Chapter 1

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.

EDA techniques are subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- Plotting the raw data (such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
- Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

A common collection of order statistics used as summary statistics are the five-number summary, sometimes extended to a seven-number summary, and the associated box plot.

1.0.1 Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

1.0.2 Introduction to Statistics : Summary Statistics

In descriptive statistics, summary statistics are used to summarize a set of observations, in order to communicate the largest amount as simply as possible. Statisticians commonly try to describe the observations in

- a measure of location, or central tendency, such as the arithmetic mean
- a measure of statistical dispersion like the standard deviation
- a measure of the shape of the distribution like skewness or kurtosis
- if more than one variable is measured, a measure of statistical dependence such as a correlation coefficient

1.0.3 Univariate Analysis

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency
- the dispersion

1.0.4 Measures of Central Tendency

The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

1.0.5 Descriptive Statistics

- Measures of Centrality
  - Mean
  - Median
- Measures of Dispersion
  - Range
  - Variance
  - Standard Deviation

1.0.6 Measures of Centrality

- Measures of centrality give one representative number for the location of the centre of the distribution of data.
- The most common measures are the ***mean*** and the ***median*** .
- We must make a distinction between a sample mean and a population mean: The sample mean is simply the average of all the items in a sample.
- The population mean (often represented by the Greek letter μ) is simply the average of all the items in a population.
- Because a population is usually very large, the population mean is usually an unknown constant.
- We will return to the matter of population means in due course. For now, we will look at sample means.

1.0.7 Sample Mean

- The sample mean is an estimator available for estimating the population mean . It is a measure of location, commonly called the average, often denoted $\bar{x}$, where x is the data set.
- Its value depends equally on all of the data which may include outliers. It may not appear representative of the central region for skewed data sets.
- It is especially useful as being representative of the whole sample for use in subsequent calculations.
- The sample mean of a data set is defined as :

$$\bar{x} = \frac{\sum x_i}{n}$$

- $\sum x_i$ is the summation of al the elements of x, and n is the sample size.

1.0.8 Using R to compute mean (and median)

When implementing this in R, we would use the following code

```
> x1=c(96, 48, 27, 72, 39, 70, 7, 68, 99 )
> sort(x1)
[1]  7 27 39 48 68 70 72 96 99
> median(x1)
[1] 68
>
> x2=c(96, 48 ,27 ,72, 39, 70, 7, 68)
> sort(x2)
[1]  7 27 39 48 68 70 72 96
> median(x2)
[1] 58
```

1.0.9 Arithmetic mean

- One of the basic quantities is the **arithmetic mean** (it is sometimes called the 'average but there are in fact other measures of average apart from the mean).
- The arithmetic mean is calculated by adding the measures of the number of observations in which you are interested and dividing by the number of observations.

$$\bar{x} = \frac{\sum x}{n}$$

- For our data set $\bar{x} = \frac{22}{5} = \mathbf{4.4}$.

1.0.10 Arithmetic Mean

The arithmetic mean is what is commonly called the average: When the word "mean" is used without a modifier, it can be assumed that it refers to the arithmetic mean. The mean is the sum of all the scores divided by the number of scores. The formula in summation notation is:

$$\mu = \sum X/N$$

where μ is the population mean and N is the number of scores.

If the scores are from a sample, then the symbol M refers to the mean and N refers to the sample size. The formula for M is the same as the formula for .

$$\mu = \sum X/N$$

The mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in skewed distributions since it can be greatly influenced by scores in the tail. Therefore, other statistics such as the median may be more informative for distributions such as reaction time or family income that are frequently very skewed.

- The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values.

  For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

  $$15, 20, 21, 20, 36, 15, 25, 15$$

  The sum of these 8 values is 167, so the mean is $167/8 = 20.875$.

- The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score no. 250 would be the median. If we order the 8 scores shown above, we would get:

  $$15, 15, 15, 20, 20, 21, 25, 36$$

  There are 8 scores and score no. 4 and no. 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

-

1.0.11 Introduction to Statistics

Population mean

When population has a finite quantity N, the population mean can be calculated as follows,
=xiNfori = 1.....N.

Sample Mean

$$x = xin\text{for}i = \{1.....n\}$$

1.0.12 Medians and modes

The median ($\tilde{x}$) is the value that separates a sample into two groups; 50% of observations are greater than the median and 50% are less than it.

The set of n numbers is arranged in ascending order, say as $x_(1), x_(2), x_(3), \ldots x_(n)$, where $x_(1)$ is the smallest of the observations and $x_(n)$ is the largest.

Computation of the median differs for samples that have an odd number size, and samples with an even number size. If sample size n is odd

$$\tilde{x} = x_{(\frac{n+1}{2})},$$

or if n is even

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}}{2}.$$

- A visual inspection of the ordered data set will be useful for a quick determination of the median.

- For example, the median of the numbers 1, 3, 4, 5 and 7 is 4 (check this by rearranging the numbers in order!) and for 1, 3, 4 and 5, the median is (3+4)/2, that is 3.5.
- The **mode** is the most frequently occurring value. There is not necessarily only one such value.
- For example, the figures 1, 2, 2, 3, 5, 9, 9, 11 have two modes: the numbers 2 and 9.

Computing Methods

There is no universal agreement on choosing the quartile values.

One standard formula for locating the position of the observation at a given percentile, y, with n data points sorted in ascending order is:

Case 1: If L is a whole number, then the value will be found halfway between positions L and L+1.

Case 2: If L is a fraction, round to the nearest whole number. (for example, L = 1.2 becomes 1).

The interquartile range is computed by subtracting the first quartile Q_1 from the third quartile Q_3.

$$IQR = Q_3 - Q_1$$

1.0.13 Measures of dispersion

It is unlikely we will only be interested in the average value of our data, we will want to know how large the spread or dispersion of values is about it.
The simplest measure of dispersion is the range. The range is simply the difference of the lowest and highest values. The range is another easy-to-understand measure, but it will clearly be very affected by a few extreme values.
Aside from the range, the most common measures of dispersion are:

- Variance
- Standard deviation
- Mean Absolute Deviation (MAD)
- Inter-quartile range.

1.0.14 Measures of Dispersion

- The first three are related to the use of the arithmetic mean, and are computed using deviations of each observation from the mean.
- The mean absolute deviation (MAD) uses the absolute values of the deviations from the mean and perhaps gives us a more intuitively understandable measure of deviation than variance and standard deviation.

1. Quantile Statistics
2. InterQuartile Range and FiveNumber Summary
3. Quartiles
4. Range
5. Standard Deviation
6. Interquartile Range (IQR)

Variance

The variance of the random variable X is defined to be:

$$V(X) = \sigma^2 = E(X^2) - E(X)^2$$

where E(X) is the expected value of the random variable X.

Notes

- the larger the variance, the further that individual values of the random variable (observations) tend to be from the mean, on average;
- the smaller the variance, the closer that individual values of the random variable (observations) tend to be to the mean, on average;
- the variance and standard deviation of a random variable are always non-negative (i.e. almost always positive, but theoretically you can get a result of zero).

1.0.15 Sample Standard Deviation

Important :The standard deviation is the square root of the variance.

The standard deviation for the sample is called s and the standard deviation for the population is called σ.

The standard deviation is often preferred to the variance as a descriptive measure because it is in the same units as the raw data e.g. if your data is measured in years, the standard deviation will also be in years whereas the variance will be in years squared.

- Standard deviation is the square root of variance
- Standard deviation is commonly used in preference to variance because it is denominated in the same units as the mean.
- For example, if dealing with time units, we could have a variance of something like 25 *square minutes* , whereas the equivalent standard deviation is 5 minutes.
- Population standard deviation is denoted σ.
- Sample standard deviation is denoted s.

The sample standard deviation will be denoted by s and the population standard deviation will be denoted by the Greek letter σ.

The sample variance will be denoted by s^2 and the population variance will be denoted by $sigma^2$.

The variance and standard deviation describe how spread out the data is. If the data all lies close to the mean, then the standard deviation will be small, while if the data is spread out over a large range of values, s will be large. Having outliers will increase the standard deviation.

One of the flaws involved with the standard deviation, is that it depends on the units that are used. One way of handling this difficulty, is called the coefficient of variation which is the standard deviation divided by the mean times 100%

1.0.16 Variance

- Variance is a measure of the dispersion of a set of data points around their mean value.
- (**Important**) Variance is a mathematical expectation of the average squared deviations from the mean.

Variance is the major measure of variability for a data set. To calculate the variance, all data values, their mean, and the number of data values are required. It is expressed in the squared unit of measurement. Its square root is the **standard deviation**. It is symbolized by σ^2 for a population and s^2 for a sample

1.0.17 Quartiles

- For an ordered set of data, which contains n items, the first and third quartiles can be identified as follows
- Q_1 is the value of the $\frac{n+1}{4}$th item
- Q_3 is the value of the $\frac{3(n+1)}{4}$th item
- If $\frac{n+1}{4}$ is not a whole number, use the two items that it is between. The first quartile is the average of those two items.
- If $\frac{n+1}{4}$ is not a whole number, use the two items that it is between. The third quartile is the average of those two items.
- Q_3 is the value of the $\frac{3(n+1)}{4}$th item

$$IQR = Q_3 - Q_1$$

1.1 Numerical Methods

The only way of describing qualitative data is using graphical methods. The first step in describing quantitative data is using graphical methods i.e. get a picture of your data.

Since quantitative data are numeric, however, we can also use numeric methods i.e. calculate a set of numbers that convey a good mental picture of the frequency distribution.

There are two main numerical descriptive measures :

(i) Measure of centrality i.e. measure of the centre of the distribution

(ii) Measure of dispersion i.e. the spread, dispersion of the data

When we know the middle of our distribution and how spread out it is about the middle, we have two numbers which create a concise numerical summary of the data.

The measure of centrality and dispersion we use depends on what our data looks like i.e. the shape of the distribution.

1.1.1 Quantile Statistics

1.1.2 Summary Analysis

A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past. Quite commonly, its purpose is to simply arrive at a few key statistics (for example, mean and standard deviation) which may then either replace the data set or be added to the data set in the form of a summary table.

1.1.3 Outliers

- An outlier is an observation in a data set which is far removed in value from the others in the data set. It is an unusually large or an unusually small value compared to the others.
- An outlier might be the result of an error in measurement, in which case it will distort the interpretation of the data, having undue influence on many summary statistics, for example, the mean and variance.
- Outliers are said to ***skew*** the distribution of values.
- If an outlier is a genuine result, it is important because it might indicate an extreme of behaviour of the process under study.
- For this reason, all outliers must be examined carefully before embarking on any formal analysis. Outliers should not routinely be removed without further justification.

1.1.4 Outliers

Compute the sample mean and median of the following data set

$$X = \{5, 6, 7, 8, 9, 11, 15, 16, 94\}$$

- The sample mean $\bar{x} = 19$

- The median of the sample is 9.
- What causes the discrepancy between mean and median?
- Which measure of centrality do you feel is more representative of the data?
  - The median - most values are between 5 and 16.

Sampling fluctuation refers to the extent to which a statistic takes on different values with different samples. That is, it refers to how much the statistic's value fluctuates from sample to sample.

A statistic whose value fluctuates greatly from sample to sample is highly subject to sampling fluctuation.

For homework 1, You were asked to compute the range of a data set.

Because the Range is the simplest measure of variability, we will include one or two short calculations in this homework.

Compute Sample Standard Deviation

Compute Sample Variance.

1.1.5 Measures of variability

Measures of variability are used to measure how spread out the data is, or how scattered the data is. The main measures of variability are 1)Range 2)Variance and Standard Deviation 3)Percentiles, Quartiles and the Interquartile Range 4)The Coefficient of Variation

1.1.6 Variance and standard deviation

The variance can only be used with the mean since it measures the degree to which the data are closely packed about the mean. If the data are tightly clustered about the mean, the variance will be relatively small. If the data are widely scattered around the mean, the variance will be relatively large.

Like the mean, we use mathematical symbols for the variance and we distinguish between the sample and population variance.

How do we calculate the variance? We can use scientific calculators or we can calculate it by hand using the following formula :

We are calculating the difference between each observation x and the mean . Some of the differences will be positive and some will be negative so we square the differences to make them all positive.

[OVERHEAD]

An easier formula to use if you are calculating the sample standard deviation by hand is

The population variance (which is rarely know) is denoted by the Greek letter (sigma squared).

Important :The standard deviation is the square root of the variance.

The standard deviation for the sample is called s and the standard deviation for the population is called .

The standard deviation is often preferred to the variance as a descriptive measure because it is in the same units as the raw data e.g. if your data is measured in years, the standard deviation will also be in years whereas the variance will be in years squared.

1.1.7 1.2.2 Measures of variability or dispersion

Measures of variability are used to measure how spread out the data is or how scattered the data is. The main measures of variability are :

$$SampleVariance : s2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

1.2 Descriptive Statistics

Descriptive Statistics

- Measures of Centrality
  - Mean
  - Median
- Measures of Dispersion
  - Range
  - Variance
  - Standard Deviation
- Quantiles
- Distribution of data (Skewed or Symmetric)

Measures of Centrality
Sample Mean

- The sample mean is an estimator available for estimating the population mean . It is a measure of location, commonly called the average, often denoted $\bar{x}$, where x is the data set.
- Its value depends equally on all of the data which may include outliers. It may not appear representative of the central region for skewed data sets.
- It is especially useful as being representative of the whole sample for use in subsequent calculations.
- The sample mean of a data set is defined as :

$$\bar{x} = \frac{\sum x_i}{n}$$

- $\sum x_i$ is the summation of al the elements of x, and n is the sample size.

Computing the sample mean
Suppose we roll a die 8 times and get the following scores: $x = \{5, 2, 1, 6, 3, 5, 3, 1\}$

What is the sample mean of the scores $\bar{x}$?

$$\bar{x} = \frac{5 + 2 + 1 + 6 + 3 + 5 + 3 + 1}{8} = \frac{26}{8} = 3.25$$

1.2.1 Measures of centrality

Measures of centrality give one representative number for the location of the centre of the data. The most commonly used measures are the mean the median the mode.

1. The mean or average is the sum of the observations divided by the total number of observations.

We use different symbols to distinguish between the mean of the sample and the mean of the population. The mean (i.e. average) value is denoted with a bar over the set name i.e. " ".

(pronounced x bar) is the sample mean.

Remark : It is very important to familiarise with notation and symbols.

It is usual to give data sets names. Commonly we will name data sets with short simple names such as " " or " ".

is the summation symbol

means sum up all the components of data set x.

The size of a sample is usually denoted as " " . Population mean

The population mean is denoted by the Greek letter (mu). We always use Greek letters when we are talking about numbers associated with the population.

where N is the size of the population.

Note we use capital letters (i.e. N) for the size of the population and lower case letters (i.e. n) for the size of the sample.

Example

A sample data set comprised of five values.

What is the sample mean value of data set " " (i.e. What is ?)

The sample mean is 44

Remark

Correctly this is called the "arithmetic mean".

There are other types of mean , such as the "geometric mean" , that feature in more advanced statistical methods.

The geometric mean could be used to compute the average rate of inflation over a period of time.

1.2.1 The median

The median is defined as the value of the number in the middle position when the data is arranged in numerical order.

It splits the distribution into two halves. The number of values greater than the median is equal to the number of values less than the median.

The important point here is that when you find where the middle of your data is i.e. the position, its the number or value in that position we are interested in.

We calculate the median for both the sample and the population by arranging the numbers in increasing order.

If there is an odd number of observations, the median is the middle number i.e. find the number in the position (k + 1) /2 where k is the number of values you have.

Example (odd number of items) Data set: 1.8 2.7 3.5 4.6 5.4

Number of items k = 5

The median is in position 3 and its value is 3.5 If there is an even number of observations the median is the mean of the two observations occupying the middle position i.e. mean of observations in position k / 2 and (k / 2 + 1) where k is the number of values you have.

Example (even number of items)
Data set: 1.8 2.7 3.5 4.6 Number of items k = 4

The median is the mean of observations in position 2 and 3 in the ordered sequence = (2.7 + 3.5) /2 = 3.1

1.2.2 Comparing Measures of Centrality

We usually use the mean or the median to describe the centre of our distribution. How do we decide between them? Take a look at the following example :

Data : 1.8 2.7 3.5 4.6 5.4 Mean = 3.6 Median = 3.5

Lets change the number 5.4 to 54. What happens to the measures of centrality?

Data : 1.8 2.7 3.5 4.6 54 Mean=13.22 Median=3.5

We can see from this example that the mean is sensitive to extreme values whereas the median is unaffected by extreme values.

If our histogram is symmetric,then we can say mean = median = mode. The most commonly used measure of centrality for a symmetric distribution is the mean. Skew If our histogram is skewed , then we can say we will use the median as our measure of centrality since it will be unaffected by the values that are causing the histogram to be skewed. When we have a left-skewed distribution then mean ¡ median. When we have a right-skewed distribution then mean ¿ median.

[OVERHEAD]

2.2.2 Measures of variability Measures of variability are used to measure how spread out the data is, or how scattered the data is. The main measures of variability are 1)Range 2)Variance and Standard Deviation 3)Percentiles, Quartiles and the Interquartile Range 4)The Coefficient of Variation

Example

Find the three quartiles and the IQR of the following data

15 34 7 12 18 9 1 42 56 28 13 24 35

First sort the data set into ascending order

1 7 9 12 13 15 18 24 28 34 35 42 56

Count how many items are in the data set (answer 13 items)

Which value is the second quartile, which is the median (answer: the 7th item, which is 18)

Q1 median of data less than or equal to median (7 items)

17 9 12 13 15 18 Answer: the 4th item, which is 12 Q3 median of data greater than or equal to median (also 7 items)

18 24 28 34 35 42 56 Answer: the 4th item of this 7, which is 34

The three quartiles are therefore Q1 = 12, Q2 = 18, Q3 = 34 The interquartile range is therefore Q3 Q1 = 22

The Coefficient of Variation [page 26]

Population C.V.

Population standard deviation

Population mean

Sample C.V.

Sample standard deviation

Sample mean

The coefficient of variation for different distribution are compared and the distribution with the largest CV value has the greatest spread.

2.