Regression, EDA about medical cost

Roysen

12/11/2020

```
library(tidyverse)
library(forcats)
library(ggridges)
library(broom.mixed)
library(corrplot)
library(modelr)
library(lme4)
library(patchwork)
library(broom)
```

读取数据

```
insurance <- read_csv("~/workspace/insurance.csv")
insurance
```

```
## # A tibble: 1,338 x 7
##      age sex      bmi children smoker region    charges
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
##  1    19 female  27.9        0 yes    southwest  16885.
##  2    18 male    33.8        1 no     southeast   1726.
##  3    28 male    33          3 no     southeast   4449.
##  4    33 male    22.7        0 no     northwest  21984.
##  5    32 male    28.9        0 no     northwest   3867.
##  6    31 female  25.7        0 no     southeast   3757.
##  7    46 female  33.4        1 no     southeast   8241.
##  8    37 female  27.7        3 no     northwest   7282.
##  9    37 male    29.8        2 no     northeast   6406.
## 10    60 female  25.8        0 no     northwest  28923.
## # … with 1,328 more rows
```

序号	变量	注释
1	age	年龄
2	sex	性别
3	bmi	身体质量指数，成人标准值（18.5-23.9），算法：kg/(m^2)
4	children	小孩数量
5	smoker	是否吸烟
6	region	地区
7	charges	投保费用

数据集有1338行（观测值），7个变量，3个字符型向量，4个数字型向量。
该数据集主要是针对医疗费用支出收集的相关数据，自变量包括用户的年龄、性别、身体质量指数、小孩数量、是否吸烟、地区。属于回归分析的范畴。

##检查缺失值

```
insurance %>%
    summarise_all(
        ~ sum(is.na(.))
    )
```

```
## # A tibble: 1 x 7
##     age   sex   bmi children smoker region charges
##   <int> <int> <int>    <int>  <int>  <int>   <int>
## 1     0     0     0        0      0      0       0
```

各个变量均没有缺失值。 #数据概览

```
summary(insurance)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

##将sex 、smoker、region变量转换为因子，并保留在insu_df数据框

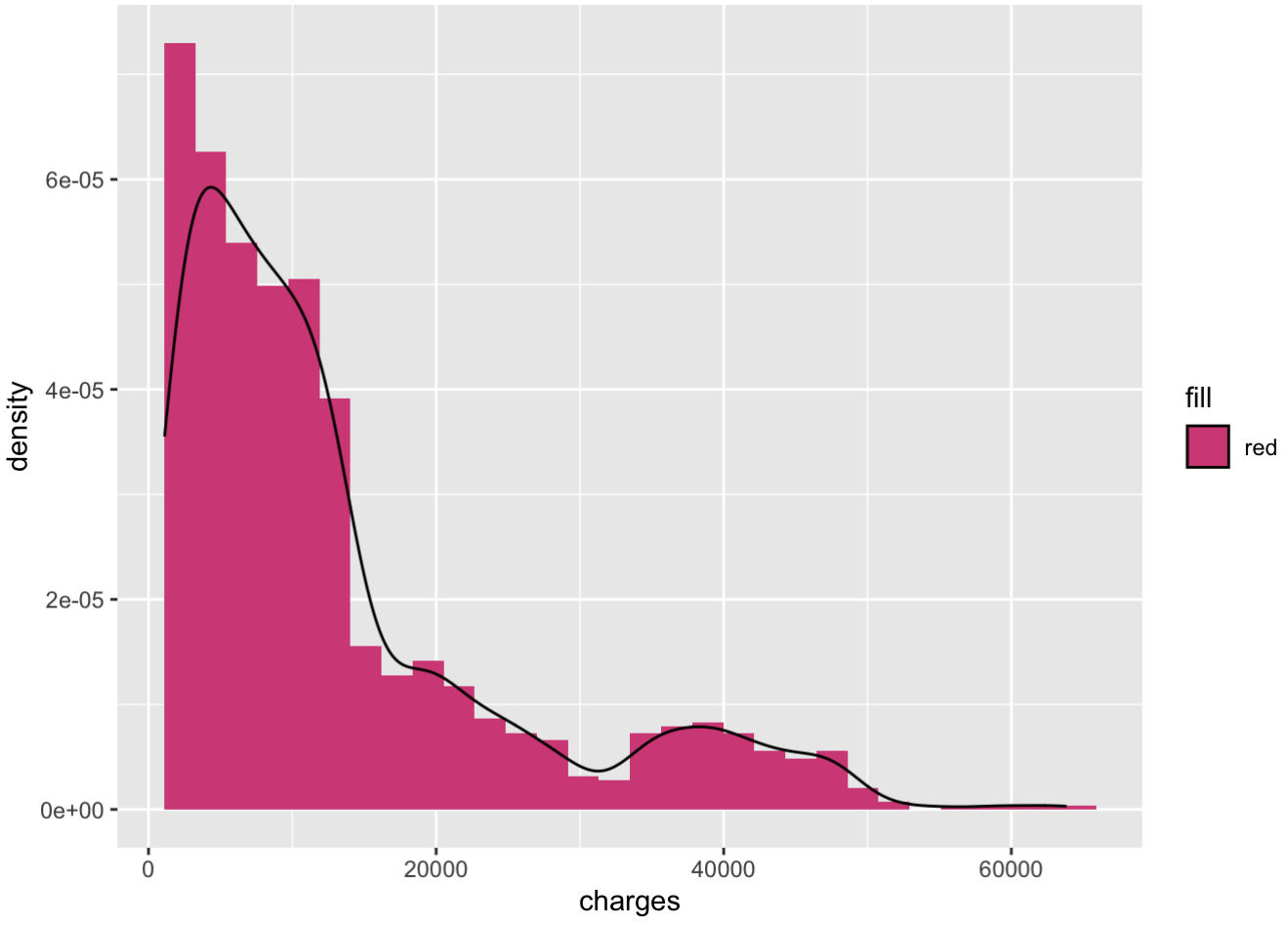
```
insu_df <- insurance %>%
    mutate(sex = factor(sex),
           smoker = factor(smoker),
           region = factor(region)
           )
insu_df
```

```
## # A tibble: 1,338 x 7
##      age sex      bmi children smoker region    charges
##    <dbl> <fct>  <dbl>    <dbl> <fct>  <fct>       <dbl>
##  1    19 female  27.9        0 yes    southwest  16885.
##  2    18 male    33.8        1 no     southeast   1726.
##  3    28 male    33          3 no     southeast   4449.
##  4    33 male    22.7        0 no     northwest  21984.
##  5    32 male    28.9        0 no     northwest   3867.
##  6    31 female  25.7        0 no     southeast   3757.
##  7    46 female  33.4        1 no     southeast   8241.
##  8    37 female  27.7        3 no     northwest   7282.
##  9    37 male    29.8        2 no     northeast   6406.
## 10    60 female  25.8        0 no     northwest  28923.
## # … with 1,328 more rows
```

#变量简单统计 #insurance变量分布图

```
insurance %>%
ggplot(aes(x = charges, y = stat(density)))+
    geom_histogram(aes(fill = "red"))+
    scale_fill_manual(values = c("#d45087"))+
    geom_density()
```

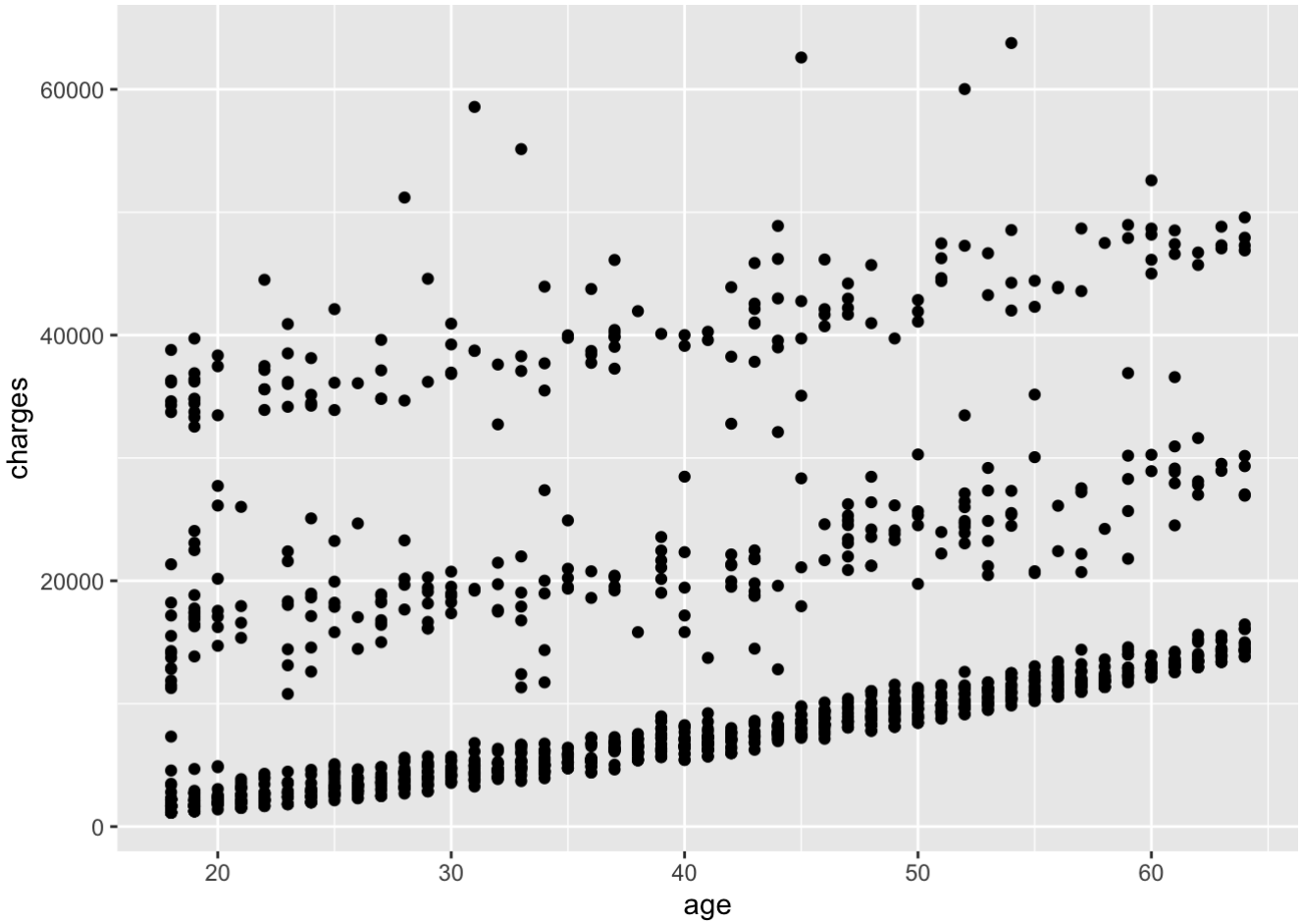
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



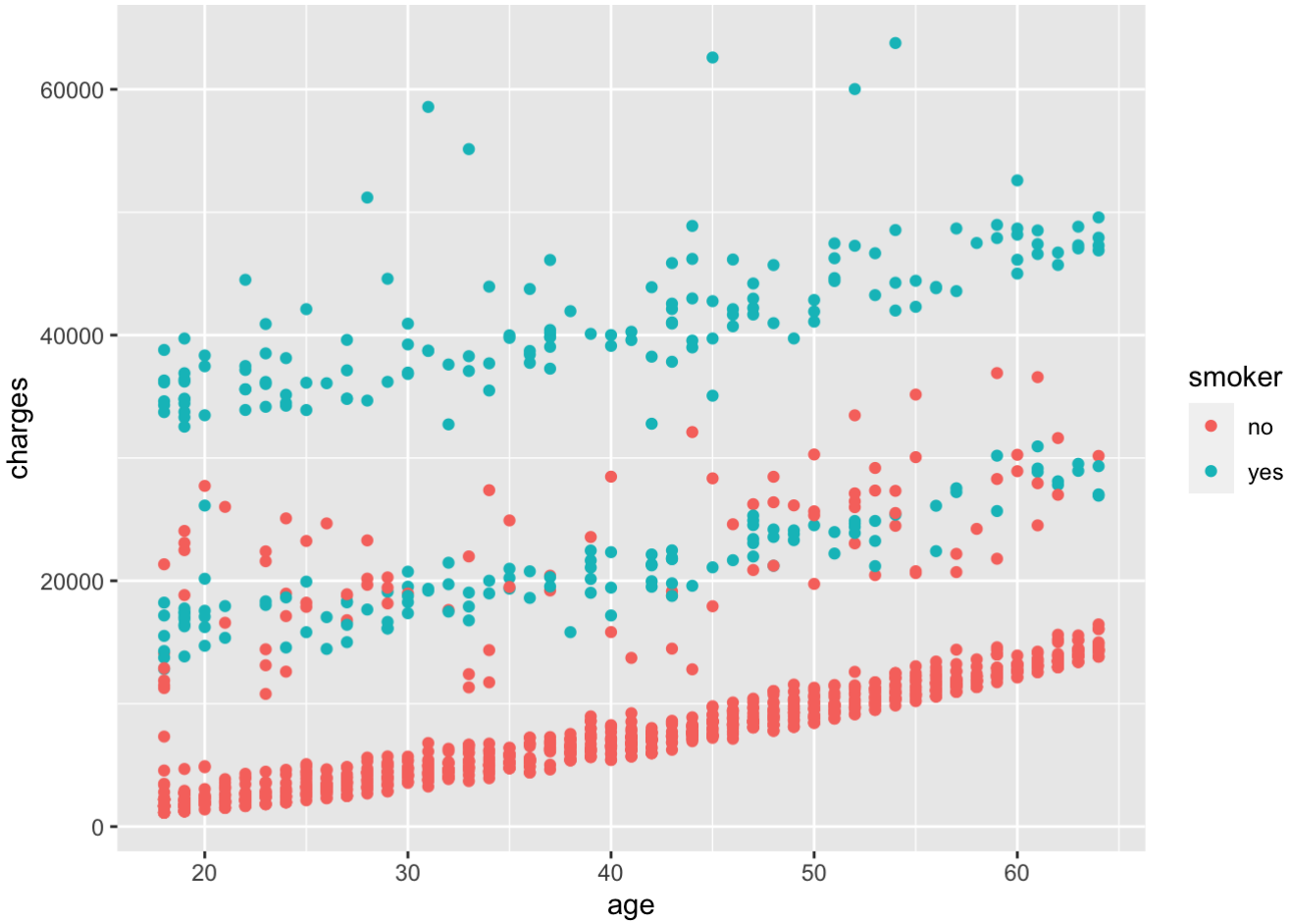
##年龄因素

```
insu_df %>%
    ggplot(aes(age, fill = "red"))+
    geom_histogram(bins = 60)
```


```
insu_df %>%
    ggplot(aes(age, charges))+
    geom_point()
```




```
insu_df %>%
    ggplot(aes(age, charges, color = smoker))+
    geom_point()
```



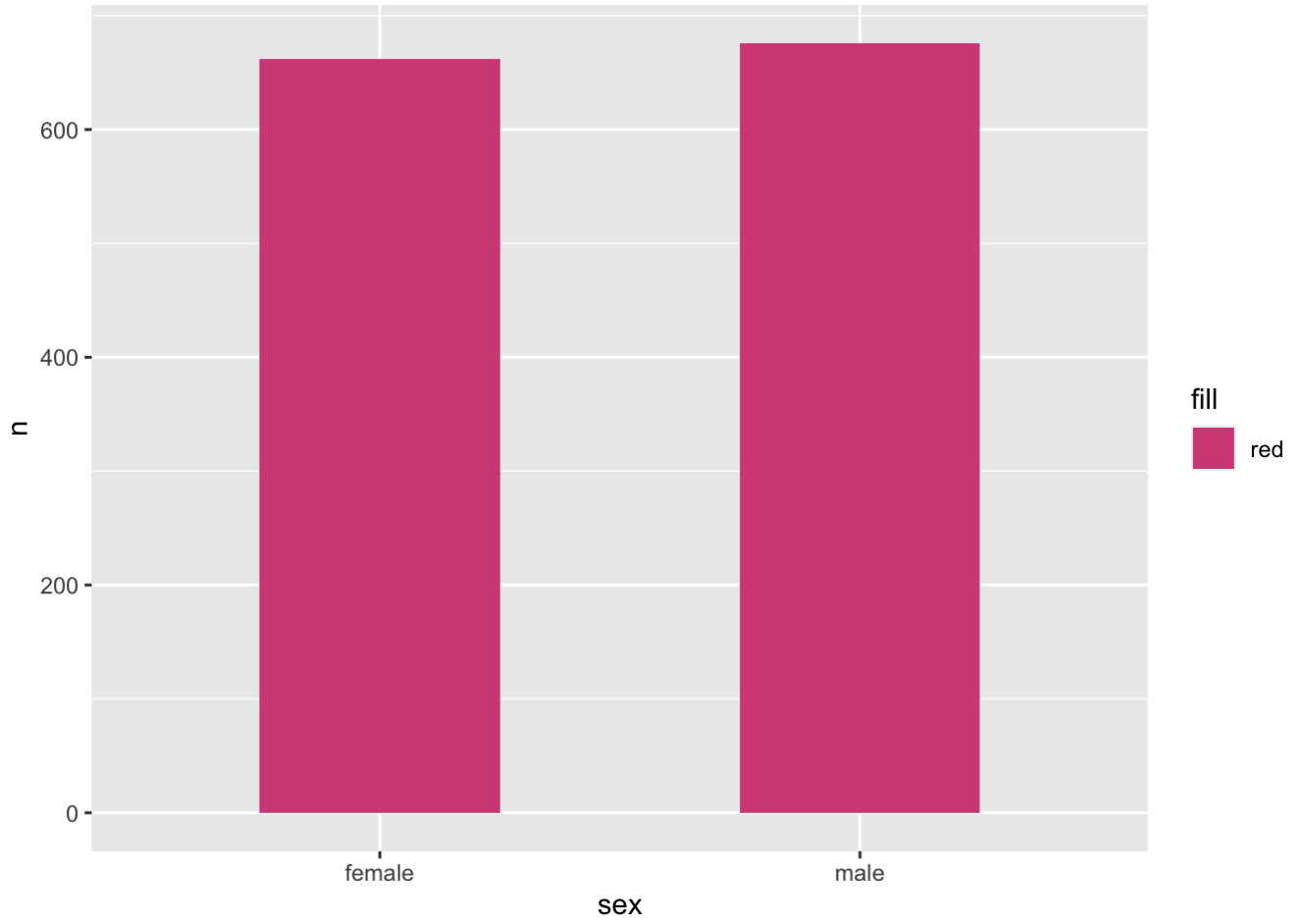
20岁以下的用户占比最高，年龄和保险支出费用呈线性关系，但是这种关系应该收到其他因素的影响，表现为三层线性关系。 经分析年龄和保险支出费用中需加入smoker这个因素，可能包含着交互效应以及多层模型。 图形表明：

医疗支出费用随着年龄增加而增加，

抽烟群体的保险支出费用比非抽烟人群的保险支出费用要高，都成线性增长趋势。

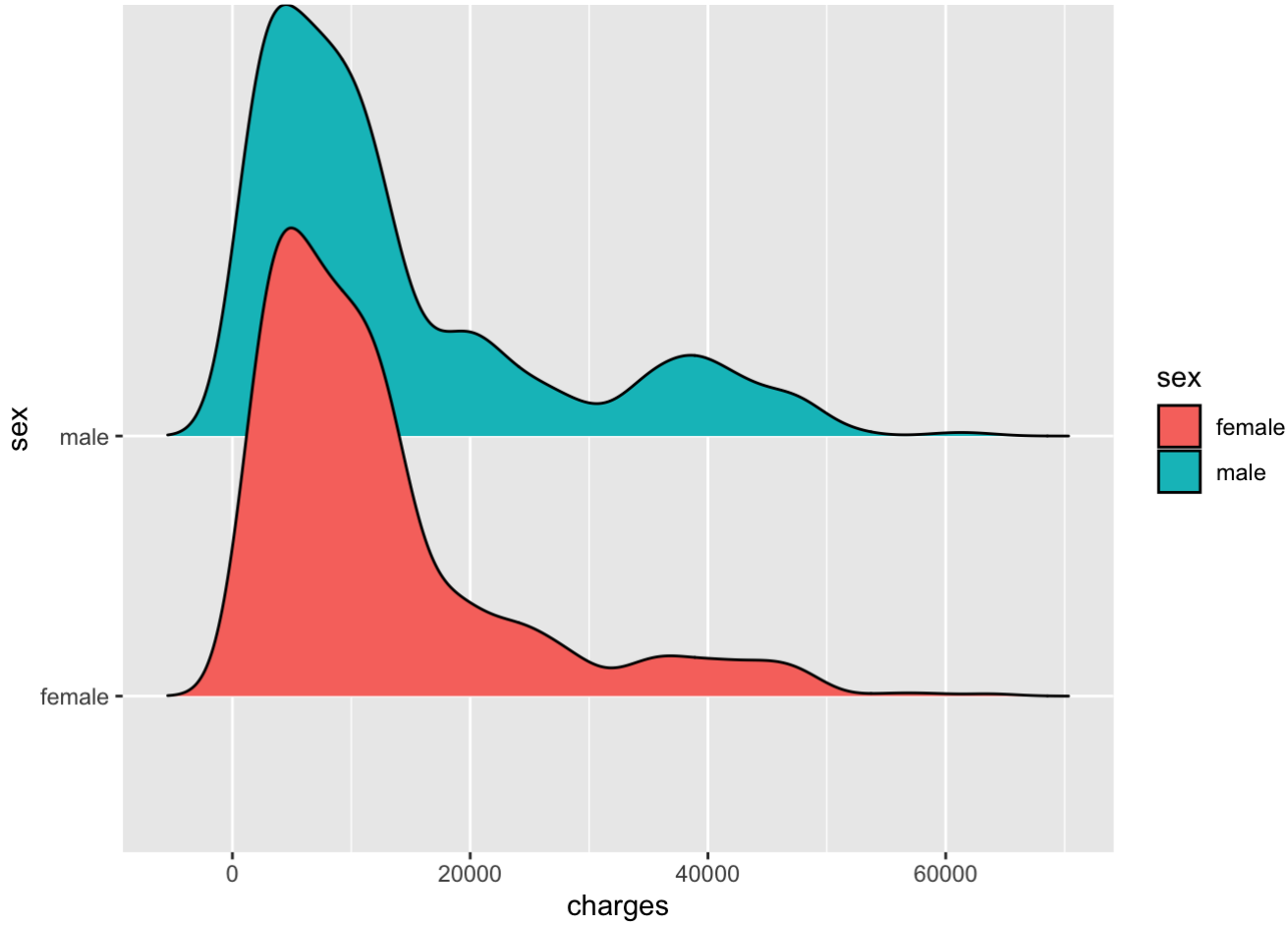
##性别因素

```
insu_df %>%
    count(sex) %>%
ggplot(aes(sex, n))+
    geom_col( aes(fill = "red", width = 0.5))+
    scale_fill_manual(values = c("#d45087"))
```

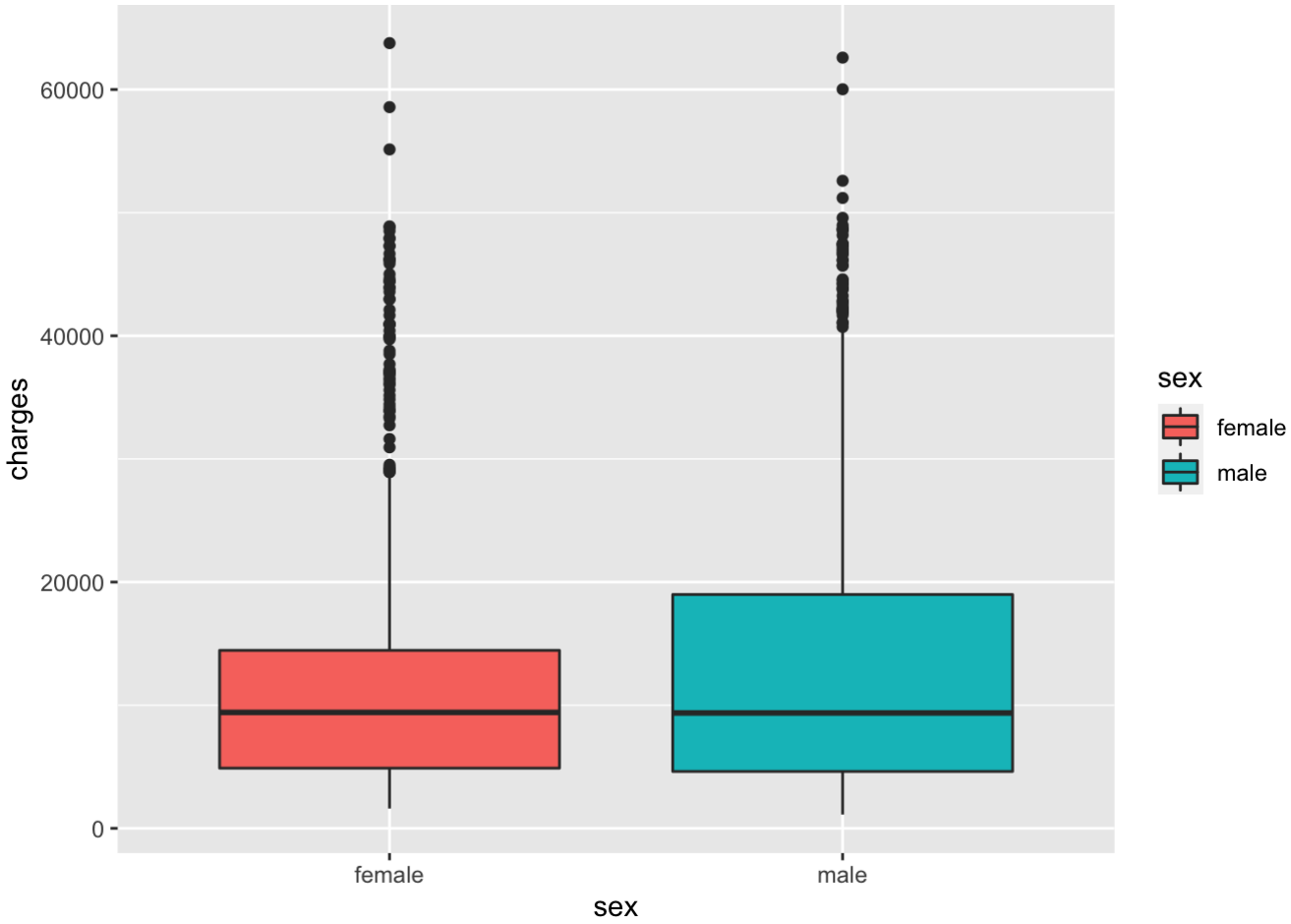



```
insu_df %>%
ggplot(aes(charges, sex, fill = sex)) +
    ggridges::geom_density_ridges()
```

```
## Picking joint bandwidth of 2190
```




```
insu_df %>%
ggplot(aes(sex, charges, fill = sex))+
    geom_boxplot()
```

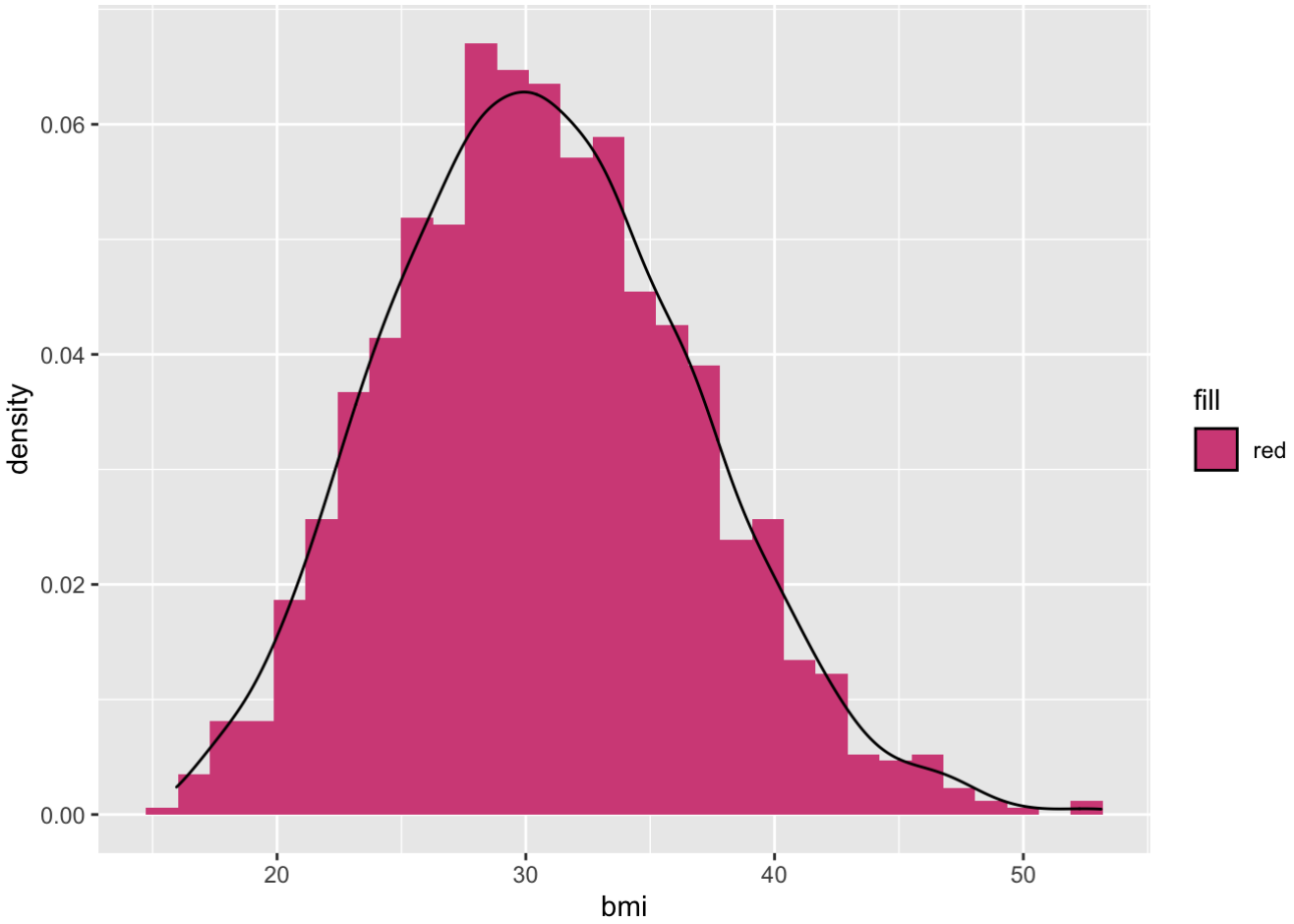


性别因素，女性医疗费用支出费用集中在（500-1500美元），男性集中在（500-2000美元）。
女性保险支出费用在（15000-30000美元）和男性支出费用在（20000-40000美元），为正常的波动范围，男性的波动幅度更大些。女性30000美元以上和男性40000以上为各自组内的异常值，表明这一部分用户群相对是少数。在异常值的部分，女性用户的波动范围很大（30000-65000美元），可能是什么原因呢？也许需要进行问卷调研。
共同点：医疗费用支出有着大致相同的曲线，分别都有两个高峰。原因同需探究。

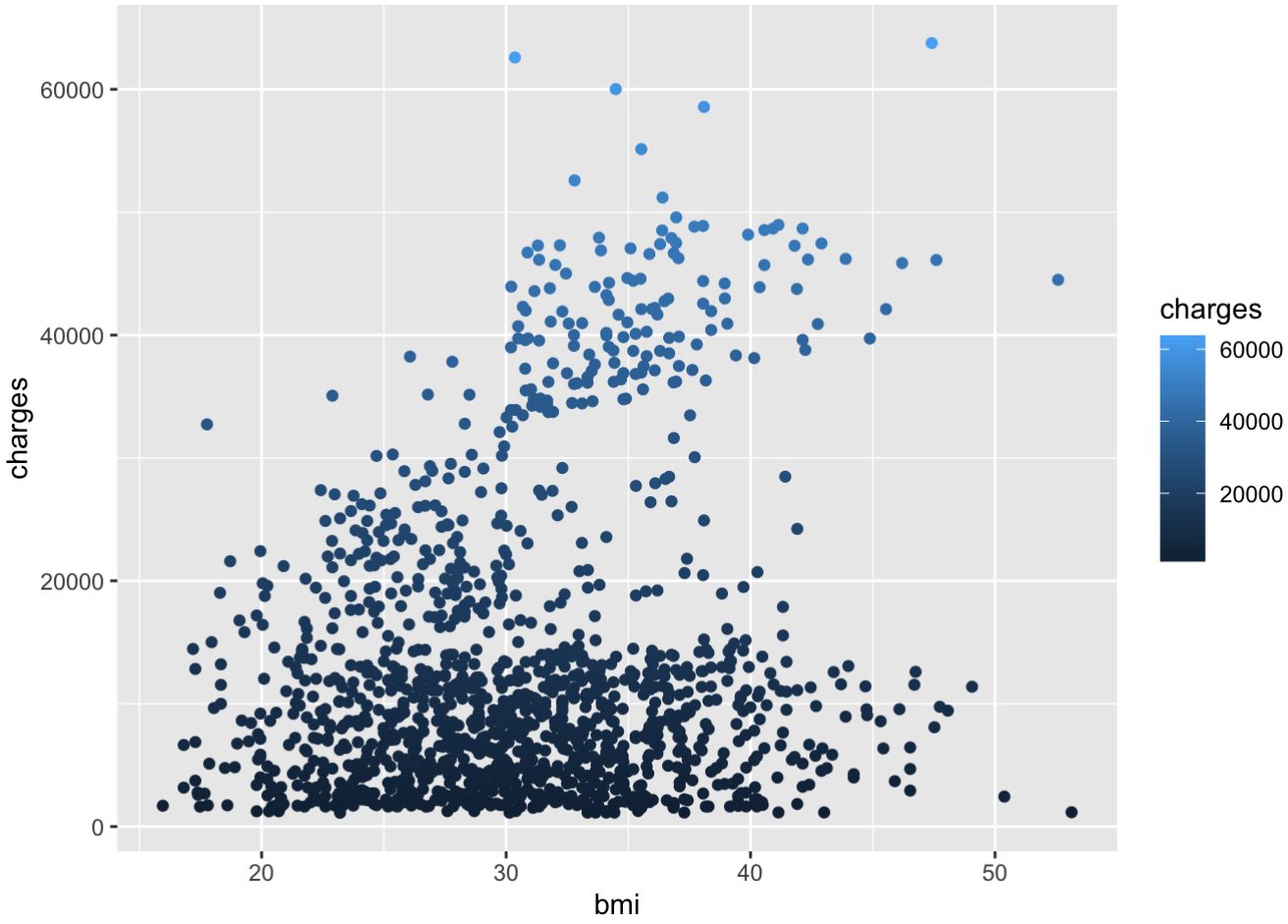
##bmi因素

```
insu_df %>%
ggplot(aes(x = bmi, y = stat(density)))+
    geom_histogram( aes(fill = "red"))+
    scale_fill_manual(values = c("#d45087"))+
    geom_density()
```

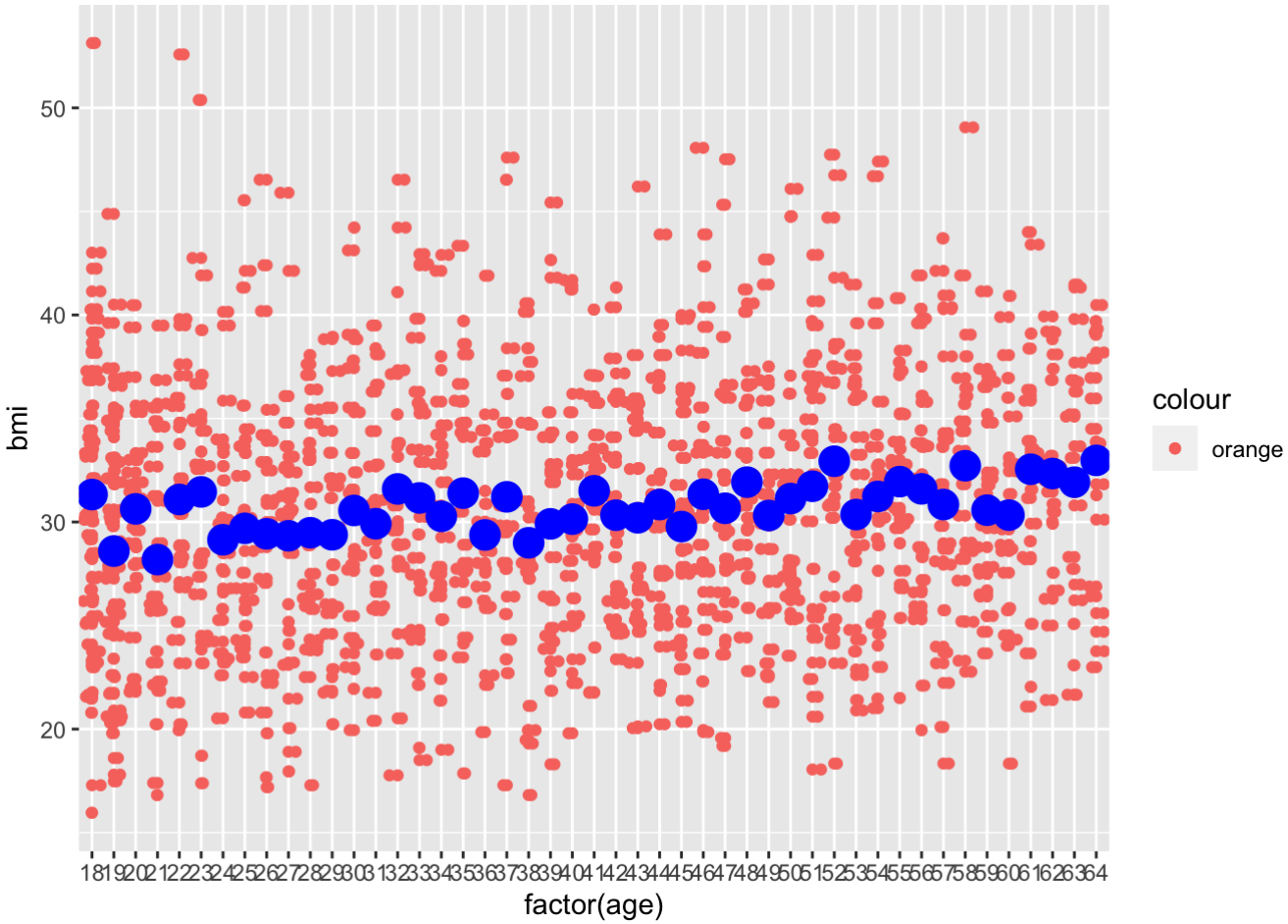
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
insu_df %>%
ggplot(aes(x = bmi, charges, color = charges))+
    geom_point()
```



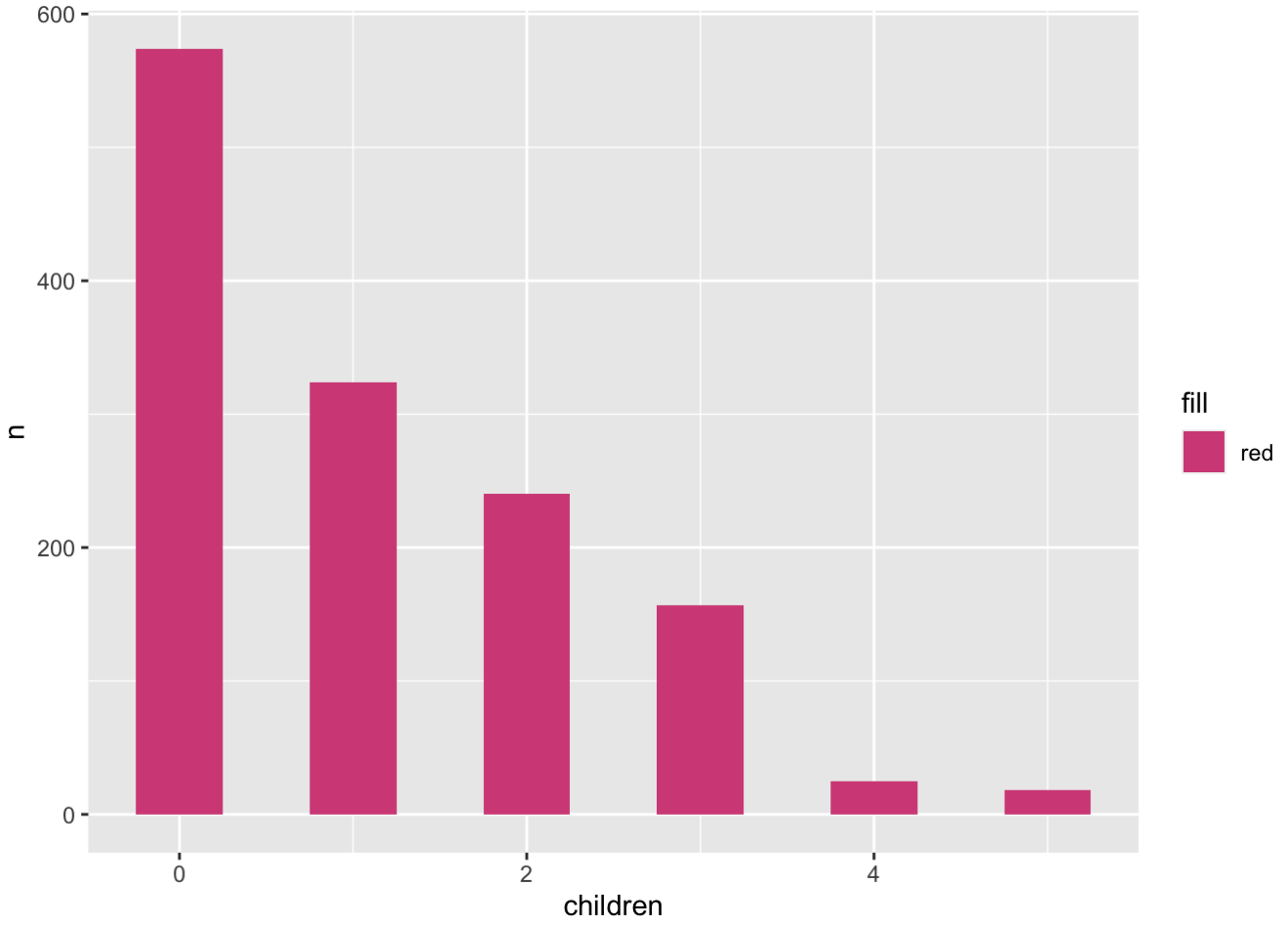

```
insu_df %>%
ggplot(aes(factor(age), bmi, color = "orange"))+
    geom_point()+
    geom_jitter()+
    stat_summary(fun.y = mean, colour = "blue", geom = "point", size = 5)
```



用户各个年龄段的平均bmi为30，bmi范围集中（25-35），在大致来说当bmi超过30，即身体明显处于肥胖状态时，保险支出费用有极大值，费用超过了30000美元。同时，各个年龄段肥胖人数的比例都接近50%，这是一个不好的现象。

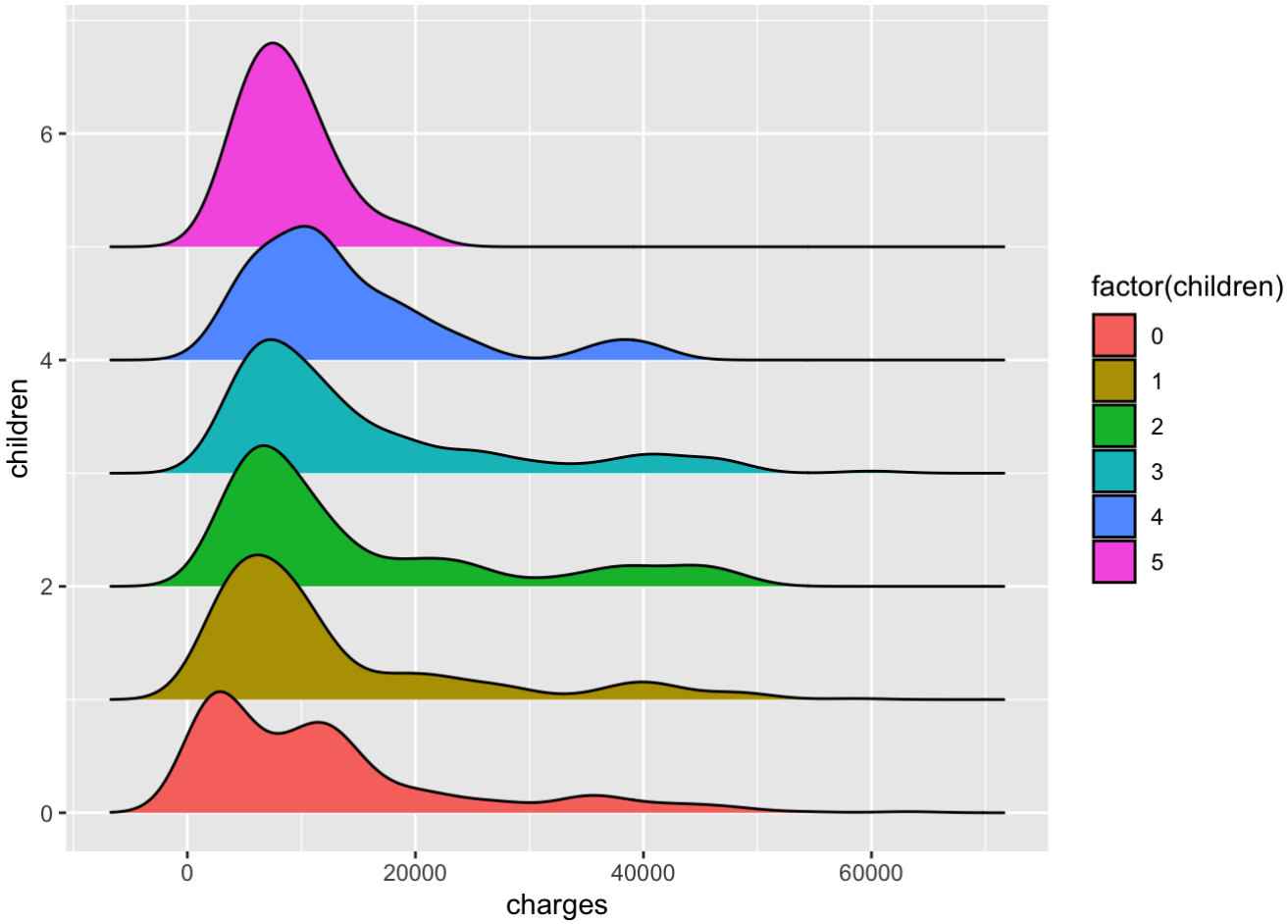
##儿童个数因素

```
insurance %>%
    count(children) %>%
    ggplot(aes(children, n))+
    geom_col( aes(fill = "red", width = 0.5))+
    scale_fill_manual(values = c("#d45087"))
```

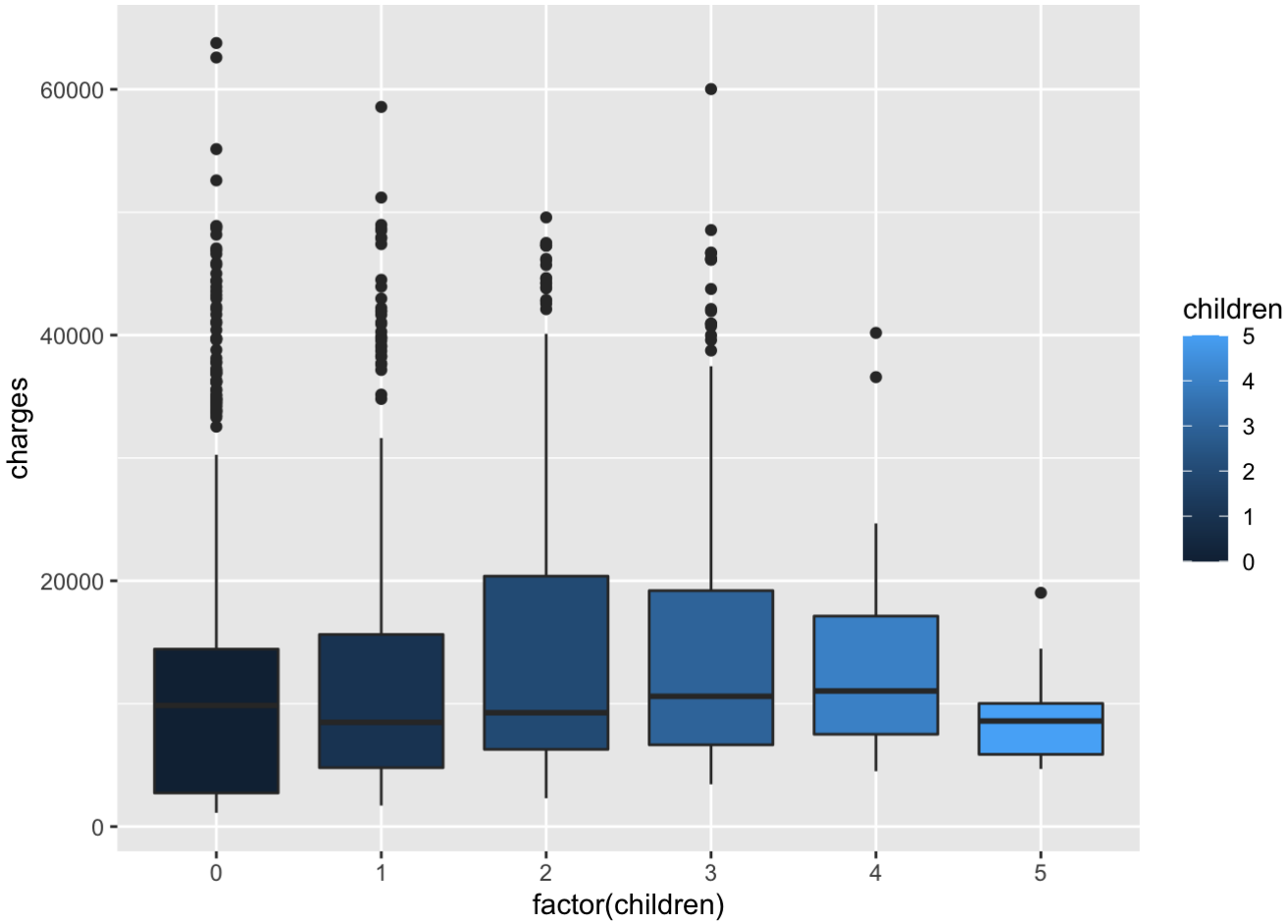



```
insurance %>%
    ggplot(aes(charges, children))+
    ggridges::geom_density_ridges(aes(fill = factor(children)))
```

```
## Picking joint bandwidth of 2610
```



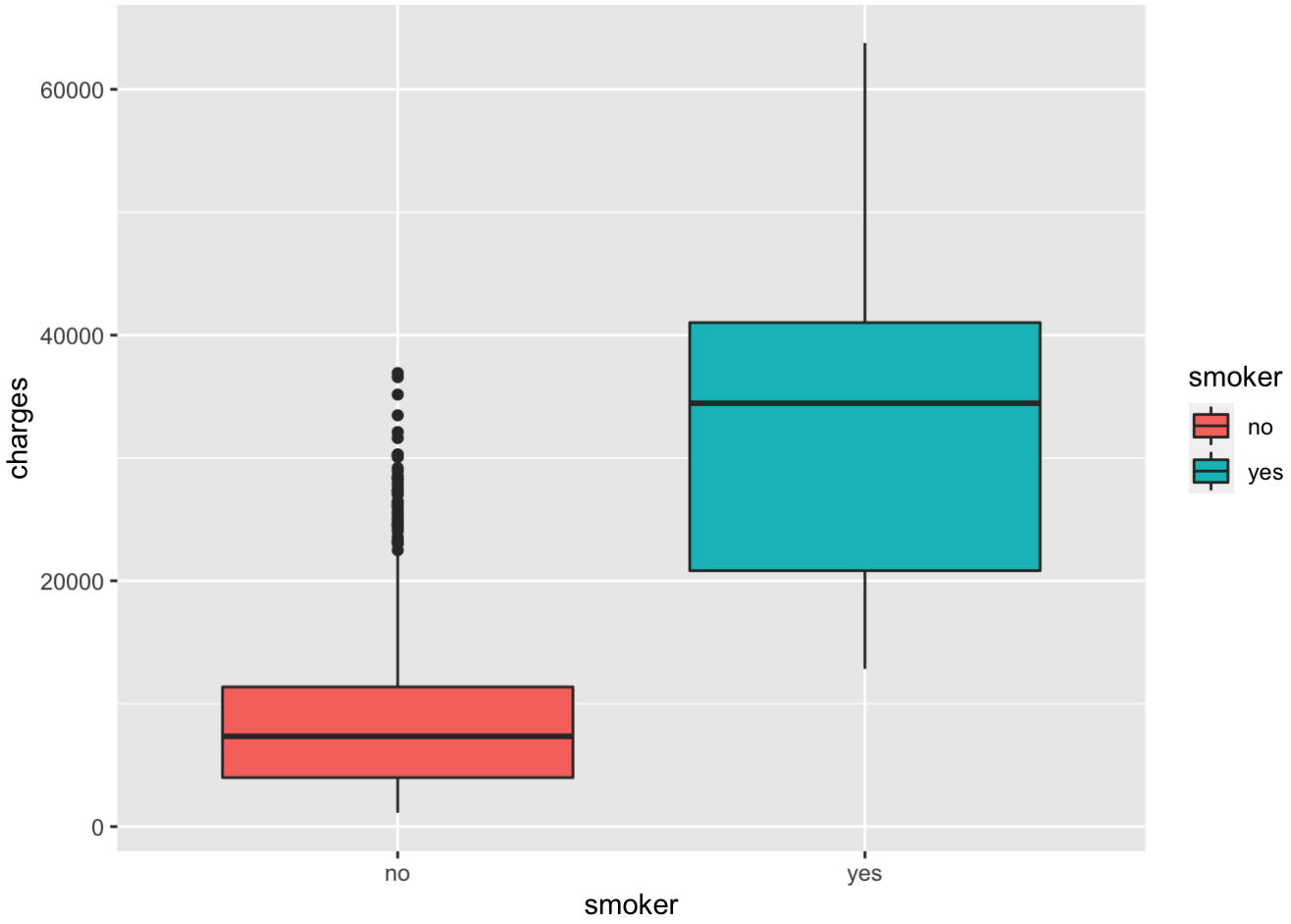

```
insurance %>%
    ggplot(aes(factor(children), charges, fill = children))+
    geom_boxplot()
```



儿童医保支出费用平均1000美元，同男女性别上并无差异，集中在20000美元以下，对于小于等于3个的儿童用户，支出费用极大值广泛存在。孩子多了各项支出也相应多了。

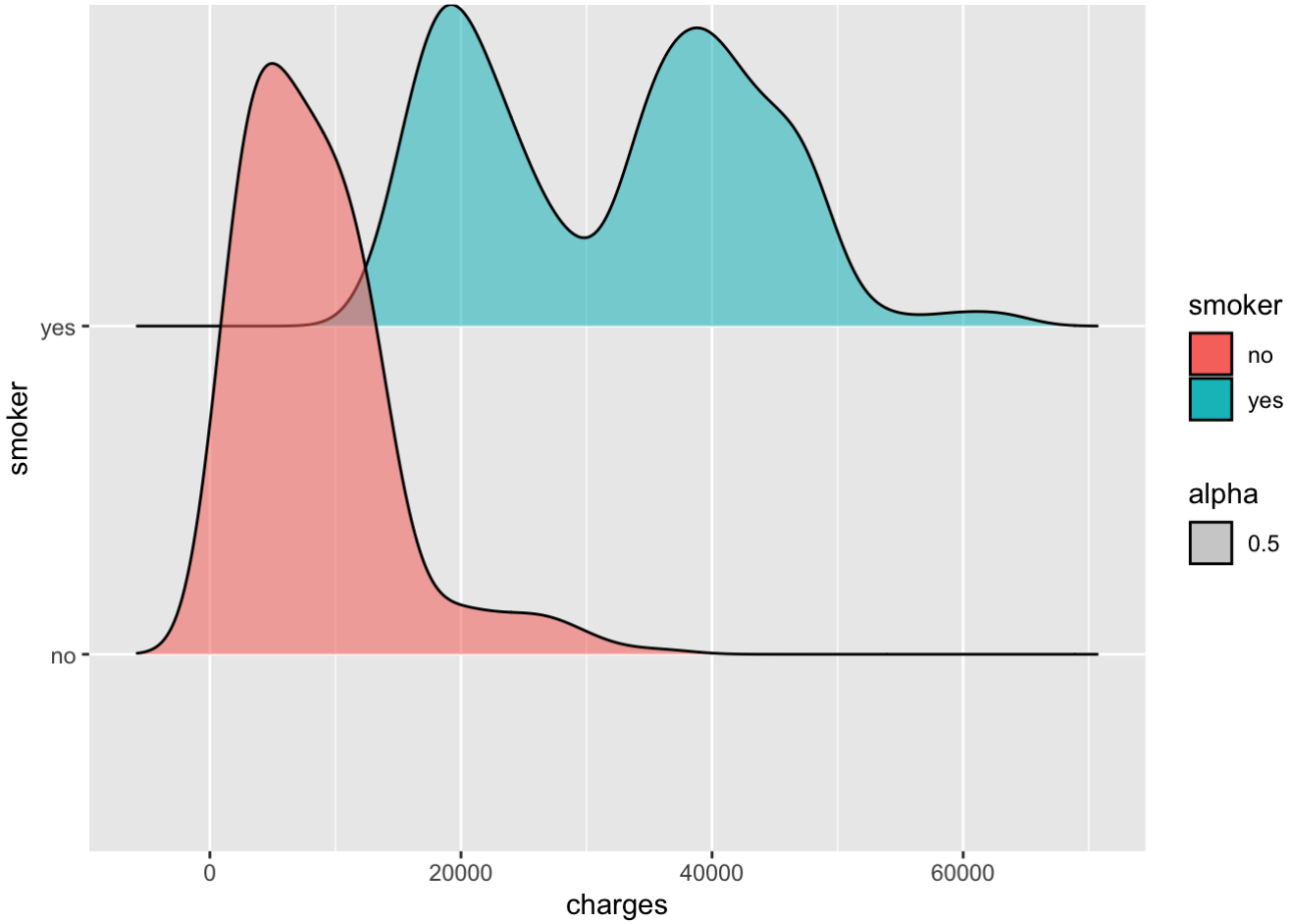
#smoker因素

```
insurance %>%
    ggplot(aes(smoker,charges, fill = smoker))+
    geom_boxplot()
```




```
insurance %>%
    ggplot(aes(charges, smoker, fill = smoker, alpha = 0.5))+
    ggridges::geom_density_ridges()
```

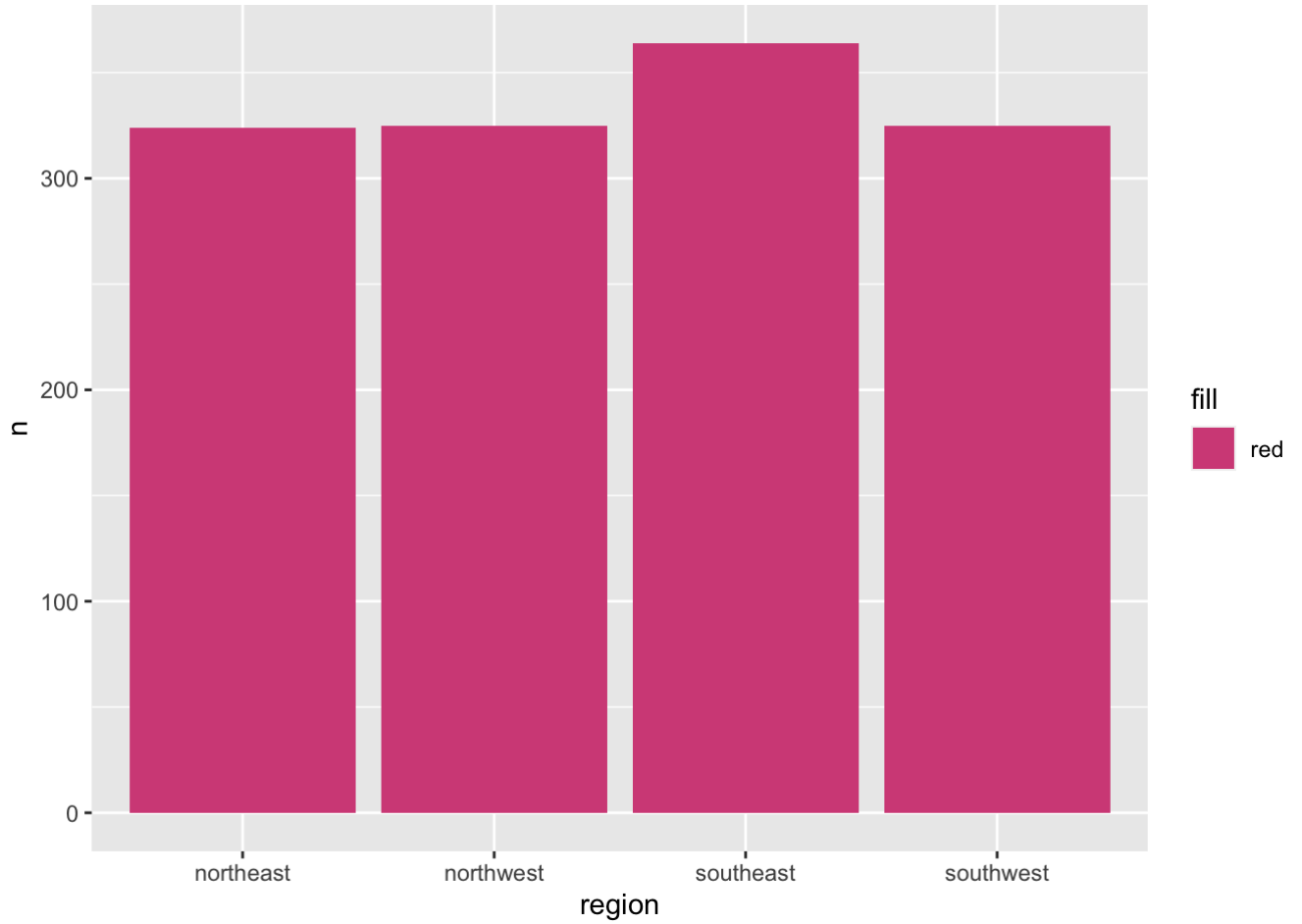
```
## Picking joint bandwidth of 2300
```



在正常的波动范围内，吸烟用户的保险支出费用远高于非吸烟用户，吸烟用户在支出费用为30000美金左右呈现山谷状，是什么因素导致这一情况呢？需要对吸烟群体进行划分分析，也许和年龄、收入等有关。

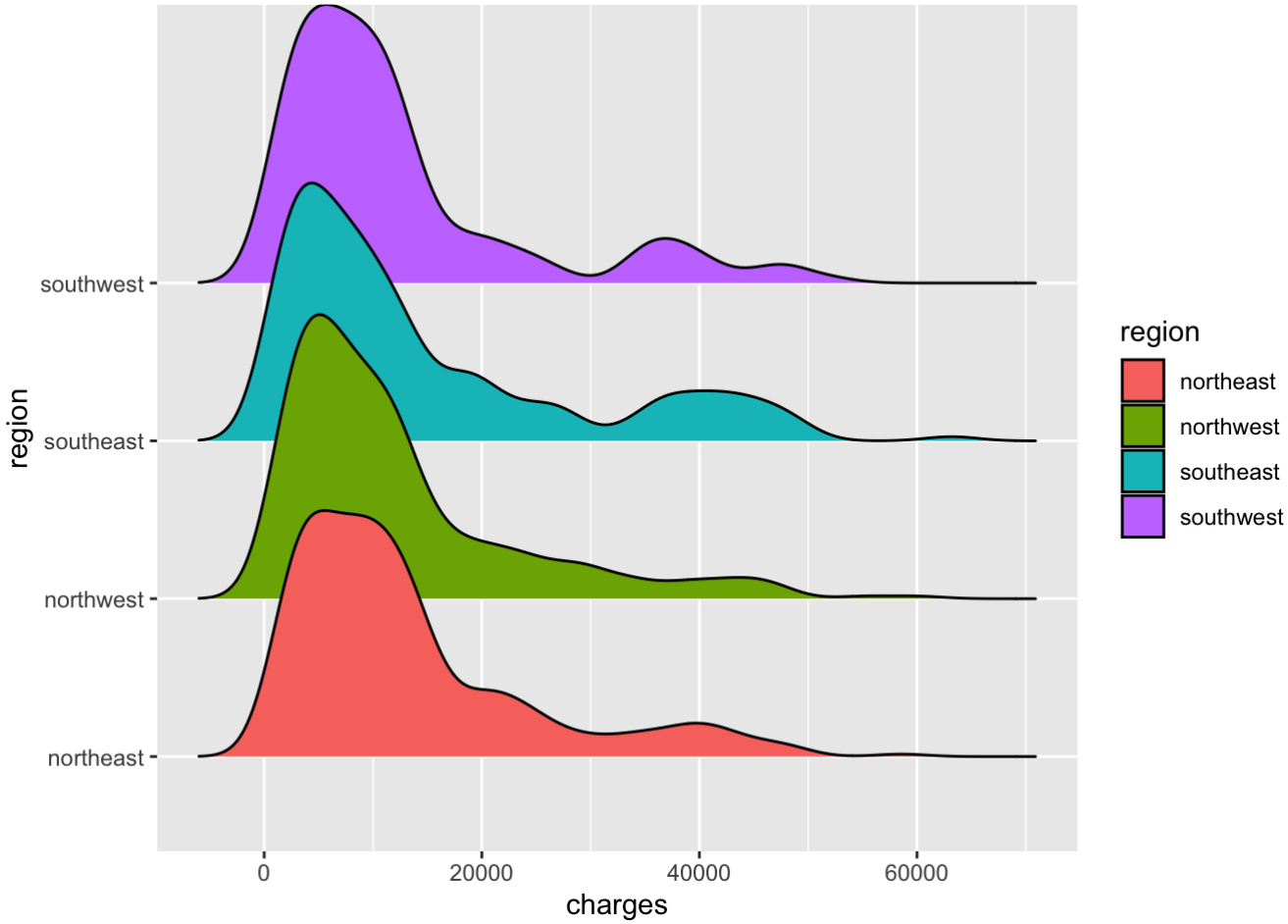
#地区因素

```
insurance %>%
    count(region) %>%
    ggplot(aes(region, n, fill = "red"))+
    geom_col()+
    scale_fill_manual(values = c("#d45087"))
```

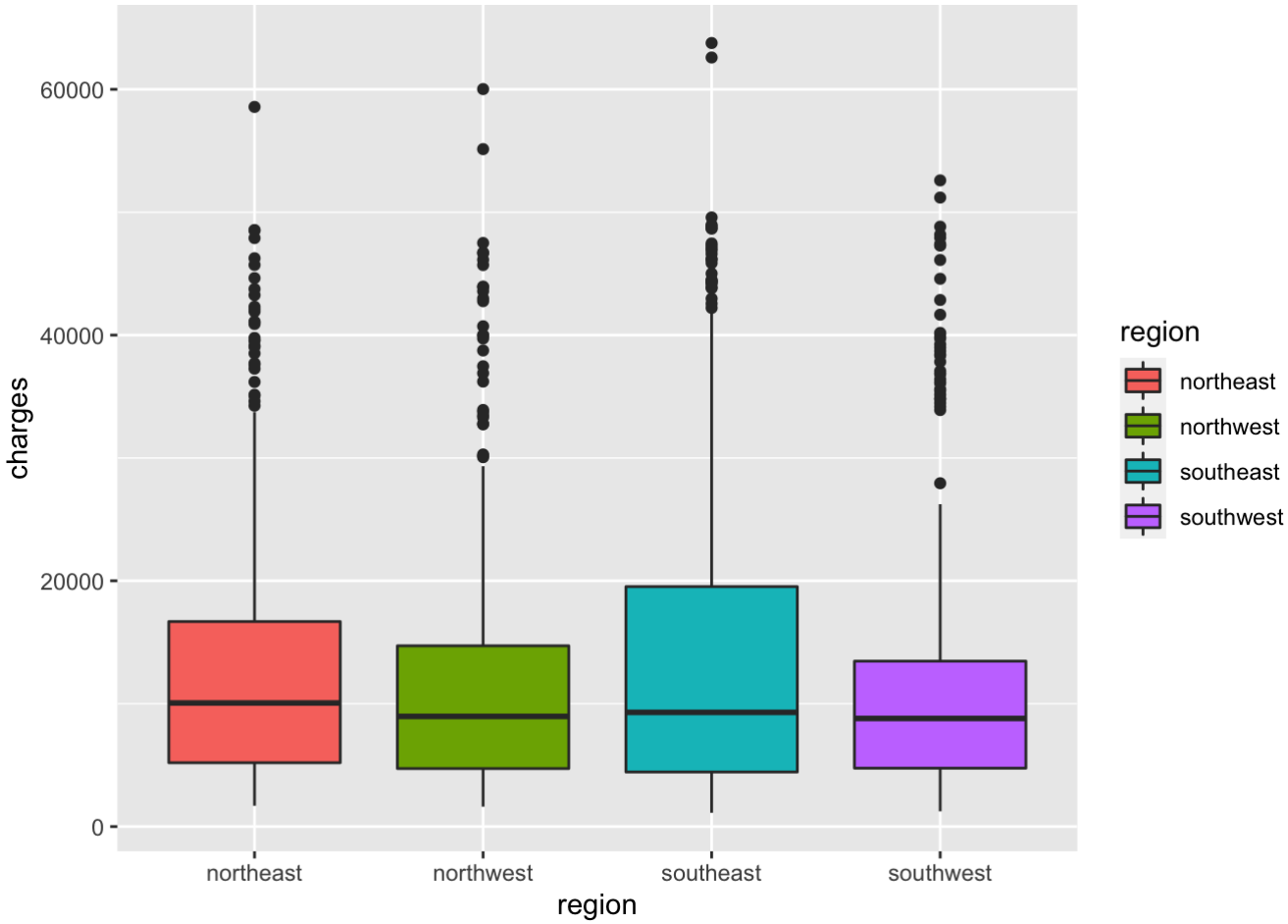



```
insurance %>%
    ggplot(aes(charges, region, fill = region))+
    ggridges::geom_density_ridges()
```

```
## Picking joint bandwidth of 2370
```



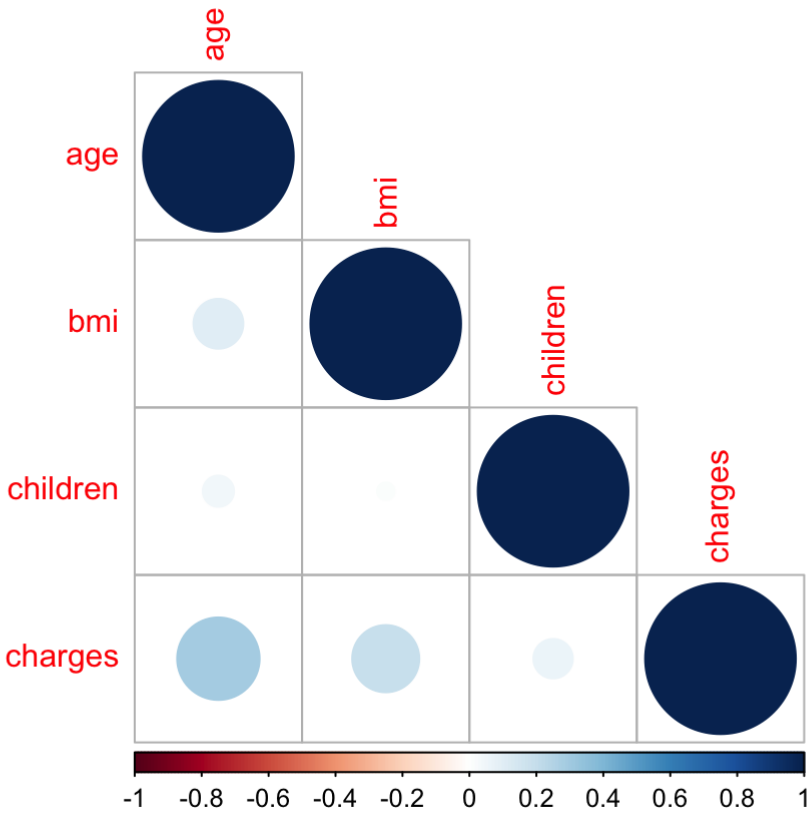

```
insurance %>%
    ggplot(aes(region, charges, fill = region))+
    geom_boxplot()
```



地区上，东南部医疗保险支出费用的极值要更高于其他地区，可能是因为地处美国的政治、经济、文化、中心–纽约及其周边。

#相关系数矩阵

```
insurance %>%
    select(-sex, -smoker, -region) %>%
    cor() %>%
    corrplot::corrplot(type = "lower")
```



charges同age、bmi、children都呈现正相关，同age的相关性要强一些。

##model

```
fit <- insu_df %>%
  mutate_at(vars(age, bmi,children, charges), scale) %>%
  lm(charges ~ age+bmi+children, data = .)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1465 -0.5775 -0.4205  0.5884  4.0155
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.732e-17  2.567e-02   0.000   1.0000
## age         2.784e-01  2.586e-02  10.767  < 2e-16 ***
## bmi         1.672e-01  2.584e-02   6.472 1.35e-10 ***
## children    5.404e-02  2.571e-02   2.102   0.0357 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9391 on 1334 degrees of freedom
## Multiple R-squared:  0.1201, Adjusted R-squared:  0.1181
## F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```
broom::tidy(fit)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 2.73e-17    0.0257  1.06e-15 1.00e+ 0
## 2 age         2.78e- 1    0.0259  1.08e+ 1 5.53e-26
## 3 bmi         1.67e- 1    0.0258  6.47e+ 0 1.35e-10
## 4 children    5.40e- 2    0.0257  2.10e+ 0 3.57e- 2
```

在显著水平为0.1%的情况下，age和bmi相对更重要，children数量的重要性相对弱一些。

```
mod2 <- lm(charges ~ age + smoker, data = insu_df)
mod2
```

```
##
## Call:
## lm(formula = charges ~ age + smoker, data = insu_df)
##
## Coefficients:
## (Intercept)          age    smokeryes
##     -2391.6        274.9      23855.3
```

```
broom::tidy(mod2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -2392.     528.      -4.53 6.52e- 6
## 2 age             275.      12.5     22.1  2.91e-92
## 3 smokeryes     23855.     433.      55.0  0.
```

#多重线性

```
mod3 <- lmer(charges ~ age + (1 + age | smoker), data = insu_df)
```

```
## boundary (singular) fit: see ?isSingular
```

```
mod3
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ age + (1 + age | smoker)
##    Data: insu_df
## REML criterion at convergence: 27234.39
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  smoker   (Intercept) 9963.72
##           age           18.18  1.00
##  Residual             6396.79
## Number of obs: 1338, groups:  smoker, 2
## Fixed Effects:
## (Intercept)          age
##        9066          287
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warning
s
```

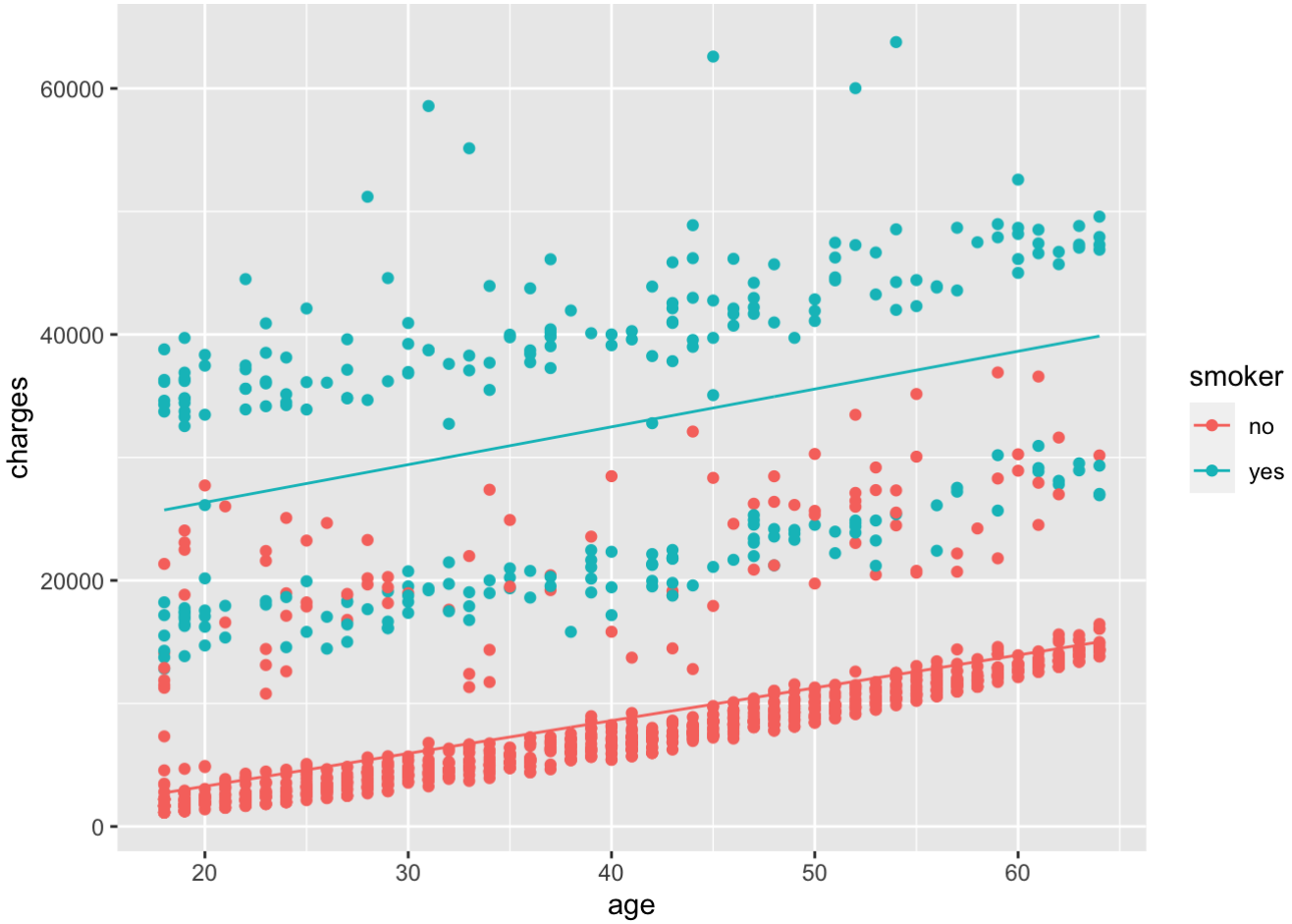
```
broom.mixed::tidy(mod3, effects = "fixed")
```

```
## # A tibble: 2 x 5
##   effect term        estimate std.error statistic
##   <chr>  <chr>          <dbl>     <dbl>     <dbl>
## 1 fixed  (Intercept)    9066.    7065.       1.28
## 2 fixed  age             287.      17.9     16.0
```

```
broom.mixed::tidy(mod3, effects = "ran_vals")
```

```
## # A tibble: 4 x 6
##   effect   group  level term        estimate std.error
##   <chr>    <chr>  <chr> <chr>          <dbl>     <dbl>
## 1 ran_vals smoker no    (Intercept) -11132.    183.
## 2 ran_vals smoker yes   (Intercept)  11132.    361.
## 3 ran_vals smoker no    age            -20.3     0.334
## 4 ran_vals smoker yes   age             20.3     0.658
```

```
insu_df %>%
  add_predictions(mod3) %>%
  ggplot(aes( age, charges, color = smoker)) +
  geom_point() +
  geom_line(aes(x = age, y = pred))
```




```
mod4 <- lmer(charges ~ age + (1 + age | sex), data = insu_df)
```

```
## boundary (singular) fit: see ?isSingular
```

```
mod4
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ age + (1 + age | sex)
##    Data: insu_df
## REML criterion at convergence: 28805.38
## Random effects:
##  Groups   Name        Std.Dev.  Corr
##  sex      (Intercept)   908.156
##           age             2.365 1.00
##  Residual             11538.881
## Number of obs: 1338, groups:  sex, 2
## Fixed Effects:
## (Intercept)          age
##      3122.8        258.7
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warning
s
```

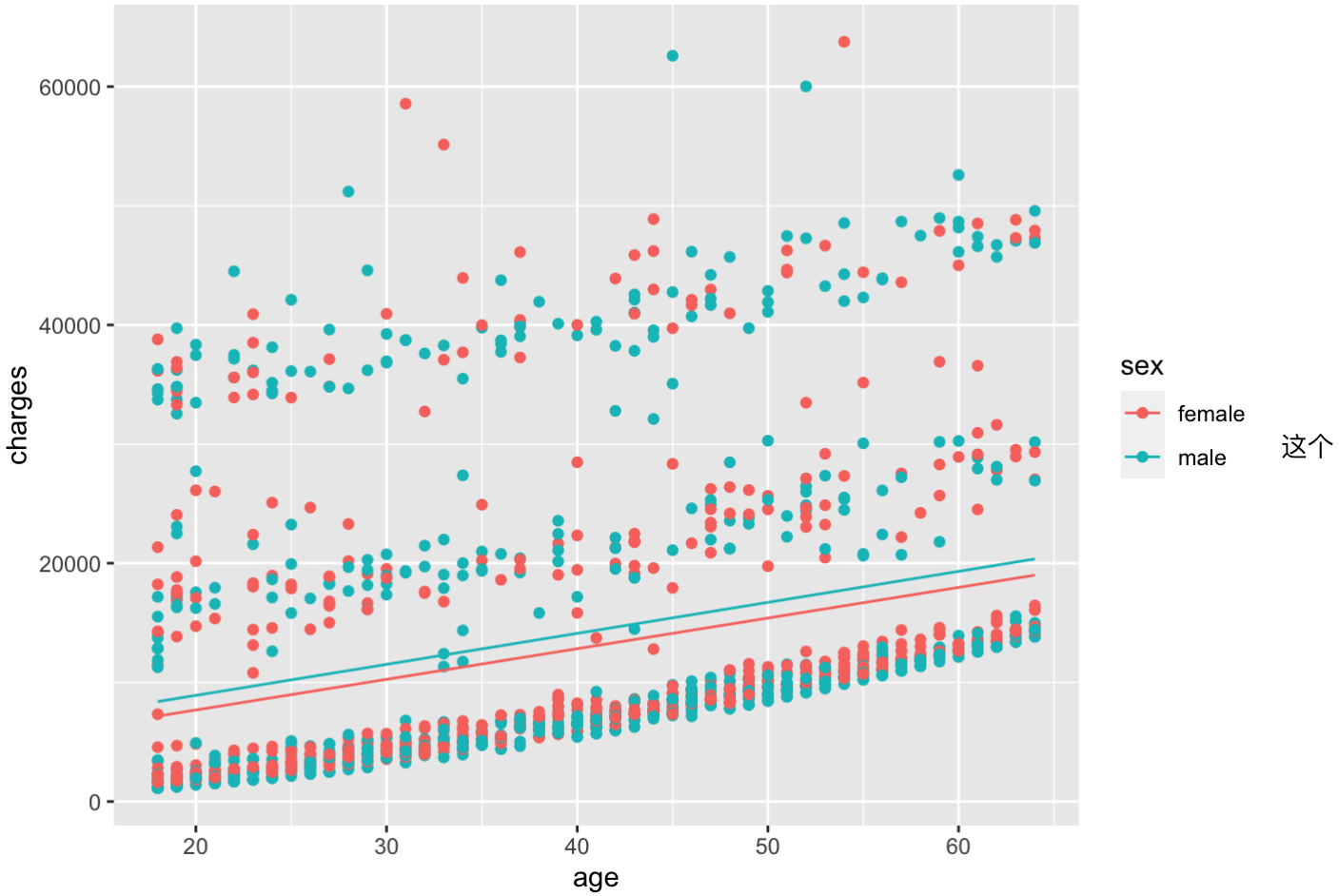
```
broom.mixed::tidy(mod4, effects = "fixed")
```

```
## # A tibble: 2 x 5
##   effect term        estimate std.error statistic
##   <chr>  <chr>          <dbl>     <dbl>     <dbl>
## 1 fixed  (Intercept)    3123.    1135.       2.75
## 2 fixed  age             259.      22.5     11.5
```

```
broom.mixed::tidy(mod4, effects = "ran_vals")
```

```
## # A tibble: 4 x 6
##   effect   group level  term        estimate std.error
##   <chr>    <chr> <chr>  <chr>          <dbl>     <dbl>
## 1 ran_vals sex   female (Intercept)  -582.     371.
## 2 ran_vals sex   male   (Intercept)   582.     368.
## 3 ran_vals sex   female age            -1.52     0.966
## 4 ran_vals sex   male   age             1.52     0.959
```

```
insu_df %>%
  add_predictions(mod4) %>%
  ggplot(aes( age, charges, color = sex)) +
  geom_point() +
  geom_line(aes(x = age, y = pred))
```



这个

模型的评价能力太不合适了。

```
mod5 <- lmer(charges ~ bmi + (1 + bmi | smoker), data = insu_df)
```

```
## boundary (singular) fit: see ?isSingular
```

```
mod5
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: charges ~ bmi + (1 + bmi | smoker)
##    Data: insu_df
## REML criterion at convergence: 27132.37
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  smoker   (Intercept) 13479.9
##           bmi           982.6  -1.00
##  Residual              6158.5
## Number of obs: 1338, groups:  smoker, 2
## Fixed Effects:
## (Intercept)          bmi
##     -3652.3        778.1
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warning
s
```

```
broom.mixed::tidy(mod5, effects = "fixed")
```

```
## # A tibble: 2 x 5
##   effect term        estimate std.error statistic
##   <chr>  <chr>          <dbl>     <dbl>     <dbl>
## 1 fixed  (Intercept)   -3652.     9571.    -0.382
## 2 fixed  bmi             778.      695.     1.12
```

```
broom.mixed::tidy(mod5, effects = "ran_vals")
```

```
## # A tibble: 4 x 6
##   effect   group  level term        estimate std.error
##   <chr>    <chr>  <chr> <chr>          <dbl>     <dbl>
## 1 ran_vals smoker no    (Intercept)    9530.     144.
## 2 ran_vals smoker yes   (Intercept)   -9530.     282.
## 3 ran_vals smoker no    bmi            -695.      10.5
## 4 ran_vals smoker yes   bmi             695.      20.5
```

该模型以不吸烟组作为参数对照，吸烟人群的bmi每增加一个单位，医疗费用支出增加982.6。

```
insu_df %>%
  add_predictions(mod5) %>%
  ggplot(aes( bmi, charges, color = smoker)) +
  geom_point() +
  geom_line(aes(x = bmi, y = pred))
```

这个模型的解释能力很好，从健康的角度考虑，吸烟与否和肥胖状况确实是健康的两大杀手，对于二者均沾的人来说，健康程度要差一些，在医疗费用的支出上高一些的可能性是合理的。

##结论 从模型上看，医疗费用支出同bmi、smoker、age这三个变量表现出较高的相关性，但这种相关性不是固定的单独线性关系，而是多重线性关系，从图形拟合中能得到验证。
形象的说这三个因素相互作用影响到了医疗费用支出的金额，医疗费用的变化收到多重因素的影响。从常识来看，一个人随着年龄的增加+肥胖严重+吸烟，那么健康问题将会非常突出，高血压、心脏病、肺病等疾病患病的概率将会大大增加，医疗费用的支出也将会大大增加。