Data Processing and Visualization using Off-the Shelf Tools

Prepared By: Kimia Aksir

What we will Learn..

- Purpose of using Off-the shelf tools
- Data Processing/Pre-processing
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data/Dimension Reduction
- Data Visualization
- Off the shelf Tool: “Excel” for Data pre-processing, Analysis and Visualization
- Excel: Pivot Table

Purpose of using Off-the Shelf Tools

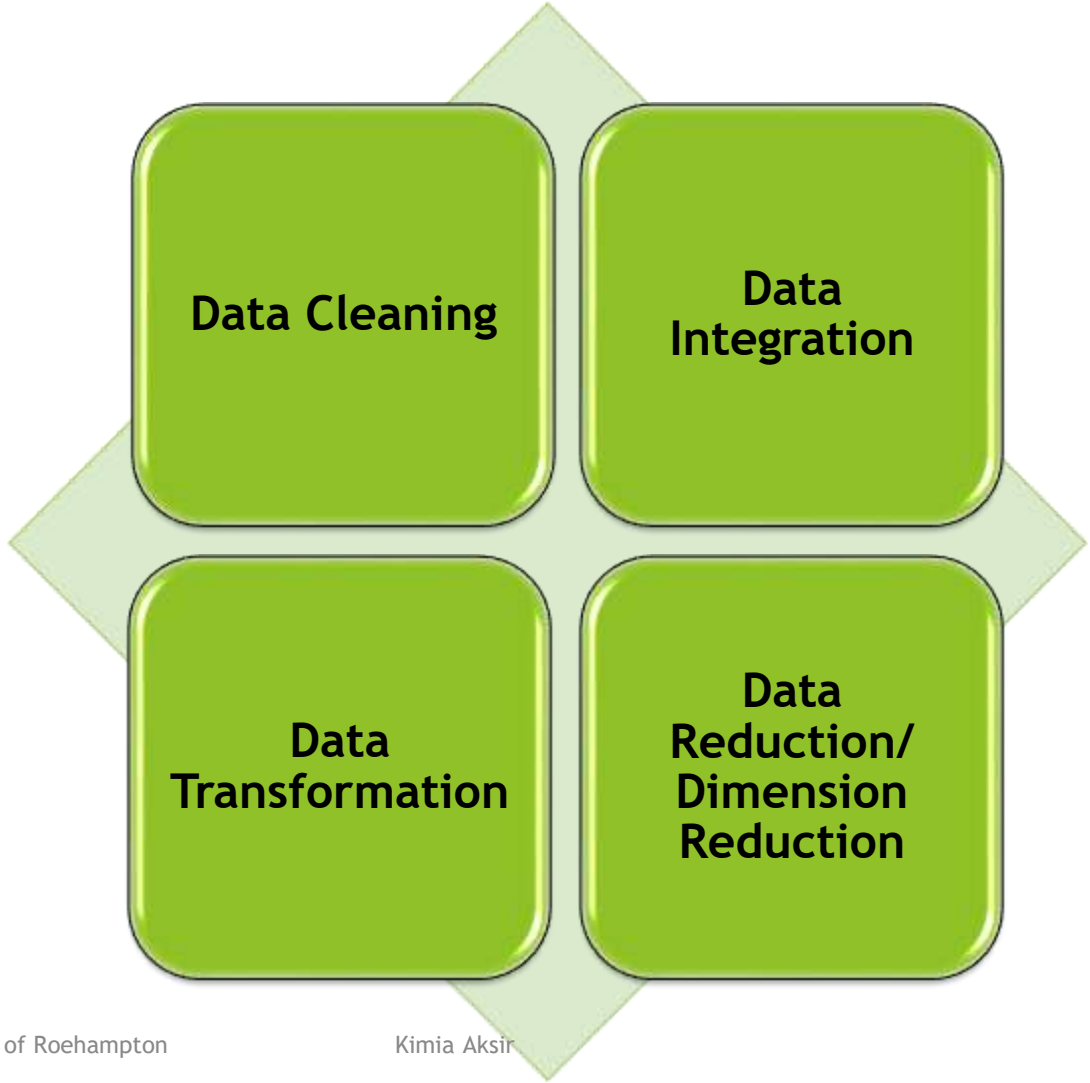
Off-the Tools can be useful to perform any small range of task, such as simple Statistical Analysis on Data, Data pre-processing, Creating useful Insights from Data and even represent the data using appropriate and meaningful Visualization.

These small range of tasks can be utilized as prerequisites of Data Analytical tasks as well depending on certain applications.

Data Preprocessing

Data preprocessing is one kind of data manipulation we do to enhance the performance of data by eliminating the unwanted or, garbage elements, formatting them wherever necessary and transforming the data in the most understandable form before they are used for any analytical purposes.

Data Preprocessing Continued…



Data Preprocessing: Data Cleaning

Missing Data

- Ignore and discard the tuple
- Add value (Manually)
- Add value (Computed Value)

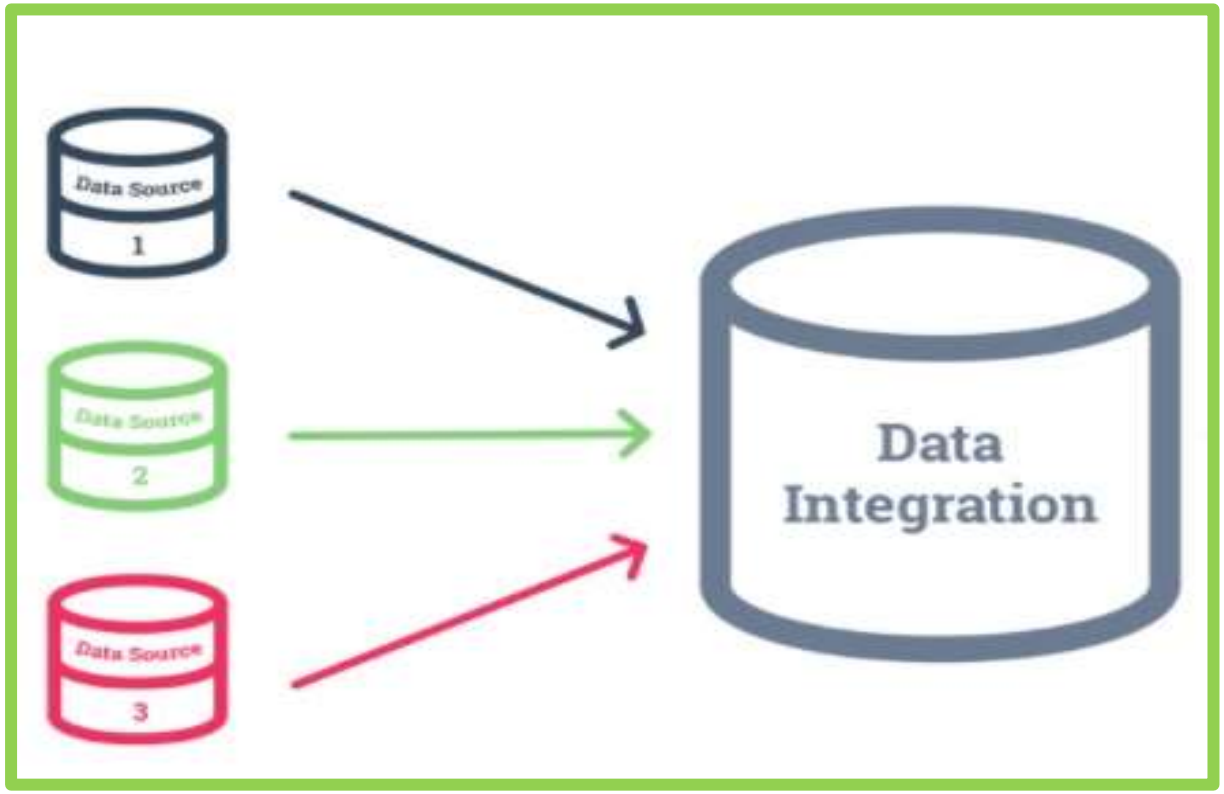
Noisy Data

- Binning
- Clustering
- Remove Noisy Data Manually

Inconsistent Data

- Using External Reference
- Using Tools

Data Preprocessing: Data Integration



"Data" to be analyzed can be collected from multiple sources. Combining these data and creating a unified view of it is Data Integration

Data Preprocessing: Transformation

Normalization

- Numeric attribute values are **scaled-up or, scaled-down** to fit within a specific range.
- Namely, **Min Max Normalization, Decimal Scaling Normalization**

Aggregation

- Aggregation becomes necessary mostly when data is **collected from multiple sources to generate a single "Aggregated" report.**

Attribute Selection

- This technique allows **creating new attribute(s) from an existing set of attributes**

Data Preprocessing: Data Reduction/ Dimension Reduction

Dimensionality Reduction

- Dimensionality reduction is used for feature selection
- This technique aims to reduce redundant feature

Data Compression

- The size of data can be reduced
- However, the compressing of data can be sometimes lossy or non-lossy

Discretization

- This is a process of converting continuous data into “range” and further these ranges corresponds to as a specific category

Data Visualization

- Several perspective of Data can be visualized easily from one presentation
- Exception in any Data/Dataset can be visualized
- Visual patterns can be analyzed.
- Data patterns can be translated into useful insights.
- Reduces time and difficulty to do “Decision Making” from Data/Dataset.

Off the shelf Tool: Excel

Source Worksheet

Long Name Filter	A YEAR (OFF)	B CustomerCountry	D ProductColor	E ProductCategory	F ProductSubcategory	I TotalCost
1	2004	Australia	Silver	Bikes	Mountain Bikes	769.49
2	2003	Australia	Multi	Clothing	Jerseys	49.99
3	2004	Australia	Blue	Bikes	Touring Bikes	2384.07
4	2004	Australia	Yellow	Clothing	Jerseys	53.99
5	2001	Australia	Silver	Bikes	Mountain Bikes	3399.99
6	2001	Australia	Silver	Bikes	Mountain Bikes	3399.99
7	2001	Australia	Black	Bikes	Mountain Bikes	3374.99
8	2001	Australia	Red	Bikes	Road Bikes	3578.27
9	2003	Australia		Accessories	Bottles and Cages	8.99
10	2004	Australia	Black	Bikes	Road Bikes	2443.35
11	2001	Australia	Red	Bikes	Road Bikes	3578.27
12	2003	Australia	Blue	Accessories	Helmets	34.99
13	2001	Australia	Red	Bikes	Road Bikes	3578.27
14	2004	Australia		Accessories	Bottles and Cages	9.99
15	2004	Australia		Accessories	Tires and Tubes	4.99
16	2003	Australia	Silver	Bikes	Mountain Bikes	2319.99
17	2004	Australia		Accessories	Tires and Tubes	35
18	2001	Australia	Red	Bikes	Road Bikes	3578.27
19	2003	Australia		Accessories	Tires and Tubes	2.29
20	2001	Australia	Black	Bikes	Road Bikes	699.0982
21	2001	Australia	Red	Bikes	Road Bikes	3578.27
22	2003	Australia	Silver	Accessories	Hydration Packs	54.99
23	2004	Australia		Accessories	Tires and Tubes	2.29

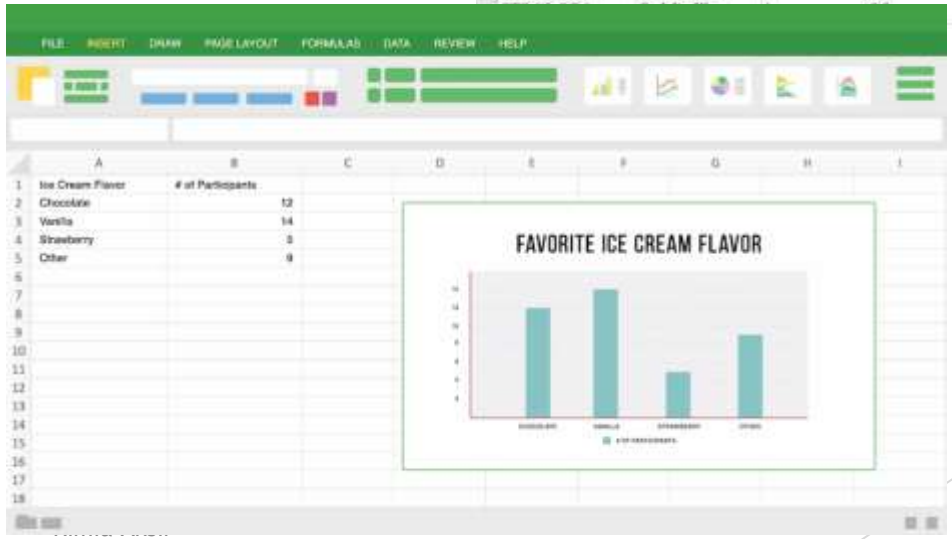
Pivot Table

Long Name	A(X) YEAR	B(Y)	C(Y) Sum of TotalCost	D(Y)
Units		Accessories	Bikes	Clothing
Comments				
F(x)=				
UserParam1				
1	2001	0	3.11421E6	0
2	2002	0	6.26258E6	0
3	2003	281089.1	6.96743E6	132728.69
4	2004	391752.38	8.79936E6	192688.76
5				
6				

Row Ascending

College Enrollment 2018-2019

Student ID	Last Name	Initial	Age	Program
ST348-245	Walton	L.	21	Drafting
ST348-246	Wilson	R.	19	Science
ST348-247	Thompson	G.	18	Business
ST348-248	James	L.	23	Nursing
ST348-249	Peterson	M.	37	Science
ST348-250	Graham	J.	20	Arts
ST348-251	Smith	F.	26	Business
ST348-252	Nash	S.	22	Arts
ST348-253	Russell	W.	19	Nursing
				Drafting



Off the shelf Tool: Excel (Data Pre-processing)

- Remove noisy Data, Data with unwanted symbols and/or, space(s)
- Selecting and treating blank cells (Missing value Handling)
- Converting one form of Data into other form (Convert Numeric Data stored as Text into numbers)
- Highlighting Errors (Formatting the cells contain Error)
- Eliminating Duplicates (Redundant Data can be removed)

Off the shelf Tool: Excel (for Data Analysis and Visualization)

- Basic and Advanced Descriptive Statistical Analysis can easily be performed in “Excel” by using readily available formula on Data.
- Visualization of Data using Chart, Combination Chart, Graph, Pivot table can be used for representing as well as representing analytical output from that Data

Excel: Pivot Table

A **pivot table** is a data processing tool used to calculate, analyze and summarize Data and facilitates to represent them in different ways such as, trends and patterns.

Pivot Table usage for Data Science tasks

- **Pivot Table is an interactive excel table:** It provides summary and insight of large amount of tabular data
- **Data Arrangement:** Appropriate Data representation possible
- **Calculation of Numeric Value:** Quick Calculation of Total, Average, Count can be done
- **Visualized Output Generation:** Can generate Chart, Graph and other insightful outcomes from data


Do you have any
QUESTIONS?