MIST Database Project

Will Rice

Table of Contents

1. Project Description 3
2. Entity-Relationship Diagram of the relational model 4
3. 10 SQL queries, descriptions, and Results 5
4. 10 MongoDB queries, descriptions, and results 7
5. Data cleansing and preparation 8
6. Issues and challenges with the data 8
7. Visualizations and key findings 9
8. Highlights of graphics and analysis 13
9. Conclusions 13

**Project Description**

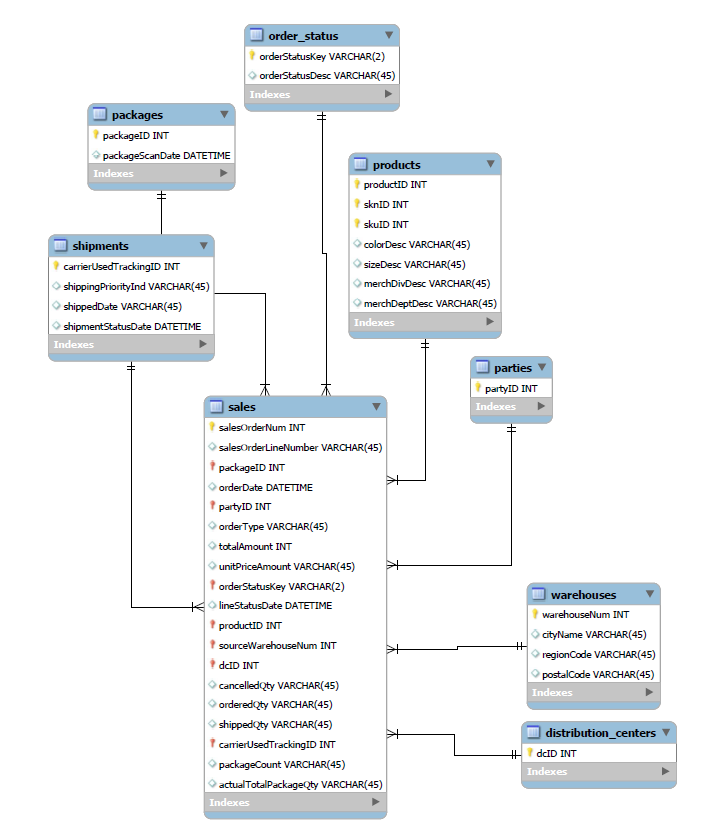
*Does Speed Matter in ECommerce?*

QVC provides a diverse range of products to a primarily U.S. customer base, shipping millions of packages each year. To stand out in the business of online retailers, shipping speed is almost a necessity to create and maintain a loyal customer base. However, QVC does not only pride itself on its expedited shipping, but it also focuses on top-notch customer service that helps create a sense of personable connections with its customers. We were given data on QVC to determine whether delivery speed truly mattered. In what ways, and how this is measured, it was up to us to decide.

We were given one large datasheet involving information on sales, packages, shipments, products, costs, distributions centers, and more. Our first task was to turn this large, rather clunky sheet of data into an entity-relation diagram, and then convert this into an appropriate relational database in the third normal form. We did this using SQL, and found the data was separated into 8 different tables to store all of the data correctly with identifiable connections within the database. We ran a few queries within SQL to get an understanding for our data, and what kind of values we were working with. This was also more to leverage our querying and database skills within SQL, and less about the physical data within it.

After this was done, we converted this data back into csv files to then import into MongoDB, a different database program that will help use various querying methods, as well as visualizations for our analysis. Again, we ran queries on sample data to ensure our use of Mongo’s syntax was correct and ready to be tested and used on real data. After some practice with a few MongoDB queries, both simple and advanced on subdocuments, arrays, and aggregate functions, we imported the real QVC sample data into MongodDB to use ofr analysis. The data cleaning and preparing was by far the most time-consuming part to ensure all the data was correctly input, and various new fields were added so that our data was available to manage the types of visual and statistical analysis we needed for our project. After uploading in our data, we started working on some visuals to help coincide with answering the key questions in the project: does delivery speed affect customer loyalty? Are there inefficiencies in the distribution process that could lead to slower delivery times? After conducting our analysis and creating some visuals, we have finalized our report regarding QVC’s ECommerce business and the different insights we discovered along the way. As you will see later, we can relay in short that the data we sampled did not show that customer loyalty (as defined by us with multiple orders) was not a result of faster delivery times, but delivery time itself could be faster if QVC placed and set up their distribution centers and warehouses in more opportune locations corresponding to the places purchasing the specific products most. This may be costing more money on distribution and losing more customers or churning more with less efficiency.

**Entity Relationship Diagram**



**SQL Queries, Descriptions, and Results**

-- Query 1: Multiple table join

SELECT s.salesOrderNum, p.productID, w.cityName

FROM `lx\_prj\_wcr\_ag`.`sales` s

JOIN `lx\_prj\_wcr\_ag`.`products` p ON s.productID = p.productID

JOIN `lx\_prj\_wcr\_ag`.`warehouses` w ON s.sourceWarehouseNum = w.warehouseNum;

-- Query 2: GROUP BY with COUNT

SELECT sourceWarehouseNum, COUNT(\*) AS TotalSales

FROM `lx\_prj\_wcr\_ag`.`sales`

GROUP BY sourceWarehouseNum;

-- Query 3: GROUP BY with HAVING

SELECT partyID, COUNT(\*) AS TotalOrders

FROM `lx\_prj\_wcr\_ag`.`sales`

GROUP BY partyID

HAVING TotalOrders > 0;

-- Query 4: ORDER BY with LIMIT

SELECT productID, colorDesc

FROM `lx\_prj\_wcr\_ag`.`products`

ORDER BY colorDesc DESC

LIMIT 5;

-- Query 5: REGEXP (Filter products with color starting with 'B')

SELECT productID, colorDesc

FROM `lx\_prj\_wcr\_ag`.`products`

WHERE colorDesc REGEXP '^B';

-- Query 6: DISTINCT (Retrieve unique warehouse regions)

SELECT DISTINCT regionCode

FROM `lx\_prj\_wcr\_ag`.`warehouses`;

-- Query 7: IN (Retrieve sales with specific product IDs)

SELECT salesOrderNum, productID

FROM `lx\_prj\_wcr\_ag`.`sales`

WHERE productID IN (1, 2, 3);

-- Query 8: AVG function (Calculate average total amount per order)

SELECT salesOrderNum, AVG(totalAmount) AS AvgTotalAmount

FROM `lx\_prj\_wcr\_ag`.`sales`

GROUP BY salesOrderNum;

-- Query 9: Subquery (Retrieve products with the highest total sales)

SELECT productID, colorDesc

FROM `lx\_prj\_wcr\_ag`.`products`

WHERE productID IN (

SELECT productID

FROM `lx\_prj\_wcr\_ag`.`sales`

GROUP BY productID

ORDER BY SUM(totalAmount) DESC

)LIMIT 1;

-- Query 10: IN (Retrieve parties with sales orders)

SELECT partyID

FROM `lx\_prj\_wcr\_ag`.`parties`

WHERE partyID IN (

SELECT DISTINCT partyID

FROM `lx\_prj\_wcr\_ag`.`sales`

);

**MongoDB Queries and Descriptions**

**Queries:**

*Querying Arrays:*

Query 1: db.parties.find({ "partyEmail.name": "John Doe" });

This query finds and displays all the documents in the parties collection that have a name in the partyEmail array as ‘John Doe.’

Query 2: db.parties.find({ "partyEmail.email": { $regex: /gmail\.com$/ } });

This query finds and displays all the documents in the parties collection that have an email address that contains the phrase ‘gmail.com’ (or, for 100% of these collections, is a Gmail address specifically).

*Querying Subdocuments:*

Query 3: db.shipments.find({ "address.state": "State 1" });

This query finds and displays all shipments that are shipping to ‘State 1’ within the state subdocument of the address field in the shipments collection.

Query 4: db.shipments.find({ "address.zipcode": { $regex: /^1/ } })

This query finds and displays all shipments that are shipping to a Zip Code that starts with the number 1 within the Zip Code subdocument of the address field in the shipments collection.

*Aggregate Queries:*

Query 5: db.order\_status.aggregate([  
 { $match: { orderStatusDesc: "Processing" } },  
 { $count: "documentCount" }  
]);

Description: This query matches all the documents in the order\_status collection that have an orderStatusDesc of ‘processing’ and counts the total documents that meet this criterion.

Query 6: db.distribution\_centers.aggregate([

{ $group: { \_id: null, avgSize: { $avg: "$size.acres" } } }

]);

Description: This query takes the average amount of the values inside the acres array in the size field of the distribution\_centers collection and displays it.

Query 7: db.sales.aggregate([

{ $group: { \_id: "$orderType", avgShippedQty: { $avg: "$shippedQty" } } },

{ $sort: { \_id: 1 } }

]);

Description: This groups the sales collection by order\_type, and then proceeds to find the average shipping quantity for all groups, and sort and display them in ascending order

Query 8: db.packages.aggregate([

{ $sort: { packageScanDate: -1 } },

{ $limit: 5 }

]);

Description: This query lists all the package collection’s documents in descending order of packagescandate.

*Simple Queries:*

Query 9: db.products.distinct("merchDeptDesc");

Description: This query identifies all the distinct values in the ‘merchDeptDesc’ in the products collection and returns general store department items.

Query 10: db.sales.find({ shippedQty: { $gt: 30 } });

Description: This query goes through the sales collection and returns all documents who had a ‘shippedQty’ greater than 30.

**Data Cleansing and Preparation**

To conduct the analysis using tables and graphs in the following section, significant data cleansing had to be done. First, the data used in the queries above were sample data that was too small a size to use in analysis and not relevant to answering some of the questions asked in the project. Thus, this sample data was removed from the existing database of collections, and new data was imported via csv files. Using Excel, we split the small sample of QVC data into our 8 collections prior to importing and took a random 1,000 cell sample of this data and created the 8 different collections with documents as seen currently in our MongoDB database. This is more than enough of a sample of data to run some analysis and create meaningful graphics via MongoDB. However, doing so was not as easy as the click of a button! Some of the different collections were not as easy as pasting the data into a different file separately to export straight from the main sheet on Excel; they needed to be prepared and have different information created to have the fully functioning collection. For instance, for the warehouse collection, we needed to use the ‘VLOOKUP’ function on Excel to match records from the QVC Warehouse file to the sample size of 1,000 records. Some ‘if’ statements were needed to generate the full list of values for all 1,000 orders for the line status code as seen in the order status sheet. Lastly, we needed to create a couple new variables within the shipments table before converting to the Mongo collection that differenced out the time it took from the order to be placed to it being delivered, as well as from when it was shipped to delivered. These variables were crucial to have prepared for Mongo as the main analysis concerns the relationship between customer loyalty and delivery speed. Without these variables created, we would not be able to do the analysis in full. After this, we input our data of 1,000 documents into the 8 collections and were ready to start our analysis.

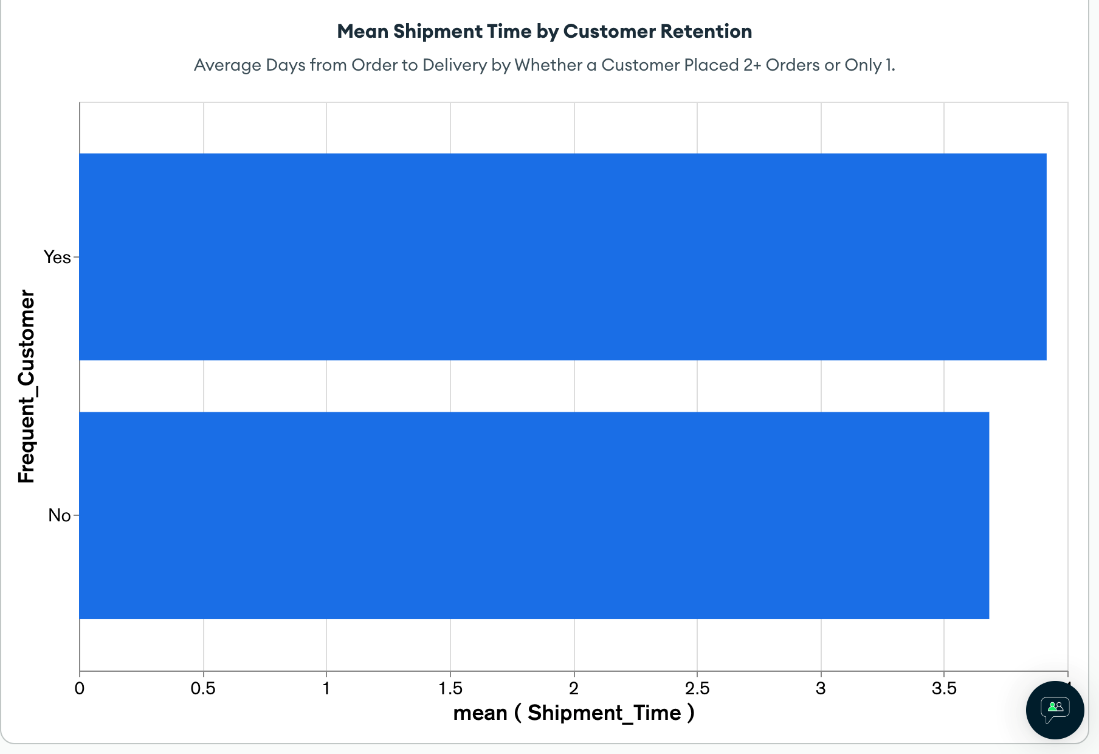
**Issues/Challenges with the Data**

Having the relationships already stored within MongoDB, and knowing what data needed to be placed into which collection helped speed up the timely data preparing process. After doing the whole process for just 25 documents in each and realizing this was not enough to do a full analysis to find trends and patterns, restarting to do 1,000 completely different values was a challenge. There were some issues with the data that included entire rows missing from the data that would have been useful to have, including the destination cities, and priority shipping index. On top of this, these data points were not possible to create manually in the time given, so some desired info was omitted. Overall, the data cleansing process involved less complicated work than it was time-consuming, this was the timeliest aspect of the analysis section of the project, and one that may get easier with more of a feel for the full process after more projects and more experience with this type of data analysis.

**Visualizations and Key Findings**

Overall, we found that in this sample data, customer retention was not necessarily a product of delivery speed. While this seems like an appropriate conclusion to jump to: places where products ordered quickest will bring back satisfied customers. However, in graphic 1, we see the very simple finding that customers who came back for at least 2 orders had longer average days between order date and delivery date than one-time customers. There is some mirroring evidence shown as well in graphic 2, that shows average unit price and total line amount spent by returning and one-time customers. In this, we see one-time customers bought more costly items and paid more than returning customers. While this could suggest that one-time customers bought all their things at once compared to returning customers buying items in spurts, it is more likely one-time bought more expensive items with even quicker checkout delivery times, and still did not return.

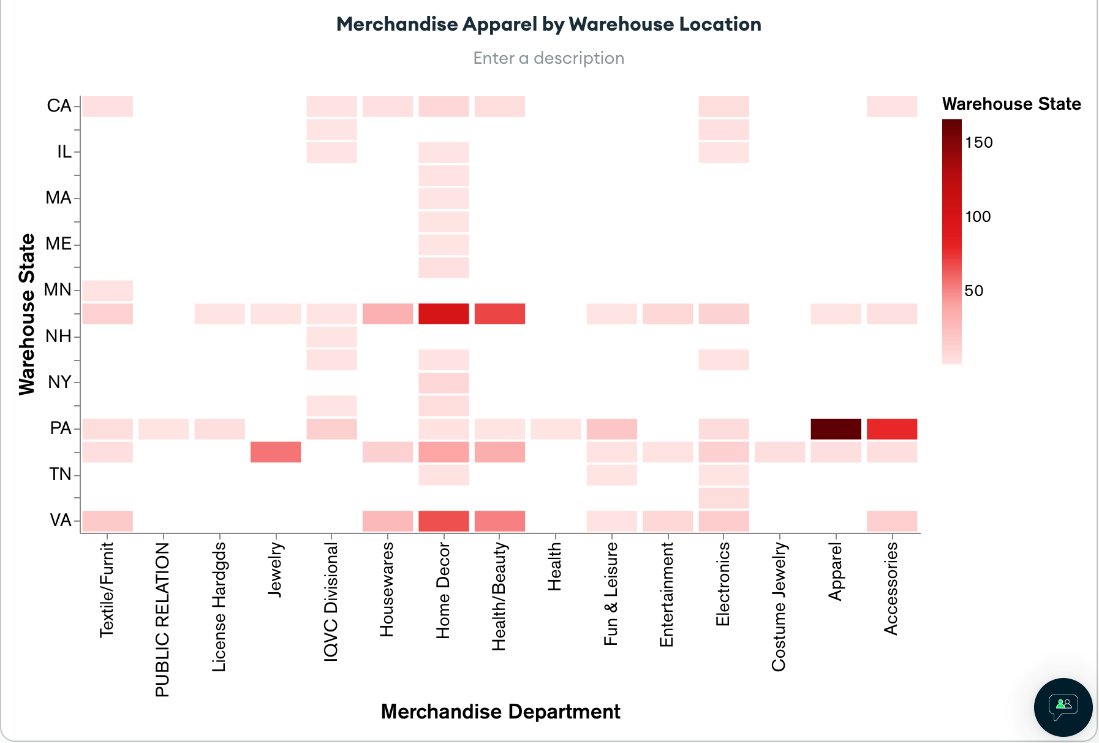
This insinuates there may be different issues between keeping customers than the delivery speed. We thought it may be interesting to see if the reason the store is not returning so many customers, or if they are not as efficient or as profitable, is due to the location of orders versus distribution in their warehouse facilities. Figures 3 and 4 are both heatmaps trying to find abnormalities in these qualities. In figure 3, we see where different merchandise departments are kept in the various locations of the warehouses. We see mostly zeros representing some warehouses are solely focused on specific items, but we do see some large quantities, specifically apparel and accessories coming from warehouses in Pennsylvania, as well as Home Décor and Health/Beauty both being prevalent in Virginia and North Carolina. Next, we found it would be interesting to see where these items were bought. We see apparel is bought heavily in Pennsylvania, but nothing comes close to how much home décor is bought by Pennsylvania, even though it does not contain it in the warehouses. The same goes for NY, FL and CA, that have little to no warehouse or distribution and yet are some of the largest states and have some of largest customer bases by state for QVC. This represents and efficiency issue, and QVC should investigate how its distribution process is run: Warehouses and distribution centers may want to be placed both more strategically nationally to cover the largest customer bases, as well as warehousing and distributing the products more tied to the customer base that purchases them. Both of these strategies would significantly lower delivery time and cost to get from start to finish and is one of the reasons QVC is currently held back. Lastly, to confirm our suspicions, we looked at total sales of items by merchandise division in figure 5. In this graphic, we found that home decor, health and beauty, and apparel are all the top-sellers, yet are made, stored, and distributed in inopportune locations for these customer baes, and these total sales would be much higher given they were placed in different or new warehouses.



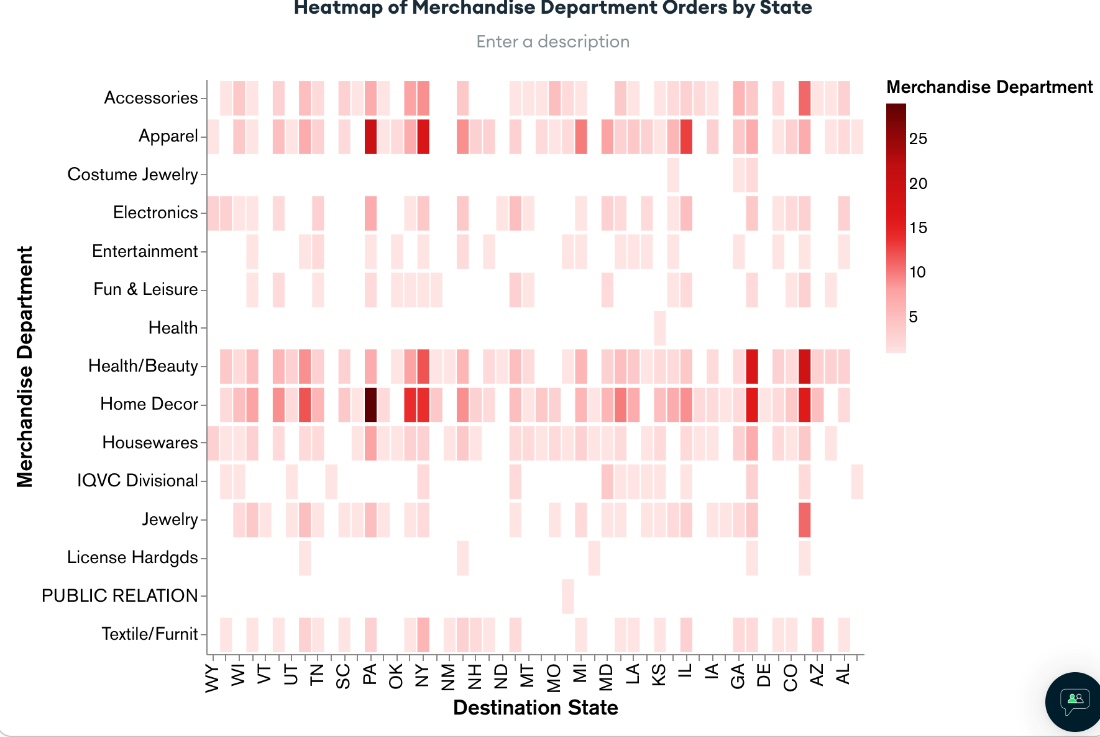
**Figure 1:** Mean Shipment Time in relation to customer loyalty, as mentioned in the main findings section



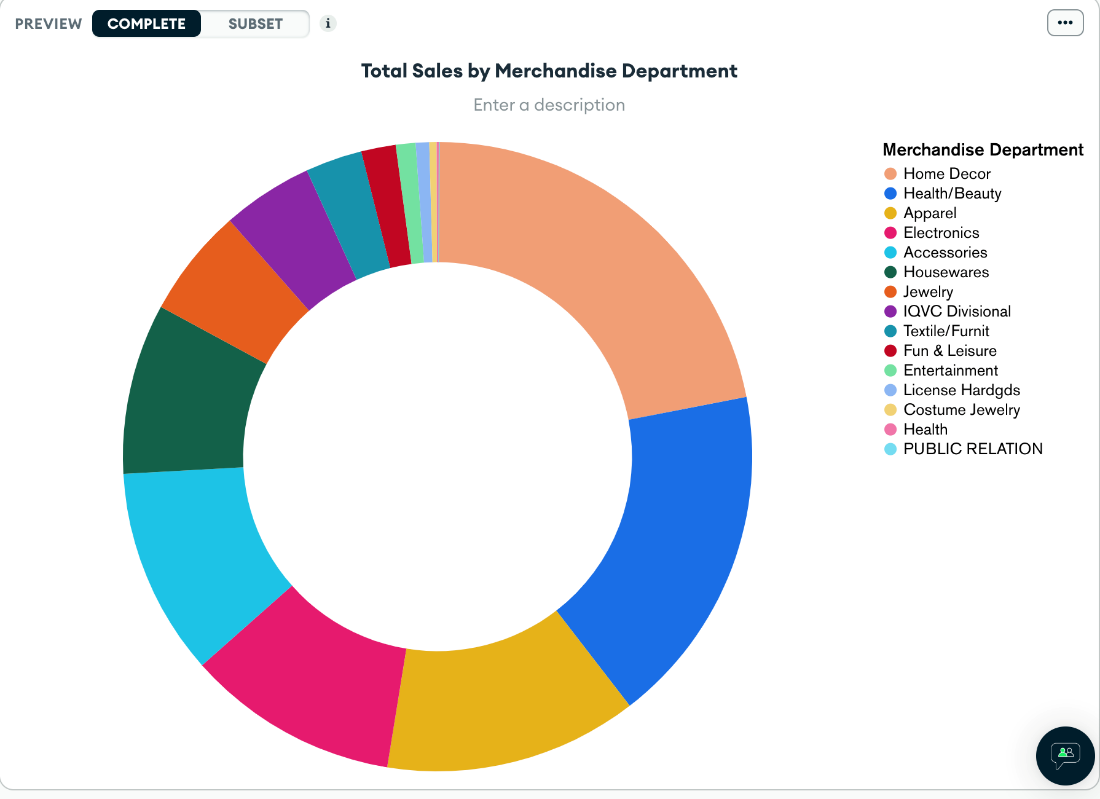
**Figure 2:** Average Unit and Line Price in relation to customer loyalty, as mentioned in the main findings section



**Figure 3:** Heatmap of merchandise distribution by warehouse location, noting apparel and home décor found in specific states.



**Figure 4:** Merchandise distribution to destination states, which shows stark differences in where apparel and home décor are made, stored, and shipped from compared to where they end.



**Figure 5:** Pie chart of total sales of each merchandise department. We see home décor, health and beauty products, and apparel are the biggest sellers, yet inefficiently sold. This could be an issue that may also be a problem with QVC’s customer retention.

**Highlights of Graphics/Analysis**

I found that the complementary use of figures 1&2 and figures 3&4 helped show more of our analysis together than had they been displayed singularly, which is a complement of the ability to make graphics that truly tell a story. The graphics were more simple than potential multi-series or combo graphs, or even geospatial location maps, but still manages to guide viewers and readers along the main points of the analysis conducted, highlighting inefficiencies and lack of customer loyalty. Our analysis in general was pointed towards the last question mentioning how delivery speed affects customer loyalty, and we branched to different questions coming from the original point, meaning that our analysis was based on an obvious train of thought and process as compared to going from section to section picking different variables to look at together. Overall, our analysis is one to just get to the main point and allow the visual to complement and illustrate the findings, as opposed to taking over as the main point, and I believe while at times it looks less appealing than sample graphics shown previously, they are simple, have little clutter, and highlight what needed to be shown.

**Conclusions**

To conclude, our project focused on using database programs such as SQL and MongoDB (as well as an intermediary data storage program in Excel) to create formal, third normal form databases and models that can be used to display trends, find patterns, and conduct analysis. Using sample QVC data, and after following data preparing processes to ensure correctness, we did an analysis of the relationship between customer loyalty and delivery times, as well as if products were moved and shipped efficiently. We found customer loyalty does not result from shipment times or form purchasing power, and that the main merchandising divisons are made, stored and distributed in locations far from the main buying hubs. This means QVC is running an inefficient distribution network and may look to strategize their warehousing and manufacturing locations in the future.

Overall, this was a great project to use real data and deal with the difficulties of messy, unclean, unrelated data, and turn it into a full report. Using SQL/NOSQL databases, as well as using the visualization tools of MongoDB helped us put our knowledge from this class to the test, and we believed we showcased our strengths in our report.