Chapter 5: Descriptive Statistics

Objectives:

- Understand the workflow in a market research study
- Know univariate and bivariate statistics and graphs
- Be able to transform data
- Learn how to handle outliers and missing data
- Learn what a codebook is
- Applications in Stata

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

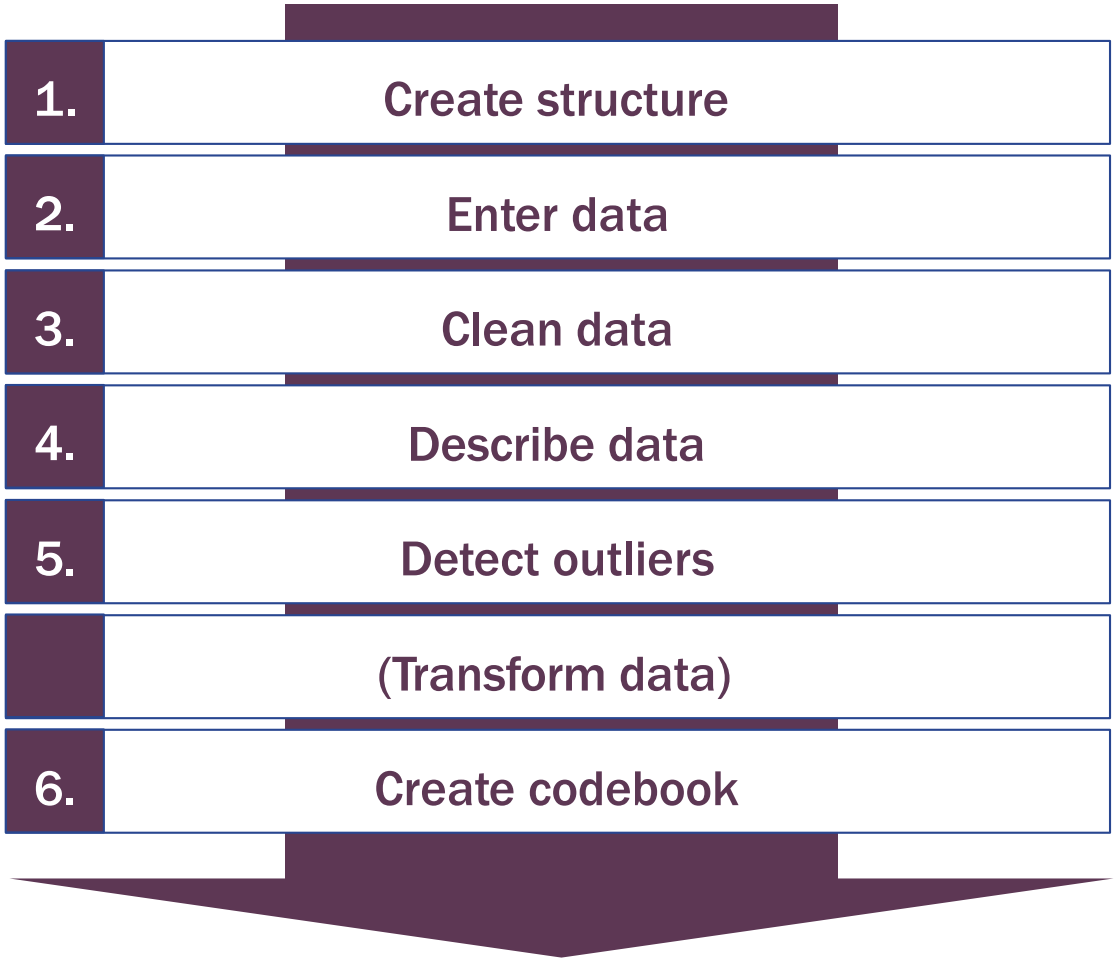
Weeks	1	2	3	4	5	6	7	8	9	10	11
Chap.1 Intro	X										
Qualtrics	X										
Chap.2 MR Process		X									
Chap.3 Data			X	X							
Chap.4 Getting Data				X							
Guest Lecture: Research Ethics					X						
Chap.5 Descriptive Statistics						X					
Chap.10 Communicating the Results							X				
Chap.6 Hypothesis Testing								X			
Chap.7 Regression Analysis									X		
Gauss-Markov and GLM										X	
Guest Lecture: Business Application											X

5. Descriptive Statistics

- Workflow and structure of a market research project
- Applications in Stata

5. Descriptive Statistics
5.1. Workflow and structure

- Market research projects are complex enough. Keep them organised with a structured workflow!
  - Helps to prevent mistakes
  - Helps to collaborate
  - Helps to stay on time
  - Helps to document the whole project



5. Descriptive Statistics
5.1. Workflow and structure

- 1. Create structure:
  - Store different types of Stata files in different directories

Subdirectory	\data	\output	\syntax	\temp	\others
Contents	Data files (.dta and .zip for backup)	Output files, e.g. .gph	do files (.do)	Revised files	Related files (.pdf, .txt, .doc)

  - Variables:
    - Short, clear, meaningful names (e.g. “satisf_room” instead of “hotel11”)
    - Explicit, more informative descriptions (e.g. “Satisfied with the room”)
  - Coding (= assigning values to specific questions)
    - Quantitative data: no special requirements
    - Qualitative data (e.g. open-ended questions):
      Collecting → grouping *(attention: subjectivity)* → recoding

5. Descriptive Statistics
5.1. Workflow and structure

- 3. Data cleaning:
  - Check for all kinds of “errors”
    - Data entry errors
      - Obvious errors (e.g. a “333” instead of a “3” on a three-point ordinal scale): Detection by descriptive statistics like minimum, maximum, range. Correction in the dataset with a new variable. Make a note in your codebook.
      - More subtle errors (e.g. a “2” instead of a “ 3” in the same scale): Detection only by checking randomly selected surveys for correctness.

5. Descriptive Statistics
5.1. Workflow and structure

- 3. Data cleaning:
  - Check for all kinds of “errors”
    - Interviewer cheats
      - Never base an interviewer‘s compensation on the number of completed questionnaires
      - Use testing techniques (next chapters)
      - Ask interviewed people if they have been interviewed
      - Quality-check the interviewer during the field phase

5. Descriptive Statistics
5.1. Workflow and structure

- 3. Data cleaning:
  - Check for errors (ctd.)
    - Missing data
      - Survey non-response (5-25%): Try to prevent it by creating proper questionnaires
      - Item non-response (2-10%):
        Check: is it a systematic (always the same question) or a random error (alternating)?

5. Descriptive Statistics
5.1. Workflow and structure

- 3. Data cleaning:
  - Check for errors (ctd.)
    - Missing data

→ Small number of items are missing? Impute these cases (giving the item the most likely value, e.g., by using Bayesian multiple imputation)

→ Higher amount of missingness: Use Multiple Imputation (MI) Predictive Mean Matching (PMM)

→ MI, PMM, and FIML How-To in Stata:

Piehlmaier, D. M. (2022). Predictive approaches to customer loyalty: the impact of missing data on the predictability of customer loyalty models. In *Handbook of Research on Customer Loyalty* (pp. 129-141). Edward Elgar Publishing.

5. Descriptive Statistics
5.1. Workflow and structure

- 4. Describe data:
  - Types of description
    - Univariate descriptives (one variable at a time)
    - Bivariate descriptives (two variables at a time)
    - Multivariate descriptives (more than two variables at a time)
  - Types of presentation:
    Univariate/bivariate/multivariate graphs/charts/tables
    are easy to understand and interpret by a general audience
  - Types of statistics:
    Univariate/bivariate/multivariate statistics
    The more variables, the more difficult to understand and interpret

5. Descriptive Statistics
5.1. Workflow and structure

- 4. Describe data:

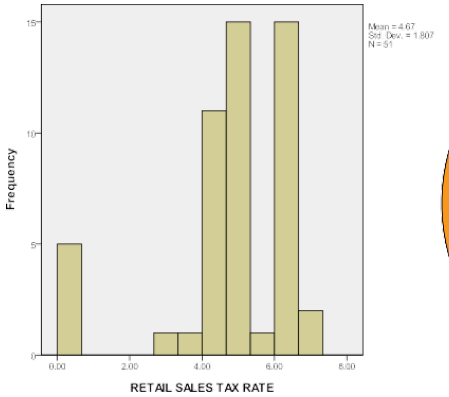
  Overview: Univariate and bivariate descriptives

# of Variables	Graphs/Charts/Tables	Statistics
univariate	Histogram Pie-Chart Frequency table Bar Chart Line Chart Box plot	Centrality: Mean, Mode, Median Dispersion: Minimum/Maximum, Range, Variance
bivariate	Scatter plots Crosstabulation	Correlations Covariance

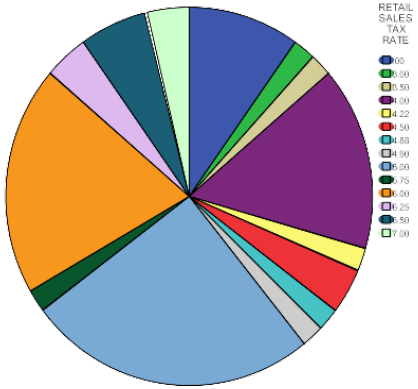
5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe data:

  Overview: <u>Univariate</u> graphs, charts, and tables



Histogram

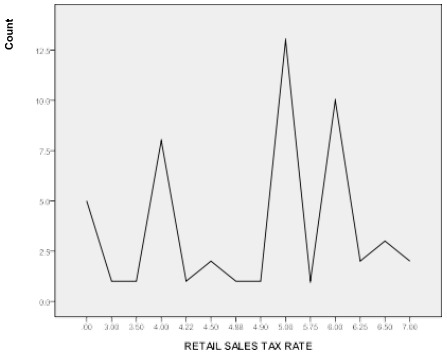


Pie Chart

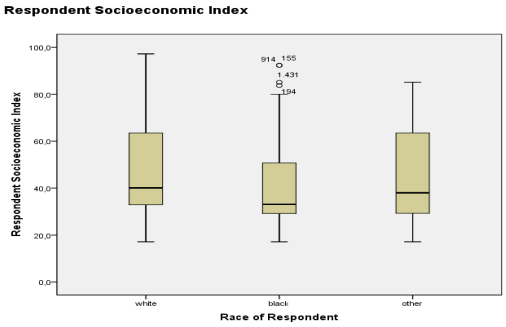
RETAIL SALES TAX RATE

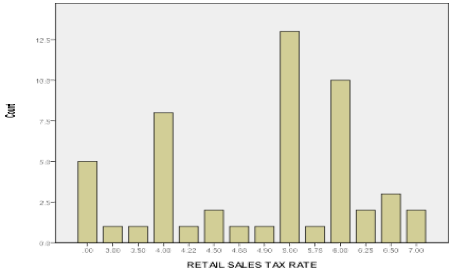
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	5	9.8	9.8	9.8
	3.00	1	2.0	2.0	11.8
	3.50	1	2.0	2.0	13.7
	4.00	8	15.7	15.7	29.4
	4.22	1	2.0	2.0	31.4
	4.50	2	3.9	3.9	35.3
	4.88	1	2.0	2.0	37.3
	4.90	1	2.0	2.0	39.2
	5.00	13	25.5	25.5	64.7
	5.75	1	2.0	2.0	66.7
	6.00	10	19.6	19.6	86.3
	6.25	2	3.9	3.9	90.2
	6.50	3	5.9	5.9	96.1
	7.00	2	3.9	3.9	100.0
	Total	51	100.0	100.0	

Frequency table



Line chart



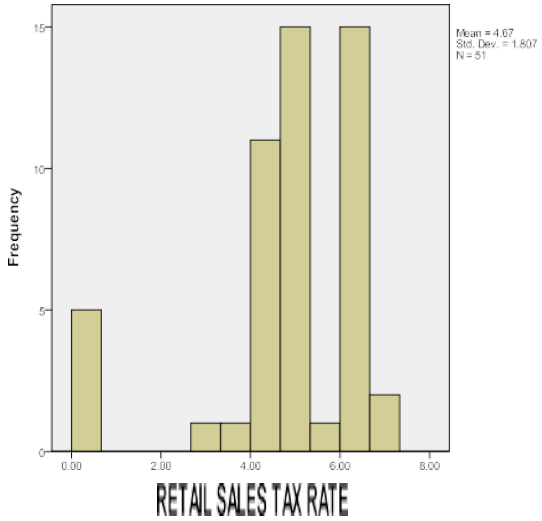


Bar chart

UNIVERSITY OF SUSSEX | BUSINESS SCHOOL

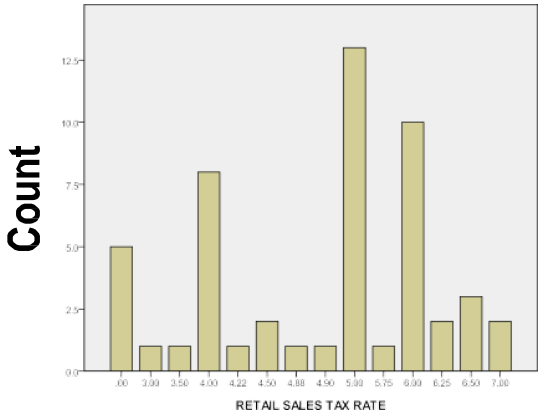
5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Univariate graphs, charts, and tables
  - Histogram:
    - How often does a certain variable value occur? A histogram creates classes for values, the height represents the class-count
    - Good for: Ratio and interval data scale
    - Good: Shows the distribution of a variable
    - Not good: With a very large number of classes



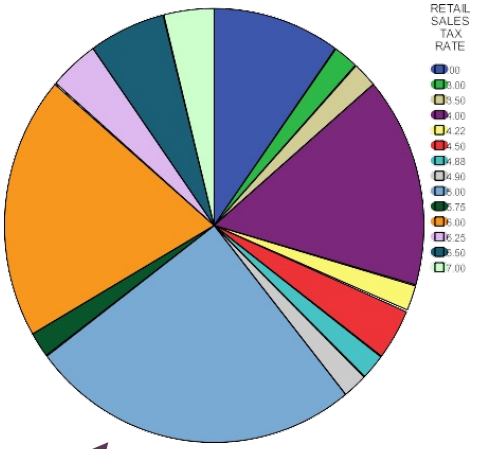
5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Univariate graphs, charts, and tables
  - Bar chart:
    - How often does a certain variable value occur? A bar chart does not create classes but shows each and every value with its frequency.
    - Good: Nominal and ordinal data scales



5. Descriptive Statistics
5.1. Workflow and Structure

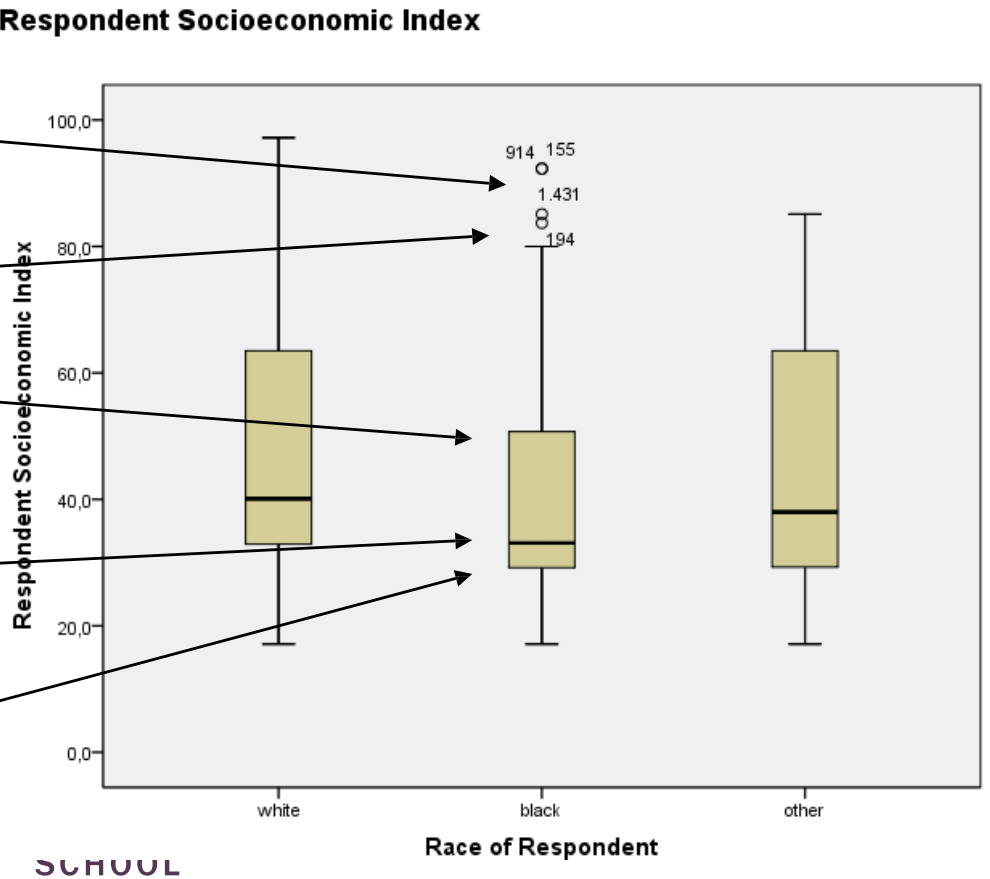
- 4. Describe Data: Univariate graphs, charts, and tables
  - **Pie chart:**
    - **Visual and easy to understand how different values are distributed**
    - **Good: less than 10 categories, interpretation based on "100%" /percentage in general**
    - **Not good: multitude of categories (<10 var.)**



5. Descriptive Statistics
5.1. Workflow and Structure

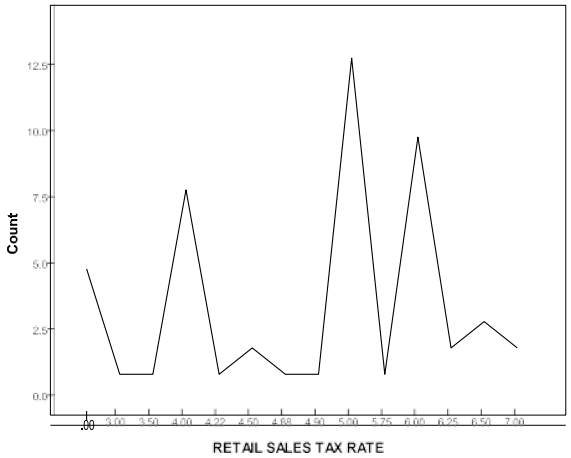
- 4. Describe Data:
  Univariate graphs, charts, and tables

- Box plots
  - Extreme values (“*”)
  - Outliers (“o”)
  - Highest value being no outlier
  - 75th percentile (75% have a value<x)
  - 50th percentile (median)
  - 25th percentile



5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data: Univariate graphs, charts, and tables
  - **Line chart:**
    - **Also describes different variable values occurring in a variable, but connects each value → continuity requirement**
    - **Good: If continuity fits the data (e.g., time series)**



5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data: Univariate graphs, charts, and tables
  - Frequency table
    - Shows all possible values of a specified variable in a dataset.

RETAIL SALES TAX RATE

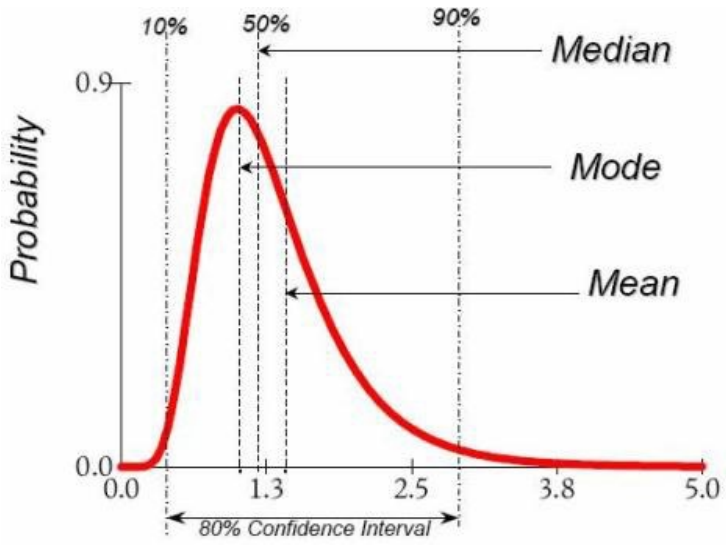
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	5	9.8	9.8	9.8
	3.00	1	2.0	2.0	11.8
	3.50	1	2.0	2.0	13.7
	4.00	8	15.7	15.7	29.4
	4.22	1	2.0	2.0	31.4
	4.50	2	3.9	3.9	35.3
	4.88	1	2.0	2.0	37.3
	4.90	1	2.0	2.0	39.2
	5.00	13	25.5	25.5	64.7
	5.75	1	2.0	2.0	66.7
	6.00	10	19.6	19.6	86.3
	6.25	2	3.9	3.9	90.2
	6.50	3	5.9	5.9	96.1
	7.00	2	3.9	3.9	100.0
	Total	51	100.0	100.0	

5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Univariate graphs, charts, and tables

Univariate statistics

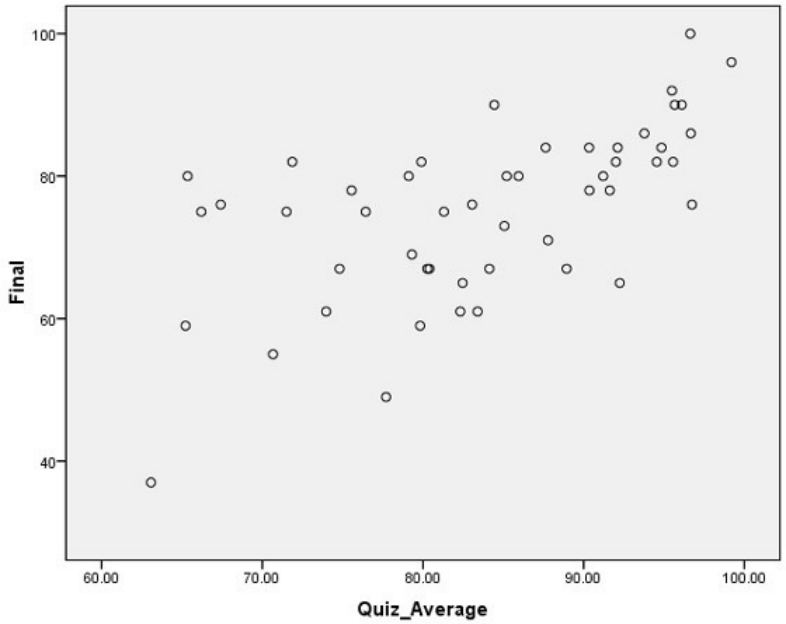
- Centrality
  - Mode
  - Mean
  - Median
- Dispersion
  - Minimum and maximum
  - Variance/standard deviation



5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:

  Overview: Bivariate graphs, charts, and tables



Scatterplot

Anxiety * Tension Crosstabulation

			Tension		
			low	high	Total
Anxiety	low	Count	5	21	26
		% within Anxiety	19.2%	80.8%	100.0%
		% within Tension	25.0%	75.0%	54.2%
		% of Total	10.4%	43.8%	54.2%
	high	Count	15	7	22
		% within Anxiety	68.2%	31.8%	100.0%
		% within Tension	75.0%	25.0%	45.8%
		% of Total	31.3%	14.6%	45.8%
Total		Count	20	28	48
		% within Anxiety	41.7%	58.3%	100.0%
		% within Tension	100.0%	100.0%	100.0%
		% of Total	41.7%	58.3%	100.0%

Crosstabulation

5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Bivariate graphs, charts, and tables
  - Scatterplot
    - Two variables, one on the x-axis, one on the y-axis (sometimes a third variable on the z-axis)
    - Shows how variables relate to each other
    - Good: For interval and ratio data scales

5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Bivariate graphs, charts, and tables
  - Crosstabulation
    - Creates a contingency table from the frequency distribution of a variables
    - Good: For all data scales

5. Descriptive Statistics
5.1. Workflow and Structure

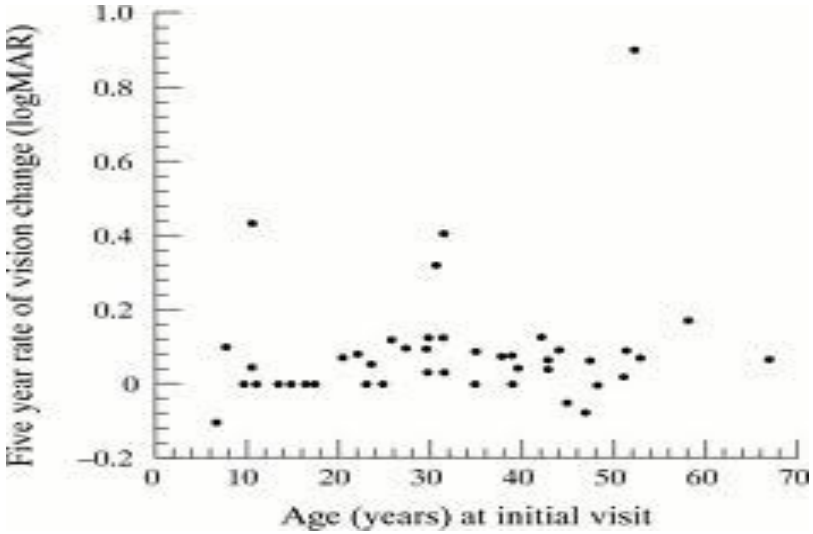
- 4. Describe Data:
  Bivariate statistics
  - Covariance
    - Describes how two variables are associated
    - Is the degree to which variables vary together
  - Correlation
    - Describes how strongly two variables are associated
      - +1 (–1) stands for a perfect positive (negative) correlation
      - 0 stands for independence of two variables (+/- 0.1)
      - In practice: < 0.30 (weak) / 0.30-0.49 (moderate) / >0.50 (strong)

5. Descriptive Statistics
5.1. Workflow and Structure

- 4. Describe Data:
  Bivariate statistics
    - Common types of correlation coefficients
      - Bravais-Pearson for interval and ratio scaled variables
      - Spearman if at least one variable is ordinally scaled
      - Kendalls Tau for few cases with ordinally scaled variables (Likert)

5. Descriptive Statistics
5.1. Workflow and Structure

- 5. Detecting outlier:
  - Extreme values that seem detached from other values
    (hint: mean >> median and mode)
  - Different forms:
    - Produced by errors during data collection / data entry
    - Real values
    - Certain extremely rare combinations of variables



5. Descriptive Statistics
5.1. Workflow and Structure

- 5. Detecting outlier:
  - Finding them:
    - Minimum / Maximum
    - Scatterplot (for rare combinations)
  - What most books say:
    - If they are real values: delete them from the data set
    - Typing or data entry error: correct if possible, otherwise delete them
    - Case of doubt: better delete them than keep
  - My recommendation:
    - Transform the variable, e.g., through log-linearization
    - Real data should never be deleted or altered (academic code of conduct)

5. Descriptive Statistics
5.1. Workflow and Structure

- Transforming data:
  - Optional in market research
  - Reason for transforming:
    - Grouping, e.g., income will be separated in two groups: high and low
    - Procedure to prepare for other statistical analyses (e.g., standardization of variables for cluster analysis)
    - Binary (dummy) variables, e.g., if some variable should get the status true (=1) or false (=0)
    - Creating constructs from more than one variable (e.g., satisfaction with the service at Subway regarding friendliness of employees and wait time)
    - Aggregation: Bringing variables from a lower level to a higher level

5. Descriptive Statistics
5.1. Workflow and Structure

- 6. Creating a codebook
  - Contains all important details regarding the data of a study
  - Content:
    - Introduction (aim of the data collection)
    - Questionnaire (consisting of all versions and all pretest-results)
    - Description of the variables (verbally, including the measurement level)
    - Summary statistics (for a fast glance)
    - Datasets (name of all the datasets used in the study)

5. Descriptive Statistics

- Workflow and structure of a market research project
- **Applications in Stata**