Exploratory Analysis of Star Wars Movie Dataset

Task 1: Data Preparation ¶

In [1]:

```
# Importing the required numpy,pandas and matplotlib packages for Processing and visual
izing the data.

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

Renaming the column headers meaningfully as required for the data preprocessing.

In [2]:

```
name=['RespondentID','Have you seen any of the 6 films in the Star Wars franchise?','Do
you consider yourself to be a fan of the Star Wars film franchise?',
      'Episode 1_seen','Episode 2_seen','Episode 3_seen','Episode 4_seen','Episode 5_se
en','Episode 6_seen','Episode 1_Rate','Episode 2_Rate','Episode 3_Rate',
      'Episode 4_Rate','Episode 5_Rate','Episode 6_Rate','Han Solo','Luke Skywalker','P
rincess Leia Organa','Anakin Skywalker','Obi Wan Kenobi','Emperor Palpatine',
      'Darth Vader','Lando Calrissian','Boba Fett','C-3P0','R2 D2','Jar Jar Binks','Pad
me Amidala','Yoda','Which character shot first?',
      'Are you familiar with the Expanded Universe?','Do you consider yourself to be a
 fan of the Expanded Universe?',
      'Do you consider yourself to be a fan of the Star Trek franchise?','Gender','Age'
,'Household Income','Education','Location']
```

In [3]:

```
data= 'StarWars.csv'
```

In [4]:

```
# Since it is CSV file it is separated by ',' and skipping the first 2 rows of original
header data and renaming it as shown above.

df= pd.read_csv(data,sep=',',header=None,names=name,skiprows=2)
```

In [5]:

```
pd.options.display.max_columns = None     # To display all columns
```

Check Data Types

Displaying the data to verify the loaded CSV file is unchanged from the original file.

In [6]:

```
df.head()
```

Out[6]:

	RespondentID	Have you seen any of the 6 films in the Star Wars franchise?	Do you consider yourself to be a fan of the Star Wars film franchise?	Episode 1_seen	Episode 2_seen	Episode 3_seen	Episode 4_seen	Episode 5_seen	Epi 6_
0	3292879998	Yes	Yes	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R
1	3292879538	No	NaN	NaN	NaN	NaN	NaN	NaN	
2	3292765271	Yes	No	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	NaN	NaN	
3	3292763116	Yes	Yes	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R
4	3292731220	Yes	Yes	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R

In [7]:

```
# Checking the size of the data frame.
df.shape
```

Out[7]:

```
(1186, 38)
```

In [8]:

```
# Checking the data types of all the columns.
df.dtypes
```

Out[8]:

```
RespondentID
int64
Have you seen any of the 6 films in the Star Wars franchise?
object
Do you consider yourself to be a fan of the Star Wars film franchise?
object
Episode 1_seen
object
Episode 2_seen
object
Episode 3_seen
object
Episode 4_seen
object
Episode 5_seen
object
Episode 6_seen
object
Episode 1_Rate                                                    f
loat64
Episode 2_Rate                                                    f
loat64
Episode 3_Rate                                                    f
loat64
Episode 4_Rate                                                    f
loat64
Episode 5_Rate                                                    f
loat64
Episode 6_Rate                                                    f
loat64
Han Solo
object
Luke Skywalker
object
Princess Leia Organa
object
Anakin Skywalker
object
Obi Wan Kenobi
object
Emperor Palpatine
object
Darth Vader
object
Lando Calrissian
object
Boba Fett
object
C-3P0
object
R2 D2
object
Jar Jar Binks
object
Padme Amidala
object
Yoda
object
Which character shot first?
```

```
object
Are you familiar with the Expanded Universe?
object
Do you consider yourself to be a fan of the Expanded Universe?
object
Do you consider yourself to be a fan of the Star Trek franchise?
object
Gender
object
Age
object
Household Income
object
Education
object
Location
object
dtype: object
```

In [9]:

```
Numerical_cols = df.columns[df.dtypes == np.number].tolist()     # Getting the list of
 Numerical variables in the dataframe.
```

In [10]:

```
categorical_cols = df.columns[df.dtypes == np.object].tolist()   # Getting the list of
 Categorical variables in the dataframe.
```

Typos

- We observed the data set,There are few typo errors which are rectified by replacing it with correct data as shown below.
- Typo errors such as `‘Yess’, ’Noo’, ’no’, ‘yes’` . Along with this case sensitive issue also is there which has been replaced all types of yes to **Yes** and all types of no to **No** using the fumction **replace()**
- And for Gender `‘male’,’Male’,’Female’,’female’,’f’` has been replaced by **male** and **female** for all other Gender typo errors

In [11]:

```
df['Do you consider yourself to be a fan of the Star Wars film franchise?']=df['Do you
 consider yourself to be a fan of the Star Wars film franchise?'].replace('Yess', 'Yes'
).replace('Noo','No')
df['Do you consider yourself to be a fan of the Expanded Universe?']=df['Do you conside
r yourself to be a fan of the Expanded Universe?'].replace('Yess', 'Yes')
df['Do you consider yourself to be a fan of the Star Trek franchise?']=df['Do you consi
der yourself to be a fan of the Star Trek franchise?'].replace('Yess', 'Yes').replace(
'Noo','No').replace('no','No').replace('yes','Yes')
df['Gender']=df['Gender'].replace('Male','male').replace('Female','female').replace('F'
,'female')
```

Extra-WhiteSpaces

- Removing the extra white spaces in the entire data set by using the **`strip()`** functions.

In [12]:

```
for col in categorical_cols:
    print(col)
    df[col] = df[col].str.strip()
    print(df[col].unique())
```

```
Have you seen any of the 6 films in the Star Wars franchise?
['Yes' 'No']
Do you consider yourself to be a fan of the Star Wars film franchise?
['Yes' nan 'No']
Episode 1_seen
['Star Wars: Episode I  The Phantom Menace' nan]
Episode 2_seen
['Star Wars: Episode II  Attack of the Clones' nan]
Episode 3_seen
['Star Wars: Episode III  Revenge of the Sith' nan]
Episode 4_seen
['Star Wars: Episode IV  A New Hope' nan]
Episode 5_seen
['Star Wars: Episode V The Empire Strikes Back' nan]
Episode 6_seen
['Star Wars: Episode VI Return of the Jedi' nan]
Han Solo
['Very favorably' nan 'Somewhat favorably'
 'Neither favorably nor unfavorably (neutral)' 'Somewhat unfavorably'
 'Unfamiliar (N/A)' 'Very unfavorably']
Luke Skywalker
['Very favorably' nan 'Somewhat favorably' 'Somewhat unfavorably'
 'Neither favorably nor unfavorably (neutral)' 'Very unfavorably'
 'Unfamiliar (N/A)']
Princess Leia Organa
['Very favorably' nan 'Somewhat favorably' 'Somewhat unfavorably'
 'Neither favorably nor unfavorably (neutral)' 'Very unfavorably'
 'Unfamiliar (N/A)']
Anakin Skywalker
['Very favorably' nan 'Somewhat favorably' 'Somewhat unfavorably'
 'Neither favorably nor unfavorably (neutral)' 'Very unfavorably'
 'Unfamiliar (N/A)']
Obi Wan Kenobi
['Very favorably' nan 'Somewhat favorably' 'Very unfavorably'
 'Neither favorably nor unfavorably (neutral)' 'Somewhat unfavorably'
 'Unfamiliar (N/A)']
Emperor Palpatine
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Very unfavorably' 'Neither favorably nor unfavorably (neutral)'
 'Somewhat unfavorably']
Darth Vader
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Somewhat unfavorably' 'Very unfavorably'
 'Neither favorably nor unfavorably (neutral)']
Lando Calrissian
['Unfamiliar (N/A)' nan 'Somewhat favorably'
 'Neither favorably nor unfavorably (neutral)' 'Very favorably'
 'Somewhat unfavorably' 'Very unfavorably']
Boba Fett
['Unfamiliar (N/A)' nan 'Somewhat unfavorably' 'Very favorably'
 'Somewhat favorably' 'Neither favorably nor unfavorably (neutral)'
 'Very unfavorably']
C-3P0
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Neither favorably nor unfavorably (neutral)' 'Somewhat unfavorably'
 'Very unfavorably']
R2 D2
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Neither favorably nor unfavorably (neutral)' 'Somewhat unfavorably'
 'Very unfavorably']
Jar Jar Binks
```

```
['Very favorably' nan 'Unfamiliar (N/A)' 'Very unfavorably'
 'Somewhat favorably' 'Somewhat unfavorably'
 'Neither favorably nor unfavorably (neutral)']
Padme Amidala
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Neither favorably nor unfavorably (neutral)' 'Somewhat unfavorably'
 'Very unfavorably']
Yoda
['Very favorably' nan 'Unfamiliar (N/A)' 'Somewhat favorably'
 'Very unfavorably' 'Neither favorably nor unfavorably (neutral)'
 'Somewhat unfavorably']
Which character shot first?
["I don't understand this question" nan 'Greedo' 'Han']
Are you familiar with the Expanded Universe?
['Yes' nan 'No']
Do you consider yourself to be a fan of the Expanded Universe?
['No' nan 'Yes']
Do you consider yourself to be a fan of the Star Trek franchise?
['No' 'Yes' nan 'no']
Gender
['male' nan 'female']
Age
['18-29' nan '500' '30-44' '> 60' '45-60']
Household Income
[nan '$0 - $24,999' '$100,000 - $149,999' '$25,000 - $49,999'
 '$50,000 - $99,999' '$150,000+']
Education
['High school degree' 'Bachelor degree' 'Some college or Associate degree'
 nan 'Graduate degree' 'Less than high school degree']
Location
['South Atlantic' 'West South Central' 'West North Central'
 'Middle Atlantic' 'East North Central' 'Pacific' nan 'Mountain'
 'New England' 'East South Central']
```

Upper Case

- Converting the entire String data to the Uppercases by using the function **`str.upper()`** to the data frame.

In [13]:

```
for col in df.columns[1:3]:
    df[col] = df[col].str.upper()

for col in df.columns[15:]:
    df[col] = df[col].str.upper()
```

Sanity Check

- Performing the Sanity checks of all attributes by getting the counts of the each unique variable.
- After getting the count, carefully examining the each attribute manually to see the presence of impossible values.

In [14]:

```
for col in Numerical_cols:
    print(col)
    print(df[col].value_counts(dropna=False))
    print('\n')
```

```
Episode 1_Rate
NaN    351
4.0    237
6.0    168
3.0    130
1.0    129
5.0    100
2.0     71
Name: Episode 1_Rate, dtype: int64


Episode 2_Rate
NaN    350
5.0    300
4.0    183
2.0    116
3.0    103
6.0    102
1.0     32
Name: Episode 2_Rate, dtype: int64


Episode 3_Rate
NaN    351
6.0    217
5.0    203
4.0    182
3.0    150
2.0     47
1.0     36
Name: Episode 3_Rate, dtype: int64


Episode 4_Rate
NaN    350
1.0    204
6.0    161
2.0    135
4.0    130
3.0    127
5.0     79
Name: Episode 4_Rate, dtype: int64


Episode 5_Rate
NaN    350
1.0    289
2.0    235
5.0    118
3.0    106
4.0     47
6.0     41
Name: Episode 5_Rate, dtype: int64


Episode 6_Rate
NaN    350
2.0    232
3.0    220
1.0    146
6.0    145
```

```
4.0     57
5.0     36
Name: Episode 6_Rate, dtype: int64
```

In [15]:

```
for categorical_col in categorical_cols:
    print(categorical_col)
    print(df[categorical_col].value_counts(dropna=False))
    print('\n')
```

```
Have you seen any of the 6 films in the Star Wars franchise?
YES    936
NO     250
Name: Have you seen any of the 6 films in the Star Wars franchise?, dtype:
int64


Do you consider yourself to be a fan of the Star Wars film franchise?
YES    552
NaN    350
NO     284
Name: Do you consider yourself to be a fan of the Star Wars film franchis
e?, dtype: int64


Episode 1_seen
Star Wars: Episode I  The Phantom Menace    673
NaN                                         513
Name: Episode 1_seen, dtype: int64


Episode 2_seen
NaN                                           615
Star Wars: Episode II  Attack of the Clones    571
Name: Episode 2_seen, dtype: int64


Episode 3_seen
NaN                                           636
Star Wars: Episode III  Revenge of the Sith    550
Name: Episode 3_seen, dtype: int64


Episode 4_seen
Star Wars: Episode IV  A New Hope    607
NaN                                  579
Name: Episode 4_seen, dtype: int64


Episode 5_seen
Star Wars: Episode V The Empire Strikes Back    758
NaN                                             428
Name: Episode 5_seen, dtype: int64


Episode 6_seen
Star Wars: Episode VI Return of the Jedi    738
NaN                                         448
Name: Episode 6_seen, dtype: int64


Han Solo
VERY FAVORABLY                                610
NaN                                           357
SOMEWHAT FAVORABLY                            151
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)    44
UNFAMILIAR (N/A)                               15
SOMEWHAT UNFAVORABLY                            8
VERY UNFAVORABLY                                1
Name: Han Solo, dtype: int64
```

```
Luke Skywalker
VERY FAVORABLY                                552
NaN                                           355
SOMEWHAT FAVORABLY                            219
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)    38
SOMEWHAT UNFAVORABLY                           13
UNFAMILIAR (N/A)                                6
VERY UNFAVORABLY                                3
Name: Luke Skywalker, dtype: int64


Princess Leia Organa
VERY FAVORABLY                                547
NaN                                           355
SOMEWHAT FAVORABLY                            210
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)    48
SOMEWHAT UNFAVORABLY                           12
UNFAMILIAR (N/A)                                8
VERY UNFAVORABLY                                6
Name: Princess Leia Organa, dtype: int64


Anakin Skywalker
NaN                                           363
SOMEWHAT FAVORABLY                            269
VERY FAVORABLY                                245
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)   135
SOMEWHAT UNFAVORABLY                           83
UNFAMILIAR (N/A)                               52
VERY UNFAVORABLY                               39
Name: Anakin Skywalker, dtype: int64


Obi Wan Kenobi
VERY FAVORABLY                                591
NaN                                           361
SOMEWHAT FAVORABLY                            159
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)    43
UNFAMILIAR (N/A)                               17
SOMEWHAT UNFAVORABLY                            8
VERY UNFAVORABLY                                7
Name: Obi Wan Kenobi, dtype: int64


Emperor Palpatine
NaN                                           372
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)   213
UNFAMILIAR (N/A)                              156
SOMEWHAT FAVORABLY                            143
VERY UNFAVORABLY                              124
VERY FAVORABLY                                110
SOMEWHAT UNFAVORABLY                           68
Name: Emperor Palpatine, dtype: int64


Darth Vader
NaN                                           360
VERY FAVORABLY                                310
SOMEWHAT FAVORABLY                            171
VERY UNFAVORABLY                              149
```

```
SOMEWHAT UNFAVORABLY                            102
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)      84
UNFAMILIAR (N/A)                                 10
Name: Darth Vader, dtype: int64


Lando Calrissian
NaN                                             366
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)     236
SOMEWHAT FAVORABLY                              223
UNFAMILIAR (N/A)                                148
VERY FAVORABLY                                  142
SOMEWHAT UNFAVORABLY                             63
VERY UNFAVORABLY                                  8
Name: Lando Calrissian, dtype: int64


Boba Fett
NaN                                             374
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)     248
SOMEWHAT FAVORABLY                              153
VERY FAVORABLY                                  138
UNFAMILIAR (N/A)                                132
SOMEWHAT UNFAVORABLY                             96
VERY UNFAVORABLY                                 45
Name: Boba Fett, dtype: int64


C-3P0
VERY FAVORABLY                                  474
NaN                                             359
SOMEWHAT FAVORABLY                              229
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)      79
SOMEWHAT UNFAVORABLY                             23
UNFAMILIAR (N/A)                                 15
VERY UNFAVORABLY                                  7
Name: C-3P0, dtype: int64


R2 D2
VERY FAVORABLY                                  562
NaN                                             356
SOMEWHAT FAVORABLY                              185
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)      57
SOMEWHAT UNFAVORABLY                             10
UNFAMILIAR (N/A)                                 10
VERY UNFAVORABLY                                  6
Name: R2 D2, dtype: int64


Jar Jar Binks
NaN                                             365
VERY UNFAVORABLY                                204
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)     164
SOMEWHAT FAVORABLY                              130
VERY FAVORABLY                                  112
UNFAMILIAR (N/A)                                109
SOMEWHAT UNFAVORABLY                            102
Name: Jar Jar Binks, dtype: int64
```

```
Padme Amidala
NaN                                             372
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)     207
SOMEWHAT FAVORABLY                              183
VERY FAVORABLY                                  168
UNFAMILIAR (N/A)                                164
SOMEWHAT UNFAVORABLY                             58
VERY UNFAVORABLY                                 34
Name: Padme Amidala, dtype: int64


Yoda
VERY FAVORABLY                                  605
NaN                                             360
SOMEWHAT FAVORABLY                              144
NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)      51
UNFAMILIAR (N/A)                                 10
VERY UNFAVORABLY                                  8
SOMEWHAT UNFAVORABLY                              8
Name: Yoda, dtype: int64


Which character shot first?
NaN                                 358
HAN                                 325
I DON'T UNDERSTAND THIS QUESTION    306
GREEDO                              197
Name: Which character shot first?, dtype: int64


Are you familiar with the Expanded Universe?
NO     615
NaN    358
YES    213
Name: Are you familiar with the Expanded Universe?, dtype: int64


Do you consider yourself to be a fan of the Expanded Universe?
NaN    973
NO     114
YES     99
Name: Do you consider yourself to be a fan of the Expanded Universe?, dtyp
e: int64


Do you consider yourself to be a fan of the Star Trek franchise?
NO     641
YES    427
NaN    118
Name: Do you consider yourself to be a fan of the Star Trek franchise?, dt
ype: int64


Gender
FEMALE    549
MALE      497
NaN       140
Name: Gender, dtype: int64


Age
```

```
45-60    291
> 60     269
30-44    268
18-29    217
NaN      140
500        1
Name: Age, dtype: int64


Household Income
NaN                    328
$50,000 - $99,999      298
$25,000 - $49,999      186
$100,000 - $149,999    141
$0 - $24,999           138
$150,000+               95
Name: Household Income, dtype: int64


Education
SOME COLLEGE OR ASSOCIATE DEGREE    328
BACHELOR DEGREE                     321
GRADUATE DEGREE                     275
NaN                                 150
HIGH SCHOOL DEGREE                  105
LESS THAN HIGH SCHOOL DEGREE          7
Name: Education, dtype: int64


Location
EAST NORTH CENTRAL    181
PACIFIC               175
SOUTH ATLANTIC        170
NaN                   143
MIDDLE ATLANTIC       122
WEST SOUTH CENTRAL    110
WEST NORTH CENTRAL     93
MOUNTAIN               79
NEW ENGLAND            75
EAST SOUTH CENTRAL     38
Name: Location, dtype: int64
```

- After examining we found out that in Age column value `500` is present which is impossible for anyone to have 500 years.
- So required action is taken by replacing `500` with `NaN` .

In [16]:

```
df['Age']=df['Age'].replace('500', np.nan)
```

Assumption

- When analysing the data set we found out that people who did not see even one episode of the StarWar series answered `No` .
- That means if the person has not seen he cannot answer the rest of the questions related to the episodes. So all other columns are empty i.e. `NA` . It is unfair to fill the empty rows with the meaningful data as they have not watched even one episode.
- So we are considering the data of those who saw atleast one movie and answered the further questions.
- The rows of that particular ID with answer `NO` for the column `'Have you seen any of the 6 films in the Star Wars franchise?'` has been removed.

In [17]:

```
# Deleting the row indexes from dataFrame for the column as mentioned below with the an
swer NO.

indexNames = df[df['Have you seen any of the 6 films in the Star Wars franchise?'] ==
"NO" ].index
df.drop(indexNames , inplace=True)
```

In [18]:

```
# Below table shows the data with the people who saw atleast one movie.
df.head()
```

Out[18]:

	RespondentID	Have you seen any of the 6 films in the Star Wars franchise?	Do you consider yourself to be a fan of the Star Wars film franchise?	Episode 1_seen	Episode 2_seen	Episode 3_seen	Episode 4_seen	Episode 5_seen	Epi 6_
0	3292879998	YES	YES	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R
2	3292765271	YES	NO	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	NaN	NaN	
3	3292763116	YES	YES	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R
4	3292731220	YES	YES	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R
5	3292719380	YES	YES	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	Star Wars: Episode IV A New Hope	Star Wars: Episode V The Empire Strikes Back	Ep R

Missing Values

- We found that there were many missing values in the original data set.
- So for all the missing values we are replacing it with the `mode` of that corresponding column to the data frame for the further analysis as shown below seperately for the Categorical and Numerical attributes.

In [19]:

```
# Checking the number of missing values in all the columns.
df.isnull().sum()
```

Out[19]:

```
RespondentID                                                                    0
Have you seen any of the 6 films in the Star Wars franchise?                    0
Do you consider yourself to be a fan of the Star Wars film franchise?    1
00
Episode 1_seen                                                                  2
63
Episode 2_seen                                                                  3
65
Episode 3_seen                                                                  3
86
Episode 4_seen                                                                  3
29
Episode 5_seen                                                                  1
78
Episode 6_seen                                                                  1
98
Episode 1_Rate                                                                  1
01
Episode 2_Rate                                                                  1
00
Episode 3_Rate                                                                  1
01
Episode 4_Rate                                                                  1
00
Episode 5_Rate                                                                  1
00
Episode 6_Rate                                                                  1
00
Han Solo                                                                        1
07
Luke Skywalker                                                                  1
05
Princess Leia Organa                                                            1
05
Anakin Skywalker                                                                1
13
Obi Wan Kenobi                                                                  1
11
Emperor Palpatine                                                               1
22
Darth Vader                                                                     1
10
Lando Calrissian                                                                1
16
Boba Fett                                                                       1
24
C-3P0                                                                           1
09
R2 D2                                                                           1
06
Jar Jar Binks                                                                   1
15
Padme Amidala                                                                   1
22
Yoda                                                                            1
10
Which character shot first?                                                     1
```

```
08
Are you familiar with the Expanded Universe?                          1
08
Do you consider yourself to be a fan of the Expanded Universe?        7
23
Do you consider yourself to be a fan of the Star Trek franchise?      1
08
Gender                                                                1
16
Age                                                                   1
17
Household Income                                                      2
61
Education                                                             1
20
Location                                                              1
18
dtype: int64
```

- Here for all the columns from 3 to 9, `True` is replaced for the person answered that particular `Episode Name` for seen episodes and `False` for the not seen episodes which is left blank.

In [20]:

```
Dict2 = {
    'Star Wars: Episode I  The Phantom Menace':True,
    'Star Wars: Episode II  Attack of the Clones':True,
    'Star Wars: Episode III  Revenge of the Sith':True,
    'Star Wars: Episode IV  A New Hope':True,
    'Star Wars: Episode V The Empire Strikes Back':True,
    'Star Wars: Episode VI Return of the Jedi':True,
     np.nan:False
}
for col in df.columns[3:9]:
    df[col] = df[col].map(Dict2)
```

In [21]:

```
# Replacing the missing columns by mode of that particular column.
for Numerical_cols in Numerical_cols:
    df[Numerical_cols].fillna(df[Numerical_cols].mode()[0],inplace= True)
```

In [22]:

```
# Replacing the missing columns by mode of that particular column.
for col in categorical_cols:
    print(col)
    df[col].fillna(df[col].mode()[0],inplace=True)
    print(df[col].unique())
```

```
Have you seen any of the 6 films in the Star Wars franchise?
['YES']
Do you consider yourself to be a fan of the Star Wars film franchise?
['YES' 'NO']
Episode 1_seen
[ True False]
Episode 2_seen
[ True False]
Episode 3_seen
[ True False]
Episode 4_seen
[ True False]
Episode 5_seen
[ True False]
Episode 6_seen
[ True False]
Han Solo
['VERY FAVORABLY' 'SOMEWHAT FAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'SOMEWHAT UNFAVORABLY'
 'UNFAMILIAR (N/A)' 'VERY UNFAVORABLY']
Luke Skywalker
['VERY FAVORABLY' 'SOMEWHAT FAVORABLY' 'SOMEWHAT UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'VERY UNFAVORABLY'
 'UNFAMILIAR (N/A)']
Princess Leia Organa
['VERY FAVORABLY' 'SOMEWHAT FAVORABLY' 'SOMEWHAT UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'VERY UNFAVORABLY'
 'UNFAMILIAR (N/A)']
Anakin Skywalker
['VERY FAVORABLY' 'SOMEWHAT FAVORABLY' 'SOMEWHAT UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'VERY UNFAVORABLY'
 'UNFAMILIAR (N/A)']
Obi Wan Kenobi
['VERY FAVORABLY' 'SOMEWHAT FAVORABLY' 'VERY UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'SOMEWHAT UNFAVORABLY'
 'UNFAMILIAR (N/A)']
Emperor Palpatine
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'VERY UNFAVORABLY' 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)'
 'SOMEWHAT UNFAVORABLY']
Darth Vader
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'SOMEWHAT UNFAVORABLY' 'VERY UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)']
Lando Calrissian
['UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'VERY FAVORABLY'
 'SOMEWHAT UNFAVORABLY' 'VERY UNFAVORABLY']
Boba Fett
['UNFAMILIAR (N/A)' 'SOMEWHAT UNFAVORABLY' 'VERY FAVORABLY'
 'SOMEWHAT FAVORABLY' 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)'
 'VERY UNFAVORABLY']
C-3P0
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'SOMEWHAT UNFAVORABLY'
 'VERY UNFAVORABLY']
R2 D2
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'SOMEWHAT UNFAVORABLY'
 'VERY UNFAVORABLY']
Jar Jar Binks
```

```
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'VERY UNFAVORABLY'
 'SOMEWHAT FAVORABLY' 'SOMEWHAT UNFAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)']
Padme Amidala
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)' 'SOMEWHAT UNFAVORABLY'
 'VERY UNFAVORABLY']
Yoda
['VERY FAVORABLY' 'UNFAMILIAR (N/A)' 'SOMEWHAT FAVORABLY'
 'VERY UNFAVORABLY' 'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)'
 'SOMEWHAT UNFAVORABLY']
Which character shot first?
["I DON'T UNDERSTAND THIS QUESTION" 'GREEDO' 'HAN']
Are you familiar with the Expanded Universe?
['YES' 'NO']
Do you consider yourself to be a fan of the Expanded Universe?
['NO' 'YES']
Do you consider yourself to be a fan of the Star Trek franchise?
['NO' 'YES']
Gender
['MALE' 'FEMALE']
Age
['18-29' '45-60' '30-44' '> 60']
Household Income
['$50,000 - $99,999' '$0 - $24,999' '$100,000 - $149,999'
 '$25,000 - $49,999' '$150,000+']
Education
['HIGH SCHOOL DEGREE' 'SOME COLLEGE OR ASSOCIATE DEGREE' 'BACHELOR DEGREE'
 'GRADUATE DEGREE' 'LESS THAN HIGH SCHOOL DEGREE']
Location
['SOUTH ATLANTIC' 'WEST NORTH CENTRAL' 'MIDDLE ATLANTIC'
 'EAST NORTH CENTRAL' 'PACIFIC' 'MOUNTAIN' 'WEST SOUTH CENTRAL'
 'NEW ENGLAND' 'EAST SOUTH CENTRAL']
```

Encoding

- For the columns 15 to 29 encoding is done as shown below for respective strings for further analysis. These columns are the rated values for all the characters in the episodes.

In [23]:

```
Dict3 = {
    'VERY FAVORABLY':5,
    'SOMEWHAT FAVORABLY':4,
    'NEITHER FAVORABLY NOR UNFAVORABLY (NEUTRAL)':3,
    'SOMEWHAT UNFAVORABLY':2,
    'UNFAMILIAR (N/A)':0,
    'VERY UNFAVORABLY':1
}
for col in df.columns[15:29]:
    df[col] = df[col].map(Dict3)
```

In [24]:

```
df.head(10)
```

Out[24]:

	RespondentID	Have you seen any of the 6 films in the Star Wars franchise?	Do you consider yourself to be a fan of the Star Wars film franchise?	Episode 1_seen	Episode 2_seen	Episode 3_seen	Episode 4_seen	Episode 5_seen	Ep 6_
0	3292879998	YES	YES	True	True	True	True	True	
2	3292765271	YES	NO	True	True	True	False	False	
3	3292763116	YES	YES	True	True	True	True	True	
4	3292731220	YES	YES	True	True	True	True	True	
5	3292719380	YES	YES	True	True	True	True	True	
6	3292684787	YES	YES	True	True	True	True	True	
7	3292663732	YES	YES	True	True	True	True	True	
8	3292654043	YES	YES	True	True	True	True	True	
9	3292640424	YES	NO	False	True	False	False	False	
10	3292637870	YES	YES	False	False	False	False	False	

Task 2: Data Exploration

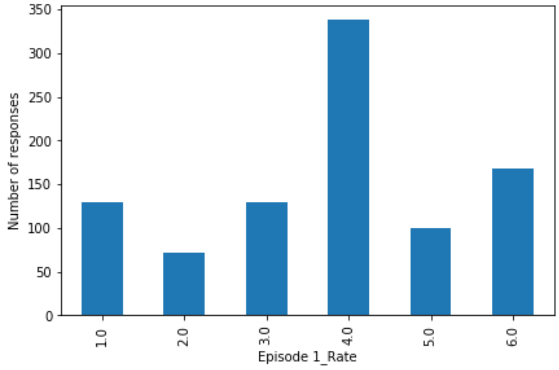
1. Explore the survey question:Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film. (Star Wars: Episode I The Phantom Menace; Star Wars: Episode II Attack of the Clones; Star Wars: Episode III Revenge of the Sith; Star Wars: Episode IV A New Hope; Star Wars: Episode V The Empire Strikes Back; Star Wars: Episode VI Return of the Jedi), then analysis how people rate Star Wars Movies

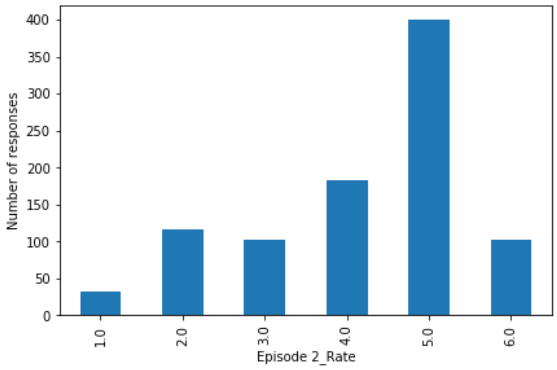
Column 10 contains the following string: 'Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film.

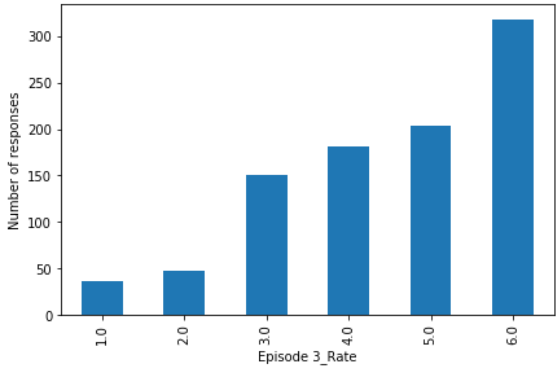
- Below chart shows that Episode 4 and 5 have higher ratings than any other StarWars parts.
- Episode 3 has got the least rating, that means people did not liked that part compared to other episodes.
- To summarise, People liked latest movies than the old ones.

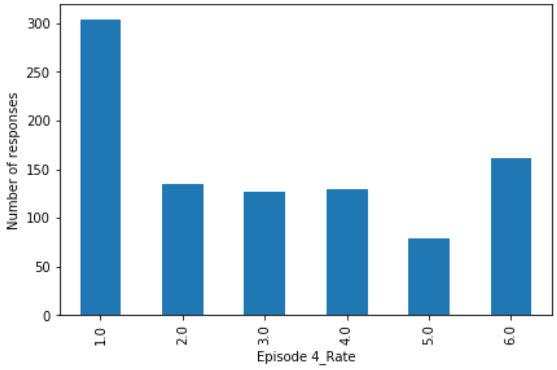
In [25]:

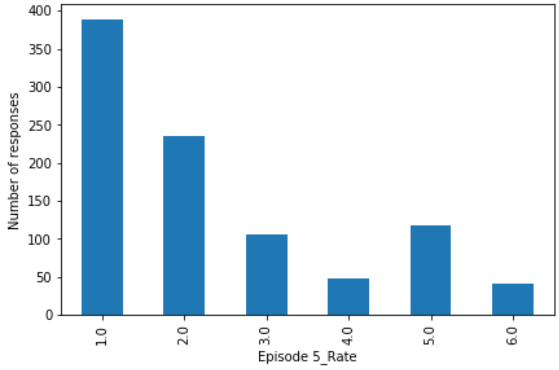
```
f = plt.figure(figsize=(15,15))
y=1

for x in range(9,15):
    col=df.columns[x]
    ax = f.add_subplot(3,2,y)
    y=y+1

    a=df[col].value_counts().sort_index()
    a.plot(kind='bar')
    plt.ylabel('Number of responses')
    plt.xlabel(col)
```

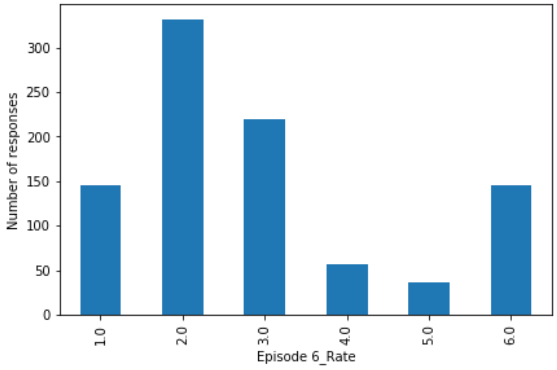

Number of responses
Episode 1_Rate
1.0
2.0
3.0
4.0
5.0
6.0


Number of responses
Episode 2_Rate
1.0
2.0
3.0
4.0
5.0
6.0


Number of responses
Episode 3_Rate
1.0
2.0
3.0
4.0
5.0
6.0

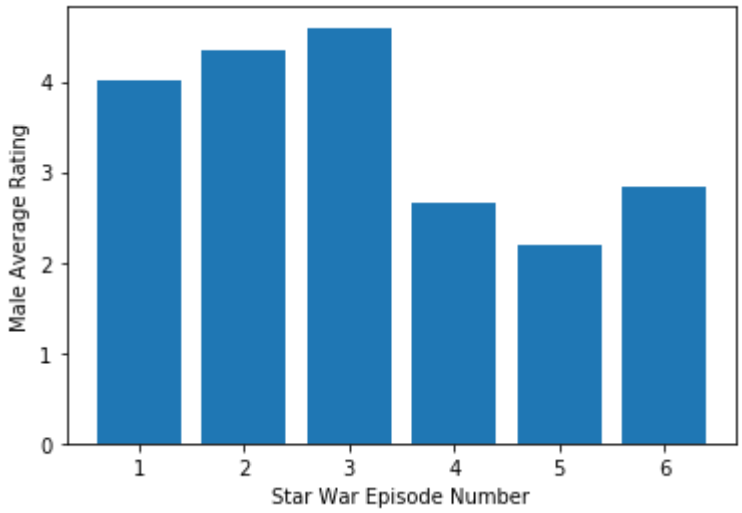

Number of responses
Episode 4_Rate
1.0
2.0
3.0
4.0
5.0
6.0

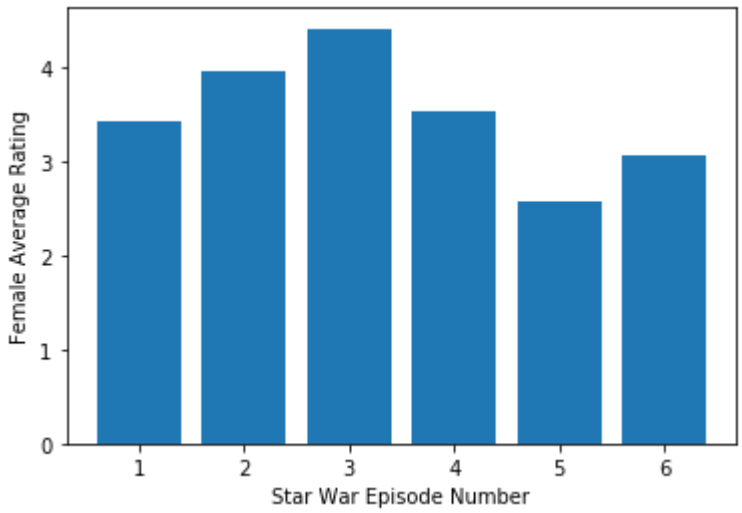

Number of responses
Episode 5_Rate
1.0
2.0
3.0
4.0
5.0
6.0


Number of responses
Episode 6_Rate
1.0
2.0
3.0
4.0
5.0
6.0

In [26]:

```
star_wars_males = males = df[df["Gender"] == "MALE"]
star_wars_females = females = df[df["Gender"] == "FEMALE"]

means_males = star_wars_males[star_wars_males.columns[9:15]].mean()
plt.bar(range(1,7), means_males)
plt.xlabel('Star War Episode Number')
plt.ylabel('Male Average Rating')
plt.show()


means_females = star_wars_females[star_wars_females.columns[9:15]].mean()
plt.bar(range(1,7), means_females)
plt.xlabel("Star War Episode Number")
plt.ylabel('Female Average Rating')
plt.show()
```





Task 2: Data Exploration

1. Explore the relationships between columns; at least 3 visualisations with plausible hypothesis

Relationship 1

Plausible hypothesis: To find out which age group has greater number of possible fans to Star Wars film franchise.

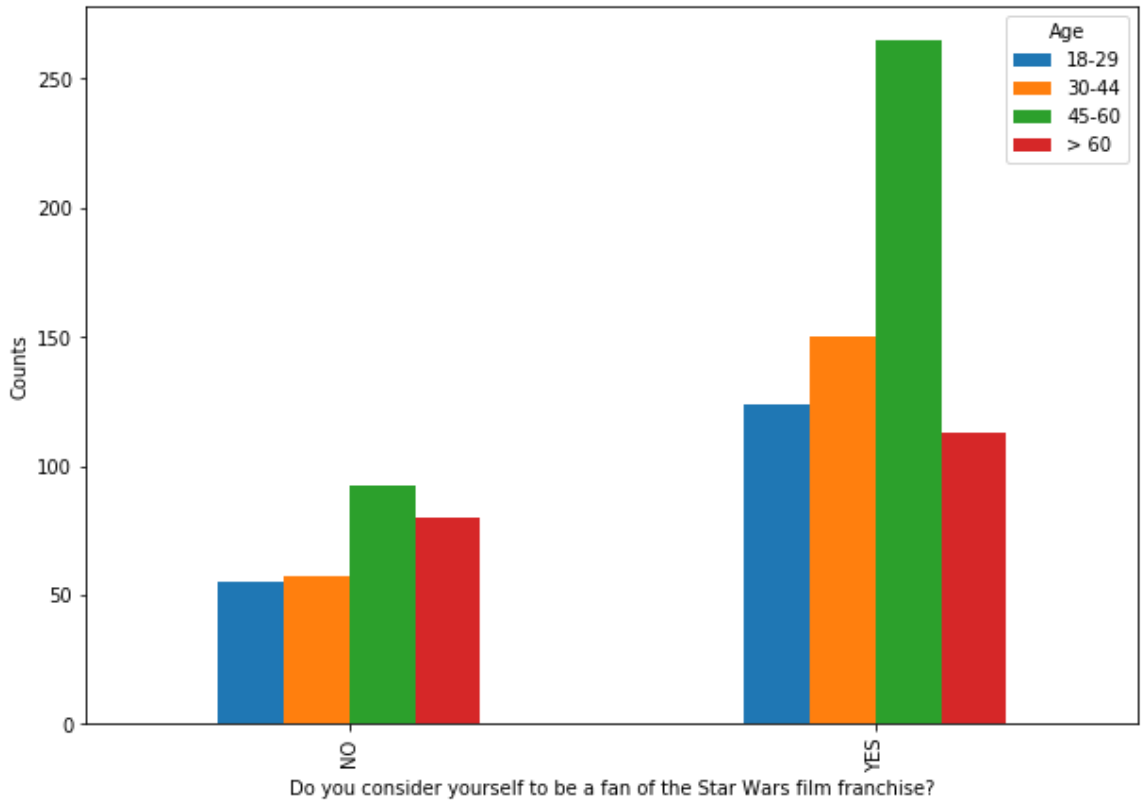
- When analysed from the below graph, Age group `45-60` has the highest number of `YES` for the above question asked in the survey.So, we can conclude that People between the age **`45-60`** are more tends to like the **Star Wars film franchise**.

In [27]:

```
x = df.groupby(['Do you consider yourself to be a fan of the Star Wars film franchise?'
,'Age'])['Do you consider yourself to be a fan of the Star Wars film franchise?'].size
().unstack()
x.plot.bar(figsize=(10,7))
plt.ylabel('Counts')
```

Out[27]:

```
Text(0, 0.5, 'Counts')
```



Relationship 2

Plausible hypothesis: After finding the age Group, now to find out which Gender are more likely to be fans of the Star Wars franchise.

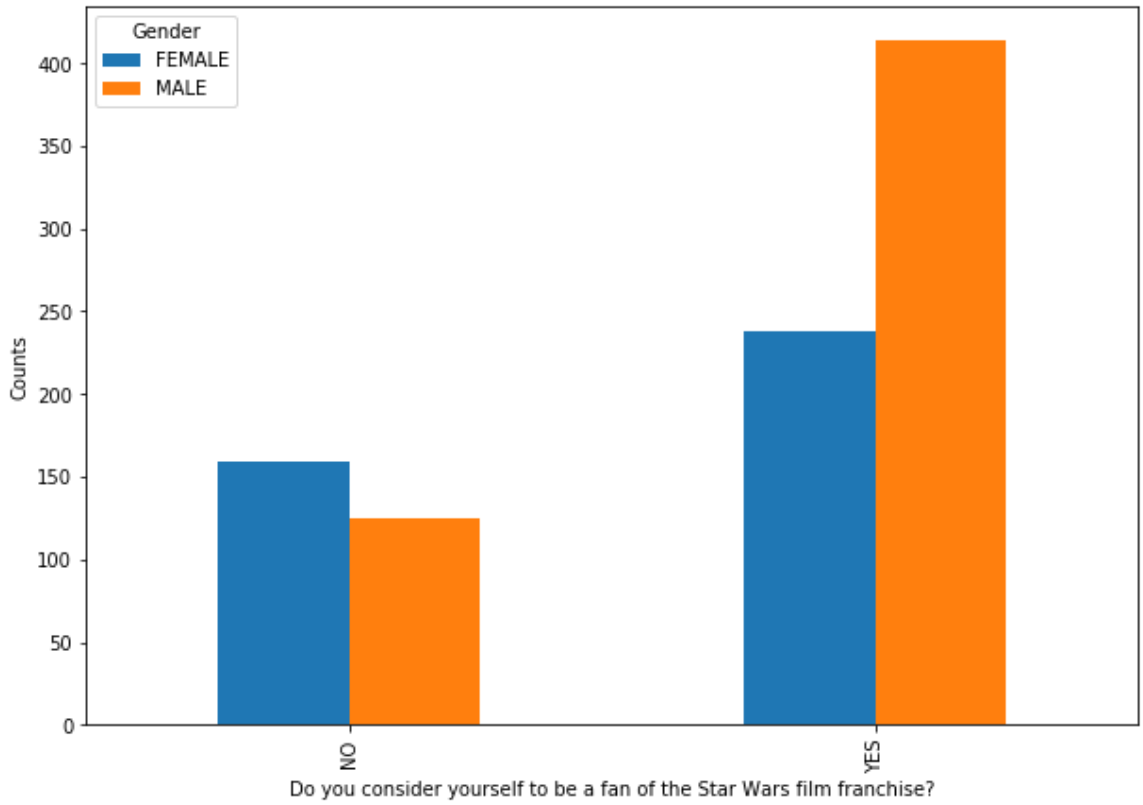
- When analysed from the below graph for **Gender** Column for the Question, we found that `MALE` has the highest number of `YES` for the above question asked in the survey.So, we can conclude that `MALE` possibly more tends to become fans for the **star war film franchise** than `FEMALE` .

In [28]:

```
x = df.groupby(['Do you consider yourself to be a fan of the Star Wars film franchise?'
,'Gender'])['Do you consider yourself to be a fan of the Star Wars film franchise?'].si
ze().unstack()
x.plot.bar(figsize=(10,7))
plt.ylabel('Counts')
```

Out[28]:

```
Text(0, 0.5, 'Counts')
```



Relationship 3

Plausible hypothesis: To find out which location has got the greater number of fans for the Star Trek franchise.

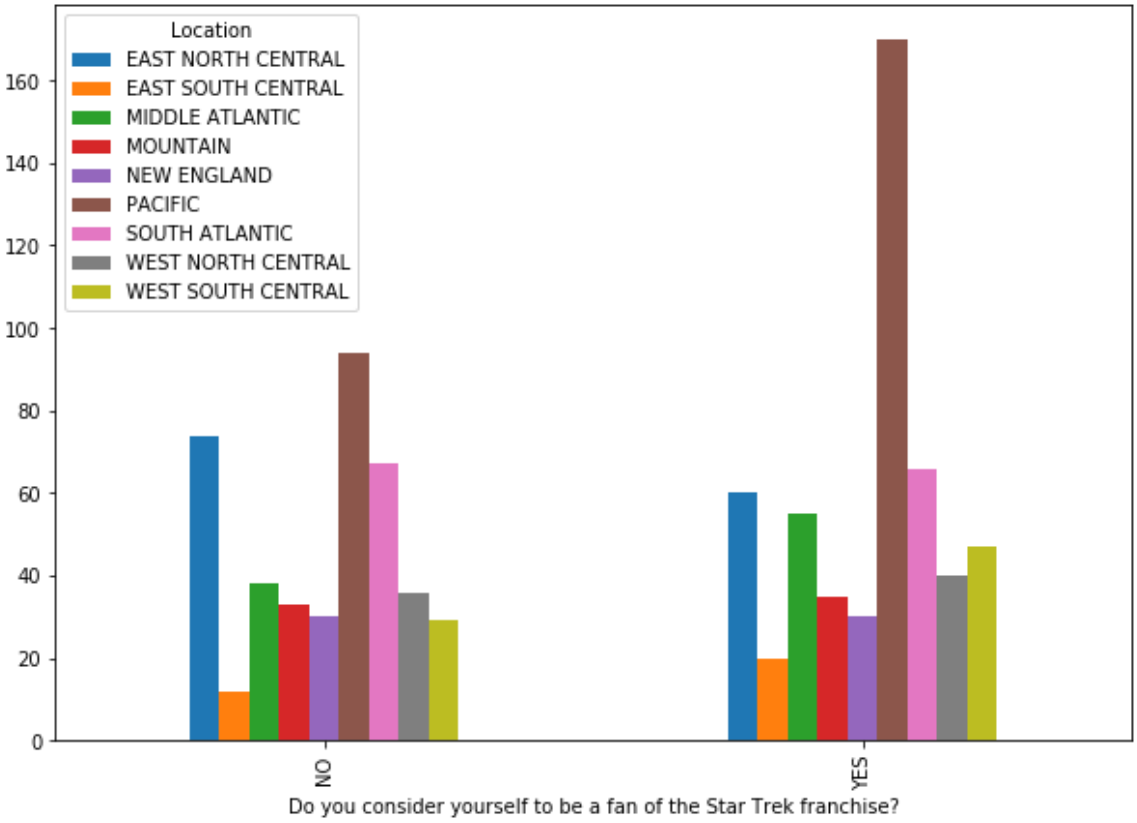
- When analysed from the below graph for **Location** Column for the Question above, we found that `PACIFIC` has the highest number of `YES` which shows **PACIFIC** seems to have a greater number of fans for the **Star Trek franchise** than other locations.

In [29]:

```
x = df.groupby(['Do you consider yourself to be a fan of the Star Trek franchise?','Location'])['Do you consider yourself to be a fan of the Star Trek franchise?'].size().unstack()
x.plot.bar(figsize=(10,7))
```

Out[29]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x225dc979e48>
```



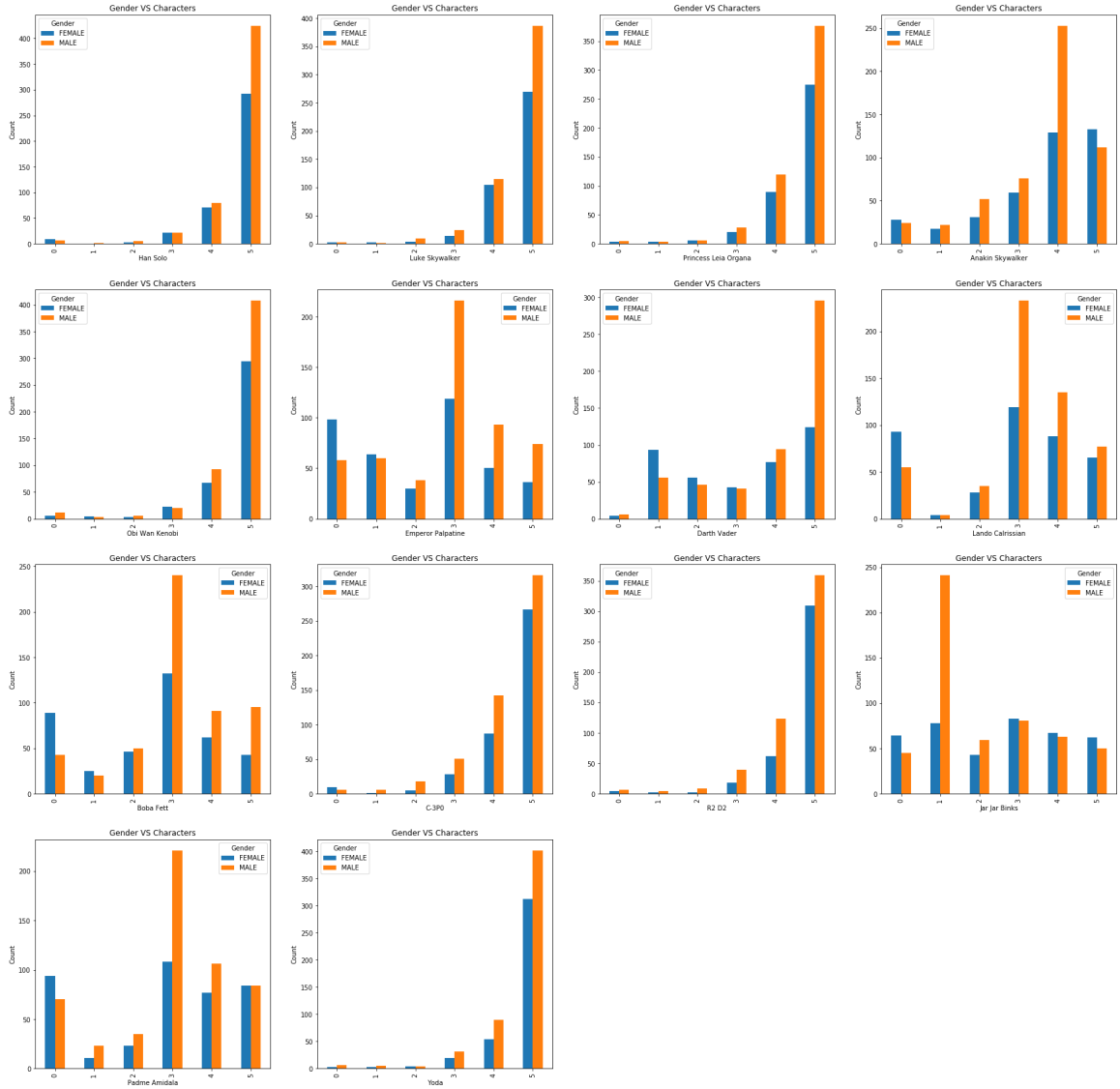
2.3 Explore a specific relationship

Here to find the relationship between **Gender** and all **Characters**, bar graph analysis is done as shown below.

- Most liked Characters are **`Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi .`**
- Most disliked character is **Jar Jar Blinks** and Males rated more than Females as least favourable character.

In [30]:

```
f = plt.figure(figsize=(30,30))
y=1

for col in df.columns[15:29]:
    ax = f.add_subplot(4,4,y)
    y=y+1
    x = df.groupby([col,'Gender'])[col].size().unstack()
    x.plot.bar(ax=ax)
    plt.title("Gender VS Characters")
    plt.ylabel("Count")
```



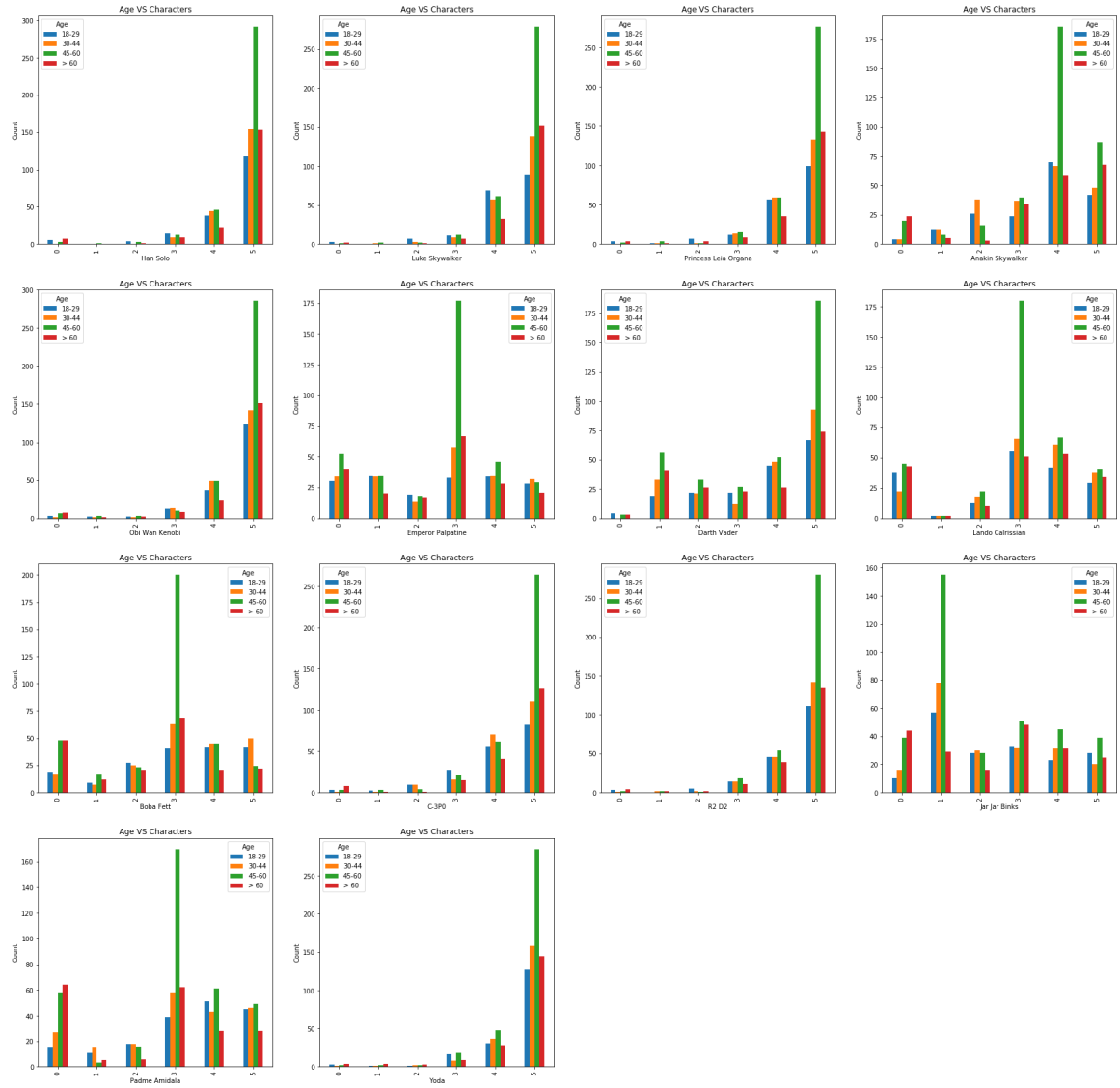
Explore a specific relationship

Here to find the relationship between **Age** and all **Characters**, bar graph analysis is done as shown below.

- Most liked Characters are `Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi` with age groups `45-60` .
- Most disliked character is `Jar Jar Blinks` and `45-60` years aged people rated more as least favourable character.

In [31]:

```
f = plt.figure(figsize=(30,30))
y=1

for col in df.columns[15:29]:
    ax = f.add_subplot(4,4,y)
    y=y+1
    x = df.groupby([col,'Age'])[col].size().unstack()
    x.plot.bar(ax=ax)
    plt.title("Age VS Characters")
    plt.ylabel("Count")
```



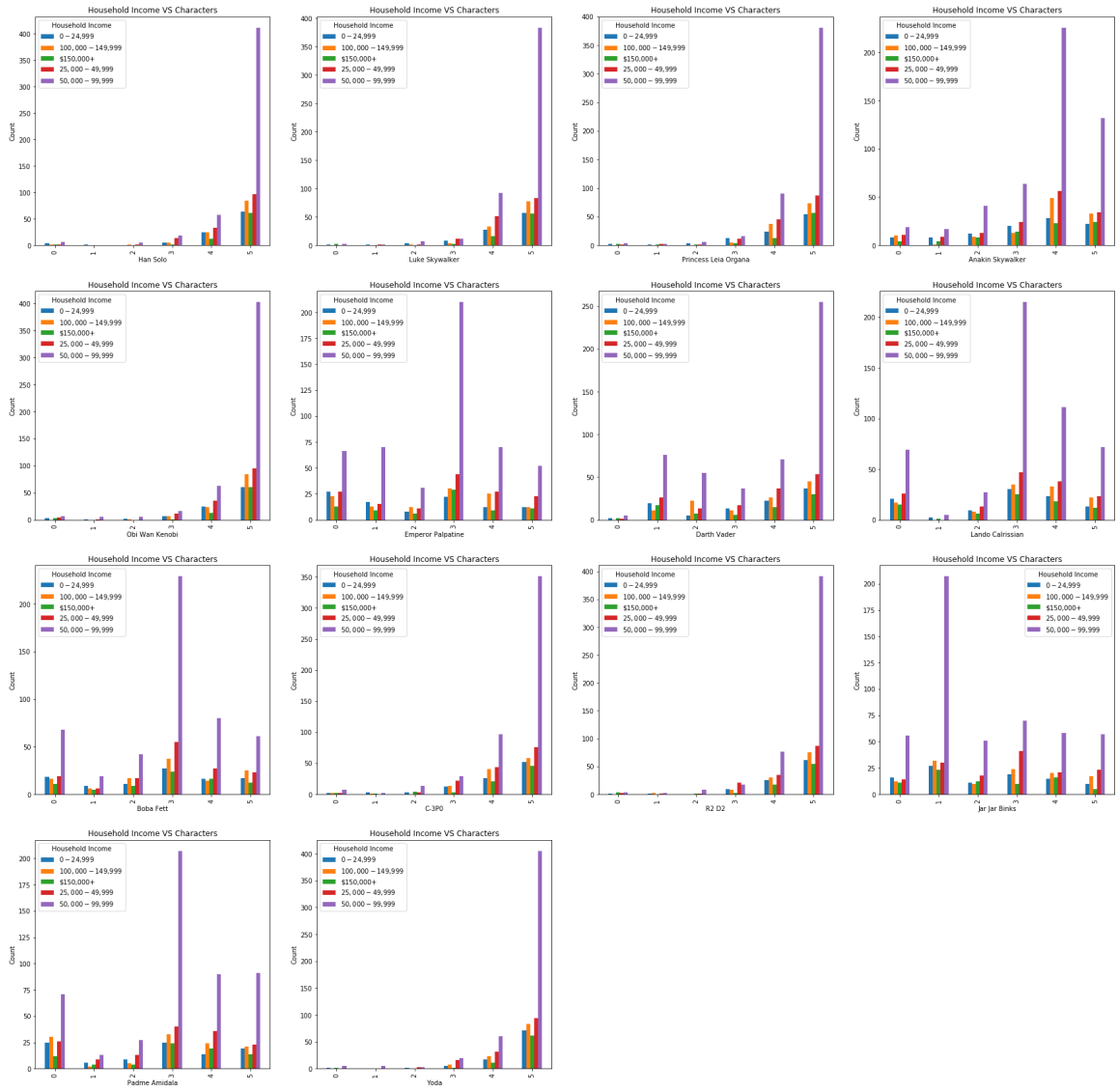
Explore a specific relationship

Here to find the relationship between **Household Income** and all **Characters**, bar graph analysis is done as shown below.

- Most liked Characters are **Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi** with household income `50,000-99,999` .
- Most disliked character is Jar Jar Blinks and `50,000-99,999` household income group rated more as least favourable character.

In [32]:

```
f = plt.figure(figsize=(30,30))
y=1

for col in df.columns[15:29]:
    ax = f.add_subplot(4,4,y)
    y=y+1
    x = df.groupby([col,'Household Income'])[col].size().unstack()
    x.plot.bar(ax=ax)
    plt.title("Household Income VS Characters")
    plt.ylabel("Count")
```



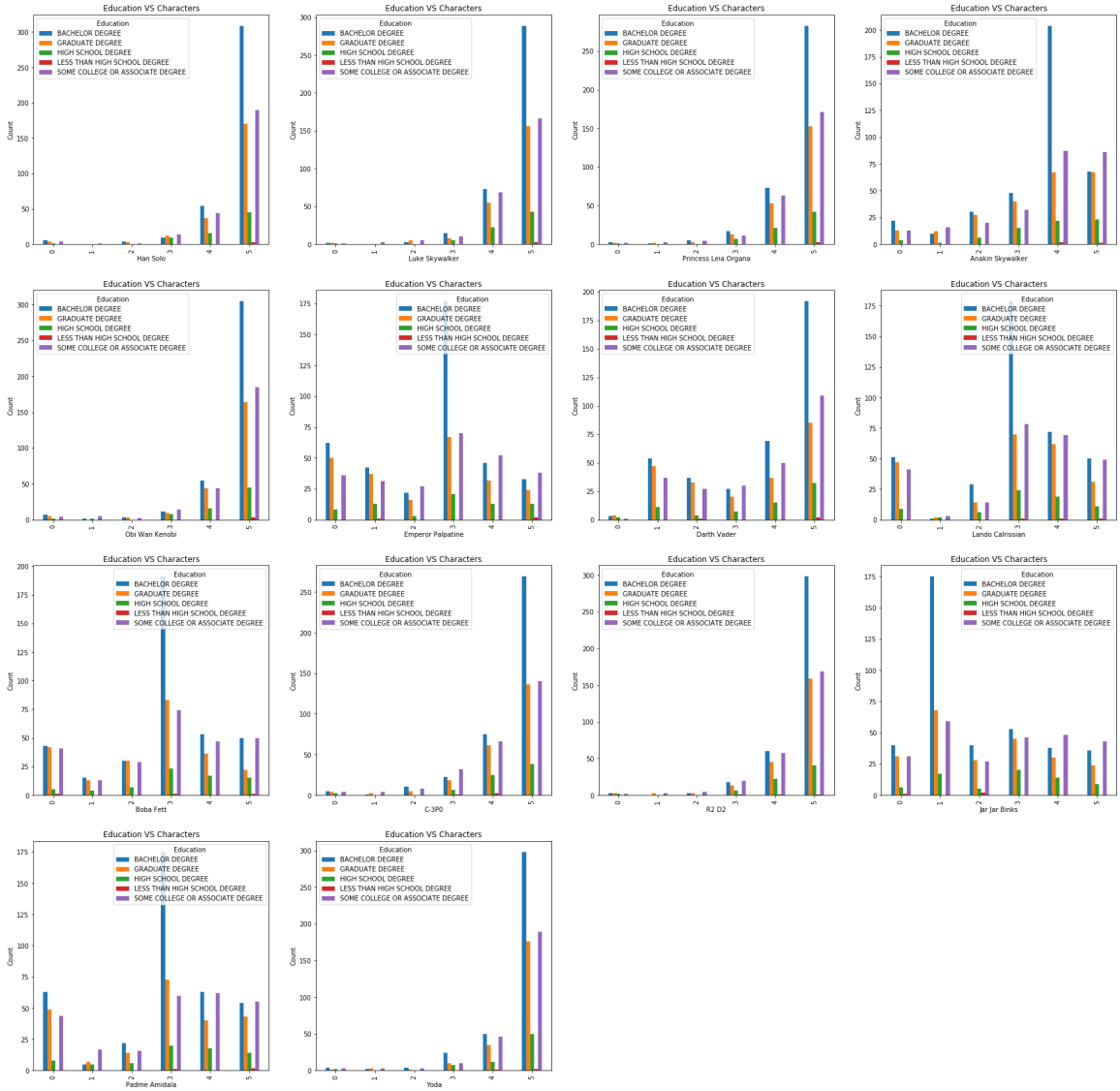
Explore a specific relationship

Here to find the relationship between **Education and all Characters**, bar graph analysis is done as shown below.

- Most liked Character is Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi with Education background `Bachelor Degree` .
- Most disliked character is Jar Jar Blinks and `Bachelor Degree` group rated most, as least favourable character.

In [33]:

```
f = plt.figure(figsize=(30,30))
y=1

for col in df.columns[15:29]:
    ax = f.add_subplot(4,4,y)
    y=y+1
    x = df.groupby([col,'Education'])[col].size().unstack()
    x.plot.bar(ax=ax)
    plt.title("Education VS Characters")
    plt.ylabel("Count")
```



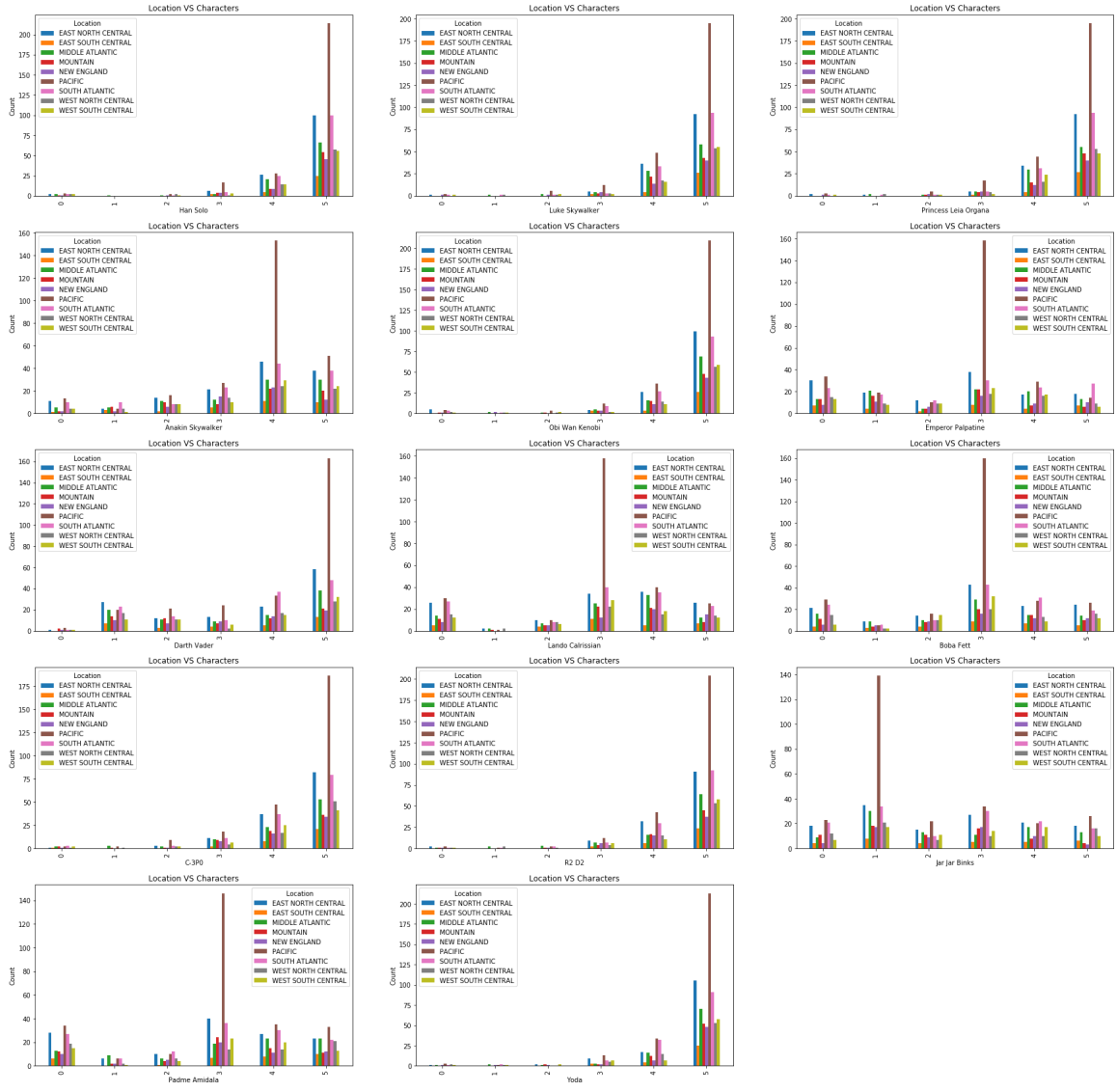
Explore a specific relationship

Here to find the relationship between **Location and all Characters**, bar graph analysis is done as shown below.

- Most liked Character is Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi with Location `PACIFIC` .
- Most disliked character is Jar Jar Blinks and `PACIFIC` location rated most, as least favourable character.

In [34]:

```
f = plt.figure(figsize=(30,30))
y=1

for col in df.columns[15:29]:
    ax = f.add_subplot(5,3,y)
    y=y+1
    x = df.groupby([col,'Location'])[col].size().unstack()
    x.plot.bar(ax=ax)
    plt.title("Location VS Characters")
    plt.ylabel("Count")
```



To `summarise` the entire analysis for characters we found that most liked characters are **Han Solo,Luke Skywalker,Yoda,Obi wan Kenobi** with people(more number of Males than Females) with the age group **45-60 years** old who have **Bachelor Degree** as Education and people from the location **PACIFIC** region rated the above mentioned characters as Highly Favourable and same section of groups rated **Jar Jar Blinks** as least favourable character in the survey. From these analysis we found the type of people who would probably like the franchise for making the movies.