# **Introduction to Career Skills in Data Analytics:**

## Introduction to Data and Data Analysis:

* Data Engineers: Build and design the data.
* Data Architect: Design data systems
* Data Analyst: Model the data, more access, automates the flow of data
* Data scientist: Process skill of analyst, engineer, and architect
* Data Literate: Ability to read, speak, listen and understand the data
* Data Fluent: The ability to create, something beyond just being able to understand read and use it.

**Data Governance***:* A framework that incorporates strategies to create solid state of data, enable accountability and provide transparency to data in the organization.

* Access information
* Source of truth
* Master data management

**Quality of Data:** Data can be trusted to produce accurate insights.

**Hallmark of quality data:**

* Completeness
* Consistency
* Validity
* Accurate

## Introduction to Business Intelligence:

* Data and business intelligence (BI) give you the information and ability to make intelligent decisions.
* KPI- Key Performance Indicators
* Store the data which is important to the business.
* Businesses need to define the metrics that help track the overall data for the organization.
* **Data Analysis:** Analyzing and capturing the original data to compare over time.
* **Business Intelligence:** Understanding where we stand on any given day.
* **Business Analytics:** Seeing and predicting future outcomes.

## Identifying Data:

**Overcoming Analysis Approach**

* Build an approach
* Think through standard questions
* Use critical thinking
* Practice Active listening

**Sample problem:** *A company has 5 products. All of these products are being purchased, but the company is losing money. As a data analyst find solution.*

**Solution:** Data is everywhere. As an analyst, find why the sales are not growing.

1. Have these products ever been profitable?
2. If they are profitable in past, at what point of time?
3. What is different about now vs then?
4. Did the wholesale cost change?
5. Did the list price change?
6. Did the cost of storing or delivery change?

* Wholesale cost – no change
* List price – no change
* Cost to deliver – no significant change
* Cost to store – steadily increasing

**Now 2nd round of questions:**

1. What can we do to reduce cost of storage?
2. What type of price increase can be justified?

**Multiple types of data:**

* Personal data
* Work data
* Real time data
* Geographical data
* Social data

**Types of company data:**

* People management
* Marketing and Sales
* Purchasing
* Warehouse
* Shipment
* Accounting

**Types of Systems:**

* Spreadsheets
* Databases
* Data warehouses – Data warehouses have refined tables from production systems. Database include hundreds of data and details with only certain fields needed for reporting. Data warehouse store data and keep it safe.
* Data Lakes – Data lakes helps organizations capture data to store before its refined for reporting needs.

## Preparing data:

Clean and prepare the data to meet business requirements.

Describing data best practices – work with duplicate data.

**Data Profiling:** We can use this approach when we have in front of us to learn about it a higher level.

* Tells how much data we have
* Helps us validate our numbers
* Shows what we are facing when we are ready to transform our data.

**Business Rule:** Defines and controls the flow of data.

For the sales data business rules for Data Profiling:

* SalesOrderID must be converted into a text data type
* SalesOrderNumber must be converted to text data type and must not contain any number.
* All date fields should not include time.
* The main account GL Number field must be included.
* The main account GL Number holds the 4 digit code for accounting and the last two digits to specify the category
* TerritoryID and Comments fields must be removed.
* The final file must be saved as CSV in order for the import into the reporting system to be successful.

## Transforming Data:

### Transforming Data in Excel:

1. Learnt how to configure the path (Csv.Document(File.Contents("F:\\_DS+DA\Career-Essentials-in-Data-Analysis-by-Microsoft-and-LinkedIn\\_1\_Intoduction to Career Skills in Data Analytics\05\_Transforming Data\Suppliers.csv"),[Delimiter=",", Columns=9, Encoding=65001, QuoteStyle=QuoteStyle.None]))
2. Changes SupplierName type of text
3. Change SupplierName to UPPERCASE – Right Click – Transform – Uppercase
4. Create a new Column from TransactionDate , extract only year.
5. Added a new column from "TotalAmount", which is [AmountExcludingTax]+[TaxAmount] and changed type to currency
6. Removed unnecessary columns.
7. Added custom column "Days", which is [TransactionDate]-[FinalizationDate] and changed type and absolute value
8. Added conditional coloum.
9. Added Pivot table, learnt and implemented.

Advance Query :

let

Source = Csv.Document(File.Contents("F:\\_DS+DA\Career-Essentials-in-Data-Analysis-by-Microsoft-and-LinkedIn\\_1\_Intoduction to Career Skills in Data Analytics\05\_Transforming Data\Suppliers.csv"),[Delimiter=",", Columns=9, Encoding=65001, QuoteStyle=QuoteStyle.None]),

#"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),

#"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"SupplierName", type text}, {"SupplierTransactionID", Int64.Type}, {"SupplierID", Int64.Type}, {"PurchaseOrderID", Int64.Type}, {"SupplierInvoiceNumber", Int64.Type}, {"TransactionDate", type date}, {"AmountExcludingTax", type number}, {"TaxAmount", type number}, {"FinalizationDate", type date}}),

#"Uppercased Text" = Table.TransformColumns(#"Changed Type",{{"SupplierName", Text.Upper, type text}}),

#"Duplicated Column" = Table.DuplicateColumn(#"Uppercased Text", "TransactionDate", "TransactionDate - Copy"),

#"Extracted Year" = Table.TransformColumns(#"Duplicated Column",{{"TransactionDate - Copy", Date.Year, Int64.Type}}),

#"Renamed Columns" = Table.RenameColumns(#"Extracted Year",{{"TransactionDate - Copy", "TransactionYear"}}),

#"Reordered Columns" = Table.ReorderColumns(#"Renamed Columns",{"SupplierName", "SupplierTransactionID", "SupplierID", "PurchaseOrderID", "SupplierInvoiceNumber", "TransactionDate", "TransactionYear", "AmountExcludingTax", "TaxAmount", "FinalizationDate"}),

#"Added Custom" = Table.AddColumn(#"Reordered Columns", "TotalAmount", each [AmountExcludingTax]+[TaxAmount]),

#"Changed Type1" = Table.TransformColumnTypes(#"Added Custom",{{"TotalAmount", Currency.Type}}),

#"Removed Other Columns" = Table.SelectColumns(#"Changed Type1",{"SupplierName", "SupplierTransactionID", "SupplierID", "PurchaseOrderID", "SupplierInvoiceNumber", "TransactionDate", "TransactionYear", "FinalizationDate", "TotalAmount"}),

#"Added Custom1" = Table.AddColumn(#"Removed Other Columns", "Days", each [TransactionDate]-[FinalizationDate]),

#"Changed Type2" = Table.TransformColumnTypes(#"Added Custom1",{{"Days", Int64.Type}}),

#"Calculated Absolute Value" = Table.TransformColumns(#"Changed Type2",{{"Days", Number.Abs, Int64.Type}}),

#"Added Conditional Column" = Table.AddColumn(#"Calculated Absolute Value", "OverUnder", each if [Days] >= 3 then "3 Days or More" else "2 Days or Less")

in

#"Added Conditional Column"

### Transforming Data in SQL:

Structured Query Language (SQL): A computer language that works with data in a relational database management system.

Microsoft SQL Server: A relational database management system with the primary function of storing and retrieving data.

1. SELECT – fields from the Table
2. FROM – table name
3. WHERE – filter data
4. GROUP BY – group rows that have same values
5. HAVING – filters groups based on specified condition
6. ORDER BY – sort data

### Transforming data in Power BI:

Power BI:

* Transforms Data
* Presents data

1. Load data in Power query in Power Editor.
2. Removed other unnecessary columns.
3. Uppercased ProductName
4. Merge Queries (Select the foreign key and select what kind of join you want to perform)
5. Expand the new table, select only necessary info
6. Use group by.
7. Exclude other tables

let

Source = Excel.Workbook(File.Contents("F:\\_DS+DA\Career-Essentials-in-Data-Analysis-by-Microsoft-and-LinkedIn\\_1\_Intoduction to Career Skills in Data Analytics\05\_Transforming Data\Power BI\Data\Products.xlsx"), null, true),

Products\_Sheet = Source{[Item="Products",Kind="Sheet"]}[Data],

#"Promoted Headers" = Table.PromoteHeaders(Products\_Sheet, [PromoteAllScalars=true]),

#"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"ProductID", Int64.Type}, {"ProductName", type text}, {"SupplierID", Int64.Type}, {"CategoryID", Int64.Type}, {"QuantityPerUnit", type text}, {"UnitPrice", type number}, {"UnitsInStock", Int64.Type}, {"UnitsOnOrder", Int64.Type}, {"ReorderLevel", Int64.Type}, {"Discontinued", type logical}}),

#"Removed Other Columns" = Table.SelectColumns(#"Changed Type",{"ProductName", "ProductID"}),

#"Reordered Columns" = Table.ReorderColumns(#"Removed Other Columns",{"ProductID", "ProductName"}),

#"Uppercased Text" = Table.TransformColumns(#"Reordered Columns",{{"ProductName", Text.Upper, type text}}),

#"Merged Queries" = Table.NestedJoin(#"Uppercased Text", {"ProductID"}, Order\_Details, {"ProductID"}, "Order\_Details", JoinKind.Inner),

#"Expanded Order\_Details" = Table.ExpandTableColumn(#"Merged Queries", "Order\_Details", {"TotalOrderAmount"}, {"Order\_Details.TotalOrderAmount"}),

#"Renamed Columns" = Table.RenameColumns(#"Expanded Order\_Details",{{"Order\_Details.TotalOrderAmount", "TotalOrderAmount"}}),

#"Grouped Rows" = Table.Group(#"Renamed Columns", {"ProductName"}, {{"Total", each List.Sum([TotalOrderAmount]), type nullable number}})

in

#"Grouped Rows"

### Common Cleaning and Transformation:

1. Spaces are characters that need to be removed.

* Leading Spaces are at the front of the field
* Trailing spaces are end of the field
* Use functions like TRIM or CLEAN to removed spaces

1. Concatenate – Combine text with one another
2. Replace Text if required
3. Change the Case of the text.
4. Remove Duplicates
5. Transform data types

### Using Built-in Function:

Learnt about built in functions

## Modelling Data:

RDMS – SQL,

Access Database

Structured and Unstructured Data

One to One – One record tied with one record between two tables

One to Many and Many to One - One record from one table tied with many records from another table

Master Data Management Data

Structured Data – Data that fits neatly into tables and spreadsheets

Unstructured Data – Doesn’t fit, all pdfs, notes, etc. Requires our brain to review and provide context

Semi-structured Data – Mix of Structured and Unstructured Data