Identitas Kelompok

Nomor Kelompok: 9

Identitas Anggota

Tuliskan identitas dari setiap anggota, meliputi NIM dan nama mahasiswa.

NIM Anggota-1: { 221112820 }, Nama Anggota-1: { Albert Putra Pratama Halawa }

NIM Anggota-2: { 221113406 }, Nama Anggota-2: { Jhon Kennedy Harefa }

NIM Anggota-3: { 221111265 }, Nama Anggota-3: { Octa Dana Rizky Lubis }

NIM Anggota-4: { 221113024 }, Nama Anggota-4: { Renaldi Aritonang }

Kontribusi Setiap Anggota

Tuliskan kontribusi dari setiap anggota di dalam pengerjaan soal UAS berbasis proyek ini.

Anggota-1: { menilai data, cleaning data, EDA, Visualisasi (distribusi nasabah, pola dan tren, order dan revenue), kesimpulan }

Anggota-2: { visualisasi persebaran Region dan kesimpulan }

Anggota-3: { analisis untuk BMI, Gender dan kesimpulan }

Anggota-4: { pertanyaan bisnis dan membuat kesimpulan }

Deskripsi Proyek

Pada UAS berbasis proyek ini, kelompok Anda harus melakukan seluruh proses analisis data, mulai dari mendefinisikan pertanyaan bisnis (pertanyaan analisis) yang ingin dijawab melalui proses analisis data hingga membuat kesimpulan dari hasil analisis (berupa visualisasi data untuk menjawab pertanyaan bisnis yang telah dibuat).

Untuk tipe A, lakukan analisis terhadap dataset insurance.csv yang berisi data transaksi asuransi. Lakukan analisis dengan tepat untuk mendapatkan insight/informasi sebanyak mungkin dari dataset tersebut. Penjelasan dari setiap kolom yang terdapat di dataset tersebut adalah sebagai berikut.

Tahap 1: Menentukan Pertanyaan Bisnis

Pada tahap ini, silahkan tentukan minimal 3 (tiga) pertanyaan bisnis yang akan dijawab berdasarkan dataset yang telah ditentukan di atas (analisis dan pahami data apa saja yang disimpan di masing-masing kolom)

1. Bagaimana dengan distribusi nasabah asuransi dalam dataset ini?

2. Berapa rata-rata jumlah anak yang dimiliki oleh nasabah?
3. Bagaimana persebaran wilayah domisili nasabah?
4. Tampilkan pola atau tren dalam premi yang dikenakan berdasarkan kombinasi faktor usia, BMI, dan status perokok?
5. bagaimana performa Order dan revenue perusahaan?

Link Project : https://drive.google.com/file/d/1a9K5dl1BvPOROv9EHwyBia7SdScRBR-x/view?usp=drive_link

Tahap 2: Menyiapkan Library

Sebelum memulai proyek, pastikan telah mengimpor semua library yang dibutuhkan untuk mengerjakan proyeknya

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Tahap 3: Mengumpulkan Data

Pada tahap ini, lakukan pengumpulan data dari dataset yang telah ditentukan di atas

```
data_insurance = pd.read_excel("./insurance.xlsx")
data_insurance.head()
```

	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

Tahap 4: Menilai Data

Pada tahap ini bisa dilakukan pengecekan tipe data, missing value, duplikasi data, keanehan pada nilai statistik, dan sebagainya

```
data_insurance.info()
```

```
<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   object 
 1   sex       1325 non-null   object 
 2   bmi       1338 non-null   float64
```

```
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1330 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(1), object(4)
memory usage: 73.3+ KB
```

```
data_insurance.isna().sum()
```

```
age          0
sex         13
bmi          0
children     0
smoker       0
region       8
charges      0
dtype: int64
```

```
print('Jumlah Duplikasi: ',data_insurance.duplicated().sum())
```

```
Jumlah Duplikasi:  1
```

```
data_insurance.describe()
```

```
                bmi     children       charges
count  1338.000000  1338.000000   1338.000000
mean     30.663397     1.094918  13270.422265
std       6.098187     1.205493  12110.011237
min      15.960000     0.000000   1121.873900
25%      26.296250     0.000000   4740.287150
50%      30.400000     1.000000   9382.033000
75%      34.693750     2.000000  16639.912515
max      53.130000     5.000000  63770.428010
```

Tahap 5: Membersihkan Data

Pada tahap ini bisa dilakukan penghapusan terhadap kolom yang tidak dibutuhkan, mengubah tipe data yang tidak sesuai, melakukan transformasi data, menambah kolom yang dibutuhkan, dan sebagainya

Menghapus Duplikasi Data

```
data_insurance.duplicated().sum()
```

```
1
```

```
data_insurance.drop_duplicates(inplace=True)
```

```
print('Jumlah Duplikasi: ',data_insurance.duplicated().sum())
```

```
Jumlah Duplikasi:  0
```

Missing Values

```
data_insurance.isna().sum()
```

```
age          0
sex         13
bmi          0
children     0
smoker       0
region       8
charges      0
dtype: int64
```

```
data_insurance[data_insurance.sex.isna()]
```

	age	sex	bmi	children	smoker	region	charges
19	30	NaN	35.300	0	yes	southwest	36837.46700
64	20	NaN	22.420	0	yes	northwest	14711.74380
136	19	NaN	34.100	0	no	southwest	1261.44200
227	58	NaN	41.910	0	no	southeast	24227.33724
273	50	NaN	27.455	1	no	northeast	9617.66245
331	52	NaN	27.360	0	yes	northwest	24393.62240
383	35	NaN	43.340	2	no	southeast	5846.91760
444	56	NaN	26.695	1	yes	northwest	26109.32905
490	19	NaN	32.900	0	no	southwest	1748.77400
563	50	NaN	44.770	1	no	southeast	9058.73030
610	47	NaN	29.370	1	no	southeast	8547.69130
639	56	NaN	33.660	4	no	southeast	12949.15540
680	21	NaN	17.400	1	no	southwest	2585.26900

```
data_insurance.dropna(inplace=True)
```

```
data_insurance.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

```
data_insurance.sex.value_counts()
```

```
sex
male      665
female    651
Name: count, dtype: int64
```

```
data_insurance.region.value_counts()
```

```
region
southeast    356
southwest    321
northeast    321
northwest    318
Name: count, dtype: int64
```

```
data_insurance.fillna(value=" ", inplace=True)

data_insurance.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Innacurate values

```
data_insurance.describe()
```

```
                bmi     children       charges
count  1316.000000  1316.000000   1316.000000
mean     30.661926     1.095745  13254.671317
std       6.089311     1.207057  12109.448424
min      15.960000     0.000000   1121.873900
25%      26.308750     0.000000   4733.635288
50%      30.400000     1.000000   9382.033000
75%      34.717500     2.000000  16604.302645
max      53.130000     5.000000  63770.428010
```

Memperbaiki Tipe Data

```
data_insurance['age'] = data_insurance['age'].apply(lambda x:
int(x.strip("'")) if isinstance(x, str) else x)

data_insurance.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1316 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1316 non-null   int64
 1   sex       1316 non-null   object
 2   bmi       1316 non-null   float64
 3   children  1316 non-null   int64
 4   smoker    1316 non-null   object
```

```
 5   region    1316 non-null   object
 6   charges   1316 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 82.2+ KB
```

Menambahkan Kolom Baru

```
data_insurance.insert(0, 'id', range(1, 1+len(data_insurance)))

data_insurance.head()
```

	id	age	sex	bmi	children	smoker	region	charges
0	1	19	female	27.900	0	yes	southwest	16884.92400
1	2	18	male	33.770	1	no	southeast	1725.55230
2	3	28	male	33.000	3	no	southeast	4449.46200
3	4	33	male	22.705	0	no	northwest	21984.47061
4	5	32	male	28.880	0	no	northwest	3866.85520

Tahap 6: Mengeksplorasi dan Menganalisis Data

Pada tahap ini bisa dilakukan pengelompokkan terhadap data berdasarkan kolom tertentu untuk dianalisis, mencari hubungan di antara kolom, melakukan filtrasi data, dan sebagainya

```
# Berikan penjelasan/kesimpulan untuk setiap analisis data yang telah dilakukan
data_insurance.sample(5)
```

	id	age	sex	bmi	children	smoker	region	charges
480	471	63	male	41.325	3	no	northwest	15555.18875
132	131	53	female	35.900	2	no	southwest	11163.56800
1122	1103	53	female	36.860	3	yes	northwest	46661.44240
153	151	42	female	23.370	0	yes	northeast	19964.74630
672	658	36	male	29.700	0	no	southeast	4399.73100

```
data_insurance.describe(include="all")
```

	id	age	sex	bmi	children	smoker	region	charges
count	1316.000000	1316.000000	1316	1316.000000	1316.000000	1316		
unique	NaN	NaN	2	NaN	NaN	2		
top	NaN	NaN	male	NaN	NaN	no		

```
freq            NaN          NaN   665          NaN          NaN
1049
mean     658.500000    39.224164   NaN    30.661926     1.095745
NaN
std      380.040787    14.057808   NaN     6.089311     1.207057
NaN
min        1.000000    18.000000   NaN    15.960000     0.000000
NaN
25%      329.750000    27.000000   NaN    26.308750     0.000000
NaN
50%      658.500000    39.000000   NaN    30.400000     1.000000
NaN
75%      987.250000    51.000000   NaN    34.717500     2.000000
NaN
max     1316.000000    64.000000   NaN    53.130000     5.000000
NaN

           region       charges
count        1316   1316.000000
unique          4           NaN
top     southeast           NaN
freq          356           NaN
mean          NaN  13254.671317
std           NaN  12109.448424
min           NaN   1121.873900
25%           NaN   4733.635288
50%           NaN   9382.033000
75%           NaN  16604.302645
max           NaN  63770.428010
```

```
# Set kolom id menjadi unik
data_insurance.id.is_unique
```

```
True
```

```
data_insurance.id.duplicated
```

```
<bound method Series.duplicated of 0          1
1          2
2          3
3          4
4          5
        ...
1333    1312
1334    1313
1335    1314
1336    1315
1337    1316
Name: id, Length: 1316, dtype: int64>
```

```
# Menambahkan kolom baru('age_group') untuk klasifikasi user
berdasarkan umur
data_insurance['age_group'] = data_insurance['age'].apply(lambda x:
"Teen" if x <=21 else ("Mature" if x > 60 else "Elderly"))

data_insurance.head()
```

```
   id  age     sex     bmi  children smoker     region      charges
age_group
0   1   19  female  27.900         0    yes  southwest  16884.92400
Teen
1   2   18    male  33.770         1     no  southeast   1725.55230
Teen
2   3   28    male  33.000         3     no  southeast   4449.46200
Elderly
3   4   33    male  22.705         0     no  northwest  21984.47061
Elderly
4   5   32    male  28.880         0     no  northwest   3866.85520
Elderly
```

```
data_insurance['age_group'].value_counts()
```

```
age_group
Elderly    1036
Teen        189
Mature       91
Name: count, dtype: int64
```

```
print("Jumlah berdasarkan jenis kelamin:")
print(data_insurance['sex'].value_counts())
print("\n")
```

```
Jumlah berdasarkan jenis kelamin:
sex
male      665
female    651
Name: count, dtype: int64
```

```
print("Average of BMI:")
print(data_insurance['bmi'].mean())
print("\n")
```

```
Average of BMI:
30.661926291793314
```

```
data_insurance.children.value_counts()
```

```
children
0    566
1    315
2    238
3    155
4     24
5     18
Name: count, dtype: int64
```

```
data_insurance.smoker.value_counts()
```

```
smoker
no     1049
yes     267
Name: count, dtype: int64
```

```
data_insurance.region.value_counts()
```

```
region
southeast    356
southwest    321
northeast    321
northwest    318
Name: count, dtype: int64
```

```
data_insurance.charges.describe()
```

```
count     1316.000000
mean     13254.671317
std      12109.448424
min       1121.873900
25%       4733.635288
50%       9382.033000
75%      16604.302645
max      63770.428010
Name: charges, dtype: float64
```

```
# menambahkan kolom (bmi_category) untuk klasifikasi kondisi berat
badan berdasarkan BMI
data_insurance['bmi_category'] = pd.cut(data_insurance['bmi'],
                                        bins=[-float('inf'), 18.5,
24.9, 29.9, float('inf')],
                                        labels=['underweight',
'normal', 'overweight', 'obese'])

bmi_category = data_insurance['bmi_category']

data_insurance.bmi_category.value_counts()
```

```
bmi_category
obese          705
```

```
overweight     372
normal         219
underweight     20
Name: count, dtype: int64
```

```
# Menambakan kolom (status) untuk klasifikasi status pernikahan user
data_insurance['status'] = data_insurance['children'].apply(lambda x:
"single" if x == 0 else "married")

status = data_insurance['status']

data_insurance.status.value_counts()
```

```
status
married    750
single     566
Name: count, dtype: int64
```

```
# klasifikasi untuk charge/premi user berdasarkan (charges)
data_insurance['charges_category'] = pd.cut(data_insurance['charges'],
                                            bins=[-float('inf'), 5000,
10000, 20000, 30000, float('inf')],
                                            labels=['very low', 'low',
'medium', 'high', 'very high'])

charges_category = data_insurance['charges_category']

print(data_insurance[['charges', 'charges_category']].head())
```

```
       charges charges_category
0  16884.92400           medium
1   1725.55230         very low
2   4449.46200         very low
3  21984.47061             high
4   3866.85520         very low
```

```
data_insurance.charges_category.value_counts()
```

```
charges_category
very low     354
medium       349
low          346
very high    159
high         108
Name: count, dtype: int64
```

```
# menghapus baris yang missing value
data_insurance.dropna(inplace=True)

data_insurance.isna().sum()
```

```
id                  0
age                 0
sex                 0
bmi                 0
children            0
smoker              0
region              0
charges             0
age_group           0
bmi_category        0
status              0
charges_category    0
dtype: int64
```

```
data_insurance.head()
```

```
   id  age     sex     bmi  children smoker     region      charges
age_group  \
0   1   19  female  27.900         0    yes  southwest  16884.92400
Teen
1   2   18    male  33.770         1     no  southeast   1725.55230
Teen
2   3   28    male  33.000         3     no  southeast   4449.46200
Elderly
3   4   33    male  22.705         0     no  northwest  21984.47061
Elderly
4   5   32    male  28.880         0     no  northwest   3866.85520
Elderly

  bmi_category   status charges_category
0   overweight   single           medium
1        obese  married         very low
2        obese  married         very low
3       normal   single             high
4   overweight   single         very low
```

Tahap 7: Memvisualisasikan Data

Pada tahap ini, sajikan informasi dalam format gambar/grafik untuk setiap pertanyaan bisnis

```
data_insurance.head()
```

```
   id  age     sex     bmi  children smoker     region      charges
age_group  \
0   1   19  female  27.900         0    yes  southwest  16884.92400
Teen
1   2   18    male  33.770         1     no  southeast   1725.55230
Teen
2   3   28    male  33.000         3     no  southeast   4449.46200
```

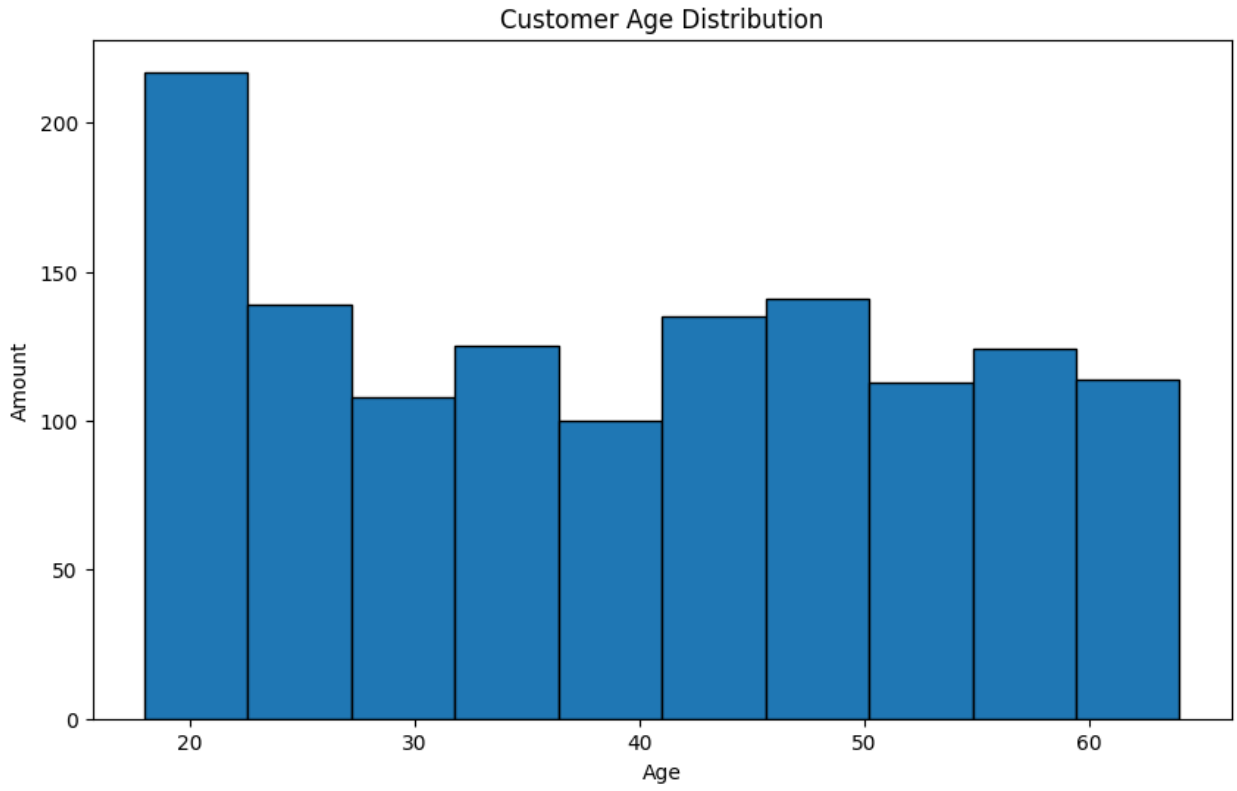
```
Elderly
3   4   33    male  22.705         0     no  northwest  21984.47061
Elderly
4   5   32    male  28.880         0     no  northwest   3866.85520
Elderly
```

	bmi_category	status	charges_category
0	overweight	single	medium
1	obese	married	very low
2	obese	married	very low
3	normal	single	high
4	overweight	single	very low

1. Distribusi Nasabah

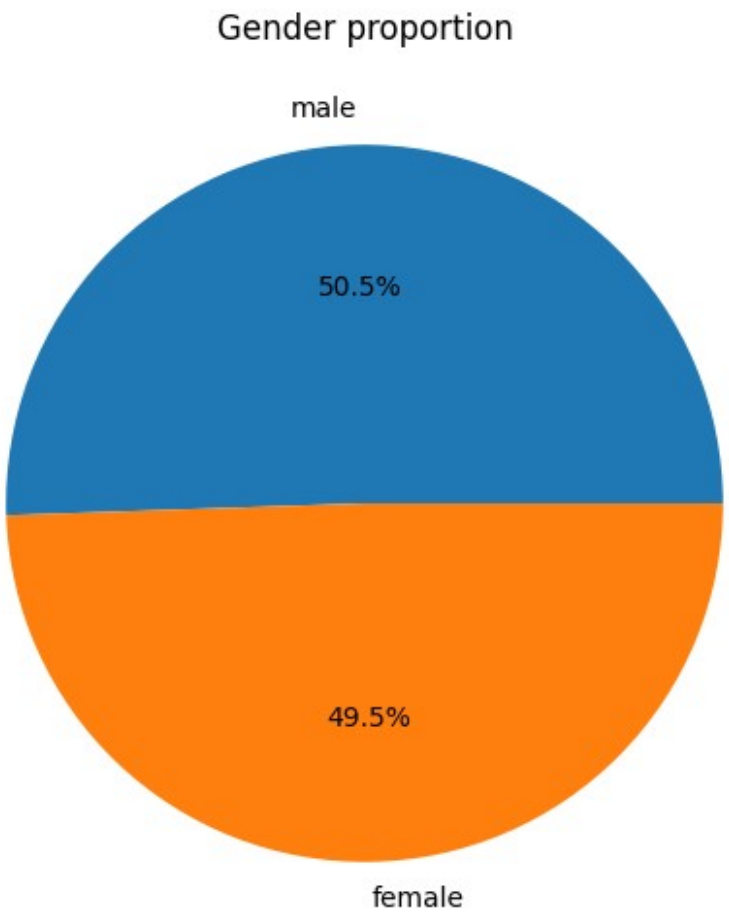
age status

```
plt.figure(figsize=(10,6))
plt.hist(data_insurance['age'], bins=10, edgecolor='black')
plt.title('Customer Age Distribution')
plt.xlabel('Age')
plt.ylabel('Amount')
plt.show()
```



Sex

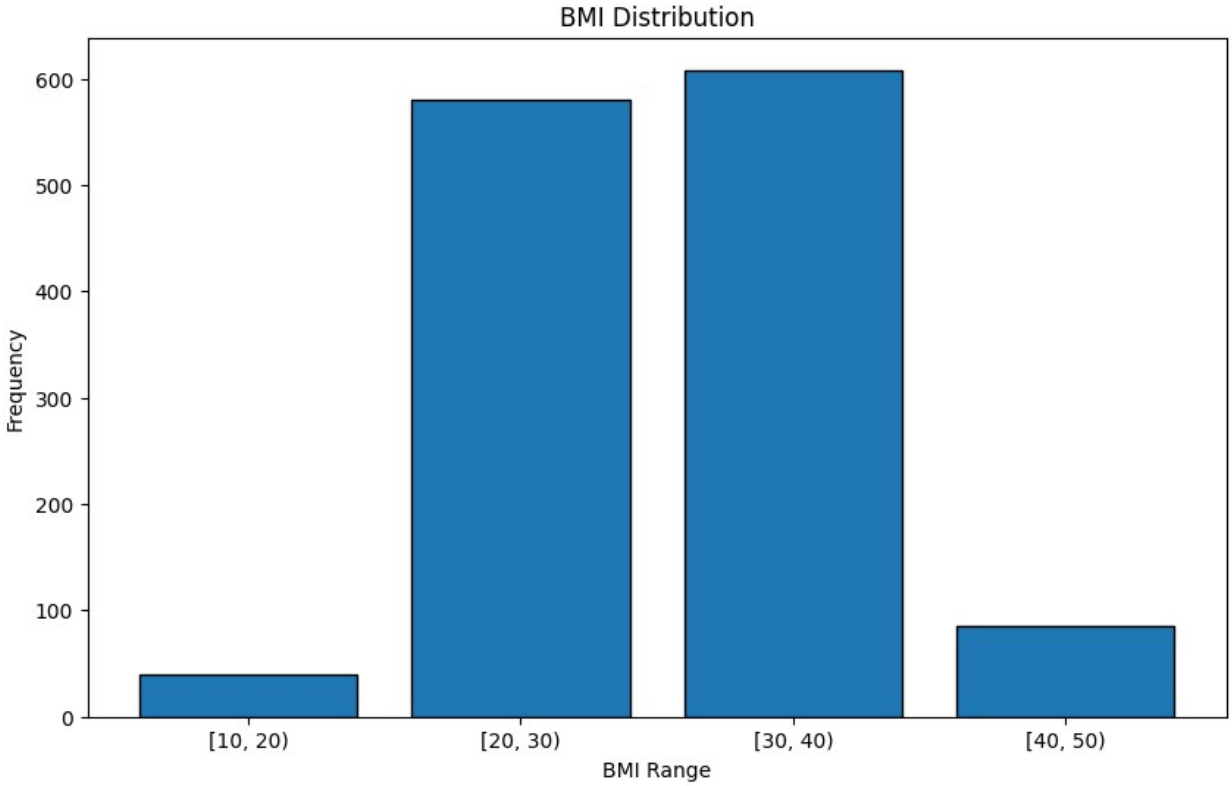
```
sex = data_insurance['sex'].value_counts()
plt.figure(figsize=(10,6))
plt.pie(sex, labels=sex.index, autopct='%1.1f%%')
plt.title('Gender proportion')
plt.show()
```



BMI

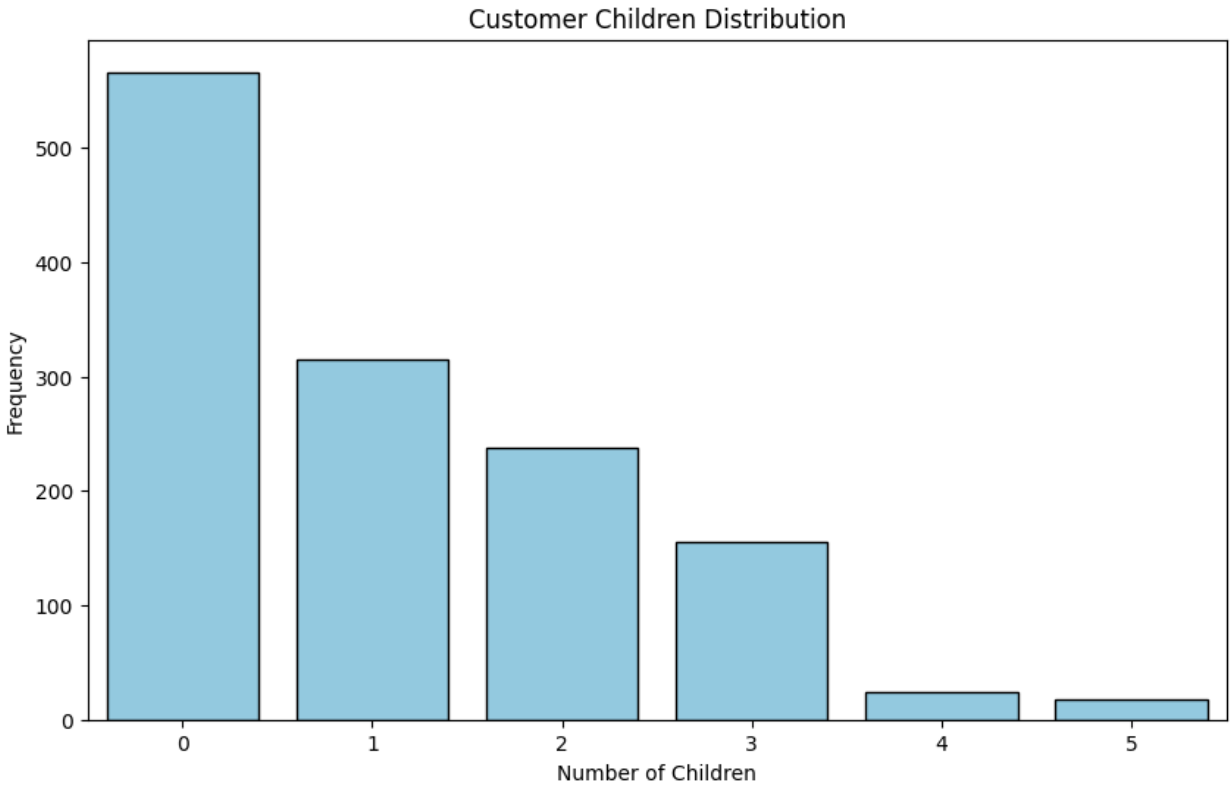
```
bins = np.arange(10, 51, 10)

bmi_counts = pd.cut(data_insurance['bmi'], bins=bins,
right=False).value_counts().sort_index()

plt.figure(figsize=(10, 6))

plt.bar(bmi_counts.index.astype(str), bmi_counts, edgecolor='black')

plt.title('BMI Distribution')
plt.xlabel('BMI Range')
```

```
plt.ylabel('Frequency')
plt.show()
```



Amount of Children

```
plt.figure(figsize=(10, 6))
sns.barplot(x=data_insurance.children.value_counts().index,
y=data_insurance.children.value_counts(), color='skyblue',
edgecolor='black')

plt.title('Customer Children Distribution')
plt.xlabel('Number of Children')
plt.ylabel('Frequency')
plt.show()
```



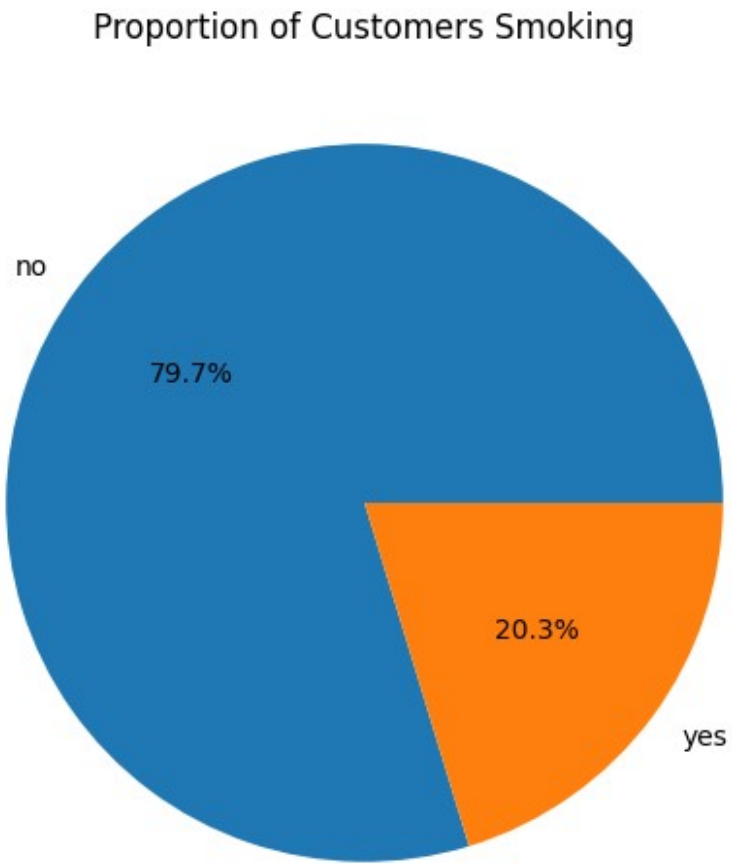

```
data_insurance.head()
```

```
   id  age     sex     bmi  children smoker     region      charges
age_group  \
0   1   19  female  27.900         0    yes  southwest  16884.92400
Teen
1   2   18    male  33.770         1     no  southeast   1725.55230
Teen
2   3   28    male  33.000         3     no  southeast   4449.46200
Elderly
3   4   33    male  22.705         0     no  northwest  21984.47061
Elderly
4   5   32    male  28.880         0     no  northwest   3866.85520
Elderly

  bmi_category   status charges_category
0   overweight   single          medium
1        obese  married        very low
2        obese  married        very low
3       normal   single            high
4   overweight   single        very low
```

Smoker

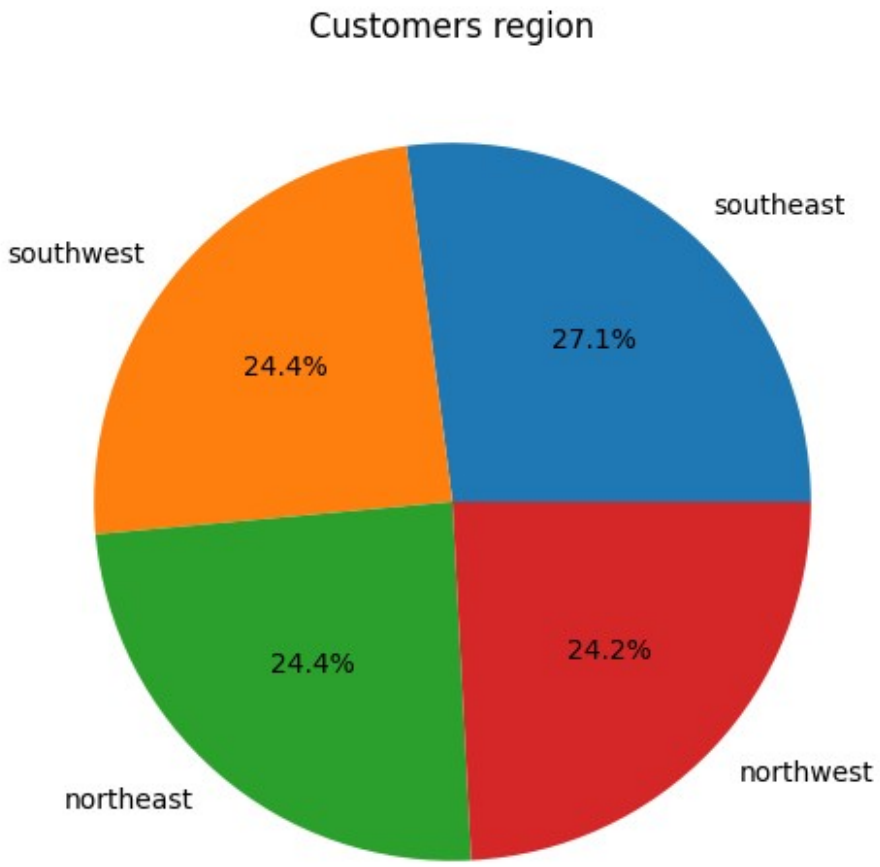
```
smoker_counts = data_insurance['smoker'].value_counts()
plt.figure(figsize=(10,6))
```

```
plt.pie(smoker_counts, labels=smoker_counts.index, autopct='%1.1f%%')
plt.title('Proportion of Customers Smoking')
plt.show()
```



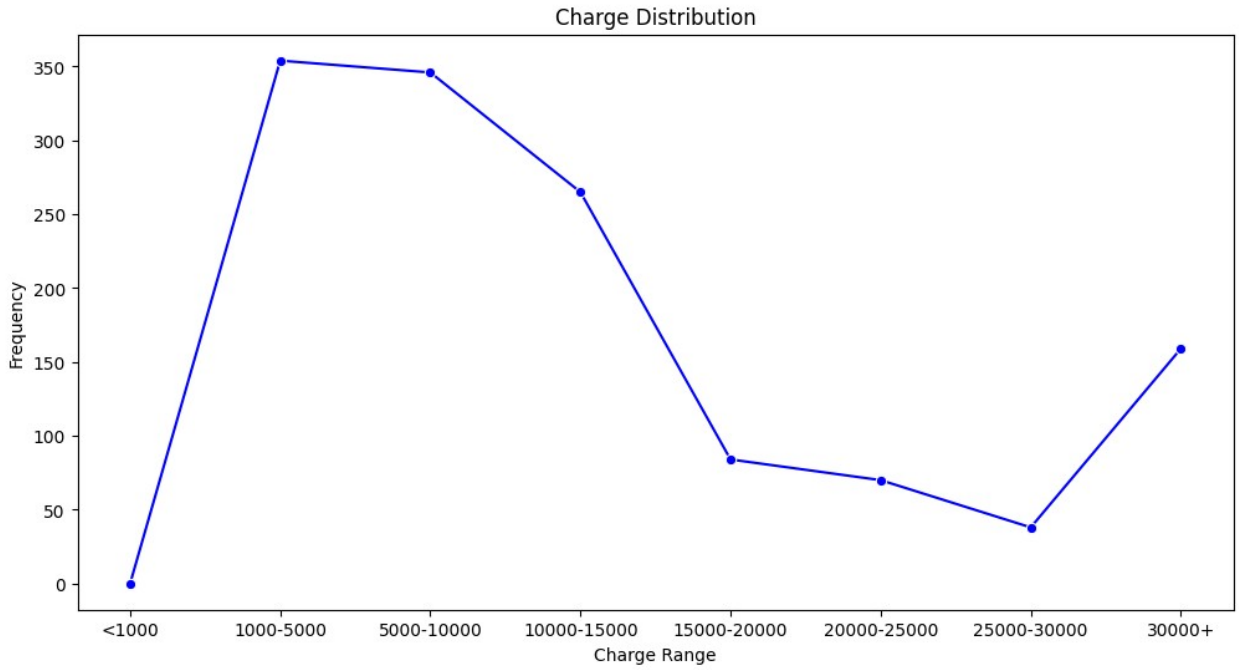
Region

```
region = data_insurance['region'].value_counts()
plt.figure(figsize=(10,6))
plt.pie(region, labels=region.index, autopct='%1.1f%%')
plt.title('Customers region')
plt.show()
```



Charges/Bill

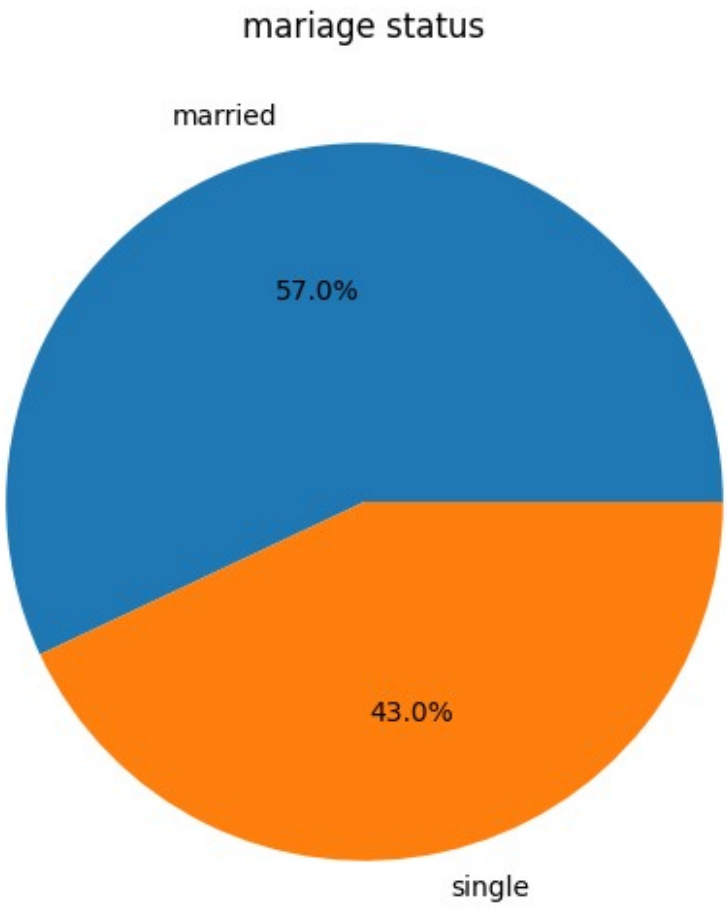
```
# Visualisasi untuk klasifikasi charge/premi nasabah dengan range
tertentu
bins = [-np.inf, 1000, 5000, 10000, 15000, 20000, 25000, 30000,
np.inf]
labels = ['<1000', '1000-5000', '5000-10000', '10000-15000', '15000-
20000', '20000-25000', '25000-30000', '30000+']

charge_counts = pd.cut(data_insurance['charges'], bins=bins,
labels=labels, right=False).value_counts().sort_index()

plt.figure(figsize=(12, 6))

sns.lineplot(x=charge_counts.index.astype(str), y=charge_counts,
marker='o', color='b')

plt.title('Charge Distribution')
plt.xlabel('Charge Range')
plt.ylabel('Frequency')
plt.show()
```



Pertanyaan 2

status

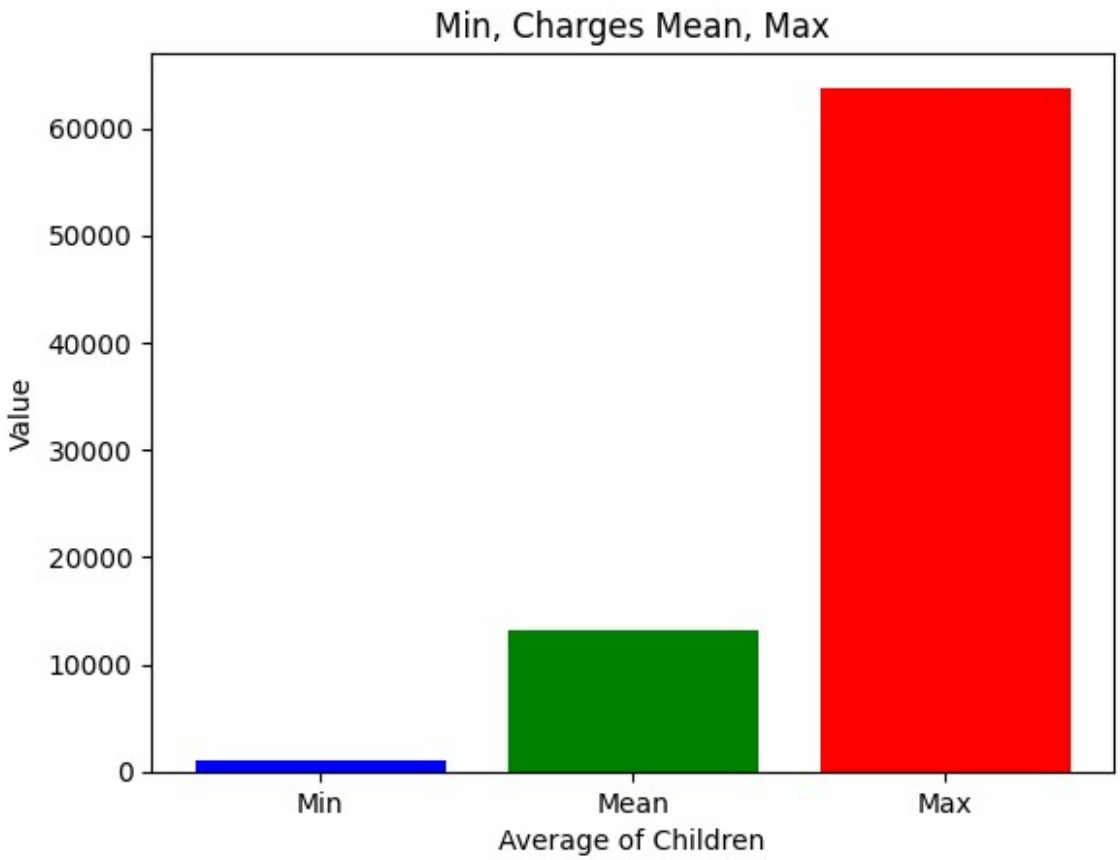
```
status = data_insurance['status'].value_counts()
plt.figure(figsize=(10,6))
plt.pie(status,labels=status.index,autopct="%1.1f%%")
plt.title('mariage status')
plt.show()
```




```
# menghitung rata-rata jumlah anak yang dimiliki nasabah
average_children = data_insurance['children'].mean()
print(average_children)
```

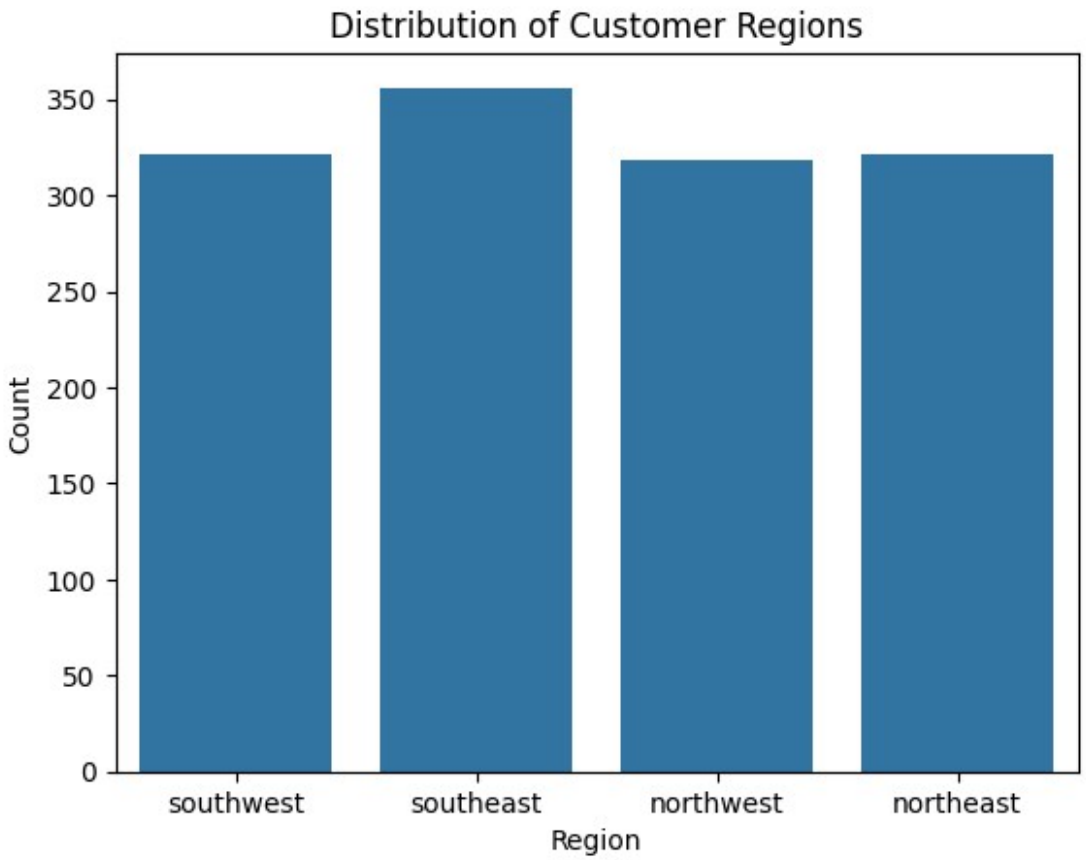
```
1.0957446808510638
```

```
# visualisasi untuk jumlah anak dari terkecil, rata-rata dan terbanyak
yang dimiliki nasabah
mean_charges = data_insurance['charges'].mean()
min_charges = data_insurance['charges'].min()
max_charges = data_insurance['charges'].max()

x = ['Min', 'Mean', 'Max']
y = [ min_charges,mean_charges, max_charges]

plt.bar(x, y, color=['blue', 'green', 'red'])
plt.title('Min, Charges Mean, Max')
plt.xlabel('Average of Children')
plt.ylabel('Value')
plt.show()
```



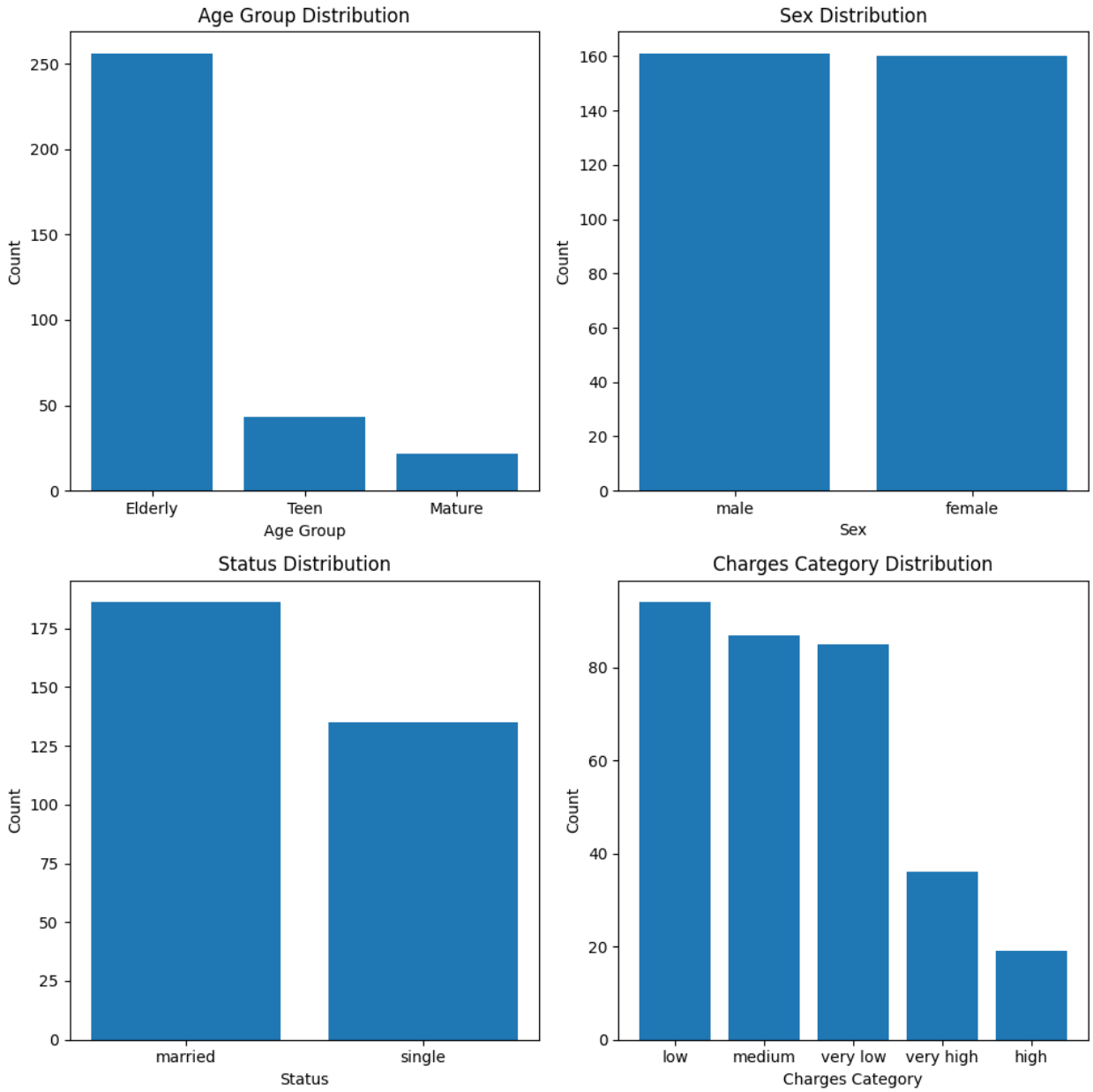
Pertanyaan 3

```
sns.countplot(x='region', data=data_insurance)

plt.title('Distribution of Customer Regions')
plt.xlabel('Region')
plt.ylabel('Count')

plt.show()
```



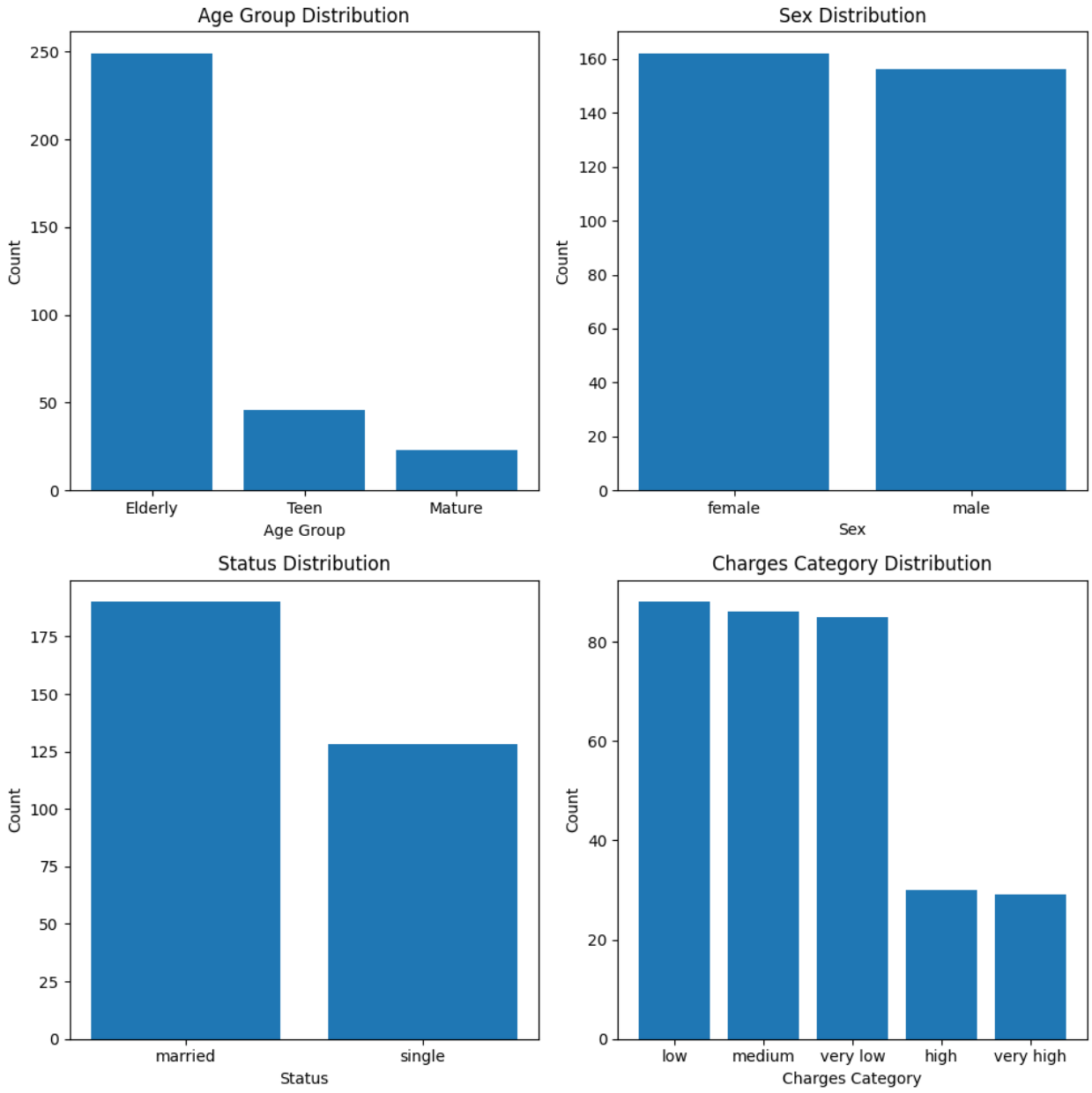

```
# menampilkan statistik jumlah nasabah berdasarkan klasifikasi umur,
gender, status dan charge/premi yang dikenakan di region southwest
southwest_df = data_insurance[data_insurance['region'] == 'southwest']
charges_category_dist =
southwest_df['charges_category'].value_counts()

age_group_dist = southwest_df['age_group'].value_counts()
sex_dist = southwest_df['sex'].value_counts()
status_dist = southwest_df['status'].value_counts()

fig, axs = plt.subplots(2, 2, figsize=(10, 10))

axs[0, 0].bar(age_group_dist.index, age_group_dist.values)
axs[0, 0].set_title('Age Group Distribution')
axs[0, 0].set_xlabel('Age Group')
axs[0, 0].set_ylabel('Count')

axs[0, 1].bar(sex_dist.index, sex_dist.values)
axs[0, 1].set_title('Sex Distribution')
axs[0, 1].set_xlabel('Sex')
axs[0, 1].set_ylabel('Count')

axs[1, 0].bar(status_dist.index, status_dist.values)
```

```
axs[1, 0].set_title('Status Distribution')
axs[1, 0].set_xlabel('Status')
axs[1, 0].set_ylabel('Count')

axs[1, 1].bar(charges_category_dist.index,
charges_category_dist.values)
axs[1, 1].set_title('Charges Category Distribution')
axs[1, 1].set_xlabel('Charges Category')
axs[1, 1].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



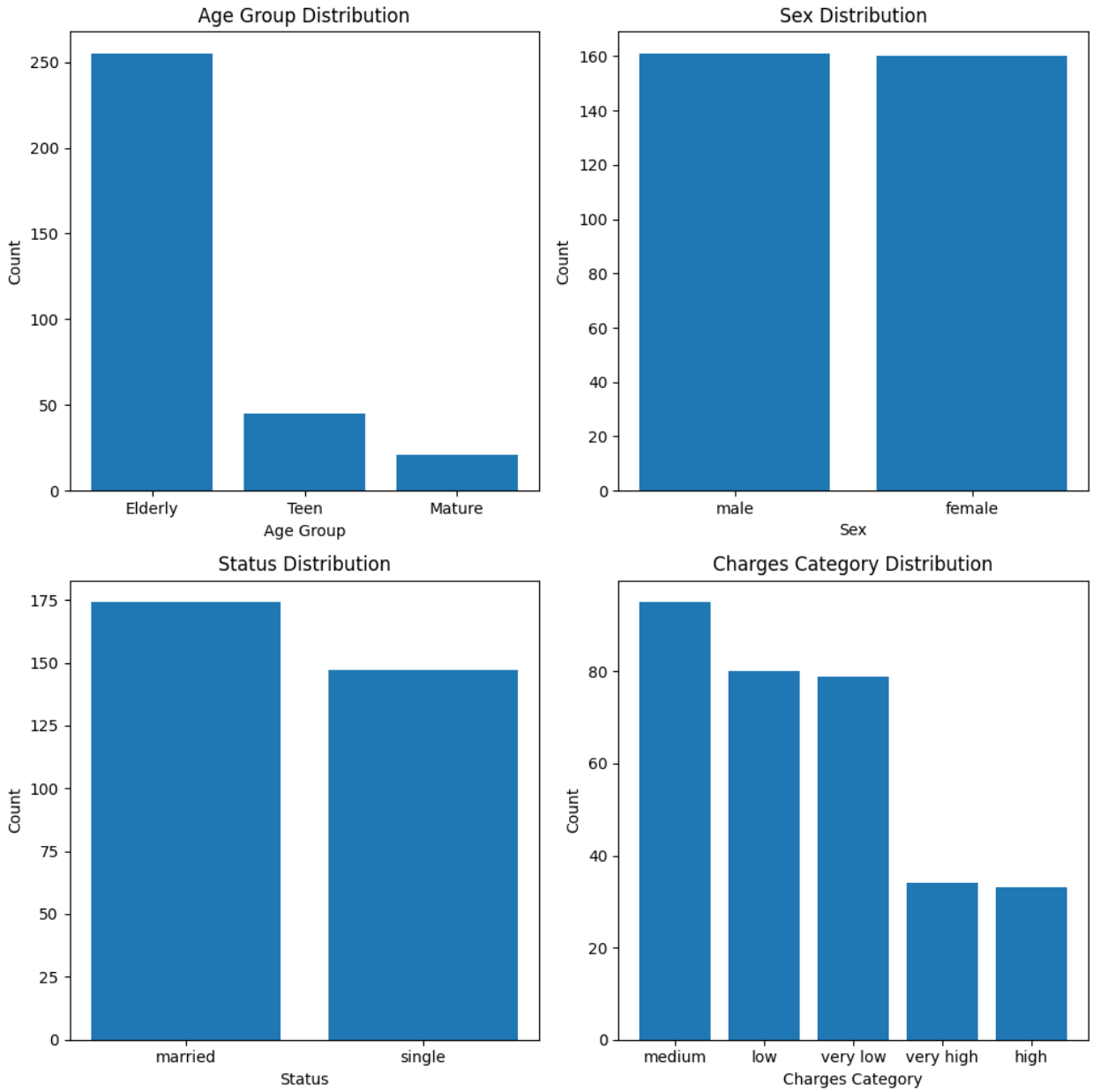

```
# menampilkan statistik jumlah nasabah berdasarkan klasifikasi umur,
gender, status dan charge/premi yang dikenakan di region northwest
northwest_df = data_insurance[data_insurance['region'] == 'northwest']
charges_category_dist =
northwest_df['charges_category'].value_counts()

age_group_dist = northwest_df['age_group'].value_counts()
sex_dist = northwest_df['sex'].value_counts()
status_dist = northwest_df['status'].value_counts()

fig, axs = plt.subplots(2, 2, figsize=(10, 10))
```

```
axs[0, 0].bar(age_group_dist.index, age_group_dist.values)
axs[0, 0].set_title('Age Group Distribution')
axs[0, 0].set_xlabel('Age Group')
axs[0, 0].set_ylabel('Count')

axs[0, 1].bar(sex_dist.index, sex_dist.values)
axs[0, 1].set_title('Sex Distribution')
axs[0, 1].set_xlabel('Sex')
axs[0, 1].set_ylabel('Count')

axs[1, 0].bar(status_dist.index, status_dist.values)
axs[1, 0].set_title('Status Distribution')
axs[1, 0].set_xlabel('Status')
axs[1, 0].set_ylabel('Count')

axs[1, 1].bar(charges_category_dist.index,
charges_category_dist.values)
axs[1, 1].set_title('Charges Category Distribution')
axs[1, 1].set_xlabel('Charges Category')
axs[1, 1].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



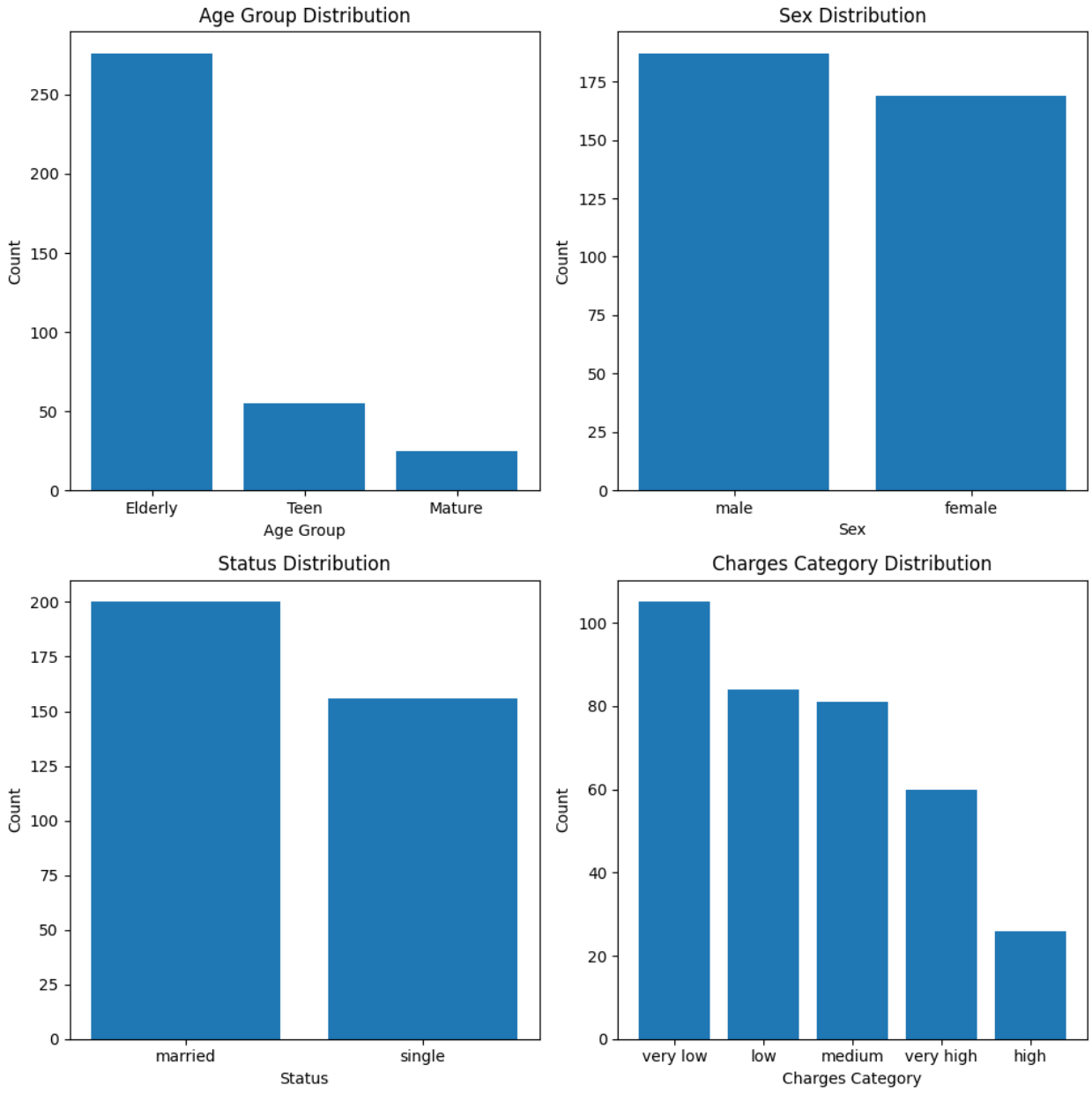

```
# menampilkan statistik jumlah nasabah berdasarkan klasifikasi umur,
gender, status dan charge/premi yang dikenakan di region northeast
northeast_df = data_insurance[data_insurance['region'] == 'northeast']
charges_category_dist =
northeast_df['charges_category'].value_counts()

age_group_dist = northeast_df['age_group'].value_counts()
sex_dist = northeast_df['sex'].value_counts()
status_dist = northeast_df['status'].value_counts()

fig, axs = plt.subplots(2, 2, figsize=(10, 10))
```

```
axs[0, 0].bar(age_group_dist.index, age_group_dist.values)
axs[0, 0].set_title('Age Group Distribution')
axs[0, 0].set_xlabel('Age Group')
axs[0, 0].set_ylabel('Count')

axs[0, 1].bar(sex_dist.index, sex_dist.values)
axs[0, 1].set_title('Sex Distribution')
axs[0, 1].set_xlabel('Sex')
axs[0, 1].set_ylabel('Count')

axs[1, 0].bar(status_dist.index, status_dist.values)
axs[1, 0].set_title('Status Distribution')
axs[1, 0].set_xlabel('Status')
axs[1, 0].set_ylabel('Count')

axs[1, 1].bar(charges_category_dist.index,
charges_category_dist.values)
axs[1, 1].set_title('Charges Category Distribution')
axs[1, 1].set_xlabel('Charges Category')
axs[1, 1].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



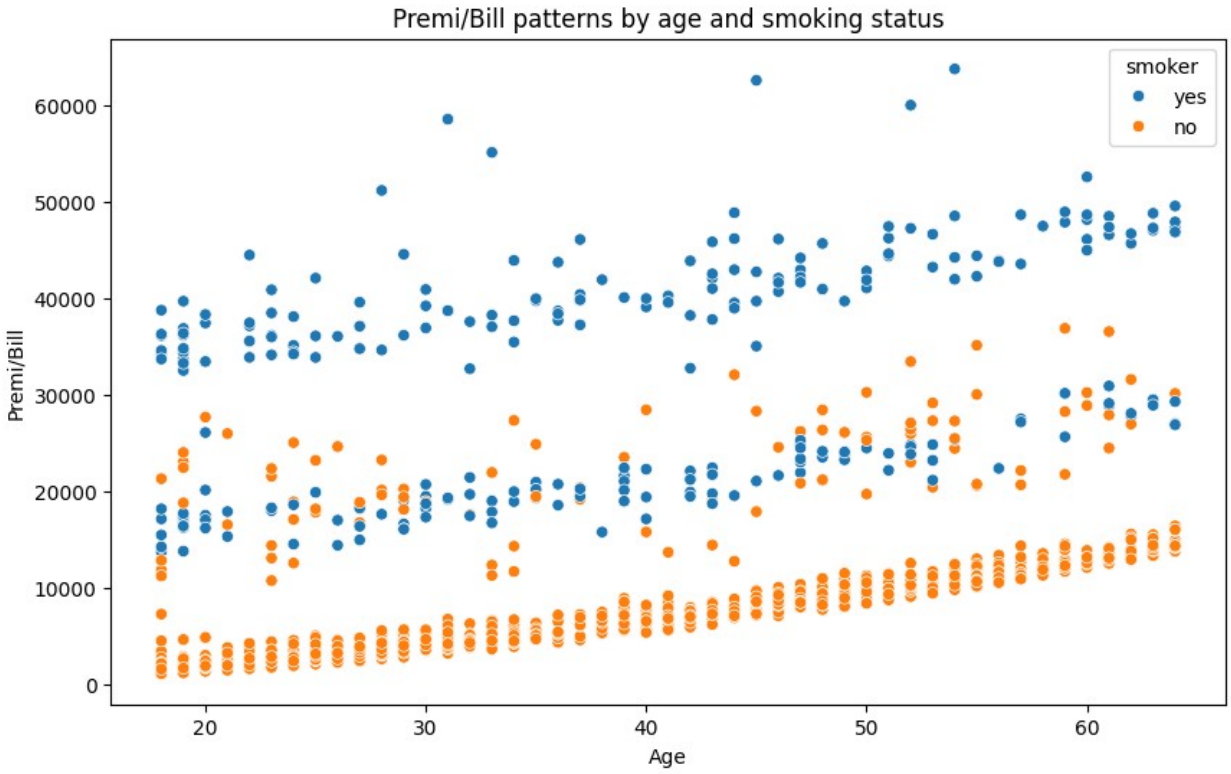

```
# menampilkan statistik jumlah nasabah berdasarkan klasifikasi umur,
gender, status dan charge/premi yang dikenakan di region southeast
southeast_df = data_insurance[data_insurance['region'] == 'southeast']
charges_category_dist =
southeast_df['charges_category'].value_counts()

age_group_dist = southeast_df['age_group'].value_counts()
sex_dist = southeast_df['sex'].value_counts()
status_dist = southeast_df['status'].value_counts()

fig, axs = plt.subplots(2, 2, figsize=(10, 10))
```

```
axs[0, 0].bar(age_group_dist.index, age_group_dist.values)
axs[0, 0].set_title('Age Group Distribution')
axs[0, 0].set_xlabel('Age Group')
axs[0, 0].set_ylabel('Count')

axs[0, 1].bar(sex_dist.index, sex_dist.values)
axs[0, 1].set_title('Sex Distribution')
axs[0, 1].set_xlabel('Sex')
axs[0, 1].set_ylabel('Count')

axs[1, 0].bar(status_dist.index, status_dist.values)
axs[1, 0].set_title('Status Distribution')
axs[1, 0].set_xlabel('Status')
axs[1, 0].set_ylabel('Count')

axs[1, 1].bar(charges_category_dist.index,
charges_category_dist.values)
axs[1, 1].set_title('Charges Category Distribution')
axs[1, 1].set_xlabel('Charges Category')
axs[1, 1].set_ylabel('Count')

plt.tight_layout()
plt.show()
```



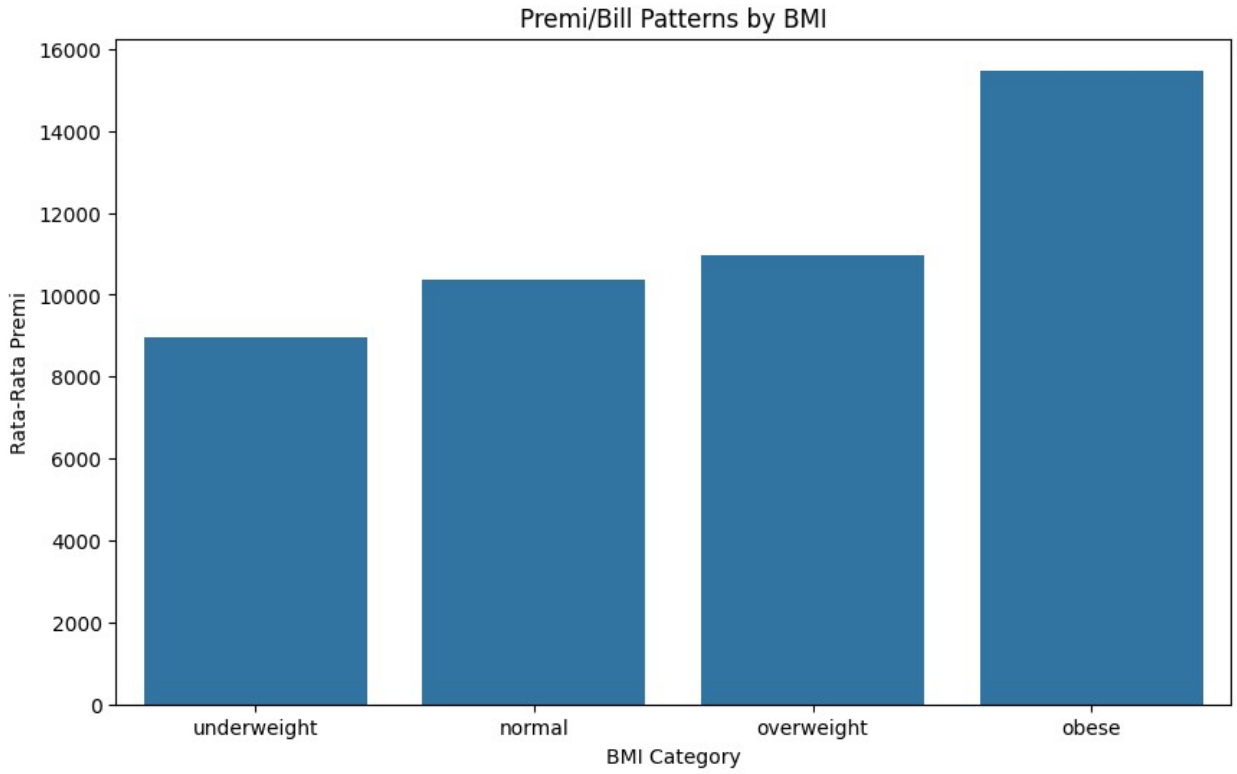
Pertanyaan 4

```
# visualisasi untuk pola charge/premi dari kolom age dan smoker
plt.figure(figsize=(10,6))
sns.scatterplot(data=data_insurance, x='age', y='charges',
hue='smoker')

plt.title('Premi/Bill patterns by age and smoking status')
plt.xlabel('Age')
plt.ylabel('Premi/Bill')
plt.show()
```

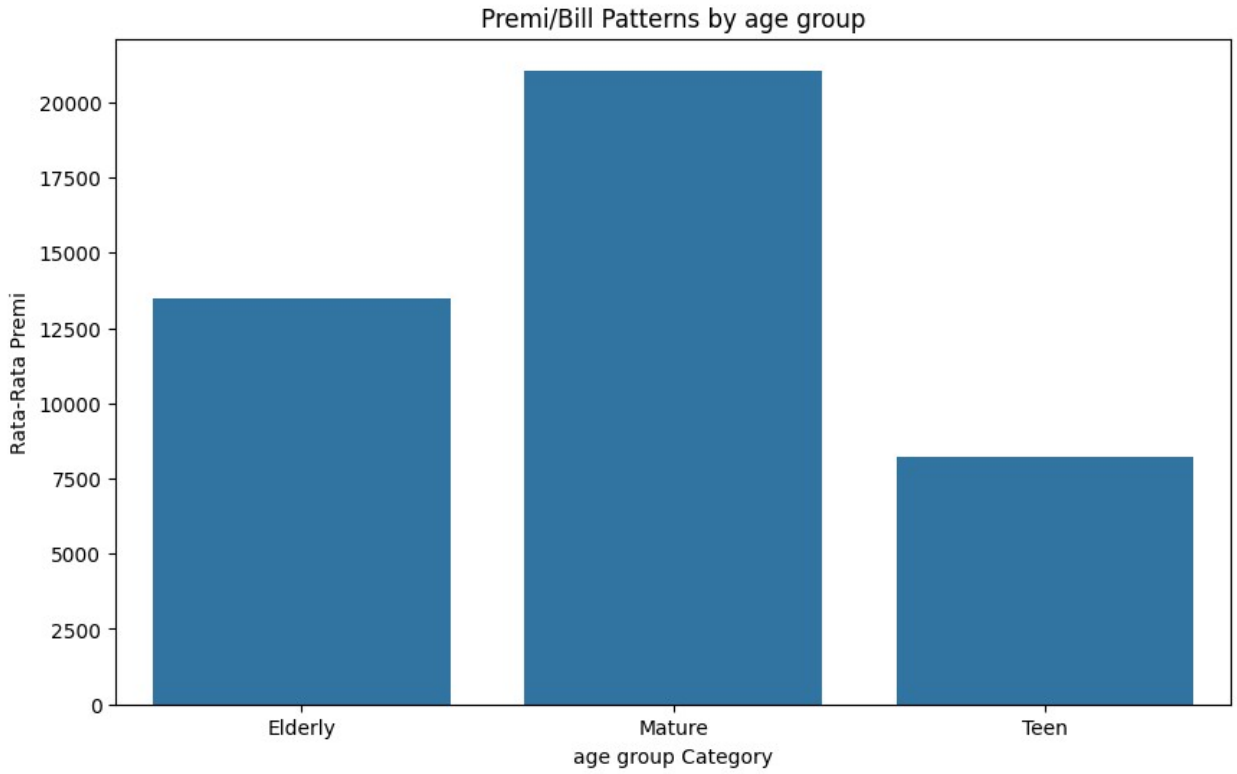



```
# visualisasi untuk klasifikasi BMI
average_charges = data_insurance.groupby('bmi_category')
['charges'].mean()

plt.figure(figsize=(10, 6))

sns.barplot(x=average_charges.index.astype(str), y=average_charges)

plt.title('Premi/Bill Patterns by BMI')
plt.xlabel('BMI Category')
plt.ylabel('Rata-Rata Premi')
plt.show()
```

```
C:\Users\Albert\AppData\Local\Temp\ipykernel_16648\3640941723.py:2:
FutureWarning: The default of observed=False is deprecated and will be
changed to True in a future version of pandas. Pass observed=False to
retain current behavior or observed=True to adopt the future default
and silence this warning.
  average_charges = data_insurance.groupby('bmi_category')
['charges'].mean()
```



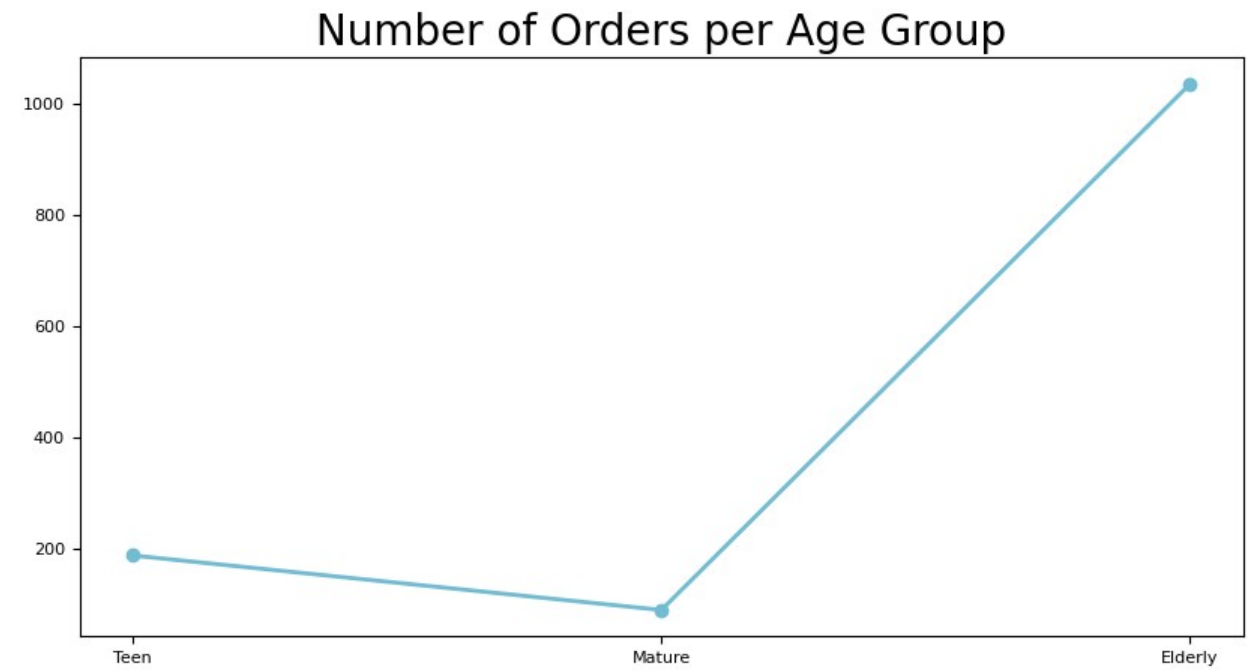

```
# visualisasi untuk klasifikasi charge/premi berdasarkan age_group
average_charges = data_insurance.groupby('age_group')
['charges'].mean()

plt.figure(figsize=(10, 6))

sns.barplot(x=average_charges.index.astype(str), y=average_charges)

plt.title('Premi/Bill Patterns by age group')
plt.xlabel('age group Category')
plt.ylabel('Rata-Rata Premi')
plt.show()
```



5. Revenue dan Order

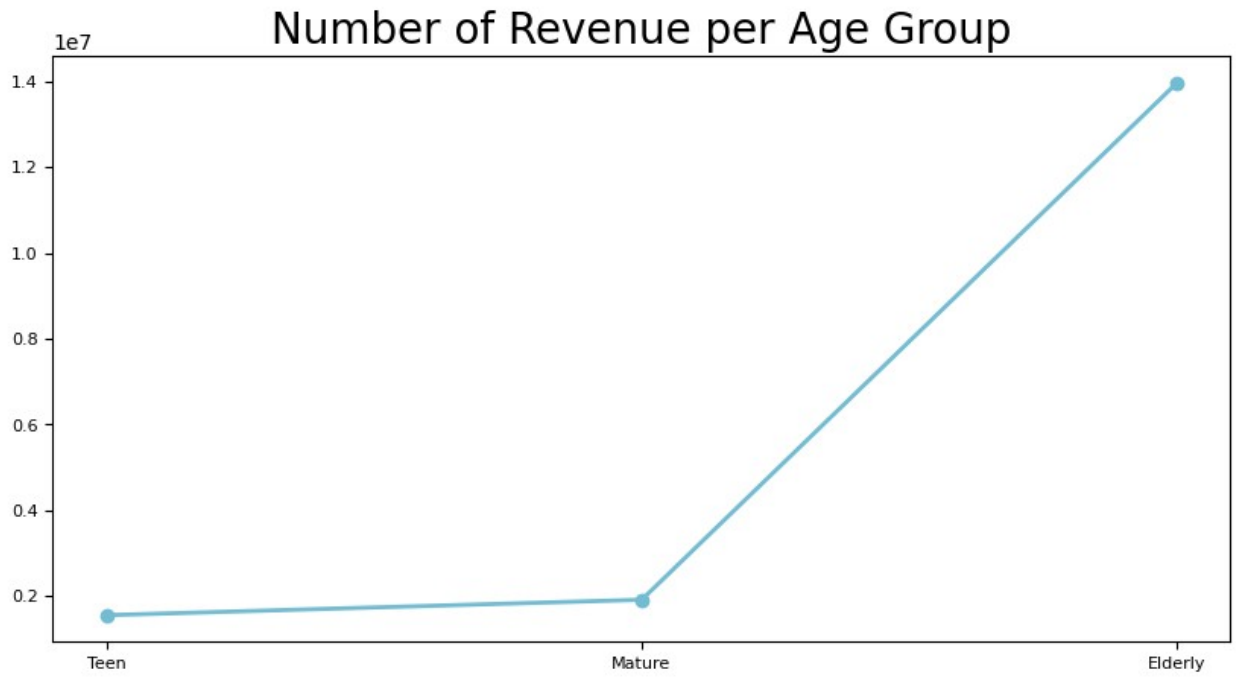
```
# visualisasi untuk jumlah order dan revenue berdasarkan id, age_group
dan charges
data_insurance['age_group'] = data_insurance['age_group']

profit = data_insurance.groupby('age_group').agg({
    "id": "count",
    "charges": "sum"
})

profit = profit.reset_index()
profit.rename(columns={"id": "order", "charges": "revenue"},
inplace=True)
age_groups_order = sorted(data_insurance['age_group'].unique())
profit['age_group'] = pd.Categorical(profit['age_group'],
categories=age_groups_order, ordered=True)
profit = profit.sort_values('age_group')

# visualisasi untuk order
profit =profit.sort_values('age_group', ascending=False)

plt.figure(figsize=(10, 5))
plt.plot(
    profit["age_group"],
    profit["order"],
    marker='o',
```

```
    linewidth=2,
    color="#72BCD4"
)

plt.title("Number of Orders per Age Group", loc="center", fontsize=20)
plt.xticks(fontsize=8)
plt.yticks(fontsize=8)
plt.show()
```



Revenue

```
# visualisasi untuk Revenue
profit = profit.sort_values('age_group', ascending=False)

plt.figure(figsize=(10, 5))
plt.plot(
    profit["age_group"],
    profit["revenue"],
    marker='o',
    linewidth=2,
    color="#72BCD4"
)

plt.title("Number of Revenue per Age Group", loc="center",
fontsize=20)
plt.xticks(fontsize=8)
plt.yticks(fontsize=8)
plt.show()
```



Tahap 8: Membuat Kesimpulan

Berikan kesimpulan untuk setiap pertanyaan berdasarkan keseluruhan tahapan yang telah dilakukan sebelumnya

Bagaimana dengan distribusi nasabah asuransi dalam dataset ini?

- kesimpulan (age status) : Dalam visualisasi menggunakan BAR PLOT menunjukkan jumlah distribusi usia pelanggan. terdapat umur 20 terdiri dari lebih dari 200 orang, umur 30 terdiri dari 100 orang, umur 40 terdiri dari kurang lebih 100 orang, umur 50 terdiri dari kurang lebih 150 dan umur 60 terdiri dari lebh dari 100.
- kesimpulan (sex status) Dalam visualisasi menggunakan PIE CHART menunjukkan persentase proporsi gender, yakni laki-laki dan perempuan. persentase male sekitar 50.5% sedangkan persentase female adalah 49.5%.
- kesimpulan (bmi status) : Dalam visualisasi menggunakan BAR PLOT menunjukkan distribusi bmi. terdapat range bmi (10,20) mencapai kurang dari 100, range bmi (20,30) mencapai kurang dari 600, range bmi (30,40) mecapai 600 lebih dan range bmi (40,50) mencapai kurang lebih 100.
- kesimpulan (amount of children) : Dalam visualisasi menggunakan BAR menunjukkan distribusi anak pelanggan. terdapat tidak mempunyai anak sebanyak lebih dari 500, 1 anak sebanyak kurang lebih 300, 2 anak sebanyak kurang lebih 200, 3 anak sebanyak kurang lebih 100.
- kesimpulan (smoker status) : Dalam visualisasi menggunakan PIE CHART menunjukkan persenan yang merokok dan tidak merokok. terdapat yang merokok sekitar 20.3%. sedangkan yang tidak merokok sekitar 79.7%.
- kesimpulan (region status) : Dalam visualisasi menggunakan PIE CHART menunjukkan persenan yang berdomisili. yakni southwest, southeast, northeast dan northwest.

terdapat pada southwest sekitar 24.4%, southeast sekitar 27.1%, northeast sekitar 24.4% dan northwest sekitar 24.2%.

- kesimpulan (charges/bill) : Dalam visualisasi menggunakan ROW menunjukkan pinjaman dan frequency. terdapat pada pinjaman <1000 = 0, pinjaman 1000-5000 sekitar 350, pinjaman 5000-10000 sekitar kurang dari 350, pinjaman 10000-15000 sekitar lebih dari 200, pinjaman 15000-20000 sekitar lebih dari 50, pinjaman 25000-30000 sekitar 50 dan pinjaman 30000+ sekitar 150.

Berapa rata-rata jumlah anak yang dimiliki oleh nasabah?

- Visualisasi menunjukkan bahwa mayoritas orang (57%) berstatus menikah dan rata-rata jumlah anak berkisar antara kurang dari 1000 hingga lebih dari 6000.

Bagaimana persebaran wilayah domisili nasabah?

- Visualisasi tersebut pada tabel age group distribution menampilkan bahwa hitungan elderly lebih tinggi daripada teen dan mature. Lalu, untuk tabel sex distribution menampilkan bahwa hitungan gender dari male sedikit lebih tinggi daripada female. Dan untuk tabel status distribution menampilkan bahwa hitungan status dari married lebih tinggi daripada single.

Tampilkan pola atau tren dalam premi yang dikenakan berdasarkan kombinasi faktor usia, BMI, dan status perokok?

- rata-rata premi asuransi cenderung meningkat seiring bertambahnya usia. Kelompok usia “Mature” memiliki rata-rata premi tertinggi, diikuti oleh “Elderly”, dan “Teen” memiliki rata-rata premi terendah. Ini menunjukkan bahwa biaya asuransi cenderung lebih tinggi untuk individu yang lebih tua. dapat simpulkan bahwa premi asuransi cenderung meningkat seiring bertambahnya usia dan BMI, dan perokok biasanya membayar premi yang lebih tinggi. Kategori BMI juga mempengaruhi premi, dengan individu ‘obese’ cenderung membayar premi tertinggi.

bagaimana performa Order dan revenue perusahaan?

- Kesimpulan: Premi asuransi dan pendapatan cenderung meningkat seiring bertambahnya usia dan BMI, dengan perokok dan individu ‘obese’ serta kelompok usia ‘Elderly’ biasanya membayar premi tertinggi dan menghasilkan pendapatan tertinggi.