Data Cleaning via Excel

The case study csv file consists of following columns: Order Time, User ID, Order Pickup Geo, Order Drop Geo, Order ID, Products, Partner Store Reach Time, Partner Start for Delivery Time, Completed/Canceled Time, Completion Flag, Order Rating, Product Amount, Delivery Charges, Discount

The data in column: Order Timestamp is in the wrong format for further analysis, and also the data is unable to convert into ‘Long Date’ format via the dropdown.

To clean the data, the following formula is used:

=CONCATENATE(LEFT(A2,10),” “,MID(A2,12,5))

A new column with the name ‘order\_timestamp’ is created and the above formula is used for the entire column.

In order to get the time (hour, minute) from the Order Time, we use another formula by creating a new column - ‘order\_timeslot’:

=TEXT(MID(A2,12,5),”hh:mm AM/PM”)

To view the month number for further analysis, we extract the month number from the first column using the formula:

=MID(A2,6,2)

We convert the month number into actual calendar month using the below formula:

=TEXT(D1,”mmm”)

In order to get on which day of the week a order has been booked, we use the formula:

=TEXT(LEFT(B2,10),”dddd”)

In the products column, it is visible that sometimes a user purchases more than 1 product. To determine how many products an order ID contain, we use the below formula:

=LEN(K2)-LEN(SUBSTITUTE(K2,",","")) + 1

In the products column, we could see each product is separated by a comma(,). Hence we count the total number of commas and 1 to it since products are one more than comma.

The columns ‘Partner Store Reach Time’, 'Partner Start for Delivery Time’, 'Completed/Canceled Time’ also are in wrong formats. Those columns are fixed using the method used for the ‘Order Timestamp’ column.

Now to calculate the difference of time between order booking time and order delivery time, we use the below formula:

=(RIGHT(R2,5) - RIGHT(B2,5)) \* 24 \* 60

In the order rating column, it could be seen that there are many blank spaces even when the completion\_flag column has ‘YES’ in its adjacent column.

For further analysis, the null values are filled with the average of the order\_rating column which is 4.8 rounded off to 5 since it’s the nearest value.

Note: The cells in order ratings must remain NULL if its adjacent Completion Flag = ‘NO’ which means the order hasn’t been completed.

To do this, we use the below formula:

=IF(S2 = ‘NO’,S2=””, IF(U2=””,5,U2))

In the above formula, firstly it is checked if the adjacent Completion flag column has ‘NO’. If so, then leave the order rating cell blank. If not, then it is checked whether the adjacent order rating column is empty, if so then we input 5 since it’s the average or else the original value from order rating column. All this data is entered in a new column.