Basic data exploration

02 July 2024 10:37

Using the .describe() function in pandas will give you 8 basic stats pertaining to the data. The

The results show 8 numbers for each column in your original dataset. The first number, the **coun** non-missing values.

Missing values arise for many reasons. For example, the size of the 2nd bedroom wouldn't be co bedroom house. We'll come back to the topic of missing data.

The second value is the **mean**, which is the average. Under that, **std** is the standard deviation, wh numerically spread out the values are.

To interpret the **min**, **25%**, **50%**, **75%** and **max** values, imagine sorting each column from lowest t (smallest) value is the min. If you go a quarter way through the list, you'll find a number that is b and smaller than 75% of the values. That is the **25%** value (pronounced "25th percentile"). The 5 defined analogously, and the **max** is the largest number.

ey are as followed:

t, shows how many rows have

llected when surveying a 1

hich measures how

to highest value. The first
igger than 25% of the values
0th and 75th percentiles are