MEDICAL REPRESENTATIVE

Aligning Medicines with the Right Doctors.

mazensafty7.github.io

mazensafty2003@gmail.com

Digital Egypt Pioneers

AGENDA

- Introduction
- Problem
- How To Solve
- Project Phases
- Data Collection
- Data Exploration (EDA)
- Data Analysis
- Data Cleaning And Preprocessing
- ML Model Developing
- Model Deployment
- Conclusion

INTRODUCTION

Medical representatives act as the primary link between pharmaceutical companies and healthcare professionals, promoting products such as drugs and medical equipment. They engage with doctors, nurses, and pharmacists to raise awareness, answer questions, and build strong relationships. A key challenge for medical representatives is convincing doctors to prescribe their company's drug over competitors with the same active ingredients. Success in this role requires effectively communicating the product's advantages and fostering trust with healthcare professionals.

PROBLEM

The current process for medical representatives is both costly and inefficient. They must visit a multitude of doctors, clinics, and hospitals, often investing significant time and resources without any guarantee that physicians will prescribe their medications. This lack of certainty not only leads to wasted efforts and increased operational costs but also hinders the ability to effectively target healthcare professionals who are more likely to be receptive to their products. Consequently, medical representatives face challenges in optimizing their outreach strategies and maximizing their impact in promoting the right medications to the right patients.

HOW TO SOLVE

Steps To Solve

- Data Exploration: Uncover insights from historical prescribing data.
- Data Analysis: Identify key factors influencing prescribing behaviors.
- Machine Learning Model: Predict prescription likelihood using advanced algorithms.

End Product

A desktop application that deploys the predictive model, providing accurate recommendations for medical representatives to effectively target doctors likely to prescribe the right medications.

PROJECT PHASES

1. Data Collection
2. Data Exploration (EDA)
3. Data Analysis
4. Data Cleaning and Preprocessing
5. Model Selection and Tuning
6. Model Deployment

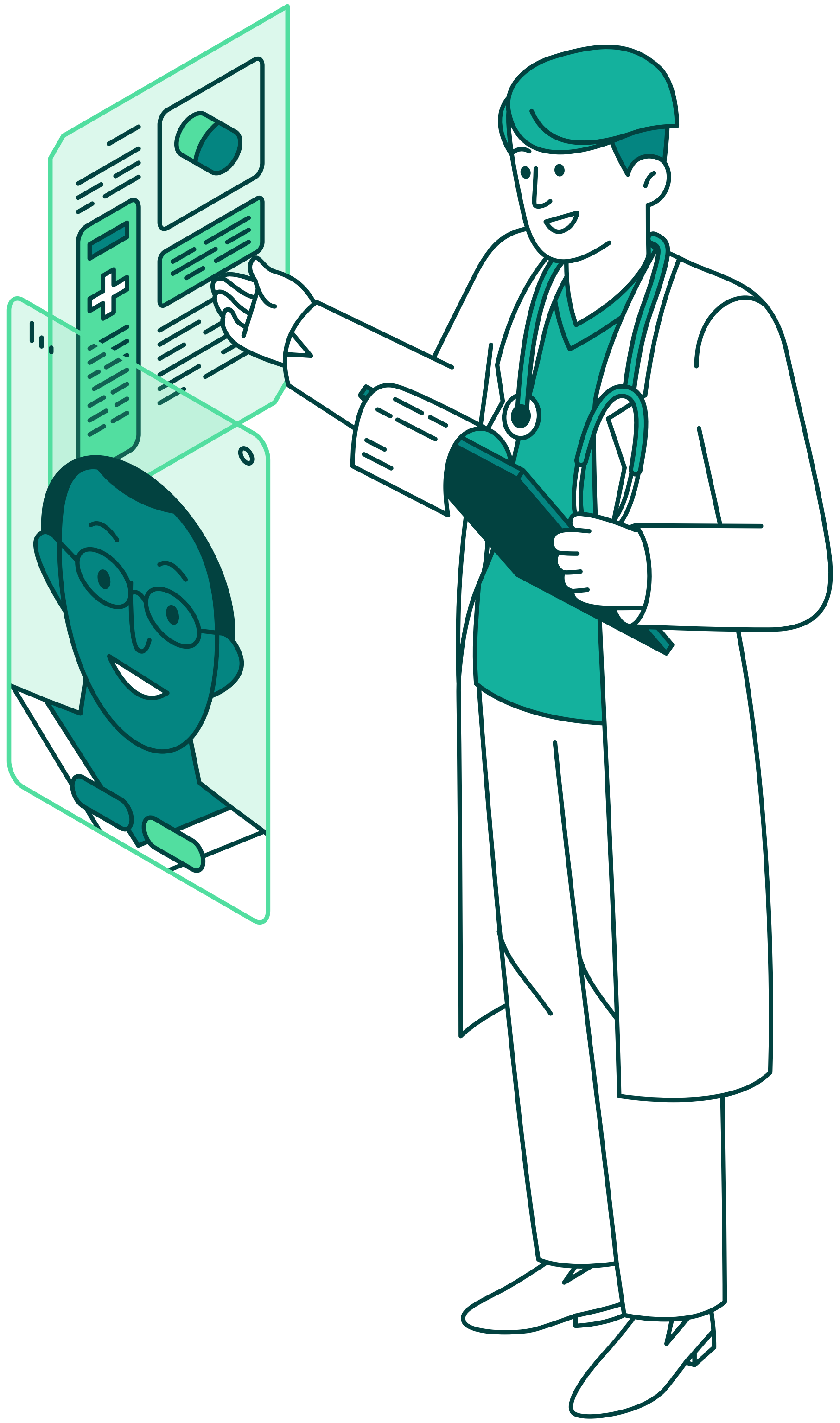
DATA COLLECTION

1. Data Source

- Company's SQL Database: Two primary tables with detailed information on medicines and doctors.

2. Data Tables Overview

- **medicine_table:**
  - id_m: Unique identifier for each medicine.
  - medicine: Commercial name, categorized as type1 to type6.
  - price: Cost per drug for patients.
- **doctor_table:**
  - id_dr: Unique identifier for each doctor.
  - exam_price: Examination fee charged by the doctor.
  - clinic_hos: Indicates whether the doctor operates in a private clinic or a hospital.
  - dr_class: Classification based on doctor popularity and patient volume, categorized as ‘a’ or ‘b’.



DATA COLLECTION

3. Data Preparation Process

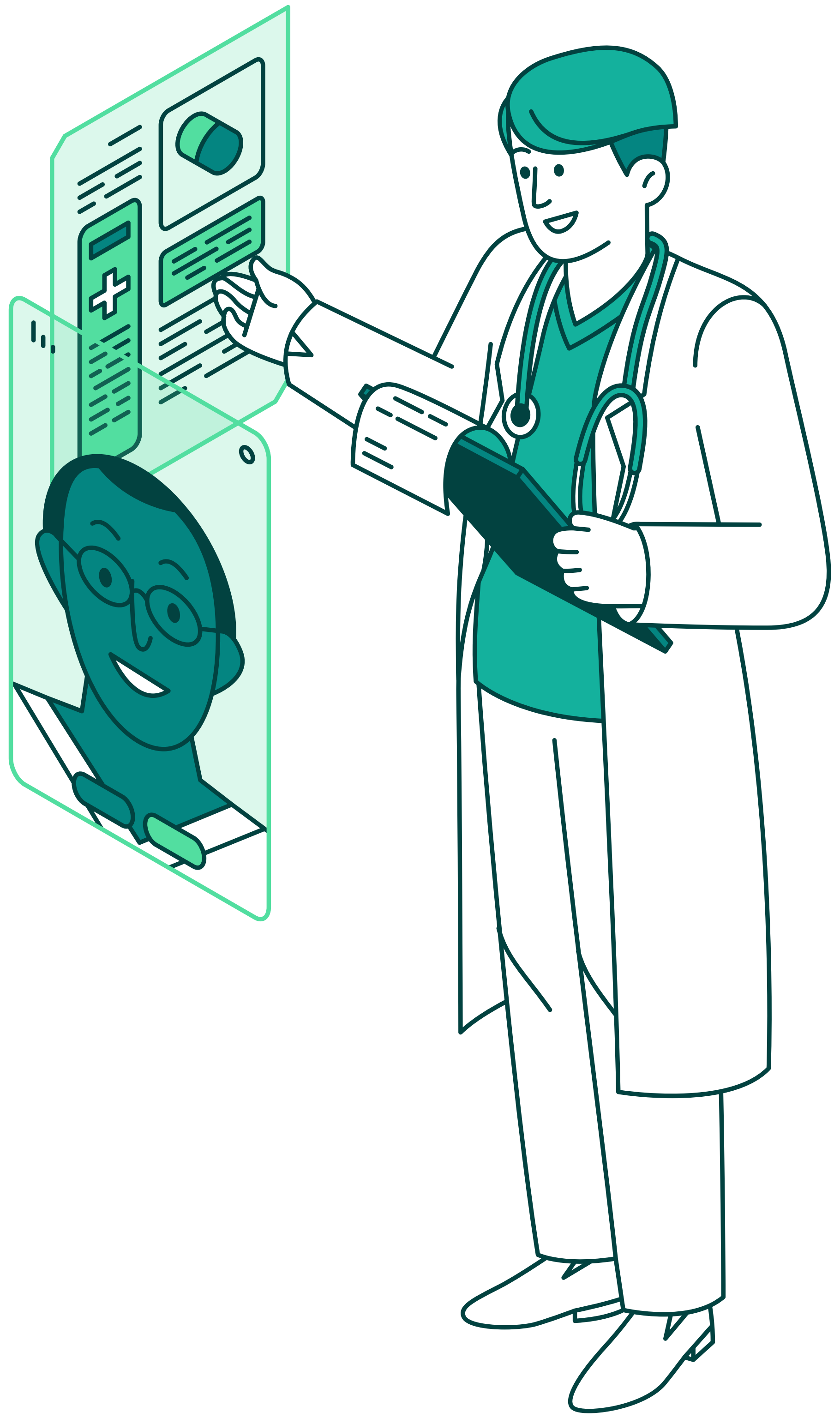
- SQL Magic: Used SQL queries to retrieve data from both tables.
- Concatenation: Merged tables into a single DataFrame.
- Data Cleaning: Dropped redundant ID columns to streamline analysis.

Convert it to DataFrame

In [6]:
```
data = data.DataFrame()
data.head()
```

Out[6]:

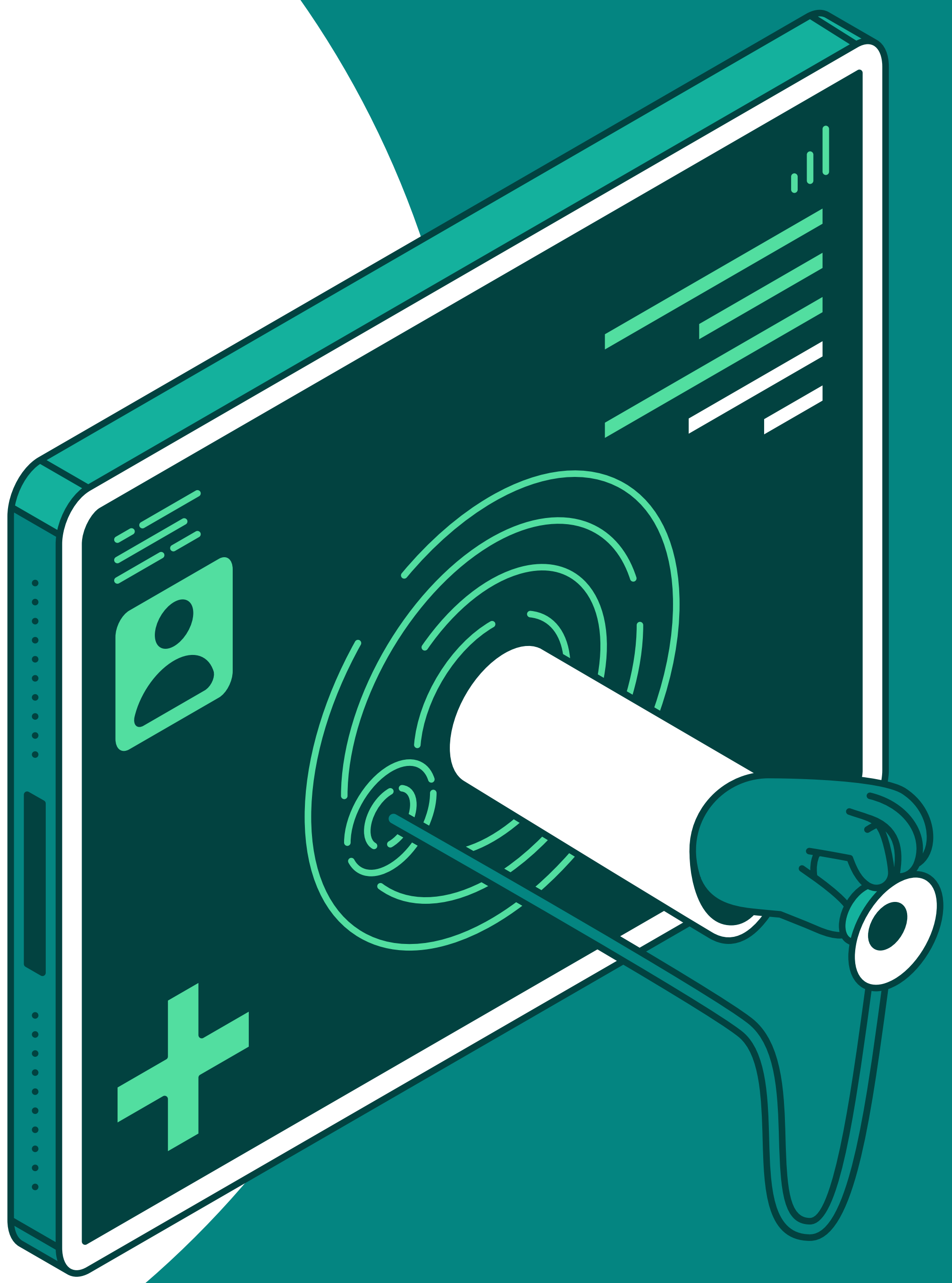
	id_m	medicine	price	id_dr	area	speciality	dr_class	exam_price	clinic_hos	write
0	1	type1	45	1	area1	chest	a	200	clinic	1
1	2	type4	36	2	area2	im	b	100	clinic	1
2	3	type1	45	3	area8	chest	a	75	hospital	1
3	4	type1	45	4	area5	chest	a	30	hospital	1
4	5	type5	29	5	area6	uro	a	220	clinic	0



DATA EXPLORATION (EDA)

1. Initial Analysis

- Row Count: 390
- Unique Categories:
  - Medicines: Types (type1 to type6).
  - Doctor Classes: Classification of doctors based on patient volume and popularity ('a' and 'b').
  - Clinic Type: Doctors working in private clinics or hospitals.
  - Specialties:
    - Chest: Chest Specialist
    - IM: Internal Medicine Specialist
    - CD: Cardiology Specialist
    - Neuro: Neurology Specialist
    - GIT: Gastrointestinal Tract Specialist
    - ENT: Ear, Nose, and Throat Specialist
    - Sur: Surgery Specialist
    - Uro: Urology Specialist
    - GP: General Practitioner
    - Or: Orthopedic Specialist
    - Vas: Vascular Specialist

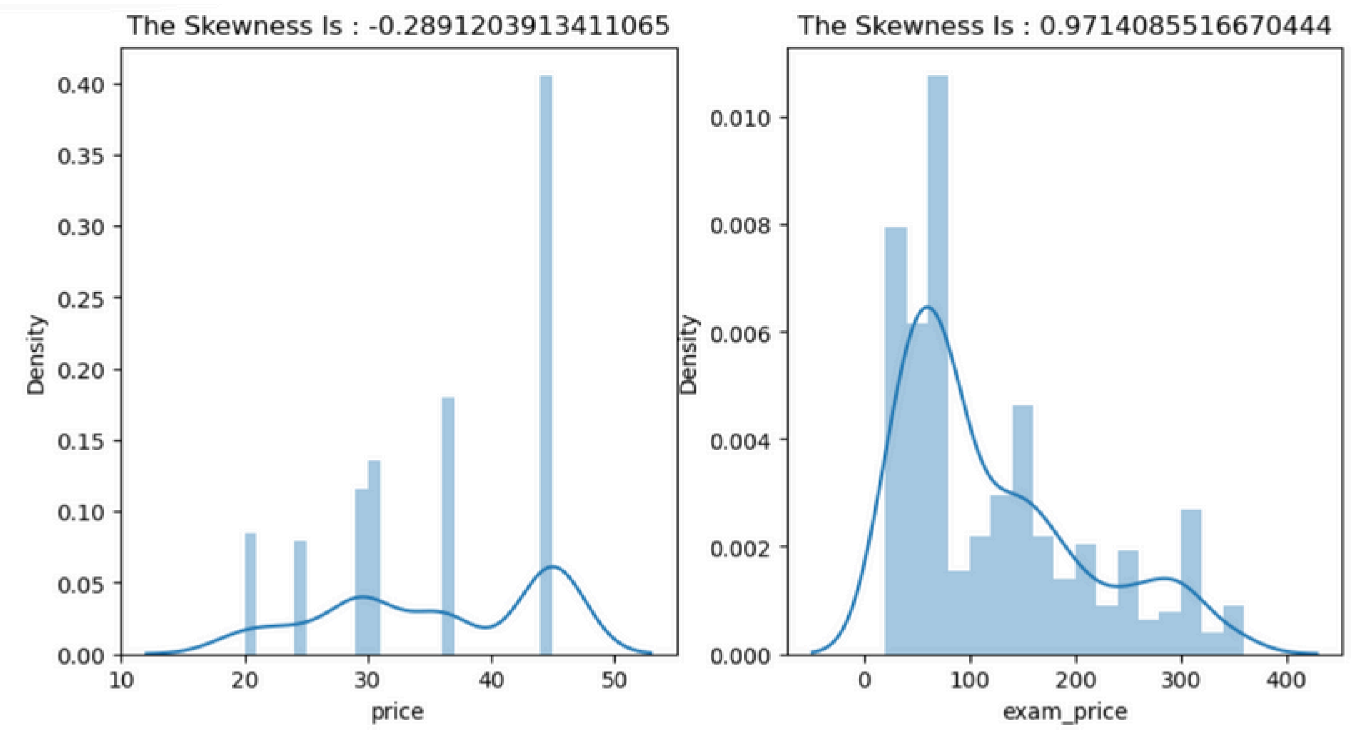


DATA EXPLORATION (EDA)

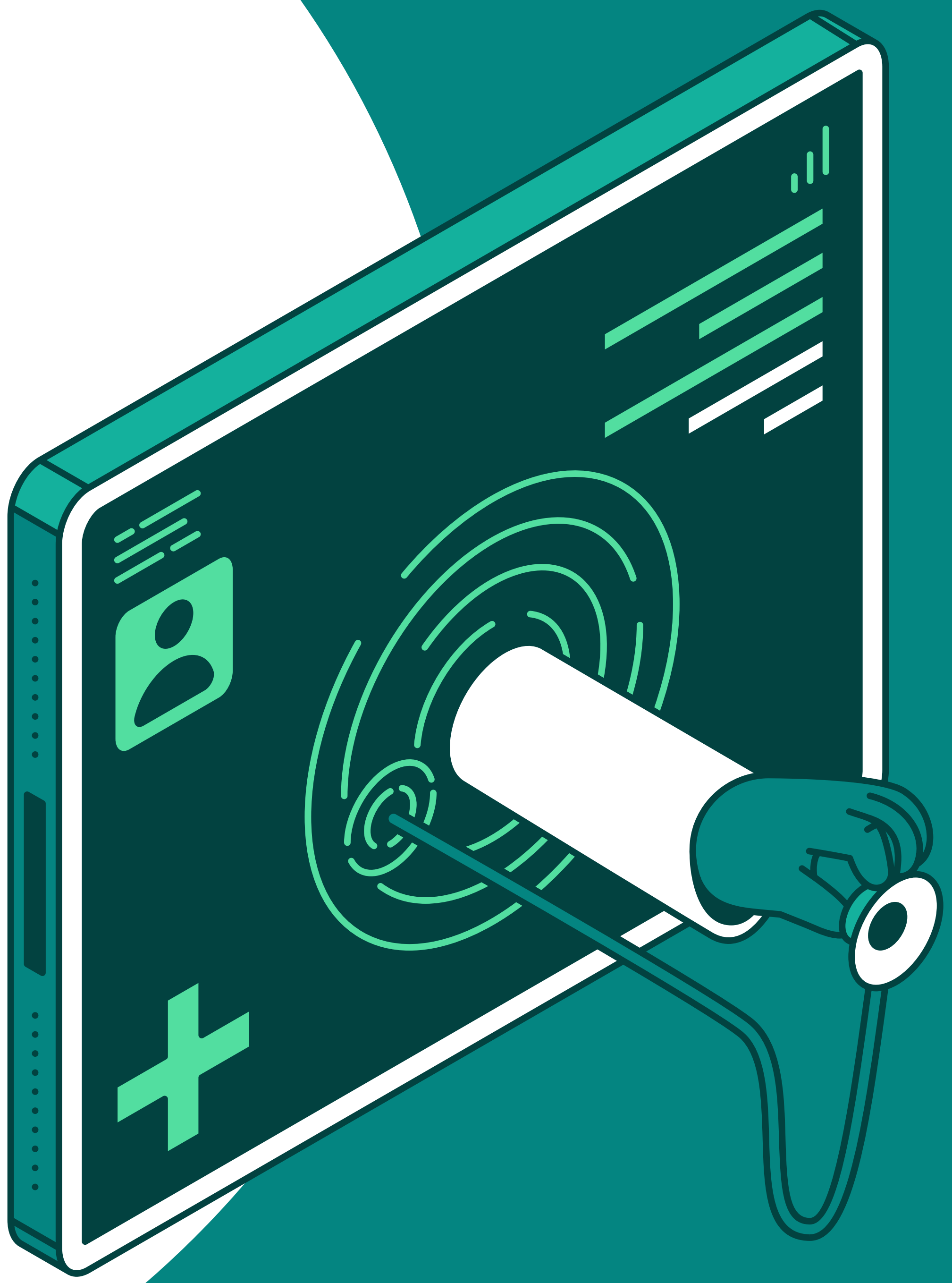
Key Descriptive Statistics

	price	exam_price	write
count	390.000000	390.000000	390.000000
mean	35.715385	121.205128	0.587179
std	8.751263	86.729844	0.492974
min	20.000000	30.000000	0.000000
25%	29.000000	50.000000	0.000000
50%	36.000000	80.000000	1.000000
75%	45.000000	170.000000	1.000000
max	45.000000	350.000000	1.000000

Distribution Insights

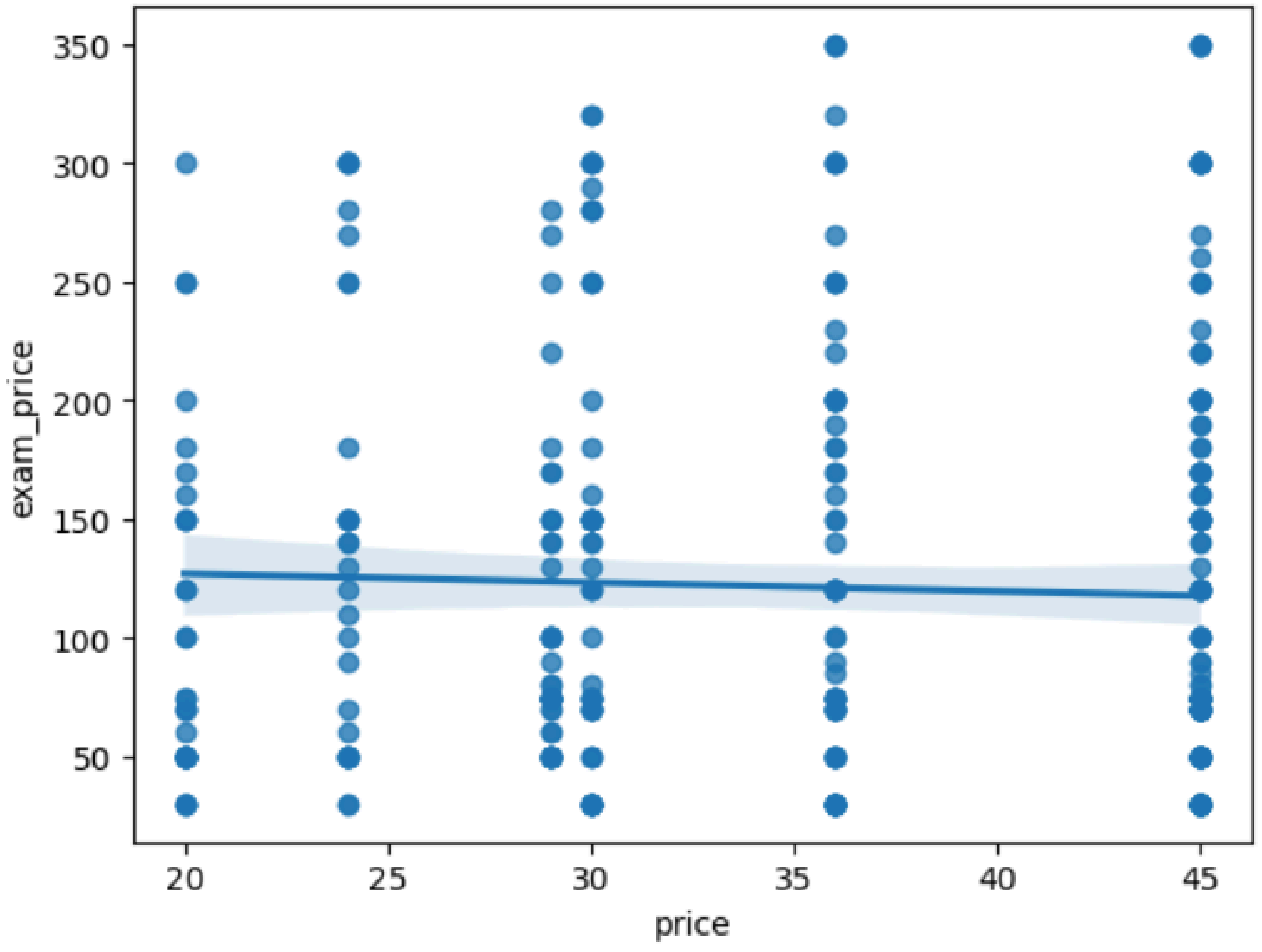


- The medicine prices demonstrate a normal distribution, evidenced by the close equality of the mean and median values. In contrast, the examination prices are right-skewed due to the generally higher fees charged by clinics compared to hospitals.



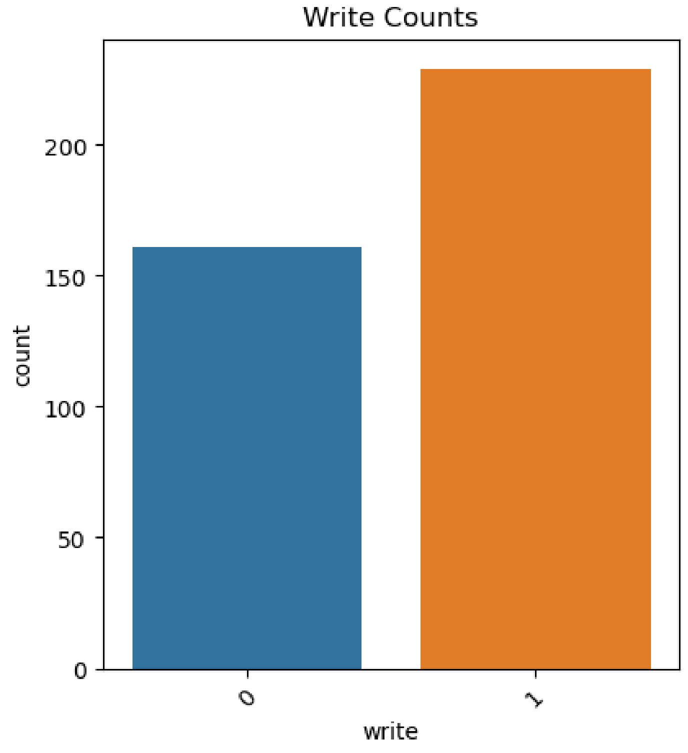
DATA EXPLORATION (EDA)

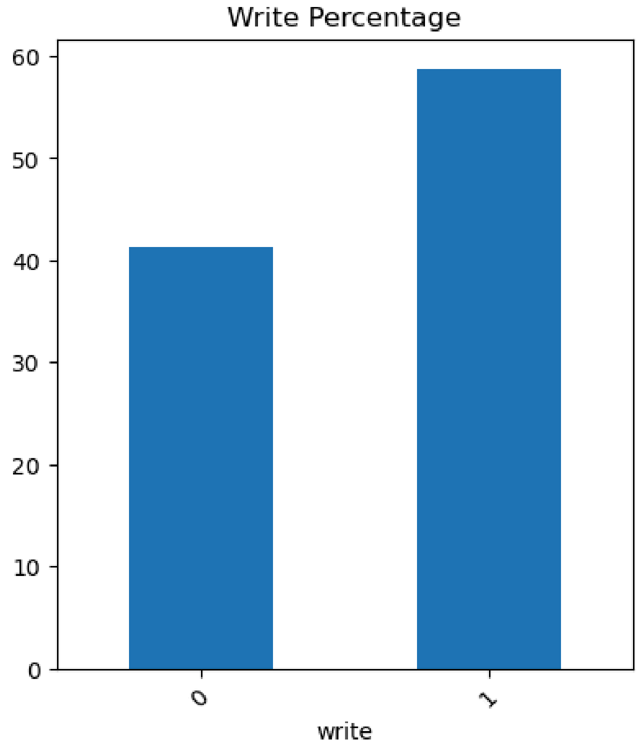
Correlation Insights



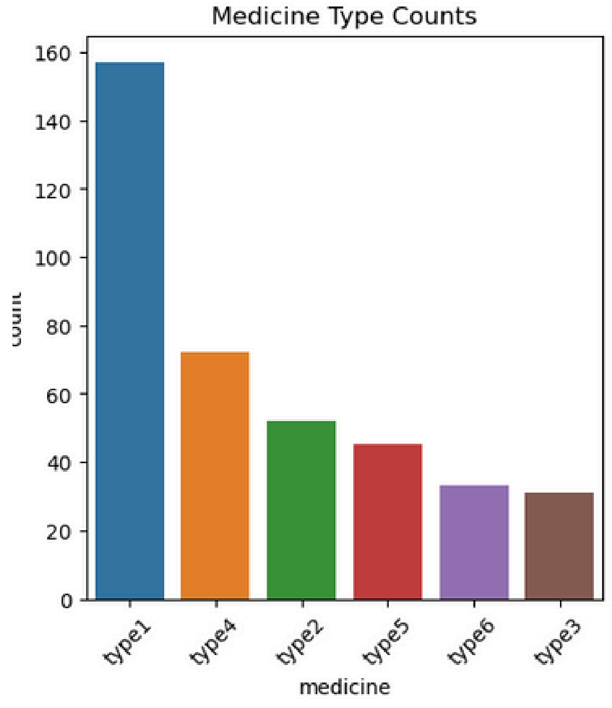
DATA EXPLORATION (EDA)

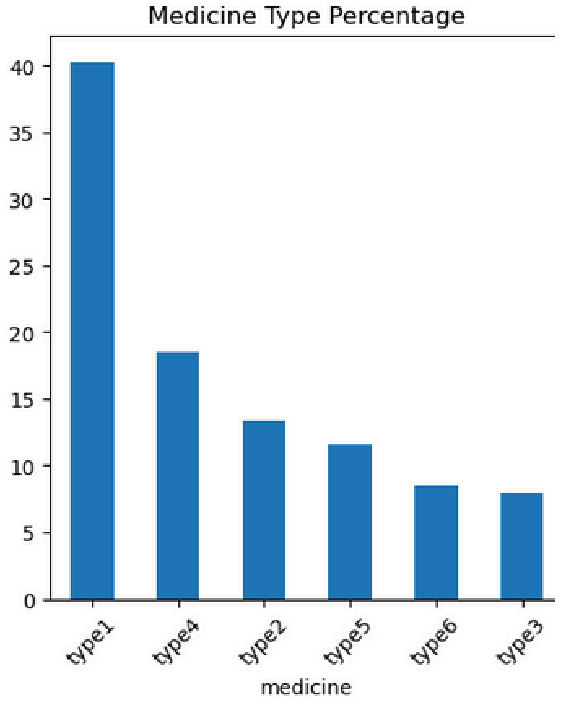
The Percentage and Counts of doctors how write is more than how didn’t in all data





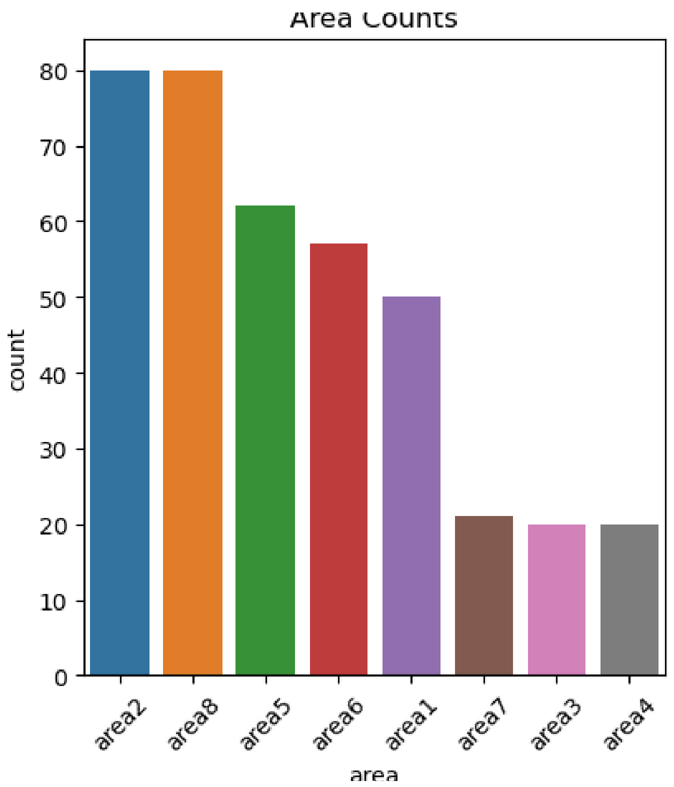
The Percentage and Counts of Type 1 more than any other types in all data

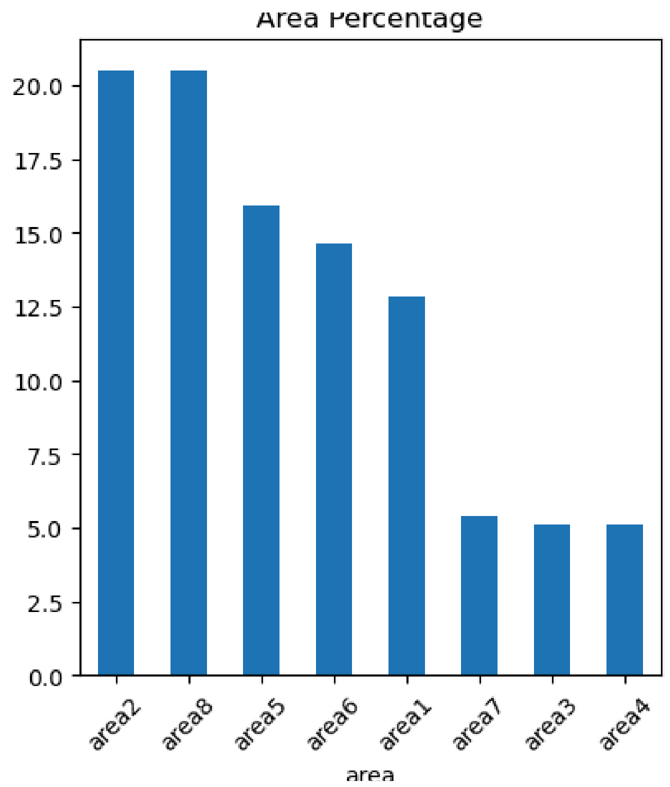




DATA EXPLORATION (EDA)

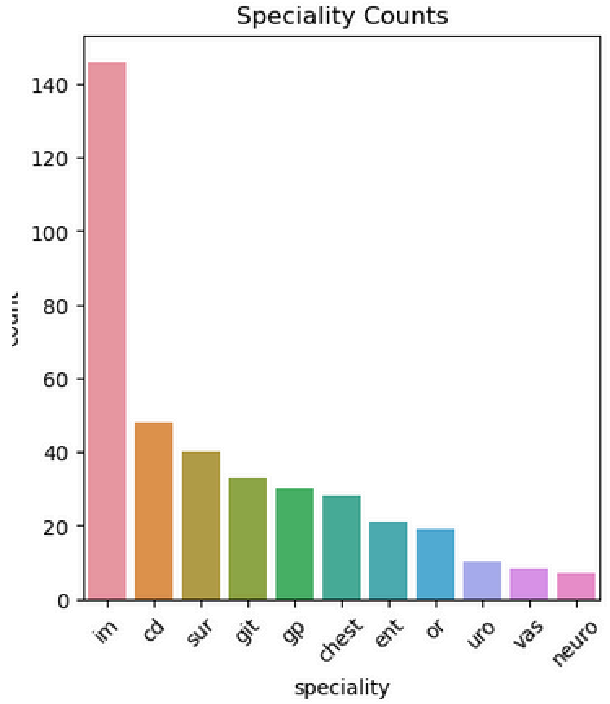
The Percentage and Counts of doctors in area 2 , 8 is more than any area

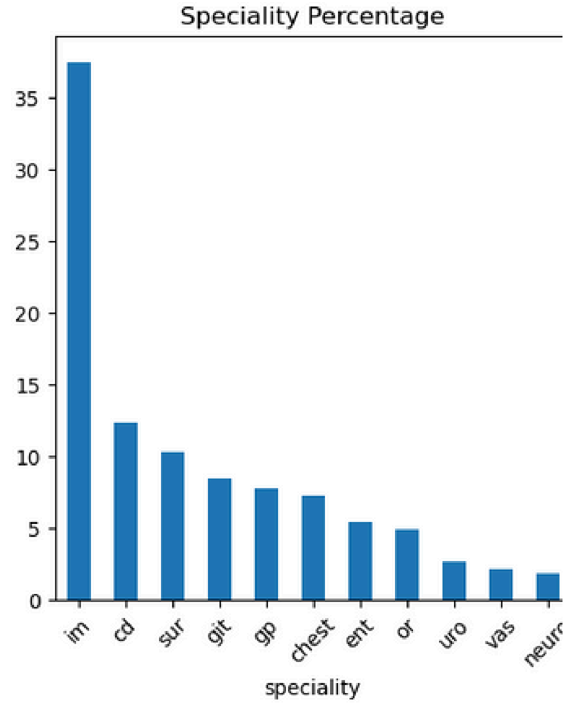




im doctors is more

Because : This prevalence is due to the fact that physicians in this specialty are highly skilled in managing a wide range of medical conditions, which makes them more represented in the data.

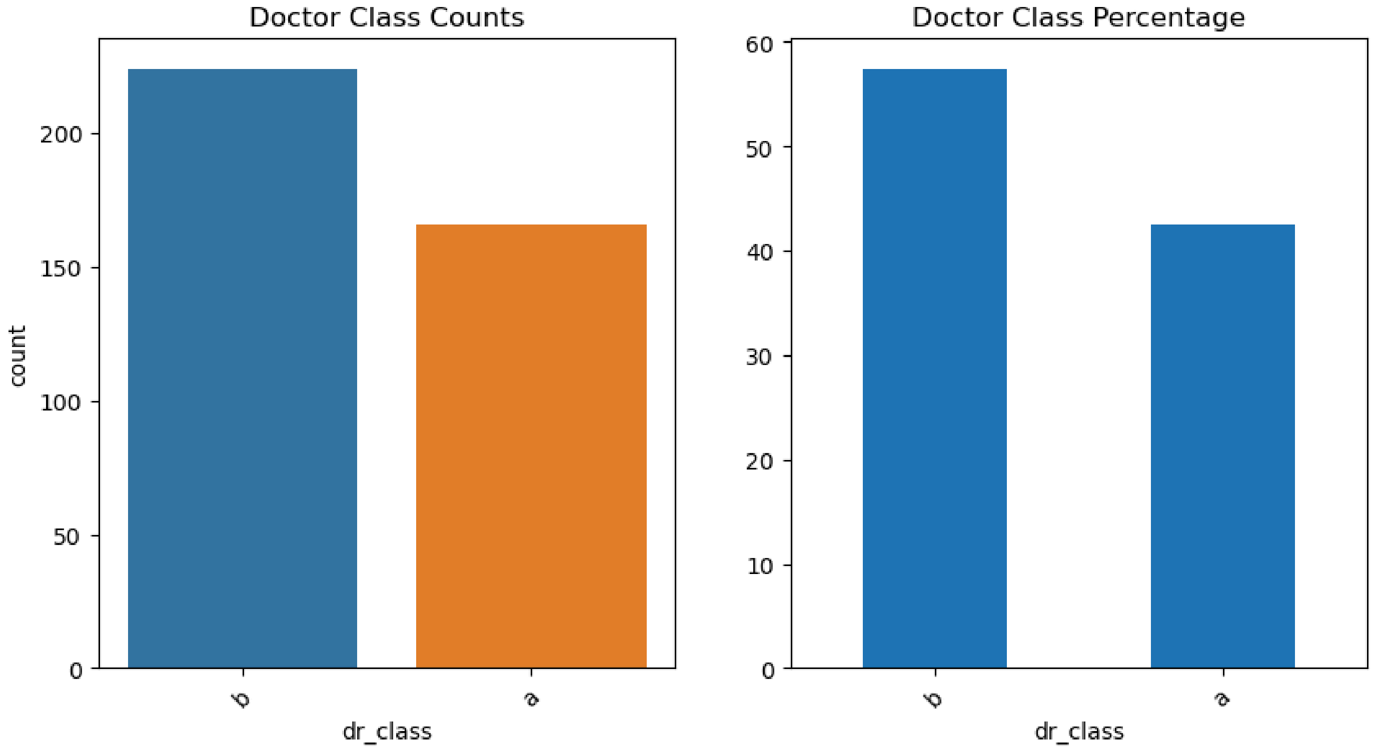


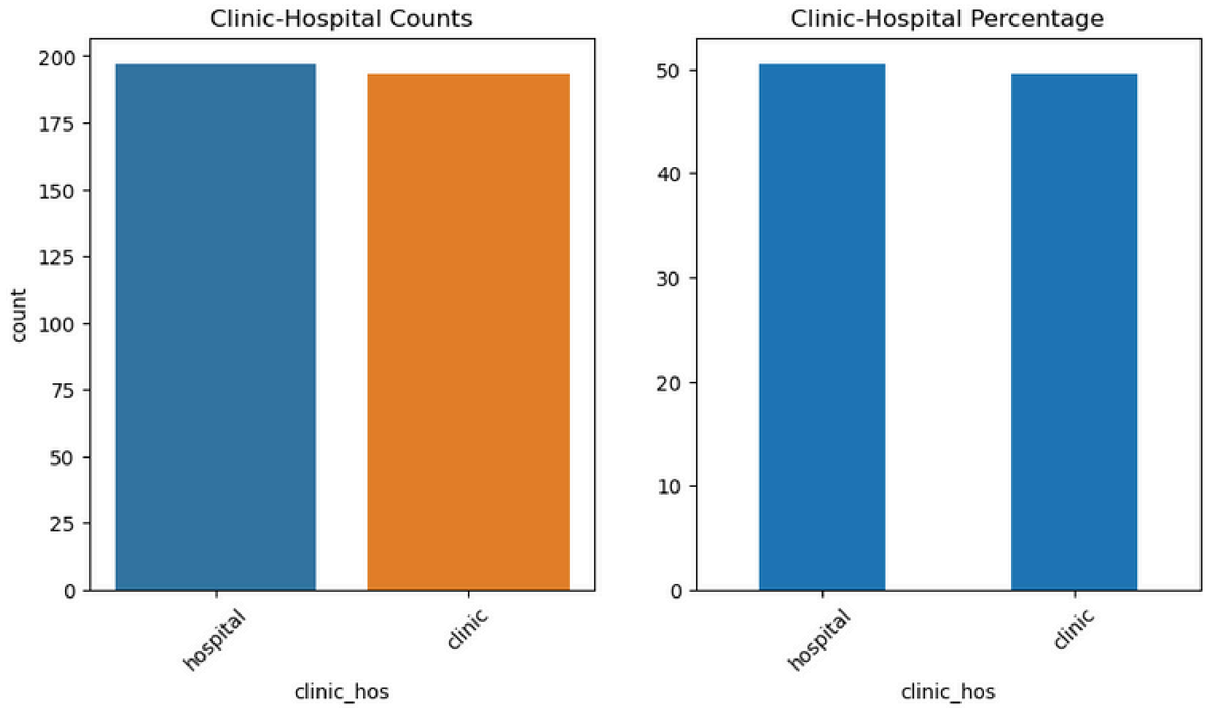


DATA EXPLORATION (EDA)

The Percentage and Counts of doctors in class b is more than class a

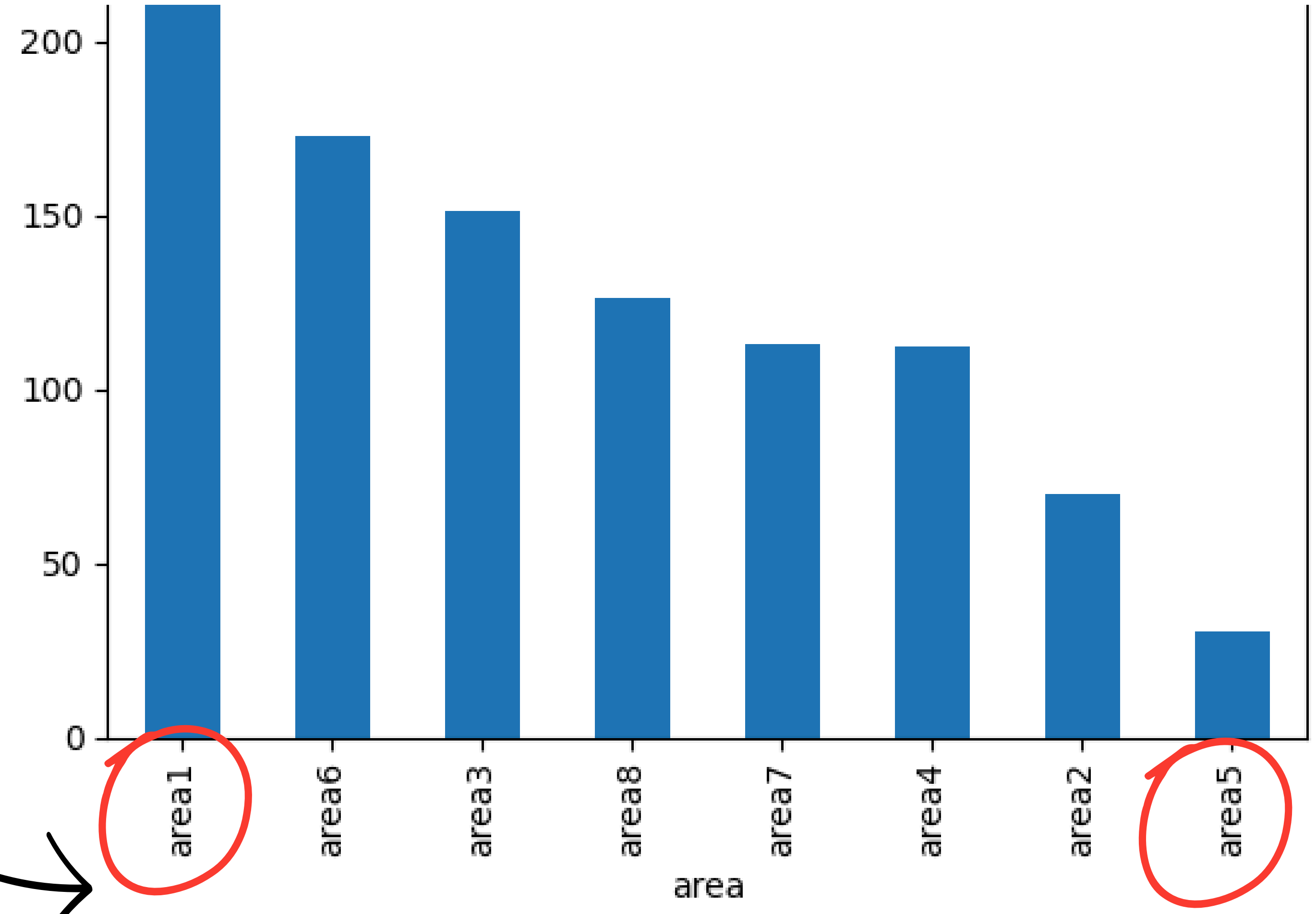
The Figure show that there is a balance between doctors in clinics and hospitals in dataset





DATA EXPLORATION (EDA)

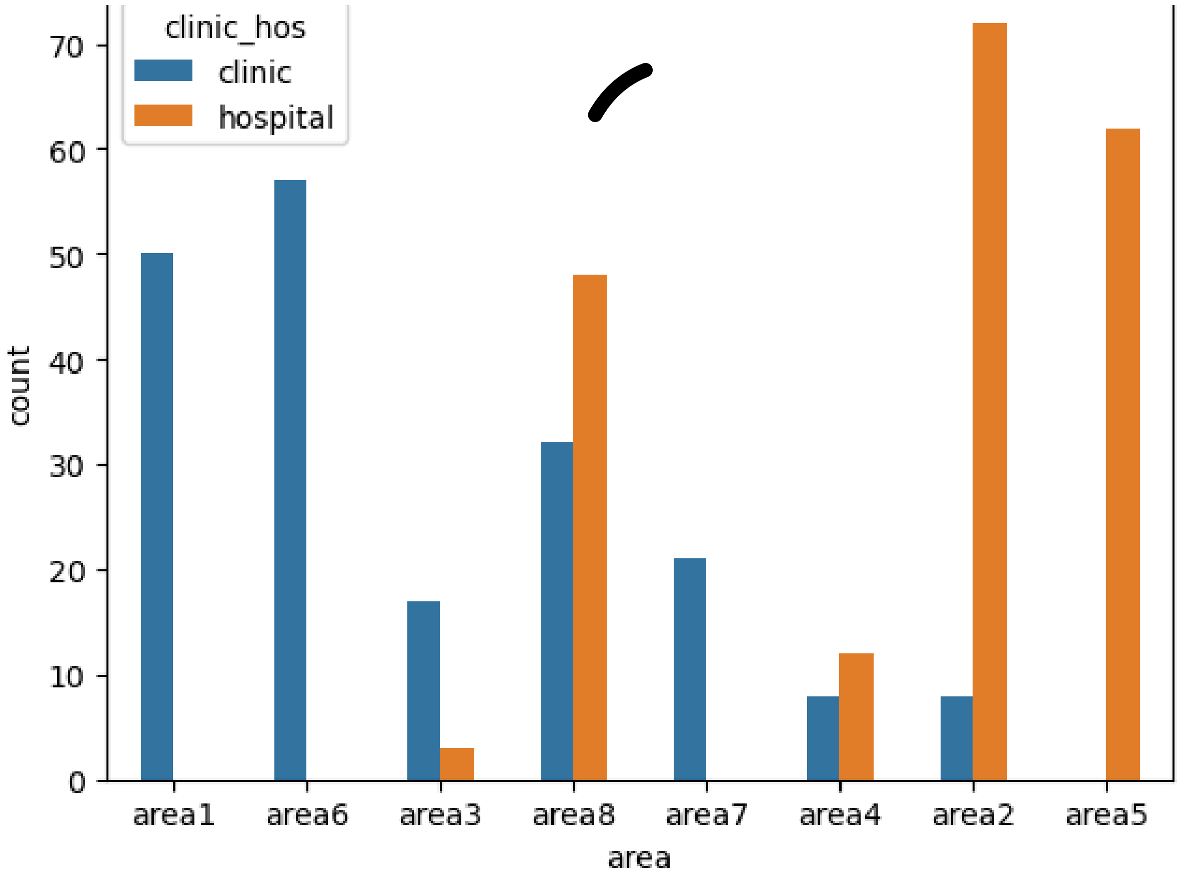
The average of examination price in each area



Most expensive area

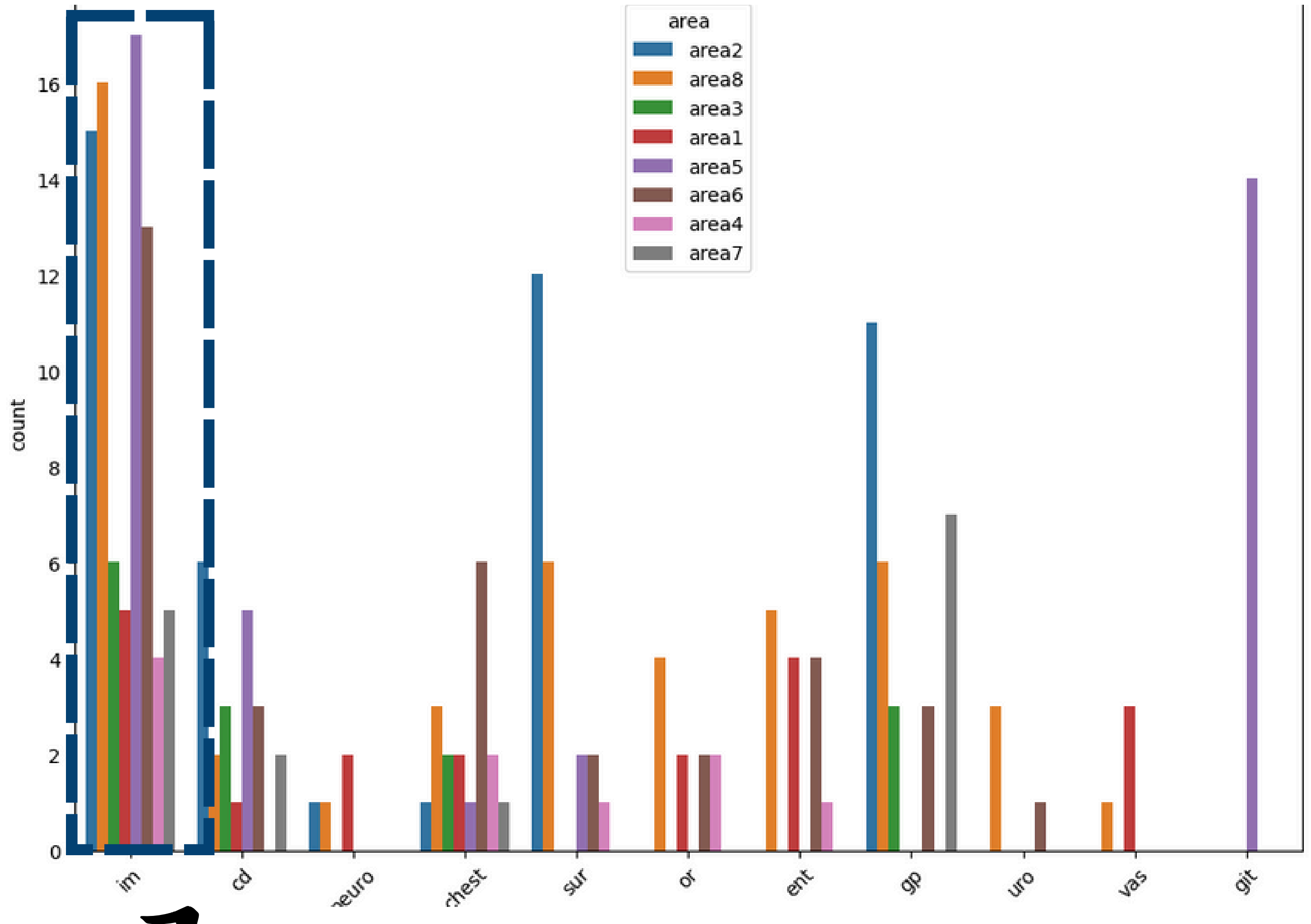
Least expensive area

Most Hospitals in least expensive areas because the exam price in hospital is very low compare to clinics

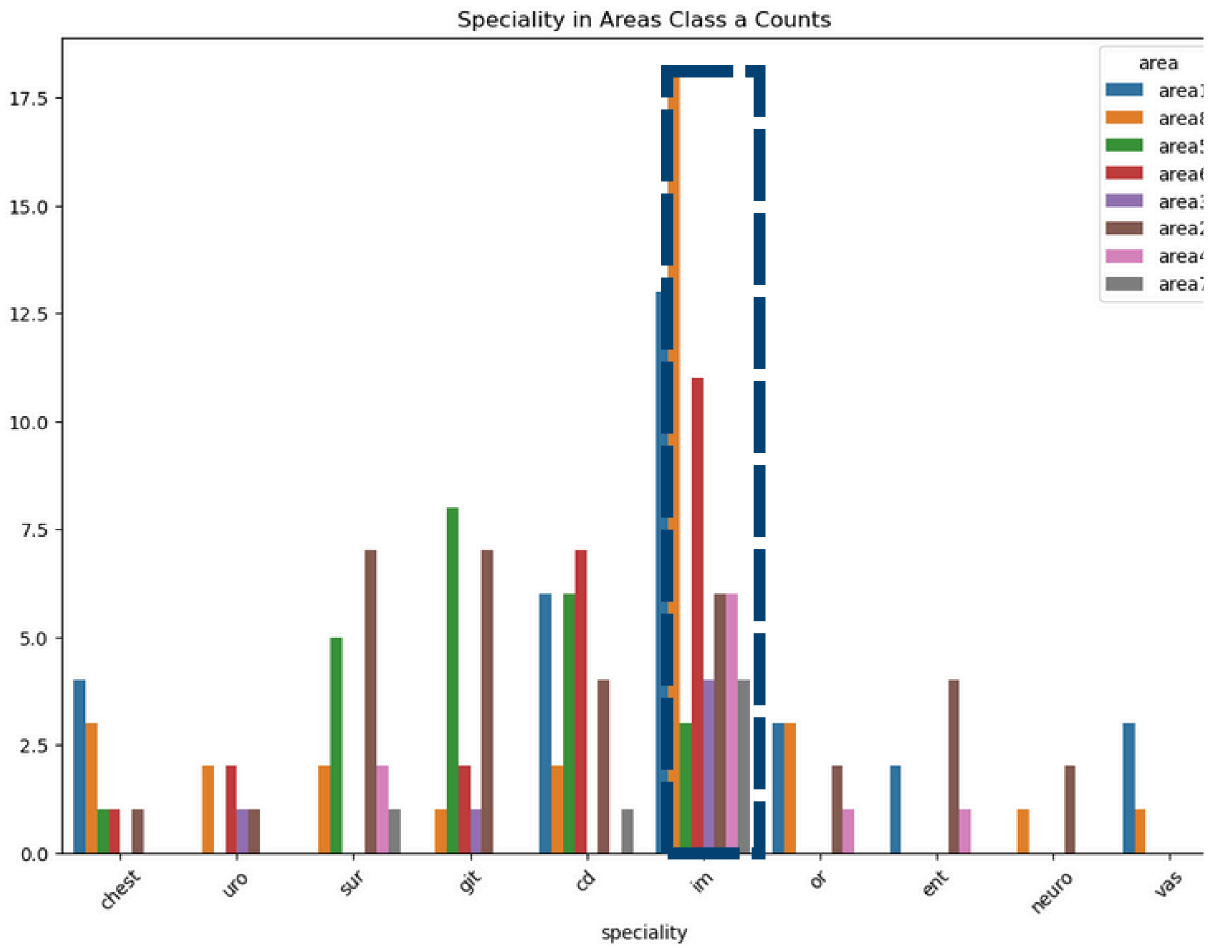


DATA EXPLORATION (EDA)

Distribution of class b doctors of each speciality in each area

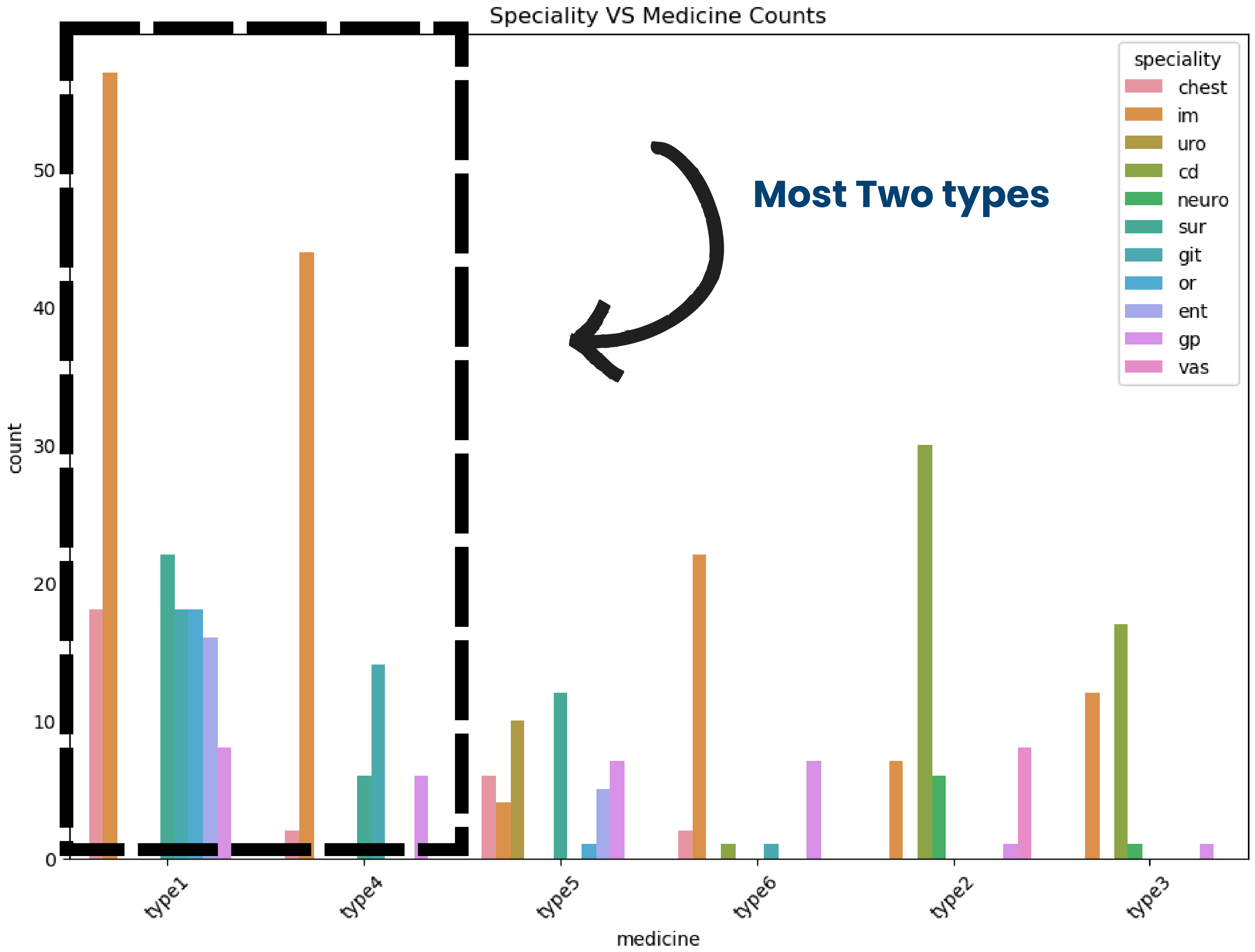


Distribution of class a doctors of each speciality in each area



im doctors is more and in all areas

DATA EXPLORATION (EDA)

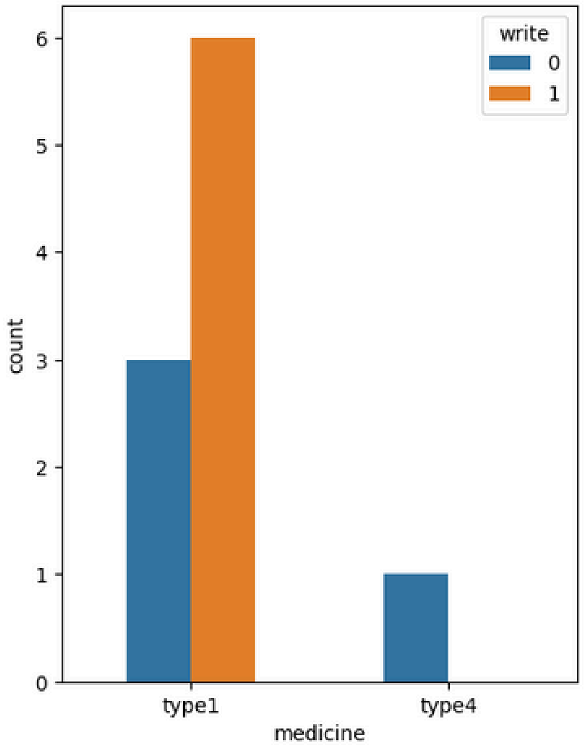


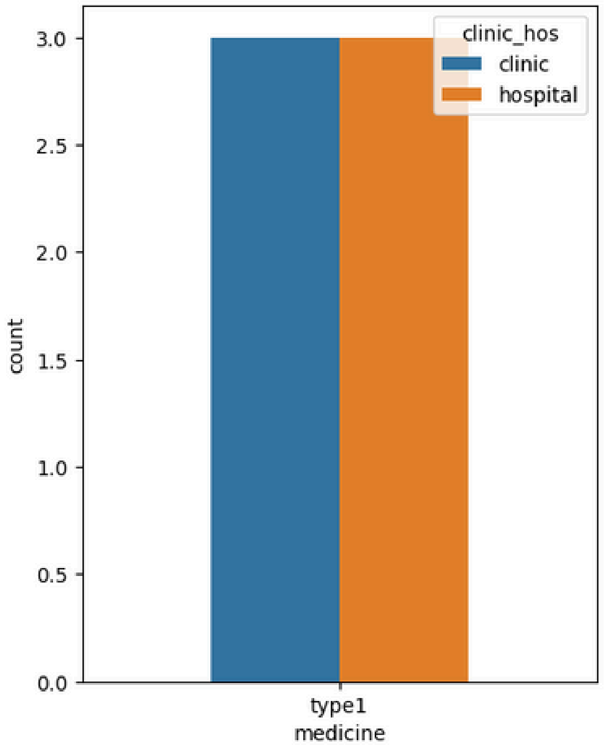
Distribution of medicines types of every doctor speciality

DATA ANALYSIS

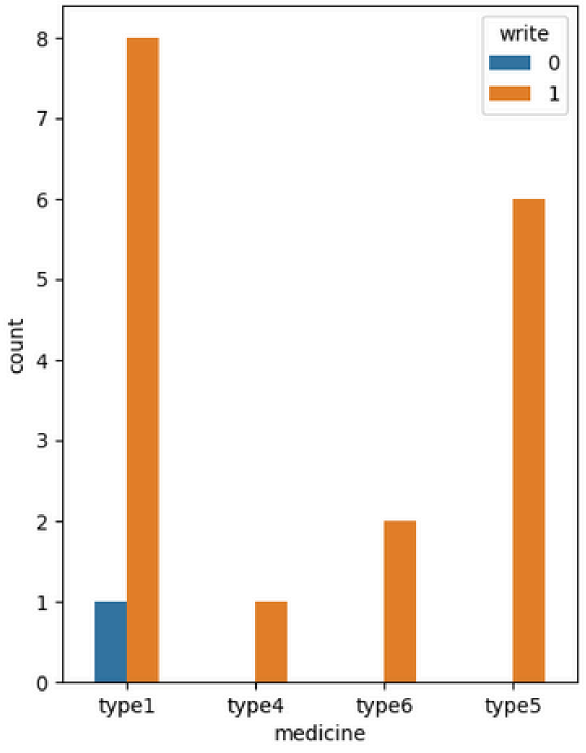
Chest Doctors

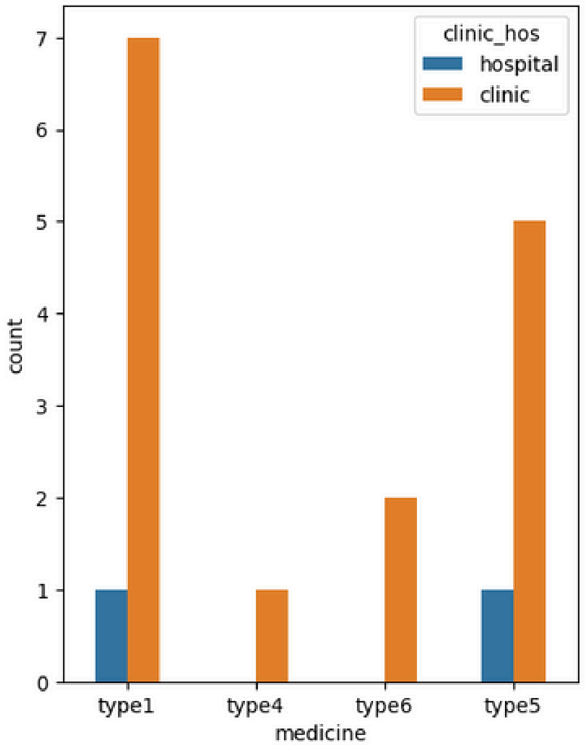
Most chest doctors in Class a write Type1 Medicine in clinics and hospitals

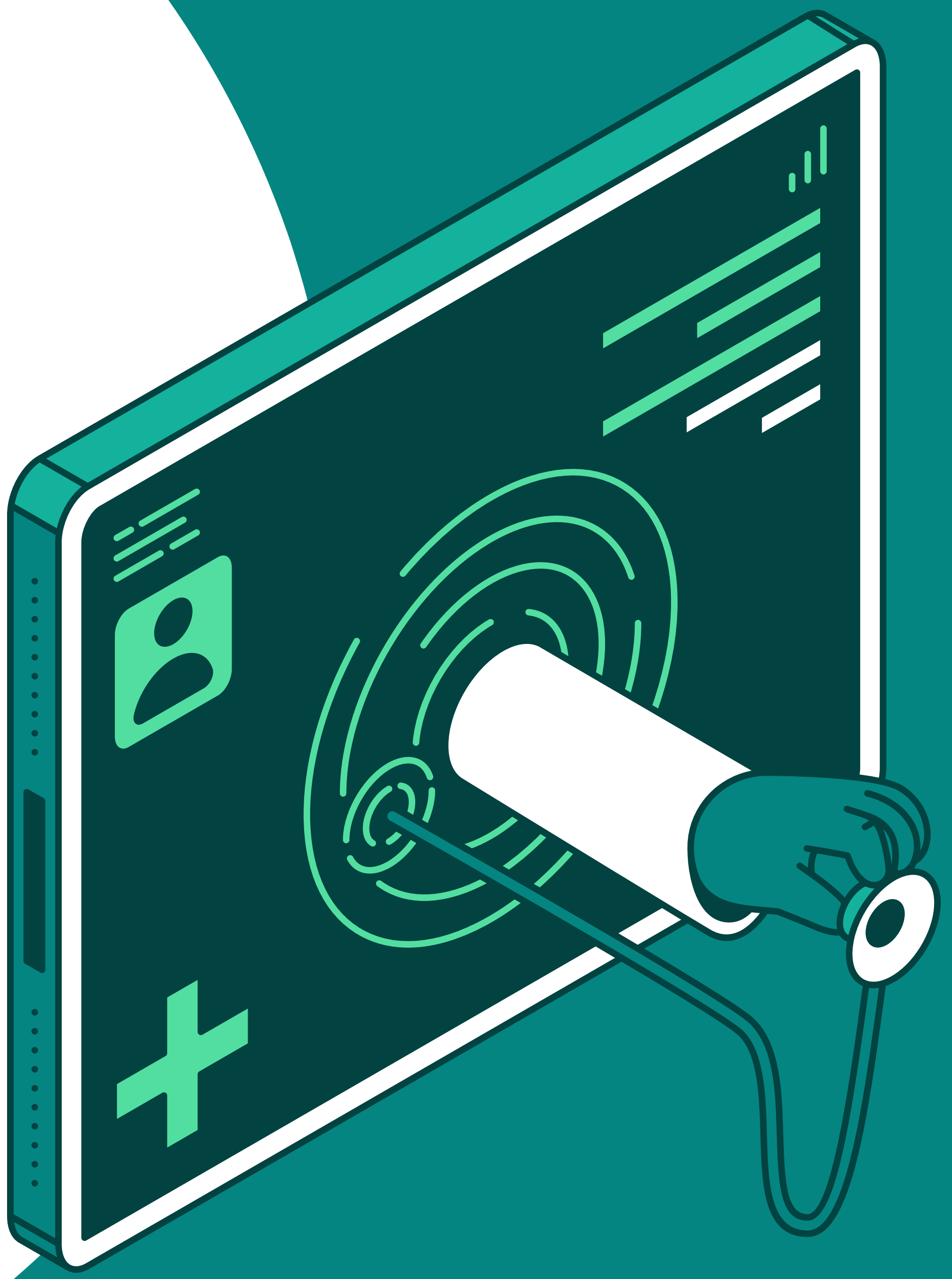




Most chest dcotors in Class b write Type1 Medicine and they also write many other cheap types 4 , 6 , 5

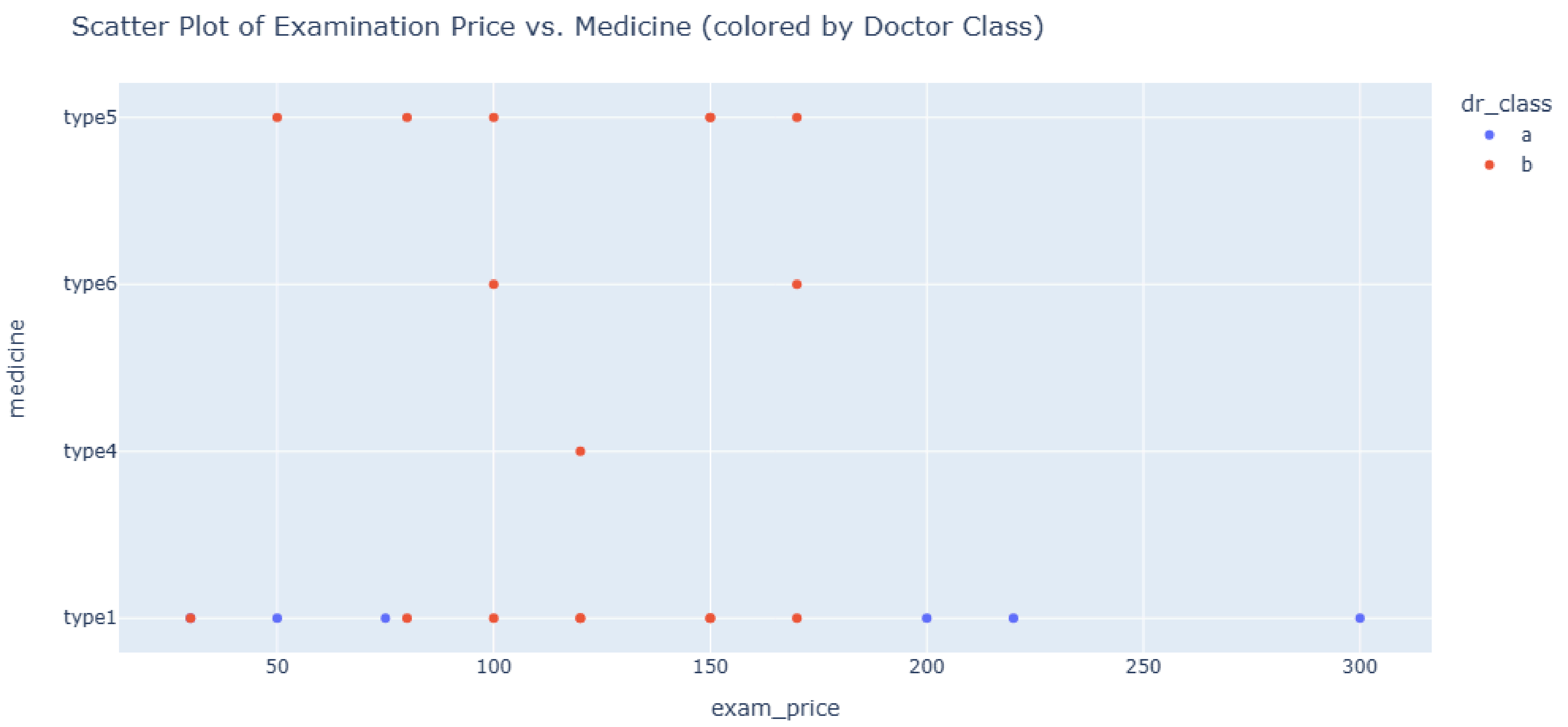


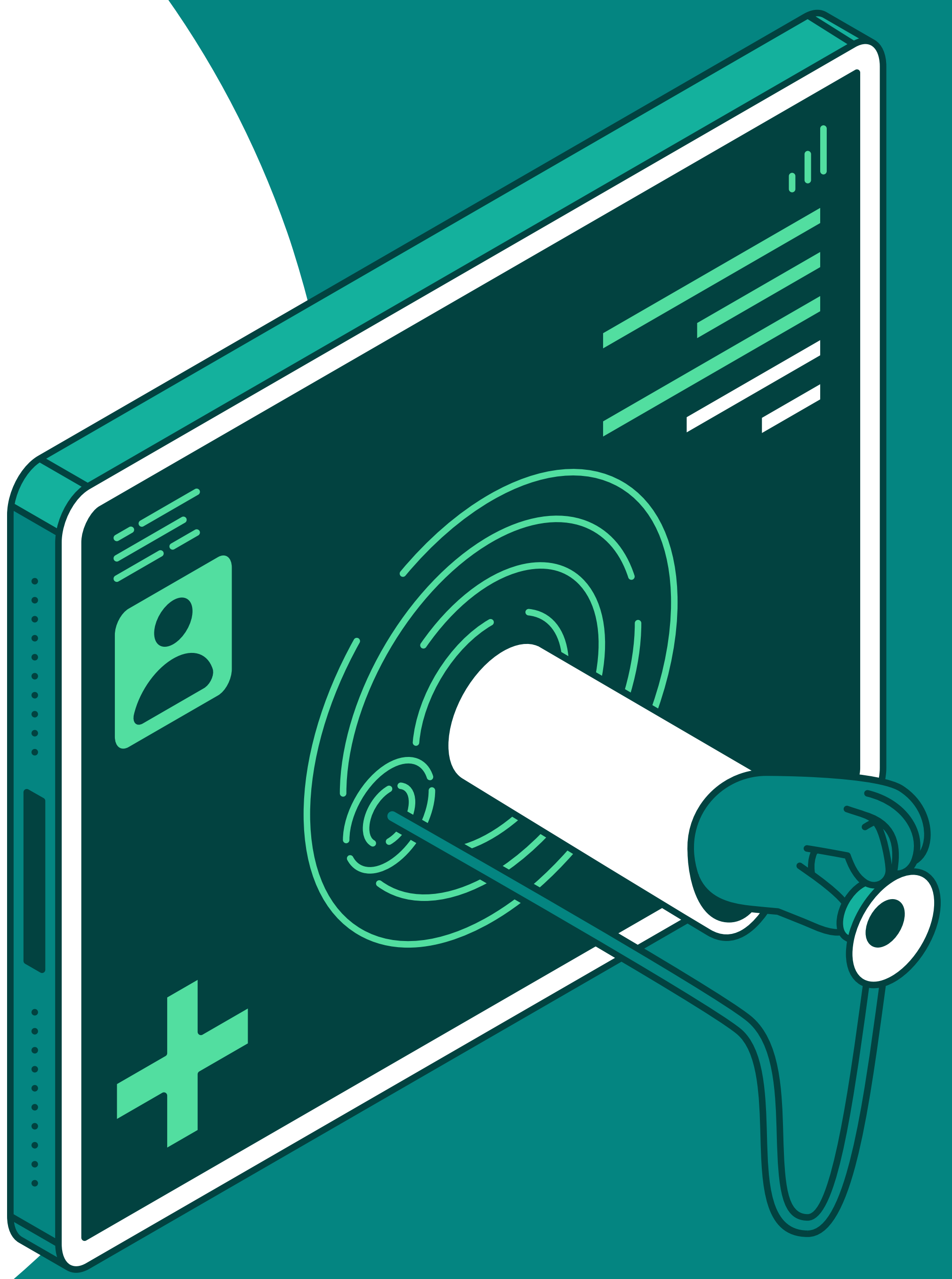




DATA ANALYSIS

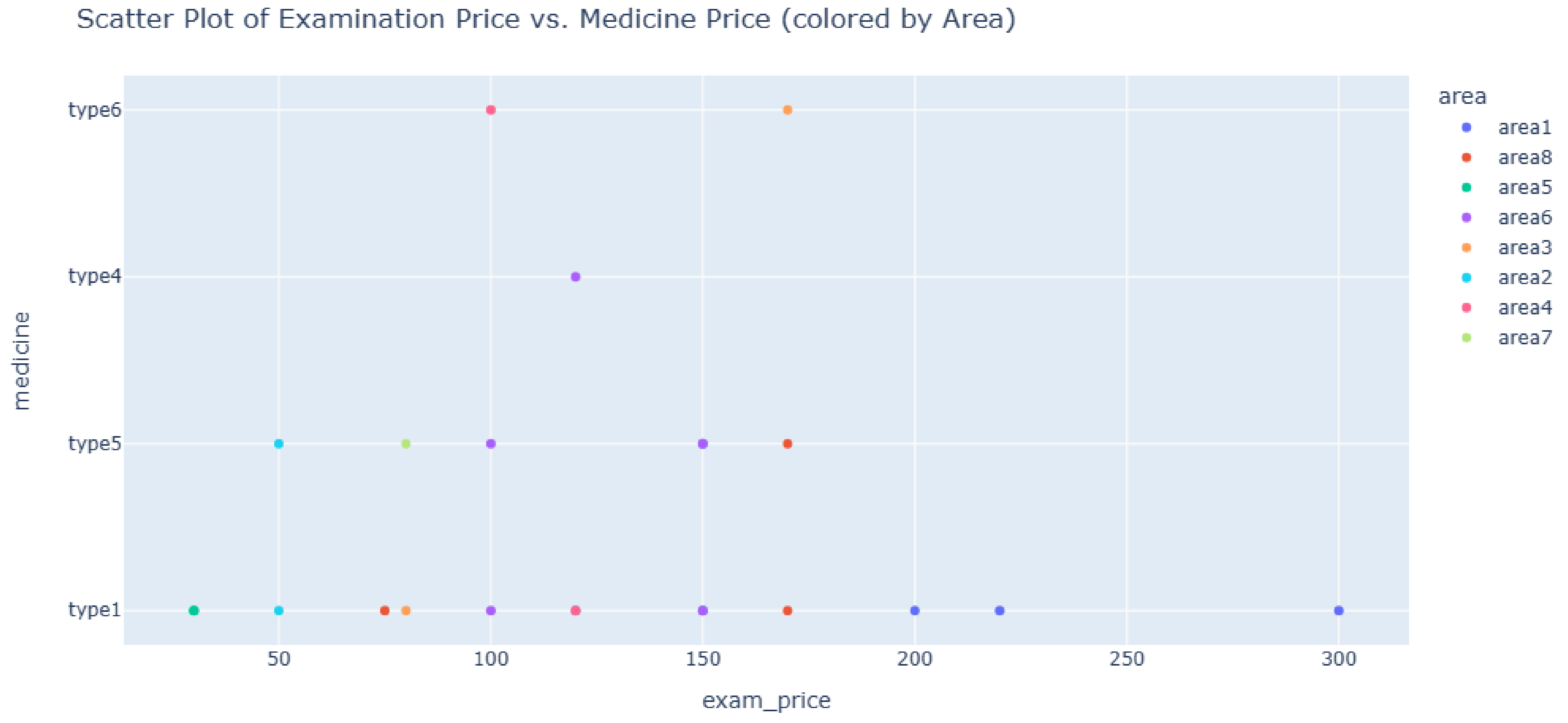
Class a write only type 1 in all exam price range hospitals or clinics
Class b write type 1 also the most but also other cheap types and in low exam price range and more points than class a





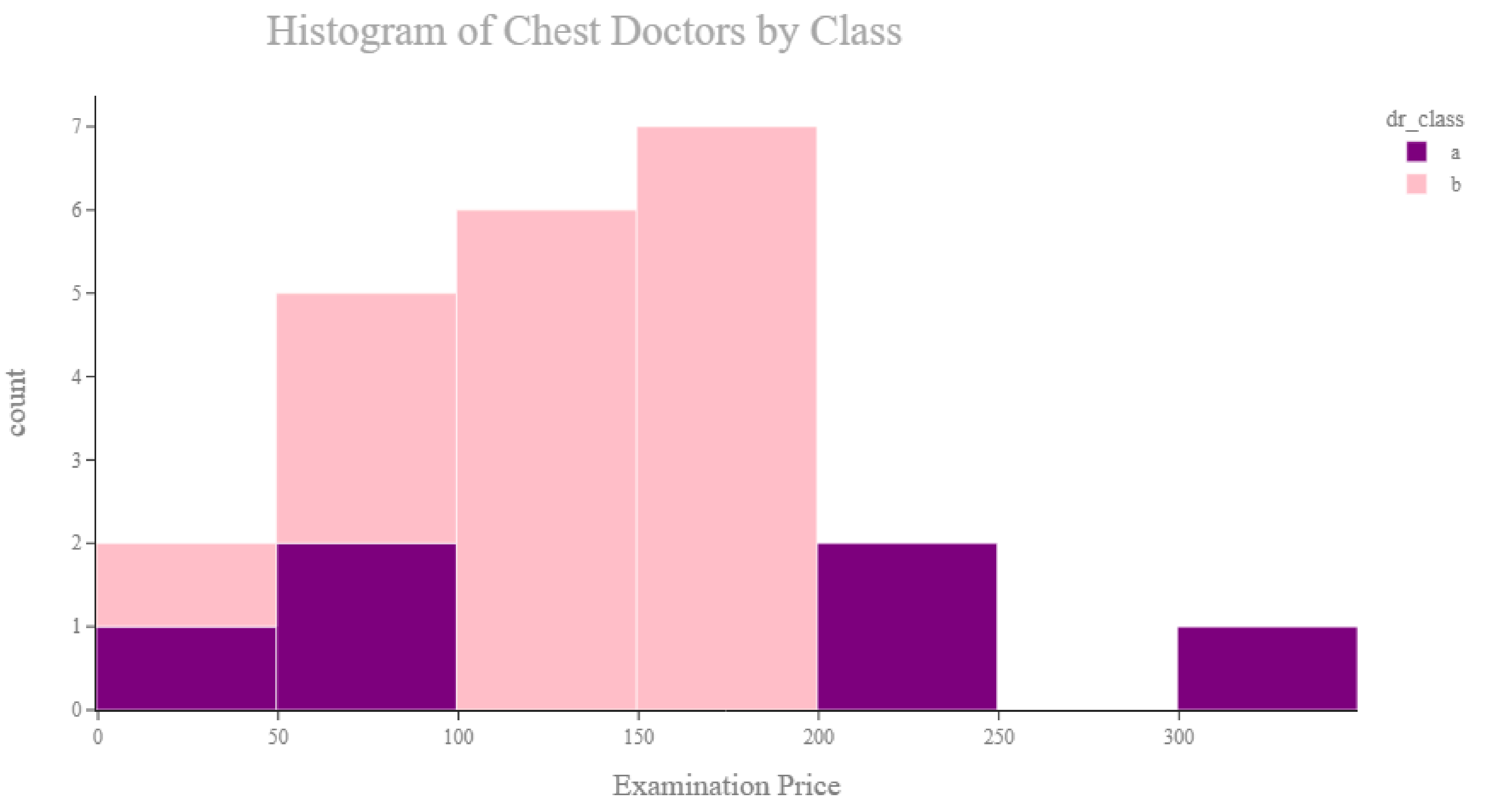
DATA ANALYSIS

Class a with the highest exam price in area 1 that because area 1 is the highest exam price avg and in area 2 , 8 in hospitals with low exam price
Class b in many areas with low range of exam prices



DATA ANALYSIS

Class a is in high range clinics and low range in hospitals
Class b is more than a and in low ranges

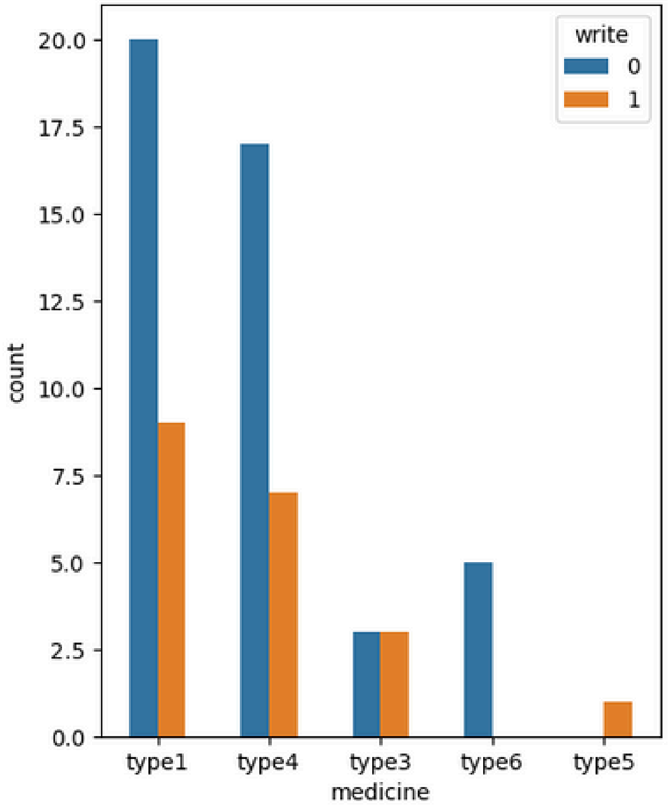


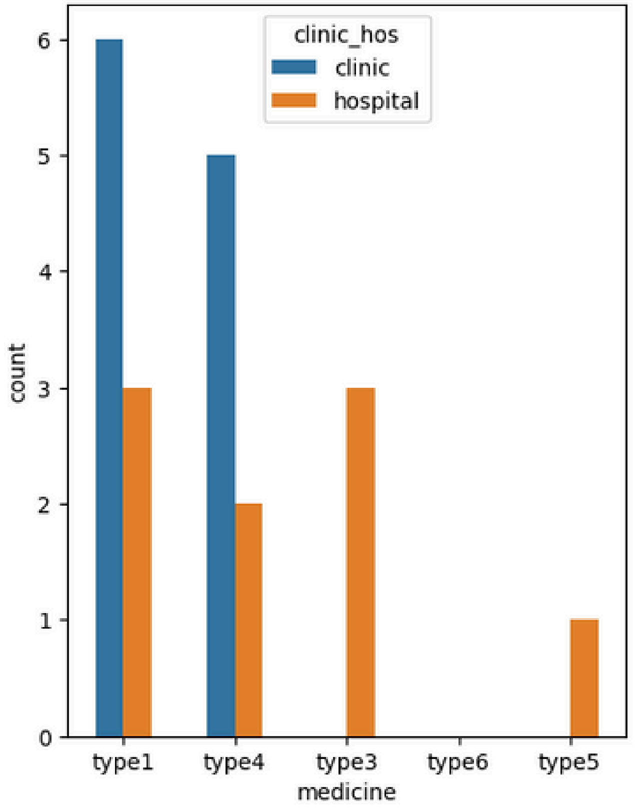
DATA ANALYSIS

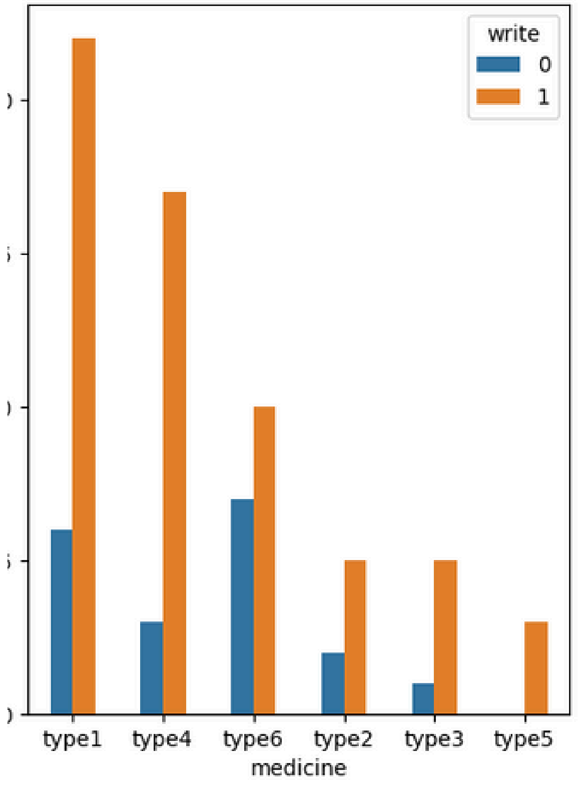
Im Doctors

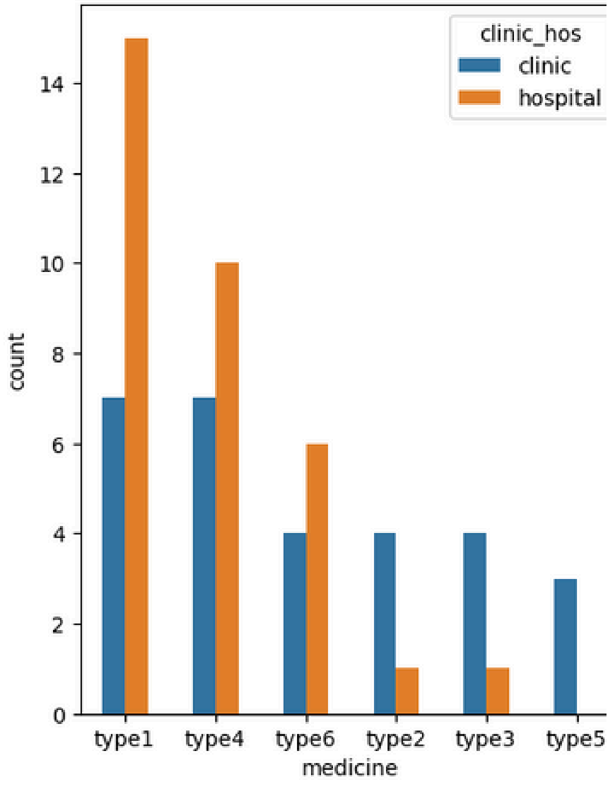
70% Im doctors in Class a write did not write
Another 30% most write Type 1 and Type 4
Type 5 written one time in hospital and 1 , 4 most in clinics
Type 3 , 5 all in hospitals

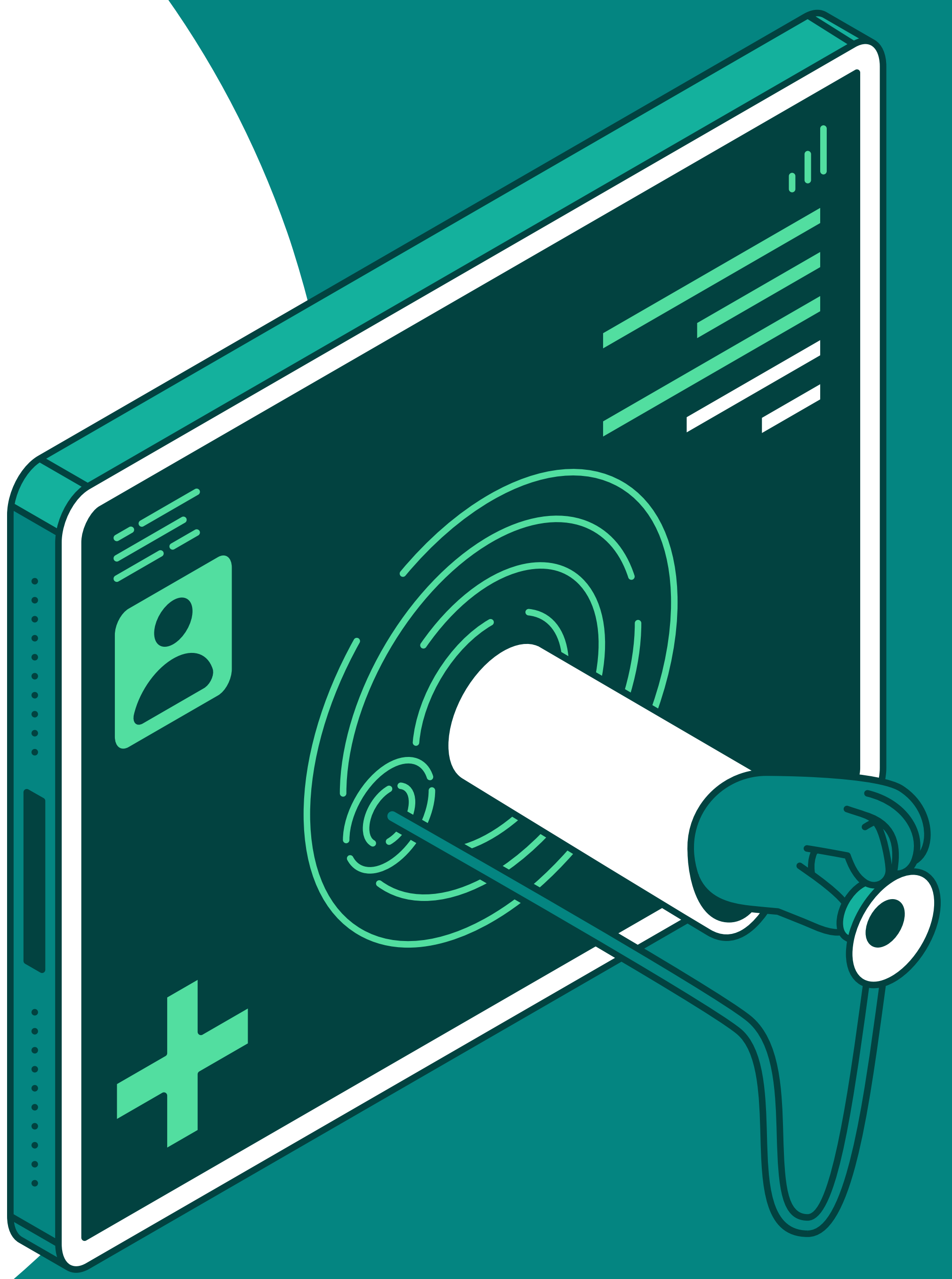
75% Im dcotors in Class b write
They write Type 1 , 4 most and also cheapest Type 6 , 2 ,3 , 5
They write Type 2 , 3 , 5 most in clinics and type 1 , 4 , 6 most in hospitals





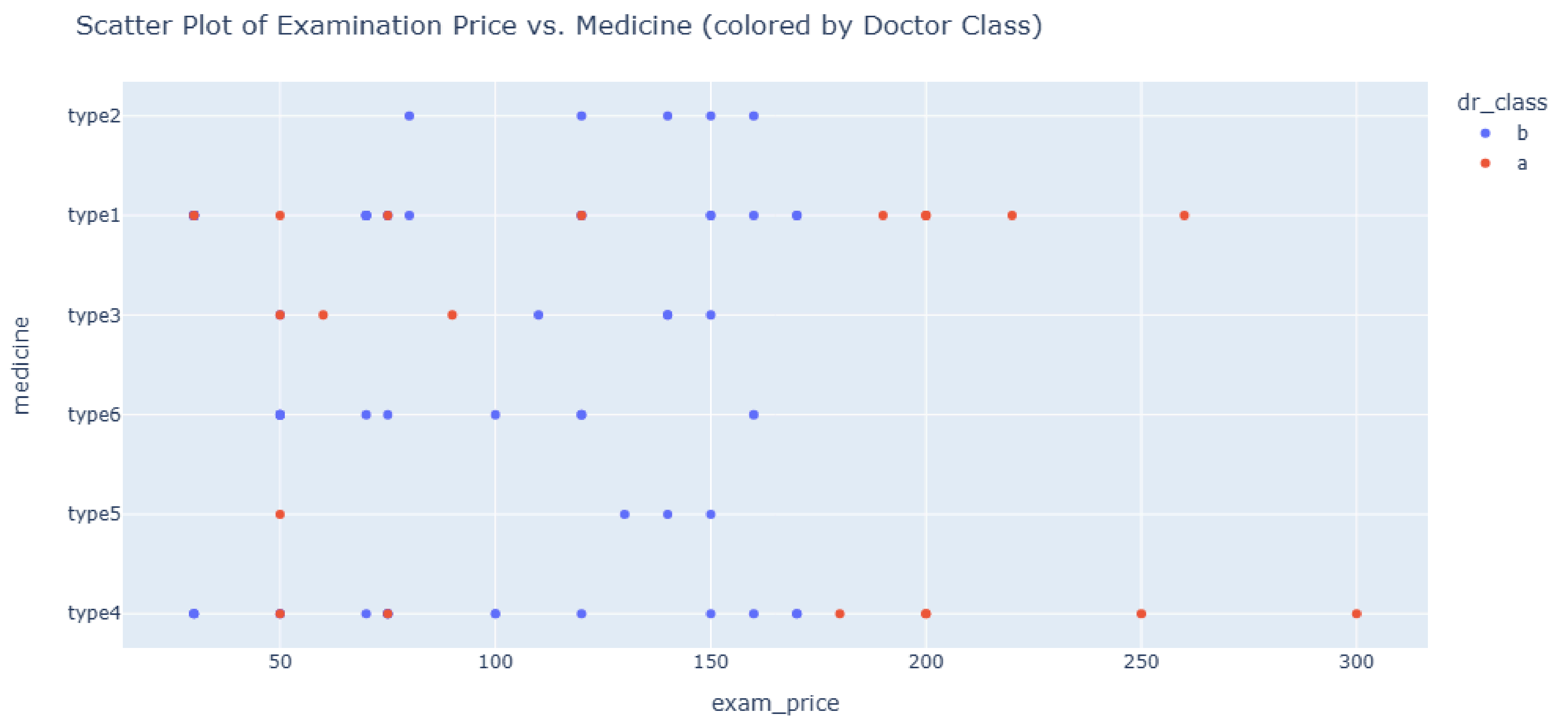






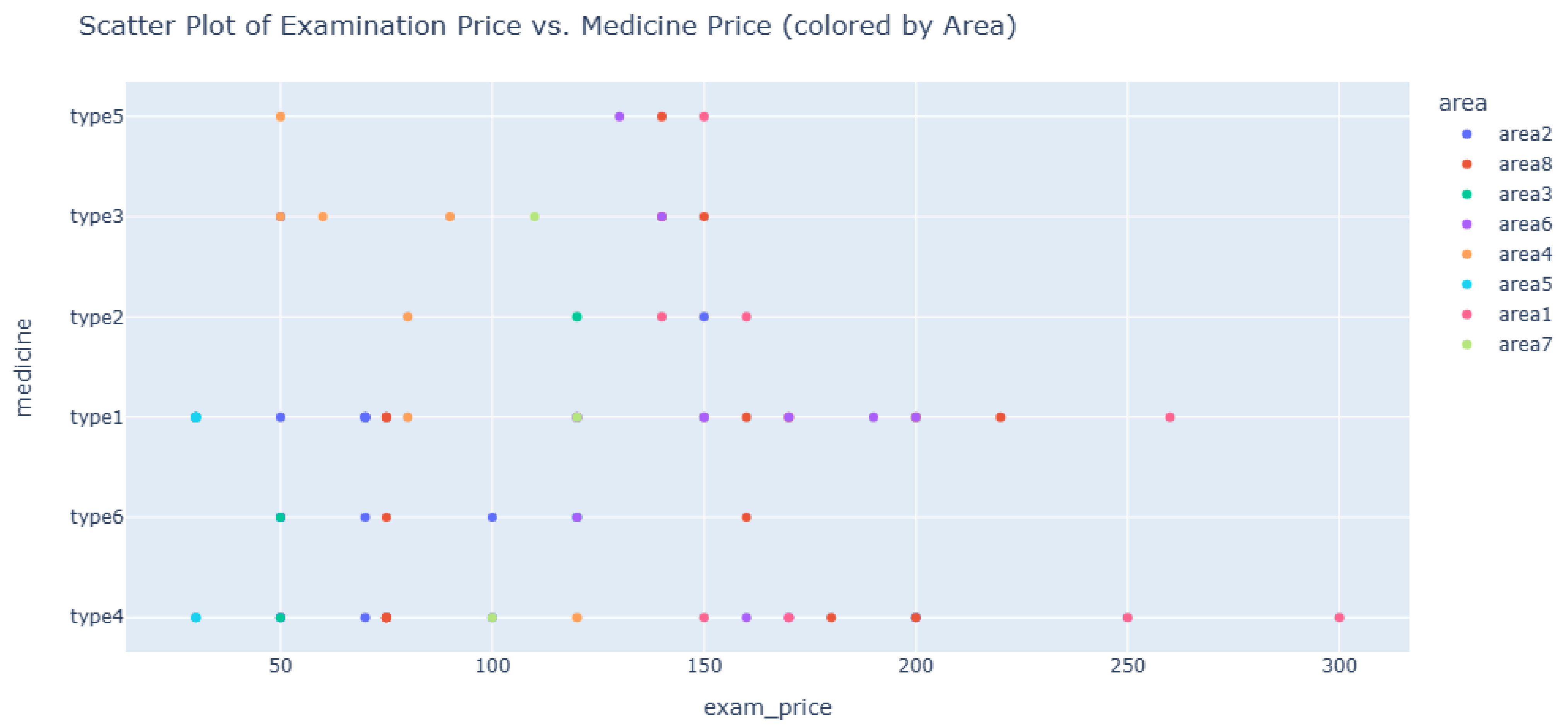
DATA ANALYSIS

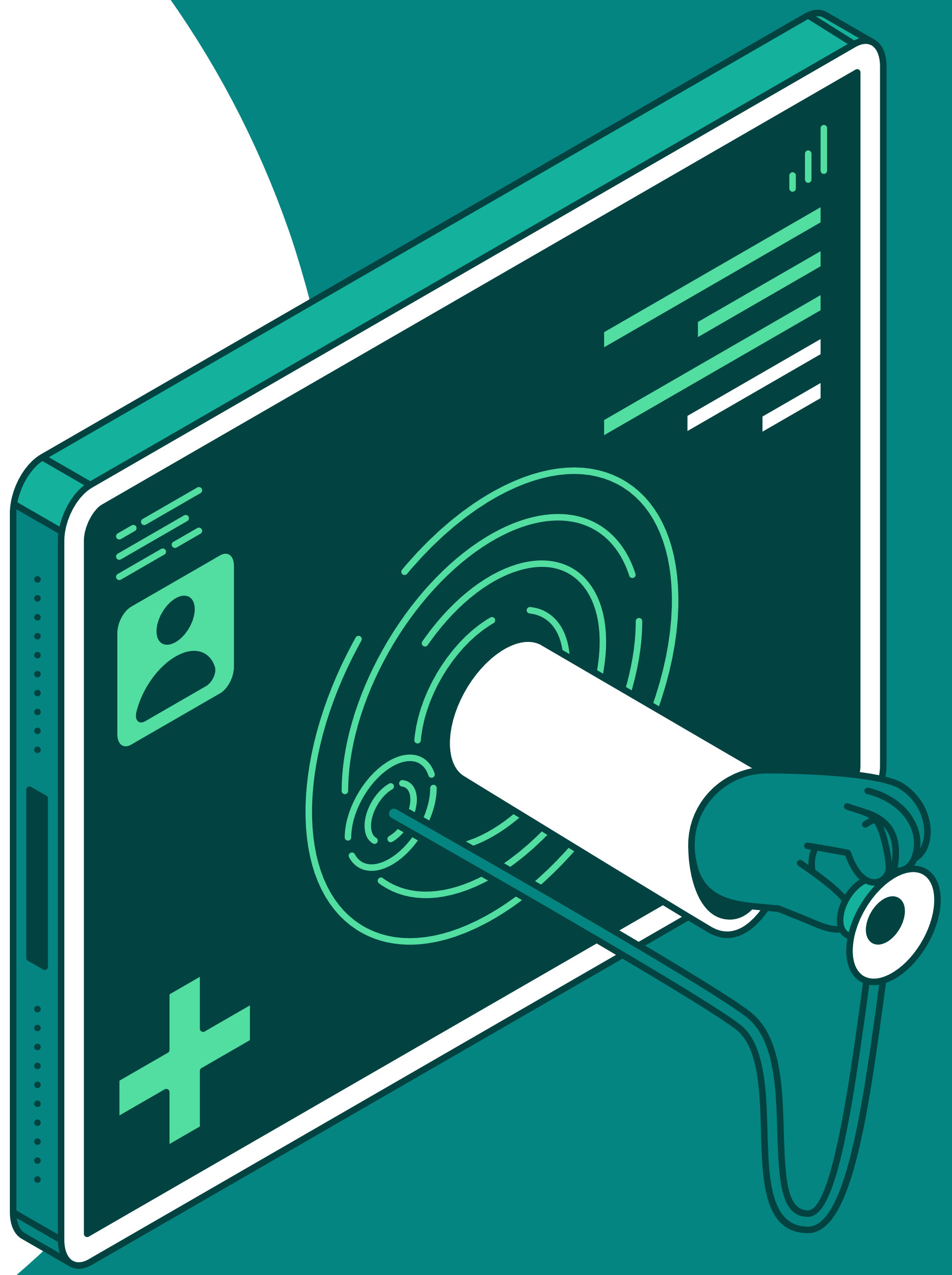
Class a with high ranges of exam price with type 1 , 4 in clinics
Class a with low ranges of exam price with type 1 , 4 and other cheap medicines in hospitals
Class be is more than class a and in low ranges with variant in medicines



DATA ANALYSIS

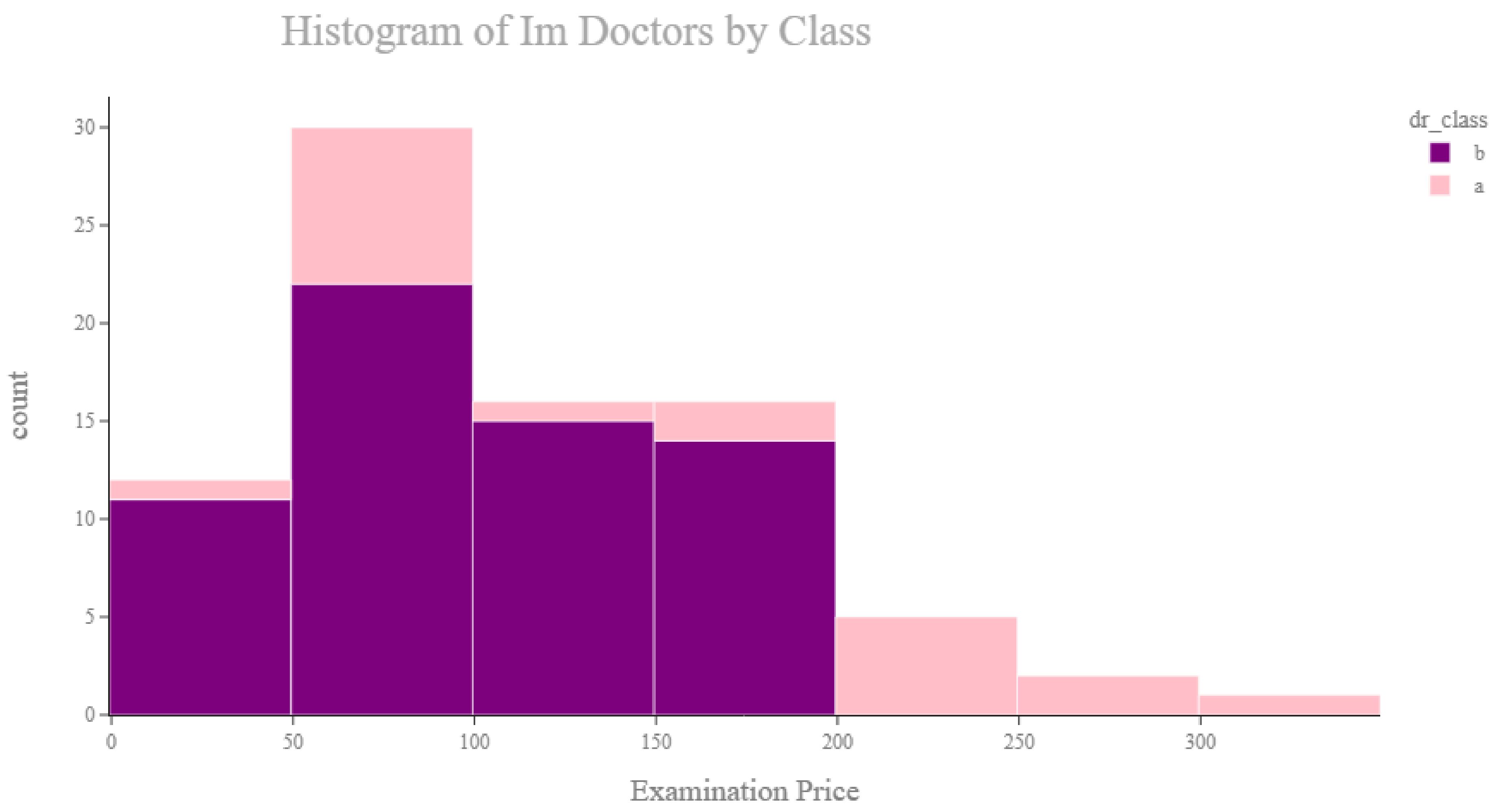
Area 1 the the most exam price and class 1
Area 5 , 2 the the lowest exam price





DATA ANALYSIS

The distribution between class a and b

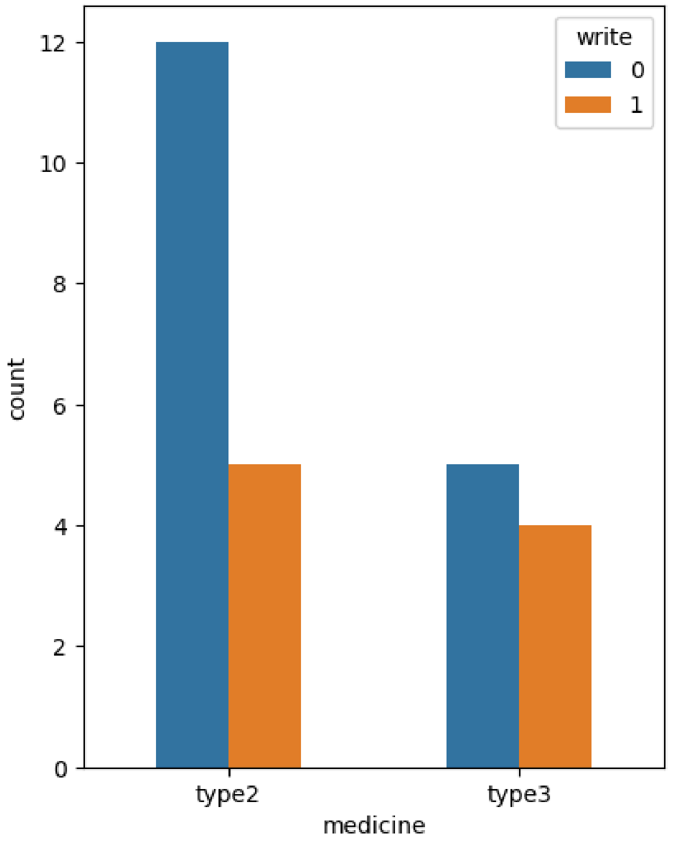


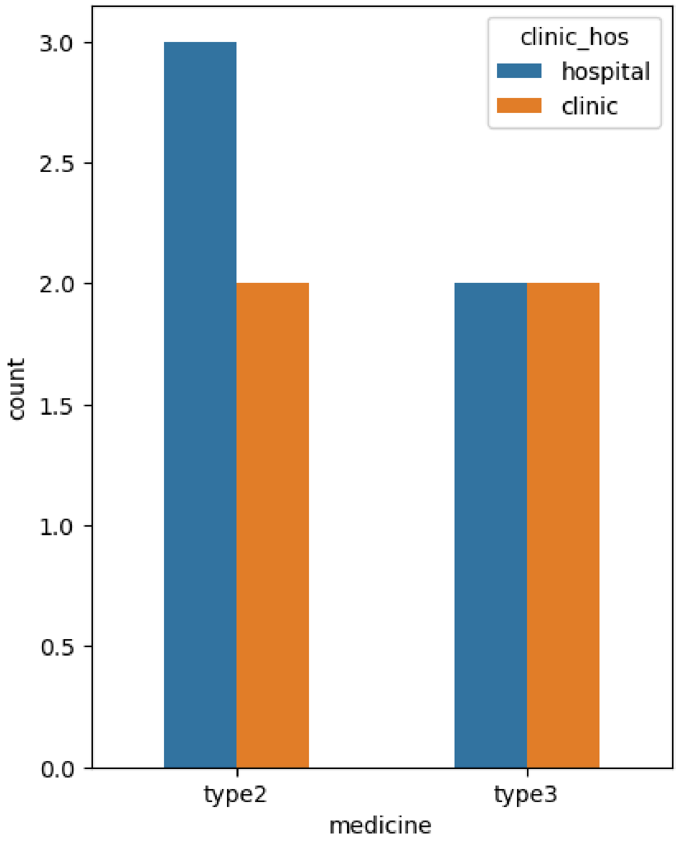
DATA ANALYSIS

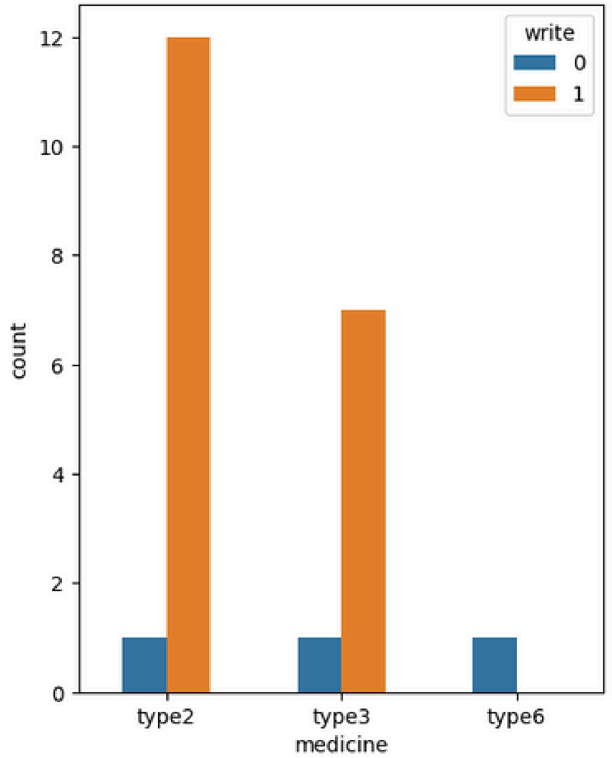
Cd Doctors

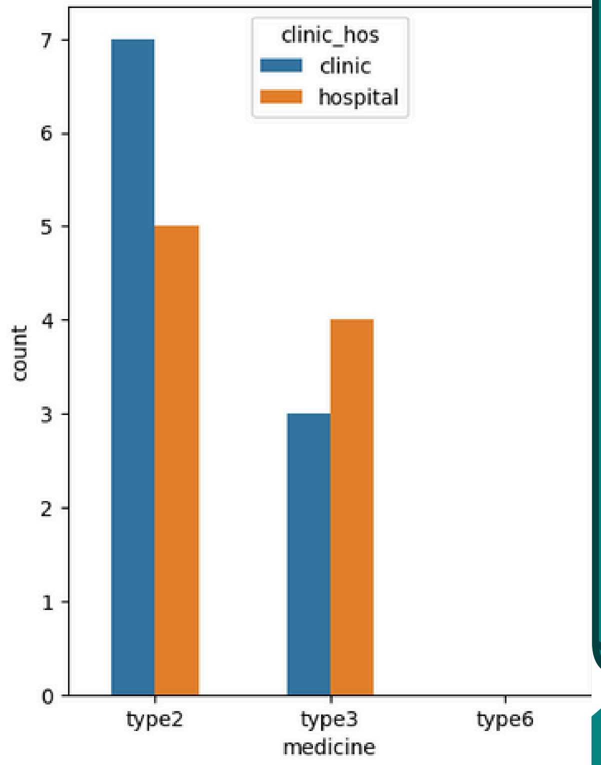
65% cd dcotors in Class a did not write
Another 35% most write Type 3 and Type 2 but the Type 3 is higher as percentage
Type 2 written most in in hospitals and Type 3 50%

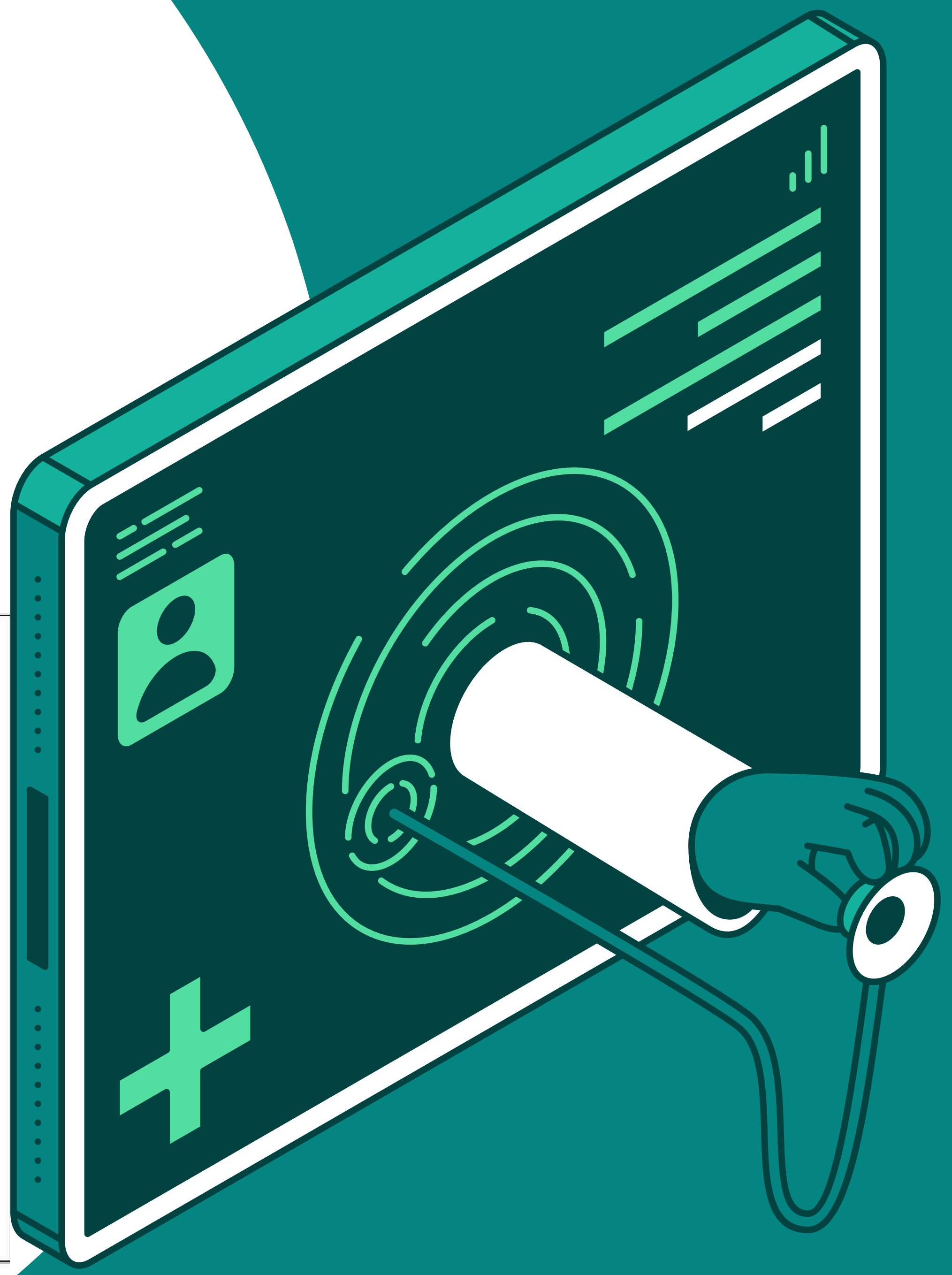
85% Im doctors in Class b write
They write type 2 most and 3
They write Type 2 More in clinics
Type 3 most in hospitals





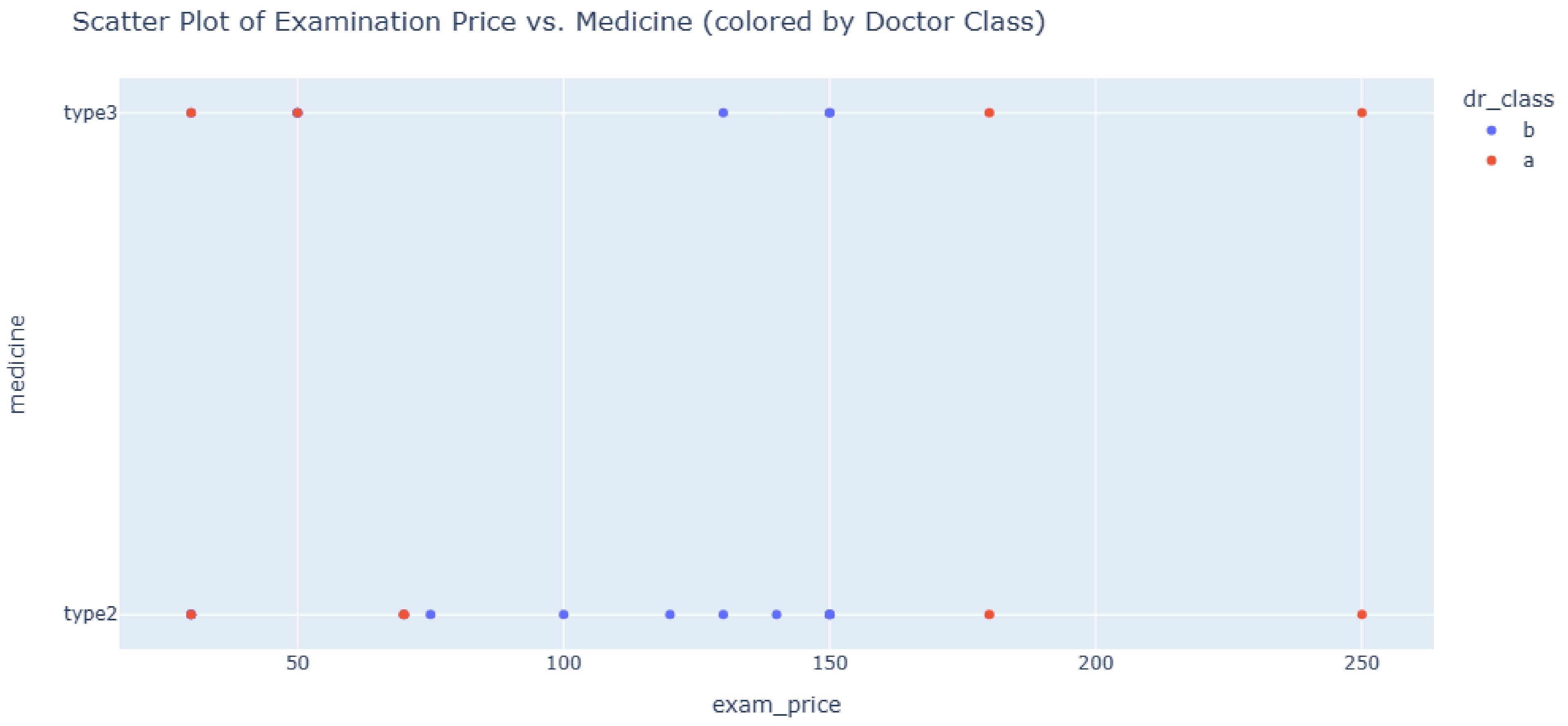






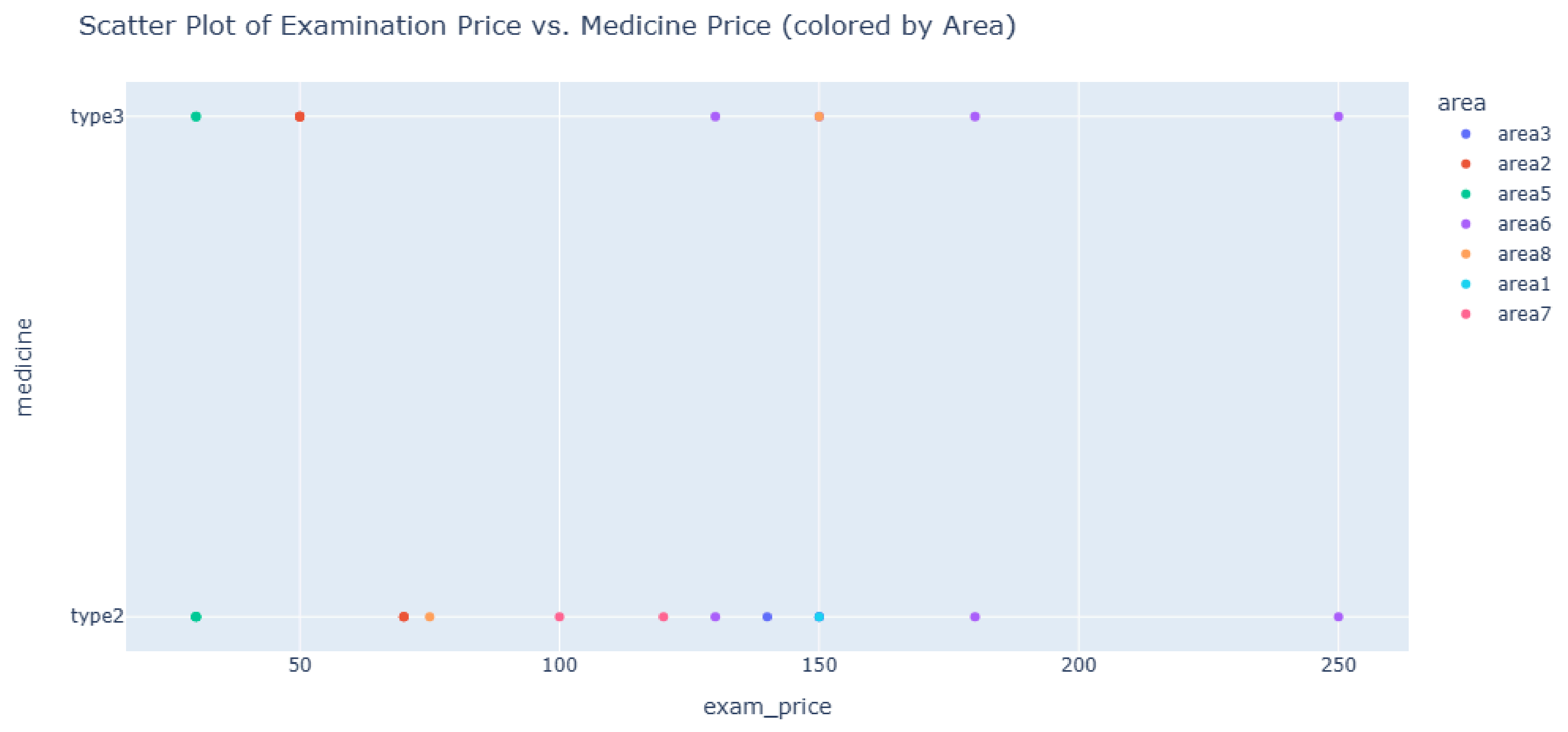
DATA ANALYSIS

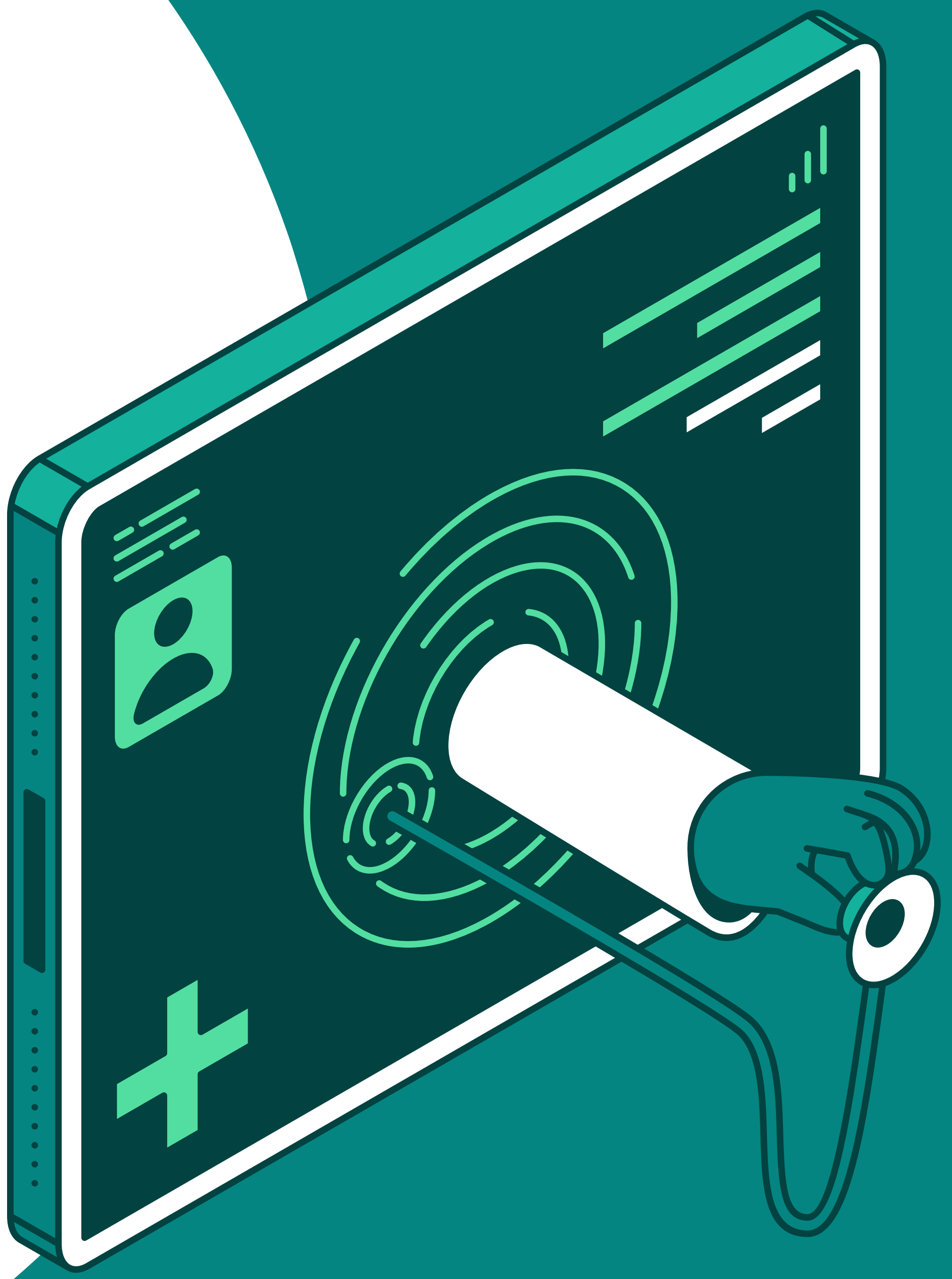
Class with high range of exam price is in clinics
Class with low range of exam price is in hospitals
Class b in low exam price range



DATA ANALYSIS

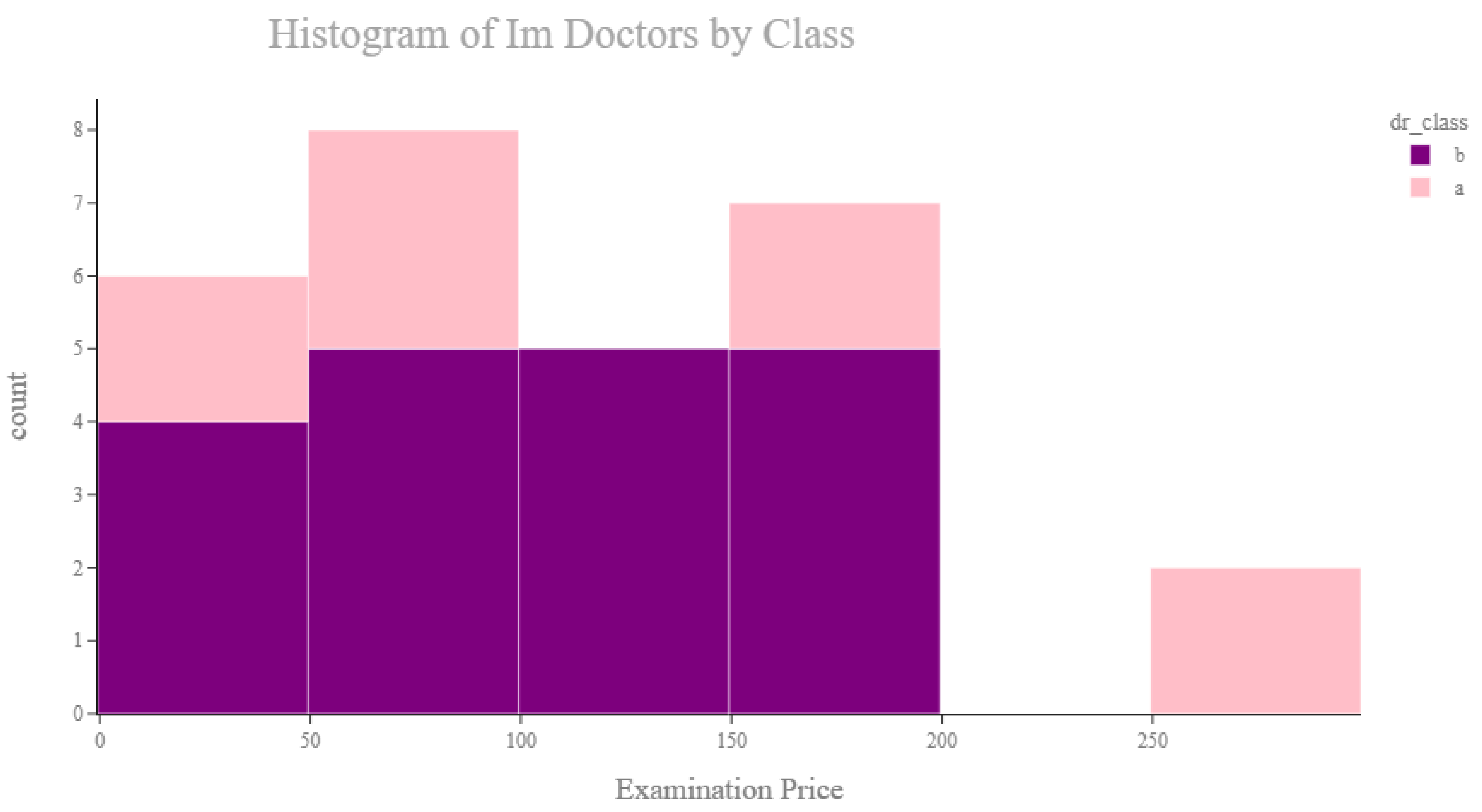
The highest exam price range area is area 6 in Class a
The lowest exam price range area is area 5 in class a but hospitals





DATA ANALYSIS

The distribution of Classes

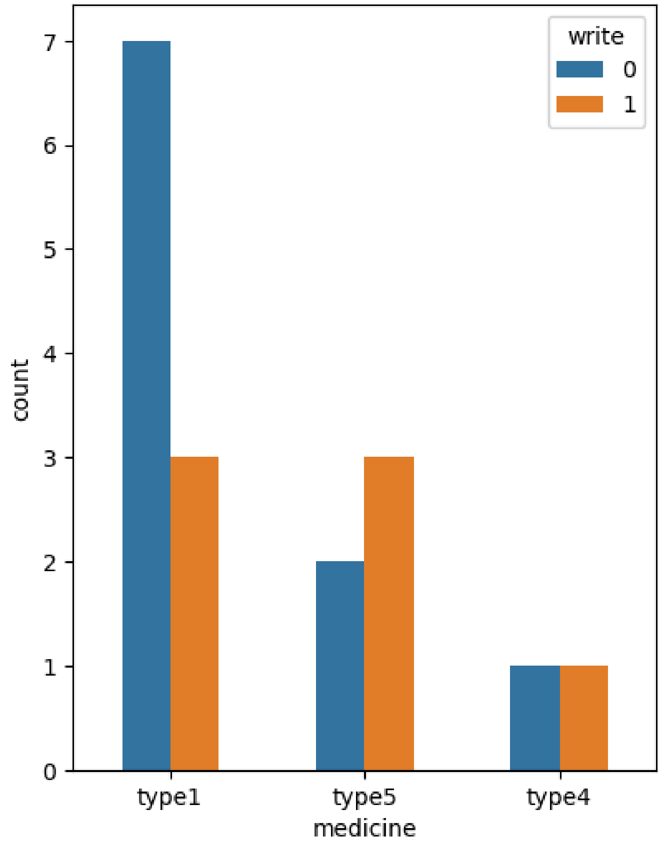


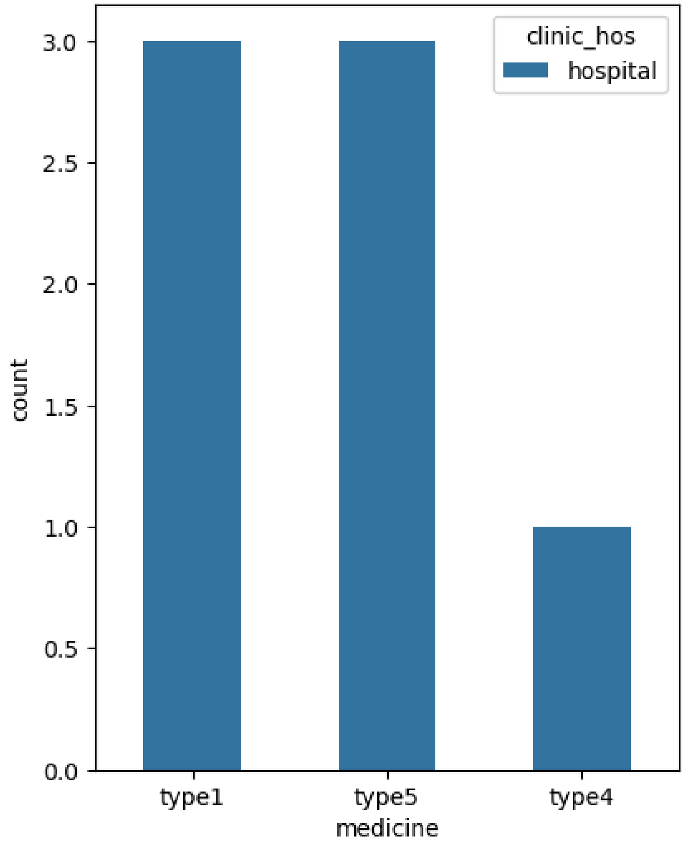
DATA ANALYSIS

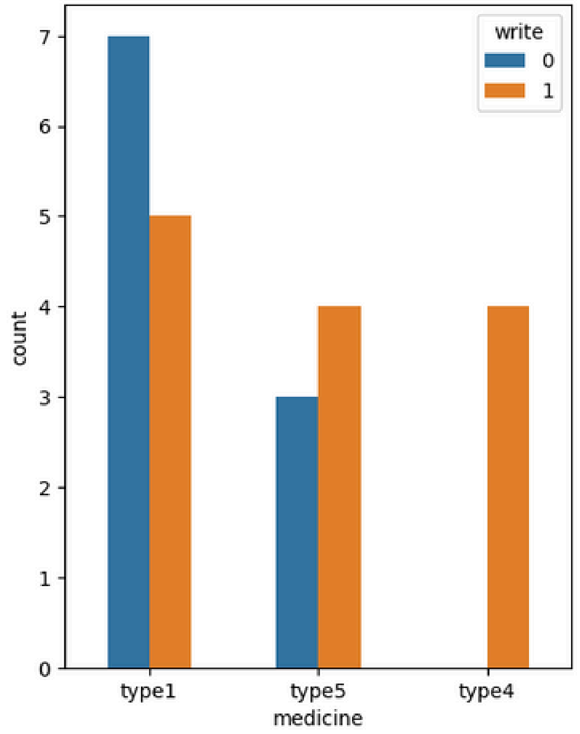
Surgery Doctors

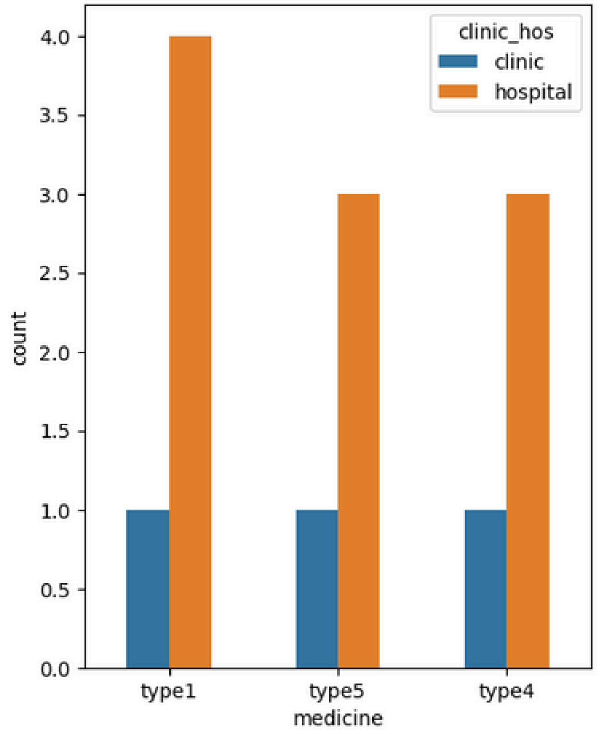
60% sur doctors in Class a did not write
Another 40% most write Type 5 is higher as percentage
All Sur Doctors in class a write in hospitals

55% sur doctors in Class b write
They write Type 4 then 5 then 1
The most in hospitals







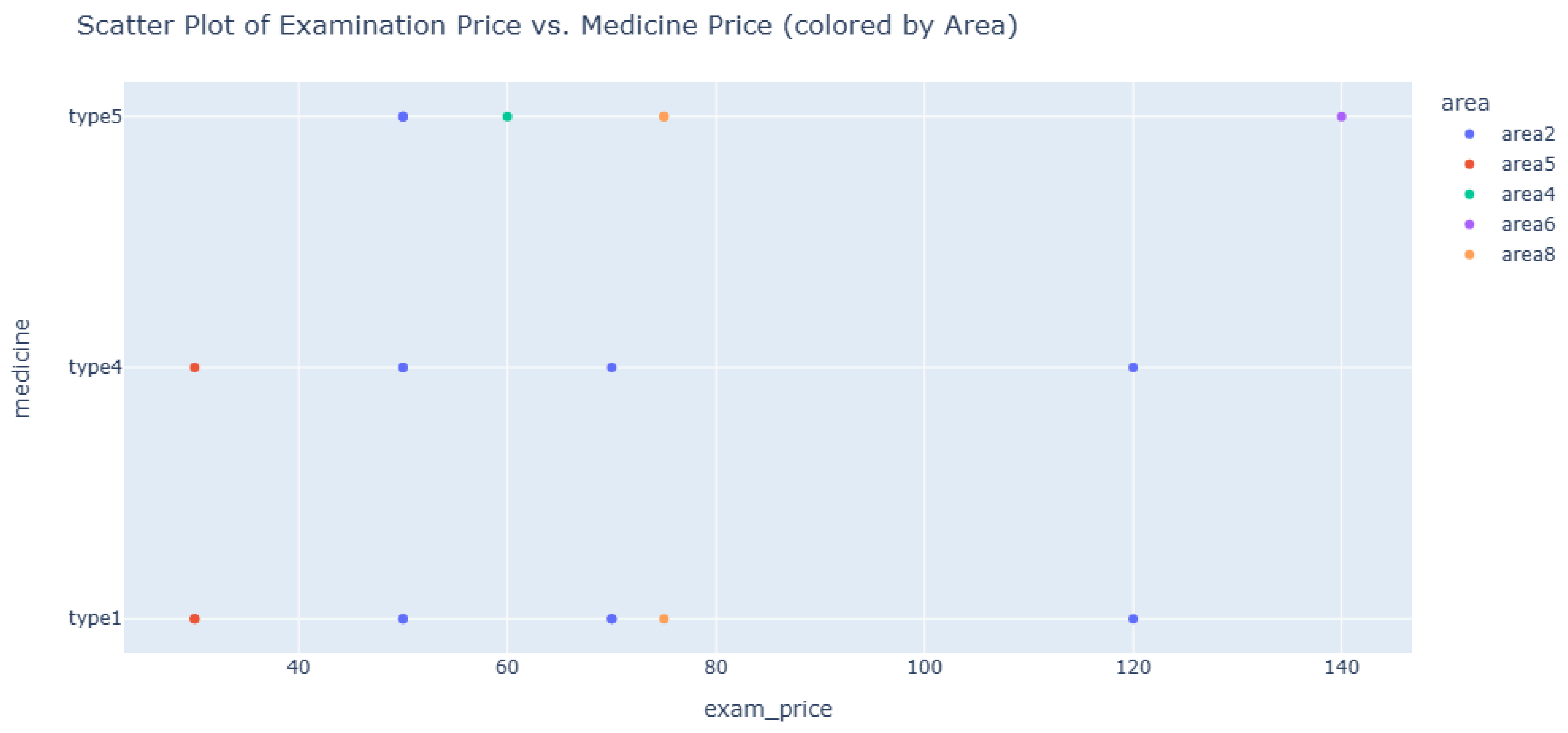


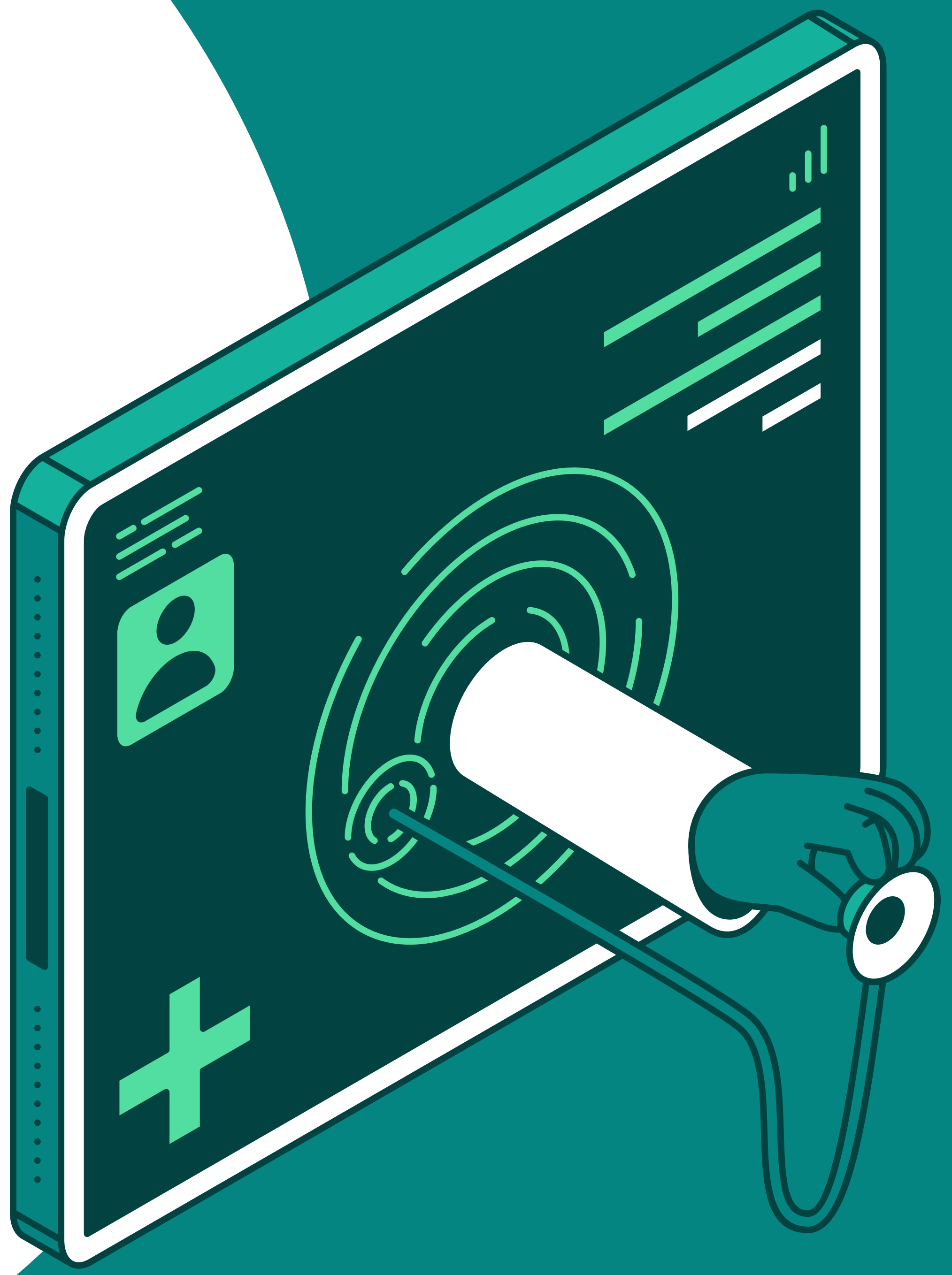
DATA ANALYSIS

Most points in low ranges because they most in hospitals just two points in high ranges with class b in clinics

DATA ANALYSIS

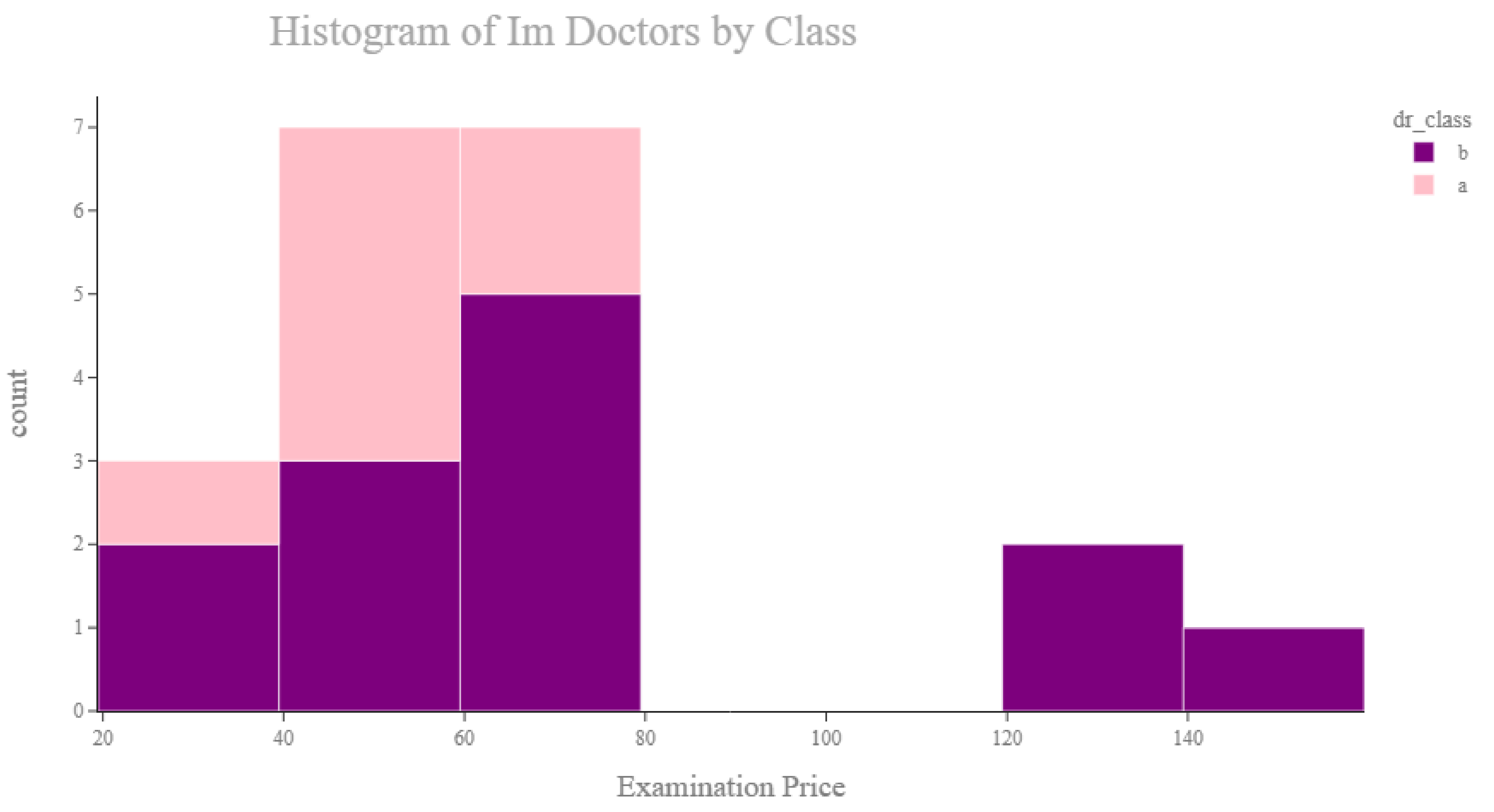
Most in Area 2





DATA ANALYSIS

Class a just hospitals with low ranges and there is few class b clinics in high ranges

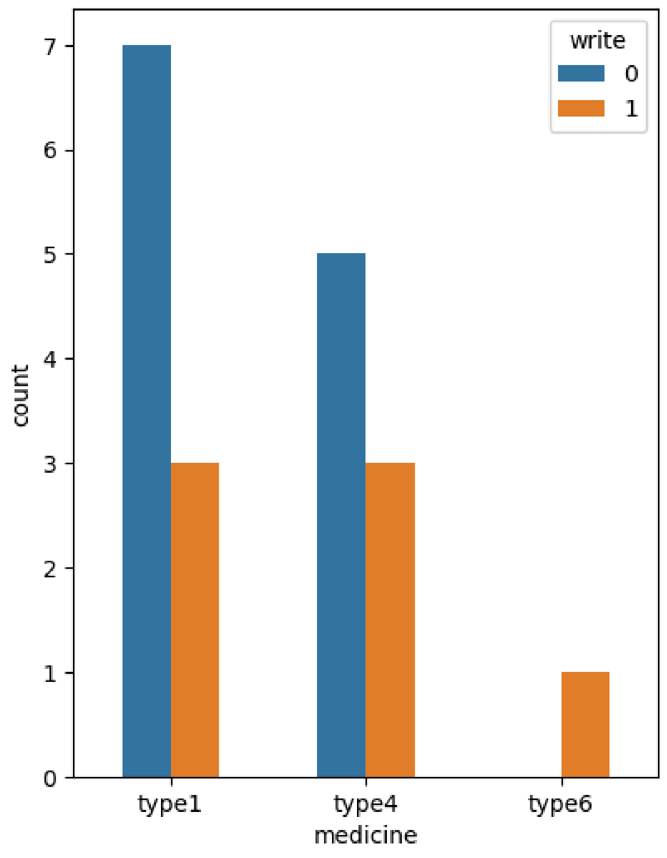


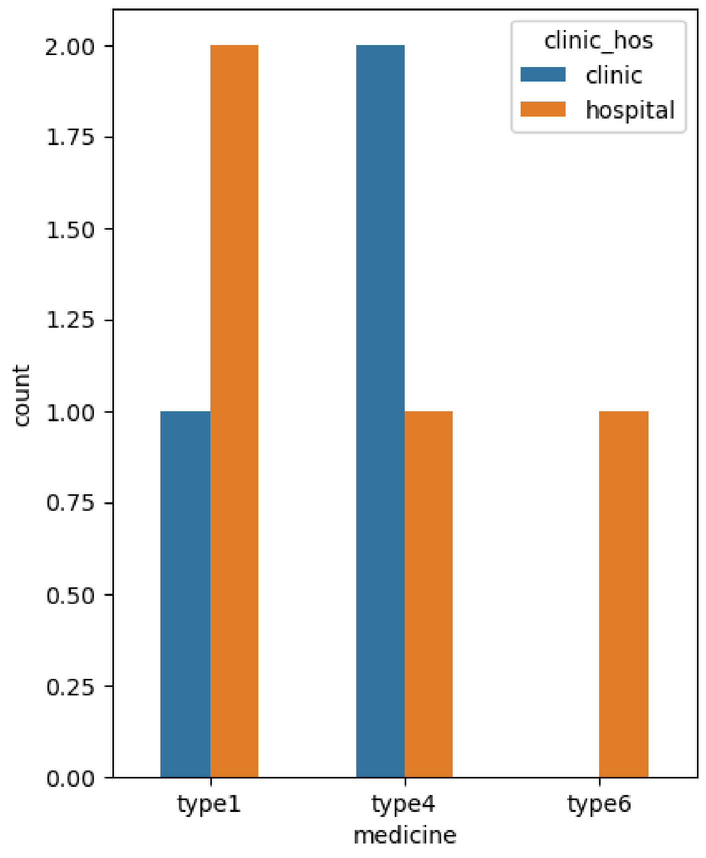
DATA ANALYSIS

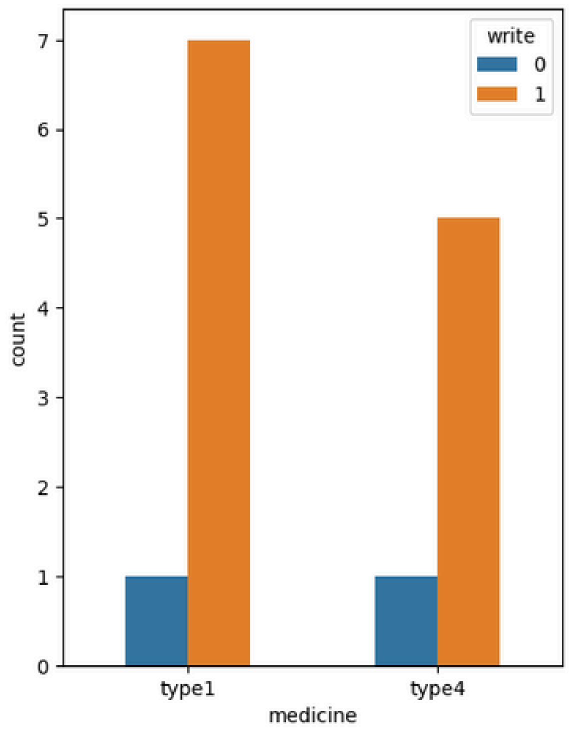
Git Doctors

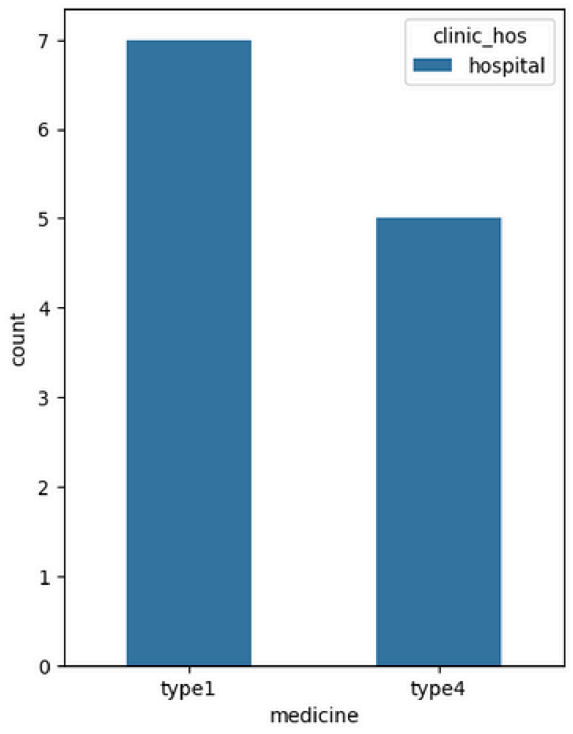
62% git dcotors in Class a did not write
Another 38% most write Type 4 and 1 , 6 just one
type 1 most in hospitals , type 4 most in clinics type 6 just in hospitals

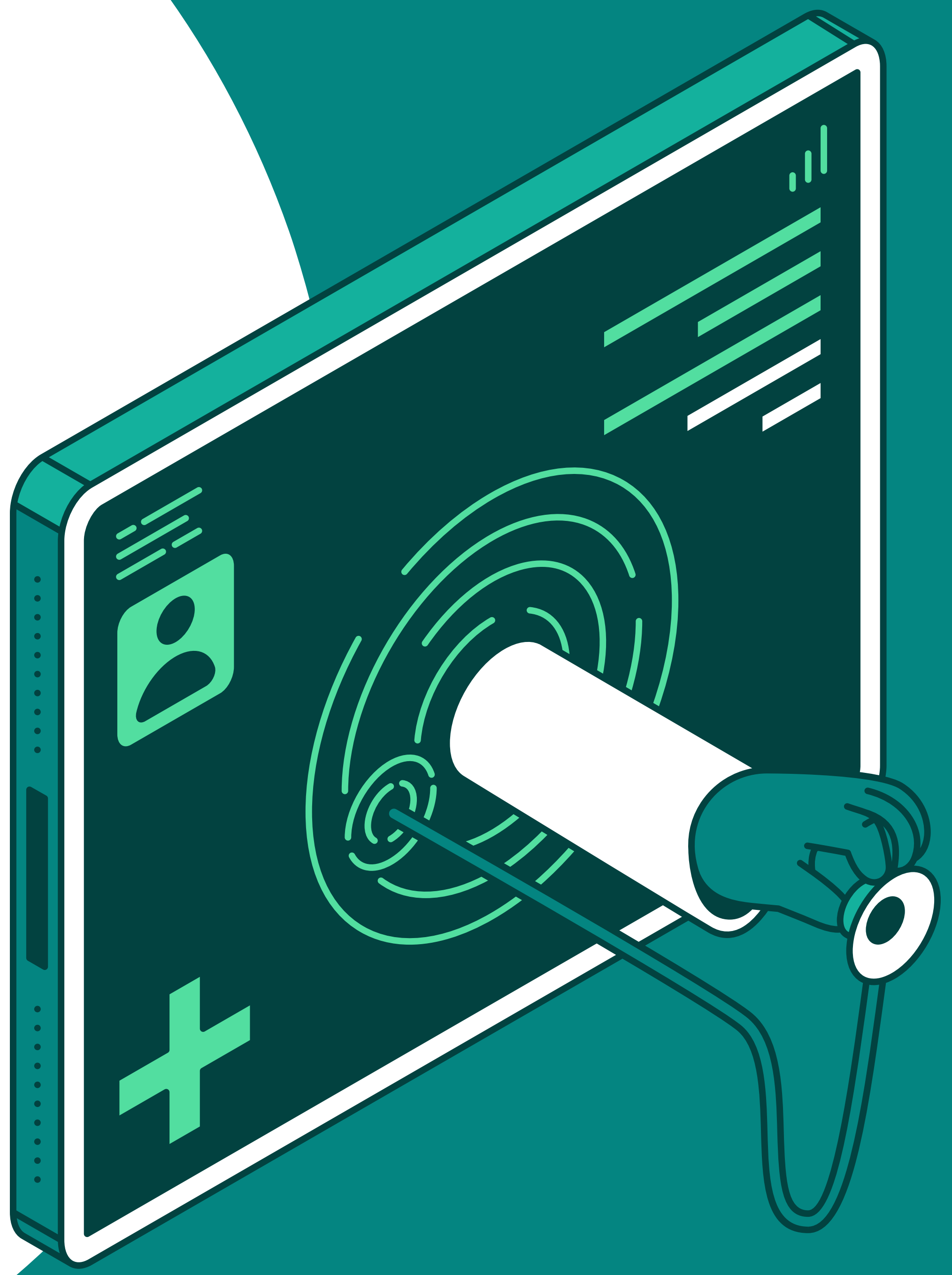
85% git doctors in Class b write
They write Type 1 most and 4
all in hospitals





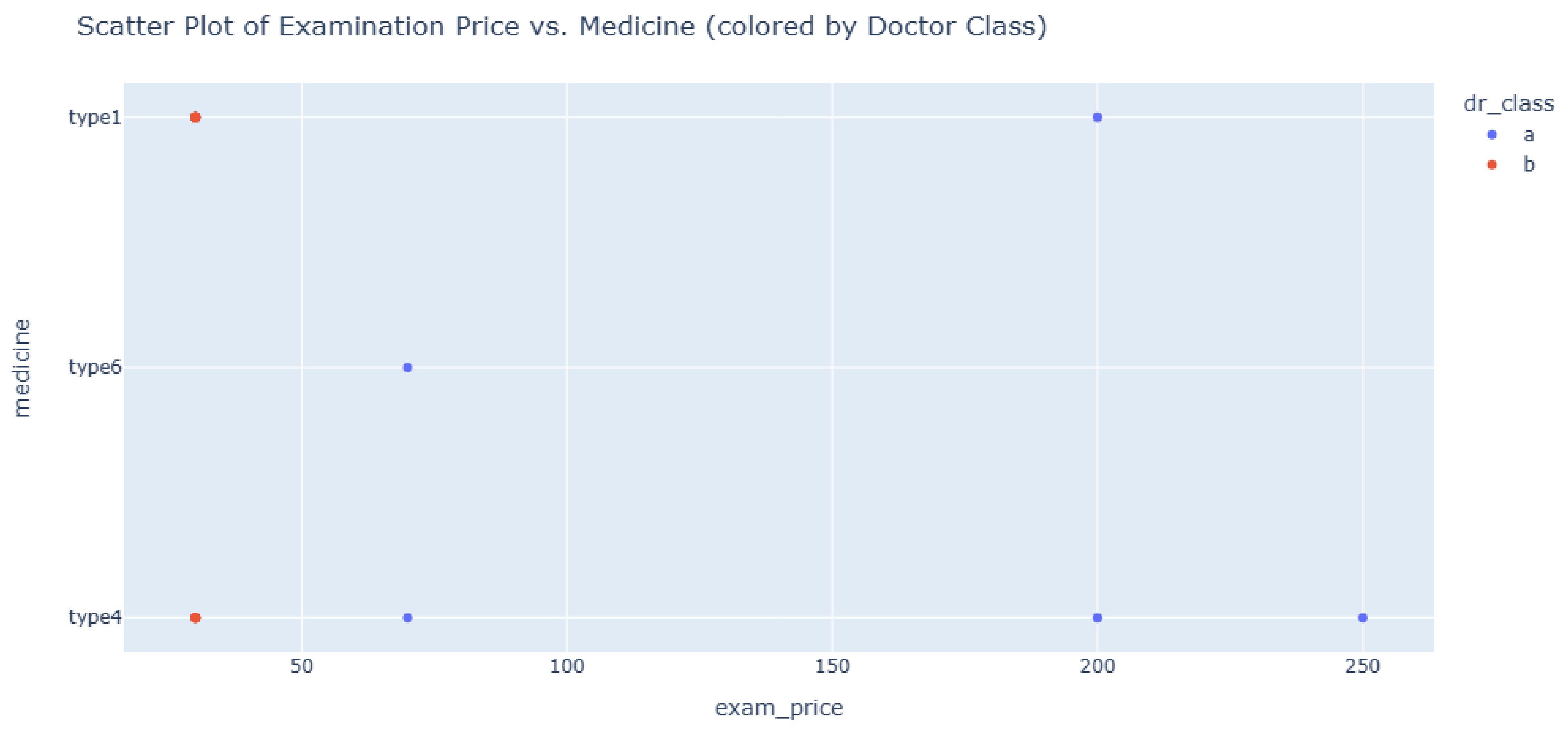


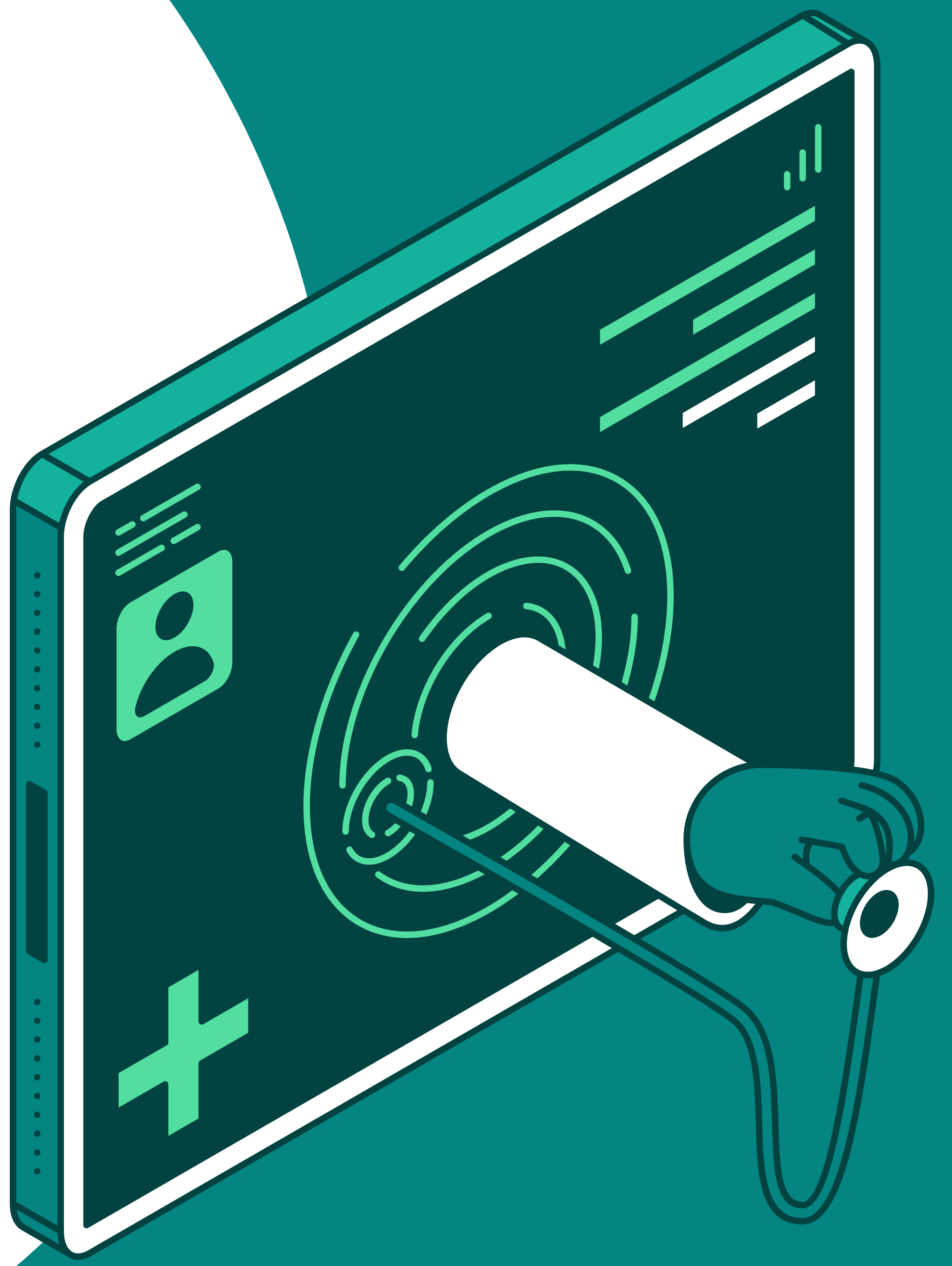




DATA ANALYSIS

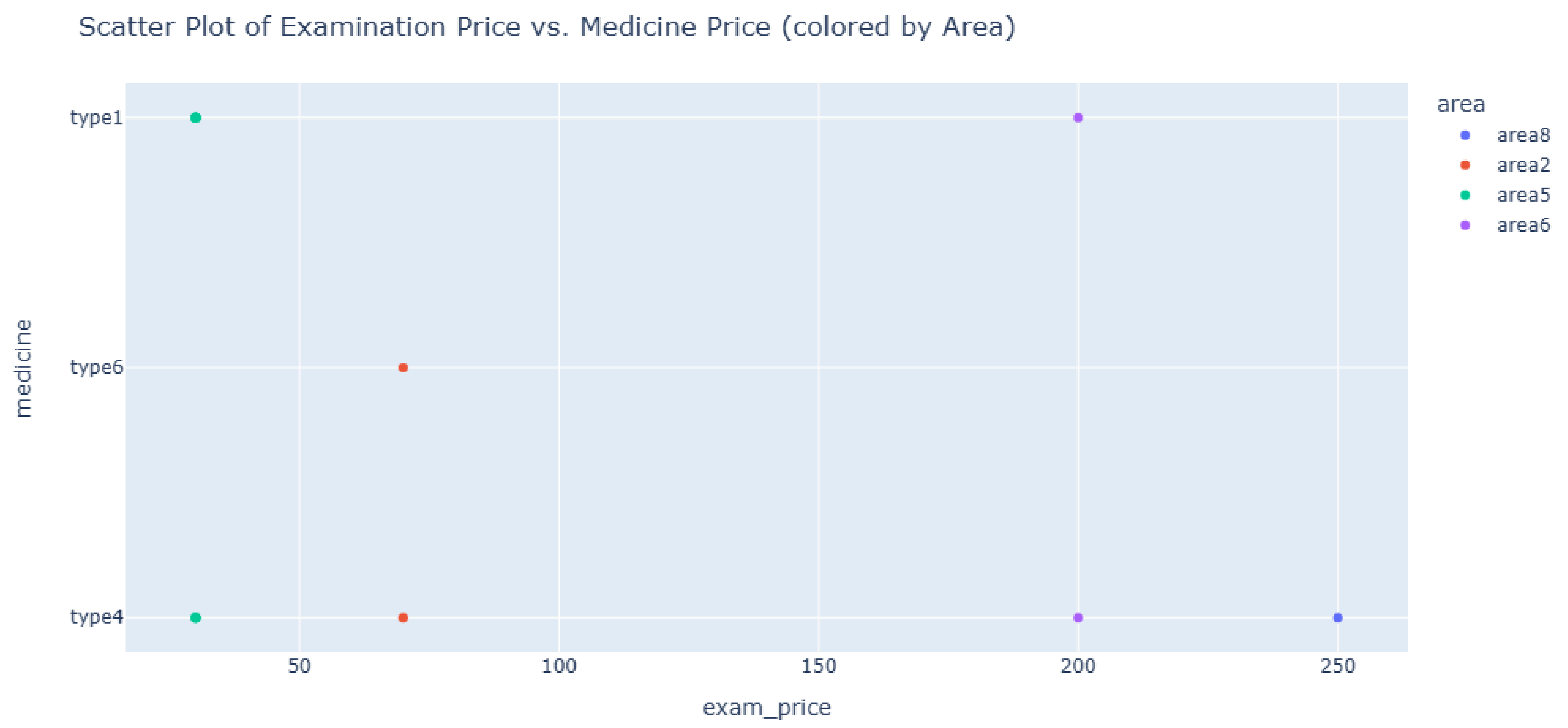
Class a in high ranges in clinics
Class b all in low ranges because of hospitals

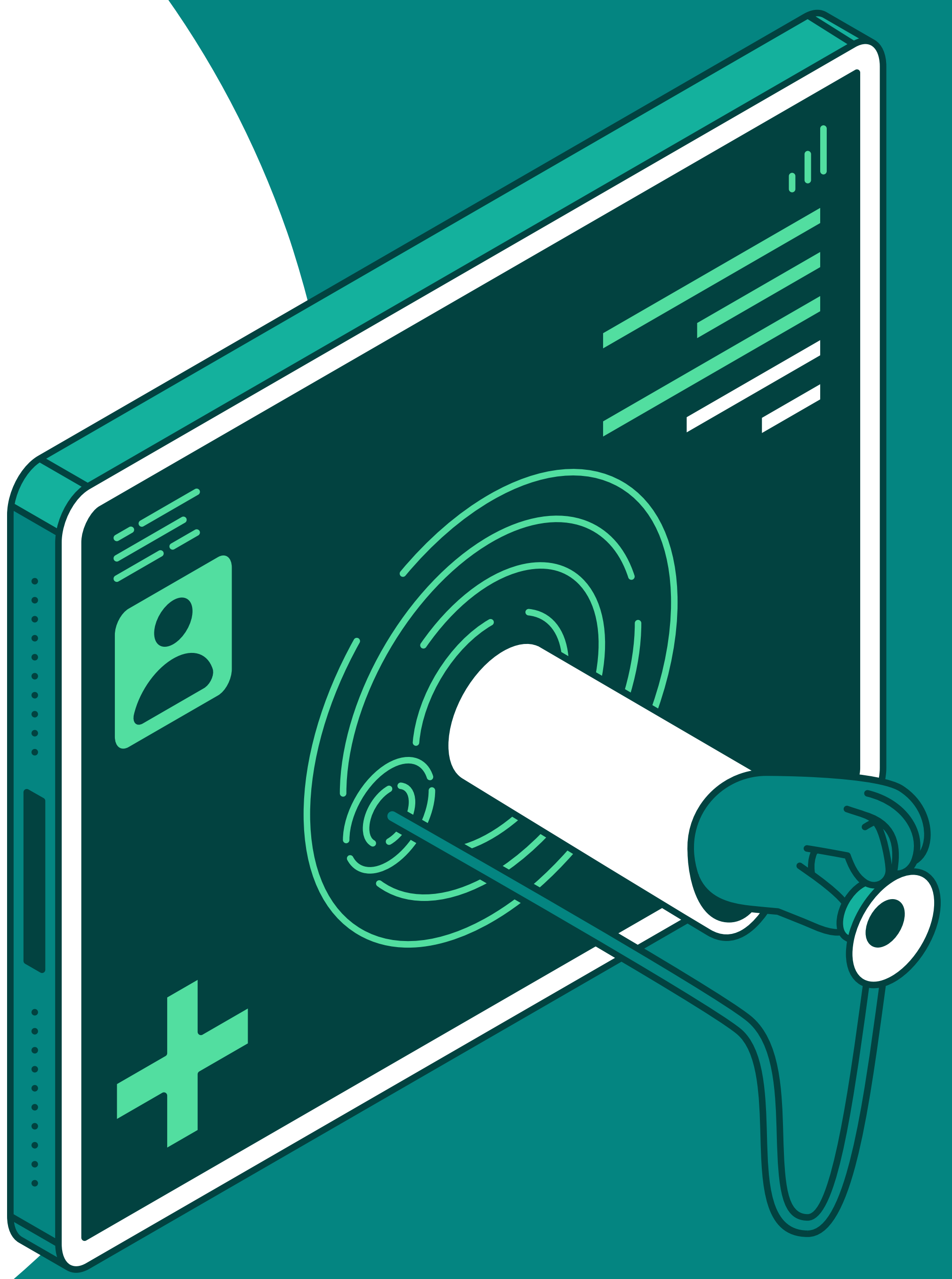




DATA ANALYSIS

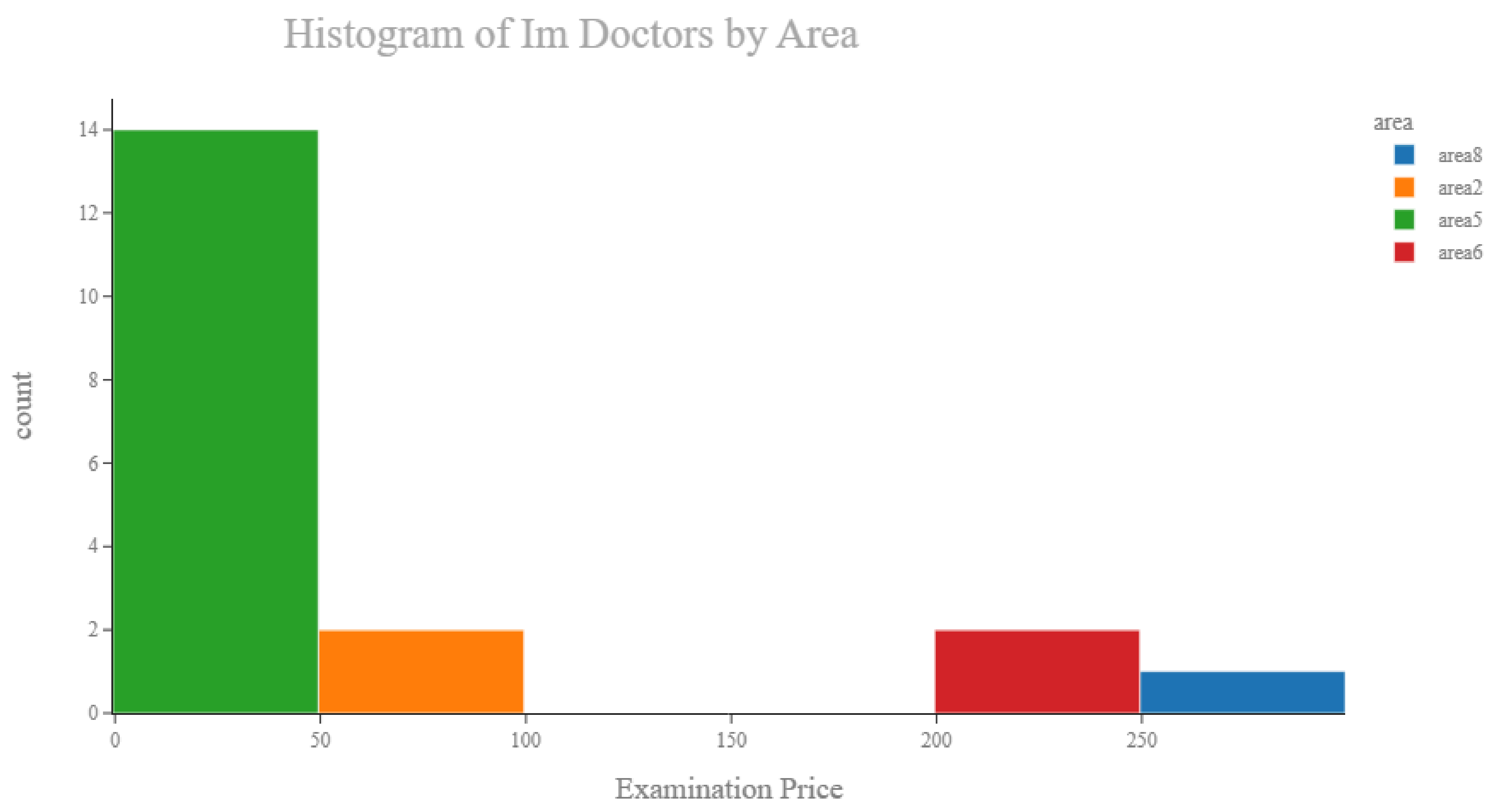
Low areas range 2 , 5 high is 6 , 8





DATA ANALYSIS

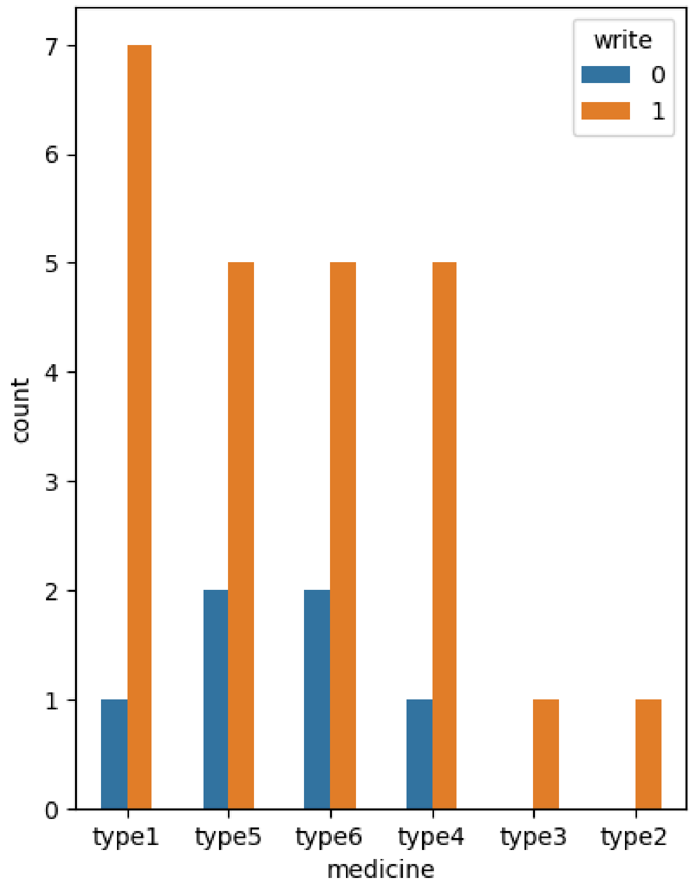
Areas Distribution most in low and area 5

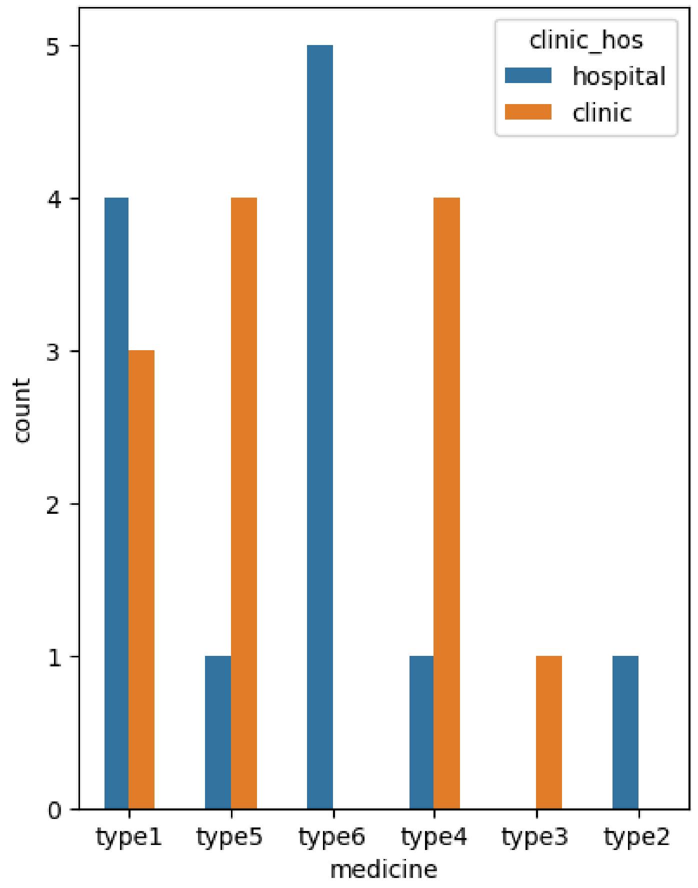


DATA ANALYSIS

Gp Doctors

There is no gp class a doctors
80% gp doctors in Class b write
They write Type 1 most
type 6 , 2 in hospitals and type 3 , 4 , 5 most in clinics





DATA ANALYSIS

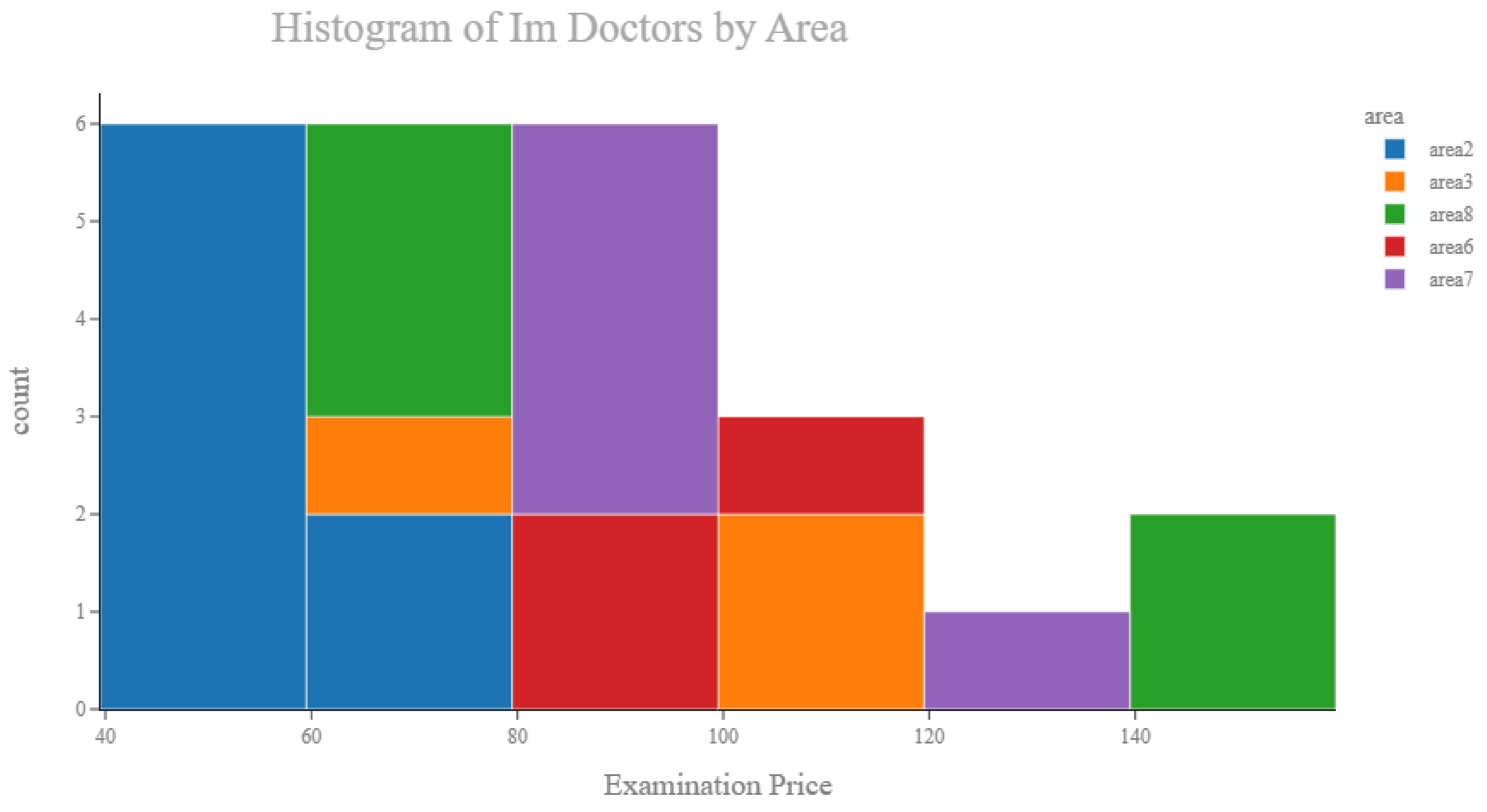
Type 1 , 4 in high ranges

DATA ANALYSIS

Area 7 , 8 in high ranges

DATA ANALYSIS

Area Distribution

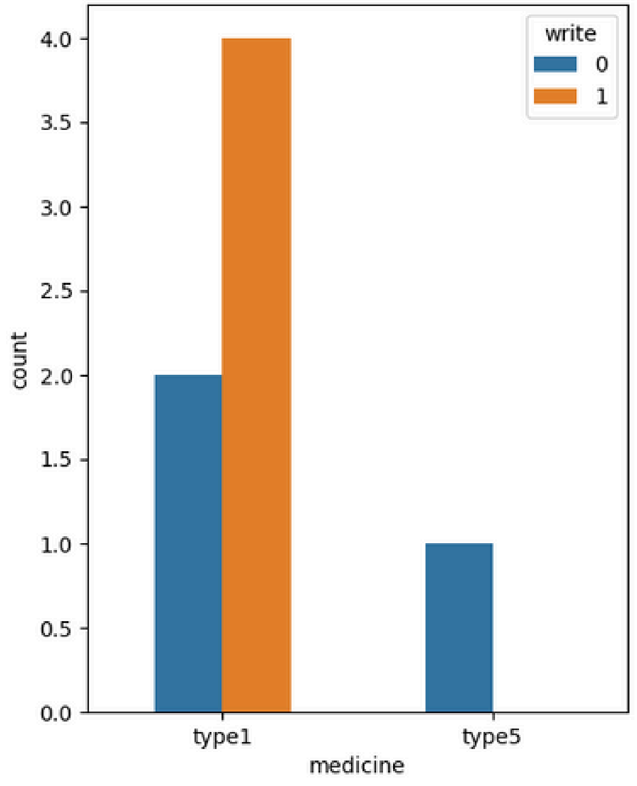


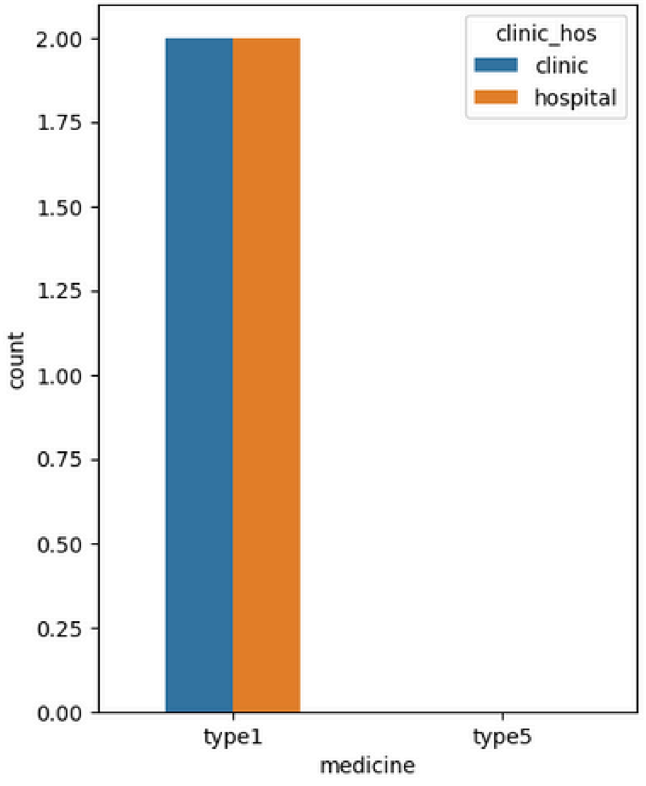
DATA ANALYSIS

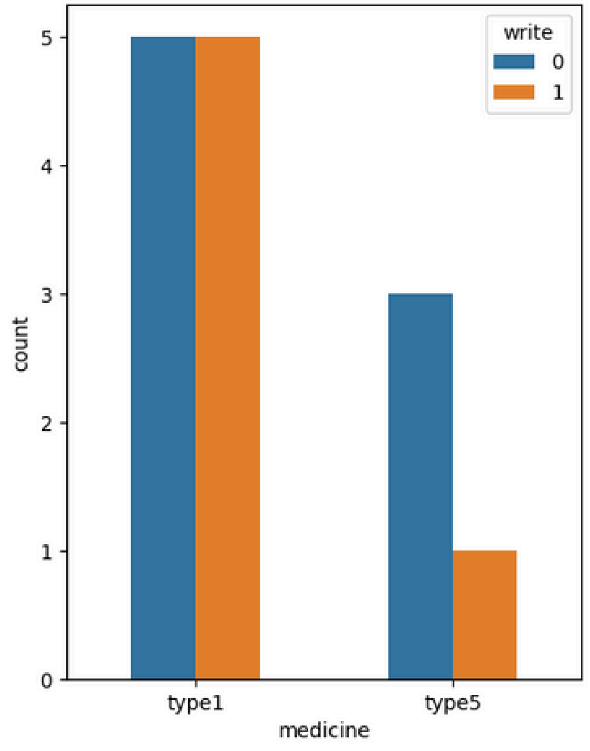
Ent Doctors

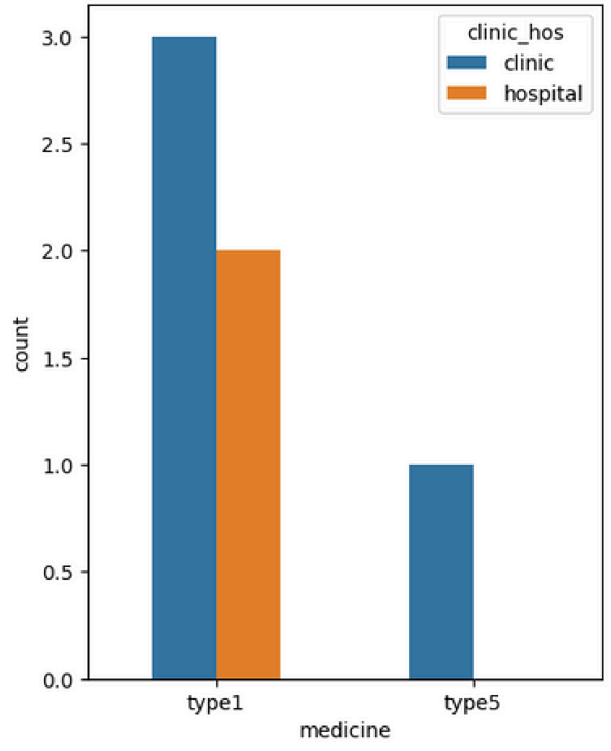
55% ent doctors in Class a write
They write type 1 most and they did not write type 5
clinics and hospitals

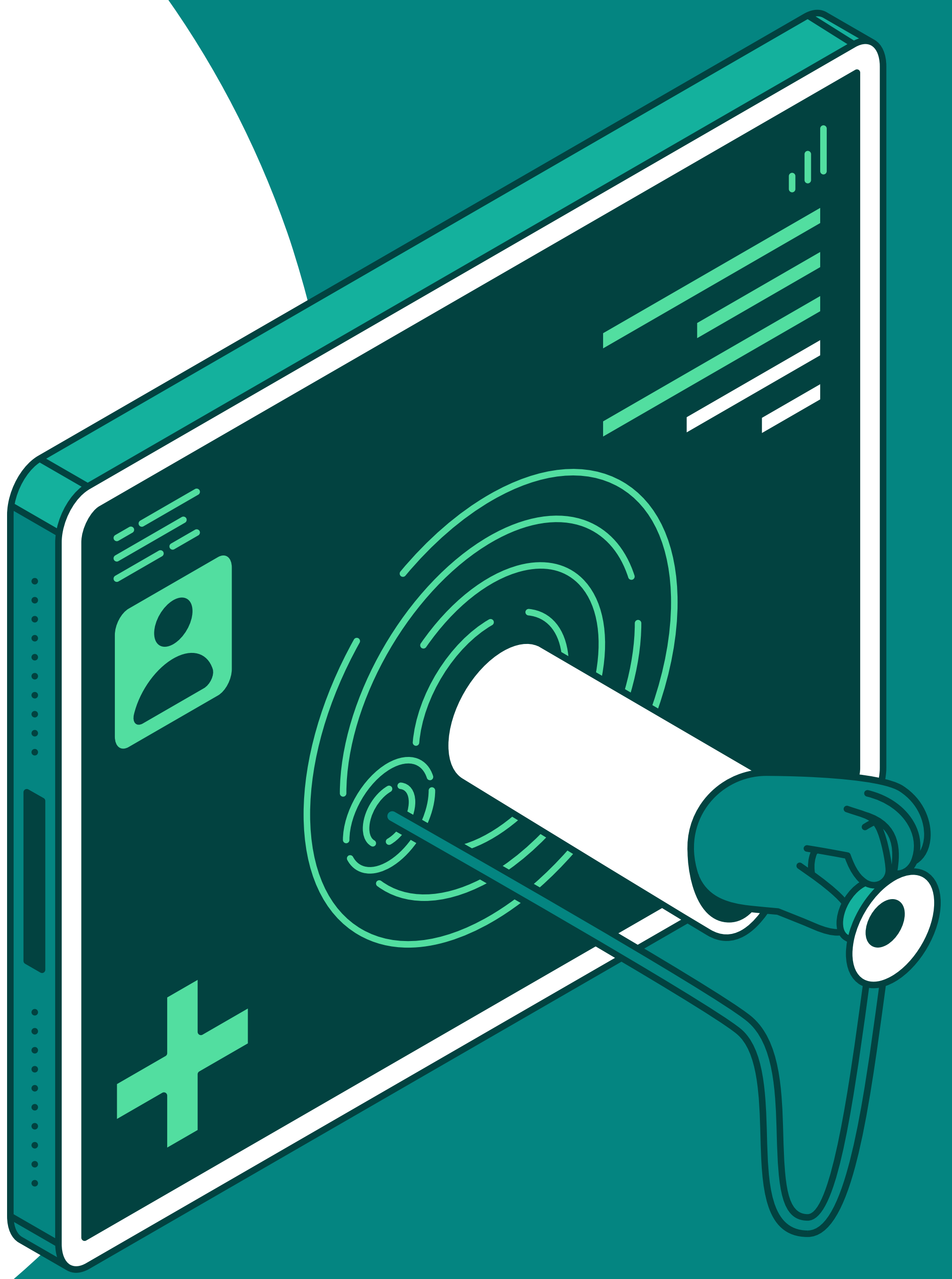
55% ent doctors in Class b did not write
They write Type 1 50%
most in clinics









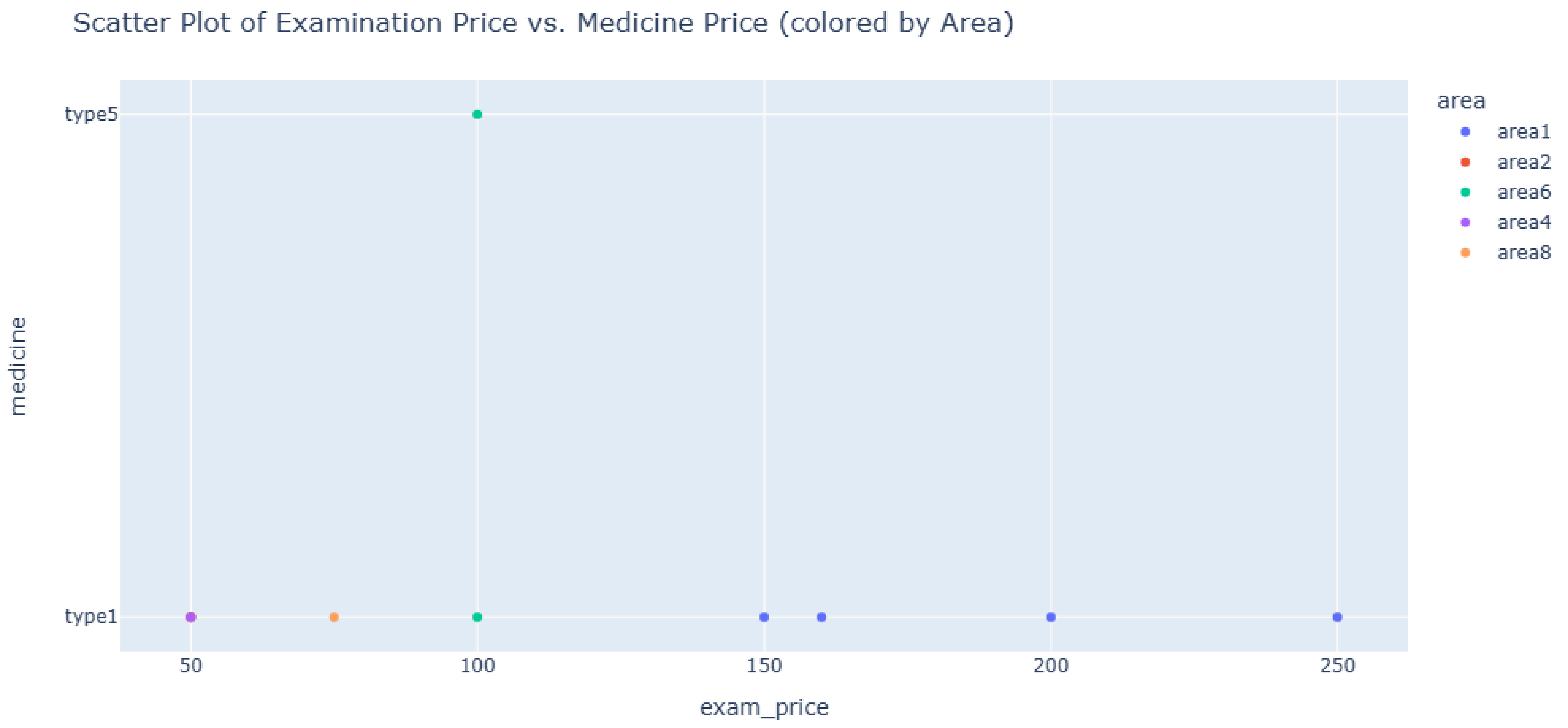


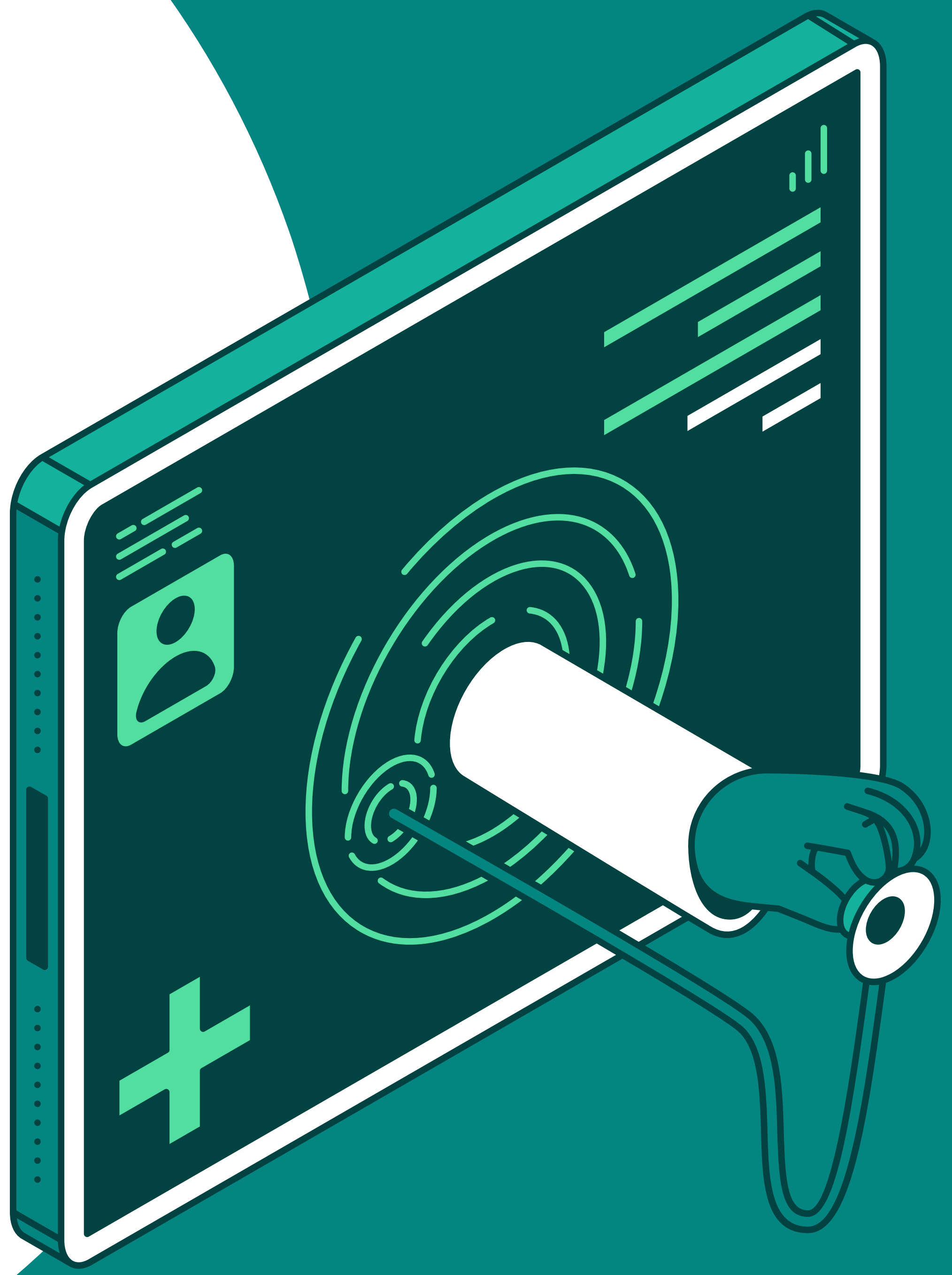
DATA ANALYSIS

Classes Distrbution

DATA ANALYSIS

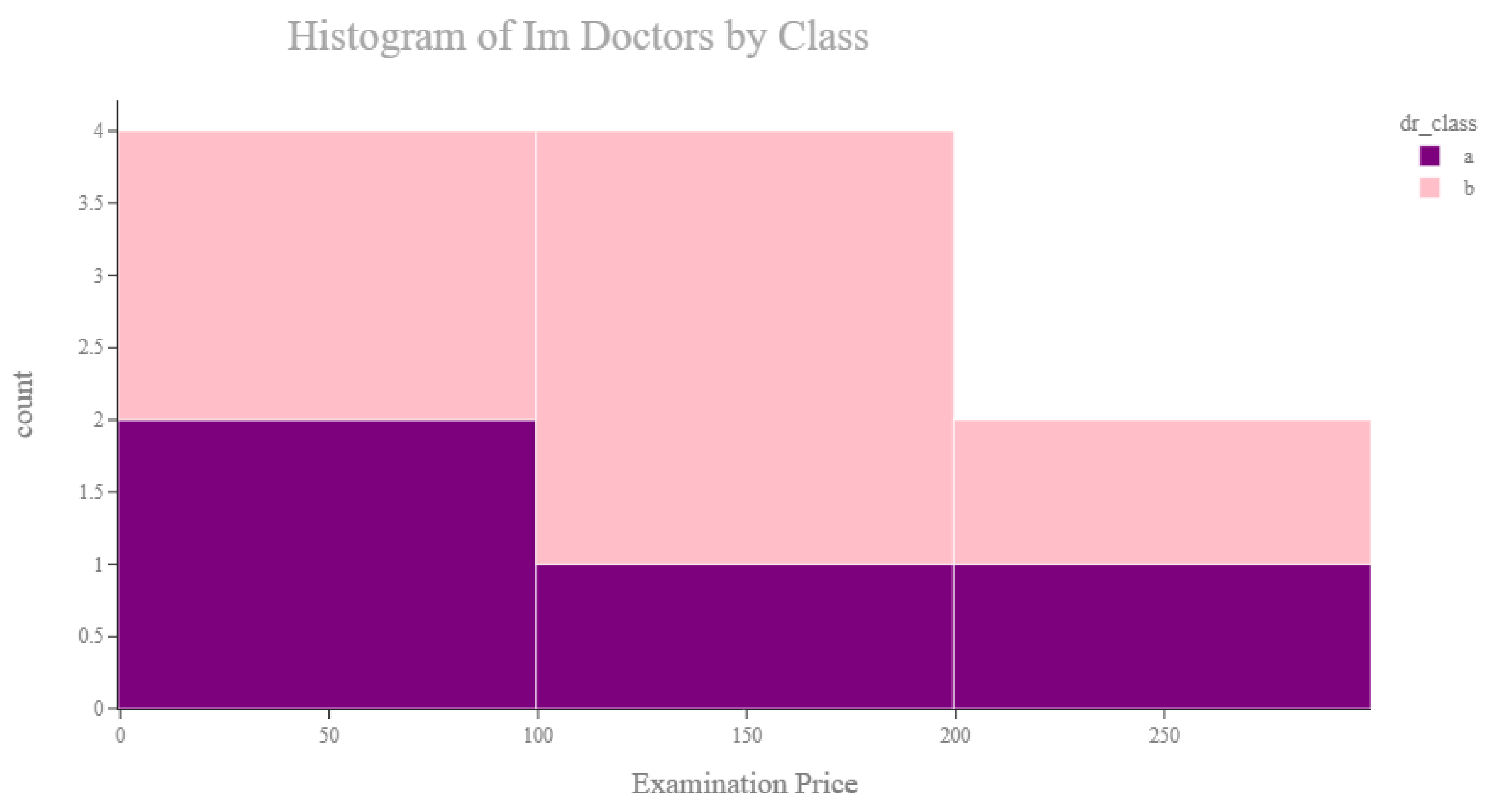
Areas Distribution





DATA ANALYSIS

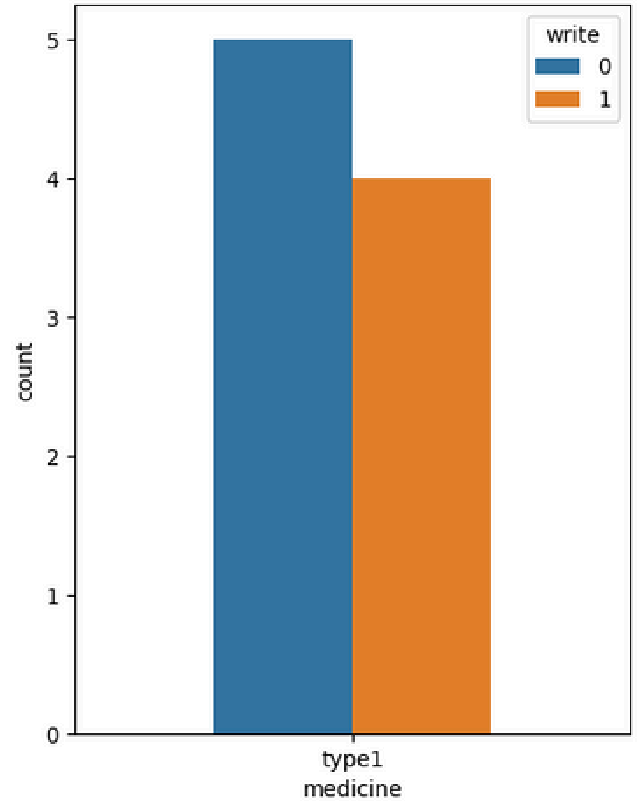
Classes Distribution

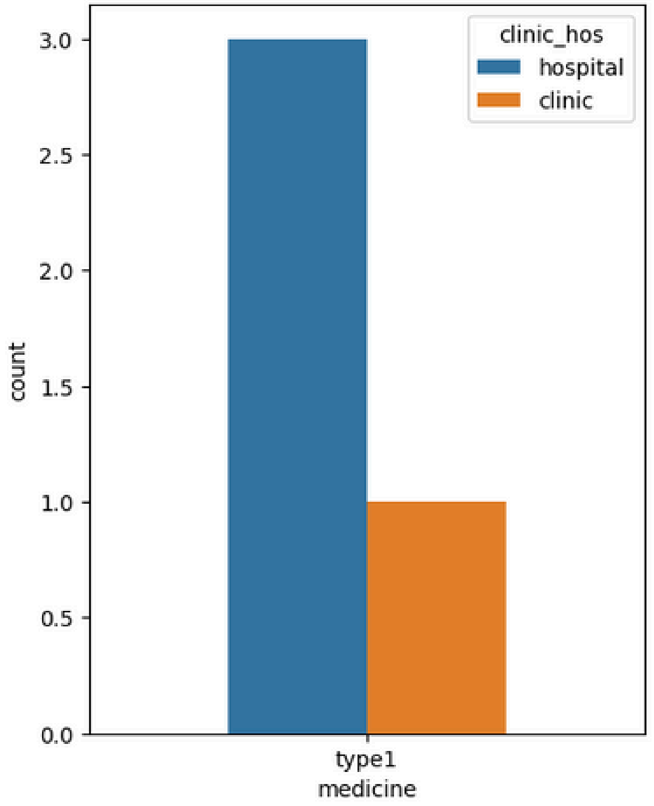


DATA ANALYSIS

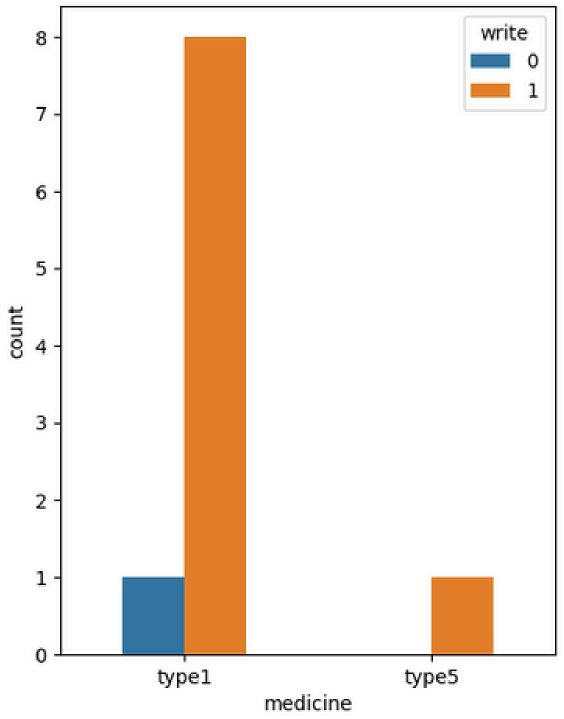
Or Doctors

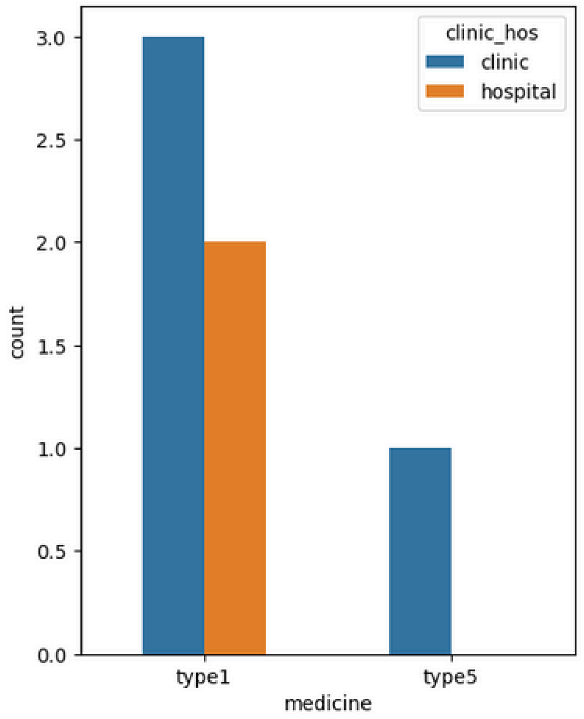
55% or doctors in Class a did not write
They write type 1 only
most in hospitals

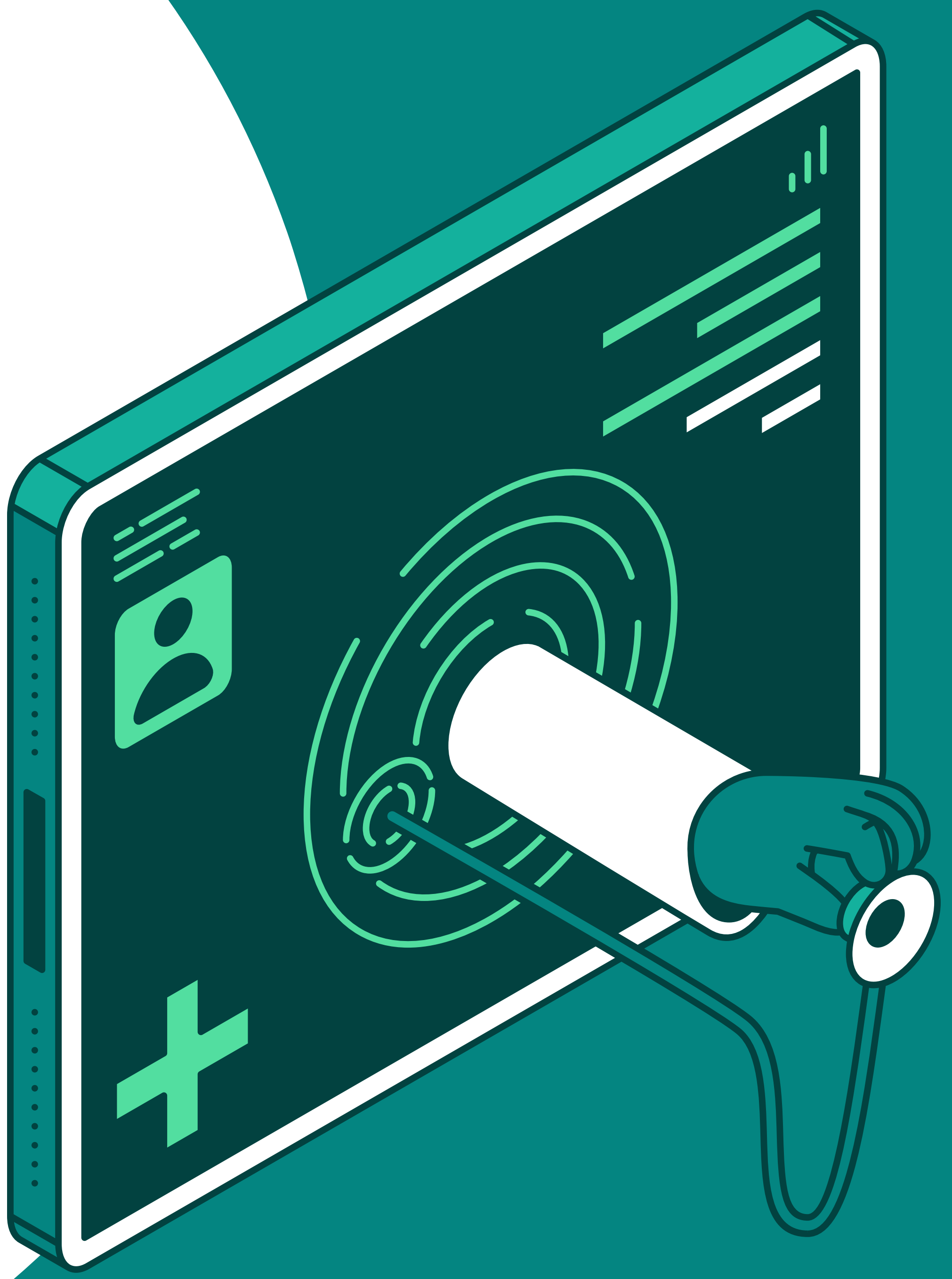




85% or doctors in Class b write
They write Type 1 and 5
most in clinics





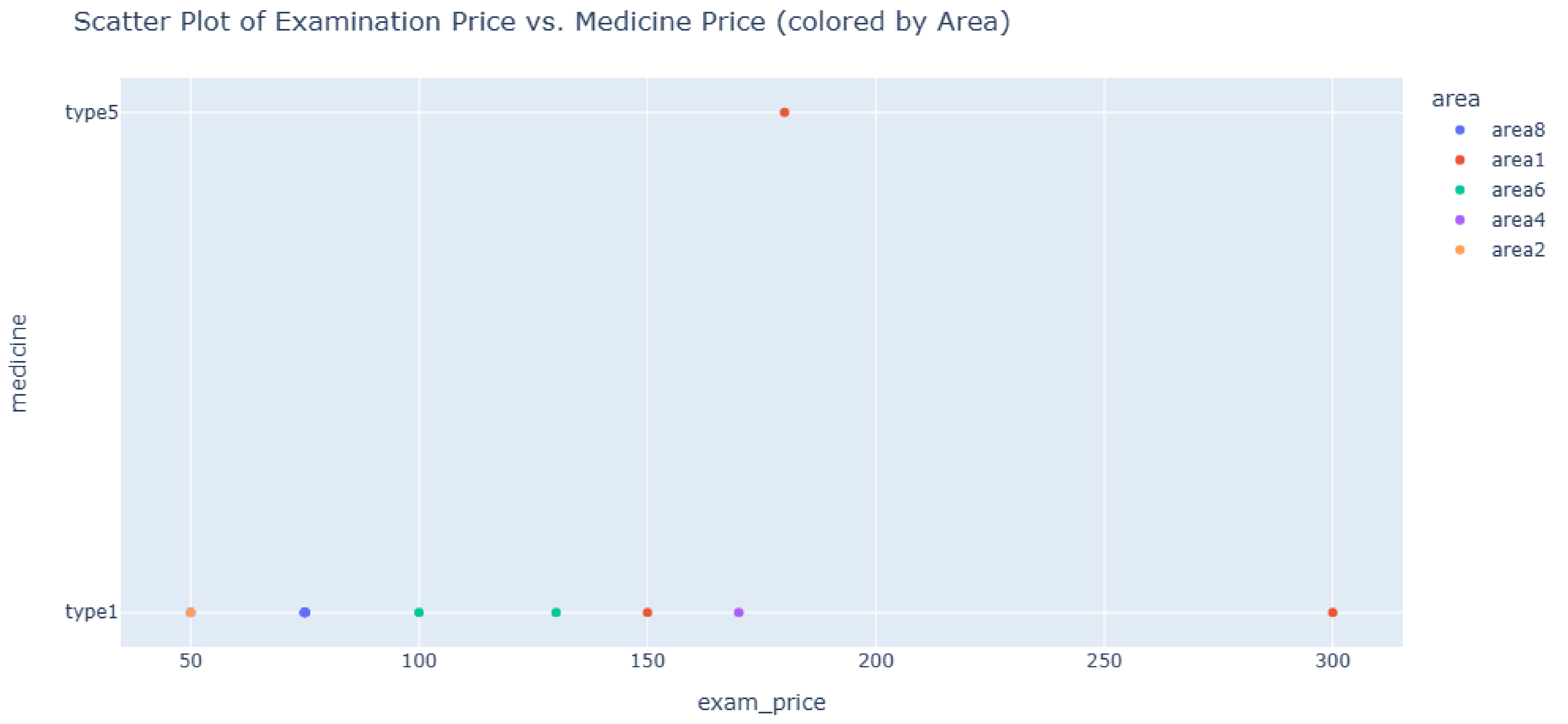


DATA ANALYSIS

Class b in low ranges

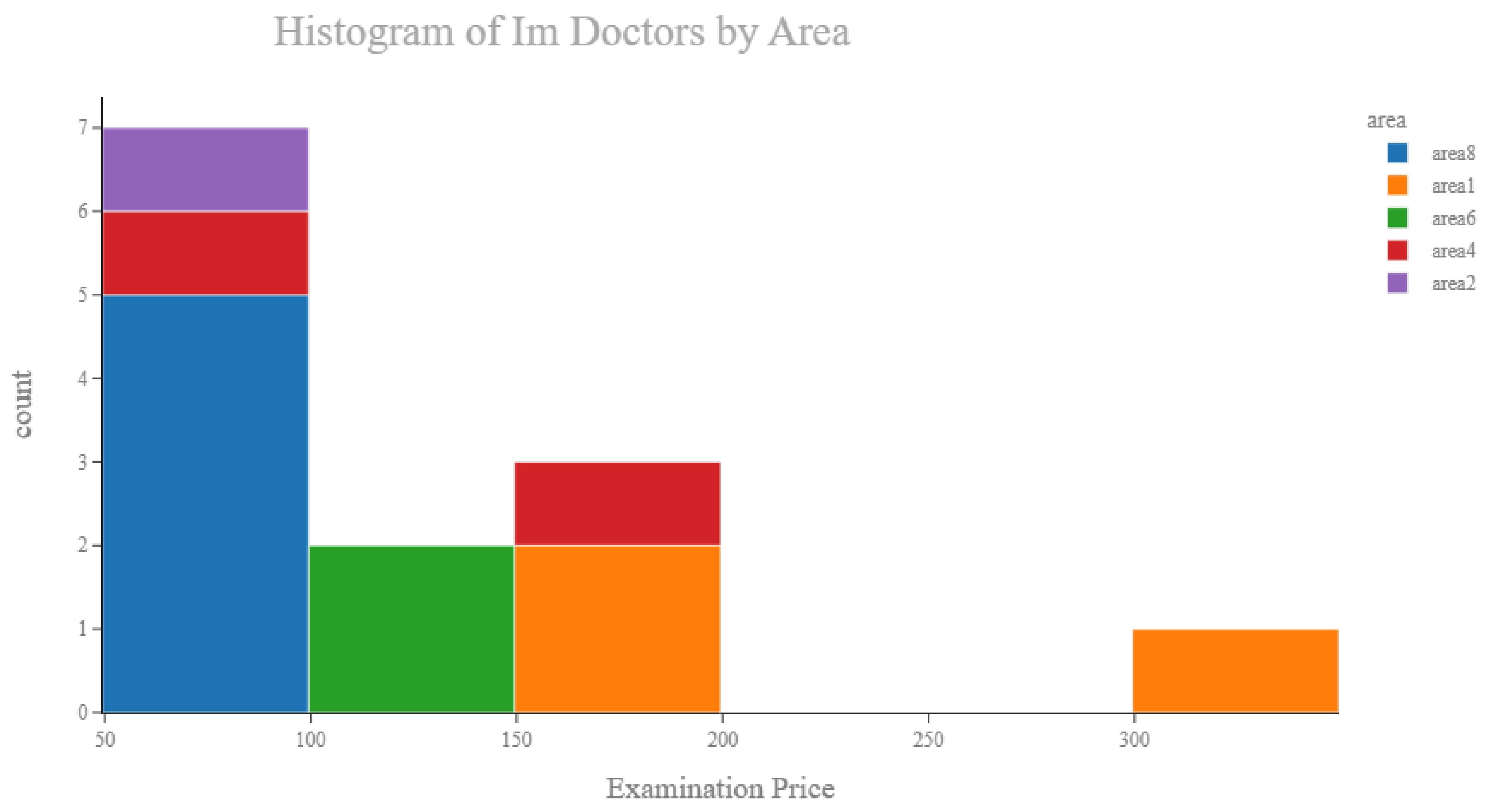
DATA ANALYSIS

Areas Distribution



DATA ANALYSIS

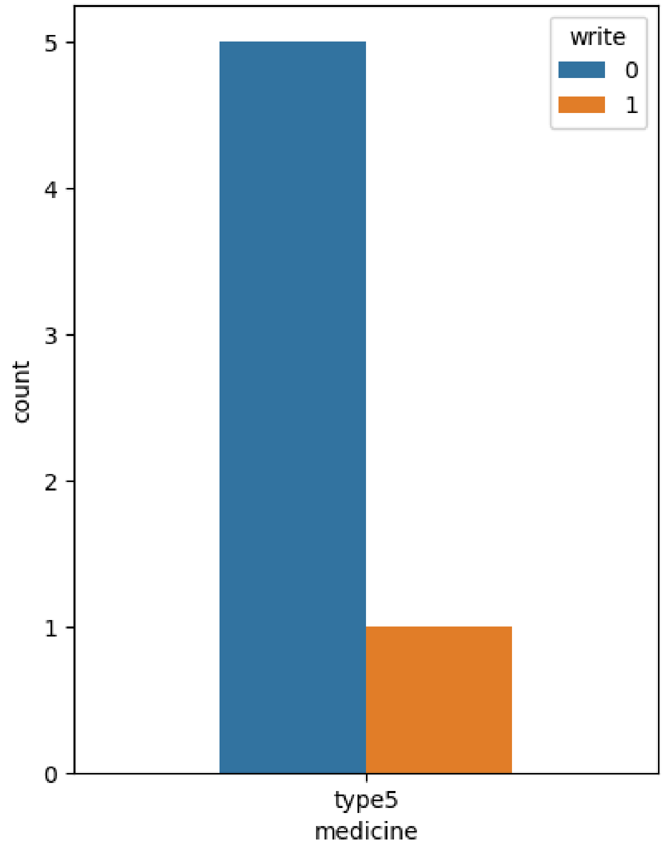
Areas Distribution

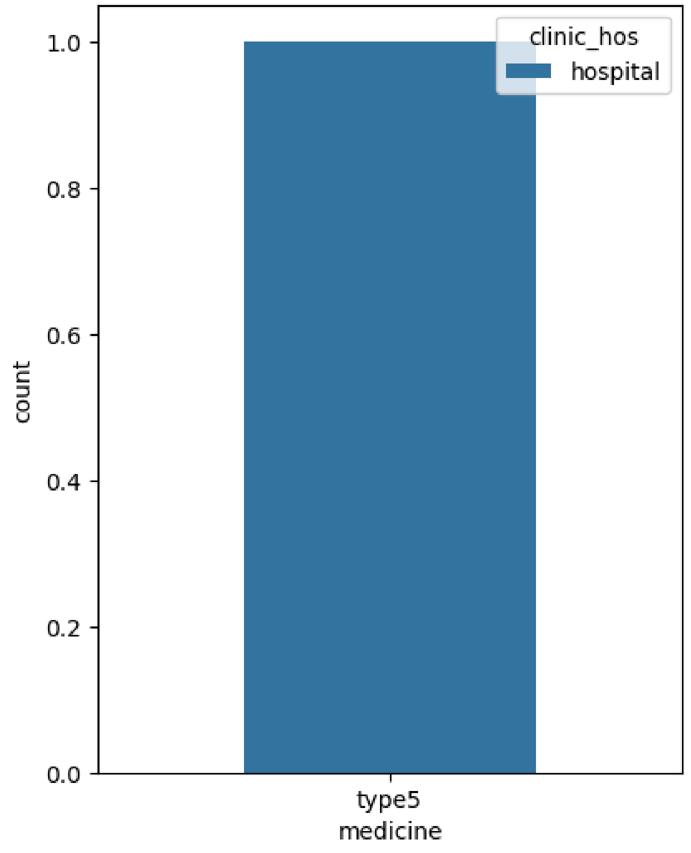


DATA ANALYSIS

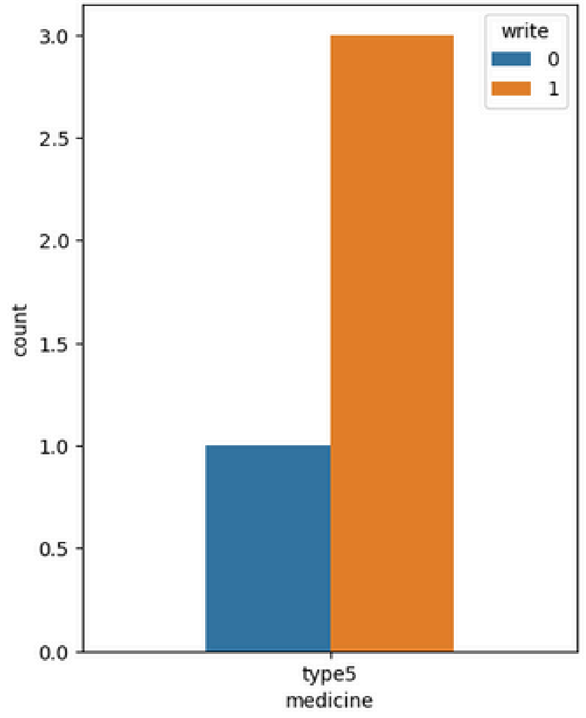
Uro Doctors

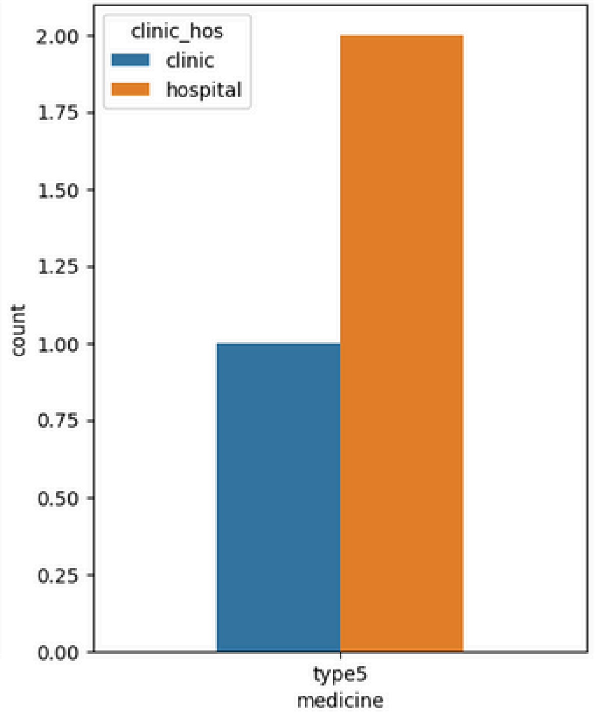
85% uro doctors in Class a did not write
They write type 5 just one in hospital





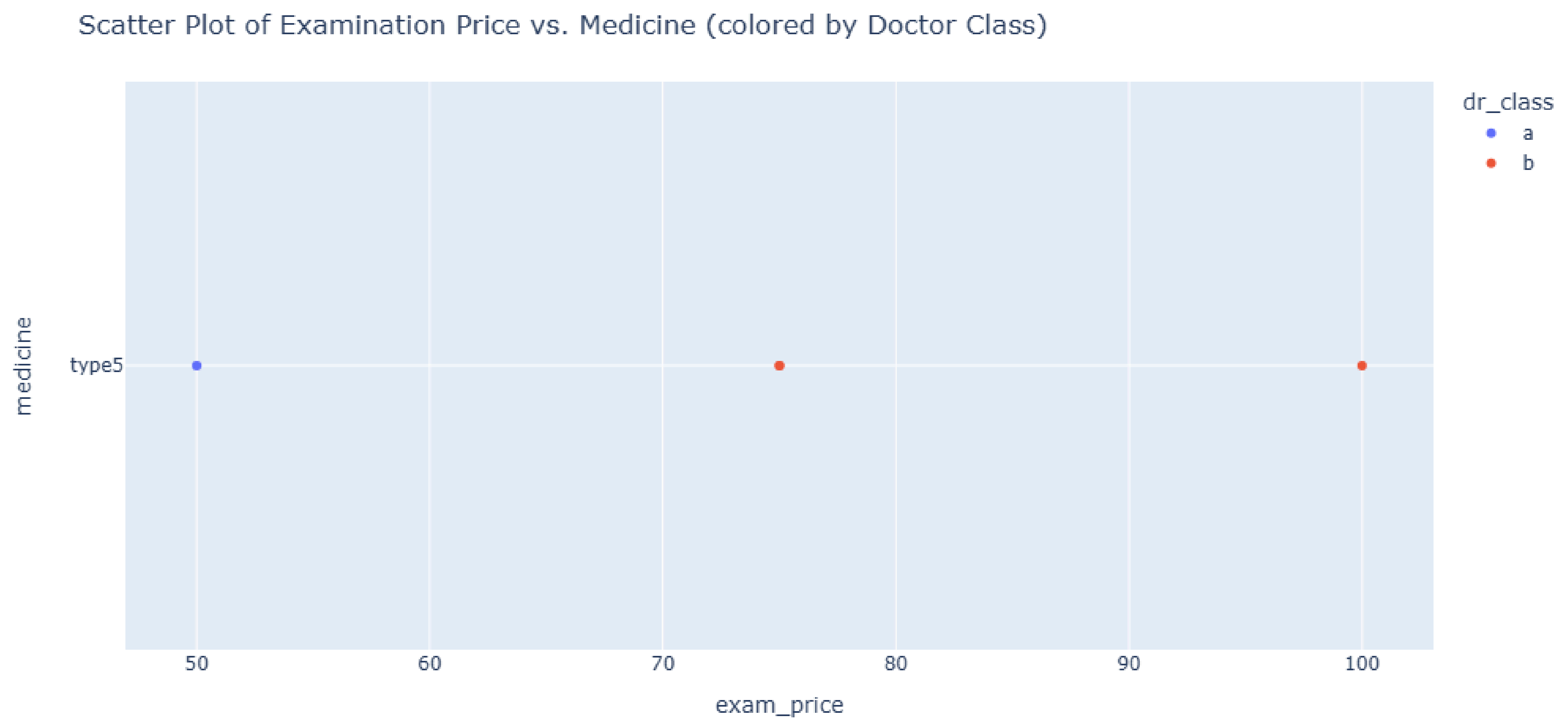
75% uro doctors in Class b write
They write only type 5
most in hospitals





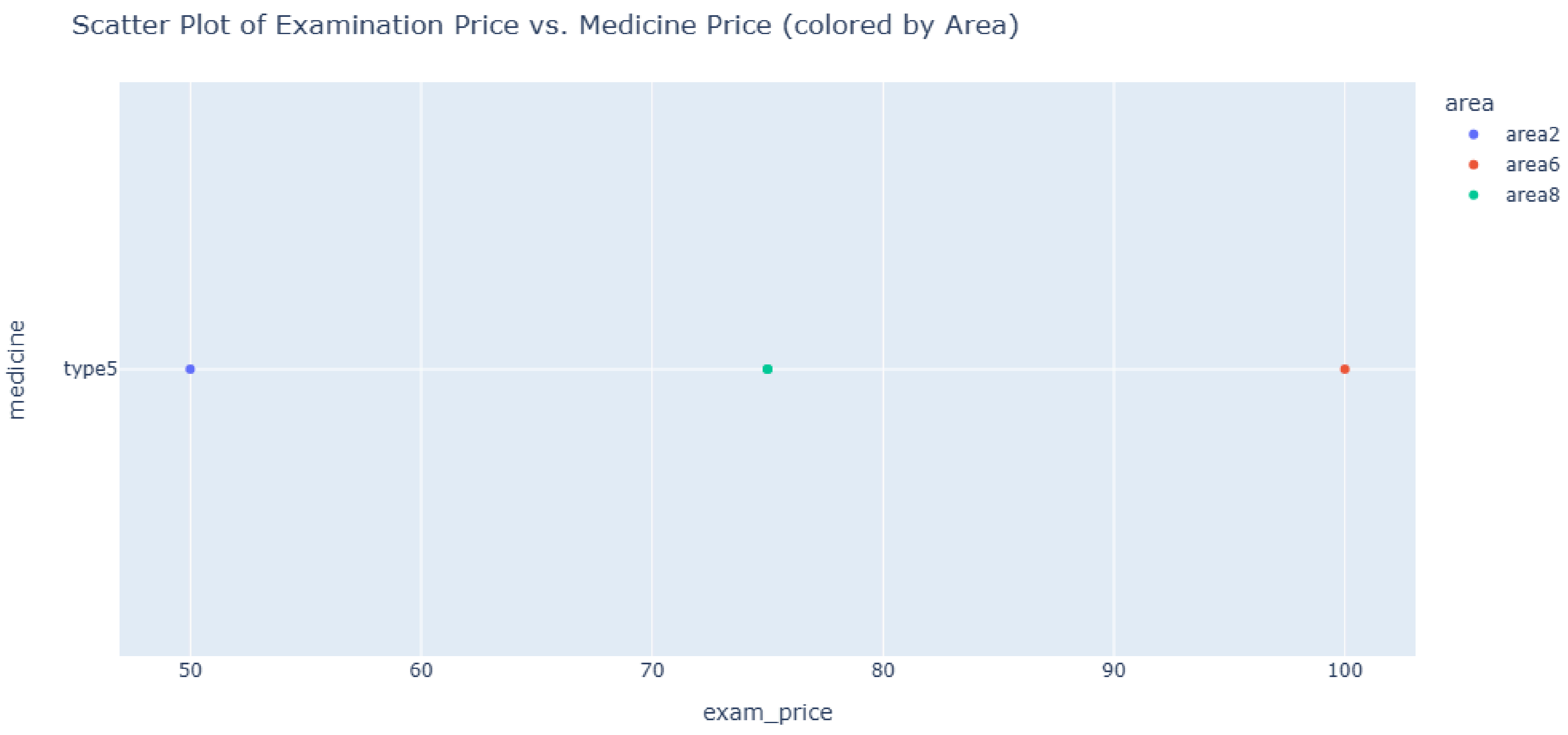
DATA ANALYSIS

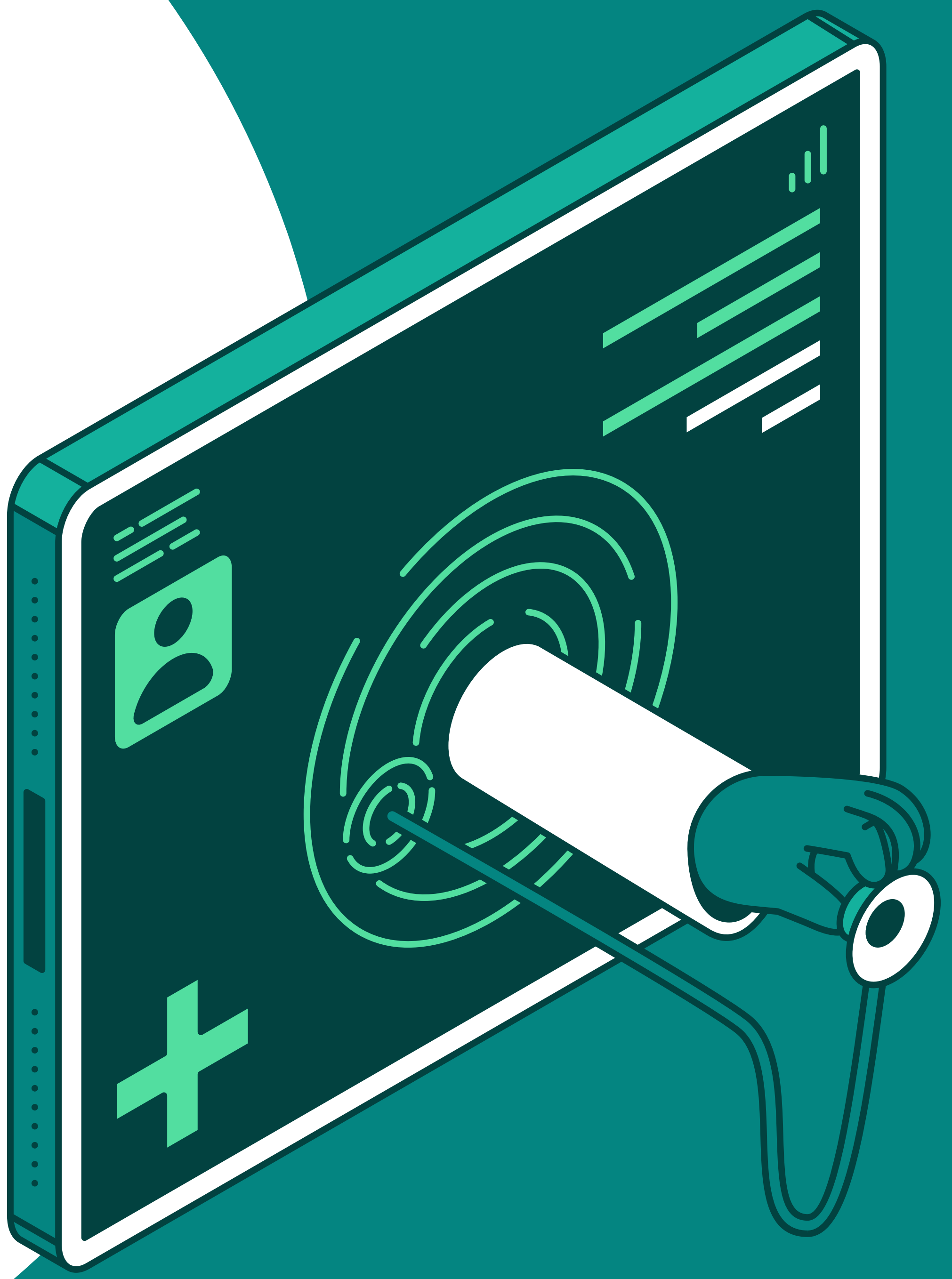
Class a just 1 in type 5 and in hospital and low ranges



DATA ANALYSIS

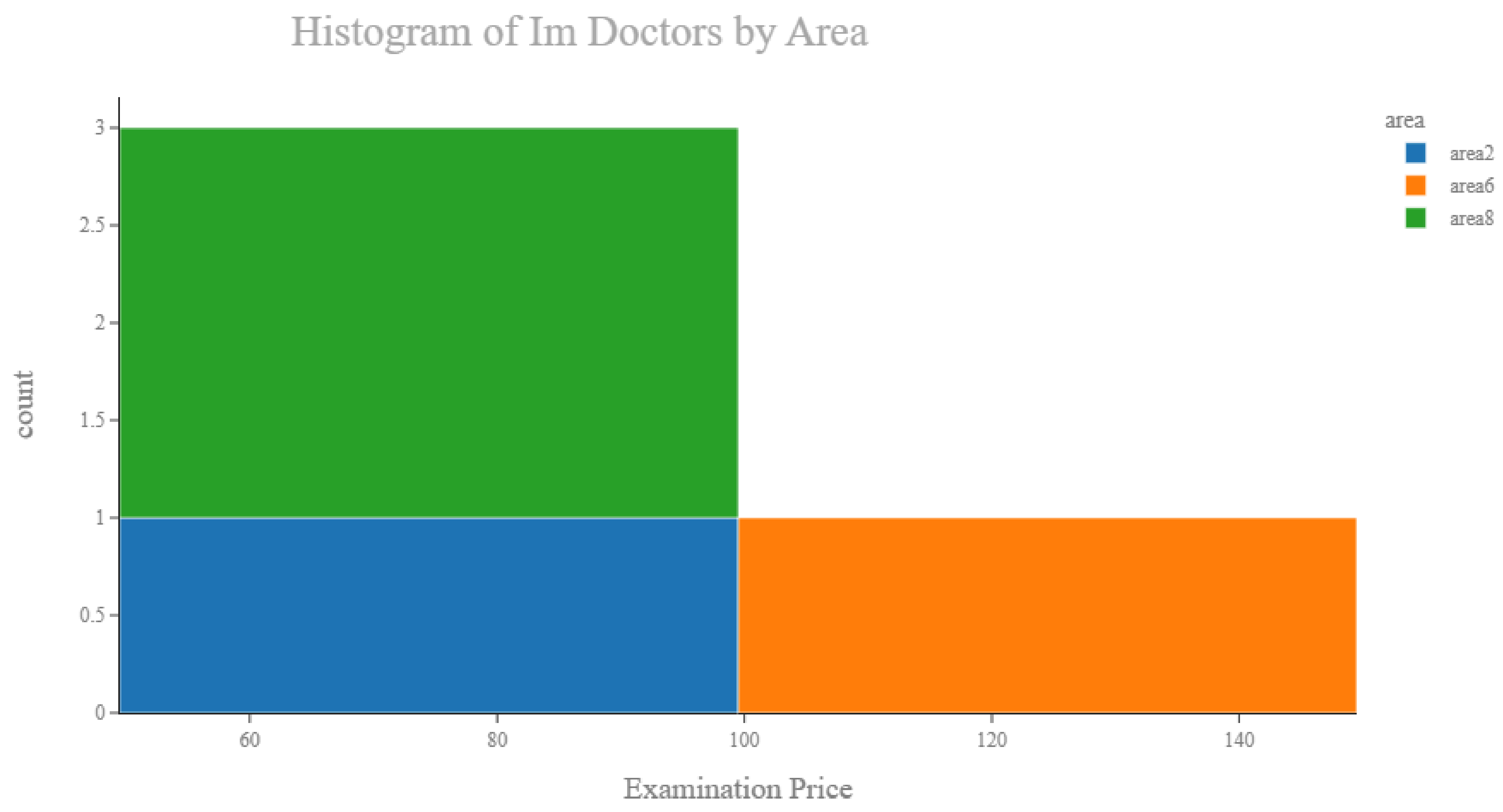
Areas Distribution

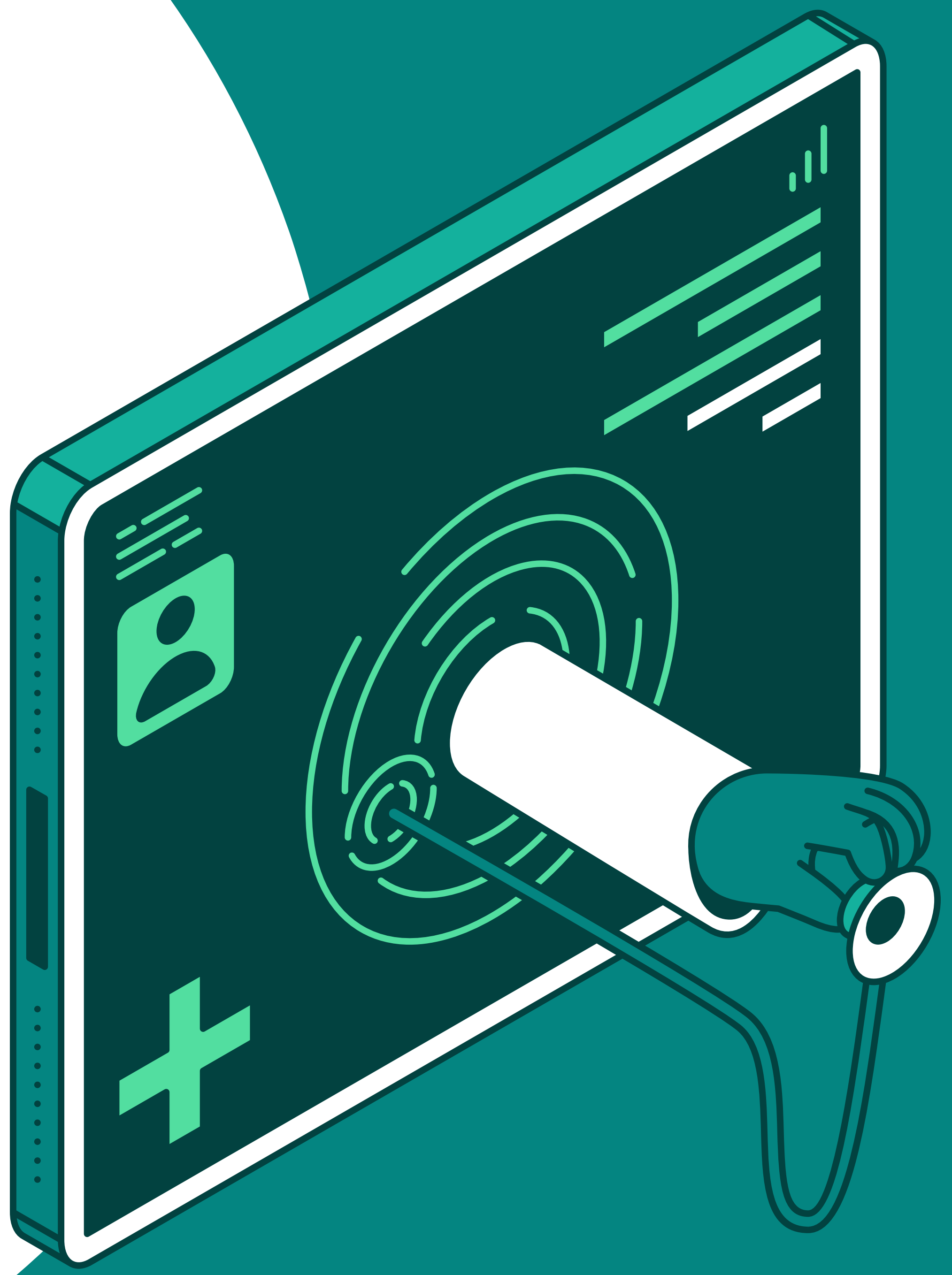




DATA ANALYSIS

Areas Distribution

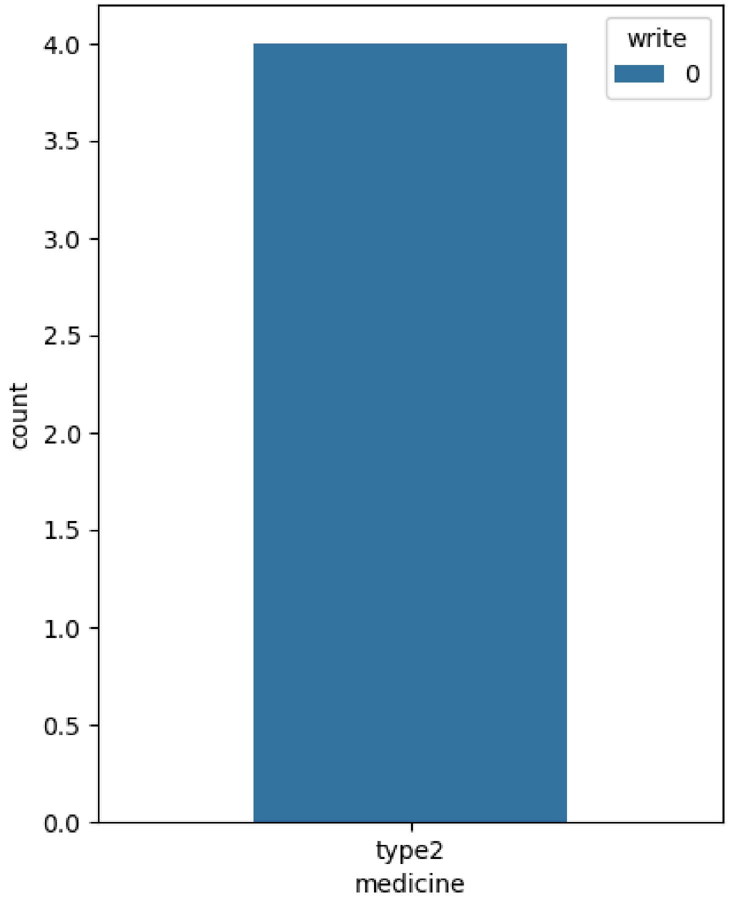




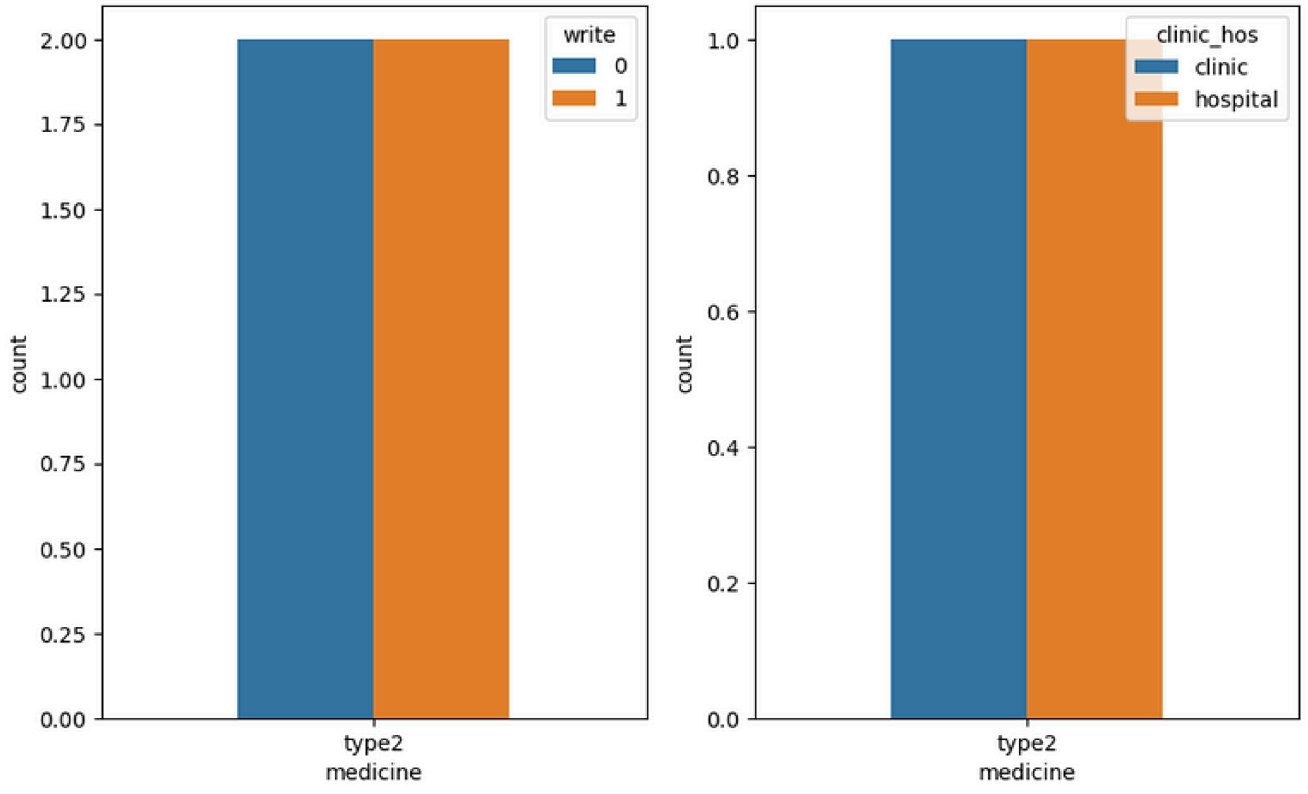
DATA ANALYSIS

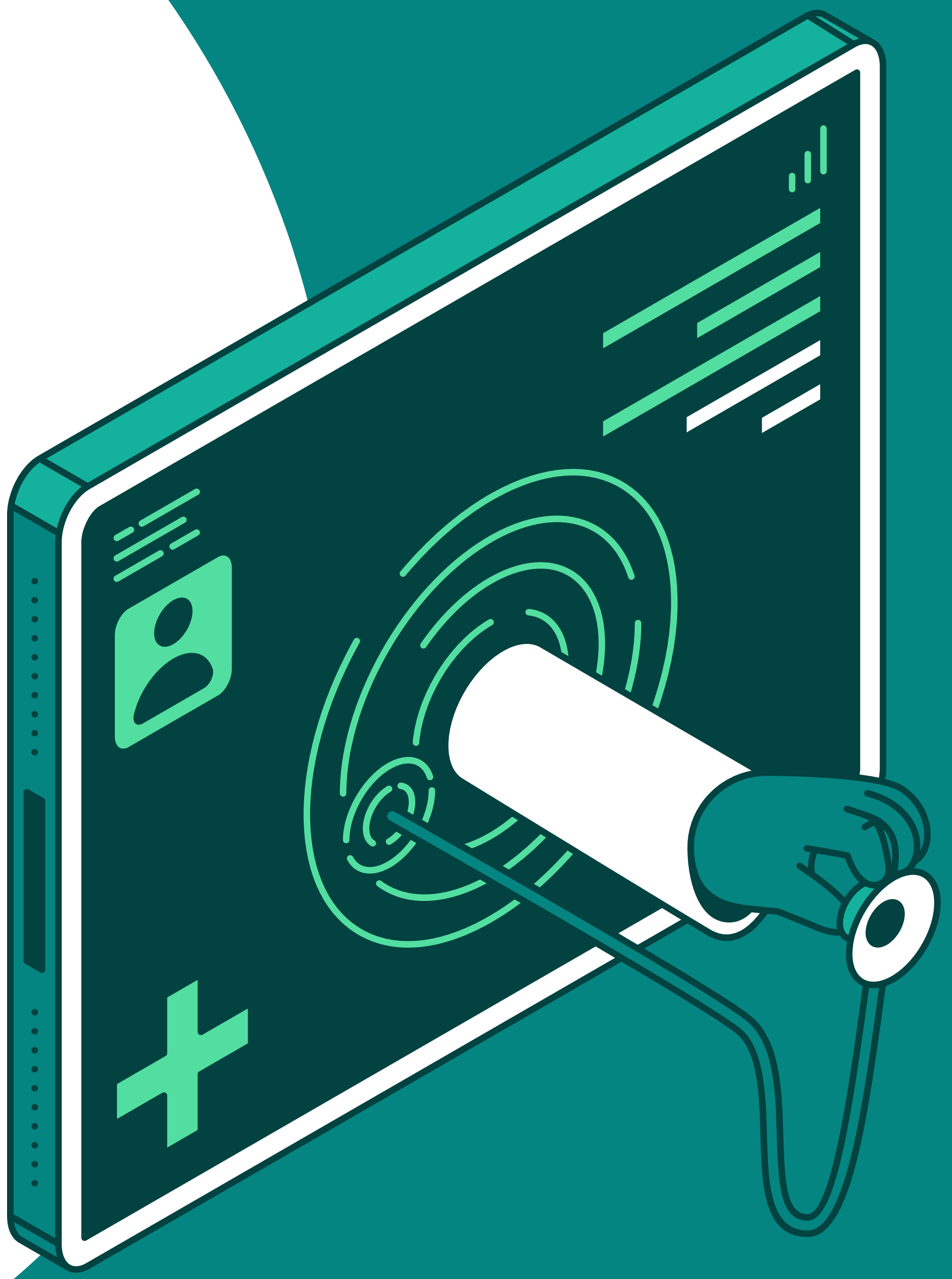
Vas Doctors

100% vas doctors in Class a did not write

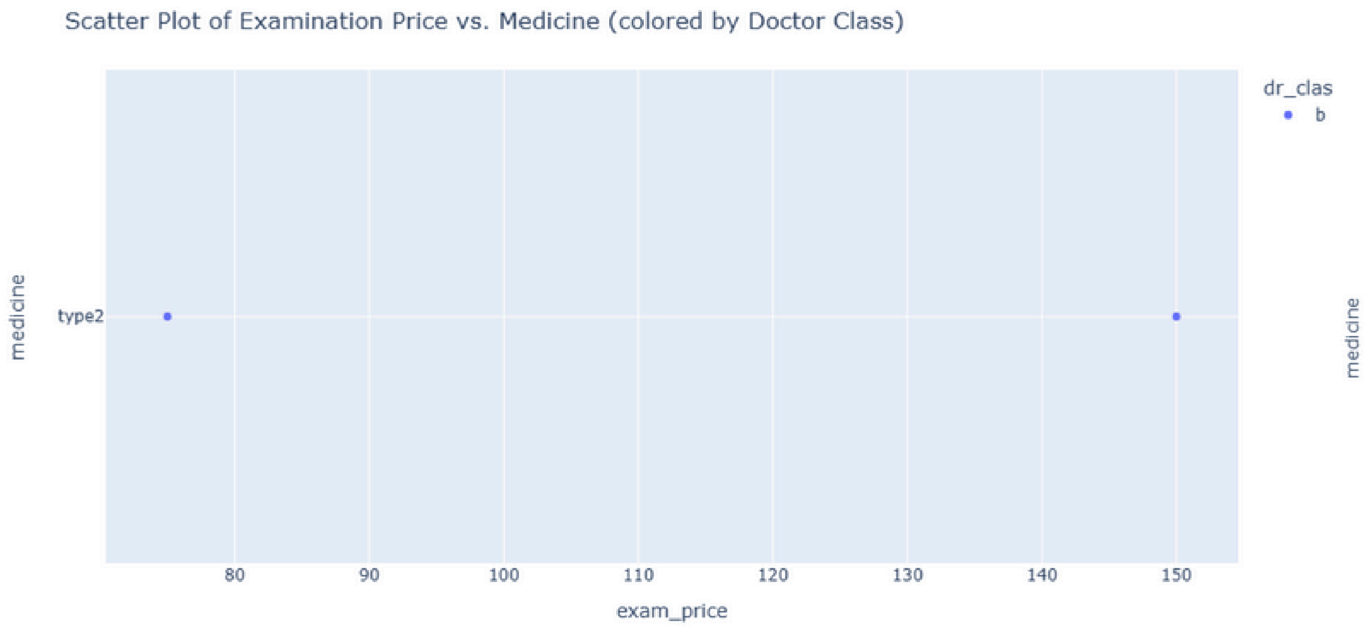


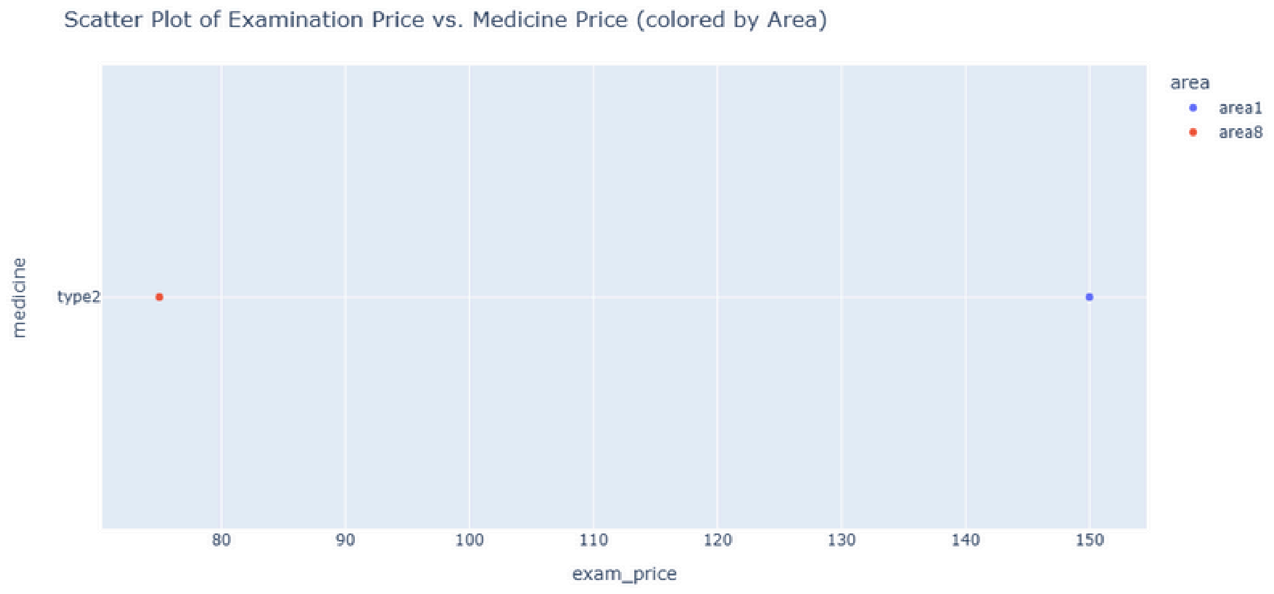
50% vas doctors in Class b write
They write Type 2





DATA ANALYSIS

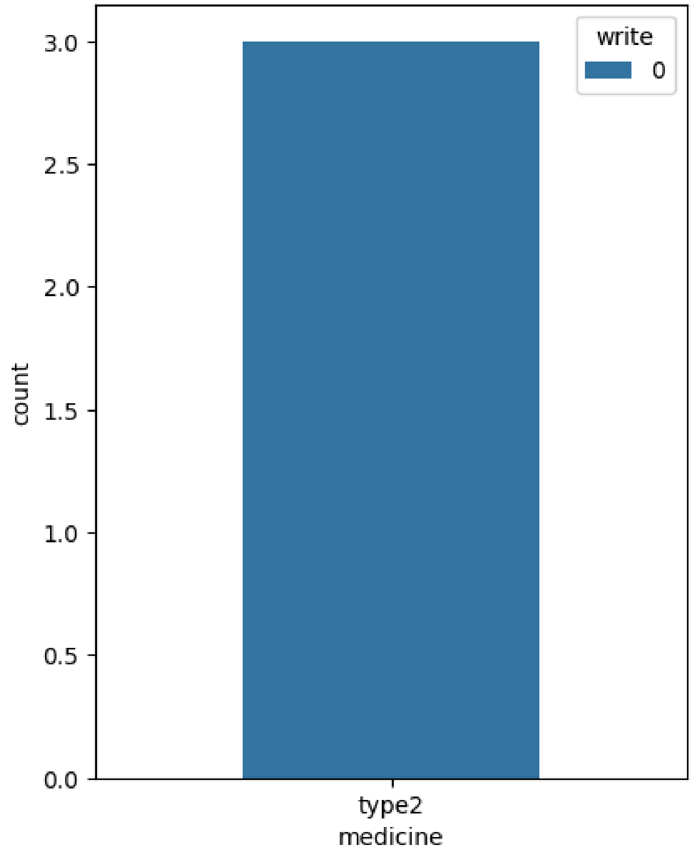




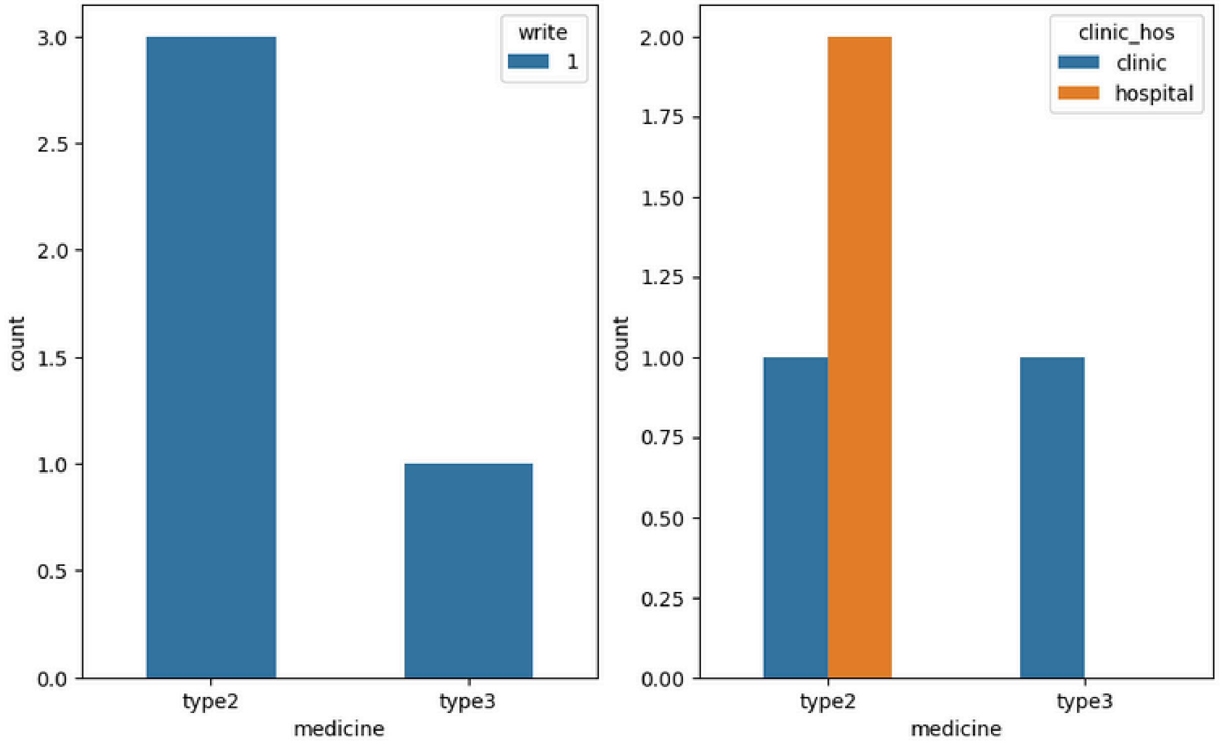
DATA ANALYSIS

neuro Doctors

100% neuro doctors in Class a did not write

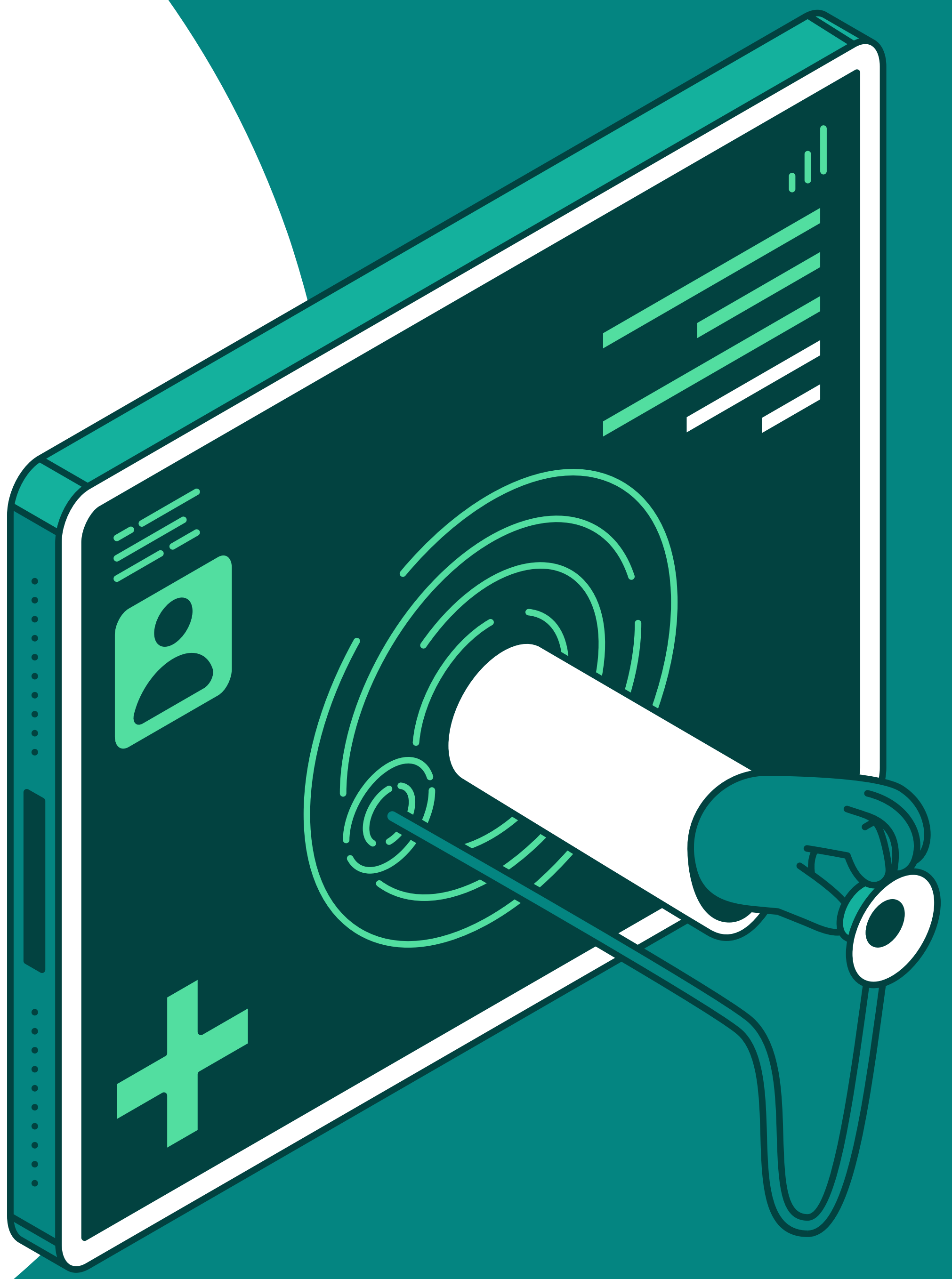


100% neuro doctors in Class b write
They write Type 2 and 3
type 3 in clinics and 2 most in hospitals



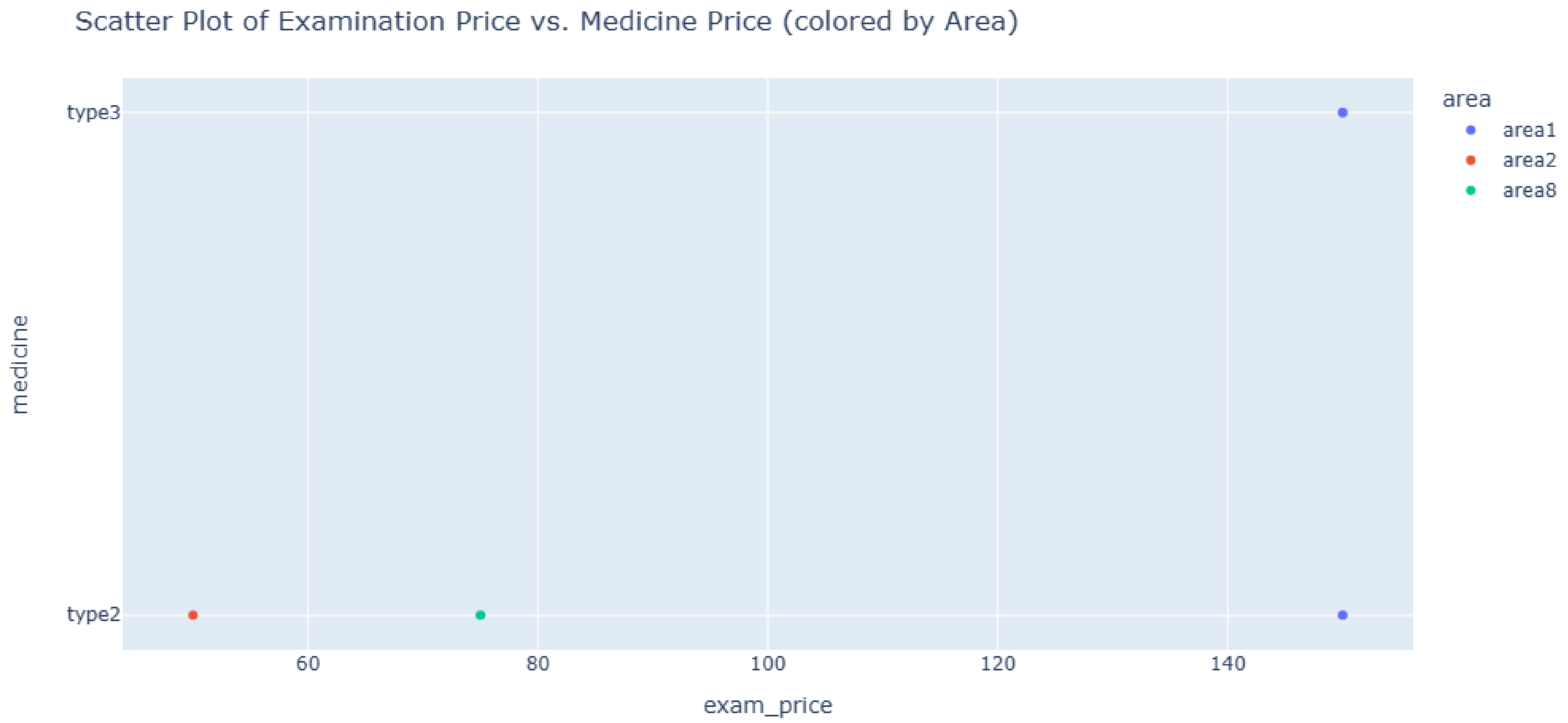
DATA ANALYSIS

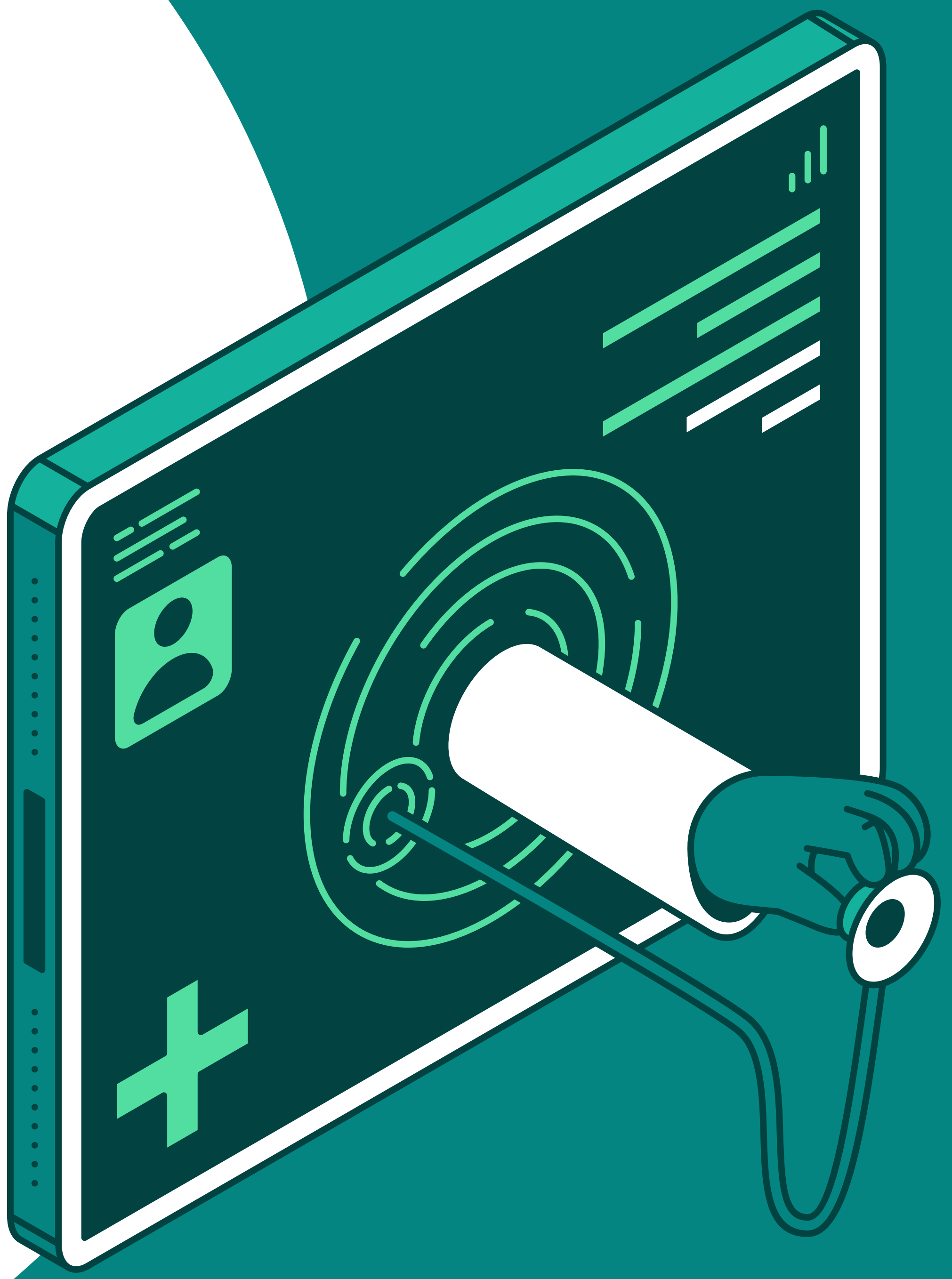
Classes Distribution



DATA ANALYSIS

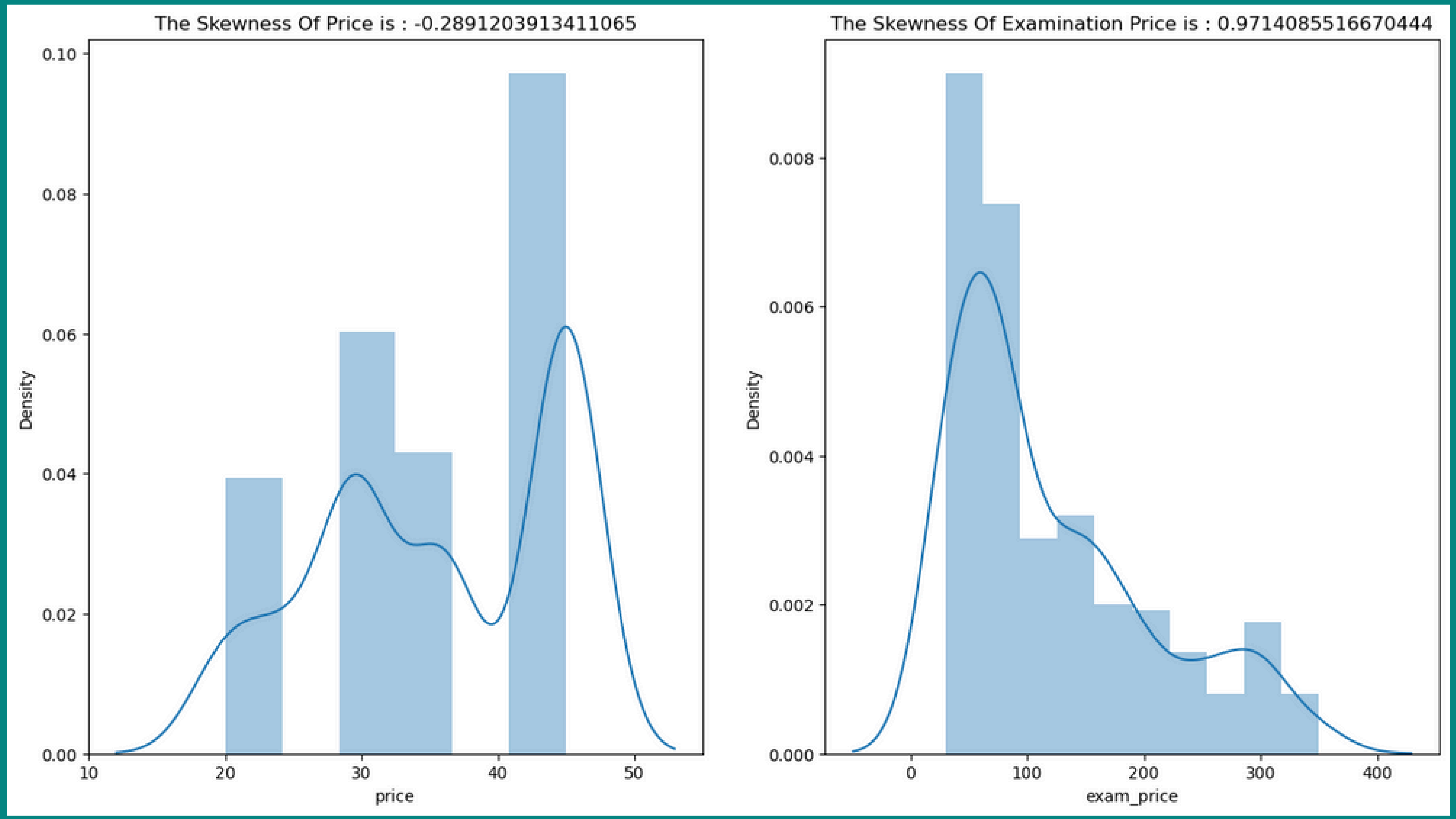
Areas Distribution





PREPARE DATA FOR ML

Skewed Data



The left plot shows the distribution of the "price" variable, which has a skewness value of approximately -0.29, indicating a distribution that is nearly symmetric. Similarly, the right plot represents the distribution of the "exam_price" variable, with a skewness value of approximately 0.97. While slightly positively skewed, it does not exhibit a high level of skewness. Based on these observations, neither variable requires log transformation, as their skewness values fall within acceptable ranges for analysis.

PREPARE DATA FOR ML

Label and One-Hot Encoding

```
encoded_df = pd.DataFrame(features_cat , columns = h_encoder.get_feature_names_out(object_col.columns) , index = data.index)
encoded_df
```

	medicine_type1	medicine_type2	medicine_type3	medicine_type4	medicine_type5	medicine_type6	area_area1	area_area2	area_area3	area_area4
0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0
2	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
385	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
386	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
387	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
388	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
389	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

I applied one-hot encoding to the categorical columns: medicine, area, and speciality, to convert them into multiple binary columns representing their unique categories. For the doctor class and hospital or clinic columns, I used label encoding to map their categories to numerical values while maintaining their ordinal or nominal relationships.

PREPARE DATA FOR ML

Data Normalization

I applied normalization to the price and exam_price columns using the MinMaxScaler from sklearn.preprocessing. This technique scales the values of these columns to a range between 0 and 1, preserving the relationships between the data points while ensuring all values fall within the same range. This is especially useful for algorithms sensitive to feature scaling.

Data Splitting

I performed data splitting to divide the dataset into training and testing sets. The target variable y is set as the write column, while the feature set X includes all columns except write. Using the train_test_split function from sklearn.model_selection, the data is split as follows:

- Training set (80%): X_train and y_train are used to train the model.
- Testing set (20%): X_test and y_test

MODEL SELCETION

1. Decision Tree Model

A Decision Tree Classifier was implemented with hyperparameter tuning using GridSearchCV and 5-fold cross-validation via ShuffleSplit.

- Parameters Tuned:
  - max_depth (3 to 8)
  - min_samples_leaf (6 to 16)
  - min_samples_split (2 to 16)
- Optimal Parameters:
  - max_depth: 4
  - min_samples_leaf: 6
  - min_samples_split: 2
- Performance:
  - Training Accuracy: 74%
  - Testing Accuracy: 81%
  - f1-score: Training (76%), Testing (85%)

2. AdaBoost Model (Best Model)

An AdaBoostClassifier was selected as the best model, utilizing a Decision Tree as the base estimator. GridSearchCV with 5-fold cross-validation tuned the hyperparameters.

- Parameters Tuned:
  - n_estimators: 85
  - learning_rate: 0.4
  - base_estimator hyperparameters (max depth, min samples leaf, min samples split)
- Optimal Parameters:
  - n_estimators: 85
  - learning_rate: 0.4
  - Base Decision Tree:
    - max_depth: 4
    - min_samples_leaf: 14
    - min_samples_split: 8
- Performance:
  - Training Accuracy: 86%
  - Testing Accuracy: 85%
  - f1-score: Training (87%), Testing (88%)
  - fb-score: Training (87%), Testing (89%)

MODEL SELCETION

3. Support Vector Machine (SVM)

An SVC (Support Vector Classifier) was also tested with polynomial kernel. Hyperparameters were tuned using GridSearchCV and 5-fold cross-validation.

- Parameters Tuned:
  - kernel: poly
  - degree: 3
  - C: 2.2
- Optimal Parameters:
  - kernel: poly
  - degree: 3
  - C: 2.2
- Performance:
  - Training Accuracy: 82%
  - Testing Accuracy: 82%
  - f1-score: Training (85%), Testing (87%)

Summary of Model Selection:

The AdaBoost Model emerged as the best-performing model, achieving the highest accuracy and f1-scores on both training and testing data. The use of GridSearchCV and cross-validation ensured optimal hyperparameters were selected for all models, improving their generalizability and performance on unseen data.

MODEL DEPLOYMENT

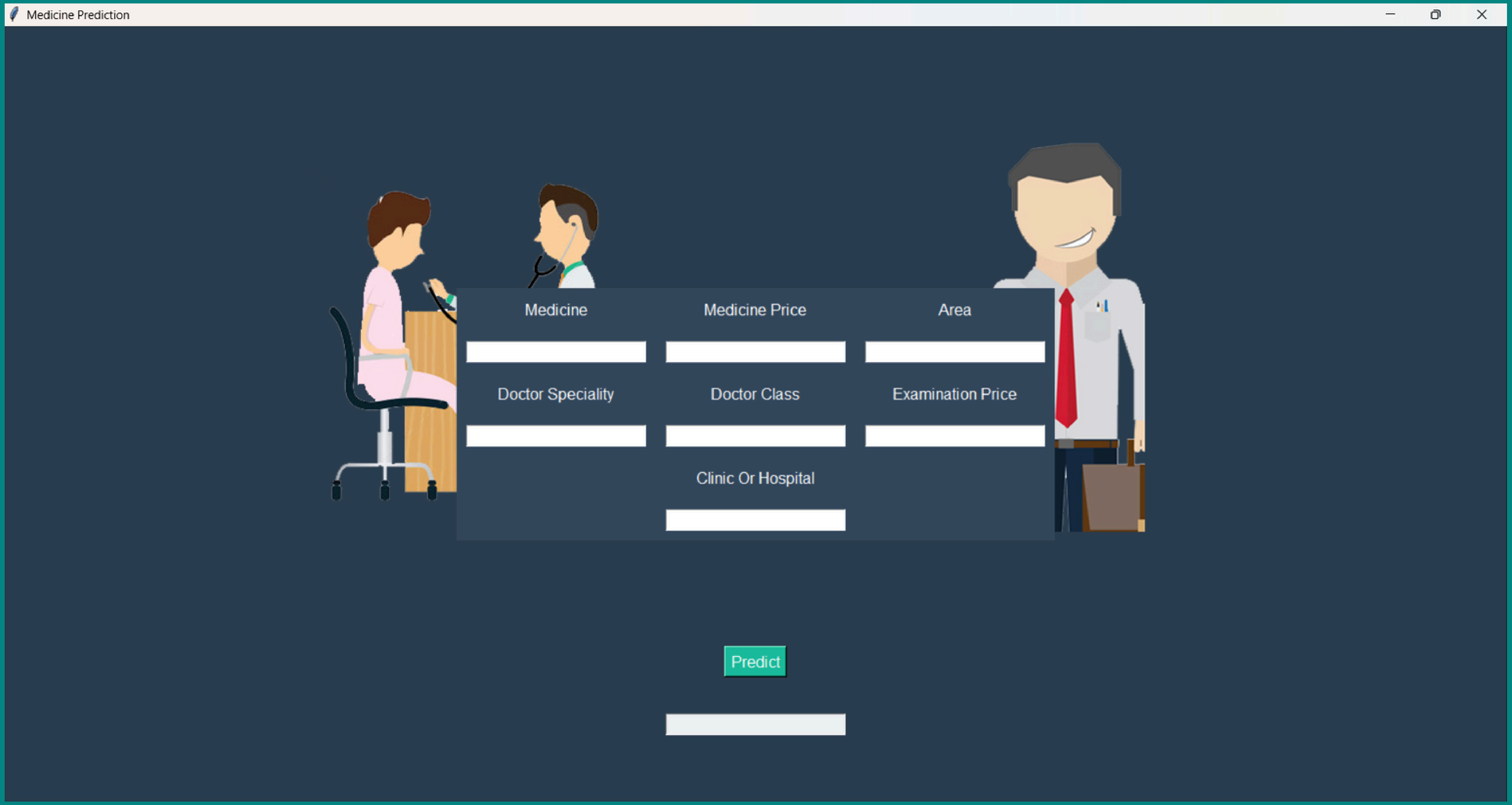
Application Description:

This is a desktop application developed using the Tkinter library in Python. The application leverages an Adaboost model to analyze medical data and predict whether a doctor will write a prescription based on various user inputs. The application incorporates the following features:

1. **Model and Preprocessing Saving:**
    - **The trained Adaboost model is saved as a file (loaded_clf.pkl) using Joblib.**
    - **Encoders and preprocessing tools such as OneHotEncoder and Scaler are also saved for reuse.**
2. **User Interface:**
    - **A user-friendly graphical interface that provides input fields for entering data such as:**
        - **Medicine name.**
        - **Medicine price.**
        - **Geographical area.**
        - **Doctor’s specialty.**
        - **Doctor’s class (A or B).**
        - **Examination price.**
        - **Clinic or hospital type (clinic or hospital).**
    - **A "Predict" button processes the input data and displays the prediction in a text output field.**
3. **Internal Workflow:**
    - **User inputs are processed through OneHotEncoder to transform categorical features into numerical values.**
    - **A Scaler is applied to normalize numerical values such as medicine price and examination price.**
    - **The pre-trained Adaboost model predicts whether a prescription will be written ("Will Write" or "Will Not Write") based on the processed input.**

This application combines simplicity in design with powerful machine learning capabilities to provide accurate predictions in a medical context.

APPLICATION



CONCLUSION

This project addresses the critical challenges faced by medical representatives by leveraging machine learning to predict a doctor's likelihood of prescribing a specific medication. Through the development of a desktop application powered by an Adaboost model, medical representatives can now make data-driven decisions, optimizing their outreach efforts and minimizing wasted time and resources.

By analyzing key features such as medication details, doctor specialties, and practice settings, the model provides valuable insights that enable representatives to focus on healthcare professionals who are more likely to prescribe their products. This not only enhances efficiency but also improves the alignment of medications with patient needs, ultimately contributing to better healthcare outcomes.

The project demonstrates the power of integrating technology into traditional workflows, paving the way for smarter, more targeted strategies in the pharmaceutical industry.