Star Wars Survey - Analysis of Target Market Demographic

This document is designed to communicate the body of a data study which utilizes Python code and other related modules with team members that may not be fluent in such syntax. All code below can be executed right here in the notebook, and the resulting output will display on screen. These markdown sections will help describe what is happening in the analysis.

This study utilizes a survey (from five-thirty-eight) assessing data on fans of the Star Wars franchise. While another study will utilize this data to assess both popular and unpopular concepts and characters for marketing purposes, this study is specifically a statistical analysis of the most favorable demographic to target in Star Wars marketing campaigns.

To run this code, the following modules are used: NumPy, Pandas, Matplotlib, Scipy.stats, and finally a few stats functions that I wrote myself. Be sure to run the code line below before looking at any other cells!

In [1]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from scipy.stats import chi2_contingency, norm, chi2, chisquare, ttest
_ind
import stat_functions as st
```

Below is the actual csv data collected from the survey. We will here both load the csv as a dataframe object, and display some relevant information to help get us acclimated with the data.

In [2]:
```
df = pd.read_csv('StarWars.csv', encoding = "ISO-8859-15")
print(df.head())
```

```
   RespondentID Have you seen any of the 6 films in the Star Wars fr
anchise?  \
0           NaN                                           Response
1  3.292880e+09                                                Yes
2  3.292880e+09                                                 No
3  3.292765e+09                                                Yes
4  3.292763e+09                                                Yes

  Do you consider yourself to be a fan of the Star Wars film franchi
se?  \
0                                           Response
1                                                Yes
2                                                NaN
3                                                 No
```

```
4                                                Yes

  Which of the following Star Wars films have you seen? Please selec
t all that apply.  \
0           Star Wars: Episode I  The Phantom Menace
1           Star Wars: Episode I  The Phantom Menace
2                                                NaN
3           Star Wars: Episode I  The Phantom Menace
4           Star Wars: Episode I  The Phantom Menace

                                    Unnamed: 4  \
0  Star Wars: Episode II  Attack of the Clones
1  Star Wars: Episode II  Attack of the Clones
2                                          NaN
3  Star Wars: Episode II  Attack of the Clones
4  Star Wars: Episode II  Attack of the Clones

                                    Unnamed: 5  \
0  Star Wars: Episode III  Revenge of the Sith
1  Star Wars: Episode III  Revenge of the Sith
2                                          NaN
3  Star Wars: Episode III  Revenge of the Sith
4  Star Wars: Episode III  Revenge of the Sith

                          Unnamed: 6  \
0  Star Wars: Episode IV  A New Hope
1  Star Wars: Episode IV  A New Hope
2                                NaN
3                                NaN
4  Star Wars: Episode IV  A New Hope

                                     Unnamed: 7  \
0  Star Wars: Episode V The Empire Strikes Back
1  Star Wars: Episode V The Empire Strikes Back
2                                           NaN
3                                           NaN
4  Star Wars: Episode V The Empire Strikes Back

                                 Unnamed: 8  \
0  Star Wars: Episode VI Return of the Jedi
1  Star Wars: Episode VI Return of the Jedi
2                                       NaN
3                                       NaN
4  Star Wars: Episode VI Return of the Jedi

  Please rank the Star Wars films in order of preference with 1 bein
g your favorite film in the franchise and 6 being your least favorit
e film.  \
0           Star Wars: Episode I  The Phantom Menace
1                                                  3
2                                                NaN
3                                                  1
4                                                  5
```

```
   ...       Unnamed: 28       Which character shot first?  \
0  ...              Yoda                          Response
1  ...    Very favorably  I don't understand this question
2  ...               NaN                               NaN
3  ...  Unfamiliar (N/A)  I don't understand this question
4  ...    Very favorably  I don't understand this question

  Are you familiar with the Expanded Universe?  \
0                                     Response
1                                          Yes
2                                          NaN
3                                           No
4                                           No

  Do you consider yourself to be a fan of the Expanded Universe?æ  \
0                                           Response
1                                                 No
2                                                NaN
3                                                NaN
4                                                NaN

  Do you consider yourself to be a fan of the Star Trek franchise?
Gender  \
0                                           Response
Response
1                                                 No
Male
2                                                Yes
Male
3                                                 No
Male
4                                                Yes
Male

        Age     Household Income                         Education
\
0  Response             Response                          Response
1     18-29                  NaN                High school degree
2     18-29         $0 - $24,999                   Bachelor degree
3     18-29         $0 - $24,999                High school degree
4     18-29  $100,000 - $149,999  Some college or Associate degree

  Location (Census Region)
0                 Response
1           South Atlantic
2       West South Central
3       West North Central
4       West North Central

[5 rows x 38 columns]
```

Below are a list of all of the questions asked in the survey (some are not assigned a name value)

```
In [3]: print(df.columns)
```

```
Index(['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 f
ranchise?',
       'Which of the following Star Wars films have you seen? Please
select all that apply.',
       'Unnamed: 4', 'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unna
med: 8',
       'Please rank the Star Wars films in order of preference with
1 being your favorite film in the franchise and 6 being your least f
avorite film.',
       'Unnamed: 10', 'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13',
       'Unnamed: 14',
       'Please state whether you view the following characters favor
ably, unfavorably, or are unfamiliar with him/her.',
       'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18', 'Unnamed: 19',
       'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22', 'Unnamed: 23',
       'Unnamed: 24', 'Unnamed: 25', 'Unnamed: 26', 'Unnamed: 27',
       'Unnamed: 28', 'Which character shot first?',
       'Are you familiar with the Expanded Universe?',
       'Do you consider yourself to be a fan of the Expanded Univers
e?æ',
       'Do you consider yourself to be a fan of the Star Trek franch
ise?',
       'Gender', 'Age', 'Household Income', 'Education',
       'Location (Census Region)'],
      dtype='object')
```

As can be seen above, each column in the survey dataframe represents relevant questions to identify potential fans and their demograpic data. However, it is clear that some cleaning is required. That said, there is enough complete data here to identify key demographic information.

```
In [4]: print(df['Do you consider yourself to be a fan of the Star Wars film f
ranchise?'].unique())
```

```
['Response' 'Yes' nan 'No']
```

The first major issue here is how this important survey question was handled. This was the second question asked in the survey. The first question was simply asking if someone had even seen a Star Wars movie. If they said no, this question was left blank. For the purposes of our market research, however, someone who has never seen the movies is not considered a fan, so instead of nan, this should be a no (someone who has never seen a movie is not a fan). Here we will include a new column, 'is_fan', which will input 'Yes' if the survey response was 'Yes', or else simply 'No'. This will give us a better understanding of who is a fan, and who is either disinterested or actively dislikes the Star Wars franchise.

Note: The answer 'Response' is irrelevant based on how the header in the dataframe was created. It is preserved for now to keep the data consistent.

In [5]:

```
df['is_fan'] = df.apply(lambda row: 'Yes' if row['Do you consider your
self to be a fan of the Star Wars film franchise?'] == 'Yes'\
                        else ('Response' if row['Do you consider yours
elf to be a fan of the Star Wars film franchise?'] == 'Response'\
                             else 'No'), axis = 1)
```

Analysis 1: Star Wars Fans by Gender

First, we will group the number of respondents into their respective gender responses, as well as whether they are fans or not. Note that this is the total number of respondents to the survey. Included as well is a briefly calculated ratio to demonstrate the percentage of each gender that are fans.

In [6]:

```
gender_sw_fans = df.groupby(['is_fan','Gender']).RespondentID.count().
reset_index()
gender_sw_fans_pivot = gender_sw_fans.pivot(columns= 'is_fan', index =
'Gender', values = 'RespondentID')
gender_sw_fans_pivot.drop('Response', axis = 1, inplace = True)
gender_sw_fans_pivot.drop('Response', axis = 0, inplace = True)
gender_sw_fans_pivot['Ratio_of_fans'] = gender_sw_fans_pivot.apply(lam
bda row: (row.Yes / (row.Yes + row.No))\
                                                                   * 1
00, axis = 1)
print(gender_sw_fans_pivot)
```

```
is_fan      No    Yes  Ratio_of_fans
Gender
Female   311.0  238.0      43.351548
Male     194.0  303.0      60.965795
```

It appears from a cursory glance that the male population is far more favorable to Star Wars than the population of women. However, let's first test this statistically against the null hypothesis that this is merely a random fluctuation of data. We will employ a chi squared test on this data.

In [7]:

```
con_list = gender_sw_fans_pivot.loc[['Female', 'Male'],['Yes', 'No']].
values.tolist()
print(con_list)
chi2, pval = st.chi2_homogeneity(con_list)
print(pval)
```

```
[[238.0, 311.0], [303.0, 194.0]]
1.2474230204340131e-08
```

As the pvalue is less that 0.05, we reject the null hypothesis and assume statistical significance. Therefore, we have evidence that the distribution of fans in Star Wars favors men more than women.

The sample seems to indicate that this difference is over 17%, which is substantial. Here we will calculate a confidence interval for the difference of proportions in these samples, to get an idea at 95% confidence of the true difference between men and women in terms of being Star Wars fans:

In [8]:

```
sums = [np.sum(i) for i in con_list]

ratios = gender_sw_fans_pivot['Ratio_of_fans'].values.tolist()
good_ratios = [i/100 for i in ratios]

print(st.diff_z_intv(good_ratios[1], sums[1], good_ratios[0], sums[0])
)
```

```
(0.11649559670562445, 0.23578933327498777)
```

It appears that, from the information we have gathered, there is good evidence to assume that the difference between men and women in terms of proportion of fans in their populations is 11%-23%.

Breakdown of fans by Age and Gender

The code below creates a count of star wars fans as they fit within age groups. They are also separated into men and women so that we can compare these categories clearly. The ratio fan/not fan as well as the total number of yes and no responses given per age group and gender are displayed below:

```
In [9]: gender_age_fan_sw = df.groupby(['is_fan','Gender', 'Age']).RespondentI
D.count().reset_index().sort_values(by = ['Gender','Age'])

gender_age_fan_sw_pivot = gender_age_fan_sw.pivot_table(columns='is_fa
n',index=['Gender', 'Age'], values = 'RespondentID')
gender_age_fan_sw_pivot.drop('Response', axis = 1, inplace = True)
gender_age_fan_sw_pivot.drop('Response', axis = 0, inplace = True)

gender_age_fan_sw_pivot['Ratio of fans'] = gender_age_fan_sw_pivot.app
ly(lambda row:(row.Yes / (row.No + row.Yes)) * 100, axis = 1)
print(gender_age_fan_sw_pivot)
```

```
is_fan           No   Yes  Ratio of fans
Gender Age
Female 18-29   64.0  50.0      43.859649
       30-44   77.0  59.0      43.382353
       45-60   77.0  74.0      49.006623
       > 60    93.0  55.0      37.162162
Male   18-29   30.0  74.0      71.153846
       30-44   41.0  91.0      68.939394
       45-60   60.0  80.0      57.142857
       > 60    63.0  58.0      47.933884
```

Cursorily, we can identify the highest number and ratio of favorable response among men in the 18-29 and 30-44 year old age ranges. We will now test statistical significance of the age group differences among men in these age groups:

```
In [10]: gender_age_no = gender_age_fan_sw_pivot['No'].values.tolist()
male_no = gender_age_no[4:]
female_no = gender_age_no[:4]
gender_age_yes = gender_age_fan_sw_pivot['Yes'].values.tolist()
male_yes = gender_age_yes[4:]
female_yes = gender_age_yes[:4]
contingency_male_lists = [[a,b] for a,b in zip(male_yes, male_no)]
print(contingency_male_lists)
chi2, pval, dof, expected = chi2_contingency(contingency_male_lists)
print(pval)
```

```
[[74.0, 30.0], [91.0, 41.0], [80.0, 60.0], [58.0, 63.0]]
0.0005426107198083925
```

The probability of the null hypothesis - that there is no significant difference in age parameters of male Star Wars fans - is about 0.05%, which is therefore rejected. We therefore find evidence for statistical significance. The ratio of fans in the 18-29 and 30-44 age groups are very close, however, and it is difficult to tell if there's any reason to assume from this data that 18-29 year olds react any more postively than 30-44 year olds. We will test this specific group now:

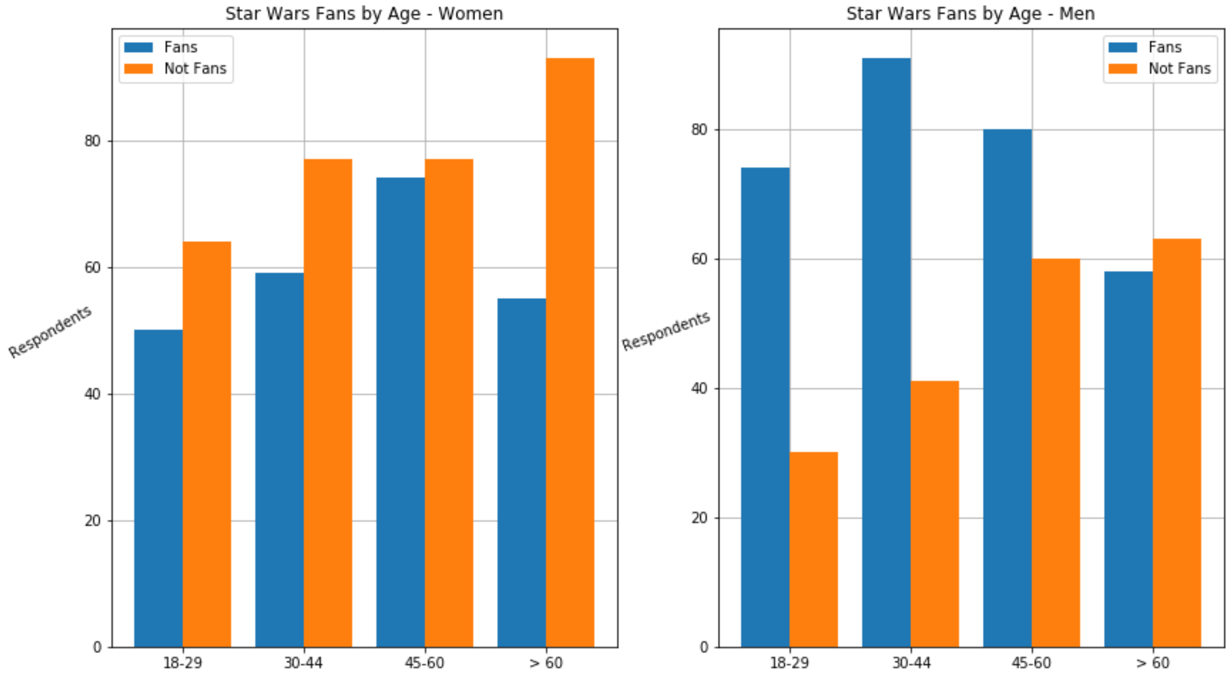
```
In [11]: zstat, pval = st.two_samp_z_test(contingency_male_lists[0], contingency_male_lists[1])
         print(pval)
```

```
0.3563396071939904
```

The probability of this null hypothesis is much higher, and we fail to reject. It is reasonably likely that there is no significant difference in the proportion of fans in both age groups, and that therefore the age range of 18-44 has a high proprotion of male fans.

Graphical representation of Age and Gender breakdown

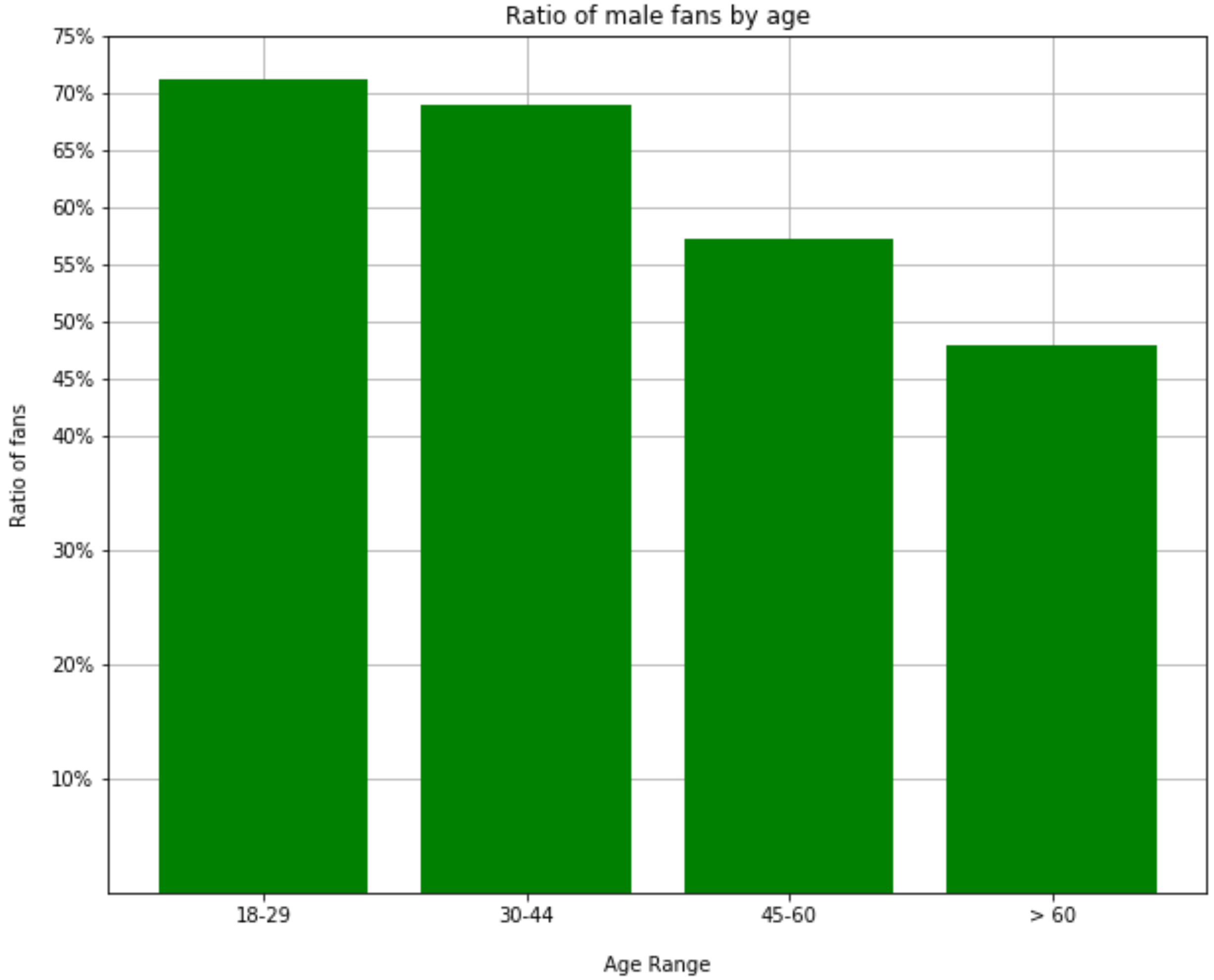
```
In [12]: age_labels = ['18-29','30-44','45-60','> 60']
         plt.figure(figsize = (14,8))
         ax = plt.subplot(1,2,1)
         ax.grid(zorder = 0)
         yesx = [2 * el + (0.8 *1) for el in range(4)]
         nox = [2 * el + (0.8 * 2) for el in range(4)]
         middle = [(a + b)/2 for a,b in zip(yesx, nox)]
         plt.bar(yesx, female_yes, zorder = 3)
         plt.bar(nox, female_no, zorder =3)
         ax.set_xticks(middle)
         ax.set_xticklabels(age_labels)
         plt.ylabel('Respondents', rotation = 30, labelpad = 20)
         plt.title('Star Wars Fans by Age - Women')
         plt.legend(['Fans', 'Not Fans'])
         ax1 = plt.subplot(1,2,2)
         ax1.grid(zorder = 0)
         plt.bar(yesx, male_yes, zorder = 3)
         plt.bar(nox, male_no, zorder = 3)
         ax1.set_xticks(middle)
         ax1.set_xticklabels(age_labels)
         plt.ylabel('Respondents', rotation = 20, labelpad = 15)
         plt.title('Star Wars Fans by Age - Men')
         plt.legend(['Fans', 'Not Fans'])
         plt.savefig('Male_female_age_breakdown.png')
         plt.show()
```



Graph representation of Male Fans by Ratio

In [13]:

```
ratio_fans = gender_age_fan_sw_pivot['Ratio of fans'].values.tolist()
male_fan_ratio= ratio_fans[4:]
plt.figure(figsize = (10,8))
ax = plt.subplot()
ax.grid(zorder = 0)
plt.bar(range(len(male_fan_ratio)), male_fan_ratio, color = 'green', z
order = 3)
ax.set_yticks([10, 20, 30, 40, 45, 50, 55, 60, 65, 70, 75])
ax.set_yticklabels(['10%','20%','30%','40%','45%','50%','55%','60%','6
5%','70%','75%'])
plt.ylabel('Ratio of fans', labelpad = 10)
ax.set_xticks(range(len(age_labels)))
ax.set_xticklabels(age_labels)
plt.xlabel('Age Range', labelpad = 15)
plt.title('Ratio of male fans by age')
plt.savefig('Ratio_of_male_fans.png')
plt.show()
```



Breakdown of fans by Income levels

An analysis of household income ranging from under $25,000 to over \$150,000 to see if a greater quantity of fans exist in any particular income bracket.

In [14]:
```
fans = df[df.is_fan == 'Yes']
income_breakdown = fans.groupby('Household Income').RespondentID.count
().reset_index().sort_values(by = 'Household Income')
income_breakdown_clean = income_breakdown[income_breakdown['Household
Income'] != 'Response']
new_index = ['$0 - $24,999', '$25,000 - $49,999','$50,000 - $99,999 ',
'$100,000 - $149,999','$150,000+']
ordered_income_breakdown = income_breakdown_clean.reindex([0, 3, 4, 1,
2]).reset_index()
income_nums = ordered_income_breakdown.RespondentID.values.tolist()
total =sum(income_nums)
ordered_income_breakdown['Proportion of Respondents'] = ordered_income
_breakdown.RespondentID.apply(lambda x: (x / total) * 100)
print(ordered_income_breakdown)
```

```
   index     Household Income  RespondentID  Proportion of Responden
ts
0      0         $0 - $24,999            61                  14.0229
89
1      3    $25,000 - $49,999            84                  19.3103
45
2      4    $50,000 - $99,999           156                  35.8620
69
3      1  $100,000 - $149,999            79                  18.1609
20
4      2             $150,000+            55                  12.6436
78
```

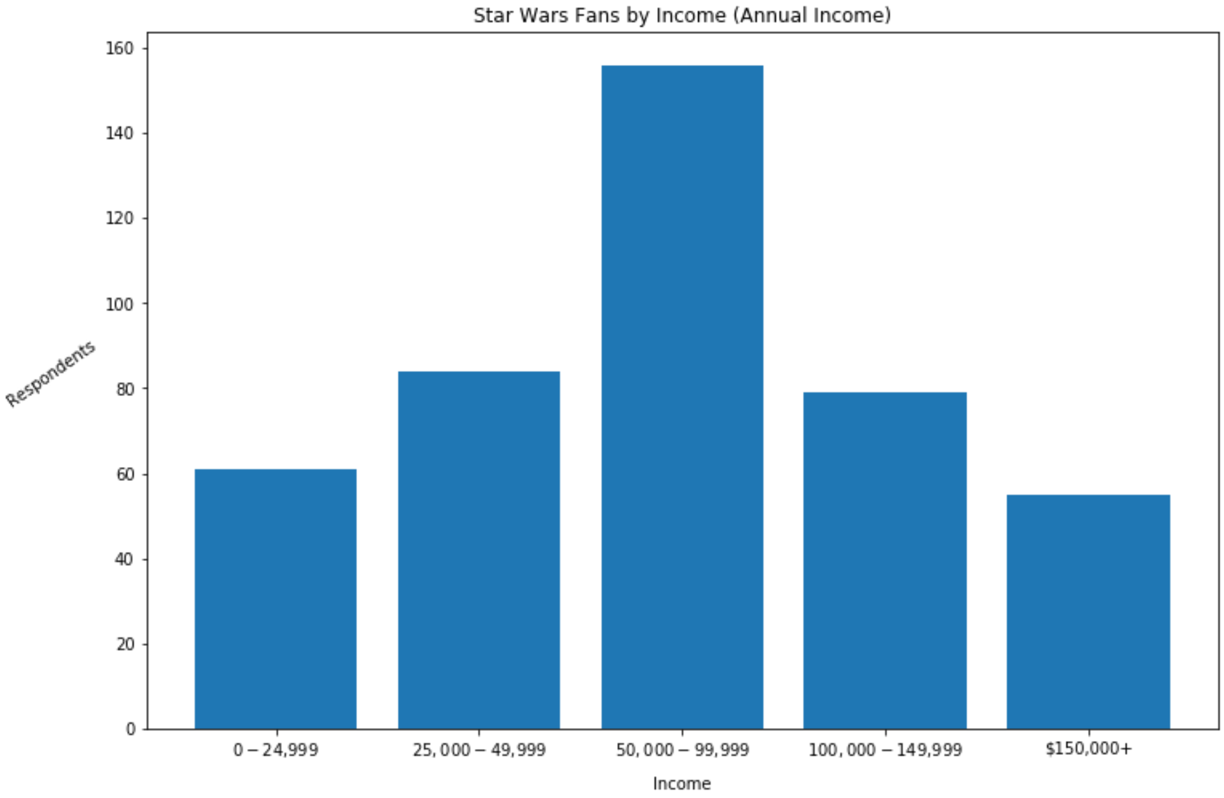
In [15]:
```
distribution = ordered_income_breakdown.RespondentID.values.tolist()
chi_2, pval = chisquare(distribution)
print(pval)
```

```
1.894685914951163e-15
```

It appears that the distribution of fans by income is not uniform, and that it follows an approximately normal distribution, with the majority of fans (about 35% of the total) in the $50,000- \$99,000 range. This is expected given the age range of 18-44, which also generally falls in this distribution of income as a whole.

Graphical Representation of Income Breakdown

```
In [16]: income_nums = ordered_income_breakdown.RespondentID.values.tolist()
         plt.figure(figsize = (12,8))
         ax = plt.subplot()
         plt.bar(range(len(income_nums)), income_nums)
         labels = ordered_income_breakdown['Household Income'].values.tolist()
         ax.set_xticks(range(len(labels)))
         ax.set_xticklabels(labels)
         plt.ylabel('Respondents', rotation = 35, labelpad = 30)
         plt.xlabel('Income', labelpad = 10)
         plt.title('Star Wars Fans by Income (Annual Income)')
         plt.show()
```



Breakdown of Fans by Education Level

In [17]:

```
education = fans.groupby('Education').RespondentID.count().reset_index
()
education_clean = education[education.Education != 'Response']

education_ordered = education_clean.reindex([3, 2, 0, 1]).reset_index(
)

total = sum(education_ordered.RespondentID.values)
education_ordered['Proportion of Respondents'] = education_ordered.Res
pondentID.apply(lambda x: (x/total) * 100)
print(education_ordered)
```

```
   index                   Education  RespondentID  \
0      3  Less than high school degree             3   
1      2          High school degree            41   
2      0             Bachelor degree           172   
3      1             Graduate degree           152   

   Proportion of Respondents  
0                   0.815217  
1                  11.141304  
2                  46.739130  
3                  41.304348  
```

The above data illustrates the number of fans as they are arranged by the highest level education degree they have earned. This is relatively unsurprising, given the age and income levels of the population sample surveyed. The largest concentration of fans have earned either a Bachelor's Degree or a Graduate Degree.

The following data will illustrate how only the men in this survey are broken down by education level:

In [18]:

```
men_fans = fans[fans.Gender == 'Male']
education_men = men_fans.groupby('Education').RespondentID.count().reset_index()
education_men_clean = education_men[education_men.Education != 'Response']
ordered_men = education_men_clean.reindex([3, 2, 0, 1]).reset_index()
total_1 = sum(ordered_men.RespondentID.values)
ordered_men['Proportion of Respondents'] = ordered_men.RespondentID.apply(lambda x: (x/total_1) * 100)
print(ordered_men)
```

```
   index                   Education  RespondentID  \
0      3  Less than high school degree             2   
1      2           High school degree            30   
2      0              Bachelor degree            90   
3      1              Graduate degree            80   

   Proportion of Respondents  
0                   0.990099  
1                  14.851485  
2                  44.554455  
3                  39.603960  
```

The proportion of fans appears to be very similar, indicating that there is no statistically different education level for men from the entire population. The total number of fans in both the whole sample and the male only education breakdown will be now compared with a chi squared homogeneity test:

In [19]:

```
pop_lists = [education_ordered['RespondentID'].values.tolist(), ordered_men['RespondentID'].values.tolist()]
print(pop_lists)
chi_2, pval = st.chi2_homogeneity(pop_lists)
print(pval)
```

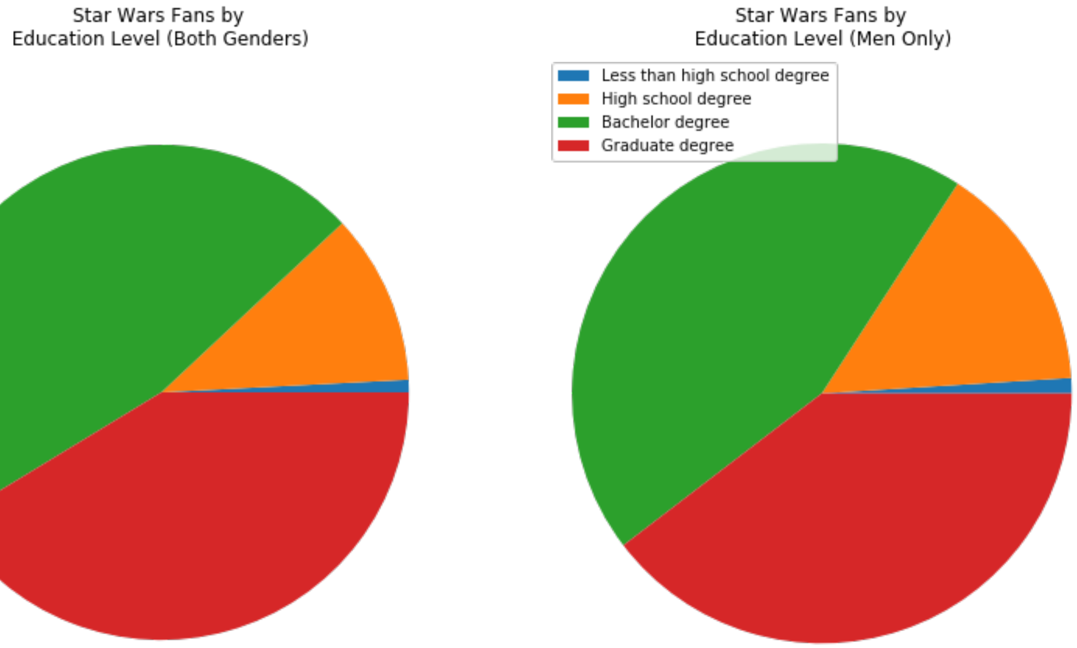
```
[[3, 41, 172, 152], [2, 30, 90, 80]]
0.6336595585471168
```

Given the probability of the null hypothesis being over 5%, we fail to reject. It appears that the distribution of male fans by education level are not significanlty different from the population distribution by education level in terms of Star Wars fans.

Graph of Education Breakdown - Whole Population and Only Male

In [20]:

```
plt.figure(figsize = (14,8))
ax = plt.subplot(1,2,1)
plt.pie(education_ordered.RespondentID.values)
plt.axis('equal')
plt.title('Star Wars Fans by\n Education Level (Both Genders)')
ax_1 = plt.subplot(1,2,2)
plt.pie(ordered_men.RespondentID.values)
plt.axis('equal')
plt.title('Star Wars Fans by\n Education Level (Men Only)')
plt.legend(ordered_men.Education.values, loc = 2)
plt.savefig('SWfans_education.png')
plt.show()
```



As can be shown, there is no significant apparent difference, and the vast majority of fans have a Bachelor's and/or Graduate Degree.

Breakdown of Fans by Region

Can it be shown that any particular region of the US has a higher concentration of fans? Here are the results as shown in the survey:

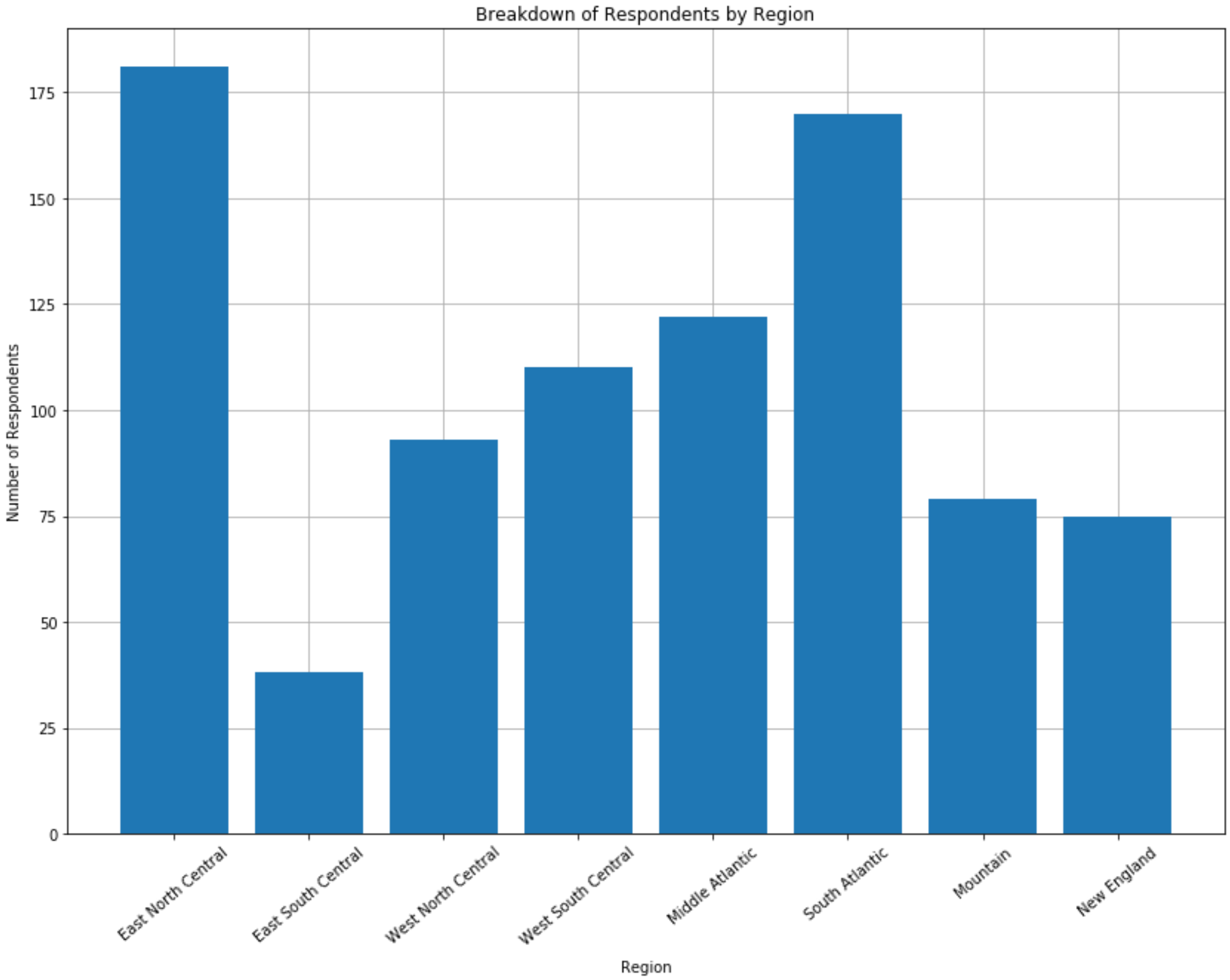
In [21]:

```
region = df.groupby('Location (Census Region)').RespondentID.count().r
eset_index()
region_clean = region[region['Location (Census Region)'] != 'Response'
]
ordered_region = region_clean.reindex([0,1,8,9,2,7,3,4]).reset_index()
print(ordered_region)
```

```
   index Location (Census Region)  RespondentID
0      0       East North Central           181
1      1       East South Central            38
2      8       West North Central            93
3      9       West South Central           110
4      2          Middle Atlantic           122
5      7           South Atlantic           170
6      3                 Mountain            79
7      4              New England            75
```

In [22]:

```
y = ordered_region.RespondentID.values.tolist()
x = range(len(y))
labels = ordered_region['Location (Census Region)'].values.tolist()
plt.figure(figsize = (14, 10))
ax = plt.subplot()
ax.grid(zorder = 0)
plt.bar(x, y, zorder = 3)
plt.xlabel('Region', labelpad = 8)
plt.ylabel('Number of Respondents')
ax.set_xticks(range(len(labels)))
ax.set_xticklabels(labels, rotation = 40)
plt.title('Breakdown of Respondents by Region')
plt.savefig('Respondents_by_region.png')
plt.show()
```



```
In [23]: regions = ordered_region['RespondentID'].values.tolist()
         print(regions)
         chi_2, pval = chisquare(regions)
         print(pval)
```

```
[181, 38, 93, 110, 122, 170, 79, 75]
2.0711017707262145e-29
```

Analyzing the distribution of fans in these regions from the samples provided, there seems to be evidence that Star Wars fans are not evenly distributed throughout these regions. The sample data suggests there to be a higher concentration of fans in the East North Central US region, as well as the South Atlantic, and the fewest fans in the East South Central region. All other regions are much closer to evenly distributed among each other.

Conclusion

Based off of the study conducted, the following demographic types are the most likely to positively respond to future marketing campaigns centered on the Star Wars franchise:

- Gender - Male
- Age range - 18-44
- Income level - $50,000 - \$99,000
- Education Level - Bachelor's Degree and higher
- Ideal Marketing Regions - East North Central, South Atlantic
- Least Ideal Marketing Region - East South Central

We can state with confidence, based on this study, that utilizing marketing strategies known to be effective with these demographic types, focused on the regions listed, are likely to be effective.

Recommendations following this survey study

This survey has suggested many potential demographics that will be more highly receptive to Star Wars marketing campaigns. However, as this is a preliminary observation, it would be most appropriate to formally test this data with an A/B test, comparing marketing strategies targeting these demographics with previous ad campaign strategies. A separate consultation would be helpful, to get information on recommending an appropriate sample size for each group. For example, if a marketing campaign has been shown to have 30% success, and there is interest in seeing no less an increase than to 35% success (about a 17% lift), we can expect results with 95% confidence with 840 samples per variation.