

# Final Project - Health Insurance Analysis

Probability - Sekolah Data Pacmann

Submitted by: Dewi Astuti (dewi-l4Zs)  
Class: Continuation (BI+DS)



# Daftar Isi

---

- I. Pembukaan
- II. Objektif
- III. Dataset
- IV. Analisa
  - 1) Descriptive Statistic Analysis
  - 2) Categorical Variables Analysis
  - 3) Continuous Variables Analysis
  - 4) Variables Correlation
  - 5) Hypothesis Testing
- V. Kesimpulan
- VI. Rekomendasi
- VII. Referensi



# Pembukaan



# Pembukaan

---

- Asuransi kesehatan adalah instrumen jaminan kesehatan di mana pengguna asuransi memiliki kewajiban membayar biaya rutin (premi) kepada penyedia asuransi. Premi tersebut diolah oleh penyedia asuransi untuk membayarkan klaim tagihan kesehatan pengguna asuransi.
- Oleh karena itu penyedia asuransi melakukan analisis yang dalam untuk menentukan besar premi dengan mempertimbangkan berbagai faktor.
- Analisa eksplorasi akan dilakukan berdasarkan variabel yang diketahui berkorelasi dengan tagihan kesehatan tersebut. Metode yang digunakan untuk menganalisa adalah:
  1. Analisis Deskriptif Statistik (Descriptive Statistic Analysis)
  2. Analisis Variabel Kategorik (Categorical Variables Analysis)
  3. Analisis Variabel Kontinu (Continuous Variables Analysis)
  4. Analisis Korelasi Variabel (Variables Correlation)
  5. Pengujian Hipotesis (Hypothesis Testing)

# Objektif



# Objektif

---

1. Menganalisa variabel-variabel yang memiliki hubungan dengan tagihan kesehatan yang diterima setiap pengguna.
2. Sepintas, bmi dan perokok cenderung menyebabkan biaya pengobatan yang tinggi bagi seseorang, sementara umur, jenis kelamin, anak-anak dan wilayah dapat berkontribusi dalam beberapa hal atau yang lain.
3. Nantinya akan bisa menghitung setiap probabilitas dari variabel-variabel yang bersangkutan dengan data-data di atas.



# Dataset

---



# Dataset

---

Dataset yang digunakan adalah tagihan asuransi kesehatan dengan **1,338 baris** yang merepresentasikan jumlah pengguna asuransi dengan variabel berikut:

- 1. Age:** umur nasabah pengguna asuransi. Tipe: numerik 18-64.
- 2. Sex:** Jenis kelamin. Tipe: teks 'female', 'male'.
- 3. BMI: Body Mass Index** dalam kg/m<sup>2</sup>. Tipe: numerik.
- 4. Children:** Jumlah anak yang ditanggung oleh asuransi. Tipe: numerik.
- 5. Smoker:** Apakah pengguna perokok atau bukan. Tipe: teks 'yes', 'no'.
- 6. Region:** Wilayah tempat tinggal pengguna. Tipe: teks 'northeast', 'southeast', 'southwest', dan 'northwest'.
- 7. Charges:** Jumlah tagihan kesehatan yang dicover asuransi. Tipe: numerik.

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.92400
18	male	33.770	1	no	southeast	1725.55230
28	male	33.000	3	no	southeast	4449.46200
33	male	22.705	0	no	northwest	21984.47061
32	male	28.880	0	no	northwest	3866.85520

# # 1. Descriptive Statistics Analysis



# Analisa Variabel Numerik

## Dari 1,338 data, didapatkan data umur

Rata-rata = 39.21 tahun.

Min = 18 tahun.

Max = 64 tahun.

## Dari 1,338 data, didapatkan data BMI

Rata-rata = 30.66.

Min = 15.96.

Max = 53.13.

## Dari 1,338 data, didapatkan data anak

Rata-rata = 1.09.

Min = 0.

Max = 5.

	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

## Dari 1,338 data, didapatkan data tagihan

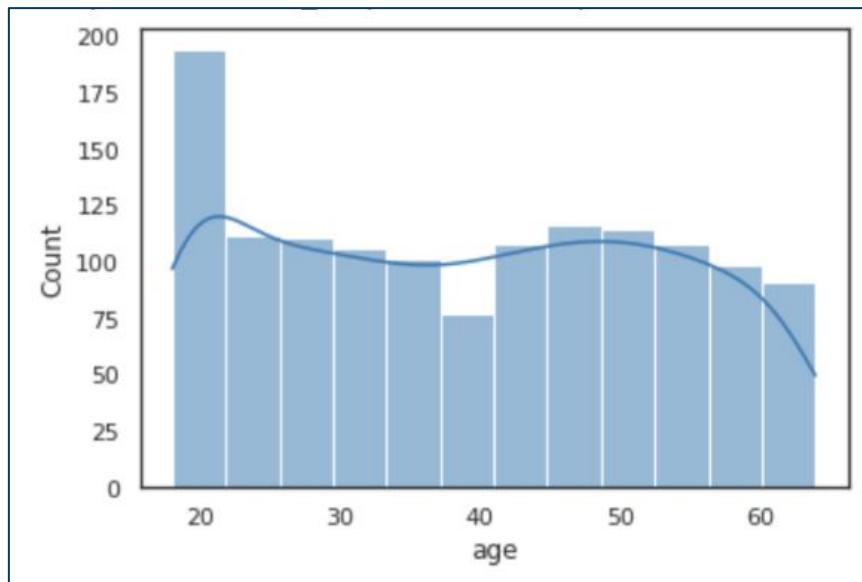
Rata-rata = \$13,270.

Min = \$1,121.

Max = \$ 63,770.

# Rata-rata Umur per Kategori

Data umur tidak terdistribusi normal, skewed positif. Dari 1,338 data, didapatkan rata-rata umur per kategori:



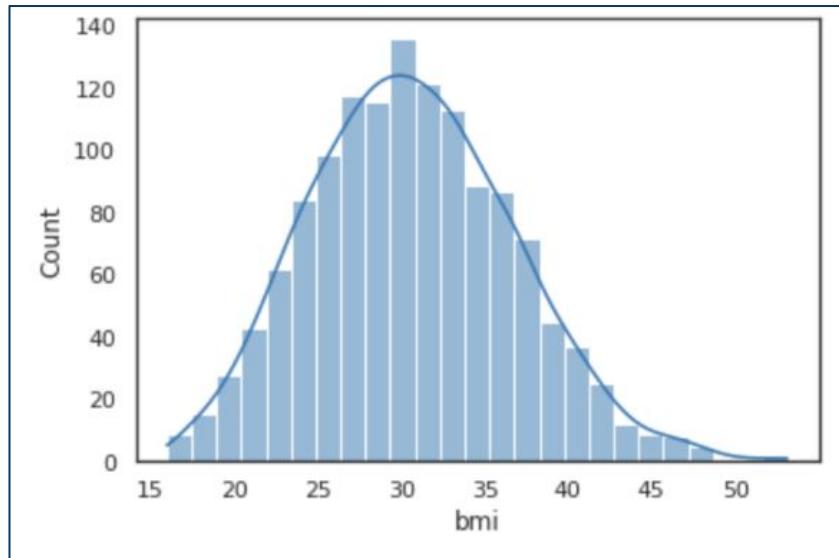
sex	age
female	39.50
male	38.92

sex	smoker	age
female	no	39.69
	yes	38.61
male	no	39.06
	yes	38.45

- Rata-rata umur perempuan lebih tinggi dibanding yang laki-laki meski tidak berbeda jauh.
- Rata-rata umur non-perokok baik yang perempuan dan laki-laki lebih tinggi dibanding yang perokok meski tidak berbeda jauh.

# Rata-rata BMI per Kategori

Data BMI terdistribusi normal dengan rata-rata BMI (telah diketahui) adalah 30.66  
Dari 1,338 data, didapatkan rata-rata BMI per kategori:



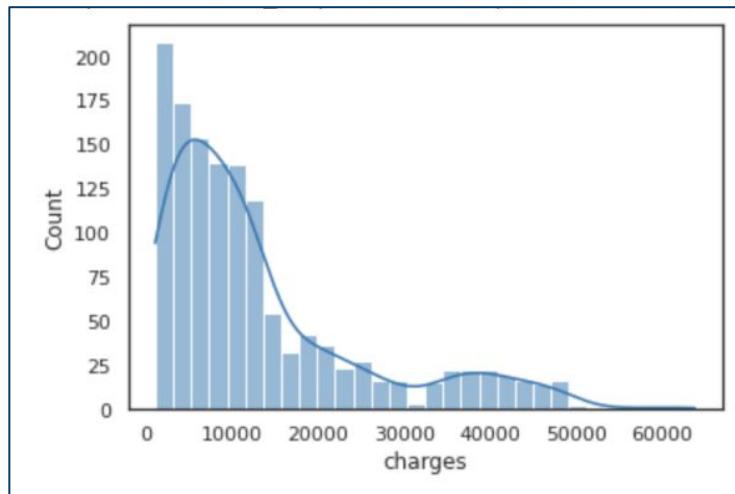
sex	bmi
female	30.38
male	30.94

sex	smoker	bmi
female	no	30.54
	yes	29.61
male	no	30.77
	yes	31.50

- Dengan standar batas ideal 25, maka rata-rata BMI tergolong *overweight*.
- Rata-rata BMI perempuan lebih rendah dibanding yang laki-laki meski tidak berbeda jauh.
- Rata-rata BMI perokok lebih tinggi dibanding non-perokok meski tidak berbeda jauh.

# Rata-rata Tagihan per Kategori

Data tagihan tidak terdistribusi normal, skewed positif. Dari 1,338 data, didapatkan rata-rata tagihan per kategori:



charges	
sex	charges
female	12570.0
male	13957.0

charges	
smoker	charges
no	8434.0
yes	32050.0

charges		
sex	smoker	charges
female	no	8762.0
female	yes	30679.0
male	no	8087.0
male	yes	33042.0

- Rata-rata tagihan perokok hampir 4 kali daripada non-perokok.
- Rata-rata tagihan perempuan lebih rendah dibanding yang laki-laki meski tidak berbeda jauh.
- Rata-rata tagihan non-perokok perempuan lebih tinggi dibanding yang non-perokok laki-laki meski tidak berbeda jauh.
- Rata-rata tagihan perokok perempuan lebih rendah dibanding yang laki-laki dan keduanya jauh lebih tinggi dari rata-rata tagihan keseluruhan.

# Variansi & Standar Deviasi

Dari 1,338 data pengguna asuransi, didapatkan variansi tagihan kesehatan dan standar deviasi.



Total pengguna yang non-perokok itu lebih banyak, tapi variansi data tagihan non-perokok \$35,925,420, jauh lebih kecil dibanding yang perokok \$133,207,311 .

Variansi kemudian diubah menjadi standar deviasi dan menunjukkan bahwa standar deviasi tagihan non-perokok adalah \$5,994 dan perokok adalah \$11,542. Standar deviasi lebih kecil dari nilai rata-rata (\$13,270) yang berarti **nilai rata-rata adalah representasi yang baik dari keseluruhan data**.

# Rata-rata Tagihan BMI>25, Perokok dan Non-Perokok

Melakukan perhitungan rata-rata tagihan pengguna dengan BMI >25 dan perokok atau non-perokok dengan menggunakan fungsi mean()).

```
bmi_over_25 = insurance.loc[insurance.bmi >25]
bmi_over_25.loc[bmi_over_25.smoker == 'yes'].charges.mean(), bmi_over_25.loc[bmi_over_25.smoker == 'no'].charges.mean()

(35116.90965694064, 8629.589609712157)
```

Rata-rata tagihan pengguna dengan BMI >25 yang perokok adalah \$35,116 **lebih tinggi hampir 4 kali lipat** dibanding pengguna dengan BMI>25 yang non-perokok yaitu \$8,629.

# #2. Categorical Variables Analysis

# Proporsion Tagihan Berdasarkan Jenis Kelamin

Mencari gender mana yang memiliki tagihan paling tinggi.

Tagihan tertinggi pengguna perempuan adalah \$ 63770

Tagihan tertinggi pengguna laki-laki adalah \$ 62593

Tagihan tertinggi ada di data perempuan dengan selisih \$1,178.

# Proporsi & Probabilitas Perokok per Gender

Dari 1,338 data pengguna asuransi, didapatkan 79.52% adalah perokok dan 20.48% adalah non-perokok.

smoker	total	percentage %
no	1064	79.52
yes	274	20.48

Mencari probability perokok per gender, kita gunakan rumus:

$P(F|S) = N(F \& S)/N(S)$  sehingga didapatkan hasil sebanyak 0.42 perokok adalah seorang perempuan.

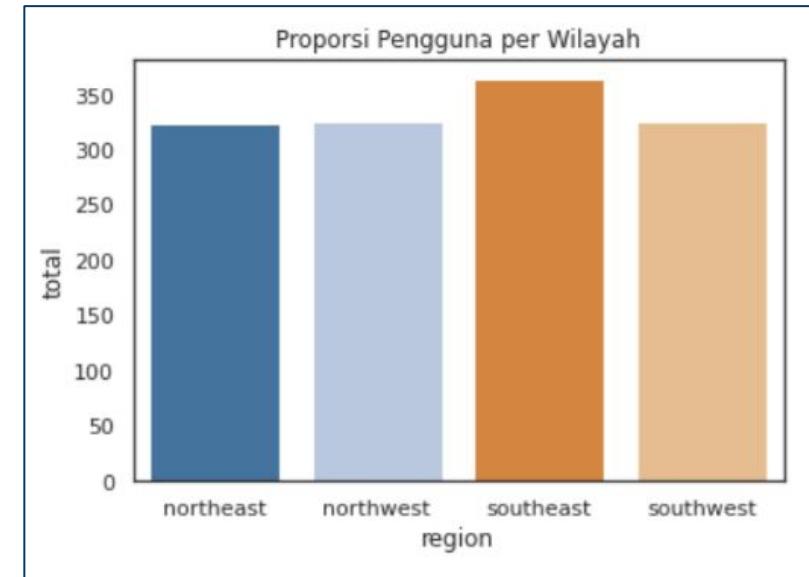
$P(M|S) = N(M \& S)/N(S)$  sehingga didapatkan hasil sebanyak 0.58 perokok adalah seorang laki-laki.

sex	total_smoker	probability
female	115	0.42
male	159	0.58

# Proporsi Pengguna per Wilayah

Dari 1,338 data pengguna asuransi, didapatkan proporsi data pengguna asuransi per wilayah.

region	total	percentage %
northeast	324	24.22
northwest	325	24.29
southeast	364	27.20
southwest	325	24.29

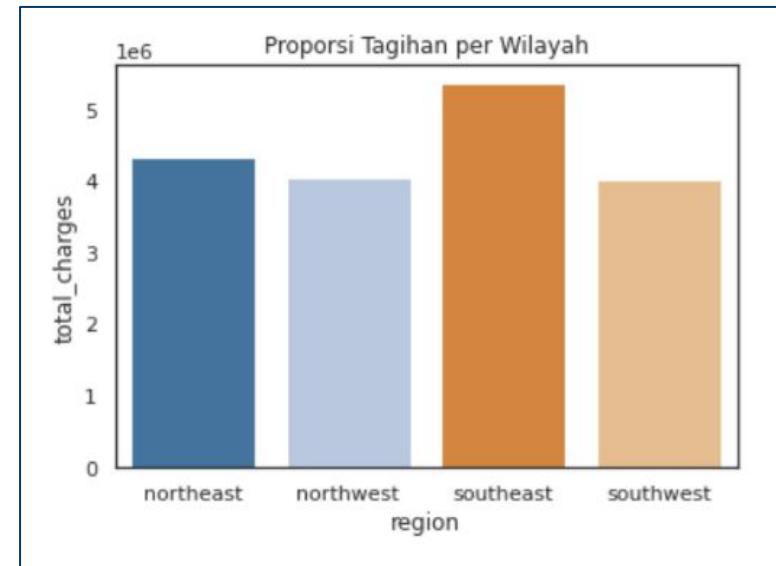


Proporsi banyak orang di tiap wilayah cukup merata dengan proporsi tertinggi di Southeast.

# Proporsi Tagihan per Wilayah

Dari 1,338 data pengguna asuransi, didapatkan proporsi tagihan di tiap wilayah.

region	total_charges	percentage %
northeast	4343669.0	24.46
northwest	4035712.0	22.73
southeast	5363690.0	30.21
southwest	4012755.0	22.60



Berkorelasi positif dengan proporsi jumlah pengguna, peluang tagihan di Southeast juga memiliki proporsi tertinggi yaitu 30%. Region northwest dan southwest memiliki proporsi 23% dan northeast 24%.

# #3. Continuous Variables Analysis

---

# Peluang Besar Tagihan Berdasarkan BMI

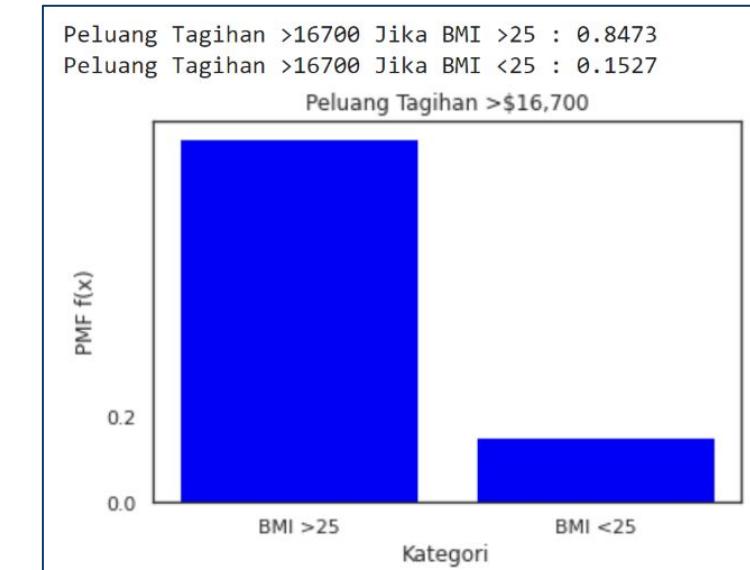
Perbandingan dari:

Peluang BMI >25, Tagihan >\$16,700 adalah  $P(\text{BMI} > 25 | T > 16,700)$

Peluang BMI <25, Tagihan >\$16,700 adalah  $P(\text{BMI} < 25 | T > 16,700)$

Hasil dari peluang seorang pengguna dengan BMI>25 mendapat tagihan >\$16,700 adalah 0.85

Hasil dari peluang seorang pengguna dengan BMI<25 mendapat tagihan >\$16,700 adalah 0.15.



Bisa disimpulkan bahwa semakin besar BMI seorang pengguna, semakin besar peluang tagihan >\$16,700.

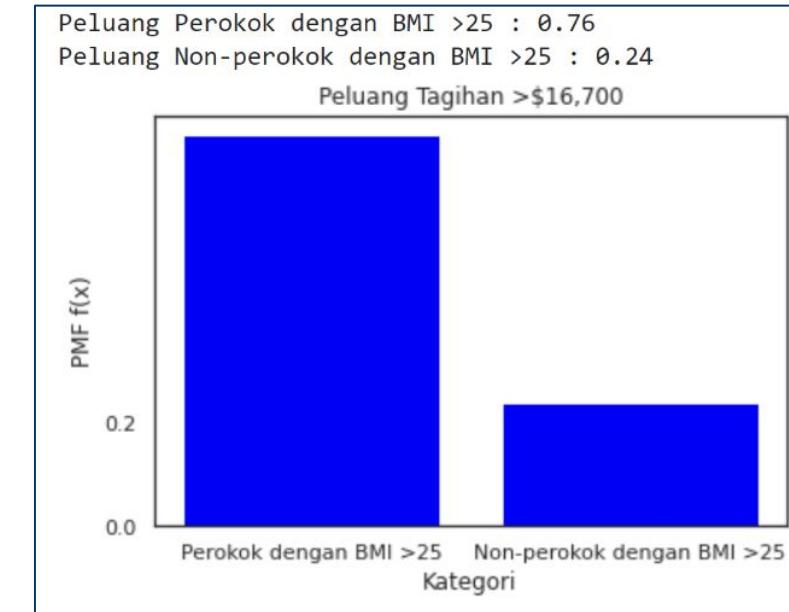
# Peluang Pengguna Perokok dengan BMI <25, Tagihan >16700

Peluang Perokok dengan BMI >25, Tagihan >\$16,700 adalah  $P(S \& \text{BMI} >25 | T>16,700)$

Peluang Non-perokok dengan BMI >25, Tagihan >\$16,700 adalah  $P(NS \& \text{BMI} >25 | T>16,700)$

Hasil dari peluang seorang perokok dengan BMI>25 mendapat tagihan >\$16,700 adalah 0.76.

Hasil dari peluang seorang non-perokok dengan BMI>25 mendapat tagihan >\$16,700 adalah 0.24.



Pengguna dengan BMI >25 yang perokok lebih mungkin mendapat tagihan >\$16,700 dibanding pengguna dengan BMI>25 yang non-perokok.

# Peluang Tagihan 16700 Jika Diketahui Perokok

Mencari CDF dengan menghitung proporsi dengan conditional CDF.

$$P(T > 16700 | S)$$

```
smoker_charges_f = len(insurance[(insurance["smoker"] == "yes") & (insurance ["charges"] >16700)])
smoker_len = len(insurance[insurance["smoker"] == "yes"])
pmf = round(smoker_charges_f / smoker_len, 2)
print(f'Peluang Tagihan 16700 Jika Perokok : {pmf}')
```

```
Peluang Tagihan 16700 Jika Perokok : 0.93
```

Artinya ada peluang 93% seorang pengguna yang merokok memiliki tagihan >\$16,700.

## #4. Variables Correlation

---

# Korelasi

Dari 1,338 data pemegang polis didapatkan koefisien korelasi ‘r’:

- Antara BMI dengan umur = 0.11
- Antara jumlah anak yang ditanggung dengan umur = 0.04
- Antara tagihan dengan umur = 0.30
- Antara tagihan dengan BMI = 0.20
- Antara tagihan dengan jumlah anak yang ditanggung = 0.07

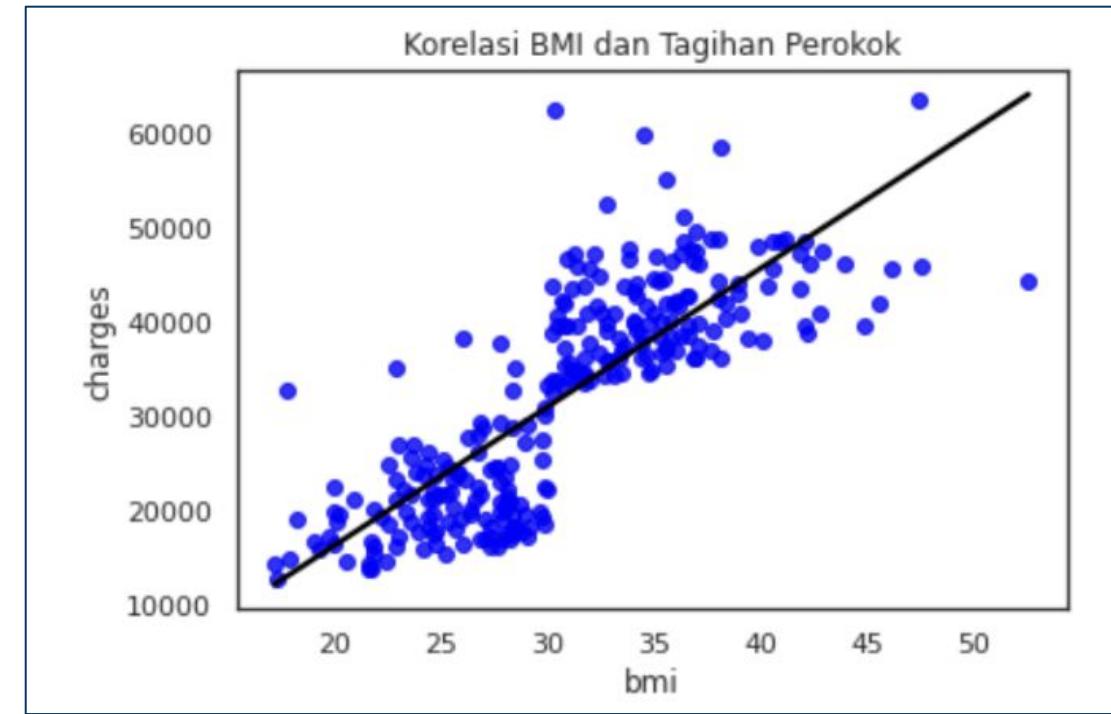
	<i>Age</i>	<i>Sex ID</i>	<i>BMI</i>	<i>Children</i>	<i>Smoker ID</i>	<i>Region ID</i>	<i>Charges</i>
<i>Age</i>	1						
<i>Sex ID</i>	(0.02)	1.00					
<i>BMI</i>	0.11	0.05	1.00				
<i>Children</i>	0.04	0.02	0.01	1.00			
<i>Smoker ID</i>	(0.03)	0.08	0.00	0.01	1.00		
<i>Region ID</i>	0.00	0.00	0.16	0.02	(0.00)	1.00	
<i>Charges</i>	0.30	0.06	0.20	0.07	0.79	(0.01)	1.00

- Umur tidak berkorelasi dengan jumlah anak yang ditanggung maupun BMI
- Korelasi antara umur dengan tagihan lebih tinggi dibanding korelasi antara BMI dengan tagihan.
- Jumlah anak yang ditanggung tidak berkorelasi dengan tagihan.
- Chart ini menunjukkan variabel yang bersifat multikolinear dan variabel yang memiliki collinearity tinggi dengan variabel target (charges).
- Ini mengungkapkan bahwa perilaku merokok diikuti oleh umur dan BMI sangat berkorelasi dengan biaya tagihan.

# Korelasi antara BMI dan Tagihan Perokok

BMI pengguna yang perokok memiliki korelasi positif dengan tagihan dengan score 0.81

```
charges      1.000000
bmi          0.806481
age          0.368224
children     0.035945
Name: charges, dtype: float64
```



Semakin tinggi BMI pengguna, semakin besar tagihannya.

# #5. Hypothesis Testing

# Tagihan Kesehatan Perokok Lebih Tinggi Dibanding Non-perokok

- H0:tagihan kesehatan perokok  $\geq$  tagihan kesehatan non-perokok
- H1: tagihan kesehatan perokok  $<$  tagihan kesehatan non-perokok
- nilai alpha/signifikansi= 0.05

Pengujian yang dilakukan adalah t-test.

Hipotesis awal adalah tagihan kesehatan perokok lebih tinggi dari non-perokok.

Hasil:

- Nilai p-value yang didapatkan adalah 1.
- Karena lebih besar dari alpha, maka kita gagal menolak null hypothesis.
- Dibutuhkan bukti statistik yang cukup untuk membuktikan klaim tersebut.

```
perokok= insurance[insurance['smoker']=='yes']
tagihan_perokok= perokok['charges']

non_perokok= insurance[insurance['smoker']=='no']
tagihan_non_perokok= non_perokok['charges']

print('variance dari tagihan perokok =', np.var(tagihan_perokok).round(2))
print('variance dari tagihan non perokok =', np.var(tagihan_non_perokok).round(2))

variance dari tagihan perokok = 132721153.14
variance dari tagihan non perokok = 35891656.0

stat, p_value= stats.ttest_ind(tagihan_perokok,
                                tagihan_non_perokok,
                                alternative='less', equal_var=False)

print(f'p-value: {p_value}')
print(f't-stats: {stat}')

p-value: 1.0
t-stats: 32.751887766341824

significance= 0.05
if p_value > significance:
    print('Gagal menolak null hypothesis')
else:
    print('menolak null hypothesis')

Gagal menolak null hypothesis
```

## Tagihan Kesehatan BMI>25 Lebih Tinggi Dibanding BMI <25

- H0: Tagihan kesehatan dengan BMI di atas 25  $\geq$  tagihan kesehatan dengan BMI di bawah 25
- H1: Tagihan kesehatan dengan BMI di atas 25 < tagihan kesehatan dengan BMI di bawah 25
- nilai alpha/signifikansi = 0.05

Pengujian yang dilakukan adalah t-test.

Hipotesis awal adalah tagihan kesehatan pengguna dengan BMI  $>25$  lebih tinggi dari pengguna dengan BMI  $<25$ .

Hasil:

- Nilai p-value yang didapatkan adalah 0.9999.
- Karena lebih besar dari alpha, maka kita gagal menolak null hypothesis.
- Dibutuhkan bukti statistik yang cukup untuk membuktikan klaim tersebut.

```
high_bmi_25= insurance[insurance['bmi'] > 25]
low_bmi_25= insurance[insurance['bmi'] < 25]

high_bmi_charges= high_bmi_25['charges']
low_bmi_charges= low_bmi_25['charges']

print('variance dari tagihan untuk bmi > 25 adalah', np.var(high_bmi_charges).round(2))
print('variance dari tagihan untuk bmi < 25 adalah', np.var(low_bmi_charges).round(2))

variance dari tagihan untuk bmi > 25 adalah 164579189.52
variance dari tagihan untuk bmi < 25 adalah 56326859.63

stat_charges, p_value_charges= stats.ttest_ind(high_bmi_charges,
                                                low_bmi_charges,
                                                alternative='less', equal_var=False)
print(f'p-value: {p_value_charges}')
print(f't-stats: {stat_charges}')

p-value: 0.999999974595514
t-stats: 5.929878344096734

significance= 0.05
if p_value > significance:
    print('Gagal menolak null hypothesis')
else:
    print('menolak null hypothesis')

Gagal menolak null hypothesis
```

# Tagihan Kesehatan Laki-laki Lebih Besar

- $H_0: \text{BMI laki-laki} = \text{BMI perempuan}$
- $H_1: \text{BMI laki-laki} \neq \text{BMI perempuan}$
- nilai alpha= 0.05

Pengujian yang dilakukan adalah t-test.

Hipotesis awal adalah tagihan kesehatan laki-laki sama dengan tagihan kesehatan perempuan.

Hasil:

- Nilai p-value yang didapatkan adalah 0.08999.
- Karena lebih besar dari alpha, maka kita gagal menolak null hypothesis.
- Dibutuhkan bukti statistik yang cukup untuk membuktikan klaim tersebut.

```
male= insurance[insurance['sex']=='male']
female= insurance[insurance['sex']=='female']

male_bmi= male['bmi']
female_bmi= female['bmi']

print('variance BMI laki-laki=', np.var(male_bmi))
print('variance BMI perempuan=', np.var(female_bmi))

variance BMI laki-laki= 37.64916073639534
variance BMI perempuan= 36.499177033798524

stat_bmi, p_value_bmi= stats.ttest_ind(male_bmi,
                                         female_bmi,equal_var=True)
print(f'p-value: {p_value_bmi}')
print(f't-stats: {stat_bmi}')

p-value: 0.08997637178984932
t-stats: 1.696752635752224

significance= 0.05
if p_value > significance:
    print('Gagal menolak null hypothesis')
else:
    print('menolak null hypothesis')

Gagal menolak null hypothesis
```

# Kesimpulan

---

# Kesimpulan

1. Berdasarkan hasil descriptive, categorical, dan continuous analysis, ditemukan bahwa faktor kebiasaan merokok pengguna asuransi memiliki korelasi yang kuat terhadap tagihan kesehatan. Perokok membayar tagihan yang lebih besar dibanding non-perokok.
2. Selain merokok, faktor BMI juga memiliki korelasi dengan tagihan, namun korelasinya cenderung lemah. Pengguna asuransi dengan  $BMI > 25$  memiliki indikasi akan membayar tagihan kesehatan lebih tinggi dibanding pengguna asuransi dengan  $BMI < 25$ .
3. Sementara, faktor umur juga memiliki korelasi positif dengan besar tagihan, di mana semakin tua seseorang, maka semakin besar tagihan tersebut, namun korelasinya cenderung lemah.
4. Faktor jenis kelamin tidak berpengaruh langsung terhadap tagihan, namun berpengaruh langsung terhadap nilai BMI dan kebiasaan merokok di mana persentase jenis kelamin seorang perokok adalah laki-laki lebih besar dibandingkan perempuan.
5. Sementara, kesimpulan dari analisis wilayah, didapati jumlah perokok dan rata-rata tagihan tertinggi ada di wilayah Southeast.

# Catatan

---

- Dapat dilakukan analisis yang lebih mendalam mengenai faktor jumlah anak yang ditanggung terhadap perilaku merokok dan tagihan kesehatan saat faktor lainnya sama.
- Dapat dilakukan analisis yang lebih mendalam mengenai riwayat penyakit yang memiliki korelasi kuat terhadap besar tagihan kesehatan.
- Dapat dilakukan uji statistik untuk beberapa macam jumlah sampel.
- Dapat dilakukan analisis regresi untuk mendapatkan hasil hubungan antara variabel untuk menentukan tagihan kesehatan.

# Referensi

- Introduction to Probability, Dimitri P. Bertsekas & John N. Tsitsiklis Chapter 4.5 Covariance and Correlation
- Probability and Statistics for Engineers and Scientist, Ronald E. Walpole et. all Chapter 10 - One and Two-Sample Tests of Hypotheses
- Rumus Varians Data Tunggal dan Varians Data Kelompok, Ibna Farabi, [Zenius.net](https://zenius.net).

