UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

# Data Science & ML Course

## Lesson #9 Exploratory Data Analysis IV

Ivanovitch Silva

October, 2018

# Agenda

- Case study: titanic
- Visualizing missing values
- Aggregate data using pivot table
- Storytelling from Seaborn

# Update from repository

```
git clone https://github.com/ivanovitchm/datascience2machinelearning.git
```

Or ….

```
git pull
```

# Case Study: Titanic

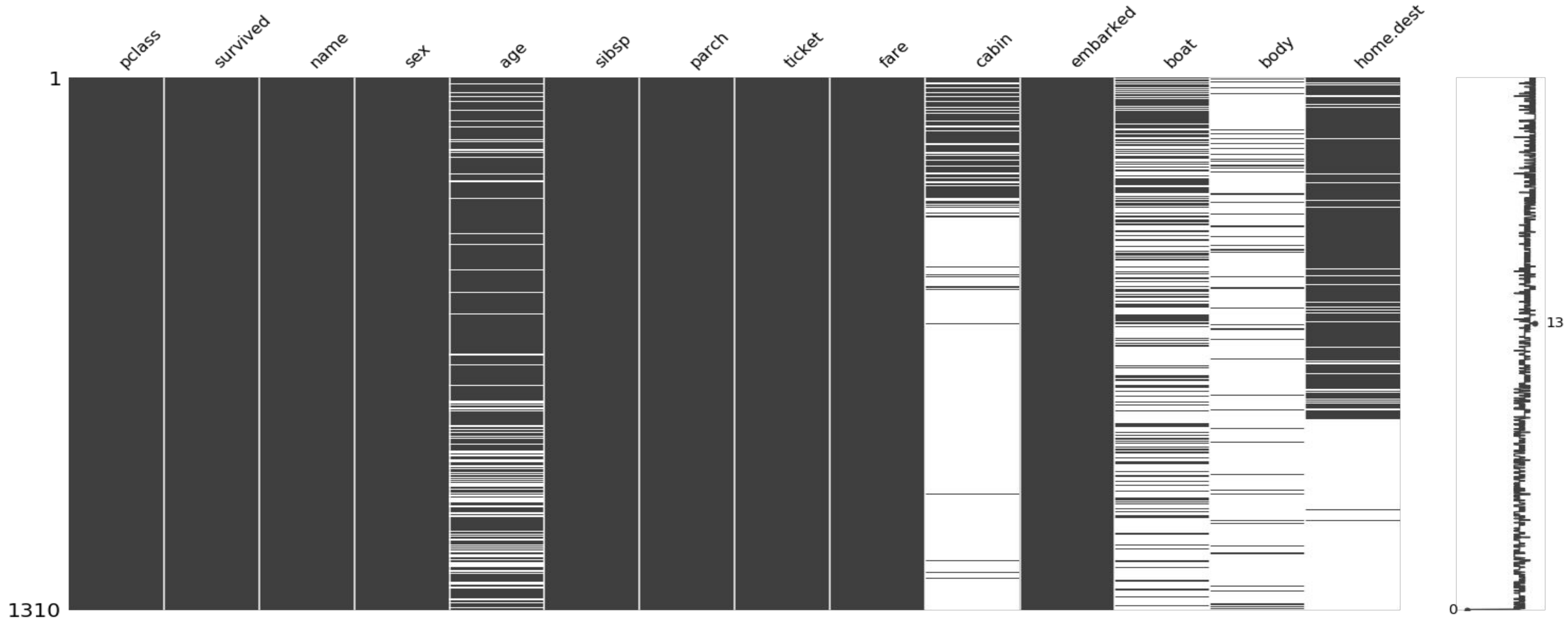
kaggle

https://www.kaggle.com/c/titanic

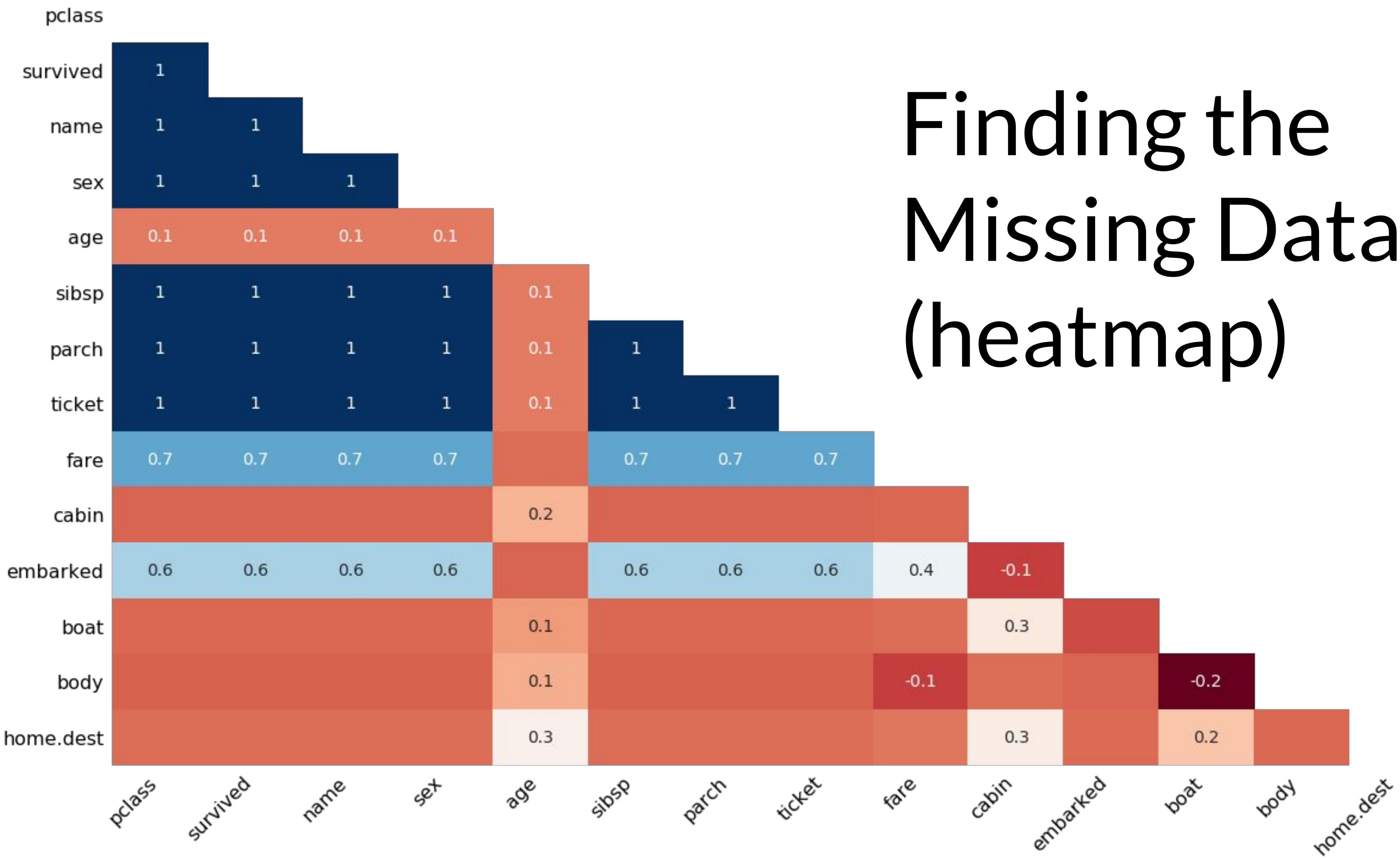
# Case Study: Titanic

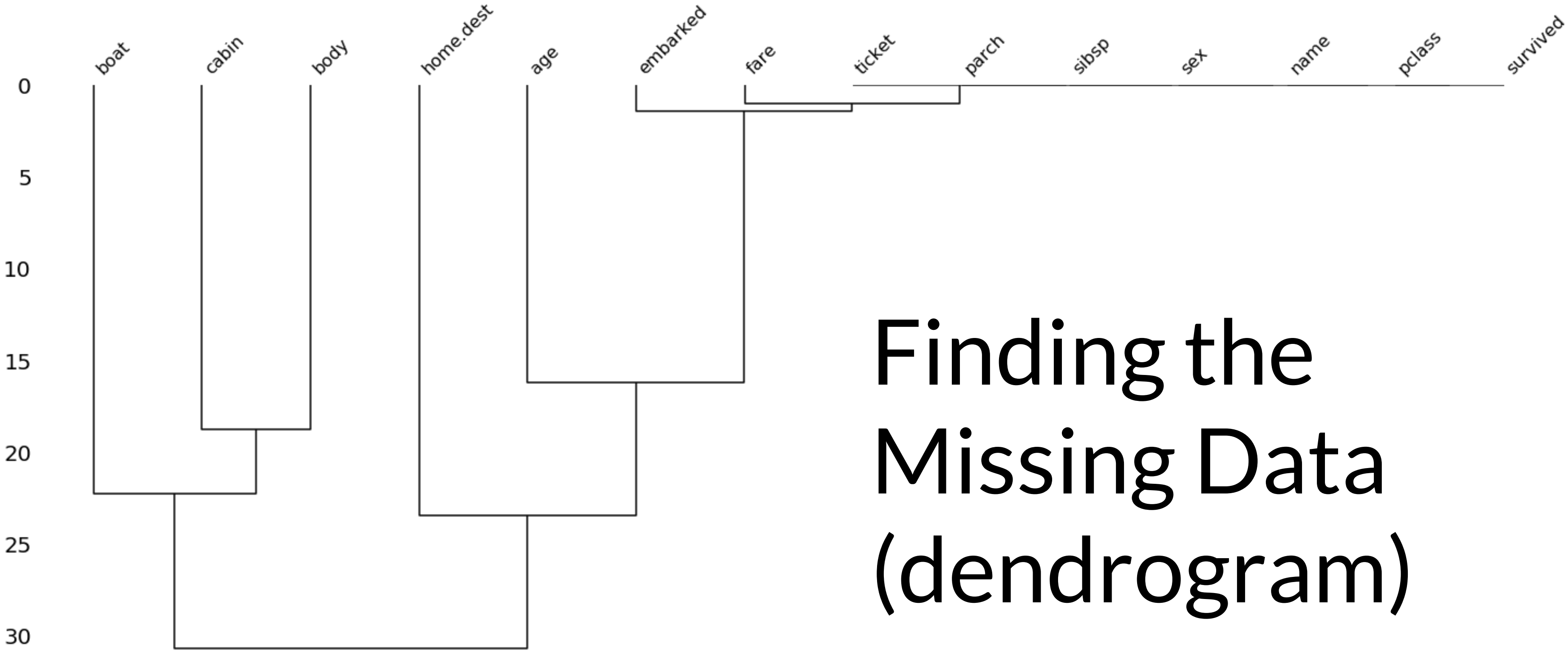
| | pclass | survived | name | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | body | home.dest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | Allen, Miss. Elisabeth Walton | female | 29.0000 | 0 | 0 | 24160 | 211.3375 | B5 | S | 2 | | St Louis, MO |
| 1 | 1 | 1 | Allison, Master. Hudson Trevor | male | 0.9167 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | 11 | | Montreal, PQ / Chesterville, ON |
| 2 | 1 | 0 | Allison, Miss. Helen Loraine | female | 2 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | | Montreal, PQ / Chesterville, ON |
| 3 | 1 | 0 | Allison, Mr. Hudson Joshua Creighton | male | 30.0000 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | 135 | Montreal, PQ / Chesterville, ON |
| 4 | 1 | 0 | Allison, Mrs. Hudson J C (Bessie Waldo Daniels) | female | 25 | 1 | 2 | 113781 | 151.5500 | C22 C26 | S | | | Montreal, PQ / Chesterville, |

# Finding the Missing Data (matrix)



# Finding the Missing Data (heatmap)






# Finding the Missing Data (dendrogram)

# missingno

https://github.com/ResidentMario/missingno

```
!conda install -c conda-forge missingno -y
```

# Calculating Summary Statistics

| | pclass | survived | name | sex | age | sibsp | parch | ticket | fare | cabin | embarked | boat | body | home.dest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1.0 | 1.0 | Allen, Miss. Elisabeth Walton | female | 29.0000 | 0.0 | 0.0 | 24160 | 211.3375 | B5 | S | 2 | NaN | St Louis, MO |
| **1** | 1.0 | 1.0 | Allison, Master. Hudson Trevor | male | 0.9167 | 1.0 | 2.0 | 113781 | 151.5500 | C22 C26 | S | 11 | NaN | Montreal, PQ / Chesterville, ON |
| **2** | 1.0 | 0.0 | Allison, Miss. Helen Loraine | female | 2.0000 | 1.0 | 2.0 | 113781 | 151.5500 | C22 C26 | S | NaN | NaN | Montreal, PQ / Chesterville, ON |

```
fares_by_class = {i:titanic_survival[titanic_survival.pclass == i].fare.mean()
                  for i in titanic_survival.pclass.unique()
                 }
```

```
{1.0: 87.50899164086687, 2.0: 21.1791963898917, 3.0: 13.302888700564957}
```

# Making Pivot Table

```
df_pivot = pd.pivot_table(titanic_survival,
                          index="pclass",
                          values=["fare","age"],
                          aggfunc=["mean"])
```

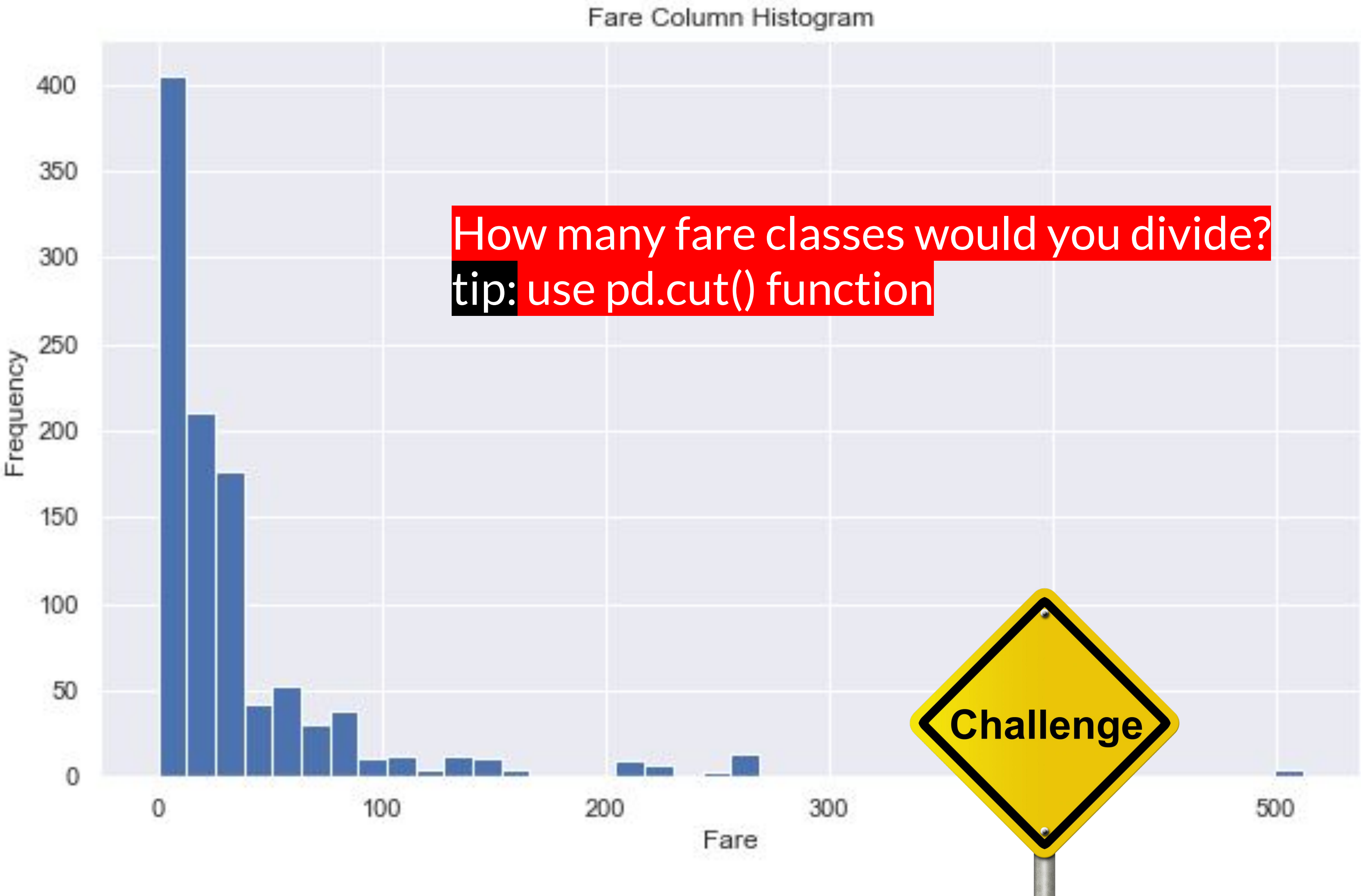
| | mean | |
|---|---|---|
| | age | fare |
| pclass | | |
| 1.0 | 39.159918 | 87.508992 |
| 2.0 | 29.506705 | 21.179196 |
| 3.0 | 24.745000 | 13.302889 |

```
df_pivot["mean"]["age"][1.0]
```

# Another way to aggregate information

```
titanic_survival["agecat"] = pd.cut(titanic_survival.age,
                         bins=[0,5,10,18,30,50,65,100],
                         labels=["Infant","Child","Teenager",
                                 "Young adult","Adult","Senior adult","Senior"])
```

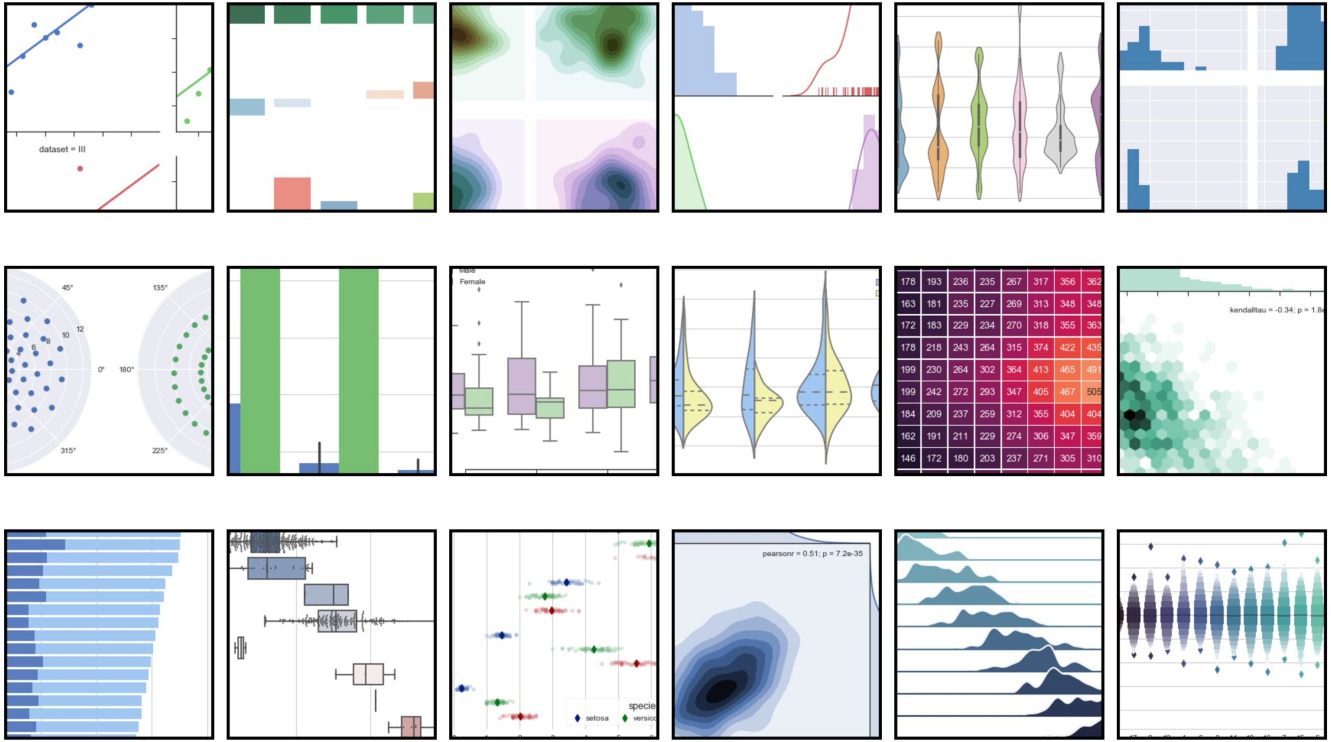
| | agecat | age |
|---|---|---|
| **0** | Young adult | 29.0000 |
| **1** | Infant | 0.9167 |
| **2** | Infant | 2.0000 |
| **3** | Young adult | 30.0000 |
| **4** | Young adult | 25.0000 |
| **5** | Adult | 48.0000 |




How many fare classes would you divide?
tip: use pd.cut() function

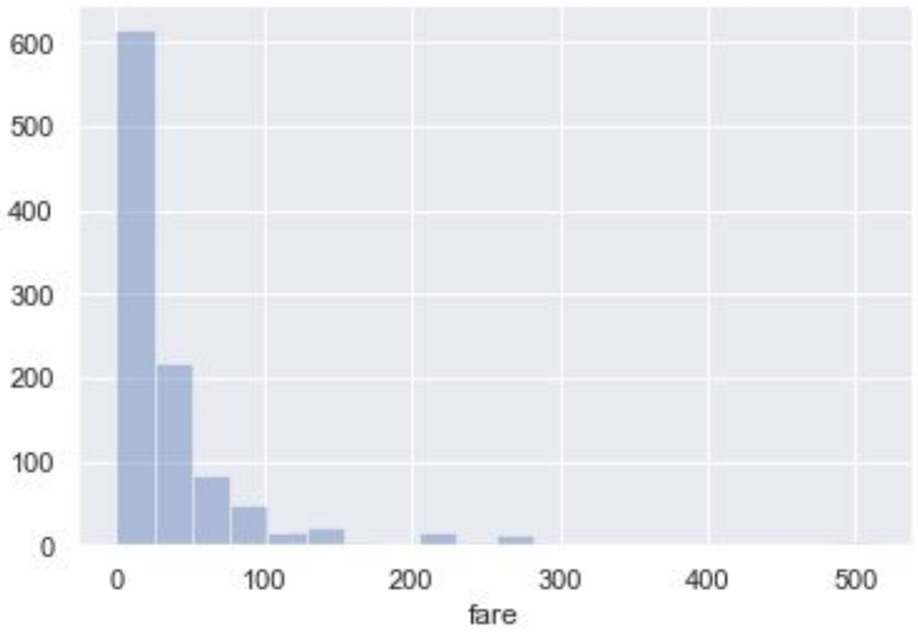
# Lesson #9 - Exploratory Data Analysis III.ipynb
## Section 1

# Storytelling from Seaborn

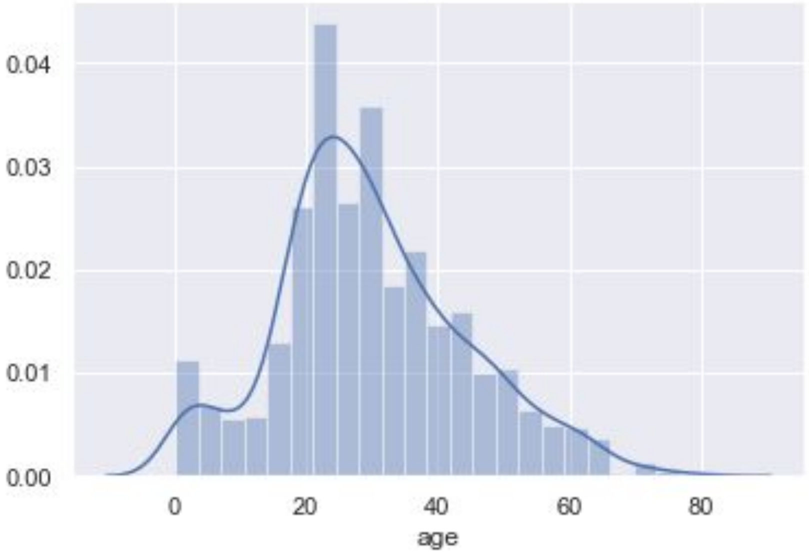


# Creating histogram in seaborn

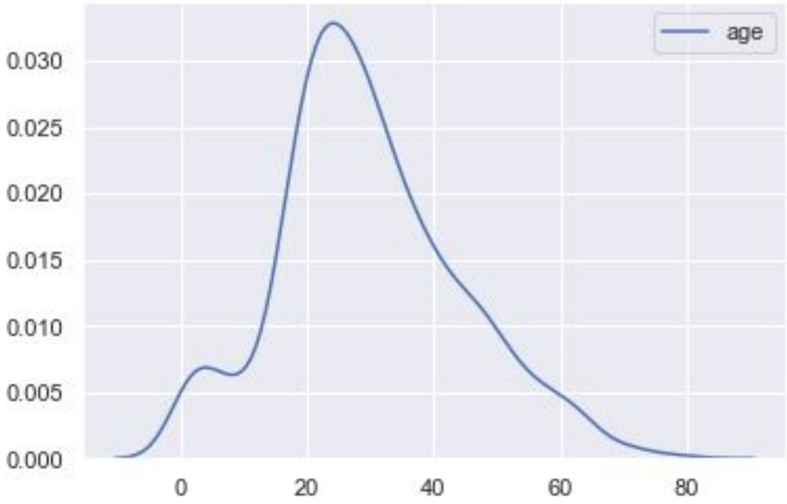
```
# seaborn is commonly imported as `sns`.
import matplotlib.pyplot as plt
import seaborn as sns

titanic.dropna(subset=["fare"],inplace=True)

#call seaborn default colors
sns.set()

# Contexts: paper, notebook, talk, and poster
sns.set_context("notebook")

# plot a univariate distribution of observations.
sns.distplot(titanic["fare"],kde=False,
             bins=20)
```



# Generating a kernel density plot

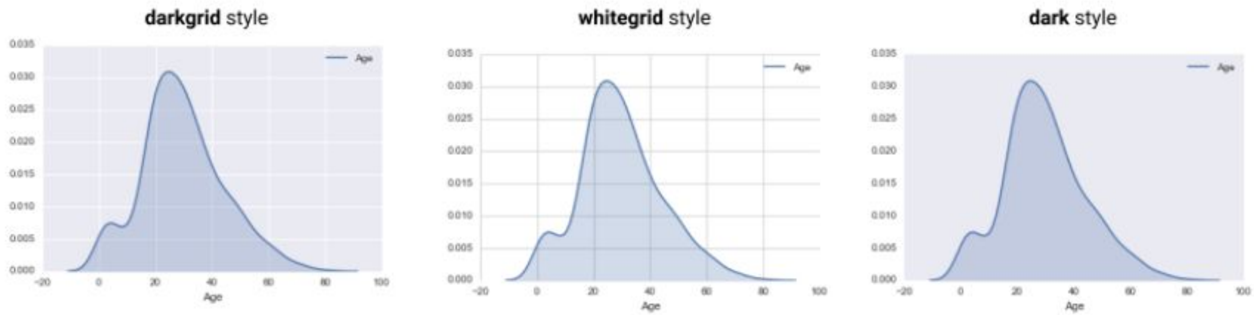



```
sns.distplot(titanic["age"])
```

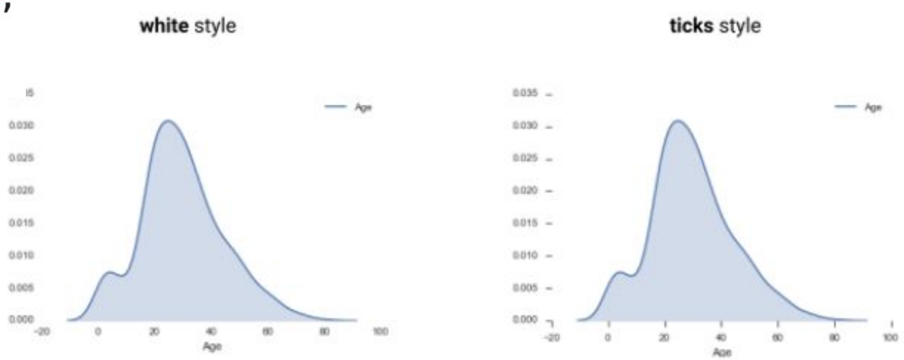



```
sns.kdeplot(titanic["age"])
```

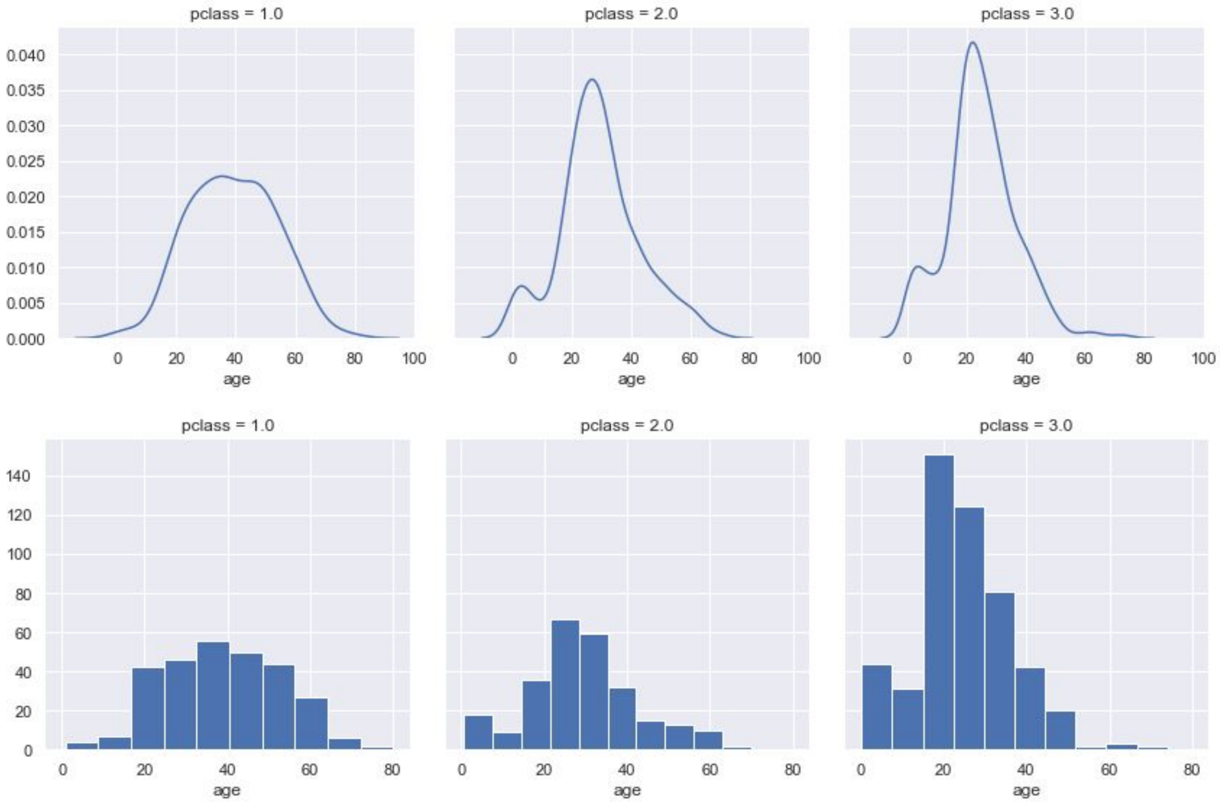
# Modifying the appearance of plots




```python
sns.set_style("ticks",
              {'xtick.bottom': False,
               'ytick.left':False}
             )
```

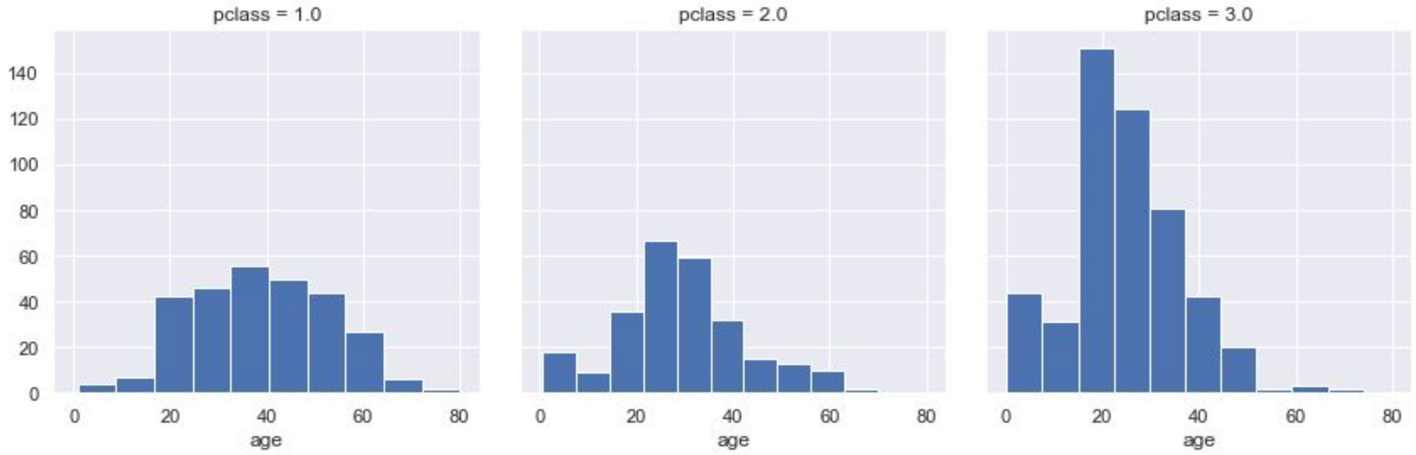


# Conditional distributions



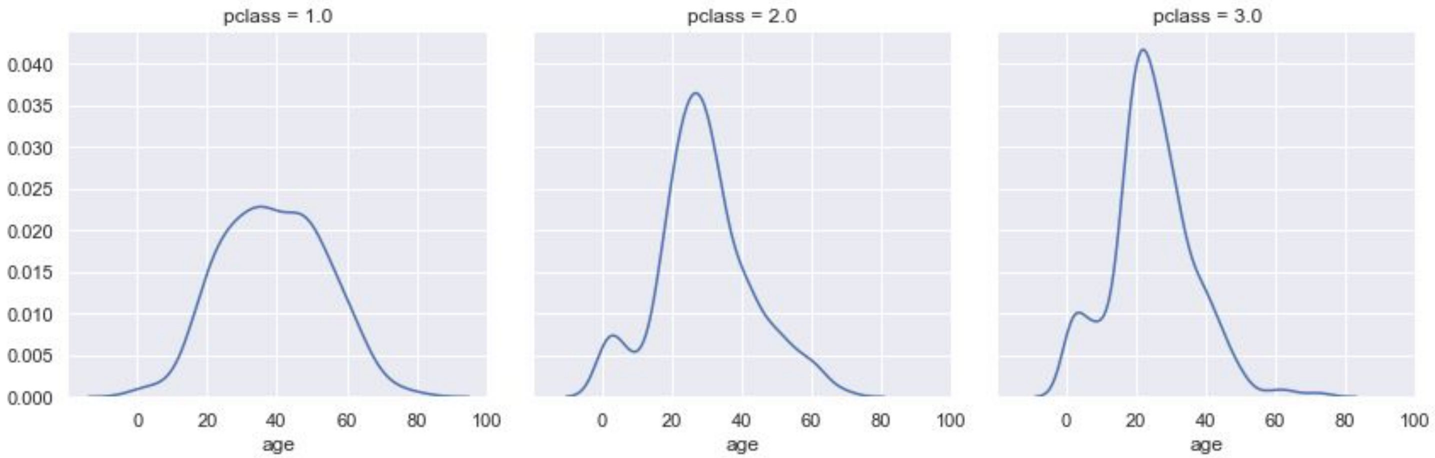
# Conditional distributions

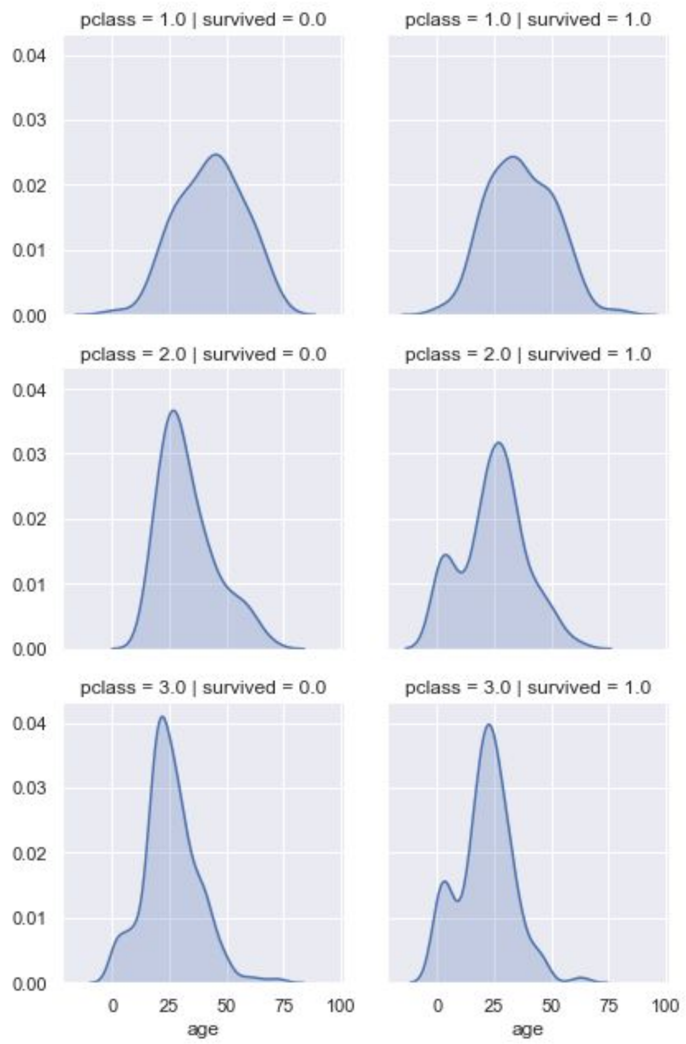
```python
# Condition on unique values of the "survived" column.
g = sns.FacetGrid(titanic, col="pclass", height=4)

# Generate a KDE plot to "age" column.
g.map(plt.hist, "age")
```



# Conditional distributions

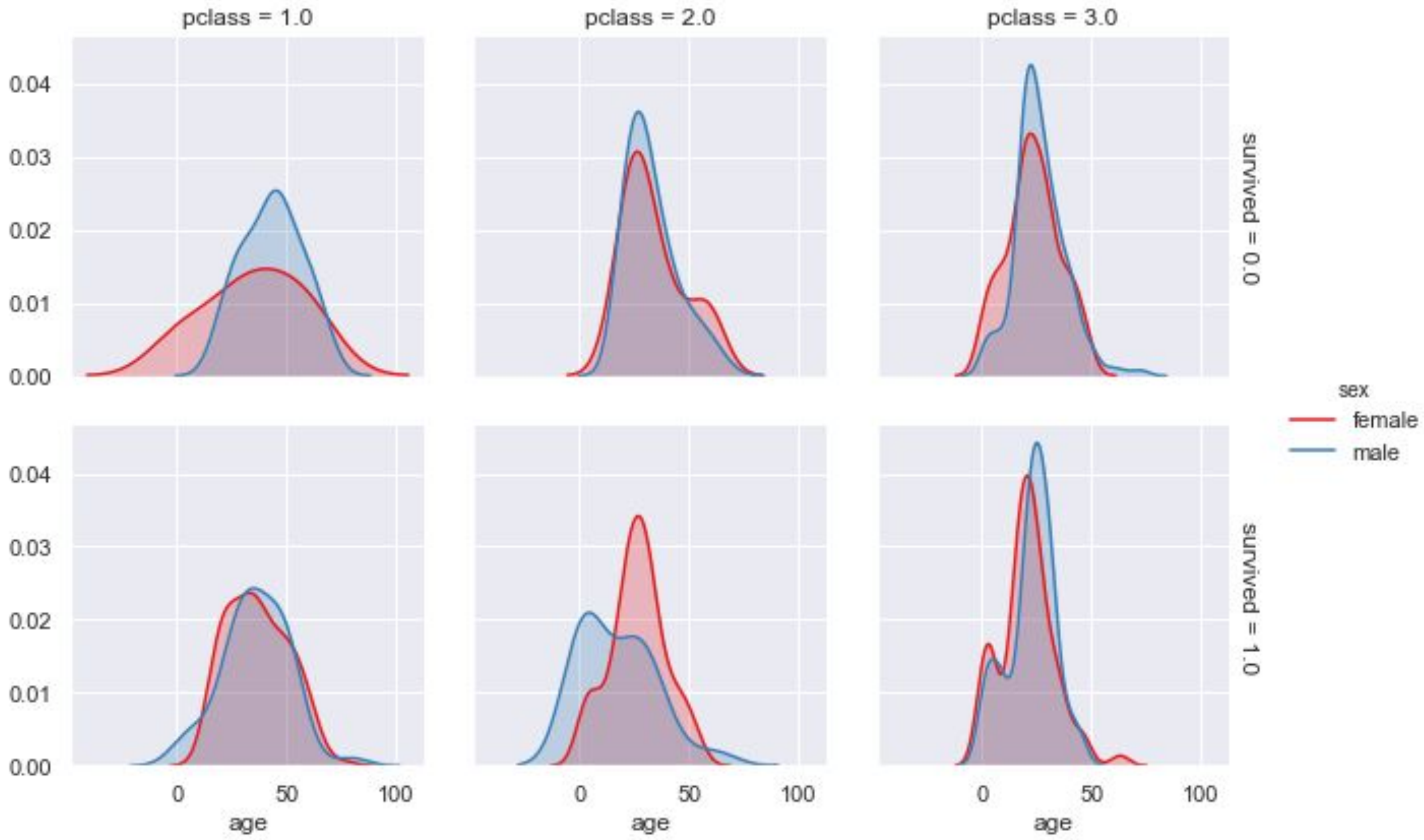
```
# Condition on unique values of the "survived" column.
g = sns.FacetGrid(titanic, col="pclass", height=4)

# Generate a KDE plot to "age" column.
g.map(sns.kdeplot, "age", shade=False)
```



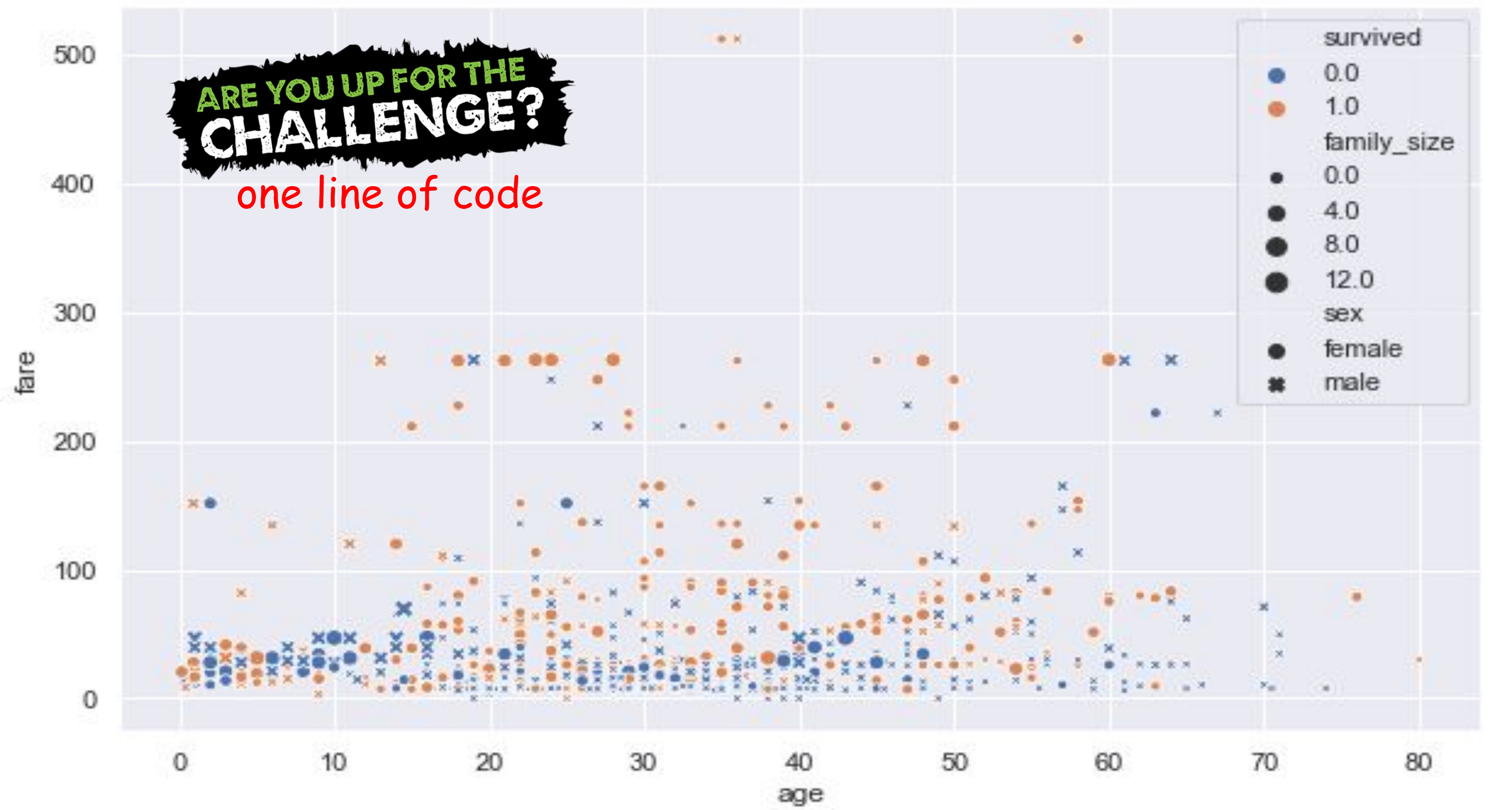



# Creating conditional plots using two conditions

```
g = sns.FacetGrid(titanic, col="survived",
                           row="pclass")
g.map(sns.kdeplot, "age", shade=True)
```

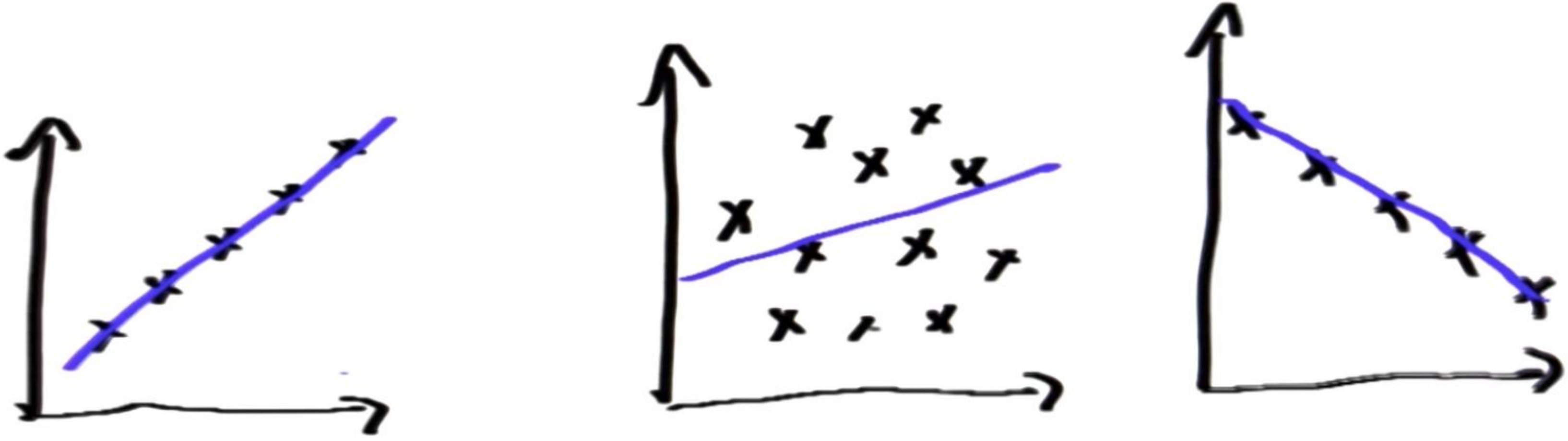



```
g = sns.FacetGrid(titanic, col="pclass", row="survived",hue="sex",palette="Set1",margin_titles=True)
g.map(sns.kdeplot, "age", shade=True).add_legend()
```


ARE YOU UP FOR THE
CHALLENGE?
one line of code
survived
0.0
1.0
family_size
0.0
4.0
8.0
12.0
sex
female
male
fare
age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80

# Measuring Correlation
# Pearson R

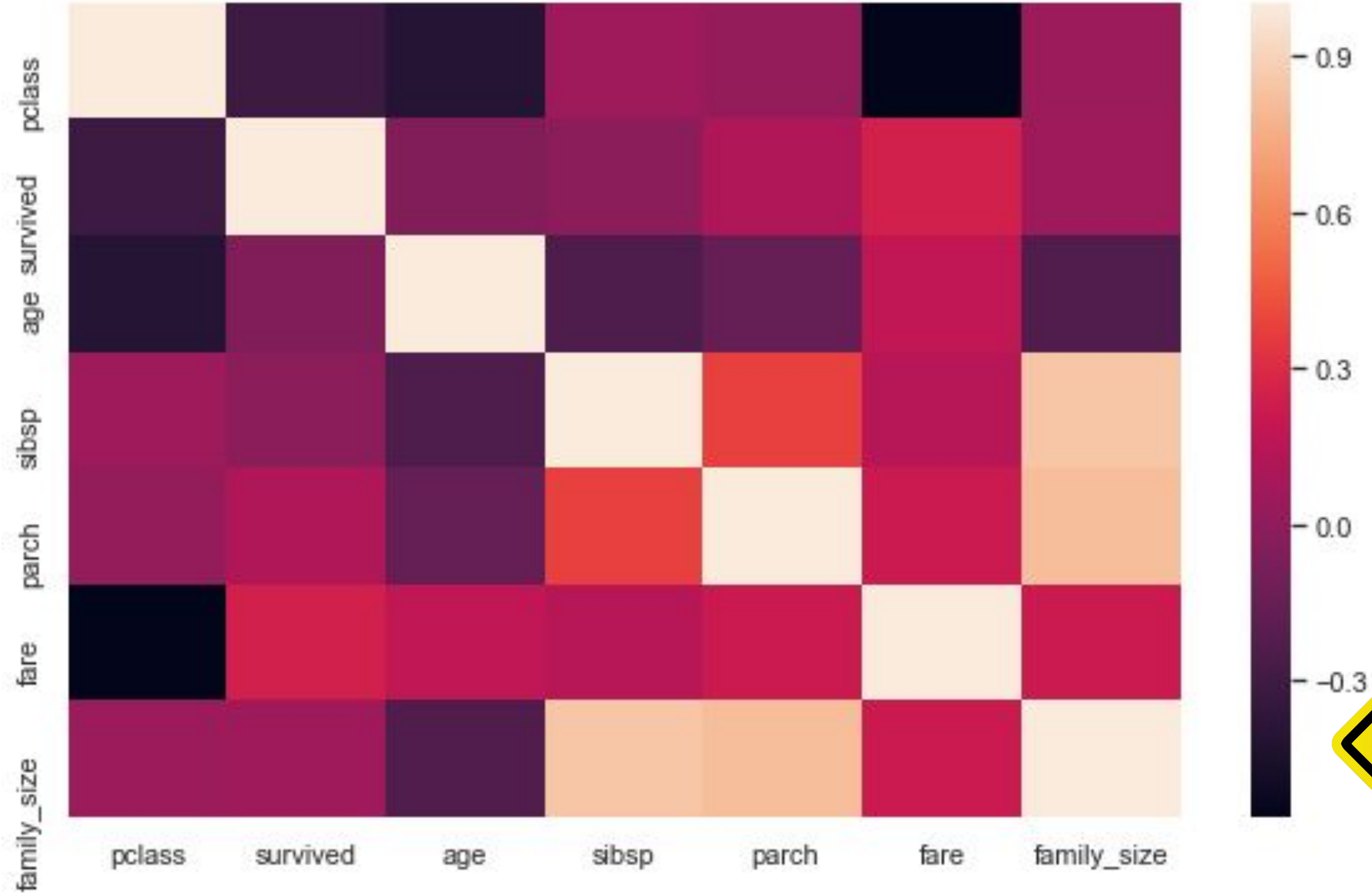
$\in [-1 \dots 1]$



$r = 1$     $0$     $-1$

```
titanic.corr()
```

| | pclass | survived | age | sibsp | parch | fare | family_size |
|---|---|---|---|---|---|---|---|
| **pclass** | 1.000000 | -0.319979 | -0.411086 | 0.047746 | 0.017685 | -0.565255 | 0.040200 |
| **survived** | -0.319979 | 1.000000 | -0.053958 | -0.012657 | 0.114091 | 0.249164 | 0.058001 |
| **age** | -0.411086 | -0.053958 | 1.000000 | -0.243139 | -0.150241 | 0.178739 | -0.239501 |
| **sibsp** | 0.047746 | -0.012657 | -0.243139 | 1.000000 | 0.374291 | 0.141184 | 0.844210 |
| **parch** | 0.017685 | 0.114091 | -0.150241 | 0.374291 | 1.000000 | 0.216723 | 0.813030 |
| **fare** | -0.565255 | 0.249164 | 0.178739 | 0.141184 | 0.216723 | 1.000000 | 0.213916 |
| **family_size** | 0.040200 | 0.058001 | -0.239501 | 0.844210 | 0.813030 | 0.213916 | 1.000000 |


pclass
survived
age
sibsp
parch
fare
family_size
0.9
0.6
0.3
0.0
−0.3


Challenge