Pacmann

Final Project

Probability Course - Sekolah Data Pacmann

Outline

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

Introduction

Background

Tren orang yang mengidap obesitas setiap tahun terus meningkat, peluang mereka untuk terkena penyakit generatif dan de-generatif pun akan terus meningkat. Dengan gaya hidup yang buruk (konsumsi makanan cepat saji, jarang berolahraga) juga menjadi faktor meningkatnya peluang seseorang terkena penyakit.

Pada analisis ini, kita akan melihat data pengguna asuransi berdasarkan gaya hidup dan juga kondisi fisik. Kita akan melihat bagaimana gaya hidup yang kita jalani mempengaruhi pengeluaran kita untuk premi asuransi.

Harapannya, kita dapat mengetahui kondisi masyarakat dan mendapatkan pelajaran dan langkah yang tepat di masa depan.

Dataset

Dataset

Features	Description
Age	Usia dari pengguna asuransi
Sex	Jenis kelamin
BMI	*Body Mass Index,* ukuran tubuh berdasarkan berat dan tinggi badan
Children	Jumlah anak yang dicover asuransi
Smoker	*Yes or No,* mengindikasikan apabila pengguna asuransi perokok atau non-perokok
Region	Daerah tempat tinggal (Northeast, Southeast, Southwest, Northwest)
Charges	Premi asuransi yang perlu dibayarkan

Descriptive Statistics Analysis

Mean of Age

```
# Rata-rata umur data
avg_age = np.round(df['age'].mean(),2)
print(f"Rata-rata umur data: {avg_age}")
```

```
Rata-rata umur data: 39.21
```

Dari 1.338 pengguna asuransi pada data yang kita punya, rata-rata dari dataset kita adalah pengguna asuransi dengan usia 39 tahun.

Variance of Charges by Smokers and Non-Smokers

```
: # Variansi charges dari perokok dan non perokok
  df.groupby(['smoker'])['charges'].var()
```

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

Variansi atau ragam dari biaya premi pengguna yang merokok dan tidak merokok berbeda. Dari data di atas, terlihat bahwa pengguna asuransi yang merokok memiliki ragam data yang lebih tinggi pada biaya preminya dibandingkan pengguna yang tidak merokok.

Age Average between Gender by smokers or non-smokers

```
# Rata-rata umur perempuan dan laki-laki perokok
df.groupby(['sex','smoker'])['age'].mean()
```

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

Usia rata-rata pengguna asuransi yang merokok lebih rendah dibandingkan non-perokok baik untuk laki-laki dan perempuan.

Average of Charges by smokers and non-smokers

```
# Rata-rata charges dari perokok dan non perokok
df.groupby(['smoker'])['charges'].mean()
```

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

Rata-rata premi asuransi perokok lebih tinggi dibandingkan non perokok.

Average of BMI by smokers and non-smokers

```
# Rata-rata bmi dari perokok dan non perokok
df.groupby(['smoker'])['bmi'].mean()
```

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

Rata-rata *Body Mass Index* perokok sedikit lebih tinggi dibandingkan non-perokok.

Average of BMI of Men and Women

```
# Rata-rata bmi dari perempuan dan laki-laki
df.groupby(['sex'])['bmi'].mean()
```

```
sex
female    30.377749
male      30.943129
Name: bmi, dtype: float64
```

Rata-rata *Body Mass Index* laki-laki lebih tinggi dibandingkan wanita.

Analysis

Dari data-data yang dianalisis, kita dapat simpulkan beberapa poin:

1. Pengguna asuransi mayoritas diisi oleh kalangan dewasa dengan rata-rata 39 tahun.
2. Perokok memiliki BMI dan tagihan premi yang lebih tinggi dibanding non-perokok
3. Laki-laki memiliki BMI yang lebih tinggi dibanding perempuan.

Categorical Variables Analysis

Proportion of smokers and non smokers

```
# Proporsi perokok atau non perokok
df['smoker'].value_counts()
```

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

Proporsi perokok lebih rendah dibanding dengan non-perokok.

Proportion of Regions

```
# Proporsi data per region
df['region'].value_counts()
```

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

Southeast memiliki pengguna asuransi yang lebih tinggi dibandingkan region lain. Selain southeast, region lain memiliki proporsi yang kurang lebih sama

Probability Charges of regions

```
# Distribusi peluang tagihan tiap region
round(df.groupby(['region'])['charges'].sum()/df['charges'].sum(),3)
```

```
region
northeast    0.245
northwest    0.227
southeast    0.302
southwest    0.226
Name: charges, dtype: float64
```

Distribusi peluang tiap-tiap region. Southeast memiliki peluang distribusi yang paling tinggi dibandingkan region lain.

Charges based on gender

```
[ ]: # Gender dengan tagihan paling tinggi
df.groupby(['sex'])['charges'].mean()
```

```
[ ]: sex
female    12569.578844
male      13956.751178
Name: charges, dtype: float64
```

Berdasarkan gender, laki-laki memiliki rata-rata biaya premi asuransi yang lebih tinggi dibanding perempuan.

Peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok

```
#peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok?
p_perempuan_perokok = round((n_perempuan_perokok / n_perokok)*100,3)
print(f"peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok: {p_perempuan_perokok}%")
```

```
peluang seseorang tersebut adalah perempuan diketahui dia adalah perokok: 41.971%
```

Peluang seseorang adalah peremuan diketahui adalah seorang perokok dari data yang kita miliki adalah sebesar 41.9%.

Peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok

```
#peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok?
p_laki_perokok = round((n_laki_perokok / n_perokok)*100,3)
print(f"peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok: {p_laki_perokok}%")
```

```
peluang seseorang tersebut adalah laki-laki diketahui dia adalah perokok: 58.029%
```

Peluang seseorang adalah laki-laki diketahui adalah seorang perokok dari data yang kita miliki adalah sebesar 58%.

Analysis

- Proporsi perokok lebih rendah dibanding non-perokok
- Region southeast adalah region dengan pengguna asuransi terbanyak
- Peluang distribusi southeast lebih tinggi dibanding region lain
- Rata-rata premi asuransi laki-laki lebih tinggi dibanding perempuan
- Peluang seorang perempuan diketahui perokok sebesar 42%
- Peluang seorang laki-laki diketahui perokok sebesar 58%

Continuous Variables Analysis

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

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

```
smoker = df[df['smoker'] == 'yes']
smoker_bmi_25 = len(smoker[(smoker['bmi'] > 25) & (smoker['charges']>16700)])
charge_over = len(df[df['charges']>16700])

pmf = round(smoker_bmi_25/charge_over,2)
pmf
```

```
0.64
```

Peluang seorang perokok dengan BMI di atas 25 mendapatkan premi lebih dari 16.700 adalah 64%.

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

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

```
charge_over =  len(df[(df['smoker'] == 'yes') & (df['charges']>16700)])
smoker = len( df[df['smoker'] == 'yes'])
pmf = round(charge_over/ smoker,2)
pmf
```

```
0.93
```

Peluang seorang dengan premi asuransi lebih dari 16.700 diketahui seorang perokok adalah 93%.

Membandingkan peluang premi asuransi

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

```
charge_over =  len(df[df['charges']>16700])
bmi_25_atas = len(df[(df['bmi'] >25) &(df['charges'] > 16700)])
bmi_25_bawah = len(df[(df['bmi'] <25) &(df['charges'] > 16700)])

atas_25 = round(bmi_25_atas / charge_over,2)
bawah_25 = round(bmi_25_bawah / charge_over,2)
print(f"peluang seseorang dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: {atas_25*100}%")
print(f"peluang seseorang dengan BMI dibawah 25 mendapat tagihan kesehatan di atas 16.7k: {bawah_25*100}%")
```

```
peluang seseorang dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: 85.0%
peluang seseorang dengan BMI dibawah 25 mendapat tagihan kesehatan di atas 16.7k: 15.0%
```

Lebih tinggi peluang seseorang dengan BMI di atas 25 mendapatkan tagihan kesehatan di atas 16.700 (85%).

Membandingkan peluang premi asuransi perokok dan non perokok

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 diatas 16.7k

```
charge_over =  len(df[(df['charges']>16700) & (df['bmi']>25)])

perokok_bmi_25_atas = len(df[(df['bmi'] > 25) & (df['smoker'] == 'yes') & (df['charges']>16700)])
non_perokok_bmi_25_atas = len(df[(df['bmi'] > 25) & (df['smoker'] == 'no') & (df['charges']>16700)])

pmf_perokok_bmi_25_atas = round(perokok_bmi_25_atas / charge_over,2)
pmf_non_perokok_bmi_25_atas = round(non_perokok_bmi_25_atas / charge_over,2)
print(f"peluang seseorang perokok dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: {pmf_perokok_bmi_25_atas
print(f"peluang seseorang non-perokok dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: {pmf_non_perokok_bmi
```

```
peluang seseorang perokok dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: 76.0%
peluang seseorang non-perokok dengan BMI diatas 25 mendapat tagihan kesehatan di atas 16.7k: 24.0%
```

Peluang perokok dengan BMI di atas 25 mendapat premi asuransi di atas 16.700 lebih tinggi dibandingkan non-perokok (76%).

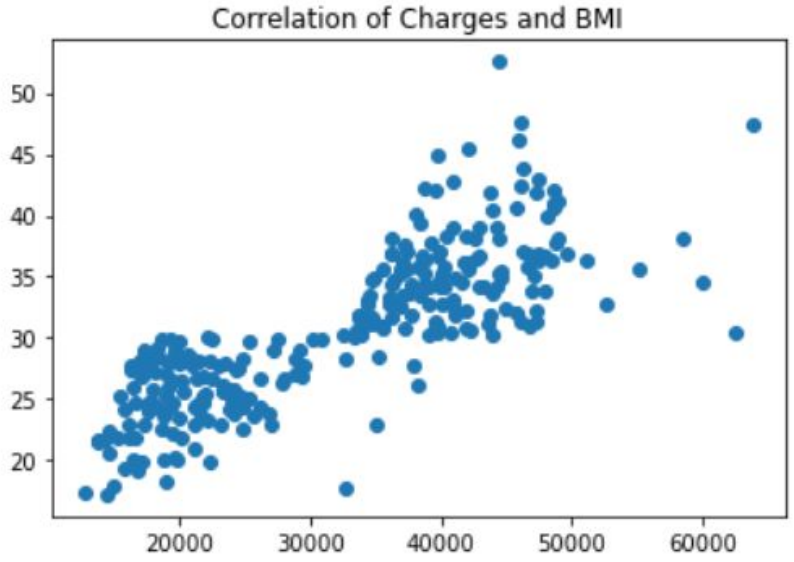
Variables Correlation

Correlation

```
corr_charges_bmi = np.corrcoef(smoker['charges'],smoker['bmi'])
corr_charges_bmi
```

```
array([[1.        , 0.80648061],
       [0.80648061, 1.        ]])
```

```
plt.scatter(smoker['charges'],smoker['bmi'])
plt.title('Correlation of Charges and BMI')
plt.show()
```



Korelasi antara premi asuransi dan Body Mass Index berkorelasi positif dan mendekati 1.

Hypothesis Testing

Tagihan premi perokok lebih tinggi daripada non perokok

Tagihan kesehatan perokok lebih tinggi daripada tagihan kesehatan non perokok

```
smoker = df[df['smoker'] == 'yes']
non_smoker = df[df['smoker'] == 'no']
a = 0.05
```

```
var_smoker = smoker['charges'].var()
var_non_smoker = non_smoker['charges'].var()
print('Variance Charges Smoker: %.4f'%(var_smoker))
print('Variance Charges Non-Smoker: %.4f'%(var_non_smoker))
```

```
Variance Charges Smoker: 133207311.2063
Variance Charges Non-Smoker: 35925420.4961
```

```
mean_smoker = smoker['charges'].mean()
print('Mean Charges Smoker: %.4f'%(mean_smoker))
mean_non_smoker = non_smoker['charges'].mean()
print('Mean Charges Non - Smoker: %.4f'%(mean_non_smoker))
```

```
Mean Charges Smoker: 32050.2318
Mean Charges Non - Smoker: 8434.2683
```

Tagihan premi perokok lebih tinggi daripada non perokok

$$H_0 : p_1 \geq p_2$$
$$H_a : p_1 < p_2$$

- p1: charges for smokers
- p2: charges for non smokers

```
stat, p = ttest_ind(smoker['charges'],non_smoker['charges'], equal_var=False,alternative = 'less')
print('Statistics: %.4f, p-value: %.4f'%(stat,p))
if p > a:
    print('Terima null hypothesis')
else:
    print('Tolak null hypothesis')
```

```
Statistics: 32.7519, p-value: 1.0000
Terima null hypothesis
```

P-value > alpha, menunjukan kita dapat menerima H0.
Kesimpulan: Tagihan premi perokok secara statistik signifikan lebih tinggi dibanding non-perokok

Tagihan Premi asuransi BMI diatas 25 lebih tinggi dibanding BMI dibawah 25

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

```
high_bmi = df[df['bmi'] > 25]
low_bmi = df[df['bmi'] < 25]
```

```
var_high_bmi = high_bmi['charges'].var()
print('Variance Charges of BMI > 25: %.4f'%(var_high_bmi))
var_low_bmi = low_bmi['charges'].var()
print('Variance Charges of BMI < 25: %.4f'%(var_low_bmi))
```

```
Variance Charges of BMI > 25: 164730179.6035
Variance Charges of BMI < 25: 56557707.4161
```

```
mean_high_bmi = high_bmi['charges'].mean()
print('Mean Charges of BMI > 25: %.4f'%(mean_high_bmi))
mean_low_bmi = low_bmi['charges'].mean()
print('Mean Charges of BMI < 25: %.4f'%(mean_low_bmi))
```

```
Mean Charges of BMI > 25: 13946.4760
Mean Charges of BMI < 25: 10282.2245
```

Tagihan Premi asuransi BMI diatas 25 lebih tinggi dibanding BMI dibawah 25

$$H_0 : p_1 \geq p_2$$
$$H_a : p_1 < p_2$$

- p1: charges for BMI > 25
- p2: charges for BMI < 25

```
stat, p = ttest_ind(high_bmi['charges'],low_bmi['charges'], equal_var=False,alternative = 'less')
print('Statistics: %.4f, p-value: %.4f'%(stat,p))
if p > a:
    print('Terima null hypothesis')
else:
    print('Tolak null hypothesis')
```

```
Statistics: 5.9299, p-value: 1.0000
Terima null hypothesis
```

P-value > alpha, menunjukan kita dapat menerima H0.
Kesimpulan: Tagihan premi bagi anggota dengan BMI > 25 secara statistik signifikan lebih tinggi dibanding BMI < 25

Tagihan asuransi laki-laki lebih tinggi dibanding perempuan

Tagihan kesehatan laki-laki lebih besar dari perempuan

```
laki = df[df['sex'] == 'male']
wanita = df[df['sex']=='female']
```

```
mean_laki = laki['charges'].mean()
print('Mean Charges Laki-laki: %.4f'%(mean_laki))
mean_wanita = wanita['charges'].mean()
print('Mean Charges Wanita: %.4f'%(mean_wanita))
```

```
Mean Charges Laki-laki: 13956.7512
Mean Charges Wanita: 12569.5788
```

```
var_laki = laki['charges'].var()
print('Variance Charges Laki-laki: %.4f'%(var_laki))
var_wanita = wanita['charges'].var()
print('Variance Charges Wanita: %.4f'%(var_wanita))
```

```
Variance Charges Laki-laki: 168247513.2882
Variance Charges Wanita: 123848048.2885
```

Tagihan asuransi laki-laki lebih tinggi dibanding perempuan

$$H_0 : p_1 \geq p_2$$
$$H_a : p_1 < p_2$$

- p1: charges for Laki - Laki
- p2: charges for Wanita

```
stat, p = ttest_ind(laki['charges'],wanita['charges'], equal_var=False,alternative = 'less')
print('Statistics: %.4f, p-value: %.4f'%(stat,p))
if p > a:
    print('Terima null hypothesis')
else:
    print('Tolak null hypothesis')
```

```
Statistics: 2.1009, p-value: 0.9821
Terima null hypothesis
```

P-value > alpha, menunjukan kita dapat menerima H0.
Kesimpulan: Tagihan laki-laki secara statistik signifikan lebih tinggi dibanding perempuan.

Conclusion

Conclusion

- Pengguna asuransi yang merokok memiliki peluang tagihan asuransi premi yang lebih tinggi dibanding non-perokok
- Pengguna asuransi dengan BMI > 25 memiliki peluang tagihan asuransi premi yang lebih tinggi dibanding pengguna asuransi BMI < 25
- *Body Mass Index* berkorelasi positif dan cenderung mendekati linier dengan tagihan asuransi
- Hindari merokok.

Reference

- T-test : Link
- CDF: Link
- Materi Pembelajaran Pacmann