Quality Control Form

Date:12/07/2021

Project: Descriptive Analytics Project

Description: Preparing and exploring offuture sales data to find patterns and trends.

Source: offuture.csv

Destination: student schema of sandbox SQL -> student.t4_offuture

Methodology:

To measure the quality of data, we took into consideration checks for completeness of data, types of columns in data, range of numerical attributes, number of empty values, number of duplicated rows, number of unique values in each column, and descriptive statistics of numerical and categorical columns.

- For completeness of data, we checked for number of rows and columns in source and destination files.
- Subsequently, type of values in each column was observed in both source and destination files. “Order date” and “Ship date” columns have type “Date” in source file but in destination file, their type has been changed to “String”. We did not opt to convert type of these columns to “Date” in destination file as it was causing error in uploading data to PostgreSQL table. Generally, it did not impact our analysis.
- Further, range of numerical columns was examined in both source and destination files by calculating minimum and maximum value in each column. In PostgreSQL, min and max function were used for this purpose.
- Next, ratio of missing values was analysed in both source and destination files. In PostgreSQL, “IS NULL” condition was used to test for missing values in each column.
- Moreover, duplicated number of rows were assessed in both source and destination files. For destination file assessment, Group By and HAVING clauses were used to filter duplicated rows.
- Following this, number of unique values in each column of source and destination data were observed. In PostgreSQL, COUNT and Distinct function were used to get unique values in each column.
- Finally, descriptive statistics of both numerical and categorical columns was analysed in source and destination data. For numerical columns, metrics such as mean, standard deviation and median were calculated. For categorical columns, frequency and name of category with maximum values was noted.

A detailed information about quality check metrics is demonstrated in the below table.

Quality Checks:

Check	Source	Destination
Completeness of data	Number of rows: 51290 Number of columns: 24	Number of rows: 51290 Number of columns: 24
Types of Columns in data	Row ID: **Integer** Order ID: **String** Order Date: **Date** Ship Date: **Date** Ship Mode: **String** Customer ID: **String** Customer Name: **String** Segment: **String** City: **String** State: **String** Country: **String** Postal Code: **Integer** Market: **String** Region: **String** Product ID: **String** Category: **String** Sub-Category: **String** Product Name: **String** Sales: **Real/float** Quantity: **Integer** Discount: **Real/float** Profit: **Real/float** Shipping cost: **Real/float** Order Priority: **String**	Row ID: **Integer** Order ID: **String** Order Date: **String** Ship Date: **String** Ship Mode: **String** Customer ID: **String** Customer Name: **String** Segment: **String** City: **String** State: **String** Country: **String** Postal Code: **Integer** Market: **String** Region: **String** Product ID: **String** Category: **String** Sub-Category: **String** Product Name: **String** Sales: **Real/float** Quantity: **Integer** Discount: **Real/float** Profit: **Real/float** Shipping cost: **Real/float** Order Priority: **String**

Range of Numerical attributes in database	Row ID: (**1, 51290)** Postal Code: **(1040, 99301)** Sales: (**0.444, 22638.48)** Quantity: (**1, 14)** Discount: (**0, 0.85)** Profit: **(-6599.978, 8399.976)** Shipping cost: **(0, 933.57)**	Row ID: (**1, 51290)** Postal Code: **(1040, 99301)** Sales: (**0.444, 22638.48)** Quantity: (**1, 14)** Discount: (**0, 0.85)** Profit: **(-6599.978, 8399.976)** Shipping cost: **(0, 933.57)**
Null rate in data	41296 missing values in Postal code attribute All other column in table have no missing value.	41296 missing values in Postal code attribute All other column in table have no missing value.
Duplicated observation in the data	No duplicated row in the data	There is no duplicated observation in the data
Number of Unique values in data attributes	Row ID: **51290** Order ID: **25035** Order Date: **1430** Ship Date: **1464** Ship Mode: **4** Customer ID: **1590** Customer Name: **795** Segment: **3** City: **3636** State: **1094** Country: **147** Postal Code: **631** Market: **7** Region: **13** Product ID: **10292** Category: **3** Sub-Category: **17** Product Name: **3788** Sales: **22995** Quantity: **14** Discount: **27** Profit: **24575** Shipping cost: **10037** Order Priority: **4**	Row ID: **51290** Order ID: **25035** Order Date: **1430** Ship Date: **1464** Ship Mode: **4** Customer ID: **1590** Customer Name: **795** Segment: **3** City: **3636** State: **1094** Country: **147** Postal Code: **631** Market: **7** Region: **13** Product ID: **10292** Category: **3** Sub-Category: **17** Product Name: **3788** Sales: **22995** Quantity: **14** Discount: **27** Profit: **24575** Shipping cost: **10037** Order Priority: **4**
Descriptive statistics for numerical attributes	**Sales** Mean: 246.4906 Standard deviation: 487.5654 Median: 85.056 **Quantity** Mean: 3.4765 Standard deviation: 2.2788	**Sales** Mean: 246.4906 Standard deviation: 487.5654 Median: 85.056 **Quantity** Mean: 3.4765 Standard deviation: 2.2788

	Median: 3 **Discount** Mean: 0.1429 Standard deviation: 0.2123 Median: 0 **Profit** Mean: 28.6109 Standard deviation:174.3409 Median: 9.24 **Shipping Cost** Mean: 26.3759 Standard deviation: 57.2968 Median:7.79	Median: 3 **Discount** Mean: 0.1429 Standard deviation: 0.2123 Median: 0 **Profit** Mean: 28.6109 Standard deviation:174.3409 Median: 9.24 **Shipping Cost** Mean: 26.3759 Standard deviation: 57.2968 Median:7.79
Descriptive statistics for categorical attributes	**Order Date** Between 01/01/2011 and 31/12/2014 **Ship Date** Between 01/01/2012 and 31/12/2014 **Ship Mode** Distinct values: First Class, Second Class, Standard Class, Same Day Top category: Standard class frequency:30775 **Segment** Distinct values: Consumer, Corporate, Home Office Top category: Consumer Frequency: 26518 **City** Top: New York City Frequency: 915 **State** Top: California Frequency: 2001 **Country** Top: United States Frequency: 9994 **Market** Distinct values: APAC, LATAM, EU, US, EMEA, Africa, Canada. Top: APAC Frequency: 11002 **Region** Distinct values: Central, South, EMEA, North, Africa, Oceania, West, Southeast Asia, East, North Asia, Central Asia, Caribbean, Canada Top: Central Frequency: 11117	**Order Date** Between 01/01/2011 and 31/12/2014 **Ship Date** Between 01/01/2012 and 31/12/2014 **Ship Mode** Distinct values: First Class, Second Class, Standard Class, Same Day Top category: Standard class frequency:30775 **Segment** Distinct values: Consumer, Corporate, Home Office Top category: Consumer Frequency: 26518 **City** Top: New York City Frequency: 915 **State** Top: California Frequency: 2001 **Country** Top: United States Frequency: 9994 **Market** Distinct values: APAC, LATAM, EU, US, EMEA, Africa, Canada. Top: APAC Frequency: 11002 **Region** Distinct values: Central, South, EMEA, North, Africa, Oceania, West, Southeast Asia, East, North Asia, Central Asia, Caribbean, Canada Top: Central Frequency: 11117

	Category Distinct Values: Office Supplies, Technology, Furniture. Top: office supplies Frequency: 31273 **Sub-Category** Distinct Values: Binders, Storage, Art, Paper, Chairs, Phones, Furnishings, Accessories, Labels, Envelops, Supplies, Fasteners, Bookcases, Copiers, Appliances, Machines, Tables. Top: Binders Frequency: 6152 **Product Name** Top: staples Frequency:227 **Order Priority** Distinct Values: Medium, High, Critical, Low Top: Medium Frequency:29433	**Category** Distinct Values: Office Supplies, Technology, Furniture. Top: office supplies Frequency: 31273 **Sub-Category** Distinct Values: Binders, Storage, Art, Paper, Chairs, Phones, Furnishings, Accessories, Labels, Envelops, Supplies, Fasteners, Bookcases, Copiers, Appliances, Machines, Tables. Top: Binders Frequency: 6152 **Product Name** Top: staples Frequency:227 **Order Priority** Distinct Values: Medium, High, Critical, Low Top: Medium Frequency:29433