Final Project Introduction to D.S

Group 11's members:

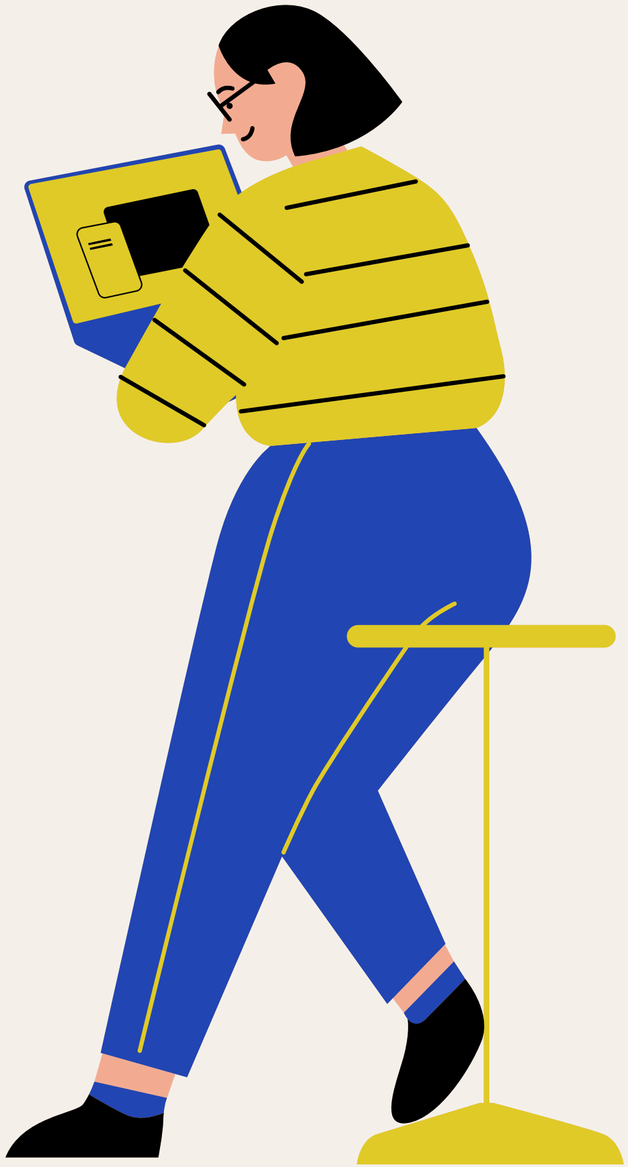
21127038 - Võ Phú Hãn
21127667 - Trương Công Gia Phát
21127743 - Trần Thái Toàn

01 - Project Planning

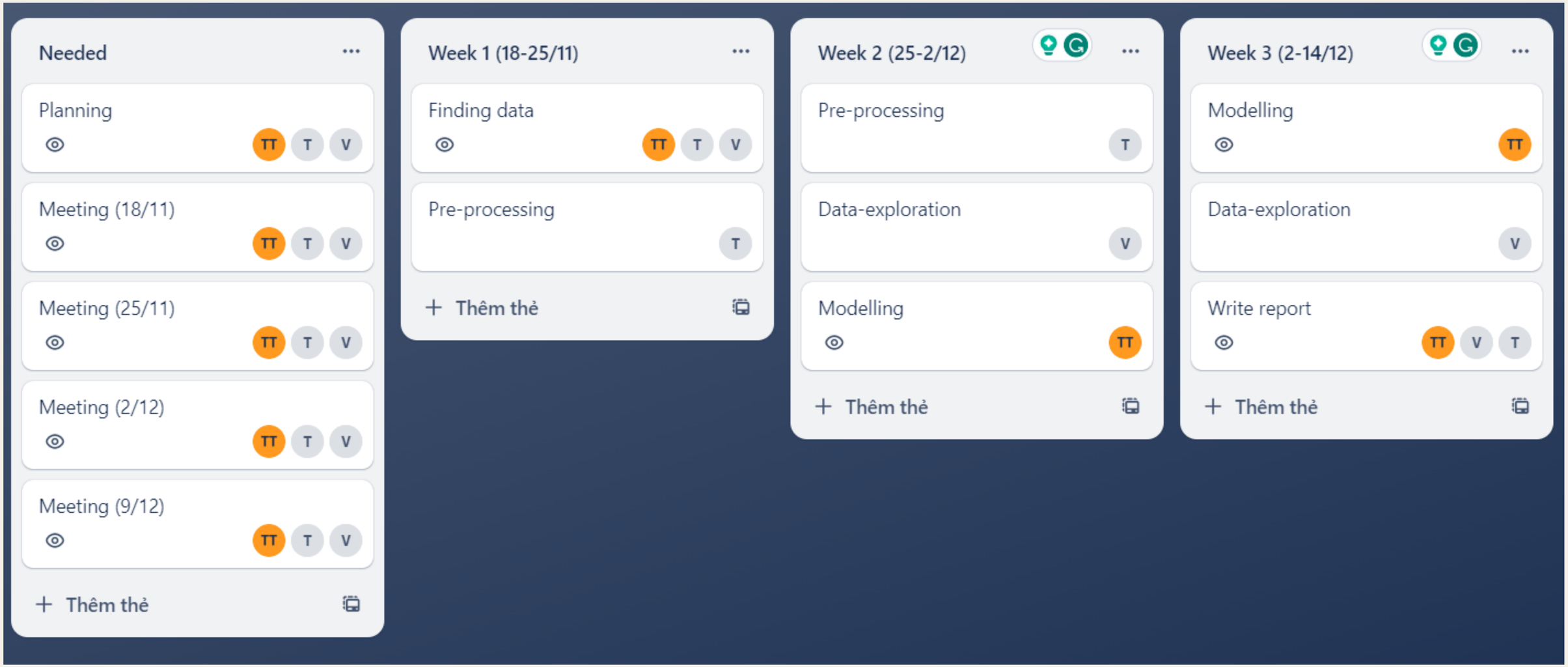
02 - Data Collection & Data Preprocessing

03 - Data Exploration

04 - Data Modelling



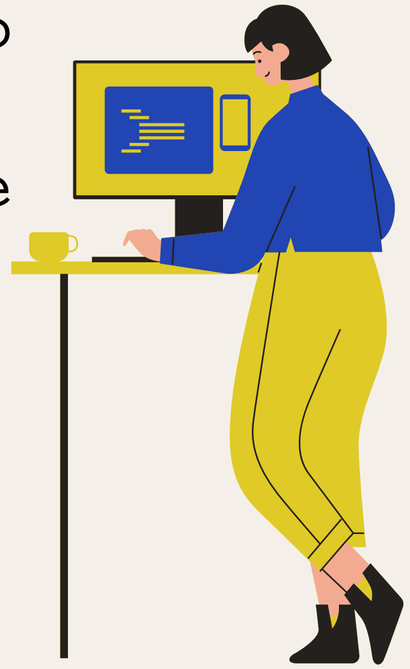
01 - Project Planning



02 - Data Collection & Data Preprocessing

https://www.communitybenefitinsight.org/?page=info.data_api

Our group chose the website followed by the link above to collect data.
This website contains information about hospitals across the United States.



2.1 - Data Collection

The Community Benefit Insight data API allows us to retrieve the following types of data:
+ Hospital data (optionally filtered by state)
+ Detailed data about a single hospital
We can get the general data for every hospital followed the link:

https://www.communitybenefitinsight.org/?page=info.data_api

If we want to retrieve detailed data for every hospital above, we use:

https://www.chttps://www.communitybenefitinsight.org/api/get_hospital_data.php?hospital_id=ID (ID is the hospital ID)

There are total 3491 hospitals but the website only allows us to make 100 requests per week for detailed information so we have to change location by changing our VPN to retrieve the latest data for every hospital.

2.2 - Data Preprocessing

The raw data we got after collecting has 3491 rows and 161 columns so our group tried to reduce the number of columns down to 30-40.

After looking into the data, we can find some problems with it:

- The dataset has a lot of columns that most of it are 0 or NaN. --> We remove column if 20% of it is 0 or NaN values..
- There are some duplicated and irrelevant columns --> We will remove all of the duplicated and irrelevant columns.
- There are a lot of flag columns.--> If the column has over 80% Yes or 80% No we will remove it.

We chose threshold 20% so we can remove as many columns as possible.

03 - Data Exploration

Basic level:

How many rows and columns are there?

--> After preprocessing, the dataset has 3491 rows and 37 columns.

What is the meaning of each column?

--> You can see in details in our Jupyter Notebook, here are 3 examples:

hospital_id:	Internal hospital ID
tot_revenue:	Total revenue
per_capita_inc:	Per capita income

03 - Data Exploration

Basic level:

Does each column have suitable datatype?

--> Yes, you can see in details in Jupyter Notebook.

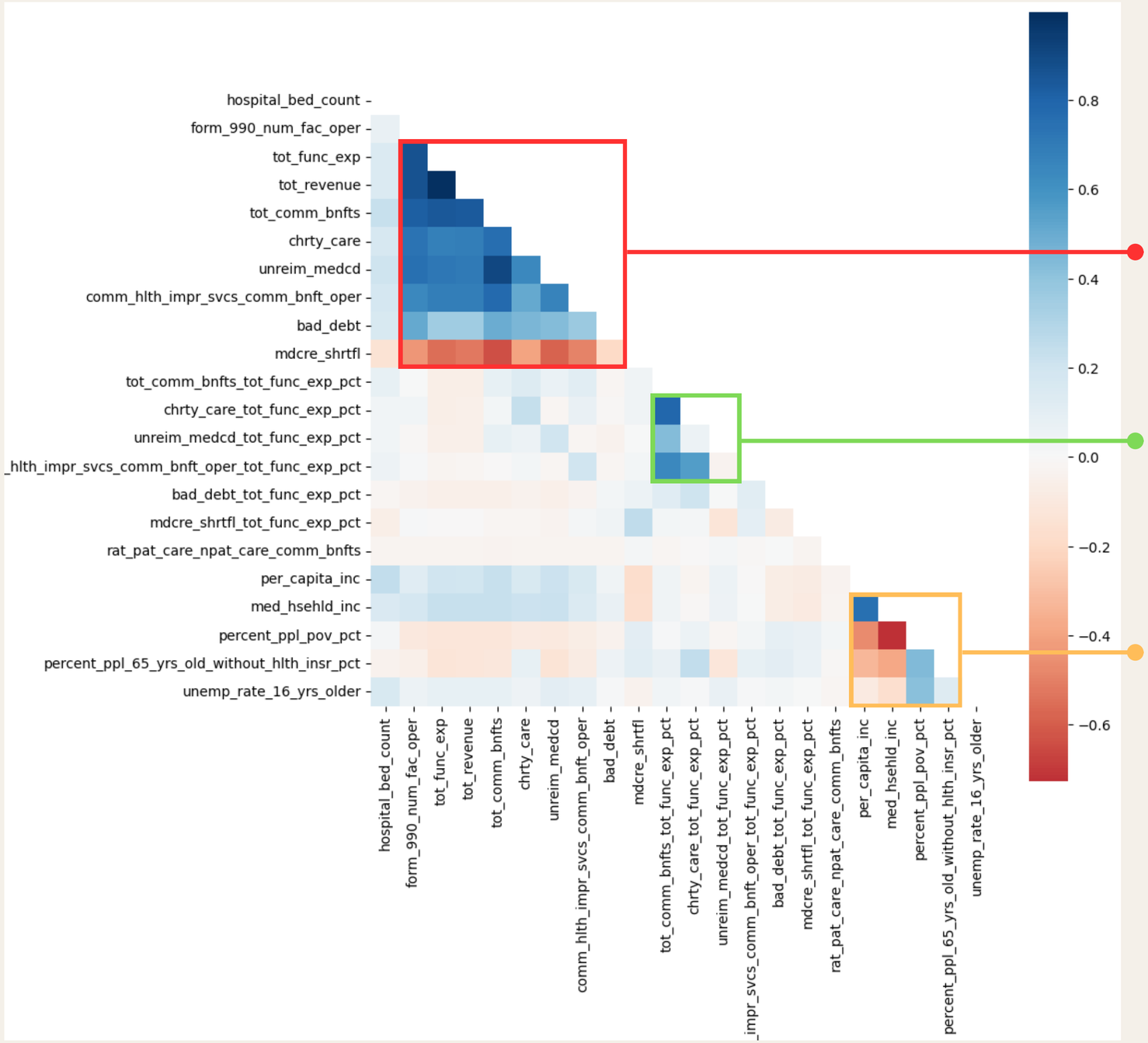
What is the distribution of data in each column?

	ein	hospital_bed_count	medicare_provider_number
missing_ratio	0.0	0.0	0.0
min	10130427.0	2.0	10007.0
lower_quartile	352528741.0	32.0	141333.0
median	476028103.0	114.0	261335.0
upper_quartile	741356589.0	275.0	390074.5
max	990269825.0	3060.0	670309.0

Data distribution in the first three numeric columns

Question 1

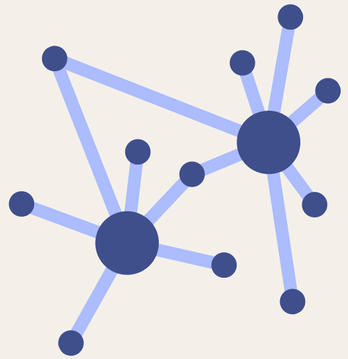
What is the correlation between numeric features in the dataset?

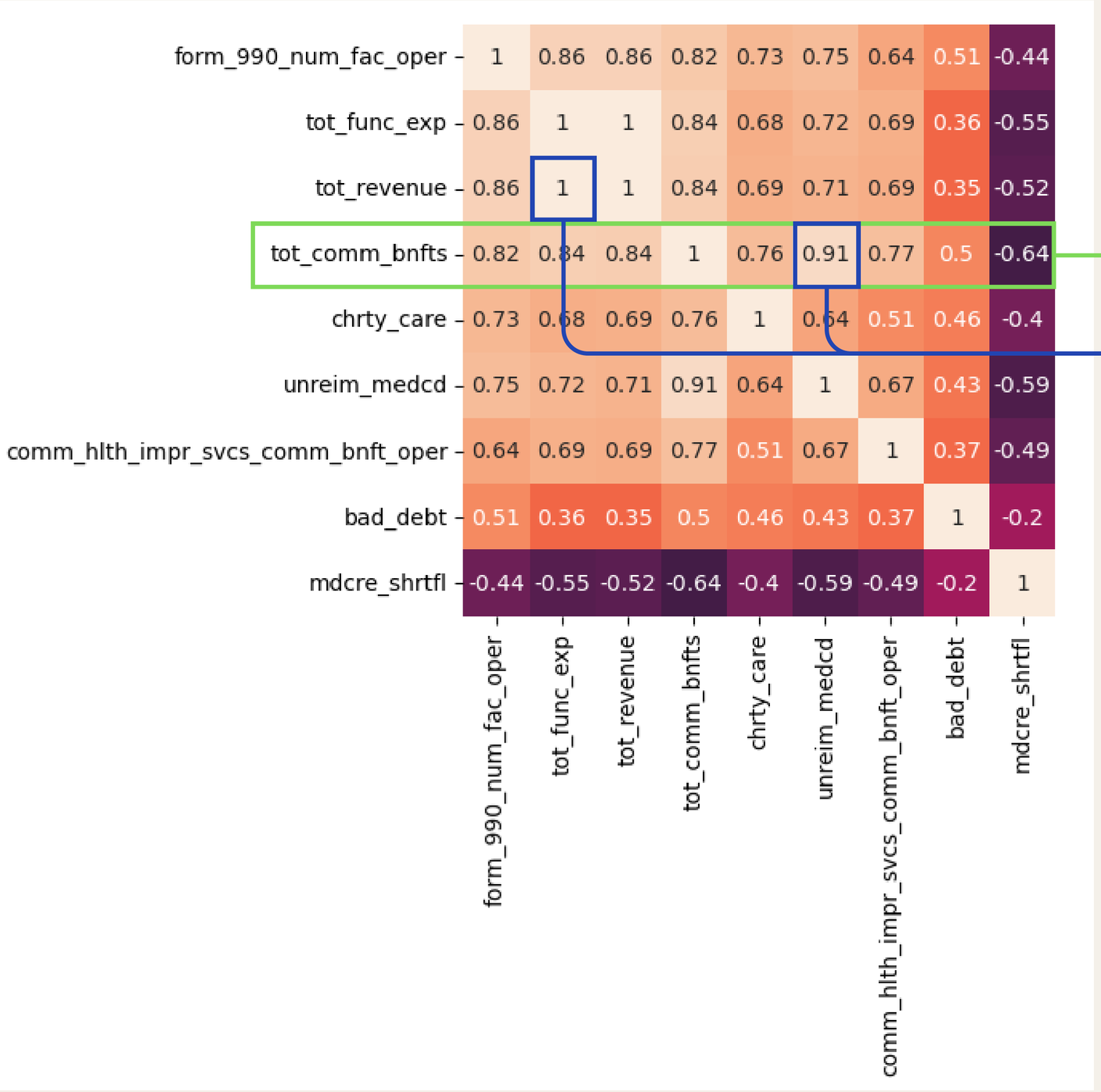


Group 1 represents *financial records* of each hospital

Group 2 represents various *percentages calculated per the total functional expenses*

Group 3 represents *socio-economic indicators*





Question 1

What is the correlation between numeric features in the dataset?

the lightest row

the strongest correlations

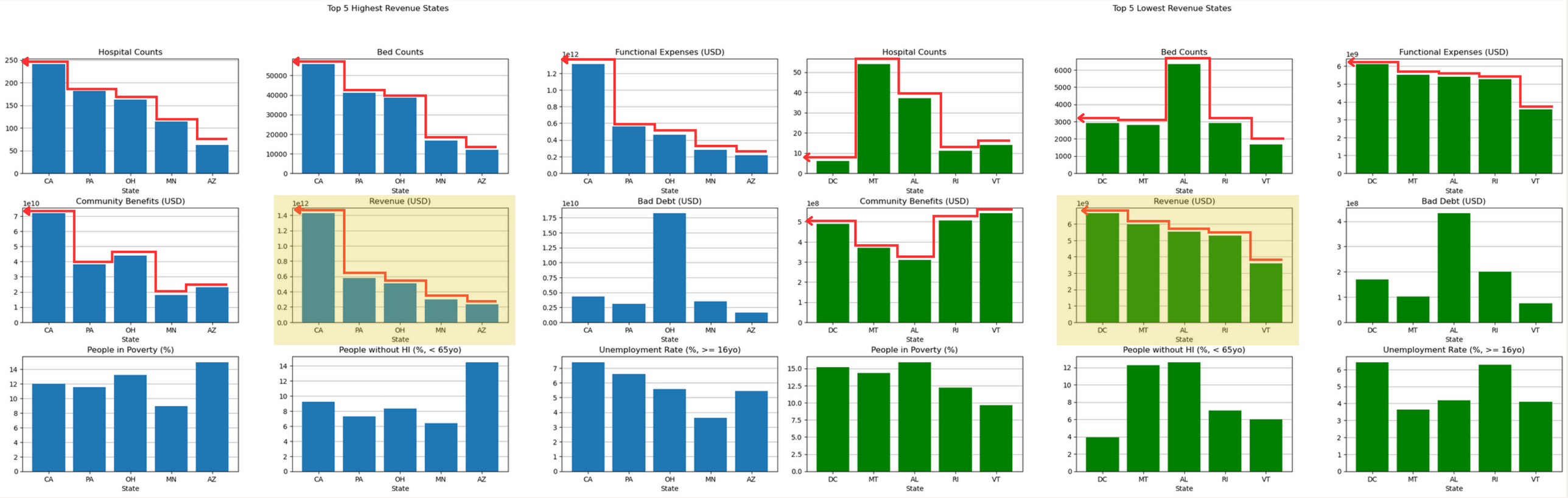
Based on these 2 correlation matrices, we will find out more about the relationships between this group of features:

- `tot_func_exp`,
- `tot_revenue`,
- `tot_comm_bnfts`,
- `chrty_care`,
- `unreim_medcd`,
- `comm_hlth_impr_svcs_comm_bnft_oper`,
- `bad_debt`,
- `mdcre_shrtfl`

and the others.

Question 2

What is the financial situation of hospitals and the quality of life in each US state?

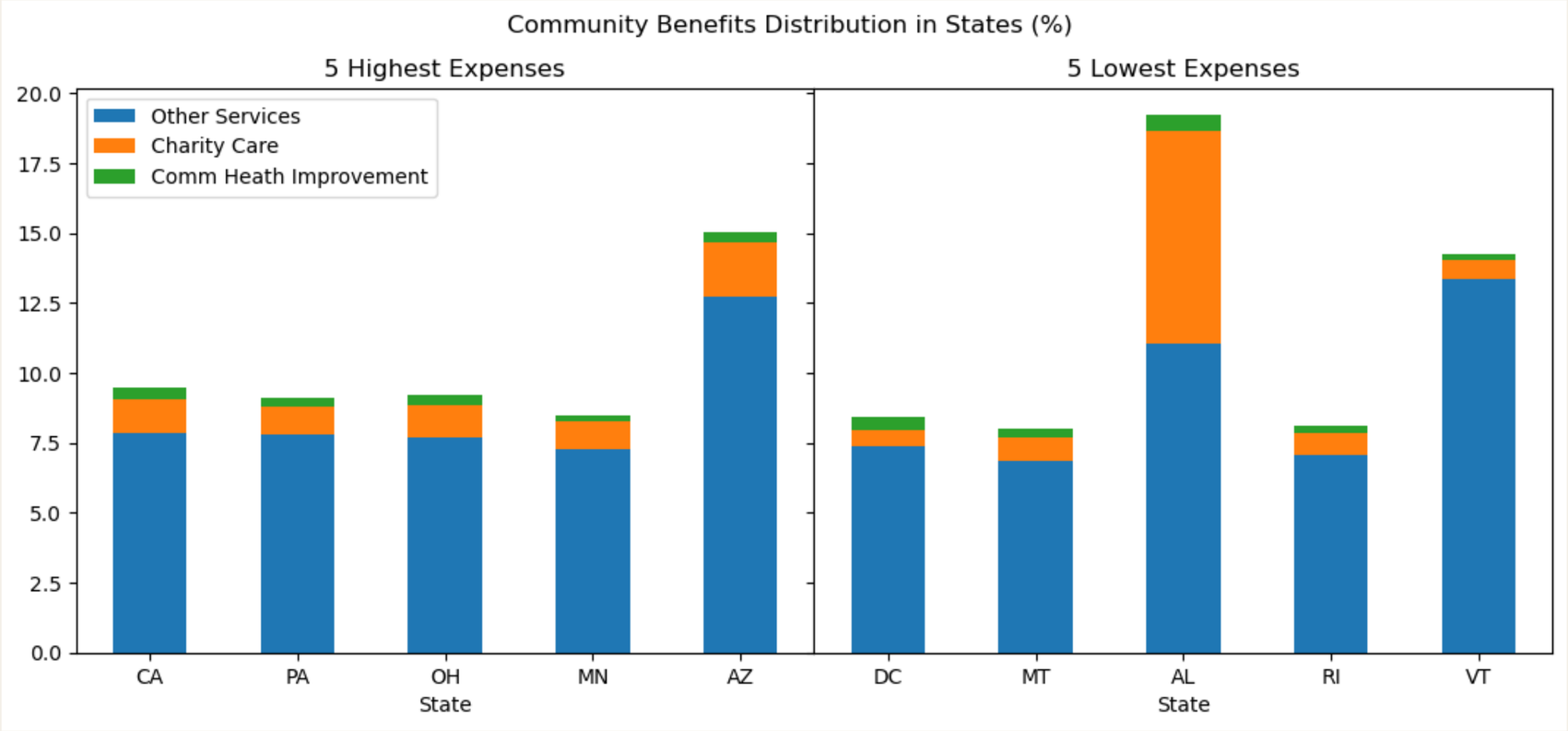


High hospital revenue states invest significantly in facilities, expenses, and community benefits, while **low revenue states** show less clarity in these investments.

Both experience high poverty, lack of health insurance, and unemployment, possibly reflecting the broader USA situation.

Question 3

How do hospitals in each state in USA allocate funds for community benefits?



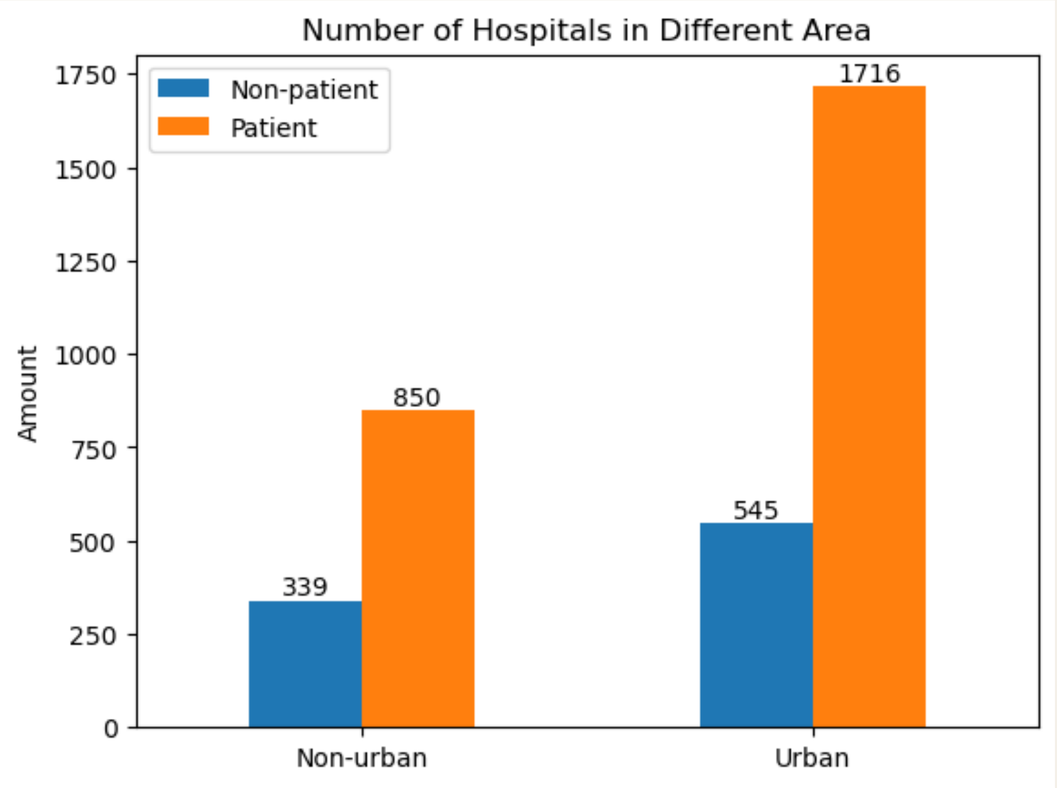
- Most states primarily contribute to their communities through charity care.
- Contribution percentages are various.

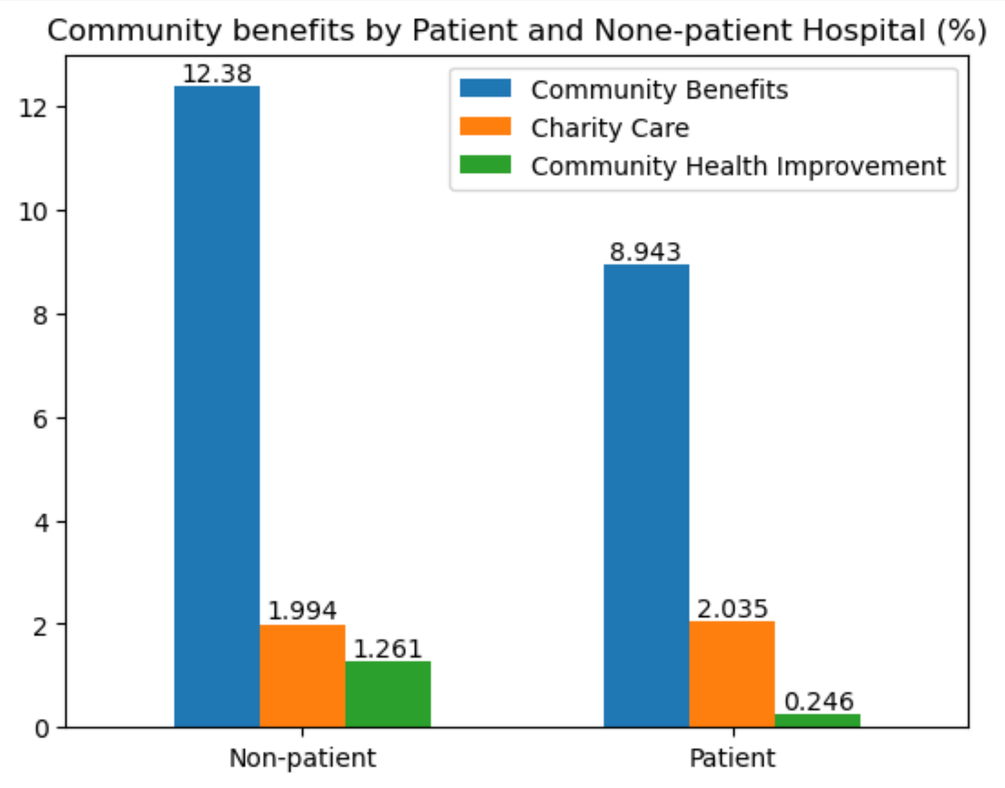
Total functional expenses don't directly correlate with the **percentage of benefits** hospitals contribute to their communities.

Question 4

What are the differences between these two types of hospital (*patient-oriented and community-oriented*) and how they contribute to the community benefits?

$$\text{Ratio of Patient Care to Non-patient Care Community Benefits} = \frac{\text{Total Patient Care Community Benefits}}{\text{Total Non-Patient Care Community Benefits}}$$

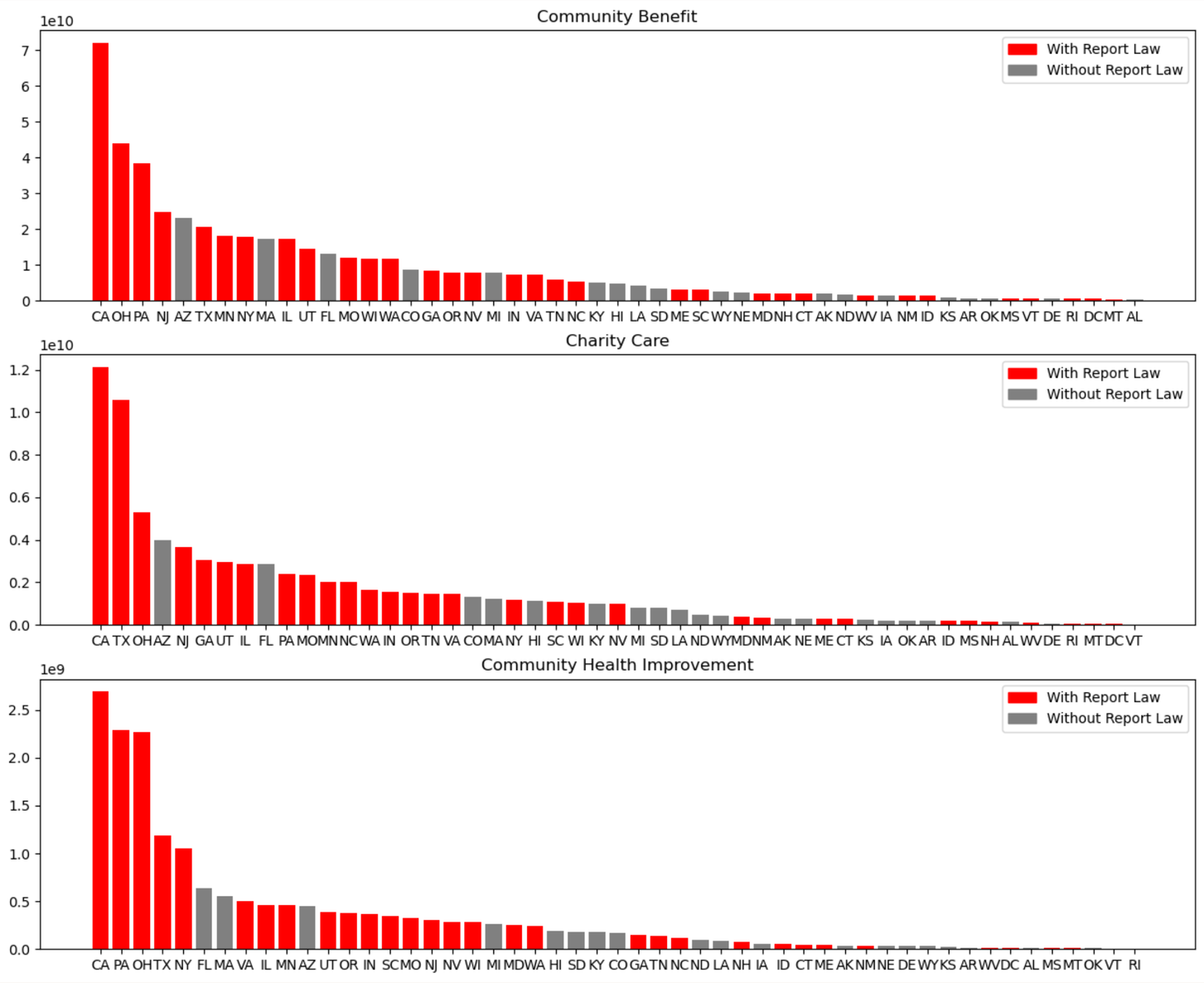




- In **non-urban** area, the community-oriented one take account of nearly 28% of the total number of hospitals, while in **urban** area, this ratio is only 24%.
- The average community benefits percentage of **community hospitals** is higher than **patient hospitals** significantly.

Question 5

What impact do **state laws requiring hospitals to report community benefits** have on the allocation of resources toward community benefit programs?



Most of states require hospital to report community benefits

↓

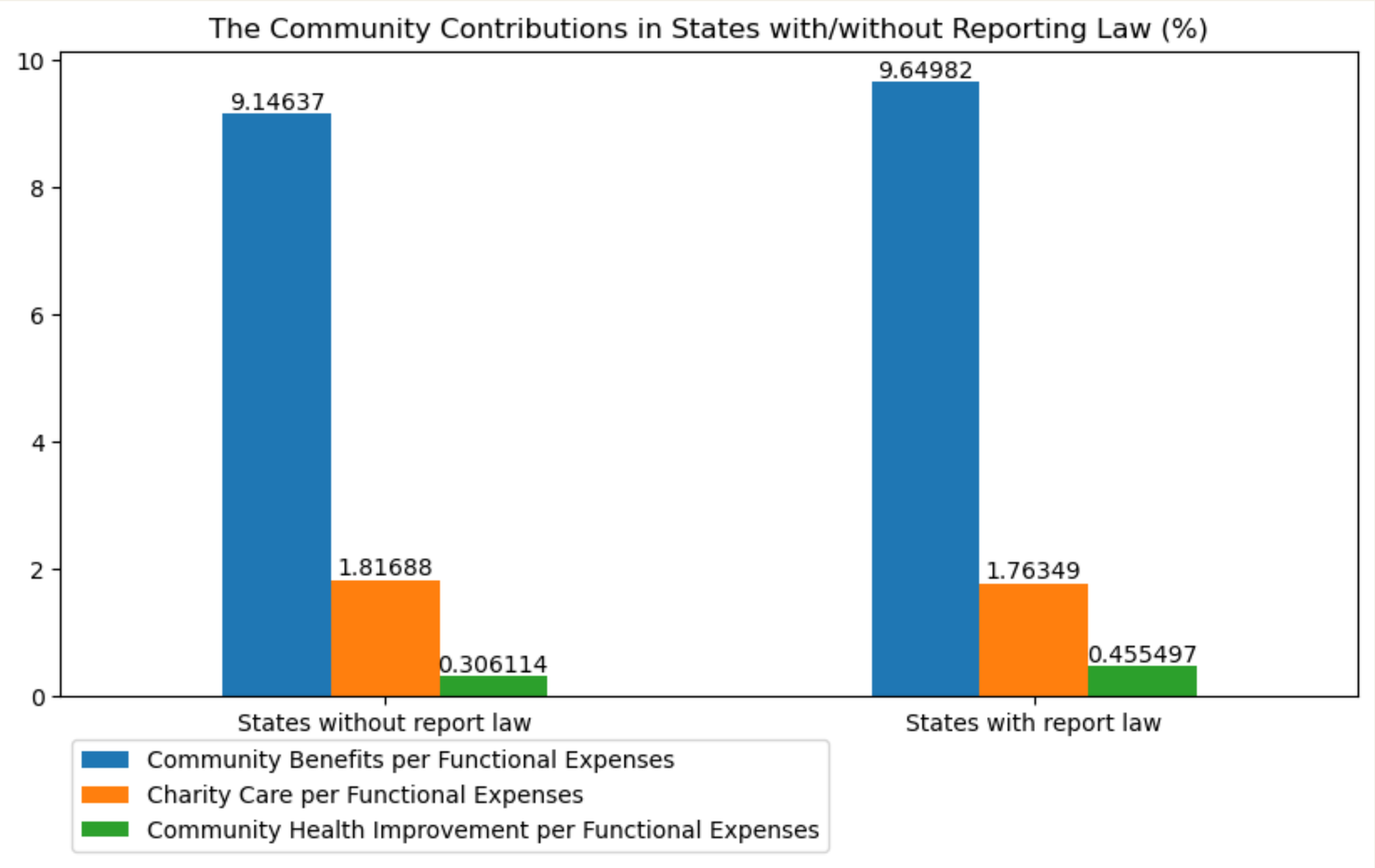
The total USD that *states with reporting laws* is also bigger than *without reporting law states*.

Does it mean that states without reporting law contribute less of their functional expenses for community?

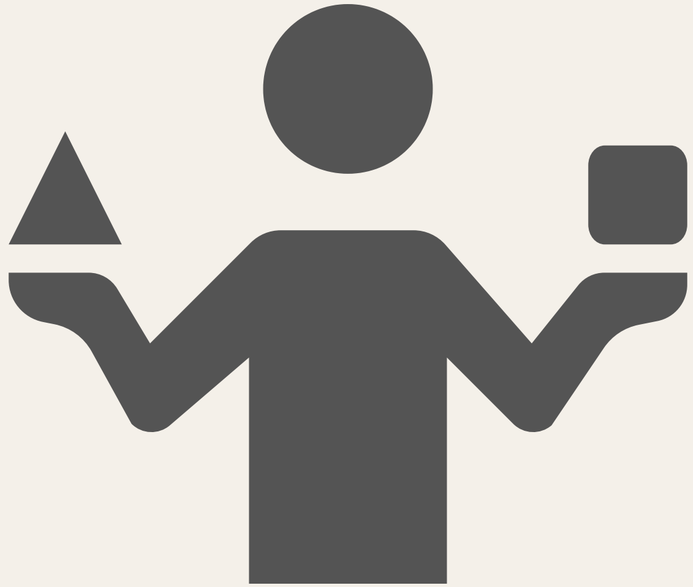
?

Question 5

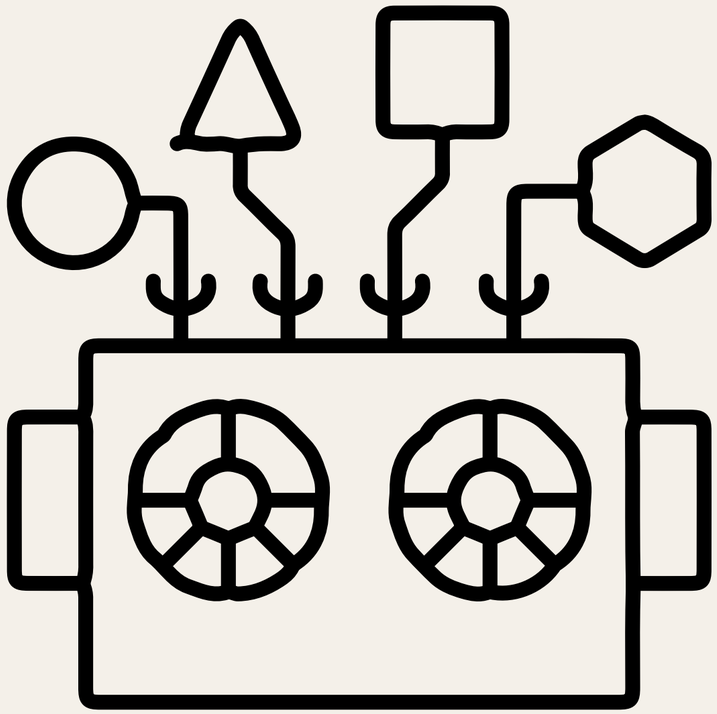
What impact do **state laws requiring hospitals to report community benefits** have on the allocation of resources toward community benefit programs?



Both **states with and without report law** contribute their percentages of functional expenses with the almost same proportion despite the states without reporting law have lower total community benefits.



04 - Data Visualization



4.1 - Problem Statement

4.2 - Data Preparation

4.3 - Modelling

4.4 - Conclusion

Initially, we will use 03 types of model, including regression, clustering, and classification.

In terms of regression, the question is predicting the total revenue of a specific hospital. Next, we will utilize the clustering model to separate hospitals into small, medium, and large groups. Finally, classification is chosen to broadcast which group is a hospital in.

By solving these questions, the locals can decide whether to invest in that hospital or not by the predicted results.

4.1 - Problem Statement

4.2 - Data Preparation

The whole data set will be pre-processed as below:

- Using the previous dataset.
- Dropping object columns (or labeling them).
- Regarding regression, assigning total revenue to y, and dropping it in the dataset.
- Splitting into training set and testing set by tools.

Data

Preparation

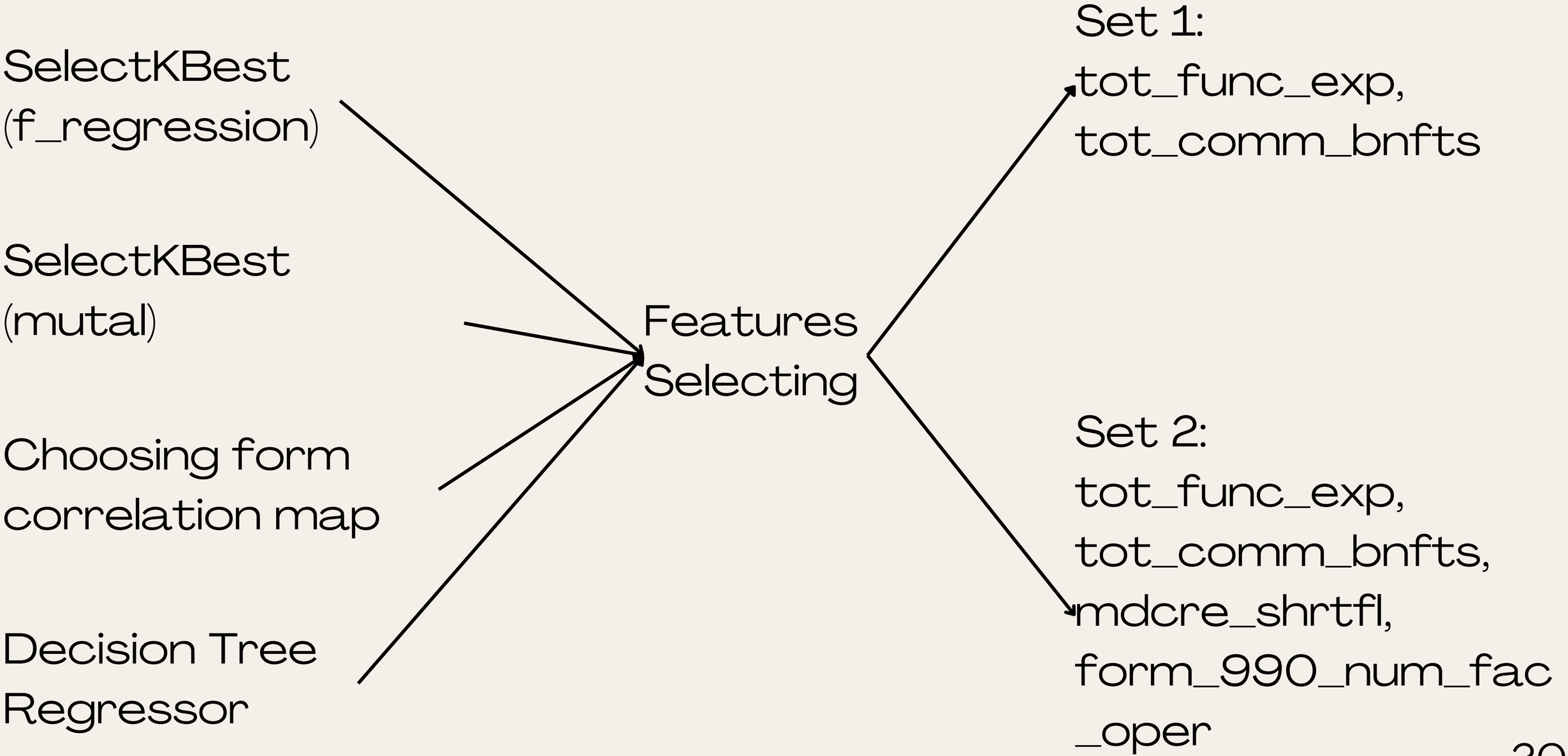
Data analysis facilitates predictive modeling and forecasting

4.3 - Modelling

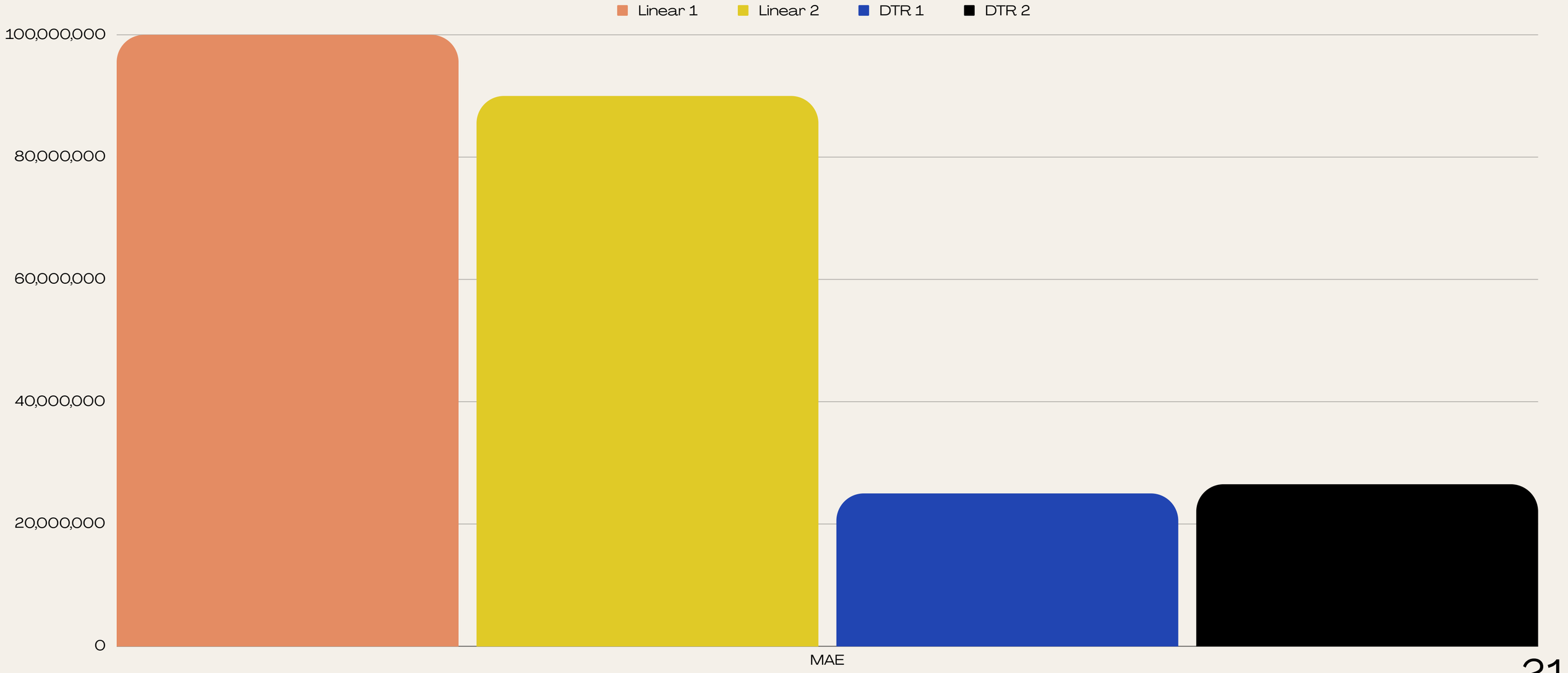
Here are models for this part:

- Regression: LinearRegression and DecisionTreeRegressor.
- Clustering: KMeans and GaussianMixture.
- Classification: RandomForestClassifier and K-Nearest Neighbors

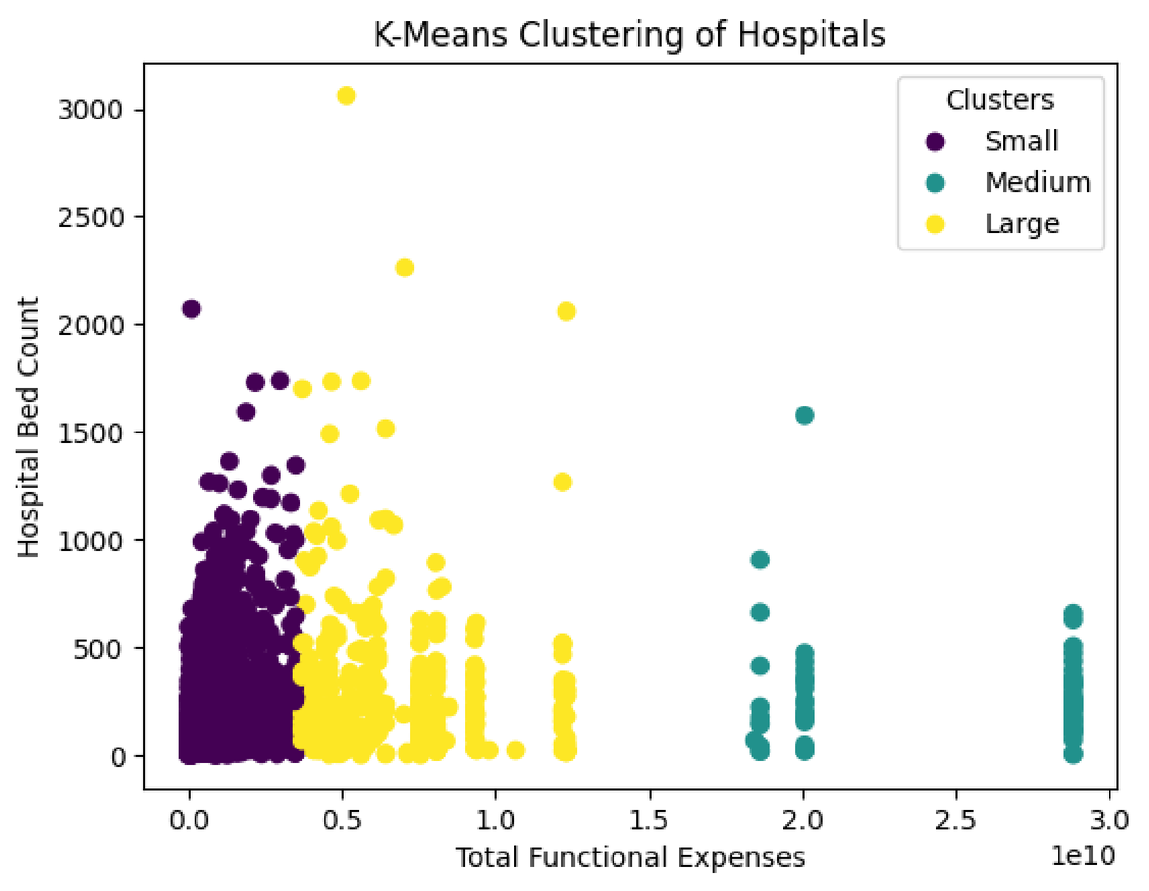
4.3.1 - Regression

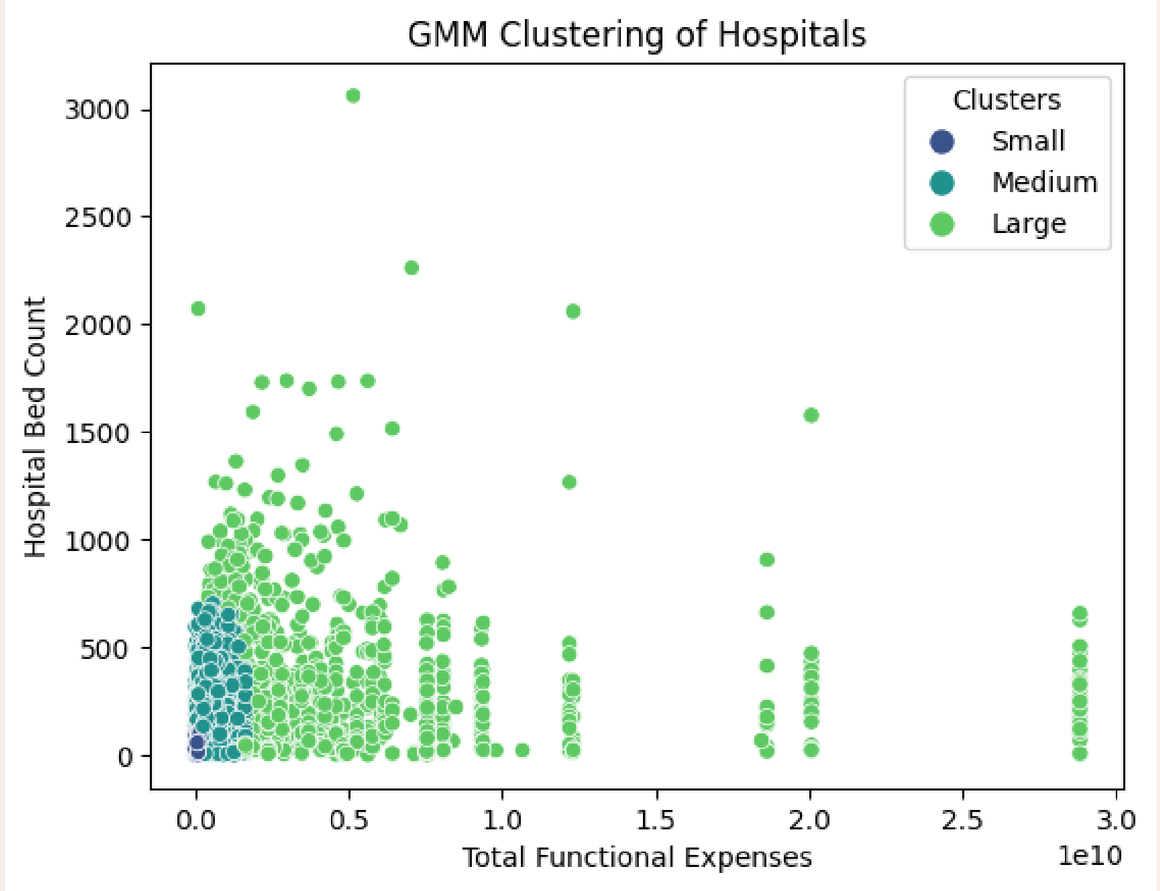


4.3.1 - Regression



4.3.2 - Clustering





4.3.3 - Classification

Selecting feature → Labeling → Modelling

Accuracy (Random Forest): 0.8896848137535817
Classification Report (Random Forest):

	precision	recall	f1-score	support
Large	0.88	0.86	0.87	294
Medium	0.94	0.94	0.94	145
Small	0.88	0.90	0.89	259
accuracy			0.89	698
macro avg	0.90	0.90	0.90	698
weighted avg	0.89	0.89	0.89	698

Accuracy (K-Nearest Neighbors): 0.9011461318051576
Classification Report (K-Nearest Neighbors):

	precision	recall	f1-score	support
Large	0.90	0.86	0.88	294
Medium	0.96	0.91	0.94	145
Small	0.87	0.94	0.91	259
accuracy			0.90	698
macro avg	0.91	0.90	0.91	698
weighted avg	0.90	0.90	0.90	698

4.4 - Conclusions

- Regression:
  - Decision Tree is better than Linear (flexibility).
  - By selecting, we can choose weighty features.
- Clustering:
  - Gaussian (more complex) is better than KMeans (simple and fast).
- Classification:
  - Good performance, both are equipvalent.
  - Each has its own benefits, depending on the data set.
- Overall, the performance may depend on the data's characteristics. training and testing set are vital, we can select valuable features by numerous tools.

Thanks

Group 11