**Exploratory Data Analysis**

**Principles of Analytic Graphics:**

1. show comparisons, evidence is always relative. Compare hypothesis A to another hypothesis,
2. Show causality, mechanism, explanation , systematic structure
3. Show multivariate data, other variables that show relationship., how does relationship change. Show as many variables as you can .
4. Integrate the evidence. data graphics should make use of many modes of evidence.
5. Describe and document the labels with appropriate sources, labels and scales.
6. think about the content , data and what you are trying to present, what is the story that is being expressed.

Why use graphs ? understand properties, find patterns, suggest modeling strategies, debug, self exploration, pre analysis
Exploratory graphs - one dimensional- 5 num summary, box plots, histograms, density plot, barplot

**Functions: one dimensional plots.**
$head()
$summary()- for 5 number summary, includes the mean as well

$boxplot(table$column, col= "blue" )
$abline(h=12) , place a line to see mean median,

$hist(, col, breaks= number of bars(adjust)) - histogram
$rug() - puts all the datapoint below
$abline(h=12, col , v = vertical line , v = median(column), led = line density )

$barplot(table, col = , main)

Two dimensional plots: Overlayed plots, Scatterplots, smooth scatterplots , Spinning plots,
Organize in a special way and use color shape size to add

**Functions: Two dimensional plots:**
$boxplot(pm25 ~region , data = pollution , color)

$par(mfrow = c() , mar= ()

$hist(subset(pollution , region == east)@pm25, color)

$with(pollution, plot(latitude, pm25, col = region))  — -SCatter plot, place line for medium
$abline(h=12, lwd = 2, lty = 2)

**LESSON 2: Plotting**
The Base Plotting System:
use plot function and annotate., Difficult to go back or translate.
liberary(datasets)
data(cars)
with(cars, plot(speed, disti)

The Lattice System:
Plots created with xyplot, specify a lot of functions , useful for conditioning plots, relationship between x and y and z. Looking for panel plots, combine variables and multiple plots. Many thins calculated automatically, good for many plots. Difficult to annotate a plot.

$library(lattice)
$state <- data.frame(state.x77, region =)
$xyplot(lifeexp ~income | region , data, state, layout c= c() )

The ggplot2 System:
language or grammar for describing the plot, strict and rigorous. useful for conditional plots and panel plots. Lot of defaults, know how to use them. Mix of both base and lattice.

$library(ggplot2)
$data(mpg)
$qplot(displ, hwy, data)

All graphics contained in base and grDevices contains all the packages,
Lattice plotting is in lattice package and grid.

The process of making a plot: Where to print ? quality ? large amount of data ? be able to dynamically resize the graphics ?

First pick the graphics system: base, lattice, ggplot2 .

Base Graphics:
Two phases- Initialize, annotate,

Plot(x,y) or hist() to initialize and launch a graphics device, window on screen ,
?par - many parameters to optimize,

library(datasets)
hist(airquality@Ozone)
with(air, plot(wind, Ozone))
box plot( Ozone ~ Month (y axis and axis),  xlab and ylab for labels) .

**Common parameters:**
pch - plotting symbol
lty : line type
lwd : line width
col : plotting color
xlab : labels
ylab : labels

par() - specify global graphics parameters:
las - orientation of the axis labels on the plot, horizontal or perpendicular to graph
bg - background color
mar: margin size
oma: outer margin size (default 0 )
mfrow: number of plots per row, column
mfcol: no of plots per column
call par("lty) to look at the defaults

**Base plotting functions:**
plot: make scatter plot or define other types of plots
lines: add lines to a plot, connect the dots
points: add points to a plot
text: add labels within the plot
title : add texts outside
mtext: add text to margin
axis: add axis or labels

$with(subset() , type ="n" , set stuff without adding anything into the data frame)

model <- lm(Ozone ~wind, air quality)  - add linear regression model to trend.
abline (model ) - add generated

par(mfrow =c(1,2), mar = c() , oma - outer margin , )  - one tow and two columns,
add multiple plotss
with(air, {
plot1,

plot2
plot3
mtext (“add outer title”)

**Graphics Devices in R**
Places to make a plot appear.
in mac quartz() , launches  a screen device.
save plotting code in a different part . dev.off() close the file

pdf(file = my plot.pdf)
with (plot)
dev.off()

Vector formats and bitmap devices
Vector : pdf, svg, win.metafile, postscripts — good for lines and solid colors.
Bitmap: png, jpeg, tiff, bmp , good for plots with large number of point s

**Lattice Plotting system:**
Main functions:
air quality <- transform(data, Month  = factor(month))
xyplot(y~x | f*g, data= , layout= c(5,1)  ) — make multidimensional panels
bwplot
histograms
stripplot
dotplot
splom
levelplot, countourplot

Return an object of class trellis, this object will have to print, R auto prints, save it to an object and

Panel functions, use for multipanel functions, each panel receives a subset, gets the xy co-orginates of the plot,

**Cluster analysis**
**Hierarchical Clustering**
define when one thing is closer to each other and when things are further apart, .
Group things together.
organize data, start with the data, start lumping them together and gradually lead to a cluster.

Make a merge point and find closest things, and make a tree like structure,. Use different distance metrics.