Pacmann

Probability Insurance Charges

Probability Course - Sekolah Data Pacmann

Outline

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

Introduction

Introduction

Asuransi kesehatan adalah salah satu hal yang patut diperhatikan karena bersangkutan dengan kebutuhan perencanaan masa depan. Pengguna asuransi kesehatan diwajibkan untuk membayar besaran uang secara rutin (premi) kepada pihak perusahaan asuransi. Premi tersebut diolah oleh perusahaan asuransi untuk membayarkan tagihan kesehatan pengguna yang tertanggung. Penentuan nilai premi menjadi tantangan tersendiri bagi pihak asuransi mengingat ada banyak faktor yang dapat mempengaruhi & meningkatkan profil resiko pengguna.

Melalui project ini, saya mencoba untuk menganalisa bagaimana hubungan variable-variable yang terdapat pada dataset untuk diambil kesimpulan yang dapat membantu keputusan bisnis.

Dataset

Dataset

Dataset yang disediakan adalah data tagihan kesehatan personal. Data ini memiliki 7 variable dengan variable charges menunjukkan besaran tagihan kesehatan. Deskripsi setiap kolom dari dataset adalah sebagai berikut:

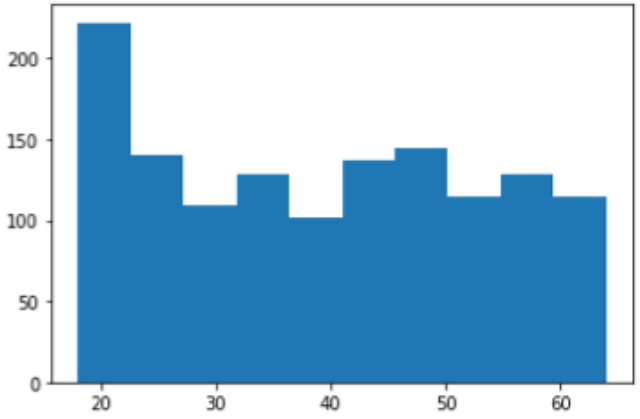
- **Age**
  Age of primary beneficiary
- **Sex**
  Insurance contractor gender, female, male
- **BMI**
  Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m2) using the ratio of height to weight, ideally 18.5 to 24.9
- **Children**
  Number of children covered by health insurance / Number of dependents
- **Smoker**
  Smoking
- **Region**
  The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **Charges**
  Individual medical costs billed by health insurance

Pacmann

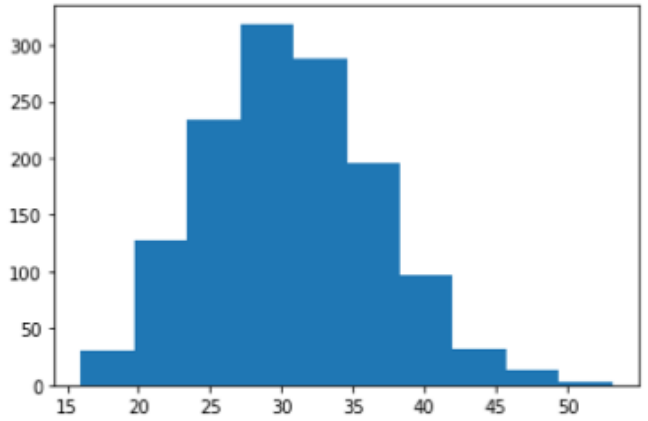
Descriptive Statistics Analysis

Rata-rata setiap variable numerical dan distribusinya

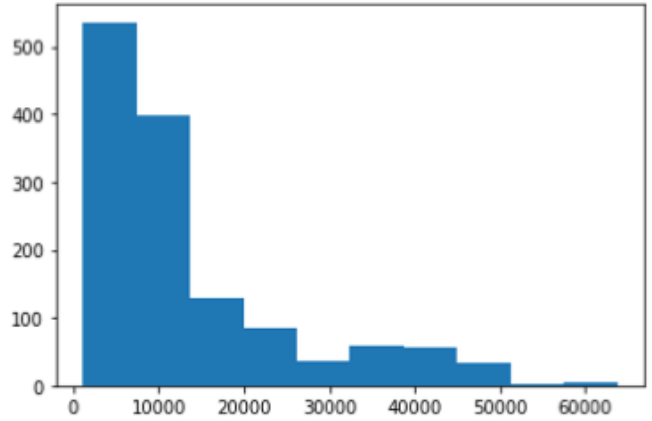
Variable	Rata-rata
Umur	39 tahun
BMI	31 kg/m^2
Tagihan	US$ 13.270



Umur



BMI



Tagihan

- Dari penjabaran di atas, ditemukan bahwa rata-rata umur nasabah adalah 39 tahun. Sedangkan rata-rata BMI-nya adalah 31 kg/m2, yang menandakan bahwa rata-rata nasabah memiliki BMI dengan kategori Obese Class I. Untuk rata-rata tagihannya sendiri berada pada US$ 13.270.

Kondisi BMI pada masing-masing genre

Genre	Jumlah	BMI terendah	Rata-rata BMI	BMI tertinggi
Laki-laki	676	15,96 kg/m^2	30,94 kg/m^2	53,13 kg/m^2
Perempuan	662	16,82 kg/m^2	30,38 kg/m^2	48,07 kg/m^2

- Dari penjabaran di atas, ditemukan bahwa perbandingan data BMI antara laki-laki dan perempuan tidak terlalu jauh. Yang patut disorot adalah rata-rata BMI yang masuk ke dalam kategori Obese Class I. Dan bahkan, BMI yang tertinggi masuk dalam kategori Obese Class III.

Tagihan kepada perokok dan non perokok

Perokok	Jumlah	Varians Tagihan	Standar Deviasi	Rata-rata Tagihan	Median
Iya	274	132.721.153	11.520	US$ 32.050	US$ 34.456
Tidak	1.064	35.891.656	5.991	US$ 8.434	US$ 7.345

- Dari penjabaran di atas, ditemukan bahwa rata-rata tagihan premi nasabah yang merokok jauh lebih besar dibandingkan dengan nasabah yang tidak merokok. Hal ini dapat dilihat dari rata-rata tagihan premi nasabah perokok yang berjumlah US$ 32.050 dengan median US$ 34.456. Sangat jauh bila dibandingkan dengan rata-rata tagihan secara keseluruhan.

- Sedangkan rata-rata tagihan nasabah yang tidak merokok dengan rata-rata tagihan premi sebesar US$ 8.434 dan median US$ 7.345 berada di bawah rata-rata secara keseluruhan.

- Selisih pada persebaran tagihan nasabah merokok pun juga lebih lebar dibandingkan dengan nasabah yang tidak merokok, dengan perbandingan 11.520 dan 5.991.

Rata-rata tagihan nasabah perokok dengan BMI di atas 25

Perokok	BMI	Rata-rata Tagihan
Iya	> 25 kg/m^2	US$ 35.117
Tidak	> 25 kg/m^2	US$ 8.630

- Melihat penjabaran di atas, dapat terlihat bahwa rata-rata tagihan mengalami peningkatan dengan nasabah yang memiliki kondisi BMI di atas 25 kg/m^2, yaitu BMI yang sudah masuk dalam kategori overweight.

Analysis

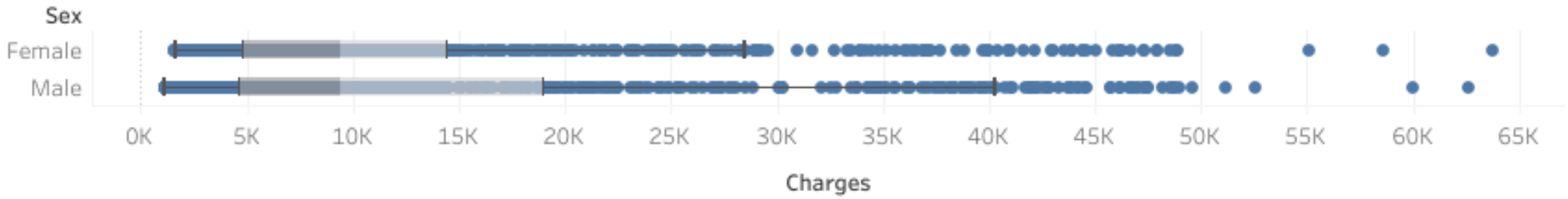
- Dataset yang dipakai, diambil dari 1.338 nasabah asuransi yang rata-rata berumur 39 tahun. Rata-rata nilai BMI mereka adalah 31 kg/m^2, yang mana termasuk ke dalam kategori obesitas. Rata-rata tagihannya yaitu US$ 13.270.

- Berdasarkan penjabaran pada slide sebelumnya, dapat diketahui bahwa terdapat perbedaan yang besar pada rata-rata tagihan antara nasabah yang perokok dan tidak merokok. Rata-rata tagihan nasabah perokok sebesar US$ 32.050, sedangkan nasabah yang tidak merokok memiliki rata-rata sebesar US$ 8.434.

- Rata-rata tagihan masih bisa meningkat lagi, apabila kita kondisikan nasabah tersebut memiliki kondisi nilai BMI di atas 25 kg/m^2, yang mana batas bawah dari kategori BMI overweight.

Categorical Variables Analysis

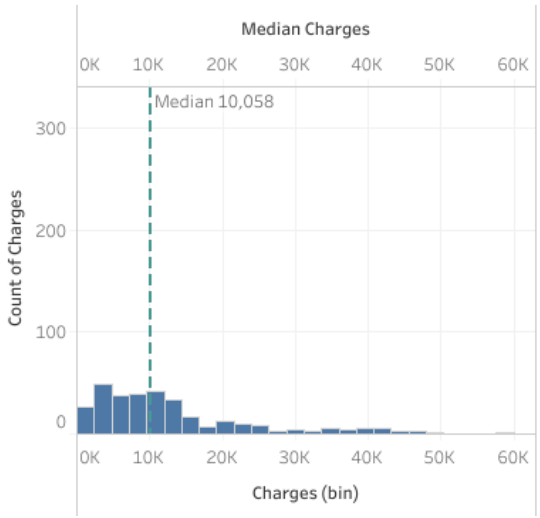
Tagihan tertinggi dimiliki oleh seorang perempuan



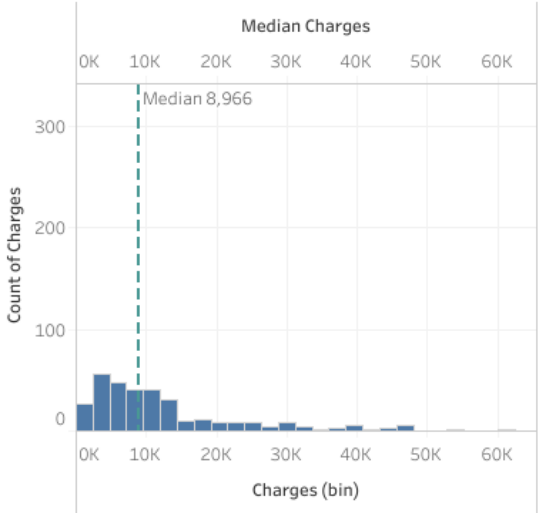
Tagihan tertinggi perempuan: US$ 63.770
Tagihan tertinggi laki-laki: US$ 62.593

- Yang menarik dari penjabaran di atas adalah bahwa tagihan tertinggi kedua gender merupakan outliers dari persebaran data tagihan premi para nasabah.

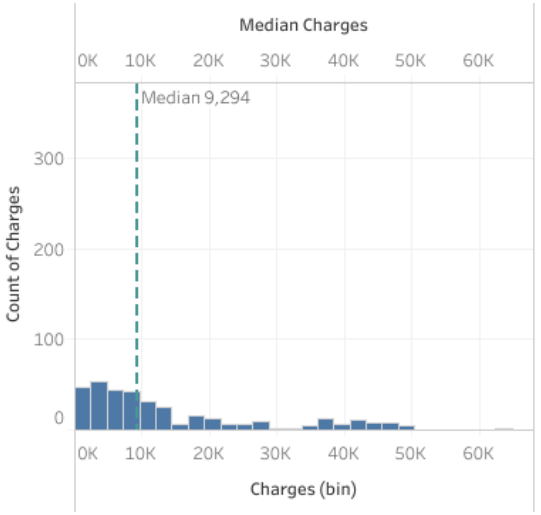
Distribusi tagihan pada masing-masing region



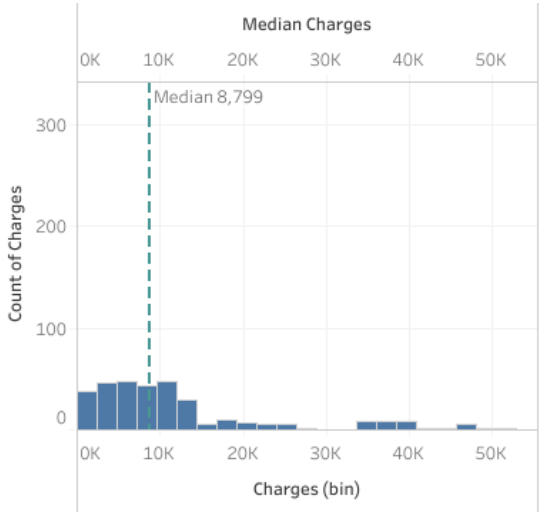
North East
Probability: 0.24



North West
Probability: 0.23



South East
Probability: 0.30

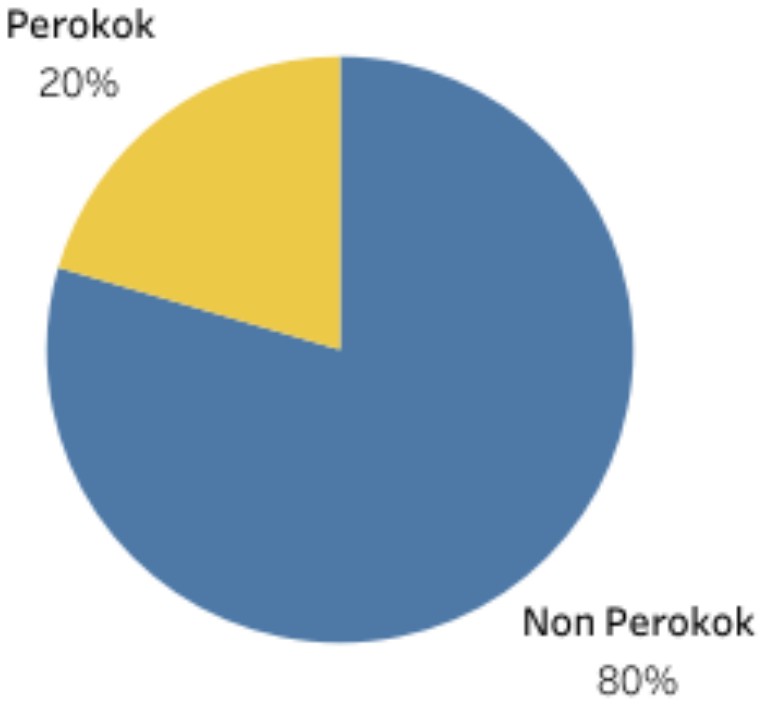


South West
Probability: 0.23

- Sejalan dengan distribusi tagihan secara keseluruhan, distribusi tagihan berdasarkan region masing-masing berbentuk eksponensial dengan skewed positive.
- Diketahui juga, bahwa nilai peluang tagihan dari region South East merupakan yang terbesar di antar region lainnya.

20% dari nasabah adalah seorang perokok

- Dari visualisasi di samping, dapat diketahui bahwa proporsi nasabah yang merokok adalah sebanyak 20% dari total nasabah.



Peluang mengetahui gender jika dia adalah seorang perokok

Gender Perokok	Jumlah	Probability
Perempuan	115	0,42
Laki-laki	159	0,58

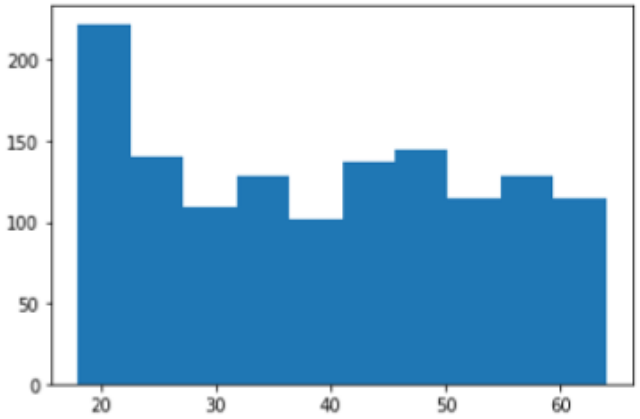
- Dari penjabaran di atas, dapat diketahui bahwa seorang perokok memiliki peluang lebih besar untuk dikenali sebagai laki-laki dibanding perempuan.

Analysis

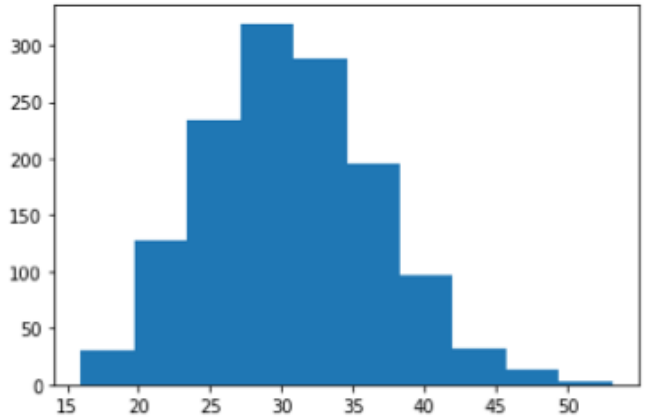
- Berdasarkan gender, persebaran tagihan premi cukup merata. Walaupun disertai beberapa outliers yang perlu menjadi catatan dalam mengambil keputusan.
- Persebaran lokasi nasabah juga cukup merata dengan region South East tampak lebih menonjol peluangnya dibanding region lainnya.
- Proporsi perokok dan non perokok tampak berbanding jauh yang menandakan bahwa nasabah dalam dataset ini mayoritas adalah non perokok.
- Dengan jumlahnya yang lebih besar dibandingkan perempuan, seorang perokok lebih berpeluang sebagai laki-laki dibandingkan perempuan.

Continuous Variables Analysis

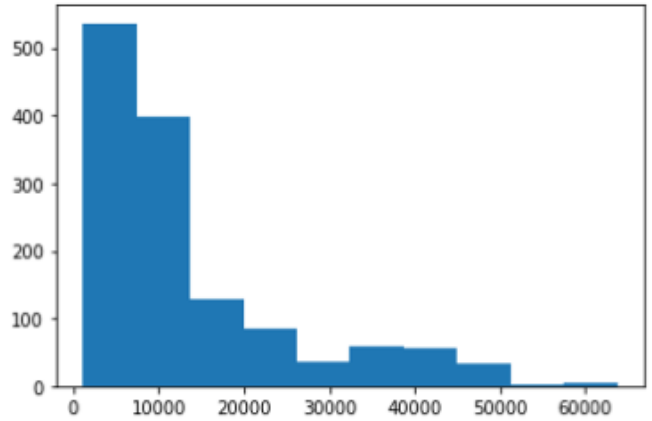
Review distribusi data variable numerical



Umur



BMI



Tagihan

- Visualisasi distribusi data ini akan membantu dalam mencari besaran peluang pada masing-masing variable disertai kondisinya.

Peluang tagihan lebih dari US$ 16.700 dengan kondisi BMI

Kondisi BMI	Probability
> 25 kg/m^2	0,23
< 25 kg/m^2	0,05

- Dari penjabaran di atas, dapat diketahui bahwa seseorang dengan kondisi BMI di atas 25 kg/m^2 kemungkinannya lebih besar untuk memiliki tagihan di atas US$ 16.700.

Peluang tagihan lebih dari US$ 16.700 dengan kondisi perokok

Perokok	Probability
Perokok	0,06
Non Perokok	0,23

- Dari penjabaran di atas, dapat diketahui bahwa seorang nasabah dengan kondisi perokok kemungkinannya jauh lebih kecil untuk memiliki tagihan di atas US$ 16.700 dibandingkan jika seorang nasabah tersebut diketahui merupakan non perokok.
- Hal ini dapat dipahami, karena proporsi perokok jauh lebih sedikit dibandingkan dengan nasabah yang tidak merokok.

Memperdalam exploratory data analysis

Bagaimana jika situasinya seseorang tersebut perokok atau non perokok dengan kondisi BMI di atas 25 kg/m^2, berapakah peluang tagihannya di atas US$ 16.700?

Perokok	BMI	Probability
Perokok	> 25 kg/m^2	0,05
Non Perokok	> 25 kg/m^2	0,19

- Dari penjabaran di atas, dapat diketahui bahwa peluang tagihan di atas US$ 16.700 mengecil jika diketahui nasabah tersebut memiliki kondisi tambahan, yaitu BMI di atas 25 kg/m^2.

Memperdalam exploratory data analysis

Bagaimana jika situasinya seseorang tersebut perokok atau non perokok dengan kondisi BMI di atas 25 kg/m^2 maupun di bawah 25 kg/m^2. Berapakah peluang tagihannya di atas US$ 16.700?

Perokok	BMI	Probability
Perokok	> 25 kg/m^2	0,05
Non Perokok	> 25 kg/m^2	0,19
Perokok	< 25 kg/m^2	0,01
Non Perokok	< 25 kg/m^2	0,04

- Dari penjabaran di atas, dapat diketahui bahwa peluang tagihan di atas US$ 16.700 mengecil jika diketahui nasabah tersebut memiliki kondisi tambahan, yaitu BMI di atas 25 kg/m^2.
- Jika dibandingkan lagi dengan kondisi nasabah dengan BMI di bawah 25 kg/m^2, nasabah yang tidak merokok dan memiliki BMI di atas 25 kg/m^2 tetap memiliki peluang terbesar untuk memiliki tagihan premi di atas US$ 16.700.

Analysis

- Berdasarkan eksplorasi data, dapat diketahui bahwa nasabah dengan kondisi BMI di atas 25 kg/m2 memiliki peluang terbesar untuk mendapatkan tagihan di atas US$ 16.700.
- Jika diperhatikan dari distribusinya yang normal dengan rata-rata 31 kg/m2, dapat dipahami jika nasabah dengan kondisi BMI di atas 25 kg/m lebih banyak dibandingkan yang di bawahnya. Sehingga, hal tersebut membuat peluangnya lebih besar.
- Hal yang sama juga ditemui pada nasabah dengan kondisi non perokok. Nasabah dengan kondisi non perokok memiliki peluang terbesar untuk mendapatkan tagihan di atas US$ 16.700.
- Dapat dipahami juga bahwa jumlah nasabah non perokok jauh lebih besar dibandingkan nasabah perokok. Sehingga, peluangnya lebih besar untuk memiliki tagihan di atas US$ 16.700.
- Jika kondisi non perokok dan memiliki BMI di atas 25 kg/m^2 digabung, peluangnya juga tetap menjadi yang terbesar untuk memiliki tagihan di atas US$ 16.700, jika dibandingkan dengan kondisi lainnya.

Pacmann

Variables Correlation

Korelasi variable tagihan dengan variable lainnya

Variabel 1	Variable 2	Korelasi
Tagihan	Usia	0,3
	BMI	0,2
	Jumlah Anak	0,07

- Dari penjabaran di atas, dapat diketahui bahwa korelasi antara tagihan dan usia merupakan korelasi yang terkuat di antara korelasi dengan variable lainnya.
- Walaupun begitu, nilai korelasinya tidak bisa dibilang kuat atau cenderung lemah.

Hypothesis Testing

Tagihan perokok lebih tinggi dari non perokok

Hasil Z-test	P-value (smaller)
46,66	1,0

- Dari hasil pengujian hipotes di atas, dapat diketahui bahwa p-value lebih dari nilai alpha.
- Karena, p-value lebih besar dari nilai alpha 0,05, sehingga tidak memiliki bukti yang cukup untuk menolak null hypothesis.
- Dengan kata lain, hipotesis bahwa tagihan perokok lebih tinggi dari non perokok **dapat diterima**.

Tagihan nasabah berdasarkan kondisi BMI

Bagaimana pengujian hipotesis bahwa tagihan nasabah dengan BMI di atas 25 lebih tinggi dibandingkan dengan tagihan nasabah dengan kondisi BMI di bawah 25.

Hasil Z-test	P-value (smaller)
4.30	0,99

- Dari hasil pengujian hipotes di atas, dapat diketahui bahwa p-value lebih dari nilai alpha.
- Karena, p-value lebih besar dari nilai alpha 0,05, sehingga tidak memiliki bukti yang cukup untuk menolak null hypothesis.
- Dengan kata lain, hipotesis bahwa tagihan nasabah dengan kondisi BMI di atas 25 lebih tinggi dibanding tagihan nasabah dengan kondisi BMI di bawah 25 **dapat diterima**.

Tagihan nasabah laki-laki dan perempuan

Bagaimana pengujian hipotesis bahwa tagihan laki-laki lebih besar dibandingkan dengan tagihan perempuan.

Hasil Z-test	P-value (smaller)
2,1	0,98

- Dari hasil pengujian hipotes di atas, dapat diketahui bahwa p-value lebih dari nilai alpha.
- Karena, p-value lebih besar dari nilai alpha 0,05, sehingga tidak memiliki bukti yang cukup untuk menolak null hypothesis.
- Dengan kata lain, hipotesis bahwa tagihan nasabah laki-laki lebih tinggi dibanding tagihan nasabah perempuan **dapat diterima**.

Conclusion

Conclusion

- Proporsi perokok sebagai faktor yang sangat mempengaruhi tagihan kesehatan cukup kecil.
- Usia menjadi faktor yang terkuat dalam mempengaruhi besar tagihan kesehatan.
- Rata-rata nasabah termasuk ke dalam obesitas, di mana hal tersebut ikut meningkatkan risiko kesehatan dari nasabah. Ke depannya, penjaringan calon nasabah melalui kondisi BMI-nya dapat diperbaiki lebih baik lagi.

Reference

- File code: https://colab.research.google.com/drive/1bYt5HBRsLvzNrsFpQB25rsr1Co8eED0R?authuser=1#scrollTo=7OdrKsLItEhY

Pacmann

Thank you!