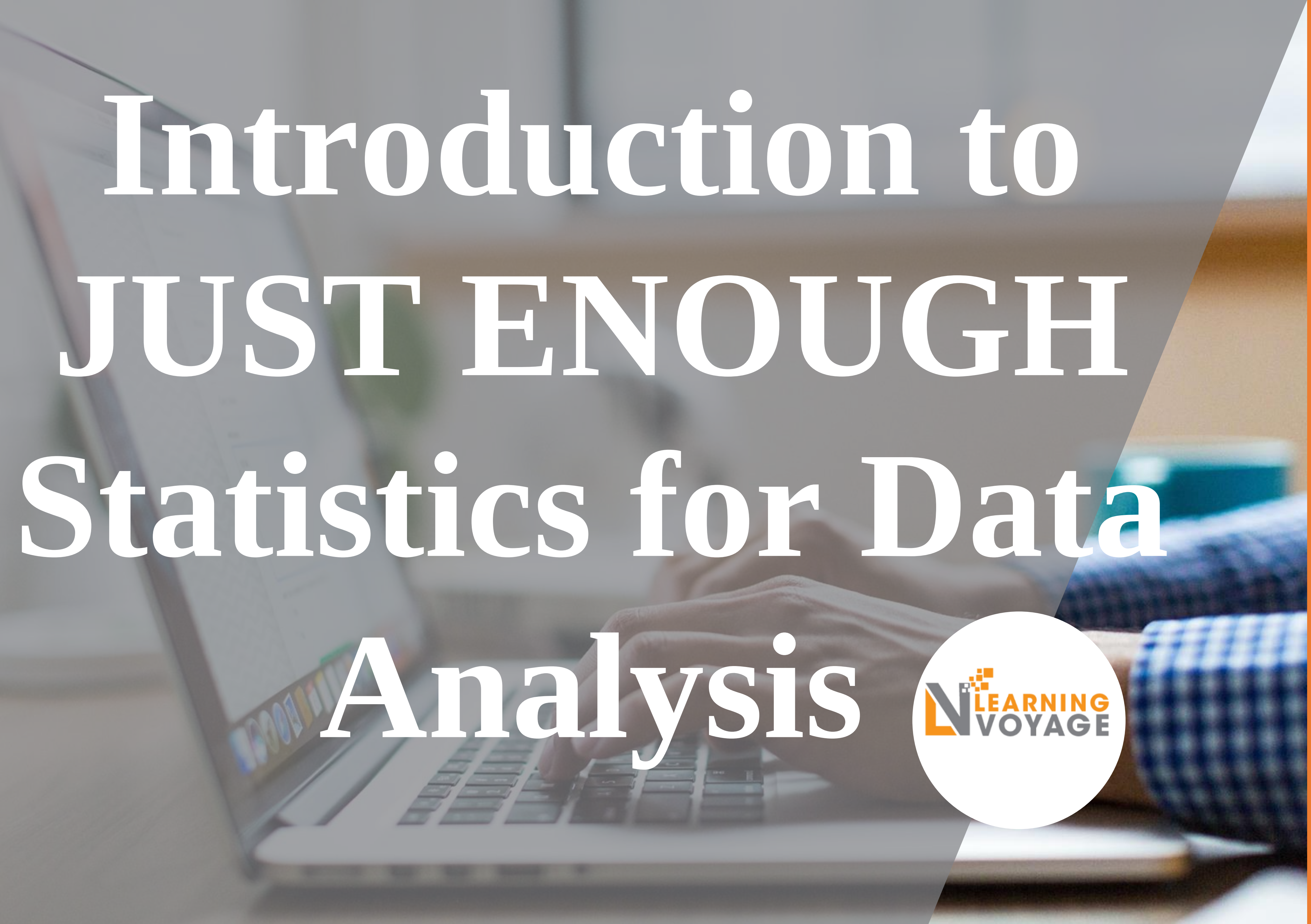
Introduction to JUST ENOUGH Statistics for Data Analysis



LEARNING VOYAGE

Theory

What is Statistics?

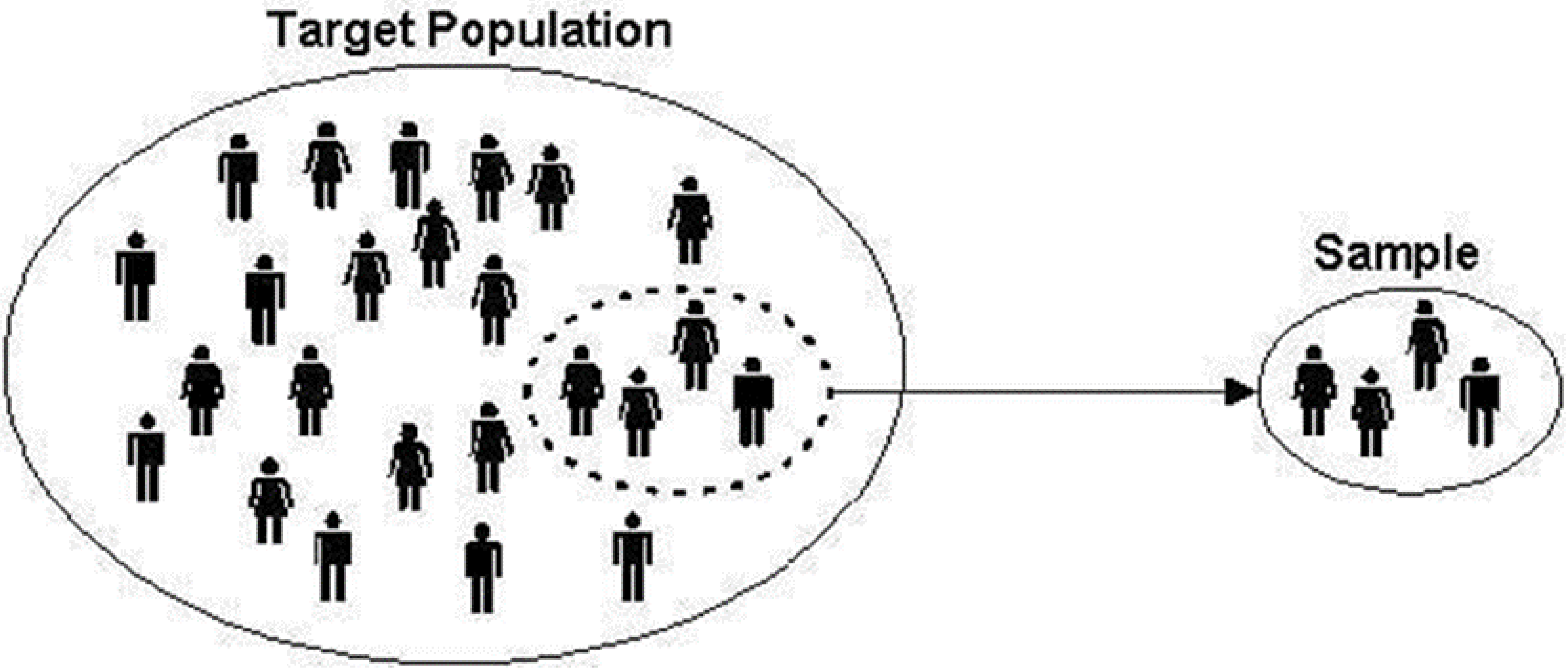
Statistics is “A telescope that allows us to study the large terrain and make it accessible to our unaided vision”

Statistics – Big Picture

Statistics provides a way of organizing data to extract information on a wider and objective basis than relying on personal experience. It is a branch of mathematics working with

- Data Gathering
- Data Understanding
- Data Analysis/Interpretation
- Data Presentation

Basic Statistical Terminology



Parameter and Statistic

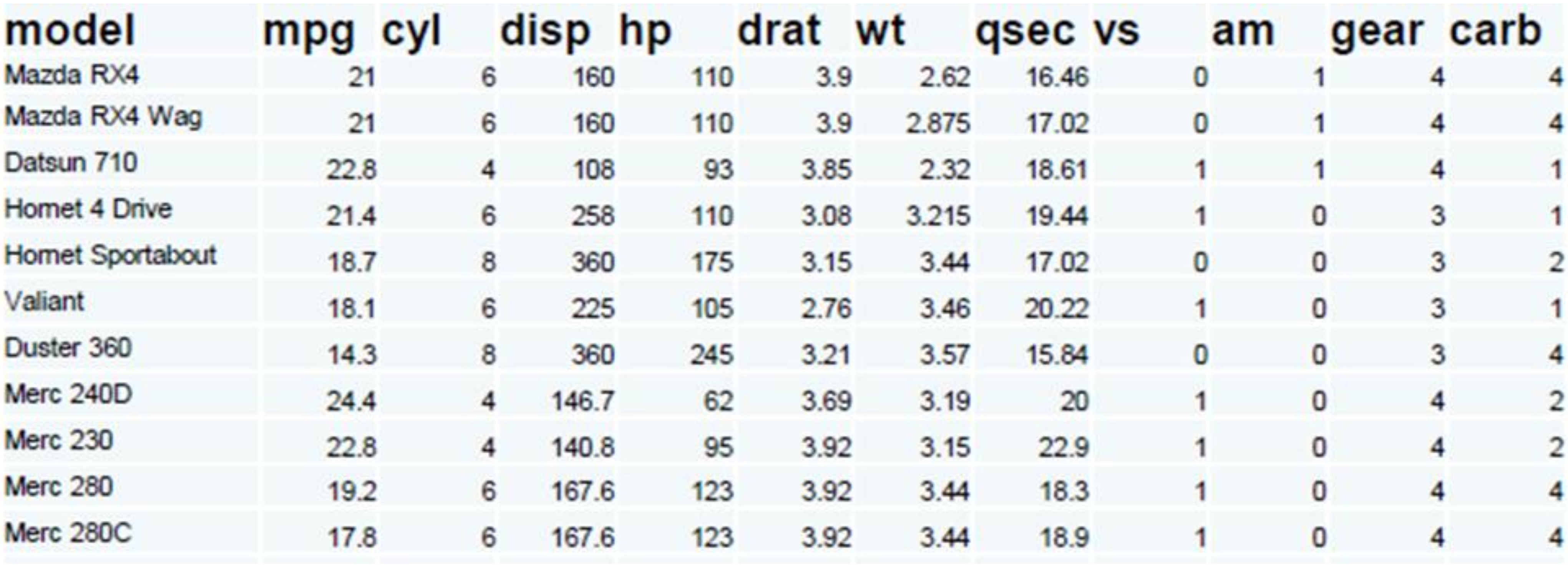
Parameter: A descriptive measure of the population. For example,

- population mean - μ
- population variance – σ2
- population standard deviation - σ

Statistic: A descriptive measure of the sample. For example,

- sample mean - xbar
- sample variance - s2
- sample standard deviation – s

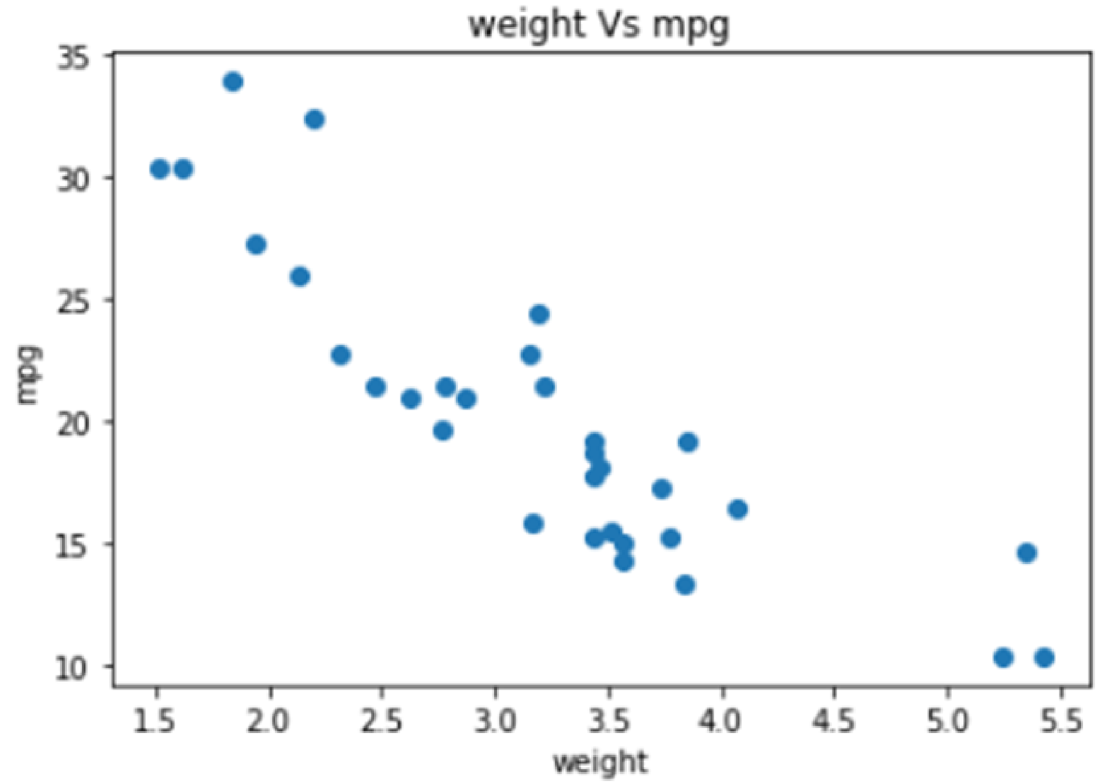
Variables and Data (Example of data)



model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Homet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Homet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4

Variables – Dependent and Independent

- An independent variable (experimental or predictor) is a variable that is being manipulated in an experiment in order to observe the effect on dependent variable (Outcome).



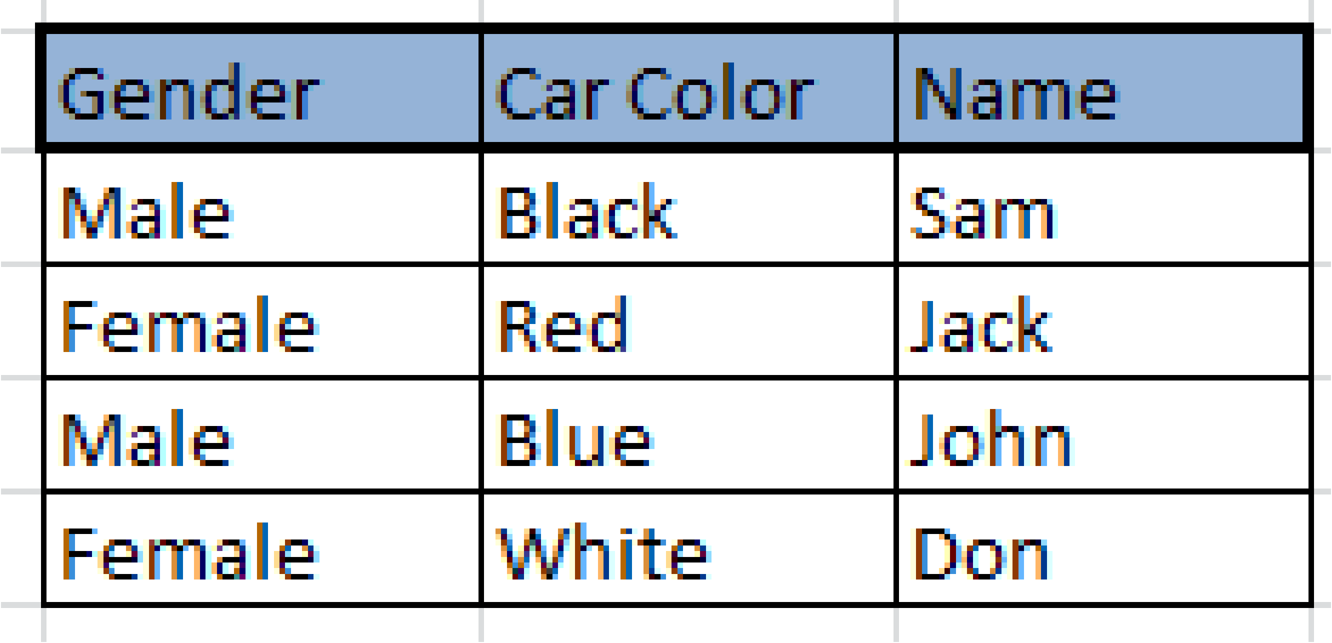
Data

Data is classified into two types Numerical and Categorical

- Categorical Data
- Numerical Data

Levels of Measurement Scales

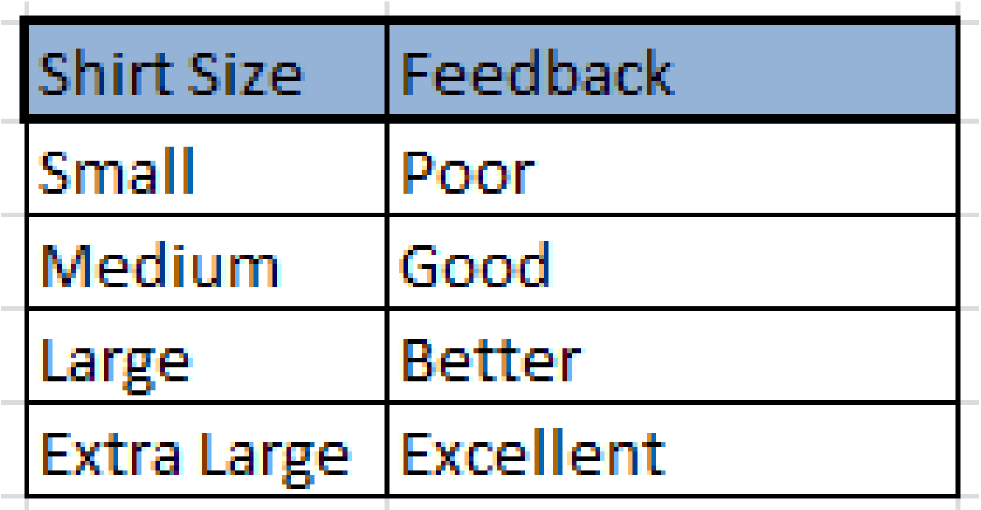
- **Nominal scale:** The nominal scale could simply be called “labels



Gender	Car Color	Name
Male	Black	Sam
Female	Red	Jack
Male	Blue	John
Female	White	Don

Levels of Measurement Scales

- **Ordinal scale:** The order of the values is what’s important and significant, but the difference between each one is not really known. Here are some examples, below



Shirt Size	Feedback
Small	Poor
Medium	Good
Large	Better
Extra Large	Excellent

Descriptive Statistics

- Descriptive statistics involves organizing, summarizing, and presenting data in an informative way.

- Descriptive statistics, unlike inferential statistics, seeks to describe the data, but does not attempt to make inferences from the sample to the whole population.

Different types of Descriptive Statistics

Descriptive statistics are broken down into two categories

- Measure of Central Tendency

- Measure of Variability (Spread)

Mean:

mean
median
mode

- Mean is a central tendency of the data i.e. a number around which a whole data is spread out.

- Formula for sample mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mean:

- Similarly, for a population data of size N, the population mean is:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Median:

Median is the value which divides the data into 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either ascending or descending order.

- May not exist as a data point in the set
- Influenced by position of items, but not their values
- Median is not influenced by extreme values

Mode

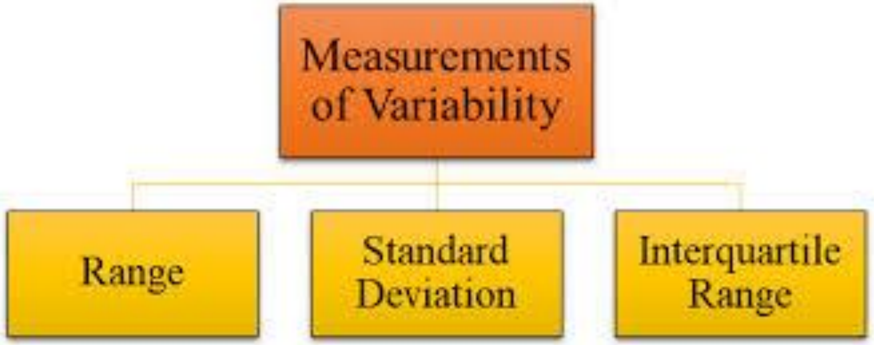
mean
median
mode

Mode: Mode is the most commonly occurring value

- Mode exists as a data point.
- Useful for qualitative data.

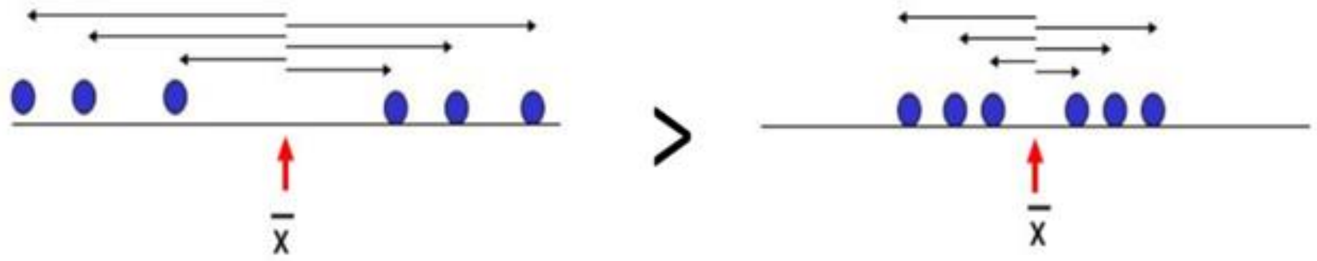
Measure of Variability (Spread / Dispersion)

- The measures that help us to know about the spread of a data set are called measures of dispersion.



Measure of Variability (Spread / Dispersion)

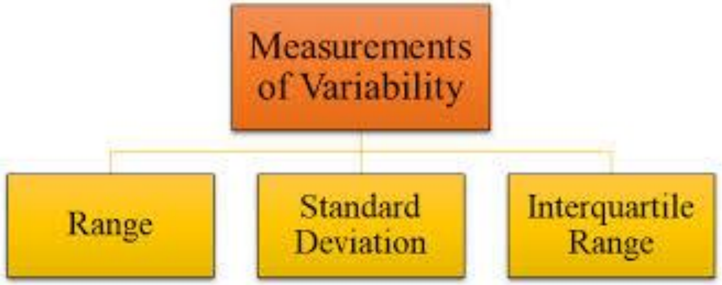
- **Standard deviation:** Standard deviation is the measurement of average distance between each quantity and mean, That is, how data is spread out from mean.
- A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.



Measure of Variability (Spread / Dispersion)

- Sample Standard Deviation is denoted by “S”

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



Measure of Variability (Spread / Dispersion)

- Population Standard Deviation is denoted by “σ” (sigma)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \, .$$

Variance

- Variance is a square of average distance between each quantity and mean.

- That is, it is a square of standard deviation.

$$\text{Variance} = (S.D.)^2$$

Variance

The variance of Population and Sample

- The variance of a population is:

Population Mean

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Population Size

- The variance of a sample is:

Sample Mean

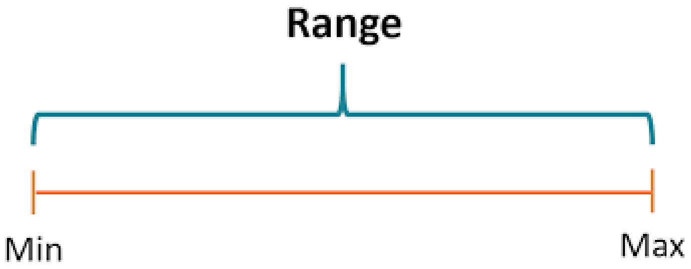
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Note! the denominator is sample size (n) minus one !

Range:

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

- It is easy to calculate.
- It is implemented for both “best” or “worst” case scenarios.
- Too sensitive for extreme values.



Levels of Measurement Scales

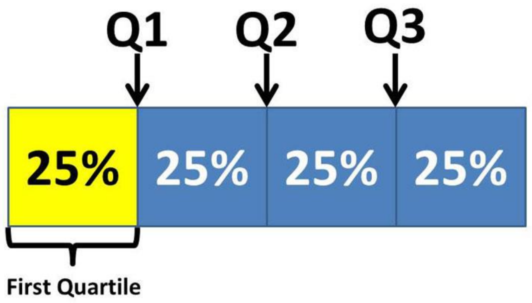
- **Percentile:** Percentile is a way to represent the position of a value in a data set.

- To calculate percentile, values in the data set should always be in ascending order.

Example:
12, 24, 41, 51, 67, 67, 85, 99

Quartile:

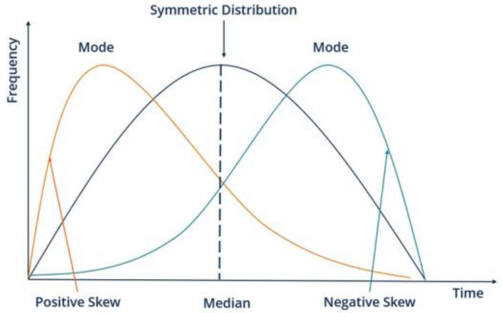
- In statistics and probability, quartile are values that divide your data into quarters provided data is sorted in an ascending order.



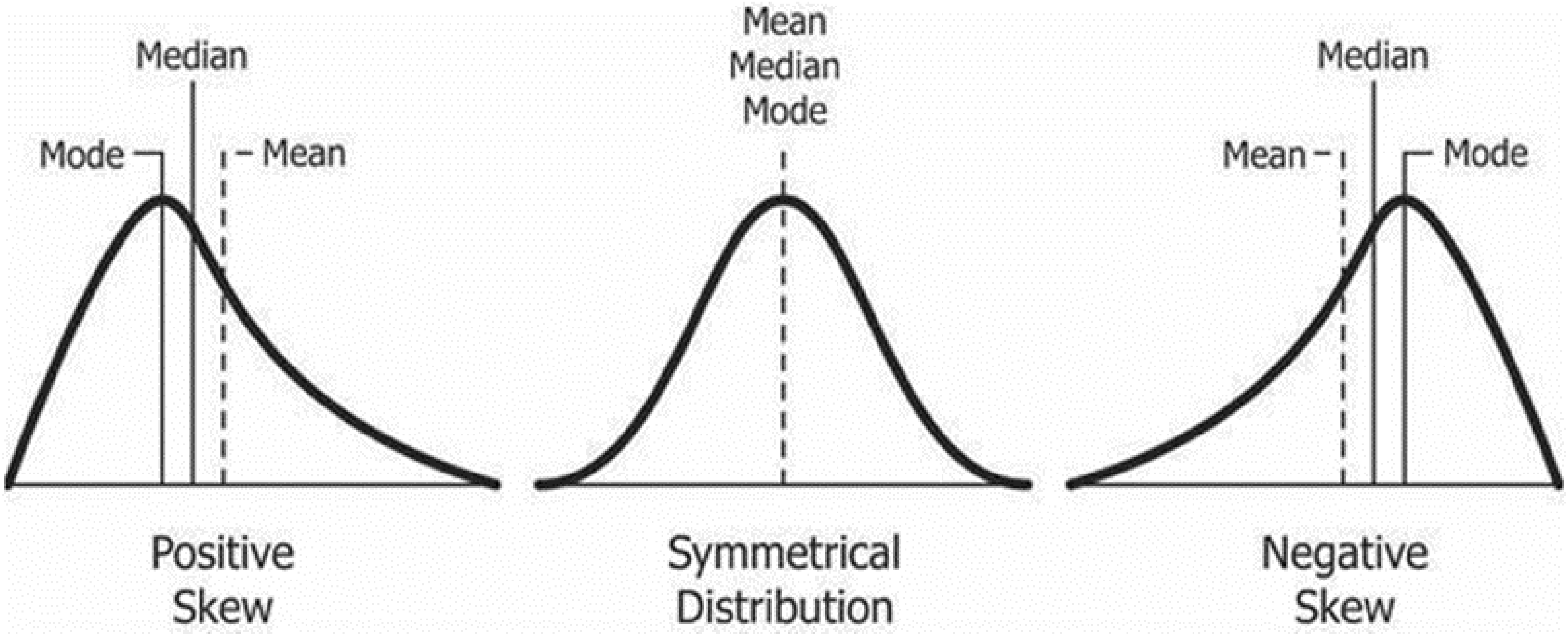
Skewness:

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

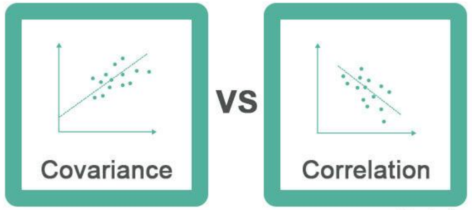
- The skewness value can be positive or negative or undefined.



Skewness:


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Covariance and Correlation



www.wallstreetmojo.com

- Covariance studies the direction between two continuous variables and Correlation studies the direction and strength between two continuous variables and helps in understanding how strongly those two continuous variables are associated with each other.

What is Covariance Matrix?

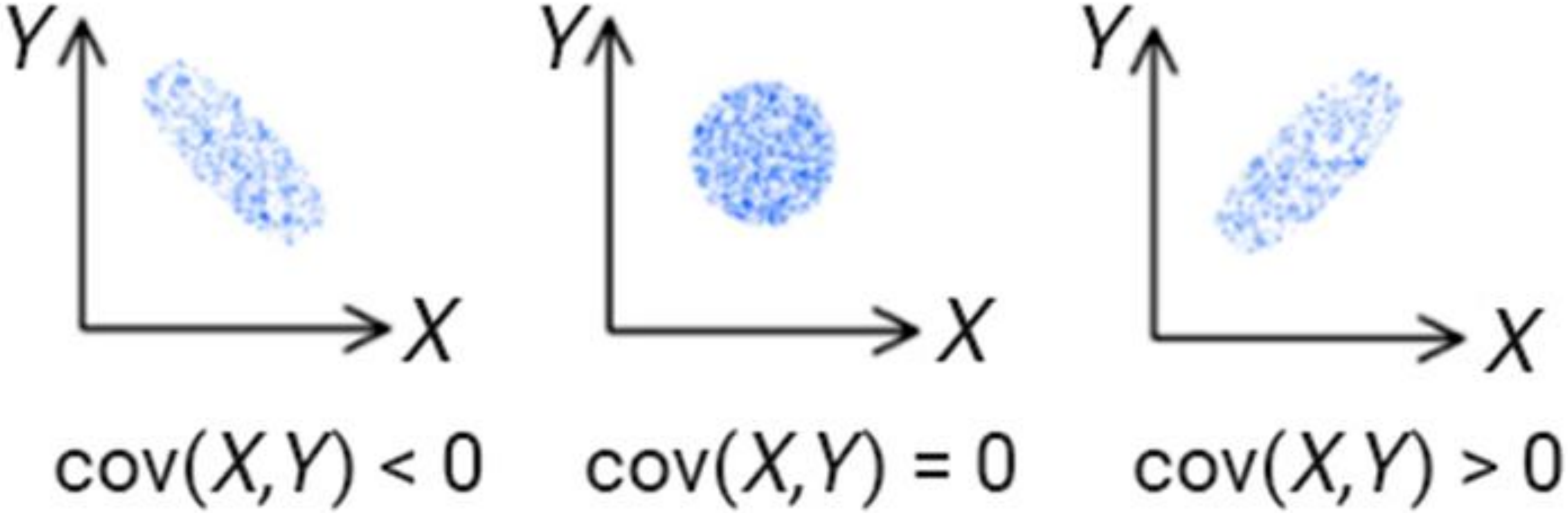
- Suppose we have two variables X and Y, then the covariance between these two variables is represented as Cov (X,Y).

- If ∑(X) and ∑(Y) are the expected values of the variables, the covariance formula can be represented as:

$$\mathbf{COV(X,\ Y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))$$

What is Covariance Matrix?

- Here are some plots that highlight how the covariance between two variables could look like in different directions.

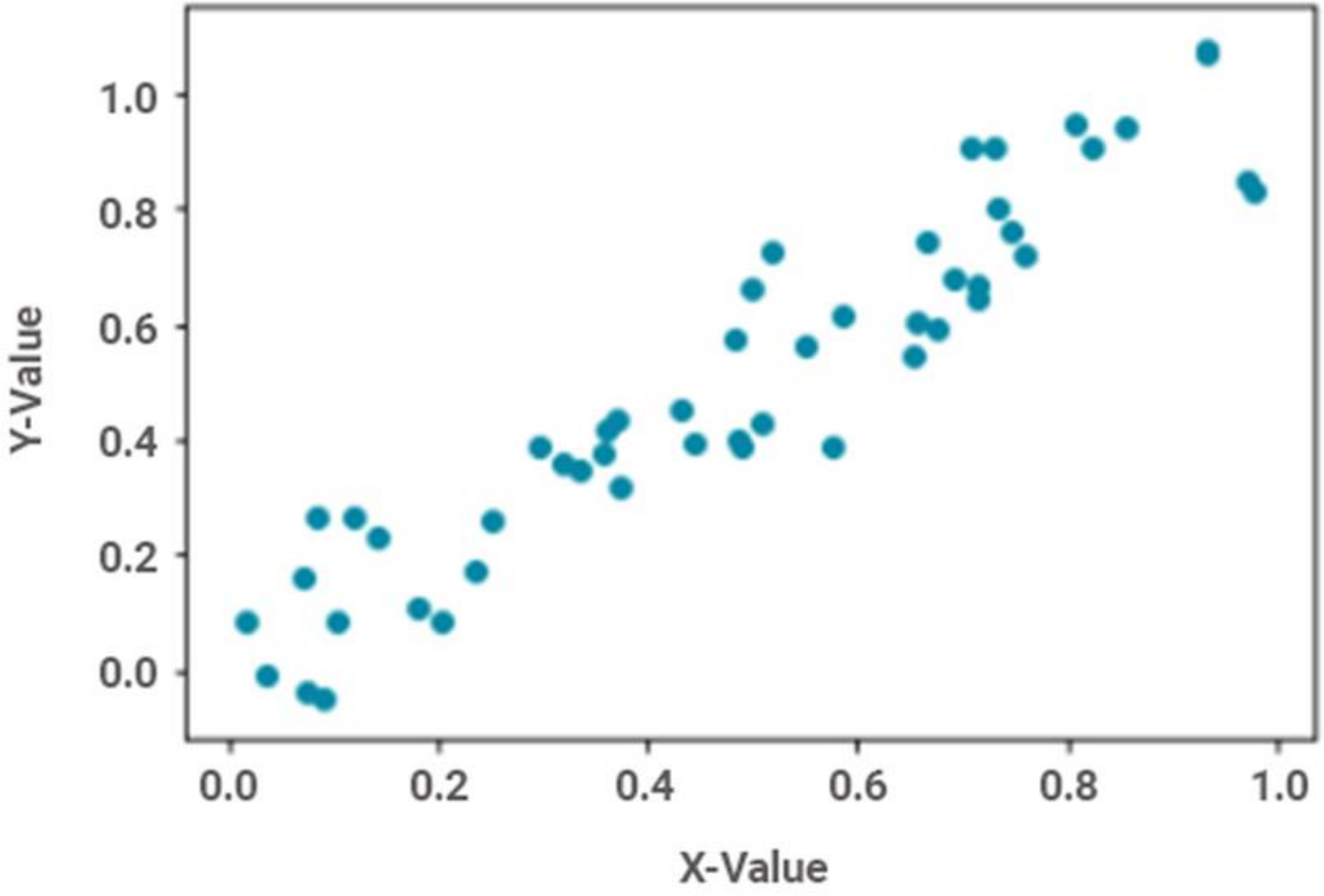


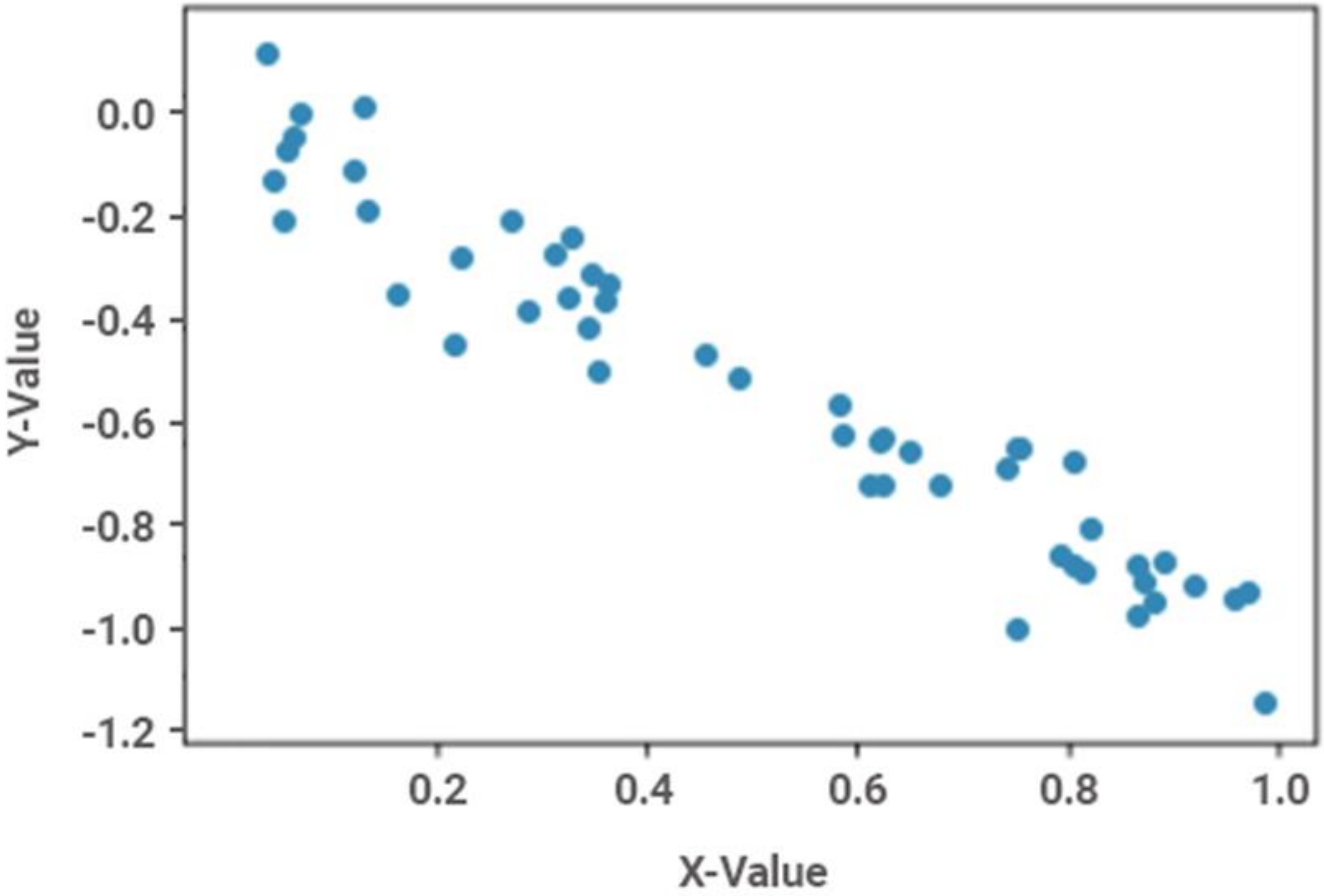
What is a Correlation Matrix?

- A correlation matrix is used to study the strength of a relationship between two variables.
- It not only shows the direction of the relationship, but also shows how strong the relationship is.
- The correlation formula can be represented as:

$$\mathrm{COR}(X,\ Y) = \frac{\mathrm{COV}(X,\ Y)}{\sqrt{\mathrm{VAR}(X)\mathrm{VAR}(Y)}}$$

What is Covariance Matrix?

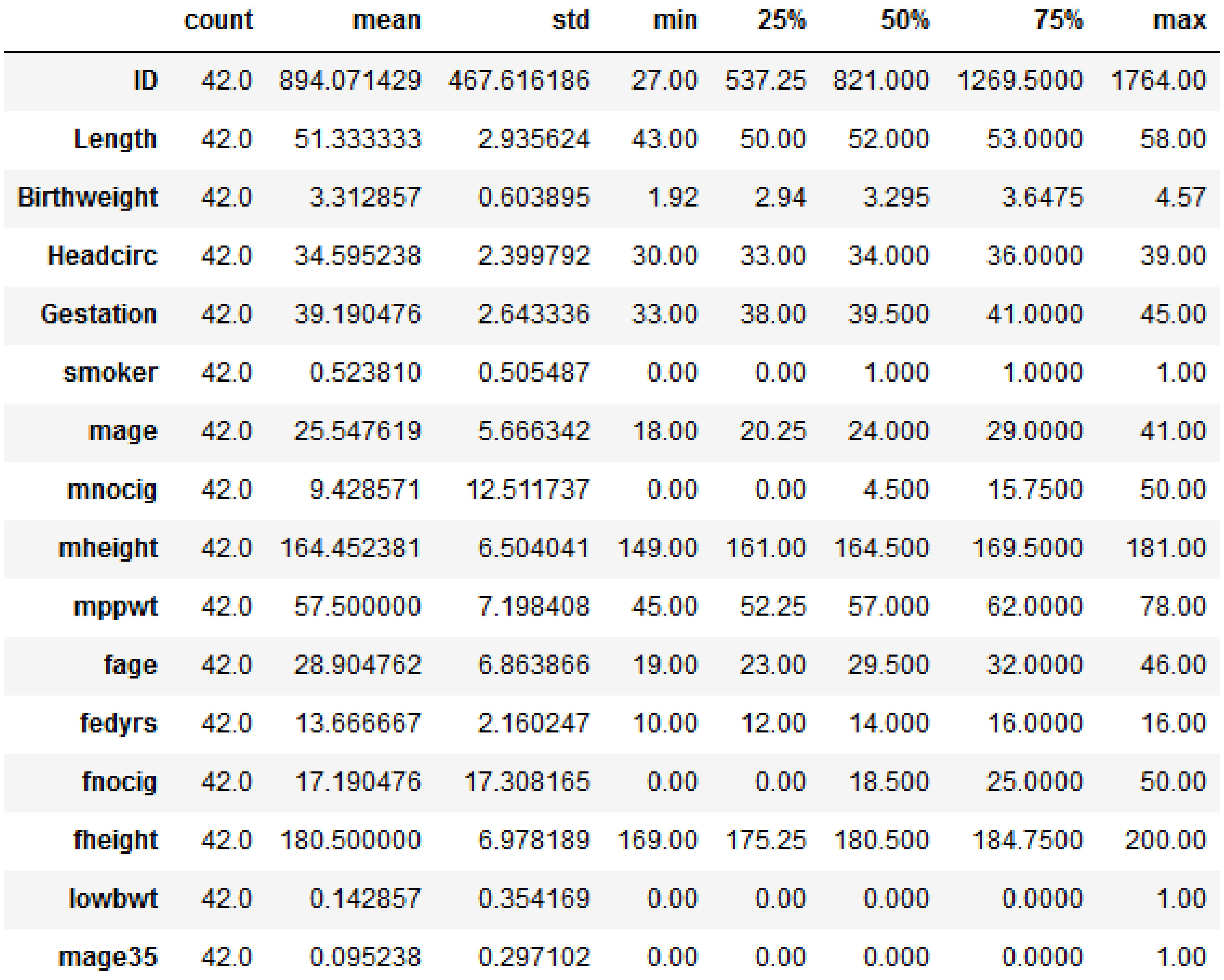



0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
Y-Value
0.2
0.4
0.6
0.8
1.0
X-Value

"Complete Lab 2"

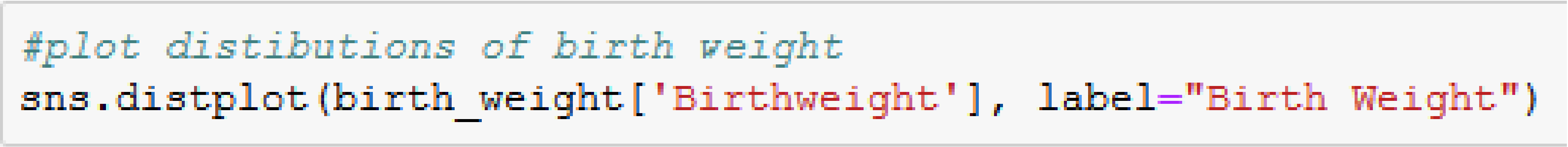
"Complete Case Study"

Descriptive Statistics

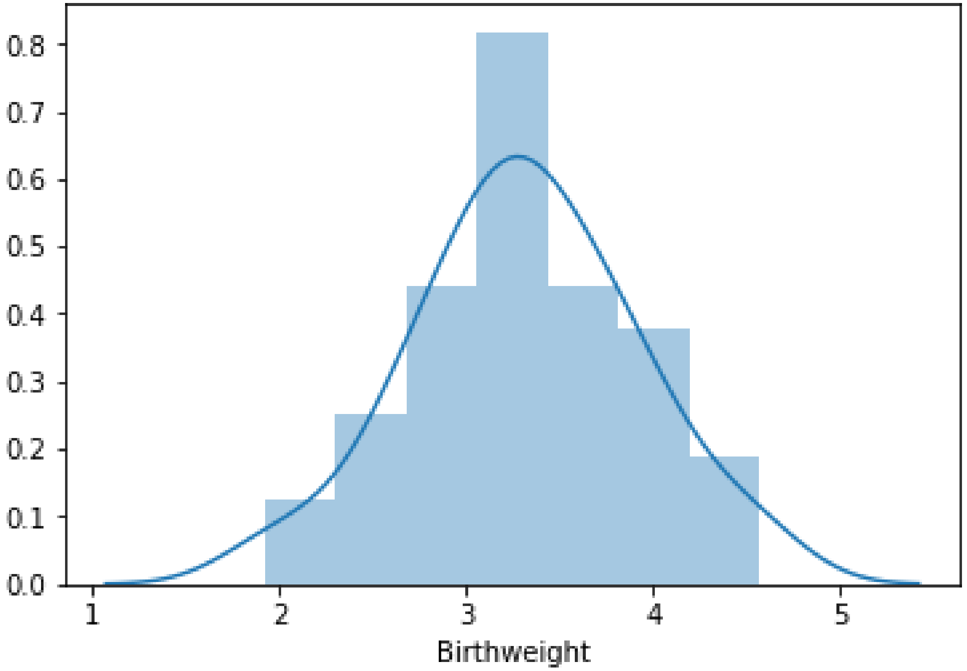


	count	mean	std	min	25%	50%	75%	max
ID	42.0	894.071429	467.616186	27.00	537.25	821.000	1269.5000	1764.00
Length	42.0	51.333333	2.935624	43.00	50.00	52.000	53.0000	58.00
Birthweight	42.0	3.312857	0.603895	1.92	2.94	3.295	3.6475	4.57
Headcirc	42.0	34.595238	2.399792	30.00	33.00	34.000	36.0000	39.00
Gestation	42.0	39.190476	2.643336	33.00	38.00	39.500	41.0000	45.00
smoker	42.0	0.523810	0.505487	0.00	0.00	1.000	1.0000	1.00
mage	42.0	25.547619	5.666342	18.00	20.25	24.000	29.0000	41.00
mnocig	42.0	9.428571	12.511737	0.00	0.00	4.500	15.7500	50.00
mheight	42.0	164.452381	6.504041	149.00	161.00	164.500	169.5000	181.00
mppwt	42.0	57.500000	7.198408	45.00	52.25	57.000	62.0000	78.00
fage	42.0	28.904762	6.863866	19.00	23.00	29.500	32.0000	46.00
fedyrs	42.0	13.666667	2.160247	10.00	12.00	14.000	16.0000	16.00
fnocig	42.0	17.190476	17.308165	0.00	0.00	18.500	25.0000	50.00
fheight	42.0	180.500000	6.978189	169.00	175.25	180.500	184.7500	200.00
lowbwt	42.0	0.142857	0.354169	0.00	0.00	0.000	0.0000	1.00
mage35	42.0	0.095238	0.297102	0.00	0.00	0.000	0.0000	1.00

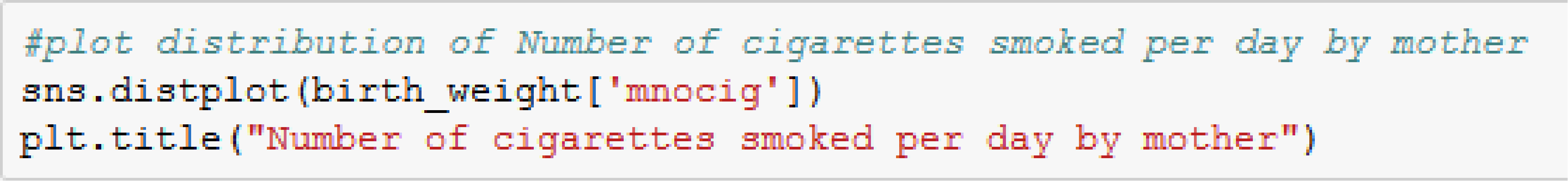
- We can analyze the distribution of the birth weight variable.



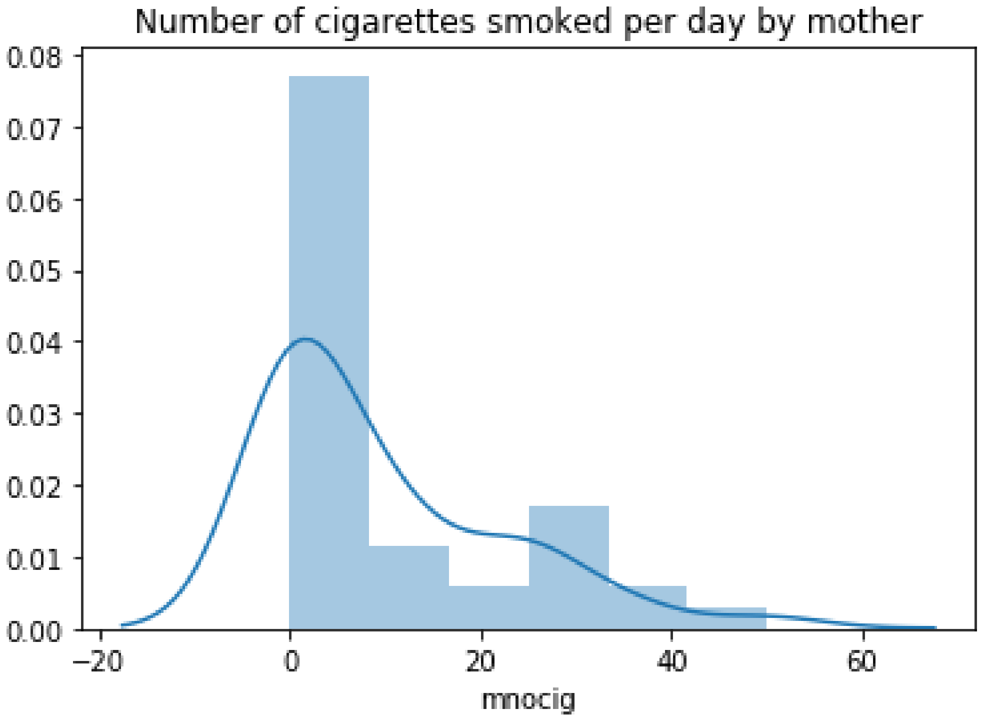
```
#plot distibutions of birth weight
sns.distplot(birth_weight['Birthweight'], label="Birth Weight")
```



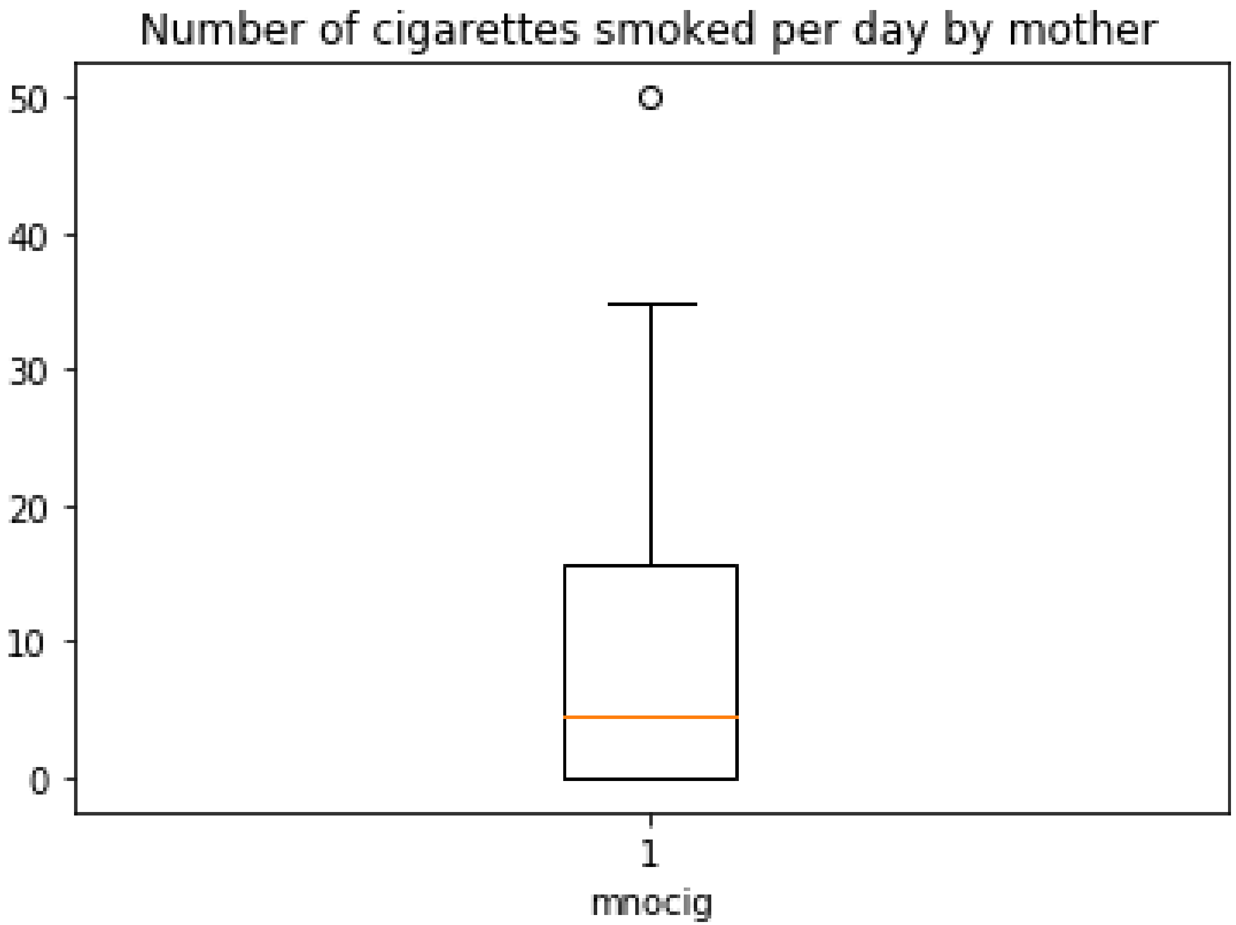
- Let’s analyze the distribution of “mnocig” (Number of cigarettes smoked per day by mother) variable



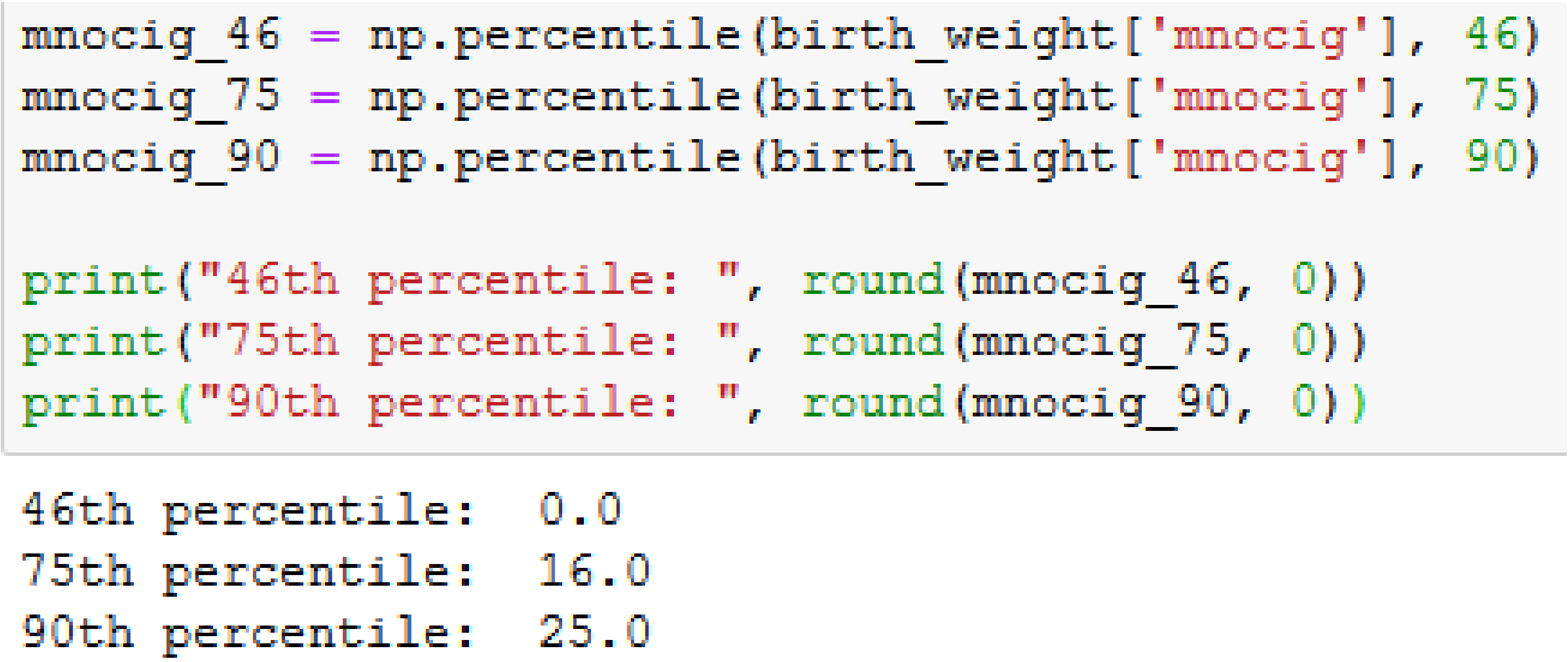
```
#plot distribution of Number of cigarettes smoked per day by mother
sns.distplot(birth_weight['mnocig'])
plt.title("Number of cigarettes smoked per day by mother")
```



Descriptive Statistics

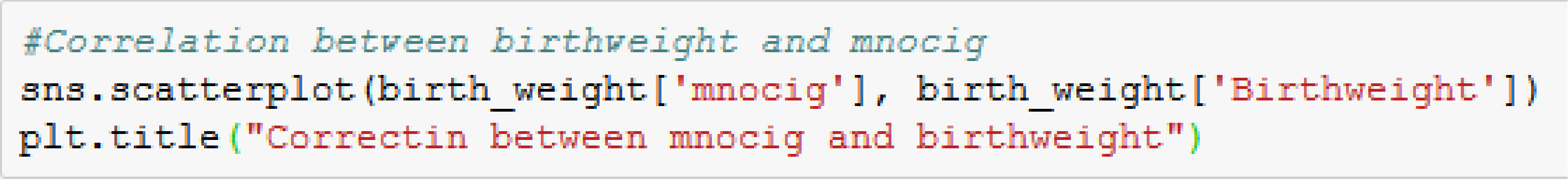


Descriptive Statistics

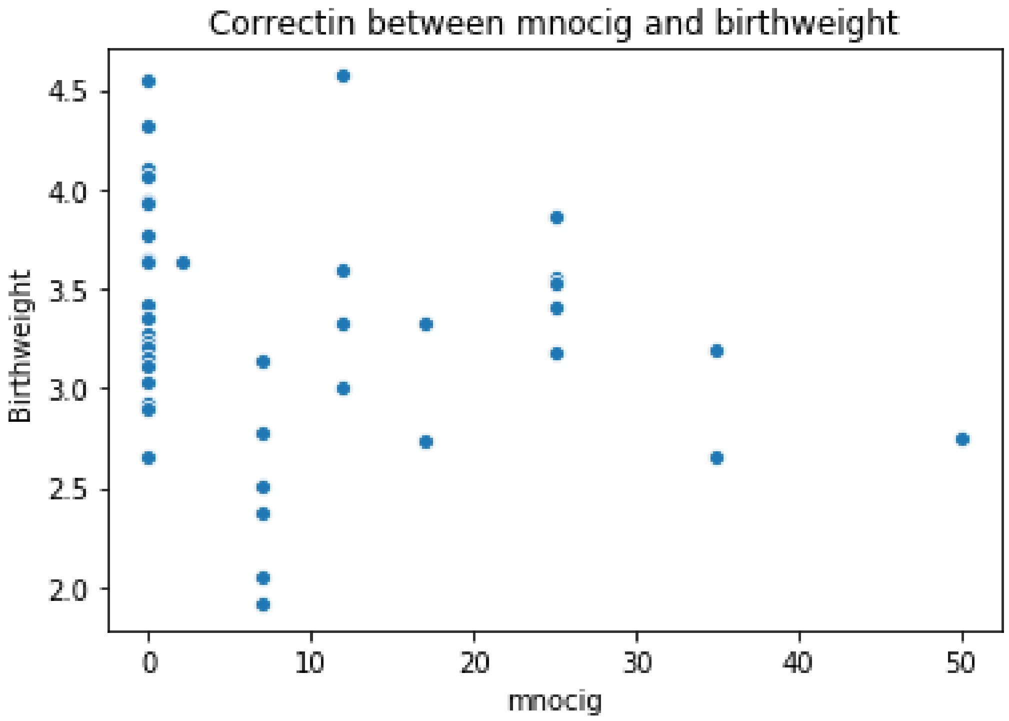


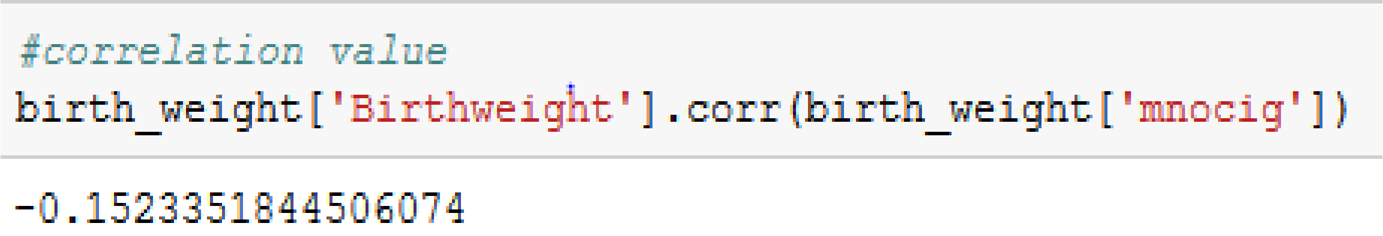
```python
mnocig_46 = np.percentile(birth_weight['mnocig'], 46)
mnocig_75 = np.percentile(birth_weight['mnocig'], 75)
mnocig_90 = np.percentile(birth_weight['mnocig'], 90)

print("46th percentile: ", round(mnocig_46, 0))
print("75th percentile: ", round(mnocig_75, 0))
print("90th percentile: ", round(mnocig_90, 0))
```

```
46th percentile:  0.0
75th percentile:  16.0
90th percentile:  25.0
```



```
#Correlation between birthweight and mnocig
sns.scatterplot(birth_weight['mnocig'], birth_weight['Birthweight'])
plt.title("Correctin between mnocig and birthweight")
```





```
#correlation value
birth_weight['Birthweight'].corr(birth_weight['mnocig'])
```

```
-0.1523351844506074
```

SUMMARY

- Statistics deals with collecting, interpreting, and drawing a conclusion from the data.
- Data is measured on different scales like nominal, ordinal, interval and ratio.
- Descriptive statistics aims to summarize a sample data with a single value with the help of mean, median and mode.

"Complete Assessment"