Data cleaning

Cleaning techniques

Data cleaning

Data overview

> We will use a dataset containing the prices of maize in three major cities in Kenya from January to March 2022.

01. The last column in the dataset, namely **exchrate**, contains values calculated by dividing the **price** column values by the **usdprice** column values.

02. At the bottom of the dataset, we calculated the sum and count of the **price**, **usdprice,** and **exchrate** columns, and the results were used to calculate the average of each, respectively.

Dataset

	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
2	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES			#DIV/0!
4	2022-01-15	Kisumu	KG	KES	36.54	0.32	113.3448873
5	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
6	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
7	2/15/2022	Nairobi	KG	KES			#DIV/0!
8	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
9	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
10	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
11	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
12	2022-03-15	KSM	5KG	KES	172.85	1.53	113.2456703
13	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
14	**Sum**				461.96	4.08	#DIV/0!
15	**Count**				9	9	9
16	**Average**				51.33	0.45	#DIV/0!
17							

01.

02.

What is data cleaning?

Data cleaning is the process of **identifying** and **correcting erroneous data** within a dataset. This improves the quality of our data in preparation for analysis.

What qualifies as erroneous data?

Data collected from different sources and in different ways are bound to have mistakes, such as being :

- Incorrect
- Incomplete
- Duplicated
- Irrelevant
- Wrongly formatted

Common data cleaning tasks

Data cleaning is not just about deleting bad data but trying to achieve **optimal accuracy** by correcting what we can.

- Removing duplicate observations
- Removing irrelevant observations
- Handling missing data
- Filtering unwanted outliers
- Fixing structural issues

Data practitioners spend **60% to 80%** of their time getting their data ready through cleaning.

Why data cleaning is important

> Data-driven decisions have become increasingly popular within organizations. As a result, there is a great need for clean and accurate data that can provide **reliable insights**.

Our **analysis results** are directly affected by the **quality of our data**, regardless of whether we follow the conventional analysis processes.

An analysis conducted with erroneous data may lead to **wrong results**. Basing business decisions and strategies on these **results** may do more harm than good.

Benefits of data cleaning

- Prevents time-consuming and costly fixes due to decisions brought about by bad data.
- Makes the analysis more efficient since the data is well organized.
- Tidy data enable more organized and effective storage.
- Helps to avoid mistakes during daily business operations involving the data.
- Reduces clutter by getting rid of unwanted data.

Missing data

Missing data refers to data points that are **incomplete** or do not contain any value.

Why it's bad for our data

- Can affect how the data will be interpreted and lead to biased results as they are not based on all values.
- Can lead to biased results and reduce the statistical power of analysis.

Possible causes

- Data entry errors
- Incomplete survey responses
- Error during data collection

Missing data can be represented in different ways, depending on how the data were collected and the software being used. It is often represented by a **blank**, **null**, or **NaN**.

Data cleaning

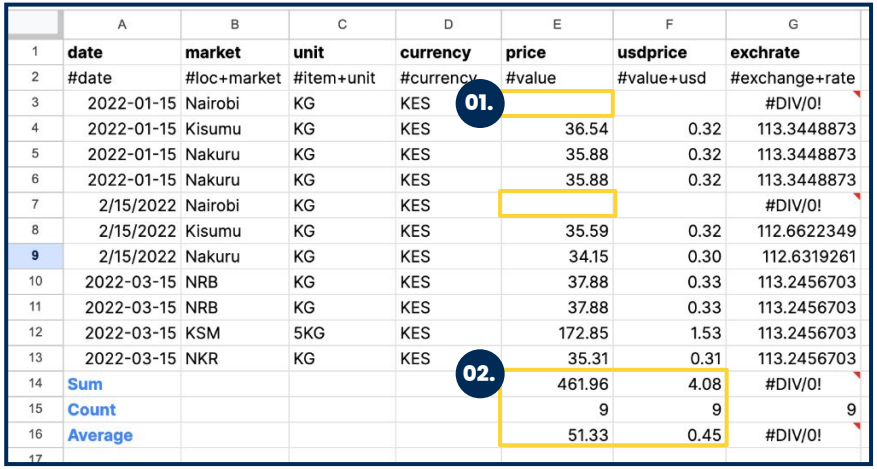
NaNs vs nulls

Both NaNs and nulls are used to **represent missing** data, but they represent slightly different values.

Nulls:

Nulls represent the **absence of any value** completely.

In Google Sheets, null values are indicated by **empty** or **blank cells**.



	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
2	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES	01.		#DIV/0!
4	2022-01-15	Kisumu	KG	KES	36.54	0.32	113.3448873
5	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
6	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
7	2/15/2022	Nairobi	KG	KES			#DIV/0!
8	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
9	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
10	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
11	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
12	2022-03-15	KSM	5KG	KES	172.85	1.53	113.2456703
13	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
14	Sum				02. 461.96	4.08	#DIV/0!
15	Count				9	9	9
16	Average				51.33	0.45	#DIV/0!
17							

01. **Null values are present** in the dataset.

02. Our calculations below are therefore **not inclusive of all the observations** in our dataset since the functions used are built to **ignore blank cells**.

Data cleaning

NaNs vs nulls

NaNs:

NaNs (Not a Number) are used to represent unrepresentable calculation results.

In Google Sheets, a NaN value appears as an **error** returned by a formula when the result is not a valid numerical value.

In a cell, this can be displayed as "**#VALUE!**" in the case where a formula attempts to perform a mathematical operation on non-numeric values or "**#DIV/0!**" where it tries to divide a zero by a zero.

	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
2	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES			#DIV/0!
4	2022-01-15	Kisumu	KG	KES	36.54		113.3448873
5	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
6	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
7	2/15/2022	Nairobi	KG	KES			#DIV/0!
8	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
9	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
10	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
11	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
12	2022-03-15	KSM	5KG	KES	172.85	1.53	113.2456703
13	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
14	Sum				461.96		#DIV/0!
15	Count				9	9	9
16	Average				51.33	0.45	#DIV/0!
17							

01.

02.

01. NaN values brought about by **dividing zero by zero**.

02. The presence of these **NaNs** in the column has affected some of our calculations that are also **returning NaN values**.

Handling missing data

During data cleaning, we search for and deal with missing values using various strategies. The appropriate strategy depends on the type and amount of missing data.

01. Imputation

The process of filling in the blanks using an **estimated value**. This can be achieved in several ways:

- Using the median or mean.
- Copying data from a similar dataset.
- Using domain knowledge.
- Using linear regression.

Limitation: Their accuracy can't be guaranteed, and they could further confuse the results by reinforcing false patterns within the data.

02. Flag as missing

Missing data may be informative in itself. We can inform our analysis of missing values by filling them in with a uniform value such as 'null'.

03. Drop observations with missing values

We can also choose to delete observations with missing values, especially if they contain too many blanks for it to be useful.

Limitation: We may end up losing valuable information, especially if we delete too much data.

Duplicate observations

Duplicate observations are entries that have been **repeated** in the dataset.

Why it's bad for our data

- Can lead to biased results.
- Can lead to an excessively large dataset, which is difficult to deal with and wastes time and storage space.
- Can skew the data, causing inaccurate and confused results.
- Can make visualization difficult to read.

Possible causes

- Data combined from multiple sources
- Error during data entry
- System glitches

Duplicate observations

	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
2	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES	38.90	0.34	113.3448873
4	2022-01-15	Kisumu	KG	KES	36.54	0.32	113.3448873
01.	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
7	2/15/2022	Nairobi	KG	KES	40.04	0.36	112.6617895
8	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
9	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
10	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
11	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
12	2022-03-15	KSM	5KG	KES	172.85	1.53	113.2456703
13	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
14	**Sum**				540.90	4.78	1244.3182
15	**Count**			02.	11	11	11
16	**Average**				49.17	0.43	113.1198
17							

Duplicate observations often occur when **two or more observations have identical values** for all or most of the variables in the dataset.

In this example, we see that the identified duplicate observations are **exact duplicates,** and we can simply **remove the second occurrence** to clean the dataset.

01. These particular observations have been duplicated.

02. Therefore, our calculations below are not accurate, as they include the duplicates. For instance, our 'Count' values are in excess by 2 because of counting the 2 duplicate entries.

Removing duplicate observations

Deduplication is the process of identifying and getting rid of the duplicate records using various strategies, including deleting and merging.

Steps to deduplication

01. **Search** for the duplicate values in Google Sheets using:
    a. The filter option in the toolbar.
    b. Data > Column stats in the toolbar.
    c. A conditional statement in a formula.
    d. Mixed references and comparisons in a formula.
02. **Handle** the duplicate observations by:
    a. Removing the duplicate, retaining only the first or last occurrence.
    b. Merging the duplicate observations.
03. Both steps 01. and 02. can be done **manually**, where we inspect duplicates one by one, or **automatically**, using built-in data cleaning functions in software.

Be **careful when removing** seemingly duplicate observations. Ensure that they do not represent distinct cases.

Unwanted outliers

Outliers are **unusual data points** that differ significantly from the rest of the values in the dataset.

Why it's bad for our data

- Can skew the analysis results (towards outliers).
- Can disturb distribution of data.
- Can affect the readability of results.

Possible causes

- Sampling errors
- Measurement errors
- Data entry errors
- Natural variation in the data

Although **outliers** may provide valuable insights into data, **unwanted outliers** can significantly impact the accuracy and validity of our analysis results.

Data cleaning

Unwanted outliers

	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
2	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES	38.90	0.34	113.3448873
4	2022-01-15	Kisumu	KG	KES	36.54	0.32	113.3448873
5	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
6	2/15/2022	Nairobi	KG	KES	40.04	0.36	112.6617895
7	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
8	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
9	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
10	2022-03-15	KSM	5KG	KES	172.85	1.53	113.2456703
11	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
12	Sum				467.14	4.13	1017.7276
13	Count				9	9	9
14	Average				51.90	0.46	113.0808

01.

02.

Can you identify a probable cause for the outlier in the example? Tip: Consider the other features in the dataset.

01. This value seems to be **significantly higher than the rest**, which means it is an **outlier**.

02. This means that our **sum and average results have been skewed** towards this outlier. We need to **investigate the reason** for this and find an appropriate way to handle the outlier.

Filtering unwanted outliers

During data cleaning, we **remove unwanted outliers** that may affect the performance of the data.

Not all outliers are errors. Outliers can provide valuable insights into the data. It's important to first **determine the validity** of removing an outlier.

01. Identifying the outliers

- **Plot the data** for visual inspection, using a scatter plot, box plot, or histogram.
- Use of **statistical measures** that are based on the distribution of data, such as the interquartile range.
- Use **domain knowledge** to determine the validity.

02. Removing the outliers

- **Delete** the entire observation containing the outlier. This is if the observation is clearly an error.
- **Replace** the outlier value with a more accurate or representative value, such as the mean.

Irrelevant observations

Irrelevant observations are observations that are **not useful** to the context or problem to be solved and therefore do not contribute to our analysis.

Why it's bad for our data

- Can make our dataset large which makes it less manageable and efficient.
- Can reduce the accuracy and effectiveness of our analysis.
- Can introduce bias.

Possible causes

- Unnecessary data picked during data collection/web scraping
- Data entry errors
- Merging of multiple datasets

Whether an observation or feature is **irrelevant** and whether or not it should be **removed** depends on the size and complexity of the dataset and the problem to be solved.

Irrelevant observations

	A	B	C	D	E	F	G
1	**date**	**market**	**unit**	**currency**	**price**	**usdprice**	**exchrate**
01.	#date	#loc+market	#item+unit	#currency	#value	#value+usd	#exchange+rate
3	2022-01-15	Nairobi	KG	KES	38.90	0.34	113.3448873
4	2022-01-15	Kisumu	KG	KES	36.54	0.32	113.3448873
5	2022-01-15	Nakuru	KG	KES	35.88	0.32	113.3448873
6	2/15/2022	Nairobi	KG	KES	40.04	0.36	112.6617895
7	2/15/2022	Kisumu	KG	KES	35.59	0.32	112.6622349
8	2/15/2022	Nakuru	KG	KES	34.15	0.30	112.6319261
9	2022-03-15	NRB	KG	KES	37.88	0.33	113.2456703
10	2022-03-15	KSM	KG	KES	34.57	0.31	113.2456703
11	2022-03-15	NKR	KG	KES	35.31	0.31	113.2456703
12	Sum			02.	328.86	2.91	1017.7276
13	Count				9	9	9
14	Average				36.54	0.32	113.0808

In this example, the irrelevant observations and features are easy to identify and simple to remove because no other features are dependent on them.

However, in many cases it will require **greater insight** to **identify** these irrelevant observations and to determine possible **interdependency**.

01. The second row in our data seems to contain some **additional metadata** that does not add any value to our data.

02. The currency column is also unnecessary. Since our dataset contains maize prices in Kenya, it is unlikely that we will have differing currencies.

Removing irrelevant observations

> We remove irrelevant observations such that we are only left with data that are **necessary** to our analysis.

The appropriate strategy for handling irrelevant observations and features depends on the **goals of our analysis** and the **structure of the data**.

In some cases, removing irrelevant observations may be appropriate, while in other cases, they **may be retained but not used** in the analysis.

It is important to note that additional observations and features often **contribute to the *story*** we are trying to tell with the data, even if they **do not influence** the analysis.

We need to be **careful** of the criteria used to identify irrelevant observations such that there are **no biases introduced**.

Structural issues

Structural errors are issues **within the data** such as inconsistent naming conventions, inconsistent formatting, typos, incorrect capitalization, and inconsistent data types.

Why it's bad for our data

- Can cause the data to be processed incorrectly and thus give incorrect or biased results.
- Can lead to mislabeled categories, which may cause us to miss out on or misinterpret key findings.

Possible causes

- Data entry errors
- Incomplete survey responses
- Error during data collection
- Incorrect importing of data
- Formula errors

The **severity of structural errors** depends on the data **structure**, how the data are being **used** in analysis, and the specific **type** of structural error.

Structural issues

	A	B	C	D	E	F
1	**date**	**market**	**unit**	**kesprice**	**usdprice**	**exchrate**
2	2022-01-15	Nairobi	KG	38.90	0.34	113.3448873
3	2022-01-15	Kisumu	KG	36.54	0.32	113.3448873
4	2022-01-15	Nakuru	KG	35.88	0.32	113.3448873
5	2/15/2022	Nairobi	KG	40.04	0.36	112.6617895
6	2/15/2022	Kisumu	KG	35.59	0.32	112.6622349
7	2/15/2022	Nakuru	KG	34.15	0.30	112.6319261
8	2022-03-15	NRB	KG	37.88	0.33	113.2456703
9	2022-03-15	KSM	KG	34.57	0.31	113.2456703
10	2022-03-15	NKR	KG	35.31	0.31	113.2456703
11	Sum			328.86	2.91	1017.7276
12	Count			9	9	9
13	Average			36.54	0.32	113.0808
14						

01.

02.

The structural issues in this example did not influence our sum, count, and average analysis. However, this is not always the case.

Incorrect file imports into Google Sheets often result in **structural issues** due to varying data types, inconsistent delimiters, etc.

01. In the date column, we have **different date formats**. This is not only untidy but may also cause confusion in our analysis.

02. For the market column, **inconsistent naming conventions** have been used. We have some rows where the full city name has been used, while others have the abbreviated form of the name. These will be treated as different categories even though they represent the same thing.

Fixing structural issues

Depending on the type of structural issue, we can apply various strategies to correct it, including correcting data entries and imports, converting to consistent formats, and imputation.

01. Standardise the data and structure

This involves keeping the data consistent throughout the dataset, for example, lowercase, uppercase, naming conventions, measurements, date formats, padding for string size, etc.

02. Convert to the correct data type

This ensures that each column has the appropriate data type and every value will be processed appropriately.

03. Spell check to correct typos

We can do this by either inspecting and then correcting them manually or automatically using programming or spelling and grammar tools.

These are only some of the ways to fix structural issues. Often, some of our other data cleaning strategies will also help with fixing structural issues.