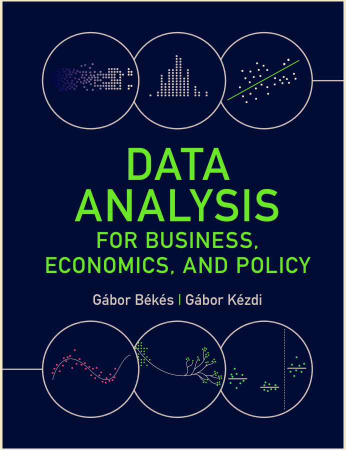
Békés-Kézdi: Data Analysis, Chapter 03: Exploratory Analysis



Data Analysis for Business, Economics, and Policy

Gábor Békés (Central European University)
Gábor Kézdi (University of Michigan)

Cambridge University Press, 2021

gabors-data-analysis.com

Central European University

Version: v3.1 License: CC BY-NC 4.0

Any comments or suggestions:
gabors.da.contact@gmail.com

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

Variable description, histograms

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: With missing values – proportion can be relative to all observations OR only observations with non-missing values (usual choice).

Probabilities and frequencies

- *Probability* is general a concept that is related to relative frequency.
- Probability is a measure of the likelihood of an *event*.
- An event is something that may or may not happen.
- Probabilities are always between zero and one.
- Probability as a generalization of relative frequencies in datasets.
- Probabilities are more general than relative frequencies as they can describe events without datasets.

The distribution and the histogram

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

- All variables have a *distribution*.
- The distribution of a variable tells 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.
- independent from values the other variables may show.

Histograms

Histogram reveals important properties of a distribution.

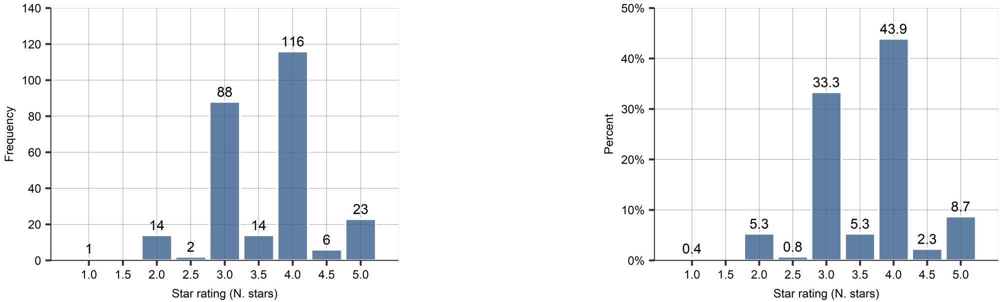
- Number and location of *modes*: these are the peaks in the distribution that stand out from their immediate neighborhood.
- Approximate regions for *center* and *tails*
- *Symmetric* or not - asymmetric distributions have a long left tail or a long right tail
- *Extreme values*: values that are very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

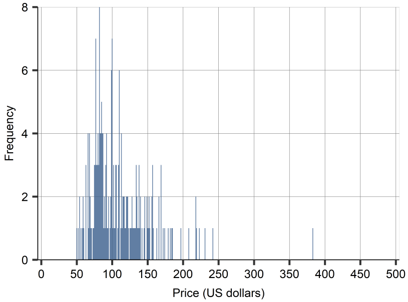
(a) Absolute frequency (count) (b) Relative frequency (percent)



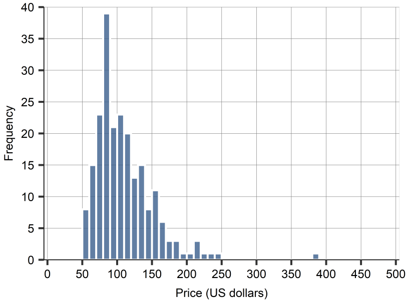
Source: `hotels-vienna` dataset. Vienna, Hotels only, for a 2017 November weekday

Hotel price histograms

(a) Histogram: individual values



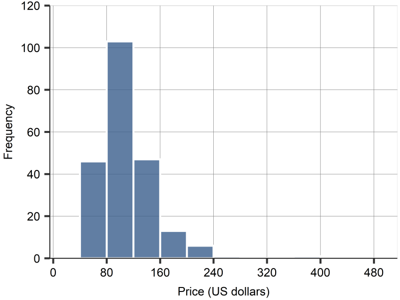
(b) Histogram: 10$ bins



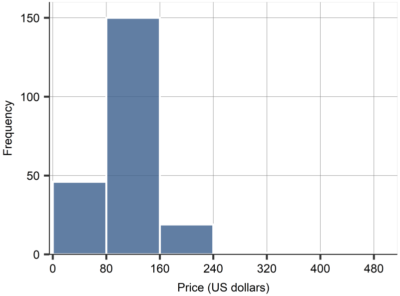
Note: *Panel (a) just shows individual values - help see where most values are. Panel (b) is a histogram with 10$ bins - more useful to capture frequencies.* Source: `hotels-vienna` dataset. Vienna, 3-4 stars hotels only, for a 2017 November weekday

Hotel price histograms

(a) Histogram: 40$ bins



(b) Histogram: 80$ bins

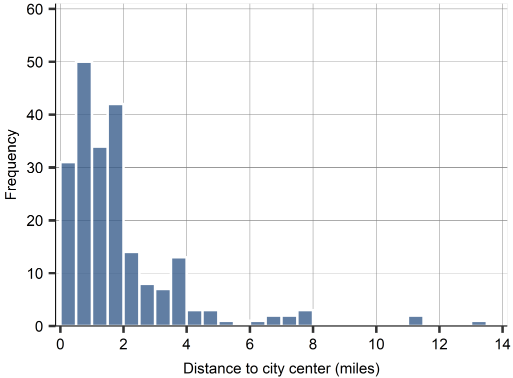


Note: *Bin size matters. Wider bins suggest a more gradual decline in frequency.*

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.

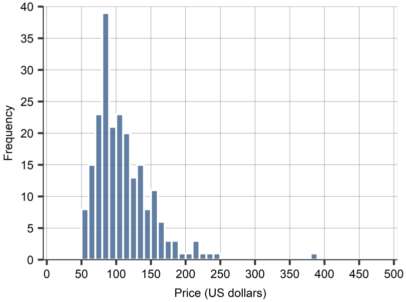


Hotel prices

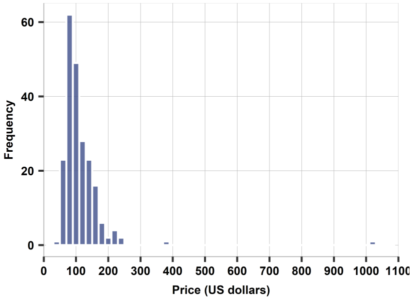
- Vienna all hotels, 3-4 stars
- Use absolute frequency (count)
- We go back to prices
- How to decide what to include? -> check observation!

Hotel price histograms

(a) Histogram: 10$ 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

EDA and cleaning - Vienna hotels

1. Start with full data N=428
2. Tabulate key qualitative variables
3. accommodation type - could be apartment, etc. Focus on hotels. N=264
4. stars - focus on 3, 3.5 4 stars, as lower bit not well covered, luxury could vary a lot. N=218
5. Look at quantitative variables, focus on extreme values.
6. Start with price. p=1012 likely error drop. keep others N= 217
7. Distance: some hotels are far away. define cutoff. drop beyond 8km N=214
8. check why hotels could be far away. Find variable city_actual. Tabulate. Realise few hotels are not in Vienna. Drop them. N=207
9. So, the final cut: Hotels, 3 to 4 stars, below 1000 euros, less than 8km from center, in Vienna actual N=207.

Summary statistics

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
  - Male/Female binary variable. Expected value could be probability / relative frequency of females.
  - Quality of hotel: 1 to 5 stars, mean can be calculated, but its meaning is less straightforward.
    - What is the assumption for getting the mean as number?

Summary statistics: The median and other quantiles

- *quantiles*: a quantile is the value that divides the observations in the dataset to two parts 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* divide the data into two parts along a certain percentage.
  - The first percentile is the value below which one percent of the observations are and 99 percent above.
- *Quartiles* divide the data into two parts along fourths.
  - 1st quartile has one quarter of the observations below and three quarters above; it is the 25^{th} percentile.
  - 2nd quartile has two quarters of the observations below and two quarters above; this is the median, and also the 50^{th} percentile.

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 are apart from each other, each standing out in its "neighborhood", but they may have different 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

- *spread of distributions* is also often used in analysis.
- Statistics that measure the spread of distributions are the range, inter-quantile ranges, the standard deviation and the variance.
- The *range* is the difference between the highest value (the maximum) and the lowest value (the minimum) of a variable.
- The *inter-quantile ranges* is the difference between two quantiles- the third quartile (the 75^{th} percentile) and the first quartile (the 25^{th} percentile).
- The 90- 10 percentile range gives the difference between the 90^{th} percentile and the 10^{th} percentile.

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.

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

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

Summary statistics: standard deviation

- 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.
- The standard deviation (SD) captures the typical difference between a randomly chosen observation and the mean.
  - Not exactly the average but similar
  - Same unit of measure (ie dollars)
- Two distributions with same mean. If SD is higher - more dispersed the data
- In Finance, SD and variance are measures of price volatility of an asset.
  - High volatility: price jumps up and down.

$$Std[x] = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}} \tag{6}$$

Using the standard deviation to define standardized values

The standard deviation is often used to re-calculate differences between values in order to express those in terms of typical distance.

$$x_{standardized} = \frac{(x - \bar{x})}{Std[x]} \tag{7}$$

- *standardized value of a variable* shows the difference from the mean in units of standard deviation.
- For example: a standardized value of one shows a value is one standard deviation larger than the mean; a standardized value of negative one shows a value is one standard deviation smaller than the mean

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*.

$$Skewness = \frac{(\bar{x} - median(x))}{Std[x]} \tag{8}$$

- 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*.

Visualizing summary statistics

- Measures of central value: Mean (average), median, other quantiles (percentiles), mode.
- Measures of spread: Range, inter-quantile range, variance, standard deviation.
- Measure of skewness: The mean–median difference.

- The box plot is a visual representation of many quantiles and extreme values.
- The violin plot mixes elements of a box plot and a density plot.

Visualizing summary statistics

Figure: Boxplot

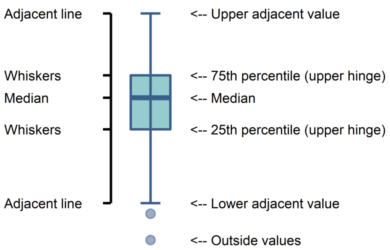
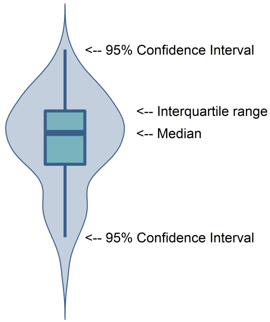


Figure: Violinplot



Density plots

- *Density plots* - also called *kernel density estimates*
- alternative to histograms - instead of bars density plots show continuous curves.
- 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.

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

- Graphical vs comparison table

London vs Vienna

Figure: Vienna Austria

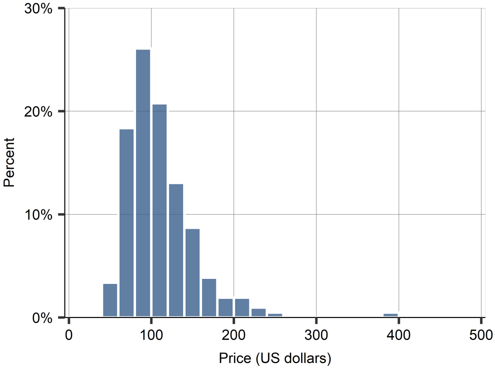
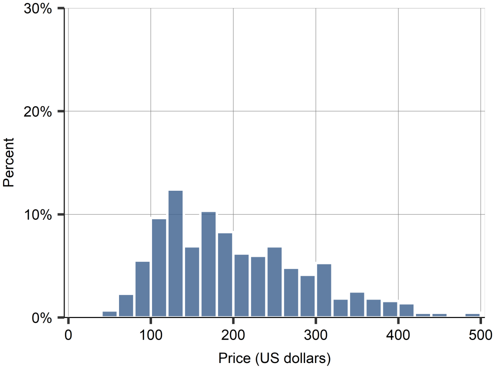
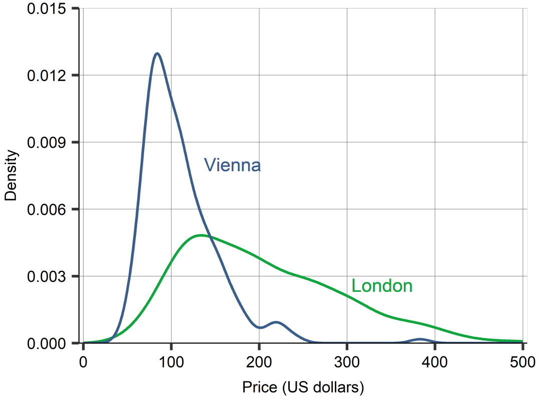


Figure: London, UK



The density plot

- Density plot
- Less reliable than histogram
- But key points good be read off
- Easy when 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

Vienna vs London

- Compare two cities, how hotel markets vary
- Graphical vs comparison table - Advantage / disadvantage?
- Both help define key messages: (1) describe and (2) explain/make sense.
  - Hotel prices in London tend to be substantially higher on average.
  - London prices are also more spread, with a minimum close to the Vienna minimum, but many hotels above 200 dollars
  - These together imply that there are many hotels in London with a price comparable to hotel prices in Vienna, but there are also many hotels with substantially higher prices

Distributions

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

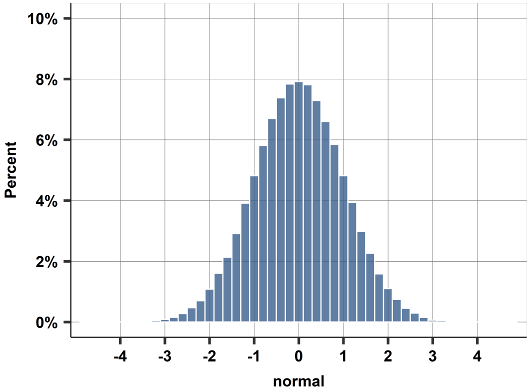
Theoretical distributions

Theoretical distributions can be helpful

- Have well-known properties!
- If variable in our data well approximated by a theoretical distribution –> attribute properties to the variable
- Real life, many variables surprisingly close to theoretical distributions.
- Will be useful when generalizing from data - Chapter 05

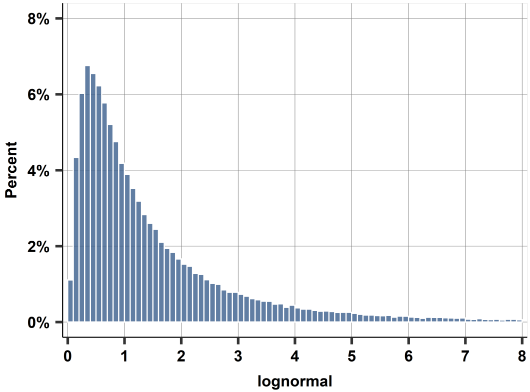
The Normal distribution

- Histogram is bell-shaped
- Outcome (event), can take any value
- Distribution is captured by two parameters
  - μ is the mean
  - σ 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.



A few more points on the Normal and log-normal

- Many many variables in real life are close to normal
- Especially when based on elementary things which are added up
- Not good approximation when
  - some reasons for non-symmetry
  - extreme values are important

- Variables are well approximated by the log-normal if they are the result of many things *multiplied* (the natural log of them is thus a sum).

Income and log-income

Figure: income

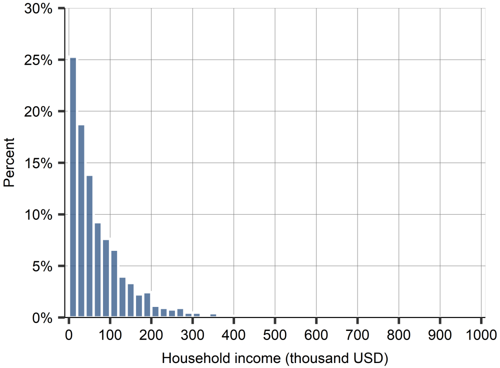
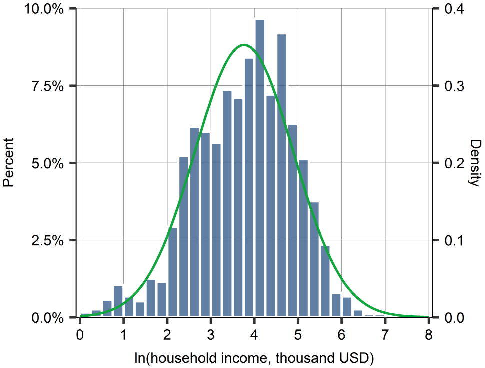
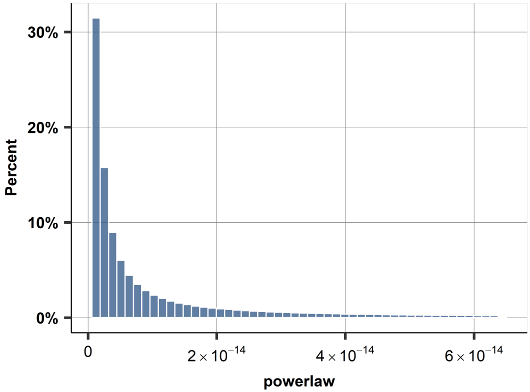


Figure: log income



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 vizualization

Data visualization: Steps

- We shall make conscious decisions and not let default settings guide us.
- Usage - what you want to show and to whom – deciding on purpose, focus, and audience

- Pick a geometric object – decide how information is conveyed: we need to choose a geometric object to visualize the information we want to show.
- Encode information – choose details of the object (color, height)
- Settle on scaffolding: – supporting features of the graph such as axes, labels, and titles.

This is a very brief overview, read more in textbook – Chapter 03

Data visualization: usage

- What is the purpose, what message 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: geoms and encoding

- Geometric object: Pick an object suitable for the information to be conveyed. A line showing value over time.
- May be one or more geoms. Dots for years and a trend line
- Encoding: Pick one encoding only Position of the line Don't apply different colors or shades

Data visualization: process

- Decide on geometric objects, and build graphs from bottom up. Advanced, dataviz experts
- Decide on a type of graph (such as bar chart), and define its elements (geoms). Top down. Most social science / business.

- Graph type – Can pick a standard object to convey information Histogram: bars to show frequency

Data visualization: scaffolding

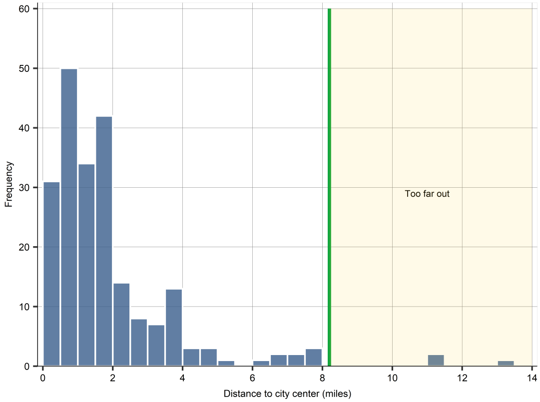
- How to present elements that support understanding.
- Make sure, a graph has
  - Title
  - Axis title and labels
  - Legend
- Content as well as format, such as font type and size.

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



AI and data exploration

GenAI is very good at describing the data

- Create histograms
- Add scaffolding
- Iterate on bin size, color, font size
- Download graph + Python code

AI and distributions

GenAI is very good at creating simulated data for distributions

- Explain the distribution, and maybe its characteristics
- Set N
- Iterate on parameters of interest
- Get explanations
- Download graph + Python code

Summary steps of EDA

1. First focus on the most important variables. Go back to look at others if subsequent analysis suggests to.
2. For qualitative variables, list relative frequencies.
3. For quantitative variables, look at histograms. May decide for transformation, find extreme values, learn about key aspects of data.
4. Check for extreme values. Decide what to do with them.
5. Look at summary statistics. It may prompt actions, such as focusing on some part of the dataset.
6. Do further exploration if necessary (time series data, comparisons across groups of observations, etc.)