Exploratory Data Analysis

1. Exploratory Data Analysis

00:00 - 00:07

Excellent job on data preparation! Let's keep going with Exploratory Data Analysis.

2. What is EDA?

00:07 - 00:26

Exploratory Data Analysis, or EDA for short, is a process that was promoted by John Tukey, a respected statistician. It consists in exploring the data and formulating hypotheses about it, and assessing its main characteristics, with a strong emphasis on visualization.

3. Data workflow

00:26 - 00:33

EDA happens after data preparation, but they can get mixed. EDA can reveal new things that need cleaning.

4. Let's dive right in

00:33 - 00:44

Let's dive right in. What can you say about these four different datasets? Well, from these lines of data, probably very little.

5. Surprise!

00:44 - 01:09

All four datasets have identical mean and variance both for the x and y features. They also have an identical correlation coefficient, and the same linear regression equation (the straight line that tries to go through all points). If you're not sure about some of the metrics listed, that's OK. The point is, they seem awfully similar! But is that the case?

6. Anscombe's quartet

01:09 - 01:18

No! Here are the four graphs. They all tell a different story, that pure metrics can't fully convey.

7. Anscombe's quartet

01:18 - 01:21

The first graph displays a linear relationship,

8. Anscombe's quartet

01:21 - 01:25

while the second one has a non-linear relationship.

9. Anscombe's quartet

01:25 - 01:31

In the third graph, we see the linear line is thrown off by one point that has an extreme y value.

10. Anscombe's quartet

01:31 - 02:02

A similar thing happens with the fourth dataset. We should have no correlation, but one extreme point is enough to display a strong one. In short, streaming through the data gives you little information. Descriptive statistics do better, but can be misleading; visualization teaches us the most. That's why EDA relies heavily on this last technique. This was an extreme example, to make a point. Let's now look at another dataset.

11. SpaceX launches

02:02 - 02:08

Let's look at SpaceX launches!

12. Knowing your data

02:08 - 02:19

I mean, let's look at the data behind SpaceX launches. The first thing to do is to know what features we're looking at. We have different information, such as the flight number or what the rocket transported. All have the correct data type.

13. Previewing your data

02:19 - 02:27

Looking at your tables helps make sense of your observations. Can you notice the missing payload mass for the first two rows?

14. Descriptive statistics

02:27 - 03:00

It's always a good idea to calculate descriptive statistics. The SpaceX dataset is mainly qualitative, but we still get a lot of information. We have a count of 55 pretty much everywhere, because we have 55 launches. The Payload Mass column shows 53 because of the two missing values we saw before. Only 1 mission failed. Most of the time, there is no attempt at landing. You could also calculate the average payload mass, or the count of launches per year. But do you know what would be best for this last option?

15. Visualize!

03:00 - 03:21

Visualization! In a glance we can see that there were no launches in 2011. The count of launches then gradually increased before doubling in 2017. 2018 is lower, but remember we only have 3 months of data for this year, so it actually looks like it's going to double again.

16. Ask more questions!

03:21 - 03:40

Now this launch count is informative, but you probably have a couple more. How about count by launch site? Rockets originally launched from Cape Canaveral Air Force Station, but in 2017 most rockets launched from Kennedy Space Center Launch Complex 39.

17. Ask more questions!

03:40 - 03:44

How about mission outcome? Just one failure in 2015!

18. Outliers

03:44 - 04:08

Another thing you do during EDA is look for outliers, that is, unusual values. Whether they are errors or valid, it's nice to know about them, as they can throw your results off. Here, we can see we have only 5 launches with a weight greater than 7,000 kg, when the average mass is closer 3800 kg.

19. Let's practice!

04:08 - 04:13

Let's check your understanding of the EDA process!