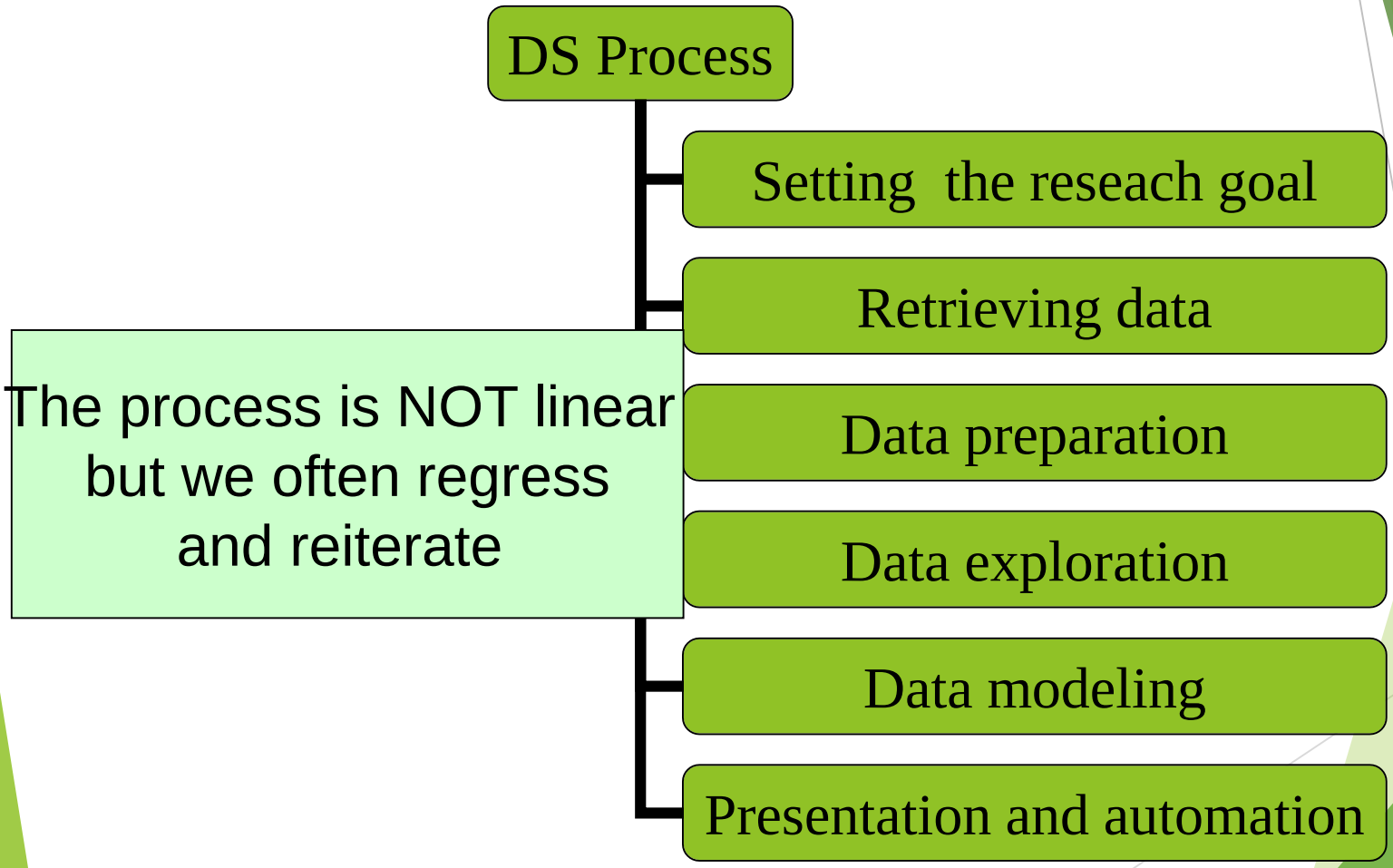

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

The Data Science Process

By Muhammad Ali Mughal

Founder & Chairperson Jawan Pakistan

Data Science Process



1. Setting the Research Goal

- A project starts by the WHAT, the Why and the HOW?
- Outcome of first phase:
  - Research goal
  - Understanding of the context
  - Project plan with well-defined milestones
  - List of deliverables
- Involves Senior personnel

Outcome: A project d

1. Setting the Research Goal

- Research goal
  - Understanding buiseness goals and context is essential
    - Data sceintists often fail to understand buisness goals
  - Take this phase seriously
    - Spend more time
    - Involve people with domain expertise

1. Setting the Research Goal

- Project charter
  - A clear research goal
  - The project mission and context
  - How analysis will be done?
  - Resources required
  - Proof that it is a achievable project or proof of concept
  - List of Deliverables
  - Timeline

2. Retrieving Data

- Start with data available internally
  - Most companies manage data about their activities
    - Databases, data warehouse, flat files, workbooks and even hard copies
  - Access to internal data depends upon company policies and politics

2. Retrieving Data

- Getting data from other sources
  - Data available online
  - Data published by companies and research labs
  - Often this data is of good quality
- Quality check
  - It is often easy to spot errors at this stage
  - Make sure that data is of right data type and it is in the right format

3. Data Preparation

- Data cleansing
  - Errors from data entry
  - Impossible values
  - Missing values
  - Outliers
  - Spaces, typos…etc
  - Error against codebook

3. Data Preparation

- Data transformation
  - Aggregating data
  - Extrapolating data
  - Derived measures
  - Creating dummies
  - Reducing number of vairables

3. Data Preparation

- Combining data
  - Merging / joining data sets
  - Set operators
  - Creating views

3. Data Preparation – Cleansing Data (removing errors in data)

Error description	Possible solution
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

Diagnostic plots can
detecting th

3. Data Preparation – Cleansing Data (removing errors in data)

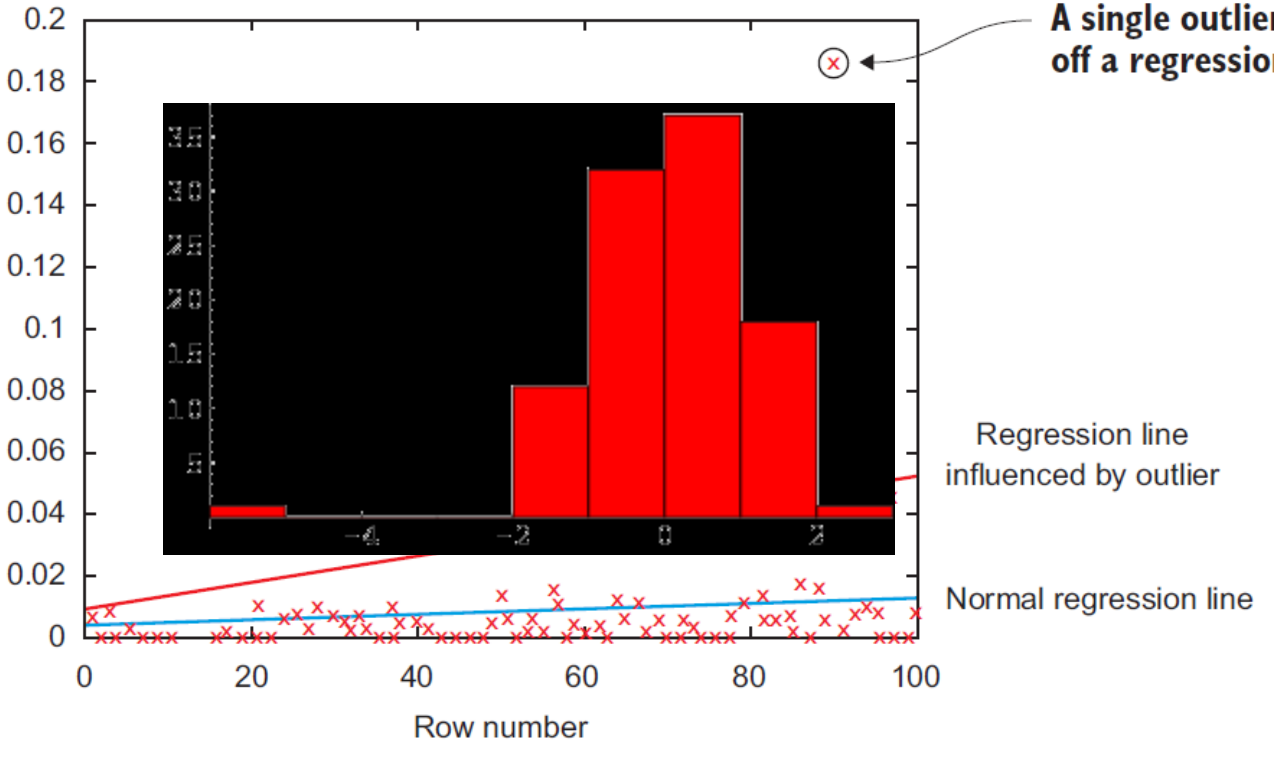
- Data entry errors
  - Typing mistakes
  - Errors during transmission
  - Errors in extract, transform, and load phase (ETL)
  - For small data sets corrections can be done manually
  - For large data sets short scripts can be written to detect and correct errors

3. Data Preparation – Cleansing Data (removing errors in data)

- Sanity checks
  - Check for incorrect values
  - Check for out of range values
  - Check for data type mismatch
  - Often these checks can be easily enforced in Programs

3. Data Preparation – Cleansing Data (removing errors in data)

- Outl
  - A
    I
  -



3. Data Preparation – Cleansing Data (removing errors in data)

An overview of techniques to handle missing data

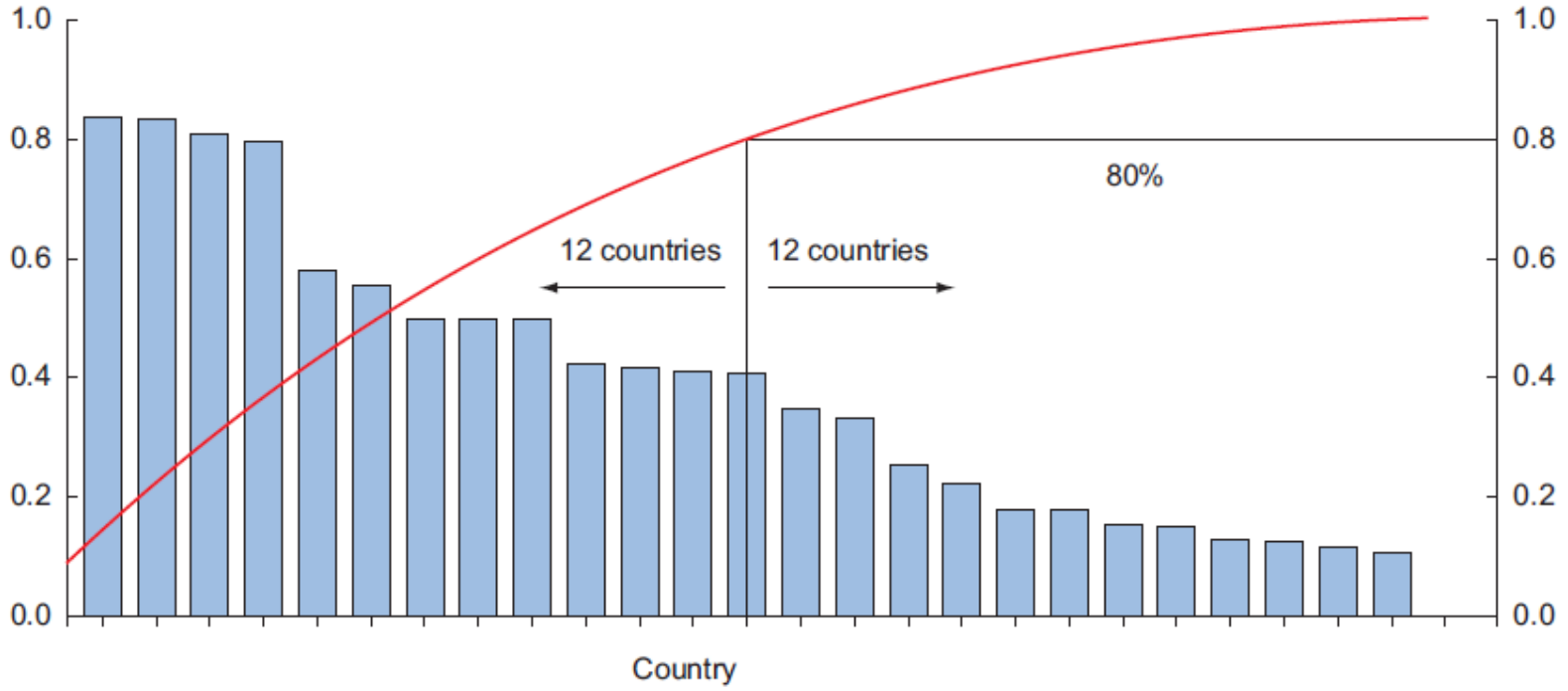
Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

4. Exploratory Data Analysis

- EDA usually employs graphical techniques to get better understanding of data, such as Histograms, Line charts, Box plots etc.
  - However, some other techniques like tabulation, clustering, simple model building can also be used
- Often undetected errors in data are discovered here

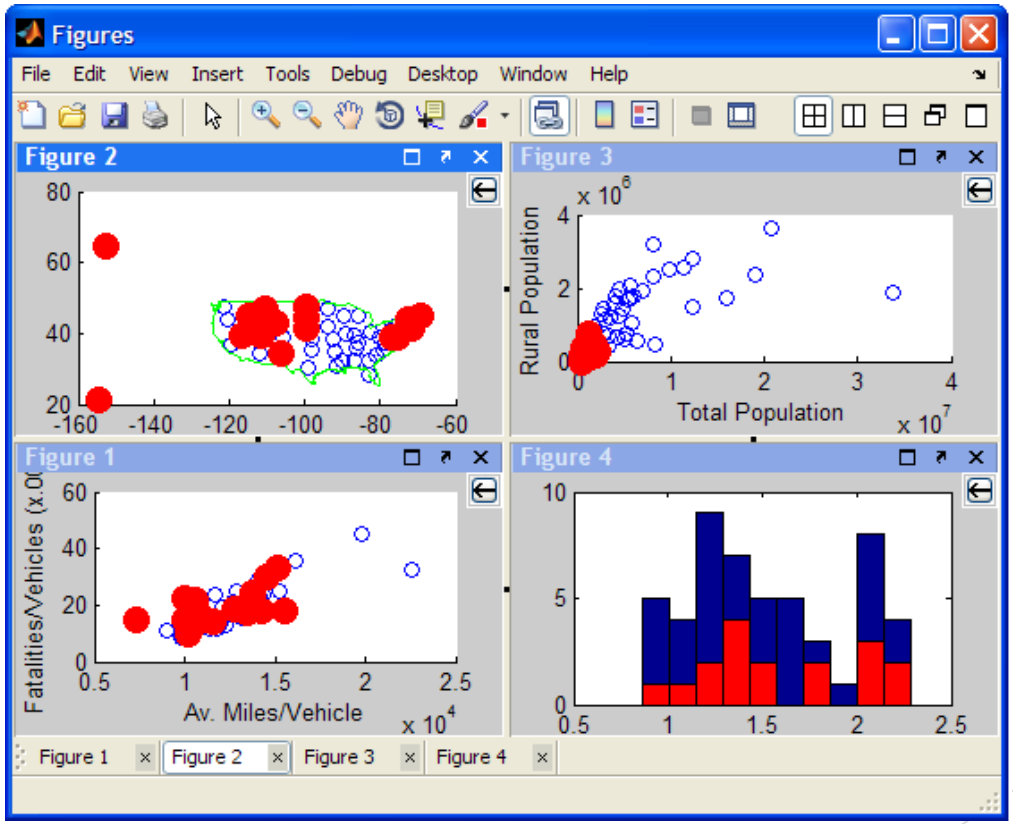
4. Exploratory Data Analysis

- Overlay plots is also a useful EDA technique, such as Pareto diagram



4. Exploratory Data Analysis

- Brushing and Linking



5. Build the Models

- The goal is to build models to do prediction, classification, or clustering on given data
- Involves machine learning, data mining and statistical techniques
- Three main steps
  - Model and variable selection
  - Model execution
  - Model diagnostic and model comparison

5. Build the Models – Model and Variable Selection

- After EDA, we have an idea of variables that could be useful in constructing the model
- Proper model selection requires expertise
- Need to consider model Performance
- Model selection must also consider project characteristics and limitations

5. Build the Models – Model and Variable Selection

- Other factors to consider
  - Must the model be moved to a production environment and, if so, would it be easy to implement?
  - How difficult is the maintenance on the model: how long will it remain relevant if left untouched?
  - Does the model need to be easy to explain?

We will start implementing our first models from Week 3

6. Presentation and Automation

- Presenting your results to the stakeholders and industrializing your analysis process for repetitive reuse and integration with other tools.
  - Needs to automate our model
- Presenting results of your hard work needs SOFT skills
  - Correctly presenting your results is as important as the modeling task itself