

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

Data Audit Review for from Sprocket Central Pty Ltd.

Dear Sir/Madam,

Thank you for providing us with the below mentioned datasets from Sprocket Central Pty Ltd. This table describes the overall statistical summary of obtained datasets. Please, let us know, if these figures are not alligned with your understanding.

Sheets	Number of records	Number of Blank Cells	Distinct customer IDs	Largest Customer ID	Latest transaction date	Earliest transaction date	Description
Transactions	19998	1542	3492	5034	30.12.2017	01.01.2017	Sales transaction records
New Customer List	998	152	1000	#N/A	#N/A	#N/A	New customers that have not been added to customer list
Customer Demographic	3998	5044	4000	4000	#N/A	#N/A	Customers in the list
Customer Address	3997	23981	4003	4003	#N/A	#N/A	Addresses of customers

Date the data received:

25.07.2020

Date the data reviewed:

27.07.2020

Date that the audit opinion made:

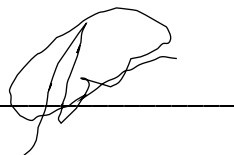
30.07.2020

The audit review and sugesstions made by:

Sitara Aghayeva

Data Analyst/KPMG Australia

Signature:



Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

Problems and Solutions

1. In New Customer Sheet and Customer Address sheet an address corresponds to more than one postcode. As each address should have one postcode, it is considered as an error. And addresses of customers number 340, 2699, 2826, 4001, 4002, 4003 in Customer Address sheet also exist in Customer Address sheet. First, it seems only 6 of new customers were added to Customer address sheet. However, we have analysed further and discovered that these 6 addresses are not completely same. The postcodes and states are different. Additionally, the same addresses have different postcodes which is hardly possible. [\(accuracy, consistency\)](#)

Customers in New Customer Sheet which the same addresses as in Customer Addresses Sheet

first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category	wealth_segment	address	postcode	state
Vinny	Incogna	Female	73	1953-02-13		Health	High Net Worth	8 Grayhawk Circle	2756	NSW
Aldridge	Poskitt	Male	84	1982-02-10	VP Sales	n/a	Mass Customer	7 Fordem Point	4161	QLD
Jenelle	Mc-Kerley	Female	40	1942-01-23	Data Coordinator	Financial Services	Mass Customer	9 Springview Terrace	4068	QLD
Ewell	Paulusch	Male	31	1998-01-15	Engineer I	Manufacturing	Mass Customer	8194 Lien Street	4032	QLD
Kariotta	Naper	Female	8	1952-04-07	VP Sales	Health	Mass Customer	87 Crescent Oaks Alley	2756	NSW
Myrtie	Ostrich	Female	70	1996-06-18	VP Quality Control	Property	Affluent Customer	320 Acker Drive	2251	NSW

Customer IDs (Matched from Customer Demographics) Addresses in Customer Address Sheet that are the same with the addresses in New Customer Sheet

customer_id	address	postcode	state	country	property_valuation
4003	320 Acker Drive	2251	NSW	Australia	7
4002	8194 Lien Street	4032	QLD	Australia	7
4001	87 Crescent Oaks Alley	2756	NSW	Australia	10
2826	9 Springview Terrace	3121	VIC	Australia	10
2699	7 Fordem Point	3057	VIC	Australia	7
340	8 Grayhawk Circle	2089	NSW	Australia	10

customer_id	first_name	last_name	gender	DOB	job_title
340	Joshuah	Purvey	Male	31.07.1973	
2699	Michal	Woltering	Male	25.05.1978	Account Representative IV
2826	Loise	Mulvany	Female	30.04.1962	Web Developer II

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

In order to compare these two datasets we have matched states and postcodes. Here 2 of the same addresses within the same country (Australia) belong to the different states. Moreover, the three of them have different postcodes. The table below shows these records in more detail.

Customer States and Postcodes match Table
(between New Customer List and Customer Addresses)

address	postcode	state	country	property_valuation	Matched postcode	Postcode = Mathced Postcode in C/A	Matched State	State = Matched State in C/A
9 Springview Terrace	4068	QLD	Australia	5	3121	FALSE	VIC	FALSE
87 Crescent Oaks Alley	2756	NSW	Australia	10	2756	TRUE	NSW	TRUE
8 Grayhawk Circle	2756	NSW	Australia	8	2089	FALSE	NSW	TRUE
7 Fordem Point	4161	QLD	Australia	5	3057	FALSE	VIC	FALSE
320 Acker Drive	2251	NSW	Australia	7	2251	TRUE	NSW	TRUE
8194 Lien Street	4032	QLD	Australia	7	4032	TRUE	QLD	TRUE

For a more deep investigation we matched the previously recorded customer with the new customers due to their addresses and we observed that they have different names. So they are not the same people recorded twice.

The customer addresses for customers No 4001, 4002, 4003 corresponds to the addresses in Customer Adress Sheet in terms of address, state and post code.

Solution: The new customer records should be added to list. Customers No 4001, 4002, 4003 have been already included in Customer Adress Sheet, these IDs should be taken into account while adding information to Customer Demographic Sheet.

The postcodes and states should be checked and verified again. The falsful data should be replaced with the right one.

The new customers that have the same addresses with the previously recorded customer should be inspected again.

2. In Customer Demoraphic Sheet the surname of the custome No 2746 is “Egle of Germany”. It does not look like a real surname. It seems like an erroneus situation. [\(accuracy\)](#)

Default column is not readable. The information it contains seems like code omitted from a source and is not converted to an understandable format. [\(accuracy\)](#)

Solution: The verification of data requirs further investigation of customer records.

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

Extract from Customer Demographic Sheet

customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category
2746	Carmelia	Egle of Germany	Female	97	25.10.1953	Actuary	Financial Services

3. Moreover, the date of birth of customer No 34 – Jephthah Bachmann is “1843-12-21” an is not in date format. It is hardly possible to have a customer 177 years old. [\(accuracy, consistency, validity\)](#)

Solution: We can easily covert it to the date format (explained below), but the verification of data requirs further investigation of customer records. To prevent such erros data validation for date should be applied such as: 1900 <= Date <= YEAR(TODAY())) - 18

Extract from Customer Demographic Sheet

customer_id	first_name	last_name	gender	past_3_years_bike_related_purchases	DOB	job_title	job_industry_category
34	Jephthah	Bachmann	U	59	1843-12-21	Legal Assistant	IT

4. In New Customer List the numbers in “past_3_years_bike_related_purchases”, “property_valuation” “postcode” columns are classified as text, while they are classified as numbers in other sheets. It is possible and reasonable to classify postcodes as text, but if we do that we should apply this method for all of our data in order to be able to compare and match them. As a contrary, the other two columns should always be classified as number, as they are numeric data. [\(accuracy, consistency, validity\)](#)

Solution: These columns can be easily converted to number using the error message itself or with the help of VALUE() formula creating a new column.

5. Product first sold date for each unique product is not the same everywhere and product IDs begin with “0”. This order generate confusions. It is better to order them beginning from 1. There are 1378 transactions for the sale of products with “0” product ID in Transactions Sheet. [\(validity\)](#)

Solution: To prevent future confusions we can add 1 to each product ID beginning with number “1” to designate them.

6. The transaction IDs do not increase with the transaction date. According to the number order the most recent transaction should have had the largest transaction number. The table below describes the order of transactions and dates. The formatted transaction numbers follows the pattern of the order from smallest to largest and the color scale moves from blue (indicates the

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

minimum) to the red (indicates the maximum). As a contrast, the transaction date column does not follow this pattern as it can be observed by the color scale disarray. [\(accuracy, consistency\)](#)

Extraction from Transactions Sheet

transaction_id	product_id	customer_id	transaction_date	online_order	order_status
1144	4	1288	21.09.2017	TRUE	Approved
1145	25	3146	18.10.2017	TRUE	Approved
1146	37	2508	04.08.2017	TRUE	Approved
1147	34	2092	28.03.2017	TRUE	Approved
1148	1	1671	14.11.2017	FALSE	Approved
1149	27	3211	12.07.2017	TRUE	Approved

Solution: The transaction IDs can be constructed as “previous ID + 1” technique. However, we do not exactly know what is applied for the transaction records.

7. In Transactions sheet a unique product ID corresponds to different product types. The table below is an extracted from Transactions sheet. The Transaction No 16118 is different, although it has the same product id with the 2 other transactions. [\(uniqueness, validity\)](#)

Solution: To prepare a product table with a primary key column (Product ID) and restrict it with a validation to permit only unique values and reject any duplicate.

In order to apply this restriction the below mentioned formula should be written choosing custom data validation: COUNTIFS(COLUMN_RANGE; FIRST_CELL_OF_COLUMN)=1

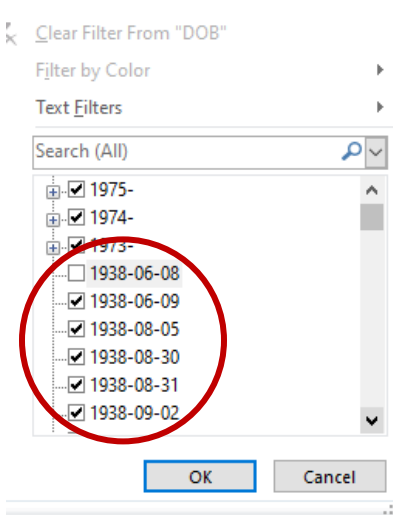
Extraction from Trasactions Sheet

transaction_id	product_id	brand	product_line	product_class	product_size	list_price	standard_cost	product_first_sold_date
6825	1	Giant Bicycles	Standard	medium	medium	1403,5	\$954,82	07.11.1991
11881	1	Giant Bicycles	Standard	medium	medium	1403,5	\$954,82	15.09.2012
16118	1	Giant Bicycles	Touring	medium	large	1873,97	\$863,95	25.07.2004

8. The date records in New Customer List within the dates of Birth (DOB) column have been not classified consistently. The dates after the year 1938 have been classified as text, not dates. It can be seen while filtering the date column. It is very obvious in the screenshot below, as the dates (1073, 1974, 1975) have “+” sign to extend them to the monthes and days. However, the dates in the year 1938 are individually designated in filter, because their type is not recognized as date in Excel. [\(accuracy, consistency, validity\)](#)

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

Solution: Add data validation to restrict the column in order to allow the input only between a particular (may be the date when company founded and the date less than, or equal to TODAY()) and apply specific date format (dd-mm-yyyy, yy-mm-dd, dd mmm yyyy, etc) to the column. It will make the inputs consistent with the other inputs in the column. To convert this column to a date format complete, a new column should be added with a formula: DATE(LEFT([@DOB]; 4); MID([@DOB]; 6; 2); RIGHT([@DOB]; 2))



9. Gender column in Customer Demographics contains F and M besides Male and Female. The data that belongs to the same category seems as different, as two of the customer records were designated F and M rather than Male or Female. It means the data in this column has not been categorised consistently and it will cause problems while summarizing and comparing them. (consistency, validity)

customer_id	first_name	last_name	gender
57	Abba	Masedon	M
1	Laraine	Medendorp	F

Some of the State names within state column located in Customer Address sheet are not abbreviated and they are recognised as a different category by Excel. (consistency, validity)

customer_id	address	postcode	state	country	property_valuation
749	760 Forest Dale Place	3068	Victoria	Australia	12
3999	1482 Hauk Trail	3064	VIC	Australia	3
147	12375 Cambridge Pass	2065	New South Wales	Australia	11
3998	736 Roxbury Junction	2540	NSW	Australia	6

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

Solution: Add data validation with a restricted list of options (QLD/NSW/VIC, Male/Female/U) and badge the cells with inappropriate input.

10. The last purchase in the list was made in 2017. We do not have sales data for the years of 2018, 2019, and 2020. [\(currency, completeness\)](#)

Solution: Automitized system to update data and to show the newly added records that need to be added to the necessary sheets and shared with necessary parties.

11. There are Rank column and Value columns. They seem interrelated. Nevertheless, the purpose of given data is not obvious and it looks irrelevant. Additionally, the Tenure and Property Value columns seems about real estate and are irrelevant to our customer data. [\(relevancy\)](#)

Solution: To review the indicators that are necessary and unneccasary and eliminate the redundant columns from the table.

12. The datasets contain N/As. There are 1542 N/As in Transaction sheet, 152 N/As in New Customer List, 1046 in Customer Demographics. Ther is not any N/A in Customer Adresses. These N/As are about name, surname, dates, order spesifications, customer information, etc. [\(completeness\)](#)

Solution: Add conditional formatting to the data to designate blank cells in order to notice them immediately and explore the sources to find missing data. To make a summary table for the data helps to observe thechanges in the number of records, N/As and other indicators promptly.

Extraction from the summary table presented above

Sheets	Number of records	Number of Blank Cells
Transactions	19998	1542
New Customer List	998	152
Customer Demographic	3998	5044
Customer Address	3997	23981

13. List Price column in Transactions Sheet does not have a currency sign. This column does not give us a complete information about the price of a product. [\(accuracy, completeness\)](#)

Solution: Using the currency or accounting type within a column consistently. Formating data as table is an easy way to manage this process. In a table the next added input takes the formatting from previous rows consisitently.

Data Validation options for Columns

Most of the validation problems cause consistency problems. For example, calegorical columns need restricted list of categories or names, numerical columns should have a data validation restricted with a particular date or number classification such as whole numbers or decimals with boudaries. According to all abovementioned problems we can offer data validation restrictions for each column.

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.

For Transactions sheet:

- a) Product ID, Customer ID, Transaction ID – only whole numbers began with 1
- b) Transaction ID – may not be duplicated
- c) Online order, order status, brand, product line, product class, product size – name from list
- d) Transaction date, Production first sold date – enter a valid date (after a particular date and before or equal TODAY)
- e) List price, standard cost – decimal

For Customer Demographics sheet:

- a) Customer ID – whole numbers began with 1 and no duplicate
- b) Gender – name from list
- c) Past 3 years bike related purchases – only whole numbers
- d) DOB (Date of Birth) - enter a valid date (between 1900 and before YEAR(TODAY) - 18)
- e) Job Title, job industry, wealth segment, deceased indicator, owns car – from list
- f) Tenure – only whole numbers

For New Customer List sheet:

- a) Gender – name from list
- b) Past 3 years bike related purchases – only whole numbers began with 1
- c) DOB (Date of Birth) - enter a valid date (between 1900 and before YEAR(TODAY) - 18)
- d) Job Title, job industry, wealth segment, deceased indicator, owns car – from list
- e) Tenure – only whole numbers
- f) Postcode – whole numbers began with 1 between 1000 and 9999
- g) State – choose from list
- h) Country – choose from list
- i) Property valuation – only whole numbers
- j) rank – whole numbers began with 1
- k) value – decimals

For Customer Addresses sheet:

- a) Customer ID – whole numbers began with 1 and no duplicate
- b) Postcode – whole numbers began with 1 between 1000 and 9999
- c) State – choose from list
- d) Country – choose from list
- e) Property valuation – whole numbers began with 1

Disclaimer: This document is prepared for only educational purposes. It is a task for Internship Program bu KPMG Australia and Internsherpa.