جامعة الطائف
TAIF UNIVERSITY

PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science

جامعة الطائف
TAIF UNIVERSITY

CHAPTER 9: COMMUNICATING DATA

- WHY DOES COMMUNICATION MATTER?
- IDENTIFYING EFFECTIVE AND INEFFECTIVE VISUALIZATIONS
  - Scatter plots
  - Line graphs
  - Bar charts
  - Histograms
  - Box plots
- WHEN GRAPHS AND STATISTICS LIE
  - Correlation versus causation


Process Page
LOGOTYPE
Title Value
60%
80%
50%
40%
70%
Your Title
Proin efficitur ornare turpis, vel pulvinar risus maximus nec. Lorem ipsum dolor sit amet.

Why does communication matter?

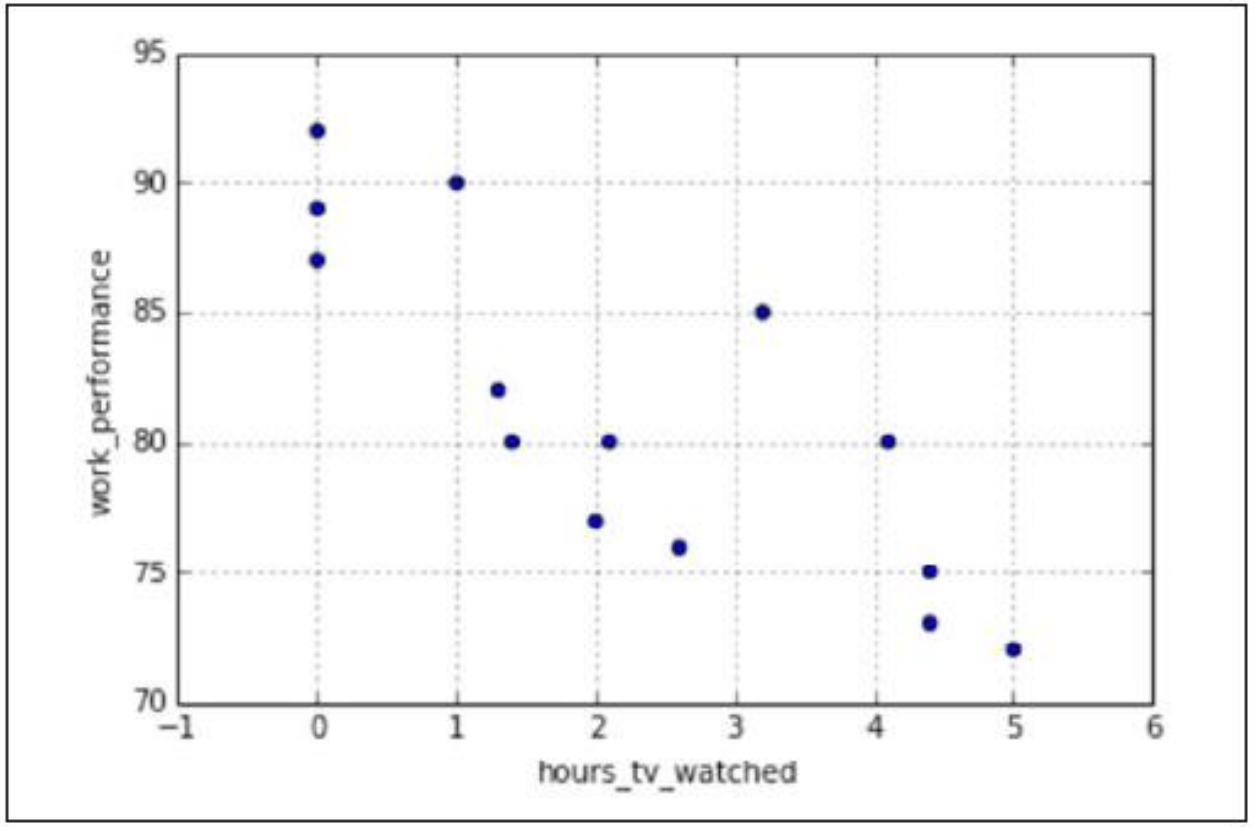
- Being able to conduct experiments and manipulate data in a coding language is not enough to conduct practical and applied data science. This is because data science is, generally, only as good as how it is used in practice.
- Communication of results is arguably as important as the results themselves.
- Generally, there are two ways of presenting results: verbal and visual.
- Of course, both the verbal and visual forms of communication can be broken down into dozens of subcategories, including slide decks, charts, journal papers, and even university lectures. However, we can find common elements of data presentation that can make anyone in the field more aware and effective in their communication skills.

Identifying effective and ineffective visualizations

- The main goal of data visualization is to have the reader quickly digest the data, including possible trends, relationships, and more.
- Ideally, a reader will not have to spend more than 5-6 seconds digesting a single visualization.
- For this reason, we must make visuals very seriously and ensure that we are making a visual as effective as possible.
- Let's look at four basic types of graphs:
  - scatter plots,
  - line graphs,
  - bar charts,
  - histograms, and
  - box plots.

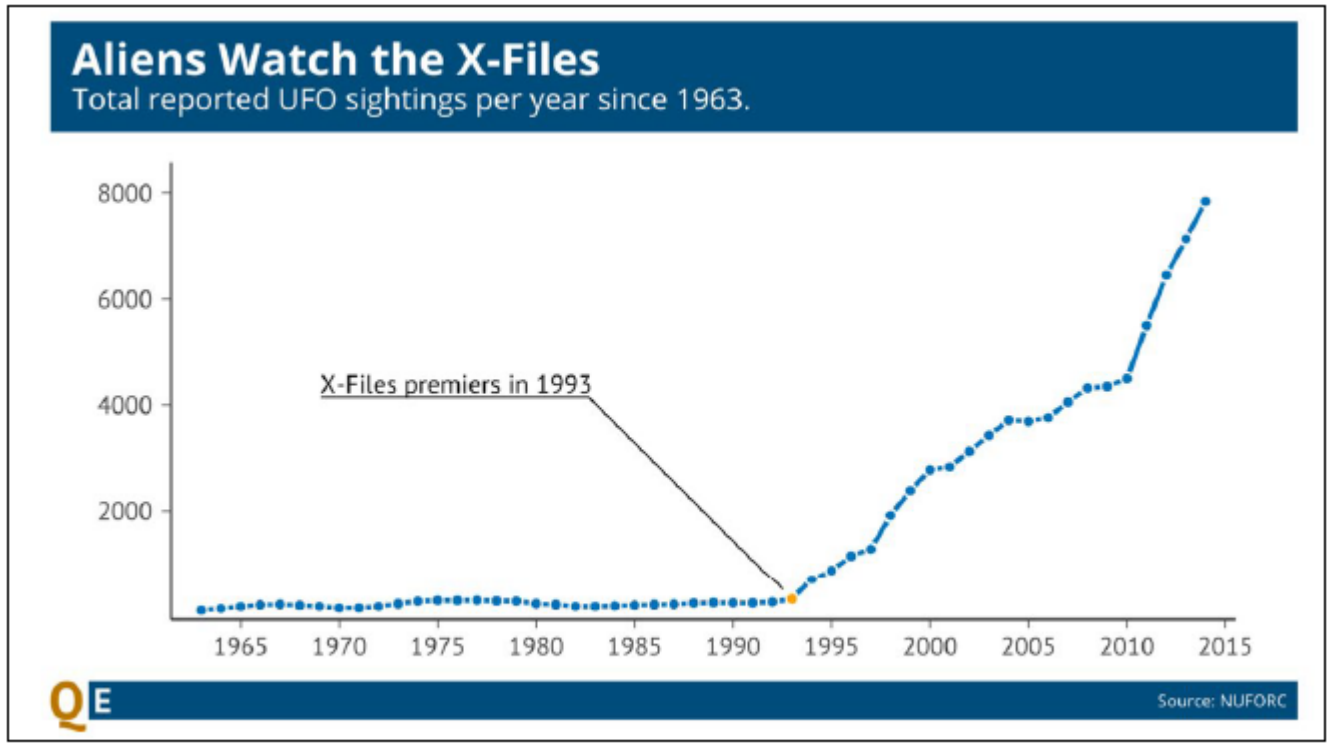
Scatter plots

- A scatter plot is probably one of the simplest graphs to create.
- It is made by creating two quantitative axes and using data points to represent observations.
- The main goal of a scatter plot is to highlight relationships between two variables and, if possible, reveal a correlation.



Line graphs

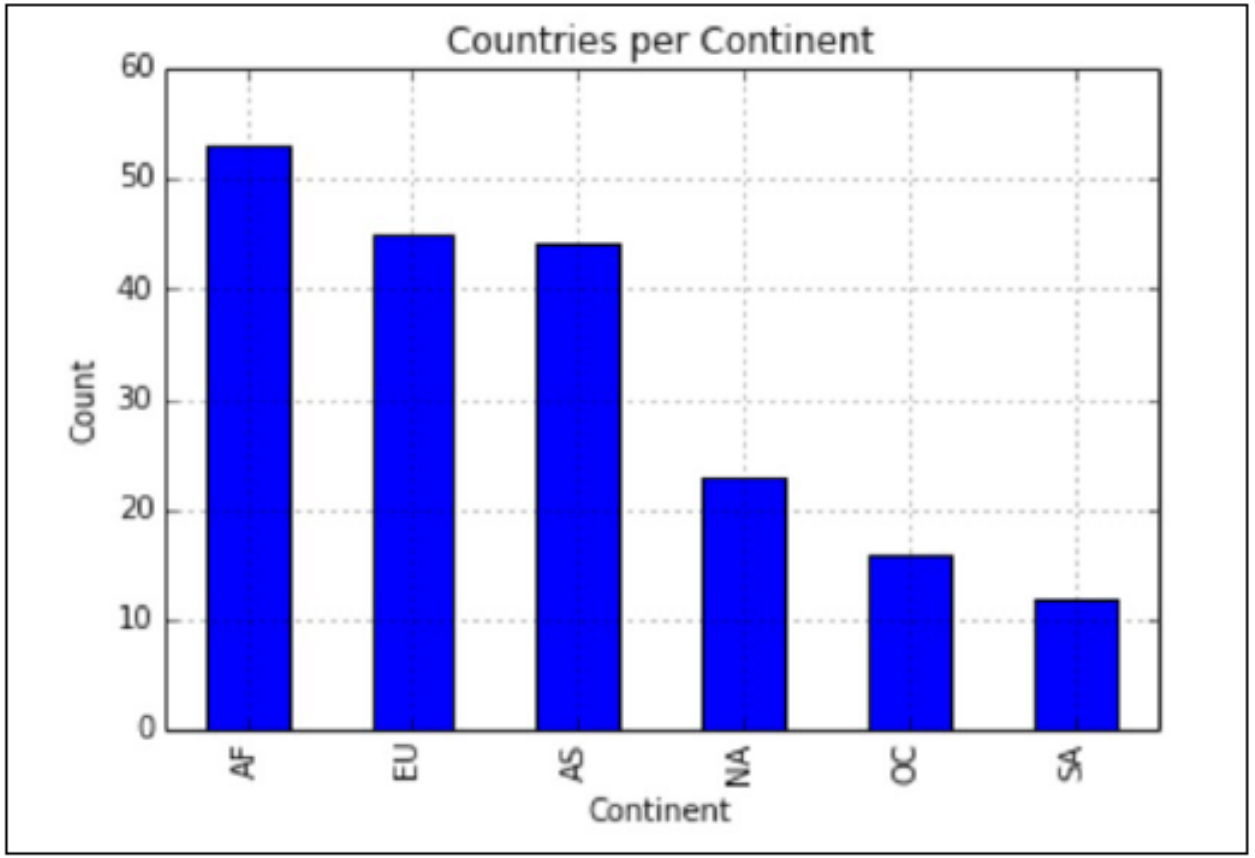
- Line graphs are, perhaps, one of the most widely used graphs in data communication.
- A line graph simply uses lines to connect data points and usually represents time on the x axis.
- Line graphs are a popular way to show changes in variables over time.
- The line graph, like the scatter plot, is used to plot quantitative variables.



Source: http://www.questionable-economics.com/what-do-we-know-about-aliens/

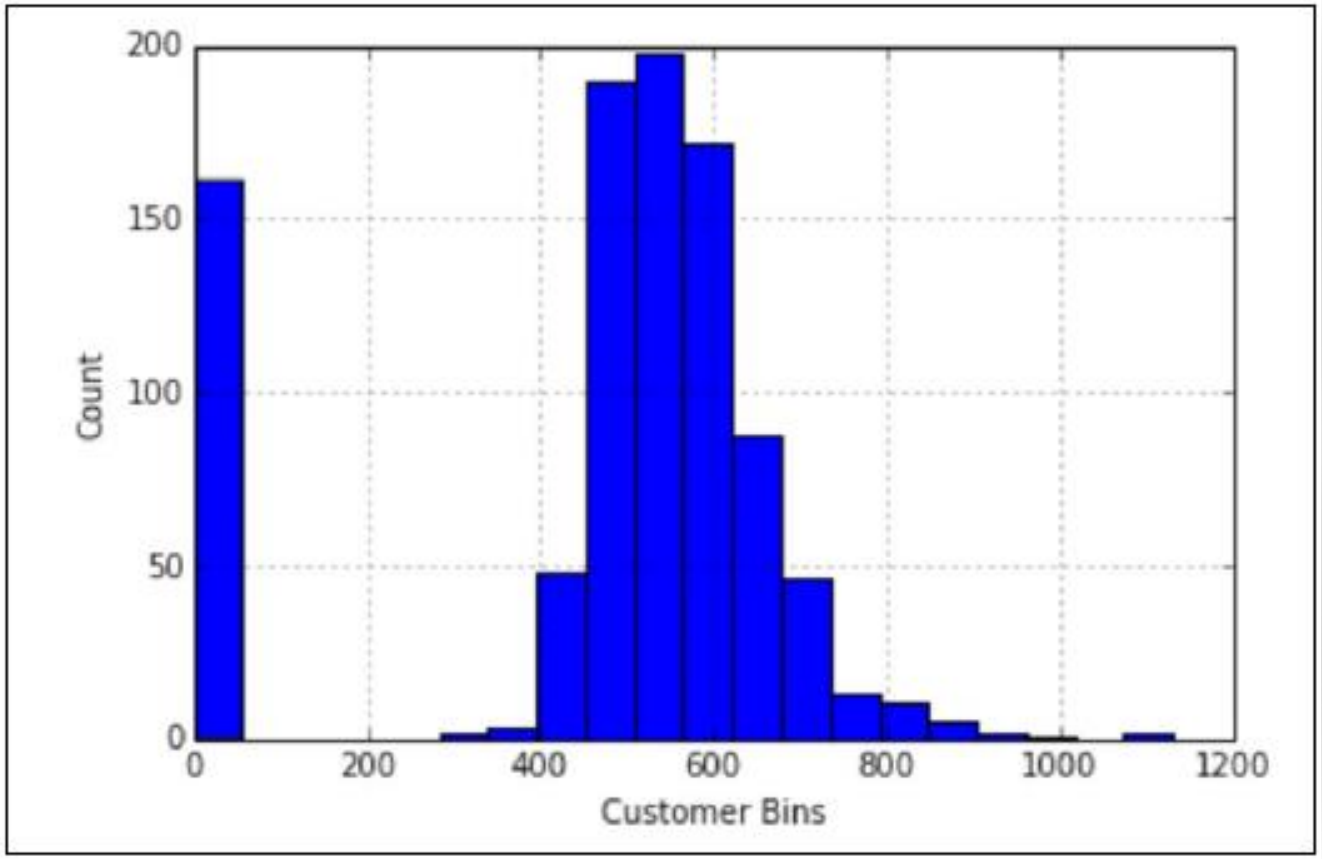
Bar charts

- We generally turn to bar charts when trying to compare variables across different groups.
- For example, we can plot the number of countries per continent using a bar chart.
- Note how the x axis does not represent a quantitative variable, in fact, when using a bar chart, the x axis is generally a categorical variable, while the y axis
- is quantitative.



Histograms

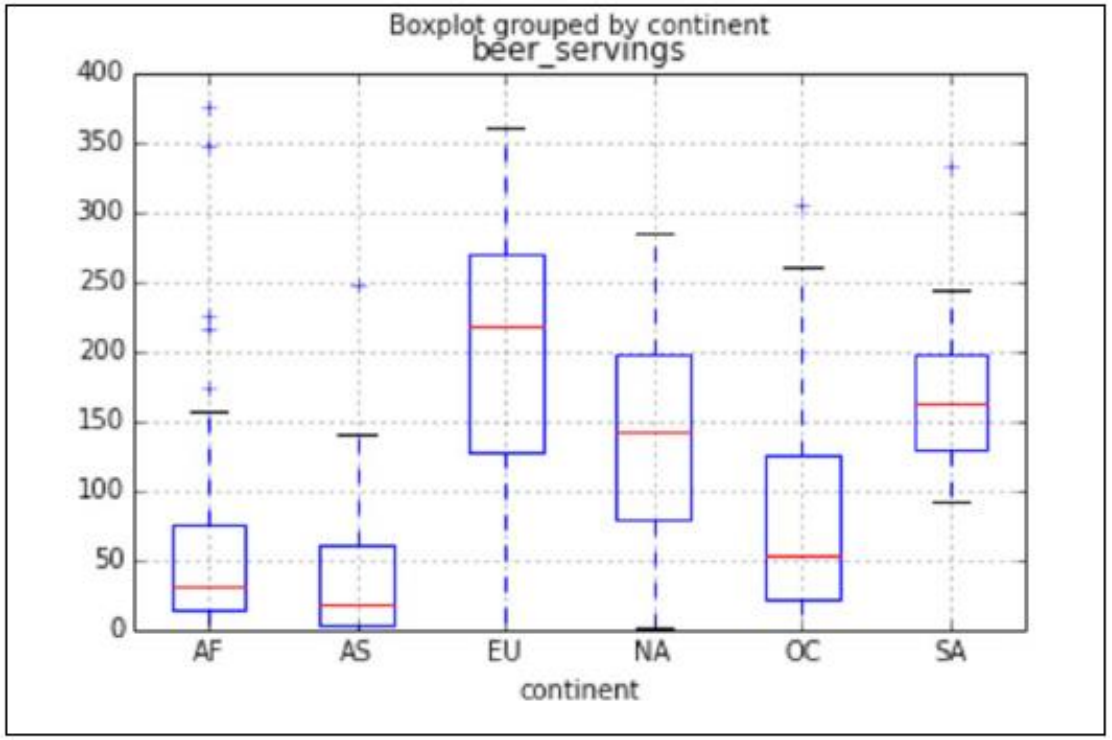
- Histograms show the frequency distribution of a single quantitative variable by splitting up the data, by range, into equidistant bins and plotting the raw count of observations in each bin.
- A histogram is effectively a bar chart where the x axis is a bin (subrange) of values and the y axis is a count.



Box plots

Box plots are also used to show a distribution of values. They are created by plotting the five number summary, as follows:

- The minimum value
- The first quartile (the number that separates the 25% lowest values from the rest)
- The median
- The third quartile (the number that separates the 25% highest values from the rest)
- The maximum value



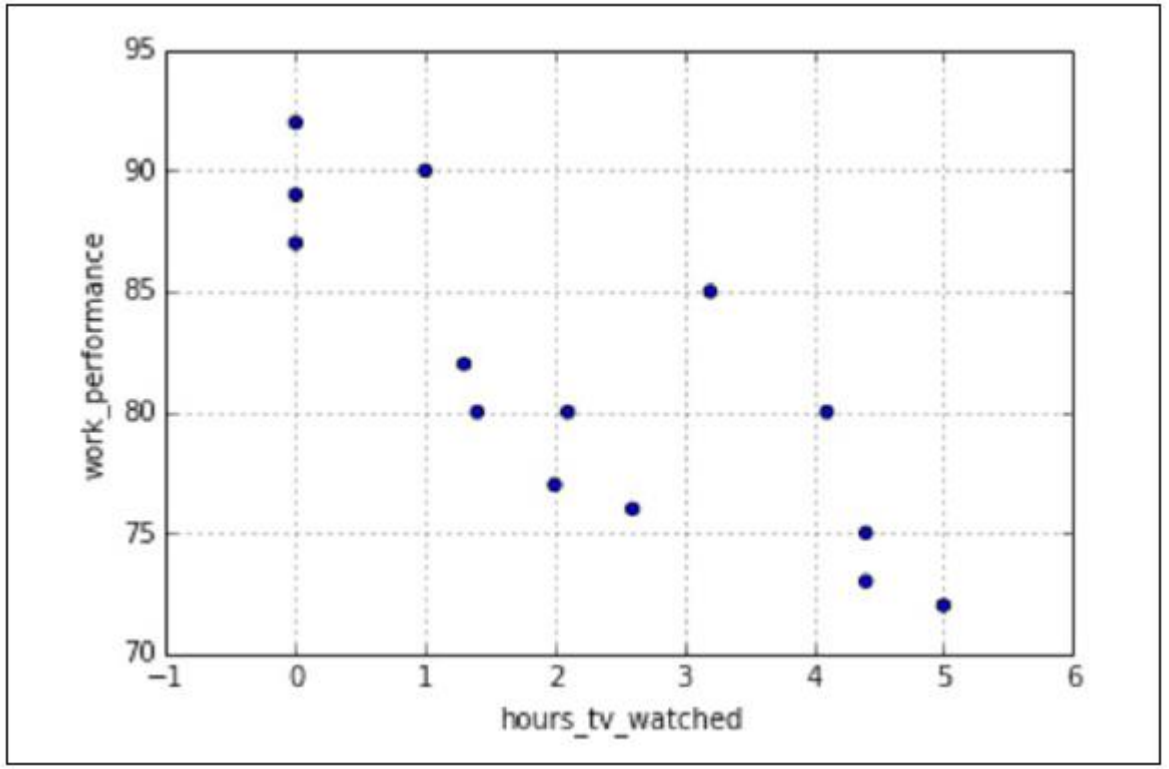
In Pandas, when we create box plots, the red line denotes the median, the top of the box (or the right if it is horizontal) is the third quartile, and the bottom (left) part of the box is the first quartile.

WHEN GRAPHS AND STATISTICS LIE

Correlation versus causation

- Correlation is a quantitative metric between -1 and 1 that measures how two variables move with each other. If two variables have a correlation close to -1, it means that as one variable increases, the other decreases, and if two variables have a correlation close to +1, it means that those variables move together in the same direction—as one increases, so does the other, and vice versa.

- Causation is the idea that one variable affects another, or leads to another.



s.zahrani@tu.edu.sa