Pacmann

Analisa Faktor-Faktor yang Berhubungan dengan Tagihan Kesehatan Pada Pengguna Asuransi

Probability Course - Sekolah Data Pacmann

Oleh: Ni Putu Budi Setianingsih
Program: Data Science Batch 9

Outline

- Introduction
- Dataset
- Descriptive Statistic Analysis
- Categorical Variables Analysis
- Continuous Variables Analysis
- Variables Correlation
- Hypothesis Testing
- Conclusion

Introduction

Introduction

- Karakteristik pengguna asuransi yang beragam.
- Banyak variabel yang mungkin berhubungan dengan profil risiko pengguna asuransi.
- Perusahaan asuransi perlu menentukan nilai premi sesuai level risiko pengguna asuransi.

Dataset

Dataset

- Memuat data personal dari 1.338 pengguna asuransi.
- Terdiri dari 7 variabel:

	age	sex	bmi	children	smoker	region	charges	bmi_category	charges_category
0	19	female	27.900	0	yes	southwest	16884.92400	over	high
1	18	male	33.770	1	no	southeast	1725.55230	over	normal
2	28	male	33.000	3	no	southeast	4449.46200	over	normal
3	33	male	22.705	0	no	northwest	21984.47061	normal	high
4	32	male	28.880	0	no	northwest	3866.85520	over	normal
...	...	...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830	over	normal
1334	18	female	31.920	0	no	northeast	2205.98080	over	normal
1335	18	female	36.850	0	no	southeast	1629.83350	over	normal
1336	21	female	25.800	0	no	southwest	2007.94500	over	normal
1337	61	female	29.070	0	yes	northwest	29141.36030	over	high

1338 rows × 9 columns

1. Usia (*Age*), numerik.
2. Jenis kelamin (*Sex*), nominal.
3. BMI, numerik.
4. Jumlah anak (*Children*), ordinal.
5. Status Perokok (*Smoker*), nominal.
6. Wilayah (*Region*), nominal.
7. Tagihan (*Charges*), numerik.

Descriptive Statistics Analysis

Mean of Age

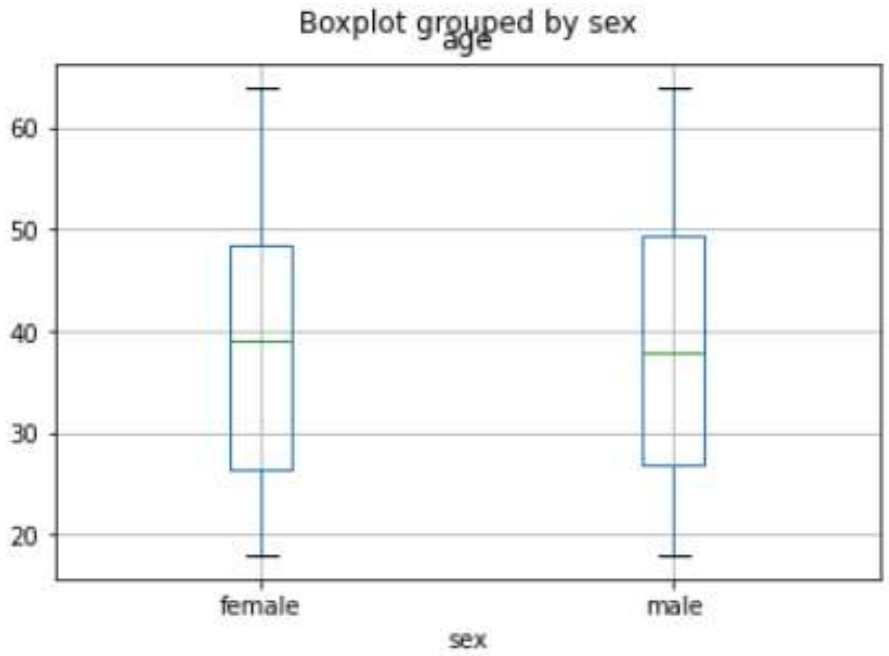
3. Berapa rata rata umur pada data tersebut?

Rata-rata = 38,61 tahun
Median/nilai tengah = 38,44 tahun

Kesimpulan:
Rata-rata umur pengguna asuransi adalah 38 tahun.

Mean of Age

6. Apakah rata rata umur perempuan dan laki-laki yang merokok sama?



Rata-rata umur perempuan merokok = 38,61
Rata-rata umur laki-laki merokok = 38,44.

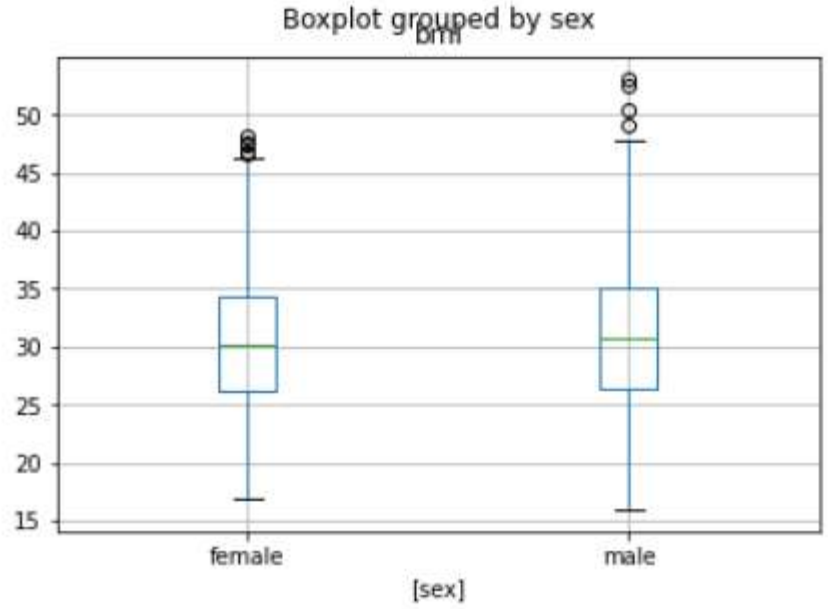
Kesimpulan:
Rata-rata umur perempuan dan laki-laki yang merokok sama, yaitu 38 tahun.

Mean of BMI

4. Berapa rata rata nilai BMI dari yang merokok?

Rata-rata BMI dari pengguna asuransi yang merokok adalah 30,71.

9. BMI mana yang lebih tinggi, seseorang laki-laki atau perempuan?



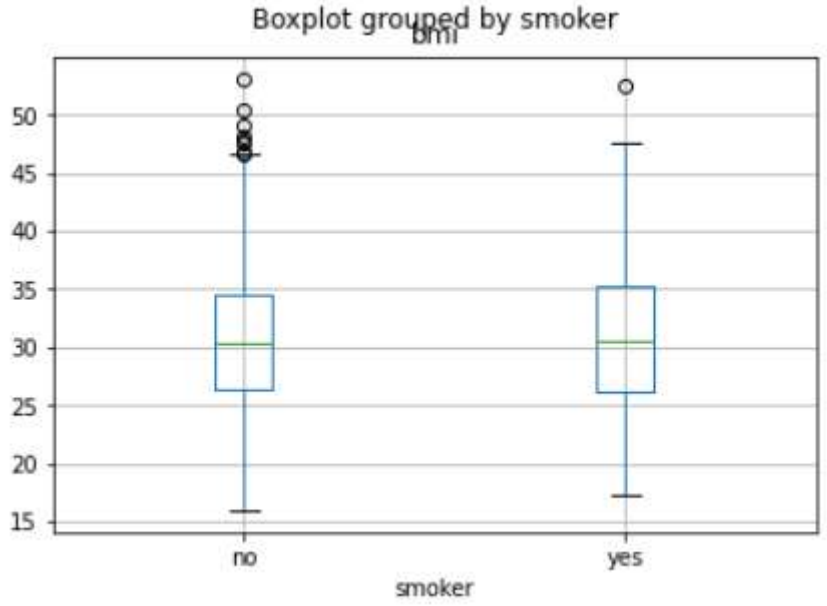
Kesimpulan:

Median/nilai tengah maupun rata-rata BMI dari pengguna laki-laki sedikit lebih tinggi daripada pengguna perempuan.

Rata-rata BMI pengguna laki-laki = 30,94 dan rata-rata BMI pengguna perempuan = 30,38.

Mean of BMI

10. BMI mana yang lebih tinggi, seseorang perokok atau non perokok?



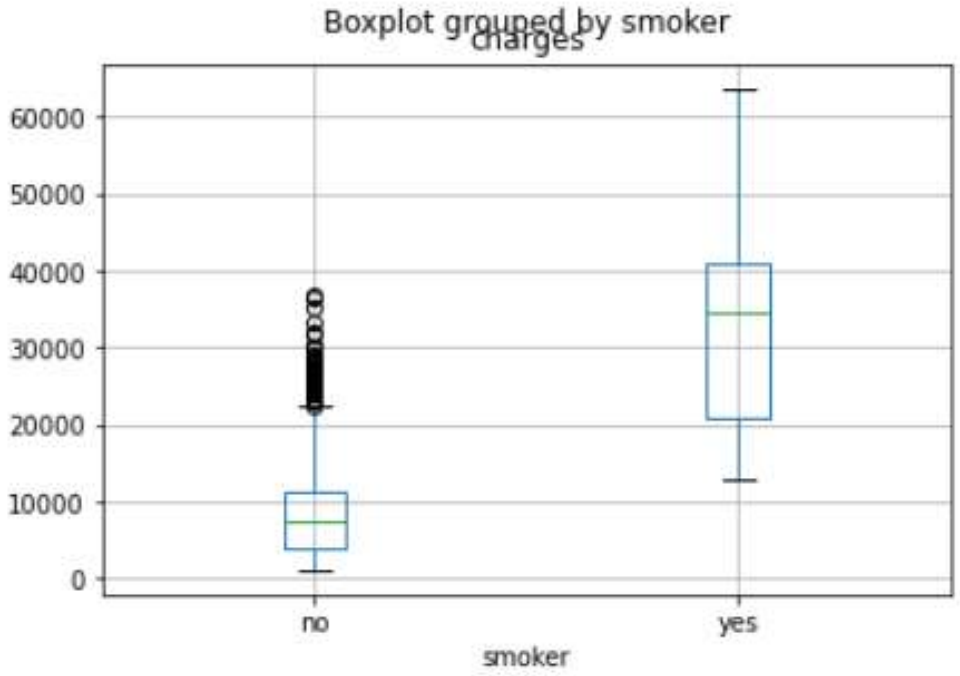
Kesimpulan:

Median/nilai tengah maupun rata-rata BMI dari pengguna perokok sedikit lebih tinggi daripada pengguna non perokok.

Rata-rata BMI pengguna perokok = 30,71 dan rata-rata BMI pengguna non perokok = 30,65.

Mean of Charges

7. Mana yang lebih tinggi, rata rata tagihan kesehatan perokok atau non merokok?
5. Apakah variansi dari data charges perokok dan non perokok sama?



Kesimpulan:

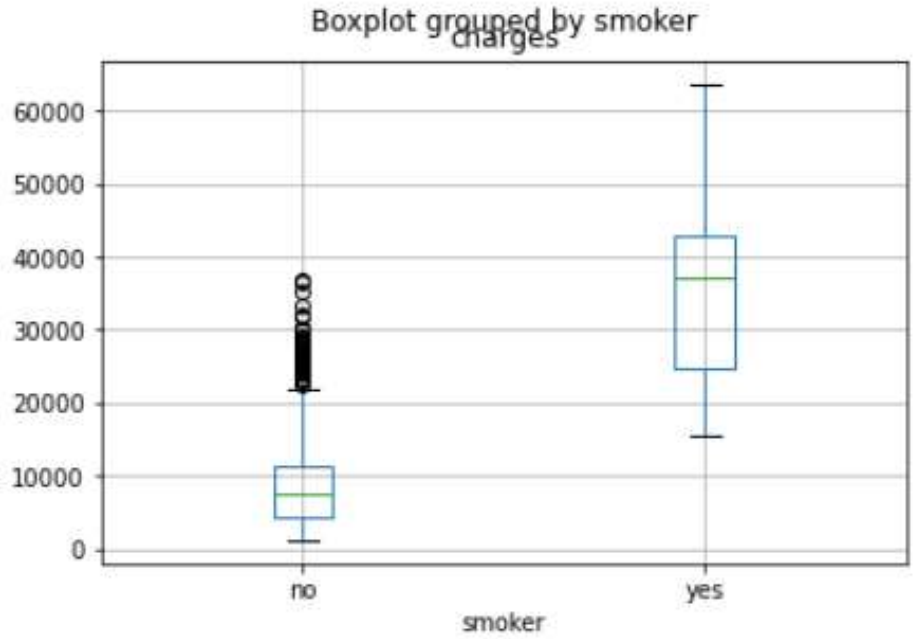
Median/nilai tengah maupun rata-rata dan variansi tagihan kesehatan perokok lebih tinggi daripada non perokok.

Rata-rata tagihan perokok = 32.050 USD
Rata-rata tagihan non perokok = 8.434 USD

Std. dev. tagihan perokok = 11.541 USD
Std. dev. tagihan non perokok = 5.993 USD

Mean of Charges

8. Mana yang lebih tinggi, rata rata tagihan kesehatan perokok dengan BMI di atas 25 atau non perokok dengan BMI di atas 25?



Kesimpulan:

Pada pengguna asuransi dengan BMI di atas 25, median/nilai tengah maupun rata-rata dan variansi tagihan kesehatan perokok lebih tinggi daripada non perokok.

Untuk pengguna asuransi dengan BMI di atas 25:
Rata-rata tagihan perokok = 35.116 USD
Rata-rata tagihan non perokok = 8.629 USD

Analysis

Berdasarkan data dari 1.338 pengguna asuransi, rata-rata umur pengguna yang merokok sama dengan rata-rata umur pengguna yang tidak merokok, yaitu sekitar 38 tahun. Rata-rata nilai BMI dari seluruh pengguna asuransi, baik laki-laki maupun perempuan, perokok maupun non perokok, menunjukkan status obesitas yaitu lebih dari 30 kg/m2.

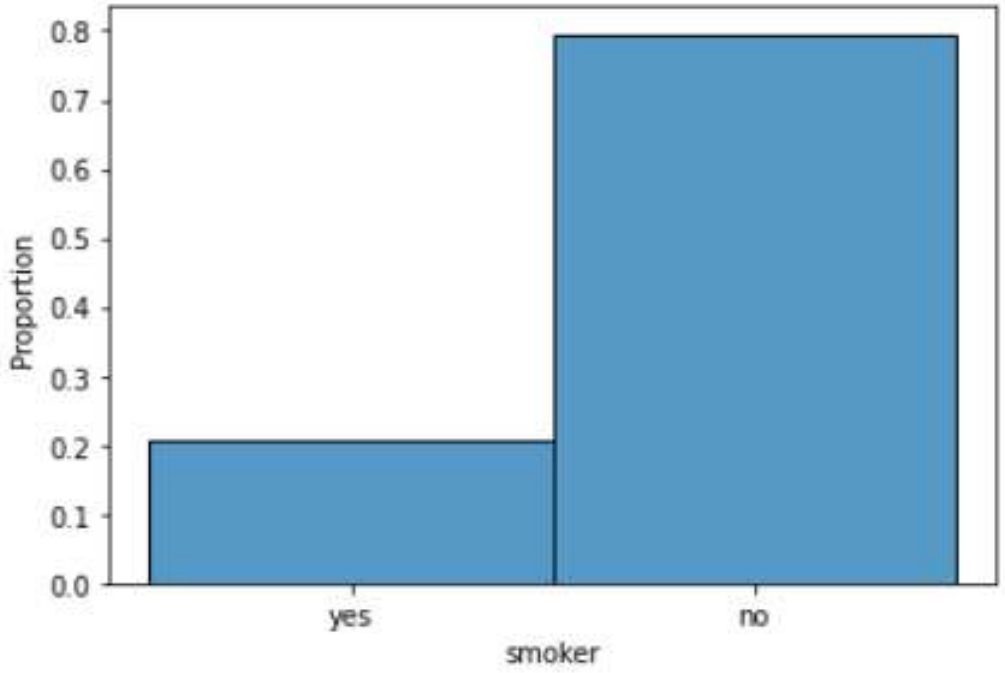
Rata-rata dan sebaran tagihan kesehatan perokok lebih tinggi daripada non perokok. Begitu juga pada pengguna dengan BMI > 25, rata-rata dan sebaran tagihan kesehatan perokok lebih tinggi daripada non perokok. Rata-rata tagihan perokok sebesar 32.050 USD dengan simpangan baku 11.541 USD. Sementara rata-rata tagihan non perokok hanya 8.434 USD dengan simpangan baku 5.993 USD.

Namun, pada boxplot terlihat nilai BMI dan tagihan kesehatan non perokok memiliki banyak data pencilan (*outlier*).

Categorical Variables Analysis

Proporsion of smokers and non smokers

4. Mana yang lebih tinggi proporsi perokok atau non perokok?

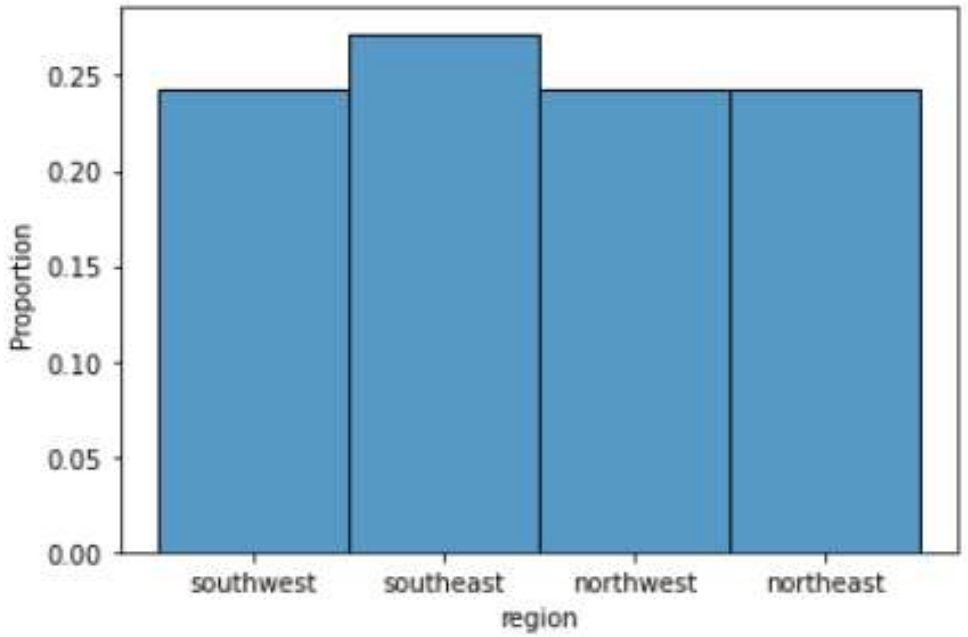


Kesimpulan:

Proporsi non perokok lebih tinggi daripada perokok.
Sekitar 80% dari data pengguna asuransi adalah non perokok.
Sisanya, 20% adalah perokok.

Proporsion of users in each region

3. Apakah setiap region memiliki proporsi data banyak orang yang sama?



Kesimpulan:

Proporsi pengguna asuransi di tiap region tidak sama. Pengguna asuransi di wilayah *South East* lebih banyak dibandingkan wilayah lainnya.

Proporsion of smokers by gender

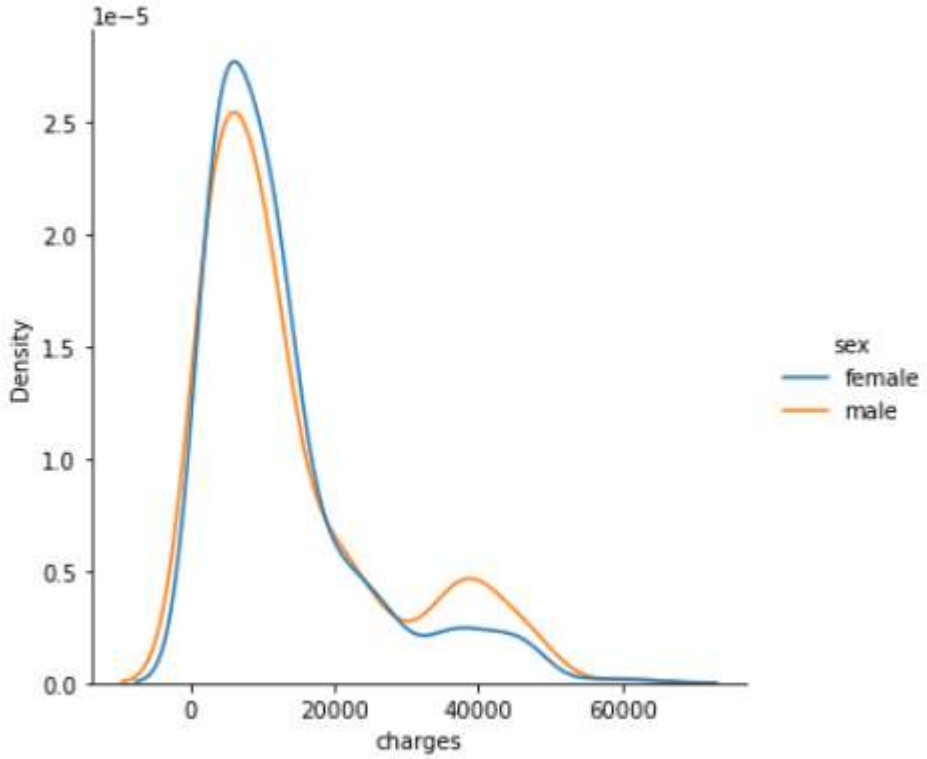
5. Berapa peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok?
Peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok adalah 0,42.

6. Berapa peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok?
Peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok adalah 0,58.

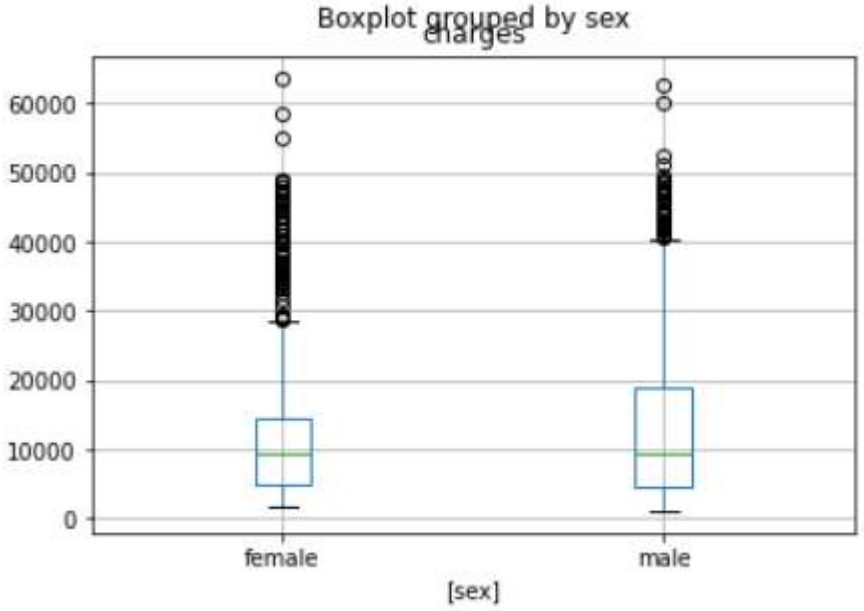
smoker	no	yes
female	547	115
male	517	159

Charges by gender

1. Gender mana yang memiliki tagihan paling tinggi? **Laki-laki.**

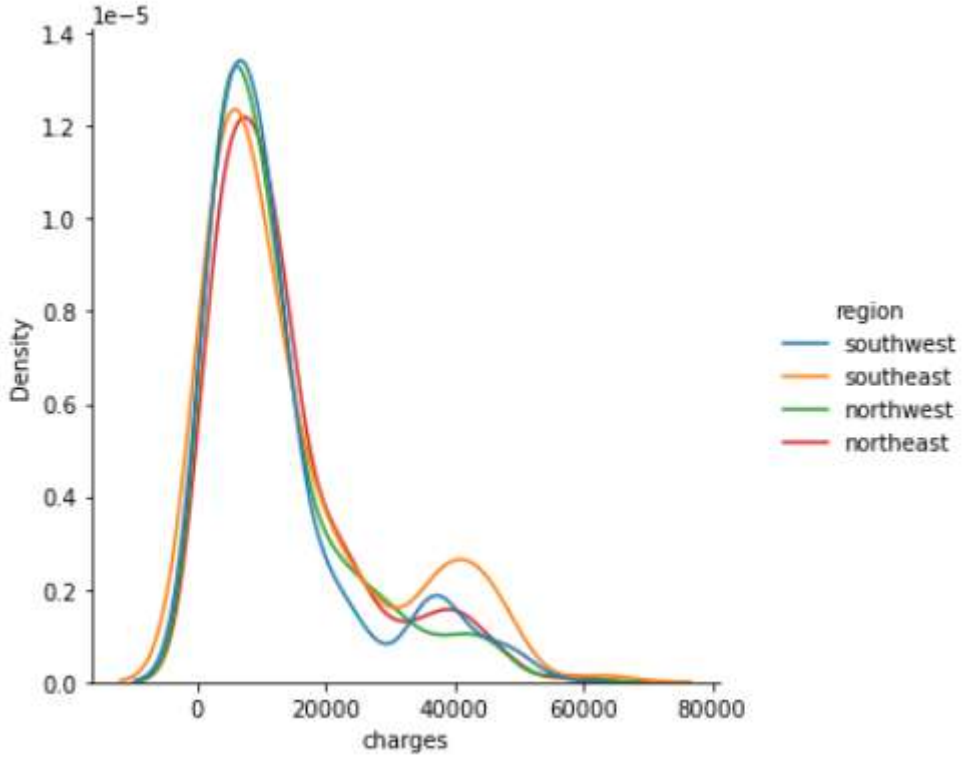


	charges	
	mean	median
sex		
female	12569.578844	9412.96250
male	13956.751178	9369.61575



Distribution in each region

2. Distribusi peluang tagihan di tiap-tiap region.



Analysis

Dalam sample ini, 79,52% pengguna asuransi adalan non perokok, dan sisanya 20,48% adalah perokok. Sementara 58% dari perokok adalah laki-laki, dan sisanya 42% dari perokok adalah perempuan.

Rata-rata tagihan kesehatan pengguna laki-laki sedikit lebih tinggi daripada pengguna perempuan. Didukung dengan sebaran data tagihan yang lebih tinggi pada pengguna laki-laki, berdasarkan *sample range* dan *interquartile range* pada boxplot.

Namun perlu diperhatikan juga pada boxplot tagihan berdasarkan jenis kelamin, baik laki-laki maupun perempuan memiliki data pencilan (*outlier*). Hal ini juga terlihat pada plot distribusi peluang tagihan, yang mana terdapat 2 puncak dengan kemiringan ke kiri. Jika dilihat dari nilai tengah/median, tagihan kesehatan perempuan justru sedikit lebih tinggi dibandingkan tagihan kesehatan laki-laki.
Pengujian hipotesis di bagian akhir nanti dapat membuktikan bahwa tagihan kesehatan laki-laki lebih tinggi daripada perempuan.

Continuous Variables Analysis

Probability of someone has high charges given he's a smoker

3. Berapa peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia adalah perokok?

Peluang seseorang acak tagihan kesehatannya diatas 16.700 USD diketahui dia adalah perokok sebesar 0,93.

smoker	no	yes
charges_category		
normal	984	20
high	80	254

Probability of someone has high charges by BMI

4. Mana yang lebih mungkin terjadi
   a. Seseorang dengan BMI di atas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau
   b. Seseorang dengan BMI di bawah 25 mendapatkan tagihan kesehatan diatas 16.7k

a. Seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.700 USD lebih mungkin terjadi dengan peluang 0,21.
Peluang kejadian b hanya 0,04.

bmi_category	normal	over
charges_category		
normal	196	808
high	51	283

Probability of someone has high charges by smoker status

5. Mana yang lebih mungkin terjadi
   a. Seseorang perokok dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.7k, atau
   b. Seseorang non perokok dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.7k

a. Seseorang perokok dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.700 USD lebih mungkin terjadi dengan peluang 0,78.
Peluang kejadian b hanya 0,06.

a.

bmi_category	normal	over
charges_category		
normal	16	4
high	39	215

b.

bmi_category	normal	over
charges_category		
normal	180	804
high	12	68

Probability of smoker has high charges

2. Mencari kemungkinan terjadi, seorang perokok dengan BMI di atas 25 akan mendapatkan tagihan kesehatan di atas 16.7k.

Peluang seseorang perokok dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.700 USD adalah 0,78.

bmi_category	normal	over
charges_category		
normal	16	4
high	39	215

Probability of charges by BMI

1. Mencari peluang seseorang dengan BMI di atas 25 mendapatkan tagihan kesehatan di bawah 16.7k.

Peluang seseorang dengan BMI di atas 25 mendapatkan tagihan kesehatan di bawah 16.700 USD adalah 0,60.

bmi_category	normal	over
charges_category		
normal	196	808
high	51	283

Analysis

- Sebanyak 93% perokok mendapatkan tagihan kesehatan di atas 16.700 USD.
- Pada sample pengguna asuransi dengan BMI di atas 25 pun, 78% perokok mendapatkan tagihan kesehatan di atas 16.700 USD. Sedangkan hanya 6% dari non perokok yang mendapatkan tagihan kesehatan di atas 16.700 USD meskipun BMI mereka di atas 25.
- Kesimpulan: **pengguna asuransi yang merokok memiliki peluang lebih tinggi untuk mendapatkan tagihan kesehatan di atas 16.700 USD** dibandingkan pengguna asuransi yang tidak merokok.

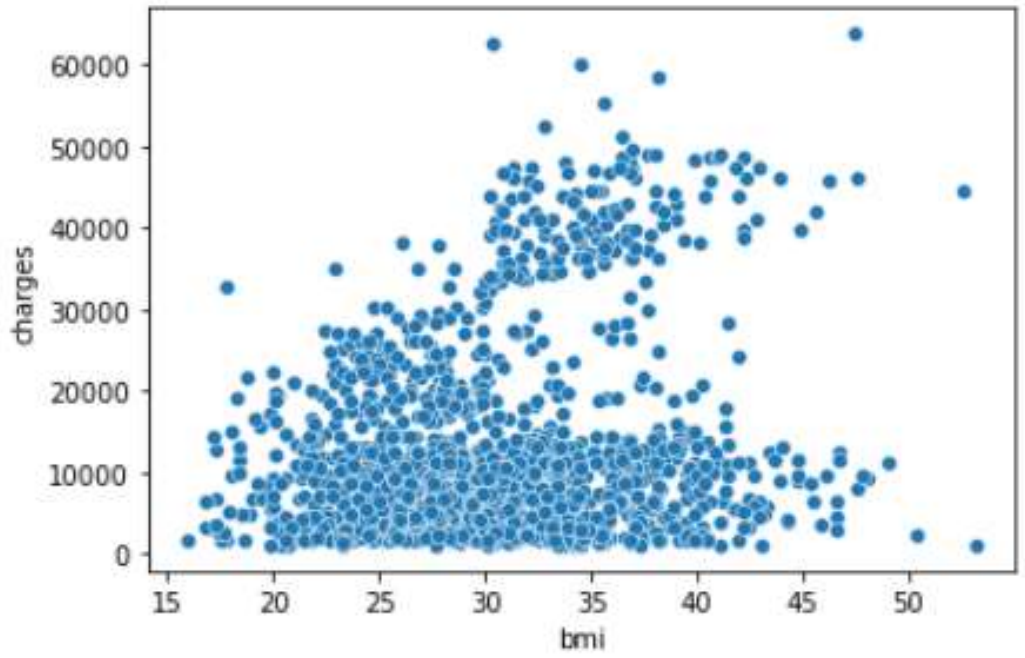
- Peluang seorang pengguna asuransi dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.700 USD relatif rendah, yaitu 0,21. Dan peluangnya mendapatkan tagihan kesehatan di bawah 16.700 USD pun lebih tinggi, yaitu 0,60. Untuk mendapatkan kesimpulan mengenai hubungan BMI dengan tagihan kesehatan, akan dilakukan analisa korelasi dan pengujian hipotesis pada bagian selanjutnya.

Variables Correlation

Correlation

Membuat scatter plot untuk melihat korelasi antara BMI dengan tagihan kesehatan.

a. Pada semua pengguna asuransi dalam sample.



	bmi	charges
bmi	1.000000	0.198341
charges	0.198341	1.000000

Correlation

b. Menguji independensi antara BMI dengan tagihan kesehatan, berdasarkan tabel kontingensi berikut.

bmi_category	normal	over
charges_category		
normal	180	804
high	12	68

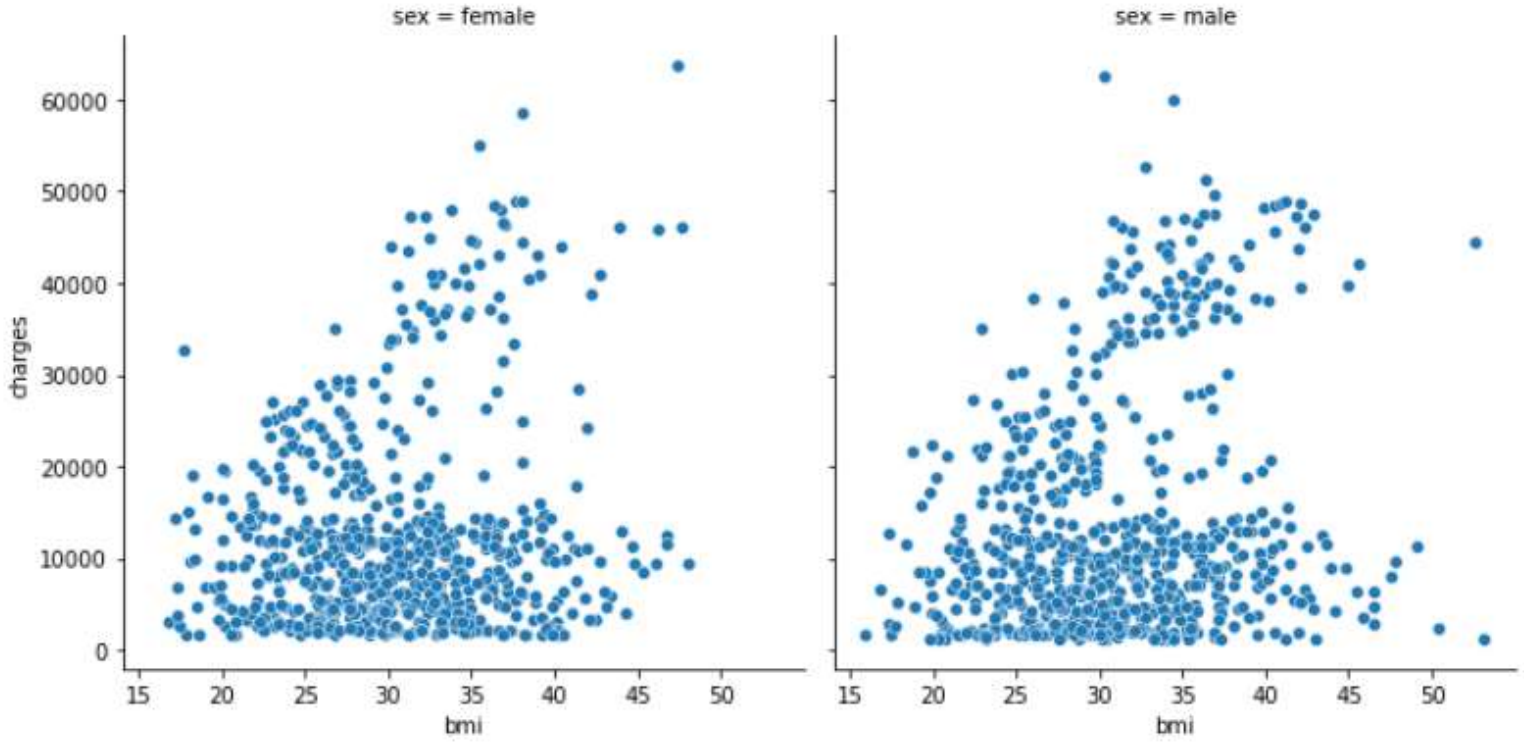
Menggunakan Chi-squared Test of Independence, diperoleh p-value = 0,098. Dengan alpha = 0,05, maka $\mathbf{H_0}$ bahwa BMI dan tagihan kesehatan adalah independen **gagal ditolak**.

Artinya, sample pengguna asuransi ini **belum cukup membuktikan adanya hubungan antara BMI dengan tagihan kesehatan penggunanya**.

Correlation

Membuat scatter plot untuk melihat korelasi antara BMI dengan tagihan kesehatan.

c. Pada pengguna asuransi perempuan dan laki-laki.

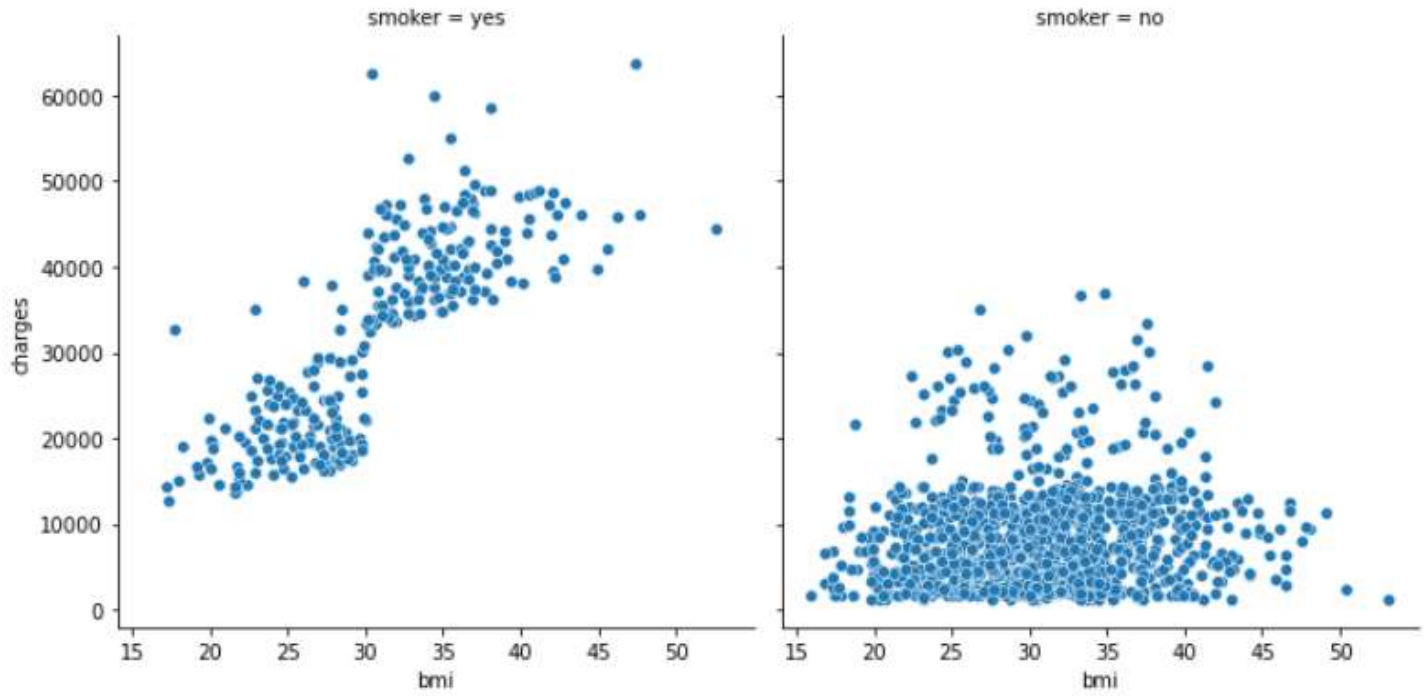


Correlation

Membuat scatter plot untuk melihat korelasi antara BMI dengan tagihan kesehatan.

d. Pada pengguna asuransi yang merokok dan yang tidak merokok.

	bmi	charges
bmi	1.000000	0.806481
charges	0.806481	1.000000



Correlation

e. Menguji independensi antara status merokok dengan tagihan kesehatan, berdasarkan tabel kontingensi berikut.

smoker	no	yes
charges_category		
normal	984	20
high	80	254

Menggunakan Chi-squared Test of Independence, diperoleh p-value = **$1{,}39 \times 10^{-184}$**. Dengan alpha = 0,05, maka $\mathbf{H_0}$ bahwa status merokok dan tagihan kesehatan adalah independen **ditolak**.

Secara sederhana, **terdapat hubungan antara status merokok dengan tagihan kesehatan**.

Analysis

- Variabel BMI memiliki korelasi yang relatif rendah terhadap variabel tagihan kesehatan pada seluruh pengguna asuransi dalam sample ini. Dengan nilai koefisien korelasi Pearson 0,198 dan hasil uji independensi Chi-squared yang gagal membuktikan adanya hubungan antara BMI dengan tagihan kesehatan.

- **Namun, pada kelompok perokok, terlihat adanya korelasi positif antara BMI dengan tagihan kesehatan** pada scatter plot.
- Dibuktikan dengan koefisien korelasi Pearson yang mendekati +1, yaitu sebesar 0,806. Sebagai tambahan, hasil uji independensi Chi-squared juga membuktikan adanya hubungan antara status merokok dengan tagihan kesehatan.

- Sampai pada analisa ini, kita dapat membangun hipotesis bahwa pengguna asuransi yang merokok cenderung menghasilkan tagihan kesehatan yang lebih besar daripada pengguna yang tidak merokok.

Hypothesis Testing

Smoker's charges are higher than non smoker's

Hipotesis

$H_0 : \mu_1 = \mu_2$ atau $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 > \mu_2$ $H_1 : \mu_1 - \mu_2 > 0$

dengan μ_1 adalah rata-rata tagihan kesehatan perokok, dan μ_2 adalah rata-rata tagihan kesehatan non perokok.

Uji Z untuk perbedaan rata-rata tagihan kesehatan dua populasi (perokok dan non perokok) menghasilkan **Z = 46,66 dan p-value = 0,0.**
Dengan alpha = 0,05, dan p-value < alpha, maka **H_0 ditolak**.

Kesimpulan: **rata-rata tagihan kesehatan perokok lebih tinggi daripada rata-rata tagihan kesehatan non perokok.**

Someone with BMI > 25 has higher charges

Hipotesis

$H_0 : \mu_1 = \mu_2$ atau $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 > \mu_2$ $H_1 : \mu_1 - \mu_2 > 0$

dengan μ_1 adalah rata-rata tagihan kesehatan pengguna asuransi dengan BMI di atas 25, dan μ_2 adalah rata-rata tagihan kesehatan pengguna asuransi dengan BMI di bawah 25.

Uji Z untuk perbedaan rata-rata tagihan kesehatan dua populasi (BMI > 25 dan BMI < 25) menghasilkan **Z = 4,32 dan p-value = $7{,}8 \times 10^{-6}$.**
Dengan alpha = 0,05, dan p-value < alpha, maka **H_0 ditolak.**

Kesimpulan: **rata-rata tagihan kesehatan pengguna asuransi dengan BMI di atas 25 lebih tinggi daripada rata-rata tagihan kesehatan pengguna asuransi dengan BMI di bawah 25.**

Male's BMI is not different from female's

Hipotesis

$H_0 : \mu_1 = \mu_2$ atau $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$ $H_1 : \mu_1 - \mu_2 \neq 0$

dengan μ_1 adalah rata-rata BMI dari pengguna laki-laki, dan μ_2 adalah rata-rata BMI dari pengguna perempuan.

Uji Z untuk perbedaan rata-rata BMI dua populasi (laki-laki dan perempuan) menghasilkan **Z = 1,69 dan p-value = 0,089**.
Dengan alpha = 0,05, dan p-value > alpha, maka **H_0 gagal ditolak**.

Kesimpulan: **statistik dari sample pengguna asuransi ini belum cukup membuktikan adanya perbedaan rata-rata BMI pada pengguna laki-laki dan perempuan. Secara sederhana, rata-rata BMI laki-laki dan perempuan sama.**

Male's charges is higher than female's

Hipotesis

$H_0 : \mu_1 = \mu_2$ atau $H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 > \mu_2$ $H_1 : \mu_1 - \mu_2 > 0$

dengan μ_1 adalah rata-rata tagihan kesehatan dari pengguna laki-laki, dan μ_2 adalah rata-rata tagihan kesehatan dari pengguna perempuan.

Uji Z untuk perbedaan rata-rata tagihan kesehatan dua populasi (laki-laki dan perempuan) menghasilkan **Z = 2,097 dan p-value = 0,018**.
Dengan alpha = 0,05, dan p-value < alpha, maka **H_0 ditolak**.

Kesimpulan: **rata-rata tagihan kesehatan pengguna laki-laki lebih tinggi daripada rata-rata tagihan kesehatan pengguna perempuan**.

Conclusion

Conclusion

1. Status merokok dari pengguna asuransi berhubungan dengan tagihan kesehatan yang akan dihasilkan. Uji hipotesis juga membuktikan bahwa rata-rata tagihan kesehatan perokok lebih tinggi daripada rata-rata tagihan kesehatan non perokok.

2. Meskipun BMI dan tagihan kesehatan seorang pengguna asuransi berkorelasi relatif rendah dalam sample ini, namun uji hipotesis membuktikan bahwa rata-rata tagihan kesehatan pengguna asuransi dengan BMI di atas 25 lebih tinggi daripada rata-rata tagihan kesehatan pengguna asuransi dengan BMI di bawah 25.

3. BMI pengguna laki-laki dan perempuan tidak berbeda berdasarkan uji hipotesis, namun rata-rata tagihan kesehatan pengguna laki-laki terbukti lebih tinggi daripada pengguna perempuan dalam sample ini.

Kesimpulan:
Perusahaan sebaiknya menyesuaikan premi asuransi berdasarkan status merokok pengguna, kategori BMI pengguna, dan jenis kelamin pengguna. Namun diperlukan analisa dan pengujian lebih lanjut untuk membuktikan pengaruh masing-masing atau ketiga variabel tersebut terhadap besarnya tagihan kesehatan, dengan mempertimbangkan adanya *confounding variable* maupun *lurking variable* dalam analisa korelasi.

Notes

- Diperlukan analisa dan pengujian lebih lanjut untuk membuktikan kesimpulan dari hasil analisa pengguna asuransi ini, dengan mempertimbangkan adanya *confounding variable* yang mempengaruhi korelasi terhadap tagihan kesehatan atau adanya *lurking variable* yang berkorelasi terhadap tagihan kesehatan namun belum termasuk dalam analisa.

- Rekomendasi untuk melanjutkan pemodelan, misalnya dengan regresi, termasuk melakukan estimasi dan uji signifikansi parameter dalam model regresi. Dengan model terbaik, perusahaan dapat menyesuaikan atau memprediksi premi setiap pengguna asuransi berdasarkan variabel-variabel yang berpengaruh terhadap besarnya tagihan kesehatan.

- Terdapat data outlier pada variabel BMI maupun tagihan kesehatan. Analyst perlu berdiskusi dengan user untuk mengonfirmasi perlakuan terhadap outlier. Dalam analisa ini, outlier tetap termasuk di dalam sample.

Reference

- Bhattacharyya, G. K. & Johnson, R. A. (2019). Statistics Principles and Methods.