IBM Course Seven

Week Three

Objectives

01

Apply Python exploratory data analysis techniques

02

Describe why and how to apply the chi-Squared Test

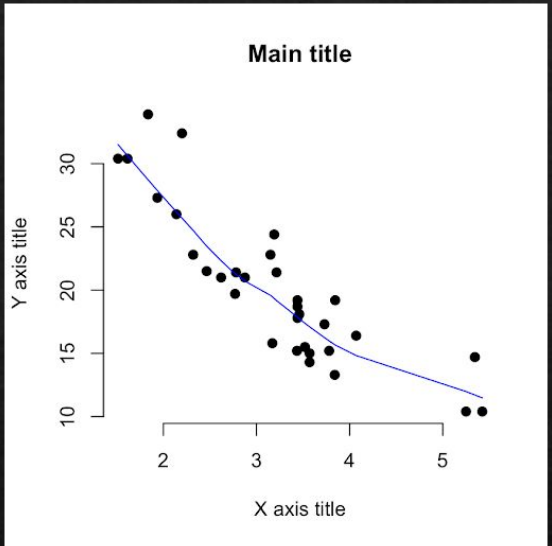
03

Implement descriptive statistics

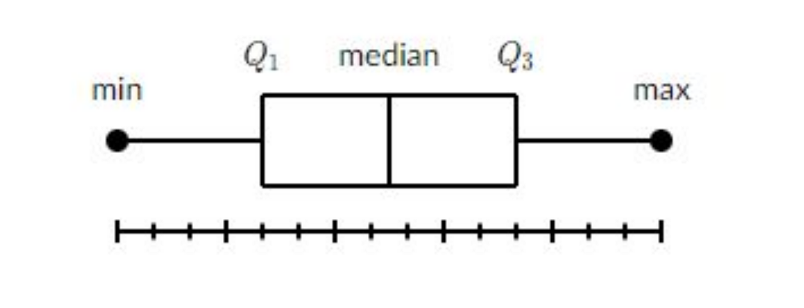
Exploratory Data Analysis

- EDA is an approach to analyzing data sets. The goal is to summarize their main characteristics with visual methods. It is used to discover patterns and anomalies, to test hypotheses, and to check assumptions.

Descriptive Statistics



- Descriptive stats describe the basic features of data by giving short summarize about the sample and measure of the data.
- You can summarize statistics using pandas describe() method.
- You can summarize the categorical data by using the value_counts() method.
- Box plot □method for graphically depicting groups of numerical data through their quartiles: minimum, first quartile, median, third quartile, and maximum.
- Scatter plot □ uses dots to represent values for two different numeric variables. They are used to observe relationships between variables. There are 3 possible correlations: positive, negative, and no correlation. n

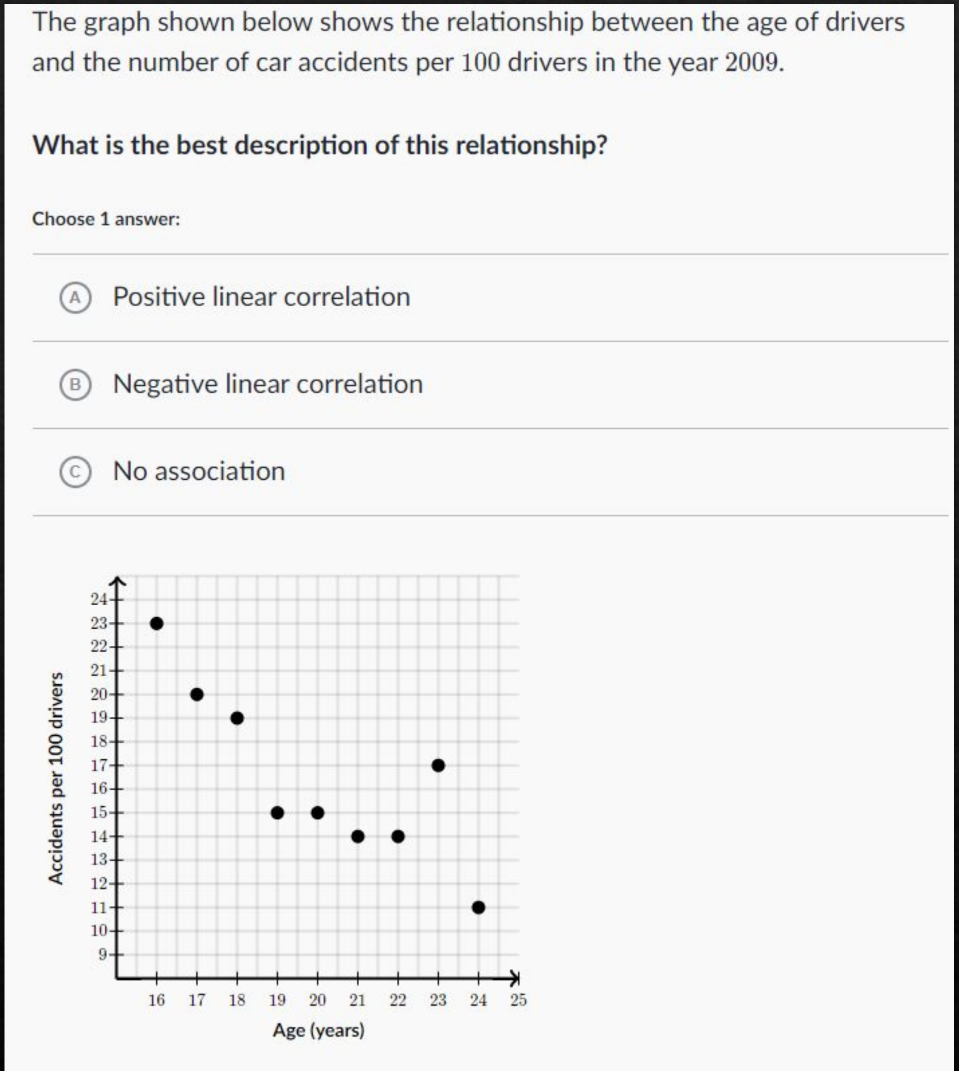


35, 29, 34, 25, 29, 28, 38, 37, 35, 30

What are the steps you would take prepare the data to be visualized in a box plot?

Practice!

Here’s a video if you get stuck.



The graph shown below shows the relationship between the age of drivers and the number of car accidents per 100 drivers in the year 2009.

What is the best description of this relationship?

Choose 1 answer:

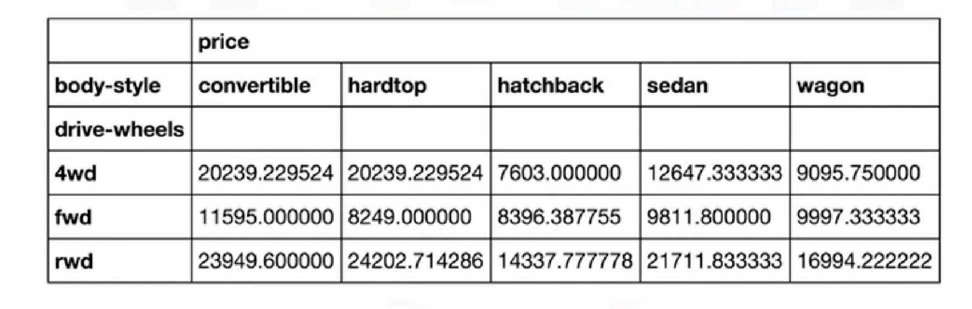
Ⓐ Positive linear correlation

Ⓑ Negative linear correlation

Ⓒ No association

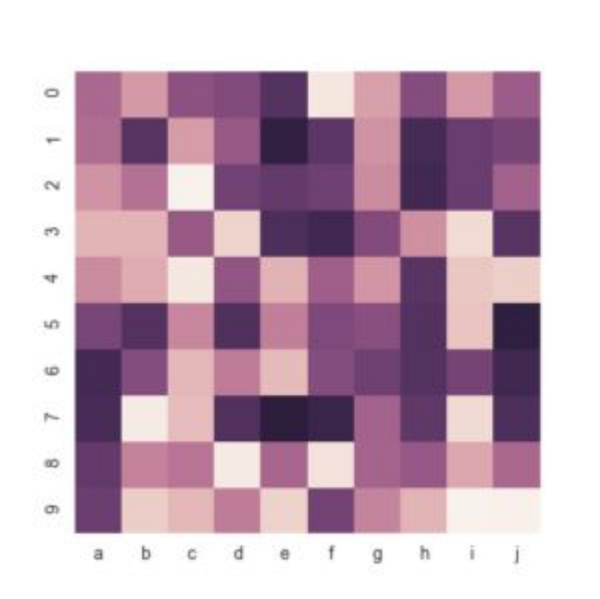
GroupBy is a function used to split data into groups based on specific criteria.

Pivot Tables



	price				
body-style	**convertible**	**hardtop**	**hatchback**	**sedan**	**wagon**
drive-wheels					
4wd	20239.229524	20239.229524	7603.000000	12647.333333	9095.750000
fwd	11595.000000	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.600000	24202.714286	14337.777778	21711.833333	16994.222222

Heatmap Plot



Correlation

- Correlation is a statistical metric for measuring to what extend variables are interdependent.
- Correlation DOES NOT IMPLY causation.
- What is the difference between correlation and causation?
- Positive linear relationship
- Negative linear relationship
- No relationship

Correlation – Statistics

- Pearson Correlation □ gives you two values: the correlation coefficient and the P-value.
- Correlation coefficient □ -1 – 1
- P-value □ 0.0001 – 0.1

<u>Chi-Square</u>

- Shows a relationship between two <u>categorical</u> variables
- Does NOT tell you what kind of relationship exists between both variables; it only indicates whether there is a relationship.
- *Watch the video linked above in your Scrum groups. Take notes while you watch. Use those notes to fill in gaps in your Coursera notes. Then, write out a 5-10 sentence summary of what a Chi-test is and how and why it is used. If you can explain it to someone else, you understand it!*

Labs and Assessments

- Take the next 2 hours to work through the lab and the quiz for Course Six Week 3. Don't forget to use the lab to test different commands and practice all concepts touched upon so far!