Data Quality Report

Hello

This e-mail was sent to you in order to address some issues we have been having in terms of data quality. I will be starting with outlining the issues we have been encountered and then more detailed analysis will be laid out to exemplify the quality issues with the data.

Data Quality Outline

Before starting out with problems of the data tables that we were presented. I have to point out the importance of the meta data. To produce a sound analysis getting the context of the data is very important. This context is almost always supplied by the meta data. Within meta data columns that a data table contains are explained so there is no questions about the data itself. Thus, we need you to provide the meta data as well.

Transactions

1. `online_order` status is missing for some observations.
2. Some entries miss product information. Affected columns are `brand, product_line, product_class, product_size, standard_cost, product_first_sold_date`.
3. Columns `transaction_date` and `product_first_sold_date` are both dates but they have inconsistent formats.

New Customer List

1. Encountered some hidden columns while reading the data. Please make sure that you expand all the columns and name them correctly.
2. Also one of the hidden columns is a duplicate of `Rank` column.
3. `job_industry_category` has missing values but the representation of the missing values should not be a text placeholder such as `n/a`.
4. There are many inconsistencies about job titles. There are Latin numbers with job titles and meaning of them are vague. There are also missing values.
5. There is a gender called "U".
6. The Date of Birth variable has inconsistent, non-accurate values and missing. Time formatting should be consistent.

Customer Demographic

1. Time formatting is inconsistent. There is one customer who was born in 1843. The same customer also has a gender of "U"

2. Gender Values are inconsistent.
3. Job Titles are ver inconsistent. They should be all lower case. There are also values with Latin numbers such as `Health Coach I` and `Health Coach III`.
4. Customer Demographic table requires a proper Missing Value place holder. In `job_industry_category` there are string "n/a" values.
5. Some customers have missing last names.
6. There is a "default" column with many random values.

Customer Address

1. There are abbreviations for state names along with long state names such as New South Vales and NSW.

Problems which are related to more than one table

1. Some customers in `CustomerDemographic` don't have their correspondents in `CustomerAddress`.
2. Also there are some customers only with addresses in `CustomerAddress`.

In Depth Analysis

Transactions

There are some entries which miss the identification of online_order status. This should be addressed in the data collection process.

```
Transactions %>%
  count(online_order, sort = T)
```

```
## # A tibble: 3 x 2
##   online_order     n
##   <lgl>        <int>
## 1 TRUE          9829
## 2 FALSE         9811
## 3 NA             360
```

```
Transactions %>%
  filter(is.na(online_order)) %>%
  head()
```

```
## # A tibble: 6 x 13
##   transaction_id product_id customer_id transaction_date    online_order
##            <dbl>      <dbl>       <dbl> <dttm>              <lgl>
## 1             98         49         333 2017-06-23 00:00:00 NA
## 2            167         90        3177 2017-04-26 00:00:00 NA
## 3            170          6         404 2017-10-16 00:00:00 NA
## 4            251         63        1967 2017-04-11 00:00:00 NA
## 5            301         78        2530 2017-03-24 00:00:00 NA
## 6            337         82        1615 2017-10-30 00:00:00 NA
## # ... with 8 more variables: order_status <chr>, brand <chr>,
## #   product_line <chr>, product_class <chr>, product_size <chr>,
## #   list_price <dbl>, standard_cost <dbl>, product_first_sold_date <date>
```

There are some entries in data table which have no product information. The missing columns are brand, product_line, product_class, product_size, standard_cost, product_first_sold_date. This problem might have arisen from a system bug. Additionally, all the missing products have the id of "0".

```
Transactions %>%
  filter(is.na(brand)) %>%
  head()
```

```
## # A tibble: 6 x 13
##   transaction_id product_id customer_id transaction_date    online_order
##            <dbl>      <dbl>       <dbl> <dttm>              <lgl>
## 1            137          0         431 2017-09-23 00:00:00 FALSE
## 2            160          0        3300 2017-08-27 00:00:00 FALSE
## 3            367          0        1614 2017-03-10 00:00:00 FALSE
## 4            407          0        2559 2017-06-14 00:00:00 TRUE
## 5            677          0        2609 2017-07-02 00:00:00 FALSE
## 6            781          0         897 2017-05-10 00:00:00 TRUE
## # ... with 8 more variables: order_status <chr>, brand <chr>,
## #   product_line <chr>, product_class <chr>, product_size <chr>,
## #   list_price <dbl>, standard_cost <dbl>, product_first_sold_date <date>
```

```
Transactions %>%
  filter(is.na(brand)) %>%
  count(product_id)
```

```
## # A tibble: 1 x 2
##   product_id     n
##        <dbl> <int>
## 1          0   197
```

Also this table needs a consistent time formatting. Variables transaction_date and product_first_sold_date are both dates but their inconsistent formatting make it harder to read and parse the data.

New Customer List

This table has some "hidden" columns! These columns could easily be missed when analyzing the data in Excel. Make sure to expand all the columns that the data table has. Additionally the Rank and one of the hidden columns ...21 are duplicated.

```
NewCustomerList %>%
  select(...21, Rank) %>%
  head()
```

```
## # A tibble: 6 x 2
##   ...21  Rank
##   <dbl> <dbl>
## 1     1     1
## 2     1     1
## 3     1     1
## 4     4     4
## 5     4     4
## 6     6     6
```

```
NewCustomerList %>%
  select(contains("...")) %>%
  head()
```

```
## # A tibble: 6 x 5
##   ...17 ...18 ...19 ...20 ...21
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  0.56 0.7   0.875 0.744     1
## 2  0.89 0.89  1.11  0.946     1
## 3  1.01 1.01  1.01  1.01      1
## 4  0.87 1.09  1.09  1.09      4
## 5  0.52 0.52  0.65  0.65      4
## 6  0.43 0.538 0.538 0.538     6
```

There is no customer_id column in this table and this might cause problems. Data creation process for new and old customers should be always the same.

`job_industry_category` has missing values but the representation of the missing values should not be a text placeholder such as `n/a`.

```
NewCustomerList %>%
  count(job_industry_category, sort = T)
```

```
## # A tibble: 10 x 2
##    job_industry_category     n
##    <chr>                 <int>
##  1 Financial Services      203
##  2 Manufacturing           199
##  3 n/a                     165
##  4 Health                  152
##  5 Retail                   78
##  6 Property                 64
##  7 IT                       51
##  8 Entertainment            37
##  9 Argiculture              26
## 10 Telecommunications       25
```

There are many inconsistencies about job titles. There are Latin numbers with job titles and meaning of them are vague, they should be merged. There are also missing values.

```
NewCustomerList %>%
  count(job_title, sort = T) %>%
  slice(80:90)
```

```
## # A tibble: 11 x 2
##    job_title                         n
##    <chr>                         <int>
##  1 Occupational Therapist            5
##  2 Programmer III                    5
##  3 Research Nurse                    5
##  4 Sales Associate                   5
##  5 Speech Pathologist                5
##  6 Tax Accountant                    5
##  7 Accountant III                    4
##  8 Budget/Accounting Analyst III     4
##  9 Community Outreach Specialist     4
## 10 Database Administrator III        4
## 11 Editor                            4
```

There is a gender called "U". Is this a bug or some placeholder for non-binary persons?

```
NewCustomerList %>%
  count(gender)
```

```
## # A tibble: 3 x 2
##   gender     n
##   <chr>  <int>
## 1 Female   513
## 2 Male     470
## 3 U         17
```

The Date of Birth variable has inconsistent and non-accurate values. Time formatting should be consistent.

Customer Demographic

Time formatting is inconsistent in this table too. ALso someone was born in 1843.

```
CustomerDemographic %>%
  select(DOB) %>%
  slice(34:38)
```

```
## # A tibble: 5 x 1
##   DOB
##   <date>
## 1 1843-12-21
## 2 1963-09-28
## 3 1977-11-09
## 4 1985-12-22
## 5 1955-10-29
```

Gender Values are inconsistent. Values entry for these types of columns should be restricted to pre-defined values.

```
CustomerDemographic %>%
  count(gender, sort = T)
```

```
## # A tibble: 6 x 2
##   gender     n
##   <chr>  <int>
## 1 Female  2037
## 2 Male    1872
## 3 U         88
## 4 F          1
## 5 Femal      1
## 6 M          1
```

Job Titles are ver inconsistent. They should be all lower case. There are also values with Latin numbers such as `Health Coach I` and `Health Coach III`.

```
CustomerDemographic %>%
  count(job_title, sort = T) %>%
  slice(90:99)
```

```
## # A tibble: 10 x 2
##    job_title                         n
##    <chr>                         <int>
##  1 Computer Systems Analyst I       15
##  2 Safety Technician II             15
##  3 Computer Systems Analyst II      14
##  4 Computer Systems Analyst IV      14
##  5 Database Administrator III       13
##  6 Software Test Engineer III       13
##  7 Account Representative IV        12
##  8 Budget/Accounting Analyst IV     12
##  9 Engineer IV                      12
## 10 Statistician II                  12
```

Customer Demographic table requires a proper Missing Value place holder. In `job_industry_category` there are string “n/a” values. Those should be proper NA values.

```
CustomerDemographic %>%
  count(job_industry_category, sort = T)
```

```
## # A tibble: 10 x 2
##    job_industry_category     n
##    <chr>                 <int>
##  1 Manufacturing           799
##  2 Financial Services      774
##  3 n/a                     656
##  4 Health                  602
##  5 Retail                  358
##  6 Property                267
##  7 IT                      223
##  8 Entertainment           136
##  9 Argiculture             113
## 10 Telecommunications       72
```

Some customers have missing last names.

```
CustomerDemographic %>%
  filter(is.na(last_name)) %>%
  head()
```

```
## # A tibble: 6 x 13
##   customer_id first_name last_name gender past_3_years_bik~ DOB        job_title
##         <dbl> <chr>      <chr>     <chr>              <dbl> <date>     <chr>
## 1           4 Talbot     <NA>      Male                  33 1961-10-03 <NA>
## 2          67 Vernon     <NA>      Male                  67 1960-06-14 Web Deve~
## 3         106 Glyn       <NA>      Male                  54 1966-07-03 Software~
## 4         139 Gar        <NA>      Male                   1 1964-07-28 Operator
## 5         197 Avis       <NA>      Female                32 1977-01-27 <NA>
## 6         211 Beitris    <NA>      Female                 6 1974-03-04 VP Marke~
## # ... with 6 more variables: job_industry_category <chr>, wealth_segment <chr>,
## #   deceased_indicator <chr>, default <chr>, owns_car <chr>, tenure <dbl>
```

There is a “default” column with many random values. Probably because of some bug in the data acquisition process.

```
CustomerDemographic %>%
  select(default) %>%
  slice_head(n = 10)
```

```
## # A tibble: 10 x 1
##    default
##    <chr>
##  1 "\"'"
##  2 "<script>alert('hi')</script>"
##  3 "43132"
##  4 "() { _; } >_[$($())] { touch /tmp/blns.shellshock2.fail; }"
##  5 "NIL"
##  6 "<U+00F0>µ <U+00F0> <U+00F0> <U+00F0>"
##  7 "â°â´âµâââ"
##  8 "(â¯Â°â¡Â°ï¼â¯ï¸µ â»ââ»)"
##  9 "0/0"
## 10 "<U+00F0>©<U+00F0>½"
```

Customer Address

New South Vales and NSW are probably the same state. The similar situation goes for VIC and Victoria.

```
CustomerAddress %>%
  count(state)
```

```
## # A tibble: 5 x 2
##   state              n
##   <chr>          <int>
## 1 New South Wales    86
## 2 NSW              2054
## 3 QLD               838
## 4 VIC               939
## 5 Victoria           82
```

Joins

These customers don't have their address information.

```
CustomerDemographic %>%
  anti_join(CustomerAddress, by = "customer_id")
```

```
## # A tibble: 4 x 13
##   customer_id first_name last_name gender past_3_years_bik~ DOB        job_title
##         <dbl> <chr>      <chr>     <chr>              <dbl> <date>     <chr>
## 1           3 Arlin      Dearle    Male                  61 1954-01-20 Recruiti~
## 2          10 Fiorenze   Birdall   Female                49 1988-10-11 Senior Q~
## 3          22 Deeanne    Durtnell  Female                79 1962-12-10 <NA>
## 4          23 Olav       Polak     Male                  43 1995-02-10 <NA>
## # ... with 6 more variables: job_industry_category <chr>, wealth_segment <chr>,
## #   deceased_indicator <chr>, default <chr>, owns_car <chr>, tenure <dbl>
```

There are also customers who have their address information but they don't exist in customer data.

```
CustomerAddress %>%
  anti_join(CustomerDemographic, by = "customer_id")
```

```
## # A tibble: 3 x 6
##   customer_id address                postcode state country   property_valuation
##         <dbl> <chr>                     <dbl> <chr> <chr>                  <dbl>
## 1        4001 87 Crescent Oaks Alley     2756 NSW   Australia                 10
## 2        4002 8194 Lien Street           4032 QLD   Australia                  7
## 3        4003 320 Acker Drive            2251 NSW   Australia                  7
```