Data Analytics Case Study 2; Datainfiniti

by Luke Watts, 28/01/2025

For my second case study, I will perform cleaning, standardizing and analysis on a set of data using Google Sheets and Google Looker Studio.

Background Scenario

I will be using a dataset from the e-commerce platform Dataworld that focuses on electronics products. The data set is provided by Datafiniti via this website link. The dataset includes information such as product names, categories, prices, sales figures, and customer reviews. By leveraging this data, we can gain valuable insights into market trends, consumer preferences, and pricing strategies. We will begin our analysis by establishing the goals & objectives of what we want to achieve from looking at this dataset.

Step 1: Ask

Let us start this process by answering these fundamental questions;

1. **Price Analysis**: What are the min, max and average prices? How do prices vary by brand and category?
2. **Availability and Condition**: What conditions are products in, and how does this impact prices?
3. **Brand and Manufacturer Insights**: Which brands and manufacturers have the most products? What are the average prices for each brand and manufacturer?
4. **Product Details**: How does product weight relate to pricing and shipping costs?
5. **Time Analysis**: How have prices changed over time?

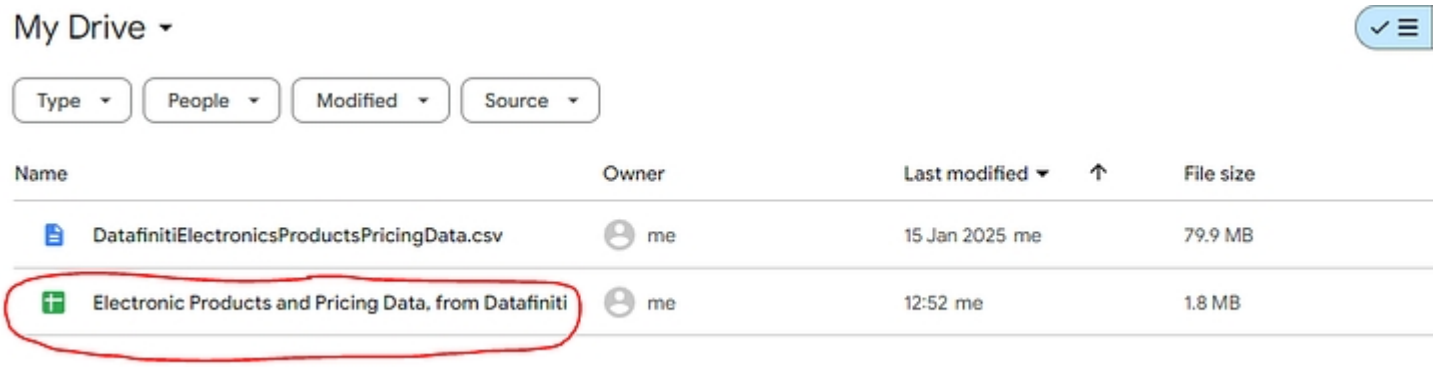
These questions will be the guide as we craft this data visualization dashboard from the data set.

Step 2: Prepare

Perform ROCCC check: Firstly, let us do our data validation due diligence. Data is considered reliable - it is unbiased, and the sample size is large. Data is original and is comprehensive- we have a full data set over a given time frame so we have all the info we need to answer the business task questions. Data is not current, as it was uploaded in 2018, but for our hypothetical analysis we can just roll back the clock and pretend we are performing this analysis 7 years in the past. Data is not cited, but has been vetted by Datafiniti.

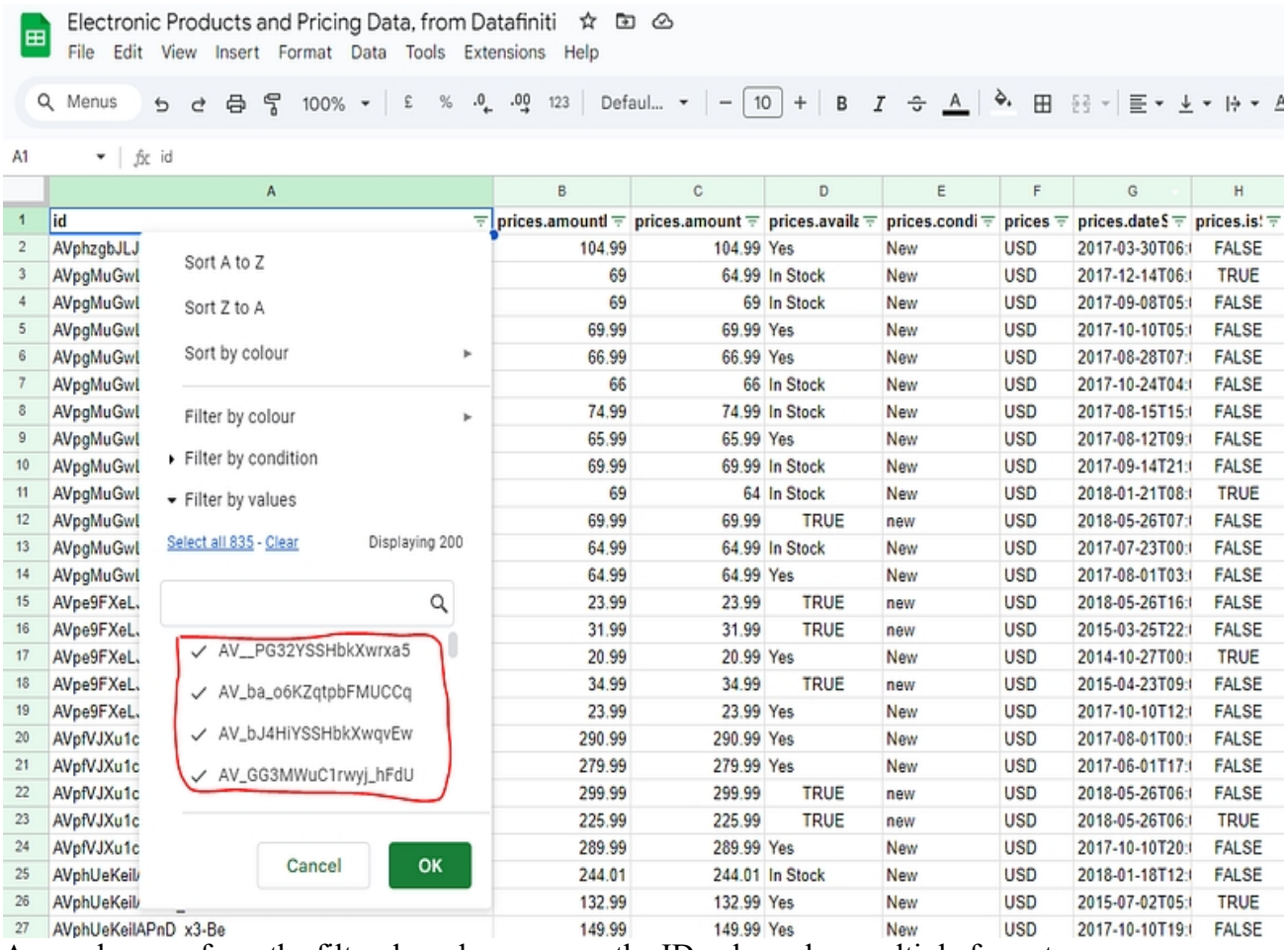
Next, the second stage is to broadly observe the dataset to gain an overview. Identify what the columns fields are, what is the data type, how large the data set is, and any other important information.

Let’s upload the dataset to google drive and change it to a Google Sheet format. Then we can begin to analyze & clean the table.



Change the .csv file to a Google sheets document

Within Google Sheets, an efficient way of checking data is by filtering columns one at a time, so let us proceed through the spreadsheet this way, looking for anomalies, formatting issues and bad data.



	A	B	C	D	E	F	G	H
1	id	prices.amountl	prices.amount	prices.availa	prices.condi	prices	prices.dateS	prices.is!
2	AVphzgbJLJ	104.99	104.99	Yes	New	USD	2017-03-30T06:(	FALSE
3	AVpgMuGwl	69	64.99	In Stock	New	USD	2017-12-14T06:(	TRUE
4	AVpgMuGwl	69	69	In Stock	New	USD	2017-09-08T05:(	FALSE
5	AVpgMuGwl	69.99	69.99	Yes	New	USD	2017-10-10T05:(	FALSE
6	AVpgMuGwl	66.99	66.99	Yes	New	USD	2017-08-28T07:(	FALSE
7	AVpgMuGwl	66	66	In Stock	New	USD	2017-10-24T04:(	FALSE
8	AVpgMuGwl	74.99	74.99	In Stock	New	USD	2017-08-15T15:(	FALSE
9	AVpgMuGwl	65.99	65.99	Yes	New	USD	2017-08-12T09:(	FALSE
10	AVpgMuGwl	69.99	69.99	In Stock	New	USD	2017-09-14T21:(	FALSE
11	AVpgMuGwl	69	64	In Stock	New	USD	2018-01-21T08:(	TRUE
12	AVpgMuGwl	69.99	69.99	TRUE	new	USD	2018-05-26T07:(	FALSE
13	AVpgMuGwl	64.99	64.99	In Stock	New	USD	2017-07-23T00:(	FALSE
14	AVpgMuGwl	64.99	64.99	Yes	New	USD	2017-08-01T03:(	FALSE
15	AVpe9FXeL.	23.99	23.99	TRUE	new	USD	2018-05-26T16:(	FALSE
16	AVpe9FXeL.	31.99	31.99	TRUE	new	USD	2015-03-25T22:(	FALSE
17	AVpe9FXeL.	20.99	20.99	Yes	New	USD	2014-10-27T00:(	TRUE
18	AVpe9FXeL.	34.99	34.99	TRUE	new	USD	2015-04-23T09:(	FALSE
19	AVpe9FXeL.	23.99	23.99	Yes	New	USD	2017-10-10T12:(	FALSE
20	AVpfVJXu1c	290.99	290.99	Yes	New	USD	2017-08-01T00:(	FALSE
21	AVpfVJXu1c	279.99	279.99	Yes	New	USD	2017-06-01T17:(	FALSE
22	AVpfVJXu1c	299.99	299.99	TRUE	new	USD	2018-05-26T06:(	FALSE
23	AVpfVJXu1c	225.99	225.99	TRUE	new	USD	2018-05-26T06:(	TRUE
24	AVpfVJXu1c	289.99	289.99	Yes	New	USD	2017-10-10T20:(	FALSE
25	AVphUeKeil	244.01	244.01	In Stock	New	USD	2018-01-18T12:(	FALSE
26	AVphUeKeil	132.99	132.99	Yes	New	USD	2015-07-02T05:(	TRUE
27	AVphUeKeilAPnD_x3-Be	149.99	149.99	Yes	New	USD	2017-10-10T19:(	FALSE

As can be seen from the filter drop down menu, the ID column has multiple formats

Step 3: Process

We will begin with cleaning the data by creating a copy of the data set within Google Sheets, and naming it "DatafinitiElectronicsProductsPricingData (CLEAN)".

We can then follow an established process that will help guide our cleaning:

1. Remove duplicate data
2. Trim whitespace
3. Check irrelevant data based on the Id column
4. Remove unnecessary columns (according to our objectives)
5. Standardizing the data format

We can perform steps 1 & 2 very quickly in Sheets by using the data cleanup option under the Data heading of the drop down menu. No duplicates were found, but whitespace was trimmed from several hundred cells.

Step 3—The Id column is the best way to check irrelevant data, since in the previous process of identifying the data we found that many Ids are in different formats and their rows contained irrelevant data.

We can use Regex combine with a filter to find the Ids that are not the same pattern.

- First we need to change all of the Ids into *Plain Text Format* so that our regex can return valid values (True/False)
- Then, we need to find a Regex Pattern that is a match to valid Id's. Let's create an additional column to help find the pattern that is a match to valid Id's, so we can then test the Regex pattern.
- From an observation of the Id column, we can see that the valid Id's always begin with "AV" or "AW", so we can use this regex pattern:

Only Id's containing AV or AW will return "TRUE" in the match id column

Once the formula has been applied to all Id's, we can filter the "match id" column to isolate just the "FALSE" results, and then delete those rows entirely. This will remove all unwanted Id's and

their associated data in the other columns, and only data that corresponds to “match id = TRUE” will remain.

Then, we can remove the temporary match id column, and the regex formula cells, as they are no longer necessary.

Step 4—Removing unnecessary columns is easy once they have been identified. Our objectives revolve around questions related to pricing, brand, product details, and time range - so we can just remove all of the columns that are not relevant to these metrics:

Full Column			Filtered Column	
1	id		1	id
2	prices.amountMax		2	prices.amountMax
3	prices.amountMin		3	prices.amountMin
4	prices.availability		4	prices.availability
5	prices.condition		5	prices.condition
6	prices.currency		6	prices.currency
7	prices.dateSeen		7	prices.isSale
8	prices.isSale		8	prices.merchant
9	prices.merchant		9	prices.shipping
10	prices.shipping		10	brand
11	prices.sourceURLs		11	categories
12	asins		12	dateAdded
13	brand		13	dateUpdated
14	categories		14	manufacturer
15	dateAdded		15	manufacturerNumber
16	dateUpdated		16	name
17	ean		17	primaryCategories
18	imageURLs		18	weight
19	keys			
20	manufacturer			
21	manufacturerNumber			
22	name			
23	primaryCategories			
24	sourceURLs			
25	upc			
26	weight			

All highlighted columns will be deleted

Step 5—Standardizing the data format will help our data to be consistent and easier to process/analyze. Now that we have all of the columns that we want to analyze, we need to identify each column's data type. Here are the data formats corresponding to each column:

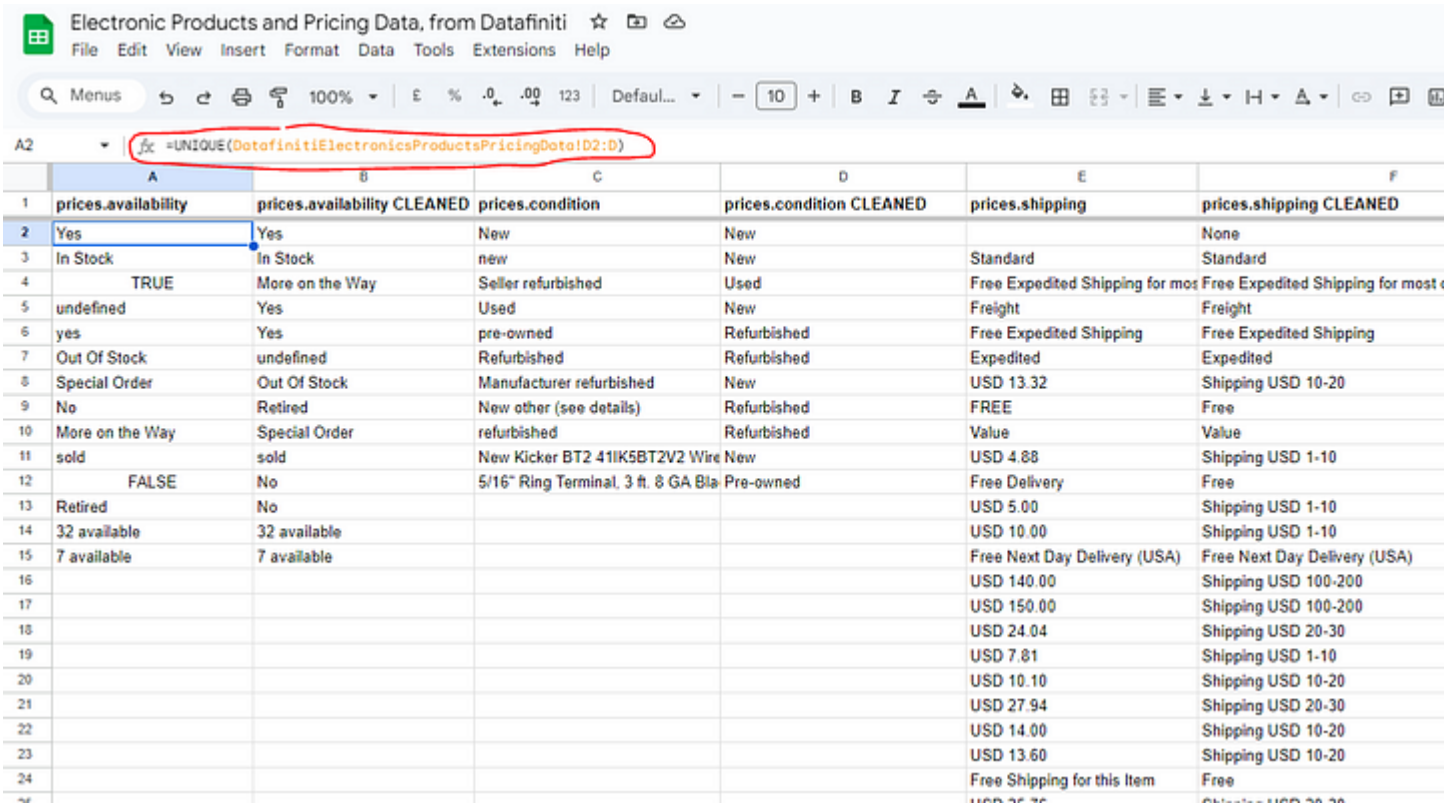
Filtered Column

No	Columns	Data Format
1	id	Plain Text
2	prices.amountMax	Currency
3	prices.amountMin	Currency
4	prices.availability	Plain Text (with data validation)
5	prices.condition	Plain Text (with data validation)
6	prices.currency	Plain Text
7	prices.isSale	Boolean
8	prices.merchant	Plain Text
9	prices.shipping	Plain Text (with data validation)
10	brand	Plain Text
11	categories	Plain Text
12	dateAdded	Date
13	dateUpdated	Date
14	manufacturer	Plain Text
15	manufacturerNumber	Plain Text
16	name	Plain Text
17	primaryCategories	Plain Text
18	weight	Number (standardizing units of measurement in pounds)

highlighted columns require further cleaning and validation to make the data set consistent

For the first three columns we can extract the unique values to make a list of validated data, and we can delete the cell values that are not in the data validation range. We can do this thusly:

1. Add another sheet for the extracted unique values, then use the "Unique" formula operator to extract the distinct values



	A	B	C	D	E	F
1	prices.availability	prices.availability CLEANED	prices.condition	prices.condition CLEANED	prices.shipping	prices.shipping CLEANED
2	Yes	Yes	New	New		None
3	In Stock	In Stock	new	New	Standard	Standard
4	TRUE	More on the Way	Seller refurbished	Used	Free Expedited Shipping for mos	Free Expedited Shipping for most
5	undefined	Yes	Used	New	Freight	Freight
6	yes	Yes	pre-owned	Refurbished	Free Expedited Shipping	Free Expedited Shipping
7	Out Of Stock	undefined	Refurbished	Refurbished	Expedited	Expedited
8	Special Order	Out Of Stock	Manufacturer refurbished	New	USD 13.32	Shipping USD 10-20
9	No	Retired	New other (see details)	Refurbished	FREE	Free
10	More on the Way	Special Order	refurbished	Refurbished	Value	Value
11	sold	sold	New Kicker BT2 41IK5BT2V2 Wire	New	USD 4.88	Shipping USD 1-10
12	FALSE	No	5/16" Ring Terminal, 3 ft. 8 GA Bla	Pre-owned	Free Delivery	Free
13	Retired	No			USD 5.00	Shipping USD 1-10
14	32 available	32 available			USD 10.00	Shipping USD 1-10
15	7 available	7 available			Free Next Day Delivery (USA)	Free Next Day Delivery (USA)
16					USD 140.00	Shipping USD 100-200
17					USD 150.00	Shipping USD 100-200
18					USD 24.04	Shipping USD 20-30
19					USD 7.81	Shipping USD 1-10
20					USD 10.10	Shipping USD 10-20
21					USD 27.94	Shipping USD 20-30
22					USD 14.00	Shipping USD 10-20
23					USD 13.60	Shipping USD 10-20
24					Free Shipping for this Item	Free

price and condition columns are imported from the original data set so no errors occur when they are updated / changed in the sheet of cleaned data

2. The Price Shipping column contains both numeric data and string data. Since we don't have enough information about the shipping rates, we can simplify this column by changing the number format to qualitative data. We will group the shipping prices by creating a shipping price range format e.g. Shipping Charge 1–10, Shipping Charge 10–20, and so forth.

3. Add another column of each price category to change the values.

D	E	F	G	H	I	J	K	L
prices.availability	prc available std	prices.condition	prc condition std	prices.currency	prices.isSale	prices.merchant	prices.shipping	prc shipping std
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
In Stock	In Stock	New	New	USD	TRUE	Walmart.com	Expedited	Expedited
In Stock	In Stock	New	New	USD	FALSE	Walmart.com	Expedited	Expedited
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
In Stock	In Stock	New	New	USD	FALSE	Walmart.com	Expedited	Expedited
In Stock	In Stock	New	New	USD	FALSE	California Electr	Value	Value
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
In Stock	In Stock	New	New	USD	FALSE	Walmart.com	Expedited	Expedited
In Stock	In Stock	New	New	USD	TRUE	Walmart.com	Standard	Standard
TRUE	Yes	new	New	USD	FALSE	Bestbuy.com		None
In Stock	In Stock	New	New	USD	FALSE	Walmart.com	Expedited	Expedited
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
TRUE	Yes	new	New	USD	FALSE	Bestbuy.com		None
TRUE	Yes	new	New	USD	FALSE	Wayfair - Walmart.com		None
Yes	Yes	New	New	USD	TRUE	Bestbuy.com	Free Shipping	Free Shipping
TRUE	Yes	new	New	USD	FALSE	Wayfair - Walmart.com		None
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
TRUE	Yes	new	New	USD	FALSE	Bestbuy.com		None
TRUE	Yes	new	New	USD	TRUE	Bestbuy.com		None
Yes	Yes	New	New	USD	FALSE	Bestbuy.com		None
In Stock	In Stock	New	New	USD	FALSE	Growkart	Value	Value
Yes	Yes	New	New	USD	TRUE	Bestbuy.com	Free Shippingon	Free Shipping o

4. Delete the old columns

Before delete the old columns, make sure we have changed the format of new columns from formulas to *Plain Text*. We can do this via copy > Paste Values Only. Below is the final table.

	A	B	C	D	E	F	G	H	I
1	id	Price Max	Price Min	Availability	Condition	Currency	On Sale	Merchant	Shipping Price
2	AWOpMDbqYSSHbkXw06lc	$38.99	$38.99	Yes	New	USD	TRUE	Bestbuy.com	None
3	AWOpMDbqYSSHbkXw06lc	$51.99	$51.99	Yes	New	USD	FALSE	Bestbuy.com	None
4	AWOpOCd2YSSHbkXw07ei	$224.99	$224.99	Yes	New	USD	TRUE	Bestbuy.com	None
5	AWOpOCd2YSSHbkXw07ei	$249.99	$249.99	Yes	New	USD	FALSE	Bestbuy.com	None
6	AWOpLyzrHh53nbDRKmTD	$1,398.00	$908.73	In Stock	New	USD	TRUE	VIPOUTLET	Expedited
7	AWOpLyzrHh53nbDRKmTD	$1,099.00	$1,099.00	In Stock	New	USD	TRUE	bhphotovideo.com	Shipping for most
8	AWOpLyzrHh53nbDRKmTD	$1,289.99	$1,289.99	Yes	New	USD	TRUE	Bestbuy.com	None
9	AWOpLyzrHh53nbDRKmTD	$1,499.99	$1,499.99	Yes	New	USD	FALSE	Bestbuy.com	None
10	AWOpLyzrHh53nbDRKmTD	$1,797.99	$1,797.99	In Stock	New	USD	FALSE	Walts TV	Standard
11	AWOpLyzrHh53nbDRKmTD	$1,897.99	$1,897.99	In Stock	New	USD	FALSE	bhphotovideo.com	Shipping for most
12	AWOpWM18YSSHbkXw09Y	$749.99	$749.99	Yes	New	USD	FALSE	Bestbuy.com	None
13	AWOmw9PhHh53nbDRKld5	$48.95	$48.95	Yes	Refurbished	USD	FALSE	superdealunlimited	Free
14	AWOmw9PhHh53nbDRKld5	$48.99	$48.99	Yes	Refurbished	USD	FALSE	geekdeal	Free
15	AWOmw9PhHh53nbDRKld5	$49.99	$49.99	Yes	Refurbished	USD	FALSE	superdealunlimited	Free
16	AWOmw9PhHh53nbDRKld5	$58.99	$58.99	Yes	New	USD	TRUE	Bestbuy.com	None
17	AWOmw9PhHh53nbDRKld5	$58.99	$58.99	Yes	New	USD	FALSE	Best Buy	None
18	AWOmw9PhHh53nbDRKld5	$59.99	$59.99	Yes	New	USD	FALSE	Bestbuy.com	None
19	AWOUVe5cuC1rwyj_rmlB	$548.00	$548.00	In Stock	New	USD	TRUE	bhphotovideo.com	Shipping for most
20	AWOUVe5cuC1rwyj_rmlB	$549.99	$549.99	Yes	New	USD	TRUE	Bestbuy.com	None
21	AWOUVe5cuC1rwyj_rmlB	$598.00	$598.00	In Stock	New	USD	FALSE	bhphotovideo.com	Shipping for most
22	AWOUVe5cuC1rwyj_rmlB	$599.99	$599.99	Yes	New	USD	FALSE	Bestbuy.com	None
23	AWOTVpQilwln0LfXm3OV	$749.00	$749.00	In Stock	New	USD	TRUE	bhphotovideo.com	Shipping for most
24	AWOTVpQilwln0LfXm3OV	$749.99	$749.99	Yes	New	USD	TRUE	Bestbuy.com	None
25	AWOTVpQilwln0LfXm3OV	$1,144.00	$810.99	In Stock	New	USD	TRUE	Tri State Camera	Value
26	AWOTVpQilwln0LfXm3OV	$899.00	$899.00	In Stock	New	USD	FALSE	bhphotovideo.com	Shipping for most
27	AWOTVpQilwln0LfXm3OV	$899.99	$899.99	Yes	New	USD	FALSE	Bestbuy.com	None
28	AWN6PKWauC1rwyj_rgtt	$449.00	$449.00	In Stock	New	USD	FALSE	bhphotovideo.com	Shipping for most
29	AWN6PKWauC1rwyj_rgtt	$479.99	$479.99	Yes	New	USD	FALSE	Bestbuy.com	None
30	AWM7tLUgHh53nbDRKC6-	$159.99	$159.99	Yes	New	USD	FALSE	Bestbuy.com	None
31	AWMjgUYxHh53nbDRJ9NI	$199.99	$199.99	Yes	New	USD	FALSE	Bestbuy.com	None
32	AWMZjqTQYSSHbkXw0Ixv	$129.95	$129.95	In Stock	New	USD	FALSE	bhphotovideo.com	Shipping for most
33	AWMZjqTQYSSHbkXw0Ixv	$129.99	$129.99	Yes	New	USD	FALSE	Bestbuy.com	None
34	AWLwDO9uYSSHbkXwz54e	$198.00	$198.00	lore on the Wa	New	USD	FALSE	bhphotovideo.com	Shipping for most

Column titles have also been cleaned up and adjusted for clarity

Next, the date columns need to be formatted. This is quite straight forward—we just use text extraction to get only the date values:

```
=LEFT(L2,LEN(L2)-10)
```

After that, change the format to usual text (not a formula) and change the format to Date. Here is the final column for date:

L	M	N	O
dateAdded	dateAddedstd	dateUpdated	dateUpdatedstd
2015-04-13T12:	4/13/2015	2018-05-12T18:	5/12/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-05-18T14:	5/18/2015	2018-06-13T19:	6/13/2018
2015-03-18T12:	3/18/2015	2018-06-13T19:	6/13/2018

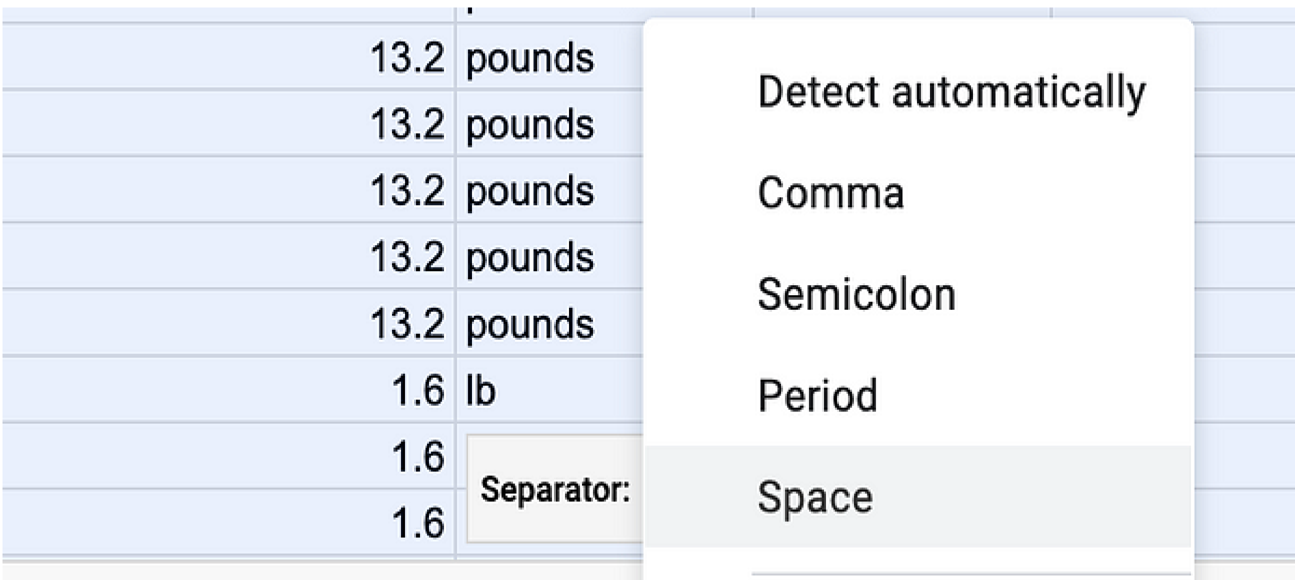
Now, we just remove the original date columns L & N.

Finally, the weight column is quite hard to standardize, because several products have more than one weight per cell…

7/26/2018	YAMAHA	NSSP1800BL	NS-SP1800BL 5 Electronics	1.6 lb 2.6 lb 1.6 lb 18.7 lb	FALSE
7/26/2018	YAMAHA	NSSP1800BL	NS-SP1800BL 5 Electronics	1.6 lb 2.6 lb 1.6 lb 18.7 lb	FALSE
7/26/2018	YAMAHA	NSSP1800BL	NS-SP1800BL 5 Electronics	1.6 lb 2.6 lb 1.6 lb 18.7 lb	FALSE
7/26/2018	YAMAHA	NSSP1800BL	NS-SP1800BL 5 Electronics	1.6 lb 2.6 lb 1.6 lb 18.7 lb	FALSE
7/26/2018	YAMAHA	NSSP1800BL	NS-SP1800BL 5 Electronics	1.6 lb 2.6 lb 1.6 lb 18.7 lb	FALSE

However, for the purpose of this beginner data analyst project, we will only take the first weight of each and standardize the units using grams (g).

- The first thing to do is to separate the numeric data from their units using the “Split text to columns” function



- Now we will convert the units to grams.

1 oz (ounce)	28.35 g
1 lb (pound)	453.6 g
1 Kg	1,000 g
1 g	1 g

- Here’s our final column result.

weight	units	unit on gram
32.8	pounds	14878.08
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4
14	pounds	6350.4

Finally, we can continue to the next step. Here is what the final table looks like:

	L	M	N	O	P	Q	R
	Date Added	Date Updated	Manufacturer	Man. Number	Product name	Categories	Weight Std (g)
m	2018-05-28	2018-06-13		PCIIDE2	StarTech - 2 Port PCI IDE Controller Adapter Card - Green	Electronics	73.839
m	2018-05-28	2018-06-13		PCIIDE2	StarTech - 2 Port PCI IDE Controller Adapter Card - Green	Electronics	73.839
ve	2018-05-28	2018-06-13		MZ-7KE512BW	Samsung - 850 PRO 512GB Internal SATA III Solid State Drive fo	Electronics	3.92
ve	2018-05-28	2018-06-13		MZ-7KE512BW	Samsung - 850 PRO 512GB Internal SATA III Solid State Drive fo	Electronics	3.92
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
/ ε	2018-05-28	2018-06-13		UN65MU7500FXZA	Samsung - 65 Class - LED - Curved - MU7500 Series - 2160p - S	Electronics	26862.9
ve	2018-05-28	2018-06-21		MZ-75E2T0B/AM	Samsung 850 EVO 2TB 2.5-Inch SATA III Internal SSD (MZ-75E2	Electronics	3.92
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
ki	2018-05-28	2018-07-02		RZ03-02260200-R3U1	Details About Razer Cynosa Chroma Rgb Gaming Keyboard Spil	Electronics	58.52
/	2018-05-24	2018-07-03		SELP18105G	Sony SELP18105G E PZ 18-105mm F4 G OSS	Electronics	480.18
/	2018-05-24	2018-07-03		SELP18105G	Sony SELP18105G E PZ 18-105mm F4 G OSS	Electronics	480.18
/	2018-05-24	2018-07-03		SELP18105G	Sony SELP18105G E PZ 18-105mm F4 G OSS	Electronics	480.18
/	2018-05-24	2018-07-03		SELP18105G	Sony SELP18105G E PZ 18-105mm F4 G OSS	Electronics	480.18
era	2018-05-24	2018-07-03		3554B002	Canon EF 100mm f/2.8L IS USM Macro Lens for Canon Digital S	Electronics	625.14
era	2018-05-24	2018-07-03		3554B002	Canon EF 100mm f/2.8L IS USM Macro Lens for Canon Digital S	Electronics	625.14
era	2018-05-24	2018-07-03		3554B002	Canon EF 100mm f/2.8L IS USM Macro Lens for Canon Digital S	Electronics	625.14
era	2018-05-24	2018-07-03		3554B002	Canon EF 100mm f/2.8L IS USM Macro Lens for Canon Digital S	Electronics	625.14
era	2018-05-24	2018-07-03		3554B002	Canon EF 100mm f/2.8L IS USM Macro Lens for Canon Digital S	Electronics	625.14
er	2018-05-19	2018-06-13		MZ-7KE1T0BW	Samsung - 850 PRO 1TB Internal SATA III Solid State Drive for L	Electronics	62.72
er	2018-05-19	2018-06-13		MZ-7KE1T0BW	Samsung - 850 PRO 1TB Internal SATA III Solid State Drive for L	Electronics	62.72
m	2018-05-07	2018-06-13		HX-HSCF-BK/AM	HyperX - Cloud Flight Wireless Stereo Gaming Headset for PC, F	Electronics	285.39
A	2018-05-02	2018-06-13		980-001300	Logitech - G560 LIGHTSYNC 2.1 Bluetooth Gaming Speakers wi	Electronics	887.88
le	2018-05-01	2018-06-16		UAJBLNBGRY	Under Armour Sport Wireless Flex Neckband In-Ear Headphones	Electronics	327.6
le	2018-05-01	2018-06-16		UAJBLNBGRY	Under Armour Sport Wireless Flex Neckband In-Ear Headphones	Electronics	327.6
le	2018-04-23	2018-06-13		WHCH700N/B	Sony - WH-CH700N Wireless Noise Canceling Over-the-Ear Hea	Electronics	240.09

Final 3 columns after standardizing, renaming and validation

Step 4: Analyzing the data

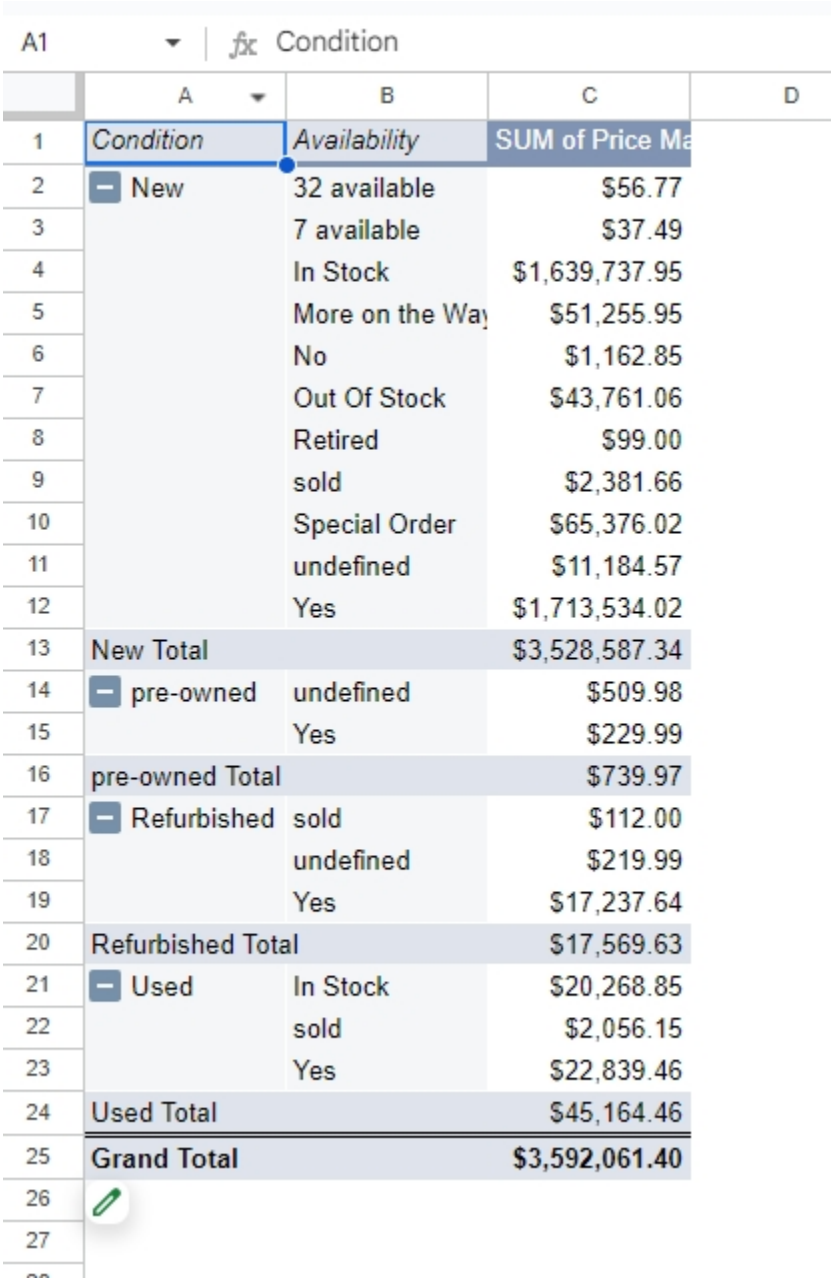
In this process, we want to conduct a simple exploratory data analysis within Google Sheets using pivot tables and some statistical formulae, in order to get some initial insights about the cleaned data. Here is the analysis framework:

- Summary Statistics: Use functions like AVERAGE(), MEDIAN(), MIN(), MAX(), and STDEV() to calculate the basic statistics for columns such as price Max, price Min, and weight. We will create a new sheet to contain this information.

A1

	A	B	C	D	E
1		**Price Max**	**Price Min**	**Weight Std (g)**	
2	**Min**	$1.00	$1.00	0.56	
3	**Max**	$6,999.99	$5,999.99	113250	
4	**Mean**	$495.59	$464.02	4761.273328	
5	**Median**	$198.99	$189.99	1087.2	
6	**Std. Dev.**	763.633317	680.5332298	8481.28711	
7					
8					
9					

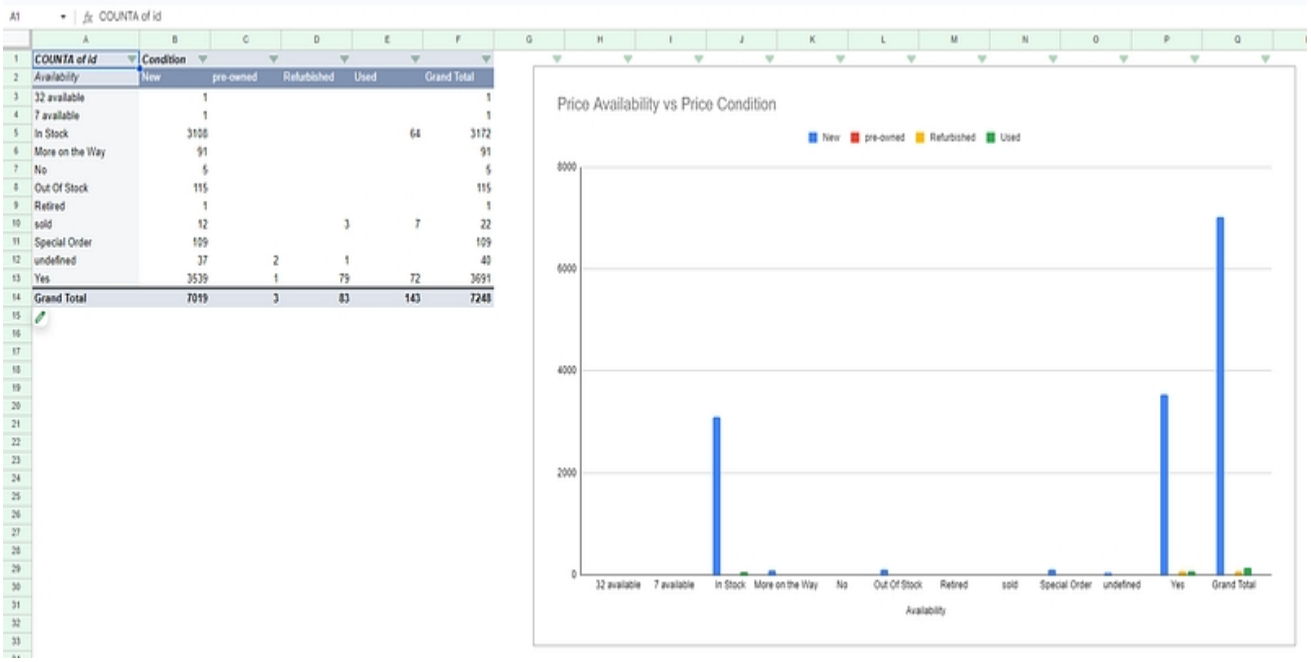
- Identify Trends: Use pivot tables to summarize and identify trends in pricing, availability, and product conditions.



A1 | fx Condition

	A	B	C	D
1	*Condition*	*Availability*	SUM of Price Ma	
2	− New	32 available	$56.77	
3		7 available	$37.49	
4		In Stock	$1,639,737.95	
5		More on the Way	$51,255.95	
6		No	$1,162.85	
7		Out Of Stock	$43,761.06	
8		Retired	$99.00	
9		sold	$2,381.66	
10		Special Order	$65,376.02	
11		undefined	$11,184.57	
12		Yes	$1,713,534.02	
13	New Total		$3,528,587.34	
14	− pre-owned	undefined	$509.98	
15		Yes	$229.99	
16	pre-owned Total		$739.97	
17	− Refurbished	sold	$112.00	
18		undefined	$219.99	
19		Yes	$17,237.64	
20	Refurbished Total		$17,569.63	
21	− Used	In Stock	$20,268.85	
22		sold	$2,056.15	
23		Yes	$22,839.46	
24	Used Total		$45,164.46	
25	**Grand Total**		**$3,592,061.40**	
26				
27				

- Compare Categories: Analyze the differences in prices and availability across different product categories using pivot tables and charts within Google Sheets.

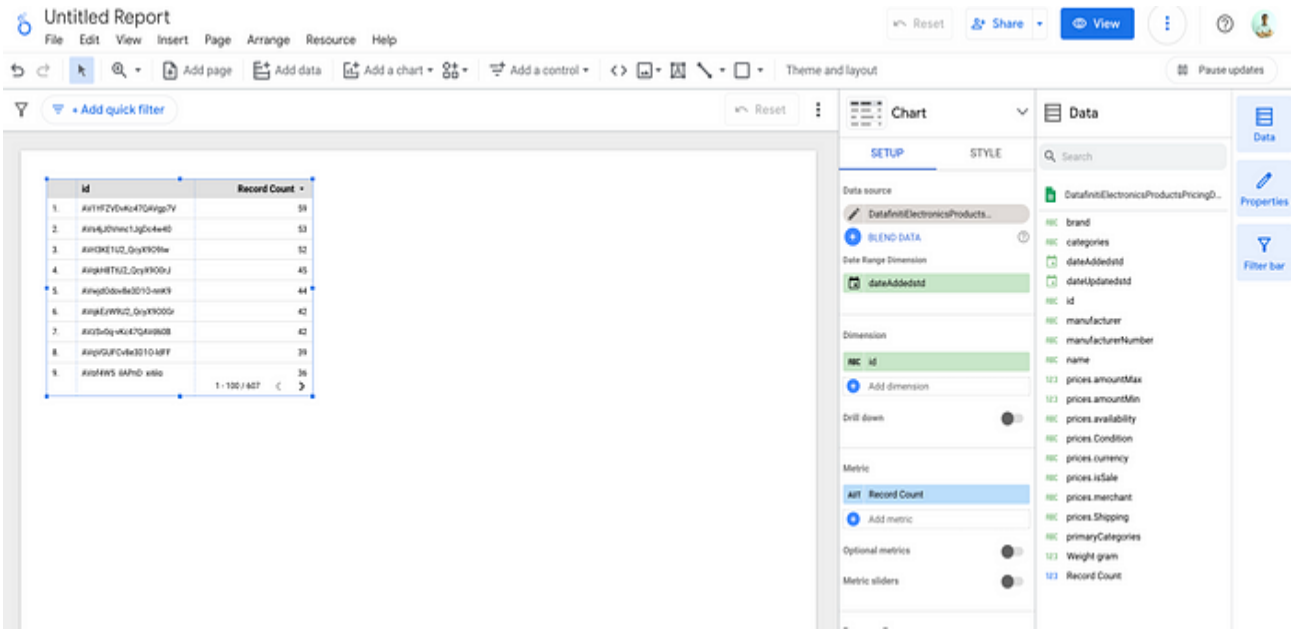


This gives us some initial insights, but we can use a more advanced Visualization package to look deeper

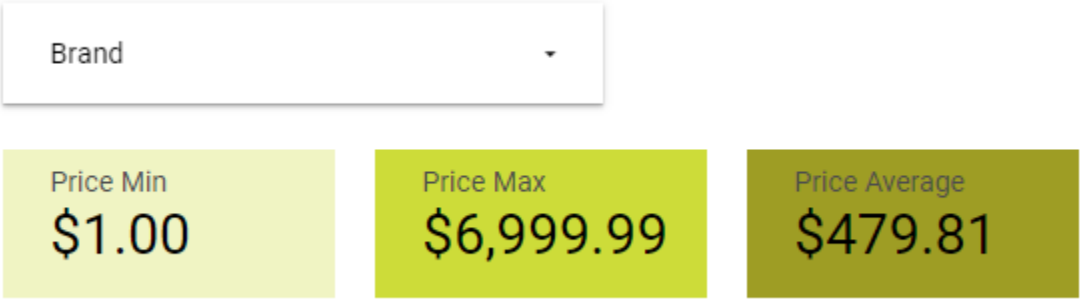
Crafting the Google Looker Studio dashboard

During our initial analysis we found that there are still a lot of information we can extract from our dataset. But, if we stick to Google Sheets it limits our ability to analyze deeper. Thus, we can use Google Looker Studio to create some better visualizations.

- Firstly, we need to make sure that the google sheet is synced with google drive.
- Then, go to https://lookerstudio.google.com and click **Create** (top right corner) > **Report**
- Then under the "Connect to Data" section choose Google Sheets, then "Authorize to access our Google Drive".
- Then we can import the data set from Google Drive, and give the report a title.



- Now, let's go back to our objectives- the first question asked was about **Price Analysis -** What are the min, max and average prices? How do prices vary by brand and category?
- To answer this, we will use the Scorecard Chart viz, combined with a Data Control (dropdown) option, using data from the brand column.

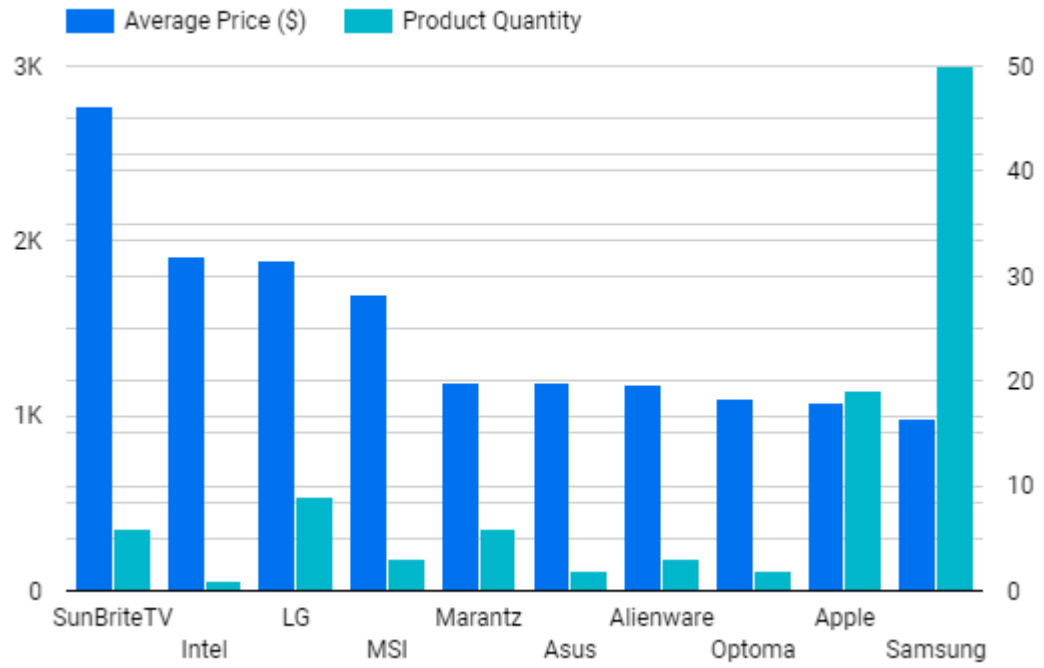


The brand dropdown element allows us to filter the other parts of the viz, and the results are updated on the fly

- The second question was **Availability and Condition -** what conditions are products in, and how does this impact prices?
- We can "Create a Table" to see the full information on this topic

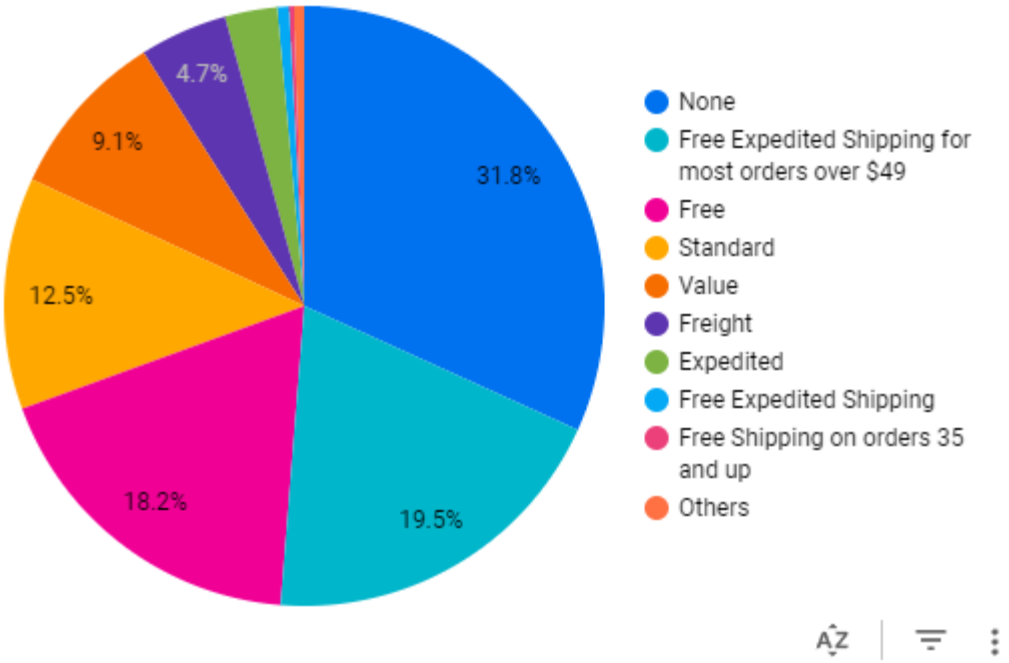
	Condition	Quantity of products ▾	Price Min	Price Max
1.	New	7,019	3,299,753.77	3,528,587.34
2.	Used	143	45,164.46	45,164.46
3.	Refurbished	83	17,569.63	17,569.63
4.	pre-owned	3	739.97	739.97

- Next is **Brand and Manufacturer Insights -** which brands and manufacturers have the most products? What are the average prices for each brand and manufacturer?
- We can use bar chart with multiple Metrics to achieve this:

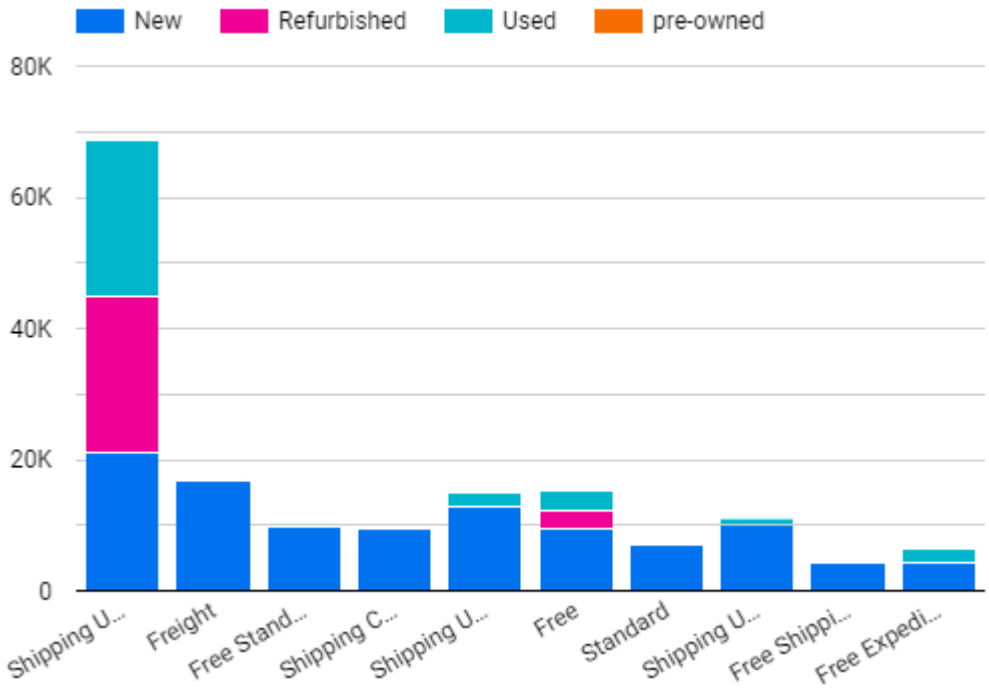


Dimension is “Brand” on the x axis, with “Av. Price” on the left y axis and “Product Quantity” on the right y axis

- **Product Details**: How does product weight relate to pricing and shipping costs?
- To answer this question, we can construct a pie chart to provide a clear breakdown of the data into segments based on shipping type:



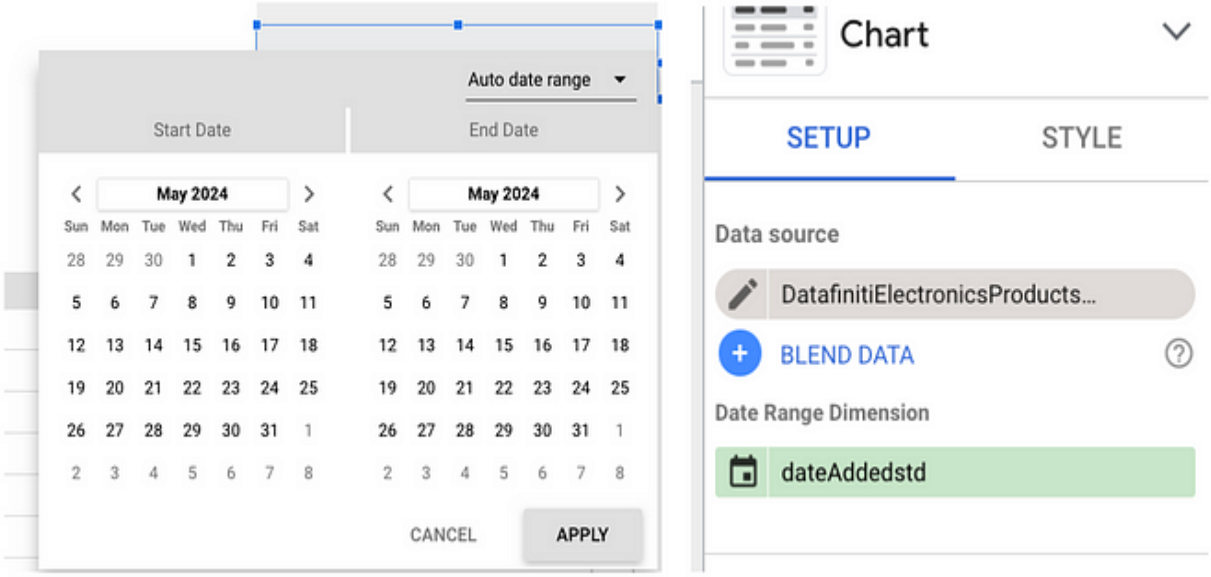
- We can use another bar chart to visualize the shipping vs product weight relationship:



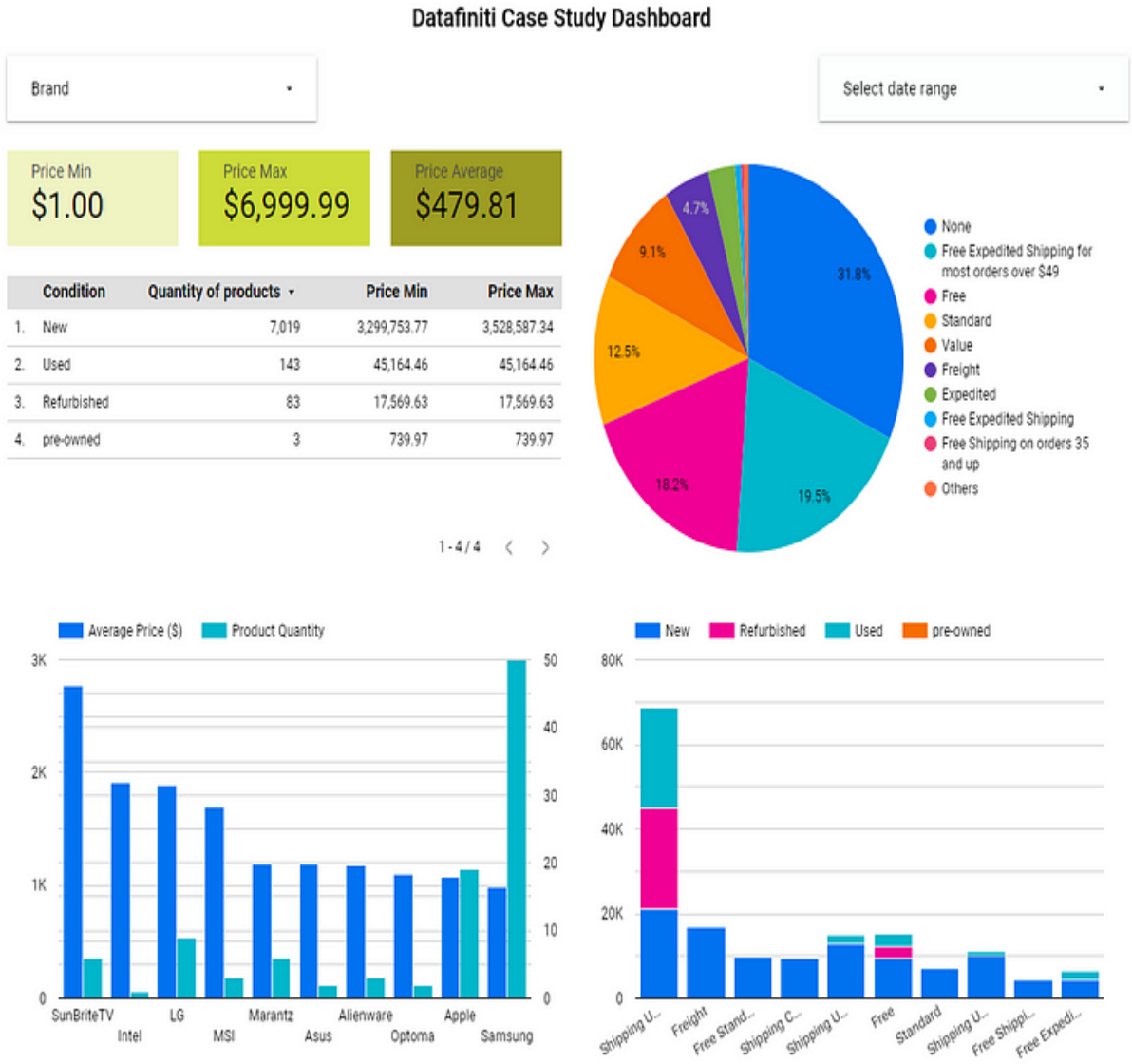
the dimension on the y axis is product weight, in grams

- The final question was **"Time Analysis**: How have prices changed over time?"

- For this, we can use the “Add a Data Control > Date Range” function from the Looker Studio menu. This allows us to filter all the visualizations and tables at once, based on the date range inputted into the data control element.



So, combining all these elements we end up with the final dashboard, which fits concisely onto one page / slide:



The dashboard can be viewed and interacted with via this link:
https://lookerstudio.google.com/reporting/0d5f0d3d-d549-4a60-8529-1483e9091721

Step 5 & 6; Share and Act:

This is the stage where we would normally share our findings from analysis with our superiors / shareholders. Then we would either ask more business questions, and proceed through the analysis stage again, or, if we felt comfortable with the findings, formulate a plan of action and proceed to enact that.

This denotes the end of my second case study :)