



# Muhammad Rafi Rizanda

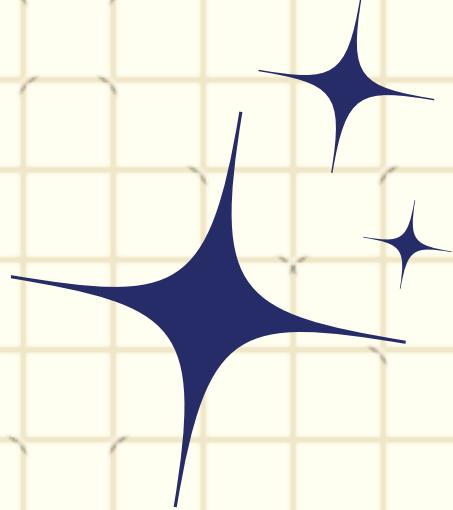
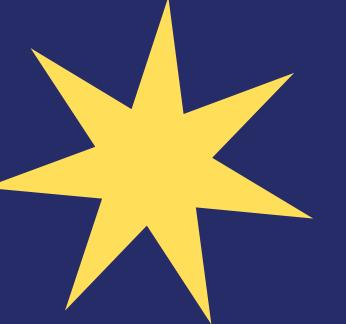
## Data Science Track

Startup Campus



Canva

# Material



Data Preprocessing meliputi :

1. Feature Engineering
2. Exploratory Data Analysis (EDA)
3. Categorical and Numerical Data Analysis
4. Automated Exploratory Data Analysis
5. Reproducible Data Analysis
6. Data Storytelling and Communication

**Feature engineering** adalah proses pemilihan, manipulasi, dan transformasi data mentah menjadi fitur yang dapat digunakan. Feature engineering juga membantu merepresentasikan masalah mendasar ke model prediktif dengan cara yang lebih baik, yang sebagai hasilnya, meningkatkan akurasi model untuk data yang tidak terlihat.

**Exploratory Data Analysis (EDA)** adalah sebuah proses menganalisa dan menyelidiki sebuah dataset untuk melihat karakteristik utamanya. EDA biasa digunakan untuk melihat insight yang mungkin tidak akan terlihat dalam proses modeling dan akan memberikan pemahaman yang lebih baik dari dataset tersebut.

# MEAN, MEDIAN, MODUS

Mean, median, dan modus merupakan tiga ukuran tendensi sentral yang artinya menggambarkan pusat suatu sebaran data. Semuanya dihitung secara berbeda, dan masing-masing memiliki kelebihan dan kekurangannya masing-masing.

Mean adalah rata-rata aritmatika dari semua nilai dalam kumpulan data. Titik di mana distribusi berada dalam keseimbangan. Titik tumpu atau titik keseimbangan dihitung sebagai mean atau mean aritmatika.

Median adalah Nilai tengah yang diukur dari nilai terendah hingga nilai tertinggi.

Modus adalah Nilai yang paling sering muncul di dataset. Jika terdapat beberapa nilai dengan frekuensi paling sering muncul sama, maka dataset bisa mempunyai lebih dari 1 modus.

# STATISTICAL FIVE SUMMARIES (Describe)

## Input

```
2 df_d['thalach'].describe()
```

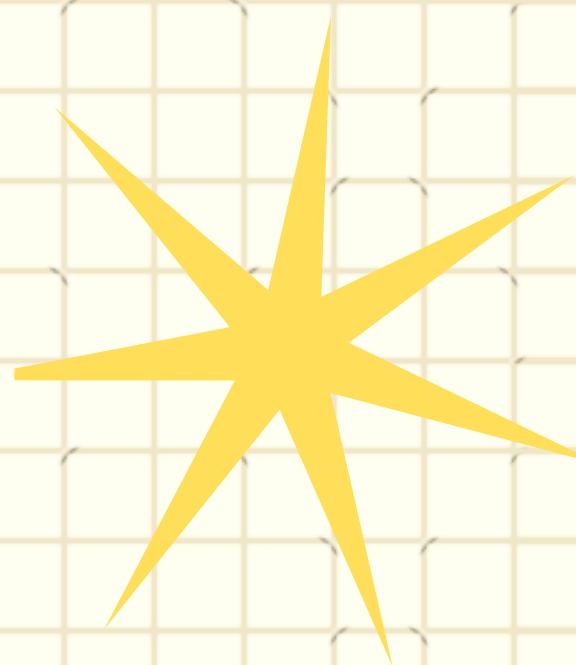
## Output

```
count      302.000000
mean       149.569536
std        22.903527
min        71.000000
25%       133.250000
50%       152.500000
75%       166.000000
max       202.000000
Name: thalach, dtype: float64
```

lima statistik adalah kumpulan lima statistik yang menggambarkan kecenderungan sentral dan penyebaran suatu kumpulan data. yaitu :

1. Minimum: Nilai terkecil dalam kumpulan data.
  2. Maksimum: Nilai terbesar dalam kumpulan data
  3. Q1 (kuartil pertama): Median dari paruh bawah kumpulan data.
  4. Q2 (kuartil atau median kedua): Nilai tengah dalam kumpulan data ketika nilai diurutkan dalam urutan menaik. Jika terdapat dua nilai tengah, maka mediannya adalah rata-rata kedua nilai tersebut.
  5. Q3 (kuartil ketiga): Median dari paruh atas kumpulan data.
- Ringkasan lima angka dapat digunakan untuk memahami distribusi kumpulan data dan mengidentifikasi outlier.

# Data Cleansing



- Imbalance data

```
1 # Split the data into training and testing sets
2 X_train, X_test, y_train, y_test = train_test_split(df_d.drop('age', axis=1), df_d['age'], test_size=0.25, random_state=42)

1 # Check if the training data is imbalanced
2 training_class_distribution = y_train.value_counts()
3 print(training_class_distribution)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	s
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0	
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0	
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0	
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0	
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
723	68	0	2	120	211	0	0	115	0	1.5	1	0	2	1	
733	44	0	2	108	141	0	1	175	0	0.6	1	0	2	1	
739	52	1	0	128	255	0	1	161	1	0.0	2	1	3	0	
843	59	1	3	160	273	0	0	125	0	0.0	2	0	2	0	
878	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0	

- Missing Value

```
def data_check(df_d) :
    missing = df_d.isnull().sum()
    return pd.DataFrame({"Missing" : missing})
data_check(df_d)
```

- Duplicated Values

```
1 df_d.duplicated().sum()
723
1 df_d.drop_duplicates(inplace = True)
2 df_d
```

- Outliers

```
1 #Check Outliers
2 sns.set(style="whitegrid", palette="colorblind")
3 fig, ax = plt.subplots(figsize=(15,10))
4 sns.boxplot(data = df_d)
5 plt.xticks(rotation=45)
```

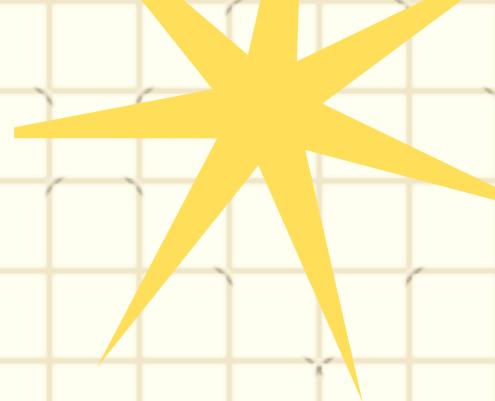
# FEATURE ENGINEERING

**Feature engineering adalah proses mengubah data mentah menjadi fitur yang dapat digunakan oleh model machine learning untuk membuat prediksi yang akurat.**

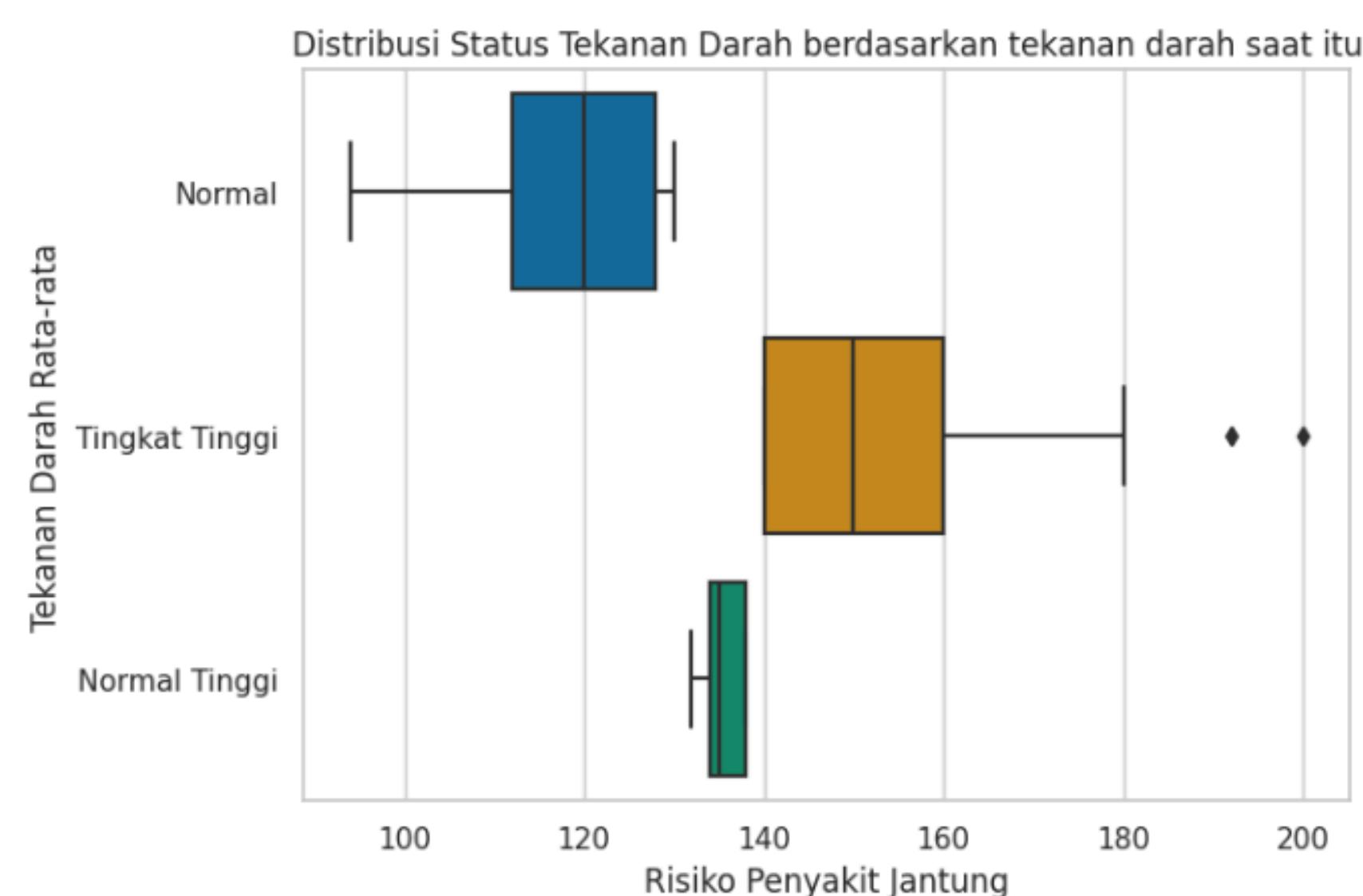


	age	sex	cp	trestbps	chol	fbs	resting	thalach	exang	oldpeak	slope	ca	thal	target	Status	Tekanan Darah	Fbs_category	thal_Status	detak_jantung_normal	Kelompok Usia	Riskis	Kelarisan	Pembangus	Jenis Kelamin	ratio_rata_tekanan_darah_klipule
0	52	1	typical angina	125	212	0	1	150	0	1.0	2	2	3	No disease	Normal	Diabetes	borderline high	True	Dewasa	1.201905		1		2.403546	
1	53	1	typical angina	140	203	1	0	150	1	3.1	0	0	3	No disease	Tingkat Tinggi	Normal	borderline high	False	Dewasa	1.309677		1		2.647509	
2	79	1	typical angina	145	174	0	1	125	1	2.6	0	0	3	No disease	Tingkat Tinggi	Diabetes	desirable	False	Lansia	1.382000		1		2.871429	
3	61	1	typical angina	140	203	0	1	150	0	0.8	2	1	3	No disease	Tingkat Tinggi	Diabetes	borderline high	True	Lansia	1.260070		1		2.426250	
4	62	0	typical angina	138	294	1	1	150	0	1.5	1	3	2	No disease	Normal Tinggi	Normal	high	False	Lansia	2.773985		0		2.325666	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
720	66	0	non-anginal pain	120	211	0	0	115	0	1.5	1	0	2	Disease	Normal	Diabetes	borderline high	False	Lansia	1.034703		0		1.794799	
721	44	0	non-anginal pain	160	141	0	1	175	0	0.6	1	0	2	Disease	Normal	Diabetes	desirable	False	Dewasa	0.805714		0		2.454545	
722	52	1	typical angina	128	256	0	1	140	1	0.8	3	1	3	No disease	Normal	Diabetes	high	False	Dewasa	1.503851		1		2.491538	
840	59	1	asymptomatic	160	273	0	0	125	0	0.8	2	0	2	No disease	Tingkat Tinggi	Diabetes	high	False	Dewasa	2.114000		1		2.711964	
870	54	1	typical angina	120	160	0	1	110	0	1.0	1	1	3	No disease	Normal	Diabetes	desirable	False	Dewasa	1.003717		1		2.322222	

302 rows × 23 columns

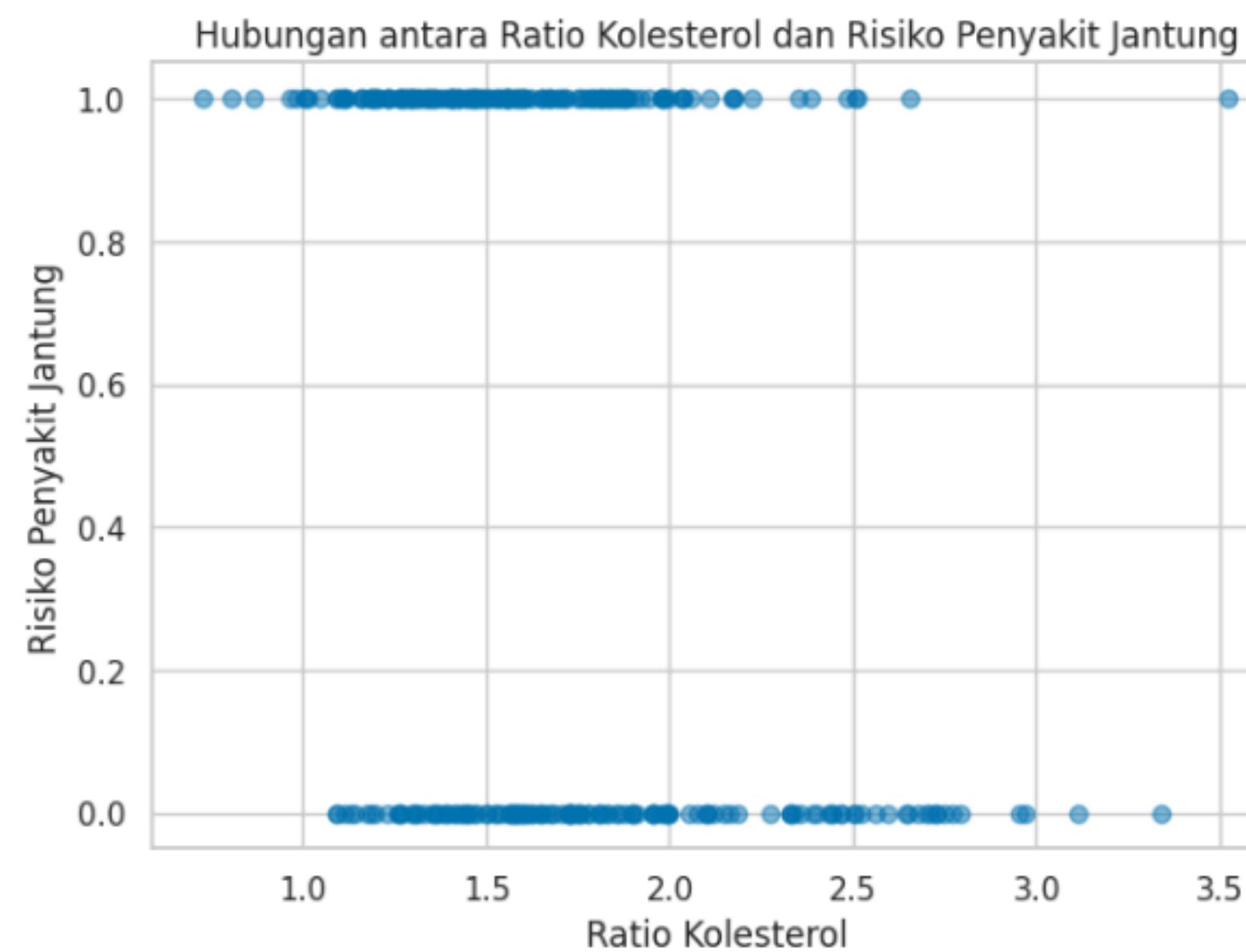


# EDA

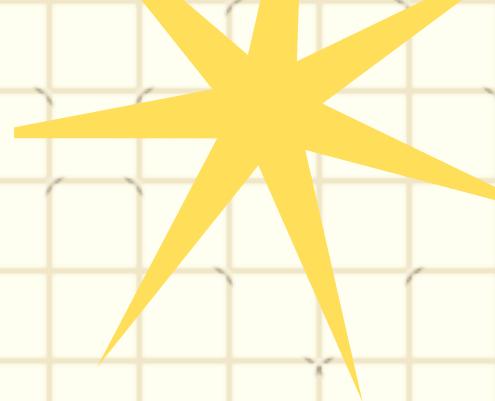


Insight dari data disamping mengenai distribusi status tekanan darah berdasarkan kondisi tekanan darah setiap baris data. saya membagi distribusi kategori menjadi 3 bagian, kondisi normal bagi tekanan darah dibawah 130. Kondisi normal tinggi untuk tekanan darah diantara 131-139 dan kondisi tingkat tinggi untuk tekanan darah diatas atau sama dengan 140

# EDA

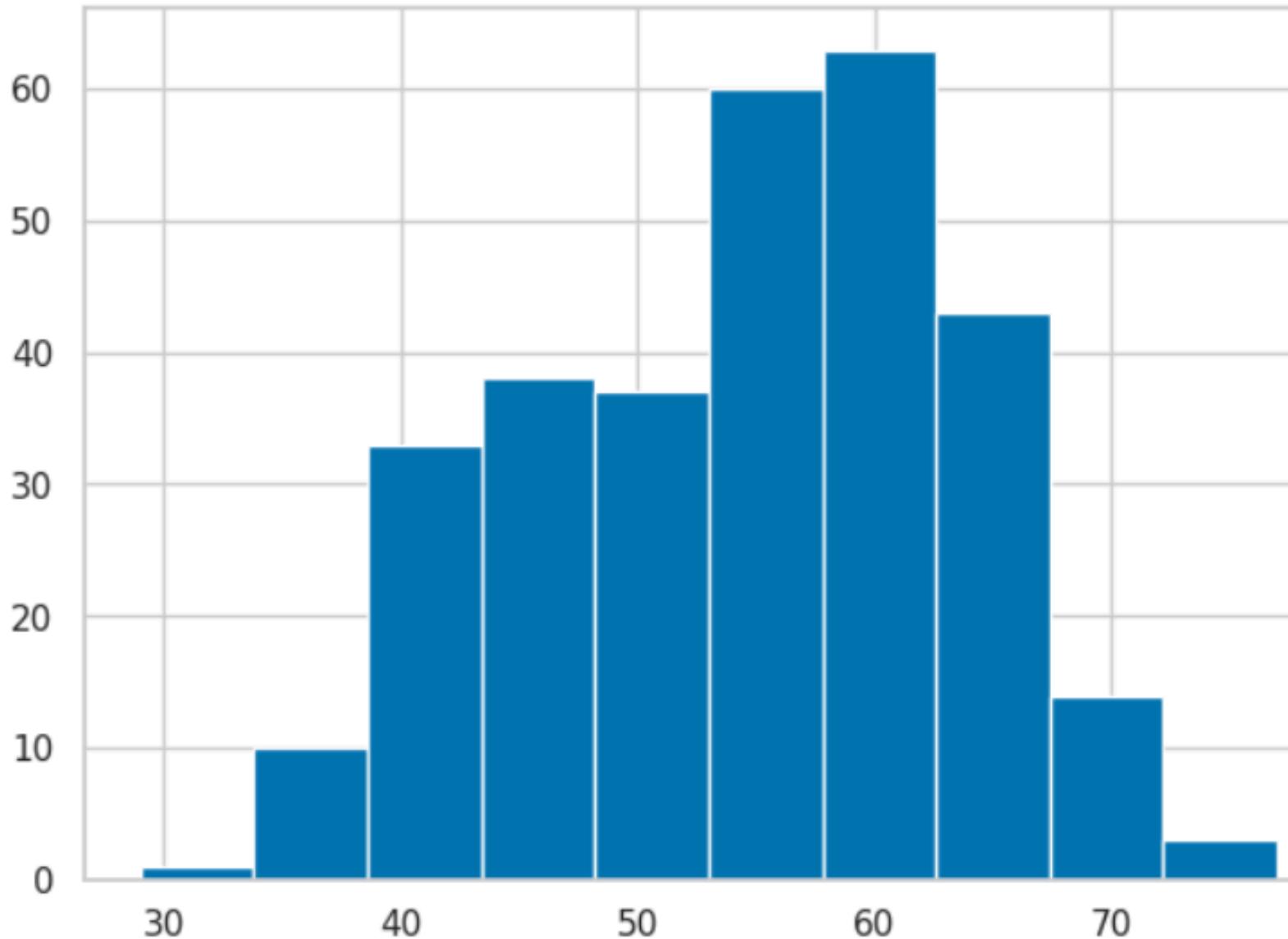


Insight dari data disamping mengenai hubungan antara ratio kolesterol dan risiko penyakit jantung setiap baris data. saya membuat distribusi penyebaran angka berdasarkan hasil pembagian kolom chol dengan kolom thalach. kemudian hasilnya akan disesuaikan dengan hasil dari kolom target setiap data



# EDA

Distribusi Usia Pasien Rumah Sakit



Insight dari data disamping mengenai distribusi usia dari pasien rumah sakit, dimana modusnya adalah umur **60**, umur dengan jumlah sedikit adalah **30**

# Data Conclusion

Berdasarkan analisis data Heart Disease Dataset, dapat disimpulkan bahwa risiko penyakit jantung lebih tinggi pada pasien dengan tekanan darah tinggi dan kolesterol tinggi.

Distribusi status tekanan darah menunjukkan bahwa mayoritas pasien memiliki tekanan darah tinggi, yaitu 62,5%. Sisanya, 37,5% memiliki tekanan darah normal atau normal tinggi. Hal ini menunjukkan bahwa tekanan darah tinggi merupakan faktor risiko penyakit jantung yang penting.

Distribusi penyebaran angka ratio kolesterol menunjukkan bahwa pasien dengan risiko penyakit jantung tinggi memiliki ratio kolesterol yang lebih tinggi. Pasien dengan risiko penyakit jantung rendah memiliki ratio kolesterol yang lebih rendah. Hal ini menunjukkan bahwa ratio kolesterol merupakan faktor risiko penyakit jantung yang penting.

Distribusi usia dari pasien rumah sakit menunjukkan bahwa pasien dengan risiko penyakit jantung tinggi cenderung lebih tua. Pasien dengan usia 60 tahun ke atas memiliki risiko penyakit jantung yang lebih tinggi. Pasien dengan usia 30 tahun ke bawah memiliki risiko penyakit jantung yang lebih rendah. Hal ini menunjukkan bahwa usia merupakan faktor risiko penyakit jantung yang penting.

Berdasarkan ketiga insight tersebut, dapat disimpulkan bahwa pasien dengan tekanan darah tinggi, kolesterol tinggi, dan usia lanjut memiliki risiko penyakit jantung yang lebih tinggi.

Kesimpulan ini dicapai dengan menggunakan metode statistik, yaitu analisis deskriptif. Analisis deskriptif digunakan untuk menggambarkan data secara umum, tanpa melakukan inferensi atau pengujian hipotesis. Dalam kasus ini, analisis deskriptif digunakan untuk menggambarkan distribusi status tekanan darah, ratio kolesterol, dan usia dari pasien rumah sakit.

va



**Link Google Collab**

