MyEduSolve

Kampus Merdeka
INDONESIA JAYA

CRISP-DM

Mini Project

Tools

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

Table of Content

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling (EDA)

Business Understanding

Departemen Kesehatan ingin mengeksplorasi data untuk mengungkap wawasan-wawasan menarik dan penting terkait kesehatan masyarakat dan dampak merokok.

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Hipotesis

Individu yang merokok cenderung memiliki risiko kesehatan yang lebih tinggi daripada yang tidak merokok dan Dampak merokok akan berbeda berdasarkan usia dan jenis kelamin individu.

Business Objectives

Mengidentifikasi Dampak merokok pada kesehatan masyarakat mempengaruhi kesehatan tekanan darah, detak jantung, dan lainnya serta Menggambarkan kelompok yang berisiko tinggi terhadap kesehatan merokok seperti usia,jenis kelamin, dan kebiasaan merokok.

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Tujuan Analisis

- Melakukan Analisis mendalam tentang hubungan antara kebiasaan merokok dan berbagai faktor kesehatan.
- Menganalisis risiko kesehatan masyarakat berdasarkan data yang ada untuk mengidentifikasi kelompok yang berisiko tinggi.
- Memberikan rekomendasi dan strategi untuk meningkatkan kualitas layanan kesehatan masyarakat terkait merokok.

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Data Understanding

Total rows 3908 dan 7 columns

Terdapat 14 missing values pada kolom cigs_per_day dan 26 missing values pada kolom chol

Terdapat 8 rows duplicate serta Ada anomali pada umur yaitu umur 1000 dan 150 lalu ada juga anomali pada heart rate yaitu -1 dan 0

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Name of Each Column

```
1 df.columns
```

```
Index(['age', 'sex', 'current_smoker', 'heart_rate', 'blood_pressure',
       'cigs_per_day', 'chol'],
      dtype='object')
```

Shape & Size

```
1 df.shape
```

```
(3908, 7)
```

```
1 df.size
```

```
27356
```

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Checking Data Type

```
 1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3908 entries, 0 to 3907
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   age              3908 non-null   int64  
 1   sex              3908 non-null   object 
 2   current_smoker   3908 non-null   object 
 3   heart_rate       3908 non-null   int64  
 4   blood_pressure   3908 non-null   object 
 5   cigs_per_day     3894 non-null   float64
 6   chol             3882 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 213.8+ KB
```

Data Describe

```
1 df.describe()
```

	age	heart_rate	cigs_per_day	chol
count	3908.000000	3908.000000	3894.000000	3882.000000
mean	53.509212	75.335466	9.275552	236.708398
std	59.655865	13.053347	12.255640	44.381001
min	32.000000	-1.000000	0.000000	113.000000
25%	42.000000	67.000000	0.000000	206.000000
50%	49.000000	75.000000	0.000000	234.000000
75%	56.000000	82.000000	20.000000	263.000000
max	1000.000000	143.000000	70.000000	696.000000

Include Object

```
1 df.describe(include = 'object')
```

	sex	current_smoker	blood_pressure
count	3908	3908	3908
unique	4	2	2317
top	female	no	130/80
freq	2076	1968	18

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Checking Missing Value

```
1 df.isnull().sum()
```

```
age                0
sex                0
current_smoker     0
heart_rate         0
blood_pressure     0
cigs_per_day      14
chol              26
dtype: int64
```

Duplicate Row

```
1 df.duplicated().sum()
```

```
8
```

Unique Value

```
1 df['sex'].unique()
```

```
array(['male', 'female', 'f', 'm'], dtype=object)
```

```
1 df['age'].unique()
```

```
array([ 54,  45,  58,  42,  57,  43,  37,  49,  55,  39,  53,
        48,  46,  40,  56,  38,  65,  41,  44,  36,  64,  68,
        52,  60,  67,  35,  34,  51,  63,  62,  59,  61,  50,
        66,  47,  70,  69, 150,  33,  32, 1000])
```

```
1 df['current_smoker'].unique()
```

```
array(['yes', 'no'], dtype=object)
```

```
1 df['heart_rate'].unique()
```

```
array([ 95,  64,  81,  90,  62,  75,  66,  65,  93,  70,  85,  58,  83,
        80,  60,  72,  71, 105,  53,  74,  63,  82,  67,  76,  68,  77,
        69,  55,  87,  86,  52,  79, 100,  78,  88,  48, 104,  92,  84,
        50,  94, 120,  98, 122, 101, 110, 107,  96,  73,  56, 103,  57,
       106,  61, 102,  89, 125,  54,  51,  91, 115,  44,  47,  45, 140,
       108,  59, 143,   0,  -1,  46, 112,  99, 130,  97])
```

```
1 df['cigs_per_day'].unique()
```

```
array([nan,  0.,  1.,  2.,  3.,  4.,  5.,  6.,  7.,  8.,  9., 10., 11.,
       12., 13., 14., 15., 16., 17., 18., 19., 20., 23., 25., 29., 30.,
       35., 38., 40., 43., 45., 50., 60., 70.])
```

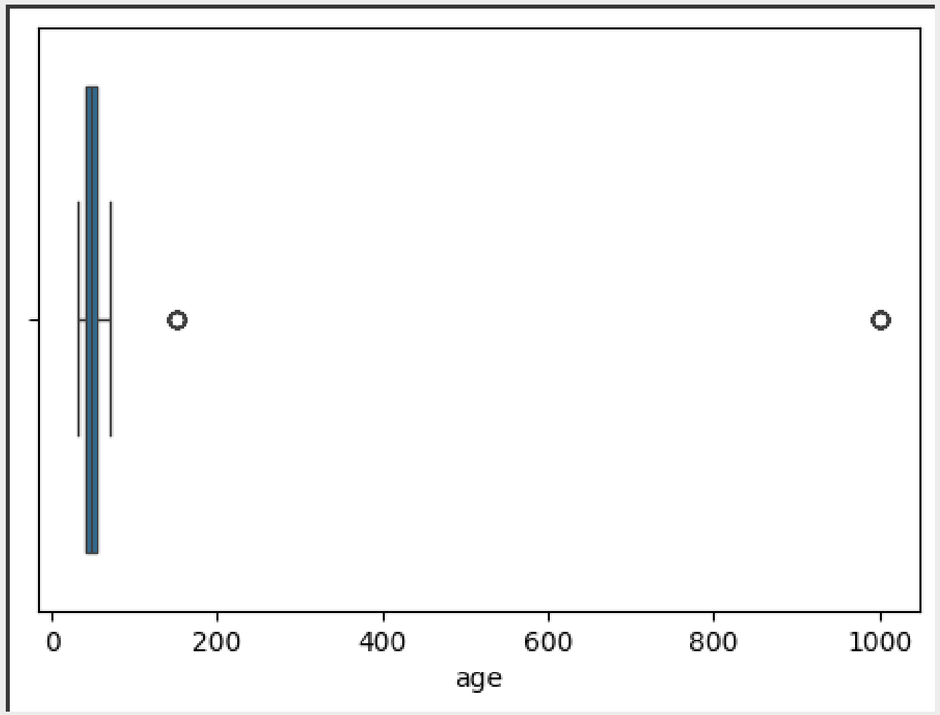
Unique Value

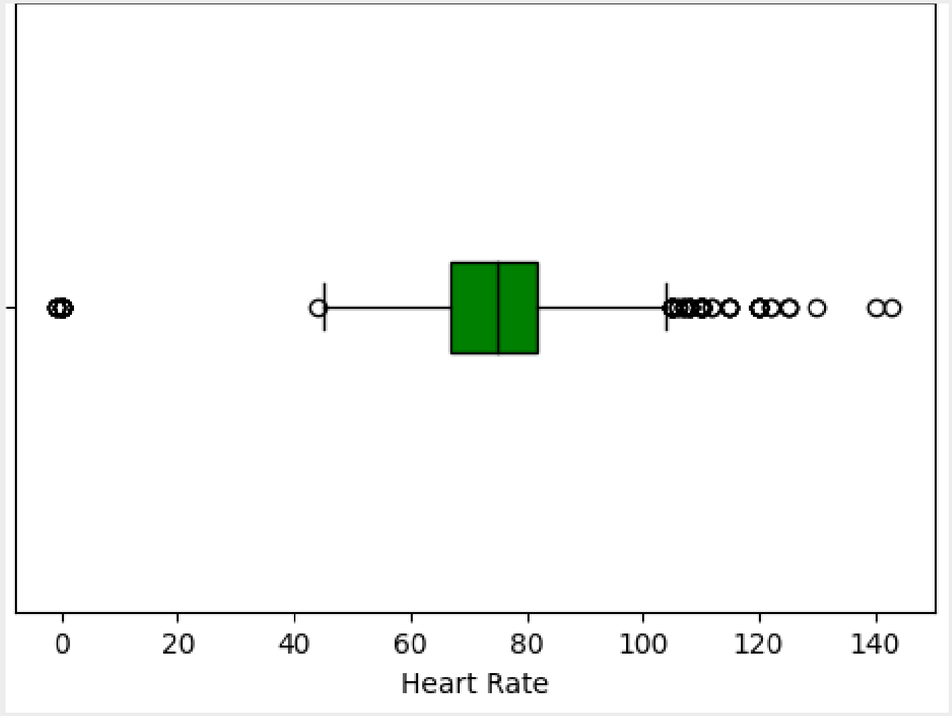
```
1 df['chol'].unique()
```

```
array([219., 248., 235., 225., 226., 223., 222., 196., 188., 256., 214.,
       285., 276., 170., 175., 240., 199., 300., 232., 167., 210., 207.,
       253., 149., 195., 169., 213., 192., 200., 228., 212., 185., 204.,
       237., 181., 227., 270., 197., 168., 215., 187., 391., 171., 249.,
       245., 202., 216., 193., 234., 230., 323., 290., 239., 203., 209.,
       314., 273., 278., 217., 182., 159., 254., 312., 229., 220., 265.,
       186., 246., 251., 177., 260., 258., 208., 282., 280., 183., 266.,
       311., 264., 301., 173., 283., 190., 176., 261., 293., 250., 211.,
       244., 231., 238., 205., 298., 287., 247., 252., 366., 198., 144.,
       305., 271., 179., 334., 201., 307., 178., 263., 304., 262., 281.,
       191., 257., 289., 221., 206., 275., 333., 236., 165., 242., 172.,
       286., 160., 241., 277., 292., 296., 180., 364., 274., 331., 320.,
       233., 332., 309., 306., 189., 156., 150., 279., 224., 288., 268.,
       302., 243., 259., 297., 218.,  nan, 303., 155., 361., 336., 325.,
       154., 294., 269., 310., 184., 267., 324., 126., 346., 295., 339.,
       272., 135., 330., 163., 382., 255., 318., 340., 291., 164., 372.,
       350., 432., 161., 162., 194., 321., 317., 327., 338., 352., 341.,
       328., 326., 308., 284., 152., 380., 299., 137., 329., 344., 153.,
       313., 319., 174., 315., 322., 158., 368., 600., 347., 316., 166.,
       354., 367., 342., 148., 355., 410., 370., 145., 335., 157., 390.,
       464., 358., 124., 385., 345., 337., 140., 398., 143., 133., 351.,
       696., 392., 359., 453., 371., 353., 405., 439., 119., 360., 113.,
       373., 363.])
```

Outlier

Terdapat Outlier pada kolom age dan Heart_rate





Noise

Data 150 dan 1000 tahun adalah noise di kolom age

```
1 df['age'].unique()
```

```
array([  54,   45,   58,   42,   57,   43,   37,   49,   55,   39,   53,
         48,   46,   40,   56,   38,   65,   41,   44,   36,   64,   68,
         52,   60,   67,   35,   24,   51,   63,   62,   59,   61,   50,
         66,   47,   70,   69,  150,   33,   32, 1000])
```

	age	sex	current_smoker
917	150	female	no
918	150	female	no
919	150	male	no
920	150	male	no
921	150	female	no
922	150	female	no
923	150	female	no
924	150	female	no
925	150	male	no
926	150	male	no
927	150	female	no
928	150	female	no
929	150	male	no
3545	1000	male	yes
3546	1000	male	yes
3547	1000	male	yes
3548	1000	male	yes
3549	1000	male	yes
3550	1000	male	yes
3551	1000	male	yes
3552	1000	male	yes
3553	1000	male	yes
3554	1000	female	yes
3555	1000	female	yes
3556	1000	male	yes
3557	1000	male	yes
3558	1000	male	yes
3559	1000	male	yes

Noise

Data 0 dan -1 adalah noise di kolom heart_rate

```
1 df['heart_rate'].unique()
```

```
array([ 95,  64,  81,  90,  62,  75,  66,  65,  93,  70,  85,  58,  83,
        80,  60,  72,  71, 105,  53,  74,  63,  82,  67,  76,  68,  77,
        69,  55,  87,  86,  52,  79, 100,  78,  88,  48, 104,  92,  84,
        50,  94, 120,  98, 122, 101, 110, 107,  96,  73,  56, 103,  57,
       106,  61, 102,  88, 125,  54,  51,  91, 115,  44,  47,  45, 140,
       108,  59, 143,   0,  -1,  46, 112,  99, 130,  97])
```

	age	sex	current_smoker	heart_rate
1771	47	male	no	0
1772	63	female	no	0
1773	46	female	no	0
1774	51	female	no	0
1775	50	male	no	0
1776	55	female	no	0
1777	61	female	no	0
1778	38	male	no	0
1779	49	female	no	0
1780	48	female	no	0
1781	41	male	no	0
1782	60	female	no	0
1783	54	female	no	0
1806	60	female	no	-1
1807	51	female	no	-1
1808	56	female	no	-1
1809	39	female	no	-1
1810	55	male	no	-1

Inconsistency

Data f dan m adalah inkonsistensi di kolom sex

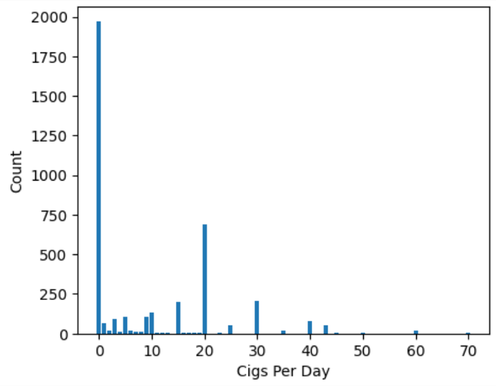
```
1 df['sex'].unique()
```

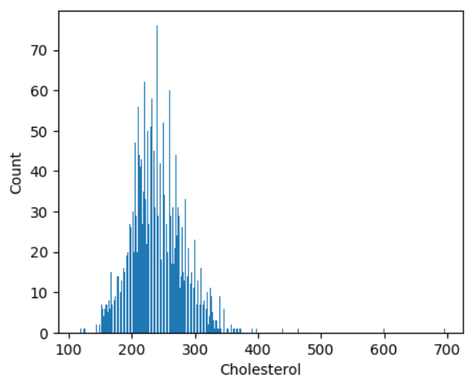
```
array(['male', 'female', 'f', 'm'], dtype=object)
```

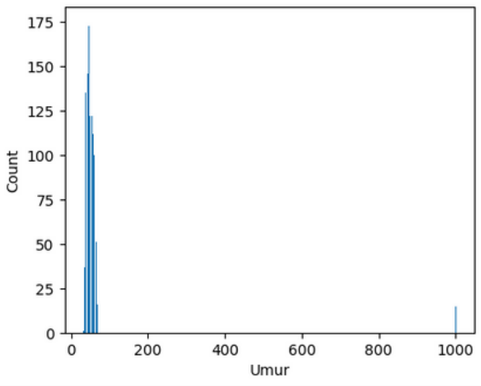
	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol
1728	40	m	no	72	123.5/83	0.0	209.0
1729	67	m	no	75	130/83	0.0	234.0
1730	56	m	no	80	127/83	0.0	194.0
1753	60	m	no	80	140.5/83	0.0	213.0
1754	40	m	no	69	149/83	0.0	213.0
1755	52	m	no	70	157.5/83	0.0	269.0
1756	38	m	no	92	121/83	0.0	235.0

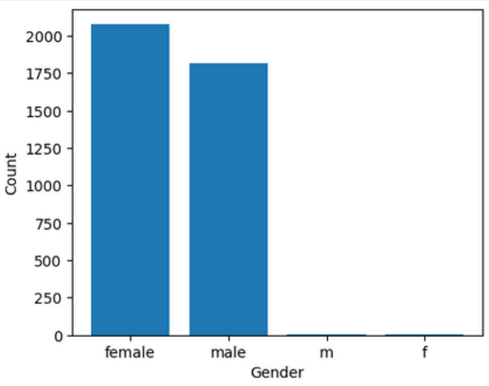
	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol
1709	41	f	no	73	125/83	0.0	237.0
1710	53	f	no	63	123/83	0.0	370.0
1711	43	f	no	90	112/83	0.0	145.0
1718	44	f	no	85	122/83	0.0	187.0
1719	56	f	no	65	125/83	0.0	273.0

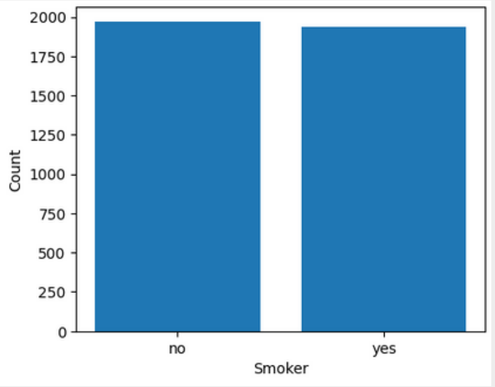
Initial EDA

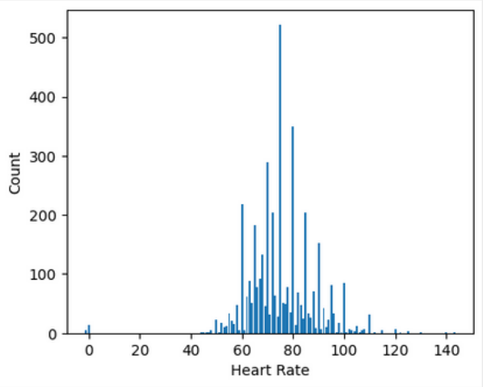












- Banyak responden yang merokok 0 batang sehari (tidak merokok) dan ada juga yang bisa merokok lebih dari 20 batang per hari hingga sekitar 60 batang per hari.
- Rata-rata cholesterol responden sebesar 236 yaitu cukup tinggi.
- Rata-rata umur dari responden adalah 53 tahun.
- Ada lebih banyak responden perempuan dibandingkan laki-laki.
- Lebih banyak responden yang sudah tidak merokok dibandingkan yang masih merokok.
- Rata-rata dari hear rate responden adalah 75 bpm

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Data Preparation

Handling Duplicated Values

Before :

```
df.duplicated().sum()
```

```
8
```

After :

```
df.duplicated().sum()
```

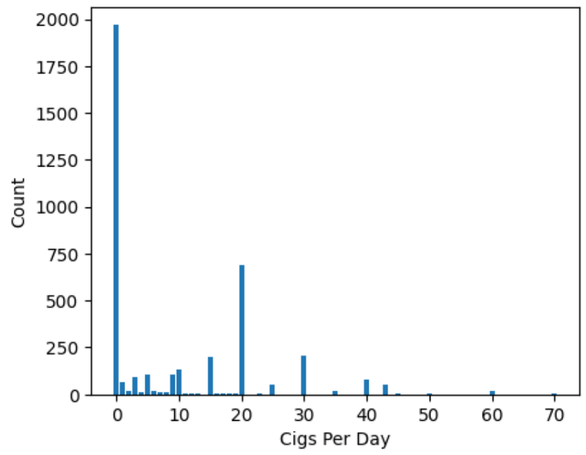
```
0
```

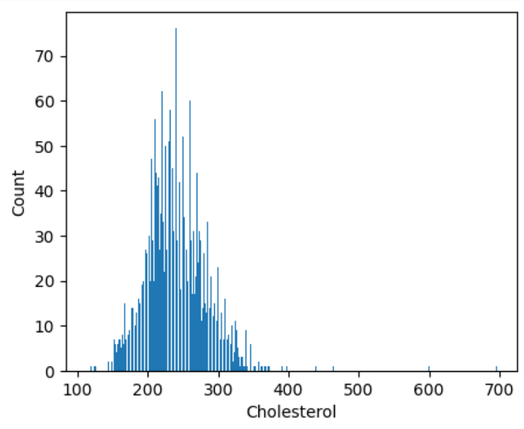
Result:

```
df.drop_duplicates()
```

	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol
0	54	male	yes	95	110/72	NaN	219.0
1	45	male	yes	64	121/72	NaN	248.0
2	58	male	yes	81	127.5/76	NaN	235.0
3	42	male	yes	90	122.5/80	NaN	225.0
4	42	male	yes	62	119/80	NaN	226.0
...	...	...	...	...	...	...	...
3895	37	male	yes	88	122.5/82.5	60.0	254.0
3896	49	male	yes	70	123/75	60.0	213.0
3897	56	male	yes	70	125/79	60.0	246.0
3898	50	male	yes	85	134/95	60.0	340.0
3899	40	male	yes	98	132/86	70.0	210.0

Handling Missing Values





Before

```
1 df.isnull().sum()
```

```
age                0
sex                0
current_smoker     0
heart_rate         0
blood_pressure     0
cigs_per_day      14
chol              26
dtype: int64
```

After :

```
df = df.dropna()
df.isnull().sum()
```

```
age                0
sex                0
current_smoker     0
heart_rate         0
blood_pressure     0
cigs_per_day       0
chol               0
dtype: int64
```

Pada Kolom Cigs_per_day dan Chol di drop karena outlier yang terdapat pada kolom tersebut sangat sedikit persentasenya dan jumlah data yang missing value hanya 40 dari 3900 baris.

Handling Noise

Drop kolom age dengan value 150 dan 1000 yang menyebabkan otlier. karena manusia umur 150-1000 merupakan peristiwa langka yang dimiliki

```
df_cleaned = df.drop(df[(df['age'] == 150) | (df['age'] == 1000)].index)
print(df_cleaned)
```

Before

```
df['age'].unique()
```

```
array([ 54,  45,  58,  42,  57,  43,  37,  49,  55,  39,  53,
        48,  46,  40,  56,  38,  65,  41,  44,  36,  64,  68,
        52,  60,  67,  35,  34,  51,  63,  62,  59,  61,  50,
        66,  47,  70,  69, 150,  33,  32, 1000])
```

After :

```
df= df_cleaned['age'].unique()
print(df)
```

```
[54 45 58 42 57 43 37 49 55 39 53 48 46 40 56 38 65 41 44 36 64 68 52
 67 35 34 51 63 62 59 61 50 66 47 70 69 33 32]
```

Handling Noise

Drop value 0 dan -1 pada kolom heart_rate karena tidak mungkin ada manusia yang memiliki detak jantung 0 dan -1

```
df_cleaned_hr = df_cleaned.drop(df_cleaned[(df_cleaned['heart_rate'] == 0)|
                                           (df_cleaned['heart_rate'] == -1)].index)
```

Before:

```
1 df['heart_rate'].unique()
```

```
array([ 95,  64,  81,  90,  62,  75,  66,  65,  93,  70,  85,  58,  83,
        80,  60,  72,  71, 105,  53,  74,  63,  82,  67,  76,  68,  77,
        69,  55,  87,  86,  52,  79, 100,  78,  88,  48, 104,  92,  84,
        50,  94, 120,  98, 122, 101, 110, 107,  96,  73,  56, 103,  57,
       106,  61, 102,  89, 125,  54,  51,  91, 115,  44,  47,  45, 140,
       108,  59, 143,   0,  -1,  46, 112,  99, 130,  97])
```

After :

```
[ 95  64  81  90  62  75  66  65  93  70  85  58  83  80  60  72  71 105
  53  74  63  82  67  76  68  77  69  55  87  86  52  79 100  78  88  48
 104  92  84  50  94 120  98 122 101 110 107  96  73  56 103  57 106  61
 102  89 125  54  51  91 115  44  47  45 140 108  59 143  46 112  99 130
  97]
```

Handling Inconsistency

Sex

Replace Value m dengan male dan value f
dengan female pada kolom sex
untuk menghasilkan data sex lebih konsisten

```
df['sex'] = df['sex'].replace({'m': 'male', 'f': 'female'})
```

Before :

```
df= df['sex'].unique()
```

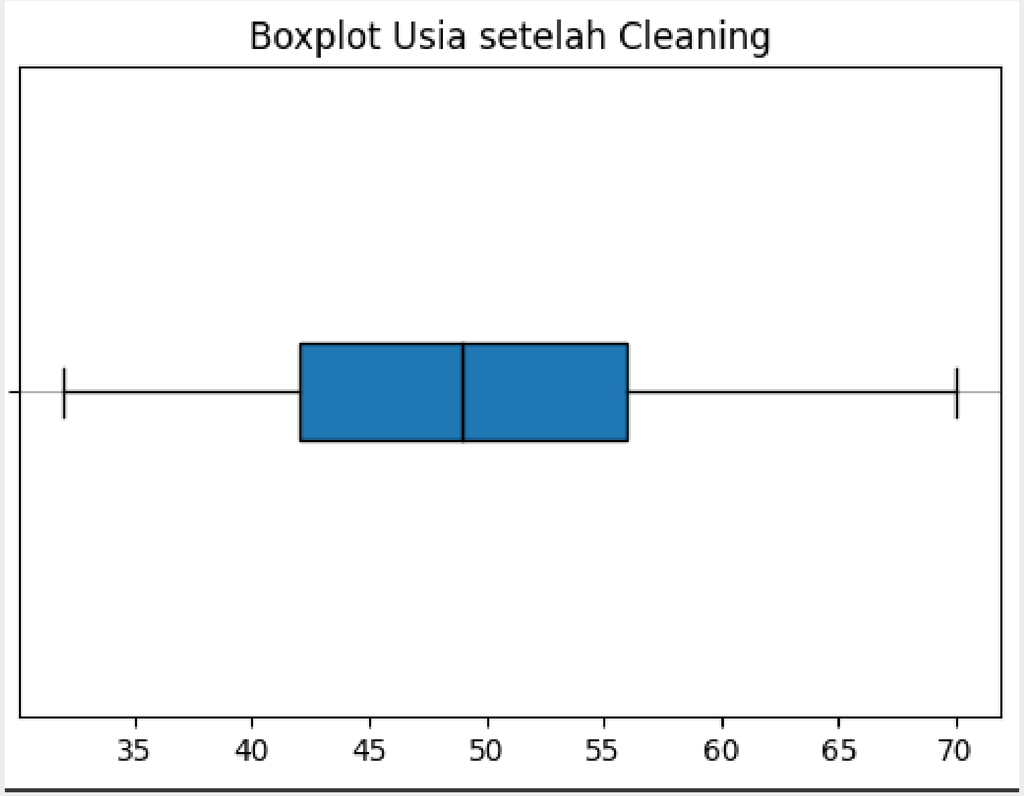
```
array(['male', 'female', 'f', 'm'], dtype=object)
```

After:

```
df['sex'].unique()
```

```
array(['male', 'female'], dtype=object)
```

Handling Outlier



Membuat Klasifikasi Tingkat Kolesterol

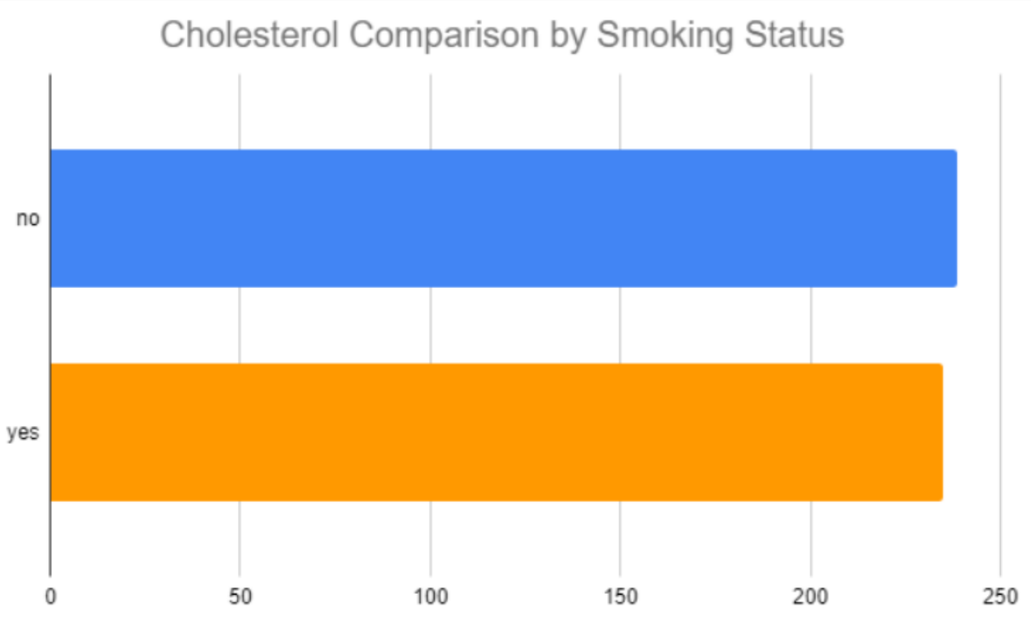
```python
#Fungsi klasifikasi kolesterol
def kolesterol(chol):
    if chol < 200:
        return 'Normal'
    elif chol < 240:
        return 'Cukup Tinggi'
    else:
        return 'Tinggi'

df_5['chol_clasification'] = df_5['chol'].apply(kolesterol)
df_5
```

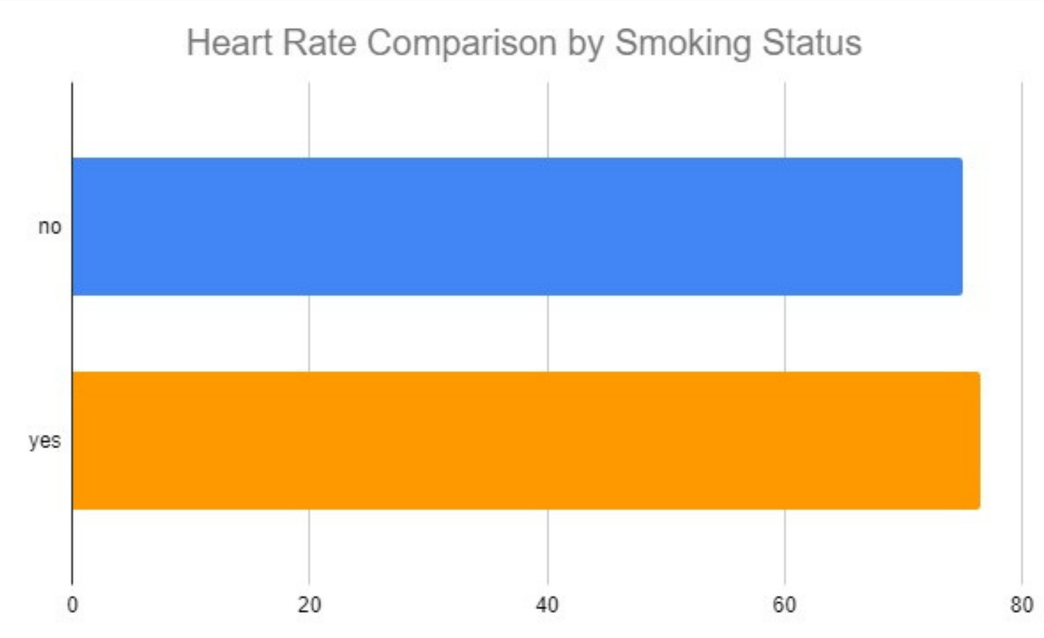
	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol	sistol	diastol	tekanan_darah	chol_clasification
14	48	male	no	75	131/52	0.0	175.0	131.0	52.0	Hipertensi Tingkat 1	Normal
15	58	female	no	75	126/52	0.0	240.0	126.0	52.0	Tinggi	Tinggi
16	46	female	no	80	102/56	0.0	199.0	102.0	56.0	Normal	Normal
17	45	female	no	75	106/58	0.0	235.0	106.0	58.0	Normal	Cukup Tinggi
18	37	female	no	81	112/60	0.0	300.0	112.0	60.0	Normal	Tinggi
...	...	...	...	...	...	...	...	...	...	...	...
3895	37	male	yes	88	122.5/82.5	60.0	254.0	122.5	82.5	Hipertensi Tingkat 1	Tinggi
3896	49	male	yes	70	123/75	60.0	213.0	123.0	75.0	Tinggi	Cukup Tinggi
3897	56	male	yes	70	125/79	60.0	246.0	125.0	79.0	Tinggi	Tinggi
3898	50	male	yes	85	134/95	60.0	340.0	134.0	95.0	Hipertensi Tingkat 1	Tinggi
3899	40	male	yes	98	132/86	70.0	210.0	132.0	86.0	Hipertensi Tingkat 1	Cukup Tinggi

Modeling (EDA)

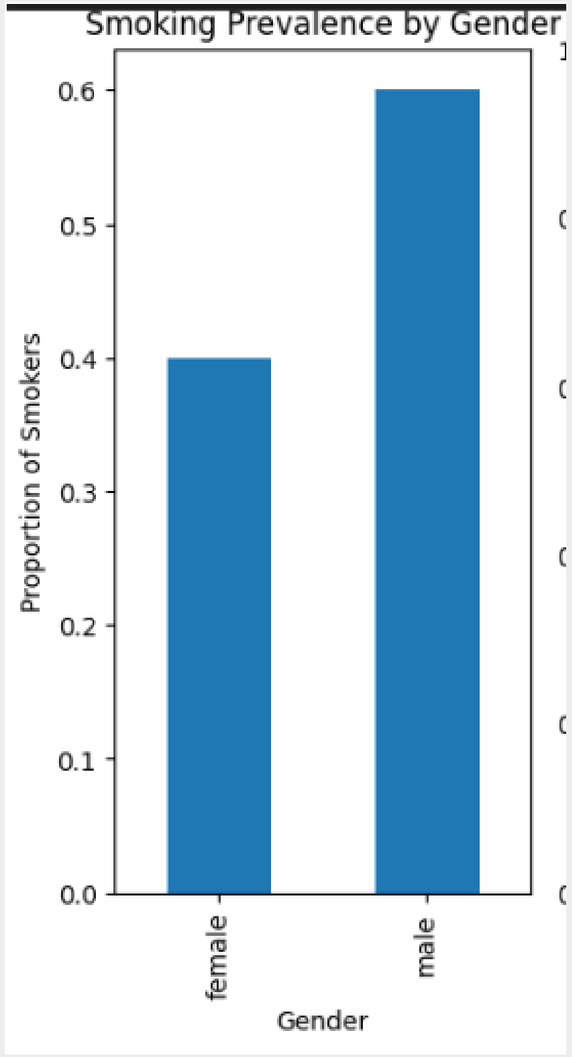
Analisis mendalam dari dataset smoker

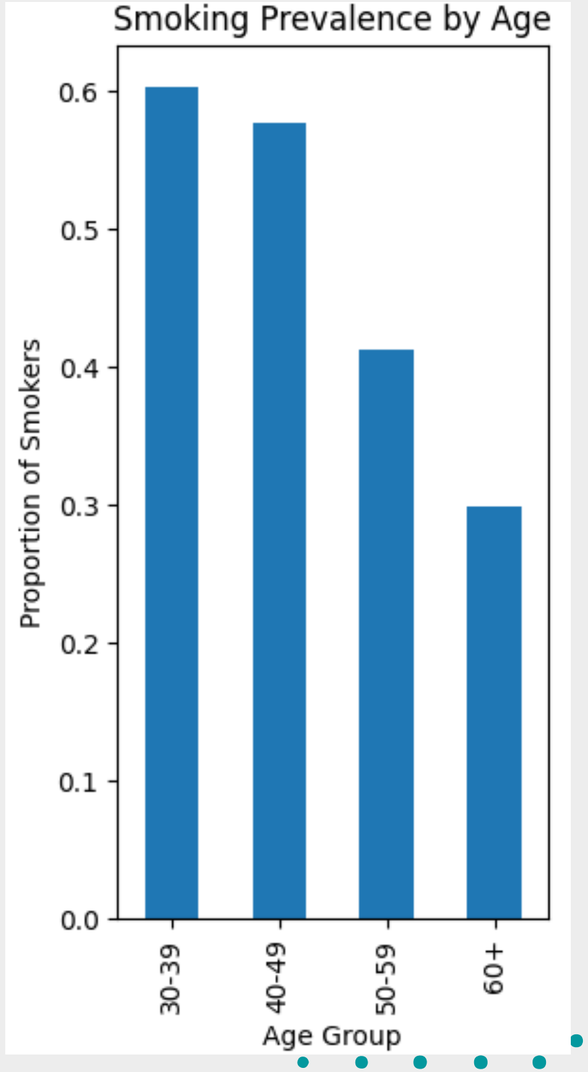
Perbandingan Tingkat Kolesterol & Detak Jantung Berdasarkan Status Merokok





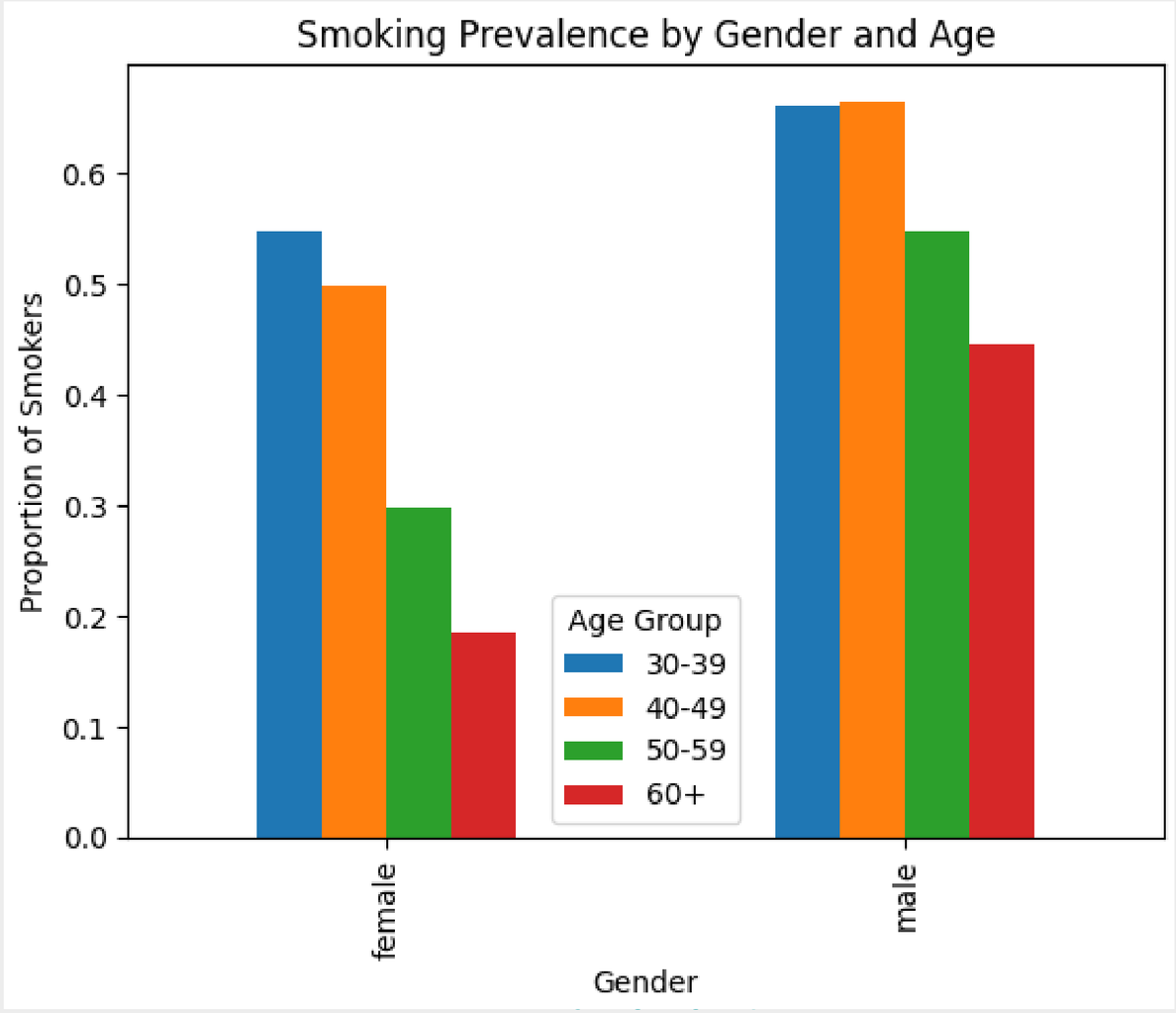
Proporsi Perokok Berdasarkan Gender & Umur



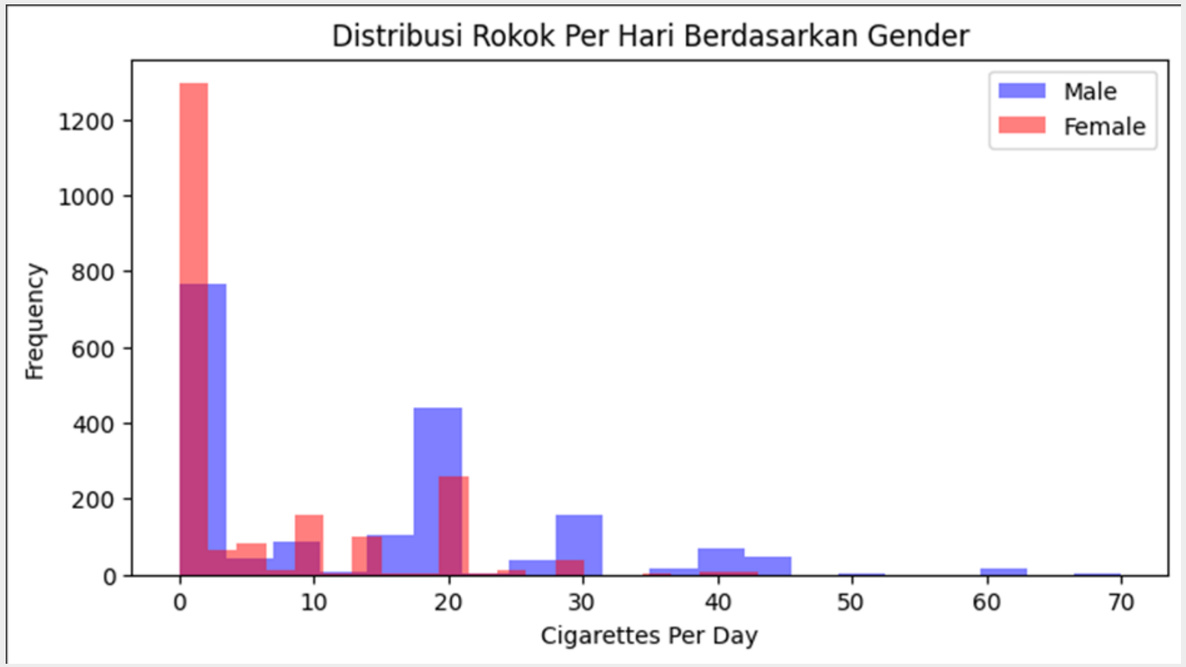


Proporsi laki-laki yang merokok lebih banyak dibandingkan proporsi perempuan yang merokok dan individu yang merokok didominasi oleh kelompok umur 30 -- 49

Proporsi Perokok Berdasarkan Gender & Umur



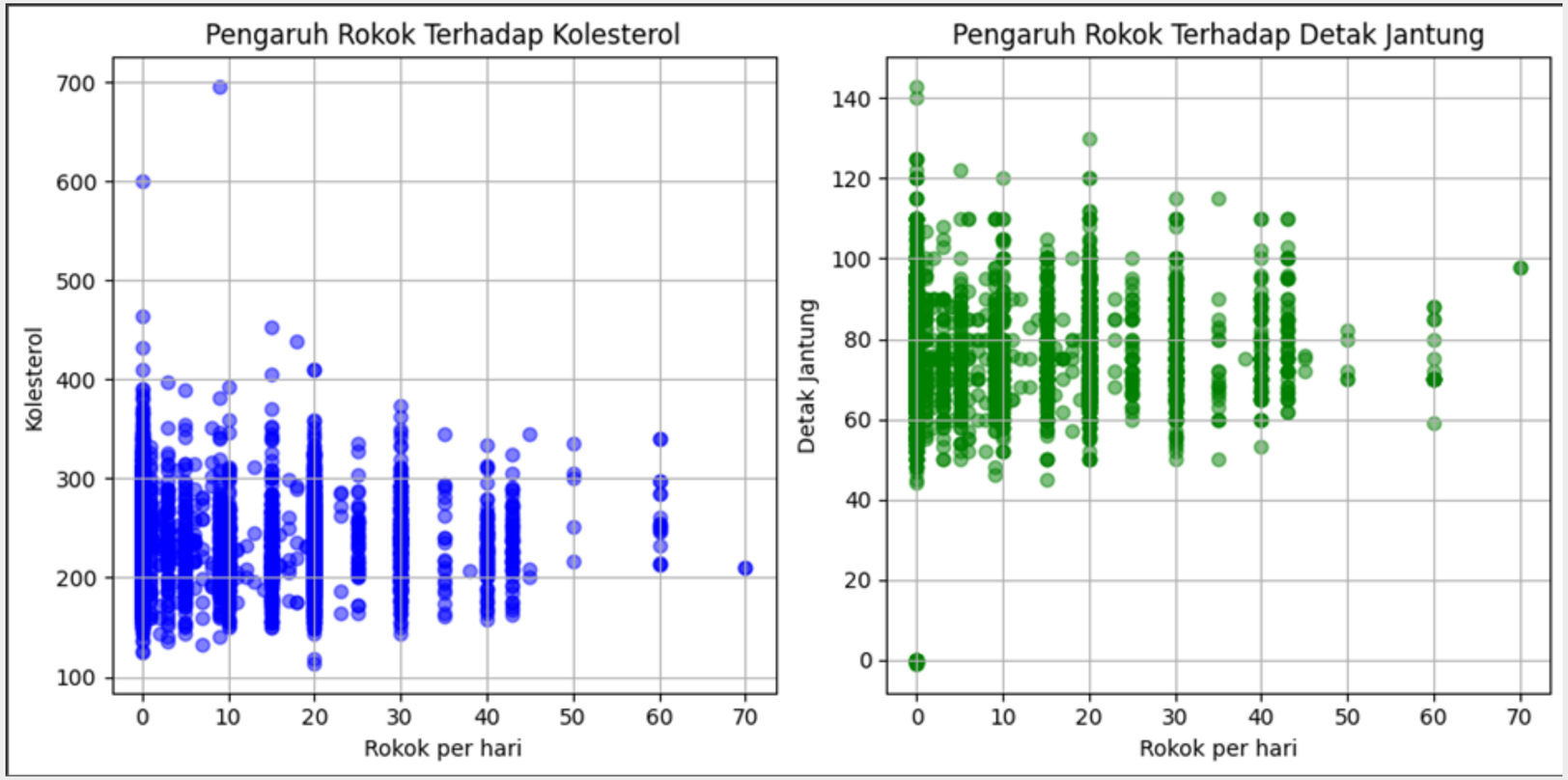
Konsumsi Rokok Per Hari Berdasarkan Gender



Jumlah rokok yang dihabiskan laki-laki terbanyak yaitu 0-20 batang dan ada juga yang menghabiskan 30-40 batang perhari sedangkan jumlah rokok yang dihabiskan oleh perempuan hanyak pa 0-20 batang.

Jumlah rokok yang dihabiskan oleh perempuan lebih sedikit dibandingkan laki-laki

Pengaruh Rokok Pada Kesehatan



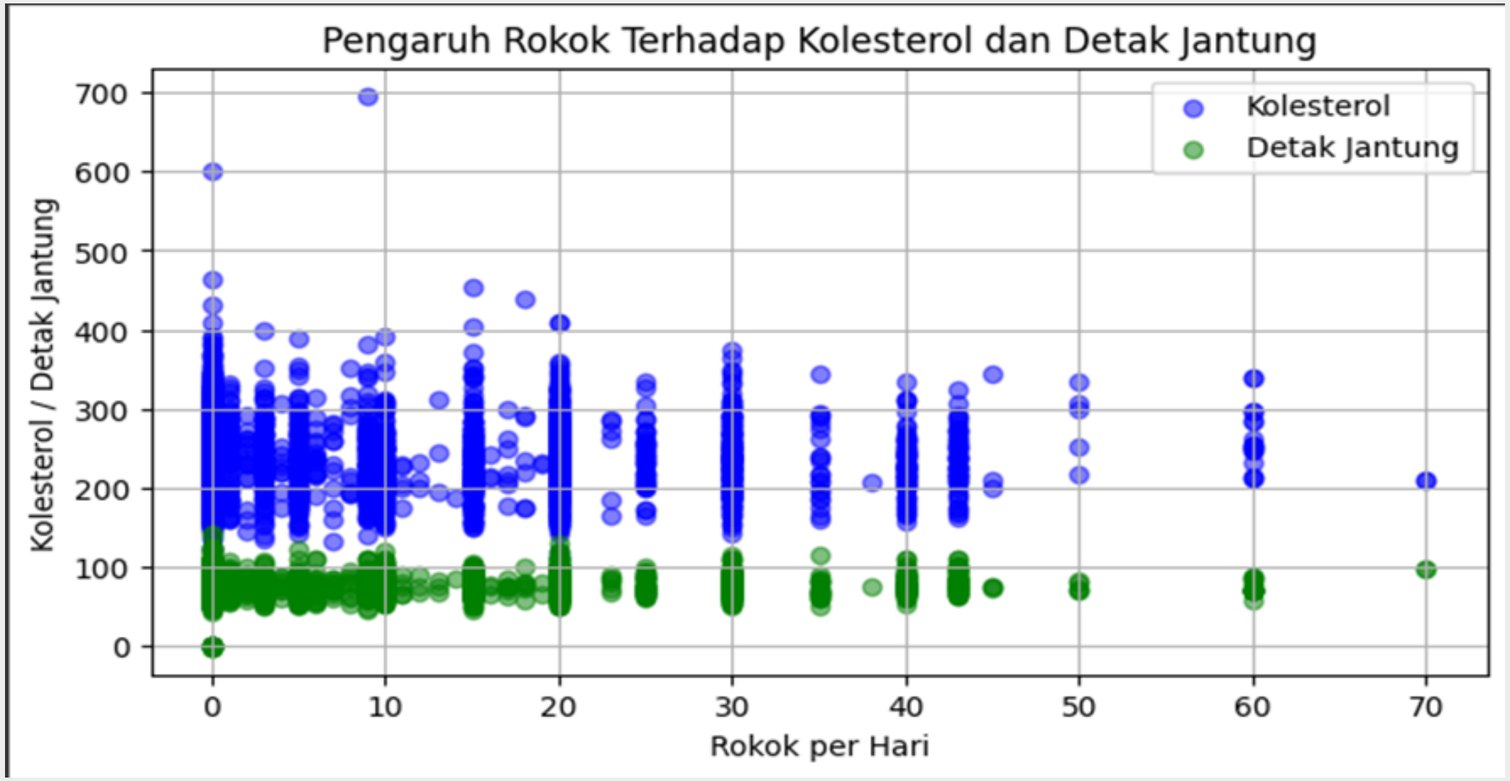
Biru= Tidak Merokok

Biru Muda= Perokok

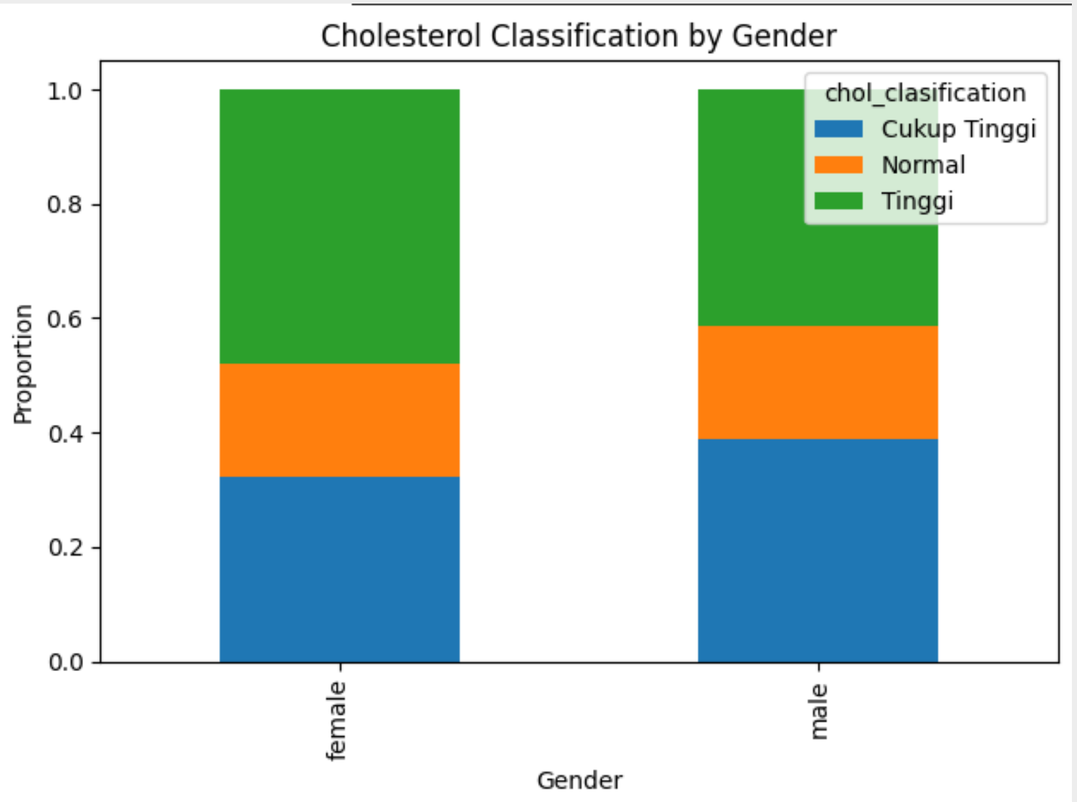
Hijau= Tidak Merokok

Hijau Muda= Perokok

Pengaruh Rokok Pada Kesehatan

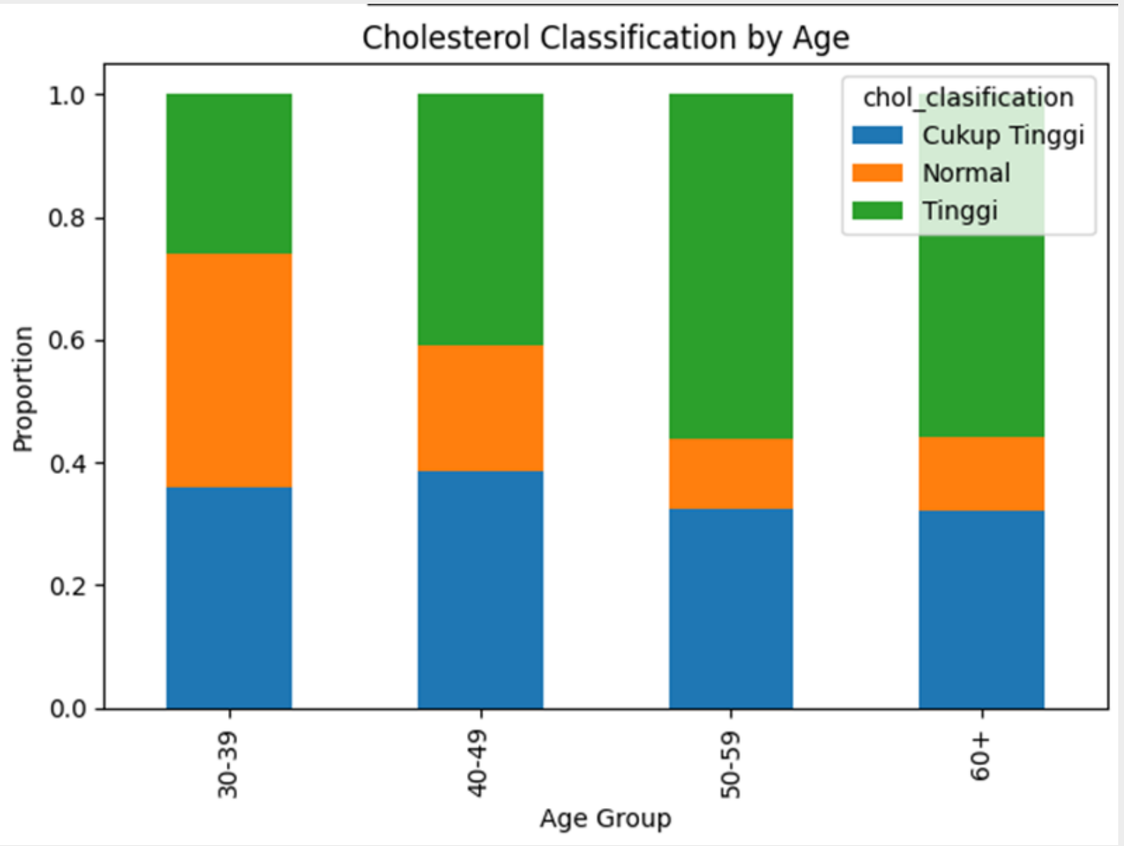


Komposisi Klasifikasi Kolestrol Berdasarkan Gender



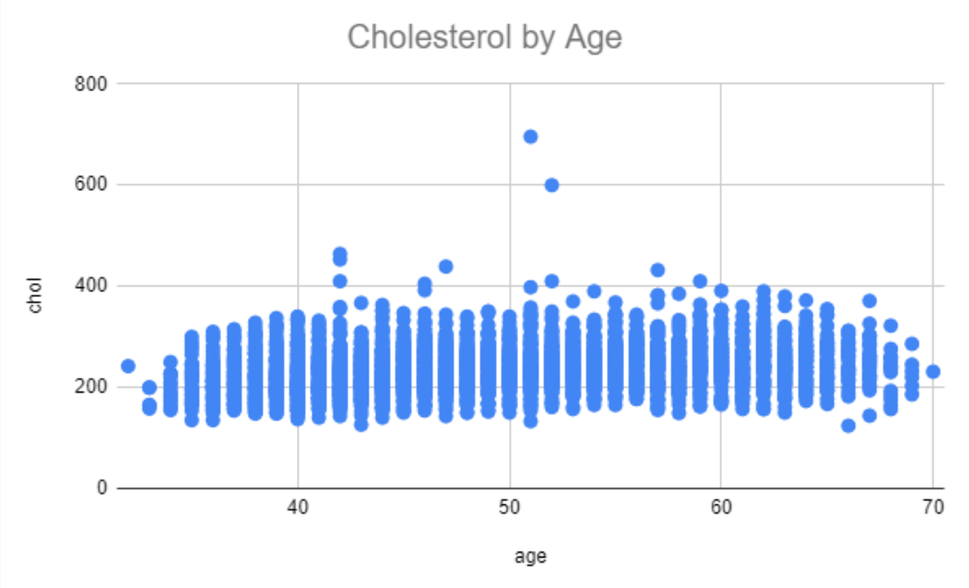
Perempuan memiliki kecenderungan untuk memiliki tingkat kolesterol lebih tinggi dibandingkan laki-laki

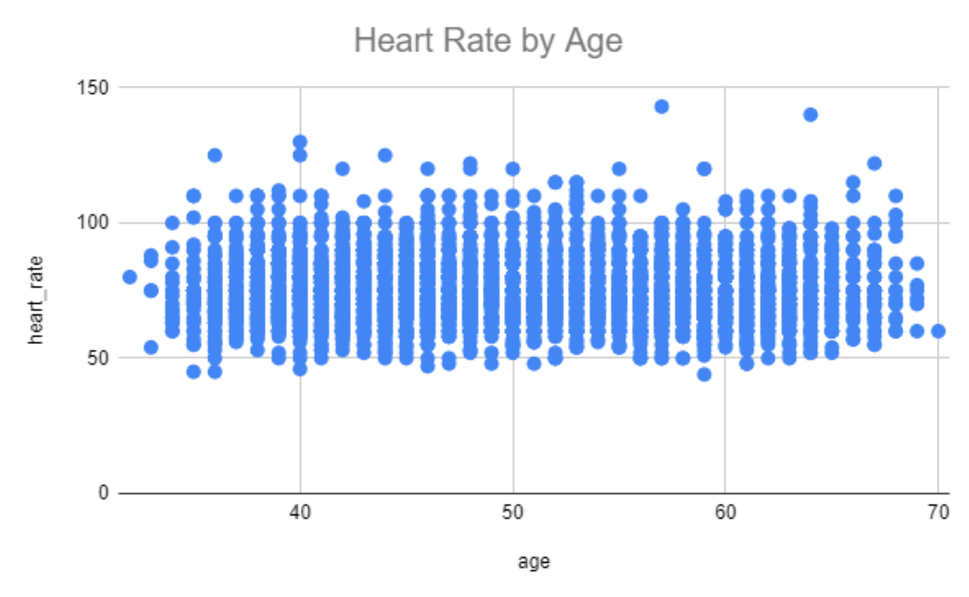
Komposisi Klasifikasi Kolestrol Berdasarkan Umur



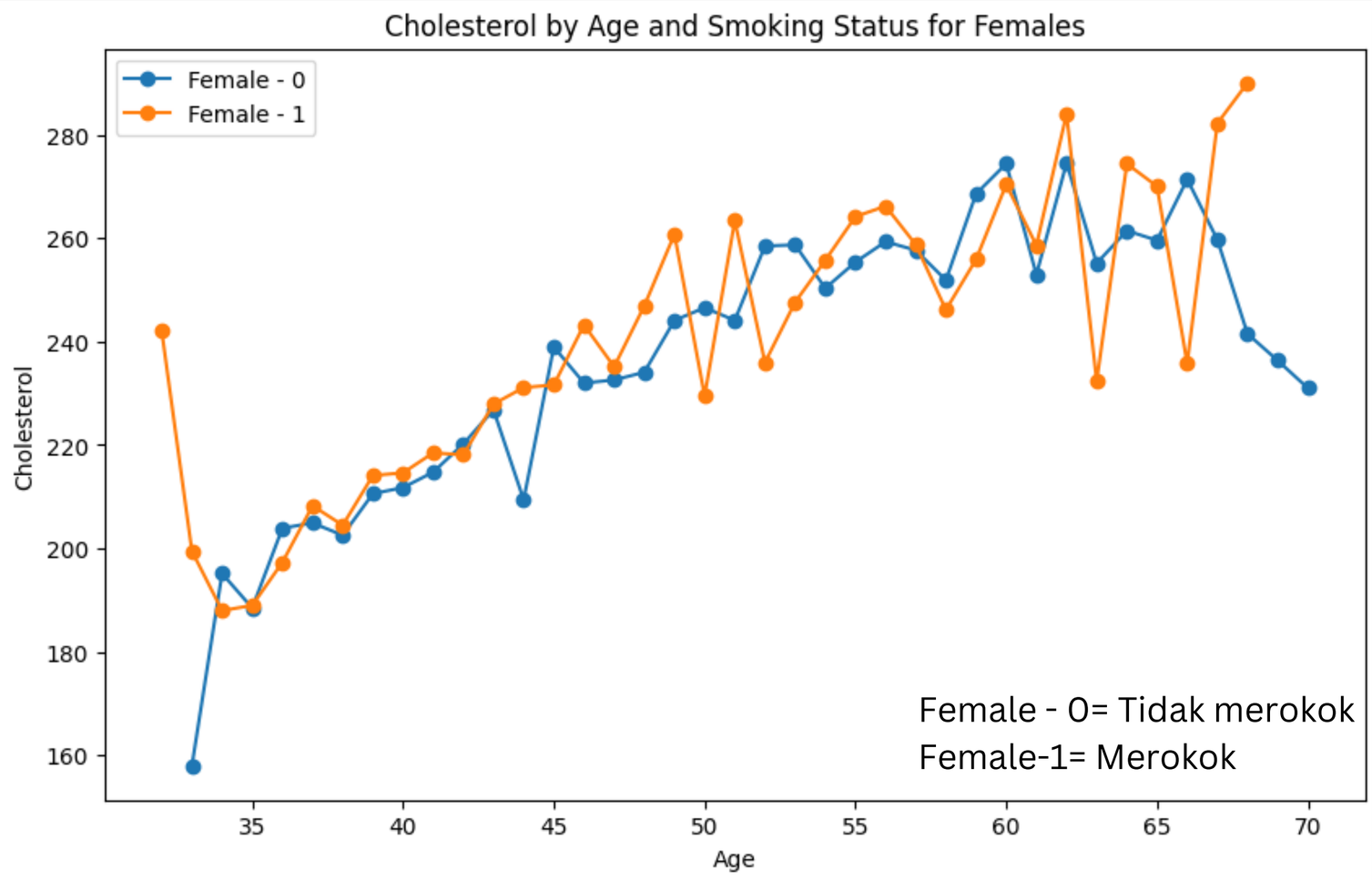
Berdasarkan Grafik Tersebut orang yang rentan terkena penyakit kolestrol yaitu usia diatas 50. Sedangkan umur 30-40 masih belum terlalu rentan untuk terkena penyakit kolestrol

Pengaruh Umur Pada Kolesterol

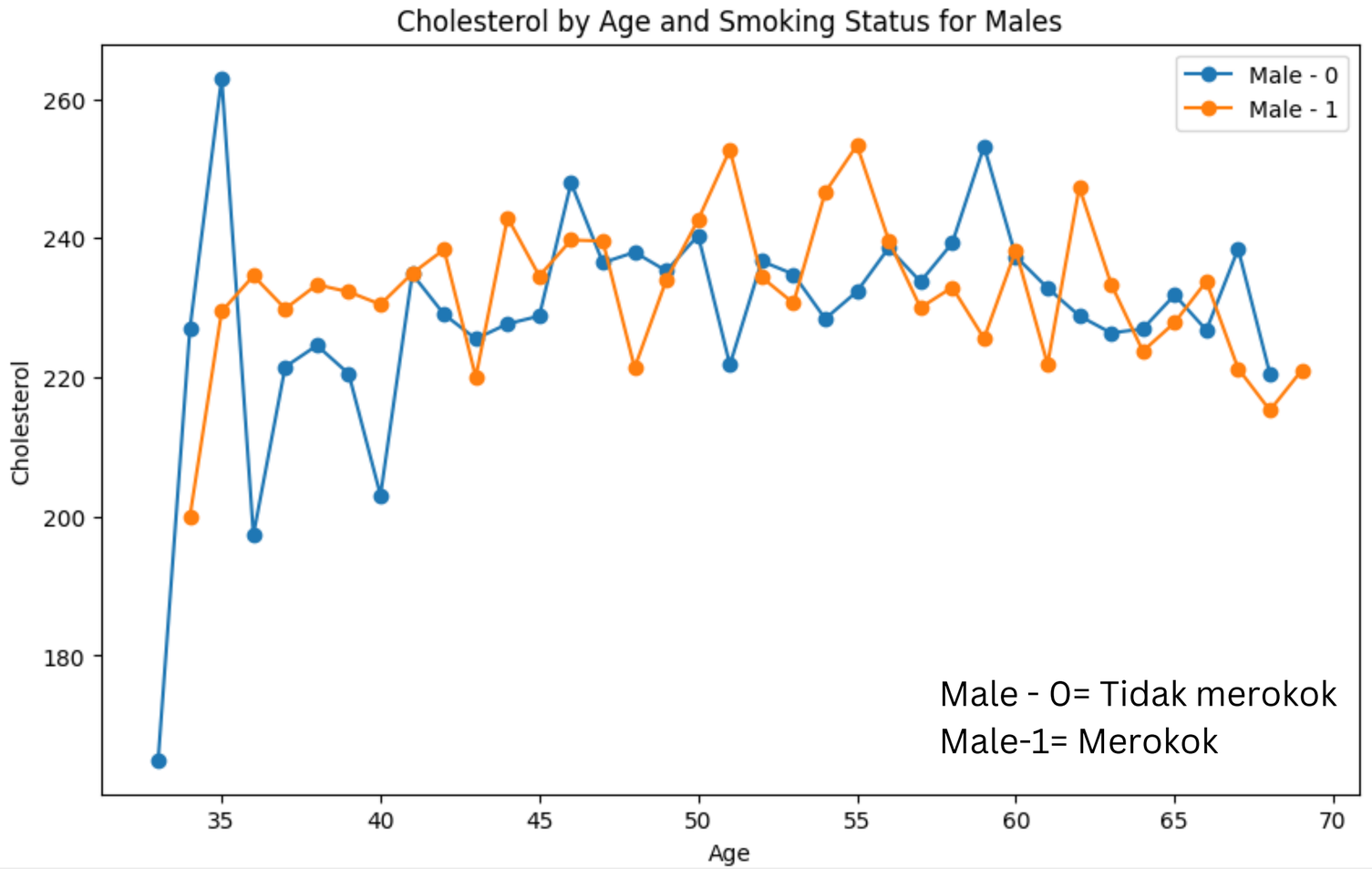




Pengaruh Rokok Pada Kesehatan Female



Pengaruh Rokok Pada Kesehatan Male



MyEduSolve

Kampus Merdeka
INDONESIA JAYA

Kesimpulannya

- Tingkat kolestertol dan detak jantung per menit tidak bisa menjadi indikator untuk menunjukan apakah konsumsi rokok mempunyai pengaruh terhadap risiko kesehatan individu.
- Perempuan dan kelompok umur 50+ lebih rentan untuk memiliki tingkat kolesterol tinggi.
- Komposisi perokok didominasi oleh kelompok umur 30 -- 49.
- Laki-laki yang merokok cenderung akan mengonsumsi rokok per hari lebih banyak dibandingkan perempuan yang merokok.
- Laki-laki memiliki kecenderungan lebih sulit untuk berhenti merokok dibandingkan perempuan.

Rekomendasi

Perlu dilakukan penelitian lebih lanjut dengan menggunakan indikator kesehatan lainnya untuk mendapatkan pemahaman yang lebih komprehensif tentang dampak merokok terhadap kesehatan individu.

Mengadakan kampanye anti-rokok yang dirancang secara khusus untuk menjangkau laki-laki, dikarenakan proporsi perokok laki-laki yang lebih tinggi dibandingkan perempuan.

Program skrining tingkat kolesterol secara teratur harus direkomendasikan untuk individu berusia 50 tahun ke atas dan perempuan, karena mereka lebih rentan terhadap tingkat kolesterol tinggi.