Project Report

Doug Laney

Aakash Kedia
Amal Sharma
Phani Sai Kamal
Shivam Jakhanwal

Percentage of Effort Contributed by Aakash Kedia: 25

Percentage of Effort Contributed by Amal Sharma: 25

Percentage of Effort Contributed by Phani Sai Kamal: 25

Percentage of Effort Contributed by Shivam Jakhanwal: 25

Signature of Student 1: Aakash Kedia

Signature of Student 2: Amal Sharma

Signature of Student 3: Phani Sai Kamal

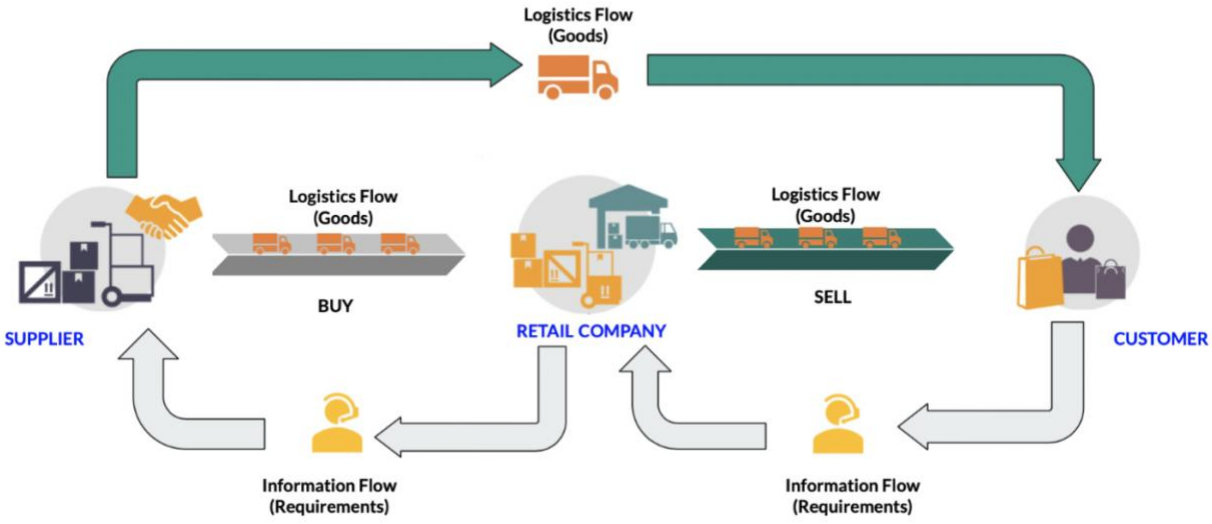
Signature of Student 4: Shivam Jakhanwal

Submission Date: 25th April, 2020

Table of Contents

Problem Setting:

In today's fast paced world, everything needs to in and within our reach at our earliest convenience. This can range from having medicines delivered to our doorsteps to having large sofas shipped halfway across the world. The problem of optimizing our logistics combined with inventory and supply side management is one from the ages. It is an ever-growing and ever-long problem that companies like UPS, FedEx and Amazon are still solving. This logistics problem roots from the travelling salesman problem, which, still hasn't been solved in the computer science world and has been defined an NP-Hard problem. Hence, most of these logistics and shipping companies still struggle with finding the best solution to saving transportation cost. In our problem setting, we have to find a way to optimize shipping costs for online orders by managing depot allocation for stocking items within regions.



a) Goods flow process of Retailer

Optimizing the Depot Allocation in order to reduce the shipment costs is a very crucial task for any E-commerce company. It is because most of the companies provide their customers with free shipping facility when the total order value reaches a certain threshold. So, the shipping costs takes up a big bite in the profits of the company. Generally, the shipping cost is dependent on Distance (i.e., the distance between the Depot location and Order Destination), Weight and Dimensions of the package. It is considered that the greater the Distance, the higher the Shipment Costs and the higher the Delivery Time. To deal with all these problems' companies tend to open multiple Depots

spread all across the nation for easier, faster and cheaper deliveries. Setting up Depots at ideal locations is just one side of the problem but the real problem arises in managing the Depot.
Depot Allocation deals with stocking of a fixed number of units of a particular item at a particular depot. In Depot Allocation, factors like Desired Consumer Base metrics, Workforce Availability, Proximity to the Customer and Logistic Services plays an important role to calculate the fixed number required and understand the workflow dynamics of a customer order. Not only the shipping costs and delivery times affect the profits of a business, but they are also the reasons because of which the online customers abandon their carts at the checkout. All these are just some of the features of Logistics that affect the businesses, there also things like Tracking facility, Priority Shipping, Insurance, Handling costs which all add up and increases the total shipping cost. The perfect allocation of an item would be mutually beneficial for the company as well as the customer, overall provides a more efficient, effective and profitable experience. In general, it is considered that the closer the Depot Location to the Customer the better the profits because by cutting down the distance an order has to travel, it can be delivered quicker and at low shipping cost, which in turn boosts the profits earned by the business. Having the right strategy in place would help in reducing the shipping cost and ultimately growing the business.

Problem Definition:

The specific problem that we want to solve is to understand customer buying patterns, forecast demand and allocate products to the right depot at the right time. This will essentially be done in an environment where we know the demand of a specific good and can determine its demand-supply fundamentals. Moreover, we will be optimizing this problem by predicating demand for a certain good based on trends and previous historic data that we have related to the good. For example: We notice a trend of Kansas City Chiefs Banners that are made in Kansas being ordered excessively in Los Angeles before the Super Bowl. Now we need manage and act accordingly to allocate our depot shipment so demand and supply for those banners is met in Los Angeles whilst the shipping cost is at it's possible minimum. How will we tackle this problem?

Optimization of Depot Inventory Allocation to drive down Shipment Costs for Online Orders based on the parameters like Frequency of Orders for a Particular item, Customer Behavior, Distance, Region or Zone, and Product Category we want to estimate the probable count of units

of a particular item to be allocated for a particular regional depot to achieve low Shipment Costs will be used as predictors. Figuring out which Depot Allocation technique to be used, or Estimating the Shipping Rates is not a simple task, as there will many changing variables along the way. These questions will help us keep better on track on how to optimize the Depot Inventory Allocation.

- Which depot an item should be allocated to optimize the shipping cost?
- How many units of a particular item should be allocated to a particular depot?
- What is the most efficient method of shipping to drive down shipping cost?
- Which depot the order item should be dispatched from?

Data Sources:

We used Retailer historical data for exploratory analysis and for predicting the outcomes. Detail of the data of provided below:

Table 1

Variable Name	Variable Description	Sample Values
ORDER_DATE	Date when Product is purchased	1/1/18
ITEM_NUM	Product serial number for tracking and order purpose in the warehouse	1011493
ORDER_QTY	Quantity of Product sold by retailer	1
SELL_PRICE	Price at which product is sold by retailer	14.99
COUPON_REDEEMED_AMT	If Promotion is provided by retailer then it will reflect here	2
SHIP_TO_STATE	Buyer State where product need to be delivered	PA
SHIP_TO_ZIP	Buyer ZIP where product need to be delivered	17522
DEPT_NUM	Department number under which product falls	20
DEPT_DESC	Department description	HEALTH & BEAUTY AIDS
ITEM_CATEGORY	Category under which product falls	DCB
ITEM_CATEGORY_DESC	Item Category Description	TOOTHBRUSH MANUAL
ITEM_DESC	Item Description	ORAL-B 3DW PULSAR TB

ITEM_COST	Item Cost Price	13.19
DEPOT_NUM	Depot/ Fulfilment center number	759
FULFILLMENT_TYPE	Depot/ Fulfilment center Type	LTL
DEPOT_NAME	Depot/ Fulfilment center Name	FlemingtonÂ
REGION	Group of states under same Region	NW
SHIP_FROM_ZIP	Depot ZIP Location	75236

Data Exploration:

We used Tableau for initial data exploration to find critical predictors responsible for evaluation and forecast of order for particular depots.

Graphs mentioned below in Figure 1 show the Top 10 revenue generated and ordered items. During this exploration we found out that All Wood Cabinet is the highest revenue generating product with $14 M followed by Safe Racks Combo 24" whereas Kirkland Signature 3PCS is the highest ordered product with over 500k plus orders in 18-19 followed by All Wood Cabinet.

Figure 1:

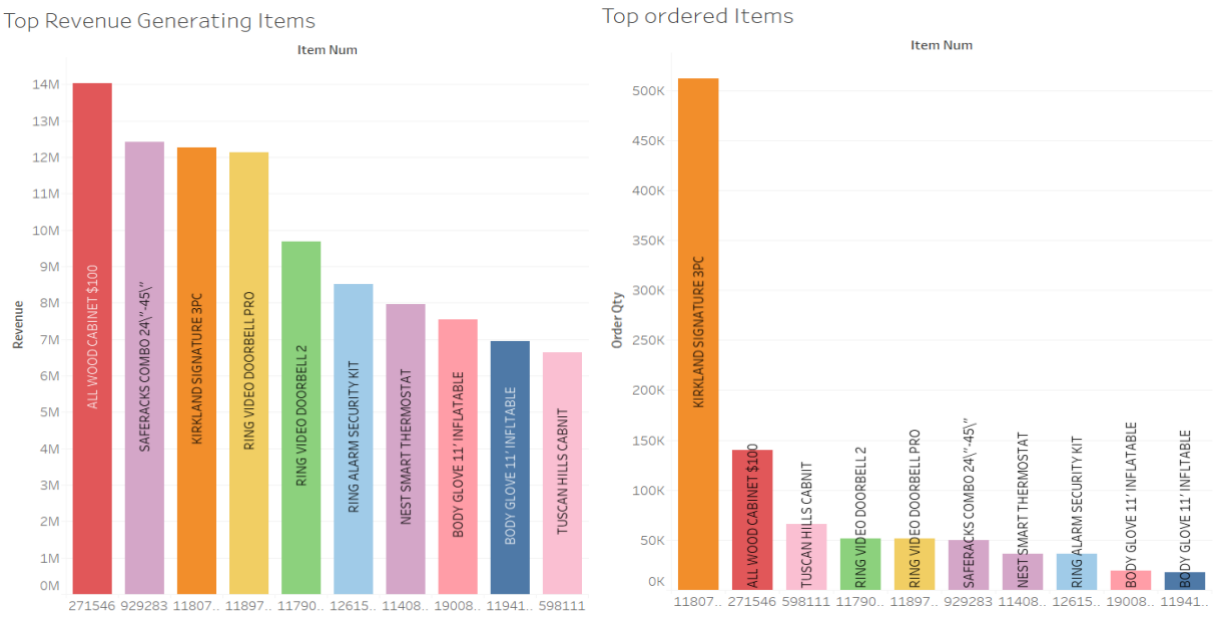
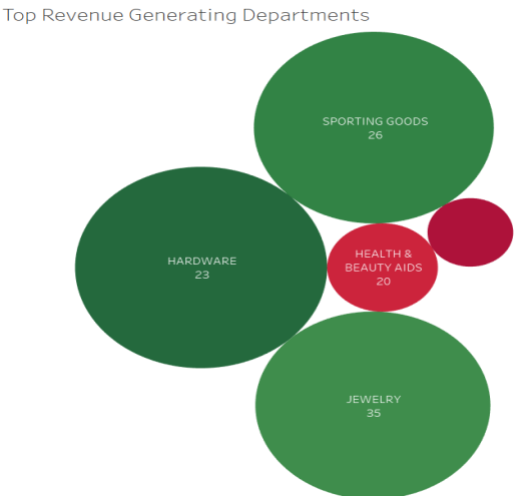


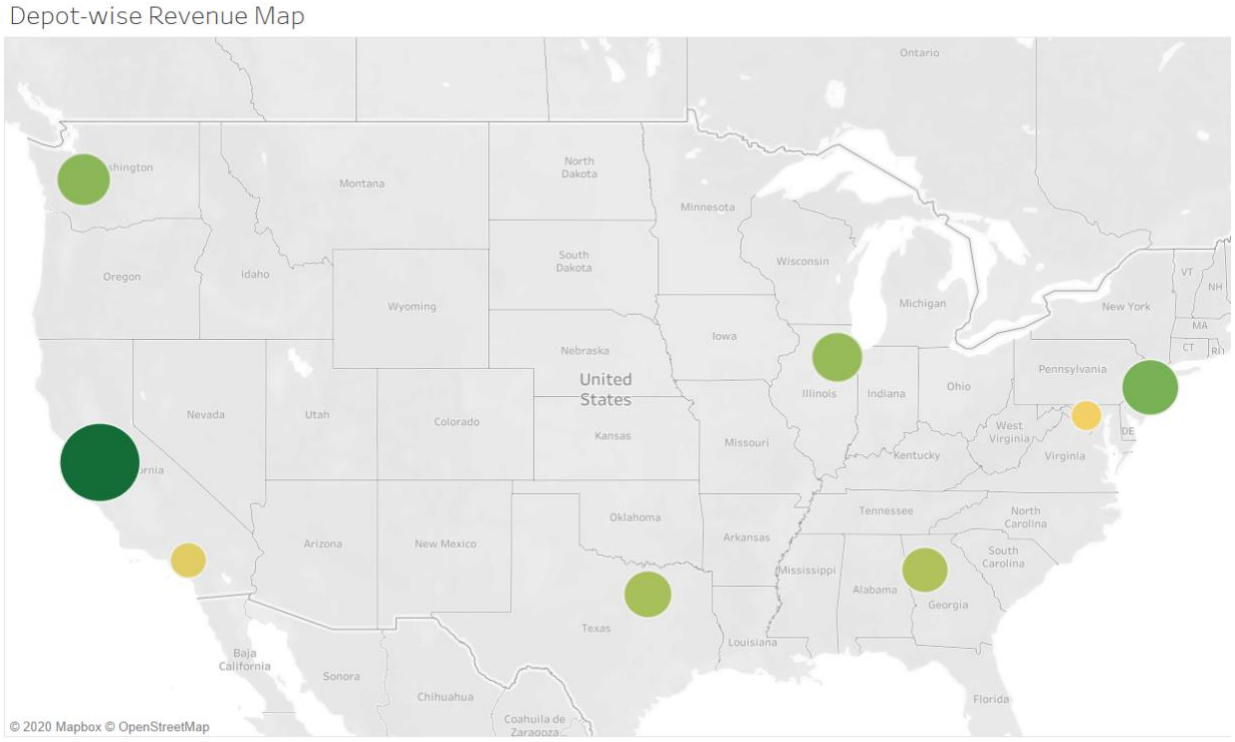
Figure 2:



By digging further deep, we investigated the departments generating highest revenue as seen in Figure 2 above. Hardware is the one contributing to almost 32% of the total revenue making it highest in top revenue generating department category followed by Jewelry and Sporting Goods that do have a significant impact as well.

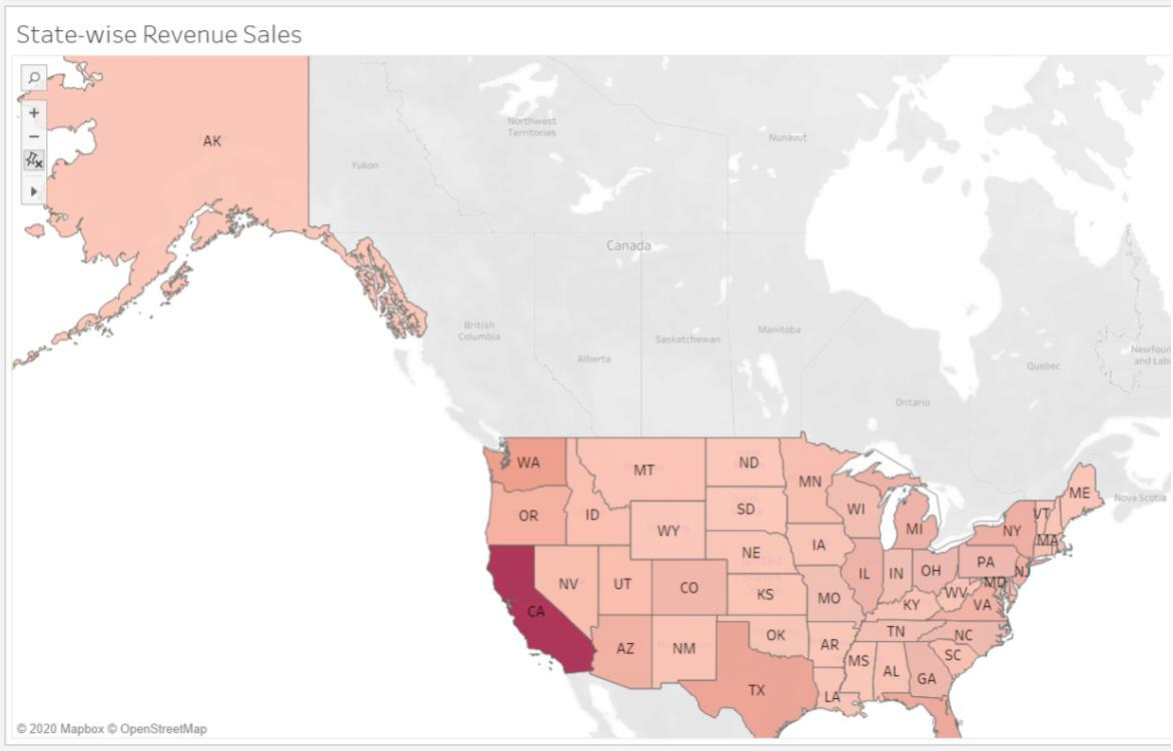
We also tried to map revenue based on depot as seen in Figure 3 below. From there we notice that the largest Circle in the Map is in California that indicates depot no 725 – Stockton is the highest revenue generating Depot all over the USA.

Figure 3:



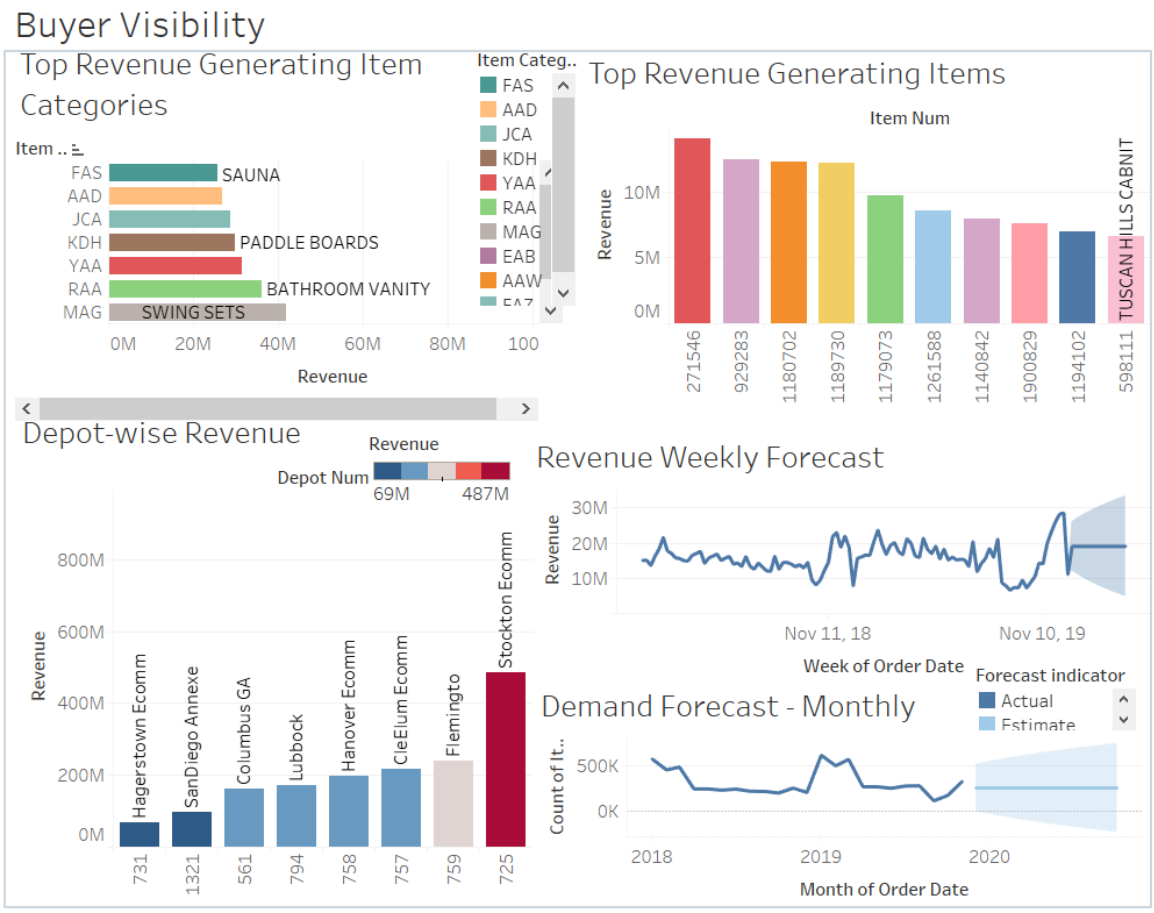
The map below, Figure 4, shows the State Wise Revenue Sales. It is clearly visible that customers in California are ordering the most items making it the highest revenue generating state in America. Other than that, we can also correlate that the depot with highest revenue is also located in same state from where the retailer gets maximum orders.

Figure 4:



Lastly, we forecasted revenue and demand for every depot and each item over the period of the next 90 days. Graphs in this interactive dashboard shown below in Figure 5 are linked to each other so that depot wise data can be segregated and categorized for better visualization.

Figure 5:
Dashboard of Depot Wise Revenue with Weekly Revenue Forecast and Monthly Demand Forecast:



Data Mining Tasks:

Data Preparation:

With about 3.5 GB data, our team spent around 70% of time in data preparation and wrangling. We were provided with raw data from one of the biggest retail giants in the world and that had to be cleaned, transformed and formatted. Below are the steps we had performed:

1. Clubbed the data from all the department for year 2019 and 2018 into single data frame.
2. Added Ship from Zip and appended against each order which was derived from GEO LOCATION data provided to us for each depot.
3. Added the revenue column which was derived from the sell price and total quantity of item sold for the particular day.
4. There were few entries in our data which had missing values like ORDER_DATE, SELL_PRICE, SHIP_TO_STATE. Handled entries like such so as to remove discrepancy in calculation.

Data Transformation:

1. Transformed Categorical Variables to Factors and Encoded Numerical Dummies for modelling.
2. Pivoted the ITEM_NUM against ORDER_DATE to plot quantity which was further used in time series analysis.
3. Grouped the data on item_num and date which was further summarized and spread to predict the demand.

Data Mining Models/Methods:

Association rules:

Association Rules are one of the important concepts used in data mining for finding strong relationship between antecedent and consequence. In a fulfilment center, all vegetables are put in a similar walkway, all dairy things are set together and cosmetics positioning shuffling from time to time. Items are intentionally positioned like this so that it lessens the client's shopping time, yet additionally helps the client to remember what significant things they may be keen on purchasing,

consequently helping stores strategically pitch other items of similar categories simultaneously. Affiliation rules help reveal every single such connection between things from large databases and are tremendously helpful in our case.

To further elaborate, we grouped items purchased on the same date and delivered in same ZIP as one unique transaction since this data was not provided. List of items with unique transaction IDs are studied as one group which helped in understanding the placement of products on aisles. Once unique transaction was retrieved, the next stage is to produce rules from the whole rundown of things and distinguish the most significant ones. We applied the Apriori algorithm to find products with strong relation.

Once we got the Antecedent and Consequence for every department, we calculated Support, Confidence and Lift which are Market Basket Analysis concepts to further validate the string association. High value of confidence suggests a strong association rule. From Table 2 below, we can see Support, Confidence and Lift for various products in the departments within our scope.

Table 2:

Dept. Num	Antecedent	Consequent	Support	Confidence	Lift
20	PREMIER VANILLA RBST FREE	PREMIER CHOCO RBST FREE	0.01978236	0.3817879	3.859528
20	KS SUPREME DIAPERS SZ 6	KS BABY WIPES 900CT	0.01385038	0.3277401	1.482337
20	KS SUPREME DIAPERS SZ 5	KS BABY WIPES 900CT	0.02100057	0.3217494	1.455241
20	KS SUPREME DIAPERS SZ 4	KS BABY WIPES 900CT	0.02325063	0.3015031	1.363669
23	DURACELL \"AAA\"-32PK	DURACELL \"AA\"-40PK	0.02218780	0.4756939	5.841298
23	RING ALARM CONTACT SENSOR	RING ALARM SECURITY KIT	0.01181229	0.3605528	4.571723
26	KS GOLF GLOVE LARGE 4PK	KIRKLAND SIGNATURE 3PC	0.01252263	0.3078486	0.8009243
28	AMAX POWER CRAZE RC GREEN	AMAX POWER CRAZE RC BLUE	0.02175957	0.5675720	12.364159
28	AMAX POWER CRAZE RC BLUE	AMAX POWER CRAZE RC GREEN	0.02175957	0.4740170	12.364159

28	AMAX HOVER STAR MATTE	AMAX HOVER STAR MATTE RED	0.01918477	0.4110886	9.523492
28	AMAX HOVER STAR MATTE RED	AMAX HOVER STAR MATTE	0.01918477	0.4444444	9.523492

Disclaimer: The following assumption have been made to uncover the transactions, **on a Particular Date, An Order from specific Zip Code can only be from One Customer/Transaction**
Note: Department Number 30 (Jewelry) has no sustainable association rules

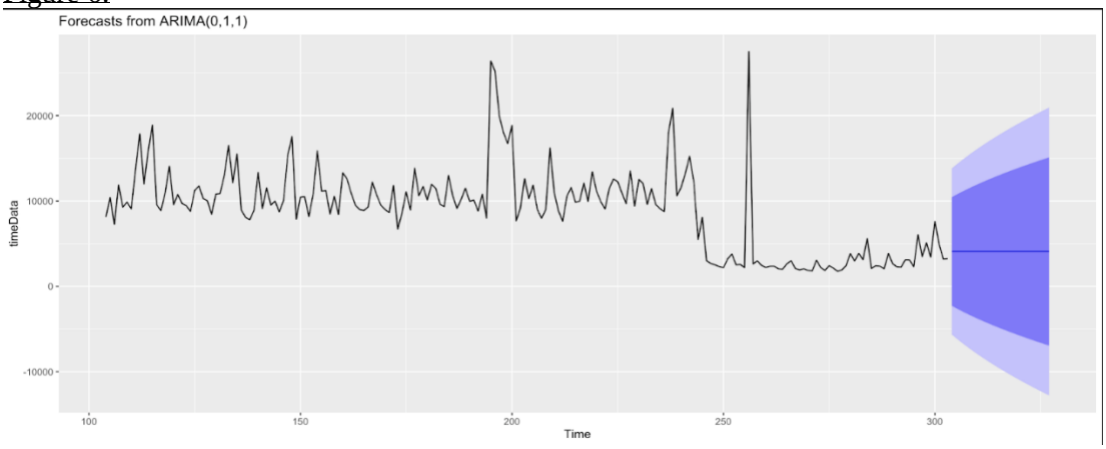
Demand Forecast:

Predictive analytics is very actively used by businesses for being prepared for unpredicted increase and decrease in demand. For the sake of time, businesses are always keen to spend money to predict the demand in advance and getting prepared for consumer consumption.

We have used time series data provided in our dataset and trained it on the Auto Regressive Integrated Moving Average (ARIMA) model to forecast next 30-days's demand as seen below in Figure 6, by analyzing the past sales. We were able to achieve MAPE of 20% which tells us the % prediction differing from actual value. MSRE was calculated as 4963.

Variables: item number, order date and order quantity

Figure 6:



Revenue Forecast:

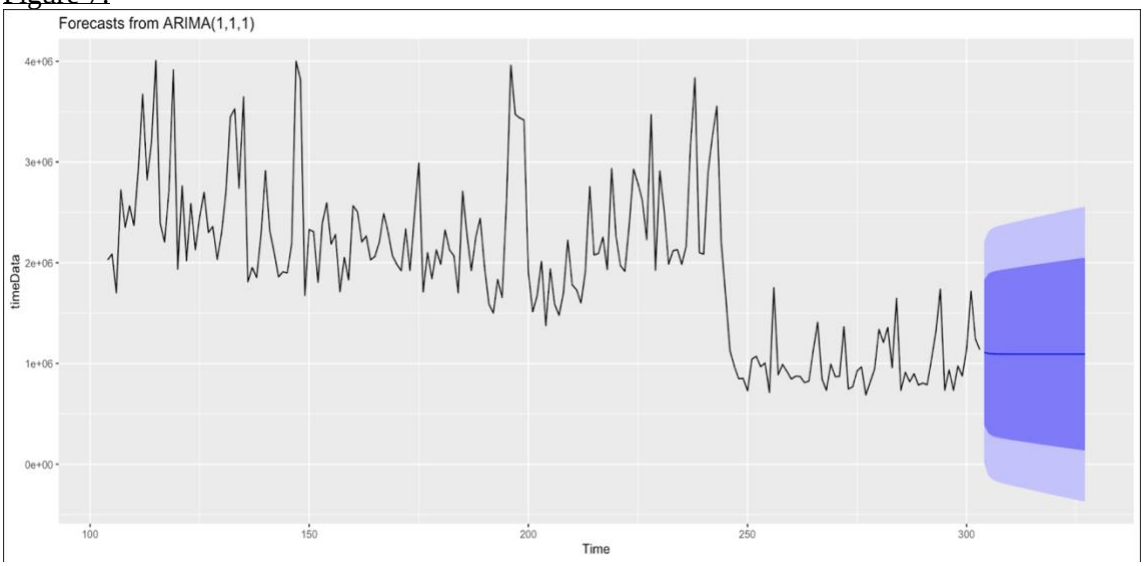
In the end it all comes down to the revenue. As revenue is something which companies put head up to, they are always keen to know the insights when there is a sudden drop or peak achieved.

Companies are always on the lookout for such events, either to replicate or avoid such scenarios as they incur huge costs or huge revenue boost both in their favor or otherwise.

For our project we feed the revenue data on daily basis to the ARIMA model and forecasted the revenue for the next 30 days as seen below in Figure 7. We were able to achieve MAPE of 10% which tells the % prediction differing from actual value. MSRE was calculated as 5570.

Variables used are item number, order date, Selling Price, Order Qty used to derive the Revenue

Figure 7:



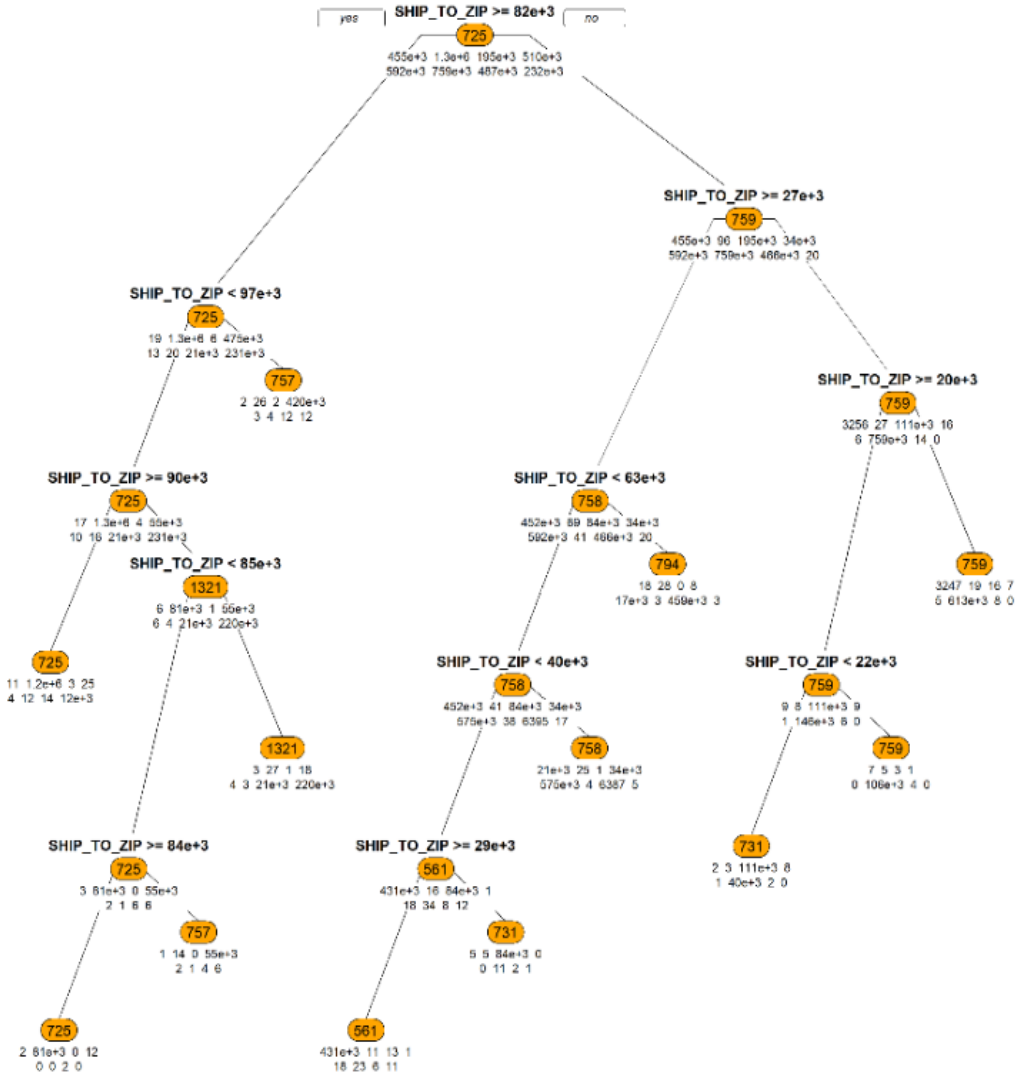
Decision Trees:

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. It is drawn upside down with its root at the top. Here features are split multiple times based on certain cut off values creating different subsets. We have mapped the path for 8 depots against features like Ship to ZIP and Item numbers. Without pruning and with only 3 features we obtained 96% of accuracy on the training data set with maximum depth up to 7 nodes. To predict the outcome in each leaf node, the average outcome of the training data in that particular node is used

It shows that when ship_to_zip is greater than or equal to 8* it will follow the certain node otherwise it will go to the different node. This is visualized below in Figure 8.

Variable used: Depot number, Ship to ZIP & Item Number

Figure 8:

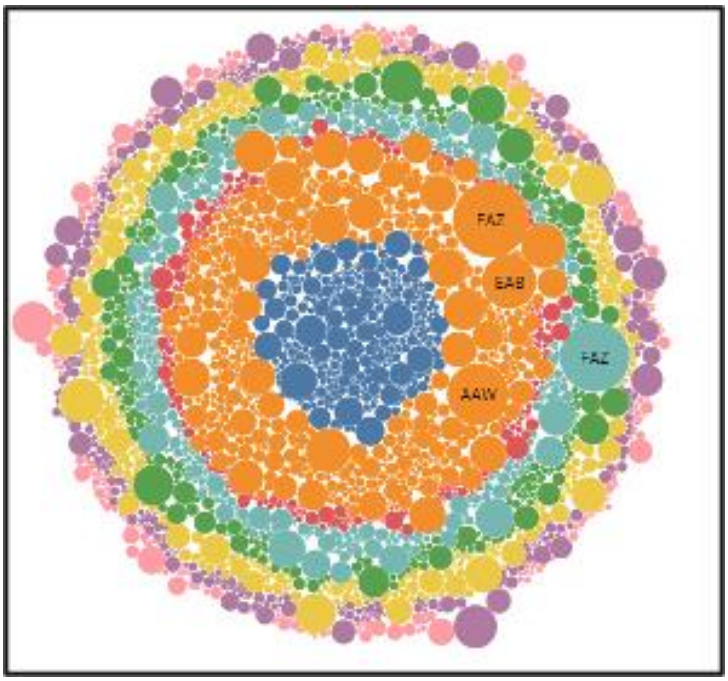


NAÏVE BAYES Classifier:

NAÏVE BAYES is the assumption of conditional independence between every pair of features given the value of the class variable. In our case, items and item categories are classified based on depot number considering features such as which item is being delivered to which states and ZIP codes. We have predicted the class of the test data set with 97% accuracy. We can see the largest Depots by revenue in the Figure 9 below:

Variable: Depot number, Ship to ZIP, Ship to State Item Category & Item Number

Figure 9:



Regression Analysis:

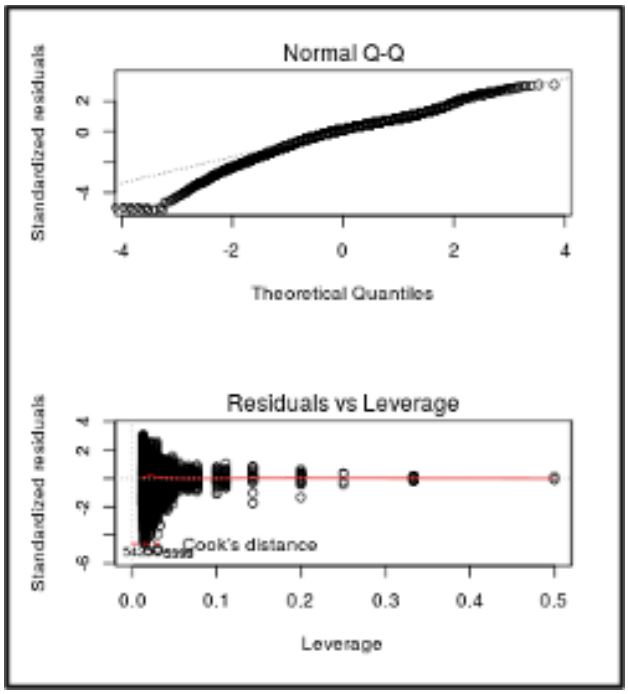
Regression analysis is a powerful statistical method that allows to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

How did it work for us?

For our scenario we were interested in predicting depot number so it was chosen as the dependent variable while the parameters like item_num, ship_to_state, ship_to_zip were chosen as independent variables as we found them to be having high correlation values with the depot number hence suspects to have greater impact when it comes to predicting the depot number. We can see this visualization in Figure 10 below.

Variable: depot_num, ship_to_zip, ship_to_state & item_number

Figure 10:



Performance Evaluation:

Association rules:

We calculated Lift, Support and Confidence for evaluating the performance of the model where Lift ratio > 1 indicates a rule that's useful in finding consequent items sets (i.e., more useful than just selecting transactions randomly) and we have received more than 1 in our case. The confidence value indicates how reliable this rule is and we have got value between 0.4~0.5

Other Models:

Table 3:

MODEL	ACCURACY	P-VALUE	R_2	RMSE	MAPE	MSRE
Demand Forecast	-	-	-	-	20%	4963
Revenue Forecast	-	-	-	-	15%	5570
Decision Trees	96.61	2.2e-16	-	-	-	-
NAÏVE BAYES Classifier	97.96	2.2e-16	-	-	-	-
Regression Analysis	-	2.2e-16	0.98	3.083e-08	-	-

After implementing the various models, it is observed that the accuracy from the Decision Tree and Naive Bayes classifier was quite impressive with 96.61% and 97.96% respectively which means that our trained model is accurate to that percent with that of valid class and have a p-value of 2.2e-16. Also, with our regression model, we were able to achieve the R_2 value of 0.98 with

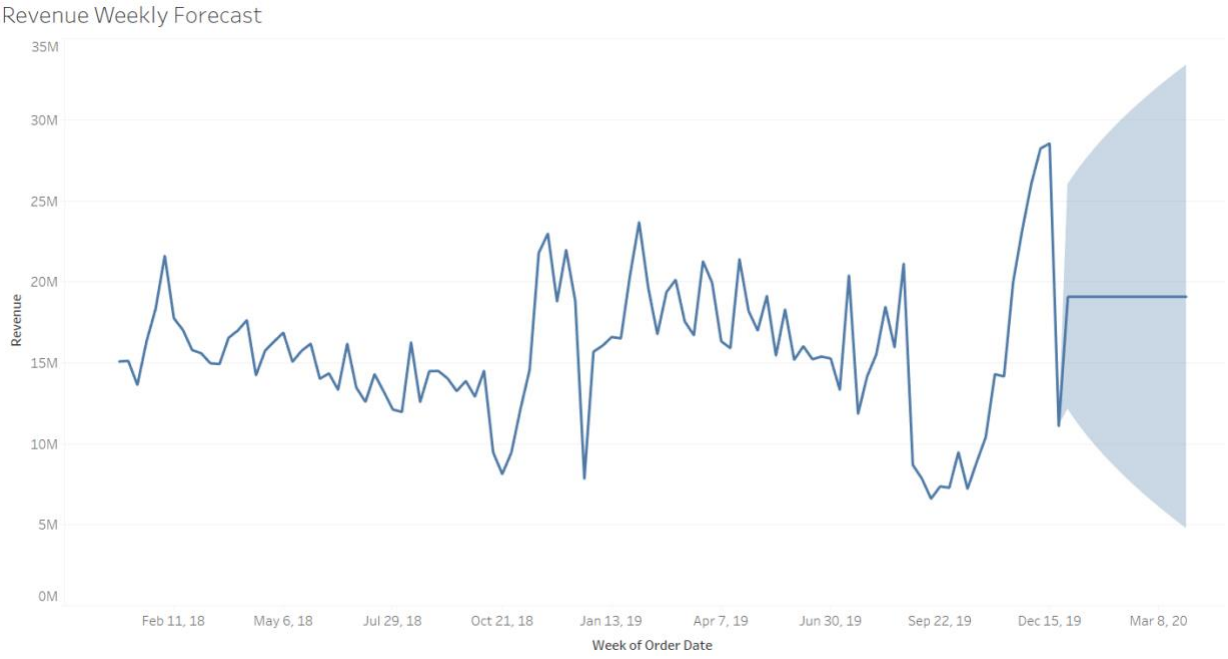
RMSE of 3.083e-08 which is splendid. Coming to the predictive analysis performed on our dataset for helping businesses make informed decisions both demands forecast and revenue forecast models had MAPE in the range between 15 ~ 20 % which opens a window of getting error fairly minimized and showcasing accurate forecasting values with MSRE 4963 and 5570 respectively.

Project Results:

Top Revenue Generating items insight:
We found out that All Wood Cabinet is the highest revenue generating product with $14 M followed by Safe Racks Combo 24" whereas Kirkland Signature 3PCS is the highest ordered product with over 500k plus orders in 18-19 followed by All Wood Cabinet.

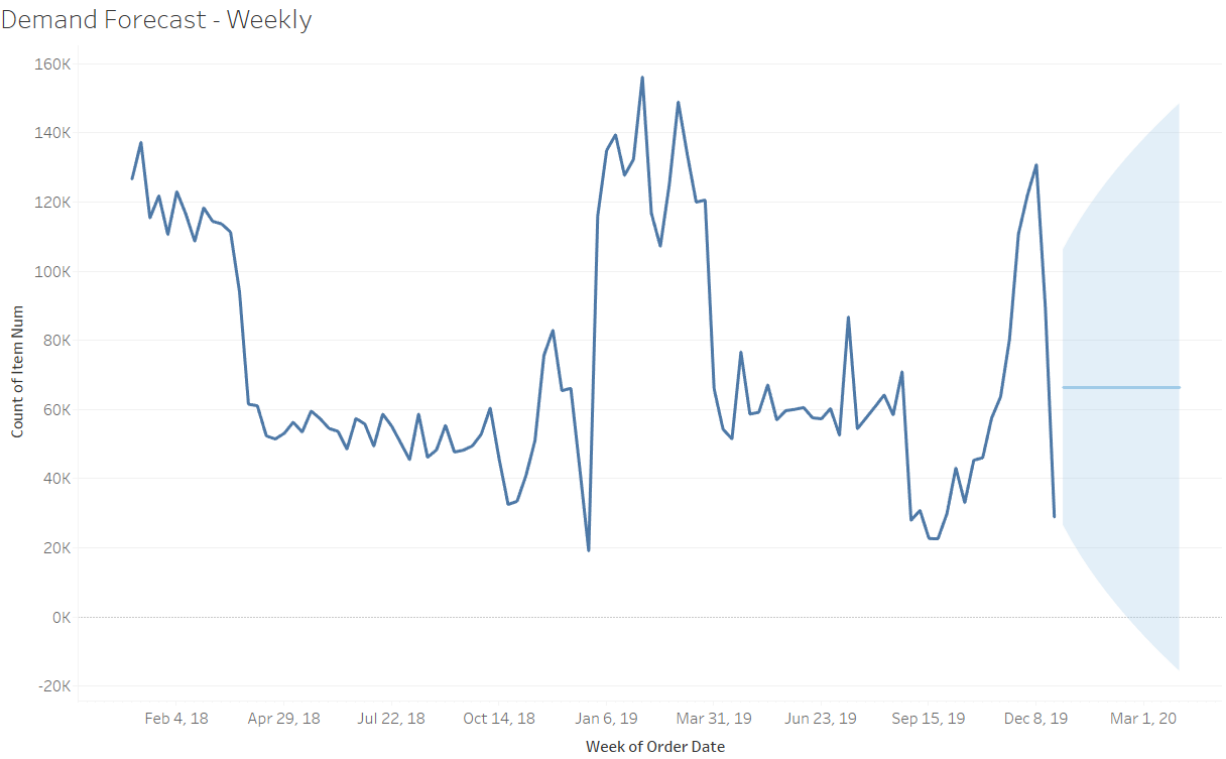
Top Revenue Generating items categories insight:
Jacuzzi and SPA items which comes under Health and Beauty AIDS contributes to almost $100 M revenue with highest in its category

Figure 11:



After performing the time series analysis on the Revenue generated weekly the trend has shown that hardware has been the most revenue generating department throughout. Although, there has been an interesting insight which is a steep fall noticed in the month of September as it was noticed that sale of hardware was declined tremendously. Further during the month of November-December it took up the pace due to festive season. It also forecasts that the span of revenue will be between $5M ~ $30M

Figure 12:



Predicting the demand incurred over the months it was noticed that the span of number of items range between 20k to just below 160k. It was observed from the previous data that there was a peak achieved in the January of 2019 which was brought down by September of 2019. Further it was also noticed from previous 2 years data high peak in the months of November and December due to holidays.

We have mapped the path for 8 depots against features like Ship to ZIP and Item numbers to find the optimum path for delivery goods for buyer visibility. Without pruning and with only 3 features we obtained 96% of accuracy on the training data set with maximum depth up to 7 nodes. To predict the outcome in each leaf node, the average outcome of the training data in that particular node is used

Impact of the Project Outcomes:

Viable demand forecasting can help supply chain network administrators by precisely anticipating item creation and the normalize the organization's income. For what reason is this significant? Having overabundance or inactive stock is hindering to the business. Retailers lose more than $1 trillion all-inclusive because of overload and unavailable circumstances. Utilizing request arranging, your business can examine if it has been meeting the month to month advancement objectives and in the event that it legitimizes the additional expenses of holding overabundance stock.

Utilizing a prescient delivery calculation, many retail chains anticipate and stock distribution center items that are normally bought by clients. Think about the instance of Amazon, it does this by breaking down the historical backdrop of their client's purchasing information. For instance, for paper towel moves, Amazon would envision the number of paper towels required in a given week and change distribution center stock in likewise manner. Utilizing prescient investigation, Amazon can advance business arrangement and excel to higher proficiency with delivery and consumer demand. With progressively precise item determining, the business will be better prepared to design the creation needs. We try to create a similar system in this project by predicting which depot item is more likely to ship a certain type of item, hence the supply chain network can be better prepared on what to stock where and in which region.

By anticipating consumer demand because of whatsoever reasons such as customer deals and special offers etc., the business can all the more likely arrange things like creation, warehousing, and delivery plans. When there is a need to do compulsory upkeep shutdowns or site reboots, the business can dodge the timespans when it gets the most requests to execute these exercises. Moreover, during the months where the business anticipates an increase in demand, the retail chain can work with the suppliers and colleagues to guarantee that stock levels stay high. In case of any foreseen satisfaction delays, it can connect with the clients early and update them on any forthcoming satisfaction issues so they can more readily get ready for this deferral.

Additionally, determining request encourages the business to anticipate any setbacks in deals. With this data, businesses can plan to store money or haggle for advances or credit terms ahead of time to address budgetary issues. Most stockroom blending and allotment or satisfaction rules are just a question of sorting out exercises as indicated by accessible data. Minute efficiencies can be discovered if data is recorded and followed up with.

Our project followed up with the data recorded and used various techniques and algorithms to understand how buyer visibility is fathomed at multinational companies. In this process we helped provide a sufficient solution to a very complex problems that most multinationals face today. Hopefully our project output helps the outcome wanted for this multinational in optimizing their depot allocation.

Bibliography:

Anshumoudgil. "Olist ECommerce-Analytics, Quasi Poisson Poly Regs." *Kaggle*, Kaggle, 7 Feb. 2020, www.kaggle.com/anshumoudgil/olist-ecommerce-analytics-clusters-poly-equation/data.

Shmueli, Galit, et al. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons, 2018.

Zaki, Mohammed J., and Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2017.

Hand, David, et al. *Principles of Data Mining*. The MIT Press, 2001.

Lander, Jared P. *R for Everyone: Advanced Analytics and Graphics*. Pearson Education, 2017.