1. The "data\_v1.csv" is the original raw data. The date formatting in the fields "Order Date" and "Ship Date" were not consistent. Some field values were in MM/DD/YYYY format whereas some were in DD-MM-YYYY format.
2. A python notebook named "data\_preprocessing.ipynb" is used to set the format as DD-MM-YYYY throughout all the date values.
3. To do so, the two date-type columns are copied and put into a separate file named “dates.csv’. A piece of python code corrects the date format and these correctly formatted date columns are then used in the main data.
4. In the same python notebook, the values in the column “Order Priority” is relabelled as “C”:”Critical”, “H”:”High”, “M”:”Medium” and “L”:”Low” for better understanding.
5. Check for NULL or NAN are performed using python and no such values are found
6. No duplicate entries are found using python
7. The fields Order ID, Units Sold, Unit Price, Unit Cost, Total Revenue cannot be negative. A sign validation for these fields is performed and no discrepancy is found.
8. This corrected formatted data along with all data is stored in a new file name "data\_v2.csv".
9. Further analysis is done in Tableau Public
10. In the Tableau environment, the default format for date is MM/DD/YYYY
11. A calculated field named “Delivery time” is created. A filter is used to filter out negative delivery time in the Tableau data-source page for analysis consistency.
12. Country and Region are put in folder “Location”
13. Order Date, Order ID, Order Priority are put in folder “Order”
14. Rest of the things are in the folder “Sales”