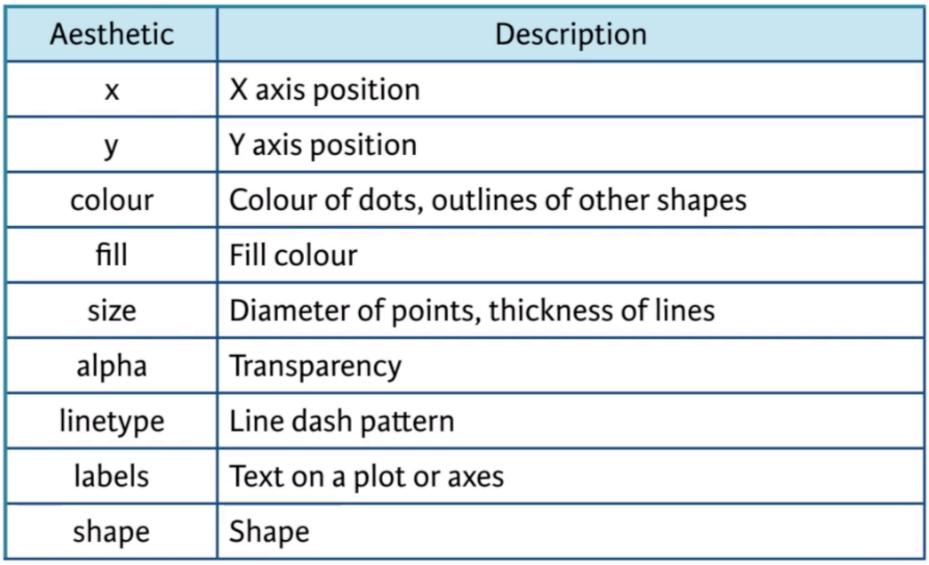
Data Visualization with R part 2

Aesthetics

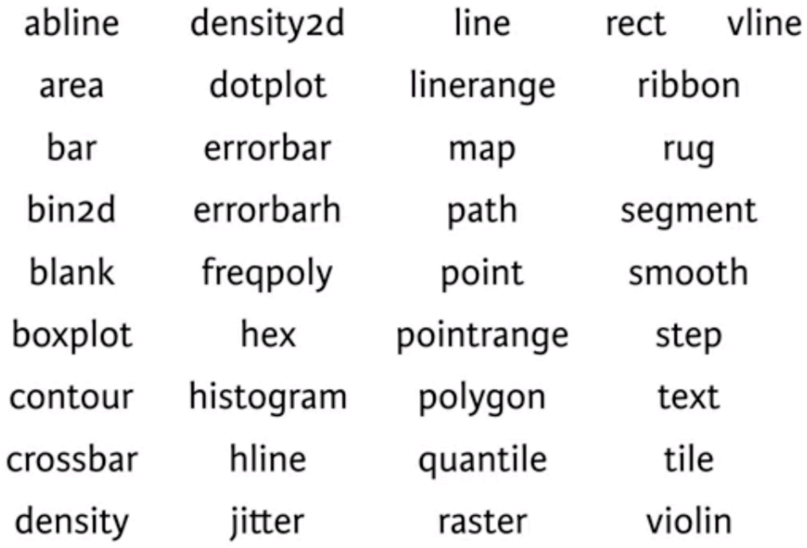
Aesthetics are an attribute of ggplot which is used to tell which variable is mapped onto it.

- They are the cornerstone of the grammar of graphics plotting concept.
- Used for converting categorical or continuous variables into visual scales through mapping that provide access to large amount of information in a very short time.
- Helps keep track the variables in the plot, example species variable can be mapped into colors.
- Add a variable for color means adding a dataframe column into visible aesthetics
- Having a proper data structure is important for mapping.
- Aesthetics attribute are called through the aes() function.
- In general it is better to use aesthetics attribute and data together in the ggplot2 function definition, unless different data sources are combined (add in geom functions then).



Aesthetic	Description
x	X axis position
y	Y axis position
colour	Colour of dots, outlines of other shapes
fill	Fill colour
size	Diameter of points, thickness of lines
alpha	Transparency
linetype	Line dash pattern
labels	Text on a plot or axes
shape	Shape

Geometrics



abline
area
bar
bin2d
blank
boxplot
contour
crossbar
density
density2d
dotplot
errorbar
errorbarh
freqpoly
hex
histogram
hline
jitter
line
linerange
map
path
point
pointrange
polygon
quantile
raster
rect
ribbon
rug
segment
smooth
step
text
tile
violin
vline

Common plot types

- Scatter plots
    - Points, jitter, abline
- Bar plots
    - Histogram, bar, error bar,
- Line plots
    - Line

dplyr package

- An important part of exploratory data analysis is summarizing data.
- Better summaries can be achieved by splitting data into groups before using the normal approximation.
- dplyr is much, much faster than other, more traditional, functions.
- It provides direct connection to and analysis within external databases permitting simpler handling of large data
- Function chaining that allows us to avoid cluttering our workspace with interim objects
- Syntax simplicity and ease of use. The code is easy to write and to follow.

Getting Started with EDA

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

What is EDA?

To explore data in a systematic way using visualization and transformation that will create a state of mind to ask the rights questions or refine the questions.

Why is it important?

Not a good idea to simply feed data to a black box!

- Helps summarize and understand data without any assumption.
- Investigate the quality of the data.
- Eliminates the wrong questions being asked
- Crucial step before machine learning/statistical modeling.
- Provides context needed to develop an appropriate model and to correctly interpret its results.
- Useful for efficient feature engineering.

EDA Cycle

- Generate questions about your data.
- Search for answers by visualising, transforming and modelling your data.
- Use what you learn to refine your questions and/or generate new questions.

Why we need to ask questions?

- To guide your investigation for understanding data.
- Focus on the specific part of the data set.
- Helps you decide which graphs, models or transformations to make.

the key to asking quality questions is to generate a large quantity of questions

The dataset that we will working on

This data set is a collection of domestic flights from three major airports of New York.

Source: https://stat.duke.edu/~mc301/data/nycflights.RData

Research Questions

1. **Is there any correlation between the departure delays and the time of the year when flights get delayed?**
2. **Can the airport origin have an association with the departure rate of the departing flights?**
3. **Is there any correlation between the arrival delays and the speed of the airplanes for the flights?**

Cheat sheet for Data Visualization