Detail Exploratory Analysis By Kamakshigari Suresh

Exploratory Data Analysis (EDA):

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

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models;
7. and determine optimal factor settings.

Typical data format and the types of EDA:

Exploratory data analysis is generally cross-classified in two ways.

First : each method is either non-graphical or graphical.

Second: each method is either univariate or multivariate (usually just bivariate).

It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

Techniques:

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

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

How Does Exploratory Data Analysis differ from Classical Data Analysis?

EDA is a data analysis approach

There are three approaches in EDA.

1. Classical

2. Exploratory (EDA)
3. Bayesian

For classical analysis, the sequence is

Problem => Data => Model => Analysis => Conclusions

For EDA, the sequence is

Problem => Data => Analysis => Model => Conclusions

For Bayesian, the sequence is

Problem => Data => Model => Prior Distribution => Analysis => Conclusions

Method of dealing with underlying model for the data distinguishes the 3 approaches:

Thus for **classical analysis**, the data collection is followed by the imposition of a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.

For **EDA**, the data collection is not followed by a model imposition; rather it is followed immediately by analysis with a goal of inferring what model would be appropriate.

Finally, for a **Bayesian analysis,** the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a dataindependent distribution on the parameters of the selected model; the analysis thus consists of formally combining both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.

Further discussion of the distinction between the classical and EDA approaches

The Exploratory Data Analysis approach does not impose deterministic or probabilistic models on the data.
On the contrary, the EDA approach allows the data to suggest admissible models that best fit the data.

Focus:

For exploratory data analysis, the focus is on the data--its structure, outliers, and models suggested by the data.

Techniques :

Classical:

Classical techniques are generally quantitative in nature.
They include ANOVA, t tests, chi-squared tests, and F tests.

Exploratory EDA:

Techniques are generally graphical. They include scatter plots, character plots, box plots, histograms, bi histograms, probability plots, residual plots, and mean plots.

Classical :

Classical techniques serve as the probabilistic foundation of science and engineering; the most important characteristic of classical techniques is that they are rigorous, formal, and "objective".

Exploratory EDA:

Exploratory EDA techniques do not share in that rigor or formality. EDA techniques make up for that lack of rigor by being very suggestive, indicative, and insightful about what the appropriate model should be. EDA techniques are

subjective and depend on interpretation which may differ from analyst to analyst, although experienced analysts commonly arrive at identical conclusions.

Data Treatment:

The EDA approach, on the other hand, often makes use of (and shows) all of the available data. In this sense there is no corresponding loss of information.

Assumptions:

Classical :

The **"good news"** of the classical approach is that tests based on classical techniques are usually very sensitive--that is, if a true shift in location, say, has occurred, such tests frequently have the power to detect such a shift and to conclude that such a shift is "statistically significant".

The **"bad news"** is that classical tests depend on underlying assumptions (e.g., normality), and hence the validity of the test conclusions becomes dependent on the validity of the underlying assumptions. Worse yet, the exact underlying assumptions may be unknown to the analyst, or if known, untested.

Thus the validity of the scientific conclusions becomes intrinsically linked to the validity of the underlying assumptions. **In practice, if such assumptions are unknown or untested, the validity of the scientific conclusions becomes suspect.**

Exploratory EDA:

Exploratory Many EDA techniques make little or no assumptions--they present and show the data--all of the data--as is, with fewer encumbering assumptions.

How Does Exploratory Data Analysis Differ from Summary Analysis?

Summary :

A summary analysis is simply a numeric reduction of a historical data set. It is quite passive. Its focus is in the past.

Exploratory :

EDA has as its broadest goal the desire to gain insight into the engineering/scientific process behind the data. Whereas summary statistics are passive and historical, EDA is active and futuristic. In an attempt to "understand" the process and improve it in the future, EDA uses the data as a "window" to peer into the heart of the process that generated the data. There is an archival role in the research and manufacturing world for summary statistics, but there is an enormously larger role for the EDA approach.

What are the EDA Goals?

The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as:

1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. **conclusions as to whether individual factors are statistically significant - Very Important**
8. optimal settings

Statistics and data analysis procedures can broadly be split into two parts:

Quantitative :

Quantitative techniques are the set of statistical procedures that yield numeric or tabular output. Examples of quantitative techniques include:

hypothesis testing
analysis of variance
point estimates and
confidence intervals
least squares regression

These and similar techniques are all valuable and are mainstream in terms of classical analysis.

Graphical On the other hand, there is a large collection of statistical tools that we generally refer to as graphical techniques.
These include:

scatter plots
histograms
probability plots

residual plots
box plots
block plots

EDA Approach Relies Heavily on Graphical Techniques:

The EDA approach relies heavily on these and similar graphical techniques. Graphical procedures are not just tools that we could use in an EDA context, they are tools that we must use. Such graphical tools are the shortest path to gaining insight into a data set in terms of testing assumptions.

model selection
model validation
estimator selection
relationship identification
factor effect determination
outlier detection

An EDA/Graphics Example:

Anscombe

Example A simple, classic (Anscombe) example of the central role that graphics play in terms of providing insight into a data set starts with the following data set:

Data

X	Y
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.84
7.00	4.82
5.00	5.68

Summary Statistics

If the goal of the analysis is to compute summary statistics plus determine the best linear fit for Y as a function of X, the results might be given as:

$N = 11$
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept = 3
Slope = 0.5
Residual standard deviation = 1.237
Correlation = 0.816

The above quantitative analysis, although valuable, gives us only limited insight into the data.

```
print('Mean of X:' , x.mean())
print('Mean of Y:' , y.mean())
print('Total Len : ', len(x))
print('Correlation:', df.corr())
```

```
Mean of X: 9.0
Mean of Y: 7.533636363636364
Total Len :  11
Correlation:          x         y
x  1.00000  0.79636
y  0.79636  1.00000
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.634
Model:	OLS	Adj. R-squared:	0.594
Method:	Least Squares	F-statistic:	15.60
Date:	Tue, 11 Aug 2020	Prob (F-statistic):	0.00335
Time:	07:25:28	Log-Likelihood:	-17.377
No. Observations:	11	AIC:	38.75
Df Residuals:	9	BIC:	39.55
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.1326	1.181	2.653	0.026	0.461	5.804
x1	0.4890	0.124	3.950	0.003	0.209	0.769

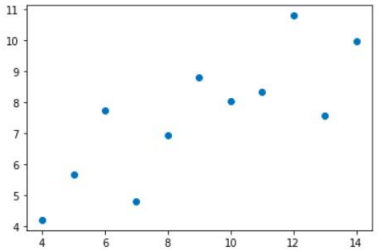
Omnibus:	0.177	Durbin-Watson:	3.176
Prob(Omnibus):	0.915	Jarque-Bera (JB):	0.370
Skew:	-0.037	Prob(JB):	0.831
Kurtosis:	2.104	Cond. No.	29.1

The above quantitative analysis, although valuable, gives us only limited insight into the data.

Scatter Plot In contrast, the following simple scatter plot of the data

```
import matplotlib.pyplot as plt
plt.scatter(x1,y1)
```

```
<matplotlib.collections.PathCollection at 0x7efec8e957b8>
```



suggests the following:

1. The data set "behaves like" a linear curve with some scatter;
2. there is no justification for a more complicated model (e.g., quadratic);
3. there are no outliers;

4. 4. the vertical spread of the data appears to be of equal height irrespective of the X-value;
5. this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

EDA Assumptions :

1. Underlying Assumptions
2. Importance
3. Testing Assumptions
4. Importance of Plots
5. Consequences

Underlying Assumptions:

Assumptions Underlying a Measurement Process:

There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

1. **random drawings;**
2. **from a fixed distribution;**
3. **with the distribution having fixed location; and**
4. **with the distribution having fixed variation.**

Assumptions for Univariate Model:

the data are uncorrelated with one another;
the random component has a fixed distribution;
the deterministic component consists of only a constant;
and the random component has fixed variation.

Techniques for Testing Assumptions:

Testing Underlying Assumptions Helps Assure the Validity of Scientific and Engineering Conclusions.

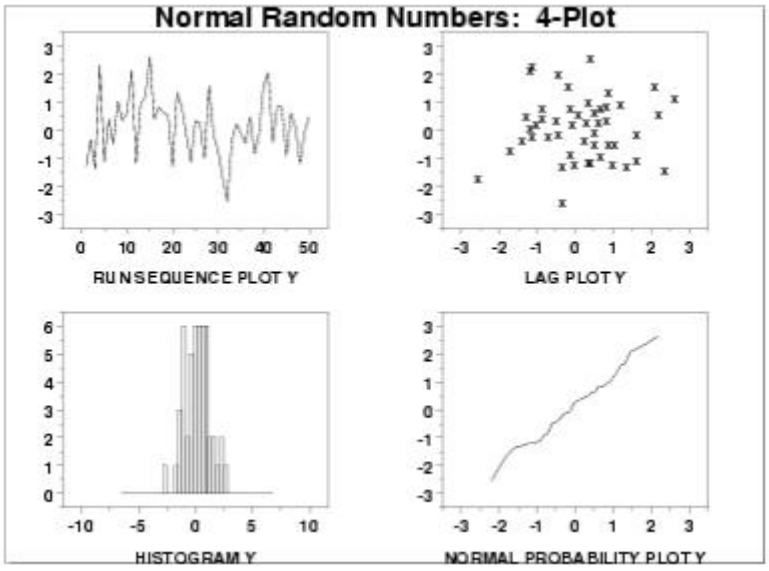
Because the validity of the final scientific/engineering conclusions is inextricably linked to the validity of the underlying univariate assumptions, it naturally follows that there is a real necessity that each and every one of the above four assumptions be routinely tested.

Four Techniques to Test Underlying Assumptions

The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:

1. run sequence plot (Yi versus i)
2. lag plot (Yi versus Yi-1)
3. histogram (counts versus subgroups of Y)
4. normal probability plot (ordered Y versus theoretical ordered Y).

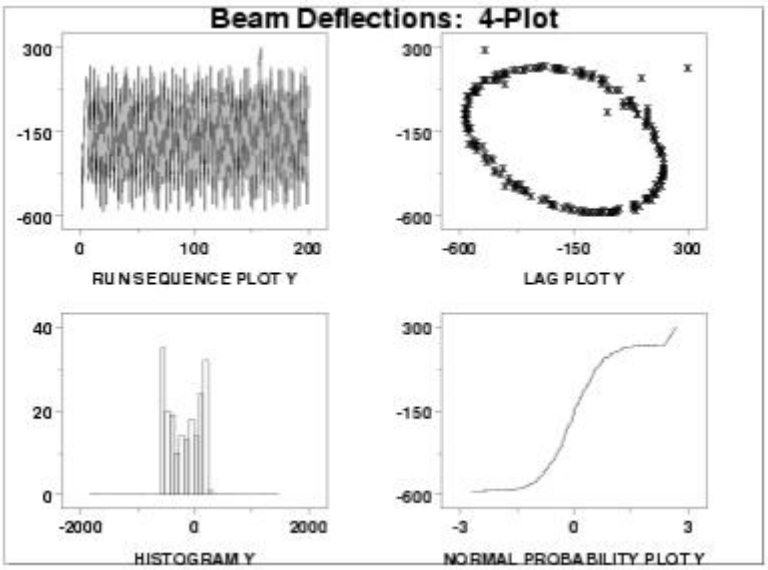
Sample Plot: Assumptions Hold



Sample Plot:

Assumptions Do Not Hold:

If one or more of the four underlying assumptions do not hold, then it will show up in the various plots as demonstrated in the following example.



Interpretation of 4-Plot:

Interpretation of EDA Plots: Flat and Equi-Banded, Random, Bell-Shaped, and Linear

The four EDA plots discussed on the previous page are used to test the underlying assumptions:

1. **Fixed Location:** If the fixed location assumption holds, then the run sequence plot will be flat and non-drifting.
2. **Fixed Variation:** If the fixed variation assumption holds, then the vertical spread in the run sequence plot will be the approximately the same over the entire horizontal axis.
3. **Randomness:** If the randomness assumption holds, then the lag plot will be structureless and random.
4. **Fixed Distribution:** If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then 1. the histogram will be bell-shaped, and 2. the normal probability plot will be linear.

Plots Utilized to Test the Assumptions :

Conversely, the underlying assumptions are tested using the EDA plots:

Run Sequence Plot: If the run sequence plot is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.

Lag Plot: If the lag plot is structureless, then the randomness assumption holds. Histogram: If the histogram is bell-shaped, the underlying distribution is symmetric and perhaps approximately normal.

Normal Probability Plot: If the normal probability plot is linear, the underlying distribution is approximately normal. If all four of the assumptions hold, **then the process is said definitionally to be "in statistical control".**