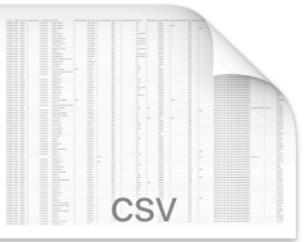


An Oscars Demographics' Data Tour

Keke Wu

01 DATA DESCRIPTION

DATA DESCRIPTION



Oscars-demographics-
DFE.csv

90 KB

AutoSave OFF Home Insert Draw Page Layout Formulas Data Review View

Calibri (Body) 12 A B C D E F G H I J K L M N O P Q R S T U V

Paste B I U Wrap Text General \$ % , < > Conditional Formatting as Table Insert Delete Format Sort & Filter Ideas

A1 fx _unit_id

1 _unit_id golden _unit_state _trusted_jud _last_judgm birthplace birthplacecc date_of_birth date_of_birt race_ethnic race_ethnic religion religionconf sexual_orient sexual_orient year_of_awd year_of_awd award biouri birthplace_g date_of_birt movie

2 670454353 FALSE finalized 3 ##### Chisinau, Mc 1 30-Sep-1895 1 White 1 Na 1 Straight 1 1927 1 Best Director http://www.imdb.com/people/320/00 Two Arali

3 670454354 FALSE finalized 3 ##### Glasgow, Sci 1 2-Feb-1886 1 White 1 Na 1 Straight 1 1930 1 Best Director http://www.imdb.com/people/626/00 The Devil

4 670454355 FALSE finalized 3 ##### Chisinau, Mc 1 30-Sep-1895 1 White 1 Na 1 Straight 1 1931 0.6667 Best Director http://www.imdb.com/people/320/00 All Quiet

5 670454356 FALSE finalized 3 ##### Chicago, Il 1 23-Feb-1899 1 White 1 Na 1 Straight 1 1932 1 Best Director http://www.imdb.com/people/544/00 Skippy

6 670454357 FALSE finalized 3 ##### Salt Lake City 1 23-Apr-1894 1 White 1 Roman Cath 1 Straight 1 1933 1 Best Director http://www.imdb.com/people/292/00 Bad Girl

7 670454358 FALSE finalized 3 ##### Glasgow, Sci 1 2-Feb-1886 1 White 1 Na 1 Straight 1 1934 1 Best Director http://www.imdb.com/people/626/00 Cavalcade

8 670454359 FALSE finalized 3 ##### Bisacquino, S 1 18-May-1897 1 White 1 Roman Cath 1 Straight 1 1935 1 Best Director http://www.imdb.com/people/459/00 Happen

9 670454360 FALSE finalized 3 ##### Bisacquino, S 1 1-Feb-1891 1 White 1 Roman Cath 1 Bisexual 1 1936 1 Best Director http://www.imdb.com/people/458/00 The Inform

10 670454361 FALSE finalized 3 ##### Bisacquino, S 1 18-May-1897 1 White 1 Roman Cath 1 Straight 1 1937 1 Best Director http://www.imdb.com/people/459/00 Mr. Deeds

11 670454362 FALSE finalized 3 ##### Los Angeles, 1 3-Oct-1899 1 White 1 Na 1 Straight 1 1938 1 Best Director http://www.imdb.com/people/380/00 The Awful

12 670454363 FALSE finalized 3 ##### Bisacquino, S 1 18-May-1897 1 White 1 Roman Cath 1 Straight 1 1939 1 Best Director http://www.imdb.com/people/459/00 You Can't

13 670454364 FALSE finalized 3 ##### Pasadena, Ci 1 23-Feb-1883 1 White 1 Na 1 Straight 1 1940 1 Best Director http://www.imdb.com/people/416/00 Gone With

14 670454365 FALSE finalized 3 ##### Cape Elizabeth 1 1-Feb-1894 1 White 1 Roman Cath 1 Bisexual 1 1941 1 Best Director http://www.imdb.com/people/458/00 The Grape

15 670454366 FALSE finalized 3 ##### Cape Elizabeth 1 1-Feb-1894 1 White 1 Roman Cath 1 Bisexual 1 1942 1 Best Director http://www.imdb.com/people/458/00 How Gree

16 670454367 FALSE finalized 3 ##### Mulhouse, H 1 1-Jul-02 1 White 1 Na 1 Straight 1 1943 1 Best Director http://www.imdb.com/people/463/00 Mrs. Miniv

17 670454368 FALSE finalized 3 ##### Budapest, H 1 24-Dec-1886 1 White 1 Na 1 Straight 1 1944 1 Best Director http://www.imdb.com/people/676/00 Casablan

18 670454369 FALSE finalized 3 ##### Los Angeles, 1 3-Oct-1898 1 White 1 Na 1 Straight 1 1945 1 Best Director http://www.imdb.com/people/380/00 Going My

19 670454370 FALSE finalized 3 ##### New York Cit 1 11-Oct-18 1 White 1 Jewish 1 Gay 1 1946 1 Best Director http://www.imdb.com/people/281/00 West Side

20 670454371 FALSE finalized 3 ##### Sucha, Galici 0.6667 22-Jun-06 1 White 1 Jewish 1 Straight 1 1946 1 Best Director http://www.imdb.com/people/982/00 The Lost V

21 670454372 FALSE finalized 3 ##### Mulhouse, H 1 1-Jul-02 1 White 1 Na 1 Straight 1 1947 1 Best Director http://www.imdb.com/people/463/00 The Best Y

22 670454373 FALSE finalized 3 ##### Istanbul, Tur 1 7-Sep-09 1 White 1 Na 1 Straight 1 1948 1 Best Director http://www.imdb.com/people/537/00 Gentlema

23 670454374 TRUE golden 82 Nevada, Mo 1 5-Aug-06 1 White 1 Na 1 Straight 1 1949 1 Best Director http://www.imdb.com/people/537/00 On the Wi

24 670454375 FALSE finalized 3 ##### Wilkes-Barre 1 11-Feb-09 1 White 1 Jewish 1 Straight 1 1950 1 Best Director http://www.imdb.com/people/723/00 A Letter to

25 670454376 FALSE finalized 3 ##### Wilkes-Barre 1 11-Feb-09 1 White 1 Jewish 1 Straight 1 1951 1 Best Director http://www.imdb.com/people/723/00 All About I

26 670454377 FALSE finalized 3 ##### Oakland, Ca 1 18-Dec-04 1 White 1 Na 1 Straight 1 1952 1 Best Director http://www.imdb.com/people/382/00 A Place In

27 670454378 FALSE finalized 3 ##### Cape Elizabeth 1 1-Feb-1894 1 White 1 Roman Cath 1 Bisexual 1 1953 1 Best Director http://www.imdb.com/people/458/00 The Quiet

28 670454379 FALSE finalized 3 ##### Vienna, Aust 1 29-Apr-07 1 White 1 Jewish 1 Straight 1 1954 1 Best Director http://www.imdb.com/people/538/00 From Here

29 670454380 FALSE finalized 3 ##### Istanbul, Tur 1 7-Sep-09 1 White 1 Na 1 Straight 1 1955 1 Best Director http://www.imdb.com/people/537/00 On the Wi

30 670454381 FALSE finalized 3 ##### Lawrence, Ks 1 30-Jan-20 1 White 1 Presbyterian 1 Straight 1 1956 1 Best Director http://www.imdb.com/people/384/00 Marty

31 670454382 FALSE finalized 3 ##### Oakland, Ca 1 18-Dec-04 1 White 1 Na 1 Straight 1 1957 1 Best Director http://www.imdb.com/people/382/00 Giant

32 670454383 FALSE finalized 3 ##### Croydon, Sur 0.6667 25-Mar-08 1 White 1 Na 1 Straight 0.6667 1958 1 Best Director http://www.imdb.com/people/462/00 The Bridg

33 670454384 FALSE finalized 3 ##### Chicago, Il 1 28-Feb-03 1 White 1 Na 1 Bisexual 1 1959 1 Best Director http://www.imdb.com/people/825/00 Gigi

34 670454385 FALSE finalized 3 ##### Mulhouse, H 1 1-Jul-02 1 White 1 Na 1 Straight 1 1960 1 Best Director http://www.imdb.com/people/463/00 Ben-Hur

35 670454386 TRUE golden 87 Sucha, Galici 1 22-Jun-06 1 White 1 Jewish 0.963 Straight 1 1961 1 Best Director http://www.imdb.com/people/281/00 The Apartm

36 670454387 FALSE finalized 3 ##### Winchester, 1 10-Sep-14 1 White 1 Na 1 Straight 1 1961 0.6667 Best Director http://www.imdb.com/people/460/00 West Side

37 670454388 FALSE finalized 3 ##### Croydon, Sur 1 25-Mar-08 1 White 1 Na 1 Straight 1 1963 1 Best Director http://www.imdb.com/people/462/00 Lawrence

38 670454389 FALSE finalized 3 ##### Shipton, York 1 5-Jun-28 1 White 1 Na 1 Bisexual 1 1964 1 Best Director http://www.imdb.com/people/249/00 Tom Jones

39 670454390 FALSE finalized 3 ##### New York Cit 1 7-Jul-1899 1 White 1 Na 0.6571 Gay 0.6571 1965 1 Best Director http://www.imdb.com/people/373/00 My Fair La

40 670454391 FALSE finalized 3 ##### Winchester, 1 10-Sep-14 1 White 1 Na 1 Straight 1 1966 1 Best Director http://www.imdb.com/people/460/00 The Sound

41 670454392 FALSE finalized 3 ##### Vienna, Aust 1 29-Apr-07 1 White 1 Jewish 1 Straight 1 1967 1 Best Director http://www.imdb.com/people/538/00 A Man for

(441,27)

DATA DESCRIPTION

Load the data

```
oscars_original = pd.read_csv('Oscars-demographics-DFE.csv', encoding = "ISO-8859-1", parse_dates=True)
for col in oscars_original.columns:
    print(col)
```

```
_unit_id
_golden
_unit_state
_trusted_judgments
_last_judgment_at
birthplace
birthplace:confidence
date_of_birth
date_of_birth:confidence
race_ethnicity
race_ethnicity:confidence
religion
religion:confidence
sexual_orientation
sexual_orientation:confidence
year_of_award
year_of_award:confidence
award
biourl
birthplace_gold
date_of_birth_gold
movie
person
race_ethnicity_gold
religion_gold
sexual_orientation_gold
year_of_award_gold
```

DATA DESCRIPTION

Data types

_unit_id
_golden
_unit_state
_trusted_judgments
_last_judgment_at
birthplace
birthplace:confidence
date_of_birth
date_of_birth:confidence
race_ethnicity
race_ethnicity:confidence
religion
religion:confidence
sexual_orientation
sexual_orientation:confidence
year_of_award
year_of_award:confidence
award
biourl
birthplace_gold
date_of_birth_gold
movie
person
race_ethnicity_gold
religion_gold
sexual_orientation_gold
year_of_award_gold

NUMERICAL

Continuous : _unit_id, _last_judgement_at_, birthplace:confidence, date_of_birth, date_of_birth:confidence, race_ethnicity:confidence, religion:confidence, sexual_orientation:confidence, year_of_award:confidence, year_of_award_gold

Discrete: _trusted_judgements, year_of_award

CATEGORICAL

Nominal : _golden, _unit_state, birthplace, race_ethnicity, religion, sexual_orientation, award, birthplace_gold, date_of_birth_gold, movie, person, race_ethnicity_gold, religion_gold, sexual_orientation_gold,

Free: biourl

DATA DESCRIPTION

Questions to answer

1. How diverse are the Oscars in terms of race & ethnicity?
2. How diverse are the Oscars in terms of gender?
3. Where are most of the winners from?
4. How likely is it to win both the Oscars and the Golden Globe?
5. Which month were most of the winners born in?
6. How old were they when they first won the Oscars?
7. How many people won more than once?

02 DATA PREPARATION

DATA PREPARATION

1. Construct a new data frame that contains only the columns of interest

```
oscars = pd.DataFrame(oscars_original, columns=['_golden', 'birthplace', 'race_ethnicity', 'date_of_birth', 'year_of_award', 'award', 'movie', 'person'])
```

	_golden	birthplace	race_ethnicity	date_of_birth	year_of_award	award	movie	person
161	False	New York City	White	14-Apr-73	2003	Best Actor	The Pianist	Adrien Brody
151	False	East Harlem, New York City	White	25-Apr-40	1993	Best Actor	Scent of a Woman	Al Pacino
243	False	Brooklyn, Ny	White	26-Mar-34	2007	Best Supporting Actor	Little Miss Sunshine	Alan Arkin
116	False	London, England	White	2-Apr-14	1958	Best Actor	The Bridge on the River Kwai	Alec Guinness
339	False	New York City	White	2-Nov-1892	1938	Best Supporting Actress	In Old Chicago	Alice Brady
79	False	Pingtung, Taiwan	Asian	23-Oct-54	2006	Best Director	Brokeback Mountain	Ang Lee
86	False	Pingtung, Taiwan	Asian	23-Oct-54	2013	Best Director	Life of Pi	Ang Lee
401	True	Los Angeles, Ca	White	4-Jun-75	2000	Best Supporting Actress	Girl, Interrupted	Angelina Jolie
421	False	Los Angeles, Ca	White	4-Jun-75	2000	Best Supporting Actress	Girl, Interrupted	Angelina Jolie
387	False	Santa Monica, Ca	White	8-Jul-51	1986	Best Supporting Actress	Prizzi's Honor	Anjelica Huston
278	False	Rome, Italy	White	7-Mar-08	1956	Best Actress	The Rose Tattoo	Anna Magnani
395	False	Winnipeg, Manitoba, Canada	White	24-Jul-82	1994	Best Supporting Actress	The Piano	Anna Paquin
285	False	Bronx, Ny	White	17-Sep-31	1963	Best Actress	The Miracle Worker	Anne Bancroft

DATA PREPARATION

1. Construct a new data frame that contains only the columns of interest

```
oscars = pd.DataFrame(oscars_original, columns=['_golden', 'birthplace', 'race_ethnicity', 'date_of_birth', 'year_of_award', 'award', 'movie', 'person'])
```

	_golden	birthplace	race_ethnicity	date_of_birth	year_of_award	award	movie	person
161	False	New York City	White	14-Apr-73	2003	Best Actor	The Pianist	Adrien Brody
151	False	East Harlem, New York City	White	25-Apr-40	1993	Best Actor	Scent of a Woman	Al Pacino
243	False	Brooklyn, Ny	White	26-Mar-34	2007	Best Supporting Actor	Little Miss Sunshine	Alan Arkin
116	False	London, England	White	2-Apr-14	1958	Best Actor	The Bridge on the River Kwai	Alec Guinness
339	False	New York City	White	2-Nov-1892	1938	Best Supporting Actress	In Old Chicago	Alice Brady
79	False	Pingtung, Taiwan	Asian	23-Oct-54	2006	Best Director	Brokeback Mountain	Ang Lee
86	False	Pingtung, Taiwan	Asian	23-Oct-54	2013	Best Director	Life of Pi	Ang Lee
401	True	Los Angeles, Ca	White	4-Jun-75	2000	Best Supporting Actress	Girl, Interrupted	Angelina Jolie
421	False	Los Angeles, Ca	White	4-Jun-75	2000	Best Supporting Actress	Girl, Interrupted	Angelina Jolie
387	False	Santa Monica, Ca	White	8-Jul-51	1986	Best Supporting Actress	Prizzi's Honor	Anjelica Huston
278	False	Rome, Italy	White	7-Mar-08	1956	Best Actress	The Rose Tattoo	Anna Magnani
395	False	Winnipeg, Manitoba, Canada	White	24-Jul-82	1994	Best Supporting Actress	The Piano	Anna Paquin
285	False	Bronx, Ny	White	17-Sep-31	1963	Best Actress	The Miracle Worker	Anne Bancroft

DATA PREPARATION

2. Remove the duplicated rows

`.drop_duplicates()`

```
oscars.drop_duplicates(subset ='movie', keep = 'last', inplace = True)
```

```
oscars
```

	_golden	birthplace	race_ethnicity	date_of_birth	year_of_award	award	movie	person
151	False	East Harlem, New York City	White	25-Apr-40	1993	Best Actor	Scent of a Woman	Al Pacino
243	False	Brooklyn, Ny	White	26-Mar-34	2007	Best Supporting Actor	Little Miss Sunshine	Alan Arkin
339	False	New York City	White	2-Nov-1892	1938	Best Supporting Actress	In Old Chicago	Alice Brady
79	False	Pingtung, Taiwan	Asian	23-Oct-54	2006	Best Director	Brokeback Mountain	Ang Lee
86	False	Pingtung, Taiwan	Asian	23-Oct-54	2013	Best Director	Life of Pi	Ang Lee
421	False	Los Angeles, Ca	White	4-Jun-75	2000	Best Supporting Actress	Girl, Interrupted	Angelina Jolie
387	False	Santa Monica, Ca	White	8-Jul-51	1986	Best Supporting Actress	Prizzi's Honor	Anjelica Huston

03 DATA ANALYSES

DATA ANALYSES

1. Oscars Winners by Race and Ethnicity

```
race_counts = oscars.race_ethnicity.value_counts()  
race_counts
```

```
White          314  
Black           13  
Hispanic         5  
Asian            3  
Multiracial      1  
Name: race_ethnicity, dtype: int64
```

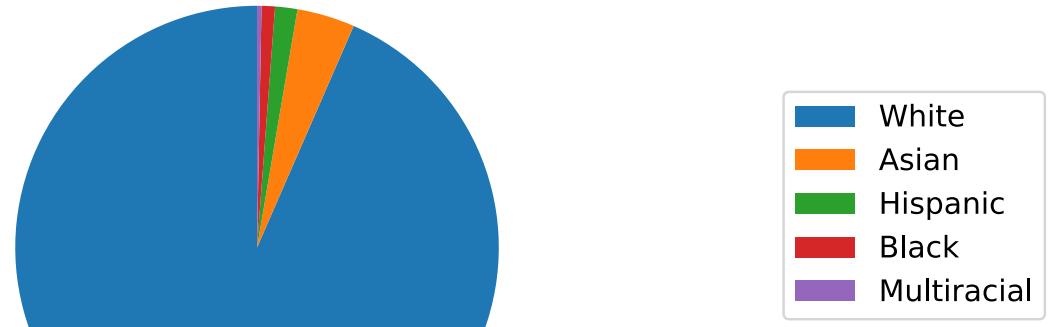
DATA ANALYSES

1. Oscars Winners by Race and Ethnicity

```
race_counts = oscars.race_ethnicity.value_counts()  
race_counts
```

```
White           314  
Black            13  
Hispanic          5  
Asian              3  
Multiracial        1  
Name: race_ethnicity, dtype: int64
```

Oscars Winners by Race or Ethnicity



Super racially skewed, right?

DATA ANALYSES

2. Oscars Winners by Award Category (Gender)

```
oscars.award.value_counts()
```

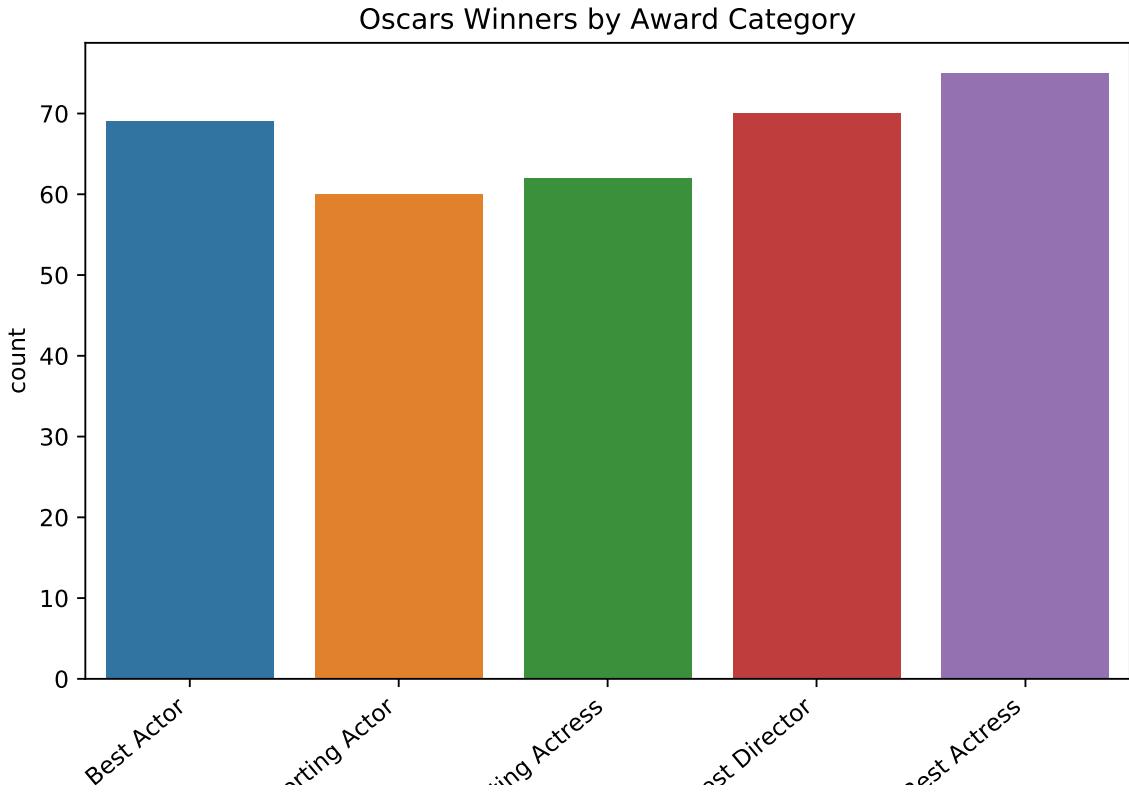
```
Best Actress          75
Best Director         70
Best Actor            69
Best Supporting Actress 62
Best Supporting Actor  60
Name: award, dtype: int64
```

DATA ANALYSES

2. Oscars Winners by Award Category (Gender)

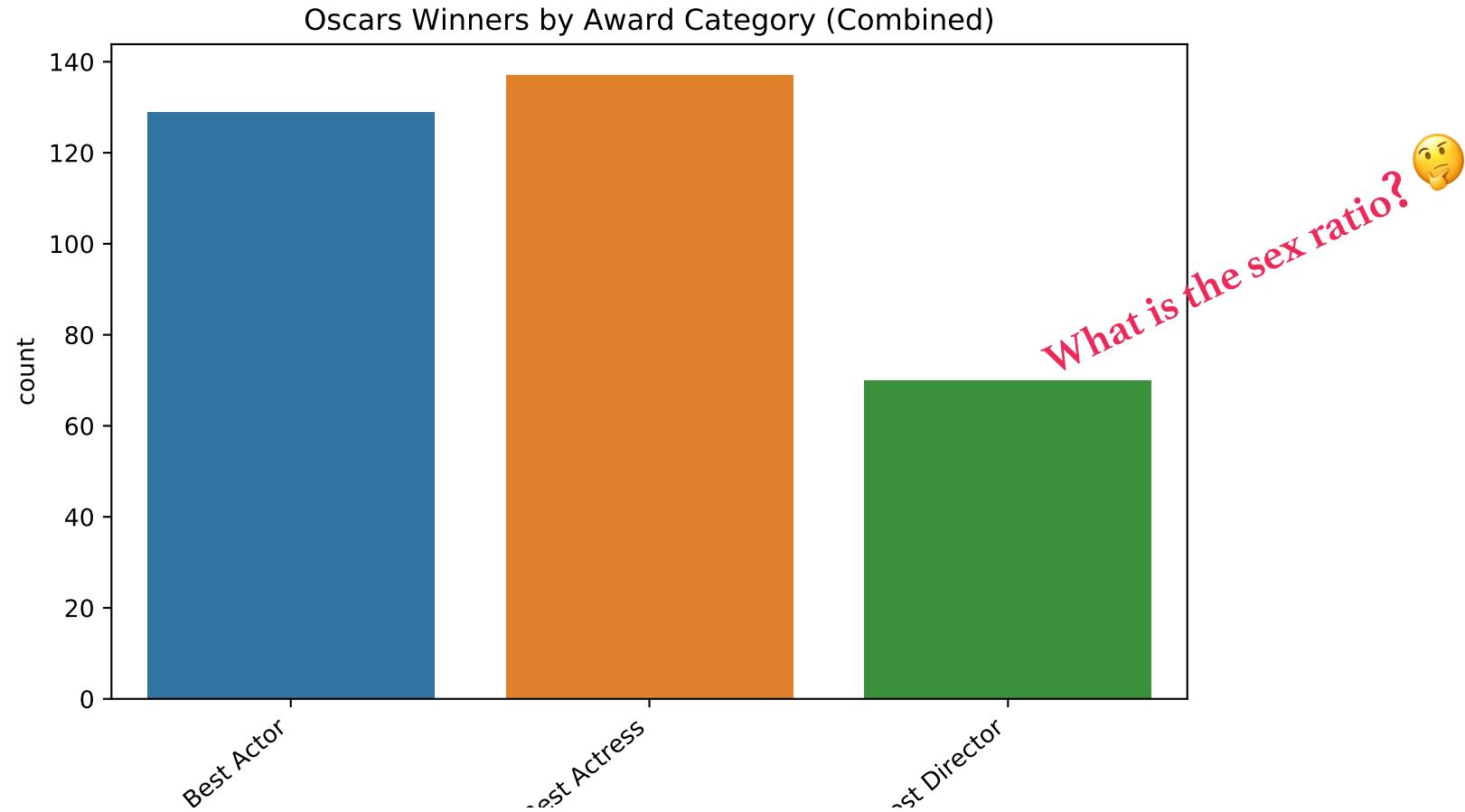
```
oscars.award.value_counts()
```

```
Best Actress          75
Best Director         70
Best Actor            69
Best Supporting Actress 62
Best Supporting Actor 60
Name: award, dtype: int64
```



DATA ANALYSES

2. Oscars Winners by Award Category (Gender)



DATA ANALYSES

3. Where are most Oscars winners from?

```
from wordcloud import WordCloud
text = str(oscars['birthplace'])
wordcloud = WordCloud(width=1000, height=500, max_words=30, colormap="Blues").generate(text)
p = plt.figure(figsize=(20,10) )
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
p.savefig("oscar_winner.pdf")
```

DATA ANALYSES

3. Where are most Oscars winners from?

```
from wordcloud import WordCloud
text = str(oscars['birthplace'])
wordcloud = WordCloud(width=1000, height=500, max_words=30, colormap="Blues").generate(text)
p = plt.figure(figsize=(20,10) )
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
p.savefig("oscar_winner.pdf")
```



DATA ANALYSES

3. Where are most Oscars winners from?

```
from wordcloud import WordCloud
text = str(oscars['birthplace'])
wordcloud = WordCloud(width=1000, height=500, max_words=30, colormap="Blues").generate(text)
p = plt.figure(figsize=(20,10) )
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.margins(x=0, y=0)
plt.show()
p.savefig("oscar_winner.pdf")
```



England

London

New York

Brooklyn

Ca

DATA ANALYSES

4. Which month were most Oscars winners born in?

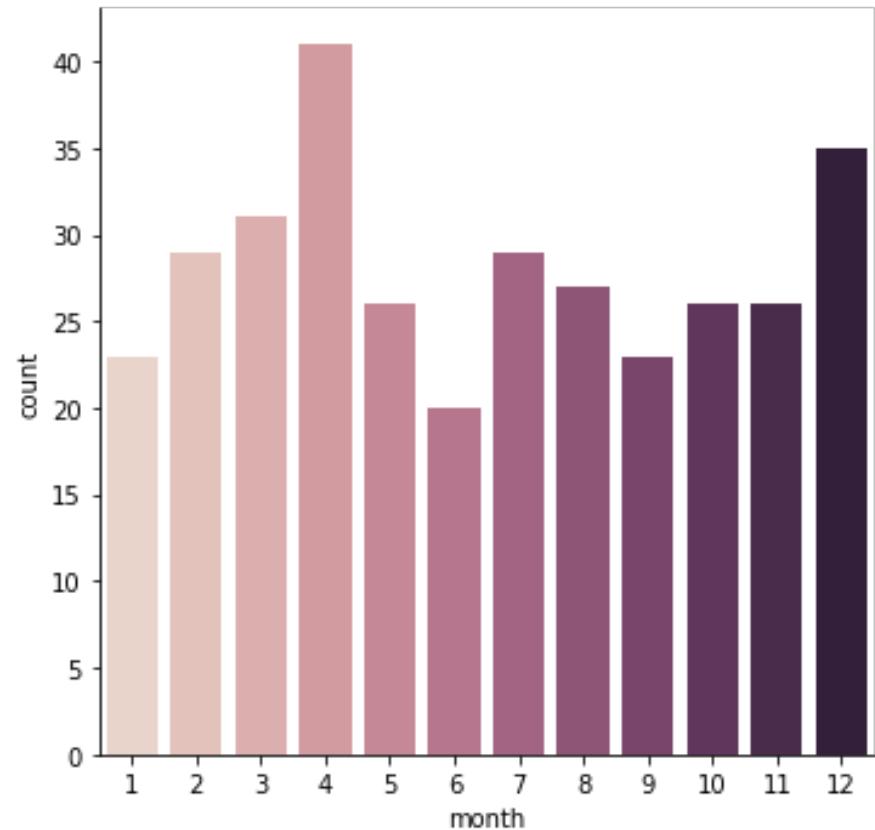
```
oscars[ 'month' ] = pd.DatetimeIndex(oscars[ 'date_of_birth' ]).month  
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```

.DatetimeIndex()

DATA ANALYSES

4. Which month were most Oscars winners born in?

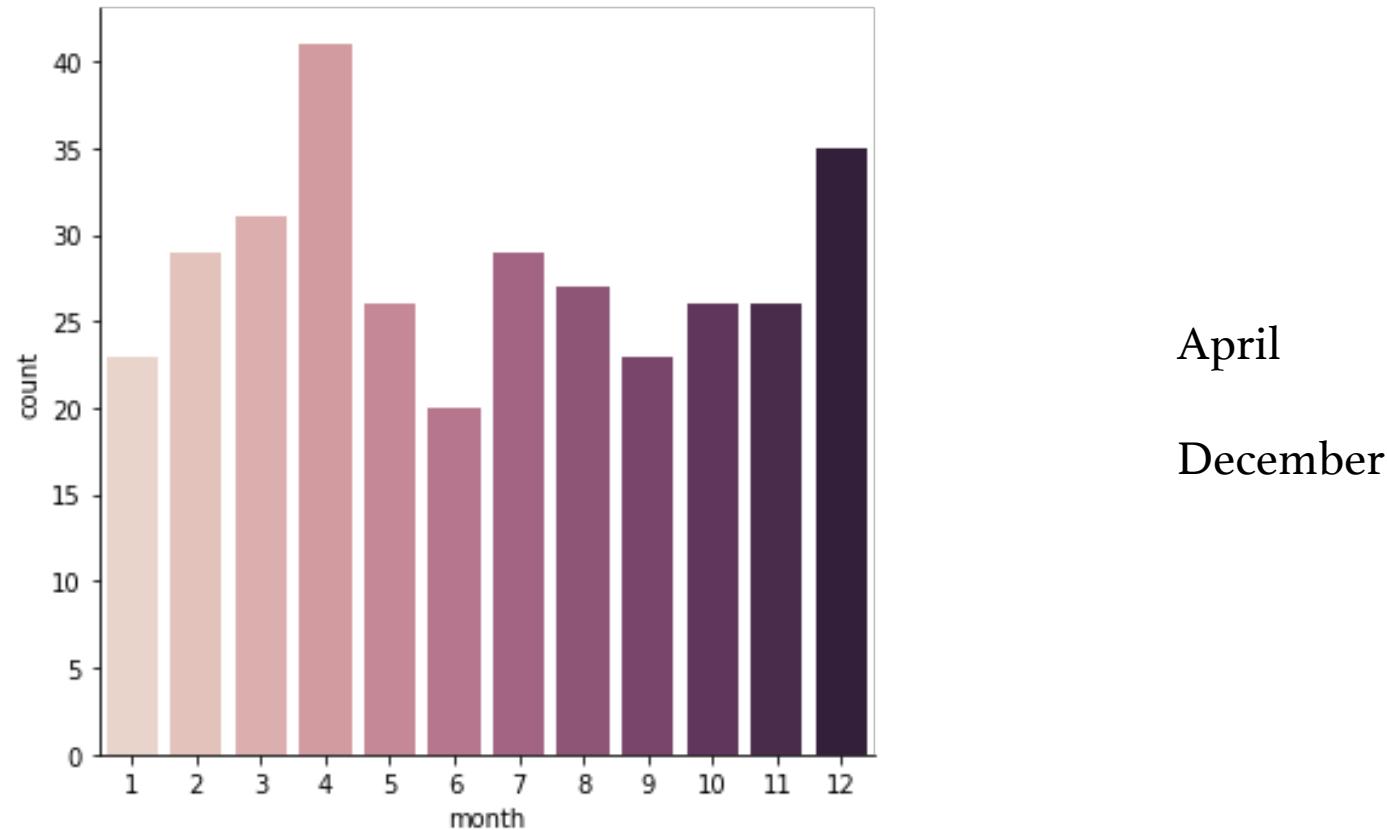
```
oscars['month'] = pd.DatetimeIndex(oscars['date_of_birth']).month  
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```



DATA ANALYSES

4. Which month were most Oscars winners born in?

```
oscars[ 'month' ] = pd.DatetimeIndex(oscars[ 'date_of_birth' ]).month  
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```



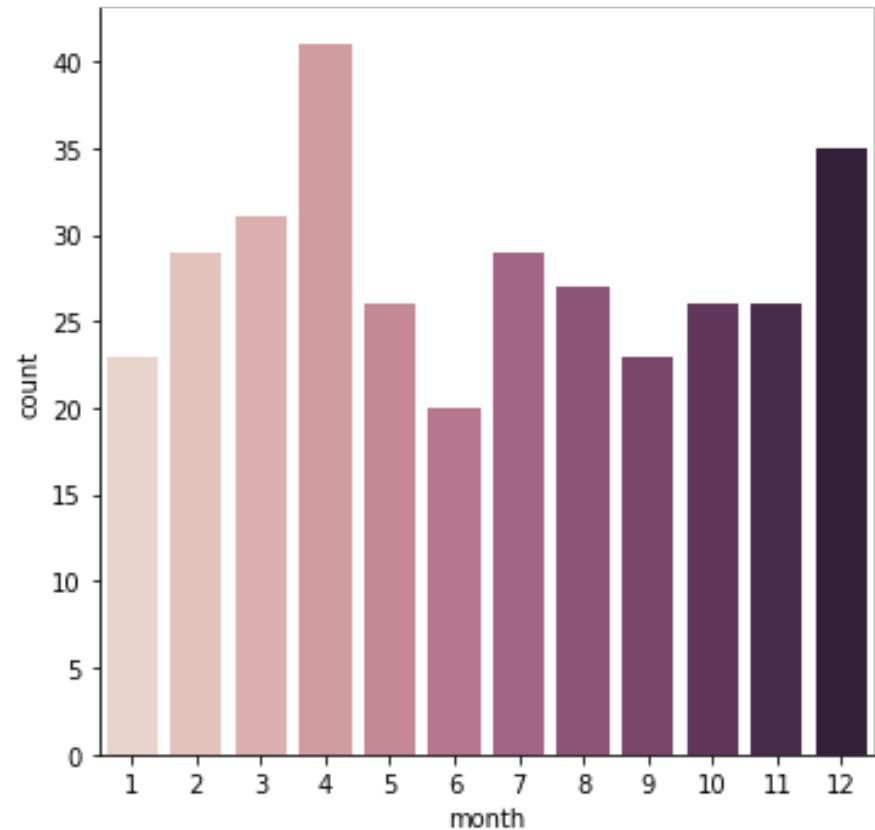
April

December

DATA ANALYSES

4. Which month were most Oscars winners born in?

```
oscars['month'] = pd.DatetimeIndex(oscars['date_of_birth']).month  
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```



April

December

```
apr = oscars.loc[oscars['month'] == 4]  
apr.shape
```

(41, 10)

DATA ANALYSES

5. What is the chance to win both the Oscars and the Golden Globe Award?

```
golden_global = oscars[oscars._golden == True].copy()
golden_oscar = len(golden_global) / len(oscars)
print("The chance of winning both the Oscars and the Golden Globe Award is" + " {:.2%}".format(golden_oscar));
```

The chance of winning both the Oscars and the Golden Globe Award is 1.49%

04 TAKEAWAYS

TAKEAWAYS

Visual Information-Seeking Mantra:

Overview first, zoom and filter, then details-on-demand.

TAKEAWAYS

Visual Information-Seeking Mantra:

Overview first, zoom and filter, then details-on-demand.

High Quality Data is the Key:

Data preparation is important, otherwise, you'll like to encounter problems like data duplication, mixed formats of data, etc., later on.

TAKEAWAYS

Visual Information-Seeking Mantra:

Overview first, zoom and filter, then details-on-demand.

High Quality Data is the Key:

Data preparation is important, otherwise, you'll like to encounter problems like data duplication, mixed formats of data, etc., later on.

1-Feb-1894

1-Feb-1894

1-Jul-02

24-Dec-1886

3-Oct-1898

11-Oct-18

22-Jun-06

1-Jul-02

7-Sep-09

5-Aug-06

11-Feb-09

11-Feb-09

18-Dec-04

Python time method **strptime()** parses a string representing a time according to a format. The return value is a `struct_time` as returned by `gmtime()` or `localtime()`.

The format parameter uses the same directives as those used by `strftime()`; it defaults to "%a %b %d %H:%M:%S %Y" which matches the formatting returned by `ctime()`.

If string cannot be parsed according to format, or if it has excess data after parsing, `ValueError` is raised.

TAKEAWAYS

Visual Information-Seeking Mantra:

Overview first, zoom and filter, then details-on-demand.

High Quality Data is the Key:

Data preparation is important, otherwise, you'll like to encounter problems like data duplication, mixed formats of data, etc., later on.

Use Visualization Across Data Analyses:

Not just to communicate your findings and results, but also a way to explore the dataset.

T H A N K S !