To: [manager@kpmg-dummy.com](mailto:manager@kpmg-dummy.com)

Subject: Sprocket Central Pty Ltd – Data Quality Assessment

Hello Sir,

I have thoroughly reviewed the datasets provided by the client in the file “KPMG\_VI\_New\_raw\_data\_update\_final” and performed the Data Quality Assessment over the given datasets –

* Customer Demographic
* Customer Addresses
* Transaction data in the past three months

I have identified certain data issues in the provided datasets and have also contributed with the best possible transformations I can think of. In order to deal with few issues, I felt that a clarification from the client will be the best way to deal with it.

Please review the analysis I have performed, and you can reach out to me for any elaboration or discussion. I would be privileged to receive feedbacks about any better methods we can implement to deal with these data issues.

Thank you for this wonderful opportunity!

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Data Issues** | **Possible Resolution/ Transformation** | **Client Confirmation/ Assumption** |
| **Transactions (Total Records = 20,000)** | | | |
| **transaction\_id** |  | **Eliminate column** |  |
| **product\_id** | For product\_id=0, records=1378 For 197 records(14%), values for columns brand, product\_line, product\_class, product\_size, standard\_cost, product\_first sold\_date are missing(50%). | Complete elimination of records with product\_id=0 might bias the data. | Client - Specifications/ industry standard for product\_id=0 |
| **transaction\_date** |  | Derive month attribute and **eliminate column**. |  |
| **online\_order** | 1. 360 Missing values (1.8%) 2. Boolean data values ['TRUE', 'FALSE', ' '] | 1. Records with missing values can be eliminated 2. Transform to **int64** ['TRUE'=1, 'FALSE'=0] |  |
| **order\_status** | String values present  ['approved', 'cancelled'] | Transform to **int64** ['approved'=1, 'cancelled'=0] |  |
| **brand** | 1. 197 missing values (<1%) 2. String values present | 1. Dependency on product\_id=0 2. nan = 0, Giant = 1, Narco = 2 OHM = 3, Solex = 4, Trek = 5 WeareA2B = 6 |  |
| **product\_line** | 2. nan = 0, standard = 1,  road = 2, mountain = 3 touring = 4 |  |
| **product\_class** | 2. nan = 0, low = 1, medium = 2 high = 3 |  |
| **product\_size** | 2. nan = 0, small = 1,  medium = 2, large = 3 |  |
| **product\_first\_sold\_date** | 1. 197 missing values (<1%) 2. int64 (5 digit) values present instead of datetime64[ns] |  | Client - Specification about 5 digit value |
| **Customer Demographics (Total Records = 4,000)** | | | |
| **last\_name** | 125 missing values (3%) | Customer info must not be ignored as the customer\_id is available. **No actions to be taken.** |  |
| **gender** | 1. Dirty inputs & inconsistency 2. String values present | Feature Transformation Male, M = 0 Female, F, Femal = 1 U = 2 |  |
| **DOB** | 1. 87 missing values (2%) 2. 1 record (customer\_id=34) contains input for the year 1843 | 1. For customer\_id=34, year of birth replaced to 1943. 2. Derive attribute **'year'** | Assume - Typo of 1943 to 1843 |
| **job\_title** | 1. 656 missing values (16%) 2. 195 unique values | Column can be eliminated altogether by ignoring data issues |  |
| **job\_industry\_category** | String values present | nan = 0, 'Manufacturing' = 1  'Financial Services'=2  'Entertainment' = 3, 'Retail' = 4  'IT' =5, 'Telecommunications' = 6 'Health' = 7, 'Argiculture' = 8, 'Property'=9 |  |
| **wealth\_segment** | String values present | Feature Encoding 'Mass Customer' = 1  'Affluent Customer' = 2  'High Net Worth' = 3 |  |
| **deceased\_indicator** | [N, Y] | Eliminate 2 records having value = Y |  |
| **default** | dirty input | **Eliminate column** |  |
| **owns\_car** | [Yes, No] | Yes=1, No=0 |  |
| **tenure** | 87 missing values (2%) | Replace by mean value. | Client - Specification for empty values (if value should be substitute by 0) |
| **Customer Address (Total Records = 3,999)** | | | |
| **address** |  | Can be eliminated due to unique values |  |
| **postcode** |  | Can be eliminated due to unique values |  |
| **state** | dirty input | New South Wales, NSW = 2 QLD = 1,  VIC, Victoria = 3 | Assume - VIC=Victoria, New South Wales=NSW |
| **country** | Single value for all records | **Eliminate column** |  |

Regards,

Prabodh Badwaik

[linkedin.com/in/prabodhbadwaik27](http://linkedin.com/in/prabodhbadwaik27)

+91 9175063905