DMI

Exploratory Data Analysis

Data Mining and Data integration in Biomedicine
Master in Bioinformatics

Janet Piñero
Medbioinformatics Solutions SL
2025-2026

Add as a comment

A	B	C	D	E	F	G	H	I
Nom	Cognoms	Número ID	github	presentati	article titl	session 1	session 2	session 3
ITXASO	ALONSO RODRÍGUEZ							
PABLO	DONATO			9				
DAVID	FRAGOSO BENTO			10				
TANNER ALEXANDER	GARCIA							
DIEGO	MAQUIEIRA VIDAL							
MONTSERRAT	PALAZÓN BALMASEDA			10				
REGINA	RODRÍGUEZ DURANT REYES			9				
MARTA	SOTELO MONTORO							
NAHIA	URRA LÓPEZ DE HEREDIA							

Refining modules to determine functionally significant clusters in molecular networks ❌

Link here

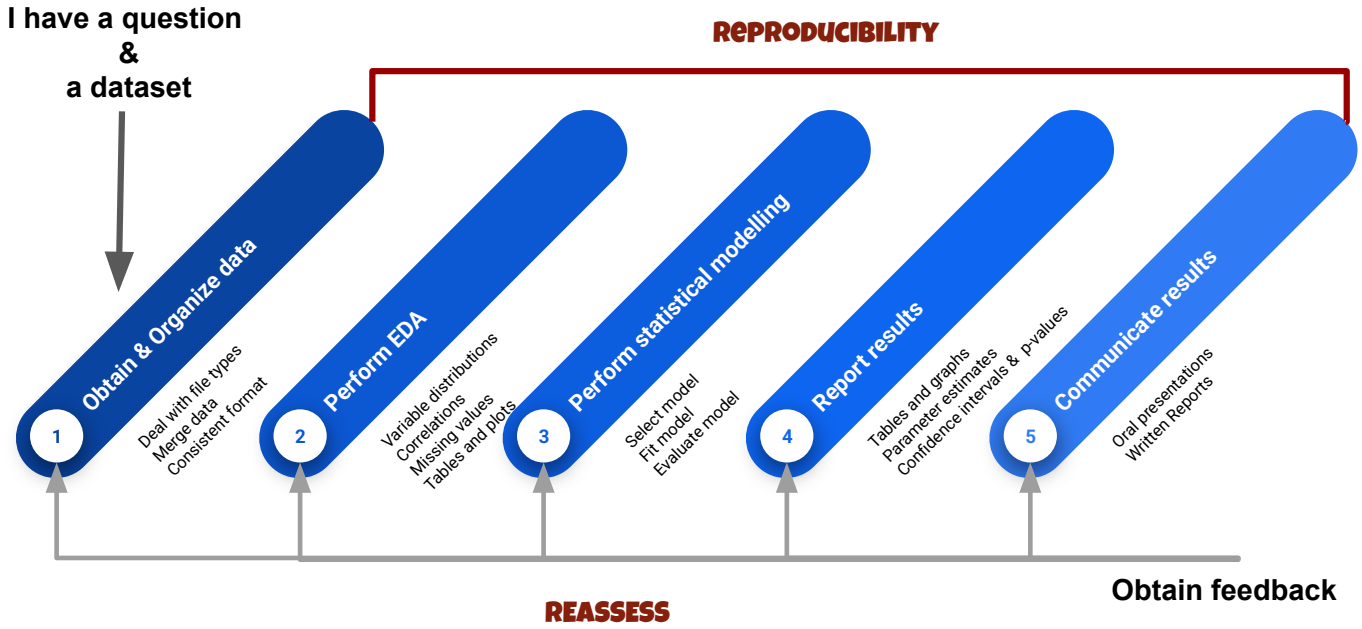
Outline

DMI

- What is exploratory data analysis (EDA)
- General steps for an EDA
- Methods for exploring Single Variables
- Methods for exploring Relationships Between Two Variables
- How to explore More Than Two Variables
- Hands-on session
- Bibliography

Recap from previous lecture

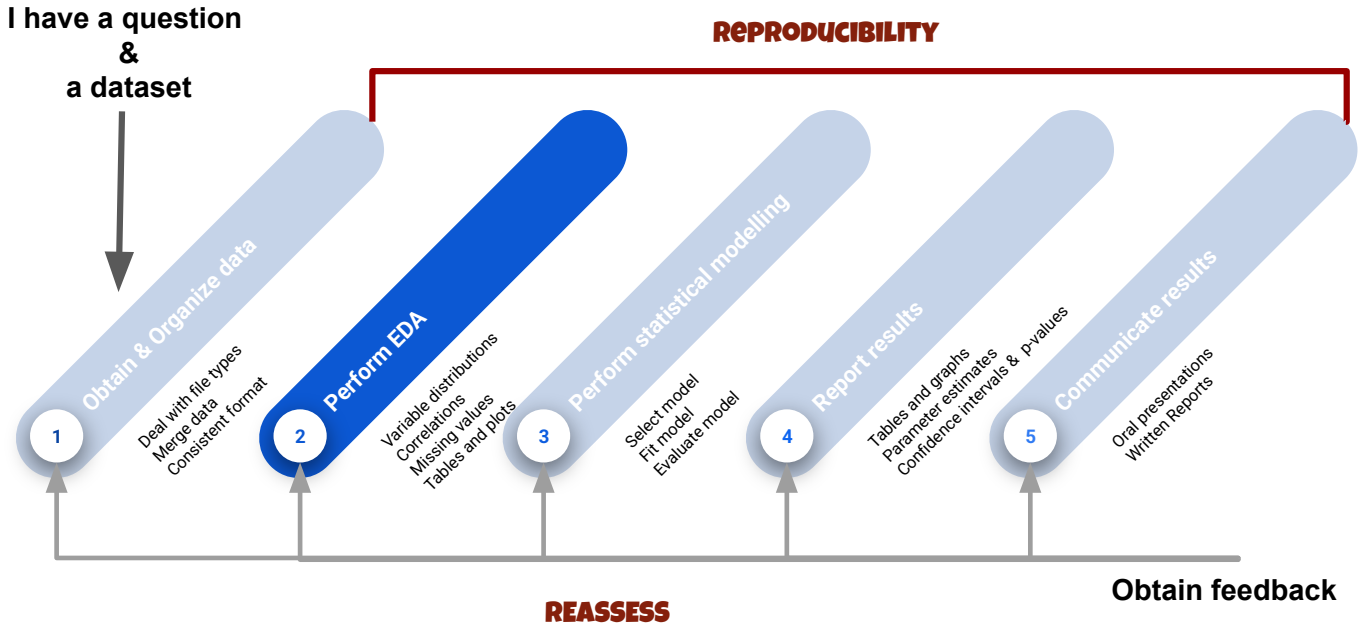
DMI



Modified from https://x.com/siminaboca/status/1298870717291917312

Recap from previous lecture

DMI



Modified from https://x.com/siminaboca/status/1298870717291917312

DMI

EDA: Getting to Know Your Data

What is EDA

DMI

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

What is EDA

DMI

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical).
- A **preliminary** exploration of the data to better understand its characteristics.

It facilitates discovering unexpected as well as conforming the expected.

Especially useful in early stages of data mining

Examples of questions to be asked during EDA

DMI

1. How many records does this dataset contain?
2. How many fields (i.e., variables) are included in each record?
3. What kinds of variables are these? (e.g., real numbers, integers, categorical…)
4. Are these variables always observed? (i.e., is missing data an issue? If so, how are missing values represented?)
5. Are the variables included in the dataset the ones we were expecting?
6. Are the values of these variables consistent with what we expect?
7. What type of variation occurs within my variables?
8. Do the variables in the dataset seem to exhibit the kinds of relationships we expect? (Indeed, what relationships do we expect, and why?)

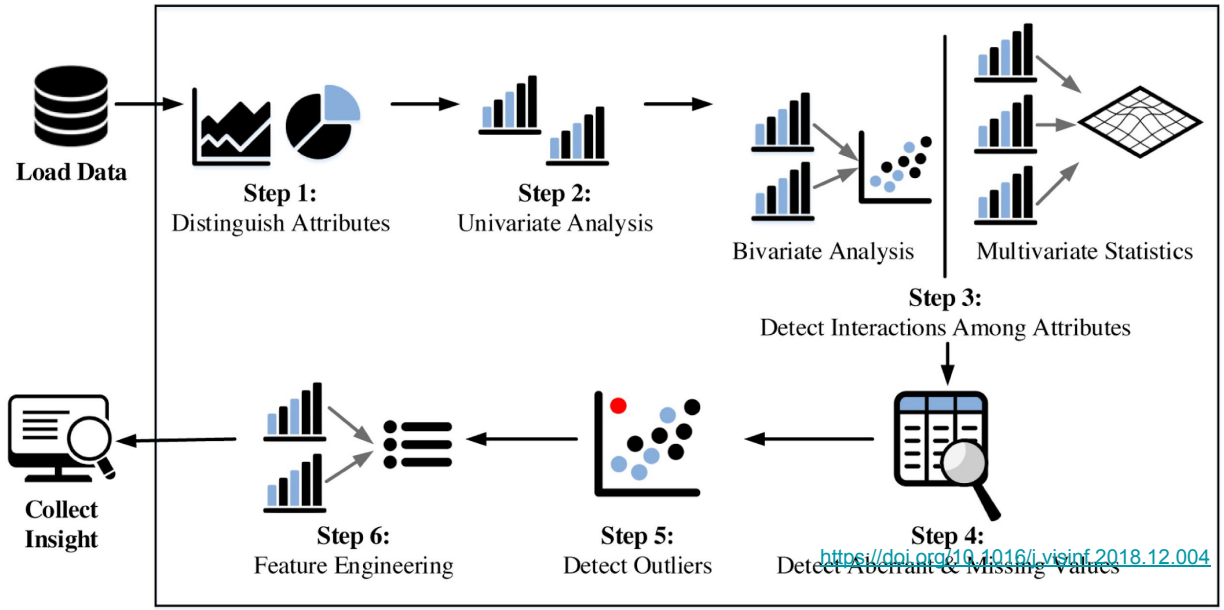
Why EDA?

DMI

- Assess data quality: missing values, coding problems
- Observe variable ranges, and their units
- Maximize insight into a dataset
- Uncover underlying structure / To see patterns in the data
- Extract important variables
- Detect outliers and anomalies / To catch mistakes
- Test underlying assumptions (e.g. normal distributions or skewed?) / To find violations of statistical assumptions
- Helps with the formulation and validation of hypotheses from the data
- Help to select the right tool for preprocessing or analysis
- Determine optimal factor settings

…and because if you don't, you will have problems later!!

The fundamental steps of the EDA



https://doi.org/10.1016/j.visinf.2018.12.004

Major Tasks in EDA

DMI

Data cleaning

- Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, or files

Data reduction

- Dimensionality reduction
- Numerosity reduction
- Data compression

Data transformation and data discretization

- Normalization
- Scaling

Data Cleaning

DMI

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., *Occupation* = " " (missing data)
- Noisy: containing noise, errors, or outliers
  - e.g., *Salary* = "−10" (an error)
- Inconsistent: containing discrepancies in codes or names, e.g.,
  - Was rating "1, 2, 3", now rating "A, B, C"
  - discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

DMI

Data is not always available

- E.g., many tuples have no recorded value for several attributes

Missing data may be due to

- Equipment malfunction
- Inconsistent with other recorded data and thus deleted
- Data were not entered due to misunderstanding
- Certain data may not be considered important at the time of entry
- Did not register history or changes of the data

Missing data may need to be inferred

Why might you choose to impute data? Most commonly, due to issues of power associated with reduced sample size

Types of Missingness

Missing Completely at Random (MCAR): there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others. The probability of being missing is the same for all cases.

Missing at Random (MAR): the probability of missingness in a variable depends only on available information (in other predictors). For example, in a registry examining depression, if male participants are less likely to complete a survey about depression severity than female participants. That is, if probability of completion of the survey is related to their sex (which is fully observed) but not the severity of their depression, then the data may be regarded as MAR

Missing Not at Random (MNAR): the probability of missingness depends on information that has not been recorded and this information also predicts the missing values. For example, in a depression registry, if participants with severe depression are more likely to refuse to complete the survey about depression severity.

Missing values in R

DMI

NULL:
- Is returned when an expression or function results in an undefined value
- Is a reserved word
- Can be the product of importing data with unknown data type.

NA:
- is a logical constant of length 1
- is an indicator for a missing value.
- is a reserved word
- can be coerced to any other data type vector
- Can be a product when importing data
- Stands for Not Available.

NaN
- stands for Not A Number
- is a logical vector of a length 1
- applies to numerical values, as well as real and imaginary parts of complex values, but not to values of integer vector.
- Is a reserved word
- Sometimes a computation will produce a result that makes little sense. Try Inf - Inf

Inf / -Inf
- stands for infinity (or negative infinity)
- is a result of storing either a large number or a product that is a result of division by zero.
- Is a reserved word
- is – in most cases – product of computations in R language and therefore very rarely a product of data import.
- It also tells you that the value is not missing and a number!
- Try 2 ^ 1024 or - 2 ^ 1024 or 1 / 0 in R

Remember: no data and 0 are not the same, 0 is a valid numerical value

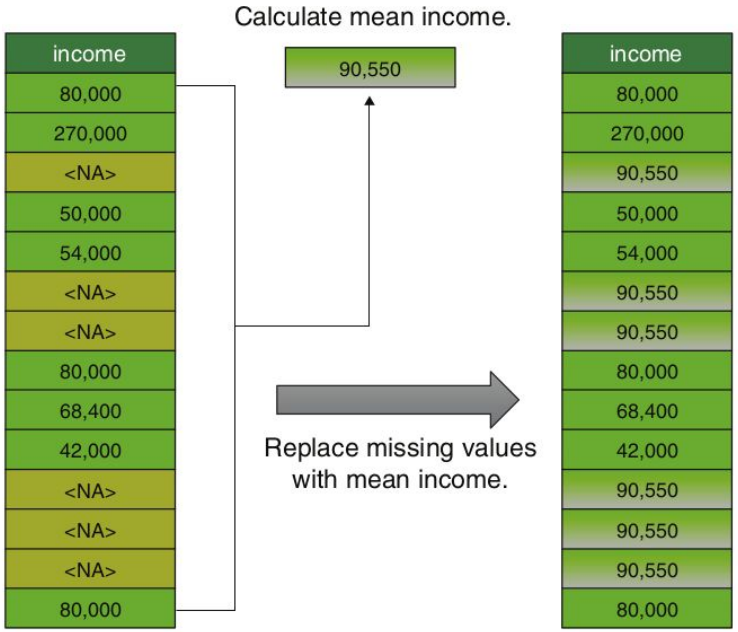
```
read.csv(file = "my_file.csv", na.strings = c("NA", " ", "-999", “NULL”))
```

How to Handle Missing Data?

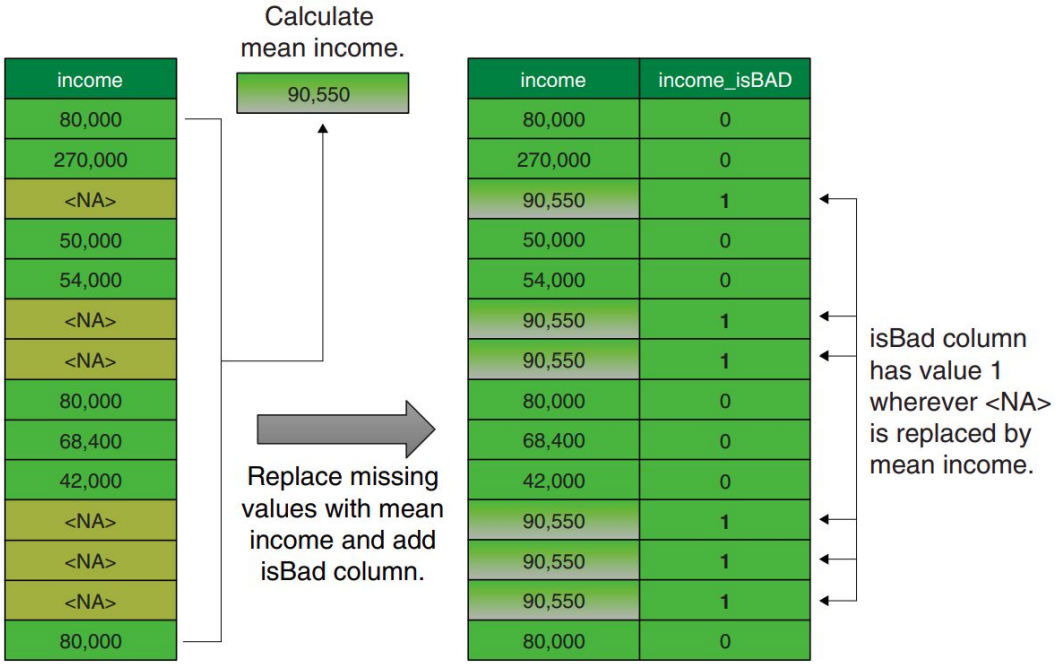
DMI

- Ignore the tuple:
- Fill in the missing value manually: tedious + infeasible?
  - usually done when class label is missing (when doing classification)
  - not effective when the % of missing values per attribute varies considerably
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean (**MCAR**)
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree, or with the value of the nearest neighbor

How to Handle Missing Data?



How to Handle Missing Data?



Missing values in R

DMI

Imputation is the process of replacing a missing value with a substituted, “best guess” value.

Some R Packages used for imputing missing values

mice: (Multivariate Imputation via Chained Equations) is one of the commonly used package by R users.

Amelia II: Uses the expectation-maximization algorithm with bootstrapping

missForest: implementation of random forest algorithm.

Hmisc: multiple purpose package useful for data analysis. Uses the simplest method for missing data imputation is imputation by mean (or median, mode, ...).

MI: Multiple imputation with diagnostics. It uses predictive mean matching method.

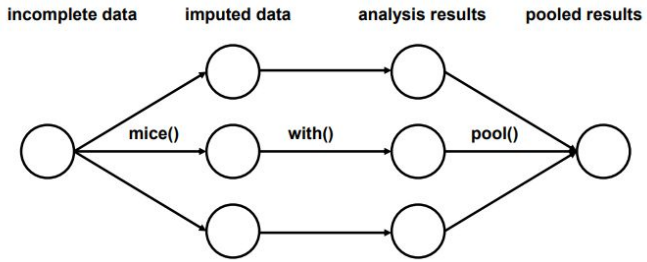
VIM: (Visualization and Imputation of Missings) implements several methods, such as kNN

… and many more!

Mice R package

DMI

1. Method: multiple imputations for a missing value that accounts for the statistical uncertainty in the imputation
2. Assumptions: the missing data is MAR. MAR occurs when a data gap is full accounted for by variables where there is complete information. *It can also work on data that is MNAR.*
3. Iterations: Multiple regression models are conducted and each variable with missing data is modeled conditionally on the responses of the other variables within the dataset. With this method, each variable is modeled according to its own distribution



Flexible Imputation of Missing Data

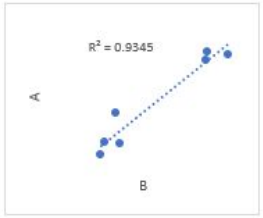
We should be suspicious of any dataset (large or small) which appears perfect. — David J. Hand

Predictive mean matching

DMI

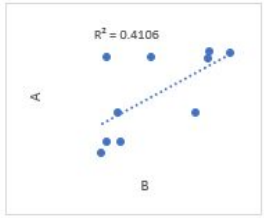
Missing data is in red. There is a strong correlation between A and B, so let's try to impute A using B and C.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
	0.80	
0.95	1.24	1.46
0.23	0.57	
0.90		1.28
0.15	0.42	
0.47	0.54	0.63
	1.14	
0.89	1.23	1.45



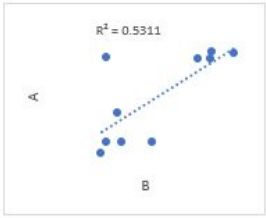
Missing data is filled in randomly. This dillutes the correlations, but allows us to impute using all available data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.90	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.47	1.14	1.28
0.89	1.23	1.45



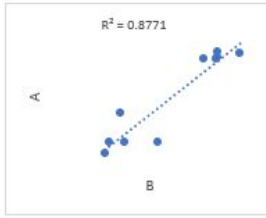
A random forest is used to predict A with B and C. Notice the correlation between A and B improved.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	0.46	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45



After Imputing B using A and C, we have achieved a correlation between A and B much closer to the original data.

A	B	C
0.93	1.40	1.53
0.24	0.46	0.76
0.24	0.80	1.53
0.95	1.24	1.46
0.23	0.57	1.28
0.90	1.24	1.28
0.15	0.42	1.53
0.47	0.54	0.63
0.89	1.14	1.28
0.89	1.23	1.45



https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html

Predictive mean matching

The predicted value of A (E[A|B,C]) is shown to the left. We are interested in imputing the bold missing value below

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

Our predicted value for the first missing sample is 0.60. The closest predicted value is 0.62. We find the closest values for all of our missing samples.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60		0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20		1.14	1.28
1.24	0.89	1.23	1.45

We then impute the value corresponding to the original data.

E[A|B,C]	A	B	C
0.73	0.93	1.40	1.53
0.62	0.24	0.46	0.76
0.60	0.24	0.80	1.53
1.39	0.95	1.24	1.46
0.36	0.23	0.57	1.28
1.27	0.90	0.46	1.28
0.15	0.15	0.42	1.53
0.65	0.47	0.54	0.63
1.20	0.89	1.14	1.28
1.24	0.89	1.23	1.45

https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html

Outliers

DMI

- "An outlier is an observation that **deviates so much** from the other observations as to arouse suspicions that it was generated by a different mechanism." Hawkings (https://www.springer.com/gp/book/9789401539968)
- An outlying observation, or 'outlier', is one that appears to **deviate markedly** from other members of the sample in which it occurs. —F.E.Grubbs, Procedures for detecting outlying observations in samples (1969)
- We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be **inconsistent with the remainder of that set of data**. —V.Barnett and T.Lewis Outliers in Statistical Data (1978)

Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature.

Example: Hadlum vs. Hadlum (1949)

DMI

- This case was analyzed in Barnett, 1978.
- The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.
- Average human gestation period is 280 days (40 weeks).
- Statistically, 349 days is an outlier.
- Very low probability for the birth of Mrs. Hadlum's child for being generated by this nominal process.
- The court ruled that the observation was valid, if extreme

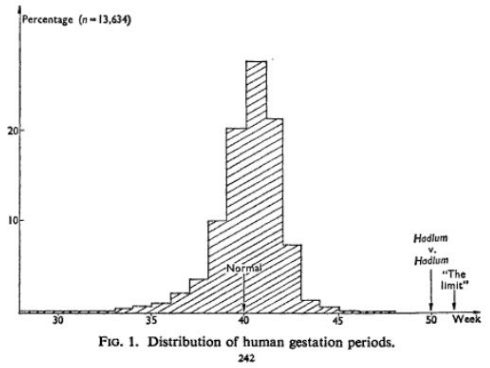


FIG. 1. Distribution of human gestation periods.
242

What causes outliers?

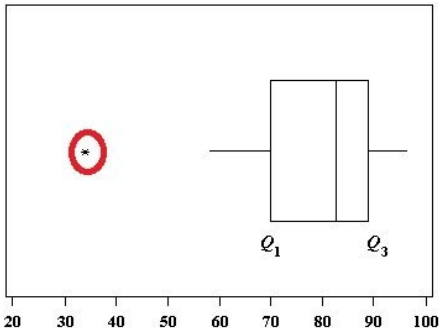
- Human errors, e.g. data entry errors
- Instrument errors, e.g. measurement errors
- Data processing errors, e.g. data manipulation
- Sampling errors, e.g. extracting data from wrong sources
- Not an error, the value is extreme, just a ‘novelty’ in the data

A dilemma

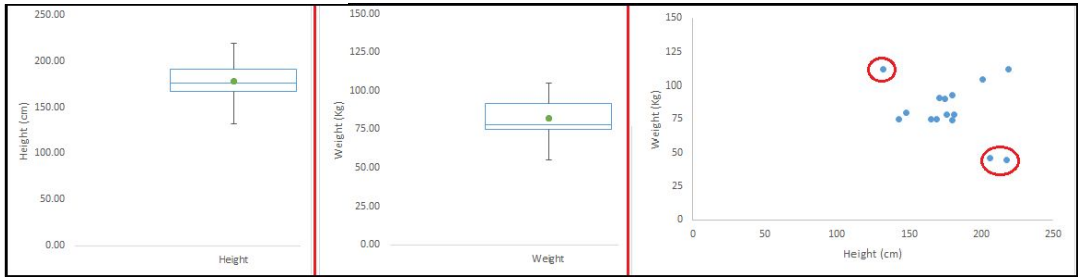
DMI

- Outliers can be genuine values
- The trade-off is between the loss of accuracy if we throw away “good” observations, and the bias of our estimates if we keep “bad” ones
- The challenge is twofold:
    1. to figure out whether an extreme value is good (genuine) or bad (error)
    2. to assess its impact on the statistics of interest

Outliers



Univariate or Multivariate

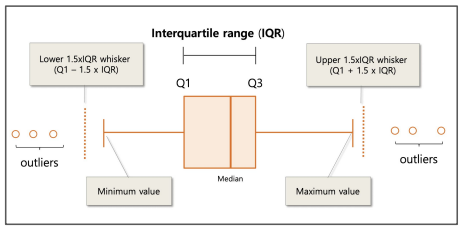


Outliers: detection Algorithms

Global Anomaly detection

1. Statistical approaches:
   - Ex, extreme values: A data point is an extreme value,

if it lies at one of the two ends of a probability distribution.



IQR criterion: Observations outside $I = [q_{0.25} - 1.5 \cdot IQR; q_{0.75} + 1.5 \cdot IQR]$

percentiles method: Observations outside the interval formed by the 2.5 and 97.5 percentiles (or 1 and 99, or the 5 and 95)

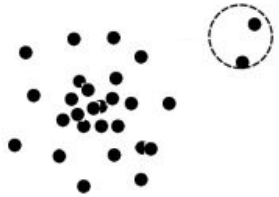
Hampel filter: Observations outside the interval formed by the median, plus or minus 3 median absolute deviations

Outliers: detection Algorithms

DMI

Global Anomaly detection

1. Statistical approaches:
   - Ex, extreme values: A data point is an extreme value, if it lies at one of the two ends of a probability distribution.
2. Clustering models: Many clustering models determine outliers as a side-product of the algorithm. It is also possible to optimize clustering models to specifically detect outliers.
3. Classification based approaches.



Outliers: detection Algorithms

DMI

Local Anomaly detection

4. Distance-based models: In these cases, the k-nearest neighbor distribution of a data point is analyzed to determine whether it is an outlier.

5. Density-based models: In these models, the local density of a data point is used to define its outlier score. Ex, Local Outlier Factor (LOF), Density-based spatial clustering of applications with noise (DBSCAN)

Outliers: in R

DMI

Using the boxplot function in R base: boxplot.stats(dataset$chol)

The outliers package provides a number of useful functions to systematically extract outliers.

The functions of most use are outlier() and scores().

The outlier() function gets the most extreme observation from the mean.

The scores() function computes the normalized score which you can use to find observation(s) that lie beyond a given value.

Dealing with outliers

For analyses you can:

- Identification
- Inclusion
- Rejection: Delete the value – crude but effective
- Accomodations:
  - Change the outlier to value ~3 SD from mean
  - Make equal to the next highest value
  - “Trim” the mean – recalculate mean from data within interquartile range
  - Impute

Classification of EDA

DMI

- Each method is either non-graphical or graphical.
  - Non-graphical methods generally involve calculation of summary statistics,
  - Graphical methods obviously summarize the data in a diagrammatic or pictorial way.
- Each method is either univariate or multivariate (usually just bivariate)
  - Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships.
  - Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables.
  - It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

DMI

Univariate EDA

Univariate EDA

DMI

Univariate analysis with numerical variables

- Measures of central tendency: Mean, Median,
- Measures of dispersion: Min, Max, Range, Quartiles, Variance, Standard deviation
- Other measures include: Skewness, Kurtosis
- Tables can be used with binning of the data

Univariate Graphical

Histogram

Kernel density plots

Box plots

Bar plots with binned data

Univariate analysis with categorical variables

- Measures of central tendency: Mode
- Frequency, percentage
- **A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.**

Univariate Graphical

Bar plots (Frequency, percentage)

Pie charts 😱

Tree maps

waffle chart

Univariate EDA: summary statistics

DMI

- Summary statistics are numbers that summarize properties of the data

  – Summarized properties include frequency, location and spread

    Examples: centrality - mean
    spread - standard deviation

  – Most summary statistics can be calculated in a single pass through the data

    Example summary()

Univariate EDA: summary statistics

In R, you'll typically use the summary() command to take your first look at the data.

Variable	Description
studyid	Study ID
treatment	Treatment group
age	Age in years
sex	Male or Female
nihss	NIH Stroke Scale Score (higher indicate more severe deficits)
location	Stroke Location - Left or Right Hemisphere
hx.isch	History of Ischemic Stroke (Yes/No)
afib	Atrial Fibrillation (1 = Yes, 0 = No)
dm	Diabetes Mellitus (1 = Yes, 0 = No)
sbp	Systolic blood pressure, in mm Hg
iv.altep	Treatment with IV alteplase (Yes/No)
time.iv	Time from stroke onset to start of IV alteplase (minutes) if iv.altep=Yes
ia.occlus	Intracranial arterial occlusion, based on vessel imaging - five categories1
extra.ica	Extracranial ICA occlusion (1 = Yes, 0 = No)
time.rand	Time from stroke onset to study randomization, in minutes
time.punc	Time from stroke onset to groin puncture, in minutes (only if Intervention)

```
> summary(dat)
   studyid                age            sex          nihss        location
 Length:500         Min.   :23.00   Female:208   Min.   :10.00   Left :269
 Class :character   1st Qu.:55.00   Male  :292   1st Qu.:14.00   Right:231
 Mode  :character   Median :65.75                Median :18.00
                    Mean   :64.71                Mean   :18.03
                    3rd Qu.:76.00                3rd Qu.:22.00
                    Max.   :96.00                Max.   :28.00

   hx.isch              afib             dm             sbp
 Length:500         Min.   :0.00   Min.   :0.000   Min.   : 78.0
 Class :character   1st Qu.:0.00   1st Qu.:0.000   1st Qu.:128.5
 Mode  :character   Median :0.00   Median :0.000   Median :145.0
                    Mean   :0.27   Mean   :0.126   Mean   :145.5
                    3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:162.5
                    Max.   :1.00   Max.   :1.000   Max.   :231.0
                                                   NA's   :1
   iv.altep            time.iv                  ia.occlus     extra.ica
 Length:500         Min.   : 42.00   A1 or A2        :  3   Min.   :0.0000
 Class :character   1st Qu.: 67.00   ICA with M1     :134   1st Qu.:0.0000
 Mode  :character   Median : 86.00   Intracranial ICA:  4   Median :0.0000
                    Mean   : 92.64   M1              :319   Mean   :0.2906
                    3rd Qu.:115.00   M2              : 39   3rd Qu.:1.0000
                    Max.   :218.00   NA's            :  1   Max.   :1.0000
                    NA's   :55                              NA's   :1
   time.rand       time.punc          treatment
 Min.   :100.0   Min.   :180   Control     :267
 1st Qu.:151.2   1st Qu.:212   Intervention:233
 Median :201.5   Median :260
 Mean   :208.6   Mean   :263
 3rd Qu.:257.8   3rd Qu.:313
 Max.   :360.0   Max.   :360
 NA's   :2       NA's   :267
```

Missing values

Univariate EDA: summary statistics, centrality

DMI

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Univariate EDA: summary statistics, dispersion

DMI

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of poin

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- However, this is also sensitive to outliers, so that other measures are often used.

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$ **average absolute deviation**

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

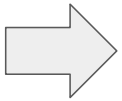
Univariate EDA: Tables for Continuous Data

Example: Ages of 10 adult leukemia patients:

35; 40; 52; 27; 31; 42; 43; 28; 50; 35

One option is to group these ages into decades and create a categorical age variable:

Age	Age group
35	31-40
40	31-40
52	51-60
27	21-30
31	31-40
42	41-50
43	41-50
28	21-30
50	41-50
35	31-40



We can then create a frequency table for this new categorical age variable.

Interval	Freq.	Rel. Freq.
21-30	2	$2/10 = 0.2 = 20\%$
31-40	4	$4/10 = 0.4 = 40\%$
41-50	3	$3/10 = 0.3 = 30\%$
51-60	1	$1/10 = 0.1 = 10\%$
Total	10	$10/10 = 1.0 = 100\%$

Univariate EDA

Graphic Displays of Basic Statistical Descriptions

Box plot: graphic display of five-number summary: `boxplot()`

Histogram: x-axis are values, y-axis represent frequencies: `hist()`

Quantile-quantile (Q-Q) plot: graphs the quantiles of one distribution against the corresponding quantiles of another. `qqnorm()`: produces a normal QQ plot of the variable

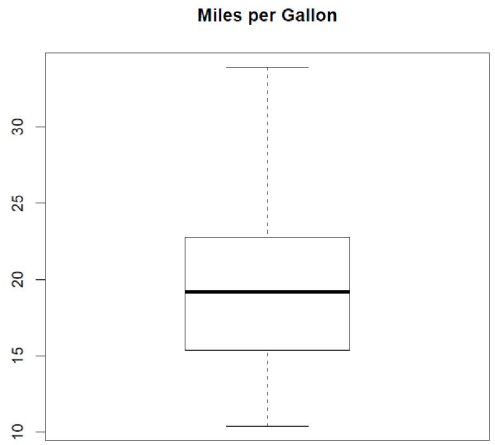
Barplot: Barplots are useful for visualizing categorical data, with the number of entries for each category being proportional to the height of the bar. `barplot()`

Density plot: The density function computes a non-parametric estimate of the distribution of a variables `density()`

Univariate EDA: box plots

DMI

boxplot(mtcars$mpg, main = "Miles per Gallon")



The five-number summary

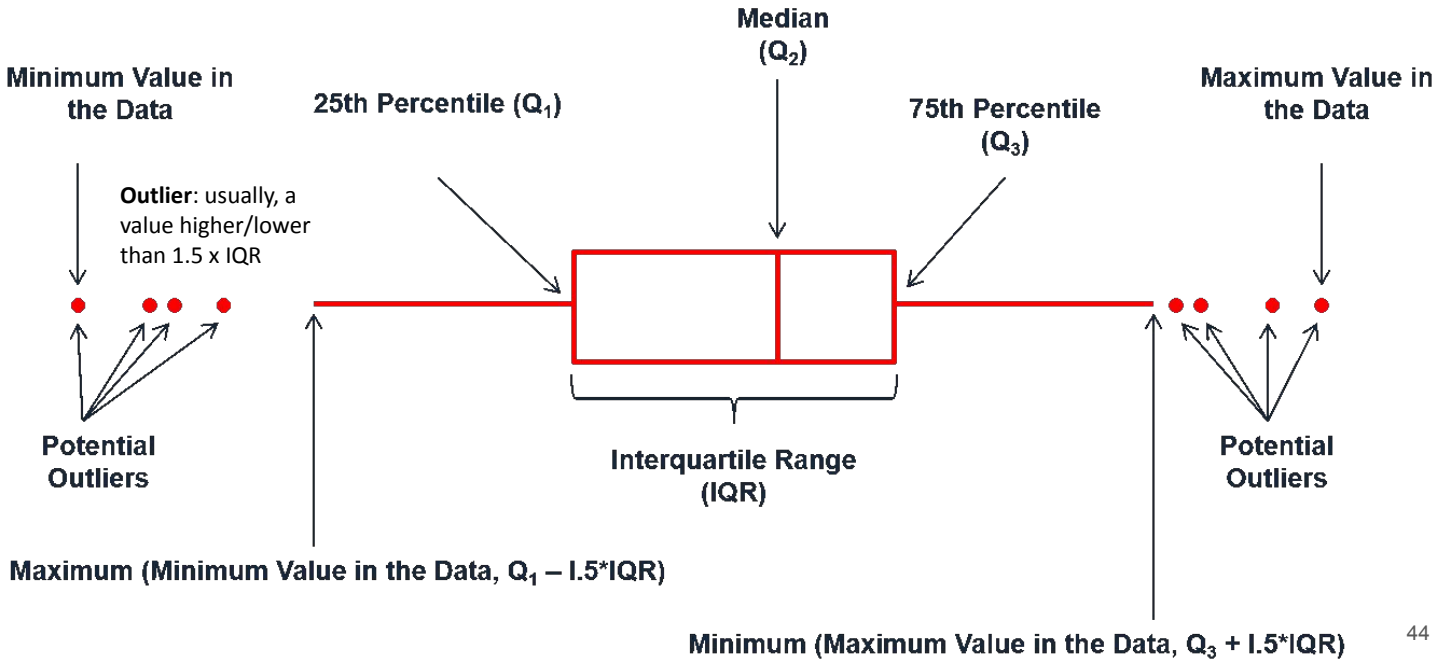
1. the sample minimum (smallest observation)
2. the lower quartile or *first quartile*
3. the median (the middle value)
4. the upper quartile or *third quartile*
5. the sample maximum (largest observation)

Tukey, 1970's

Univariate EDA: box plots

DMI

Assumption: Data is generated by a Gaussian process. Therefore it is normally distributed about the mean.



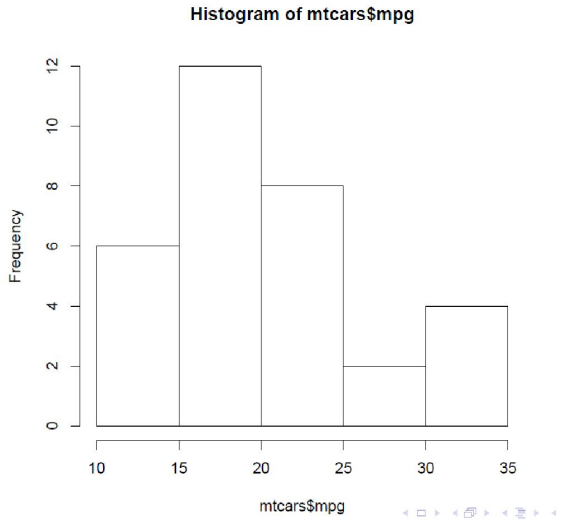
Univariate EDA: histograms

DMI

For continuous data, plots the distribution of values

hist(mtcars$mpg)

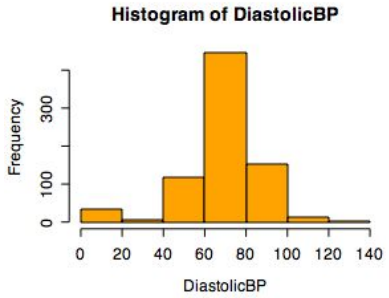
- Usually shows the distribution of values of a single variable
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height (area) of each bar indicates the number of objects
- Shape of histogram depends on the number of bins



Univariate EDA: histograms

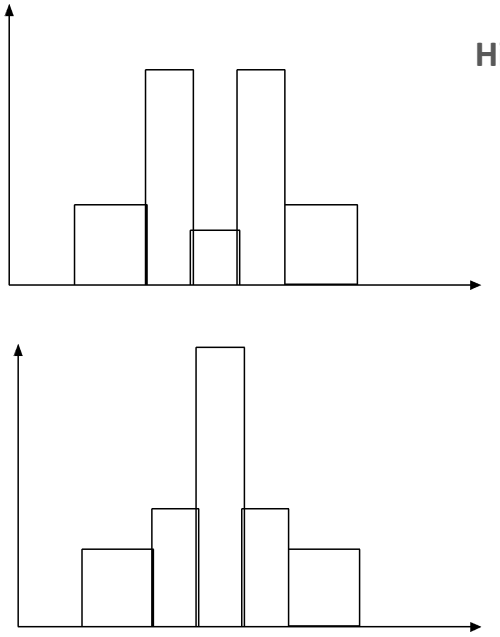
DMI

- Histogram:
  - Shows center, variability, skewness, modality,
  - outliers, or strange patterns.
  - Bin width and position matter



Univariate EDA: histograms

DMI

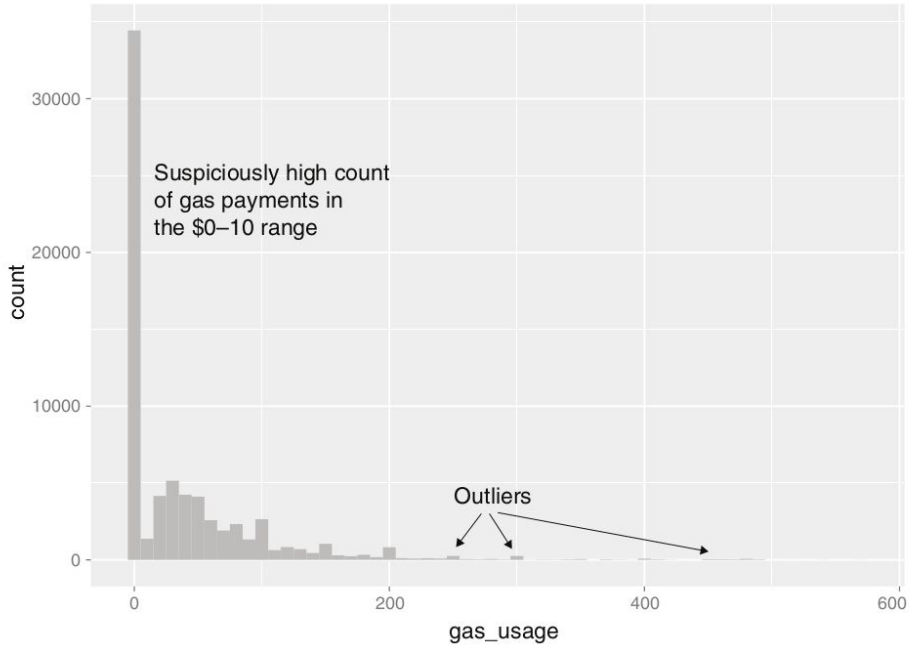


Histograms Often Tell More than Boxplots

- The two histograms shown in the left may have the same box plot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Univariate EDA: histograms

DMI



Univariate EDA: histograms

DMI

Issues with Histograms

- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - But ‘small multiples’ can be effective

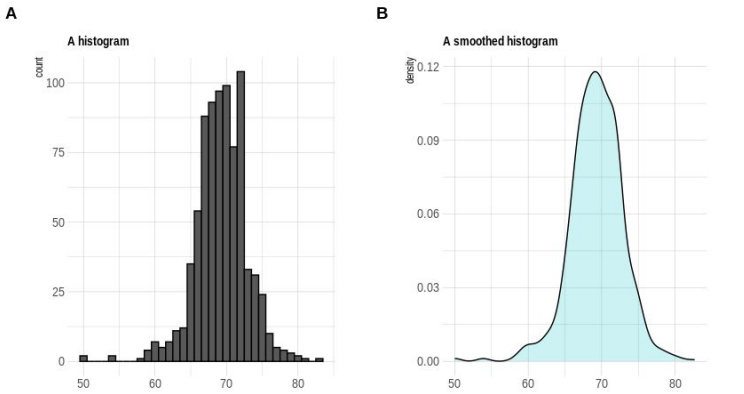
Univariate EDA: density plots

DMI

Smoothed Histograms

- To visualize the underlying probability distribution of the data by drawing an appropriate continuous curve
- The curve needs to be estimated from the data, and the most commonly used method for this estimation procedure is called kernel density estimation.
- The most widely used kernel is a Gaussian kernel (i.e., a Gaussian bell curve), but there are many other choices.

Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.



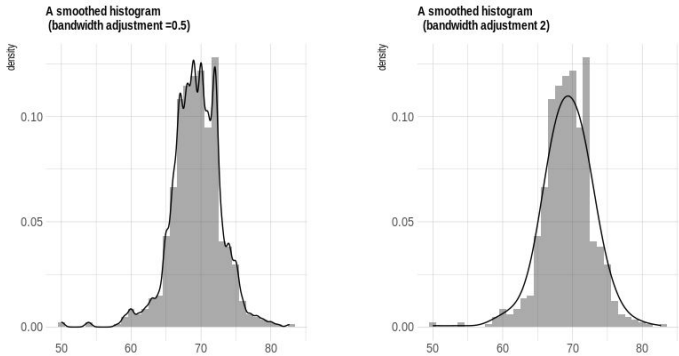
You can think of a density plot as a continuous histogram of a variable, except the area under the density plot is rescaled to equal one.

Univariate EDA: density plots

DMI

Smoothed Histograms

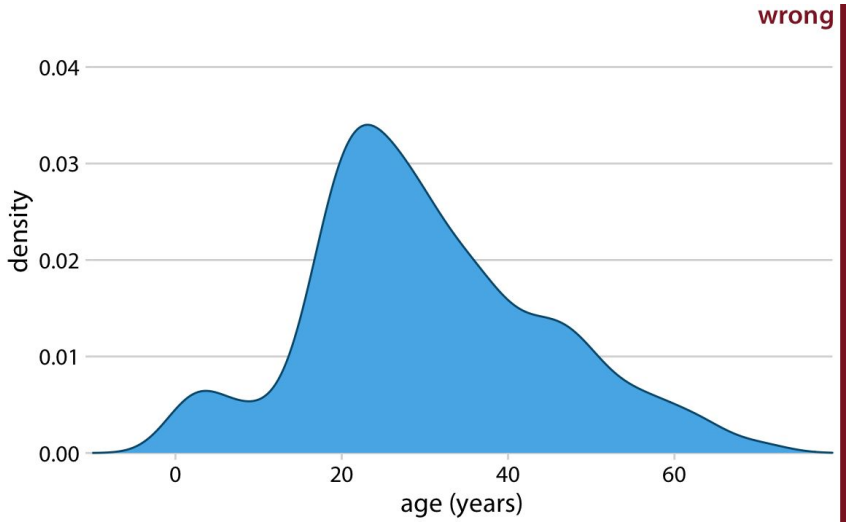
Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.



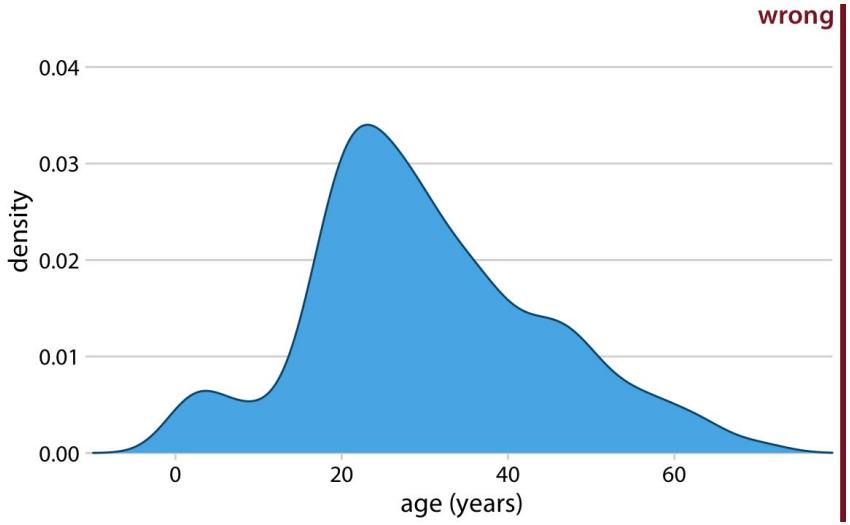
➔ Bandwidth choice is an art

➔ Usually want to try several

Univariate EDA: density plots



Univariate EDA: density plots



Always verify that your density estimate does not predict the existence of nonsensical data values.

Univariate EDA: Q-Q plots

DMI

A Q-Q (quantile-quantile) plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

- Two data sets come from populations with a common distribution.
- **If a data set follows a particular distribution**

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight

By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

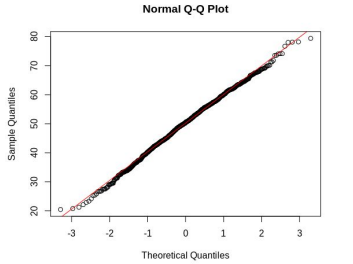
Univariate EDA: Q-Q plots

Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

In R, there are two functions to create Q-Q plots: **qqnorm** and **qqplot**.

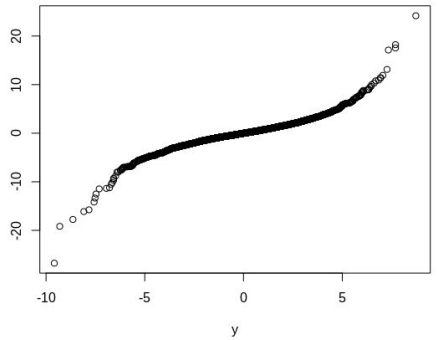
qqnorm creates a Normal Q-Q plot

```
x <- rnorm(1000, 50, 10)
qqnorm(x)
qqline(x, col = "red")
```

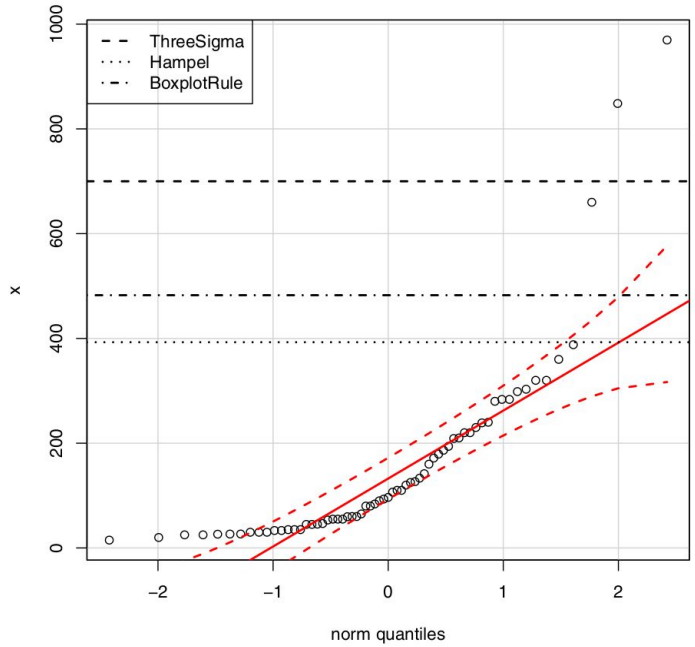


qqplot creates a Q-Q plot for any distribution.

```
y <- rlogis(10000)# Random
values according to logistic
distribution
z <- rt(10000, 3)      # Random
values according to student t
distribution
qplot(y, z)
```



Q-Q plots & outliers

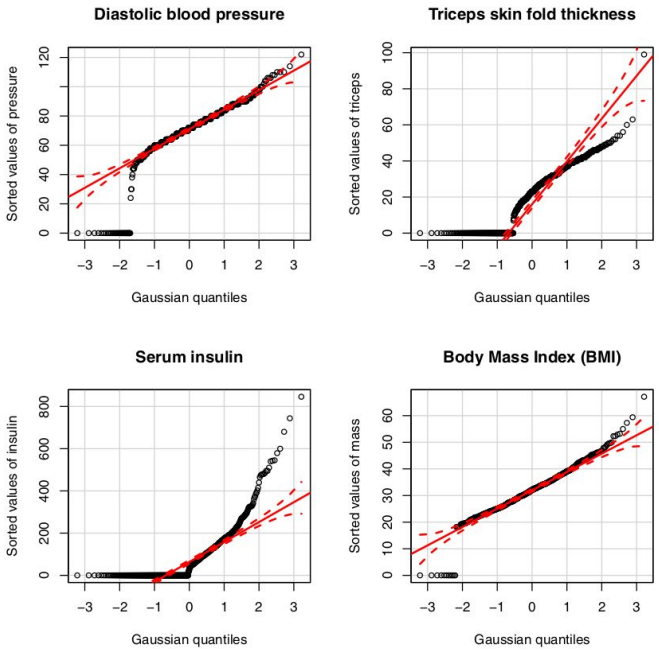


Three Sigma: for normally distributed data, almost all observed data will fall within three standard deviations

Hampel: Observations outside the interval formed by the median, plus or minus 3 median absolute deviations

Boxplot: Values below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered potential outliers.

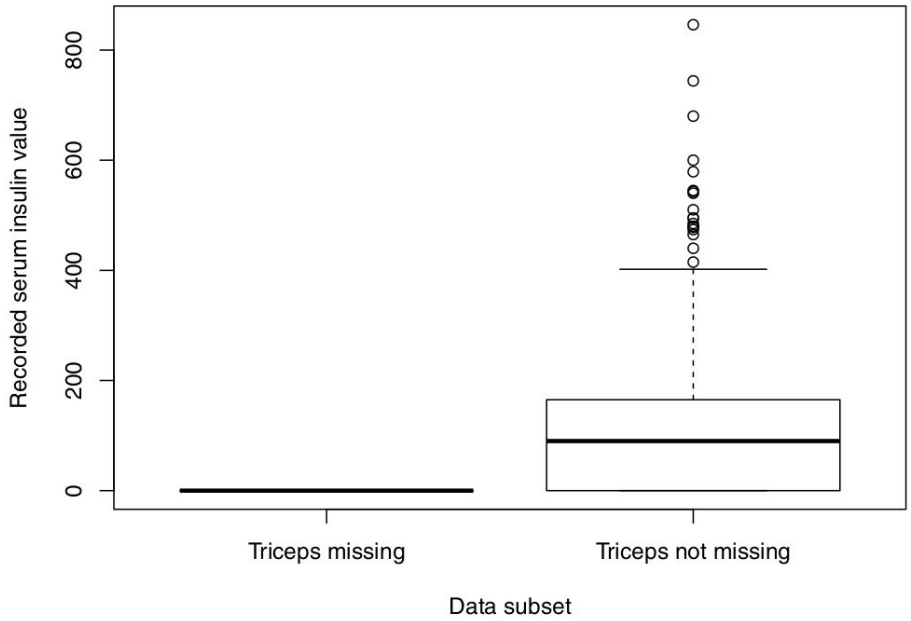
Q-Q plots & missing data



Pima Indians diabetes dataset

Missing data revisited

DMI



Univariate EDA: categorical data

DMI

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of a an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Univariate EDA: categorical data

DMI

- Frequency Table: Categories with counts
- Relative Frequency Table: Percentage in each category

Cancer by Site, 2001

	Colon	Breast	Prostate	Lung	Urinary	Total
Freq.	135,400	193,700	198,100	169,500	87,500	**784,200**
Relative Freq.	17.26%	24.70%	25.26%	21.61%	11.16%	100%

Frequency Table: Categories with counts

Relative Frequency Table: Percentage in each category, e.g.,

$$17.26\% = \frac{135400}{784200} \times 100$$

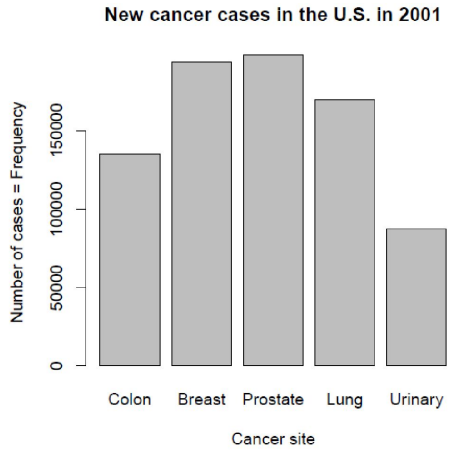
Note: percentages sum to 100%.

Univariate EDA: categorical data

DMI

Graphing a Frequency Table - Bar Charts

Plot the number of observations in each category:



Differences between histograms and bar charts

- Histograms are used to show distributions of variables while bar charts are used to compare variables
- Histograms plot binned quantitative data while bar charts plot categorical data
- Bars can be reordered in bar charts but not in histograms
- Histograms differ from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts

DMI

Bivariate EDA

Bivariate EDA

DMI

Bivariate Analysis finds out the relationship between two variables.

Look for association between variables

Perform bivariate analysis for any combination of categorical and continuous variables.

The combination can be:

- Categorical & Categorical
- Categorical & Continuous
- Continuous & Continuous

Bivariate EDA

DMI

Bivariate analysis with numerical variables

Correlation: Correlation is a bivariate analysis that measures the extent that two variables are related (“correlated”) to one another. The value of the correlation coefficient varies between +1 and -1.

Bivariate analysis with categorical variables

Cross-tabulation: frequency tables

Advanced techniques include:

Cluster analysis

Principal component analysis (PCA)

Bivariate Graphical

Scatterplot

Box plot

Bar plots

Bivariate EDA: Categorical & Categorical

DMI

Two-way table

```
> table(mtcars$cyl, mtcars$am)

    0  1
4   3  8
6   4  3
8  12  2
> table(cylinders = mtcars$cyl, manual= mtcars$am)
         manual
cylinders  0  1
        4  3  8
        6  4  3
        8 12  2
```

mtcars dataset: The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Bivariate EDA: Categorical & Categorical

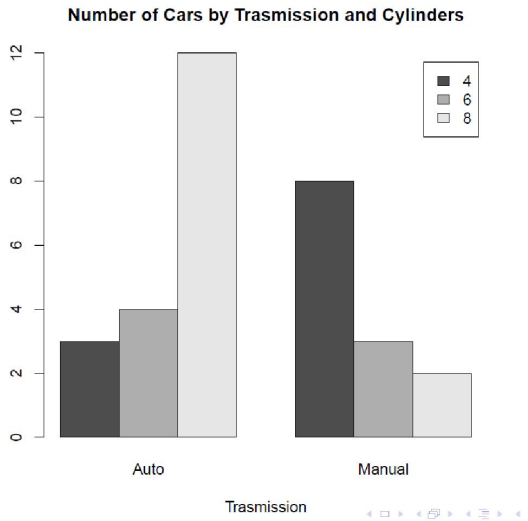
DMI

Barplots

Use the table function to create a two-way frequency table, and

plotting options to group bars

```
counts <- table(mtcars$cyl, mtcars$am)
colnames(counts) <- c("Auto", "Manual")
barplot(counts,
    + main = "Number of Cars by Transmission and Cylinders",
    + xlab = "Transmission",
    + beside = TRUE,
    + legend = rownames(counts))
```

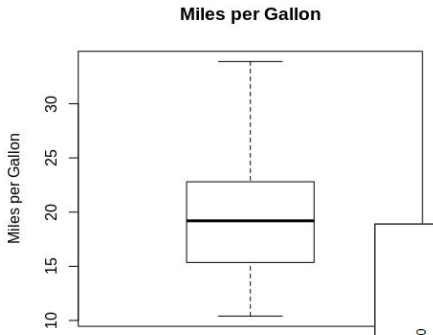


Bivariate EDA: Categorical & Continuous

DMI

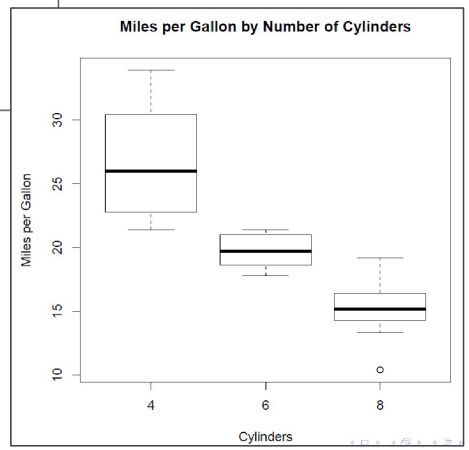
Box plots

```
boxplot(mtcars$mpg,
 main = "Miles per Gallon", ylab="mpg")
```



The boxplot function can also take a formula as an argument mpg conditional on cyl

```
boxplot(mpg ~ cyl,
  + data = mtcars,
  + main = "Miles per Gallon by Number of Cylinders",
  + xlab = "Cylinders",
  + ylab = "Miles per Gallon")
```



Bivariate EDA: Categorical & Continuous

DMI

Box plots

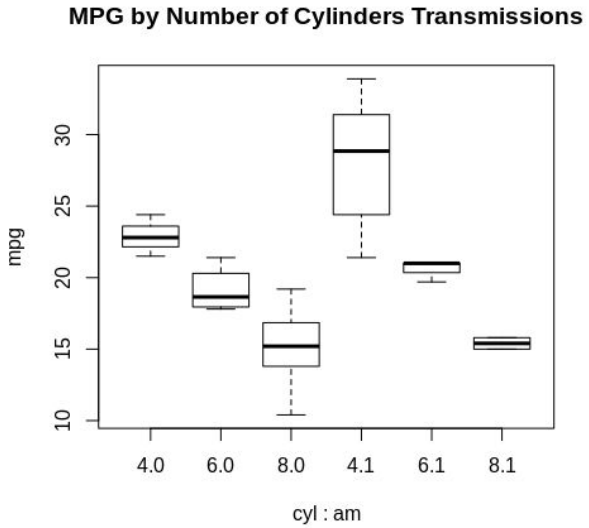
More than 2 variables!
(2 categorical, 1 continuous)

```
# Expand the formula
boxplot(mpg ~ cyl + am,
+ data = mtcars,
+ main = "MPG by Number of Cylinders &
          Transmissions")
```

mpg: Miles/gallon

cyl: Number of cylinders (4,6,8)

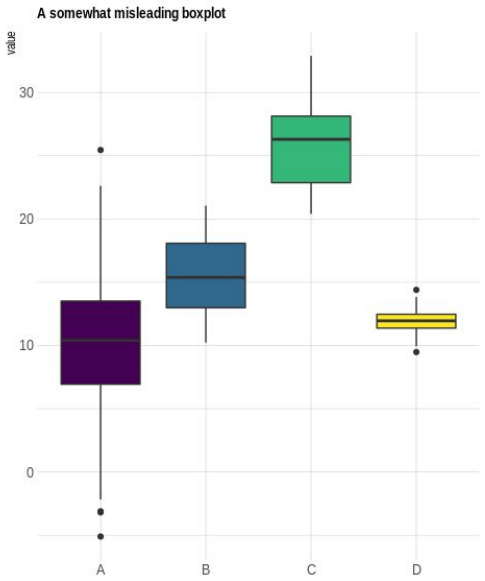
am: Transmission (0 = automatic, 1 = manual)



Bivariate EDA: Categorical & Continuous

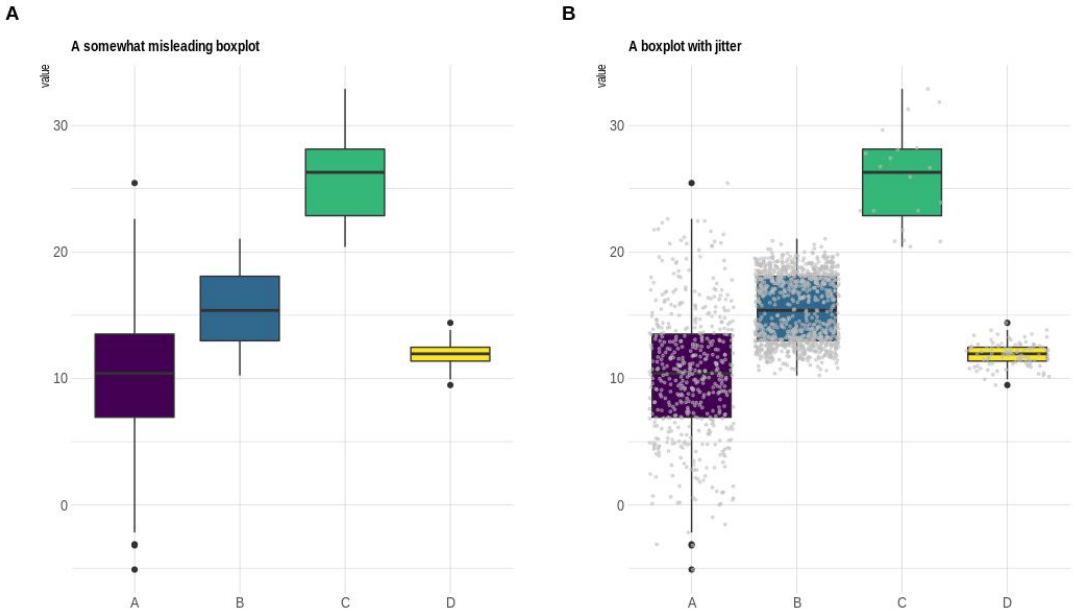
DMI

A



Bivariate EDA: Categorical & Continuous

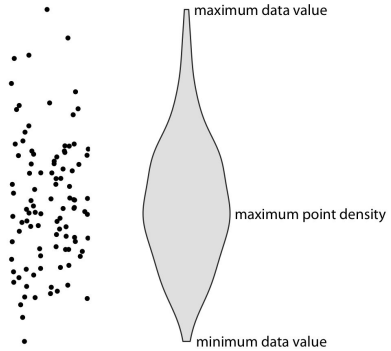
DMI



Don't use boxplots for small numbers of observations, just plot the data!

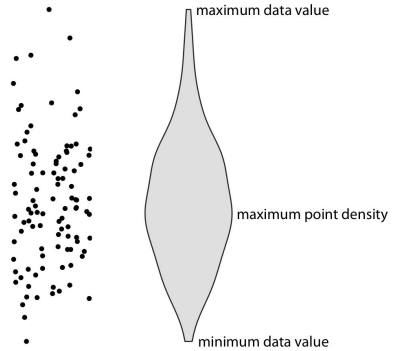
Violin plots

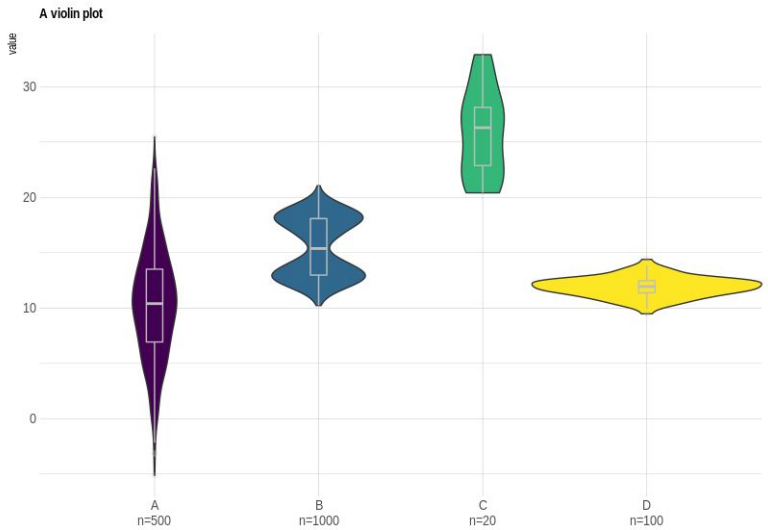
DMI



Before using violins to visualize distributions, verify that you have sufficiently many data points in each group to justify showing the point densities

Violin plots



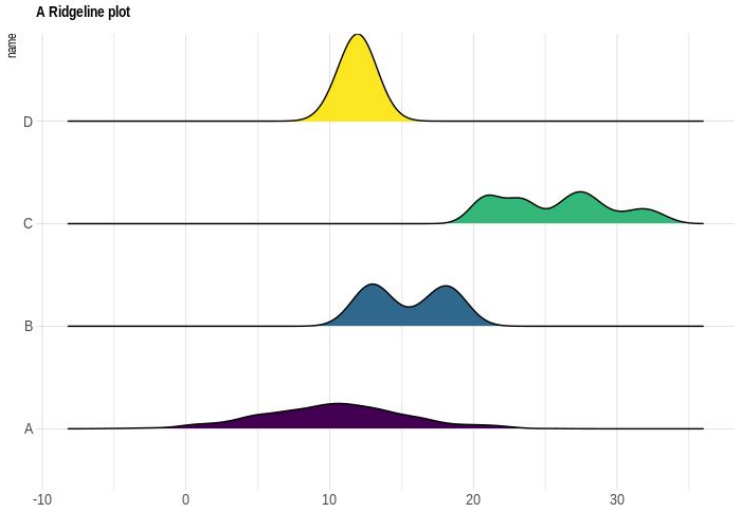


Before using violins to visualize distributions, verify that you have sufficiently many data points in each group to justify showing the point densities

Ridgeline plots

DMI

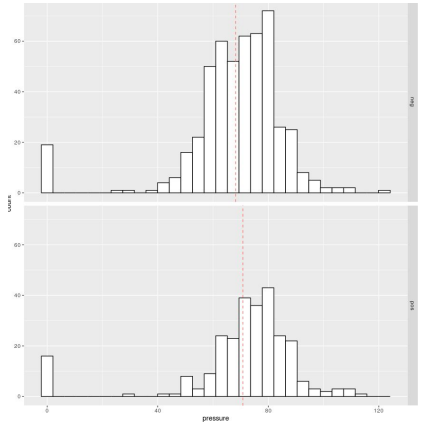
- Ridgeline plots make sense when the number of group to represent is medium to high, and thus a classic window separation would take to much space. Indeed, the fact that groups overlap each other allows to use space more efficiently. If you have less than ~6 groups, dealing with other distribution plots is probably better.
- It works well when there is a clear pattern in the result, like if there is an obvious ranking in groups. Otherwise group will tend to overlap each other, leading to a messy plot not providing any insight.

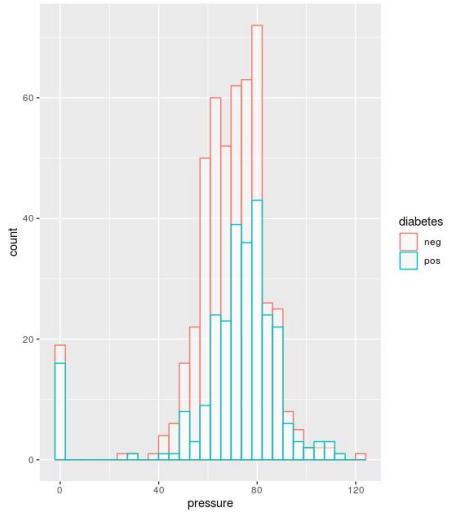


Bivariate EDA: Categorical & Continuous

- If one variable is categorical, use small multiple histograms

```
ggplot(PimaIndiansDiabetes, aes(x=pressure, color=diabetes)) +
  geom_histogram(fill="white", alpha=0.5, position="identity")
```





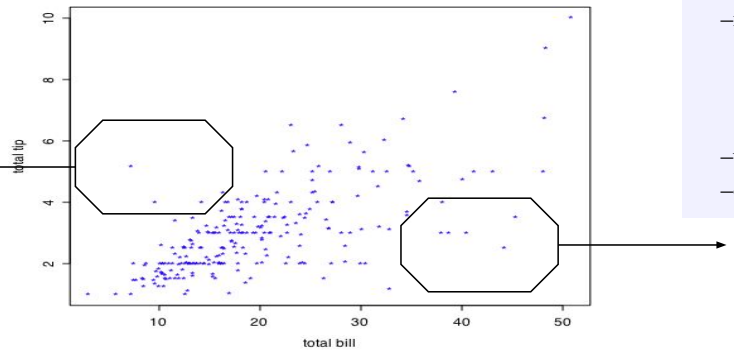
Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

- Attribute values determine the position
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- It is useful to have arrays of scatter plots to compactly summarize the relationships of several pairs of attributes

interesting?



useful to answer:
–x,y related?
•linear
•quadratic
•other
–variance(y) depend on x?
–outliers present?

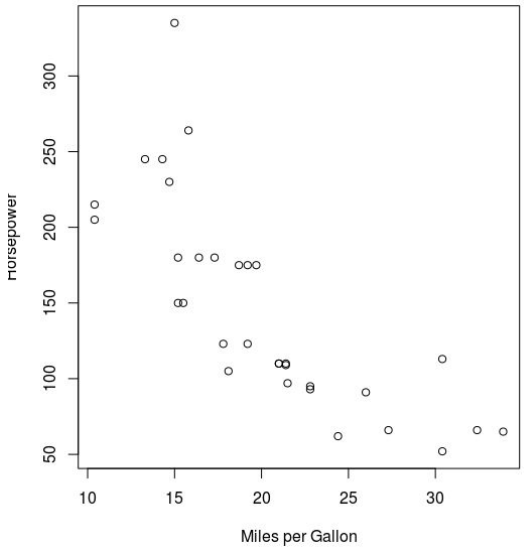
interesting?

Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

```
plot(mtcars$mpg, mtcars$hp,
  xlab = "Miles per Gallon",
  ylab = "Horsepower")
```

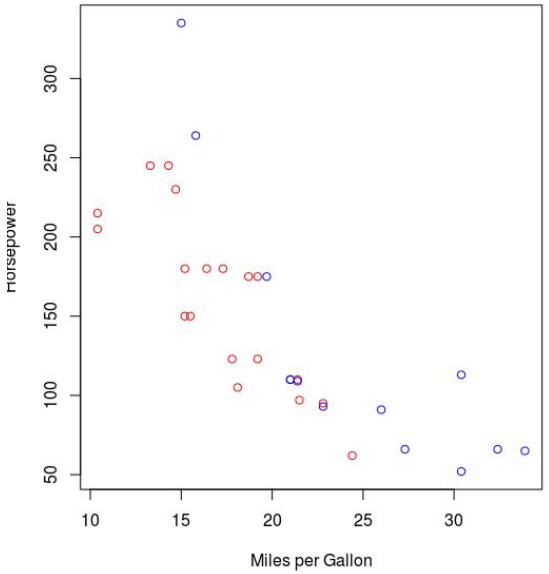


Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

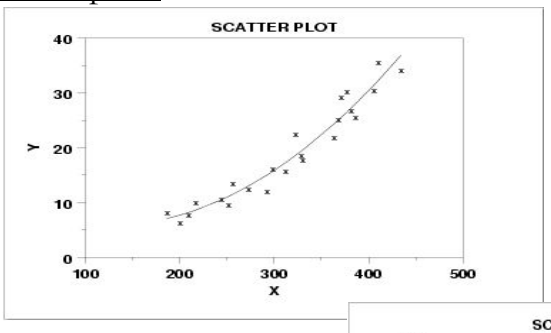
```
# create a vector for conditional color coding
colorcode <- ifelse(mtcars$am == 0, "red", "blue")

plot(mtcars$mpg,  mtcars$hp,
        xlab = "Miles per Gallon",
        ylab = "Horsepower",
        col = colorcode)
```



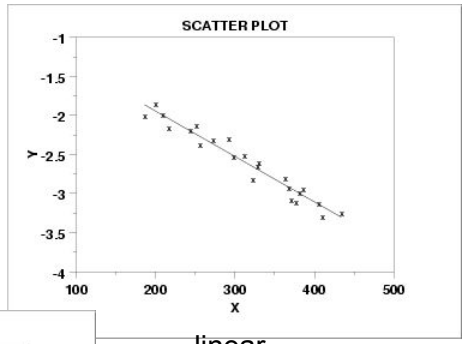
Bivariate EDA: Continuous & Continuous

DMI

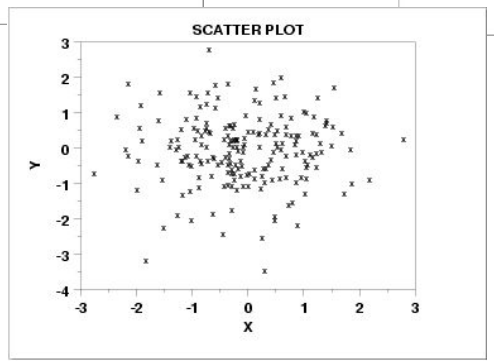
Scatter plots



quadratic



linear



No apparent relation

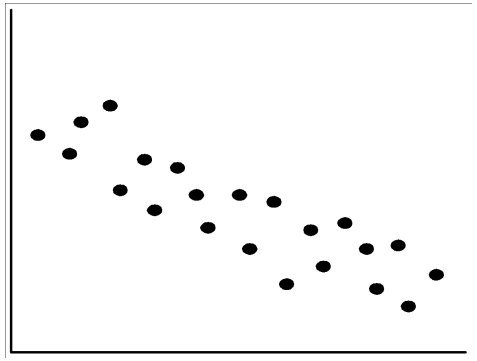
Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

positively correlated data

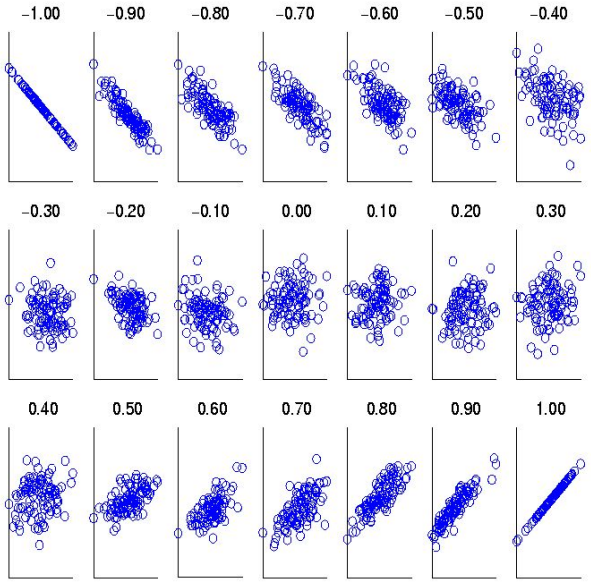
negative correlated data



Bivariate EDA: Continuous & Continuous

DMI

Scatter plots



Visualizing Changes of Correlation Coefficient

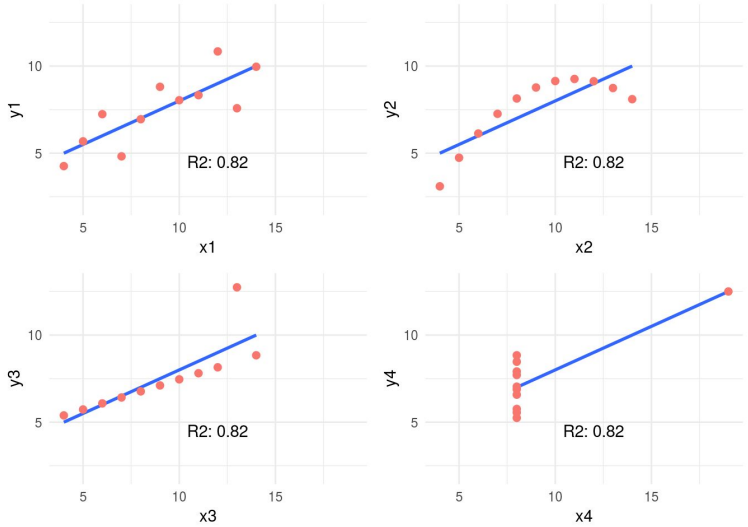
Correlation coefficient value range: [–1, 1]

A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1

Bivariate EDA: Continuous & Continuous

DMI

Scatter plots



Beware!

4-different plots, having the same mean for every x and y variable (9 and 7.501 respectively), and the same degree of correlation. You can check all the measures by typing summary(anscombe_data).

This is why is so important to plot relationships when analyzing correlations!!

For more info on Anscombe's quartet: Wikipedia:

Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

- But can be bad with lots of data

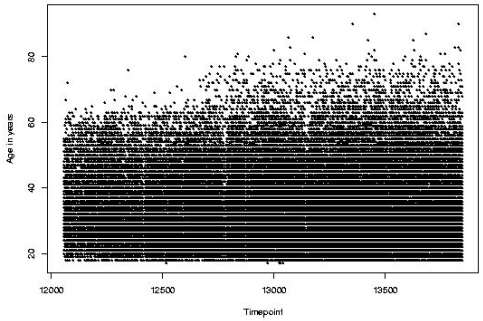


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Overplotting is when the data or labels in a data visualization overlap, making it difficult to see individual data points in a data visualization.

Overplotting typically occurs when there are either a large number of data points and/or a small number of unique values in the dataset.

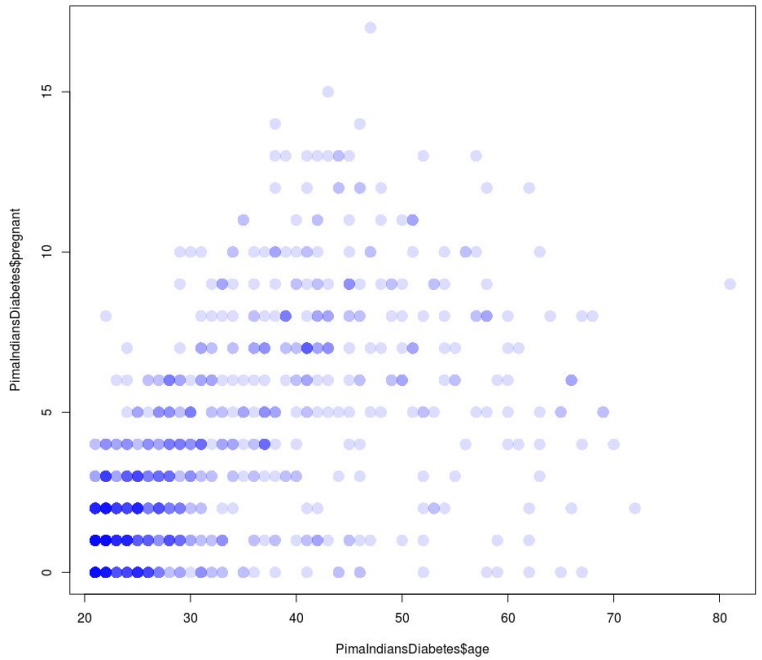
Bivariate EDA: Continuous & Continuous

DMI

Scatter plots

Transparent plotting

```
plot( PimaIndiansDiabetes$age,
      PimaIndiansDiabetes$pregnant,
      col="#0000ff22", pch=16,cex=2)
```



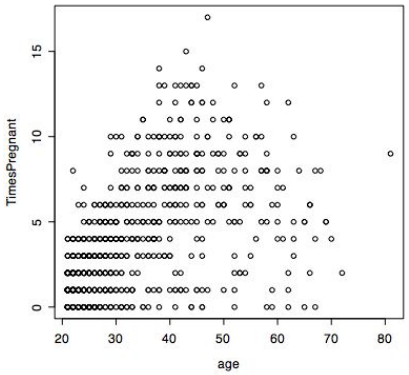
Bivariate EDA: Continuous & Continuous

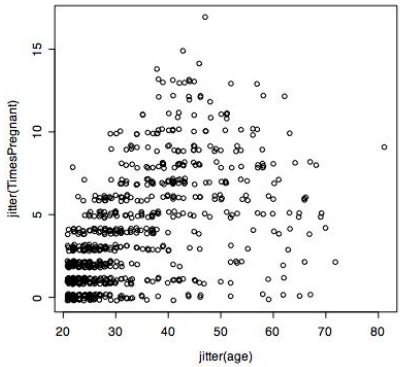
DMI

Scatter plots

Jittering points helps too:

- plot(PimaIndiansDiabetes$age,PimaIndiansDiabetes$pregnant)
- plot(jitter(PimaIndiansDiabetes$age),jitter(PimaIndiansDiabetes$pregnant))



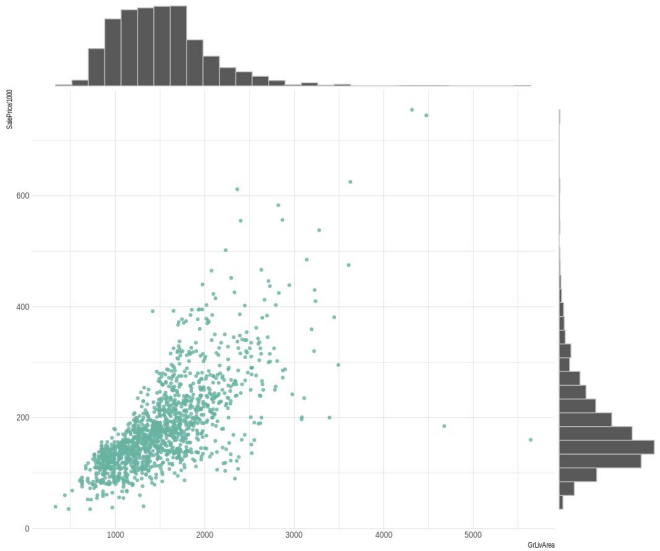


Bivariate EDA: Continuous & Continuous

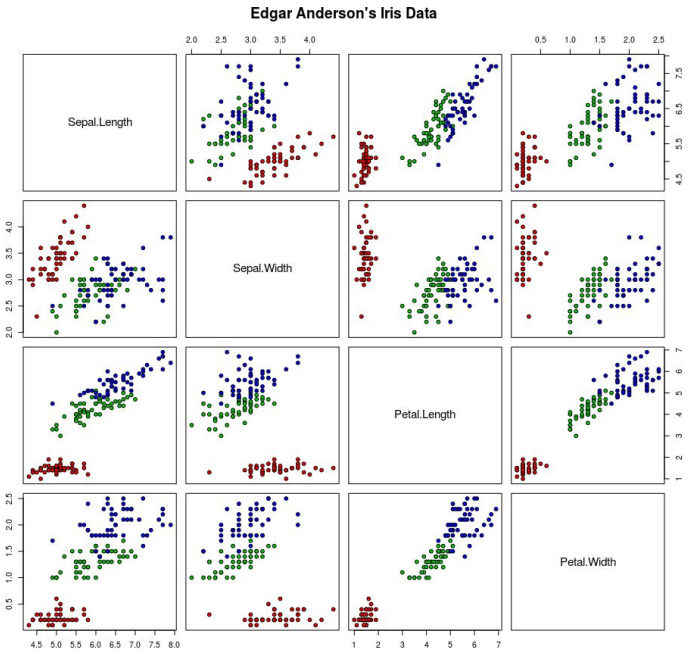
DMI

Scatter plots

```
data <- fread("https://raw.githubusercontent.com/jpinero/DMI_2021/main/datasets/house/train.csv")
library(ggExtra)
# create a ggplot2 scatterplot
p <- data %>%
  ggplot( aes(x=GrLivArea, y=SalePrice/1000)) +
    geom_point(color="#69b3a2", alpha=0.8) +
    theme_ipsum() +
    theme(
      legend.position="none"
    )

# add marginal histograms
  ggExtra::ggMarginal(p, type = "histogram", color="grey")
```



EDA: more than 2 variables



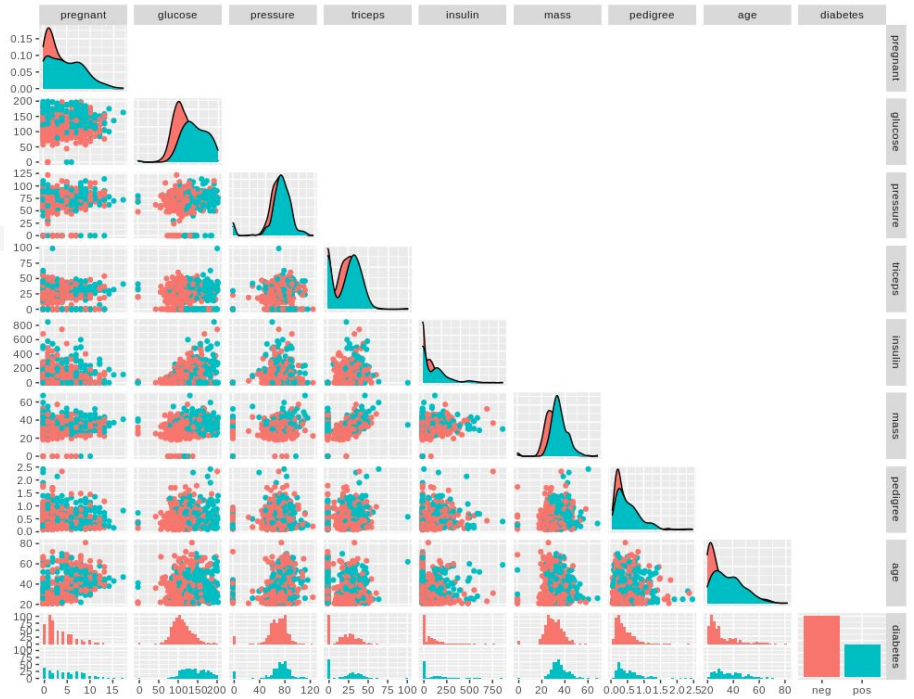
Pairwise scatterplots of Iris Attributes

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in 1936

```
pairs(iris[1:4], main = "Iris Data", pch = 21, bg =
c("red", "green3", "blue")[unclass(iris$Species)])
```

EDA: more than 2 variables

```
library(mlbench)
library("GGally")

data("PimaIndiansDiabetes")
ggpairs(PimaIndiansDiabetes,
```

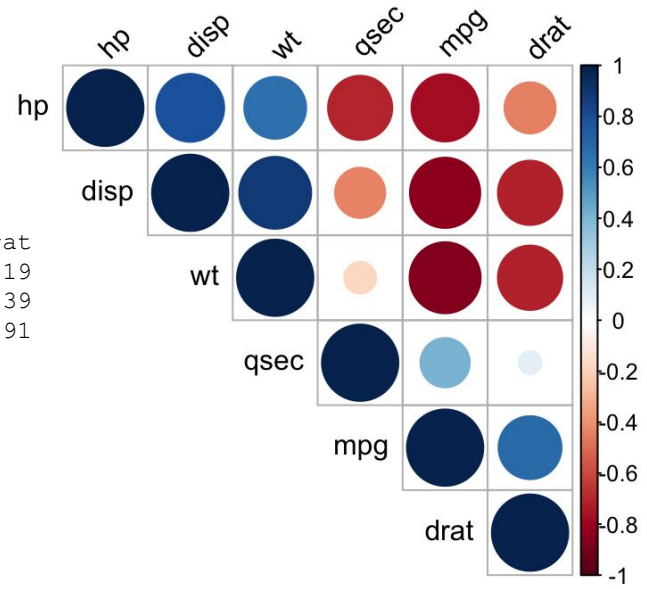


EDA: more than 2 variables

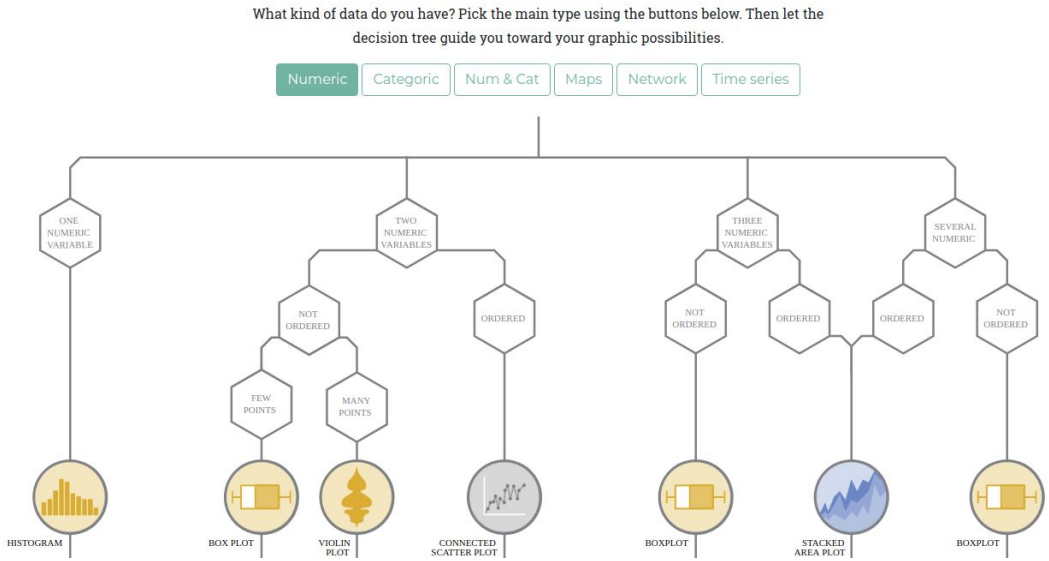
DMI

Correlation plots

```
res <- cor( mtcars[, c(1,3,4,5,6,7)])
res[1:3, 1:4]
             mpg       disp         hp       drat
mpg    1.0000000 -0.8475514 -0.7761684  0.6811719
disp -0.8475514  1.0000000  0.7909486 -0.7102139
hp   -0.7761684  0.7909486  1.0000000 -0.4487591

library(corrplot)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

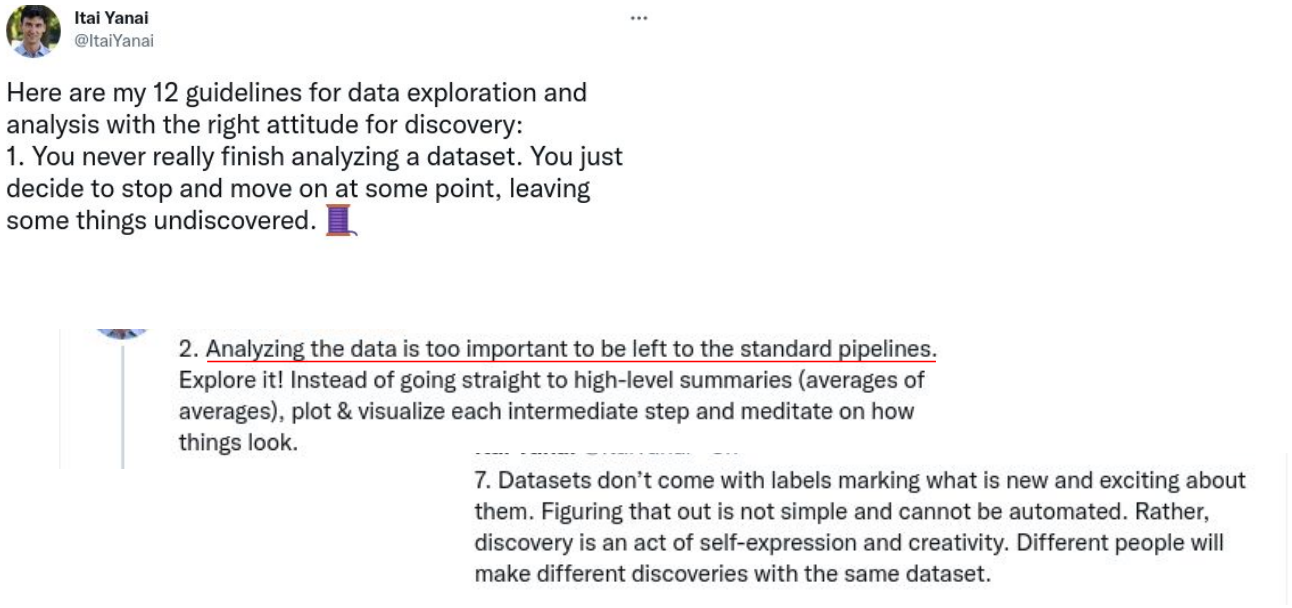


Choosing your plot



https://www.data-to-viz.com/

Guidelines for data exploration



Itai Yanai
@ItaiYanai

Here are my 12 guidelines for data exploration and analysis with the right attitude for discovery:
1. You never really finish analyzing a dataset. You just decide to stop and move on at some point, leaving some things undiscovered. 🧵

2. Analyzing the data is too important to be left to the standard pipelines. Explore it! Instead of going straight to high-level summaries (averages of averages), plot & visualize each intermediate step and meditate on how things look.

7. Datasets don't come with labels marking what is new and exciting about them. Figuring that out is not simple and cannot be automated. Rather, discovery is an act of self-expression and creativity. Different people will make different discoveries with the same dataset.

https://x.com/ItaiYanai/status/1612627199332433922

EDA checklist

DMI

- Did you plot univariate and multivariate summaries of the data?
- Did you check for outliers?
- Did you identify the missing data code?
- Is each variable one column? Is each observation one row?
- Do different data types appear in each table?
- Did you record the recipe for moving from raw to tidy data?
- Did you create a code book?
- Did you record all parameters, units, and functions applied to the data?
- Did you identify missing values?
- Did you make univariate plots (histograms, density plots, boxplots)?
- Did you consider correlations between variables (scatterplots)?
- Did you check the units of all data points to make sure they are in the right range?
- Did you try to identify any errors or miscoding of variables?
- Did you consider plotting on a log scale?
- ….

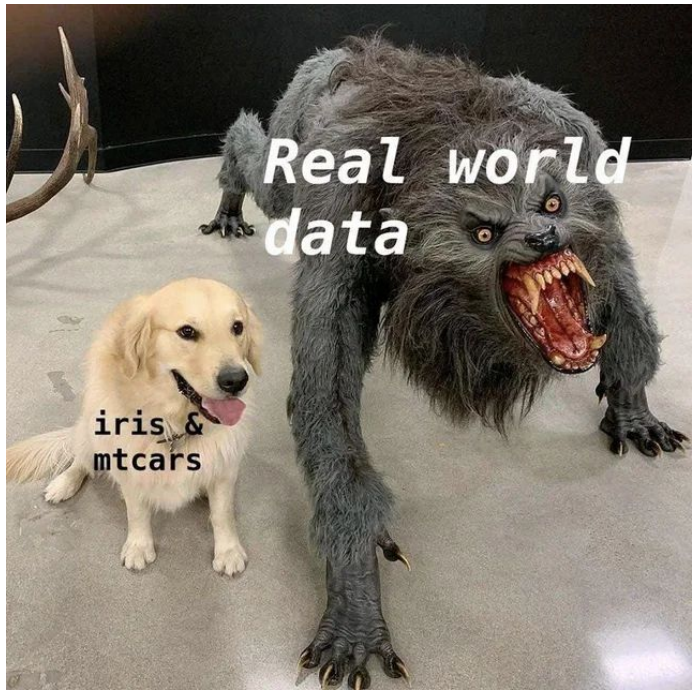
R packages for EDA

DMI

- DataExplorer
- GGally
- SmartEDA
- tableone
- vtable
- summarytools

And always keep in mind…

DMI



Bibliography

From the book R for Data Science (Garrett Grolemund, Hadley Wickham)

The chapter

https://r4ds.had.co.nz/exploratory-data-analysis.html

Hands-on session

Go to this page and accept the assignment

Exploratory Data Analysis

DMI

"There are no routine statistical questions, only questionable statistical routines." — Sir David Cox

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." — John Tukey