Program整体时间安排：5-6小时（1天时间）

Data quality assessment: time - 2 hours (search for resources, completing and checking with Templete)

*the importance of optimizing the quality of customer datasets cannot be underestimated. The better the quality of the dataset, the better chance you will be able to use it to drive company growth*.

identifying the data quality issues and strategies to mitigate these issues

Data quality issues & strategies:

Framework:

* Accuracy
* Completeness
* Consistency
* Currency
* Relevancy
* Validity
* Uniqueness

Knowledge:

Tenure refers to the number of months that a customer has subscribed for.

deceased\_indicator: dead

Improvement:

1. Basis summary statistics firstly
2. Classify the issues, not just list the problems one by one
3. Give the mitigation and recommendation
4. Identify the main dataset: Customer Demographic

Email:

Dear Client,

Thank you for providing these datasets. After checking them, there are some data quality issues and we also provide suggestions about dealing with them. The issues and suggestions for 3 datasets are below:

1. Transactions: 20000 records.
   1. transaction\_date: from 2017/1 to 2017/12. Only values between 2017/10-2017/12 are needed, which is 5079 records.
   2. online\_order: 98.205% completeness. For the last 3 months, the completeness is 98.17%.
   3. order\_status: 98.205% completeness. For the last 3 months, the completeness is 98.17%.
   4. brand: 99.02% completeness. For the last 3 months, completeness is 99.114%. And there is consistent in product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date.
   5. list\_price: the format can be changed into the Currency with dollars.
   6. product\_first\_sold\_date: wrong format and it should be changed into date.
2. Customer Demographic: 4000 records
   1. last\_name: 96.875% completeness.
   2. gender: 97.8% completeness. Duplicate values: F, Female, and Famel all represent female; M and Male are both male.
   3. DOB: the customer\_id of 34 has the wrong DOB value, which is 1843-12-21. The completeness is 97.825%. The values of customers with gender U are blank.
   4. job\_title: 87.575% completeness.
   5. job\_industry\_category: 83.6% completeness and 16.4% of values are n/a.
   6. default: useless and can be deleted.
   7. tenure: The completeness is 97.825%. The values of customers with blank DOB are blank.
3. Customer Address: 3999 records
   1. state: New South Wales and NSW are duplicated.
4. The Customer Address lacks the information of customer\_id 3,10,22,23 and has information of 4001, 4002, 4003. But CustomerDemographic doesn’t have records for customer\_id 4001, 4002, 4003. The Transactions has records for customer\_id between 1 to 3500, and 5034.

We are looking forward to your reply.

Sincerely,

Holly