Name : Nidhi Surti

KPMG Virtual Internship Project

Task 1 : Data Quality Assessment

The client provided KPMG with 3 datasets

1. Customer Demographic

2. Customer Addresses

3. Transaction Data in past 3 months

In [1]:
```
#Import the necessary libraries
import pandas as pd
```

In [2]:
```
#Reading the data
df = pd.ExcelFile("C:/Users/Dell/Downloads/KPMG_VI_New_raw_data_update_final.x
```

In [3]:
```
#Reading each file separately
Transactions = pd.read_excel(df, "Transactions")
NewCustomerList = pd.read_excel(df, "NewCustomerList")
CustomerDemographic = pd.read_excel(df, "CustomerDemographic")
CustomerAddress = pd.read_excel(df, "CustomerAddress")
```

In [4]:
```
#Exploring Transactions dataset
Transactions.head(5)
```

Out[4]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	U
0	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	pro
1	1	2	2950	2017-02-25 00:00:00	False	Approved	Solex	
2	2	3	3120	2017-05-21 00:00:00	True	Approved	Trek Bicycles	
3	3	37	402	2017-10-16 00:00:00	False	Approved	OHM Cycles	
4	4	88	3135	2017-08-31 00:00:00	False	Approved	Norco Bicycles	

5 rows × 26 columns

In [5]:
```
Transactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20001 entries, 0 to 20000
Data columns (total 26 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Note: The data and information in this document is reflective of a hypot
hetical situation and client. This document is to be used for KPMG Virtual In
ternship purposes only.   20001 non-null  object
 1   Unnamed: 1
20001 non-null  object
 2   Unnamed: 2
20001 non-null  object
 3   Unnamed: 3
20001 non-null  object
 4   Unnamed: 4
19641 non-null  object
 5   Unnamed: 5
20001 non-null  object
 6   Unnamed: 6
19804 non-null  object
 7   Unnamed: 7
19804 non-null  object
 8   Unnamed: 8
19804 non-null  object
 9   Unnamed: 9
19804 non-null  object
 10  Unnamed: 10
20001 non-null  object
 11  Unnamed: 11
19804 non-null  object
 12  Unnamed: 12
19804 non-null  object
 13  Unnamed: 13
0 non-null      float64
 14  Unnamed: 14
0 non-null      float64
 15  Unnamed: 15
0 non-null      float64
 16  Unnamed: 16
0 non-null      float64
 17  Unnamed: 17
0 non-null      float64
 18  Unnamed: 18
0 non-null      float64
 19  Unnamed: 19
0 non-null      float64
 20  Unnamed: 20
0 non-null      float64
 21  Unnamed: 21
0 non-null      float64
 22  Unnamed: 22
0 non-null      float64
 23  Unnamed: 23
0 non-null      float64
 24  Unnamed: 24
```

```
0 non-null      float64
 25  Unnamed: 25
0 non-null      float64
dtypes: float64(13), object(13)
memory usage: 4.0+ MB
```

In [6]:
```
#Using only the required columns
Transactions = Transactions.iloc[:,0:13]
Transactions.head()
```

Out[6]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	U
0	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	prc
1	1	2	2950	2017-02-25 00:00:00	False	Approved	Solex	
2	2	3	3120	2017-05-21 00:00:00	True	Approved	Trek Bicycles	
3	3	37	402	2017-10-16 00:00:00	False	Approved	OHM Cycles	
4	4	88	3135	2017-08-31 00:00:00	False	Approved	Norco Bicycles	

In [7]:
```
Transactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20001 entries, 0 to 20000
Data columns (total 13 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Note: The data and information in this document is reflective of a hypot
hetical situation and client. This document is to be used for KPMG Virtual In
ternship purposes only.   20001 non-null  object
 1   Unnamed: 1
20001 non-null  object
 2   Unnamed: 2
20001 non-null  object
 3   Unnamed: 3
20001 non-null  object
 4   Unnamed: 4
19641 non-null  object
 5   Unnamed: 5
20001 non-null  object
 6   Unnamed: 6
19804 non-null  object
 7   Unnamed: 7
19804 non-null  object
 8   Unnamed: 8
19804 non-null  object
 9   Unnamed: 9
19804 non-null  object
 10  Unnamed: 10
20001 non-null  object
 11  Unnamed: 11
19804 non-null  object
 12  Unnamed: 12
19804 non-null  object
dtypes: object(13)
memory usage: 2.0+ MB
```

In [8]:
```
#Checking the shape of the data
Transactions.shape
```

Out[8]:
```
(20001, 13)
```

In [9]:
```
#Check for the null values
Transactions.isnull().sum()
```

Out[9]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.       0
Unnamed: 1
0
Unnamed: 2
0
Unnamed: 3
0
Unnamed: 4
360
Unnamed: 5
0
Unnamed: 6
197
Unnamed: 7
197
Unnamed: 8
197
Unnamed: 9
197
Unnamed: 10
0
Unnamed: 11
197
Unnamed: 12
197
dtype: int64
```

There are missing values in 6 columns . They can be dropped or treated as the nature of analysis

In [10]:
```
#Checking for duplicate values
Transactions.duplicated().sum()
```

Out[10]:
```
0
```

There are no duplicate values so the data is unique

In [11]:
```
#Check for the uniqueness of each column
Transactions.nunique()
```

Out[11]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.     20001
Unnamed: 1
102
Unnamed: 2
3495
Unnamed: 3
365
Unnamed: 4
3
Unnamed: 5
3
Unnamed: 6
7
Unnamed: 7
5
Unnamed: 8
4
Unnamed: 9
4
Unnamed: 10
297
Unnamed: 11
104
Unnamed: 12
101
dtype: int64
```

In [12]:
```
#Check for the columns
Transactions.columns
```

Out[12]:
```
Index(['Note: The data and information in this document is reflective of a hy
pothetical situation and client. This document is to be used for KPMG Virtual
Internship purposes only. ',
       'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5',
       'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9', 'Unnamed: 10',
       'Unnamed: 11', 'Unnamed: 12'],
      dtype='object')
```

In [13]:
```
Transactions['Unnamed: 1'].value_counts()
```

Out[13]:
```
0             1378
3              354
1              311
35             268
38             267
              ...
16             136
8              136
100            130
47             121
product_id       1
Name: Unnamed: 1, Length: 102, dtype: int64
```

In [14]:
```
Transactions['Unnamed: 2'].value_counts()
```

Out[14]:
```
2476    14
1068    14
2183    14
3048    13
1913    13
        ..
1544     1
1325     1
872      1
2291     1
2328     1
Name: Unnamed: 2, Length: 3495, dtype: int64
```

In [15]:
```
Transactions['Unnamed: 3'].value_counts()
```

Out[15]:
```
2017-08-18 00:00:00    82
2017-02-14 00:00:00    82
2017-10-15 00:00:00    76
2017-01-31 00:00:00    73
2017-12-19 00:00:00    71
                       ..
2017-12-07 00:00:00    37
2017-03-29 00:00:00    36
2017-09-25 00:00:00    35
2017-10-19 00:00:00    32
transaction_date        1
Name: Unnamed: 3, Length: 365, dtype: int64
```

In [16]:
```
Transactions['Unnamed: 4'].value_counts()
```

Out[16]:
```
True            9829
False           9811
online_order       1
Name: Unnamed: 4, dtype: int64
```

In [17]:
```
Transactions['Unnamed: 5'].value_counts()
```

Out[17]:
```
Approved        19821
Cancelled         179
order_status        1
Name: Unnamed: 5, dtype: int64
```

In [18]:
```
Transactions['Unnamed: 6'].value_counts()
```

Out[18]:
```
Solex             4253
Giant Bicycles    3312
WeareA2B          3295
OHM Cycles        3043
Trek Bicycles     2990
Norco Bicycles    2910
brand                1
Name: Unnamed: 6, dtype: int64
```

In [19]:
```
Transactions['Unnamed: 7'].value_counts()
```

Out[19]:
```
Standard        14176
Road             3970
Touring          1234
Mountain          423
product_line        1
Name: Unnamed: 7, dtype: int64
```

In [20]:
```
Transactions['Unnamed: 8'].value_counts()
```

Out[20]:
```
medium           13826
high              3013
low               2964
product_class        1
Name: Unnamed: 8, dtype: int64
```

In [21]:
```
Transactions['Unnamed: 9'].value_counts()
```

Out[21]:
```
medium          12990
large            3976
small            2837
product_size        1
Name: Unnamed: 9, dtype: int64
```

In [22]:
```
Transactions['Unnamed: 10'].value_counts()
```

Out[22]:
```
2091.47    465
1403.50    396
71.49      274
1231.15    235
1890.39    233
           ...
875.99       1
877.44       1
880.30       1
883.91       1
1148.41      1
Name: Unnamed: 10, Length: 297, dtype: int64
```

In [23]:
```
Transactions['Unnamed: 11'].value_counts()
```

Out[23]:
```
388.92           465
954.82           396
53.62            274
161.6            235
260.14           233
                ...
206.35           114
standard_cost      1
667.4000244        1
312.7350159        1
270.2999878        1
Name: Unnamed: 11, Length: 104, dtype: int64
```

In [24]:
```
Transactions['Unnamed: 12'].value_counts()
```

Out[24]:
```
33879                    234
41064                    229
37823                    227
39880                    222
38216                    220
                        ...
42404                    168
41922                    166
37659                    163
34586                    162
product_first_sold_date    1
Name: Unnamed: 12, Length: 101, dtype: int64
```

In [25]:
```
Transactions['Unnamed: 3'].head(20)
```

Out[25]:
```
0        transaction_date
1     2017-02-25 00:00:00
2     2017-05-21 00:00:00
3     2017-10-16 00:00:00
4     2017-08-31 00:00:00
5     2017-10-01 00:00:00
6     2017-03-08 00:00:00
7     2017-04-21 00:00:00
8     2017-07-15 00:00:00
9     2017-08-10 00:00:00
10    2017-08-30 00:00:00
11    2017-01-17 00:00:00
12    2017-01-05 00:00:00
13    2017-02-26 00:00:00
14    2017-09-10 00:00:00
15    2017-06-11 00:00:00
16    2017-10-10 00:00:00
17    2017-04-03 00:00:00
18    2017-06-02 00:00:00
19    2017-04-06 00:00:00
Name: Unnamed: 3, dtype: object
```

The values in the Unnamed: 3 is columns are correct as it shows everything happening at the different day at the different time

Exploring the NewCustomerList dataset

In [26]:
```
NewCustomerList.head(5)
```

Out[26]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed
0	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_t
1	Chickie	Brister	Male	86	1957-07-12	Gene Manag
2	Morly	Genery	Male	69	1970-03-22	Structu Engin
3	Ardelis	Forrester	Female	10	1974-08-28 00:00:00	Senior C Account
4	Lucine	Stutt	Female	64	1979-01-28	Acco Representat

5 rows × 23 columns

In [27]:
```
NewCustomerList.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1001 entries, 0 to 1000
Data columns (total 23 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Note: The data and information in this document is reflective of a hypot
hetical situation and client. This document is to be used for KPMG Virtual In
ternship purposes only.   1001 non-null   object
 1   Unnamed: 1
972 non-null    object
 2   Unnamed: 2
1001 non-null   object
 3   Unnamed: 3
1001 non-null   object
 4   Unnamed: 4
984 non-null    object
 5   Unnamed: 5
895 non-null    object
 6   Unnamed: 6
836 non-null    object
 7   Unnamed: 7
1001 non-null   object
 8   Unnamed: 8
1001 non-null   object
 9   Unnamed: 9
1001 non-null   object
 10  Unnamed: 10
1001 non-null   object
 11  Unnamed: 11
1001 non-null   object
 12  Unnamed: 12
1001 non-null   object
 13  Unnamed: 13
1001 non-null   object
 14  Unnamed: 14
1001 non-null   object
 15  Unnamed: 15
1001 non-null   object
 16  Unnamed: 16
1000 non-null   float64
 17  Unnamed: 17
1000 non-null   float64
 18  Unnamed: 18
1000 non-null   float64
 19  Unnamed: 19
1000 non-null   float64
 20  Unnamed: 20
1000 non-null   float64
 21  Unnamed: 21
1001 non-null   object
 22  Unnamed: 22
1001 non-null   object
dtypes: float64(5), object(18)
memory usage: 180.0+ KB
```

In [28]:
```
#Dropping the float values
NewCustomerList.drop(['Unnamed: 16','Unnamed: 17','Unnamed: 18','Unnamed: 19',
```

In [29]:
```
#Checking the shape of the dataset
NewCustomerList.shape
```

Out[29]:
```
(1001, 18)
```

In [30]:
```
#Checking for null values
NewCustomerList.isnull().sum()
```

Out[30]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.       0
Unnamed: 1
29
Unnamed: 2
0
Unnamed: 3
0
Unnamed: 4
17
Unnamed: 5
106
Unnamed: 6
165
Unnamed: 7
0
Unnamed: 8
0
Unnamed: 9
0
Unnamed: 10
0
Unnamed: 11
0
Unnamed: 12
0
Unnamed: 13
0
Unnamed: 14
0
Unnamed: 15
0
Unnamed: 21
0
Unnamed: 22
0
dtype: int64
```

There are missing values in 4 columns. They can be dropped or treated according to the nature of analysis

In [31]:
```
#Checking for duplicate values
NewCustomerList.duplicated().sum()
```

Out[31]:
```
0
```

There are no duplicate values

In [32]:
```
#Checking for uniquess of each column
NewCustomerList.nunique()
```

Out[32]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.      941
Unnamed: 1
962
Unnamed: 2
4
Unnamed: 3
101
Unnamed: 4
962
Unnamed: 5
185
Unnamed: 6
10
Unnamed: 7
4
Unnamed: 8
2
Unnamed: 9
3
Unnamed: 10
24
Unnamed: 11
1001
Unnamed: 12
523
Unnamed: 13
4
Unnamed: 14
2
Unnamed: 15
17
Unnamed: 21
325
Unnamed: 22
325
dtype: int64
```

In [33]:
```
#Exploring the columns
NewCustomerList.columns
```

Out[33]:
```
Index(['Note: The data and information in this document is reflective of a hy
pothetical situation and client. This document is to be used for KPMG Virtual
Internship purposes only. ',
       'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5',
       'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9', 'Unnamed: 10',
       'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13', 'Unnamed: 14',
       'Unnamed: 15', 'Unnamed: 21', 'Unnamed: 22'],
      dtype='object')
```

In [34]:
```
NewCustomerList['Unnamed: 1'].value_counts()
```

Out[34]:
```
Borsi         2
Hallt         2
Eade          2
Shoesmith     2
Sissel        2
             ..
Clee          1
Fraschetti    1
Dowyer        1
Terlinden     1
Roseman       1
Name: Unnamed: 1, Length: 962, dtype: int64
```

In [35]:
```
NewCustomerList['Unnamed: 2'].value_counts()
```

Out[35]:
```
Female    513
Male      470
U          17
gender      1
Name: Unnamed: 2, dtype: int64
```

There are 17 unknown/unspecified gender

In [36]:
```
NewCustomerList['Unnamed: 3'].value_counts()
```

Out[36]:
```
60                                    20
59                                    18
70                                    17
42                                    17
37                                    16
                                      ..
9                                      5
92                                     5
85                                     4
20                                     3
past_3_years_bike_related_purchases    1
Name: Unnamed: 3, Length: 101, dtype: int64
```

In [37]:
```
NewCustomerList['Unnamed: 4'].value_counts()
```

Out[37]:
```
1951-11-28            2
1979-07-28            2
1961-07-31            2
1941-07-21            2
1959-09-18            2
                     ..
1995-12-09            1
1954-10-19            1
1974-06-08 00:00:00   1
1965-04-22            1
1951-09-16            1
Name: Unnamed: 4, Length: 962, dtype: int64
```

In [38]:
```
NewCustomerList['Unnamed: 5'].value_counts()
```

Out[38]:
```
Associate Professor             15
Software Consultant             14
Environmental Tech              14
Chief Design Engineer           13
VP Sales                        12
                                ..
Computer Systems Analyst III     1
Database Administrator IV        1
Computer Systems Analyst II      1
Health Coach I                   1
Safety Technician IV             1
Name: Unnamed: 5, Length: 185, dtype: int64
```

In [39]:
```
NewCustomerList['Unnamed: 6'].value_counts()
```

Out[39]:
```
Financial Services       203
Manufacturing            199
Health                   152
Retail                    78
Property                  64
IT                        51
Entertainment             37
Argiculture               26
Telecommunications        25
job_industry_category      1
Name: Unnamed: 6, dtype: int64
```

In [40]:
```
NewCustomerList['Unnamed: 7'].value_counts()
```

Out[40]:
```
Mass Customer        508
High Net Worth       251
Affluent Customer    241
wealth_segment         1
Name: Unnamed: 7, dtype: int64
```

In [41]:
```
NewCustomerList['Unnamed: 8'].value_counts()
```

Out[41]:
```
N                    1000
deceased_indicator      1
Name: Unnamed: 8, dtype: int64
```

In [42]:
```
NewCustomerList['Unnamed: 9'].value_counts()
```

Out[42]:
```
No          507
Yes         493
owns_car      1
Name: Unnamed: 9, dtype: int64
```

In [43]:
```
NewCustomerList['Unnamed: 10'].value_counts()
```

Out[43]:
```
9         79
13        74
11        68
10        63
12        61
5         60
7         60
17        59
15        58
8         55
14        54
16        49
6         45
4         36
18        36
19        34
3         26
21        24
20        22
2         15
22        12
1          8
0          2
tenure     1
Name: Unnamed: 10, dtype: int64
```

In [44]:
```
NewCustomerList['Unnamed: 11'].value_counts()
```

Out[44]:
```
4 Manufacturers Crossing    1
19 Debs Parkway             1
38407 Sutteridge Circle     1
56334 Vera Crossing         1
77785 Veith Lane            1
                           ..
266 Lakewood Terrace        1
62 Dryden Junction          1
1 Raven Way                 1
432 Ronald Regan Court      1
73 Riverside Trail          1
Name: Unnamed: 11, Length: 1001, dtype: int64
```

In [45]:
```
NewCustomerList['Unnamed: 12'].value_counts()
```

Out[45]:
```
2232    9
2145    9
3977    7
3029    7
2168    7
       ..
2539    1
2518    1
2422    1
3930    1
4183    1
Name: Unnamed: 12, Length: 523, dtype: int64
```

In [46]:
```
NewCustomerList['Unnamed: 13'].value_counts()
```

Out[46]:
```
NSW      506
VIC      266
QLD      228
state      1
Name: Unnamed: 13, dtype: int64
```

In [47]:
```
NewCustomerList['Unnamed: 14'].value_counts()
```

Out[47]:
```
Australia    1000
country         1
Name: Unnamed: 14, dtype: int64
```

In [48]:
```
NewCustomerList['Unnamed: 15'].value_counts()
```

Out[48]:
```
9                    173
8                    161
7                    136
10                   116
6                     69
11                    61
5                     57
4                     53
3                     51
12                    46
2                     42
1                     30
7                      2
property_valuation     1
11                     1
8                      1
6                      1
Name: Unnamed: 15, dtype: int64
```

In [49]:
```
NewCustomerList['Unnamed: 21'].value_counts()
```

Out[49]:
```
760    13
259    12
386     9
133     9
455     9
       ..
355     1
574     1
12      1
641     1
249     1
Name: Unnamed: 21, Length: 325, dtype: int64
```

In [50]:
```
NewCustomerList['Unnamed: 22'].value_counts()
```

Out[50]:
```
0.6375    13
1.0625    12
0.8925     9
1.2375     9
0.5        9
          ..
0.867      1
Value      1
0.476      1
0.63       1
0.94       1
Name: Unnamed: 22, Length: 325, dtype: int64
```

Exploring CustomerDemographic Dataset

In [51]:
```
CustomerDemographic.head()
```

Out[51]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0	customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB
1	1	Laraine	Medendorp	F	93	1953-10-12 00:00:00
2	2	Eli	Bockman	Male	81	1980-12-16 00:00:00
3	3	Arlin	Dearle	Male	61	1954-01-20 00:00:00
4	4	Talbot	NaN	Male	33	1961-10-03 00:00:00

5 rows × 26 columns

In [52]:
```
CustomerDemographic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4001 entries, 0 to 4000
Data columns (total 26 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Note: The data and information in this document is reflective of a hypot
hetical situation and client. This document is to be used for KPMG Virtual In
ternship purposes only.   4001 non-null   object
 1   Unnamed: 1
4001 non-null   object
 2   Unnamed: 2
3876 non-null   object
 3   Unnamed: 3
4001 non-null   object
 4   Unnamed: 4
4001 non-null   object
 5   Unnamed: 5
3914 non-null   object
 6   Unnamed: 6
3495 non-null   object
 7   Unnamed: 7
3345 non-null   object
 8   Unnamed: 8
4001 non-null   object
 9   Unnamed: 9
4001 non-null   object
 10  Unnamed: 10
3699 non-null   object
 11  Unnamed: 11
4001 non-null   object
 12  Unnamed: 12
3914 non-null   object
 13  Unnamed: 13
0 non-null      float64
 14  Unnamed: 14
0 non-null      float64
 15  Unnamed: 15
0 non-null      float64
 16  Unnamed: 16
0 non-null      float64
 17  Unnamed: 17
0 non-null      float64
 18  Unnamed: 18
0 non-null      float64
 19  Unnamed: 19
0 non-null      float64
 20  Unnamed: 20
0 non-null      float64
 21  Unnamed: 21
0 non-null      float64
 22  Unnamed: 22
0 non-null      float64
 23  Unnamed: 23
0 non-null      float64
 24  Unnamed: 24
```

```
0 non-null      float64
 25  Unnamed: 25
0 non-null      float64
dtypes: float64(13), object(13)
memory usage: 812.8+ KB
```

In [53]:
```
#Checking for null values
CustomerDemographic.isnull().sum()
```

Out[53]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.        0
Unnamed: 1
0
Unnamed: 2
125
Unnamed: 3
0
Unnamed: 4
0
Unnamed: 5
87
Unnamed: 6
506
Unnamed: 7
656
Unnamed: 8
0
Unnamed: 9
0
Unnamed: 10
302
Unnamed: 11
0
Unnamed: 12
87
Unnamed: 13
4001
Unnamed: 14
4001
Unnamed: 15
4001
Unnamed: 16
4001
Unnamed: 17
4001
Unnamed: 18
4001
Unnamed: 19
4001
Unnamed: 20
4001
Unnamed: 21
4001
Unnamed: 22
4001
Unnamed: 23
4001
Unnamed: 24
4001
Unnamed: 25
4001
dtype: int64
```

There are missing values in 7 columns. They can be dropped or treated according to the nature of analysis

In [54]:
```
#Checking for duplicate data
CustomerDemographic.duplicated().sum()
```

Out[54]:
```
0
```

There are no duplicate values

In [56]:
```
CustomerDemographic.columns
```

Out[56]:
```
Index(['Note: The data and information in this document is reflective of a hy
pothetical situation and client. This document is to be used for KPMG Virtual
Internship purposes only. ',
       'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5',
       'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9', 'Unnamed: 10',
       'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13', 'Unnamed: 14',
       'Unnamed: 15', 'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18',
       'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22',
       'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25'],
      dtype='object')
```

In [57]:
```
#Checking for uniqueness of each column
CustomerDemographic.nunique()
```

Out[57]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.     4001
Unnamed: 1
3140
Unnamed: 2
3726
Unnamed: 3
7
Unnamed: 4
101
Unnamed: 5
3449
Unnamed: 6
196
Unnamed: 7
10
Unnamed: 8
4
Unnamed: 9
3
Unnamed: 10
91
Unnamed: 11
3
Unnamed: 12
23
Unnamed: 13
0
Unnamed: 14
0
Unnamed: 15
0
Unnamed: 16
0
Unnamed: 17
0
Unnamed: 18
0
Unnamed: 19
0
Unnamed: 20
0
Unnamed: 21
0
Unnamed: 22
0
Unnamed: 23
0
Unnamed: 24
0
Unnamed: 25
0
dtype: int64
```

In [58]:
```
CustomerDemographic['Unnamed: 1'].value_counts()
```

Out[58]:
```
Tobe      5
Max       5
Timmie    5
Kippy     4
Kim       4
         ..
Vivien    1
Lilli     1
Myles     1
Reeva     1
Easter    1
Name: Unnamed: 1, Length: 3140, dtype: int64
```

In [59]:
```
CustomerDemographic['Unnamed: 2'].value_counts()
```

Out[59]:
```
Pristnor    3
Ramsdell    3
Lamming     2
Cator       2
Dahlman     2
           ..
Melonby     1
Peasey      1
Halfhyde    1
Kegley      1
Roseman     1
Name: Unnamed: 2, Length: 3726, dtype: int64
```

In [60]:
```
CustomerDemographic['Unnamed: 3'].value_counts()
```

Out[60]:
```
Female    2037
Male      1872
U           88
Femal        1
gender       1
M            1
F            1
Name: Unnamed: 3, dtype: int64
```

Certain categories are not correctly titled. The names in these categories are re-named.

In [61]:
```
#Re-naming the categories
CustomerDemographic['Unnamed: 3'] = CustomerDemographic['Unnamed: 3'].replace(
```

In [62]:
```
CustomerDemographic['Unnamed: 3'].value_counts()
```

Out[62]:
```
Female         2039
Male           1873
Unspecified      88
gender            1
Name: Unnamed: 3, dtype: int64
```

In [63]:
```
CustomerDemographic['Unnamed: 4'].value_counts()
```

Out[63]:
```
19                                    56
16                                    56
67                                    54
20                                    54
2                                     50
                                      ..
85                                    27
95                                    27
86                                    27
92                                    24
past_3_years_bike_related_purchases    1
Name: Unnamed: 4, Length: 101, dtype: int64
```

In [64]:
```
CustomerDemographic['Unnamed: 5'].value_counts()
```

Out[64]:
```
1978-01-30    7
1978-08-19    4
1964-07-08    4
1962-12-17    4
1976-09-25    4
             ..
1958-06-23    1
1984-03-25    1
1976-10-24    1
1960-04-25    1
1976-04-29    1
Name: Unnamed: 5, Length: 3449, dtype: int64
```

In [65]:
```
CustomerDemographic['Unnamed: 6'].value_counts()
```

Out[65]:
```
Business Systems Development Analyst    45
Social Worker                           44
Tax Accountant                          44
Internal Auditor                        42
Legal Assistant                         41
                                        ..
Health Coach I                           3
Health Coach III                         3
Research Assistant III                   3
Developer I                              1
job_title                                1
Name: Unnamed: 6, Length: 196, dtype: int64
```

In [66]:
```
CustomerDemographic['Unnamed: 7'].value_counts()
```

Out[66]:
```
Manufacturing            799
Financial Services       774
Health                   602
Retail                   358
Property                 267
IT                       223
Entertainment            136
Argiculture              113
Telecommunications        72
job_industry_category      1
Name: Unnamed: 7, dtype: int64
```

In [67]:
```
CustomerDemographic['Unnamed: 8'].value_counts()
```

Out[67]:
```
Mass Customer        2000
High Net Worth       1021
Affluent Customer     979
wealth_segment          1
Name: Unnamed: 8, dtype: int64
```

In [68]:
```
CustomerDemographic['Unnamed: 9'].value_counts()
```

Out[68]:
```
N                    3998
Y                       2
deceased_indicator      1
Name: Unnamed: 9, dtype: int64
```

In [69]:
```
CustomerDemographic['Unnamed: 10'].value_counts()
```

Out[69]:
```
100                                              113
1                                                112
-1                                               111
-100                                              99
â°â´âµâêâ                                        53
                                                 ...
/dev/null; touch /tmp/blns.fail ; echo            30
âªâªtestâª                                        29
ì¸ëë°í ë¥´                                       27
,ãã»:*:ã»ãâ( â» Ï â» )ãã»:*:ã»ãâ                  25
default                                           1
Name: Unnamed: 10, Length: 91, dtype: int64
```

In [70]:
```
CustomerDemographic = CustomerDemographic.drop('Unnamed: 10', axis=1)
```

The values are inconsistent therefore dropping the column

In [71]:
```
CustomerDemographic.head(5)
```

Out[71]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0	customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB
1	1	Laraine	Medendorp	Female	93	1953-10-12 00:00:00
2	2	Eli	Bockman	Male	81	1980-12-16 00:00:00
3	3	Arlin	Dearle	Male	61	1954-01-20 00:00:00
4	4	Talbot	NaN	Male	33	1961-10-03 00:00:00

5 rows × 25 columns

In [72]:
```
CustomerDemographic['Unnamed: 11'].value_counts()
```

Out[72]:
```
Yes         2024
No          1976
owns_car       1
Name: Unnamed: 11, dtype: int64
```

In [73]:
```
CustomerDemographic['Unnamed: 12'].value_counts()
```

Out[73]:
```
7          235
5          228
11         221
10         218
16         215
8          211
18         208
12         202
9          200
14         200
6          192
4          191
13         191
17         182
15         179
1          166
3          160
19         159
2          150
20          96
22          55
21          54
tenure       1
Name: Unnamed: 12, dtype: int64
```

In [74]:
```
CustomerDemographic['Unnamed: 13'].value_counts()
```

Out[74]:
```
Series([], Name: Unnamed: 13, dtype: int64)
```

In [75]:
```
CustomerDemographic['Unnamed: 14'].value_counts()
```

Out[75]:
```
Series([], Name: Unnamed: 14, dtype: int64)
```

In [76]:
```
CustomerDemographic['Unnamed: 15'].value_counts()
```

Out[76]:
```
Series([], Name: Unnamed: 15, dtype: int64)
```

In [77]:
```
CustomerDemographic['Unnamed: 16'].value_counts()
```

Out[77]:
```
Series([], Name: Unnamed: 16, dtype: int64)
```

In [78]:
```
CustomerDemographic['Unnamed: 17'].value_counts()
```

Out[78]:
```
Series([], Name: Unnamed: 17, dtype: int64)
```

In [79]:
```
CustomerDemographic['Unnamed: 18'].value_counts()
```

Out[79]:
```
Series([], Name: Unnamed: 18, dtype: int64)
```

In [80]:
```
CustomerDemographic['Unnamed: 19'].value_counts()
```

Out[80]:
```
Series([], Name: Unnamed: 19, dtype: int64)
```

In [81]:
```
CustomerDemographic['Unnamed: 20'].value_counts()
```

Out[81]:
```
Series([], Name: Unnamed: 20, dtype: int64)
```

In [82]:
```
CustomerDemographic['Unnamed: 21'].value_counts()
```

Out[82]:
```
Series([], Name: Unnamed: 21, dtype: int64)
```

In [83]:
```
CustomerDemographic['Unnamed: 22'].value_counts()
```

Out[83]:
```
Series([], Name: Unnamed: 22, dtype: int64)
```

In [84]:
```
CustomerDemographic['Unnamed: 23'].value_counts()
```

Out[84]:
```
Series([], Name: Unnamed: 23, dtype: int64)
```

In [85]:
```
CustomerDemographic['Unnamed: 24'].value_counts()
```

Out[85]:
```
Series([], Name: Unnamed: 24, dtype: int64)
```

In [86]:
```
CustomerDemographic['Unnamed: 25'].value_counts()
```

Out[86]:
```
Series([], Name: Unnamed: 25, dtype: int64)
```

Exploring Customer Address Dataset

In [87]:
```
CustomerAddress.head(5)
```

Out[87]:

	Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unna
0	customer_id	address	postcode	state	country	property_valuation	NaN	
1	1	060 Morning Avenue	2016	New South Wales	Australia	10	NaN	
2	2	6 Meadow Vale Court	2153	New South Wales	Australia	10	NaN	
3	4	0 Holy Cross Court	4211	QLD	Australia	9	NaN	
4	5	17979 Del Mar Point	2448	New South Wales	Australia	4	NaN	

5 rows × 26 columns

In [88]:
```
CustomerAddress.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 26 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Note: The data and information in this document is reflective of a hypot
hetical situation and client. This document is to be used for KPMG Virtual In
ternship purposes only.   4000 non-null   object
 1   Unnamed: 1
4000 non-null   object
 2   Unnamed: 2
4000 non-null   object
 3   Unnamed: 3
4000 non-null   object
 4   Unnamed: 4
4000 non-null   object
 5   Unnamed: 5
4000 non-null   object
 6   Unnamed: 6
0 non-null      float64
 7   Unnamed: 7
0 non-null      float64
 8   Unnamed: 8
0 non-null      float64
 9   Unnamed: 9
0 non-null      float64
 10  Unnamed: 10
0 non-null      float64
 11  Unnamed: 11
0 non-null      float64
 12  Unnamed: 12
0 non-null      float64
 13  Unnamed: 13
0 non-null      float64
 14  Unnamed: 14
0 non-null      float64
 15  Unnamed: 15
0 non-null      float64
 16  Unnamed: 16
0 non-null      float64
 17  Unnamed: 17
0 non-null      float64
 18  Unnamed: 18
0 non-null      float64
 19  Unnamed: 19
0 non-null      float64
 20  Unnamed: 20
0 non-null      float64
 21  Unnamed: 21
0 non-null      float64
 22  Unnamed: 22
0 non-null      float64
 23  Unnamed: 23
0 non-null      float64
 24  Unnamed: 24
```

```
0 non-null      float64
 25  Unnamed: 25
0 non-null      float64
dtypes: float64(20), object(6)
memory usage: 812.6+ KB
```

In [89]:
```
#Checking for null values.
CustomerAddress.isnull().sum()
```

Out[89]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.        0
Unnamed: 1
0
Unnamed: 2
0
Unnamed: 3
0
Unnamed: 4
0
Unnamed: 5
0
Unnamed: 6
4000
Unnamed: 7
4000
Unnamed: 8
4000
Unnamed: 9
4000
Unnamed: 10
4000
Unnamed: 11
4000
Unnamed: 12
4000
Unnamed: 13
4000
Unnamed: 14
4000
Unnamed: 15
4000
Unnamed: 16
4000
Unnamed: 17
4000
Unnamed: 18
4000
Unnamed: 19
4000
Unnamed: 20
4000
Unnamed: 21
4000
Unnamed: 22
4000
Unnamed: 23
4000
Unnamed: 24
4000
Unnamed: 25
4000
dtype: int64
```

In [90]:
```
CustomerAddressN = CustomerAddress.dropna()
```

In [91]:
```
CustomerAddressN.head()
```

Out[91]:

Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	U

0 rows × 26 columns

In [92]:
```
#Checking for null values.
CustomerAddressN.isnull().sum()
```

Out[92]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.     0.0
Unnamed: 1
0.0
Unnamed: 2
0.0
Unnamed: 3
0.0
Unnamed: 4
0.0
Unnamed: 5
0.0
Unnamed: 6
0.0
Unnamed: 7
0.0
Unnamed: 8
0.0
Unnamed: 9
0.0
Unnamed: 10
0.0
Unnamed: 11
0.0
Unnamed: 12
0.0
Unnamed: 13
0.0
Unnamed: 14
0.0
Unnamed: 15
0.0
Unnamed: 16
0.0
Unnamed: 17
0.0
Unnamed: 18
0.0
Unnamed: 19
0.0
Unnamed: 20
0.0
Unnamed: 21
0.0
Unnamed: 22
0.0
Unnamed: 23
0.0
Unnamed: 24
0.0
Unnamed: 25
0.0
dtype: float64
```

There are no null values

In [93]:
```
#Checking for duplicate values
CustomerAddressN.duplicated().sum()
```

Out[93]:
```
0
```

There is no duplicated values

In [94]:
```
#Checking for uniqueness of each column
CustomerAddressN.nunique()
```

Out[94]:
```
Note: The data and information in this document is reflective of a hypothetic
al situation and client. This document is to be used for KPMG Virtual Interns
hip purposes only.     0
Unnamed: 1
0
Unnamed: 2
0
Unnamed: 3
0
Unnamed: 4
0
Unnamed: 5
0
Unnamed: 6
0
Unnamed: 7
0
Unnamed: 8
0
Unnamed: 9
0
Unnamed: 10
0
Unnamed: 11
0
Unnamed: 12
0
Unnamed: 13
0
Unnamed: 14
0
Unnamed: 15
0
Unnamed: 16
0
Unnamed: 17
0
Unnamed: 18
0
Unnamed: 19
0
Unnamed: 20
0
Unnamed: 21
0
Unnamed: 22
0
Unnamed: 23
0
Unnamed: 24
0
Unnamed: 25
0
dtype: int64
```

Exploring the columns

In [96]:
```
CustomerAddressN.columns
```

Out[96]:
```
Index(['Note: The data and information in this document is reflective of a hy
pothetical situation and client. This document is to be used for KPMG Virtual
Internship purposes only. ',
       'Unnamed: 1', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4', 'Unnamed: 5',
       'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9', 'Unnamed: 10',
       'Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13', 'Unnamed: 14',
       'Unnamed: 15', 'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18',
       'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22',
       'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25'],
      dtype='object')
```

In [99]:
```
CustomerAddress['Unnamed: 1'].value_counts()
```

Out[99]:
```
3 Mariners Cove Terrace    2
3 Talisman Place           2
64 Macpherson Junction     2
614 Shopko Trail           1
56145 Porter Lane          1
                          ..
59231 Claremont Place      1
78 Rockefeller Park        1
294 Lawn Junction          1
18939 Upham Hill           1
005 Bunker Hill Lane       1
Name: Unnamed: 1, Length: 3997, dtype: int64
```

In [100]:
```
CustomerAddress['Unnamed: 2'].value_counts()
```

Out[100]:
```
2170    31
2155    30
2145    30
2153    29
2770    26
        ..
2400     1
3144     1
2257     1
2429     1
2794     1
Name: Unnamed: 2, Length: 874, dtype: int64
```

In [101]:
```
CustomerAddress['Unnamed: 3'].value_counts()
```

Out[101]:
```
NSW                2054
VIC                 939
QLD                 838
New South Wales      86
Victoria             82
state                 1
Name: Unnamed: 3, dtype: int64
```

In [102]:
```
CustomerAddress['Unnamed: 4'].value_counts()
```

Out[102]:
```
Australia    3999
country         1
Name: Unnamed: 4, dtype: int64
```

In [103]:
```
CustomerAddress['Unnamed: 5'].value_counts()
```

Out[103]:
```
9                    647
8                    646
10                   577
7                    493
11                   281
6                    238
5                    225
4                    214
12                   195
3                    186
1                    154
2                    143
property_valuation     1
Name: Unnamed: 5, dtype: int64
```

In [104]:
```
CustomerAddress['Unnamed: 6'].value_counts()
```

Out[104]:
```
Series([], Name: Unnamed: 6, dtype: int64)
```

In [105]:
```
CustomerAddress['Unnamed: 7'].value_counts()
```

Out[105]:
```
Series([], Name: Unnamed: 7, dtype: int64)
```

In [106]:
```
CustomerAddress['Unnamed: 8'].value_counts()
```

Out[106]:
```
Series([], Name: Unnamed: 8, dtype: int64)
```

In [107]:
```
CustomerAddress['Unnamed: 9'].value_counts()
```

Out[107]:
```
Series([], Name: Unnamed: 9, dtype: int64)
```

In [108]:
```
CustomerAddress['Unnamed: 10'].value_counts()
```

Out[108]:
```
Series([], Name: Unnamed: 10, dtype: int64)
```

In [109]:
```
CustomerAddress['Unnamed: 11'].value_counts()
```

Out[109]:
```
Series([], Name: Unnamed: 11, dtype: int64)
```

In [110]:
```
CustomerAddress['Unnamed: 12'].value_counts()
```

Out[110]:
```
Series([], Name: Unnamed: 12, dtype: int64)
```

```
In [111]: CustomerAddress['Unnamed: 13'].value_counts()
Out[111]: Series([], Name: Unnamed: 13, dtype: int64)

In [112]: CustomerAddress['Unnamed: 14'].value_counts()
Out[112]: Series([], Name: Unnamed: 14, dtype: int64)

In [113]: CustomerAddress['Unnamed: 15'].value_counts()
Out[113]: Series([], Name: Unnamed: 15, dtype: int64)

In [114]: CustomerAddress['Unnamed: 16'].value_counts()
Out[114]: Series([], Name: Unnamed: 16, dtype: int64)

In [115]: CustomerAddress['Unnamed: 17'].value_counts()
Out[115]: Series([], Name: Unnamed: 17, dtype: int64)

In [116]: CustomerAddress['Unnamed: 18'].value_counts()
Out[116]: Series([], Name: Unnamed: 18, dtype: int64)

In [117]: CustomerAddress['Unnamed: 19'].value_counts()
Out[117]: Series([], Name: Unnamed: 19, dtype: int64)

In [118]: CustomerAddress['Unnamed: 20'].value_counts()
Out[118]: Series([], Name: Unnamed: 20, dtype: int64)

In [119]: CustomerAddress['Unnamed: 21'].value_counts()
Out[119]: Series([], Name: Unnamed: 21, dtype: int64)

In [120]: CustomerAddress['Unnamed: 22'].value_counts()
Out[120]: Series([], Name: Unnamed: 22, dtype: int64)

In [121]: CustomerAddress['Unnamed: 23'].value_counts()
Out[121]: Series([], Name: Unnamed: 23, dtype: int64)

In [122]: CustomerAddress['Unnamed: 24'].value_counts()
Out[122]: Series([], Name: Unnamed: 24, dtype: int64)

In [123]: CustomerAddress['Unnamed: 25'].value_counts()
Out[123]: Series([], Name: Unnamed: 25, dtype: int64)
```

All the columns appear to be consistent and correct information