Lecture 3 Exploratory Data Analysis
EDA

Tamer Çetin

Learning Outcomes

- Understand the importance and uses of EDA
- Understand distributions, visualize them, and describe the most important features
- Identify situations when extreme values matter and make decisions about extreme values
- Produce graphs and tables of presentation that are focused, informative, and easy to read
- Know the main features of the most important theoretical distributions and assess whether they are good approximations of distributions of variables in actual data

Motivation

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub

```
Data columns (total 81 columns):
Id                1460 non-null int64
MSSubClass        1460 non-null int64
MSZoning          1460 non-null object
LotFrontage       1201 non-null float64
LotArea           1460 non-null int64
Street            1460 non-null object
Alley             91 non-null object
LotShape          1460 non-null object
LandContour       1460 non-null object
Utilities         1460 non-null object
LotConfig         1460 non-null object
LandSlope         1460 non-null object
Neighborhood      1460 non-null object
Condition1        1460 non-null object
Condition2        1460 non-null object
BldgType          1460 non-null object
HouseStyle        1460 non-null object
OverallQual       1460 non-null int64
OverallCond       1460 non-null int64
YearBuilt         1460 non-null int64
YearRemodAdd      1460 non-null int64
```

Motivation

- EDA aims;
  to describe variables in data
  to understand potential problems with data/variables
  to help additional cleaning data
  to prepare data for the ultimate analysis
- How to explore the data and check whether it is clean enough for (further) analysis?
  Tools such as frequency, probabilities, distribution, and histogram!
- How to describe data and present the key features?
  Target and covariates!
- How should you start the analysis itself?
  By knowing data: prediction/classification!

Exploratory data analysis (EDA) - describing variables

EDA is an essential first step of data analysis!

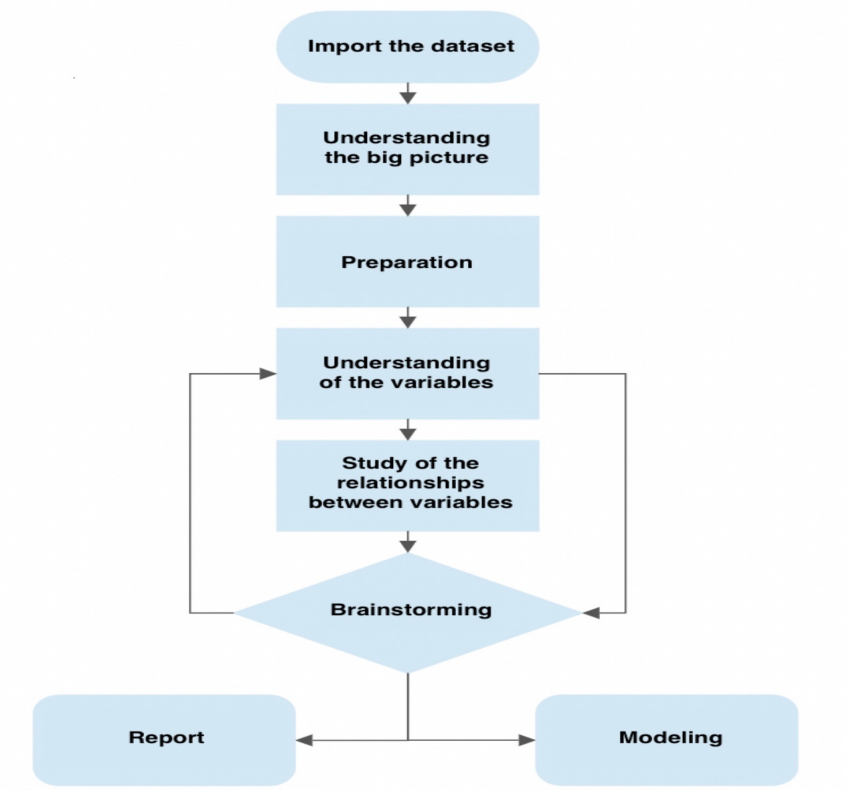
Reasons to do EDA!

1. To provide informative description of variables (descriptive analysis)
2. To know if data is clean and ready for the analysis (part of iterative data cleaning process)
3. To guide subsequent analysis (providing descriptive analysis/informative description of variables for further analysis)
4. To give context such as feature importance (for interpretation)

   Feature importance does not to need to be at the end of the analysis!!!
5. To ask additional questions (for specifying the (research) question)

   Quite often, EDA uncovers patterns for a variable of interest!!!
6. To answer the business/research questions using simple tools
   The end of the analysis: Rare but possible!

Exploratory data analysis (EDA)



Key tasks in EDA: 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

⇓

- to understand variables you work with
- to make comparisons

Key tasks in EDA: describe variables

```
loan.isnull().sum()
```

```
id                                0
member_id                         0
loan_amnt                         0
funded_amnt                       0
funded_amnt_inv                   0
                                 ...
tax_liens                        39
tot_hi_cred_lim               39717
total_bal_ex_mort             39717
total_bc_limit                39717
total_il_high_credit_limit    39717
Length: 111, dtype: int64
```

```
loan.isnull().sum()/len(loan)*100
```

```
id                              0.0000
member_id                       0.0000
loan_amnt                       0.0000
funded_amnt                     0.0000
funded_amnt_inv                 0.0000
                                   ...
tax_liens                       0.0982
tot_hi_cred_lim               100.0000
total_bal_ex_mort             100.0000
total_bc_limit                100.0000
total_il_high_credit_limit    100.0000
Length: 111, dtype: float64
```

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: When a variable has missing values, absolute frequency is usual choice!
- Relative frequency is useful to see probability.
  A measure of the likelihood of an event!
- An event is the occurrence of a particular value of a variable.
  Manager of the firm is female:
  This event occurs various times in data so its probability is relative frequency.

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)

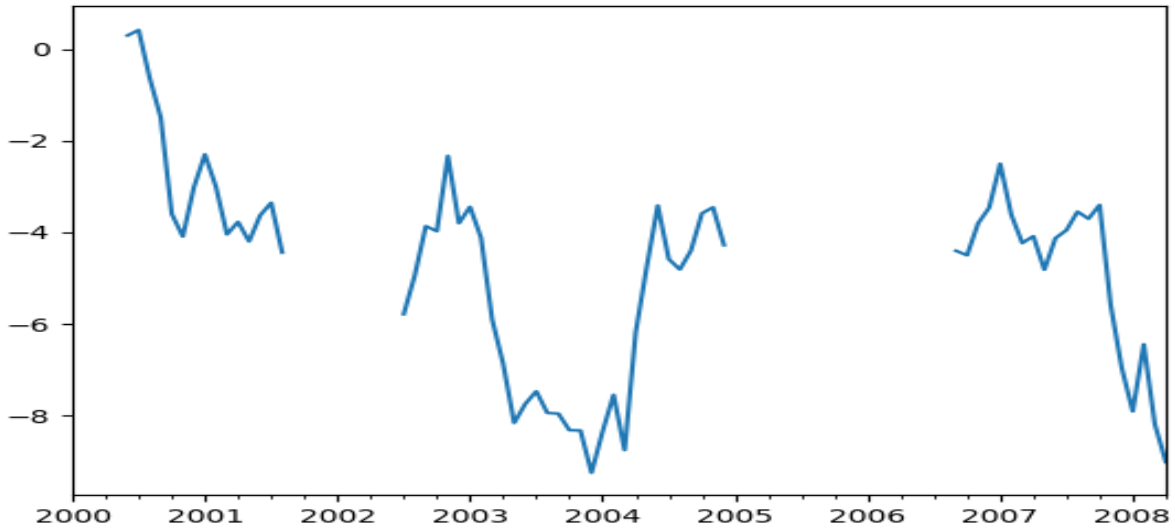
Missing Values

- The best and most time-consuming way to eliminate missing values is to fill them in yourself, provided it is possible to obtain accurate information through further research.
- Missing values in categorical variables
- Delete the observations with missing value
- But in what conditions
- Replace Missing Values with the Most Frequent Value
- Develop a model to predict missing values
- If the variables in question are qualitative (nominally scaled), missing values can be avoided by creating a new class.

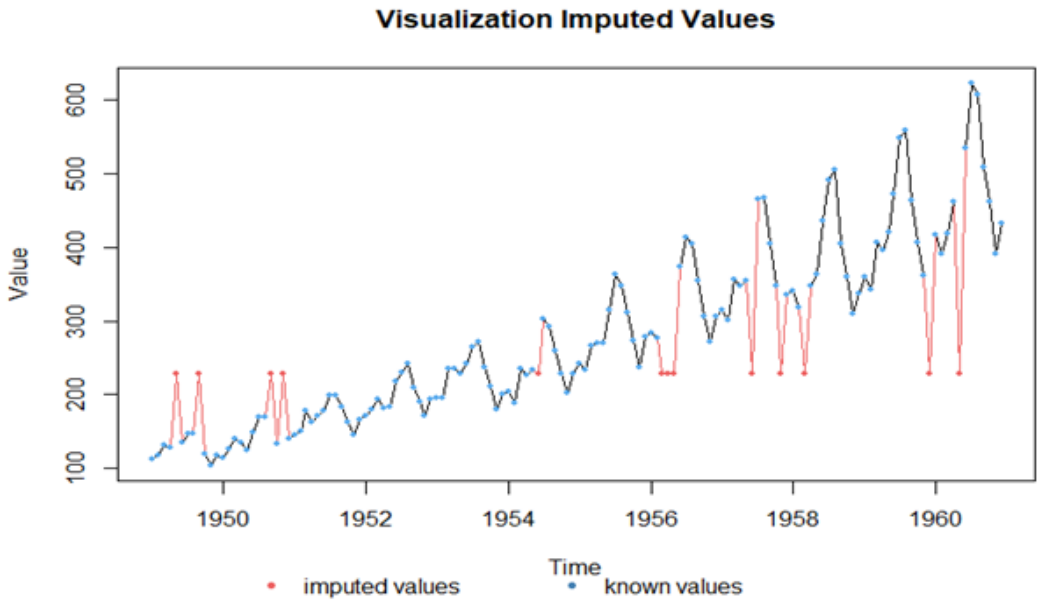
Missing Values

- Missing values in numerical variables
- If missing values are more than significant amount of data, remove them
- Make sure omitted data points are non-systematic!
- Calculate the arithmetic mean of variable and use the mean value to impute missing data using the mean

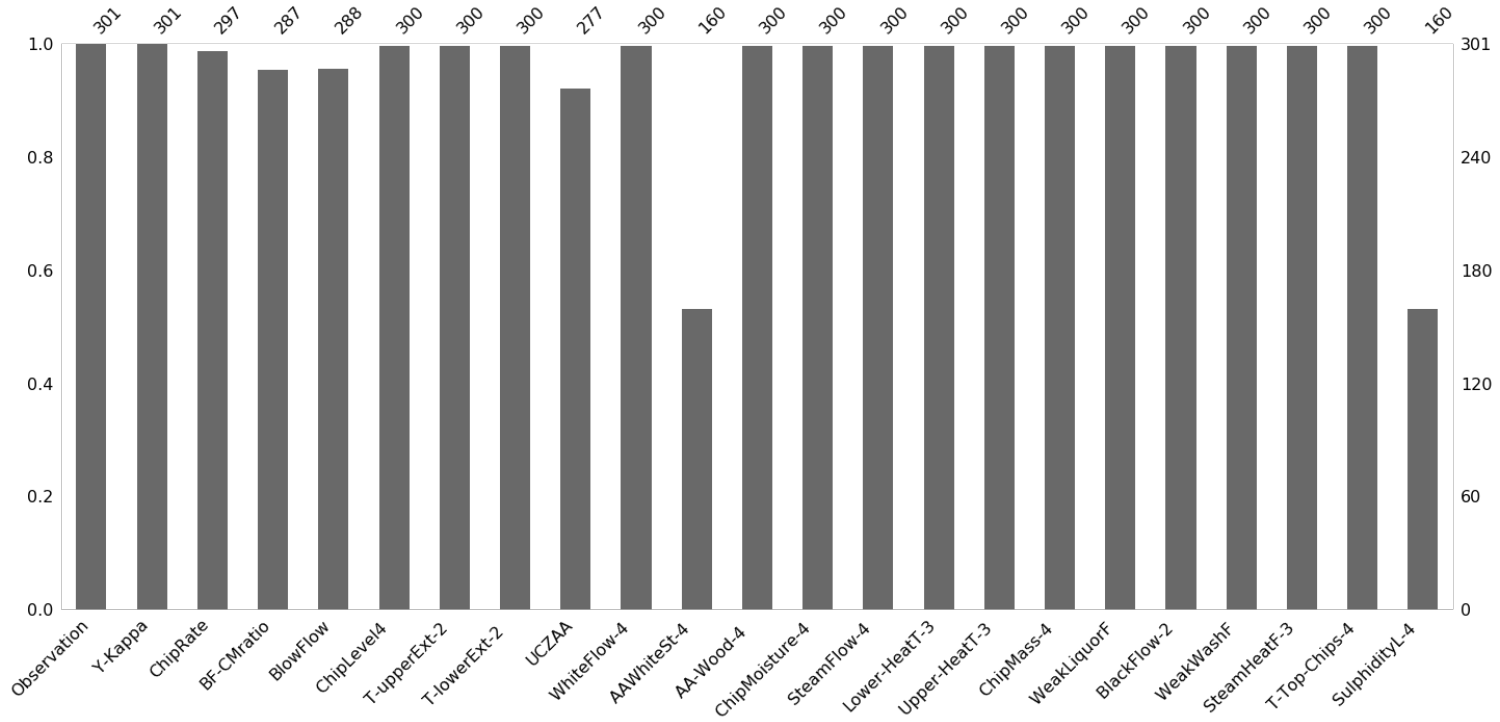
Missing Values: Charts/Time Series Data



Missing Values: Charts/Time Series Data



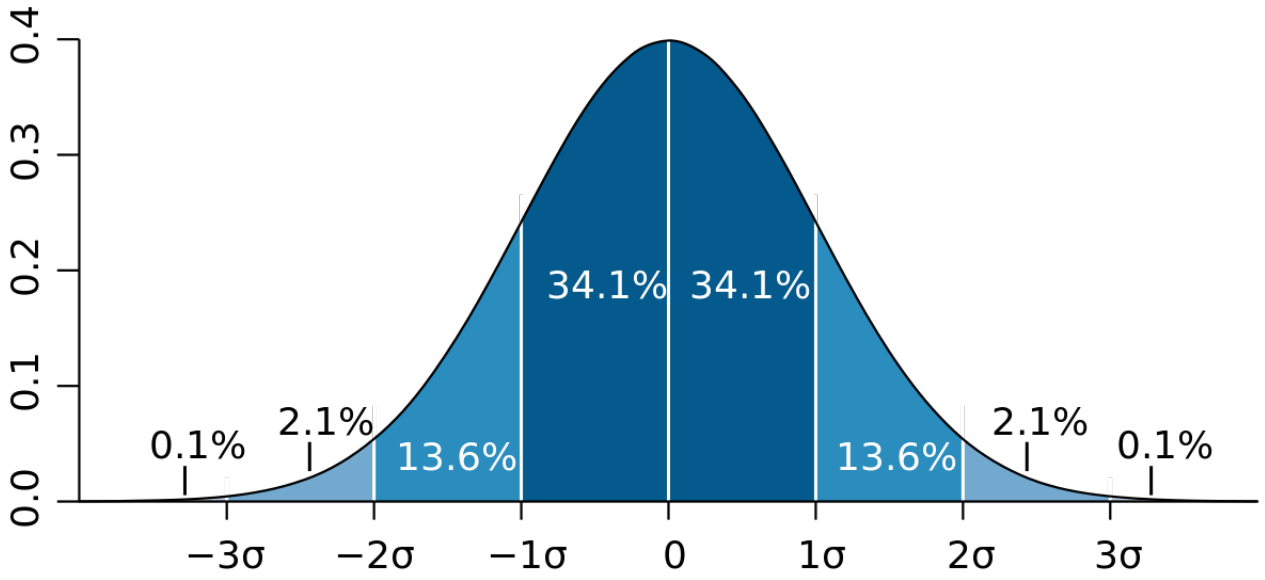
Missing Values: Bar Graphs



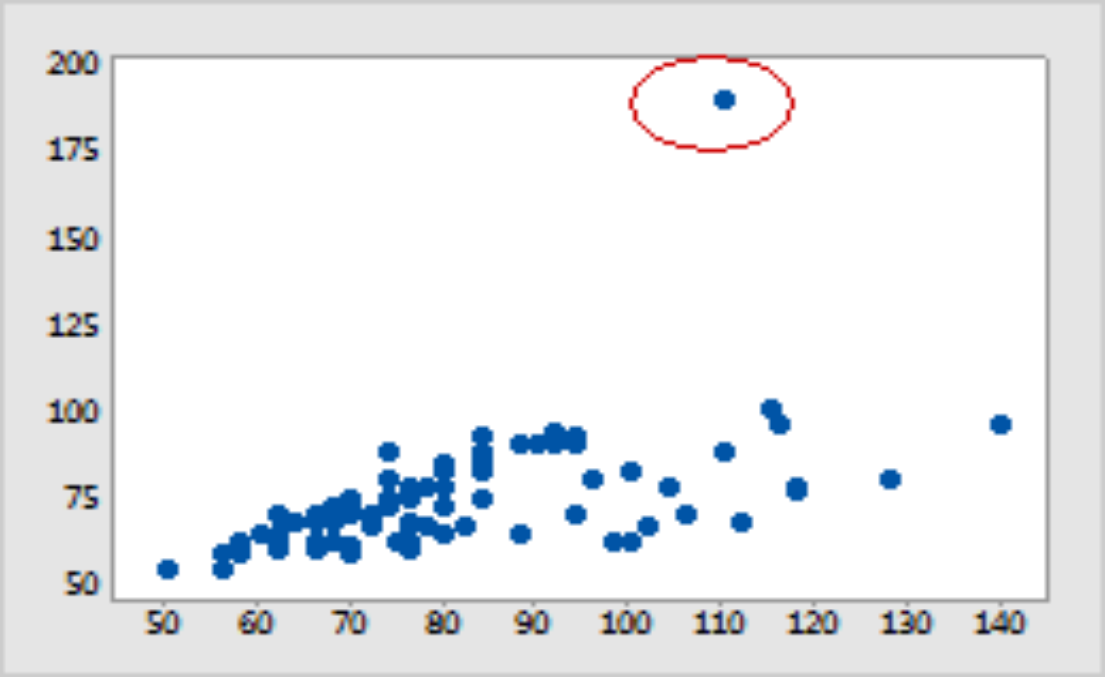
Outliers and Obviously Incorrect values

- A problem similar to missing values is that of obviously incorrect values/outliers.
- Standard datasets often contain both.
- Outliers must be removed from data.
- Why?
- High squared-root errors
- Overfitting

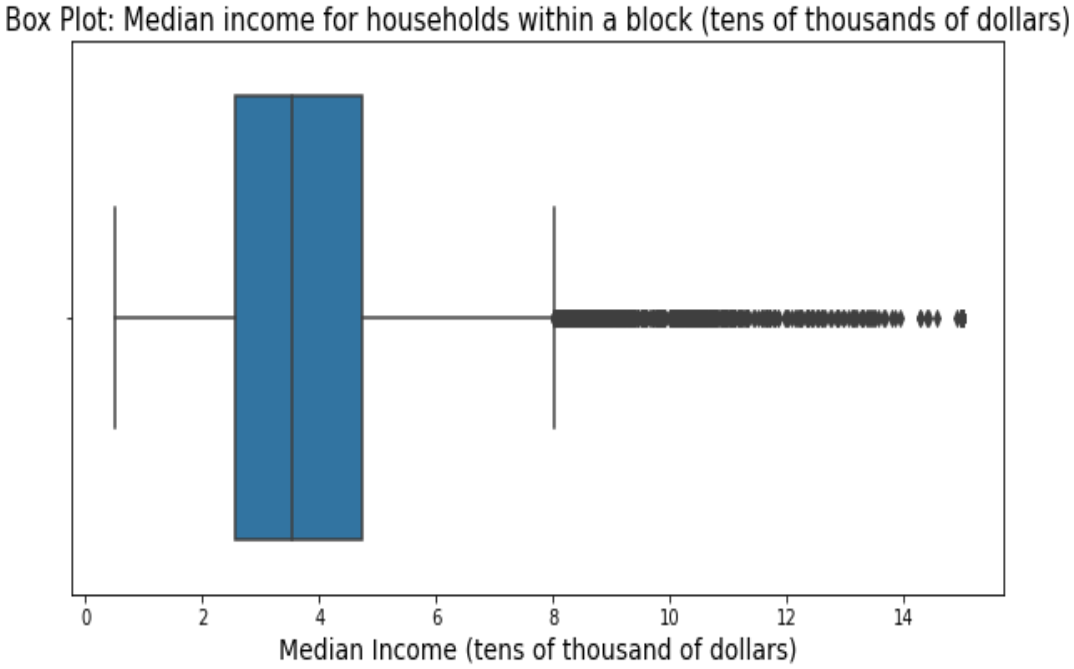
Some Theory on Outliers



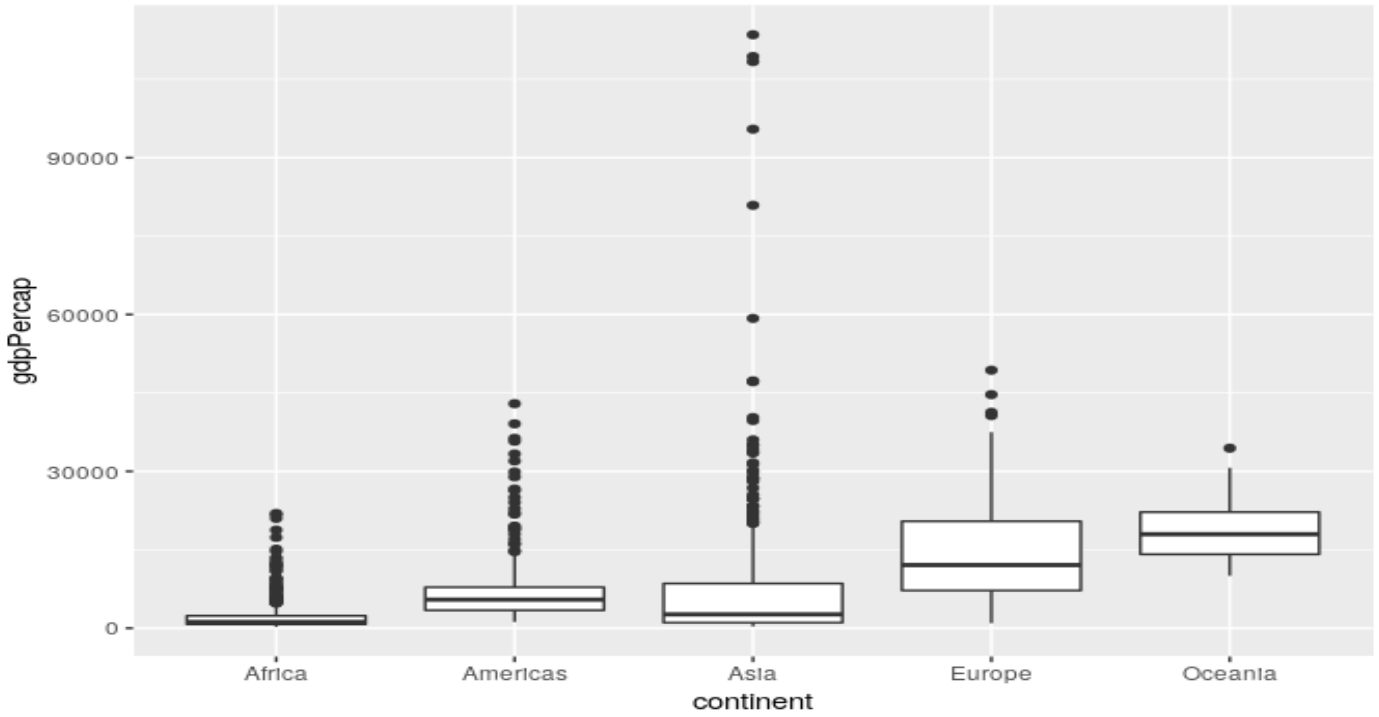
Checking Outliers: Scatter Plots



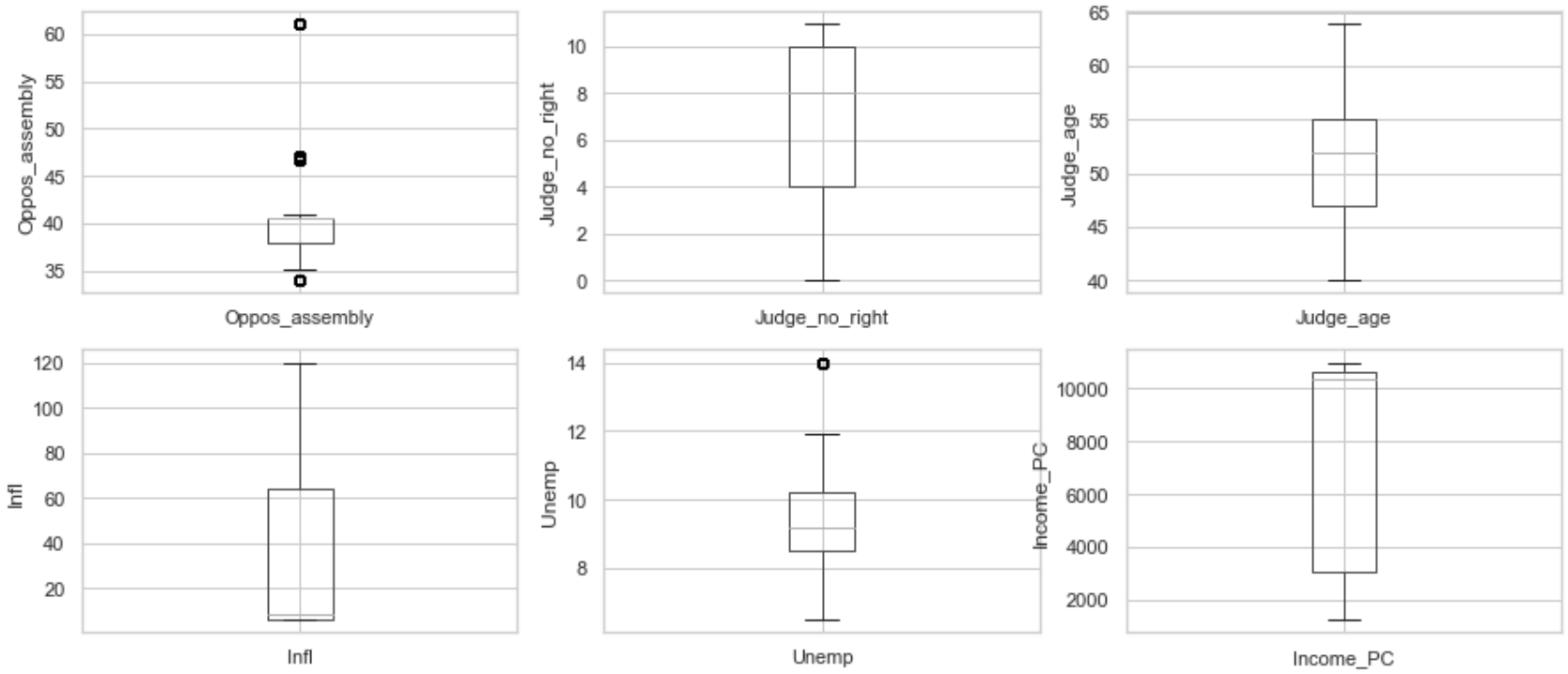
Checking Outliers: Box Plots



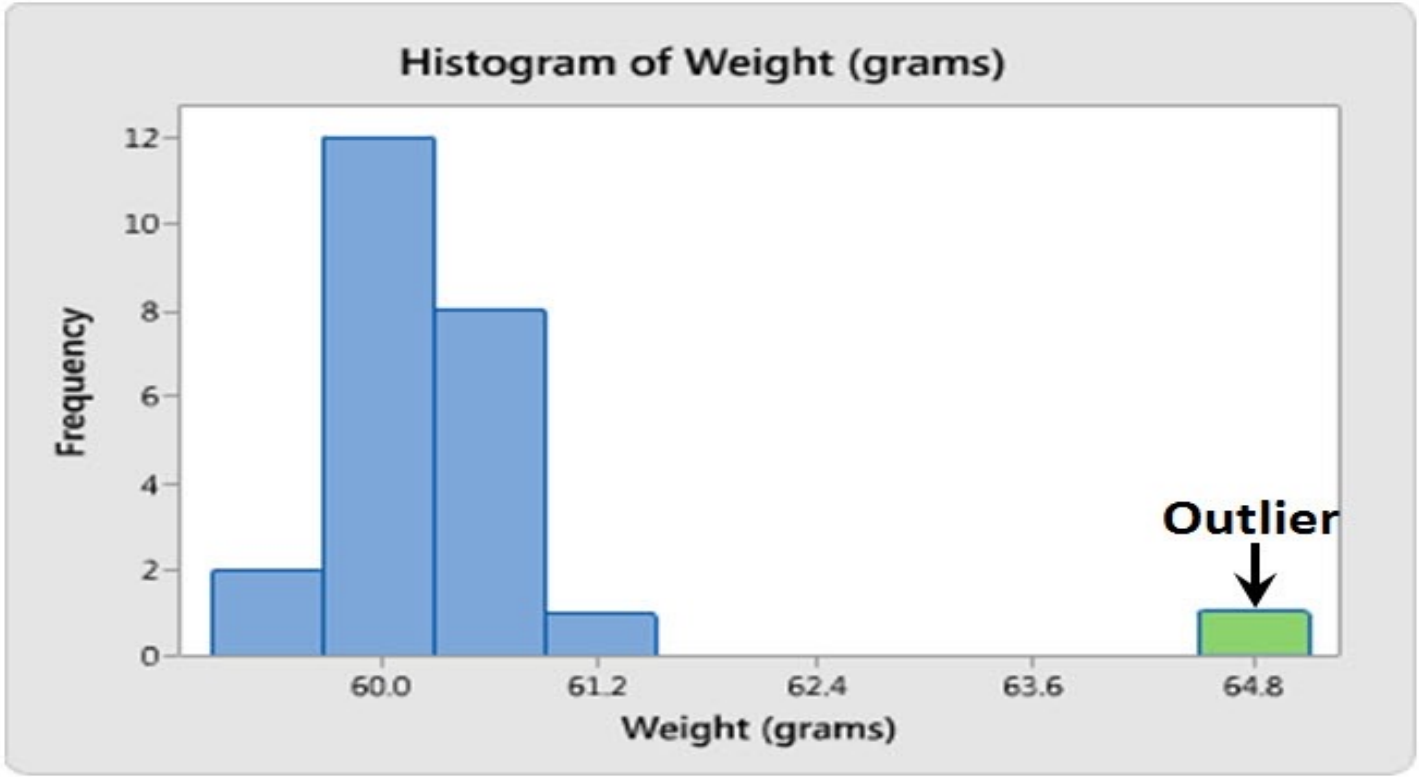
Checking Outliers: Boxplots



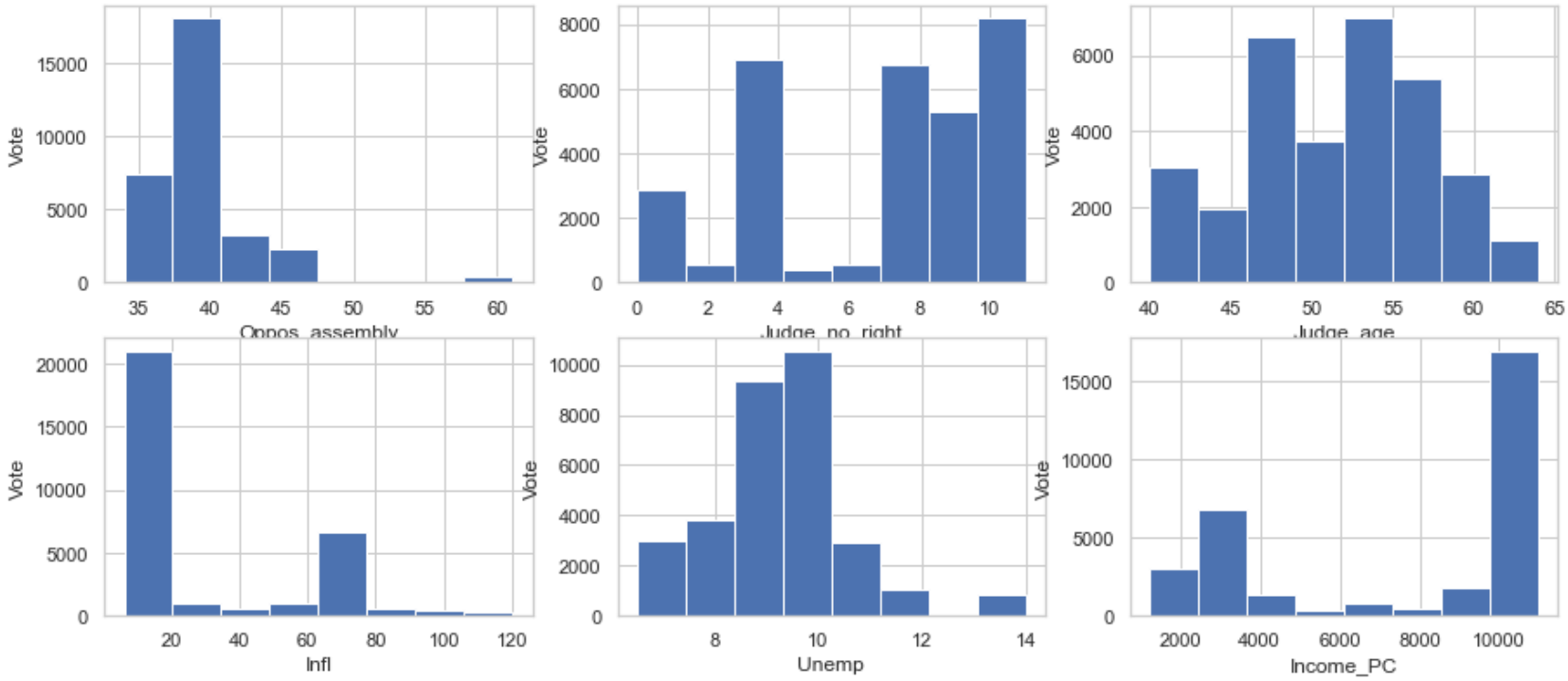
Checking Outliers: Boxplots



Checking Outliers: Histograms

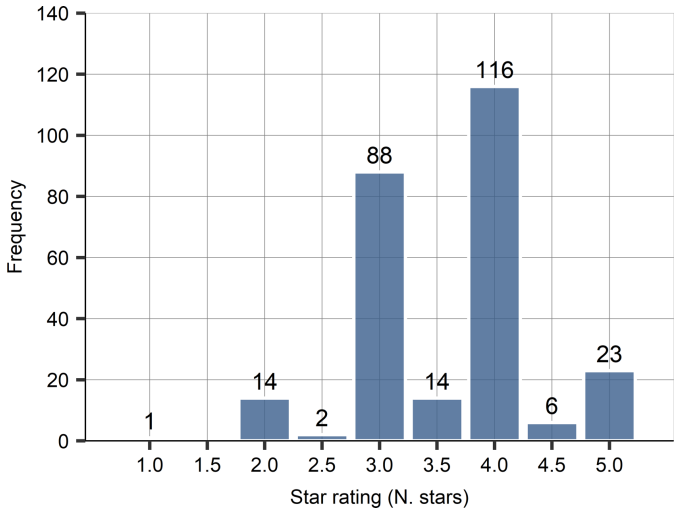


Checking Outliers: Histograms

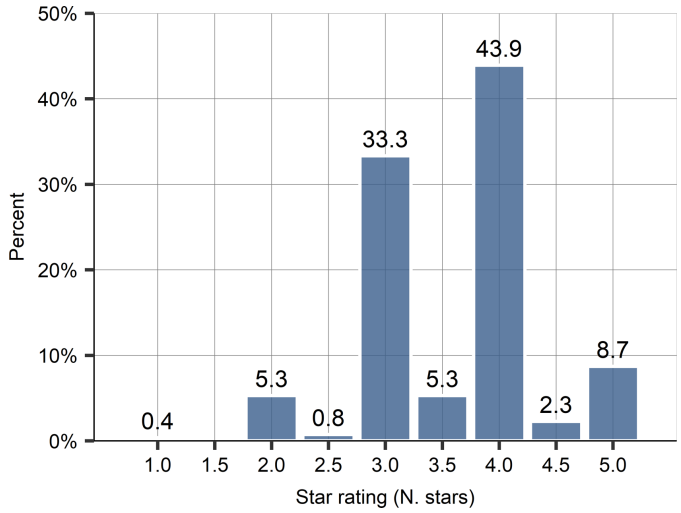


Hotel price histograms

(a) Absolute frequency (count)



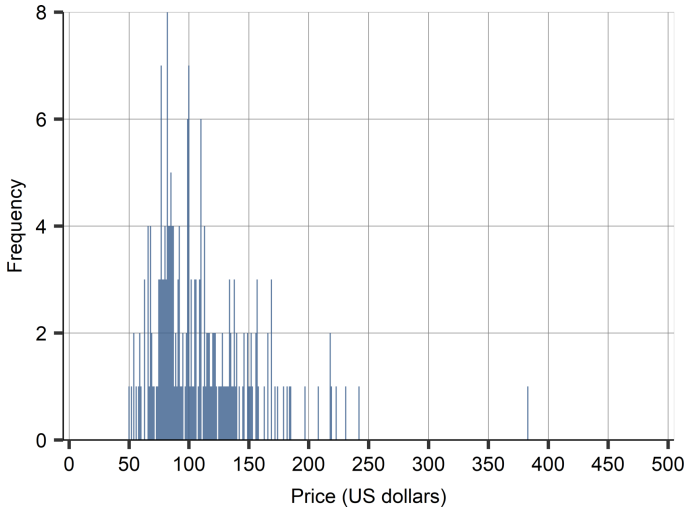
(b) Relative frequency (percent)



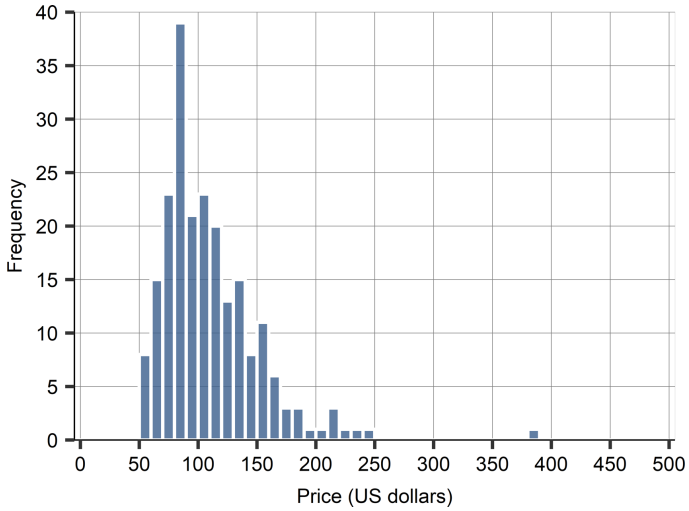
Source: hotels-vienna dataset. Vienna, Hotels only, for a 2017 November weekday

Hotel price histograms

(a) Histogram: individual values



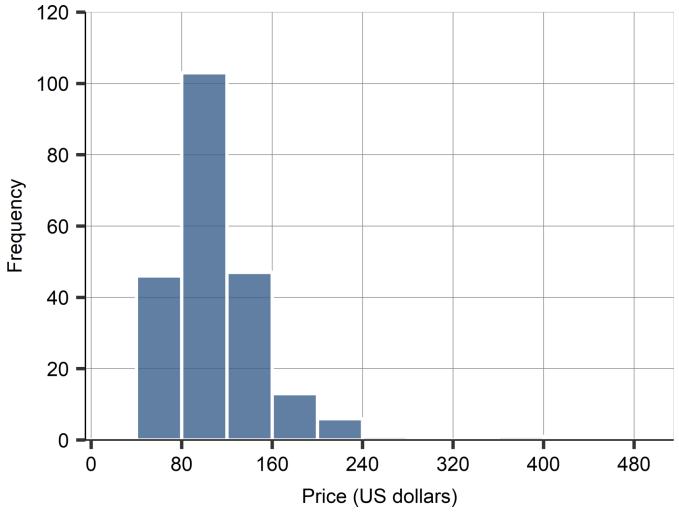
(b) Histogram: 20$ bins



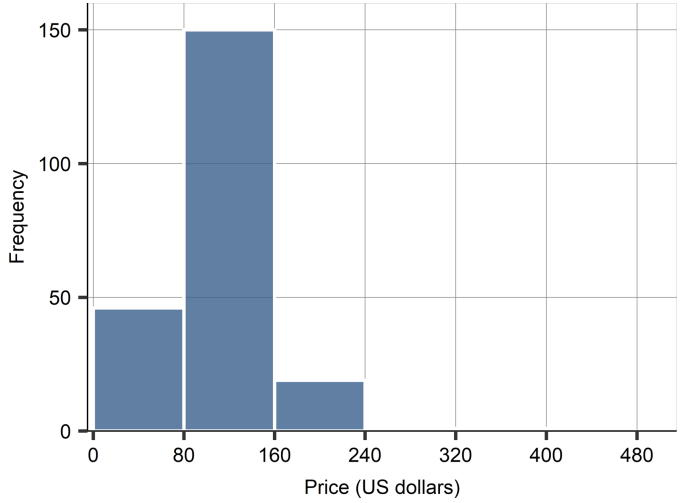
Note: *Panel (a) just shows individual values - help see where most values are. Panel (b) is a histogram with 20$ bins - more useful to capture frequencies.* Source: hotels-vienna dataset. Vienna, 3-4 stars hotels only, for a 2017 November weekday

Hotel price histograms

(a) Histogram: 40$ bins



(b) Histogram: 80$ bins

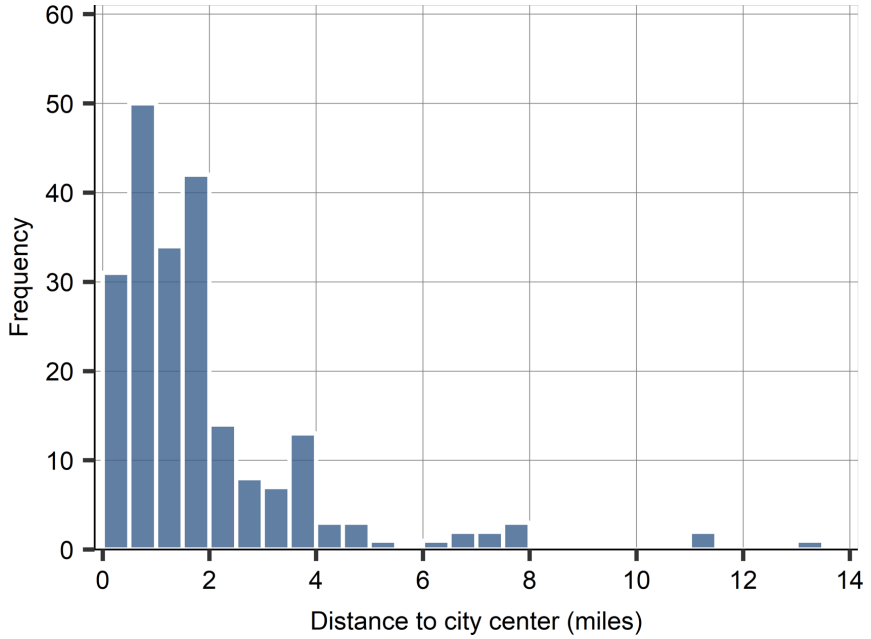


Note: *Bin size matters. Wider bins suggest a more gradual decline in frequency.*

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.

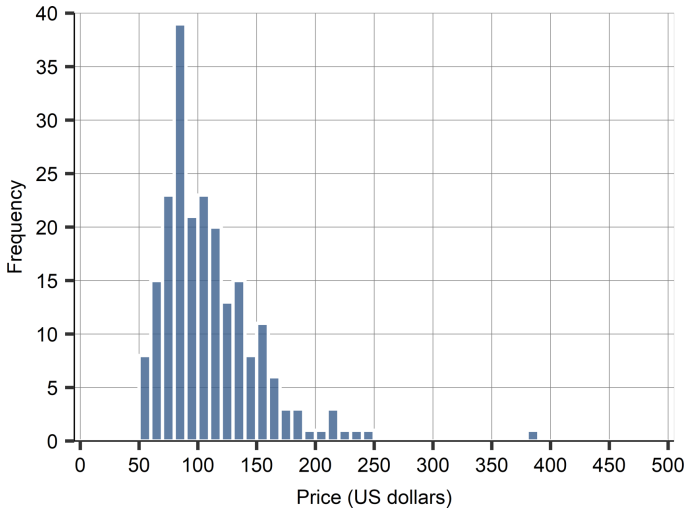


Hotel prices

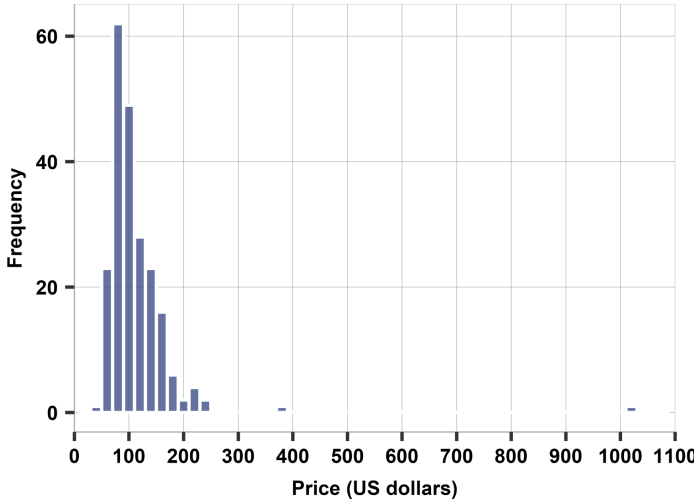
- Vienna all hotels, 3-4 stars
- Use absolute frequency (count)
- We go back to prices
- How to decide what to include? -> check observation!

Hotel price histograms

(a) Histogram: 20$ bins as seen



(b) Histogram: including extreme value above 1000$



Source: hotels-viennadataset. 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 data.
- Basic *summary statistics* describe the most important features of distributions of variables.

Panel A: Call Option Contracts

	ttm	mness	embed_lev	iv	delta	gamma	vega	theta	$r^{\Delta}_{S\,pot}$
Mean	129	1.05	32.02	0.16	0.20	0.002	196.85	-55.10	-3.14%
Median	91	0.97	25.99	0.15	0.18	0.002	155.54	-44.19	-2.31%
Std. Dev.	99	0.67	21.00	0.07	0.15	0.002	159.82	45.86	1.33%
No. Obs.	24,749	24,749	24,749	24,749	24,749	24,749	24,749	24,749	24,749

Panel B: Put Option Contracts

	ttm	mness	embed_lev	iv	delta	gamma	vega	theta	$r^{\Delta}_{S\,pot}$
Mean	123	-1.16	19.10	0.26	-0.15	0.001	168.03	-70.60	-5.18%
Median	91	-1.20	16.73	0.24	-0.10	0.001	123.02	-59.06	-4.26%
Std. Dev.	99	0.66	11.08	0.10	0.14	0.002	148.34	48.94	1.48%
No. Obs.	52,341	52,341	52,341	52,341	52,341	52,341	52,341	52,341	52,341

Summary statistics

Table A1: Descriptive Statistics

	October 1998 mean
Head of household:	
... Is indigenous	0.41
... Age	41.13
... Education (HS or higher)	0.005
... Is male	0.94
... Is an agricultural worker	0.65
Household size	
... Number of children less than 6 years old	1.97
... Number of children 6-16 years old	2.81
... Number of adults 17+ years old	2.54
Log monthly average per capita consumption (log pesos)	5.08
Average number of days a school-age child misses school	0.32
Average number of days a young child is sick	1.07
Assigned to treatment group	0.61
N	6537

Notes: Table shows the average levels in October 1998 of households matched to November 1999 survey sample. HS education defined as 12 years or more of education. Number of days a young child is sick, and number of days a school-age child misses school, are computed as an average over the number of children in the respective age group in the household. Sample restricted to only households with children in the targeted categories for health and schooling intervention (0-5 y.o., 6-16 y.o.) during the November 1999 survey.

Table 1. Summary Statistics

	Full sample	Obtained a loan	Did Not Obtain a Loan
Applied before deadline	0.085	1	0.01
Obtained a loan before deadline	0.074	1	0
Loan amount in Rand	110	1489	0
	(536)	(1351)	(0)
Loan in default		0.12	
Got outside loan and did not apply with Lender	0.22	0.00	0.24
Maturity = 4 months		0.81	
Offer rate	7.93	7.23	7.98
Last loan amount in Rand	1118	1158	1115
	(829)	(835)	(828)
Last maturity = 4 months	0.93	0.91	0.93
Low risk	0.14	0.30	0.12
Medium risk	0.10	0.21	0.10
High risk	0.76	0.50	0.78
Female	0.48	0.49	0.48
Predicted education (years)	6.85	7.08	6.83
	(3.25)	(3.30)	(3.25)
Number previous loans with Lender	4.14	4.71	4.10
	(3.77)	(4.09)	(3.74)
Months since most recent loan with Lender	10.4	6.19	10.8
	(6.80)	(5.81)	(6.76)
Race = African	0.85	0.85	0.85
Race = Indian	0.03	0.03	0.03
Race = White	0.08	0.08	0.08
Race = Mixed ("Coloured")	0.03	0.04	0.03
Gross monthly income in Rand	3416	3424	3416
	(19657)	(2134)	(20420)
Number of observations	53194	3944	49250

Means or proportions, with standard deviations in parentheses.

Summary statistics: Sample mean

The most used statistic is the *mean*:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

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

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

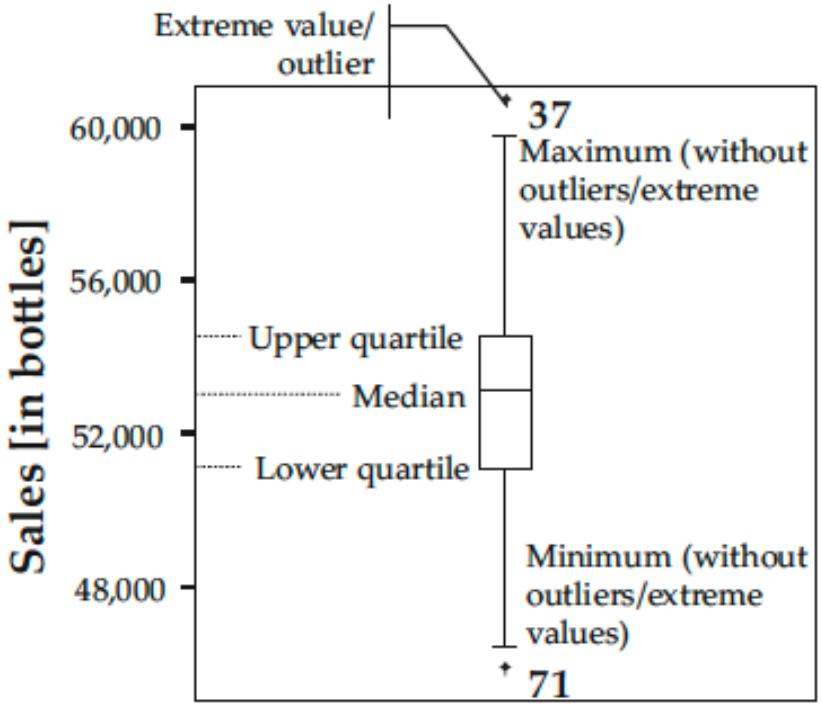
Summary statistics: The median and other quantiles

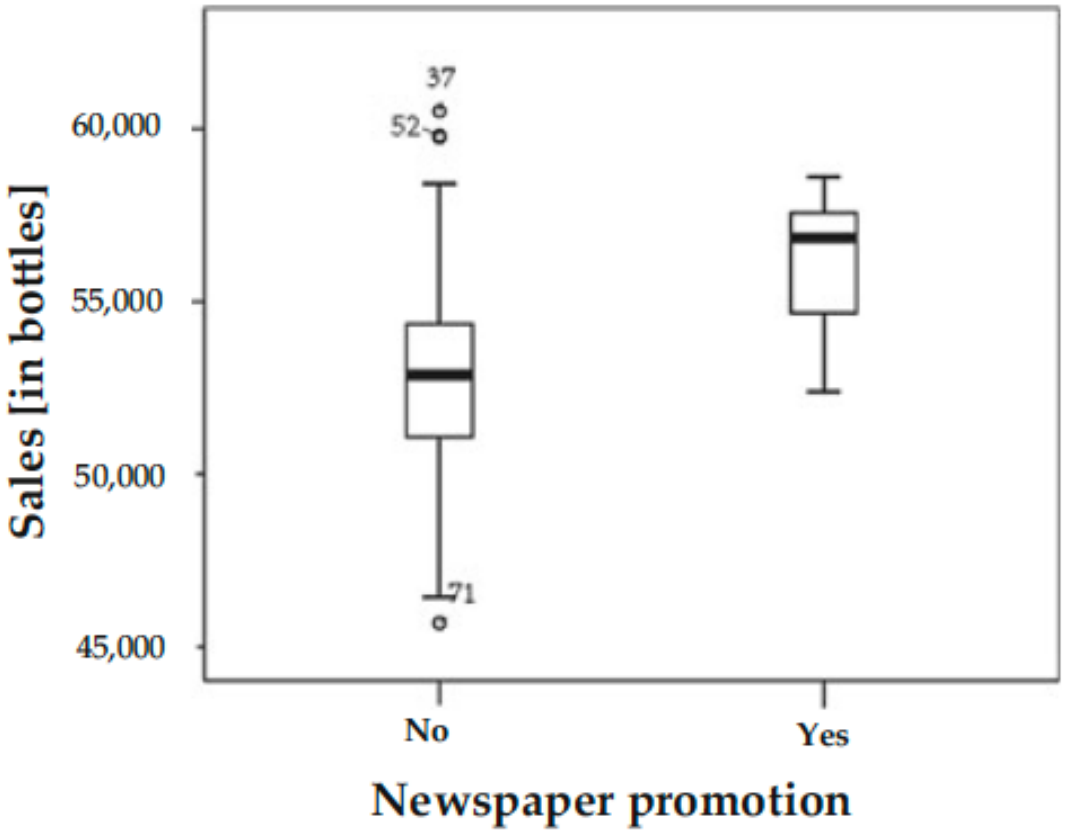
- *Quantiles*: a quantile is the value that divides the observations in the dataset to two or more equal parts in specific proportions
- The *median* is the middle value of the distribution - half the observations have lower value and the other half have higher value
- *Percentiles* divide the data into parts along a certain percentage

  Percentiles (100-quantiles): 99 percentiles split the data into 100 parts
- *Quartiles* divide the data into two parts along fourths

  Quartiles (4-quantiles): Three quartiles split the data into four parts
  Deciles (10-quantiles): Nine deciles split the data into 10 parts

Summary statistics: The median and other quantiles



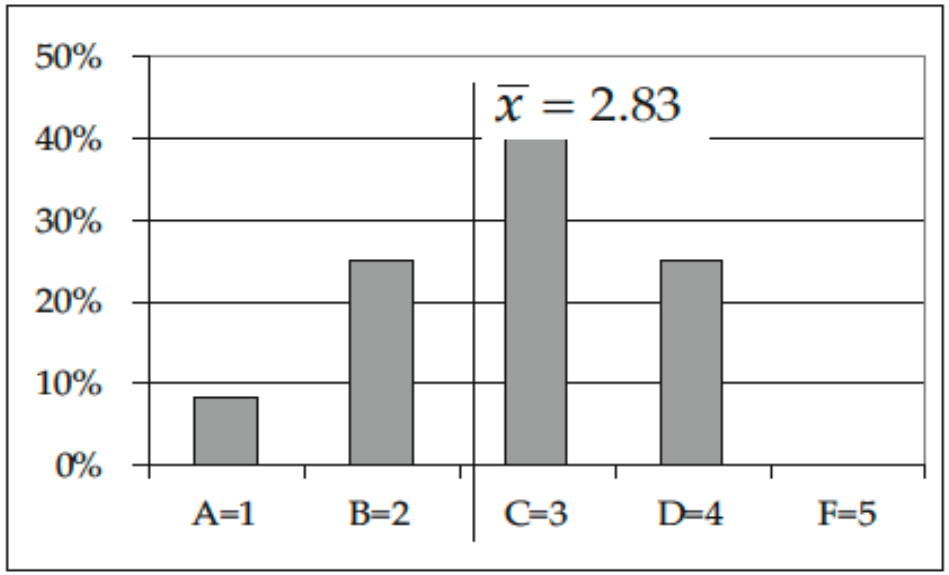


Summary statistics: The mode

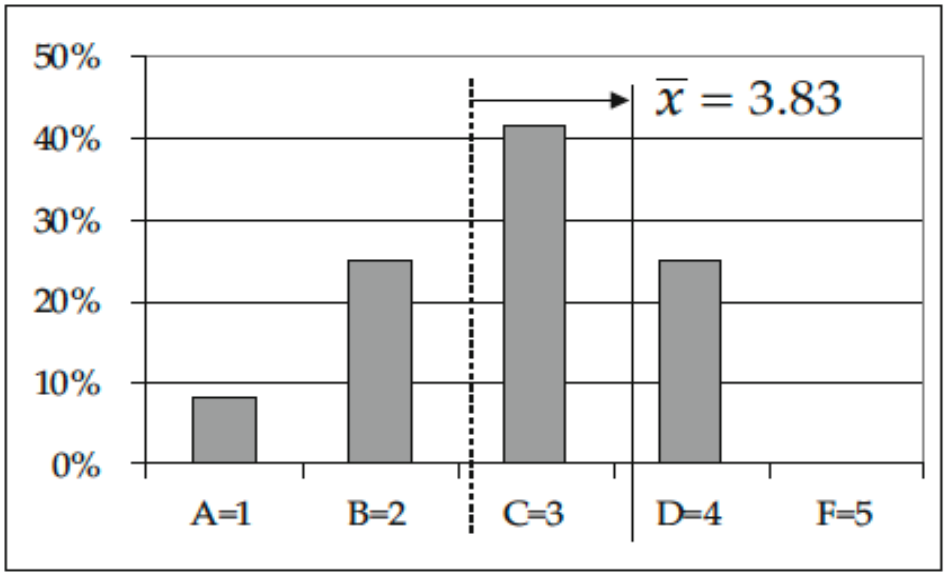
- The *mode* is the value with the highest frequency in the data.
- It identifies the value that appears most frequently in a distribution.
  The mode is the "champion" of the distribution.
- Some distributions are unimodal, others have multiple modes.
- Multiple modes are apart from each other, each 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.



Part 1

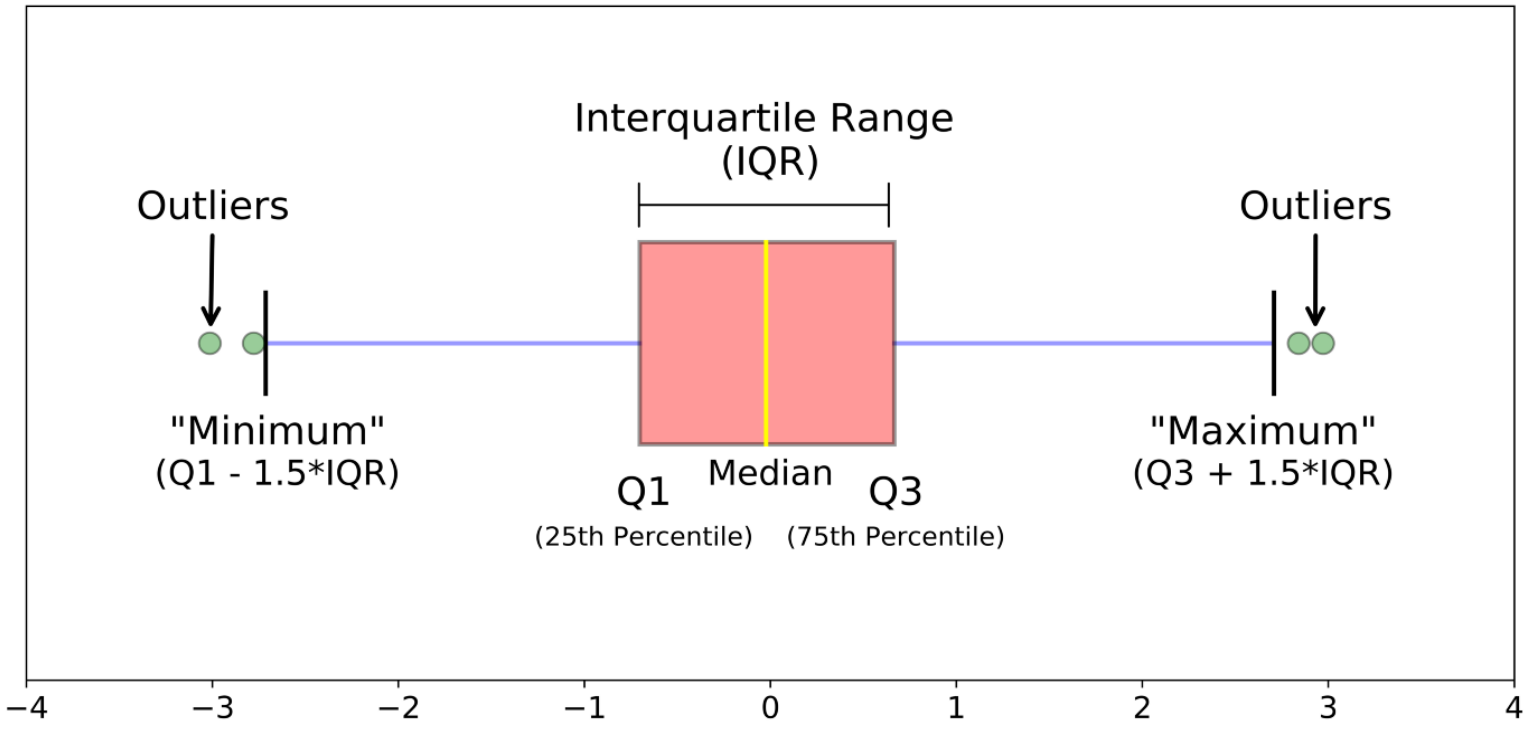


Part 2

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-quartile range* 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.

Some Theory on Outliers



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.
- 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.

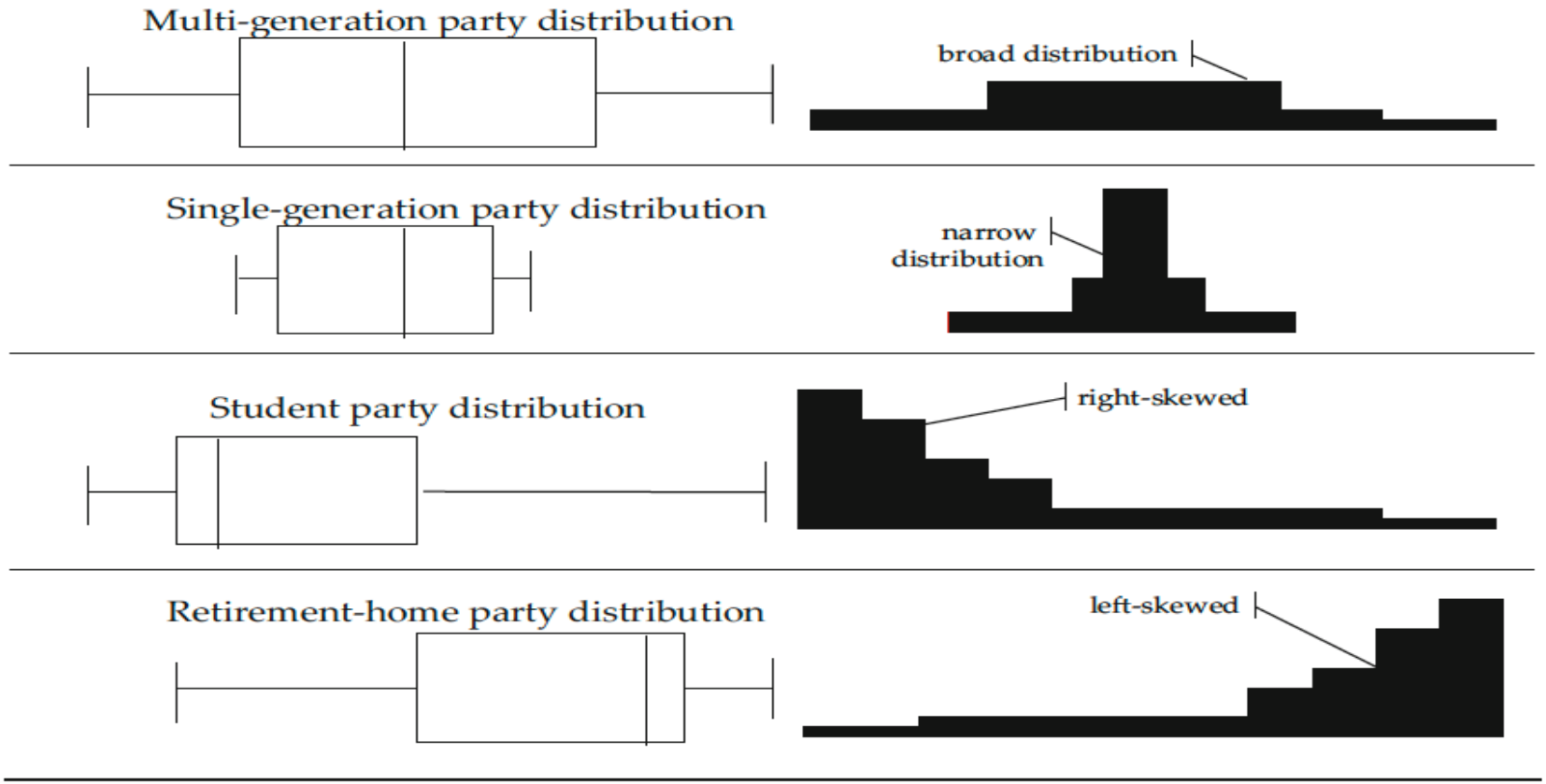
$$Var(x)_{emp} = S^2_{emp} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

$$Sdt(x) = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$

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: The median and other quantiles



Summary statistics: skewness measure

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

$$Skewness = \frac{(\bar{x} - median(x))}{Sdt(x)} \tag{6}$$

- 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*.

Income and log-income

Figure: income

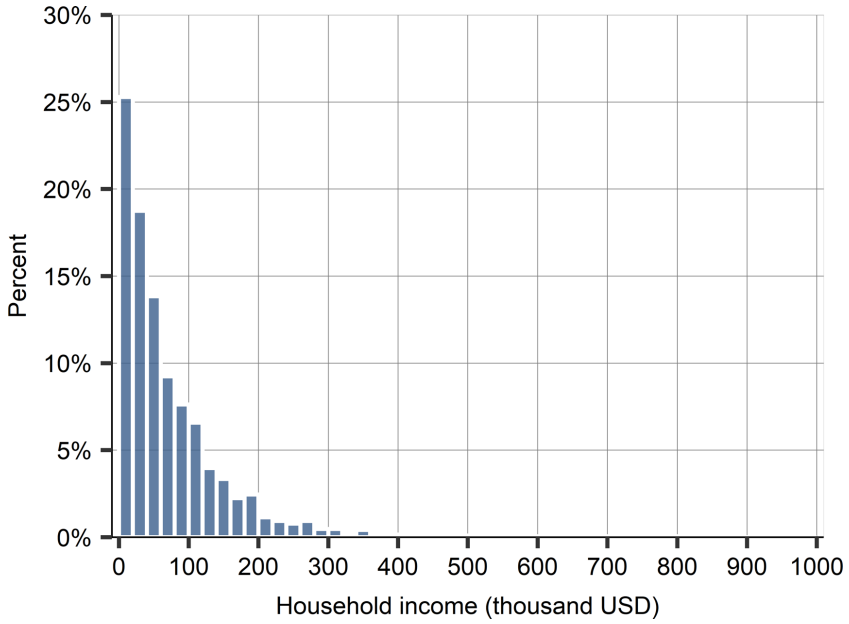
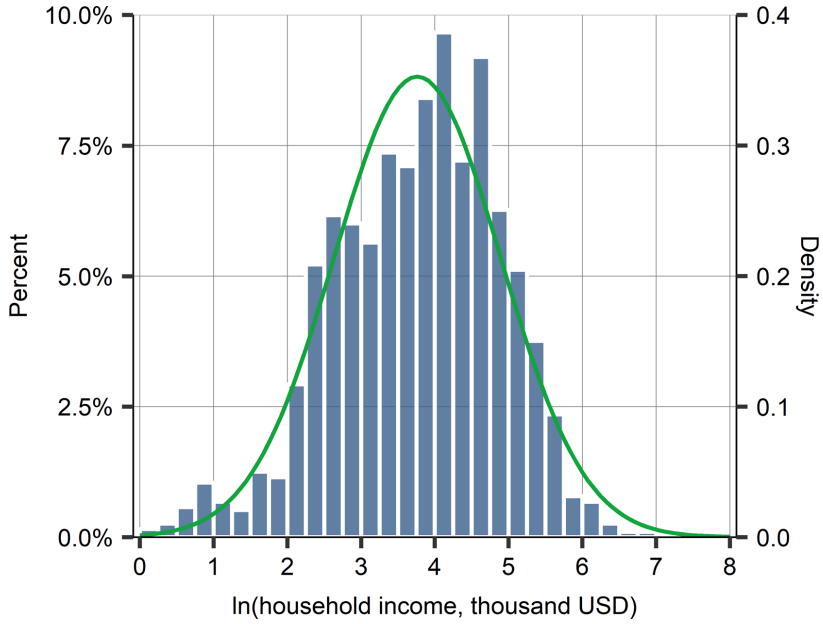


Figure: log income



Source: https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations

Data visualization: encoding

Converting categories in qualitative features to numbers as numerical variables

- 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: encoding ---> one-hot encoding

To create dummies for binary variables!!!

Original categorical column

Origin
USA
Japan
Europe
USA
Europe

One-Hot encoded columns

Origin_USA	Origin_Japan	Origin_Europe
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

Data visualization: encoding ---> label encoding

Only ordinal variables!!!

Original categorical column

Education
High School
Primary School
Master Degree
Bachelor Degree
High School

Label encoded column

Education
2
1
4
3
2

Data visualization: issues with encoding

- Sparsity: data with high amount of zeros.
- It is not good for ML algorithms!

1. Overfitting
2. Ml models focus on the importance of dense data but not sparse data
3. Space cost
4. Time cost
5. Some ML algorithms perform badly/poorly on sparse datasets

Data visualization: issues with encoding

- How to handle sparsity

1. Convert the feature from sparse to dense

   PCA

   Feature hashing: data is binned into desired number of outputs

   Perform feature selection and feature extraction

   Use t-SNE (t-Distributed stochastic neighbor embedding)
2. Remove features from the model
3. Use ML algorithms less sensitive to sparse datasets
4. Time cost
5. Some ML algorithms perform badly/poorly on sparse datasets

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.