

Exploratory Data Analysis (EDA) on Intermediate Colleges of Pakistan

Identifying Performance Trends and Regional
Disparities

Done by Basit Siddiqui

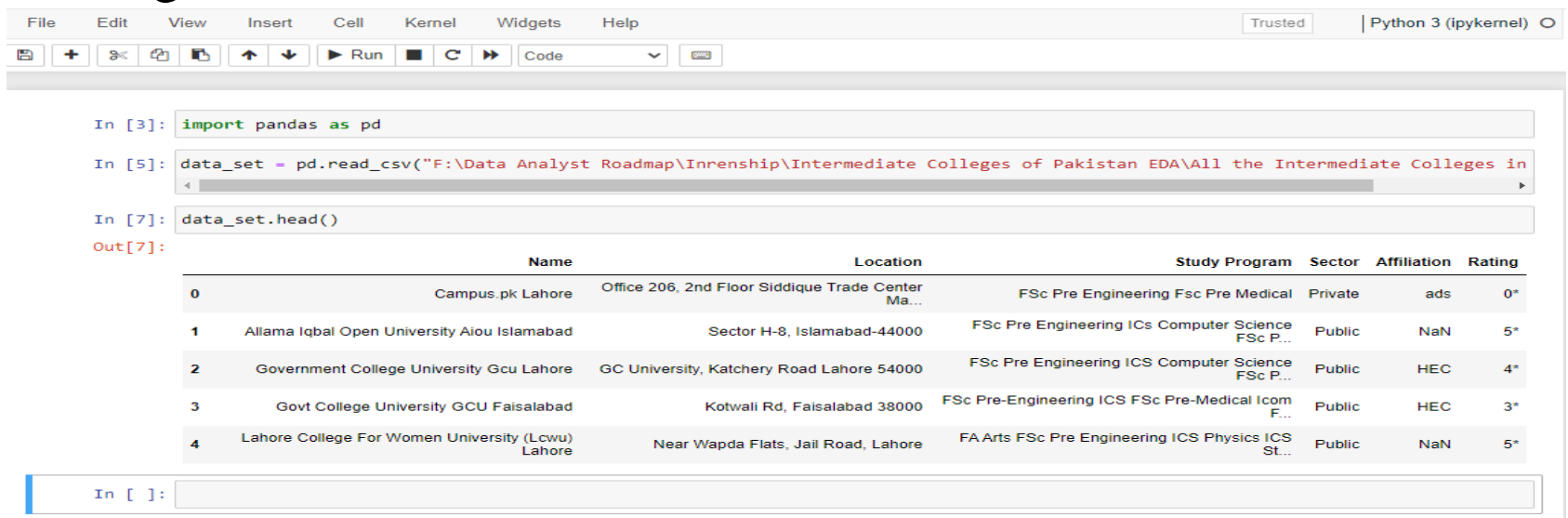
**Internship Project of CodexCue Software
Solutions**

Project Overview

- Analyzed educational data from Intermediate Colleges across Pakistan.
- Used Exploratory Data Analysis (EDA) to identify trends and regional disparities.
- Focused on academic performance, faculty quality, and resource allocation.

Dataset Overview

- Columns: Name, Location, Study Program, Sector, Affiliation, Rating.



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [3]: import pandas as pd
```

```
In [5]: data_set = pd.read_csv("F:\Data Analyst Roadmap\Inrenship\Intermediate Colleges of Pakistan EDA\All the Intermediate Colleges in
```

```
In [7]: data_set.head()
```

Out[7]:

	Name	Location	Study Program	Sector	Affiliation	Rating
0	Campus.pk Lahore	Office 206, 2nd Floor Siddique Trade Center Ma...	FSc Pre Engineering Fsc Pre Medical	Private	ads	0*
1	Allama Iqbal Open University Aiou Islamabad	Sector H-8, Islamabad-44000	FSc Pre Engineering ICs Computer Science FSc P...	Public	NaN	5*
2	Government College University Gcu Lahore	GC University, Katchery Road Lahore 54000	FSc Pre Engineering ICS Computer Science FSc P...	Public	HEC	4*
3	Govt College University GCU Faisalabad	Kotwali Rd, Faisalabad 38000	FSc Pre-Engineering ICS FSc Pre-Medical Icom F...	Public	HEC	3*
4	Lahore College For Women University (Lcwu) Lahore	Near Wapda Flats, Jail Road, Lahore	FA Arts FSc Pre Engineering ICS Physics ICS St...	Public	NaN	5*

In []:

- Data Cleaning: Handled missing values, corrected inconsistencies, and verified data types.

Data Cleaning & Initial Exploration

- Cleaned the dataset by handling missing values and correcting inconsistencies.

Check for the missing values

```
In [8]: missing_values = data_set.isnull().sum()  
print(missing_values)
```

```
Name          0  
Location       4  
Study Program  0  
Sector        529  
Affiliation    1475  
Rating         0  
dtype: int64
```

```
In [ ]: |
```

The Affiliation column has missing values in 90.49% of the data, You can consider removing this column.

Affiliation have 90% missing values so will have to drop this column.

```
In [5]: data_set.drop(columns=['Affiliation'], inplace=True)
```

```
In [6]: data_set.head()
```

Out[6]:

	Name	Location	Study Program	Sector	Rating
0	Campus.pk Lahore	Office 206, 2nd Floor Siddique Trade Center Ma...	FSc Pre Engineering Fsc Pre Medical	Private	0*
1	Allama Iqbal Open University Aiou Islamabad	Sector H-8, Islamabad-44000	FSc Pre Engineering ICS Computer Science FSc P...	Public	5*
2	Government College University Gcu Lahore	GC University, Katchery Road Lahore 54000	FSc Pre Engineering ICS Computer Science FSc P...	Public	4*
3	Govt College University GCU Faisalabad	Kotwali Rd, Faisalabad 38000	FSc Pre-Engineering ICS FSc Pre-Medical Icom F...	Public	3*
4	Lahore College For Women University (Lcwu) Lahore	Near Wapda Flats, Jail Road, Lahore	FA Arts FSc Pre Engineering ICS Physics ICS St...	Public	5*

```
In [ ]:
```

In location column, and sector column have small missing values, fill those values using most frequent value (mode)

```
In [7]: #use mode() to add missing values in location colum
data_set['Location'].fillna(data_set['Location'].mode()[0], inplace=True)
```

```
In [8]: #use mode() to add missing values in Sector column as well.
data_set['Sector'].fillna(data_set['Sector'].mode()[0], inplace=True)
```

```
In [10]: print(data_set.isnull().sum())
```

```
Name      0
Location  0
Study Program  0
Sector     0
Rating     0
dtype: int64
```

```
In [ ]:
```

To solve the location and city problem for analysis, extract city in location column and add one more column in our dataset name (city).

Correct Inconsistencies (if required)

```
In [11]: print(data_set['Location'].unique())
```

```
['Office 206, 2nd Floor Siddique Trade Center Main Boulevard Gulberg III Lahore '
'Sector H-8, Islamabad-44000' 'GC University, Katchery Road Lahore 54000'
... 'Riphah International College Swat'
'Riphah International College Dina' 'Riphah International College Dargai']
```

```
In [ ]:
```

```
In [13]: import re
def extract_city(location):
    cities = {
        'Lahore': r'Lahore',
        'Islamabad': r'Islamabad',
        'Karachi': r'Karachi',
        'Swat': r'Swat',
        'Dina': r'Dina',
        'Dargai': r'Dargai',
    }
    for city, pattern in cities.items():
        if re.search(pattern, location, re.IGNORECASE):
            return city
    return 'Unknown' #for just default/Unknown values

data_set['City'] = data_set['Location'].apply(extract_city)
```

```
In [14]: print(data_set['City'].unique())
```

```
['Lahore' 'Islamabad' 'Unknown' 'Karachi' 'Dina' 'Swat' 'Dargai']
```

```
In [15]: data_set.head()
```

Out[15]:

	Name	Location	Study Program	Sector	Rating	City
0	Campus.pk Lahore	Office 206, 2nd Floor Siddique Trade Center Ma...	FSc Pre Engineering Fsc Pre Medical	Private	0*	Lahore
1	Allama Iqbal Open University Aiou Islamabad	Sector H-8, Islamabad-44000	FSc Pre Engineering ICs Computer Science FSc P...	Public	5*	Islamabad
2	Government College University Gcu Lahore	GC University, Katchery Road Lahore 54000	FSc Pre Engineering ICS Computer Science FSc P...	Public	4*	Lahore
3	Govt College University GCU Faisalabad	Kotwali Rd, Faisalabad 38000	FSc Pre-Engineering ICS FSc Pre-Medical Icom F...	Public	3*	Unknown
4	Lahore College For Women University (Lcwu) Lahore	Near Wapda Flats, Jail Road, Lahore	FA Arts FSc Pre Engineering ICS Physics ICS St...	Public	5*	Lahore

```
In [ ]:
```

Verified data types and performed basic descriptive statistics.

Remove non-numeric characters like '*' and convert to numeric (Rating Column)

```
In [18]: data_set['Rating'] = data_set['Rating'].str.replace('*', '', regex=False).astype(float)
```

```
In [20]: data_set['Rating'] = pd.to_numeric(data_set['Rating'], errors='coerce')
```

```
In [21]: print(data_set['Rating'].describe())
```

```
count    1630.000000
mean       1.118865
std        1.825516
min         0.000000
25%         0.000000
50%         0.000000
75%         2.000000
max         5.000000
Name: Rating, dtype: float64
```

```
In [ ]: |
```

In Rating column have non-numeric data type so it will remove it or change it numeric data type.

Now the dataset is completely clean

Remove non-numeric characters like '*' and convert to numeric (Rating Column)

```
In [18]: data_set['Rating'] = data_set['Rating'].str.replace('*', '', regex=False).astype(float)
```

```
In [20]: data_set['Rating'] = pd.to_numeric(data_set['Rating'], errors='coerce')
```

```
In [21]: print(data_set['Rating'].describe())
```

```
count    1630.000000
mean      1.118865
std       1.825516
min       0.000000
25%      0.000000
50%      0.000000
75%      2.000000
max       5.000000
Name: Rating, dtype: float64
```

```
In [22]: data_set.head()
```

Out[22]:

	Name	Location	Study Program	Sector	Rating	City
0	Campus.pk Lahore	Office 206, 2nd Floor Siddique Trade Center Ma...	FSc Pre Engineering Fsc Pre Medical	Private	0.0	Lahore
1	Allama Iqbal Open University Aiou Islamabad	Sector H-8, Islamabad-44000	FSc Pre Engineering ICS Computer Science FSc P...	Public	5.0	Islamabad
2	Government College University Gcu Lahore	GC University, Katchery Road Lahore 54000	FSc Pre Engineering ICS Computer Science FSc P...	Public	4.0	Lahore
3	Govt College University GCU Faisalabad	Kotwali Rd, Faisalabad 38000	FSc Pre-Engineering ICS FSc Pre-Medical Icom F...	Public	3.0	Unknown
4	Lahore College For Women University (Lcwu) Lahore	Near Wapda Flats, Jail Road, Lahore	FA Arts FSc Pre Engineering ICS Physics ICS St...	Public	5.0	Lahore

```
In [ ]:
```


Academic Results Analysis

- Analyzed academic performance across different regions.

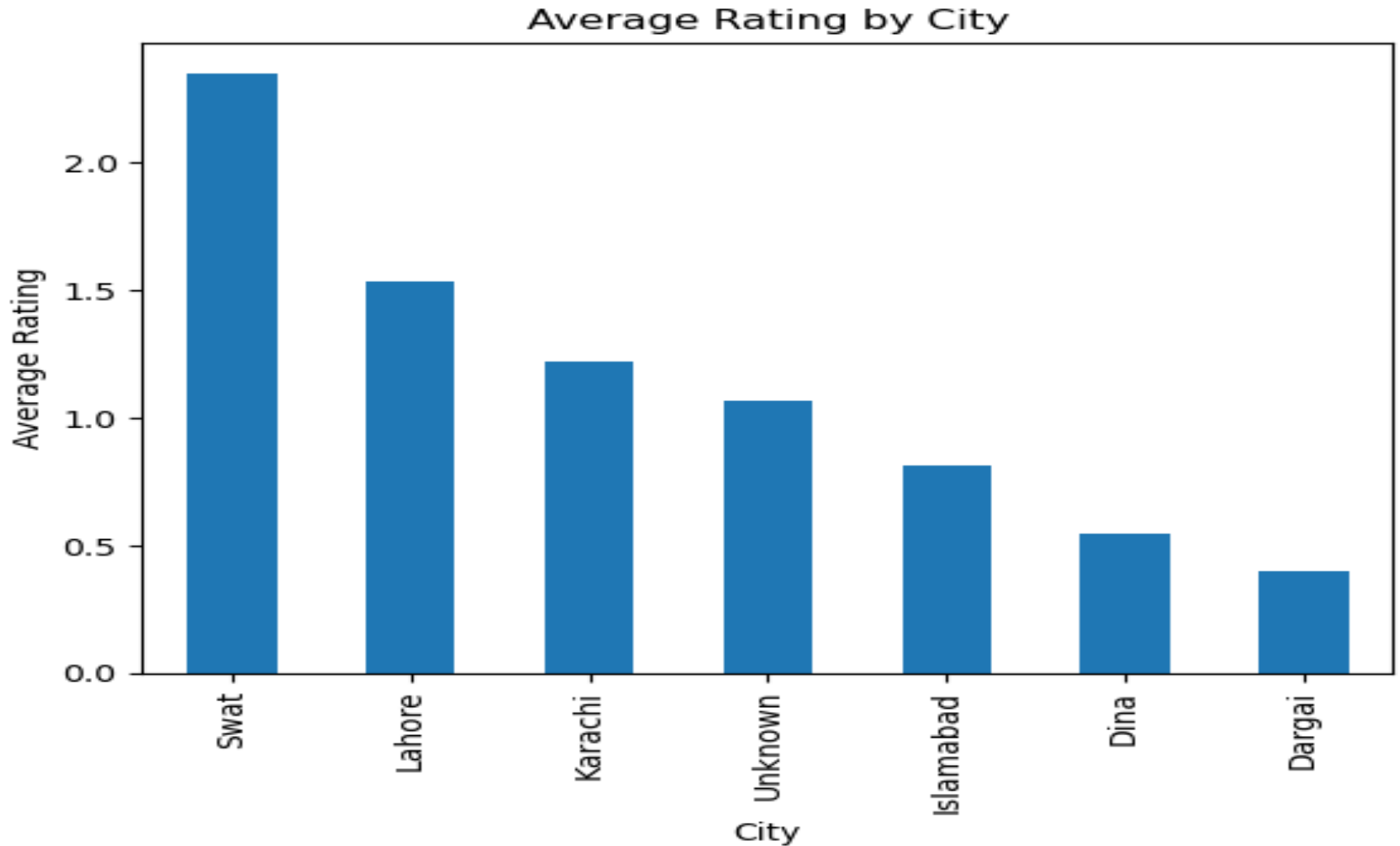
Analyze Ratings Across Regions For Academic Results Analysis

```
In [18]: city_performance = data_set.groupby('City')['Rating'].mean().sort_values(ascending=False)  
print(city_performance)
```

```
City  
Swat      2.350000  
Lahore    1.536036  
Karachi   1.219907  
Unknown   1.065039  
Islamabad 0.811475  
Dina      0.545455  
Dargai    0.400000  
Name: Rating, dtype: float64
```

```
In [ ]:
```

Visualized trends in ratings across cities like Swat, Lahore, and Islamabad etc.



- Identified Swat as the highest performing region and Dargai as the lowest.
- The analysis shows significant regional disparities in ratings. Swat has the highest average rating above 2.0, indicating relatively better performance. In contrast, Dargai has the lowest average rating at 0.4, suggesting potential areas for improvement.

Faculty Quality & Sector Comparison

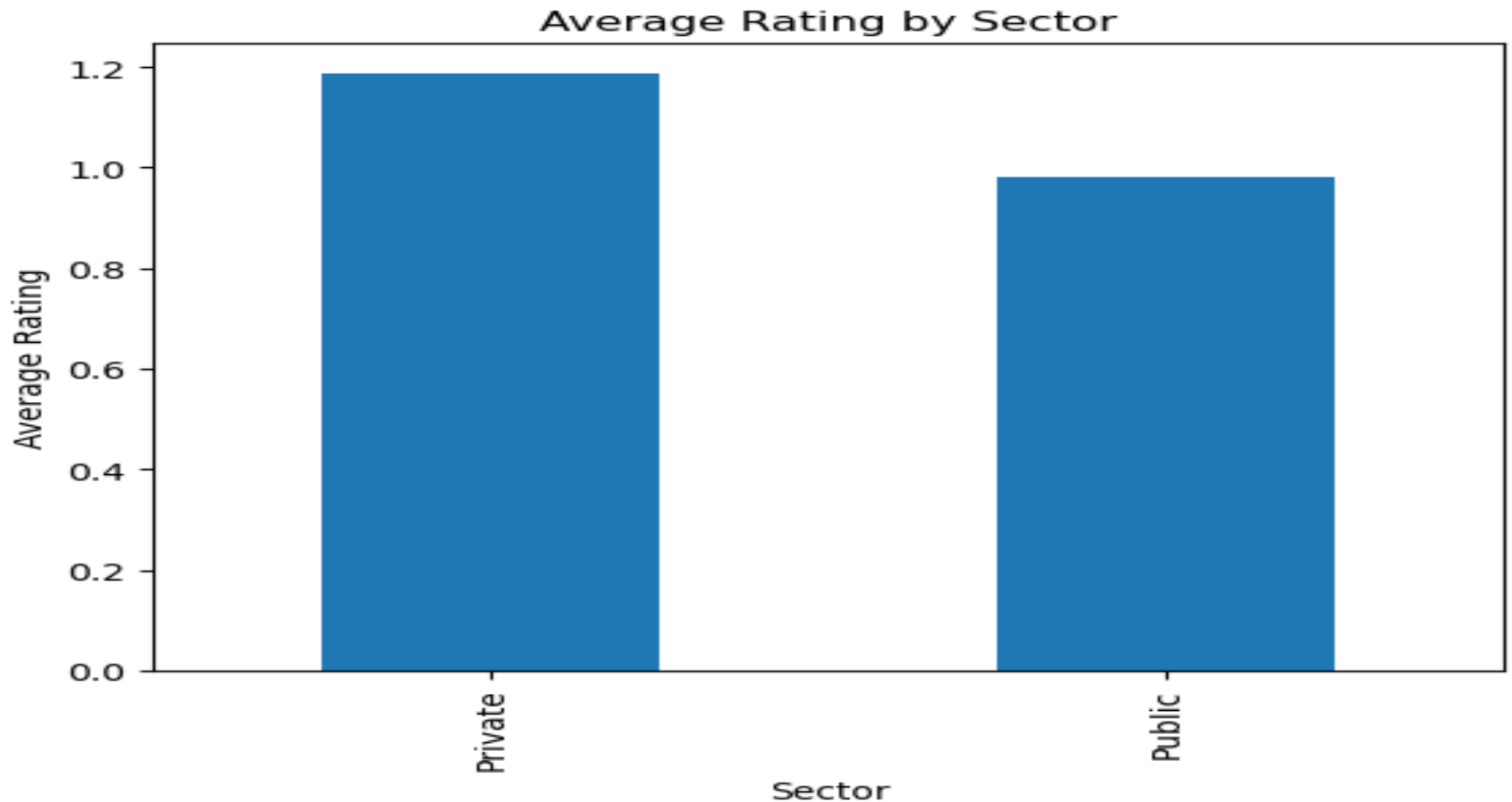
- Compared faculty quality based on sector (Public vs. Private)

Faculty Quality Analysis

```
In [21]: Faculty_sector_performance = data_set.groupby('Sector')['Rating'].mean().sort_values(ascending=False)  
print(Faculty_sector_performance)
```

```
Sector  
Private    1.186755  
Public     0.981447  
Name: Rating, dtype: float64
```

```
In [ ]: |
```



- Found that private institutions had slightly higher ratings overall.
- Noted that lower performance in some regions was not solely due to sector differences.

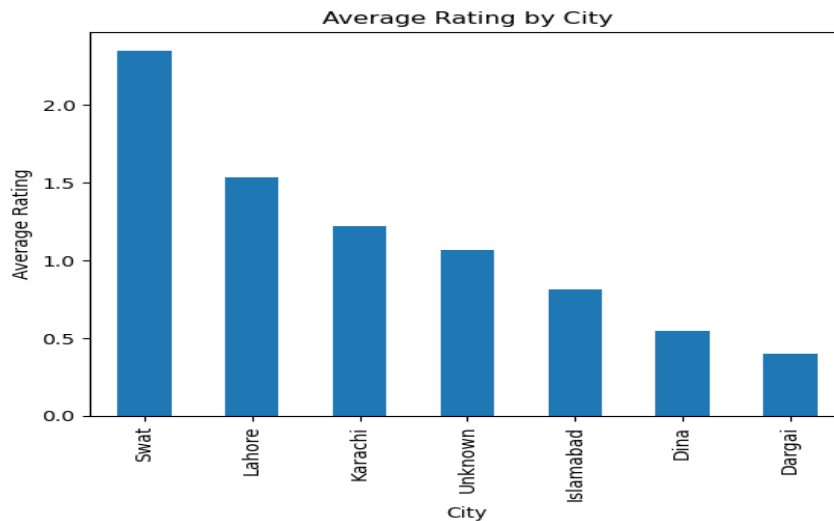
Regional Disparity Analysis

- Compared performance across regions and highlighted disparities.

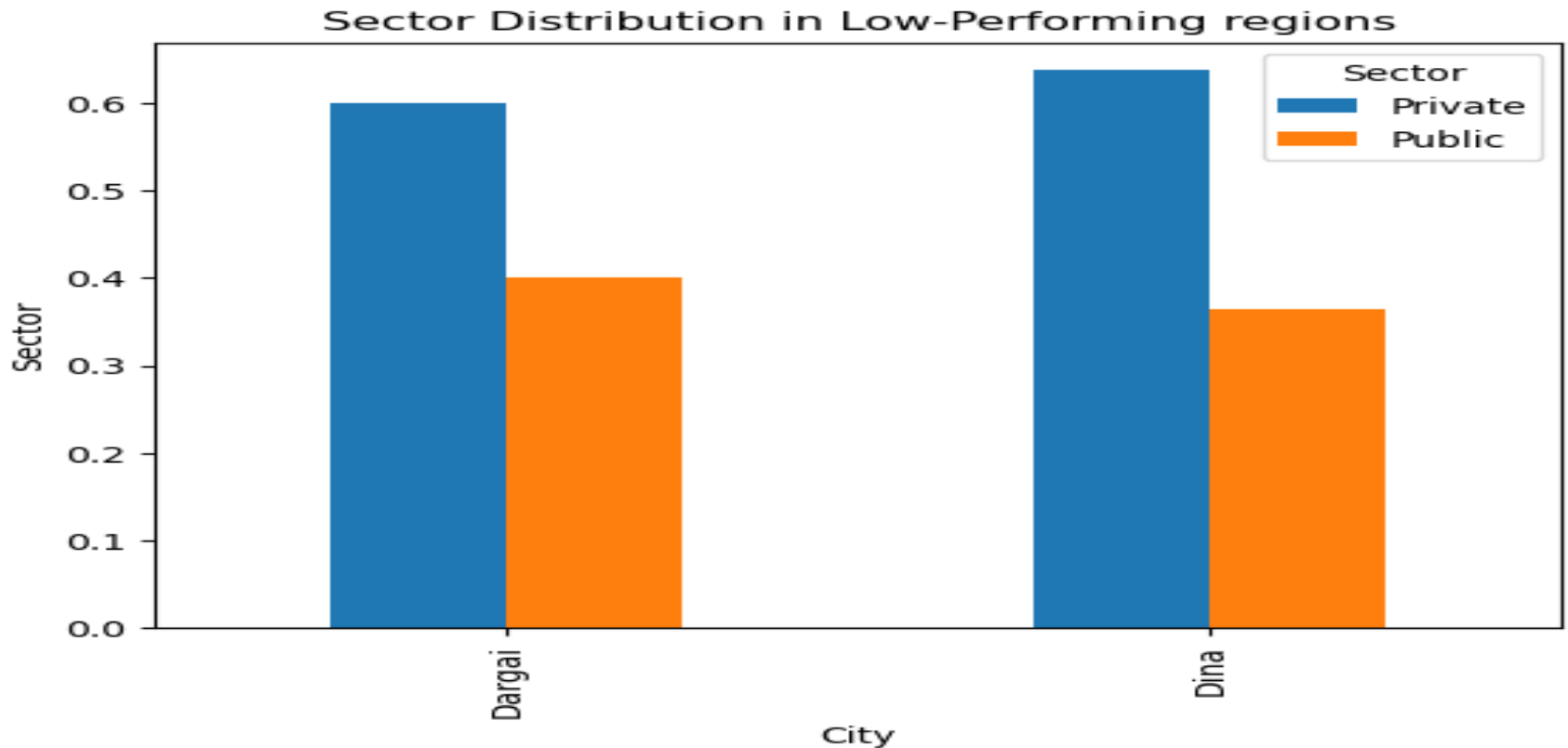
In [28]: `#Compare average ratings by city`

```
city_comparison = data_set.groupby('City')['Rating'].mean().sort_values(ascending=False)
city_comparison.plot(kind='bar', title='Average Rating by City')
plt.ylabel('Average Rating')
```

Out[28]: `Text(0, 0.5, 'Average Rating')`



In [26]:



The analysis shows that Dargai and Dina, which have the lowest ratings, are primarily served by private sector institutions 60% in Dargai and 63.6% in Dina.

- Noted significant underperformance in Dargai and Dina despite private sector dominance.
- Suggested resource allocation and faculty quality as possible factors.

Recommendations

- Improve resource allocation in low-performing regions.
- Implement targeted faculty development programs.
- Replicate successful practices from high-performing regions like Swat.
- Boost public sector performance through better resource management.

Conclusion

- The EDA revealed key regional and sector-based disparities in intermediate education.
- Recommendations focus on addressing these gaps to improve educational outcomes across Pakistan.