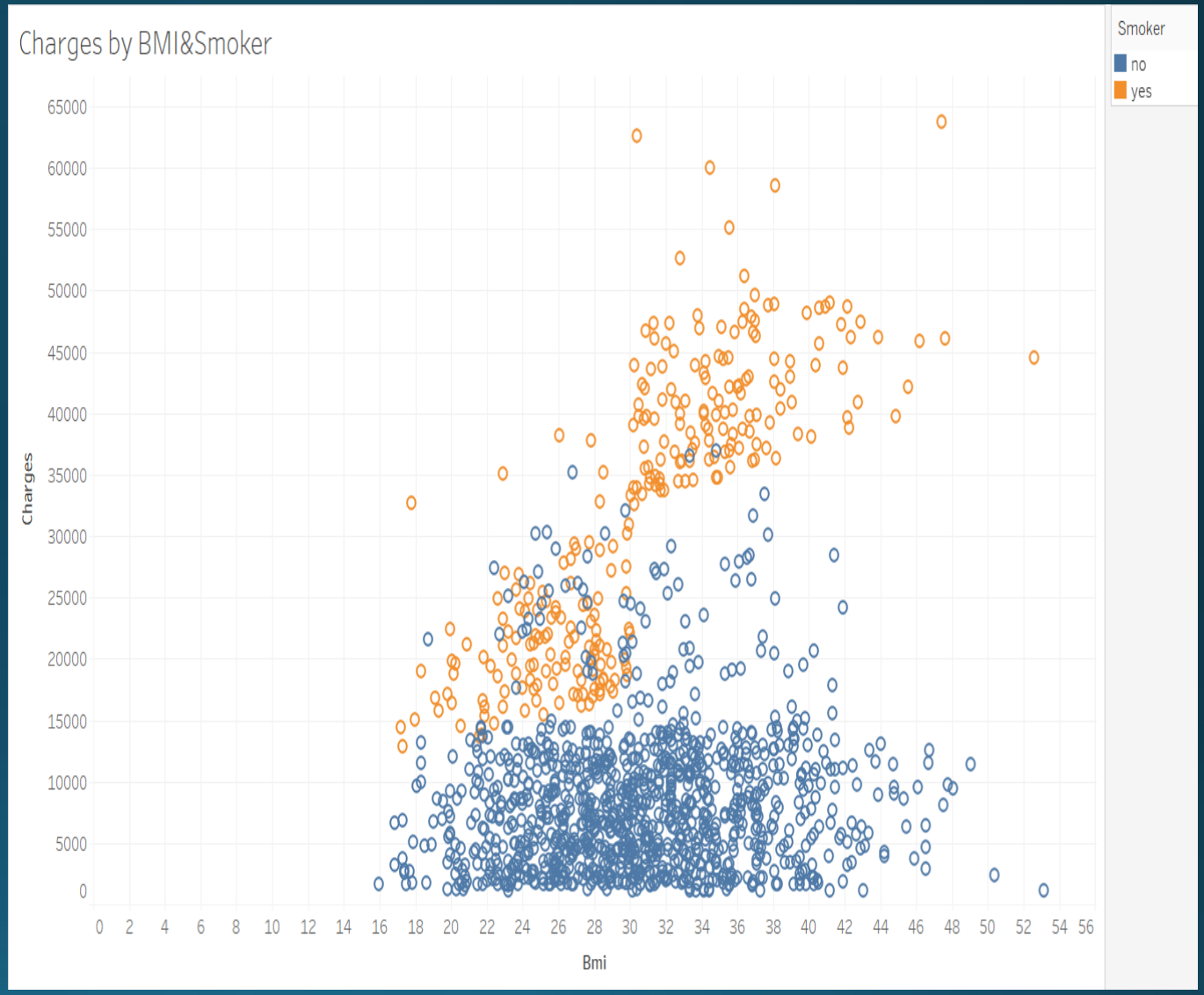
Charges by Age&Smokers:

- Smokers are generally charge a much higher rate.
- Charges above 30000 are usually from smokers and below 15000 are generally non-smokers.
- Anything in between could be from smoker or non-smoker.



BMI:

- Despite the BMI indicator is used to measure health risk for an individual, the feature is not as important as knowing whether the individual is a smoker or non-smoker.
- Smoker tends to incur a much higher charge as compared to non-smoker. When the BMI of a smoker goes beyond 30, the charges increases to a minimum of 30000. Non-smoker with BMI > 30 generally have charges incurred below 30000.



correlation:

- Next, we try to understand the correlation between dependent and explanatory variables. A correlation matrix would be very useful to see the pair wise relationships with numerical variables.

	age	bmi	children	charges
age	1.0000000	0.1092719	0.04246900	0.29900819
bmi	0.1092719	1.0000000	0.01275890	0.19834097
children	0.0424690	0.0127589	1.00000000	0.06799823
charges	0.2990082	0.1983410	0.06799823	1.00000000

- There appears to be moderately positive correlation between age, bmi and insurance charges. It seems that as age/ bmi increases, corresponding insurance charges too see a rise. On a side note, age also appears to be weakly positively correlated to insurance charges.

Model building:

Model 1:

m1 <- lm(formula = charges ~., data = data)

Insights:

- Here, the intercept value does not make much sense as it represents the insurance charges when all other variables are 0 and bmi is not possible.
- The model summary reveals several variables that are insignificant toward predicting the target variable - sexmale, and regionnorthwest.

```
Call:
lm(formula = charges ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- Coefficients of age, bmi, children and smoker are positively related to insurance charges. For each year of increase in age, charges are expected to rise by USD 256.9 given all other variables stay constant.

- Similarly, every unit rise in BMI would hike insurance charges by USD 339.2. Also, an additional child on plan could make insurance dearer by USD 475.5.

- For instance, insurance could be costlier to smoker by USD 23848 in comparison to non-smoker.

- Similarly, it would be cheaper by USD 131 for males than females and USD 353 cheaper to a person from north west than a person from north east US.

- Residual value is the difference between the actual value and the predicted value. Median residual value is USD -982 while maximum is around USD 30000.
- This implies that the model underestimates the insurance charges by around USD 982 for most number of cases. Also, most residual values lie between 1st and 3rd Quartiles i.e. USD -2848 to 1393.
- Sex variable has quite high p-value which implies that is unlikely to be a good predictor of insurance charges.
- The R-Squared value lies between [0, 1] with values nearer to 1 the better. Multiple R-Squared = 0.75 implies that model explains about 75% of the variance in predicted values.
- Adjusted R-Squared values penalize the additional independent variables and thus are a better indicator than R-Squared alone. This value of 0.75 represents that model is significant enough to make good prediction of charges given these explanatory variables.

The regression model developed till now assumed only linear relationship between the variable. However, in real world, scenario can be quite complicated. We need to factor in these complications and bring about subtle modifications in the model.

Transforming BMI value to categorical value:

- It is observed that, insurance charges don't vary much for people with average or slightly above average BMI. But, it can really take off for obese and morbidly obese people. We need to consider this logic and add a categorical variable that would indicate if BMI is above a certain high value say, 30.

data$bmi30 <- ifelse(data$bmi > 30, 1, 0)

Interactive effects of smoking and obesity:

- It is regarded that smoking and obesity are both individually detrimental to health. However, we suspect that their combined effect would potentially be more harmful and thus affect the insurance charges. We, therefore, consider an interaction of these two variables in the model (Smoker * BMI30).
- The remodel regression and the summary of the results is,

m1_updated <- lm(charges ~ age + children + bmi + sex + bmi30 + bmi30*smoker + region, data = data)

- An extremely low p-value corresponding to newly included variables like bmi30, interaction between smoking and obesity, indicate that these variables are statistically significant and more likely to affect the insurance charges.

- Adjusted R-Squared has risen from 75% to 86%.

- Residual standard error has fallen down from 6062 to 4457.

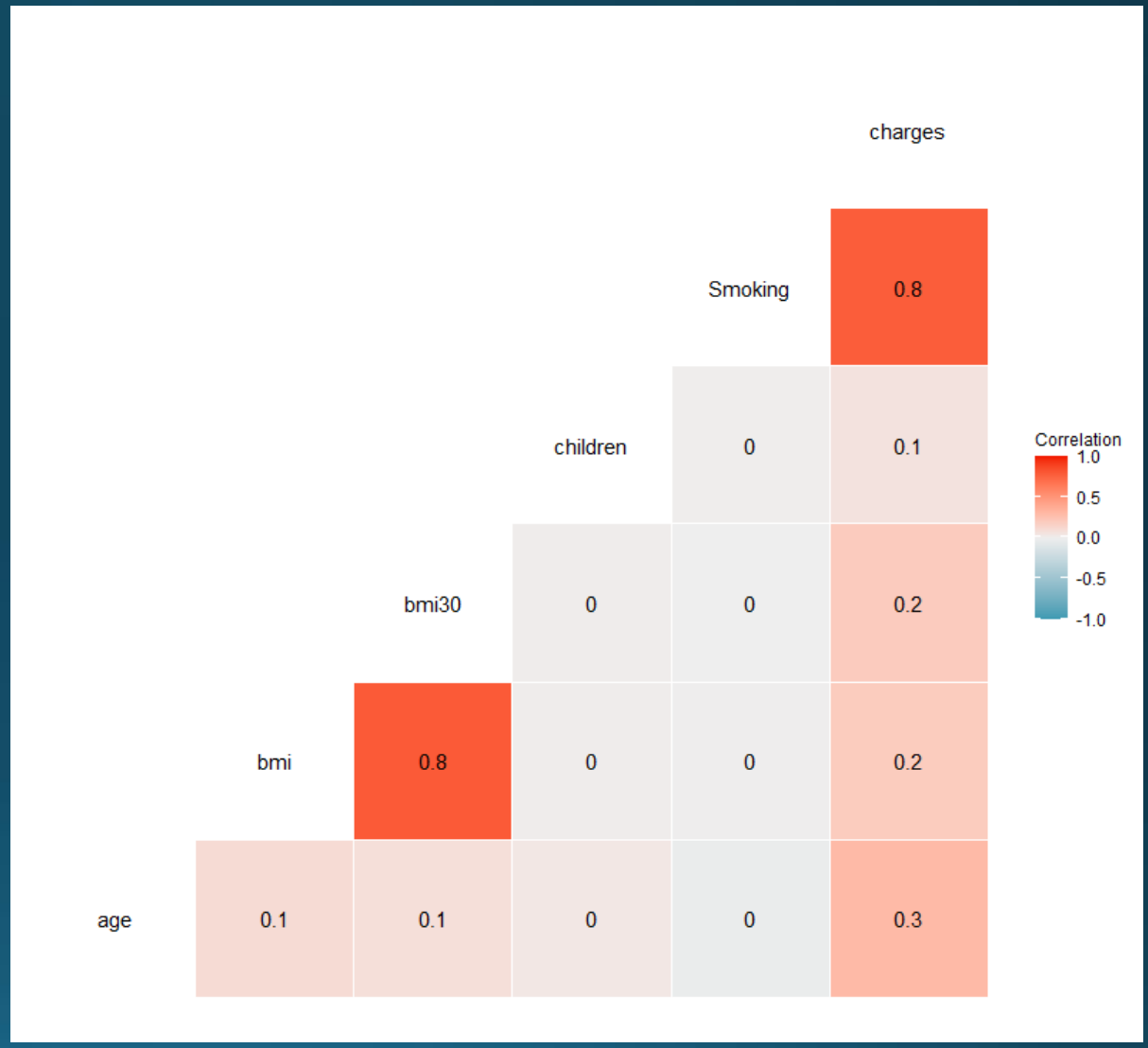
```
Call:
lm(formula = charges ~ age + children + bmi + sex + bmi30 + bmi30 *
    smoker + region, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4081.3 -1830.2 -1263.2  -464.7 24813.0

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4622.33     953.77  -4.846 1.41e-06 ***
age                263.64       8.75  30.130  < 2e-16 ***
children           508.97     101.31   5.024 5.76e-07 ***
bmi                108.96      34.35   3.172 0.001549 **
sexmale           -470.14     244.98  -1.919 0.055185 .
bmi30             -803.06     423.19  -1.898 0.057964 .
smokeryes        13413.21     439.59  30.513  < 2e-16 ***
regionnorthwest   -263.74     350.20  -0.753 0.451514
regionsoutheast   -822.28     352.57  -2.332 0.019837 *
regionsouthwest  -1165.72     351.45  -3.317 0.000935 ***
bmi30:smokeryes 19909.67     605.92  32.859  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4457 on 1327 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8646
F-statistic: 854.5 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Correlation between the newly added variables:

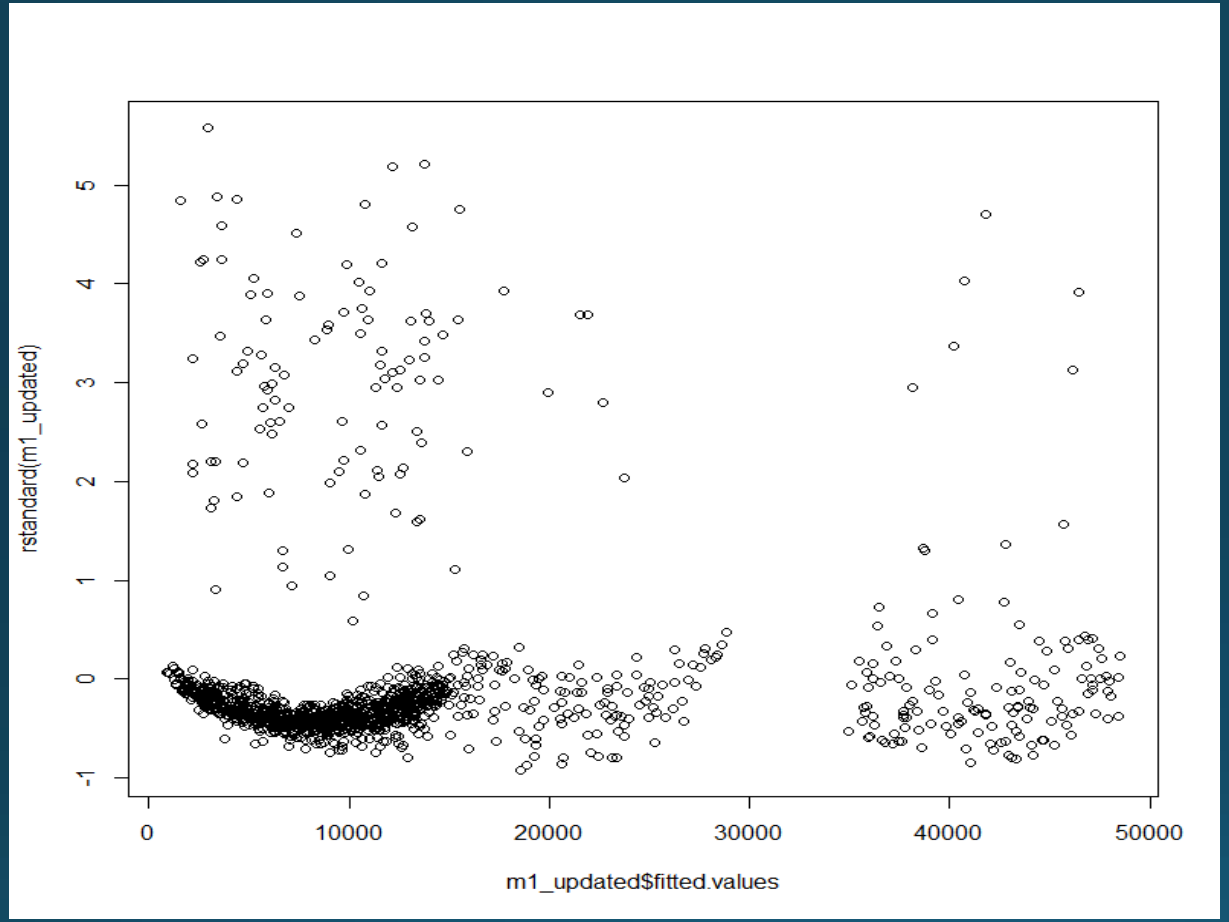
- Let’s add another variable smoking(smoker=“yes”) and see the correlation.

- In the improved model, smoking, bmi30 all seem to be positively correlated with the other existing variables.

- Our model is good enough.



standardised residual plot for the Model:



- Residuals is the difference between actual data and the prediction.
- Not all but most of the values are close to 0 which means the actual and the predicted values are closer.

Thank you!