KPMG VIRTUAL INTERNSHIP PROJECT

In [1]:
```
#TASK: 1 - Data Quality Assessment
```

Assessment of data quality and complete in preparation for analysis. The client provided KPMG with 3 datasets: 1.Customer Demographic 2.Customer Addresses 3.Transactions data in the past 3 months

In [2]:
```
# Importing the required libraries
import pandas as pd
import numpy as np
```

Reading the data

In [3]:
```
data = pd.ExcelFile("KPMG1.xlsx")
```

Reading each file separately

In [4]:
```
Transactions = pd.read_excel(data, 'Transactions')
NewCustomerList = pd.read_excel(data, 'NewCustomerList')
CustomerDemographic = pd.read_excel(data, 'CustomerDemographic')
CustomerAddress = pd.read_excel(data, 'CustomerAddress')
```

```
C:\Users\ELCOT\AppData\Local\Temp\ipykernel_8212\3814060680.py:2: FutureWarning: Inferring datetime64[ns] from data containing strings is deprecated and will be removed in a future version. To retain the old behavior explicitly pass Series(data, dtype=datetime64[ns])
  NewCustomerList = pd.read_excel(data, 'NewCustomerList')
C:\Users\ELCOT\AppData\Local\Temp\ipykernel_8212\3814060680.py:3: FutureWarning: Inferring datetime64[ns] from data containing strings is deprecated and will be removed in a future version. To retain the old behavior explicitly pass Series(data, dtype=datetime64[ns])
  CustomerDemographic = pd.read_excel(data, 'CustomerDemographic')
```

Exploring Transactions Data Set

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

Out[5]:

	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	product_line	product_class	product_size	list_price	standard_cost	product_first_sold_date
0	1	2	2950	2017-02-25	0.0	Approved	Solex	Standard	medium	medium	71.49	53.62	41245.0
1	2	3	3120	2017-05-21	1.0	Approved	Trek Bicycles	Standard	medium	large	2091.47	388.92	41701.0
2	3	37	402	2017-10-16	0.0	Approved	OHM Cycles	Standard	low	medium	1793.43	248.82	36361.0
3	4	88	3135	2017-08-31	0.0	Approved	Norco Bicycles	Standard	medium	medium	1198.46	381.10	36145.0
4	5	78	787	2017-10-01	1.0	Approved	Giant Bicycles	Standard	medium	large	1765.30	709.48	42226.0

In [6]:
```
Transactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20000 entries, 0 to 19999
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   transaction_id           20000 non-null  int64
 1   product_id               20000 non-null  int64
 2   customer_id              20000 non-null  int64
 3   transaction_date         20000 non-null  datetime64[ns]
 4   online_order             19640 non-null  float64
 5   order_status             20000 non-null  object
 6   brand                    19803 non-null  object
 7   product_line             19803 non-null  object
 8   product_class            19803 non-null  object
 9   product_size             19803 non-null  object
 10  list_price               20000 non-null  float64
 11  standard_cost            19803 non-null  float64
 12  product_first_sold_date  19803 non-null  float64
dtypes: datetime64[ns](1), float64(4), int64(3), object(5)
memory usage: 2.0+ MB
```

In [7]:
```
#Using only the required columns
Transactions = Transactions.iloc[:, 0:13]
Transactions.head()
```

Out[7]:

	transaction_id	product_id	customer_id	transaction_date	online_order	order_status	brand	product_line	product_class	product_size	list_price	standard_cost	product_first_sold_date
0	1	2	2950	2017-02-25	0.0	Approved	Solex	Standard	medium	medium	71.49	53.62	41245.0
1	2	3	3120	2017-05-21	1.0	Approved	Trek Bicycles	Standard	medium	large	2091.47	388.92	41701.0
2	3	37	402	2017-10-16	0.0	Approved	OHM Cycles	Standard	low	medium	1793.43	248.82	36361.0
3	4	88	3135	2017-08-31	0.0	Approved	Norco Bicycles	Standard	medium	medium	1198.46	381.10	36145.0
4	5	78	787	2017-10-01	1.0	Approved	Giant Bicycles	Standard	medium	large	1765.30	709.48	42226.0

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

Out[8]: (20000, 13)

In [9]:
```
#Checking for duplicate values
Transactions.duplicated().sum()
```

Out[9]: 0

There are no duplicate values, so the data is unique.

In [10]:
```
#check for uniqueness of each column
Transactions.nunique()
```

Out[10]:
```
transaction_id             20000
product_id                   101
customer_id                 3494
transaction_date             364
online_order                   2
order_status                   2
brand                          6
product_line                   4
product_class                  3
product_size                   3
list_price                   296
standard_cost                103
product_first_sold_date      100
dtype: int64
```

In [11]:
```
Transactions.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[11]:
```
transaction_id               0
product_id                   0
customer_id                  0
transaction_date             0
online_order               360
order_status                 0
brand                      197
product_line               197
product_class              197
product_size               197
list_price                   0
standard_cost              197
product_first_sold_date    197
dtype: int64
```

In [12]:
```
#There are 7 missing values in columns. They can be dropped or nature analysis
```

Exploring the columns

In [13]:
```
Transactions.columns
```

Out[13]:
```
Index(['transaction_id', 'product_id', 'customer_id', 'transaction_date',
       'online_order', 'order_status', 'brand', 'product_line',
       'product_class', 'product_size', 'list_price', 'standard_cost',
       'product_first_sold_date'],
      dtype='object')
```

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

Out[14]:
```
Approved     19821
Cancelled      179
Name: order_status, dtype: int64
```

In [15]:
```
Transactions.online_order = Transactions.online_order.fillna('Nan')
```

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

Out[16]:
```
Solex             4253
Giant Bicycles    3312
WeareA2B          3295
OHM Cycles        3043
Trek Bicycles     2990
Norco Bicycles    2910
Name: brand, dtype: int64
```

In [17]:
```
Transactions.brand = Transactions.brand.fillna('Nan')
```

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

Out[18]:
```
Standard    14176
Road         3970
Touring      1234
Mountain      423
Name: product_line, dtype: int64
```

In [19]:
```
Transactions.product_line = Transactions.product_line.fillna('Nan')
```

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

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

In [21]:
```
Transactions.product_class = Transactions.product_class.fillna('Nan')
```

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

Out[22]:
```
medium    12990
large      3976
small      2837
Name: product_size, dtype: int64
```

In [23]:
```
Transactions.product_size = Transactions.product_size.fillna('Nan')
```

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

Out[24]:
```
388.920000    465
954.820000    396
53.620000     274
161.600000    235
260.140000    233
             ...
151.960000    124
206.350000    114
312.735016      1
270.299988      1
667.400024      1
Name: standard_cost, Length: 103, dtype: int64
```

In [25]:
```
Transactions.standard_cost = Transactions.standard_cost.fillna('Nan')
```

In [26]:
```
Transactions['product_first_sold_date']
```

Out[26]:
```
0        41245.0
1        41701.0
2        36361.0
3        36145.0
4        42226.0
          ...
19995    37823.0
19996    35560.0
19997    40410.0
19998    38216.0
19999    36334.0
Name: product_first_sold_date, Length: 20000, dtype: float64
```

In [27]:
```
Transactions.product_first_sold_date = Transactions.product_first_sold_date.fillna('Nan')
```

In [28]:
```
Transactions.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[28]:
```
transaction_id             0
product_id                 0
customer_id                0
transaction_date           0
online_order               0
order_status               0
brand                      0
product_line               0
product_class              0
product_size               0
list_price                 0
standard_cost              0
product_first_sold_date    0
dtype: int64
```

Exploring New Customer List Data Set

In [29]:
```
NewCustomerList.head(5)
```

Out[29]:

	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category	wealth_segment	deceased_indicator	owns_car	...	state	country	property_valuation	Unnamed: 16	Unnamed: 17	Unnamed: 18
0	Chickie	Brister	Male	86	1957-07-12	General Manager	Manufacturing	Mass Customer	N	Yes	...	QLD	Australia	6	0.47	0.5875	0.734375
1	Morly	Genery	Male	69	1970-03-22	Structural Engineer	Property	Mass Customer	N	No	...	NSW	Australia	11	0.78	0.7800	0.975000
2	Ardelis	Forrester	Female	10	1974-08-28	Senior Cost Accountant	Financial Services	Affluent Customer	N	No	...	VIC	Australia	5	0.43	0.4300	0.430000
3	Lucine	Stutt	Female	64	1979-01-28	Account Representative III	Manufacturing	Affluent Customer	N	Yes	...	QLD	Australia	1	0.55	0.6875	0.687500
4	Melinda	Hadlee	Female	34	1965-09-21	Financial Analyst	Financial Services	Affluent Customer	N	No	...	NSW	Australia	9	0.56	0.5600	0.700000

5 rows × 23 columns

In [30]:
```
NewCustomerList.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 23 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   first_name                           1000 non-null   object
 1   last_name                            971 non-null    object
 2   gender                               1000 non-null   object
 3   past_3_years_bike_related_purchases  1000 non-null   int64
 4   DOB                                  983 non-null    datetime64[ns]
 5   job_title                            894 non-null    object
 6   job_industry_category                835 non-null    object
 7   wealth_segment                       1000 non-null   object
 8   deceased_indicator                   1000 non-null   object
 9   owns_car                             1000 non-null   object
 10  tenure                               1000 non-null   int64
 11  address                              1000 non-null   object
 12  postcode                             1000 non-null   int64
 13  state                                1000 non-null   object
 14  country                              1000 non-null   object
 15  property_valuation                   1000 non-null   int64
 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   int64
 21  Rank                                 1000 non-null   int64
 22  Value                                1000 non-null   float64
dtypes: datetime64[ns](1), float64(5), int64(6), object(11)
memory usage: 179.8+ KB
```

In [31]:
```
#Dropping the unnamed columns
NewCustomerList.drop(['Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18',
       'Unnamed: 19', 'Unnamed: 20'], axis=1, inplace=True)
```

In [32]:
```
#Checking the shape of the dataset
NewCustomerList.shape
```

Out[32]: (1000, 18)

In [33]:
```
#Checking for null values
NewCustomerList.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[33]:
```
first_name                               0
last_name                               29
gender                                   0
past_3_years_bike_related_purchases      0
DOB                                     17
job_title                              106
job_industry_category                  165
wealth_segment                           0
deceased_indicator                       0
owns_car                                 0
tenure                                   0
address                                  0
postcode                                 0
state                                    0
country                                  0
property_valuation                       0
Rank                                     0
Value                                    0
dtype: int64
```

There are 4 missing values in columns. They can be dropped or nature analysis

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

Out[34]:
```
Sissel       2
Minshall     2
Borsi        2
Shoesmith    2
Sturch       2
            ..
O'Moylane    1
Axtens       1
Moxted       1
Conrad       1
Duffill      1
Name: last_name, Length: 961, dtype: int64
```

In [35]:
```
NewCustomerList.last_name = NewCustomerList.last_name.fillna('Nan')
```

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

Out[36]:
```
1998-02-05    2
1978-01-15    2
1977-11-08    2
1951-11-28    2
1979-07-28    2
             ..
1945-08-08    1
1943-08-27    1
1999-10-24    1
1976-01-24    1
1955-10-02    1
Name: DOB, Length: 958, dtype: int64
```

In [37]:
```
NewCustomerList.DOB = NewCustomerList.DOB.fillna('Nan')
```

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

Out[38]:
```
Associate Professor         15
Environmental Tech          14
Software Consultant         14
Chief Design Engineer       13
Assistant Manager           12
                            ..
Accountant II                1
Programmer IV                1
Administrative Officer       1
Accounting Assistant III     1
Web Developer I              1
Name: job_title, Length: 184, dtype: int64
```

In [39]:
```
NewCustomerList.job_title = NewCustomerList.job_title.fillna('Nan')
```

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

Out[40]:
```
Financial Services    203
Manufacturing         199
Health                152
Retail                 78
Property               64
IT                     51
Entertainment          37
Argiculture            26
Telecommunications     25
Name: job_industry_category, dtype: int64
```

In [41]:
```
NewCustomerList.job_industry_category = NewCustomerList.job_industry_category.fillna('Nan')
```

In [42]:
```
#Checking for duplicate values
NewCustomerList.duplicated().sum()
```

Out[42]: 0

In [43]:
```
#Checking for uniquess of each column
NewCustomerList.nunique()
```

Out[43]:
```
first_name                              940
last_name                               962
gender                                    3
past_3_years_bike_related_purchases     100
DOB                                     959
job_title                               185
job_industry_category                    10
wealth_segment                            3
deceased_indicator                        1
owns_car                                  2
tenure                                   23
address                                1000
postcode                                522
state                                     3
country                                   1
property_valuation                       12
Rank                                    324
Value                                   324
dtype: int64
```

In [44]:
```
NewCustomerList.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[44]:
```
first_name                             0
last_name                              0
gender                                 0
past_3_years_bike_related_purchases    0
DOB                                    0
job_title                              0
job_industry_category                  0
wealth_segment                         0
deceased_indicator                     0
owns_car                               0
tenure                                 0
address                                0
postcode                               0
state                                  0
country                                0
property_valuation                     0
Rank                                   0
Value                                  0
dtype: int64
```

Exploring Customer Demographic Data Set

In [45]:
```
CustomerDemographic.head()
```

Out[45]:

	customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category	wealth_segment	deceased_indicator	default	owns_car	tenure
0	1	Laraine	Medendorp	F	93	1953-10-12	Executive Secretary	Health	Mass Customer	N	"'	Yes	11.0
1	2	Eli	Bockman	Male	81	1980-12-16	Administrative Officer	Financial Services	Mass Customer	N	<script>alert('hi')</script>	Yes	16.0
2	3	Arlin	Dearle	Male	61	1954-01-20	Recruiting Manager	Property	Mass Customer	N	2018-02-01 00:00:00	Yes	15.0
3	4	Talbot	NaN	Male	33	1961-10-03	NaN	IT	Mass Customer	N	() { _; } >_[$(())] { touch /tmp/blns.shellsh...	No	7.0
4	5	Sheila-kathryn	Calton	Female	56	1977-05-13	Senior Editor	NaN	Affluent Customer	N	NIL	Yes	8.0

In [46]:
```
CustomerDemographic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 13 columns):
 #   Column                               Non-Null Count  Dtype
---  ------                               --------------  -----
 0   customer_id                          4000 non-null   int64
 1   first_name                           4000 non-null   object
 2   last_name                            3875 non-null   object
 3   gender                               4000 non-null   object
 4   past_3_years_bike_related_purchases  4000 non-null   int64
 5   DOB                                  3913 non-null   datetime64[ns]
 6   job_title                            3494 non-null   object
 7   job_industry_category                3344 non-null   object
 8   wealth_segment                       4000 non-null   object
 9   deceased_indicator                   4000 non-null   object
 10  default                              3698 non-null   object
 11  owns_car                             4000 non-null   object
 12  tenure                               3913 non-null   float64
dtypes: datetime64[ns](1), float64(1), int64(2), object(9)
memory usage: 406.4+ KB
```

In [47]:
```
#Checking for null values
CustomerDemographic.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[47]:
```
customer_id                              0
first_name                               0
last_name                              125
gender                                   0
past_3_years_bike_related_purchases      0
DOB                                     87
job_title                              506
job_industry_category                  656
wealth_segment                           0
deceased_indicator                       0
default                                302
owns_car                                 0
tenure                                  87
dtype: int64
```

In [48]:
```
CustomerDemographic['last_name'].value_counts()
```

Out[48]:
```
Pristnor     3
Ramsdell     3
Eager        2
Zisneros     2
Dredge       2
            ..
Whittock     1
Carwithen    1
Blas         1
Dearlove     1
Oldland      1
Name: last_name, Length: 3725, dtype: int64
```

In [49]:
```
CustomerDemographic.last_name = CustomerDemographic.last_name.fillna('Nan')
```

In [50]:
```
CustomerDemographic['DOB'].value_counts()
```

Out[50]:
```
1978-01-30    7
1964-07-08    4
1962-12-17    4
1978-08-19    4
1977-05-13    4
             ..
1989-06-16    1
1998-09-30    1
1985-03-11    1
1989-10-23    1
1991-11-05    1
Name: DOB, Length: 3448, dtype: int64
```

In [51]:
```
CustomerDemographic.DOB = CustomerDemographic.DOB.fillna('Nan')
```

In [52]:
```
CustomerDemographic['job_title'].value_counts()
```

Out[52]:
```
Business Systems Development Analyst    45
Tax Accountant                          44
Social Worker                           44
Internal Auditor                        42
Recruiting Manager                      41
                                        ..
Database Administrator I                 4
Health Coach I                           3
Health Coach III                         3
Research Assistant III                   3
Developer I                              1
Name: job_title, Length: 195, dtype: int64
```

In [53]:
```
CustomerDemographic.job_title = CustomerDemographic.job_title.fillna('Nan')
```

In [54]:
```
CustomerDemographic['job_industry_category'].value_counts()
```

Out[54]:
```
Manufacturing         799
Financial Services    774
Health                602
Retail                358
Property              267
IT                    223
Entertainment         136
Argiculture           113
Telecommunications     72
Name: job_industry_category, dtype: int64
```

In [55]:
```
CustomerDemographic.job_industry_category = CustomerDemographic.job_industry_category.fillna('Nan')
```

In [56]:
```
CustomerDemographic['default'].value_counts()
```

Out[56]:
```
100                                        113
1                                          112
-1                                         111
-100                                        99
Ù¡Ù¢Ù£                                       53
                                          ...
testâ testâ«                                31
/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
Name: default, Length: 90, dtype: int64
```

In [57]:
```
CustomerDemographic.default = CustomerDemographic.default.fillna('Nan')
```

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

Out[58]:
```
7.0     235
5.0     228
11.0    221
10.0    218
16.0    215
8.0     211
18.0    208
12.0    202
9.0     200
14.0    200
6.0     192
13.0    191
4.0     191
17.0    182
15.0    179
1.0     166
3.0     160
19.0    159
2.0     150
20.0     96
22.0     55
21.0     54
Name: tenure, dtype: int64
```

In [59]:
```
CustomerDemographic.tenure = CustomerDemographic.tenure.fillna('Nan')
```

In [60]:
```
#Checking for duplicate data
CustomerDemographic.duplicated().sum()
```

Out[60]: 0

In [61]:
```
#Checking for uniqueness of each column
CustomerDemographic.nunique()
```

Out[61]:
```
customer_id                            4000
first_name                             3139
last_name                              3726
gender                                    6
past_3_years_bike_related_purchases     100
DOB                                    3449
job_title                               196
job_industry_category                    10
wealth_segment                            3
deceased_indicator                        2
default                                  91
owns_car                                  2
tenure                                   23
dtype: int64
```

In [62]:
```
CustomerDemographic.apply(lambda x: sum(x.isnull()),axis=0)
```

Out[62]:
```
customer_id                            0
first_name                             0
last_name                              0
gender                                 0
past_3_years_bike_related_purchases    0
DOB                                    0
job_title                              0
job_industry_category                  0
wealth_segment                         0
deceased_indicator                     0
default                                0
owns_car                               0
tenure                                 0
dtype: int64
```

Exploring Customer Address Data Set

In [63]:
```
CustomerAddress.head(5)
```

Out[63]:

	customer_id	address	postcode	state	country	property_valuation
0	1	060 Morning Avenue	2016	New South Wales	Australia	10
1	2	6 Meadow Vale Court	2153	New South Wales	Australia	10
2	4	0 Holy Cross Court	4211	QLD	Australia	9
3	5	17979 Del Mar Point	2448	New South Wales	Australia	4
4	6	9 Oakridge Court	3216	VIC	Australia	9

In [64]:
```
CustomerAddress.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3999 entries, 0 to 3998
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   customer_id         3999 non-null   int64
 1   address             3999 non-null   object
 2   postcode            3999 non-null   int64
 3   state               3999 non-null   object
 4   country             3999 non-null   object
 5   property_valuation  3999 non-null   int64
dtypes: int64(3), object(3)
memory usage: 187.6+ KB
```

In [65]:
```
#Checking for null values.
CustomerAddress.isnull().sum()
```

Out[65]:
```
customer_id           0
address               0
postcode              0
state                 0
country               0
property_valuation    0
dtype: int64
```

In [66]:
```
#There are no null values.
```

In [67]:
```
#Checking for duplicate values
CustomerAddress.duplicated().sum()
```

Out[67]: 0

In [68]:
```
#Checking for uniqueness of each column
CustomerAddress.nunique()
```

Out[68]:
```
customer_id           3999
address               3996
postcode               873
state                    5
country                  1
property_valuation      12
dtype: int64
```

Exploring the columns

In [69]:
```
CustomerAddress['postcode'].value_counts()
```

Out[69]:
```
2170    31
2155    30
2145    30
2153    29
3977    26
        ..
3808     1
3114     1
4721     1
4799     1
3089     1
Name: postcode, Length: 873, dtype: int64
```

In [70]:
```
CustomerAddress['state'].value_counts()
```

Out[70]:
```
NSW                2054
VIC                 939
QLD                 838
New South Wales      86
Victoria             82
Name: state, dtype: int64
```

In [71]:
```
CustomerAddress['country'].value_counts()
```

Out[71]:
```
Australia    3999
Name: country, dtype: int64
```

In [72]:
```
CustomerAddress['property_valuation'].value_counts()
```

Out[72]:
```
9     647
8     646
10    577
7     493
11    281
6     238
5     225
4     214
12    195
3     186
1     154
2     143
Name: property_valuation, dtype: int64
```

All the columns appear to have consistent and correct information.