**MIHIR PAHWA**

[**mihirsp707@gmail.com**](mihirsp707@gmail.com)

**INTITAL ANALYSIS**

SUBJECT : Problems with given dataset.

Sir/Ma’am,

The dataset which you sent us had a few problems. I have NOT deleted any data while cleaning because I need clarifications on a few points listed as below—

* **Sheet 2- Transactions**
  + There are rows in *brands*, *orders*, *product\_class* and other columns which are unknown or Null. There are 195 null values .
  + The product\_first\_sold\_date column looks wrong and does not resemble date format. I haven’t removed it yet.
* **Sheet 3- New Customer**
  + I have assumed that property valuation was in millions so example, Min value of Property valuation = 0.1 million. If this is incorrect you can tell me so. I will be happy to correct it.
  + I have removed deceased and country columns as they are of no significance as they have only one unique values respectively
  + DOB column has some 70 null values, along with *job\_title*(103) and *job\_industry\_category*(159)
* **Sheet 4 – Customer Demographics**
  + DOB columns cannot be parsed because it’s a string. I’m working on it.
  + *Job\_title* has 506 nulls, *job \_industry\_category* has 656 nulls
  + Also , The DEFAULT column is not clear to me, I think it’s a xml or html script, I will talk to my seniors and get back to you on that.

**I HAVE ADDED THE R SCRIPTS NEEDED AT THE END OF EVERY FILE FOR YOU TO CLEAN THE DATASET IF YOU NEED IT, JUST IN CASE.**

**YOU CAN ALSO USE Summary() function included at end of every script for more info.**

**UNCOMMENT-> RUN SCIPT-> NULL VALUES WILL BE REMOVED**

|  |  |
| --- | --- |
| **transaction\_id** | **20000** |
| **Customer\_id** | **5034** |
| **Approved\_orders** | **19821** |
|  |  |