Day 12 | Mode - imputation

Import Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Import Dataset

In [3]:
```
df = pd.read_csv('train.csv',usecols=['GarageQual','FireplaceQu','SalePrice'])
```

In [4]:
```
df.head()
```

Out[4]:

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

In [6]:
```
df.sample(10)
```

Out[6]:

	FireplaceQu	GarageQual	SalePrice
822	Gd	TA	225000
958	NaN	TA	185000
486	NaN	TA	156000
658	Gd	TA	97500
1254	Gd	TA	165400
238	NaN	TA	318000
69	TA	TA	225000
921	NaN	NaN	145900
1405	Gd	TA	275000
32	NaN	TA	179900

Check missing (null) value

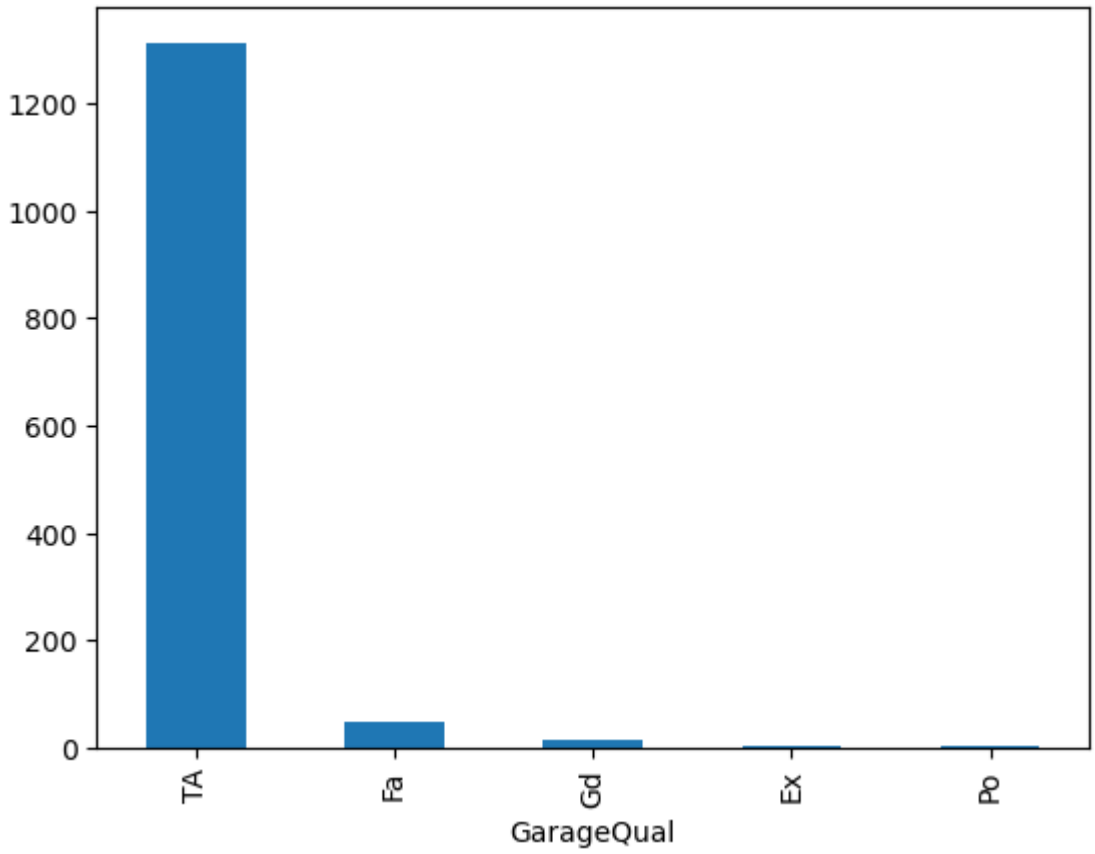
In [7]:
```
df.isnull().mean()*100
```

Out[7]:
```
FireplaceQu    47.260274
GarageQual      5.547945
SalePrice       0.000000
dtype: float64
```

Plot Bar Garage Value

In [8]:
```
df['GarageQual'].value_counts().plot(kind='bar')
```

Out[8]:
```
<Axes: xlabel='GarageQual'>
```



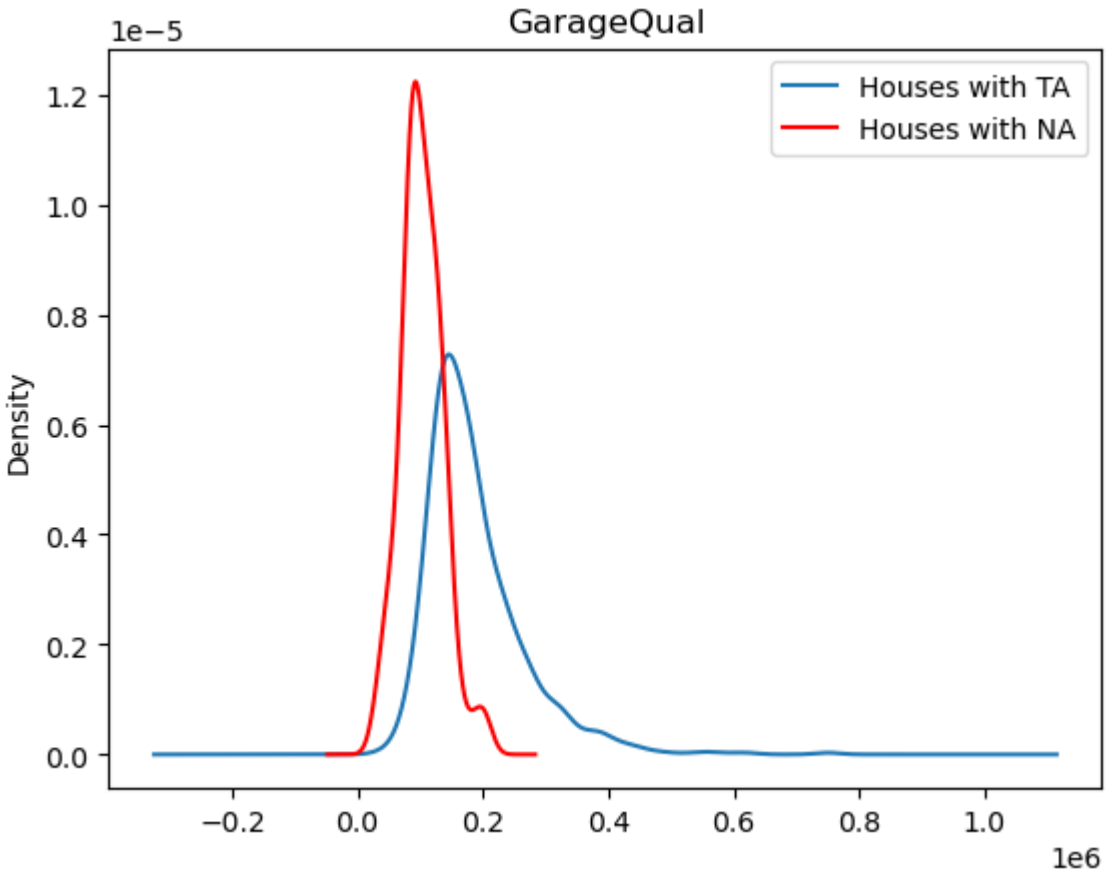
In [9]:
```
df['GarageQual'].mode()
```

Out[9]:
```
0    TA
Name: GarageQual, dtype: object
```

kde Plot | Compare Houses with TA | Null

In [10]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
df[df['GarageQual']=='TA']['SalePrice'].plot(kind='kde', ax=ax)
df[df['GarageQual'].isnull()]['SalePrice'].plot(kind='kde', ax=ax, color='red')
lines, labels = ax.get_legend_handles_labels()
labels = ['Houses with TA', 'Houses with NA']
ax.legend(lines, labels, loc='best')
plt.title('GarageQual')
```

Out[10]:
```
Text(0.5, 1.0, 'GarageQual')
```



Store variable TA in temp

In [11]:
```
temp = df[df['GarageQual']=='TA']['SalePrice']
```

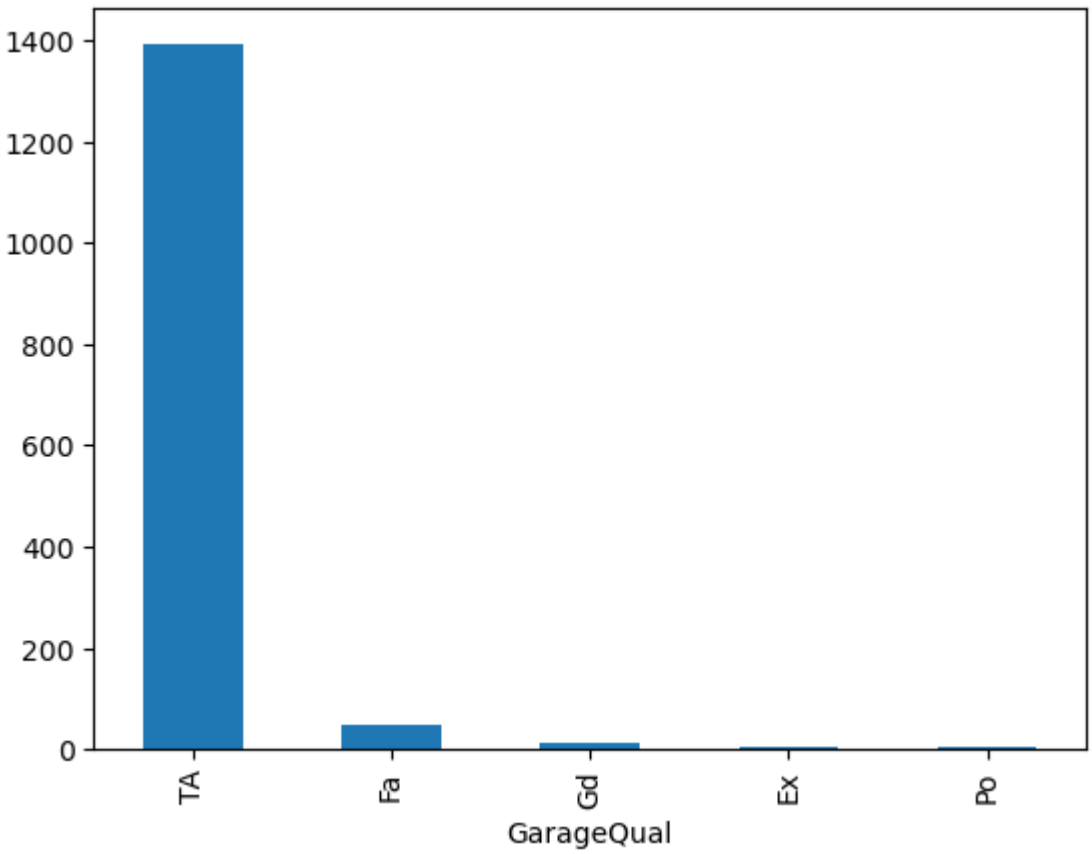
Replace missing value with TA

In [12]:
```
df['GarageQual'].fillna('TA', inplace=True)
```

Review Bar Plot Changes

In [13]:
```
df['GarageQual'].value_counts().plot(kind='bar')
```

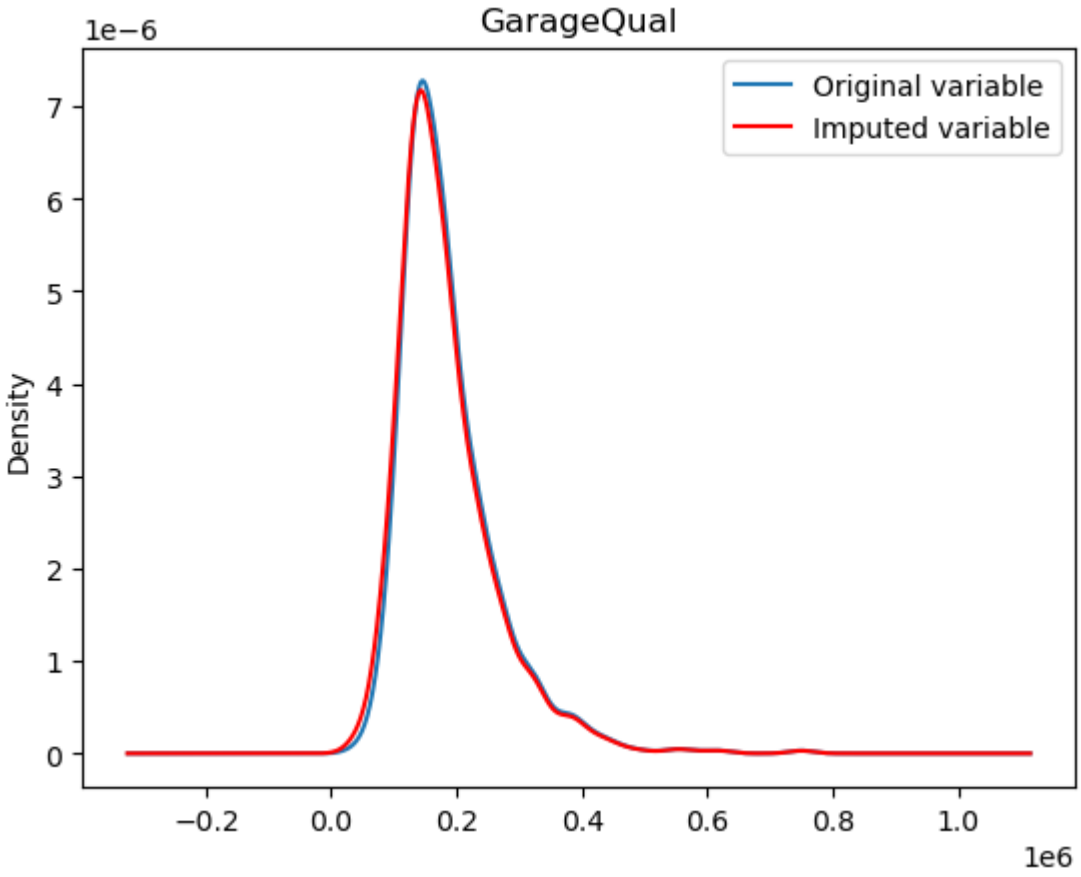
Out[13]: <Axes: xlabel='GarageQual'>



Draw plot again | After Imputation

In [14]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
temp.plot(kind='kde', ax=ax)
# distribution of the variable after imputation
df[df['GarageQual'] == 'TA']['SalePrice'].plot(kind='kde', ax=ax, color='red')
lines, labels = ax.get_legend_handles_labels()
labels = ['Original variable', 'Imputed variable']
ax.legend(lines, labels, loc='best')
# add title
plt.title('GarageQual')
```

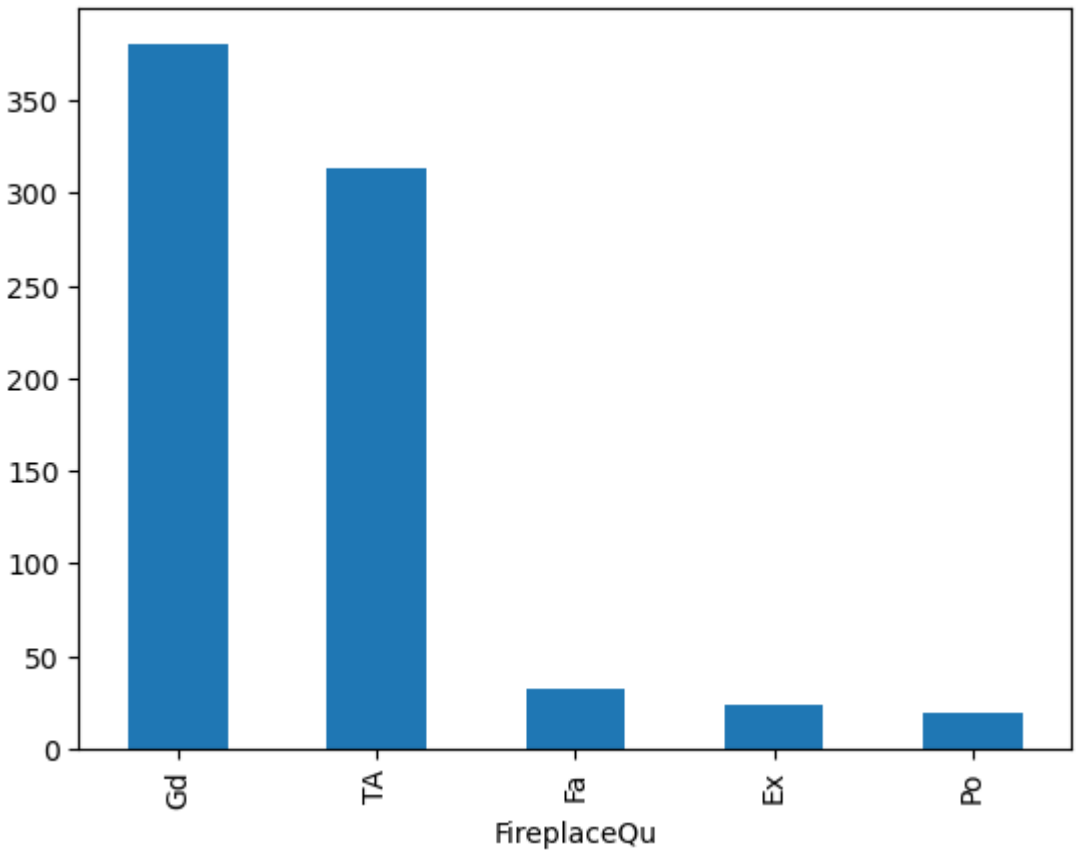
Out[14]:
```
Text(0.5, 1.0, 'GarageQual')
```



Plot Bar Fire Place

In [15]:
```
df['FireplaceQu'].value_counts().plot(kind='bar')
```

Out[15]:
```
<Axes: xlabel='FireplaceQu'>
```



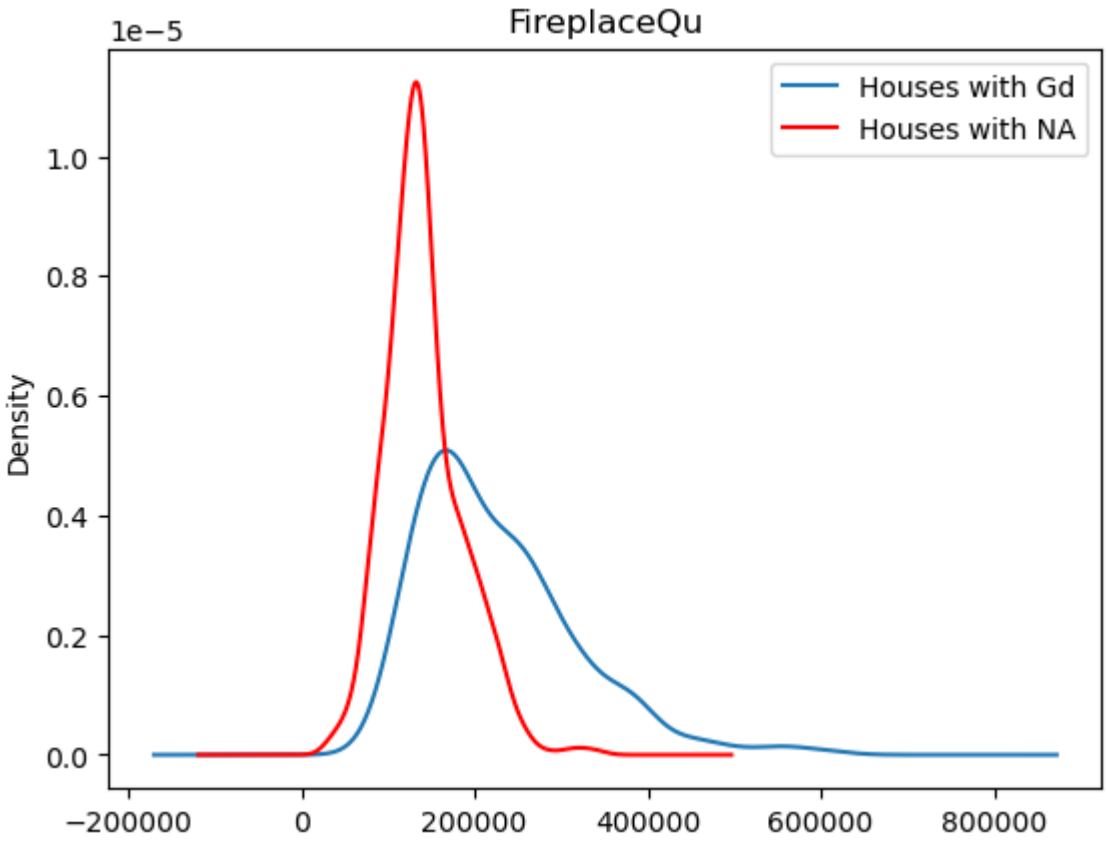
In [16]:
```
df['FireplaceQu'].mode()
```

Out[16]:
```
0    Gd
Name: FireplaceQu, dtype: object
```

kde Plot | Replace House with GD and NA

In [17]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
df[df['FireplaceQu']=='Gd']['SalePrice'].plot(kind='kde', ax=ax)
df[df['FireplaceQu'].isnull()]['SalePrice'].plot(kind='kde', ax=ax, color='red')
lines, labels = ax.get_legend_handles_labels()
labels = ['Houses with Gd', 'Houses with NA']
ax.legend(lines, labels, loc='best')
plt.title('FireplaceQu')
```

Out[17]: Text(0.5, 1.0, 'FireplaceQu')



Store Temp Variable

In [18]:
```
temp = df[df['FireplaceQu']=='Gd']['SalePrice']
```

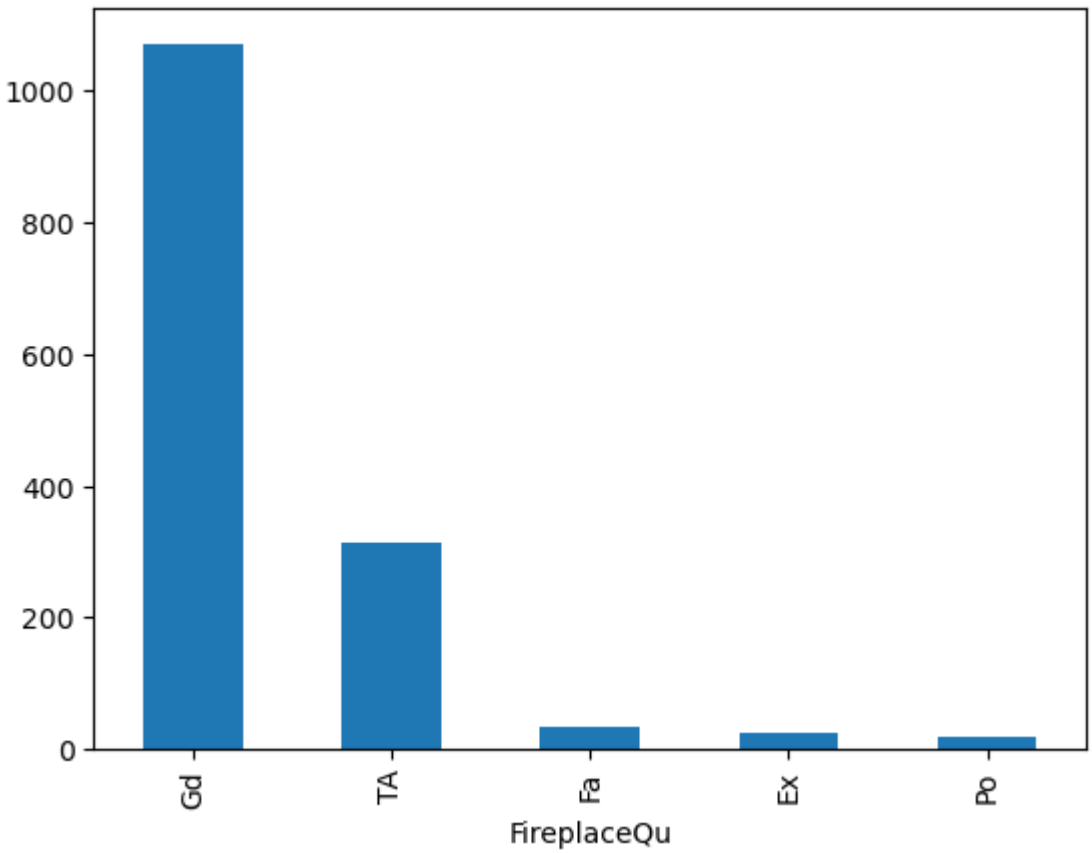
Replace missing value with GD

In [19]:
```
df['FireplaceQu'].fillna('Gd', inplace=True)
```

Draw Bar Plot

In [20]:
```
df['FireplaceQu'].value_counts().plot(kind='bar')
```

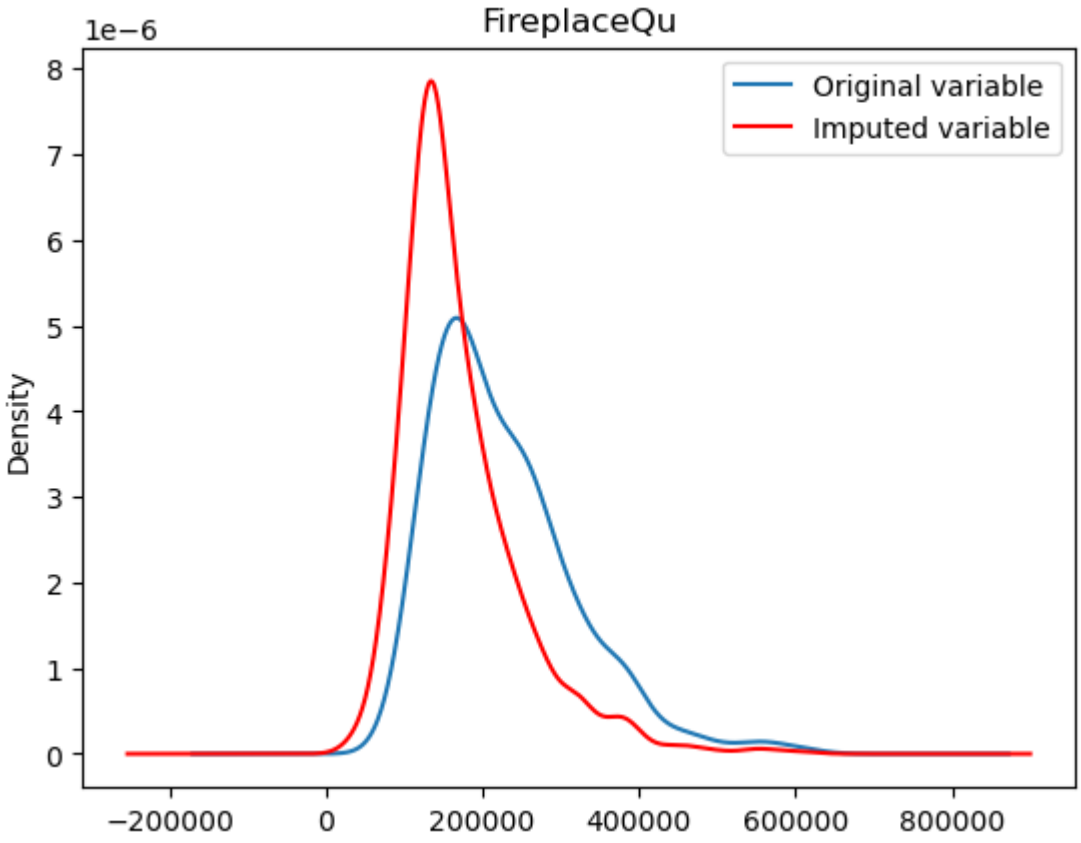
Out[20]:
```
<Axes: xlabel='FireplaceQu'>
```



Draw plot again | After Imputation

In [21]:
```
fig = plt.figure()
ax = fig.add_subplot(111)
temp.plot(kind='kde', ax=ax)
# distribution of the variable after imputation
df[df['FireplaceQu'] == 'Gd']['SalePrice'].plot(kind='kde', ax=ax, color='red')
lines, labels = ax.get_legend_handles_labels()
labels = ['Original variable', 'Imputed variable']
ax.legend(lines, labels, loc='best')
# add title
plt.title('FireplaceQu')
```

Out[21]:
```
Text(0.5, 1.0, 'FireplaceQu')
```



In [ ]: