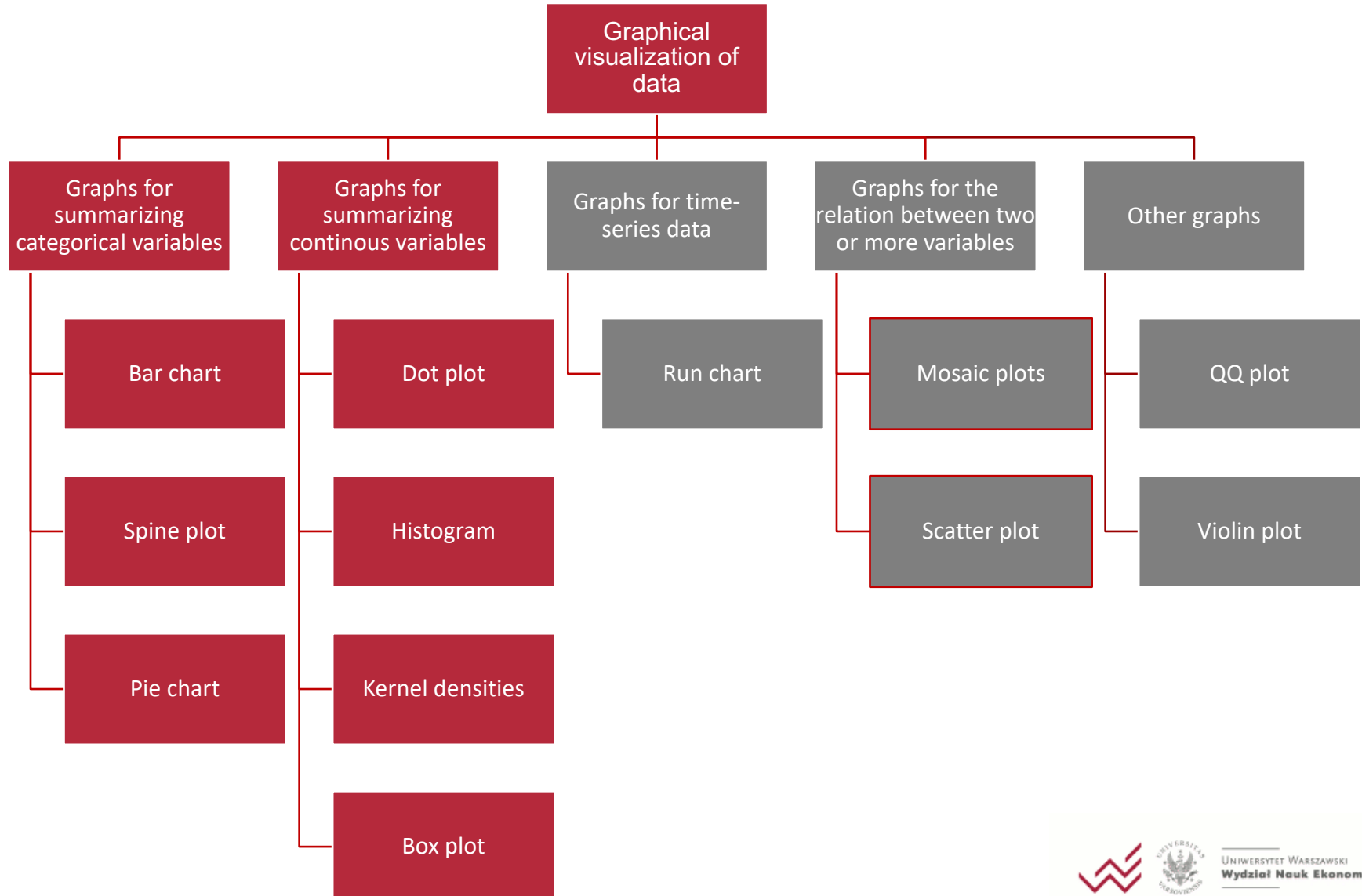


Graphical analysis of data (II)

Marcin Chlebus, Ewa Cukrowska-Torzewska
Faculty of Economic Sciences
University of Warsaw

Lecture 5

Types of graphs

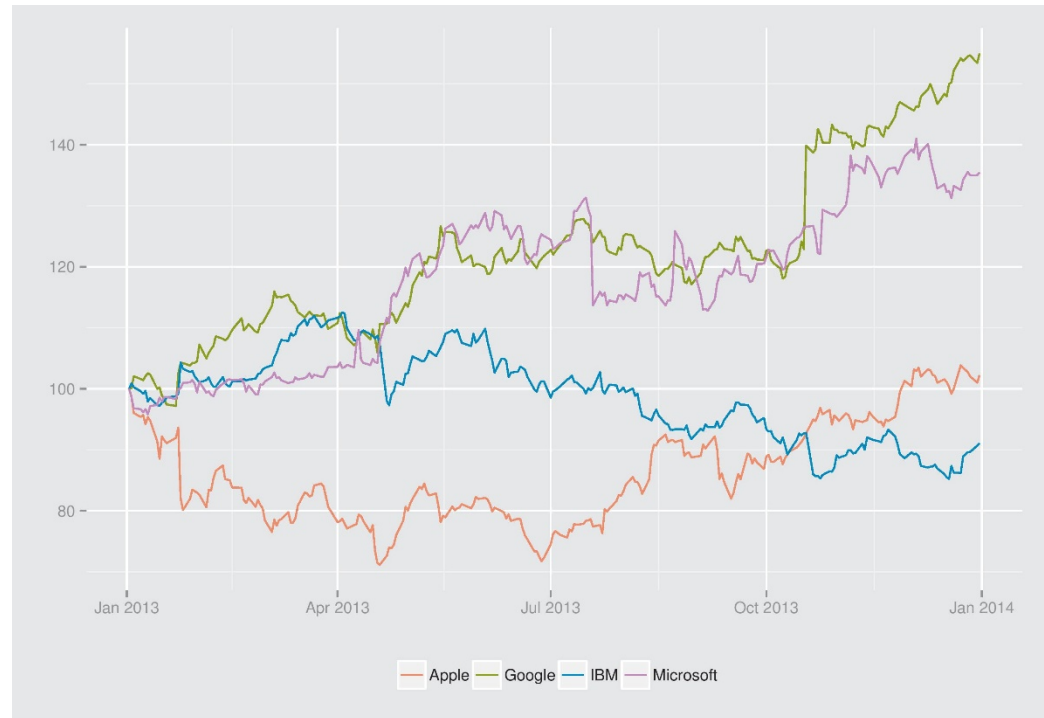


Run charts

- A run chart is a line graph of data plotted over time
- Outliers can be easily detected
- It is used to detect patterns and trends of a given process over time, **auto-correlation** in the data and whether the process is **stationary**

Autocorrelation – linear dependence of a variable with itself at two points in time.

Stationary process – the distribution does not change if it is shifted over time (i.e. the mean, variance and covariance are the constant over time), e.g. the White noise.



Exercise 1:

Use data on air quality that is available in R.

Graph the temperature and wind data over time.

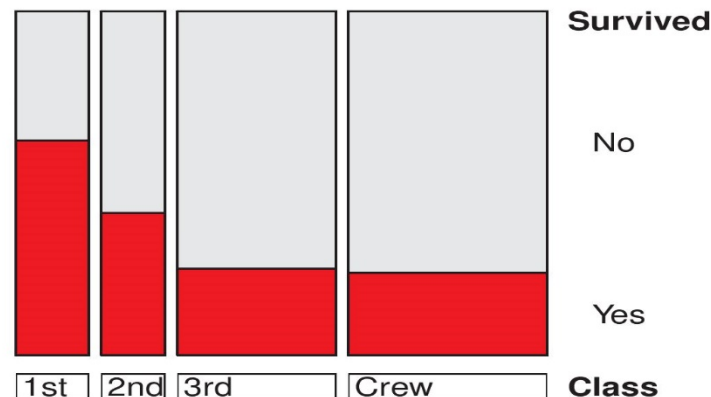
What can say about the stationarity and autocorrelation in these data?

Mosaic plot

- It is used to present the relation between two or more categorical variables
- The frequencies of the contingency table are represented by a collection of rectangles (or tiles) whose area is proportional to the cell frequency

Gender	Survived	Total
Male	No	1364
	Yes	367
Female	No	126
	Yes	344

Survived	1st Class	2nd Class	3rd Class	Crew
No	122	167	528	673
Yes	203	118	178	212



Mosaic plot

Gender	Survived	1st Class	2nd Class	3rd Class	Crew
Male	No	118	154	422	670
	Yes	62	25	88	192
Female	No	4	13	106	3
	Yes	141	93	90	20



Exercise 2:

Install package „car” and use the dataset „Salaries” from that package.

The dataset contains salaries of the nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S.

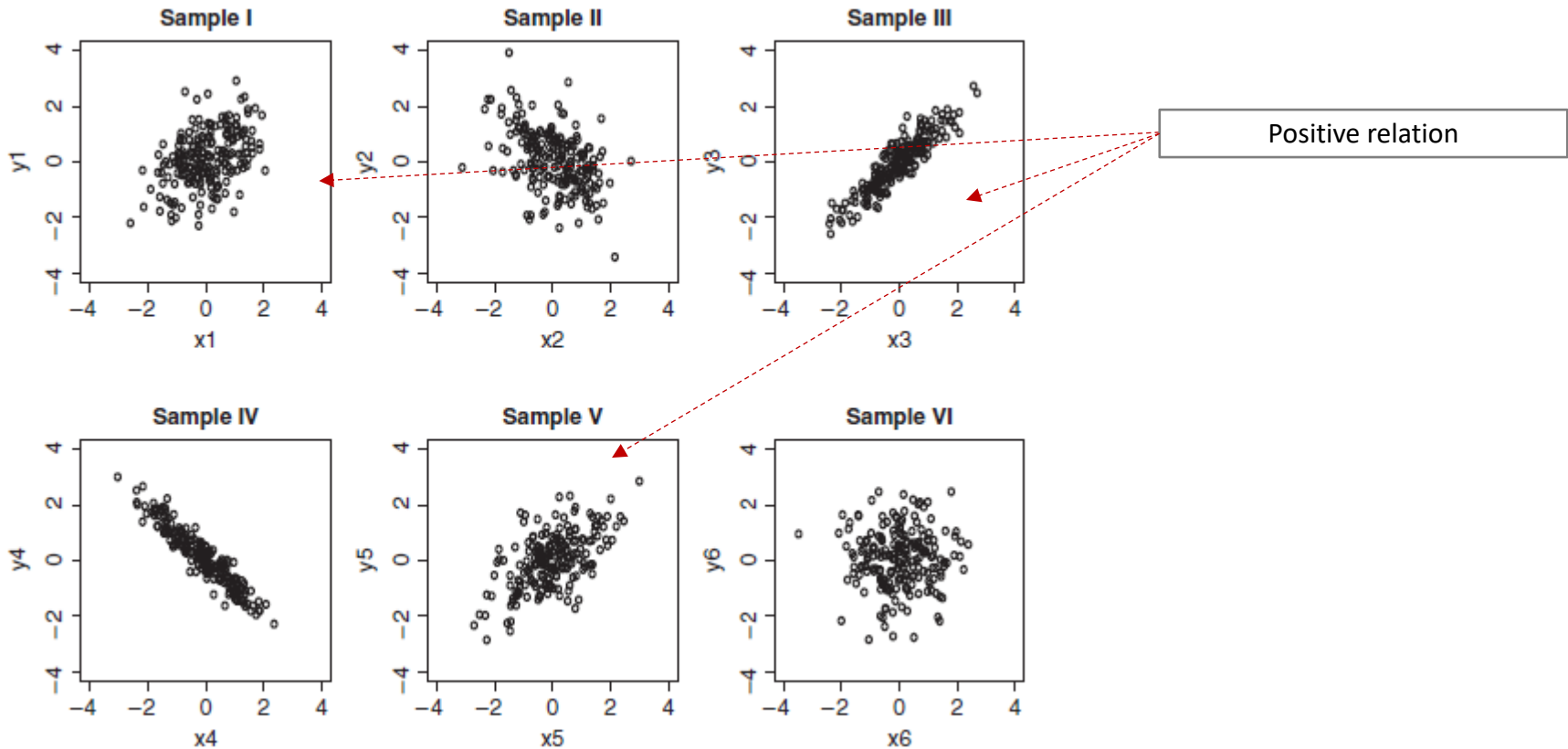
Create the mosaic plot for the discipline and sex

Create the mosaic plot for the discipline, rank and sex

Interpret the graphs.

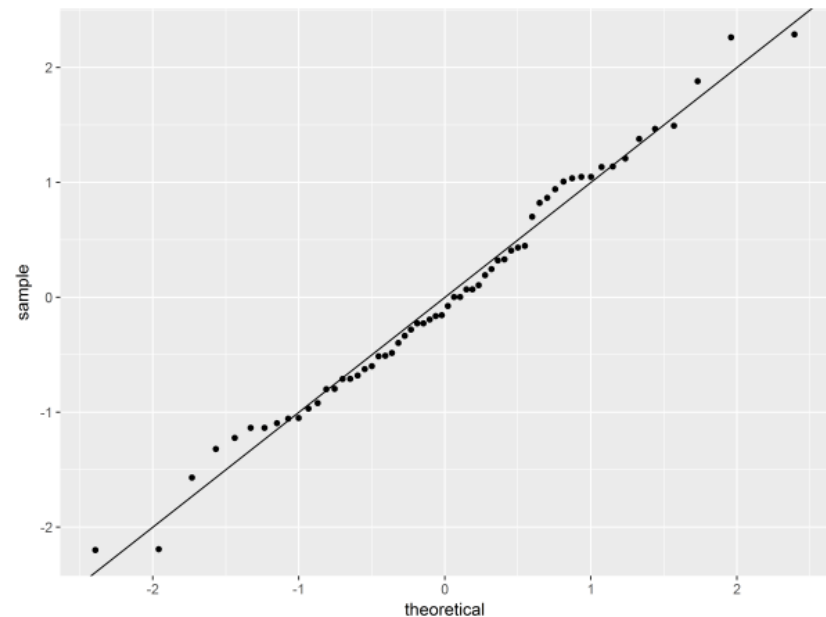
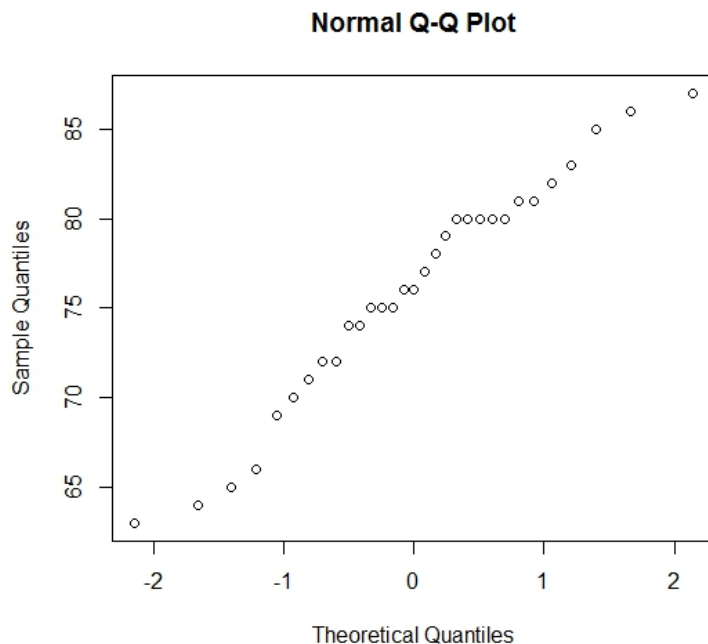
Scatter plot

- It is used to detect dependence between two continuous variables
- It maps each variable to an x- or y-axis coordinate



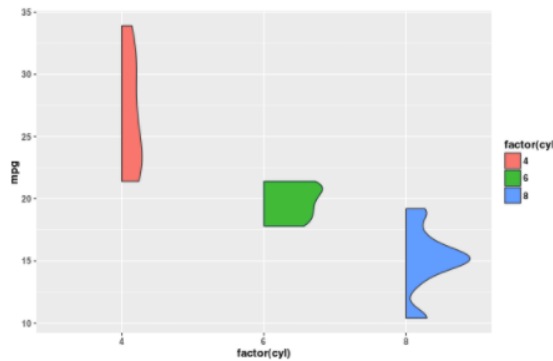
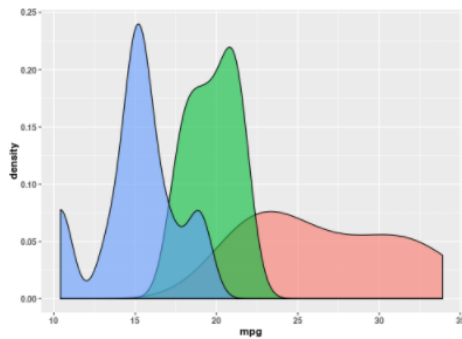
QQ plot (quantile-quantile plot)

- It is used to assess whether the data come from a given theoretical distribution (e.g. Normal, lognormal, exponential, etc)
- It is a scatterplot in which the quantiles for the variable observed in the data are plotted against the theoretical quantiles of a specified theoretical distribution
- If both the sets of the quantiles come from the same distribution the points should form a line

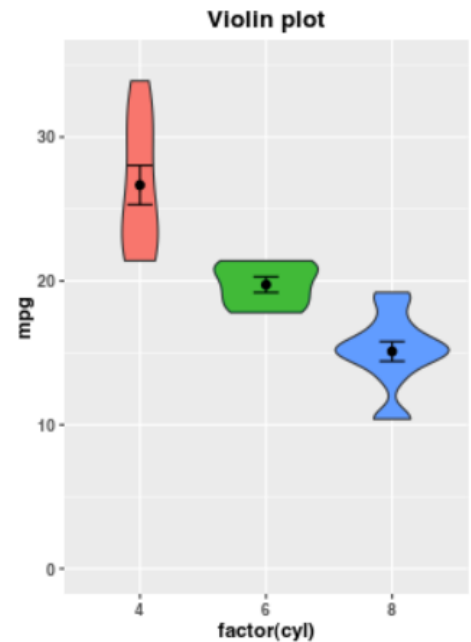
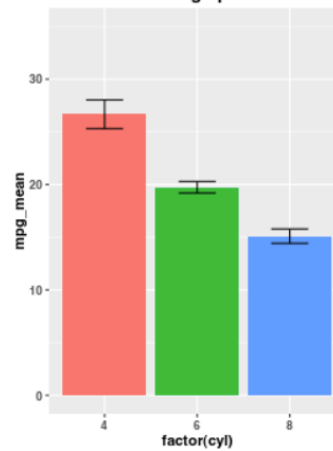


Violin plot

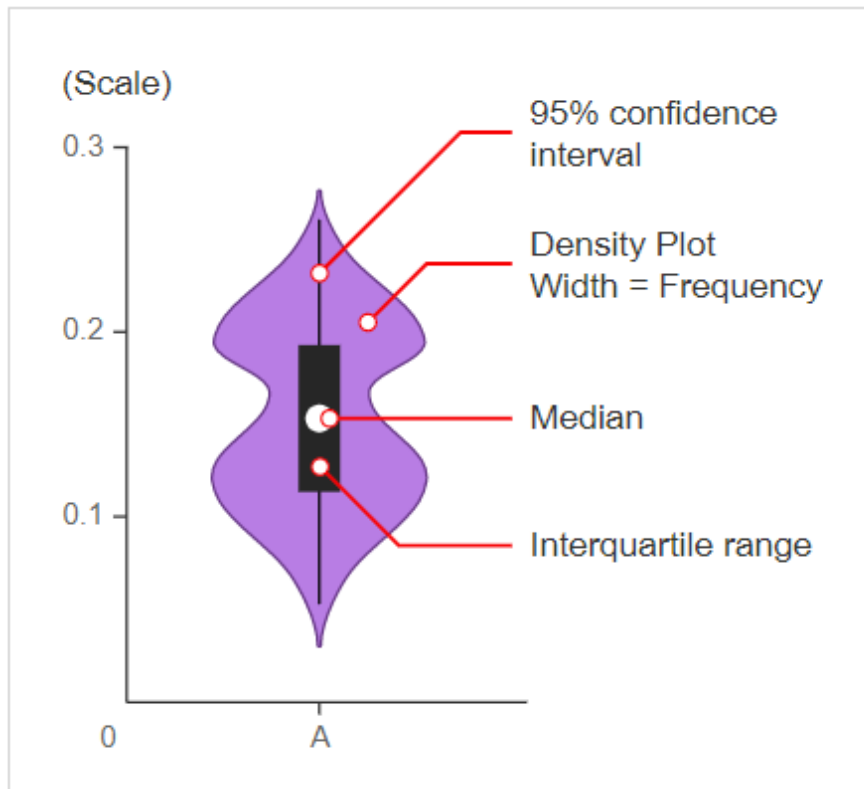
- The plots are combination of box plots and density plots
- It is used to summarize the distribution and the density of a given variable



Bar graph



Violin plot





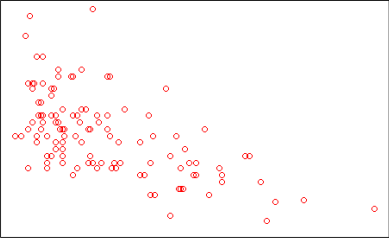
Source: https://datavizcatalogue.com/methods/violin_plot.html

Exercise 3:

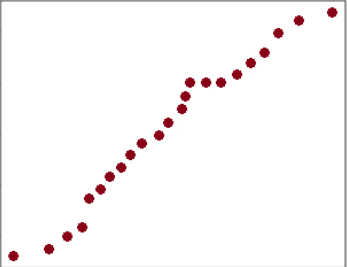
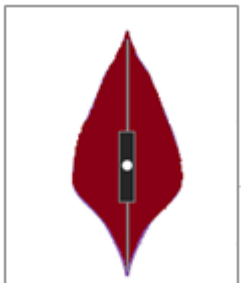
Use data on diamonds that are available in R. Using the data:

- depict the relation between the number of carats and the price
- depict the relation between the number of carats, the price and the quality of the cut
- check graphically whether the price of the diamonds
- summarize the distribution and the density of the diamonds price by cut quality.

Data visualization in R

	Standard approach	Ggplot package
Run chart 	<pre>library(qicharts) qic()</pre>	<pre>ggplot(data=., aes(x=time variable, y)) + geom_line()</pre>
Mosaic plot 	<pre>library(vcd) doubledecker(x ~ y1 + y2, data = .)</pre> <p>Or</p> <pre>table(x,y) → table mosaicplot(table)</pre>	<pre>ggplot(data = .) + geom_mosaic()</pre>
Scatter plot 	<pre>plot(x, y)</pre>	<pre>ggplot(data=., aes(x=., y=.)) + geom_point()</pre>

Data visualization in R

	Built in functions	Ggplot package
QQ plot 	For normal distribution: qqnorm(.) qqline(.)	<code>ggplot(data=., aes(sample=.))+stat_qq()</code>
Violin plot 	<code>vioplot()</code>	<code>ggplot(data=., aes(x=., y=.)) + geom_violin()</code>

Bibliography

Antony Unwin, Martin Theus, Heike Hofmann, Graphics of Large Datasets Visualizing a Million. Springer 2006.

Winston Chang. Practical Recipes for Visualizing Data. R Graphics Cookbook. O'Reilly 2012.

For data visualization using ggplot:

<http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>

Thank you for your attention

Time for practice!