EXPLORATORY DATA ANALYSIS (EDA)

Introducción a la Ciencia de Datos

Coral del Val Muñoz

Dept. Ciencias de la Computación e Inteligencia Artificial,
Universidad de Granada
Dept. Molecular Biophysics, German Cancer Research Center Heidelberg, Alemania

Index

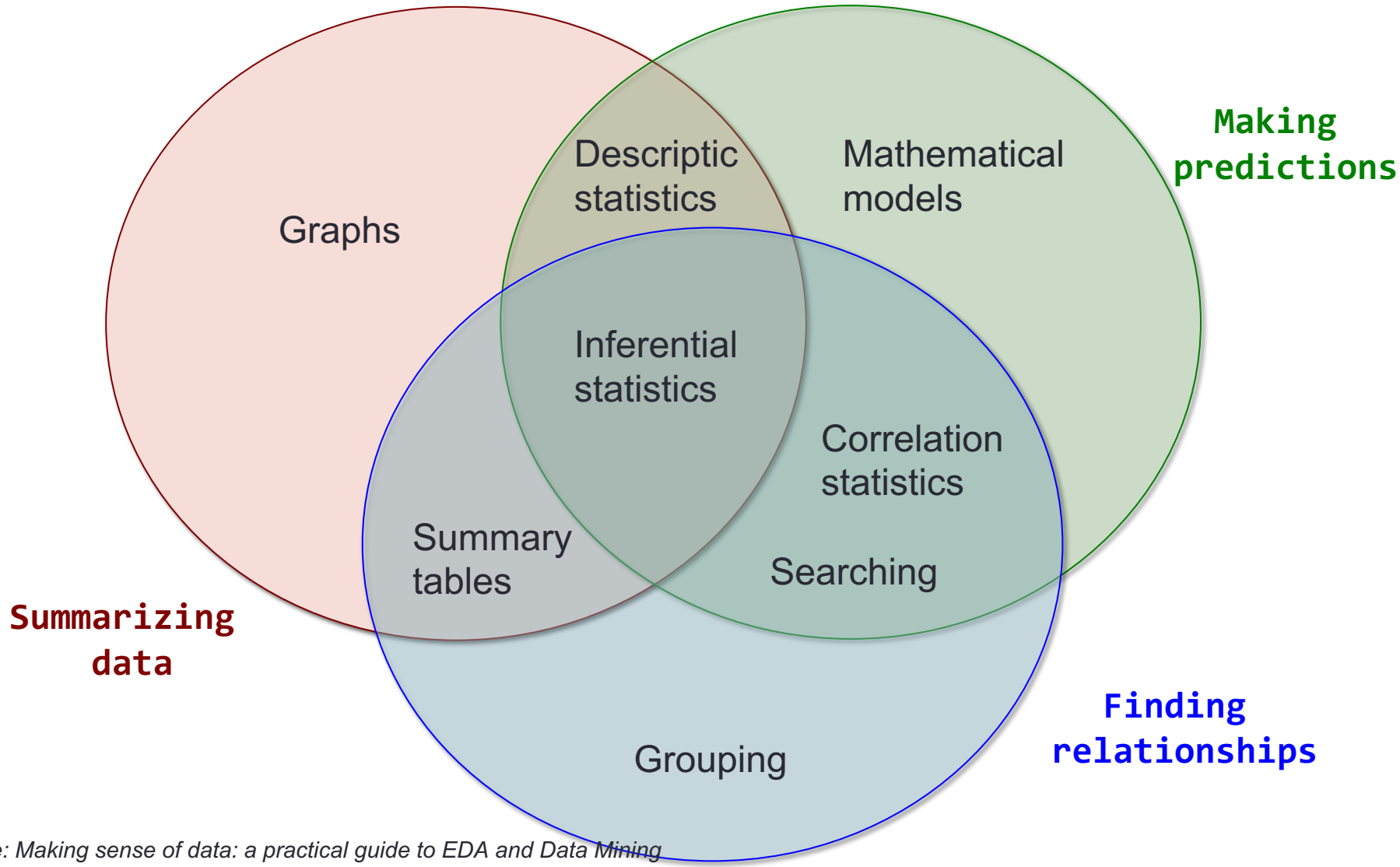
What is EDA?

- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

EDA is specially useful

- **Gain insight** into the data set
- **Discover patterns** in the data
- **Extract** important variables
- Detect **outliers** and **anomalies**
- To find violations of statistical assumptions (e.g. normal distributions or skewed?)
- Identify useful raw data & transforms (e.g. log(x))
- Generate hypothesis

Data Analysis tasks and methods



Source: Making sense of data: a practical guide to EDA and Data Mining

EDA techniques

Most EDA techniques consist of:

- Organizing and tidying data
- Plotting raw data (e.g. histograms, probability plots, scatter plots).
- Plotting simple statistics (e.g. mean plots, standard deviation plots, box plots)
- Use those plots to maximize our natural pattern-recognition abilities.

EDA tasks 1

Numeric Variables	categorical Variables	Explore distributions	Explore normality
• descriptic statistic	• fischer's test • chi-square test	• skewness • kurtosis • std, var	• QQ plots • shapiro-wilkinson test

EDA tasks 2

Compare groups	Explore correlations	Explore data	Explore
• box plots • non-parametric tests • kruskal-wallis test	• different methods	• simple linear models • non-linear models	• missing values • outliers

Understand your raw data

Using descriptive statistics

- Data types
- Data dimensionality
- Class distribution
- Data summary
- Standard deviation
- Skewness & Kurtosis
- Correlation

Using data visualization

- Univariate visualization
  - Histograms, density plots
  - Box and wisker plots
  - Bar plots
  - Missing values plots
- Multivariate Visualization
  - Correlation
  - Scatterplot matrix

Data Types

Using descriptive statistic

- Categorical
  - **Binary**
    Two categories
    `0=male,`
    `1=female`
  - **Nominal**
    More categories
    `1=green,`
    `2=blue,`
    `3=red,`
    `4=orange`
  - **Ordinal**
    Order (rank) matters
- Quantitative
  - Discrete numeric
  - Continuous uninterrupted

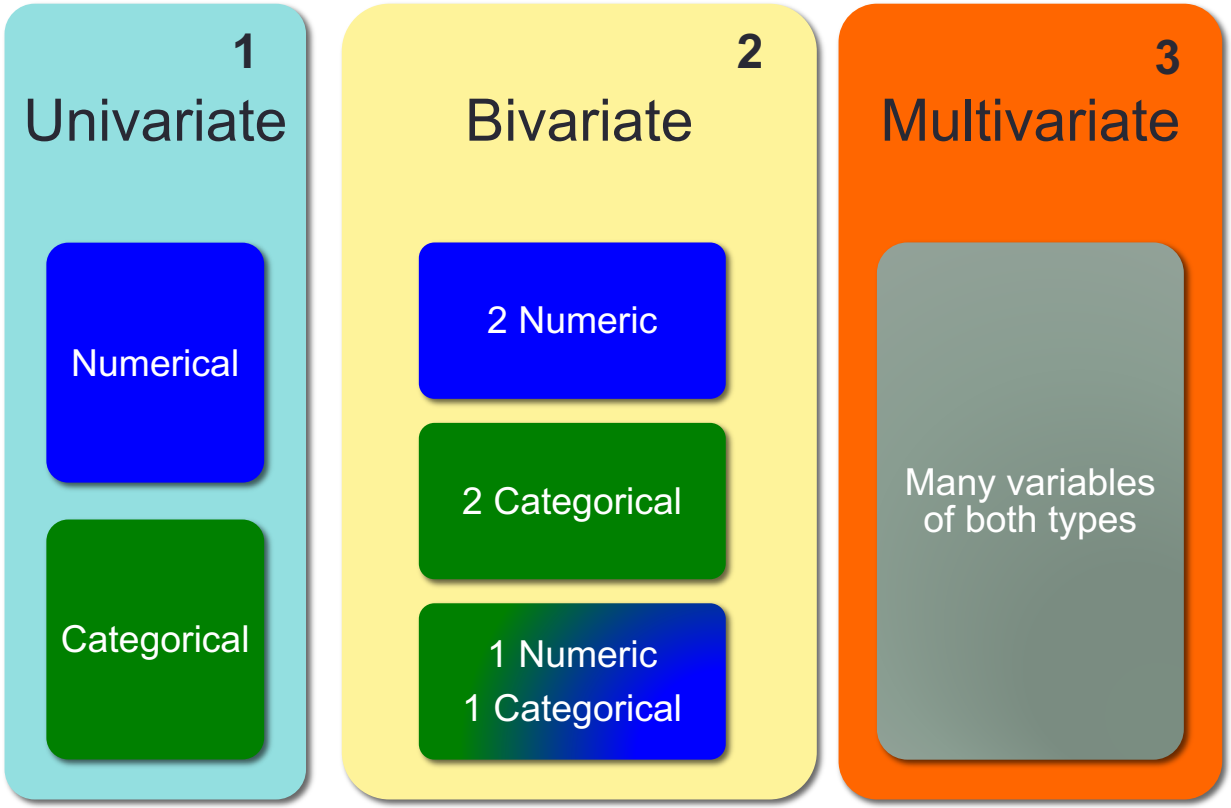
Data dimensionality

Using descriptive statistic

Univariate:
Measurement made on one variable per Subject

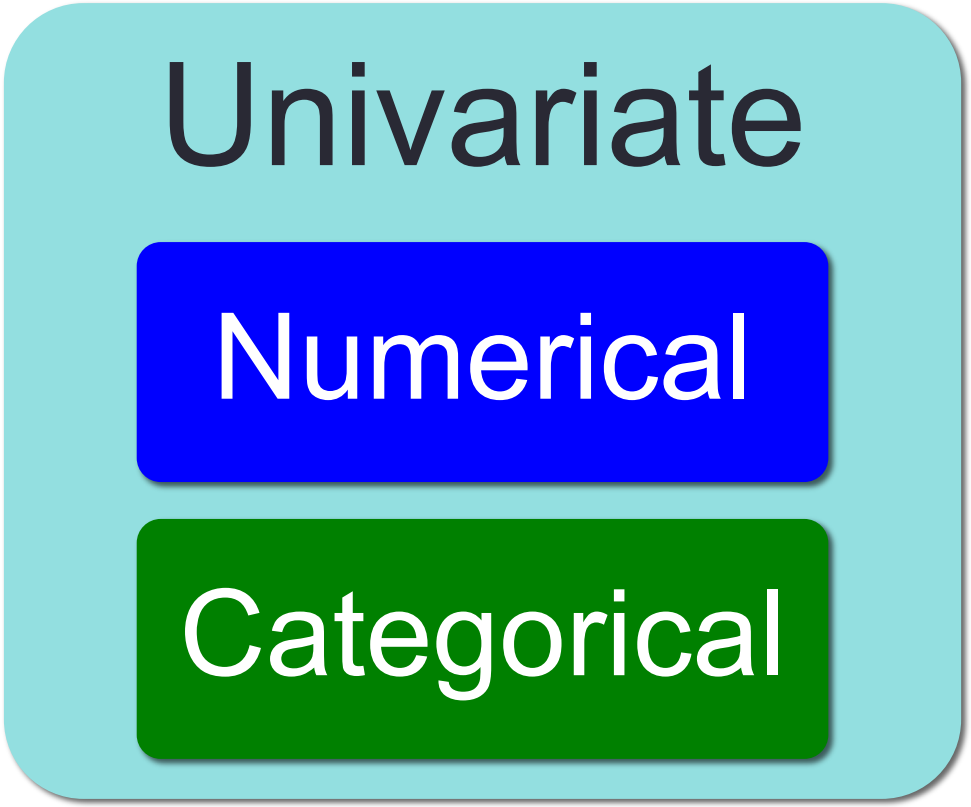
Bivariate:
Measurement made on two variables per Subject

Multivariate:
Measurement made on many variables per subject



1. Univariate analysis

Using descriptive statistic



Examining distributions of numerical data

- In order to convert raw data into useful information we need to:

'quantify' a data set, using a set of *summary statistics* (e.g. mean, median, variance, standard deviation)

Using descriptive statistic

Numerical summaries

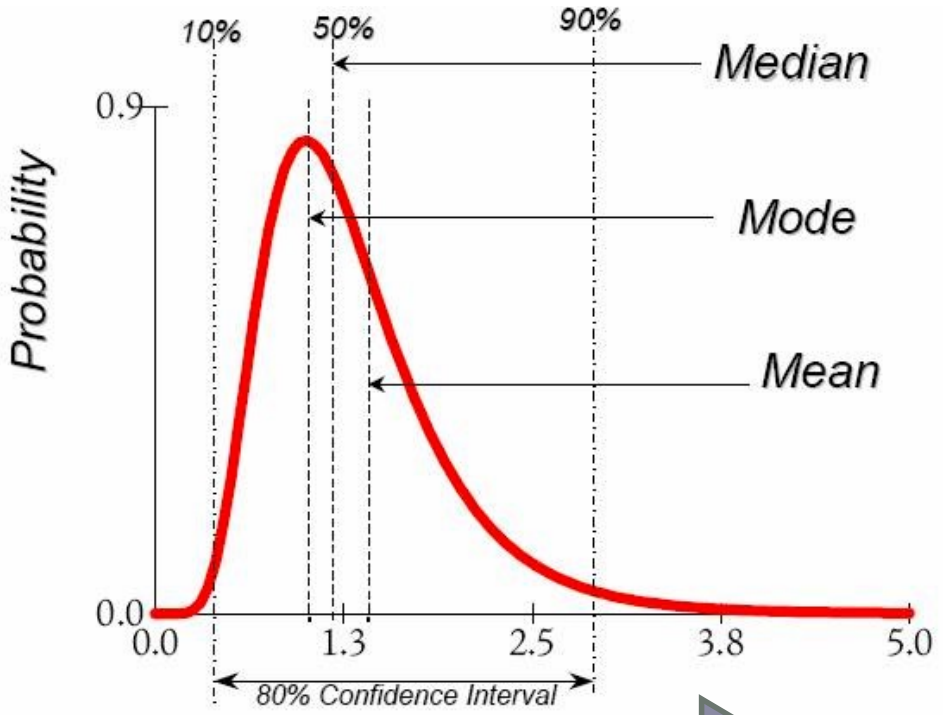
Using descriptive statistic

- **Central Tendency measures.** They are computed to give a “center” around which the measurements in the data are distributed
- **Variation or Variability measures**. They describe “data spread” or how far away the measurements are from the center.
- **Relative Standing measures**. They describe the relative position of specific measurements in the data.

Univariate analysis: Central tendency measures

- `mean()`
- `median()`
- `mode()`

- When using most of these functions remember to use argument **`na.rm = T`**

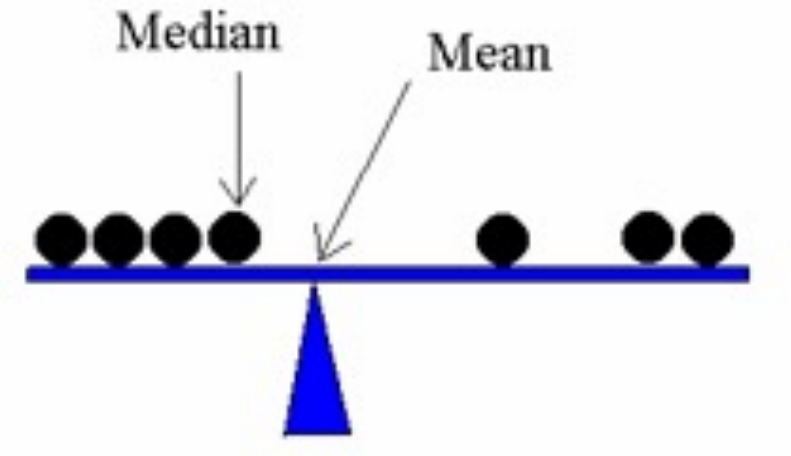


source
http://herdingcats.typepad.com/photos/uncategorized/statistics.jpg

Using descriptive statistic

Mean or Median?

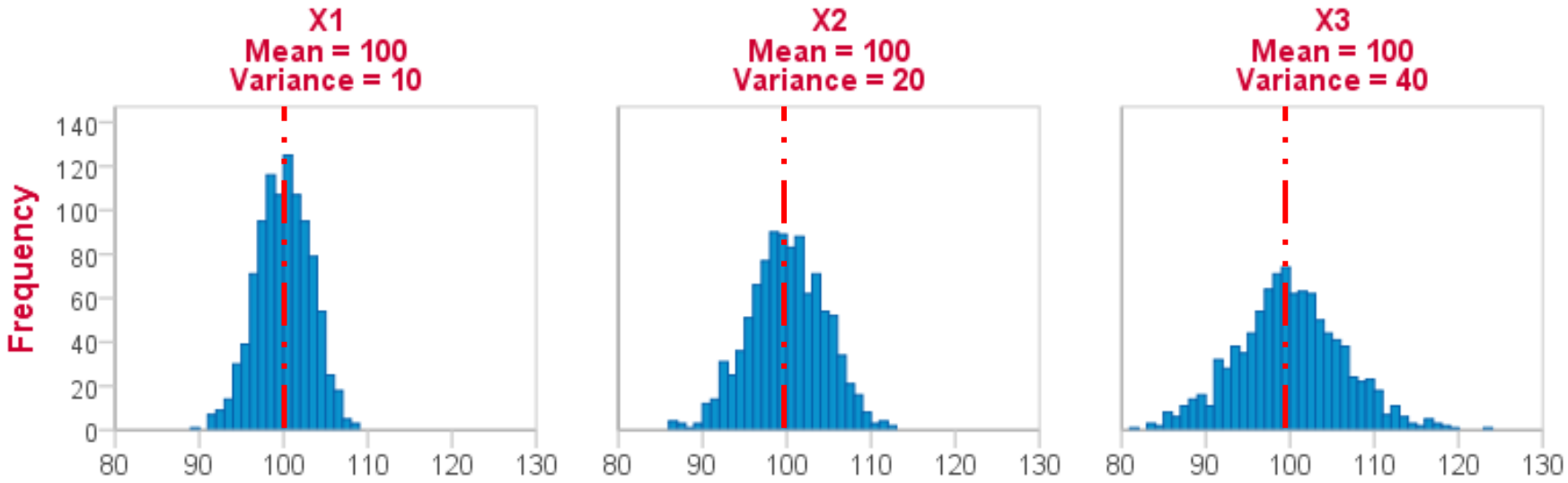
- Mean is best for symmetric distributions without outliers.
- The mean is not a robust tool since it is largely influenced by outliers.
- Median is useful for skewed distributions or data with outliers. It derives at central tendency since it is much more robust and sensible.



Source: https://onlinecourses.science.psu.edu

Using descriptive statistic

Univariate analysis: dispersion measures

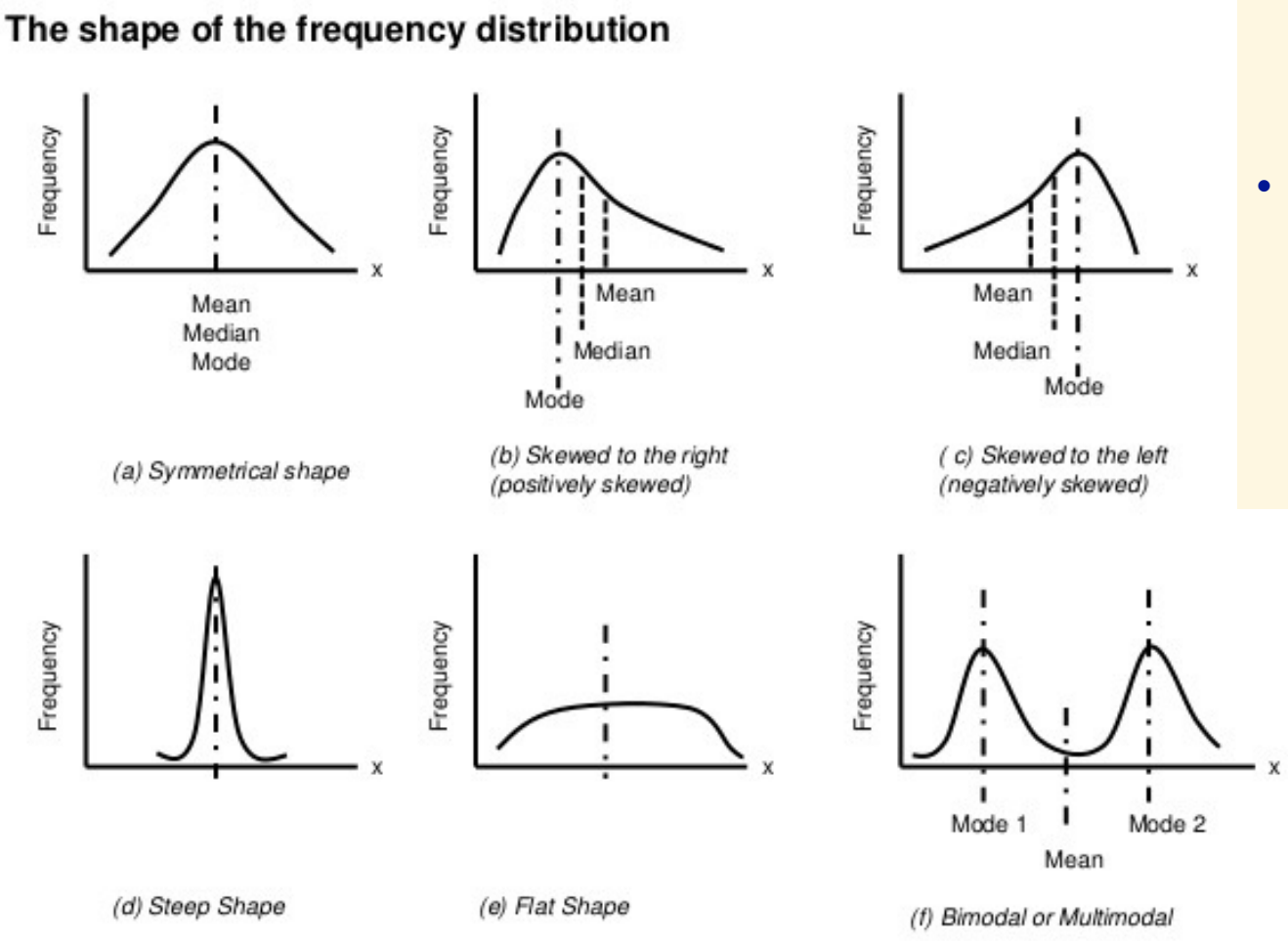


Ffrom: https://www.spss-tutorials.com/descriptive-statistics-one-metric-variable/

Using descriptive statistic

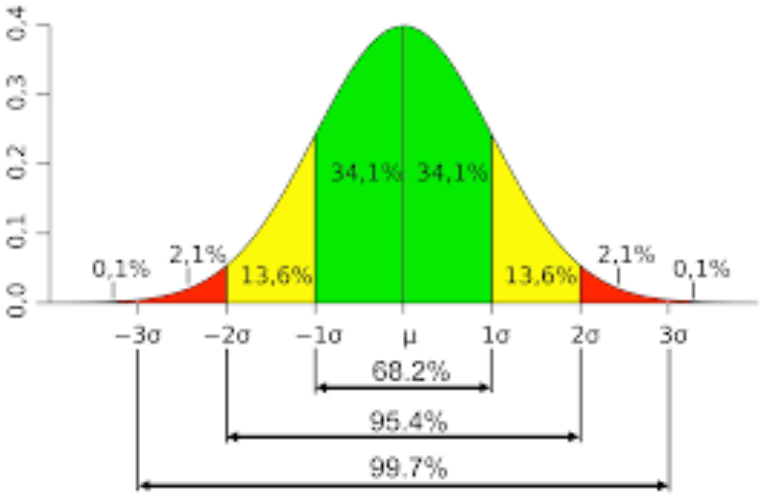
Common distributions

- With enough measurements, most variables are distributed normally
- But in order to fully describe data we need to introduce the idea of a standard deviation



Univariate analysis: dispersion measures

- **Variance:** The variance describes the spread of the data and measures how much the values of a variable differ from the mean. `var()`
- **Standard deviation:** it indicates how far a number of values lie apart from the mean. `std()`
- They measure the same in different scales
- When using most of these functions remember to use argument **`na.rm = T`**

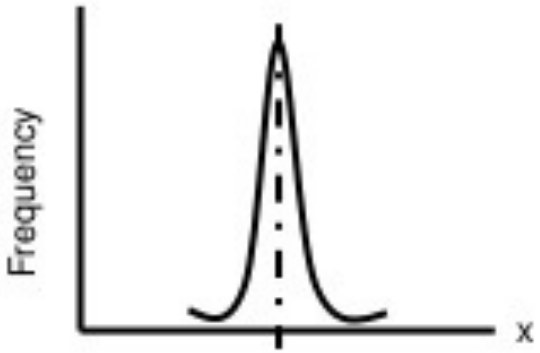


source http://projectmanager.com.au/can-you-use-standard-deviation-in-project-management/

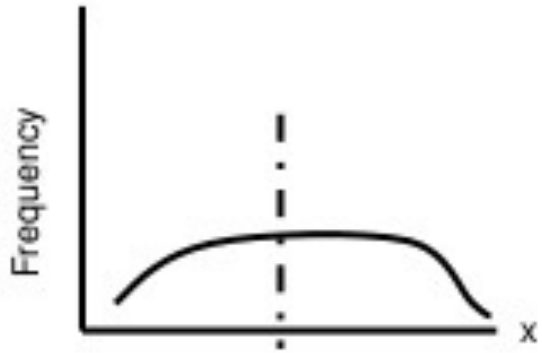
Interpreting Standard Deviation (SD)

SD lets you know about the distribution of scores around the mean.

- High SDs (relative to the mean) indicate the scores are spread out
- Low SDs tell you that most scores are very near the mean.



Low SD



High SD

Using descriptive statistic

Standarized scores (Z)

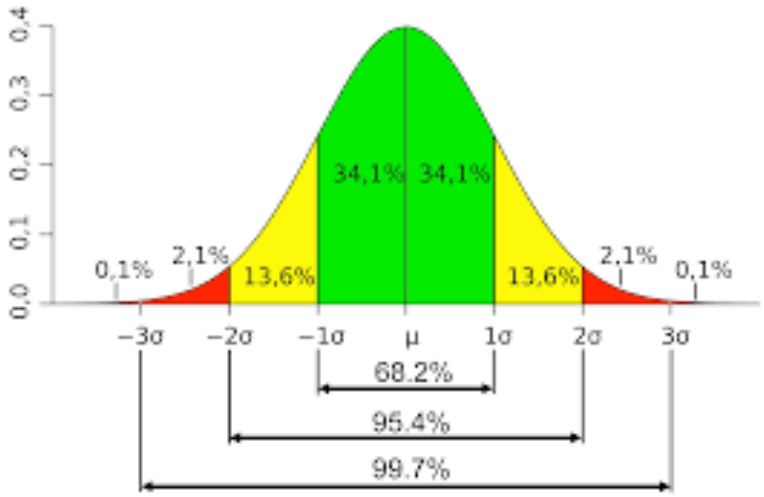
Used to force the scores onto a normal distribution

Subtract the mean from each score and divide by SD

$$Z = (X - mean)/SD$$

ALL Z-scores have a mean of 0 and SD of 1.

This allows the proportion of scores anywhere in the distribution.



source http://projectmanager.com.au/can-you-use-standard-deviation-in-project-management/

Normally distributed data, approximately 95% of the values lie within 2 sd of the mean.

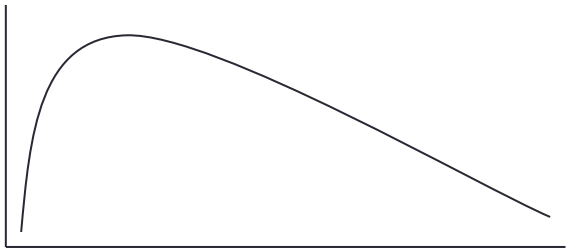
Using descriptive statistic

Skewness: symmetry of the distribution

Skewness is a measure of asymmetry.

- Values for skewness close to zero indicate that the shape of a frequency distribution for a variable approximates a normal distribution

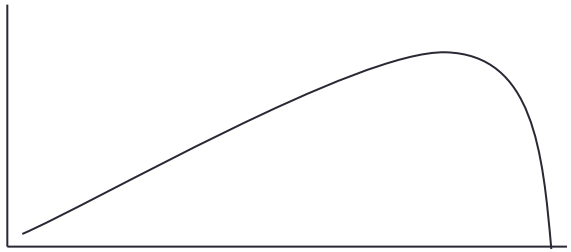
Positive skew



skewness score is positive;

Positive skew is reduced by using the square root or log transformation

Negative skew



skewness score is negative

Negative skew is reduced by squaring the data values

Using descriptive statistic

Skewness in R: measure of symetry

```
#Skewness and Kurtosis
library(moments)

> skewness(airquality$Ozone, na.rm=T)
[1] 1.225681

> agostino.test(airquality$Ozone)

D'Agostino skewness test

data:  airquality$Ozone
skew = 1.2257, z = 4.6564, p-value = 3.219e-06
alternative hypothesis: data have a skewness
```

Far away from zero. Suggest that variable is not normally distributed

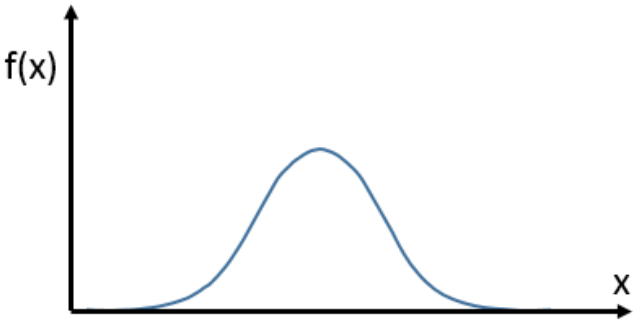
Small value <0.05, variable not normally distributed, accepted alternative hypothesis

Using descriptive statistic

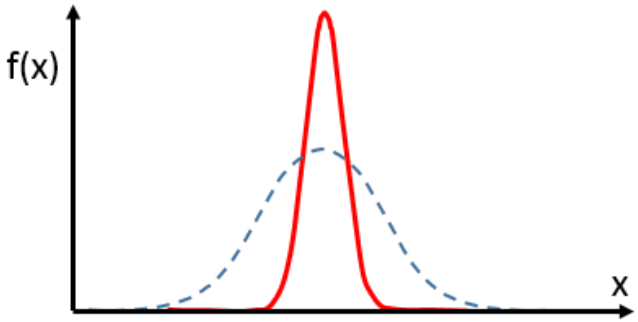
Using descriptive statistic

Kurtosis

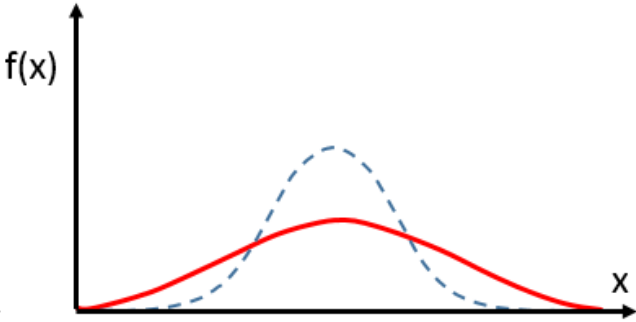
- The type of peak the distribution has can be characterized by a measurement called *kurtosis*.
- Values of kurtosis close to zero indicate that the shape of a frequency distribution approximates a normal distribution



Zero kurtosis
Gaussian distribution



Positive kurtosis



Negative kurtosis

Kurtosis in R: measure of outliers and heavy tails

```
#Kurtosis
library(moments)

> kurtosis(airquality$Ozone, na.rm=T)
[1] 4.184071

> anscombe.test(airquality$Ozone)

Anscombe-Glynn kurtosis test

data:  airquality$Ozone
kurt = 4.1841, z = 2.2027, p-value = 0.02762
alternative hypothesis: kurtosis is not equal to 3
```

Far away from zero. Suggest that variable is not normally distributed

Small value <0.05, variable not normally distributed, accepted alternative hypothesis: Probable presence of outliers

Using descriptive statistic

Univariate analysis: dispersion measures

- **Minimum**: `min()`
- **Maximum:** `max()`
- **Range:** The maximun difference in the data. `range()`

- When using most of these functions remember to use argument

```
na.rm = T
```

Using descriptive statistic

Using descriptive statistic

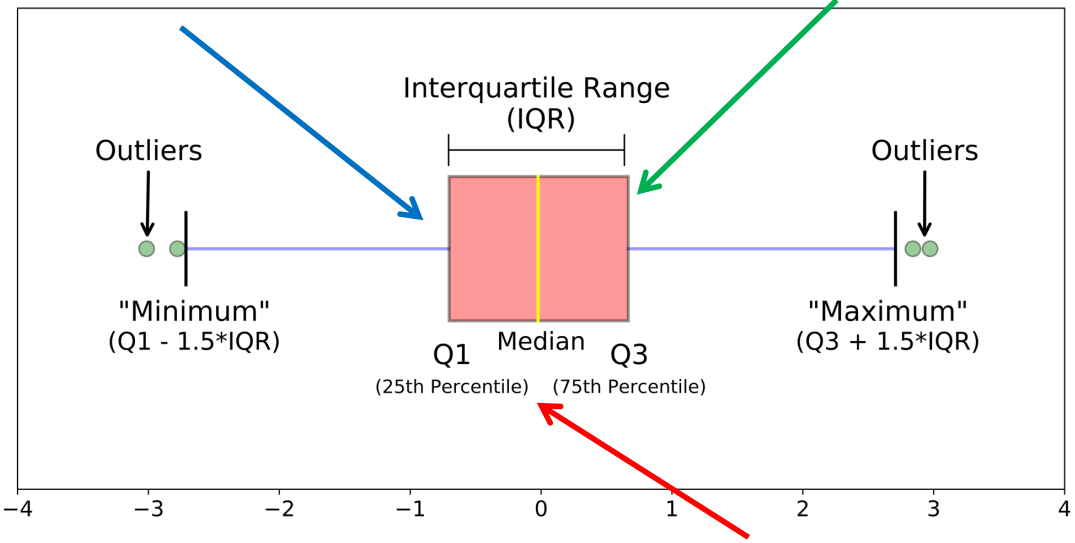
Univariate analysis: quartiles

- The quartiles of a population or a sample are the three values which divide the distribution or observed data into even fourths. `quantiles()`

- Quantiles divide data into equally size groups

0.25 or 50% quantile)

0.75 or 50% quantile)



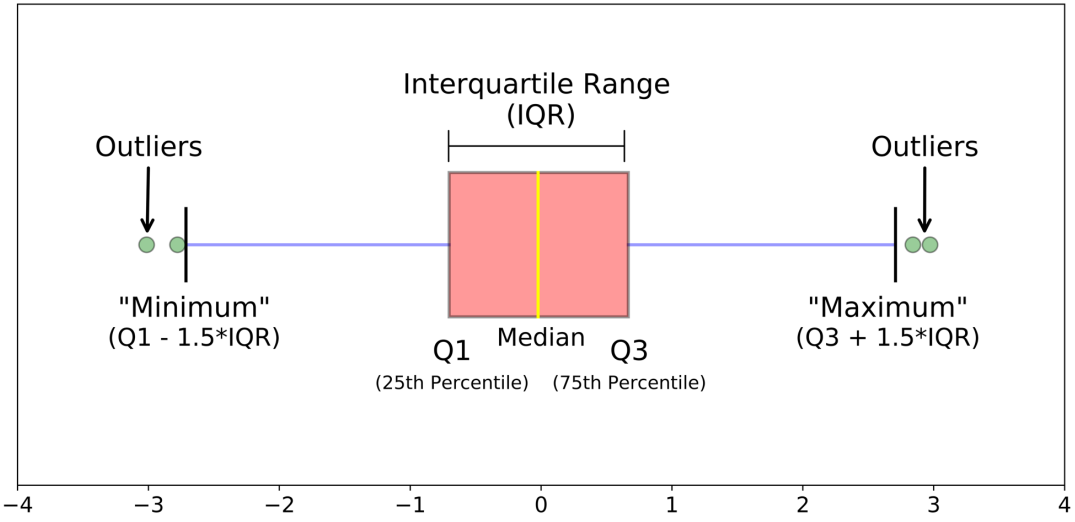
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Median 0.5 or 50% quantile)

Using descriptive statistic

Univariate analysis: quantiles

- In R quantiles have 9 different ways of being calculated. `quantiles()`
- `Small datasets tricky, large datasets they will do fair`
- remember to use argument **`na.rm = T`**



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Using descriptive statistic

Univariate analysis: percentiles (aka quantiles)

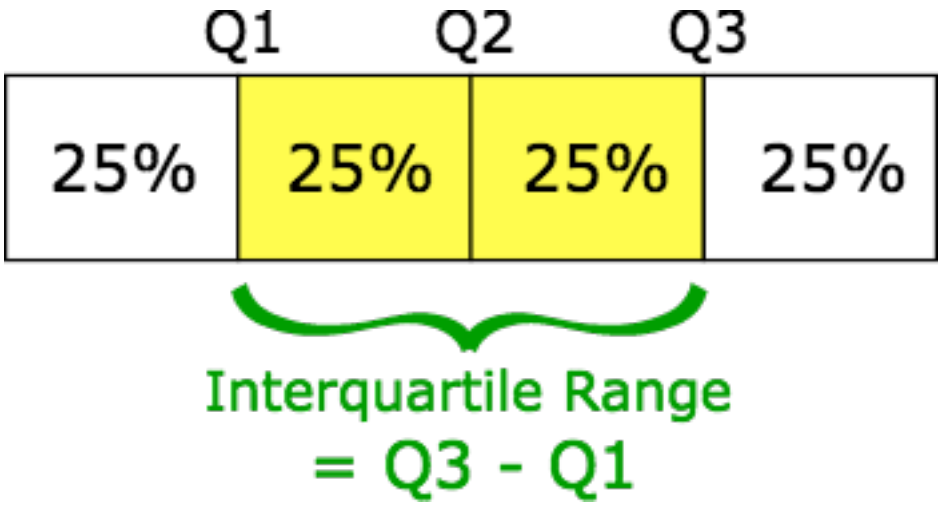
The *p*-quantile has the property that *p*% of the observations are less than or equal to it.

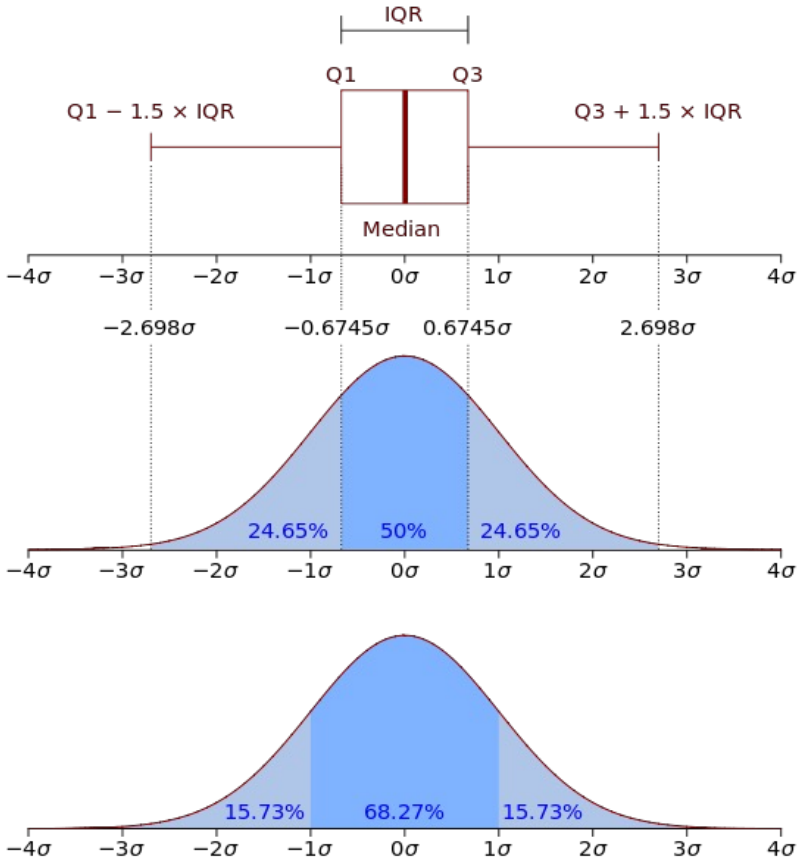
```
> set.seed(100)
> x <- rnorm(100, mean=0, sd=1)
> quantile(x)
        0%        25%        50%        75%       100%
-2.2719255 -0.6088466 -0.0594199  0.6558911  2.5819589
> quantile(x, probs=c(0.1, 0.2, 0.9))
       10%        20%        90%
-1.1744996 -0.8267067  1.3834892
```

Using descriptive statistic

Univariate analysis: inter quartile range

- The inter quartile range (IQR) is a more robust measure of spread `IQR()`
- remember to use argument **`na.rm = T`**





source wikipedia

Short Description of Descriptive Statistics and R Functions

Parameter	Description	R function
Mean	arithmetic average	`mean()`
Median	middle value, 50% quantile	`median()`
Mode	most frequent value	`mode()`
Standard Deviation	variation around mean	`sd()`
Quantiles	percent rank of values, such that all values are <= p	`quantile()`

Univariate analysis: Five numbers summary

The `summary()` function prints some basic descriptive statistics (including the count of missing values) for not only one, but also multiple variables, for example:

```
> summary(rage)

      rage
 Min.   :18.00
 1st Qu.:35.00
 Median :48.00
 Mean   :49.62
 3rd Qu.:64.00
 Max.   :97.00
 NA's   :3
```

R packages for summaries

```
library(skimr)
```

Basic descriptive stats of numeric variables and counts for categorical variables

```
skim(diamonds)
```

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>
1	carat	0	1	0.7979397	0.4740112
2	depth	0	1	61.7494049	1.4326213
3	table	0	1	57.4571839	2.2344906
4	price	0	1	3932.7997219	3989.4397381
5	x	0	1	5.7311572	1.1217607
6	y	0	1	5.7345260	1.1421347
7	z	0	1	3.5387338	0.7056988

R packages for summaries

```
library(gtsummary)
```

tbl_summary() :

summarizes categorical variables by **counts and percentages**,
summarizes numeric variables by **median and IQR**

add_p(): **conducts statistical tests with all variables and provides p-values**

numeric variables:

two groups comparison: non-parametric Wilcoxon rank sum test
more than two groups:Kruskal-Wallis rank sum test

categorical variables:

Fischer`s exact test when n<5
Pearson’s Chi-sqared test when n=>5

See example

Hmisc package

- In order to know more about the dataset such as the missing values, distribution of numerical variables, and distinct values of categorical variables, we can use an additional package called Hmisc

```
> library (Hmisc)
> describe(iris)
```

```
5  Variables      150  Observations
---------------------------------------------------------------------------------------------
Sepal.Length
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75
     150        0       35    0.998    5.843   0.9462    4.600    4.800    5.100    5.800    6.400
     .90      .95
   6.900    7.255

lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
Species
       n  missing distinct
     150        0        3

Value          setosa versicolor  virginica
Frequency          50         50         50
Proportion      0.333      0.333      0.333
---------------------------------------------------------------------------------------------
```

Explore normality: a word about Normal distribution

- Means and variances are ways to describe a *distribution* of scores.
- Knowing about your distributions is one of the best ways to understand your data
- A NORMAL (Gaussian) distribution is the most common assumption of statistics, thus it **is important to check if your data are normally distributed.**

Check normality!!!!

If the distributions of our variables are not approximately normal we violate the assumptions of most used statistical tests

We must examine each variable's distribution and make adjustments when necessary so that normality assumptions are met.

Distribution	Test type	2 groups	> 2 groups
Normal	Parametric	T-test	ANOVA
Not normal	Non-parametric	Mann-Withney	Kruskal-Wallis

How to check normality?

- Visual inspection, QQplots
- **significance test** comparing the sample distribution to a normal one
- There are several **normality tests:**
  - **Kolmogorov-Smirnov (K-S)**
  - **Shapiro-Wilk's test**. * recommended better power
- 8

How to check normality?

- Visual inspection, **QQplots**
- **Significance test** comparing the sample distribution to a normal one

  There are several **normality tests:**

  - **Kolmogorov-Smirnov (K-S)**
  - **Shapiro-Wilk's test**. * recommended better power

- Both are needed

QQ-plot

- The quantile-quantile plots are often used to determine whether a dataset is normally distributed

- The QQ-plot shows the theoretical quantiles versus the empirical quantiles. If the distribution assumed (theoretical one) is indeed the correct one, we should observe a straight line.

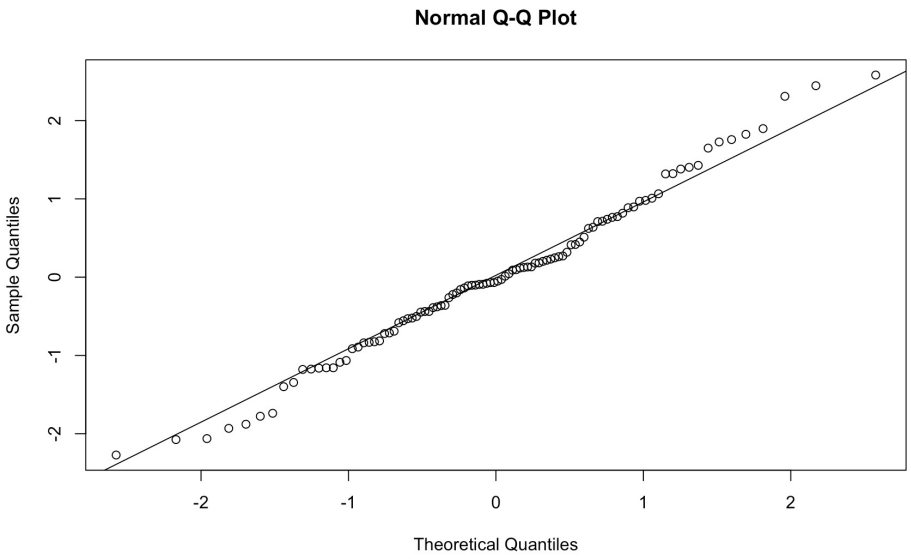
QQ-plot

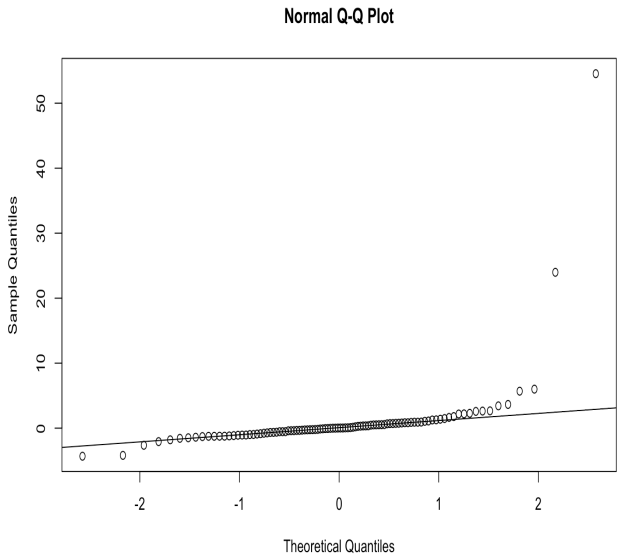
Distribucion normal

```
x<-rnorm(100, mean=0, sd=1)
library(ggplot2)
#create Q-Q plot
ggplot(df, aes(sample=x)) +
stat_qq() +
stat_qq_line()
```

T Distribucion with 2 df

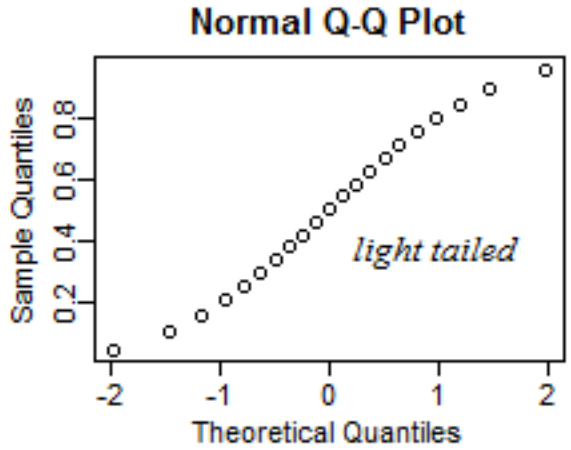
```
set.seed(100)
x<-rt(100,df=2)
qqnorm(x)
qqline(x)
```

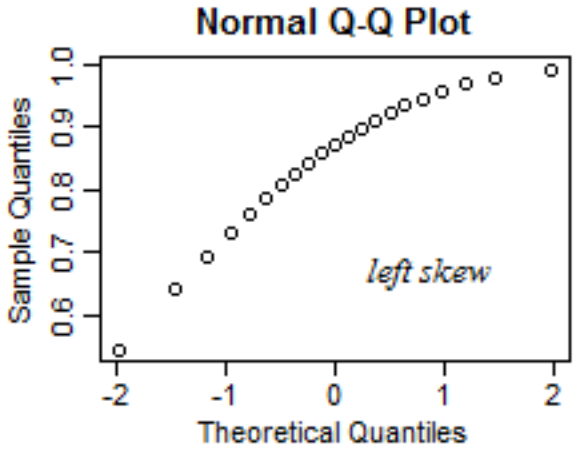


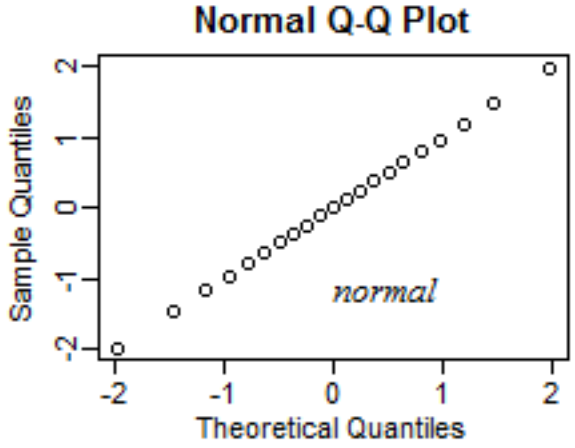


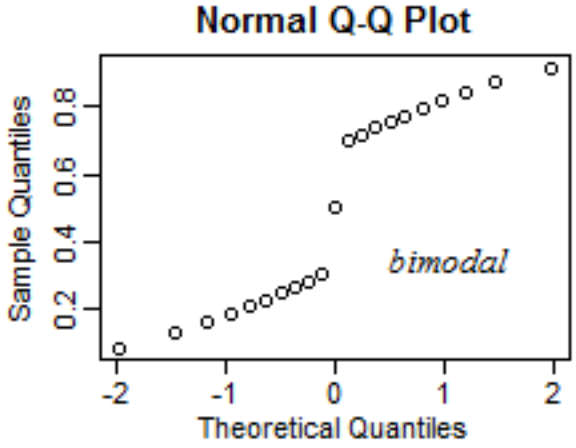
qqnorm() computes theortical normal quantiles

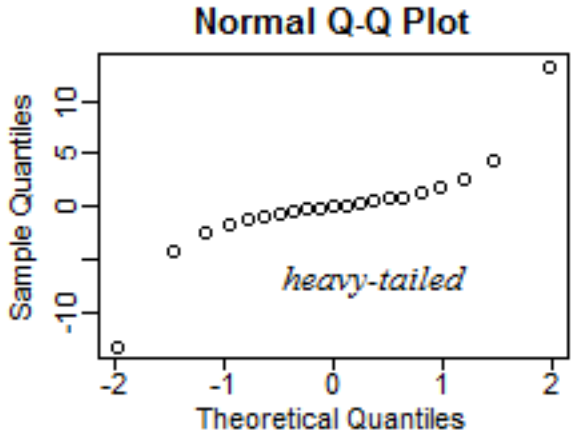
QQ-plot

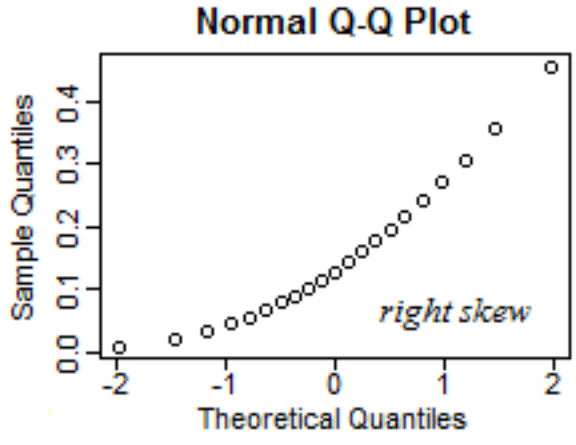












Source: https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot

Normality test

```
shapiro.test(murders$population)

        Shapiro-Wilk normality test
data:  murders$population
W = 0.7174, p-value = 1.372e-08
```

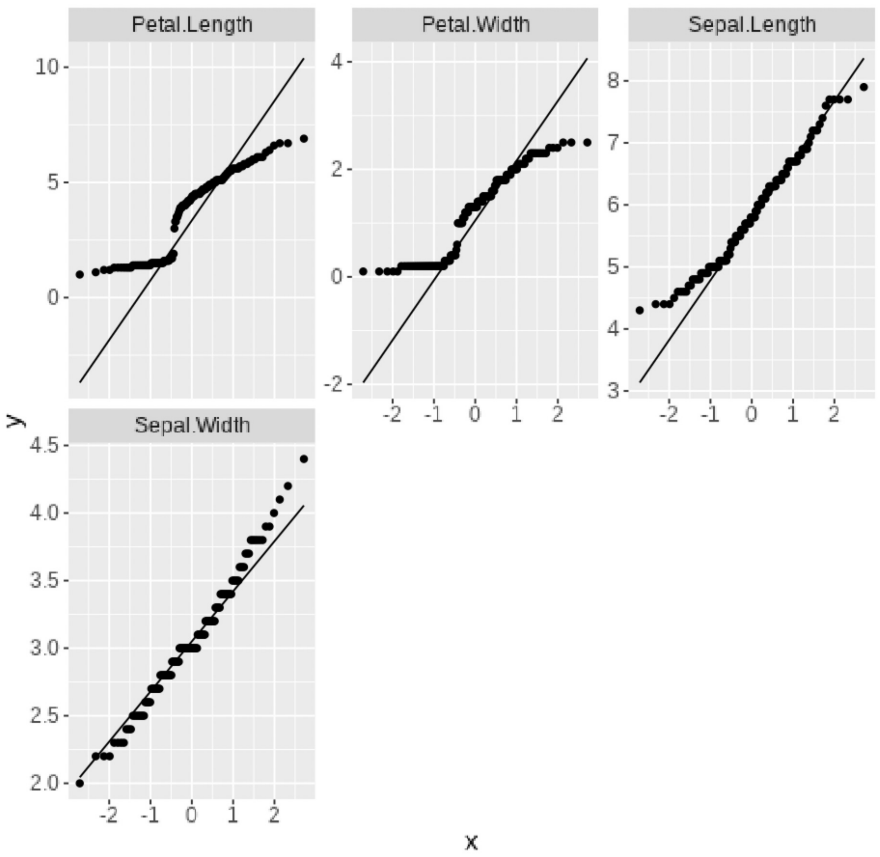
, the **p-value < 0.05** implies that the distribution of the data are significantly different from normal distribution. In other words, we CAN NOT assume the normality.

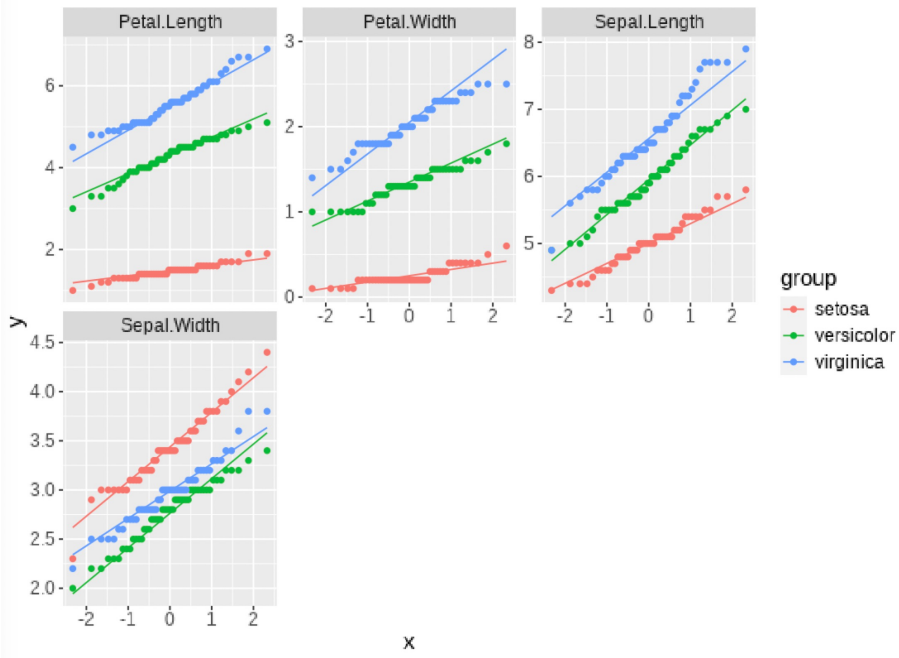
The null hypothesis of these tests is that “sample distribution is normal”.

If the test is **significant**, the distribution is non-normal.

R packages for normality

```
library (DataExplorer)
        plot_qq(iris)
        plot_qq(iris, by="Species")
```



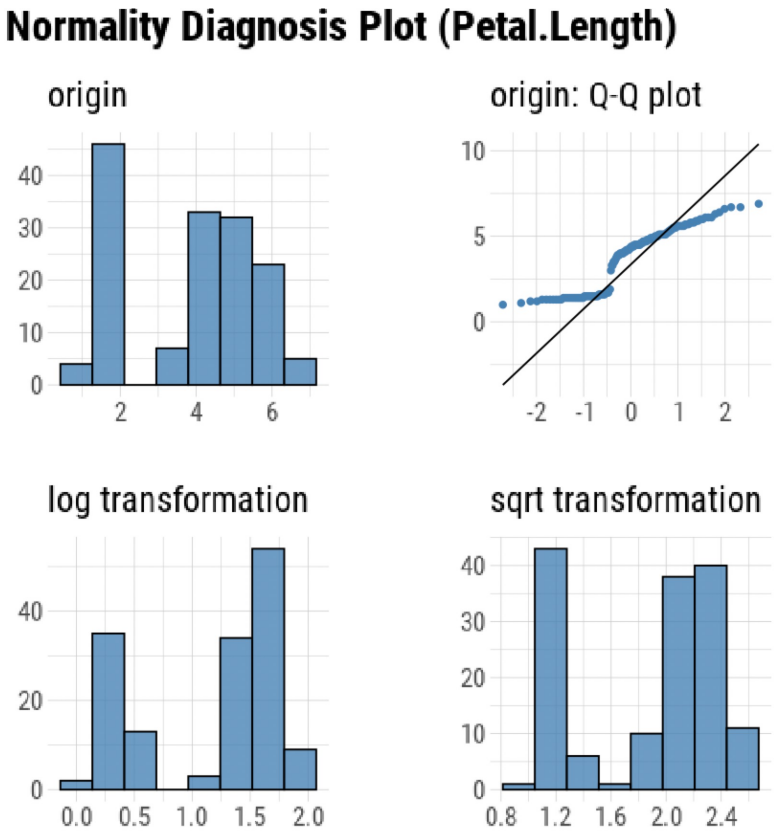


R packages for normality

```
library(dlookr)
```

plot_normality(): *QQ-plot, original histogram, histogram with two common tranformations if normality assumption is not met.*

```
iris %>%
group_by(Species) %>%
plot_normality(Petal.Length)
```



Comparison of Statistical Analysis Tools for Normally and Non-Normally Distributed Data

Tools for Normally Distributed Data	R functions	Equivalent Tools for Non-Normally Distributed Data	R functions
T-test	`t.test()`	Mann-Whitney test; Mood's median test; Kruskal-Wallis test	`wilcox.test()`; `mod.medtest()`; `kruskal.test()`
ANOVA	`aov()`	Mood's median test; Kruskal-Wallis test	`mod.medtest()`; `kruskal.test()`
Paired t-test	`t.test(x, y, paired = TRUE)`	One-sample sign test	`SIGN.test()`
F-test; Bartlett's test	`var.test()`; `bartlett.test()`	Levene's test	`Levene.test()`

Statistical Distributions

- The names of the R functions for distributions comprise two parts:
  - the first letter indicates the "function group",
  - the second indicates the distribution.

distribution	R name	*distribution*	R name	*distribution*	R name
normal	`norm`	t	`t`	χ^2	`chisq`
exponential	`exp`	f	`f`	uniform	`unif`
log-normal	`lnorm`	beta	`beta`	gamma	`gamma`
logistic	`logis`	weibull	`weibull`	cauchy	`cauchy`
geometric	`geom`	binomial	`binom`	hypergeometric	`hyper`
poisson	`pois`	negative binomial	`nbinom`		

- *Common distributions have their corresponding R "names":*

Contingency tables

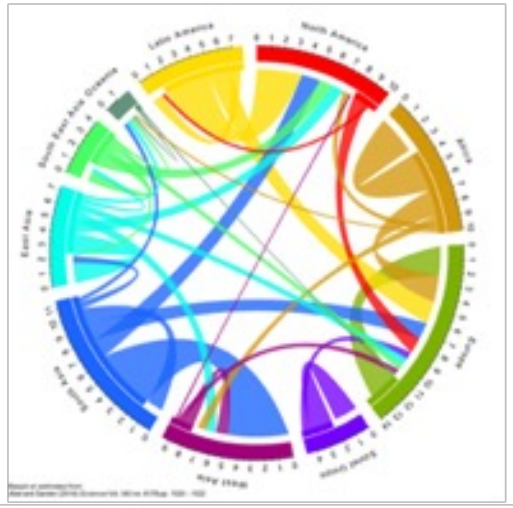
Categorical

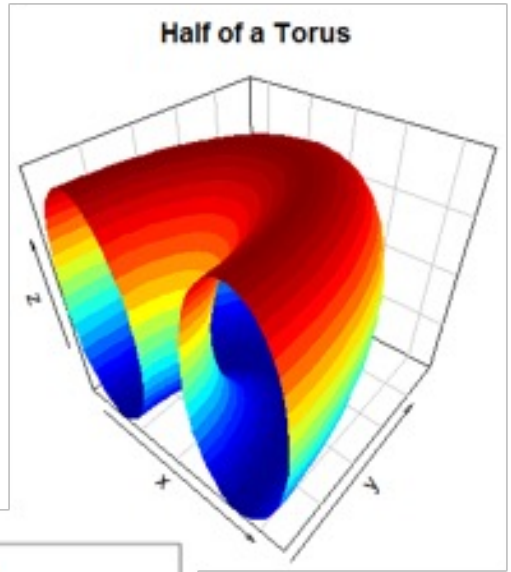
- For categorical data we can use contingency tables.

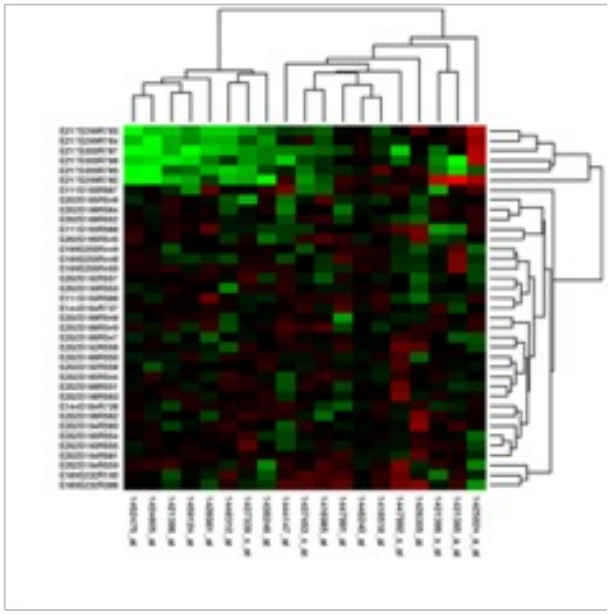
```
table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

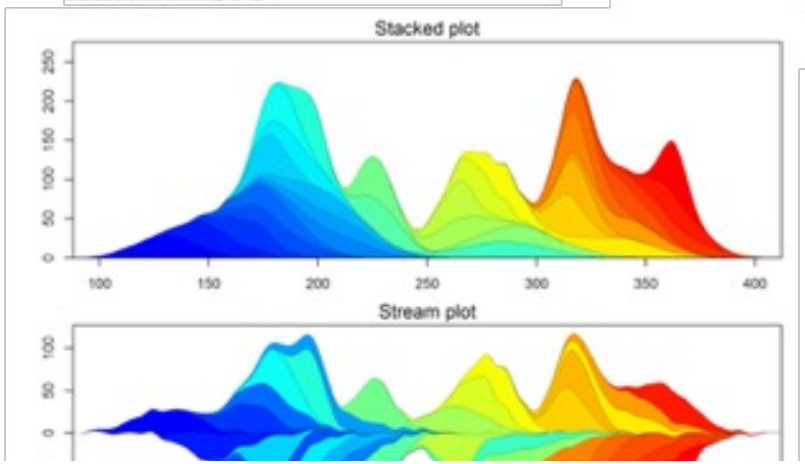
Data Visualization

Gallery of plots









Visualization

Why graphs in EDA?

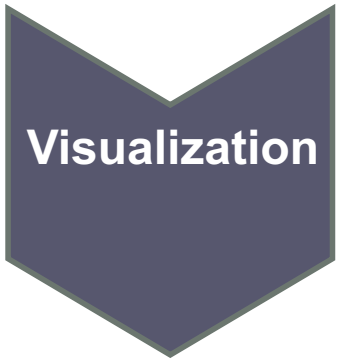
- To understand data properties
- To find patterns
- To suggest modelling strategies
- To “debug” data
- To communicate results

Data Visualization

- A graphic should display as much information as it can
- Strive for clarity.
- Make the data stand out:
  - Avoid too many superimposed elements, such as too many curves in the same graphing space.
  - Find the right aspect ratio and scaling to properly bring out the details of the data.
  - Avoid having the data all skewed to one side or the other of your graph.

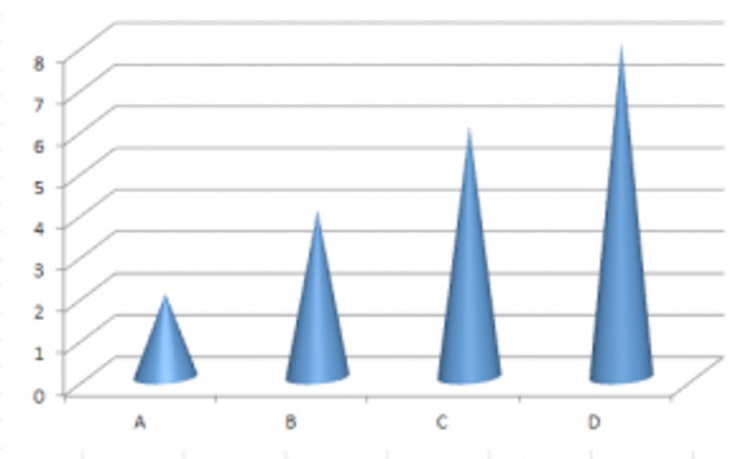
Differences between graphs for EDA and for results communication

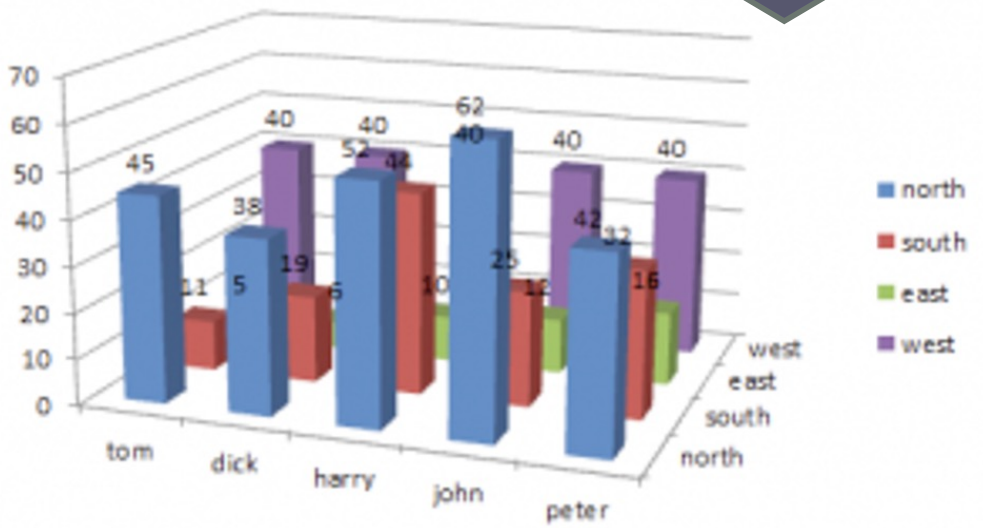
- They are made quickly and in big numbers
- They are thought to get understanding of the data
- Axis and legends are cleaned up later

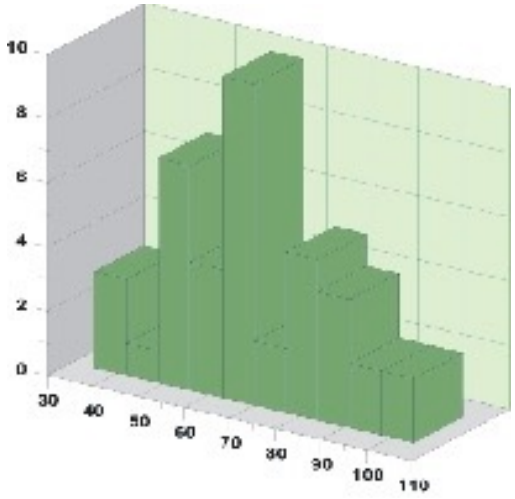


Visualization

Examples of bad plots



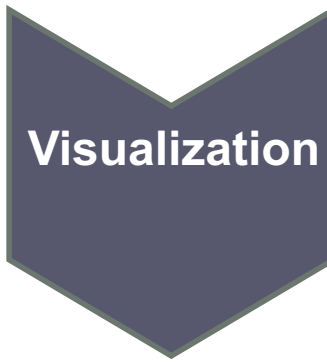




https://www.forbes.com/sites/naomirobbins/2012/05/30/winner-of-the-bad-graph-contest-announced-2/#35e9d5632e06

Univariate data: Graphical analyisis

- Univariate plots are plots of individual attributes without interactions.
- The goal is to learn something about the distribution, central tendency and spread of each attributes.
  - Histogram
  - Density plots
  - Box plots
  - Bar plots



Histogram

- Each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
- It gives a general impression of the distribution's shape

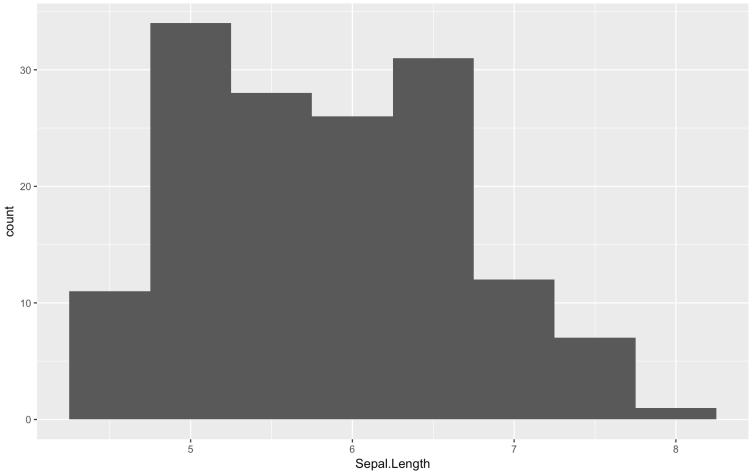
- To construct a histogram, define the range of data for each bar (called a **bin**).
- Generally one will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution.
- It is often worthwhile to try a few different bin sizes/numbers especially with small samples

Histogram

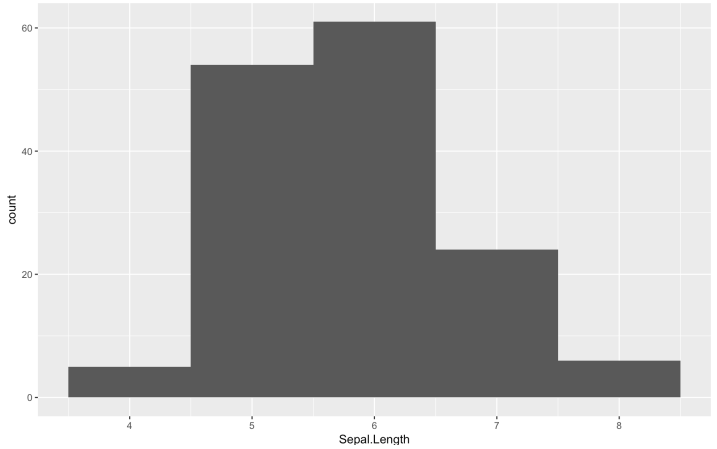
- With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

```
> ggplot(iris, aes(Sepal.Length))
+ geom_histogram(binwidth = 0.5)
> ggplot(iris, aes(Sepal.Length))
+ geom_histogram(binwidth = 1)
> ggplot(iris, aes(Sepal.Length))
+ geom_histogram(binwidth = 2)
```

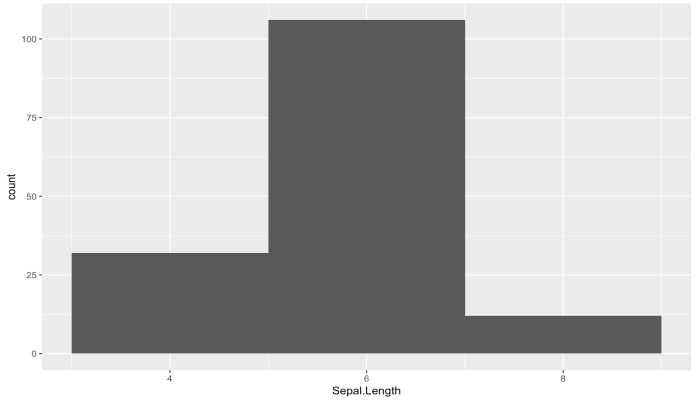
Bin=0.5



Bin=1



Bin=2



Issues with Histograms

- For small data sets, histograms can be misleading.
- 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
  - Be careful with axes and scales

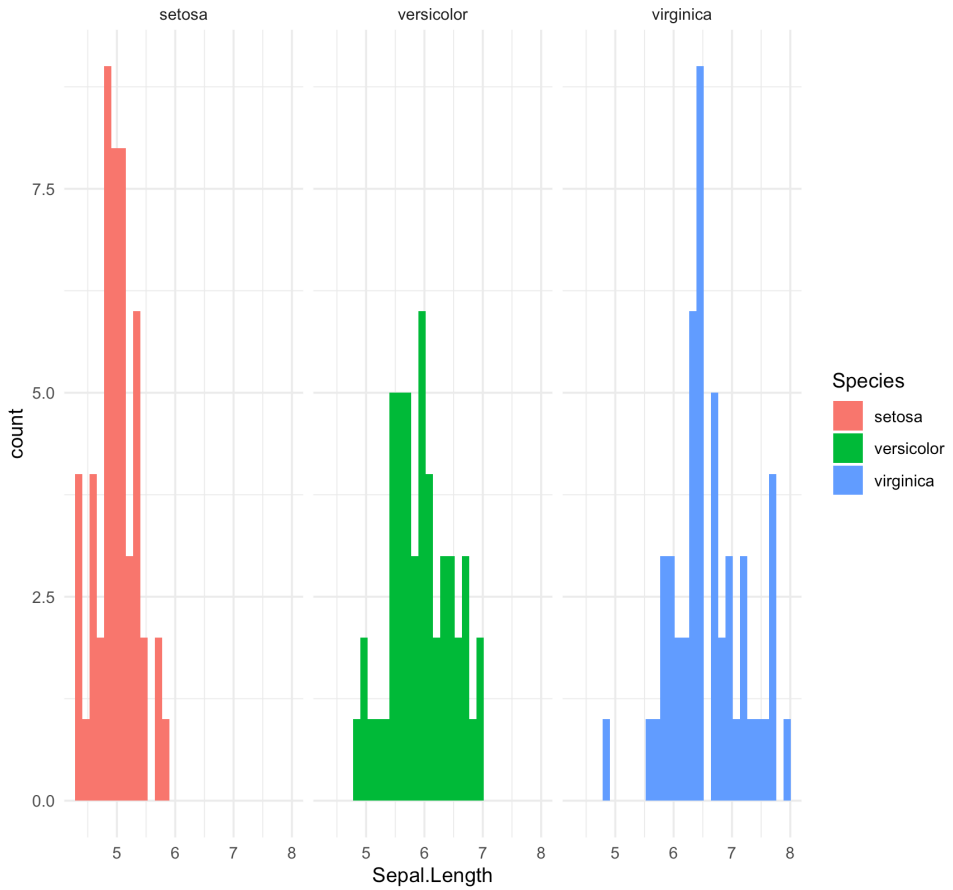
Histogram

```
ggplot(data=iris, aes(x=Sepal.Length, fill=Species))
+ geom_histogram() +
+ theme_minimal() +
+ facet_wrap(~Species)
```

Histograms effectively only work with 1 variable at a time

But ‘small multiples’ can be effective

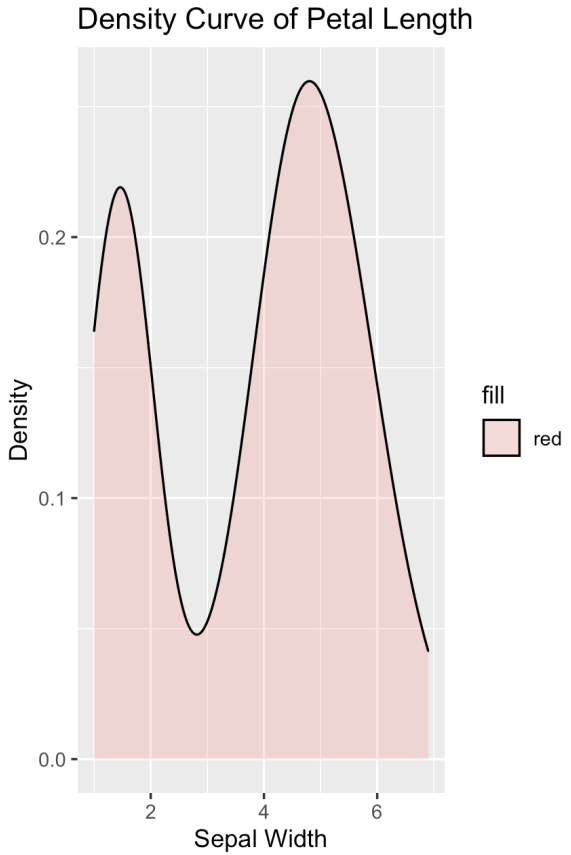
Be careful with axes and scales



Density curve

- We can smooth out the histograms to lines using a density plot.
- These are useful for a more abstract depiction of the distribution of each variable.

```
ggplot(data=iris, aes(x=Petal.Length, fill="red"))+
geom_density(stat="density", alpha=I(0.2)) +
xlab("Sepal Width") +  ylab("Density") +
ggtitle("Density Curve of Petal Length")
```



Bar graphs

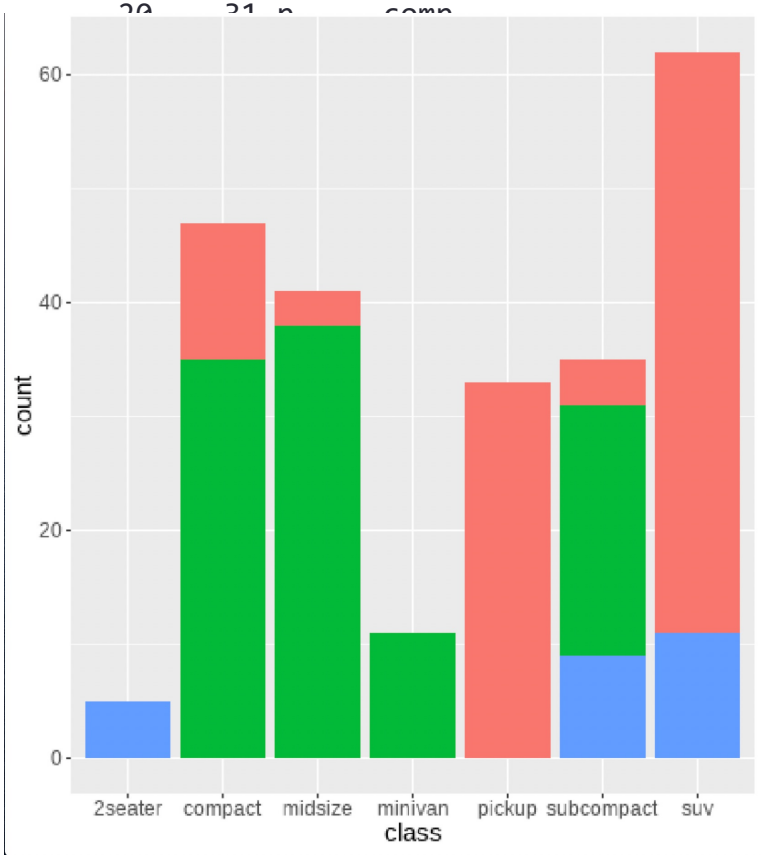
- Bar charts are a common visual tool for displaying a single categorical variable.
- Categories are listed on the x-axis, and frequencies or proportions on the y-axis.
- The height of each bar represents either counts or percentages
- Easier to compare categories with bar graph than with pie chart

Visualization

Bar graphs

```
> mpg
 manufacturer model      displ  year   cyl trans      drv      cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     comp…
 2 audi         a4           1.8  1999     4 manual(m…  f        21    29 p     comp…
 3 audi         a4           2    2008     4 manual(m…  f        20    31 p     comp…
 4 audi         a4           2    2008     4 auto(av)   f
 5 audi         a4           2.8  1999     6 auto(l5)   f
 6 audi         a4           2.8  1999     6 manual(m…  f
 7 audi         a4           3.1  2008     6 auto(av)   f
```

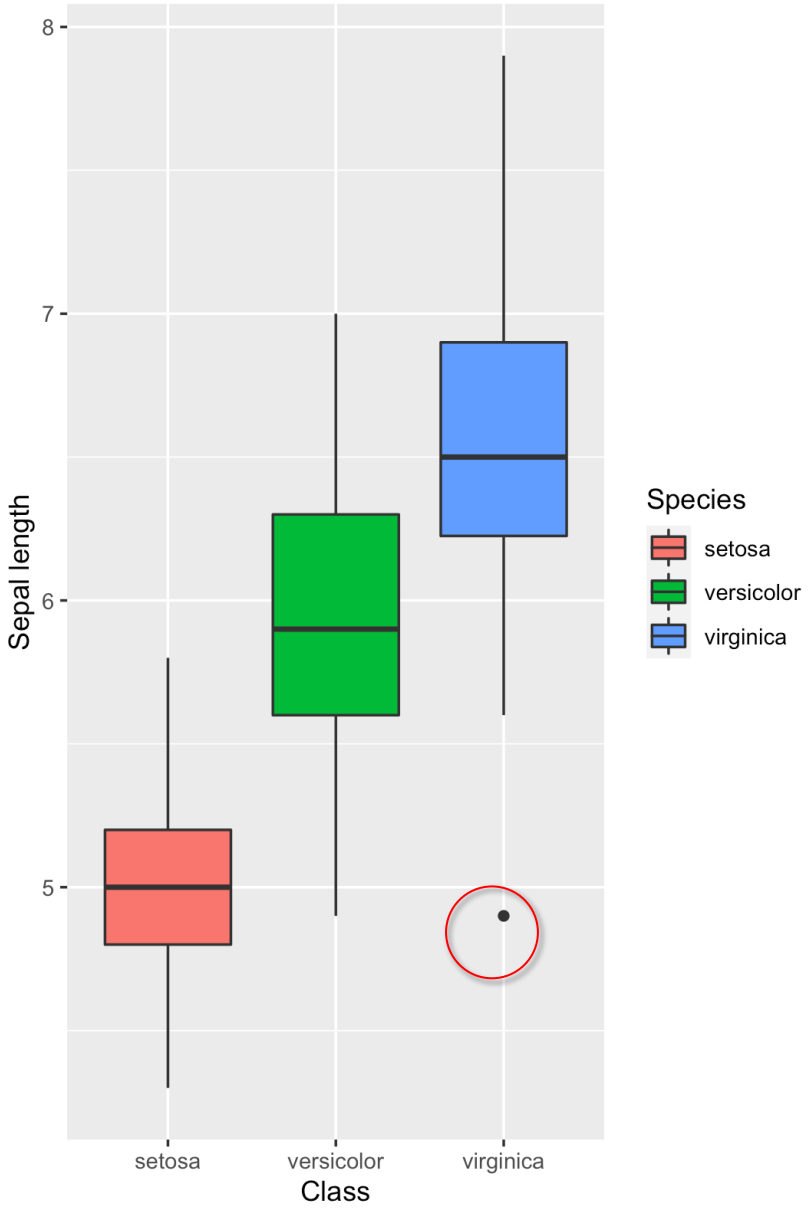
```
ggplot(mpg, aes(class,fill=drv))+
   geom_bar()
```



Boxplot

- Boxplots show robust measures of location and spread as well as providing information about symmetry of a frequency distribution and outliers.
- Is the way to visualize the five-number summary.

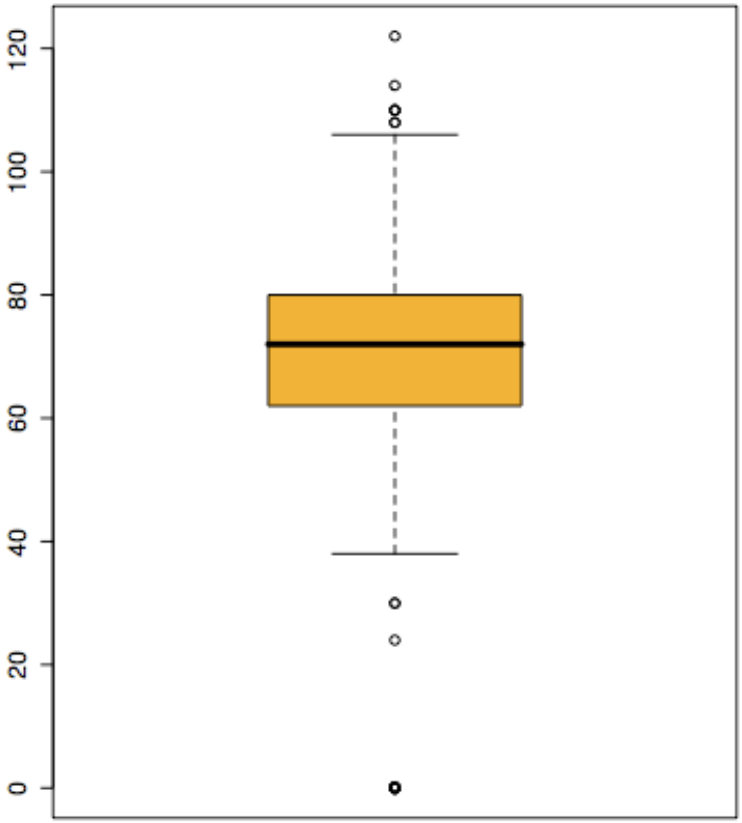
```
library(ggplot2)
ggplot(iris, aes(x=Species , y=Sepal.Length)) +
        geom_boxplot(aes(fill=Species), position = "dodge") +
        xlab("Class") +
        ylab("Sepal length")
```



Boxplots

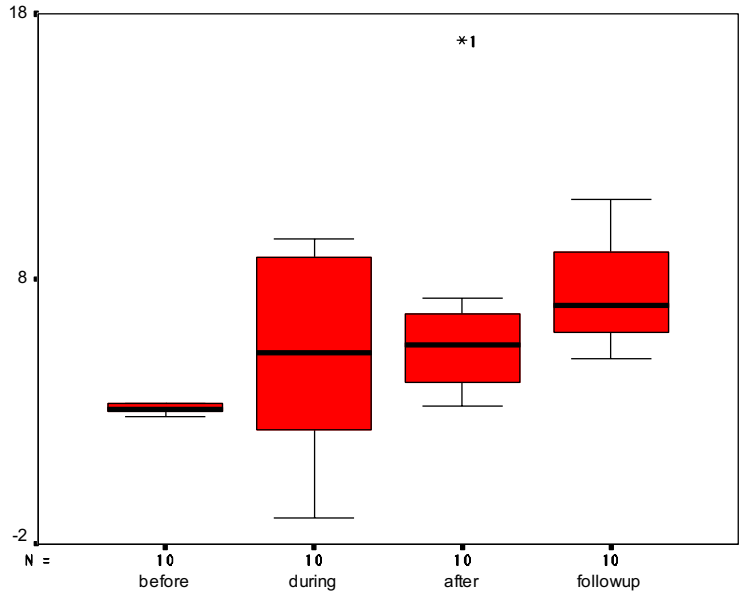
Negative issues:

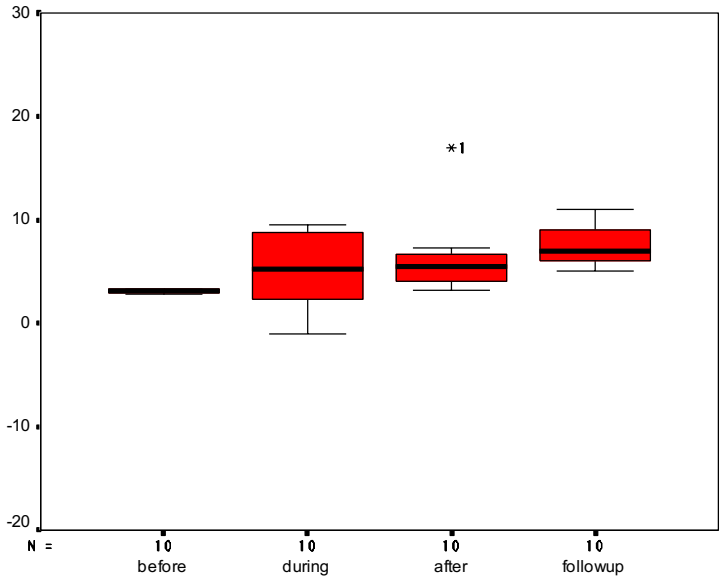
- Overplotting
- Hard to tell distributional shape
- no standard implementation in software (many options for whiskers, outliers)



Scales

- It is very important to pay attention to the *scale* that you are using when you are plotting.

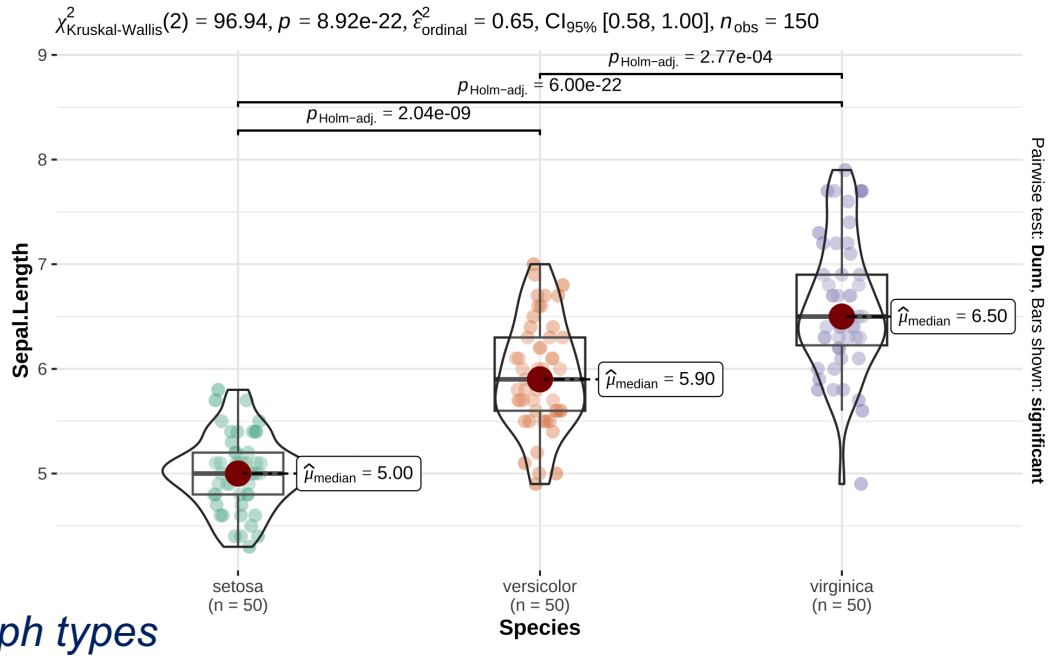




Significative differences in plots?. ggstatsplot

- Are there significant differences between grouos
- Between which groups are signficant differences (pairwise comparisons)
- Is it still significant the difference if we use an adjusted p-value?

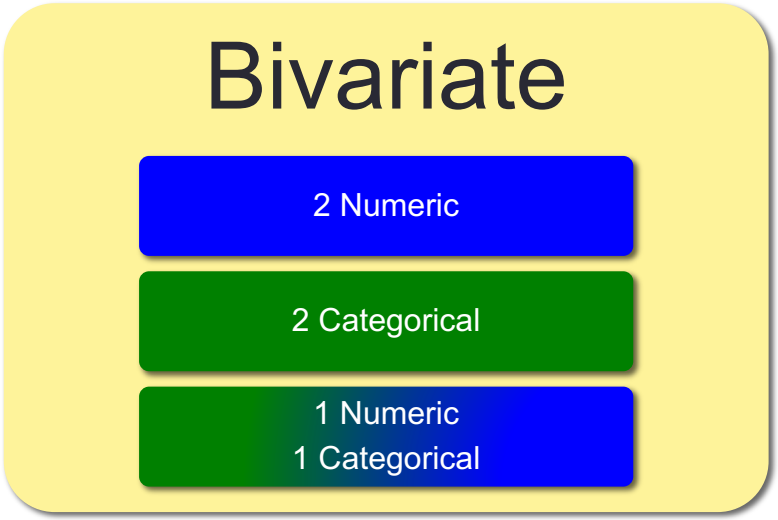
```
library(ggstatsplot)

ggbetweenstats(
  data = iris,
  x    = Species,
  y    = Sepal.Length,
  type = "np")
```



Different functions for different graph types

2. Bivariate analysis

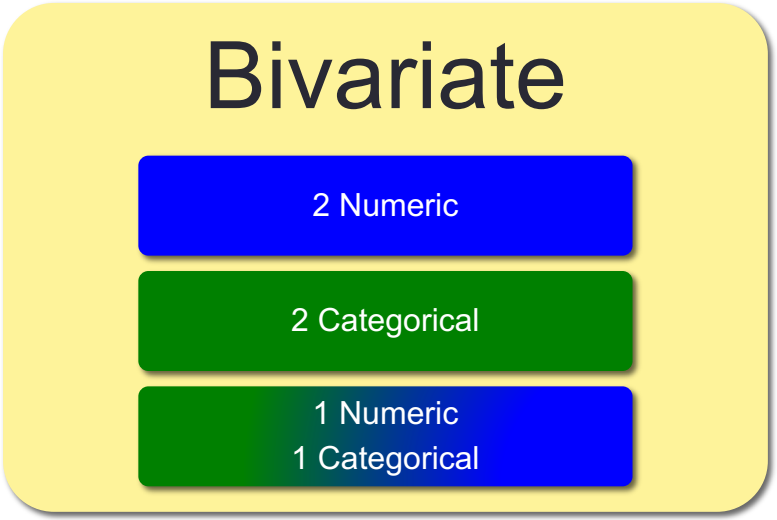
Using descriptive statistic



Bivariate Quantitative analysis

Bivariate analysis include:

- Crosstabs
- Covariance
- Correlation



Advanced techniques include:

- Cluster analysis
- Analysis of variance (ANOVA)
- Factor analysis
- Principal component analysis (PCA)

Using descriptive statistic

Bivariate Quantitative: Contingency tables

2 Categorical variables

Contingency tables provide a way to display the frequencies and relative frequencies of observations classified according to two categorical variables.

- The elements of one category are displayed across the columns; the elements of the other category are displayed over the rows.

- The basic table types supported by crosstab() are:
  - *f*requency - frequency count
  - *r*ow.pct - proportion within row
  - *c*ol.pct - proportion within column
  - *j*oint.pct - proportion within final 2 dimensions of table
  - *t*otal.pct - proportion of entire table

Using descriptive statistic

Bivariate Quantitative: Contingency tables

```
library(gmodels)
data(infert, package = "datasets")
CrossTable(infert$education, infert$induced, prop.t=TRUE, prop.r=TRUE,
prop.c=TRUE)
```

```
 Cell Contents
|-------------------------|
|                       N |
|              Expected N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

infert$education	infert$induced: 0	1	2	Row Total
0-5yrs	4	2	6	12
	1.232	0.506	9.898	
	0.333	0.167	0.500	0.048
	0.028	0.029	0.162	
	0.016	0.008	0.024	
6-11yrs	78	27	15	120
	1.121	1.059	0.471	
	0.650	0.225	0.125	0.484
	0.545	0.397	0.405	
	0.315	0.109	0.060	
12+ yrs	61	39	16	116
	0.518	1.627	0.099	
	0.526	0.336	0.138	0.468
	0.427	0.574	0.432	
	0.246	0.157	0.065	
Column Total	143	68	37	248
	0.577	0.274	0.149	

Test available:
Chiq, Fisher and Macnemar

Bivariate Quantitative: Covariation

2 Numerical variables

- The covariance expresses how much two numeric variables “change together” and the nature of that relationship, whether it is positive or negative

- The R commands `cov()` is used for the sample covariance; you need only to supply the two corresponding vectors of data.

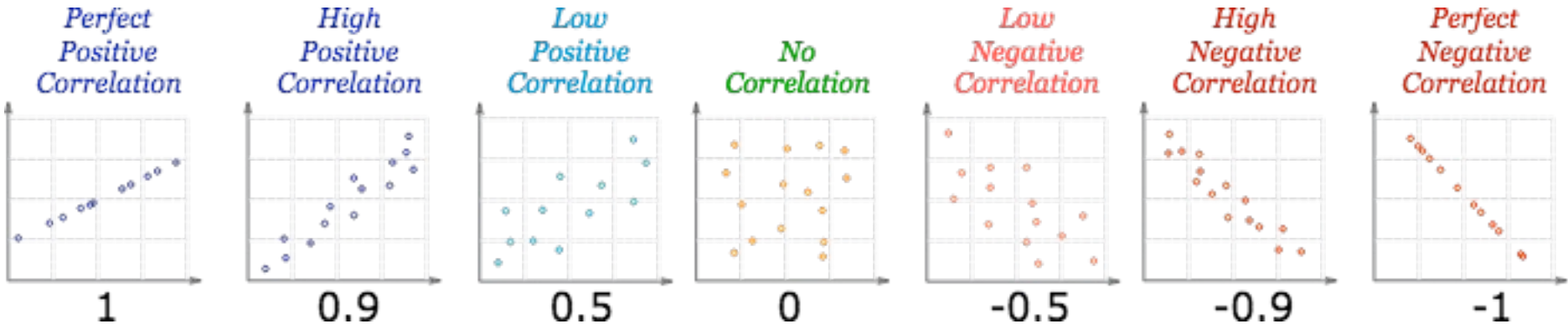
```
xdata <- c(2,4.4,3,3,2,2.2,2,4)
ydata <- c(1,4.4,1,3,2,2.2,2,7)
cov(xdata,ydata)
[1] 1.479286
```

- A positive `cov()` result shows a positive linear relationship—as x increases, y increases.
- A negative result, it shows a negative linear relationship
- A covariance = 0 indicates that there is no linear relationship

Bivariate Quantitative: Correlation

2 Numerical variables

- Correlation allows you to interpret the covariance further by identifying both the direction and the strength of any association.



Technically, independence implies zero correlation, but the reverse is not necessarily true.

Source: https://www.mathsisfun.com/data/correlation.html

Bivariate Quantitative: Correlation

2 Numerical variables

- `cor()` computes the correlation coefficient
- `cor.test()` test for association/correlation between paired samples.
- It returns both the correlation coefficient and the significance level(or p-value) of the correlation .

```
cor(x, y, method = c("pearson", "kendall", "spearman"))
cor.test(x, y, method=c("pearson", "kendall", "spearman"))
```

Pearson's corr. Coefficient is the most common
Kendall NP. Advised when the data does not come from a bivariate normal distributio
Spearman more robust if we know there are outliers

Bivariate Quantitative: Correlation

Pearson correlation coefficient

- Both variables should be normally distributed
- Both variables should have a linear relationship
- Assumes that data is equally distributed about the regression line (homoscedasticity)

```
cor(x, y, method = c("pearson"))
cor.test(x, y, method=c("pearson))
```

Using descriptive statistic

Bivariate Quantitative: Correlation example

Using descriptive statistic

```
# Extract the p.value
res$p.value
# Extract the correlation coefficient
res$estimate
```

```
library(datasets)
my_data <- mtcars
res <- cor.test(my_data$wt, my_data$mpg,
                method = "pearson")
```

```
data:  my_data$wt and my_data$mpg
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 -0.9338264 -0.7440872
sample estimates:
       cor
-0.8676594
```

t is the t-test statistic value
(t = -9.559

df degrees of freedom (df= 30),

p-value significance level of the t-test
p-value = 1.294e-10 < alpha=0.05)
➔ significative correlated

conf.int is the confidence interval of the correlation coefficient at 95% (-0.9338264 ,-0.7440872)

Bivariate Quantitative: Correlation

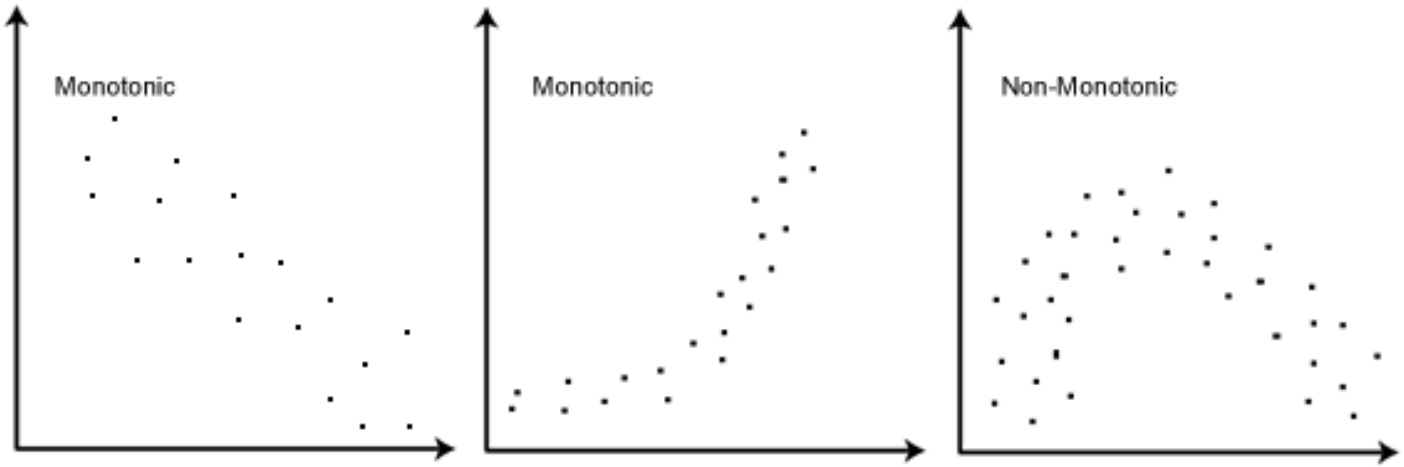
Spearman correlation coefficient

- It is s the nonparametric version of the Pearson product-moment correlation.
- It measures the strength and direction of association between two ranked variables.
- You need two variables that are either ordinal, interval or ratio (e.g ordering levels of education)
- Spearman's correlation determines the strength and direction of the **monotonic relationship** between your two variables

Using descriptive statistic

Monotonic relationship

- A monotonic relationship is a relationship that does one of the following:
  - (1) as the value of one variable increases, so does the value of the other variable;
  - or (2) as the value of one variable increases, the other variable value decreases.



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Bivariate Quantitative: Correlation

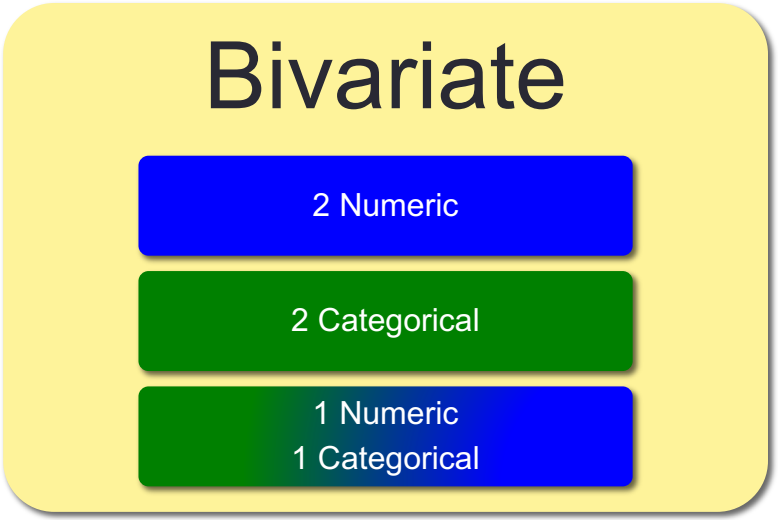
Kendall rank correlation coefficient

- Kendall rank correlation is a non-parametric test
- It assess statistical associations based on the ranks of the data.
- It does not carry any assumptions about the distribution of the data
- The interpretations of Kendall's tau and Spearman's rank correlation coefficient are very similar and thus invariably lead to the same inferences.

Test and typical questions

Question	Test
Are two variables (n = 2) are correlated (i.e., associated)	**Correlation test between two variables**
Are multiple variables (n > 2) are correlated	**Correlation matrix between multiple variables**
Are two groups (n = 2) of samples differ from each other	**Comparing the means of two groups:**
	Student's t-test (parametric)
	Wilcoxon rank test (non-parametric)
Are multiple groups (n >= 2) of samples differ from each other	**Comparing the means of more than two groups**
	ANOVA test (analysis of variance, parametric): extension of t-test to compare more than two groups.
	Kruskal-Wallis rank sum test (non-parametric): extension of Wilcoxon rank test to compare more than two groups
Are the variability of two samples differ	**Comparing the variances:**
	Comparing the variances of two groups: F-test (parametric)
	Comparison of the variances of more than two groups:
	Bartlett's test (parametric)
	Levene's test (parametric)
	Fligner-Killeen test (non-parametric)

Bivariate analysis Visualization



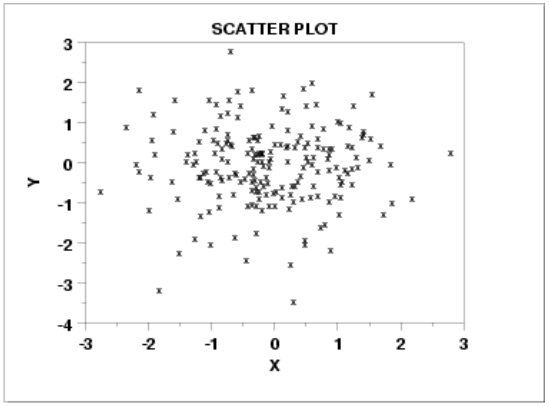
Variable 1	Variable 2	Display Example
Categorical	Categorical	Mosaic plot, association plots
Categorical	Continuous	Box plots
Continuous	Continuous	Scatter plots

Scatter plot

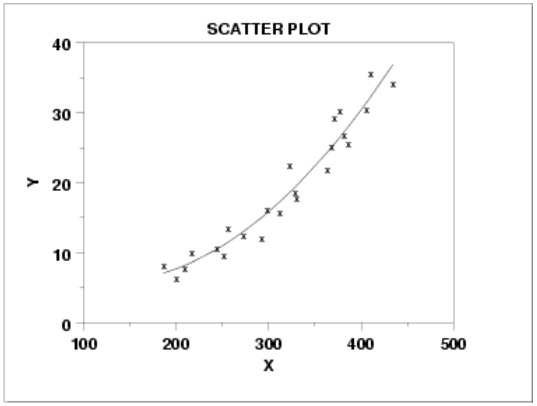
- Scatter plots reveal relationships or association between two numerical variables
- They can provide answers to the following questions:
  - Are variables X and Y related?
  - Are variables X and Y linearly related?
  - Are variables X and Y non-linearly related? Quadratic? Other?
  - Does the variation in Y change depending on X?
  - Are there outliers?

Scatter plot

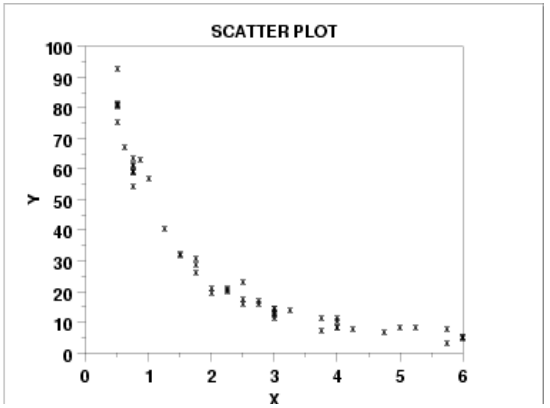
No Relationship



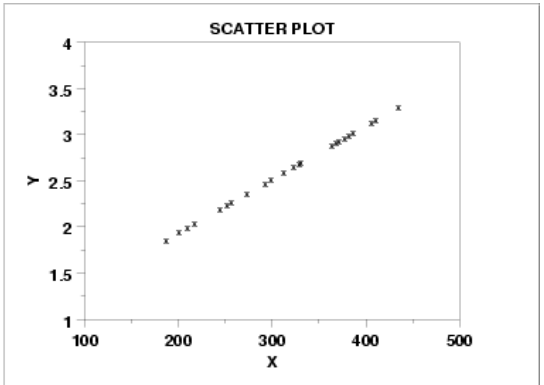
cuadratic



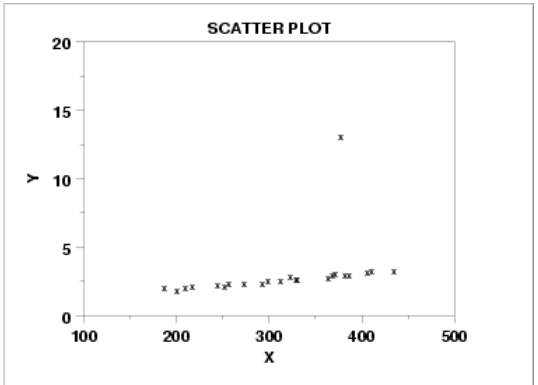
exponential



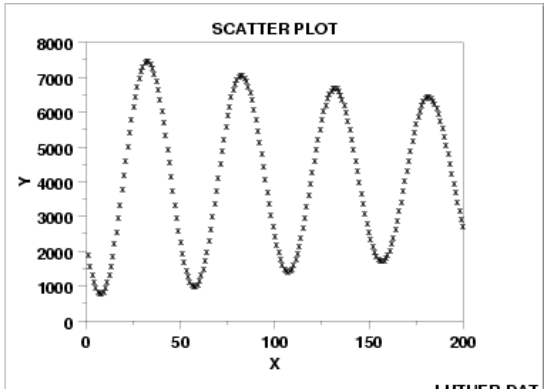
perfect lineal correlation



outliers

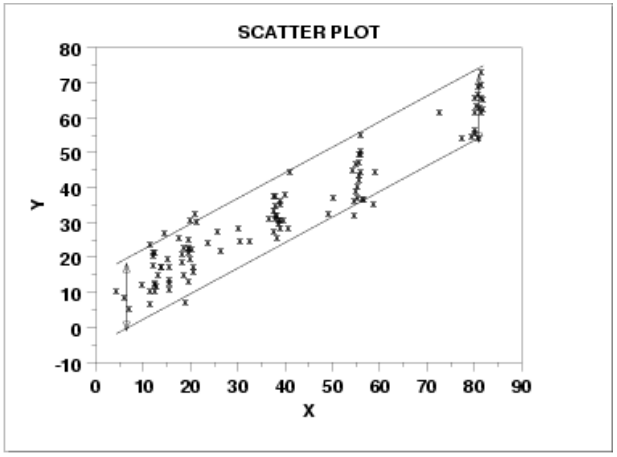


sinusoidal



Scatter plot

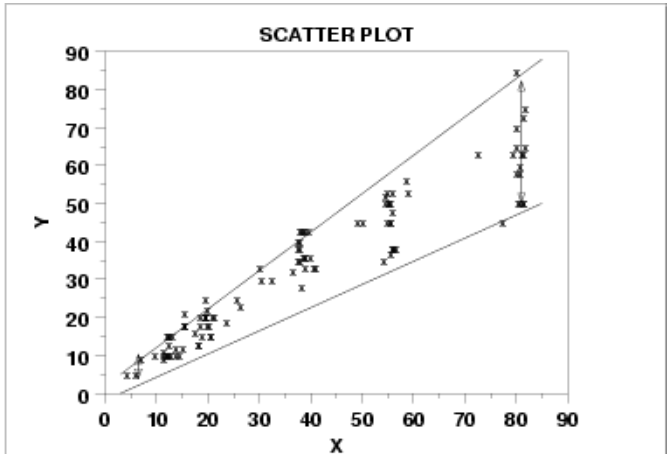
Homoscedastic



Variation of Y does Not depend on X

- Y (+- 10 units) regardless X
- Important: assumption from regression

Heteroscedastic



Variation of Y does depend on X

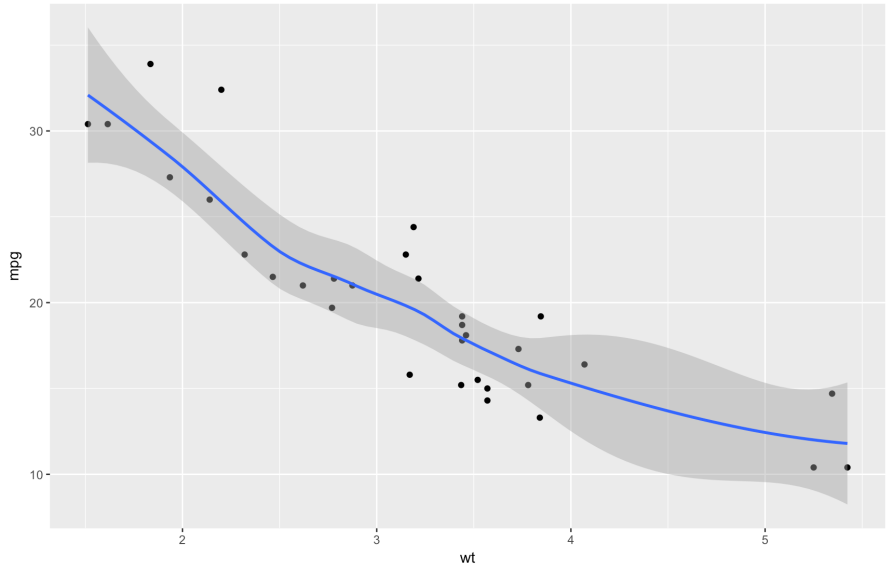
- The variation of Y is not constant,
- Implies necessity of proper weighting or Y transformation

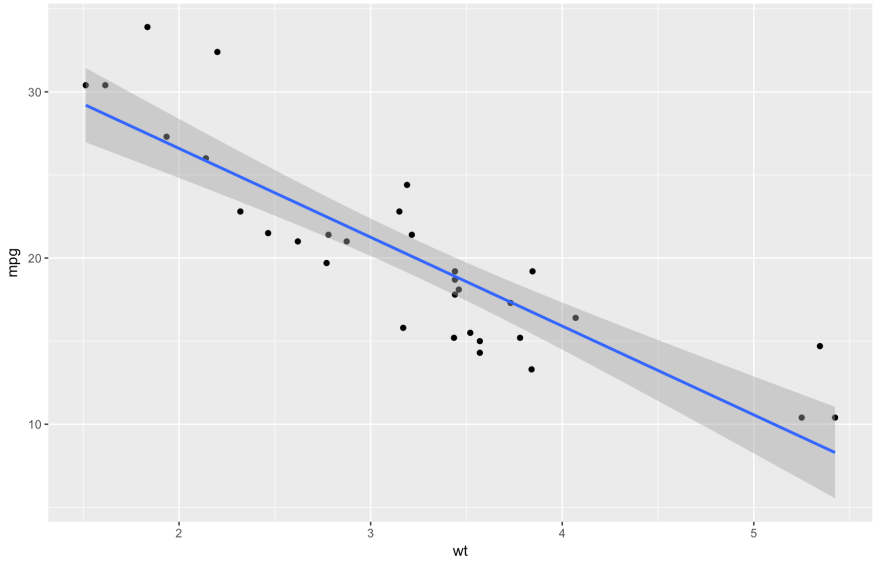
The Box-Cox normality plot can help determine a suitable transformation.
https://rcompanion.org/handbook/I_12.html

Plotting two continuous variables: smoothers (ggplot2)

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17
Valiant           18.1   6  225 105 2.76 3.460 20
```

```
# Add the regression line with
# confidence interval
ggplot(mtcars, aes(x=wt, y=mpg))
+ geom_point()
+ geom_smooth(method=lm)
```





Two variables –continuous. Large datasets

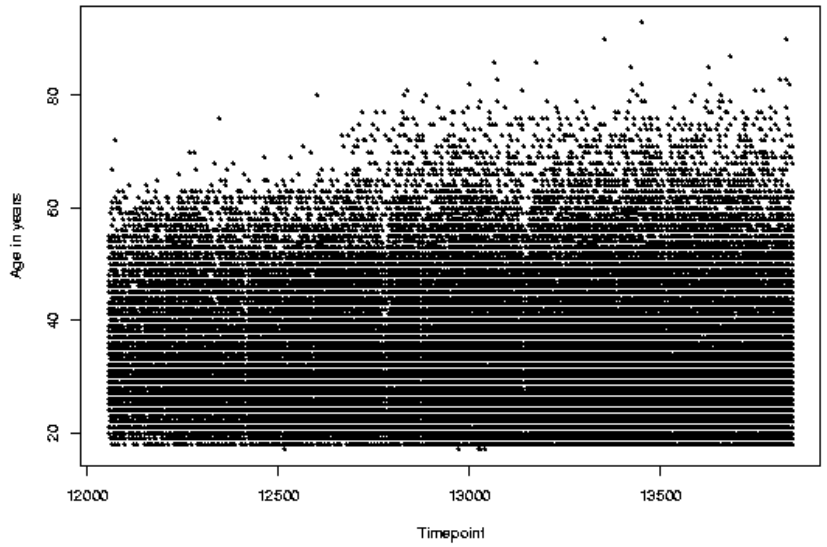
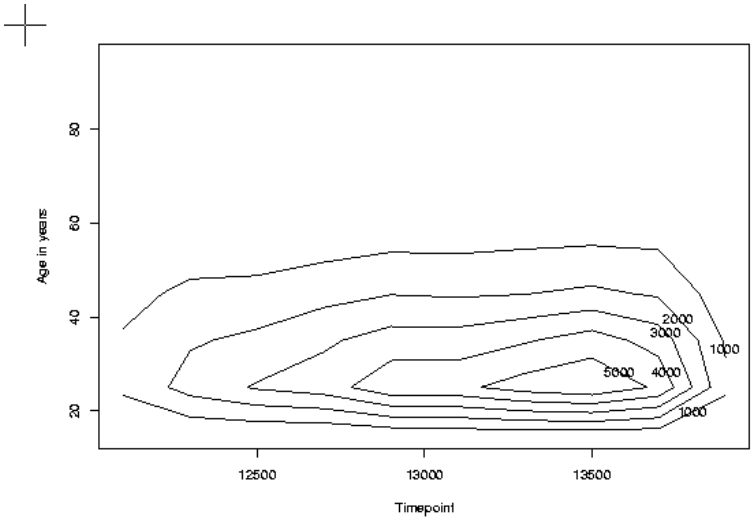


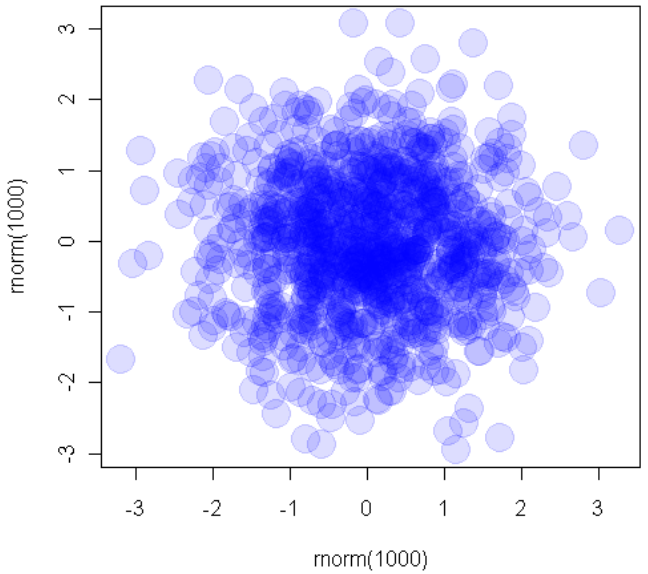
Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Two variables –continuous. Large datasets



Contour plots

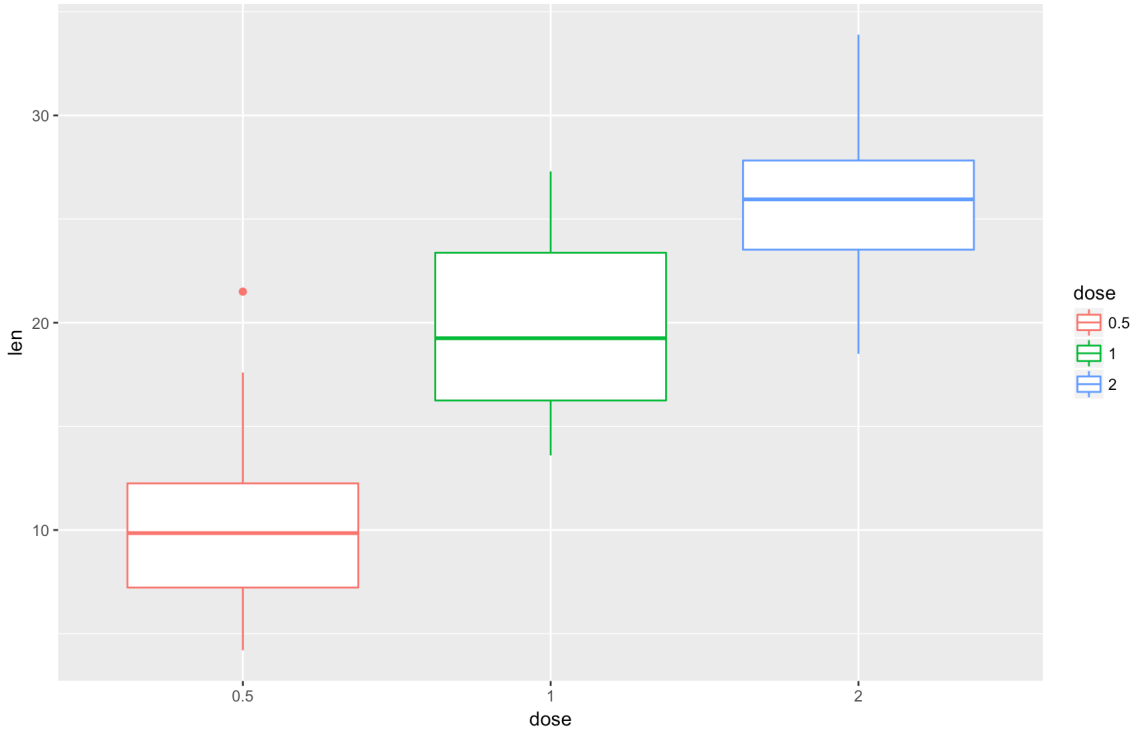
```
ggplot(df, aes(x = x, y = y)) +
    geom_density_2d()
```



Alpha blending: transparent plotting

```
X=rnorm(1000),
Y= rnorm(1000)

ggplot(aes(x, y, color ="#0000ff22”))+
    geom_point(size = 3, alpha = 0.5)
```

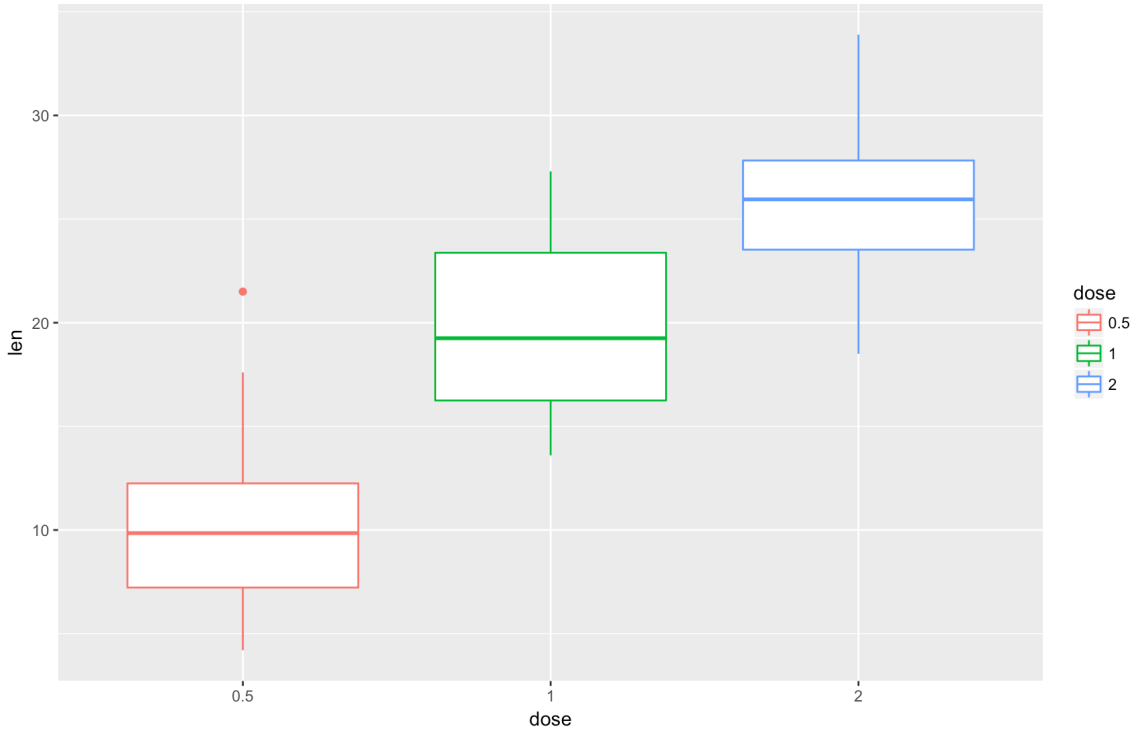
Boxplot: two variables




```
ToothGrowth$dose<-
as.factor(ToothGrowth$dose)
> head(ToothGrowth)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
```

Two variables: 1 continuous, 1 ordinal => factor conversion (example)

Boxplot: two variables




```
ToothGrowth$dose<-
as.factor(ToothGrowth$dose)
> head(ToothGrowth)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
```

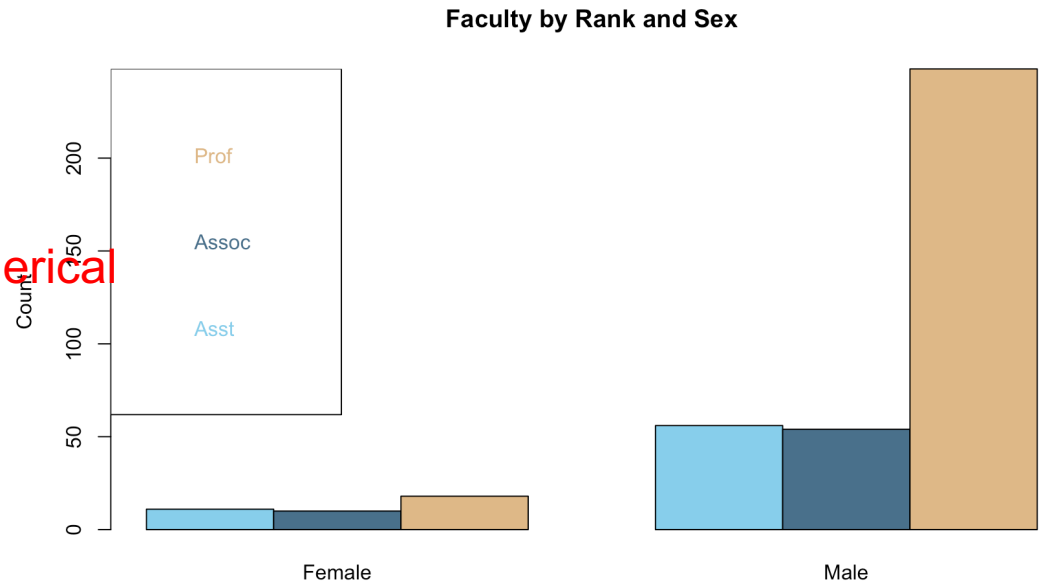
Two variables: 1 continuous, 1 ordinal => factor conversion (example)
1 numerical, 1 categorical

Bar plots, stacked plots

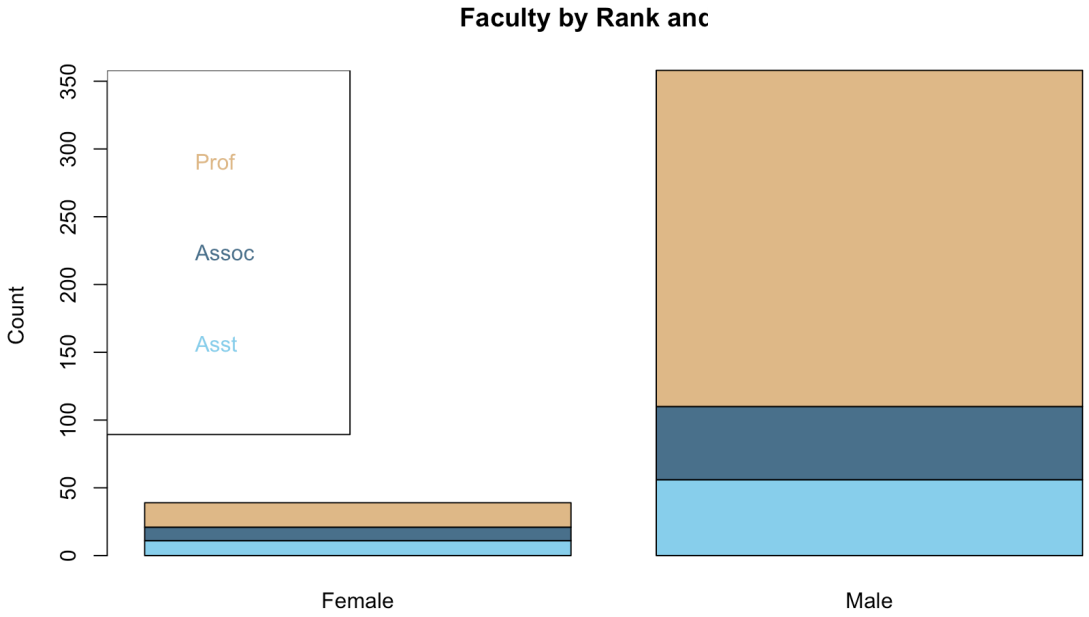
```
library(car)
attach(Salaries)
rankcount = table(rank) #get counts & save in vector rankcount
rank2 = table(rank,sex)
barplot(rank2, ylab = "Count", names.arg = c("Female","Male"),
        main = "Faculty by Rank and Sex",
        col = c("skyblue","skyblue4","burlywood"),
        sub = "c. Stacked plot")
legend("topleft", c("Prof","Assoc","Asst"),
       text.col = c("burlywood","skyblue4","skyblue"))

barplot(rank2, ylab = "Count", names.arg = c("Female","Male"),
        main = "Faculty by Rank and Sex",
        col = c("skyblue","skyblue4","burlywood"),
        sub = "d. Grouped plot", beside = T)
legend("topleft", c("Prof","Assoc","Asst"),
       text.col = c("burlywood","skyblue4","skyblue"))
```

Bar plots

Two variables categorical y numerical



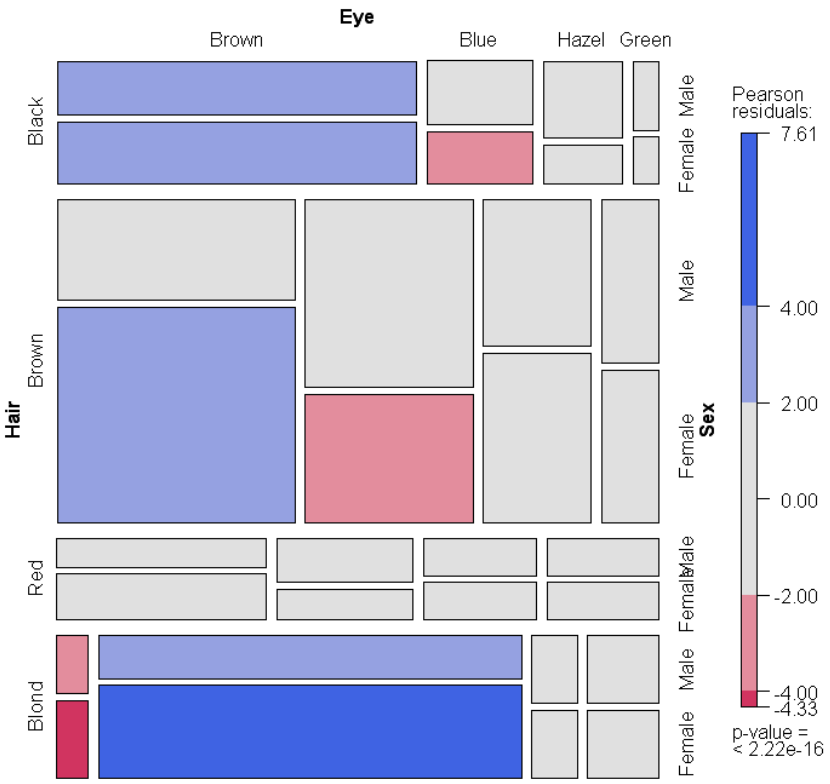
d. Grouped plot



c. Stacked plot

Visualization of two categorical data: mosaic plot

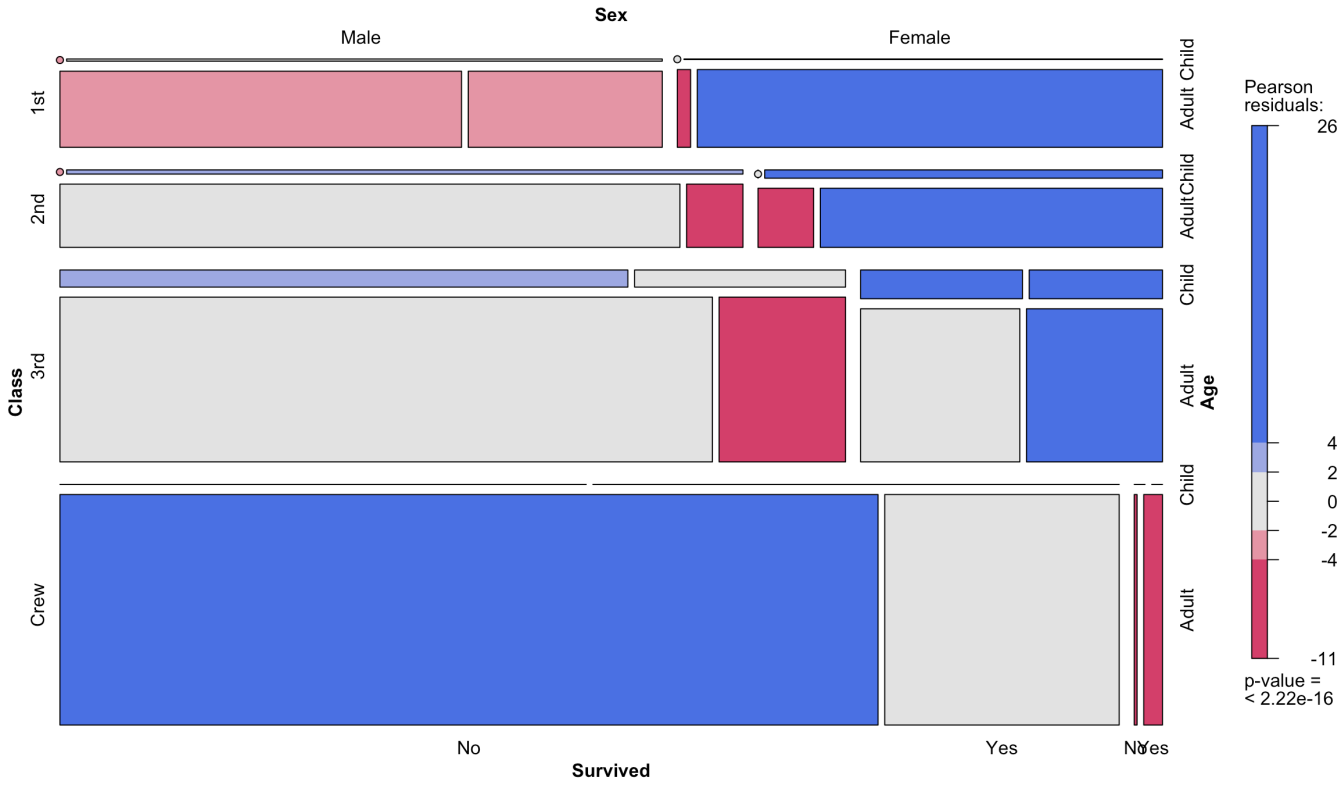
- Mosaic plots provide an ideal method both for visualizing contingency tables
- At each stage of plot creation, the rectangles are split parallel to one of the two axes.
- The important encoding is length.



Visualization of two categorical data: mosaic plot

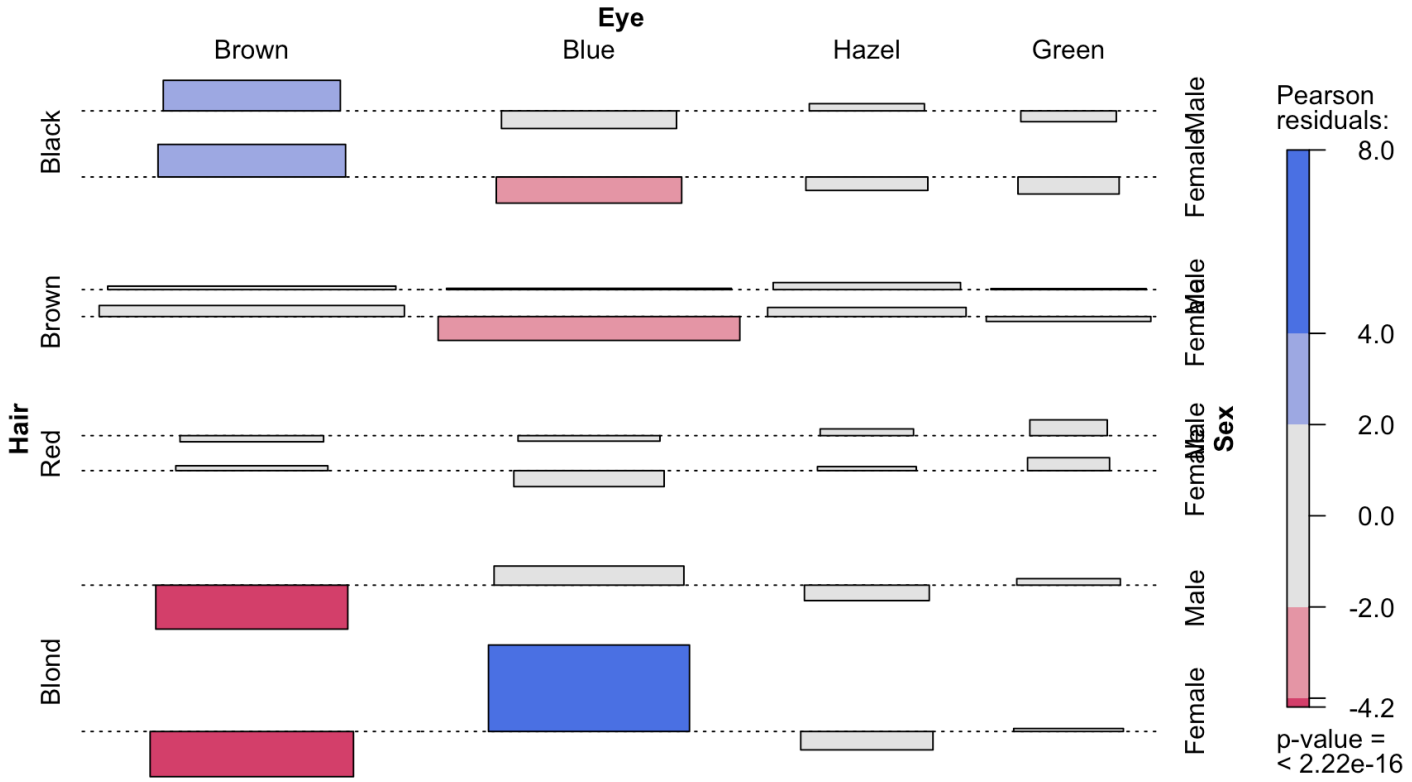
- In order to produce a mosaic plot it is necessary to have:
  - A contingency table containing the data.
  - A preferred ordering of the variables, with the “response” variable last.

```
library(vcd)
mosaic(Titanic
,shade=TRUE)
```



Visualization of two categorical data: Association plot

- Counts are represented by rectangles proportional in size to the counts of the combinations they represent



Multivariate Quantitative

Bivariate analysis include:

- Crosstabs
- Covariance
- Correlation

Advanced techniques include:

- Cluster analysis
- Analysis of variance (ANOVA)
- Factor analysis
- Principal component analysis (PCA)

Multivariate Graphical

- Matrix Scatterplot
- Profile Plot
- Correlations for Multivariate Data

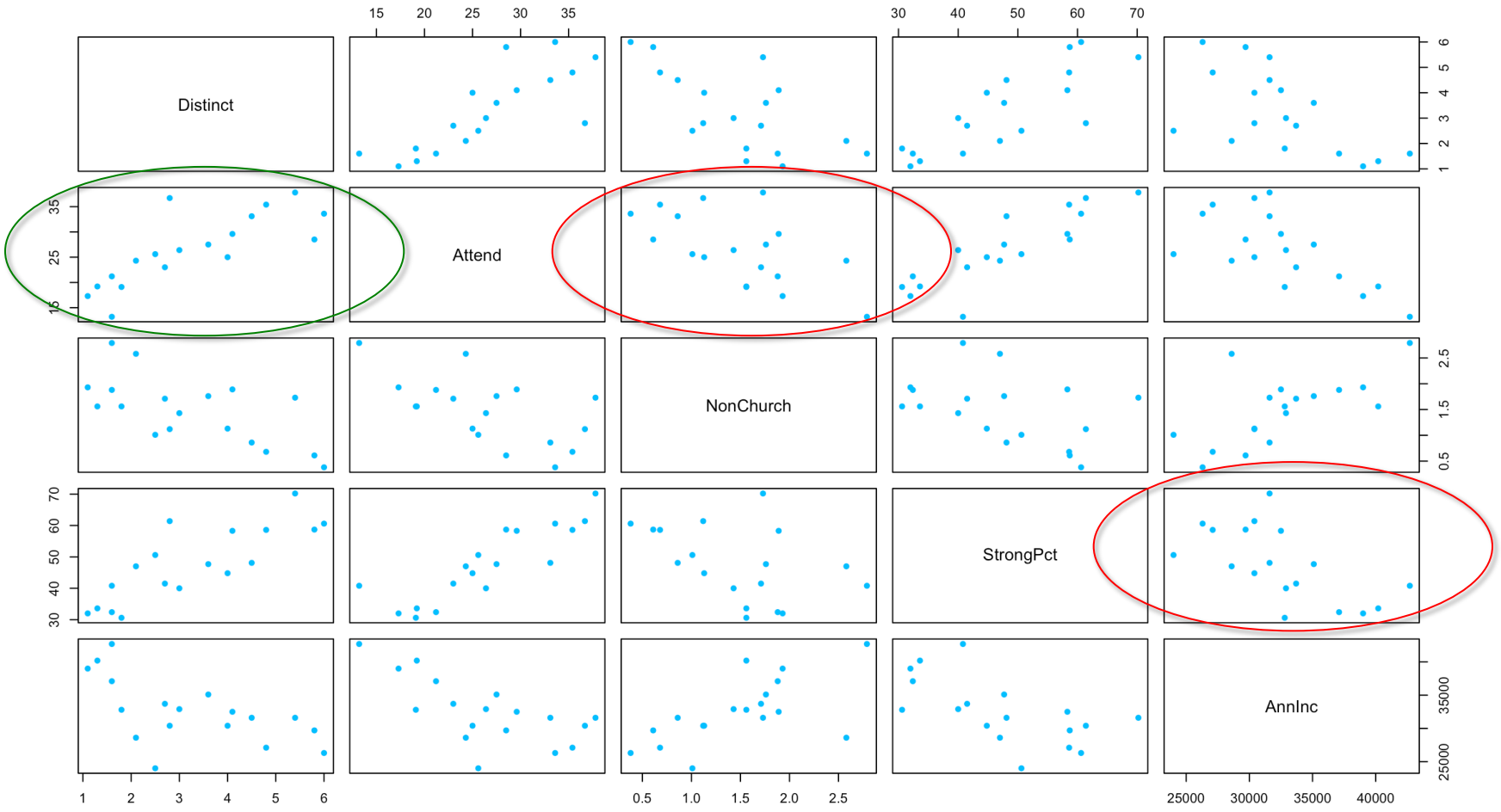
Scatterplot matrix for numerical variables

- Sometimes helps to look at the relationships of each of the possible pairs of variables first. R provides a shortcut command, pairs()

```
install.packages("Sleuth2")
library(Sleuth2)
attach(ex1713)
head(ex1713)
pairs(~ Distinct + Attend + NonChurch + StrongPct + AnnInc,
      pch = 16, col = "deepskyblue")
```

the variable names are typed as a formula, beginning with the ~ symbol

Scatterplot matrix for numerical variables: pairs()



Corrgram/Correlogram

- Type of graph related to the scatter plot matrix.
- The individual scatter plots are replaced by symbols that represent numbers measuring the amount of linear correlation between two quantitative variables.
- first necessary to make a correlation matrix by using the `cor()` function:

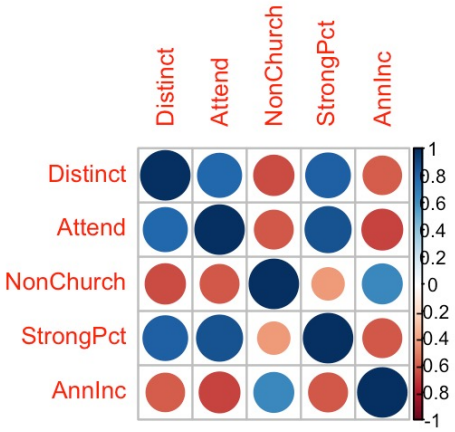
```
> install.packages("corrplot")
> library(Sleuth2)
> attach(ex1713)
> y = cor(ex1713[, 2:6]) # use all rows and columns 2-6
```

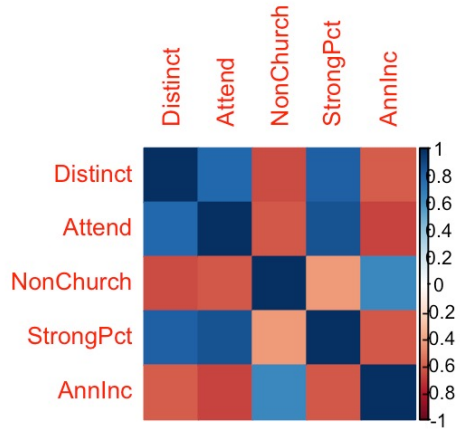
```
             Distinct     Attend  NonChurch  StrongPct     AnnInc
Distinct    1.0000000  0.7891067 -0.6585883  0.8127124 -0.6003892
Attend      0.7891067  1.0000000 -0.6107342  0.8649691 -0.6766143
NonChurch -0.6585883 -0.6107342  1.0000000 -0.4218525  0.6458747
StrongPct   0.8127124  0.8649691 -0.4218525  1.0000000 -0.6146261
AnnInc     -0.6003892 -0.6766143  0.6458747 -0.6146261  1.0000000
```

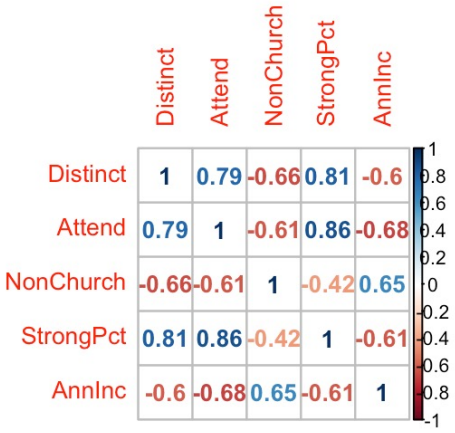
Corrgram/Correlogram

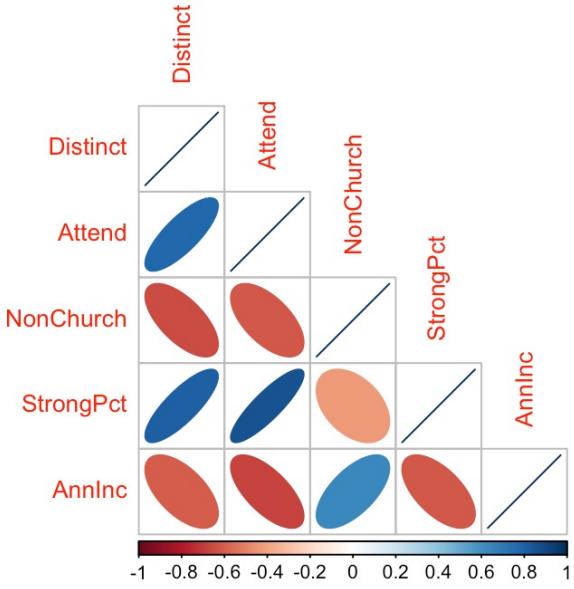
```
> install.packages("corrplot")
> library(Sleuth2)
> library("corrplot")
> attach(ex1713)
# use all rows and columns 2-6
> y = cor(ex1713[, 2:6])
> par(mfrow = c(2,2))
> corrplot(y)  # default method is "circle"
> corrplot(y, method = "color")
> corrplot(y, method = "number")
> corrplot(y, method = "ellipse", type = "lower")
```

Corrgram/Correlogram









Generalized Pairs Matrix with Mixed Quantitative and Categorical Variables

- Datasets with both quantitative and categorical variables are quite common.
- it is still possible to produce a meaningful display of all the pairwise plots of variables with `ggpairs()` from the GGally package or `gpairs()` from the gpairs package.

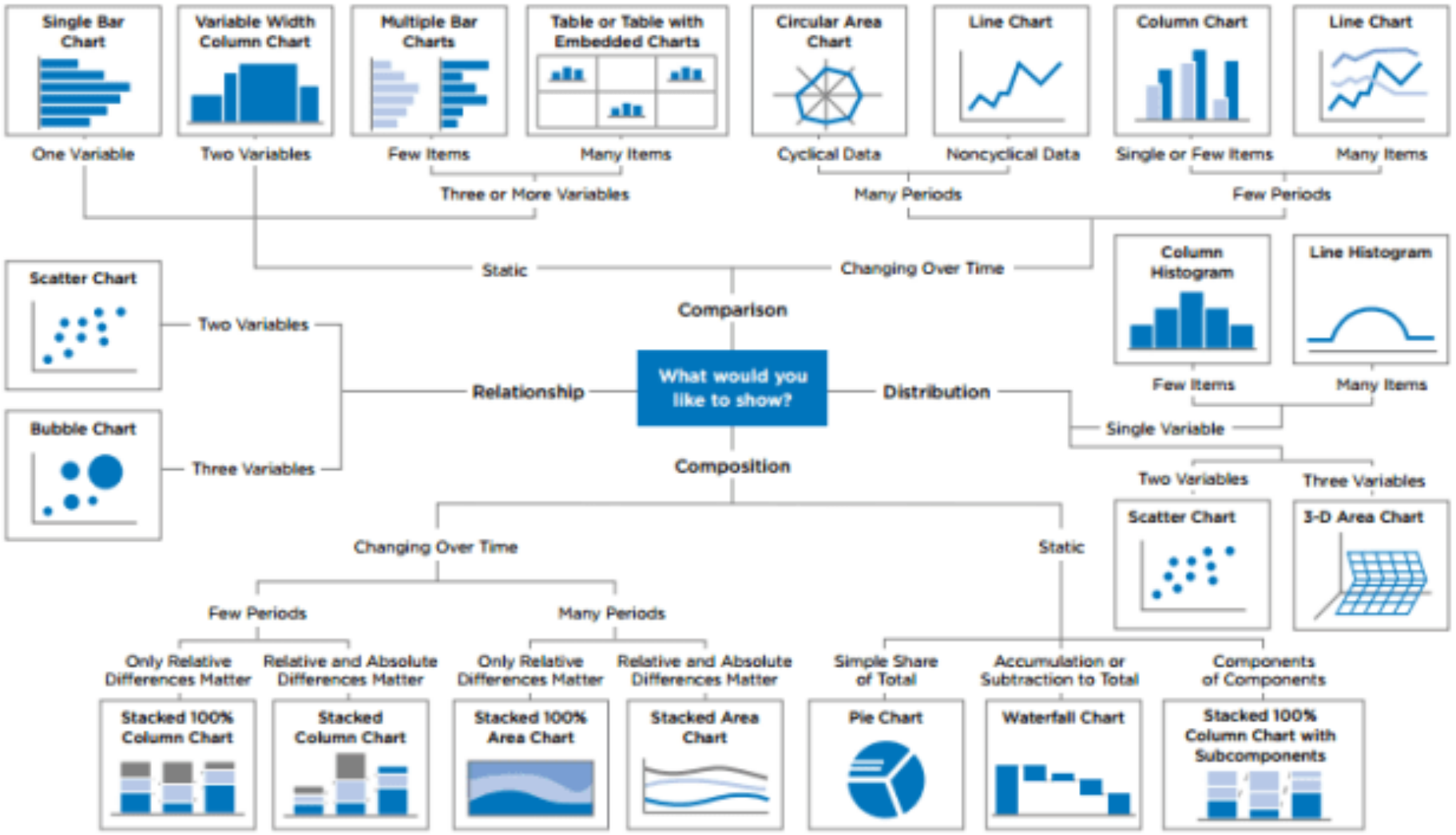
Generalized Pairs Matrix with Mixed Quantitative and Categorical Variables

```
library(GGally)
library(ggplot2)
data(tips, package="reshape")
ggpairs(data=tips, # data.frame with variables
        columns=1:3, # columns to plot, default to all.
        title="tips data", # title of the plot)
```

To determine the best graph…

- How many variables do you want to show in a single chart?
- How many data points will you display for each variable?
- Will you display values over a period of time, or among items or groups?

SELECTING THE APPROPRIATE CHART FOR STRATEGY PRESENTATIONS



Source:https://conorneill.com/2018/01/11/selecting-the-right-chart-for-your-presentation/

Exercises