Pacmann

Insurance

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 sebagai proteksi diri seseorang dari ketidakpastian di masa depan. Salah satu asuransi populer yang digunakan oleh maysrakat adalah asuransi kesehetan. Alasan mengapa asuransi tersebut populer karena pembiayaan terhadap fasilitas kesehatan termasuk obat-obatan tidaklah sedikit. Untuk itu, perusahaan biasanya memiliki pendataan dalam *profling* pengguna. Data insurance yang dibagikan berisikan profile dari setiap penggunanya.

Dataset

Dataset

- Data yang digunakan berupa insurance.csv dengan total baris mencapai 1338 dan 7 kolom. Adapun, kolom-kolom tersebut di antaranya:
    1. age: Age of primary beneficiary
    2. sex: Insurance contractor gender, female, male
    3. 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
    4. children: Number of children covered by health insurance / Number of dependents
    5. smoker: Smoking:
    6. region The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
    7. charges: Individual medical costs billed by health insurance

Descriptive Statistics Analysis

Summary Statisktik Secara Keseluruhan

Summary statistik

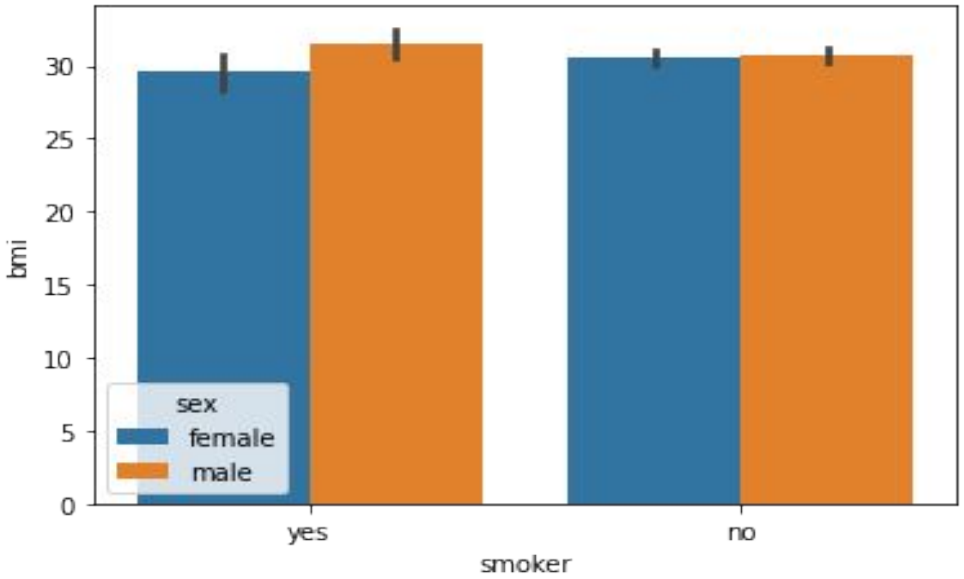
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Informasi data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

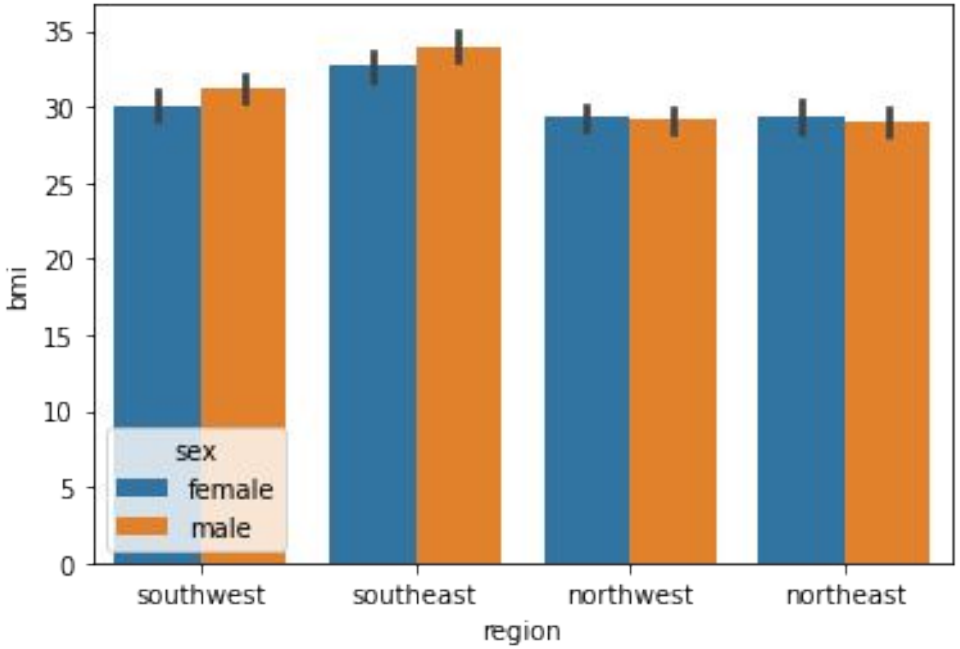
Categorical Variables Analysis

Bmi orang yang merokok/tidak berdasarkan gender



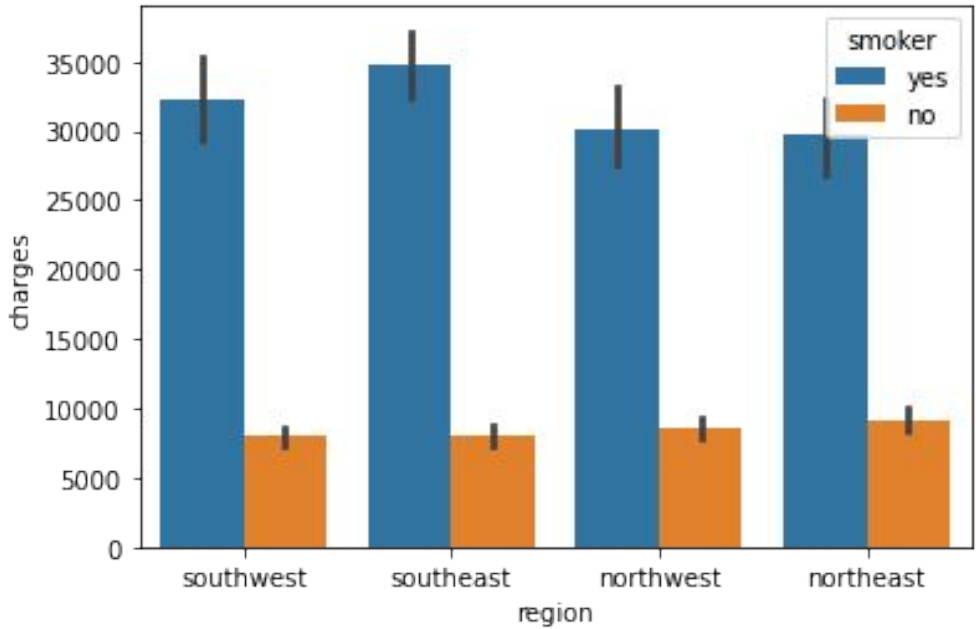
- Untuk orang yang merokok, bmi mereka lebih tinggi dibandingkan yang tidak.
- Berbeda dengan orang yang tidak merokok, bmi mereka kurang lebih sama

Peroleh BMI di setiap region berdasarkan gender



- Secara umum BMI paling besar dimiliki region southeast
- Untuk southwest dan southeast porsi bmi lebih besar dimiliki oleh pria sedangkannorthwest dan northeast dimiliki oleh wanita

Bmi orang yang merokok/tidak berdasarkan gender



- Secara garis besar, orang yang merokok di semua region memiliki total charges yang lebih besar dibanding tidak merokok

Probabilitas dari visualisasi di atas

Berdasarkan visualisasi sebelumnya, terdapat pertanyaan yang dijawab melalui probabilitas dan ekspetasi, sebagai berikut:

1. Berapa peluang ia laki-laki jika ia merokok?
2. Berapa peluang ia perempuan jika ia merekok?
3. Dari biaya charges yang dikeluarkan, mana ekspetasinya lebih antara orang yang merokok atau tidak
4. Manakah ekspetasi charges yang paling besar (bill insurance)?

```
sex_smoker["male"]["yes"]/np.sum(sex_smoker.sum())
```

```
0.11883408071748879
```

```
sex_smoker["female"]["yes"]/np.sum(sex_smoker.sum())
```

```
0.08594917787742899
```

```
[59] bill_insurance_smoker = df_insurance.pivot_table(index="smoker", values="charges",aggfunc="mean")
     bill_insurance_smoker
```

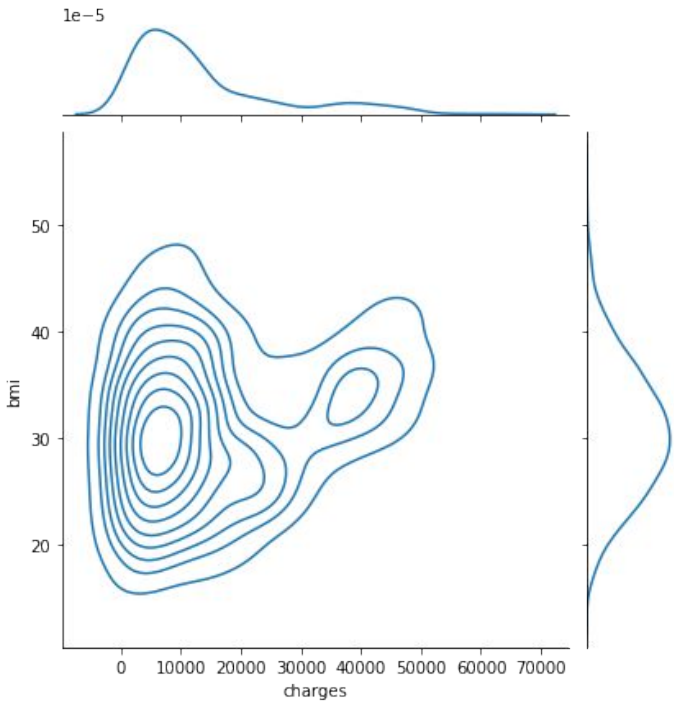
smoker	charges
no	8434.268298
yes	32050.231832

```
bill_insurance_smoker = df_insurance.pivot_table(index="region", values="charges",aggfunc="mean")
bill_insurance_smoker
```

region	charges
northeast	13406.384516
northwest	12417.575374
southeast	14735.411438
southwest	12346.937377

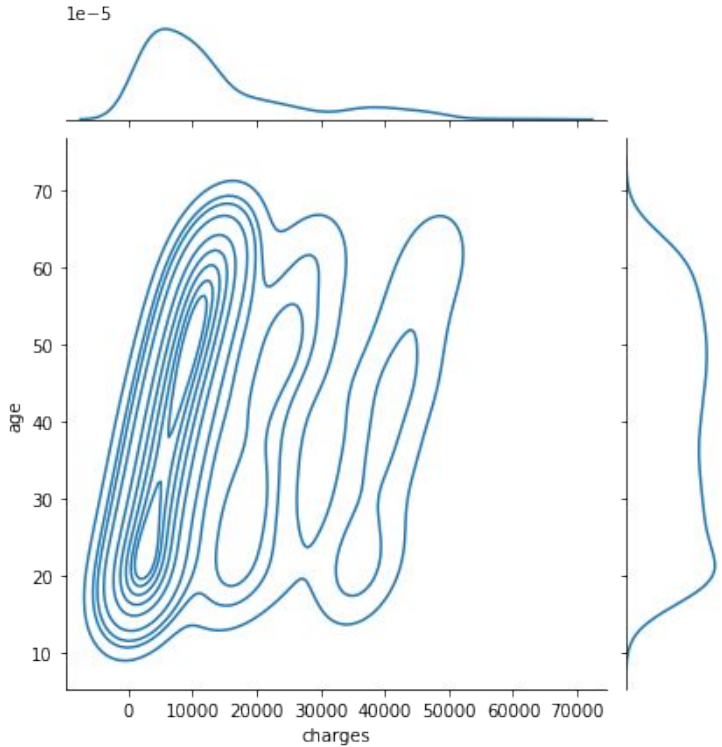
Continuous Variables Analysis

BMI vs Charges



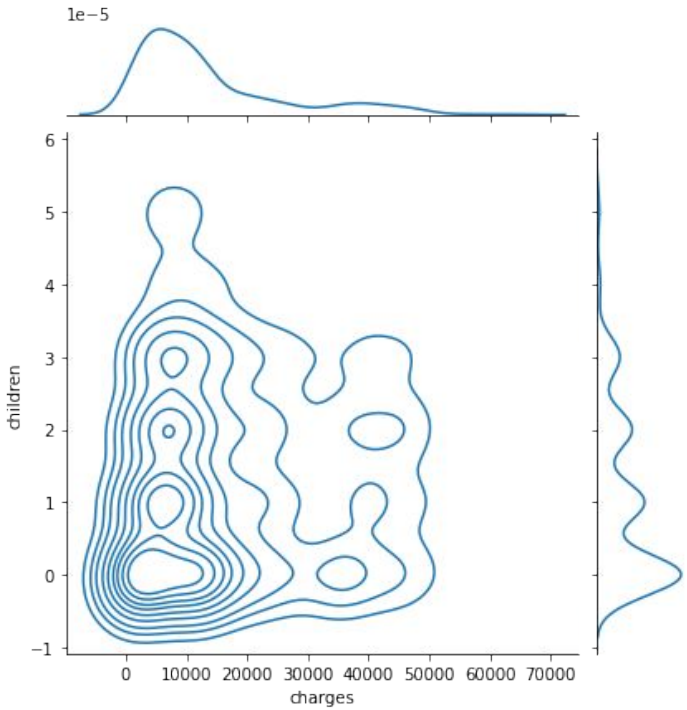
- Hasil cdf bukan probabilitas, dari grafik di atas dapat disimpulkan bahwa persebaran bmi di rangw 20-40 dan charges 0-30.000 lebih banyak terlihat dari luasannya dari kontur join disribusinya

Age vs Charges



- Dari grafik di atas dapat disimpulkan bahwa persebaran age di range 10-70 dan charges 0-10.000 lebih banyak terlihat dari luasannya dari kontur join disribusinya

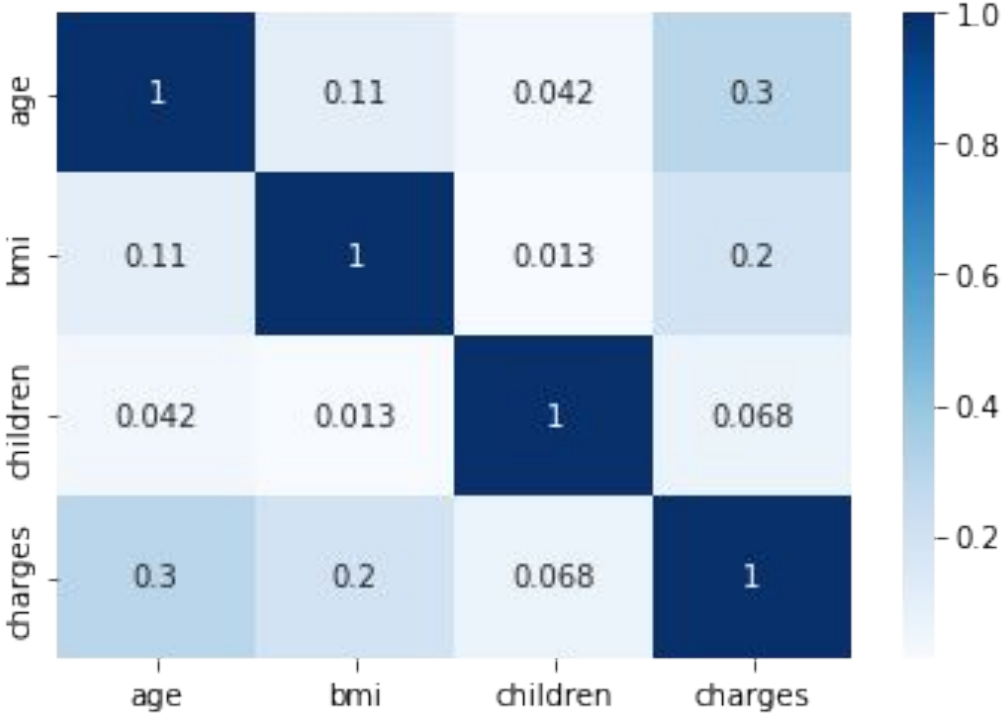
Children vs Charges



- Dari grafik di atas dapat disimpulkan bahwa persebaran jumlah anak di rangw 0-3 dan charges 0-20.000 lebih banyak terlihat dari luasannya dari kontur join disribusinya

Variables Correlation

Correlation



Hypothesis Testing

Pengaruh Smoker terhadap Charges dan BMI

Hipotesis charges

Hipotesis Null : $\mu_1 - \mu_2 = 0$, charges orang merokok sama dengan orang tidak merokok

Hipotesis Alternative : $\mu_1 - \mu_2 ! = 0$, Charges orang merokok tidak sama dengan orang tidak merokok

Hipotesis BMI

Hipotesis Null : $\mu_1 - \mu_2 = 0$, perolehan BMI orang merokok sama dengan orang tidak merokok

Hipotesis Alternative : $\mu_1 - \mu_2 ! = 0$, perolehan BMI orang merokok tidak tidak sama dengan orang yang merokok

Result #1

```
t=46.665, p=0.000
```

Tolak H0 karena p-value < alpha (0.05)), maka charges dengan tipe yes (orang merokok) tidak sama dengan no (orang tidak merokok).

Result #2

```
t=0.137, p=0.891
```

Gagal Tolak H0 (terima H0) karena p-value > alpha (0.05), maka belum cukup bukti untuk mengatakan bahwa BMI orang yang merokok tidak sama dengan orang yang tidak merokok

Conclusion

Conclusion

- Baik secara peluang atau perhitungan biasa jumlah laki-laki yang merokok pada data insurance tersebut lebih besar dibandingkan wanita
- Secara visualisasi orang yang merokok, mayoritas lebih besar biaya charges nya dibandingkan orang yang tidak merokok
- Eksptasi region berdasarkan biaya charge-nya paling besar berada di southeast
- Persebaran bmi di range 20-40 dan charges 0-30.000 lebih banyak terlihat dari luasannya
- Bahwa persebaran age di range 10-70 dan charges 0-10.000 lebih banyak terlihat dari luasannya dari kontur join distribusinya
- Bahwa persebaran jumlah anak di range 0-3 dan charges 0-20.000 lebih banyak terlihat dari luasannya dari kontur join distribusinya
- Hipotesis charges dengan tipe yes (orang merokok) tidak sama dengan no (orang tidak merokok) terbukti
- Hipotesis BMI orang yang merokok tidak sama dengan orang yang tidak merokok belum terbukti (kurang cukup bukti)

Notes

- Masih banyak hipotesis yang perlu diuji dari hasil probability maupun visualisasi, seperti (apakah charges orang merokok lebih besar dari orang yang tidak merokok).

- dsb..