Task 1: Data Quality Assessment

Assessment of data quality and completeness in preparation for analysis. The client provided KPMG with 3 datasets:

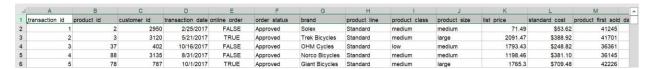
- Customer Demographic

A A	В	С	D	E		F	G	Н	I.	J	K	L	M	
customer id	first name	last name	gender	past 3 years 1	bik DOB		job title	job industry cate	wealth segment	deceased indica	default	owns car	tenure	
	1 Laraine	Medendorp	F	9	93 19	953-10-12	Executive Secret	Health	Mass Customer	N		Yes		11
	2 Eli	Bockman	Male	8	31 19	980-12-16	Administrative Of	Financial Service	Mass Customer	N	<script>alert('hi')</td><td>Yes</td><td></td><td>16</td></tr><tr><td></td><td>3 Arlin</td><td>Dearle</td><td>Male</td><td>6</td><td>51 19</td><td>954-01-20</td><td>Recruiting Manag</td><td>Property</td><td>Mass Customer</td><td>N</td><td>1-Feb</td><td>Yes</td><td></td><td>15</td></tr><tr><td></td><td>4 Talbot</td><td></td><td>Male</td><td>3</td><td>33 19</td><td>961-10-03</td><td></td><td>IT</td><td>Mass Customer</td><td>N</td><td>() { _: } >_[\$(\$())]</td><td>No</td><td></td><td>7</td></tr><tr><td></td><td>5 Sheila-kathryn</td><td>Calton</td><td>Female</td><td></td><td>56 19</td><td>77-05-13</td><td>Senior Editor</td><td>n/a</td><td>Affluent Custome</td><td>N</td><td>NIL</td><td>Yes</td><td></td><td>8</td></tr></tbody></table></script>			

- Customer Addresses

4	А	В	С	D	Е	F	G	Н	l l	J	K	L
1	customer_id	address	postcode	state	country	property_valuation	า					
2		1 060 Morning Aver	201	6 New South Wales	Australia	10						
3		6 Meadow Vale C	215	New South Wales	Australia	10						
4		4 0 Holy Cross Cou	421	1 QLD	Australia	9						
5		17979 Del Mar P	244	New South Wales	Australia	4						

Transactions data in the past 3 months



The task was to draft an email to the client identifying the data quality issues and strategies to mitigate the issues. Below are the main points of the letter:

- o In the table Transactions:
- Missing Data: There are blank spaces at the following columns: online_order, order_status, brand, product_line, product_class, product_size, standart_cost and product_first_sold_date.

When not possible to rescue the missing data, these will be deleted. They represent less than 1% of the data and will not affect the results of our training model.

- Inconsistency issues: The data was cleaned

transaction date is a object not a date

standard cost is a object not a float (number)

product first sold date is a float (number) not a date

- In the table CustomerDemographic:
- There are blank spaces at the following columns: last_name, DOB, job_industry_category, job_title and tenure. If possible, rescue the values.
- -Validity issues: DOB is an object not a date. The data was cleaned.
- -Gender column has different values for Male and Female (M, U, Femal...). The data will be cleaned. We recommend applying a single character option ('F', 'M', 'U' for others) or a drop-down list.
- The *Default* column was excluded, because it does not have any important value.
 - o In the table CustomerAdress:

a drop-down list with 'Other' to states out of Australia.								

- Standardize the name of the states in the column *state:* VIC or Victoria, NSW or New South Wales. The data will be cleaned to avoid multiple representations of the same value. We recommend an input message with the rule or