CIS 4400

Final Project

William Hall

Group #1

Paul Sung Rhee, Daria Gurova

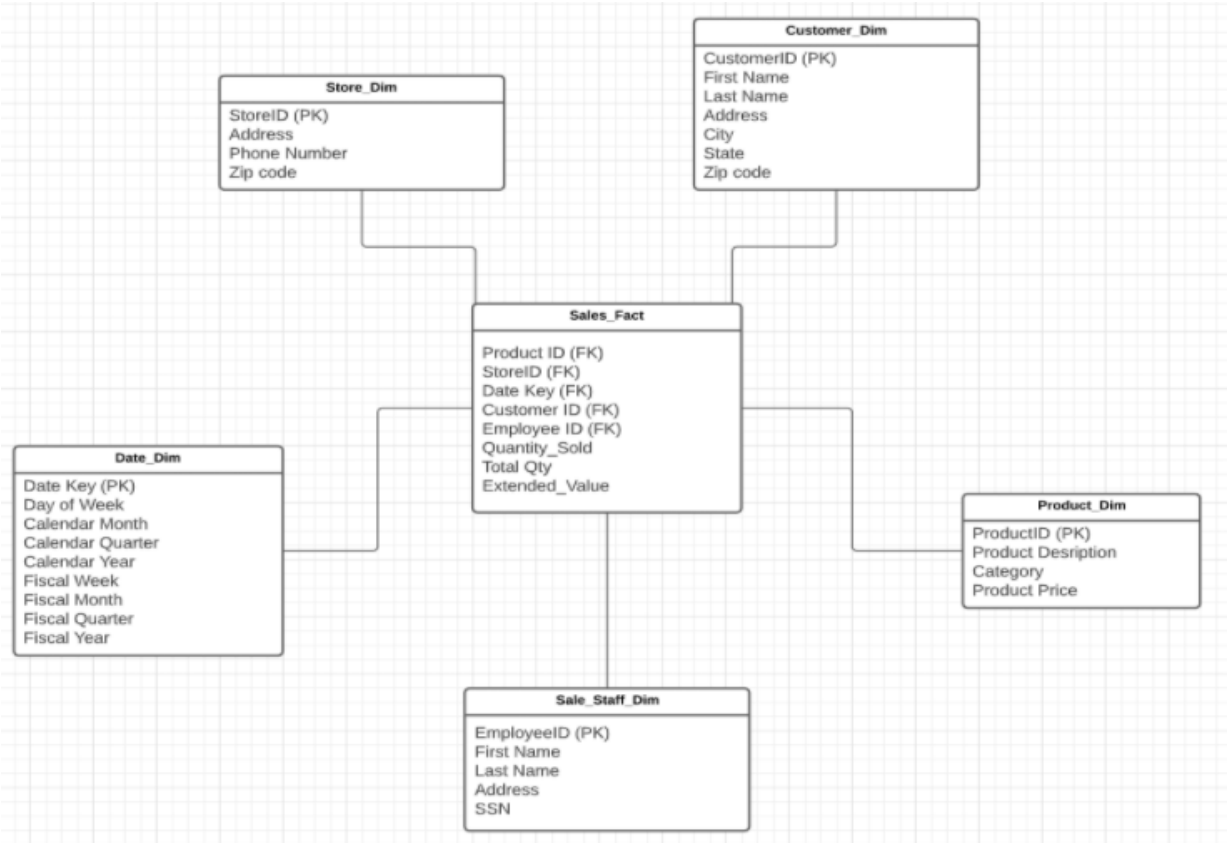
The Business/Organization & Opportunity

This business main products consist of office supplies and office furniture. From our company's opening in 2016, we have built our original store into a chain of 9 scattered around the 5 boroughs. We also have a successful online store that has sales from all over the country.

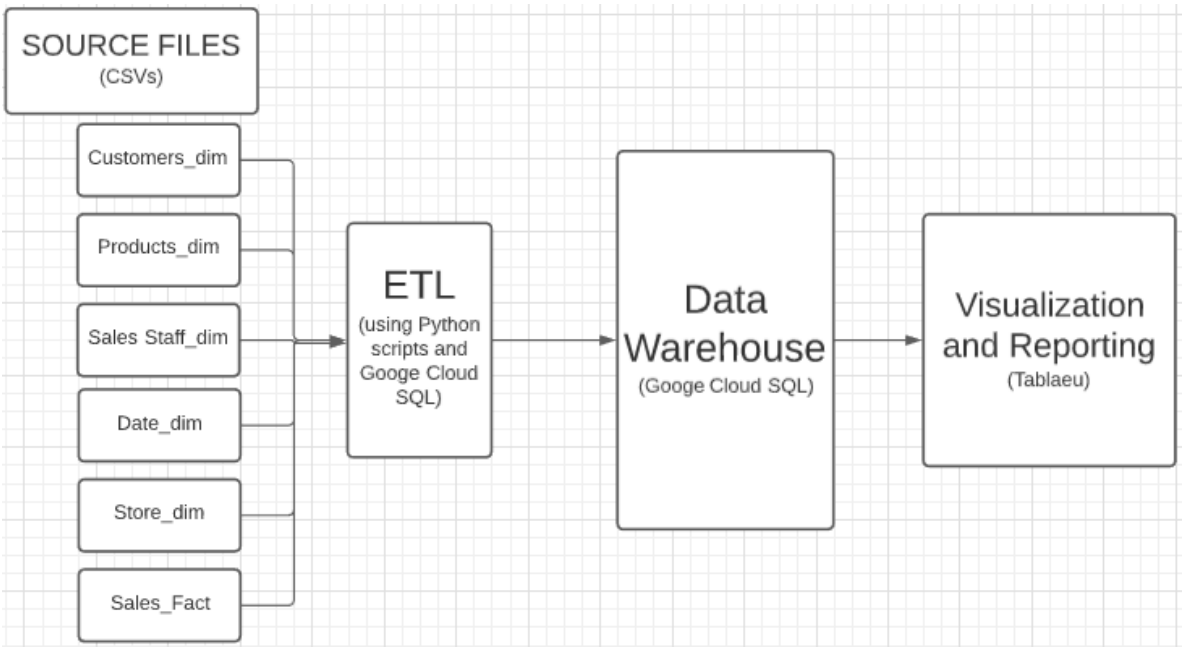
Problem

Our company's old system that we used since its opening has been flawed with unorganized data. Currently, we are utilizing new tools to better capture and organize. In this dry run using our new tools, we are analyzing our first 100,000 transactions to focus on improvements to increase sales and make better decisions in the future.

Logical Data Model



Architecture Diagram



1. SOURCE – We are currently using flat files of company information from 2016-2020. This data was created by the Mockaroo data generators at mockaroo.com.
2. ETL – Source files will be cleaned and transformed to be loaded using Python scripts to the Data Warehouse (Google Cloud SQL).
3. DATA WAREHOUSE - Google Cloud SQL virtual machine.
4. DATA MINING/VISUALIZATION – We will use Tableau for visualizations and reporting.

Detailed Design

1. Flat Files/CSV's – These are our source files that have all of our company information from 2016-2020.
2. PYTHON – We will be using Python to load our transformed source files to the Google Cloud.
3. GOOGLE CLOUD SQL – This will be used for our data warehouse and report querying tool. Alternatives include Oracle and MongoDB.
4. Tableau – Will be used for visualizations and BI reporting. Alternatives include Python (using Matplotlib and Bokeh), R or PowerBI

Risks

Main "risk zones" for this project mainly consist of:

1. Names and addresses of our customers using our online store
2. Credit/Debit card information linked to our customer accounts
3. The type of products our customers are purchasing

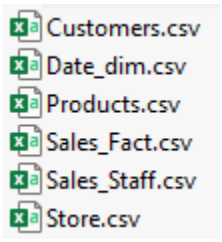
Solutions to these risks could be remedied by using security software to prevent the theft of data. Data encryption must also be used to encrypt data in the case of theft by unauthorized users.

Documentation of the ETL Process

NOTE: Related files can be found in the GITHUB Repository (https://github.com/whall411/CIS4400-Final-Project.git)

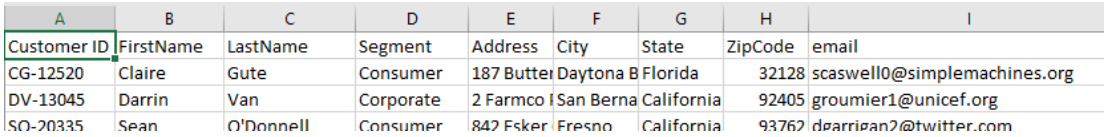
EXTRACT

1. Source files from our business
   a.



TRANSFORM

1. Data Cleansing to be performed
   a. Customers



A	B	C	D	E	F	G	H	I
Customer ID	FirstName	LastName	Segment	Address	City	State	ZipCode	email
CG-12520	Claire	Gute	Consumer	187 Butter	Daytona B	Florida	32128	scaswell0@simplemachines.org
DV-13045	Darrin	Van	Corporate	2 Farmco I	San Berna	California	92405	groumier1@unicef.org
SO-20335	Sean	O'Donnell	Consumer	842 Esker	Fresno	California	93762	dgarrigan2@twitter.com

To:

A	B	C	D	E	F	G	H
CustomerID	First Name	Last Name	Segment	Address	City	State	ZipCode
1	Claire	Gute	Consumer	187 Butter	Daytona B	Florida	32128
2	Darrin	Van	Corporate	2 Farmco I	San Berna	California	92405
3	Sean	O'Donnell	Consumer	842 Esker	Fresno	California	93762

b. Products

A	B	C	D	E	F
RowID	Product ID	Category	Sub-Category	Product Name	Price
1	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Co	130.98
2	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabri	243.98
3	OFF-LA-10000240	Office Sur	Labels	Self-Adhesive Ad	7.31

To:

A	B	C	D	E
ProductID	Category	Sub Category	Product Description	Price
FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collect	130.98
FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Uph	243.98
OFF-LA-10000240	Office Sup	Labels	Self-Adhesive Address	7.31
FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series	191.5155

c. Sales Fact Table

A	B	C	D	E	F	G
Order_Date	Store	Employee_ID	custid	Product_ID	Quantity	Sales
1/3/2017	1	122	294	OFF-AR-10004511	2	$8.56
1/3/2017	1	265	267	OFF-BI-10003196	6	$6.73
1/3/2017	1	186	191	FUR-CH-10002372	2	$113.37
1/3/2017	1	275	44	OFF-BI-10002225	2	$41.28

To:

A	B	C	D	E	F	G
Order_Date	StoreID	Employee ID	Customer ID	Product ID	Quantity_Sold	Sales
1/3/2017	1	122	294	OFF-AR-10004511	2	$8.56
1/3/2017	1	265	267	OFF-BI-10003196	6	$6.73
1/3/2017	1	186	191	FUR-CH-10002372	2	$113.37

d. Sales Staff

A	B	C	D	E
Employee_ID	first name	last name	Address	Soc sec
1	Ysabel	McIlharga	63 Melvin	649-58-300
2	Emelia	Higgonet	24 Burnin	577-61-209
3	Olga	Heinert	76 Mallor	562-36-364

To:

A	B	C	D	E
Employee ID	First Name	Last Name	Address	SSN
1	Ysabel	McIlharga	63 Melvin	649-58-3009
2	Emelia	Higgonet	24 Burning	577-61-2091
3	Olga	Heinert	76 Mallory	562-36-3640
4	Egan	Matteyi	34 Corscot	101-55-8333

e. Stores

A	B	C	D
store id	Add	phone	zip code
1	2 Havey Tr	722-797-92	10455
2	4573 Men	736-952-85	10027
3	718 Burnir	229-332-70	11747
4	05 Di Lore	192-245-95	11432

To:

A	B	C	D
StoreID	Address	Phone Number	Zip Code
1	2 Havey Tr	722-797-9284	10455
2	4573 Men	736-952-8946	10027
3	718 Burnir	229-332-7041	11747

f. Date Dim

A	B	C	D	E	F	G	H	I	J
DATE_KEY	DAY_OF_WEEK	CALENDAR_MONTH	CALENDAR_QUARTER	CALENDAR_YEAR	FISCAL_WEEK	FISCAL_MONTH	FISCAL_MONTH_NAME	FISCAL_QUARTER	FISCAL_YEAR
1/25/2018 0:00	THURSDAY	201801	20181	2018	201804	201801	Jan-18	20181	2018
1/27/2018 0:00	SATURDAY	201801	20181	2018	201804	201801	Jan-18	20181	2018
1/29/2018 0:00	MONDAY	201801	20181	2018	201805	201801	Jan-18	20181	2018
1/31/2018 0:00	WEDNESDAY	201801	20181	2018	201805	201801	Jan-18	20181	2018

LOAD

1. To begin the loading process, you will need to connect to the existing Google Cloud Platform instance
   a. https://console.cloud.google.com/compute/instances
2. From there you must connect to the BASH environment by clicking on the SSH button in our instance.
3. While in BASH, enter the following to enter the MYSQL environment. Enter your password after.

```
w_hall411@hw2-instance-1:~$ mysql -u root -p --host 127.0.0.1
```

4. Setting a new database for project.

```
MySQL [(none)]> create database final_project;
Query OK, 1 row affected (0.015 sec)

MySQL [(none)]> use final_project;
Database changed
```

5. Upload cleaned source files in BASH

```
wdbxny411@final-project:~$ dir
Customers.csv  Date_dim.csv  Products.csv  Sales_Fact.csv  Sales_Staff.csv  Sales_Staff_(1).py  cloud_sql_proxy
Customers.py   Date_dim.py   Products.py   Sales_Fact.py   Sales_Staff.py   Store.csv           venv
```

6. Now we must create a python script to convert the cleaned source files to SQL. I am using the *to_sql* from sqlalchemy.
   a. .py files are located in GITHUB (https://github.com/whall411/CIS4400-Final-Project.git)
7. Run the py script in BASH and return to the MYSQL environment. In MYSQL display your tables and make a

simple query to ensure the .py script was successful

```
MySQL [final_project]> show tables;
+-------------------------+
| Tables_in_final_project |
+-------------------------+
| Customers               |
+-------------------------+
1 row in set (0.002 sec)
```

```
Database changed
MySQL [final_project]> select * from Customers limit 5;
+-------+------------+------------+-----------+-----------+--------------------+---------------+-------------+---------+
| index | CustomerID | First Name | Last Name | Segment   | Address            | City          | State       | ZipCode |
+-------+------------+------------+-----------+-----------+--------------------+---------------+-------------+---------+
|     0 |          1 | Claire     | Gute      | Consumer  | 187 Butternut Court | Daytona Beach | Florida     |   32128 |
|     1 |          2 | Darrin     | Van       | Corporate | 2 Farmco Point     | San Bernardino | California  |   92405 |
|     2 |          3 | Sean       | O'Donnell | Consumer  | 842 Esker Court    | Fresno        | California  |   93762 |
|     3 |          4 | Brosina    | Hoffman   | Consumer  | 9 Vermont Alley    | Pittsburgh    | Pennsylvania |   15210 |
|     4 |          5 | Andrew     | Allen     | Consumer  | 0834 Waxwing Parkway | Detroit     | Michigan    |   48211 |
+-------+------------+------------+-----------+-----------+--------------------+---------------+-------------+---------+
5 rows in set (0.002 sec)
```

8. Repeat with the other source files

```
Database changed
MySQL [final_project]> show tables;
+-------------------------+
| Tables_in_final_project |
+-------------------------+
| Customers               |
| Date_dim                |
| Products                |
| Sales_Fact              |
| Sales_Staff             |
| Store                   |
+-------------------------+
6 rows in set (0.001 sec)
```

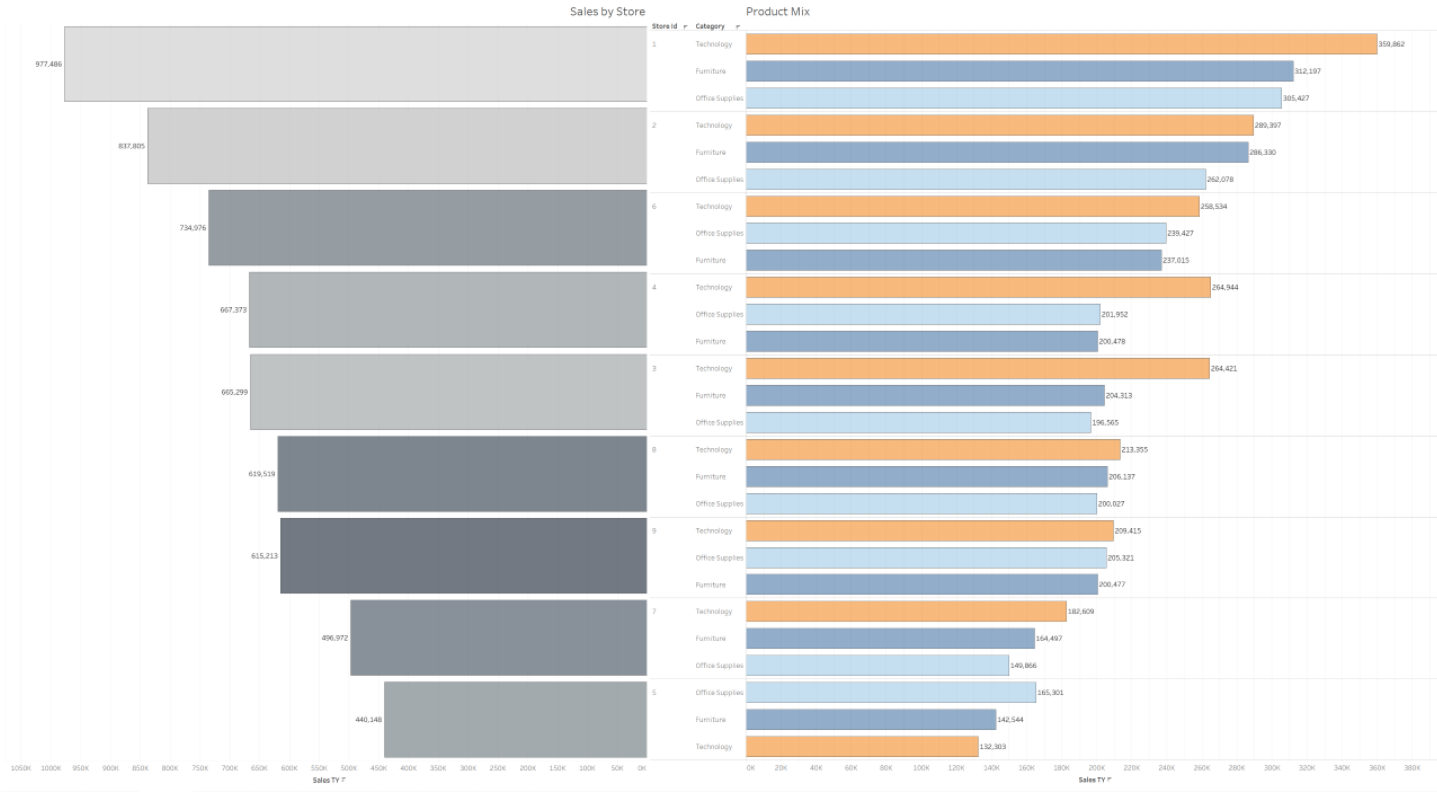
Schema

The schema in this case doesn't apply because the sqlachemy package automates the whole process. Using it with pandas, sqlachemy only requires the column names, the chosen table name, and the database it is going to.

Analysis

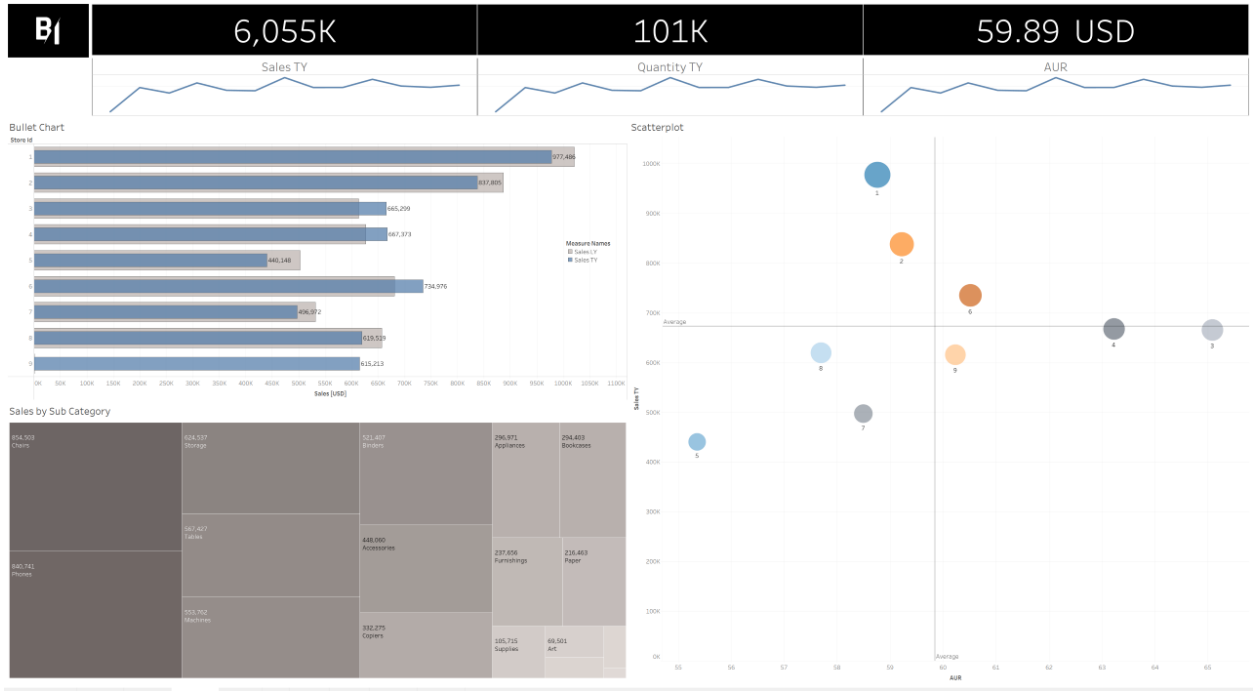
The visualizations presented below tell a different story about the strengths and weaknesses of our company. It is important to provide a visual representation about the performance of our company to quickly make decisions.

Sales by Store and Product Mix



The above visualization provides us information about the performance of each store ranked from highest to lowest. The next visualization, the product mix, is used to show how the popularity of each segment of each store.

Bullet Chart/Scatterplot/Sales by Sub-Category

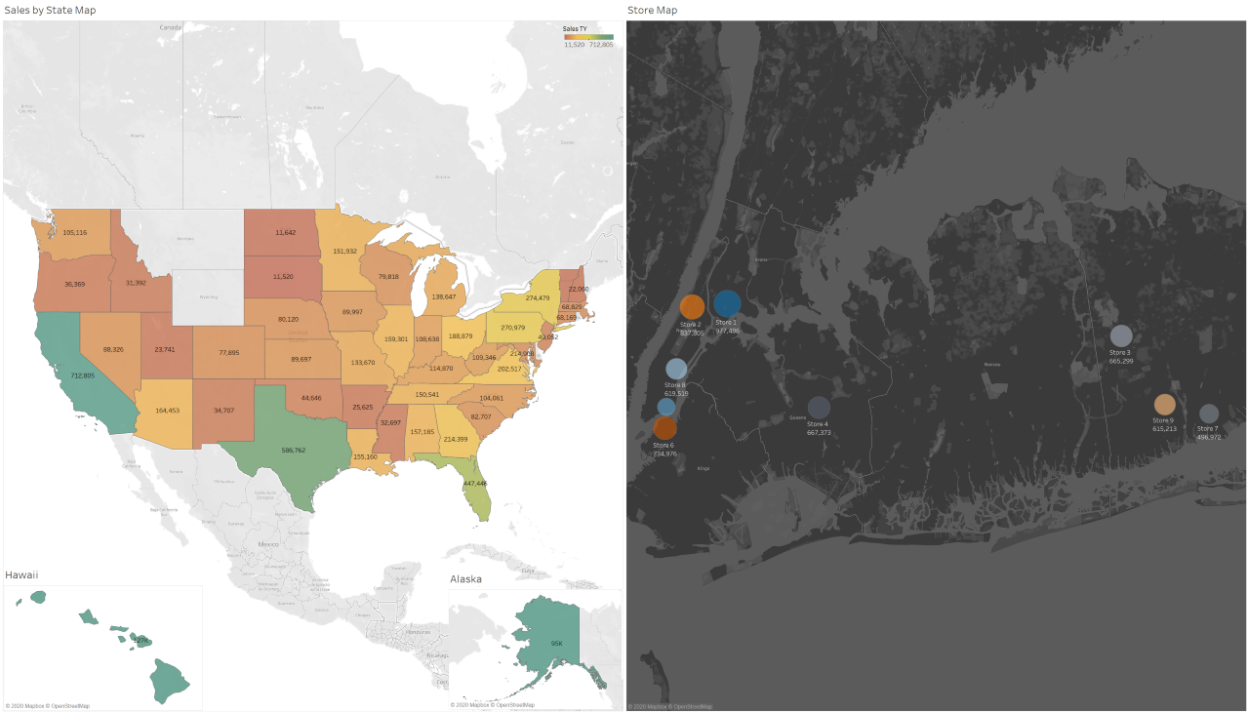


In the bullet chart, the purpose of this shows the performance of sales from the previous year to the current year. The blue bars represent the current the current year and the grey bars represent the previous year.

The Scatterplot displays the sales performance compared to the average of the company.

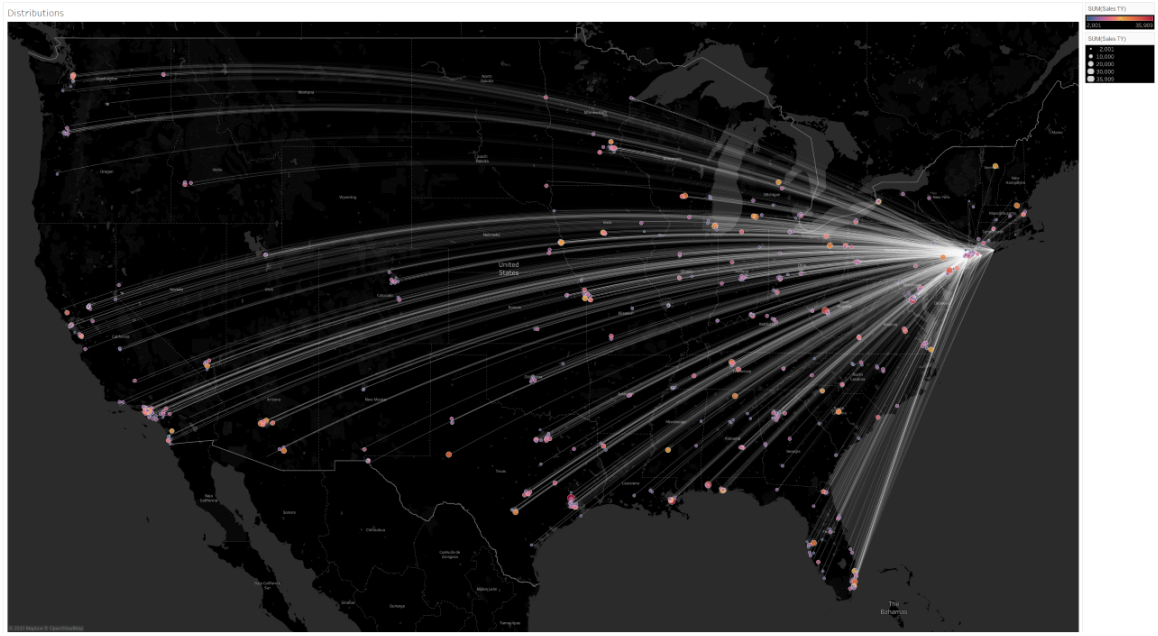
The Sales by Sub-Category represents the most popular type of sub-category in our stores. The bigger the square, the higher performing.

States with the most sales/Store Map



The right map represents the states with the highest and lowest number of purchases. According to the map, it appears that most of our sales originate from California. The store map on the left displays our store locations in NYC and Long Island. The size of the location's circle represents its sales performance.

Origin and Destination of Sales



The final visualization displays the areas that have purchased our products. This allows us to make critical decisions as to where our company should expand to.

Conclusion

In this project, I worked on the ETL process, created the architecture diagram, and generated some risks and solutions associated with the project. Although I would like to say that the ETL process was the most difficult (due to some pesky errors), It is the part I enjoyed the most. I am finally comfortable writing Python scripts and building tables using sqlachemy. I can say that I learned how a basic data warehouse is set up, and very curious to see a larger enterprise sized warehouse one day.

If I could do this project all over again, I would've better prepared myself by knowing more analytic and visualization tools like Tableau to make analysis of the data I imported. Also, I would've been more comfortable with other technologies that perform the same functions.

In the future, I would like to test my ETL skills further using massive amounts of data. I believe if I work on a large-scale data warehouse, I will be able to understand it much better. Generally, this was a positive experience. I was confused in the beginning but getting my hands dirty and creating one showed me a lot.