



# Introduction To Statistics

Statistics is a field of maths that deals with the collection, analysis, explanation and presentation of data. Statistical methods can be used to find solutions to problems which contain numerical data. These mathematical formulas or Statistical methods used to find solutions are known as quantitative models. Thus, we can define Statistics as a branch of mathematics which uses different quantitative models to produce a solution for a real-world problem containing numerical data. Statistics can be used to solve real-world complex problems using data. Knowledge of Statistics can help a person to draw some very interesting inferences from data.

**Real world examples of application of Statistics:**

- Collection of data in data-surveys
- Filling of missing values in data (Data Cleaning)
- Analysis of Data
- Finding useful inferences from data
- Presentation of data using charts and tables ( Boxplot, Histogram etc)

**Role of Statistics in Data Science**

*Josh Wills, a former head of data engineering at Slack, said “A data scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”*

Data is a term that we usually hear these days, in today’s modern world data is among one of the most common terms that is being used by people in their day to day life. Enormous amount of data is being generated by people all around the world on a daily basis. Now for some people this data is useless but for some it is as valuable as gold. Data Scientists use this unstructured data to find some really interesting inferences which can help their companies in many ways. Example using data to specifically target a particular audience, using data for early detection of diseases, using data for political campaigns etc. These are some of the real-world scenarios in which data is being used to enhance the outcomes.

But finding these inferences isn’t an easy job, the data that we gather is unstructured and requires a lot of pre-processing. Only after preprocessing the data can this data be used for visualizations and training models. Here statistics comes into play, having a good knowledge of statistics can help a programmer in various steps of Data Analysis.

It can be used in data pre-processing, for example seeing the unstructured data an analyst can determine whether to drop the rows which don't contain any value (NaN rows) or to fill them with mean, mode etc (Statistical methods).

After data pre-processing, an analyst can find various inferences from data even without visualising the data example using .describe() function is python:

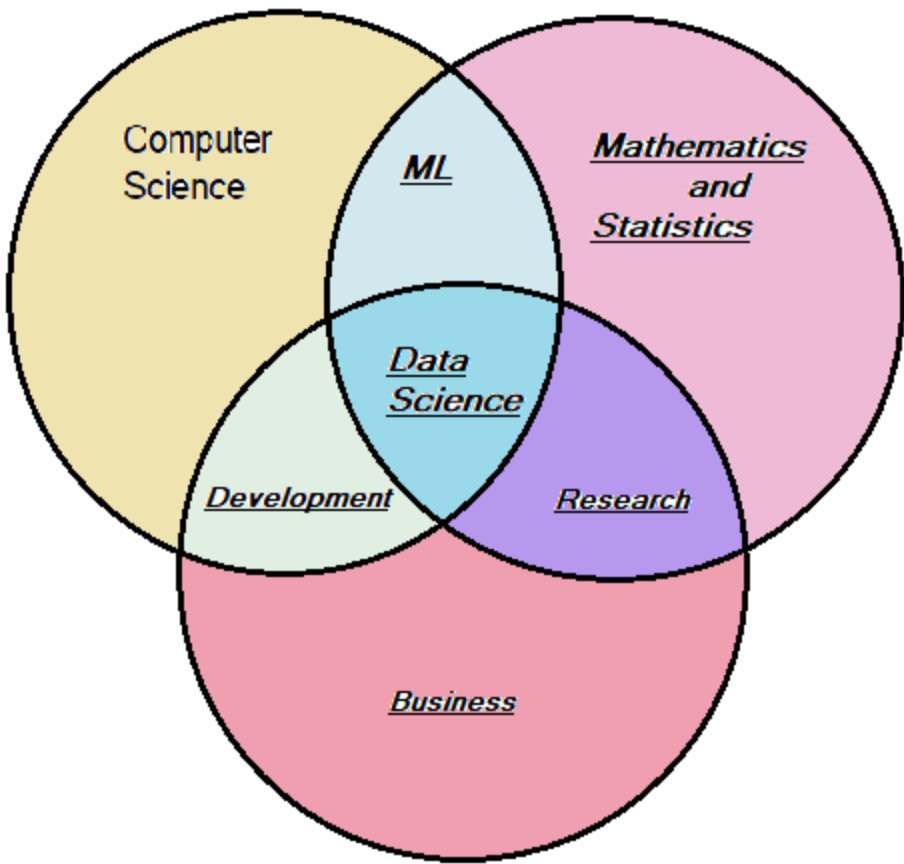
```
1 df.describe()
```

	price	year	mileage
count	2499.000	2499.000	2499.000
mean	18767.671	2016.714	52298.685
std	12116.095	3.443	59705.516
min	0.000	1973.000	0.000
25%	10200.000	2016.000	21466.500
50%	16900.000	2018.000	35365.000
75%	25555.500	2019.000	63472.500
max	84900.000	2020.000	1017936.000

Now for a normal programmer these values won’t make any sense but for a person with knowledge of stats, a lot of inferences can be drawn from this simple function.

Statistical visualizations like boxplot, histogram, line, etc can be very important to convey these inferences to the stakeholders.





And finally, when we train a ML model, Statistics can be of great help in evaluating our model, pre-processing the data before fitting in the model etc.



Overall, statistics is essential in data science because it provides a framework for understanding and analyzing data. Without statistics, data scientists would not be able to make sense of the vast amounts of data that is generated in today's world, nor would they be able to use that data to make informed decisions.

Some of the key applications of statistics in data science include:

1. Descriptive statistics: It is a set of statistical techniques that are used to summarize and describe the characteristics of a dataset. This includes measures of central tendency (such as the mean, median, and mode), measures of variability (such as the range and standard deviation), and measures of distribution (such as histograms and frequency tables).

-  Dash
-  Articles
-  Videos
-  Quiz

Confidence intervals are another inferential statistical technique used to estimate population parameters. A confidence interval is a range of values that is likely to contain the true population parameter with a certain level of confidence. For example, a 95% confidence interval for the population mean would be a range of values that is expected to contain the true population mean with 95% confidence.



If you are facing any issue on this page. Please let us know.