COMP 499 Introduction to Data Analytics

Lecture 9 — Exploratory Data Analysis

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

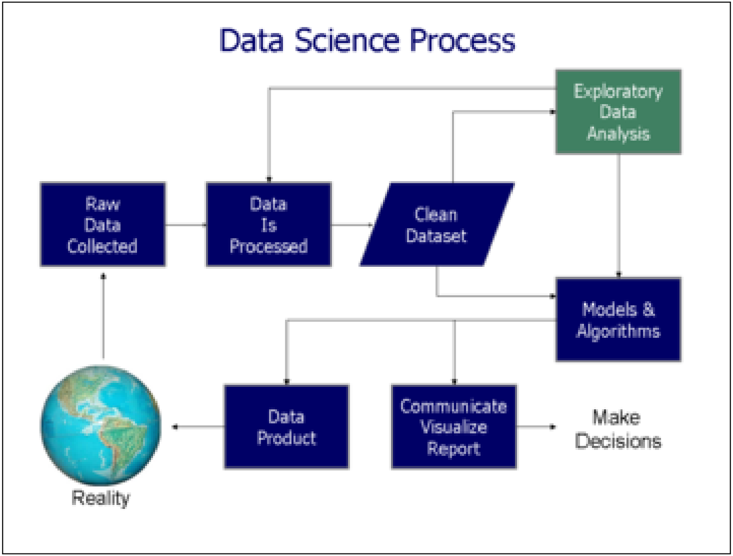
gregb@cs.concordia.ca

Exploratory Data Analysis (EDA)

Outline of Lecture

- EDA: Concepts, Steps, Methods
- Skewness and Kurtosis
- Regression: Curve Fitting
- Dimension reduction: PCA
- Clustering
- Feature Engineering

Data Analytics



wikipedia

Exploratory Data Analysis

Tukey 1977 book

John Tukey (1977), Exploratory Data Analysis, Addison-Wesley.

NIST Engineering Statistics Handbook

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; and
7. determine optimal factor settings.

The EDA approach is not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

`https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm`

Exploratory Data Analysis

NIST Engineering Statistics Handbook

EDA is an approach to data analysis
that postpones the usual assumptions about what kind of model
the data follow
with the more direct approach of
allowing the data itself
to reveal its underlying structure and model.

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Exploratory Data Analysis (EDA)

- get a general sense of the data
- interactive and visual
  - (cleverly/creatively) exploit human visual power to see patterns
    - 1 to 5 dimensions (e.g. spatial, color, time, sound)
  - e.g. plot raw data/statistics, reduce dimensions as needed
- data-driven (model-free)
- especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g. log(x))
- *http://www.itl.nist.gov/div898/handbook/eda/eda.htm*

- Bottom line: it is always well worth looking at your data!

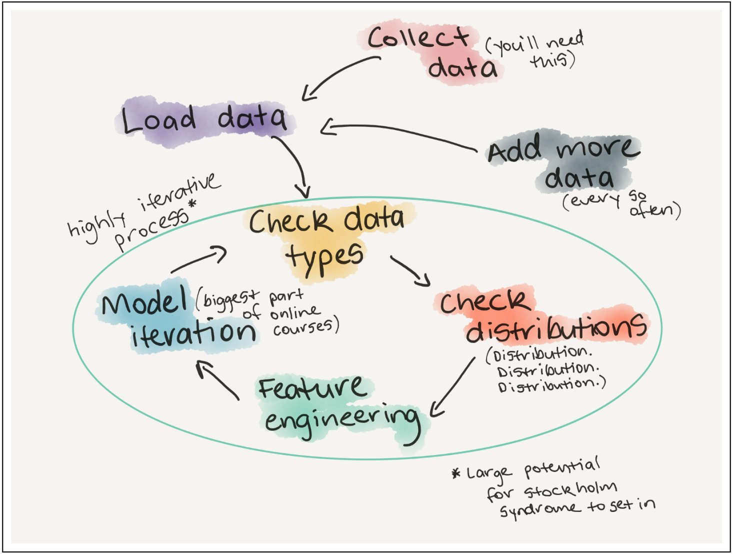
EDA Checklist

1. What question(s) are you trying to solve (or prove wrong)?
2. What kind of data do you have and how do you treat different types?
3. What's missing from the data and how do you deal with it?
4. Where are the outliers and why should you care about them?
5. How can you add, change or remove features to get more out of your data?

Daniel Bourke, A Gentle Introduction to Exploratory Data Analysis,

https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184

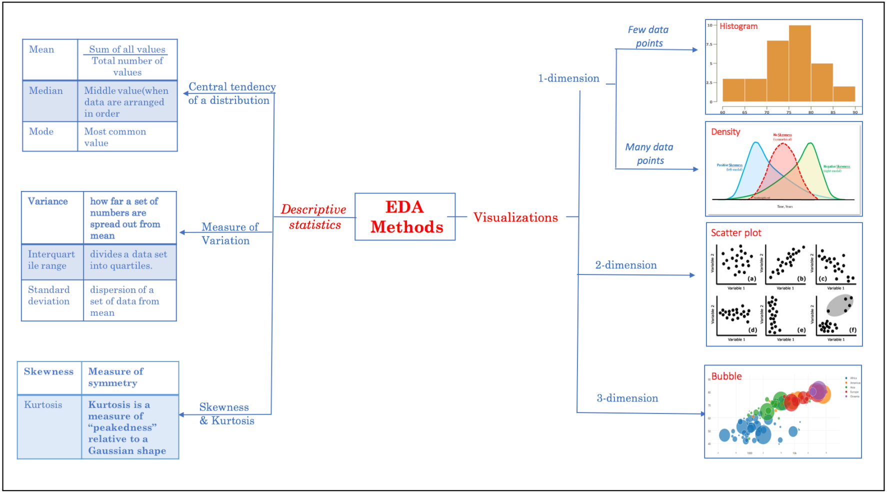
EDA Circle of Life



Daniel Bourke, A Gentle Introduction to Exploratory Data Analysis,

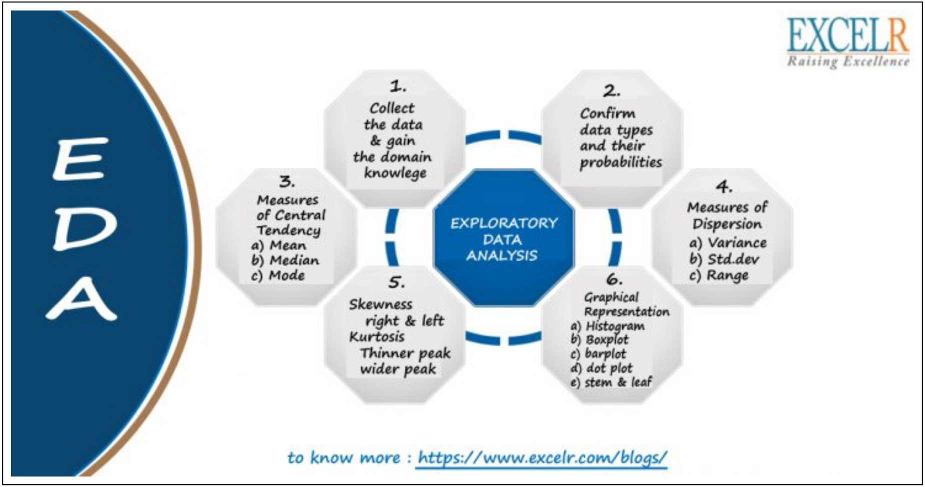
https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184

EDA Methods



Mean
Sum of all values / Total number of values
Median
Middle value(when data are arranged in order
Mode
Most common value
Central tendency of a distribution
Variance
how far a set of numbers are spread out from mean
Interquartile range
divides a data set into quartiles.
Standard deviation
dispersion of a set of data from mean
Measure of Variation
Skewness
Measure of symmetry
Kurtosis
Kurtosis is a measure of "peakedness" relative to a Gaussian shape
Skewness & Kurtosis
Descriptive statistics
EDA Methods
Visualizations
1-dimension
Few data points
Many data points
2-dimension
3-dimension
Histogram
Density
Scatter plot
Bubble

EDA Steps


EDA
EXCELR
Raising Excellence
1. Collect the data & gain the domain knowlege
2. Confirm data types and their probabilities
3. Measures of Central Tendency
a) Mean
b) Median
c) Mode
4. Measures of Dispersion
a) Variance
b) Std.dev
c) Range
5. Skewness right & left
Kurtosis Thinner peak wider peak
6. Graphical Representation
a) Histogram
b) Boxplot
c) barplot
d) dot plot
e) stem & leaf
EXPLORATORY DATA ANALYSIS
to know more : https://www.excelr.com/blogs/

EDA Key Concepts

What are the key concepts about EDA?

- 2 types of Data Analysis
  - *Confirmatory* data analysis
  - *Exploratory* data analysis
- 4 objectives of EDA
  - *Discover* Patterns
  - *Spot* Anomalies
  - *Frame* Hypothesis
  - *Check* Assumptions
- 2 methods for exploration
  - *Univariate* Analysis
  - *Bivariate* Analysis
- Stuff done during EDA
  - *Trends*
  - *Distributions*
  - *Mean*
  - *Median*
  - *Outlier*
  - *Spread measurement (SD)*
  - *Correlations*
  - *Hypothesis testing*
  - *Visual exploration*

EDA: Skewness and Kurtosis

Besides analyses to characterize central tendency and variability ...
a further characterization of the data includes **skewness** and **kurtosis**.

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
That is, data sets with high kurtosis tend to have heavy tails, or outliers.
Data sets with low kurtosis tend to have light tails, or lack of outliers.
A uniform distribution would be the extreme case.

Detecting Skewness and Kurtosis

The **histogram** is an effective graphical technique for showing both the skewness and kurtosis of data set.

`https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm`

Process: Exploratory Data Analysis

Exploratory Data Analysis

Learn about the properties of the data

Steps for Exploratory Data Analysis

- Descriptive statistics: mean/median and variance, quantiles, outliers
- Correlation
- Fitting curves and distributions
- Dimension reduction
- Clustering

Regression: Curve Fitting

Regression Analysis

a set of statistical processes for estimating the relationships among variables

helps one understand how the typical value of the dependent variable changes
when any one of the independent variables is varied,s
while the other independent variables are held fixed.

Linear Regression

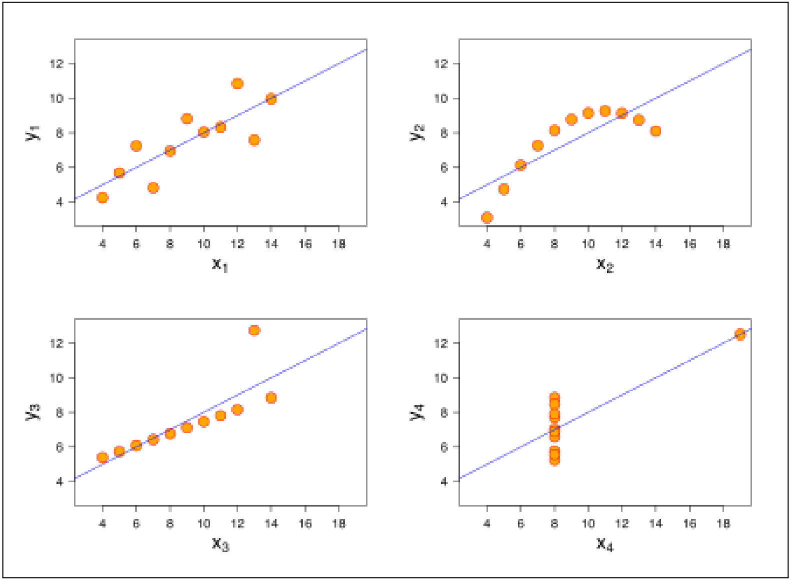
fit a line to (x,y) data
y is dependent variable, x is independent variable

Curve Fitting

Can fit other forms of curves to data

Regression: Curve Fitting

Anscombe's Quartet



Dimension reduction: PCA

Principal Component Analysis (PCA)

Aim: to identify the combinations of variables that explain the variability in the data set

Method

Transform original set of correlated variables
into
set of orthogonal (independent) variables

- linear combination of original variables
- first principal component accounts for as much of variability as possible
- second PC accounts for as much of remaining variability as possible
- etc

Map to PC for Dimension Reduction

Clustering

Clustering

brings together *"similar"* observations

Distances

Many potential distances
Euclidean distance
Manhattan distance
Cosine distance

k-Means Clustering

Creates k clusters, pre-defined k
Start with k random centroids
Iteratively assign points to nearest centroid,
and recompute centroids

Agglomerative Clustering

Start each point is cluster
Iteratively merge closest clusters

Clusters define Nominal Dimension

Clustering: Consistency of Data

Cluster/sort data values

To bring together
duplicate and similar data values
to make it easy to see differences/errors
(See OpenRefine video 1 of 3)

Cluster observations

To bring together
duplicate and similar observations
to make it easy to see differences/errors

Check for consistency

Differences need to be investigated

Feature Engineering

Feature

A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Any attribute could be a feature, as long as it is useful to the model.

Process of Feature Engineering

- Brainstorming or Testing features;
- Deciding what features to create;
- Creating features;
- Checking how the features work with your model;
- Improving your features if needed;
- Go back to brainstorming/creating more features until the work is done.

See video 3, Ryan Baker, Coursera, Big Data Week 3 Feature Engineering

`https://www.youtube.com/watch?v=drUToKxEAUA`

Feature Creation

Aggregation

Basic aggregation operators

- sum
- mean, media, mode
- frequency

Other

- binning

Transformation

Apply a transformation to features

- normalization, unification, resolution, regularization
- log
- feature split
- scaling

Feature Creation: Binning

Numerical Data to Categorical Data

Example: Age

Define **bins**:

`Infant` for *age* between 0 – 4
`Child` for *age* between 5 – 12
`Teen` for *age* between 13 – 19
`YoungAdult` for *age* between 20 – 29
`Adult` for *age* between 30 – 44
`Mature` for *age* between 45 – 64
`Senior` for *age* between 65 – 79
`Elderly` for *age* 80 and over

Feature Creation: Splitting

Feature Splitting

Example: Name split to FirstName, LastName

Example: Date 2019-06-21 split to Year, Month, Day

Python featuretools

name	type	description
num_true	aggregation	Finds the number of 'True' values in a boolean.
percent_true	aggregation	Finds the percent of 'True' values in a boolean feature.
time_since_last	aggregation	Time since last related instance.
num_unique	aggregation	Returns the number of unique categorical variables.
avg_time_between	aggregation	Computes the average time between consecutive events.
all	aggregation	Test if all values are 'True'.
min	aggregation	Finds the minimum non-null value of a numeric feature.
mean	aggregation	Computes the average value of a numeric feature.
seconds	transform	Transform a Timedelta feature into the number of seconds.
second	transform	Transform a Datetime feature into the second.
and	transform	For two boolean values, determine if both values are 'True'.
month	transform	Transform a Datetime feature into the month.
cum_sum	transform	Calculates the sum of previous values of an instance for each value in a time-dependent entity.
percentile	transform	For each value of the base feature, determines the percentile in relation
time_since_previous	transform	Compute the time since the previous instance.
cum_min	transform	Calculates the min of previous values of an instance for each value in a time-dependent entity.

Feature Contribution

Correlation Example

r^2 measures how much of variation is explained by linear regression

Contribution to Model

When building a model from your dataset,
does the technique allow you
to know the contribution of each feature?

Compare with PCA

PCA finds principal orthogonal components
components are ranked by contribution
components are defined as combinations of features