



SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority



Population by Details in Saudi Arabia

Amjad Alomani

Aisha Aloumi

Instructor: Ali El-Kassas

1- We used the Pandas library in Python and import the library & Load the selected dataset into your preferred data analysis environment.

like what is shown below:

```
# Load the selected dataset into your
import pandas as pd

df = pd.read_csv('population.csv')
```

2- Examine the first few rows of a dataset to understand their contents and determine the number of rows.

```
df.head(10)

df = df.sample(10000)
df
```

	Age_Group	Age_Id	Gender	Governorate	Nationality	Region	Pop
51350	90-94	91	Female	Al Andah	Saudi	Jazan	7
28880	45-49	48	Female	Al Tawal	Non-Saudi	Jazan	15
7232	10-14	12	Female	Al Harjah	Saudi	Aseer	202
49012	85-89	85	Male	Ar Rayn	Non-Saudi	Ar Riyadh	1
14238	20-24	24	Male	Ahad Ruffaydah	Non-Saudi	Aseer	508
...
13543	20-24	22	Male	Thar	Saudi	Najran	79
45387	75-79	77	Female	BiJurashi	Saudi	Al Bahah	42
28138	45-49	47	Male	Al Khafji	Saudi	Eastern Region	234
36006	60-64	60	Female	Ash Shamli	Non-Saudi	Hail	4
53316	95-99	97	Female	Al Ahsa	Saudi	Eastern Region	17

10000 rows x 7 columns

3- Explore the structure of the dataset:

```
df.dtypes

df.shape

df.columns

df.index
```

4- Check for and handle any duplicate entries.

```
df.duplicated()

df.duplicated().sum()
```

5- Handle missing values appropriately.

```
df.isnull().sum()
Age_Group      0
Age_Id         0
Gender         0
Governorate    0
Nationality    0
Region         0
Pop            0
dtype: int64
```

6- Conduct univariate analysis to understand the distribution of individual variables.

[illegible]

7- value counts + using groupby

```
df['Gender'].value_counts()

Female    5822
Male      4978
Name: Gender, dtype: int64

df['Nationality'].value_counts()

Saudi      5462
Non-Saudi  4538
Name: Nationality, dtype: int64

import pandas as pd

nat= df.groupby(['Nationality', 'Gender']).size()
for nationality in df['Nationality'].unique():
    print("Nationality:", nationality)
    print(gender_counts[nationality])

Nationality: Saudi
Gender
Female    2787
Male      2675
dtype: int64
Nationality: Non-Saudi
Gender
Female    2235
Male      2303
dtype: int64

total_population = df['Pop'].sum()
total_population

5509243
```

```
import pandas as pd

max_age_by_region = df.groupby('Region')['Age_Id'].max()
region_with_highest_ages = max_age_by_region.idxmax()
region_with_highest_ages

'Al Bahah'

df['Age_Id'].max()

110
```

```
import pandas as pd
rbc=df['Region'].value_counts()
rbc
```

Ar Riyadh	1471
Makkah Al Mukarramah	1197
Jazan	1157
Aseer	1114
Al Qaseem	895
Eastern Region	838
Al Madinah Al Munawwarah	664
Al Bahah	618
Hail	584
Tabuk	487
Najran	452
Northern Borders	263
Al Jawf	260

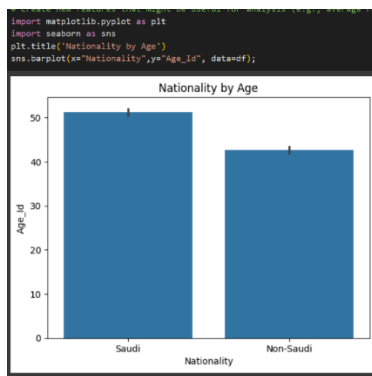
```
Name: Region, dtype: int64
```

```
import pandas as pd
cal = df.groupby('Region')['Pop'].sum()
cal
```

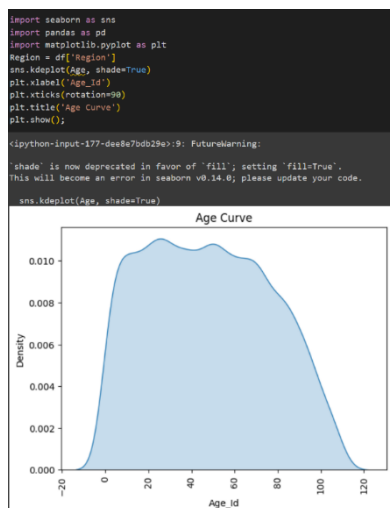
Region	
Al Bahah	59656
Al Jawf	92543
Al Madinah Al Munawwarah	441927
Al Qaseem	250482
Ar Riyadh	1031189
Aseer	388947
Eastern Region	954288
Hail	153089
Jazan	245158
Makkah Al Mukarramah	1541595
Najran	116708
Northern Borders	63330
Tabuk	188331

```
Name: Region, dtype: int64
```

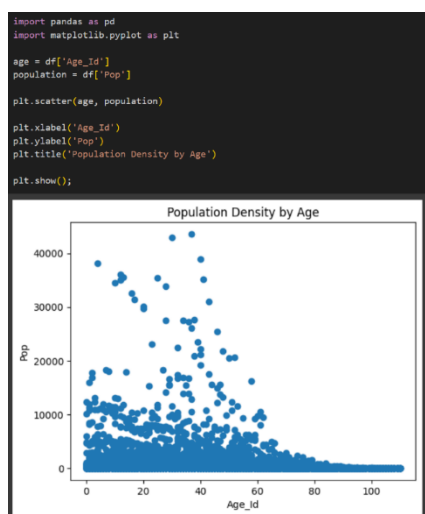
8- Perform bivariate analysis to explore relationships between variables



9- Conduct univariate analysis to understand the distribution of individual variables:



10- Visualize the data using appropriate plot



```
import pandas as pd

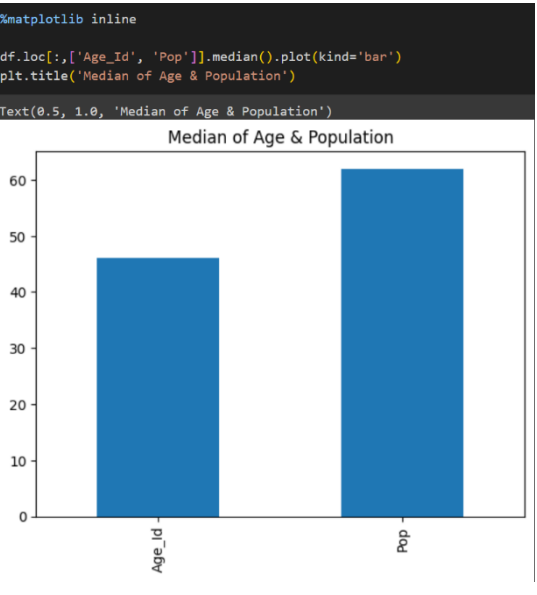
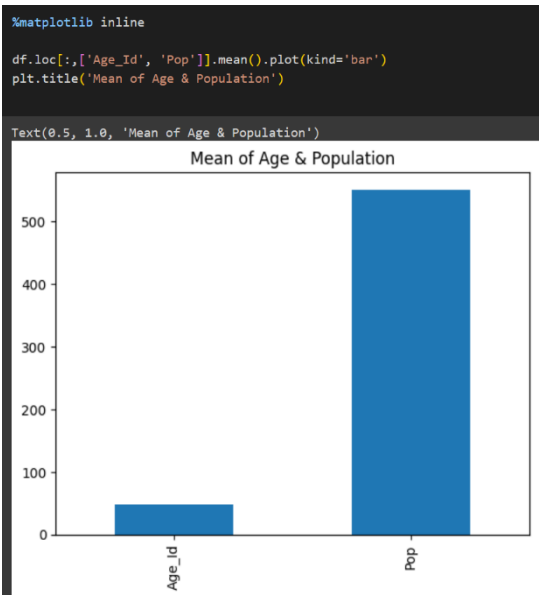
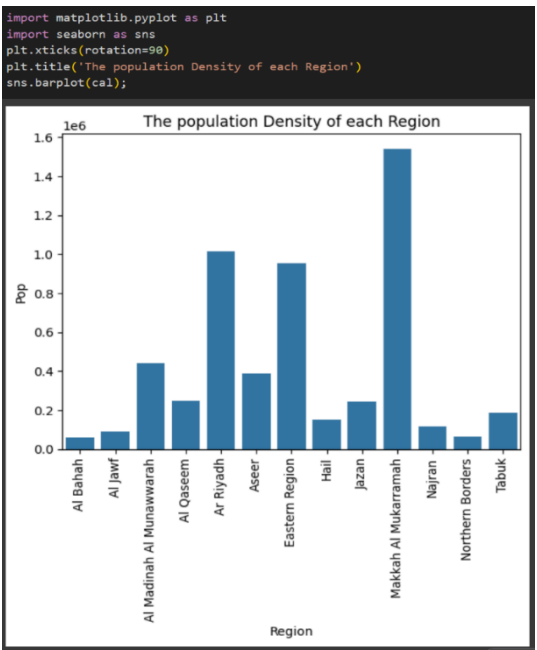
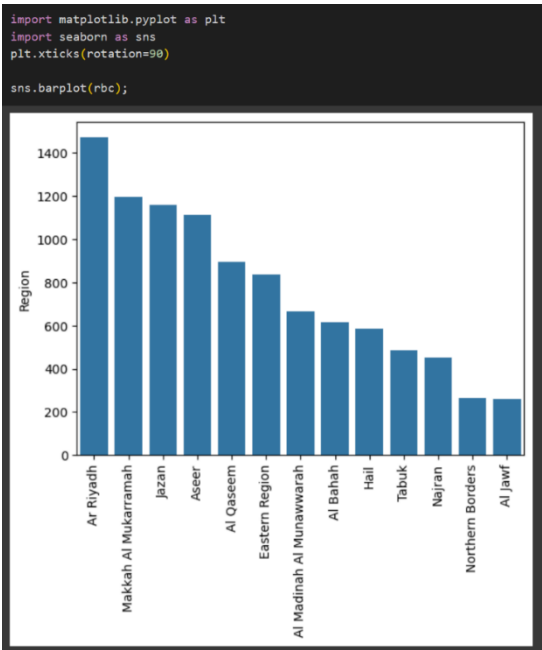
age_bins = [0, 18, 30, 50, float('inf')]
age_labels = ['0-17', '18-29', '30-49', '50+']
df['Age_Group'] = pd.cut(df['Age_Id'], bins=age_bins, labels=age_labels)

age_group_population = df.groupby('Age_Group')['Pop'].sum()
age_group_with_max_population = age_group_population.idxmax()

print("Age group with maximum population:", age_group_with_max_population)
```

Age group with maximum population: 30-49

11- analysis:



12- Create new features that might be useful for analysis:

```
import pandas as pd
df['Age_Population_Relationship'] = df['Age_Id'] / df['Pop']
df
```

	Age_Group	Age_Id	Gender	Governorate	Nationality	Region	Pop	Age_Population_Relationship
51350	90-94	91	Female	Al Aridah	Saudi	Jazan	7	13.000000
28880	45-49	48	Female	At Tuwal	Non-Saudi	Jazan	15	3.200000
7232	10-14	12	Female	Al Harjah	Saudi	Aseer	202	0.059406
49012	85-89	85	Male	Ar Rayn	Non-Saudi	Ar Riyadh	1	85.000000
14238	20-24	24	Male	Ahad Rufaydah	Non-Saudi	Aseer	508	0.047244
...
13543	20-24	22	Male	Thar	Saudi	Najran	79	0.278481
45387	75-79	77	Female	Biljurashi	Saudi	Al Bahah	42	1.833333
28138	45-49	47	Male	Al Khafji	Saudi	Eastern Region	234	0.200855
36006	60-64	60	Female	Ash Shamli	Non-Saudi	Hail	4	15.000000
53316	95-99	97	Female	Al Ahsa	Saudi	Eastern Region	17	5.705882

10000 rows x 8 columns

13- Calculate summary statistics for relevant variables:

```
min_age = df['Age_Id'].min()
max_age = df['Age_Id'].max()
print("Minimum Age:", min_age)
print("Maximum Age:", max_age)

Minimum Age: 0
Maximum Age: 110

mean_age = df['Age_Id'].mean()
median_age = df['Age_Id'].median()
print("Mean Age:", mean_age)
print("Median Age:", median_age)

Mean Age: 47.2988
Median Age: 46.0

age_range = max_age - min_age
print("Age Range:", age_range)

Age Range: 110

age_groups = pd.cut(df['Age_Id'], bins=[0, 18, 30, 50, float('inf')], labels=['0-17', '18-29', '30-49', '50+'])
age_group_counts = age_groups.value_counts()
print("Population Distribution by Age Group:")
print(age_group_counts)
```

Population Distribution by Age Group:

50+	4513
30-49	2181
0-17	1891
18-29	1304

Name: Age_Id, dtype: int64

14-Age score for each region:

```
import pandas as pd
age_by_region = df.groupby('Region')['Age_Id'].unique()
for region, age in age_by_region.items():
    print("Region:", region)
    print("Age_Id:", age)
```

Region: Al Bahah
Age_Id: [65 41 13 78 50 3 34 35 22 33 39 85 60 76 17 8 11 53
20 43 0 7 69 4 103 52 62 5 100 24 23 94 58 67 49 99
19 95 1 18 81 9 51 66 6 77 71 54 10 97 73 55 38 89
70 40 44 45 28 30 63 32 91 48 59 46 98 14 90 31 2 84
104 93 110 92 16 47 42 88 25 86 57 79 12 26 29 36 61 72
82 27 96 68 105 21 83 64 75 56 87 15 102 80 37 101]

Region: Al Jawf
Age_Id: [27 69 19 16 2 28 13 96 52 61 84 11 1 14 71 30 66 8
34 60 10 20 58 12 82 15 35 44 25 81 65 39 31 21 6 23
98 45 79 56 51 26 4 57 22 106 24 47 36 75 73 3 76 89
50 17 5 49 78 70 9 72 38 40 68 88 101 48 92 67 18 46
62 54 0 91 53 42 99 90 77 33 64 63 109 29 105 80 7 43
95 83 55 74 110 41 100 102 37]

Region: Al Madinah Al Munawwarah
Age_Id: [88 8 46 65 40 52 17 59 24 31 75 48 51 108 67 101 3 39
56 36 35 80 72 28 87 70 10 73 38 105 18 16 83 84 30 33
42 66 0 13 49 1 12 61 96 4 20 94 47 34 53 92 6 57
68 44 82 15 45 41 23 81 55 85 29 64 109 90 62 7 91 14
69 95 77 19 32 89 54 58 9 21 11 25 63 107 50 86 106 100
27 97 60 22 98 74 103 43 102 71 76 26 78 2 104 93 37 79
99 5]

Region: Al Qaseem
Age_Id: [23 3 56 70 12 24 64 37 14 21 50 55 42 57 79 71 11 89
87 51 53 17 59 46 19 45 33 80 2 9 74 47 61 27 72 36
68 8 10 54 34 4 16 43 20 25 90 82 98 22 28 101 15 60
75 84 91 49 76 52 38 44 41 67 77 88 6 29 107 58 106 63
48 1 5 81 83 39 92 62 40 30 109 104 95 26 65 102 94 73
78 31 86 66 85 97 35 13 0 99 18 32 93 96 100 7 105 69
108]

Region: Ar Riyadh
Age_Id: [85 37 46 41 47 60 81 13 55 28 14 72 40 7 44 83 62 30
49 35 22 5 100 9 109 54 23 6 79 12 74 58 70 20 17 88

What were the problems faced in this project?

- 1 .Finding Saudi-specific data containing 10,000 rows.
- 2 .Slow query performance due to the data volume.

What is the largest population according to age groups?

Youth and children.

Why is the number of non-Saudi residents very close to the number of Saudi residents?

Because of the economic development in the city and also tourism.

What is the most densely populated city?

- 1 .Mecca
- 2 .Riyadh
- 3 .Eastern Province

What factors drive seniors to choose living in rural areas away from crowded cities?

Seniors' preference for rural living can be explained by the greater tranquility and safety, as well as the opportunity for a healthier lifestyle and less stressful living environment provided by nature and rural settings.

Reference:

<https://open.data.gov.sa/ar/home>

General Authority for Statistics in the Kingdom of Saudi Arabia