Pacmann

Analisa Dataset Asuransi Kesehatan

Probability Course - Sekolah Data Pacmann

Outline

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

Introduction

Introduction

Pada proyek kali ini kita akan coba mengeksplorasi *dataset* asuransi kesehatan untuk melihat hubungan antar variabel-variabel pada data pengguna dengan tagihan kesehatan

Dataset

Dataset

- age
- sex
- bmi
- children
- smoker
- region
- charges

Descriptive Statistics Analysis

Mean of Age

- Berapa rata rata umur pada data tersebut?

```
Rerata Umur : 39.21
```

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

	rerata umur laki-laki	rerata umur perempuan
	38.92	39.50
perokok	38.45	38.61
non perokok	39.06	39.69

Mean of BMI

- Berapa rata rata nilai BMI dari yang merokok?

```
Rerata BMI              : 30.66
Rerata BMI perokok      : 30.71
tRerata BMI non perokok : 30.65
```

Mean of Charges

- Apakah variansi dari data charges perokok dan non perokok sama?

	tagihan perokok	tagihan non perokok
rerata	32050.23	8434.27
var	133207311.21	35925420.50
std	11541.55	5993.78

- Mana yang lebih tinggi, rata rata tagihan kesehatan perokok yang BMI nya diatas 25 atau non perokok yang BMI nya diatas 25?

rerata tagihan perokok bmi > 25	rerata tagihan non perokok bmi > 25
35116.91	8633.96

Analysis

Dataset yang dipakai, diambil dari 1338 orang dengan rata-rata umur 39 tahun yang rataan ini kurang lebih sama atau berdekatan dengan rata-rata umur pada ada laki-laki ataupun perempuan yang perokok ataupun non perokok. Rata-rata nilai BMI mereka adalah 30.71, yang ini sudah termasuk dalam kategori obesitas, walaupun nilainya tidak jauh dari rentang overweght (25 - 29,9). Dari sini mungkin akan timbul pertanyaan apakah kemudian rerata BMI yang diatas normal ini akan berpengaruh pada tagihan kesehatan?

Untuk rata-rata tagihan terlihat perbedaan yang cukup besar antara data perokok dengan yang bukan perokok. Dimana rata-rata perokok memiliki tagihan 32.050 dengan sebaran sebesar 11.541, sedangkan yang bukan perokok memiliki tagihan 8434 dengan sebaran 5993. Dari hal tersebut terlihat bahwa variabel perokok cukup mempengaruhi besarnya tagihan kesehatan.

Categorical Variables Analysis

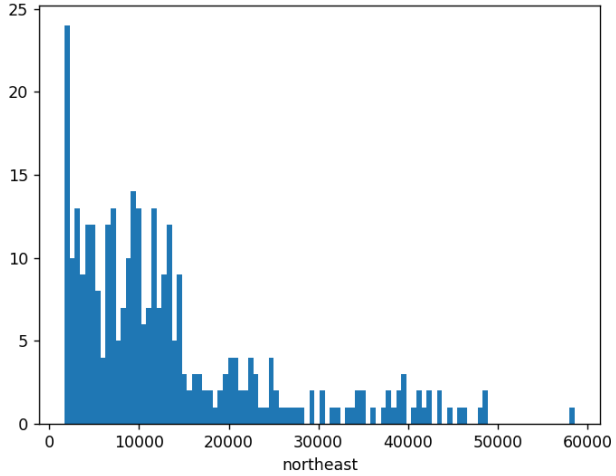
Proporsion of smokers and non smokers

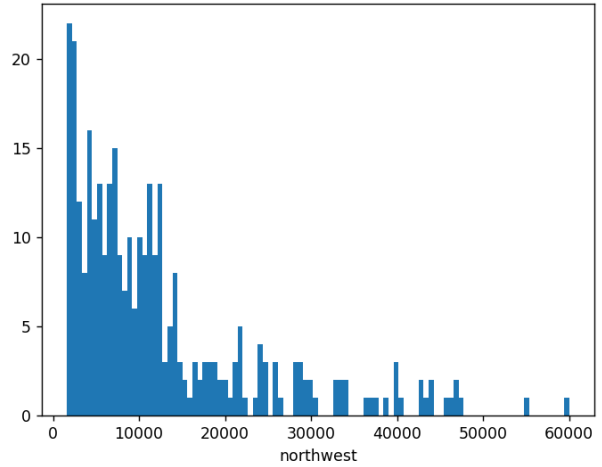
- Mana yang lebih tinggi proporsi perokok atau non perokok

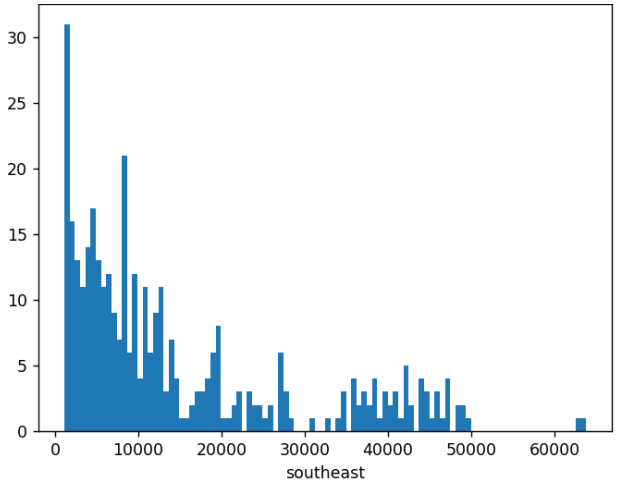
	perokok	non perokok
jumlah	274	1064
proporsi	0.20	0.80

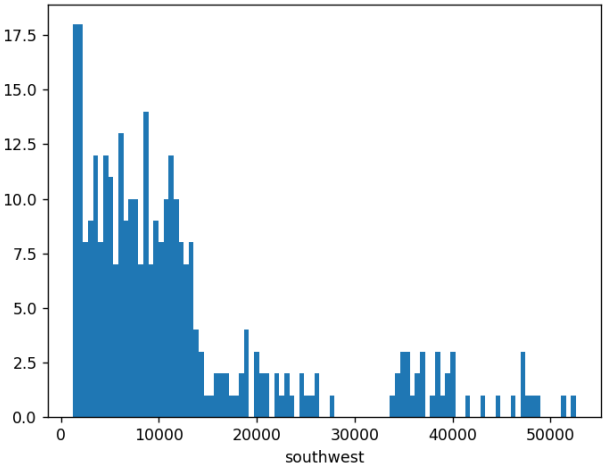
Proporsion of charges in each region

- Distribusi peluang tagihan di tiap-tiap region









Proporsion of charges based on sex

- Jenis kelamin mana yang memiliki tagihan paling tinggi?

	tagihan laki2	tagihan perempuan
rerata	13956.75	12569.58
maksimal	62592.87	63770.43
median	9369.62	9412.96

Etc..

- Berapa peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok?
- Berapa peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok?

	perempuan	laki-laki
jika perokok	0.42	0.58

Analysis

Pada bagian ini kita bisa melihat bahwa tidak terlihat adanya perbedaan yang berarti antara tagihan kesehatan dengan jenis kelamin ataupun wilayah. Namun hal menarik pada bagian ini adalah proporsi jumlah perokok yang hanya ¼ dari yang tidak merokok, padahal jika dibagian sebelumnya terlihat bahwa rerata tagihan perokok hampir 4 x dari rerata tagihan non perokok.

Continuous Variables Analysis

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

- Mencari kemungkinan terjadi, seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700

```
Peluang bmi >= 25 mendapat tagihan > 16.7k : 0.26
```

- Berapa peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia adalah perokok

```
Peluang tagihan > 16.7k jika perokok : 0.93
```

BMI vs Smokers

- Mana yang lebih mungkin terjadi
  - Seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau
  - Seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k

```
==========================
Peluang tagihan > 16.7k
==========================
bmi < 25      bmi >= 25
  0.21          0.26
```

- Mana yang lebih mungkin terjadi
  - Seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau
  - Seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan

```
====================================
Peluang tagihan > 16.7k & bmi >= 25
====================================
perokok          non_perokok
  0.98             0.08
```

Analysis

Dari perbandingan diatas kita dapat melihat bahwa variabel perokok lebih lebih mempunyai kaitan yang cukup besar pada tagihan kesehatan dibandingkan dengan variabel BMI

Variables Correlation

Correlation

Hypothesis Testing

Smoker's charges are higher than non smoker's

- Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok

```
H0 : Tagihan kesehatan perokok <= Tagihan kesehatan non perokok
Ha : Tagihan kesehatan perokok > Tagihan kesehatan non perokok
   z_kritis      t_uji kesimpulan
0     1.645   8.788974   Tolak Ho
1     1.645   9.798209   Tolak Ho
2     1.645   7.225025   Tolak Ho
3     1.645   8.717845   Tolak Ho
4     1.645   8.694976   Tolak Ho
5     1.645   9.072342   Tolak Ho
6     1.645   8.008495   Tolak Ho
7     1.645  13.126667   Tolak Ho
8     1.645   7.260771   Tolak Ho
9     1.645  11.886064   Tolak Ho
```

Hypothesis Testing #2 and Answer

- Tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25

```
H0 : Tagihan kesehatan BMI diatas sama dengan 25 <= tagihan BMI dibawah 25
Ha : Tagihan kesehatan BMI diatas  25 > tagihan BMI dibawah 25
   z_kritis     t_uji kesimpulan
0     1.645  1.128593  Terima H0
1     1.645  0.891912  Terima H0
2     1.645  1.713206   Tolak Ho
3     1.645 -0.082010  Terima H0
4     1.645  1.345015  Terima H0
5     1.645  0.196113  Terima H0
6     1.645  1.732717   Tolak Ho
7     1.645  0.654518  Terima H0
8     1.645  1.660870   Tolak Ho
9     1.645 -0.237519  Terima H0
```

Etc..

- Tagihan kesehatan laki-laki lebih besar dari perempuan

```
H0 : Tagihan kesehatan laki-laki <= perempuan
Ha : Tagihan kesehatan laki-laki > perempuan
   z_kritis     t_uji kesimpulan
0     1.645 -0.261701  Terima H0
1     1.645  1.146294  Terima H0
2     1.645  1.369589  Terima H0
3     1.645 -0.368122  Terima H0
4     1.645 -0.208090  Terima H0
5     1.645 -0.369833  Terima H0
6     1.645  1.728887   Tolak Ho
7     1.645  1.460162  Terima H0
8     1.645  1.762738   Tolak Ho
9     1.645  1.550484  Terima H0
```

Conclusion

Conclusion

Dari eksplorasi data ini terlihat jelas bahwa variabel perokok mempunyai hubungan (relasi) yang kuat dan positif terhadap besarnya tagihan kesehatan, sedang yang lainnya relatif mempunyai hubungan yang lemah (seperti bmi) bahkan tidak ada hubungan terhadap tagihan kesehatan. Oleh karena itu ada baiknya pihak asuransi memberikan perhatian khusus dan mengambil langkah tertentu terkait faktor perokok ini, seperti mengkampanyekan atau menggalakkan hidup sehat dengan tidak merokok dan iainnya

Notes

- Pengambilan sample yang lebih banyak pada uji hipotesa
- Penggunaan metode uji hipotesis lainnya seperti z test, dan lainnya

Reference

- Sekolah data pacmann (siswa.pacmann.ai, live class, etc)
- https://docs.scipy.org
- https://pandas.pydata.org/docs/
- https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm