Pacmann

Insurance Project

Probability Course - Sekolah Data Pacmann

Outline

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

Introduction

Introduction

- Project ini dibuat untuk melakukan analisis data pada tagihan kesehatan.
- Tujuan dari project ini adalah untuk mengetahui variable yang dapat mempengaruhi besaran tagihan kesehatan.
- Melalui project dilakukan analisis deskriptif statistic, analisis variable kategorikal, analisis variable continuous, analisis korelasi, dan pengujian hipotesis.

Dataset

Dataset

- Dataset yang akan digunakan adalah dataset tagihan kesehatan.
- Terdapat 7 variable dalam data set ini yaitu age, sex, bmi, children, smoker, region, charges

Data preprocessing:

1. Melakukan identifikasi terhadap jenis data
2. Mencari null dan duplicate value (ditemukan satu duplicate value dan telah dihilangkan dari dataset)
3. Mengubah categorical value dengan dummy variabel

Descriptive Statistics Analysis

Mean of Age

- Hasil yang diperoleh: Rata rata usia dari data adalah 39,22 tahun

	age	bmi	children	charges	sex_male	smoker_yes
count	1337.000000	1337.000000	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487	0.504862	0.204936
std	14.044333	6.100468	1.205571	12110.359656	0.500163	0.403806
min	18.000000	15.960000	0.000000	1121.873900	0.000000	0.000000
25%	27.000000	26.290000	0.000000	4746.344000	0.000000	0.000000
50%	39.000000	30.400000	1.000000	9386.161300	1.000000	0.000000
75%	51.000000	34.700000	2.000000	16657.717450	1.000000	0.000000
max	64.000000	53.130000	5.000000	63770.428010	1.000000	1.000000

Rata-rata usia dari pengguna yang merokok

- Apabila di breakdown ke dalam gender rata-rata usia dari pengguna yang merokok hampir sama yaitu kurang lebih berusia 38 tahun
- rata-rata usia wanita yang merokok adalah 38.608696 sedangkan pria yang merokok berusia 38.446541

Sex	Smoker	Age_mean
Female	No	39.691042
	Yes	38.608696
Male	No	39.100775
	Yes	38.446541

Rata-rata nilai BMI

- BMI digunakan untuk memahami kondisi tubuh seseorang apakah ideal atau tidak. Berdasarkan hasil perhitungan diperoleh informasi bahwa rata-rata **BMI perokok** sebesar **30,7085**. Apabila di breakdown ke dalam kategori perokok, maka rata-rata BMI perokok lebih besar dibandingkan rata-rata BMI nonperokok

- Apabila di breakdown ke dalam gender, maka rata-rata BMI pria lebih besar dibandingkan rata-rata BMI wanita

Sex	Bmi_mean
Female	30.377749
Male	30.943652

Smoker	Bmi_mean
No	30.651853
Yes	30.708449

Rata-rata nilai BMI dari pengguna yang merokok

- Rata-rata BMI pria yang merokok **lebih besar** yaitu sebesar 31.504182 sedangkan Wanita yang merokok rata-rata BMI nya sebesar 29.608261

Sex	Smoker	Bmi_mean
Male	No	30.770930
	Yes	31.504182
Female	No	30.539525
	Yes	29.608261

Rata-Rata tagihan kesehatan perokok dan nonperokok

- Berdasarkan hasil perhitungan, **rata-rata tagihan kesehatan perokok lebih besar 279,73 persen** dibandingkan non perokok

Smoker	Charges_mean
No	8440.660307
Yes	32050.231832

Varians dari perokok dan non perokok

- Berdasarkan hasil perhitungan, **varians charges dari perokok lebih besar** dibandingkan non perokok

Smoker	Charges_var
No	3.588195e+07
Yes	1.327212e+08

Smoker	Charges_std
No	5990.154240
Yes	11520.466707

Analysis

Berdasarkan hasil analisis dari 1337 orang dengan rata-rata usia kurang lebih 39 tahun dengan rata-rata BMI sebesar 30,66 kg/m2 menunjukkan bahwa rata-rata berat banda orang-orang dalam data masuk ke dalam **kategori obesitas*.** Kategori perokok memiliki rata-rata bmi yang lebih tinggi dibandingkan non perokok.

Tagihan Kesehatan yang dibayarkan perokok juga lebih besar dibandingkan non perokok yaitu sebesar 32.050 atau lebih besar 279,73 persen dibandingkan non perokok. Apabila dilihat sebarannya, sebaran tagihan dari perokok lebih besar dibandingkan non perokok.

**) BMI interpretation source: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#InterpretedAdults*

Categorical Variables Analysis

Tagihan berdasarkan gender

- Berdasarkan hasil perhitungan maksimum tagihan pada tiap gender, diperoleh bahwa tagihan tertinggi ada pada pria .

Sex	Charges_max
Female	12569.57884
Male	13974.99886

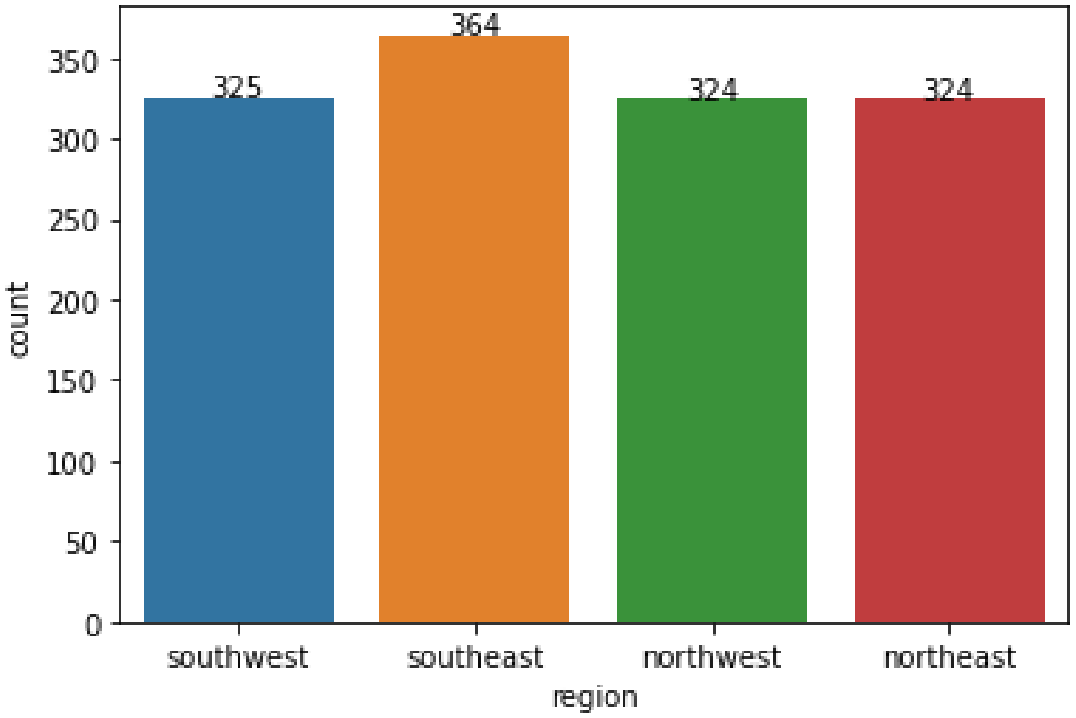
Peluang distribusi tagihan di setiap daerah

- Peluang distribusi tertinggi ada pada daerah Southeast sebesar 27%.

Region	Peluang Distribusi
southeast	27%
southwest	24%
northeast	24%
northwest	24%

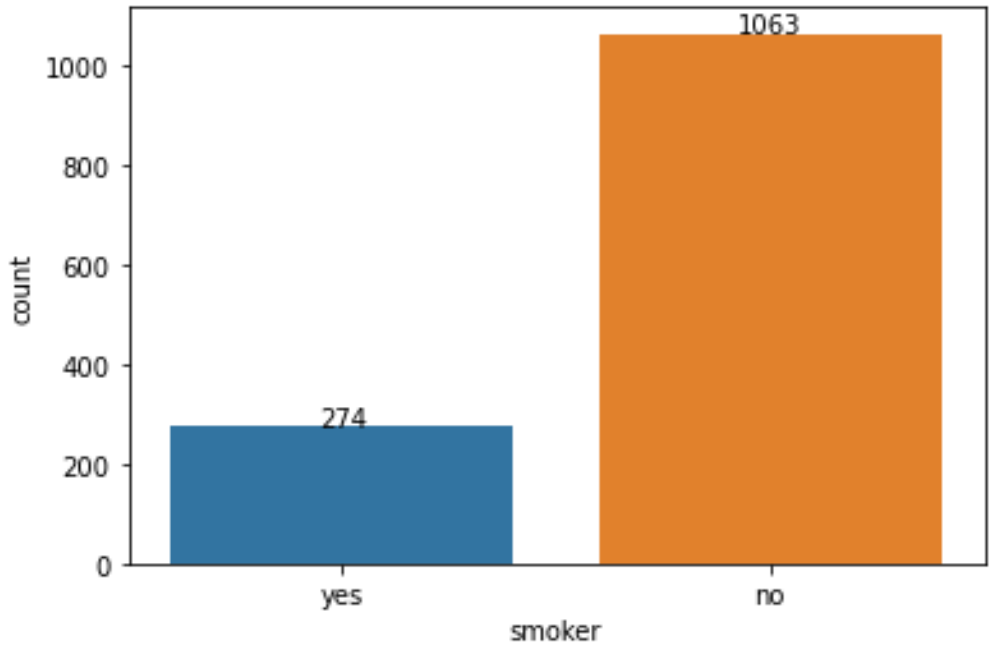
Proporsi orang pada setiap region

- Proporsi pada setiap region tidak sama, proporsi terbesar ada pada daerah southeast



Proporsi perokok dan non perokok

- Jumlah perokok lebih kecil dibandingkan jumlah yang tidak merokok.



Peluang perokok berdasarkan gender

- Peluang dari perempuan dan seorang perokok adalah 41,97 persen
- Peluang dari laki-laki dan seorang perokok adalah 58,03

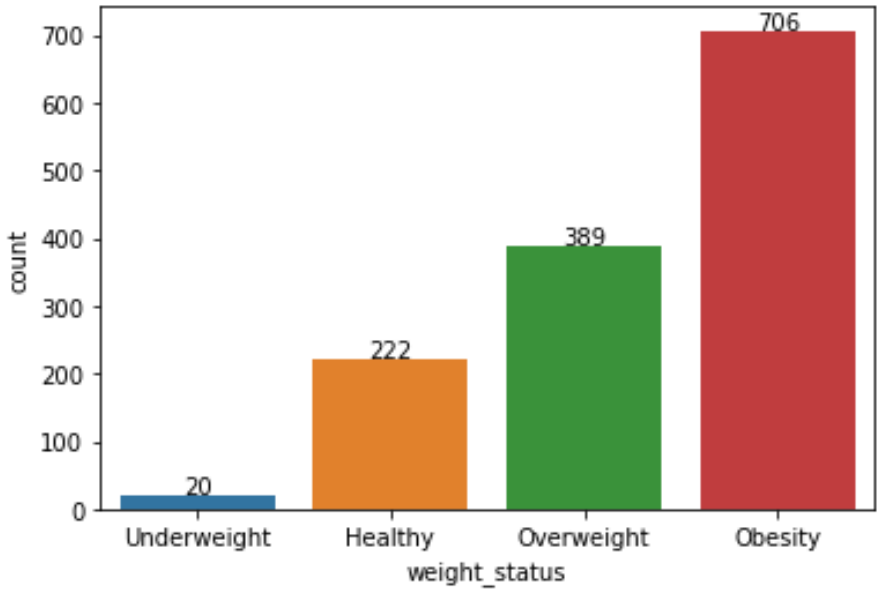
Sex	Smoker	prob
Male	No	0.485419
	Yes	0.580292
Female	No	0.514581
	Yes	0.419708

Analysis

- Peluang distribusi tertinggi ada di southeast yang mana distribusinya juga merupakan yang paling besar
- Jumlah perokok lebih sedikit dibandingkan non perokok
- Peluang seorang laki laki merupakan perokok lebih besar dibandingkan perempuan seorang perokok

Continuous Variables Analysis

Peluang tagihan



BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Healthy Weight
25.0 – 29.9	Overweight
30.0 and Above	Obesity

- Peluang seseorang yang underweight dan memiliki tagihan Kesehatan > 16,7k adalah 0,14 %
- Peluang seseorang yang sehat dan memiliki tagihan Kesehatan > 16,7k adalah 3,59 %
- Peluang seseorang yang overweight dan memiliki tagihan Kesehatan > 16,7k adalah 14,29%
- Peluang seseorang yang obesitas dan memiliki tagihan Kesehatan > 16,7k adalah 16,08%

Peluang tagihan

- Peluang seseorang dengan tagihan diatas 16,7k adalah perokok sebesar 18,99 persen
- Peluang seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k sebesar 21,17%. Peluang ini lebih besar dibandingkan seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k yang sebesar 3,81 %.
- Peluang seorang perokok dengan BMI diatas 25 mendapatkan tagihan Kesehatan diatas 16.7k sebesar 16,08 %. Peluang ini lebih besar dibandingkan peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan Kesehatan diatas 16.7k yang sebesar 5,09%

Analysis

- Berdasarkan hasil perhitungan, peluang seseorang membayar tagihan > 16,7k pada seseorang berat badan diatas 25 lebih besar dibandingkan orang yang berat badannya dibawah 25
- peluang perokok membayar tagihan dalam jumlah yang tinggi lebih besar dibandingkan non perokok
- Peluang seorang perokok dengan BMI diatas 25 mendapatkan tagihan Kesehatan diatas 16.7k lebih besar dibandingkan peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan Kesehatan diatas 16.7k.

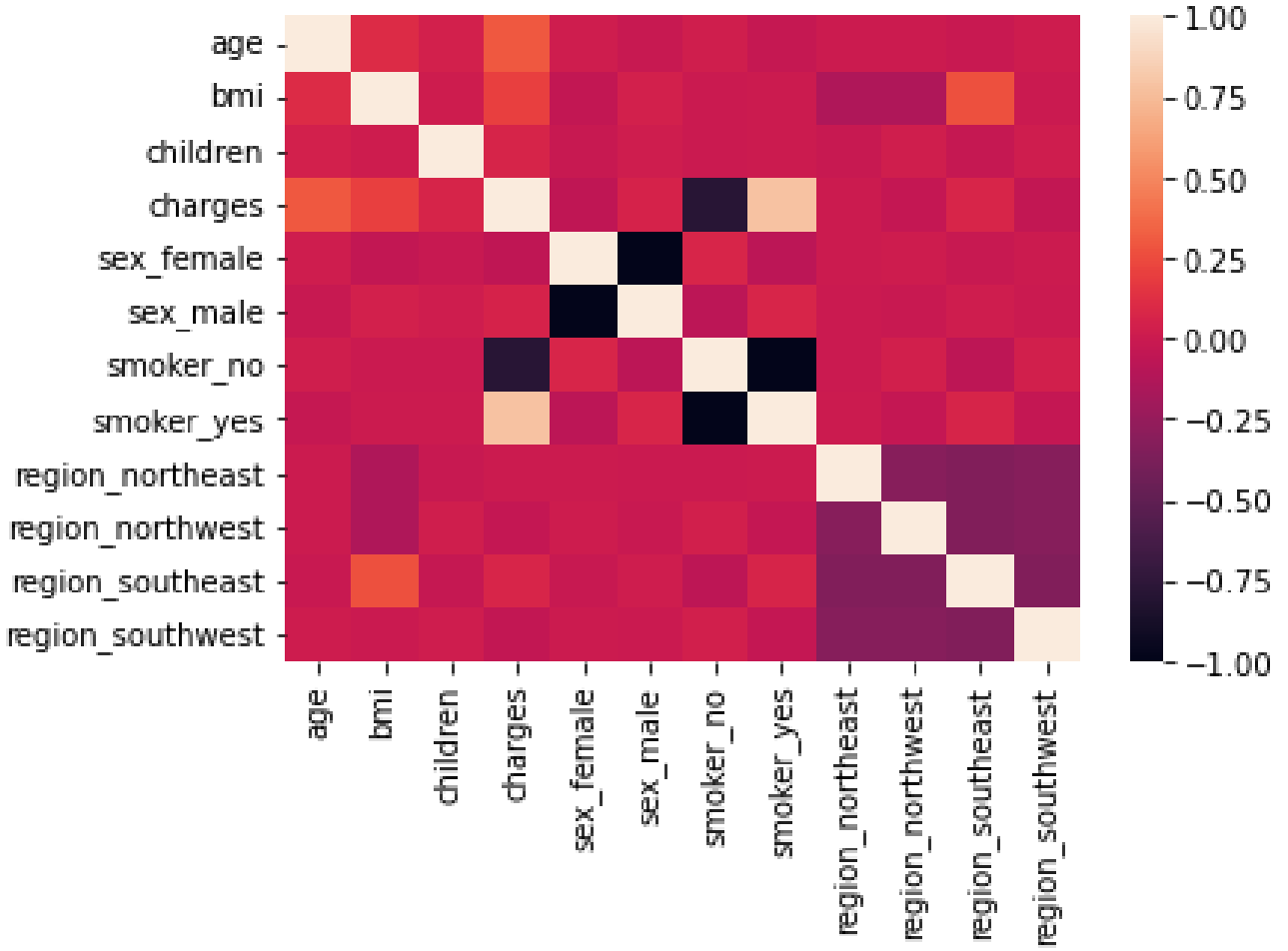
Variables Correlation

Correlation

Berdasarkan hasil korelasi, korelasi paling kuat ada pada besar tagihan dengan perokok dengan nilai korelasi 0,79.

Selanjutnya, umur dan bmi juga berkolerasi kuat terhadap tagihan.

Disisi lain, non perokok berkolerasi negative dengan jumlah tagihan



Hypothesis Testing

Kerangka kerja

Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok

H0: μ(perokok) > μ(nonperokok)
H1: μ(perokok) ≤ μ(nonperokok)
α = 0,05

- Nilai p-value yang dapatkan adalah 1.0
- Karena lebih dari alpha, maka kita gagal menolak klaim μ(perokok) > μ(nonperokok) , karena belum ada cukup bukti statistik untuk menolak klaim tersebut

Tagihan kesehatan dengan BMI di atas 25 lebih tinggi daripada tagihan Kesehatan dengan BMI di bawah 25

Pacmann

H0: $\mu(BMI > 25) > \mu(BMI < 25)$
H1: $\mu(BMI > 25) \leq \mu(BMI < 25)$
$\alpha = 0{,}05$

- Nilai p-value yang dapatkan adalah 0,7606
- Karena lebih dari alpha, maka kita gagal menolak klaim $\mu(BMI > 25) > \mu(BMI < 25)$, karena belum ada cukup bukti statistik untuk menolak klaim tersebut

Hypothesis Testing BMI laki-laki dan perempuan sama

H0: μ(BMI male) = μ(BMI female)
H1: μ(BMI male) ≠ μ(BMI female)
α = 0,05

- Nilai p-value yang dapatkan adalah 0,0002
- Karena kurang dari alpha, maka kita menolak klaim μ(BMI male) = μ(BMI female) , karena belum ada cukup bukti statistik untuk menerima klaim tersebut

Conclusion

Conclusion

Berdasarkan hasil Analisa, perokok membayar tagihan kesehatan lebih besar dibandingkan non perokok yaitu sebesar 32.050 atau lebih besar 279,73 persen dibandingkan non perokok.

Tagihan kesehatan bagi orang dengan BMI di atas 25 lebih tinggi daripada tagihan Kesehatan dengan BMI di bawah 25.

BMI laki-laki tidak sama dengan perempuan

Notes

- Untuk selanjutnya dapat menguji hipotesis untuk mengevaluasi korelasi antara besaran tagihan dengan perokok

Reference

- https://levelup.gitconnected.com/cozy-up-with-your-data-6aedfb651172
- https://towardsdatascience.com/11-simple-code-blocks-for-complete-exploratory-data-analysis-eda-67c2817f56cd
- https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#InterpretedAdults