OFFUTURE

THE FUTURE OF OFFICE SUPPLIES

Quality Control Form

Team Name	JOIN
Team Members	Egle, Aksha, James, Rachel
Client Name	Offuture
Date of file received	2024-04-24
Name of file received	1 - Offuture 2011 - 2014
File Format received	.csv
Size of file (KB) received	12.1 MB
Encoding of the file	UTF-8
Recorded Number of Columns	24
Recorded Number of Rows	51290
Name of Schema of Destination Table	student
Name of Destination Table	offuture

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

Table of contents

[SQL01] Count of Rows 3
[SQL02] Count of Distinct Rows 3
[SQL03] Count of Columns 3
[SQL04] Sum of Column Sums 3
[SQL05] Sum of Row Sums 3
[SQL06] Sum, Min, Max Per Column 4
[SQL07] Count of Distinct Entries Per Column 5
[SQL08] Count of Nulls 6
[SQL09] Count of correct separator "-" locations and dashes 7
[SQL10] Eyeball Check 7
[SQL11] Count of Rows Where Order Date Is Before Or Equal To Ship Date 8
[SQL12] Maximum precision (no. of decimal places) 8
[Manual check] Date format Check 8
General Notes 8

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL01] Count of Rows

This table compares the number of rows in csv (source) and sql (target) files.

Source	Target	Outcome
51290	51290	PASS

[SQL02] Count of Distinct Rows

This table compares the number of unique rows in csv (source) and sql (target) files.

Source	Target	Outcome
51290	51290	PASS

[SQL03] Count of Columns

This table compares the number of columns in csv (source) and sql (target) files.

Source	Target	Outcome
24	24	PASS

[SQL04] Sum of Column Sums

This table compares the sum of numeric column sums in csv (source) and sql (target) files.

Source	Target	Outcome
1331006116.61916	1331006116.61916	PASS

[SQL05] Sum of Row Sums

This table compares the sum of numeric row sums in csv (source) and sql (target) files.

Source	Target	Outcome
1331006116.61916	1331006116.61916	PASS

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL06] Sum, Min, Max Per Column

This table compares the sum, lowest (min) and highest (max) entries in numeric columns in csv (source) and sql (target) files. Since it is not possible to find a sum of columns, which display date, they are marked as N/A.

Column name	Sum		Min		Max		Outcome
	Source	Target	Source	Target	Source	Target	
Row ID	1315357695	1315357695	1	1	51290	51290	PASS
Order Date	N/A	N/A	01/01/2011	2011-01-01	31/12/2014	2014-12-31	PASS
Ship Date	N/A	N/A	03/01/2011	2011-01-03	07/01/2015	2015-01-07	PASS
Postal Code	551572652	551572652	1040	1040	99301	99301	PASS
Sales	*12642501.9098802	12642501.91	0.444	0.444	22638.48	22638.48	PASS
Quantity	178312	178312	1	1	14	14	PASS
Discount	*7329.72799999923	7329.728	0	0	0.85	0.85	PASS
Profit	*1467457.29127998	1467457.291	-6599.978	-6599.978	8399.976	8399.976	PASS
Shipping Cost	1352820.69	1352820.69	0	0	933.57	933.57	PASS

Note: * The sums of Sales, Discount and Profit in the csv (source) file are more precise than in the sql (target) file. We speculate that it happens because the csv file has been opened with Excel, which supports a higher precision by default, compared to SQL interface – DBeaver.

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL07] Count of Distinct Entries Per Column

This table compares the number of unique entries in each column in csv (source) and sql (target) files.

Column name	Source	Target	Outcome
Row ID	51290	51290	PASS
Order ID	25035	25035	PASS
Order Date	1430	1430	PASS
Ship Date	1464	1464	PASS
Ship Mode	4	4	PASS
Customer ID	1590	1590	PASS
Customer Name	795	795	PASS
Segment	3	3	PASS
City	3636	3636	PASS
State	1094	1094	PASS
Country	147	147	PASS
Postal Code	631	631	PASS
Market	7	7	PASS
Region	13	13	PASS
Product ID	10292	10292	PASS
Category	3	3	PASS
Sub-Category	17	17	PASS
Product Name	3788	3788	PASS
Sales	22995	22995	PASS
Quantity	14	14	PASS
Discount	27	27	PASS
Profit	24575	24575	PASS
Shipping Cost	10037	10037	PASS
Order Priority	4	4	PASS

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL08] Count of Nulls

This table compares the number of missing values in csv (source) and sql (target) files.

Column name	Source	Target	Outcome
Row ID	0	0	PASS
Order ID	0	0	PASS
Order Date	0	0	PASS
Ship Date	0	0	PASS
Ship Mode	0	0	PASS
Customer ID	0	0	PASS
Customer Name	0	0	PASS
Segment	0	0	PASS
City	0	0	PASS
State	0	0	PASS
Country	0	0	PASS
Postal Code	41296	41296	PASS
Market	0	0	PASS
Region	0	0	PASS
Product ID	0	0	PASS
Category	0	0	PASS
Sub-Category	0	0	PASS
Product Name	0	0	PASS
Sales	0	0	PASS
Quantity	0	0	PASS
Discount	0	0	PASS
Profit	0	0	PASS
Shipping Cost	0	0	PASS
Order Priority	0	0	PASS

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL09] Count of correct separator "-" locations and dashes

This table compares the number of cases, wherethe separator "-" locations have been placed correctly, according to the format of Order ID, Customer ID and Product ID entries, in (source) and sql (target) files. This quality check also compares the number of separators "-" for each ID between (source) and sql (target) files.

Column name	Number of correct separator "-" locations		Number of separator "-"		Outcome
	Source	Target	Source	Target	
Order ID	51290	51290	2	2	PASS
Customer ID	51290	51290	1	1	PASS
Product ID	51290	51290	2	2	PASS

[SQL10] Eyeball Check

This table displays the row ID' s of randomly selected rows, which were visually compared between csv (source) and sql (target) files.

Row ID	Outcome
23988	PASS
11407	PASS
4728	PASS
15850	PASS
46370	PASS

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL11] Count of Rows Where Order Date Is Before Or Equal To Ship Date

This table compares the number of rows, where Order Date is before or equal to ship date in csv (source) and sql (target) files. This control check ensures that the date entries are logical and consistent.

Source	Target	Outcome
51290	51290	PASS

[SQL12] Maximum precision (no. of decimal places)

This table compares the maximumprecision of numeric (excluding date type) columns in csv (source) and sql (target) files. The values represent the number of decimal places.

Column name	Source	Target	Outcome
Row ID	0	0	PASS
Sales	*5	*5	PASS
Quantity	0	0	PASS
Discount	3	3	PASS
Profit	*5	*5	PASS
Ship Cost	2	2	PASS

Note: * Sales and Profit entries have a higher precision than 2, which would be appropriate for monetary figures.

[Manual check] Date format Check

This table compares the date format used for Order Date and Ship Date columns in csv (source) and sql (target) files.

Column name	Source	Target	Outcome
Order Date	*DD/MM/YYYY	*D/M/YYYY	PASS
Ship Date	*DD/MM/YYYY	*D/M/YYYY	

Note: * The day format in csv (source) file takes 2 places [DD], while in sql (target) file takes just one [D]. The difference lies in the data type used in the sql (target) file, where it has been assigned as variable character for the ease of import.

OFFUTURE
THE FUTURE OF OFFICE SUPPLIES

[SQL13] Duplicate Product ID check

This table displays the duplicate Product ID's and the number of Product Names for each ID.

Product ID	Number of Product Names
OFF-PA-10001971	2
OFF-EN-10001832	2
OFF-SU-10004662	2
OFF-AVE-10002102	2

General Notes

- Sales and profit precision is higher than 2, which would be appropriate for monetary figures. We speculate this might have happened due to currency exchange.
- Sum check for columns Sales, Discount, and Profit show there are values, which have a decimal place of more than 2:
    - Sales – up to 4 decimal places
    - Discount – up to 8 decimal places
    - Profit – up to 3 decimal places
- Although the dataset includes information from collected from 2011 to 2014, shipping date goes after 2015.
- There are products, having inconsistent categories, for example, a product, which goes by name "Staples" has a lot of categories, whereas other staples have just one category.
- Some inconsistencies can be found within Markets. For example, Austria is included in both EU and EMEA markets and Mongolia belonging to both APAC and EMEA markets.
- Region column is inconsistent with its values and definition. We notice names of markets, continents or global directions as region values.