Chapter 1.

Exploratory Data Analysis

It is essential and very often overlooked that the first step of any formal statistical analysis is to *get to know the data*. Or we might call it making acquaintance or making friends with data. "Seeing is believing." The advice is to see or read the data with our bare eyes. This sounds to be too primitive but many times this simple procedure can be very rewarding. You might find far more useful information than you expect by merely scrolling down the data. "Take a rough look." is the first step.

More formally, this initial step of data analysis is called exploratory analysis—just to explore data, often without particular purposes other than describing or visualizing the data. The methods can be crudely divided into two categories: numerical or graphical.

1.1. Numerical methods though simple statistics.

The methods described here are largely about the observations of a single variable. Before giving details, we first set some convention about notations. for any

Consider n observations $(x_1, ..., x_n)$ of a single variable The typical statistics for single variable observations are:

a). for describing the distribution of the data: quantiles (percentiles or order statistics): minimum, first quartile, median, third quartile, maximum. The order statistics is the same as the original observations, except now arranged in "order". They are commonly denoted as $x_{(1)} \leq ... \leq x_{(n)}$, and $x_(k)$ is the k-th smallest observation. Order statistics can be viewed as the quantiles of the distribution of the data. For any distribution F of quantiles

a). for describing location or center of the data: sample mean $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$ and sample median

$$\hat{m} = \begin{cases} x_{(n/2+1/2)} & \text{if } n \text{ is odd} \\ 1/2(x_{(n/2-1)} + x_{(n/2)}) & \text{if } n \text{ is even;} \end{cases}$$

where $x_{(1)} \leq ... \leq x_{(n)}$ are the so called order statistic. We may say median is the value such that half of the observations are smaller than it and the other half greater than it. Note that $\bar{x}$ is the value minimize the spread by l_2 distance and $\hat{m}$ is the value minimize the spread by l_1 distance. Therefore it is fair to use them as proxies of "center" of the data.

b). for describing spread of the data: sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

sample standard deviation is $s = \sqrt{s^2}$; range: $x_{(n)} - x_{(1)}$, inter-quartile range: the difference between 25th percentile, which is called first quartile, and the 75th percentile, called the third quartile. The range, although intuitive and straightforward, is heavily influenced by outliers. Standard deviation is by far the most widely used measure of the spread of data.

The estimated density function is commonly used to describe the distribution of the variable. A simple approach is the kernel density estimation:

$$\hat{f}(t) = \frac{1}{nb}\sum_{i=1}^{n} k((x_i - t)/b).$$

Here $b > 0$ is the bandwidth and k is a kernel function, which is usually nonnegative and must integrate to 1. It is not difficult to generalize this kernel density estimation to several variables (For many variables, this estimation faces the difficulty of the "curse-of-dimensionality").

The most natural estimator of the cumulative distribution function of the population is the empirical distribution

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le t),$$

where $I(\cdot)$ is an indicator function. Therefore $\hat{F}(t)$ is simply the percentage of observations taking value less than or equal t.

For many variables, the pair wise relations are usually described by the sample variance variance matrix (also called covariance matrix) or correlation matrix. Suppose we have n observations of p variables. A conventional way of describing the data using matrix is

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & \vdots & ... & \vdots \\ x_{31} & x_{32} & ... & x_{np} \end{pmatrix} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

With this notation, x_{ij} is the i-th observation of the j-th variable. and $\mathbf{x}_j$ is n-vector representing all observations of the j-th variable. We shall also use $x_i = (x_{i1}, ..., x_{ip})$ to denote the i-th observations.

The sample variance and correlation matrix is then

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & ... & s_{1p} \\ s_{21} & s_{22} & ... & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & ... & s_{pp} \end{pmatrix} \qquad \mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & ... & r_{1p} \\ r_{21} & r_{22} & ... & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & ... & r_{pp} \end{pmatrix}$$

where $s_{kl} = 1/(n-1)\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$ and $r_{kl} = s_{kl}/(s_{kk}s_{ll})^{1/2}$ are the sample covariance and correlation between $\mathbf{x}_k$ and $\mathbf{x}_l$, the k-th and the l-th variable.

1.2. The graphical methods.

There are several popular and simple plots. The histogram demonstrates the counts over each subinterval with a given width. The box plot specifies the five quartiles: minimum, 25th percentile, median, 75th percentile and maximum. The density plot can be viewed as a smoothed version of the histogram. A slight uncertainty is one can choose the bin width for histogram and likewise the bandwidth for the density plot. The large the bin width or the bandwidth, the smoother the curve.

The the quantile to quantile plot, called qqplot, is used to check whether two variables have similar distribution, or checking whether the distribution of one variable is close to certain given distribution. If two variables have distributions close to each other, their quantiles at the same levels should also be close to each other. Therefore the plot of the quantiles of one variable against the corresponding ones of the other variable should be close to a straight line at 45 degree through origin. Notice that qqplot only shows the distribution similarity of dissimilarity. It does not reveal their dependence.

In finance, a commonly asked question whether the data, typically referring to the returns data, are normal. This is a typical application of the qqnorm plots. Often we wish to check whether a variable follows certain given distribution, which may be suspected based on, for example, domain knowledge or past experience. The most commonly used one is the qqnorm plot, which plots the quantiles of the data against the corresponding quantiles of the standard normal distribution, i.e., it plots

$$(t_1, x_{(1)}),\ (t_2, x_{(2)}),\ ...\ , (t_n, x_{(n)})$$

where t_k is the percentile of the standard normal distribution at level $(k-1/2)/n$. Using Φ to denote the cumulative distribution function for the standard normal distribution, then $t_k = \Phi^{-1}((k - 1/2)/n)$. If the plot is close to a straight line with slope b and intercept a, then we can conclude

that the variable follows a distribution close to normal distribution with mean a and variable b^2. If, for example, the plots are increases sharply above the line on the far right, it implies the variables may have many outliers.

To check the dependence or relation between two variables, one can use scatter plot. Some of the relations, such as positive/negative correlation, linear/nonlinear relation can be identified visually. For several, say p, variables, one can make p by p matrices of plots. Each plot is a scatter plot of one of the p variables against another variable. In R, this can be done by `pairs`.

1.3. Data transformation.

Arbitrary transformation of data is dangerous and the resulting transformed data could entirely lose the original meaning. However, certain "restrictive" transformation can bring the data the desired properties without much distorting the data. Usually such transformations are monotone.

Very often the observations are heavily skewed and the skewness may have adverse effect in the subsequent analysis. A monotone transformation can reduce the skewness. The power transformation, or the Box-Cox transformation, is, for positive x,

$$y = f(x, \lambda) = \begin{cases} (x^\lambda - 1)/\lambda & \text{for } \lambda > 0 \\ \log(x) & \text{for } \lambda = 0 \end{cases}$$

For data skewed to the right, i.e, with a heavy right tail, a power transformation with small λ reduces the right-skewness and brings out a more symmetrized data. Likewise, for data skewed to the left, a power transformation with small λ reduces the right-skewness and brings out a more symmetrized data.

Among the power transformations, those with $\lambda \leq 1$ are concave transformations as the transformation function is easily verified to be concave. The concave transformation can stablize the variance if the small observations are less variable than the large observations, which is often the case. Consider log-transformation as an example. The small observations, after transformation will become more variable, since the log function has large derivatives for small values of the variables. At the same time, the large observations, after transformation will become less variable, since the log function has small derivatives for small values of the variables.

1.4. An example.

Consider the closing price Shanghai index from year 2001 to 2016, which is plotted in Figure 1.1.

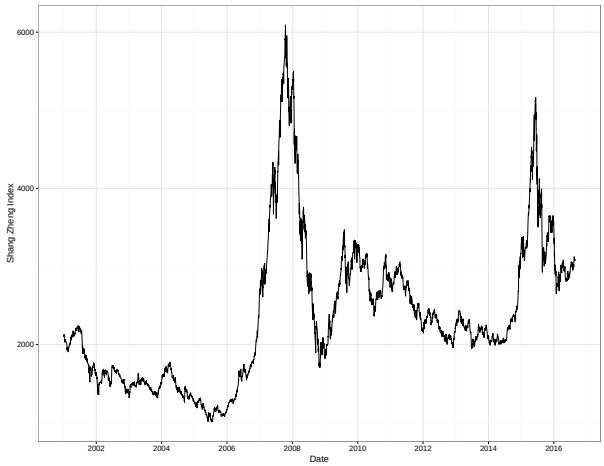


Figure 1.1. *The Shanghai Index from 2001-2016*

The summary statistics are

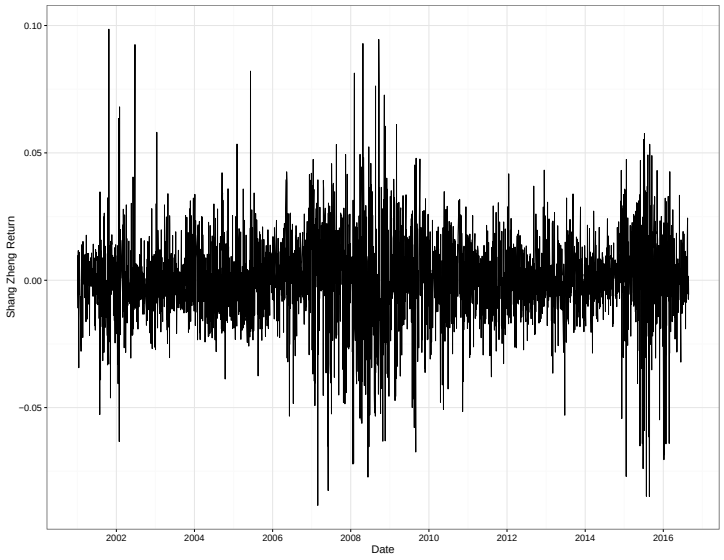


Figure 1.2. *The plots about the returns of Shanghai Index*

minimum	1st quartile	median	3rd quartile	maximum
1628	2199	2373	2903	6092

And the mean and standard deviation are 2373 and 941. This is actually a time series data, but we treat them as a simple random sample from a population.

Normally the returns data are of more interests, which are shown in Figure 1.2.

The summary statistics for daily returns of Shanghai Index are

minimum	1st quartile	median	3rd quartile	maximum
-0.0884100	-0.0073590	0.0005536	0.0084160	0.0985700

The sample mean and standard deviation are 0.0002382 and 0.01662418. Assume that the returns are normally distributed with the sample mean and variance. The chance for an observation to be as small as -0.0884100 would be about one day out of 1.9 million years, assuming 250 trading days a year.

Figure 1.3 shows the histogram, boxplot density plot and qqnorm plot. Notice that there are many outliers as seen in the qqnorm plot which is curved upward in the right and downward in the left.

As the normality assumption of the index returns does not seem to hold. It may be interesting to explore the possibility that the returns may follow a t-distribution. The qqplots with respect to t-distribution for a number of degrees of freedom is show in Figures 1.4-1.5.

It appears from the plots t-distribution with degree of freedom 8 is close to the best fit by t-distributions. A quantitative test, such as Kolmogorov-Smirnov test may be carried.

Consider now six stocks prices data from July 2014 to July 2015: with stock codes: 601398 601939 601288 600028 601857 601088. The prices and returns are plotted in Figure 1.6. and 1.7.

The covarariance matrix is

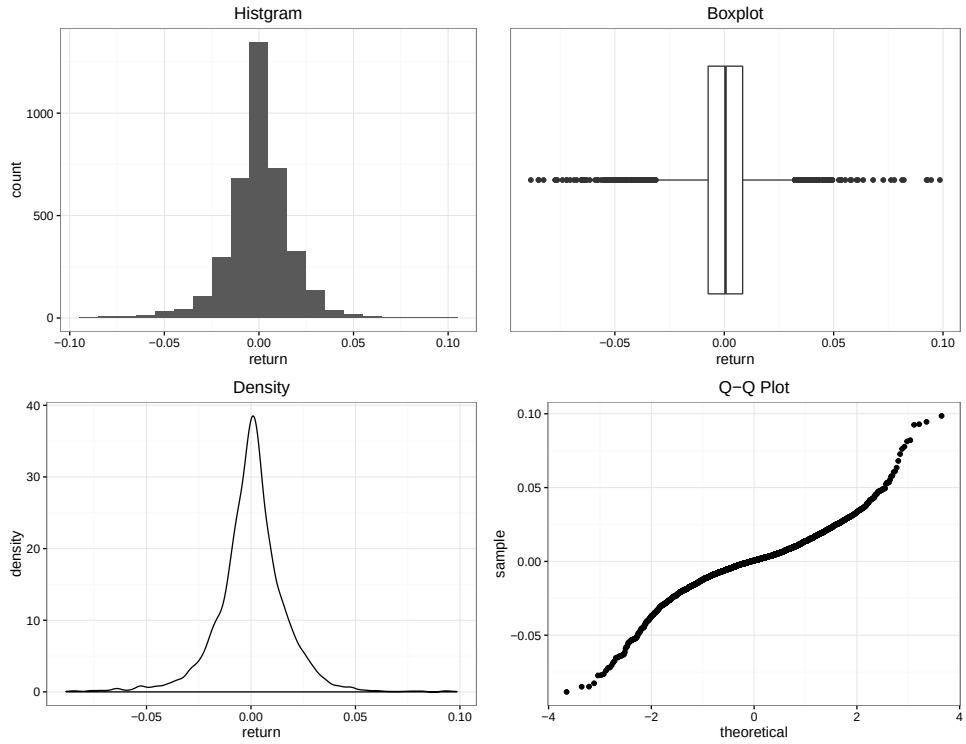


Figure 1.3. *Various distribution plots for Shanghai index returns.*

	601398	601939	601288	600028	601857	601088
601398	0.000418	0.000427	0.000394	0.000332	0.000336	0.000355
601939	0.000427	0.000594	0.000460	0.000402	0.000403	0.000471
601288	0.000394	0.000460	0.000455	0.000364	0.000364	0.000414
600028	0.000332	0.000402	0.000364	0.000555	0.000520	0.000501
601857	0.000336	0.000403	0.000364	0.000520	0.000606	0.000466
601088	0.000355	0.000471	0.000414	0.000501	0.000466	0.000833

And the correlation matrix is

	601398	601939	601288	600028	601857	601088
601398	1.000	0.856	0.903	0.690	0.668	0.602
601939	0.856	1.000	0.885	0.700	0.672	0.669
601288	0.903	0.885	1.000	0.723	0.693	0.672
600028	0.690	0.700	0.723	1.000	0.897	0.736
601857	0.668	0.672	0.693	0.897	1.000	0.655
601088	0.602	0.669	0.672	0.736	0.655	1.000

It is seen that they are highly correlated, not a surprise, since they are all positively correlated with the market returns. Figure 1.5 shows their pairwise scatter plot, where the positive correlation is quite obvious.

Exercise 1.1 ⋆⋆ Exercise 1 of RU Chapter 4.

Exercise 1.2 ⋆⋆ Exercise 2 of RU Chapter 4.

Exercise 1.3 ⋆⋆ Exercise 3 of RU Chapter 4.

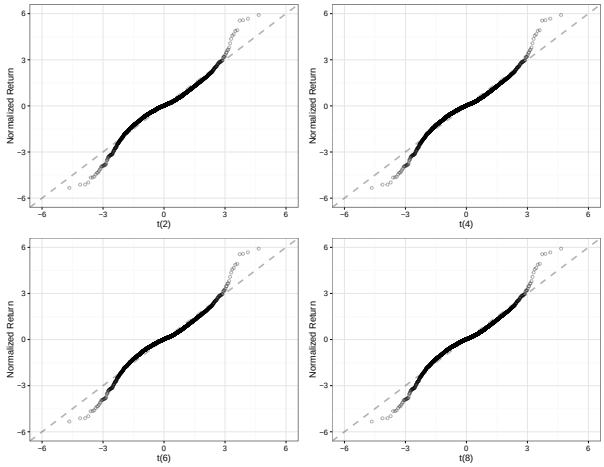


Figure 1.4. *qqplots with respect to t-distributions .*

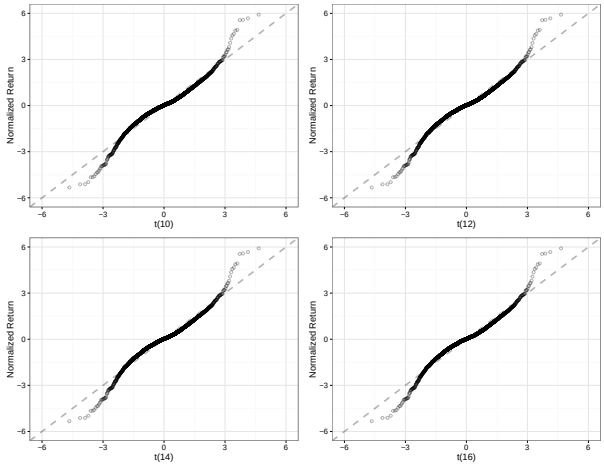
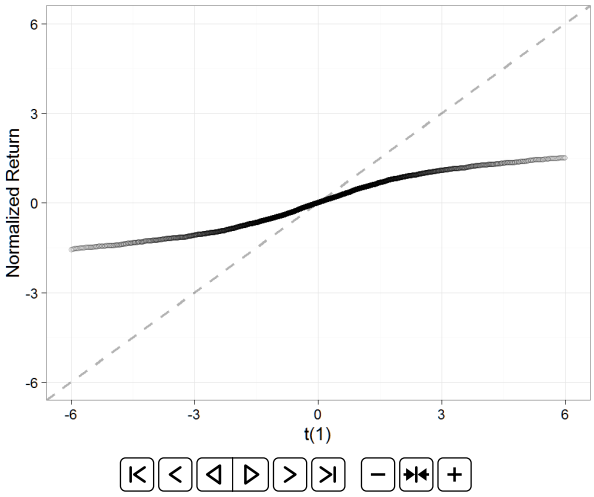


Figure 1.5. *qqplots with respect to t-distributions.*

Exercise 1.4 ⋆⋆ Exercise 4 of RU Chapter 4.

Exercise 1.5 ⋆⋆ Show that, for two random variables X and Y, if Y and $aX + b$ have the same distribution, then the $q_Y(t) = aq_X(t) + b$ for all $t \in (0, 1)$, where $q_Y(t)$ and q_X are the qunatiles of Y and X, respectively.



Animation 1.1. *qqplots with respect to t-distributions.*

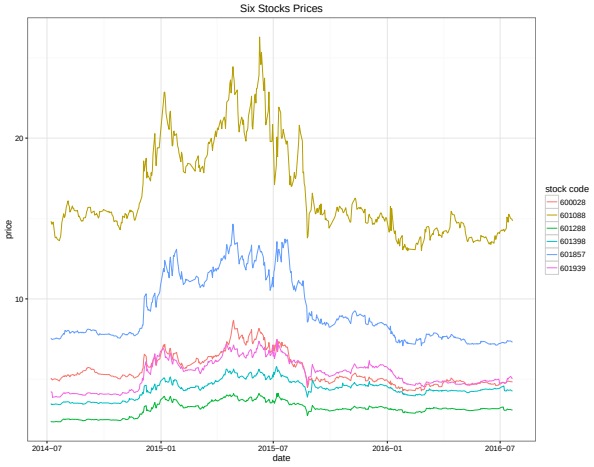


Figure 1.6. *Prices of six stocks.*

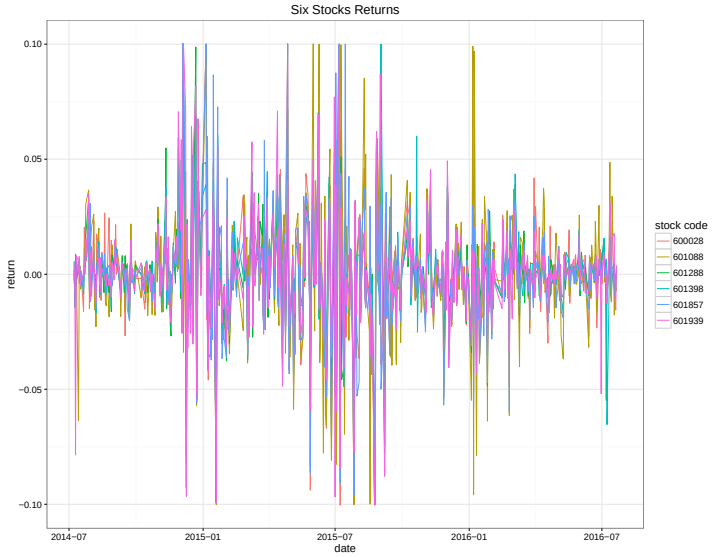


Figure 1.7. *The plots about the returns of the six stocks.*

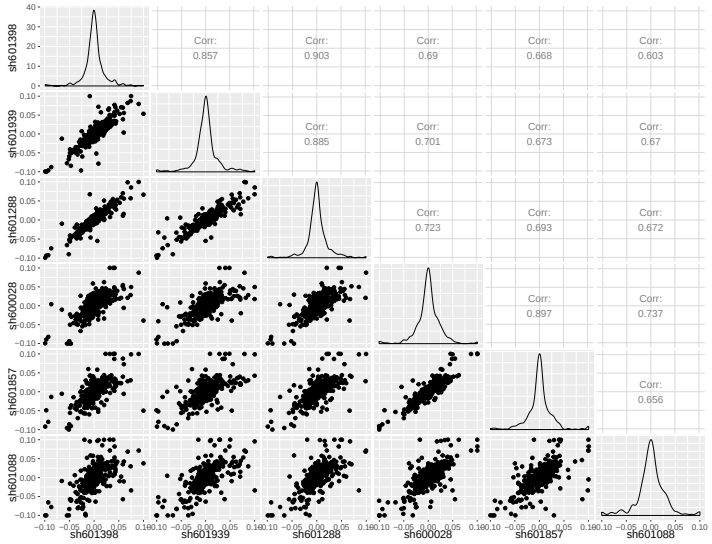


Figure 1.8. *The pairwise plots of the returns of the six stocks.*