Pacmann

Analisa Tagihan Asuransi

Probability Course - Sekolah Data Pacmann

Outline

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

Introduction

Introduction

Polis asuransi memiliki banyak pengguna dari berbagai kalangan, tidak terbatas tua-muda, lelaki-perempuan maupun indikator kesehatan lainnya. Masing-masing kelompok pengguna juga memiliki jumlah klaim (tagihan) yang bervariasi.

Namun adakah implikasi tertentu dari berbagai kelompok pengguna tersebut terhadap jumlah klaim? Project ini bertujuan untuk menjawab pertanyaan tersebut dan memetakan korelasi dari berbagai kelompok yang ada.

Pihak asuransi dapat menggunakan hasil analisis dari project ini untuk memperkirakan kemungkinan kelompok dengan charge yang lebih besar sehingga bisa meng-adjust proporsi pengguna asuransi.

Dataset

Dataset

Data yang digunakan adalah 1338 rincian pengguna asuransi dengan kolom sebagai berikut:

- Age = Umur pengguna asuransi
- BMI = Body mass index, berat dalam kg/(tinggi dalam m)^2
- Sex = Jenis kelamin pengguna asuransi
- Children = Jumlah anak dari pengguna asuransi
- Smoker = Status merokok dari pengguna asuransi
- Region = Region dimana pengguna asuransi tinggal
- Charges = Jumlah klaim pengguna asuransi (yang dianggap sebagai tagihan oleh asuransi)

Bentuk data:

- Age = Integer
- BMI = Float
- Sex = String(male, female)
- Children = Integer
- Smoker = Boolean
- Region = String
- Charges = Float (dalam USD)

Descriptive Statistics Analysis

Rata-rata Umur Pengguna Asuransi

- Dengan menjumlahkan nominal seluruh umur pengguna asuransi dibagi jumlah count dari pengguna asuransi.
- Hasil: **39.20**

Rata-rata umur pengguna asuransi ialah 39 tahun

Rata-rata nilai BMI dari pengguna yang merokok

Data pengguna asuransi yang merokok dipisahkan lalu dicari rata-rata dari BMI nya

- Hasil: **30.70**

Nilai BMI dari pengguna asuransi yang merokok termasuk kategori tinggi, yang mana itu tidak sehat. Normalnya orang yang sehat memiliki BMI di kisaran 18-30.

Variance tagihan dari perokok dan non perokok

Data perokok dan non perokok dipisahkan kemudian masing-masing dicari variance dari tagihannya.

- Hasil:
  - Variance Tagihan Perokok: **132,721,153.14**
  - Variance Tagihan non Perokok: **35,891,656.00**

Terlihat varians tagihan perokok jauh lebih tinggi dari varians non perokok. Ini menunjukkan range tagihan perokok lebih tinggi.

Rata rata umur perempuan dan laki-laki yang merokok

Data perokok dipilih, kemudian laki-laki dan perempuan dipisah lalu dicari rata-rata umurnya.

- Hasil:
  - Rata-rata umur perokok laki-laki = **38.92 tahun**
  - Rata-rata umur perokok perempuan = **39.50 tahun**

Laki-laki dan perempuan yang merokok memiliki rata-rata umur yang berdekatan.

Tagihan Perokok Dengan BMI Kecil dan Besar

Data pengguna yang BMI >25 dipilih, kemudian dipisahkan antara perokok dan non perokok, lalu masing-masing dicari rata-rata tagihannya.

- Hasil:
  - Rata-rata tagihan kesehatan perokok = **$32,050.23**
  - Rata-rata tagihan kesehatan non perokok = **$8,434.27**

Tagihan perokok dengan BMI >25 jauh lebih tinggi dibandingkan non perokok dengan BMI>25.

Analysis

Data dari 1338 pengguna asuransi menunjukkan rata-rata umur dari pengguna asuransi sekitar 39 tahun.

Ditemukan perokok diantara pengguna asuransi rata-rata memiliki BMI 30.7, ini menunjukkan bahwa kelompok rata-rata perokok tidak memiliki postur badan ideal.

Perokok juga memiliki varians tagihan sekitar 4 kali lipat dibanding yang tidak merokok, yang berarti tagihan kesehatan perokok bisa mencapai 4 kali lebih besar dibanding non perokok apabila memiliki rata-rata tagihan yang sama.

Sebaran umur dari perokok antara laki-laki dan perempuan berbeda tipis dengan rata-rata terpaut 1,5 tahun menunjukkan bahwa jenis kelamin tidak mempengaruhi fase orang (dalam usia) untuk merokok.

Jika diperhatikan kelompok dengan BMI diatas 25, ditemukan bahwa rata-rata tagihan perokok 4 kali lebih besar dibanding non perokok. Ini menunjukkan perokok memiliki resiko kesehatan (resiko keuangan untuk pihak asuransi) yang jauh lebih tinggi dari non perokok.

Categorical Variables Analysis

Tagihan Maksimal Gender

Data dikelompokkan berdasarkan gender lalu dicari tagihan tertinggi tiap gender.

- Hasil:
  - Laki-lak: **$63,770.43**
  - Perempuan: **$62,592.87**

Tidak ada perbedaan signifikan antara laki-laki dan perempuan.

Distribusi jumlah tagihan di tiap-tiap region

Mencari proporsi jumlah tagihan pengguna asuransi pada tiap-tiap region.

- Hasil:

region	charges
northeast	0.24
northwest	0.23
southeast	0.30
southwest	0.23

Distribusi proporsi tagihan cukup seimbang pada tiap-tiap region.

Distribusi jumlah pengguna di tiap-tiap region

Mencari proporsi jumlah pengguna asuransi pada tiap-tiap region.

- Hasil:

region	people
northeast	324
northwest	325
southeast	364
southwest	325

Distribusi pengguna cukup seimbang pada tiap-tiap region.

Proporsi jumlah perokok

Mencari proporsi jumlah perokok asuransi dari seluruh pengguna.

- Hasil:

smoker	people
no	1064
yes	274

Mayoritas pengguna asuransi tidak merokok

Peluang perempuan jika diketahui seseorang perokok

Menghitung kemungkinan perempuan pada kelompok orang yang merokok

- Hasil: 0.42

Jika seseorang perokok, maka 42% kemungkinan ia adalah perempuan.

Peluang laki-laki jika diketahui seseorang perokok

Menghitung kemungkinan laki-laki pada kelompok orang yang merokok

- Hasil: 0.58

Jika seseorang perokok, maka 58% kemungkinan ia adalah laki-laki.

Analysis

Antara laki-laki dan perempuan tidak memiliki perbedaan yang tinggi pada tagihan maksimalnya. Masing-masing hanya terpaut $1,200. Ini menandakan laki-laki dan perempuan dapat terpapar resiko penyakit dengan biaya pengobatan yang maksimalnya sama.

Sebaran pengguna asuransi pada tiap region cukup merata. Hanya region southwest yang terlihat mencolok perbedaannya. Ini juga dialami oleh proporsi tagihan, yang mana southwest memiliki proporsi tagihan yang lebih tinggi dibanding yang lain. Ini menunjukkan penjualan polis yang cukup seimbang pada masing-masing region.

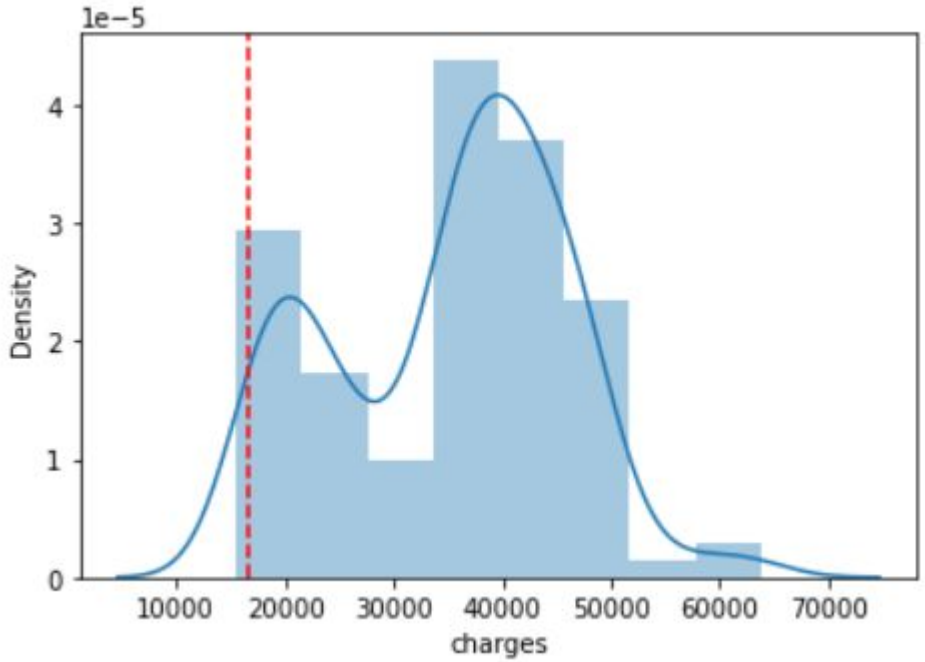
Sementara dari status merokok peserta asuransi, mayoritas tidak merokok. Adapun dari kelompok perokok, 58% adalah laki-laki sementara 42% adalah perempuan. Ini menunjukkan perokok tidak timpang berbeda antar gender.

Continuous Variables Analysis

Peluang Tagihan Pada Kelompok Tertentu

Peluang seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700.

Hasil: 0.95

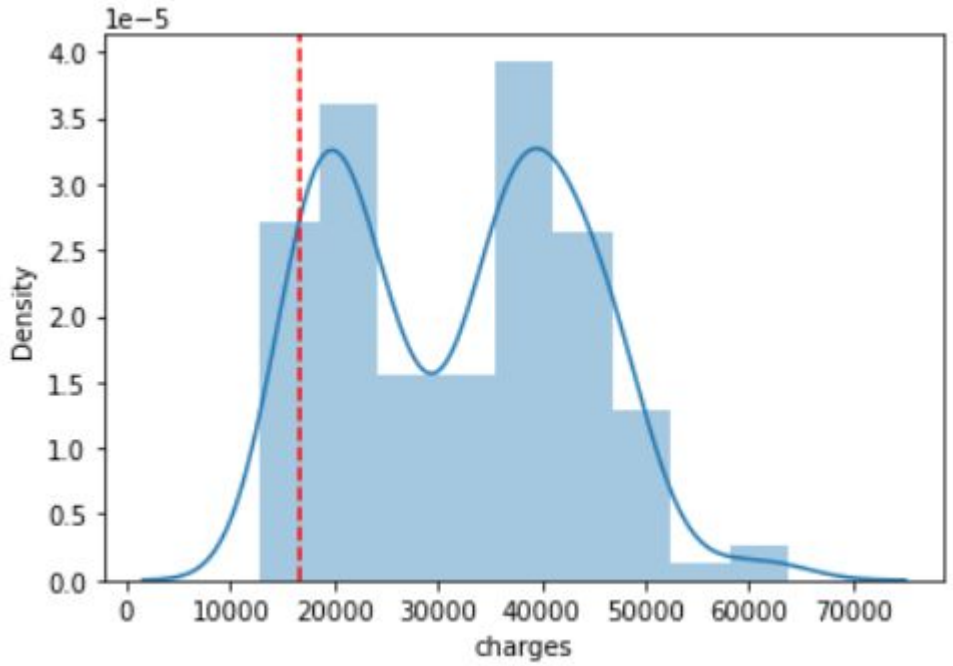


Analisis: Ini menunjukkan bahwa perokok dengan BMI >25 memiliki kemungkinan sangat besar untuk memiliki tagihan diatas 16.7k

Peluang Tagihan Pada Kelompok Tertentu

Peluang seorang perokok mendapatkan tagihan kesehatan di atas 16.700.

Hasil: 0.91

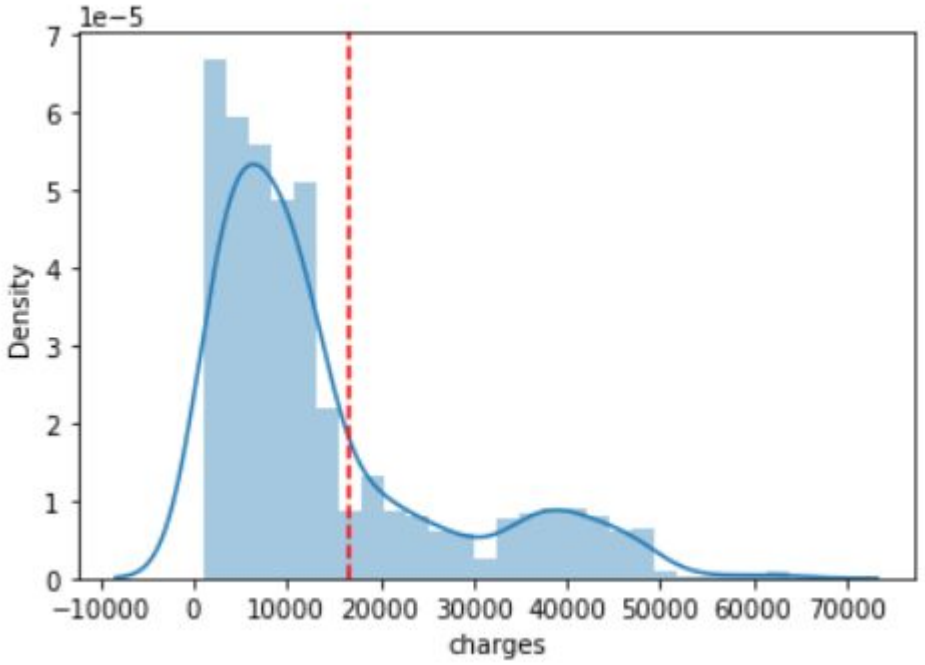


Analisis: Ini menunjukkan bahwa perokok dengan BMI berapapun memiliki kemungkinan sangat besar untuk memiliki tagihan diatas 16.7k. Pada dasarnya perokok memiliki resiko kesehatan yang cukup tinggi.

Kemungkinan terjadi lebih besar

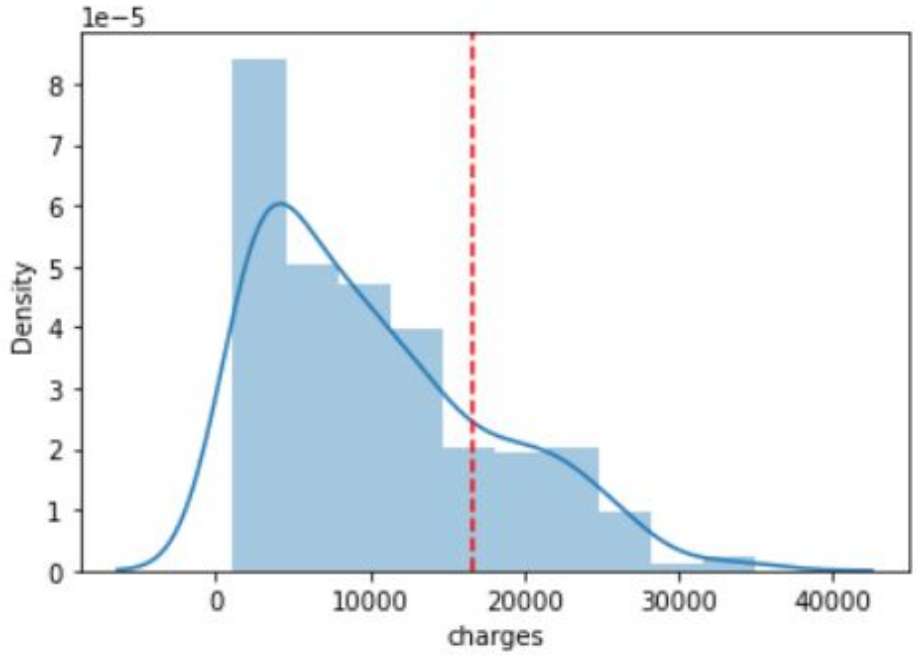
Seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k

Hasil: 0.415



Seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k

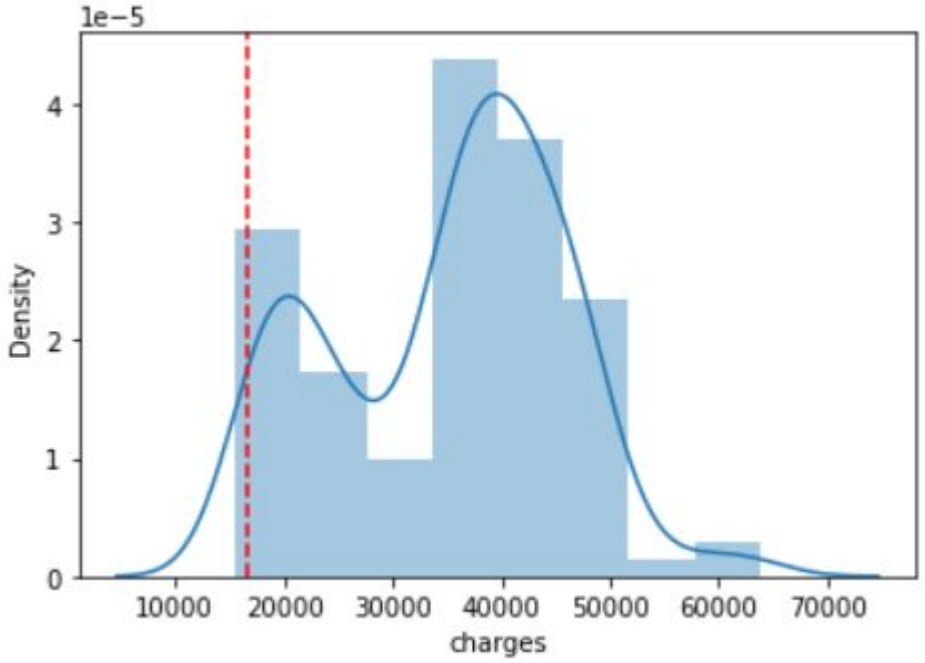
Hasil: 0.196



Kemungkinan terjadi lebih besar

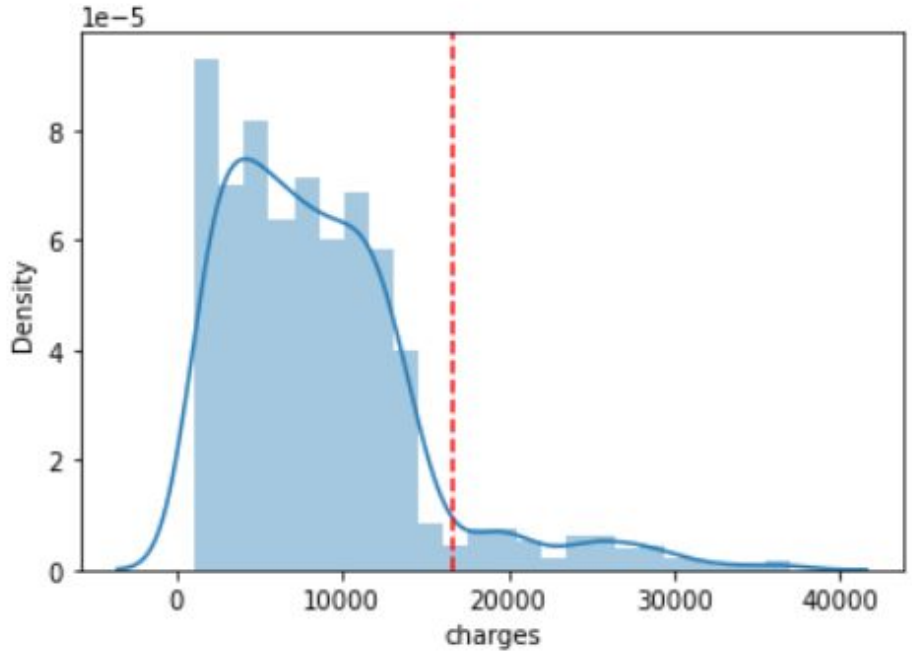
Seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k

Hasil: 0.958



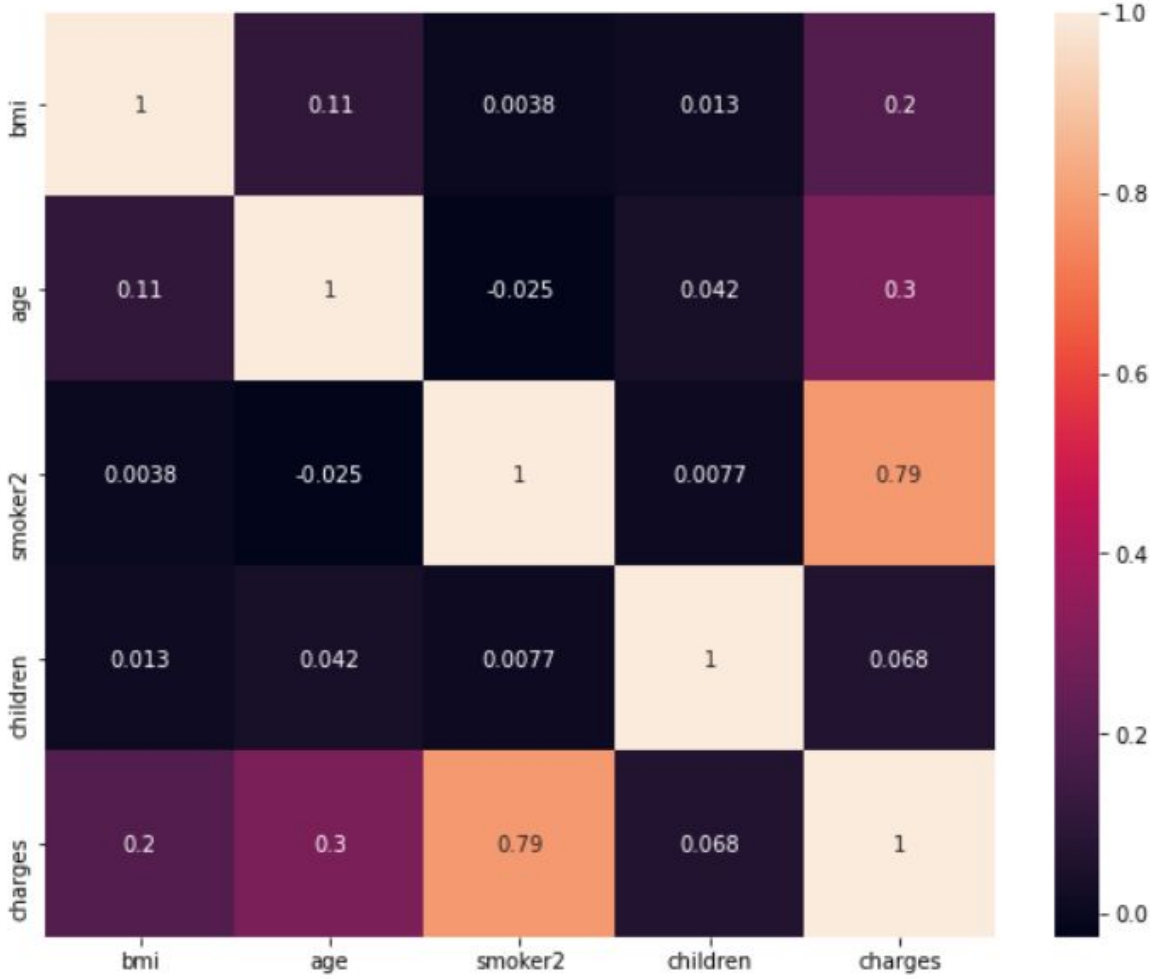
Seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k

Hasil: 0.090



Variables Correlation

Correlation



Tabel heatmap menunjukkan bahwa umur, BMI dan jumlah anak tidak memiliki korelasi yang kuat terhadap naik turunnya tagihan.

Sementara smoker memiliki korelasi yang sangat kuat.

Artinya resiko kesehatan terbesar dibawa oleh status perokoknya seseorang. Jauh lebih besar dibanding variabel-variabel lainnya.

Hypothesis Testing

Tagihan Perokok

- Tes hipotesis dengan tingkat signifikansi 95% pada data perokok dan sampel non perokok menunjukkan:
    - Nilai p-value yang didapatkan adalah 1.0
    - Karena lebih dari alpha, maka klaim bahwa tagihan perokok lebih tinggi dibanding non perokok **diterima**, karena ada cukup bukti statistik untuk membuktikan klaim tersebut

Tagihan & BMI

- Tes untuk menguji apakah tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25.
- Hipotesis:
  - H0: Tagihan BMI besar = Tagihan BMI kecil
  - H1: Tagihan BMI besar < Tagihan BMI kecil

- Tes hipotesis dengan tingkat signifikansi 95% pada data pengguna dengan BMI >25 dan pengguna dengan BMI <25 menunjukkan:

  - Nilai p-value yang didapatkan adalah 0.99
  - Karena lebih dari alpha, maka klaim bahwa tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25 **diterima**, karena ada cukup bukti statistik untuk membuktikan klaim tersebut

BMI & Gender

- Tes untuk menguji apakah BMI laki-laki dan perempuan sama.
- Hipotesis:
  - H0: BMI laki-laki = BMI perempuan
  - H1: BMI laki-laki ≠ BMI perempuan

- Tes hipotesis dengan tingkat signifikansi 95% pada data pengguna menunjukkan:

  - Nilai p-value yang didapatkan adalah 0.08037162
  - Karena lebih dari alpha, maka klaim bahwa BMI laki-laki dan perempuan sama **diterima**, karena ada cukup bukti statistik untuk membuktikan klaim tersebut

Conclusion

Conclusion

- Status merokok seseorang berpengaruh sangat besar terhadap tagihan kesehatannya.
- Umur, diluar dugaan, tidak begitu pengaruh terhadap tagihan kesehatan seseorang. Ini menunjukkan banyak juga usia muda yang memiliki kondisi kesehatan yang kurang baik.
- Asuransi secara umum memiliki pengguna yang tersebar dengan baik, dengan tingkat pengguna beresiko tinggi (merokok) yang masih moderat.

Notes

- Tes hipotesis perlu dilakukan normalisasi data. Saya saat ini masih terbatas kemampuan untuk melakukan hal tersebut. Hal ini membuat p value dari test 1 dan 2 agak kurang intuitif.