|  |
| --- |
| Course: Data Science [Z15811] Class: Data Exploration [03]  Subject: Discussion topics |

1. When visualizing the distribution of a single variable, a histogram can be used. How is it calculated?

Verdeel de variabele warden in gelijk verdeelde intervals genaamd ‘bins’

1. How can you detect that a continuous variable contains unusual values (so-called outliers)?

* Sorteren van de dataset
* Grafieken maken (boxplots, histograms, scatterplots voor outliers te vinden)
* Z-scores gebruiken

Afbeelding met tekst

Automatisch gegenereerde beschrijving

1. Which visualization can you use to explore the distribution of a continuous variable, broken down by a categorical variable? How can you make the comparison the easiest?

Histogram en density lines, dit maakt het makkelijk voor een boxplot te gebruiken.

1. Another alternative to display the distribution of a continuous variable broken down by a categorical variable is the boxplot. Which summaries does a boxplot contain?

Min

Max

Q1 (25ste percentiel)

Q3 (75ste percentiel)

Median

1. How is the 75th percentile of a variable in a dataset calculated?

Maal 0.75, afronden naar beneden

1. Does a boxplot contain the mean value of the variable you’re displaying?

Nee, het gebruikt percentielen.

1. What is the difference between a mean and a median?

Mean = gemiddelde  
Median = mediaan (midden)

Sketch a histogram for a continuous variable whose:

* 1. mean and the median are roughly the same

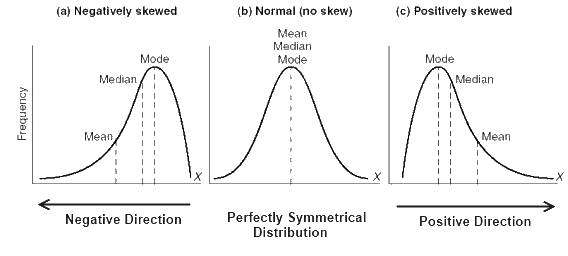
Not skewed

* 1. mean is larger than its median

Skewed to the left

* 1. mean is smaller than its median

Skewed to the right



1. Which visualization can you use to display the covariation between categorical variables?

Geom\_count()

1. What’s the meaning of a negative covariance/correlation? What is the meaning of positive covariance/correlation?

Negative means that if a variable changes with +1 StandardDeviation the other changes + r\*StandardDeviation