In [20]:

```
import pandas as pd
star_wars = pd.read_csv("star_wars.csv", encoding='ISO-8859-1')
```

In [22]:

```
star_wars = star_wars[pd.notnull(star_wars['RespondentID'])]
```

In [23]:

```
yes_no = {
    'Yes' : True,
    'No' : False
}

for col in [
    '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?'
    ]:
    star_wars[col] = star_wars[col].map(yes_no)
```

Out[23]:

	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?	Which of the following Star Wars films have you seen? Please select all that apply.	Unnamed: 4	Unnamed: 5	Unnamed: 6	U 7
1	3292879998	True	True	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	S E T E S E

2	3292879538	False	NaN	NaN	NaN	NaN	NaN	N
3	3292765271	True	False	Star Wars: Episode I The Phantom Menace	Star Wars: Episode II Attack of the Clones	Star Wars: Episode III Revenge of the Sith	NaN	N
4	3292763116	True	True	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	S E T E S E
5	3292731220	True	True	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	S E T E S E

5 rows × 38 columns

In [25]:

```
import numpy as np

true_false = {
    "Star Wars: Episode I  The Phantom Menace": True,
    np.nan: False,
    "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
}

for col in star_wars.columns[3:9]:
    star_wars[col] = star_wars[col].map(true_false)
```

In [27]:

```
star_wars = star_wars.rename(columns={
        "Which of the following Star Wars films have you seen? Please select all tha
        "Unnamed: 4": "seen_2",
```

```
        "Unnamed: 5": "seen_3",
        "Unnamed: 6": "seen_4",
        "Unnamed: 7": "seen_5",
        "Unnamed: 8": "seen_6"
        })
```

Out[27]:

	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?	seen_1	seen_2	seen_3	seen_4	seen_5	seen
1	3292879998	True	True	True	True	True	True	True	True
2	3292879538	False	NaN	False	False	False	False	False	False
3	3292765271	True	False	True	True	True	False	False	False
4	3292763116	True	True	True	True	True	True	True	True
5	3292731220	True	True	True	True	True	True	True	True

5 rows × 38 columns

5 rows × 38 columns

In [32]:

```
star_wars = star_wars.rename(columns={
        "Please rank the Star Wars films in order of preference with 1 being your fa
        "Unnamed: 10": "ranking_2",
        "Unnamed: 11": "ranking_3",
        "Unnamed: 12": "ranking_4",
        "Unnamed: 13": "ranking_5",
        "Unnamed: 14": "ranking_6",
        })

star_wars[star_wars.columns[9:15]] = star_wars[star_wars.columns[9:15]].astype(float
```

Out[32]:

	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?	seen_1	seen_2	seen_3	seen_4	seen_5	seen
1	3292879998	True	True	True	True	True	True	True	True
2	3292879538	False	NaN	False	False	False	False	False	False
3	3292765271	True	False	True	True	True	False	False	False
4	3292763116	True	True	True	True	True	True	True	True
5	3292731220	True	True	True	True	True	True	True	True

5 rows × 38 columns

To this point, the data was cleaned up by taking the following steps.

- Converted Yes/No to True/False in the question of whether the respondent saw one of the movies.
- Renamed columns for which movie they saw and the ranking columns to a more descriptive name.
- Removed records with no respondent ID.
- Ran an analysis to determine which movie ranked the highest.

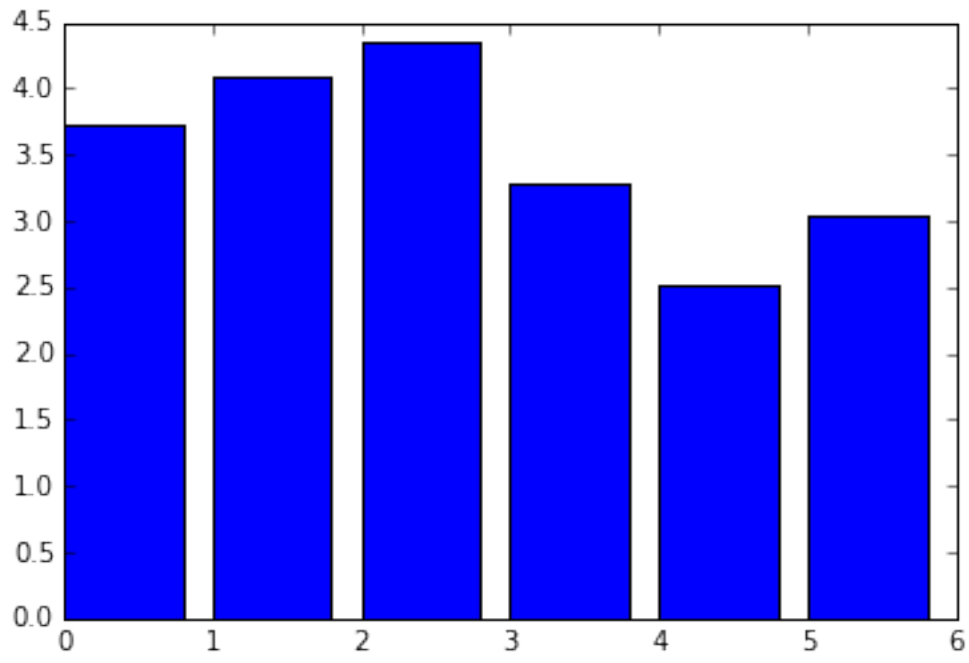
In [34]:

```
%matplotlib inline
import matplotlib.pyplot as plt

plt.bar(range(6),star_wars[star_wars.columns[9:15]].mean())
```

Out[34]:

```
<Container object of 6 artists>
```



In [35]:

```
star_wars[star_wars.columns[9:15]].mean()
```

Out[35]:

```
ranking_1    3.732934
ranking_2    4.087321
ranking_3    4.341317
ranking_4    3.272727
ranking_5    2.513158
ranking_6    3.047847
dtype: float64
```

The fifth movie, The Empire Strikes Back was the highest ranked movie. with 2.51 out six with 1 being the highest ranked.

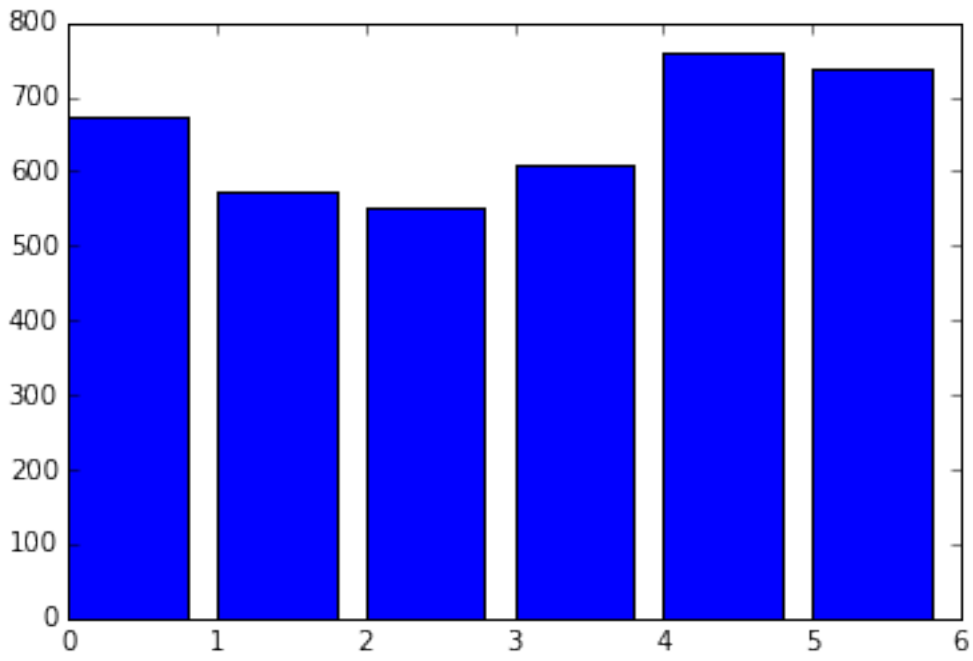
This movie contains multiple mini plots compared to the other movies. In addition, it is the first interaction between Darth Vader and Luke Skywalker which is a major plotline in the series.

In [36]:

```
plt.bar(range(6),star_wars[star_wars.columns[3:9]].sum())
star_wars[star_wars.columns[3:9]].sum()
```

Out[36]:

```
seen_1    673
seen_2    571
seen_3    550
seen_4    607
seen_5    758
seen_6    738
dtype: int64
```



The last two movies show to have been seen more than the first three. One reason for this could be that the last three were the originals and have been around for a longer time.

The first movie was the third highest and may be because it was

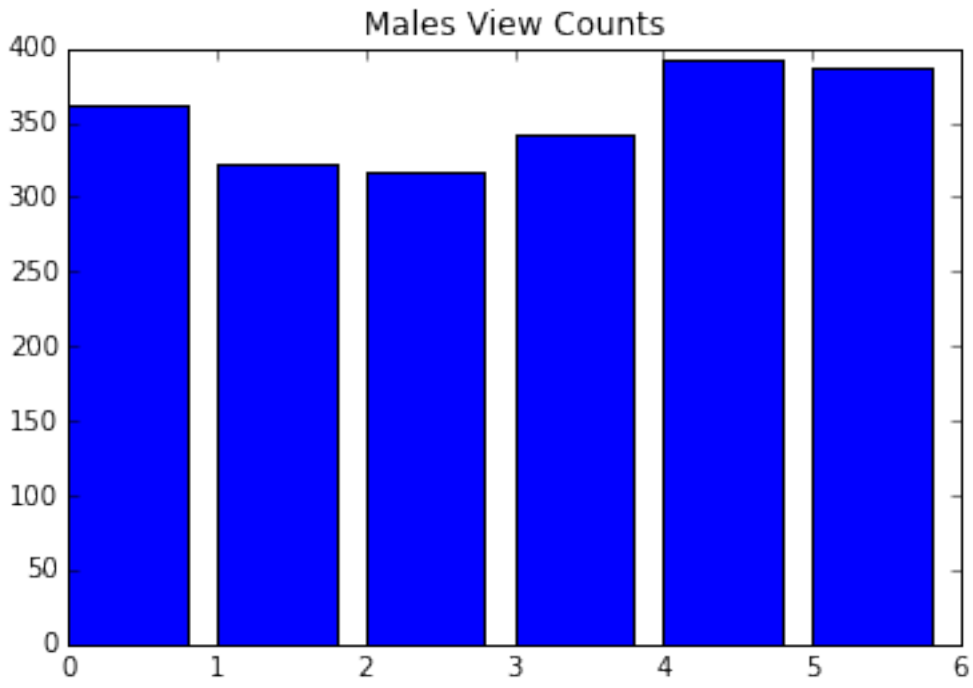
In [46]:

```
males = star_wars[star_wars['Gender'] == 'Male']
females = star_wars[star_wars['Gender'] == 'Female']

plt.bar(range(6),males[males.columns[3:9]].sum())
plt.title("Males View Counts")
males[males.columns[3:9]].sum()
```

Out[46]:

```
seen_1    361
seen_2    323
seen_3    317
seen_4    342
seen_5    392
seen_6    387
dtype: int64
```

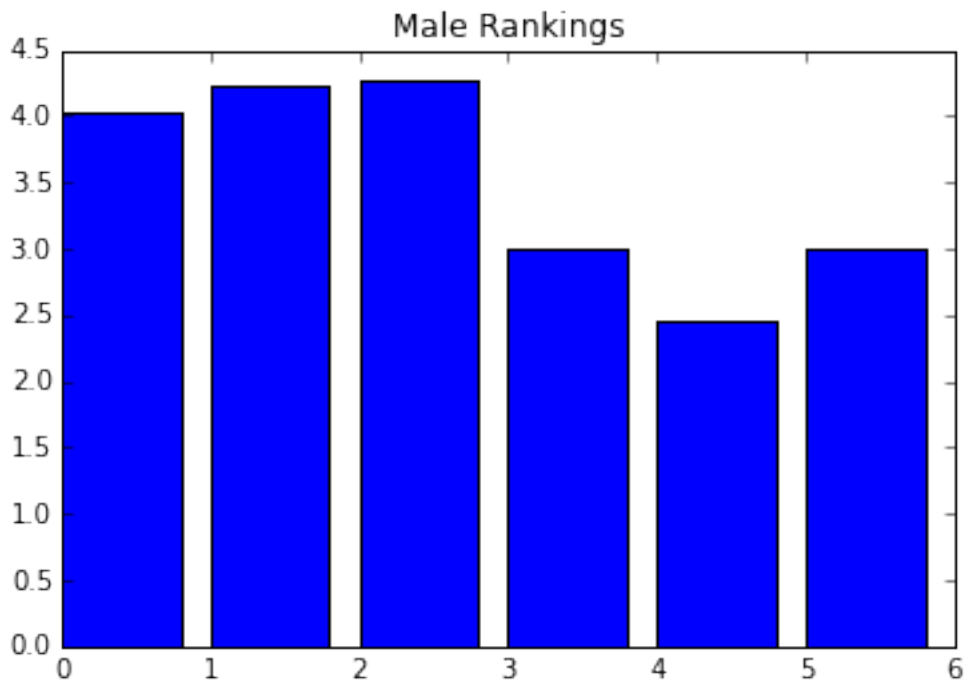


In [45]:

```
plt.bar(range(6),males[males.columns[9:15]].mean())
plt.title("Male Rankings")
males[males.columns[9:15]].mean()
```

Out[45]:

```
ranking_1    4.037825
ranking_2    4.224586
ranking_3    4.274882
ranking_4    2.997636
ranking_5    2.458629
ranking_6    3.002364
dtype: float64
```

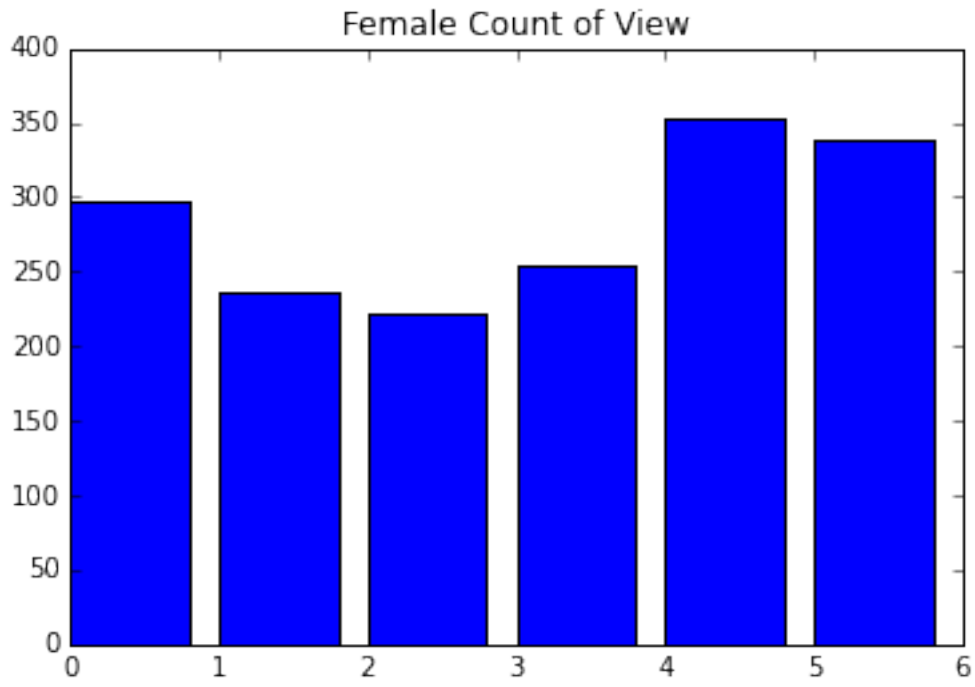


In [44]:

```
plt.bar(range(6),females[females.columns[3:9]].sum())
plt.title("Female Count of View")
females[females.columns[3:9]].sum()
```

Out[44]:

```
seen_1    298
seen_2    237
seen_3    222
seen_4    255
seen_5    353
seen_6    338
dtype: int64
```

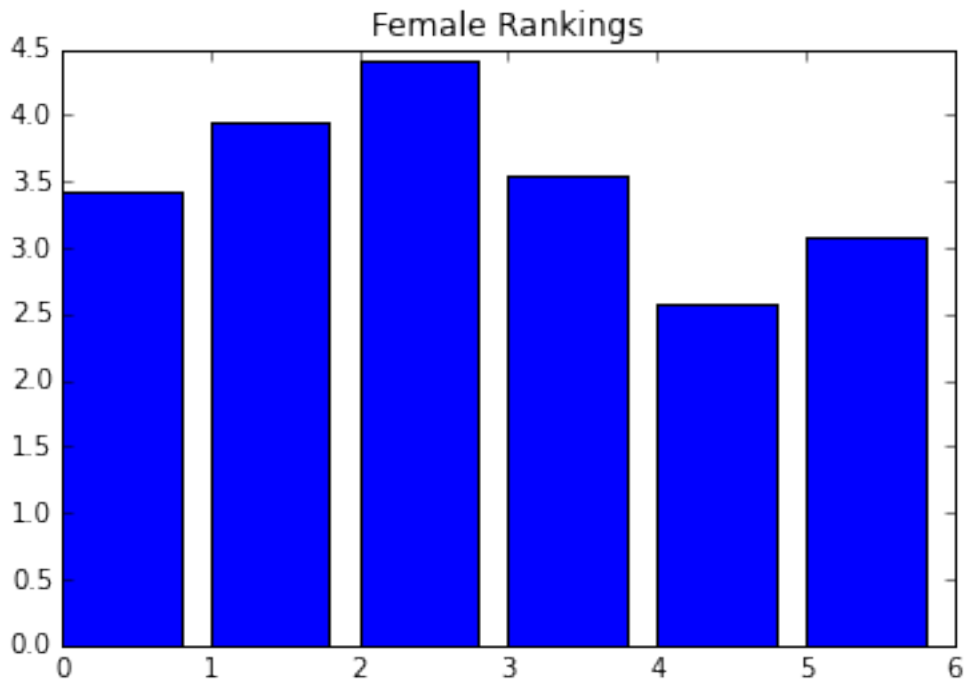


In [43]:

```
plt.bar(range(6),females[females.columns[9:15]].mean())
plt.title("Female Rankings")
females[females.columns[9:15]].mean()
```

Out[43]:

```
ranking_1    3.429293
ranking_2    3.954660
ranking_3    4.418136
ranking_4    3.544081
ranking_5    2.569270
ranking_6    3.078086
dtype: float64
```



More men viewed the original movies and ranked them higher. More women viewed the newer movies but ranked them lower. The newer movies were more realistic and a little more violent. Especially the third movie in which Anikan is severely injured at the end.

In [ ]: