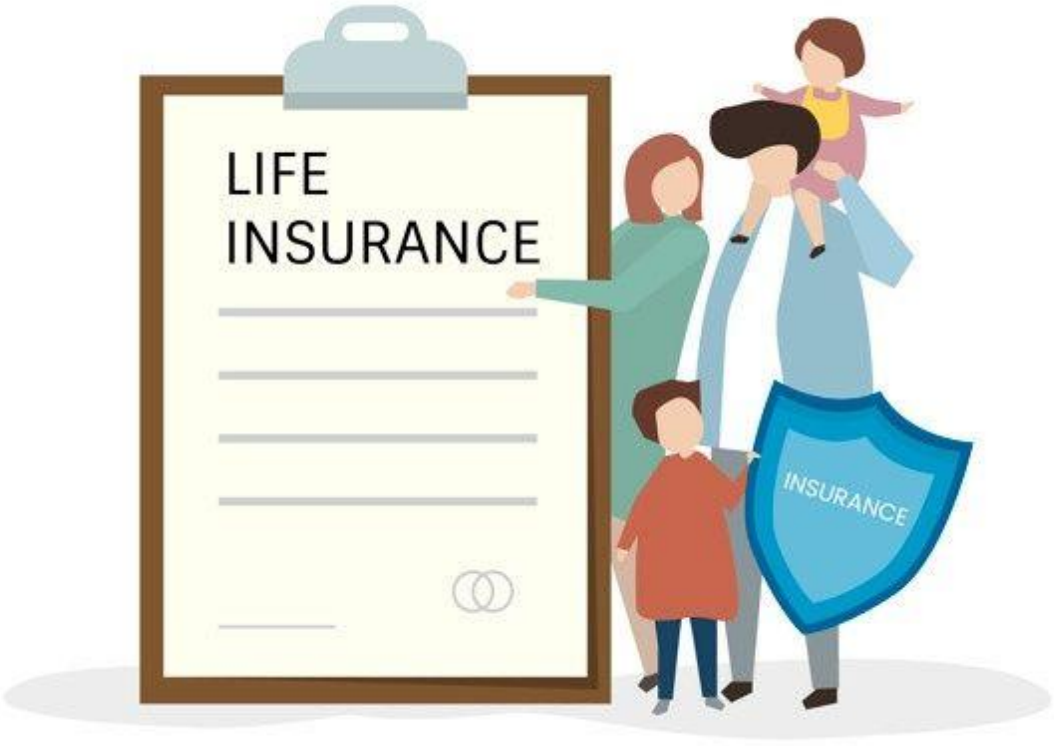
INSURANCE DATA REPORTING AND ANALYSIS



Arranged by:

Assita Multi Hasna

1. Introduction

Asuransi kesehatan adalah salah satu jenis asuransi yang memberikan jaminan kepada penggunanya untuk mengganti sebagian atau seluruh biaya perawatan atas terjadinya resiko kesehatan atau penyakit. Pengguna asuransi diwajibkan untuk membayar premi kepada perusahaan asuransi agar bisa mendapatkan manfaat asuransi kesehatan. Penentuan besaran nilai premi yang harus dibayarkan menjadi hal yang harus diperhatikan oleh perusahaan asuransi dikarenakan banyaknya faktor yang mempengaruhi dan meningkatkan faktor resiko kesehatan pengguna.

Sehubungan dengan project final kelas Probability, maka dengan ini saya akan menganalisa variable-variabel yang memiliki hubungan dengan tagihan kesehatan yang diterima oleh setiap pengguna. Dengan menggunakan dataset "Insurance.csv" yang berisi beberapa data personal pengguna seperti umur, gender, tempat tinggal pengguna, banyak anak tertanggung asuransi, nilai bmi dan keadaan pengguna sebagai perokok atau non perokok. Untuk selanjutnya akan digunakan python sebagai *tools* untuk membantu menganalisis dataset.

Berikut deskripsi singkat terkait informasi kolom pada dataset Insurance.csv:

- Age → Umur dari pengguna atau penerima manfaat asuransi kesehatan
- Sex → Jenis kelamin dari pengguna (Female, Male)
- BMI → Body mass index, ukuran yang digunakan untuk menunjukan kategori berat badan seseorang. Idealnya BMI seseorang adalah 18.5 sampai 24.9
- Children → Jumlah anak yang tertanggung
- Smoker → Pengguna sebagai perokok atau non perokok
- Region → Tempat tinggal pengguna yang tersebar di area northeast, southeast, southwest, northwest.
- Charges → Tagihan biaya kesehaatn yang harus dibayarkan kepada perusahaan asuransi.

2. Question and Answer

- Menampilkan dataset “Insurance.csv”

```
insurance = pd.read_csv("insurance.csv")
insurance
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

```
insurance.describe()
```

	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

- **Langkah #1 - Analisa Descriptive Statistic**

1) Berapa rata-rata umur pada data tersebut?

```
insurance.mean()["age"]
```

```
39.20702541106129
```

Rata-rata umur pada dataset insurance adalah 39.20702541106129

2) Berapa rata-rata nilai BMI dari yang merokok?

```
insurance.groupby(["smoker"]).mean()["bmi"]
```

```
smoker
no     30.651795
yes    30.708449
Name: bmi, dtype: float64
```

Rata-rata nilai BMI dari yang merokok adalah 30.708449

3) Apakah variansi dari data charges perokok dan non perokok sama?

```
insurance.groupby(["smoker"]).var()["charges"]
```

```
smoker
no     3.592542e+07
yes    1.332073e+08
Name: charges, dtype: float64
```

Variansi dari data charges perokok dan non perokok tidak sama.

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

```
insurance.groupby(["smoker", "sex"]).mean()["age"]
```

```
smoker  sex
no      female    39.691042
        male      39.061896
yes     female    38.608696
        male      38.446541
Name: age, dtype: float64
```

Rata-rata umur perempuan dan laki-laki yang merokok hampir sama yaitu untuk perempuan 38.608696 dan untuk laki-laki 38.44654.

5) Mana yang lebih tinggi, rata-rata tagihan kesehatan perokok atau non merokok?

```
insurance.groupby(["smoker"]).mean()["charges"]
```

```
smoker
no      8434.268298
yes    32050.231832
Name: charges, dtype: float64
```

Rata-rata tagihan kesehatan perokok lebih tinggi yaitu 32050.231832

➔ Dataset “Insurance.csv” diambil dari 1338 orang yang memiliki rata-rata umur 39.20702541106129 tahun. Rata-rata nilai BMI perokok adalah 30.708449 , tidak jauh berbeda dengan rata-rata nilai BMI keseluruhan yaitu 30.663397. Tetapi lebih tinggi dibandingkan dengan nilai BMI ideal yaitu 18.5 sampai 24.9.

➔ Variansi dari data charges atau tagihan perokok dan non perokok tidak sama. Tagihan perokok 1.332073e+08 lebih tinggi dibandingkan tagihan non perokok yaitu 3.592542e+07. Kemudian, untuk rata-rata nilai tagihan perokok yaitu sebesar 32050.231832 juga lebih tinggi dibandingkan dengan nilai tagihan non perokok 8434.268298.

➔ Rata-rata umur perempuan dan laki-laki yang merokok hampir sama yaitu untuk perempuan 38.608696 dan untuk laki-laki 38.44654. Tidak jauh berbeda dengan nilai rata-rata umur keseluruhan yaitu 39.20702541106129 tahun.

- **Langkah #2 - Analisa Variabel Kategorik (PMF)**

1) Gender mana yang memiliki tagihan paling tinggi?

```
insurance.groupby(["sex"]).max()["charges"]
```

```
sex
female    63770.42801
male      62592.87309
Name: charges, dtype: float64
```

Female atau perempuan yang memiliki tagihan paling tinggi yaitu sebesar 63770.42801

2) Apakah setiap region memiliki proporsi data banyak orang yang sama?

```
insurance["region"].value_counts()
```

```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

```
southeast = 364
southwest = 325
northwest = 324
northeast = 324

total_region = southeast + southwest + northeast + northwest

p_se = southeast/total_region
print(f"Proporsi banyaknya orang di region southeast adalah {p_se: .2f}")
p_sw = southwest/total_region
print(f"Proporsi banyaknya orang di region southwest adalah {p_sw: .2f}")
p_ne = northeast/total_region
print(f"Proporsi banyaknya orang di region northeast adalah {p_ne: .2f}")
p_nw = northwest/total_region
print(f"Proporsi banyaknya orang di region northwest adalah {p_nw: .2f}")
```

Proporsi banyaknya orang di region southeast adalah 0.27

Proporsi banyaknya orang di region southwest adalah 0.24

Proporsi banyaknya orang di region northeast adalah 0.24

Proporsi banyaknya orang di region northwest adalah 0.24

Tidak semua region memiliki proporsi data banyaknya orang yang sama.

3) Mana yang lebih tinggi proporsi perokok atau non perokok?

```
insurance["smoker"].value_counts()
```

```
no     1064
yes     274
Name: smoker, dtype: int64
```

```
no = 1064
yes = 274

total_smoker = no + yes

p_no = no/total_smoker
print(f"Proporsi non perokok adalah {p_no: .2f}")
p_yes = yes/total_smoker
print(f"Proporsi perokok adalah {p_yes: .2f}")
```

```
Proporsi non perokok adalah  0.80
Proporsi perokok adalah  0.20
```

Non perokok memiliki proporsi yang lebih tinggi yaitu 0.80

4) Berapa peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok?

```
summary_data(data_file_name="insurance.csv")
```

```
Hasil Data Operasi      no    yes
--------------------  ----  -----
male                   517    159
female                 547    115
```

```
n_perokok = 159 + 115
n_female_perokok = 115
peluang_female_perokok = n_female_perokok/n_perokok

print(f"peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok adalah : {peluang_female_perokok:.2f}")
```

```
peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok adalah : 0.42
```

Peluangnya sebesar 0.42

5) Berapa peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok?

```
n_perokok = 159 + 115
n_male_perokok = 159
peluang_male_perokok = n_male_perokok/n_perokok

print(f"peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok adalah : {peluang_male_perokok:.2f}")
```

```
peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok adalah : 0.58
```

Peluangnya sebesar 0.58

- ➔ Dari dataset “Insurance.csv” dapat diketahui bahwa gender female atau perempuan adalah yang memiliki tagihan paling tinggi yaitu sebesar 63770.42801 sedangkan male atau laki-laki memiliki tagihan sebesar 62592.87309, tidak berbeda jauh dengan tagihan tertinggi.
- ➔ Tidak semua region memiliki proporsi banyaknya data orang yang sama. Ada tiga region yang memiliki proporsi data banyak orang yang sama yaitu southwest, northeast dan northwest dengan 0.24% disetiap regionnya. Berbeda dengan yang lain, southeast memiliki proporsi data sebesar yaitu 0 0.27%.
- ➔ Non perokok memiliki proporsi yang lebih tinggi yaitu 0.80. Sedangkan proporsi perokok hanya sebesar 0.20.
- ➔ Peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok yaitu 0.58. Sedangkan peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok yaitu 0.42. Jadi peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok itu lebih besar dibandingkan dengan perempuan.

- **Langkah #3 - Analisa Variabel Kontinu (CDF)**

1) Mencari peluang besar tagihan berdasarkan BMI. Peluang seorang dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 30000

```
insurance[(insurance.bmi>25)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.90	0	yes	southwest	16884.9240
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.00	3	no	southeast	4449.4620
4	32	male	28.88	0	no	northwest	3866.8552
5	31	female	25.74	0	no	southeast	3756.6216
...	...	...	...	...	...	...	...
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

1091 rows × 7 columns

```
insurance[(insurance.bmi>25)&(insurance.charges>30000)]
```

	age	sex	bmi	children	smoker	region	charges
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
29	31	male	36.300	2	yes	southwest	38711.00000
30	22	male	35.600	0	yes	southwest	35585.57600
...	...	...	...	...	...	...	...
1301	62	male	30.875	3	yes	northwest	46718.16325
1303	43	male	27.800	0	yes	southwest	37829.72420
1308	25	female	30.200	0	yes	southwest	33900.65300
1313	19	female	34.700	2	yes	southwest	36397.57600
1323	42	female	40.370	2	yes	southeast	43896.37630

159 rows × 7 columns

```
total_data_bmi = 1091
soal_1_a = 159

p_soal_1_a = soal_1_a/total_data_bmi

print(f"peluang seorang dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 30000 adalah : {p_soal_1_a:.2f}")
```

```
peluang seorang dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 30000 adalah : 0.15
```

Peluang seorang dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 30000 adalah 0.15

2) Mencari kemungkin terjadi, seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700.

```
insurance[(insurance.smoker == "yes")&(insurance.bmi>25)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1308	25	female	30.200	0	yes	southwest	33900.65300
1313	19	female	34.700	2	yes	southwest	36397.57600
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

219 rows × 7 columns

```
insurance[(insurance.smoker == "yes")&(insurance.bmi>25)&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1308	25	female	30.200	0	yes	southwest	33900.65300
1313	19	female	34.700	2	yes	southwest	36397.57600
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

215 rows × 7 columns

```
total_data_bmiSmoker = 219
soal_2 = 215

p_soal_2 = soal_2/total_data_bmiSmoker

print(f"peluang seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700 adalah : {p_soal_2:.2f}")
```

```
peluang seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700 adalah : 0.98
```

Peluang seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.700 adalah 0.98

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

```
insurance[(insurance.smoker == "yes")]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1313	19	female	34.700	2	yes	southwest	36397.57600
1314	30	female	23.655	3	yes	northwest	18765.87545
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

274 rows × 7 columns

```
insurance[(insurance.smoker == "yes")&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1313	19	female	34.700	2	yes	southwest	36397.57600
1314	30	female	23.655	3	yes	northwest	18765.87545
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

254 rows × 7 columns

```
total_data_perokok = 274
soal_3 = 254

p_soal_3 = soal_3/total_data_perokok

print(f"peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia perokok adalah : {p_soal_3:.2f}")
```

```
peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia perokok adalah : 0.93
```

Peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia perokok adalah 0.93

4) Mana yang lebih mungkin terjadi

a. Seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau

```
insurance[(insurance.bmi>25)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.90	0	yes	southwest	16884.9240
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.00	3	no	southeast	4449.4620
4	32	male	28.88	0	no	northwest	3866.8552
5	31	female	25.74	0	no	southeast	3756.6216
...	...	...	...	...	...	...	...
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

1091 rows × 7 columns

```
insurance[(insurance.bmi>25)&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
9	60	female	25.840	0	no	northwest	28923.13692
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
...	...	...	...	...	...	...	...
1313	19	female	34.700	2	yes	southwest	36397.57600
1318	35	male	39.710	4	no	northeast	19496.71917
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

283 rows × 7 columns

```
total_data_atas25 = 1091
soal_4_a = 283

p_soal_4_a = soal_4_a/total_data_atas25

print(f"peluang seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : {p_soal_4_a:.2f}")
```

```
peluang seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.26
```

Peluang seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.26

b. Seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k

```
insurance[(insurance.bmi<25)]
```

	age	sex	bmi	children	smoker	region	charges
3	33	male	22.705	0	no	northwest	21984.47061
15	19	male	24.600	1	no	southwest	1837.23700
17	23	male	23.845	0	no	northeast	2395.17155
26	63	female	23.085	0	no	northeast	14451.83515
28	23	male	17.385	1	no	northwest	2775.19215
...	...	...	...	...	...	...	...
1304	42	male	24.605	2	yes	northeast	21259.37795
1306	29	female	21.850	0	yes	northeast	16115.30450
1314	30	female	23.655	3	yes	northwest	18765.87545
1316	19	female	20.600	0	no	southwest	1731.67700
1328	23	female	24.225	2	no	northeast	22395.74424

245 rows × 7 columns

```
soal_4b = insurance[(insurance.bmi<25)&(insurance.charges>16700)]
insurance[(insurance.bmi<25)&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
3	33	male	22.705	0	no	northwest	21984.47061
58	53	female	22.880	1	yes	southeast	23244.79020
62	64	male	24.700	1	no	northwest	30166.61817
69	28	male	23.980	3	yes	southeast	17663.14420
85	45	male	22.895	2	yes	northwest	21098.55405
98	56	male	19.950	0	yes	northeast	22412.64850
128	32	female	17.765	2	yes	northwest	32734.18630
140	34	male	22.420	2	no	northeast	27375.90478
153	42	female	23.370	0	yes	northeast	19964.74630
156	48	male	24.420	0	yes	southeast	21223.67580
219	24	female	23.210	0	no	southeast	25081.76784
224	42	male	24.640	0	yes	southeast	19515.54160
235	40	female	22.220	2	yes	southeast	19444.26580

```
print(len(soal_4b))
```

```
51
```

```
total_data_bwh25 = 245
soal_4_b = 51

p_soal_4_b = soal_4_b/total_data_bwh25

print(f"peluang seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : {p_soal_4_b:.2f}")
```

```
peluang seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.21
```

Peluang seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.21. Yang lebih mungkin terjadi adalah seseorang dengan BMI

diatas 25 mendapatkan tagihan kesehatan diatas 16.7k. Karena peluangnya lebih besar yaitu sebesar 0.26

5) Mana yang lebih mungkin terjadi

a. Seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k, atau

```
insurance[(insurance.smoker == "yes")&(insurance.bmi>25)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1308	25	female	30.200	0	yes	southwest	33900.65300
1313	19	female	34.700	2	yes	southwest	36397.57600
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

219 rows × 7 columns

```
insurance[(insurance.smoker == "yes")&(insurance.bmi>25)&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
11	62	female	26.290	0	yes	southeast	27808.72510
14	27	male	42.130	0	yes	southeast	39611.75770
19	30	male	35.300	0	yes	southwest	36837.46700
23	34	female	31.920	1	yes	northeast	37701.87680
...	...	...	...	...	...	...	...
1308	25	female	30.200	0	yes	southwest	33900.65300
1313	19	female	34.700	2	yes	southwest	36397.57600
1321	62	male	26.695	0	yes	northeast	28101.33305
1323	42	female	40.370	2	yes	southeast	43896.37630
1337	61	female	29.070	0	yes	northwest	29141.36030

215 rows × 7 columns

```
total_data_bmiSmoker = 219
soal_5_a = 215

p_soal_5_a = soal_5_a/total_data_bmiSmoker

print(f"peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : {p_soal_5_a:.2f}")
```

```
peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.98
```

Peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.98

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

```
insurance[(insurance.smoker == "no")&(insurance.bmi>25)]
```

	age	sex	bmi	children	smoker	region	charges
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.00	3	no	southeast	4449.4620
4	32	male	28.88	0	no	northwest	3866.8552
5	31	female	25.74	0	no	southeast	3756.6216
6	46	female	33.44	1	no	southeast	8240.5896
...	...	...	...	...	...	...	...
1332	52	female	44.70	3	no	southwest	11411.6850
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450

872 rows × 7 columns

```
insurance[(insurance.smoker == "no")&(insurance.bmi>25)&(insurance.charges>16700)]
```

	age	sex	bmi	children	smoker	region	charges
9	60	female	25.840	0	no	northwest	28923.13692
45	55	male	37.300	0	no	southwest	20630.28351
102	18	female	30.115	0	no	northeast	21344.84670
115	60	male	28.595	0	no	northeast	30259.99556
138	54	female	31.900	3	no	southeast	27322.73386
...	...	...	...	...	...	...	...
1195	19	female	27.930	3	no	northwest	18838.70366
1206	59	female	34.800	2	no	southwest	36910.60803
1211	39	male	34.100	2	no	southeast	23563.01618
1258	55	male	37.715	3	no	northwest	30063.58055
1318	35	male	39.710	4	no	northeast	19496.71917

68 rows × 7 columns

```
total_data_bmiNon = 872
soal_5_b = 68

p_soal_5_b = soal_5_b/total_data_bmiNon

print(f"peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : {p_soal_5_b:.2f}")
```

```
peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.08
```

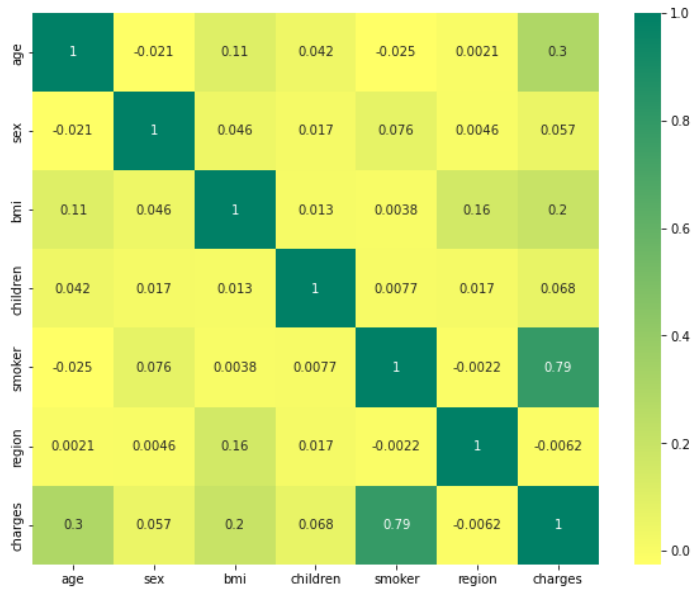
Peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.08

Yang lebih mungkin terjadi adalah peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k yaitu sebesar 0.98

- ➔ Secara acak dicari peluang seorang dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 30k adalah 0.15. Dapat disimpulkan bahwa sekitar 15% seseorang yang memiliki BMI diatas 25 mempunyai besaran tagihan diatas 30k.
- ➔ Peluang seorang perokok dengan BMI diatas 25 akan mendapatkan tagihan kesehatan di atas 16.7k adalah 0.98. Dapat disimpulkan bahwa 98% dari sesorang yang merokok dan mempunyai BMI diatas 25 mempunyai tagihan tagihan diatas 16.7k
- ➔ Peluang seseorang acak tagihan kesehatannya diatas 16.7k diketahui dia perokok adalah 0.93. Bisa disimpulkan bahwa 93% yang memiliki tagihan diatas 16.7k adalah seorang perokok.
- ➔ Peluang seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.26. Peluang seseorang dengan BMI dibawah 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.21. Yang lebih mungkin terjadi adalah seseorang dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k. Karena peluangnya lebih besar yaitu sebesar 0.26 atau 26%.
- ➔ Peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah 0.98. Peluang seseorang non perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k adalah : 0.08. Yang lebih mungkin terjadi adalah peluang seseorang perokok dengan BMI diatas 25 mendapatkan tagihan kesehatan diatas 16.7k yaitu sebesar 0.98 atau 98%.

- Langkah #4 - Analisa Korelasi Variabel

```
corr = insurance.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, cmap = 'summer_r')
```



Berdasarkan table korelasi heatmap, dapat disimpulkan bahwa:

1) Charges memiliki nilai korelasi 0.79 terhadap smoker. Ini artinya Charges atau tagihan mempunyai hubungan korelasi positif yang kuat terhadap smoker atau faktor merokok tidaknya pengguna asuransi.
2) Charges memiliki nilai korelasi 0.2 terhadap bmi. Ini artinya Charges atau tagihan mempunyai hubungan korelasi positif yang lemah terhadap bmi atau faktor kategori berat badan pengguna asuransi.
3) Charges memiliki nilai korelasi 0.3 terhadap age(umur). Ini artinya Charges atau tagihan mempunyai hubungan korelasi positif yang lemah terhadap age atau faktor umur pengguna asuransi.
4) Faktor sex, children dan region hampir tidak memiliki korelasi dengan charges (tagihan).

- **Langkah #5 - Pengujian Hipotesis**

1) Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok

```
Ho = "Tagihan perokok dan non perokok sama"
Ha = "Tagihan perokok dan non perokok tidak sama"

nonPerokok = insurance[insurance["smoker"] == 0]
perokok = insurance[insurance["smoker"] == 1]
tagihan_nonPerokok = nonPerokok["charges"]
tagihan_perokok = perokok["charges"]

print(f'Jumlah perokok adalah : {perokok.shape[0]}')
print(f'Variance tagihan perokok adalah : {np.var(tagihan_perokok)}')
print(f'Jumlah non perokok adalah : {nonPerokok.shape[0]}')
print(f'Variance tagihan non perokok adalah: {np.var(tagihan_nonPerokok)}')
```

```
Jumlah perokok adalah : 274
Variance tagihan perokok adalah : 132721153.13625304
Jumlah non perokok adalah : 1064
Variance tagihan non perokok adalah: 35891656.00316425
```

```
from scipy.stats import ttest_ind

t_statistic, p_value = ttest_ind(tagihan_perokok, tagihan_nonPerokok, equal_var=False)
print(f't_statistic: {t_statistic}\np_value: {p_value}')
```

```
t_statistic: 32.751887766341824
p_value: 5.88946444671698e-103
```

Ho = "Tagihan perokok dan non perokok sama"

Ha = "Tagihan perokok dan non perokok tidak sama"

Level signifikansi : 0.05

p-value : 5.88946444671698e-103

Dari hasil diatas dapat diketahui nilai dari p-value 5.88946444671698e-103 lebih rendah dari level signifikansi yaitu 0.05, artinya tolak null hypothesis. Dalam hal ini menunjukan bahwa tagihan yang dibayar oleh perokok dan non perokok tidak sama. Perokok membayar tagihan asuransi lebih tinggi dibandingkan dengan non perokok.

2) Tagihan kesehatan dengan BMI diatas 25 lebih tinggi daripada tagihan kesehatan dengan BMI dibawah 25

```
Ho = "Tagihan kesehatan BMI diatas 25 dan dibawah 25 sama"
Ha = "Tagihan kesehatan BMI diatas 25 dan dibawah 25 tidak sama"

bmi_lebih_25 = insurance[insurance["bmi"] > 25]
bmi_kurang_25 = insurance[insurance["bmi"] < 25]
tagihan_lebih_25 = bmi_lebih_25["charges"]
tagihan_kurang_25 = bmi_kurang_25["charges"]

print(f'Jumlah bmi lebih dari 25 adalah : {bmi_lebih_25.shape[0]}')
print(f'Variance tagihan bmi lebih dari 25 adalah : {np.var(tagihan_lebih_25)}')
print(f'Jumlah bmi kurang dari adalah : {bmi_kurang_25.shape[0]}')
print(f'Variance tagihan bmi kurang dari adalah: {np.var(tagihan_kurang_25)}')
```

```
Jumlah bmi lebih dari 25 adalah : 1091
Variance tagihan bmi lebih dari 25 adalah : 164579189.5213265
Jumlah bmi kurang dari adalah : 245
Variance tagihan bmi kurang dari adalah: 56326859.63068615
```

```
from scipy.stats import ttest_ind

t_statistic, p_value = ttest_ind(tagihan_lebih_25, tagihan_kurang_25, equal_var=False)
print(f't_statistic: {t_statistic}\np_value: {p_value}')
```

```
t_statistic: 5.929878344096734
p_value: 5.080897303161378e-09
```

Ho = "Tagihan kesehatan BMI diatas 25 dan dibawah 25 sama"

Ha = "Tagihan kesehatan BMI diatas 25 dan dibawah 25 tidak sama"

Level signifikansi : 0.05

p-value : 0.000018

Dari hasil diatas dapat diketahui nilai dari p-value 0.000018 lebih rendah dari level signifikansi yaitu 0.05, artinya tolak null hypothesis. Dalam hal ini menunjukan bahwa tagihan kesehatan BMI diatas 25 dan dibawah 25 tidak sama. Pengguna asuransi yang memiliki BMI diatas 25 membayar tagihan asuransi lebih tinggi dibandingkan dengan pengguna asuransi yang memiliki BMI dibawah 25.

3) BMI laki-laki dan perempuan sama

```
Ho = "BMI laki-laki dan perempuan sama"
Ha = "BMI laki-laki dan perempuan tidak sama"

laki = insurance[insurance['sex'] == 1]
perempuan = insurance[insurance['sex'] == 0]
bmi_laki = laki['bmi']
bmi_perempuan = perempuan['bmi']

print(f'Jumlah laki-laki adalah : {laki.shape[0]}')
print(f'Variance bmi laki-laki adalah : {np.var(bmi_laki)}')
print(f'Jumlah perempuan adalah: {perempuan.shape[0]}')
print(f'Variance bmi perempuan adalah: {np.var(bmi_perempuan)}')
```

```
Jumlah laki-laki adalah : 676
Variance bmi laki-laki adalah : 37.6491607363954
Jumlah perempuan adalah: 662
Variance bmi perempuan adalah: 36.49917703379856
```

```
from scipy.stats import ttest_ind

t_statistic, p_value = ttest_ind(bmi_laki, bmi_perempuan, equal_var=False)
print(f't_statistic: {t_statistic}\np_value: {p_value}')
```

```
t_statistic: 1.697027933124022
p_value: 0.08992430667834876
```

Ho = "BMI laki-laki dan perempuan sama"

Ha = "BMI laki-laki dan perempuan tidak sama"

Level signifikansi : 0.05

p-value : 0.08992430667834876

Dari hasil diatas dapat diketahui nilai dari p-value 0.08992430667834876 lebih tinggi dari level signifikansi yaitu 0.05, artinya gagal tolak null hypothesis. Dalam hal ini menunjukan bahwa BMI laki-laki dan perempuan relatif sama. Berdasarkan data BMI dari laki-laki dan perempuan tidak memiliki perbedaan yang signifikan.

3. Kesimpulan

Berdasarkan analisis dataset “Insurance.csv” dapat disimpulkan bahwa faktor smoker atau kebiasaan merokok pengguna asuransi memiliki korelasi yang kuat terhadap charges atau tagihan asuransi kepada para pengguna. Dalam hal ini perokok membayar tagihan yang lebih besar dibandingkan dengan non perokok. Selain smoker, faktor BMI dan age juga memiliki korelasi dengan charges namun korelasinya cenderung lemah. Pengguna asuransi dengan BMI diatas 25 juga memiliki indikasi akan membayar tagihan asuransi lebih tinggi dibandingkan dengan pengguna asuransi yang memiliki BMI dibawah 25. Berdasarkan gender, untuk BMI laki-laki dan perempuan relatif sama karena tidak memiliki perbedaan yang signifikan.

4. Saran Perbaikan
   - Sebaiknya lebih detail dalam pembahasan analisis hasil.
   - Memahami soal dengan lebih baik agar pertanyaan dapat dijawab dengan tepat.

5. Referensi

   - Gambar cover https://pin.it/5XEP0YB
   - https://medium.com/@ugursavci/complete-exploratory-data-analysis-using-python-9f685d67d1e4