



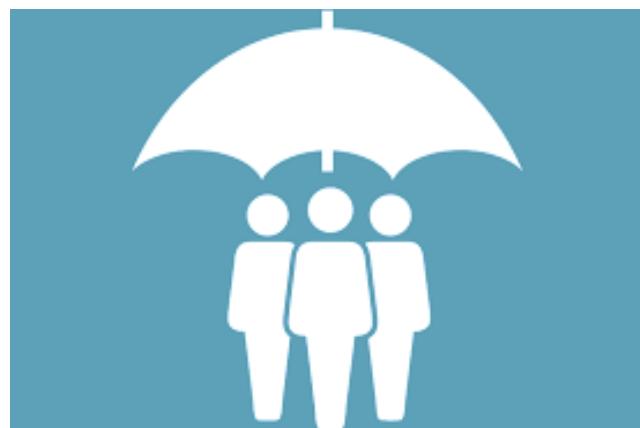
تأثیر خصوصیات افراد بر هزینه بیمه آنها

پروژه درس یادگیری آماری

ساجده اسماعیلزاده و مهشید میرهاشمی

استاد: دکتر ریخته گران

ترم اول سال تحصیلی ۱۴۰۰-۱۳۹۹



فاز اول

مصورسازی داده‌ها

Data Visualization

یک شرکت بیمه برای کسب درآمد، نیاز دارد حق بیمه‌ای که سالیانه دریافت می‌کند از هزینه‌های درمانی بیمه شدگانش بیشتر باشد. بنابراین بیمه‌ها، هزینه و زمان زیادی در تهیه مدل‌هایی که هزینه تخمینی بیمه را برای جمعیت بیمه شده به طور دقیق پیش‌بینی کند، سرمایه گذاری می‌کنند.

تخمین هزینه‌های پزشکی دشوار است، زیرا معمولاً پرهزینه‌ترین شرایط به ظاهر تصادفی هستند. اما برخی شرایط برای برخی اقسام جامعه شیوع بیشتری دارد. برای مثال سرطان ریه در میان افراد سیگاری و بیماری قلبی در میان افراد چاق بیشتر است.

هدف از این تجزیه و تحلیل، استفاده از داده‌های بیمار برای تخمین میانگین هزینه بیمه است. از این برآوردها می‌توان با توجه به هزینه‌های درمان، حق بیمه سالانه را بالاتر یا پایین تربرد.

در حال حاضر در کشور ما، بیمه‌های متفاوت متناسب با شرایط افراد وجود ندارد. این پروژه به بررسی اهمیت در نظر گرفتن شرایط افراد برای تعیین نوع بیمه می‌پردازد که هم به نفع شرکت بیمه و هم به نفع بیمه شونده می‌باشد.

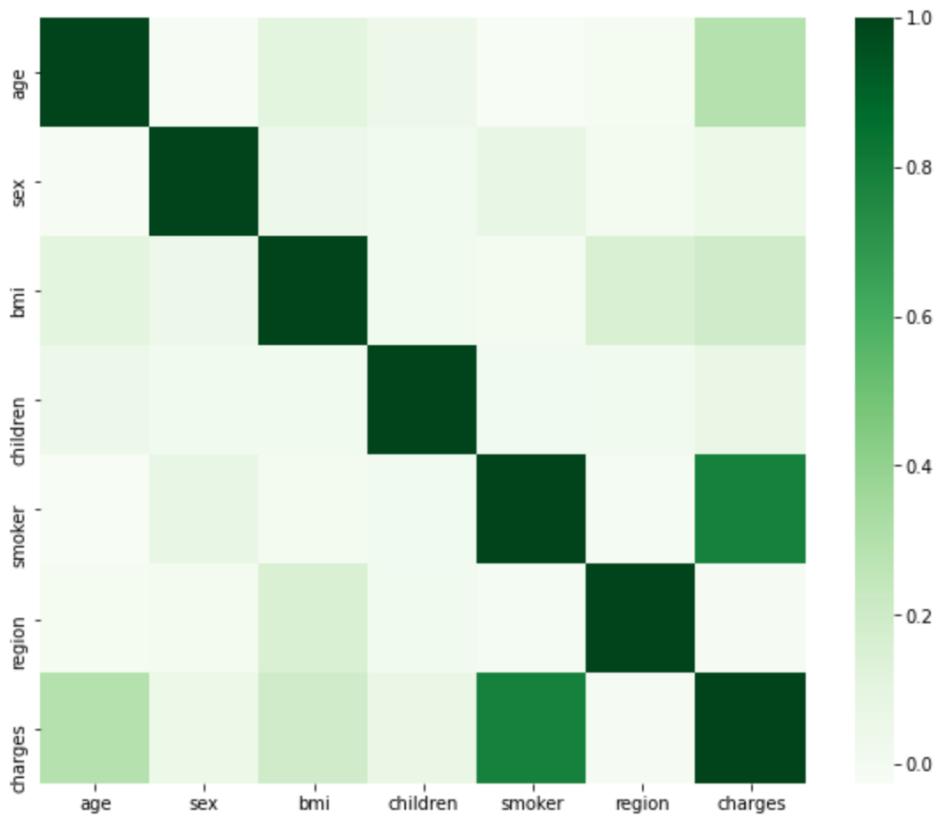
این مجموعه داده شامل اطلاعات ۱۳۳۹ فرد در کشور آمریکا است که ۷ فاکتور مختلف (متغیرها) در این افراد بررسی شده است.

<https://github.com/stedy/Machine-Learning-with-R-datasets>

معرفی متغیرها

نام متغیر	نوع متغیر	توضیح متغیر
Age	عددی و صحیح	سن
Sex	رده‌ای (مرد یا زن بودن)	جنسیت
BMI	عددی و اعشاری	شاخص توده بدنی
Children	عددی و صحیح	تعداد فرزندان (تحت تکفل بیمه)
Smoker	رده‌ای (سیگاری بودن یا نبودن)	سیگاری بودن یا نبودن
Region	رده‌ای (شمال شرق، شمال غرب، جنوب شرق و جنوب غرب)	منطقه زندگی
Charges	عددی و اعشاری	هزینه سالیانه بیمه

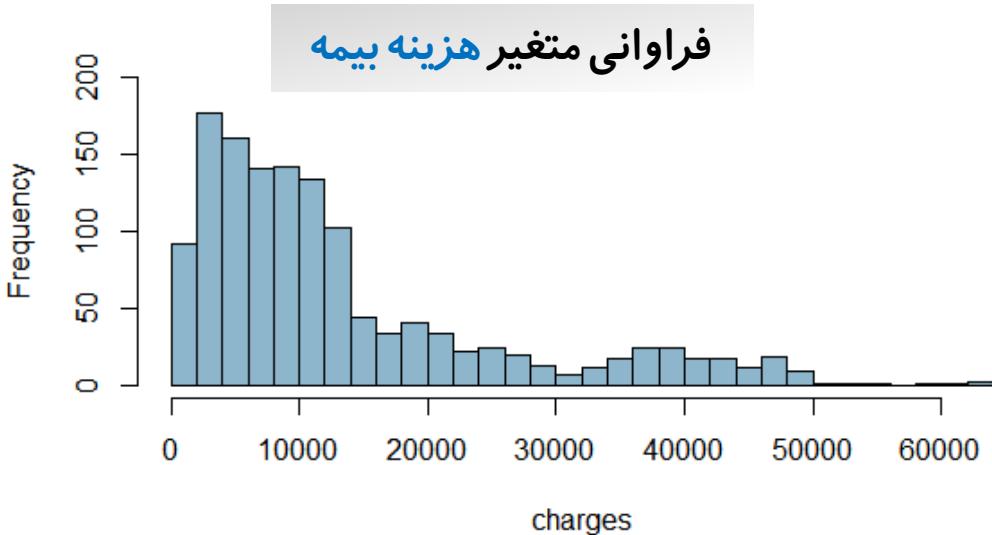
age	sex	bmi	children	smoker	region	charges	
<int>	<chr>	<dbl>	<int>	<chr>	<chr>	<dbl>	
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31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



در این شکل، همبستگی^۱ میان متغیرها را بررسی می‌کنیم. ضریب همبستگی عددی بین صفر و یک است و هرچه این عدد بیشتر باشد، میزان همبستگی دو متغیر بیشتر است. مطابق شکل، بیشترین میزان همبستگی بین متغیرهای [سیگاری بودن](#) و [هزینه بیمه](#) است.

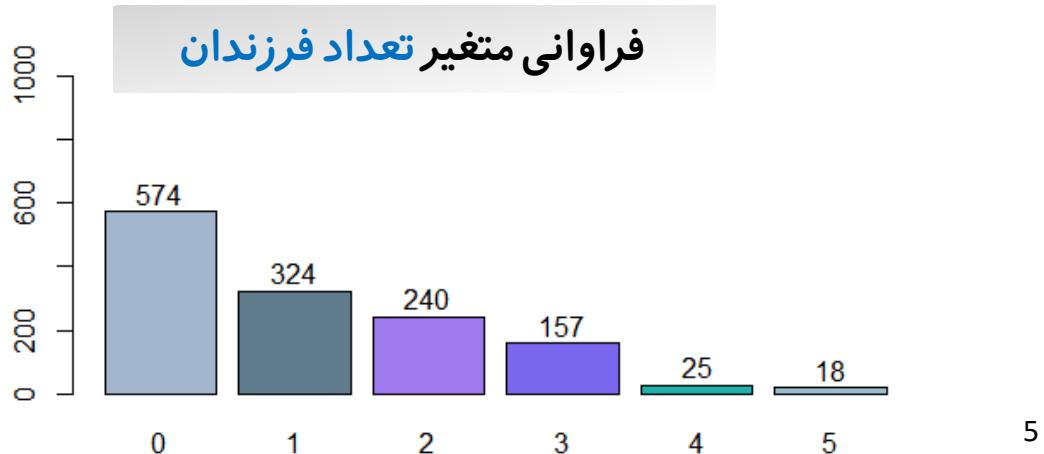
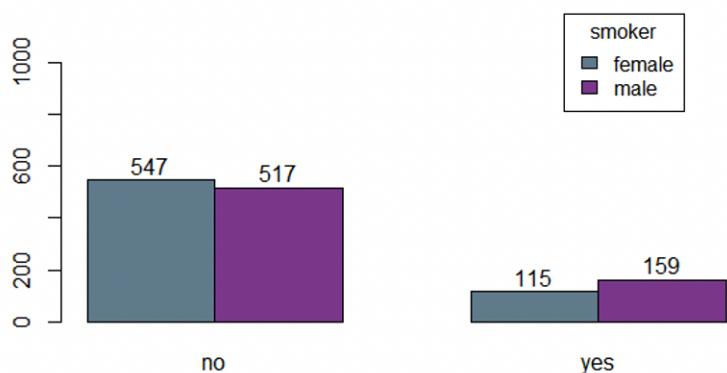
طبق بررسی داده‌ها، در این مجموعه داده هیچ داده گمشده‌ای وجود ندارد!

¹ corolation

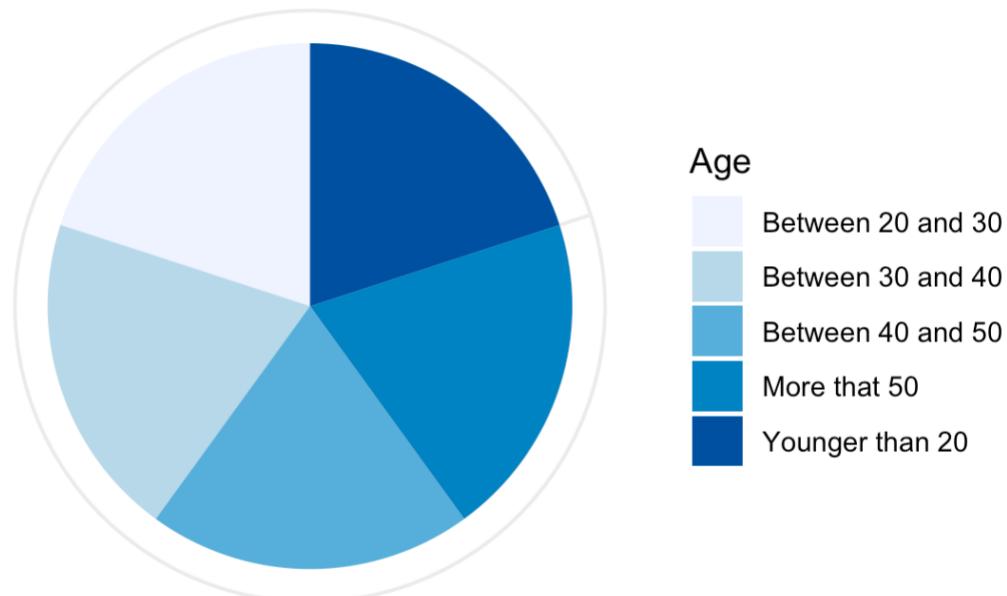


همان‌طور که در شکل بالا پیداست، وزن داده‌ها بیشتر در حدود ۵۰۰۰ تا ۱۰۰۰ بوده است. همچنین با توجه به داده‌ها، ۲۰۰۰۰ آستانه مناسبی برای تقسیم بندی افراد به کم هزینه و پرهزینه خواهد بود.

فراوانی متغیر سیگاری بودن به تفکیک جنسیت

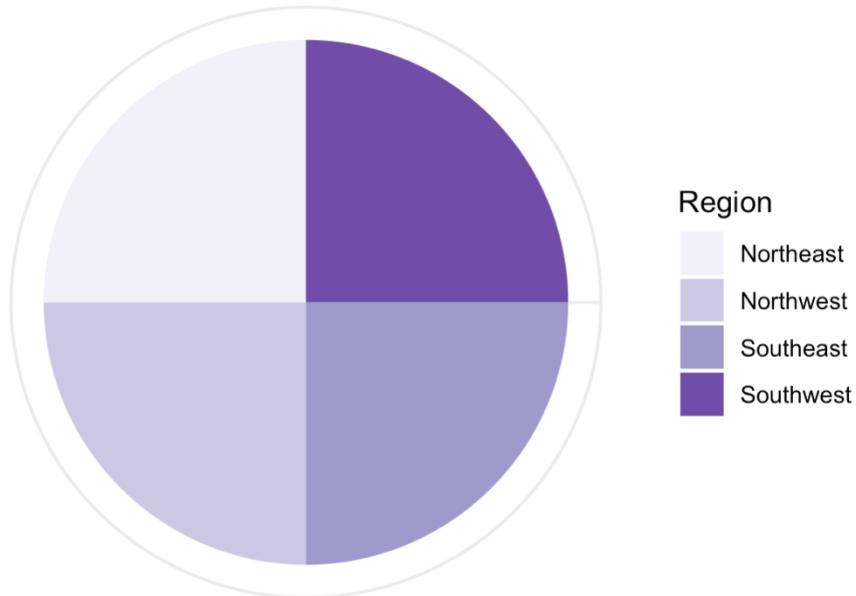


فراوانی متغیر سن افراد

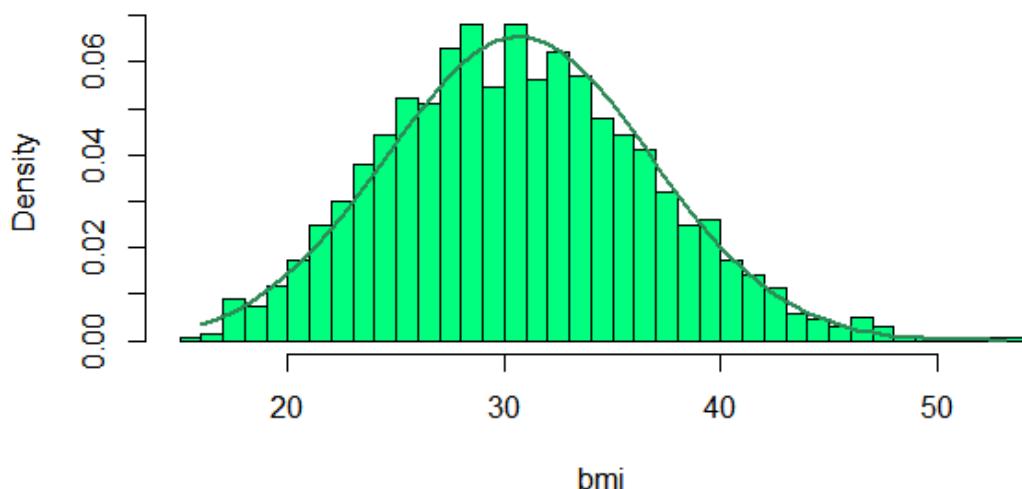


این دو نمودار نشان می‌دهند پراکندگی دو متغیر سن و منطقه زندگی افراد تقریباً برابر است.

فراوانی متغیر منطقه زندگی

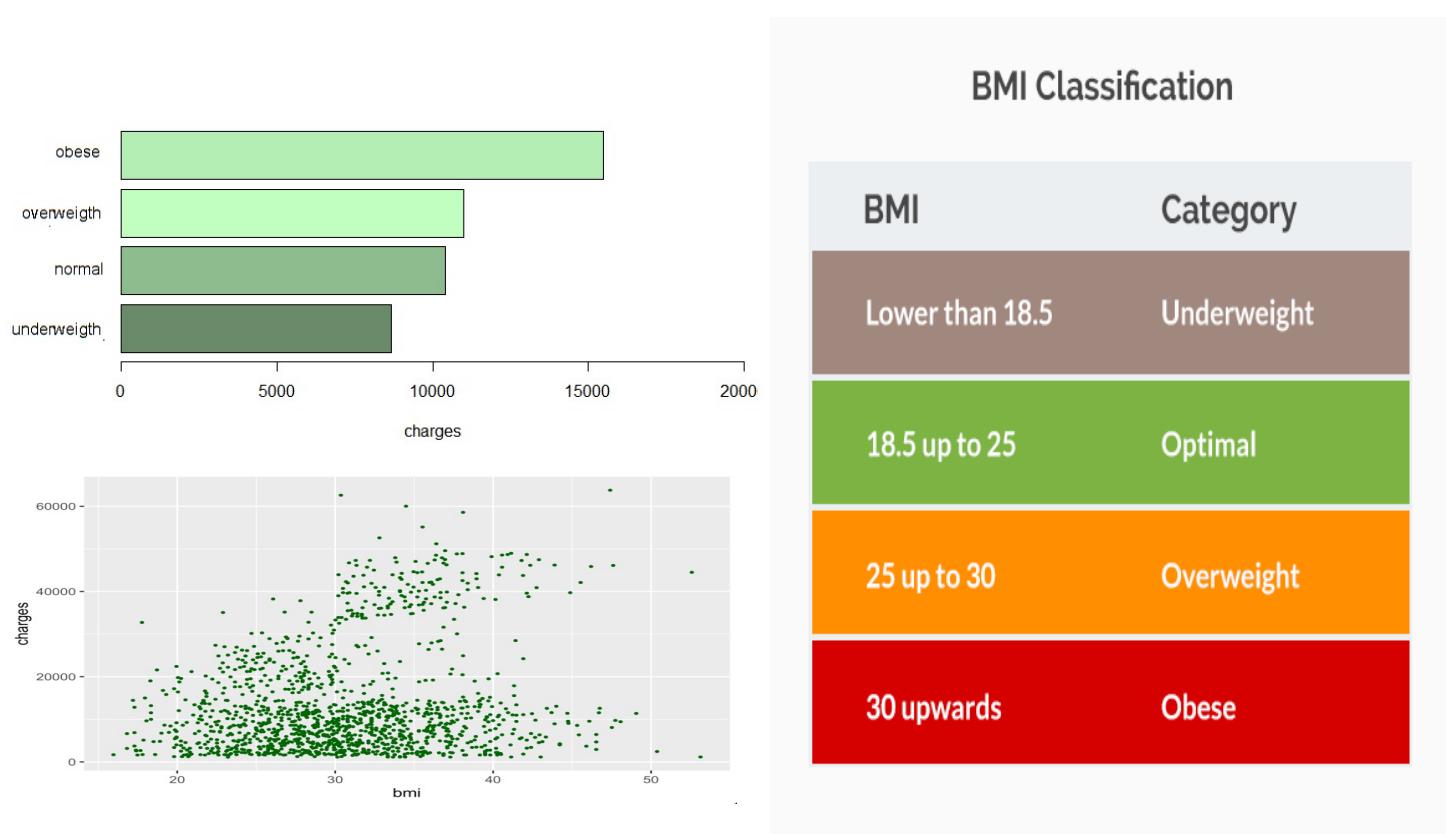


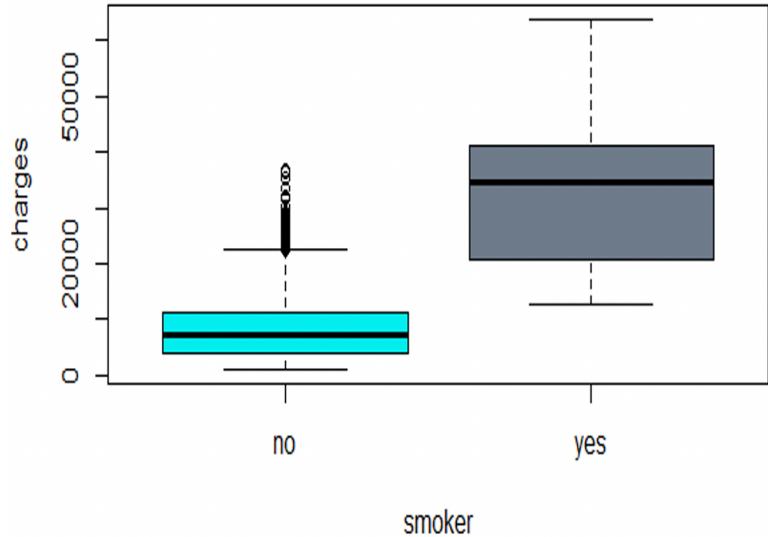
فراوانی متغیر BMI



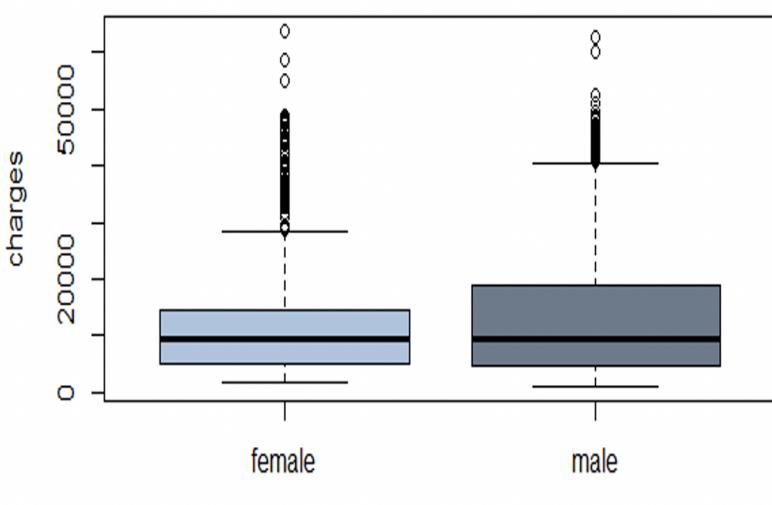
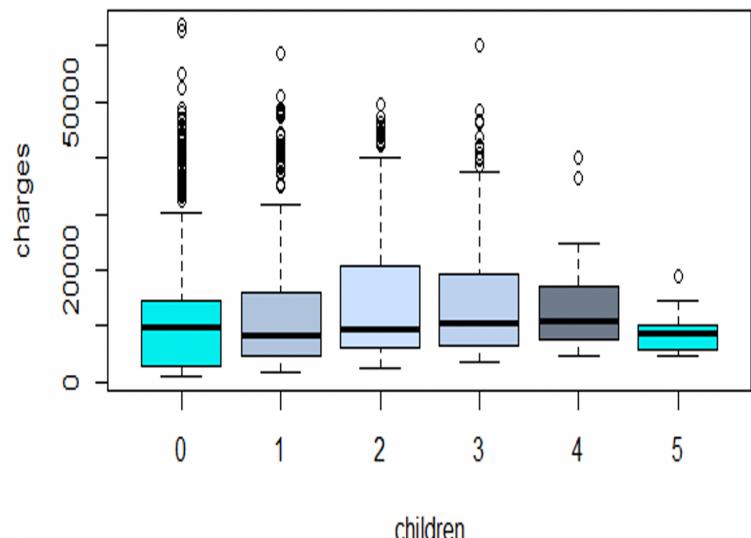
از اینجا به بعد به بررسی و تحلیل متغیرهای متفاوت روی متغیر **هزینه بیمه** می پردازیم.

با توجه به نمودار و جدول زیر، **هزینه بیمه** مطابق انتظار ما با افزایش **BMI**، زیاد می شود. همچنین با توجه به طبقه‌بندی این متغیر در زیر، می‌توان دید که برای **BMI** سی به بالا، مقدار **هزینه بیمه** به طور قابل توجهی افزایش دارد.

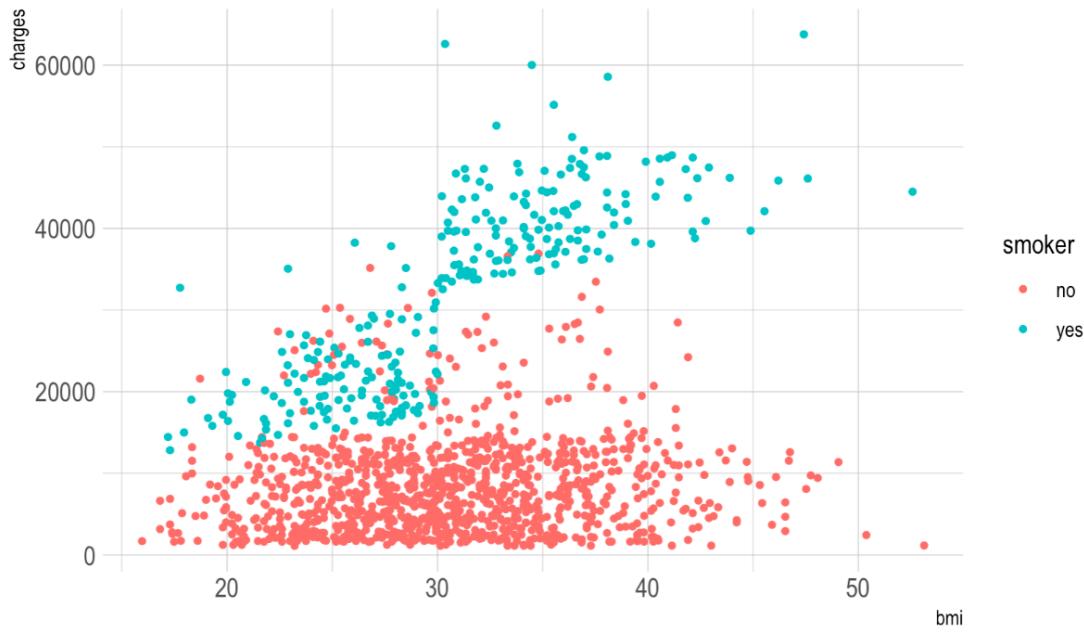




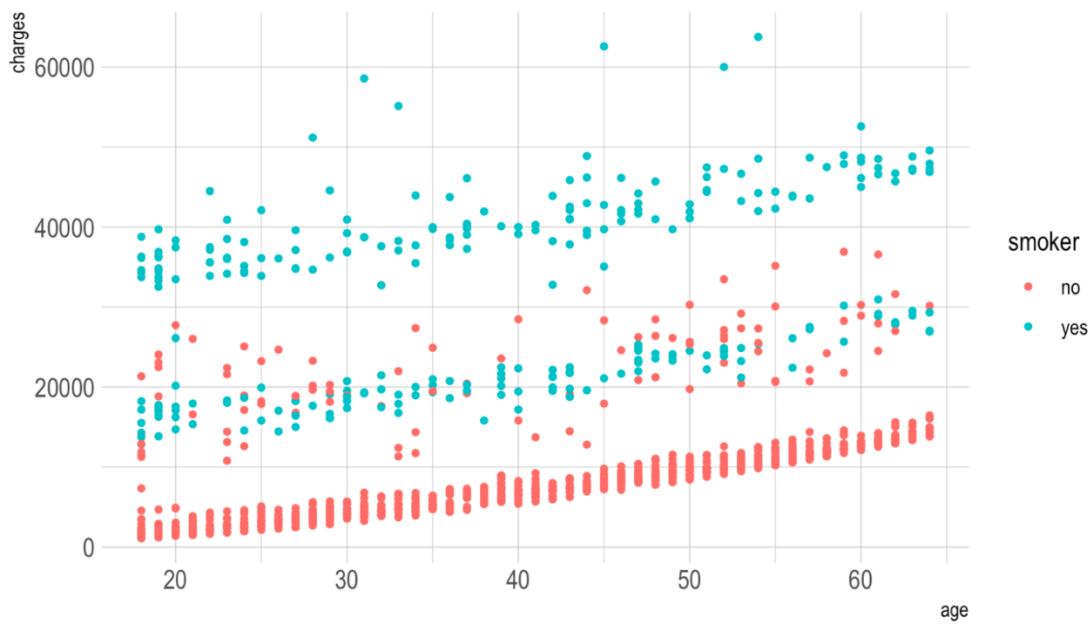
نمودار جعبه‌ای تاثیر سیگاری بودن بر هزینه درمانی
میزان هزینه افراد سیگاری خیلی بیشتر از غیرسیگاری هاست. همچنین میانه در افراد سیگاری به سمت بالاست پس بیشتر داده‌ها مقادیر کمتر از میانه اختیار کرده‌اند و چولگی منفی دارد. از طرف دیگر، میانه در افراد غیرسیگاری در وسط مستطیل قرار دارد، پس داده‌ها متقارن‌اند. همچنین در این متغیر داده‌های پرت و دورافتاده زیاد می‌باشند.

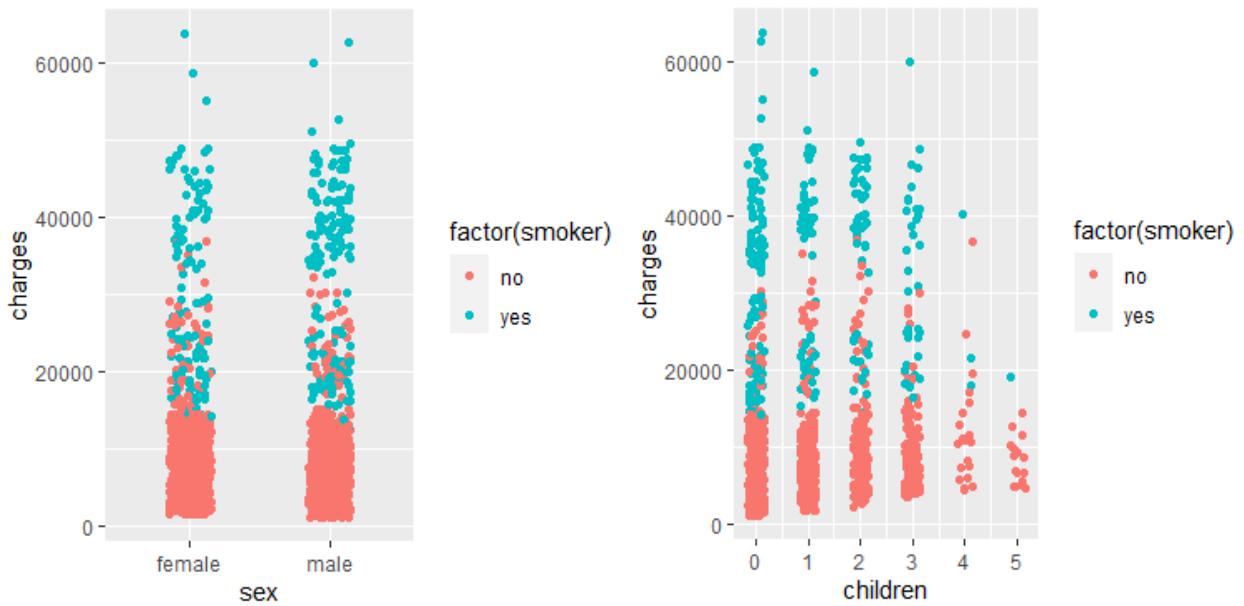


نمودار جعبه‌ای تاثیر جنسیت بر هزینه درمانی
میانه هزینه درمانی در دو گروه مرد و زن تقریباً برابر است. اما در گروه زنان میانه در وسط مستطیل قرار دارد. پس داده‌ها قرینه هستند. در گروه مردان، میانه به سمت پایین مستطیل نزدیک‌تر است. پس داده‌ها سمت مقادیر بزرگ‌تر از میانه هستند و چولگی مثبت است.

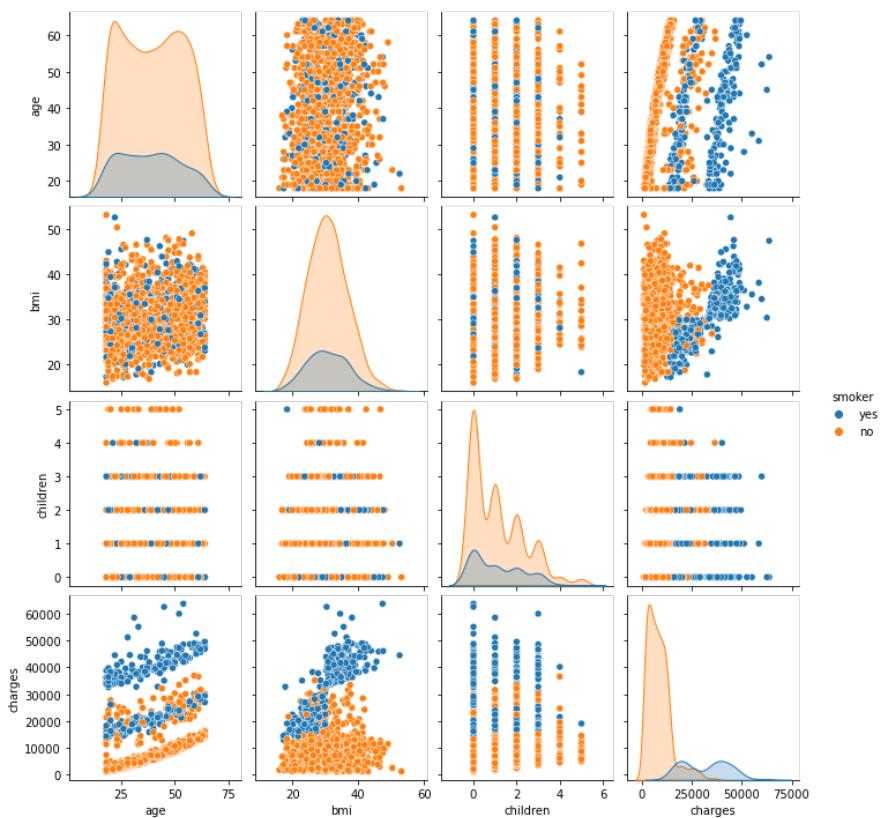


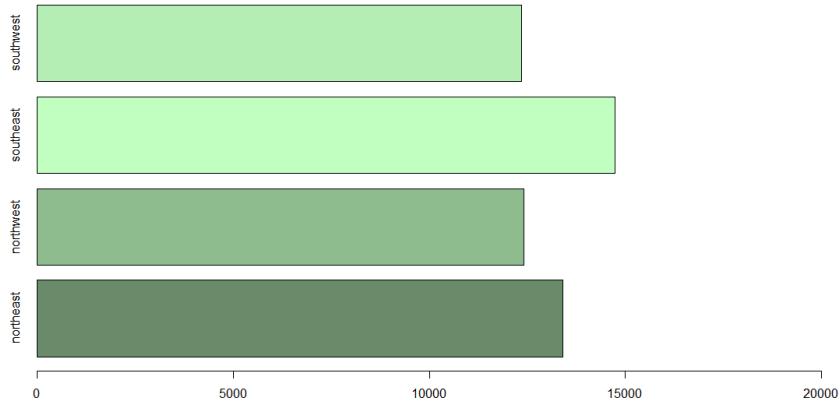
این دو نمودار تاثیر BMI و سن را با تفکیک سیگاری بودن بر هزینه بیمه نشان می‌دهد. همان‌طور که قبل نشان دادیم، همبستگی متغیر سیگاری بودن و هزینه بیمه زیاد است و این همبستگی، توجیه دو دسته شدن داده‌ها در این دو شکل است. در شکل بالا بین BMI افراد سیگاری و هزینه بیمه رابطه‌ی خطی وجود دارد و در شکل پایین می‌توان دید که، هم‌بین سن افراد سیگاری و هزینه بیمه و هم‌بین سن افراد غیرسیگاری و هزینه بیمه رابطه‌ی خطی وجود دارد.





در نمودارهای بالا نیز می‌توان تاثیر بالای سیگاری بودن بر هزینه بیمه را با توجه به دو دسته شدن داده‌ها در دو متغیر جنسیت و تعداد فرزندان مشاهده کرد. در نمودار زیر نیز، می‌توان دید که متغیر سیگاری بودن تقریباً بر همه متغیرها تاثیر گذاشته است و متغیر مهمی برای ما محسوب می‌شود. همچنین با توجه به نمودار فراوانی هزینه بیمه، می‌توان دید داده‌ها دو مده هستند. به همین دلیل ممکن است در مدل رگرسیونی، مانده‌ها نرمال نشوند.





این نمودار ، نمایان‌گر میانگین **هزینه بیمه** در هر یک از چهار **منطقه** کشور است. با توجه به این که تعداد افراد از هر **منطقه** تقریباً یکسان بود، می‌توان نتیجه گرفت **منطقه جنوب غرب، هزینه بیمه بالاتری داشته‌اند.**

نتیجه نهایی:

متغیر **سیگاری بودن** موثرترین متغیر روی متغیر **هزینه بیمه** است و پس از آن **متغیر BMI** و **سن**، بیشترین تاثیر را دارند. با توجه به هدف شرکت بیمه ، متغیر **هزینه بیمه**، متغیری مناسب و کاربردی برای پیش‌بینی است و مدل مناسب برای این تخمین ، با در نظر گرفتن پراکندگی داده‌ها و هدف ، رگرسیون خطی است. اگرچه ممکن است رگرسیون خطی به دلیل نرمال نشدن مانده‌ها، تفاسیر دقیقی نداهد. در این صورت می‌توان از روش‌های دیگری چون درخت تصمیم یا شبکه عصبی استفاده کرد.

تحلیل های پیشنهادی

- بررسی برای انتخاب بهترین متغیر هدف با توجه به داده ها و نیاز شرکت
- انتخاب هزینه درمان به عنوان متغیر هدف و تخمین آن با مدل های یادگیری با ناظارت (رگرسیون، طبقه بندی، درخت تصمیم و ...)
- با توجه به بررسی های انحصار شده، نوعی رابطه ای خطی بین داده ها دیده می شود. پس می توان از رگرسیون خطی چندگانه برای پیش بینی **هزینه بیمه** استفاده کرد.
- بررسی موثرترین متغیرها روی متغیر هدف (**هزینه بیمه**) و تفسیر مناسب تاثیر هر متغیر روی **هزینه بیمه**
- طبقه بندی افراد بر اساس متغیرهای گسسته (جنسیت، سیگاری بودن یا نبودن و منطقه زندگی) و تخمین **هزینه بیمه** برای هر طبقه به صورت جداگانه.
- با توجه به نمودار فراوانی **هزینه بیمه**، از مقدار ۲۰۰۰۰ دلار کاهش قابل توجهی دیده می شود (یا اینکه می توانیم بگوییم مد دوم داده هاست) و همچنین این افراد هزینه ای زیادی را به بیمه متحمل می شوند. پس شناسایی این افراد برای شرکت بیمه می تواند مهم باشد . با استفاده از روش های طبقه بندی می توان این افراد را شناسایی کرد .
- بررسی بالاتر بودن میانگین **هزینه بیمه** افراد سیگاری و غیرسیگاری و مقایسه آن ها
- با توجه به اینکه متغیر **سیگاری بودن** و **BMI** بالای ۳۰ در میزان **هزینه بیمه** تاثیر زیادی دارد، می توان با استفاده از درخت های تصمیم، تاثیر گذاری آنها را در هزینه ای بیمه بهتر مشاهده کرد و با استفاده از درخت تصمیم برای **هزینه ای بیمه** پیش بینی هایی به دست آورد . همچنین می توان با استفاده از درخت تصمیم، طبقه بندی گفته شده در مورد ششم را نیز انحصار داد و مقادیر تاثیرگذار در این طبقه بندی را به طور دقیق تر بررسی کرد .
- بررسی صحت اطلاعات وارد شده به خصوص در مورد متغیرهای موثر به کمک خوش بندی
- خوش بندی مناسب برای داده ها و پیدا کردن بهترین مدل روی هر خوش

فاز دوم

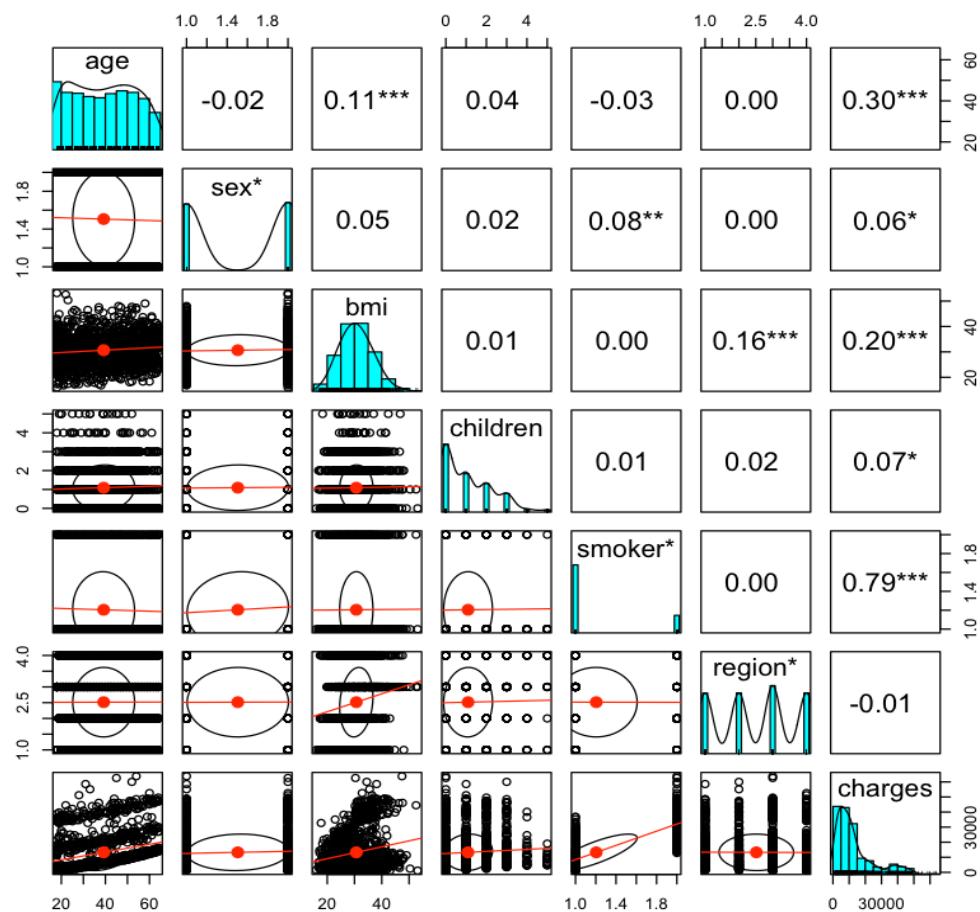
رگرسیون خطی

Linear Regression

هدف از این گزارش، پیش‌بینی مقدار **هزینه بیمه** به روش رگرسیون، با متغیرهای **سن، جنسیت، شاخص توده بدنی (BMI)، تعداد فرزندان، سیگاری بودن و منطقه زندگی** افراد می‌باشد.

ابتدا به طور رندم، ۷۰ درصد داده‌ها را به عنوان **داده آموزشی** و ۳۰ درصد را به عنوان **داده آزمایشی** در نظر می‌گیریم.

تمام مدل‌هایی که روی داده‌ها اجرا می‌کنیم، تنها روی داده‌های آموزشی ما هستند و از داده‌آزمایشی تنها برای بررسی دقیق مدل استفاده می‌کنیم و هیچ دخالتی در انتخاب مدل نهایی ما ندارند و در نهایت، با استفاده از مجموعه اعتبارسنج، مدل‌ها را با یکدیگر مقایسه می‌کنیم.
در ابتدا نگاهی کلی به فراوانی متغیرها و همبستگی آن‌ها با یکدیگر در نمودار زیر خواهیم داشت.



با توجه به ردیف آخر، در شکل‌های دوم، چهارم و ششم، خط‌های قرمز رنگ دارای شبیه قابل قبولی نیستند و ممکن است رابطه‌ی خطی با متغیر **هزینه بیمه** نداشته باشند. همچنین این متغیرها دارای کوواریانس کمی هستند.

با توجه به این نمودار، کمترین میزان همبستگی بین متغیرها و متغیر هدف، در منطقه مشاهده می شود. بنابراین این را در نظر داریم که ممکن است در حضور بقیه متغیرها، این متغیر قابل چشم پوشی باشد و بتوان آن را حذف کرد.

با استفاده از روش backward selection مدل پیشنهادی با متغیر های مورد نظر به شرح زیر است.

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
  data = 1)

Coefficients:
            (Intercept)      smokeryes        bmi       children        age
              -11578.4          23961.1         331.7          479.5         261.7
regionnnorthwest  regionsoutheast regionsouthwest
              -1035.1           -1706.9          -1243.5
```

مدل پیشنهادی را با روش forward selection به دست می آوریم.

```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
  data = 1)

Coefficients:
            (Intercept)      smokeryes        age        bmi       children
              -11578.4          23961.1         261.7         331.7          479.5
regionnnorthwest  regionsoutheast regionsouthwest
              -1035.1           -1706.9          -1243.5
```

حال با رویکرد ترکیبی نیز مدل پیشنهادی را به دست می آوریم.

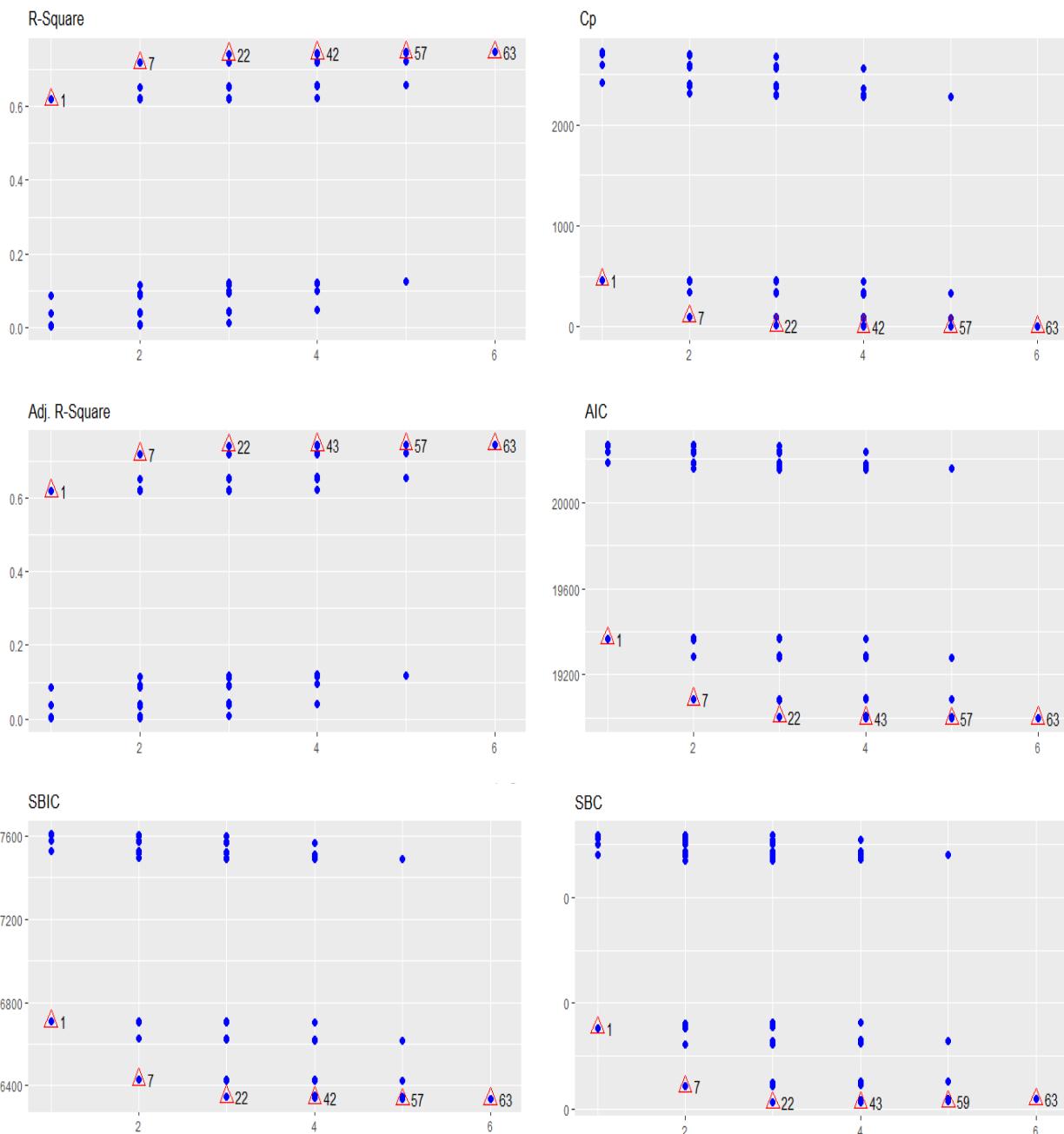
```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
  data = 1)

Coefficients:
            (Intercept)      smokeryes        age        bmi       children
              -11845.1          23924.7         262.1         307.7          465.0
```

نتایج این رویکردها، با آستانه p_value در ازای مقدار ۰.۰۱ به دست آمده اند، همچنین رویکردهای فوق با مقایسه ای مدل ها از طریق AIC نیز بررسی شد و نتایج مشابهی را به همراه داشت.

حال با توجه به کم بودن متغیر ها می توانیم از روش بهترین زیر مجموعه ای ممکن نیز استفاده کرده و نتایج را مشاهده کنیم. اما این روش در مواردی که متغیر توضیحی زیاد می باشد ، بسیار زمان برو و به عبارتی غیر ممکن است.

نتایج این رویکرد به شرح زیر است:



همان طور که در شکل ها مشاهده می شود، بهترین مدل با چهار متغیر توضیحی در رویکرد های مختلف، مدل شماره ۴۲ و ۴۳ می باشد . مدل شماره ۴۲ مدلی است که با متغیر های سیگاری

بودن، سن و منطقه برآش داده شده است و مدل شماره ۴۳ همان مدلی است که با رویکرد ترکیبی به دست آوردهیم.

مدل‌هایی که تاکنون به دست آورده‌ایم، مدل‌های ساده خطی بوده و توان دوم یا اثر متقابل بین متغیرها در آن‌ها اعمال نشده است، در صورتی که وجود این موارد می‌تواند در اثربخشی مدل، تاثیر ویژه‌ای داشته باشد. به همین دلیل با در نظر گرفتن این‌که تعداد متغیرها زیاد نیست، با اجرای مدل روی تمامی متغیرها، قدم به قدم جلو رفته و مدل را کامل می‌کنیم.

مدل اول:

```

Coefficients:
(Intercept)      sexmale      smokeryes      bmi
                 -11584.92       19.26      23958.77      331.56
regionnorthwest regionsoutheast regionsouthwest children
                 -1035.41      -1706.81      -1244.36      479.53
age
                 261.73

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

Residuals:
    Min      1Q Median      3Q     Max
-11673  -2912   -998   1374  30063

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11584.92    1208.54 -9.586 < 2e-16 ***
sexmale       19.26     404.82   0.048  0.96206
smokeryes    23958.77    501.71  47.754 < 2e-16 ***
bmi          331.56     34.47   9.620 < 2e-16 ***
regionnorthwest -1035.41    578.97 -1.788  0.07405 .
regionsoutheast -1706.81    572.49 -2.981  0.00294 **
regionsouthwest -1244.36    580.97 -2.142  0.03247 *
children        479.53    164.29   2.919  0.00360 **
age             261.73     14.45  18.109 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6146 on 927 degrees of freedom
Multiple R-squared:  0.7472,    Adjusted R-squared:  0.745
F-statistic: 342.5 on 8 and 927 DF,  p-value: < 2.2e-16

```

این مدل، با حضور تمامی متغیرهاست. میزان Adj-R-square به ما نشان می‌دهد ۷۴/۵ درصد از دلیل میزان [هزینه بیمه](#) با حضور متغیرهای بالا با ضرایب مشخص شده بوده به صورت رابطه خطی ذکر شده قابل توجیه است.

در جدول صفحه‌ی قبل، می‌بینیم که متغیر [جنسیت](#)، بالاترین p_value را داشته و نتیجه می‌گیریم که در حضور بقیه متغیرها، می‌توان متغیر [جنسیت](#) را حذف کرد. این موضوع از لحاظ شهودی هم قابل درک است. زیرا به نظر نمی‌رسد جنسیت افراد، تاثیر خاصی در میزان [هزینه بیمه](#) آن‌ها داشته باشد!

نکته دیگری که در این جدول قابل توجه است، زیاد بودن p_value متغیرهای [منطقه](#) و تعداد [فرزنдан](#) است. پس می‌توانیم با استفاده از آزمون فرض، بررسی کنیم که آیا لازم است هر سه تای این متغیرها حذف شوند یا به بیان دیگر، آیا ضریب هر سه متغیر در مدل مА صفر خواهد بود یا خیر.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	930	35375178637	NA	NA	NA
2	927	34683892998	3	691285639	6.158688 0.0003806634

در جدول بالا، مقدار p_value محاسبه شده برای حذف هر سه متغیر [جنسیت](#) و [منطقه](#) و تعداد [فرزندان](#)، تقریباً کم است و به این معناست که داده‌های ما از فرض (حذف هر سه متغیر) حمایت نمی‌کنند.

حال با استفاده از آزمون فرض، بررسی می‌کنیم که آیا لازم است دو متغیر [جنسیت](#) و [منطقه](#) حذف شوند یا خیر.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	930	35375178637	NA	NA	NA
2	931	35375216509	-1	-37872.26	0.0009956473 0.9748346

در جدول، مقدار p_value محاسبه شده برای حذف هر دو متغیر [جنسیت](#) و [منطقه](#)، تا حد مطلوبی زیاد است و به این معناست که داده‌های ما از فرض حذف هر دو متغیر حمایت کرده و حذف این دو، مدل بهتری ارائه می‌دهد.

همان‌طور که انتظار می‌رفت به مدلی رسیدیم که با استفاده از رویکرد ترکیبی و روش بهترین زیرمجموعه ممکن بدست آورده بودیم.

حال مدل جدید را با ۴ متغیر موثر بر متغیر هدف، اجرا می‌کنیم.

مدل دوم:

```
Coefficients:
(Intercept) smokeryes      bmi      children      age
-11845.1     23924.7       307.7     465.0       262.1

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

Residuals:
    Min      1Q  Median      3Q      Max 
-12154.9 -2943.2 - 961.6 1280.6 29380.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11845.12   1152.79 -10.275 < 2e-16 ***
smokeryes    23924.70    500.22  47.828 < 2e-16 ***
bmi          307.66     32.75   9.394 < 2e-16 ***
children     465.00     164.53   2.826  0.00481 ** 
age           262.06     14.49   18.086 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6164 on 931 degrees of freedom
Multiple R-squared:  0.7446,    Adjusted R-squared:  0.7435 
F-statistic: 678.6 on 4 and 931 DF,  p-value: < 2.2e-16
```

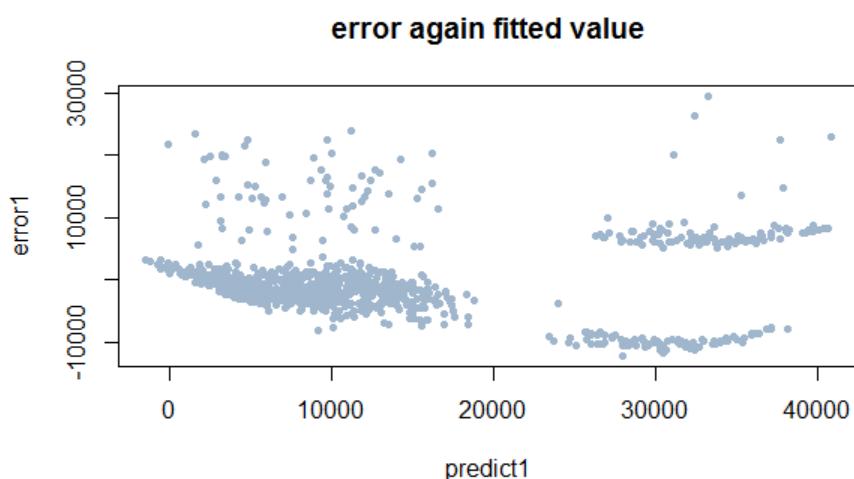
همان‌طور که در جدول بالا قابل مشاهده است، مقدار `p_value`‌ها تا حد مطلوبی کم هستند. تنها متغیر **تعداد فرزندان**، با بقیه متغیرها تفاوت قابل توجهی دارد که به دلیل آزمون فرضی که انجام شد، نیازی به حذف این متغیر نیست و این مدل، بهترین مدل خطی بدون اثر متقابل متغیرها و توان‌های بالاتر آن‌هاست.

در این مدل، مقدار `Adj-R-square` از مدل قبلی کمی کمتر است. یعنی میزان منطبق بودن داده‌ها بر مدل جدید ما (بدون متغیر **منطقه** و **جنسيت**) ۷۴.۳۵ درصد است. اما در ادامه می‌بینیم که مدل دوم، بهتر از مدل اول عمل می‌کند و مقدار `Adj-R-square` با در نظر گرفتن تفسیری که دارد،

آنچنان قابل اعتماد نیست. چون با اضافه کردن هر متغیر به مدل، این مقدار زیاد می شود. اما لزوماً متغیرهای بیشتر، مدل بهتری ارائه نمی دهند.

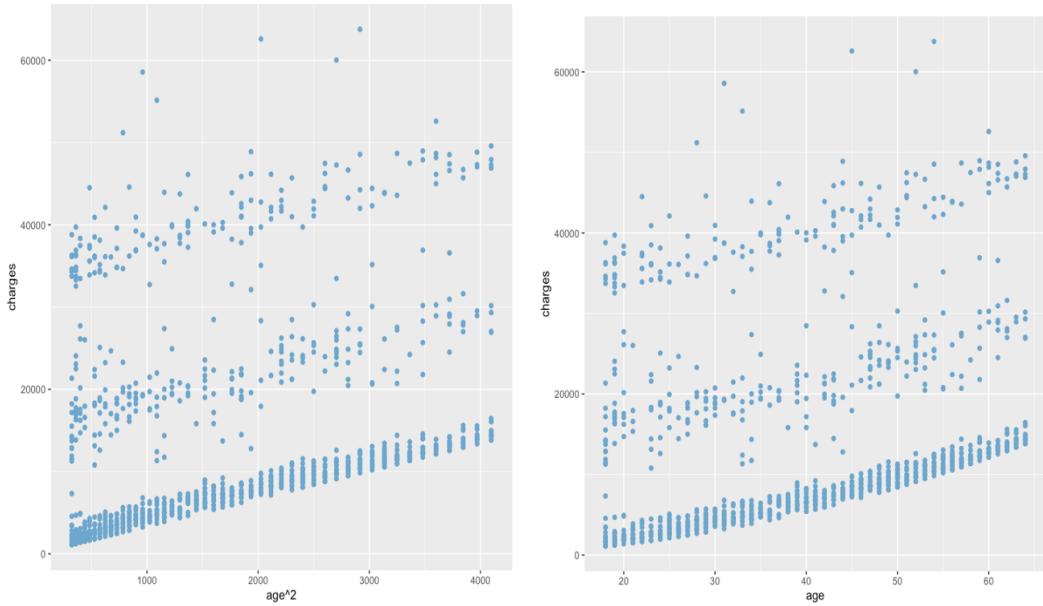
با توجه به بررسی های انحصاری شده در قسمت مصورسازی داده ها، موثر ترین متغیرها بر متغیر هدف به ترتیب **سیگاری بودن** ، **سن** و **BMI** هستند . بنابراین در ادامه، توجه ویژه ای به این سه متغیر خواهیم داشت.

با توجه به این نمودار، می توان دید که نوعی رابطه‌ی غیرخطی بین باقی مانده های مدل قبل دیده می شود و احتمالاً این ناشی از وجود نوعی رابطه‌ی غیرخطی بین داده ها می باشد .



با بررسی نمودار تاثیر **سن** بر **هزینه بیمه**، می توان فهمید این نمودار کاملاً به صورت خطی نبوده و کمی انجنا دارد. به همین دلیل، این متغیر را به توان دو رسانده و تاثیر آن را بر متغیر هدف بررسی می کنیم . این بررسی ها مقدمه ای برای شروع مدل سوم است.

همان طور که در شکلها زیر مشخص است، رابطه متغیر هدف با توان دوم سن، خطی بوده و به نظر می‌رسد مدل مناسب تری ارائه دهد.



مدل جدید را با ۴ متغیر موثر بر متغیر هدف و توان دوم متغیر سن، در نظر می‌گیریم. با اجرای این مدل می‌توان دید که p_value متغیر سن زیاد خواهد شد. پس مدلی با اضافه کردن توان دوم سن و حذف خود سن خواهیم داشت.

مدل سوم:

```

Call:
lm(formula = charges ~ smoker + bmi + children + I(age^2), data = train)

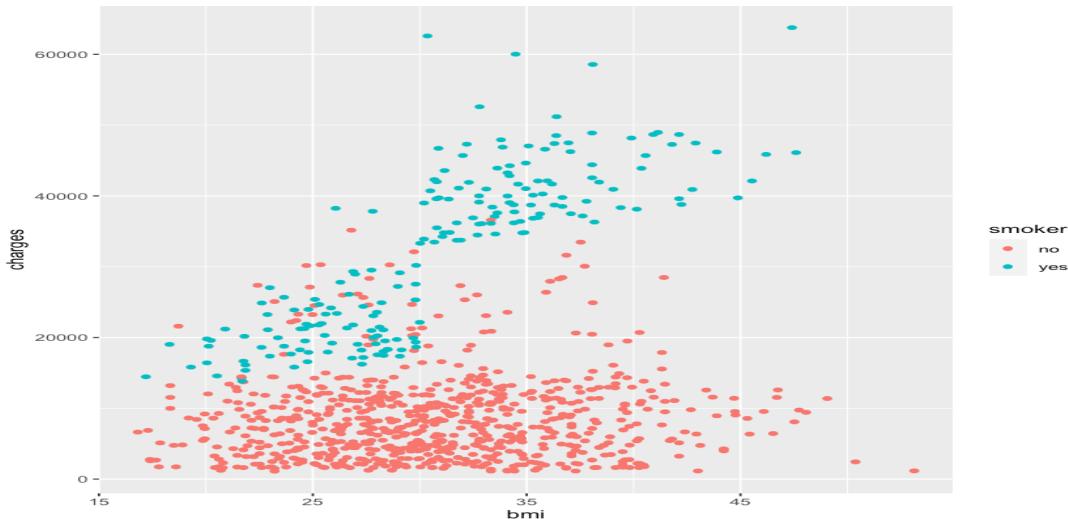
Residuals:
    Min      1Q  Median      3Q     Max 
-11446.8 -2954.6 - 978.9  1175.7 30060.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7282.0602   1066.3473  -6.829 1.54e-11 ***
smokeryes   23943.8245    498.0850  48.072 < 2e-16 ***
bmi         302.8325    32.6313   9.280 < 2e-16 ***
children    595.9954    163.6883   3.641 0.000287 ***
I(age^2)      3.2971     0.1793  18.386 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6138 on 931 degrees of freedom
Multiple R-squared:  0.7468,    Adjusted R-squared:  0.7457 
F-statistic: 686.5 on 4 and 931 DF,  p-value: < 2.2e-16

```

شکل زیر، پراکندگی داده‌ها را بر اساس مقدار **BMI** و هزینه **بیمه** به تفکیک **سیگاری بودن** نمایش می‌دهد. در این شکل تا حد زیادی داده‌ها به دو دسته **سیگاری** و **غیرسیگاری** تقسیم شده‌اند. به همین دلیل، این اثر متقابل را نیز وارد مدل کرده و بررسی می‌کنیم.



مدل چهارم که علاوه بر ۴ متغیر اولیه موثر بر متغیر هدف شامل اثر متقابل سیگاری بودن و **BMI** می‌باشد را اجرا می‌کنیم.

مدل چهارم:

```
Coefficients:
(Intercept)      smokeryes        bmi       children      age
              -2526.03     -22488.73     -6.13        516.37    268.82
smokeryes:bmi  1500.76

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

Residuals:
    Min      1Q   Median      3Q      Max 
-10995.4 -1931.2 -1355.4  -445.6 30133.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2526.03    994.25  -2.541   0.0112 *  
smokeryes  -22488.73   2007.26 -11.204  < 2e-16 *** 
bmi          -6.13      29.14   -0.210   0.8334    
children     516.37     130.24   3.965   7.9e-05 *** 
age          268.82     11.47   23.435  < 2e-16 *** 
smokeryes:bmi 1500.76    63.63   23.586  < 2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4879 on 930 degrees of freedom
Multiple R-squared:  0.8402,    Adjusted R-squared:  0.8393 
F-statistic: 978 on 5 and 930 DF,  p-value: < 2.2e-16
```

همان طور که در جدول صفحه قبل مشاهده می‌کنیم، ۸۳.۹۳ درصد از دلیل هزینه‌ی بیمه توسط مدل فوق قابل توجیه است.

در شکل صفحه قبل دیدیم متغیر سیگاری بودن و BMI، اثر متقابل دارند. اما این از مقدار ۳۰ به بالا در BMI تفکیک داده‌ها براساس سیگاری بودن یا نبودن بسیار واضح تراست. به همین دلیل ممکن است حضور اثر متقابل این دو مدل دقیق‌تری به ما ارائه دهد. توجه داریم که در صورت حضور اثر متقابل دو متغیر، مطابق Hierarchical Principle خود آن دو متغیر نیز باید در مدل حضور داشته باشند. بنابراین متغیر مقادیر ۳۰ به بالای BMI را نیز اضافه می‌کنیم.

مدل پنجم:

```

Coefficients:
(Intercept)      smokeryes          bmi           children
-4313.58         12907.92          67.89          504.94
age              bmi30             smoker_yes   bmi30:smoker_yes
268.70           -25.29            NA             572.89

Call:
lm(formula = charges ~ smoker + bmi + children + age + (bmi30) *
(smoker_yes), data = train)

Residuals:
    Min      1Q Median      3Q     Max
-16200  -1809  -1257    -468  23221

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4313.58    1297.54 -3.324 0.000921 ***
smokeryes  12907.92    533.16  24.210 < 2e-16 ***
bmi          67.89     48.59   1.397 0.162749
children    504.94    119.87   4.212 2.77e-05 ***
age          268.70    10.56   25.451 < 2e-16 ***
bmi30       -25.29    17.14  -1.475 0.140505
smoker_yes  NA        NA      NA      NA
bmi30:smoker_yes 572.89    20.26  28.270 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4490 on 929 degrees of freedom
Multiple R-squared:  0.8648,    Adjusted R-squared:  0.8639
F-statistic: 990.1 on 6 and 929 DF,  p-value: < 2.2e-16

```

در صد اطمینان در این مدل نیز با افزایش اندکی به ۸۶.۳۹ رسیده است.

تا اینجا ۵ مدل بررسی کردیم. اکنون ترکیبی از مدل سوم و چهارم را اجرا کرده و بررسی می‌کنیم.

پس مدل ششم که علاوه بر ۴ متغیر موثر بر متغیر هدف و توان دوم متغیر سن، شامل اثر متقابل سیگاری بودن و BMI می‌باشد را در نظر می‌گیریم. با حضور توان دوم سن، p_value متغیر سن بالا می‌رود. پس این متغیر را نیز حذف می‌کنیم.

مدل ششم:

```
call:  
lm(formula = charges ~ smoker + bmi + children + bmi * smoker +  
  I(age^2), data = train)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-10389.2 -1710.9 -1349.5  -866.5 30832.6  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 2.150e+03 9.317e+02  2.308  0.0212 *  
smokeryes -2.241e+04 1.995e+03 -11.231 < 2e-16 ***  
bmi        -1.054e+01 2.897e+01  -0.364  0.7162  
children    6.506e+02 1.293e+02   5.031 5.87e-07 ***  
I(age^2)    3.376e+00 1.417e-01  23.820 < 2e-16 ***  
smokeryes:bmi 1.499e+03 6.324e+01  23.699 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4849 on 930 degrees of freedom  
Multiple R-squared:  0.8421,    Adjusted R-squared:  0.8413  
F-statistic: 992.3 on 5 and 930 DF,  p-value: < 2.2e-16
```

اکنون مشابه این مدل، ترکیب مدل سوم و پنجم را نیز بررسی می‌کنیم.

در نتیجه مدل هفتم ما، شامل ۳ متغیر اولیه، توان دوم متغیر سن و اثر متقابل مقادیر بالای ۳۰ در BMI و سیگاری بودن است.

مدل هفتم:

```
call:  
lm(formula = charges ~ smoker + bmi + children + I(age^2) + bmi30 *  
    smoker, data = train)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-15384.5 -1643.0 -1246.3   -769.2  23961.0  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 154.1745 1252.6488  0.123  0.9021  
smokeryes 12917.4501  529.0004 24.419 < 2e-16 ***  
bmi         72.3214   48.2057  1.500  0.1339  
children    638.9426  118.8397  5.377 9.61e-08 ***  
I(age^2)     3.3769   0.1302 25.935 < 2e-16 ***  
bmi30       -29.0832  17.0112 -1.710  0.0877 .  
smokeryes:bmi30 573.4287  20.1068 28.519 < 2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4455 on 929 degrees of freedom  
Multiple R-squared:  0.8669, Adjusted R-squared:  0.866  
F-statistic: 1008 on 6 and 929 DF, p-value: < 2.2e-16
```

ابتدا این ۷ مدل را به طور خلاصه بیان کرده و سپس از روی جدول مرتبط با آن‌ها، بهترین مدل را انتخاب می‌کنیم.

مدل اول: حضور هر ۶ متغیر در رگرسیون خطی ساده
مدل دوم: حذف متغیر [منطقه](#) و [جنسيت](#) و رگرسیون خطی ساده با ۴ متغیر
مدل سوم: اضافه کردن توان دوم متغیر [سن](#) به مدل دوم و حذف متغیر [سن](#)
مدل چهارم: اضافه کردن اثر متقابل [سیگاری بودن](#) و [BMI](#) به مدل دوم
مدل پنجم: اضافه کردن اثر متقابل [سیگاری بودن](#) و [BMI](#) سی به بالا به مدل دوم
مدل ششم: ترکیب مدل سوم و چهارم
مدل هفتم: ترکیب مدل سوم و پنجم

	مدل اول	مدل دوم	مدل سوم	مدل چهارم	مدل پنجم	مدل ششم	مدل هفتم
Adjust R-square	0.745	0.7435	0.7455	0.8393	0.8639	0.8411	0.8659
P-value	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
MSE-train	37775771	37997010	37708654	23800814	20163443	23536692	19871328
MSE_validation LOOSV	38228851	38284612	38031361	23952917	20277677	23709535	20002346
MSE_validation K=5	38150008	38349552	38468715	24019900	20321536	23527275	20001358
MSE_validation K=10	38455928	38610830	38132638	24040963	20297039	23776060	19998989
AIC	18997.77	18999.26	18993.13	18562.41	18408.17	18552.96	18395.5
BIC	19046.18	19028.31	19027.02	18596.3	18446.9	18591.69	18439.08

با توجه به موارد بالا، بهترین مدل، مدل هفتم است که در همه موارد مطلوب‌تر از بقیه مدل‌ها به نظر می‌رسد. پس مدل نهایی ما با حضور متغیرهای **سیگاری بودن**، **BMI**، **تعداد فرزندان**، **توان دوم سن**، اثر متقابل **سیگاری بودن** و **BMI** سی به بالاست و مدل به صورت زیر می‌باشد.

$$\text{Charges} = 154.174 + 12917.450 \times \text{smoker} + 638.943 \times \text{children} + 72.321 \times \text{bmi} + 3.377 \times \text{age}^2 + 573.429 \times \text{bmi}^{30} * \text{smoker}$$

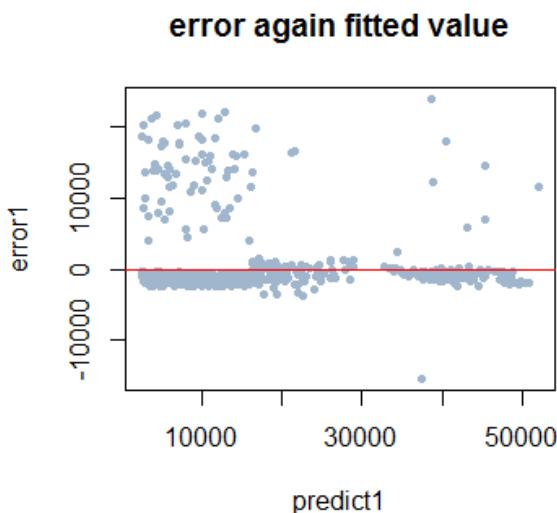
زمانی که همه متغیرها صفر باشند، Intercept مقدار **هزینه بیمه** را پیش‌بینی می‌کند. اما این ضریب با شهود ما مطابقت ندارد. زیرا هیچ داده‌ای با سن و BMI صفر وجود ندارد. به همین دلیل Intercept در داده‌های ما تفسیر درستی نخواهد داشت.

ضرایب بقیه متغیرها، کاهش یا افزایش **هزینه بیمه** با تغییر هر متغیر را نشان می‌دهند. برای مثال با افزایش یک سال به **سن** فرد و در نظر گرفتن توان دوم آن، در صورت ثابت ماندن بقیه متغیرها، به طور میانگین **هزینه بیمه** ۳.۳۷۶۹ دلار افزایش خواهد داشت. همچنین اگر همه متغیرها ثابت بمانند، با افزایش هر **فرزنده**، **هزینه بیمه** به طور میانگین ۳۸.۹۴۳ دلار افزایش دارد. به همین صورت می‌توان ضرایب بقیه متغیرها را نیز تفسیر کرد.

در مورد ضرایب متغیر گستته‌ی **سیگاری** بودن نیز می‌توان گفت با ثابت ماندن بقیه متغیرها، یک فرد **سیگاری** به طور میانگین، ۱۲۹۱۷/۴۵ دلار **هزینه بیمه** بیشتری دارد که مقدار بسیار قابل توجهی است!

متوجه **هزینه بیمه** سالیانه افرادی که **سیگاری** هستند و **BMI** آن‌ها از ۳۰ بیشتر است به طور میانگین، ۵۷۳/۴۲۹+۷۲۰.۳۲۱ دلار افزایش می‌یابد.

اکنون قصد داریم میزان دقت مدل نهایی را بررسی کنیم.



نمودار بالا، تفاوت تخمین داده **هزینه بیمه** با خطا را نشان می‌دهد. همان‌طور که در نمودار مشخص است، در مقادیر پایین‌تر هزینه درمان، تعداد قابل توجهی داده پرت وجود دارد ولی در مقادیر بالاتر، مدل تخمین نسبتاً مناسبی دارد.

فواصل اطمینان ضرایب:

(Intercept)	2.5 %	97.5 %
smokeryes	-2304.17480	2612.523743
bmi	11879.27590	13955.624324
children	-22.28334	166.926099
I(age^2)	405.71723	872.167908
bmi30	3.12140	3.632463
smokeryes:bmi30	-62.46806	4.301719
	533.96873	612.888719

همان طور که گفته شد تفسیر عرض از مبدا معنا و مفهوم خاصی ندارد . در مورد متغیر **سیگاری بودن** می توان گفت اگر فردی **سیگاری** باشد و سایر متغیر ها ثابت بماند ، تخمین می زنیم حداقل ۱۱۸۷۹.۲۷ دلار و حداکثر ۱۳۹۵۵/۶۲ دلار **هزینه بیمه** افزایش یابد . برای **BMI** با ثابت بودن بقیه می متغیرها ، تخمین می زنیم اگر یک واحد به **BMI** اضافه شود ، حداقل ۲۲.۲۸ دلار از **هزینه بیمه** کم و حداکثر ۱۶۶.۹۳ دلار به **هزینه بیمه** اضافه می شود . برای **تعداد فرزندان** با ثابت بودن بقیه می متغیرها ، اگر یک نفر به **تعداد فرزندان** اضافه شود ، تخمین می زنیم حداقل ۴۰۵.۷۱ دلار و حداکثر ۸۷۲.۱۷ دلار به **هزینه بیمه** افزوده می شود . اگر به متغیر **سن** یک واحد اضافه شود ، با در نظر گرفتن توان دوم آن و ثابت ماندن بقیه می متغیرها ، تخمین می زنیم حداقل ۱۲.۳ دلار و حداکثر ۳/۶۳ دلار به **هزینه بیمه** اضافه شود . اگر فردی **سیگاری** باشد و **BMI** او از ۳۰ بیشتر باشد ، در صورت ثابت بودن بقیه متغیرها ، پیش بینی می شود حداقل ۲۲.۲۸ دلار و حداکثر ۵۳۳.۹۶+۶۱۲.۸۹ دلار به **هزینه بیمه** افزوده می شود . به تمام تفاسیر بالا ، حدود ۹۵ درصد اطمینان داریم .

بررسی هم خطی چندگانه:

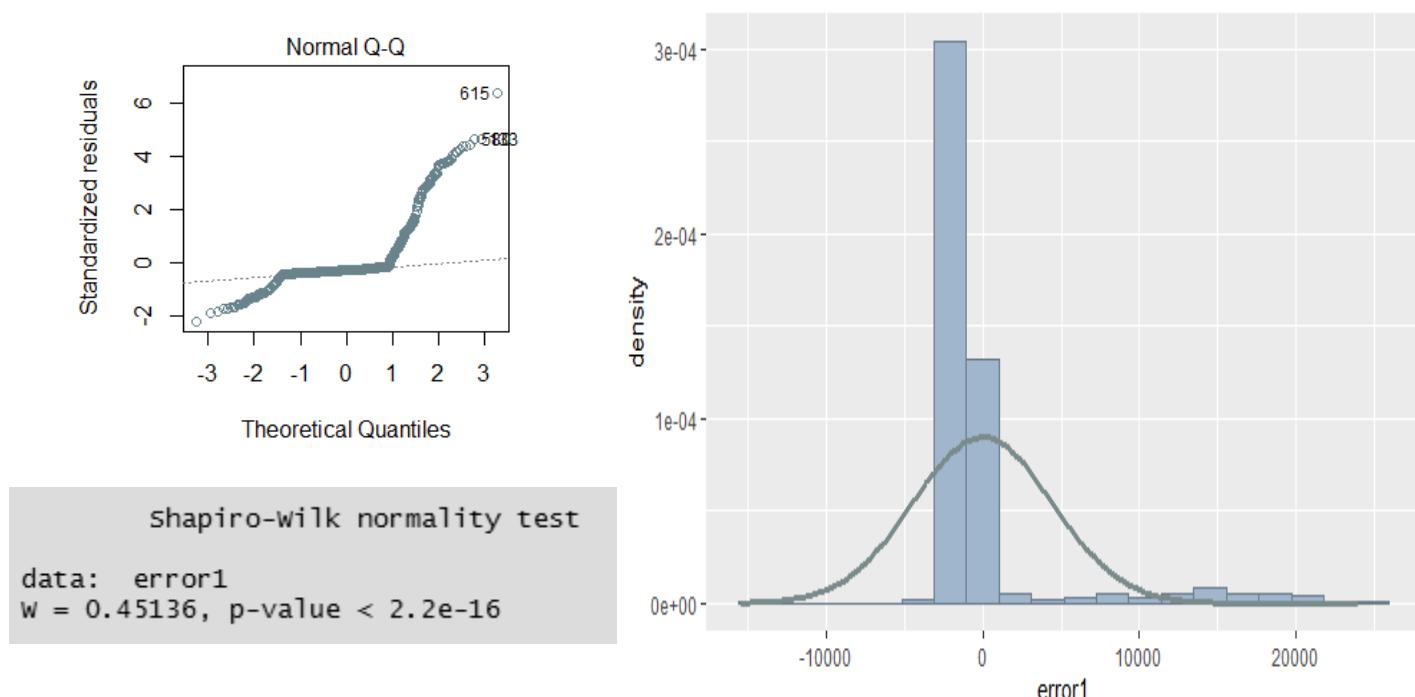
با توجه به جدول فوق از آنجایی که عامل تورم واریانس از مقدار آستانه ای ۵ فراتر نرفته است ، ارتباط قوی بین متغیرها وجود ندارد .

smoker	2.143206
bmi	4.178086
children	1.000409
I(age^2)	1.009815
bmi30	4.398967
smoker:bmi30	2.421528

نرمال بودن مانده ها:

میانگین مانده ها عدد بسیار کوچکی بوده و می توان نتیجه گرفت که تقریبا نزدیک به صفر هستند. همچنین با توجه به شکل آزمون شاپیرو والک به این نتیجه می رسیم که داده ها از فرض نرمال بودن مانده ها، حمایت نمی کنند. در نتیجه مانده ها نرمال نمی باشند ولی با رسم نمودار هیستوگرام می توان دید که توزیع داده ها به گونه ای است که در مرکز بیشتر تجمع کرده اند.

نرمال نبودن مانده ها باعث می شود آزمون های فرض و فواصل اطمینان، آنچنان قابل اعتماد نباشند. همچنین نرمال کردن متغیرها نیز در نرمال بودن مانده ها تاثیرگذار نبود. به همین دلیل، برای این داده ها، رگرسیون روش مناسبی نخواهد بود.

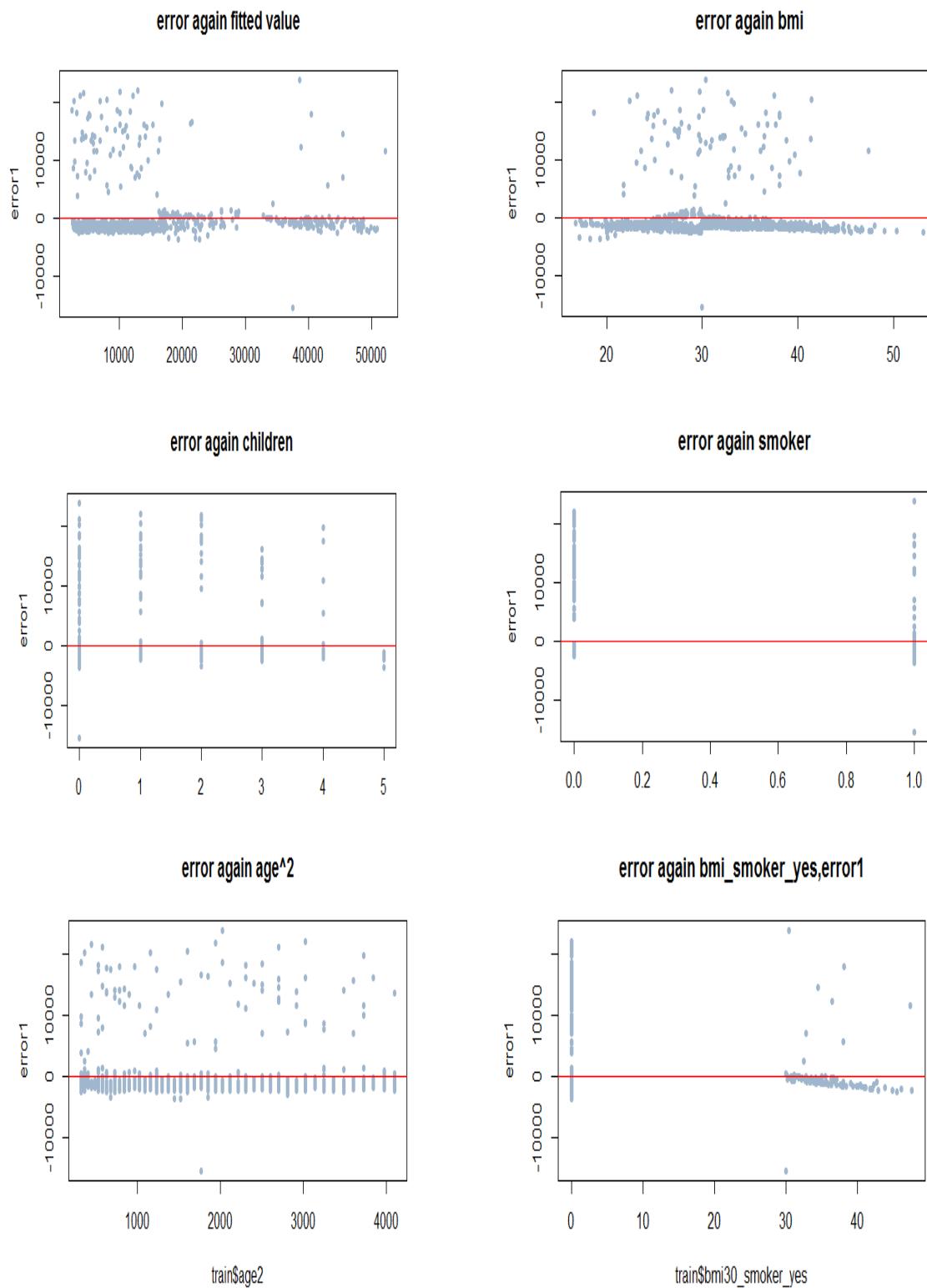


داده های پرت و تاثیرگذار:

مطابق بررسی ها، تعداد ۱۹۴ داده اهرم گون و ۴۷ داده، پرت هستند. بین این ها، ۴ داده مشترک وجود دارد که می توانند روی نتیجه نهایی تاثیر ویژه ای داشته باشند. اما با بررسی آماره کوک و DFbeta، هیچ داده تاثیرگذاری یافت نمی شود.

بررسی ثبات واریانس کواریانس:

نودارهای زیر، برای بررسی ثبات واریانس می‌باشند. همان‌طور که مشاهده می‌شود، روند خاصی در نودارها وجود ندارد. در نتیجه مانده‌ها دارای ثبات واریانس می‌باشند.



فاز سوم

طبقه بندی

Classification

هدف از این گزارش، پیش‌بینی افرادی است که **هزینه بیمه** آن‌ها از مقدار آستانه‌ای ۲۰۰۰ دلار بیشتر است با استفاده از روش‌های مختلف طبقه‌بندی، با متغیرهای **سن**، **جنسیت**، **شاخص توده بدنی (BMI)**، **تعداد فرزندان**، **سیگاری بودن و منطقه** زندگی افراد می‌باشد.

ابتدا به طور رندم، ۷۰ درصد داده‌ها را به عنوان **داده آموزشی** و ۳۰ درصد را به عنوان **داده آزمایشی** در نظر می‌گیریم.

تمام مدل‌هایی که روی داده‌ها اجرا می‌گیریم، تنها روی داده‌های آموزشی ما هستند و از داده‌آزمایشی تنها برای بررسی دقت مدل استفاده می‌کنیم.

در این فاز قصد داریم با استفاده از روش‌های رگرسیون لجستیک، تحلیل ممیزی خطی (LDA)، تحلیل ممیزی درجه ۲ (QDA) و K_نzdیک ترین همسایگی (KNN) مدل‌هایی برای پیش‌بینی برآش داده و در آخر این مدل‌ها را با یکدیگر مقایسه کرده و مدلی مناسب با داده‌ها گزارش دهیم.

به همین منظور ابتدا مدلی با روش رگرسیون لجستیک برآش می‌دهیم. اولین مدل با تمام متغیرهای **سن**، **جنسیت**، **شاخص توده بدنی (BMI)**، **تعداد فرزندان**، **سیگاری بودن و منطقه** است.

مدل اول:

```

Call:
glm(formula = hight_charges ~ age + sex + children + region +
    bmi + smoker, data = train1)

Deviance Residuals:
    Min      1Q      Median      3Q      Max 
-0.76231 -0.11054 -0.04132  0.04785  1.05004 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2857206  0.0552830 -5.168 2.89e-07 ***
age          0.0031466  0.0006546  4.807 1.79e-06 ***
sexmale      0.0234723  0.0183632  1.278  0.2015    
children     0.0038343  0.0075430  0.508  0.6113    
regionnorthwest -0.0052669  0.0259294 -0.203  0.8391    
regionsoutheast -0.0575064  0.0267573 -2.149  0.0319 *  
regionsouthwest -0.0367266  0.0263178 -1.396  0.1632    
bmi          0.0077531  0.0016225  4.778 2.05e-06 ***
smokeryes    0.7346823  0.0228526 32.149 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.07815937)

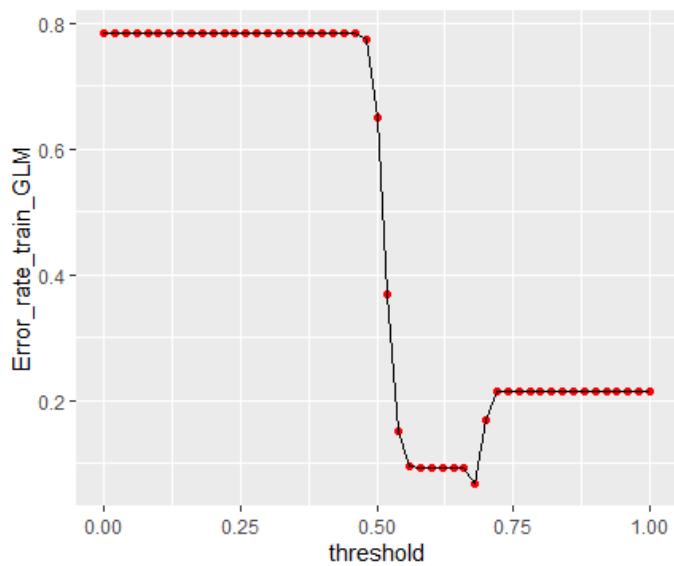
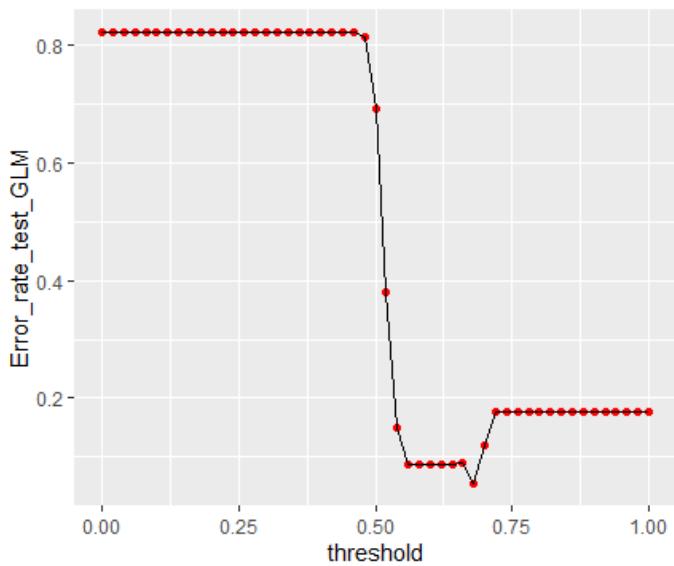
Null deviance: 158.453 on 936 degrees of freedom
Residual deviance: 72.532 on 928 degrees of freedom
AIC: 281.63

Number of Fisher Scoring iterations: 2

```

همان طور که مشاهده می کنید ، متغیر های جنسیت ، تعداد فرزندان و منطقه ، مقدار p_value از ۰.۰۱ بیشتر است و ممکن است به حضور آنها در حضور سایر متغیر ها نیاز نباشد قبل از بررسی این موضوع و برآراش مدل جدید به بررسی جوانب این مدل می پردازیم .

ابتدا باید آستانه‌ی تصمیم‌گیری را تعیین کنیم.



با توجه به شکل‌ها می‌توان دید هم در داده‌های آموزشی و هم داده‌های آزمایشی کمترین خطا در نقطه ۰/۶۸ رخ می‌دهد ، پس آستانه‌ی تصمیم‌گیری را این نقطه قرار می‌دهیم .

جدول زیر پیش‌بینی‌ها را در مقابل مقدار واقعی به ما می‌دهد .

Default Status \ Prediction		0	1
0	327	19	
1	3	52	

در جدول فوق می‌توان دید که مدل برآراش داده شده در کل ۲۲ نفر را از ۴۰۱ نفر اشتباه تشخیص داده است ، پیش‌بینی شده که ۳ نفر ، میزان هزینه بیمه بیشتری از مقدار آستانه‌ای دارند ، اما این‌طور نبوده است و پیش‌بینی شده که ۱۹ نفر ، میزان هزینه بیمه کمتری از مقدار آستانه‌ای دارند در صورتی خلاف آن بوده است .

حال با استفاده از آزمون wald بررسی می کنیم آیا می توان سه متغیر جنسیت، تعداد فرزندان و منطقه را هم زمان از مدل حذف کرد یا خیر.

```
> wald.test(b = coef(GLM), varb = vcov(GLM), Terms = 3:7)
chi-squared test:
x2 = 8.281, df = 5, P(> x2) = 0.1414
```

با توجه به نتایج به دست آمده، از آنحایی که p_value کوچکتر از ۰.۰۵ نیست، فرض صفر بودن ضرایب متغیرهای جنسیت، تعداد فرزندان و منطقه رد نمی شود و آزمون از فرض ماحمایت می کند. یعنی به حضور این متغیرها در حضور سایر متغیرها نیاز نیست.

حال مدل دوم را با حذف متغیرهای جنسیت، تعداد فرزندان و منطقه برازش می دهیم.

مدل دوم:

```
Call:
glm(formula = hight_charges ~ bmi + smoker + age, data = train1)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-0.77014 -0.10544 -0.04268  0.04072  1.06747 

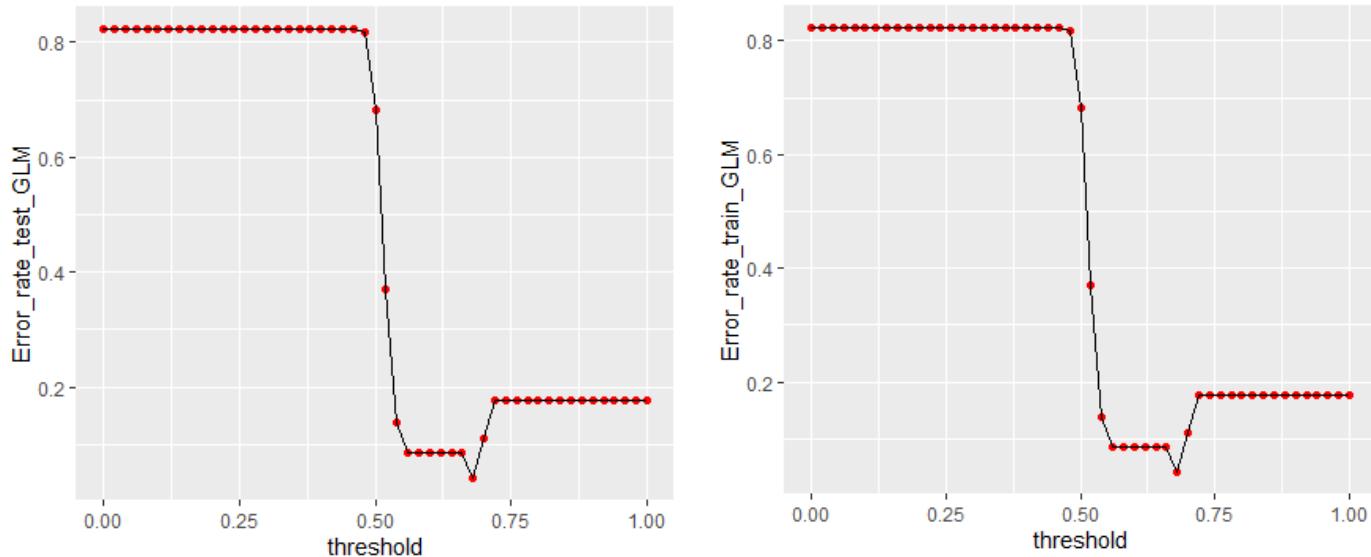
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2682833  0.0524435 -5.116 3.79e-07 ***
bmi         0.0067221  0.0015511  4.334 1.63e-05 ***
smokeryes   0.7369402  0.0227672 32.368 < 2e-16 ***
age          0.0032612  0.0006531  4.994 7.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.07843427)

Null deviance: 158.453  on 936  degrees of freedom
Residual deviance: 73.179  on 933  degrees of freedom
AIC: 279.95

Number of Fisher scoring iterations: 2
```

حال باید آستانه‌ی تصمیم‌گیری را تعیین کنیم.



در این نمودارها می‌توان دید هم در داده‌های آموزشی و هم در داده‌های آزمایشی، کمترین خطای نقطه‌ی ۶۸٪ رخ می‌دهد. پس آستانه‌ی تصمیم‌گیری را این نقطه قرار می‌دهیم.

جدول پیش‌بینی در مقابل مقادیر واقعی:

Default Status \ Prediction		0	1
0	329	16	
1	1	55	

در جدول فوق می‌توان دید که مدل برآذش داده شده در کل ۱۷ نفر را از ۴۰ نفر اشتباہ تشخیص داده است، ۱ نفر پیش‌بینی شده که میزان [هزینه بیمه](#) او از مقدار آستانه‌ای فراتر می‌رود اما اینطور نبوده و ۱۶ نفر نیز پیش‌بینی شده که میزان [هزینه بیمه](#) آن‌ها از مقدار آستانه‌ای فراتر نمی‌رود در صورتی خلاف آن بوده است.

حال مدل سوم را با استفاده از روش تحلیل ممیزی خطی، LDA، برآذش می دهیم.
از آنجایی که نمی توان متغیر های کیفی را وارد این مدل کرد مگر در مواردی که از توزیع هایی خاص استفاده کنیم، مجبور شدیم در این مدل تنها از سن و BMI برای برآذش مدل استفاده کنیم و با توجه به حذف متغیر [سیگاری بودن](#) که متغیری مهم حساب می شود انتظار می رود این مدل نتیجه ای بهتری نسبت به رگرسیون لجستیک به ما ندهد.

مدل سوم:

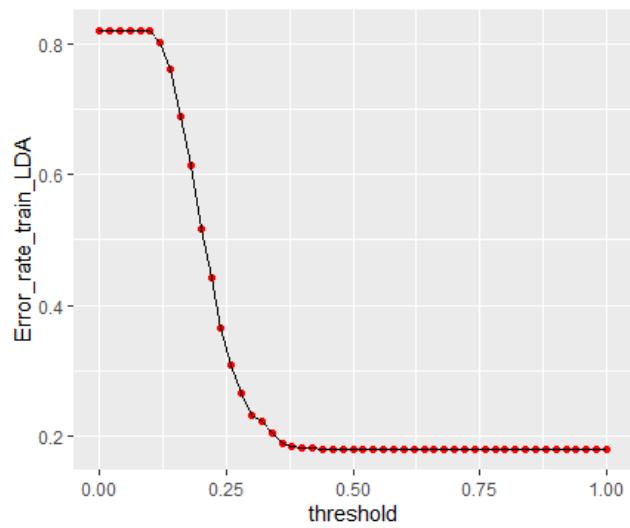
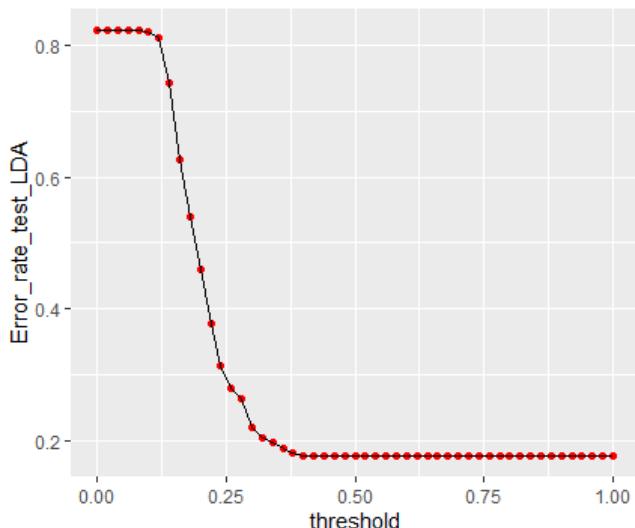
```
call:
lda(hight_charges ~ age + bmi, data = train1)

Prior probabilities of groups:
 0     1 
0.7844184 0.2155816

Group means:
      age      bmi
0 38.62721 30.31393
1 42.05941 31.88205

Coefficients of linear discriminants:
    LD1
age 0.04532371
bmi 0.11949623
```

حال با استفاده از نمودارهای زیر، آستانه ۰/۴۴ را برای تصمیم گیری انتخاب می کنیم.



جدول پیش بینی در مقابل مقادیر واقعی :

Default Status \ Prediction	0	1
0	330	71
1	0	0

در جدول فوق می توان دید که مدل LDA، ۷۱ نفر را از ۴۰۱ نفر اشتباہ تشخیص داده است و برای هر ۷۱ نفر نیز پیش بینی شده که میزان **هزینه بیمه** آنها از مقدار آستانه ای فراتر نمی رود در صورتی خلاف آن بوده است.

حال مدل چهارم را با استفاده از روش تحلیل ممیزی درجه ۲، QDA ، برازش می دهیم.
در این مدل نیز مانند روش تحلیل ممیزی خطی نمی توان متغیر های کیفی را وارد مدل کرد، مگر در مواردی که از توزیع هایی خاص استفاده کنیم. بنابراین مجبور شدیم در این مدل تنها از **سن** و **BMI** برای برازش مدل استفاده کنیم و مشابه قبل با حذف متغیر **سیگاری بودن**، انتظار می رود این مدل نیز نتیجه‌ی بهتری نسبت به رگرسیون لجستیک به ما ندهد.

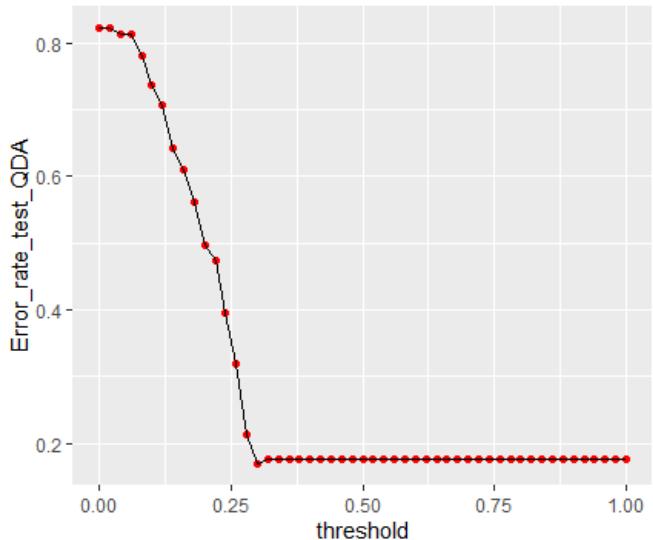
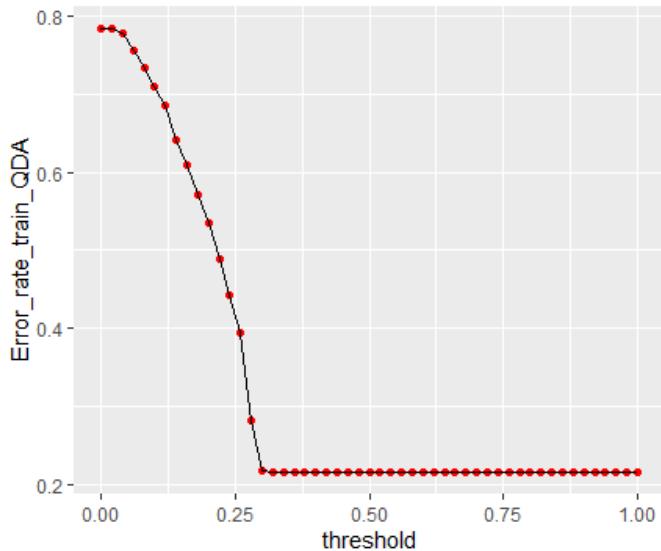
مدل چهارم:

```
call:
qda(hight_charges ~ age + bmi, data = train1)

Prior probabilities of groups:
      0      1
0.7844184 0.2155816

Group means:
      age      bmi
0 38.62721 30.31393
1 42.05941 31.88205
```

حال با استفاده از نمودارهای زیر آستانه ۳۲.۰ را برای تصمیم گیری انتخاب می کنیم .



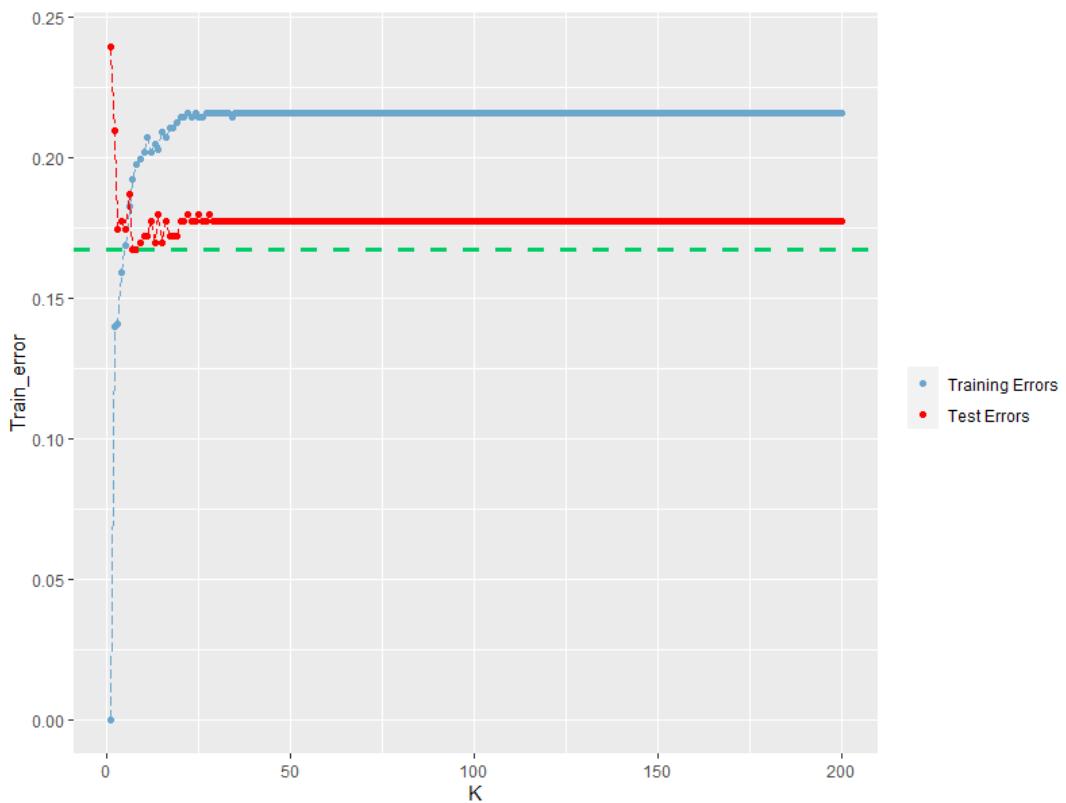
جدول پیش بینی در مقابل مقادیر واقعی :

Default Status Prediction \	0	1
0	329	67
1	1	4

در جدول فوق می توان دید که در مدل QDA، ۶۸ نفر را از ۴۰۱ نفر اشتباه تشخیص داده است. برای ۱ نفر پیش بینی شده که میزان **هزینه بیمه** از مقدار آستانه فراتر رفته در صورتی که در واقعیت فراتر نرفته است و ۶۷ نفر دیگر نیز پیش بینی شده که میزان **هزینه بیمه** آنها از مقدار آستانه ای فراتر نمی رود، در صورتی خلاف آن بوده است.

حال مدل پنجم را با استفاده از روش **K_Nزدیک ترین همسایگی**، KNN، برازش می دهیم.

ابتداًی مقدار K مناسب را با استفاده از شکل زیر انتخاب می کنیم.



با توجه به شکل، برای $K=5$ ، مقدار MSE آزمایشی به حداقل خود می رسد. اما می توان دید که برای $K=5$ ، هر دو MSE آموزشی و آزمایشی نقریباً به هم نزدیک هستند. پس این حالت را نیز در نظر گرفته و بررسی می کنیم.

مدل پنجم:

مدل را برای $K=5$ ، بررسی می کنیم. خروجی این روش به دلیل ناپارامتری بودن آن، تنها پیش بینی ها می باشد.

جدول پیش بینی در مقابل مقادیر واقعی :

Default Status \ Prediction	0	1
0	329	59
1	11	12

در جدول فوق ، می توان دید که مدل KNN ، ۷۰ نفر را از ۴۰۱ نفر اشتباہ تشخیص داده است. برای ۱۱ نفر پیش بینی شده که میزان هزینه بیمه از مقدار آستانه فراتر می رود در صورتی که در واقعیت فراتر نرفته است و ۵۹ نفر دیگر نیز پیش بینی شده که میزان هزینه بیمه آنها از مقدار آستانه ای فراتر نمی رود در صورتی خلاف آن بوده است .

مدل ششم:

جدول پیش بینی در مقابل مقادیر واقعی برای $K=8$ ، به صورت زیر است :

Default Status \ Prediction	0	1
0	325	58
1	5	13

در جدول فوق می توان دید که مدل KNN ، ۶۳ نفر را از ۴۰۱ نفر اشتباہ تشخیص داده است ، برای ۵ نفر پیش بینی شده که میزان هزینه بیمه از مقدار آستانه فراتر می رود در صورتی که در واقعیت فراتر نرفته است و ۵۸ نفر دیگر نیز پیش بینی شده که میزان هزینه بیمه آنها از مقدار آستانه ای فراتر نمی رود ، در صورتی خلاف آن بوده است .

در پایان، قصد داریم بین این ۶ مدل، بهترین مدل را با توجه به تفاسیر و AIC، خطاهای MSE های آزمایشی و آموزشی انتخاب کنیم.

ابتدا این ۶ مدل را به طور خلاصه بیان کرده و سپس از روی جدول مرتبط با آنها، بهترین مدل را انتخاب می‌کنیم.

مدل اول: حضور هر ۶ متغیر در رگرسیون لجستیک

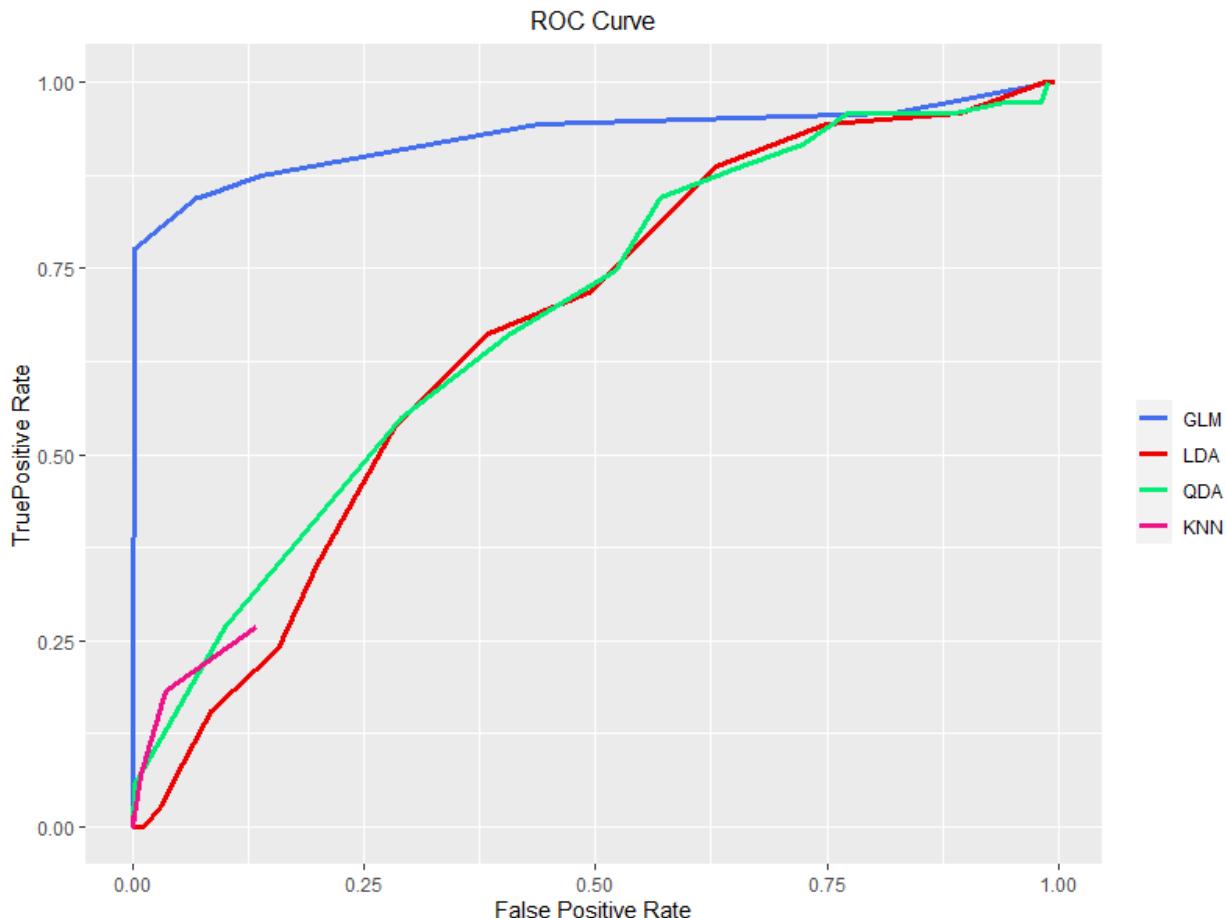
مدل دوم: حذف متغیرهای جنسیت، تعداد فرزندان و منطقه و رگرسیون لجستیک با ۳ متغیر

مدل سوم: مدل با روش LDA

مدل چهارم: مدل با روش QDA

مدل پنجم: مدل با روش KNN و $K=5$

مدل ششم: مدل با روش KNN و $K=8$



با توجه به موارد بالا، بهترین مدل، مدل دوم است که در اکثر موارد مطلوب‌تر از بقیه مدل‌ها به نظر می‌رسد. همچنین تفاوت MSE آموزشی و آزمایشی در این مدل کمتر از ۰.۰۵ می‌باشد و می‌توان نتیجه گرفت که در این مدل بیش برآذش نداریم.

همچنین با توجه به نمودار Roc، می‌توان دید که سطح زیر نمودار در رگرسیون لجستیک بیشتر از بقیه‌ی حالت‌های است و این امر انتخاب مدل نهایی ما را تایید می‌کند.

	مدل اول	مدل دوم	مدل سوم	مدل چهارم	مدل پنجم k=5	مدل ششم k=8
AIC	281.63	279.95	_____	_____	_____	_____
MSE_test	0.05486284	0.04239401	0.1770574	0.1695761	0.1745636	0.1620948
1-MSE_test	0.9451372	0.957606	0.7844184	0.8304239	0.8254364	0.8379052
MSE_train	0.09178228	0.08964781	0.2155816	0.2187834	0.167556	0.1974386
1-MSE_train	0.9082177	0.9103522	0.7844184	0.7812166	0.832444	0.8025614
Pos.pred.val	0.9454545	0.9821429	NAN	NAN	0.7586207	0.5217391
Neg.pred.val	0.9450867	0.9536232	0.8229426	0.8229426	0.84	0.8439153
خطای نوع اول	0.009090909	0.003030303	0	0	0.02857143	0.03333333
خطای نوع دوم	0.2676056	0.2253521	1	1	0.3267327	0.1690141

پس با همه‌ی این تفاسیر مدل نهایی ما به صورت زیر است:

$$\ln \frac{p(y = 1 | X = x)}{1 - p(y = 1 | X = x)} = -0.2682833 + 0.067221 \times \text{BMI} \\ + 0.7369402 \times \text{Smoker} + 0.0032612 \times \text{Age}$$

برای اینکه در تفاسیر از \ln استفاده نکنیم، e^{β_i} ها را تفسیر می‌کنیم.

از آنجایی که عرض از مبدا، معنا و مفهوم خاصی ندارد به تفسیر بقیه‌ی ضرایب می‌پردازیم.

$e^{\beta_1} = 1.07$: بخت اینکه **هزینه‌ی بیمه** از مقدار آستانه‌ای ۲۰۰۰۰ دلار فراتر رود برای فردی که به **BMI** او یک واحد اضافه شود و سایر متغیرهای توضیحی ثابت بماند ۱.۰۷ برابر بخت رخداد اتفاق، زمانی است که یک واحد به **BMI** اضافه نشده است.

$e^{\beta_2} = 2.09$: بخت اینکه **هزینه‌ی بیمه** از مقدار آستانه‌ای ۲۰۰۰۰ دلار فراتر رود برای فردی که به **سیگاری** باشد و سایر متغیرهای توضیحی ثابت بماند ۲.۰۹ برابر بخت رخداد اتفاق، زمانی است که **سیگاری** نباشد!

$e^{\beta_3} = 1.003$: بخت اینکه **هزینه‌ی بیمه** از مقدار آستانه‌ای ۲۰۰۰۰ دلار فراتر رود، برای فردی که به **سن** او یک واحد اضافه شود و سایر متغیرهای توضیحی ثابت بماند ۱.۰۰۳ برابر بخت رخداد اتفاق، زمانی است که یک واحد به **سن** اضافه نشده است.

حال به تفسیر برآورد فاصله‌ای ضرایب در مدل نهایی می‌پردازیم:

	2.5 %	97.5 %
(Intercept)	-0.371070674	-0.165496011
bmi	0.003681960	0.009762262
smokeryes	0.692317260	0.781563202
age	0.001981292	0.004541206

همان‌طور که گفته شد، تفسیر برای عرض از مبدا معنا و مفهومی ندارد.

BMI : بخت اینکه هزینه‌ی بیمه از مقدار آستانه‌ای ۲۰۰۰ دلار فراتر رود برای فردی که به **BMI** او یک واحد اضافه شود و سایر متغیرهای توضیحی ثابت بماند، حداقل 1.00^3 برابر و حداکثر 1.00^1 برابر بخت رخداد اتفاق، زمانی است که یک واحد به **BMI** اضافه نشده است و به این حرف خود درصد اطمینان داریم.

Smoker سیگاری باشد و سایر متغیرهای توضیحی ثابت بماند، حداقل 1.00^1 برابر و حداکثر 2.00^{18} برابر بخت رخداد اتفاق، زمانی است که سیگاری نباشد و به این حرف خود 95^5 درصد اطمینان داریم!

Age : بخت اینکه هزینه‌ی بیمه از مقدار آستانه‌ای ۲۰۰۰ دلار فراتر رود برای فردی که به **سن** او یک واحد اضافه شود و سایر متغیرهای توضیحی ثابت بماند، حداقل 1.00^1 برابر و حداکثر 1.00^4 برابر بخت رخداد اتفاق، زمانی است که یک واحد به **سن** اضافه نشده است و به این حرف خود درصد اطمینان داریم.

فاز چهارم

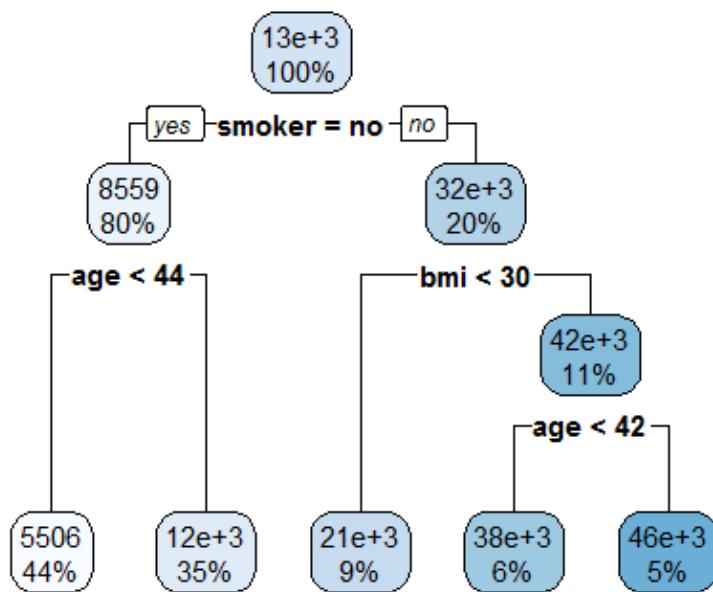
درخت تصمیم

Decision Tree

هدف از این گزارش، پیش‌بینی **هزینه بیمه** با استفاده از درخت های تصمیم و مقایسه ای نتایج آن با رگرسیون خطی می باشد . برای این منظور ابتدا داده ها را به دو دسته ای آموزشی و آزمایشی تقسیم می کنیم.

مدل اول:

ابتدا درختی با ساده ترین حالت ممکن برآذش می دهیم .



حال به تفسیر این درخت از چپ به راست می پردازیم . در ابتدا می توان دید که متغیر **سیگاری بودن** مهم ترین متغیر در تعیین میزان **هزینه بیمه** افراد است و افراد سیگاری **هزینه بیمه** بیشتری دارند .

اولین شاخه ای درخت نشان دهنده ای این است که فردی **سیگاری نباشد** و سن وی از 44 سال کمتر باشد، متوسط **هزینه بیمه** او، ۵۵۰۶ دلار پیش‌بینی می شود. توجه کنید که ۴۶ درصد از کل داده های آموزشی در این شاخه قرار دارند .

دومین شاخه نشان می دهد که اگر فرد **سیگاری نباشد** و سن او از ۴۴ سال بیشتر باشد ، پیش بینی می شود که متوسط **هزینه بیمه** وی ۱۲۰۰۰ دلار باشد. ۳۵ درصد کل داده های آموزشی در این دسته قرار دارند .

سومین شاخه افرادی که سیگاری باشند و **BMI** آنها از ۳۰ کمتر است را نشان داده و پیش‌بینی می‌شود که متوسط هزینه‌ی بیمه آنها ۲۱۰۰۰ باشد. همچنین ۹ درصد کل داده‌های آموزشی در این دسته قرار دارند.

چهارمین شاخه نشان دهنده افرادی است که سیگاری باشند و **BMI** آنها از ۳۰ بالاتر و **سن** آنها کمتر از ۴۲ سال باشد، پیش‌بینی می‌شود که متوسط هزینه‌ی بیمه آنها ۳۸۰۰۰ دلار باشد. ۶ درصد کل داده‌های آموزشی در این دسته قرار دارند.

در آخر برای افرادی که سیگاری هستند و **BMI** آنها از ۳۰ بالاتر و **سن** آنها بیشتر از ۴۲ سال است، پیش‌بینی می‌شود که متوسط هزینه‌ی بیمه آنها ۴۶۰۰۰ دلار باشد. همچنین ۵ درصد کل داده‌های آموزشی در این دسته قرار دارند.

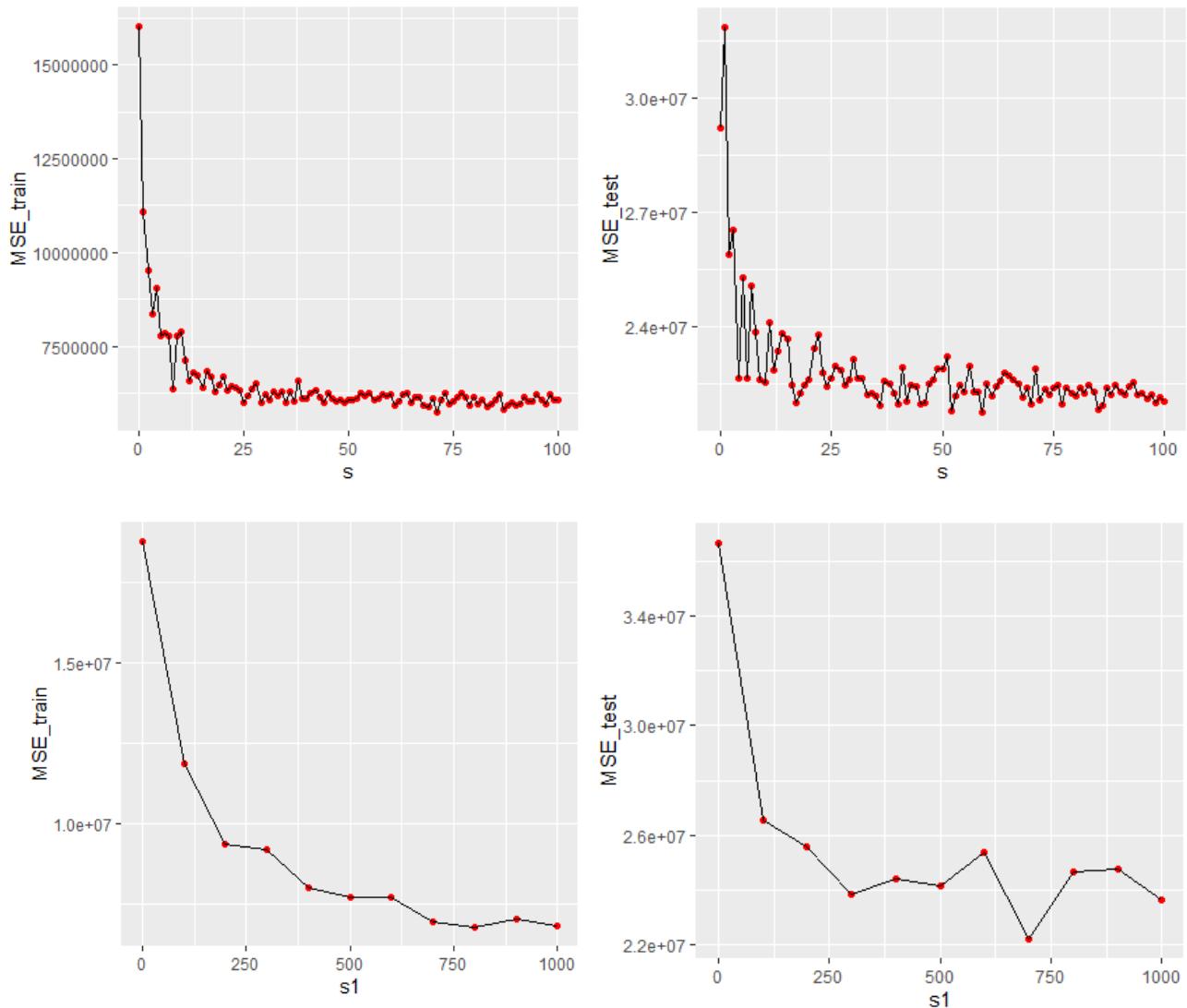
درجول‌های زیر می‌توان پیش‌بینی ۶ داده‌ی اول آموزشی و آزمایشی را در مقایسه با مقدار واقعی در یک نگاه مشاهده کرد.

train	1	2	3	4	5	6
prediction	5506.289	12376.124	21204.291	38207.814	5506.289	12376.124
charges	1526.312	11729.680	23887.663	33732.687	6799.458	7804.160

test	1	2	3	4	5	6
prediction	12376.124	21204.291	12376.124	5506.289	12376.124	5506.289
charges	11454.022	21677.283	9778.347	5757.413	10381.479	1725.552

همان‌طور که مشاهده می‌کنید این درخت تفسیرپذیری بالایی دارد اما برای پیش‌بینی کردن مناسب نیست، پس با استفاده از روش‌های **RandomForest**، **Bagging**، **Boosting** و سعی می‌کنیم قابلیت پیش‌بینی آن را افزایش دهیم. ابتدا به بررسی روش **Bagging** می‌پردازیم.

برای انتخاب تعداد درخت‌ها، از نمودارهای زیر استفاده می‌کنیم.



مدل دوم:

باتوجه به دو نمودار اول، از آنجایی که روی داده‌ی آموزشی از ۵۰ به بعد تغییرزیادی در MSE دیده نمی‌شود، کمترین مقدار MSE آزمایشی را که با ۵۹ درخت است، برای برآش درخت انتخاب می‌کنیم.

train	1	2	3	4	5	6
Prediction	2768.285	11819.727	23679.835	34061.466	8610.770	10050.549
charges	1526.312	11729.680	23887.663	33732.687	6799.458	7804.160
test	1	2	3	4	5	6
prediction	11671.570	21094.591	12178.720	10906.614	12483.153	4326.218
charges	11454.022	21677.283	9778.347	5757.413	10381.479	1725.552

در دو نمودار دوم، تعداد مناسب برای درخت ها ۷۰۰ تا است و نتایج مدل به صورت زیر خواهد بود.

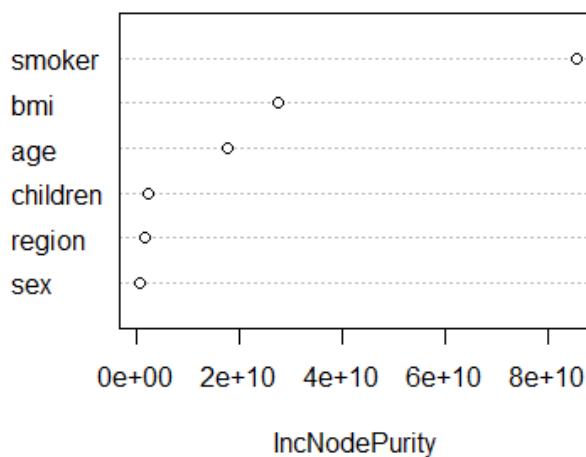
مدل سوم:

train	1	2	3	4	5	6
Prediction	4012.018	12061.609	23697.378	33993.422	8955.946	9662.201
charges	1526.312	11729.680	23887.663	33732.687	6799.458	7804.160

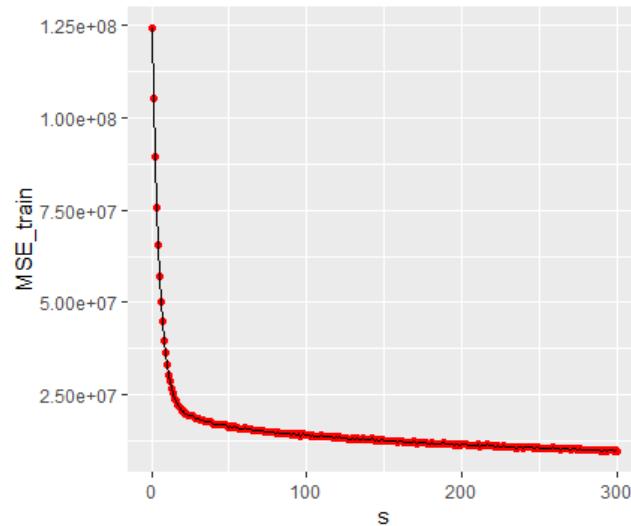
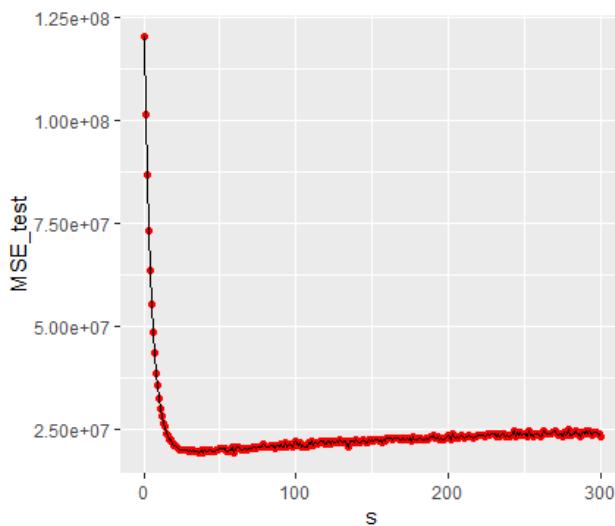
test	1	2	3	4	5	6
prediction	12074.502	21267.032	12045.303	10866.574	11409.558	4455.145
charges	11454.022	21677.283	9778.347	5757.413	10381.479	1725.552

با توجه به نمودار زیر دو متغیر **سیگاری بودن** و **BMI** در این روش بیشترین تاثیر را دارند.

bagging



حال با دو متغیری که به صورت رندوم از متغیرها انتخاب می شود، **RandomForest** را تشکیل می دهیم. با توجه به نمودار های زیر تعداد درخت ها را ۹۰۰ انتخاب می کنیم.



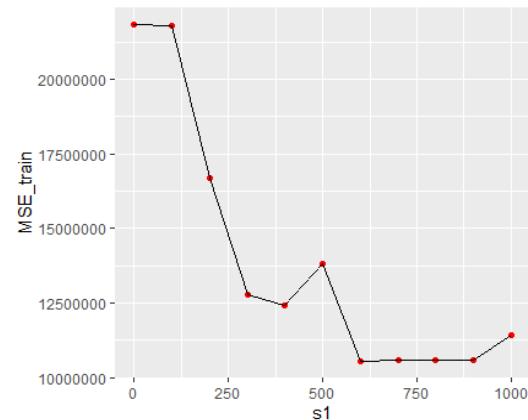
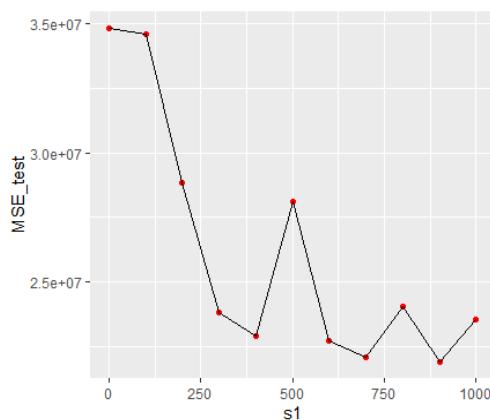
مدل چهارم:

نتایج به صورت زیر است :

train	1	2	3	4	5	6
prediction	4831.683	11863.555	24278.046	33026.596	7680.019	9398.886
charges	1526.312	11729.680	23887.663	33732.687	6799.458	7804.160

test	1	2	3	4	5	6
prediction	12392.002	23123.450	12005.162	9984.100	12071.997	3791.142
charges	11454.022	21677.283	9778.347	5757.413	10381.479	1725.552

اکنون از روش Boosting درخت مناسب را انتخاب می کنیم. قبل از اینکه روشن فوچ را اجرا کنیم، بهترین تعداد درخت را از روی نمودارهای زیر انتخاب می کنیم.



مدل پنجم:

با توجه به بررسی های انحصار شده تعداد درخت ها را ۳۷۱ انتخاب کردیم و نتایج به صورت زیر است.

train	1	2	3	4	5	6
prediction	3085.183	12667.551	23637.502	34642.333	8301.229	9993.774
charges	1526.312	11729.680	23887.663	33732.687	6799.458	7804.160

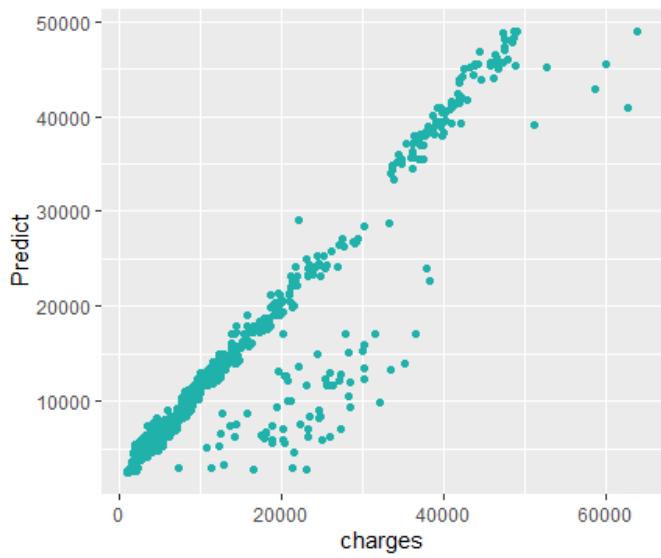
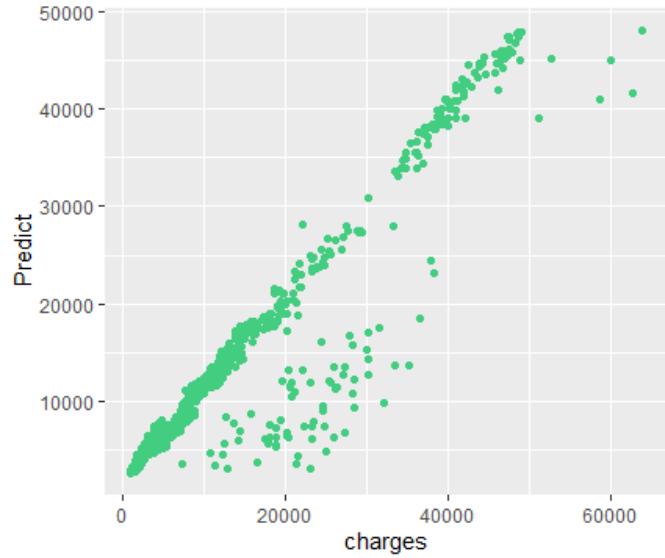
test	1	2	3	4	5	6
prediction	13138.941	23301.445	10911.639	7188.066	11081.389	3468.247
charges	11454.022	21677.283	9778.347	5757.413	10381.479	1725.552

اکنون نگاهی به MSE آموزشی و آزمایشی ۵ مدل بررسی شده خواهیم داشت.

	مدل اول	مدل دوم	مدل سوم	مدل چهارم	مدل پنجم
MSE_train	24082083	5875832	5882591	9767783	17860995
MSE_test	23824182	22577965	22083630	20788395	19369143

با توجه به جدول فوق مدل پنجم را به عنوان بهترین مدل انتخاب می کنیم زیرا هم از لحاظ MSE آزمایشی مدل بهتری است و هم چون MSE آموزشی و آزمایشی تقریباً به هم نزدیک اند می توان به این مدل از نظر بیش برازش بیشتر اطمینان کرد.

حال قصد داریم به مقایسه این مدل و مدل نهایی معرفی شده در فاز دوم بپردازیم.



سمت راست نمودار مدل پنجم، و سمت چپ، نمودار نهايی رگرسیون را نمایش می‌دهد. با توجه به این دونمودار می‌توان دید که هرچند این دو مدل تفاوت چندانی در پیش‌بینی ندارند اما با این حال مدلی که در این فاز معرفی شده کمی بهتر عمل می‌کند و همچنین مشکل نرمال نبودن مانده‌ها در مدل فوق تاثیری ندارد. در نتیجه، استفاده از این مدل بیشتر توصیه می‌شود.

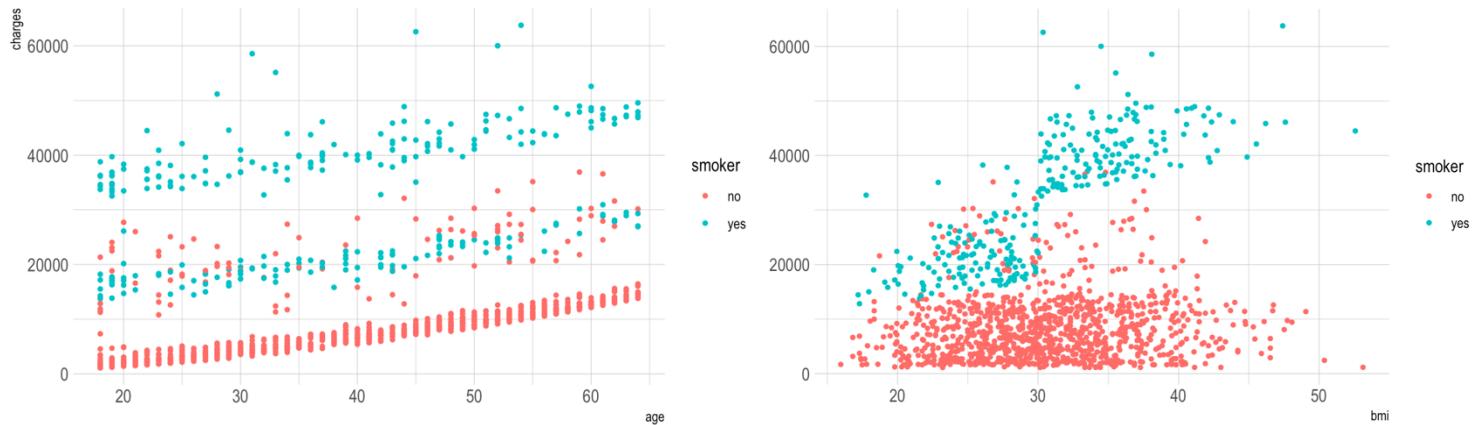
فاز پنجم

خوشه‌بندی

Clustering

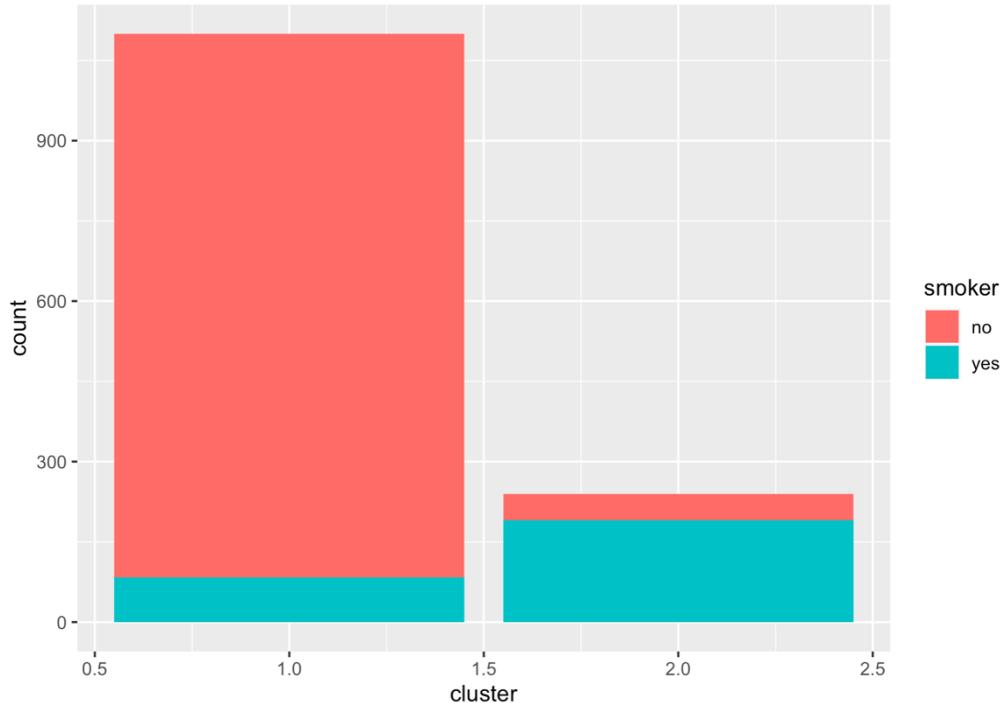
در شرکت های بیمه، اطلاعات افراد در تخمین [هزینه بیمه](#) بسیار موثر اند. به همین دلیل، درستی این اطلاعات برای این شرکت ها حائز اهمیت هستند. به دلایل مختلفی فرد بیمه شونده ممکن است اطلاعاتی را سهوی یا عمدی اشتباه وارد کند و تاثیر ویژه‌ای بر تخمین نهایی او داشته باشد.

در این بخش سعی داریم با کمک **تکنیک خوش‌بندی**، به بررسی درستی متغیر **سیگار کشیدن** بپردازیم. همان‌طور که در بخش‌های گذشته بررسی شد، این متغیر، یکی از موثر ترین متغیرها بر متغیر هدف ماست و سیگاری بودن یا نبودن، مطابق نمودارهای زیر، تاثیر بسزایی در تخمین ما خواهد داشت.



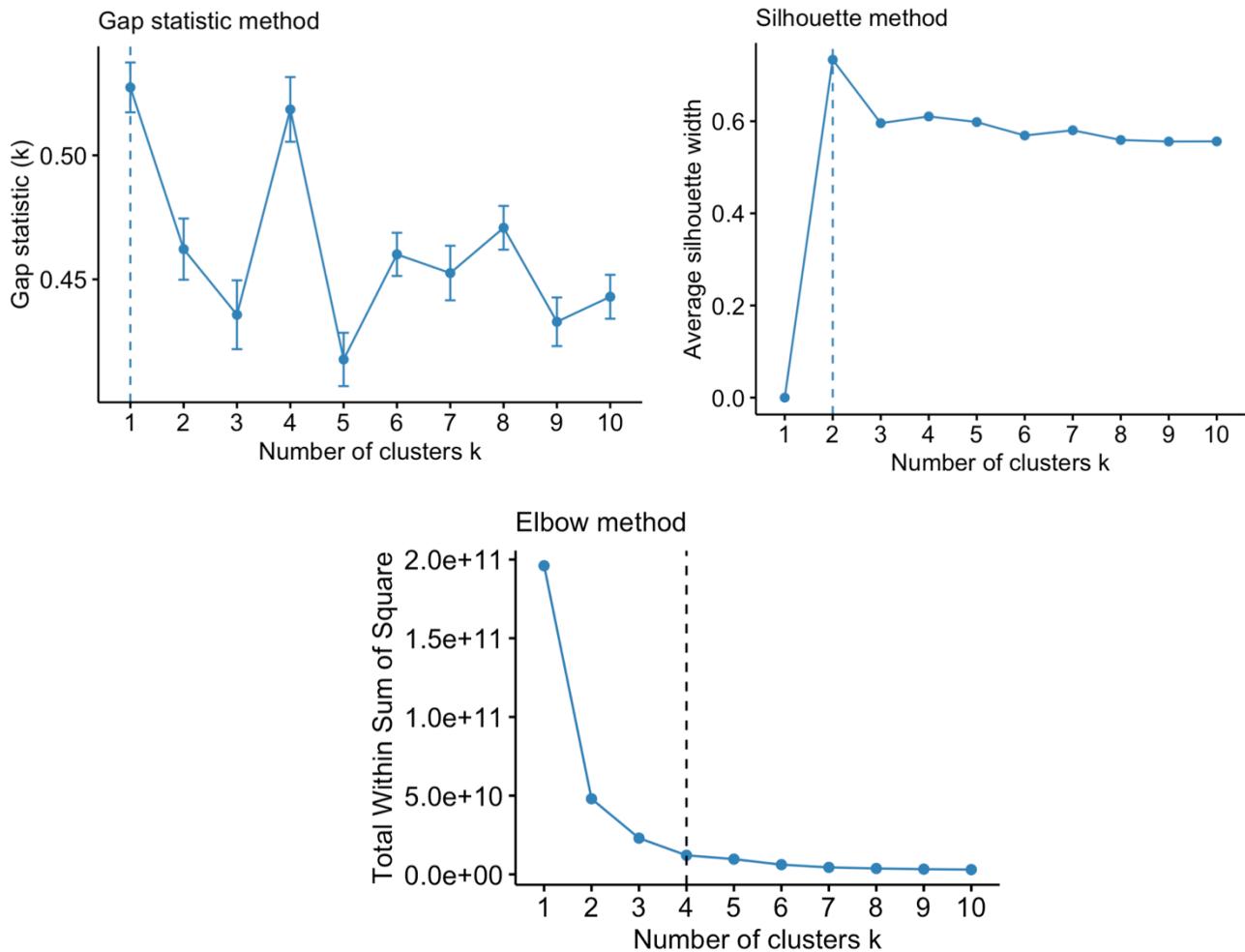
برای خوش‌بندی داده‌ها، با حذف متغیر **سیگار** شروع کرده و داده‌ها را با روش **k-means** به دو خوش‌بندی می‌کنیم. با توجه به پراکندگی داده‌ها در نمودارهای بالا، انتظار داریم در یک خوش‌بندی افراد سیگاری و در خوش‌بندی دیگر افراد غیرسیگاری را مشاهده کنیم.

نمودار زیر نشان دهنده داده های هر خوش به تفکیک **سیگاری** بودن یا نبودن است. همان طور که انتظار داشتیم، داده ها تقریبا با معیار سیگاری بودن یا نبودن تقسیک شده اند.



در خوش اول حدود ۹۲ درصد **غیرسیگاری** و در خوش دوم، حدود ۷۹ درصد **سیگاری** داریم. جالب توجه است که حدود ۲۱ درصد افراد در خوش دوم، اطلاعاتی شبیه به افراد **سیگاری** دارند در حالی که **غیرسیگاری** ثبت شده اند. اگر تاثیر متغیرهای دیگر در متغیر هدف را در نظر نگیریم، احتمال این که این ۲۱ درصد به اشتباه اطلاعات خود را وارد کرده باشند وجود داشته و شرکت بیمه باید فاکتورهای دیگری را برای کم کردن احتمال خطا در نظر بگیرد.

می دانیم یکی از معایب خوش‌بندی به روش k-means، مشخص بودن عدد k پیش از خوش‌بندی است. در صورتی که ممکن است اطلاعی از تعداد مناسب برای خوش‌ها نداشته باشیم. به همین دلیل، نیاز به روشنی داریم که قبل از این متدها، برای ما تعداد مناسب خوش‌ها را مشخص کند. در اینجا از سه روش Elbow Method و Silhouette Method، Gap Statistic Method استفاده می‌کنیم.

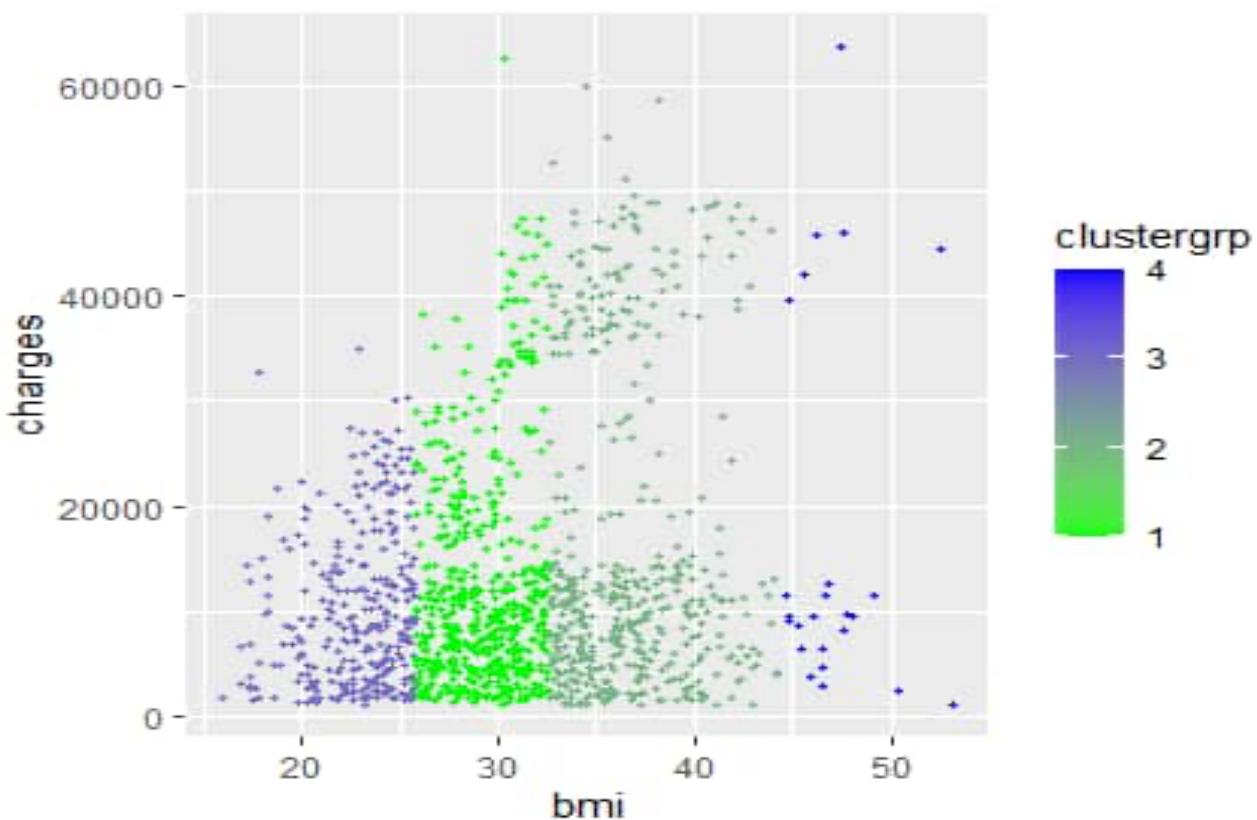


در روش Gap Statistic Method، تعداد مناسب خوش‌ها عدد یک محاسبه شده که به این معناست که داده‌های ماخوش‌بندی نیاز ندارند. روش دوم یعنی Silhouette Method تعداد مناسب خوش‌ها را عدد دو محاسبه کرده که در صفحات قبل بررسی کردیم. روش آخر، Elbow Method، تعداد مناسب برای خوش‌ها را چهار اعلام کرده است. در ادامه به بررسی این تعداد خوش‌بندی می‌پردازیم.

در بخش‌های قبلی دیدیم که متغیر **BMI** بر متغیر هدف تاثیر زیادی دارد. همچنین با در نظر گرفتن اثر متقابل **BMI** و **سیگاری بودن**، مدل بسیار بهتری در رگرسیون داشتیم. به همین دلیل در نظر داریم چهار خوش‌کردن داده‌ها، بر مبنای **BMI** باشد. شکل زیر نشان‌دهنده وضعیت سلامتی افراد براساس **BMI** است.

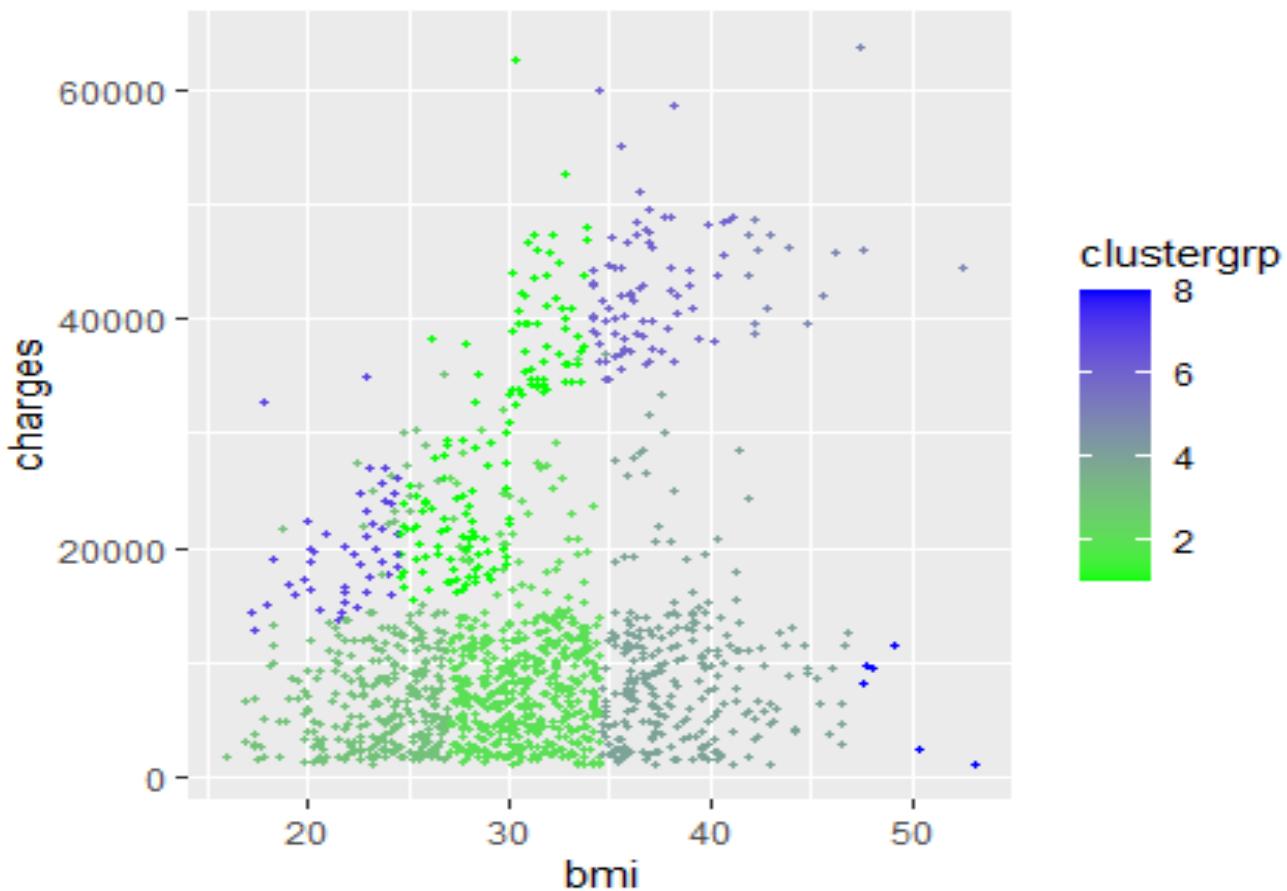


اکنون با روش خوش‌بندی **سلسه مراتبی**، داده‌های **BMI** را به ۴ خوش‌خواه تقسیم می‌کنیم.



همان طور که انتظار داشتیم، داده ها بر اساس **BMI** تفکیک شده اند که این موضوع می تواند در تخمین دقیق تر تاثیرگذار باشد. در بعضی خوش ها کمتر بودن پراکندگی داده ها بسیار کم کننده هستند. در بخش رگرسیون، دیدیم که اثر متقابل **BMI** و **سیگاری بودن**، تاثیر ویژه ای در بهبود عملکرد مدل داشت. به همین دلیل ممکن است ۸ دسته کردن داده ها، که شامل تمام حالات برای ۴ دسته **BMI** و ۲ دسته **سیگار**، نتیجه مناسب تری در خوش بندی ارائه دهد.

با توجه به پراکندگی **هزینه بیمه** بر حسب **BMI** به تفکیک **سیگاری** بودن یا نبودن و اینکه **BMI** را می توان به ۴ دسته اصلی تقسیم کرد، قصد داریم با خوش بندی داده ها به ۸ دسته، افراد با خصوصیات مشابه را در یک خوش قرار دهیم. به طور مثال افراد **سیگاری** با **BMI** کمتر از ۵/۱۸ در یک دسته، افراد **غیرسیگاری** با **BMI** کمتر از ۵/۱۸ در دسته ای دیگر و ... حال با اجرای خوش بندی سلسله مراتبی روی داده ها و در نظر گرفتن $k=8$ (تعداد خوش ها) نتایج روی نمودار **هزینه بیمه** بر حسب **BMI** به صورت زیر است:



همان طور که در شکل می بینید دسته بندی فوق به شکلی که انتظار می رفت انعام شده و در هر دسته افراد با مشخصات تقریباً یکسان قرار دارند. حال با توجه به این نتایج وقتی مورد جدید به شرکت بیمه مراجعه کرد می توان دید در کدام دسته قرار می گیرند و حدوداً [هزینه بیمه](#) او را تخمین زد و متناسب با مشخصات وی، بیمه مناسبی به او پیشنهاد داد.

تمام موارد بررسی شده در این بخش برای چندخوشه کردن داده ها می تواند کار ما را برای بررسی داده ها و یافتن مدل مناسب راحت تر کند و به عنوان پیش پردازشی برای مدل اصلی محسوب شود.