Healthcare Data Analysis

About this file: The dataset is intended for educational and non-commercial use. It is entirely synthetic and does not contain real patient data. This dataset consists of 10,000 records, each representing a synthetic patient healthcare record. It includes various attributes, such as patient demographics, medical conditions, admission details, and many more.

Dataset Information:

Each column provides specific information about the patient, their admission, and the healthcare services provided, making this dataset suitable for various data analysis and modeling tasks in the healthcare domain. Here's a brief explanation of each column in the dataset -

Name: This column represents the name of the patient associated with the healthcare record.

Age: The age of the patient at the time of admission, expressed in years.

Gender: Indicates the gender of the patient,

either 1 for "Male" or 2 for "Female."

Blood Type: The patient's blood type, which can be one of the common blood types (1 for "A+", 2 for "A-", 3 for "B+", 4 for "B-", 5 for "O+", 6 for "O-", 7 for "AB+", and 8 for "AB-").

Medical Condition: This column specifies the primary medical condition or diagnosis associated with the patient, such as

1 for "Diabetes," 2 for "Hypertension," 3 for "Asthma," 4 for “Arthritis”, 5 for “Obesity”, 6 for “Cancer”.

Date of Admission: The date on which the patient was admitted to the healthcare facility.

Doctor: The name of the doctor responsible for the patient's care during their admission.

Hospital: Identifies the healthcare facility or hospital where the patient was admitted.

Insurance Provider: This column indicates the patient's insurance provider, which can be one of several options, including

1 for "Aetna," 2 for "Blue Cross," 3 for "Cigna," 4 for "UnitedHealthcare," and 5 for "Medicare."

Billing Amount: The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.

Room Number: The room number where the patient was accommodated during their admission.

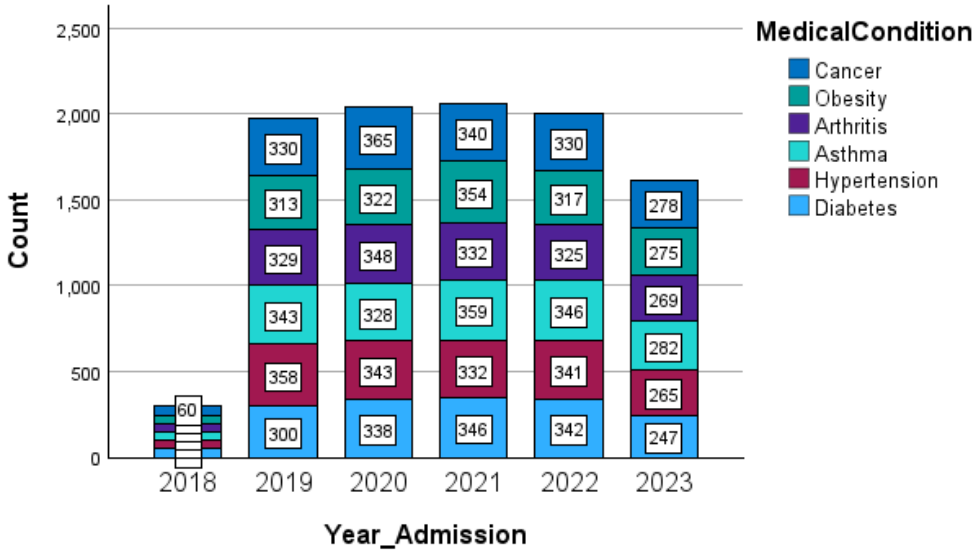
Admission Type: Specifies the type of admission, which can be 1 for "Emergency," 2 for "Elective," or 3 for "Urgent," reflecting the circumstances of the admission.

Discharge Date: The date on which the patient was discharged from the healthcare facility, based on the admission date and a random number of days within a realistic range.

Medication: Identifies a medication prescribed or administered to the patient during their admission. Examples include 1 for "Aspirin," 2 for "Ibuprofen," 3 for "Penicillin," 4 for "Paracetamol," and 5 for "Lipitor."

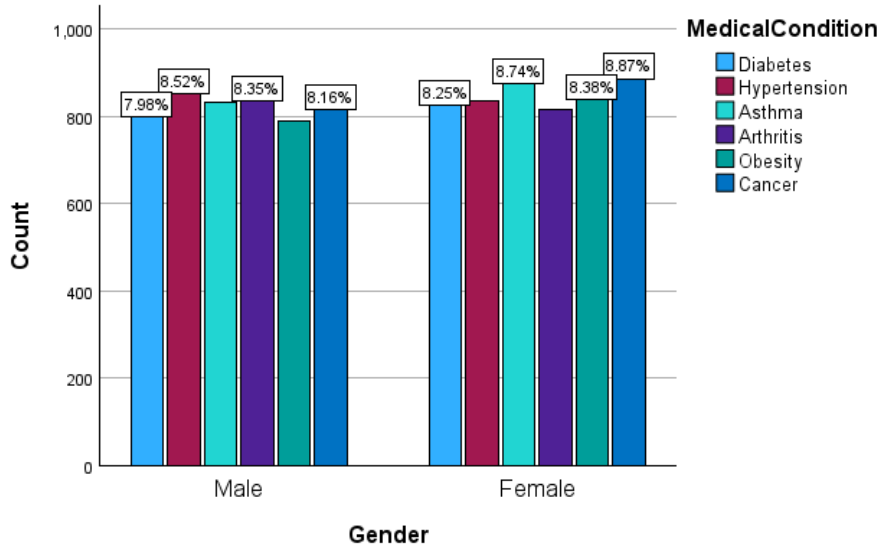
Test Results: Describes the results of a medical test conducted during the patient's admission. Possible values include 1 for "Normal," 2 for "Abnormal," or 3 for "Inconclusive," indicating the outcome of the test.

1) How many the number of patients with different medical conditions have been recorded.



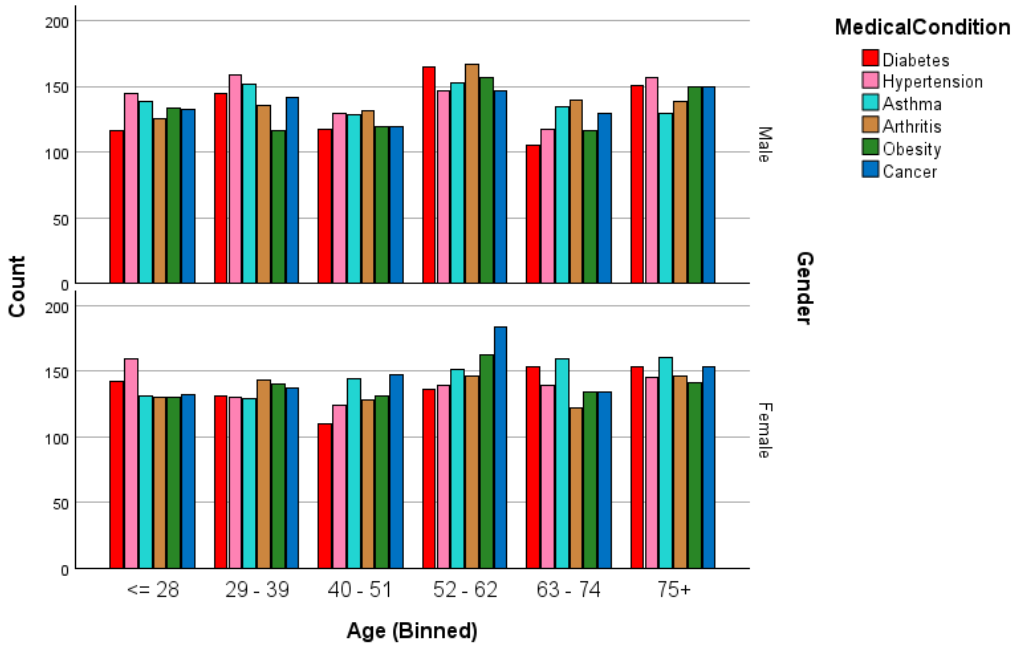
This bar chart shows that maximum number of patients were recorded in year 2021 and in the year 2018 there was minimum number of patients were accounted. Furthermore, maximum patients were accounted in year 2020, which was decreasing over the years. Asthma and Diabetes patients were maximum in year 2021.

2) Variation in number of different medical condition patients gender wise.



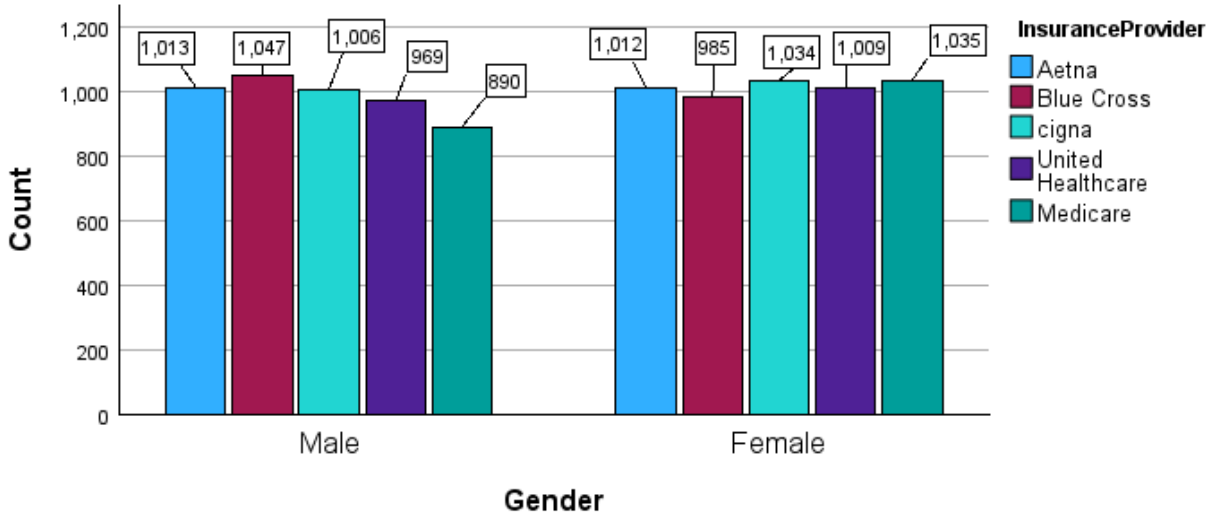
In this clustered column chart, it is clearly visible that 8.87% females were coping with Cancer disease out of all patients, this percentage was only 8.16% in males. I females' percentage of medical conditions of Diabetes, Asthma, Obesity and Cancer were recorded more than their male counterpart. Hypertension and Arthritis patients' percentage was more in male than females.

3) Which age group people are more likely to get different medical condition along with their genders.



As clearly visible from this column chart, Hypertension was maximum in 75+ years males and in females under 28 years group. Overall, in males' percentage of diseases were recorded more than females.

4) How many patients are there with different Insurance provider along with their gender?



Maximum number of males have joined Blue Cross insurance company. Cigna and Medicare have same number of females who are receiving insurance benefits from these company. Medicare has least number of males connected with them.

5) Billing amount is normally distributed amongst all patients. (Explore)

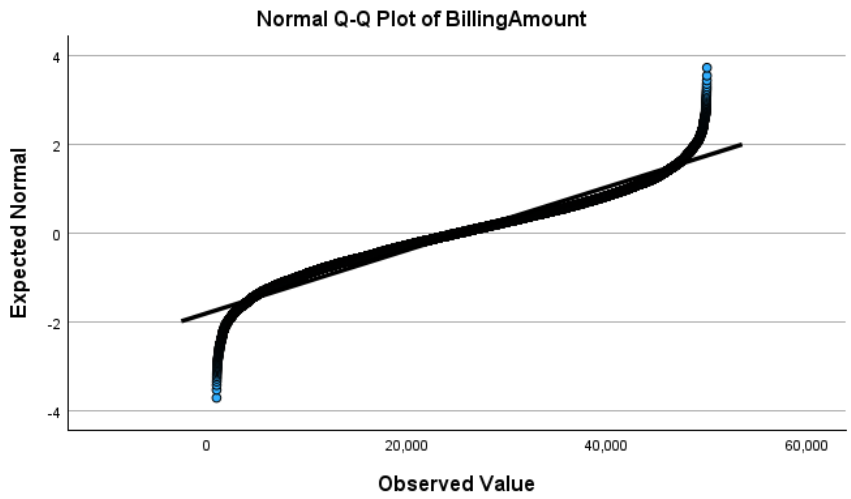
Null Hypothesis – Distribution is normal.

Alternate Hypothesis – Distribution is not normal.

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
BillingAmount	.059	10000	<.001

a. Lilliefors Significance Correction



As shown in this table, p-value of normality test is 0.001 which is less than significant value that is 0.05. hence, we **reject the null hypothesis.** This means billing amount is not normally distributed amongst the patients.

6) **Research Hypothesis – It is assumed that on an average patient is spending 30,000 dollars on his medication and hospital expenses. (One sample t-test)**

Statistical Hypothesis – The mean of sample and mean of population is same.

Null Hypothesis – There is no significant difference between sample mean and population mean. $\mu = \$30,000$

Alternate Hypothesis – There is significant difference between sample mean and population mean. $\mu \neq \$30,000$

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
BillingAmount	10000	25516.80677777	14067.292709243	140.672927092

One-Sample Test

	Test Value = 35000						
			Significance			95% Confidence Interval of the Difference	
	t	df	One-Sided p	Two-Sided p	Mean Difference	Lower	Upper
BillingAmount	-67.413	9999	<.001	<.001	-9483.193222228	-9758.94047167	-9207.44597278

As the p value of One sample T-test (-67.413) is 0.001 which is less than significant level (0.05), we **reject null hypothesis**. This define that there is significant difference between sample and population mean. The sample mean is 25,500 dollars which is different from population mean.

7) **Research Hypothesis – Male patients are paying more bill on their medication than female patients. (Independent sample t-test)**

Null Hypothesis – There is no significant difference between male and female billing amount.

Alternate Hypothesis - There is significant difference between male and female billing amount.

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
BillingAmount	Male	4925	25550.21593313	14106.549000909	201.010000365
	Female	5075	25484.38508514	14030.405749752	196.948286830

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means					
					Significance			
	F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Differ
	.431	.512	.234	9998	.408	.815	65.830847992	281.39081
			.234	9985.474	.408	.815	65.830847992	281.41365

As p-value is 0.512 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no significant difference between male and female billing amount.

8) **Research Hypothesis – Age of Cancer patient is more than any other disease patients. (One way ANOVA)**

Null Hypothesis – There is no significant difference between average age for different diseases.

Alternate Hypothesis - There is significant difference between average age for different diseases.

ANOVA

Age

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1154.406	5	230.881	.602	.699
Within Groups	3835740.745	9994	383.804		
Total	3836895.152	9999			

Post Hoc Tests

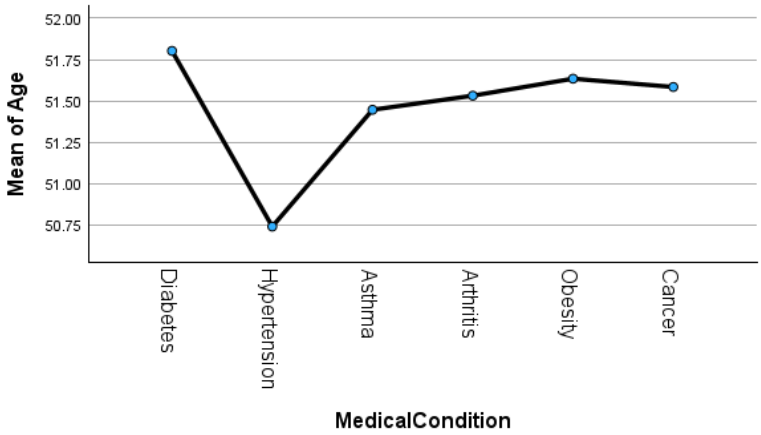
Multiple Comparisons

Dependent Variable: Age

Tukey HSD

(I) MedicalCondition	(J) MedicalCondition	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Diabetes	Hypertension	1.065	.681	.623	-.88	3.01
	Asthma	.357	.679	.995	-1.58	2.29
	Arthritis	.271	.685	.999	-1.68	2.22
	Obesity	.168	.687	1.000	-1.79	2.13
	Cancer	.219	.680	1.000	-1.72	2.16
Hypertension	Diabetes	-1.065	.681	.623	-3.01	.88
	Asthma	-.708	.672	.900	-2.62	1.21
	Arthritis	-.793	.678	.851	-2.73	1.14
	Obesity	-.896	.681	.776	-2.84	1.04

As the p-value is 0.699 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no significant difference between the average age for different medical conditions.



There is little difference between mean of Diabetes and Hypertension patients' age otherwise all the disease has same mean. Hence, age of all disease patient is almost same.

9) **Research Hypothesis – There is no variation in billing amount based on medical condition and medication. (Two-way ANOVA Linear Model Univariate)**

Null Hypothesis – There is no variation in billing amount based on medical condition and medication.

Alternate Hypothesis - There is a variation in billing amount based on medical condition and medication.

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
BillingAmount	Based on Mean	1.383	29	9970	.083
	Based on Median	1.355	29	9970	.097
	Based on Median and with adjusted df	1.355	29	9945.719	.097
	Based on trimmed mean	1.381	29	9970	.083

As the p-value of Levene's test is 0.097 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no variation in billing amount for different medical conditions and medications.

Tests of Between-Subjects Effects

Dependent Variable: BillingAmount

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6994267566.604[a]	29	241181640.228	1.220	.193
Intercept	6492928001131.895	1	6492928001131.895	32831.898	<.001
MedicalCondition	889228828.466	5	177845765.693	.899	.480
Medication	1451118152.587	4	362779538.147	1.834	.119
MedicalCondition * Medication	4508906356.900	20	225445317.845	1.140	.299
Error	1971695085384.507	9970	197762796.929		
Total	8489763634292.481	10000			
Corrected Total	1978689352951.112	9999			

a. R Squared = .004 (Adjusted R Squared = .001)

As the p-value is 0.480 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no variation in billing amount for different medical conditions.

10) Research Hypothesis – Proportion of emergency admission is more in patients of above 62 age group than patients below age group 39. (Chi-Square Test - Crosstab)

Null Hypothesis – There is no significant association in admission type and age.

Alternate Hypothesis - There is a significant association in admission type and age.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	10.201[a]	10	.423
Likelihood Ratio	10.189	10	.424
Linear-by-Linear Association	.052	1	.820

N of Valid Cases	10000		

As the p-value of Pearson Chi-Square (10.201) is 0.423 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no significant association in admission type and age.

Age (Binned) * Admission Type Crosstabulation

			Admission Type			
			Emergency	Elective	Urgent	Total
Age (Binned)	<= 28	Count	540	526	546	1612
		% within Age (Binned)	33.5%	32.6%	33.9%	100.0%
		% within AdmissionType	16.0%	16.2%	16.1%	16.1%
	29 - 39	Count	563	528	566	1657
		% within Age (Binned)	34.0%	31.9%	34.2%	100.0%
		% within AdmissionType	16.7%	16.3%	16.7%	16.6%
	40 - 51	Count	511	505	511	1527
		% within Age (Binned)	33.5%	33.1%	33.5%	100.0%
		% within AdmissionType	15.2%	15.6%	15.1%	15.3%
	52 - 62	Count	649	605	598	1852
		% within Age (Binned)	35.0%	32.7%	32.3%	100.0%
		% within AdmissionType	19.3%	18.7%	17.6%	18.5%
	63 - 74	Count	493	508	580	1581
		% within Age (Binned)	31.2%	32.1%	36.7%	100.0%
		% within AdmissionType	14.6%	15.7%	17.1%	15.8%
	75+	Count	611	570	590	1771
		% within Age (Binned)	34.5%	32.2%	33.3%	100.0%
		% within AdmissionType	18.1%	17.6%	17.4%	17.7%
Total		Count	3367	3242	3391	10000
		% within Age (Binned)	33.7%	32.4%	33.9%	100.0%
		% within AdmissionType	100.0%	100.0%	100.0%	100.0%

This crosstab describes that 33.5% emergency, 32.6% elective and 33.9% urgent admissions were recorded in hospital of under 28 years patient. Maximum number of admissions in emergency was noticed in 52-62 years age group.

11) Research Hypothesis – Billing amount and age are positively correlated if medical condition is factored out. (Partial Correlation)

Null Hypothesis – There is no significant relation between billing amount and age provided that medical condition is factored out.

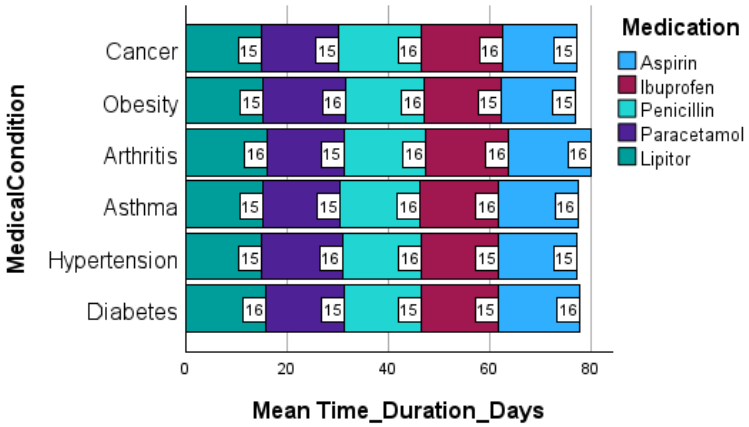
Alternate Hypothesis - There is no significant relation between billing amount and age provided that medical condition is factored out.

Correlations

Control Variables			BillingAmount	Age
MedicalCondition	BillingAmount	Correlation	1.000	-.009
		Significance (2-tailed)	.	.344
		df	0	9997
	Age	Correlation	-.009	1.000
		Significance (2-tailed)	.344	.
		df	9997	0

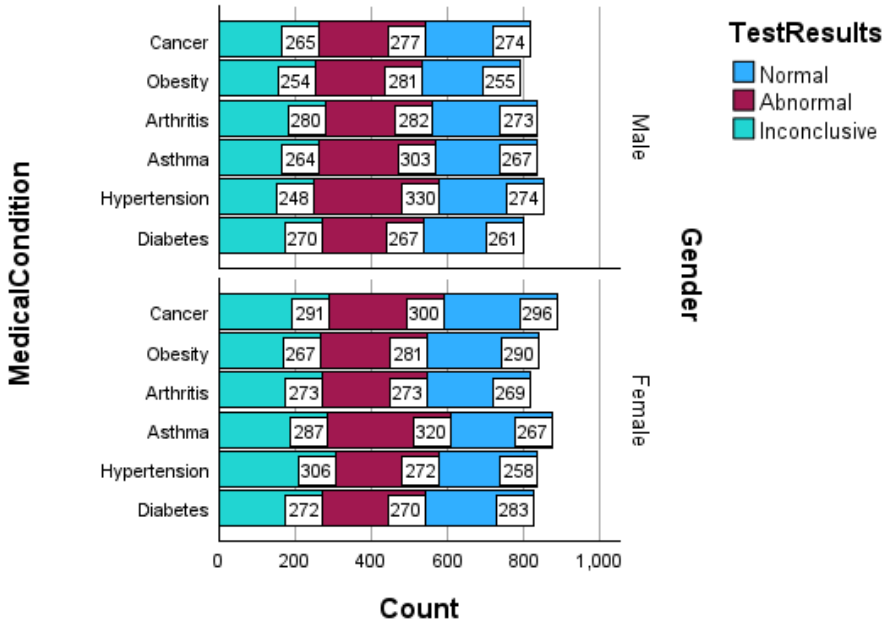
As the p-value of Pearson correlation is 0.344 which is more than significant value 0.05 which means we **fail to reject null hypothesis**. Therefore, there is no significant relation in billing amount and age even if we control the medical condition of patient.

12) Time duration taken by any medication to cure any medical condition.



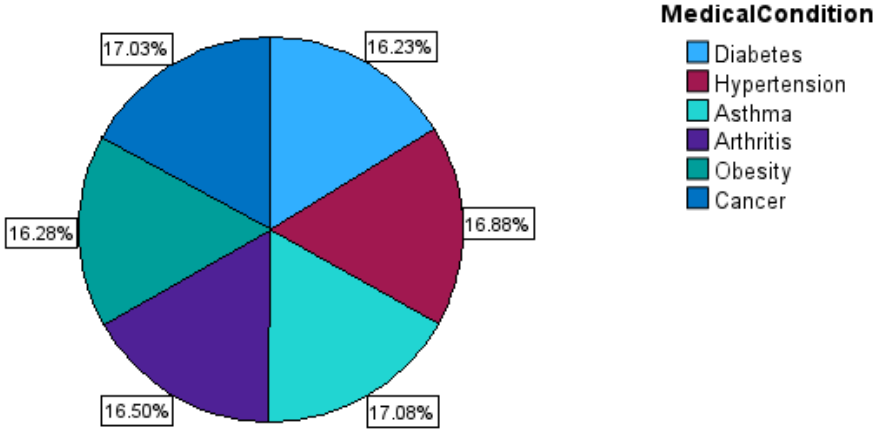
Time duration taken by Penicillin and Ibuprofen medicine to cure disease is more than any other medicine.

13) Number of patients in each medical condition along with test results gender wise.



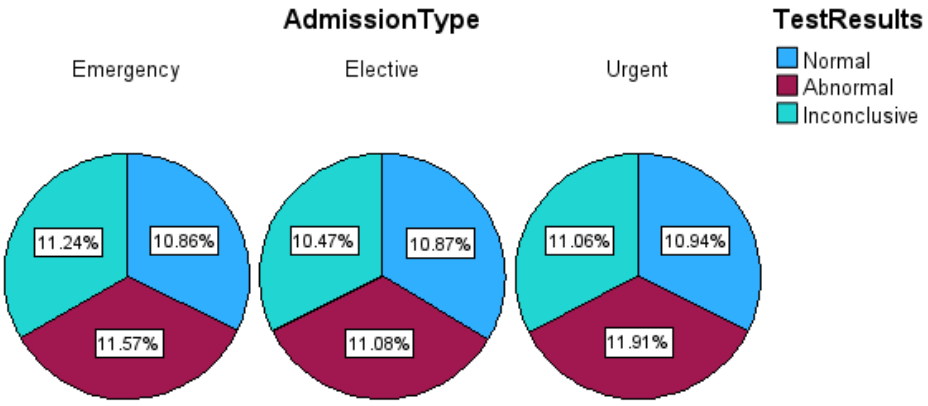
There are a greater number of abnormal test results has been seen in male for different medical conditions whereas in females, for Cancer, Arthritis and Asthma more test results are abnormal.

14) Percentage of patients with different medical conditions.



Maximum number of patients has been accounted for Asthma disease with 17.08% and Diabetes patients are second highest number of patients with percentage of 17.03%.

15) Test result and admission type



As shown in this pie chart, abnormal results are more in percentage for all types of admissions in hospital and normal results are least in number.

https://www.kaggle.com/code/hainescity/healthcare-dataset-eda/notebook

Rajvinder Kaur