B Cover Page Templat

School of Informatics & IT
TEMASEK POLYTECHNIC

DIPLOMA IN BIG DATA & ANALYTICS

DATA WAREHOUSING AND BUSINESS INTELLIGENCE (CDA2C01)

AY 2021/2022 APRIL SEMESTER

Project – Individual Report

Practical Class: P03

Tutor: Ms Nami

Submitted By: Edzran Hisham

Matric Number: 2004986B

Date: 8/8/2021

"By submitting this work, I am / we are declaring that I am / we are the originator(s) of this work and that all other original sources used in this work has been appropriately acknowledged.

I / We understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my/ our own without proper acknowledgement.

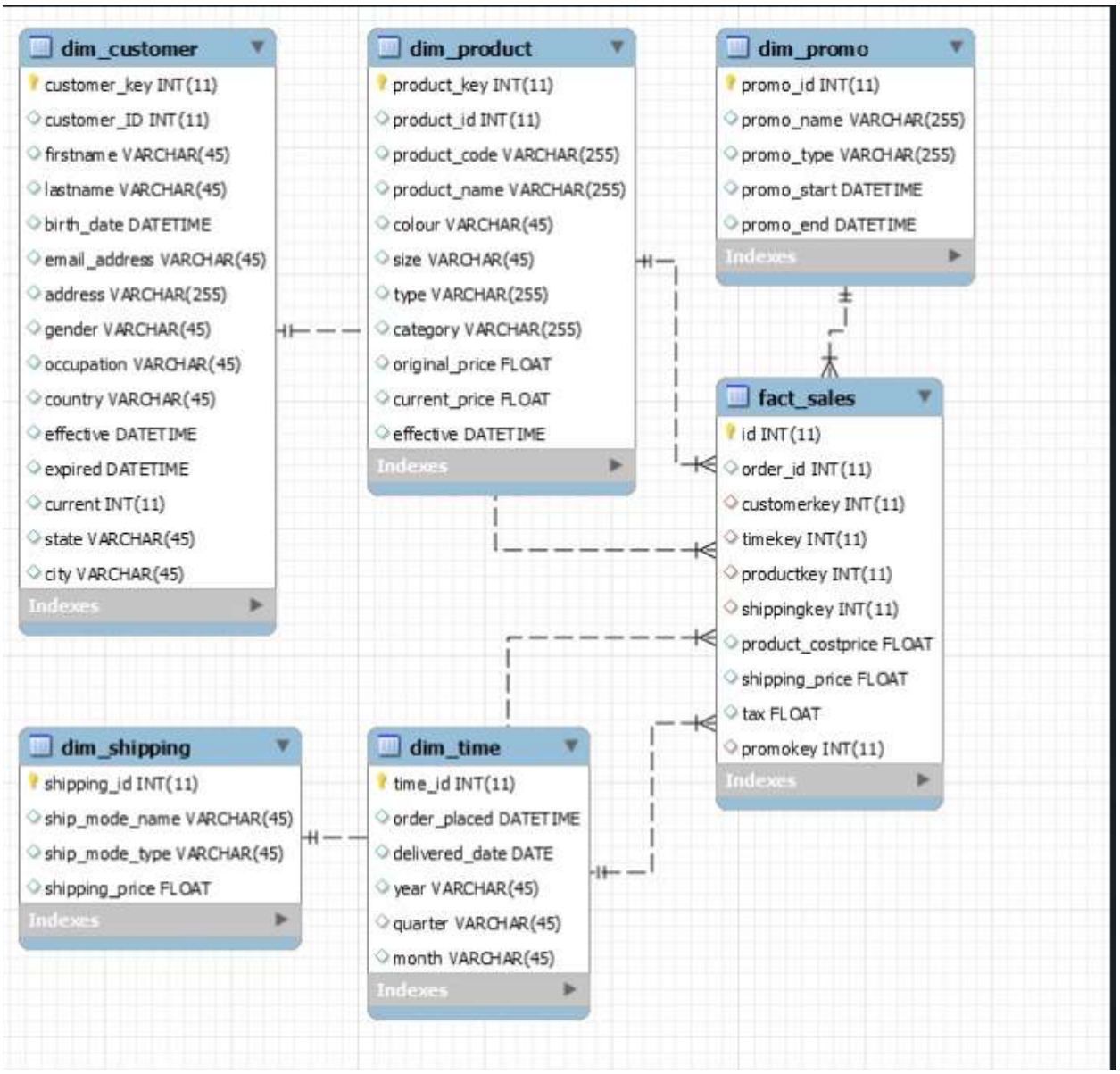
I / We also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism."

Edzran Hisham 

Hello. In this report, I will be documenting the steps and procedures I will take in order to design and perform my ETL process.

Here is the data warehouse schema that our company Asadi is looking to use to integrate all of their data in.

Asadi star schema



Data Dictionary

Column	Data Type	Description
customer_key	INT(11)	the sarugate key of the customer data
customer_ID	INT(11)	the ID of the customer
firstname	VARCHAR (45)	firstname of customer
lastname	VARCHAR (45)	lastname of customer
birth_date	DATETIME	birthdate of customer
email_address	VARCHAR (45)	email address of customer
address	VARCHAR (255)	address of customer
gender	VARCHAR (45)	gender of customer
occupation	VARCHAR (45)	occupation of customer
country	VARCHAR (45)	country of origin of customer
effective	DATETIME	the effective date of the data
expired	DATETIME	the expired date of the data
current	INT(11)	values of 0 or 1. 1 means current, 0 means not current
state	VARCHAR (45)	the state of the customer
city	VARCHAR (45)	city of the customer

Column	Data Type	Description
Product_key	INT(11)	the sarrugate key of the product data
Product_ID	INT(11)	the ID of the product
product_code	VARCHAR (255)	the product code of the product
Product_Name	VARCHAR (255)	the name of the product
Colour	VARCHAR (45)	the colour of the product
Size	VARCHAR (45)	the size of the product
Type	VARCHAR (255)	the type of the prooduct
Category	VARCHAR (255)	the category of the product
current_price	FLOAT	the current price of the product (SCD type 3)
original_price	FLOAT	the original price of the product (SCD type 3)
effective	DATETIME	the date of time where the current price takes effect

Column	Data Type	Description
shipping_ID	INT(11)	the primary key of shipping data
ship_mode_name	VARCHAR (255)	the name of the shipping name
ship_mode_type	VARCHAR (255)	the name of the shipping mode
shipping_price	FLOAT	the usual price of the shipping

Column	Data Type	Description
time_ID	INT(11)	the primary key of time data
order_placed	DATETIME	the date where the order is placed
delivered_date	DATE	the date where the order is delivered
year	VARCHAR (45)	the year of the order placed
quarter	VARCHAR (45)	the quarter of the order placed
month	VARCHAR (45)	the month of the order placed

Column	Data Type	Description
promo_ID	INT(11)	the primary key of promotion data
promo_name	VARCHAR (45)	the name of the promotion
promo_type	VARCHAR (45)	the type of the prromotion
promo_start	DATETIME	the start date of the promotion
promo_end	DATETIME	the end date of the promotion

Column	Data Type	Description
ID	INT(11)	the primary key of the fact_sales data
order_id	INT(11)	the id of the order (digenerate dimension)
customerKey	INT(11)	the foreign key connected to customer data
TimeKey	INT(11)	the foreign key connected to time data
ProductKey	INT(11)	the foreign key connected to product data
ShippingKey	INT(11)	the foreign key connected to shipping data
promoKey	INT(11)	the foreign key connected to promotion data
product_costprice	FLOAT	the product price after discount
shipping_price	FLOAT	the shipping price after discount
tax	FLOAT	the tax of the transit

Datasets

Here are the data sets that the company is looking to integrate:

Asadi_products

Format: xlsx

The product catalog was generated from Asadi's e-commerce website and exported in xlsx format.

1	product_id	product_code	product_name	colour	size	type	category	price
2	1	AQ3308-06	Air Force 1 Blue	Blue	US9	casual	shoe	250
3	2	AQ3789-72	Curry 7	Black	US7	basketbal	shoe	119.99
4	3	AQ6831-01	Kobe 1	Gold	US11	basketbal	shoe	450
5	4	SV3741-17	Dri-fit Shirt	White	M	sport	shirt	59.99
6	5	PX5831-43	Sweatpants	Grey	33	casual	pants	7999
7	6	PX2111-23	Running shorts	Black	28	sport	pants	59.99
8	7	SV8282-27	Dri-fit Tank-top	White	L	sport	shirt	59.99
9	8	CP7391-28	Cap	Black	M	casual	accessory	89.99
10	9	AQ3271-37	Zion 1	Red	US17	basketbal	shoe	350
11	10	BS1283-31	Fifa world cup football	White	7	football	ball	99.99
12	11	BS1592-12	Fiba international basketbal	Orange	7	basketbal	ball	79.99
13	11	BS1592-12	Fiba international basketbal	Orange	7	basketbal	ball	79.99
14	12	SV2724-07	Ronaldo 7 jersey	White	L	football	shirt	69999
15	13	SV2345-19	Messi 10 jersey	Purple	S	football	shirt	69.99
16	14	SV8214-10	Rooney 10 jersey	Red	M	football	shirt	69.99
17	15	SV2840-23	Jordan 23 jersey	Red	M	basketbal	shirt	69.99
18	16	SV2743-30	Curry 30 jersey	Yellow	S	basketbal	shirt	69.99
19	16	SV2743-30	Curry 30 jersey	Yellow	S	basketbal	shirt	69.99
20	16	SV2743-30	Curry 30 jersey	Yellow	S	basketbal	shirt	69.99
21	17	SV7339-34	Giannis 34 jersey	Green	XL	basketbal	shirt	69.99
22	18	CP2278-82	Wristband	White	S	sport	accessory	34.99
23	19	CP3729-12	Arm sleeve phone holder	Black	L	sport	accessory	17.99
24	20	PX3813-13	Compression tights	Black	30	running	pants	34.99
25	20	PX3813-13	Compression tights	Black	30	running	pants	34.99
26	20	PX3813-13	Compression tights	Black	30	running	pants	34.99
27	21	SK3222-12	Arsenal Bag	Red	15L	Sports	Bags	45
28	22	SK3221-13	Adventure Backpack Small	Black	20L	Sports	Bags	60
29	23	SK3452-15	Modern Utility Three-Way B	Grey	25L	casual	Bags	130

This data is in excel(xlsx) format. The data contains the catalog of product items from Asadi's ecommerce stores. The data is not clean and the types of bad data that can be found are:

- Duplicate entries
- Inconsistency in sizing
- Error in pricing value

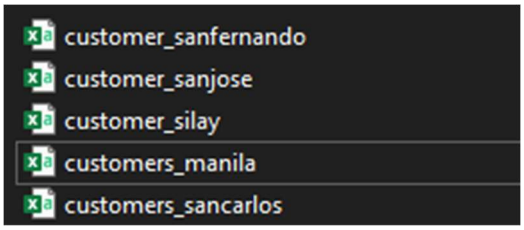
The types of transformation I would like to perform are:

- Adding of surrogate key and type 3 SCD to track product information changes
- Joining of product_costprice table to calculate profit

<u>Asadi customer</u>

Format: CSV

Asadi also have various retail stores across the Philippines. They have outlets in Manila, Silay, San Carlos, San Jose and San Feranando. Each retail store keeps track of their own customers and therefore we would have to join the datasets together.



Manila dataset:

1	customer_	firstname	lastname	birth_date	email_add	address	gender	occupatio	country	state	city
2	1	Reina	Pollock	Mar. 12, 1!	Reina@gn	658 Gonza	M	Tour Guid	Phillipines		Manila
3	2	James	Kubota	Jul. 24, 19	James@g	UCPB Buil	Male	Tour Guid	Phillipines		Manila
4	3	Zandra	Arai	Oct. 7, 198	Zandra@g	No. 39 Pla	M	Student	Phillipines		Manila
5	4	Stefany	Luthy	3-May-68	Stefany@	3/F FEMII	F	Tour Guid	Phillipines		Manila
6	5	Young	Dirks	Sep. 15, 1!	Young@g	9684 Sunn	F	Tour Guid	Phillipines		Manila
7	6	Mariam	Tremper	Aug. 20, 1!	Mariam@	Racine, W	Female	Mechanic	Phillipines		Manila
8	7	Frances	Willie	########	Frances@	9996 Shor	F	Tour Guid	Phillipines		Manila
9	8	Normand	Manion	Feb. 5, 19	Normand	Laurel, MI	F	Tour Guid	Phillipines		Manila
10	9	Coletta	Pabon	Aug. 3, 19	Coletta@	7946 Hawt	M	Accountar	Phillipines		Manila
11	10	Becky	Cotnoir	Aug. 18, 1!	Becky@gr	Zion, IL 60	F	Tour Guid	Phillipines		Manila
12	11	Santo	Belair	Mar. 24, 1!	Santo@gn	7189 Walt	M	Tour Guid	Phillipines		Manila
13	12	Lauralee	Chevere	Sep. 10, 1!	Lauralee@	Toms Rive	M	Tour Guid	Phillipines		Manila
14	13	Cathern	Most	Sep. 11, 1!	Cathern@	50 Beechv	M	Tour Guid	Phillipines		Manila
15	14	Terese	Foy	Jan. 20, 19	Terese@g	Wake For	M	Data Scier	Phillipines		Manila
16	15	Windy	Carbonell	Oct. 23, 1!	Windy@g	640 Yukon	Female	Data Scier	Phillipines		Manila
17	16	Bernardo	Randle	Aug. 23, 1!	Bernardo(	Barrington	M	Data Scier	Phillipines		Manila
18	17	Lyda	Degroot	Mar. 5, 19!	Lyda@gm	153 Hamil	F	Data Scier	Phillipines		Manila
19	18	Rosena	Fiorenza	Jul. 10, 19!	Rosena@	Irmo, SC 2	M	Data Scier	Phillipines		Manila
20	19	Kay	Marth	Feb. 18, 1!	Kay@gma	50 NE. Livi	M	Lawyer	Phillipines		Manila
21	20	Armanda	Gilbertsor	########	Armanda(	Aliquippa	F	Lawyer	Phillipines		Manila
22	21	Mi	Wilson	Feb. 29, 1!	Mi@gmail	246 Rocka	F	Lawyer	Phillipines		Manila

The type of bad data that can be found are:

- Empty column (State)
- Incorrect datetime format
- Inconsistent gender values

The types of transformation I would like to perform are:

- Adding of surrogate key to keep track of historical data in the future
- Calculate new column for analysis : Age
- Join and combine the various datasets from different retail stores
- Drop column “Country” as all the customers in the datasets reside in Philippines.

San Carlos dataset:

ID	Name	Lastname	birthday	email	address	gender	job	country	state
22	Kaylene	Rhem	Nov. 14, 1	Kaylene@	Clearwate	F	Engineer	Phillipines	
23	Sheena	Mable	Aug. 30, 1	Sheena@	7536 Cher	M	Accountar	Phillipines	
24	Carmelia	Hosch	Jun. 14, 19	Carmelia@	Crofton, M	F	Student	Phillipines	
25	Lia	Yepez	Apr. 26, 19	Lia@yaho	390 Lafaye	M	Accountar	Phillipines	
26	Russ	Tinoco	Dec. 14, 19	Russ@yah	Zeeland, M	F	Teacher	Phillipines	

Additionally, the dataset format is similar, however the column names are different. Also, there is no "City" column. I will take this into consideration when I am joining as well and ensure I match the values.

San Jose dataset:

1	ID	Name	Lastname	birthday	email	address	gender	job	country	state
2	36	Yetta	Goering	Jul. 3, 1963	Yetta@gm	Lakeville,	M	Lawyer	Phillipines	
3	37	Katheleer	Kilbourn	13-May-82	Katheleer	806 Elizab	Male	Lawyer	Phillipines	
4	38	Jeni	Heatwole	Dec. 20, 1989	Jeni@gma	Stratford,	M	Student	Phillipines	
5	39	Sheron	Duhaime	Mar. 19, 1972	Sheron@g	8389 Stud	F	Student	Phillipines	
6	40	Ardath	Musson	Apr. 27, 1988	Ardath@g	Bethpage,	F	Student	Phillipines	
7	41	Linnea	Greenleaf	Jul. 1, 1982	Linnea@g	5 Locust St	F	Accountar	Phillipines	
8	42	Alayna	Cavins	Aug. 12, 1966	Alayna@g	Independ	F	Student	Phillipines	
9	43	Herb	Applebau	Aug. 27, 1998	Herb@gm	9513 East	M	Accountar	Phillipines	

San Fernando:

1	ID	firstname	lastname	date of bi	email	address	gender	occupatio	country	state
2	44	Maribeth	Depriest	Aug. 8, 19	Maribeth(	554 Victor	F	Mechanic	Phillipines	
3	45	Alexis	Vashon	Dec. 16, 19	Alexis@ya	Winchest	Female	Lawyer	Phillipines	
4	46	Harland	Vanderfor	Nov. 20, 1	Harland@	527 Shub I	F	Mechanic	Phillipines	
5	47	Gia	Gundlach	Jul. 24, 19	Gia@yahc	Bridgewat	Male	Accountar	Phillipines	
6	48	Latricia	Olivo	Feb. 22, 19	Latricia@y	7286 Hanc	M	Mechanic	Phillipines	
7	49	Betsy	Cintron	Jul. 13, 19	Betsy@ya	Fitchburg,	F	Accountar	Phillipines	
8	50	Galen	Coy	Mar. 18, 19	Galen@ya	13 Shadov	Male	Data Scier	Phillipines	

Silay dataset:

1	ID	Name	Lastname	d.o.b	email	address	gender	job	country	state
2	27	Kaylee	Collingtor	Feb. 10, 19	Kaylee@g	Ocean Spr	M	Data Scier	Phillipines	
3	28	Laurel	Malta	Jul. 22, 19	Laurel@gr	41 South J	F	Student	Phillipines	
4	29	Brenton	Gwynn	Aug. 24, 19	Brenton@	Longview,	M	Tour Guid	Phillipines	
5	30	Sindy	Shoemake	Jan. 28, 19	Sindy@ya	35 Oklaho	M	Student	Phillipines	
6	31	Nilda	Hooton	Dec. 18, 19	Nilda@ya	Irvington,	F	Mechanic	Phillipines	
7	32	Casimira	Jetton	Oct. 2, 198	Casimira@	329 Glen F	F	Mechanic	Phillipines	
8	33	Stanley	Schlatter	Jun. 11, 19	Stanley@	De Pere, V	M	Lawyer	Phillipines	
9	34	Whitley	Lombardi	Oct. 13, 19	Whitley@	585 Peach	Female	Mechanic	Phillipines	
10	35	Bobette	Lajeuness	Jan. 18, 19	Bobette@	Adrian, M	F	Tour Guid	Phillipines	

asadi_promo

Format: MS Access

MS Access is a multi-user platform. A database containing the promotions and the shipping rate is shared amongst the company's Marketing team and shipping partners.

This is done so that the data is known and shared amongst other related departments that use the information and also the Marketing team can customize and alter their rates and the changes would be seen by other departments.

The data is relatively clean and no data cleaning is needed.

Promotions

ID	promo_nam	promo_type	promo_start	promo_end	promo_disc
1	12.12 Sale	12% off all items	12/12/2020	12/14/2020	12
2	Back 2 School	30% off all bags and shoes	3/1/2021	4/15/2021	30
3	Mid-Year Sale	30% off all items	6/1/2020	7/1/2020	30
6	8.8 National Da	Discounts of $50, min spending $200	8/8/2020	8/10/2020	50

Asadi_shipping

Format: MS Access

The shipping information is shared amongst Asadi and shipping partners. This is done so that any changes in the shipping rates would be reflected and Asadi may monitor the rate changes as well. \

The data is also relatively clean and no data cleaning is needed.

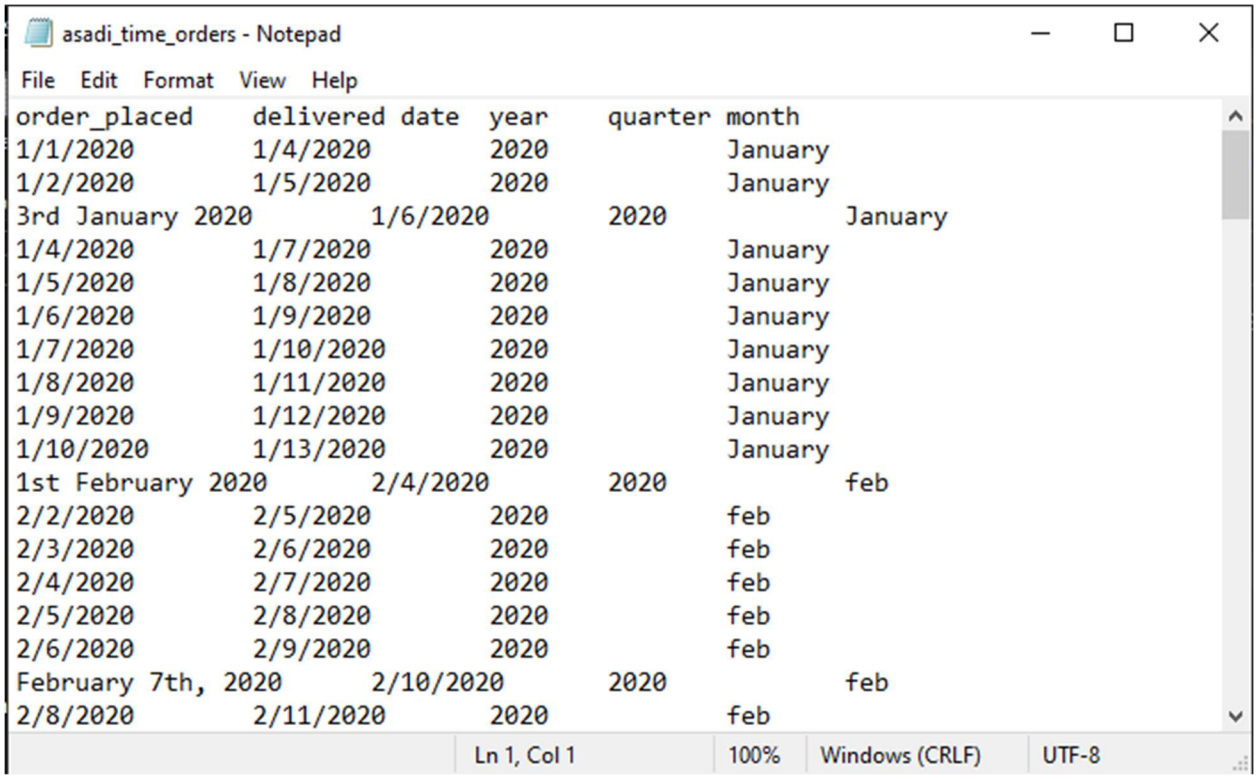
Shipping

shipping_ID	ship_mode_	ship_mode_	shipping_price_per_kilo
1	Fedex	Air	$5.00
2	UPS	Sensitive Air	$7.00
3	DHL	Sensitive Air	$7.00
4	Ninjavan	Land	$2.50
5	J&T Express	Land	$2.50
6	Ocean Networl	Sensitive Sea	$0.75
7	Orient Oversea	Sea	$0.50

Asadi_time

Format: txt.

This dataset is printed out from the ecommerce POS system. It contains the dates an item is ordered and successfully delivered:



The data contains bad data types such as:

- Incorrect datetime format in order_placed and delivered_date columns
- Inconsistency in Month naming
- Empty column (Quarter)
- Missing data in delivered_date column

Fact_orders

Format: mySQL

This data is in mySQL format as all orders come from Asadi's POS system.

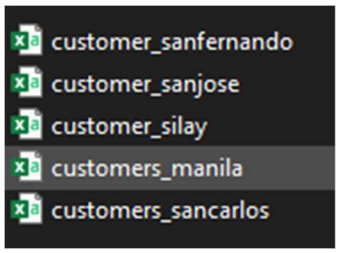
id	order_id	customerkey	timekey	productkey	shippingkey	product_costprice	shipping_price	tax	promokey
1	1	16	1	23	3	NULL	NULL	NULL	1
2	2	32	2	6	5	NULL	NULL	NULL	3
3	3	15	3	6	6	NULL	NULL	NULL	6
4	4	45	4	11	4	NULL	NULL	NULL	2
5	5	36	5	18	5	NULL	NULL	NULL	1
6	6	30	6	4	7	NULL	NULL	NULL	6
7	7	15	7	18	2	NULL	NULL	NULL	3
8	8	1	8	16	6	NULL	NULL	NULL	2
9	9	47	9	18	2	NULL	NULL	NULL	6
10	10	16	10	13	4	NULL	NULL	NULL	6
11	11	3	11	18	3	NULL	NULL	NULL	1
12	12	47	12	7	2	NULL	NULL	NULL	2
13	13	39	13	8	2	NULL	NULL	NULL	3
14	14	12	14	14	4	NULL	NULL	NULL	1
15	15	1	15	11	6	NULL	NULL	NULL	6
16	16	43	16	2	6	NULL	NULL	NULL	3
17	17	37	17	9	6	NULL	NULL	NULL	2
18	18	26	18	9	6	NULL	NULL	NULL	1
19	19	3	19	17	7	NULL	NULL	NULL	6
20	20	33	20	9	7	NULL	NULL	NULL	3
21	21	13	21	2	1	NULL	NULL	NULL	1
22	22	7	22	12	1	NULL	NULL	NULL	1
23	23	20	23	10	4	NULL	NULL	NULL	2
24	24	43	24	16	6	NULL	NULL	NULL	3
25	25	26	25	1	3	NULL	NULL	NULL	2
26	26	16	26	1	4	NULL	NULL	NULL	1
27	27	22	27	22	5	NULL	NULL	NULL	6
28	28	16	28	15	1	NULL	NULL	NULL	2
29	29	3	29	21	4	NULL	NULL	NULL	1

Some transformations I would like to make are:

- Dropping of Tax column as Asadi produces its good locally and does not import thus there is no need to pay for import taxes
- Calculating shipping_price column
- Calculating product_costprice column

Cleaning and Transformation in TIBCO

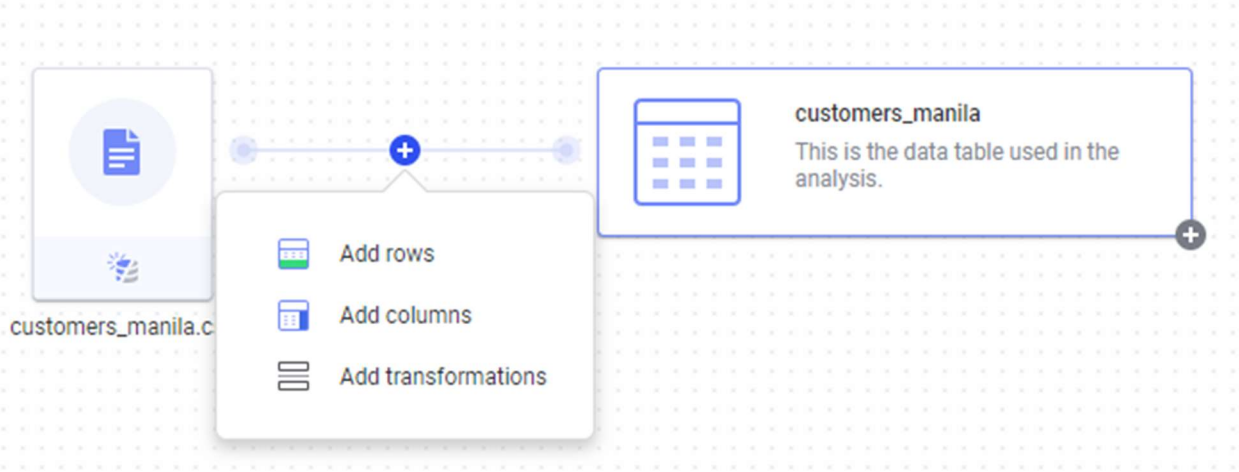
Dataset #1 (customers)



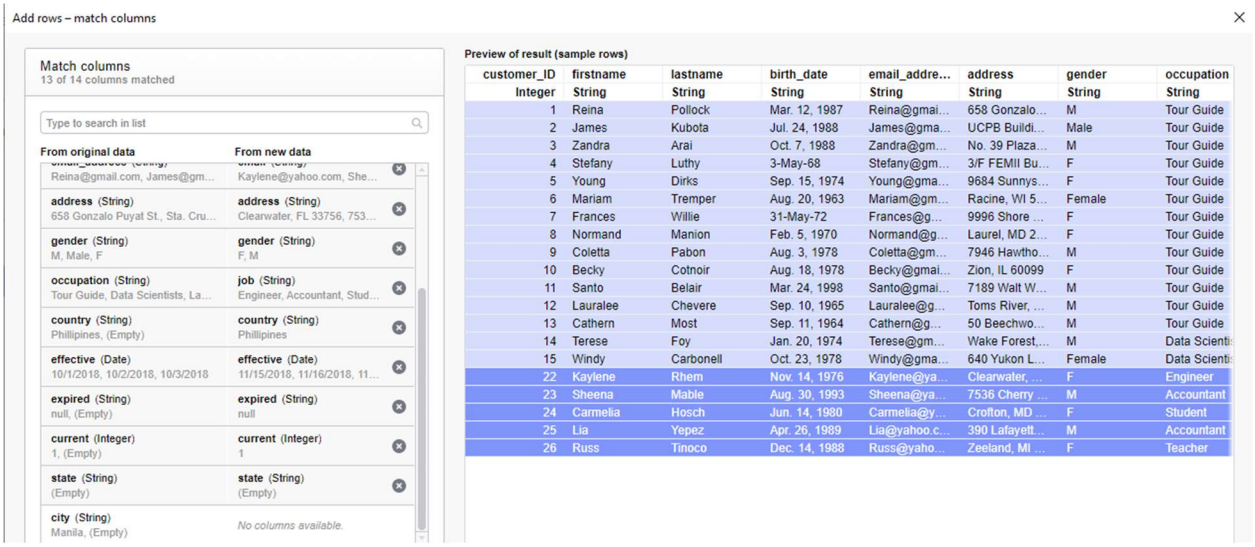
Appending all retail outlets datasets together.

I will start data cleaning in TIBCO by importing the customer datasets first. I will start with the Manila dataset.

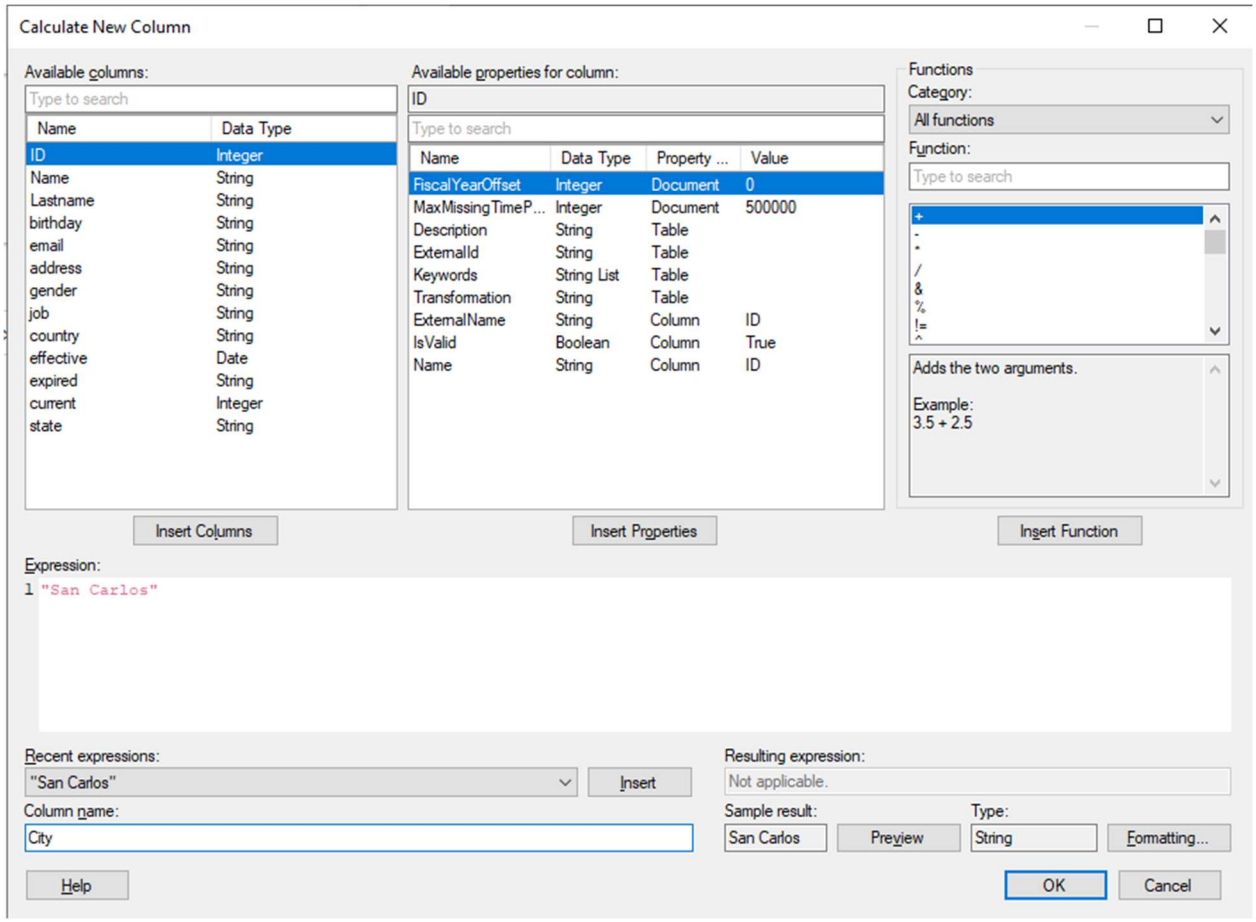
After importing, I will start appending the other datasets from the various retail stores across the Philippines. This is done by clicking on the plus sign and clicking Add rows, > Browse Local Files



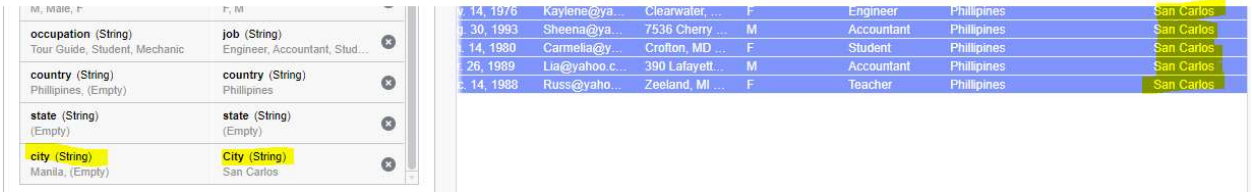
The first data set I will append to customers_Manila is customers_sancarlos. As the column names do not match, I would have to manually match columns as such:



Additionally, the customers_sancarlos data does not have the city column. Thus I will transform the data to **add a new column "City"** which corresponds to the "City" column in the Manila dataset as well.



The results are as shown:



After successfully appending, I take a look at the data in table format. One error I spotted was that between the manila data and san carlos data, there were many **empty rows**. :

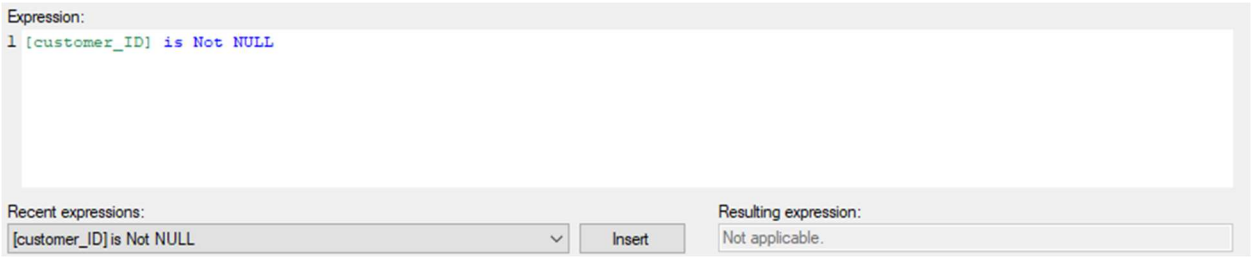
18	Rosena	Fiorenza	Jul. 10, 1992	Rosena@gmai...	Irmo, SC 29063	M	Data Scientists	Phillipines	Manila
19	Kay	Marth	Feb. 18, 1986	Kay@gmail.co...	50 NE. Livings...	M	Lawyer	Phillipines	Manila
20	Armanda	Gilbertson	15-May-62	Armanda@gm...	Aliquippa, PA ...	F	Lawyer	Phillipines	Manila
21	Mi	Wilson	Feb. 29, 1992	Mi@gmail.com	246 Rockaway...	F	Lawyer	Phillipines	Manila

22	Kaylene	Rhem	Nov. 14, 1976	Kaylene@yah...	Clearwater, FL...	F	Engineer	Phillipines	San Carlos
23	Sheena	Mable	Aug. 30, 1993	Sheena@yaho...	7536 Cherry H...	M	Accountant	Phillipines	San Carlos
24	Carmelia	Hosch	Jun. 14, 1980	Carmelia@yah...	Crofton, MD 2...	F	Student	Phillipines	San Carlos

In Filters, I discovered that there were empty values in every column and unticking them would remove all the empty rows between my manila dataset and san carlos data. Therefore, I will try to delete empty values from my dataset.

In the “Data in analysis” tab, I can see that each column has the same amount of empty values which are 429.

To delete empty rows, I add a transformation “Filter Rows” and use the expression:

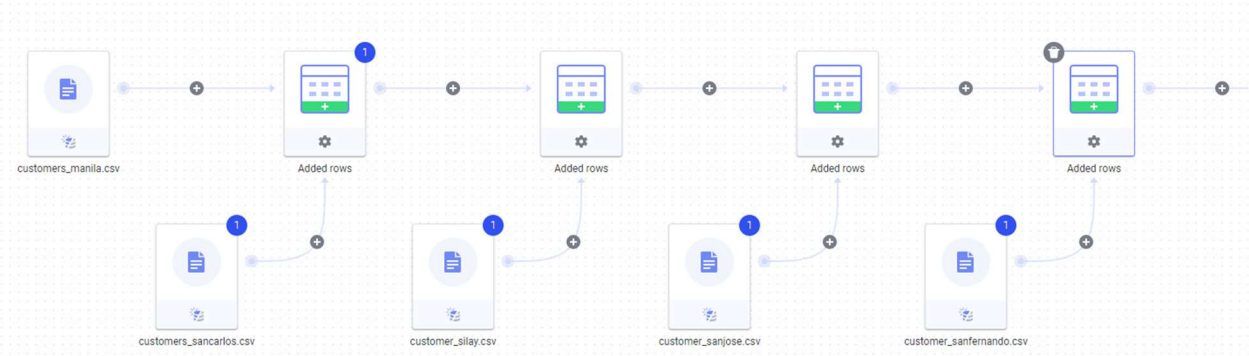


My dataset is now properly appended without empty values in the middle and I have 50 rows of customer data. Now I will continue appending the other datasets from the other retail stores across Philippines.

14	Terese	Foy	Jan. 20, 1974	Terese@gmail...	Wake Forest, ...	M	Data Scientists	Phillipines	Manila
15	Windy	Carbonell	Oct. 23, 1978	Windy@gmail....	640 Yukon Lane	Female	Data Scientists	Phillipines	Manila
16	Bernardo	Randle	Aug. 23, 1963	Bernardo@gm...	Barrington, IL ...	M	Data Scientists	Phillipines	Manila
17	Lyda	Degroot	Mar. 5, 1995	Lyda@gmail.c...	153 Hamilton ...	F	Data Scientists	Phillipines	Manila
18	Rosena	Fiorenza	Jul. 10, 1992	Rosena@gmai...	Irmo, SC 29063	M	Data Scientists	Phillipines	Manila
19	Kay	Marth	Feb. 18, 1986	Kay@gmail.co...	50 NE. Livings...	M	Lawyer	Phillipines	Manila
20	Armanda	Gilbertson	15-May-62	Armanda@gm...	Aliquippa, PA ...	F	Lawyer	Phillipines	Manila
21	Mi	Wilson	Feb. 29, 1992	Mi@gmail.com	246 Rockaway...	F	Lawyer	Phillipines	Manila
22	Kaylene	Rhem	Nov. 14, 1976	Kaylene@yah...	Clearwater, FL...	F	Engineer	Phillipines	San Carlos
23	Sheena	Mable	Aug. 30, 1993	Sheena@yaho...	7536 Cherry H...	M	Accountant	Phillipines	San Carlos
24	Carmelia	Hosch	Jun. 14, 1980	Carmelia@yah...	Crofton, MD 2...	F	Student	Phillipines	San Carlos
25	Lia	Yepez	Apr. 26, 1989	Lia@yahoo.co...	390 Lafayette ...	M	Accountant	Phillipines	San Carlos
26	Russ	Tinoco	Dec. 14, 1988	Russ@yahoo....	Zeeland, MI 4...	F	Teacher	Phillipines	San Carlos

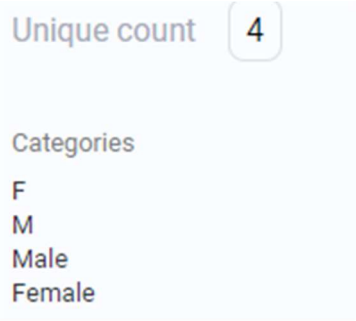
50 of 50 rows

My data canvas after appending all customer datasets from all the retail outlets:

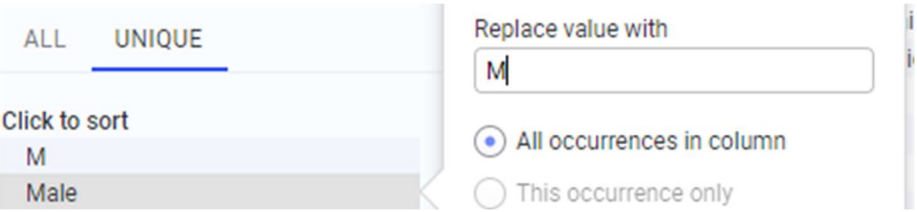


Data inconsistency in Gender column.

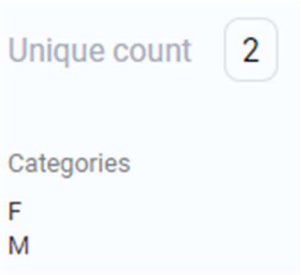
There should only be 2 unique values. M or F. Therefore I will try to change all Male and Female values to M and F.



Under the “Unique” tab, I will select the “Male” value and change all occurrences to the value “M”, and repeat it for the value “Female” to “F”.

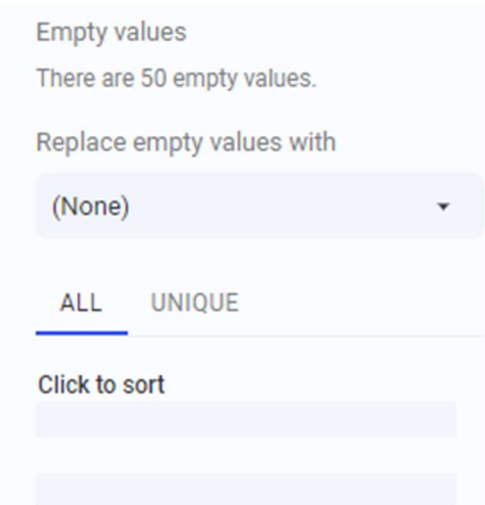


Data in Gender column is now consistent:

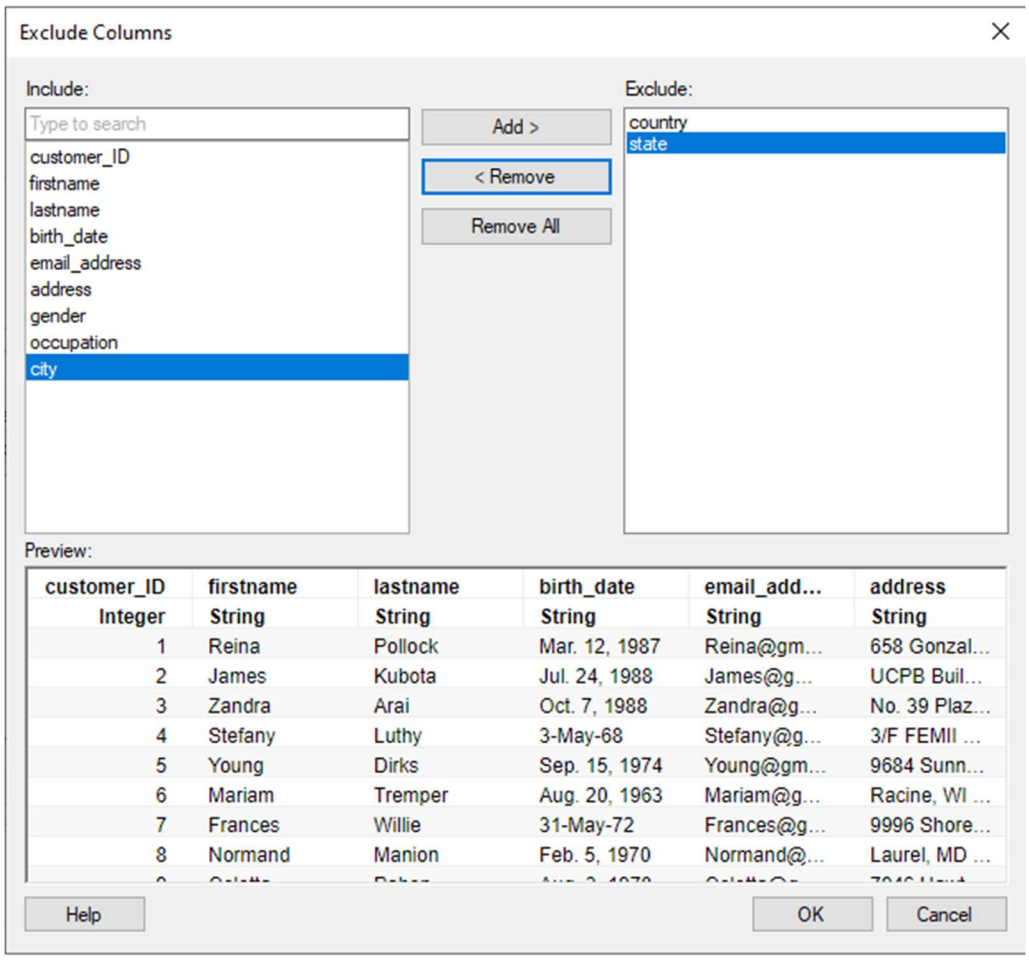


Excluding country and state

I can see that the entire “State” column is empty. It also looks like all customers are based in the Philippines. These columns are not useful in my analysis and thus I will remove it completely.



In my data canvas, I will add the transformation “Exclude Columns”, and exclude column “Country” and “State” as shown below:



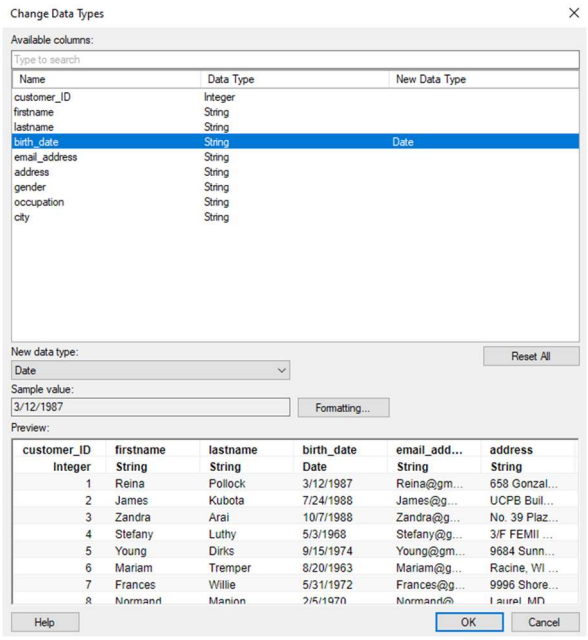
Reformat birth_date column

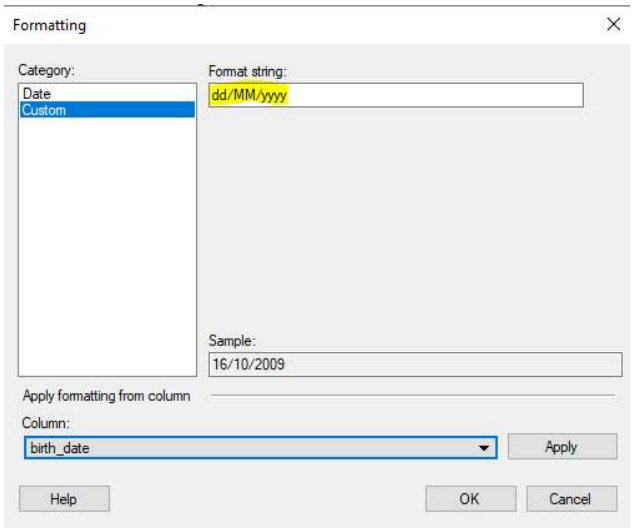
The current data type for birth_date is string and the date format is inconsistent as shown below:

birth_date
String
Mar. 12, 1987
Jul. 24, 1988
Oct. 7, 1988
3-May-68

Therefore I will try to make the date format consistent and change the data type to datetime in DD:MM:YYYY format as that is the normal day-month-year notation Philippines use.

Using the “Change Data Type” transformation in the data canvas, I can select birth_date column and change its data type into Date. Then in “Formatting” I will change the Date notation to DD:MM:YYYY.

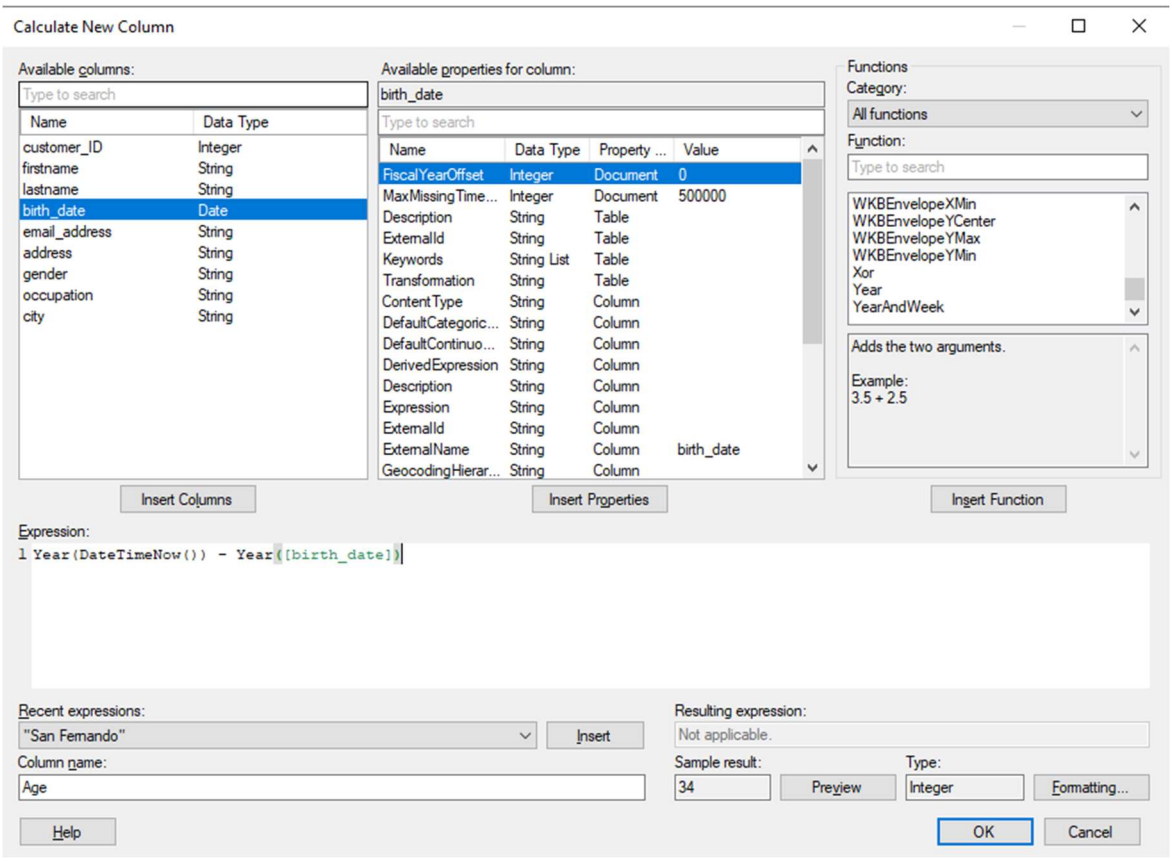




Reformatted birth_date column:

Calculate new column "Age" for analysis

In the data canvas, I can add transformation and use "Calculate a new column". The expression to calculate age from birth_date column is:



Here is the result:

First 100 rows

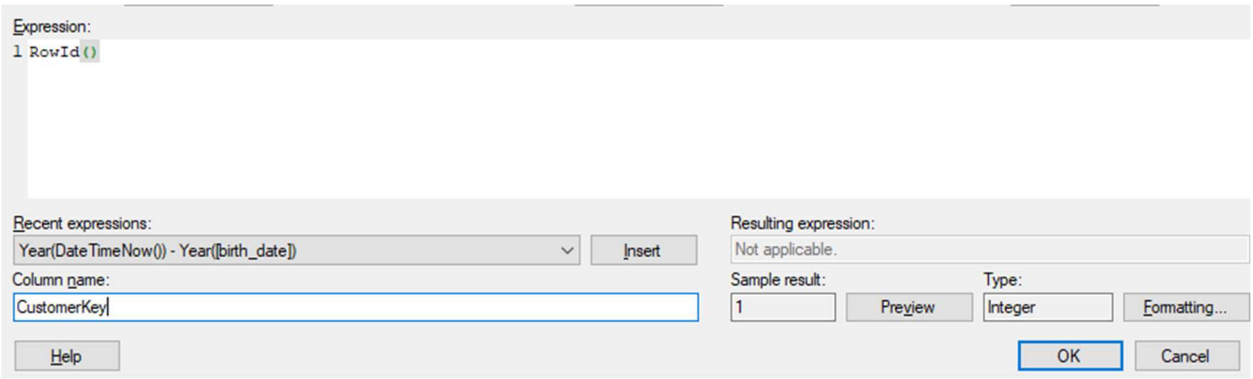
Age	birth_date
Integer	Date
34	12/03/1987
33	24/07/1988
33	07/10/1988
53	03/05/1968

Create surrogate key and insert SCD type 2 dimension

As it is important to capture historical customer information for analysis, a Type 2 Slowly Changing Dimension would be implemented into the customer data. A surrogate key would also be created in the customer data as the data would be later stored in MySQL.

Surrogate key

We can do this by creating a new column an using the function RowID() to generate unique row id numbers for each row.



Result:

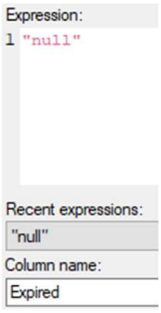
CustomerKey
1
2
3
4
5
6
7
8

Type 2 SCD

In the data canvas, I will add transformation and calculate 3 new columns which are:

Effective:

Expired:



Current:

As the company is only starting to store historical data for its customers, all customer data is assumed to be new. Thus all customer data has a current value of 1 and none has expired. When new customer information is updated, a new row with the updated information would be created.

The type 2 SCD columns are as shown:

Effective	Expired	Current
06/08/2021 7:11:45 ...	null	1
06/08/2021 7:11:45 ...	null	1
06/08/2021 7:11:45 ...	null	1
06/08/2021 7:11:45 ...	null	1

Export to csv

All customer data is appended and cleaned, and I will export as csv file to be loaded into mySQL data warehouse later.

Dataset #2 (products)

Taking a look at the dataset, data cleaning is needed as there are:

- Duplicate rows
- Pricing value for certain items to be high

Adjusting pricing value

A sweatpants does not make sense to cost $7999 and a jersey to be $69999

5	PX5831-43	Sweatpants	Grey	33	casual	pants	7999.00
12	SV2724-07	Ronaldo 7 jersey	White	L	football	shirt	69999.00

Similar jersey's cost $69.99.

Ronaldo 7 jersey	White	L	football	shirt	69999.00
Messi 10 jersey	Purple	S	football	shirt	69.99
Rooney 10 jersey	Red	M	football	shirt	69.99
Jordan 23 jersey	Red	M	basketball	shirt	69.99
Curry 30 jersey	Yellow	S	basketball	shirt	69.99

Therefore,I will adjust the pricing for the jersey to be $69.99 and the sweatpants to be $79.99 as it that is a reasonable price.

Removing duplicate rows

Duplicate rows found in the data are:

16	SV2743-30	Curry 30 jersey
16	SV2743-30	Curry 30 jersey
16	SV2743-30	Curry 30 jersey

20	PX3813-13	Compression tights
20	PX3813-13	Compression tights
20	PX3813-13	Compression tights

11	BS1592-12	Fiba international basketball
11	BS1592-12	Fiba international basketball

An easy remedy for this is to just delete the duplicated rows and leave 1 row. I can do this by highlighting the rows and deleting them:

20	PX3813-13	Compression tights
20	PX3813-13	Compression tights
20	PX3813-13	Compression tights

The data is now cleaned of any duplicated rows.

asadi_products(xlsx) - Sheet1

product_id	product_code	product_name	colour	size	type	category	price
1	AQ3308-06	Air Force 1 Blue	Blue	US9	casual	shoe	250.00
2	AQ3789-72	Curry 7	Black	US7	basketball	shoe	119.99
3	AQ6831-01	Kobe 1	Gold	US11	basketball	shoe	450.00
4	SV3741-17	Dri-fit Shirt	White	M	sport	shirt	59.99
5	PX5831-43	Sweatpants	Grey	33	casual	pants	79.99
6	PX2111-23	Running shorts	Black	28	sport	pants	59.99
7	SV8282-27	Dri-fit Tank-top	White	L	sport	shirt	59.99
8	CP7391-28	Cap	Black	M	casual	accessory	89.99
9	AQ3271-37	Zion 1	Red	US17	basketball	shoe	350.00
10	BS1283-31	Fifa world cup football	White	7	football	ball	99.99
11	BS1592-12	Fiba international basketball	Orange	7	basketball	ball	79.99
12	SV2724-07	Ronaldo 7 jersey	White	L	football	shirt	69.99
13	SV2345-19	Messi 10 jersey	Purple	S	football	shirt	69.99
14	SV8214-10	Rooney 10 jersey	Red	M	football	shirt	69.99
15	SV2840-23	Jordan 23 jersey	Red	M	basketball	shirt	69.99
16	SV2743-30	Curry 30 jersey	Yellow	S	basketball	shirt	69.99
17	SV7339-34	Giannis 34 jersey	Green	XL	basketball	shirt	69.99
18	CP2278-82	Wristband	White	S	sport	accessory	34.99
19	CP3729-12	Arm sleeve phone holder	Black	L	sport	accessory	17.99
20	PX3813-13	Compression tights	Black	30	running	pants	34.99
21	SK3222-12	Arsenal Bag	Red	15L	Sports	Bags	45.00
22	SK3221-13	Adventure Backpack Small	Black	20L	Sports	Bags	60.00
23	SK3452-15	Modern Utility Three-Way Bag	Grey	25L	casual	Bags	130.00

Join data (add new columns)

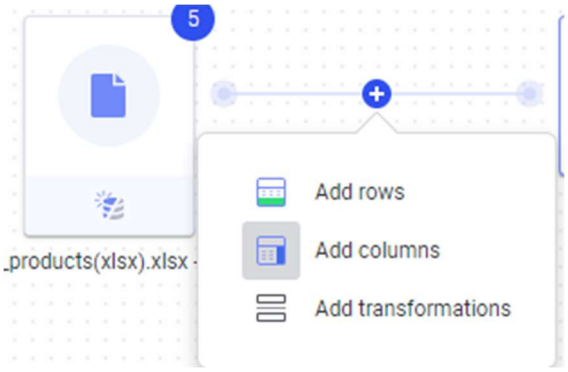
A customer may want to know how much are the shipping costs of their products. As the shipping rates are decided by the product's weight, I will get the weight of each product from Asadi's e-commerce website and join the column with the current product table.

I will also get the cost price of manufacturing each product so to be able to calculate profit made from each product.

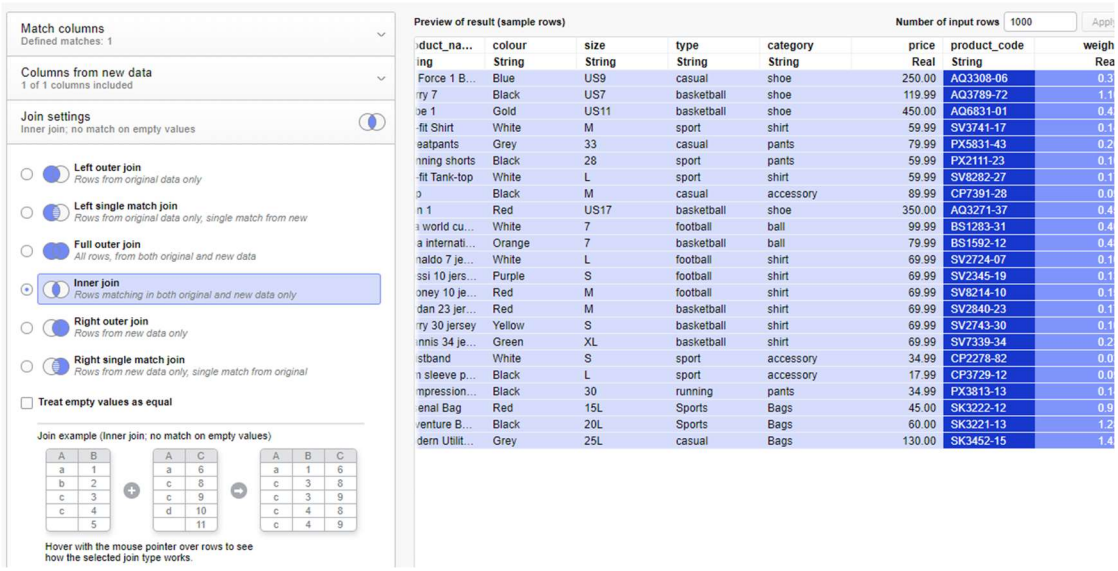
The dataset we will be joining is additional_info_product data. The data consists of the product code, its respective weight in kg and cost price:

product_code	weight	cost_price
AQ3308-06	0.37	50
AQ3789-72	1.1	23.998
AQ6831-01	0.42	90
SV3741-17	0.141748	11.998
PX5831-43	0.198447	15.998
PX2111-23	0.099223	11.998

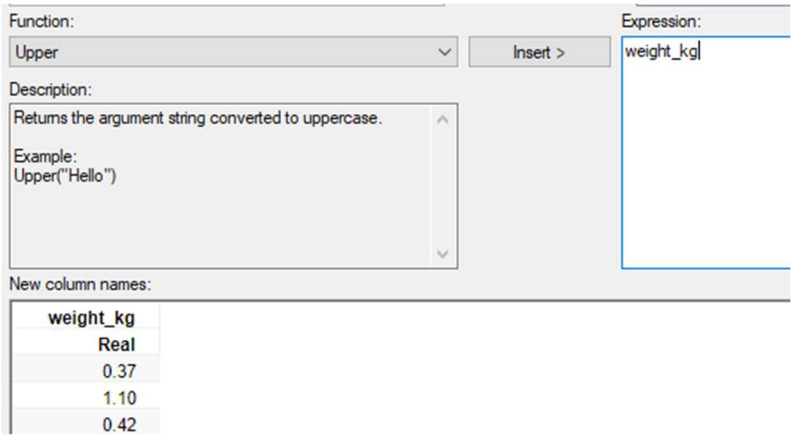
I will first start by adding new columns in my data canvas and open up the product_weight data.



As there is a already a similar product_code column in the original dataset, I will use the appropriate join method which is Inner join which selects rows that match both original and new data.



I will also rename weight column > weight_kg just to be clearer.



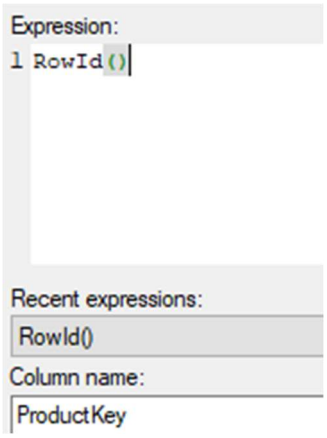
Adding surrogate key and Type 3 Slowly Changing Dimension

As product information does not have frequent changes, a type 3 SCD would be implemented in order for Asadi to track history value without increasing table size as new information is updated.

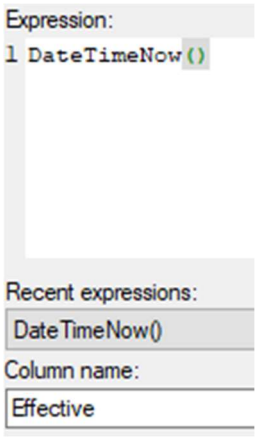
As the dynamic of pricing could affect sales, the SCD would store the historical values of pricing.

The columns to be added for a type 3 SCD are:

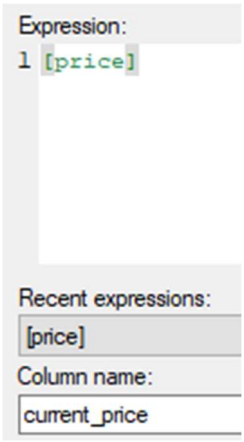
ProductKey(surrogate key):



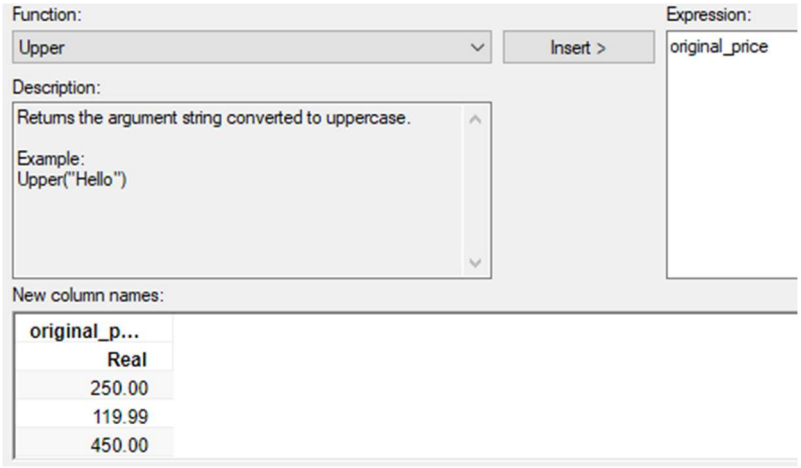
Effective:



original_price:



And original_price which I will be renamed from original price column:



The product table is now cleaned and transformed:

asadi_products(xlsx) - Sheet1

ProductKey	product_id	product_code	product_name	colour	size	type	category	weight_kg	Effective	cost_price	original_price	current_price
1	1	AQ3308-06	Air Force 1 Blue	Blue	US9	casual	shoe	0.37	08/08/2021 1:...	50.00	250.00	250.00
2	2	AQ3789-72	Curry 7	Black	US7	basketball	shoe	1.10	08/08/2021 1:...	24.00	119.99	119.99
3	3	AQ6831-01	Kobe 1	Gold	US11	basketball	shoe	0.42	08/08/2021 1:...	90.00	450.00	450.00
4	4	SV3741-17	Dri-fit Shirt	White	M	sport	shirt	0.14	08/08/2021 1:...	12.00	59.99	59.99
5	5	PX5831-43	Sweatpants	Grey	33	casual	pants	0.20	08/08/2021 1:...	16.00	79.99	79.99
6	6	PX2111-23	Running shorts	Black	28	sport	pants	0.10	08/08/2021 1:...	12.00	59.99	59.99
7	7	SV8282-27	Dri-fit Tank-top	White	L	sport	shirt	0.17	08/08/2021 1:...	12.00	59.99	59.99
8	8	CP7391-28	Cap	Black	M	casual	accessory	0.09	08/08/2021 1:...	18.00	89.99	89.99
9	9	AQ3271-37	Zion 1	Red	US17	basketball	shoe	0.45	08/08/2021 1:...	70.00	350.00	350.00
10	10	BS1283-31	Fifa world cup...	White	7	football	ball	0.40	08/08/2021 1:...	20.00	99.99	99.99
11	11	BS1592-12	Fiba internatio...	Orange	7	basketball	ball	0.48	08/08/2021 1:...	16.00	79.99	79.99
12	11	BS1592-12	Fiba internatio...	Orange	7	basketball	ball	0.48	08/08/2021 1:...	16.00	79.99	79.99
13	12	SV2724-07	Ronaldo 7 jers...	White	L	football	shirt	0.16	08/08/2021 1:...	14.00	69.99	69.99
14	13	SV2345-19	Messi 10 jersey	Purple	S	football	shirt	0.13	08/08/2021 1:...	14.00	69.99	69.99
15	14	SV8214-10	Rooney 10 jer...	Red	M	football	shirt	0.15	08/08/2021 1:...	14.00	69.99	69.99
16	15	SV2840-23	Jordan 23 jers...	Red	M	basketball	shirt	0.17	08/08/2021 1:...	14.00	69.99	69.99
17	16	SV2743-30	Curry 30 jersey	Yellow	S	basketball	shirt	0.16	08/08/2021 1:...	14.00	69.99	69.99
18	16	SV2743-30	Curry 30 jersey	Yellow	S	basketball	shirt	0.16	08/08/2021 1:...	14.00	69.99	69.99
19	16	SV2743-30	Curry 30 jersey	Yellow	S	basketball	shirt	0.16	08/08/2021 1:...	14.00	69.99	69.99
20	17	SV7339-34	Giannis 34 jer...	Green	XL	basketball	shirt	0.23	08/08/2021 1:...	14.00	69.99	69.99
21	18	CP2278-82	Wristband	White	S	sport	accessory	0.03	08/08/2021 1:...	7.00	34.99	34.99
22	19	CP3729-12	Arm sleeve ph...	Black	L	sport	accessory	0.09	08/08/2021 1:...	3.60	17.99	17.99
23	20	PX3813-13	Compression ...	Black	30	running	pants	0.14	08/08/2021 1:...	7.00	34.99	34.99
24	20	PX3813-13	Compression ...	Black	30	running	pants	0.14	08/08/2021 1:...	7.00	34.99	34.99
25	20	PX3813-13	Compression ...	Black	30	running	pants	0.14	08/08/2021 1:...	7.00	34.99	34.99
26	21	SK3222-12	Arsenal Bag	Red	15L	Sports	Bags	0.91	08/08/2021 1:...	9.00	45.00	45.00
27	22	SK3221-13	Adventure Bac...	Black	20L	Sports	Bags	1.28	08/08/2021 1:...	12.00	60.00	60.00
28	23	SK3452-15	Modern Utility ...	Grey	25L	casual	Bags	1.42	08/08/2021 1:...	26.00	130.00	130.00

I will export it and save it as products_asadi in xlsx format to be loaded into mySQL later on.

<u>Dataset #3 (time)</u>

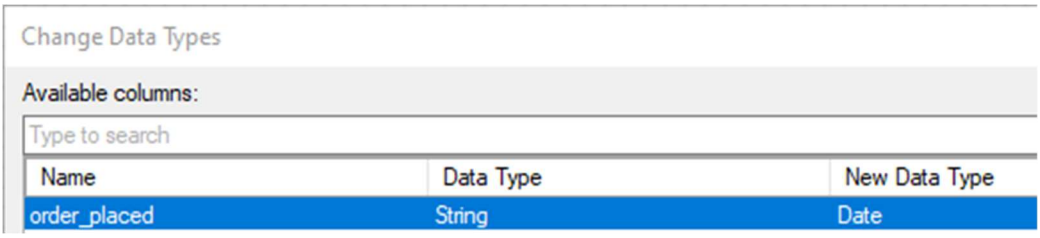
Taking a look at the dataset, data cleaning is needed for:

order_placed	delivered date	year	quarter	month
1/1/2020	1/4/2020	2020		January
1/2/2020	1/5/2020	2020		January
3rd January 2...	1/6/2020	2020		January
1/4/2020	1/7/2020	2020		January
1/5/2020	1/8/2020	2020		January
1/6/2020		2020		January
1/7/2020	1/10/2020	2020		January
1/8/2020	1/11/2020	2020		January
1/9/2020		2020		January
1/10/2020	1/13/2020	2020		January
1st February 2...	2/4/2020	2020		feb
2/2/2020	2/5/2020	2020		feb
2/3/2020	2/6/2020	2020		feb
2/4/2020	2/7/2020	2020		feb
2/5/2020		2020		feb

- Incorrect datetime format in order_placed and delivered_date columns
- Missing data in delivered_date column
- Inconsistency in Month naming
- Empty column (Quarter)

I will start by importing the txt file into TIBCO as a new data table.

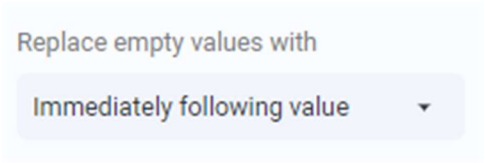
Changing order_placed column dtype to Date and filling in missing values



As shown above, I changed the data type of the column was originally a string. After changing, missing values were introduced in the row as there were originally wrong formatting of dates in the column.

There are about 4 missing values after changing the data type:

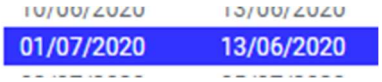
Looking at the dataset, it looks like there were orders placed almost every single day and each order placed is not far apart from each. Only a day or two away. Therefore, I feel that it is appropriate for me to replace the missing values with the value immediately after:



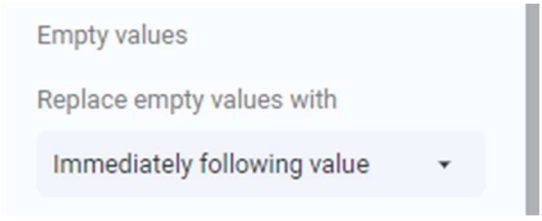
Filling in missing value for delivered_date:

I could not replace the missing values with the value immediately before as after looking through my data, there would be records which would not make sense such as this:

Where order_placed date is later than delivered_date.



Therefore for the missing values of delivered_date, replacing the empty values with the value immediately after was suitable as checking through my data, all delivered_date are dated few days after dates in order_placed which was what we want.



Calculating empty quarter column

quarter	mc
	Ja
	Ja
	Ja
	Ja
	Ja
	Ja
	Ja
	Ja
	Ja
	Ja

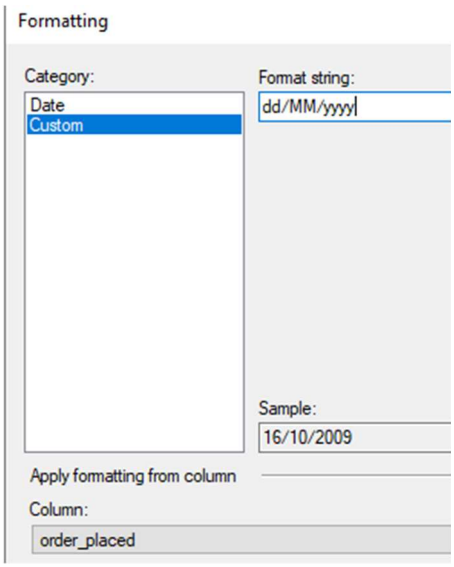
The quarter column was populated with the respective quarter by using the expression below:

```
case
when [month] = "January" then "Q1"
when [month] = "February" then "Q1
when [month] = "March" then "Q1"
when [month] = "April" then "Q2"
when [month] = "May" then "Q2"
when [month] = "June" then "Q2"
when [month] = "July" then "Q3"
when [month] = "August" then "Q3"
when [month] = "September" then "Q3"
when [month] = "October" then "Q4"
when [month] = "November" then "Q4"
when [month] = "December" then "Q4"
ELSE "0"
end
```

Changing delivered_date format to dd/MM/yyyy

The data is currently in M/D/YYYY format and to standardize the formatting with order_placed column, I will change the formatting to dd/MM/yyyy

Configuration:



Before:

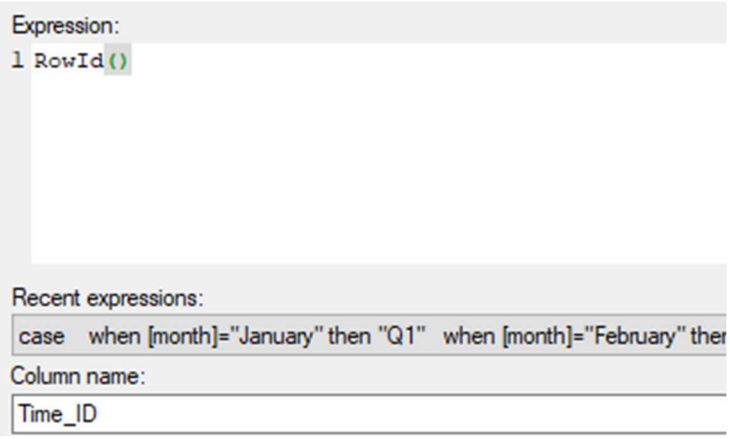
delivered date
1/4/2020
1/5/2020
1/6/2020
1/7/2020
1/8/2020
1/8/2020
1/10/2020
1/11/2020

After:

order_placed	delivered date
01/01/2020	04/01/2020
02/01/2020	05/01/2020
04/01/2020	06/01/2020
04/01/2020	07/01/2020
05/01/2020	08/01/2020

Adding index column Time_ID

I will create a new column called Time_ID and create unique row IDs for each row as shown:



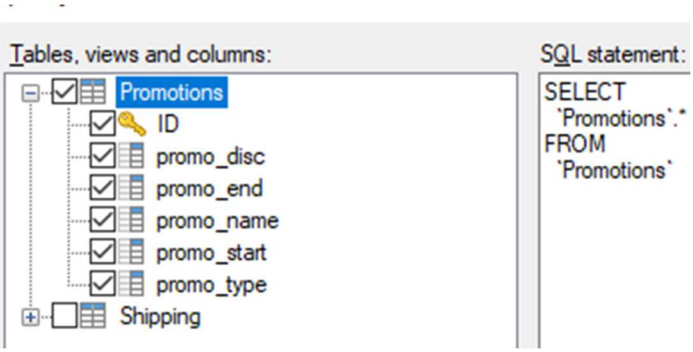
Export cleaned data to csv

The time dataset is now cleaned and transformed and can be exported into csv as time_orders_asadi.csv.

Dataset #4 (Promotions)

Format: MS Access

I will now import the promotions data from MS Access into TIBCO and name it asadi_promotions. The shipping data is also stored in MS Access, however I will import it separately later on.



Taking a look at the promo dataset, it is relatively clean:

asadi_promotions

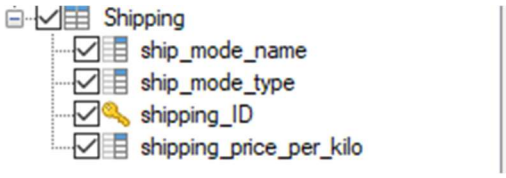
ID	promo_name	promo_type	promo_end	promo_start	promo_disc
1	12.12 Sale	12% off all items	12/14/2020 12:00:00 AM	12/12/2020 12:00:00 AM	12
2	Back 2 School	30% off all bags and sh...	4/15/2021 12:00:00 AM	3/1/2021 12:00:00 AM	30
3	Mid-Year Sale	30% off all items	7/1/2020 12:00:00 AM	6/1/2020 12:00:00 AM	30
6	8.8 National Day Sale	Discounts of $50, min ...	8/10/2020 12:00:00 AM	8/8/2020 12:00:00 AM	50

I would like to make use of the promo_disc column to calculate a discount column which could be useful in analysis. However, I will perform when I have loaded all data into the data-warehouse. (mySQL)

Dataset #5 (Shipping)

Format: MS Access

I will now import the promotions data from MS Access into TIBCO and name it asadi_shipping.



Taking a look at the shipping dataset, it is also relatively clean:

asadi_shipping

shipping_ID	ship_mode_name	ship_mode_type	shipping_price_per_kilo
1	Fedex	Air	5.00
2	UPS	Sensitive Air	7.00
3	DHL	Sensitive Air	7.00
4	Ninjavan	Land	2.50
5	J&T Express	Land	2.50
6	Ocean Network Express	Sensitive Sea	0.75
7	Orient Overseas Container Line	Sea	0.50

I would like to make use of the column shipping_price_per_kilo to calculate a shipping_cost column for the different products which could be useful for customers to see how much it would cost to ship certain items. I will perform this when I have loaded the data into the data-warehouse (mySQL).

Dataset #6 (fact_orders)

id	order_id	customerkey	timekey	productkey	shippingkey	product_costprice	shipping_price	tax	promokey
1	1	16	1	23	3	NULL	NULL	NULL	1
2	2	32	2	6	5	NULL	NULL	NULL	3
3	3	15	3	6	6	NULL	NULL	NULL	6
4	4	45	4	11	4	NULL	NULL	NULL	2
5	5	36	5	18	5	NULL	NULL	NULL	1
6	6	30	6	4	7	NULL	NULL	NULL	6
7	7	15	7	18	2	NULL	NULL	NULL	3
8	8	1	8	16	6	NULL	NULL	NULL	2
9	9	47	9	18	2	NULL	NULL	NULL	6
10	10	16	10	13	4	NULL	NULL	NULL	6
11	11	3	11	18	3	NULL	NULL	NULL	1
12	12	47	12	7	2	NULL	NULL	NULL	2
13	13	39	13	8	2	NULL	NULL	NULL	3
14	14	12	14	14	4	NULL	NULL	NULL	1
15	15	1	15	11	6	NULL	NULL	NULL	6
16	16	43	16	2	6	NULL	NULL	NULL	3
17	17	37	17	9	6	NULL	NULL	NULL	2
18	18	26	18	9	6	NULL	NULL	NULL	1
19	19	3	19	17	7	NULL	NULL	NULL	6
20	20	33	20	9	7	NULL	NULL	NULL	3
21	21	13	21	2	1	NULL	NULL	NULL	1
22	22	7	22	12	1	NULL	NULL	NULL	1
23	23	20	23	10	4	NULL	NULL	NULL	2
24	24	43	24	16	6	NULL	NULL	NULL	3
25	25	26	25	1	3	NULL	NULL	NULL	2
26	26	16	26	1	4	NULL	NULL	NULL	1
27	27	22	27	22	5	NULL	NULL	NULL	6
28	28	16	28	15	1	NULL	NULL	NULL	2
29	29	3	29	21	4	NULL	NULL	NULL	1

Data cleaning needed:

- Drop column Taxes as all goods of Asadi are produced in Philippines itself and they do not have to pay tax for importing goods.
- Calculating shipping_price

Firstly, I will export this data from mySQL to csv in order to perform data cleaning and transformation.

Populate product_costprice column

Inserting a new column called product_code and

taking the respective columns(product_key, product_code) from products_asadi xlsx file, using =VLOOKUP(), I will populate the product_code column and ensure that it corresponds to the correct product_key used.

Columns from products_asadi: =VLOOKUP method used

A	B	C
ProductKey	product_id	product_code
1	1	AQ3308-06
2	2	AQ3789-72
3	3	AQ6831-01
4	4	SV3741-17
5	5	PX5831-43
6	6	PX2111-23
7	7	SV8282-27
8	8	CP7391-28
9	9	AQ3271-37
10	10	BS1283-31
11	11	BS1592-12
12	12	SV2724-07
13	13	SV2345-19
14	14	SV8214-10
15	15	SV2840-23
16	16	SV2743-30
17	17	SV7339-34
18	18	CP2278-82
19	19	CP3729-12
20	20	PX3813-13
21	21	SK3222-12
22	22	SK3221-13
23	23	SK3452-15

productkey	shippingkey	product_code	product_costprice
2	7	=VLOOKUP(E2,$E$33:$F$56,2,FALSE)	
18	1		6.998
1	5		50
7	6		11.998
20	4		6.998
5	5		15.998
3	1		90
2	3		23.998
14	4		13.998
4	3		11.998
8	2		17.998
20	5		6.998
9	1		70
15	7		13.998
19	3		3.598
5	3		15.998
11	4		15.998
10	1		19.998
9	1		70
19	5		3.598
18	5		6.998
10	2		19.998
4	1		11.998
9	2		70
18	4		6.998
4	1		11.998
11	2		15.998
22	5		12
20	3		6.998
21	2		9

product_key	product_code		product_costprice
1	AQ3308-06		50
2	AQ3789-72		23.998
3	AQ6831-01		90

Once done, I can now populate product_costprice column by using the same VLOOKUP function and using the table array below:

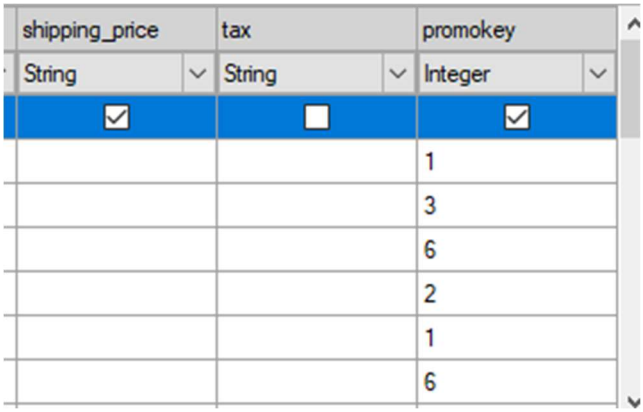
=VLOOKUP method used:

	shippingkey	product_code	product_costprice	shipping_price
2	7	AQ3789-72	=VLOOKUP(G2,$F$33:$G$56,2,FALSE)	
18	1	CP2278-82	VLOOKUP(lookup_value, **table_array**, col_in	
1	5	AQ3308-06	50	
7	6	SV8282-27	11.998	
20	4	PX3813-13	6.998	
5	5	PX5831-43	15.998	
3	1	AQ6831-01	90	
2	3	AQ3789-72	23.998	
14	4	SV8214-10	13.998	
4	3	SV3741-17	11.998	
8	2	CP7391-28	17.998	
20	5	PX3813-13	6.998	
9	1	AQ3271-37	70	
15	7	SV2840-23	13.998	
19	3	CP3729-12	3.598	
5	3	PX5831-43	15.998	
11	4	BS1592-12	15.998	
10	1	BS1283-31	19.998	
9	1	AQ3271-37	70	
19	5	CP3729-12	3.598	
18	5	CP2278-82	6.998	
10	2	BS1283-31	19.998	
4	1	SV3741-17	11.998	
9	2	AQ3271-37	70	
18	4	CP2278-82	6.998	
4	1	SV3741-17	11.998	
11	2	BS1592-12	15.998	
22	5	SK3221-13	12	
20	3	PX3813-13	6.998	
21	2	SK3222-12	9	

y	product_code	product_costpr ce
1	AQ3308-06	50
2	AQ3789-72	23.998

Now that the product_costprice column is populated, I will delete the product_code column as it is of no use anymore. And import the data set into TIBCO to calculate the shipping_price column and drop tax column.

In the configuration window when importing, I will untick tax column to exclude it from the data:



To calculate shipping column I would need to take a look at the product_key column and shipping_key column. Shipping cost is calculated by weight_kg(product table)*ship_price_per_kilo(shipping table):

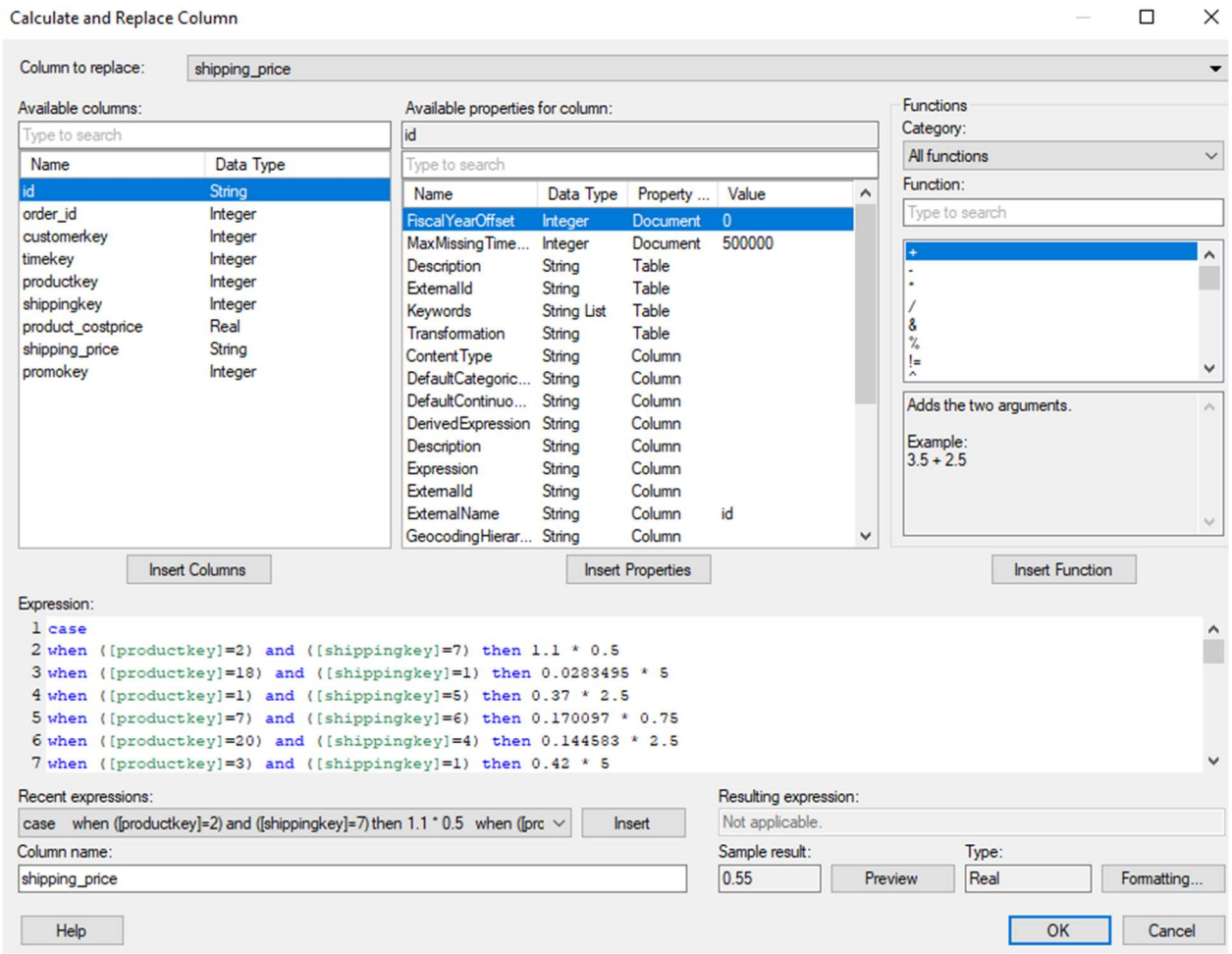
asadi_shipping

shipping_ID	ship_mode_n...	ship_mode_ty...	shipping_price_per_...
1	Fedex	Air	5.00
2	UPS	Sensitive Air	7.00
3	DHL	Sensitive Air	7.00
4	Ninjavan	Land	2.50
5	J&T Express	Land	2.50
6	Ocean Networ...	Sensitive Sea	0.75
7	Orient Overse...	Sea	0.50

asadi_products(xlsx) - Sheet1

product_id	product_code	product_name	colour	size	type	category	original_price	weight_kg
1	AQ3308-06	Air Force 1 Blue	Blue	US9	casual	shoe	250.00	0.37
2	AQ3789-72	Curry 7	Black	US7	basketball	shoe	119.99	1.10

Therefore in the data canvas for fact_orders data, I will insert a transformation, and calculate and replace the values for shipping_price using the above two information as a guide, here is part of the expression:



After inserting the column shipping_price is populated and ready to be exported and imported back into mySQL datawarehouse.

id	order_id	customerkey	timekey	productkey	shippingkey	product_costprice	shipping_price	promokey
1	1	16	1	2	7	23.998	0.55000000	1
2	2	28	2	18	1	6.998	0.14174750	3
3	3	34	3	1	5	50.000	0.92500000	6
4	4	48	4	7	6	11.998	0.12757275	2
5	5	2	5	20	4	6.998	0.36145750	1
6	6	29	6	5	5	15.998	0.49611750	6
7	7	19	7	3	1	90.000	2.10000000	3
8	8	23	8	2	3	23.998	7.70000000	2
9	9	3	9	14	4	13.998	0.37563000	6
10	10	25	10	4	3	11.998	0.99223600	6
11	11	28	11	8	2	17.998	0.59534020	1
12	12	35	12	20	5	6.998	0.36145750	2
13	13	11	13	9	1	70.000	2.26796000	3

Importing cleaned data into mySQL data warehouse

Create database

Before starting, I will create a database called asadi_dw.

Using SQLizer to convert files into SQL databases

Importing cleaned data (products_asadi)

SQLizer

Easily convert files into SQL databases

All done!

Your file was converted successfully.

Download: asadi_products.sql

```
CREATE TABLE IF NOT EXISTS asadi_products (
    `product_id` INT,
    `product_code` VARCHAR(9) CHARACTER SET utf8,
    `product_name` VARCHAR(29) CHARACTER SET utf8,
    `colour` VARCHAR(6) CHARACTER SET utf8,
    `size` VARCHAR(4) CHARACTER SET utf8,
    `type` VARCHAR(10) CHARACTER SET utf8,
    `category` VARCHAR(9) CHARACTER SET utf8,
    `original_price` NUMERIC(5, 2),
    `weight_kg` NUMERIC(8, 7),
    `ProductKey` INT,
    `Effective` DATETIME,
    `current_price` NUMERIC(5, 2)
);
INSERT INTO asadi_products VALUES
    (1,'AQ3308-06','Air Force 1 Blue','Blue','US9','casual','shoe',250,0.37,1,'2021-07-08 00:12:48',250),
    (2,'AQ3789-72','Curry 7','Black','US7','basketball','shoe',119.99,1.1,2,'2021-07-08 00:12:48',119.99),
    (3,'AQ6831-01','Kobe 1','Gold','US11','basketball','shoe',450,0.42,3,'2021-07-08 00:12:48',450),
    (4,'SV3741-17','Dri-fit Shirt','White','M','sport','shirt',59.99,0.141748,4,'2021-07-08 00:12:48',59.99),
    (5,'PX5831-43','Sweatpants','Grey','33','casual','pants',79.99,0.198447,5,'2021-07-08 00:12:48',79.99),
```

Inserting the created script into mySQL:

```
CREATE TABLE IF NOT EXISTS asadi_products (
    `product_id` INT,
    `product_code` VARCHAR(9) CHARACTER SET utf8,
    `product_name` VARCHAR(29) CHARACTER SET utf8,
    `colour` VARCHAR(6) CHARACTER SET utf8,
    `size` VARCHAR(4) CHARACTER SET utf8,
    `type` VARCHAR(10) CHARACTER SET utf8,
    `category` VARCHAR(9) CHARACTER SET utf8,
    `original_price` NUMERIC(5, 2),
    `weight_kg` NUMERIC(8, 7),
    `ProductKey` INT,
    `Effective` DATETIME,
    `current_price` NUMERIC(5, 2)
);
INSERT INTO asadi_products VALUES
    (1,'AQ3308-06','Air Force 1 Blue','Blue','US9','casual','shoe',250,0.37,1,'2021-07-08 00:12:48',250),
    (2,'AQ3789-72','Curry 7','Black','US7','basketball','shoe',119.99,1.1,2,'2021-07-08 00:12:48',119.99),
    (3,'AQ6831-01','Kobe 1','Gold','US11','basketball','shoe',450,0.42,3,'2021-07-08 00:12:48',450),
    (4,'SV3741-17','Dri-fit Shirt','White','M','sport','shirt',59.99,0.141748,4,'2021-07-08 00:12:48',59.99),
    (5,'PX5831-43','Sweatpants','Grey','33','casual','pants',79.99,0.198447,5,'2021-07-08 00:12:48',79.99),
    (6,'PX2111-23','Running shorts','Black','28','sport','pants',59.99,0.0992233,6,'2021-07-08 00:12:48',59.99),
    (7,'SV8282-27','Dri-fit Tank-top','White','L','sport','shirt',59.99,0.170097,7,'2021-07-08 00:12:48',59.99),
    (8,'CP7391-28','Cap','Black','M','casual','accessory',89.99,0.0850486,8,'2021-07-08 00:12:48',89.99),
    (9,'AQ3271-37','Zion 1','Red','US17','basketball','shoe',350,0.453592,9,'2021-07-08 00:12:48',350),
    (10,'BS1283-31','Fifa world cup football','White','7','football','ball',99.99,0.396893,10,'2021-07-08 00:12:48',99.99),
```

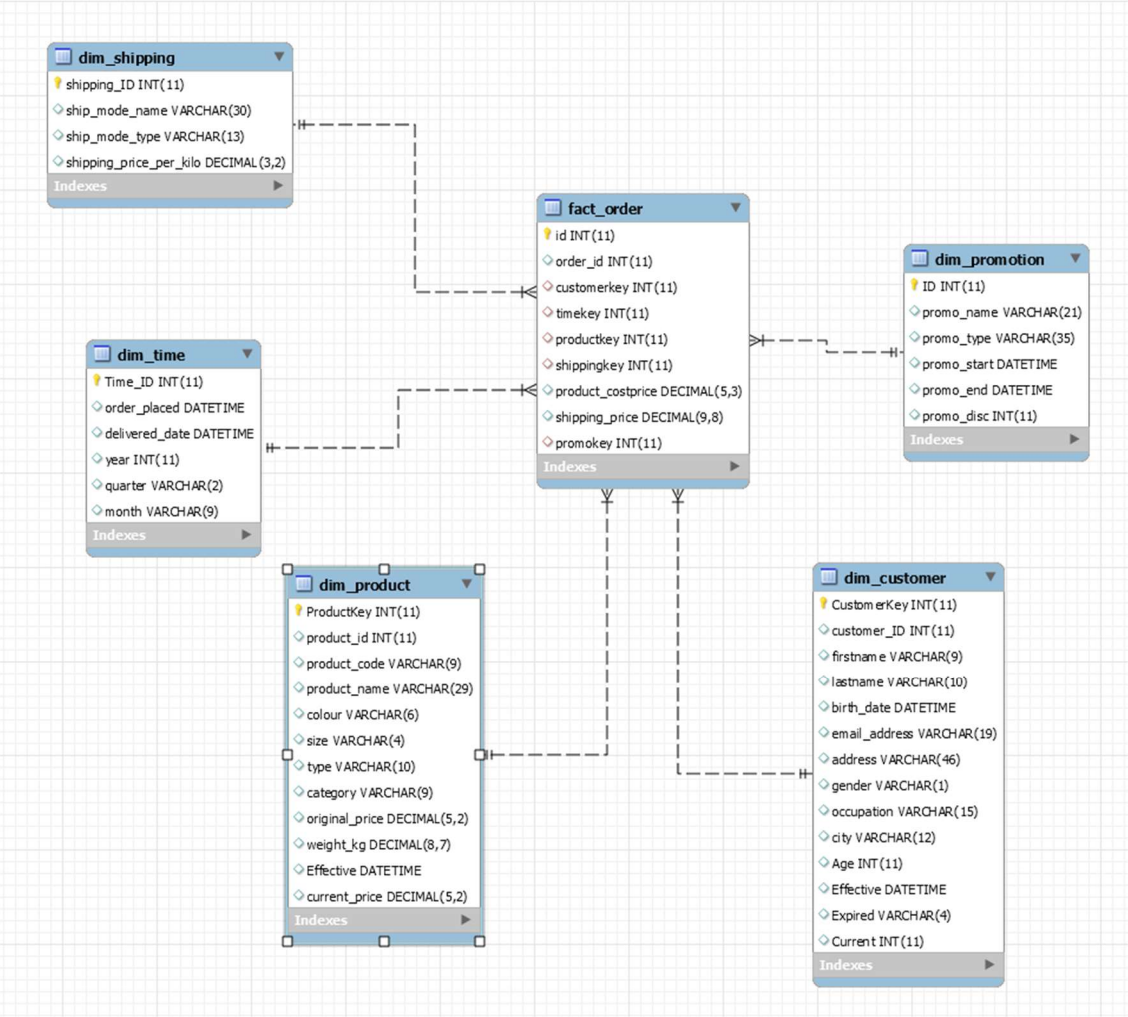
Result:

product_id	product_code	product_name	colour	size	type	category	original_price	weight_kg	ProductKey	Effective	current_price
1	AQ3308-06	Air Force 1 Blue	Blue	US9	casual	shoe	250.00	0.3700000	1	2021-07-08 00:12:48	250.00
2	AQ3789-72	Curry 7	Black	US7	basketball	shoe	119.99	1.1000000	2	2021-07-08 00:12:48	119.99
3	AQ6831-01	Kobe 1	Gold	US11	basketball	shoe	450.00	0.4200000	3	2021-07-08 00:12:48	450.00
4	SV3741-17	Dri-fit Shirt	White	M	sport	shirt	59.99	0.1417480	4	2021-07-08 00:12:48	59.99
5	PX5831-43	Sweatpants	Grey	33	casual	pants	79.99	0.1984470	5	2021-07-08 00:12:48	79.99
6	PX2111-23	Running shorts	Black	28	sport	pants	59.99	0.0992233	6	2021-07-08 00:12:48	59.99
7	SV8282-27	Dri-fit Tank-top	White	L	sport	shirt	59.99	0.1700970	7	2021-07-08 00:12:48	59.99
8	CP7391-28	Cap	Black	M	casual	accessory	89.99	0.0850486	8	2021-07-08 00:12:48	89.99
9	AQ3271-37	Zion 1	Red	US17	basketball	shoe	350.00	0.4535920	9	2021-07-08 00:12:48	350.00
10	BS1283-31	Fifa world cup football	White	7	football	ball	99.99	0.3968930	10	2021-07-08 00:12:48	99.99

After importing the data, I will set ProductKey as the primary key since the data is a type 3 SCD. I will also do the same for customer_asadi dataset where it uses a type 2 SCD. I will set CustomerKey column as the primary key.

I will now repeat these steps to continue loading the other cleaned data in. For data in MS Access which are the promotion and shipping data, I will use TIBCO to export into csv first so as to be compatible for converting using the SQLizer website.

Here is the final result:



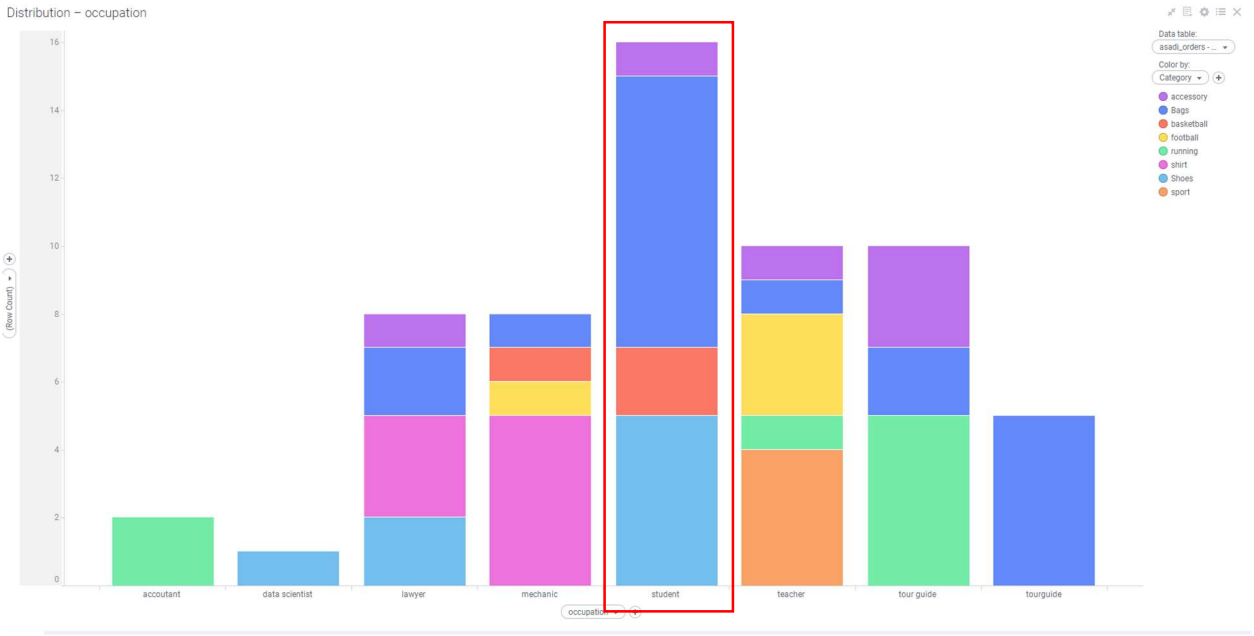
Visualizations

Problem Statement

Asadi would like to increase revenue through cross-selling.

What are the products that are bought frequently together?

Distribution of product category bought by occupation

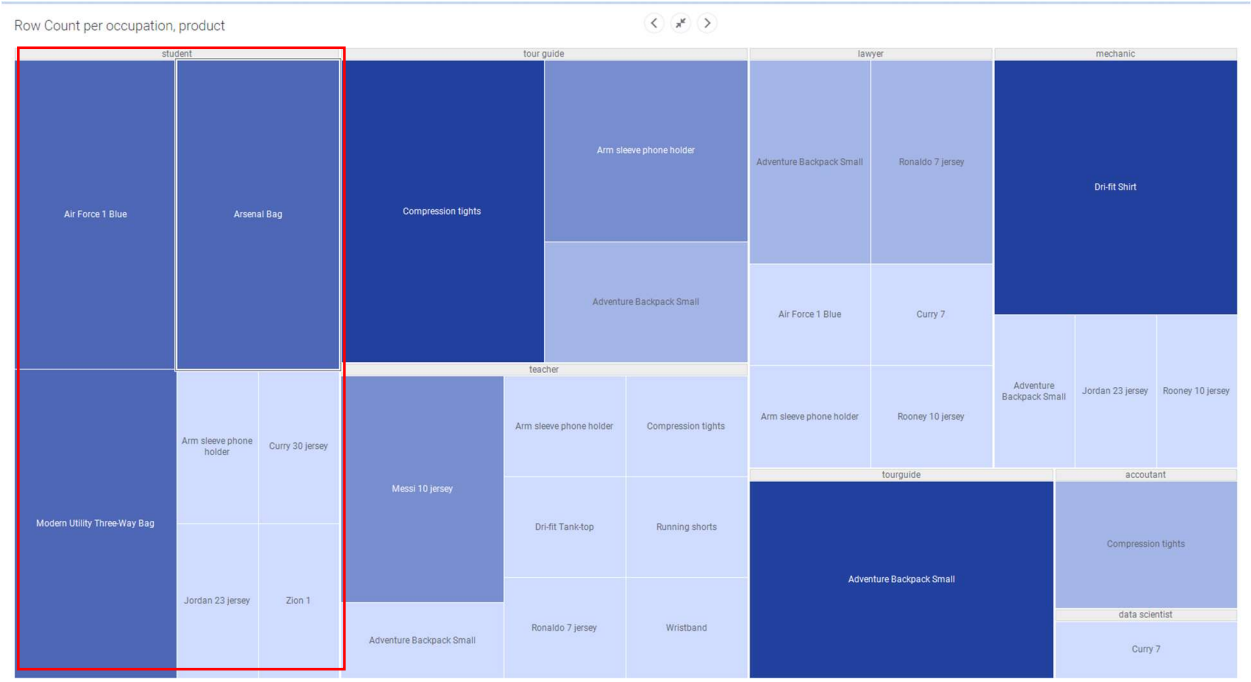


From the stacked-bar chart, we can generally see that students are the most frequent customer for Asadi, followed by teachers, and tour guides.

In the student bar, we can see that the 2 most popular category that students purchase are Bags(dark blue) and Shoes(light blue).

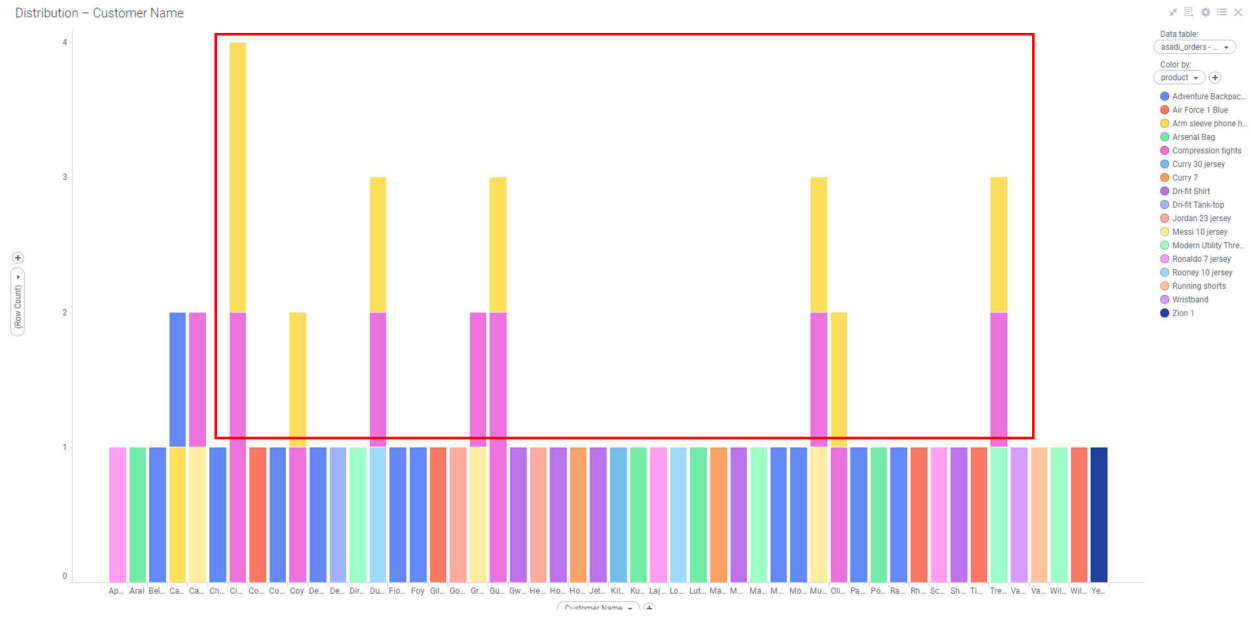
The takeaway for Asadi from this chart is – Asadi increase revenue by targeting students(most frequent type of customer) by coming up with promotions that include bags and shoes together. In the shipping table, they already have a Back 2 School promotional discount and selling these 2 category products could be implemented in the promotion to attract students.

Treemap of occupation by product



We can take a deeper look at the type of products that students purchase here. On the left top corner, we can see that the shoe and bags that students purchase together are the “Air Force 1 Blue”, “Arsenal Bag” and “Modern Utility Three-Way Bag”. These shoes and bags are popular amongst students and thus, a promotion that sells these items together could increase revenue.

Distribution customer by product



One obvious observation here is that customers who bought compression tights, have also bought an Arm Sleeve phone holder. We can see this by the yellow bar(Arm Sleeve phone holder) and pink bars(Compression tights) being purchased by the same customers.

Therefore, the takeaway for Asadi would be to place these two items together in their retail outlets as these items are normally bought one after the other. Customers who see these two items together have a higher chance of buying both, thus increasing revenue for Asadi.

<u>Focusing by highlighting tour guide bar</u>

To be more in depth, customers who purchase Compression Tights and Arm Sleeve Phone Holders are mainly tour guides. This makes sense as tour guides are often out and about. With this information, Asadi can create more products that cater to tour guides