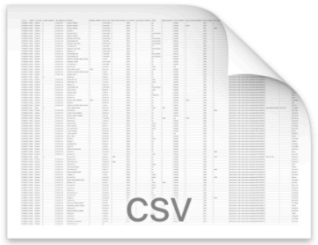
INFO 5871 Fall 19 Intro to Data Science Midterm

An Oscars Demographics' Data Tour

Keke Wu

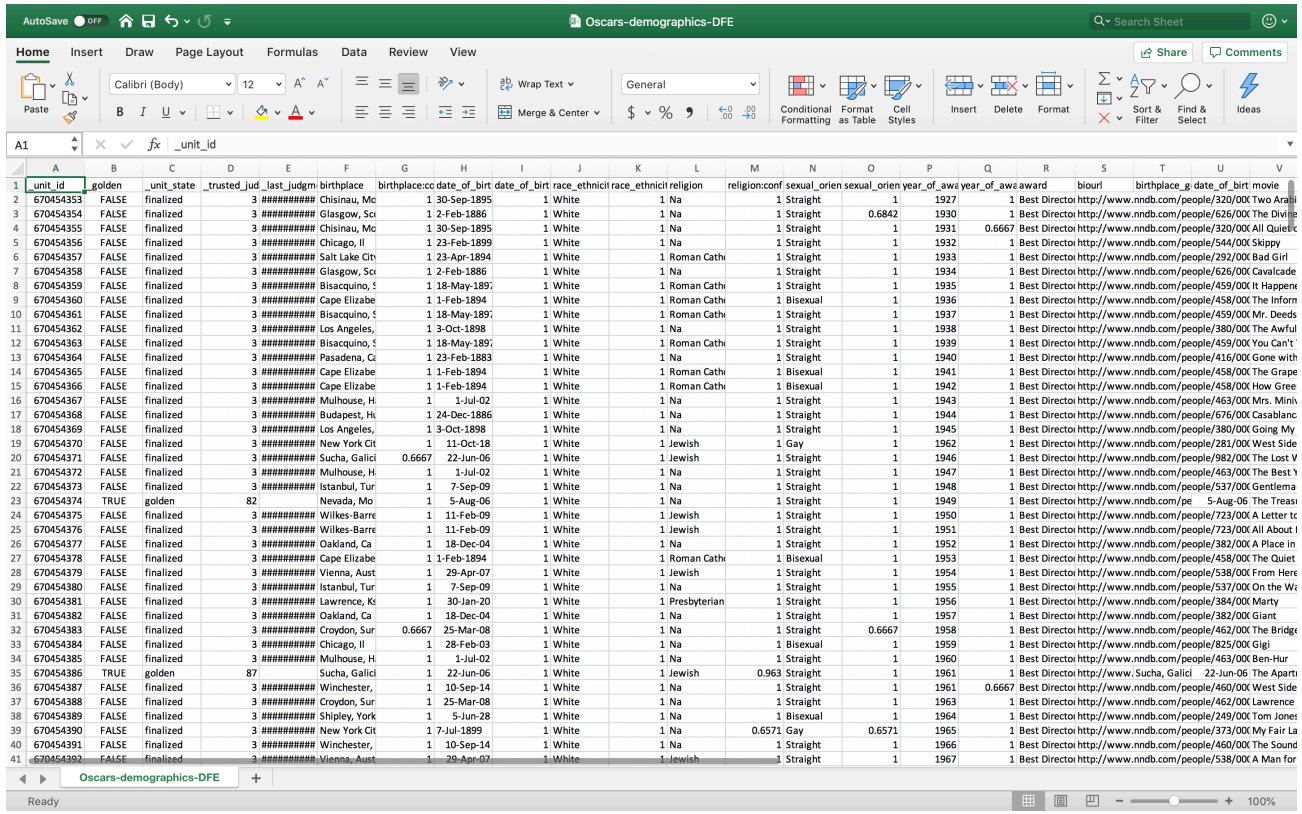
01 DATA DESCRIPTION

DATA DESCRIPTION



Oscars-demographics-DFE.csv

90 KB



(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_oridentation: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_oridentation, 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
```

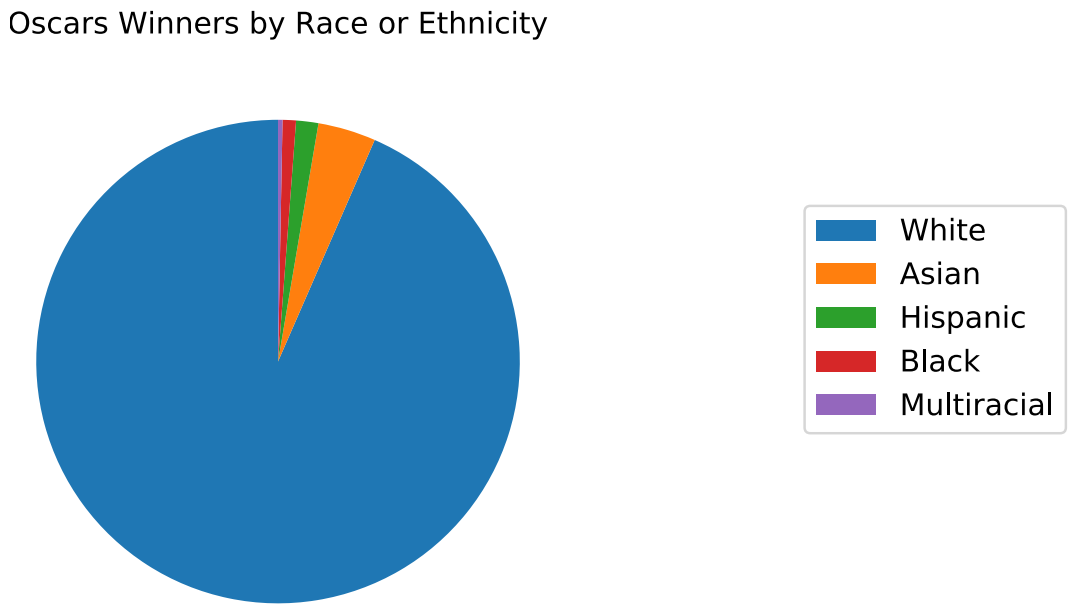
```
White           314
Black            13
Hispanic          5
Asian             3
Multiracial       1
Name: race_ethnicity, dtype: int64
```

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
```



😱 Super racially skewed, right?

DATA ANALYSES

2. Oscars Winners by Award Category (Gender)

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

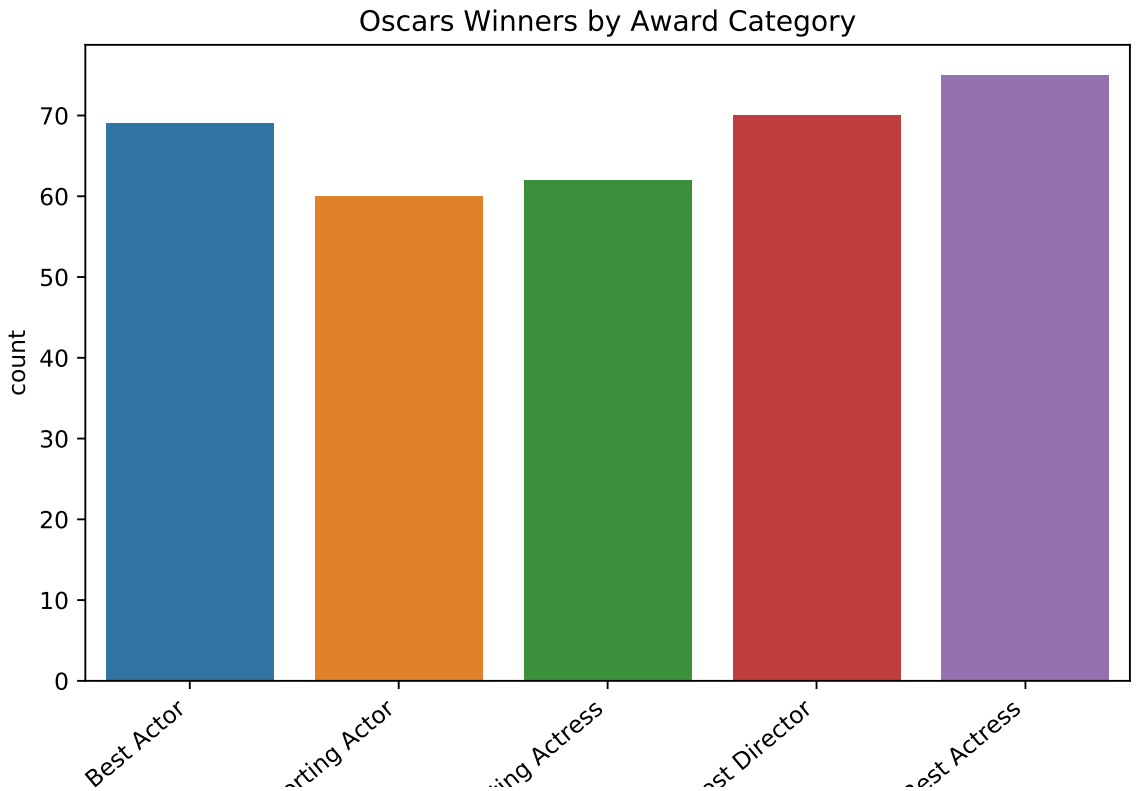
```
Best Actress                75
Best Director               70
Best Actor                  69
Best Supporting Actress     62
Best Supporting Actor       60
Name: award, dtype: int64
```

DATA ANALYSES

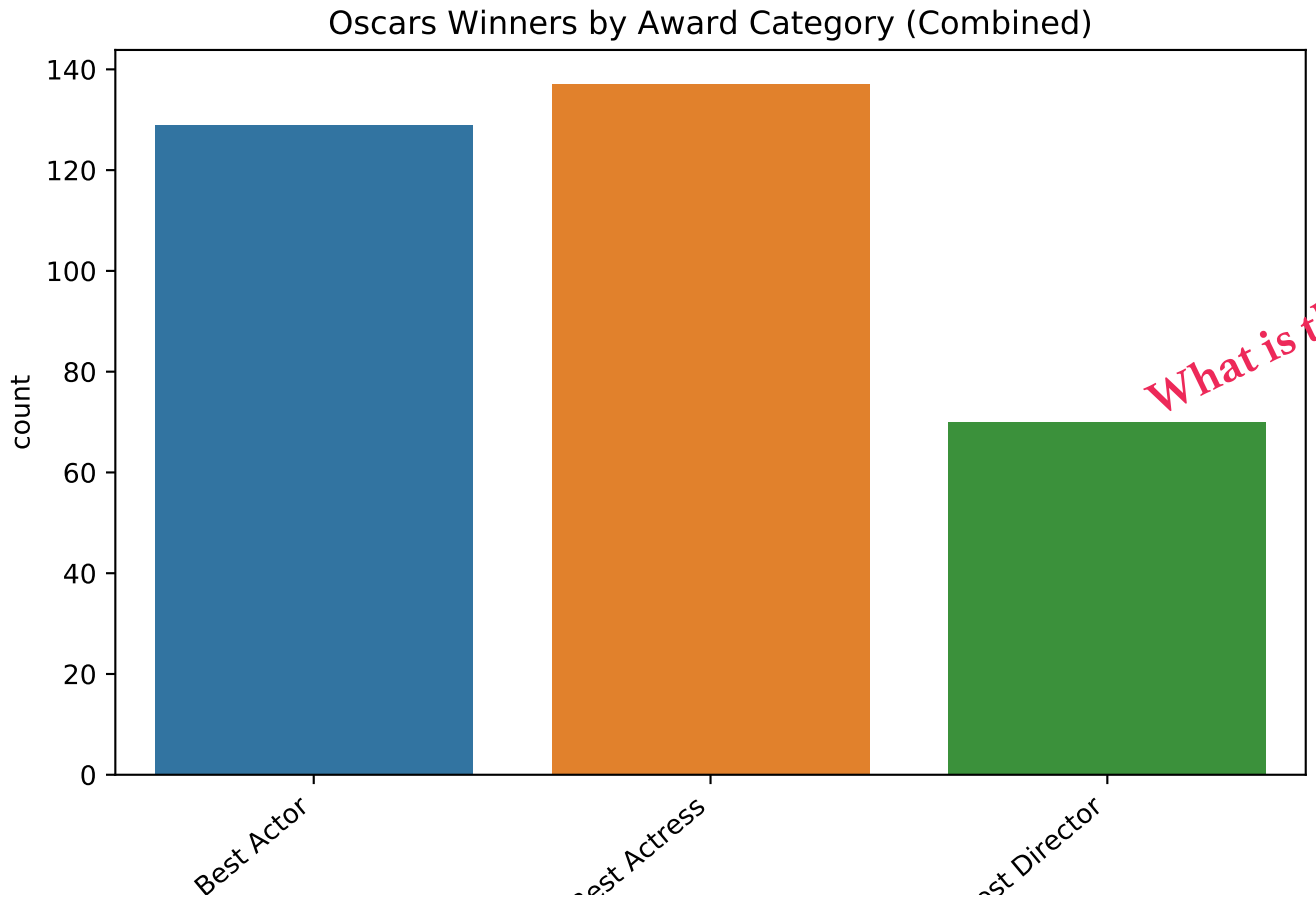
2. Oscars Winners by Award Category (Gender)

```
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
```



2. Oscars Winners by Award Category (Gender)



What is the sex ratio? 🤔

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?

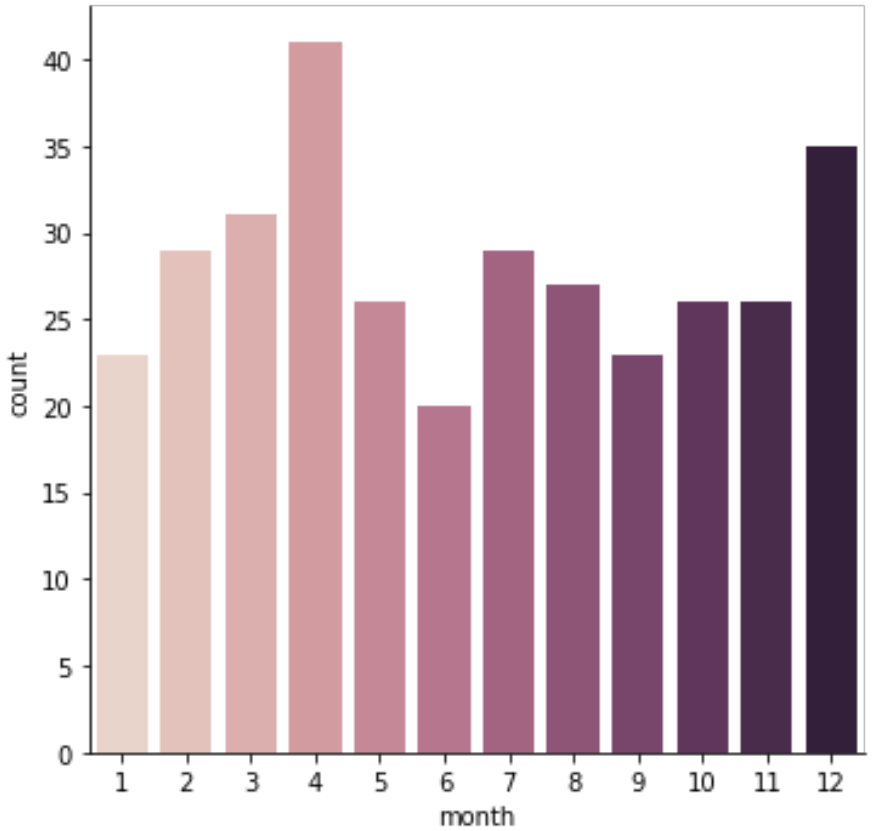
```
oscars['month'] = pd.DatetimeIndex(oscars['date_of_birth']).month
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```

.DatetimeIndex()

DATA ANALYSES

4. Which month were most Oscars winners born in?

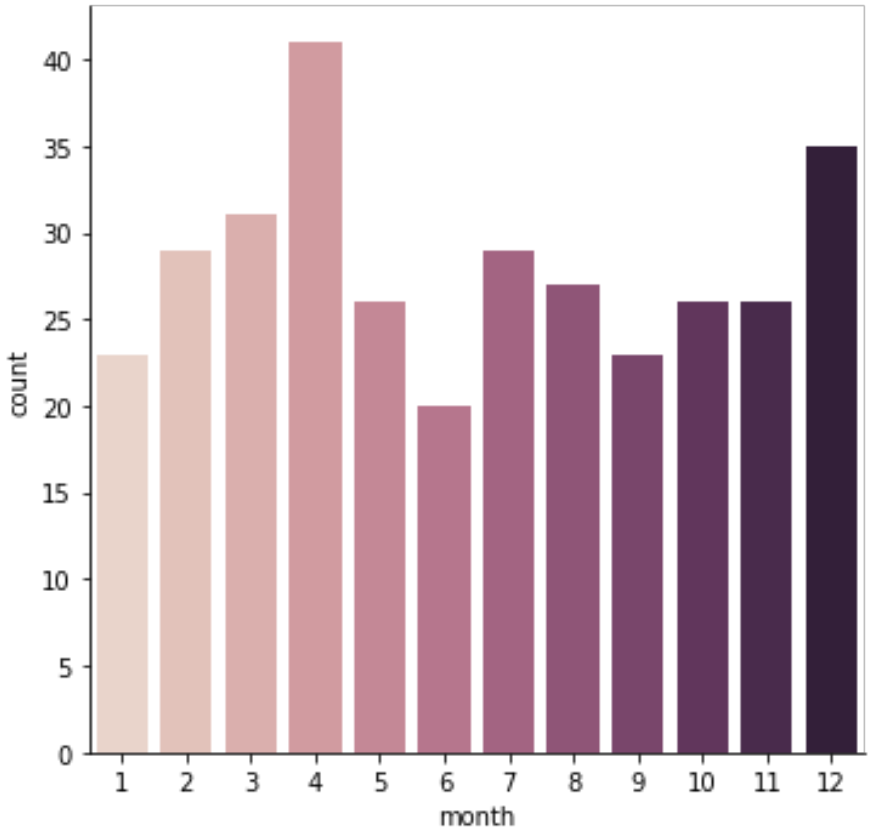
```
oscars['month'] = pd.DatetimeIndex(oscars['date_of_birth']).month
sns.catplot(x="month", kind='count', palette="ch:.10", data=oscars);
```



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);
```



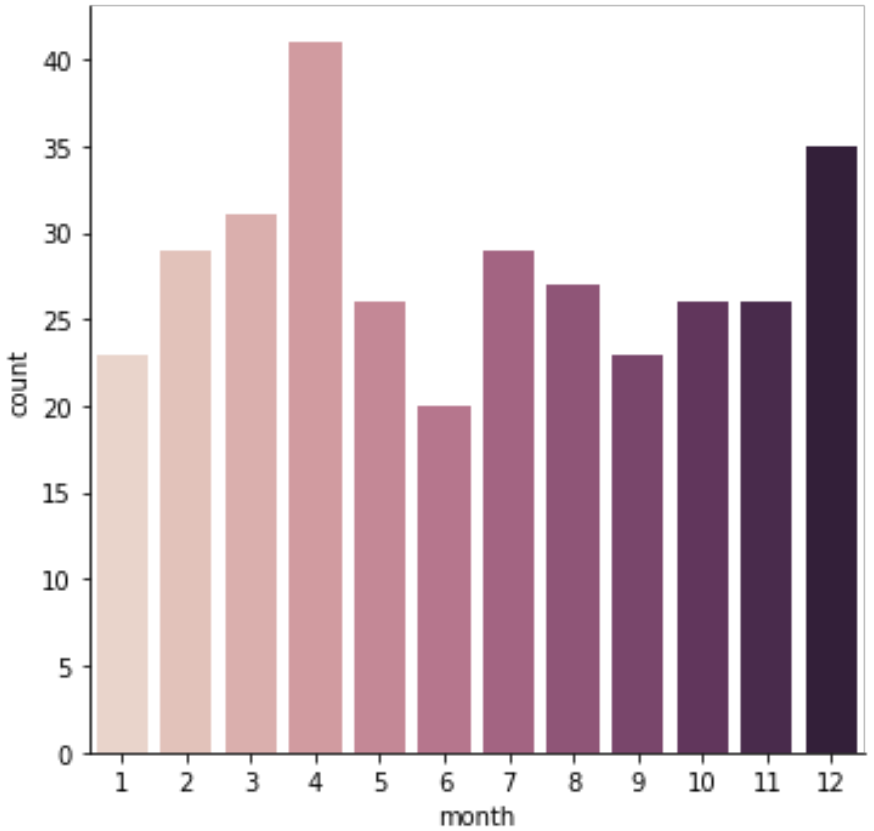
April

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

Pythom 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 !