



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

PROGRAMMING FOR BIOINFORMATICS

(SECB3203)

PROJECT PROGRESS 3

PROJECT TITLE:

**ANALYZING HEART ATTACK DETECTION THROUGH VARIOUS
CLASSIFICATION ALGORITHMS**

Lecturer:

DR. NIES HUI WEN

Group Members:

No	Name	Matric No
1	JELIZA JUSTINE A/P SEBASTIN	A21EC0034
2	NUR AISYAH FATIHAH BINTI MOHAMED ROZI	A21EC0107
3	SITI NURKAMILAH BINTI SAIFUL BAHARI	A21EC0131

Descriptive statistic

Descriptive statistics typically involve calculating various measures to describe the data's central tendency, variability, and distribution. By transposing the DataFrame using `transpose()`, you'll be able to see all columns with their respective statistics.

Display basic statistics of the dataset

```
summary_statistics = df.describe().transpose()
```

```
print(summary_statistics)
```

[303 rows x 14 columns]							
	count	mean	std	...	50%	75%	max
age	303.0	54.366337	9.082101	...	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	...	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	...	1.0	2.0	3.0
trtbps	303.0	131.623762	17.538143	...	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	...	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	...	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	...	1.0	1.0	2.0
thalachh	303.0	149.646865	22.905161	...	153.0	166.0	202.0
exng	303.0	0.326733	0.469794	...	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	...	0.8	1.6	6.2
slp	303.0	1.399340	0.616226	...	1.0	2.0	2.0
caa	303.0	0.729373	1.022606	...	0.0	1.0	4.0
thall	303.0	2.313531	0.612277	...	2.0	3.0	3.0
output	303.0	0.544554	0.498835	...	1.0	1.0	1.0

Basic Grouping

The basics of grouping in data analysis involve dividing a dataset into groups based on one or more categorical variables and performing operations independently on each group. So, since the objective is to compare the performance of the machine learning models (Gaussian Naïve Bayes and Logistic Regression) in terms of accuracy, precision, recall, and F1-score for heart attack detection we are grouping by the output(0= less chance of heart attack while 1= more chance of heart attack). This could help analyze the model's performance in distinguishing what factors led to getting a heart attack.

```
grouped_data = df.groupby(output).mean()
```

```
print(grouped_data)
```

	age	sex	cp	trtbps	chol	fbs	...	thalach	exng	oldpeak	slp	caa	thall
output							...						
0	56.601449	0.826087	0.478261	134.398551	251.086957	0.159420	...	139.101449	0.550725	1.585507	1.166667	1.166667	2.543478
1	52.496970	0.563636	1.375758	129.303030	242.230303	0.139394	...	158.466667	0.139394	0.583030	1.593939	0.363636	2.121212

So for further explanation, we can see the average age for a chance of getting a heart attack is 52.49 while the average age for not getting a heart attack is 56.60.

As our data is numerical and our objective is to compare Linear Regression and Gaussian Naïve Bayes, we will focus on using Pearson collection instead of the ANOVA test.

Correlation Matrix with Numbers Heatmap

```
import pandas as pd
```

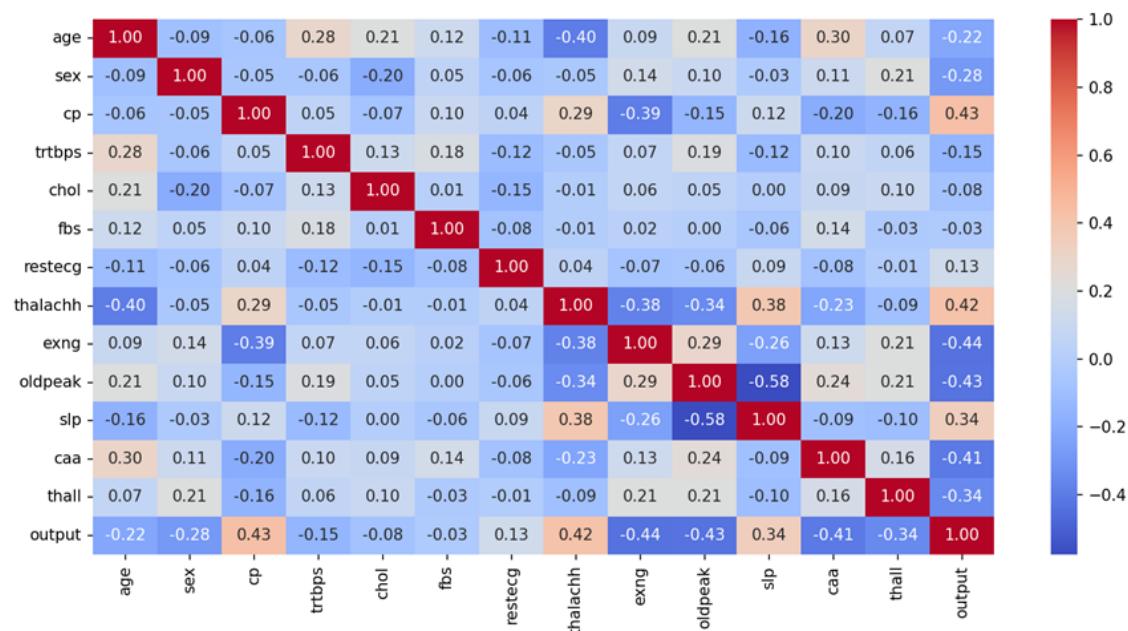
```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
correlation_matrix = df.corr()
```

```
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
```

```
plt.show()
```



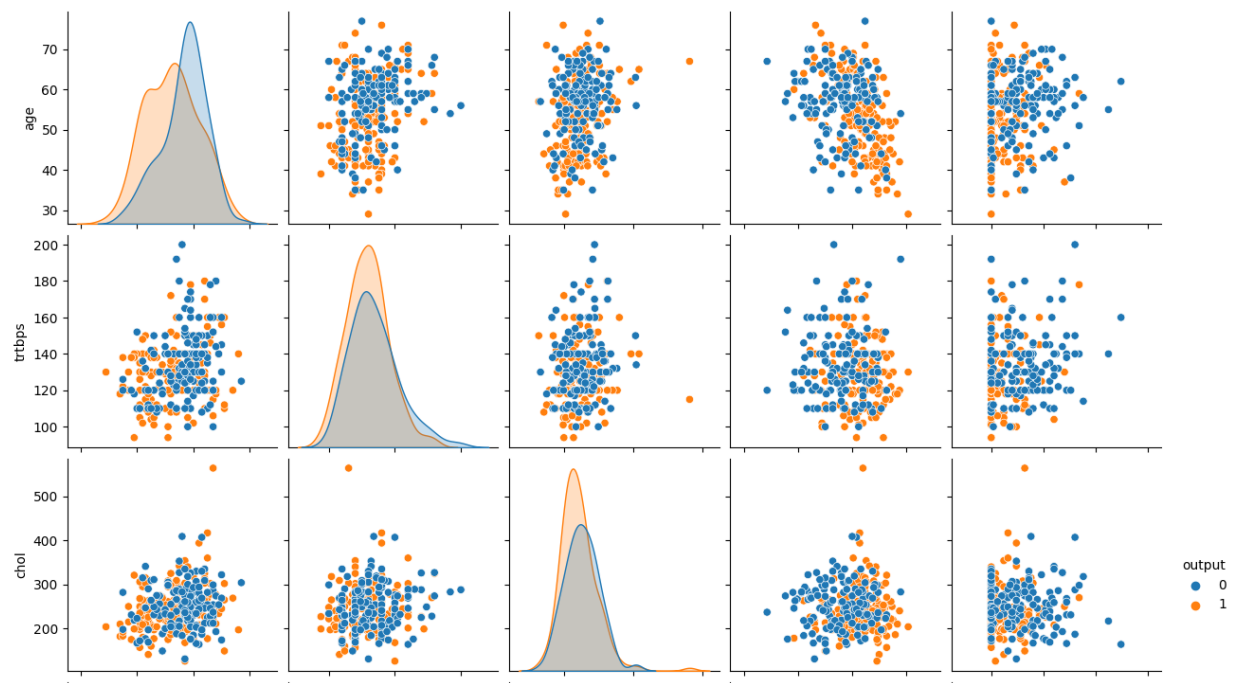
Numeric Feature Analysis

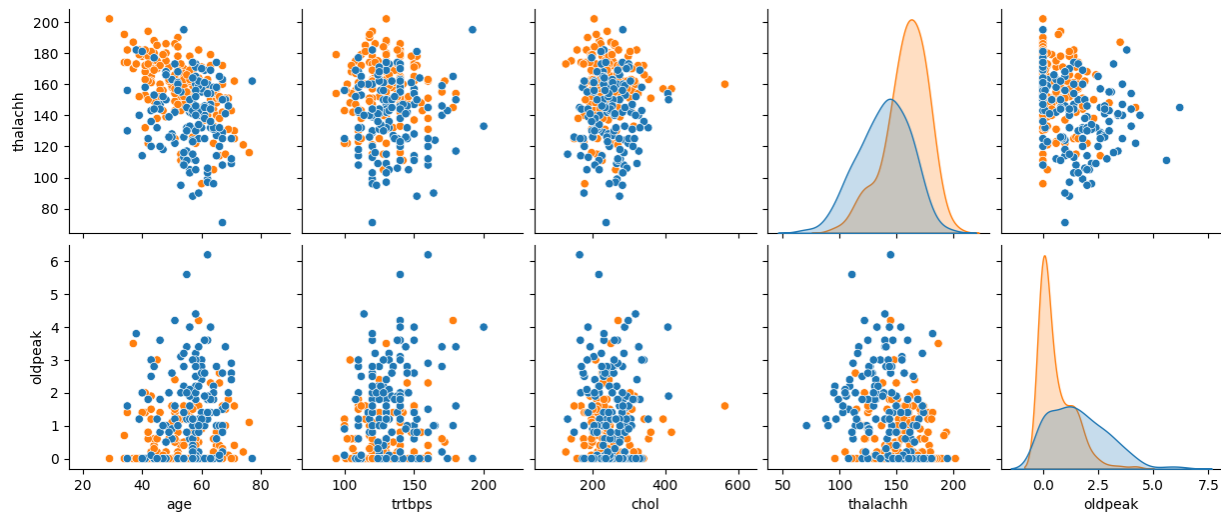
→ Bivariate data analysis with scatter plot

Pairplot allows us to plot pairwise relationships between variables within a dataset. This creates a nice visualization and helps us understand the data by summarizing a large amount of data in a single figure.

```
numeric_list = ["age", "trtbps", "chol", "thalachh", "oldpeak", "output"]
```

```
df_numeric = df.loc[:, numeric_list]  
sns.pairplot(df_numeric, hue = "output", diag_kind = "kde")  
plt.show()
```





The dataset shows the relationships between pairs of six numeric variables: age, trtbps (resting blood pressure), chol (cholesterol), thalachh (maximum heart rate achieved), oldpeak (ST segment depression), and output (the target variable, which is 0 for less chance of heart attack and 1 for more chance of heart attack).

Each row and column of the plot corresponds to one of the numeric variables. The diagonal plots in the middle of the chart are density plots (kde plots) that show the distribution of each variable. The off-diagonal plots are scatter plots that show the relationship between each pair of variables. The color of the points in the scatter plots corresponds to the target variable (output), with red indicating a higher chance of a heart attack and blue indicating a lower chance.

The observations from the pairplot are as follows:

- Age vs. trtbps: There is a weak positive correlation, suggesting that as age increases, resting blood pressure tends to slightly increase as well.
- Age vs. thalachh: There is a weak negative correlation, indicating that as age increases, maximum heart rate tends to slightly decrease.
- trtbps vs. thalachh: There is a weak negative correlation, suggesting that individuals with higher resting blood pressure tend to have lower maximum heart rates.
- thalachh vs. oldpeak: There is a weak positive correlation, indicating that individuals with higher maximum heart rates tend to have slightly higher ST segment depression.
- chol vs. oldpeak: There is a weak positive correlation, suggesting that individuals with higher cholesterol tend to have slightly higher ST segment depression.

Pearson Correlation

The Pearson Correlation measures the linear dependence between two variables X and Y. The resulting coefficient is a value between -1 and 1 inclusive, where:

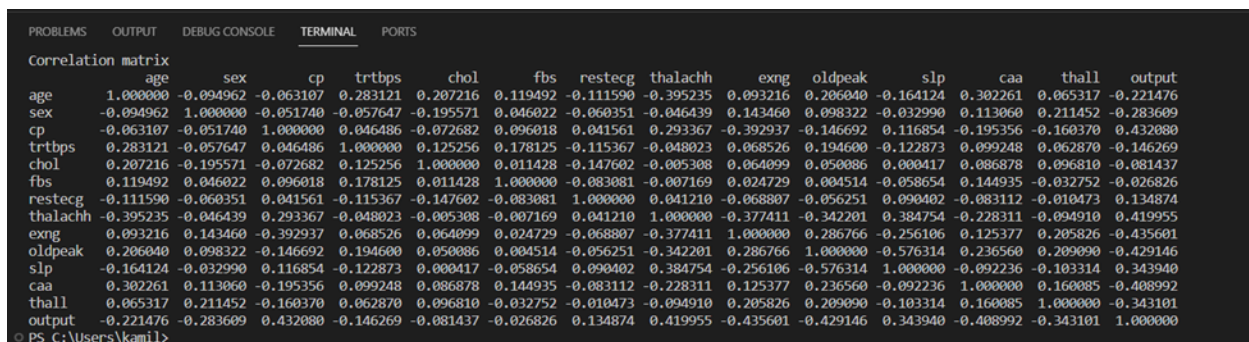
- 1: Total positive linear correlation.
- 0: No linear correlation, the two variables most likely do not affect each other.
- -1: Total negative linear correlation.

Pearson Correlation is the default method of the function "corr"

```
numeric_columns = df.select_dtypes(include=['float64', 'int64'])
```

```
correlation_matrix = numeric_columns.corr()
```

```
print(correlation_matrix)
```



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.000000	-0.094962	-0.063107	0.283121	0.207216	0.119492	-0.111590	-0.395235	0.093216	0.206040	-0.164124	0.302261	0.065317	-0.221476
sex	-0.094962	1.000000	-0.051740	-0.057647	-0.195571	0.046022	-0.060351	-0.046439	0.143460	0.098322	-0.032990	0.113060	0.211452	-0.283609
cp	-0.063107	-0.051740	1.000000	0.046486	-0.072682	0.096018	0.041561	0.293367	-0.392937	-0.146692	0.116854	-0.195356	-0.160370	0.432080
trtbps	0.283121	-0.057647	0.046486	1.000000	0.125256	0.178125	-0.115367	-0.048023	0.068526	0.194600	-0.122873	0.099248	0.062870	-0.146269
chol	0.207216	-0.195571	-0.072682	0.125256	1.000000	0.011428	-0.147602	-0.005308	0.064099	0.050086	0.000417	0.086878	0.096810	-0.081437
fbs	0.119492	0.046022	0.096018	0.178125	0.011428	1.000000	-0.083081	-0.007169	0.024729	0.004514	-0.058654	0.144935	-0.032752	-0.026826
restecg	-0.111590	-0.060351	0.041561	-0.115367	-0.147602	-0.083081	1.000000	0.041210	-0.068807	-0.056251	0.090402	-0.083112	-0.010473	0.134874
thalachh	-0.395235	-0.046439	0.293367	-0.048023	-0.005308	-0.007169	0.041210	1.000000	-0.377411	-0.342201	0.384754	-0.228311	-0.094910	0.419955
exng	0.093216	0.143460	-0.392937	0.068526	0.064099	0.024729	-0.068807	-0.377411	1.000000	0.286766	-0.256106	0.125377	0.205826	-0.435601
oldpeak	0.206040	0.098322	-0.146692	0.194600	0.050086	0.004514	-0.056251	-0.342201	0.286766	1.000000	-0.576314	0.236560	0.209090	-0.429146
slp	-0.164124	-0.032990	0.116854	-0.122873	0.000417	-0.058654	0.090402	0.384754	-0.256106	-0.576314	1.000000	-0.092236	-0.103314	0.343940
caa	0.302261	0.113060	-0.195356	0.099248	0.086878	0.144935	-0.083112	-0.228311	0.125377	0.236560	-0.092236	1.000000	0.160085	-0.408992
thall	0.065317	0.211452	-0.160370	0.062870	0.096810	-0.032752	-0.010473	-0.094910	0.205826	0.209090	-0.103314	0.160085	1.000000	-0.343101
output	-0.221476	-0.283609	0.432080	-0.146269	-0.081437	-0.026826	0.134874	0.419955	-0.435601	-0.429146	0.343940	-0.408992	-0.343101	1.000000

P-value:

The P-value is the probability value that the correlation between these two variables is statistically significant. Normally, we choose a significance level of 0.05, which means that we are 95% confident that the correlation between the variables is significant.

By convention, when

- the p-value is < 0.001 : we say there is strong evidence that the correlation is significant.
- the p-value is < 0.05 : there is moderate evidence that the correlation is significant.
- the $0.05 < \text{p-value} < 0.1$: there is weak evidence that the correlation is significant.
- the p-value is > 0.1 : there is no evidence that the correlation is significant.

Here we compare the p-value of the output(classification of getting a heart attack) with other features


```

44 from scipy import stats
45
46 pearson_coef, p_value = stats.pearsonr(df['age'], df['output'])
47 print("The Pearson Correlation Coefficient and P-value of 'age' and 'output'")
48 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
49 print()
50 pearson_coef, p_value = stats.pearsonr(df['sex'], df['output'])
51 print("The Pearson Correlation Coefficient and P-value of 'sex' and 'output'")
52 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
53 print()
54 pearson_coef, p_value = stats.pearsonr(df['cp'], df['output'])
55 print("The Pearson Correlation Coefficient and P-value of 'cp' and 'output'")
56 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
57 print()
58 pearson_coef, p_value = stats.pearsonr(df['trtbps'], df['output'])
59 print("The Pearson Correlation Coefficient and P-value of 'trtbps' and 'output'")
60 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
61 print()
62 pearson_coef, p_value = stats.pearsonr(df['chol'], df['output'])
63 print("The Pearson Correlation Coefficient and P-value of 'chol' and 'output'")
64 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
65 print()
66 pearson_coef, p_value = stats.pearsonr(df['fbs'], df['output'])
67 print("The Pearson Correlation Coefficient and P-value of 'fbs' and 'output'")
68 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
69 print()
70 pearson_coef, p_value = stats.pearsonr(df['restecg'], df['output'])
71 print("The Pearson Correlation Coefficient and P-value of 'restecg' and 'output'")
72 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
73 print()
74 pearson_coef, p_value = stats.pearsonr(df['thalachh'], df['output'])
75 print("The Pearson Correlation Coefficient and P-value of 'thalachh' and 'output'")
76 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
77 print()

```

```

78 pearson_coef, p_value = stats.pearsonr(df['exng'], df['output'])
79 print("The Pearson Correlation Coefficient and P-value of 'exng' and 'output'")
80 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
81 print()
82 pearson_coef, p_value = stats.pearsonr(df['oldpeak'], df['output'])
83 print("The Pearson Correlation Coefficient and P-value of 'oldpeak' and 'output'")
84 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
85 print()
86 pearson_coef, p_value = stats.pearsonr(df['slp'], df['output'])
87 print("The Pearson Correlation Coefficient and P-value of 'slp' and 'output'")
88 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
89 print()
90 pearson_coef, p_value = stats.pearsonr(df['caa'], df['output'])
91 print("The Pearson Correlation Coefficient and P-value of 'caa' and 'output'")
92 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
93 print()
94 pearson_coef, p_value = stats.pearsonr(df['thall'], df['output'])
95 print("The Pearson Correlation Coefficient and P-value of 'thall' and 'output'")
96 print( "The Pearson Correlation Coefficient is", pearson_coef, " with a P-value of P = ", p_value)
97

```


Result:

```
The Pearson Correlation Coefficient and P-value of 'age' and 'output'
The Pearson Correlation Coefficient is -0.221475827766562 with a P-value of P = 0.00010394837285416845

The Pearson Correlation Coefficient and P-value of 'sex' and 'output'
The Pearson Correlation Coefficient is -0.28360935779586227 with a P-value of P = 5.402435780432229e-07

The Pearson Correlation Coefficient and P-value of 'cp' and 'output'
The Pearson Correlation Coefficient is 0.43207959156640785 with a P-value of P = 3.6273838772544235e-15

The Pearson Correlation Coefficient and P-value of 'trtbts' and 'output'
The Pearson Correlation Coefficient is -0.1462686638415544 with a P-value of P = 0.010926538861949197

The Pearson Correlation Coefficient and P-value of 'chol' and 'output'
The Pearson Correlation Coefficient is -0.08143720051844142 with a P-value of P = 0.15803697464249777

The Pearson Correlation Coefficient and P-value of 'fbs' and 'output'
The Pearson Correlation Coefficient is -0.026825970565970136 with a P-value of P = 0.6424070490676519

The Pearson Correlation Coefficient and P-value of 'restecg' and 'output'
The Pearson Correlation Coefficient is 0.13487444702864623 with a P-value of P = 0.019033607668275408

The Pearson Correlation Coefficient and P-value of 'thalachh' and 'output'
The Pearson Correlation Coefficient is 0.4199550436638698 with a P-value of P = 2.4761460479236786e-14

The Pearson Correlation Coefficient and P-value of 'exng' and 'output'
The Pearson Correlation Coefficient is -0.4356007617136187 with a P-value of P = 2.04646758906984e-15

The Pearson Correlation Coefficient and P-value of 'oldpeak' and 'output'
The Pearson Correlation Coefficient is -0.42914583288673835 with a P-value of P = 5.814566948031603e-15

The Pearson Correlation Coefficient and P-value of 'slp' and 'output'
The Pearson Correlation Coefficient is 0.3439395324893869 with a P-value of P = 8.221388831030162e-10

The Pearson Correlation Coefficient and P-value of 'caa' and 'output'
The Pearson Correlation Coefficient is -0.40899197975692636 with a P-value of P = 1.3173455415475193e-13

The Pearson Correlation Coefficient and P-value of 'thall' and 'output'
The Pearson Correlation Coefficient is -0.34310071238956447 with a P-value of P = 9.089044024818302e-10
PS C:\Users\kamil>
```

'age' and 'output':

- Correlation: -0.2215 (moderate negative correlation)
- P-value: 0.0001 (highly significant)

'sex' and 'output':

- Correlation: -0.2836 (moderate negative correlation)
- P-value: 5.4024e-07 (highly significant)

'cp' and 'output':

- Correlation: 0.4321 (strong positive correlation)
- P-value: 3.6274e-15 (highly significant)

'trtbs' and 'output':

- Correlation: -0.1463 (weak negative correlation)
- P-value: 0.0109 (significant)

'chol' and 'output':

- Correlation: -0.0814 (weak negative correlation)
- P-value: 0.1580 (not significant)

'fbs' and 'output':

- Correlation: -0.0268 (very weak negative correlation)
- P-value: 0.6424 (not significant)

'restecg' and 'output':

- Correlation: 0.1349 (weak positive correlation)
- P-value: 0.0190 (significant)

'thalachh' and 'output':

- Correlation: 0.4200 (strong positive correlation)
- P-value: 2.4761e-14 (highly significant)

'exng' and 'output':

- Correlation: -0.4356 (strong negative correlation)
- P-value: 2.0465e-15 (highly significant)

'oldpeak' and 'output':

- Correlation: -0.4291 (strong negative correlation)
- P-value: 5.8146e-15 (highly significant)

'slp' and 'output':

- Correlation: 0.3439 (moderate positive correlation)
- P-value: 8.2214e-10 (highly significant)

'caa' and 'output':

- Correlation: -0.4090 (strong negative correlation)

- P-value: 1.3173×10^{-13} (highly significant)

'thall' and 'output':

- Correlation: -0.3431 (moderate negative correlation)
- P-value: 9.0890×10^{-10} (highly significant)

Conclusion

We now have a better idea of what our data looks like and which variables are important to take into account when predicting heart attack detection. We have narrowed it down to the following variables:

- age
- sex
- cp
- thalachh
- exng
- oldpeak
- slp
- caa
- thall

As we now move into building machine learning models to automate our analysis, feeding the model with variables that meaningfully affect our target variable will improve our model's prediction performance.