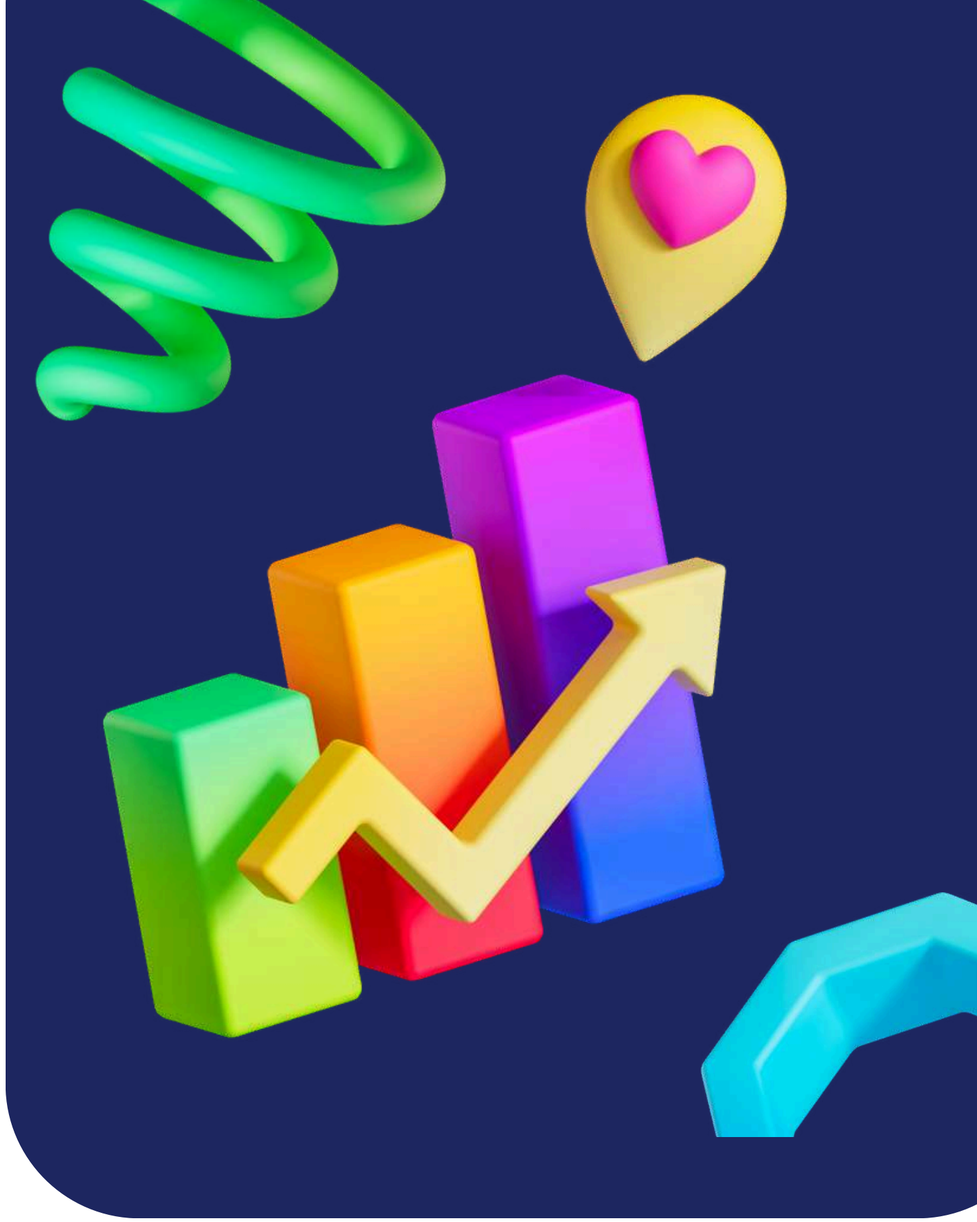
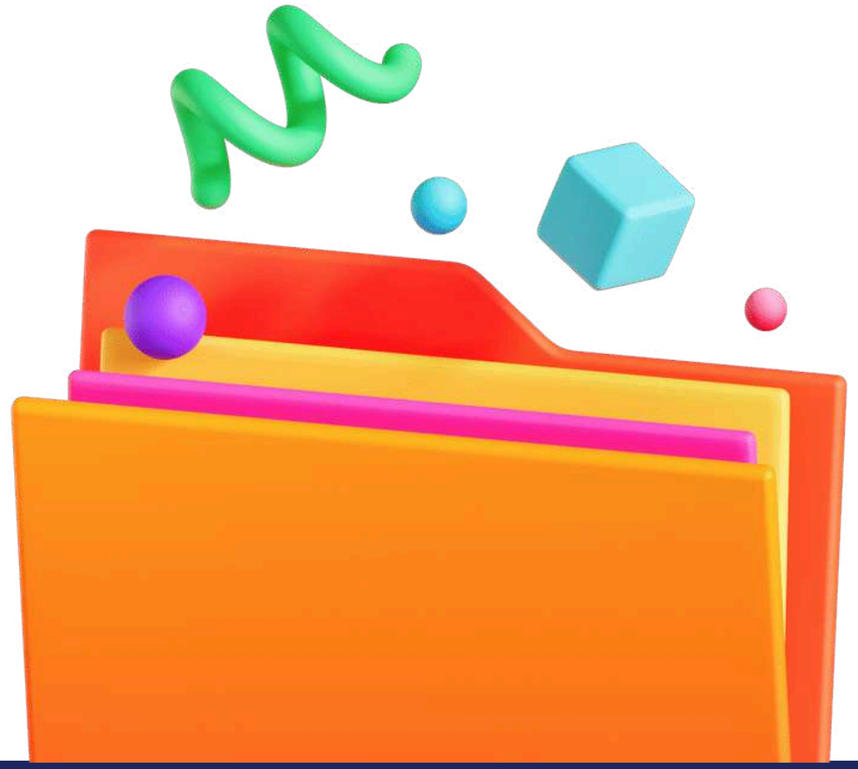
MyEduSolve

Kampus Merdeka
INDONESIA JAYA

CRISP - DM



CONTENTS



MyEduSolve

Kampus Merdeka
INDONESIA JAYA

BUSINESS UNDERSTANDING

Business Understanding

Latar Belakang

Sebuah Instansi Kesehatan melakukan survey terkait kondisi kesehatan masyarakat perokok dan non-perokok untuk mengetahui kondisi kesehatan mereka yang meliputi detak jantung, dan kolesterol sehingga mendapatkan gambaran mengenai kesehatan mereka

Business Questions

1. Apakah ada Kaitan antara gender terhadap kecenderungan merokok?
2. Apakah terdapat perbedaan kondisi kesehatan perokok dan non perokok?
3. Apakah ada Kaitan antara usia perokok terhadap tingkat kolesterol & detak jantung mereka?

Business Understanding

Objective Business

1. Analisis Kaitan Gender terhadap kecenderungan merokok
2. Analisis perbedaan tingkat kesehatan antara perokok dan bukan perokok
3. Analisis Kaitan Usia Perokok terhadap tingkat kolesterol dan detak jantung.

Tujuan Proyek

Melakukan analisis mendalam untuk mengetahui bagaimana perbedaan kondisi kesehatan antara perokok dengan non-perokok.

MyEduSolve

Kampus Merdeka
INDONESIA JAYA

DATA UNDERSTANDING

DATA EXPLORATION

Nama Kolom

```
[ ]  df.columns

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

Data Shape
(Baris , Kolom)

```
[ ]  df.shape

     (3908, 7)
```

Data Size

```
[ ]  df.size

     27356
```

DATA EXPLORATION

```
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
```

- Dari informasi disamping, dapat dilihat secara lengkap mengenai nama kolom, jumlah baris, dan tipe data tiap kolom.
- terdapat missing value pada beberapa kolom seperti **cigs_per_day & chol**

DATA DESCRIPTION

```
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

Deskripsi Kolom data bertipe Numerik

```
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

Deskripsi Kolom data bertipe Kategorikal

MISSING VALUES & DUPLICATE

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

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

Kolom data yang memiliki Missing Value pada dataset adalah **cigs_per_day** & **chol**

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

```
8
```

8 duplikat value ditemukan pada dataset yang dimiliki

UNIQUE VALUE

```
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])
```

```
print(df.sex.unique())
```

```
['male' 'female' 'f' 'm']
```

```
print(df.heart_rate.unique())
```

```
[ 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]
```

```
print(df.current_smoker.unique())
```

```
['yes' 'no']
```

```
print(df.blood_pressure.unique())
```

```
['110/72' '121/72' '127.5/76' ... '153.5/105' '104/73.5' '134/95']
```

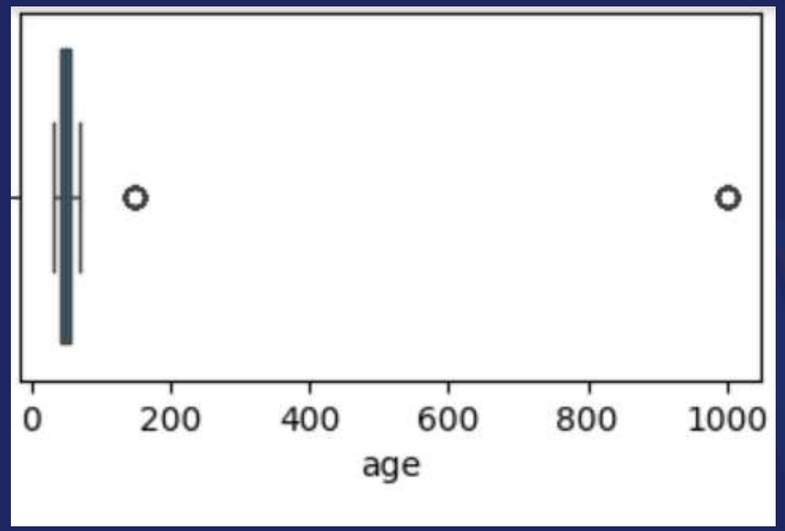
```
print(df.cigs_per_day.unique())
```

```
[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.]
```

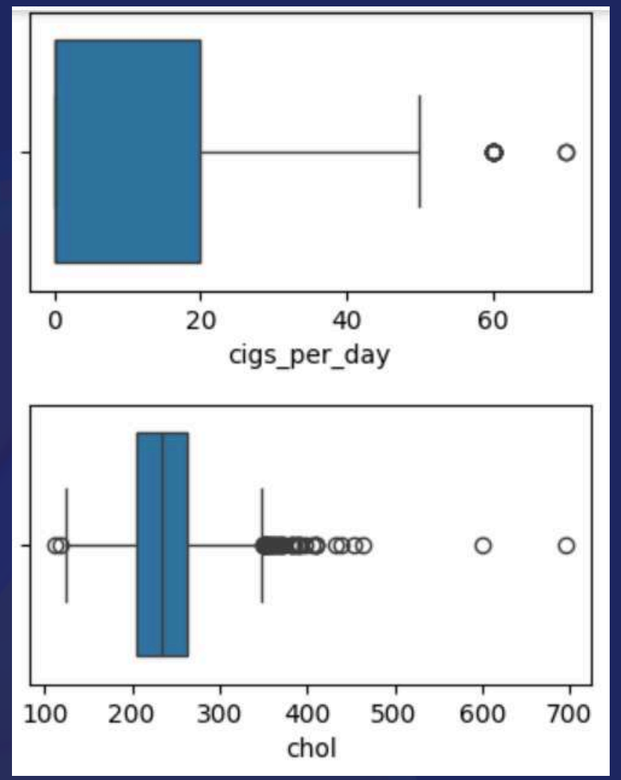
```
print(df.chol.unique())
```

```
[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 2 nilai pada kolom **Age** yang nilainya jauh di atas mayoritas nilai pada kolom tersebut.



- pada kolom **cigs_per_day** terdapat 2 nilai yang berada diatas nilai mayoritas pada kolom tersebut
- pada kolom **chol** terdapat cukup banyak data pencilan yang melebihi upper dan lower bound yang ada.

INCONSISTENT DATA

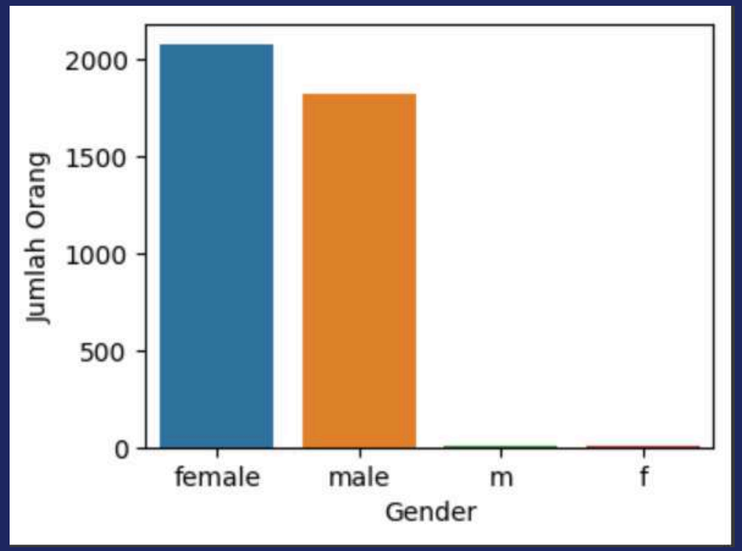
```
df['sex'].unique()
```

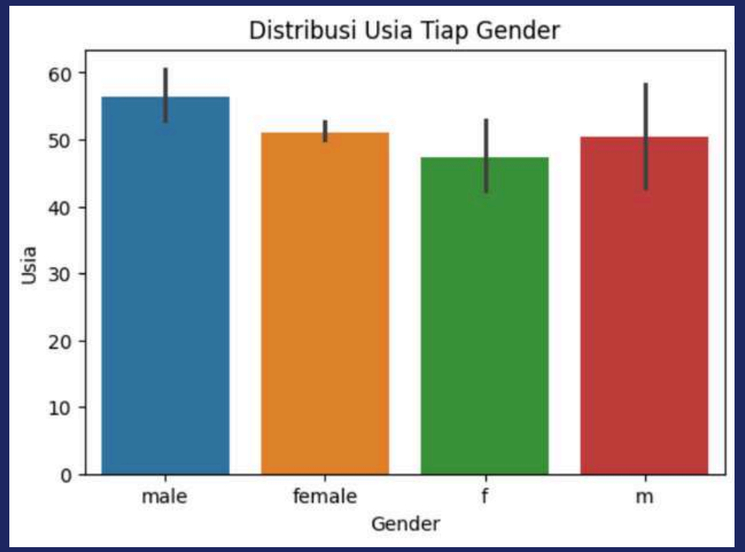
```
array(['male', 'female', 'f', 'm'], dtype=object)
```

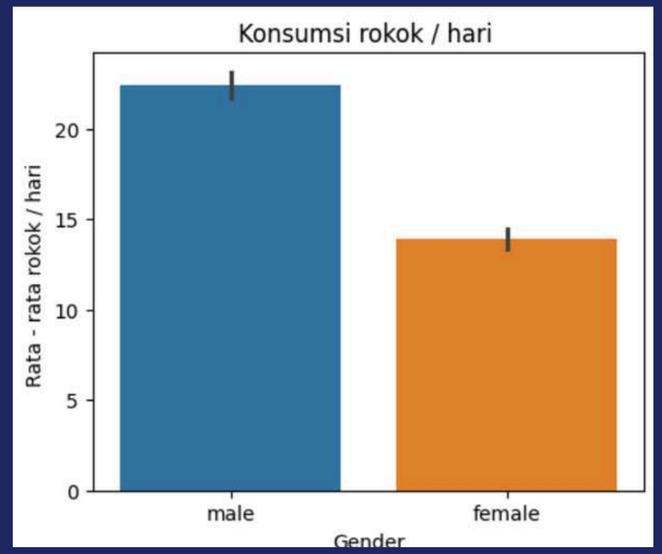
rs baru itu kita cri

Terdapat penulisan yang tidak konsisten pada kolom jenis kelamin dari subjek yang analisa

INITIAL EDA







- Jumlah **perempuan** pada subyek analisa memiliki porsi **terbanyak** dengan rincian female (2076 orang) + f (5 orang) serta Male (1820 orang) + m (7 orang)
- Rata-rata usia male untuk subyek penelitian berusia 56 tahun, female (51 tahun), f (47 tahun), dan m (50 tahun)
- Rata-rata konsumsi rokok perhari bagi subyek analisa dengan status merokok adalah **aktif** diduduki oleh gender **Male** di posisi teratas dengan rata-rata konsumsi 22 batang/hari dan Female 14 batang/hari

DATA PREPARATION

DATA PREPARATION

Handling Missing Value

Cigs Per day

Check nilai skew

```
[8]  df_smooker['cigs_per_day'].skew()

     1.2903488748302245
```

Distribusi tidak normal karena nilai skew lebih besar dari 0,5 sehingga missing value pada kolom **cigs_per_day** akan diisi dengan median

Mengisi dengan nilai median

```
0s  # Mengisi null value dengan median
    med = df_smooker['cigs_per_day'].median()
    df_smooker['cigs_per_day'] = df_smooker['cigs_per_day'].fillna(med)
```

DATA PREPARATION

Handling Missing Value

Check nilai skew

```
[8]  df_smooker['chol'].skew()

     0.8992783738829574
```

Mengisi dengan nilai median

```
[14] # Mengisi null value dengan median
     med1 = df_smooker['chol'].median()
     df_smooker['chol'] = df_smooker['chol'].fillna(med1)
```

Chol Column

Distribusi tidak normal karena nilai skew lebih besar dari 0,5 sehingga missing value pada kolom **chol** akan diisi dengan median

DATA PREPARATION

Missing Value Before & After Filling

Before

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

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

DATA PREPARATION

	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol
3892	59	male	yes	70	153.5/105	60.0	298.0
3893	48	male	yes	70	104/73.5	60.0	252.0
3894	46	male	yes	70	121/82	60.0	285.0
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
3900	59	male	yes	70	153.5/105	60.0	298.0
3901	48	male	yes	70	104/73.5	60.0	252.0
3902	46	male	yes	70	121/82	60.0	285.0
3903	37	male	yes	88	122.5/82.5	60.0	254.0
3904	49	male	yes	70	123/75	60.0	213.0
3905	56	male	yes	70	125/79	60.0	246.0
3906	50	male	yes	85	134/95	60.0	340.0
3907	40	male	yes	98	132/86	70.0	210.0

Handling Duplicate

```
df_smooker.duplicated().sum()
```

```
8
```

Terdapat **8 rows duplicate data** yang perlu dihapus untuk membersihkan data sehingga analisis data yang akan dilakukan menjadi lebih akurat

DATA PREPARATION

Handling Duplicate

Drop Duplicates Data

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

After Handling Duplicate

```
df_smooker.duplicated().sum()
```

```
0
```

```
df_smooker[df_smooker.duplicated(subset=['age', 'sex', 'current_smoker', 'heart_rate', 'blood_pressure', 'cigs_per_day', 'chol'], keep=False)]
```

	age	sex	current_smoker	heart_rate	blood_pressure	cigs_per_day	chol

DATA PREPARATION

Data type correction

Mengubah tipe data **cigs_per_day** dan **chol** karena umumnya konsumsi rokok atau tingkat kolesterol seseorang tidak berbentuk float

```
<class 'pandas.core.frame.DataFrame'>
Index: 3814 entries, 0 to 3887
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             3814 non-null   int64
 1   sex             3814 non-null   object
 2   current_smoker  3814 non-null   object
 3   heart_rate      3814 non-null   int64
 4   blood_pressure  3814 non-null   object
 5   cigs_per_day    3814 non-null   float64
 6   chol            3814 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 238.4+ KB
```

Before

```
<class 'pandas.core.frame.DataFrame'>
Index: 3814 entries, 0 to 3887
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             3814 non-null   int64
 1   sex             3814 non-null   object
 2   current_smoker  3814 non-null   object
 3   heart_rate      3814 non-null   int64
 4   blood_pressure  3814 non-null   object
 5   cigs_per_day    3814 non-null   int64
 6   chol            3814 non-null   int64
dtypes: int64(4), object(3)
memory usage: 238.4+ KB
```

After

DATA PREPARATION

Handling Inconsistent Data

Mengubah data **f** menjadi female dan **m** menjadi male

Before

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

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

After

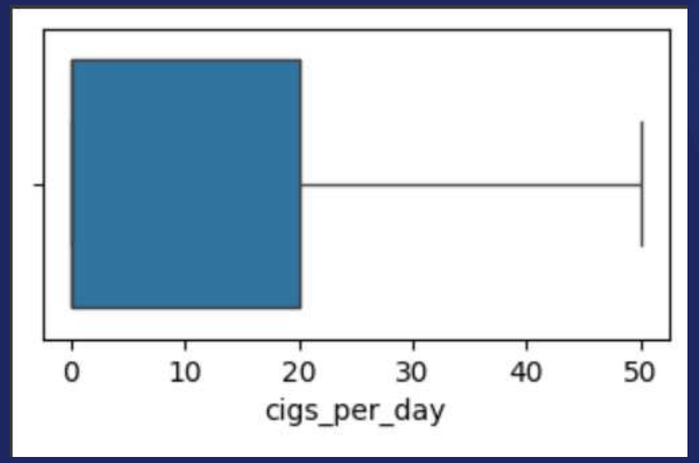
```
df['sex'] = df['sex'].replace({'f': 'female', 'm': 'male'})
```

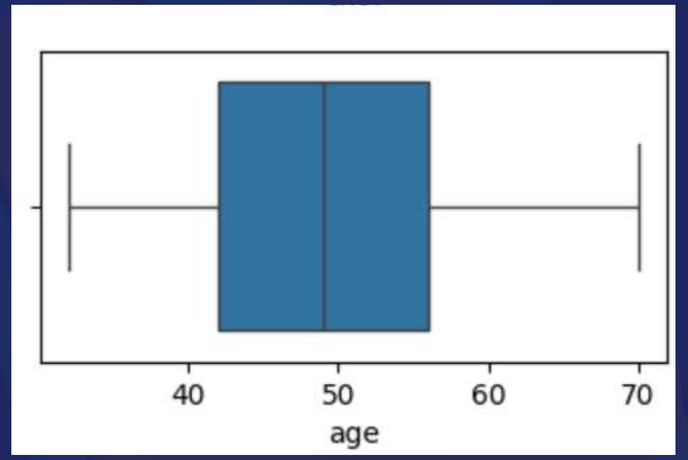
```
df['sex'].unique()
```

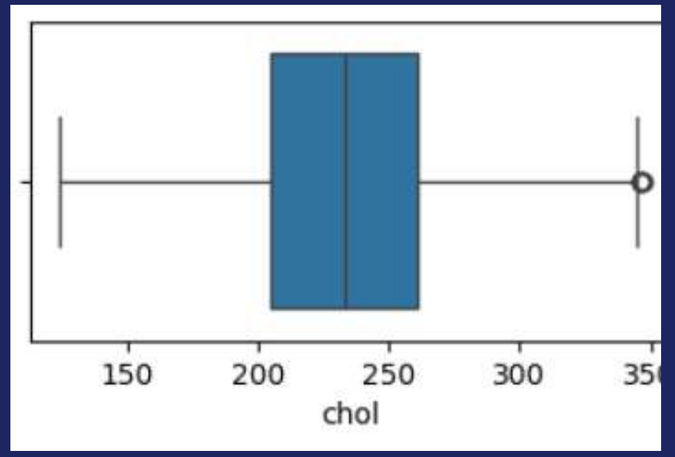
```
array(['male', 'female'], dtype=object)
```

DATA PREPARATION

Handling Outlier







Melakukan Drop pada Outlier data yang dimiliki karena jumlahnya yang tidak terlalu banyak dan untuk memberikan hasil analisa yang lebih akurat dan terfokus

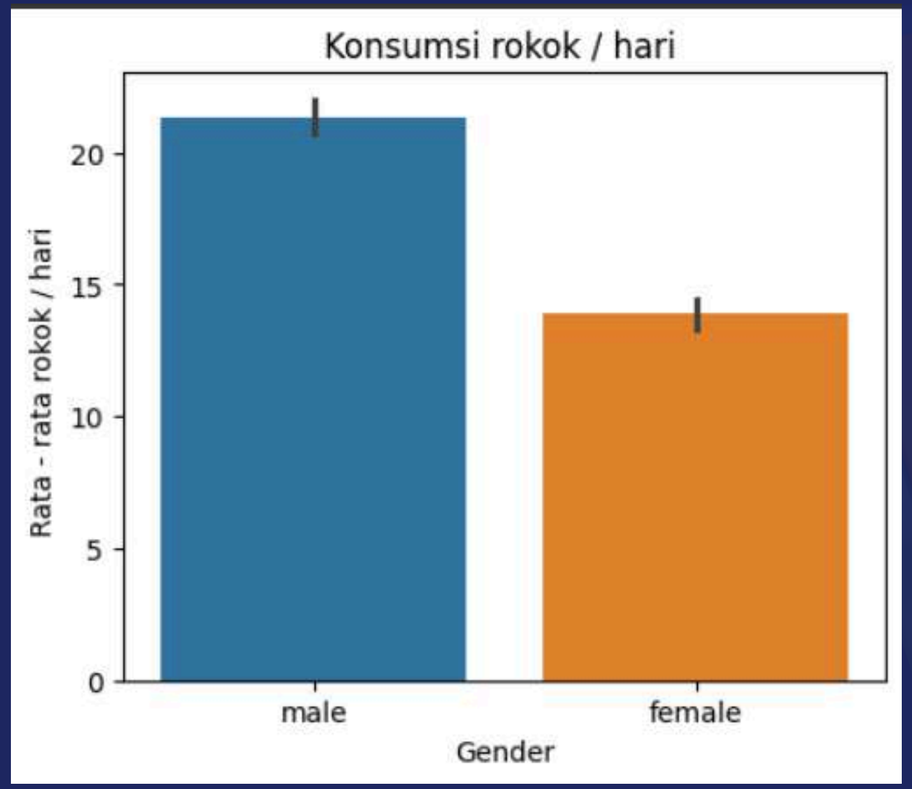
MyEduSolve

Kampus Merdeka
INDONESIA JAYA

MODELING (EDA)

MODELING (EDA)

Apakah ada kaitan antara gender terhadap kecenderungan merokok?

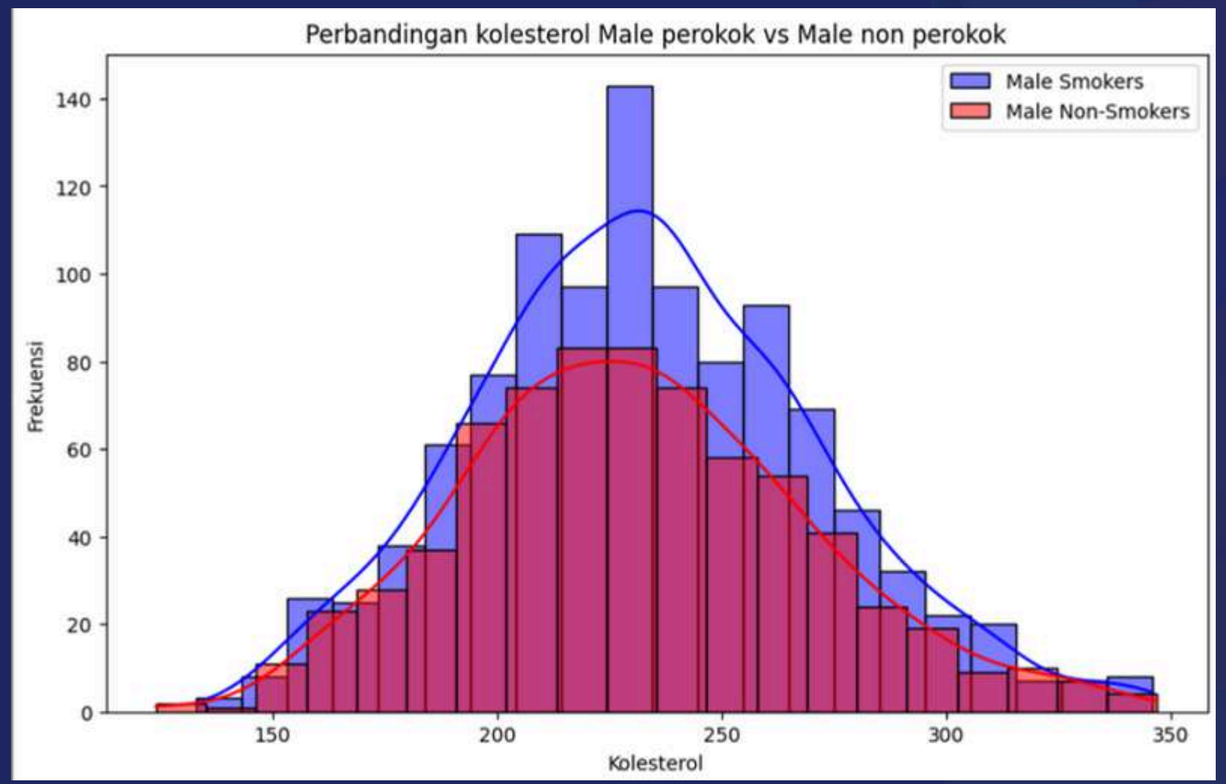



```
Rata-rata cigs_per_day berdasarkan gender:
sex
female    13.881628
male      21.349251
Name: cigs_per_day, dtype: float64
```

Dari rata-rata konsumsi rokok perharinya, perokok male mengonsumsi rokok lebih banyak dibanding female yaitu 21 buah/hari sedangkan rata-rata konsumsi rokok female adalah 14 buah/hari

MODELING (EDA)

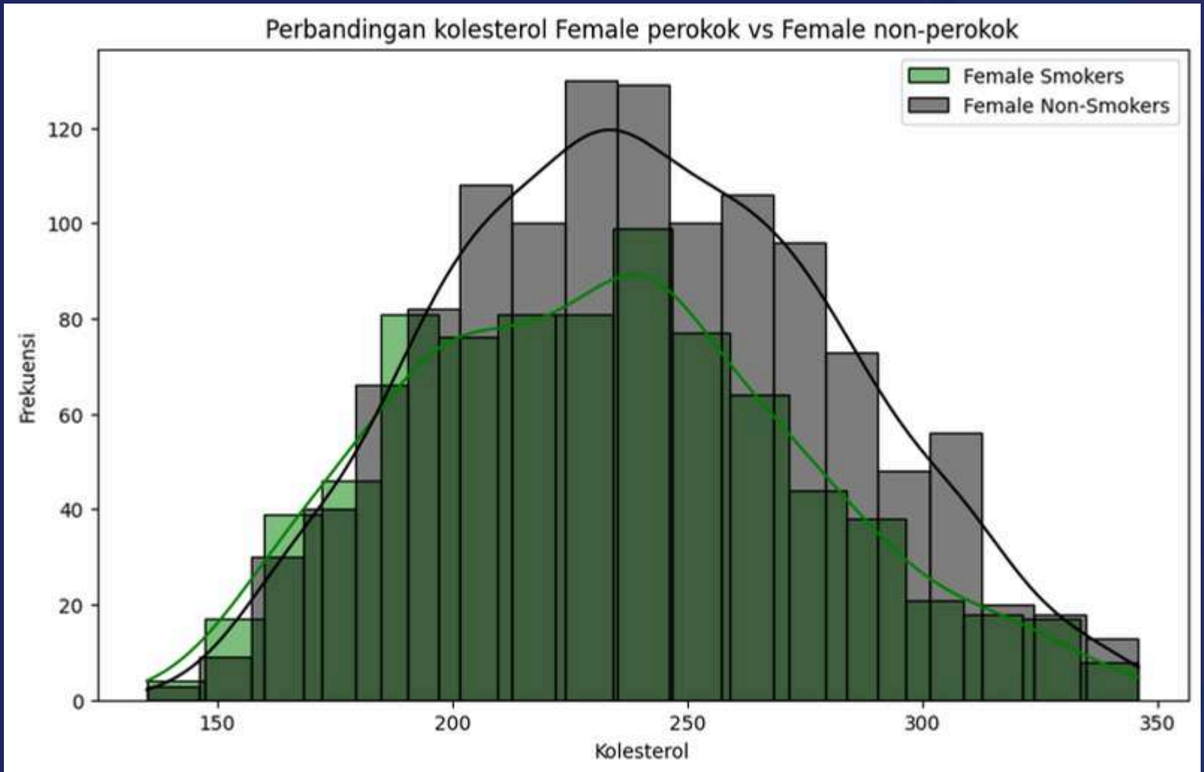
Apakah terdapat perbedaan kondisi kesehatan perokok dan non perokok?



Dari sisi kolesterol, terlihat bahwa tingkat kolesterol dari male perokok aktif dan male non-perokok memiliki tingkat kolesterol yang relatif hampir sama. namun, jumlah dari male perokok dengan kolesterol tinggi cenderung lebih banyak dibanding non-perokok. sehingga dapat diasumsikan bahwa rokok memiliki kaitan dengan tinggi rendahnya kolesterol.

MODELING (EDA)

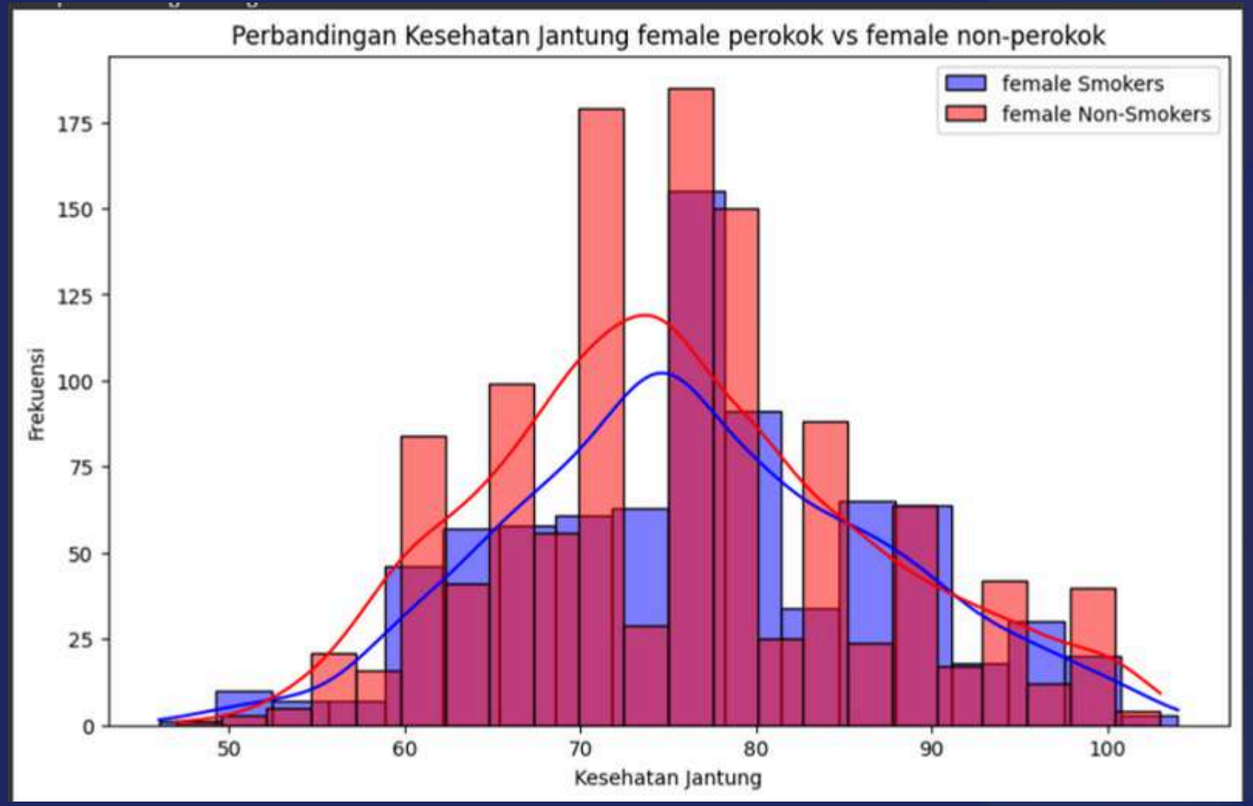
Apakah terdapat perbedaan kondisi kesehatan perokok dan non perokok?



Untuk subyek female, terlihat bahwa tingkat kolesterol dari female perokok aktif dan female non-perokok tidak memiliki perbedaan yang terlalu signifikan. namun, jumlah female non-perokok justru memiliki frekuensi terbanyak dari segi kolesterol tinggi .

MODELING (EDA)

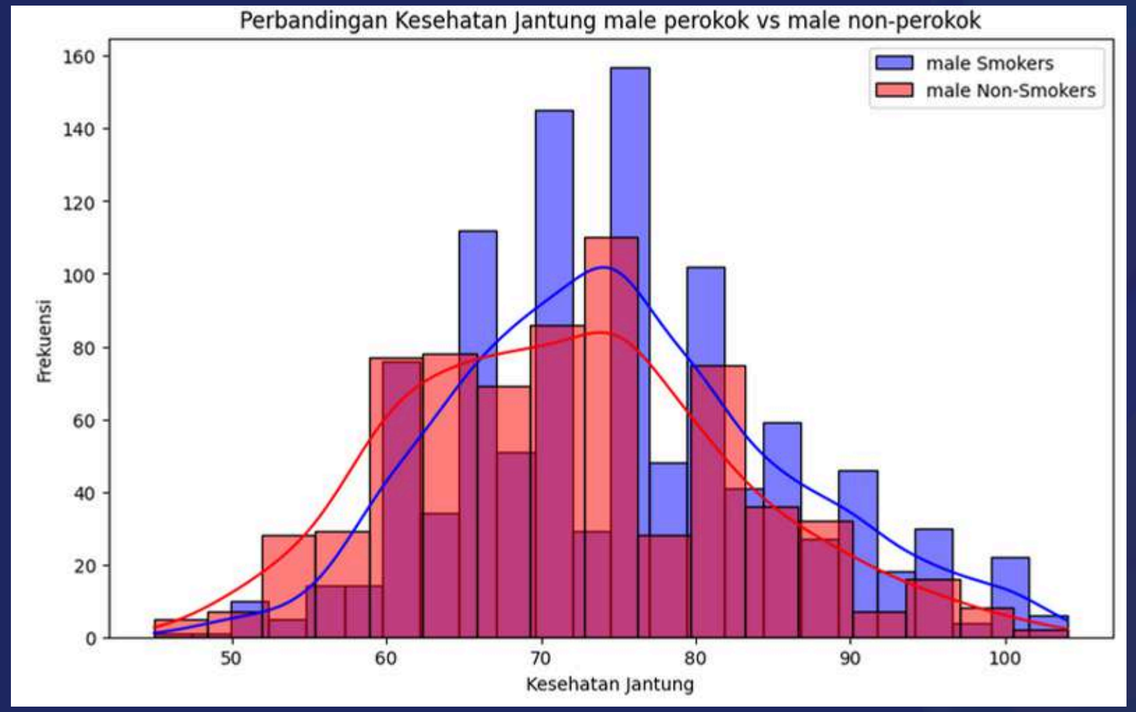
Apakah terdapat perbedaan kondisi kesehatan perokok dan non perokok?



Untuk subyek female, detak jantung dari female perokok cenderung berada diangka normal, artinya kemungkinan rokok kurang berkaitan dengan detak jantung seseorang.

MODELING (EDA)

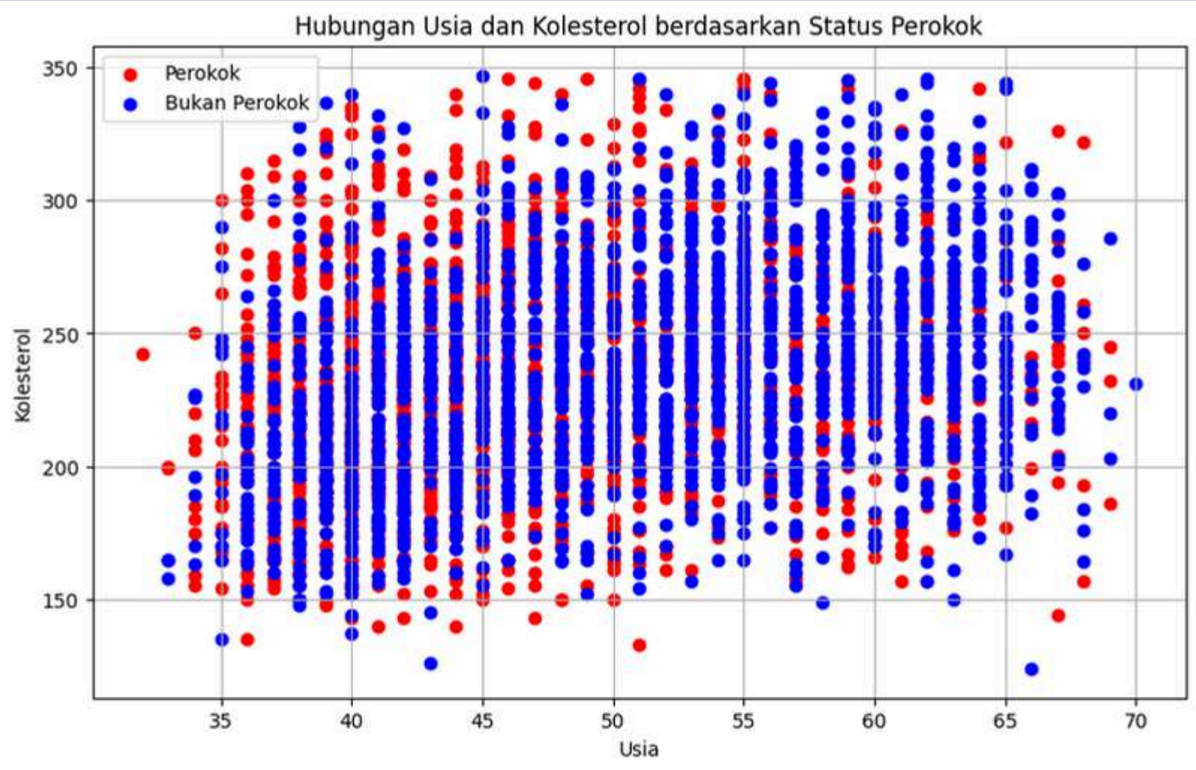
Apakah terdapat perbedaan kondisi kesehatan perokok dan non perokok?



Untuk subyek male, dapat dilihat dari chart disamping menyatakan bahwa status perokok dan tidak itu kurang berkaitan dengan tingkat detak jantung seseorang dikarenakan pada detak jantung normal (70-80) justru lebih banyak orang dengan status perokok.

MODELING (EDA)

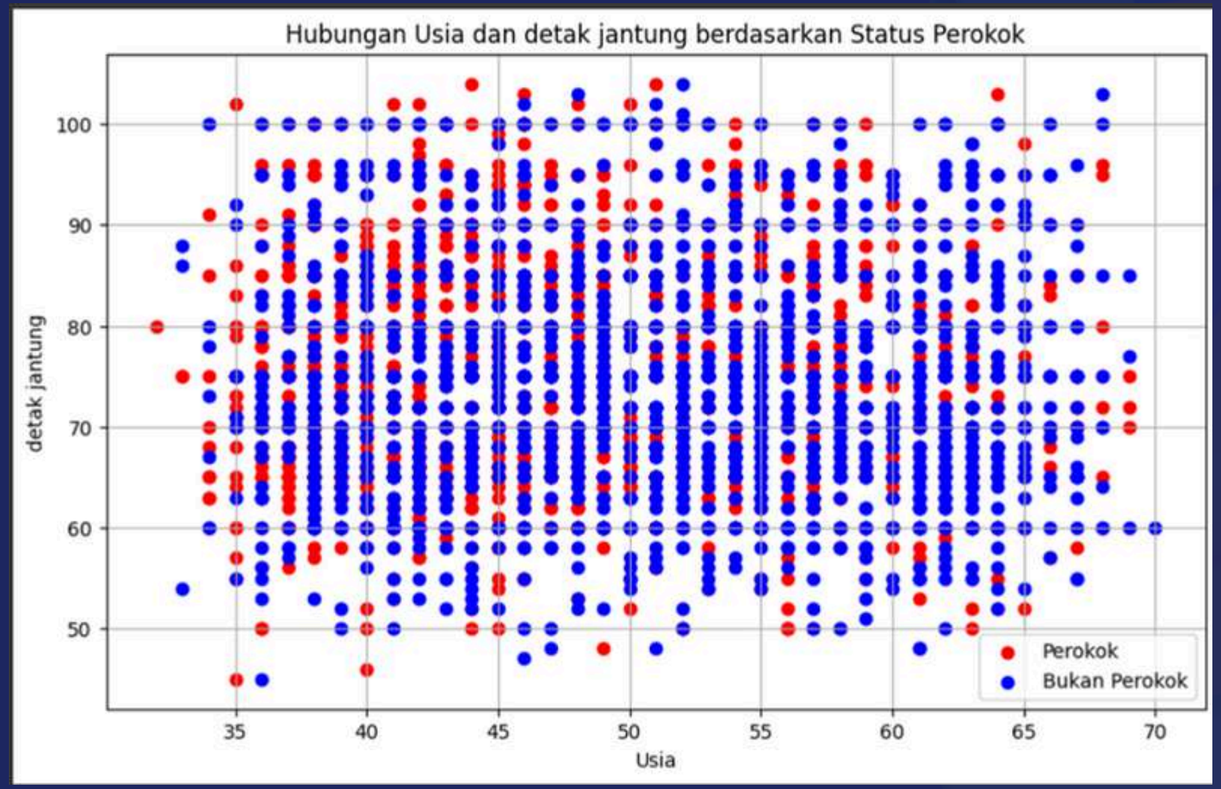
Apakah terdapat kaitan antara usia perokok dan tingkat kolesterolnya



Grafik ini menunjukkan bahwa usia dan kolesterol memiliki korelasi positif (korelasi = 0.3). Dapat dilihat bahwa usia memiliki kaitan dengan kolesterol, di mana semakin tinggi umur seseorang maka semakin rentan meningkatnya kadar kolesterolnya.

MODELING (EDA)

Apakah terdapat kaitan antara usia perokok dan detak jantungnya



Grafik ini menunjukkan bahwa usia dan detak jantung memiliki korelasi negatif (korelasi = - 0.1). Dapat dilihat bahwa usia memiliki kaitan dengan detak jantung, di mana semakin tinggi umur seseorang maka semakin rendah pula detak jantung/ menit yang dimilikinya.

EXECUTIVE SUMMARY

Rangkuman dari analisis pada dataset Smoker_dataset adalah sebagai berikut :

- Male menempati urutan **teratas** dengan rata - rata konsumsi rokok paling banyak per harinya sekitar 21 buah/hari sedangkan female hanya mengonsumsi rokok rata - rata 14 buah/hari.

- Faktor **Usia** berkaitan dengan tingkat kolestrol seseorang (korelasi = 0.3), di mana **semakin tinggi usia** seseorang maka **semakin rentan meningkatnya kadar kolesterol. Namun,** rokok juga memiliki kaitan terhadap tingkat kolesterol, di mana pada usia muda tingkat kolesterol perokok lebih tinggi daripada orang yang tidak merokok

- **Usia** juga berkorelasi dengan detak jantung seseorang(korelasi = -0.1) , di mana **semakin tinggi usia** seseorang maka **semakin rendah** detak jantung yang dimilikinya.

- **Dari sisi Kolesterol & detak jantung,** Tingkat kolesterol perokok pada gender male memiliki frekuensi yang tinggi dibandingkan dengan male non-perokok namun keduanya memiliki rata - rata kolesterol yang hampir sama diangka 230(tidak ada perbedaan signifikan). Sedangkan pada female frekuensi kolesterol tinggi justru dimiliki oleh **female non perokok**, Hal ini bisa saja disebabkan oleh faktor lain yang menyebabkab tingkat kolesterol pada non-perokok tinggi seperti dari segi makanan. status perokok atau tidak **kurang berkaitan** dengan detak jantung seseorang, karena pada data menunjukkan seseorang dengan status perokok justru cenderung memiliki detak jantung diangka normal.

- **Kesimpulannya** adalah **peningkatan usia** memiliki korelasi dengan kolesterol dan detak jantung seseorang, di mana saat **usia semakin tinggi** maka kolesterol juga cenderung akan meningkat serta detak jantung akan semakin melemah. faktor ia sebagai perokok aktif juga turut andil dalam **peningkatan** kolesterol seseorang namun **kurang berkaitan** dengan detak jantung yang dimilikinya namun dalam konteks kolesterol & detak jantung, **tidak ada perbedaan kesehatan yang signifikan** antara perokok dan non perokok.

REKOMENDASI

berikut adalah rekomendasi dari analisis yang telah dilakukan :

- **Kampanye Edukasi Kesehatan** oleh departemen kesehatan yang bertujuan untuk meningkatkan kesadaran masyarakat tentang pentingnya menjaga kesehatan kardiovaskular seiring bertambahnya usia. Kampanye ini harus mencakup informasi tentang risiko kolesterol tinggi, merokok, dan perubahan detak jantung yang terkait dengan usia.

- **Mengajak media massa** untuk berpartisipasi dalam penyebaran informasi dan pesan kesehatan yang berkaitan dengan merokok dan kesehatan kardiovaskular.

- **Pemeriksaan Kesehatan Berkala** dengan memperluas akses masyarakat terhadap pemeriksaan kesehatan berkala, terutama bagi orang-orang yang berusia lanjut. Departemen kesehatan dapat bekerja sama dengan pusat kesehatan lokal untuk menyediakan layanan pemeriksaan kolesterol dan detak jantung yang mudah diakses dan terjangkau.

- **Memberi saran kepada pemerintah setempat** untuk membuat aturan pembatasan merokok di tempat umum dan lingkungan kerja untuk mengurangi paparan asap rokok bagi perokok pasif.

Attachment Details

Dataset

Google Collab

Thank You Everyone

Kelompok 1 DA - MES