campusx-official / **100-days-of-machine-learning** Public

Code Issues 3 Pull requests Actions Projects Security Insights

main

100-days-of-machine-learning / day37-handling-missing-categorical-data / **missing-category-imputation.ipynb**

campusx-official Add files via upload

History

1 contributor

```
294 lines (294 sloc) | 24 KB
```

In [22]:

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

In [23]:

```
df = pd.read_csv('train.csv',usecols=['GarageQual','FireplaceQu','SalePrice'
```

In [24]:

```
df.head()
```

Out[24]:

	FireplaceQu	GarageQual	SalePrice
0	NaN	TA	208500
1	TA	TA	181500
2	TA	TA	223500
3	Gd	TA	140000
4	TA	TA	250000

In [25]:

```
df.isnull().mean()*100
```

Out[25]:

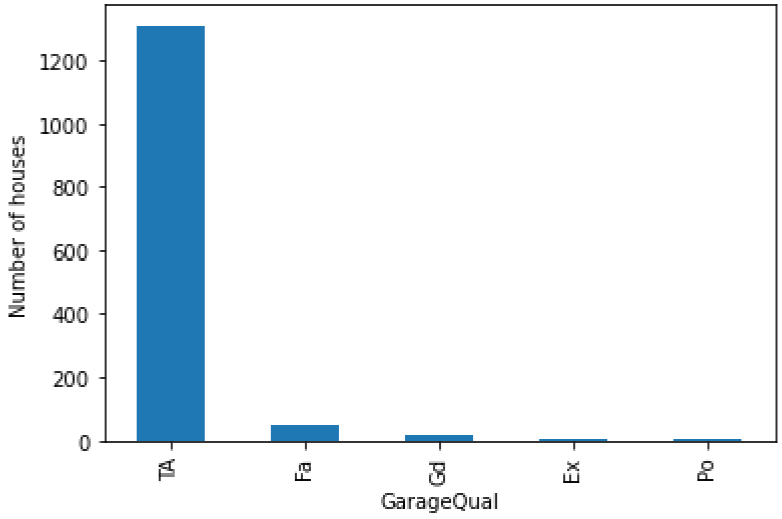
```
FireplaceQu    47.260274
GarageQual      5.547945
SalePrice       0.000000
dtype: float64
```

In [26]:

```
df['GarageQual'].value_counts().sort_values(ascending=False).plot.bar()
plt.xlabel('GarageQual')
plt.ylabel('Number of houses')
```

Out[26]:

```
Text(0, 0.5, 'Number of houses')
```



In [27]:

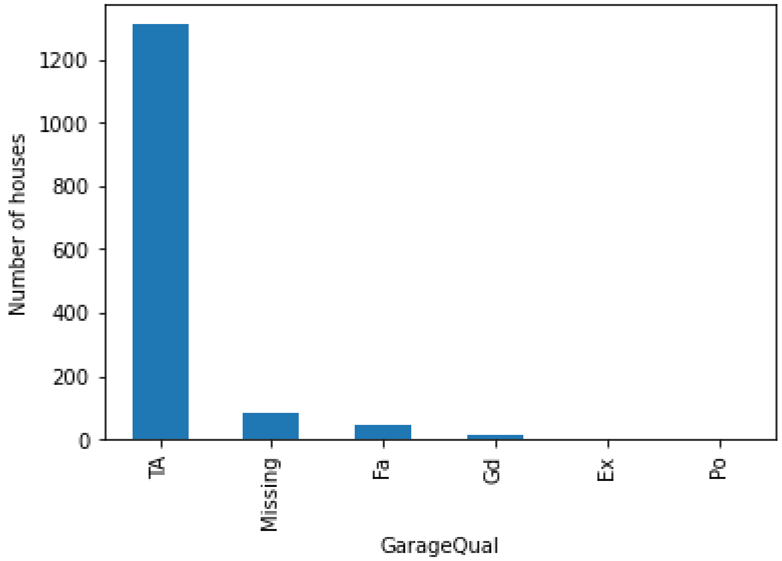
```
df['GarageQual'].fillna('Missing', inplace=True)
```

In [28]:

```
df['GarageQual'].value_counts().sort_values(ascending=False).plot.bar()
plt.xlabel('GarageQual')
plt.ylabel('Number of houses')
```

Out[28]:

```
Text(0, 0.5, 'Number of houses')
```



In [29]:

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(df.drop(columns=['SalePrice
```

In [30]:

```
from sklearn.impute import SimpleImputer
```

In [31]:

```
imputer = SimpleImputer(strategy='constant',fill_value='Missing')
```

In [32]:

```
X_train = imputer.fit_transform(X_train)
X_test = imputer.transform(X_train)
```

In [33]:

```
imputer.statistics_
```

Out[33]:

```
array(['Missing', 'Missing'], dtype=object)
```

In [ ]: