In [4]:

```
import pandas as pd
import numpy as np
from pyspark.sql import SparkSession
import pyspark.sql.functions as F
```

In [5]:

```
# create a SparkSession
spark = SparkSession.builder.appName("House Rent").getOrCreate()
```

In [7]:

```
df1 = spark.read.csv("D:\sem-6\DSPL\DSPL_LAB-main\House_Rent_main6-main1.csv",
```

In [8]:

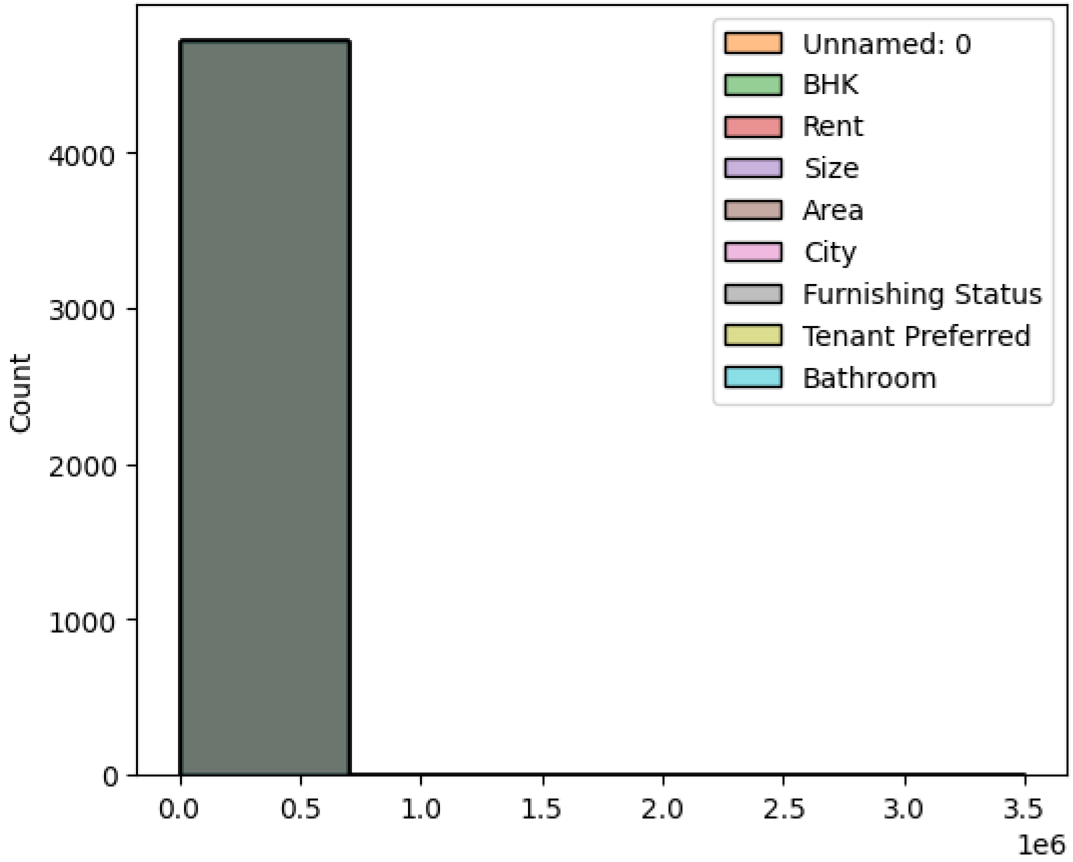
```
df = df1
df.printSchema() # print the schema of the DataFrame
```

```
root
 |-- _c0: integer (nullable = true)
 |-- Unnamed: 0: integer (nullable = true)
 |-- BHK: integer (nullable = true)
 |-- Rent: integer (nullable = true)
 |-- Size: integer (nullable = true)
 |-- Floor: string (nullable = true)
 |-- Area: integer (nullable = true)
 |-- Area Locality: string (nullable = true)
 |-- City: integer (nullable = true)
 |-- Furnishing Status: integer (nullable = true)
 |-- Tenant Preferred: integer (nullable = true)
 |-- Bathroom: integer (nullable = true)
```

In [9]:
```
# visualize the data
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(6,5))
sns.histplot(df.toPandas(), bins=5)
```

```
C:\Users\zaidk\AppData\Local\Programs\Python\Python310\lib\site-packages\seab
orn\distributions.py:163: UserWarning: The label '_c0' of <matplotlib.patche
s.Patch object at 0x000001E684567C40> starts with '_'. It is thus excluded fr
om the legend.
  ax_obj.legend(handles, labels, title=self.variables["hue"], **legend_kws)
```

Out[9]:
```
<Axes: ylabel='Count'>
```

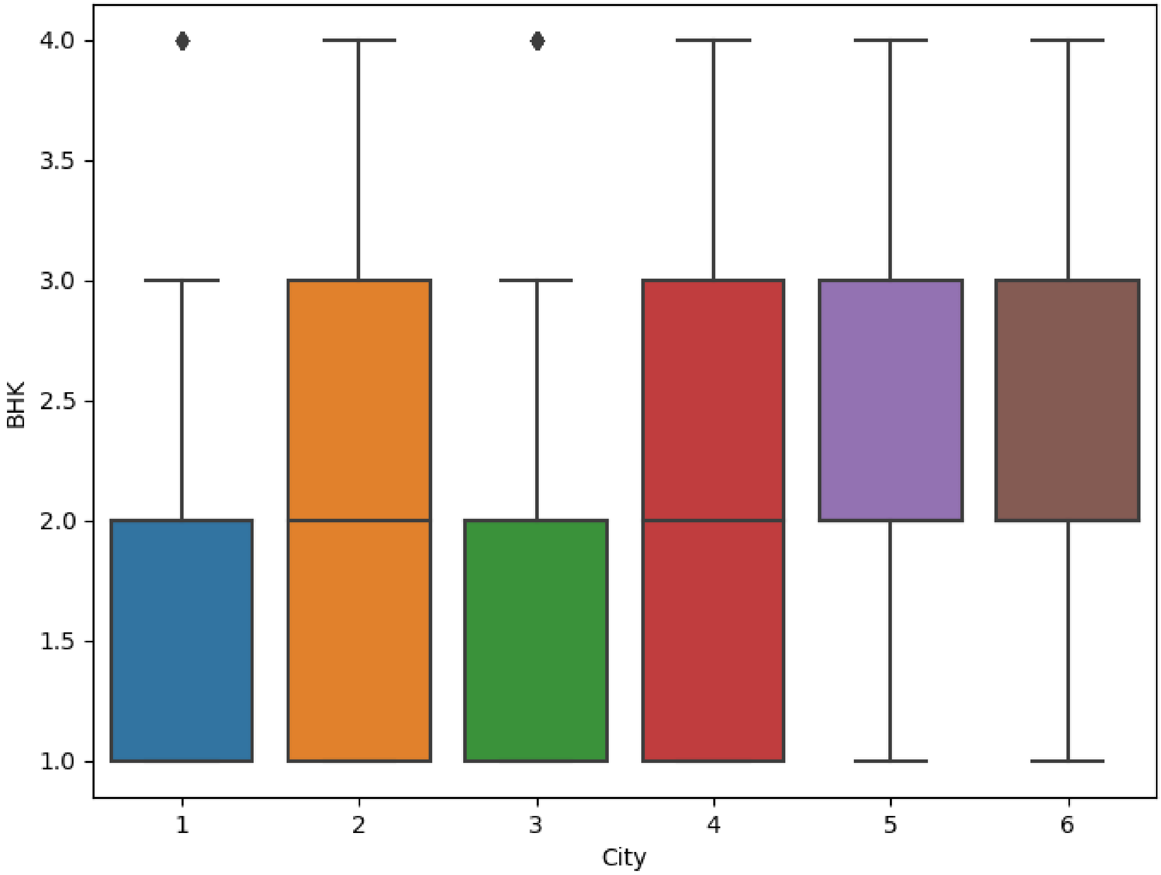


In [10]:
```
df.head()
```

Out[10]:
```
Row(_c0=0, Unnamed: 0=0, BHK=2, Rent=10000, Size=1100, Floor='Ground out of
2', Area=1, Area Locality='Bandel', City=1, Furnishing Status=1, Tenant Prefe
rred=1, Bathroom=2)
```

In [14]:

```
# visualize the data
plt.figure(figsize=(8,6))
sns.boxplot(data=df.toPandas(), x="City", y="BHK")
plt.show()
```



In [15]:

```
# stop the SparkSession
spark.stop()
```

In [19]:

```
import pandas as pd
df2 = pd.read_csv("House_Rent_main6-main1.csv")
```

In [20]:
```
df2.head(4)
```

Out[20]:

	Unnamed: 0.1	Unnamed: 0	BHK	Rent	Size	Floor	Area	Area Locality	City	Furnishing Status	Ter Prefe
0	0	0	2	10000	1100	Ground out of 2	1	Bandel	1	1	
1	1	1	2	20000	800	1 out of 3	1	Phool Bagan, Kankurgachi	1	2	
2	2	2	2	17000	1000	1 out of 3	1	Salt Lake City Sector 2	1	2	
3	3	3	2	10000	800	1 out of 2	1	Dumdum Park	1	1	

In [23]:
```
df2.drop(['Floor'], axis=1, inplace=True)
```

In [24]:
```
df2.head()
```

Out[24]:

	Unnamed: 0.1	Unnamed: 0	BHK	Rent	Size	Area	Area Locality	City	Furnishing Status	Tenant Preferred	Bat
0	0	0	2	10000	1100	1	Bandel	1	1	1	
1	1	1	2	20000	800	1	Phool Bagan, Kankurgachi	1	2	1	
2	2	2	2	17000	1000	1	Salt Lake City Sector 2	1	2	1	
3	3	3	2	10000	800	1	Dumdum Park	1	1	1	
4	4	4	2	7500	850	2	South Dum Dum	1	1	2	

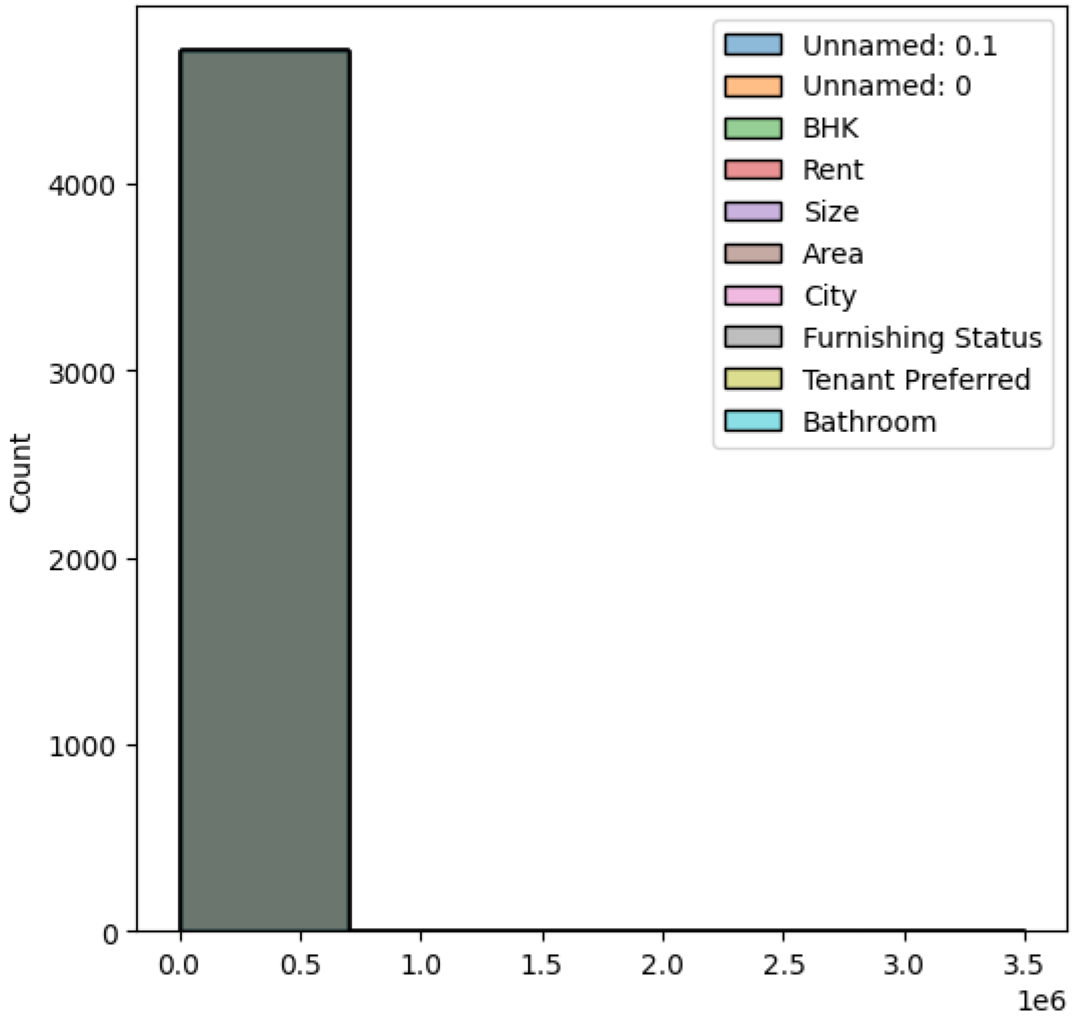
In [25]:
```
# visualize the data
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [26]:

```
plt.figure(figsize=(6,6))
sns.histplot(df, bins=5)
```

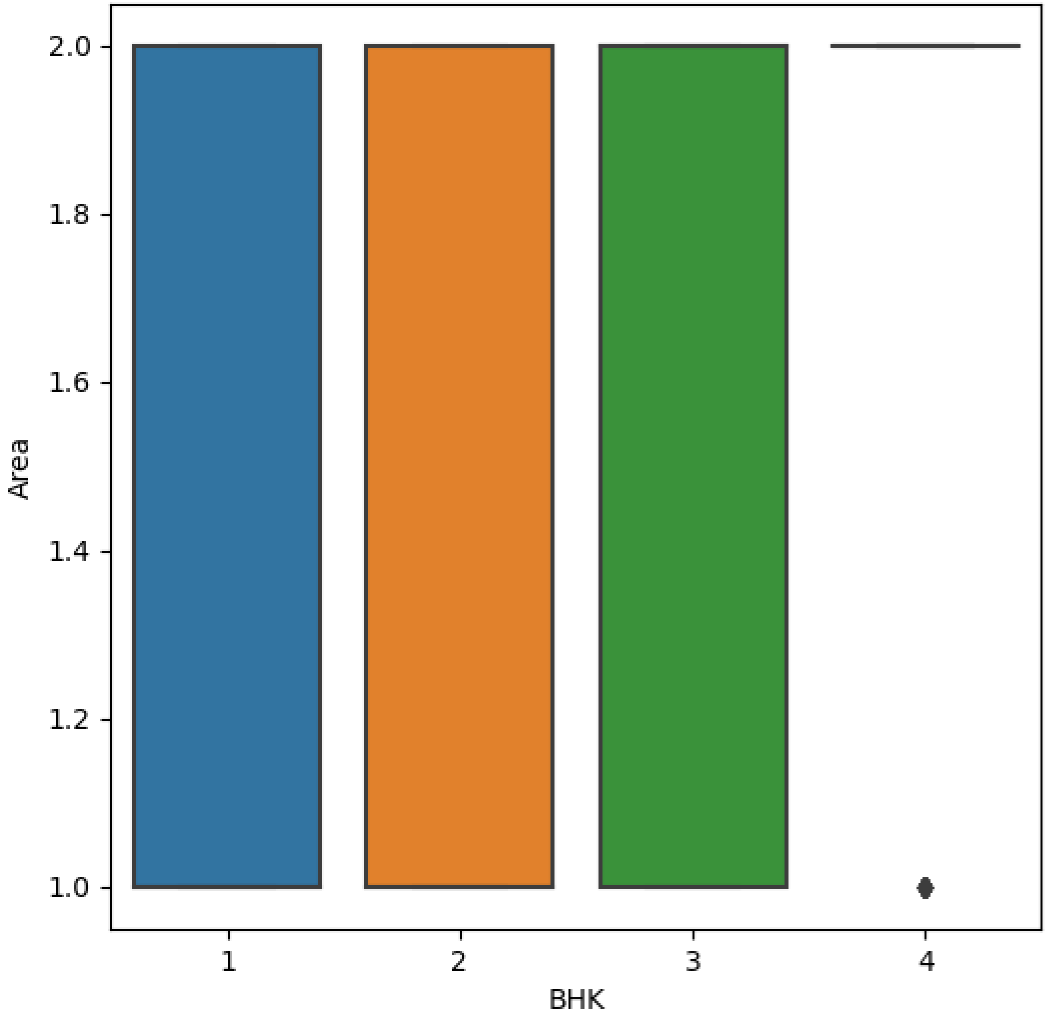
Out[26]:

```
<Axes: ylabel='Count'>
```



In [33]:

```
# visualize the data
plt.figure(figsize=(6,6))
sns.boxplot(data=df, x="BHK", y="Area")
plt.show()
```



In [ ]: