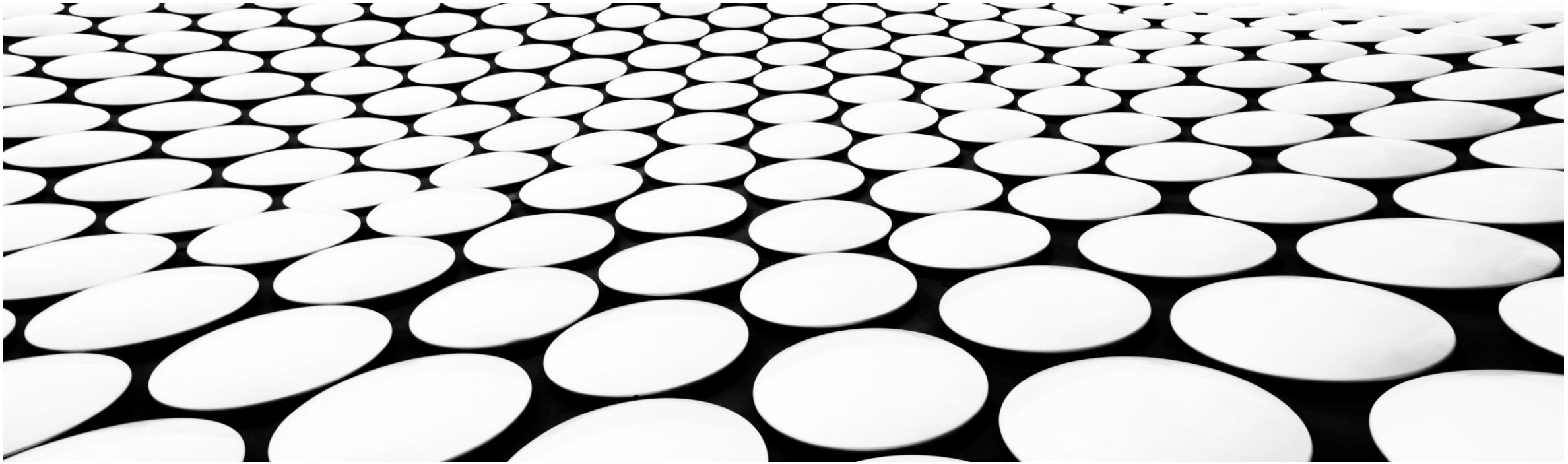
EXPLORATORY DATA ANALYSIS

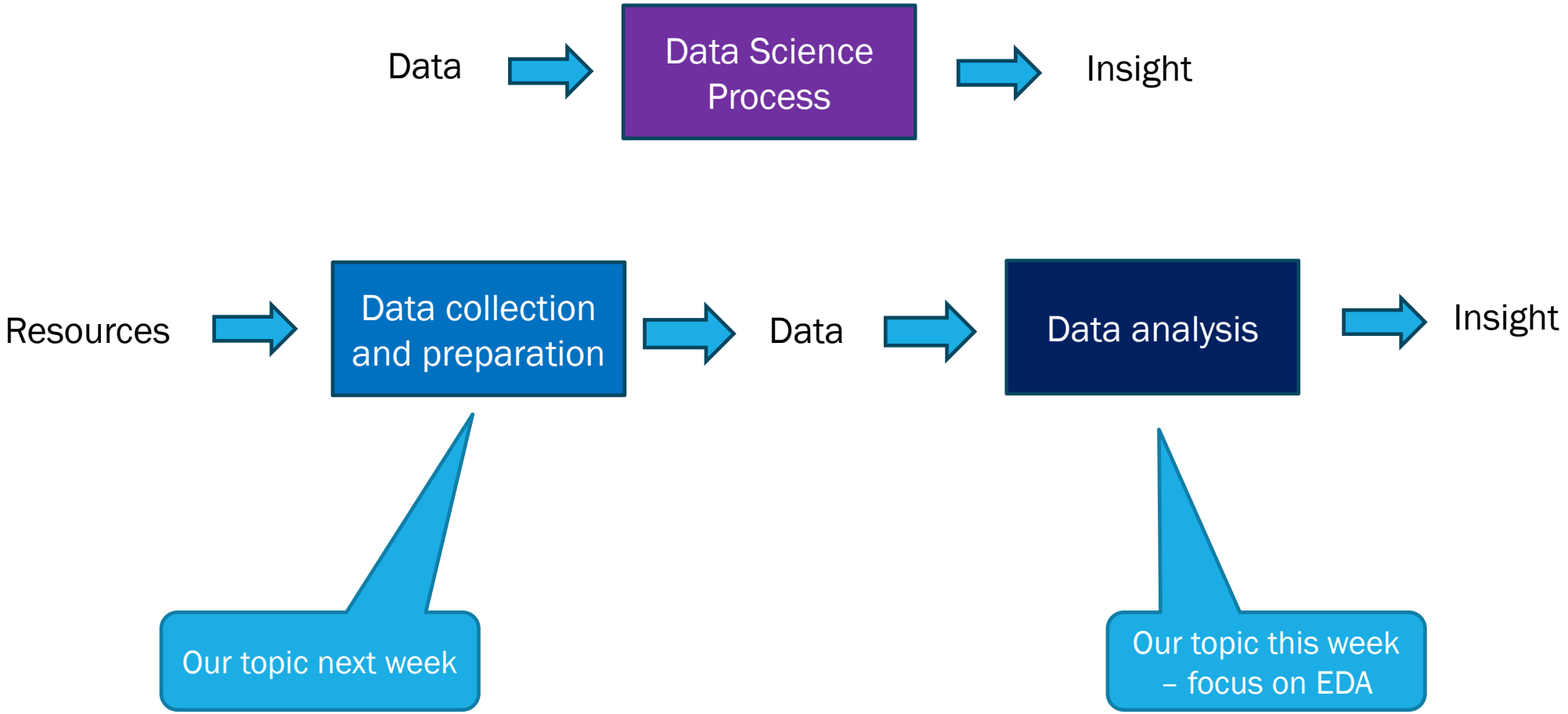
Story behind the data: Titanic dataset



EXPLORATORY DATA ANALYSIS

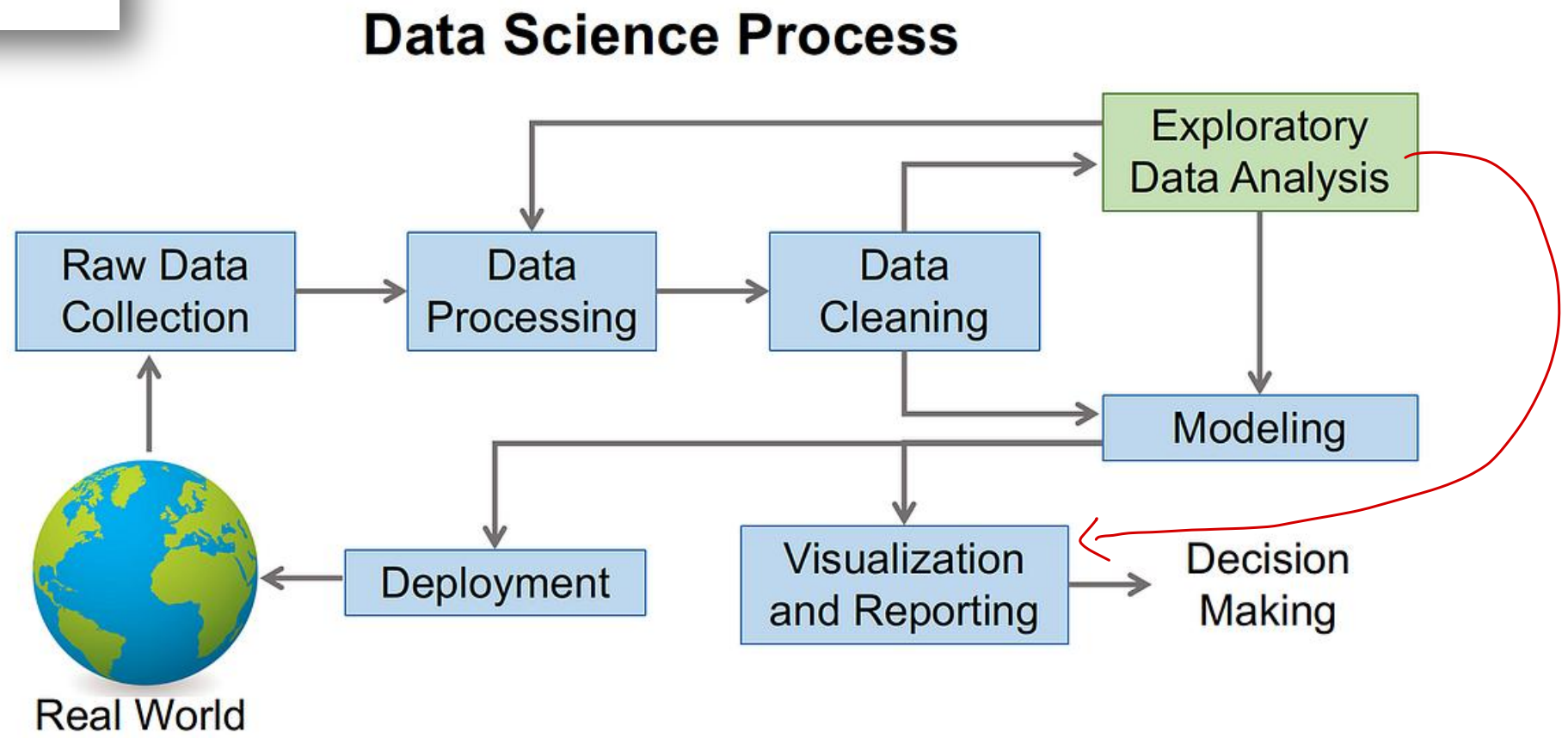
- Descriptive statistics
- Diagnostic statistics
- Case Study 1 – Titanic

Friday Case Study 2 +
Examples for Assignment 1

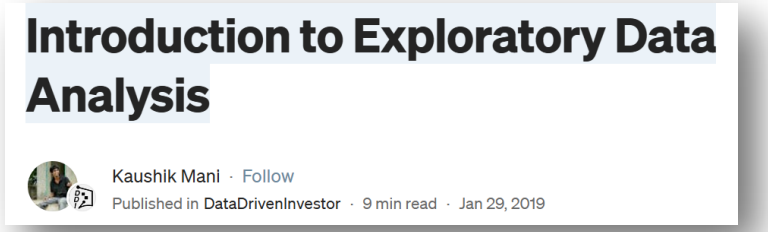

Data
Data Science Process
Insight
Resources
Data collection and preparation
Data
Data analysis
Insight
Our topic next week
Our topic this week – focus on EDA

Data Science Process: A Comprehensive Guide

Abhijit · Follow
6 min read · Jan 15, 2024



Source



Introduction to Exploratory Data Analysis

Kaushik Mani · Follow
Published in DataDrivenInvestor · 9 min read · Jan 29, 2019

… one of the most important components to any data science experiment that doesn't get as much importance as it should is Exploratory Data Analysis (EDA).

In short, EDA is "A first look at the data". It is a critical step in analyzing the data from an experiment. It is used to understand and summarize the content of the dataset to ensure that the features which we feed to our machine learning algorithms are refined and we get valid, correctly interpreted results.

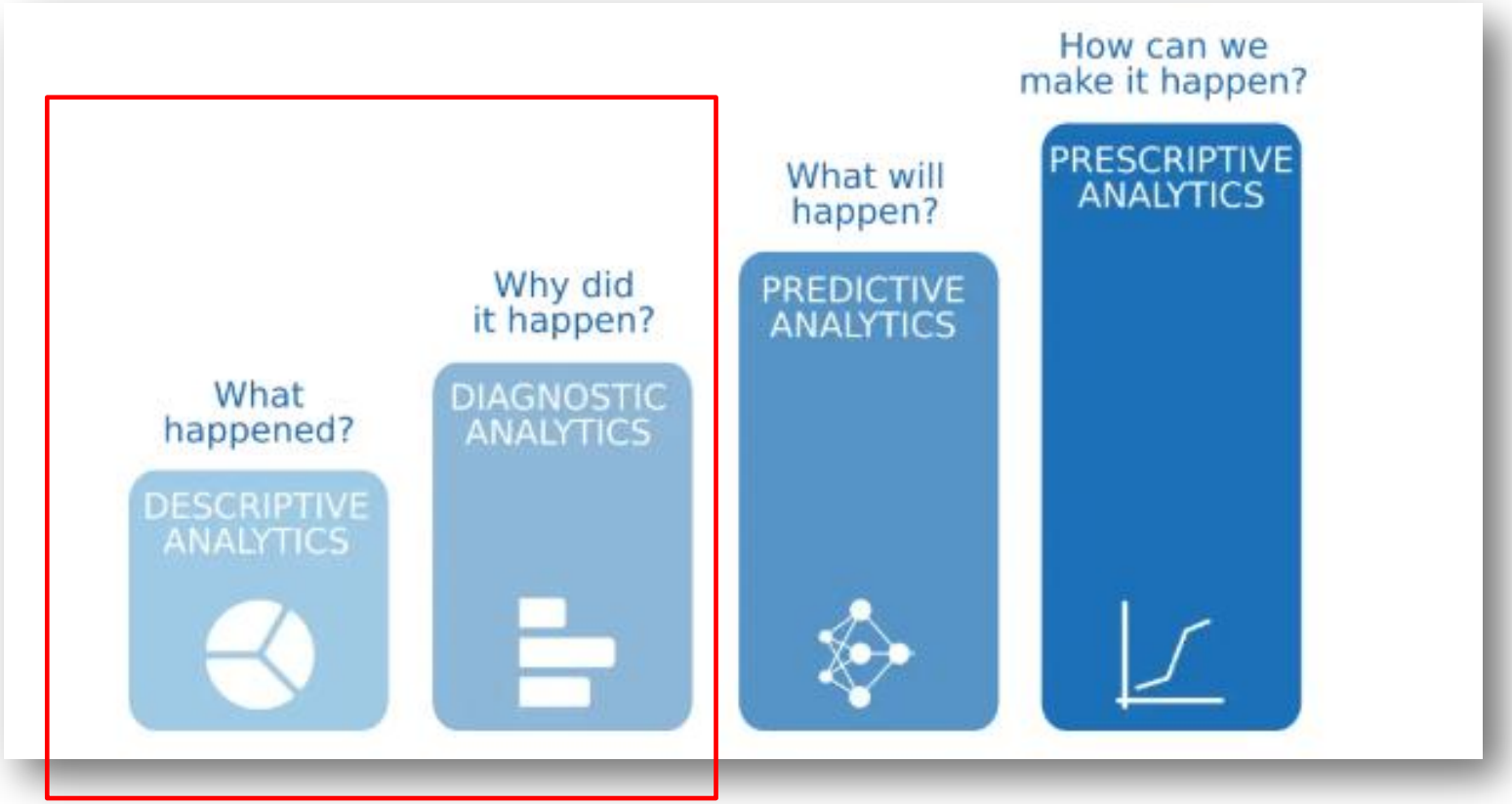
Or that we present valid descriptive insights in reports

Source

What is advanced analytics?

Tutorials

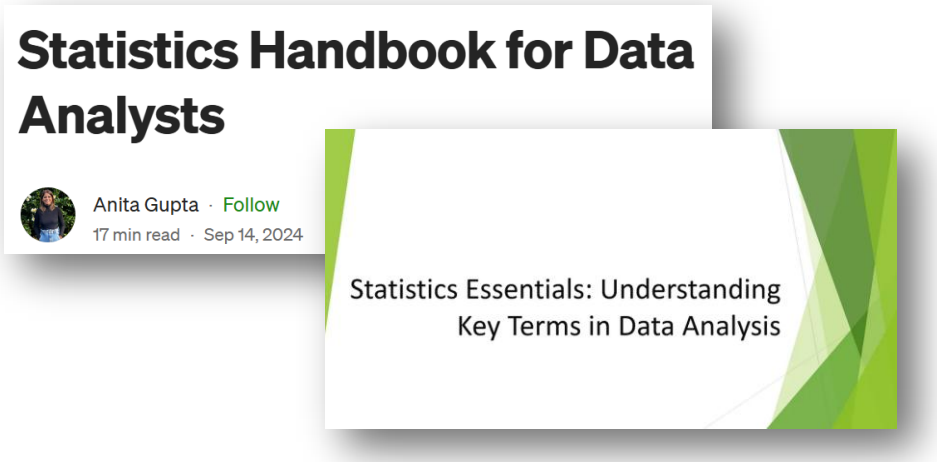
Pablo Martín August 31, 2023



neuraldesigner

Source

Descriptive Statistics



Measures of central tendency
1. Mean
2. Median
3. Mode

Measures of dispersion
1. Range
2. Variance
3. Standard deviation

Source

Mean: The average value.

$$\text{Mean}(\mu) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example

The average salary across United States is $60,000.

Median: The middle value in a sorted dataset. If the dataset contains an even number of values, the median is the average of the two middle values.

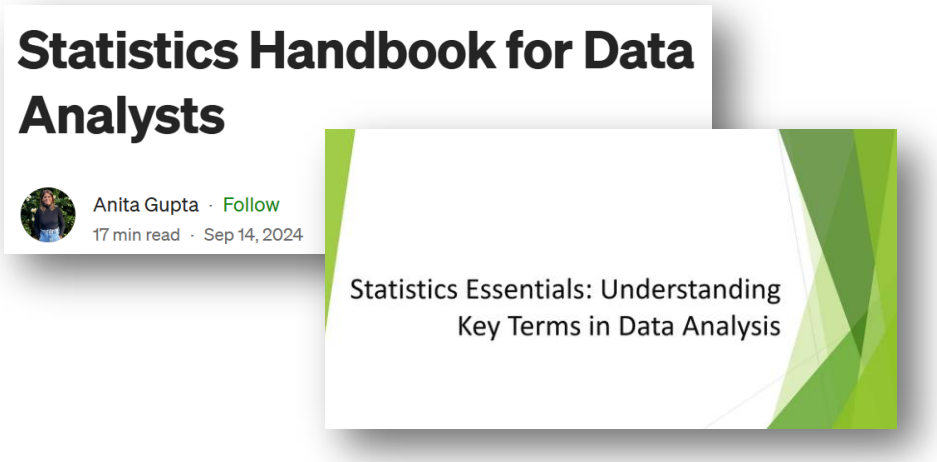
Example

The median divides the data points: 50% are above it, and 50% are below it. So next time someone says the median salary is $50,000, it means 50% of people earn below that and 50% earn above it.

Mode: The most frequently occurring value.

Example

If you do a poll and ask what food students want to order for lunch, and the majority vote for pizza, then pizza is the mode — the highest occurring value.



Measures of central tendency

1. Mean
2. Median
3. Mode

Measures of dispersion

1. Range
2. Variance
3. Standard deviation

Source

Range is from the minimum to the maximum value.

Example

If you're looking to buy a house, you might be considering houses in the price range of $500,000 to $800,000. We use ranges all the time in day-to-day conversations.

Variance is **a measure of dispersion**, meaning it is a measure of how far a set of numbers is spread out from their average value.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

S^2 = Variance
n = The Number of data Point
X_i = Each of the values of the data
$\bar{X}$ = The Mean of X_i

Equation for a sample

Standard deviation is the square root of the variance.

This allows the dispersion metric to be in the same unit as the observed data. For example, we now have dollars instead of dollars squared, making interpretation more useful.

$$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

n = Number of Observations
X_i = Value of the one Observation
$\bar{X}$ = Mean Value of Observations

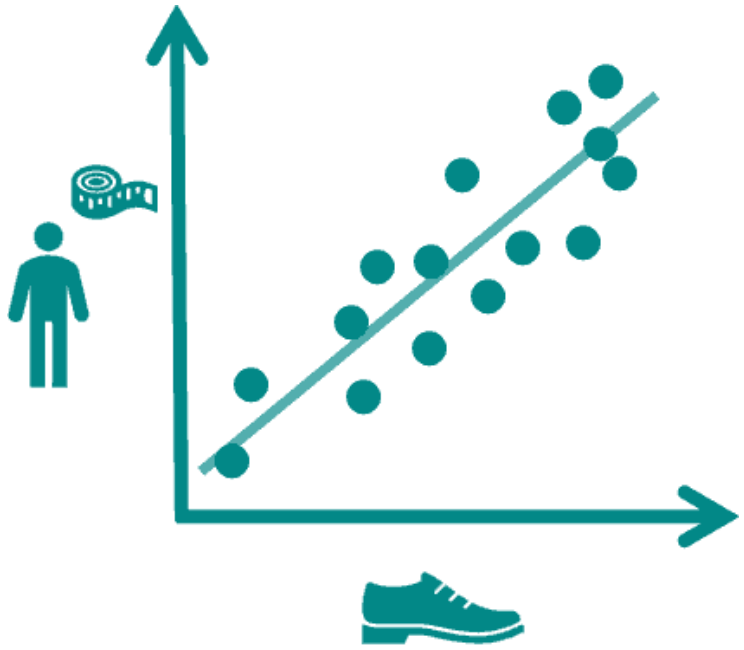
Equation for a sample

Diagnostic statistics

DATAtab

Correlation analysis

Correlation analysis is a statistical method used to evaluate the relationship between two variables, such as the association between body size and shoe size.



Source

DATAtab

Pearson correlation analysis

With the Pearson correlation analysis you get a statement about the linear correlation between metric scaled variables. The respective **covariance** is used for the calculation.

With the help of correlation analysis two statements can be made:

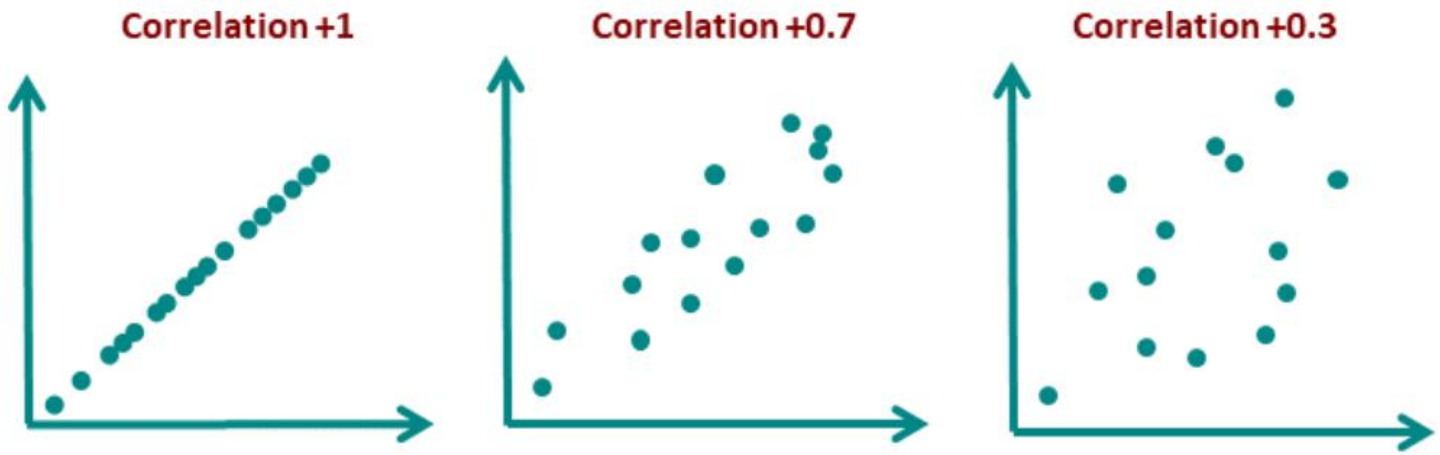
- one about the direction
- and one about the strength

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Source

DATAtab



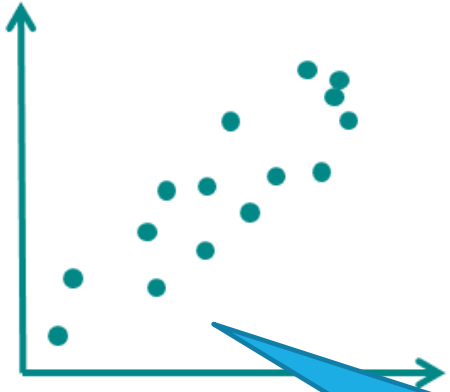
$\|r\|$	Strength of correlation
0.0 < 0.1	no correlation
0.1 < 0.3	little correlation
0.3 < 0.5	medium correlation
0.5 < 0.7	high correlation
0.7 < 1	very high correlation

Source

Requirements for causality

1

There is a significant correlation



2

Timely

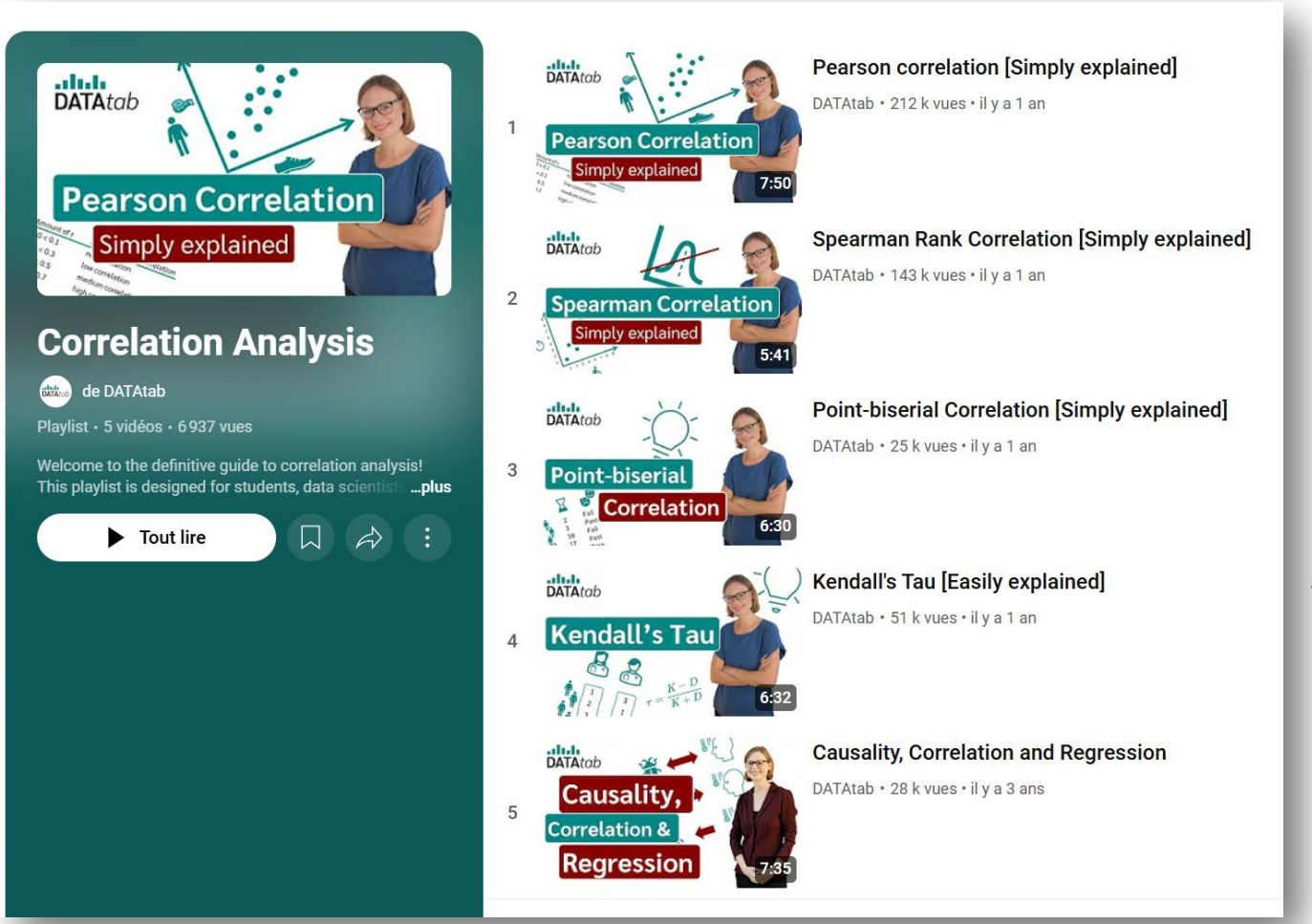
Variable A surveyed

Variable B surveyed

Theory

Profound theory

If we don't even have correlation… for sure we can affirm there is NO causality



Choice varies based on conditions (parametric: metric data normally distributed)

Source

Case Study 1
Titanic Dataset

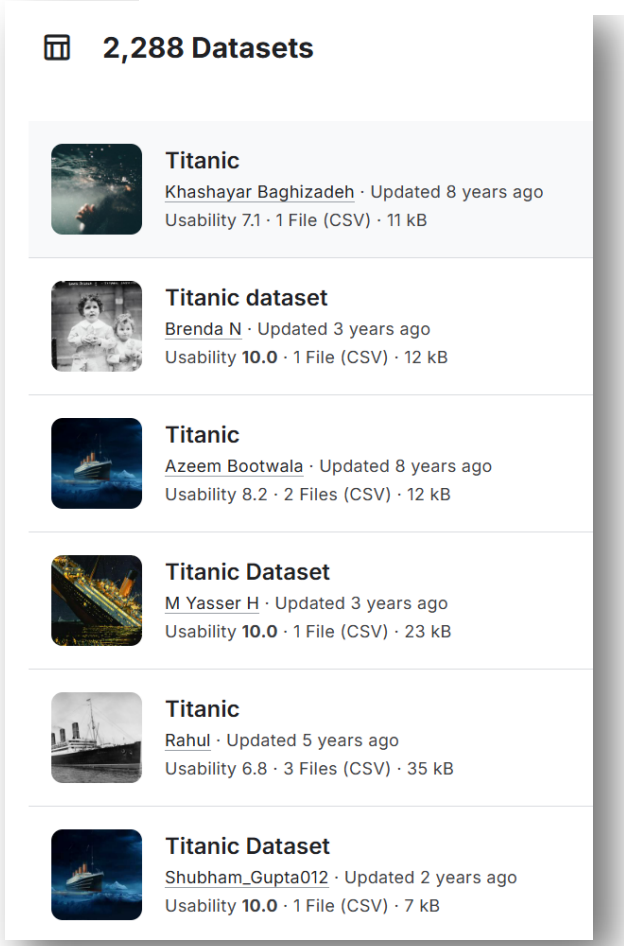
```
# Import libraries for visualization
import seaborn as sns
import matplotlib.pyplot as plt

# Loading the dataset
titanic = sns.load_dataset('titanic')
```

```
# Number of samples, Number of attributes
titanic.shape
```

```
(891, 15)
```

Dataset available within seaborn

kaggle



Lots of other titanic datasets!!

```
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
```

A first view of attributes and data types

Siblings (sibsp)

- Brother, sister
- Step-brother, step-sister
- Husband, wife

Parents/Children (parch)

- Mother, father
- Son, daughter
- Step-children

```
titanic.head(10)
```

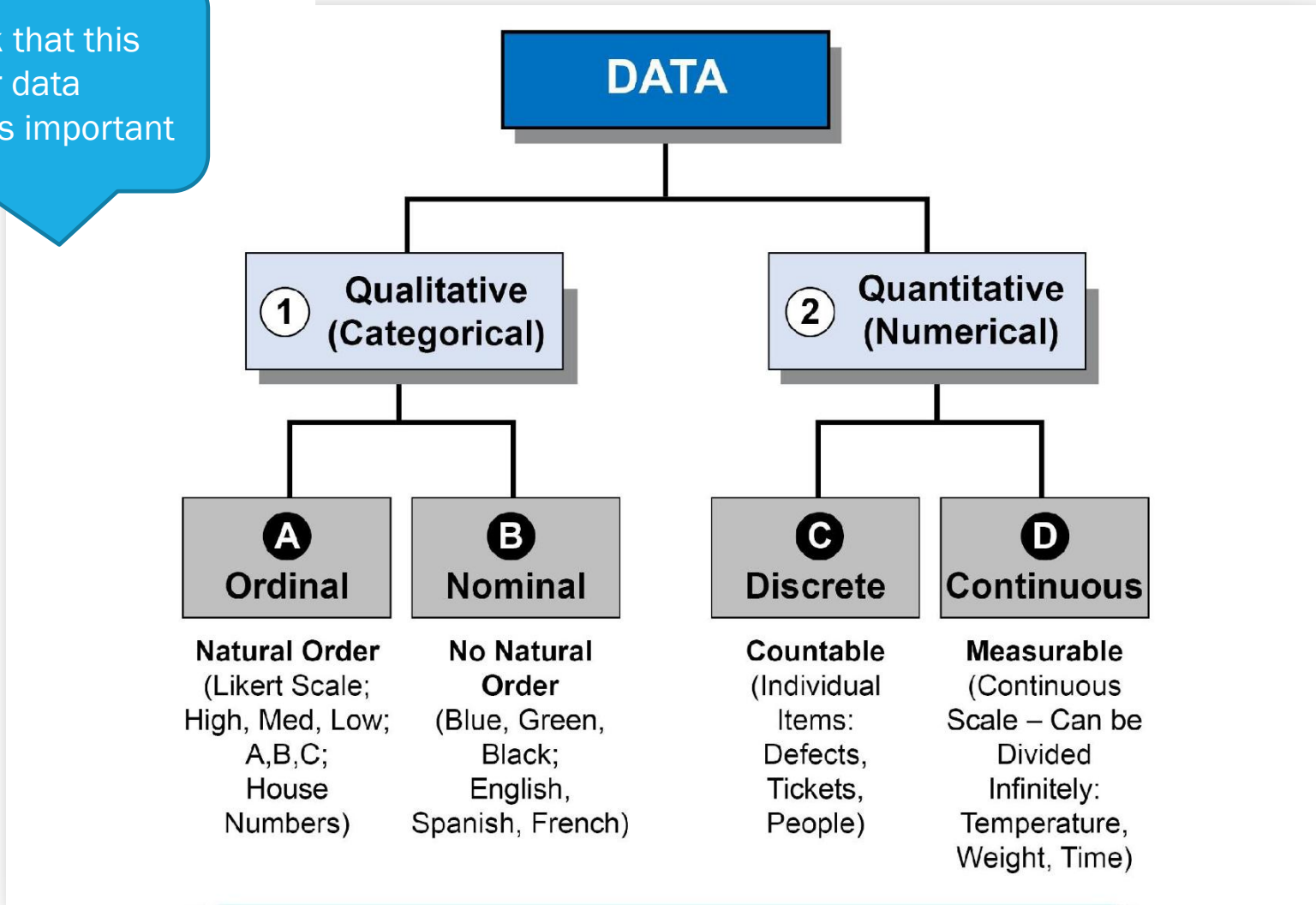
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

Measurement	Format	Volume
Quantitative (Numerical) Qualitative (Categorical)	Structured Unstructured Semi-Structured	Big Data Small Data

Collection	Source	Time
Primary Data Secondary Data	Internal Data External Data	Historical Data Real-Time Data

How would we characterize our current dataset?

We saw last week that this dimension for data characterization was important



Source

```
titanic.head(10)
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

Feature analysis

```
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
```

A view at missing data also… we'll do data cleaning later in the semester

A view at perhaps redundant attributes?

Importance of self-explanatory labels when documentation is lost…

Descriptive statistics

```
titanic.describe()
```

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Central tendency

Dispersion

```
titanic.describe(include=["object", "category"])
```

	sex	embarked	class	who	deck	embark_town	alive
count	891	889	891	891	203	889	891
unique	2	3	3	3	7	3	2
top	male	S	Third	man	C	Southampton	no
freq	577	644	491	537	59	644	549

Mode is on categorical data

Dispersion on categorical data

```
titanic.value_counts('embark_town')
```

embark_town	count
Southampton	644
Cherbourg	168
Queenstown	77

```
titanic.value_counts('class')
```

class	count
Third	491
First	216
Second	184

```
titanic.value_counts('who')
```

who	count
man	537
woman	271
child	83

```
titanic.value_counts('class')
```

class	count
Third	491
First	216
Second	184

```
titanic.value_counts('pclass')
```

pclass	count
3	491
1	216
2	184

Redundancy

Sometimes a numerical feature is really a categorical feature

```
titanic[["age"]].describe()
```

	age
count	714.000000
mean	29.699118
std	14.526497
min	0.420000
25%	20.125000
50%	28.000000
75%	38.000000
max	80.000000

```
titanic.value_counts('age')
```

age	count
24.0	30
22.0	27
18.0	26
19.0	25
30.0	25
...	...
53.0	1
66.0	1
70.5	1
74.0	1
80.0	1

In general value_counts on a numerical data does not make sense

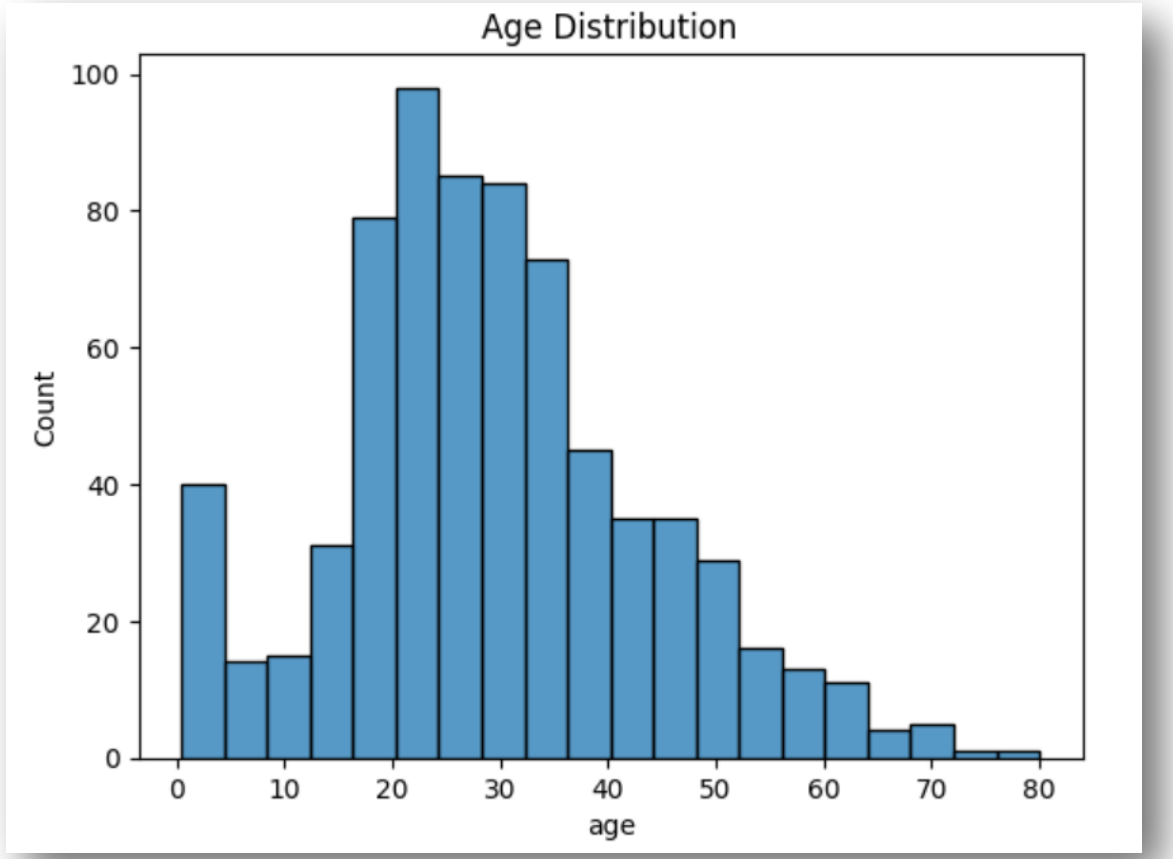
Better with graphs for visualisation of dispersion

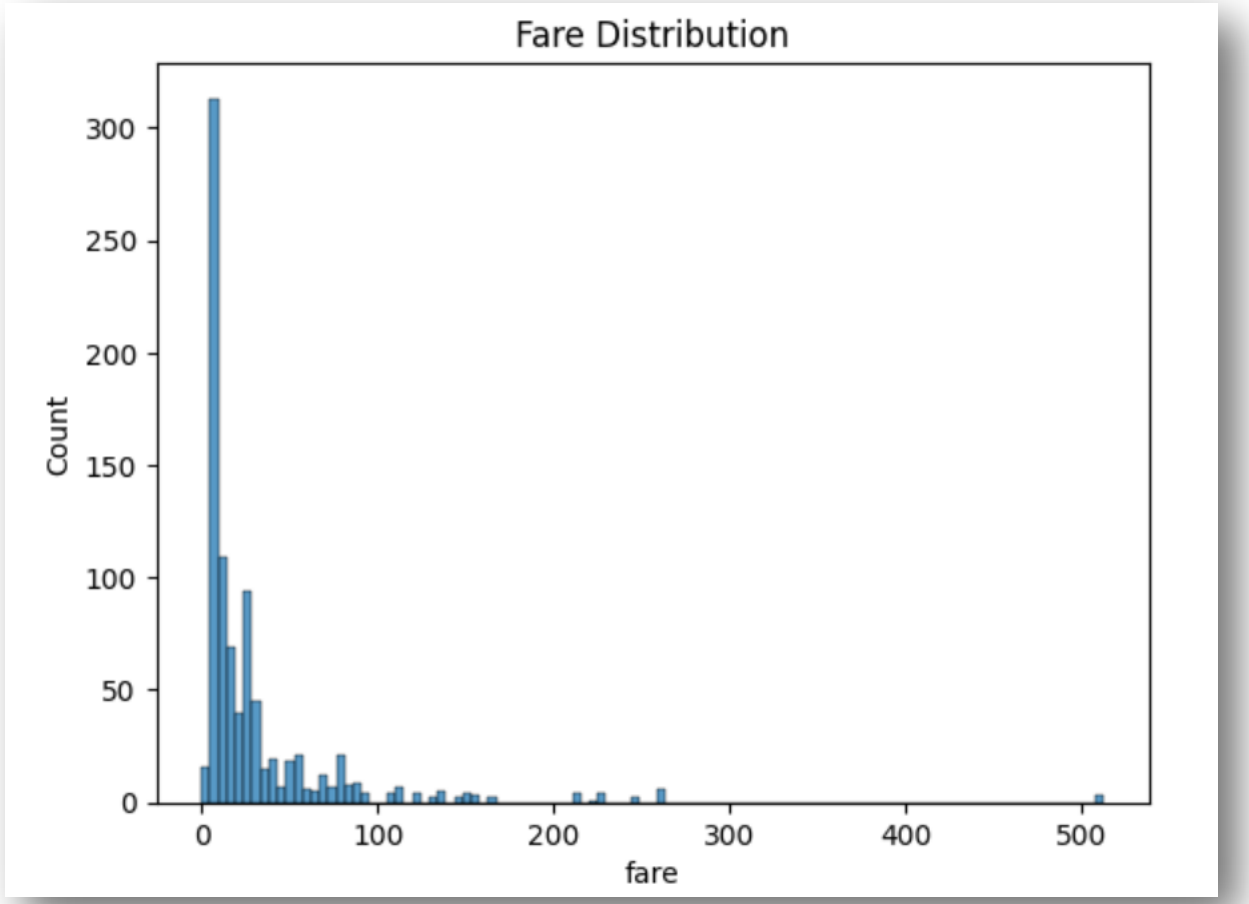
Univariate analysis Numerical Data

Histograms

A histogram is a bar chart that displays the frequency of a numerical variable's values. It is created by dividing the data into intervals, called bins, and counting the number of observations that fall within each bin.

```
sns.histplot(data=titanic, x="age")
plt.title("Age Distribution")
plt.show()
```





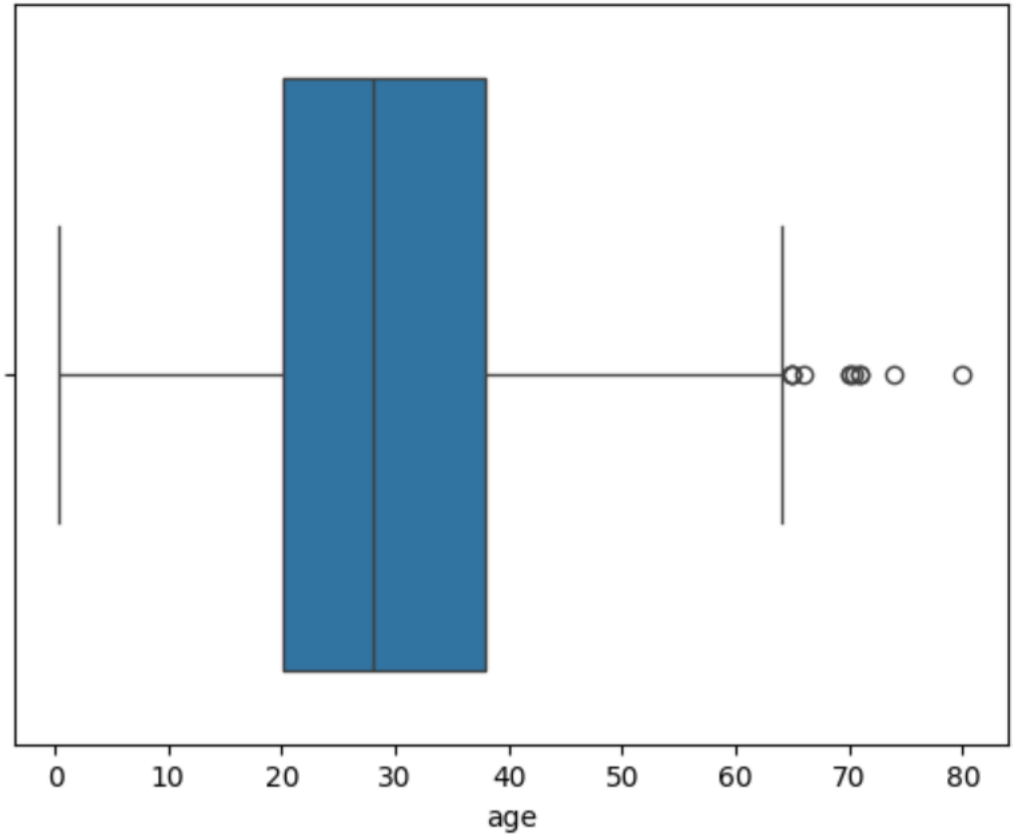
Further study of distributions…

Perhaps an outlier?

Box Plots

A box plot is a graphical summary of a numerical distribution showing its median, quartiles, spread, and potential outliers.

```
sns.boxplot(x="age", data=titanic)
```

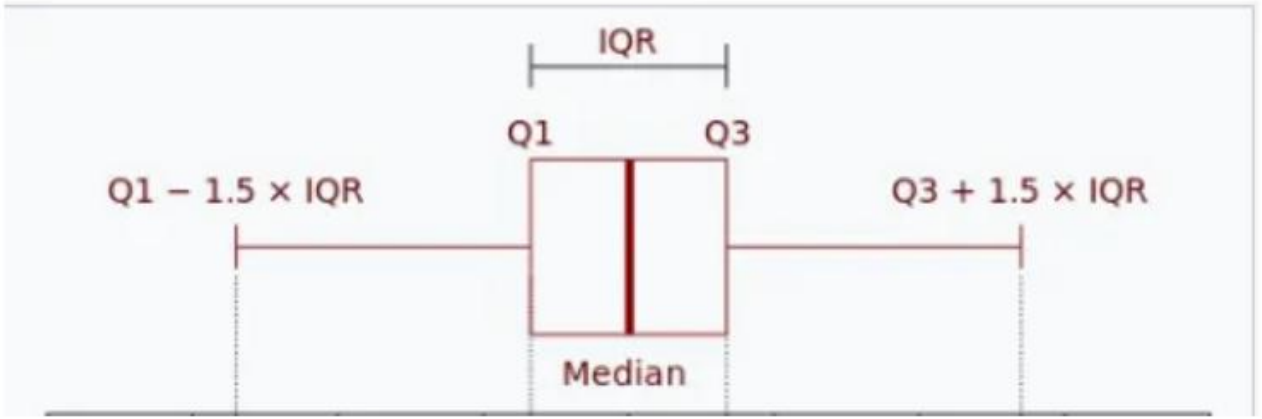


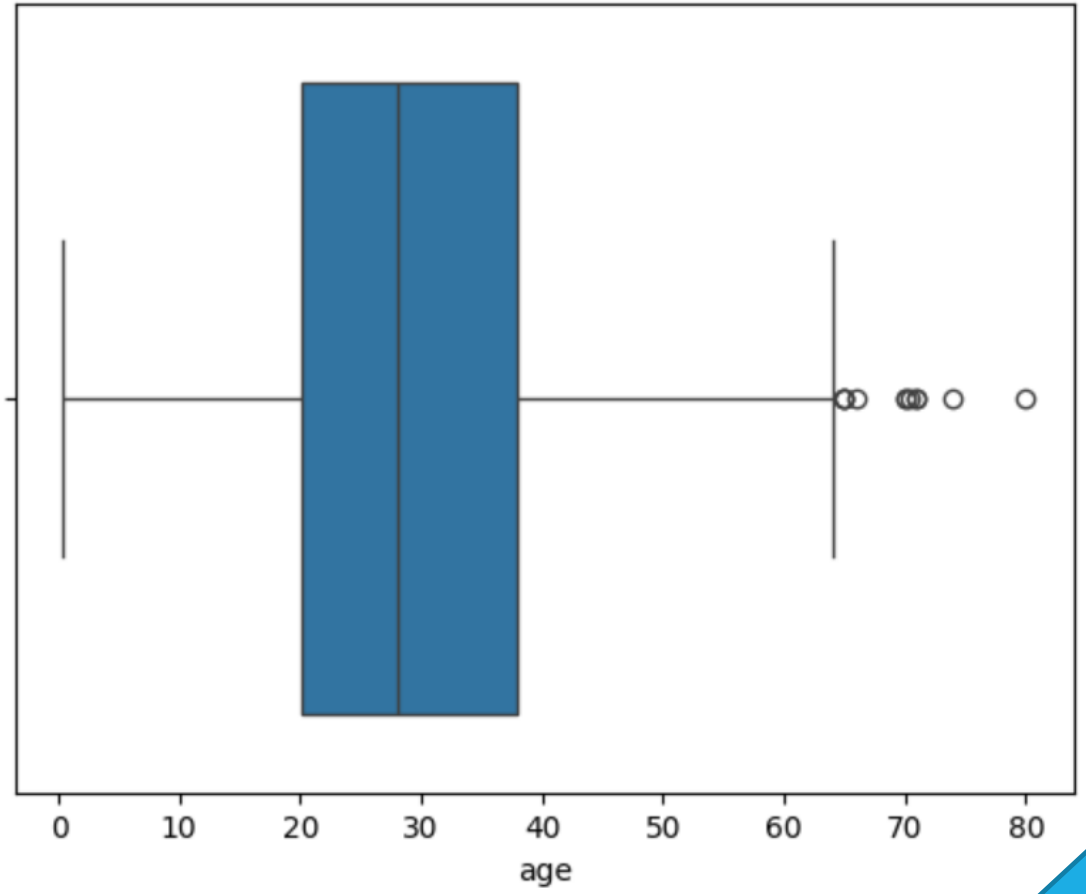
• The **middle line** represents the **median**, which means that 50% of the data points lie below this line, and 50% lie above.

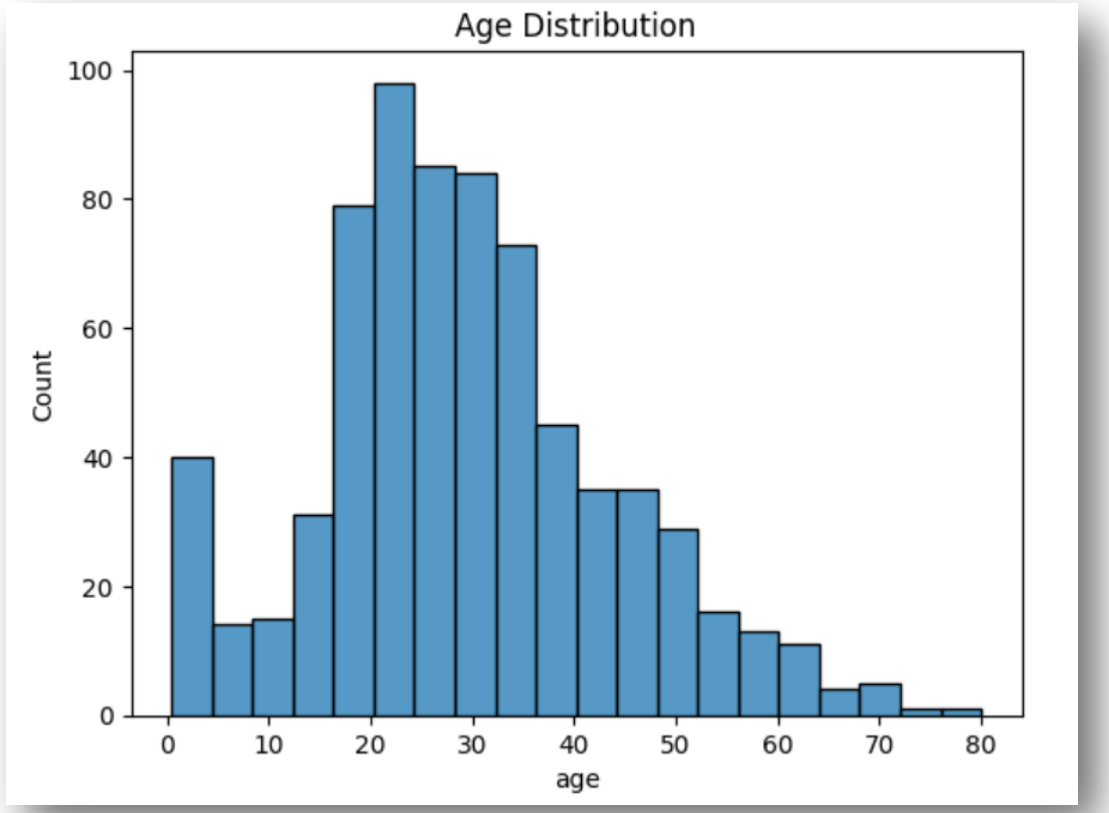
• Then we have the **first quartile (Q1)** and the **third quartile (Q3)**. The median is the second quartile.

• The range between Q1 and Q3 is called the **interquartile range (IQR)**, which represents 50% of the data.

• The **whiskers** extend from Q1 and Q3 to the **minimum** and **maximum** values within a defined range.







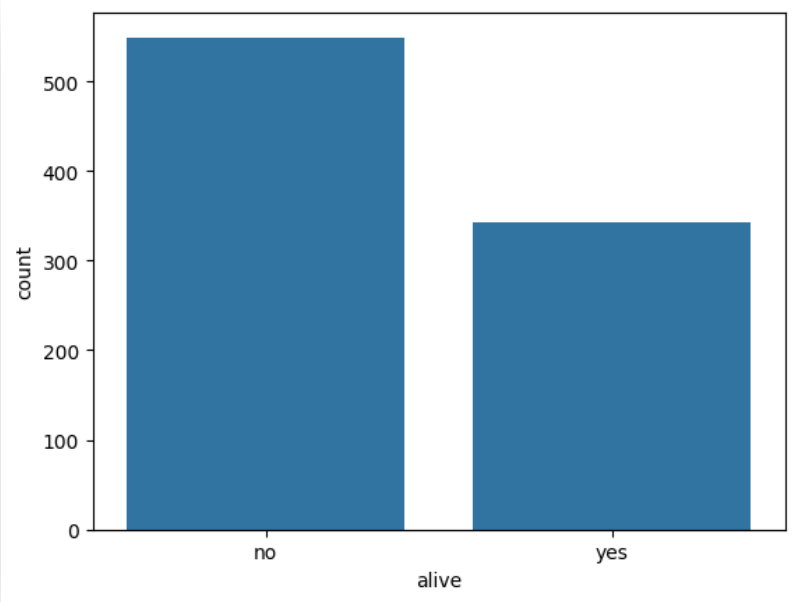
Often used side by side

Univariate analysis
Categorical Data

Count Plots

A count plot is a type of bar plot that displays the count of observations in each category.

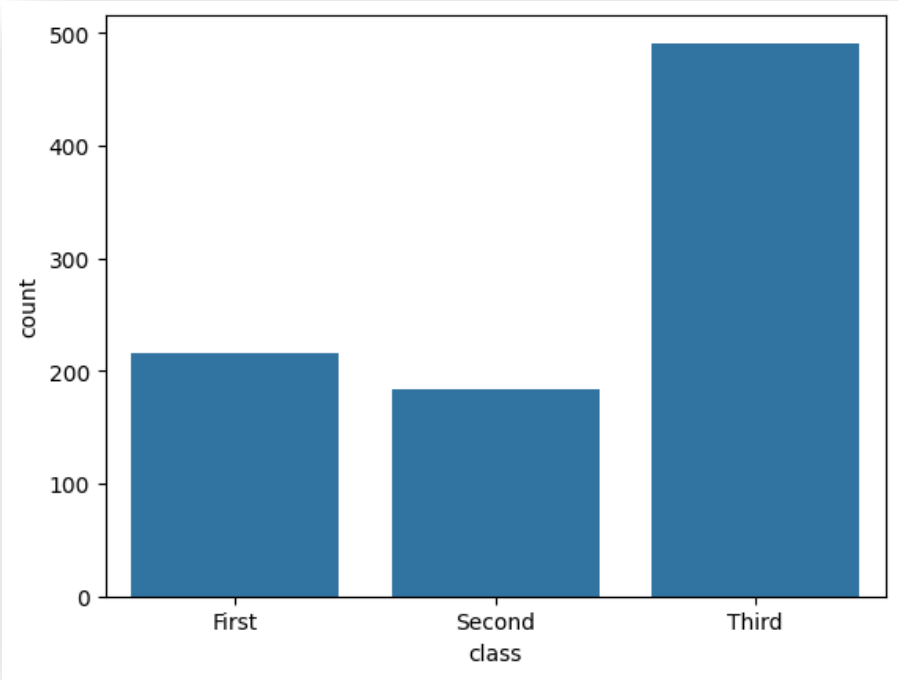
```
sns.countplot(x="alive", data=titanic)
plt.show()
```

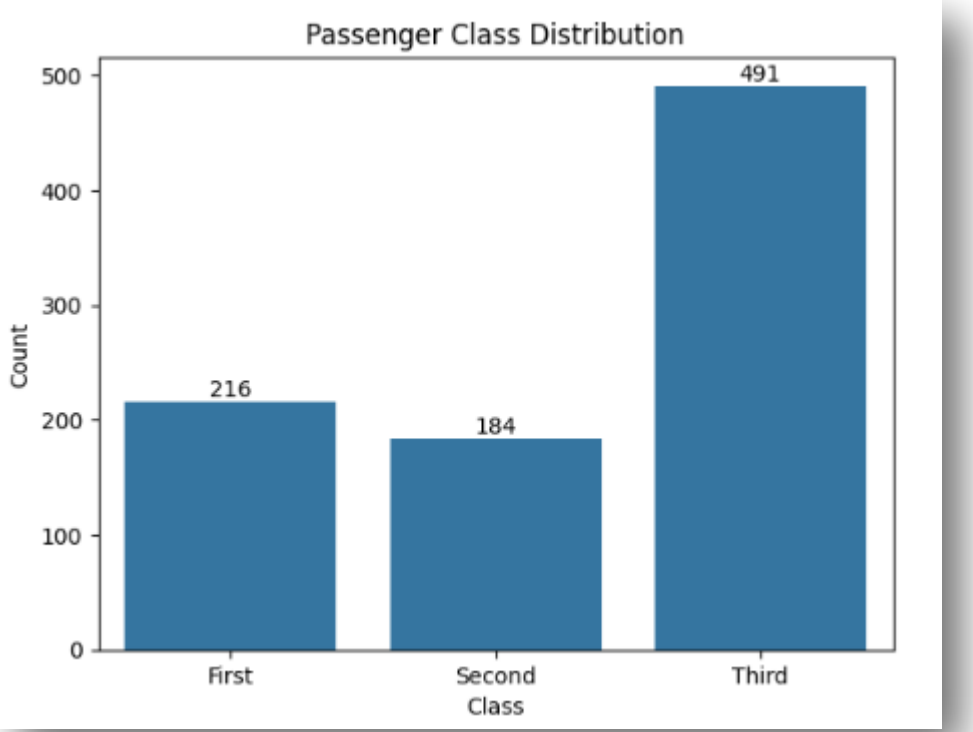


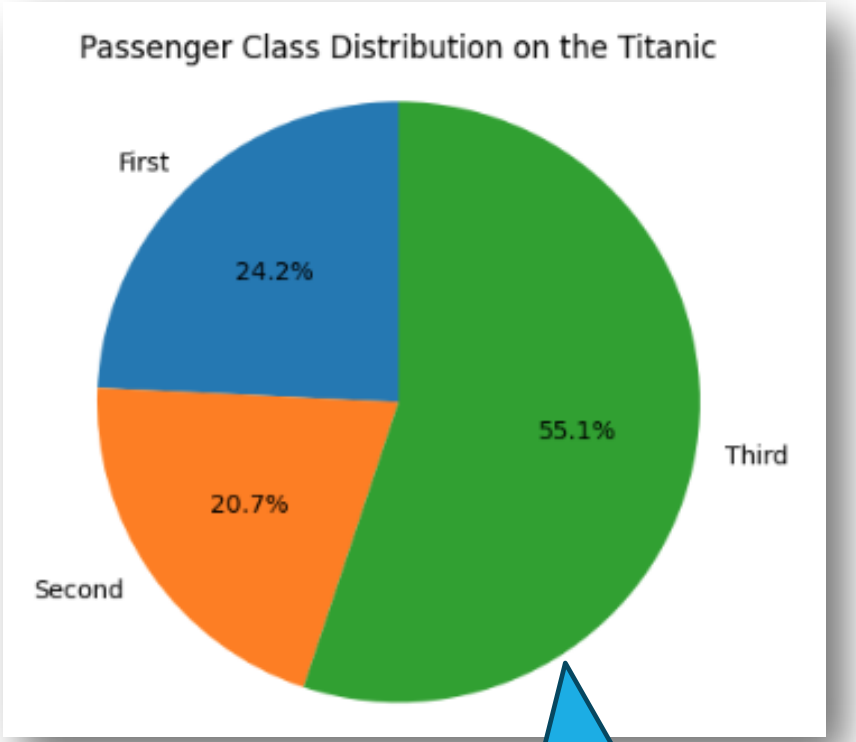
Data obtained about 891 passengers on the Titanic shows that most passengers died in the incident.

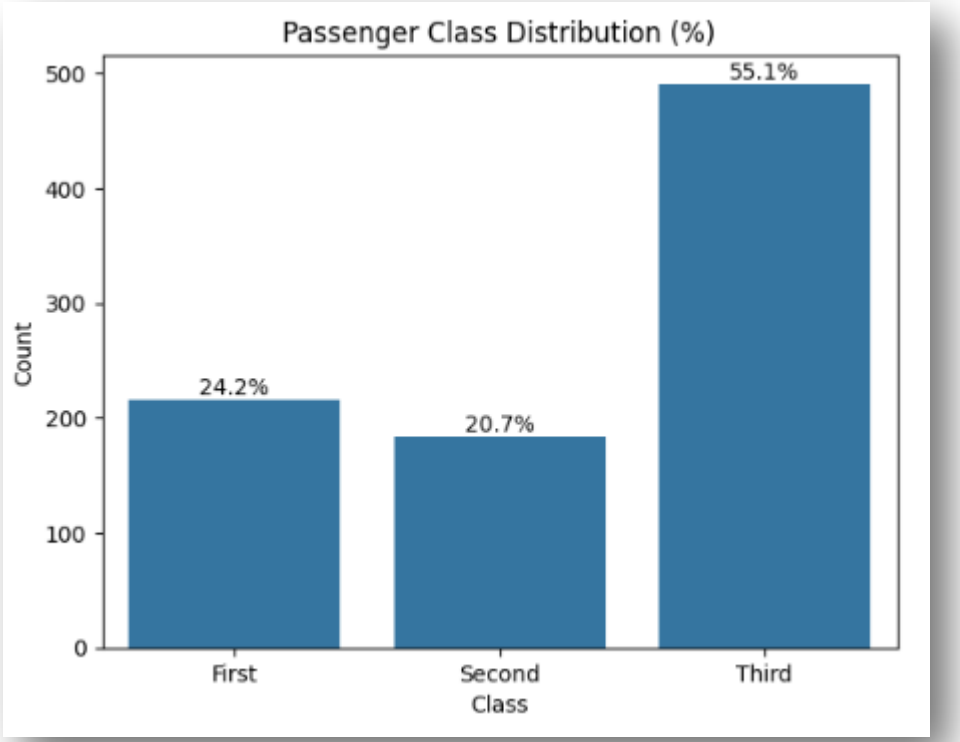
Communication Skills

Good practice to write a sentence (insignt) from what you observe in the graph.



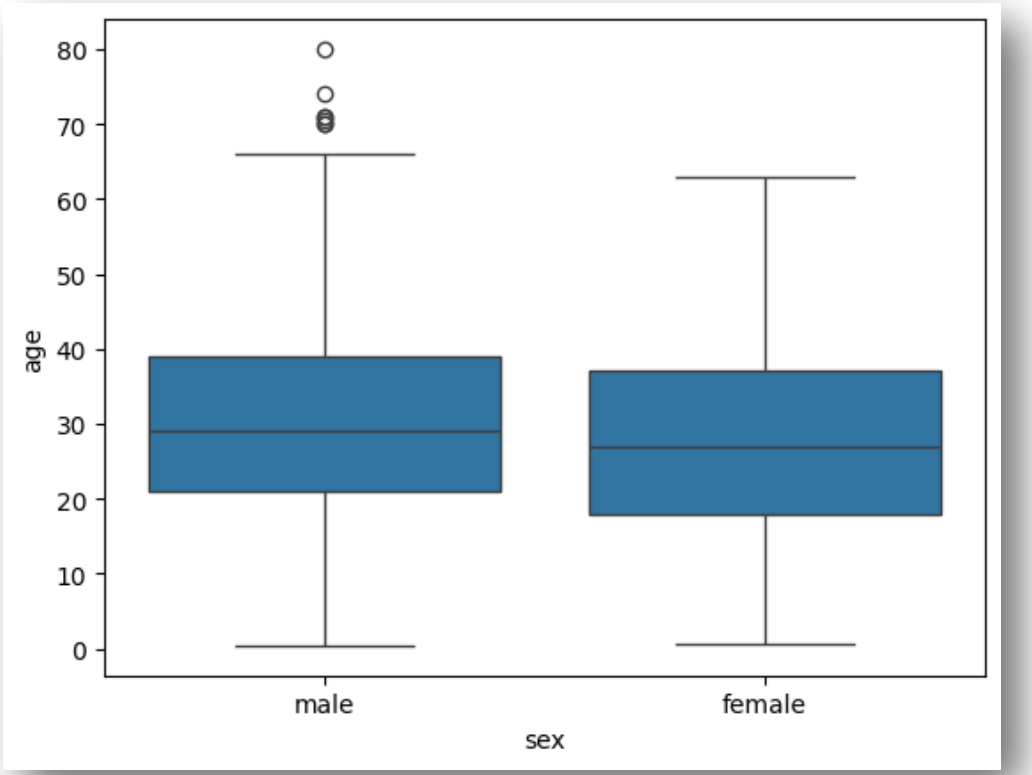






What about Pie Chart?

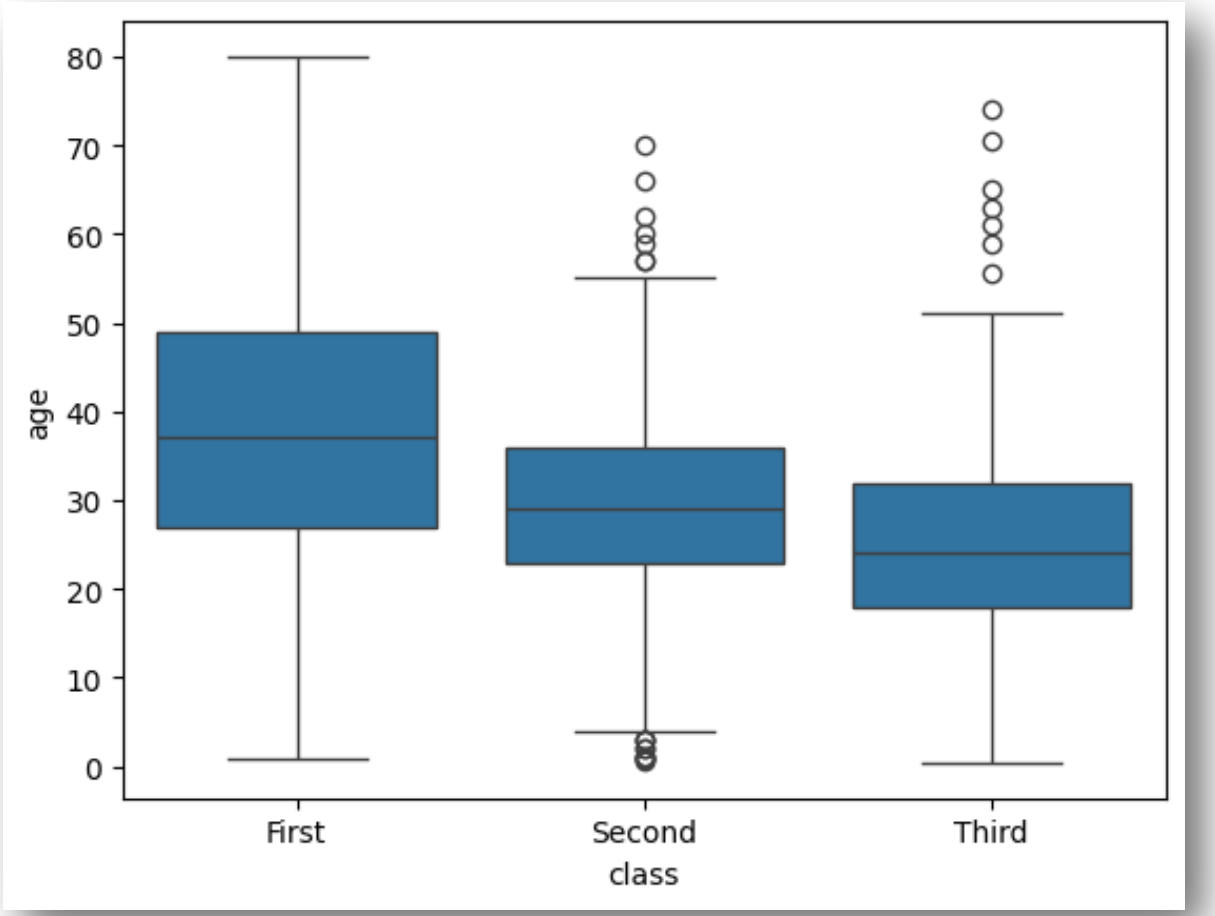
Bivariate analysis Numerical/Categorical



Shows central tendency statistics of one variable according to values of a second variable

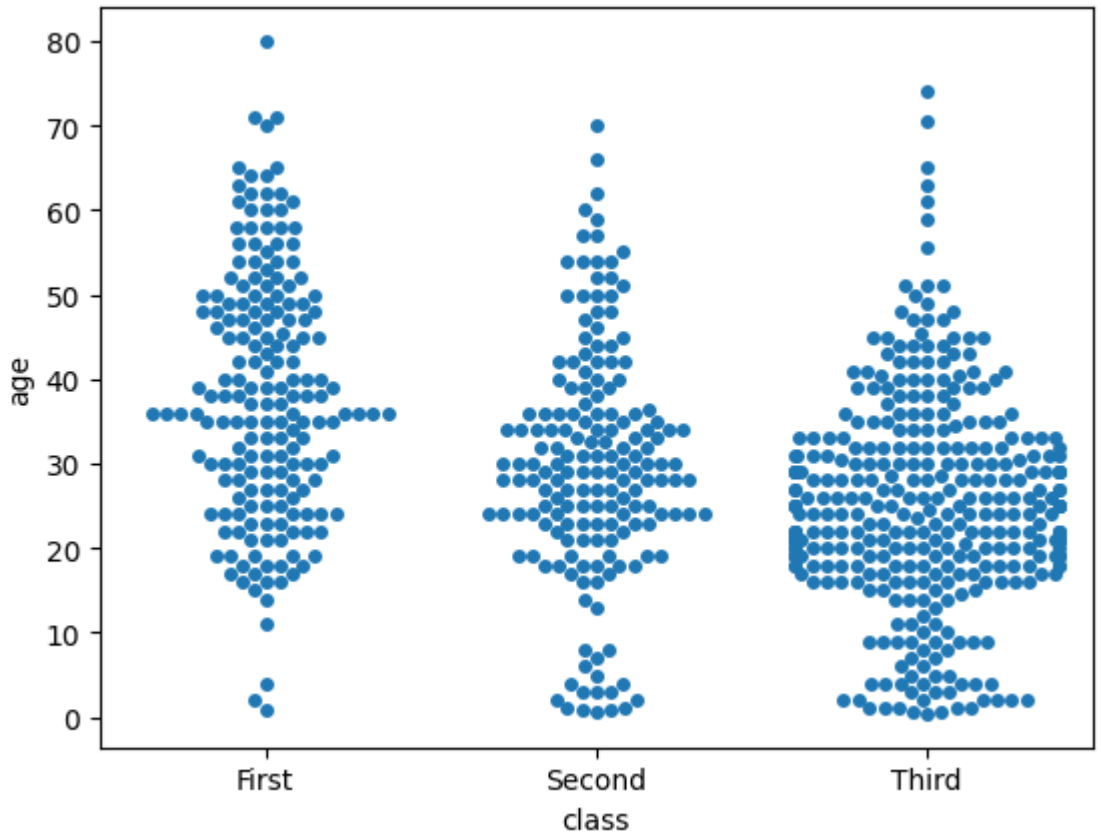
```
sns.boxplot(x="sex", y="age", data=titanic)
```

```
sns.boxplot(x="class", y="age", data=titanic)
```

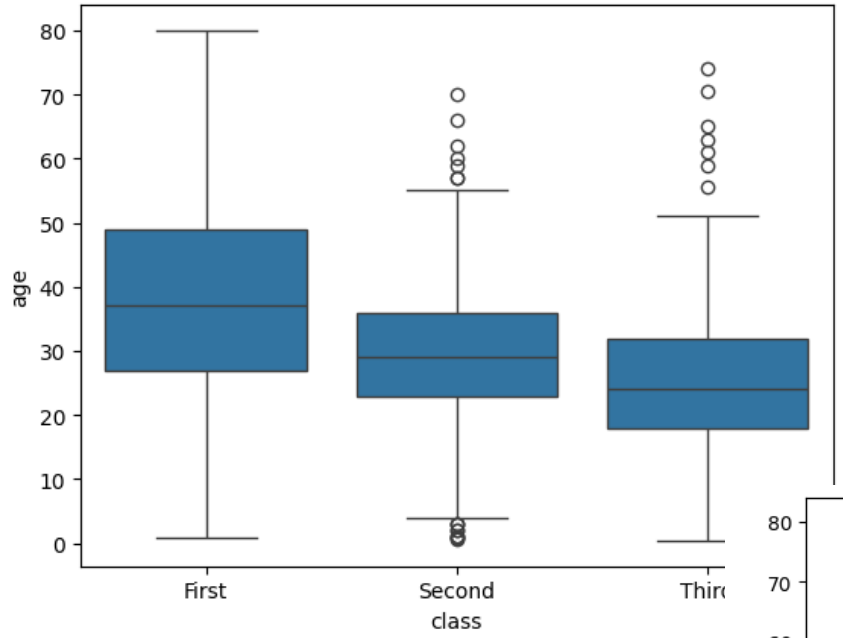


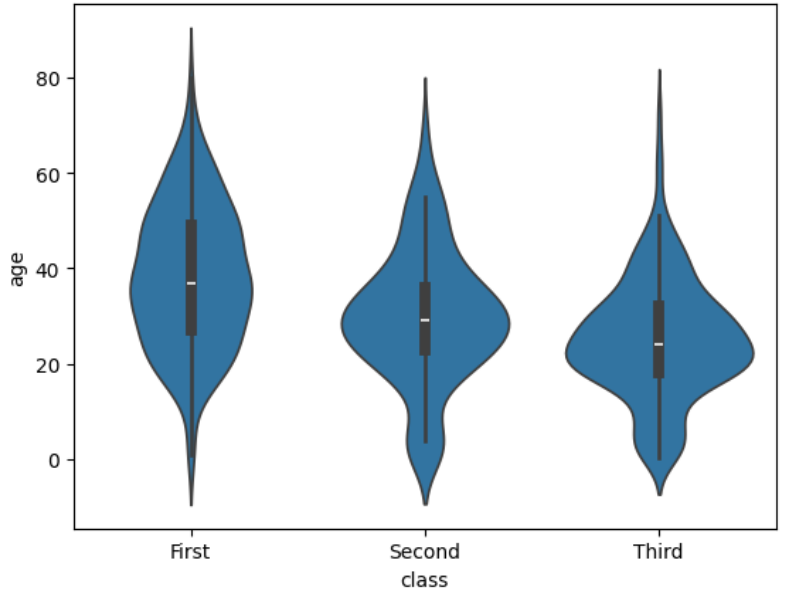
Insight?

```
sns.swarmplot(x="class", y="age", data=titanic)
```



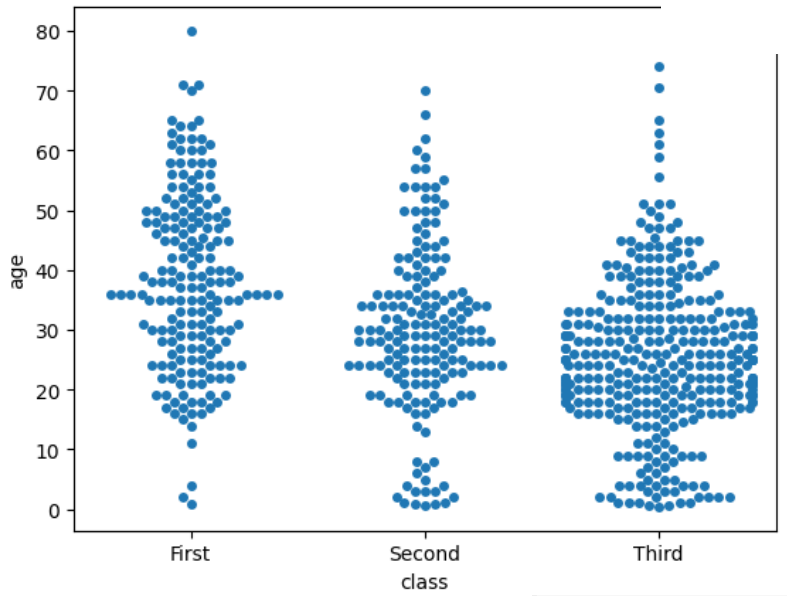
What do you think?






```
sns.violinplot(x="class", y="age", data=titanic)
```

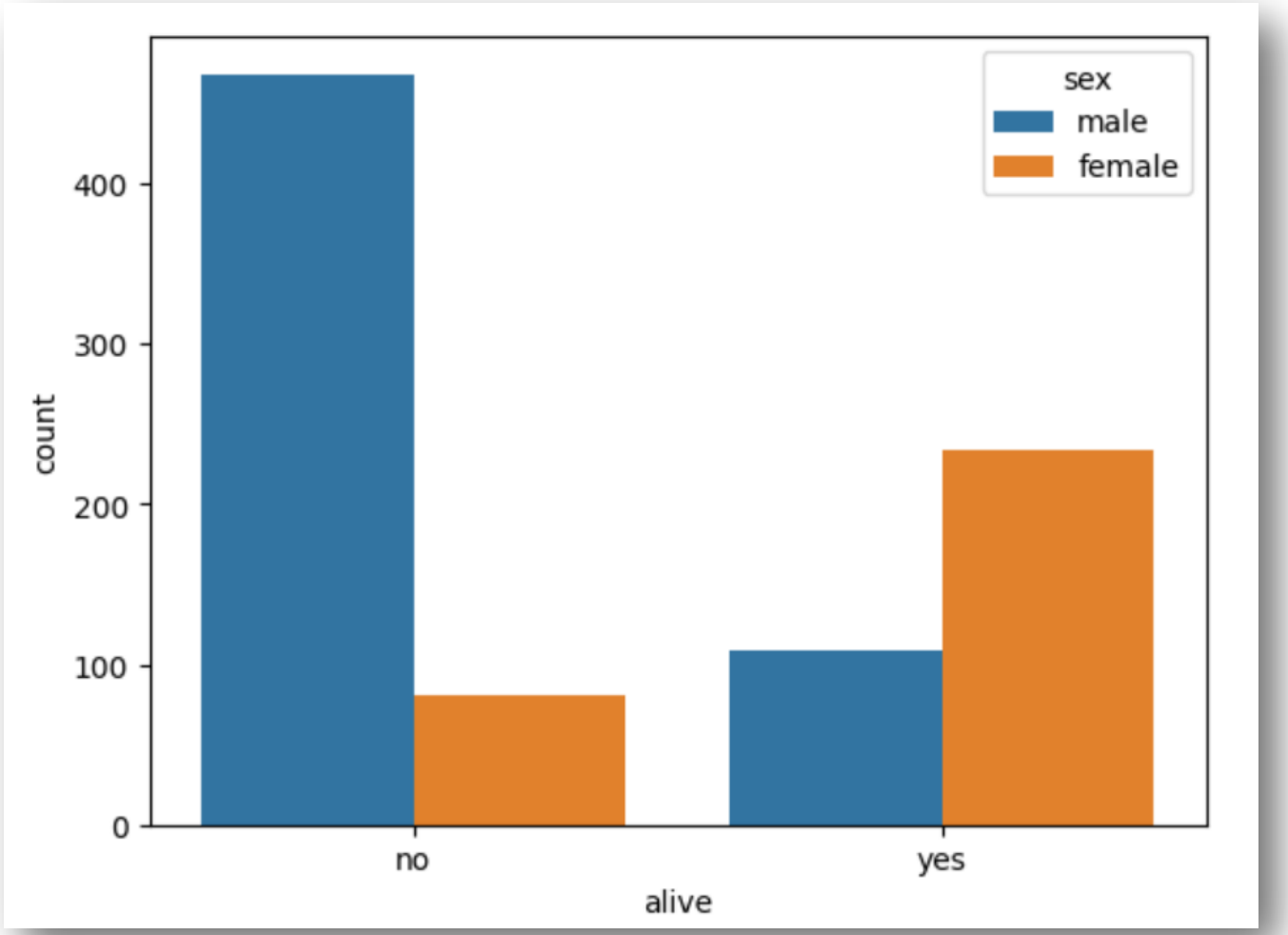
Which would you put in the report? Only a visual (preference) difference?



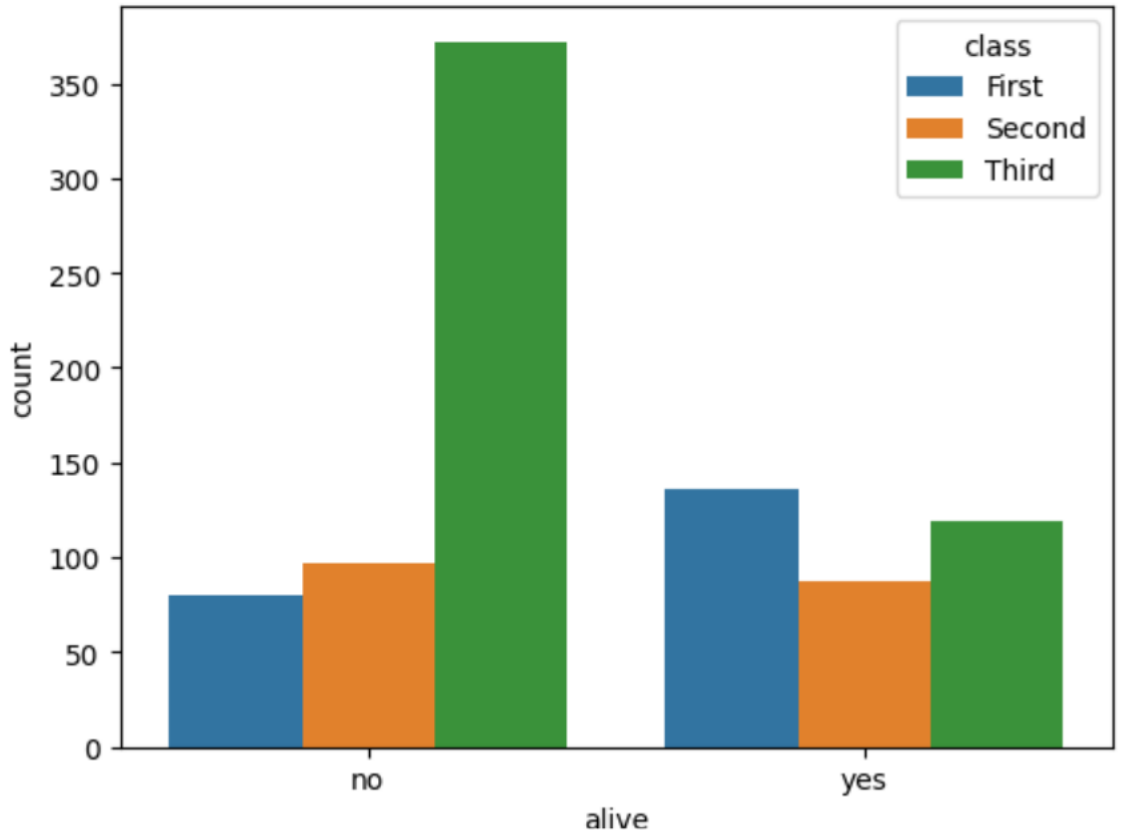

```
sns.swarmplot(x="class", y="age", data=titanic)
```

Bivariate analysis Categorical/Categorical

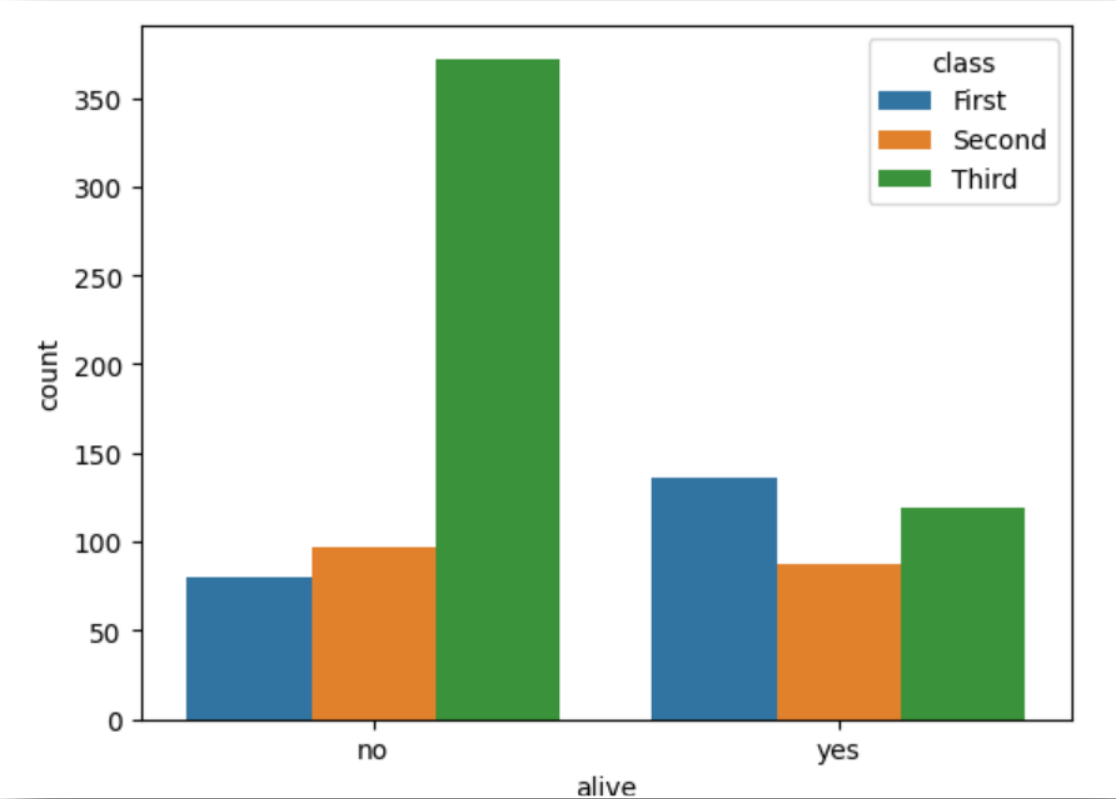
```
sns.countplot(x="alive", hue="sex", data=titanic)
plt.show()
```



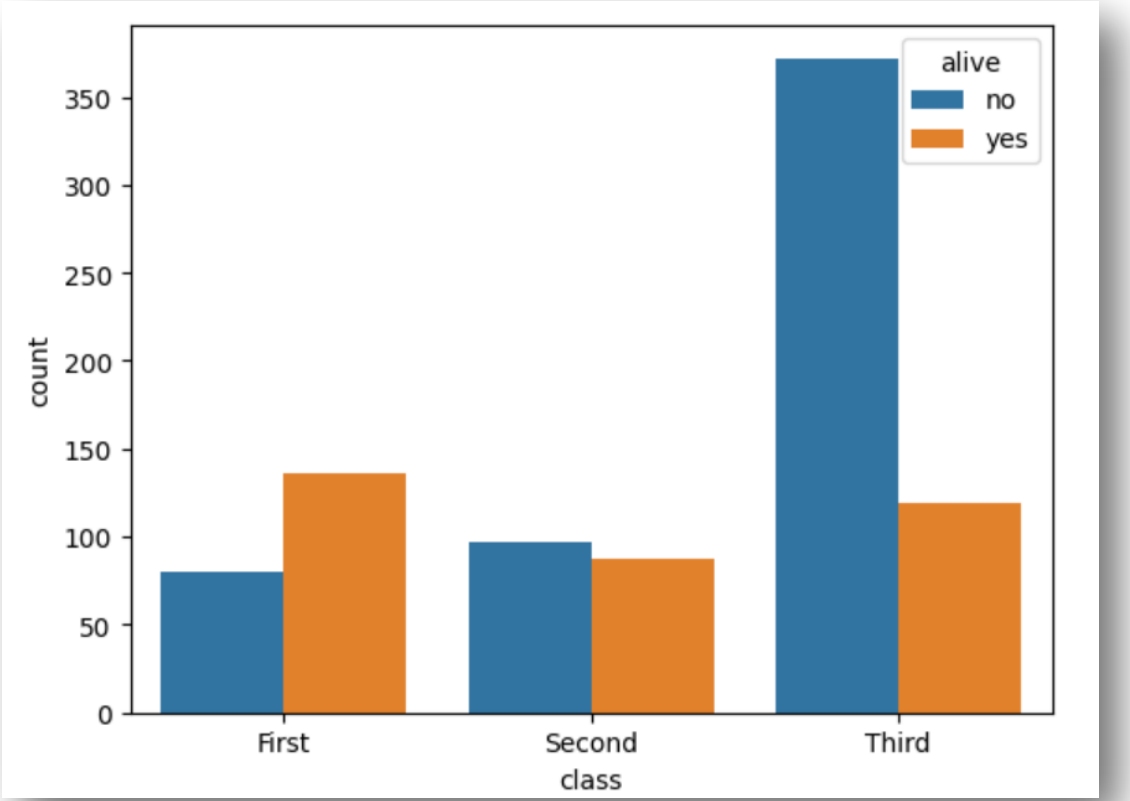
```
sns.countplot(x="alive", hue="class", data=titanic)
plt.show()
```



```
sns.countplot(x="alive", hue="class", data=titanic)
plt.show()
```

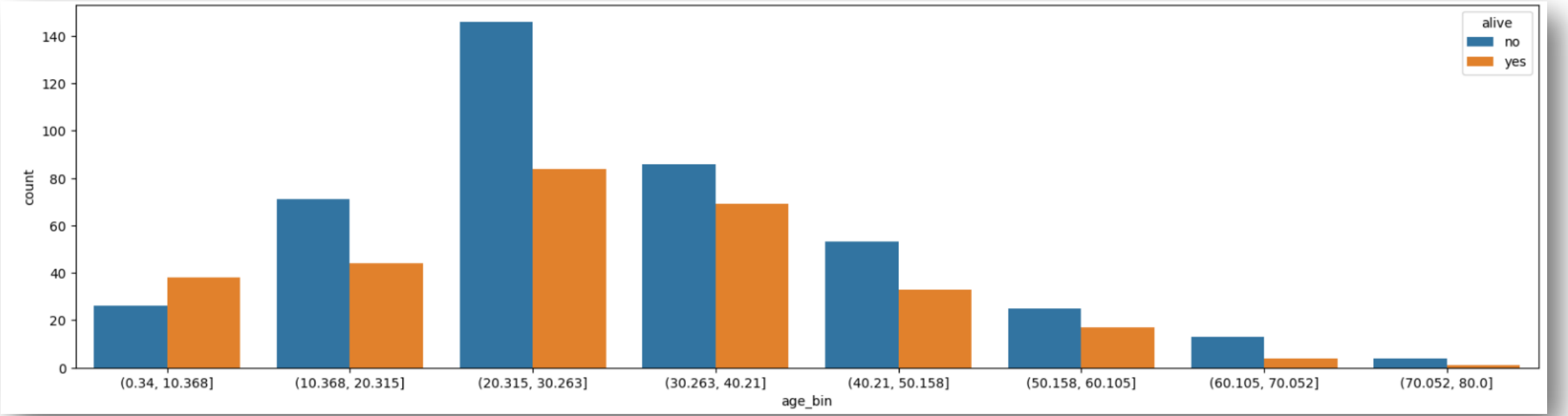



```
sns.countplot(x="class", hue="alive", data=titanic)
plt.show()
```



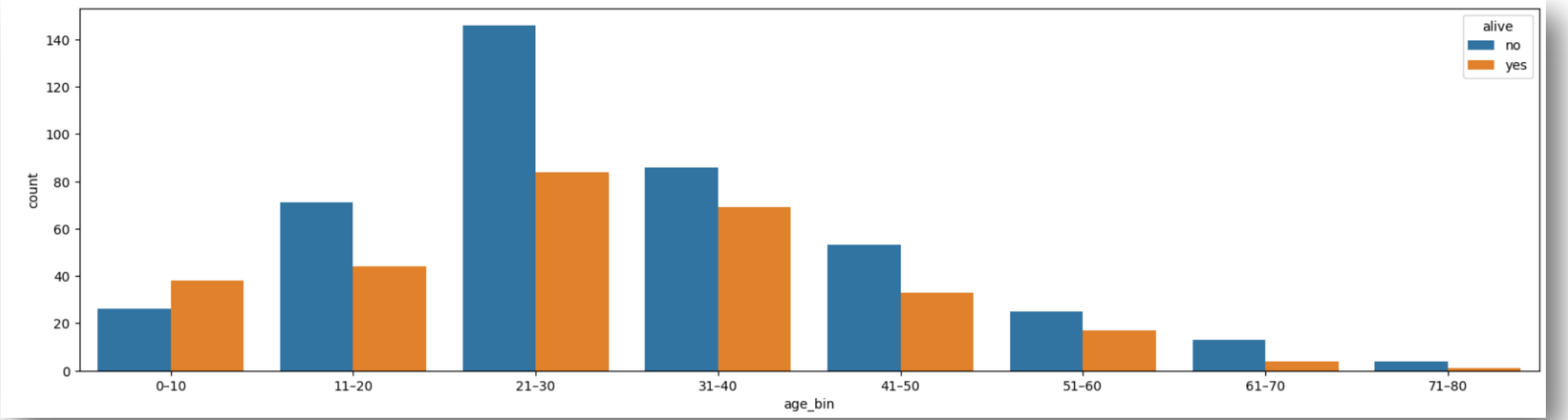
Bivariate analysis Categorical/Numerical

```
titanic['age_bin'] = pd.cut(titanic['age'], 8)
fig = plt.figure(figsize=(20,5))
sns.countplot(x="age_bin", hue="alive", data=titanic)
```



```python
bins = [0, 10, 20, 30, 40, 50, 60, 70, 80]
labels = ["0-10", "11-20", "21-30", "31-40", "41-50", "51-60", "61-70", "71-80"]

titanic["age_bin"] = pd.cut(
    titanic["age"],
    bins=bins,
    labels=labels,
    right=True,       # intervals are (a, b]
    include_lowest=True
)

fig = plt.figure(figsize=(20,5))
sns.countplot(x="age_bin", hue="alive", data=titanic)
```

Instead of using the bins like here titanic['age_bin'] = pd.cut(titanic['age'], 8), how can I make my own bins of 0-10, 11-20, etc

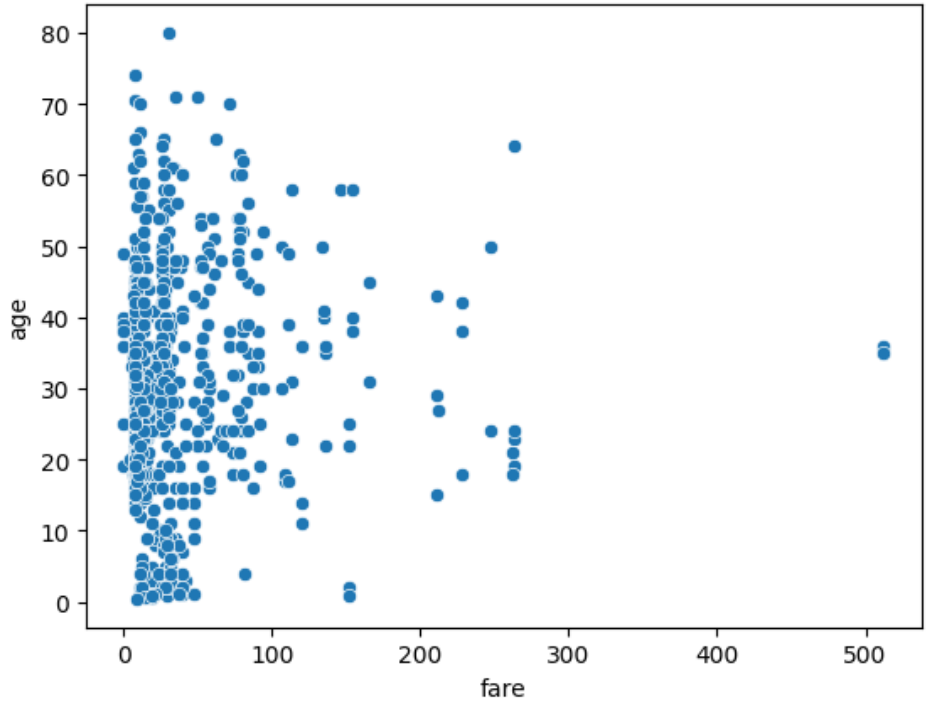


Bivariate analysis Numerical/Numerical

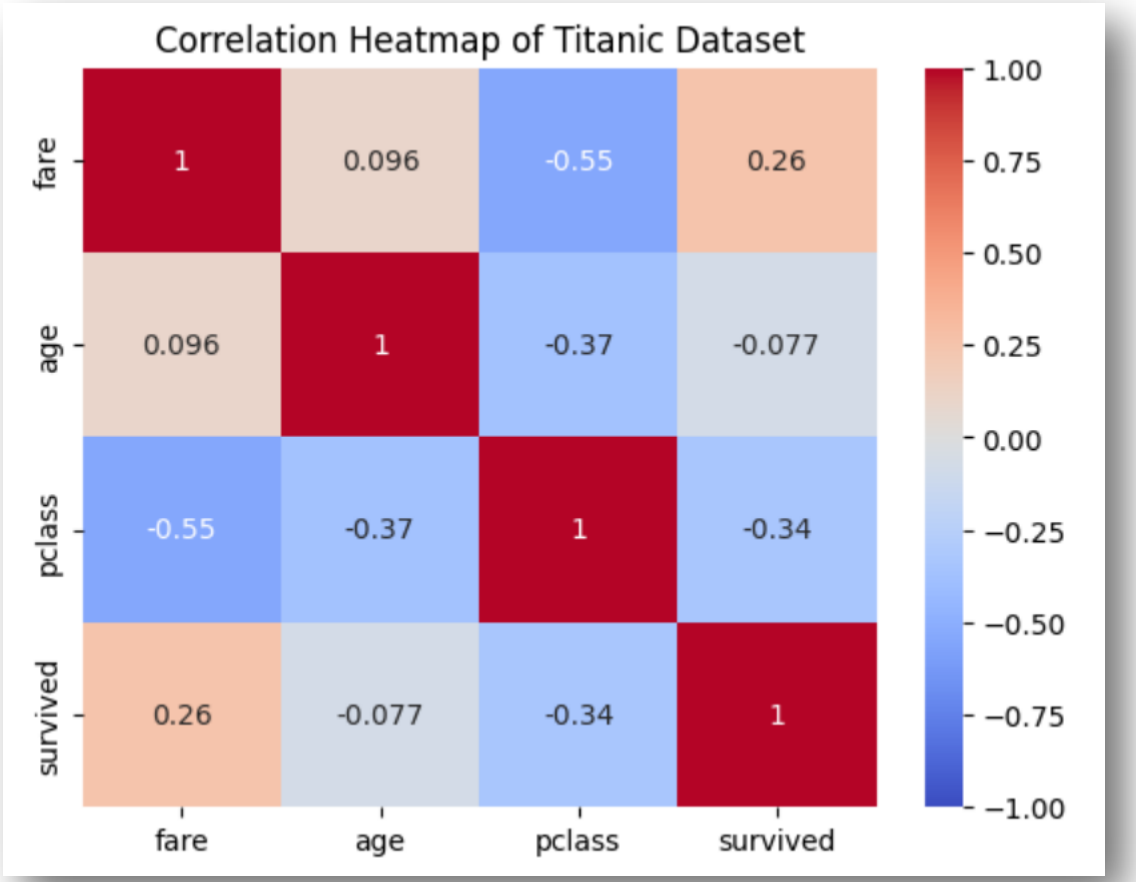
Scatter Plots

A **scatter plot** shows the relationship between two continuous variables, making it great for visualizing correlation.

```
sns.scatterplot(x="fare", y="age", data=titanic)
```



```python
cols = ["fare", "age", "pclass", "survived"]
corr = titanic[cols].corr()

sns.heatmap(
    corr,
    vmin=-1,
    vmax=1,
    center=0,
    cmap="coolwarm",
    annot=True
)

plt.title("Correlation Heatmap of Titanic Dataset")
plt.show()
```



Univariate Analysis

- Numerical feature
- Categorical feature

Bivariate Analysis

- Numerical/Categorical
- Categorical/Categorical
- Categorical/Numerical
- Categorical/Categorical

EXPLORATORY DATA ANALYSIS

- Descriptive statistics
- Diagnostic statistics
- Case Study 1 – Titanic

Friday Case Study 2 +
Examples for Assignment 1