DATA ANALYSIS

CASE STUDY ON SCHOOLS OF PUNE DISTICT IN 2019-2020

KAPALLI HEMESH RAJU

2211CS010274

GROUP-4

Introduction to the UDISE+ 2019-20 School Enrollment Dataset

The Unified District Information System for Education Plus (UDISE+) dataset for the year 2019-20 provides comprehensive insights into school enrollment statistics for Pune, Maharashtra. This dataset captures key details such as school locations (urban/rural), management types (government, private, etc.), and student enrollment numbers across different grade levels, from pre-primary to class 12. With 2,024 records and 41 attributes, it offers a detailed view of student distribution, gender ratios, and institutional management patterns. Analyzing this dataset helps understand trends in student enrollment, highlight disparities, and support data-driven decision-making in the education sector.

Description

The CSV file contains school enrollment data for Pune, Maharashtra, for the academic year 2019-20, with 2024 entries and 41 columns. It includes location details (state_name, district_name, udise_block_name), school information (sch_category_id, sch_mgmt_name), and enrollment numbers for boys and girls from pre-primary to class 12. The dataset categorizes schools by management type and tracks student distribution across different grades. The enrollment data is split by gender for each class, providing insights into student demographics across various school blocks in Pune. Let me know if you need specific analysis or visualization.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv("UDISE_plus-19-20_enrol_pune_mah.csv")
df
```

Out[2]:

	ac_year	st_code	state_name	dt_code	district_name	block_cd	udise_block_nam
0	2019-20	27	Maharashtra	2725	PUNE	272501	AMBEGAON
1	2019-20	27	Maharashtra	2725	PUNE	272502	BARAMAT
2	2019-20	27	Maharashtra	2725	PUNE	272503	BHOI
3	2019-20	27	Maharashtra	2725	PUNE	272504	DAUNI
4	2019-20	27	Maharashtra	2725	PUNE	272505	HAVEL
...	...	...	...	...	...	...	.
2019	2019-20	27	Maharashtra	2725	PUNE	272508	KHEI
2020	2019-20	27	Maharashtra	2725	PUNE	272511	PURANDAI
2021	2019-20	27	Maharashtra	2725	PUNE	272505	HAVEL
2022	2019-20	27	Maharashtra	2725	PUNE	272518	Pune Cit
2023	2019-20	27	Maharashtra	2725	PUNE	272507	JUNNAI

2024 rows × 41 columns

Reading .csv file

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2024 entries, 0 to 2023
Data columns (total 41 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   ac_year           2024 non-null   object
 1   st_code           2024 non-null   int64 
 2   state_name        2024 non-null   object
 3   dt_code           2024 non-null   int64 
 4   district_name     2024 non-null   object
 5   block_cd          2024 non-null   int64 
 6   udise_block_name  2024 non-null   object
 7   loc_name          2024 non-null   object
 8   sch_category_id   2024 non-null   int64 
 9   tr_cat_name       2024 non-null   object
 10  school_category   2024 non-null   object
 11  sch_mgmt_id       2024 non-null   int64 
 12  sch_mgmt_name     2024 non-null   object
 13  caste_id          2024 non-null   int64 
 14  caste_name        2024 non-null   object
 15  pre_primary_boy   2024 non-null   int64 
 16  pre_primary_girl  2024 non-null   int64 
 17  class1_boy        2024 non-null   int64 
 18  class2_boy        2024 non-null   int64 
 19  class3_boy        2024 non-null   int64 
 20  class4_boy        2024 non-null   int64 
 21  class5_boy        2024 non-null   int64 
 22  class6_boy        2024 non-null   int64 
 23  class7_boy        2024 non-null   int64 
 24  class8_boy        2024 non-null   int64 
 25  class9_boy        2024 non-null   int64 
 26  class10_boy       2024 non-null   int64 
 27  class11_boy       2024 non-null   int64 
 28  class12_boy       2024 non-null   int64 
 29  class1_girl       2024 non-null   int64 
 30  class2_girl       2024 non-null   int64 
 31  class3_girl       2024 non-null   int64 
 32  class4_girl       2024 non-null   int64 
 33  class5_girl       2024 non-null   int64 
 34  class6_girl       2024 non-null   int64 
 35  class7_girl       2024 non-null   int64 
 36  class8_girl       2024 non-null   int64 
 37  class9_girl       2024 non-null   int64 
 38  class10_girl      2024 non-null   int64 
 39  class11_girl      2024 non-null   int64 
 40  class12_girl      2024 non-null   int64 
dtypes: int64(32), object(9)
memory usage: 648.4+ KB
```

file info

In [4]:
```
df = df.drop_duplicates()
df = df.dropna()
```

->cleaning the data

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2024 entries, 0 to 2023
Data columns (total 41 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   ac_year           2024 non-null   object
 1   st_code           2024 non-null   int64 
 2   state_name        2024 non-null   object
 3   dt_code           2024 non-null   int64 
 4   district_name     2024 non-null   object
 5   block_cd          2024 non-null   int64 
 6   udise_block_name  2024 non-null   object
 7   loc_name          2024 non-null   object
 8   sch_category_id   2024 non-null   int64 
 9   tr_cat_name       2024 non-null   object
 10  school_category   2024 non-null   object
 11  sch_mgmt_id       2024 non-null   int64 
 12  sch_mgmt_name     2024 non-null   object
 13  caste_id          2024 non-null   int64 
 14  caste_name        2024 non-null   object
 15  pre_primary_boy   2024 non-null   int64 
 16  pre_primary_girl  2024 non-null   int64 
 17  class1_boy        2024 non-null   int64 
 18  class2_boy        2024 non-null   int64 
 19  class3_boy        2024 non-null   int64 
 20  class4_boy        2024 non-null   int64 
 21  class5_boy        2024 non-null   int64 
 22  class6_boy        2024 non-null   int64 
 23  class7_boy        2024 non-null   int64 
 24  class8_boy        2024 non-null   int64 
 25  class9_boy        2024 non-null   int64 
 26  class10_boy       2024 non-null   int64 
 27  class11_boy       2024 non-null   int64 
 28  class12_boy       2024 non-null   int64 
 29  class1_girl       2024 non-null   int64 
 30  class2_girl       2024 non-null   int64 
 31  class3_girl       2024 non-null   int64 
 32  class4_girl       2024 non-null   int64 
 33  class5_girl       2024 non-null   int64 
 34  class6_girl       2024 non-null   int64 
 35  class7_girl       2024 non-null   int64 
 36  class8_girl       2024 non-null   int64 
 37  class9_girl       2024 non-null   int64 
 38  class10_girl      2024 non-null   int64 
 39  class11_girl      2024 non-null   int64 
 40  class12_girl      2024 non-null   int64 
dtypes: int64(32), object(9)
memory usage: 648.4+ KB
```

convert colums to string

In [6]:
```
df['ac_year'] = df['ac_year'].astype(str)
df['district_name'] = df['district_name'].astype(str)
df['school_category'] = df['school_category'].astype(str)
```

Making columns for total no of boys and girls

In [7]:
```
df['total_boys'] = df[[col for col in df.columns if 'boy' in col]].sum(axis=1)
df['total_girls'] = df[[col for col in df.columns if 'girl' in col]].sum(axis=1)
df['total_enrollment'] = df['total_boys'] + df['total_girls']
```

In [8]:
```
df['total_enrollment']
```

Out[8]:
```
0       1025
1       1923
2       1065
3       4276
4       5082
        ...
2019      14
2020       7
2021       0
2022       0
2023       0
Name: total_enrollment, Length: 2024, dtype: int64
```

In [9]:
```
df['total_boys']
```

Out[9]:
```
0        525
1        932
2        562
3       2295
4       2589
        ...
2019       0
2020       7
2021       0
2022       0
2023       0
Name: total_boys, Length: 2024, dtype: int64
```

In [10]:
```
df['total_girls']
```

Out[10]:
```
0        500
1        991
2        503
3       1981
4       2493
        ...
2019      14
2020       0
2021       0
2022       0
2023       0
Name: total_girls, Length: 2024, dtype: int64
```

In [11]:
```
df
```

Out[11]:

	ac_year	st_code	state_name	dt_code	district_name	block_cd	udise_block_nam
0	2019-20	27	Maharashtra	2725	PUNE	272501	AMBEGAON
1	2019-20	27	Maharashtra	2725	PUNE	272502	BARAMAT
2	2019-20	27	Maharashtra	2725	PUNE	272503	BHOI
3	2019-20	27	Maharashtra	2725	PUNE	272504	DAUNI
4	2019-20	27	Maharashtra	2725	PUNE	272505	HAVEL
...	...	...	...	...	...	...	.
2019	2019-20	27	Maharashtra	2725	PUNE	272508	KHEI
2020	2019-20	27	Maharashtra	2725	PUNE	272511	PURANDAI
2021	2019-20	27	Maharashtra	2725	PUNE	272505	HAVEL
2022	2019-20	27	Maharashtra	2725	PUNE	272518	Pune Cit
2023	2019-20	27	Maharashtra	2725	PUNE	272507	JUNNAI

2024 rows × 44 columns

In [12]:
```
u=df[['udise_block_name','total_enrollment','total_boys','total_girls']]
u
```

Out[12]:

	udise_block_name	total_enrollment	total_boys	total_girls
0	AMBEGAON	1025	525	500
1	BARAMATI	1923	932	991
2	BHOR	1065	562	503
3	DAUND	4276	2295	1981
4	HAVELI	5082	2589	2493
...	...	...	...	...
2019	KHED	14	0	14
2020	PURANDAR	7	7	0
2021	HAVELI	0	0	0
2022	Pune City	0	0	0
2023	JUNNAR	0	0	0

2024 rows × 4 columns

-> it gives us the total no of students in 'udise_block_name'.(Area)

->it gives us the total no of boys in 'udise_block_name'.(Area)

->it gives us the total no of girls in 'udise_block_name'.(Area)

In [13]:
```
s = df[['total_boys', 'total_girls', 'total_enrollment']].describe()
print(s)
```

```
        total_boys   total_girls  total_enrollment
count   2024.000000   2024.000000       2024.000000
mean     573.855237    499.026680       1072.881917
std     1359.564479   1140.338777       2489.982811
min        0.000000      0.000000          0.000000
25%       23.000000     19.000000         43.000000
50%      112.000000     92.000000        207.000000
75%      479.250000    425.000000        903.500000
max    21360.000000  16618.000000      37978.000000
```

In [14]:
```
a=df['total_boys'].sum()
a
```

Out[14]: 1161483

-> There are total '11,61,483' boys are enrolled in pune(district)

In [15]:
```
b=df['total_girls'].sum()
b
```

Out[15]: 1010030

-> There are total '10,10,030' girls are enrolled in school

In [16]:
```
t=df['total_enrollment'].sum()
t
```

Out[16]: 2171513

-> There are total '21,71,513' student enrolled in school in pune district

In [17]:
```
per_boys=(a/t)*100
print("Precentage of boys:",per_boys)
per_grils=(b/t)*100
print("Precentage of girls:",per_grils)
```

```
Precentage of boys: 53.48726901473765
Precentage of girls: 46.51273098526235
```

-> In pune, there are 53% of boys and 47% of girls in schools

In [18]:
```
diff=a-b
per_diff=(diff/t)*100
print('Diffrents between boys and girls',diff)
print("Precentage of differents betweens boys and girls:",per_diff)
```

```
Diffrents between boys and girls 151453
Precentage of differents betweens boys and girls: 6.974538029475301
```

-> There are nearly 7% more boys 'student' in pune

In [19]:
```
e=df['sch_mgmt_name'].value_counts()
e
```

Out[19]:
```
sch_mgmt_name
Private Unaided (Recognized)                          820
Government Aided                                      684
Local body                                            316
Unrecognized                                           96
Tribal Welfare Department                              32
Social welfare Department                              28
Kendriya Vidyalaya / Central School                    28
Department of Education                                 8
Madarsa recognized (by Wakf board/Madarsa Board)        4
Other Central Govt. Schools                             4
Jawahar Navodaya Vidyalaya                              4
Name: count, dtype: int64
```

->There are 820 private schools in pune

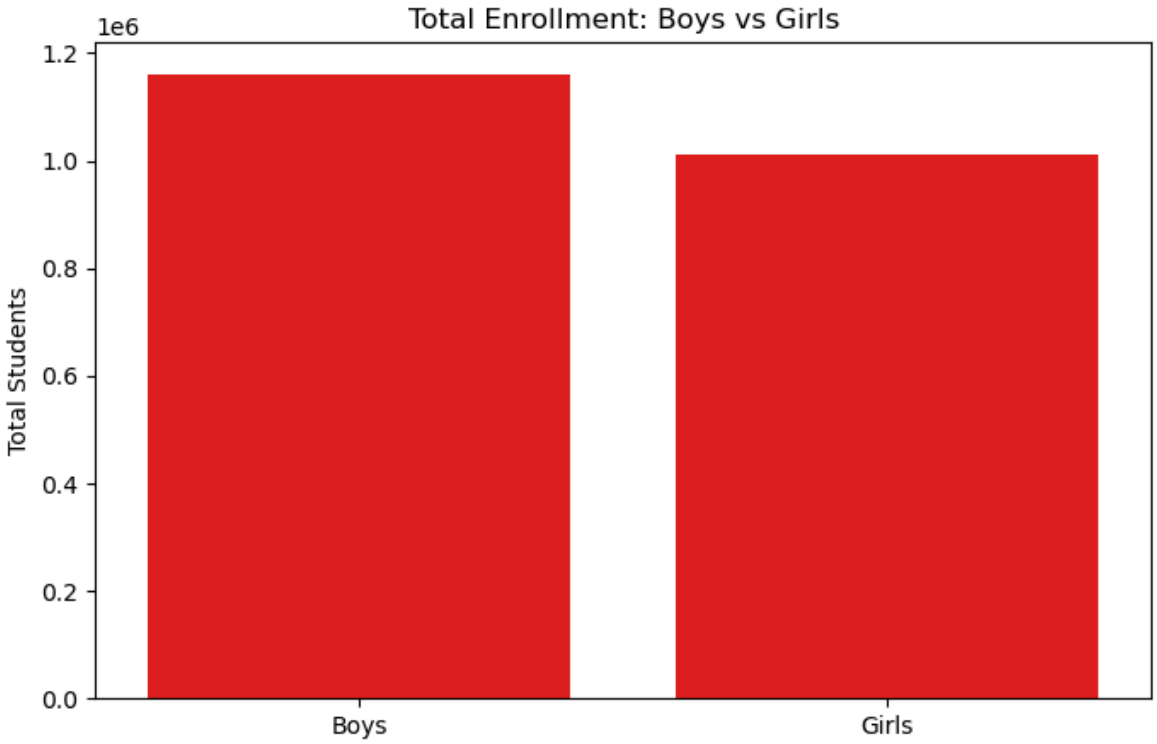
In [20]:
```
f=df['loc_name'].value_counts()
f
```

Out[20]:
```
loc_name
Urban    1144
Rural     880
Name: count, dtype: int64
```

Data Visualzation

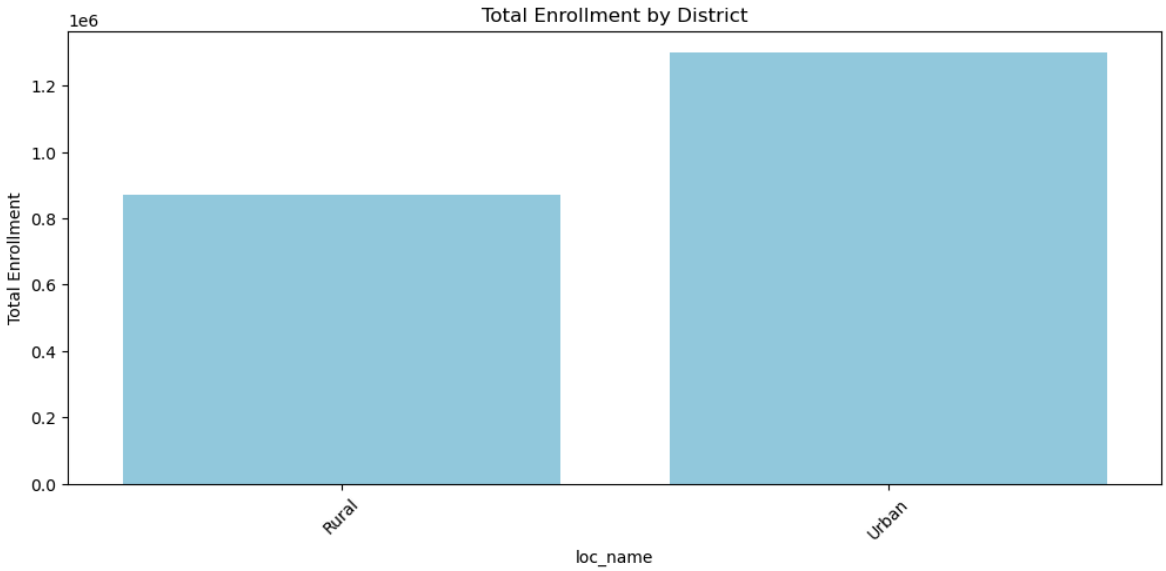
Seaborn

In [21]:
```
plt.figure(figsize=(8, 5))
sns.barplot(x=['Boys', 'Girls'], y=[df['total_boys'].sum(), df['total_girls'].su
plt.title('Total Enrollment: Boys vs Girls')
plt.ylabel('Total Students')
plt.show()
```



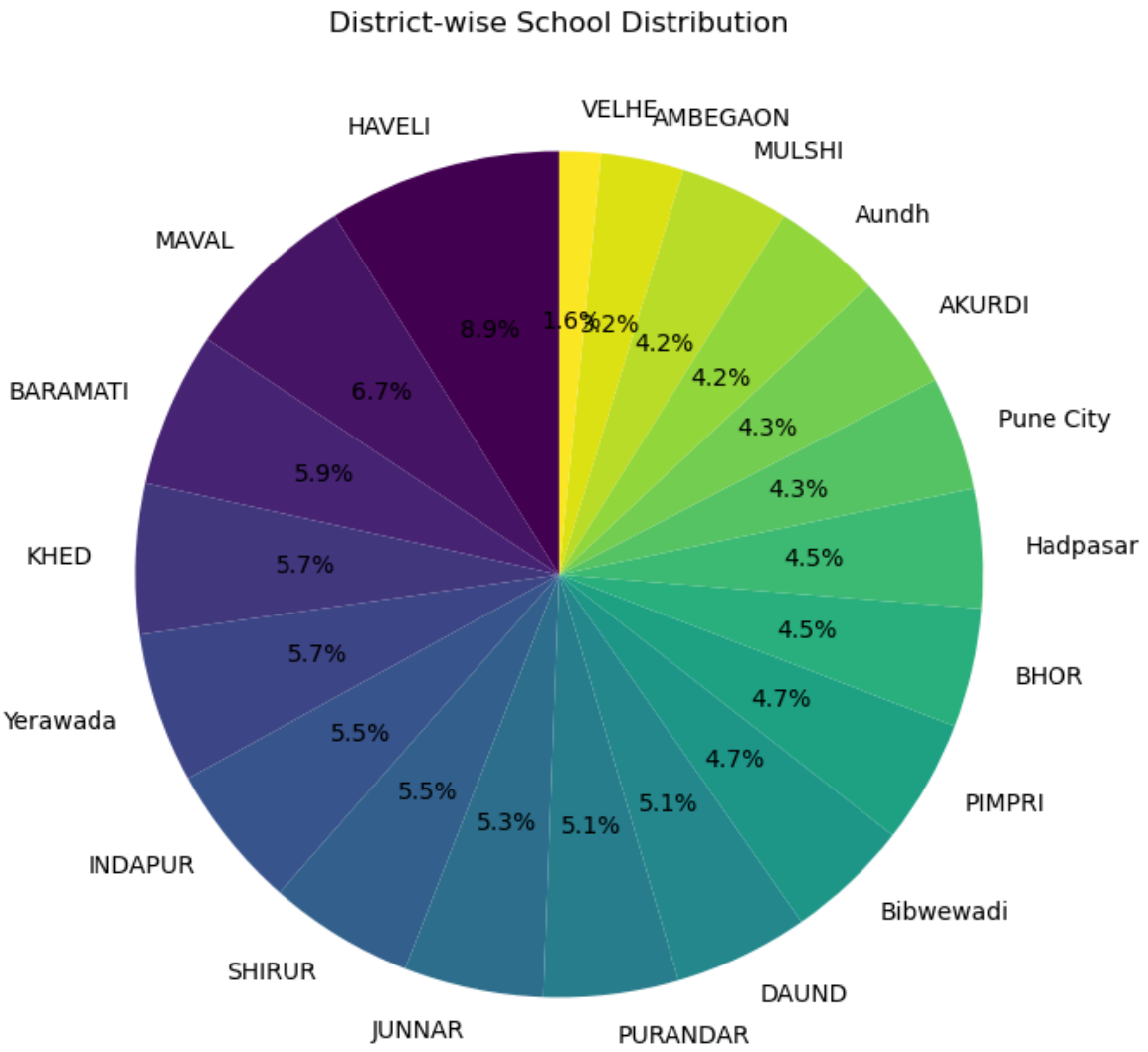
-> There are more Boys then Girls

In [22]:
```
district_enrollment = df.groupby('loc_name')['total_enrollment'].sum().reset_ind
plt.figure(figsize=(12, 5))
sns.barplot(data=district_enrollment, x='loc_name', y='total_enrollment',color='
plt.title('Total Enrollment by District')
plt.xlabel('loc_name')
plt.ylabel('Total Enrollment')
plt.xticks(rotation=45)
plt.show()
```



-> urban has more schools

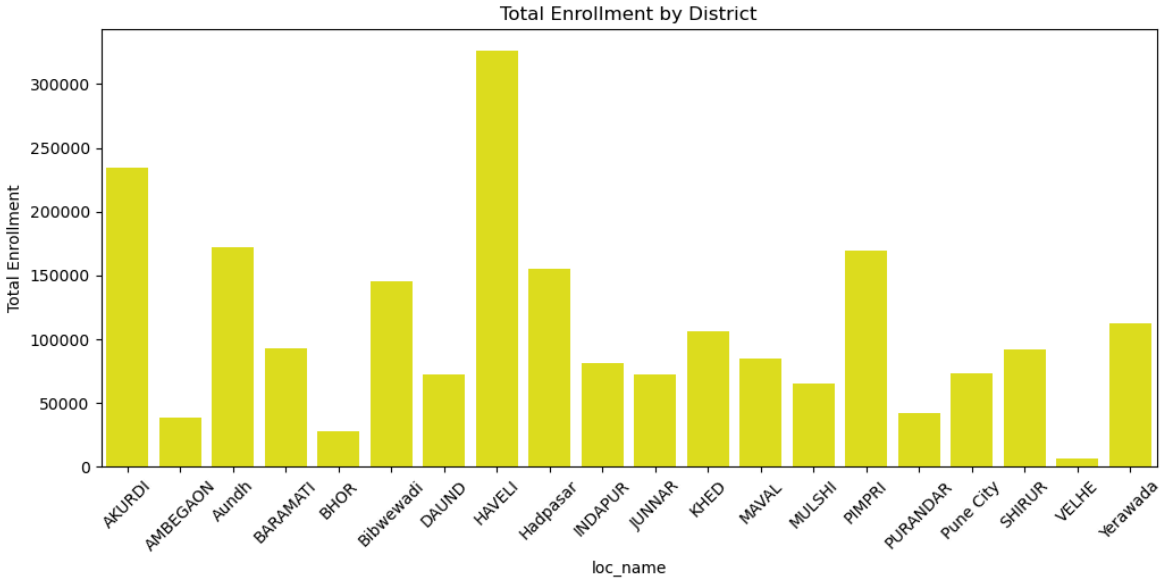
In [23]:
```
district_counts = df['udise_block_name'].value_counts()
plt.figure(figsize=(8, 8))
district_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90, colormap='vir
plt.title('District-wise School Distribution')
plt.ylabel('')
plt.show()
```



-> HAVELI has more schools in pune.

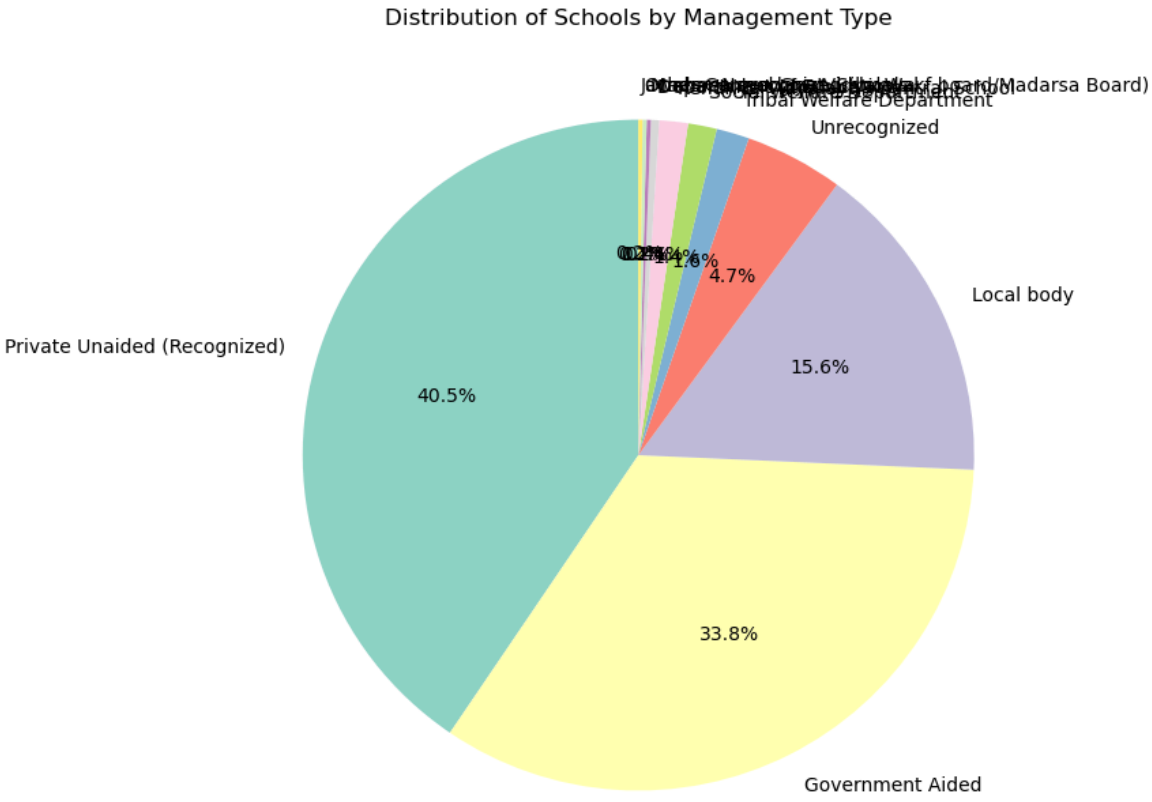
In [ ]:
```
```

In [24]:
```
district_enrollment = df.groupby('udise_block_name')['total_enrollment'].sum().r
plt.figure(figsize=(12, 5))
sns.barplot(data=district_enrollment, x='udise_block_name', y='total_enrollment'
plt.title('Total Enrollment by District')
plt.xlabel('loc_name')
plt.ylabel('Total Enrollment')
plt.xticks(rotation=45)
plt.show()
```



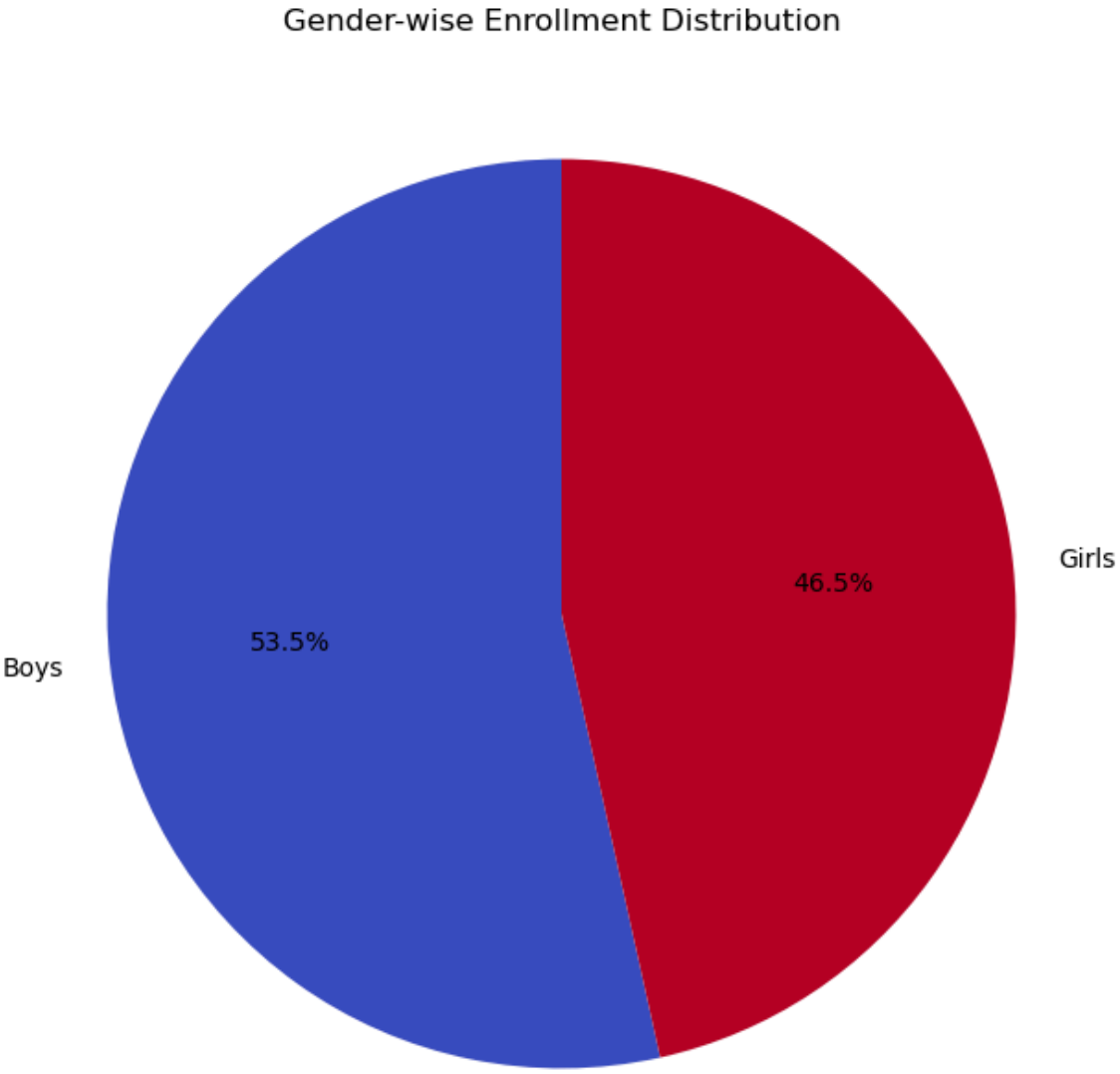
-> area wise total enrollment

In [25]:
```
def plot_pie_chart(data, title):
    plt.figure(figsize=(8, 8))
    data.plot(kind='pie', autopct='%1.1f%%', startangle=90, colormap='Set3')
    plt.title(title)
    plt.ylabel('')
    plt.show()
sch_mgmt_counts = df['sch_mgmt_name'].value_counts()
plot_pie_chart(sch_mgmt_counts, 'Distribution of Schools by Management Type')
```



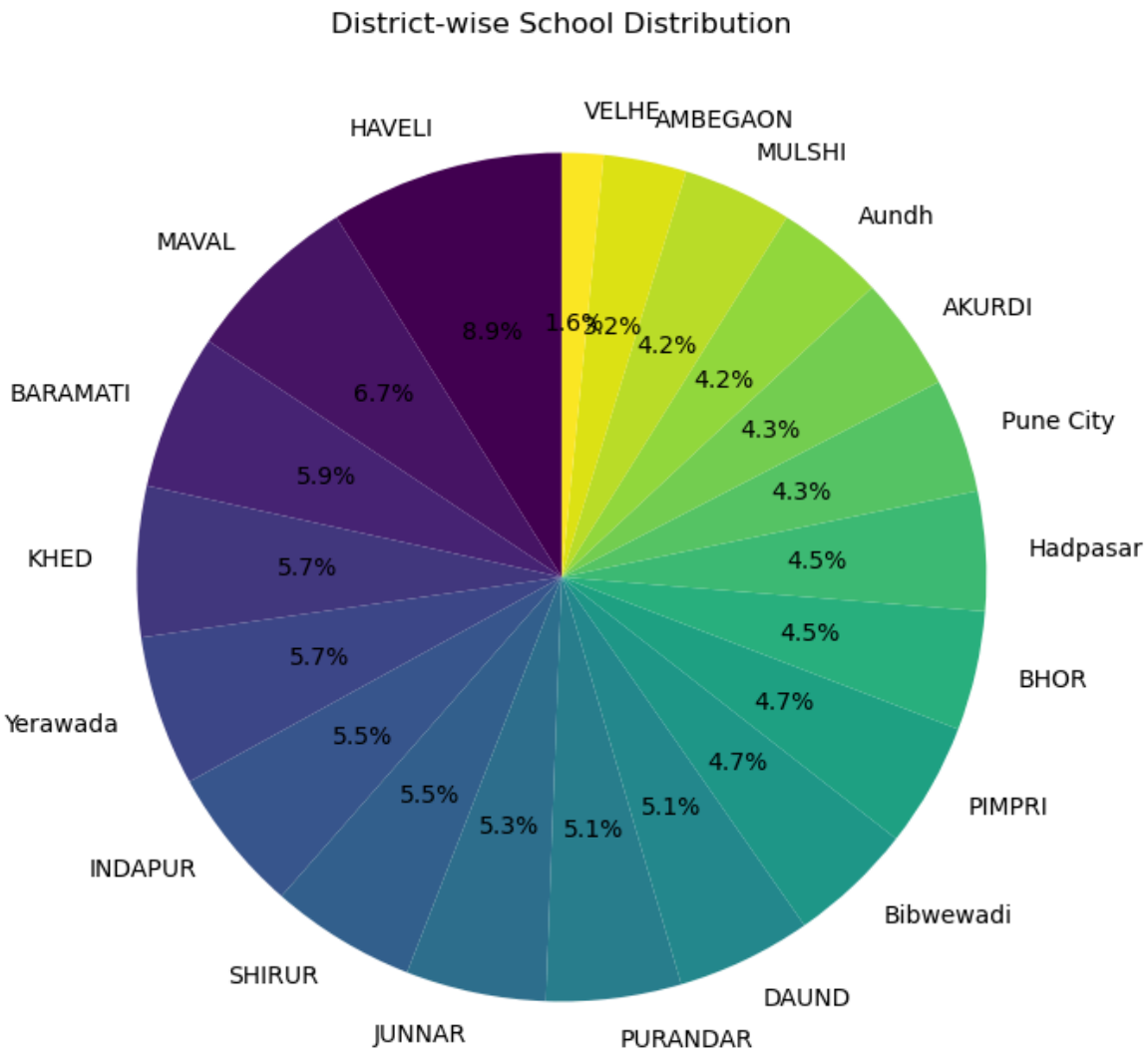
-> there are more private schools in pune.

In [26]:
```
gender_counts = pd.Series({'Boys': df['total_boys'].sum(), 'Girls': df['total_gi
plt.figure(figsize=(8, 8))
gender_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90, colormap='coolw
plt.title('Gender-wise Enrollment Distribution')
plt.ylabel('')
plt.show()
```

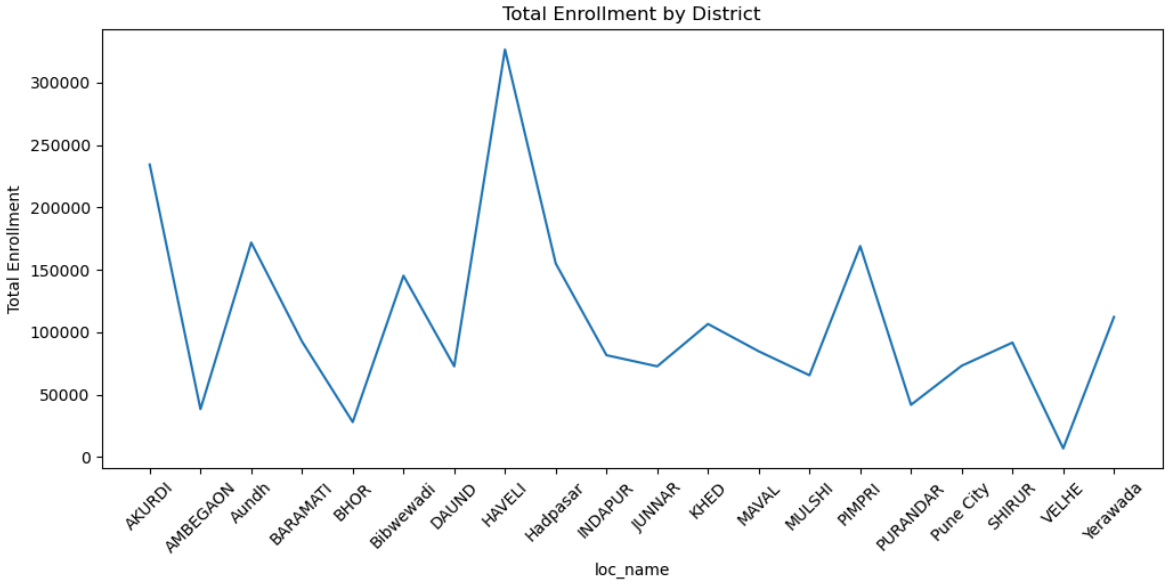


-> There are more boys in schools.

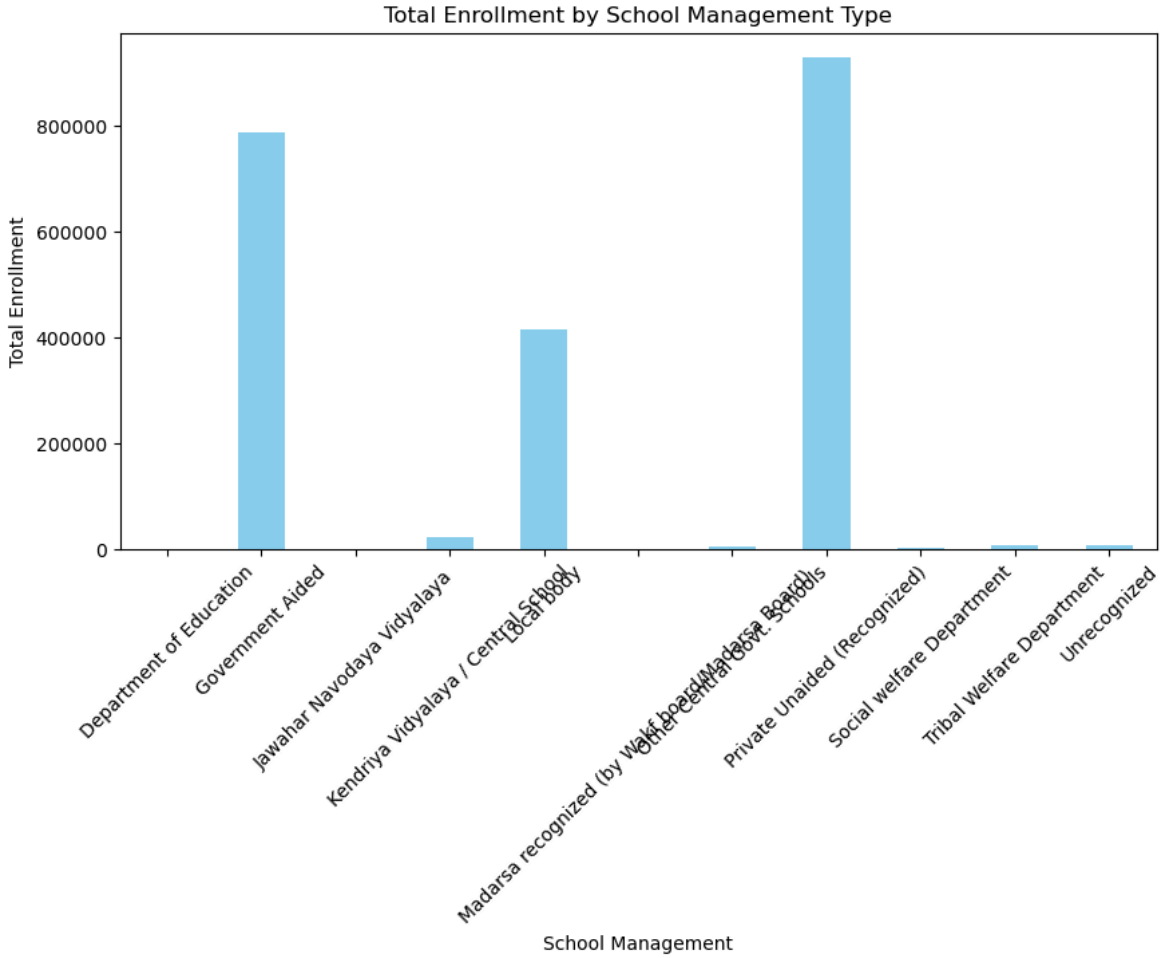
In [27]:
```
district_counts = df['udise_block_name'].value_counts()
plt.figure(figsize=(8, 8))
district_counts.plot(kind='pie', autopct='%1.1f%%', startangle=90, colormap='vir
plt.title('District-wise School Distribution')
plt.ylabel('')
plt.show()
```



In [28]:
```
district_enrollment = df.groupby('udise_block_name')['total_enrollment'].sum().r
plt.figure(figsize=(12, 5))
sns.lineplot(data=district_enrollment, x='udise_block_name', y='total_enrollment
plt.title('Total Enrollment by District')
plt.xlabel('loc_name')
plt.ylabel('Total Enrollment')
plt.xticks(rotation=45)
plt.show()
```

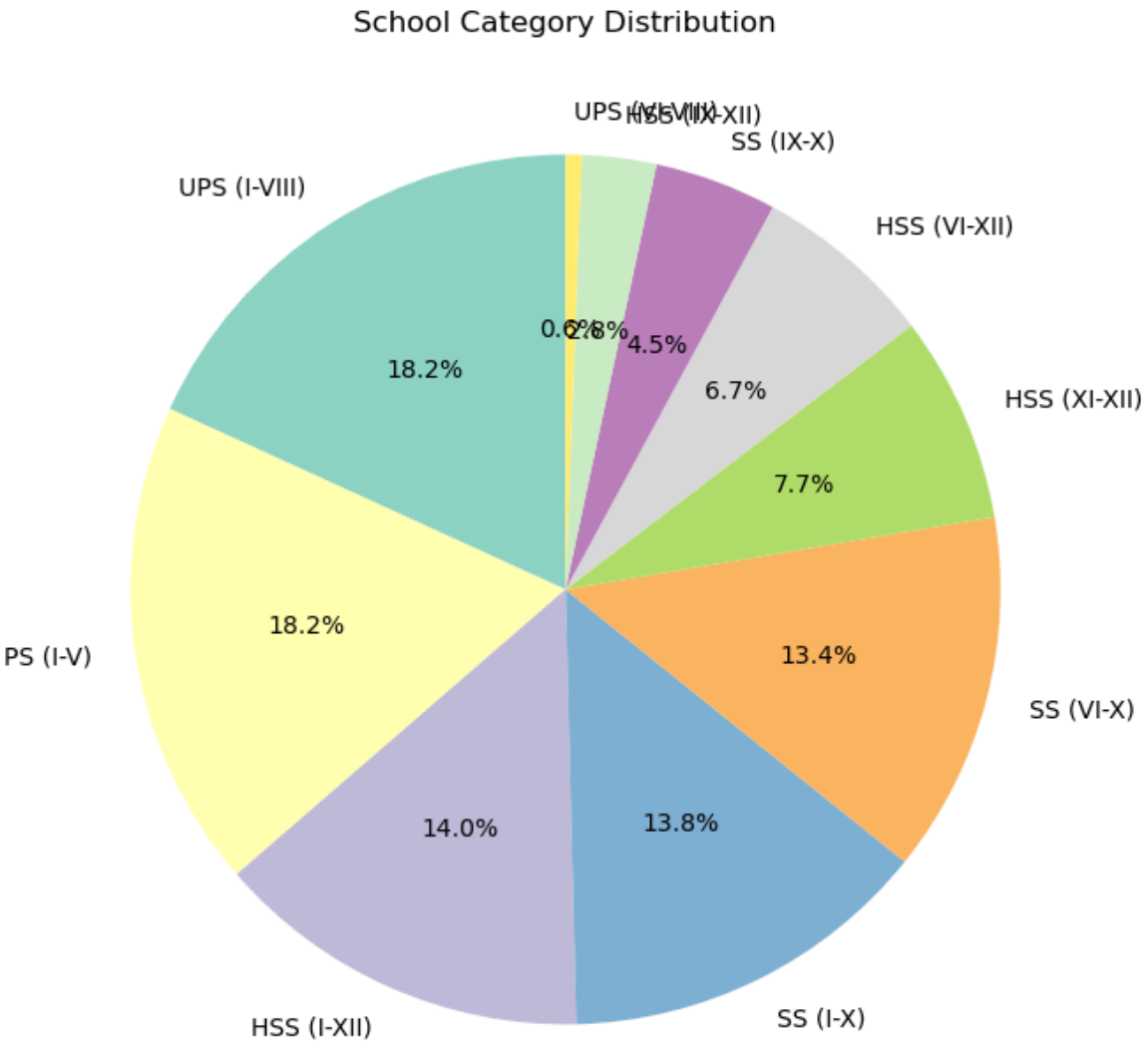


In [29]:
```
plt.figure(figsize=(10, 5))
df.groupby('sch_mgmt_name')['total_enrollment'].sum().plot(kind='bar', color='sk
plt.title('Total Enrollment by School Management Type')
plt.xlabel('School Management')
plt.ylabel('Total Enrollment')
plt.xticks(rotation=45)
plt.show()
```

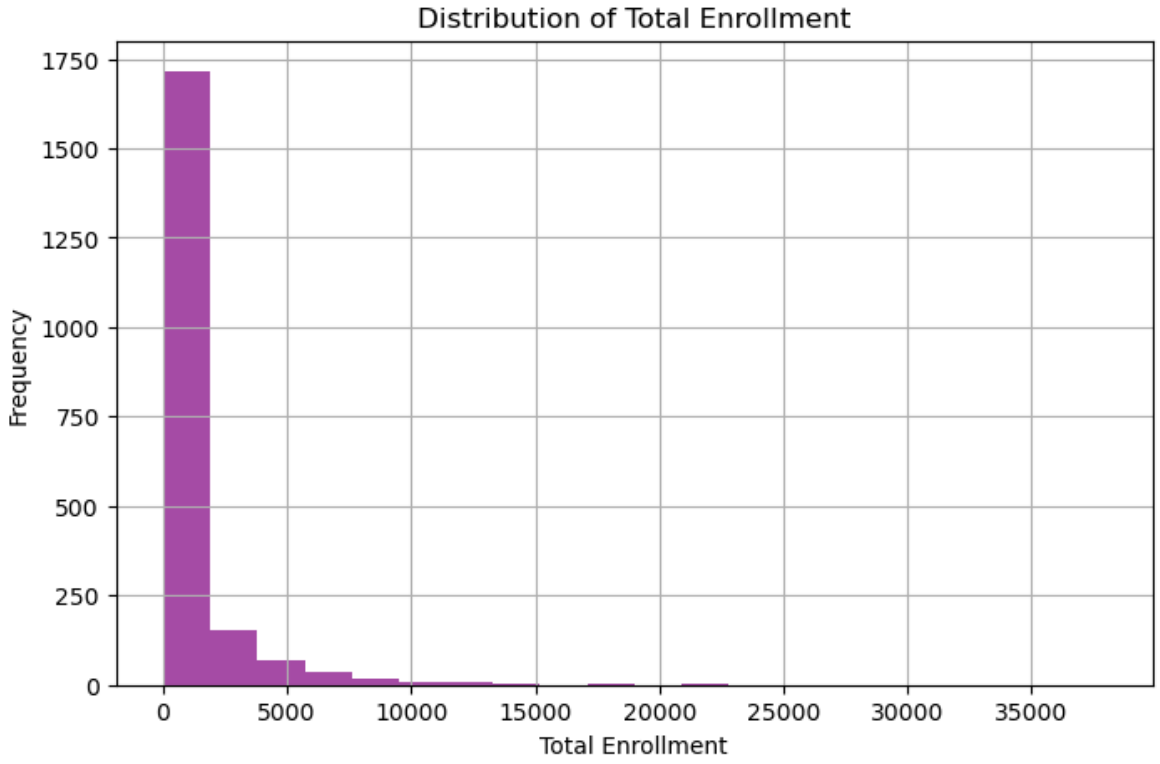


In [30]:
```
plt.figure(figsize=(8, 8))
df['school_category'].value_counts().plot(kind='pie', autopct='%1.1f%%', startan
plt.title('School Category Distribution')
```

```
plt.ylabel('')
plt.show()
```



In [31]:
```
plt.figure(figsize=(8, 5))
plt.hist(df['total_enrollment'], bins=20, color='purple', alpha=0.7)
plt.title('Distribution of Total Enrollment')
plt.xlabel('Total Enrollment')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```

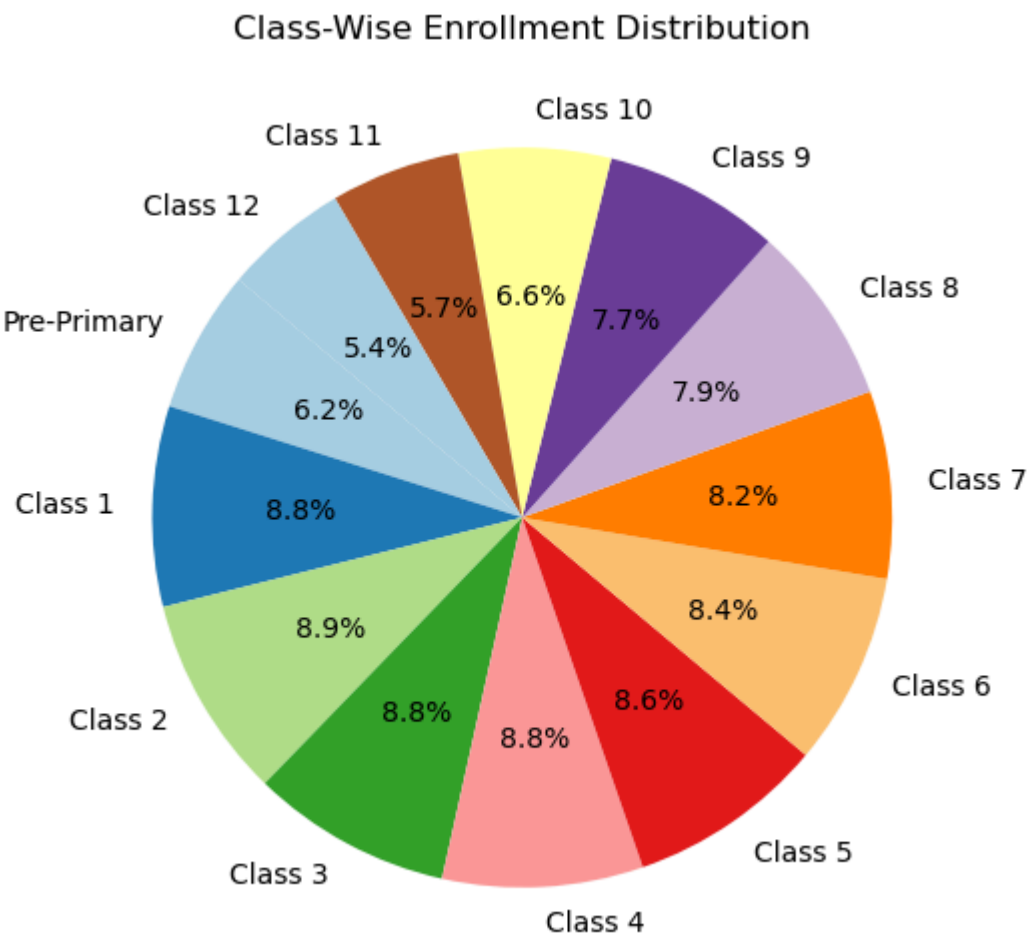


In [32]:
```
df["total_pre_primary"] = df["pre_primary_boy"] + df["pre_primary_girl"]
class_wise_enrollment = {"Pre-Primary": df["total_pre_primary"].sum()}

for i in range(1, 13):
    df[f"total_class{i}"] = df[f"class{i}_boy"] + df[f"class{i}_girl"]
    class_wise_enrollment[f"Class {i}"] = df[f"total_class{i}"].sum()
```

In [33]:
```
classes = list(class_wise_enrollment.keys())
enrollment_numbers = list(class_wise_enrollment.values())

plt.figure(figsize=(10, 6))
plt.pie(enrollment_numbers, labels=classes, autopct='%1.1f%%', startangle=140, c
plt.title("Class-Wise Enrollment Distribution")
plt.show()
```



-> In 2 class there are more students and 12 class has less students

conclusion:

Total Enrollment: 2,171,513 students across all schools

Gender Distribution: 1,161,483 boys and 1,010,030 girls, with a gender ratio of approximately 1.15 boys per girl.

Class-Wise Enrollment: The most populated class is Class 2 (192,592 students), while the least populated is Class 12 (118,248 students)

Pre-Primary Enrollment: 135,706 students are enrolled in pre-primary education.

School Management Distribution: The majority of students are enrolled in Private Unaided (Recognized) schools, with 928,468 students.

Rural vs Urban Enrollment: 872,442 students are enrolled in rural schools, while 1,299,071 students are in urban schools.

In [ ]: