Initial Data

Original Data in Google Sheets:

order_id	order_date	ship_date	ship_mode	customer_name	segment	state	country	market	region	product_id	category	sub_category	product_name	sales	quantity	discount	profit	shipping_cost	order_priority	year
AG-2011-2040	1/1/2011	1/6/2011	Standard Class	Toby Braunhardt	Consumer	Constantine	Algeria	Africa	Africa	OFF-TEN-10000025	Office Supplies	Storage	Tenex Lockers, Blue	408	2	0	138.14	35.46	Medium	2011
IN-2011-47883	1/1/2011	1/8/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	OFF-SU-10000618	Office Supplies	Supplies	Acme Trimmer, High Speed	120	3	0.1	36.036	9.72	Medium	2011
HU-2011-1220	1/1/2011	1/5/2011	Second Class	Annie Thurman	Consumer	Budapest	Hungary	EMEA	EMEA	OFF-TEN-10001585	Office Supplies	Storage	Tenex Box, Single Width	66	4	0	29.64	8.17	High	2011
IT-2011-3647632	1/1/2011	1/5/2011	Second Class	Eugene Moren	Home Office	Stockholm	Sweden	EU	North	OFF-PA-10001492	Office Supplies	Paper	Enermax Note Cards, Premium	45	3	0.5	-26.055	4.82	High	2011
IN-2011-47883	1/1/2011	1/8/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	FUR-FU-10001447	Furniture	Furnishings	Eldon Light Bulb, Duo Pack	114	5	0.1	37.77	4.70	Medium	2011
IN-2011-47883	1/1/2011	1/8/2011	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	OFF-PA-10001968	Office Supplies	Paper	Eaton Computer Printout Paper, 8.5 x 11	55	2	0.1	15.342	1.80	Medium	2011
CA-2011-1510	1/1/2011	1/6/2011	Standard Class	Magdalene Morse	Consumer	Ontario	Canada	Canada	Canada	TEC-OK-10002759	Technology	Machines	Okidata Inkjet, Wireless	514	1	0	5.12	24.10	Medium	2011
IN-2011-79397	1/1/2011	1/3/2011	Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania	OFF-AP-10000304	Office Supplies	Appliances	Hoover Microwave, White	276	1	0.1	110.412	125.32	Critical	2011
ID-2011-80230	1/1/2011	1/9/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-CO-10004182	Technology	Copiers	Hewlett Wireless Fax, Laser	912	4	0.4	-319.404	107.1	Low	2011
ID-2011-46881	1/1/2011	1/7/2011	Standard Class	Lindsay Williams	Corporate	Nineveh	Iraq	EMEA	EMEA	FUR-NEW-10002791	Furniture	Chairs	Novimex Swivel Stool, Set of Two	867	6	0	259.32	81.26	High	2011
IN-2011-61519	1/1/2011	1/7/2011	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	OFF-ST-10003020	Office Supplies	Storage	Tenex Lockers, Industrial	338	3	0.45	-122.8005	33.75	High	2011
IN-2011-61519	1/1/2011	1/7/2011	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	FUR-TA-10002797	Furniture	Tables	Chromcraft Round Table, Adjustable Height	211	1	0.55	-70.3995	21.52	High	2011
ES-2011-4969686	1/1/2011	1/7/2011	Standard Class	Dorothy Dickinson	Consumer	England	United Kingdom	EU	North	FUR-BO-10000720	Furniture	Bookcases	Dania Corner Shelving, Traditional	854	7	0	290.43	12.56	Medium	2011
IN-2011-33652	1/1/2011	1/9/2011	Standard Class	Dennis Pardue	Home Office	Sarawak	Malaysia	APAC	Southeast Asia	TEC-CO-10000594	Technology	Copiers	Hewlett Fax and Copier, Laser	693	1	0	50.13	10.40	Medium	2011
ID-2011-80230	1/1/2011	1/9/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000214	Furniture	Chairs	Hon Rocking Chair, Set of Two	159	2	0.4	-99.876	10.07	Low	2011
MA-2011-160234	1/1/2011	1/7/2011	Standard Class	Stewart Visinsky	Consumer	Guatemala	Guatemala	LATAM	Central	TEC-PH-10002667	Technology	Phones	Nokia Headset, VoIP	195	4	0	64.88	8.43	Medium	2011
IN-2011-770	1/1/2011	1/7/2011	Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-RIN-10003081	Office Supplies	Appliances	Breville Coffee Grinder, Black	123	2	0	42.9	8.41	Medium	2011
ID-2011-80230	1/1/2011	1/9/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-AC-10002863	Technology	Accessories	Belkin Numeric Keypad, Bluetooth	69	2	0.4	3.42	8.34	Low	2011
ID-2011-80230	1/1/2011	1/9/2011	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000968	Furniture	Chairs	SAFCO Chairmat, Black	69	2	0.4	-20.412	8.17	Low	2011
ID-2011-12596	1/1/2011	1/8/2011	Standard Class	Chris McAfee	Consumer	Nakhon Ratchasima	Thailand	APAC	Southeast Asia	OFF-ST-10002066	Office Supplies	Storage	Smead File Cart, Blue	195	2	0.47	-45.9018	7.74	Medium	2011
IN-2011-79397	1/1/2011	1/3/2011	Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania	OFF-LA-10003396	Office Supplies	Labels	Avery Color Coded Labels, Laser Printer Compatible	36	3	0.1	4.743	7.46	Critical	2011
IN-2011-7080	1/1/2011	1/8/2011	Second Class	Nat Gilpin	Corporate	Razavi Khorasan	Iran	EMEA	EMEA	OFF-BIC-10000182	Office Supplies	Art	BIC Sketch Pad, Water Color	52	1	0	7.77	5.91	High	2011

There are a total of 21 columns and 51,291 rows in this above google sheet data that was provided initially.

It's a list of all the variables (total of 21) bunched together in this list format to show all the data that has been collected for sales. This is highly inefficient due to multiple repetition and we can utilize a relational database to showcase a cleaner and efficient way of presenting data for the purpose of analytics. In order to understand the importance of translating this list data to relational database, we need to identify the problems working with List Data (Google Sheets)

Problems of List Data:

- Redundancy: Same data appearing multiple times
    - Ex: Out of 21 variables, at least 18 of them are consistently repeating. This is taking too much storage space
- Multiple Themes: Having too much info/concepts in one row.
    - Ex: A given row tells us every single information we have to know about that specific line order, but we rarely want all the information at the same time. It's always better to divide variables into different tables
- Anomalies
    - Update Anomaly: If we wanted to update the category type of a specific product, we would have to make changes to every single product in this list and it would be very time consuming.
    - Deletion Anomaly: If we want to remove a specific order from the list, it would end up deleting every information related to this order where some columns might be needed for future use.

- Insertion Anomaly: If we want to add a new city and country to the list of countries that have a participating market, there would be NULL values for every other column because an order hasn't taken place.

Relational Database helps solve all of these issues by dividing variables into different tables, while ensuring that the relationships between tables exist. This helps connect all of the tables and variables between each other. The next step would be to identify the Functional Dependencies (Value of 1 or more attributes helps find another attribute) for the purpose of Data Normalization. This is the process to make sure the relational database is well informed and is not susceptible to data anomalies.

Functional Dependencies:

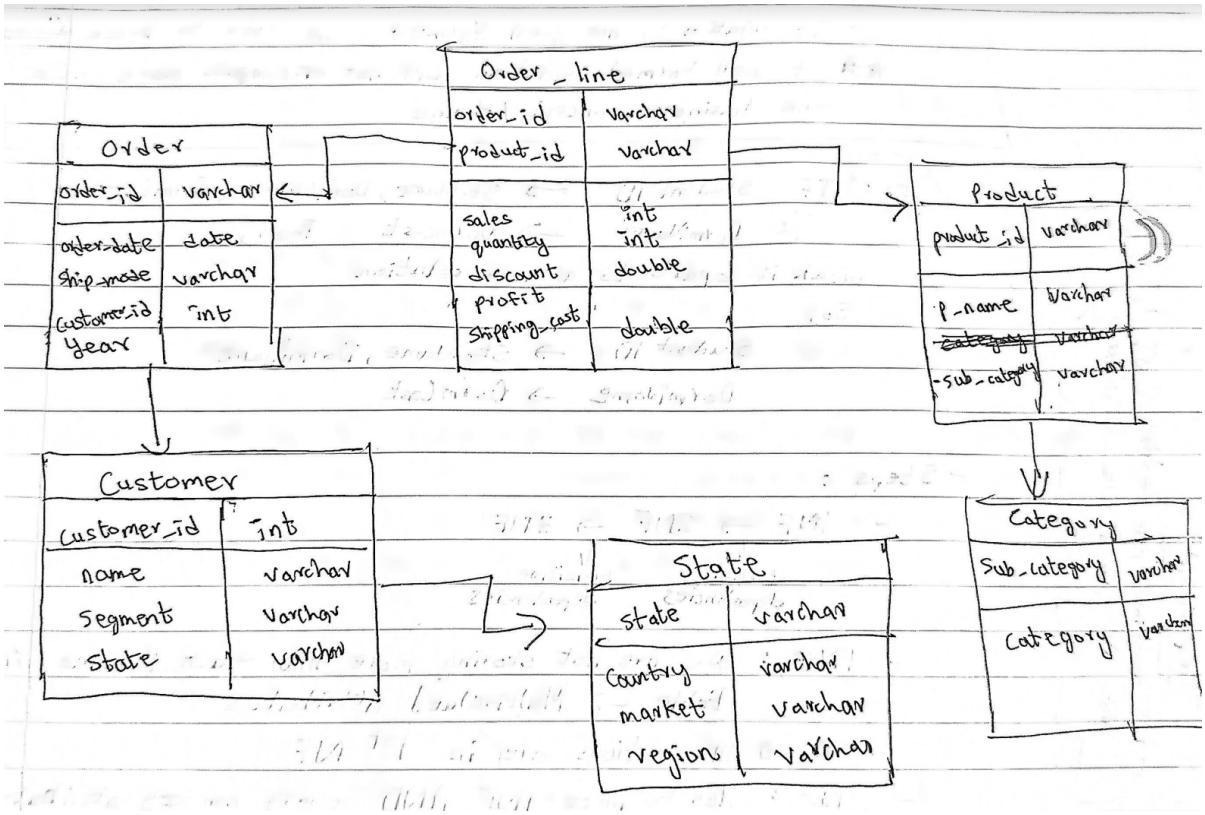
- Every determinant should be a candidate key (has the potential to become a primary key).
- If a relation is not well formed, we need to break into multiple tables.
- A well informed relation will not encompass more than one business concept/theme
- Steps to complete this process: 1NF => 2NF => 3NF
    - 1NF: Do not store more than 1 value in a table (Multivalued attributes)
    - 2NF: Has to meet 1NF AND every non key attribute must be dependent on the primary key. No partial dependencies.
    - 3NF: PK determinant for another attribute but that attribute is determinant for another attribute. This is a transitive dependency so, we break it into smaller relations.

From all the variables in the google sheet, the following functional dependencies were created to create a relationship between every variable:

- {order_id} ⇒ {customer_id, first_name, last_name, state, country, segment, market, region, ship_date, ship_mode, order_date, order_priority, year}
- {customer_id} ⇒ {first_name, last_name, state, country, segment, market, region}
- {state} ⇒ {country, market, region}
- {country} ⇒ {market}
- {product_id} ⇒ {product_name, category, sub_category}
- {sub_category} ⇒ {category}
- {order_id, product_id} ⇒ {sales, quantity, discount, profit, shipping cost}

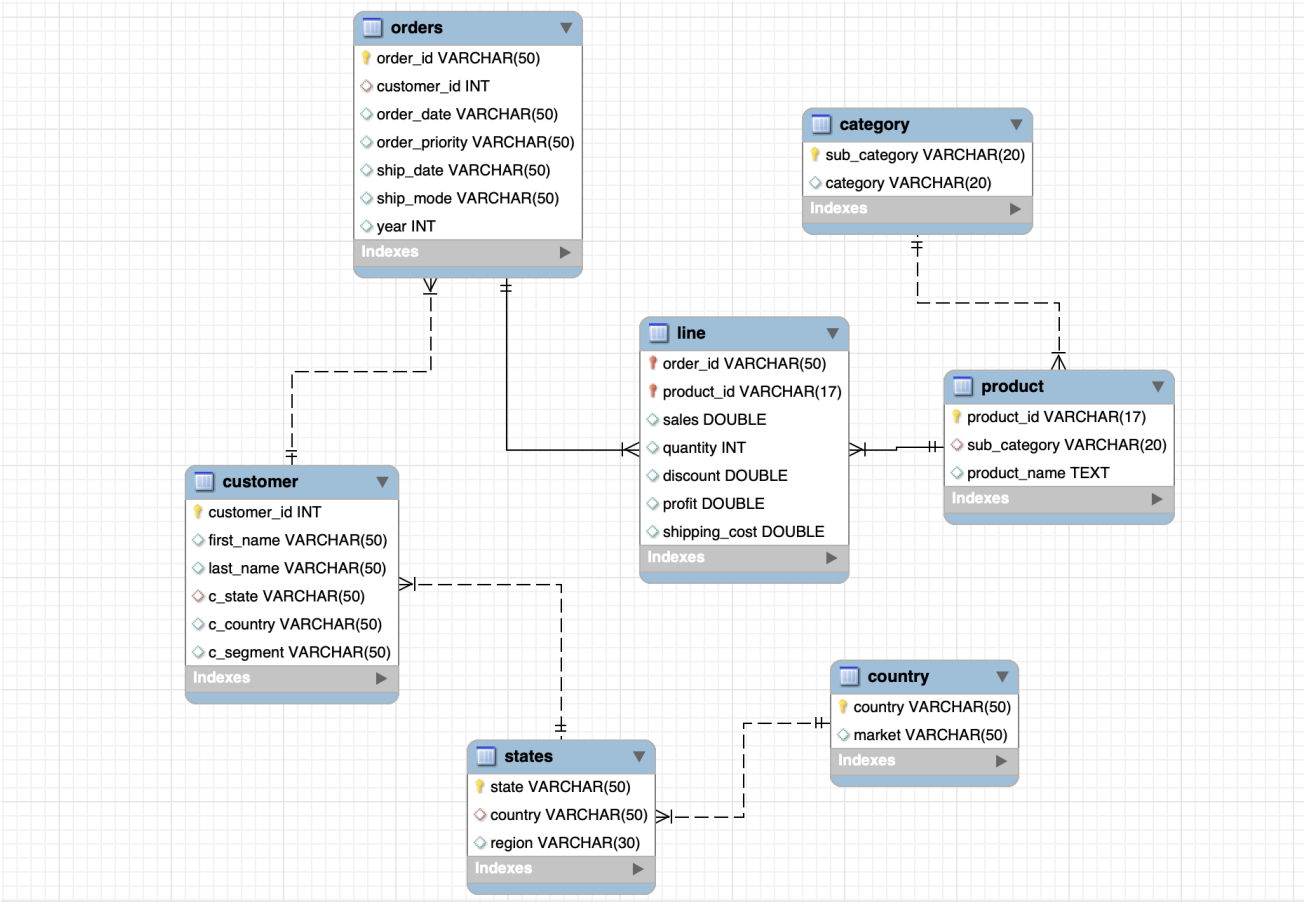
Data Normalization

The following is my initial take on Data Normalization to showcase the relationships between each table:



ER Diagram

The next step would be to create a well functioning ER Diagram by identifying the primary and foreign keys for each table. The following ER Diagram shows all the properties for every table, and how it creates a well formed relational database.



Things to add to this document:

- *How Customer ID was created*
- *How each table was separately made in Excel before converting to csv files for importing in MySQL*
- *Link to that Spreadsheet to show the preliminary table*

Visual Representation of Tables

For visual purposes, I have provided the images for the first 10 values for each table below through querying in MySQL Workbench.

Category Table:

sub_category	category
Accessories	Technology
Appliances	Office Supplies
Art	Office Supplies
Binders	Office Supplies
Bookcases	Furniture
Chairs	Furniture
Copiers	Technology
Envelopes	Office Supplies
Fasteners	Office Supplies
Furnishings	Furniture
NULL	NULL

Country Table:

country	market
Afghanistan	APAC
Albania	EMEA
Algeria	Africa
Angola	Africa
Argentina	LATAM
Armenia	EMEA
Australia	APAC
Austria	EU
Azerbaijan	EMEA
Bahrain	EMEA
NULL	NULL

Line Table:

order_id	product_id	sales	quantity	discount	profit	shipping_co...
AE-2011-9160	OFF-FEL-10001405	82.674	2	0.7	-157.086	5.69
AE-2011-9160	TEC-EPS-10004171	78.408	6	0.7	-88.992	3.87
AE-2013-1130	FUR-BUS-10003055	224.748	6	0.7	-232.272	60.08
AE-2013-1130	OFF-ACC-10004278	4.248	1	0.7	-4.692	0.1
AE-2013-1530	OFF-STI-10000114	16.668	2	0.7	-29.472	1.41
AE-2013-1530	OFF-TEN-10002817	6.966	1	0.7	-8.604	1.75
AE-2014-2840	OFF-ROG-10001340	42.48	1	0.7	-75.06	8.04
AE-2014-3830	OFF-AVE-10000357	3.159	1	0.7	-4.971	0.25
AE-2014-3830	OFF-AVE-10004827	3.429	1	0.7	-3.891	0.25
AE-2014-3830	OFF-BIC-10002270	77.256	8	0.7	-82.584	4.08
NULL	NULL	NULL	NULL	NULL	NULL	NULL

Customer Table:

customer_id	first_name	last_name	c_state	c_country	c_segment
1000001	Aaron	Bergman	Bangkok	Thailand	Consumer
1000002	Aaron	Bergman	Campania	Italy	Consumer
1000003	Aaron	Bergman	Centre	Cameroon	Consumer
1000004	Aaron	Bergman	Copperbelt	Zambia	Consumer
1000005	Aaron	Bergman	Da Nang	Vietnam	Consumer
1000006	Aaron	Bergman	England	United Kingdom	Consumer
1000007	Aaron	Bergman	Guatemala	Guatemala	Consumer
1000008	Aaron	Bergman	Hubei	China	Consumer
1000009	Aaron	Bergman	Ile-de-France	France	Consumer
1000010	Aaron	Bergman	Jalisco	Mexico	Consumer
NULL	NULL	NULL	NULL	NULL	NULL

Product Table:

product_id	sub_category	product_name
FUR-ADV-10000002	Furnishings	Advantus Photo Frame, Duo Pack
FUR-ADV-10000108	Furnishings	Advantus Clock, Erganomic
FUR-ADV-10000183	Furnishings	Advantus Photo Frame, Black
FUR-ADV-10000188	Furnishings	Advantus Stacking Tray, Erganomic
FUR-ADV-10000190	Furnishings	Advantus Frame, Duo Pack
FUR-ADV-10000571	Furnishings	Advantus Frame, Erganomic
FUR-ADV-10000600	Furnishings	Advantus Clock, Duo Pack
FUR-ADV-10000847	Furnishings	Advantus Stacking Tray, Black
FUR-ADV-10001283	Furnishings	Advantus Frame, Black
FUR-ADV-10001440	Furnishings	Advantus Door Stop, Black
NULL	NULL	NULL

States Table:

state	country	region
'Ajman	United Arab Emirates	EMEA
'Amman	Jordan	EMEA
'Asir	Saudi Arabia	EMEA
Abia	Nigeria	Africa
Abruzzi	Italy	South
Abuja Capital Territory	Nigeria	Africa
Aceh	Indonesia	Southeast Asia
Acre	Brazil	South
Ad Dawhah	Qatar	EMEA
Adamaoua	Cameroon	Africa

Orders Table:

order_id	customer_id	order_date	order_prior...	ship_date	ship_mode	year
AE-2011-9160	1017350	2011-10-03	Medium	2011-10-07	Standard Class	2011
AE-2013-1130	1008005	2013-10-14	High	2013-10-14	Same Day	2013
AE-2013-1530	1014442	2013-12-31	High	2014-01-03	Second Class	2013
AE-2014-2840	1017256	2014-11-05	Critical	2014-11-08	First Class	2014
AE-2014-3830	1009086	2014-12-13	Medium	2014-12-19	Standard Class	2014
AE-2014-4120	1011269	2014-03-07	Medium	2014-03-11	Standard Class	2014
AG-2011-1070	1021974	2011-01-10	Medium	2011-01-15	Second Class	2011
AG-2011-1390	1007169	2011-08-16	Medium	2011-08-21	Standard Class	2011
AG-2011-1440	1006262	2011-12-15	Medium	2011-12-20	Standard Class	2011
AG-2011-2040	1022076	2011-01-01	Medium	2011-01-06	Standard Class	2011
NULL	NULL	NULL	NULL	NULL	NULL	NULL