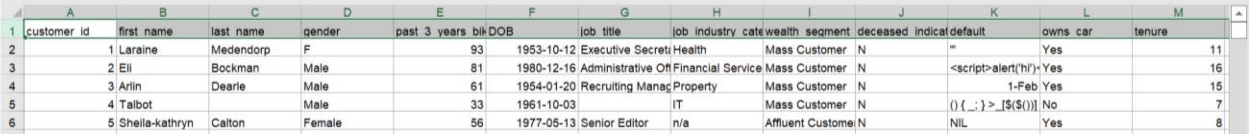
Task 1: Data Quality Assessment

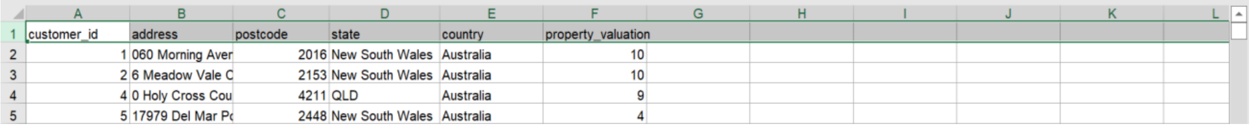
Assessment of data quality and completeness in preparation for analysis. The client provided KPMG with 3 datasets:

- Customer Demographic



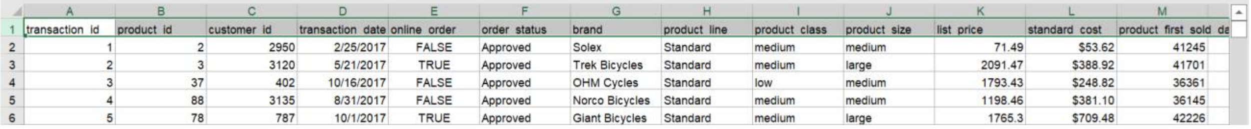
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	customer id	first name	last name	gender	past 3 years bik	DOB	job title	job industry cate	wealth segment	deceased indicat	default	owns car	tenure
2	1	Laraine	Medendorp	F	93	1953-10-12	Executive Secret	Health	Mass Customer	N	"	Yes	11
3	2	Eli	Bockman	Male	81	1980-12-16	Administrative Off	Financial Service	Mass Customer	N	<script>alert('hi')<	Yes	16
4	3	Arlin	Dearle	Male	61	1954-01-20	Recruiting Manag	Property	Mass Customer	N	1-Feb	Yes	15
5	4	Talbot		Male	33	1961-10-03		IT	Mass Customer	N	() { _; } >_[$($())]	No	7
6	5	Sheila-kathryn	Calton	Female	56	1977-05-13	Senior Editor	n/a	Affluent Custome	N	NIL	Yes	8

- Customer Addresses



	A	B	C	D	E	F	G	H	I	J	K	L
1	customer_id	address	postcode	state	country	property_valuation						
2	1	060 Morning Aver	2016	New South Wales	Australia	10						
3	2	6 Meadow Vale C	2153	New South Wales	Australia	10						
4	4	0 Holy Cross Cou	4211	QLD	Australia	9						
5	5	17979 Del Mar Pc	2448	New South Wales	Australia	4						

- Transactions data in the past 3 months



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	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 da
2	1	2	2950	2/25/2017	FALSE	Approved	Solex	Standard	medium	medium	71.49	$53.62	41245
3	2	3	3120	5/21/2017	TRUE	Approved	Trek Bicycles	Standard	medium	large	2091.47	$388.92	41701
4	3	37	402	10/16/2017	FALSE	Approved	OHM Cycles	Standard	low	medium	1793.43	$248.82	36361
5	4	88	3135	8/31/2017	FALSE	Approved	Norco Bicycles	Standard	medium	medium	1198.46	$381.10	36145
6	5	78	787	10/1/2017	TRUE	Approved	Giant Bicycles	Standard	medium	large	1765.3	$709.48	42226

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:

- 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.

  - In the table CustomerAdress:

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