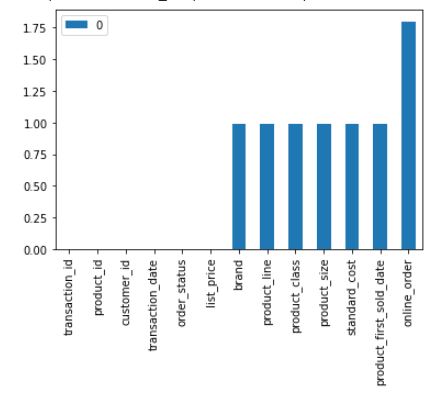
Hello Sir,

I have gone through the datasets you have provided and have found some data quality issues. I think this is very important to mention as the quality of data can affect the analysis that will be done in the future

First of all, for the dataset you have provided, there is no metadata. So interpreting or understanding some of the columns has become pretty difficult and we would prefer dropping such columns if we cannot get any meaningful info from that column.

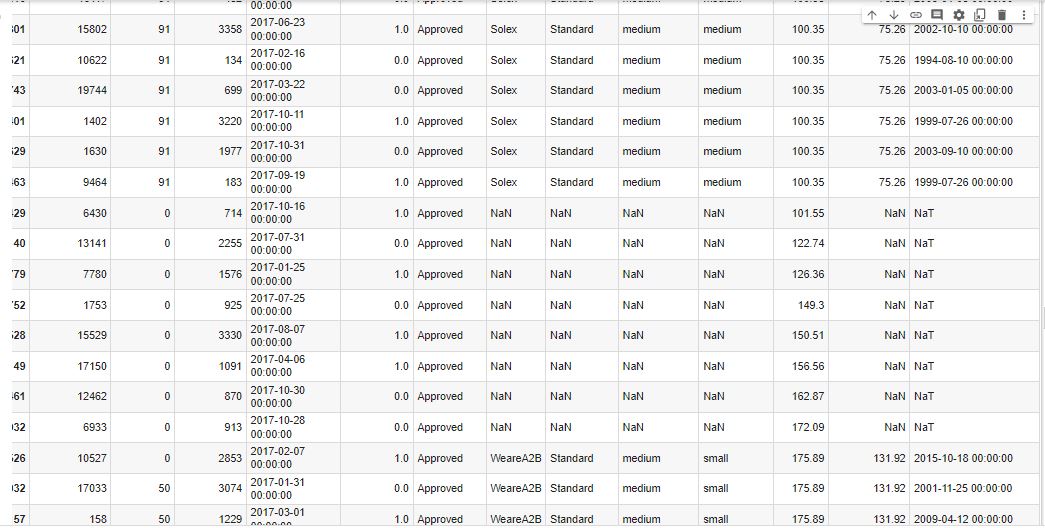
Regarding the **Transactions** dataset

* I was told that the dataset contains transaction data of **3 months** but according to my analysis it has **12 months** of transaction data. So our observations will be based on 12 months of data.
* The dataset contains **missing values**. Here I am attaching a screenshot from my analysis to show what percentage of values are missing from the respective columns



The column **online\_order** has the highest percentage of missing values

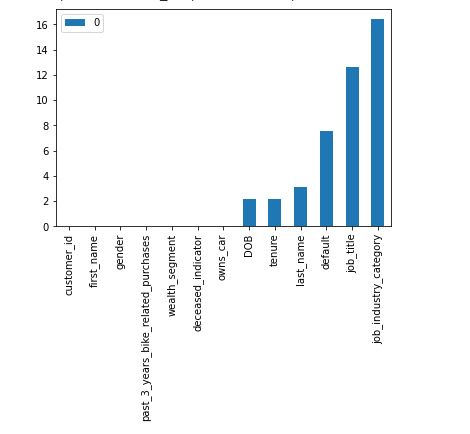
* There is a strange observation regarding missing values. **197 rows** have been found who have values missing from the columns **brand,product\_line, product\_class, product\_size ,standard\_cost ,product\_first\_sold\_date** at the same time. And all these involve the **product\_id 0**. Out of these 197 rows, **2 rows** also have the value from the column **online\_order** missing. The list prices of these products also seem very uncommon and ‘non standard’. I am attaching a screenshot so that you can understand what I want to convey



* It is very unclear what the column **product\_first\_sold\_date** here refers to. Does this refer to the first time that product was launched? If that is so then that column seems to be faulty as the the product with the same **product\_id**, **brand,product\_line, product\_class, product\_size ,standard\_cost** and **list\_price** has different launch dates. Or does this refer to the first time that customer brought that specific product from your company? Even in that case that column seems faulty. We need to clearly understand what that refers to. Or else we would prefer working without that column, as it is not giving much useful information.
* There is a **customer id 5034** that has 3 transactions but we do not have any info about that customer in the customer demographics dataset.

Regarding the **CustomerDemographic** dataset

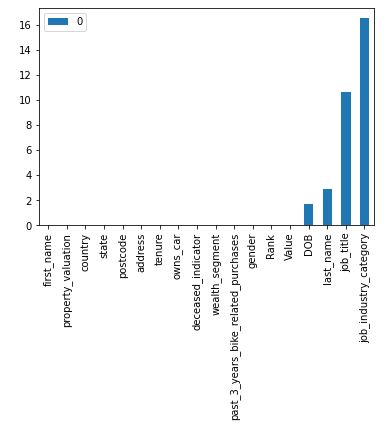
* We found some columns which were having **missing values**. I am attaching the screenshot showing the percentage of values missing from each column



* Gender column apparently had **6 unique values**, upon further analysis it was found they were **'F', 'Male', 'Female', 'U', 'Femal', 'M'**. I guess it must be some mistake. For our future analysis we assumed F=Female=Femal=F and Male=M=M and U is perhaps unknown or the customer did not wish to disclose.
* Customer ID 34 whose full name is Jephthah Bachmann has his DOB listed **1843-12-21**. This seems a mistake
* It is very difficult to understand what the **default column** wants to convey. For our analysis we will be dropping that column.
* For 2 customer ids 1521 and 2054, we found that their **ages are almost equal to their tenure** which is absurd.

Regarding the **NewCustomerList** dataset

* We found some columns which were having **missing values**. I am attaching the screenshot showing the percentage of values missing from each column



* Other than that we did not find much issue in this dataset.

For the **CustomerAddress** dataset

* We saw that for the column state the unique values have been repeated. The unique values were **'New South Wales' ,'QLD' ,'VIC', 'NSW', 'Victoria'**. We know that New South Wales=NSW and Victoria=VIC
* For the customer ids **3, 10, 22** and **23** in the CustomerDemographics table, there is no address that has been given.
* Other than that we didn’t find any other issue here. There were no missing values as well.