Hi There,

Greeting for the day!!!

I am DARSHAN DESAI from KPMG Analytics Team (Virtual Internship program), I have gone through your dataset and found some discrepancies as follows.

|  |  |
| --- | --- |
| Customer Sales Dataset | |
| Total Missing Values - 315 | These Values under following columns as –  Last\_name – 29, DOB – 17, job\_title – 106, job\_industry\_category - 165 |
| Transactions Dataset | |
| Total Missing Values - 1542 | These values under following columns as –  Online\_order – 360 missing and brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date each having 197 missing values |
| Customer Demographic Dataset | |
| Missing Values - 1461 | These Values are missing under following columns as – last\_name – 125, DOB – 87, job\_title – 506, job\_industry\_category – 656, tenure - 87 |
| Date of birth issues - 1 | One entry contains DOB as 1843 which is an issue |
| Column issue | Default column in dataset contains vague information |

* Data inconsistency is there in the dataset which affects the training dataset and eventually affects the whole analysis.  
  To solve this one can use either of string or numeric type selection only this will reduce the issue.  
    
  Also, in gender and address column we can use the drop-down list or radio button (if certain states only allowed) that will help to reduce human prone errors in terms of shortcuts. For example there is a gender named "U' which will get removed by the analysis team.
* In the dataset at the date of the birth column, the restriction should be enforced like the maximum age should be less than 120 years or likewise.  
  As we can see the one column contains birth year as 1843 years which is completely outlier to data.
* The data is not in sync amongst each sheet as there is no customer\_id in the customer demographic.  
  This will be treated as an outlier in the dataset.   
  To solve this maintain data in sync.

Kindly, look at the above table which contains quality issues over a given dataset which could hamper the further analytics.

Thanks and Regards,

Darshan Desai