Exploratory Data Analysis and Feature Engineering

Quan Zhang, Ph.D.

Exploratory Data Analysis (EDA)

A first look at the data

Data collected from experiments is not “clean”

EDA is a critical first step in analyzing the data from an experiment

What EDA does

1. detection of mistakes
2. checking of assumptions
3. preliminary selection of appropriate models
4. determining relationships among the explanatory variables
5. assessing the direction and rough size of relationships between explanatory and outcome variables.

Loosely speaking, exploratory data analysis (EDA) includes any method of looking at data that does not require formal statistical modeling and inference

Typical data format

The collected data is often in the form of a matrix

- One row for each observation
- One column for each subject id, outcome variable, and explanatory variable
- Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable

Types of EDA

Exploratory data analysis is generally cross-classified in two ways

1. Either non-graphical or graphical
2. Either univariate or multivariate (usually just bivariate)

Specifically,

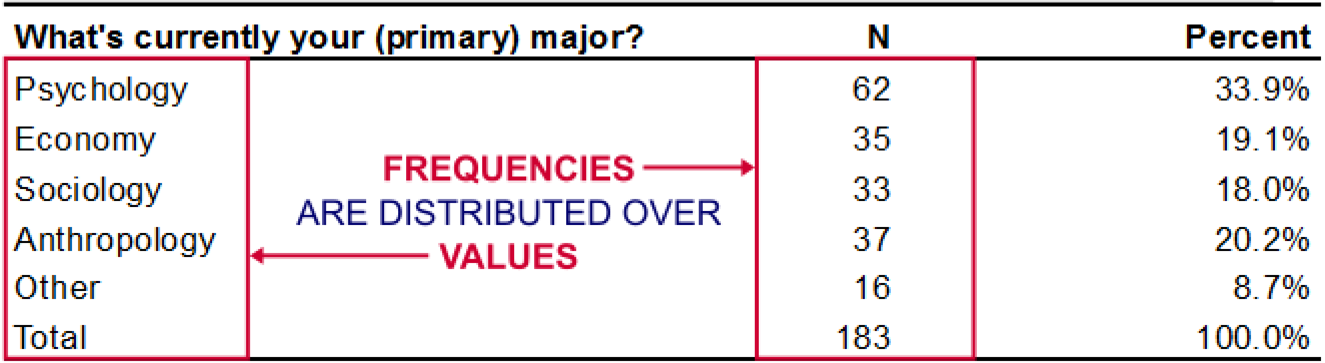
- Non-graphical methods generally involve calculation of summary statistics
- graphical methods summarize the data in a pictorial way
- Univariate methods look at one variable (data column) at a time
- multivariate methods look at two or more variables at a time to explore relationships (* perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA)

Univariate non-graphical EDA

1. The usual goal of univariate non-graphical EDA is to better appreciate the “sample distribution” and also to make some tentative conclusions about the population distribution(s)
2. Outlier detection is also a part of this analysis.

Univariate non-graphical EDA: categorical data

- a categorical variable is simply the range of values and the frequency (or relative frequency) of occurrence for each value.
- EDA strategy: tabulation of the frequencies, usually along with calculation of the fraction



What's currently your (primary) major?	N	Percent
Psychology	62	33.9%
Economy	35	19.1%
Sociology	33	18.0%
Anthropology	37	20.2%
Other	16	8.7%
Total	183	100.0%

FREQUENCIES ARE DISTRIBUTED OVER VALUES

Univariate non-graphical EDA: quantitative data

Univariate EDA for a quantitative (continuous) variable is a way to make preliminary assessments about the population distribution of the variable

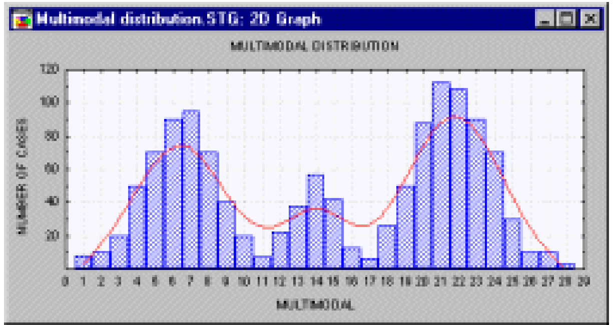
The characteristics of the sample/population distribution:

- Population: center, spread, modality (number of peaks in the pdf), shape (including “heaviness of the tails”), and outliers
- Sample statistics: sample mean, sample variance, sample standard deviation, sample skewness and sample kurtosis (kurtosis measures tail extremity; i.e., either existing outliers (for the sample kurtosis) or propensity to produce outliers)
- These sample statistics tell us more detailed information of a histogram

Univariate non-graphical EDA: quantitative data

Central tendency:

- AKA “location” of a distribution, describing typical or middle values
- Useful statistics for the central tendency include mean, median, and sometimes mode
- Mean is the most commonly used
- Median is makes more sense (more robust) if the tail is heavy (there are extreme values)
- Mode is rarely used:
    - Describes most likely or frequently occurring value
    - Can be useful if the distribution is of multi-modality



Univariate non-graphical EDA: quantitative data

Spread

- The variance and standard deviation are two useful measures of spread
  - The variance is the mean of the squares of the individual deviations.
  - The standard deviation is the square root of the variance.
  - For Normally distributed data, approximately 95% of the values lie within 2 sd of the mean.
- Interquartile range (IQR) is also useful
  - One quarter of the data fall below the first quartile, denoted by Q1
  - Three quarters of the data fall below the third quartile, denoted by Q3
  - Interquartile range (IQR) = Q3-Q1
  - The middle half of data fall within the IQR
  - If the data are more spread out, then the IQR tends to increase, and vice versa
  - **The IQR is a more robust measure of spread than the variance or standard deviation**

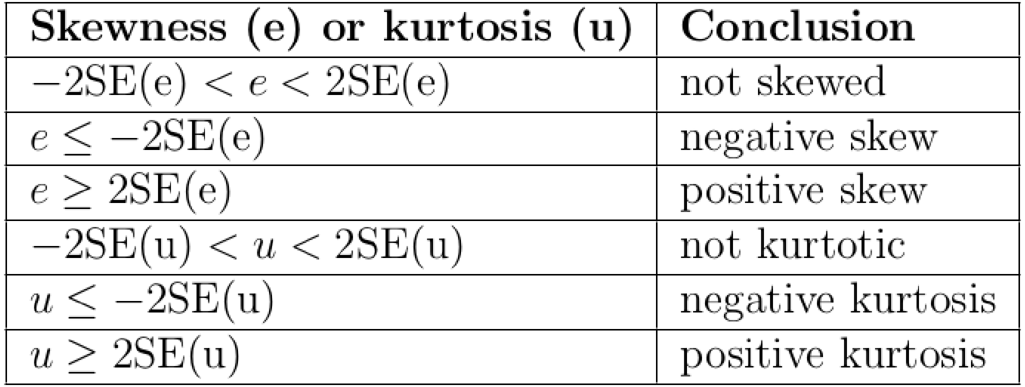
Univariate non-graphical EDA: quantitative data

Skewness: $\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$

- A measure of asymmetry

Kurtosis: $\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$

- A measure of tailedness
- rarely used



Skewness (e) or kurtosis (u)	**Conclusion**
$-2\mathrm{SE(e)} < e < 2\mathrm{SE(e)}$	not skewed
$e \leq -2\mathrm{SE(e)}$	negative skew
$e \geq 2\mathrm{SE(e)}$	positive skew
$-2\mathrm{SE(u)} < u < 2\mathrm{SE(u)}$	not kurtotic
$u \leq -2\mathrm{SE(u)}$	negative kurtosis
$u \geq 2\mathrm{SE(u)}$	positive kurtosis

where e and u are estimates of skewness and kurtosis with standard errors SE(e) and SE(u)

Univariate non-graphical EDA: quantitative data

- If the quantitative variable, say age, does not have too many distinct values, a tabulation, as we used for categorical data
- But for categorical variables, none of these sample statistics make any sense

Univariate graphical EDA

A more complete picture of the sample distribution

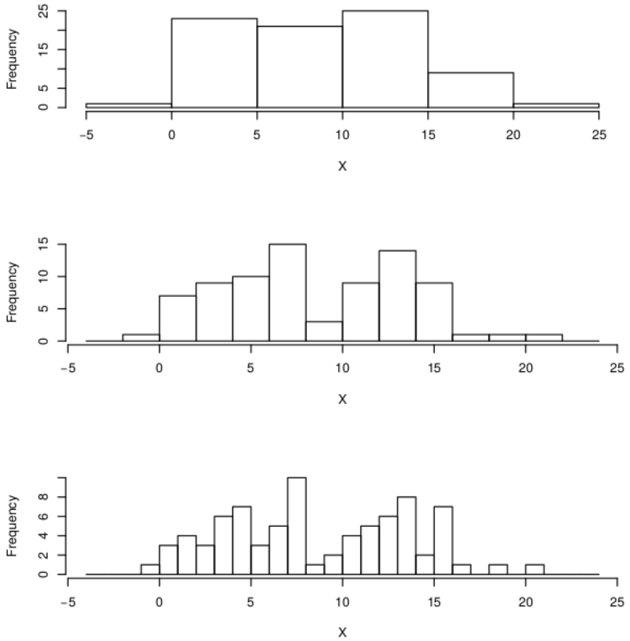
- Histogram
- Boxplot
- QQ plot (commonly used in hypothesis tests instead of modeling)

Univariate graphical EDA: histogram

- Histograms make sense for both categorical and quantitative data
- Useful for ordinal data (like ratings between 1 and 5)
  - Ordinal data is often treated as quantitative data
- Histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

Univariate graphical EDA: histogram

- Generally you will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution

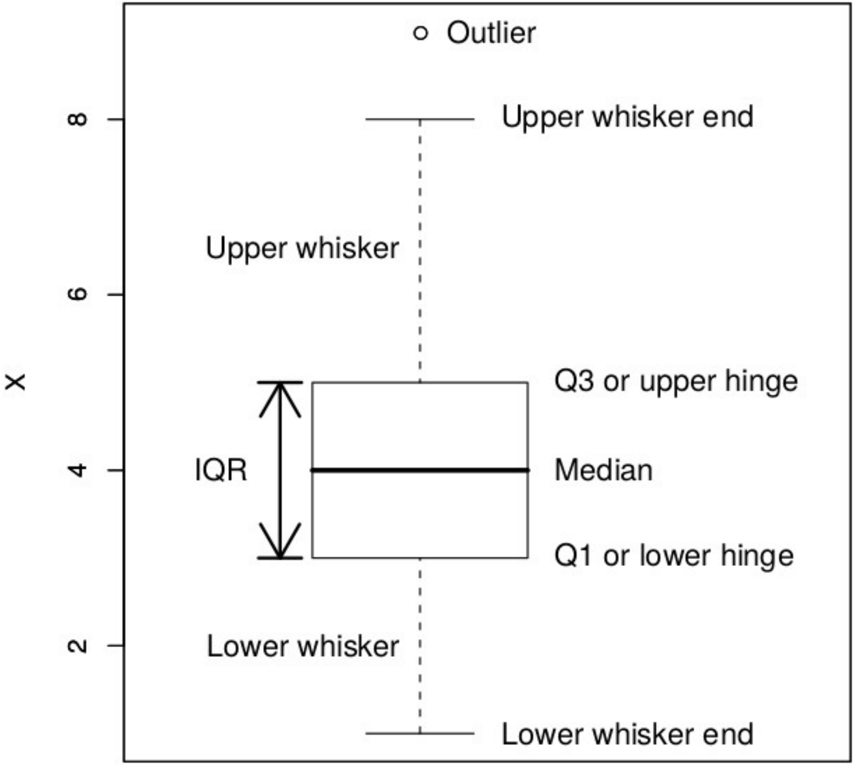


Univariate graphical EDA: boxplot

- good at presenting information about the central tendency, symmetry and skew, as well as outliers,
- Can be misleading about aspects such as multimodality

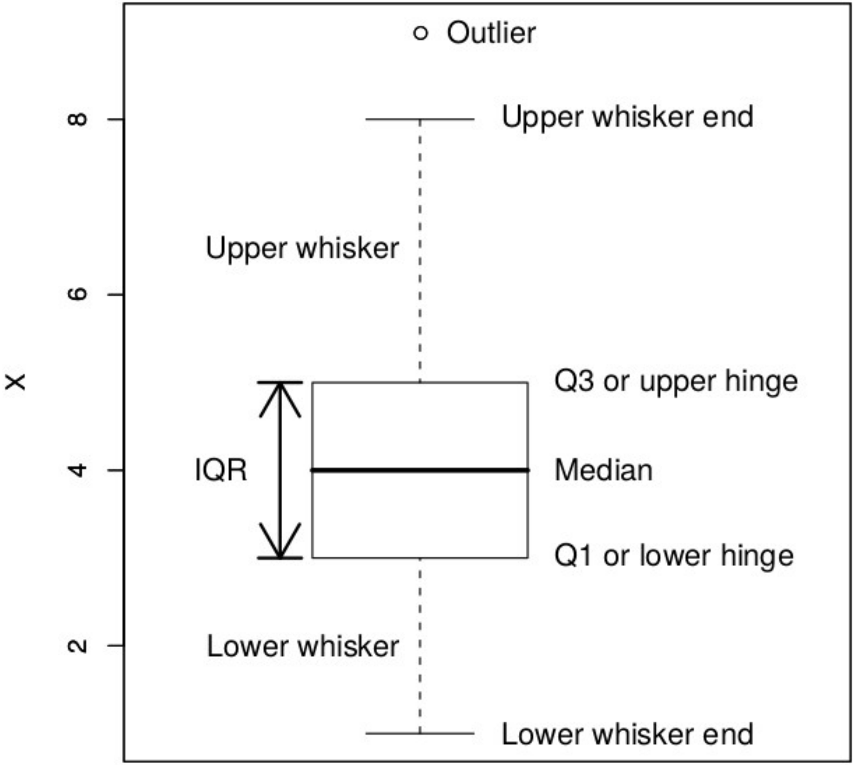
Univariate graphical EDA: boxplot

- Each whisker is drawn out to the most extreme data point that is less than 1.5 IQRs beyond the corresponding hinge



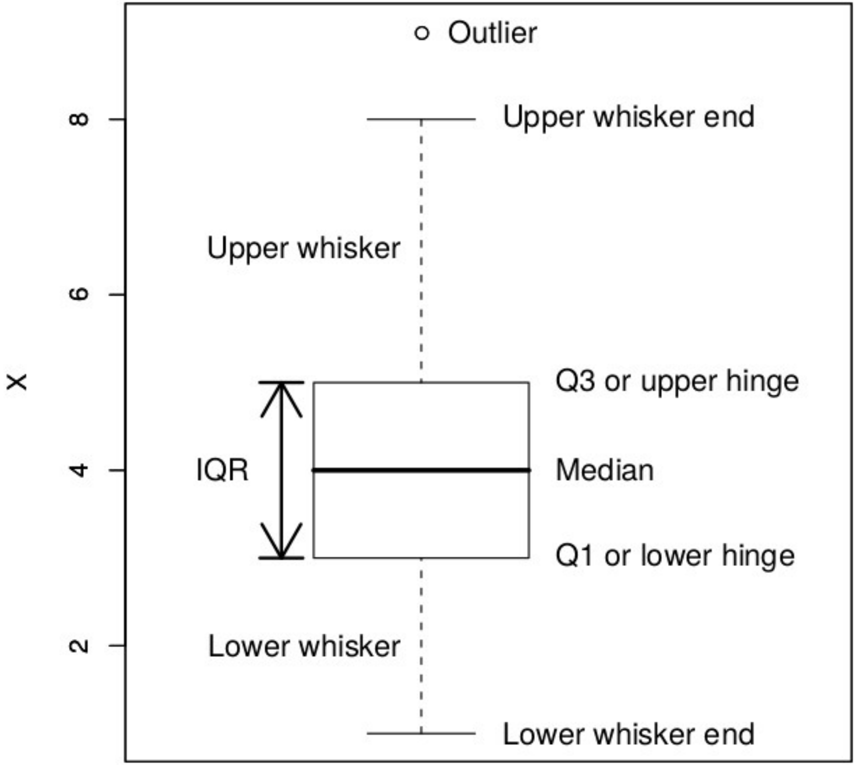
Univariate graphical EDA: boxplot

- Any data value more than 1.5 IQRs beyond its corresponding hinge in either direction is considered an “outlier” and is individually plotted
- The term “outlier” is not well defined in statistics; the definition varies depending on the purpose and situation



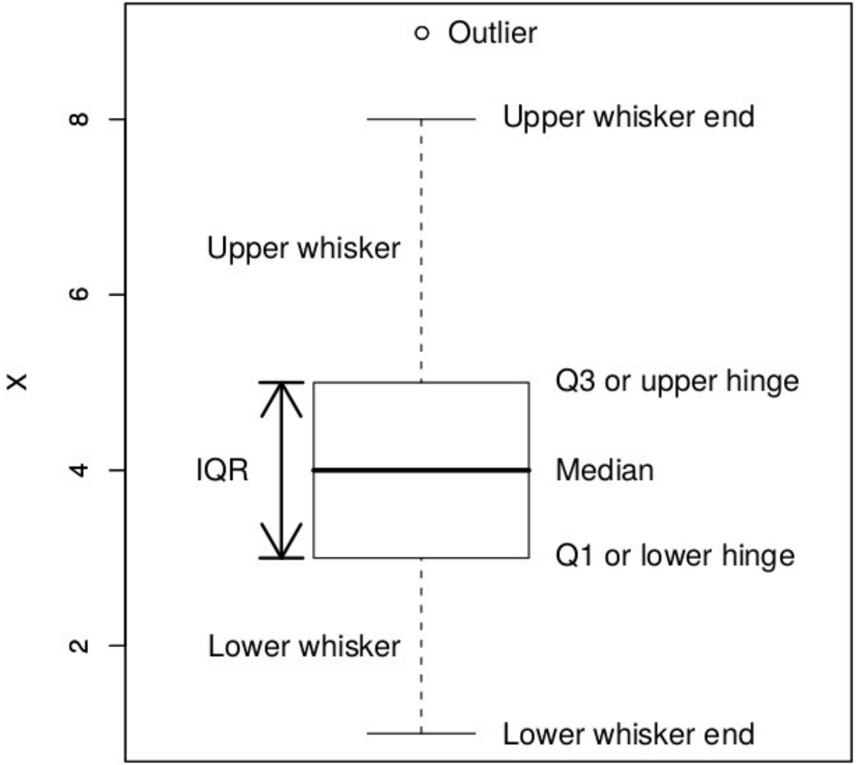
Univariate graphical EDA: boxplot

- Symmetry is appreciated by noticing if the median is in the center of the box and if the whiskers are the same length as each other.



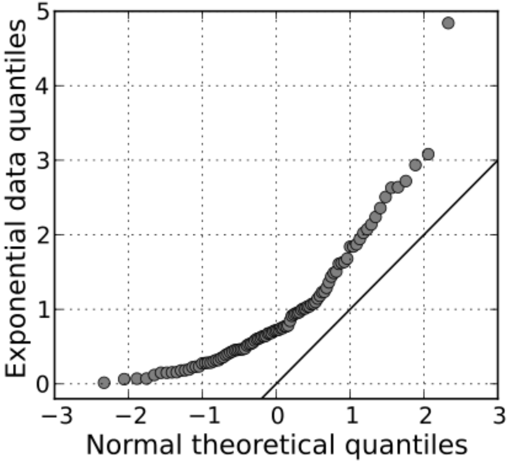
Univariate graphical EDA: boxplot

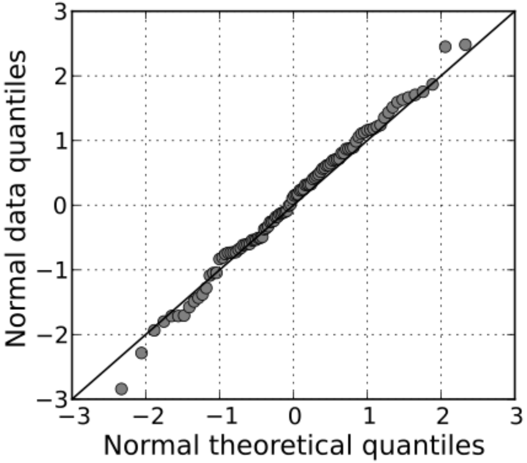
- In a boxplot, many outliers suggests heavy tails (positive kurtosis), or possibly many data entry errors
- Boxplots are excellent EDA plots because they rely on robust statistics Like median and IQR rather than more sensitive ones such as mean and sd.



Univariate graphical EDA: quantile-quantile plot

- Quantile-quantile (QQ) plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate)





Univariate graphical EDA: quantile-quantile plot

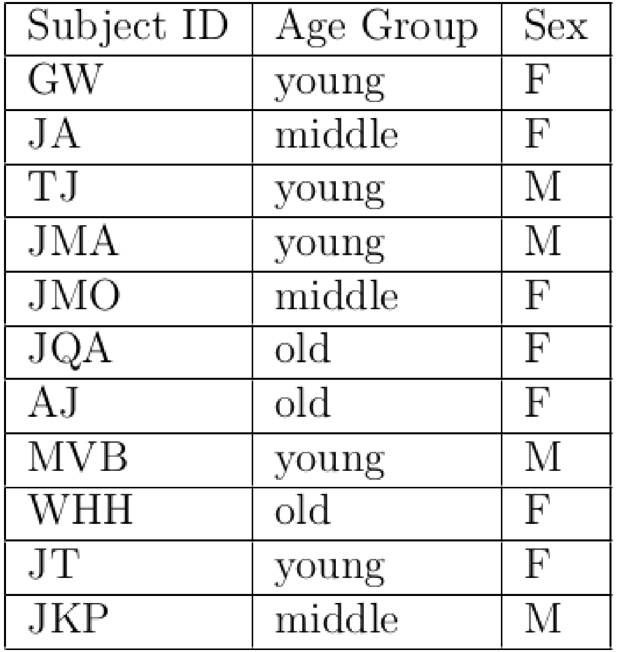
- Quantile-quantile (QQ) plot is often used in hypothesis testing problems
  - Two-sample t-test requires approximately normal data
- In modeling (regression, logistic...) we don’t have to assume distributions of predictors.
- An especially useful application: examining the normality of residuals in a linear regression

Multivariate non-graphical EDA

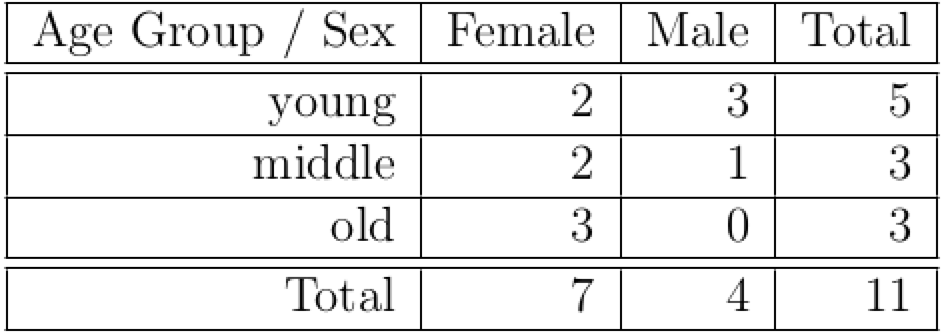
- generally show the relationship between two or more variables in the form of either cross-tabulation or statistics
    - Cross-tabulation: for categorical data (and quantitative data with only a few different values)
    - Univariate statistics by category: for a pair of quantitative and categorical variables
    - Covariance and correlation matrices: for multiple quantitative variables
    - Correlation between two categorical variables is out of the scope of this class (See https://rpubs.com/hoanganhngo610/558925 for detail)

Multivariate non-graphical EDA: Cross-tabulation

- Cross-tabulation: for categorical data (and quantitative data with only a few different values)



Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M



Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Multivariate non-graphical EDA: covariance and correlation

- The sample covariance/correlation is a measure of how much two quantitative variables “co-vary”, i.e., how much (and in what direction) should we expect one variable to change when the other changes
    - Positive covariance/correlations values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa
    - Negative covariances/correlations suggest that when one variable is above its mean, the other is below its mean
    - Covariances/correlations near zero suggest that the two variables vary independently
      Independence implies zero covariance/correlation, but the reverse is not necessarily true

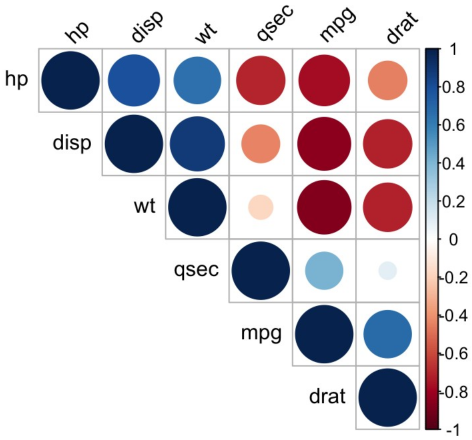
$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s_x s_y}$$

Multivariate non-graphical EDA: covariance and correlation matrices

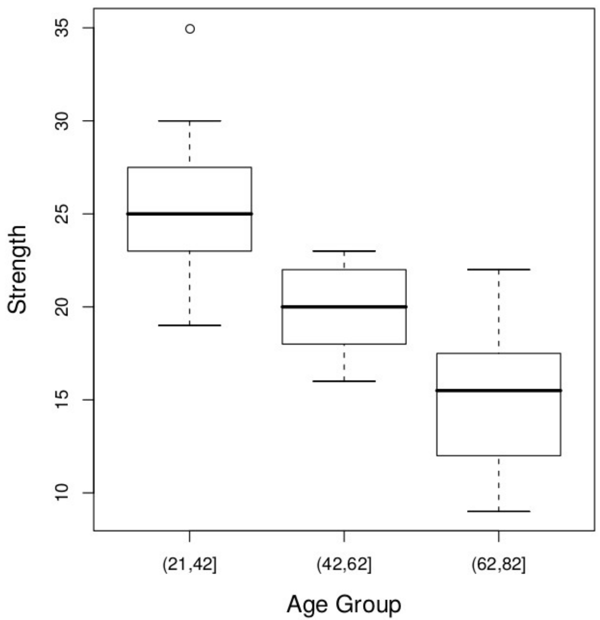
- For multiple quantitative variables X1, X2, X3,..., a covariance (correlation) matrix has element (i,j) as the covariance (correlation) between Xi and Xj.
- Visualize a correlation matrix in R

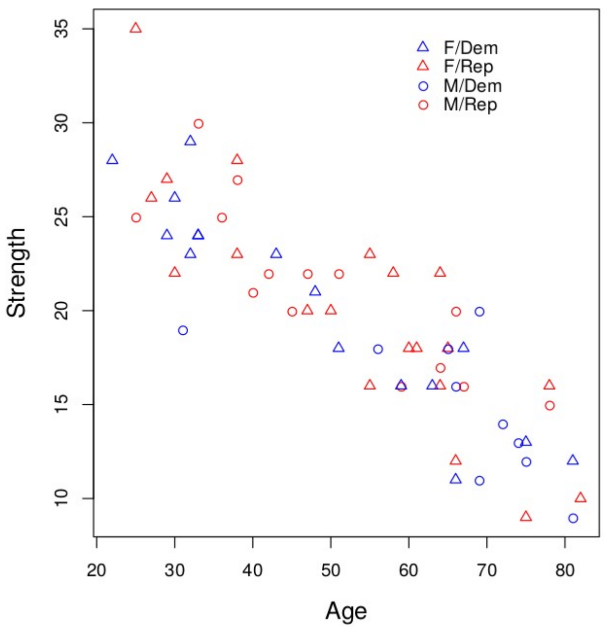
```
library(corrplot)

corrplot(...)
```



Multivariate graphical EDA

- Univariate graphs by category: for a quantitative and a categorical variable
    - Side-by-side box plots, scatterplots (for a pair of quantitative variables in different groups)





Summary of EDA

You should always perform appropriate EDA before further analysis of your data.

Perform whatever steps are necessary to **become more familiar with your data**, **check for obvious mistakes**, **learn about variable distributions**, and **learn about relationships between variables**.

EDA is not an exact science – it is a very important art!

In this class, whenever you have mistakes or missing data, simply remove these observations. But in your future career, you should always explore where the mistakes and missing data are from if there are a lot of them.

Feature engineering: Exaggerate Signals that Matter

- One of the most creative parts of the data science process—literally the art of data science

- There are two types of data scientists: **model-centric** and **feature-centric**

- The model-centric data scientists believe that throwing a large amount of labeled data and computational resources (e.g., GPUs) will automatically learn the right features (in the lower layers) as well as the mapping between those features and the output.

- The feature-centric data scientists believe in systematically and painstakingly creating meaningful features to make the modeling stage simple.

Model-centric data science

- The creative process takes the form of designing the right architecture—nature and type of layers in the deep learning models as opposed to designing individual features.

- This model-centric deep learning approach works well in domains such as text, vision, speech, and time series data where

(1) the space of possible features is very large, which makes it impractical to explore it through traditional feature engineering

(2) the amount of data is substantial enough to learn the large number of parameters in deep learning models

(3) we need a hierarchy of features and not just a single layer of features (CNN)

Feature-centric data science

- Feature-centric data scientists marry their deep understanding of the data (acquired from insights stage) with substantial appreciation of the domain knowledge (acquired from domain experts) to build features

- These features are highly interpretable, semantically deeper than the original data, and cover all potential aspects of input-output mapping

- This traditional approach to data science is more useful when domain knowledge and interpretability of model output is as important as prediction accuracy.

- In social sciences, interpretation is always important and feature engineering is necessary

Feature engineering

Feature transformation

- E.g., log transformation for features that have skew distributions or large variance

Feature normalization

- Even after proper transformations, the raw inputs might be in different ranges and their values in different units
- Example: value of a house, one might need features such as number of rooms and bathrooms (count), area of the house (square foot), distance from nearest school or places of interest (kilometers), prices of nearby houses sold recently (money), and age of the house (years)
- The features might need to be transformed to some min-max range (so min is always 0 and max is always 1) or zero mean and variance one
- Coefficients are learning the relative importance of these features
- (Carefully) Remove outliers before normalization

Feature engineering

Creating invariant features

- Often the raw data contains variances in it that are not related to the problem at hand
- For example, speech recognition problems have accent variances; images might have illumination, pose, rotation, and scale variances
- The final “data” that we see (e.g., sound of a word spoken by a person) is a “joint” of the actual signal in it (e.g., the actual word spoken) with additional factors (accent, tonal quality, loudness, etc.)
- Keeping what is essential for the task (signal) and ignoring what is not (noise) is the key to good feature engineering
- Understanding and removing these variances is perhaps the most complex part of feature engineering and requires deep understanding of the domain, possible sources of such variances, and the tools to remove these variances

Feature engineering

Ratio Features

- Many features contain variances that can be removed simply by dividing them with other features

- E.g., in credit models, instead of using total debt it is better to use debt-to-income ratio, instead of using total-payment a better feature would be the percent of equated monthly installments paid, and instead of total-credit-taken, percent of credit limit reached might be better features.

Feature engineering

Output feature ratios

- Output features might also have biases that must be corrected for before trying to predict them
- E.g., in forecasting sales, instead of predicting the raw sales count, we might want to predict deviation from the expected sales given the context (city, season, etc.)
- In movie or product ratings, the ratings data has inherent “consumer bias.” A critical consumer will typically rate most products say 1–3 out of 5 and hardly give a rating of 5, while a generous customer might rate most products between 3 and 5. Now a rating of 4 on a certain product does not mean the same thing for these two customers. It should be “calibrated” correctly to remove individual customer’s rating biases to make them “comparable” across customers.

Feature engineering

Creating new features

- E.g., four features in a credit card fraud prevention problem: location and time of the last and the current transaction.

- What new features can we create?

Feature engineering

Creating new features

- E.g., four features in a credit card fraud prevention problem: location and time of the last and the current transaction.
- What new features can we create?
- A common sense domain knowledge: there should be sufficient time between two distant transactions
- Create velocity: ratio of distance between current and previous transaction to time between current and previous transaction
- If velocity is too large, possibly a fraud

Feature engineering

Defining response variable

- The response variable may not be obvious
- E.g., in churn prediction, we might have to define churn in terms of future user behavior (did not make any purchase in the last *how many* months)
- E.g., In credit modeling, we might define a high-risk customer as someone who missed his last *how many* minimum payments
- In problems where future is to be predicted based on current and past observation, defining the future output to be predicted becomes very critical.

Feature engineering

Setting the right defaults

- A default value is typically associated with a feature if no meaningful value can be assigned
- For numeric features, often such default values are zero
- Default values should be carefully set, e.g., a sequence of events (browsings, purchases, returns). Suppose a feature for the first-occurrence of a type of event.

(Browsing, purchase, purchase, return)

First occurrence for browsing: 1.   First occurrence for purchase: 2.

First occurrence of return: 4

How about (Browsing, purchase, purchase)?

If we pick a default value of 0 for no occurrence of returns, it will confuse the model. A better default might be the length of the data (3, in this example) plus a constant or a big number

Feature engineering

Imputing missing features

- Missing data is ubiquitous
- Missing data imputation or removal? It depends on the data size and expert opinion
- Substituting the wrong defaults or simple average value of a feature may not always work

Feature engineering

Feature selection

- Once a large number of features have been engineered, we might decide not to use all of them together in the same model because some of them might be highly correlated with each other.
- Model-agnostic approach: Lasso, best subset…
- EDA