



Data Munging

Areej Alasiry

Lecture Outline

- Data Collection
- Data Cleaning
- Descriptive Statistics
- Data Visualisation
- Exploratory Data Analysis (EDA)

Data Collection

Collecting Data

- It is the most critical issue in data science
- Hunting for data (who has the data that I need?)
 - Companies (business and privacy issues)
 - Government (Freedom of Information Act (FOI))
 - Academic data set
 - Collect your own dataset
- Scraping
 - Spidering
 - Scrapping
- Logging

Objectives of Data Exploration

- Data Understanding
- Data Preparation
- Data Science Tasks
- Interpreting the results

Descriptive Statistics

Datasets

- IRIS Dataset
 - Features
 - Observation
 - Dimension
- Types of data
 - Numeric / Continuous
 - Categorical / Nominal

Descriptive Statistics

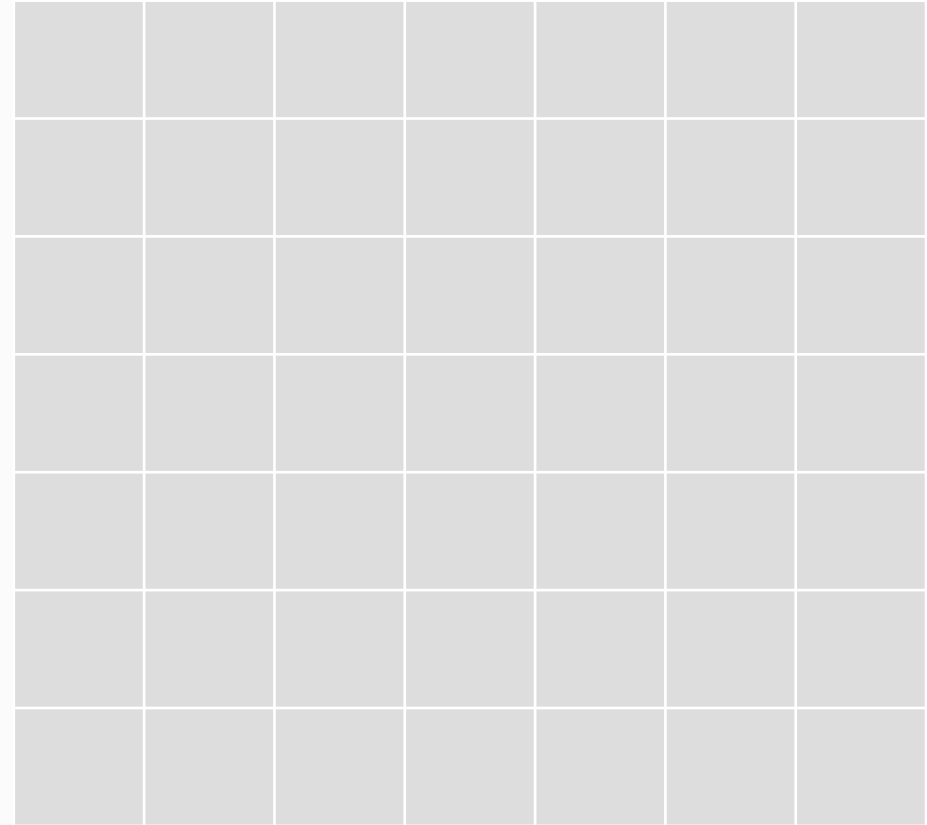
- Central Tendency
 - Mean
 - Median
 - Mode
- Spread
 - Range, variance, and standard deviation
- Shape of dataset distribution
 - Symmetry, Skewness, and kurtosis

Univariate Exploration

- Central Tendency
 - Mean
 - Median
 - Mode
- Spread
 - Range
 - Deviation (Variance, and Standard Deviation)

Multivariate Exploration

- **Central Data Point**
 - Hypothetical data point with the most typical attribute values
- **Correlation**
 - The statistical relationship between two variables



Data Visualisation

- Motivation
 - Dense information
 - Relationships
- Univariate Visualisation
- Multivariate Visualisation
- High dimensional data visualisation



Data Visualisation

Univariate Visualisation

- Histogram

From: Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande 2019

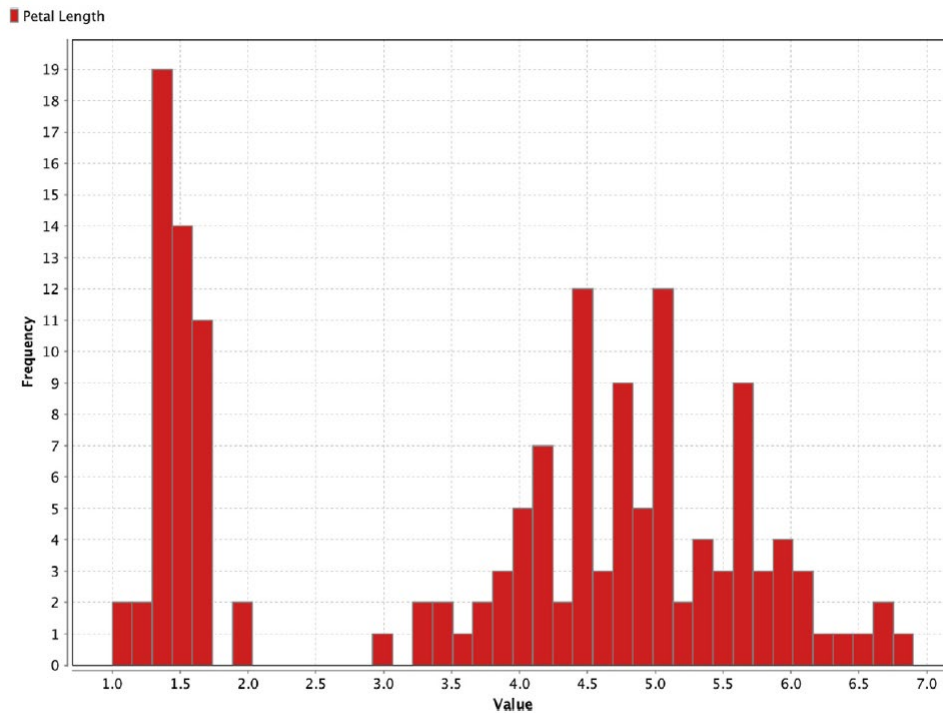


FIGURE 3.5

Histogram of petal length in Iris dataset.

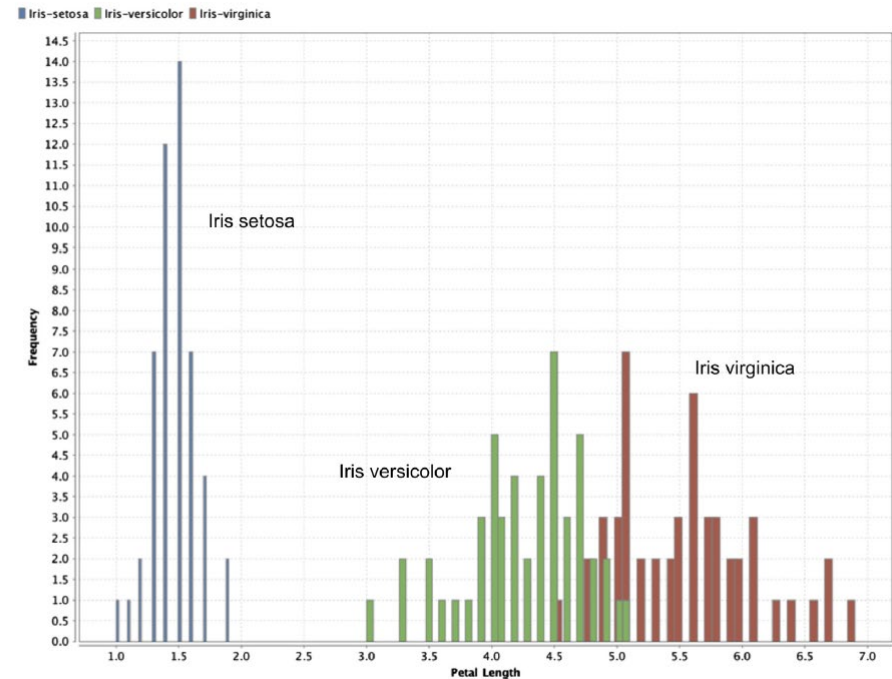
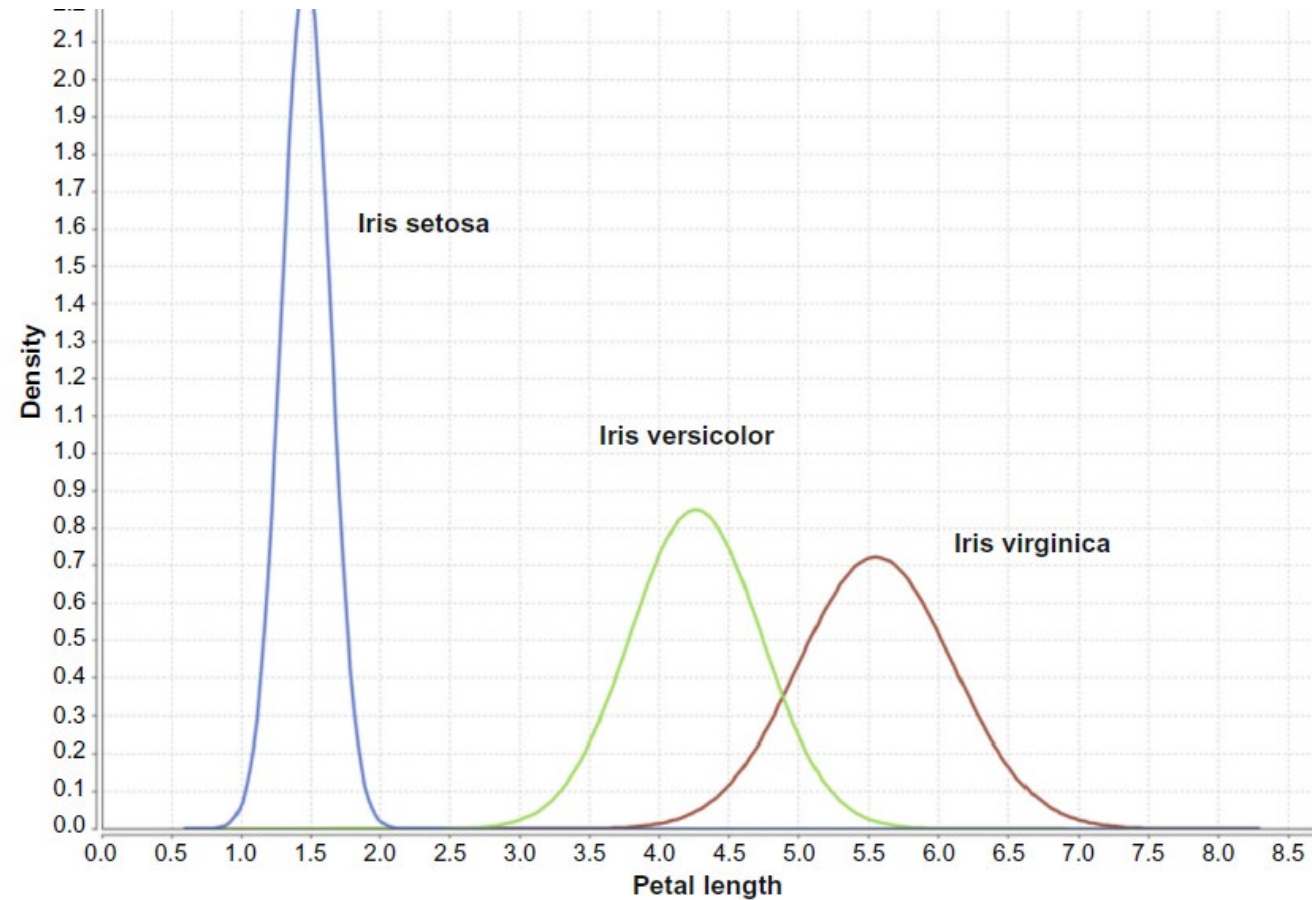


FIGURE 3.6

Class-stratified histogram of petal length in Iris dataset.

Univariate Visualisation

- Distribution Chart



From: Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande 2019

FIGURE 3.9

Distribution of petal length in Iris dataset.

Multivariate Visualisation

- Scatter Plot

From: Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande 2019

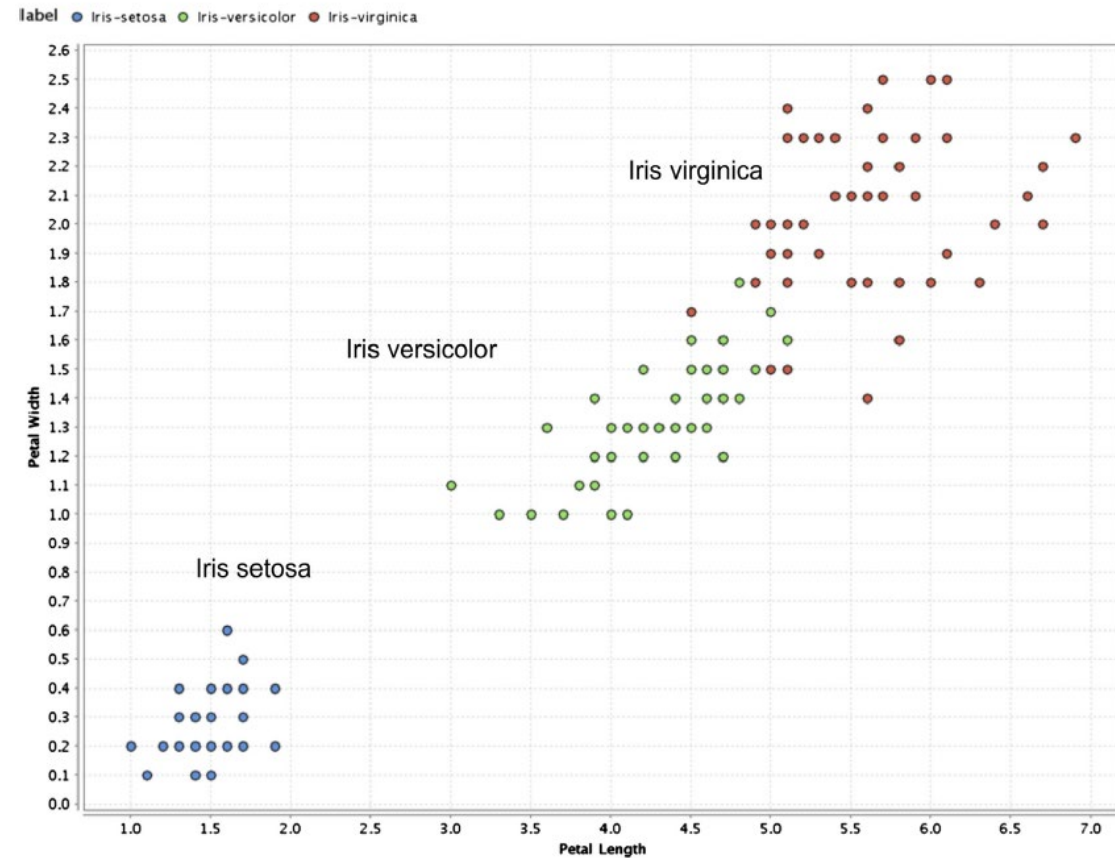


FIGURE 3.10
Scatterplot of Iris dataset.

Multivariate Visualisation

- Density Chart

From: Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande 2019

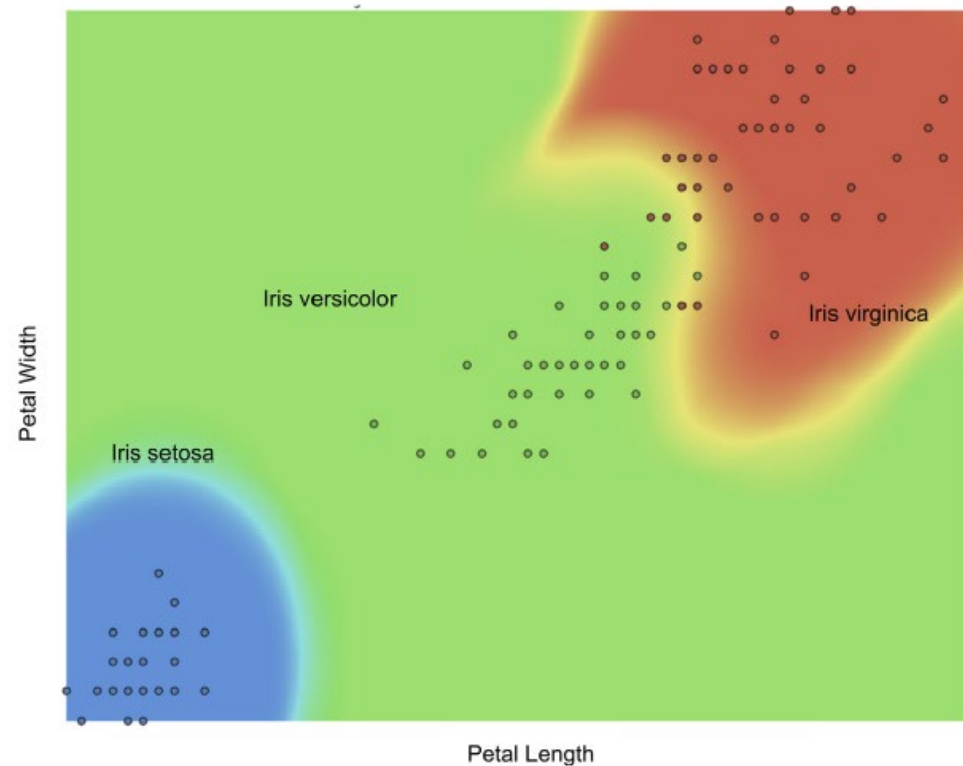


FIGURE 3.14

Density chart of a few attributes in the Iris dataset.

High Dimensional Data Visualisation

- Parallel Chart

From: Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande 2019

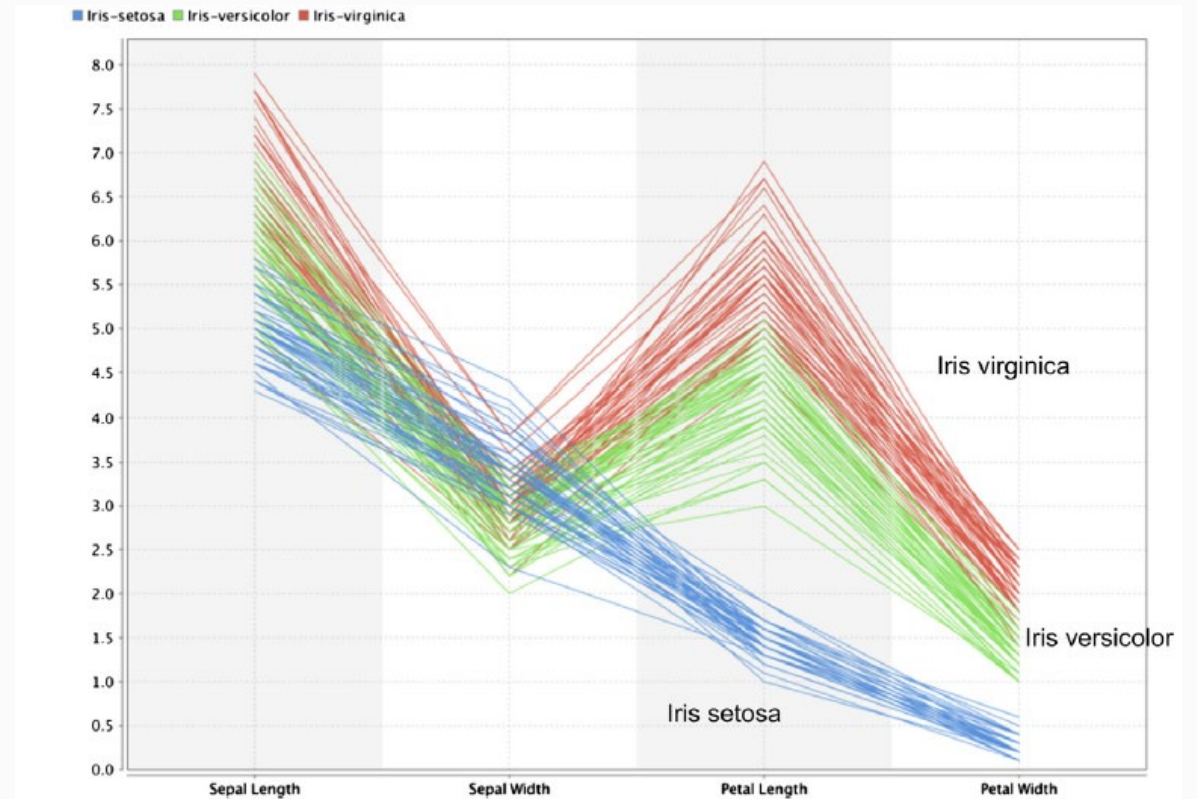


FIGURE 3.15

Parallel chart of Iris dataset.

Cleaning Data

Errors and Artifacts

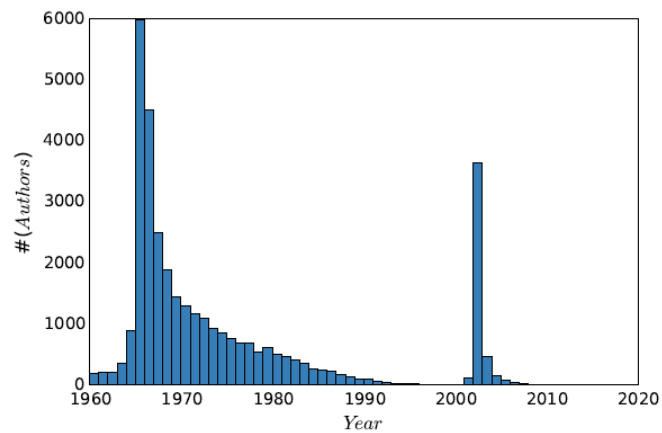


Figure 3.2: What artifacts can you find in this time series, counting the number of author's names first appearing in the scientific literature each year?

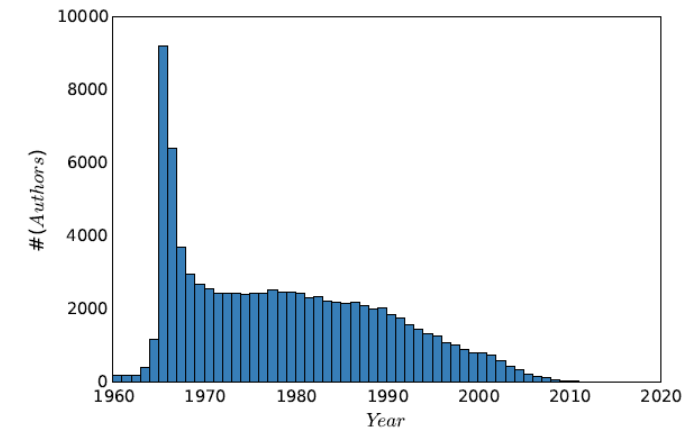




Figure 3.3: The cleaned data removes these artifacts, and the resulting distribution looks correct.

From: The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.

Data Compatibility

- Example: Comparison of weights in pounds versus weights in kilograms!
- Unit Conversion
- Numerical Representation
- Name Unification
- Date/Time Unification
- Financial Unification

Missing Values

- What should we do with missing values?
 -  Drop all records
 - Estimate the missing values
 - Heuristic-based imputation
 -  Mean value imputation
 - Random value imputation
 - Imputation by nearest neighbour
 - Imputation by interpolation

Outlier Detection

- Often created by data entry mistakes
- Identify the largest and smallest values in each column
- Visual inspection can be used to identify outliers
- We could exclude the records with the outliers

Data Exploration Roadmap

- Organise the dataset
- Central Tendency
- Spread of each attribute
- Visualise the distribution of each attribute
- Pivot the data
- Investigate outliers
- Examine the relationships between attributes
- Visualise high-dimensional datasets

Session Mind Map

References

- Data Science: Concepts and Practice. 2nd Edition by Vijay Kotu and Bala Deshpande, ISBN: 978-0-12-814761-0©2019. [Chapter 3]
- *The Data Science Design MANUAL. Steven S. Skiena, ISBN: 978-3-319-55444-0 ©2017.[Chapter 3]*