Final Project Milestone 3

Wyatt Rasmussen

DSC 540

Topic: Movies, this set of data will be pulled from Wikipedia

In [1]:

```
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
```

In [2]:

```
fd = open("List of Golden Globe winners - Wikipedia.html", "r")
soup = BeautifulSoup(fd)
fd.close()
```

In [3]:

```
all_tables = soup.find_all("table")
print("Total number of tables are {} ".format(len(all_tables)))
```

```
Total number of tables are 5
```

In [4]:

```
data_table = soup.findAll("table", {"class": 'wikitable'})
print(type(data_table))
```

```
<class 'bs4.element.ResultSet'>
```

In [5]:

```
table = data_table[2]
# table
# commented out table because the output file looked awful
```

In [6]:

```
headers = table.tbody.findAll('tr', recursive=False)[0]
headers_list = [th for th in headers.findAll('th')]
print(len(headers_list))
```

```
8
```

In [7]:

```
headers_list
```

Out[7]:

```
[<th>Year
 </th>,
 <th>Drama
 </th>,
 <th>Musical/Comedy
 </th>,
 <th>Drama Actor
 </th>,
 <th>Musical/Comedy Actor
 </th>,
 <th>Drama Actress
 </th>,
 <th>Musical/Comedy Actress
 </th>,
 <th>Director
 </th>]
```

In [8]:

```
headersList = headers_list[1:]
headersList
```

Out[8]:

```
[<th>Drama
 </th>,
 <th>Musical/Comedy
 </th>,
 <th>Drama Actor
 </th>,
 <th>Musical/Comedy Actor
 </th>,
 <th>Drama Actress
 </th>,
 <th>Musical/Comedy Actress
 </th>,
 <th>Director
 </th>]
```

In [9]:

```
rows = table.findChildren(['tr'])
```

In [10]:

```
dataRows = []
for tr in rows:
    td = tr.find_all('td')
    row = [tr.text for tr in td]
    dataRows.append(row)
```

In [11]:

```
# dataRows
# commented out the dataRows because output file looked awful
```

In [12]:

```
goldenGlobeWinner = pd.DataFrame(dataRows, columns = ['Drama', 'Musical/Comedy',
                                       'Drama Actress', 'Musical/Comedy Actress'
goldenGlobeWinner.head()
```

Out[12]:

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comedy Actress
0	None	None	None	None	None	None
1	A Place in the Sun\n	An American in Paris\n	Fredric March,Death of a Salesman\n	Danny Kaye,On the Riviera\n	Jane Wyman,The Blue Veil \n	June Allyson,Too Young to Kiss\n
2	The Greatest Show on Earth\n	With a Song in My Heart\n	Gary Cooper,High Noon\n	Donald O'Connor,Singin' in the Rain\n	Shirley Booth,Come Back, Little Sheba \n	Susan Hayward,With a Song in My Heart\n

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comedy Actress	
3	The Robe\n	No award\n	Spencer Tracy,The Actress\n	David Niven,The Moon Is Blue\n	Audrey Hepburn,Roman Holiday \n	Ethel Merman,Call Me Madam\n	Z
4	On the Waterfront \n	Carmen Jones\n	Marlon Brando,On the Waterfront\n	James Mason,A Star Is Born\n	Grace Kelly,The Country Girl \n	Judy Garland,A Star Is Born\n	t

Data Transformation 1

Dropping the first row of data since it is all empty

In [13]:

```
goldenGlobeWinner = goldenGlobeWinner.drop(labels=0, axis=0)
goldenGlobeWinner.head()
```

Out[13]:

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comed Actres
1	A Place in the Sun\n	An American in Paris\n	Fredric March,Death of a Salesman\n	Danny Kaye,On the Riviera\n	Jane Wyman,The Blue Veil \n	June Allyson,To Young to Kiss\
2	The Greatest Show on Earth\n	With a Song in My Heart\n	Gary Cooper,High Noon\n	Donald O'Connor,Singin' in the Rain\n	Shirley Booth,Come Back, Little Sheba \n	Susa Hayward,With Song in M Heart\
3	The Robe\n	No award\n	Spencer Tracy,The Actress\n	David Niven,The Moon Is Blue\n	Audrey Hepburn,Roman Holiday \n	Ethe Merman,Call M Madam\
4	On the Waterfront \n	Carmen Jones\n	Marlon Brando,On the Waterfront\n	James Mason,A Star Is Born\n	Grace Kelly,The Country Girl \n	Judy Garland, Star Is Born\
5	East of Eden\n	Guys and Dolls\n	Ernest Borgnine,Marty\n	Tom Ewell,The Seven Year Itch\n	Anna Magnani,The Rose Tattoo \n	Jea Simmons,Guy and Dolls\

Data Transformation 2

Adding a year column for the year the award was won

In [14]:

```
goldenGlobeWinner['Year'] = range(1952, len(goldenGlobeWinner) + 1952)
goldenGlobeWinner.head()
```

Out[14]:

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comed Actres
1	A Place in the Sun\n	An American in Paris\n	Fredric March,Death of a Salesman\n	Danny Kaye,On the Riviera\n	Jane Wyman,The Blue Veil \n	June Allyson,To Young to Kiss\

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comed Actres
2	The Greatest Show on Earth\n	With a Song in My Heart\n	Gary Cooper,High Noon\n	Donald O'Connor,Singin' in the Rain\n	Shirley Booth,Come Back, Little Sheba \n	Susa Hayward,With Song in M Heart\
3	The Robe\n	No award\n	Spencer Tracy,The Actress\n	David Niven,The Moon Is Blue\n	Audrey Hepburn,Roman Holiday \n	Ethe Merman,Call M Madam\
4	On the Waterfront \n	Carmen Jones\n	Marlon Brando,On the Waterfront\n	James Mason,A Star Is Born\n	Grace Kelly,The Country Girl \n	Judy Garland, Star Is Born\
5	East of Eden\n	Guys and Dolls\n	Ernest Borgnine,Marty\n	Tom Ewell,The Seven Year Itch\n	Anna Magnani,The Rose Tattoo \n	Jea Simmons,Guy and Dolls\

Data Transformation 3

This data transformation removes the \n from all data

In [15]:

```
goldenGlobeWinner = goldenGlobeWinner.replace('\n',' ', regex=True)
goldenGlobeWinner.head()
```

Out[15]:

	Drama	Musical/Comedy	Drama Actor	Musical/Comedy Actor	Drama Actress	Musical/Comedy Actress
1	A Place in the Sun	An American in Paris	Fredric March,Death of a Salesman	Danny Kaye,On the Riviera	Jane Wyman,The Blue Veil	June Allyson,Too Young to Kiss
2	The Greatest Show on Earth	With a Song in My Heart	Gary Cooper,High Noon	Donald O'Connor,Singin' in the Rain	Shirley Booth,Come Back, Little Sheba	Susan Hayward,With a Song in My Heart
3	The Robe	No award	Spencer Tracy,The Actress	David Niven,The Moon Is Blue	Audrey Hepburn,Roman Holiday	Ethel Merman,Call Me Madam
4	On the Waterfront	Carmen Jones	Marlon Brando,On the Waterfront	James Mason,A Star Is Born	Grace Kelly,The Country Girl	Judy Garland,A Star Is Born
5	East of Eden	Guys and Dolls	Ernest Borgnine,Marty	Tom Ewell,The Seven Year Itch	Anna Magnani,The Rose Tattoo	Jean Simmons,Guys and Dolls

Data Transformation 4

Separating the movie from the actor, actresses, and director awards.

In [16]:

```
new = goldenGlobeWinner["Drama Actor"].str.split(",", n = 1, expand = True)
new.head()
```

Out[16]:

	0	1
1	Fredric March	Death of a Salesman
2	Gary Cooper	High Noon
3	Spencer Tracy	The Actress
4	Marlon Brando	On the Waterfront
5	Ernest Borgnine	Marty

In [17]:

```
goldenGlobeWinner['Drama Actor (Actor)']= new[0]
goldenGlobeWinner['Drama Actor (Movie)']= new[1]
goldenGlobeWinner.drop(labels=["Drama Actor"], axis=1, inplace=True)
goldenGlobeWinner.head()
```

Out[17]:

	Drama	Musical/Comedy	Musical/Comedy Actor	Drama Actress	Musical/Comedy Actress	Directoı
1	A Place in the Sun	An American in Paris	Danny Kaye,On the Riviera	Jane Wyman,The Blue Veil	June Allyson,Too Young to Kiss	Laslc Benedek,Deatl of a Salesmar
2	The Greatest Show on Earth	With a Song in My Heart	Donald O'Connor,Singin' in the Rain	Shirley Booth,Come Back, Little Sheba	Susan Hayward,With a Song in My Heart	Cecil B DeMille,The Greatest Show on Eartl
3	The Robe	No award	David Niven,The Moon Is Blue	Audrey Hepburn,Roman Holiday	Ethel Merman,Call Me Madam	Frec Zinnemann,Fron Here to Eternity
4	On the Waterfront	Carmen Jones	James Mason,A Star Is Born	Grace Kelly,The Country Girl	Judy Garland,A Star Is Born	Elia Kazan,Or the Waterfron
5	East of Eden	Guys and Dolls	Tom Ewell,The Seven Year Itch	Anna Magnani,The Rose Tattoo	Jean Simmons,Guys and Dolls	Joshua Logan,Picnic

Repeating the process for the Musical/Comedy Actor Awards

In [18]:

```
new1 = goldenGlobeWinner["Musical/Comedy Actor"].str.split(",", n = 1, expand =
```

In [19]:

```
goldenGlobeWinner['Musical/Comedy Actor (Actor)']= new1[0]
goldenGlobeWinner['Musical/Comedy Actor (Movie)']= new1[1]
goldenGlobeWinner.drop(labels=["Musical/Comedy Actor"], axis=1, inplace=True)
goldenGlobeWinner.head()
```

Out[19]:

	Drama	Musical/Comedy	Drama Actress	Musical/Comedy Actress	Director	Year	Drama Actor (Actor)
1	A Place in the Sun	An American in Paris	Jane Wyman,The Blue Veil	June Allyson,Too Young to Kiss	Laslo Benedek,Death of a Salesman	1952	Fredric March

	Drama	Musical/Comedy	Drama Actress	Musical/Comedy Actress	Director	Year	Drama Actor (Actor)
2	The Greatest Show on Earth	With a Song in My Heart	Shirley Booth,Come Back, Little Sheba	Susan Hayward,With a Song in My Heart	Cecil B. DeMille,The Greatest Show on Earth	1953	Gary Cooper
3	The Robe	No award	Audrey Hepburn,Roman Holiday	Ethel Merman,Call Me Madam	Fred Zinnemann,From Here to Eternity	1954	Spencer Tracy
4	On the Waterfront	Carmen Jones	Grace Kelly,The Country Girl	Judy Garland,A Star Is Born	Elia Kazan,On the Waterfront	1955	Marlon Brando
5	East of Eden	Guys and Dolls	Anna Magnani,The Rose Tattoo	Jean Simmons,Guys and Dolls	Joshua Logan,Picnic	1956	Ernest Borgnine

Repeating again for Drama Actress

In [20]:

```
new2 = goldenGlobeWinner["Drama Actress"].str.split(",", n = 1, expand = True)
```

In [21]:

```
goldenGlobeWinner['Drama Actress (Actress)']= new2[0]
goldenGlobeWinner['Drama Actress (Movie)']= new2[1]
goldenGlobeWinner.drop(labels=["Drama Actress"], axis=1, inplace=True)
goldenGlobeWinner.head()
```

Out[21]:

	Drama	Musical/Comedy	Musical/Comedy Actress	Director	Year	Drama Actor (Actor)	Drama Actor (Movie)	Mus A
1	A Place in the Sun	An American in Paris	June Allyson,Too Young to Kiss	Laslo Benedek,Death of a Salesman	1952	Fredric March	Death of a Salesman	
2	The Greatest Show on Earth	With a Song in My Heart	Susan Hayward,With a Song in My Heart	Cecil B. DeMille,The Greatest Show on Earth	1953	Gary Cooper	High Noon	Dor
3	The Robe	No award	Ethel Merman,Call Me Madam	Fred Zinnemann,From Here to Eternity	1954	Spencer Tracy	The Actress	
4	On the Waterfront	Carmen Jones	Judy Garland,A Star Is Born	Elia Kazan,On the Waterfront	1955	Marlon Brando	On the Waterfront	.
5	East of Eden	Guys and Dolls	Jean Simmons,Guys and Dolls	Joshua Logan,Picnic	1956	Ernest Borgnine	Marty	

Repeating again for Musical/Comedy Actress

In [22]:

```
new3 = goldenGlobeWinner["Musical/Comedy Actress"].str.split(",", n = 1, expand
```

In [23]:

```
goldenGlobeWinner['Musical/Comedy Actress (Actress)']= new3[0]
```

```
goldenGlobeWinner['Musical/Comedy Actress (Movie)']= new3[1]
goldenGlobeWinner.drop(labels=["Musical/Comedy Actress"], axis=1, inplace=True)
goldenGlobeWinner.head()
```

Out[23]:

	Drama	Musical/Comedy	Director	Year	Drama Actor (Actor)	Drama Actor (Movie)	Musical/Comedy Actor (Actor)	Mus A
1	A Place in the Sun	An American in Paris	Laslo Benedek,Death of a Salesman	1952	Fredric March	Death of a Salesman	Danny Kaye	(
2	The Greatest Show on Earth	With a Song in My Heart	Cecil B. DeMille,The Greatest Show on Earth	1953	Gary Cooper	High Noon	Donald O'Connor	Sing
3	The Robe	No award	Fred Zinnemann,From Here to Eternity	1954	Spencer Tracy	The Actress	David Niven	The
4	On the Waterfront	Carmen Jones	Elia Kazan,On the Waterfront	1955	Marlon Brando	On the Waterfront	James Mason	/
5	East of Eden	Guys and Dolls	Joshua Logan,Picnic	1956	Ernest Borgnine	Marty	Tom Ewell	Th

Lastly repeating for Director

In [24]:

```
new4 = goldenGlobeWinner["Director"].str.split(",", n = 1, expand = True)
```

In [25]:

```
goldenGlobeWinner['Director (Director)']= new4[0]
goldenGlobeWinner['Director (Movie)']= new4[1]
goldenGlobeWinner.drop(labels=["Director"], axis=1, inplace=True)
goldenGlobeWinner.head()
```

Out[25]:

	Drama	Musical/Comedy	Year	Drama Actor (Actor)	Drama Actor (Movie)	Musical/Comedy Actor (Actor)	Musical/Comedy Actor (Movie)	I A (Ac
1	A Place in the Sun	An American in Paris	1952	Fredric March	Death of a Salesman	Danny Kaye	On the Riviera	W
2	The Greatest Show on Earth	With a Song in My Heart	1953	Gary Cooper	High Noon	Donald O'Connor	Singin' in the Rain	S
3	The Robe	No award	1954	Spencer Tracy	The Actress	David Niven	The Moon Is Blue	/ He
4	On the Waterfront	Carmen Jones	1955	Marlon Brando	On the Waterfront	James Mason	A Star Is Born	
5	East of Eden	Guys and Dolls	1956	Ernest Borgnine	Marty	Tom Ewell	The Seven Year Itch	M

Data Transformation 5

Setting the year as the index

In [26]:

```
goldenGlobeWinner = goldenGlobeWinner.set_index('Year')
goldenGlobeWinner.head()
```

Out[26]:

	Drama	Musical/Comedy	Drama Actor (Actor)	Drama Actor (Movie)	Musical/Comedy Actor (Actor)	Musical/Comedy Actor (Movie)	Dra Actre (Actre
Year							
1952	A Place in the Sun	An American in Paris	Fredric March	Death of a Salesman	Danny Kaye	On the Riviera	Ja Wym
1953	The Greatest Show on Earth	With a Song in My Heart	Gary Cooper	High Noon	Donald O'Connor	Singin' in the Rain	Shir Bo
1954	The Robe	No award	Spencer Tracy	The Actress	David Niven	The Moon Is Blue	Aud Hepb
1955	On the Waterfront	Carmen Jones	Marlon Brando	On the Waterfront	James Mason	A Star Is Born	Gra K
1956	East of Eden	Guys and Dolls	Ernest Borgnine	Marty	Tom Ewell	The Seven Year Itch	Ar Magn

Final Dataframe

In [27]:

```
goldenGlobeWinner
```

Out[27]:

	Drama	Musical/Comedy	Drama Actor (Actor)	Drama Actor (Movie)	Musical/Comedy Actor (Actor)	Musical/Comedy Actor (Movie)	(A
Year							
1952	A Place in the Sun	An American in Paris	Fredric March	Death of a Salesman	Danny Kaye	On the Riviera	
1953	The Greatest Show on Earth	With a Song in My Heart	Gary Cooper	High Noon	Donald O'Connor	Singin' in the Rain	
1954	The Robe	No award	Spencer Tracy	The Actress	David Niven	The Moon Is Blue	I
1955	On the Waterfront	Carmen Jones	Marlon Brando	On the Waterfront	James Mason	A Star Is Born	Gra
1956	East of Eden	Guys and Dolls	Ernest Borgnine	Marty	Tom Ewell	The Seven Year Itch	

	Drama	Musical/Comedy	Drama Actor (Actor)	Drama Actor (Movie)	Musical/Comedy Actor (Actor)	Musical/Comedy Actor (Movie)	(A
Year							
...	...	...	...	...	...	...	
2017	Moonlight	La La Land	Casey Affleck	Manchester by the Sea	Ryan Gosling	La La Land	
2018	Three Billboards Outside Ebbing, Missouri	Lady Bird	Gary Oldman	Darkest Hour	James Franco	The Disaster Artist	McD
2019	Bohemian Rhapsody	Green Book	Rami Malek	Bohemian Rhapsody	Christian Bale	Vice	Glen
2020	1917	Once Upon a Time in Hollywood	Joaquin Phoenix	Joker	Taron Egerton	Rocketman	Z
2021	Nomadland	Borat Subsequent Moviefilm	Chadwick Boseman	Ma Rainey's Black Bottom	Sacha Baron Cohen	Borat Subsequent Moviefilm	Ar

70 rows × 12 columns

```
In [ ]:
```