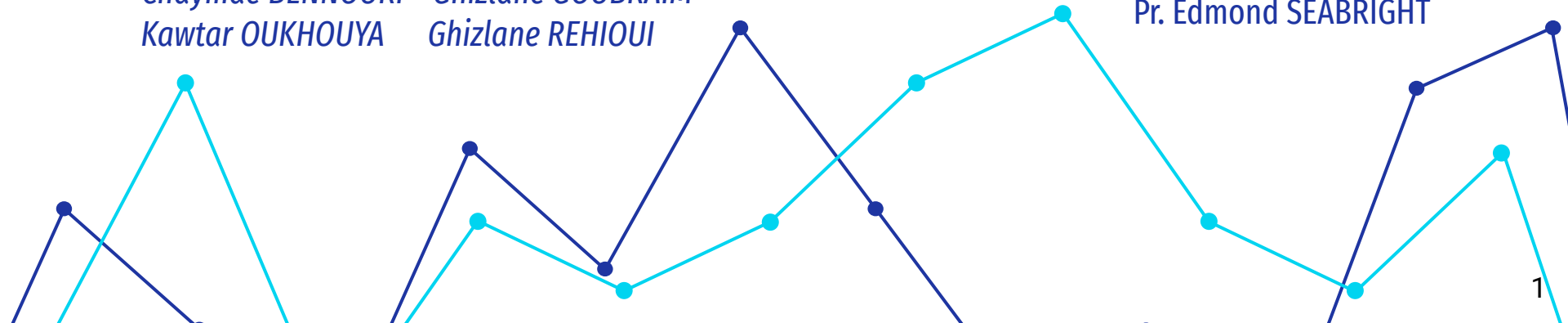
The effect of smoking, bmi, age and their interactions on the insurance charges

By:

Chaymae BENNOURI *Ghizlane GOUBRAIM*
Kawtar OUKHOUYA *Ghizlane REHIOUI*

Professor:
Pr. Edmond SEABRIGHT



Context

Context

Ahmed claims that his insurance company is unfair and he decides to sue them.

Research questions

What variables affect insurance charges more? Is it the region, age, children, smoking, gender or the BMI?

What is the effect of the interactions between region, age, children, smoking, gender and the BMI on the insurance charges?

Dataset variables

01

Age

age of primary beneficiary

02

Gender

insurance contractor gender(female, male)

03

BMI

Body mass index

04

Children

Number of children covered by health insurance

05

Smoker

Smoking habit

06

Region

the beneficiary's residential area in the US

07

Charges

Individual medical costs billed by health insurance

Dataset variables

Our data is from **Kaggle** on an **unknown US** insurance company.
Our data frame is composed of 7 columns and 1338 observations (rows).

	age	gender	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622

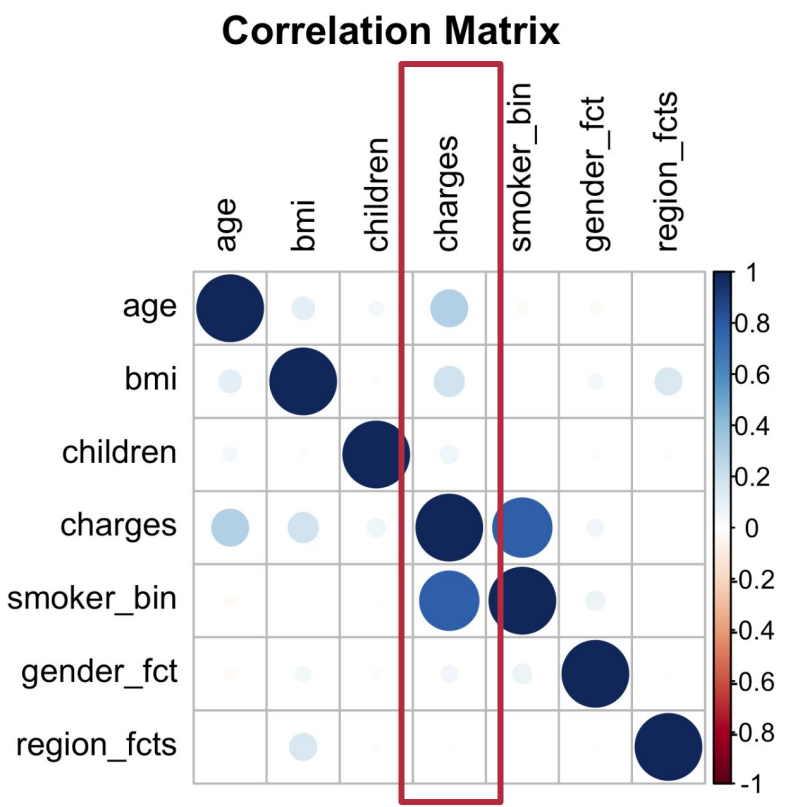
Transformations

Before we start our analysis, we will do some transformations and new variables:

- ***smoker*** → ***smoker_bin :*** factor variable (0 for no / 1 for yes)
- ***gender*** → ***gender_fct :*** factor variable (0 for female / 1 for male)
- ***region*** → ***region_fcts :*** factor variable where Northeast is 1, Northwest is 2, Southeast is 3 and Southwest is 4.
- ***age_centered :*** Centered the age on the minimum age of data (18 y.o)
- ***bmi_centered :*** Centered bmi around the mean BMI of the data (30.66 kg/m^2)
- ***bmi_cat :*** a new integer variable BMI category (-1 for underweight [bmi <= 18.5], 0 for healthy [bmi between 30 and 18.5], +1 for overweight and obese [bmi >= 30])

NOTE: *The bmi_cat was based on the Center for Disease Control and Prevention's categorization*

Correlation between variables



Highest correlation for *charges* with:

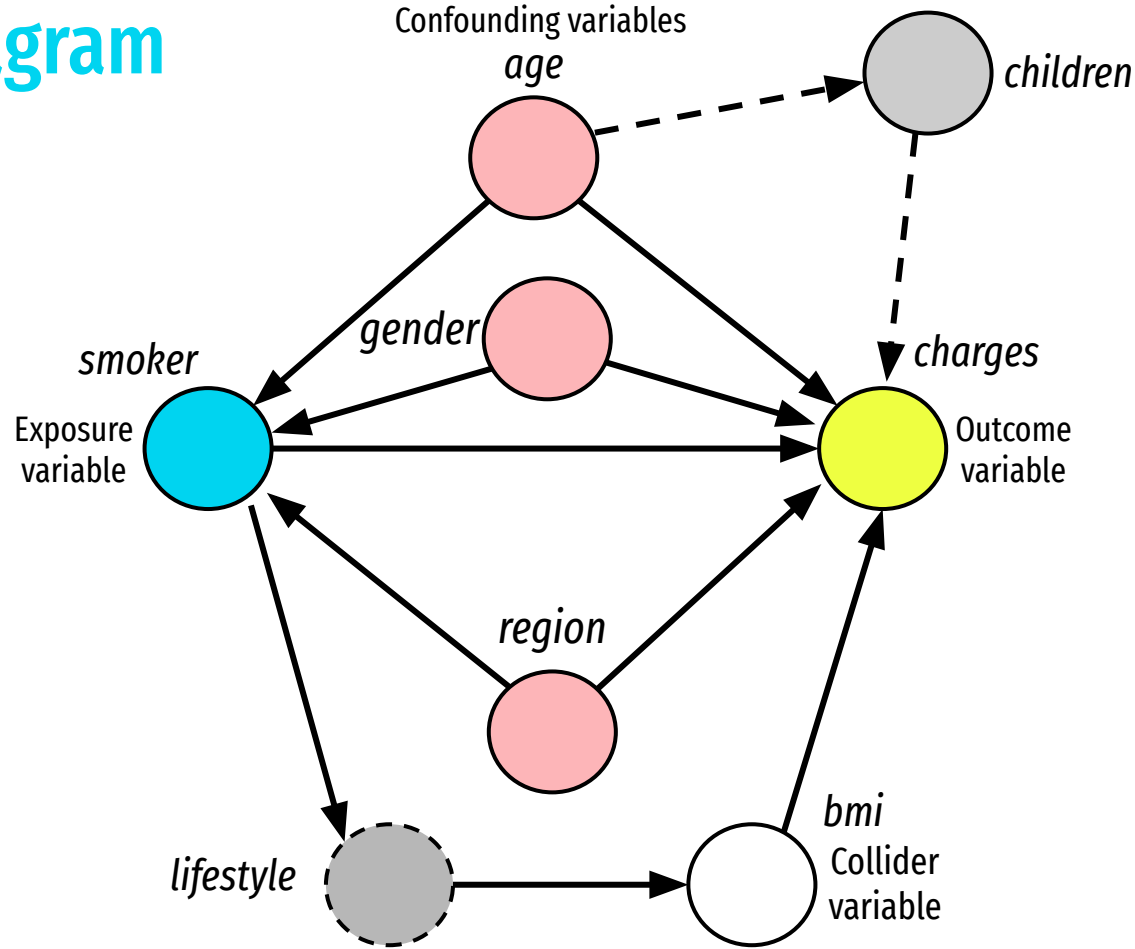
- *smoker_bin : 0.7872*
- *age : 0.2990*
- *bmi : 0.1983*

Least correlation for *charges* with:

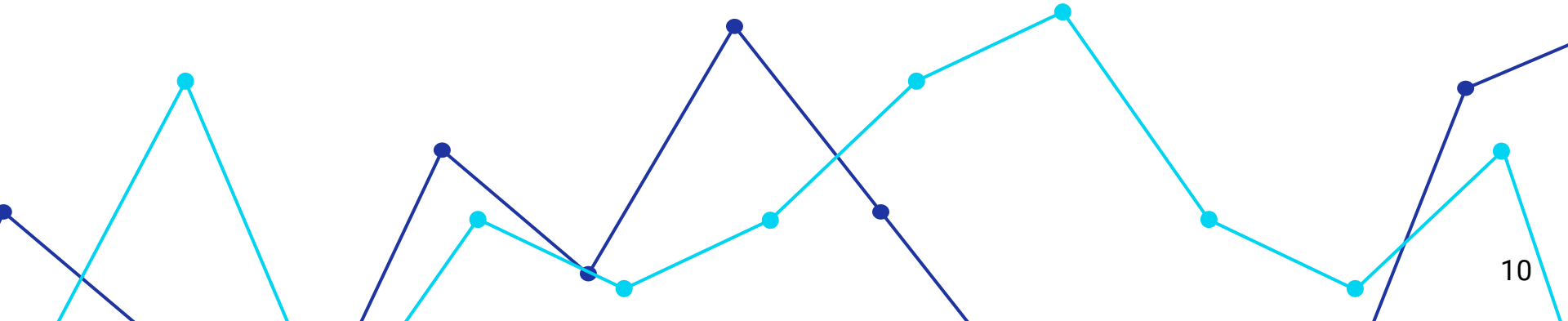
- *children : 0.0679*
- *gender_fct : 0.0573*
- *region_fcts : -0.006208*

We also check for multicollinearity between the independent variables: no correlation > 0.9
→ No risk of multicollinearity!

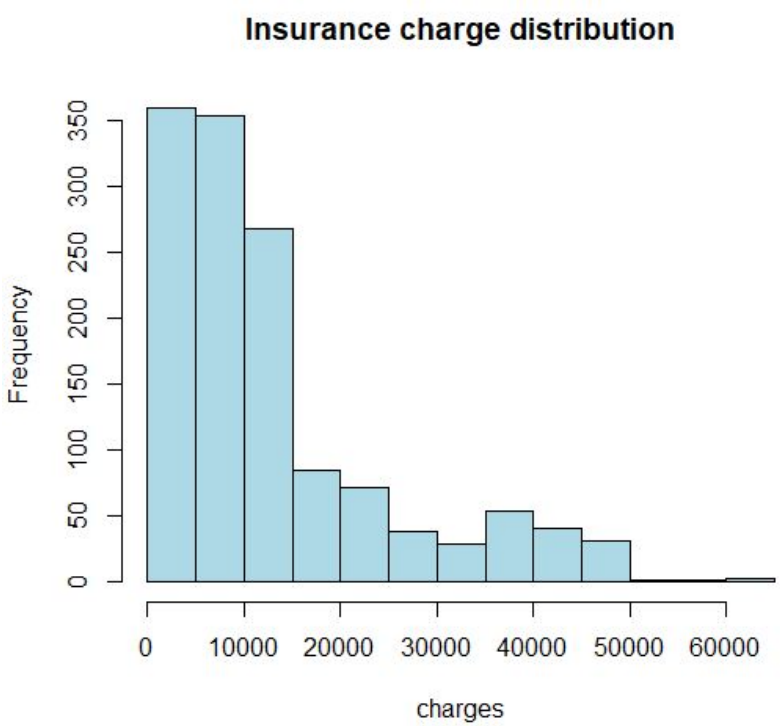
Causal diagram

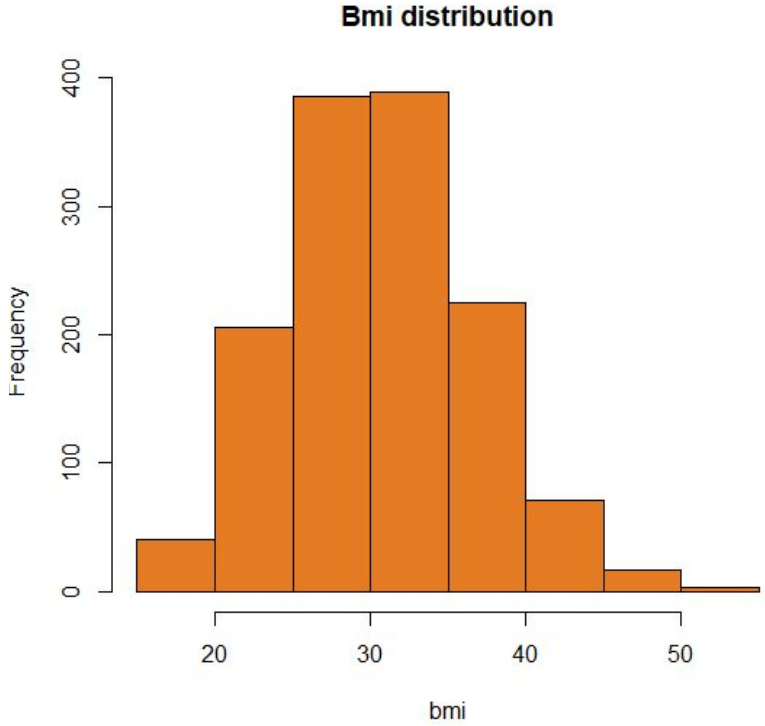


Exploratory Data Analysis

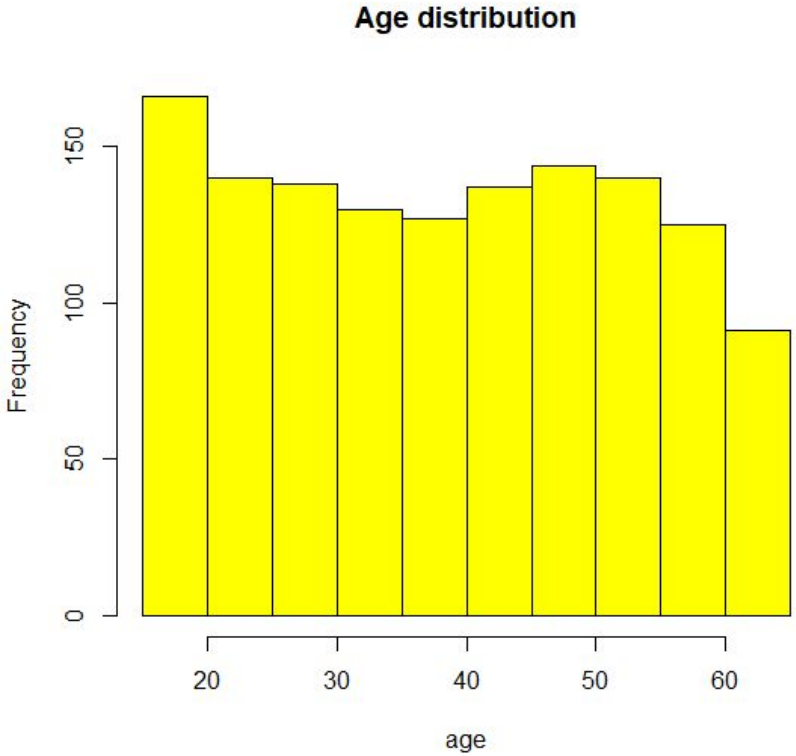


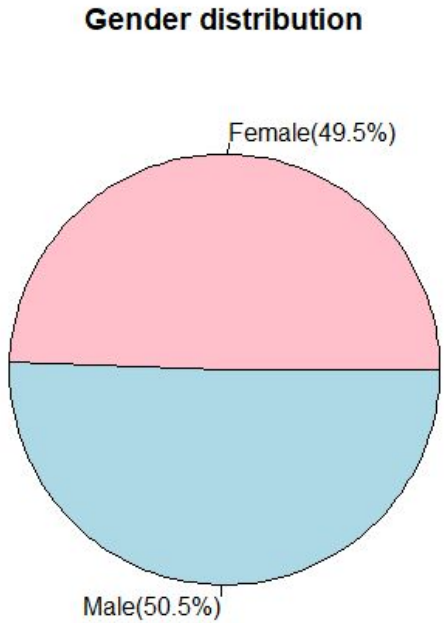
EDA - PART 1


Insurance charge distribution
Frequency
0
50
100
150
200
250
300
350
0
10000
20000
30000
40000
50000
60000
charges

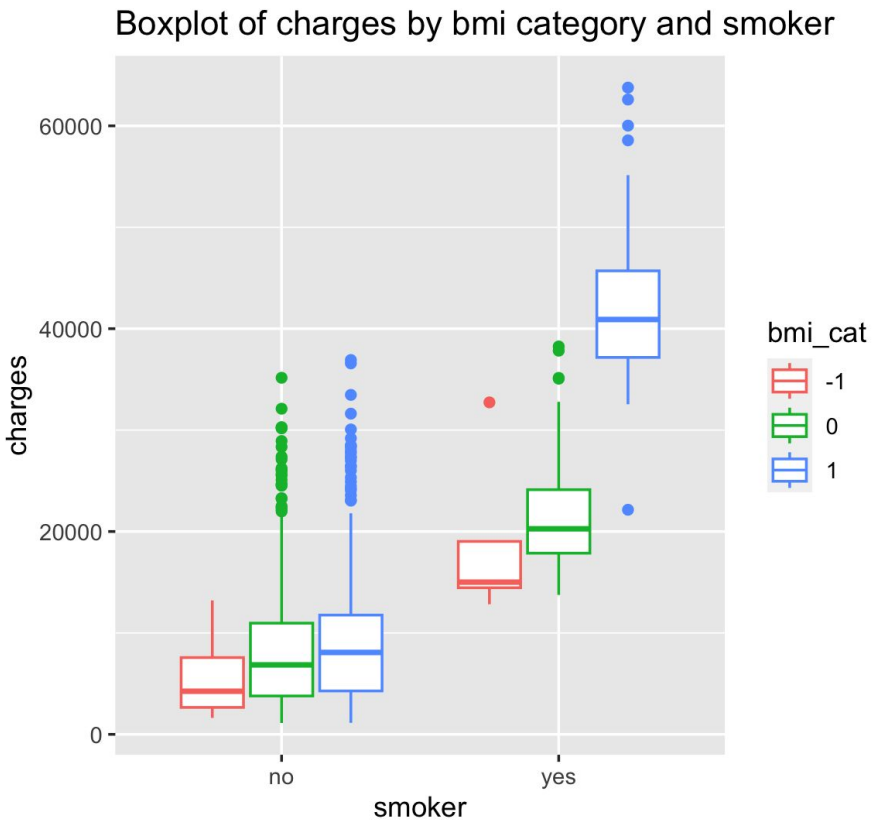

Bmi distribution
Frequency
0
100
200
300
400
20
30
40
50
bmi

EDA - PART 1


Age distribution
Frequency
0
50
100
150
20
30
40
50
60
age


Gender distribution
Female(49.5%)
Male(50.5%)

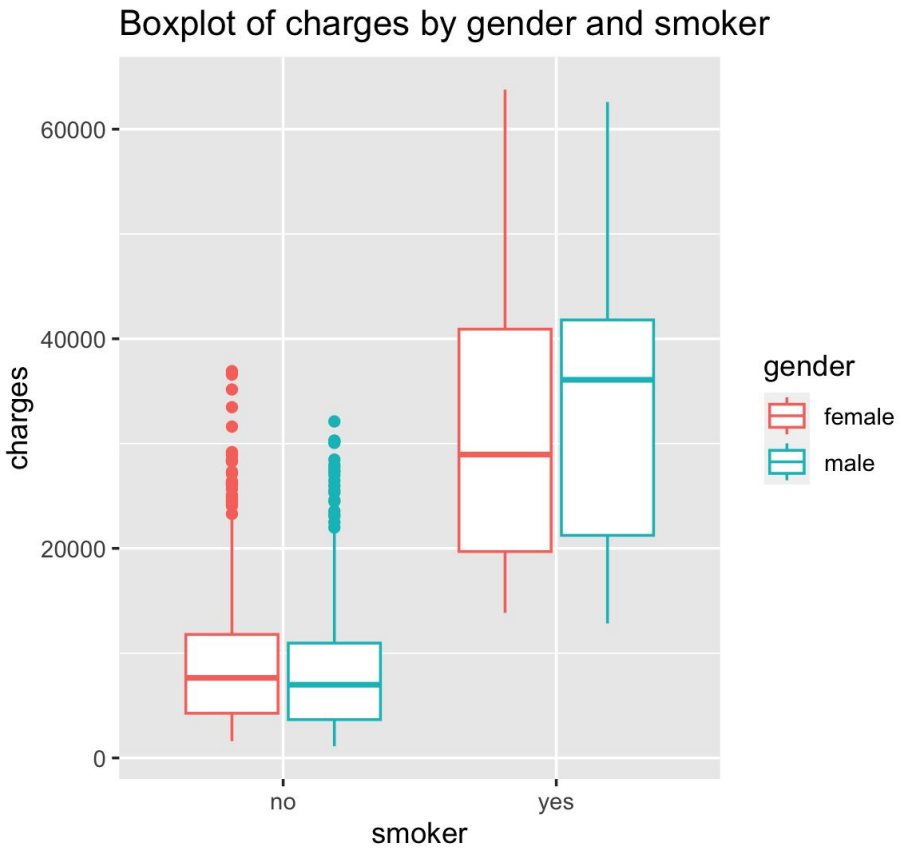
EDA - PART 2



- Smokers generally have higher charges
- Smoker + overweight -> higher charges
- Underweight + smoker > underweight + non smoker
- Normally, we expect the healthy people to have the least insurance charges.

➔ There is interaction between smokers and bmi_cat.

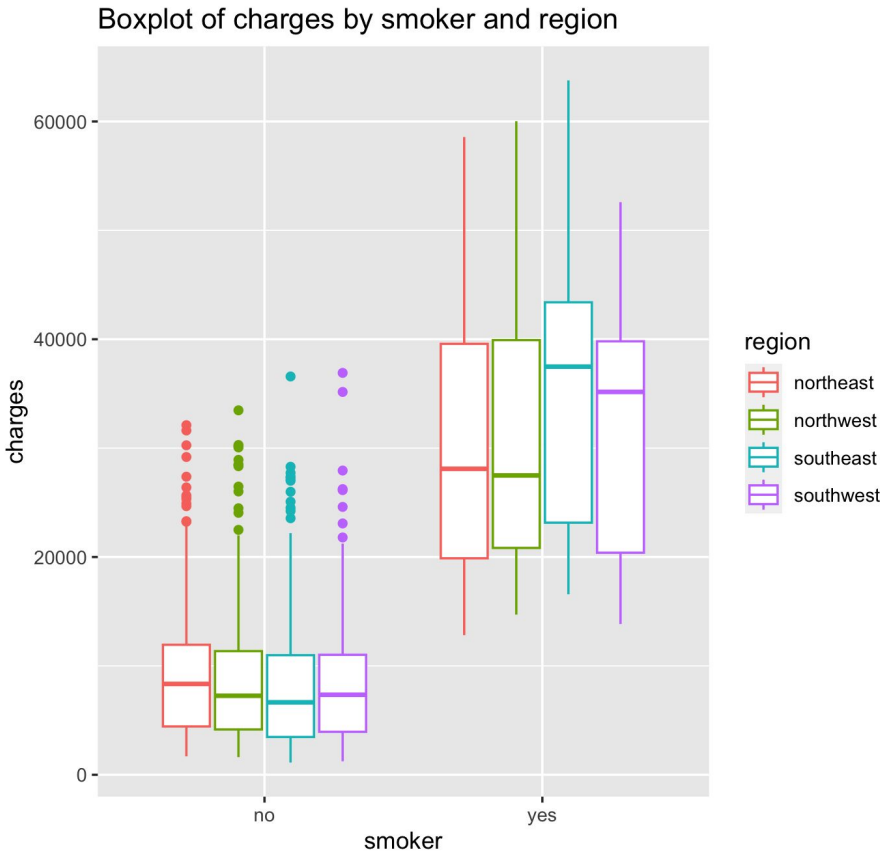
EDA - PART 2



- Smokers have higher insurance charges regardless of the gender
- Male smokers have higher charges than female smokers
- Non smoker's charges are slightly equal between Females and Males

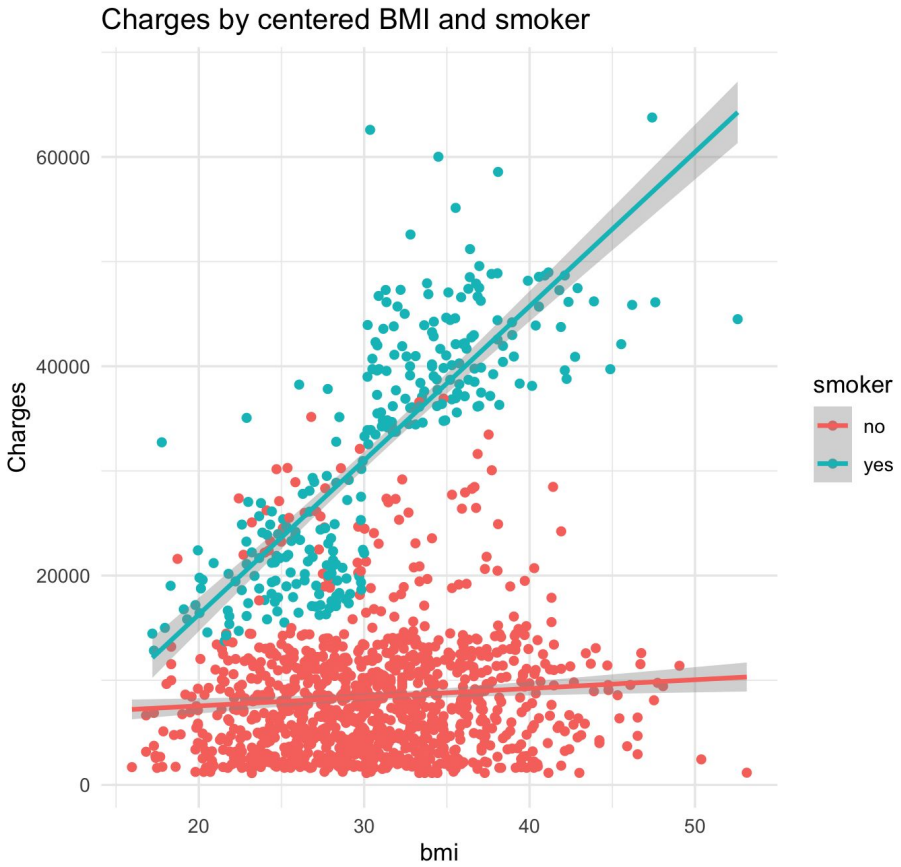
➔ There is interaction between smokers and gender

EDA - PART 2



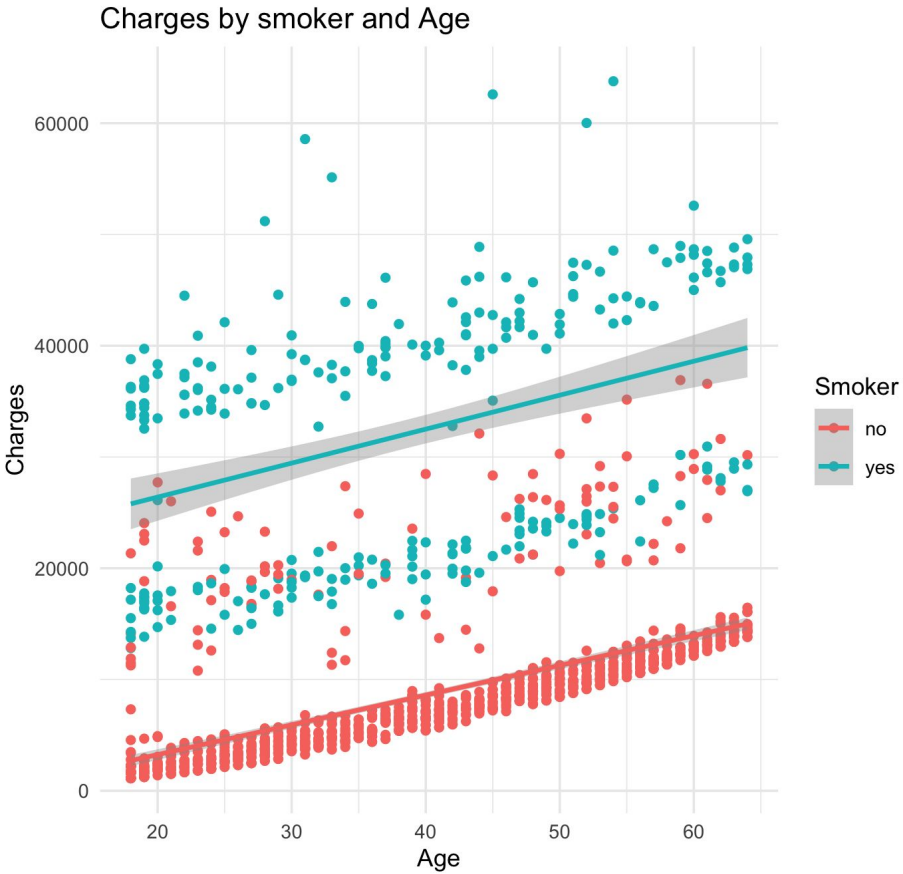
- No smokers have approximately the same charges in different regions
- Smoking in the south regions is more expensive than the north regions

EDA - PART 2



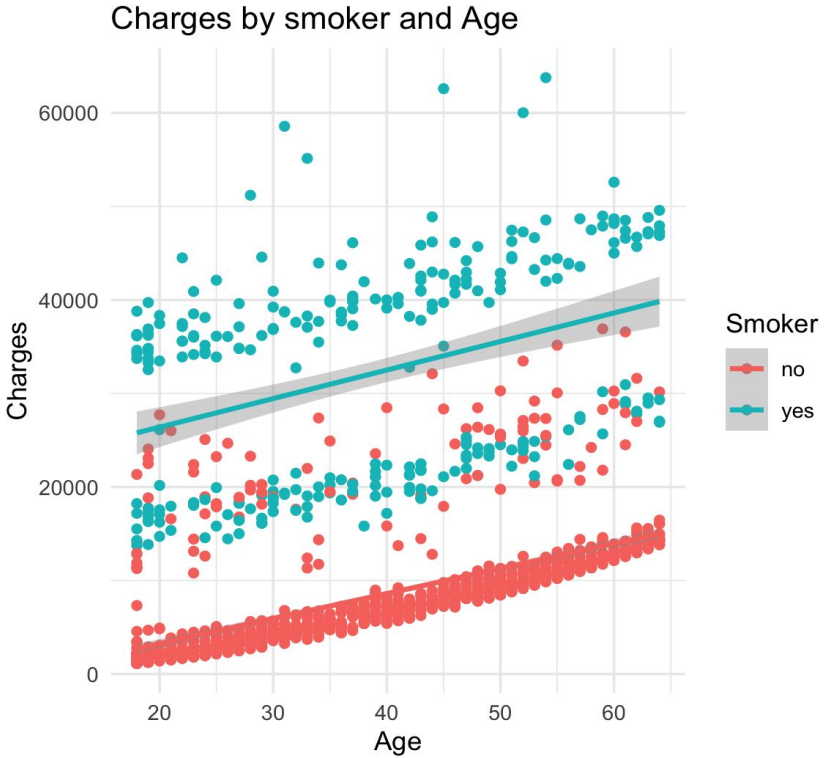
- Charges increase faster as bmi increases for a smoker than for a non smoker.
- In general, there is a linear relationship between smoker and charges.
- This shows that there is probably an interaction between bmi and smoking that affects the charges.

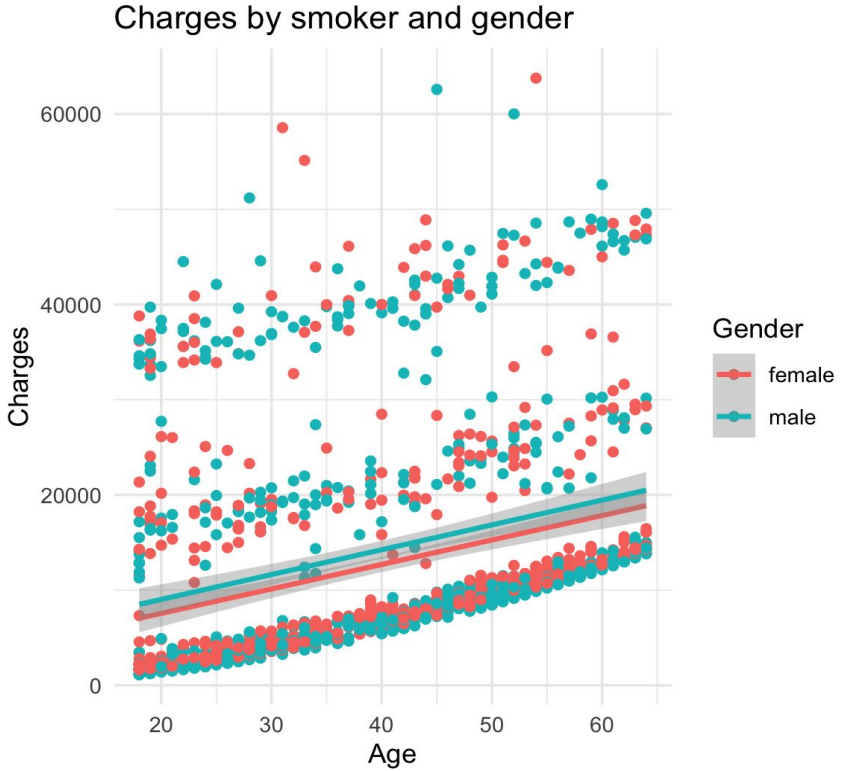
EDA - PART 2



- Generally, smokers have higher insurance charge.
- Insurance charges increase as the age increases as well.
- **Special cases :** There is some cases when the non smokers have high insurance charge, and the smokers have less insurance charges.
- This shows that there are other variables that interact with smoking habit to affect the charges.

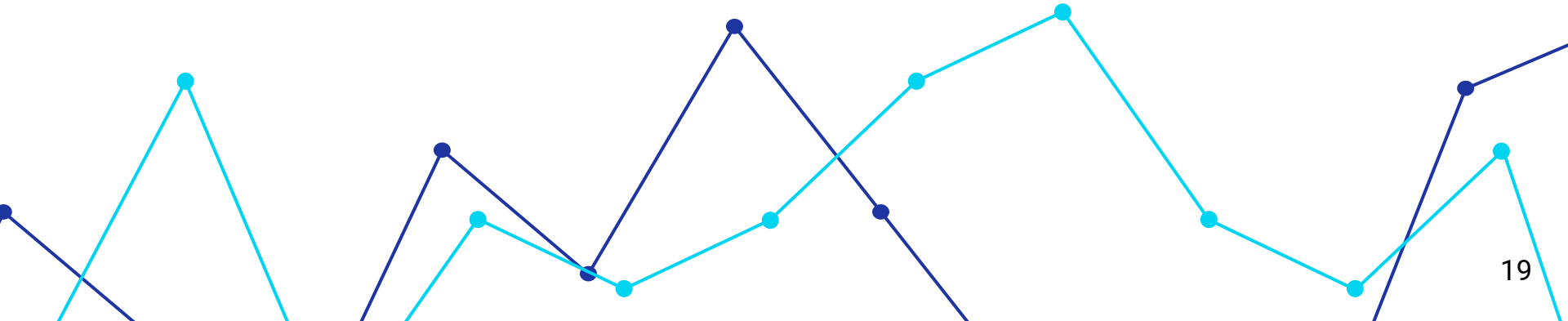
EDA - PART 2





→ There is no visually clear interaction between gender and age

Models



MODEL 1: Multiple linear Regression without Interactions

```
glm(formula = insurance$charges ~ insurance$children + insurance$smoker +
    insurance$gender + insurance$bmi_centered + insurance$age_centered +
    insurance$region)

Deviance 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)                3085.7      485.3   6.359 2.79e-10 ***
insurance$children          475.5      137.8   3.451 0.000577 ***
insurance$smokeryes       23848.5      413.1  57.723  < 2e-16 ***
insurance$gendermale       -131.3      332.9  -0.394 0.693348
insurance$bmi_centered      339.2       28.6  11.860  < 2e-16 ***
insurance$age_centered      256.9       11.9  21.587  < 2e-16 ***
insurance$regionnorthwest  -353.0      476.3  -0.741 0.458769
insurance$regionsoutheast -1035.0      478.7  -2.162 0.030782 *
insurance$regionsouthwest  -960.0      477.9  -2.009 0.044765 *
---
```

Model 1: Including all the variables & without interactions

- A childless female of age 18 with an average bmi, doesn't smoke and lives in the northeast, pays 3,085.7 $.

MODEL 1: Multiple linear Regression without Interactions

```
glm(formula = insurance$charges ~ insurance$children + insurance$smoker +
    insurance$gender + insurance$bmi_centered + insurance$age_centered +
    insurance$region)

Deviance 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)                3085.7      485.3   6.359 2.79e-10 ***
insurance$children          475.5      137.8   3.451 0.000577 ***
insurance$smokeryes       23848.5      413.1  57.723  < 2e-16 ***
insurance$gendermale       -131.3      332.9  -0.394 0.693348
insurance$bmi_centered      339.2       28.6  11.860  < 2e-16 ***
insurance$age_centered      256.9       11.9  21.587  < 2e-16 ***
insurance$regionnorthwest  -353.0      476.3  -0.741 0.458769
insurance$regionsoutheast -1035.0      478.7  -2.162 0.030782 *
insurance$regionsouthwest  -960.0      477.9  -2.009 0.044765 *
---
```

Model 1: Including all the variables & without interactions

- For every child, this female pays an additional amount of 475.5$

MODEL 1: Multiple linear Regression without Interactions

```
glm(formula = insurance$charges ~ insurance$children + insurance$smoker +
    insurance$gender + insurance$bmi_centered + insurance$age_centered +
    insurance$region)

Deviance 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)                3085.7      485.3   6.359 2.79e-10 ***
insurance$children          475.5      137.8   3.451 0.000577 ***
insurance$smokeryes       23848.5      413.1  57.723  < 2e-16 ***
insurance$gendermale       -131.3      332.9  -0.394 0.693348
insurance$bmi_centered      339.2       28.6  11.860  < 2e-16 ***
insurance$age_centered      256.9       11.9  21.587  < 2e-16 ***
insurance$regionnorthwest  -353.0      476.3  -0.741 0.458769
insurance$regionsoutheast -1035.0      478.7  -2.162 0.030782 *
insurance$regionsouthwest  -960.0      477.9  -2.009 0.044765 *
---
```

Model 1: Including all the variables & without interactions

- For a childless male of age 18 with an average bmi, doesn't smoke and lives in the northeast, pays 3,085.7 - 131.3 = 2,954.4 $

MODEL 1: Multiple linear Regression without Interactions

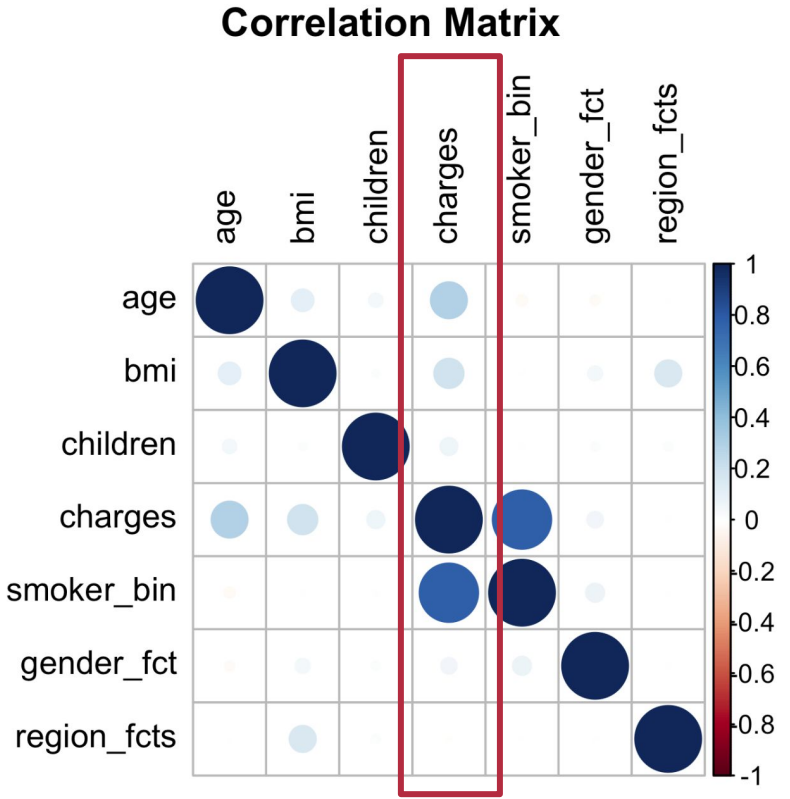
```
(Dispersion parameter for gaussian family taken to be 36749084)

    Null deviance: 1.9607e+11  on 1337  degrees of freedom
Residual deviance: 4.8840e+10  on 1329  degrees of freedom
AIC: 27116

Number of Fisher Scoring iterations: 2
```

Highest correlation for *charges* with:

- *smoker_bin : 0.7872*
- *age : 0.2990*
- *bmi : 0.1983*

Least correlation for *charges* with:

- *children : 0.0679*
- *gender_fct : 0.0573*
- *region_fcts : -0.006208*



MODEL 2: Multiple linear Regression without Interactions

```
glm(formula = insurance$charges ~ insurance$smoker + insurance$bmi_centered +
    insurance$age_centered)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12415.4   -2970.9    -980.5    1480.0   28971.8

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             2887.50     316.29   9.129   <2e-16 ***
insurance$smokeryes    23823.68     412.87  57.703   <2e-16 ***
insurance$bmi_centered   322.62      27.49  11.737   <2e-16 ***
insurance$age_centered   259.55      11.93  21.748   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37116356)

    Null deviance: 1.9607e+11  on 1337  degrees of freedom
Residual deviance: 4.9513e+10  on 1334  degrees of freedom
AIC: 27124

Number of Fisher Scoring iterations: 2
```

Model 2: using most correlated variables with our outcome variable

- A non smoker person with an average bmi and age of 18 pays: 2,887.50$

MODEL 2: Multiple linear Regression without Interactions

```
glm(formula = insurance$charges ~ insurance$smoker + insurance$bmi_centered +
    insurance$age_centered)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12415.4   -2970.9    -980.5    1480.0   28971.8

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              2887.50     316.29   9.129   <2e-16 ***
insurance$smokeryes     23823.68     412.87  57.703   <2e-16 ***
insurance$bmi_centered    322.62      27.49  11.737   <2e-16 ***
insurance$age_centered    259.55      11.93  21.748   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37116356)

    Null deviance: 1.9607e+11  on 1337  degrees of freedom
Residual deviance: 4.9513e+10  on 1334  degrees of freedom
AIC: 27124

Number of Fisher Scoring iterations: 2
```

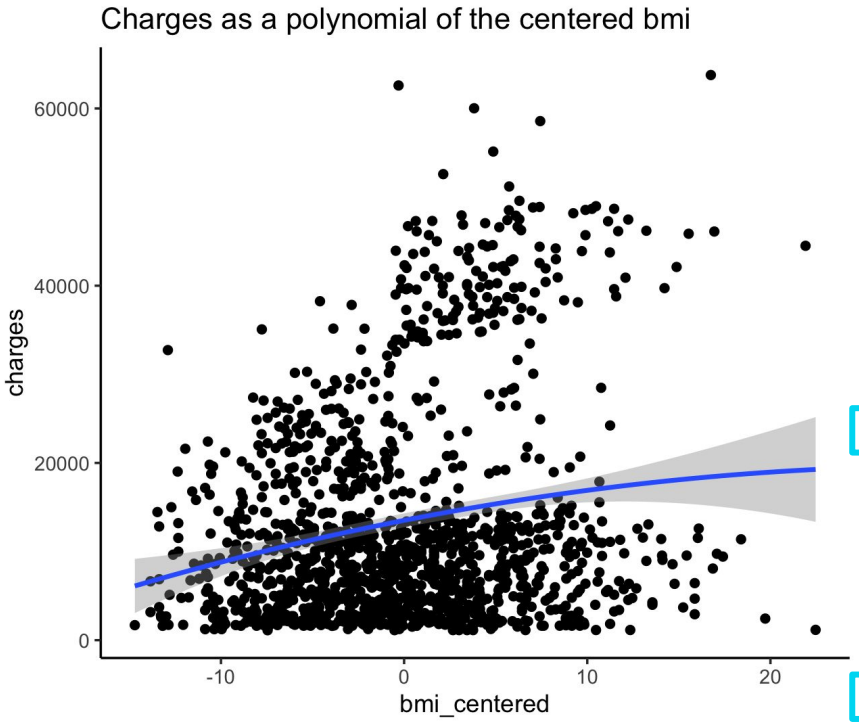
Model 2: using most correlated variables with our outcome variable

- A smoker person with an average bmi and age of 18 pays: 2,887.50+23,823.68=26,711.18$

Approximately the same AIC with model 1

MODEL 3: Non linear regression: Polynomial

We anticipate that charges could be best represented by a polynomial function of the bmi.




```
# Representing charges as a polynomial of bmi
ply_chr_bmi<- glm(insurance$charges ~ insurance$bmi_centered
                  + I(insurance$bmi_centered^2))
summary(ply_chr_bmi)
```

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-18100   -8152   -3847    4843   49198

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 13517.995    402.341  33.598  < 2e-16 ***
insurance$bmi_centered        405.396     54.386   7.454 1.62e-13 ***
I(insurance$bmi_centered^2)    -6.662      6.397  -1.041    0.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 140979720)

    Null deviance: 1.9607e+11  on 1337  degrees of freedom
Residual deviance: 1.8821e+11  on 1335  degrees of freedom
AIC: 28908
```

Previous model AIC: 27124

MODEL 3: Polynomial

```
ply_chr_bmi_all<-
  glm(insurance$charges
        ~ insurance$bmi_centered
        + I(insurance$bmi_centered^2)
        + insurance$children
        + insurance$smoker
        + insurance$gender
        + insurance$age_centered
        + insurance$region)
summary(ply_chr_bmi_all)
```

When we add the other variables to this model, we get a lower AIC compared to all models

→ This model is better than the previous ones

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-10750   -2937   -1209    1693   29670

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     3402.143    505.142   6.735 2.43e-11 ***
insurance$bmi_centered           350.809     29.034  12.083  < 2e-16 ***
I(insurance$bmi_centered^2)       -7.288      3.288  -2.217 0.026807 *
insurance$children               477.127    137.604   3.467 0.000542 ***
insurance$smokeryes            23863.392    412.601  57.837  < 2e-16 ***
insurance$gendermale            -132.716    332.457  -0.399 0.689812
insurance$age_centered           255.404     11.899  21.464  < 2e-16 ***
insurance$regionnorthwest       -418.118    476.483  -0.878 0.380369
insurance$regionsoutheast       -999.225    478.262  -2.089 0.036872 *
insurance$regionsouthwest      -1012.946    477.827  -2.120 0.034199 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 36641170)

    Null deviance: 1.9607e+11  on 1337  degrees of freedom
Residual deviance: 4.8659e+10  on 1328  degrees of freedom
AIC: 27113

Number of Fisher Scoring iterations: 2
```

MODEL 4: Multiple linear Regression with Interactions

```
Call:
lm(formula = insurance$charges ~ insurance$bmi_centered * insurance$smoker_bin +
    insurance$age_centered)

Residuals:
     Min       1Q   Median       3Q      Max
-14595.4  -2015.2  -1319.2   -290.5  29313.7

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               2729.636    254.838  10.711   <2e-16 ***
insurance$bmi_centered                       7.109     25.058   0.284    0.777
insurance$smoker_bin1                    23783.372    332.569  71.514   <2e-16 ***
insurance$age_centered                     266.758      9.617  27.739   <2e-16 ***
insurance$bmi_centered:insurance$smoker_bin1 1430.920     53.217  26.888   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4907 on 1333 degrees of freedom
Multiple R-squared:  0.8363,    Adjusted R-squared:  0.8358
F-statistic:  1702 on 4 and 1333 DF,  p-value: < 2.2e-16
```

- The predicted charges for a non smoker 18 years old with an average bmi : 2,729.6$

charges = 2,729.6 + 7.109*bmi_centered + 23,783.4*smoker_bin + 266.7*age_centered + 1,430.9*bmi_centered*smoker_bin

MODEL 4: Multiple linear Regression with Interactions

```
Call:
lm(formula = insurance$charges ~ insurance$bmi_centered * insurance$smoker_bin +
    insurance$age_centered)

Residuals:
     Min       1Q   Median       3Q      Max
-14595.4  -2015.2  -1319.2   -290.5  29313.7

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 2729.636    254.838  10.711   <2e-16 ***
insurance$bmi_centered                         7.109     25.058   0.284    0.777
insurance$smoker_bin1                      23783.372    332.569  71.514   <2e-16 ***
insurance$age_centered                       266.758      9.617  27.739   <2e-16 ***
insurance$bmi_centered:insurance$smoker_bin1 1430.920     53.217  26.888   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4907 on 1333 degrees of freedom
Multiple R-squared:  0.8363,    Adjusted R-squared:  0.8358
F-statistic:  1702 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The predicted charges for a 1 unit increase in bmi for an 18yo non smoker: 2,729.6+7.11=2,736.71$

charges = 2,729.6 + 7.109*bmi_centered + 23,783.4*smoker_bin + 266.7*age_centered + 1,430.9*bmi_centered*smoker_bin

MODEL 4: Multiple linear Regression with Interactions

```
Call:
lm(formula = insurance$charges ~ insurance$bmi_centered * insurance$smoker_bin +
    insurance$age_centered)

Residuals:
     Min       1Q   Median       3Q      Max
-14595.4  -2015.2  -1319.2   -290.5  29313.7

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               2729.636    254.838  10.711   <2e-16 ***
insurance$bmi_centered                       7.109     25.058   0.284    0.777
insurance$smoker_bin1                    23783.372    332.569  71.514   <2e-16 ***
insurance$age_centered                     266.758      9.617  27.739   <2e-16 ***
insurance$bmi_centered:insurance$smoker_bin1 1430.920     53.217  26.888   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4907 on 1333 degrees of freedom
Multiple R-squared:  0.8363,    Adjusted R-squared:  0.8358
F-statistic:  1702 on 4 and 1333 DF,  p-value: < 2.2e-16
```

- For a smoker with an extra unit of bmi means they will pay: 7.109+23,783.4+1,430.9=25,221.41$

charges = 2,729.6 + 7.109*bmi_centered + 23,783.4*smoker_bin + 266.7*age_centered + 1,430.9*bmi_centered*smoker_bin

Limitations

01 Salary

02 Job type & risk

03 Living expenses

05 Lifestyle

06 Chronic health issues & disabilities

Case Closed!

Ahmed is unable to prove his claim due to lack of data. He chooses to change his insurance company.