✦ Member-only story

Understanding Data Cleaning: The Unseen Hero of Data Science

Ritesh Gupta · Follow
Published in Python in Plain English · 5 min read · Oct 29, 2024

Uncovering the Hidden Power of Data Cleaning for Accurate Insights

Data cleaning is a foundational step in the data science workflow that often goes unnoticed but is crucial for the success of any data project. Think of it like tidying up a messy room before you start organizing or decorating — without clean data, any analysis, visualization, or model built on it will be unreliable. In fact, data scientists spend around 70–80% of their time on data cleaning! Let's dive into the why and how of data cleaning, with examples to illustrate its importance. 🚀

Why is Data Cleaning Important? 🤔

Imagine you're a chef preparing a meal. 🧑‍🍳🍳 Would you start with dirty, spoiled ingredients? Probably not. Similarly, data scientists need clean, high-quality data for analysis and model building. Unclean data can lead to:

- **Incorrect Analysis:** Bad data can skew results, leading to poor decisions.
- **Unreliable Models:** Machine learning models trained on flawed data are less accurate.
- **Misleading Insights:** With messy data, patterns and trends may appear differently than they truly are.

Common Types of Data Quality Issues 🔍

- **Missing Values** 🌫️: Some records may lack information for specific variables, causing incomplete data that can reduce model accuracy.
- **Duplicate Data** 📂: Duplicate entries are common and can bias analysis or inflate counts, leading to skewed results.

- **Outliers** 📈: Outliers are data points that are significantly different from the majority and can distort averages, correlations, or model performance.
- **Inconsistent Formatting** 📝: Data might be formatted inconsistently, such as dates listed in multiple formats (e.g., “DD/MM/YYYY” and “MM/DD/YYYY”) or text entries with different capitalizations.
- **Invalid Data** 🚫: Sometimes data entries contain incorrect values, like negative ages or future birthdates, which don’t make logical sense.

The Data Cleaning Process 🛠️

Data cleaning generally involves several steps. Let’s walk through each of these steps with some examples.

1. Handling Missing Values 🌫️

Missing values can occur for various reasons, such as data entry errors, system errors, or skipped survey questions. Here are a few techniques to handle them:

- **Dropping Missing Values:** If a variable has only a few missing values, we can remove those records. However, if many records have missing values, dropping them may lead to significant data loss.

```
# Example in Python: Dropping rows with missing values
df.dropna(inplace=True)
```

Imputing Missing Values: In cases where the missing data is essential, we might fill it with the mean, median, or mode (for numerical data) or the most frequent category (for categorical data).

```
# Example in Python: Imputing missing values with mean
df['age'].fillna(df['age'].mean(), inplace=True)
```

2. Removing Duplicates 📂

Duplicates often result from errors in data collection or merging datasets. Removing duplicates is necessary to prevent double-counting and misleading results.

```
2. Removing Duplicates 📂
Duplicates often result from errors in data collection or merging datasets. Remo
```

Consider a retail dataset where each transaction should be unique. If duplicates are present, revenue and sales figures might be inflated.

3. Handling Outliers 📈

Outliers can have various causes, from human error to exceptional cases that deserve special attention. Here are some ways to manage outliers:

- **Removing Outliers:** If outliers are likely due to data entry errors, they can be removed.
- **Transforming Data:** Applying transformations (like log transformations) can reduce the impact of outliers on analysis.
- **Using Robust Statistics:** Using metrics less sensitive to outliers, like the median, can help reduce their influence.

```
# Example in Python: Removing outliers based on a threshold
df = df[df['column'] < threshold_value]
```

For example, if you have age data in a customer dataset and spot an age entry of 200, it's likely an error that should be removed.

4. Standardizing Formats 📝

Inconsistent formatting can cause significant issues during analysis. Standardizing formats includes ensuring consistency in date formats, units of measurement, and text data capitalization.

For instance, if a dataset has a "Date" column with both "YYYY-MM-DD" and "MM/DD/YYYY" formats, you can standardize them into a single format for uniformity.

```
# Example in Python: Converting dates to a single format
df['date'] = pd.to_datetime(df['date'], format='%Y-%m-%d')
```

Similarly, if you're analyzing data on customer regions and have both "NY" and "New York" entries, it's essential to make them consistent.

5. Addressing Invalid Data 🚫

Invalid data refers to entries that don't make logical sense, like negative prices or impossible dates. Identifying and correcting these values is crucial.

For example, if a dataset has customer ages and an entry is "-5", you'd need to correct or remove this data.

```
# Example in Python: Removing invalid data based on conditions
df = df[df['age'] > 0]
```

Real-World Example: E-commerce Sales Dataset 🛒

Let's say you're analyzing an e-commerce sales dataset with information on customers, purchases, and dates. Here's how data cleaning might look in this case:

1. **Missing Values:** Fill missing "customer age" values with the median age of customers.
2. **Duplicates:** Remove duplicate transactions to avoid inflating sales figures.
3. **Outliers:** Check for outliers in transaction amounts and decide if they're legitimate high-value purchases or errors.
4. **Standardizing Formats:** Ensure all dates are in the same format, such as "YYYY-MM-DD".
5. **Invalid Data:** Remove entries where "quantity purchased" is negative or where the "purchase date" is in the future.

After these steps, your dataset is clean, consistent, and ready for further analysis. 📊✨

Data Cleaning Tools 🧰

Many tools help automate data cleaning processes. Here are a few popular ones:

- **Python (Pandas, NumPy)**: Widely used for data cleaning, with versatile functions for handling missing values, outliers, and more.
- **R**: Offers similar functionality for data manipulation and cleaning.
- **OpenRefine**: A powerful, open-source tool specifically designed for data cleaning and transformation tasks.
- **Microsoft Excel**: Provides basic data cleaning capabilities, like removing duplicates, filtering, and simple transformations.

Final Thoughts 🌟

Data cleaning may not be the most glamorous part of data science, but it's a critical step that can make or break the success of your analysis. By investing time in cleaning data, you ensure that the insights drawn from it are trustworthy and that any machine learning models built on it are as accurate as possible. 🏆

So, next time you're diving into a new dataset, remember to appreciate the humble art of data cleaning. It's the hidden hero behind every successful data project! 🦸‍♂️✨

In Plain English 🚀

Thank you for being a part of the ***In Plain English*** *community! Before you go:*

- Be sure to **clap** and **follow** the writer 👏
- Follow us: **X** | **LinkedIn** | **YouTube** | **Discord** | **Newsletter** | **Podcast**
- **Create a free AI-powered blog on Differ.**
- More content at **PlainEnglish.io**

Data Science | Machine Learning | Deep Learning | NLP | Python

Published in Python in Plain English

31K Followers · Last published 11 hours ago

Follow

New Python content every day. Follow to join our 3.5M+ monthly readers.

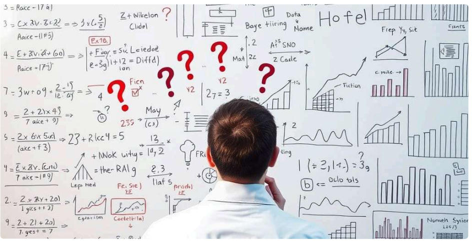
Written by Ritesh Gupta

3.6K Followers · 28 Following

Follow

Data Scientist, I write Article on Machine Learning| Deep Learning| NLP | Open CV | AI Lover ❤️

More from Ritesh Gupta and Python in Plain English

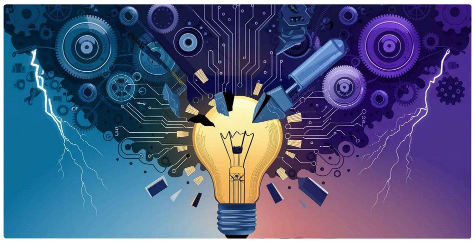


Ritesh Gupta

Can You Handle These 25 Toughest Data Science Interview Questions?

The role of a Data Scientist demands a unique blend of skills, including statistics, machine...

Sep 25 7

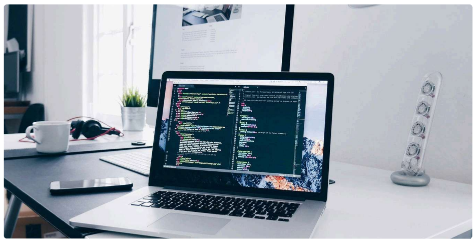


PY In Python in Plain English by Abdur Rahman

5 Overrated Python Libraries (And What You Should Use Instead)

Traditional Devs, Look Away — This One's Not for You!

Nov 3 20



PY In Python in Plain English by Abdur Rahman

8 Uncommon but Extremely Useful Python Libraries for 2025

You'll regret not knowing this earlier

Nov 5 7

PY In Python in Plain English by Ritesh Gupta

7 GitHub Repos to Transform You into a Pro ML/AI Engineer

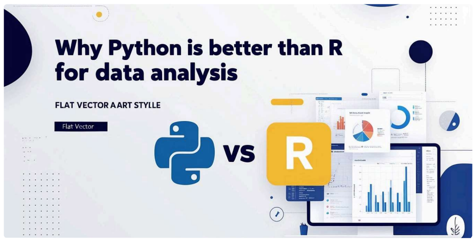
Hands-On Guides, Tools, and Frameworks to Fast-Track Your AI Journey

Nov 5 1

See all from Ritesh Gupta

See all from Python in Plain English

Recommended from Medium

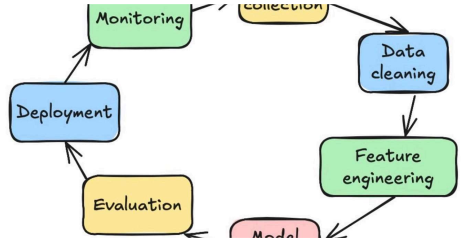


PY In Python in Plain English by Mayur Koshti

Why Python is Better than R for Data Analysis

Python and R for data analysis would involve breaking down each relevant aspect, feature...

Nov 15 7



tds In Towards Data Science by Haden Pelletier

Every Step of the Machine Learning Life Cycle Simply Explained

A comprehensive guide to the ML life cycle with examples in Python

3d ago 3

Lists

Predictive Modeling w/ Python

20 stories · 1691 saves

Practical Guides to Machine Learning

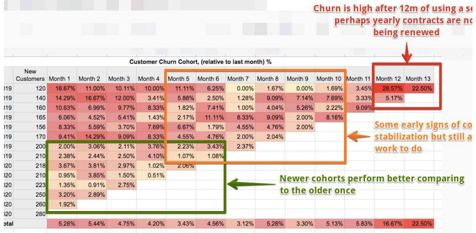
10 stories · 2057 saves

Natural Language Processing

1839 stories · 1461 saves

Coding & Development

11 stories · 920 saves

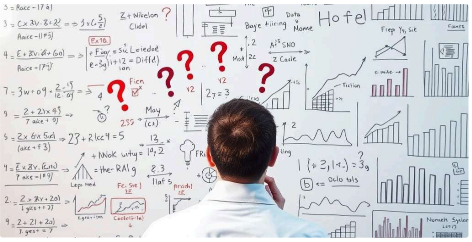

Churn is high after 12m of using a se
perhaps yearly contracts are no
being renewed
Customer Churn Cohort, (relative to last month) %
Some early signs of co
stabilization but still a
work to do
Newer cohorts perform better comparing
to the older once

In Stackademic by Muhammad Muhsi Sidik

Data Visualization for Customer Retention: Unlocking Insights wit...

Customer retention is the lifeblood of any business striving for long-term success....

Nov 20

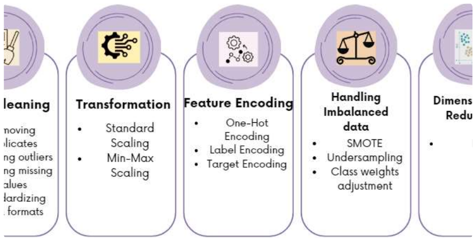


Ritesh Gupta

Can You Handle These 25 Toughest Data Science Interview Questions?

The role of a Data Scientist demands a unique blend of skills, including statistics, machine...

Sep 25 7



In Code Like A Girl by Niveatha Manickavasagam

Top 5 Data Preprocessing Techniques: Beginner to...

A Comprehensive Guide to Clean, Transform, and Optimize Data Effectively

Nov 18 1

In DataDrivenInvestor by Bex T.

Firecrawl: How to Scrape Entire Websites With a Single Command...

And turn them into LLM-ready data

4d ago 3

See more recommendations