Data Visualization

Revision Class 01 - Basic Plots and Random Variables

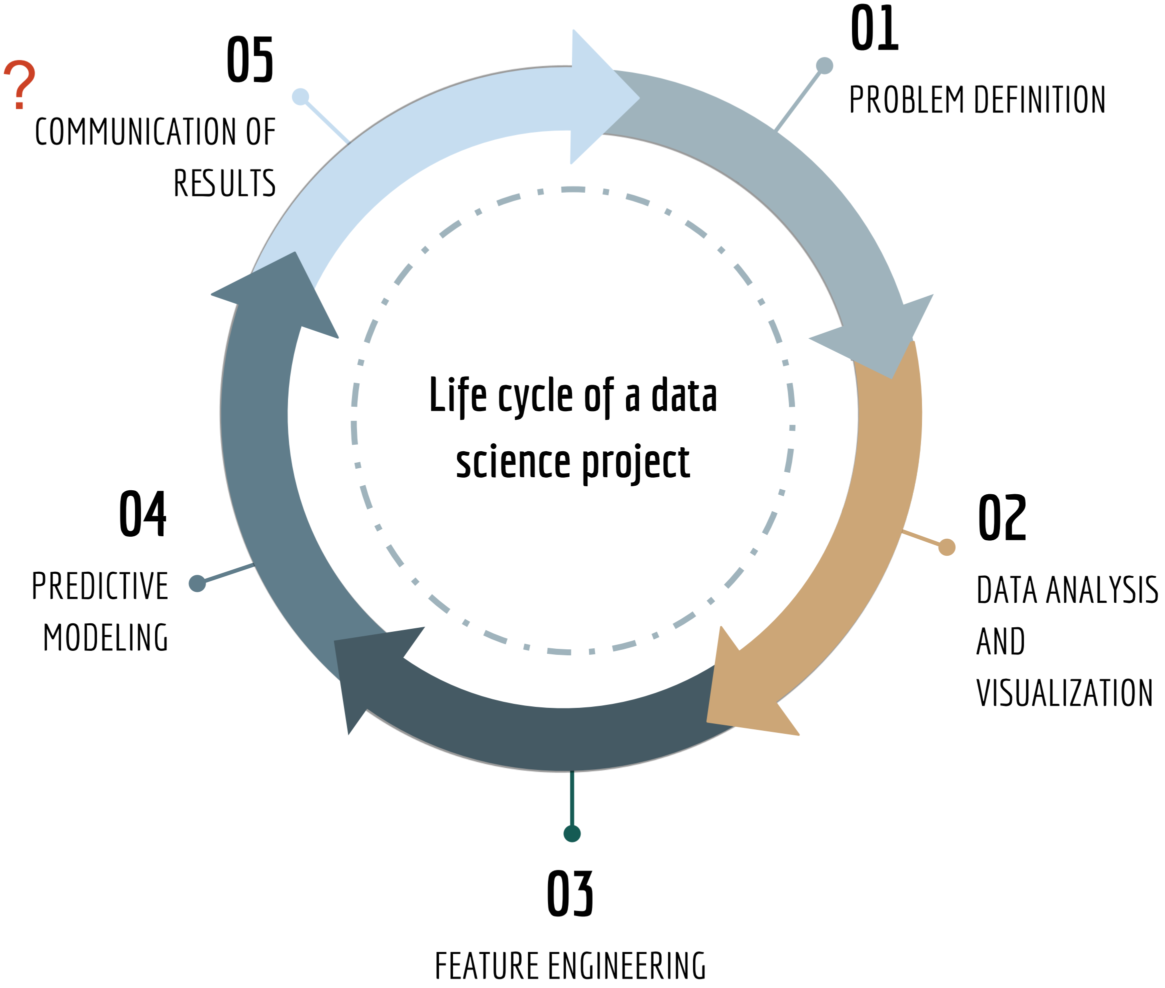
DigitalLab@LaPlataforme_

What is Data Science?

Data science is a discipline that aims to **develop a product based on data**.

Uses approaches from the **data analysis** and **machine learning**.

Visualization plays an important role on steps: **02**, **04** and **05.**



Data Visualization

Data visualization is relevant in the data science process as it helps to:

- **Identify** relevant information and **properties in our dataset**.

Data Visualization

Data visualization is relevant in the data science process as it helps to:

- **Identify** relevant information and **properties in our dataset**.
- Detect patterns and **correlations between variables**.

Data Visualization

Data visualization is relevant in the data science process as it helps to:

- **Identify** relevant information and **properties in our dataset**.
- Detect patterns and **correlations between variables**.
- Experiment and **provide answers to hypothesis** during our research process.

Data Visualization

Data visualization is relevant in the data science process as it helps to:

- **Identify** relevant information and **properties in our dataset**.
- Detect patterns and **correlations between variables**.
- Experiment and **provide answers to hypothesis** during our research process.
- Recognize machine learning model **relevant features**.

Data Visualization

Data visualization is relevant in the data science process as it helps to:

- **Identify** relevant information and **properties in our dataset**.
- Detect patterns and **correlations between variables**.
- Experiment and **provide answers to hypothesis** during our research process.
- Recognize machine learning model **relevant features**.
- **Communicate results** to team members.

Random Variable

A **random variable** (r.v.) X is a **function** X: Ω → R where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

Random Variable

A **random variable** (r.v.) X is a **function** X: Ω → R where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

Intuitively, a r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.

Random Variable

A **random variable** (r.v.) X is a **function** X: Ω → R where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

Intuitively, a r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.

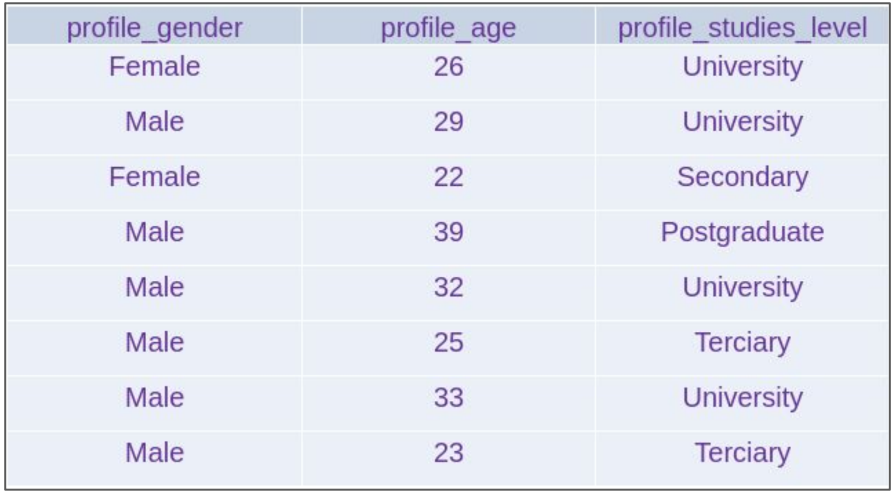
The random variables can be of different types:

- Numerical
  - Continuous
  - Discrete (Infinite or finite set of numerable values)
- Categorical
- Ordinal

Random Variable

A **random variable** (r.v.) X is a **function** $X: \Omega \rightarrow R$ where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

A r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.



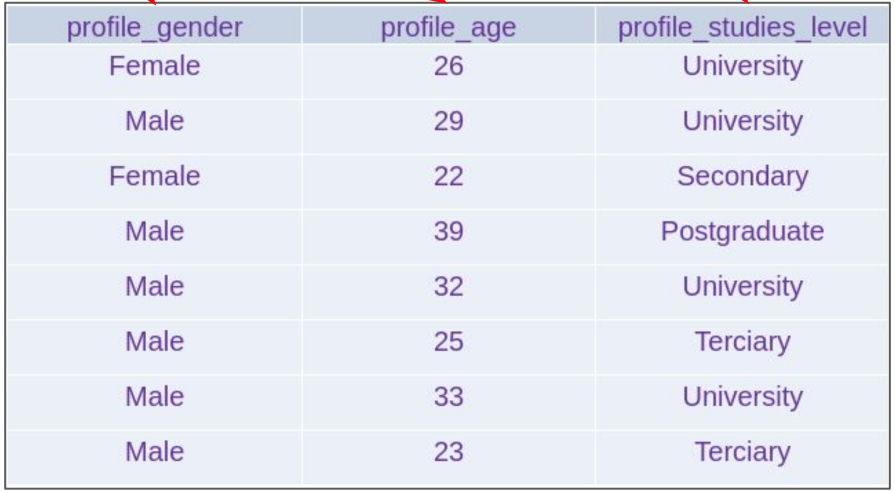
profile_gender	profile_age	profile_studies_level
Female	26	University
Male	29	University
Female	22	Secondary
Male	39	Postgraduate
Male	32	University
Male	25	Terciary
Male	33	University
Male	23	Terciary

Random Variable

A **random variable** (r.v.) X is a **function** $X: \Omega \to R$ where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

A r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.

Columns (Random Variables)

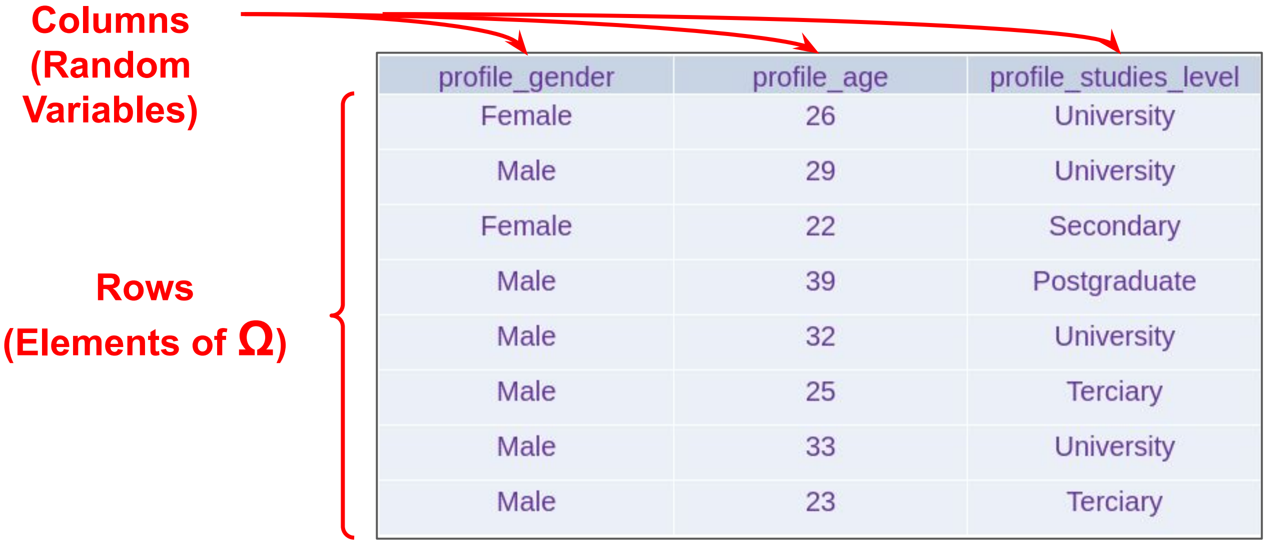


profile_gender	profile_age	profile_studies_level
Female	26	University
Male	29	University
Female	22	Secondary
Male	39	Postgraduate
Male	32	University
Male	25	Terciary
Male	33	University
Male	23	Terciary

Random Variable

A **random variable** (r.v.) X is a **function** X: Ω → R where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

A r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.



Columns (Random Variables)

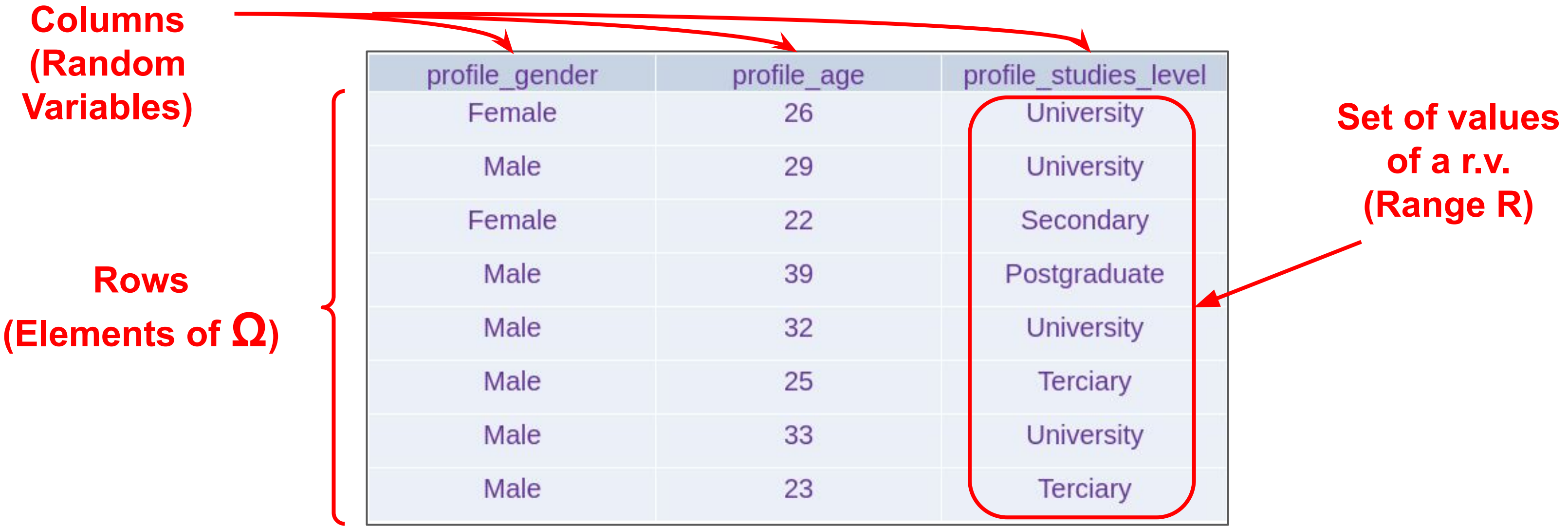
Rows (Elements of Ω)

profile_gender	profile_age	profile_studies_level
Female	26	University
Male	29	University
Female	22	Secondary
Male	39	Postgraduate
Male	32	University
Male	25	Terciary
Male	33	University
Male	23	Terciary

Random Variable

A **random variable** (r.v.) X is a **function** X: Ω → R where **Ω is the state space** and **R** is the set of values that the variable can take called **Range**.

A r.v. is **equivalent to a column** of your dataset after applying 0 or more filters.



Columns (Random Variables)

Rows (Elements of Ω)

Set of values of a r.v. (Range R)

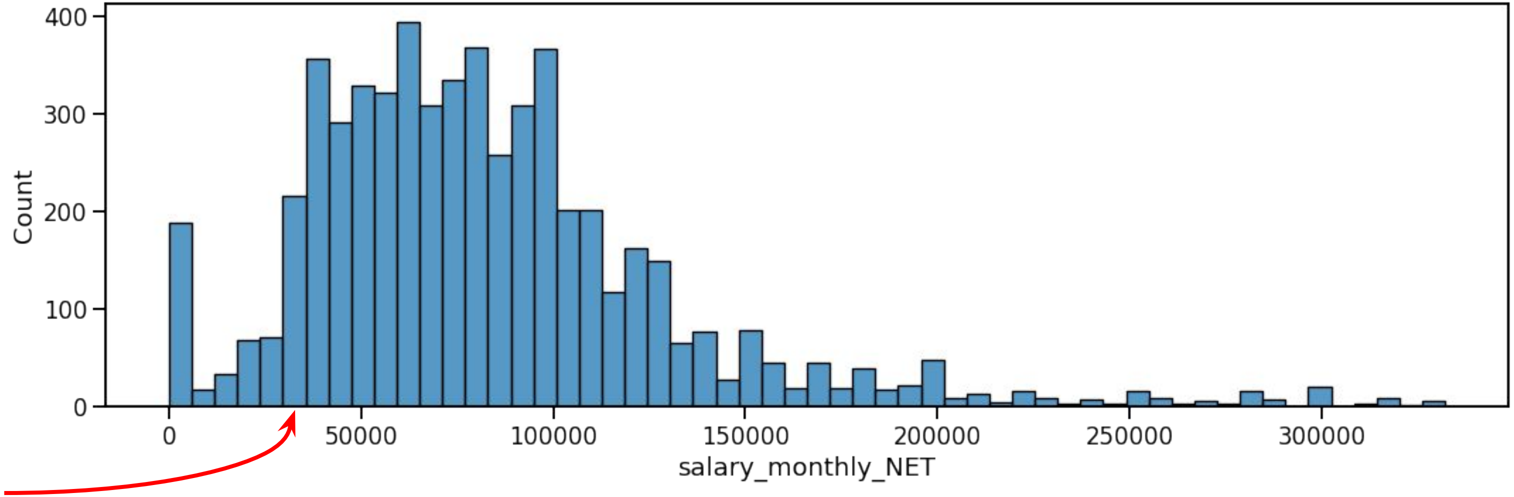
profile_gender	profile_age	profile_studies_level
Female	26	University
Male	29	University
Female	22	Secondary
Male	39	Postgraduate
Male	32	University
Male	25	Terciary
Male	33	University
Male	23	Terciary

Basic Plots: Histogram

Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).

Basic Plots: Histogram

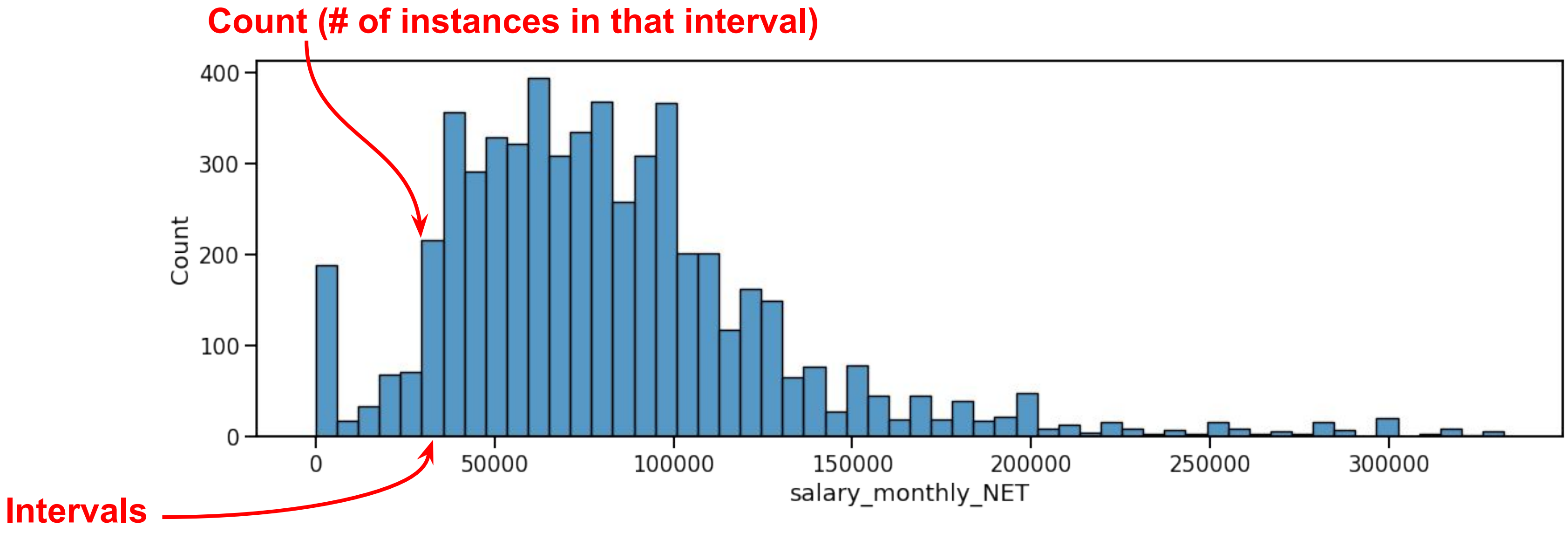
Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).



Intervals

Basic Plots: Histogram

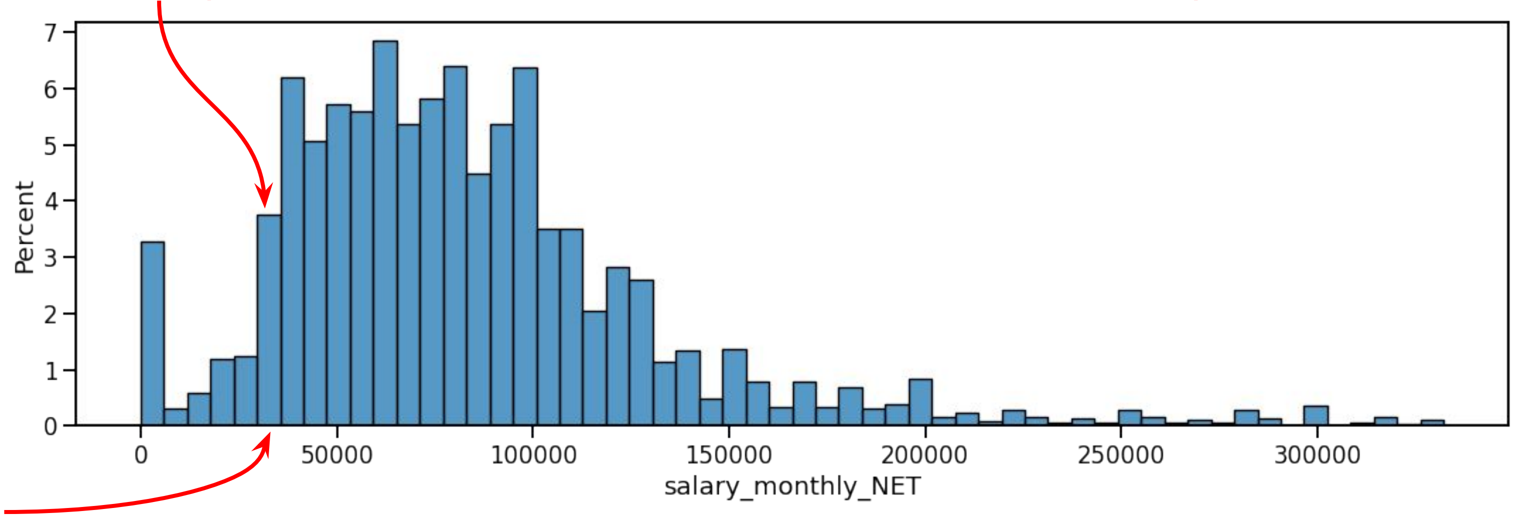
Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).



Basic Plots: Histogram

Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).

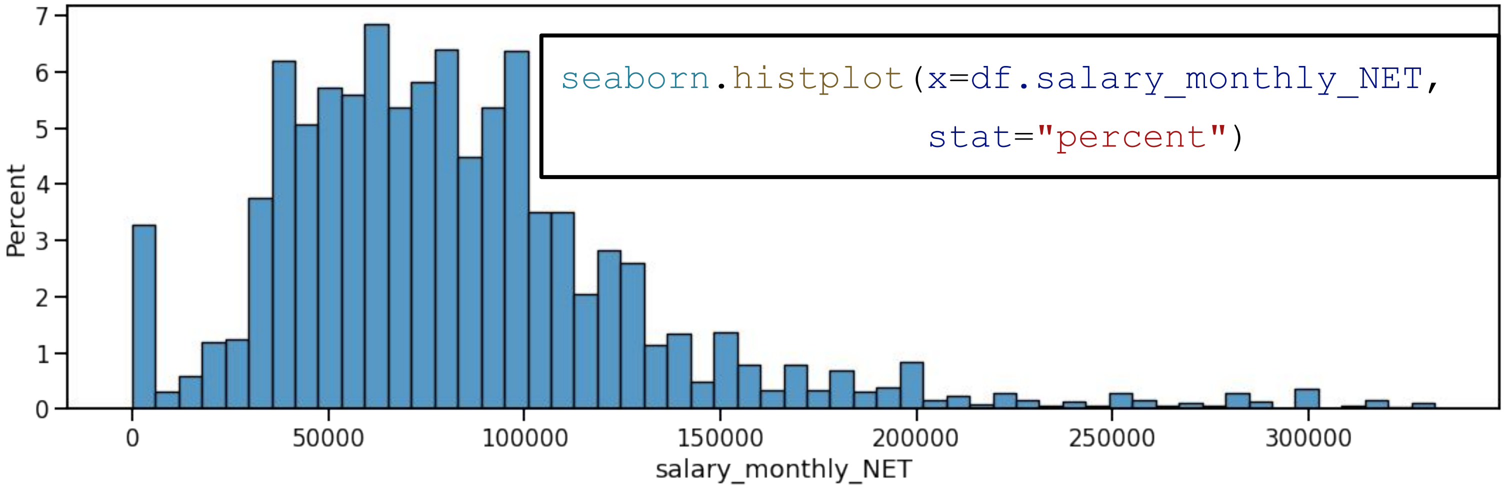
Percent (# of instances in that interval / total # of instances * 100)



Intervals

Basic Plots: Histogram

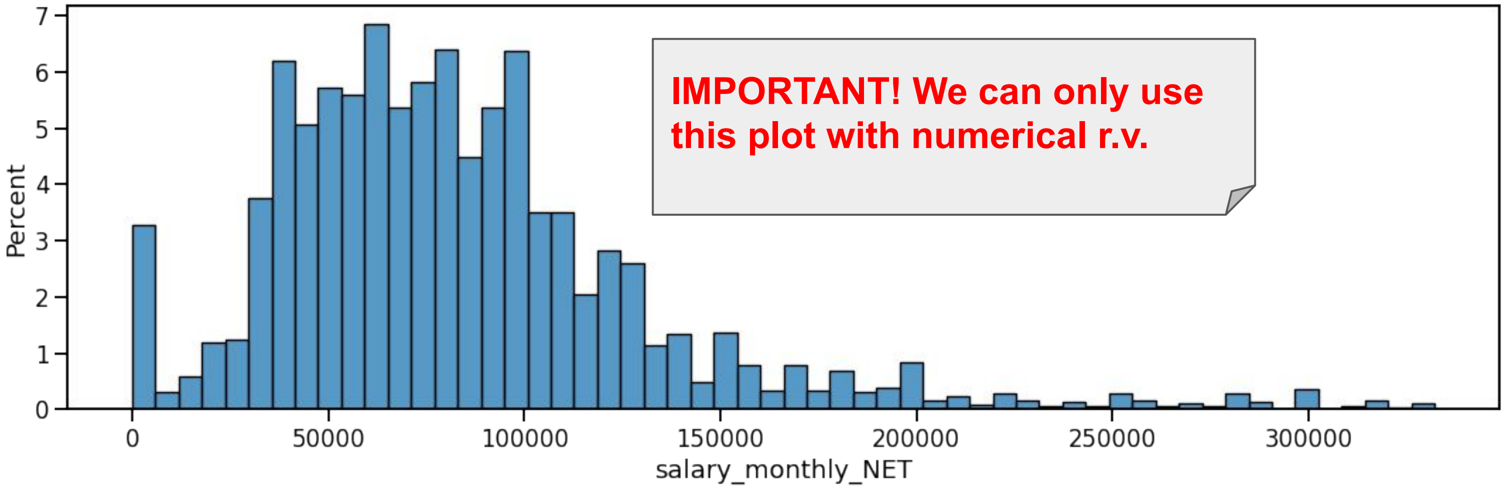
Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).



```
seaborn.histplot(x=df.salary_monthly_NET,
                 stat="percent")
```

Basic Plots: Histogram

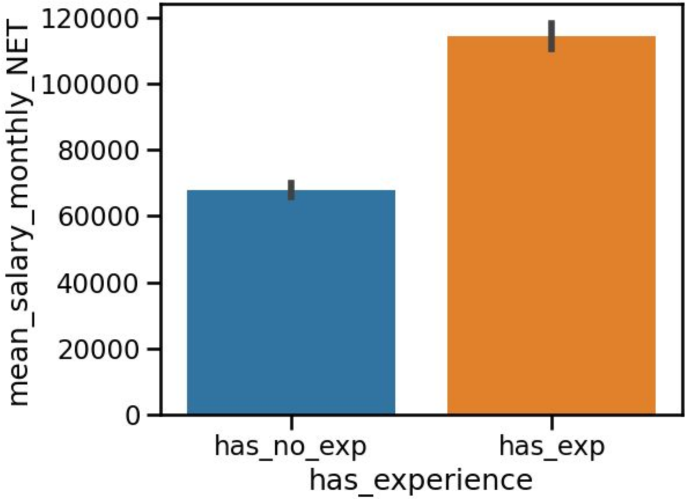
Mark **equally sized intervals** on a **horizontal** measurement axis. **Above each interval**, draw a rectangle whose **height is the corresponding count** (or relative frequency, density, percent, etc.).



IMPORTANT! We can only use this plot with numerical r.v.

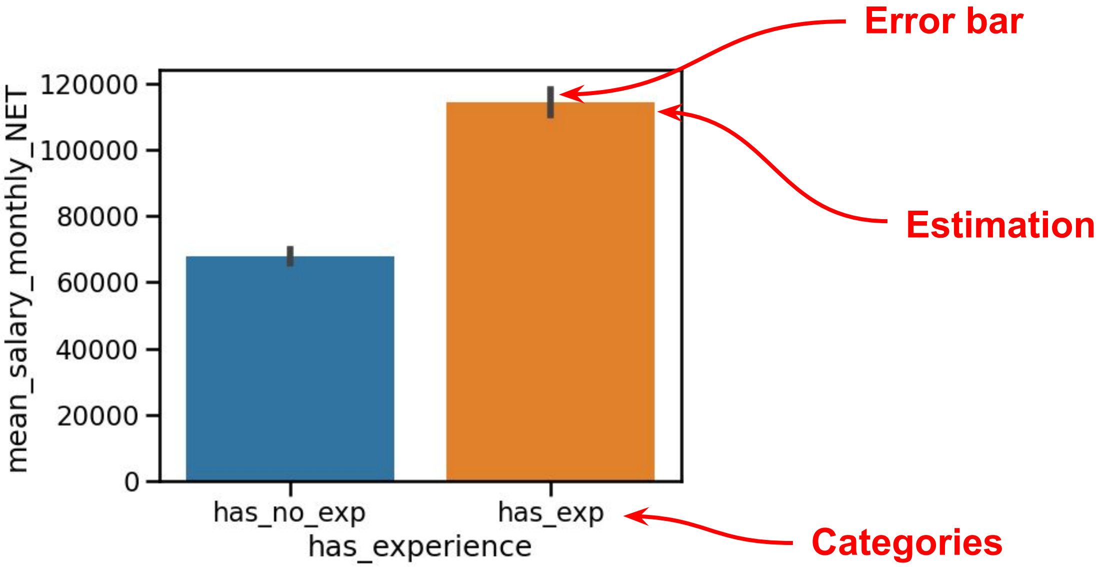
Basic Plots: Barplot

It represents an **estimate of central tendency for a numeric variable** with the height of each rectangle and provides some indication of the **uncertainty around that estimate** using error bars.



Basic Plots: Barplot

It represents an **estimate of central tendency for a numeric variable** with the height of each rectangle and provides some indication of the **uncertainty around that estimate** using error bars.



Error bar

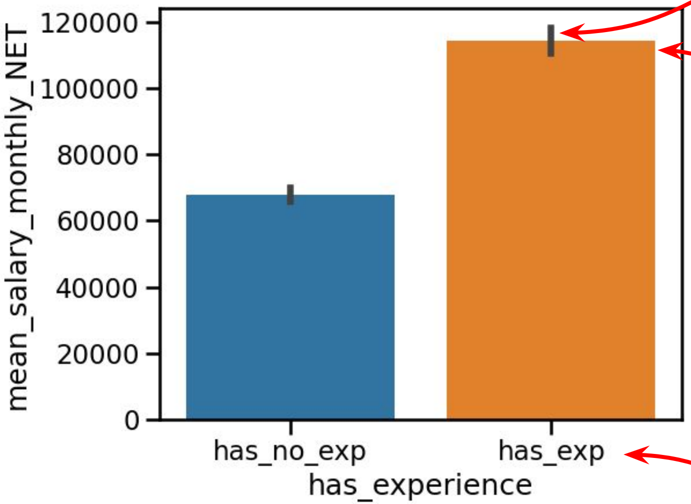
Estimation

Categories

Basic Plots: Barplot

It represents an **estimate of central tendency for a numeric variable** with the height of each rectangle and provides some indication of the **uncertainty around that estimate** using error bars.

```
seaborn.barplot(
    data=df,
    x="has_experience",
    y="salary_monthly_NET")
```



Error bar

Estimation

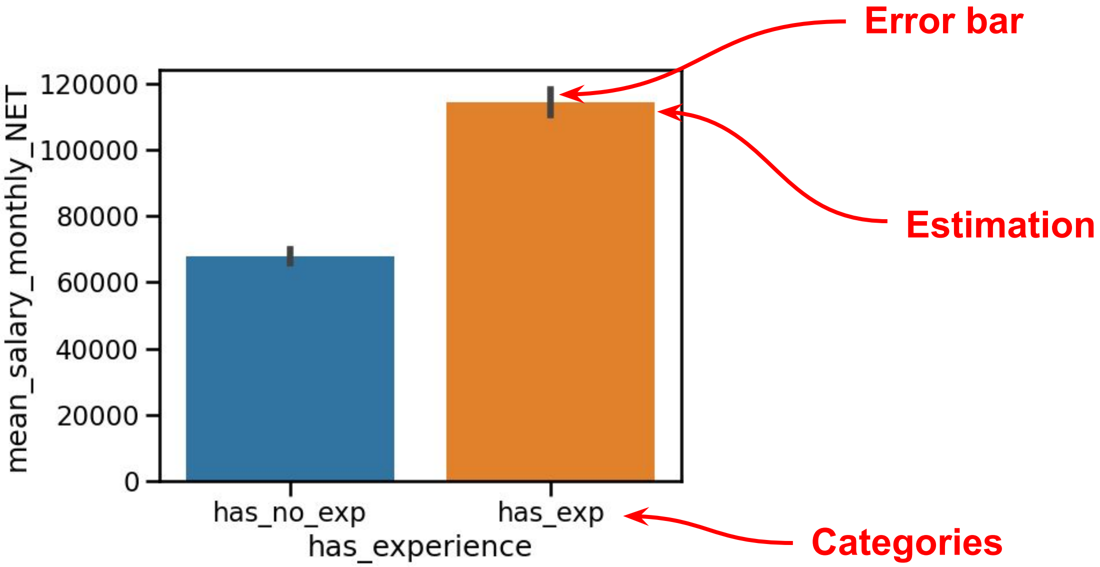
Categories

Basic Plots: Barplot

It represents an **estimate of central tendency for a numeric variable** with the height of each rectangle and provides some indication of the **uncertainty around that estimate** using error bars.

```
seaborn.barplot(
     data=df,
     x="has_experience",
     y="salary_monthly_NET")
```

IMPORTANT! We can only use this plot with numerical r.v. in combination with a categorical one.



Error bar

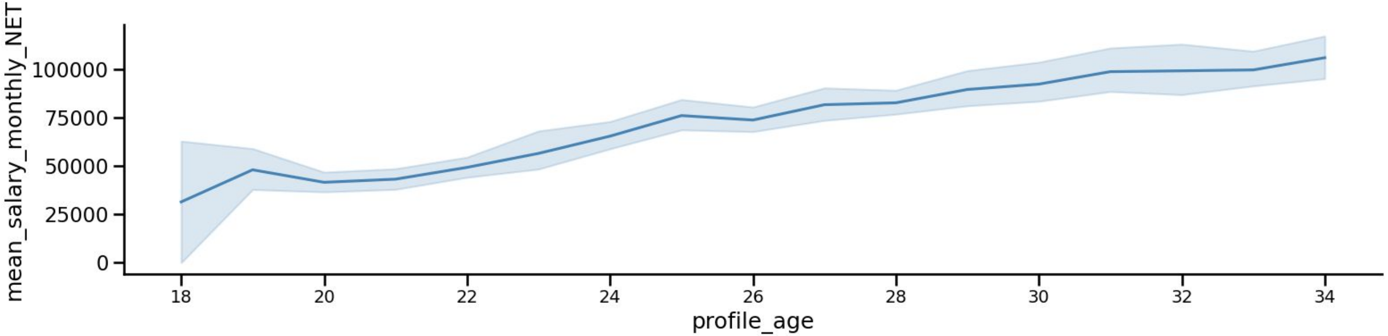
Estimation

Categories

Basic Plots: Lineplot

It is useful when you want to **understand changes in one variable as a function of time**, or a similarly continuous variable.

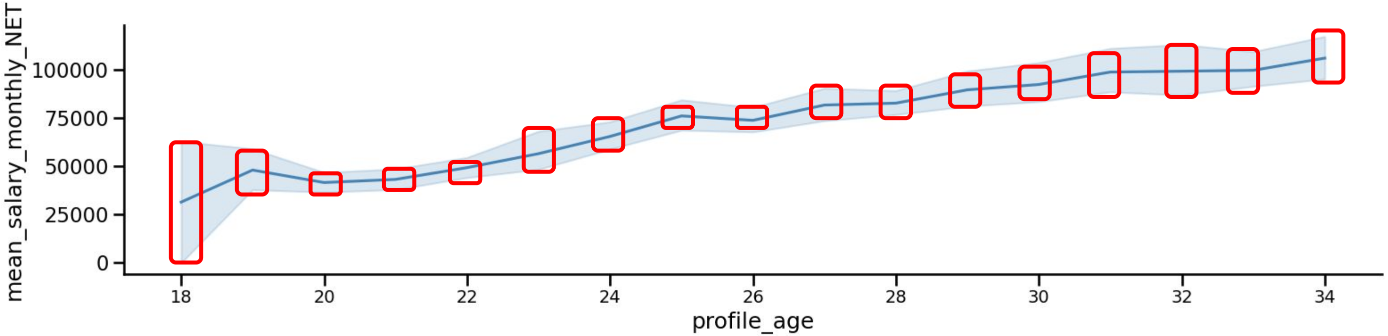
The plot **aggregates over multiple y values at each value of x** and shows an estimate of the central tendency and a confidence interval for that estimate.



Basic Plots: Lineplot

It is useful when you want to **understand changes in one variable as a function of time**, or a similarly continuous variable.

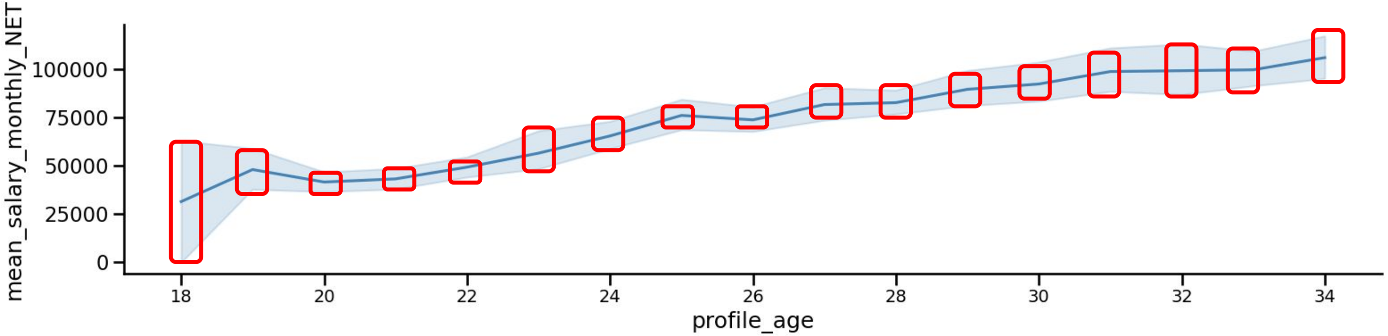
The plot **aggregates over multiple y values at each value of x** and shows an estimate of the central tendency and a confidence interval for that estimate.



Basic Plots: Lineplot

It is useful when you want to **understand changes in one variable as a function of time**, or a similarly continuous variable.

The plot **aggregates over multiple y values at each value of x** and shows an estimate of the central tendency and a confidence interval for that estimate.



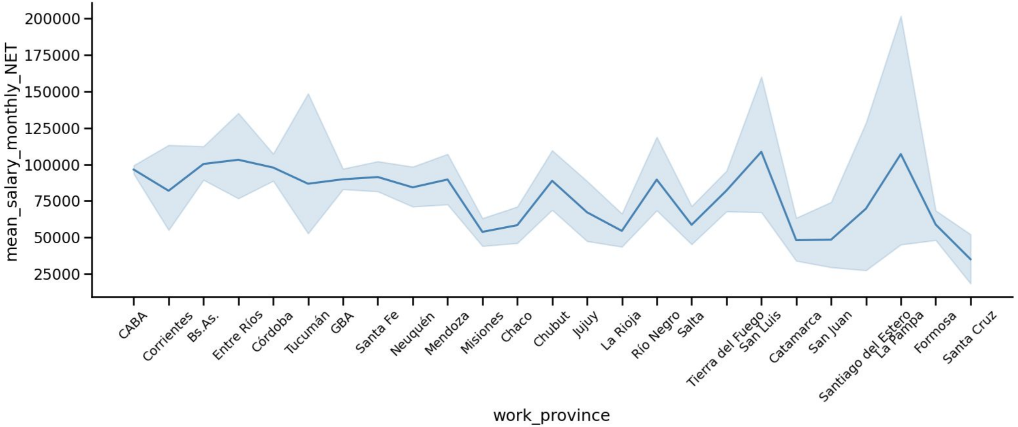
```
seaborn.lineplot(data=df, x="profile_age", y="salary_monthly_NET")
```

Basic Plots: Lineplot

It is useful when you want to **understand changes in one variable as a function of time**, or a similarly continuous variable.

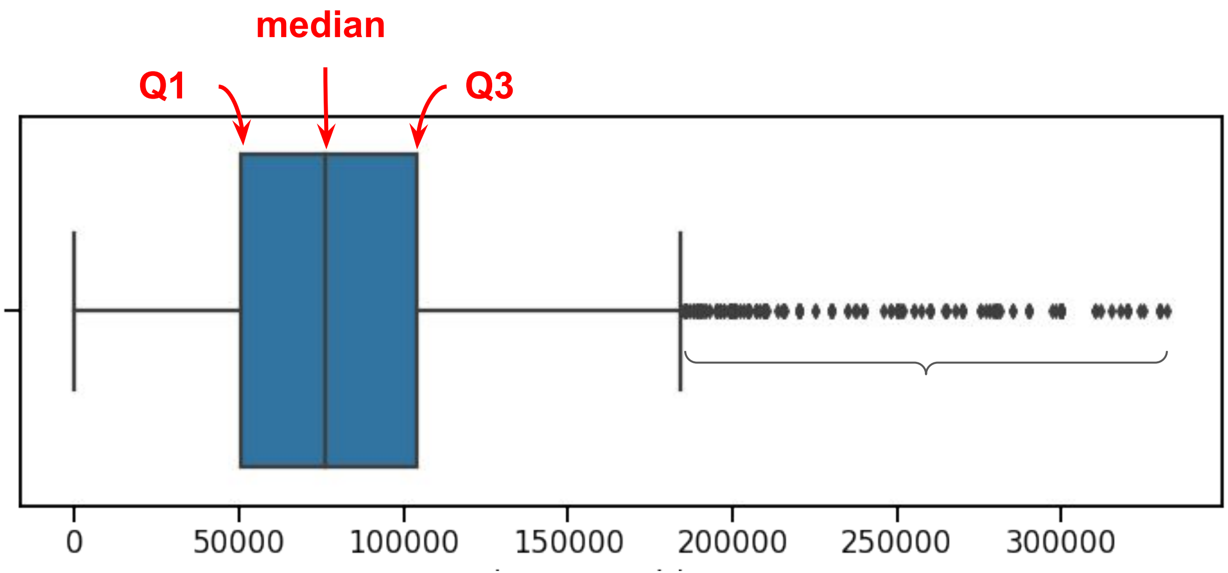
The plot **aggregates over multiple y values at each value of x** and shows an estimate of the central tendency and a confidence interval for that estimate.

IMPORTANT! Don’t use a categorical r.v. on the x axis.



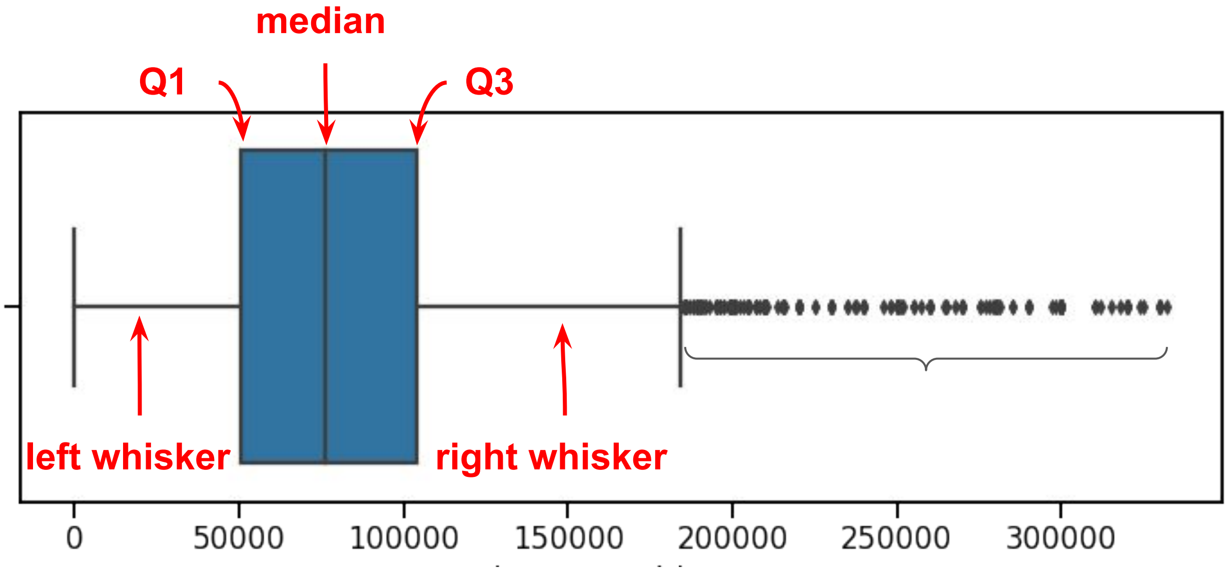
Basic Plots: Boxplot

The **box shows the quartiles** of the dataset while the **whiskers extend to show the rest of the distribution**, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.



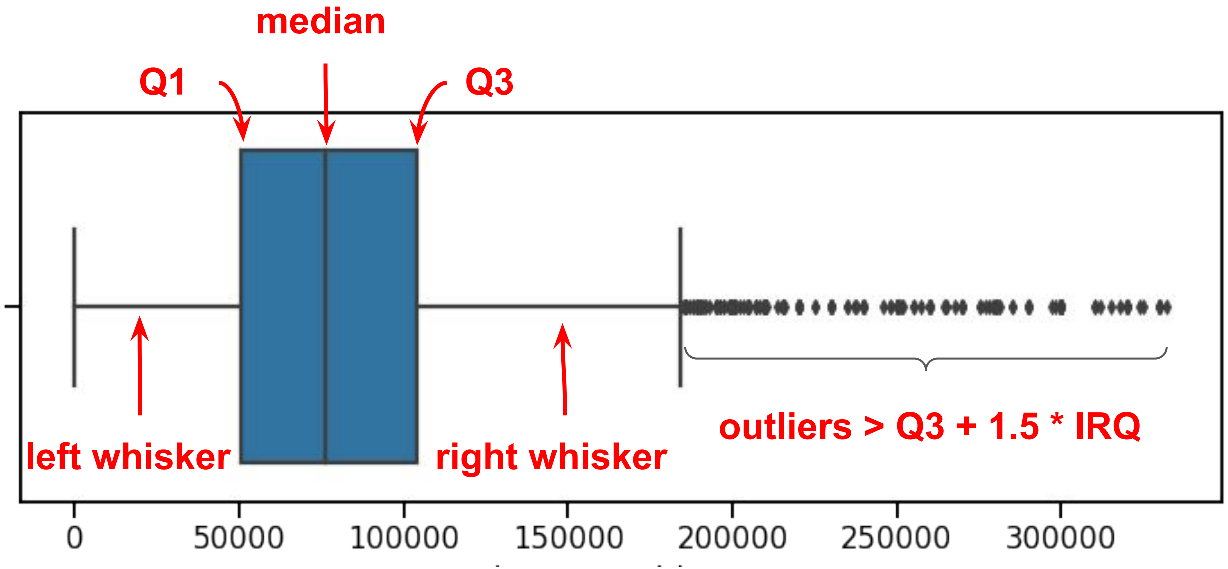
Basic Plots: Boxplot

The **box shows the quartiles** of the dataset while the **whiskers extend to show the rest of the distribution**, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.



Basic Plots: Boxplot

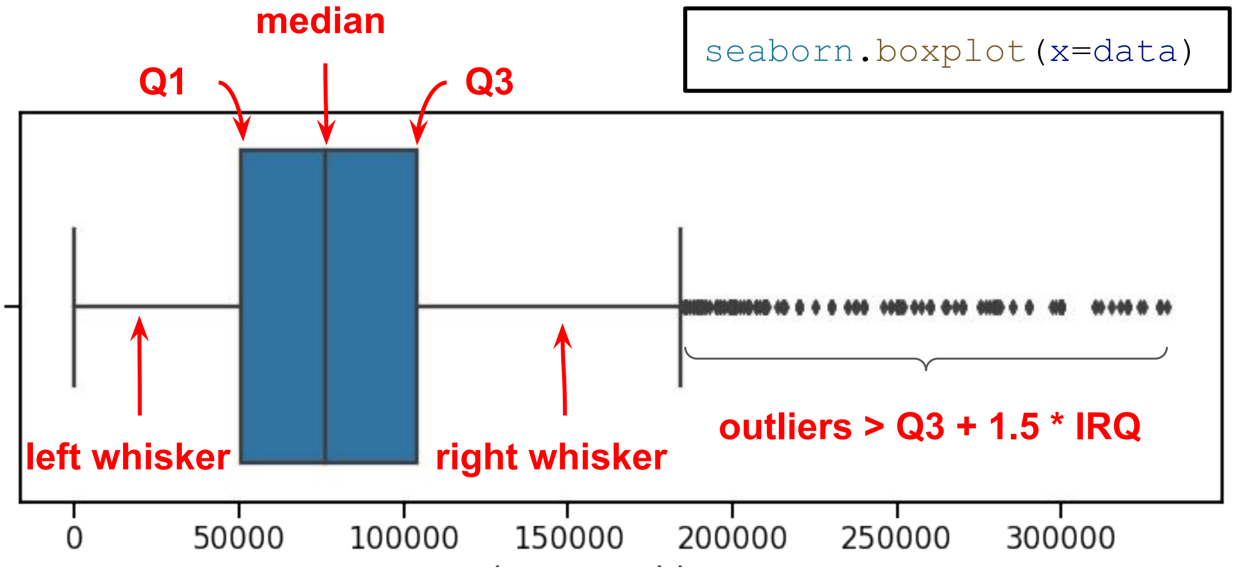
The **box shows the quartiles** of the dataset while the **whiskers extend to show the rest of the distribution**, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.



Basic Plots: Boxplot

The **box shows the quartiles** of the dataset while the **whiskers extend to show the rest of the distribution**, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.

```
seaborn.boxplot(x=data)
```



Probabilities

A probability **P** is a function **takes an state space Ω** and **returns a real number** between 0 and 1. At the same time, it has to **hold some properties**. Basically, for each subset A of Ω , P(A) is a number such as:

Probabilities

A probability **P** is a function **takes an state space Ω** and **returns a real number** between 0 and 1. At the same time, it has to **hold some properties**. Basically, for each subset A of Ω , P(A) is a number such as:

- $\mathbf{0} \leq \mathbf{P}(A) \leq 1$
- $\mathbf{P}(\boldsymbol{\Omega}) = 1$
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$, for A and B disjoints
- $\mathbf{P}(\cup_i A_i) = \sum_i \mathbf{P}(A_i)$ for A_1, A_2 ,... disjoints

Probabilities

A probability **P** is a function **takes an state space Ω** and **returns a real number** between 0 and 1. At the same time, it has to **hold some properties**. Basically, for each subset A of Ω , P(A) is a number such as:

- $\mathbf{0} \leq \mathbf{P}(A) \leq 1$
- $\mathbf{P}(\mathbf{\Omega}) = 1$
- $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$, for A and B disjoints
- $\mathbf{P}(\cup_i A_i) = \sum_i \mathbf{P}(A_i)$ for A_1, A_2 ,... disjoints

Events can be thought as **restrictions applied to one or several r.v.** Conditional probability between the two events is defined as:

$\mathbf{P}(A|B) = \mathbf{P}(A \text{ and } B) / P(B)$

$\mathbf{P}(A|B) = |A \text{ and } B| / |B|$

Common Operations on Dataframes

We can apply certain operations on a dataframe. The simplest ones are **projections** and **filterings**.

Common Operations on Dataframes

We can apply certain operations on a dataframe. The simplest ones are **projections** and **filterings**.

Projections: Put in brackets the name of the column we want to project.

```
df["profile_gender"], df["profile_age"], df[["profile_gender", "profile_age"]]
```

Common Operations on Dataframes

We can apply certain operations on a dataframe. The simplest ones are **projections** and **filterings**.

Projections: Put in brackets the name of the column we want to project.

```
df["profile_gender"], df["profile_age"], df[["profile_gender", "profile_age"]]
```

Filterings: Create a Pandas Series of booleans and give it as input to a dataframe of the same shape.

```
df[
    (df["profile_gender"] == "Male") &
    (df["profile_age"] < 30)
]
```

Condition to filter