Exploratory Data Analysis

MKT 566

Instructor: Davide Proserpio

What we will learn

How to use visualization to explore your data in a systematic way (also called **Exploratory Data Analysis** or **EDA**)

- Generate questions about your data
- Search for answers by visualizing, transforming, and modelling your data
- Use what you learn to refine your questions and/or generate new questions.

(Partially based on Chapter 7 of R for Data Science)

EDA Goal

- Develop an understanding of your data
- The easiest way to do this is to **use questions** as tools to guide your investigation
- EDA is fundamentally a **creative process**

Summary statistics

Computing **summary statistics** is also very useful and should be done at the beginning of any analysis

- Mean
- Standard deviation
- Min
- Max
- Median
- Number of missing values

These descriptive statistics can provide valuable insights into your data!

Data cleaning

During EDA, you should perform what is called **data cleaning**, which involves things like

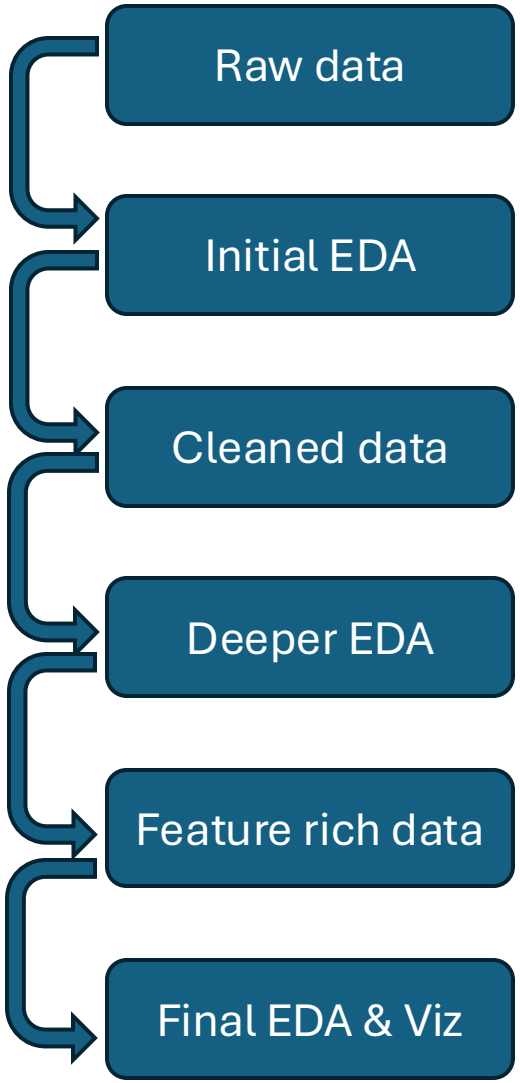
- Finding and removing erroneous data
- Decide what to do with outliers
- Decide what to do with missing data
- …

A few more things: Feature engineering

Very often you will create new variables using existing ones, this process is referred to as **feature engineering,** e.g.:

- Extract the week number from a date variable
- Sum up total ad spend across all advertising channels
- Compute the cumulative average review ratings for all products
- …

General workflow



Spot errors, outliers, missing data → Data Cleaning

Uncover patterns, group effects, perform feature engineering

Elea McDonnell Feit • 1st
Professor of Marketing and Associate Dean of Research
3w • Edited •

Rule 1: Never trust your data. There are all sorts of oddities in any data set and you want to learn about them.

Rule 2: Summarize and plot your data. Over and over. All kinds of summaries. Not just the default summaries in the software you are using. Ask questions like, "How many subscriptions are there that last more than 10 years?" and "How prominent is the holiday seasonality?"

Rule 3: Compare the summaries to your priors. When something in the data seems seems off, pause and consider the consequences for your analysis. Keep a list of these oddities in your lab notebook.

Rule 4: Chase oddities that might be consequential. Form a hypothesis that explains what is going on. Then think of a way to test it. Do more summaries. Look at the documentation again. Ask someone. Observe the process the data describes. Don't stop until you are pretty sure you know what is going on.

How do we explore the data?

Variation and covariation

- There is no rule about which questions you should ask to guide your research.
- However, two types of questions will always be useful for making discoveries within your data. You can loosely word these questions as:
  - What type of **variation** occurs **within** my variables?
  - What type of **covariation** occurs **between** my variables?

What is variation?

Variation is the tendency of the values of a variable to change from measurement to measurement

- We are interested in “within-the-same-variable” patterns

We can observe variation

- For the **same continuous variable** over two different measures, e.g.
  - Measure temperature at 10 am and 4 pm
- For the **same categorical variable** across different “subjects”
  - Eye color across individuals

How to visualize variation

The best way to understand patterns of variation is to visualize:

- The **distribution** of the variable's values
- How the **variable evolves** over repeated observations (e.g., when we have repeated observations over time, i.e., panel data or time series)

Visualizing distributions

Different charts depending on the type of variable

- Continuous → density/histogram
- Categorical/discrete → bar chart

Visualizing distributions

Example with the marketing dataset from the library ‘datarium’

```
> head(marketing)
  youtube facebook newspaper sales
1  276.12    45.36     83.04 26.52
2   53.40    47.16     54.12 12.48
3   20.64    55.08     83.16 11.16
4  181.80    49.56     70.20 22.20
5  216.96    12.96     70.08 15.48
6   10.44    58.68     90.00  8.64
```

Visualizing distributions

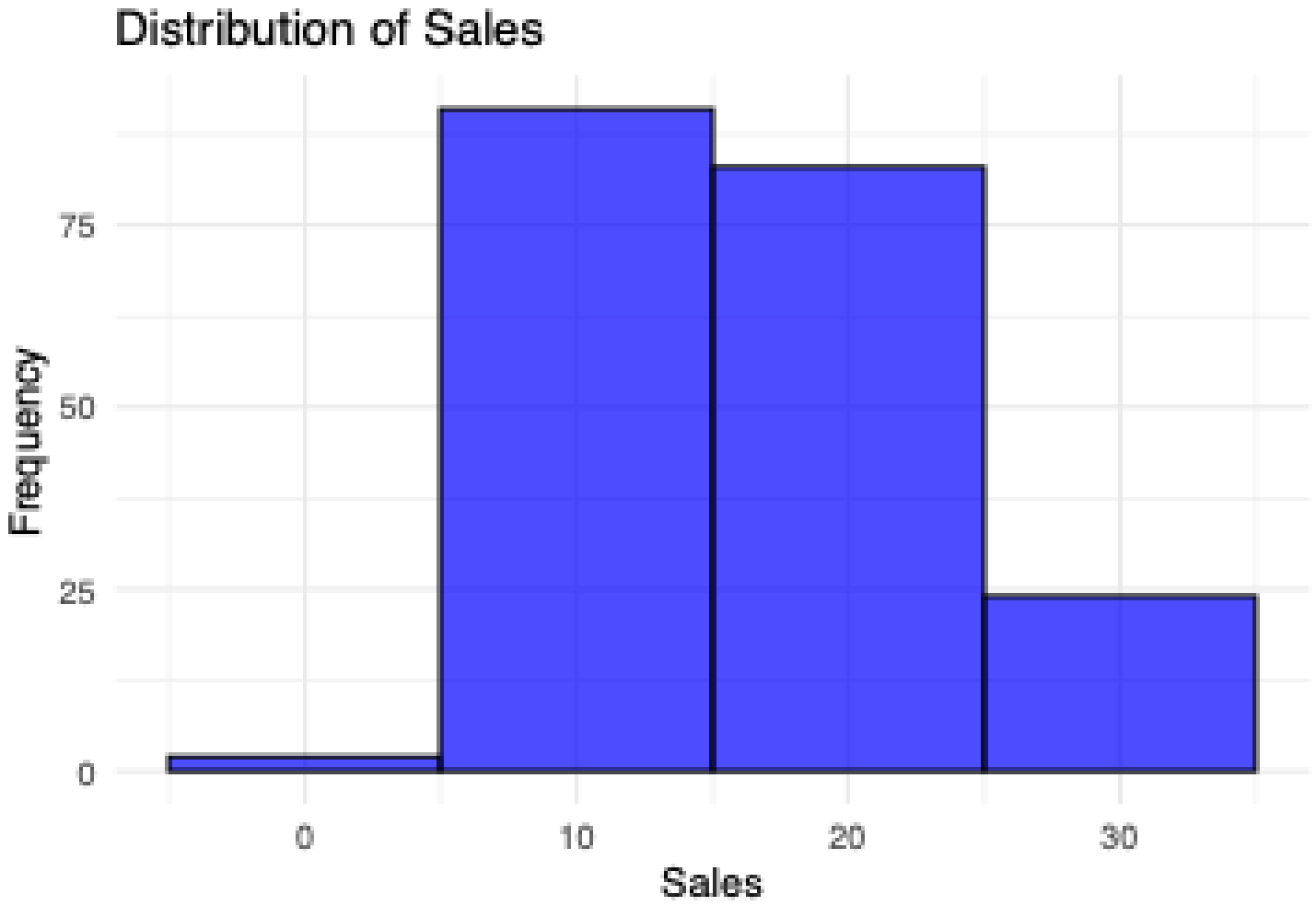
Example with the marketing dataset from the library ‘datarium’

```
> summary(marketing)
    youtube          facebook       newspaper          sales      
 Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92  
 1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45  
 Median :179.70   Median :27.48   Median : 30.90   Median :15.48  
 Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83  
 3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88  
 Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40  
```

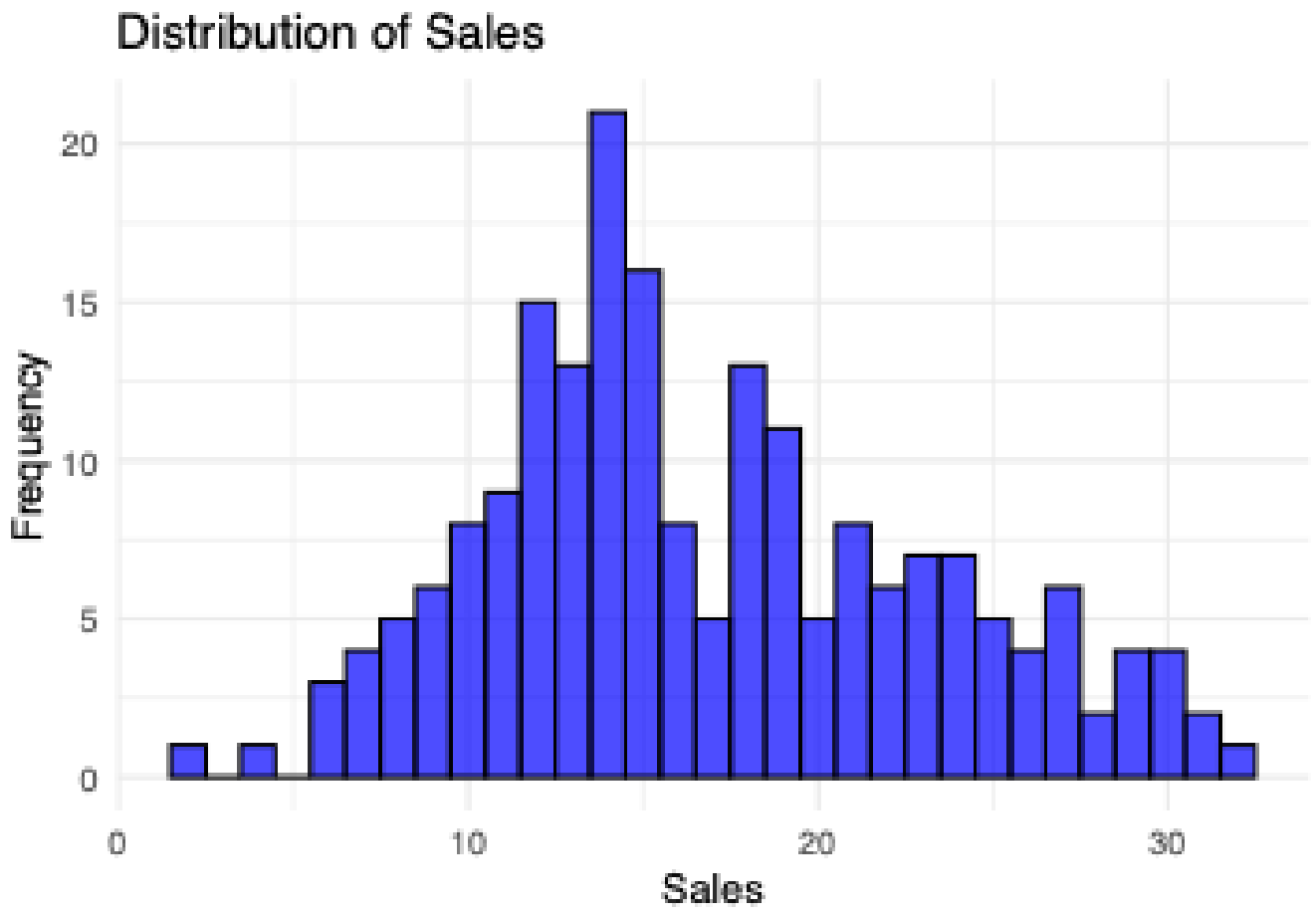
Visualizing distributions

Let’s look at the distributions of the four variables

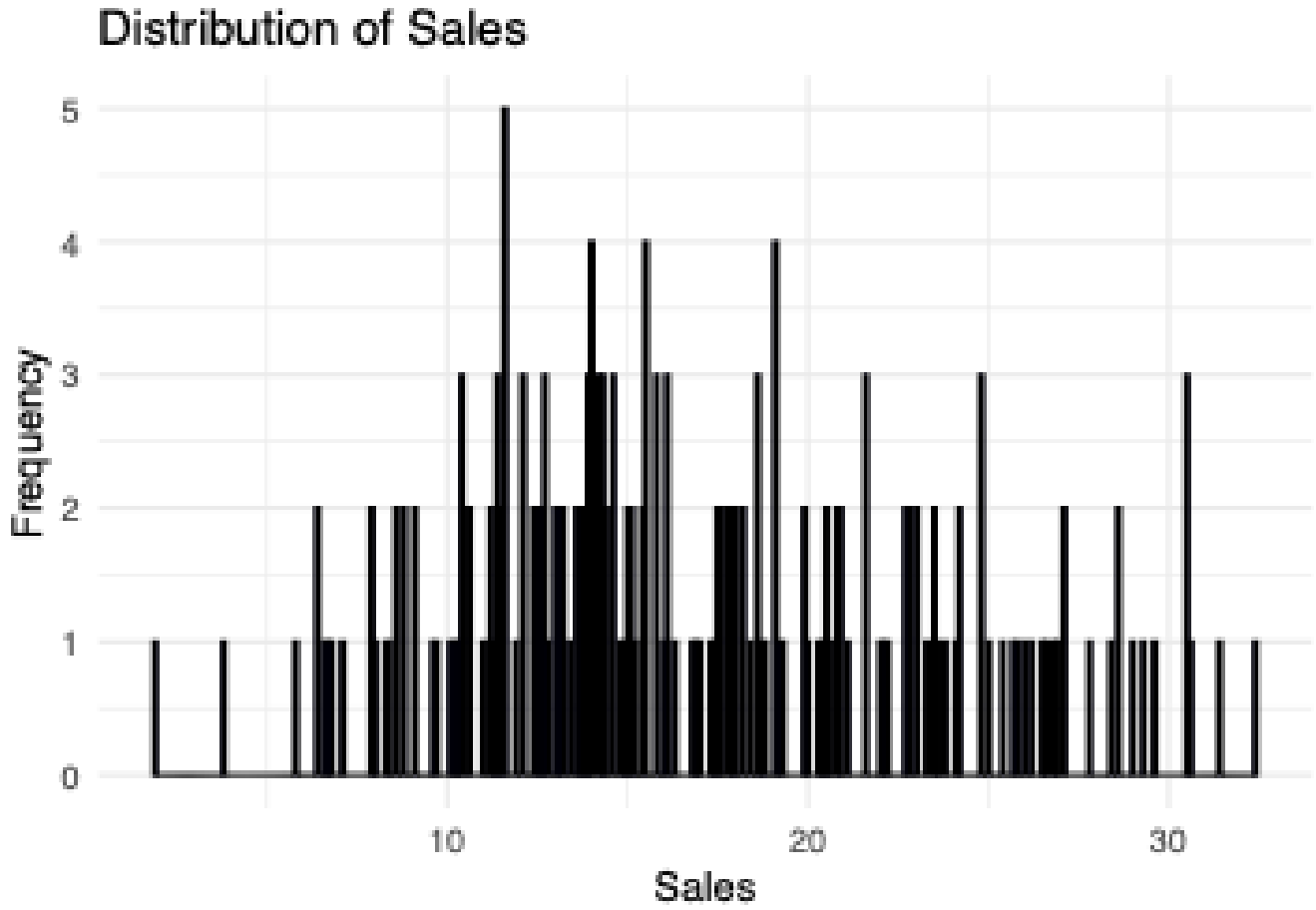
Visualizing distributions



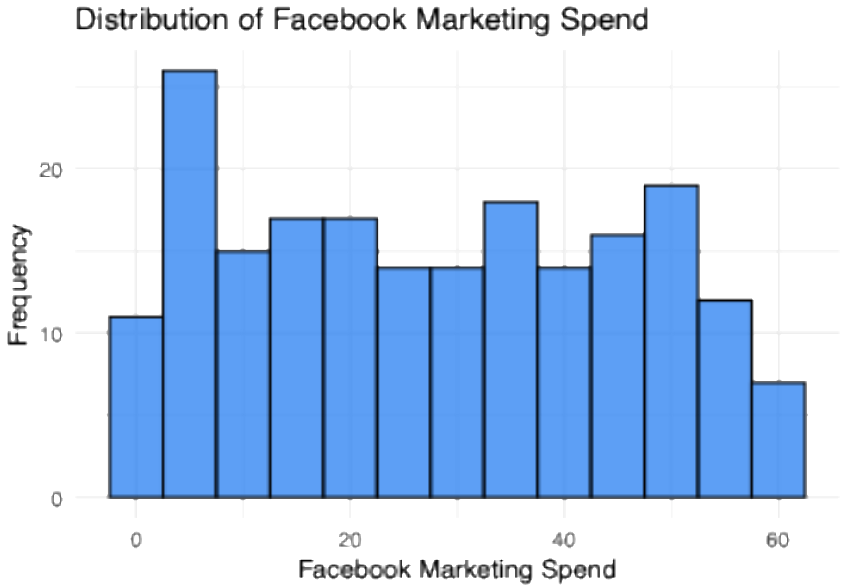
Visualizing distributions

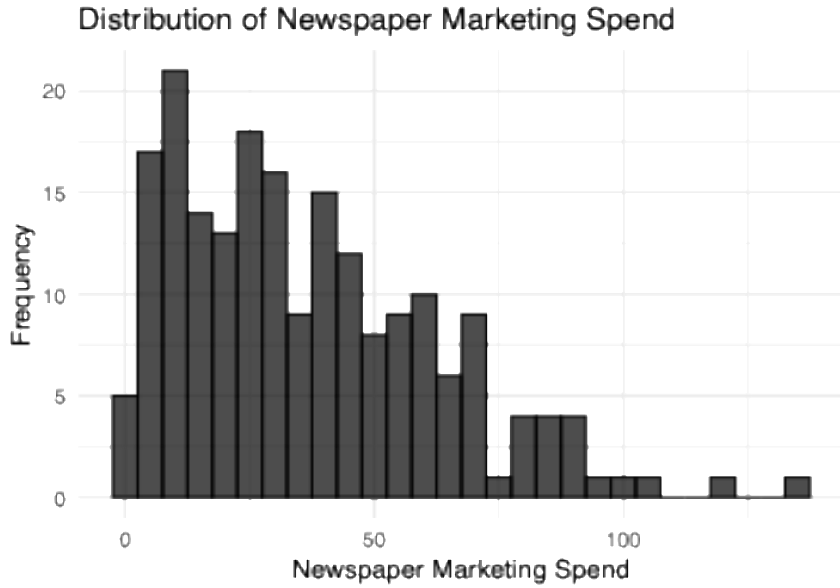


Visualizing distributions



Visualizing distributions





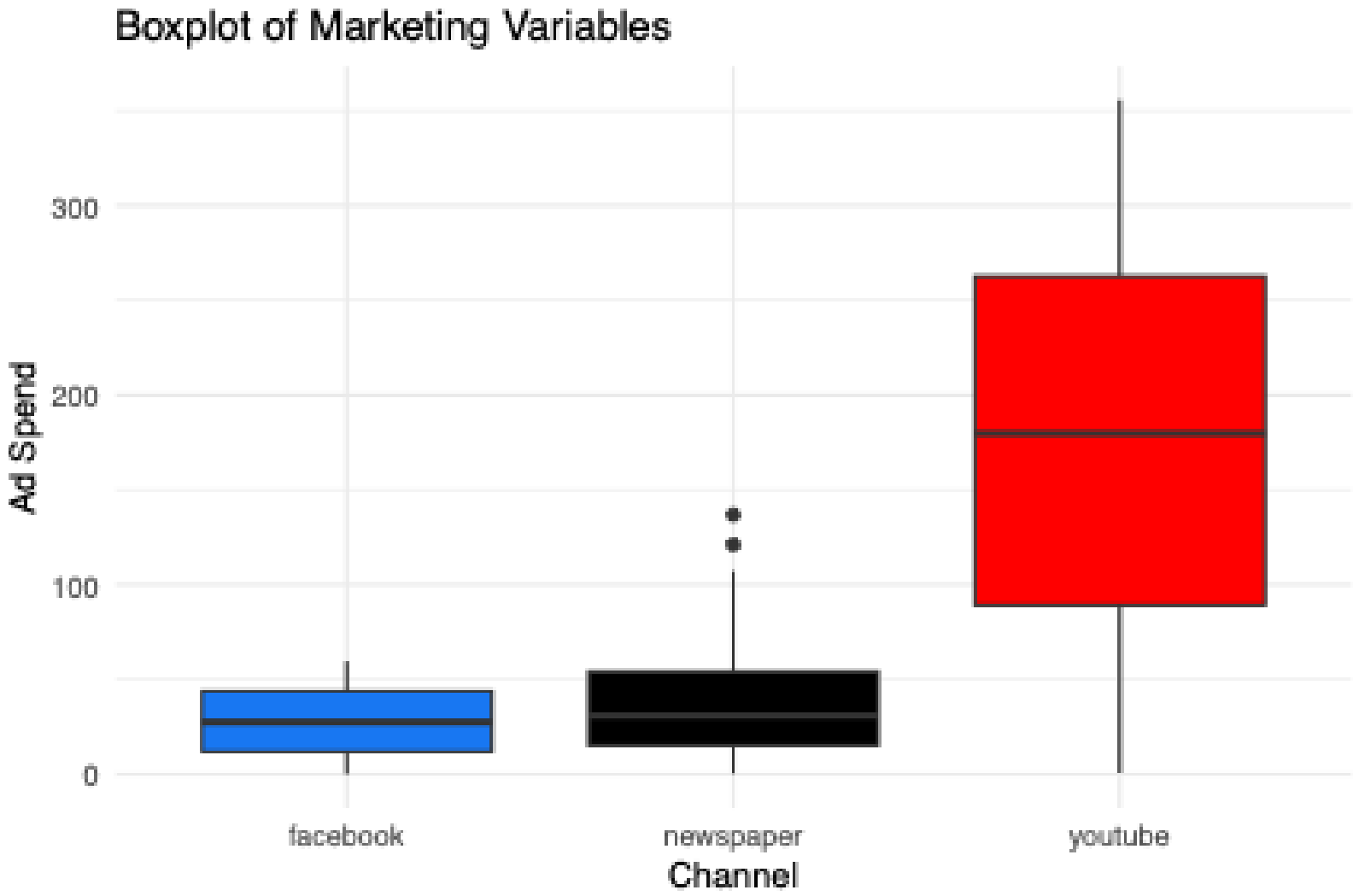
What can we learn from distribution charts

- Common vs rare values (high vs low bars)
- Unusual patterns

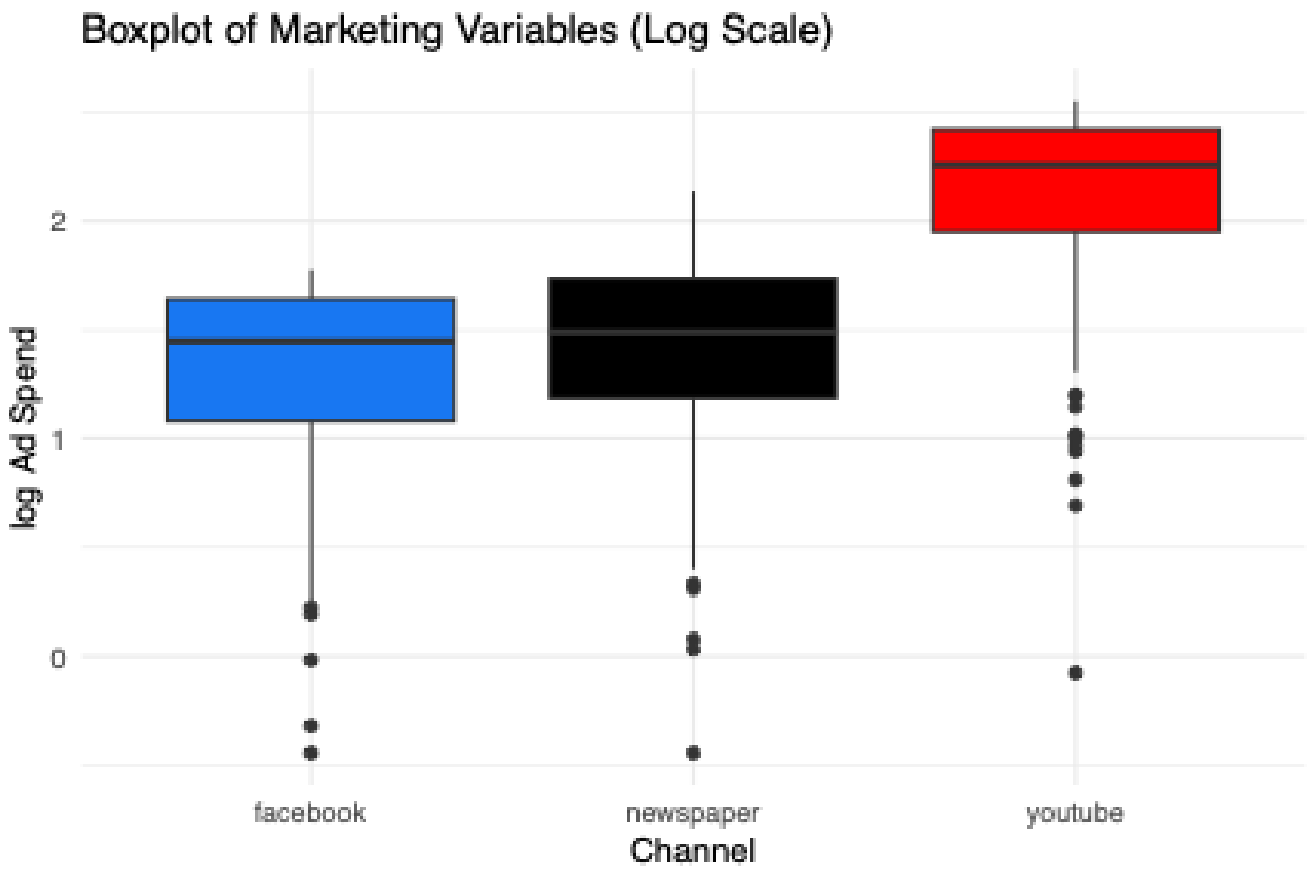
How to specifically look for outliers

- Box plots or scatter plots are often more useful
- Box plots automatically flag outliers using a mathematical rule
  - Any point beyond 1.5 × IQR (Interquartile Range)
  - IQR = Quartile 3 – Quartile 1
  - Lower bound = Q1 – 1.5 IQR
  - Upper bound = Q3 + 1.5 IQR

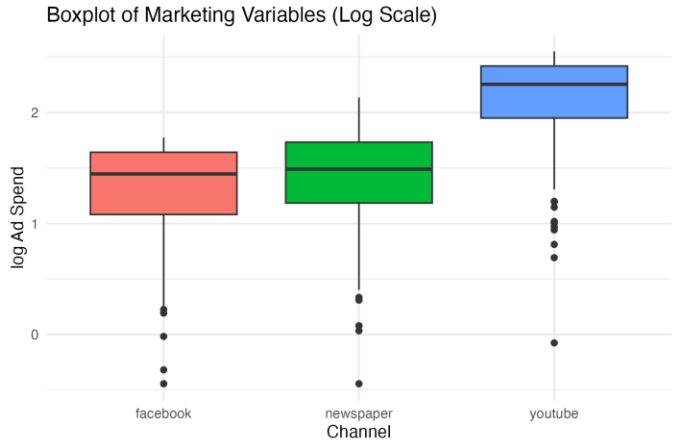
Example with ad spend



Example with logged ad spend



Example with logged ad spend



Log makes small values more visible

- It expands small values and compresses large ones

When and why use log:

- To handle skewed data (common in marketing spend, sales, and clicks distributions).
- To make patterns at the low end visible when a few big values dominate.
- To turn multiplicative growth into a straight line (e.g., exponential trends).

Why log matter in EDA

- **Log makes small values more visible**: you don't lose them at the bottom of the plot.
- It can also reveal **negative outliers** (very low spend, clicks, sales) that were hidden on the linear scale.
- It emphasizes **relative differences** (ratios) instead of absolute differences.
  - Log(a) – log(b) = log(a/b)
  - Example: sales increase from $100 to $110:
    - Actual % change: $\frac{(110-100)}{100} = 0.10 \rightarrow 10\%$
    - Using logs: log(110) - log(100) = log(1.1) ~ 0.095 → 9.5%

Best practice for outliers

- It’s good practice to repeat your analysis with and without the outliers
- If they have minimal effect on the results, and you can’t figure out why they’re there, it’s reasonable to replace them with missing values and move on.
- If they have a substantial effect on your results, you shouldn’t drop them without justification. You’ll need to figure out what caused them (e.g., a data entry error) and disclose that you removed them in your write-up.
- Generally, rely on robust statistics.

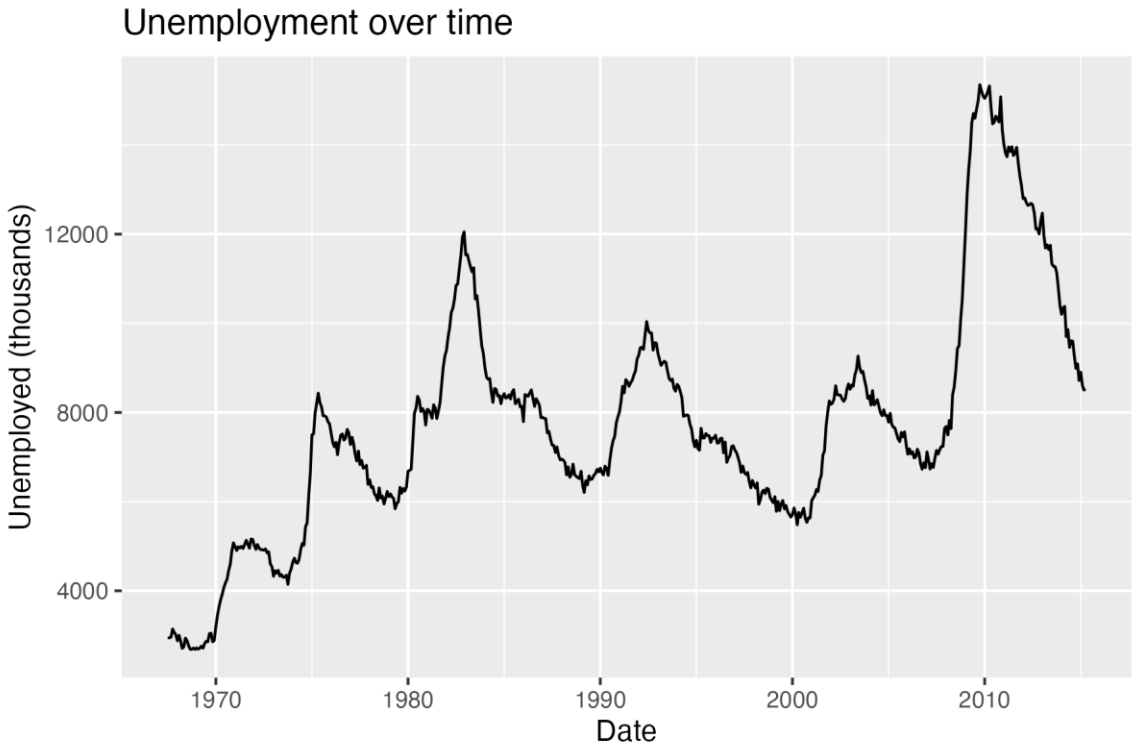
Visualizing trends

Example with ggplot2's Economics dataset:

- This dataset was produced from US economic time series data available from https://fred.stlouisfed.org/, containing info about several economic indicators

```
> head(economics)
# A tibble: 6 × 6
  date         pce    pop psavert uempmed unemploy
  <date>     <dbl>  <dbl>   <dbl>   <dbl>    <dbl>
1 1967-07-01  507. 198712    12.6     4.5     2944
2 1967-08-01  510. 198911    12.6     4.7     2945
3 1967-09-01  516. 199113    11.9     4.6     2958
4 1967-10-01  512. 199311    12.9     4.9     3143
5 1967-11-01  517. 199498    12.8     4.7     3066
6 1967-12-01  525. 199657    11.8     4.8     3018
```

Visualizing trends

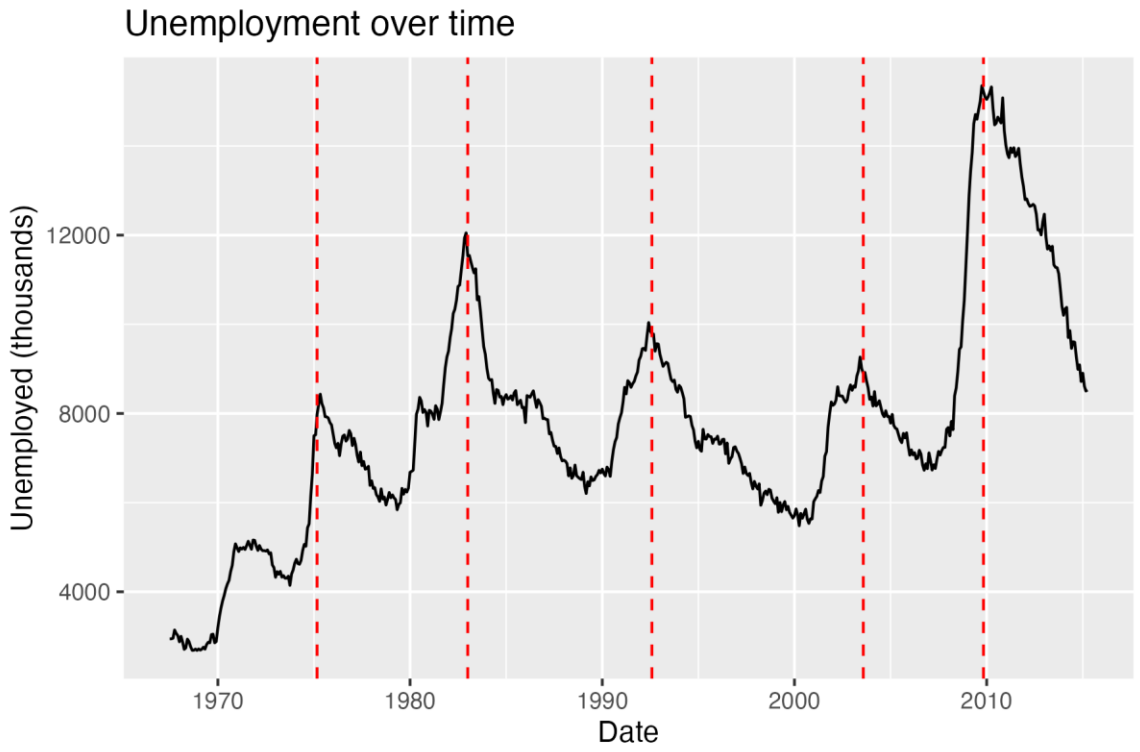


What can we learn from trends?

- Long-Run Directional Trends
- Seasonality & Cyclicality
- Anomalies & Outliers

What can we learn from trends?

- Long-Run Directional Trends
- Seasonality & Cyclicality
- Anomalies & Outliers



What can we learn from trends?

- Long-Run Directional Trends
- Seasonality & Cyclicality
- Anomalies & Outliers