Hello,

This is Kshitij Gurbhele from KPMG Data analytics(Virtual Internship) team. We have reviewed the databases provided by your company , and we found some errors in it which are listed as follows:

|  |  |  |
| --- | --- | --- |
| Table Name | No. of Records | Missing records |
| Transactions | (20000, 13) | |  |  | | --- | --- | | Online\_order | 360 | | Brand | 197 | | Product\_line | 197 | | Product\_class | 197 | | Product\_size | 197 | | Standard\_cost | 197 | | product\_first\_sold\_date | 197 | |
| NewCustomerList | (1000, 23) | |  |  | | --- | --- | | Last\_name | 29 | | DOB | 17 | | Job\_title | 106 | | job\_industry\_category | 165 | |
| CustomerDemographic | (3999, 17) | |  |  | | --- | --- | | Last\_name | 125 | | DOB | 87 | | Job\_title | 506 | | job\_industry\_category | 656 | | default | 302 | | tenure | 87 | |
| CustomerAddress | (3999, 6) | 0 |

Missing values of:-

* Online order should be dropped by considering them fake or unauthorised.
* Brand to be replaced by mode or dropped depending on needs of predictive model.
* Product line, product class, product size to be replaced by mode.
* Standard cost to be replaced by mean.
* product\_first\_sold\_date replace by mode.
* Last name to be assigned some common value
* Job title,job industry category to be replaced bymode.
* Default to be dropped because of garbage values present in it.
* Tenure to be replaced by mean.

Some other data cleaning done is as follows:

* product\_first\_sold\_date dtype changed from float to datetime.
* New customer list contained category “U” under gender that was further changed to “Unspecified” for convenience.
* CustomerDemographic had category “F”, “Femal” & “M” to “Female” and “Male” respectively.
* Customer id’s were added to NewCustomerList succeeding the table CustomerDemographic and then both were merged to form a single final list.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Regards,

Kshitij Gurbhele