**DATA QUALITY ASSESSMENT**

Respected Sir/madam,

I have looked through each of the data sets provided and run some basic checks on them using R. I have summarized my results with regards to the problems present in the data, along with possible solutions(in green), below. The metrics used for assessment include, but are not limited to, accuracy, completeness, uniformity, validity and relevancy.

Transactions Data

1. No associated meta data or data description. For instance, there is no information about the units of prices or any description of what each column means leading to issues in Relevancy.

**The solution for this needs to come from the client end as only data collectors are aware of correct data descriptions.**

1. **-**  Around 3% of the data contains at least one missing value/column entry. Under normal circumstances this does not amount to much – but in a sample size of 20000 this amounts to around 500 observations. It is reasonable to assume this does impact analysis and predictions quite significantly.

**-**  Majority of missing values are regarding the brand and details like size , class, std cost etc. and also on whether the order was online or not. This could be significant while trying to analyse online vs offline purchase trends.

**Missing values can be imputated if possible, else ignored. It is unreasonable to expect the client to trace the same person and re-gather information.**

1. From the earliest and latest date values we can conclude that the values are 3 years old - not up to date – currency issue.

**To obtain accurate and reliable findings it is essential to have recent data. If possible, the client is requested to provide us with the same.**

1. Besides the ones listed above, there were a few extremely minor consistency issues, not worth mentioning. There is one metric which we cannot verify – accuracy, since how the data was collected is not something we are aware of.

Customer Demographic Data

1. No associated meta data or data description.
2. **-** Around 14% of the data contains at least one missing value/column entry. This is significant especially since there are only 4000 observations in total of which more than 800 have missing values.

**-** Last names, job titles and categories are predominantly missing. This will hinder customer segmentation and the understanding purchase preferences of each demographic.

**As mentioned before, the only solution is to impute missing values where possible. If not, the entire observation must be omitted.**

1. **-** One DOB is 1843 - which is impossible and inaccurate.

**-** Further, there appear to be around 82 entries having DOBs in the years 2002,2001 and 2000 having rather descriptive and senior job titles – these people only 15-16 years of age during the time of transactions(which are all in 2017), which seems implausible since they cannot logically hold senior/managerial titles at such ages. This brings into question the accuracy of data provided.

**-** Many people of this age group also appear to own homes and cars - again dubious. This could lead to severe inaccuracies in understanding the customer demographic, especially given the large number of missing values.

**There is no solution to this besides advocating for better data collection practices. The onus is not only on the client – sometimes customers provide incorrect data regardless of the practices followed.**

Customer Address Data

There are no major inaccuracies in this data set.