HEART DISEASE PREDICTION

By: Group A

Participants

Sopika Kanagaratnam

Sowjanya Nagulapati

Chukwunwendu Chizoba

Mamadou Aliou Diallo

Introduction

Heart diseases are the collection of diseases that are related to the heart or blood vessels. Blood is pumped and circulated to all tissues of the body by the heart. Important organs of the body such as the brain and kidneys suffer if the heart's pumping mechanism becomes inefficient. If the heart stops functioning at least for some minutes that will lead to Death. The World Health Organization (WHO) has estimated that 17.9 million deaths occur worldwide every year due to heart disease (Goenka et al., 2009)

Due to poor prevention, the number of heart disease deaths continues to increase, although most heart disease deaths can be prevented and may be avoided. Different academics have created support systems for clinical decisions to help speed up and simplify the detection of heart disease in today's digital age. Many researchers have used data mining and machine learning approaches to develop clinical decision-support systems for heart disease prediction.

The aim of this report is to analyze the heart disease data set and provide a detailed report with findings from our analysis.

Data set Description

The heart disease data set consists of **14 variables and 270 observations**. The data set categorizes the patients whether they have heart disease or not. The literature says that initially, the data set contained 76 features, and published studies chose only 14 features that are most relevant in predicting heart diseases.

Categorical variable – 9 (Sex, Chestpaintype, FBSover120, ECG result, Excerciseangina, SlopeofST, Numberofvesselfluro, Thallium, Heart disease)

Numerical Variable – 5 (Age, BP, Cholesterol, MaxHR, STSexdepression)

Attribute	Description	Domain of value
Age	Age in years	29 to 77
Sex	Sex	Male (1) Female (0)
Cp	Chest pain type	Typical angina (1) Atypical angina (2) Non-anginal (3) Asymptomatic (4)
Trestbps	Resting blood sugar	94 to 200 mm Hg
Chol	Serum cholesterol	126 to 564 mg/dl
Fbs	Fasting blood sugar	>120 mg/dl True (1) False (0)
Restecg	Resting ECG result	Normal (0) ST-T wave abnormality (1) LV hypertrophy (2)
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise induced angina	Yes (1) No (0)
Oldpeak	ST depression induced by exercise relative to rest	0 to 6.2
Slope	Slope of peak exercise ST segment	Upsloping (1) Flat (2) Downsloping (3)
Ca	Number of major vessels coloured by fluoroscopy	0–3
Thal	Defect type	Normal (3) Fixed defect (6) Reversible defect (7)
Num	Heart disease	0–4

Figure 1: Variable Description

Data cleaning

1. Checking for missing values

Statistics

		Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
N	Valid	270	270	270	255	253	270	270	261	270	266	270	270	270	270
	Missing	0	0	0	15	17	0	0	9	0	4	0	0	0	0
Mean		54.43	.75	3.32	131.09	248.76	.15	1.02	157.89	.33	1.042	1.59	.67	4.70	
Median		55.00	1.00	3.00	130.00	245.00	.00	2.00	154.00	.00	.800	2.00	.00	3.00	
Mode		54	1	4	120	234	0	2	162	0	.0	1	0	3	
Std. Deviation		9.109	.965	2.660	18.116	51.406	.356	.998	100.789	.471	1.1416	.614	.944	1.941	
Variance		82.975	.931	7.074	328.204	2642.612	.127	.996	10158.515	.222	1.303	.377	.891	3.766	
Minimum		29	0	1	94	126	0	0	71	0	.0	1	0	3	
Maximum		77	11	44	200	564	1	2	1380	1	6.2	3	3	7	

Figure 2: Missing value check

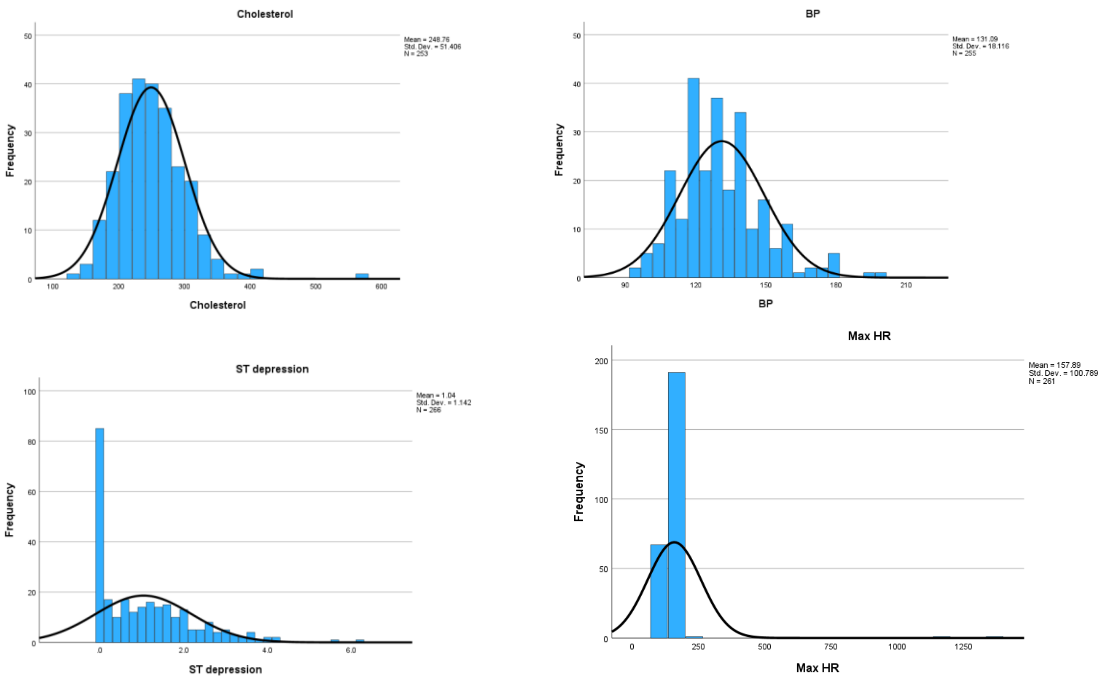
It is observed that from the above table attributes of BP, Cholesterol, MaxHR, and ST depression contain missing values. As the percentage of total missing values in each attribute is less than 10%, that will not affect the results of the analysis much. However, to replace missing values normality is checked.

Normality check for replacing missing values

Statistics

		Cholesterol	BP	ST depression	Max HR
N	Valid	253	255	266	261
	Missing	17	15	4	9
Mean		248.76	131.09	1.042	157.89
Median		245.00	130.00	.800	154.00
Mode		234	120	.0	162
Std. Deviation		51.406	18.116	1.1416	100.789
Variance		2642.612	328.204	1.303	10158.515
Skewness		1.200	.757	1.274	10.630
Std. Error of Skewness		.153	.153	.149	.151
Minimum		126	94	.0	71
Maximum		564	200	6.2	1380

Figure 3: Skewness



Skewness values of the variables which have missing values are checked to determine whether the mean or median is to be replaced. Since the skewness values show higher values and the normal plots show right skewness, the **median values** of those variables were replaced.

2. Checking for out-of-range values

Statistics

		Age	Sex	Chest pain type	FBS over 120	EKG results	Exercise angina	Slope of ST	Number of vessels fluro	Thallium	Heart Disease	SMEAN(BP)	SMEAN (Cholesterol)	SMEAN (MaxHR)	SMEAN (STdepression)
N	Valid	270	270	270	270	270	270	270	270	270	270	270	270	270	270
	Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Minimum		29	0	1	0	0	0	1	0	3		94.0	126.0	71.0	.00
Maximum		77	11	44	1	2	1	3	3	7		200.0	564.0	1380.0	6.20

It can be observed that the range of the attribute sex is 0-11 and the range of the chest pain type is 1-44. Frequency tables for these two variables were obtained to further analysis.

Sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	86	31.9	31.9	31.9
	1	182	67.4	67.4	99.3
	10	1	.4	.4	99.6
	11	1	.4	.4	100.0
	Total	270	100.0	100.0	

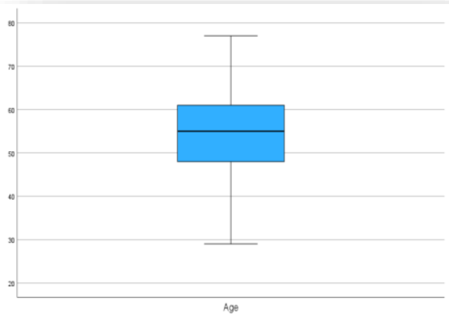
Chest pain type

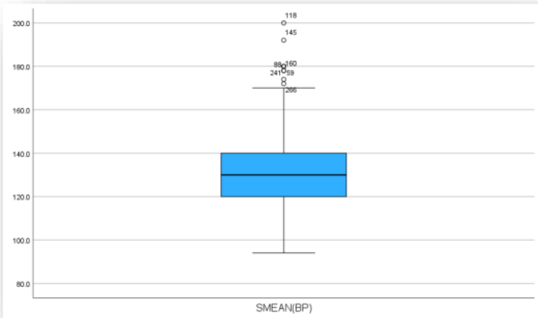
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	20	7.4	7.4	7.4
	2	42	15.6	15.6	23.0
	3	79	29.3	29.3	52.2
	4	128	47.4	47.4	99.6
	44	1	.4	.4	100.0
	Total	270	100.0	100.0	

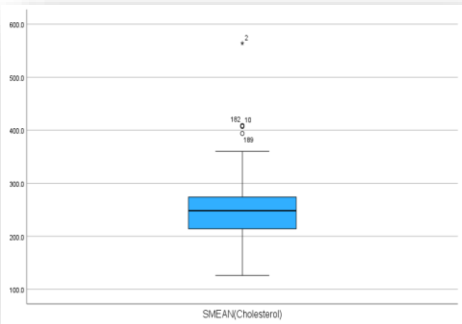
It can be confirmed that there is a mistake in the observations with the categories 10 and 11 in sex and it is converted into 1 as considering that was a typing error. Same as this the category of 44 in chest pain type was converted into 4.

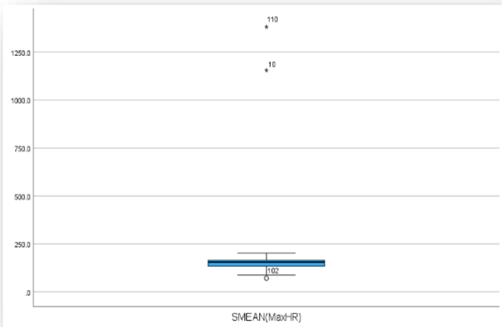
3. Checking for Outliers

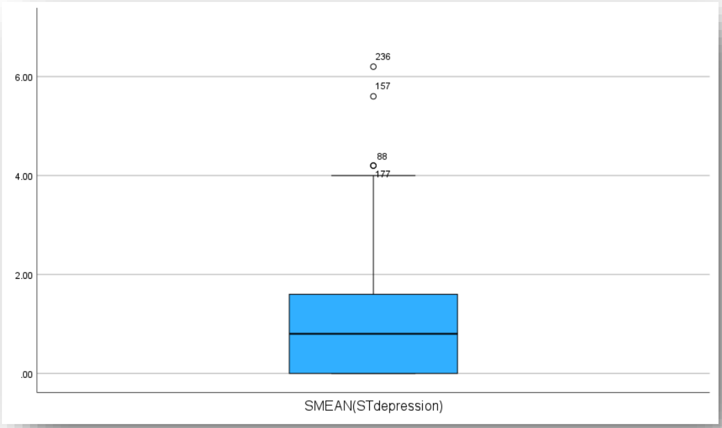
Boxplot is plotted to check the outlier observations.



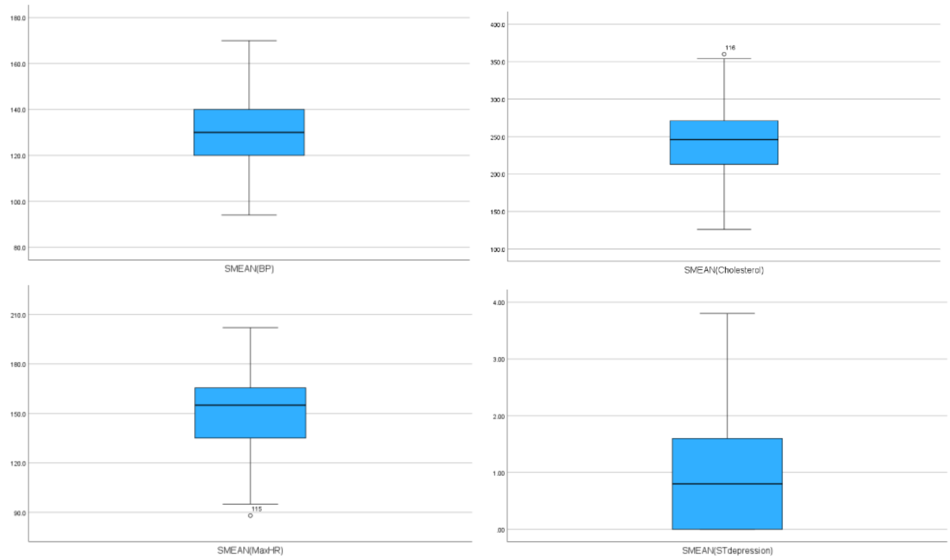








Variable Age does not contain any outliers while BP, Cholesterol, MaxHR, and ST Depression contain some outliers. Since the number of outliers is a small amount, it is eliminated from the data set. Below are the box plots after eliminating the outliers from the data.



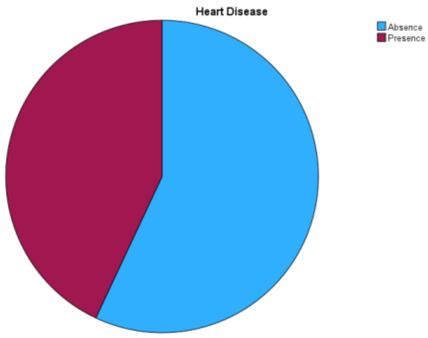
Now the data set is clean and ready for analysis.

Descriptive Analysis

Descriptive Analysis is the type of analysis of data that helps describe, show, or summarize data points in a constructive way such that patterns might emerge that fulfill every condition of the data.

Heart Disease

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absence	143	57.0	57.0	57.0
	Presence	108	43.0	43.0	100.0
	Total	251	100.0	100.0	

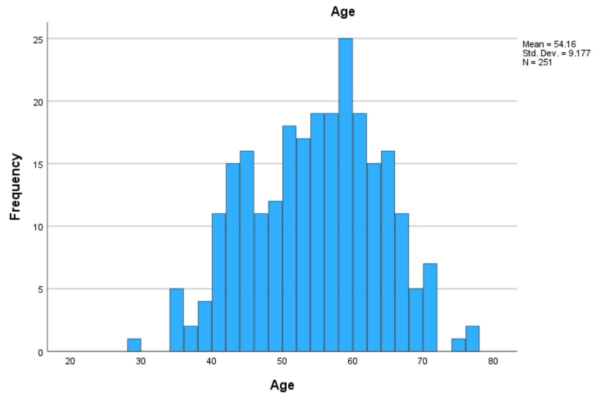


The data set categorizes as 143 (57%) patients without heart disease and 108 (43%) patients with heart disease. Thus, the presence of heart disease in patients is less than the absence of the disease.

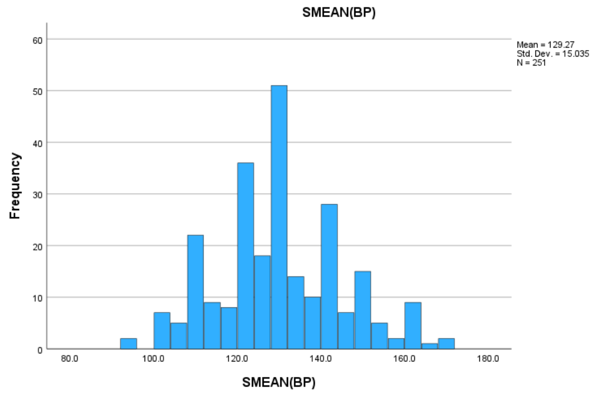
Statistics

		Age	SMEAN(BP)	SMEAN (Cholesterol)	SMEAN (MaxHR)	SMEAN (STdepression)
N	Valid	251	251	251	251	251
	Missing	0	0	0	0	0
Mean		54.16	129.272	245.570	149.976	.9596
Median		54.00	130.000	246.000	155.000	.8000
Mode		54[a]	120.0	248.8	157.9[a]	.00
Std. Deviation		9.177	15.0349	42.6954	22.2742	.99235
Variance		84.217	226.047	1822.894	496.140	.985
Minimum		29	94.0	126.0	88.0	.00
Maximum		77	170.0	360.0	202.0	3.80
Percentiles	25	47.00	120.000	213.000	134.000	.0000
	50	54.00	130.000	246.000	155.000	.8000
	75	61.00	140.000	271.000	166.000	1.6000

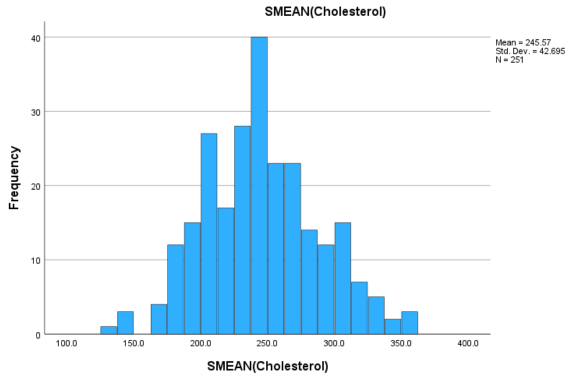
a. Multiple modes exist. The smallest value is shown



The age of the patients included in the data set is within the range of 29 -77 and most of the patients between 50s to 60s. The mean age is 54.16. The standard deviation of the age is 9.177. Since this is a low value, most of the data points are clustered around the mean



The BP of the patients fluctuated within the range between 94 – 170 and the mean BP is 129.27. The first 50% of the patients in the data set have BP values below 130.



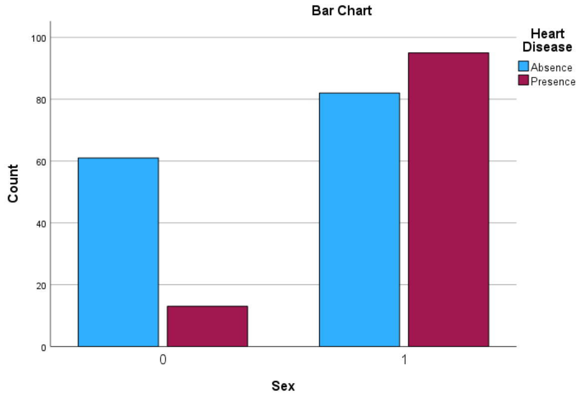
The Cholesterol of the patients included in the data set fluctuated from 126 to 360. The average cholesterol is 245.57. The standard deviation is 42.69 which is a high value. There for the data points are more spread out. Patients with the cholesterol above 271 is only the 25% of the total patients.

Heart disease categorization with Sex

Sex * Heart Disease Crosstabulation

Count

		Heart Disease		
		Absence	Presence	Total
Sex	0	61	13	74
	1	82	95	177
Total		143	108	251



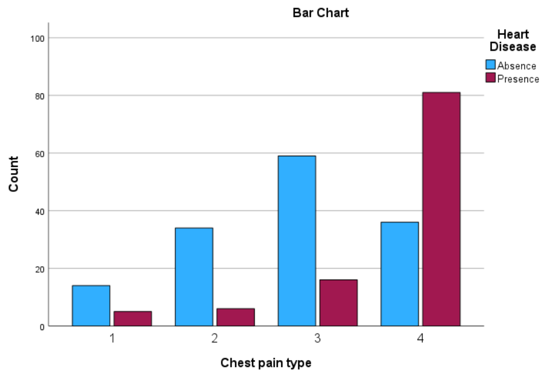
13 (17.56%) females out of 74 have heart disease while 95 (53.67%) males out of 177 have heart disease. Therefore, male patients with heart disease are higher than that of females in the data set.

Heart disease categorization with Chest pain

Chest pain type * Heart Disease Crosstabulation

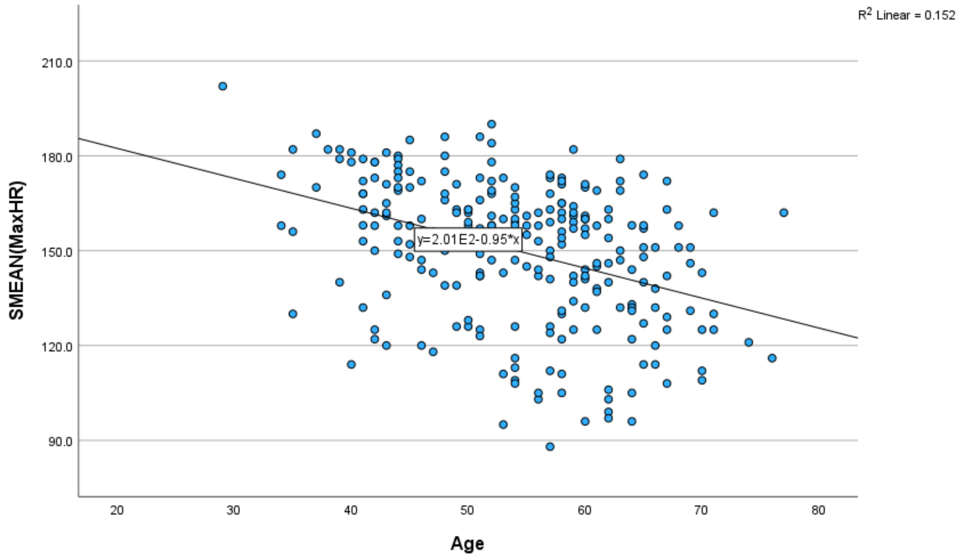
Count

		Heart Disease		
		Absence	Presence	Total
Chest pain type	1	14	5	19
	2	34	6	40
	3	59	16	75
	4	36	81	117
Total		143	108	251



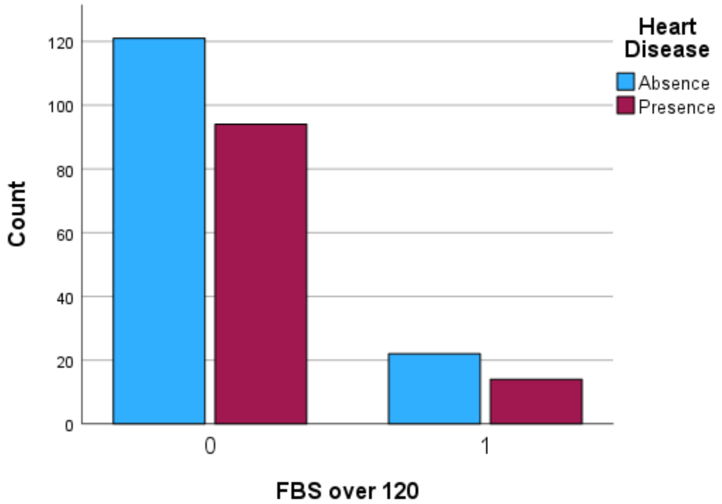
117 patients out of 251 have chest pain type 4 and 81 patients with type 4 chest pain have heart disease. Comparatively the number of patients with chest pain types 1,2, and 3 who have heart disease is less than those who haven’t heart disease. Thus, chest pain type 4 is an indication of a patient having heart disease.

Scatter plot for Age Vs MaxHR



It can be seen that from the above scatter plot achieved maximum heart rate decreases when the patient's age increases. So, there is a negative relationship between age and Heart rate achieved.

FBS VS Heart Disease



Fasting blood sugar or FBS a diabetes indicator with FBS >120 mg/d is considered diabetic (True class). Here, we observe that the number for class 1(True), is lower compared to class 0 (False). However, if we look closely, there are a higher number of heart disease patients without diabetes. This provides an indication that FBS might not be a strong feature differentiating between heart disease and non-disease patients.

Normality Test

Mean, Median and Mode

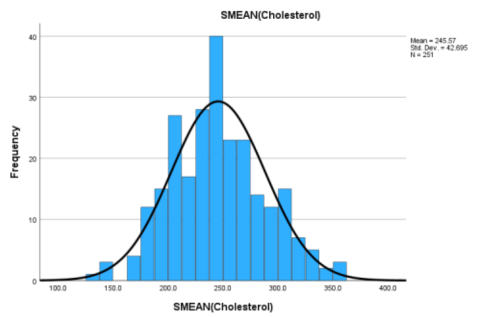
Statistics

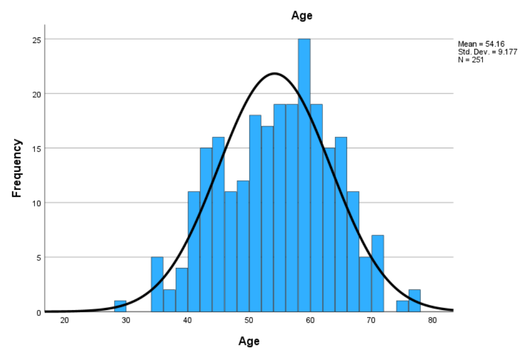
		Age	SMEAN(BP)	SMEAN (Cholesterol)	SMEAN (MaxHR)	SMEAN (STdepression)
N	Valid	251	251	251	251	251
	Missing	0	0	0	0	0
Mean		54.16	129.272	245.570	149.976	.9596
Median		54.00	130.000	246.000	155.000	.8000
Mode		54[a]	120.0	248.8	157.9[a]	.00
Std. Deviation		9.177	15.0349	42.6954	22.2742	.99235
Variance		84.217	226.047	1822.894	496.140	.985
Skewness		-.121	.206	.156	-.568	.844
Std. Error of Skewness		.154	.154	.154	.154	.154
Kurtosis		-.547	-.189	-.127	-.282	-.205
Std. Error of Kurtosis		.306	.306	.306	.306	.306

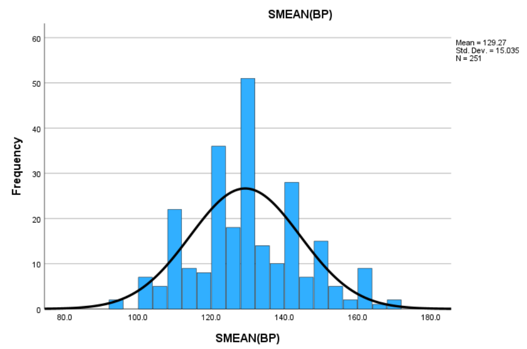
a. Multiple modes exist. The smallest value is shown

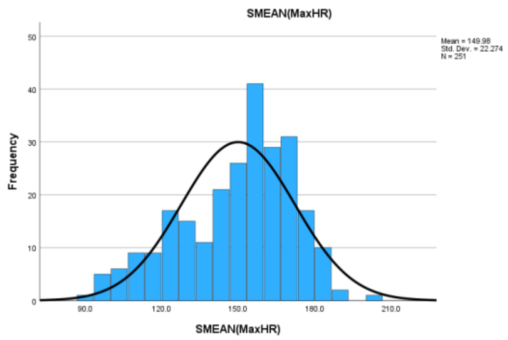
The first thing we can check for normality is to check if the mean, median, and mode values are equal. For the above continuous variables in our data set these three values are approximately equal for Age and Cholesterol. The other three variables have differences between mean, mode, and median.

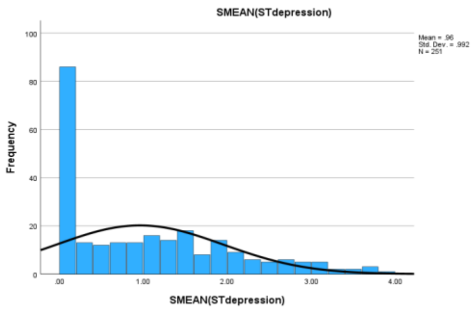
Histogram





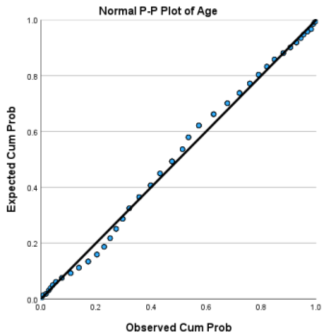


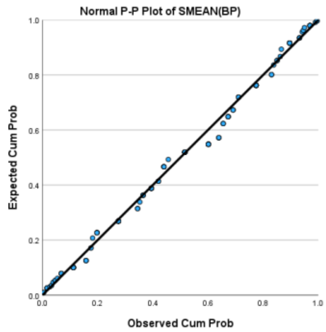


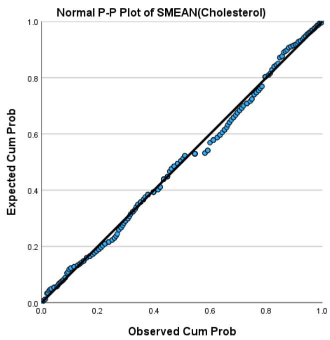


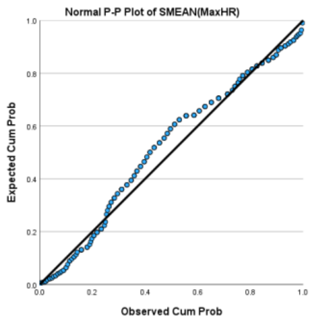
When plotting the histogram with a normal curve it can be observed that except for ST depression other four variables show an approximately symmetric normal curve.

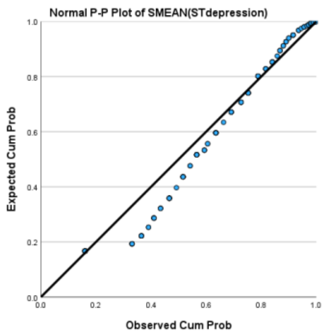
P-P Plots





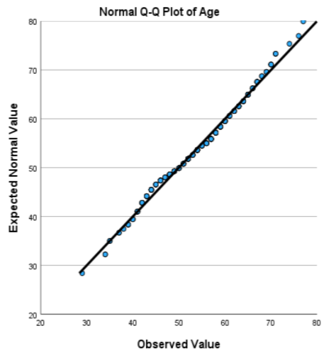


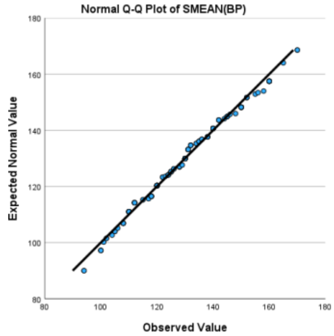


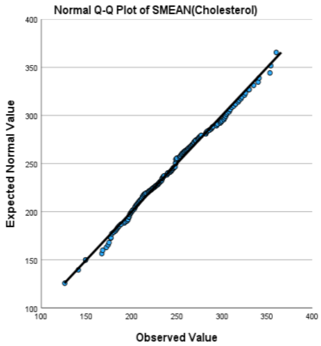


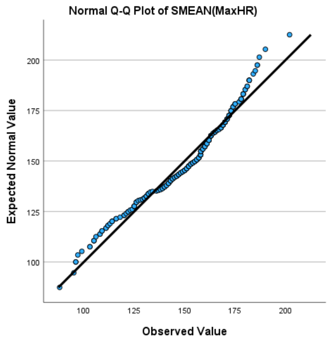
P-P plots of the variables age, BP, and Cholesterol form an approximate straight line while MaxHR and STdepression deviate from the straight line.

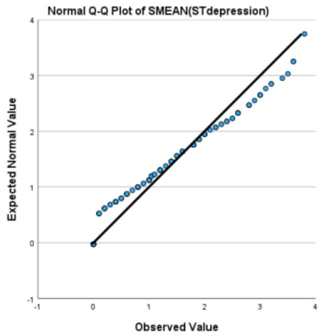
Q-Q Plots











Age, cholesterol, and BP form an approximate straight line when plotting the Q_Q plot while MaxHR and STDepression deviate from the line in the line starting and ending.

Skewness and Kurtosis Z-values

Age => Skewness/Standard Error = -0.121/0.154 = -0.785

Kurtosis/Standard Error =-0.547/0.306 = 1.787

BP => Skewness = 0.206/0.154 = 1.337

Kurtosis = -0.189/0.306 = -0.6176

Cholesterol = > Skewness = 0.156/0.154 =1.012

-0.127/0.306=-0.415

MaxHR => Skewness = -0.568/0.154= **-3.688**

Kurtosis = 0.282/0.306=0.921

ST Depression => Skewness = 0.844/0.154= **5.48**

Kurtosis = -0.205/0.306=0.669

For sample sizes between 50 to 300, Z-value of +- 3.29 is sufficient to establish the normality. In our data set, variables Age, BP, and cholesterol have values for skewness and Kurtosis in the range between -3.29 to +3.29. The skewness of MaxHR and ST depression take the z-values of skewness out of this range.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.072	251	.003	.989	251	.062
SMEAN(BP)	.081	251	<.001	.985	251	.010
SMEAN(Cholesterol)	.056	251	.052	.994	251	.421
SMEAN(MaxHR)	.105	251	<.001	.965	251	<.001
SMEAN(STdepression)	.167	251	<.001	.870	251	<.001

a. Lilliefors Significance Correction

Based on the Shapiro- Wilk test, the significance value of age is 0.062 (>0.05) and for Cholesterol 0.421(>0.05). That means these two variables are approximately normally distributed. When looking into the result of the Kolmogorov test the only variable Cholesterol has a significance value of 0.052 which is greater than 0.05. Based on this test the only attribute that was normally distributed is Cholesterol.

Based on the visual and statistical normality test outcome, we conclude that the variable cholesterol follows a normal distribution and Variables Age, BP, MaxHR, and STDepression are not normally distributed.

- **what is the probability of the patients whose BP is more than 150?**
  P (X > 150) = P (Z > 1.37865) = 1 - 0.9147 = 8.53%

The probability of the patients who have BP levels more than 150 is 8.53%

- **what is the probability of the patients whose Cholesterol level is more than 330?**
  P (X > 330) = P (Z > 1.97749) = 1 - 0.9756 = 2.44%

The probability of the patients who have cholesterol levels more than 330 is 2.44%

Data Transformation for Normality

Variables in the data set following normal distribution are essential for most of the tests we conduct for further analysis. Since the variables Age, BP, MaxHR, and STDepression fail to follow the normal distribution tests for normality check was performed after data transformation.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Trf_age	.032	250	.200*	.997	250	.961

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The p-value for both Kolmogorov and Shapiro_wilk tests is greater than the significance value of 0.05. So, we can confirm that the variable age is normally distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Trf_MaxHR	.020	250	.200*	.999	250	.999

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

For the variable maxHR, the p-value of the normality tests shows greater than 0.05 after transformation. Thus, this follows the normal distribution.

The variables BP and STdepression don't follow the normal distribution even after the data transformation.

Power and effect size

Tests of Between-Subjects Effects

Dependent Variable: Heart Disease

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	23.476[a]	71	.331	1.555	.010	.382	110.426	1.000
Intercept	376.791	1	376.791	1772.360	<.001	.908	1772.360	1.000
Sex	4.941	1	4.941	23.243	<.001	.115	23.243	.998
Age	9.095	40	.227	1.069	.373	.193	42.780	.931
Sex * Age	5.121	30	.171	.803	.757	.119	24.089	.724
Error	38.054	179	.213					
Total	680.000	251						
Corrected Total	61.530	250						

a. R Squared = .382 (Adjusted R Squared = .136)

b. Computed using alpha = .05

Power of sex = 0.998 and power of age = 0.931 when predicting heart disease. There is a 72% possibility to discover if there is any interaction between sex and age. 6% -14% is a good effect size for medium sample sizes. In this data set, the effect of the sex variable is 11% and the effect of the sex is 19%. This indicates that the difference between the groups in age and sex is meaningful when finding heart disease.

Parametric Test

Parametric test: T-test.

One sample T-test and unpaired samples T-test are conducted in this report.

1. <u>One sample T-test:</u>

Is there any difference in the population mean and sample mean of cholesterol?

Hypothesis:

H0: The population mean value of the cholesterol is equal to the sample mean value of the cholesterol ($\mu1 = \mu2$)

H1: The population mean value of cholesterol is not equal to the sample mean value of cholesterol ($\mu1 \neq \mu2$)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
SMEAN(Cholesterol)	62	242.835	44.1112	5.6021

One-Sample Test

	Test Value = 245.57						
			Significance			95% Confidence Interval of the Difference	
	t	df	One-Sided p	Two-Sided p	Mean Difference	Lower	Upper
SMEAN(Cholesterol)	-.488	61	.314	.627	-2.7352	-13.937	8.467

The two-sided P-value of the one-sample T-test is 0.627 which is >0.05. That means The H0 is accepted. There is no significant difference in the mean value of cholesterol in the population and sample.

2. <u>Unpaired sample T-test</u>

Relationship between the cholesterol level and sex?

Hypothesis:

H0: The average cholesterol value is equal for females and males in the heart disease data (µm = µf)

H1: The average cholesterol value is not equal for females and males in the heart disease data (µm ≠ µf)

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
SMEAN(Cholesterol)	0	74	252.406	47.3302	5.5020
	1	74	243.585	43.7915	5.0907

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
SMEAN(Cholesterol)	Equal variances assumed	1.017	.315	1.177	146	.121	.241	8.8206	7.4958	-5.9937	23.6349
	Equal variances not assumed			1.177	145.127	.121	.241	8.8206	7.4958	-5.9945	23.6356

As the observation of males and females in our data set is not equal, we reduced the sample size of males to females to do an unpaired sample t-test. From the result of the above T-test, it can be seen that the two-sided p-value = 0.241(>0.05) and the calculated t -value = 1.177 < critical t-value =1.660. That means H0 is accepted and H1 is rejected. Therefore, there is no significant difference in the mean value of cholesterol in males and females. So, a patient cholesterol level is not influenced by their sex.

Relationship between the BP level and age?

H0: Average blood pressure of patients aged above 55 is equal to the average blood pressure of patients aged below 55 (µaboove55 = µbelow55)

H1: Average blood pressure of patients aged above 55 is not equal to the average blood pressure of patients aged below 55 (µaboove55 ≠ µbelow55)

Group Statistics

	Trf_age	N	Mean	Std. Deviation	Std. Error Mean
Trf_BP	>= 55.00	119	133.4180	15.61565	1.43148
	< 55.00	119	125.2821	13.58391	1.24523

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Trf_BP	Equal variances assumed	1.837	.177	4.288	236	<.001	<.001	8.13586	1.89730	4.39805	11.87367
	Equal variances not assumed			4.288	231.559	<.001	<.001	8.13586	1.89730	4.39768	11.87404

As the two-sided p-value is less than 0.05 we reject the null hypothesis and accept the alternative hypothesis. There is sufficient evidence to prove that the average blood pressure of patients aged above 55 is different from the average blood pressure of patients aged 55 and below. So BP level is influenced by the patient's age.

Is the average age of the person having heart disease equal to that the person who doesn't have heart disease?

A- Group of patients with heart disease
B- Group of patients without heart disease

H0: The average age of the person having heart disease is equal to the person who doesn't have heart disease (uA = uB)

H1: The average age of the person having heart disease is not equal to the person who doesn't have heart disease (uA ≠ uB)

Group Statistics

	Heart Disease	N	Mean	Std. Deviation	Std. Error Mean
Trf_age	Presence	107	55.9981	7.97963	.77142
	Absence	107	52.4830	9.48257	.91671

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Trf_age	Equal variances assumed	1.573	.211	2.934	212	.002	.004	3.51507	1.19810	1.15334	5.87679
	Equal variances not assumed			2.934	205.986	.002	.004	3.51507	1.19810	1.15295	5.87719

P-Value is less than the significance value of 0.05. i.e we can reject H0 and accept H1. Therefore, the average age of the person having heart disease is not equal to the person who doesn't have heart disease. So age is one of the influential factor of identifying heart disease.

Parametric test: ANOVA

1. One-way ANOVA

Has the chest pain type been influenced by cholesterol?

Hypothesis:

H0: There are no significant differences in the mean cholesterol value for each type of chest pain group. ($\mu 1= \mu 2= \mu 3= \mu 4$)

H1: There are significant differences in the mean value of cholesterol for each type of chest pain group. (atleast one $\mu i \neq \mu j$ i,j=1,2,3,4)

Descriptives

SMEAN(Cholesterol)

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum	Between-Component Variance
1		19	238.250	34.7232	7.9660	221.514	254.986	182.0	298.0	
2		40	247.107	37.8241	5.9805	235.010	259.204	195.0	325.0	
3		75	239.347	46.6637	5.3883	228.611	250.083	126.0	360.0	
4		117	250.223	42.6029	3.9386	242.422	258.024	149.0	354.0	
Total		251	245.570	42.6954	2.6949	240.263	250.878	126.0	360.0	
Model	Fixed Effects			42.6441	2.6917	240.269	250.872			
	Random Effects				3.0769	235.778	255.362			6.5814

ANOVA

SMEAN(Cholesterol)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6549.631	3	2183.210	1.201	.310
Within Groups	449173.805	247	1818.517		
Total	455723.436	250			

As for the outcome of the ANOVA, the p-value = 0.31 which is greater than the significance value of 0.05. That means H0 is accepted. Therefore, there are no significant differences in the mean cholesterol in each type of chest pain group. So, we can conclude that there is no influence on chest pain by cholesterol.

2. Two-way ANOVA

Has heart rate been influenced by sex and the number of vessels colored by fluoroscopy?

Hypothesis:

H0: There are no significant differences in the mean heart rate between the groups of numberofvesselsfluro

H1: There are significant differences in the mean heart rate between the groups of numberofvesselsfluro

H0: There are no significant differences in the mean heart rate between the groups of sex

H1: There are significant differences in the mean heart rate between the groups of sex

H0: numberofvesselsfluro has no effect on the effect of sex

H1: numberofvesselsfluro has an effect on the effect of sex

Tests of Between-Subjects Effects

Dependent Variable: Trf_MaxHR

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	16727.025[a]	7	2389.575	5.642	<.001	.140	39.495	.999
Intercept	1514120.414	1	1514120.414	3575.040	<.001	.937	3575.040	1.000
Numberofvesselsfluro	9447.007	3	3149.002	7.435	<.001	.084	22.306	.985
Sex	40.507	1	40.507	.096	.757	.000	.096	.061
Numberofvesselsfluro * Sex	1971.053	3	657.018	1.551	.202	.019	4.654	.406
Error	102493.158	242	423.525					
Total	5742292.033	250						
Corrected Total	119220.183	249						

a. R Squared = .140 (Adjusted R Squared = .115)

b. Computed using alpha = .05

The p-value of numberofvesselsfluro is less than the significance value of 0.05. So, the H0 is rejected and H1 is accepted i.e there are significant differences in the mean heart rate between the groups of numberofvesselsfluro.

Since the p-value of sex is > 0.05, H0 is accepted i.e there are no significant differences in the mean heart rate between the groups of sex.

Also, numberofvesselsfluro has no effect on the effect of sex as the p-value > 0.05. Therefore, there is no interaction between sex and the number of vessels colored by fluoroscopy.

Conclusion:

From the parametric test, we can conclude that a patient BP level is influenced by age and also age is a significant factor in identifying heart disease. Also, sex has no effect on the cholesterol level and heart rate of a patient.

Non-Parametric test

In this dataset, it is possible to perform only the following two non-parametric tests.

1. Kruskal Wallis Test
2. Mann Whitney U test

The other tests Wilcoxon test and Friedman test are not possible to apply as we don't have dependent samples.

Non-parametric test: Kruskal-Wallis test

Is there any difference in age between the different chest pain groups?

H0: There is no significant difference in the patient's age across different levels of chest pain.

H1: There is a significant difference in the patient's age across different levels of chest pain.

Ranks

	Chest pain type	N	Mean Rank
Age	1	19	142.11
	2	40	103.01
	3	75	120.56
	4	117	134.73
	Total	251	

This test can be applied only if we have more than two independent samples. Here as we have four categorize, the test can be applied.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Age is the same across categories of Chest pain type.	Independent-Samples Kruskal-Wallis Test	.070	Retain the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

As we accept H0, there is no significant difference in the patient's age across different levels of chest pain. So, age is not a influencing factor in determining chest pain type.

Is there any difference in age between the group of slopeofST?

H0: There is no significant difference in the patient's age across different groups of slopeofST

H1: There is a significant difference in the patient's age across different groups of slopeofST

Ranks

	Slope of ST	N	Mean Rank
Age	1	126	114.80
	2	111	137.53
	3	14	135.43
	Total	251	

Test Statistics[a,b]

	Age
Kruskal-Wallis H	6.042
df	2
Asymp. Sig.	.049

a. Kruskal Wallis Test

b. Grouping Variable: Slope of ST

Since the p-value < significance value of 0.05 we fail to reject H1. Therefore, there is a significant difference in the patient's age across different groups of slopeofST. So, Age is an influencing factor in determining SlopeofST of a patient.

Non-parametric test: Mann-Whitney U test

Is STdepression influence heart disease?

H0: There is no significant difference in the patient's STdepression across the heart disease categories

H1: There is a significant difference in the patient's STdepression across the heart disease categories.

Ranks

	Heart Disease	N	Mean Rank	Sum of Ranks
ST depression	Presence	105	155.74	16352.50
	Absence	142	100.53	14275.50
	Total	247		

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of ST depression is the same across categories of Heart Disease.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

The p-value of this test is less than a 5% significance level. That means we reject the Null hypothesis and accept the Alternative Hypothesis. So, there is a significant difference in the patient's STdepression across the heart disease categories. That means that patients with heart disease and without heart disease definitely have changes in STdepression.

Is maximum heart rate influence heart disease?

H0: There is no significant difference in the patient's maximum heart rate across the heart disease categories

H1: There is a significant difference in the patient's maximum heart rate across the heart disease categories

Ranks

	Heart Disease	N	Mean Rank	Sum of Ranks
Max HR	Presence	104	88.12	9164.50
	Absence	138	146.66	20238.50
	Total	242		

Test Statistics[a]

	Max HR
Mann-Whitney U	3704.500
Wilcoxon W	9164.500
Z	-6.441
Asymp. Sig. (2-tailed)	<.001

a. Grouping Variable: Heart Disease

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Max HR is the same across categories of Heart Disease.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

The p-value of this test is less than a 5% significance level. That means we reject the null hypothesis and accept the alternative hypothesis. So, there is a significant difference in the patient's maximum heart rate across the heart disease categories. That means the patients with heart disease and without heart disease definitely have differences in the maximum heart rate.

From these non-parametric tests, we can conclude that STdepression and maximum heart rate are important variables in predicting heart disease.

Correlation

Correlation analysis is a statistical technique that shows how variables are related to each other.

<u>Pearson Correlation</u>

Is there any correlation between the variables Age, Cholesterol, sex, Max HR, and heart disease?

H0: There is no correlation between the variables under consideration

H1: There is a correlation between the variables under consideration

Correlations

		Age	Cholesterol	Sex	Max HR	Heart Disease
Age	Pearson Correlation	1	.188**	-.082	-.395**	.202**
	Sig. (2-tailed)		.004	.196	<.001	.001
	N	251	235	251	242	251
Cholesterol	Pearson Correlation	.188**	1	-.107	-.054	.172**
	Sig. (2-tailed)	.004		.100	.422	.008
	N	235	235	235	227	235
Sex	Pearson Correlation	-.082	-.107	1	-.085	.332**
	Sig. (2-tailed)	.196	.100		.186	<.001
	N	251	235	251	242	251
Max HR	Pearson Correlation	-.395**	-.054	-.085	1	-.421**
	Sig. (2-tailed)	<.001	.422	.186		<.001
	N	242	227	242	242	242
Heart Disease	Pearson Correlation	.202**	.172**	.332**	-.421**	1
	Sig. (2-tailed)	.001	.008	<.001	<.001	
	N	251	235	251	242	251

**. Correlation is significant at the 0.01 level (2-tailed).

Pearson correlation can be used for the variables which are normally distributed. It can be observed from the above table that the p-value for all four variables related to the heart disease variable is less than 5%. That means we can accept the H0, there is a correlation between the variables. The variable maxHR is 42% negatively correlated with heart disease and the variables age, cholesterol, and sex are 20%, 17%, and 33% positively correlated with heart disease respectively.

Spearman Correlation

Is there any correlation between the variables BP, ST Depression, FBS over 120, EKG Results, and heart disease?

H0: There is no correlation between the variables under consideration

H1: In contrast, the alternative hypothesis assumes that there is a correlation

Correlations

			Heart Disease	BP	ST depression	FBS over 120	EKG results
Spearman's rho	Heart Disease	Correlation Coefficient	1.000	.057	.390**	-.034	.196**
		Sig. (2-tailed)	.	.366	<.001	.590	.002
		N	251	251	247	251	251
	BP	Correlation Coefficient	.057	1.000	.138*	.096	.124
		Sig. (2-tailed)	.366	.	.030	.128	.051
		N	251	251	247	251	251
	ST depression	Correlation Coefficient	.390**	.138*	1.000	-.020	.102
		Sig. (2-tailed)	<.001	.030	.	.760	.110
		N	247	247	247	247	247
	FBS over 120	Correlation Coefficient	-.034	.096	-.020	1.000	.062
		Sig. (2-tailed)	.590	.128	.760	.	.329
		N	251	251	247	251	251
	EKG results	Correlation Coefficient	.196**	.124	.102	.062	1.000
		Sig. (2-tailed)	.002	.051	.110	.329	.
		N	251	251	247	251	251

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Spearman correlation can be applied to non-normally distributed data. The above table shows that the p-value of the variables ST depression and EKG results related to the variable heart disease is less than 5% therefore H0 can be accepted. So, there is a 39% and 20% positive correlation respectively. BP and FBS over 120 are not correlated with the response variable heart disease as the p-value is greater than 5%.

Regression

Changes in one or more explanatory variables can be associated with changes in the dependent variable in a regression model.

Simple linear regression

Does variable Age is a significant predictor for Cholesterol?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.188[a]	.035	.031	43.432

a. Predictors: (Constant), Age

b. Dependent Variable: Cholesterol

From the above table, R^2=3.5% R^2=3.5% reveals that the regression model explains 3.5% of the variability observed in the target variable. A regression model is generally better if there is more variance observed. Adjusted R^2 indicates that the model perfectly predicts the values in the target field by only 3.1%.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16042.001	1	16042.001	8.504	.004[b]
	Residual	439507.685	233	1886.299		
	Total	455549.685	234			

a. Dependent Variable: Cholesterol

b. Predictors: (Constant), Age

Coefficients[a]

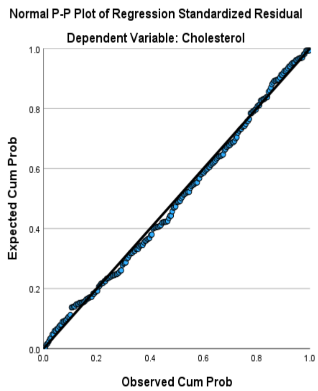
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	196.938	16.842		11.693	<.001
	Age	.895	.307	.188	2.916	.004

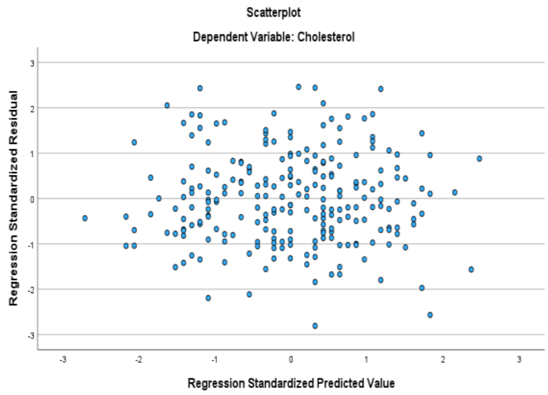
a. Dependent Variable: Cholesterol

Since the p-value of the model is less than 5%, the model is effective, and the predictor variable age significantly predicts the dependent variable cholesterol.

The model can be written as follows:

Y(Cholesterol) = 0.895*Age + 196.938





The above plot shows the 3.5% variability across the straight line. The Scatter plot indicates the strength and direction of the associated variables.

Since our predictor variable is a categorical variable and there is no correlation with more than two numerical variables multiple regression model cannot be used to predict.

Logistic regression

Logistic regression is a method to determine the cause-effect relationship between independent variables with dependent variables. In this, the output variable must be a categorical variable.

Can the variables Chest pain type, Cholesterol, EKG results, ST depression, Number of vessels Fluro, Thallium, Max HR, and Age significantly predict heart disease?

Dependent Variable Encoding

Original Value	Internal Value
Absence	0
Presence	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			Heart Disease		Percentage Correct
Observed			Absence	Presence	
Step 0	Heart Disease	Absence	131	0	100.0
		Presence	93	0	.0
	Overall Percentage				58.5

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.343	.136	6.384	1	.012	.710

Block 0 gives information about our data. The data without heart disease 131 cases and 93 cases with heart disease. The overall percentage is 58.5% and the p-value is 0.012 which is less than 5% this means the Target value is predicted correctly.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	166.197	13	<.001
	Block	166.197	13	<.001
	Model	166.197	13	<.001

The Chi-square test significantly predicts the target value because p- the value is less than 5%.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	137.855[a]	.524	.705

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

From the above summary table, we can conclude that the cox & snell R square test and Nagelkerke R Square test give p-value of 0.524 and 0.705 which is greater than 5% means the Independent variables cannot predict the Dependent variable Heart Disease. Compared to both tests, the Nagelkerke R Square test is most considerable.

Classification Table[a]

Observed			Predicted: Heart Disease: Absence	Predicted: Heart Disease: Presence	Predicted: Percentage Correct
Step 1	Heart Disease	Absence	117	14	89.3
		Presence	18	75	80.6
	Overall Percentage				85.7

a. The cut value is .500

From the above table, it can be seen that the model correctly predicts the cases when the heart disease absence is 117 out of 131. 14 cases are wrongly predicted as present. Meanwhile, the cases correctly predicted when heart disease presence is 75 out of 93. 18 cases are wrongly predicted as absent when the actual case is present. The overall percentage is about 85.7% accurate.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex	1.170	.626	3.491	1	.062	3.223
	Age	.002	.030	.004	1	.951	1.002
	Chest pain type	.633	.239	7.036	1	.008	1.883
	BP	.486	.360	1.818	1	.178	1.626
	Cholesterol	.014	.006	6.236	1	.013	1.014
	FBS over 120	-.666	.657	1.029	1	.310	.514
	EKG results	.723	.248	8.505	1	.004	2.061
	Max HR	-.015	.013	1.287	1	.257	.986
	Exercise angina	.930	.494	3.538	1	.060	2.533
	ST depression	.580	.281	4.260	1	.039	1.786
	Slope of ST	.149	.458	.106	1	.745	1.161
	Number of vessels fluro	1.141	.294	15.113	1	<.001	3.131
	Thallium	.522	.135	14.982	1	<.001	1.686
	Constant	-15.149	5.245	8.342	1	.004	.000

a. Variable(s) entered on step 1: Sex, Age, Chest pain type, BP, Cholesterol, FBS over 120, EKG results, Max HR, Exercise angina, ST depression, Slope of ST, Number of vessels fluro, Thallium.

From the above table, we can conclude that the significant p-value is less than 5% for the variables Chest pain type, Cholesterol, EKG results, ST depression, Number of vessels fluro, and Thallium. So, those variables significantly predict the dependent variable of heart disease. Other variables are not significant in predicting heart disease.

Conclusion

From the correlation and regression analysis, we conclude that the variables Chest pain type, Cholesterol, EKG results, ST depression, Number of vessels fluro, and Thallium significantly predict heart disease and the other variables in the data set are not influential factors for heart disease.

References

Gárate-Escamila, A.K., El Hassani, A.H. and Andrès, E., 2020. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, p.100330.

Guidi, G., Pettenati, M.C., Melillo, P. and Iadanza, E., 2014. A machine learning system to improve heart failure patient assistance. IEEE journal of biomedical and health informatics, 18(6), pp.1750-1756.

Goenka, S., Prabhakaran, D., Ajay, V.S. and Reddy, K.S., 2009. Preventing cardiovascular disease in India–translating evidence to action. Current science, pp.367-377.

ANALYSIS OF CO2 EMISSION DATA

Introduction

Carbon dioxide emissions are the primary driver of global climate change. It's widely recognized that to avoid the worst impacts of climate change, the world needs to urgently reduce emissions. But how this responsibility is shared between regions, countries, and individuals has been an endless point of contention in international discussions.

This debate arises from the various ways in which emissions are compared: as annual emissions by country; emissions per person; historical contributions; and whether they adjust for traded goods and services. These metrics can tell very different stories.

Dataset Description:

The data is about CO2 emissions worldwide. This contains different kinds of indexes regarding co2 emissions which is useful for measuring key values for this issue. The dataset contains 25204 observations with 58 variables for 245 countries from 1750 to 2020.

Missing values

There are no missing values in the ISO code, Country, and Year variables. All the other variables have missing values. We present the missing values country-wise for each variable whose percentage of missing values is less than 50% and greater than 10%. We assume that it will have no meaning when taking the variables in which missing values are more than 50% and it is normal to have missing values less than 10%.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
year	25204	100.0%	0	0.0%	25204	100.0%
co2	23949	95.0%	1255	5.0%	25204	100.0%
consumption_co2	3976	15.8%	21228	84.2%	25204	100.0%
co2_growth_prct	24931	98.9%	273	1.1%	25204	100.0%
co2_growth_abs	23585	93.6%	1619	6.4%	25204	100.0%
trade_co2	3976	15.8%	21228	84.2%	25204	100.0%
co2_per_capita	23307	92.5%	1897	7.5%	25204	100.0%
consumption_co2_per_capita	3976	15.8%	21228	84.2%	25204	100.0%
share_global_co2	23949	95.0%	1255	5.0%	25204	100.0%
cumulative_co2	23949	95.0%	1255	5.0%	25204	100.0%
share_global_cumulative_co2	23949	95.0%	1255	5.0%	25204	100.0%
co2_per_gdp	15389	61.1%	9815	38.9%	25204	100.0%
consumption_co2_per_gdp	3761	14.9%	21443	85.1%	25204	100.0%
co2_per_unit_energy	9141	36.3%	16063	63.7%	25204	100.0%
coal_co2	17188	68.2%	8016	31.8%	25204	100.0%
cement_co2	12248	48.6%	12956	51.4%	25204	100.0%
flaring_co2	4382	17.4%	20822	82.6%	25204	100.0%
gas_co2	8845	35.1%	16359	64.9%	25204	100.0%
oil_co2	20539	81.5%	4665	18.5%	25204	100.0%
other_industry_co2	1999	7.9%	23205	92.1%	25204	100.0%
cement_co2_per_capita	12218	48.5%	12986	51.5%	25204	100.0%
coal_co2_per_capita	16860	66.9%	8344	33.1%	25204	100.0%
flaring_co2_per_capita	4381	17.4%	20823	82.6%	25204	100.0%
gas_co2_per_capita	8835	35.1%	16369	64.9%	25204	100.0%
oil_co2_per_capita	20181	80.1%	5023	19.9%	25204	100.0%
other_co2_per_capita	1999	7.9%	23205	92.1%	25204	100.0%
trade_co2_share	3976	15.8%	21228	84.2%	25204	100.0%
share_global_cement_co2	12248	48.6%	12956	51.4%	25204	100.0%
share_global_coal_co2	17188	68.2%	8016	31.8%	25204	100.0%
share_global_flaring_co2	4382	17.4%	20822	82.6%	25204	100.0%
share_global_gas_co2	8845	35.1%	16359	64.9%	25204	100.0%
share_global_oil_co2	20539	81.5%	4665	18.5%	25204	100.0%
share_global_other_co2	1999	7.9%	23205	92.1%	25204	100.0%
cumulative_cement_co2	12248	48.6%	12956	51.4%	25204	100.0%
cumulative_coal_co2	17188	68.2%	8016	31.8%	25204	100.0%
cumulative_flaring_co2	4382	17.4%	20822	82.6%	25204	100.0%
cumulative_gas_co2	8845	35.1%	16359	64.9%	25204	100.0%
cumulative_oil_co2	20539	81.5%	4665	18.5%	25204	100.0%
cumulative_other_co2	1999	7.9%	23205	92.1%	25204	100.0%
share_global_cumulative_cement_co2	12248	48.6%	12956	51.4%	25204	100.0%
share_global_cumulative_coal_co2	17188	68.2%	8016	31.8%	25204	100.0%
share_global_cumulative_flaring_co2	4382	17.4%	20822	82.6%	25204	100.0%
share_global_cumulative_gas_co2	8845	35.1%	16359	64.9%	25204	100.0%
share_global_cumulative_oil_co2	20539	81.5%	4665	18.5%	25204	100.0%
share_global_cumulative_other_co2	1999	7.9%	23205	92.1%	25204	100.0%
total_ghg	5208	20.7%	19996	79.3%	25204	100.0%
ghg_per_capita	5155	20.5%	20049	79.5%	25204	100.0%
methane	5211	20.7%	19993	79.3%	25204	100.0%
methane_per_capita	5157	20.5%	20047	79.5%	25204	100.0%
nitrous_oxide	5211	20.7%	19993	79.3%	25204	100.0%
nitrous_oxide_per_capita	5157	20.5%	20047	79.5%	25204	100.0%
population	22878	90.8%	2326	9.2%	25204	100.0%
gdp	13538	53.7%	11666	46.3%	25204	100.0%
primary_energy_consumption	8690	34.5%	16514	65.5%	25204	100.0%
energy_per_capita	8681	34.4%	16523	65.6%	25204	100.0%
energy_per_gdp	6803	27.0%	18401	73.0%	25204	100.0%

CO2	
Country	Missing value %
Puerto R	99.0%
Kuwaiti	96.7%
Leeward	90.1%
French W	87.5%
French E	87.3%
Christma	72.5%
Ryukyu I	69.6%
St. Kitt	62.5%
Eritrea	59.8%
Panama C	57.7%
Micrones	48.2%
Antarcti	38.2%
Ireland	35.7%
Saint He	22.6%
Belgium	12.3%

Coal_CO2	
Country	Missing value %
Brunei	98.9%
Bahrain	96.6%
Benin	93.7%
French E	91.5%
Leeward	91.5%
Guyana	90.1%
Rwanda	90.1%
French W	88.9%
Cape Ver	88.7%
Jordan	88.7%
Paraguay	88.7%
Vanuatu	88.1%
Trinidad	86.6%
Haiti	84.5%
Barbados	83.9%
Ryukyu I	82.6%
Bermuda	81.7%
Yemen	81.7%
Papua Ne	80.3%
Sierra L	77.5%
Costa Ri	76.1%
Guatemal	72.5%
Senegal	69.8%
South Su	69.0%
Sudan	69.0%
Cambodia	68.2%
United A	67.7%
Faeroe I	67.6%
Guadelou	67.6%
Angola	66.2%
Suriname	63.4%
El Salva	62.0%
Saint Pi	62.0%
Bolivia	61.3%
Honduras	60.6%
Ethiopia	60.0%
Dominica	59.7%
Reunion	59.2%
Iraq	55.3%
Macao	53.7%
Greenlan	53.5%
Mauritan	51.6%

CO2_growth_prct	
Country	Missing value %
Micrones	50.0%

CO2_per_gdp	
Country	Missing value %
World	94.1%
Armenia	79.1%
Azerbaij	79.1%
Belarus	79.1%
Estonia	79.1%
Georgia	79.1%
Kazakhst	79.1%
Kyrgyzst	79.1%
Latvia	79.1%
Lithuani	79.1%
Moldova	79.1%
Tajikist	79.1%
Turkmeni	79.1%
Ukraine	79.1%
Uzbekist	79.1%
Slovakia	78.9%
North Ko	74.1%
Czechia	69.6%
Russia	69.1%
Poland	59.7%
Bosnia a	50.7%
Croatia	50.7%
Monteneg	50.7%
North Ma	50.7%
Serbia	50.7%
Slovenia	50.7%

Oil_CO2	
Country	Missing value %
Puerto R	99.0%
Kuwaiti	96.7%
Leeward	90.1%
French W	88.9%
French E	87.3%
Christma	72.5%
Ryukyu I	69.6%
Eritrea	67.1%
St. Kitt	62.5%
Denmark	59.6%
Ireland	58.5%
Belgium	58.0%
Panama C	57.7%
Eswatini	56.3%
Turkey	52.6%

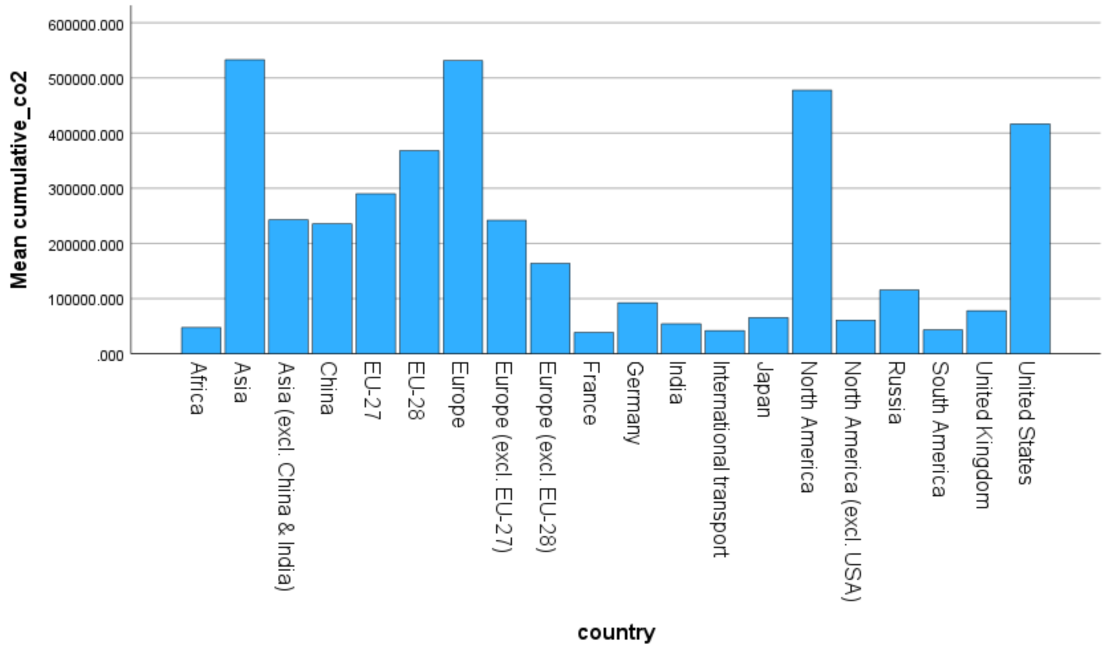
gdp	
Country	Missing value %
World	94.1%
Armenia	79.1%
Azerbaij	79.1%
Belarus	79.1%
Estonia	79.1%
Georgia	79.1%
Kazakhst	79.1%
Kyrgyzst	79.1%
Latvia	79.1%
Lithuani	79.1%
Moldova	79.1%
Tajikist	79.1%
Turkmeni	79.1%
Ukraine	79.1%
Uzbekist	79.1%
Slovakia	78.9%
North Ko	74.1%
Czechia	69.6%
Russia	69.1%
Poland	59.7%

Population	
Country	Missing value %
North Am	53.0%
Micrones	48.2%

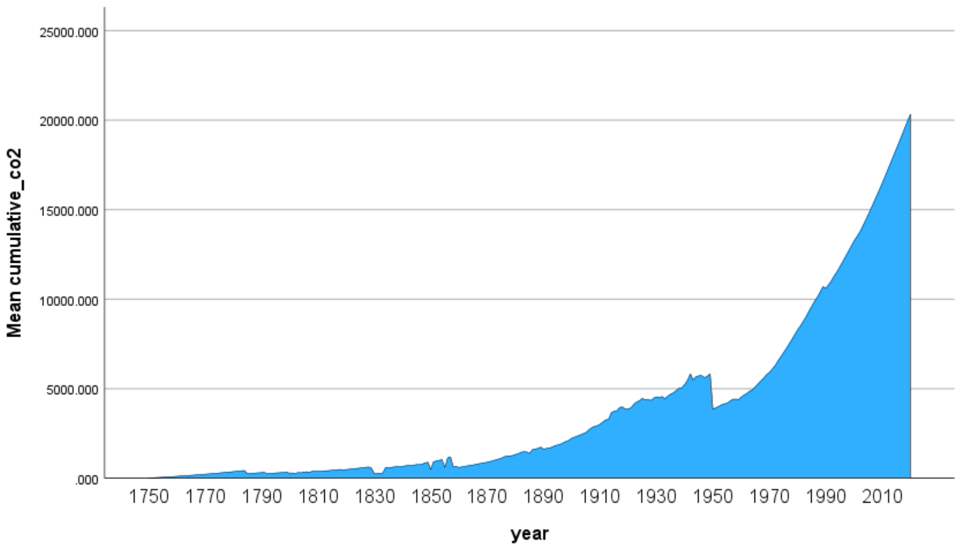
Coal_CO2_per_capita	
Country	Missing value %
Brunei	98.9%
Bahrain	96.6%
Benin	93.7%
Guyana	90.1%
Rwanda	90.1%
Cape Ver	88.7%
Jordan	88.7%
Paraguay	88.7%
Vanuatu	88.1%
Trinidad	86.6%
Haiti	84.5%
Barbados	83.9%
Bermuda	81.7%
Yemen	81.7%
Papua Ne	80.3%
Sierra L	77.5%
Costa Ri	76.1%
Guatemal	72.5%
Senegal	69.8%
South Su	69.0%
Sudan	69.0%
Cambodia	68.2%
United A	67.7%
Faeroe I	67.6%
Guadelou	67.6%
Angola	66.2%
Suriname	63.4%
El Salva	62.0%
Saint Pi	62.0%
Bolivia	61.3%
Honduras	60.6%
Ethiopia	60.0%
Dominica	59.7%
Reunion	59.2%
Iraq	55.3%
Macao	53.7%
Greenlan	53.5%
Mauritan	51.6%
Cameroon	50.7%

The missing value is analyzed by the country for each variable and excluded countries with a higher percentage of missing values in our primary variables.

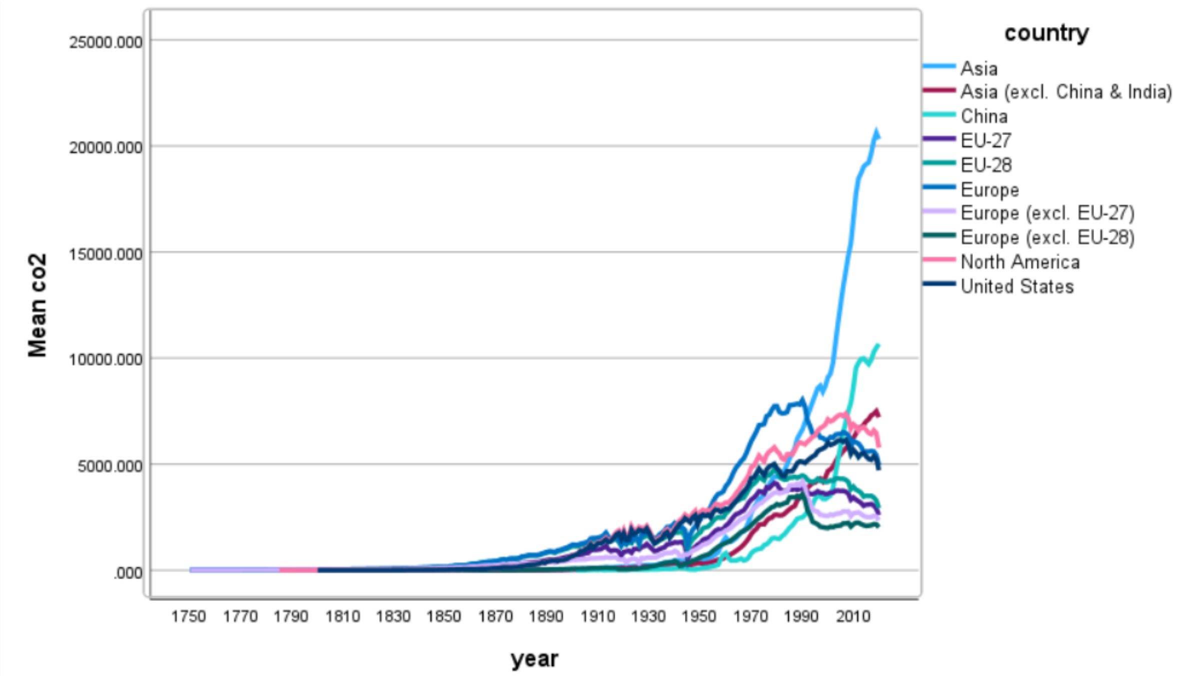
Descriptive Analysis



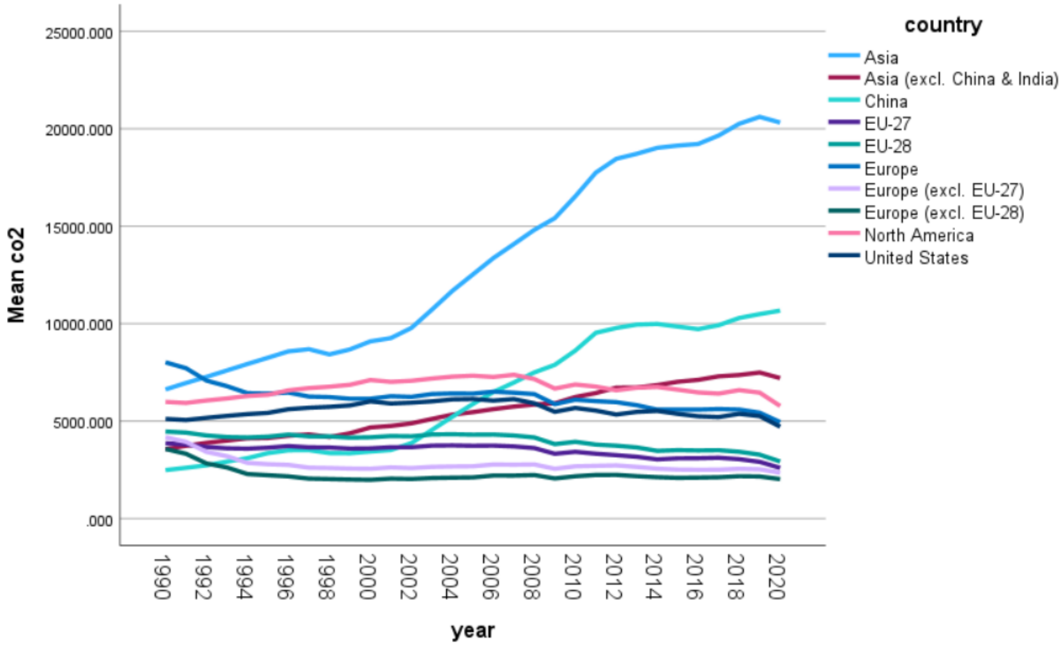
The above chart contains the top 20 countries of cumulative production-based CO2 emissions until the year 2020. From the above figure, it can be seen who has contributed most to global co2 emissions. Asia, Europe, North America, and the United States have emitted the most to date in the order.



The chart describes how cumulative production-based emissions of CO2 grow since the first year of data availability till 2020 across the world. There was a drop in the 1950 and then it increases till date.



The above chart shows the distribution of annual production-based emissions of CO_2 only for the countries that have higher cumulative CO_2 emissions. Asia emits the highest amount of CO_2 since 1992 and emitted nearly 20000 million tonnes in 2020. China is in the 2nd place and emitted about 10000 million tonnes in 2020.



The above chart shows the distribution of annual production-based emissions of CO_2 only for the countries that have higher cumulative CO_2 emissions from 1990. What becomes clear when we look at emissions across the world today is that the countries with the highest emissions over history are not always the biggest emitters today. Europe was the highest emitter before 1990 and the amount has started falling since 1990 and it contributed only 5000 million tones in 2020. China and Asia have contributed a small amount in history, and they are in the first two places in the last decade.

Checking for Normality

The selected variables are checked for normality.

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
co2	.411	13410	<.001
cumulative_co2	.418	13410	<.001
co2_per_gdp	.175	13410	<.001
population	.387	13410	<.001
gdp	.401	13410	<.001
co2_per_capita	.280	13410	<.001
share_global_cumulative_co2	.417	13410	<.001
share_global_co2	.408	13410	<.001

a. Lilliefors Significance Correction

As shown in the above table, the p-value of all the tested variables is less than the significance value, these variables are not normally distributed.

Therefore, we cannot conduct parametric tests for these variables.

Nonparametric test: Mann-Whitney test

The variable CO2 is categorized into two categories the above-average and the below-average.

Is a country's Population an influential factor in CO2 emission?

H0: There is no significant difference in the distribution of population across the categories of CO2.
H1: There is a significant difference in the distribution of population across the categories of CO2.

Ranks

	CO2_category	N	Mean Rank	Sum of Ranks
population	1.00	13332	8085.56	107796642.50
	2.00	7700	14725.50	113386385.50
	Total	21032		

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of population is the same across categories of CO2_category.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

As the p-value is less than the significance value H0 is rejected. So, there is a significant difference in the distribution of population across the categories of CO2. Countries that have high populations have high production-based CO2 emissions.

Is a country's PPP an influential factor in CO2 emission?

H0: There is no significant difference in the distribution of GDP across the categories of CO2.
H1: There is a significant difference in the distribution of GDP across the categories of CO2.

Ranks

	CO2_category	N	Mean Rank	Sum of Ranks
gdp	1.00	6837	3785.88	25884042.00
	2.00	6084	9467.22	57598539.00
	Total	12921		

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of gdp is the same across categories of CO2_category.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

As the p-value is less than the significance value, H0 is rejected. So, there is a significant difference in the distribution of GDP across the categories of CO2. Therefore, a country's GDP is an influencing factor in CO2 emission.

Conclusion

The global CO_2 emissions data set is analyzed. To do that, missing values are checked and then some descriptive analysis is conducted to draw useful insights from this data. Then Normality check was done. As the primary variables do not follow the normal distribution, non-parametric tests were applied. According to this study, a country's GDP and population have the greatest influence on CO_2 emissions.

References

https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv