# Lecture 3 : Exploratory Data Analysis

Sakib Anwar

AN7914 Data Analytics and Modelling

University of Winchester

2024

# Motivation

Understand the market conditions for hotels in Vienna, using prices.

- How should you start the analysis itself?
- How to describe the data and present the key features?
- How to explore the data and check whether it is clean enough for (further) analysis?

# Exploratory data analysis (EDA) - describing variables

5 reason to do EDA!

1. To check data cleaning (part of iterative process)
2. To guide subsequent analysis (for further analysis)
3. To give context of the results of subsequent analysis (for interpretation)
4. To ask additional questions (for specifying the (research) question)
5. Offer simple, but possibly important answers to questions.

# Key tasks: describe variables

Look at key variables

- what values they can take and
- how often they take each of those values.
- are there extreme values

Describe what you see

- Descriptive statistics - key features summarized

$\Downarrow$

- to understand variables you work with
- to make comparisons

# Frequency of values

- The *frequency* or more precisely, *absolute frequency* or *count*, of a value of a variable is simply the number of observations with that particular value.
- The *relative frequency* is the frequency expressed in relative, or percentage, terms: the *proportion* of observations with that particular value among all observations.
- *Practical Note on Missing Values:* When calculating proportions, decide whether to include all observations or just those with non-missing values. The latter is commonly preferred.

# The distribution and the histogram

A key part of EDA is to look at the (empirical) distribution of the most important variables.

- All variables have a *distribution*.
- The distribution of a variable gives the frequency of each value of the variable in the data.
- May be expressed in terms of absolute frequencies (number of observations) or relative frequencies (percent of observations).
- The distribution of a variable completely describes the variable as it occurs in the data.
- Each variable's distribution is analyzed *independently*, not influenced by other variables' distributions.

# Histograms

Histogram reveals important properties of a distribution.

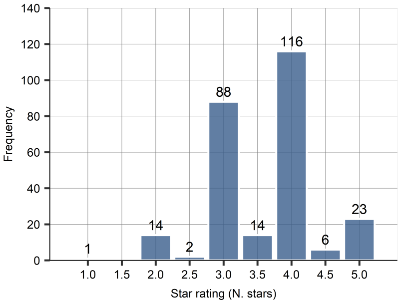
- Number and location of *modes*: these are the peaks in the distribution that stand out from their immediate neighborhood.
- Approximate regions for *center* and *tails*
- *Symmetric* or not - asymmetric distributions have a long left tail or a long right tail
- *Extreme values*: values that are very different from the rest. Extreme values are at the far end of the tails of histograms.

# Extreme values

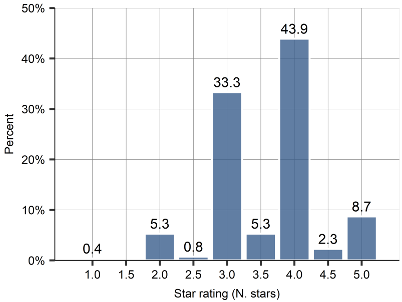
- Some variables have extreme values: substantially larger or smaller values for one or a handful of observations than the values for the rest of the observations.
- Need conscious decision.
  - Is this an error? (drop or replace)
  - Is this not an error but not part of what we want to talk about? (drop)
  - Is this an integral feature of the data? (keep)

# Hotel price histograms

(a) Absolute frequency (count)



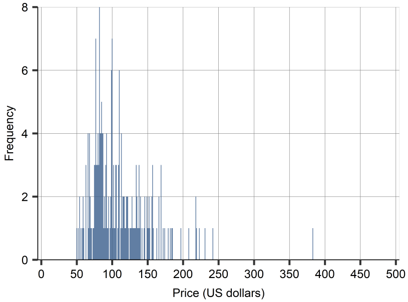

(b) Relative frequency (percent)



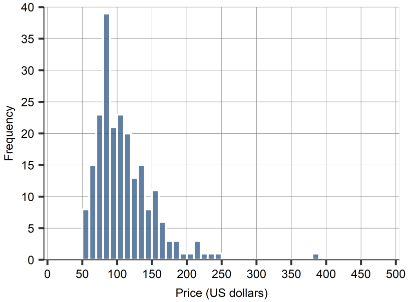

Source: `hotels-vienna` dataset. Vienna, Hotels only, for a 2017 November weekday

# Hotel price histograms

(a) Histogram: individual values



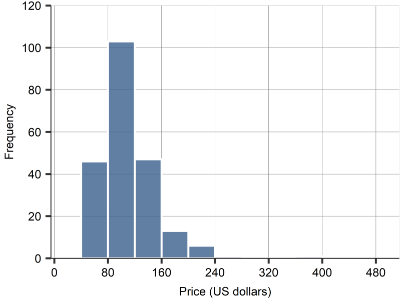

(b) Histogram: 20$ bins



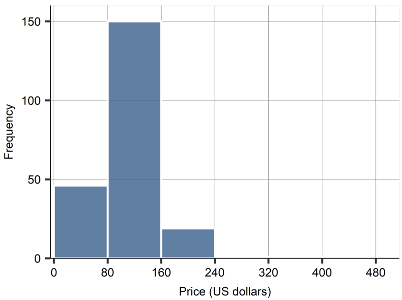

Note: *Panel (a) just shows individual values - help see where most values are. Panel (b) is a histogram with 20$ bins - more useful to capture frequencies.* Source: `hotels-vienna` dataset. Vienna, 3-4 stars hotels only, for a 2017 November weekday

# Hotel price histograms

(a) Histogram: 40$ bins




(b) Histogram: 80$ bins

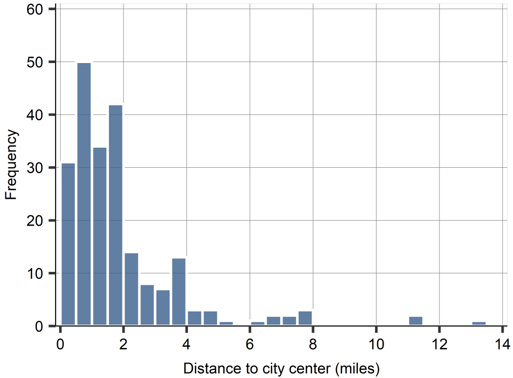



Note: *Bin size matters. Wider bins suggest a more gradual decline in frequency.*

# Hotel density plot

- Vienna all hotels, 3-4 stars
- Use absolute frequency (count)
- For this histogram we use 0.5-mile-wide bins. This way we can see the extreme values in more detail
- Dropped very far - likely not Vienna

Figure: Histogram of distance to the city center.

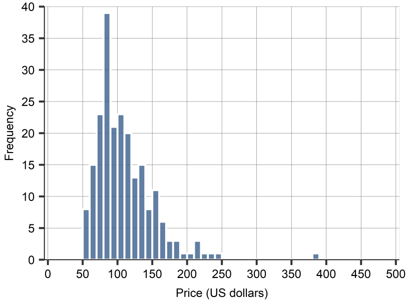


# Hotel prices

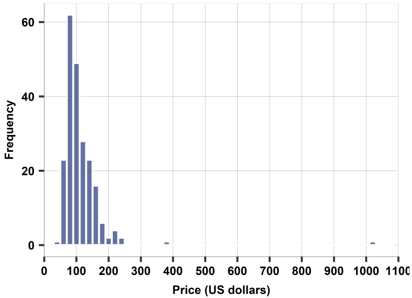
- Vienna all hotels, 3-4 stars
- Use absolute frequency (count)
- We go back to prices
- How to decide what to include? -> check observation!

# Hotel price histograms

(a) Histogram: 20$ bins as seen




(b) Histogram: including extreme value above 1000$




Source: `hotels-vienna` dataset. Vienna, 3-4 stars hotels only, for a 2017 November weekday

# Summary statistics

- For any given variable, a *statistic* is a meaningful number that we can compute from a dataset.
- Basic *summary statistics* describe the most important features of distributions of variables.

- Many of you know this. I briefly cover it

# Summary statistics: Sample mean

The most used statistic is the *mean*:

$$\bar{x} = \frac{\sum x_i}{n} \tag{1}$$

where $x_i$ is the value of variable $x$ for observation $i$ in the dataset that has $n$ observations in total. Two key features

$$\overline{x + a} = \overline{x} + a \tag{2}$$

$$\overline{x \cdot b} = \overline{x} \cdot b \tag{3}$$

# The Expected value

- The expected value is the value that one can expect for a randomly chosen observation
- The notation for the expected value is $E[x]$.
- For a quantitative variable, the expected value is the mean
- For a qualitative variable, it can only be determined if transformed to a number

# Summary statistics: The median and other quantiles

- *Quantiles*: a quantile is a value that divides the observations in a dataset into two or more groups in specific proportions.
- The *median* is the middle value of the distribution - half the observations have lower value and the other half have higher value.
- *Percentiles* are values that divide a dataset into 100 equal parts, each representing a percentage of data points that fall below it
- *Quartiles* are a type of quantiles which divide the number of data points into four parts, or quarters, of more-or-less equal size.

# Summary statistics: The mode

- The *mode* is the value with the highest frequency in the data.
- Some distributions are unimodal, others have multiple modes.
- Multiple modes can be observed, each distinct within its own ‘neighborhood’ of values, yet they may exhibit differing frequencies

# Summary statistics: central tendency

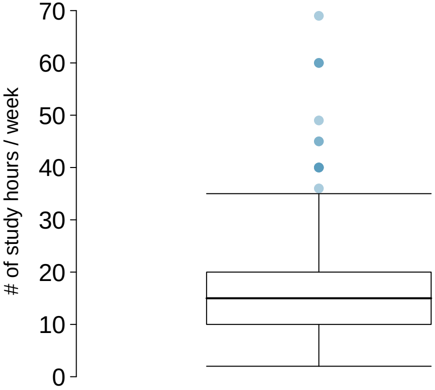
- The mean, median and mode are different statistics for the *central value* of the distribution
- Central tendency.
  - The mode is the most frequent value
  - The median is the middle value
  - The mean is the value that one can expect for a randomly chosen observation.

# Summary Statistics: Spread of Distributions

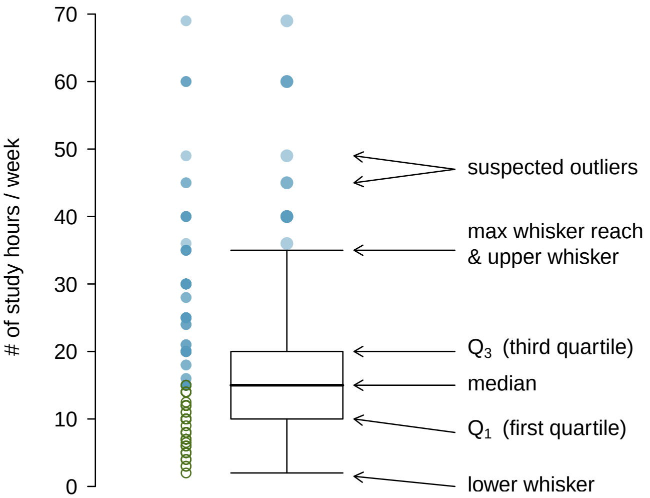
- Analyzing the *spread of distributions* is crucial for understanding data variability.
- Key statistics for measuring distribution spread include:
  - *Range:* The difference between the maximum and minimum values in a dataset.
  - *Inter-Quartile Range (IQR):* The difference between the third quartile (75th percentile) and the first quartile (25th percentile), showing the spread of the middle 50% of the data.
  - *90-10 Percentile Range:* The difference between the 90th and 10th percentiles, highlighting the spread of the central 80% of the data.
  - *Standard Deviation and Variance:* Measures that quantify the dispersion of data points around the mean.
- These statistics provide insights into the distribution's overall variability and identify potential outliers.

# Box Plot

The box in a box plot represents the middle 50% of the data, and the thick line in the box is the median.



# Anatomy of a box plot



# Whiskers and Outliers

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.
- The maximum upper whisker reach is calculated as $Q3 + 1.5 \times IQR$, and the maximum lower whisker reach is $Q1 - 1.5 \times IQR$.
- For example, if the IQR (Interquartile Range) is $20 - 10 = 10$, then:
  - The maximum upper whisker reach would be $20 + 1.5 \times 10 = 35$.
  - The maximum lower whisker reach would be $10 - 1.5 \times 10 = -5$.
- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

# Summary statistics: standard deviation

- The most widely used measure of spread is the *standard deviation*. Its square is the *variance*.
- Variance is the average squared difference of each observed value from the mean.
- The standard deviation captures the typical difference between a randomly chosen observation and the mean.
- The variance is a less intuitive measure. At the same time, the variance is easier to work with, because it is a mean value itself.

$$Var[x] = \frac{\sum(x_i - \overline{x})^2}{n} \tag{4}$$

$$Std[x] = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \tag{5}$$

# Summary statistics: skewness

- A distribution is *skewed* if it isn't symmetric.
- It may be skewed in two ways, having *a long left tail* or having *a long right tail*.
- Example: hotel price distributions having a long right tail - such as in hotel price distribution.
- Skewness and the prevalence of extreme values are related. With distributions with long tails, values far away from all other values are more likely.
- When extreme values are important for the analysis, skewness of distributions is important, too.

# Summary statistics: skewness measure

Simplest measure is *mean–median measure of skewness*.

$$Skewness = \frac{(\bar{x} - median(x))}{Std[x]} \tag{6}$$

- When the distribution is symmetric its mean and median are the same.
- When it is skewed with a long right tail the mean is larger than the median: the few very large values in the right tail tilt the mean further to the right.
- When a distribution is skewed with a long left tail the mean is smaller than the median
- To make this measure comparable across various distributions use a standardized measure
- If multiplied by 3, and then it's called *Pearson's second measure of skewness*.

# Vienna vs London

- Compare two cities, how hotel markets vary
- Vienna, London
- 3-4 star hotels, only "Hotels" (no apartments), below 1000 dollars.
- Focus on actual city=Vienna and actual city=London (exclude nearby related villages).
- Use `hotels-europe` dataset.
- N=207 for Vienna, N=435 for London

# London vs Vienna

Figure: Vienna Austria

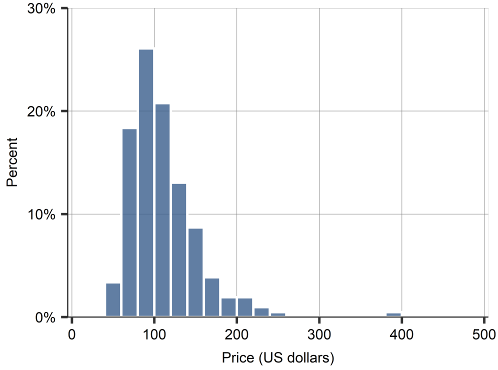
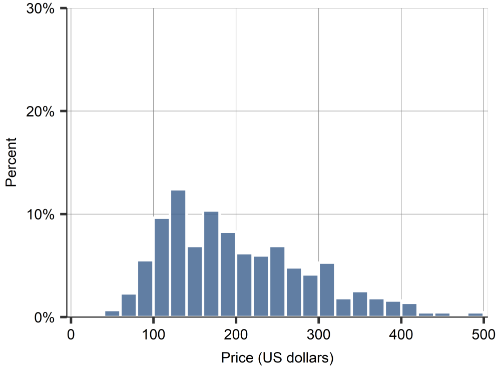



Figure: London, UK

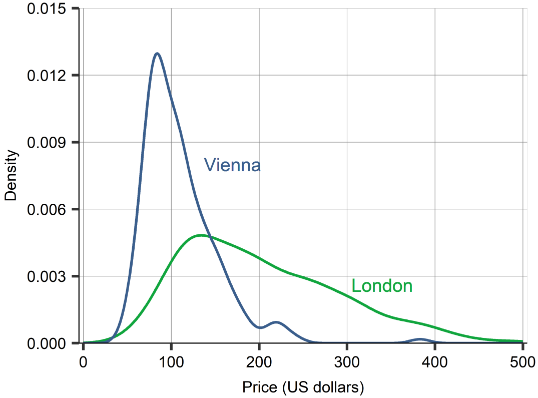


# Density plots

- *Density plots* - also called *kernel density estimates*
- Alternative to histograms - instead of bars density plots show continuous curves.
- We may think of them as curves that wrap around the corresponding histograms.
- Density plots complementing histograms - some believe density plots allow for easier comparison of distributions across groups in the data.

# The density plot

- Density plot
- Less reliable than histogram
- But easy to read the key points
- Easy for comparison



# Case study hotels: descriptive statistics

Table: Descriptive statistics for hotel prices in two cities.

| City | N | Mean | Median | Min | Max | Std | Skew |
|---|---|---|---|---|---|---|---|
| London | 435 | 202.36 | 186 | 49 | 491 | 88.13 | 0.186 |
| Vienna | 207 | 109.98 | 100 | 50 | 383 | 42.22 | 0.236 |

Source: `hotels-europe` dataset. Vienna and London, weekday, November 2017

# Theoretical distributions

Theoretical distributions are distributions of variables with idealized properties.

- Show frequencies for theoretical distributions and not for empirical ones.
- The likelihood of each value in a more abstract setting - hypothetical "dataset" or "population," or the abstract space of the possible realizations of events.
- Theoretical distributions are fully captured by few *parameters*: these are statistics determine the whole distributions

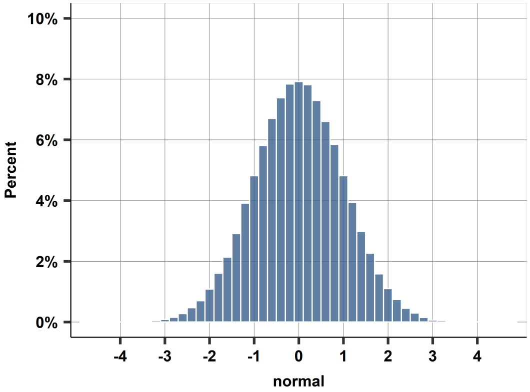
# Theoretical Distributions

Theoretical distributions provide valuable insights into data analysis due to their well-defined properties.

- *Well-Known Properties:* Theoretical distributions have specific, well-understood characteristics.
- *Approximation of Real Data:* When a variable in our dataset closely matches a theoretical distribution, we can infer the distribution's properties for our variable. This helps in understanding the behaviour and underlying patterns of the data.
- *Common in Real Life:* Many real-life variables align closely with theoretical distributions, often more than initially expected.
- *Utility in Generalization:* Recognizing these patterns is crucial for making broader inferences from our data, as will be discussed in more detail in Chapter 5/Lecture 6 (to be confirmed).

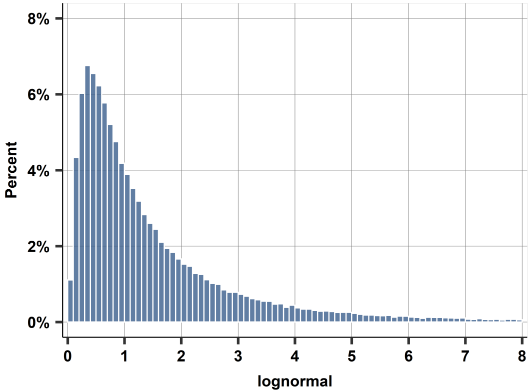
# The Normal distribution

- Histogram is bell-shaped
- Outcome (event), can take any value
- Distribution is captured by two parameters
  - $\mu$ is the mean
  - $\sigma$ the standard deviation
- Symmetric = median, mean (and mode) are the same.
- Example: height of people, IQs, ect.



# The log-normal distribution

- Asymmetrically distributed with long right tails.
- start from a normally distributed RV $(x)$, transform it: $(e^x)$ and the resulting variable is distributed log-normal.
- Always non-negative
- Example distributions of income, or firm size.



# Insights on Normal and Log-Normal Distributions

- *Prevalence of Normal Distribution:* A significant number of variables in real life approximate a normal distribution, especially those that result from the sum of many elementary factors.
- *Limitations of Normal Distribution:* However, the normal distribution might not be a good fit in situations where:
  - There is asymmetry in the data.
  - Extreme values play a crucial role and cannot be ignored.
- *Log-Normal Distribution:* Variables resulting from multiplicative processes often fit a log-normal distribution, as the log transformation turns multiplication into addition, aligning more with a normal distribution.

# Income and log-income

Figure: income

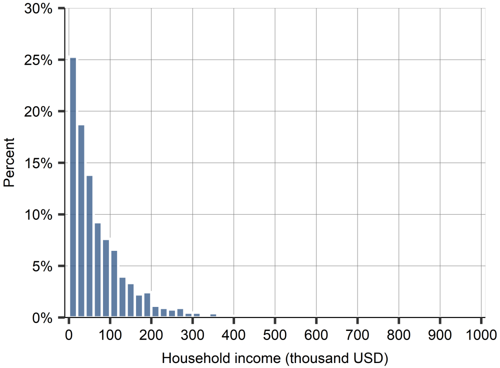
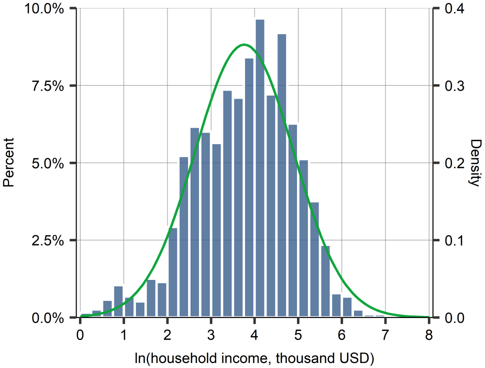
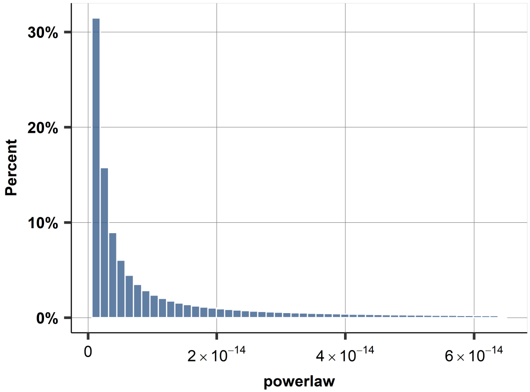



Figure: log income



# The power law distribution

- Also called as Pareto distribution
- Very large extreme values - well approximated
- Relative frequency of close-by values are the same along large and small values
- Real world: many examples, but often not the whole distribution
- Example: frequency of words, city population, wealth



# Data visualization

- We shall make conscious decisions and not let default settings guide us.
- We break our preferences into four layers:
  1. usage - what you want to show and to whom – deciding on purpose, focus, and audience
  2. encoding - key elements of the graphs showing how you want to present the message
  3. scaffolding - how you want to present elements that support understanding such as axis and legends
  4. annotation - adding anything else you want to emphasize

This is a very brief overview, more in Chapter 03 of Gabor Data Analysis Book

# Data visualization: usage

- What message do you want to convey and to whom?
- As a general principle, one graph should convey one message.
- Be explicit about the purpose of the graph and the target audience: general audience vs specialist
- For a specialist audience, more complicated graphs are okay.

# Data visualization: encoding

- How to show what you want to show.
- Picking the type of graph, additional features to it, and colors.
- Type of graph: bar chart, scatterplot, etc.
- How to denote information (dots, lines, bars)
- Have an additional encoding to help comparison: bar chart for separate groups encoded with colors.
- One information, one encoding - use size or color but not both.

# Data visualization: scaffolding

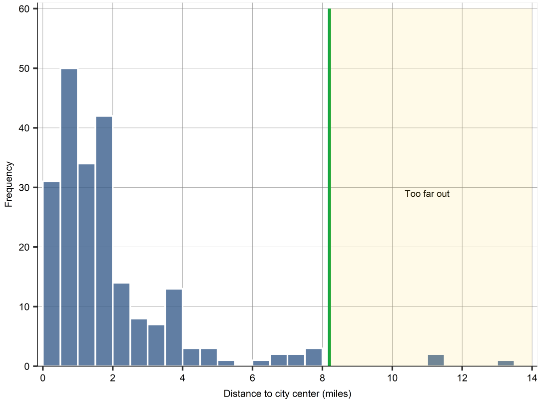
- How to present elements that support understanding.
- Make sure that a graph has
  - Title
  - Axis title and labels
  - Legend
- Content as well as format, such as font type and size.

# Data visualization: annotations

- If there is something else we want to emphasize?
- Additional information can help put the graph into context or emphasize some part of it
- Colored area, circled observations, arrow+text, etc

# Data visualization: example

- Usage: to show distribution for general audience
- Encoding is bars (histogram), bin size set at 20
- Axes labelled with title + grid
- annotation: far away hotels



# Summary

- Always check your key variables
- Look at summary statistics, understand key features such as central tendency, spread and skewness
- Look at histograms to get a broader picture of the distribution, see if multiple modes or extreme values.
- EDA helps describe the data, and plan the analysis
- Data vizualization matters, makes sense to do it carefully.