DATA SCIENCE & BUSINESS ANALYTICS

FINANCE AND ECONOMICS IN SPORT

FALL 2022

WHAT IS DATA SCIENCE?

Study of Data

- Developing methods of collecting data
- Analyzing data
- Interpreting the analysis results

To extract useful information

WHAT IS ANALYTICS?

"The extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to derive decisions and actions."

(Davenport & Harris, 2007)

BUSINESS ANALYTICS

The use of statistics and math to extract meaningful insights from data to make better organizational decisions

- Interpretation of historical data to identify trends and patterns
- Forecasting future outcomes
- Identifying which outcome will yield the best results in a given scenario
- Budgeting / forecasting / product development

SPORT ANALYTICS

USE OF ANALYTICS IN THE SPORT INDUSTRY

- Segment existing fans and estimate their value
- Predict consumer demand
- Examine effectiveness of marketing activities
- Assess value of athletes to their brand
- Analyze athletes' performance
- Predict and prevent player injuries

WHAT ARE DATA?

- Information in a variety of forms
  - Numbers, words, pictures, video, measurements and observations

- Important questions to ask prior to data collection:
  - What is relevant?
  - What are the sources of data?
  - How much data are needed?
  - How to ensure quality?

TYPES OF DATA ANALYSIS

Quantitative Methods

- Testing theories using numbers
- Economic indicators
- Finance performance
- Stock price

Qualitative Methods

- Testing theories using language
- Interviews
- Conversations
- Newspapers
- Media broadcasts

GENERATING AND TESTING THEORIES

- **Theory**
  - A hypothesized general principle or set of principles that explain known findings about a topic and from which new hypotheses can be generated.
- **Hypothesis**
  - A prediction from a theory
- **Falsification**
  - The act of disproving a theory or hypothesis.

COLLECTING DATA TO TEST THEORIES

- **Dependent Variable**
  - The proposed effect
  - An outcome variable
  - Measured not manipulated (in experiments)
- **Independent Variable**
  - The proposed cause
  - A predictor variable
  - A manipulated variable (in experiments)

LEVELS OF MEASUREMENT

Categorical Variable	Entities are divided into distinct categories.	Example
Binary	There are only two categories.	Win or Loss Alive or Dead
Nominal	There are more than two categories.	North Carolina Virginia New York
Ordinal	The same as a nominal variable but the categories have a logical order.	Fist place Second place Third place

LEVELS OF MEASUREMENT

Continuous Variable	Entities get a distinct score.	Example
Interval	Equal intervals on the variable represent equal differences in the property being measured.	The difference between 6 and 8 is equivalent to the difference between 13 and 15.
Ratio	The same as an interval variable, but the ratios of scores on the scale must also make sense.	A score of 10 on a GDP scale means that the country is, in reality, twice as rich as a country scoring 5.

TYPES OF STATISTICAL ANALYSIS

- **Descriptive statistics**

  Summarize and describe data via frequencies, central tendency, measures of dispersion and distribution characteristics.

  (e.g., Stock price of Nike, Total revenue of Division I schools)

- **Inferential statistics**

  Rely on a sample to make inferences about a population via inductive reasoning.

  - Sampling method is extremely important.
  - Relationship between dependent variable and independent variable
  - Representativeness of data
  - Generalizability of data

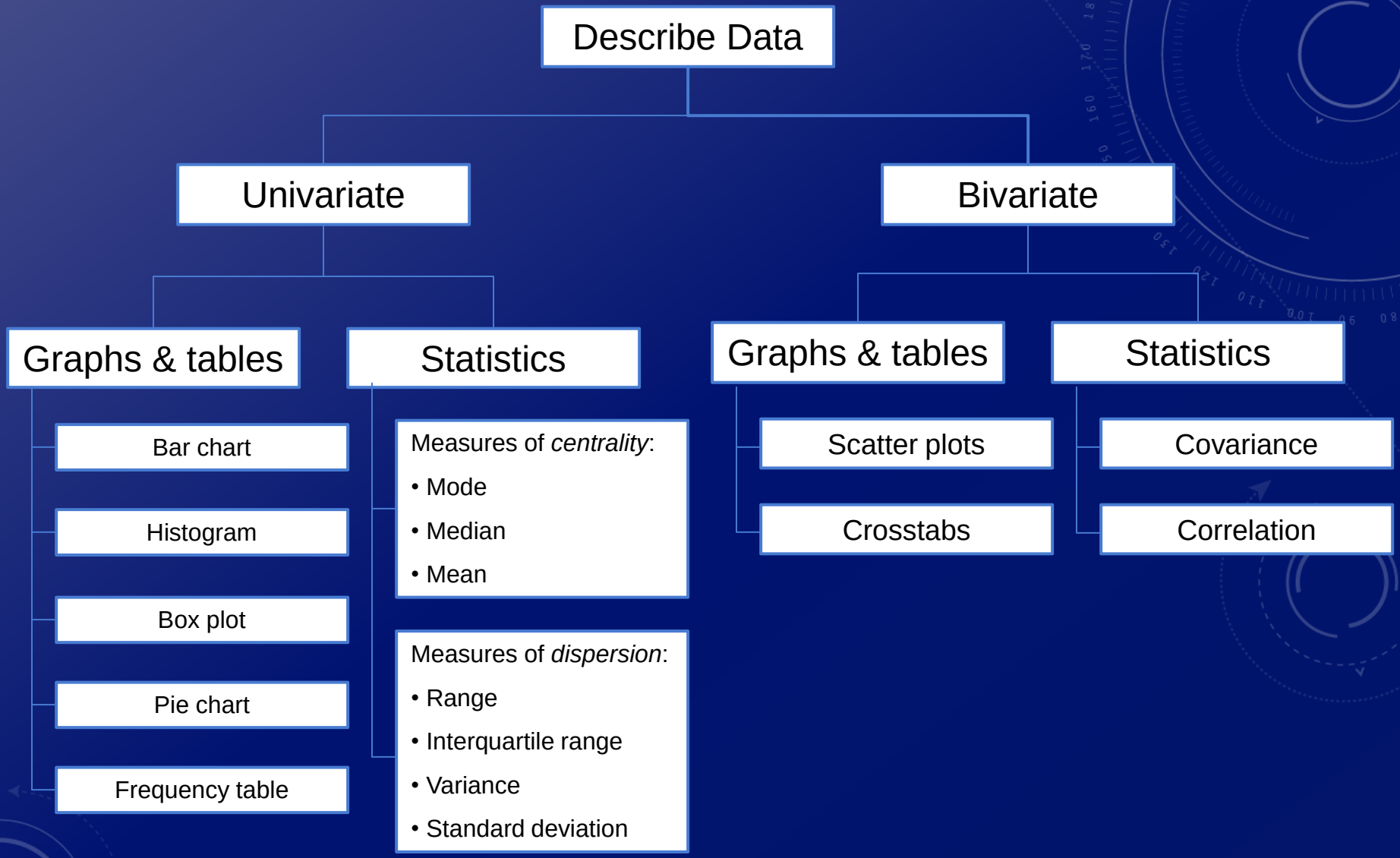
DESCRIPTIVE STATISTICS

WORK FLOW OF DATA MANAGEMENT

Enter Data	• Enter data by hand in *Jamovi* or *Excel*: Risk of making errors • Get data from external sources (*U.S. Census Bureau / Bureau of Economic Analysis*) • Checks of consistency are very important.
Clean Data	• Check data for suspicious response patterns • Check data for entry errors • Check data for outliers • Check for missing data

Describe Data	• Univariate analysis
	  • Mean / Median /Mode
	  • Box plot
	• Bivariate analysis
	  • Correlations
	  • Scatter plot

DESCRIBE DATA



UNIVARIATE & BIVARIATE STATISTICS

Univariate

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median

Range

Interquartile range

Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation

$$s = \sqrt{s^2}$$

Bivariate

Covariance

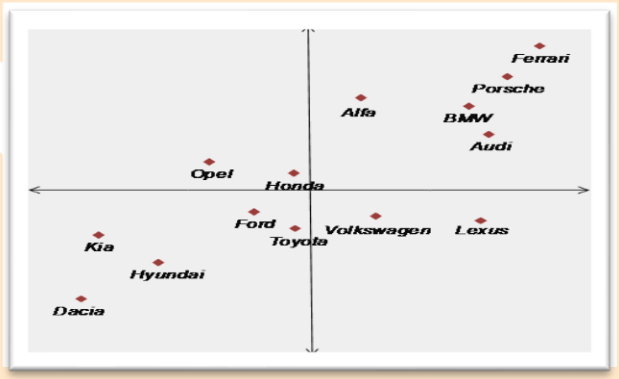
$$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlation (Pearson)

$$r = \frac{Cov(x, y)}{s_x \times s_y}$$

Scatter plots

Crosstabs (Contingency tables, Pivot tables)



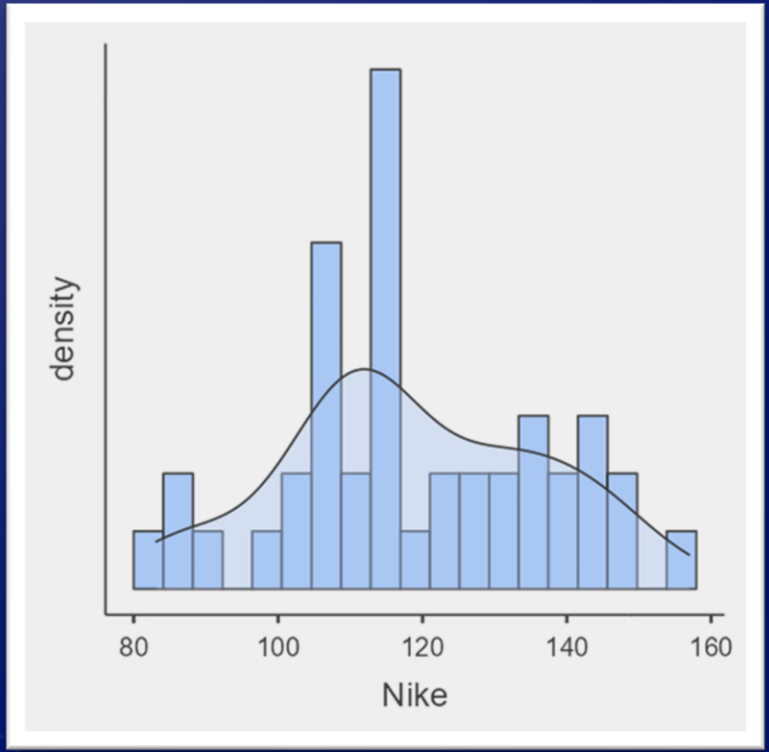
UNIVARIATE STATISTICS

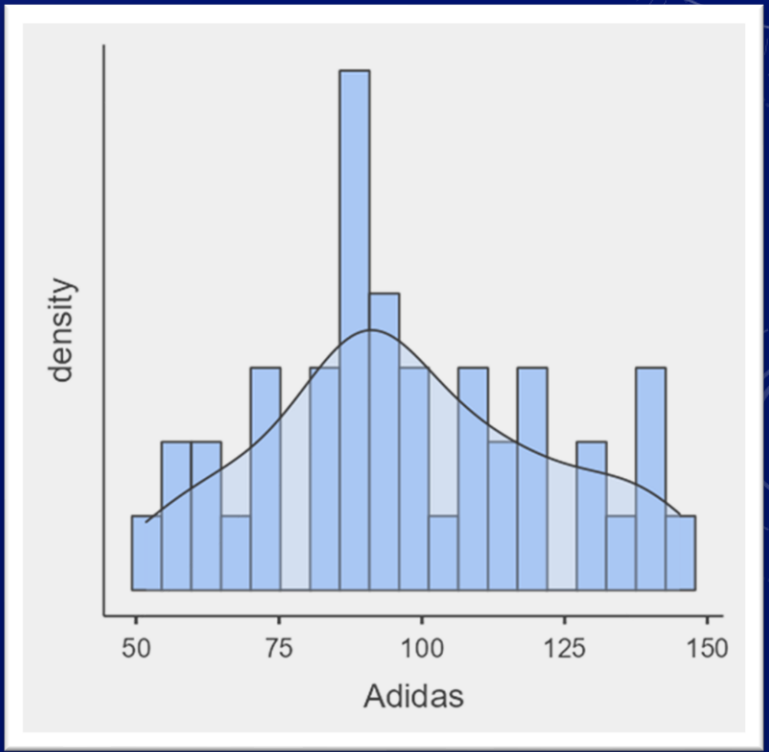
- In *Jamovi,* Analyses ➡ Exploration ➡ Descriptives
- Central Tendency
  - mean / median / mode
- Dispersion
  - standard deviation / minimum / maximum
- Distribution
  - skewness / kurtosis

UNIVARIATE STATISTICS

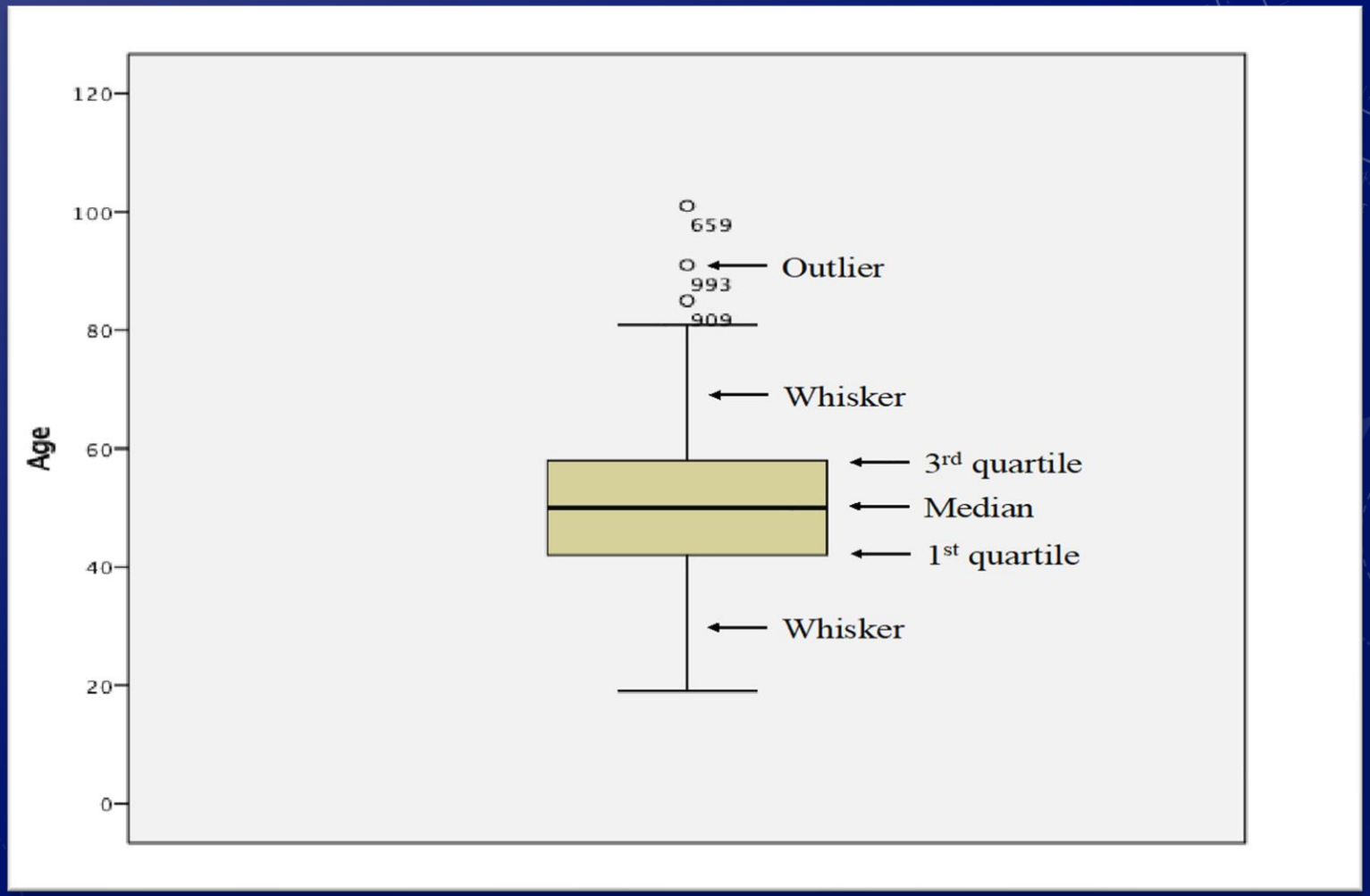
	Nike	UA	Adidas	DKS
N	42	42	42	42
Missing	0	0	0	0
Mean	119	12.3	97.5	100
Median	115	10.2	93.5	105
Mode	143	6.65 [a]	51.8 [a]	73.8 [a]
Standard deviation	18.0	4.31	24.9	12.1
Minimum	83.1	6.65	51.8	73.8
Maximum	157	19.9	145	117
Skewness	0.0615	0.394	0.176	-0.822
Std. error skewness	0.365	0.365	0.365	0.365
Kurtosis	-0.545	-1.48	-0.649	-0.569
Std. error kurtosis	0.717	0.717	0.717	0.717
Shapiro-Wilk W	0.974	0.871	0.971	0.886
Shapiro-Wilk p	0.450	< .001	0.365	< .001

UNIVARIATE GRAPHS: HISTOGRAM

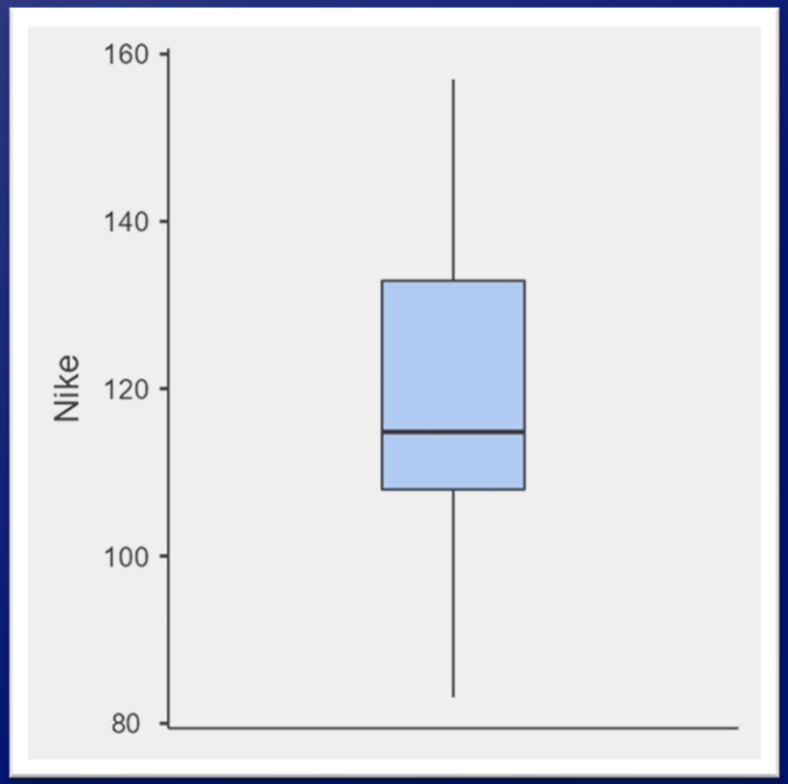


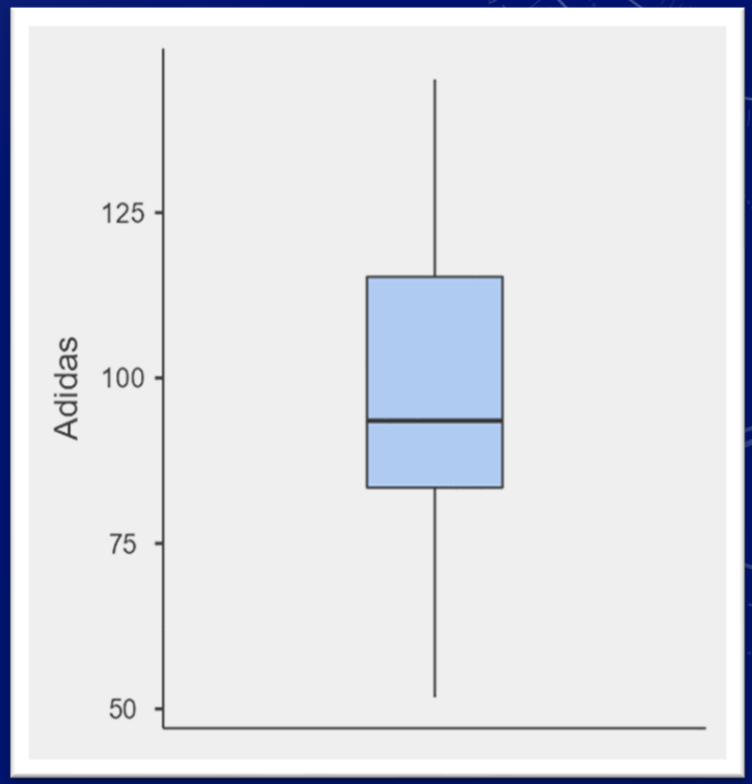


UNIVARIATE GRAPHS: BOX PLOT



UNIVARIATE GRAPHS: BOX PLOT





BIVARIATE STATISTICS: CORRELATION

- As a rule of thumb (Cohen 1988), an absolute correlation …
  - **Below 0.30: weak relationship,**
  - **Between 0.30 and 0.49: moderate relationship**
  - **Above 0.49: strong relationship**

- Spearman's correlation coefficient is used for calculating correlations between two variables that are both **interval or ratio-scaled**:
- Spearman's correlation coefficient and Kendall's tau when at least one variable for determining the correlation is **measured on an ordinal scale**.
- Contingency coefficient, Cramer's V, and Phi for variables measured on a **nominal scale**. These statistical measures are used with crosstabs.

BIVARIATE STATISTICS: CORRELATION

- In *Jamovi,* Analyses → Regression → Correlation Matrix
- Correlation Coefficients
  - Pearson
  - Spearman
  - Kendall's tau-b

BIVARIATE STATISTICS: CORRELATION

Correlation Matrix

		Nike	Adidas	UA	DKS
Nike	Pearson's r	—			
	p-value	—			
	Spearman's rho	—			
	p-value	—			
Adidas	Pearson's r	0.962			
	p-value	< .001	—		
	Spearman's rho	0.950	—		
	p-value	< .001	—		
UA	Pearson's r	0.951	0.945	—	
	p-value	< .001	< .001	—	
	Spearman's rho	0.972	0.978	—	
	p-value	< .001	< .001	—	
DKS	Pearson's r	0.275	0.173	0.371	—
	p-value	0.039	0.137	0.008	—
	Spearman's rho	0.267	0.139	0.211	—
	p-value	0.044	0.190	0.090	—

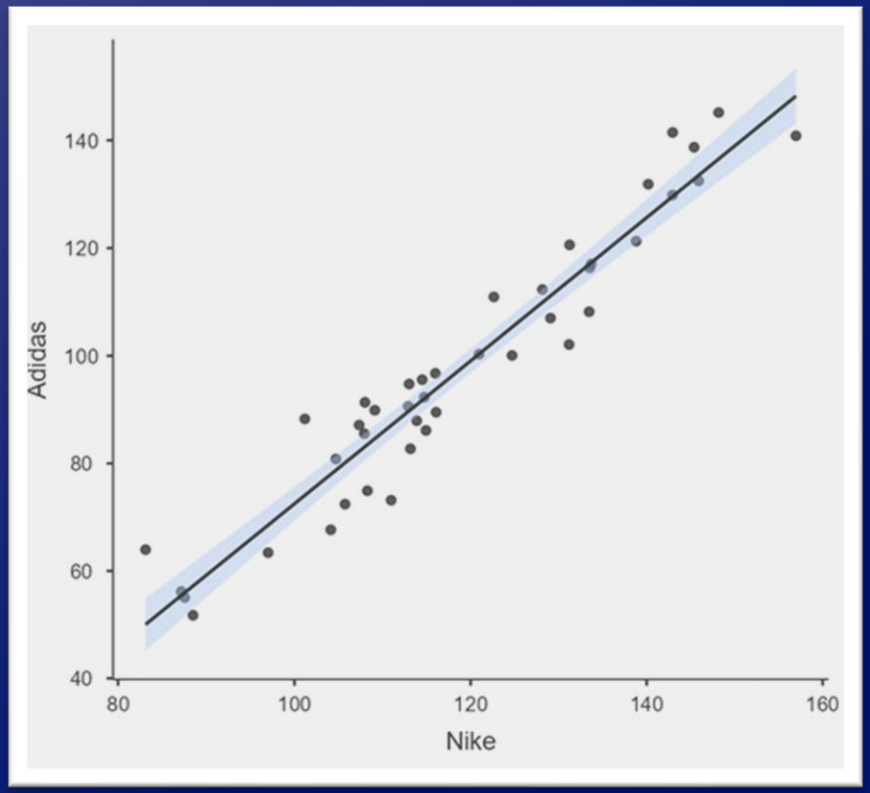
Note. H_a is positive correlation

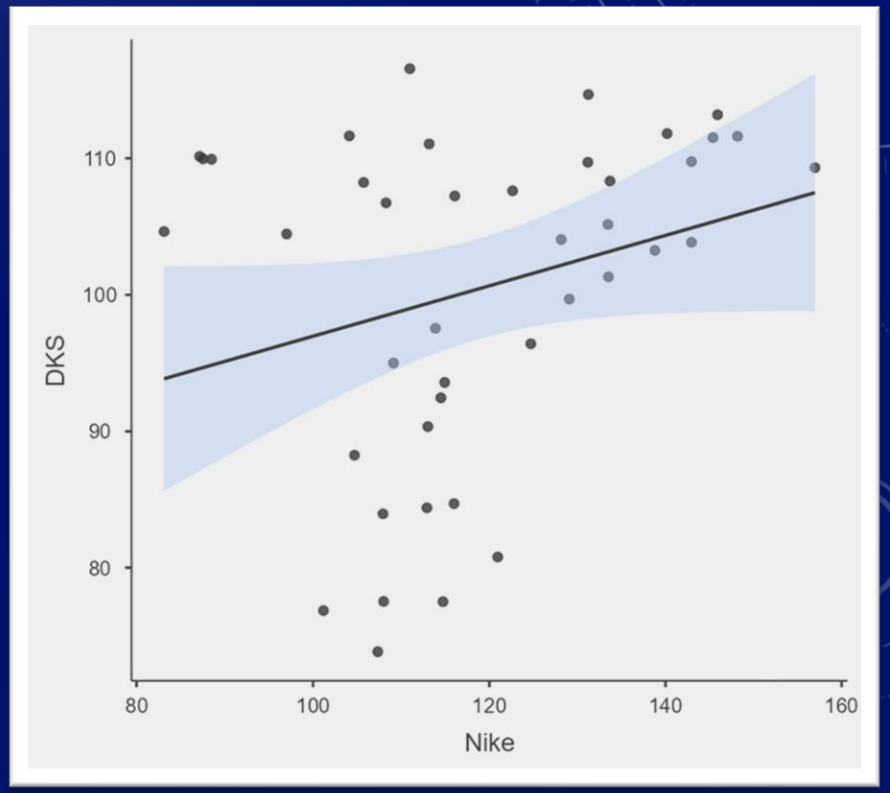
P-value <0.05, suggesting that the correlation is statistically significant.

BIVARIATE STATISTICS: SCATTER PLOT

- In *Jamovi,* Analyses ⟹ Exploration ⟹ Scatterplot
- Regression Line
  - Non
  - Linear
  - Smooth
  - Standard error

BIVARIATE GRAPHS: SCATTER PLOT





CLASS ACTIVITY

- Review quiz
- Class assignment
  - Measure of centrality
  - Measure of dispersion
  - Correlation
  - Scatterplot