Final Project

Tujuan:

Tugas ini bertujuan untuk memahami lebih dalam dan menerapkan model regresi linear untuk menyelesaikan permasalahan nyata.

Cakupan tugas:

1. Pemahaman kontekstual, berupa interpretasi kuantitas yang dipelajari dalam konteks contoh permasalahan nyata
2. Pengolahan data, analisis dan interpretasi hasil

Due Date Pengumpulan Laporan di Emas : Minggu, 17 Desember 2023 pukul 23.59 WIB

Anggota kelompok:

No	Nama	NPM	Kontribusi	Tingkat kontribusi
1	Safira Ramadhani	2206026656	Aktif dalam mencari data, melakukan *pre-processing*, melakukan analisis, menyusun laporan, dan menyusun powerpoint	100%
2	Gian Sinar Katulistiwa	2206031675	Aktif dalam mencari data, melakukan *pre-processing*, melakukan analisis, menyusun laporan, dan menyusun powerpoint	100%
3	Renata Shaula Alfino Ritonga	2206815812	Aktif dalam mencari data, melakukan *pre-processing*, melakukan analisis, menyusun laporan, dan menyusun powerpoint	100%
4	Golda Aurelia Silalahi	2206826173	Aktif dalam mencari data, melakukan *pre-processing*, melakukan analisis, menyusun laporan, dan menyusun powerpoint	100%

5	Amira Shohifa	2206829130	Aktif dalam mencari data, melakukan *pre-processing*, melakukan analisis, menyusun laporan, dan menyusun powerpoint	100%

Instruksi:

Carilah data dengan permasalahan yang dapat dimodelkan dengan regresi linear. Data memuat minimal 200 pengamatan, dengan pengukuran numerik maupun kategorik. Dapat menggunakan data pada Project 1. Prosedur yang dilakukan mencakup pemilihan model terbaik (dengan metrik yang sesuai), seleksi variabel, pengecekan asumsi (dan langkah mengatasi masalah jika ada pelanggaran asumsi), pastikan bahwa tidak ada pitfalls pada regresi yang diajukan.

Lakukan pengolahan data (jika diperlukan lakukan pre-processing data terlebih dahulu), dengan prosedur yang tepat, kemudian lakukan analisis dan interpretasi hasilnya. Ikuti langkah – langkah berikut.

Bagian 1. Pendahuluan

1.1 Rumusan Masalah

Biaya perawatan kesehatan yang terus meningkat telah menjadi masalah serius di banyak negara. Pengeluaran ini dipengaruhi oleh faktor-faktor seperti perkembangan teknologi medis, biaya obat-obatan yang terus naik, dan tingginya permintaan akan layanan kesehatan. Oleh karena itu, pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi biaya medis sangat penting.

Selain itu, faktor demografis seperti usia, jenis kelamin, lokasi geografis, indeks massa tubuh (BMI), jumlah anak yang ditanggung oleh asuransi kesehatan, serta kebiasaan merokok juga memiliki pengaruh yang signifikan terhadap besarnya biaya medis yang dikeluarkan oleh seseorang.

Prediksi biaya asuransi kesehatan mengenai biaya asuransi kesehatan yang akan dikeluarkan dengan faktor-faktor di atas dapat membantu perusahaan asuransi mengatur premi, mengelola risiko keuangan, dan merencanakan alokasi sumber daya secara efektif. Analisis prediksi biaya asuransi kesehatan yang tepat dapat memberikan landasan yang kuat dalam mendukung keberlanjutan sistem asuransi kesehatan serta memberikan manfaat signifikan bagi peserta asuransi dengan penawaran premi yang terjangkau.

1.2 Sumber Data

A. Permasalahan

Permasalahan pada data yang kami ambil merupakan permasalahan regresi linear mengenai prediksi biaya asuransi.

B. Sumber

Kami menggunakan dataset yang bersumber dari laman Kaggle dengan link sebagai berikut: https://www.kaggle.com/datasets/mirichoi0218/insurance

C. Ukuran Data

Data memuat 1338 pengamatan dengan 7 jumlah pengukuran (kolom data).

D. Skala/Tipe Data dan Arti Pengukuran Data Tersebut

Data yang digunakan merupakan dataset yang berisi 7 jumlah pengukuran, yaitu:

No.	Pengukuran	Tipe	Keterangan	Variabel
1	age	*integer* (numerik)	usia nasabah asuransi kesehatan	variabel prediktor
2	sex	*object*	jenis kelamin (laki-laki atau perempuan)	variabel prediktor

3	bmi	*float* (numerik)	indeks massa tubuh	variabel prediktor
4	children	*object*	jumlah anak atau tanggungan yang termasuk dalam cakupan asuransi kesehatan	variabel prediktor
5	smoker	*object*	penerima asuransi merokok atau tidak	variabel prediktor
6	region	*object*	area tempat tinggal penerima asuransi kesehatan (northeast, southeast, southwest, dan northwest)	variabel prediktor
7	charges	*float* (numerik)	biaya medis yang ditagih oleh asuransi kesehatan	variabel respon

Maka, data yang akan digunakan dalam pemodelan memuat 1338 pengamatan dengan tujuh variabel, yaitu satu variabel respon kuantitatif, dua variabel prediktor kuantitatif, dan empat variabel prediktor kualitatif.

E. Tujuan

Tujuan dari analisis regresi yang akan dilakukan adalah untuk mengetahui faktor apa saja yang berpengaruh terhadap tagihan biaya yang harus dibayar ke suatu asuransi kesehatan pada saat nasabah melakukan pengobatan dan melakukan prediksi tagihan biaya yang harus dibayar ke suatu asuransi kesehatan berdasarkan faktor-faktor tertentu.

Bagian 2. Pre-processing (jika ada) dan analisis deskriptif

Kami melakukan beberapa tahap pre-processing data sebelum nantinya akan melakukan proses pemodelan.

2.1 Load Data

Tahap pertama kami melakukan *import library* yang dibutuhkan dalam proses *pre-processing*. Data kami yang diambil dari laman Kaggle disimpan pada google drive. Pengambilan file dari google drive melalui google collab dapat menggunakan kode berikut. Dataset kami disimpan dalam variabel **data**.

Berikut *import library* yang dibutuhkan.

```
[1]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     import plotly.express as px
     from scipy.stats import zscore
     from sklearn.model_selection import train_test_split
     from sklearn.preprocessing import OneHotEncoder
     from sklearn.compose import ColumnTransformer
```

Berikut dilakukan *load data* file insurance beserta menampilkan lima data pertamanya.

```
[4]  #Melakukan load data
     data = pd.read_csv('/content/insurance_asli.csv')
     data.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	1	southwest	16884.92400
1	18	male	33.770	1	0	southeast	1725.55230
2	28	male	33.000	3	0	southeast	4449.46200
3	33	male	22.705	0	0	0rthwest	21984.47061
4	32	male	28.880	0	0	0rthwest	3866.85520

2.2 Melihat Informasi Data dan Mengecek Missing Values

Pada tahap ini, kami akan melihat tipe data masing-masing variabel, mengecek ada atau tidaknya *missing values* pada kolom atau baris data tersebut, dan melihat jumlah entri dari masing-masing variabel dengan kode berikut.

Berikut adalah informasi dari dataset kami.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

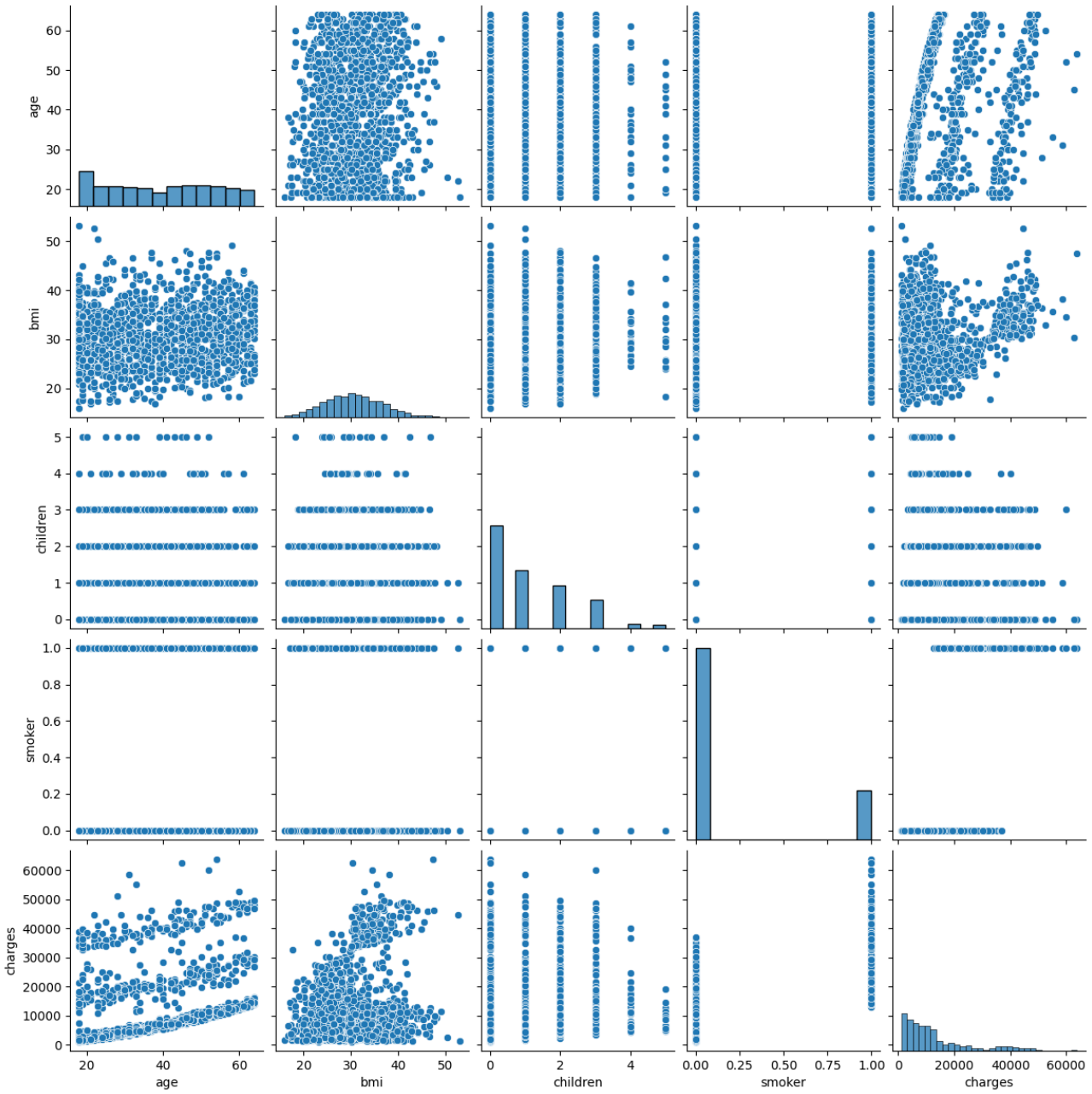
Dapat dilihat, data memiliki 1338 pengamatan dengan 6 variabel prediktor, yaitu **"age"**, **"sex"**, **"bmi"**, **"children"**, **"smoker"**, dan **"region"**, serta 1 variabel respon, yaitu **"charges"**. Masing-masing variabel tidak memiliki *missing values* ditandai dengan *"non-null"* pada kolom *"Non-Null Count"*.

Kami juga melihat nilai statistika deskriptif dari variabel numerik **"age"**, **"bmi"**, **"children"**, serta **"charges"** dengan *output* sebagai berikut.

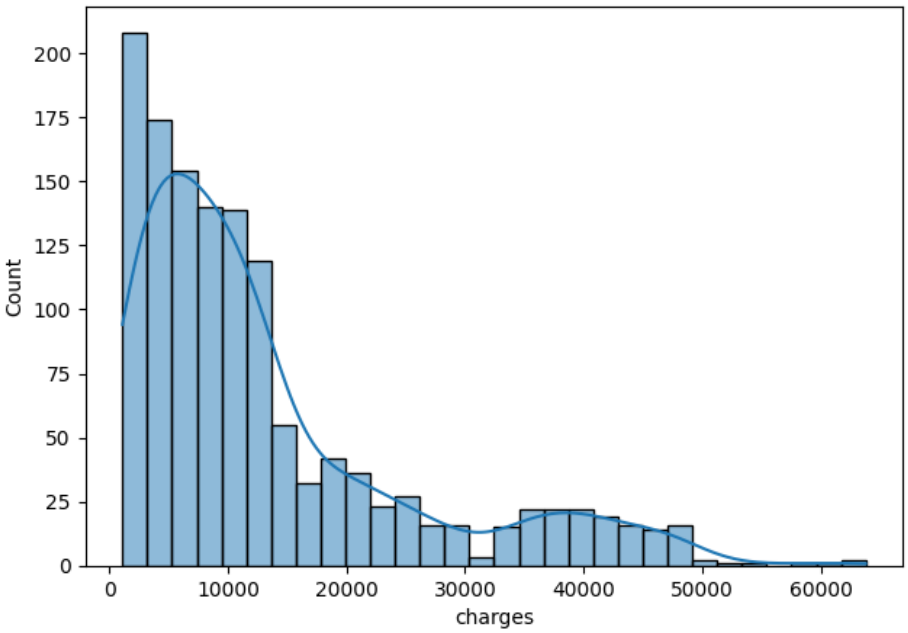
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

2.3 Visualisasi

Berikut bentuk visualisasi *pairplot* dari data.



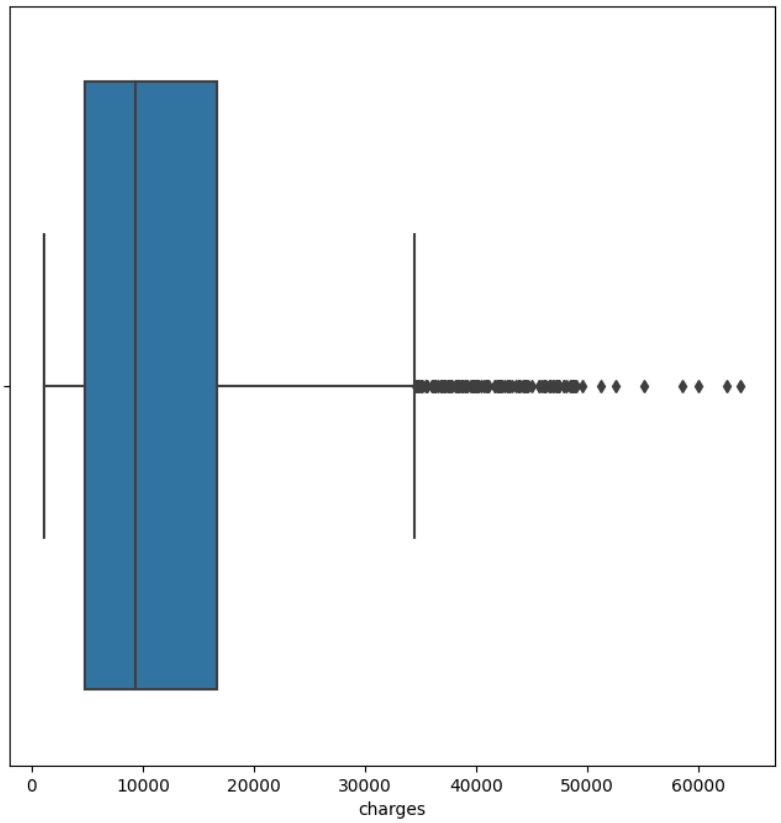
Akan dilihat juga visualisasi histogram dari kuantitas setiap nilai-nilai pada variabel respon **"charges"**.



2.4 Mengecek Outliers Variabel Numerik

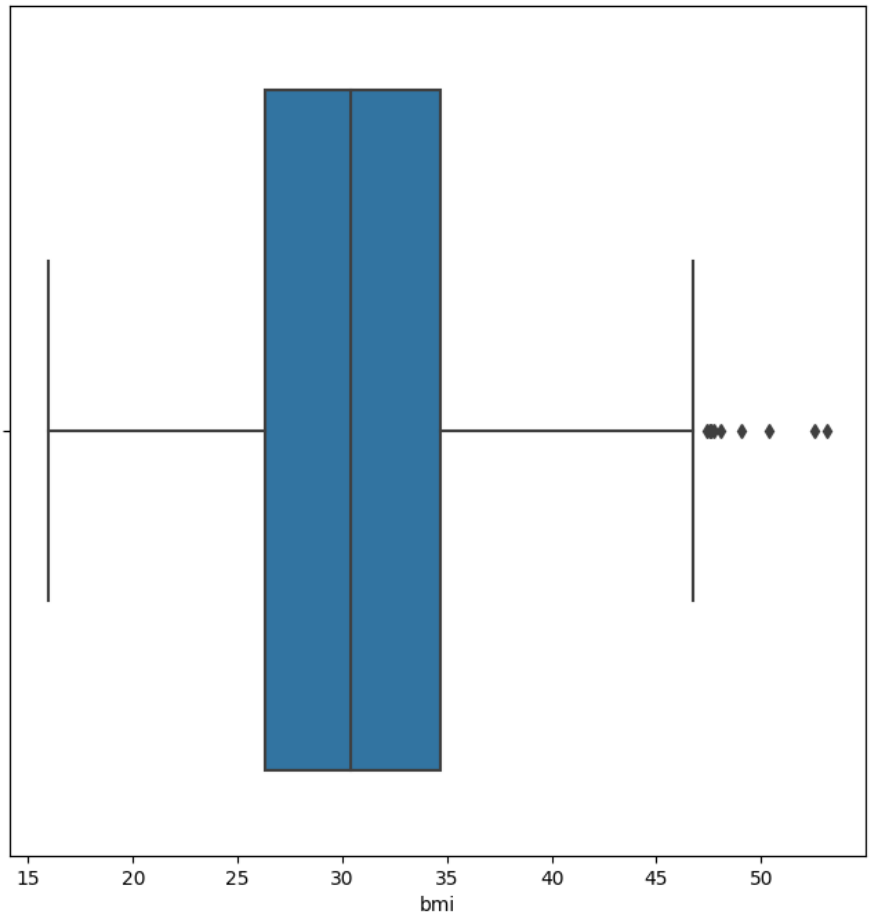
Kami akan melakukan pengecekan Outliers pada data yang bertipe numerik yaitu variabel **"charges"**, **"bmi"**, **"age"**, dan **"children"**. Kami melihat keberadaan *outliers* dengan menggunakan boxplot dan melihat nilai - nilai *outliers*.

a. **Untuk variabel "charges"**



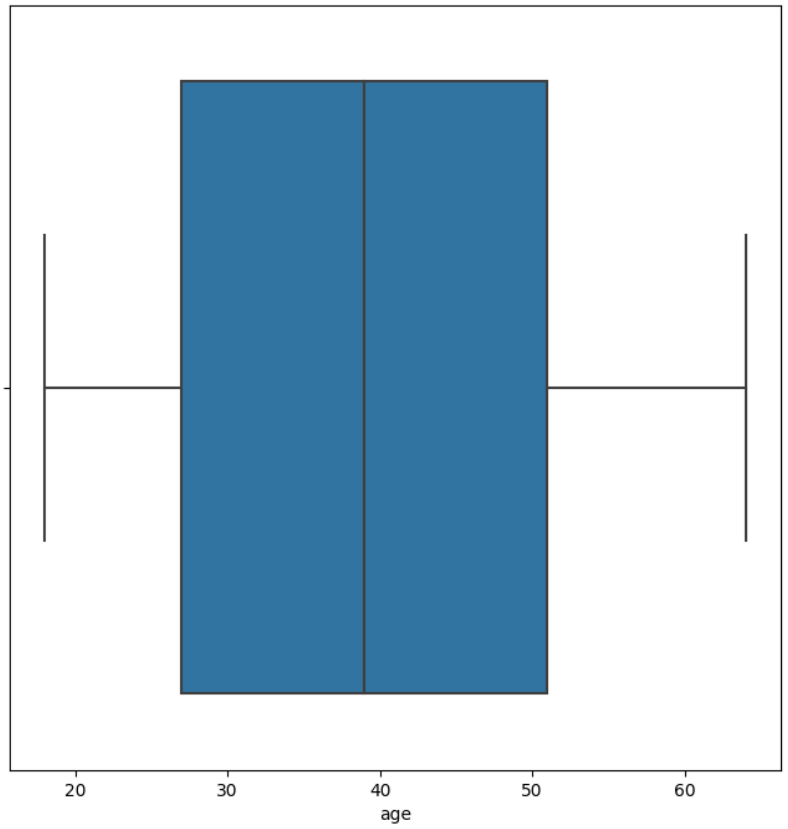
```
Nilai-nilai Outlier:
14      39611.75770
19      36837.46700
23      37701.87680
29      38711.00000
30      35585.57600
           ...
1300    62592.87309
1301    46718.16325
1303    37829.72420
1313    36397.57600
1323    43896.37630
Name: charges, Length: 139, dtype: float64
```

b. Untuk variabel "bmi"



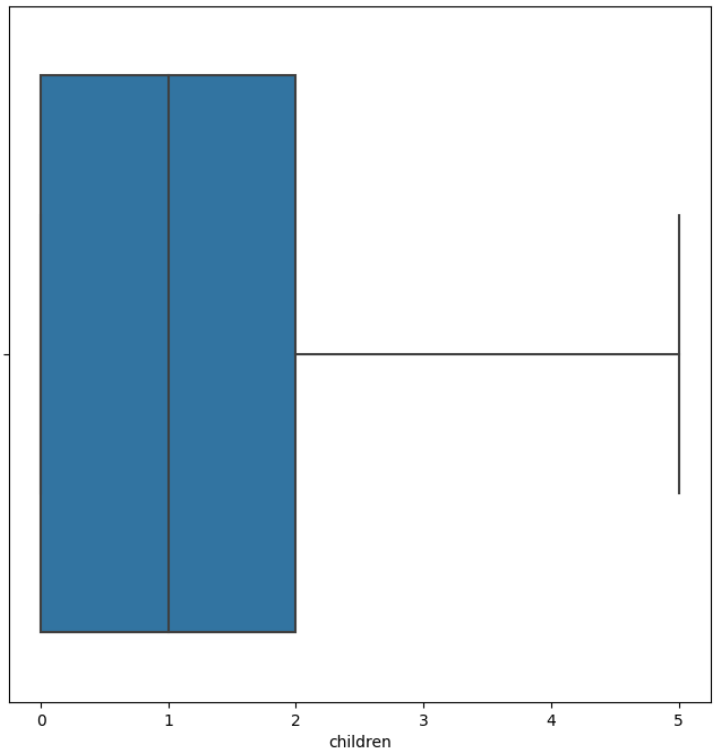
```
Nilai-nilai Outlier:
116     49.06
286     48.07
401     47.52
543     47.41
847     50.38
860     47.60
1047    52.58
1088    47.74
1317    53.13
Name: bmi, dtype: float64
```

c. Untuk variabel “age”




```
Nilai-nilai Outlier:
Series([], Name: age, dtype: int64)
```

d. Untuk variabel "children"




```
Nilai-nilai Outlier:
Series([], Name: children, dtype: int64)
```

Berdasarkan *output* pada tahap ini, dapat dilihat bahwa variabel yang memiliki outliers adalah variabel **"charges"** dan variabel **"bmi"**.

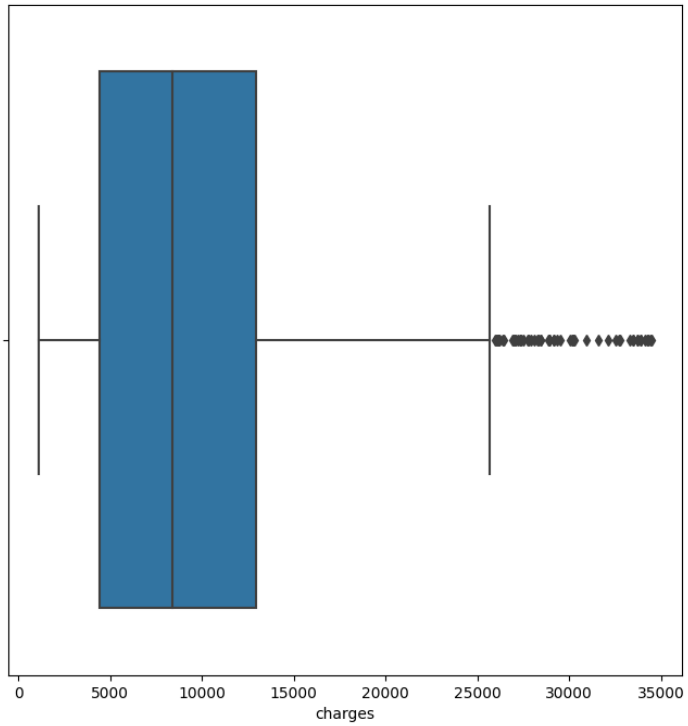
2.5 Menghilangkan Outliers

Pada tahap ini akan dilakukan penghapusan outliers untuk variabel **"charges"** dan variabel **"bmi"**.

a. Untuk variabel "charges"

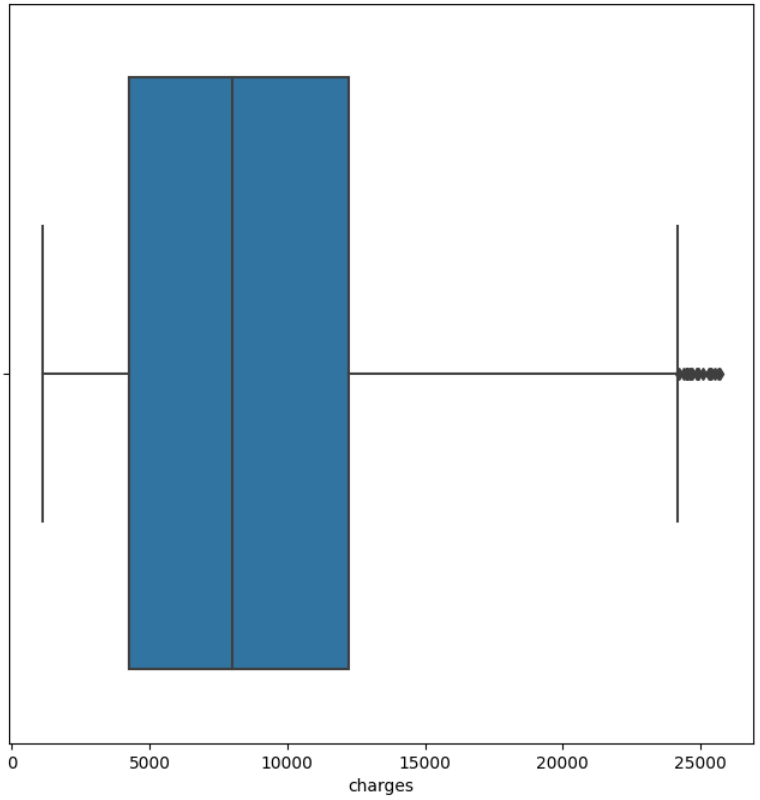
Pada variabel **"charges"**, penghapusan *outliers* dilakukan sebanyak 5 kali

1) Setelah penghapusan *outliers* yang pertama, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



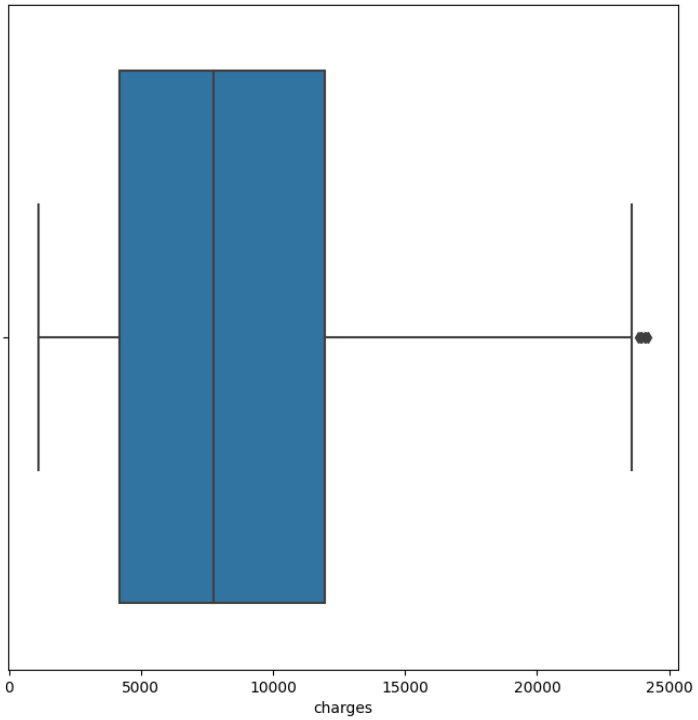
Karena masih terdapat *outliers*, akan dilakukan penghapusan *outlier* yang kedua.

2) Setelah penghapusan *outliers* yang kedua, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



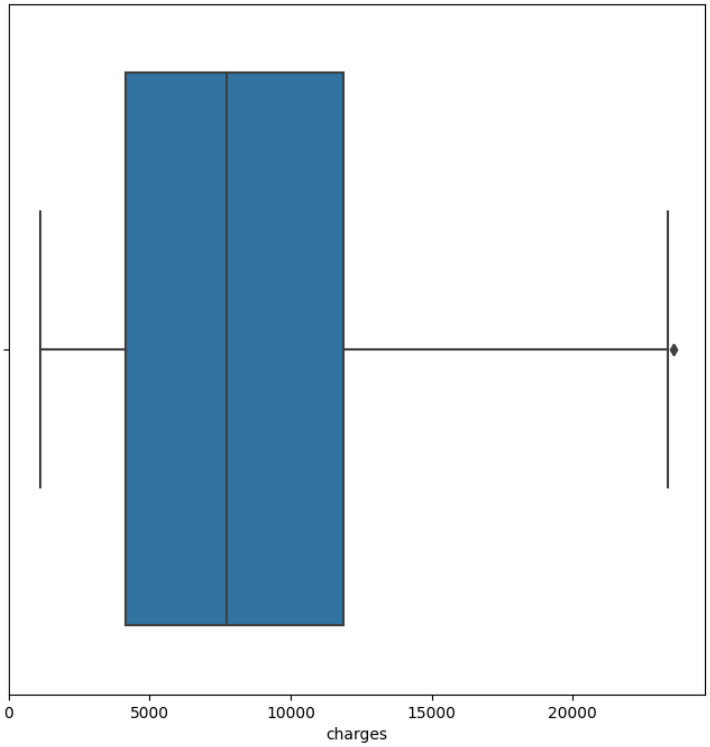
Karena masih terdapat *outliers*, akan dilakukan penghapusan *outlier* yang ketiga.

3) Setelah penghapusan *outliers* yang ketiga, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



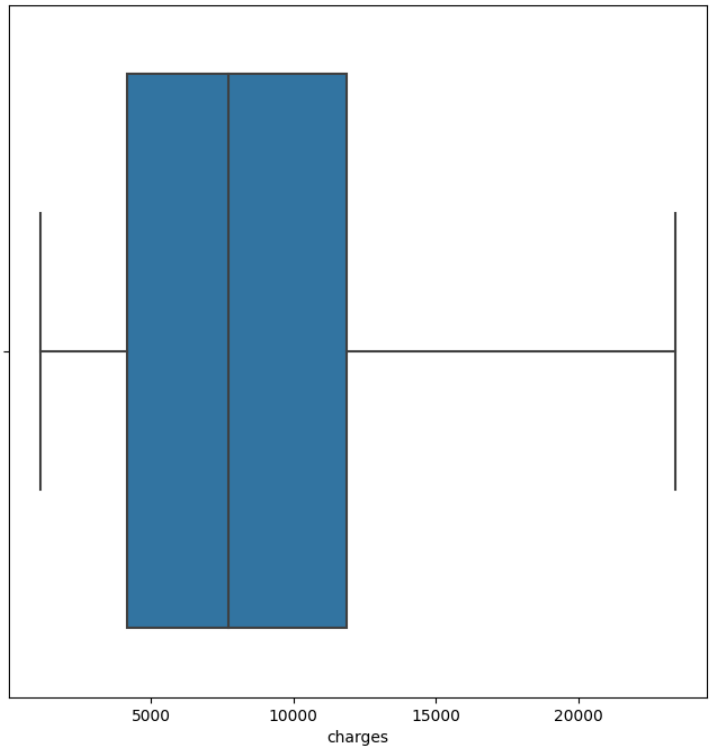
Karena masih terdapat *outliers*, akan dilakukan penghapusan *outlier* yang keempat.

4) Setelah penghapusan *outliers* yang keempat, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



Karena masih terdapat *outliers*, akan dilakukan penghapusan *outlier* yang kelima.

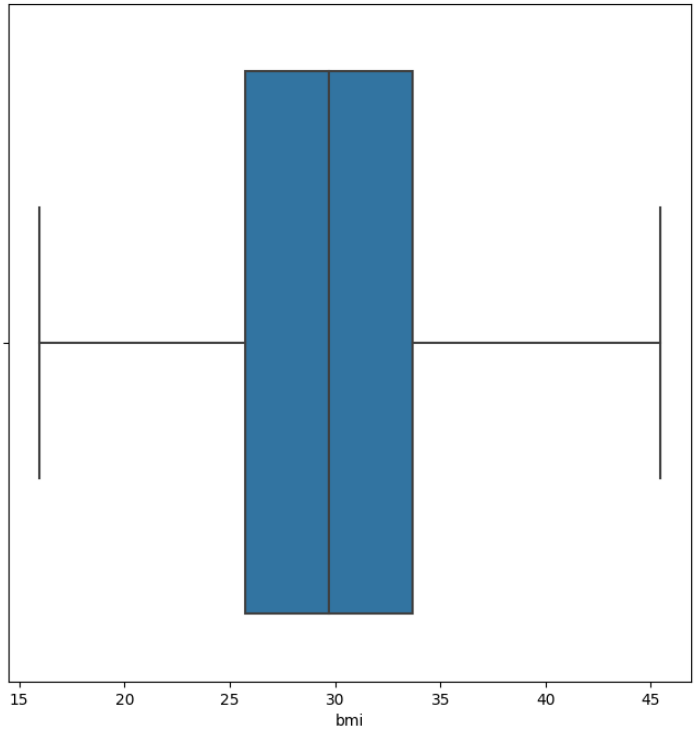
5) Setelah penghapusan *outliers* yang kelima, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



Didapat sudah tidak terdapat *outliers* lagi pada variabel “charges”.

b. Untuk variabel “bmi”

Pada variabel **“bmi”** dilakukan penghapusan *outliers* sebanyak satu kali. Setelah penghapusan *outliers*, dilakukan visualisasi *boxplot* kembali dengan hasil sebagai berikut.



Dapat dilihat bahwa sudah tidak terdapat *outliers* lagi di variabel “bmi”.

Setelah dilakukan penghapusan *outliers* dari variabel **"charges"** dan variabel **"bmi"**, akan diperiksa kembali informasi dari dataset dengan hasil sebagai berikut.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1116 entries, 0 to 1336
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1116 non-null   int64
 1   sex       1116 non-null   object
 2   bmi       1116 non-null   float64
 3   children  1116 non-null   int64
 4   smoker    1116 non-null   object
 5   region    1116 non-null   object
 6   charges   1116 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 69.8+ KB
```

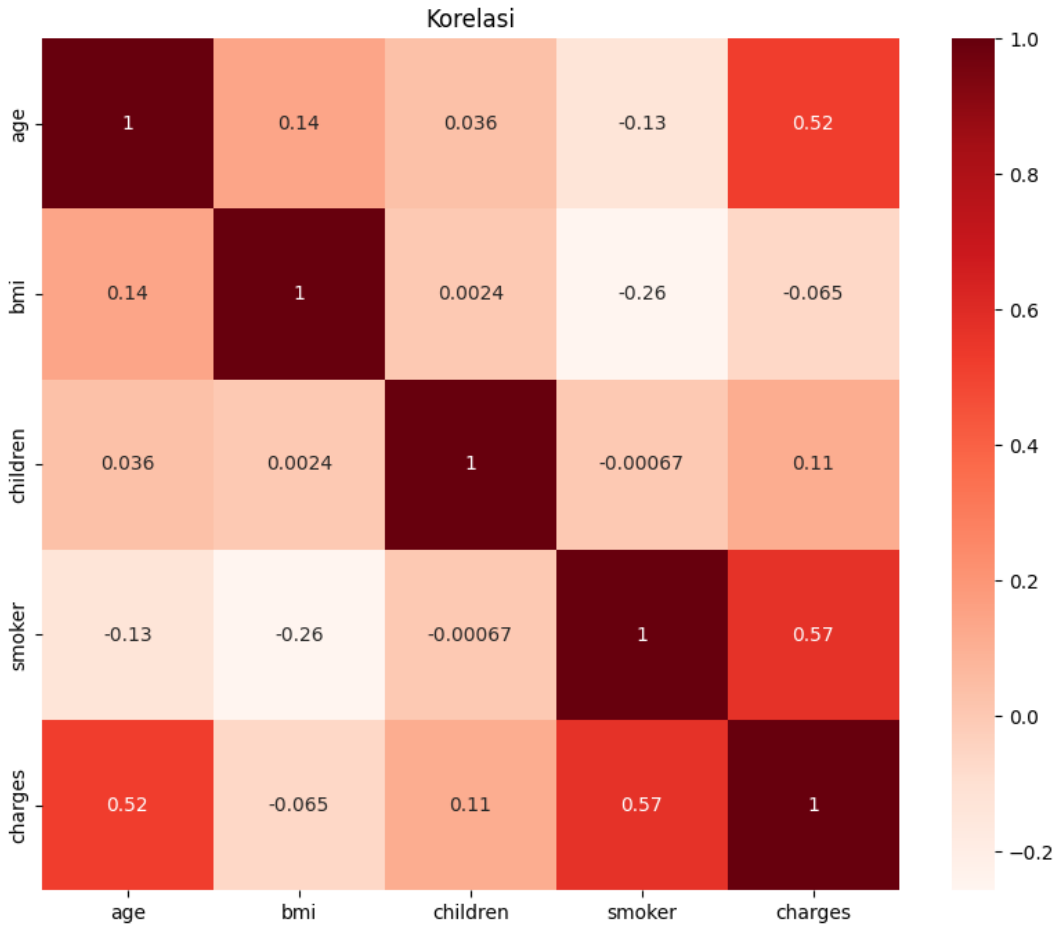
Dapat dilihat, jumlah pengamatan yang awalnya berjumlah 1338 pengamatan, berkurang menjadi 1116 pengamatan saja. Hal ini dikarenakan pada saat penghapusan *outliers*, baris data yang memiliki *outliers* juga dihapus.

2.6 Melihat Korelasi Antar Variabel

Akan dilakukan pengecekan korelasi antara variabel prediktor **"smoker"**, **"age"**, **"children"**, dan **"bmi"** terhadap variabel respon **"charges"**. Didapatkan *output* sebagai berikut.

```
Correlation with charges variable:
charges     1.000000
smoker      0.569117
age         0.520632
children    0.112928
bmi        -0.064507
Name: charges, dtype: float64
```

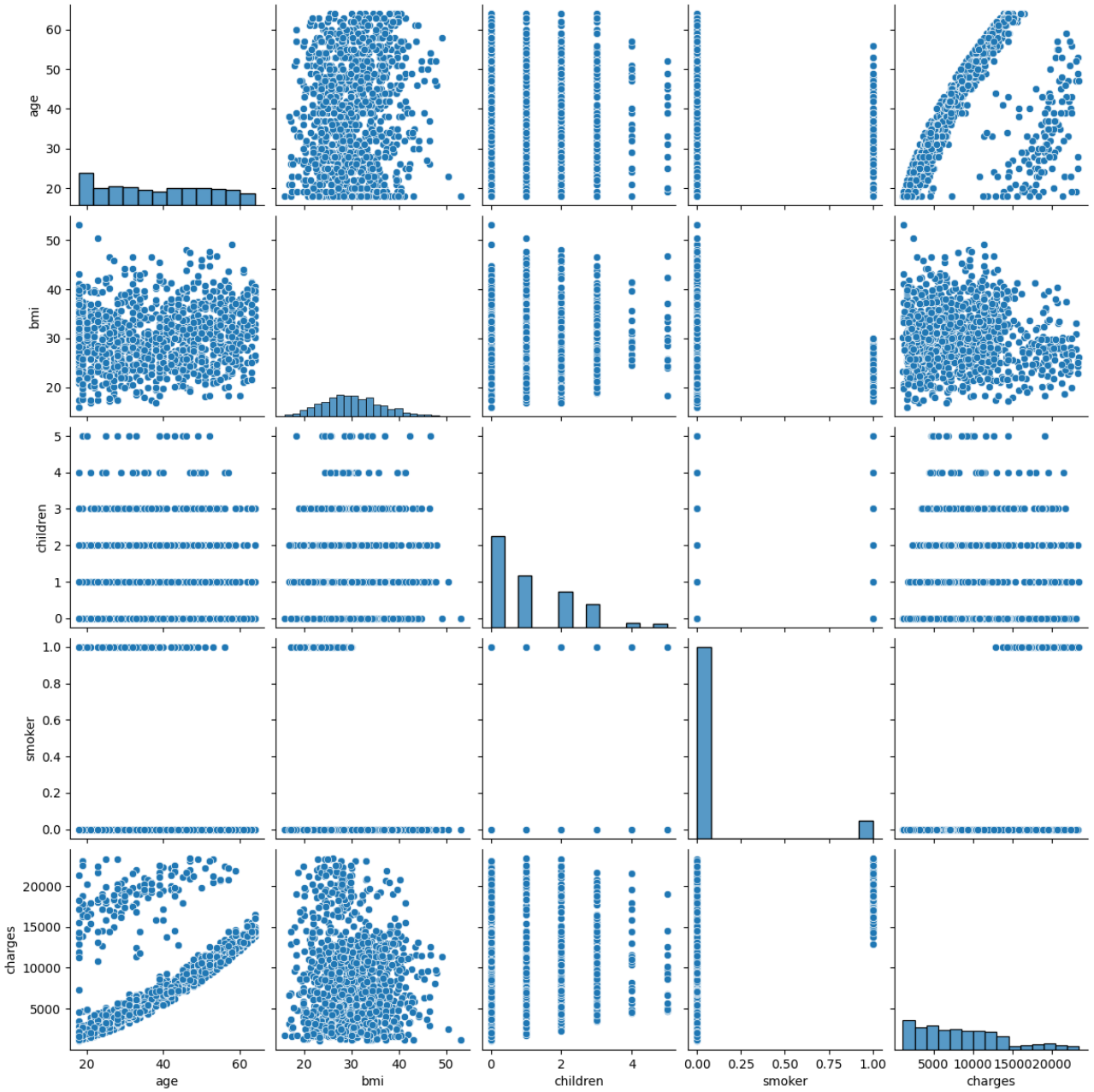
Kemudian, akan diperiksa juga korelasi antara masing-masing variabel prediktor numerik dengan menggunakan visualisasi *heatmap*.

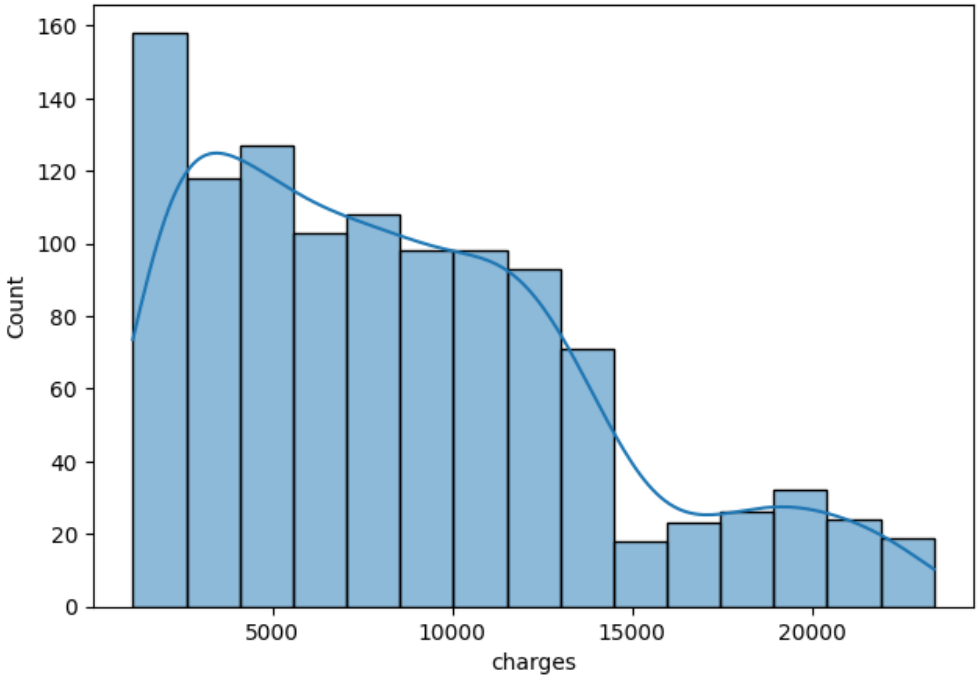


Dapat dilihat bahwa nilai korelasi antara variabel prediktor terhadap variabel prediktor yang lain tidak besar, yang menandakan tidak terdapat multikolinearitas antara variabel prediktor.

2.7 Visualisasi Setelah Penghapusan *Outlier*

Akan dilihat lagi visualisasi *pairplot* dari data dan visualisasi histogram dari kuantitas setiap nilai-nilai pada variabel respon setelah dilakukan penghapusan *outliers*.





2.8 Penyimpanan dataset

Setelah dilakukan *pre-processing* dilakukan penyimpanan *dataset* yang nantinya akan digunakan untuk pemodelan di *software RStudio*.

```
[38] # Penyimpanan dataset yang telah dilakukan preprocessing
     data_encoded.to_csv('data_prepro.csv', index=False)
```

2.9 Hipotesis yang Diajukan

Dari hasil *pre-processing,* kami akan mengajukan model regresi linier yang dapat digunakan untuk melakukan prediksi terhadap variabel respon **'charges'**. Berikut adalah hipotesis untuk teknik regresi pada proses selanjutnya.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$, yang artinya model tidak berguna

H_1 : Minimal salah satu dari $\beta_j \neq 0; j = 1, 2, \ldots, 6$, yang artinya model berguna

Dengan j = 1, 2, 3, 4, 5, 6, Berturut-turut merepresentasikan variabel: age, sex, bmi, children, smoker, dan region.

Bagian 3. Pemodelan

Pada bagian ini, kami akan mengajukan 3 model regresi linear terbaik yang berbeda dengan rincian sebagai berikut.

a. Model 1

Pada model 1, kami akan mengajukan model regresi linear yang dilakukan dari data mentah (bukan hasil *pre-processing*) dengan menggunakan metode *backward elimination* dan dengan menerapkan fungsi step().

```
> #backward elimination
> step(model_all, direction = "backward")
Start:  AIC=23316.43
charges ~ age + sex + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
- sex       1 5.7164e+06 4.8845e+10 23315
<none>                  4.8840e+10 23316
- region    3 2.3343e+08 4.9073e+10 23317
- children  1 4.3755e+08 4.9277e+10 23326
- bmi       1 5.1692e+09 5.4009e+10 23449
- age       1 1.7124e+10 6.5964e+10 23717
- smoker    1 1.2245e+11 1.7129e+11 24993

Step:  AIC=23314.58
charges ~ age + bmi + children + smoker + region

           Df  Sum of Sq        RSS   AIC
<none>                  4.8845e+10 23315
- region    3 2.3320e+08 4.9078e+10 23315
- children  1 4.3596e+08 4.9281e+10 23325
- bmi       1 5.1645e+09 5.4010e+10 23447
- age       1 1.7151e+10 6.5996e+10 23715
- smoker    1 1.2301e+11 1.7186e+11 24996

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

Coefficients:
    (Intercept)              age              bmi         children
       -11990.3            257.0            338.7            474.6
      smokeryes  regionnorthwest  regionsoutheast  regionsouthwest
        23836.3           -352.2          -1034.4           -959.4
```

Kemudian, didapat *summary* dari model 1 sebagai berikut.

```
> model1 <- lm(formula = charges ~ age + bmi + children + smoker + region, data = insurance)
> summary(model1)

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

Residuals:
     Min       1Q   Median       3Q      Max
-11367.2  -2835.4   -979.7   1361.9  29935.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
age                256.97      11.89  21.610  < 2e-16 ***
bmi                338.66      28.56  11.858  < 2e-16 ***
children           474.57     137.74   3.445 0.000588 ***
smokeryes        23836.30     411.86  57.875  < 2e-16 ***
regionnorthwest   -352.18     476.12  -0.740 0.459618
regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
regionsouthwest   -959.37     477.78  -2.008 0.044846 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Berdasarkan *summary* dari model 1 di atas, dapat disimpulkan model regresi linear untuk model 1 adalah sebagai berikut.

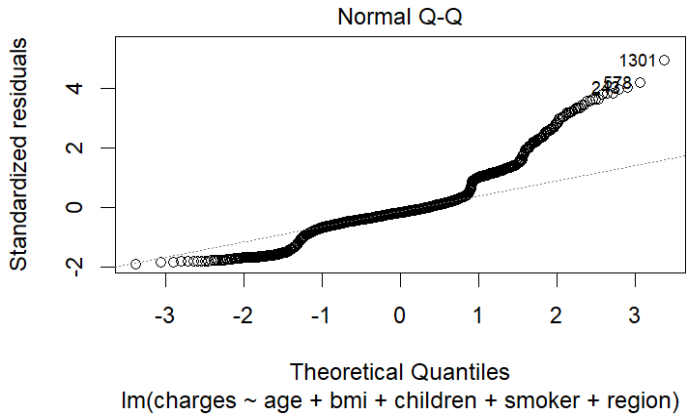
$$Y = -11990.27 + 256.97x_{age} + 338.66x_{bmi} + 474.57x_{children} + 23836.30x_{smoker\_yes} - 352.18x_{region\_northwest} - 1034.36x_{region\_southeast} - 959.37x_{region\_southwest}$$

Dengan **"smoker_yes"** adalah variabel **"smoker"** yang dijadikan dummy dengan **"smoker_no"** menjadi *base level*. Kemudian variabel **"region_northwest"**, **"region_southeast"**, **"region_southwest"** adalah variabel **"region"** yang dijadikan dummy dengan **"region_northeast"** menjadi *base level*.

Model 1 memiliki nilai *Adjusted R-Squared* mencapai 0.7496. Artinya, model regresi dijelaskan oleh variabel prediktor sebesar 0.7496. Sedangkan, 0.2504 sisanya dijelaskan oleh variabel prediktor lain di luar model regresi linear yang kami ajukan. Model 1 dipilih karena ketika menggunakan *backward elimination* dengan fungsi step(), nilai AIC (*Akaike Information Center*) terkecil berada pada model tersebut.

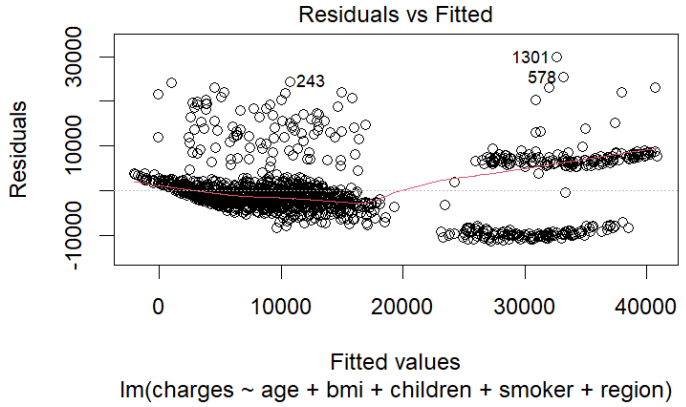
Akan diperiksa apakah asumsi normalitas pada residual, linearitas, homoskedastisitas, dan tidak adanya multikolinearitas pada model 1 terpenuhi atau tidak.

1) Asumsi Normalitas pada Residual



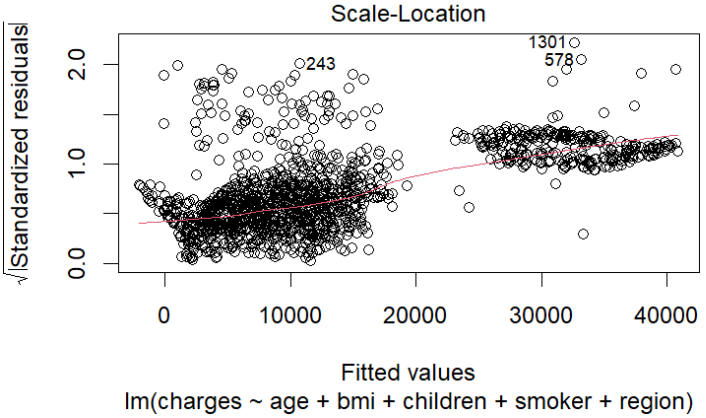
Dapat dilihat bahwa titik-titik dalam plot mengikuti garis diagonal. Sehingga asumsi normalitas pada residual terpenuhi.

2) Asumsi Linearitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola khusus. Sehingga asumsi linearitas terpenuhi.

3) Asumsi Homoskedastisitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola tertentu. Sehingga, asumsi homoskedastisitas terpenuhi.

4) Asumsi Tidak Ada Multikolinearitas

```
> vif(model1)
                age                 bmi            children           smokeryes 
             1.0162              1.1042              1.0037              1.0064 
regionnorthwest regionsoutheast regionsouthwest 
             1.5188              1.6522              1.5294 
```

Dapat dilihat nilai VIF (*Variance Inflation Factor*) untuk setiap variabel prediktor di model 1 semuanya bernilai < 4 yang menandakan tidak adanya multikolinearitas. Sehingga asumsi tidak ada multikolinearitas terpenuhi.

b. Model 2

Pada model 2, kami akan mengajukan model regresi linear yang dilakukan dari data mentah (bukan hasil *pre-processing*). Berikut adalah model 2 yang akan kami ajukan.

```
> insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
> model3 <- lm(formula = charges ~ age + bmi + children + smoker + bmi30*smoker +
 region, data = insurance)
> summary(model3)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-18422.1  -1845.2  -1276.1   -441.8  24566.2 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -4967.297    954.272  -5.205 2.24e-07 ***
age               263.663      8.812  29.920  < 2e-16 ***
bmi               114.090     34.595   3.298 0.001000 ** 
children          516.796    102.056   5.064 4.69e-07 ***
smokeryes       13383.160    444.301  30.122  < 2e-16 ***
bmi30            -869.805    426.244  -2.041 0.041485 *  
regionnorthwest  -264.050    352.801  -0.748 0.454328    
regionsoutheast  -823.426    355.196  -2.318 0.020588 *  
regionsouthwest -1221.147    354.076  -3.449 0.000581 ***
smokeryes:bmi30 19744.706    610.253  32.355  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4490 on 1328 degrees of freedom
Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8625 
F-statistic: 933.2 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Pada model ini, kami membuat variabel **“bmi30”** yang merupakan pembagian 2 faktor dari variabel **“bmi”**, yaitu ketika nilai **“bmi”** kurang dari 30 dan ketika nilai **“bmi”** lebih dari 30. Nilai **“bmi”** kurang dari 30, berarti nasabah penerima asuransi kesehatan tidak memiliki permasalahan obesitas dan ditandai dengan nilai “0”. Nilai **“bmi”** lebih dari 30, berarti nasabah penerima asuransi kesehatan memiliki permasalahan obesitas yang nantinya akan membutuhkan kebutuhan medis lebih banyak dibandingkan yang tidak obesitas, sehingga ditandai dengan nilai “1”.

Kemudian, kami menambahkan interaksi antara variabel **“bmi30”** dengan variabel **“smoker”** karena kami mengasumsikan ketika seorang nasabah asuransi kesehatan memiliki permasalahan obesitas dan juga merupakan perokok, akan membutuhkan kebutuhan medis lebih banyak dibandingkan dengan nasabah asuransi kesehatan yang tidak memiliki permasalahan obesitas tetapi bukan merupakan perokok.

Sehingga, model regresi untuk model 2 adalah sebagai berikut.

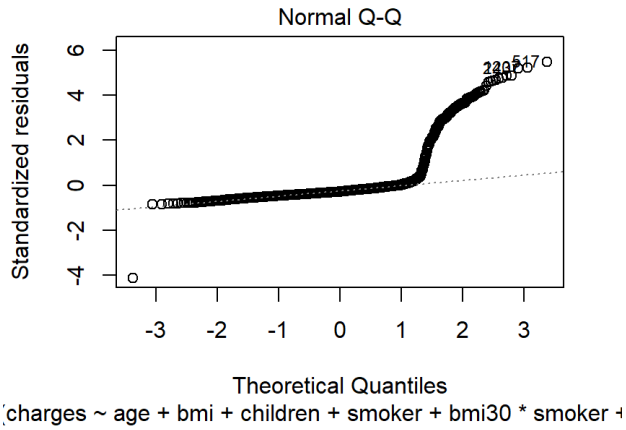
$$
\begin{aligned}
Y = \; & -4967.297 + 263.663x_{age} + 114.090x_{bmi} + 516.796x_{children} \\
& + 13383.160x_{smoker\_yes} - 869.805x_{bmi30} - 264.050x_{region\_northwest} \\
& - 823.426x_{region\_southeast} - 1221.147x_{region\_southwest} \\
& + 19744.706x_{smoker\_yes:bmi30}
\end{aligned}
$$

Dengan **“smoker_yes”** adalah variabel **“smoker”** yang dijadikan dummy dengan **“smoker_no”** menjadi *base level*. Kemudian variabel **“region_northwest”**, **“region_southeast”**, **“region_southwest”** adalah variabel **“region”** yang dijadikan dummy dengan **“region_northeast”** menjadi *base level*.

Model 2 memiliki nilai *Adjusted R-Squared* mencapai 0.8625. Artinya, model regresi dijelaskan oleh variabel prediktor sebesar 0.8625. Sedangkan, 0.1375 sisanya dijelaskan oleh variabel prediktor lain di luar model regresi linear yang kami ajukan. Model 2 dipilih karena kami mempertimbangkan kemungkinan adanya interaksi dan pengaplikasian di kondisi nyata.

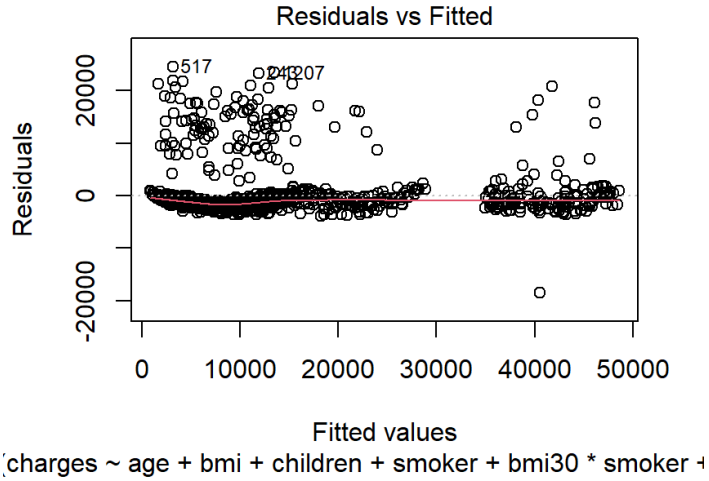
Akan diperiksa apakah asumsi normalitas pada residual, linearitas, homoskedastisitas, dan tidak adanya multikolinearitas pada model 2 terpenuhi atau tidak.

1) Asumsi Normalitas pada Residual



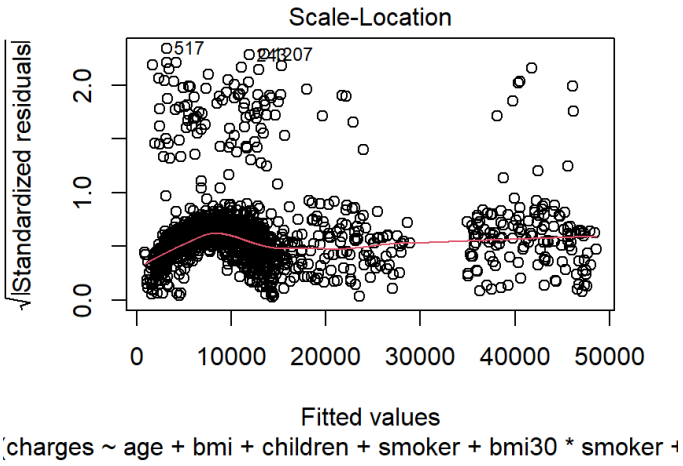
Dapat dilihat bahwa titik-titik dalam plot mengikuti garis diagonal. Sehingga asumsi normalitas pada residual terpenuhi.

2) Asumsi Linearitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola khusus. Sehingga asumsi linearitas terpenuhi.

3) Asumsi Homoskedastisitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola tertentu. Sehingga, asumsi homoskedastisitas terpenuhi.

4) Asumsi Tidak Ada Multikolinearitas

```
> vif(model3)
                 age                  bmi             children
              1.0167               2.9520               1.0039
           smokeryes                bmi30      regionnorthwest
              2.1337               3.0051               1.5193
regionsoutheast      regionsouthwest      smokeryes:bmi30
              1.6584               1.5303               2.3885
```

Dapat dilihat nilai VIF (*Variance Inflation Factor*) untuk setiap variabel prediktor di model 2 semuanya bernilai < 4 yang menandakan tidak adanya multikolinearitas. Sehingga, asumsi tidak ada multikolinearitas terpenuhi.

c. Model 3

Pada model 3, kami akan mengajukan model regresi linear yang dilakukan dari data mentah (bukan hasil *pre-processing*). Pada model 3, kami melakukan transformasi variabel dependen "charges" menjadi bentuk logaritmik. Hal ini dikarenakan kami mencoba untuk mencari model yang lebih bagus dengan mengganti bentuk variabel dependen **"charges"** yang mulanya linear menjadi logaritmik.

Berikut adalah model 3 yang akan kami ajukan.

```
> modeltransform<- lm(log(charges) ~ age + bmi + children + sex + smoker +
 region, data = insurance)
> summary(modeltransform)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.07186 -0.19835 -0.04917  0.06598  2.16636

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
age              0.0345816  0.0008721  39.655  < 2e-16 ***
bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
children         0.1018568  0.0100995  10.085  < 2e-16 ***
sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared:  0.7679,    Adjusted R-squared:  0.7666
F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Berdasarkan *summary* di atas, model regresi linear untuk model 3 adalah

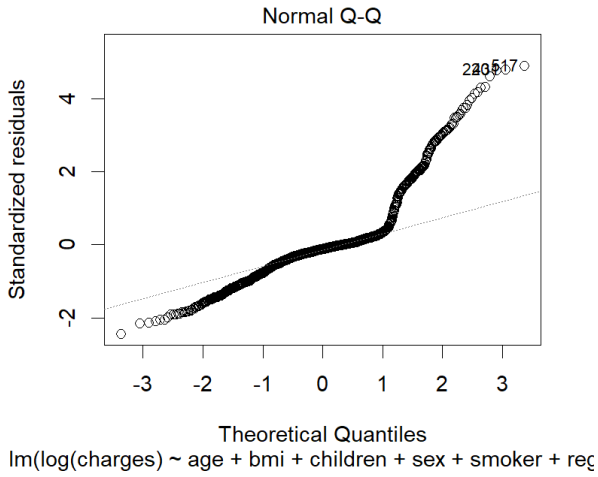
$$Y = 7.0305581 + 0.0345816x_{age} + 0.0133748x_{bmi} + 0.1018568x_{children} - 0.0754164x_{sex\_male} + 1.5543228x_{smoker\_yes} - 0.0637876x_{region\_northwest} - 0.1571967x_{region\_southeast} - 0.1289522x_{region\_southwest}$$

Dengan **“sex_male”** adalah variabel **“sex”** yang dijadikan dummy dengan **“sex_female”** menjadi *base level*. Kemudian, **“smoker_yes”** adalah variabel **“smoker”** yang dijadikan dummy dengan **“smoker_no”** menjadi *base level*. Selain itu, variabel **“region_northwest”**, **“region_southeast”**, **“region_southwest”** adalah variabel **“region”** yang dijadikan dummy dengan **“region_northeast”** menjadi *base level*.

Model 3 memiliki nilai *Adjusted R-Squared* mencapai 0.7666. Artinya, model regresi dijelaskan oleh variabel prediktor sebesar 0.7666. Sedangkan, 0.2334 sisanya dijelaskan oleh variabel prediktor lain di luar model regresi linear yang kami ajukan. Model 3 dipilih karena kami mempertimbangkan *Adjusted R-Squared* tertinggi apabila *outliers* pada data dihapus.

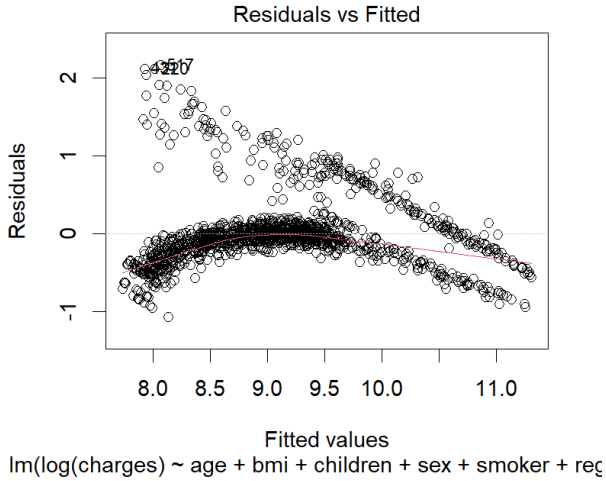
Akan diperiksa apakah asumsi normalitas pada residual, linearitas, homoskedastisitas, dan tidak adanya multikolinearitas pada model 3 terpenuhi atau tidak.

1) Asumsi Normalitas pada Residual



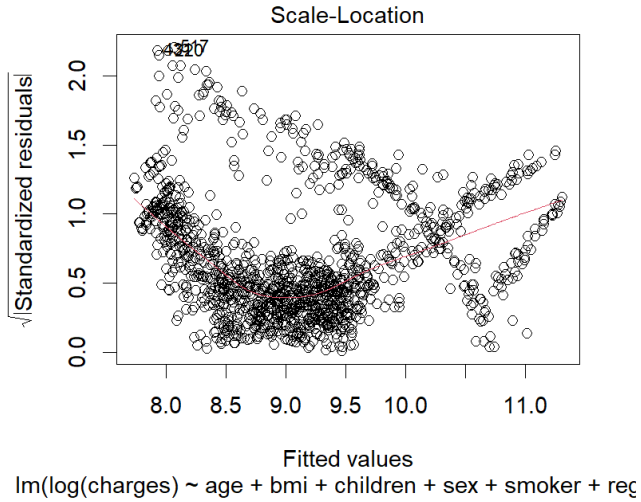
Dapat dilihat bahwa titik-titik dalam plot mengikuti garis diagonal. Sehingga asumsi normalitas pada residual terpenuhi.

2) Asumsi Linearitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola khusus. Sehingga asumsi linearitas terpenuhi.

3) Asumsi Homoskedastisitas



Dapat dilihat bahwa titik-titik pada plot tersebar merata tanpa pola tertentu. Sehingga, asumsi homoskedastisitas terpenuhi.

4) Asumsi Tidak Ada Multikolinearitas

```
vif(modeltransform)
          age           bmi      children       sexmale
       1.0168        1.1066        1.0040        1.0089
    smokeryes regionnorthwest regionsoutheast regionsouthwest
       1.0121        1.5188        1.6522        1.5294
```

Dapat dilihat nilai VIF (*Variance Inflation Factor*) untuk setiap variabel prediktor di model 2 semuanya bernilai < 4 yang menandakan tidak adanya multikolinearitas. Sehingga asumsi tidak ada multikolinearitas.

d. Model Terbaik

Dari ketiga model regresi linear yang telah kami ajukan, kami memilih model 2 untuk dijadikan model yang terbaik. Hal ini didasarkan pada nilai *Adjusted R-Squared* tertinggi dimiliki oleh model 2, yaitu senilai 0.8625. Sehingga model terbaik untuk permasalahan pada data kami adalah sebagai berikut.

$$Y = -4967.297 + 263.663x_{age} + 114.090x_{bmi} + 516.796x_{children} + 13383.160x_{smoker\_yes} - 869.805x_{bmi30} - 264.050x_{region\_northwest} - 823.426x_{region\_southeast} - 1221.147x_{region\_southwest} + 19744.706x_{smoker\_yes:bmi30}$$

Bagian 4. Pengolahan data dan analisis hasil

4.1 Asumsi Model

Berdasarkan uji asumsi dengan menggunakan plot untuk model 2 yang sudah dilakukan sebelumnya, didapatkan bahwa model 2 sudah memenuhi asumsi normalitas, linearitas, dan homoskedastisitas. Namun, karena uji asumsi dengan menggunakan plot terkadang bersifat subjektif, maka akan dilakukan lagi uji asumsi lanjutan dengan menggunakan perhitungan yang lebih akurat.

a. Uji Normalitas Residual

Akan digunakan *Asymptotic one-sample Kolmogorov-Smirnov Test*.

- Tujuan:

  Untuk mengetahui model regresi terbaik yang kami ajukan memenuhi asumsi normalitas atau tidak.

- Hipotesis:

  H_0: Residual berdistribusi normal.

  H_1: Residual tidak berdistribusi normal.

- Taraf Signifikansi:

  $\alpha = 0.05$

- Statistik Uji:

  Dengan menggunakan *software RStudio* diperoleh *output* sebagai berikut.

```
> residuals <- residuals(model3)
> ks.test(model3$residuals,ecdf(model3$residuals))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  model3$residuals
D = 0.0014948, p-value = 1
alternative hypothesis: two-sided
```

- Aturan Keputusan:

  H_0 ditolak jika nilai $p - value < \alpha$. Karena $p - value = 1 > 0.05 = \alpha$, maka H_0 tidak ditolak.

- Kesimpulan:

  Dengan taraf signifikansi $\alpha = 0.05$, maka dapat disimpulkan bahwa residual data berdistribusi normal.

Dengan menggunakan *Asymptotic one-sample Kolmogorov-Smirnov Test*, dapat disimpulkan bahwa asumsi normalitas residual terpenuhi.

b. Uji Linearitas

Akan digunakan *Goldfeld-Quandt Test*.

- Tujuan:

  Untuk mengetahui model regresi terbaik yang kami ajukan memenuhi asumsi linearitas atau tidak.

- Hipotesis:

  H_0: Tidak ada hubungan non-linear antara variabel independen dan variabel dependen.

  H_1: Terdapat hubungan non-linear antara variabel independen dan variabel dependen.

- Taraf Signifikansi:

  $\alpha = 0.05$

- Statistik Uji:

  Dengan menggunakan *software RStudio* diperoleh *output* sebagai berikut.

```
> gq_test <- gqtest(model3)
> gq_test

        Goldfeld-Quandt test

data:  model3
GQ = 0.83556, df1 = 659, df2 = 659, p-value = 0.9894
alternative hypothesis: variance increases from segment 1 to 2
```

- Aturan Keputusan:

  H_0 ditolak jika nilai $p - value < \alpha$. Karena $p - value = 0.9894 > 0.05 = \alpha$, maka H_0 tidak ditolak.

- Kesimpulan:

  Dengan taraf signifikansi $\alpha = 0.05$, maka dapat disimpulkan bahwa tidak ada hubungan non-linear antara variabel independen dan variabel dependen.

Dengan menggunakan *Goldfeld-Quandt Test*, dapat disimpulkan bahwa asumsi linearitas terpenuhi.

c. Uji Homoskedastisitas

Akan digunakan *Breusch-Pagan Test*.

- Tujuan:

  Untuk mengetahui model regresi terbaik yang kami ajukan memenuhi asumsi homoskedastisitas atau tidak.

- Hipotesis:

  H_0: Sebaran data cukup seragam.

  H_1: Sebaran data tidak seragam.

- Taraf Signifikansi:

  $\alpha = 0.05$

- Statistik Uji:

  Dengan menggunakan *software RStudio* diperoleh *output* sebagai berikut.

```
> bptest(model3)

        studentized Breusch-Pagan test

data:  model3
BP = 4.9534, df = 9, p-value = 0.8384
```

- Aturan Keputusan:

  H_0 ditolak jika nilai $p - value < \alpha$. Karena $p - value = 0.8384 > 0.05 = \alpha$, maka H_0 tidak ditolak.

- Kesimpulan:

  Dengan taraf signifikansi $\alpha = 0.05$, maka dapat disimpulkan bahwa seragam data cukup seragam.

Dengan menggunakan *Breusch-Pagan Test*, dapat disimpulkan bahwa asumsi homoskedastisitas terpenuhi.

4.2 Multikolinearitas

Salah satu metode statistika yang bisa digunakan untuk menguji multikolinearitas adalah menggunakan *Variance Inflation Factor* (VIF). Sederhananya, VIF adalah salah satu cara untuk mengukur efek multikolinearitas di antara prediktor dalam model. Akan dicek multikolinearitas untuk setiap variabel.

```
> vif(model3)
                age                 bmi            children
             1.0167              2.9520              1.0039
          smokeryes               bmi30    regionnorthwest
             2.1337              3.0051              1.5193
regionsoutheast    regionsouthwest    smokeryes:bmi30
             1.6584              1.5303              2.3885
```

Dari hasil yang diperoleh, semua variabel memiliki nilai VIF < 4, sehingga bebas dari indikasi multikolinearitas.

4.3 Analisis Model

Dari model regresi terbaik yang kami pilih, berikut *summary* dari model tersebut.

```
> insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
> model3 <- lm(formula = charges ~ age + bmi + children + smoker + bmi30*smoker +
 region, data = insurance)
> summary(model3)

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

Residuals:
     Min       1Q   Median       3Q      Max
-18422.1  -1845.2  -1276.1   -441.8  24566.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4967.297    954.272  -5.205 2.24e-07 ***
age               263.663      8.812  29.920  < 2e-16 ***
bmi               114.090     34.595   3.298 0.001000 **
children          516.796    102.056   5.064 4.69e-07 ***
smokeryes       13383.160    444.301  30.122  < 2e-16 ***
bmi30            -869.805    426.244  -2.041 0.041485 *
regionnorthwest  -264.050    352.801  -0.748 0.454328
regionsoutheast  -823.426    355.196  -2.318 0.020588 *
regionsouthwest -1221.147    354.076  -3.449 0.000581 ***
smokeryes:bmi30 19744.706    610.253  32.355  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4490 on 1328 degrees of freedom
Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8625
F-statistic: 933.2 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Dengan model ini, dapat dilihat bahwa nilai $p - value = 2.2e - 16 < 0.05 = \alpha$, sehingga H_0 model pada bagian 2 ditolak. Dengan demikian, dengan taraf signifikansi $\alpha = 0.05$, dapat disimpulkan bahwa model regresi terbaik yang kami ajukan berguna.

4.4 Korelasi Setiap Variabel dari Model

Selanjutnya, kami akan menganalisis korelasi antara variabel prediktor dan variabel respons. Ini dapat dilakukan dengan menggunakan model yang dibuat untuk mengetahui korelasi nilai parameter yang diperkirakan untuk masing-masing variabel. Estimasi parameter menunjukkan nilai positif, dan estimasi parameter menunjukkan nilai negatif.

- **"age"** dengan **"charges"** memiliki korelasi linier yang positif. Berarti, semakin tua usia seseorang, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin tinggi.
- **"bmi"** dengan **"charges"** memiliki korelasi linier yang positif. Berarti, semakin tinggi *body mass index* seseorang, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin tinggi.

- **"children"** dengan "charges" memiliki korelasi linier yang positif. Berarti, semakin banyaknya jumlah anak seseorang, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin tinggi.
- **"smoker_yes"** dan **"charges"** memiliki korelasi linier yang positif. Berarti, individu yang memiliki kebiasaan merokok, tagihan biaya asuransi kesehatan yang harus dibayar akan lebih tinggi.
- **"region_northwest"** dan "charges" memiliki korelasi linier yang negatif. Berarti, wilayah yang semakin terletak di sebelah barat laut, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin rendah.
- **"region_southeast"** dan **"charges"** memiliki korelasi linier yang negatif. Berarti, wilayah yang semakin terletak di sebelah tenggara, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin rendah.
- **"region_southwest"** dan **"charges"** memiliki korelasi linier yang negatif. Berarti, wilayah yang semakin terletak di sebelah barat daya, tagihan biaya asuransi kesehatan yang harus dibayar akan semakin rendah.
- Interaksi antara **"smoker_yes"** dengan **"bmi30"** memiliki korelasi linier yang positif. Berarti, individu yang memiliki kebiasaan merokok dan memiliki permasalahan obesitas, tagihan biaya asuransi kesehatan yang harus dibayar akan lebih tinggi.

4.5 Variabel yang paling mempengaruhi

Dalam menentukan variabel prediktor mana yang paling mempengaruhi variabel respon, dapat dilihat dari nilai korelasi yang tinggi terhadap variabel respon dan nilai VIF yang cenderung rendah. Sebelumnya, telah dilakukan proses untuk mencari nilai korelasi antara variabel prediktor dan variabel respon dengan hasil sebagai berikut.

```
Correlation with charges variable:
charges     1.000000
smoker      0.569117
age         0.520632
children    0.112928
bmi        -0.064507
Name: charges, dtype: float64
```

Dapat dilihat bahwa variabel prediktor **"smoker"** dan **"age"** merupakan dua variabel prediktor tertinggi yang memiliki korelasi yang tinggi dengan variabel respon **"charges"**. Kemudian, berikut nilai VIF dari masing-masing variabel prediktor pada model regresi terbaik yang kami ajukan.

```
> vif(model3)
            age             bmi        children
         1.0167          2.9520          1.0039
      smokeryes           bmi30 regionnorthwest
         2.1337          3.0051          1.5193
regionsoutheast regionsouthwest smokeryes:bmi30
         1.6584          1.5303          2.3885
```

Dapat dilihat bahwa variabel prediktor **"children"** dan **"age"** merupakan dua variabel prediktor tertinggi yang memiliki nilai VIF yang rendah. Variabel prediktor **"smoker"** memiliki nilai VIF yang lebih tinggi, yaitu 2.1337. Sebenarnya, nilai VIF 2.1337 masih belum menjadi masalah multikolinearitas. Namun, ketika nilai VIF = 1 memiliki arti tidak ada multikolinearitas dengan variabel prediktor lain. Karena variabel prediktor **"age"** memiliki nilai VIF = 1.0167 dan memiliki interaksi yang cukup tinggi dengan variabel respon **"charges"**, maka dapat ditarik kesimpulan bahwa variabel prediktor yang paling mempengaruhi variabel respon **"charges"** adalah variabel **"age"**.

Bagian 5. Penutup

Setelah dilakukan analisis terhadap pengolahan data yang kami dapatkan, dapat diambil kesimpulan bahwa tidak semua variabel prediktor yang ada mempengaruhi variabel respon. Variabel prediktor yang tidak mempengaruhi variabel respon dan tidak kami masukkan ke dalam model adalah variabel **"sex"**. Variabel-variabel prediktor yang mempengaruhi variabel respon walaupun tingkat korelasinya bervariasi dari kecil hingga besar adalah umur, indeks massa tubuh, memiliki anak/tanggungan atau tidak, merupakan individu perokok atau tidak, memiliki permasalahan obesitas atau tidak, serta area tempat tinggal.

Berdasarkan analisis korelasi dengan variabel respon **"charges"** dan nilai VIF, dapat dilihat bahwa variabel prediktor yang paling mempengaruhi model adalah variabel **"age"** atau umur dari penerima bantuan asuransi kesehatan. Kemudian, hasil *Adjusted R-Square* dari model regresi terbaik yang kami pilih juga sudah menunjukkan korelasi yang kuat dengan nilai 0.8625. Masing-masing variabel prediktornya juga memiliki signifikansi yang tinggi, dapat dilihat dari nilai $Pr(>|t|)$ yang bernilai kurang dari $\alpha = 0.05$, kecuali pada variabel prediktor **"region_northwest"**. Hal ini berarti, variabel prediktor **"region_northwest"** kurang memiliki pengaruh yang signifikan terhadap variabel respon **"charges"**.

Bagian 6. Lampiran

Link Google Drive:
https://drive.google.com/drive/folders/1utPYw0rGaPw46PcOlchSXA5t1eDO-VSr

Link Google Colab (Preprocessing):
https://colab.research.google.com/drive/13BRYEcUQhePLgiMQXK2o-OdFsEfAzMig?usp=sharing

Link Kaggle: https://www.kaggle.com/datasets/mirichoi0218/insurance/

Penilaian Presentasi:

- Dilakukan pada sesi kelas masing-masing di Minggu kedua periode UAS.
- Durasi waktu Max 7 menit.
- Komponen penilaian mengacu pada rubrik yang ada (lihat file excel: Rubrik penilaian presentasi final)
- Urutan presentasi kelompok dilakukan secara acak