Project Proposal
Supplier and Market Place Data Consolidation (ETL Project)

Names of Group Members

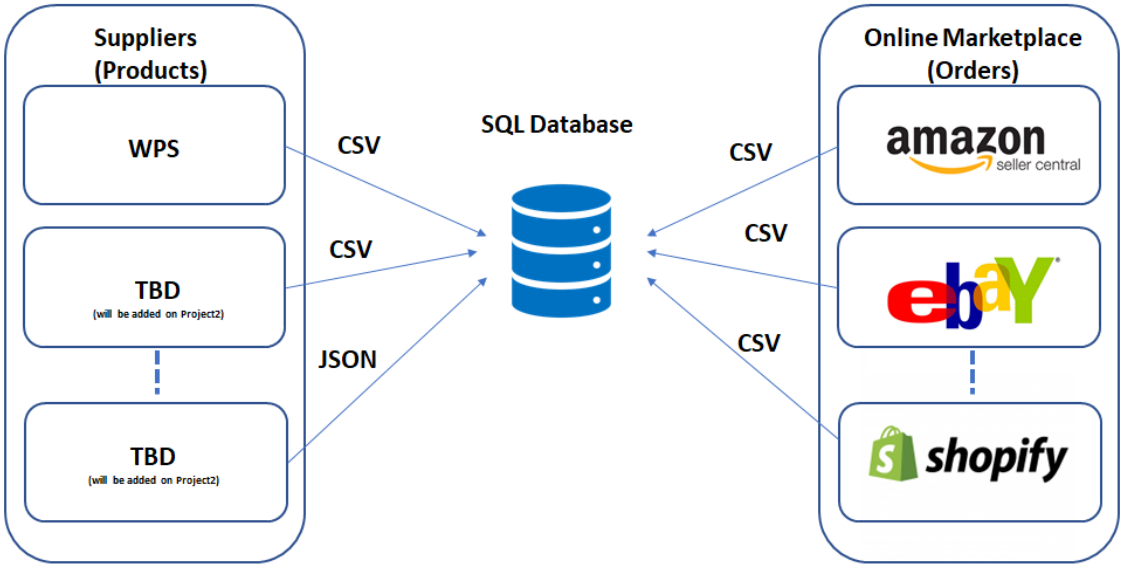
- Shadi Askarzadeh
- Samira Hajimohammadzadeh
- Shadi Salama
- Warren Vince Sy

Synopsis of Project:

Online Sellers have no single view of their Data. All the sales data are in silo for every marketplace they sell. On top of that supplier data is on its own Silo as well. Goal for this project is to consolidate data from different marketplaces and suppliers so that the seller/customer can have a single view of their data.

ETL Project

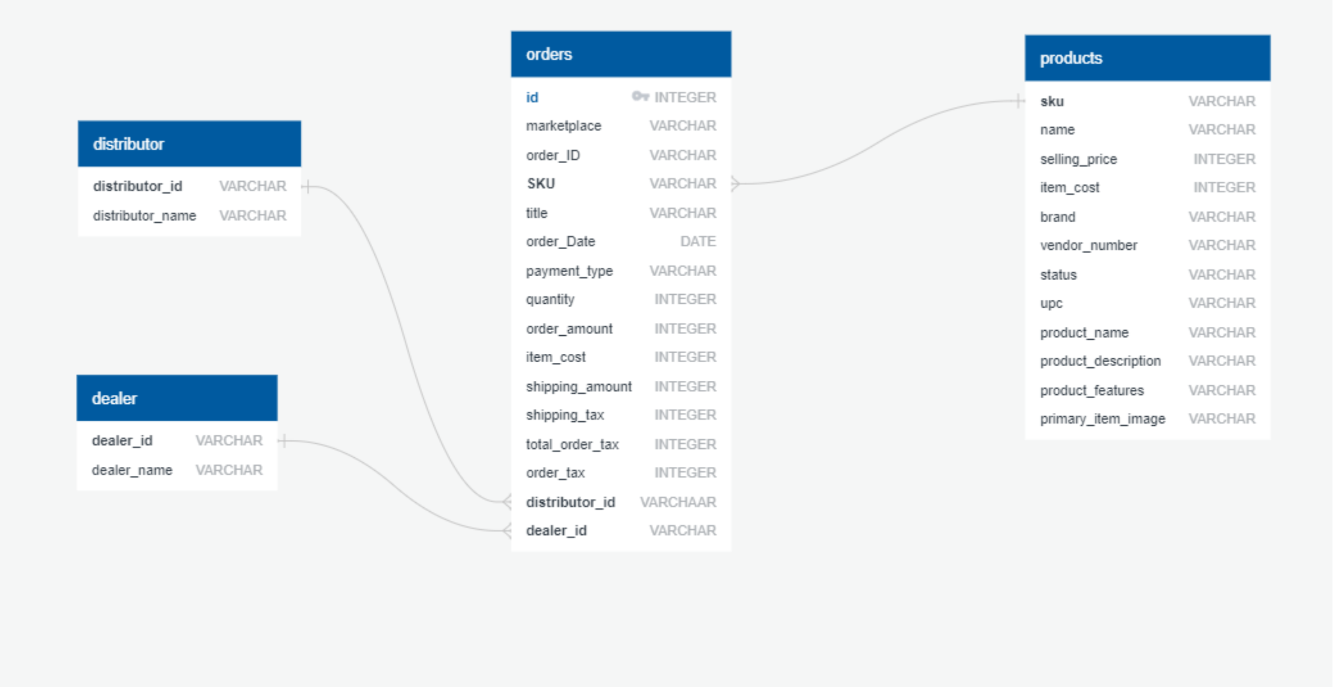
ETL project diagram



Datasets Used:

For this ETL project, we will be extracting data from different market places (eBay, amazon, Shopify) and suppliers (WPS) and saving it on a single database using relational tables.

Database Architecture: *"Relational Database"*



distributor	
distributor_id	VARCHAR
distributor_name	VARCHAR

dealer	
dealer_id	VARCHAR
dealer_name	VARCHAR

orders	
id	INTEGER
marketplace	VARCHAR
order_ID	VARCHAR
SKU	VARCHAR
title	VARCHAR
order_Date	DATE
payment_type	VARCHAR
quantity	INTEGER
order_amount	INTEGER
item_cost	INTEGER
shipping_amount	INTEGER
shipping_tax	INTEGER
total_order_tax	INTEGER
order_tax	INTEGER
distributor_id	VARCHAAR
dealer_id	VARCHAR

products	
sku	VARCHAR
name	VARCHAR
selling_price	INTEGER
item_cost	INTEGER
brand	VARCHAR
vendor_number	VARCHAR
status	VARCHAR
upc	VARCHAR
product_name	VARCHAR
product_description	VARCHAR
product_features	VARCHAR
primary_item_image	VARCHAR

We created foure tables in the database

Orders Table – It will store all the order data from the different market places. We also added two additional columns (distributor_id and dealer_id) which are foreign keys to two master tables. We added these tables as placeholders to be able to accommodate new dealers and distributors.

Products Table – It will store all the products the sellers are selling on the market place.

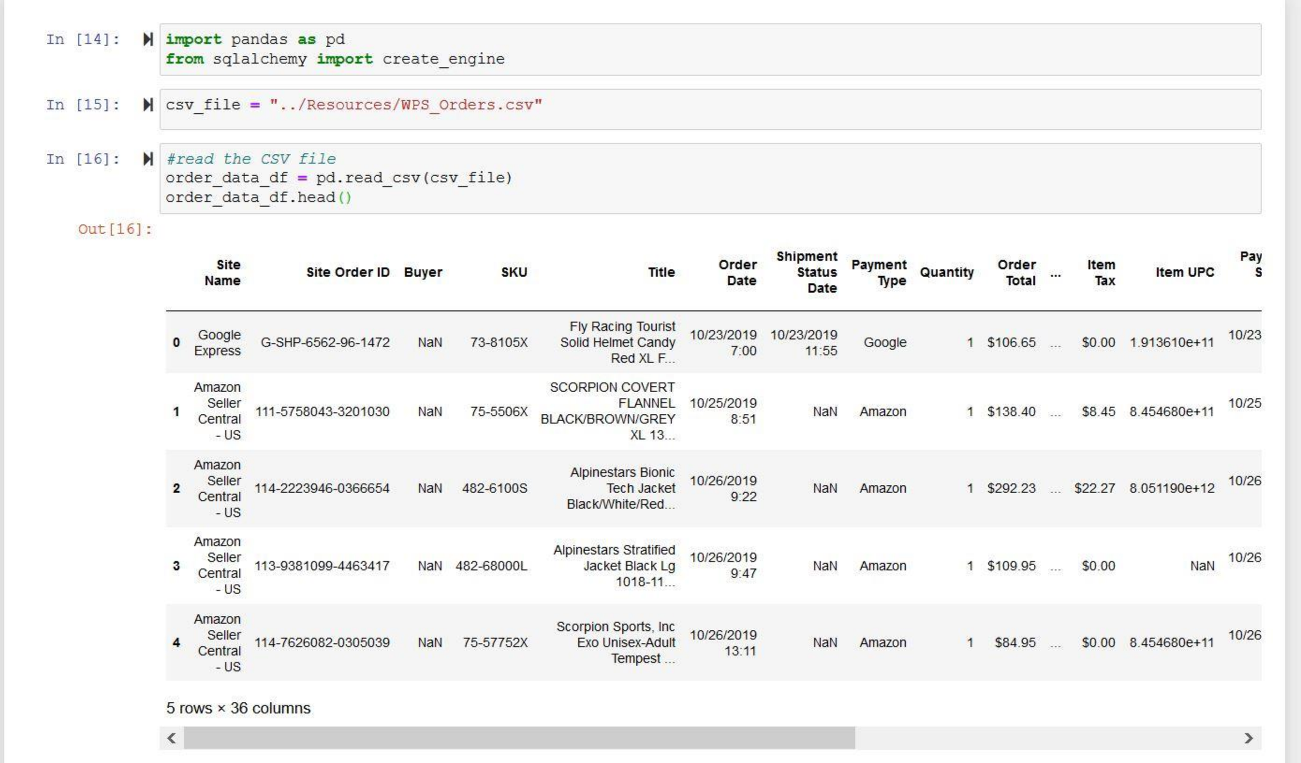
Dealer Table – It will store all Dealer Data. We will add more dealer related information on project 2.

Distributor Table - It will store all Distributor Data. We will add more distributor related information on project 2.

Data Cleansing

Set up Orders table:

1. Extract Data from CSV file and convert it to Data Frame



```
In [14]:  import pandas as pd
          from sqlalchemy import create_engine
```

```
In [15]:  csv_file = "../Resources/WPS_Orders.csv"
```

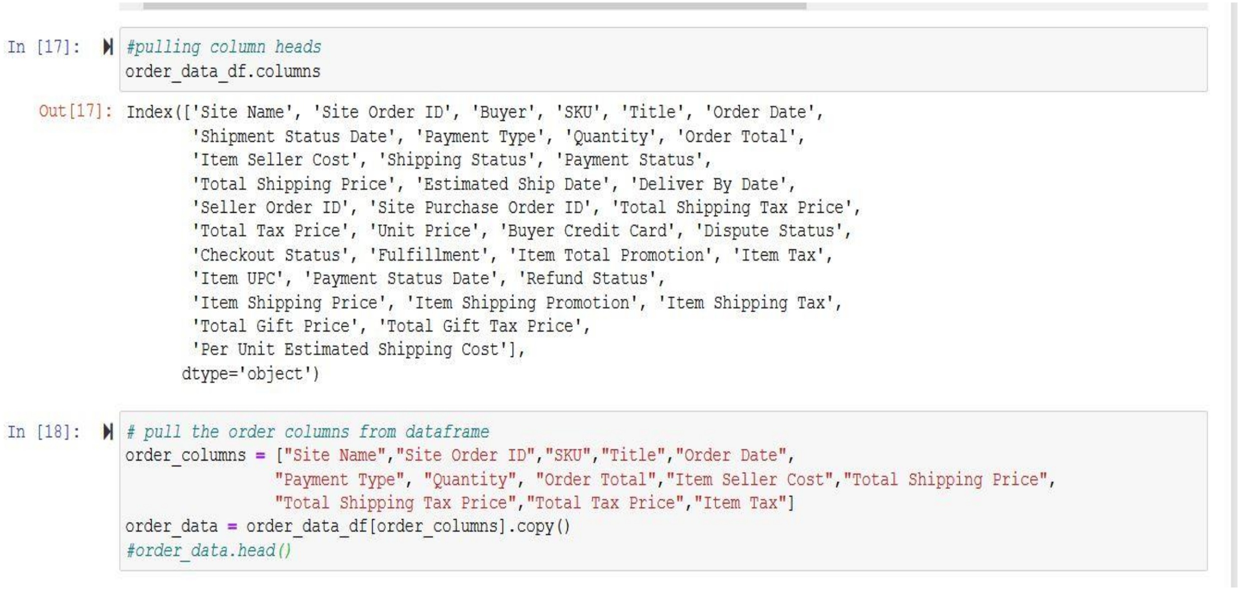
```
In [16]:  #read the CSV file
          order_data_df = pd.read_csv(csv_file)
          order_data_df.head()
```

Out[16]:

	Site Name	Site Order ID	Buyer	SKU	Title	Order Date	Shipment Status Date	Payment Type	Quantity	Order Total	...	Item Tax	Item UPC	Pay S
0	Google Express	G-SHP-6562-96-1472	NaN	73-8105X	Fly Racing Tourist Solid Helmet Candy Red XL F...	10/23/2019 7:00	10/23/2019 11:55	Google	1	$106.65	...	$0.00	1.913610e+11	10/23
1	Amazon Seller Central - US	111-5758043-3201030	NaN	75-5506X	SCORPION COVERT FLANNEL BLACK/BROWN/GREY XL 13...	10/25/2019 8:51	NaN	Amazon	1	$138.40	...	$8.45	8.454680e+11	10/25
2	Amazon Seller Central - US	114-2223946-0366654	NaN	482-6100S	Alpinestars Bionic Tech Jacket Black/White/Red...	10/26/2019 9:22	NaN	Amazon	1	$292.23	...	$22.27	8.051190e+12	10/26
3	Amazon Seller Central - US	113-9381099-4463417	NaN	482-68000L	Alpinestars Stratified Jacket Black Lg 1018-11...	10/26/2019 9:47	NaN	Amazon	1	$109.95	...	$0.00	NaN	10/26
4	Amazon Seller Central - US	114-7626082-0305039	NaN	75-57752X	Scorpion Sports, Inc Exo Unisex-Adult Tempest ...	10/26/2019 13:11	NaN	Amazon	1	$84.95	...	$0.00	8.454680e+11	10/26

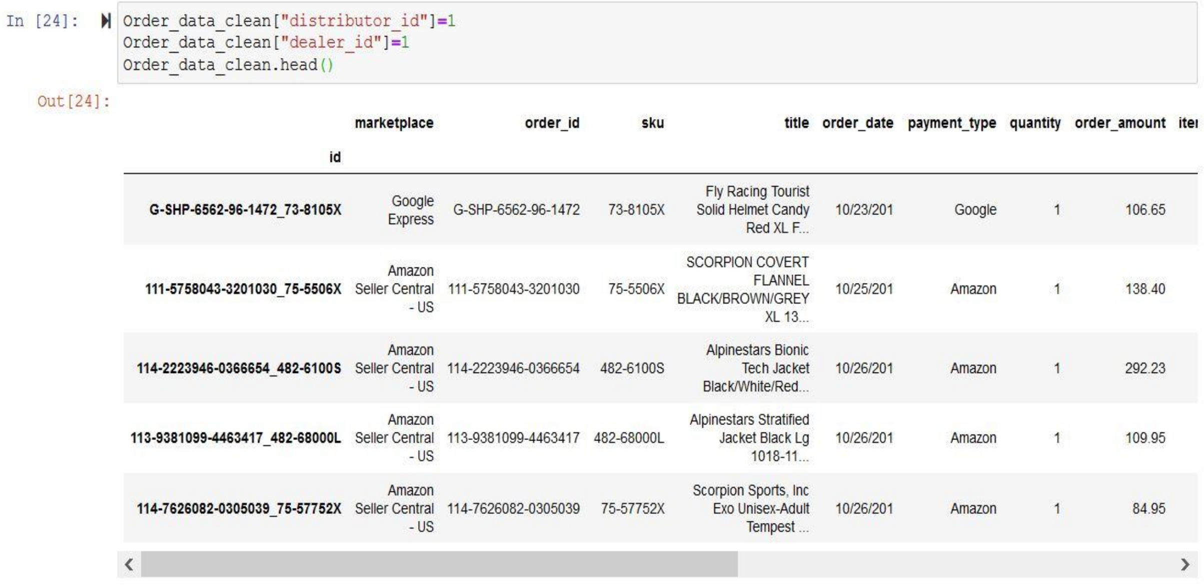
5 rows × 36 columns

2. Clean data before loading to database



```
In [17]:  #pulling column heads
          order_data_df.columns

Out[17]: Index(['Site Name', 'Site Order ID', 'Buyer', 'SKU', 'Title', 'Order Date',
               'Shipment Status Date', 'Payment Type', 'Quantity', 'Order Total',
               'Item Seller Cost', 'Shipping Status', 'Payment Status',
               'Total Shipping Price', 'Estimated Ship Date', 'Deliver By Date',
               'Seller Order ID', 'Site Purchase Order ID', 'Total Shipping Tax Price',
               'Total Tax Price', 'Unit Price', 'Buyer Credit Card', 'Dispute Status',
               'Checkout Status', 'Fulfillment', 'Item Total Promotion', 'Item Tax',
               'Item UPC', 'Payment Status Date', 'Refund Status',
               'Item Shipping Price', 'Item Shipping Promotion', 'Item Shipping Tax',
               'Total Gift Price', 'Total Gift Tax Price',
               'Per Unit Estimated Shipping Cost'],
              dtype='object')
```

```
In [18]:  # pull the order columns from dataframe
          order_columns = ["Site Name","Site Order ID","SKU","Title","Order Date",
                          "Payment Type", "Quantity", "Order Total","Item Seller Cost","Total Shipping Price",
                          "Total Shipping Tax Price","Total Tax Price","Item Tax"]
          order_data = order_data_df[order_columns].copy()
          #order_data.head()
```



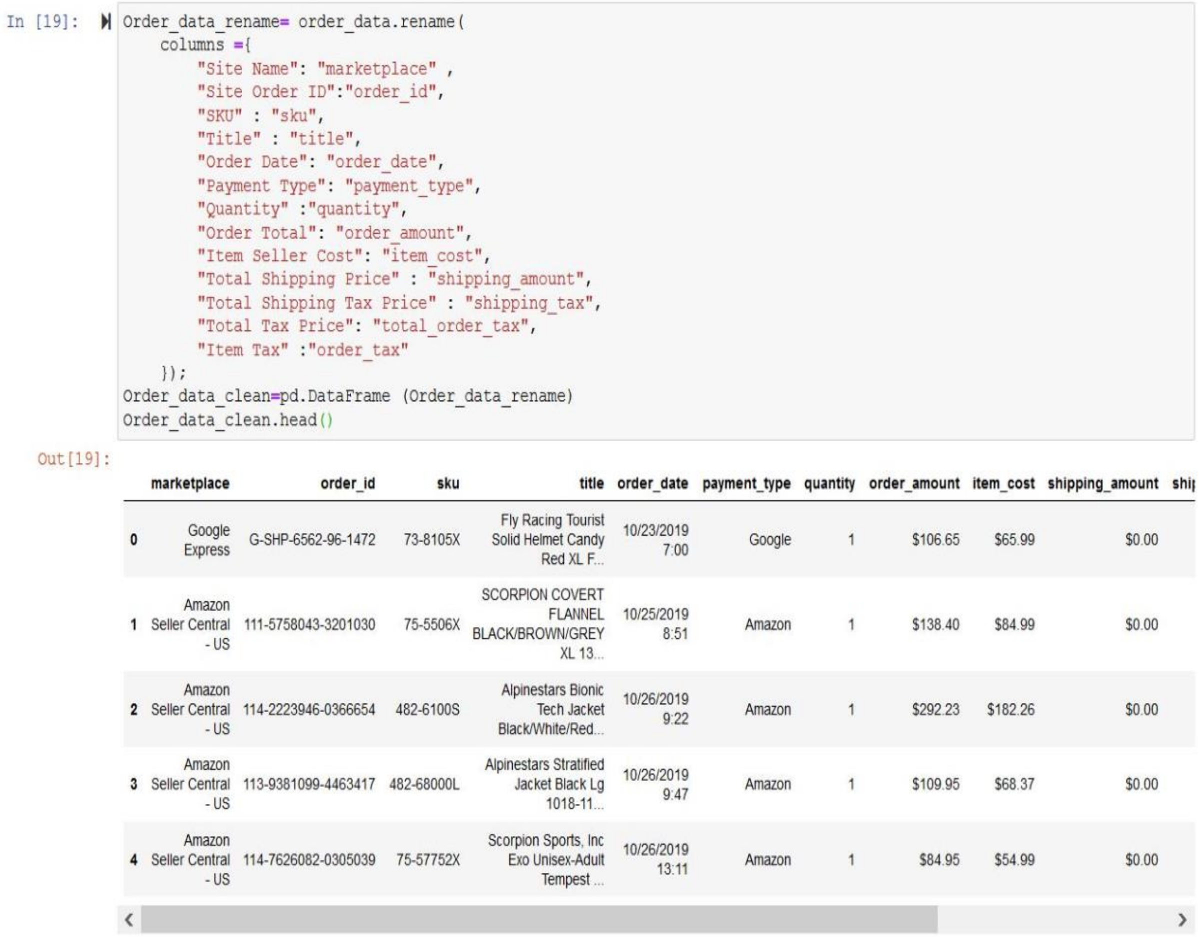
In [24]:

```
Order_data_clean["distributor_id"]=1
Order_data_clean["dealer_id"]=1
Order_data_clean.head()
```

Out[24]:

	marketplace	order_id	sku	title	order_date	payment_type	quantity	order_amount	iter
id									
G-SHP-6562-96-1472_73-8105X	Google Express	G-SHP-6562-96-1472	73-8105X	Fly Racing Tourist Solid Helmet Candy Red XL F...	10/23/201	Google	1	106.65	
111-5758043-3201030_75-5506X	Amazon Seller Central - US	111-5758043-3201030	75-5506X	SCORPION COVERT FLANNEL BLACK/BROWN/GREY XL 13...	10/25/201	Amazon	1	138.40	
114-2223946-0366654_482-6100S	Amazon Seller Central - US	114-2223946-0366654	482-6100S	Alpinestars Bionic Tech Jacket Black/White/Red...	10/26/201	Amazon	1	292.23	
113-9381099-4463417_482-68000L	Amazon Seller Central - US	113-9381099-4463417	482-68000L	Alpinestars Stratified Jacket Black Lg 1018-11...	10/26/201	Amazon	1	109.95	
114-7626082-0305039_75-57752X	Amazon Seller Central - US	114-7626082-0305039	75-57752X	Scorpion Sports, Inc Exo Unisex-Adult Tempest ...	10/26/201	Amazon	1	84.95	

CLEAN DATA



In [19]:

```
Order_data_rename= order_data.rename(
    columns ={
        "Site Name": "marketplace" ,
        "Site Order ID":"order_id",
        "SKU" : "sku",
        "Title" : "title",
        "Order Date": "order_date",
        "Payment Type": "payment_type",
        "Quantity" :"quantity",
        "Order Total": "order_amount",
        "Item Seller Cost": "item_cost",
        "Total Shipping Price" : "shipping_amount",
        "Total Shipping Tax Price" : "shipping_tax",
        "Total Tax Price": "total_order_tax",
        "Item Tax" :"order_tax"
    });
Order_data_clean=pd.DataFrame (Order_data_rename)
Order_data_clean.head()
```

Out[19]:

	marketplace	order_id	sku	title	order_date	payment_type	quantity	order_amount	item_cost	shipping_amount	shi
0	Google Express	G-SHP-6562-96-1472	73-8105X	Fly Racing Tourist Solid Helmet Candy Red XL F...	10/23/2019 7:00	Google	1	$106.65	$65.99	$0.00	
1	Amazon Seller Central - US	111-5758043-3201030	75-5506X	SCORPION COVERT FLANNEL BLACK/BROWN/GREY XL 13...	10/25/2019 8:51	Amazon	1	$138.40	$84.99	$0.00	
2	Amazon Seller Central - US	114-2223946-0366654	482-6100S	Alpinestars Bionic Tech Jacket Black/White/Red...	10/26/2019 9:22	Amazon	1	$292.23	$182.26	$0.00	
3	Amazon Seller Central - US	113-9381099-4463417	482-68000L	Alpinestars Stratified Jacket Black Lg 1018-11...	10/26/2019 9:47	Amazon	1	$109.95	$68.37	$0.00	
4	Amazon Seller Central - US	114-7626082-0305039	75-57752X	Scorpion Sports, Inc Exo Unisex-Adult Tempest ...	10/26/2019 13:11	Amazon	1	$84.95	$54.99	$0.00	

In [23]:

```
#take out the time from the "order date" column
Order_data_clean['order_date']=Order_data_clean['order_date'].map(lambda x: str(x)[0:9])
Order_data_clean.head()
```

Out[23]:

	marketplace	order_id	sku	title	order_date	payment_type	quantity	order_amount	ite
id									
G-SHP-6562-96-1472_73-8105X	Google Express	G-SHP-6562-96-1472	73-8105X	Fly Racing Tourist Solid Helmet Candy Red XL F...	10/23/201	Google	1	106.65	
111-5758043-3201030_75-5506X	Amazon Seller Central - US	111-5758043-3201030	75-5506X	SCORPION COVERT FLANNEL BLACK/BROWN/GREY XL 13...	10/25/201	Amazon	1	138.40	
114-2223946-0366654_482-6100S	Amazon Seller Central - US	114-2223946-0366654	482-6100S	Alpinestars Bionic Tech Jacket Black/White/Red...	10/26/201	Amazon	1	292.23	
113-9381099-4463417_482-68000L	Amazon Seller Central - US	113-9381099-4463417	482-68000L	Alpinestars Stratified Jacket Black Lg 1018-11...	10/26/201	Amazon	1	109.95	
114-7626082-0305039_75-57752X	Amazon Seller Central - US	114-7626082-0305039	75-57752X	Scorpion Sports, Inc Exo Unisex-Adult Tempest ...	10/26/201	Amazon	1	84.95	

In [20]:

```
print(Order_data_clean.dtypes)
```

```
marketplace        object
order_id           object
sku                object
title              object
order_date         object
payment_type       object
quantity            int64
order_amount       object
item_cost          object
shipping_amount    object
shipping_tax       object
total_order_tax    object
order_tax          object
dtype: object
```

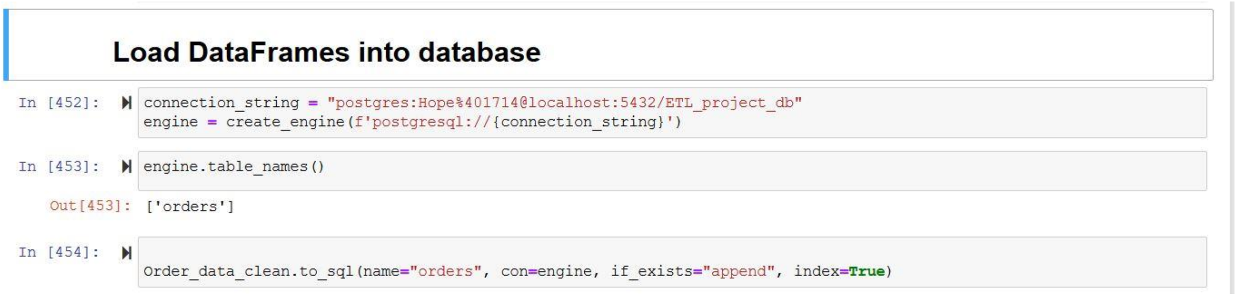
In [21]:

```
Order_data_clean["order_amount"] = Order_data_clean["order_amount"].str.replace(",","")
Order_data_clean["order_amount"] = Order_data_clean["order_amount"].str.replace("$","").astype(float)


Order_data_clean["item_cost"] = Order_data_clean["item_cost"].str.replace(",","")
Order_data_clean["item_cost"] = Order_data_clean["item_cost"].str.replace("$","").astype(float)

Order_data_clean["shipping_amount"] = Order_data_clean["shipping_amount"].str.replace(",","")
Order_data_clean["shipping_amount"] = Order_data_clean["shipping_amount"].str.replace("$","").astype(float)

Order_data_clean["shipping_tax"] = Order_data_clean["shipping_tax"].str.replace(",","")
Order_data_clean["shipping_tax"] = Order_data_clean["shipping_tax"].str.replace("$","").astype(float)

Order_data_clean["total_order_tax"] = Order_data_clean["total_order_tax"].str.replace(",","")
Order_data_clean["total_order_tax"] = Order_data_clean["total_order_tax"].str.replace("$","").astype(float)

Order_data_clean["order_tax"] = Order_data_clean["order_tax"].str.replace(",","")
Order_data_clean["order_tax"] = Order_data_clean["order_tax"].str.replace("$","").astype(float)
```

In [22]:

```
# create the "id" column by concatenate the order_ID & SKU column
Order_data_clean["id"] = Order_data_clean['order_id'].map(str)+'_'+Order_data_clean['sku']
Order_data_clean.drop_duplicates('id', inplace=True)
Order_data_clean.set_index("id", inplace = True)
Order_data_clean.head()
```

3. Load the data to SQL database. The tables type has been created in the pgAdmin in advance.

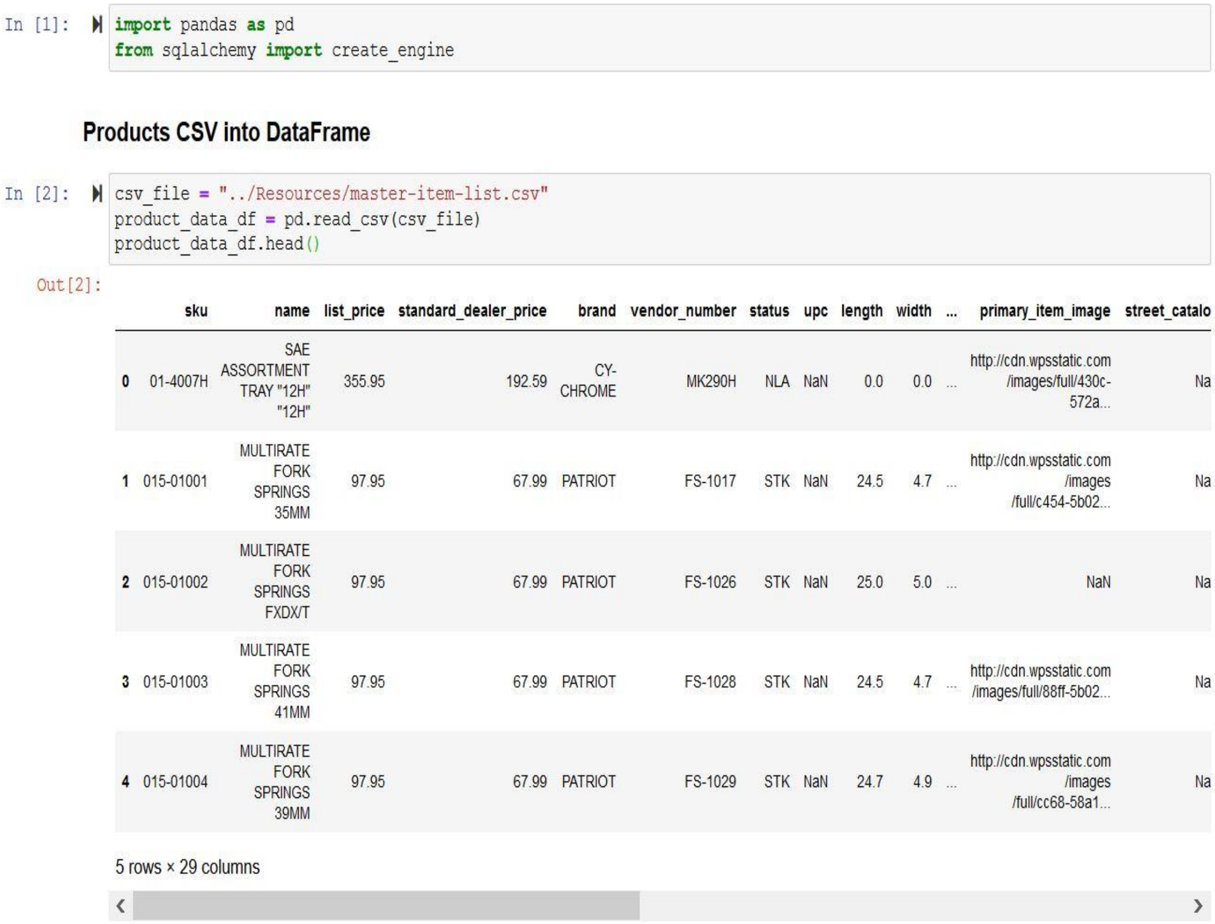


Load DataFrames into database

```
In [452]:  connection_string = "postgres:Hope%401714@localhost:5432/ETL_project_db"
           engine = create_engine(f'postgresql://{connection_string}')

In [453]:  engine.table_names()

Out[453]:  ['orders']

In [454]:  
           Order_data_clean.to_sql(name="orders", con=engine, if_exists="append", index=True)
```

Set up products table (Supplier File)***:***

1. Extract Data from CSV file and convert it to Data Frame



```
In [1]:  import pandas as pd
         from sqlalchemy import create_engine
```

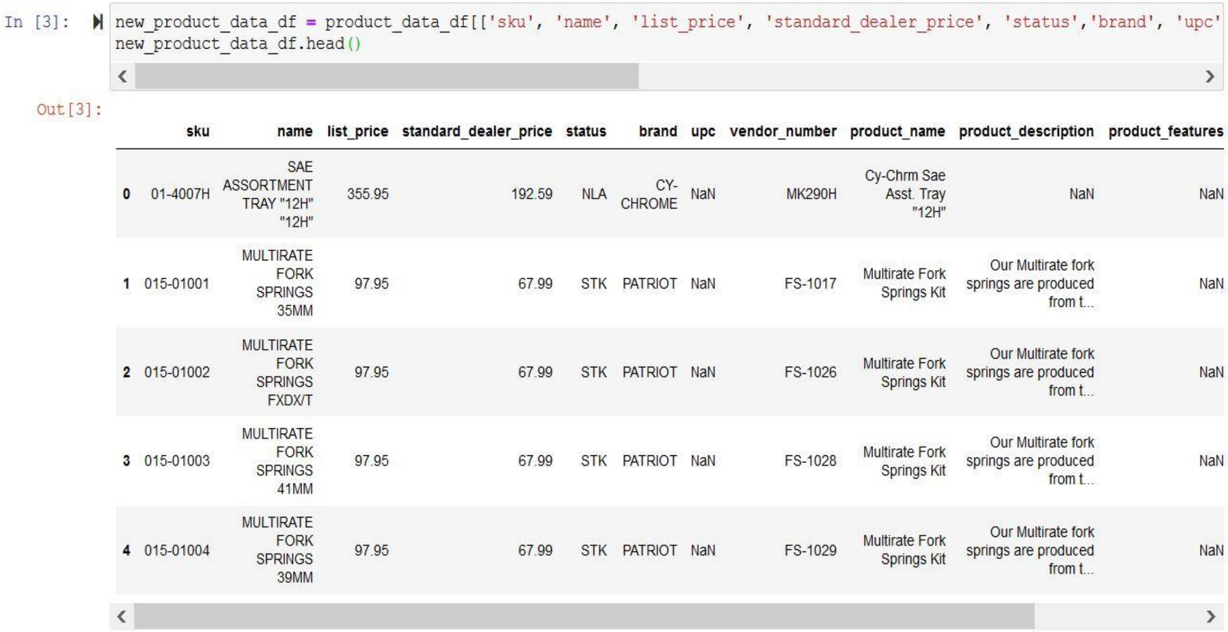
Products CSV into DataFrame

```
In [2]:  csv_file = "../Resources/master-item-list.csv"
          product_data_df = pd.read_csv(csv_file)
          product_data_df.head()

Out[2]:
```

	sku	name	list_price	standard_dealer_price	brand	vendor_number	status	upc	length	width	...	primary_item_image	street_catalo
0	01-4007H	SAE ASSORTMENT TRAY "12H" "12H"	355.95	192.59	CY-CHROME	MK290H	NLA	NaN	0.0	0.0	...	http://cdn.wpsstatic.com /images/full/430c-572a...	Na
1	015-01001	MULTIRATE FORK SPRINGS 35MM	97.95	67.99	PATRIOT	FS-1017	STK	NaN	24.5	4.7	...	http://cdn.wpsstatic.com /images /full/c454-5b02...	Na
2	015-01002	MULTIRATE FORK SPRINGS FXDX/T	97.95	67.99	PATRIOT	FS-1026	STK	NaN	25.0	5.0	...	NaN	Na
3	015-01003	MULTIRATE FORK SPRINGS 41MM	97.95	67.99	PATRIOT	FS-1028	STK	NaN	24.5	4.7	...	http://cdn.wpsstatic.com /images/full/88ff-5b02...	Na
4	015-01004	MULTIRATE FORK SPRINGS 39MM	97.95	67.99	PATRIOT	FS-1029	STK	NaN	24.7	4.9	...	http://cdn.wpsstatic.com /images /full/cc68-58a1...	Na

5 rows × 29 columns

New product data with select columns



In [3]:

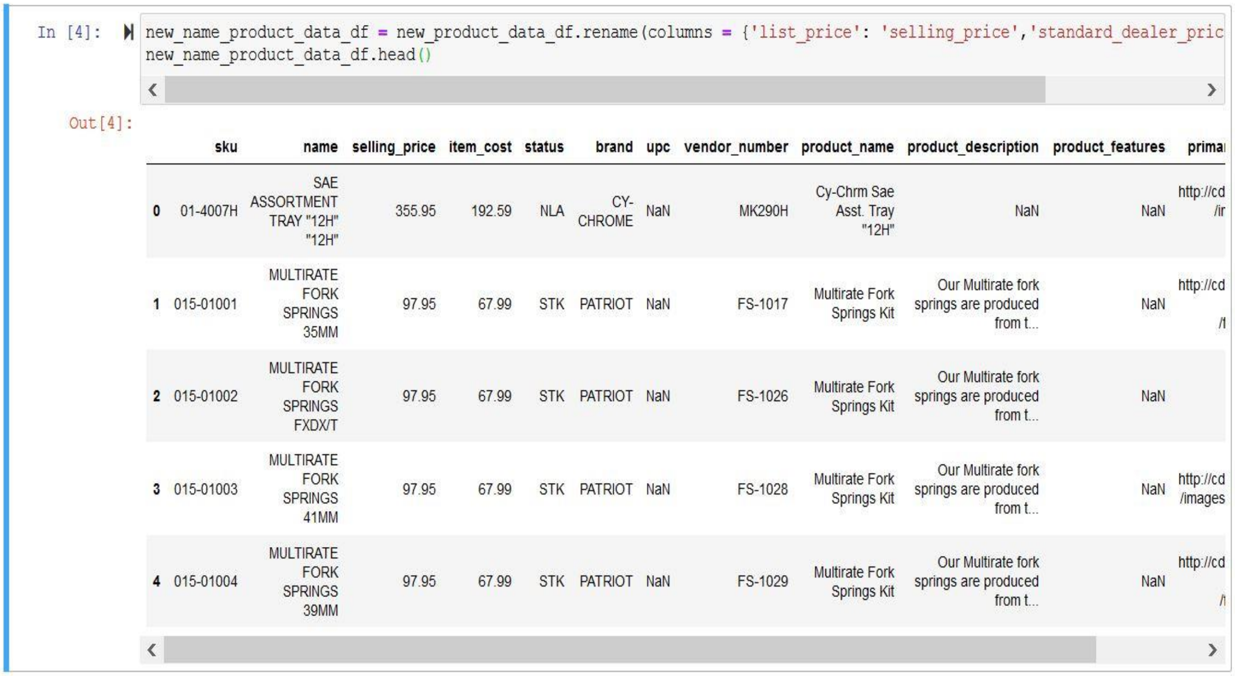
```
new_product_data_df = product_data_df[['sku', 'name', 'list_price', 'standard_dealer_price', 'status','brand', 'upc'
new_product_data_df.head()
```

Out[3]:

	sku	name	list_price	standard_dealer_price	status	brand	upc	vendor_number	product_name	product_description	product_features
0	01-4007H	SAE ASSORTMENT TRAY "12H" "12H"	355.95	192.59	NLA	CY-CHROME	NaN	MK290H	Cy-Chrm Sae Asst. Tray "12H"	NaN	NaN
1	015-01001	MULTIRATE FORK SPRINGS 35MM	97.95	67.99	STK	PATRIOT	NaN	FS-1017	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN
2	015-01002	MULTIRATE FORK SPRINGS FXDX/T	97.95	67.99	STK	PATRIOT	NaN	FS-1026	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN
3	015-01003	MULTIRATE FORK SPRINGS 41MM	97.95	67.99	STK	PATRIOT	NaN	FS-1028	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN
4	015-01004	MULTIRATE FORK SPRINGS 39MM	97.95	67.99	STK	PATRIOT	NaN	FS-1029	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN

2. Clean data before loading to database

Rename column name



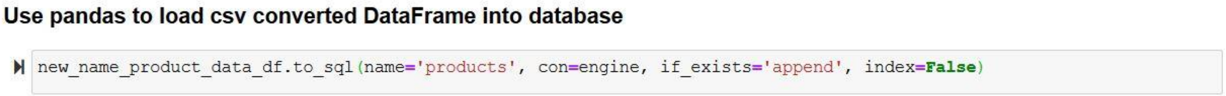
In [4]:

```
new_name_product_data_df = new_product_data_df.rename(columns = {'list_price': 'selling_price','standard_dealer_pric
new_name_product_data_df.head()
```

Out[4]:

	sku	name	selling_price	item_cost	status	brand	upc	vendor_number	product_name	product_description	product_features	primai
0	01-4007H	SAE ASSORTMENT TRAY "12H" "12H"	355.95	192.59	NLA	CY-CHROME	NaN	MK290H	Cy-Chrm Sae Asst. Tray "12H"	NaN	NaN	http://cd /ir
1	015-01001	MULTIRATE FORK SPRINGS 35MM	97.95	67.99	STK	PATRIOT	NaN	FS-1017	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN	http://cd /t
2	015-01002	MULTIRATE FORK SPRINGS FXDX/T	97.95	67.99	STK	PATRIOT	NaN	FS-1026	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN	
3	015-01003	MULTIRATE FORK SPRINGS 41MM	97.95	67.99	STK	PATRIOT	NaN	FS-1028	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN	http://cd /images
4	015-01004	MULTIRATE FORK SPRINGS 39MM	97.95	67.99	STK	PATRIOT	NaN	FS-1029	Multirate Fork Springs Kit	Our Multirate fork springs are produced from t...	NaN	http://cd /t

3. Load data to database



Use pandas to load csv converted DataFrame into database

```
new_name_product_data_df.to_sql(name='products', con=engine, if_exists='append', index=False)
```