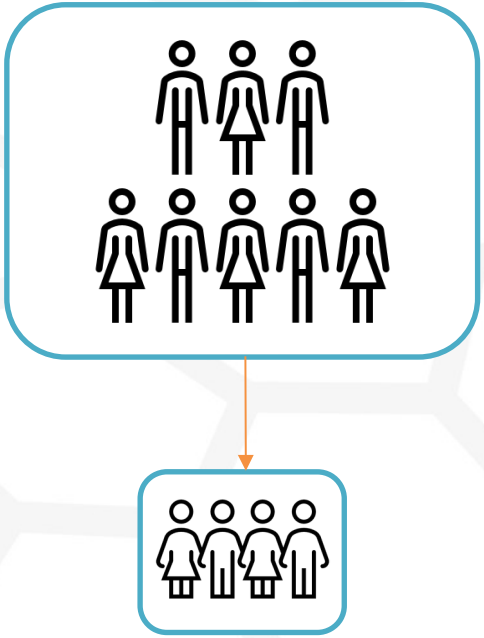
Intro to Stats

Population, Sample & Observation

- The units on which we measure data—such as people, cities, animals — are called **observations**.
- The collection of *all* observations that we are interested in is called a **population.**
- If we consider a selection of observations, then these observations are called a **sample**.



Population
All students
at
a college

Sample
A few selected students from the college

Summary Statistics

- A statistic is a measure of some attribute of sampled data. They can be used to concisely describe features of a dataset.

- The choice of summary statistic depends on whether we are looking at numerical or categorical data.

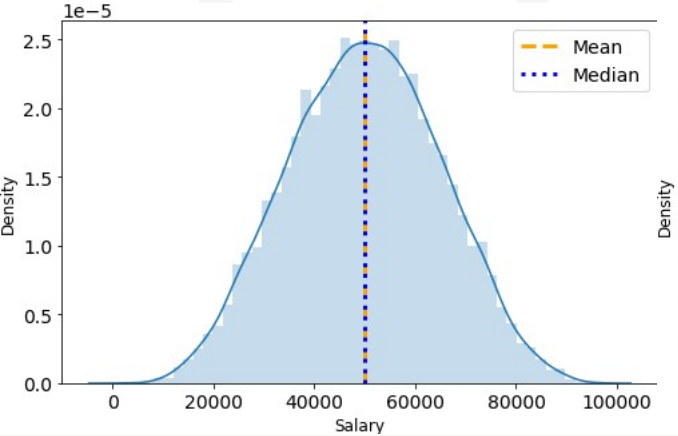
Summary Statistics - Measures of Central Tendency
Numerical Data (Univariate)

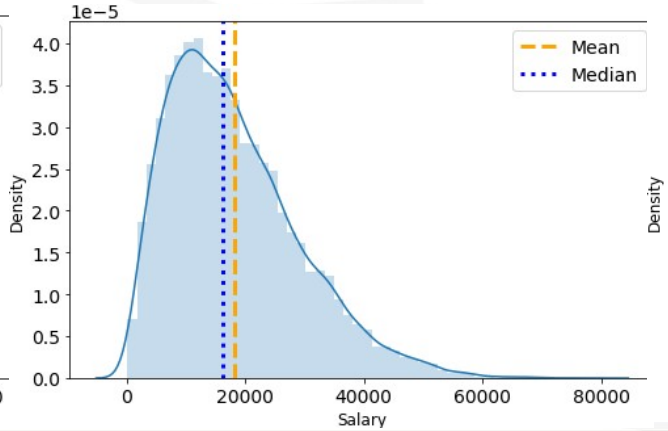
- **Mean -** The sum of the data values divided by the number of observations $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

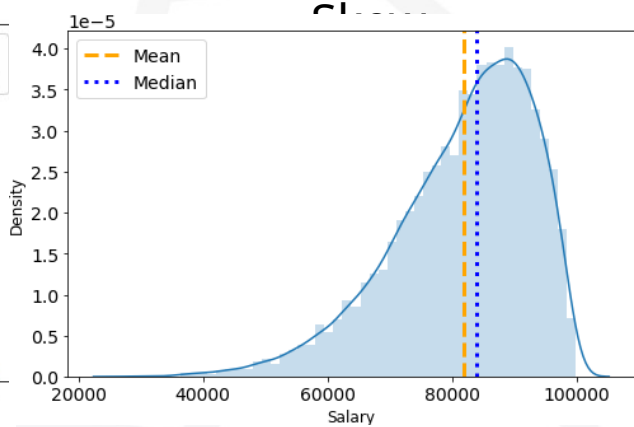
- **Median -** The 50th percentile of the data (the point below which 50% of the observations fall)

Positive Skew

Negative Skew

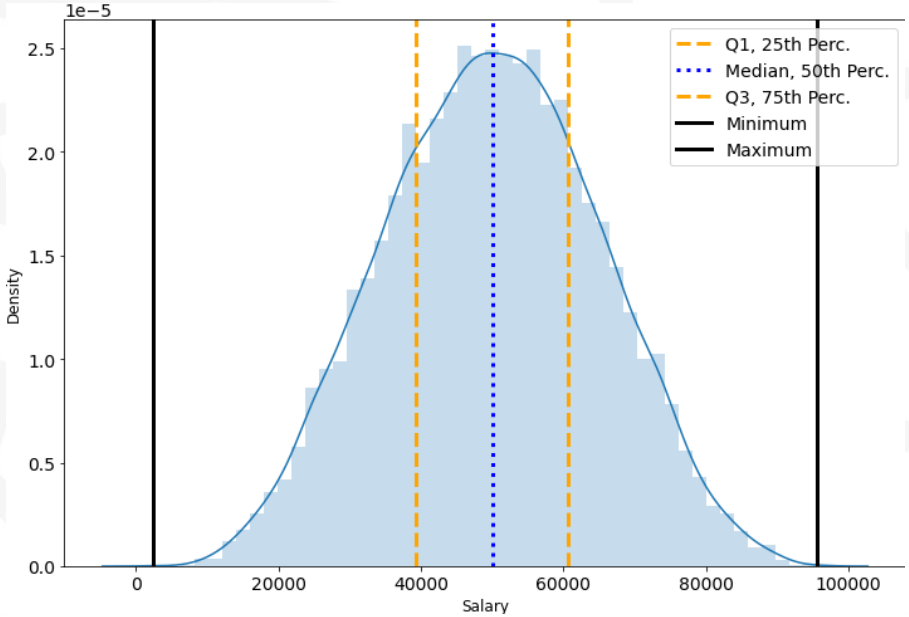






Summary Statistics - Percentiles

A **percentile** is a value at which a certain percentage of the data fall below. For example, the 75th percentile is the value at which 75% of the data fall below.



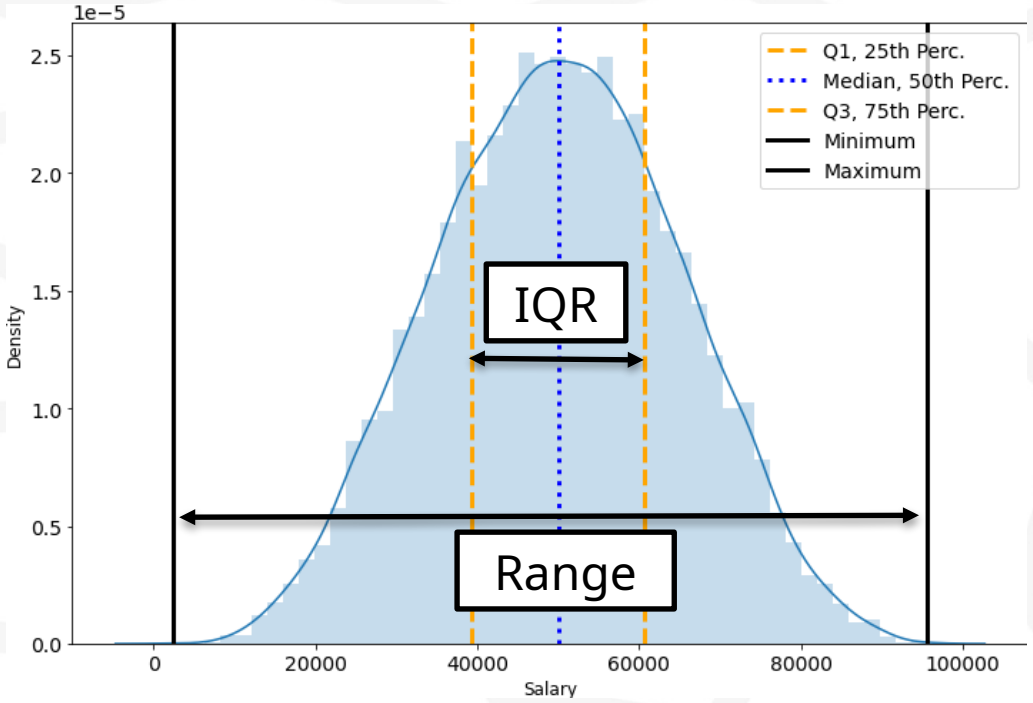
Certain percentiles have special names:

- The *25th percentile* is also called the 1st quartile.
- The *50th percentile* is also called the 2nd quartile or the median.
- The *75th percentile* is also called the 3rd quartile.

Summary Statistics - Measures of Spread (1/2) Numerical Data (Univariate)

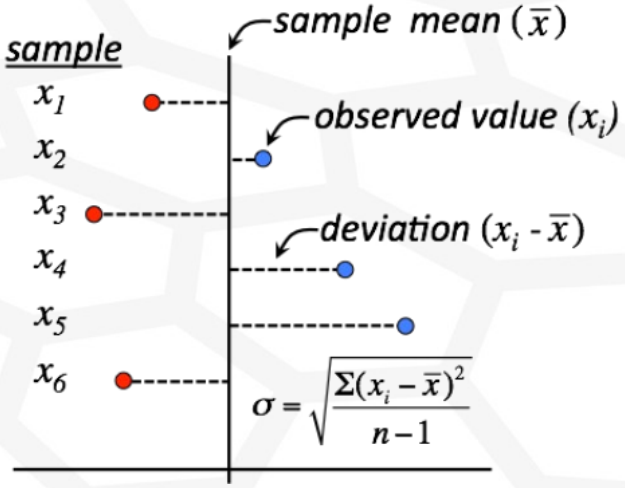
- **Range -** the difference between the minimum and maximum of the data

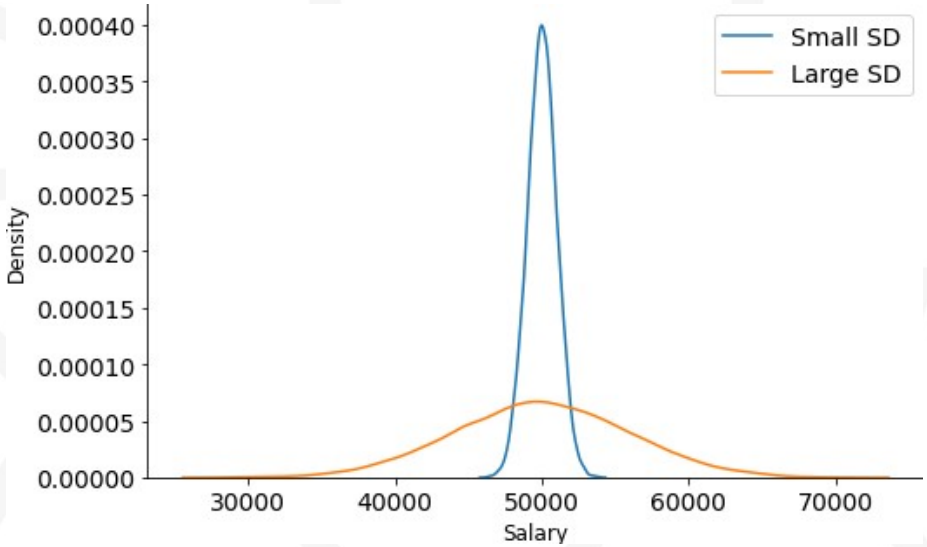
- **Interquartile Range -** the difference between the 1st and 3rd quartiles of the data.



Summary Statistics - Measures of Spread (2/2)
Numerical Data (Univariate)

Standard Deviation - The sum of the data values divided by the number of observations





Example

Say we collected the following test scores from 10 students:

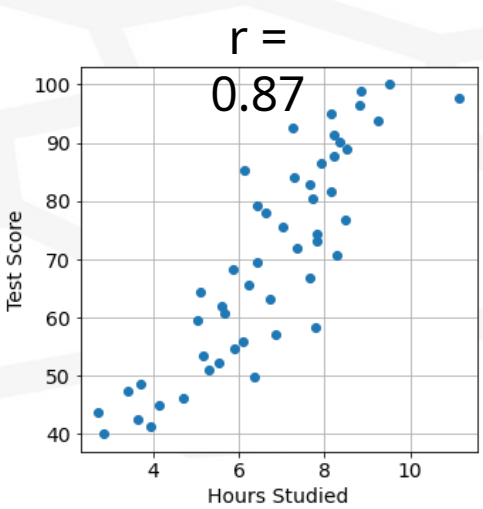
85, 90, 92, 73, 95, 60, 89, 78, 99, 90

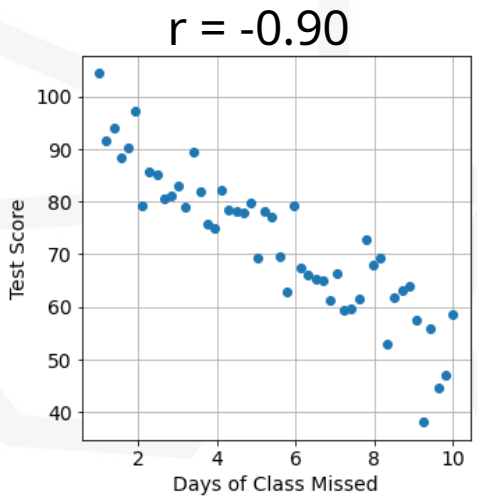
Calculate the following:

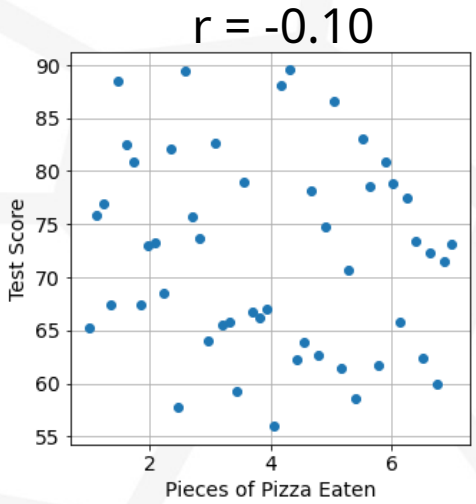
- Mean
- Median
- Range

Summary Statistics – Correlation Numerical Data (Bivariate)

Correlation between **two continuous variables** can be thought of as the strength of the linear relationship between the variables. The correlation coefficient is a number between -1 and +1 that quantifies the strength and direction of the relationship.







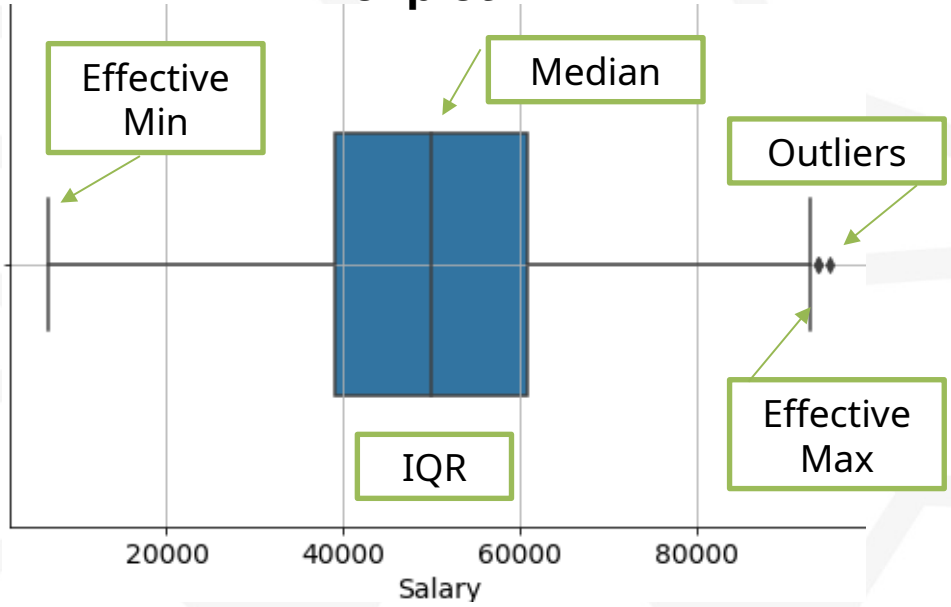
Summary Statistics – Categorical Data

Frequency
Relative Frequency
Percentage

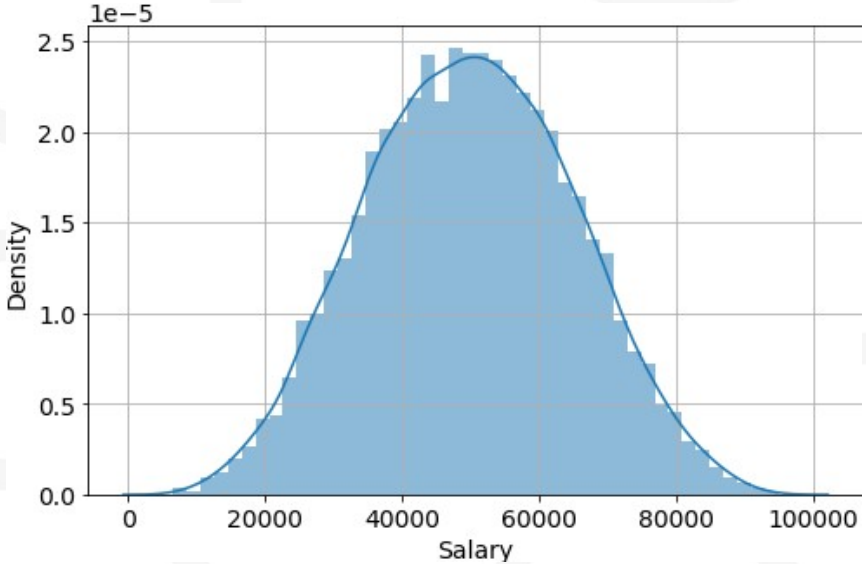
Degree	Frequency	Relative Frequency	Percentage
High School	2	0.050	5.0
Bachelor's	7	0.175	17.5
MBA	20	0.500	50.0
Master's	3	0.075	7.5
Law	4	0.100	10.0
PhD	4	0.100	10.0
	40		

Visualizations – Numerical Data (Univariate)

Boxplot

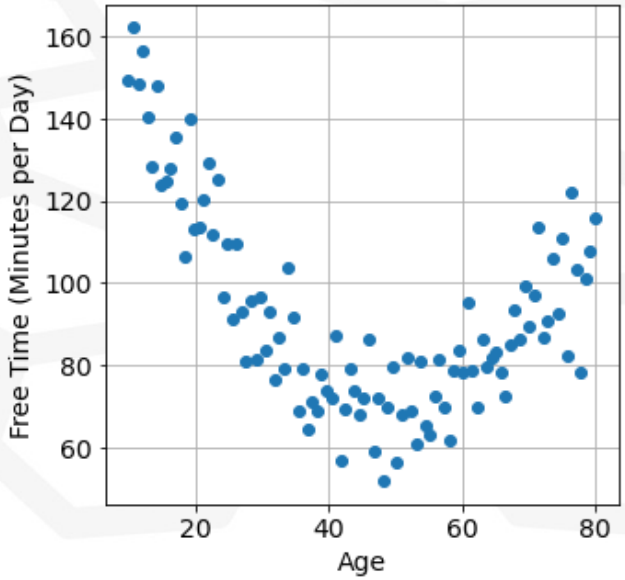


Histogram

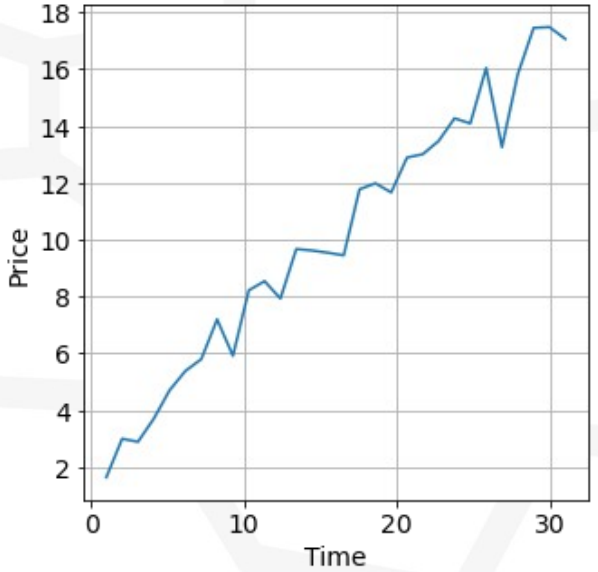


Visualizations – Numerical (Bivariate)

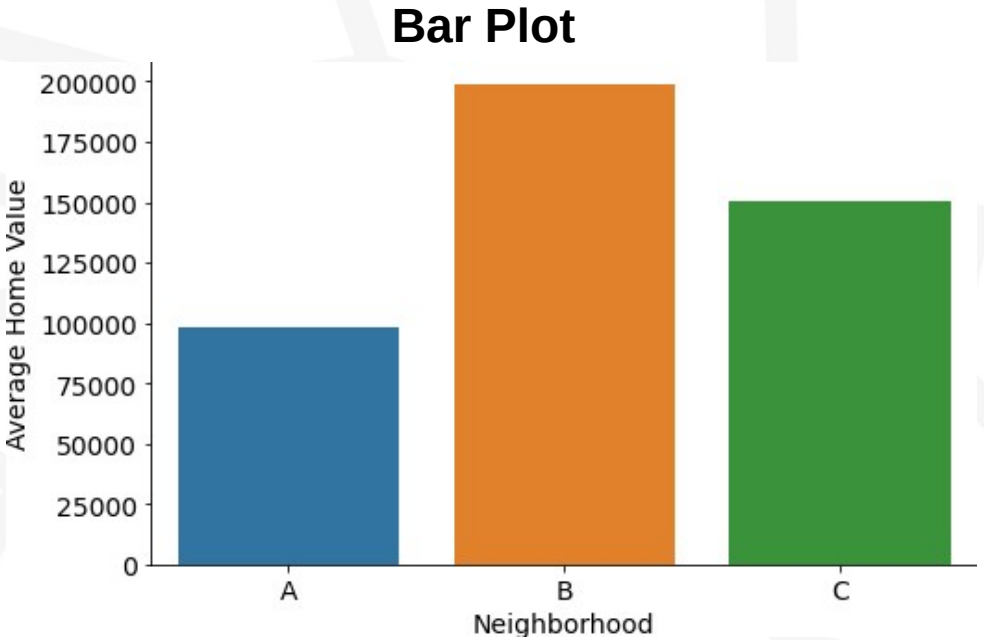
Scatter Plot



Line Plot



Visualizations – Categorical Data (Univariate)



Things to Look Out for In Visualizations