Dear Mr. Smith,

In this email, I will highlight data quality concerns and provide solutions to these problems. You may read the summary quickly, but you can also investigate further using my Jupiter notebook and the details below.

I hope my suggestions are useful, and please do not hesitate to contact me or the Lighthouse & Innovation team at KPMG if you have any further concerns or questions.

Sincerely,  
Ngoc Nguyen  
KPMG’s Lighthouse Innovation Intern

—----------------------—-------------------------------------------—----------------------—------------------  
I’ll evaluate the dataset quality following 7 dimensions: Accuracy, Completeness, Consistency, and Currency, Relevancy, Validity, and uniqueness

**Evaluate process:** In order to explore the datasets, I will utilize Python along with helpful packages and tools. Follow this link for more information and my evaluation procedure with source code:

***Summary:***

|  | Transactions | New CustomerList | Customer  Demographic | Customer  Address |
| --- | --- | --- | --- | --- |
| Accuracy | ‘​​'product\_first\_sold\_date' is wrong format | Multiple columns are wrong format  ‘Property\_valuation’ bonded with NaN colum | BOB inaccurate | Accurate |
| Completeness | Missing profit columns  5 NaN columns  Customer\_id is incomplete | Missing values in multiple columns | Missing ‘age’ column  Missing values in multiple columns | No missing value founded |
| Consistency | Consistent | ‘Postcode’ is not consistent | “Gender’ and ‘default’ are not consistent. | ‘State’ is inconsitency |
| Currency | Value up to Dates | Value up to Dates | Value up to Dates | Value up to Dates |
| Relevancy | columns are relevant | ‘deceased\_indicator’ is irrelevant | ‘default’ is irrelevant | columns are relevant |
| Validity | Valid | Valid | Valid | Valid |
| Uniqueness | No duplicated records | No duplicated records | No duplicated records | No duplicated records |

**Mitigate suggestion:**

* Improve the accuracy by:
* Re-format 'product\_first\_sold\_date', ‘past\_3\_years\_bike\_related\_purchases’, ‘postcode’’ and ‘property\_valuation’ to day/time
* Investigate “NaN’ columns and name proper header to them or drop if necessary
* Find inaccurate record in ‘BOB” and fix it.
* Make the dataset more complete by:
* Adding ‘profit’ in Transaction and ‘age’ column in Customer Demographic
* Re-investigate and refill missing values in all columns mention detailly the detail part below. Drop them if necessary for natural anaylis
* Improve the consistency by:
* Make “gender”, “defualt” and “State” more consistent by filter and change uncommon cases to more common cases. For example, filter out “New South Wales” to “NSW", “Victoria" to “VIC".
* I have one recommendation that we should set the decimal place in the consistent value, such as 2 or 3.
* For Currency, please keep record new customer's transaction and infomations and update them.
* For Relevancy:
* Try to decode the ‘default’, for example, using encoding = ‘utf=8’ with python. Otherwise, drop it. Also drop the ‘deceased\_indicator’
* About validity and Uniqueness, I think we’re in the good track of them.

**Detail:**

Accuracy

* Transaction: All the columns in transaction dataset have correct values except ‘​​'product\_first\_sold\_date' is not saved in the day/time format.
* NewCusomterList: From the excel file, I could see ‘past\_3\_years\_bike\_related\_purchases’, ‘postcode’ and ‘property\_valuation’ datas were stored as string instead of numerous data types except the row 30th in ‘postcode’ and ‘property\_valuation’. Furthermore. The ‘property\_valuation’ is bonded with ‘NaN’ column next to it.
* CustomerDemographic: DOB inaccurate
* CustomerAddress: Have correct values on all columns

Completeness

* Transaction: contain 5 ‘NaN’ columns ( no headers and description telling information about that 5 columns. Also only 3495/20001 customer\_id is completed. In addition: ‘online\_order’ has 360 null values, ‘brand’, ‘product\_line’, ‘product\_class’, ‘product\_size’, ‘standard\_cost’ and ‘product\_first\_sold\_date’ has the same of 197 null values.
* NewCustomerList: There are 29 missing customers’ lastname; 17, 106,165 null values in ‘DOB’, ‘job\_title’ and ‘job\_industry\_category’ respectively.
* CustomerDemographic: Missing ‘Age’ column; 125 missing customer’s last name. About nulls value: 87 in ‘DOB’, 506 in ‘job\_title’, 656 in ‘job\_industry\_category’, 302 in ‘default’ and 87 in ‘tenure’.3
* CustomerAddress: There are no missing values in CustomerAddress

Consistency

* Transaction: Throughout my experiment, the data in Transaction’s columns are consistent.
* NewCustomerList: The ‘postcode’ was saved by string datatype instead of numerous. However there are 5 times the missing values were filled by 4000.
* CustomerDemographic: “Gender’ and ‘default’ are not consistent.
* CustomerAddress: ‘State’ is inconsitency

Currency: As mentioned, Sprocket Central Pty Ltd provides the data in the past 3 months. So the datas are currency.

Relevancy

* Transaction: All the columns are relevant
* NewCustomerList: the ‘deceased\_indicator’ is irrelevant to me
* CustomerDemographic: The ‘default’ colum is irrelevant
* CustomerAddress: Relevant.

Validity

* Transaction: Most of columns contain valid values
* NewCustomerList: Most of columns contain valid values
* CustomerDemographic: ‘default’ contains not allowable values.
* CustomerAddress: Most of columns contain valid values

Uniqueness: after using .duplicated() and .nunique() methods in Python, all the datasets yields 0. As the result the dataset is unique.