DATA ASSESSING AND CLEANING

THiNK ABOUT THE DATA

Before cleaning data, you need to understand what it represents and the story it holds.

TYPES OF UNCLEAN DATA

Dirty Data:

Data with quality or content issues, such as:

- Duplicated Data
- Missing Data
- Corrupt Data
- Inaccurate Data

Messy Data:

Data with structural or organization issues, including violations of these principles:

- Each variable forms a column
- Each observation forms a row
- Each observational unit forms a table

EXAMPLE OF DiRTY DATA

ID	NAME	AGE	SALARY
101	John	29	50000
102	Jane	32	60000
103	John	29	50000
104	Sam	NaN	55000
105	Alice	28	error

Issues:

- **Duplicated Data:** Row 1 and Row 3 are duplicates
- **Missing Data:** Age is missing in Row 4
- **Corrupt Data:** Salary has an "error" value in Row 5

ID	NAME	AGE	SALARY
101	John	29	50000
102	Jane	32	60000
104	Sam	30	55000
105	Alice	28	52000

Fixes Applied:

- **Removed Duplicates:** Row 3 is deleted.
- **Filled Missing Values:** Estimated Sam’s Age as 30.
- **Corrected Errors:** Fixed Alice’s Salary to 52000.

EXAMPLE OF MESSY DATA

NAME	CONTACT	INCOME_23	INCOME_24
John	john@email.com	50000	52000
Jane	jane@email.com	60000	62000

Issues:

- **Multiple Variables in One Column:** Contact mixes email information instead of having a dedicated Email column**.**
- **Wide Format:** Income_2023 and Income_2024 should be one "Year" column and one "Income" column.

NAME	Email	YEAR	INCOME
John	john@email.com	2023	50000
John	john@email.com	2024	52000
Jane	jane@email.com	2023	60000
Jane	jane@email.com	2024	62000

STEPS TO REMEMBER:

1. Write a summary for your data

2. Write column descriptions.

3. Add any additional information if noticed.

THiNGS WE DO WHiLE ASSESSiNG THE DATA

We **discover** and then **document**

STEPS TO ASSESS THE DATA

1. Manual Inspection:

- **How:** Review data manually using tools like Google Sheets or Excel.
- **Why:** Helps identify obvious errors, patterns, or anomalies that automated checks might miss.
- **Tip:** Scanning the data visually can reveal hidden trends, inconsistencies, or outliers.

2. Programmatic Inspection:

- **How:** Use programming languages like Python or SQL to analyze data systematically.
- **Why:** Enables you to handle large datasets, automate checks, and perform in-depth analysis.
- **Common Techniques:**
  - **Python:** Use libraries like pandas, numpy, and matplotlib for data inspection.
  - **SQL:** Write queries to filter, group, and summarize data.

NOTE:

Assessing data is an ongoing process, you won't find all the patterns in one go. You need to review it repeatedly to uncover deeper insights.

WHAT NEXT?

Once you identify all the mistakes and patterns in the data, you can label them as either dirty data or messy data.

Examples:

- Spelling Mistake: **Dirty Data**
- Missing Data: **Dirty Data**
- Contact and email together in a column: **Messy Data**
- Column values are mixture of abbreviations and names (e.g.,state: BLR, Bihar, Delhi): **Dirty Data**.

DATA QUALITY DIMENSIONS

Only for dirty data

One more level of labeling is done for dirty data

Types

- **Completeness:** For missing data
- **Validity:** eg, duplicate patient id or height in negative
- **Accuracy:** Data is valid but not accurate. e.g., weight of an adult is 1 kg.
- **Consistency:** Both valid and accurate but written differently. e.g., New York and NY

ORDER OF SEVERiTY

Completeness>Validity>Accuracy>Consistency

DATA CLEANiNG ORDER

1.Dirty Data-> Completeness

2.Messy Data

3.Dirty Data-> Validity

4.Dirty Data->Accuracy

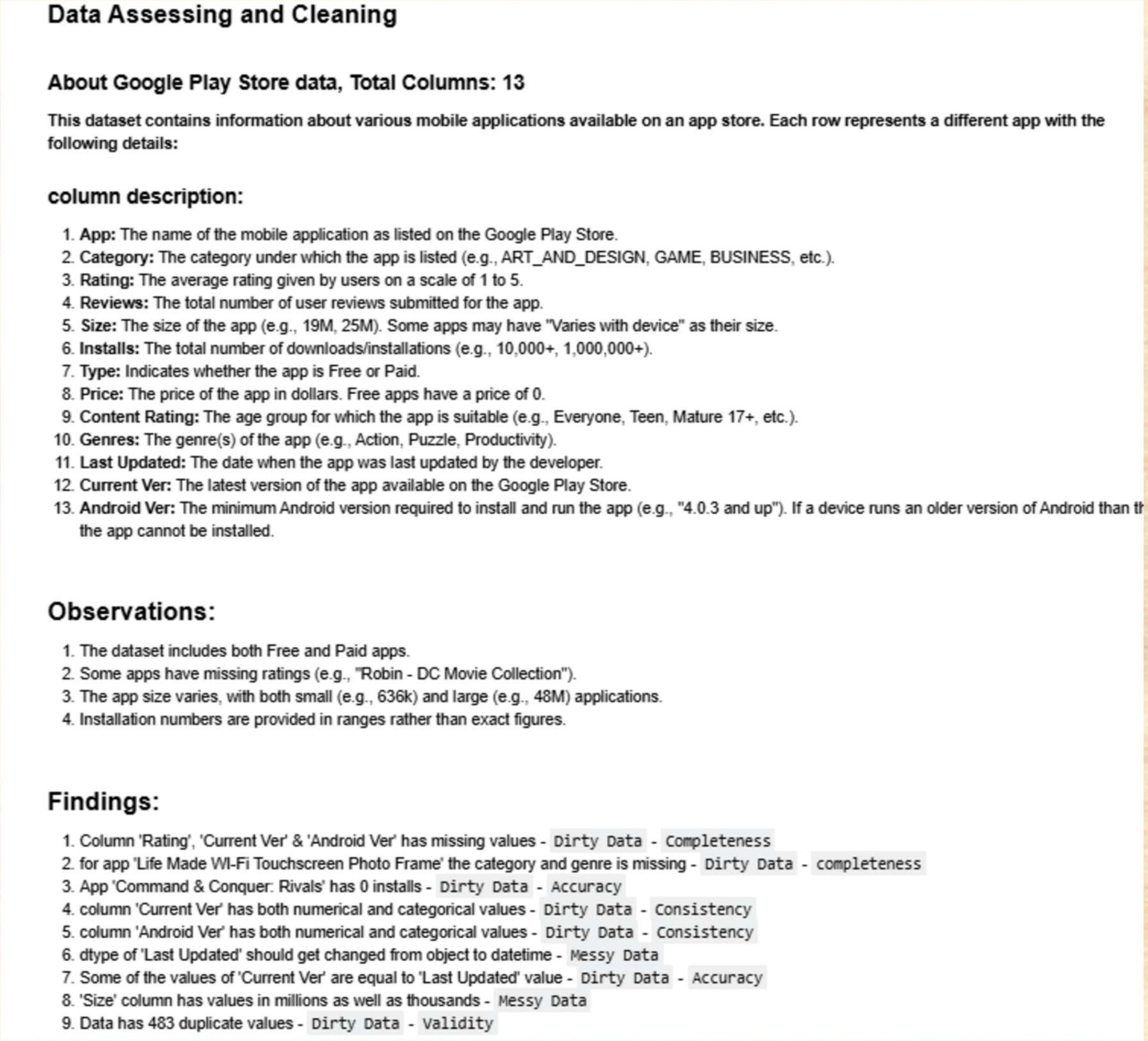
5.Dirty Data-> Consistency

STEPS INVOLVED IN CLEANING

1. Define the problem.
2. Plan and Write the Solution
3. Apply and Test the Solution
4. Review and Iterate

Always create a copy of your dataset before starting the cleaning process to preserve the original data.

Sample



Data Assessing and Cleaning

About Google Play Store data, Total Columns: 13

This dataset contains information about various mobile applications available on an app store. Each row represents a different app with the following details:

column description:

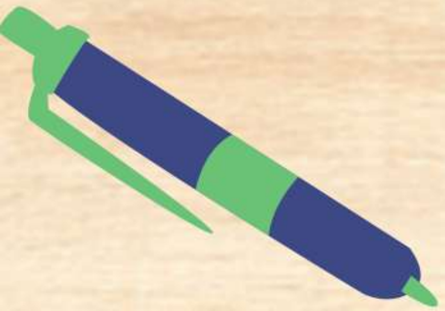
1. **App:** The name of the mobile application as listed on the Google Play Store.
2. **Category:** The category under which the app is listed (e.g., ART_AND_DESIGN, GAME, BUSINESS, etc.).
3. **Rating:** The average rating given by users on a scale of 1 to 5.
4. **Reviews:** The total number of user reviews submitted for the app.
5. **Size:** The size of the app (e.g., 19M, 25M). Some apps may have "Varies with device" as their size.
6. **Installs:** The total number of downloads/installations (e.g., 10,000+, 1,000,000+).
7. **Type:** Indicates whether the app is Free or Paid.
8. **Price:** The price of the app in dollars. Free apps have a price of 0.
9. **Content Rating:** The age group for which the app is suitable (e.g., Everyone, Teen, Mature 17+, etc.).
10. **Genres:** The genre(s) of the app (e.g., Action, Puzzle, Productivity).
11. **Last Updated:** The date when the app was last updated by the developer.
12. **Current Ver:** The latest version of the app available on the Google Play Store.
13. **Android Ver:** The minimum Android version required to install and run the app (e.g., "4.0.3 and up"). If a device runs an older version of Android than th the app cannot be installed.

Observations:

1. The dataset includes both Free and Paid apps.
2. Some apps have missing ratings (e.g., "Robin - DC Movie Collection").
3. The app size varies, with both small (e.g., 636k) and large (e.g., 48M) applications.
4. Installation numbers are provided in ranges rather than exact figures.

Findings:

1. Column 'Rating', 'Current Ver' & 'Android Ver' has missing values - `Dirty Data` - `Completeness`
2. for app 'Life Made WI-Fi Touchscreen Photo Frame' the category and genre is missing - `Dirty Data` - `completeness`
3. App 'Command & Conquer: Rivals' has 0 installs - `Dirty Data` - `Accuracy`
4. column 'Current Ver' has both numerical and categorical values - `Dirty Data` - `Consistency`
5. column 'Android Ver' has both numerical and categorical values - `Dirty Data` - `Consistency`
6. dtype of 'Last Updated' should get changed from object to datetime - `Messy Data`
7. Some of the values of 'Current Ver' are equal to 'Last Updated' value - `Dirty Data` - `Accuracy`
8. 'Size' column has values in millions as well as thousands - `Messy Data`
9. Data has 483 duplicate values - `Dirty Data` - `Validity`



Now that all the issues are identified, work on solving them one by one in the right order.