Meta : Introduction to Data Analyst.

Process of Collecting, cleaning, Organizing, analyzing and interpreting data to uncover insights and make informed decision.

1] Collection : - Gathering Data from various Sources.
- Data Can be in Structured or unstructured format.
- Data Should be accurate & relevent

2] Cleaning :- removing duplicates, inconsistencies or errors in dataset.
(Scrubbing)
- ensure data is accurate, Consistent and ready for analysis.

3] Organizing :- Sorting and Categorizing data into meaningful groups.
- Helps in understanding data quickly & identifying trends and patterns.

4] Analyzing :- Using Statistical & Mathematical methods to uncover insights.
- Uses algorithm, graphs, Prediction models. to discover pattern, trends and make forecasts based on historical data.

5] Interpreting:- ○ Presenting insights in a easy to understand way
○ Includes, charts, dashboard, Storytelling to Support decision making.

☆ Diffentiation between Data Analyst and Data Scientist

○ Data Analyst:-

- Understand and Visualize Structured data.
- Work:- Reporting, Dashboard, Data Cleaning
- Skills:- Basic Stats, Excel, SQL, python, R.
- Tools: Excel, BI Tools, SQL, R.
- Goal:- Find Insights for business decisions

○ Data Scientist.

- Build model using all types of data
- Work:- Predictive modeling, automation, ML algo.
- Skills: Advance ~~Excel~~, Stats, ML, python, TenserFlow, Spark.
- Tools: Python, TenserFlow, Hadoop, Spark, MySQL.
- Goal: Build Systems that predict and automate.

☆ OSEMN Framework.

Obtain : Gather data from relevant Sources
Scrub : Cleaning & preparing data.
Explore : Search for patterns.
Model : Generate Predictions & Insights.
iNterpret: Present & Communicate your insights.

o SMART Goals:
Specific, Measurable, Achievable, Relevant, Time-Bound.

- Knowing the business goal is Crucial for creating a plan and measuring progress.
- Writting down the goal helps clarify what the business aims to achieve.
- Identifying "key performance indicators" (KPIs) help measure whether the goal was achieved.
- KPI: Measurable value that can help you track your progress towards your goal.
  They are: Measurable, Directional, Directly related to your goal.

☆ Obtaining Data.

1] Freely accessible, open-source databases

2] Data Specific to your Company.
- Data Collected by the Company
- Data the Company Subscribes to.

3] Data you intentionally Collect.

o Common Data Formats.
i] Numeric Data (Quantitative) [Table Storing]
ii] Tent Data [(Unstructured) [NLP]
iii] Visual Data (imges, videos)

o Sampled Data:- Data from a Subset of a larger population or a larger dataset that is used to represent the entire dataset or population.

o Sampling Consideration:
- Sample Size
- Representativeness
- Generalizability.

- First Party Data

Data Collected by a business directly from it's Customers, Website Visitors or other internal Sources.

Includes: Surveys, Transaction data, and so on.

- Third Party Data.

Data Gathered by outside parties that are not affiliated with the business itself. Data about Company collected by another Comany. Have less Control on data Collection.

- Evaluating the Validity of Data Sources

1 Source Creadibility checking
↳ Authorship, publication date.

2 Methodology
↳ Sample Size, Sampling method,, Data Collection

3 Objectivity
↳ Bias, Coflicts of interest,

4 Accuracy
↳ Consistency, Errors

5 Relevance.
↳ Scope, Meaningful Content.

★ Scrubbing Your Data Clean.

i] Removing Duplicates

ii] Format Records.

iii] Handling Missing Values.

- -> Fill in missing value
- -> Delete record with missing value.

iv] Check for wrong values.

4 2-8-25

A Exploring Data

1 The language of Data
2 Visualizing data
3 Examine Variable distributions
4 Examine Variable relationships
5 Feature engineering.

o Summary Statistics:
Mathematical tools that Combine large amounts of data into a Single number that Says Something about all of the data as a whole.

o Key Characteristics of Creating Visualization:
- Clear & descriptive Title.
- x & Y axis relationships
- legends explanation.
- labels on axis
- colors, visuals, shaps of plotting points

o Common Charts
i] Bar charts :- Comparing Categorical data

ii] Line charts :- Showing trends overtime

iii] Scatter plots :- Show relationship between two variables (x & y)

- Data Distribution:
  What data in dataset looks like or how it is 'Spread' when all values are plotted on graph.

- "Binning" is technique to treat numeric data as Categorical data by divided number in buckets (Bins). Each Bin has some range.

- Common Types of Data distribution.

i) Normal Distribution: A Symmetric bell-Shaped Curve where most data points cluster around the mean.

ii) Bimodal Distribution: A distribution with two distinct peaks or models, like a Camel with two bumps

iii) Log-Normal Distribution: A Skewed distribution with the peak Shifted to One Side, usually ~~a~~ with long tail on the right.

iv) Exponential Distribution: A rapid decline from a high peak near zero, used to model time between random events

v) Uniform Distribution: A flat distribution where all values occur with equal frequency.

- Data RelationShips:
How different data points interact and influence each other. (Statistical relationShip).

- Type of RelationShips:(corelation)

i] Positive RelationShips :. one increase Other increases

ii] Negative RelationShips: one increases Othe decreases

iii] None Corelation:- No relation between two variables (No impact)

- Corerelation Coefficients
Numerical measures of Strength and direction of Corelation. (-1 to 1)

- Feature Engineering:-
A process where we Create new features or modify Enisting ones to better Understand our data.

- Encoding: Process of turning a String of data into numerical data by mapping each Unique String to Unique number.

★ Modeling Data.

Discover hidden patterns in data by using data from the past to predict the future.

o Models:

Mathematical tools used to recognize patterns in data and get insights on what might happen in the future.

o Phases of modeling.

- o Training: When the model learns a relationship
- o Testing: When the model's learnings are tested.

o Types of models:

i] Linear Regression -> Numerical nature, forcasting price of stock, trends, etc

ii] Classification -> Predict the output as categorical class which it belongs to, can be binary like True or False, or Multiclass problem.

iii] Clustering -> Split data into groups or segment with similar characteristics. often used to make customer base & divide them into smaller niche audience which can help for more focus advertising.

o Common Modeling Algorithms

1 Linear Regression:-
- Form of Regression model.
- Simple, less data-intensive.
•

2 Decision Tree:-
- Tree like Structure
- Segment data based on Series of binary decisions.
- Random Forest.
- Can be Classification or Regression model.

3 Neural Networks
- Use Complex networks of 'neurons' to make predictions.
- learning Complex relationships
- involves numerious variable & pattern.

★ Interpreting Data

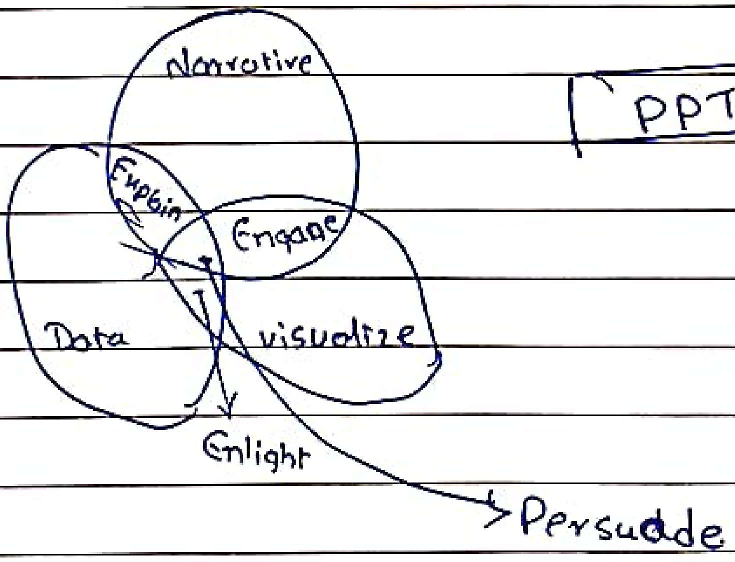
- The interpret Stage translate your analytical findings back to a business Content.

- Try & answer the business questions driving the entire project.

- Question to Ask:

1 What was objective for this analysis?
2 How does data answer my questions?
3 What other learning do i have?
4 How can i apply this to business Content?
5 How Confident Should i be?

- Explain, Enlighten, and Engage.



PPT

Model

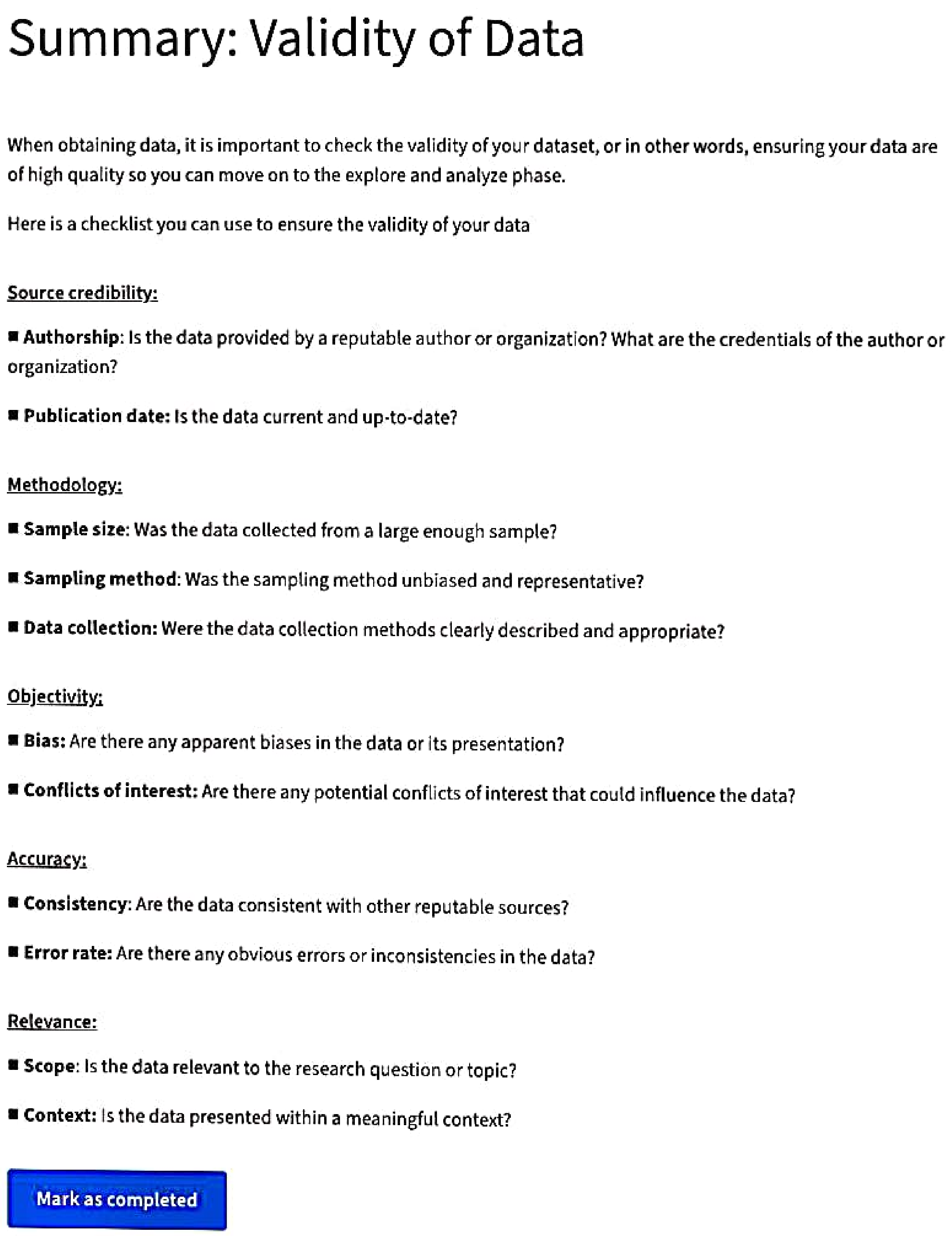
Generate predictions and insights

- Select a model type for your goals (often in cooperation with a partner)
- <u>Categories of models include:</u>
    - Classification - Is this “A” or “B”?
    - Regression - How much or how many?
    - Clustering - What natural segments can we find in our data?

iNterpret

Help others to understand the results of your analysis

- Build visualizations
- Construct stories
- Create presentations of your findings



Summary: Validity of Data

When obtaining data, it is important to check the validity of your dataset, or in other words, ensuring your data are of high quality so you can move on to the explore and analyze phase.

Here is a checklist you can use to ensure the validity of your data

Source credibility:

- **Authorship**: Is the data provided by a reputable author or organization? What are the credentials of the author or organization?

- **Publication date:** Is the data current and up-to-date?

Methodology:

- **Sample size**: Was the data collected from a large enough sample?

- **Sampling method**: Was the sampling method unbiased and representative?

- **Data collection:** Were the data collection methods clearly described and appropriate?

Objectivity:

- **Bias:** Are there any apparent biases in the data or its presentation?

- **Conflicts of interest:** Are there any potential conflicts of interest that could influence the data?

Accuracy:

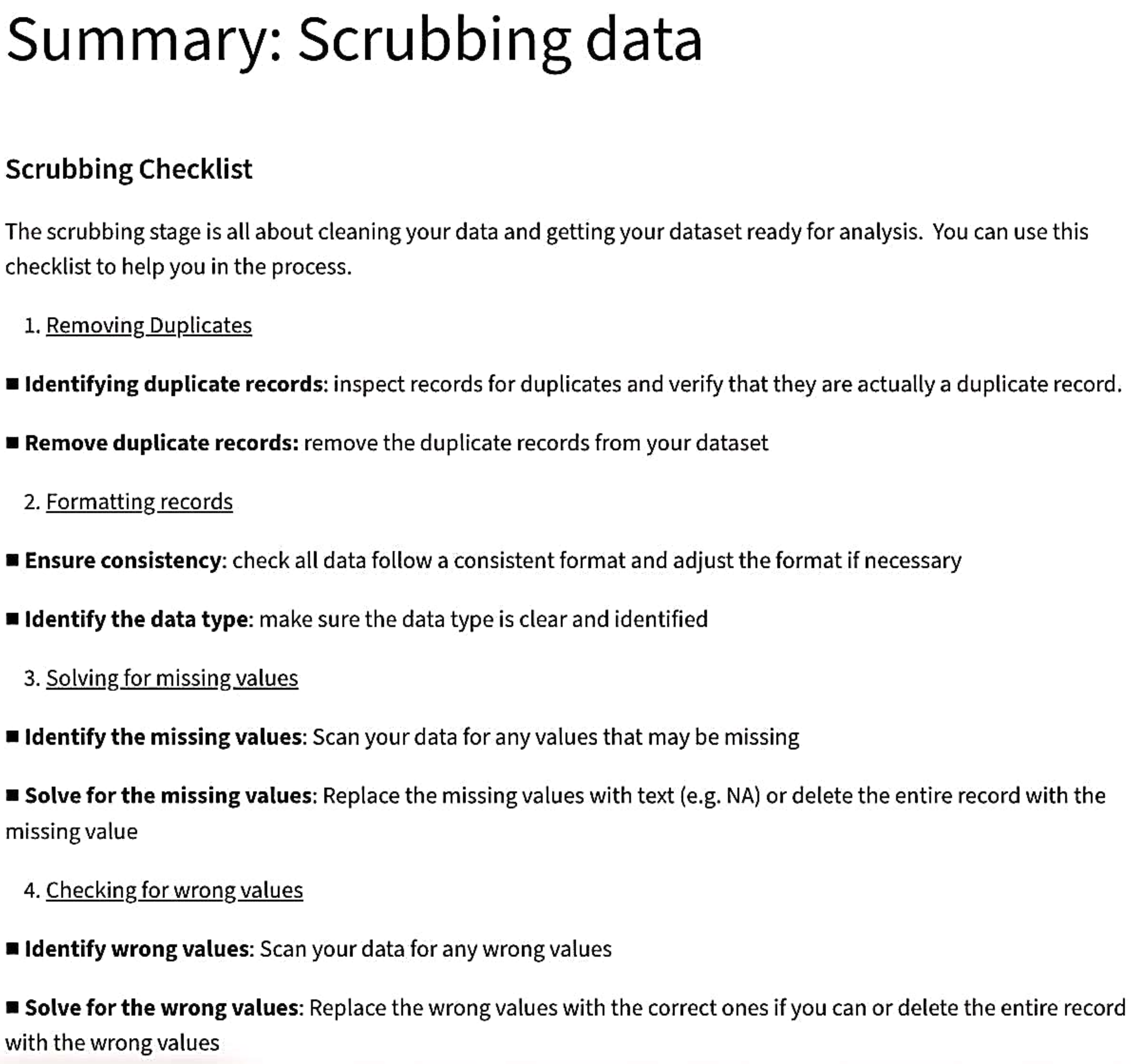
- **Consistency**: Are the data consistent with other reputable sources?

- **Error rate:** Are there any obvious errors or inconsistencies in the data?

Relevance:

- **Scope**: Is the data relevant to the research question or topic?

- **Context:** Is the data presented within a meaningful context?



Summary: Scrubbing data

Scrubbing Checklist

The scrubbing stage is all about cleaning your data and getting your dataset ready for analysis. You can use this checklist to help you in the process.

1. Removing Duplicates

■ **Identifying duplicate records**: inspect records for duplicates and verify that they are actually a duplicate record.

■ **Remove duplicate records:** remove the duplicate records from your dataset

2. Formatting records

■ **Ensure consistency**: check all data follow a consistent format and adjust the format if necessary

■ **Identify the data type**: make sure the data type is clear and identified

3. Solving for missing values

■ **Identify the missing values**: Scan your data for any values that may be missing

■ **Solve for the missing values**: Replace the missing values with text (e.g. NA) or delete the entire record with the missing value

4. Checking for wrong values

■ **Identify wrong values**: Scan your data for any wrong values

■ **Solve for the wrong values**: Replace the wrong values with the correct ones if you can or delete the entire record with the wrong values

Explore Checklist

What is your data telling you?

■ **Inspect your data**: If your dataset isn't too large, read through your data to assess whether interesting information jumps out

■ **Use summary statistics**: Evaluate your data by summarizing it (categorize, use statistics like average, deviation, etc.)

■ **Inspect a random sample of your data**: if your dataset is too large, a random sample may give you so information

Visualizing data

■ **Visualize your data** using bar charts, line charts or scatter plots to examine information hidden in your

Bar charts **Line charts** **Scatter plots**

Examine variable distributions

■ **Inspect the distribution of your data**

- Categorize the data
- Plot the categorized data

Common data distributions:

Normal **Bimodal** **Log-normal** **Exponential** **Uniform**

Learn more about your data:

■ **Evaluate the minimum**

■ **Evaluate the maximum**

■ **Evaluate the mode**

■ **Evaluate the standard deviation**

Examine variable relationships

■ **Visualize variables to understand their correlation**

Common visualizations:

Scatter plot **Line chart**

■ Calculate the correlation coefficient to understand the strength of th

Summary Reading: iNterpreting D
Storytelling

iNterpret Checklist

Step 1: Understand the results of your analysis

Ask the following questions:

■ **What was the objective for this analysis?**

■ **How does the data answer my questions?**

■ **What other learnings do I have?**

■ **How can I apply this to a business context?**

■ **How confident should I be?**

- How wrong is the model?
- How likely is the model to be correct?
- What scenarios cause the model to be incorrect?

Step 2: Explain your findings

Build a presentation with these key components: