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

- 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?

Covariation

- **Covariation** is the tendency for the values of two or more variables to vary together in a related way
- The best way to spot covariation is to **visualize the relationship between two or more variables**
- How you do that should again depend on the type of variables involved

Visualizing covariation

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
```

A categorical and continuous variable

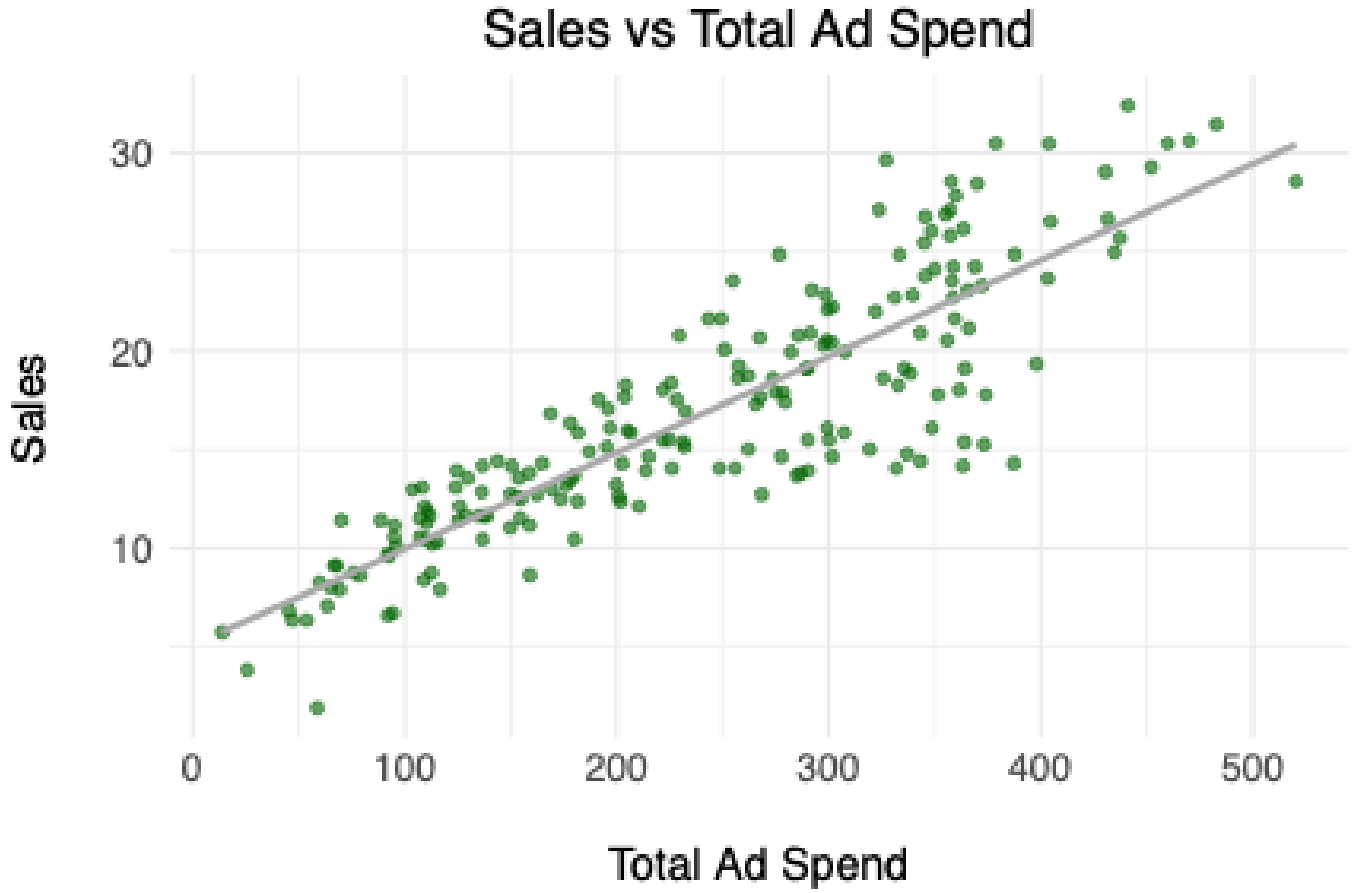
How can we visualize sales by ad spend?

Two continuous variables

How can we visualize sales by ad spend?

Two continuous variables

How can we visualize sales by ad spend?

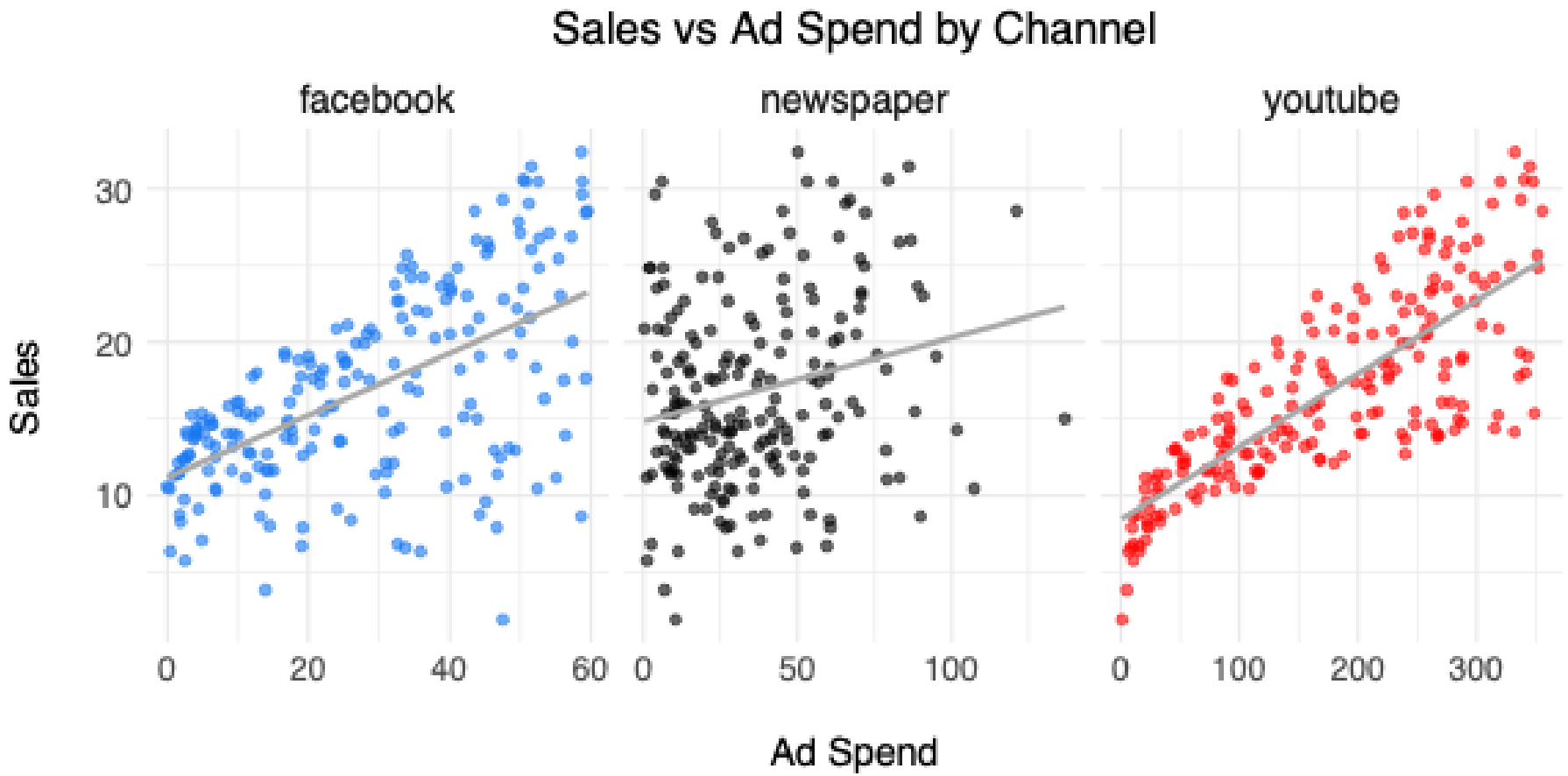


Two continuous variables

Can we do a more informative viz?

Two continuous variables

Can we do a more informative viz?



One categorical and one continuous variable

Let’s use the simulated marketing dataset we explored last week

```
> head(df)
   CustomerID   Age Gender  Device Channel Ad_Spend Clicks Purchases Revenue
        <int> <int> <char>  <char>  <char>    <num>  <int>     <int>   <num>
1:          1    54      M  Mobile  Social   718.60     95         6  149.16
2:          2    18      F  Mobile  Search   233.00     34         1   22.22
3:          3    42      F  Mobile  Search   122.51     18         0    0.00
4:          4    27      F Desktop  Social   198.78     19         1   13.22
5:          5    53      F  Mobile  Social   145.19     19         4  150.48
6:          6    35      M Desktop   Video   125.74      9         0    0.00
```

One categorical and one continuous variable

Which viz can we use to explore the relationship between purchases and gender?

One categorical and one continuous variable

Which viz can we use to explore the relationship between purchases and gender?