```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```python
path="/content/housing.csv"
```

```python
df = pd.read_csv('/content/housing.csv')
```

```python
df
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	45
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	35
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	35
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	34
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	34
...	...	...	...	...	...	...	...	...	
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	7
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	7
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	9
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	8
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	8

20640 rows × 10 columns

```python
print(df.isnull().sum())
```

```
longitude             0
latitude              0
housing_median_age    0
total_rooms           0
total_bedrooms      207
population            0
households            0
median_income         0
median_house_value    0
ocean_proximity       0
dtype: int64
```

```python
df.fillna(df.mean(), inplace=True)
```

```python
df.dropna(inplace=True)
```

```python
print(df.duplicated().sum())
df.drop_duplicates(inplace=True)
```

```
0
```

```python
for col in df.select_dtypes(include=np.number).columns:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    LB = Q1 - 1.5 * IQR
    UB = Q3 + 1.5 * IQR
    outliers = df[(df[col] < LB) | (df[col] > UB)]
    df = df.drop(outliers.index)
```

```python
df.to_csv("/content/housing.csv", index=False)
```

```python
print(f"Col: {col}")
print(f"Original data points: {len(df[col])}")
print(f"Data points after IQR filtering: {len(outliers)}\n")
```

```
Col: median_house_value
Original data points: 15857
Data points after IQR filtering: 47
```

```python
plt.figure(figsize=(10, 6))
plt.bar(df[col].index, df[col], label=col)
plt.xlabel('Index')
plt.ylabel(col)
plt.title(f'Bar graph for {col}')
plt.show()
```

➡

Bar graph for median_house_value