Healthcare dataset
Dummy data with Multi Category Classification Problem

Extracted from: https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data

---

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 :

1. **Name:** This column represents the name of the patient associated with the healthcare record.
2. **Age:** The age of the patient at the time of admission, expressed in years.
3. **Gender:** Indicates the gender of the patient, either "Male" or "Female."
4. **Blood Type:** The patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).
5. **Medical Condition:** This column specifies the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.
6. **Date of Admission:** The date on which the patient was admitted to the healthcare facility.
7. **Doctor:** The name of the doctor responsible for the patient's care during their admission.
8. **Hospital:** Identifies the healthcare facility or hospital where the patient was admitted.
9. **Insurance Provider:** This column indicates the patient's insurance provider, which can be one of several options, including "Aetna," "Blue Cross," "Cigna," "UnitedHealthcare," and "Medicare."
10. **Billing Amount:** The amount of money billed for the patient's healthcare services during their admission. This is expressed as a floating-point number.
11. **Room Number:** The room number where the patient was accommodated during their admission.
12. **Admission Type:** Specifies the type of admission, which can be "Emergency," "Elective," or "Urgent," reflecting the circumstances of the admission.
13. **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.
14. **Medication:** Identifies a medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."

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

Potential correlation candidates:

Correlation can only be calculated for numerical data

1. Age → Condition; Age →Billing amount
2. Medication → Condition; Medication → Billing amount; Medication → Test Results
3. Condition → Billing amount; Condition → Days Admitted, Condition → Gender, Blood type → Condition
4. Test Results → Billing amount; (Does more money really amount to successful treatment?)

Sorting candidate column: AGE or DAYS ADMITTED

Note: Date table: Annually, Quarterly, Monthly, Weekly

General stats:

- Insurance provided by an insurance company according to date table
- Cases of each Condition according to date table
- Increase in medical conditions with age
- Average billing amount (Different b/w median and mean)
- Total amount spent in healthcare sector
- Average Days Admitted according to the condition
- Number of Cases according to the blood group

Data Cleaning Steps

1. Convert csv to xlsx

Initial formatting:

2. heading colours,
3. align centred the entries
4. fixes in the name column
    - Convert names to proper lowercase
    - Removing honorifics (Dr Mr Mrs)

- Removing dot special character from name
- Convert the name into proper name with first letter uppercase

5. Creating a S.no column,
6. Convert billing amount to number, shorten to 2 decimal places
7. Basic fixes in the table structure
8. Remove special characters in hospital name
9. Create a Days Admitted column
10. Sort the data according to age
11. Convert Male to M and Female to F

Data Processing

I. Age Analysis worksheet

Create an extra Age to Condition worksheet with the following data outcomes

columns:

- name
- age
- condition
- Billing amount

1. Extract a list of unique medical conditions from the CONDITION column.
2. Count the total number of people in one condition.
3. Make a pivot table with Age in rows and CONDITION in columns, Put Count of Age in values show as percentage.
4. Group Age according to class interval size 10.
5. Correct the formatting and colours of the pivot table.
6. Add the conditional formatting in the pivot table as heat map.
7. More details to be added in the form of charts.

Insights Found in this sheet:

By the below data we can clearly deduce that in this dataset:

1. most of the people having a Medical Condition fall in the ranges from 35 to 67 and furthermore most of the people in this range are suffering from Obesity and Diabetes.
2. According to the heatmap, second range of people lie in the 68-78 range for Arthritis and in the 57 – 67 range for Hypertension.
3. We can also derive that young aged people do have the least number of medical conditions.
4. People in the range of age 79-89 are not suffering in less numbers, but the number of these people is small in this dataset to begin with.
5. From the perspective of relation between Age and Medical conditions this data closely mimics real world population as observed.
6. When trying to find a correlation between Age and Billing amount the score is -0.003832 which is almost no correlation, which clearly suggests that in this data Age does not affect Billing amount in any way.

Total cases	Number
Arthritis	9308
Asthma	9185
Cancer	9227
Diabetes	9304
Hypertension	9245
Obesity	9231
Total	**55500**

Count of Age	Conditions						
Age Intervals	Arthritis	Asthma	Cancer	Diabetes	Hypertension	Obesity	Grand Total
13-23	8.95%	8.60%	8.80%	8.61%	8.76%	8.95%	8.78%
24-34	15.65%	16.41%	16.06%	15.63%	15.79%	16.13%	15.94%
35-45	16.43%	16.06%	16.09%	16.51%	16.25%	16.04%	16.23%
46-56	16.49%	16.11%	16.20%	16.29%	16.02%	16.99%	16.35%
57-67	15.47%	16.02%	16.41%	17.20%	16.54%	16.53%	16.36%
68-78	16.71%	16.28%	15.80%	15.78%	16.17%	15.27%	16.00%
79-89	10.30%	10.53%	10.63%	9.98%	10.47%	10.09%	10.33%
Grand Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Correlation Analysis between Age and Billing amount:

	Age	*Billing amount*
Age	1	
Billing Amount	-0.003832	1

II. Medication/Medicine Analysis worksheet

Columns:

- Name
- Condition
- Medication
- Billing amount

From the below pivot table we can deduce the following insights about the medications:

For Arthritis → most prescribed: Aspirin ; Least prescribed: Ibuprofen

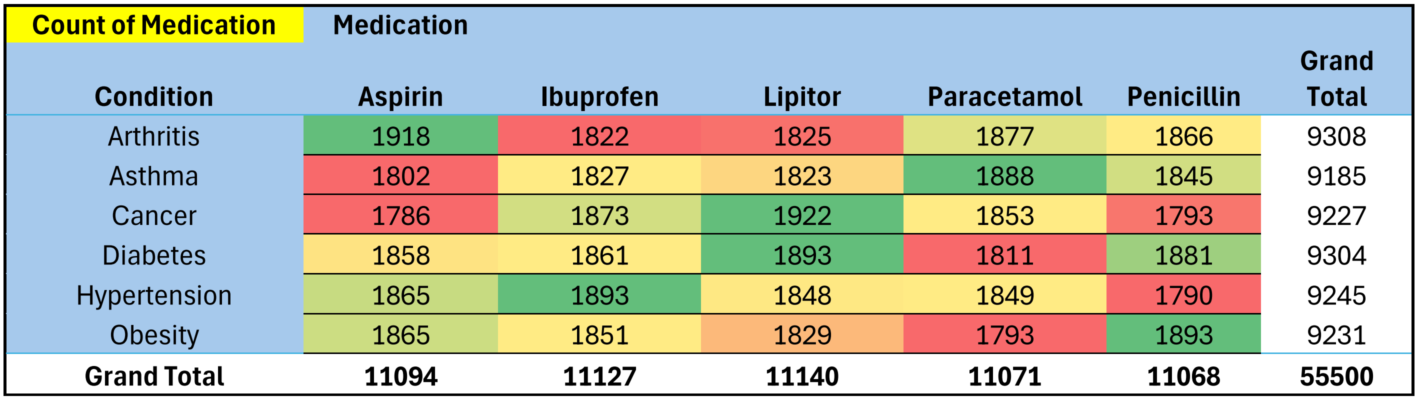
For Asthma →most prescribed: Paracetamol ; Least prescribed: Aspirin

For Cancer →most prescribed: Lipitor ; Least prescribed: Aspirin

For Diabetes →most prescribed: Lipitor ; Least prescribed: Paracetamol

For Hypertension→most prescribed: Ibuprofen ; Least prescribed: Penicillin

For Obesity →most prescribed: Penicillin ; Least prescribed: Paracetamol



Count of Medication	Medication					
Condition	Aspirin	Ibuprofen	Lipitor	Paracetamol	Penicillin	Grand Total
Arthritis	1918	1822	1825	1877	1866	9308
Asthma	1802	1827	1823	1888	1845	9185
Cancer	1786	1873	1922	1853	1793	9227
Diabetes	1858	1861	1893	1811	1881	9304
Hypertension	1865	1893	1848	1849	1790	9245
Obesity	1865	1851	1829	1793	1893	9231
Grand Total	**11094**	**11127**	**11140**	**11071**	**11068**	**55500**

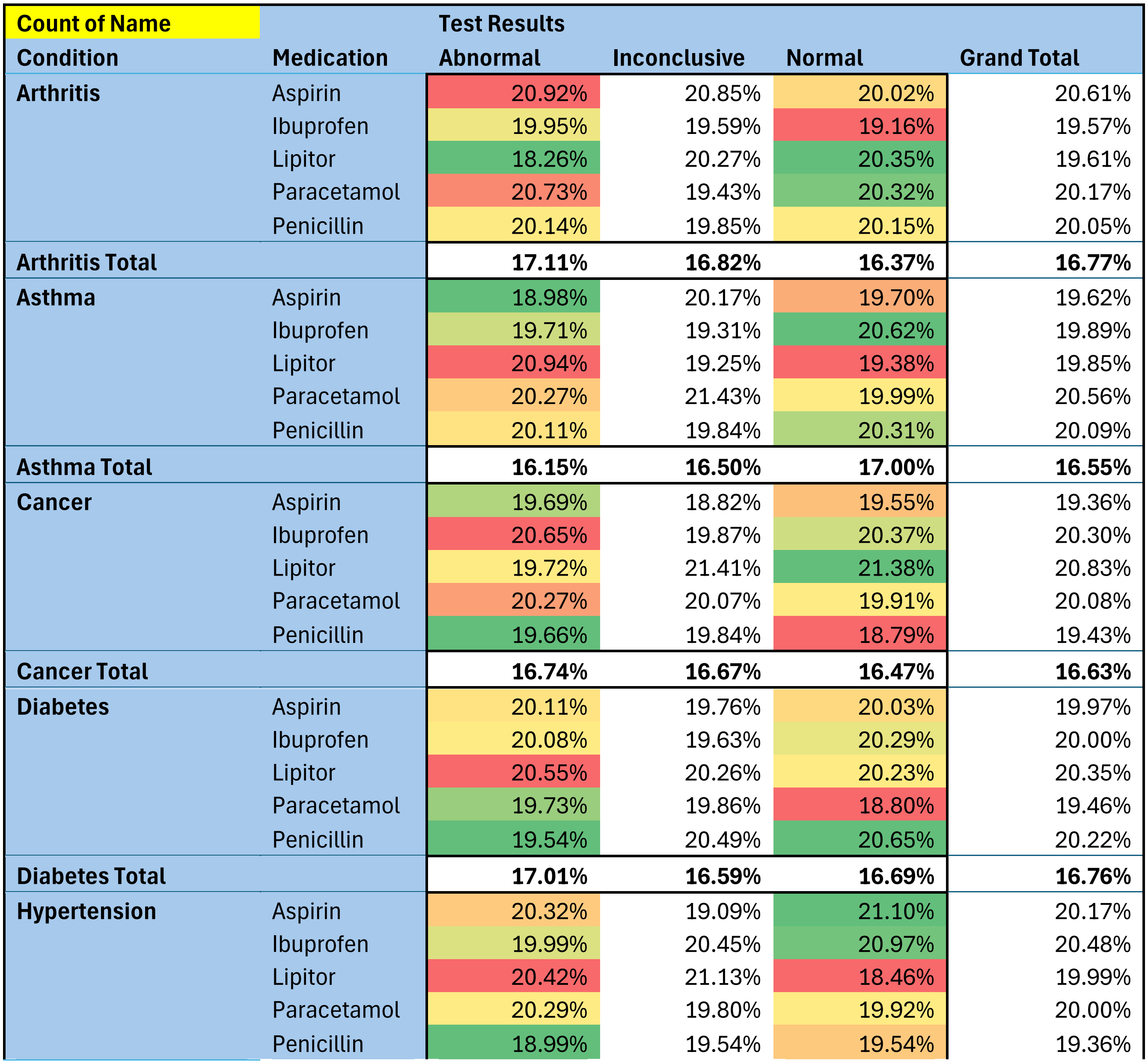
From the pivot table given below we have a lot of interesting findings :

We can not derive any meaning from the results that are Inconclusive. We need to highlight the medications which are the least in ABNORMAL and the most in NORMAL.
In this sense, a medical condition sequentially gives us certain findings .

For

- Arthritis ; Lipitor is the best medicine according to test results
- Asthma; Ibuprofen
- Cancer ; Lipitor
- Diabetes; Penicillin
- Hypertension; Ibuprofen
- Obesity ; Aspirin

But above data has a lot of conflicts that suggests that a lot of doctors also prescribe medications that are not optimized for the best test results. Only for cancer and hypertension, have the doctors prescribed Lipitor and Ibuprofen respectively which is also fits for best test results.



Count of Name		Test Results			
Condition	**Medication**	**Abnormal**	**Inconclusive**	**Normal**	**Grand Total**
Arthritis	Aspirin	20.92%	20.85%	20.02%	20.61%
	Ibuprofen	19.95%	19.59%	19.16%	19.57%
	Lipitor	18.26%	20.27%	20.35%	19.61%
	Paracetamol	20.73%	19.43%	20.32%	20.17%
	Penicillin	20.14%	19.85%	20.15%	20.05%
Arthritis Total		**17.11%**	**16.82%**	**16.37%**	**16.77%**
Asthma	Aspirin	18.98%	20.17%	19.70%	19.62%
	Ibuprofen	19.71%	19.31%	20.62%	19.89%
	Lipitor	20.94%	19.25%	19.38%	19.85%
	Paracetamol	20.27%	21.43%	19.99%	20.56%
	Penicillin	20.11%	19.84%	20.31%	20.09%
Asthma Total		**16.15%**	**16.50%**	**17.00%**	**16.55%**
Cancer	Aspirin	19.69%	18.82%	19.55%	19.36%
	Ibuprofen	20.65%	19.87%	20.37%	20.30%
	Lipitor	19.72%	21.41%	21.38%	20.83%
	Paracetamol	20.27%	20.07%	19.91%	20.08%
	Penicillin	19.66%	19.84%	18.79%	19.43%
Cancer Total		**16.74%**	**16.67%**	**16.47%**	**16.63%**
Diabetes	Aspirin	20.11%	19.76%	20.03%	19.97%
	Ibuprofen	20.08%	19.63%	20.29%	20.00%
	Lipitor	20.55%	20.26%	20.23%	20.35%
	Paracetamol	19.73%	19.86%	18.80%	19.46%
	Penicillin	19.54%	20.49%	20.65%	20.22%
Diabetes Total		**17.01%**	**16.59%**	**16.69%**	**16.76%**
Hypertension	Aspirin	20.32%	19.09%	21.10%	20.17%
	Ibuprofen	19.99%	20.45%	20.97%	20.48%
	Lipitor	20.42%	21.13%	18.46%	19.99%
	Paracetamol	20.29%	19.80%	19.92%	20.00%
	Penicillin	18.99%	19.54%	19.54%	19.36%

Hypertension Total		**16.17%**	**16.84%**	**16.97%**	**16.66%**
Obesity	Aspirin	20.15%	19.76%	20.71%	20.20%
	Ibuprofen	20.31%	20.28%	19.56%	20.05%
	Lipitor	19.60%	20.12%	19.73%	19.81%
	Paracetamol	19.09%	19.76%	19.43%	19.42%
	Penicillin	20.85%	20.09%	20.58%	20.51%
Obesity Total		**16.81%**	**16.57%**	**16.51%**	**16.63%**
Grand Total		**100.00%**	**100.00%**	**100.00%**	**100.00%**

From the table below we can infer that all the medications have almost the same amount of average billing amount hence one medication is not more or less expensive than another.

Ibuprofen costs slightly more than other medications.

Row Labels	**Average of Billing Amount**	
Aspirin	$	25,594.26
Ibuprofen	$	25,735.58
Lipitor	$	25,342.47
Paracetamol	$	25,533.47
Penicillin	$	25,490.92
Grand Total		**25539.3161**

III. Medical Condition Analysis worksheet

Columns:

- Name
- Days Admitted
- Condition
- Billing amount
- Gender
- Blood Type

Row Labels	**Avg of Days Admitted**	**Average of Billing Amount**	
Arthritis	16	$	25,497.33
Asthma	16	$	25,635.25
Cancer	15	$	25,161.79
Diabetes	15	$	25,638.41
Hypertension	15	$	25,497.10
Obesity	15	$	25,805.97
Grand Total	**16**	**$**	**25,539.32**

This table shows less about the population insights and more about this particular dataset because it shows that the digits are more or less evenly spread when it comes to relation with the medical condition, not necessarily showing true facts.

Given below is the pivot table that aims to display the distribution of medical condition in Male vs Female:

Count of Name	Gender		
Condition	F	M	Grand Total
Arthritis	4686	4622	9308
Asthma	4553	4632	9185
Cancer	4602	4625	9227
Diabetes	4651	4653	9304
Hypertension	4612	4633	9245
Obesity	4622	4609	9231
Grand Total	**27726**	**27774**	**55500**

On general inspection of this data, we can derive that similar to the previous insight that we found this data is also more or less distributed equally among the attributes and gender does not affect the probability of having a certain medical condition.

However, we can look at the maximum and minimum values :

- Females having arthritis is the most common in the dataset .
- Females having Asthma is the least common.

Below is the pivot table which attempts to relate blood type with condition.

Non-interestingly enough, medical conditions are also evenly distributed across all the types of blood.

- A+ having Diabetes is the most common.
- B+ having Hypertension is the least common.

Count of Name	Blood Type								
Condition	A-	A+	AB-	AB+	B-	B+	O-	O+	Grand Total
Arthritis	1153	1116	1192	1130	1169	1201	1149	1198	9308
Asthma	1173	1135	1134	1189	1119	1108	1154	1173	9185
Cancer	1134	1185	1198	1112	1144	1196	1150	1108	9227
Diabetes	1167	1213	1139	1173	1151	1188	1122	1151	9304
Hypertension	1199	1128	1125	1215	1173	1103	1145	1157	9245
Obesity	1143	1179	1157	1128	1188	1149	1157	1130	9231
Grand Total	**6969**	**6956**	**6945**	**6947**	**6944**	**6945**	**6877**	**6917**	**55500**

IV. Bill Analysis worksheet

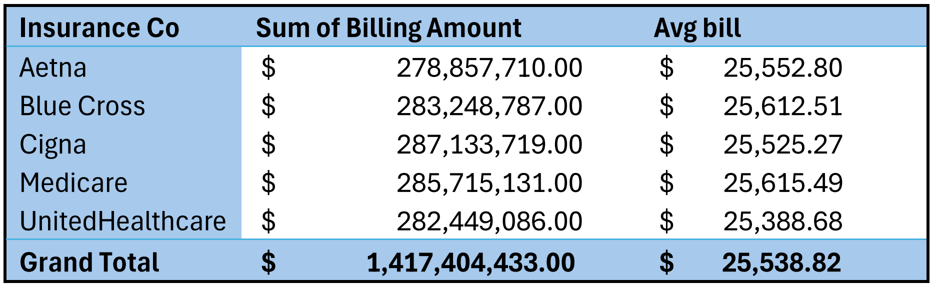
Columns:

- Name
- Test Results
- Billing amount
- Insurance provider

Bill can be analysed in the following key ways:

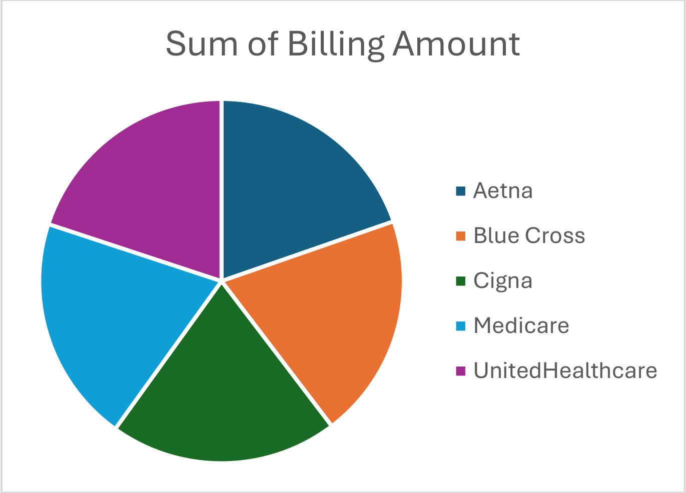
- Which insurance company gave what amount of insurance in the form of a pivot table.

  The pivot table of insurance provider and the sum of billing amount displays that all the insurance companies paid almost the same amount of insurance and have similar bill average which means there are very less outliers in the data.

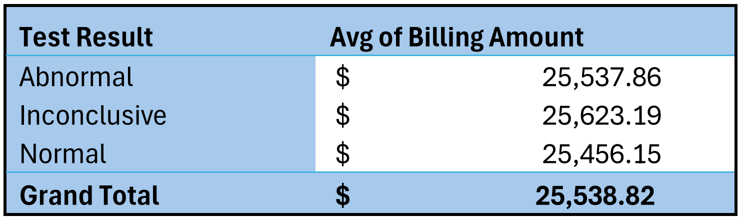


Insurance Co	Sum of Billing Amount	Avg bill
Aetna	$ 278,857,710.00	$ 25,552.80
Blue Cross	$ 283,248,787.00	$ 25,612.51
Cigna	$ 287,133,719.00	$ 25,525.27
Medicare	$ 285,715,131.00	$ 25,615.49
UnitedHealthcare	$ 282,449,086.00	$ 25,388.68
Grand Total	**$ 1,417,404,433.00**	**$ 25,538.82**

- Pie chart that can display which company gave how much of insurance as a fraction of the whole amount. The pie chart clearly shows that out of the total amount which is 1.4B $ all 5 of the companies have 1/5 share each of the amount of insurance paid.



- A pivot table of Test result and Average of billing amount which can show if amount of money spend can increase your chances of a normal test result. The last pivot table shows that Test results can't be necessarily related with the billing amount because almost equal amount of money was paid which resulted in three different test results.



Test Result	Avg of Billing Amount
Abnormal	$ 25,537.86
Inconclusive	$ 25,623.19
Normal	$ 25,456.15
Grand Total	**$ 25,538.82**

V. Conclusion

By conducting the above study of different attributes, we can conclude that the data was equally distributed amongst all the attributes and no solid conclusions can be drawn because the data is not skewed towards any particular outcome favouring or unfavouring the results. This study told us the methods by which we can conduct studies on similar datasets which need to be cleaned, processed and ultimately explored to find interesting insights about the population behaviour in the data represented in the form of a sample.

This framework can be subsequently used to create a dashboard representing different stats on one single page.