In [1]:
```
!pip install pandas
# !pip install basic-image-eda
# !pip install scikit-image
# !pip install matplotlib
import pandas as pd
```

```
Looking in indexes: https://pypi.org/simple, https://pypi.ngc.nvidia.com
Requirement already satisfied: pandas in /opt/conda/lib/python3.8/site-packages (1.4.1)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.8/site-packages (from panda
s) (2021.3)
Requirement already satisfied: python-dateutil>=2.8.1 in /opt/conda/lib/python3.8/site-packages
(from pandas) (2.8.2)
Requirement already satisfied: numpy>=1.18.5 in /opt/conda/lib/python3.8/site-packages (from pand
as) (1.21.2)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.8/site-packages (from python-da
teutil>=2.8.1->pandas) (1.16.0)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviou
r with the system package manager. It is recommended to use a virtual environment instead: http
s://pip.pypa.io/warnings/venv
```

In [2]:
```
train = pd.read_csv('data/train22A/train2_new.csv') # reading the csv file
```

In [3]:
```
train.head() # printing first five rows of the file
```

Out[3]:

	name	upperLength	clothesStyles	hairStyles	lowerLength	lowerStyles	shoesStyles	t
0	img_qh_train2_00000006345.jpg	LongSleeve	Solidcolor	Short	Skirt	Solidcolor	Sandals	
1	img_qh_train2_00001008825.jpg	LongSleeve	multicolour	Short	Trousers	Solidcolor	Sneaker	
2	img_qh_train2_00002004117.jpg	LongSleeve	multicolour	Long	Trousers	Solidcolor	Sneaker	
3	img_qh_train2_00003002526.jpg	LongSleeve	multicolour	Short	Trousers	Solidcolor	Sneaker	
4	img_qh_train2_00004004439.jpg	LongSleeve	Solidcolor	Long	Trousers	Solidcolor	Sneaker	

5 rows × 30 columns

In [4]:
```
train.columns
```

Out[4]:
```
Index(['name', 'upperLength', 'clothesStyles', 'hairStyles', 'lowerLength',
       'lowerStyles', 'shoesStyles', 'towards', 'upperBlack', 'upperBrown',
       'upperBlue', 'upperGreen', 'upperGray', 'upperOrange', 'upperPink',
       'upperPurple', 'upperRed', 'upperWhite', 'upperYellow', 'lowerBlack',
       'lowerBrown', 'lowerBlue', 'lowerGreen', 'lowerGray', 'lowerOrange',
       'lowerPink', 'lowerPurple', 'lowerRed', 'lowerWhite', 'lowerYellow'],
      dtype='object')
```

In [5]:
```
train.drop(['lowerLength',
       'lowerStyles', 'shoesStyles', 'towards', 'lowerBlack',
       'lowerBrown', 'lowerBlue', 'lowerGreen', 'lowerGray', 'lowerOrange',
       'lowerPink', 'lowerPurple', 'lowerRed', 'lowerWhite', 'lowerYellow'], axis=1, inplace=Tru
```

In [19]:
```
train.head()
```

Out[19]:

	name	upperLength	clothesStyles	hairStyles	upperBlack	upperBrown	upperBlue	up
0	img_qh_train2_00000006345.jpg	LongSleeve	Solidcolor	Short	NaN	NaN	NaN	

	name	upperLength	clothesStyles	hairStyles	upperBlack	upperBrown	upperBlue	up
1	img_qh_train2_00001008825.jpg	LongSleeve	multicolour	Short	0.4	NaN	NaN	
2	img_qh_train2_00002004117.jpg	LongSleeve	multicolour	Long	NaN	NaN	NaN	
3	img_qh_train2_00003002526.jpg	LongSleeve	multicolour	Short	0.8	NaN	NaN	
4	img_qh_train2_00004004439.jpg	LongSleeve	Solidcolor	Long	NaN	NaN	1.0	

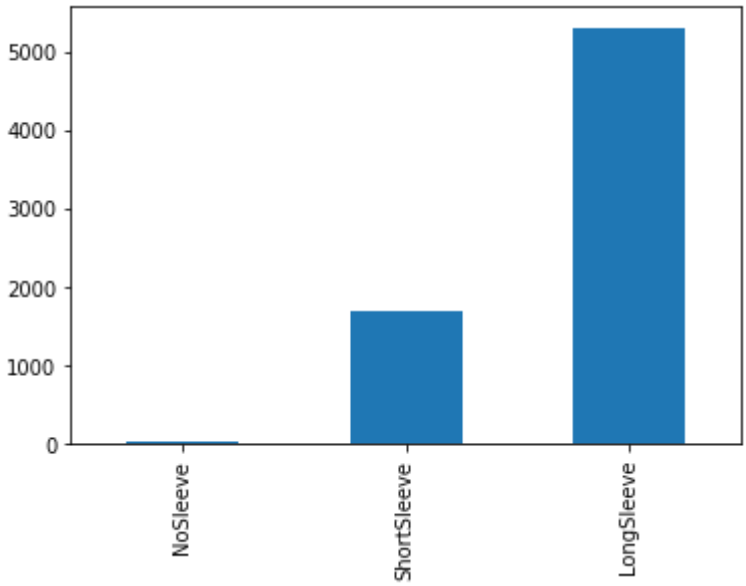
In [20]:
```
train.columns
```

Out[20]:
```
Index(['name', 'upperLength', 'clothesStyles', 'hairStyles', 'upperBlack',
       'upperBrown', 'upperBlue', 'upperGreen', 'upperGray', 'upperOrange',
       'upperPink', 'upperPurple', 'upperRed', 'upperWhite', 'upperYellow'],
      dtype='object')
```

In [6]:
```
# print(train.groupby('upperLength').count()['name'])
# train['upperLength'].hist()
print('upperLength: ')
print(train['upperLength'].value_counts(ascending=True))
train['upperLength'].value_counts(ascending=True).plot.bar()
```

```
upperLength:
NoSleeve         12
ShortSleeve    1686
LongSleeve     5302
Name: upperLength, dtype: int64
<AxesSubplot:>
```

Out[6]:

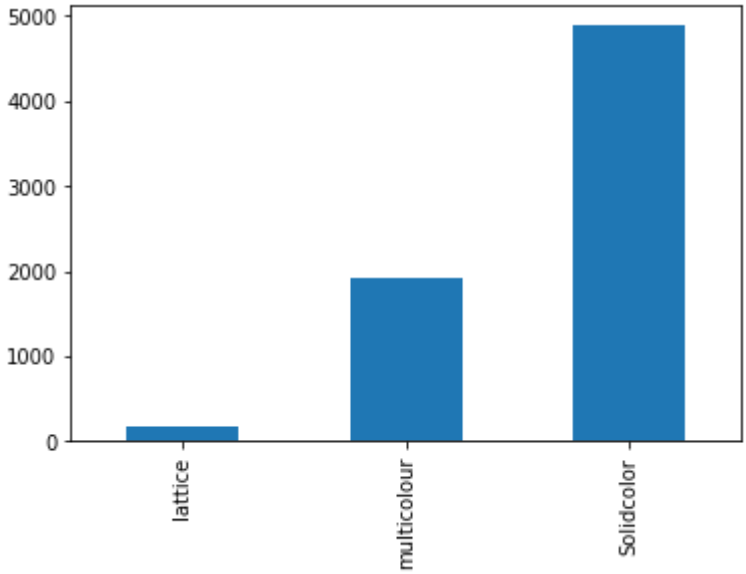


In [7]:
```
# print(train.groupby('clothesStyles').count()['name'])
# train['clothesStyles'].hist()

print('clothesStyles: ')
print(train['clothesStyles'].value_counts(ascending=True))
train['clothesStyles'].value_counts(ascending=True).plot.bar()
```

```
clothesStyles:
lattice         183
multicolour    1930
Solidcolor     4887
Name: clothesStyles, dtype: int64
<AxesSubplot:>
```

Out[7]:



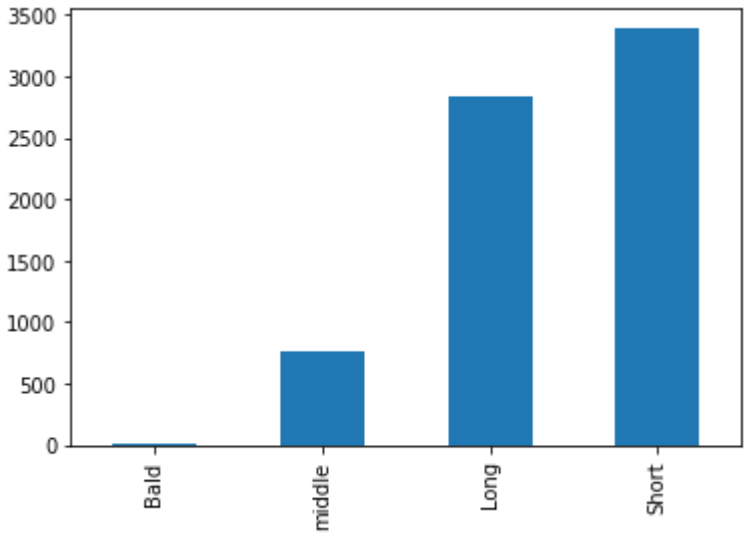
In [8]:

```
# print(train.groupby('hairStyles').count()['name'])
# train['hairStyles'].hist()

print('hairStyles: ')
print(train['hairStyles'].value_counts(ascending=True))
train['hairStyles'].value_counts(ascending=True).plot.bar()
```

```
hairStyles:
Bald        12
middle     764
Long      2839
Short     3385
Name: hairStyles, dtype: int64
<AxesSubplot:>
```

Out[8]:



In [9]:

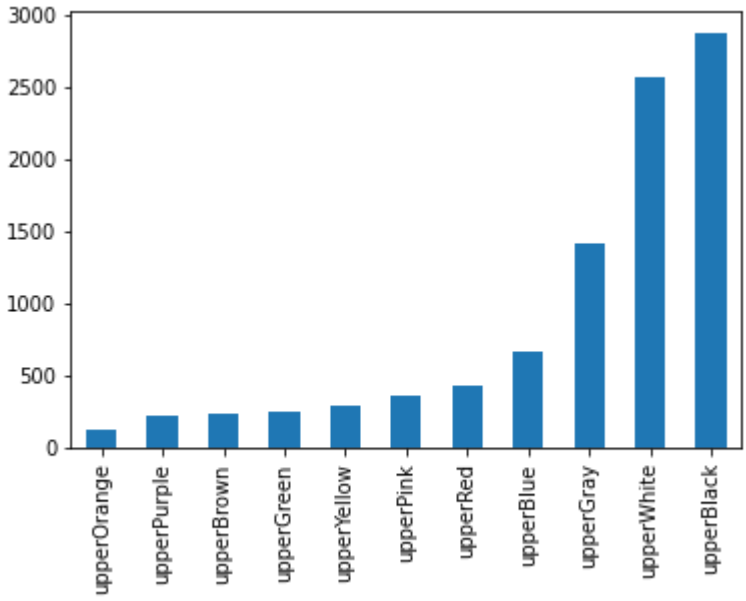
```
new_df = train.copy(deep=True)
new_df.drop(['name','upperLength','clothesStyles', 'hairStyles'], axis=1, inplace=True)
uppercolorsCount = new_df.count().sort_values()
print(uppercolorsCount)
uppercolorsCount.plot.bar()
```

```
upperOrange      117
upperPurple      213
upperBrown       237
upperGreen       246
upperYellow      293
upperPink        364
upperRed         424
upperBlue        662
upperGray       1414
```

```
upperWhite     2563
upperBlack     2877
dtype: int64
```

Out[9]:

```
<AxesSubplot:>
```



In [21]:

```
clothesStyles_group = train.groupby('clothesStyles').count()
clothesStyles_group.drop(['name','upperLength', 'hairStyles'], axis=1, inplace=True)
clothesStyles_group.head()
```

Out[21]:

clothesStyles	upperBlack	upperBrown	upperBlue	upperGreen	upperGray	upperOrange	upperPink	upperPurp
Solidcolor	1545	114	366	132	725	67	237	1
lattice	133	13	12	2	59	0	6	
multicolour	1199	110	284	112	630	50	121	

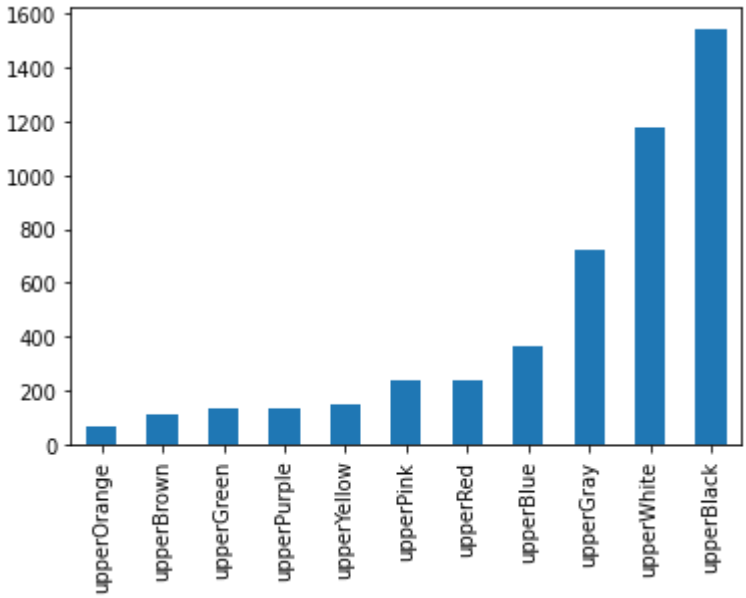
In [22]:

```
print(clothesStyles_group.loc['Solidcolor'].sort_values())
clothesStyles_group.loc['Solidcolor'].sort_values().plot.bar()
```

```
upperOrange      67
upperBrown      114
upperGreen      132
upperPurple     136
upperYellow     148
upperPink       237
upperRed        239
upperBlue       366
upperGray       725
upperWhite     1181
upperBlack     1545
Name: Solidcolor, dtype: int64
```

Out[22]:

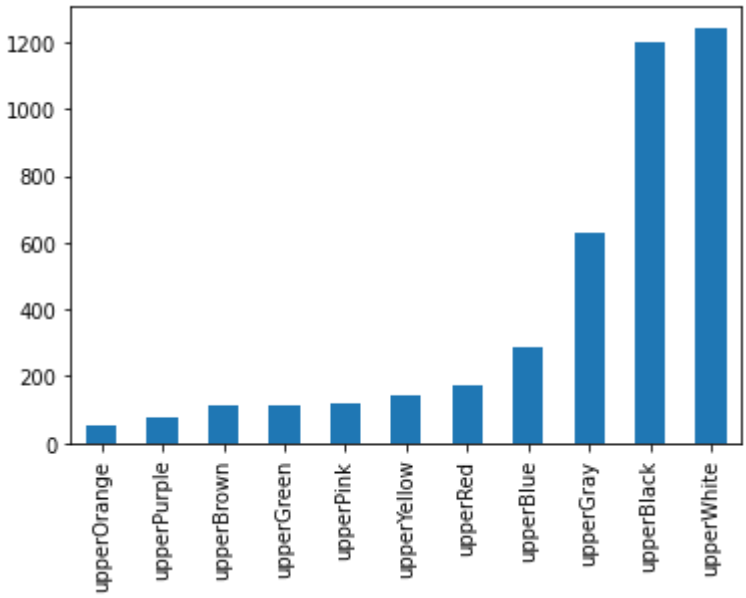
```
<AxesSubplot:>
```



In [23]:
```
print(clothesStyles_group.loc['multicolour'].sort_values())
clothesStyles_group.loc['multicolour'].sort_values().plot.bar()
```

```
upperOrange       50
upperPurple       76
upperBrown       110
upperGreen       112
upperPink        121
upperYellow      140
upperRed         170
upperBlue        284
upperGray        630
upperBlack      1199
upperWhite      1243
Name: multicolour, dtype: int64
<AxesSubplot:>
```

Out[23]:



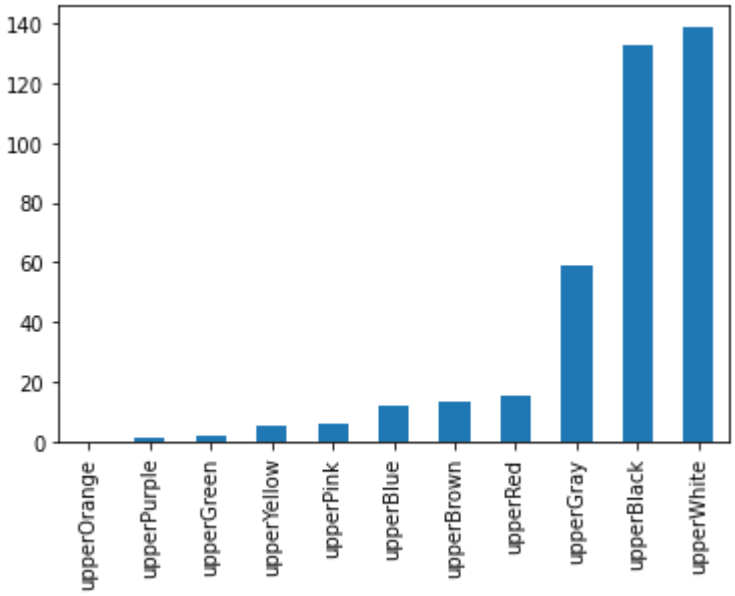
In [24]:
```
print(clothesStyles_group.loc['lattice'].sort_values())
clothesStyles_group.loc['lattice'].sort_values().plot.bar()
```

```
upperOrange       0
upperPurple       1
upperGreen        2
upperYellow       5
upperPink         6
upperBlue        12
upperBrown       13
upperRed         15
```

```
upperGray       59
upperBlack     133
upperWhite     139
Name: lattice, dtype: int64
```

Out[24]:

```
<AxesSubplot:>
```



In [25]:

```
color_count_df = train.copy(deep=True)
colors = ['upperBlack',
       'upperBrown', 'upperBlue', 'upperGreen', 'upperGray', 'upperOrange',
       'upperPink', 'upperPurple', 'upperRed', 'upperWhite', 'upperYellow']
def apply_color_count(series):
    count = 0
    for i in colors:
        if series[i] > 0:
            count = count + 1
    return count

color_count_df["color_count"] = color_count_df.apply(apply_color_count, axis=1)

color_count_df = color_count_df.drop(['upperLength', 'hairStyles', 'upperBlack',
       'upperBrown', 'upperBlue', 'upperGreen', 'upperGray', 'upperOrange',
       'upperPink', 'upperPurple', 'upperRed', 'upperWhite', 'upperYellow'], axis=1)

color_count_df.head()
```

Out[25]:

	name	clothesStyles	color_count
0	img_qh_train2_00000006345.jpg	Solidcolor	1
1	img_qh_train2_00001008825.jpg	multicolour	3
2	img_qh_train2_00002004117.jpg	multicolour	2
3	img_qh_train2_00003002526.jpg	multicolour	2
4	img_qh_train2_00004004439.jpg	Solidcolor	1

In [26]:

```
for index, row in color_count_df.iterrows():
    if row['clothesStyles'] == 'Solidcolor' and row['color_count'] > 1:
        print('[error label Solidcolor] name: ', row['name'])
    if row['clothesStyles'] == 'lattice' and row['color_count'] < 2:
        print('[error label lattice] name: ', row['name'])
    if row['clothesStyles'] == 'multicolour' and row['color_count'] < 2:
        print('[error label multicolour] name: ', row['name'])
```

```
[error label multicolour] name:  img_qh_train2_02368001422.jpg
[error label multicolour] name:  img_qh_train2_03848006214.jpg
```

```
[error label multicolour] name:  img_qh_train2_03917003653.jpg
[error label multicolour] name:  img_qh_train2_03941004803.jpg
[error label Solidcolor] name:  img_qh_train2_03952001195.jpg
[error label Solidcolor] name:  img_qh_train2_03999004246.jpg
```

In [27]:

```
color_frequency_df = color_count_df.copy(deep=True)
def apply_color_frequency_item(series, frequency):
    if series['color_count'] == frequency:
        return 1
    else:
        return 0

color_frequency_df["one"] = color_frequency_df.apply(apply_color_frequency_item, args=(1,), axis
color_frequency_df["two"] = color_frequency_df.apply(apply_color_frequency_item, args=(2,), axis
color_frequency_df["three"] = color_frequency_df.apply(apply_color_frequency_item, args=(3,), ax

color_frequency_df = color_frequency_df.drop(['name', 'color_count'], axis=1)
color_frequency_df.head()
```

Out[27]:

	clothesStyles	one	two	three
0	Solidcolor	1	0	0
1	multicolour	0	0	1
2	multicolour	0	1	0
3	multicolour	0	1	0
4	Solidcolor	1	0	0

In [28]:

```
color_frequency_df = color_frequency_df.groupby('clothesStyles').sum()
color_frequency_df.head()
```

Out[28]:

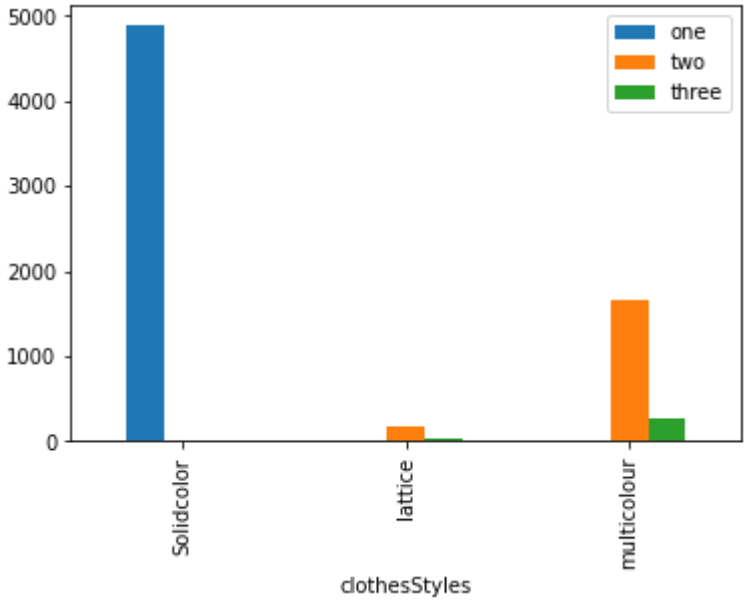
clothesStyles	one	two	three
Solidcolor	4885	1	1
lattice	0	164	19
multicolour	4	1652	269

In [29]:

```
color_frequency_df.plot.bar()
```

Out[29]:

```
<AxesSubplot:xlabel='clothesStyles'>
```


5000
4000
3000
2000
1000
0
one
two
three
Solidcolor
lattice
multicolour
clothesStyles