Webscraping & Analytics of Practo Website

Objective:

In [ ]:
```
To scrape Practo website (20 pages) and get details Doctor Name, Department,
Experience, Hospital, Fees, Waiting time, Recommendation percent and Reviews.
Conduct Exploratory Data Analysis and get meaningful insights.
```

Method

In [ ]:
```
Practo is an Online website that hosts Doctor listings across all major cities in Indi
I have scraped the data of this website for 20 pages by filtering it with 'Best Docto
in 'Kukatpally', 'Hyderabad'.
It is an infinite scroll website, I inspected the developer tools and see the data tra
highest at a particular 'xhr' record.
Fetched the URL, and it hosts the entire info I needed. I converted that webpage url t
JSON object and grabbed the key-value pairs.,
Each page consists of 10 results and thus I analyzed results of 200 results. I extract
the data like
Doctor Name, Department, Experience, Hospital, Fees, Wait time, Recommendation percent
and Reviews. Extracted the data and loaded the data into a
Pandas Dataframe and performed further analysis and visualization.
```

In [2]:
```
baseurl = 'https://www.practo.com/hyderabad/doctors/kukatpally'
```

In [3]:
```
import requests
import pandas as pd
import json
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:
```
url = 'https://www.practo.com/marketplace-api/mweb/provider-seo/v2?page=1&reach_versi
r=requests.get(url)
data=r.json()
#print(data)
info=data['doctors']['entities']
Doctors = [info[key]['doctor_name'] for key in info.keys()]
Exp = [info[key]['experience_years'] for key in info.keys()]
Hospital= [info[key]['practice']['name'] for key in info.keys()]
Department=[info[key].get('specialization_info', {}).get('specialization') for key in
Fees= [info[key]['consultation_fees'] for key in info.keys()]
wait_time=[info[key]['wait_time'] for key in info.keys()]
Recommend_percent=[info[key]['recommendation_percent'] for key in info.keys()]
Reviews=[info[key]['reviews_count'] for key in info.keys()]
#print(Doctors)
#print(Exp)
#print(Hospital)
#print(f'Department: ', Department)
#print(Fees)
#print(f'wait time is ', wait_time)
#print(Recommend_percent)
#print(Reviews)
```

In [5]:
```
for page in range(2,21):
    url = 'https://www.practo.com/marketplace-api/mweb/provider-seo/v2?page={}&reach_v
    r=requests.get(url)
    data=r.json()
    info=data['doctors']['entities']
    Doctors.extend([info[key]['doctor_name'] for key in info.keys()])
    Exp.extend([info[key]['experience_years'] for key in info.keys()])
    Hospital.extend([info[key]['practice']['name'] for key in info.keys()])
    Department.extend([info[key].get('specialization_info', {}).get('specialization')
    Fees.extend([info[key]['consultation_fees'] for key in info.keys()])
    wait_time.extend([info[key]['wait_time'] for key in info.keys()])
    Recommend_percent.extend([info[key]['recommendation_percent'] for key in info.keys
    Reviews.extend([info[key]['reviews_count'] for key in info.keys()])
```

In [6]:
```
print(len(wait_time))
```

```
200
```

In [7]:
```
df=pd.DataFrame()
df['Doctors'] = Doctors
df['Experience'] = Exp
df['Hospital'] = Hospital
df['Department'] = Department
df['Fees'] = Fees
df['Waiting'] = wait_time
df['Recommend_percent'] = Recommend_percent
df['Reviews'] = Reviews
```

In [8]:
```
df
```

Out[8]:

	Doctors	Experience	Hospital	Department	Fees	Waiting	Recommend_percent	R
0	Dr. K. Suma	34	Prasad Infertility Solutions	Gynecologist/Obstetrician	500	15.0	93.0	
1	Dr. Ajay Kumar Paruchuri	19	Ajay Orthopedic Clinic	Spine Surgeon	600	30.0	96.0	
2	Dr. S. Bala Ram Babu	22	Dr. Bala's Sexology Clinic	Sexologist	600	30.0	87.0	
3	Dr. Meka Satyanarayana	40	Aswini Clinic	General Physician	800	60.0	96.0	
4	Dr. Ravinder Raju	24	Positive Homeopathy	Homoeopath	300	30.0	71.0	
...	...	...	...	...	...	...	...	
195	Dr. Arun Kumar	9	Arya Dental	Dentist	200	15.0	100.0	
196	Dr. S.Ziauddhin	10	Arya Dental	Dentist	200	NaN	NaN	
197	Dr. Chandra Sheker M (Physiotherapist)	10	Nightingales Home Health Services	Physiotherapist	650	NaN	NaN	
198	Dr. Divya Krupa N (Physiotherapist)	10	Nightingales Home Health Services	Physiotherapist	650	NaN	NaN	
199	Dr. Sunil Reddy A	16	Idea Clinics	Psychiatrist	1250	NaN	NaN	

200 rows × 8 columns

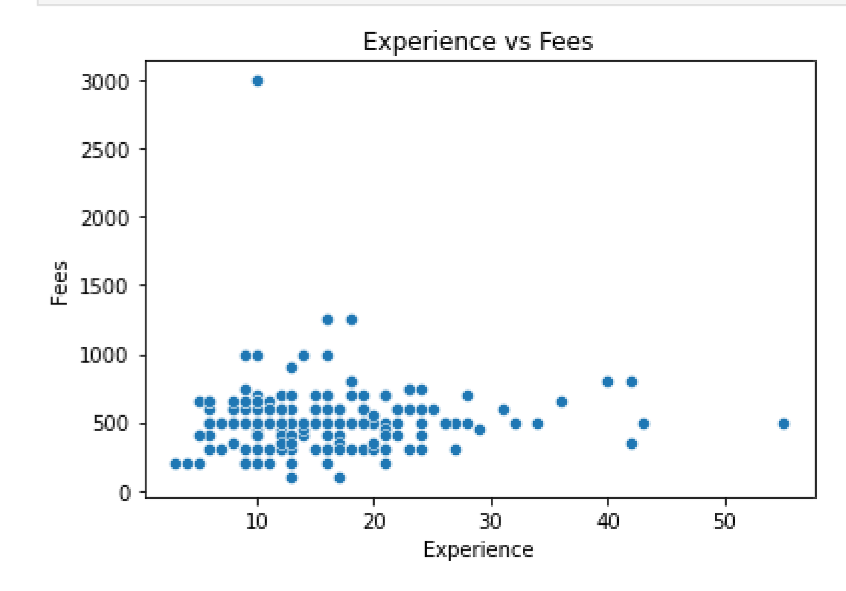
In [10]:
```
df.dtypes
```

Out[10]:
```
Doctors                object
Experience              int64
Hospital               object
Department             object
Fees                    int64
Waiting               float64
Recommend_percent     float64
Reviews                 int64
dtype: object
```

Questions

1. Does Fee vary with Experience? 2. What is the maximum fees and which department is having it? 3. What is the average fee? 4. Which Departments are hosted on Practo webpage? 5. Which Hospitals are hosted on Practo webpage? 6. What is the average fees by department? 7. Top 7 departments by average fees? 8. Plot count of doctors by department? 9. Count of Doctors by Experience?1. Does Fee vary with Experience?

In [13]:
```
sns.scatterplot(x='Experience', y='Fees', data=df);
plt.title('Experience vs Fees');
```



From the above visualization, it is clear that the consultation fee remained almost same for Doctors between 10-20 years

3000 is an outlier. Hence we will exclude that data point in calculating average fees

In [14]:
```
df['Fees'].max()
```

Out[14]:
```
3000
```

In [15]:
```
#Filtering the dataframe without the row containing max fee as 3000
new_df=df[df['Fees']!=3000]
new_df
```

Out[15]:

	Doctors	Experience	Hospital	Department	Fees	Waiting	Recommend_percent	R
0	Dr. K. Suma	34	Prasad Infertility Solutions	Gynecologist/Obstetrician	500	15.0	93.0	
1	Dr. Ajay Kumar Paruchuri	19	Ajay Orthopedic Clinic	Spine Surgeon	600	30.0	96.0	
2	Dr. S. Bala Ram Babu	22	Dr. Bala's Sexology Clinic	Sexologist	600	30.0	87.0	
3	Dr. Meka Satyanarayana	40	Aswini Clinic	General Physician	800	60.0	96.0	
4	Dr. Ravinder Raju	24	Positive Homeopathy	Homoeopath	300	30.0	71.0	
...	...	...	...	...	...	...	...	
195	Dr. Arun Kumar	9	Arya Dental	Dentist	200	15.0	100.0	
196	Dr. S.Ziauddhin	10	Arya Dental	Dentist	200	NaN	NaN	
197	Dr. Chandra Sheker M (Physiotherapist)	10	Nightingales Home Health Services	Physiotherapist	650	NaN	NaN	
198	Dr. Divya Krupa N (Physiotherapist)	10	Nightingales Home Health Services	Physiotherapist	650	NaN	NaN	
199	Dr. Sunil Reddy A	16	Idea Clinics	Psychiatrist	1250	NaN	NaN	

199 rows × 8 columns

2. What is the maximum fees and which department is having it?

In [16]:
```
Max_Fees=new_df['Fees'].max()
print(f'The max fees is ', Max_Fees)
```

```
The max fees is  1250
```

In [17]:
```
df.loc[df.Fees == 1250,'Department']
```

Out[17]:
```
187    Psychiatrist
199    Psychiatrist
Name: Department, dtype: object
```

3. What is the average fee?

In [18]:
```
Average_fee=new_df['Fees'].mean()
print(f'The average fee of the Doctors in the given dataset is: ', Average_fee)
```

```
The average fee of the Doctors in the given dataset is:  503.00502512562815
```

4. Which Departments are hosted on Practo webpage?

In [19]:
```
df['Department'].value_counts()
```

Out[19]:
```
Dentist                                35
General Physician                      17
Gynecologist/Obstetrician              16
Physiotherapist                        13
Orthopedist                            13
Homoeopath                             12
Dermatologist                          11
General Surgeon                        11
Psychiatrist                            6
Ear-Nose-Throat (ENT) Specialist        6
Pediatrician                            5
Pulmonologist                           4
Neurosurgeon                            4
Endocrinologist                         4
Cardiologist                            4
Plastic Surgeon                         3
Neurologist                             3
Speech Therapist                        3
Ayurveda                                3
Gastroenterologist                      2
Psychologist                            2
Hair Transplant Surgeon                 2
Laparoscopic Surgeon                    2
Nephrologist                            2
Spine Surgeon                           2
Sexologist                              2
Oncologist                              2
Radiologist                             1
Endocrine Surgeon                       1
Ophthalmologist                         1
Hematologist                            1
Bariatric Surgeon                       1
Internal Medicine                       1
Dietitian/Nutritionist                  1
Diabetologist                           1
Urologist                               1
Name: Department, dtype: int64
```

5. Which Hospitals are hosted on Practo webpage?

In [20]:
```
df['Hospital'].value_counts()[:10]
```

Out[20]:
```
Apollo Clinic                          12
OMNI Hospitals                         10
Amor Hospitals                          9
Idea Clinics                            9
Nightingales Home Health Services       9
Vitality Hospital                       6
Sree Manju Hospitals                    6
Solitaire Family Dentistry              6
Positive Homeopathy                     4
LN Dental                               4
Name: Hospital, dtype: int64
```

6. What is the average fees by department?

In [21]:
```
df.groupby('Department')['Fees'].mean().sort_values(ascending=False)
```

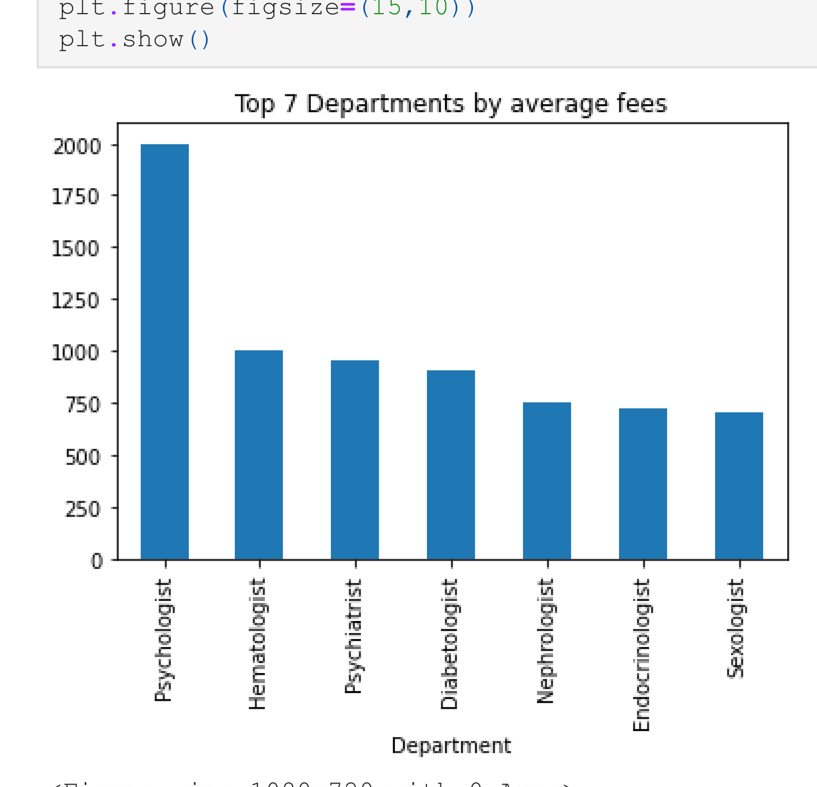
Out[21]:
```
Department
Psychologist                          2000.000000
Hematologist                          1000.000000
Psychiatrist                           950.000000
Diabetologist                          900.000000
Nephrologist                           750.000000
Endocrinologist                        725.000000
Sexologist                             700.000000
Speech Therapist                       650.000000
Ear-Nose-Throat (ENT) Specialist       608.333333
Radiologist                            600.000000
Plastic Surgeon                        600.000000
Hair Transplant Surgeon                600.000000
Dermatologist                          563.636364
Cardiologist                           550.000000
Endocrine Surgeon                      550.000000
Spine Surgeon                          550.000000
Gastroenterologist                     550.000000
Pulmonologist                          550.000000
Physiotherapist                        538.461538
Neurologist                            533.333333
Orthopedist                            526.923077
Neurosurgeon                           525.000000
General Physician                      523.529412
Pediatrician                           520.000000
General Surgeon                        509.000000
Urologist                              500.000000
Laparoscopic Surgeon                   500.000000
Oncologist                             500.000000
Dietitian/Nutritionist                 500.000000
Gynecologist/Obstetrician              478.125000
Ayurveda                               466.666667
Homoeopath                             345.833333
Dentist                                305.714286
Ophthalmologist                        300.000000
Internal Medicine                      200.000000
Bariatric Surgeon                       99.000000
Name: Fees, dtype: float64
```

In [22]:
```
type(df.groupby('Department')['Fees'].mean().sort_values(ascending=False))
```

Out[22]:
```
pandas.core.series.Series
```

In [ ]:
```
7. Top 7 departments by average fees?
```

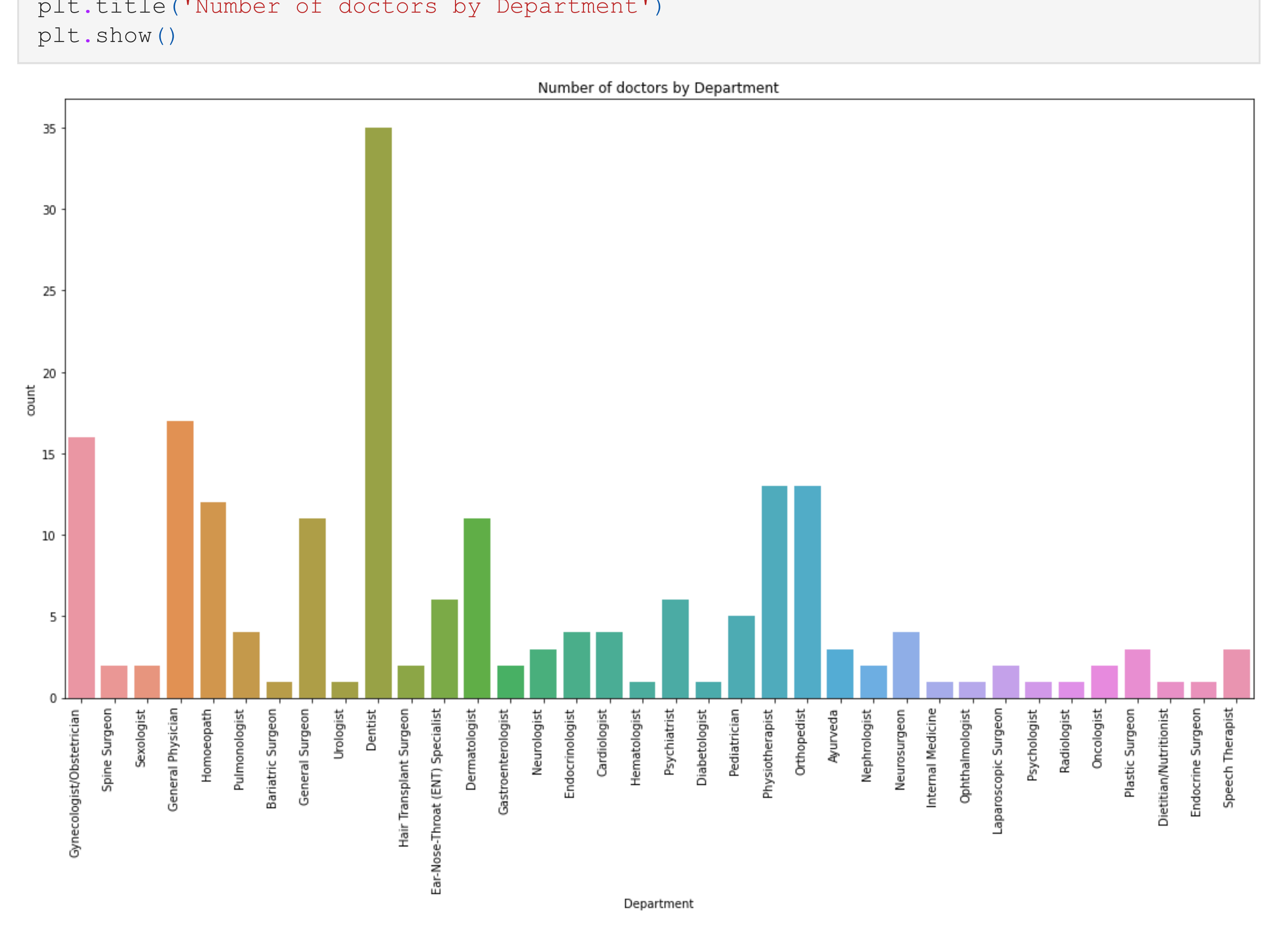
In [26]:
```
(df.groupby('Department')['Fees'].mean().sort_values(ascending=False)[:7].plot(kind='
plt.title('Top 7 Departments by average fees')
plt.figure(figsize=(15,10))
plt.show()
```




```
<Figure size 1080x720 with 0 Axes>
```

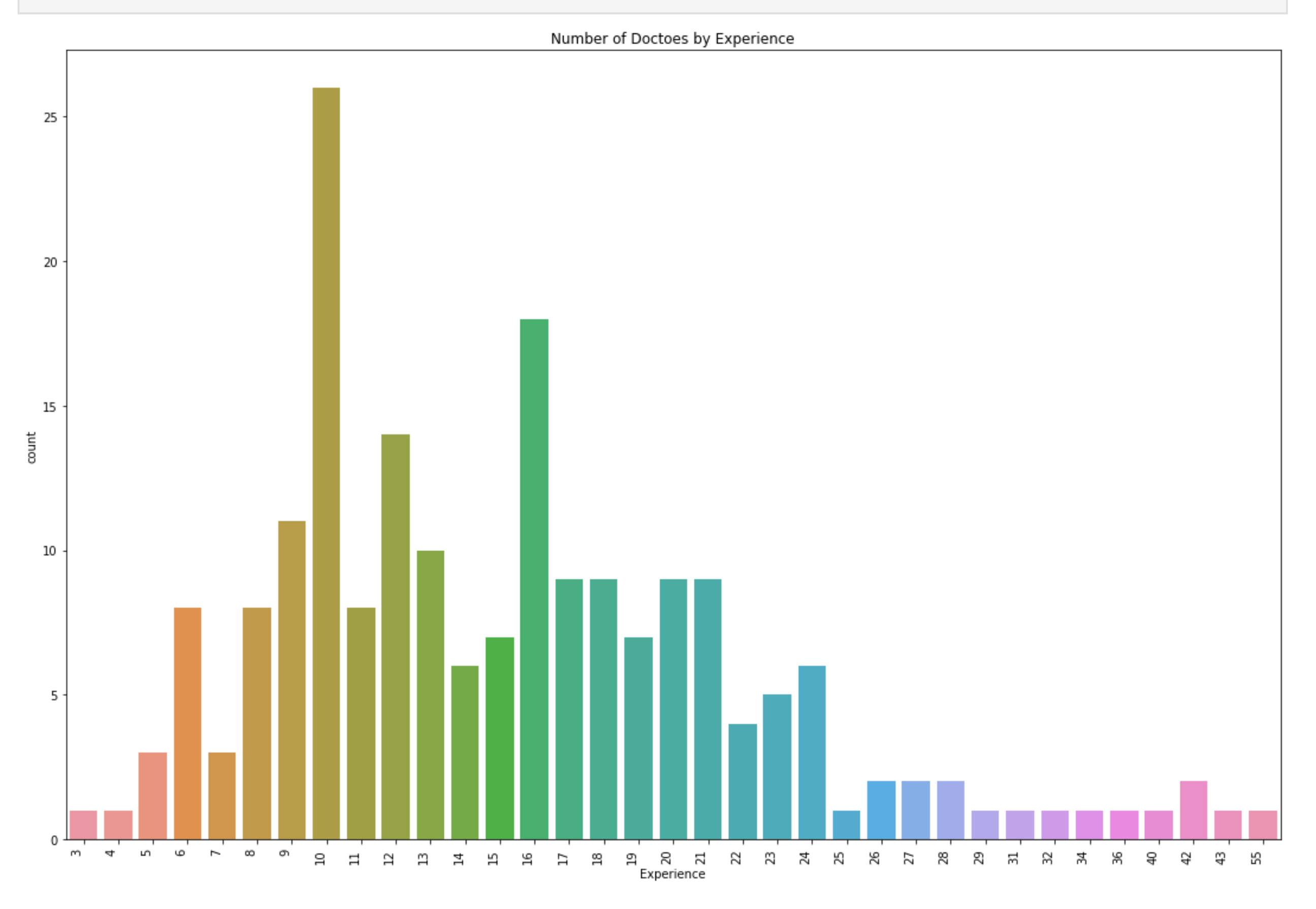
8. Plot count of doctors by department?

In [27]:
```
plt.figure(figsize=(15,10))
ax=sns.countplot(x='Department', data=new_df)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90, ha="right")
plt.tight_layout()
plt.title('Number of doctors by Department')
plt.show()
```



9. Count of Doctors by Experience?

In [28]:
```
plt.figure(figsize=(15,10))
ax=sns.countplot(x='Experience', data=new_df)
ax.set_xticklabels(ax.get_xticklabels(), rotation=90, ha="right")
plt.tight_layout()
plt.title('Number of Doctoes by Experience')
plt.show()
```



Conclusion

The average fees of a doctor as per the given dataset is 503 INR.

Doctor fee remained the same for Doctors with 10 years and 20 years experience

The Maximum fees as per the data is 1250 and Psychiartist Department is charging that fees.

Top 7 departments by average fees are Psychologist, Hematologist, Psychiatrist, Diabetologist, Nephrologist, Endocrinologist, Sexologist

Hospitals hosted on Practo website are Apollo Clinic, OMNI Hospitals, Nightingales Home Health Services, Idea Clinics, Amor Hospitals etc.

Departments hosted on Practo Dentistry, General Physician, Gyneacologist, Physitherapist, Orthopedist, General Surgeon, Dermatologist, Homeopath etc.