Data visualization in exploratory data analysis

Edoardo Milotti

Advanced Statistics for Physics

Phys. Dept., University of Trieste

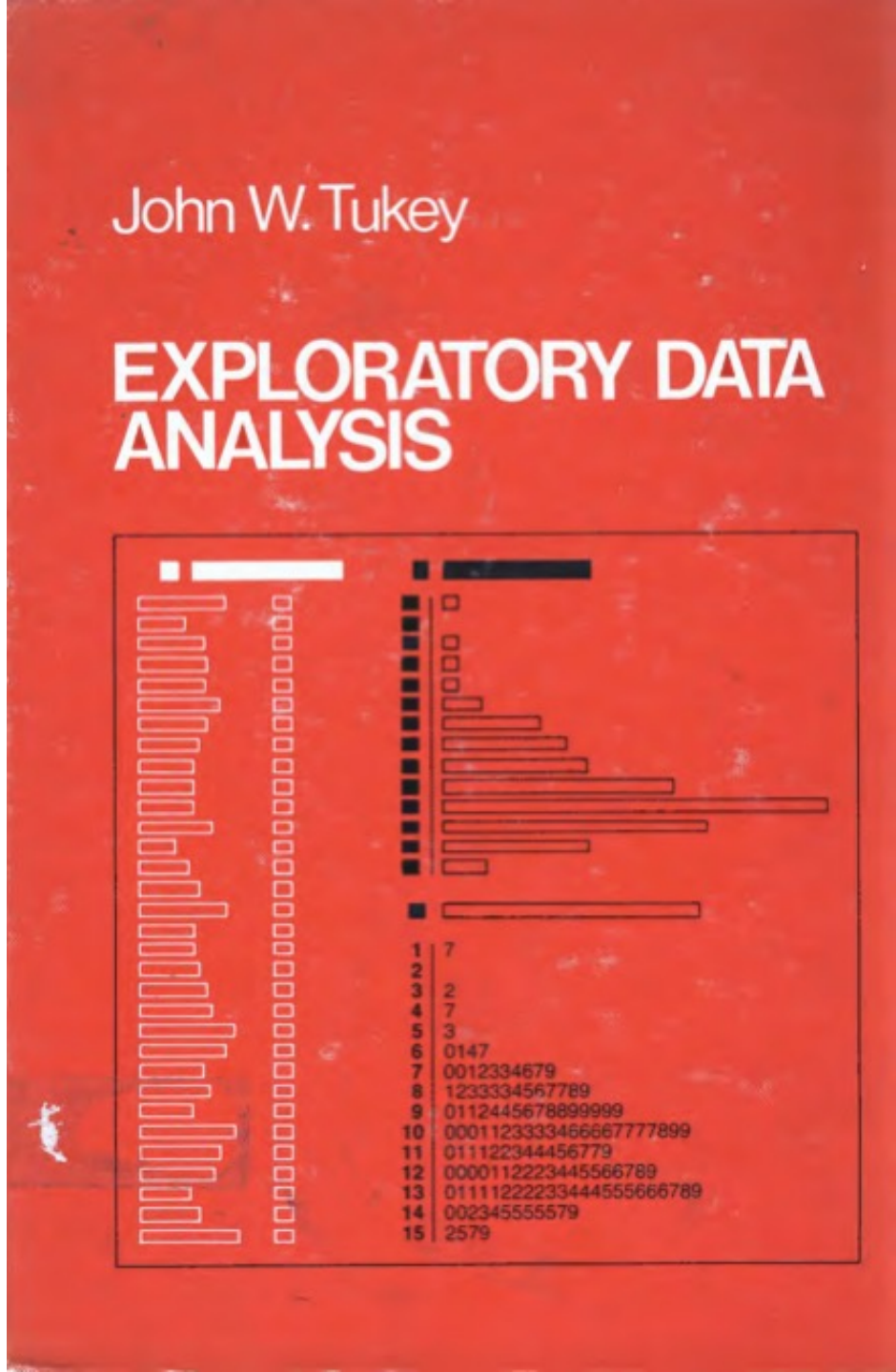
J. W. Tukey: " Exploratory Data Analysis", Addison-Wesley (1977)

This book is based on an important principle:

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

Learning first what you can do will help you to work more easily and effectively.

This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, not with confirmation.



What is Exploratory Data Analysis?

(https://www.ibm.com/topics/exploratory-data-analysis)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

Exploratory Data Analysis Tools

(https://www.ibm.com/topics/exploratory-data-analysis)

Specific statistical functions and techniques you can perform with EDA tools include:

- **Clustering and dimension reduction techniques**, which help create graphical displays of high-dimensional data containing many variables.
- **Univariate visualization** of each field in the raw dataset, with summary statistics.
- **Bivariate visualizations** and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at.
- **Multivariate visualizations**, for mapping and understanding interactions between different fields in the data.
- **K-means Clustering** is a clustering method in unsupervised learning where data points are assigned into K groups, i.e., the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- **Predictive models**, such as linear regression, use statistics and data to predict outcomes.

Florence Nightingale, the pioneer statistician

https://www.sciencemuseum.org.uk/objects-and-stories/florence-nightingale-pioneer-statistician

THE CRIMEAN WAR

The Crimean War (1853–6) was the first high-technology conflict of the modern age—an age of railways, telegraph wires, photography and high-explosive shells.

And it was a war of shocking statistics, with tens of thousands of soldiers dying.

Wrecked battery, Crimea, 1855. Photograph by James Robertson.

Plate from The Illustrated London News, 24 February 1855, showing Nightingale at work.
Science Museum Group Collection

Florence Nightingale and her nurses saw soldiers suffering from frostbite, dysentery, cholera and typhus living in ‘utterly chaotic, unsanitary and inhumane living conditions’.

‘There were no blankets, beds, furniture, food, or cooking utensils, but there were rats and fleas everywhere’, historian Eileen Magnello has recounted.

On top of that, the nurses found inadequate medical records. There was no systematic recording or reporting, hundreds of soldiers were buried without a record being made of their deaths, and a bureaucratic inertia prevented nurses and administrators from spotting obvious flaws in the system.

Florence Nightingale studied mathematics from an early age as her parents had strongly endorsed women's education.

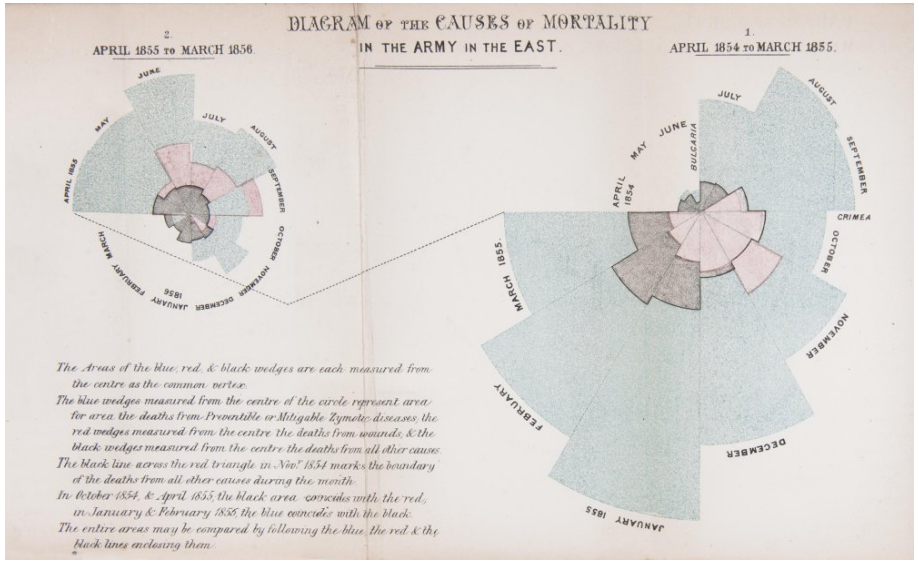
Years before she began her formal mathematical training at the age of twelve, she had developed skills in collecting, organising and presenting data.

Her devotion to mathematical practice and the study of statistics drove her throughout her subsequent career in nursing and medical reform.

Nightingale set about collecting statistics in Crimea. She treated this activity—counting the number of soldiers killed, injured or diseased—in the same way biologists collected specimens of butterflies and fossils on field trips.

Like many collectors, she employed other people to collect statistics on her behalf, running a team of data-gatherers at the Crimea.

Photograph of Florence Nightingale by the London Stereoscopic & Photographic Company Ltd.



When Nightingale returned from the Crimea to London in 1856, she set about publicising her statistical findings as well as her proposed medical reforms. But she was aware of the limited effect one person could have on practices within the armed forces and the nursing profession.

Her mission was to reach the people who could put her reforms into practice: MPs, government officials and army officers, few of whom had statistical or scientific training.

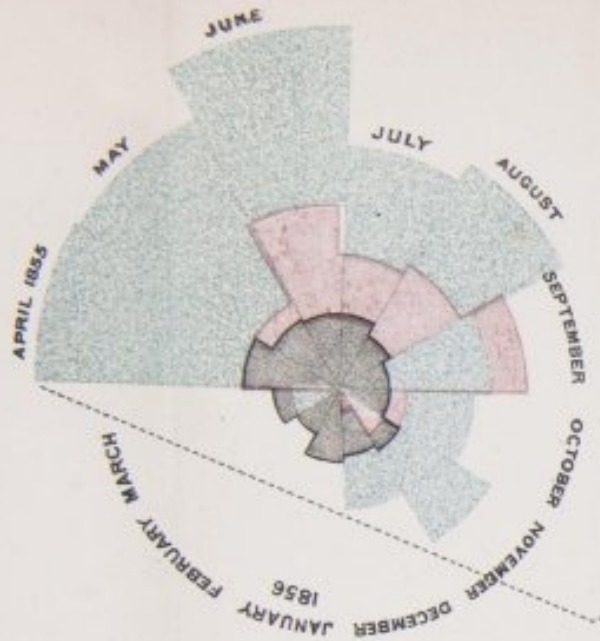
One of Nightingale's most significant innovations was a diagram which showed the causes of soldiers' deaths over two successive years in the Crimea.

- The first year (shown on the right of the diagram) was 1854–5, following her arrival in the region.
- The second (on the left) was 1855–6, after she had implemented a series of reforms to the hospital and nursing practices.
- In her diagram, each wedge represented a month, and the area of the wedge showed the number of soldiers who had died that month.
- The blue area showed deaths from preventable diseases picked up in the terrible conditions at the Crimea.
- Red sections showed deaths from battlefield wounds.
- Black areas were deaths from other causes.

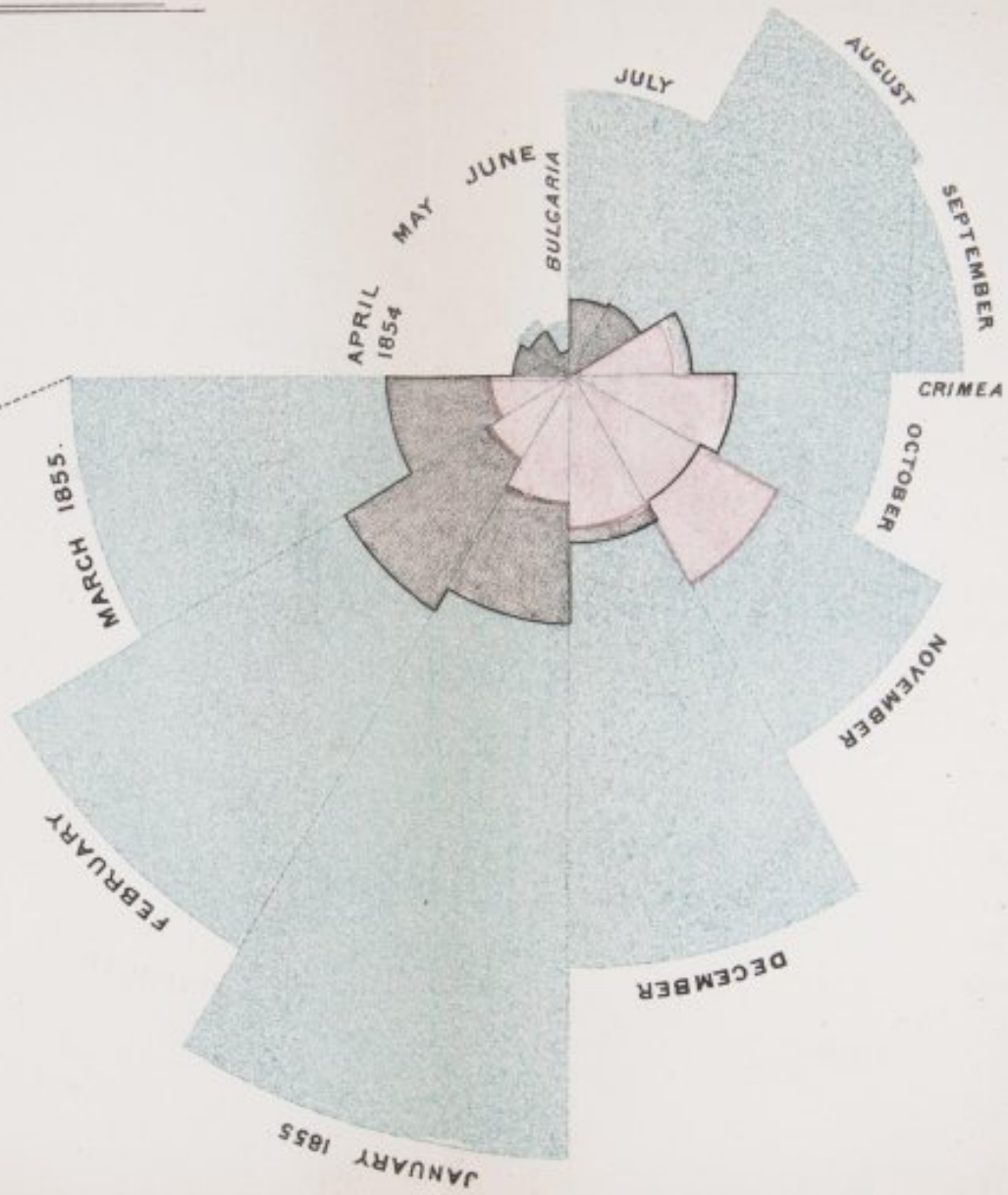
Readers could see two things. The first was that the reforms Nightingale implemented and campaigned for had made a huge positive difference to mortality. The second, and more shocking result, was that more soldiers died from preventable diseases during the war than from injuries.

DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.



1.
APRIL 1854 TO MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Novr. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red, in January & February 1855, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

By 1856, Florence Nightingale had transformed hospital care in the Crimean War—her next step was to use statistics to convince the British army and government of the need for widespread reform.

Today, we are used to seeing statistics presented in graphical form. Infographics are common in newspapers, magazines and online. However, in 1850s Britain, the approach was revolutionary.

While most statisticians provided data in tables of numbers, Nightingale was one of a small group of mathematicians who seized on the power of graphics to describe statistical findings to a non-specialist readership.

With her mortality diagram, Nightingale wanted MPs and army officials to get a quick visual understanding of the scale of the problem, counteracting their entrenched belief that soldiers died from wounds rather than unsanitary hospitals.

Lithograph showing Nightingale in discussion with an officer at the barracks hospital, Scutari, Crimea, 1856.
Science Museum Group Collection

Nightingale needed exposure to get her message out. In 1856, she began collaborating with the radical writer, journalist and sociologist Harriet Martineau.

Alongside fictional work, Martineau was noted for her early work on economics and taxation.

She was also leader-writer for the left-wing Daily News, publishing on social issues such as women's rights, education, poor-law and the abolition of slavery.

In 1859, Martineau agreed to write a popular book revealing Nightingale's findings to the public.

It included the now-famous polar-area diagram of soldier fatalities in a fold-out page at the front of the book.

Nightingale was concerned that her work would be censored by the British army for the effect it would have on troop morale. She advised Martineau that her reform proposals would need to be disguised in the book's narrative.

In spite of this, the message imparted in 'England And Her Soldiers' was hard-hitting. In her preface, Martineau wrote, 'we sustained a fearful misfortune in the last war', describing the 'mismanagement, helplessness, and doomed condition' of soldiers there.

The book was, she said, 'a grave work' and, following its publication in 1859, it was deemed unsuitable for distribution to the libraries at army barracks. Nightingale's concerns over censorship had been well founded.

Harriet Martineau’s book ‘England And Her Soldiers’, with Florence Nightingale’s pioneering statistical diagrams, was widely read outside the army rank-and-file. It undoubtedly advanced Nightingale’s cause. The two reformers worked together for many years on other projects.

In the decades that followed, statistical graphics became common as a means to share complex data with non-mathematicians.

Nightingale was elected the first female member of the Statistical Society (now the Royal Statistical Society) in 1858, just two years after returning from the Crimea.

Her influence on nursing and the presentation of statistics was profound.

https://www.youtube.com/watch?v=jbkSRLYSojo