**1. Importing required libraries and dataset**

Table of Hospitalisation details (first 5 rows):

Customer ID year month date children charges Hospital tier City tier \

0 Id2335 1992 Jul 9 0 563.84 tier - 2 tier - 3

1 Id2334 1992 Nov 30 0 570.62 tier - 2 tier - 1

2 Id2333 1993 Jun 30 0 600.00 tier - 2 tier - 1

3 Id2332 1992 Sep 13 0 604.54 tier - 3 tier - 3

4 Id2331 1998 Jul 27 0 637.26 tier - 3 tier - 3

State ID

0 R1013

1 R1013

2 R1013

3 R1013

4 R1013

Table of Medical Examinations:

Customer ID BMI HBA1C Heart Issues Any Transplants Cancer history \

0 Id1 47.410 7.47 No No No

1 Id2 30.360 5.77 No No No

2 Id3 34.485 11.87 yes No No

3 Id4 38.095 6.05 No No No

4 Id5 35.530 5.45 No No No

NumberOfMajorSurgeries smoker

0 No major surgery yes

1 No major surgery yes

2 2 yes

3 No major surgery yes

4 No major surgery yes

Table of Names:

Customer ID name

0 Id1 Hawks, Ms. Kelly

1 Id2 Lehner, Mr. Matthew D

2 Id3 Lu, Mr. Phil

3 Id4 Osborne, Ms. Kelsey

4 Id5 Kadala, Ms. Kristyn

# 2. Check for missing values

Number of missing values (NaN) for each column in Hospitalisation details:

Customer ID 0

year 0

month 0

date 0

children 0

charges 0

Hospital tier 0

City tier 0

State ID 0

dtype: int64

Number of missing values (NaN) for each column in Medical Examinations:

Customer ID 0

BMI 0

HBA1C 0

Heart Issues 0

Any Transplants 0

Cancer history 0

NumberOfMajorSurgeries 0

smoker 0

dtype: int64

Number of missing values (NaN) for each column in Names:

Customer ID 0

name 0

dtype: int64

# 3. Find percentage of rows containing trivial value (?) and drop them if possible

Percentage of rows with trivial values (?) in table of Hospitalisation Details:

0.597524541186513

Percentage of rows with trivial values (?) in table of Medical Examinations:

0.08565310492505353

Percentage of rows with trivial values (?) in table of Names:

0.0

# 4. Encode ordinal and categorical variables into numeric representations to enable further process and machine learning model processing

# 5. Remove irrelevant or low-impact columns to improve model performance and reduce noise

# 6. Cleaning NumberOfMajorSurgeries variable

# 7. Create a new column for age of patients and calcuate it from patient\_age

First five rows of data after adding column for patinet age

Customer ID year month date children charges Hospital tier City tier \

0 Id2335 1992 Jul 9 0 563.84 1.0 2.0

1 Id2334 1992 Nov 30 0 570.62 1.0 0.0

2 Id2333 1993 Jun 30 0 600.00 1.0 0.0

3 Id2332 1992 Sep 13 0 604.54 2.0 2.0

4 Id2331 1998 Jul 27 0 637.26 2.0 2.0

State ID\_R1011 State ID\_R1012 State ID\_R1013 patient age

0 0.0 0.0 1.0 33

1 0.0 0.0 1.0 33

2 0.0 0.0 1.0 32

3 0.0 0.0 1.0 33

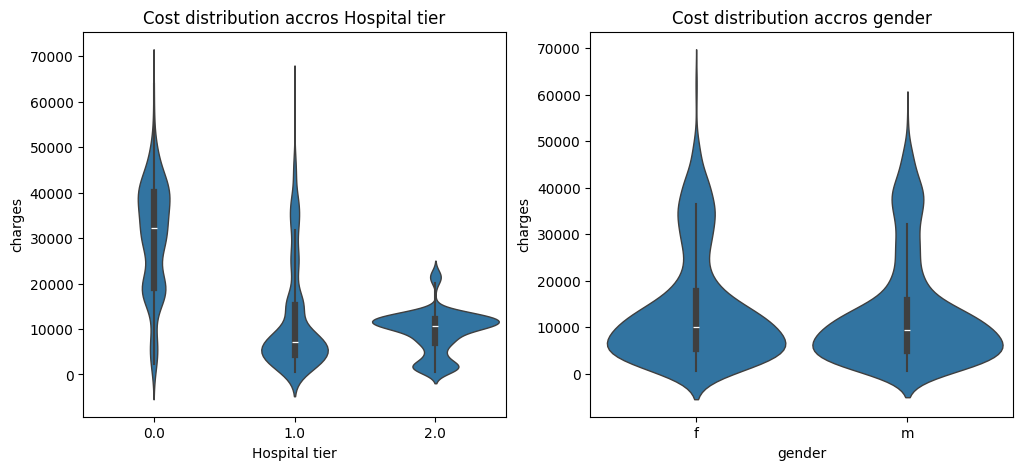
4 0.0 0.0 1.0 27

# 8. Create a new column for gender of patients

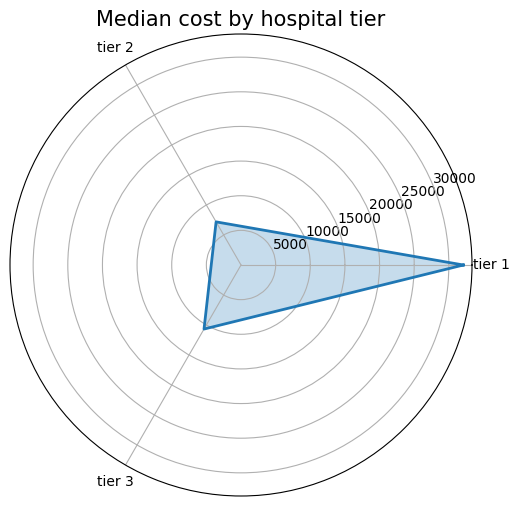
# 9. histogram, box plot and swarm plot of cost

# 10. Cost Distribution of charges across hospital tier and gender

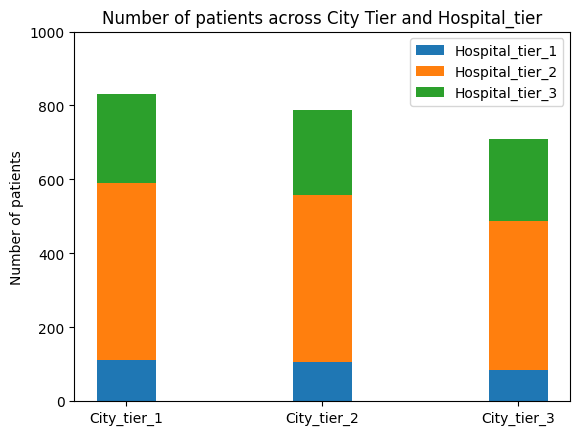
Text(0.5, 1.0, 'Cost distribution accros gender')



# 11. Median cost by hospital tier



# 12. Number of patients across city tier and hospital tier

****