PROJECT REPORT

ON

ANALYSIS OF SALES FOR SUPERSTORE

Submitted
TO
The Maharaja Sayajirao University of Baroda
In Partial fulfilment for the Degree of
Masters of Science in Statistics

Prepared By:

KINJAL PANARA

YUTI TAILOR

https://www.bibirbal.com

https://www.msubaroda.ac.in/

Date of Submission: *24-06-20*

CERTIFICATE

This is to certify that **Kinjal Panara and Yuti Tailor** have successfully and satisfactorily completed the project titled:

"Analysis of Sales for Superstore"

as a team in the academic year 2020-2021 and have submitted the work to the Department of Statistics in fourth semester as a partial fulfilment for the degree of Master of Science in Statistics and have represented their original work.

I wish them a grand success in future.

Dr. Deepa Kandpal
(Mentor)
Professor at Department of Statistics,
Faculty of Science,
The Maharaja Sayajirao University of Baroda

Prof. V.A.Kalamkar
Head of Department,
Department of Statistics,
Faculty of Science,
The Maharaja Sayajirao University of Baroda

Contents

1. Problem Statement

Context: The context of this project is to do predictive analysis of the sales for superstore data using Statistical Phenomena as a student of Masters of Statistics and interns in BiBirbal family. This project is part of our curriculum in the final year of masters.

Objectives: This study first investigates the secondary data provided by BiBirbal and then works the question that the company wants to know. The main questions are to forecast sales for the next year and on which category of the product and which segment of the customers company should focus on. And we answer these questions with statistical justifications.

Method: Firstly, we investigate the data through EDA and data have been pre-processed as per the requirement of the methods. In forecasting we tried to fit suitable models by checking their assumptions and evaluating them on the bases of RMSE. Some models are working on the data but not fulfilling our objective partially are also mentioned in the analysis. Other questions are answered by ABC inventory analysis.

Conclusions: We have tried to fit different models to different categories and segments of the data. For each of them there are different models like ARIMA, SARIMA, Holt-Winters's method etc which provide the best forecast among them and use them to answer the question for sales forecasting. Some multivariate models have been applied to the data like Random forest and VAR but they fail to meet the requirements. Some model's assumptions were violated so we use alternative approaches to answer questions like ABC inventory analysis and Market Basket Analysis.

2. ACKNOWLEDGEMENT

We would first like to express our deep sense of gratitude and thanks to our Head of the Department Prof. Vipul Kalamkar sir, by whom we got a chance to be part of the BiBirbal family as interns and our project guide Dr. Deepa Kandpal ma'am.

From BIBirbal (https: www.bibirbal.com) Mr. Sangharsh Sapre sir who always helped us, welcomed our questions, kept us motivated and gave us a lot of recommendations and suggestions. We would not have reached this phase, if it were not for him permanent support, advice, and guidance. And Mr. Ajay Joshi sir for their exceptional guidance, supervision and encouragement Throughout this project.

And finally, we would like to thank our parents and friends for their tremendous support and encouragement.

3. Introduction

We are students of M.Sc. Statistics at Maharaja Sayajirao University of Baroda. In our final academic year of masters, we have to do one statistical project as part of the curriculum. Through our Head of the Department Prof. Vipul Kalmkar, we got a chance to be part of the BiBirbal family as interns. They provided us with the project. Our project guide assigned through the department is Dr. Deepa Kandpal ma'am. From BiBirbal Mr. Sangharsh Sapre sir and Mr. Ajay Joshi has guided us throughout the project.

4. About Data

Data was provided to us by BiBirbal. Hence, we have to work with secondary data throughout the analysis. Some questions were asked that are our objective for the project. Data contains sales of superstore for the year 2016,2017 and 2018. It is about the sales of products related to office supplies from a superstore. The data contains sales of various products for 3 Categories ('Furniture' 'Office-supplies' and 'Technology') for 3 years (2016, 2017 & 2018) across India as a market region for 3 segments of customers ('Consumer', 'Home Office', 'Corporate'). The question was asked about forecasting the sales of categories of products and focus areas for the different types of customer segments and categories. [For further details of the data see appendix 1]

5. Objectives

There are few questions that the company wants to know and we are asked to answer them with statistical justifications. Here are those questions as our main objectives.

- Which type of customers should the company focus on for the next 3 years?
- Does the company need to continue to offer diversified products across a range of customers or should it consolidate products and/or customer base?
- You may choose to focus on a specific sub category or product. Please forecast sales quantity for most selling items/products/categories for 2019.

First, we have forecasted the sales with the help of time series models. Firstly, we visualise our time series then we decompose our time series and test for the stationarity then we try to build different models and evaluate all models. From that we have chosen the best model and forecast sales from it. We had done most of the work in python.

For the 2nd and 3rd objective first, we try to answer that question by building a model and predicting the answer. For that we will try to build different regression models after checking the assumption and linearity. Then fit the model accordingly and predict from it. If any of the models won't meet the expectation, that is, the tried model won't meet assumptions or the power of the estimator is low then we go for ABC inventory analysis and from that we answer that question.

6. Methodology

Understanding the data

Firstly, we try to **understand the data** and what it is trying to tell us.

In the **Data Cleaning** part, we check if there are any missing values in the data or the data contains any irrelevant duplicate values or columns. For e.g., name of customers, salesperson and products in sales data. In any case we won't be using them, if needed we will use their Id's. And if the unit price of the product is not mentioned and we have sales, quantities and discounts we can get the price of a particular item.

Understand the data

- Data Cleaning
- Check data type of each variables
- Exploratory Data analysis

Then we move forward to checking the **type of the variables**. For e.g., In sales data Profit will be quantitative variable and categories will be qualitative variable. Now we will see what if our data involves categorical variables. Then let's understand this by an example. Say e.g. In Sales data say we have 3 categories of the products under the column of categories namely "Technology", "Office Supply" and "Furniture". Then the following are the ways.

Ways to work with Categorical Variable

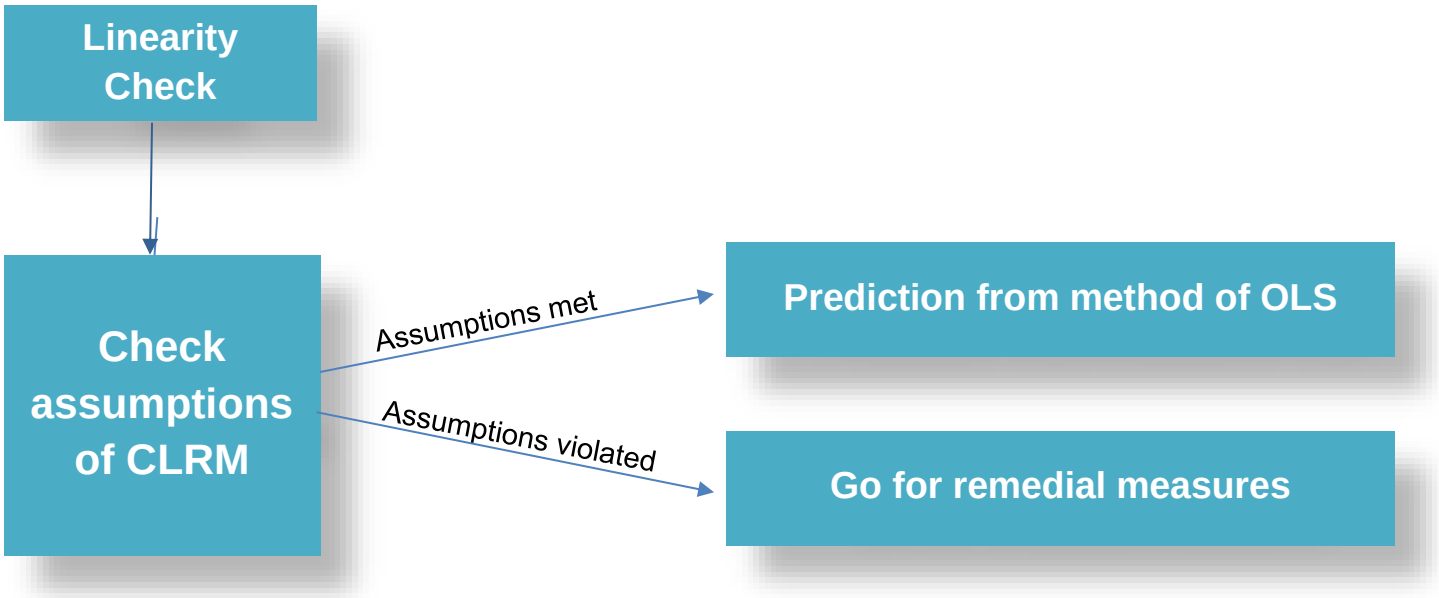
Grouping Data	Dummy Variable	Encoding
E.g., group data by one category say Technology and then analyse	E.g., one variable with 3 categories will make 3 dummy variables Each contain only values 0 or 1	E.g., one variable with 3 categories encoded as; Technology = 1 Office Supply = 2 Furniture = 3

Now in **Exploratory Data Analysis** we will talk about relationships between the variables. As of now we won't bother about feature selection much as we are not fitting the model right now, we just

want to know their behaviour with each other. That can be achieved by either scatter plot or correlation between the variables. And to make it presentable we can use scatter matrix and correlation matrix. Which can easily be done in python. From this we can tell if the data is linear or not and it will be helpful in deciding models and methods to be used in prediction.

Working with CLRM:

Now we will see if we can work with the Classical Linear Regression Model. CLRM has few assumptions that need to be checked. Following are the steps to be done to achieve that.



If we see some level of linear relationship among them then we can apply our simple linear regression. Now to do so first we have to decide which features we want to have in the model and need to check the assumptions of that. So first we select the features by forward regression method and we have done that in python so it tells us which feature we should prioritize. After getting that we have to check the assumptions of CLRM (Classical Linear Regression Model). Software R provides us with a package called “gvlma” to check the assumptions of CLRM. We just have to provide a model which we want to fit and a single line code will tell us which assumption is satisfied and which is violated. And if the assumptions are violated. we have to see if any remedial methods are providing satisfactory results or not.

Working with Time Series:

Working with time series data is different than working with normal data. Firstly, we need to make the data consisting of time points, a time series data. Then some prior steps need to be done as below.

While working with time series data we have to make sure that the time stamp in the data is equidistant. i.e., Hourly, Daily, Weekly, Monthly etc.
For e.g. We have data of sales, which contain dates on which the order has been booked and we have data of 3 years; 2016, 2017 and 2018. So, we have to make this data equidistance; Monthly or Quarterly to have enough meaningful time points. So, if we decided to make them monthly data then we will have data of 36 months.

Data Pre-processing

- Can forecast from both univariate as well as multivariate TS model.
- Prerequire steps before fitting models for TS forecasting;
- **Make time point equidistance** e.g., Monthly, Quarterly, etc.
- **Decomposition of time series** will tell us what type of variation is present in the time series data
- **Stationarity test** done using Augmented Dickey-Fuller test as stationarity is assumption of some time series models

As we begin working with endogenous data and start to develop forecasting models, it helps to identify and isolate factors working within the system that influence behaviour. The decomposition of a time series attempts to isolate individual components such as error, trend, and seasonality (ETS). Stats models in python provide a seasonal decomposition tool we can use to separate out the different components. This lets us see quickly and visually what each component contributes to the overall behaviour.

Then we test stationarity in our time-series data. Stationarity means that the statistical properties of a time series (or rather the process generating it) do not change over time. Stationarity is important because many useful analytical tools and statistical tests and models rely on it.

Modelling:

Time series modelling cannot be done similarly as other modelling in python. Especially when we are talking about the Train-Test-Split part of it. Normally we use train_test_split. It is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. But this function assumes that data points are randomly distributed and select data points randomly to make these subsets. But in time series we desire to forecast future time points from the past time points. For e.g. Given sales of 2016-2018 we want to forecast sales for 2019. Here sales are a time dependent variable. It is not random data points.

So, we need to use a specific method to split the data called Walk-Forward validation. Let's understand this method with the above example. Say we have data of 3 years 2016-2018 and want to forecast for 2019. Then we will take data from 2016 and 2017 as training data sets and 2018 as testing data sets. Then we can predict the sales for 2018 and evaluate the model on the basis of that. Then the best model will be used to forecast sales for 2019 based on the evaluation of all models.

Univariate Time Series Models:

Now as we know Time-Series data can be handled by both Univariate modelling as well as Multivariate modelling. So, let's first understand univariate models.

Univariate Time Series Forecasting

- **Independent variable:** X (Time)
- **Dependent variable:** Y (Any variable)

- **Stationary Time Series:**
  Auto Regressive model of order p [AR(p)]
  Auto Regressive Moving Average [ARMA]

- **Non-Stationary Time Series:**
  When TS exhibits TREND:
  Auto Regressive Integrated Moving Average [ARIMA]
  When TS exhibits SEASONAL Variation:
  Seasonal Auto Regressive Integrated Moving Average [SARIMA]
  When TS exhibits both TREND and SEASONAL Variation:
  Holt-Winter Method [H-W Method]

In the Univariate time series model, there are only two variables. One is an independent variable which contains time points and the other is a dependent variable which depends on the time and we want to predict.

If the time series is stationary then we can just apply the simplest models to Auto Regressive models of any order, say p AR(p) and Auto Regressive Moving averages ARMA.

If data are non-stationary then we can apply the ARIMA, SARIMA and Holt-Winter Methods as per the type of the variation contained in the time series.

ARIMA: 'Auto Regressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. It is useful when time series exhibits a trend.

SRIMA: 'Seasonal Auto Regressive Integrated Moving Average', is a forecasting algorithm. It is useful when the time series exhibits Seasonal variation.

Holt-Winter Method: It is Triple Exponential Smoothing – used for forecasting data with trend and/or seasonality.

Multivariate Time Series Models:

Now we will see which multivariate models are used to forecast time series.

Multivariate Time Series Forecasting

- **Independent variable:** X (more than 2 X, including time)
- **Dependent variable:** Y (Any variable)
- Vector Auto-Regression [VAR] can be used to predict multiple time series variables using single model.
- Random Forest Regression can be used to forecast TS data.

In the Multivariate time series model, there are more than two variables. One is an independent variable which contains time points and other independent variables. Another is a dependent variable which depends on the time and other independent variables, and we want to predict.

First is VAR. Which is the most commonly used method for multivariate time series forecasting – Vector Auto-Regression (VAR). In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables. Another is Random Forest Regression. It can also be used to forecast TS data.

Evaluation of the models:

Now if applicable, one may try several models from these models listed, including Univariate as well as Multivariate models. Then to choose which model is best from this; That is to evaluate these models one may use AIC (The Akaike information criterion) or RMSE (Root Mean Square Error). For both of them, as the value is less the better model is.

7. Additional:

In addition to modelling, to fulfil our requirement of the objectives we have used two additional methods.

Now the alternative way to get information from the data without modelling it, is ABC-inventory analysis and Market Basket analysis.

Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identifies patterns of co-occurrence (that is when two or more things take place together). Market Basket Analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased. This is generally used to know the cross-selling opportunity of the products.

ABC analysis divides an inventory into three categories ("A items" with very tight control and accurate records, "B items" with less tightly controlled and good records, and "C items" with the simplest controls possible and minimal records) to determine levels of importance. This is a categorisation method used to see which are the fast-moving product or slow-moving products.

8. Data Pre-processing

Given data is fairly cleaned and does not contain any missing values. We just need to add one more column of variable that is of Unit Price of the particular order, given enough information was enough to find the unit price. Hence, we calculate that simply in excel by using formula;
Unit Price = (Sales / Quantity) / (1-Discount)

9. Software Environment and Tool used

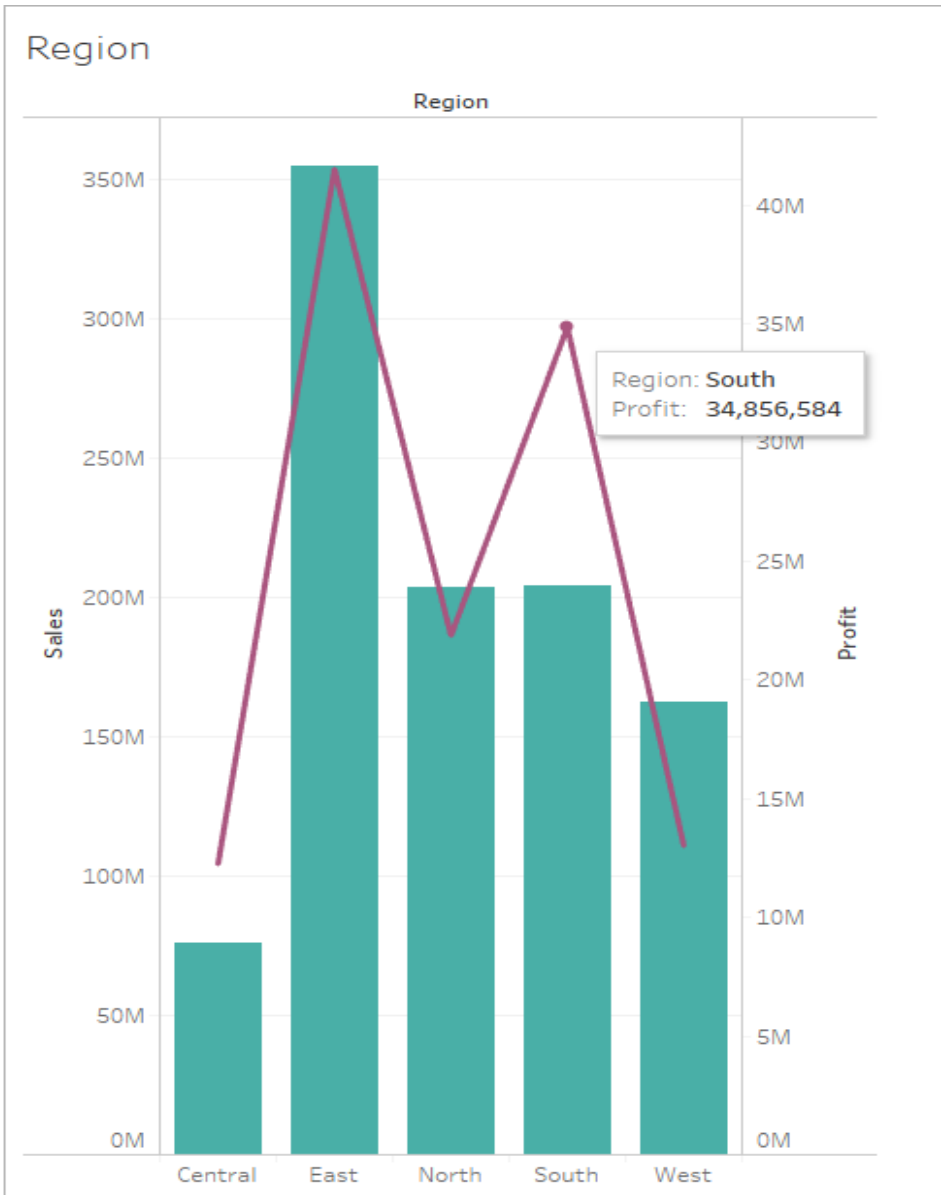
Most of the work in this project has been done by Microsoft Excel, Python's working environment Jupiter, Tableau and somewhat in R-programming and SPSS.

10. Statistical Analysis

Data Visualization:

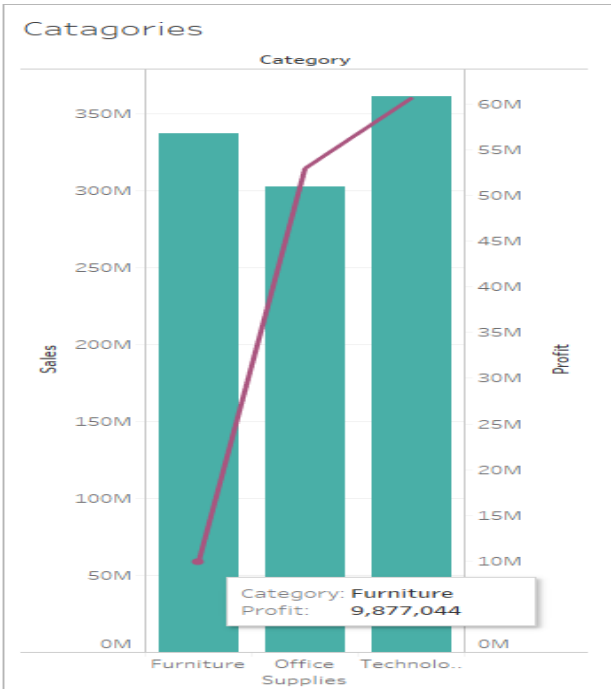
Before any data analysis, first we will do exploratory data analysis.

Exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.



So, let's see what data is trying to tell us. First, we want to see regional behaviour so we plot sales as well as profit of each region. Where the line is showing profit of each region and bars are showing sales of each region.

From this graph we can see that, if we concentrate on sales, that region East is doing well in it and regions North and South are doing almost the same. But by looking at the line we got to know that though the region south has medium sales but the amount of profit gained by that region is really good. It's an obvious thing that more sales don't mean good for the company. So, one should focus on profit more than the sales.



Now we have the main 3 categories of the product. So, let's plot them as well.

To see the behaviour of our products by it's categories we plot it with respect to sales of each category as well as profit gain from each category. We have mainly three categories of the products. Each bar in the graph represents the categories.

From this graph we can see that Technology is providing a good amount of sales and profit but as we discussed above, Furniture's sales are good but profit from that category is really low. That means this category is causing a high amount of loss.

Next, we have 3 segments of customers. Let's see their behaviour.

As per discussed above, the segment's behaviour is different. Segment consumers have the most sales as well as the most profit and the home office has the minimum amount of them but, good amount of profit compared to its sales.
Now, let's go forward to the questions that the company wants to know. So, let's solve them one by one.

Another thing to be noticed is that the rows of the given data contain the information about the purchase of one particular product, the quantity of that item might be anything but the information is only for one product. So, if one person has ordered 5 different products then the order ID of that will be the same for 5 rows of different products. Now one might buy from different categories of the product. So, to understand the buying pattern we go for Market Basket Analysis. That which products are getting sold together more often. And that might be helpful in further analysis.

Objective 1. To study the data and try to predict 2019 sales for 3 categories.

Whenever we are working with time series data, we need to make sure of a few things. Firstly, we need to make data ready for the forecasting. There are some prerequisite steps like a few data preprocessing steps then building a model to forecast the sales and compare these models to find which model works best. Most of the work for this objective is done in the python. These are the steps to forecast the sales. [For detailed background working code see the link of python file in appendix number 2]

Data Pre-processing

- **Make time point equidistant**

While working with time series data we have to make sure that time data points are equidistant or not. That is data should be of hourly, monthly, quarterly or yearly. We have added two more rows to make data equidistant, quarterly and monthly in python.

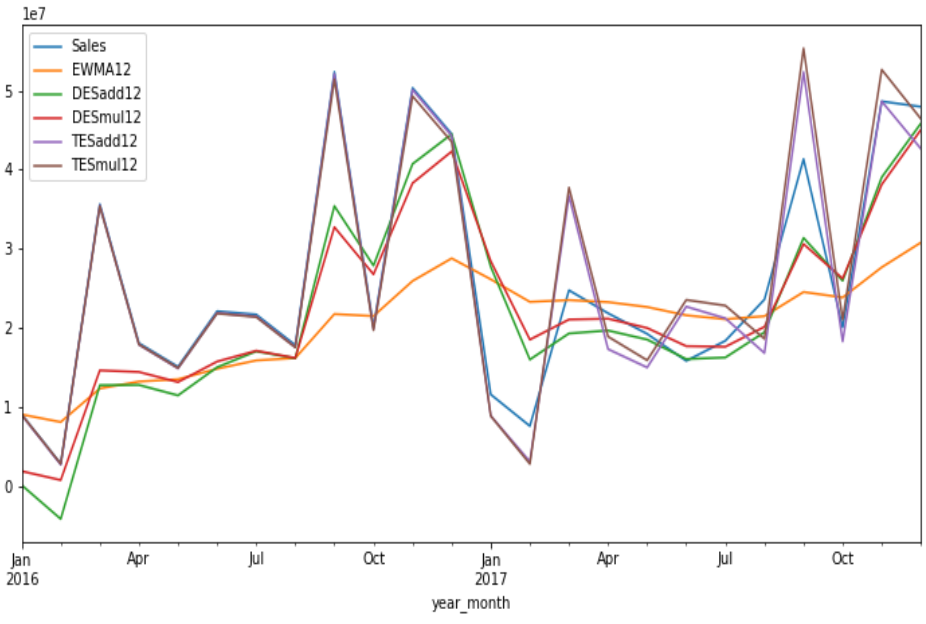
Code:

```
df['Quarter year'] = pd.PeriodIndex(df['Order_Date'], freq='Q')

df['month_year']=df['Order_Date'].apply(lambda x: x.strftime('%Y-%m'))
```

- **Smoothing the Time Series**

Smoothing techniques are kinds of data pre-processing techniques to remove noise from a data set. This allows important patterns to stand out.



The idea behind data smoothing is that it can identify simplified changes to help predict different trends and patterns. [1]

Our data is of sales and by plotting our data seems to have seasonality hence we go for triple exponential smoothing. That is called the Holt-Winter method. By

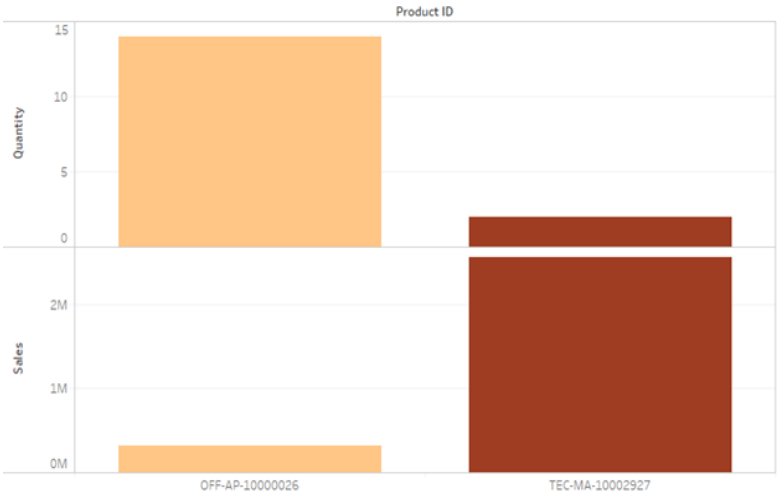
plotting we can see that it does give better results than the single and double exponential smoothing for the same reason.

Model Building

Here, we have time series data for sales of 2016 to 2018 and we want to predict the sales for 2019.

If we want to go one level deeper to understand the sales behaviour then we can see how much one particular product is getting sold. That is it's quantity. Sometimes one might be interested in forecasting quantity rather than sales because of many reasons say how much one particular item should be stored for next quarter. As the revenue from one cell phone will be like revenue from 100 stationary items.

Following graph is an example of that:



Sum of Quantity and Sum of Sales of each Product ID. The data was filtered on Category, which keeps Office Supplies and Technology. The view is filtered on Product ID, which keeps OFF-AP-10000026 (Product Name: Tripp Lite Isotel 6 Outlet Surge Protector with Fax/Modem Protection) and TEC-MA-10002927(Product Name: Canon imageCLASS MF7460 Monochrome Digital Laser Multifunction Copier). Here we can see that item from the office supply (1st bar with light color) have sales lower than 5 lakhs but the quantity at which they are being sold is around 15. Where in the case of item from technology (2nd bar with dark shade) have sales higher than 2 million but the quantity at which they are getting sold is less than 5.

But here our objective is to forecast sales so let's focus on that but if one wants then may proceed in a similar line.

Visualizing TS data:

As we begin working with endogenous data and start to develop forecasting models, firstly we visualise the time series. We see what the behaviour of the sales is.

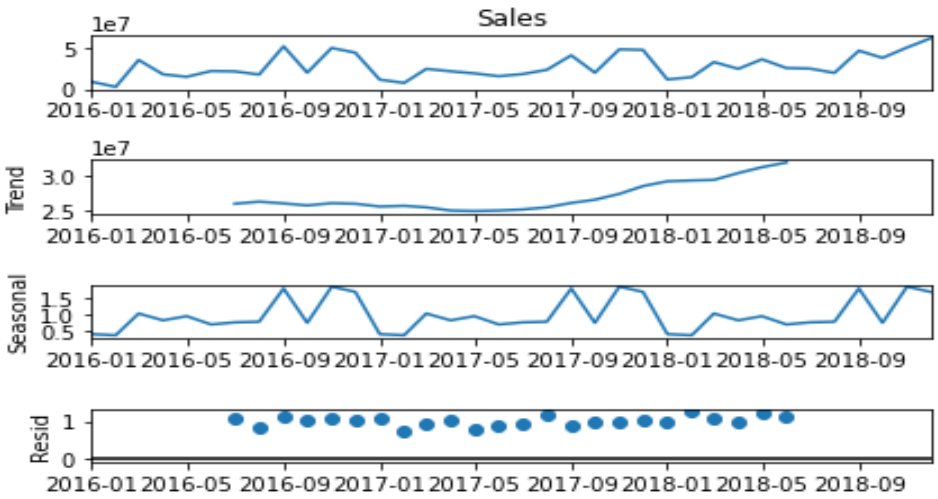
Visualization plays an important role in time series analysis and forecasting. Plots of the raw sample data can provide valuable diagnostics to identify temporal structures like trends, cycles, and seasonality that can influence the choice of model.

These are Time vs. Sales plots.

Now from the graph we have some trend as well as seasonality factor in the data. Let's try to decompose these time series to know the contribution of these factors in the time series. it helps to identify and isolate factors working within the system that influence behaviour.

The decomposition of a time series attempts to isolate individual components such as error, trend, and seasonality (ETS).

Stats models provide a seasonal decomposition tool we can use to separate out the different components. This lets us see quickly and visually what each component contributes to the overall behaviour.



We apply an additive model when it seems that the trend is more linear and the seasonality and trend components seem to be constant over time.

A multiplicative model is more appropriate when we are increasing (or decreasing) at a non-linear rate. From this graph we can see that our trend is increasing in non-linear patterns. So, for overall data some non-linear models should work.

Testing Normality:

Many of the statistical techniques have the assumption that the data are normally distributed. That is our data is a sample coming from a normal distribution. So, as moving towards

the population data it will be normally distributed. Hence before building the model, we need to test whether our data is normally distributed or not.

Here we are using the Shapiro-Wilks test.

Hypothesis: - H0: The sample belongs to normal distribution.

H1: The sample doesn't belong to normal distribution.

Test Statistics: -

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})}$$

Where, xi are sample data points, x^- is average.

Critical Region: - If α > p-value then we reject the null hypothesis with 5% level of significance.

Then we conclude that the sample doesn't belong to a normal distribution. If α < p-value then o.w.

Test for Stationarity:

Stationarity means that the statistical properties of a time series (or rather the process generating it) do not change over time. Stationarity is important because many useful analytical tools and statistical tests and models rely on it.

We will test it by the Dicky-Fuller Test.

Hypothesis: - H0: The data are stationary.

H1: The data are not stationary.

Critical Region: - If α > p-value then we reject the null hypothesis.

Then we conclude that the data are not stationary. If α < p-value then o.w.

Different TS Models

In the Multivariate time series model, there are more than two variables. One is an independent variable which contains time points and other independent variables. Another is a dependent variable which depends on the time and other independent variables, and we want to predict.

First is VAR. Which is the most commonly used method for multivariate time series forecasting – Vector Auto-Regression (VAR). In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables. Another is Random Forest Regression. It can also be used to forecast TS data.

But the requirement of our objective is not satisfied by these models. So, we have to move forward to the univariate models

ARIMA: 'Auto Regressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. It is useful when time series exhibits a trend.

SRIMA: 'Seasonal Auto Regressive Integrated Moving Average', is a forecasting algorithm. It is useful when the time series exhibits Seasonal variation.

Holt-Winter Method: It is Triple Exponential Smoothing – used for forecasting data with trend and/or seasonality.

Here with the help of above steps we have fit these models as per our model fighting technique for time series forecasting. Firstly, we fit the model on whole data but we have categories of the product and their behaviour may not be same in nature so we fit the model on each category. Now sales of each category will be forecasted individually. That means we will try different models mentioned above. Below is the evaluation of these models.

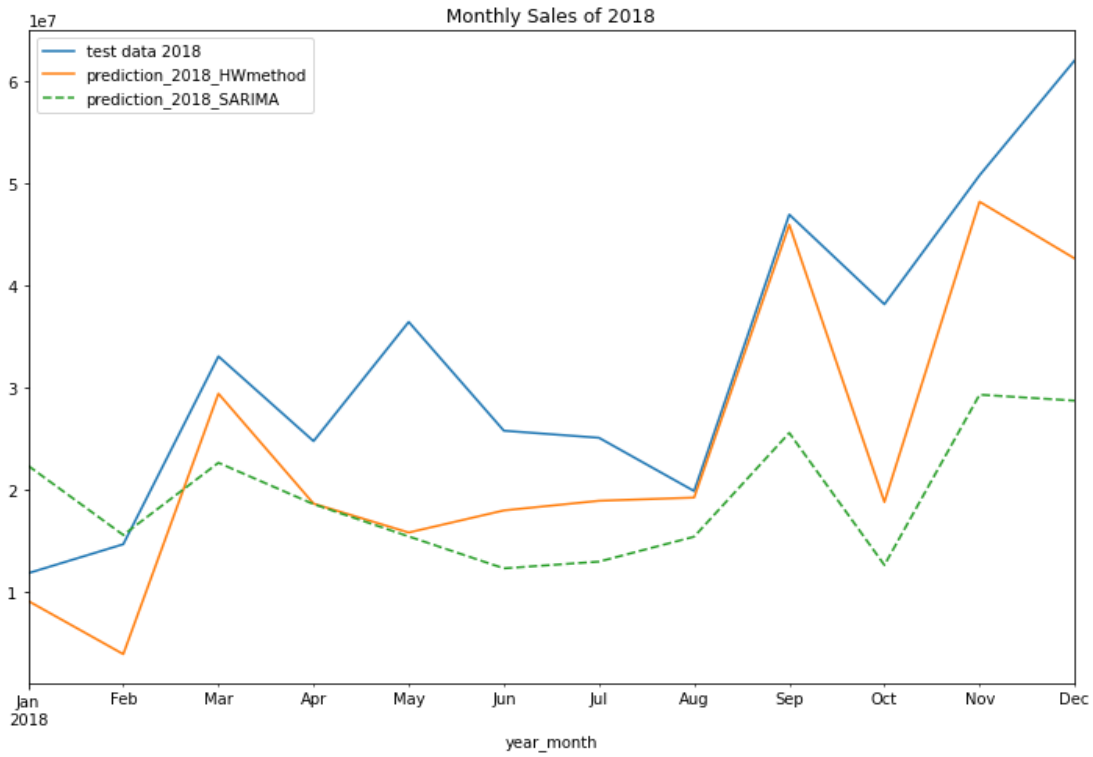
Model fitting is done in python.

Evaluation of Models

Let's compare these models by plotting original and forecasted sales of 2018 by all the models. Let's see which model has done the best forecasting.

For Full data

First, we have fitted the model for full data. Here data contains seasonality so we have fitted the model SARIMA and Holt-Winters's method.



Here, from this graph, the blue line indicates the original sales data of 2018 and we want to compare these models' predictions for sales of 2018 which are indicated by the different lines. Now we can see that the orange line is near to the blue line. So, we may say that Holt-Winters's method shows the most accurate prediction among all models. But we do need a statistical justification for these.

Generally, two methods are used to evaluate the model. They are **The Akaike information criterion (AIC)** and **Root Mean Square Error (RMSE) [**Note: There are few other ways too, to compare the models.]

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

- the number of independent variables used to build the model.
- the maximum likelihood estimates of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. [3]

Root Mean Square Error (RMSE) is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

In RMSE, the errors are squared before they are averaged. This basically implies that RMSE assigns a higher weight to larger errors. This indicates that RMSE is much more useful when large errors are present and they drastically affect the model's performance. It avoids taking the absolute value of the error and this trait is useful in many mathematical calculations.

We have used the RMSE method. In this metric also, lower the value, better is the performance of the model. Below is the RMSE of the models for general sales forecasting.

Model	RMSE
H-W Method	11019628.69
SARIMA	17625317.5

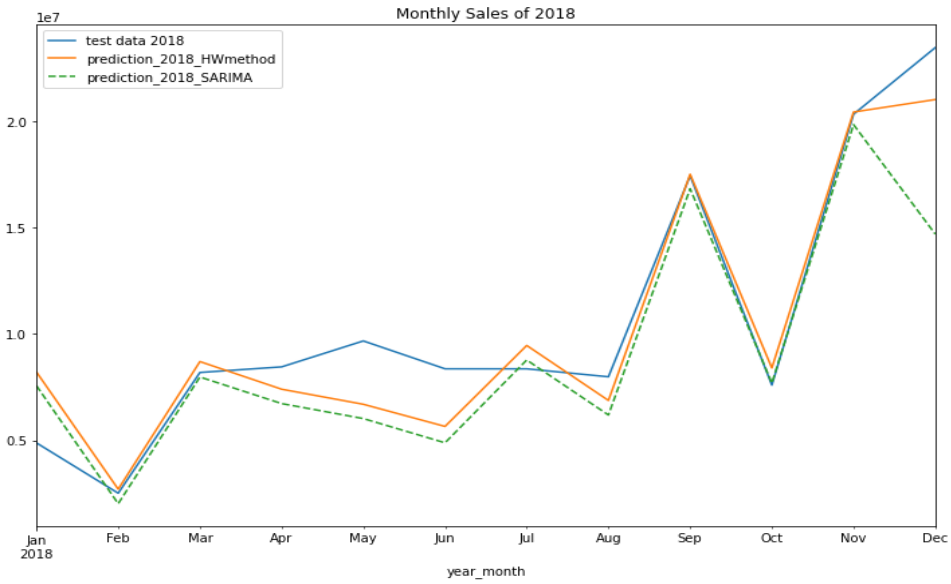
Here we can see that Holt-Winters method has the lowest RMSE. Hence it is the best model for forecasting sales on overall data.

Similarly, we want to forecast sales category wise. As the data of Furniture, Office supplies and Technology have seasonality in it, so we go with the seasonal model, that is H-W Method and SARIMA model.

Furniture:

Data of furniture also contain seasonal effects hence that also behave like full data.

Here, from this graph, the blue line indicates the original sales data of 2018 and we want to compare these models' predictions for sales of 2018 which are indicated by the different lines. Now this graph

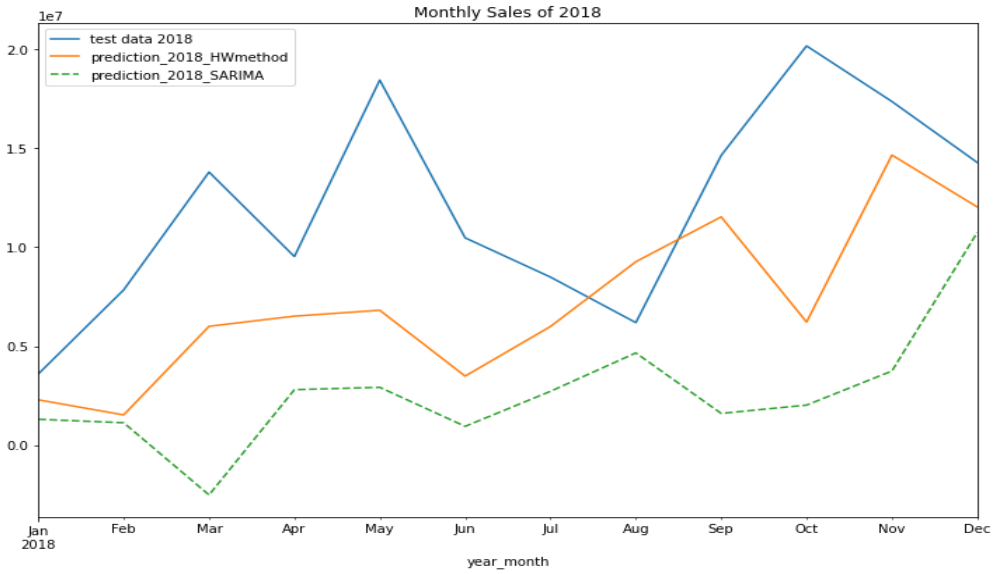


is a bit confusing as in starting the orange and green line are near to the blue line. In that case it is better to go for RMSE. Below are the values of RMSE of particular models.

Model	RMSE
H-W Method	1750708.00
SARIMA	3121926.32

RMSE from ARIMA model is lowest among these so it is best and we will choose this model for final forecasting of the sales for 2019 for category Furniture.

Technology:



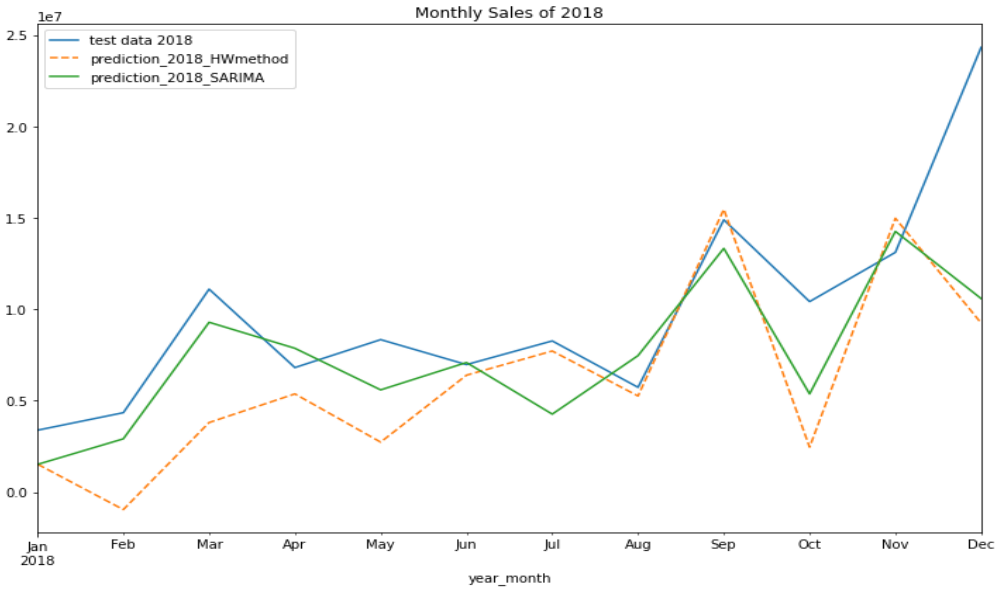
Here, from this graph, the blue line indicates the original sales data of 2018 and we want to compare these models' predictions for sales of 2018 which are indicated by the different lines. Now the purple line is nearest to blue but green is also showing similar patterns and near data points.

So, let's confirm this by RMSE values.

Model	RMSE
H-W Method	6612971.30
SARIMA	10909798.95

The ARIMA model has the lowest RMSE value. So, this is the best model among all and we will use this model for forecasting 2019 sales of categories Technology.

Office Supplies:



Here, from this graph, the blue line indicates the original sales data of 2018 and we want to compare these models' predictions for sales of 2018 which are indicated by the different lines. Now this plot is really confusing and we can't say anything about the best model from this so let's confirm by RMSE values.

Model	RMSE
H-W Method	5868906.58
SARIMA	4603334.27

The ARIMA model has the lowest RMSE value. So, this is the best model among all and we will use this model for forecasting 2019 sales of categories office supply.

Now, we first divided our data into 3 parts according to category (that is Furniture, Technology and Office Supplies) and forecasting Sales of it individually, combining those forecasted Sales of 3 categories, which is final forecasted Sales of our given full data.

Forecasting for 2018 (Sales in crores)

Month	**Forecasted Sales of 2018** for Furniture (H-W Method)	**Forecasted Sales of 2018** for Technology (H-W Method)	**Forecasted Sales of 2018** for Office Supplies (SARIMA)	Sum of 3 forecasted Categories Sales of 2018	Original Sales of 2018	*Forecasted 2018 Sales for full data (H-W Method)*
January	0.820	0.229	0.152	1.201	1.186	*0.915*
February	0.270	0.152	0.291	0.713	1.470	*0.394*
March	0.870	0.600	0.929	2.399	3.309	*2.945*
April	0.740	0.652	0.786	2.178	2.480	*1.868*
May	0.669	0.682	0.559	1.910	3.647	*1.585*
June	0.565	0.349	0.709	1.623	2.582	*1.801*
July	0.946	0.599	0.426	1.971	2.512	*1.896*
August	0.687	0.927	0.746	2.360	1.991	*1.927*
September	1.751	1.153	1.333	4.237	4.698	*4.597*
October	0.840	0.622	0.537	1.999	3.820	*1.884*
November	2.044	1.466	1.425	4.935	5.082	*4.824*
December	2.103	1.204	1.058	4.365	6.207	*4.267*

Now we have our best model to forecast for the future. Here, we have done monthly forecasting for sales of 2018 in general. Above table contains the forecasted values for three categories separately for each month. Fifth column is the sum of sales of these categories and the last column is forecasting from the full data.

Forecasting for 2019 (Sales in crores)

Month	Forecasted Sales of 2019 for Furniture (H-W Method)	Forecasted Sales of 2019 for Technology (H-W Method)	Forecasted Sales of 2019 for Office Supplies (SARIMA)	Sum of 3 forecasted Categories Sales of 2019	*Forecasted 2019 Sales for full data (H-W Method)*
January	0.616	0.433	0.292	1.341	*2.116*
February	0.226	0.586	0.415	1.227	*2.296*
March	1.013	1.457	1.090	3.560	*4.084*
April	0.807	0.916	0.707	2.430	*3.283*
May	0.815	1.350	0.781	2.946	*4.201*
June	0.897	0.788	0.694	2.379	*3.112*
July	0.943	0.725	0.715	2.383	*3.060*
August	0.754	0.704	0.611	2.069	*2.715*
September	1.964	1.534	1.434	4.932	*5.435*
October	0.922	1.403	0.938	3.263	*4.368*
November	2.127	1.709	1.321	5.157	*5.795*
December	2.267	1.929	2.135	6.331	*6.290*

We have forecasted sales for 2019 in two ways. One is forecasted for full data and other is forecasted in categories and sum of them. Now to see which forecasting provides better results we have evaluated both on the basis of RMSE which is done in python only. And we get to know that the model with sum of these 3 categories perform better then overall model to forecast sales for 2018. [The comparison graph for the same is provided in the appendix 3].

Now we have our best model to forecast for the future. Here, we have done monthly forecasting for sales of 2019 in general but for better understanding we have presented quarters and quarterly sales' 95 % Confidence Interval (C.I) for mean. and by following the above 4 steps we can forecast sales for different categories of products too.

Quarters of 2019	95% C.I. for Sales (in crores)
1	(0.15, 3.92)
2	(2.13, 3.03)
3	(0.87, 5.37)
4	(2.69, 7.13)

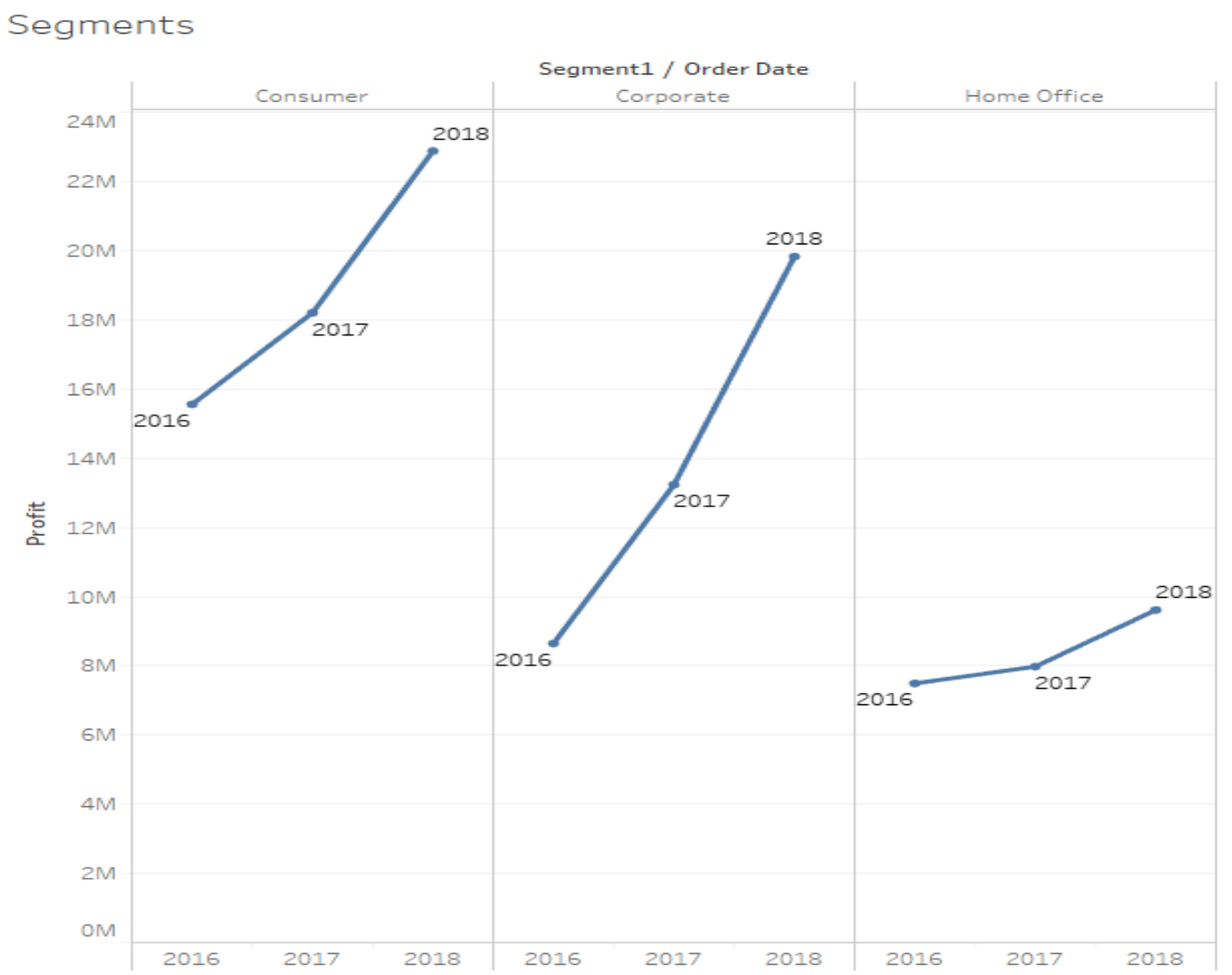
Objective 2. Which type of customers should the company focus on for the next 3 years?

Customers are divided into mainly 3 types called segments.

1. Consumer
2. Home Office
3. Corporate

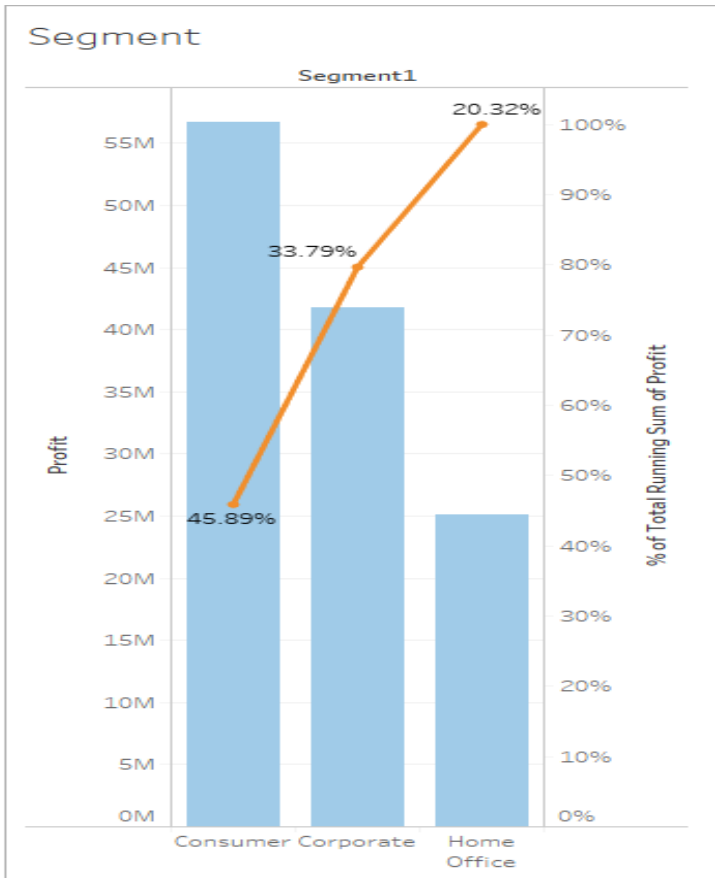
Company should focus on the customers by whom it gets a good amount of profit as well as loss.

Below plot shows profit gain of each segment over given years that are 2016,2017,2018. Company should focus on the customers who are loyal to them and provide a good amount of profit over the years as well as the customers causing loss to the company.



We can see that in all segments the rate of profit increases throughout given years and from segment Consumer we have maximum profit and from segment home office we have minimum profit.

Now with the help of the pareto chart we will decide the segment to be prioritized and to be focused.



This is the graph of segment vs profit and from this graph we can see that around 45% of the profit is gained by the consumer segment.

Now let us focus on the segment consumer. Though it provides the most gain in the profit there are many customers causing loss. We can categorize each customer according to the profit margin. Say customers with a profit percentage more than 1% will be classified in Class A, between 1% to 0.25% will be in Class B and less than 0.25% will be in Class C. So that we can focus accordingly.

Now customers of class A should be taken care of by the company as they provide a good amount of profit. And category C is mostly causing loss to the company. The calculation has been simply done in excel.

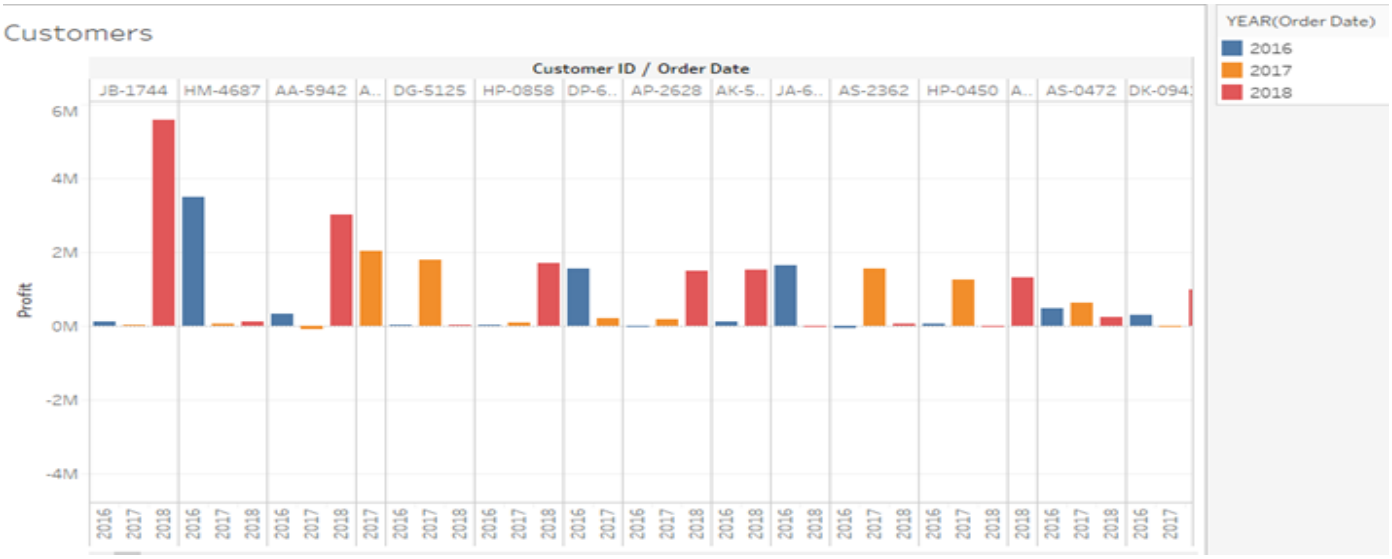
No. of customers in each class (Consumer)

A	B	C	**Total**
36	45	66	147

Likewise, we can see other segments according to the company's interest.

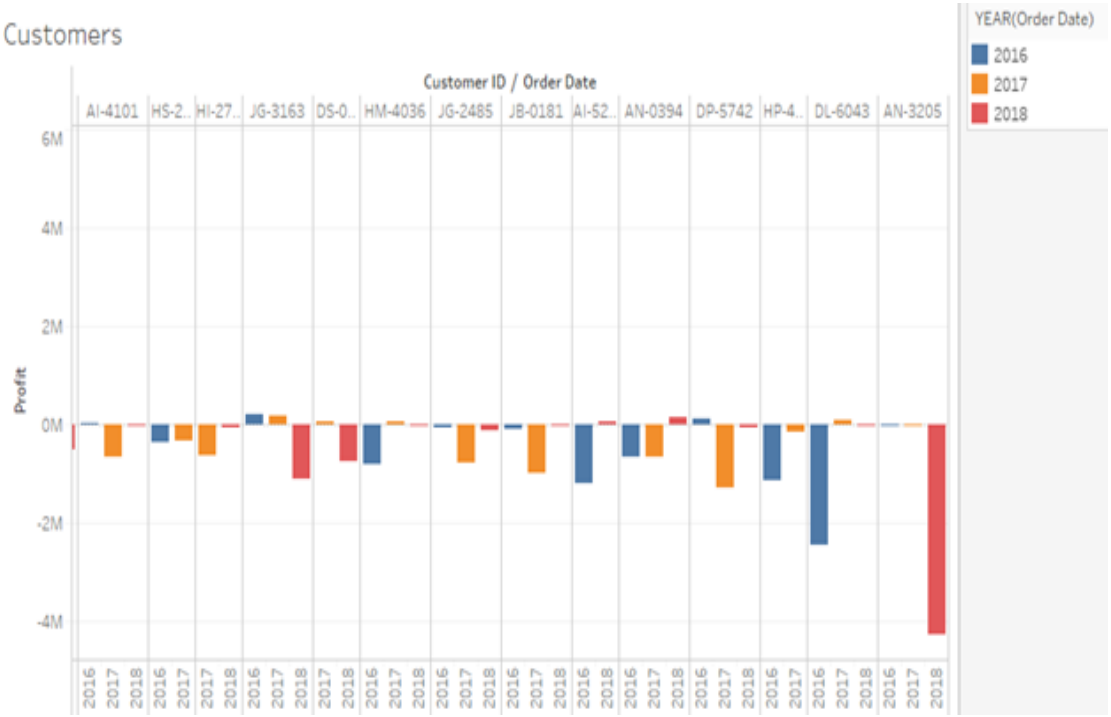
Now let’s focus on each customer’s behaviour through their id.

This is the plot of customer id vs profit gain by them sorted in decreasing order so the 1st graph of each segment shows the customers with maximum profit and the 2nd graph shows the customers with minimum profit.



Each customer's profit per year is in one column and from that we can see that these are the customers who have been loyal to the company and have provided profit most of the time and sometimes really good amounts of it.

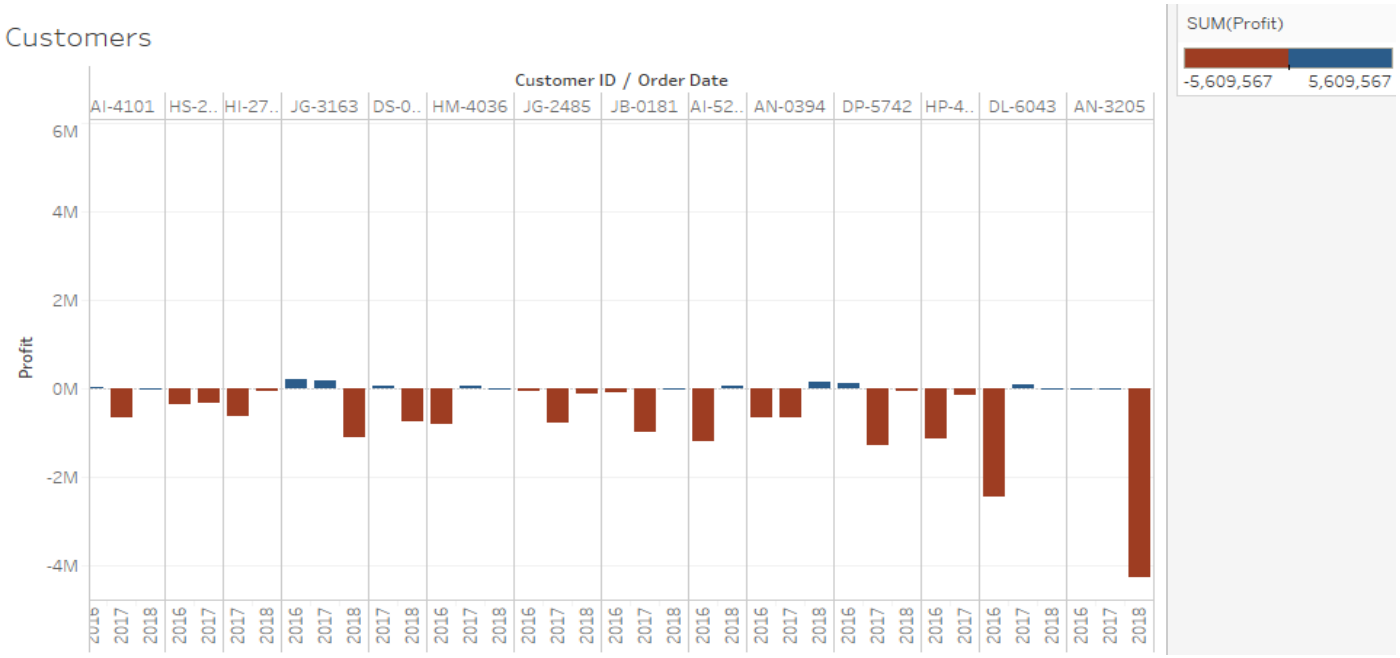
Above graph is sorted in decreasing order of total profit gain by each customer and we can see that we got maximum profit in 2018 plus the profit gain is more in 2018. Hence, we can see that the company's profit margin is rising.



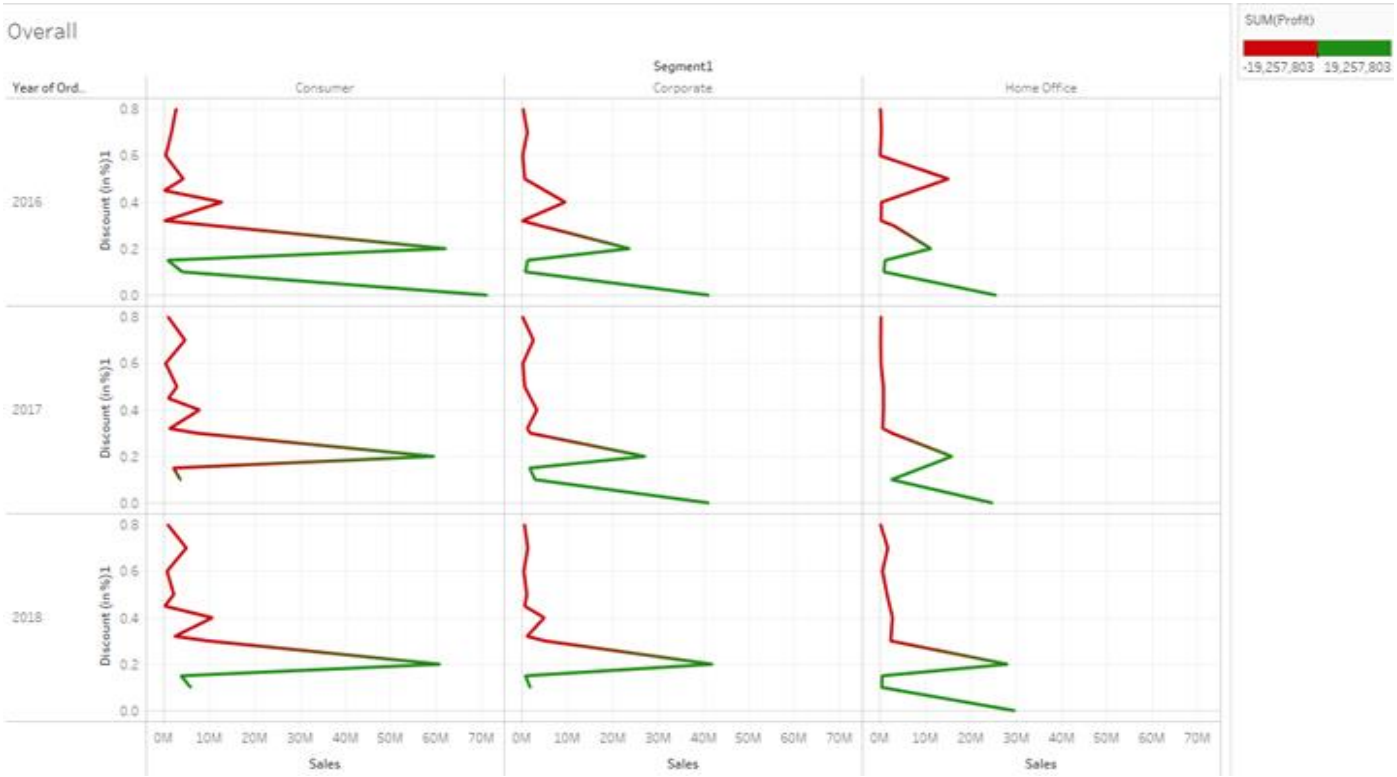
Now let's see about loss. Below graph is of customers with minimum profit.

Maximum loss is also from the year 2018 but other than the maximum loss from the one customer in year 2017 loss was frequent.

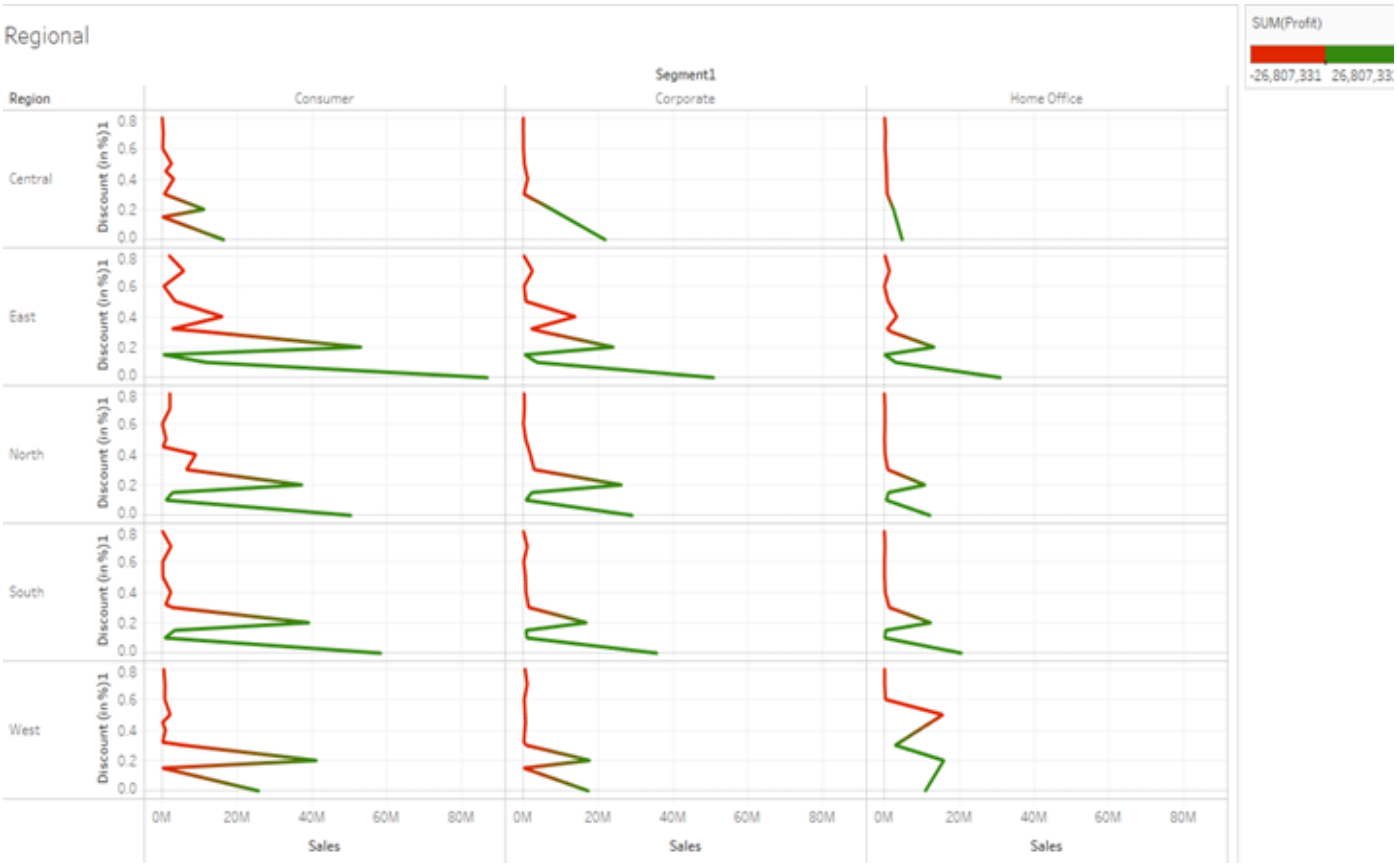
Note that a customer causing loss is a frequent customer but may provide a small amount of profit in past or following years. That can be seen from the below graph. Where red indicates loss and blue profit.



Below graph shows overall behaviour of segments with amount of sales and Discount over 3 years, where colour shows the degree of profit. (Red for loss and green for profit)



From the above graph we can see that overall years and segments the more discount the company provides the less amount of profit they gain, sometimes losses too. Specially to consumers it is advisable to provide a small amount of discount for good amount of sales as well as small amount of profit. High amount of profit results in loss more often.



Above graph for sales in each region of India with colour indicating profit or loss. That is red shows loss and green shows profit. From this also we can see that in most of the cases high discounts result in loss. For a company it is safe to give a discount below 2% but, in the region, central and west it still results in loss.

There is a negative correlation of degree (-0.22) between the variables profit and discount but that is not strong enough to say that more amount of discount will always end up being a loss for the company. But there is correlation between them, of low degree but it is there so we can't just ignore that fact.

[Note: -0.22 is overall correlation on whole data. Individually the correlation for each segment separately is also between [-0.19, -0.25].]

Hence, the company should focus on the customers which are profitable for the company with no or less amount of discount as well as a loyal customer.

Objective 3. Does the company need to continue to offer diversified products across a range of customers or should it consolidate products and/or customer base?

Products are divided into mainly 3 types called Category.

1. Furniture
2. Office Supplies
3. Technology

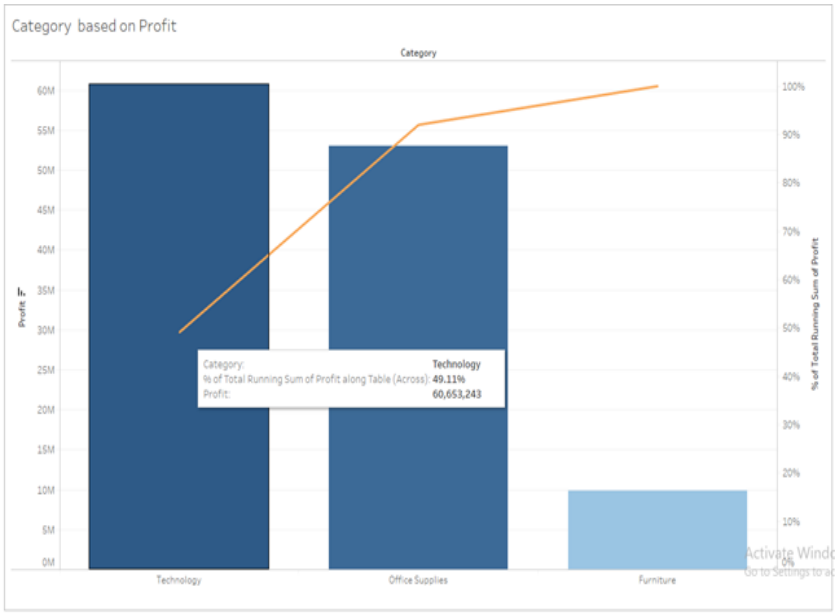
And these Category are divided into Sub-Category as

1. Furniture: Chairs, Furnishings, Bookcases and Tables
2. Office Supplies: Binders, paper, storage, Appliances, Envelopes, Art, Labels, Fasteners and Supplies
3. Technology: Phone, Copies, Accessories and Machines

Company should focus on the Products by which it gets a good amount of profit as well as loss.

Plot shows profit gain of each Category over given years that are 2016, 2017 and 2018

From the above graph we see that Furniture gives minimum profit than the other two Category and in Technology Category and Office Supplies Category the rate of profit increases throughout given years.



Now let us focus on Category Technology. Though it provides the most gain in the profit there are many products causing loss. We can categorize each product according to the profit margin.

Say products with a profit percentage more than 1% will be classified in Class A, between 1% to 0.25% will be in Class B and less than 0.25% will be in Class C. So that we can focus accordingly.

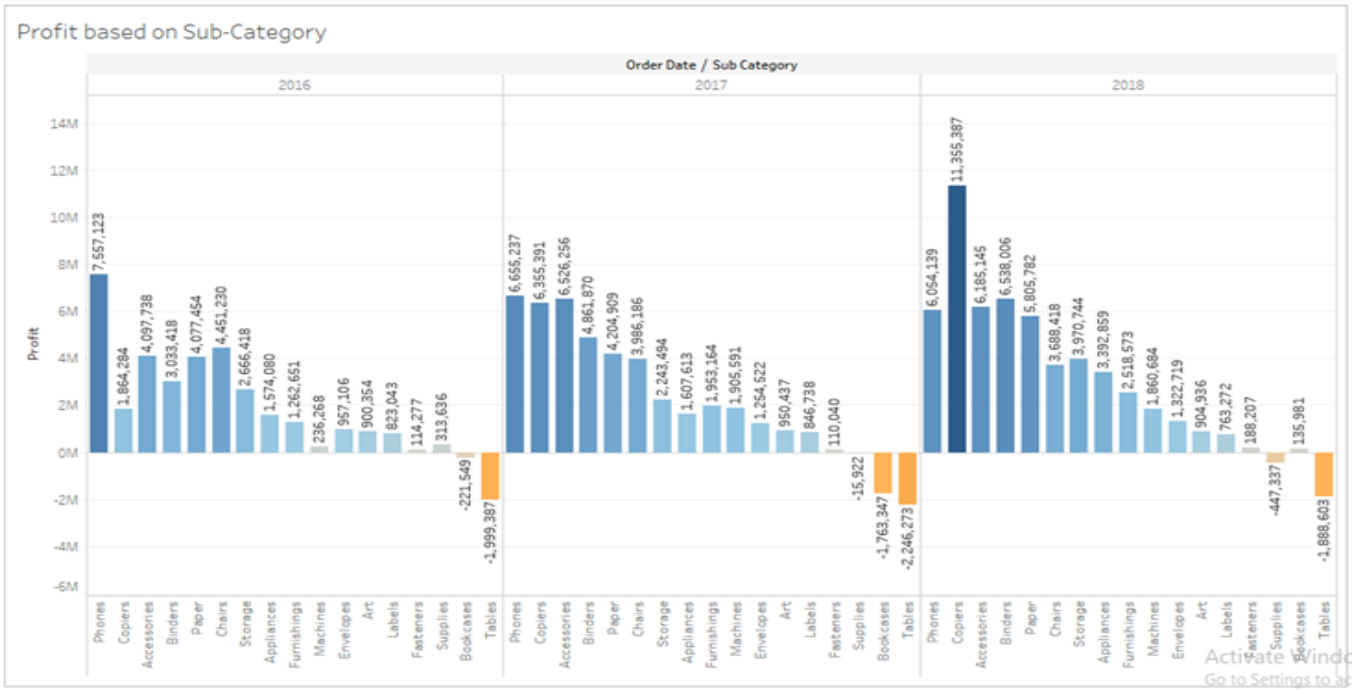
Now products of class A should be taken care of by the company as they provide a good amount of profit. And class C is mostly causing loss to the company. The calculation has been simply done in excel.

No. of products in each class (Technology)

A	B	C	**Total**
25	80	279	384

Likewise, we can see other Categories according to the company's interest.

<u>Plot shows profit gain of each Sub-Category over given years that are 2016,2017,2018</u>



We see that Tables do not make profit over given years that are 2016,2017,2018.

Rate of profit increases throughout given years

- Technology: Copies
- Office Supplies: Binders, paper, Appliances, Envelopes
- Furniture: Furnishings,

Rate of profit decreases throughout given years

- Technology: Phone (although it makes profit)
- Office Supplies: Supplies (it makes loss)
- Furniture: Chairs (although it makes profit)

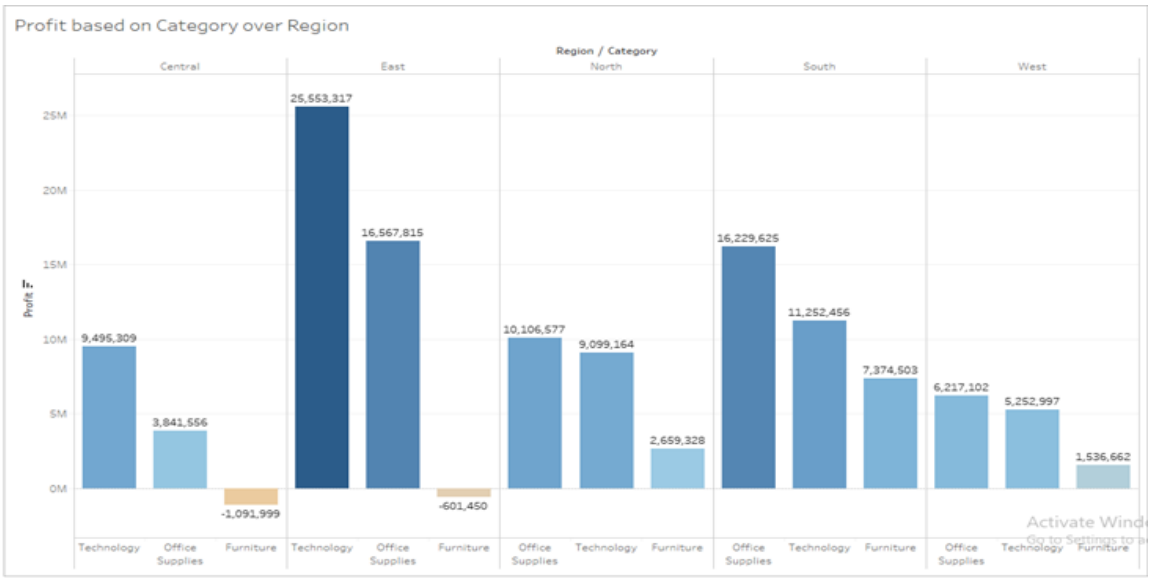
Rate of profit 1st decreases (in 2017) and then increases (in 2018) in given years

- Technology: --
- Office Supplies: storage, Fasteners
- Furniture: Bookcases, Tables

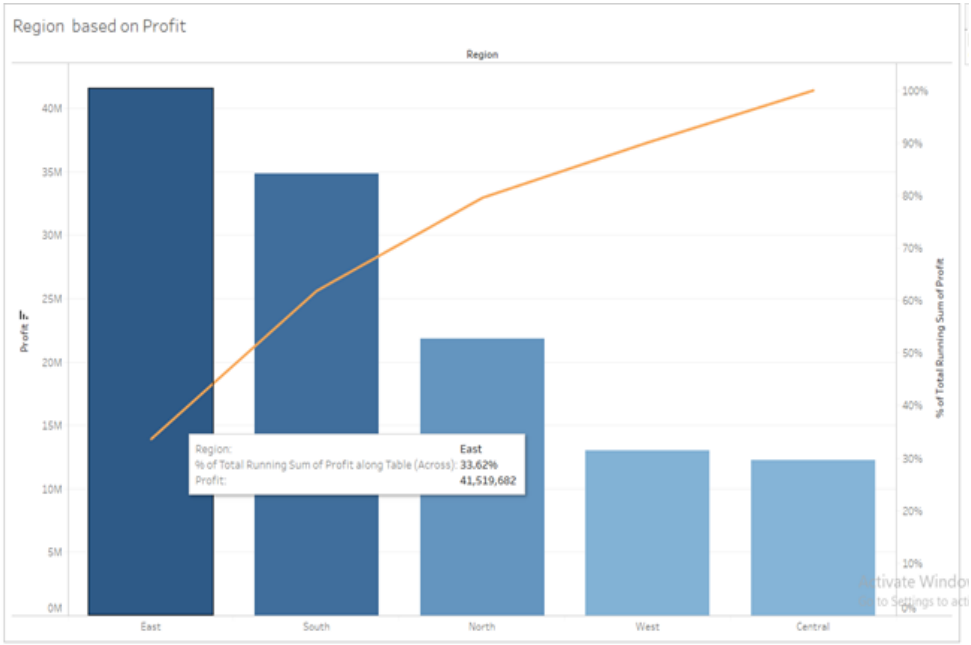
Rate of profit 1st increases (in 2017) and then decreases (in 2018) in given years

- Technology: Accessories, Machines
- Office Supplies: Art, Labels,
- Furniture: --

Plot shows profit gain of each Category over given Region: Central, East, North, South and West



From this graph we see that, In North, South and West profit gain by each Category but in Central and East Furniture makes loss.



Now let us focus on Region East. Though it provides the most gain in the profit there are many products causing loss. We can categorize each product according to the profit margin. Say products with a profit percentage more than 1% will be classified in Class A, between 1% to 0.25% will be in Class B and less than 0.25% will be in Class C. So that we can focus accordingly.

Now products of class A should be taken care of by the company as they provide a good amount of profit. And class C is mostly causing loss to the company. The calculation has been simply done in excel.

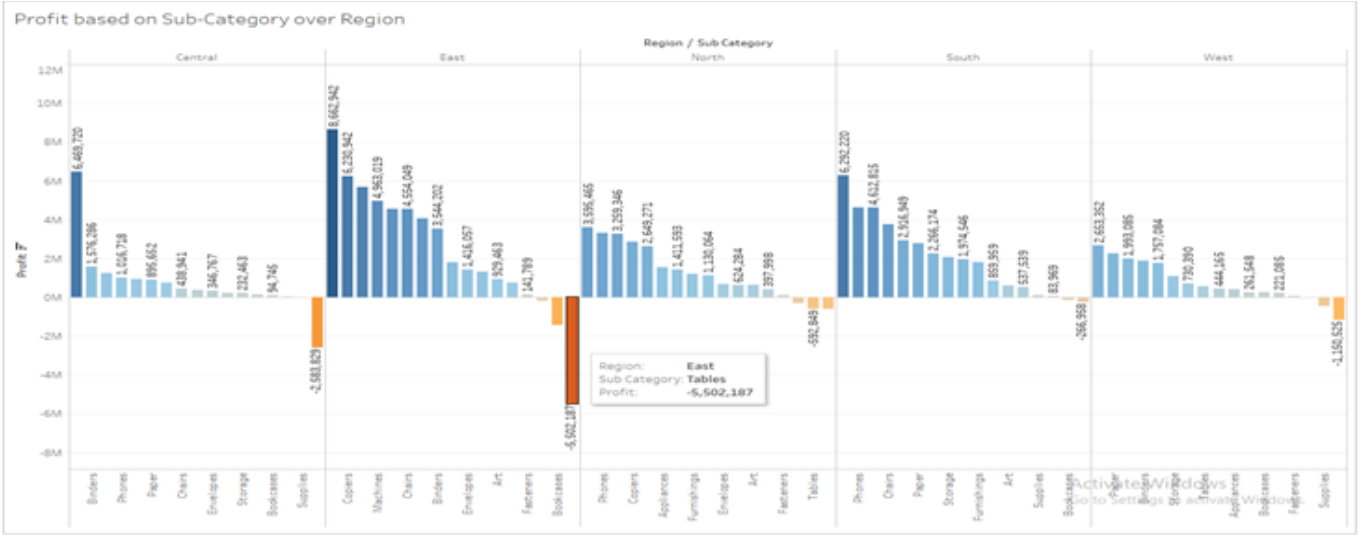
No. of products in each class (East Region)

A	B	C	**Total**
24	101	1193	1318

Likewise, we can

see other regions according to the company's interest.

Plot shows profit gain of each Sub-Category over given Region: Central, East, North, South and West



From above graph we see that,

In Central: Supplies and Tables make loss.

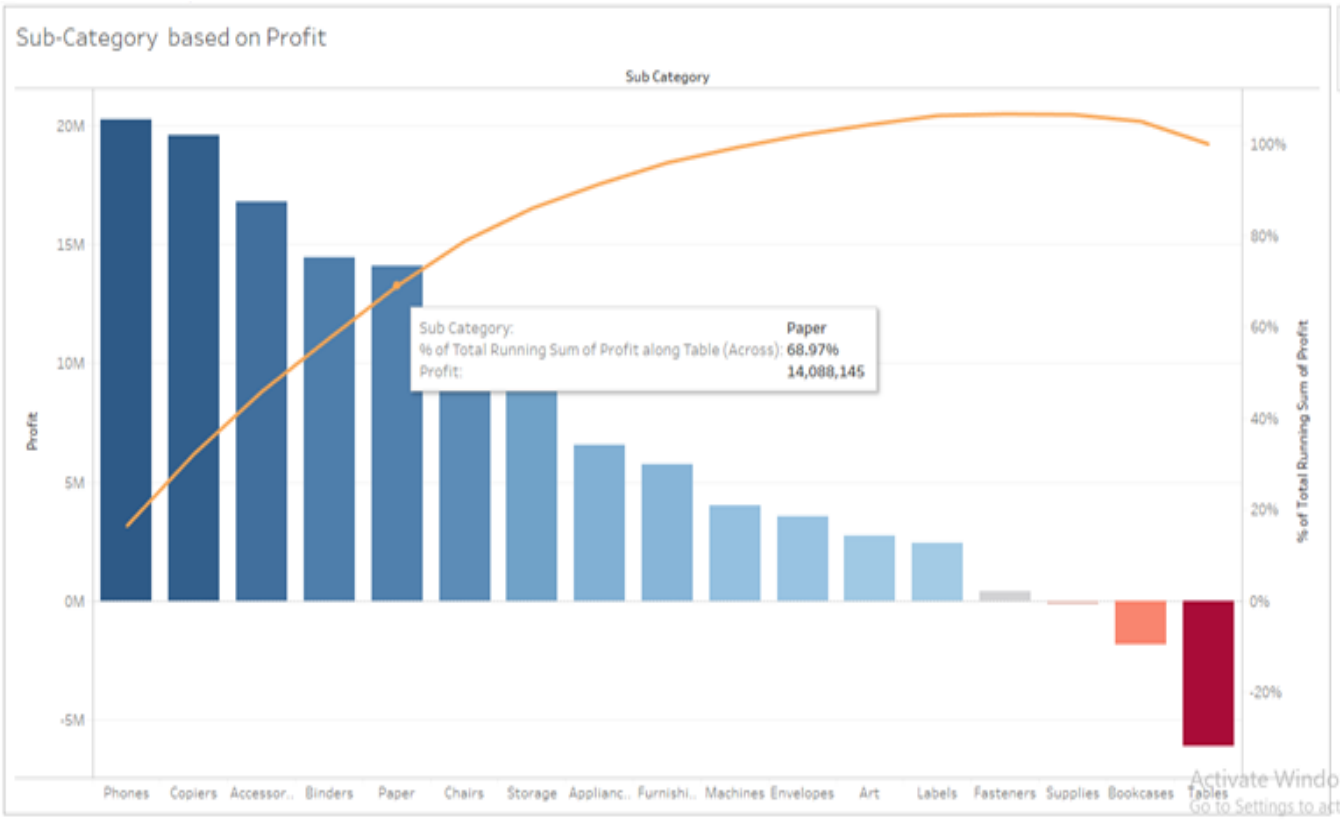
In the East: Supplies, Bookcases and Tables make losses.

In the North: Machines, Tables and Bookcases make losses.

In the South: Bookcases and Machines make losses.

In the West: Furnishings, Supplies and Machines make losses.

<u>This is the plot of Sub-Category vs profit gain by them sorted in decreasing order so it shows the Sub-Category with maximum profit (over the given period of time).</u>



From the above graph we see that 68.97% profit is given by Phones, Copiers, Accessories, Binders and Papers from overall profit.

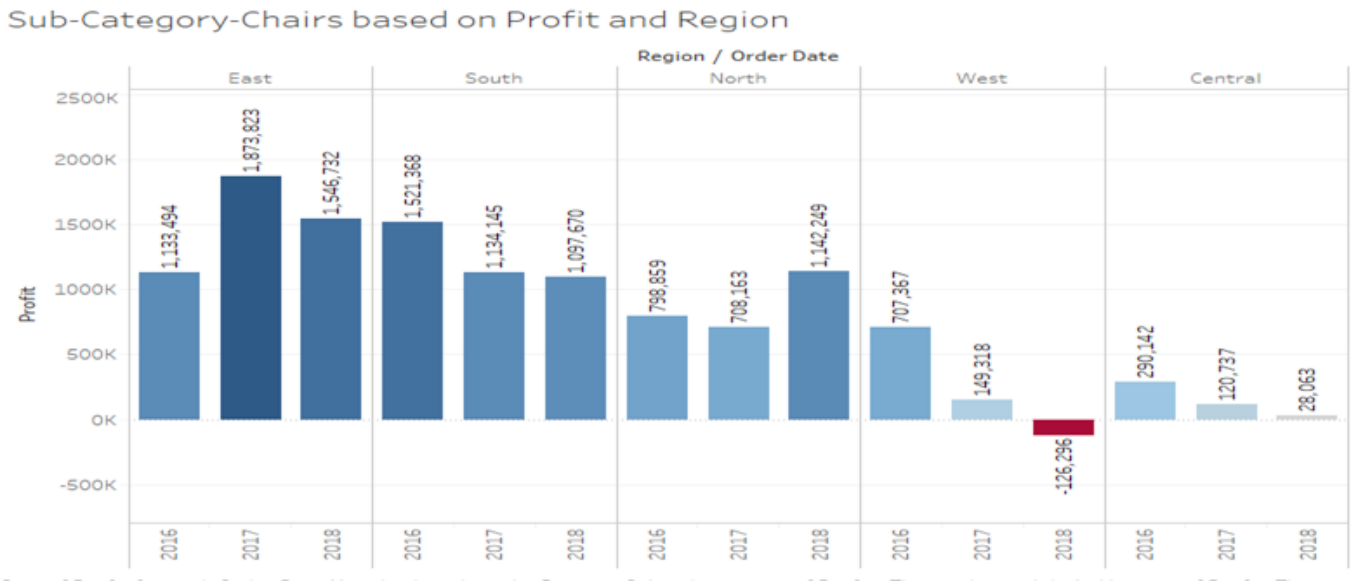
For 2016: 73.22% profit is given by Phones, Chairs, Accessories, Papers and Binders from overall profit.

For 2017: 72.53% profit is given by Phones, Accessories, Copiers, Binders and Papers from overall profit.

For 2018: 68.65% profit is given by Copiers, Binders Accessories, Phones, and Papers from overall profit.

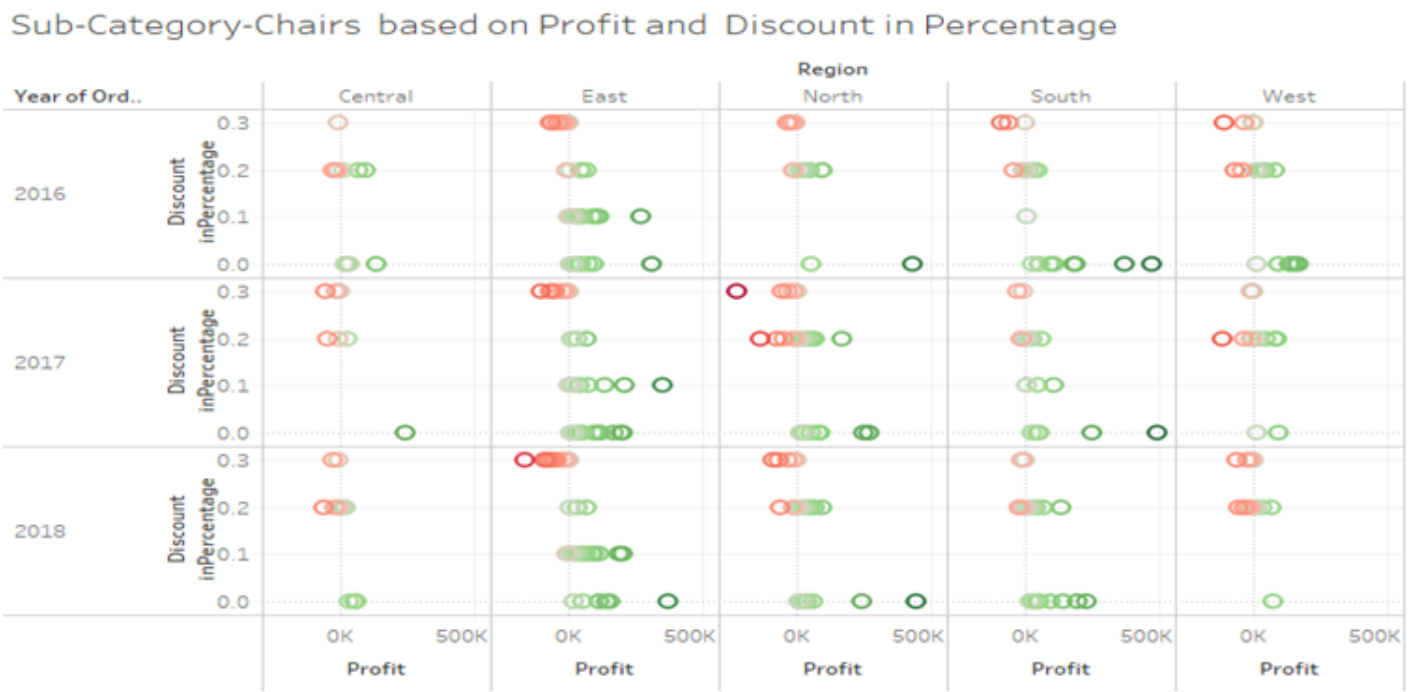
Furniture: Chairs, Furnishings, Bookcases and Tables

❖ For Chairs



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by the sum of Profit. The data is filtered on Sub-Category, which keeps Chairs.

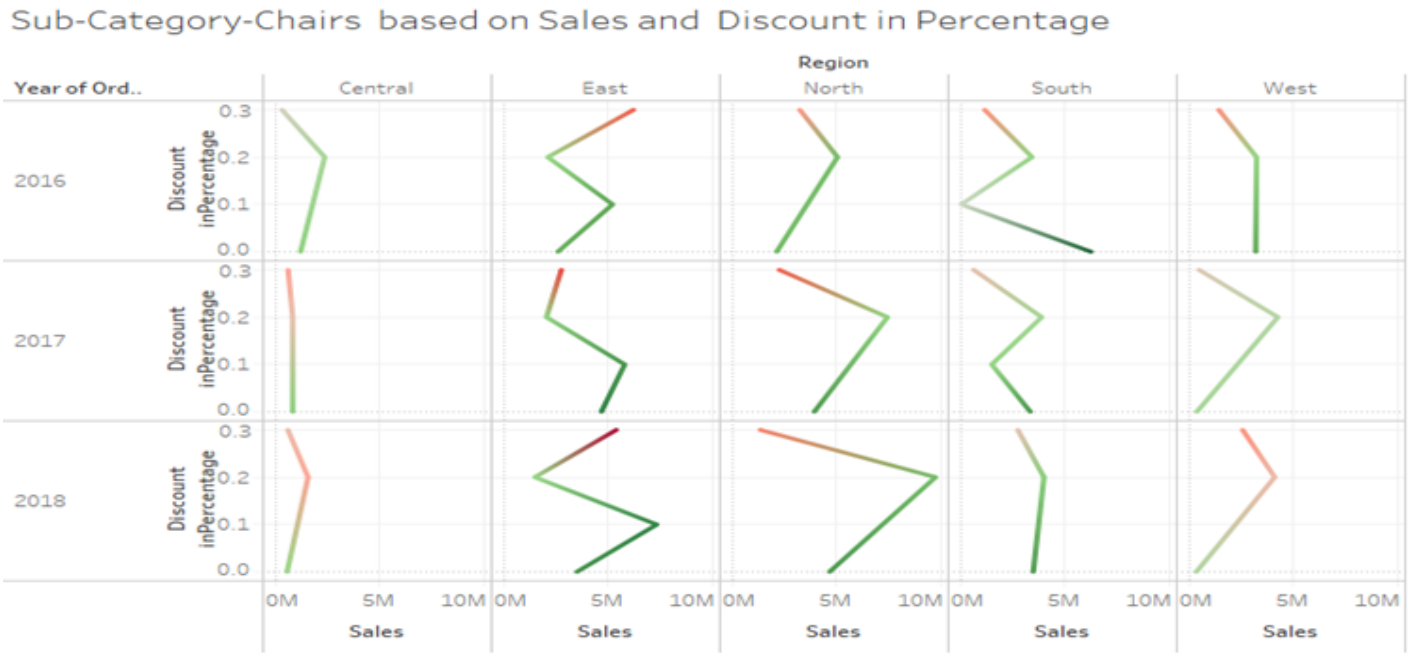


Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Chairs.

2017 and 2018 in the East Region have more profit, from the graph we see that with less discount on product we gain more profit.

And in 2018 in the West, we get losses due to giving more discounts.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Chairs.

From above graph we see that,

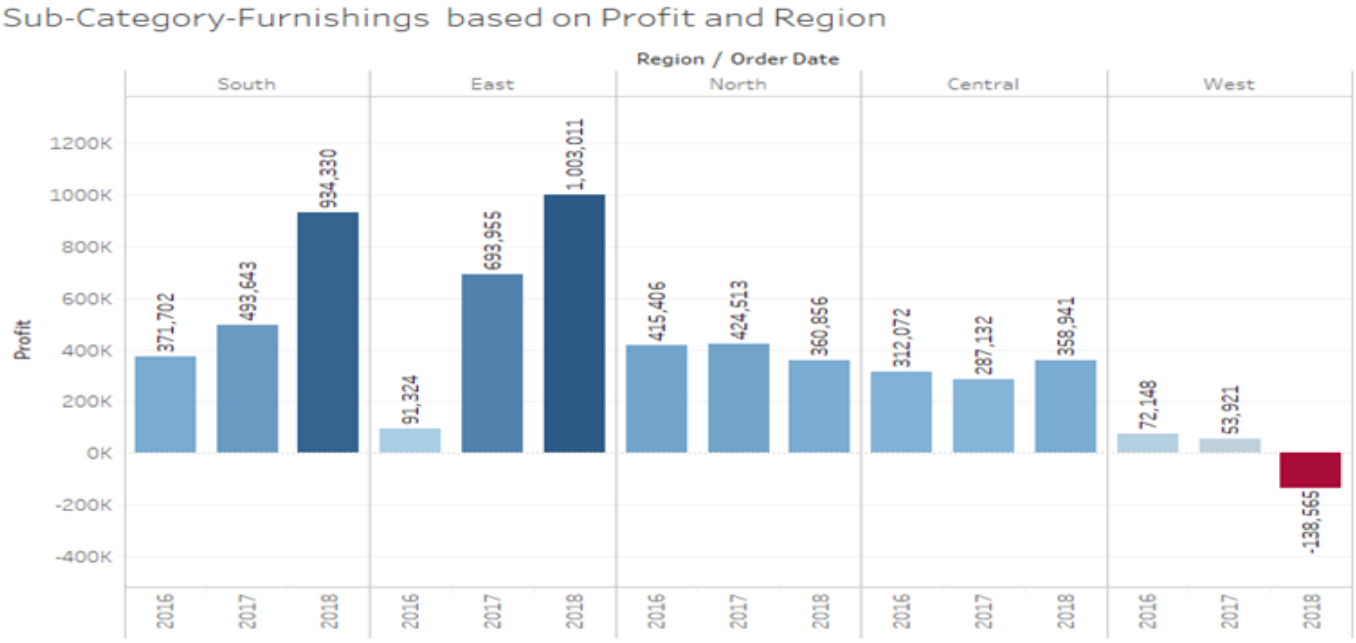
For Central Region: there is not much more effect on sales as we increase discount, but as increases in discount leads to decreases in profit

For the East Region: as we increase discounts from 0 % to 10 %, our sales also increase and its leads to making profit.

As we increase discounts from 10 % to 20%, our sales decrease. And as we increase discounts from 20% to 30 %, our sales increases, but as increases in discounts leads to decreases in profit.

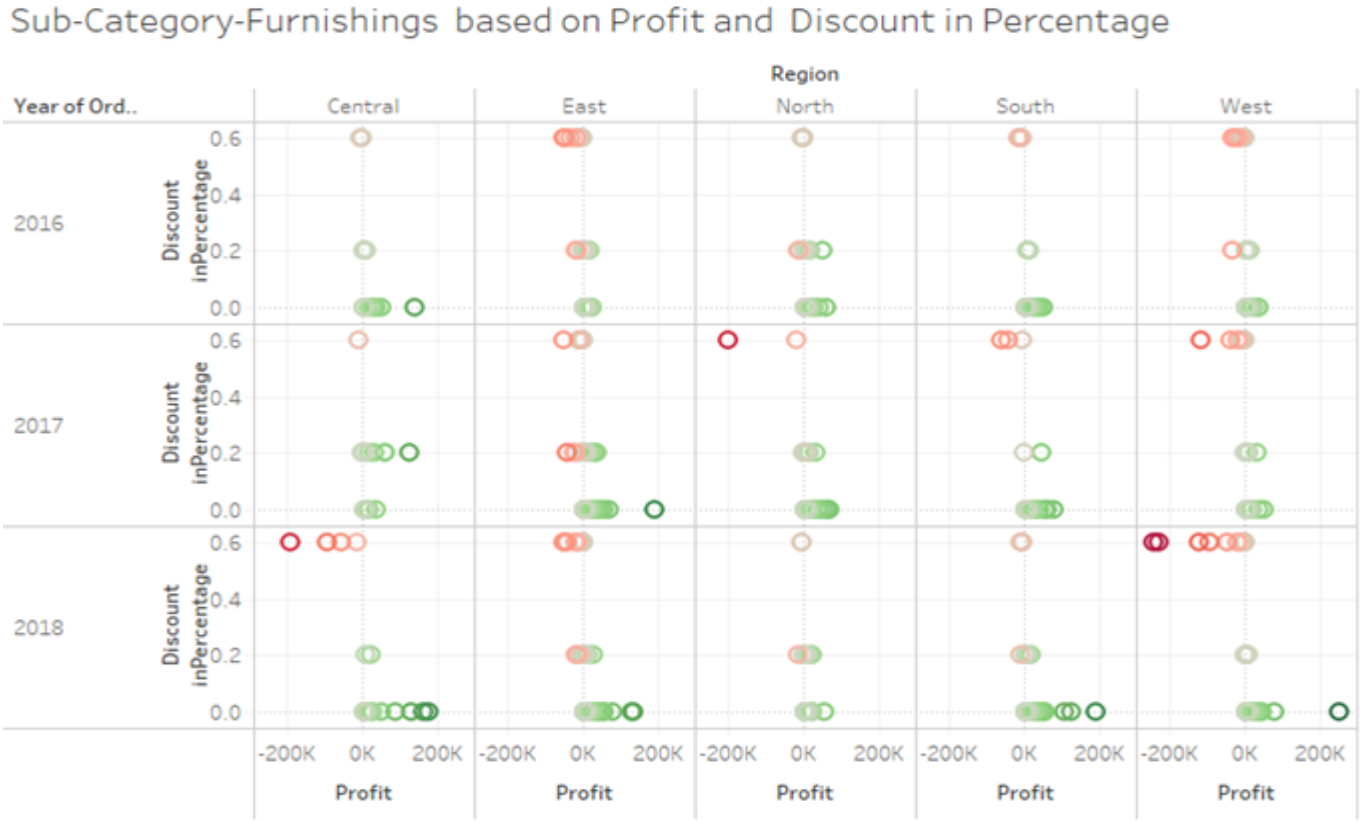
For the North Region: As we increase discounts from 10 % to 20%, our sales increase and give profit.

❖ For Furnishings



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Furnishings.

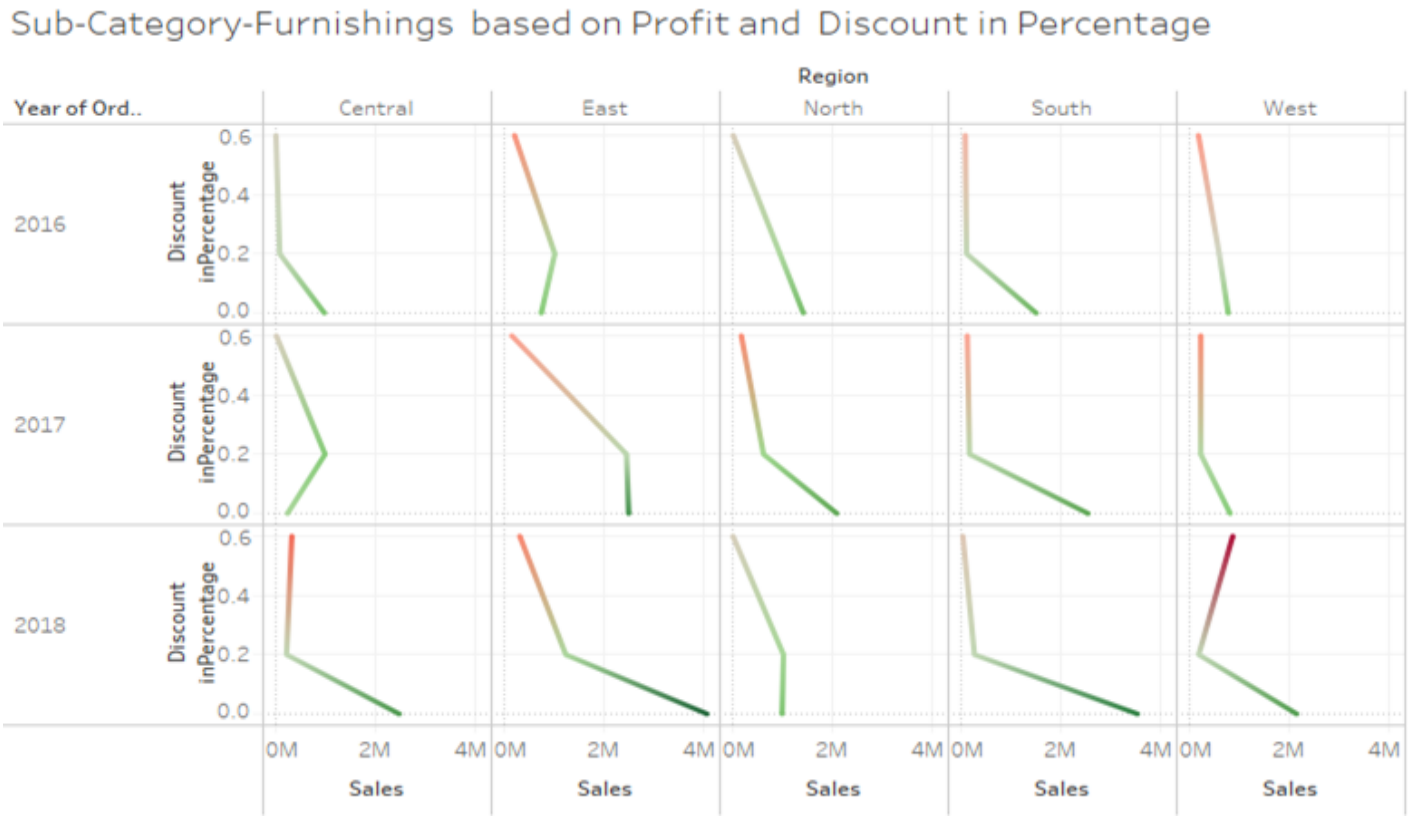


Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Furnishings.

In the South and East region our profit increases with less discount.

And in 2018 in the West, we get losses due to giving more discounts.

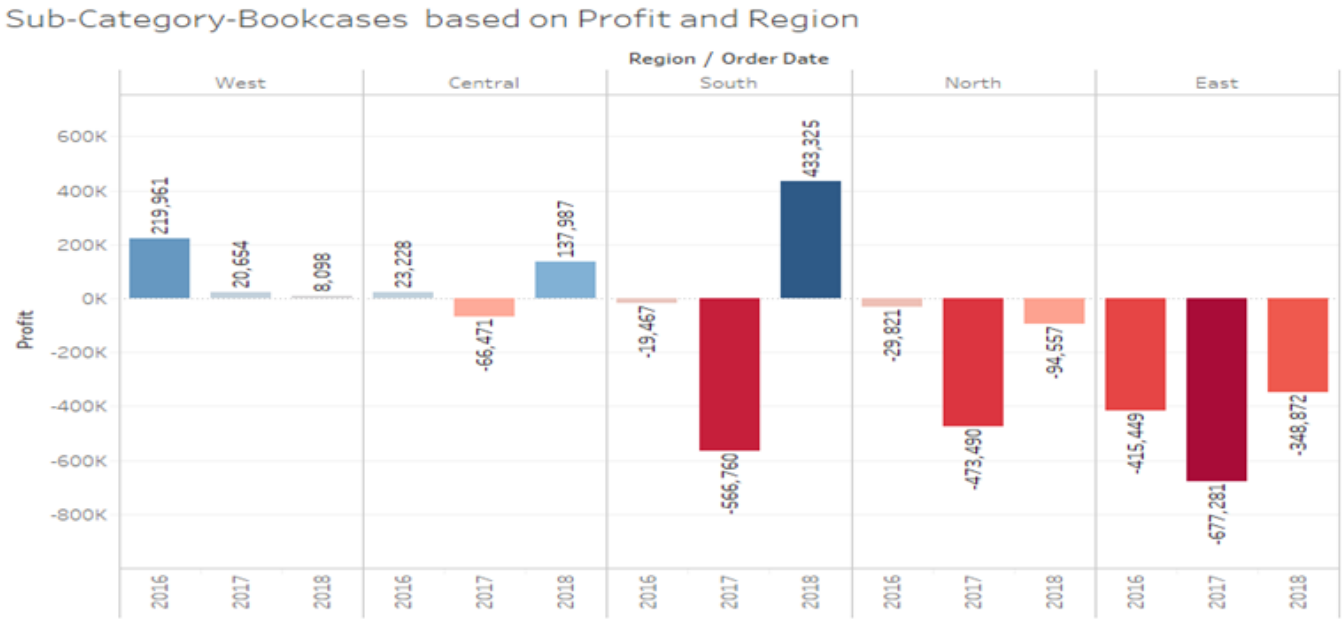
Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Furnishings.

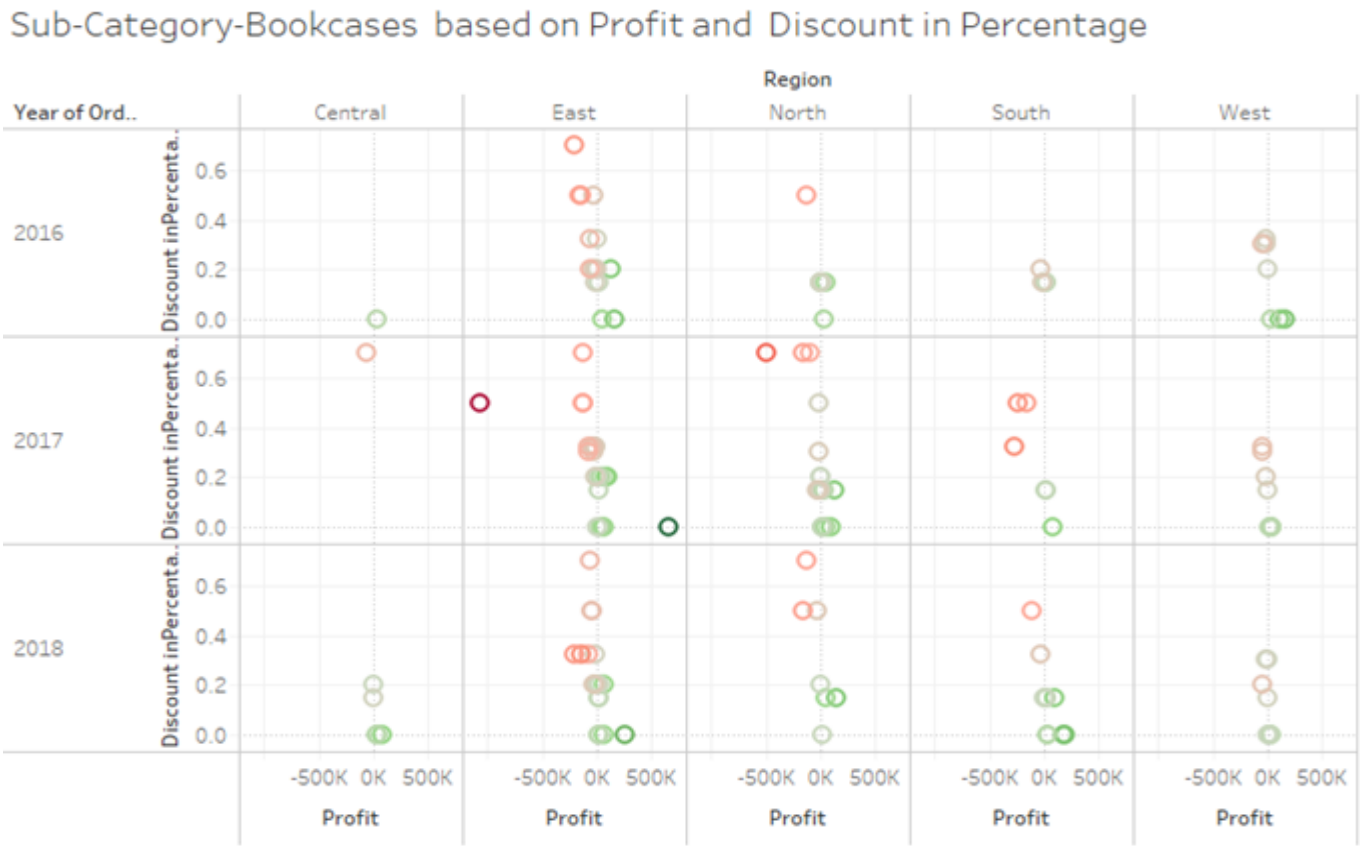
From the above graph we see that, in all regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in profit and hence making losses.

❖ For Bookcases



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by the sum of Profit. The data is filtered on Sub-Category, which keeps Bookcases.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Bookcases.

In the North and East region, we are facing losses in 2016,2017 and 2018 due to giving high discounts.

And in 2018 in the West, we get losses due to giving more discounts.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Bookcases.

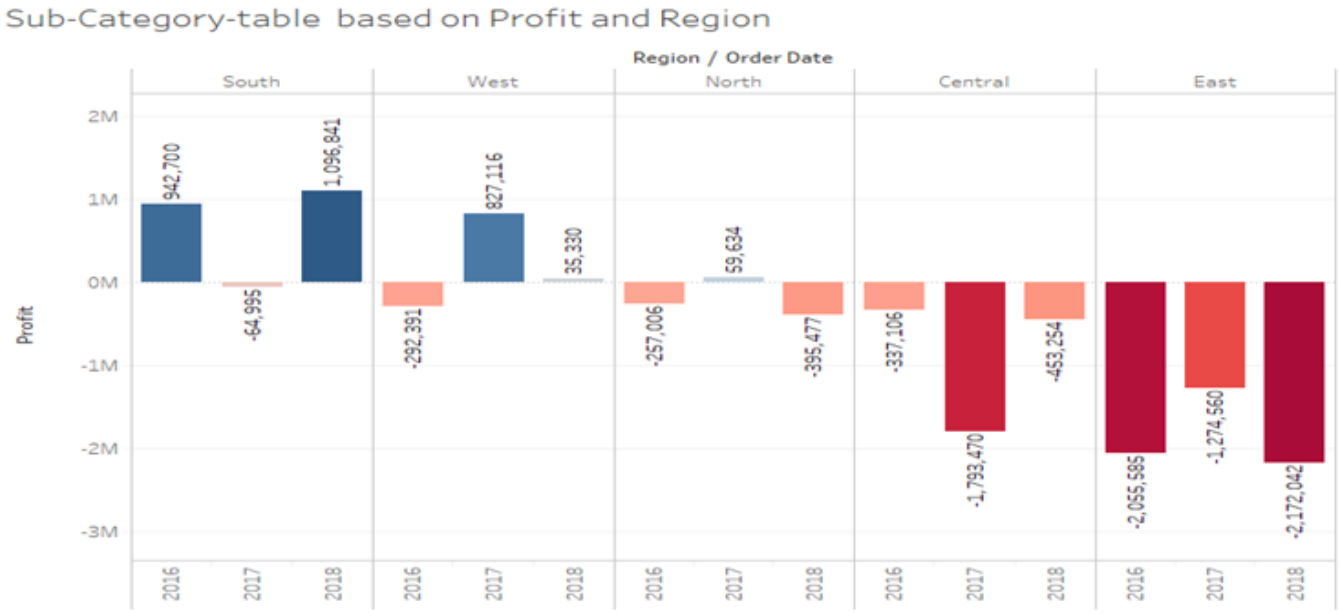
From above graph we see that,

For the East Region: increasing discount from 0 % to 15 %, our sales decrease and from 15 % to 20%, our sales increase leads to making profit.

For the North Region: increasing discounts from 0 % to 15 %, our sales increases and making profit.

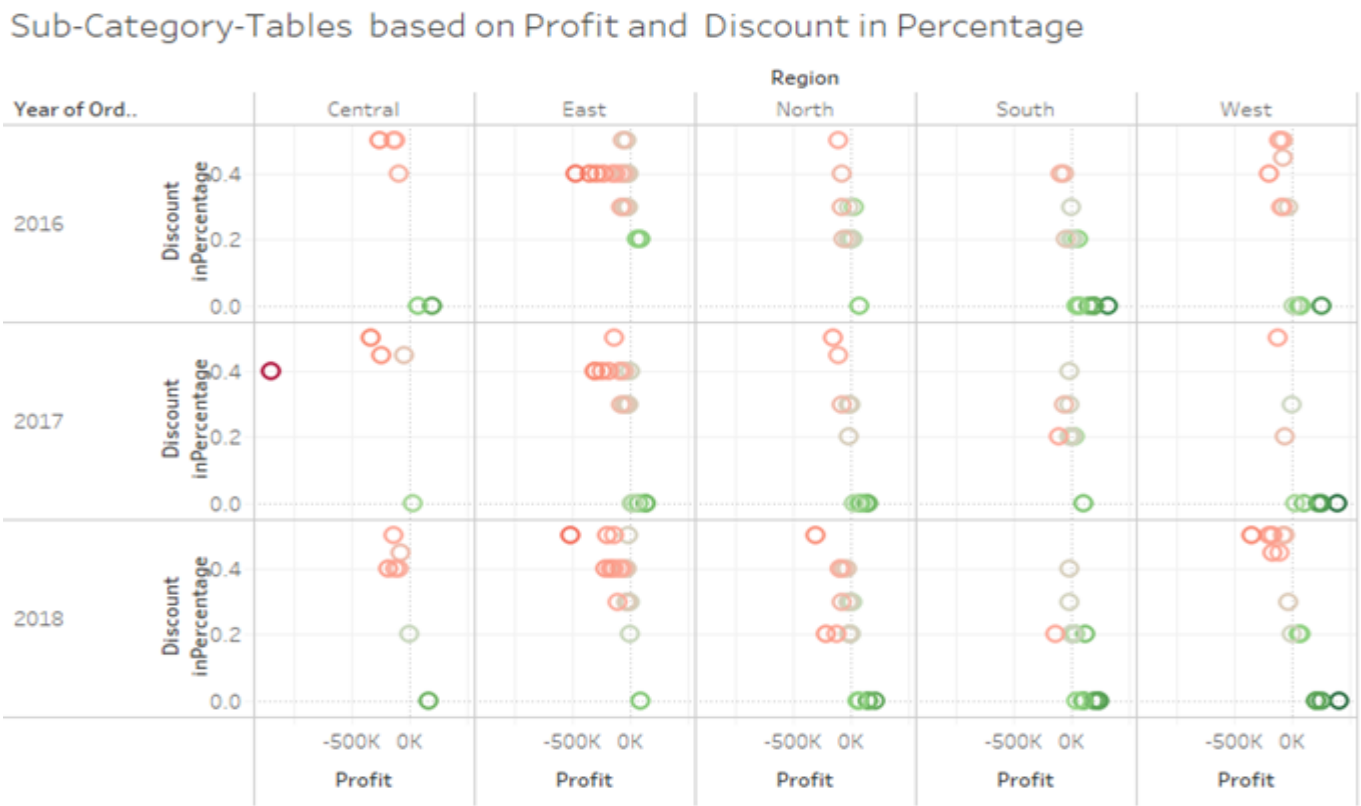
For the West Region: without discounts our sales are more and making profit, as we increase discounts leads to failing profit.

❖ For Tables



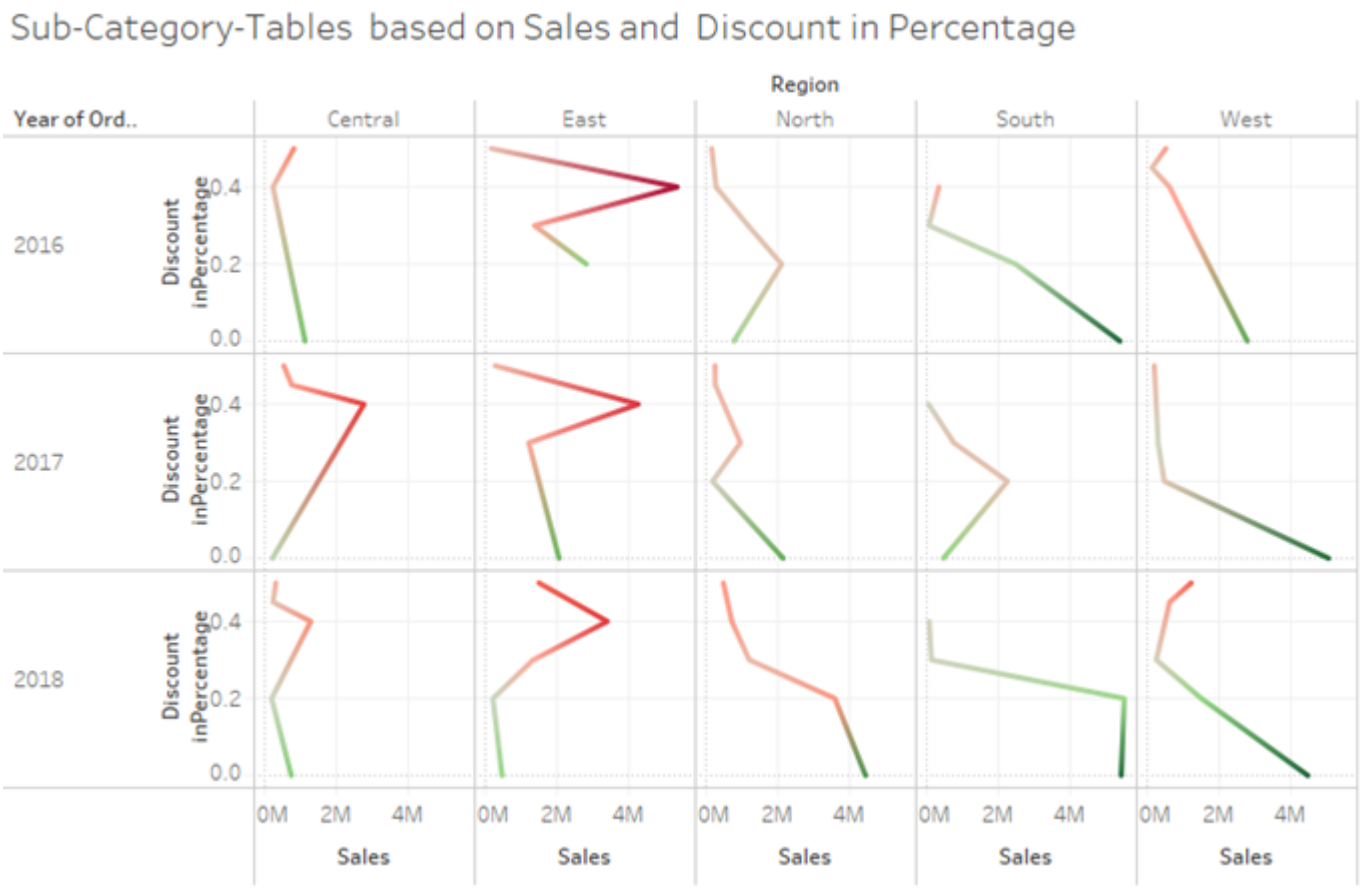
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Tables.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Tables.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Tables.

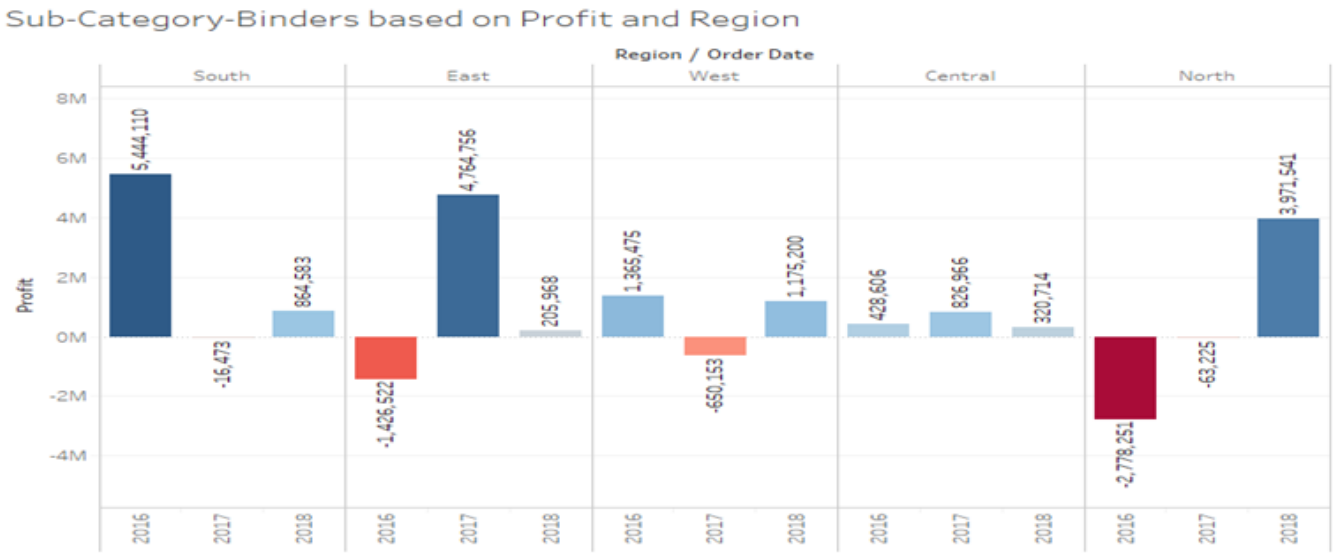
From above graph we see that,

For Central and East Region: increasing discount from 20% to 40% our sales are increased but it makes more loss.

For the North and West Region: without a discount our sales are more and making profit, as we increase the discount leads to failing profit.

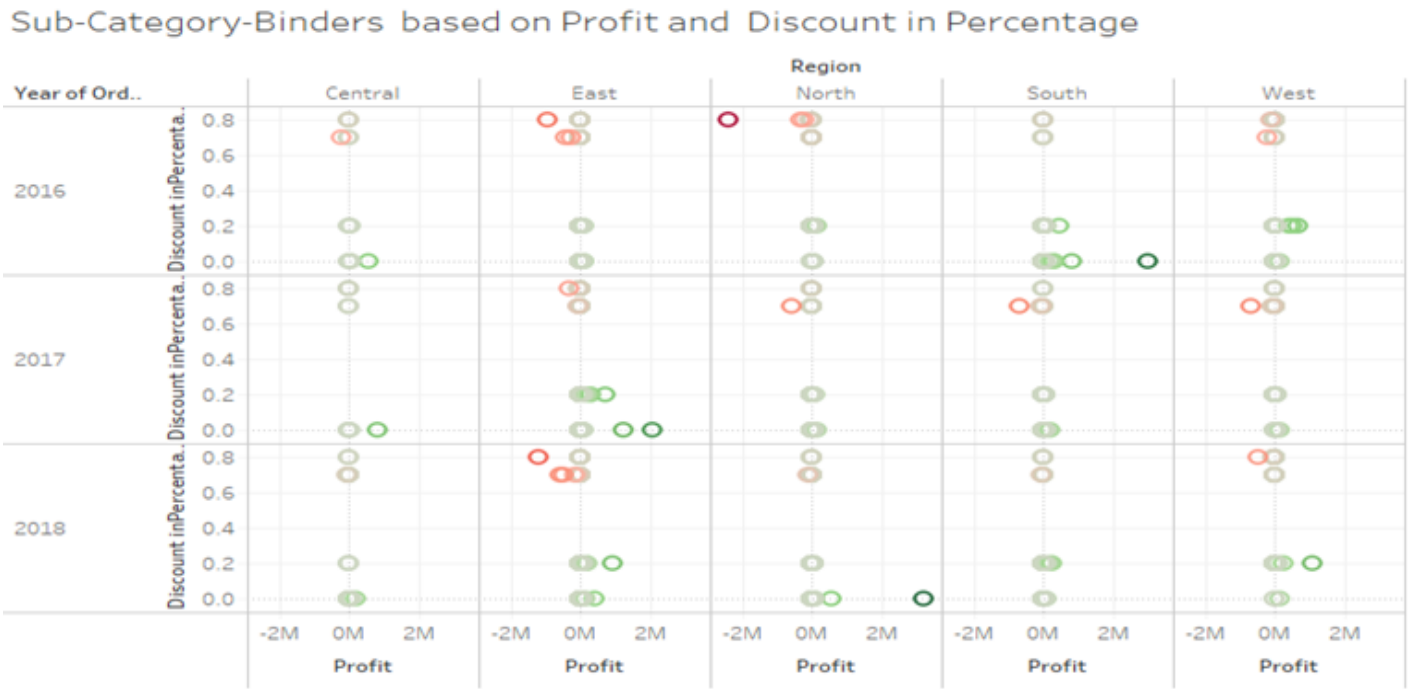
Office Supplies: Binders, paper, storage, Appliances, Envelopes, Art, Labels, Fasteners and Supplies

❖ For Binders



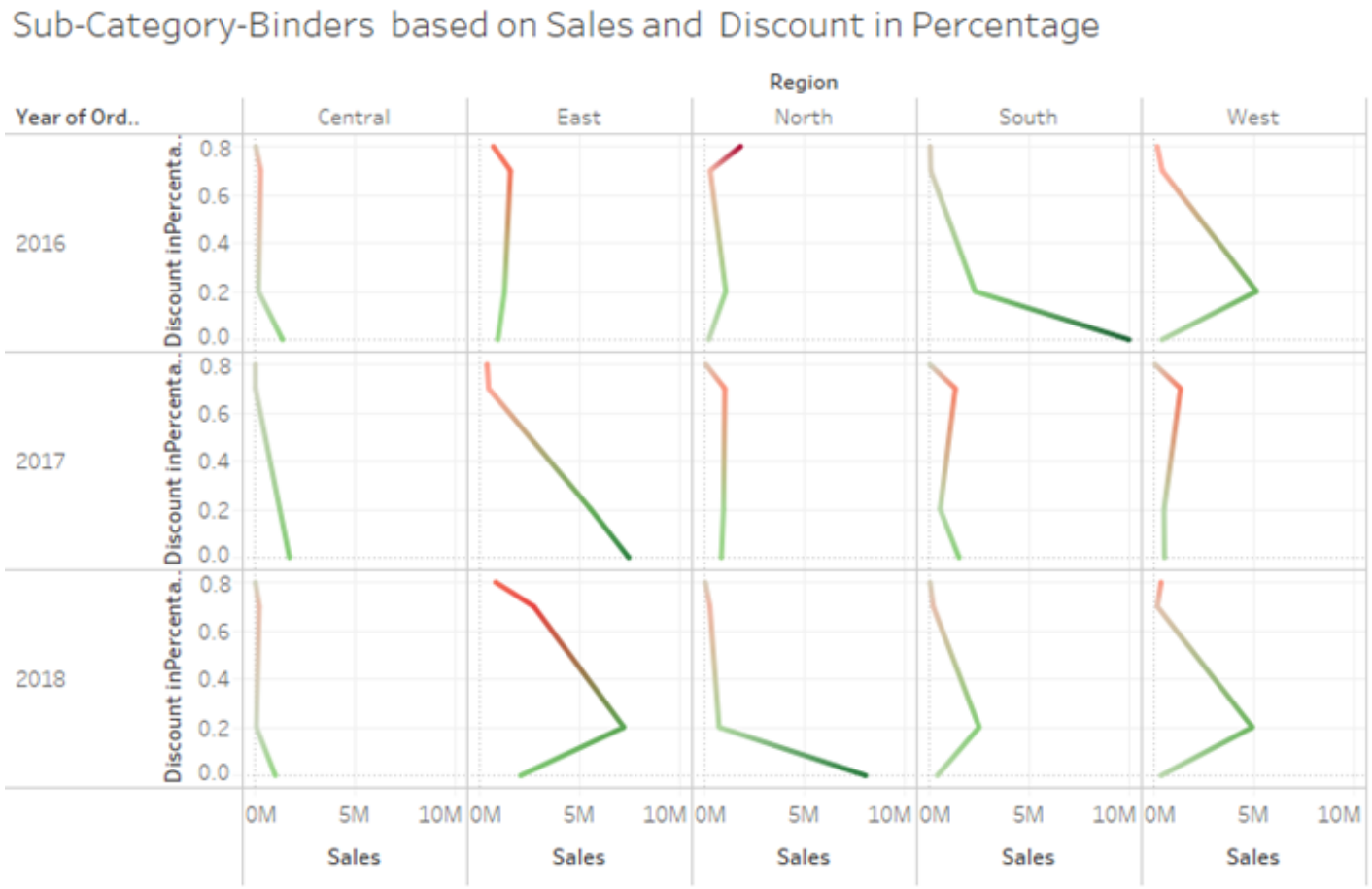
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Binders.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Binders.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Binders.

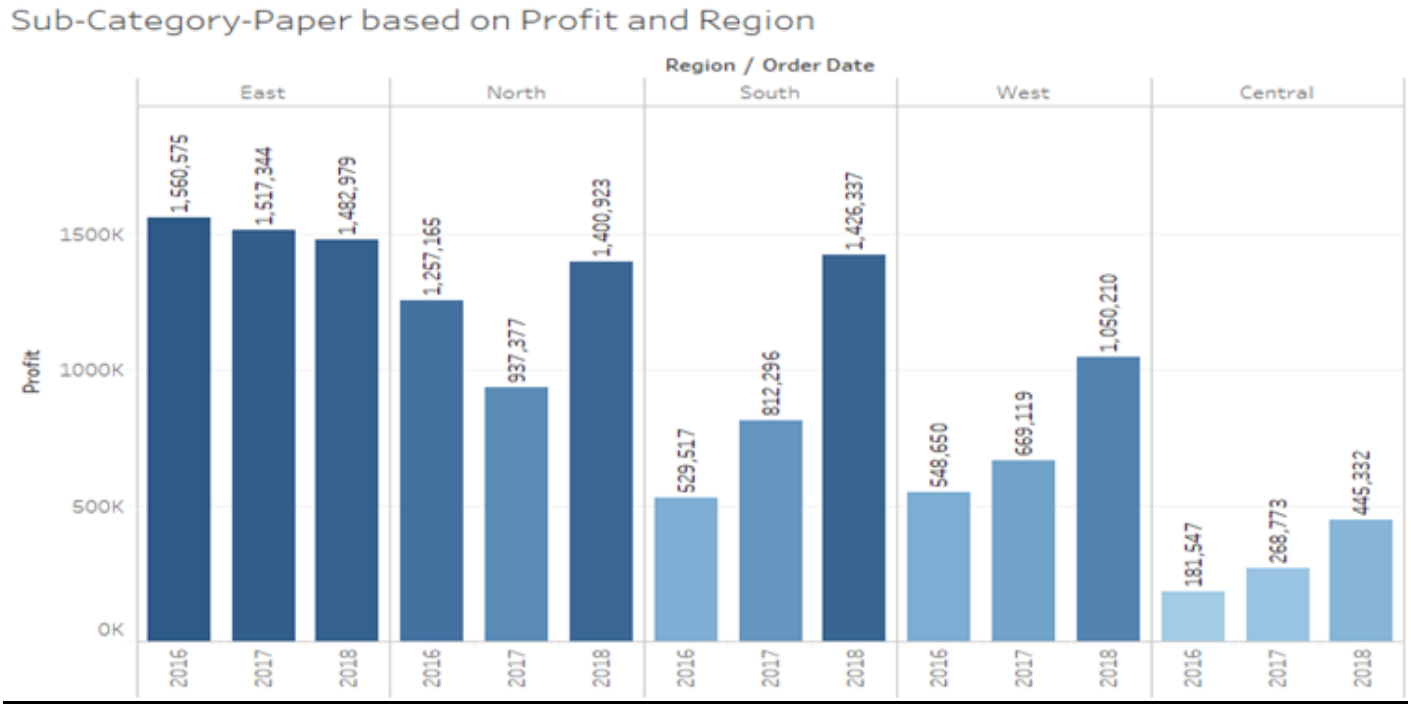
From above graph we see that,

For the West Region: with 20% discounts our sales are increasing and making profit, as we increase discounts leads to failing profit.

For East Region: increasing discount from 0 % to 20 %, our sales increases for 2018 and making profit.

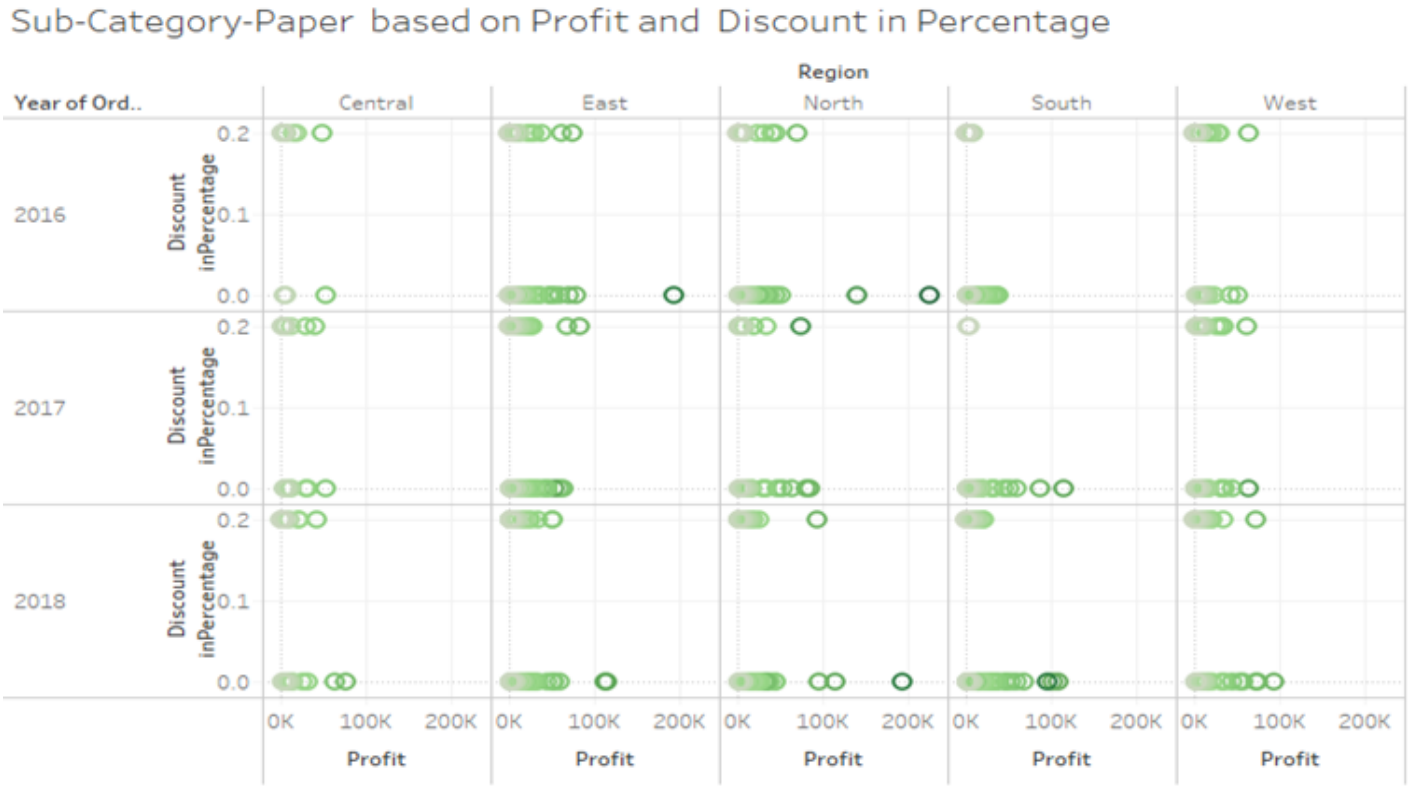
For the North Region: without discounts our sales are more and making profit, as we increase discounts leads to failing profit.

❖ For Paper



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Paper.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Paper.

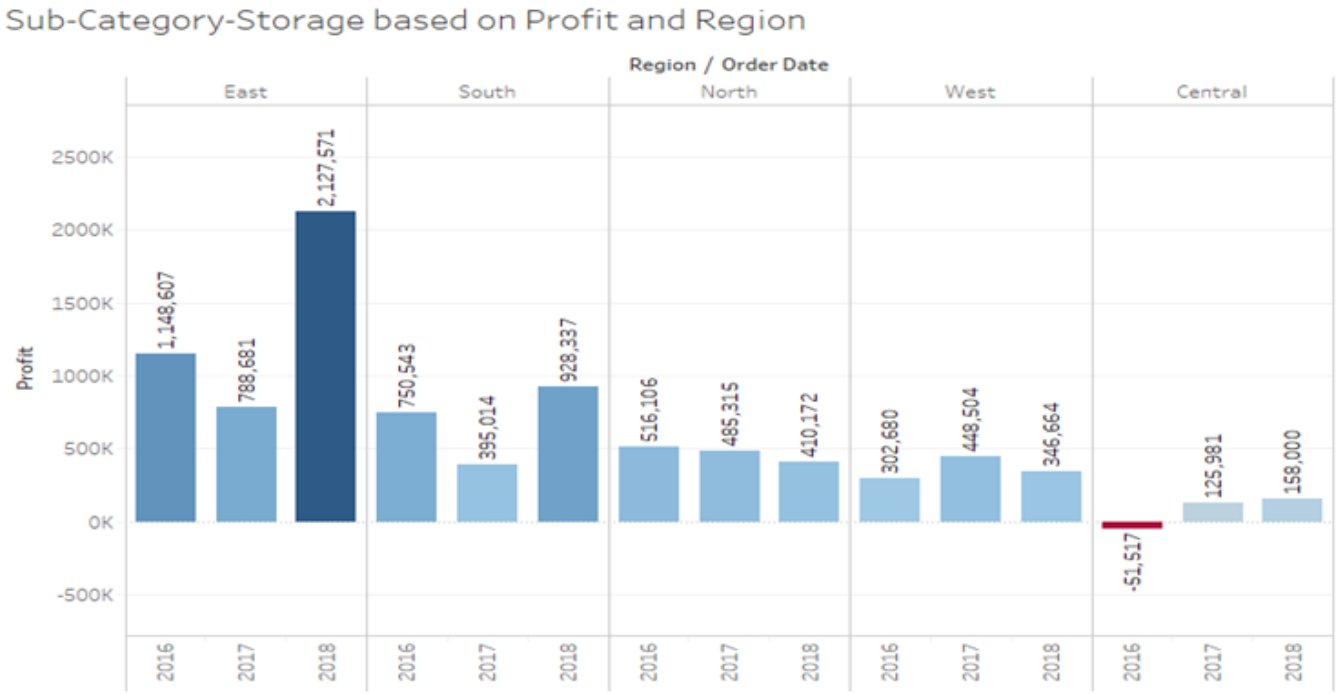
Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Paper.

From above graph we see that,

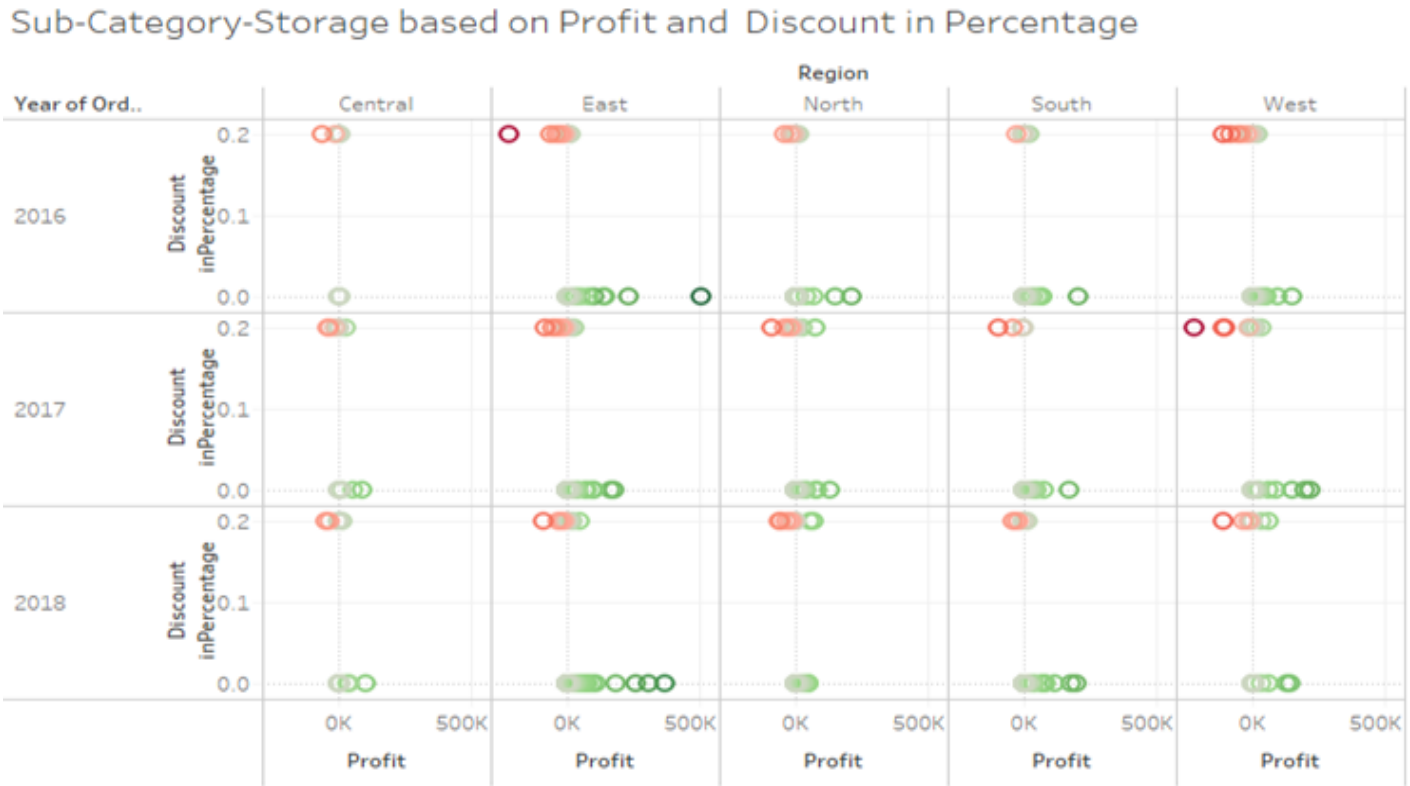
In all regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and hence minimizing profit.

❖ For Storage



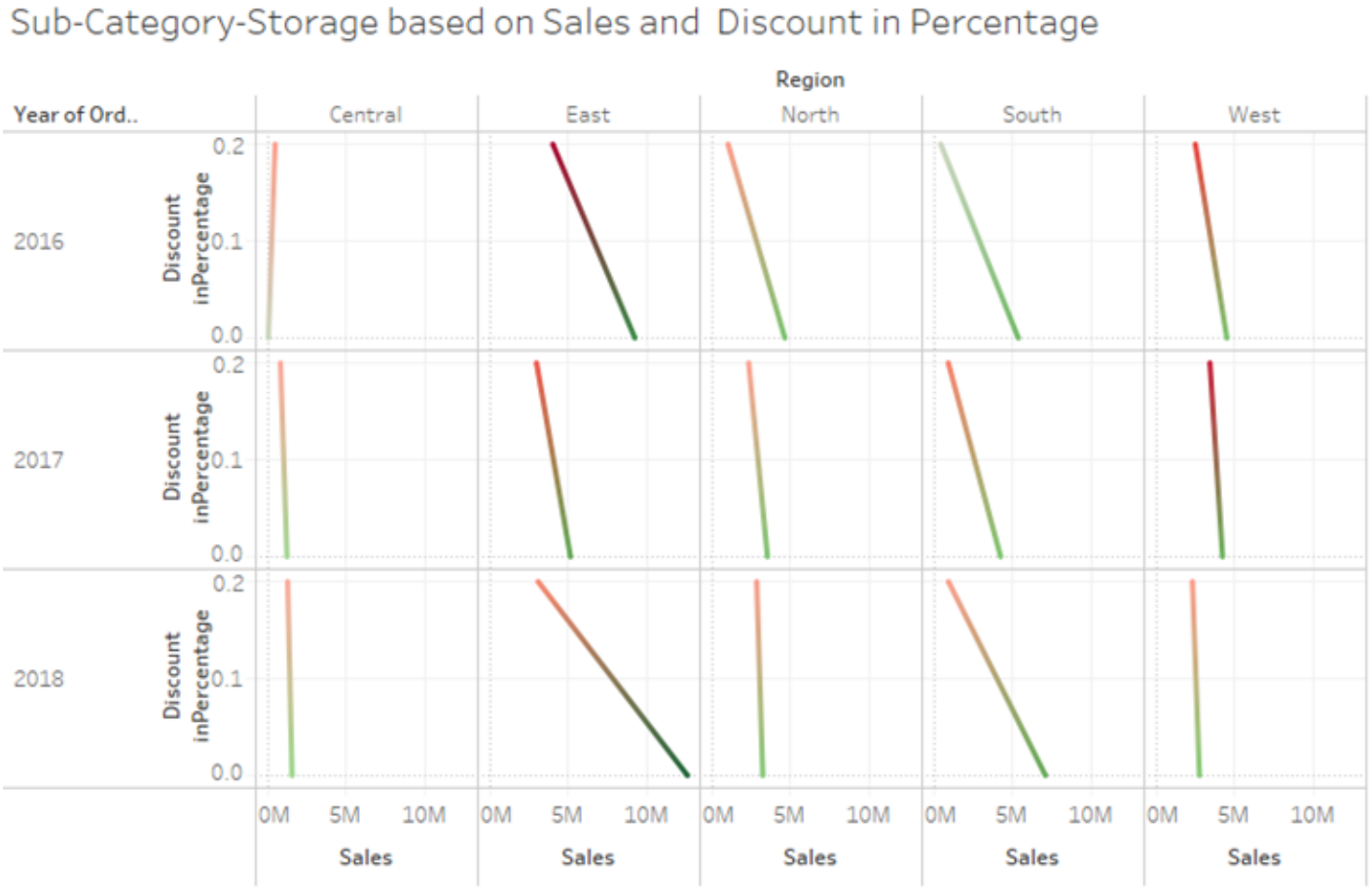
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Storage.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Storage.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

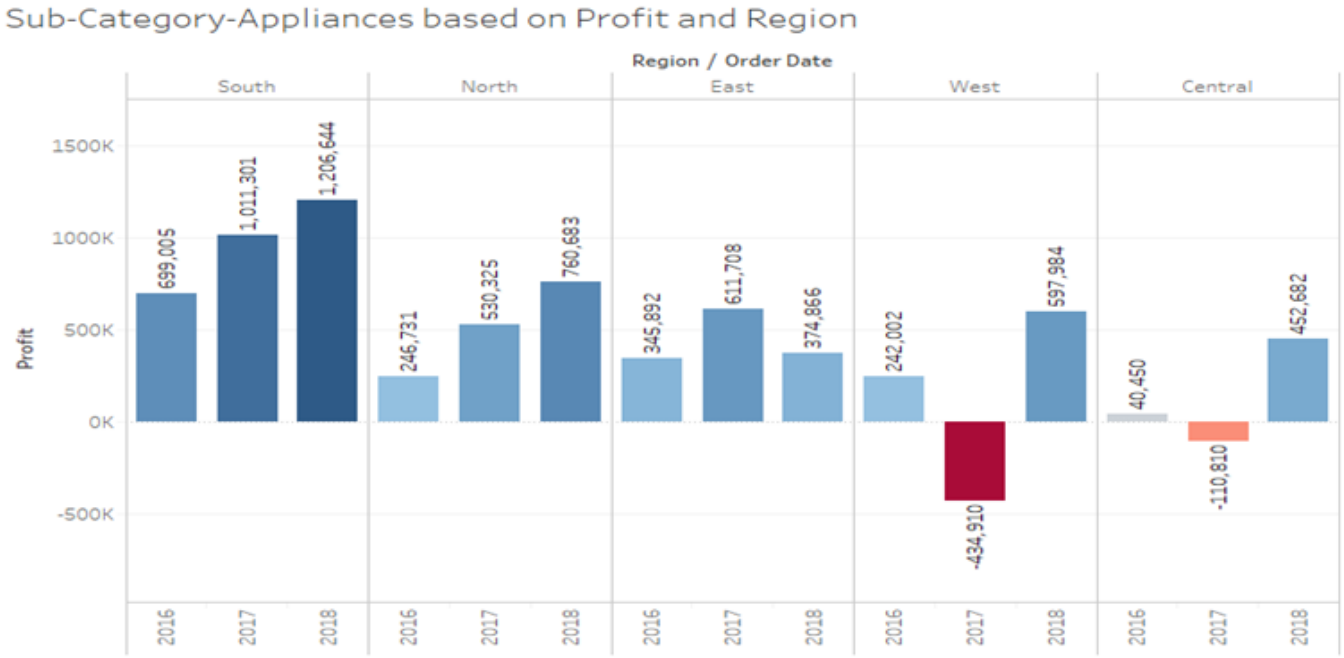


Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Storage.

From above graph we see that,

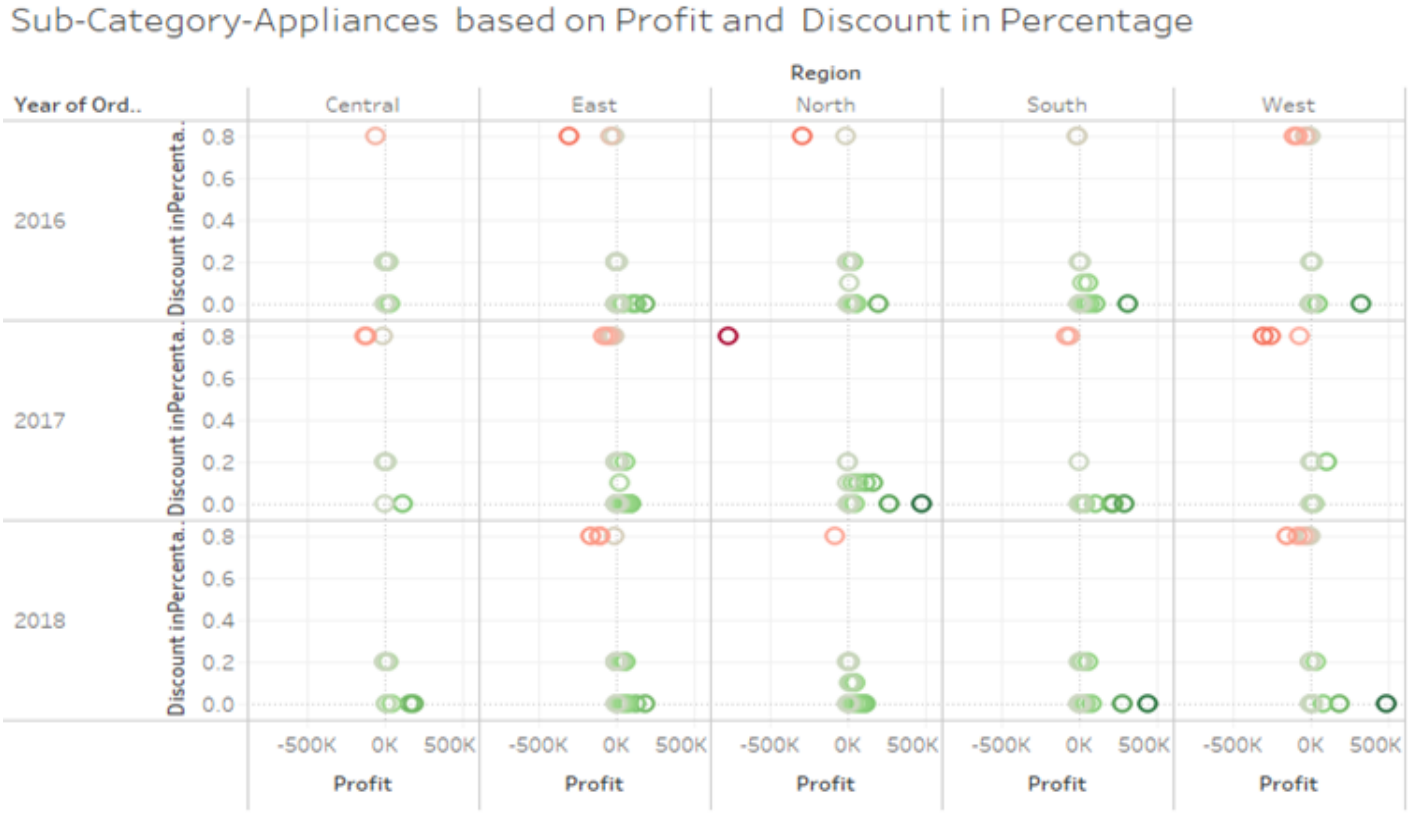
In all regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and decreases in profit & hence making losses.

❖ For Appliances



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Appliances.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Appliances.

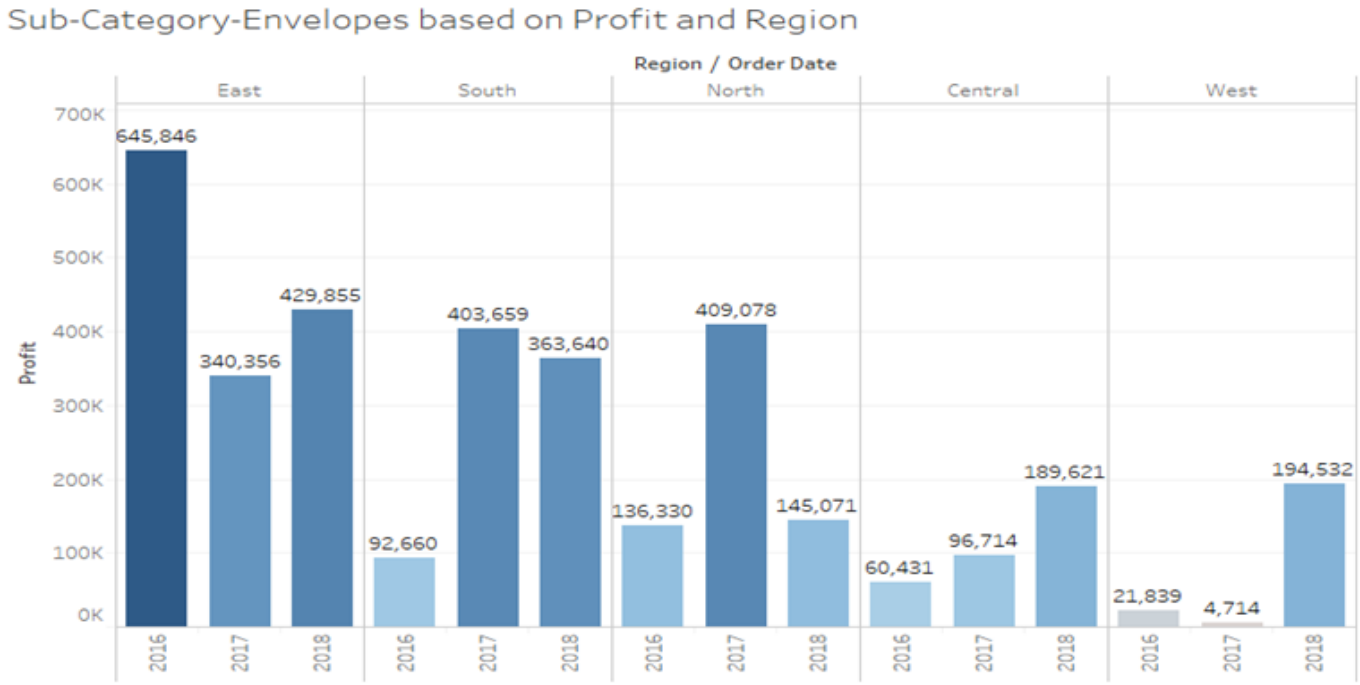
Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Appliances.

From above graph we see that,

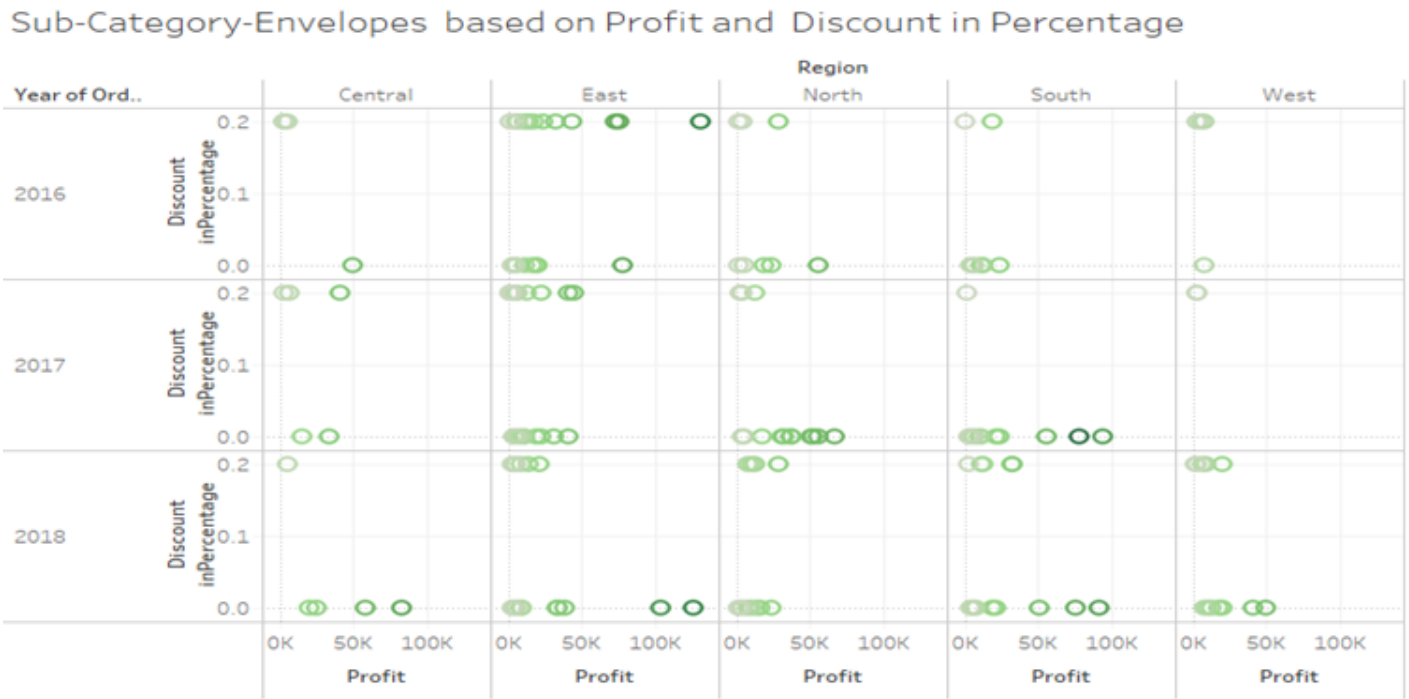
In all regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and decreases in profit & hence making losses.

❖ For Envelopes



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Envelopes.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Envelopes.

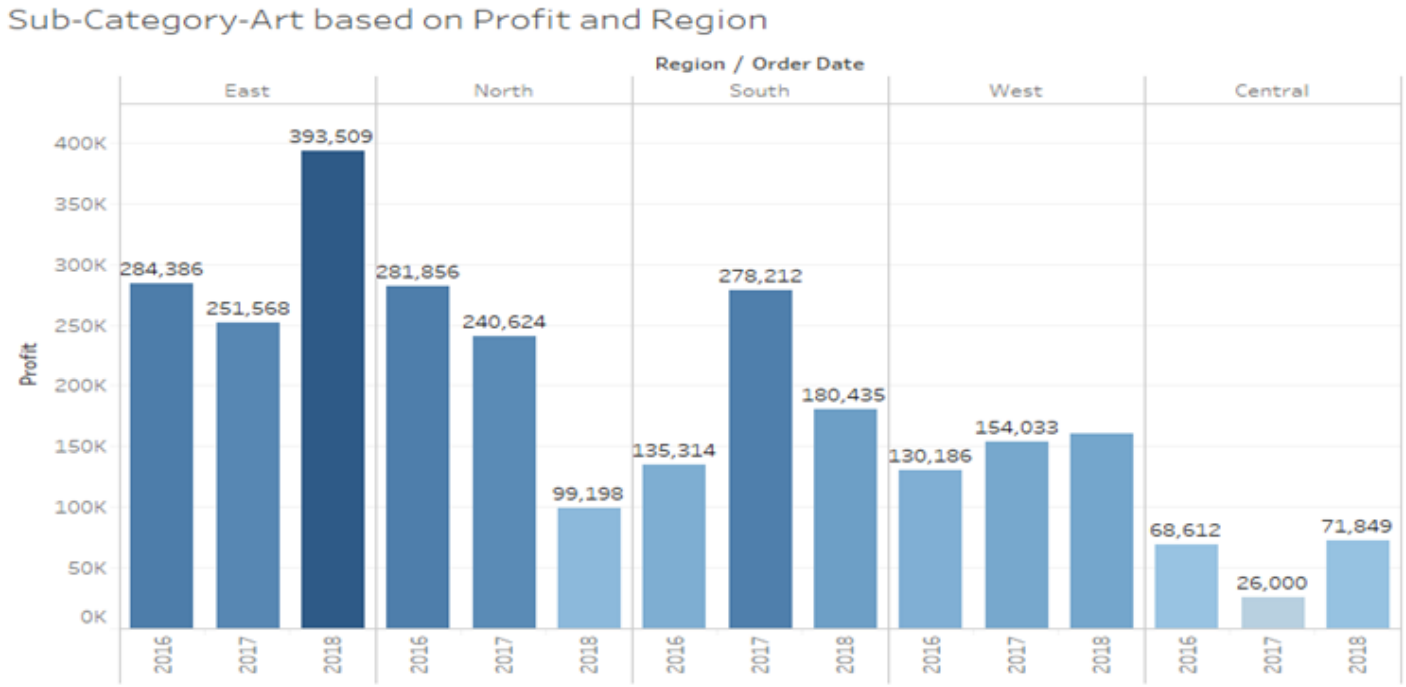
Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Envelopes.

From above graph we see that,

For Central and East Region: without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and hence minimizing profit for 2018.

❖ For Art

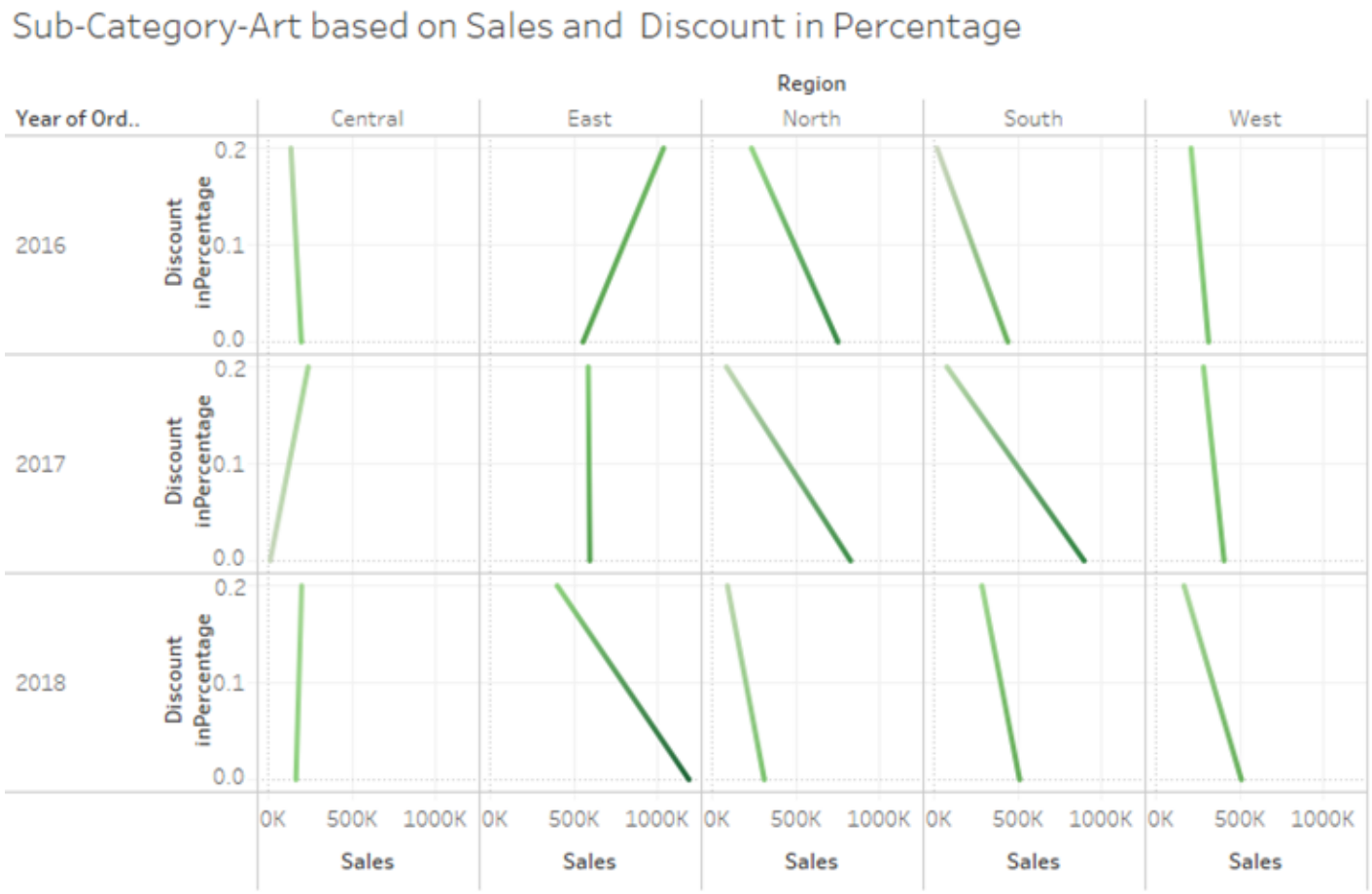


Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Art.

Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Art.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

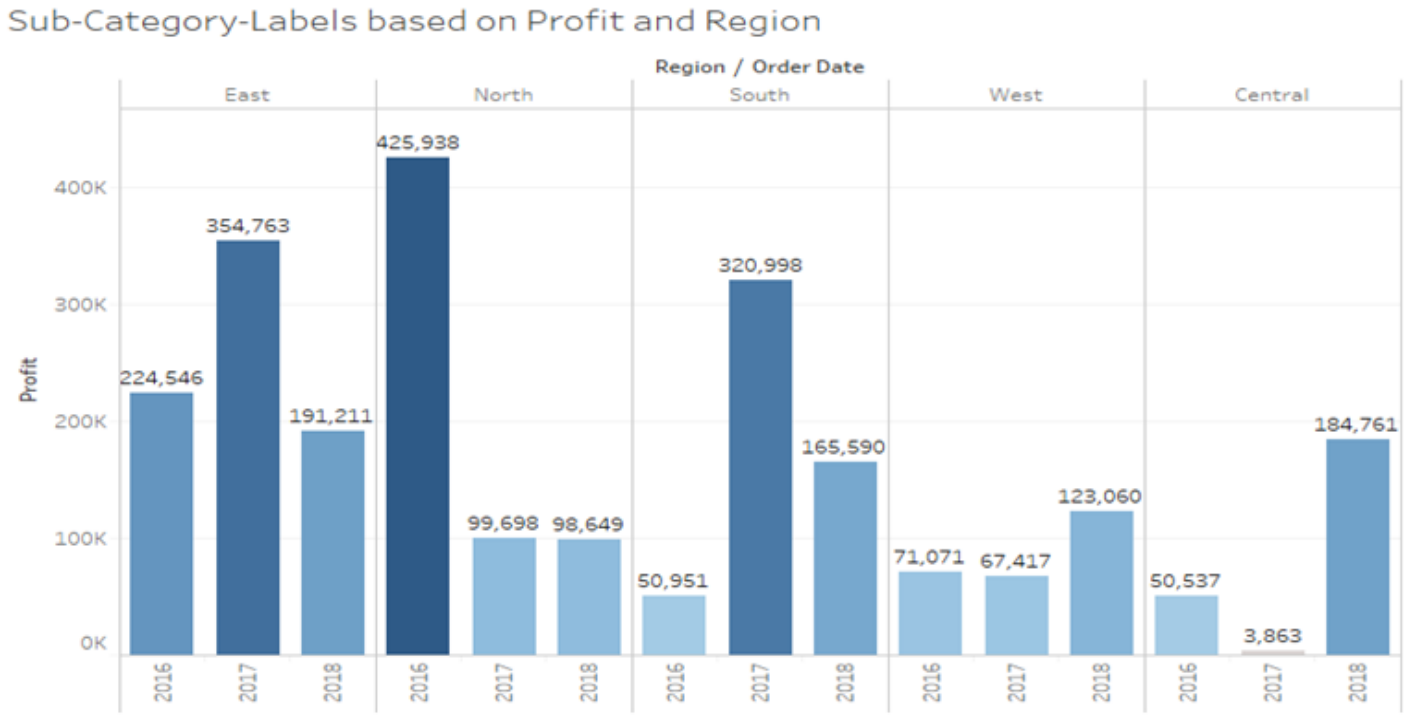


Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Art.

From above graph we see that,

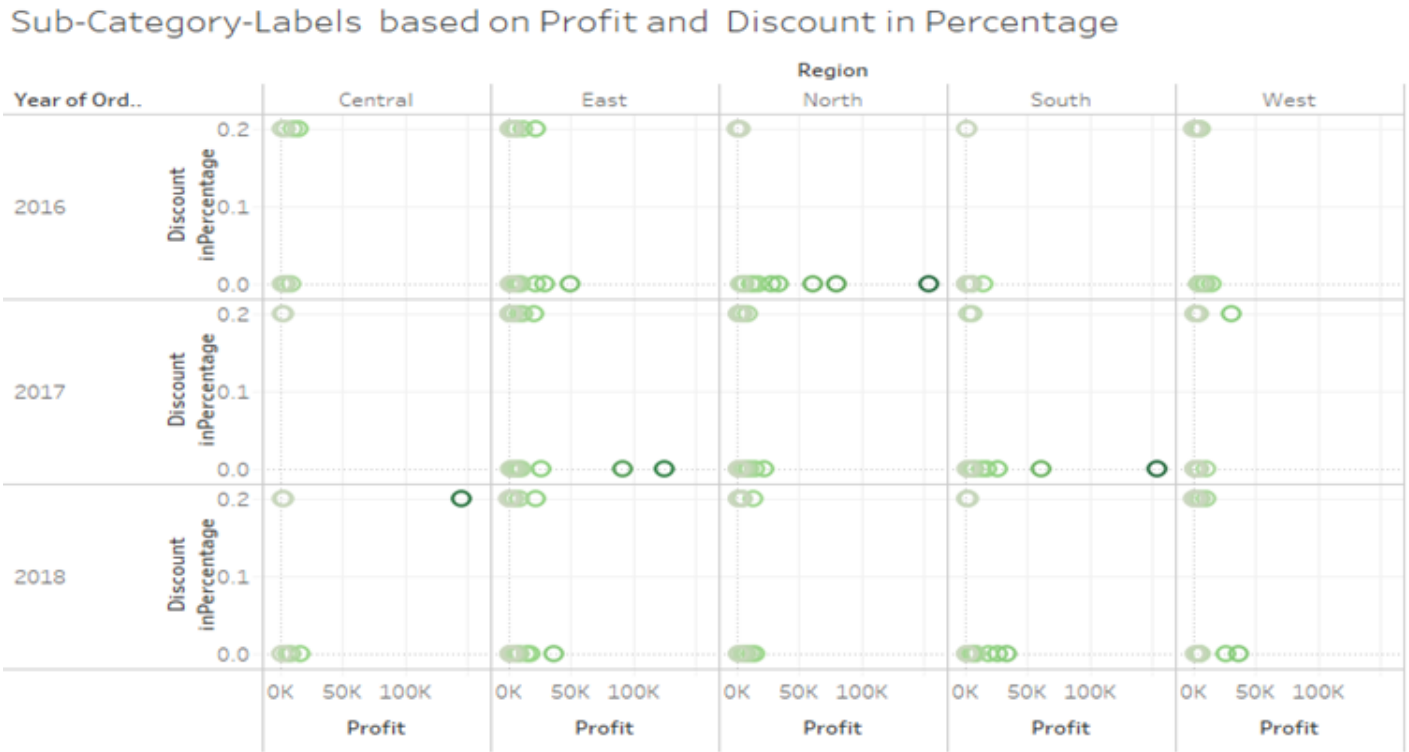
For East Region: without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and hence minimizing profit for 2018.

❖ For Labels



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Labels.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Labels.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

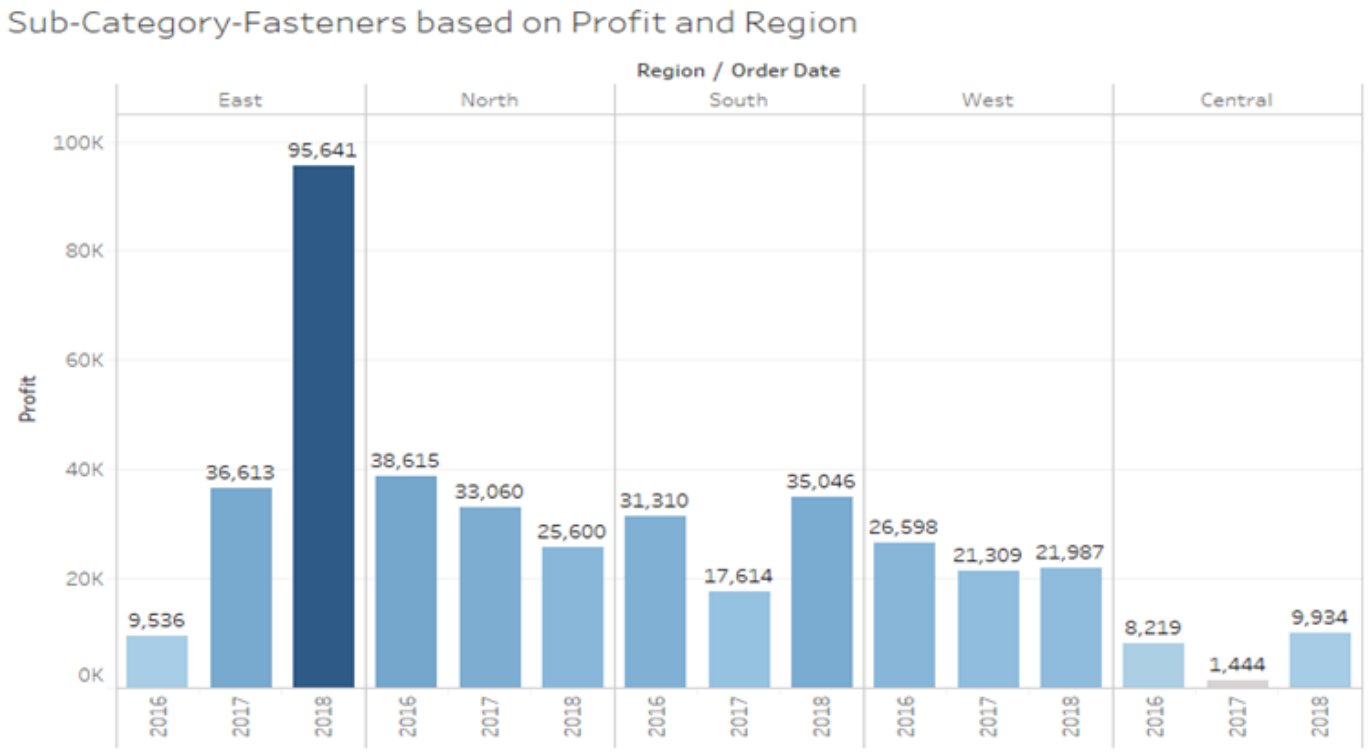
Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Labels.

From above graph we see that,

For the Central Region: as we increase discounts from 0 % to 20 %, our sales also increase and it leads to making profit in 2018.

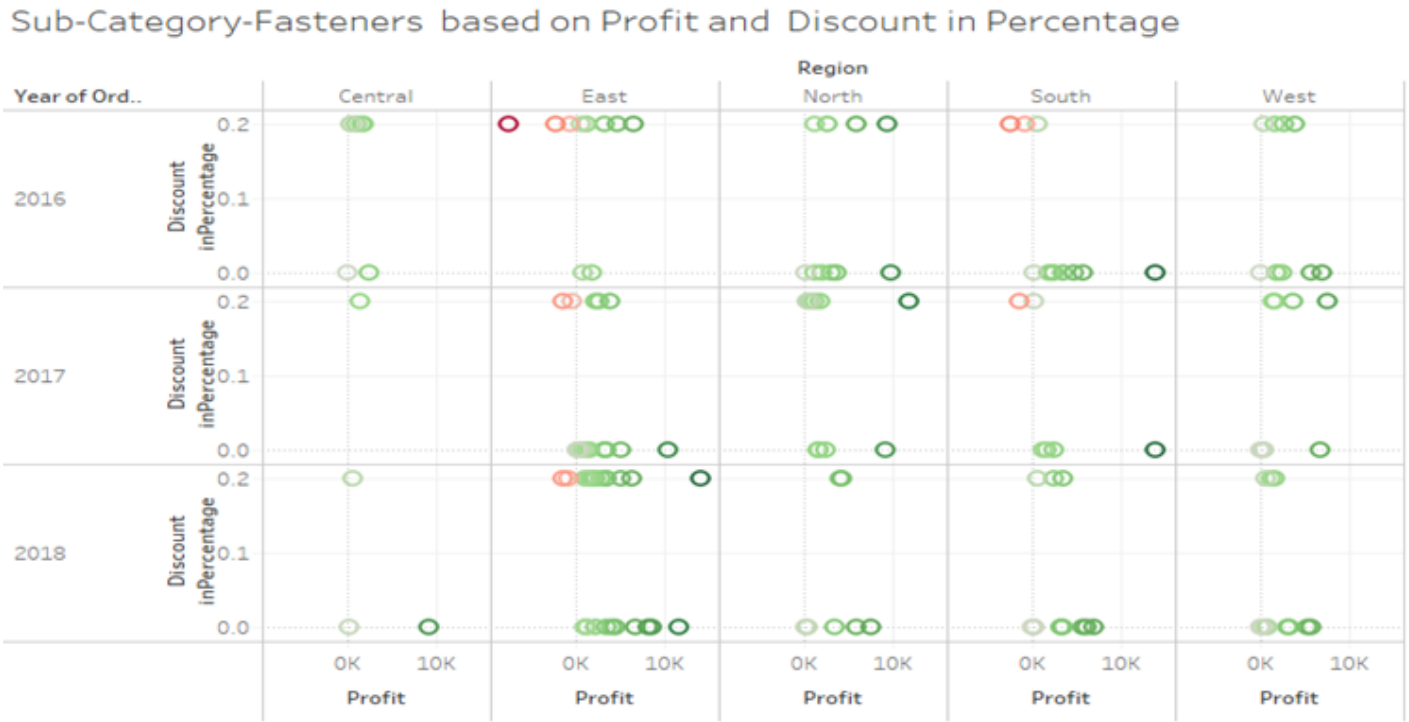
In all other regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and minimizing profit.

❖ For Fasteners



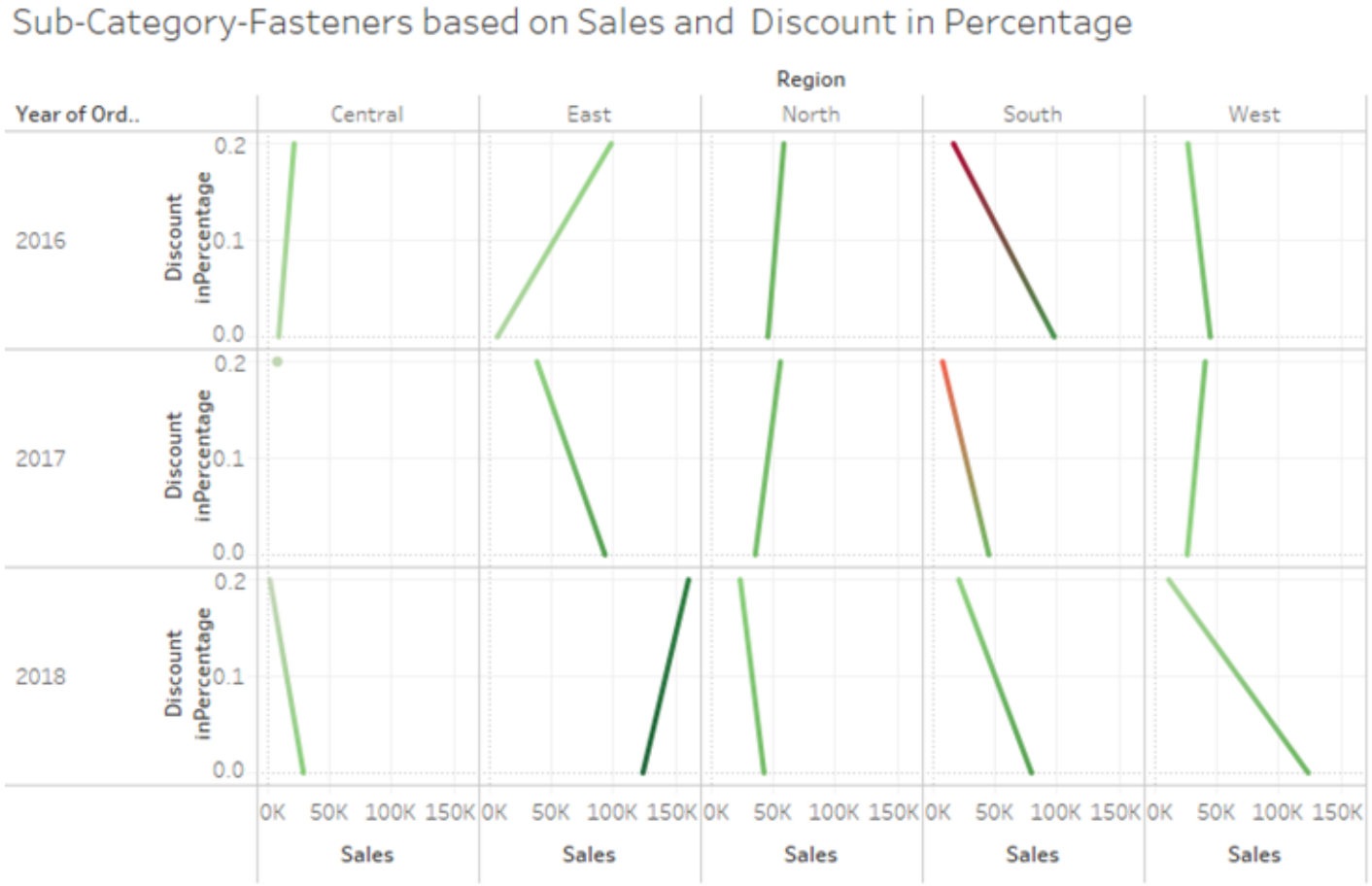
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Fasteners.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Fasteners.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Fasteners.

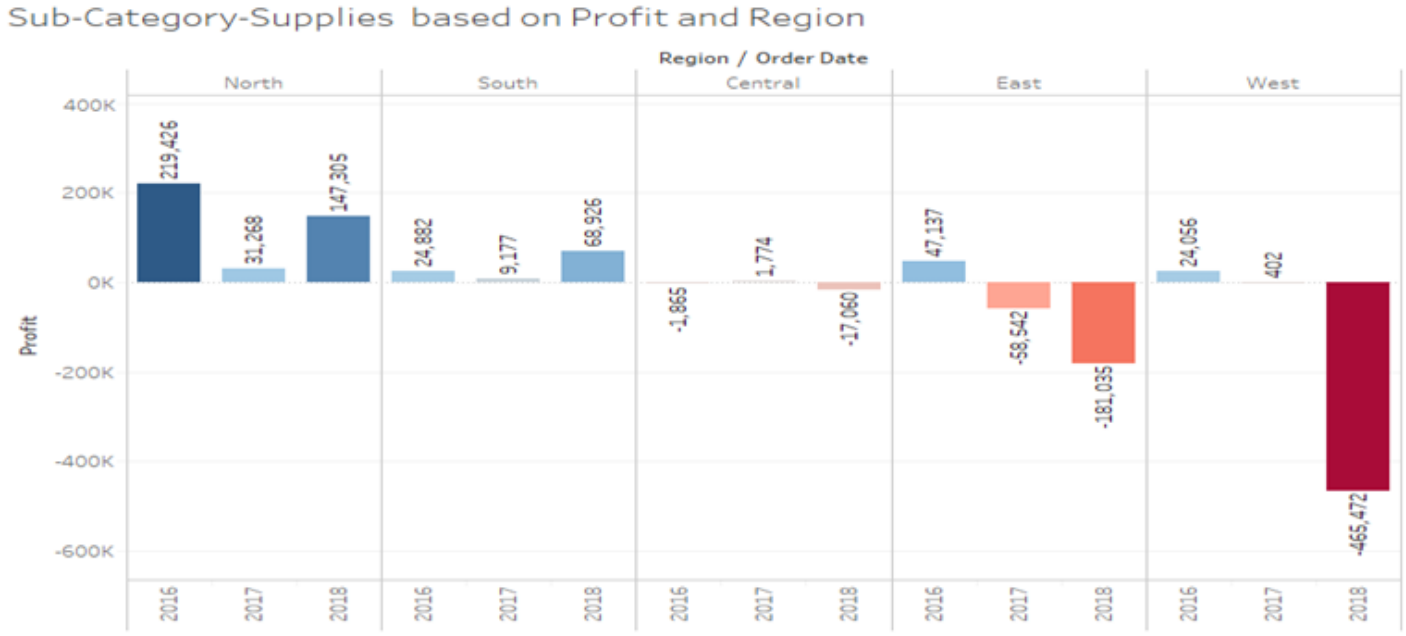
From above graph we see that,

For the East Region: increasing discount from 0 % to 20 %, our sales increases and making profit in 2016 & 2018.

In all other regions without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and minimizing profit.

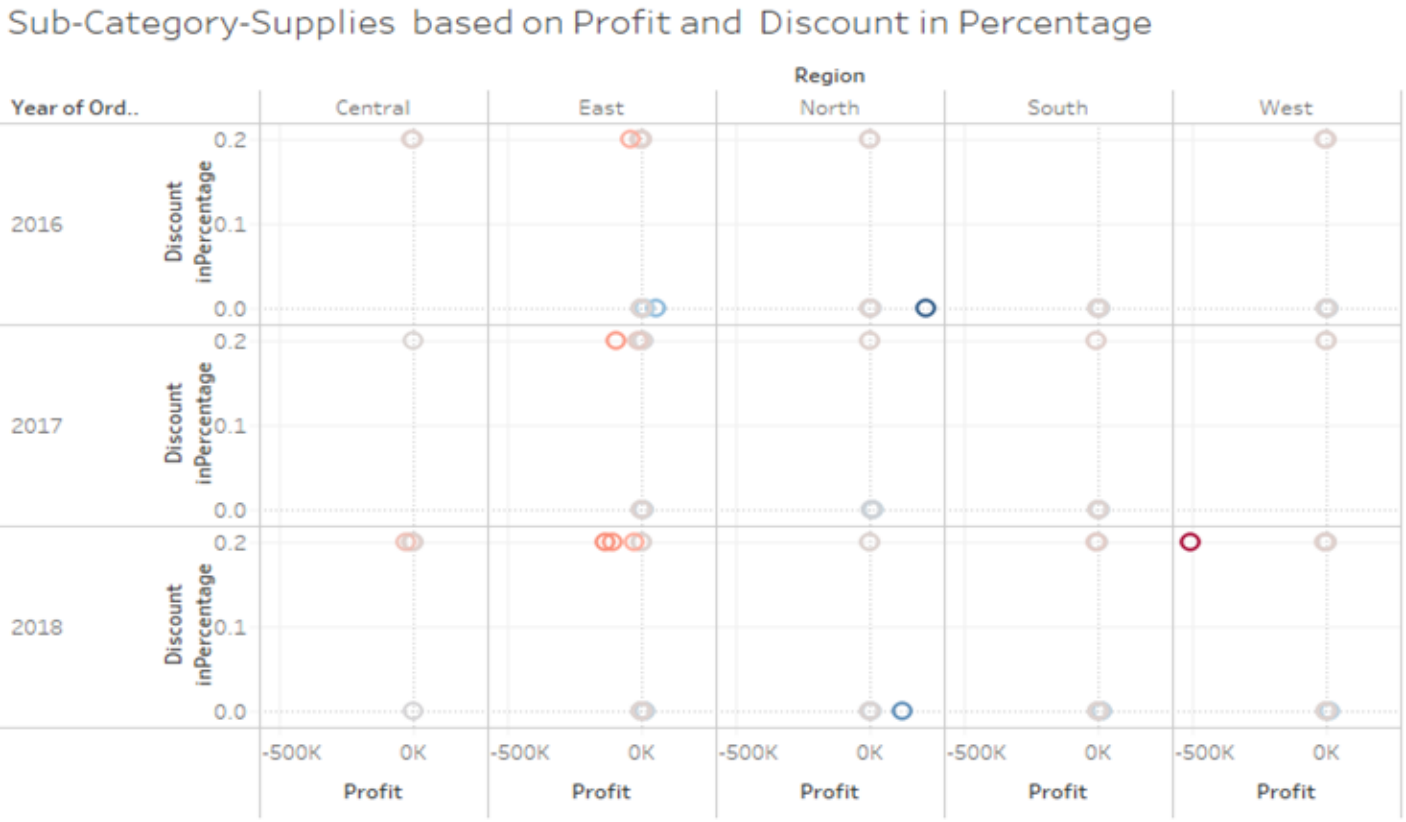
For the South Region: increases in discount leads to decreases in profit and making losses.

❖ For Supplies



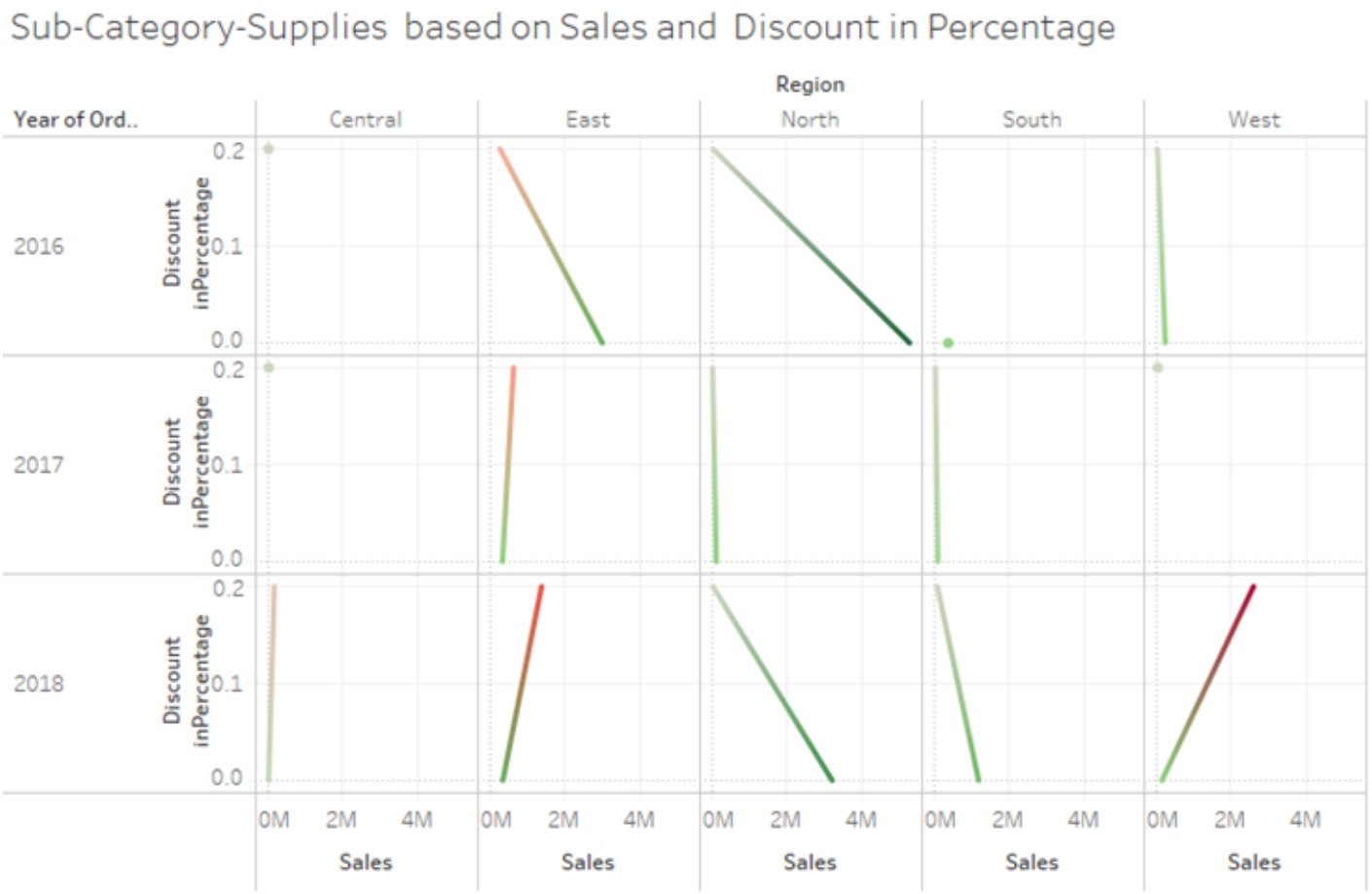
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Supplies.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Supplies.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Supplies.

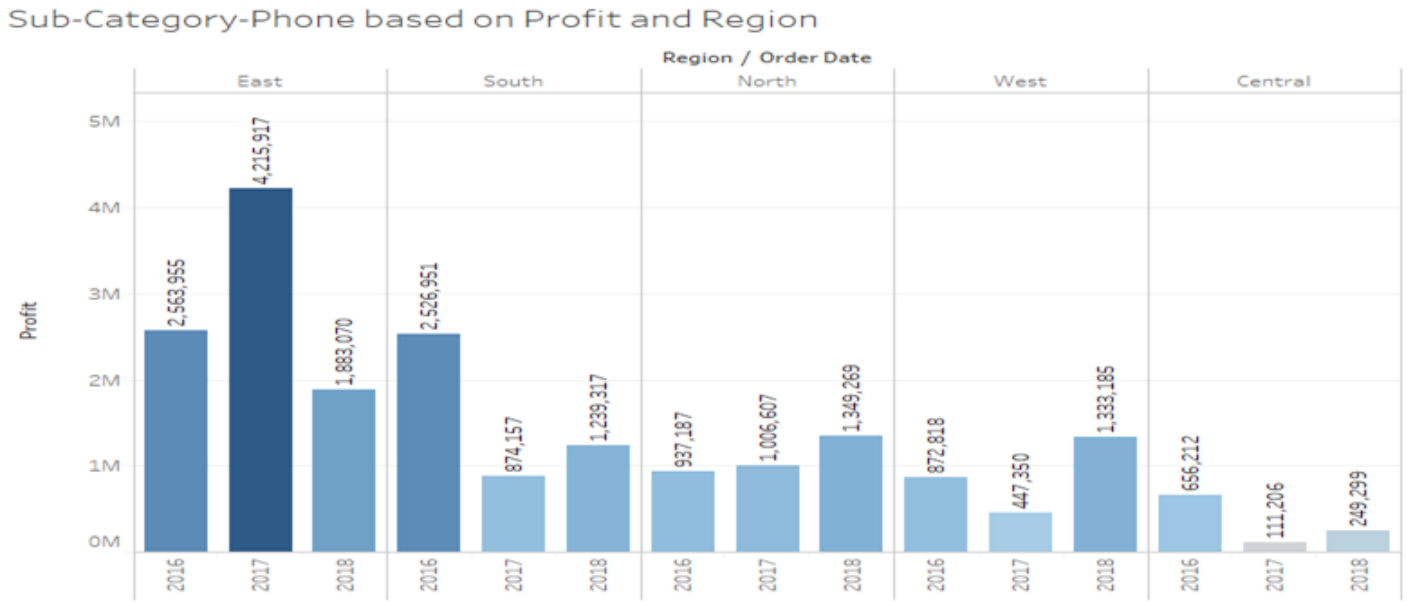
From above graph we see that,

For East and West Region: increasing discounts from 0 % to 20 %, our sales increases and leads to failing profit.

For the North and South Region: without discount there is more sales and it's also making profit, but as increases in discount leads to decreases in sales and minimizing profit.

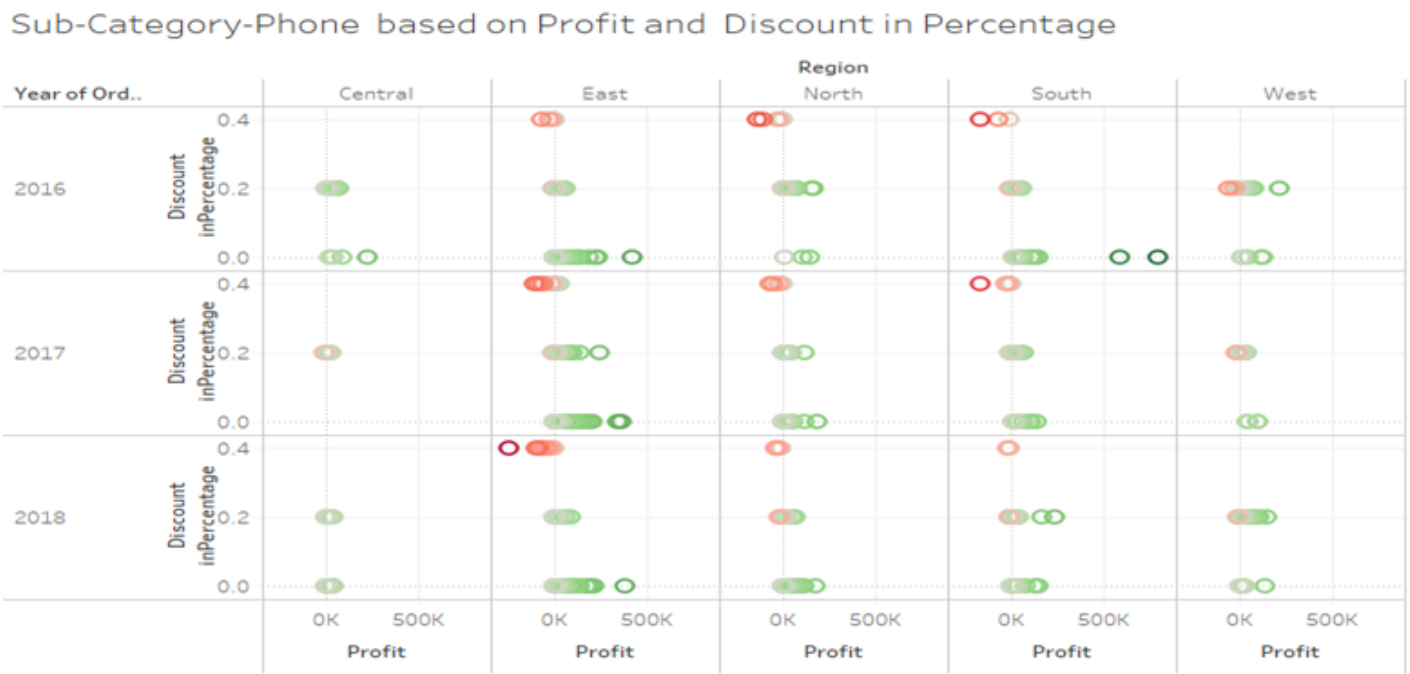
Technology: Phone, Copies, Accessories and Machines

❖ For Phone



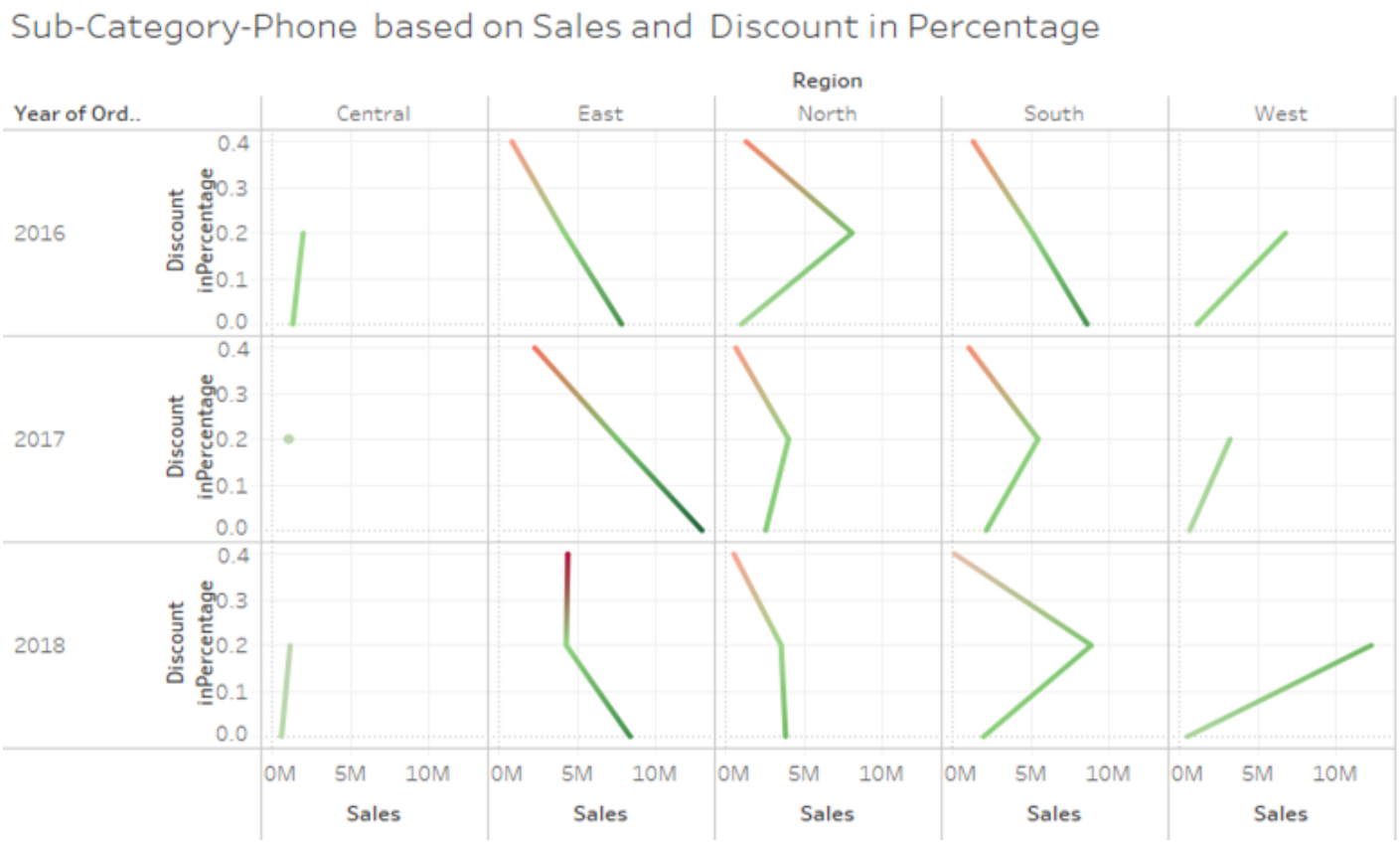
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by the sum of Profit. The data is filtered on Sub-Category, which keeps Phone.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Phone.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Phone.

From above graph we see that,

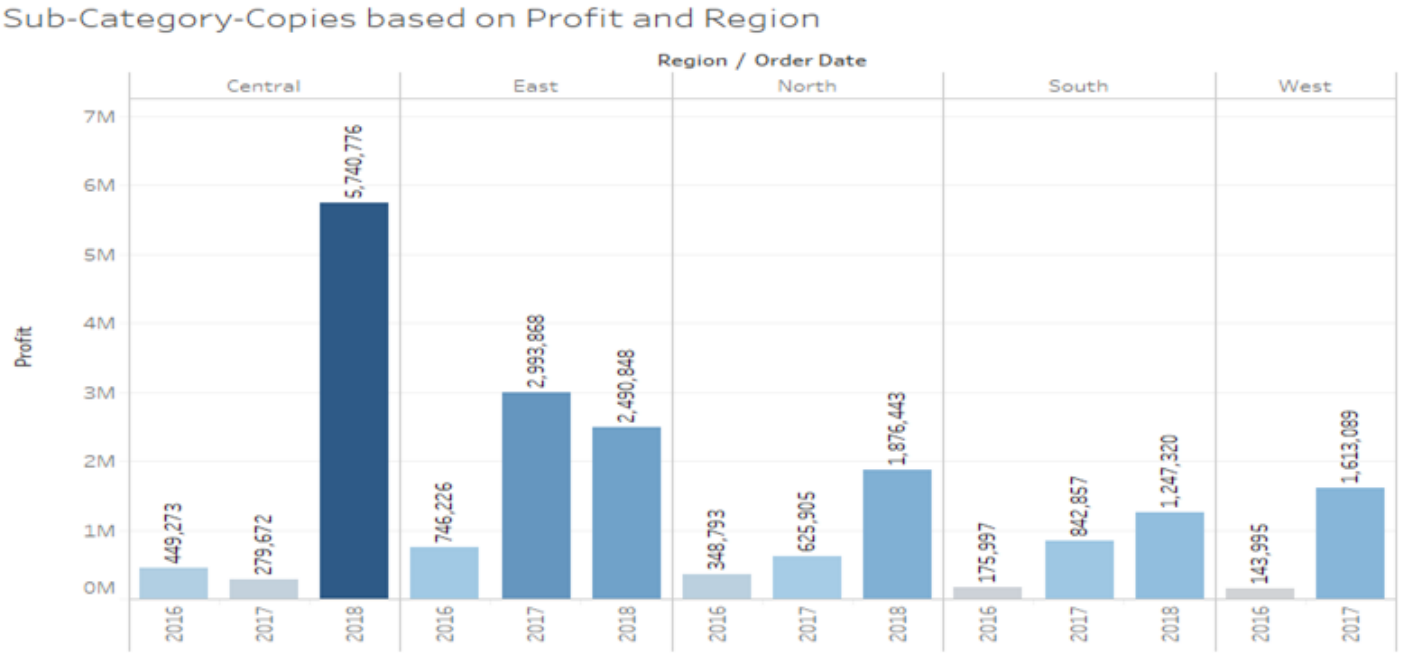
For the East Region: without discount our sales are more and making profit, increasing discount our sales is decreasing and it makes losses.

For the West Region: with increasing discounts our sales are increasing and making profit.

For the South Region: increasing the discount from 0% to 20% our sales increases year by year and it makes profit.

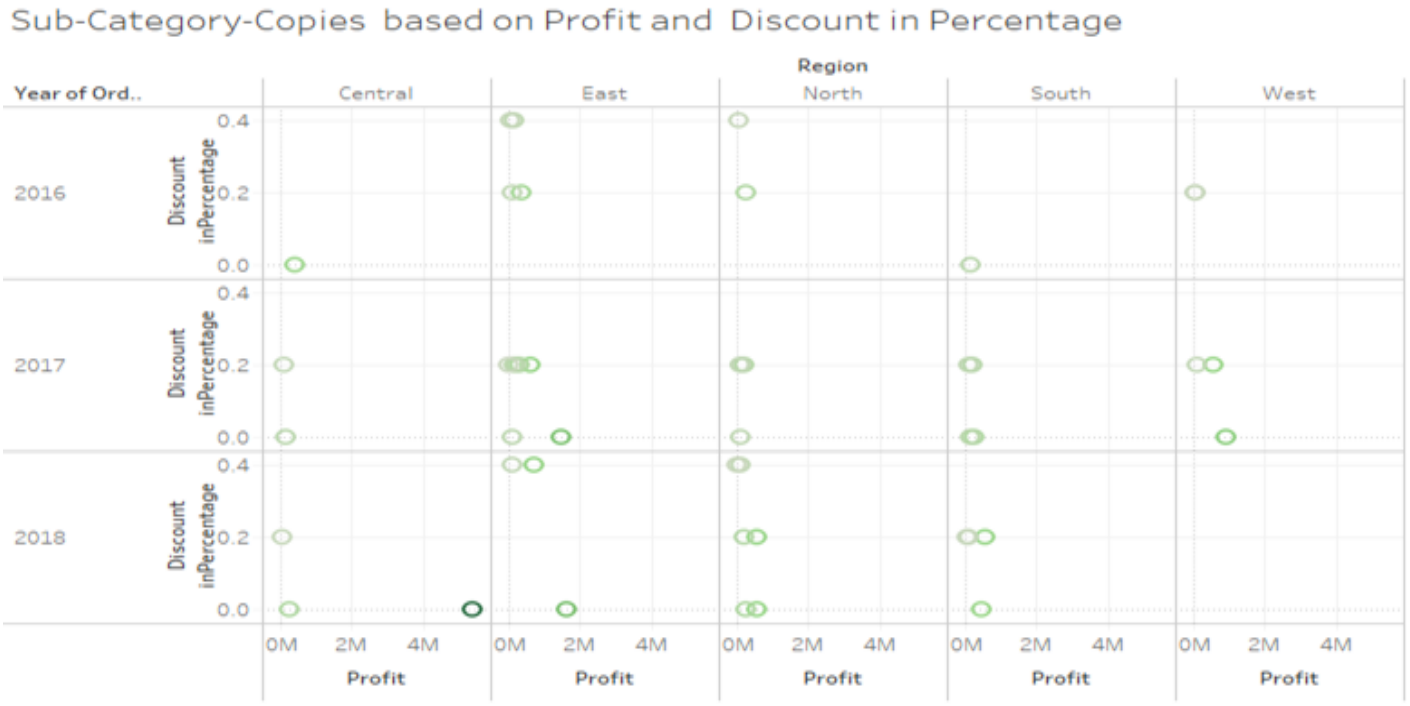
For the North Region: increasing discount from 0% to 20% our sales decrease year by year.

❖ For Copies



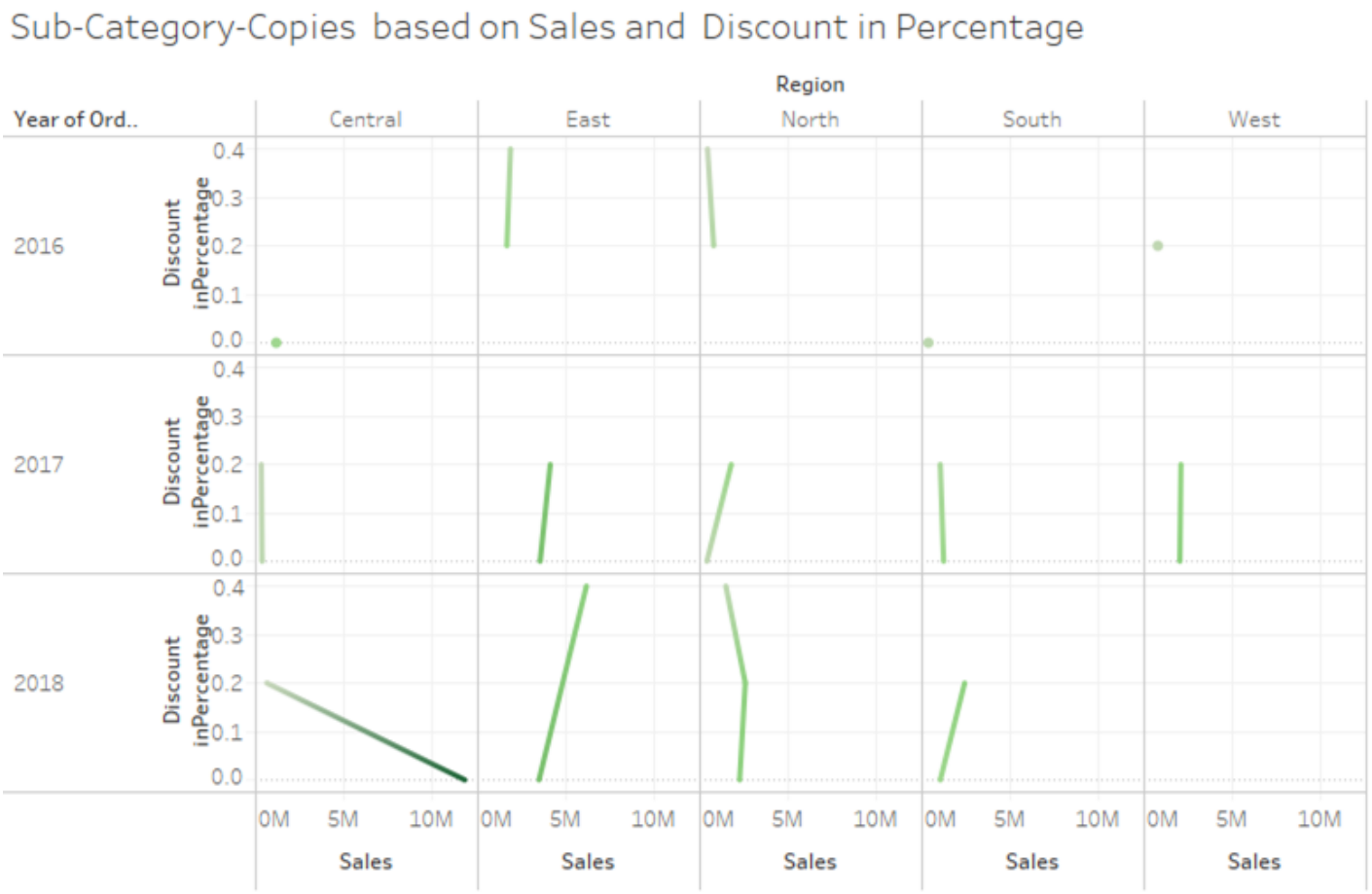
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by the sum of Profit. The data is filtered on Sub-Category, which keeps Copies.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Copies.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



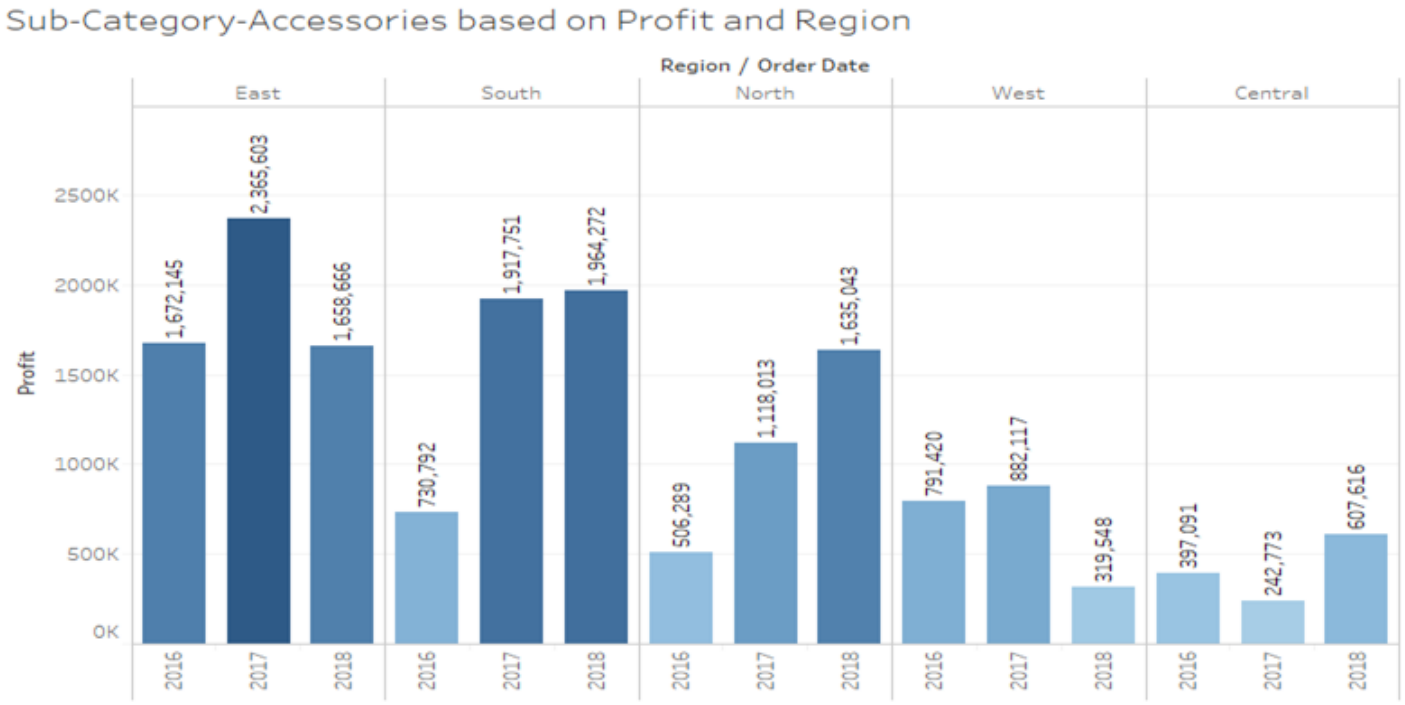
Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Copies.

From above graph we see that,

For the Central Region: without discount our sales are more and making profit, increasing discount our sales is decreasing and its minimizing profit.

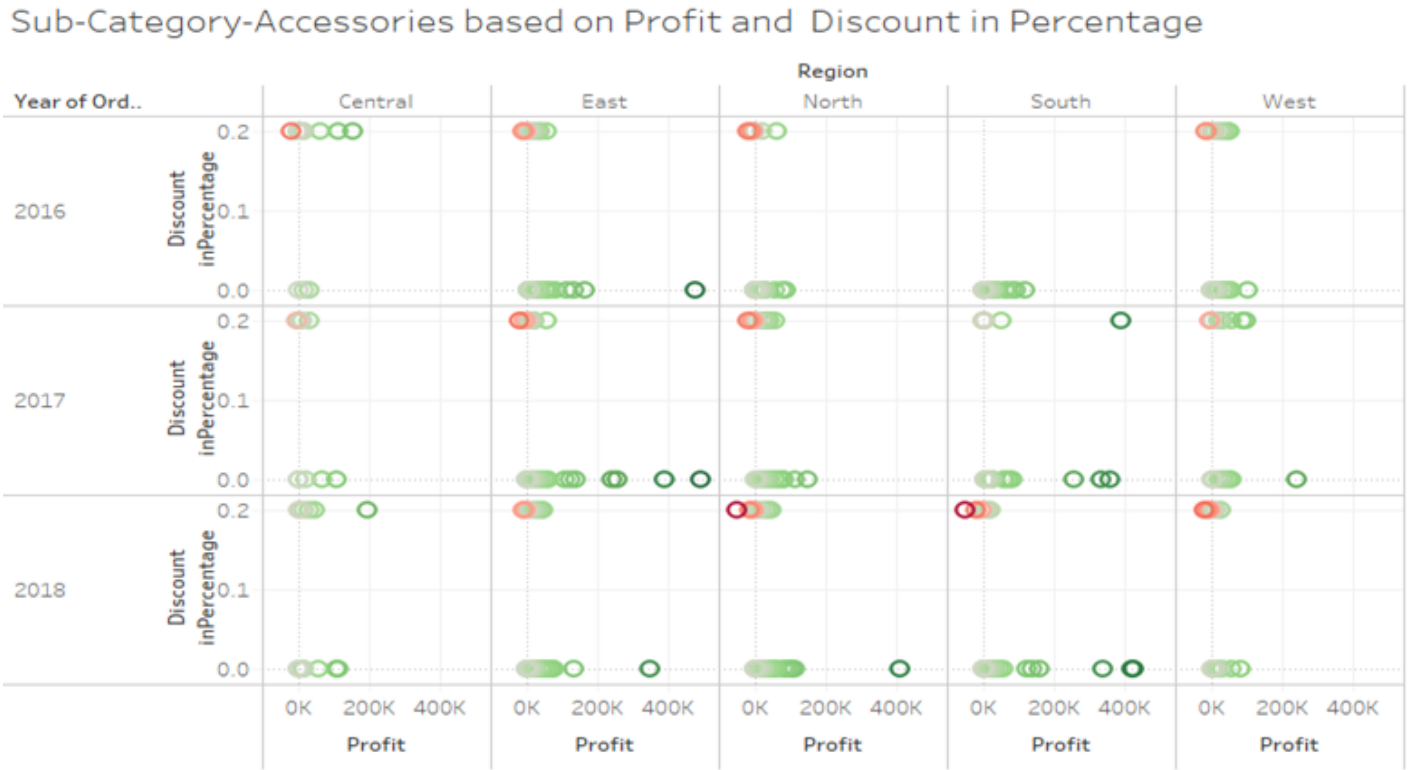
For the East Region: with increasing discounts our sales increase and make profit.

❖ For Accessories



Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by the sum of Profit. The data is filtered on Sub-Category, which keeps Accessories.



Above graph is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Accessories.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).

Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Accessories.

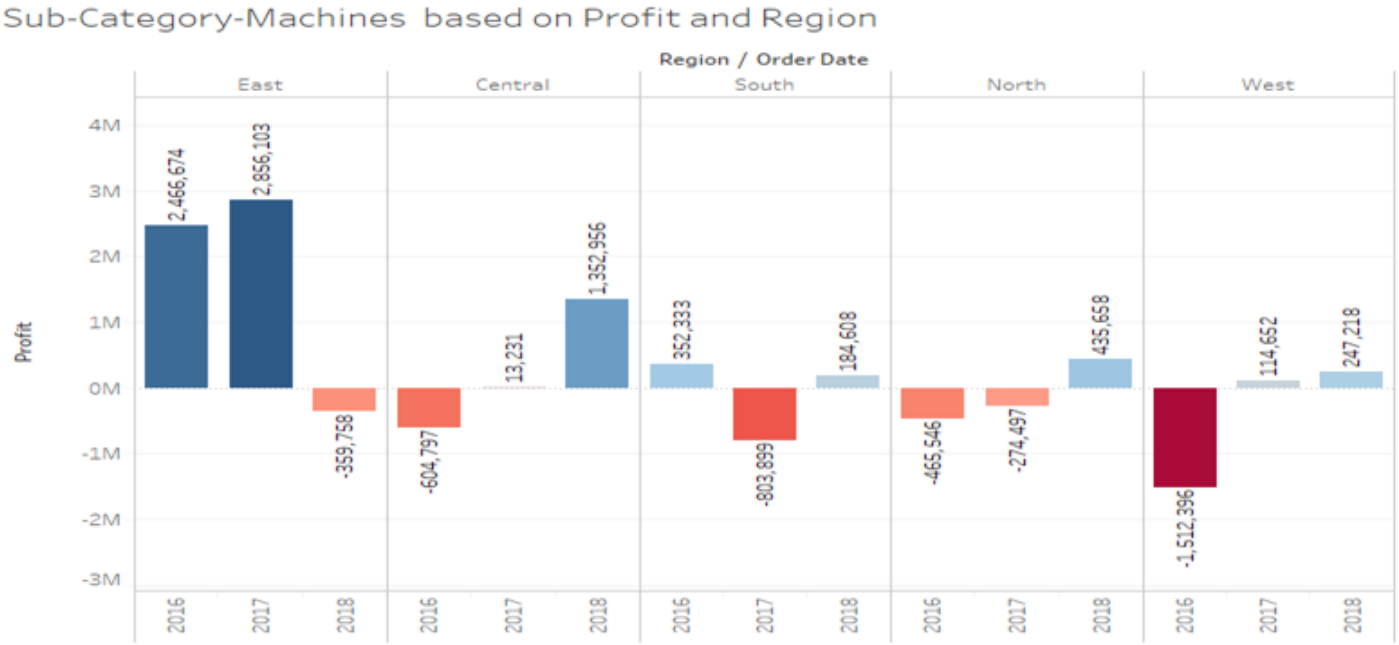
From above graph we see that,

For the East and North Region: without discount our sales are more and making more profit, increasing discount our sales is decreasing and minimizing profit.

For the South Region: without discount our sales are more and making more profit, increasing discount our sales is decreasing and failing profit hence making losses.

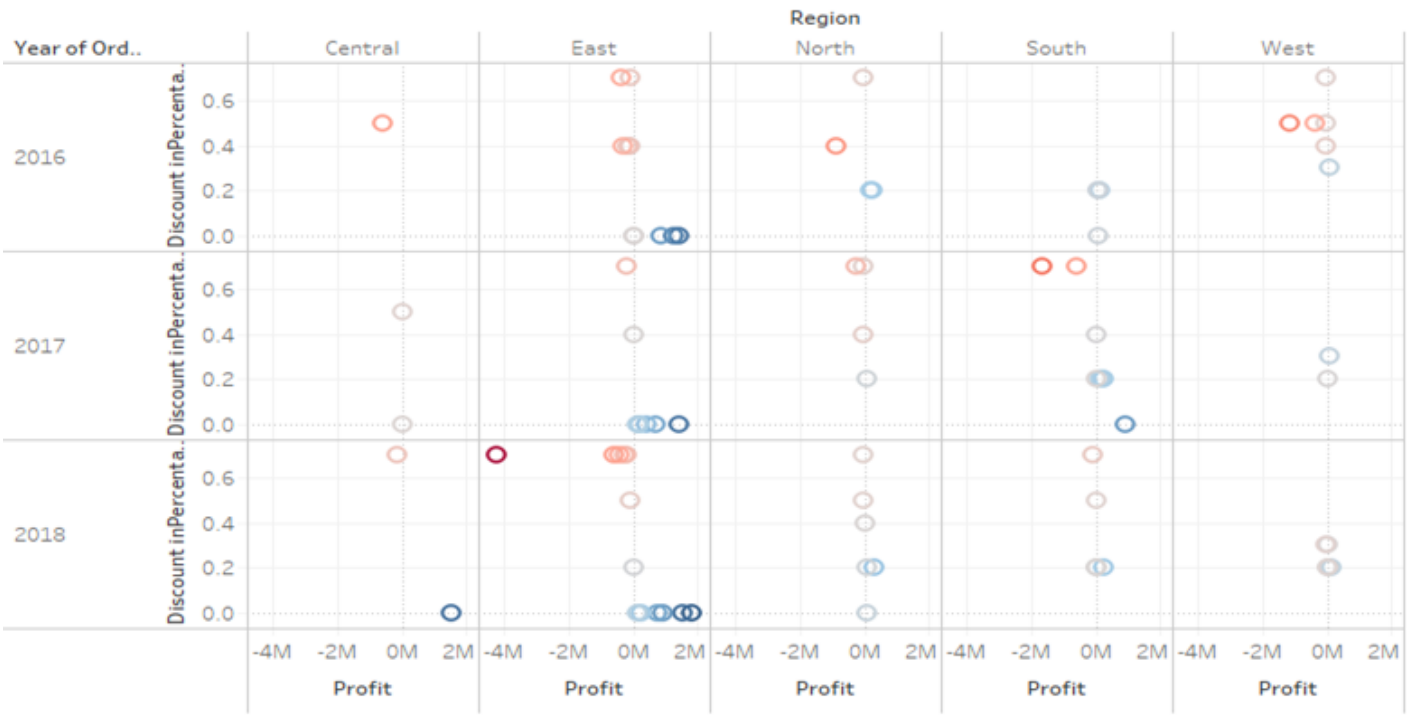
For the Central Region: with increasing discounts our sales are increasing and making profit.

❖ For Machines



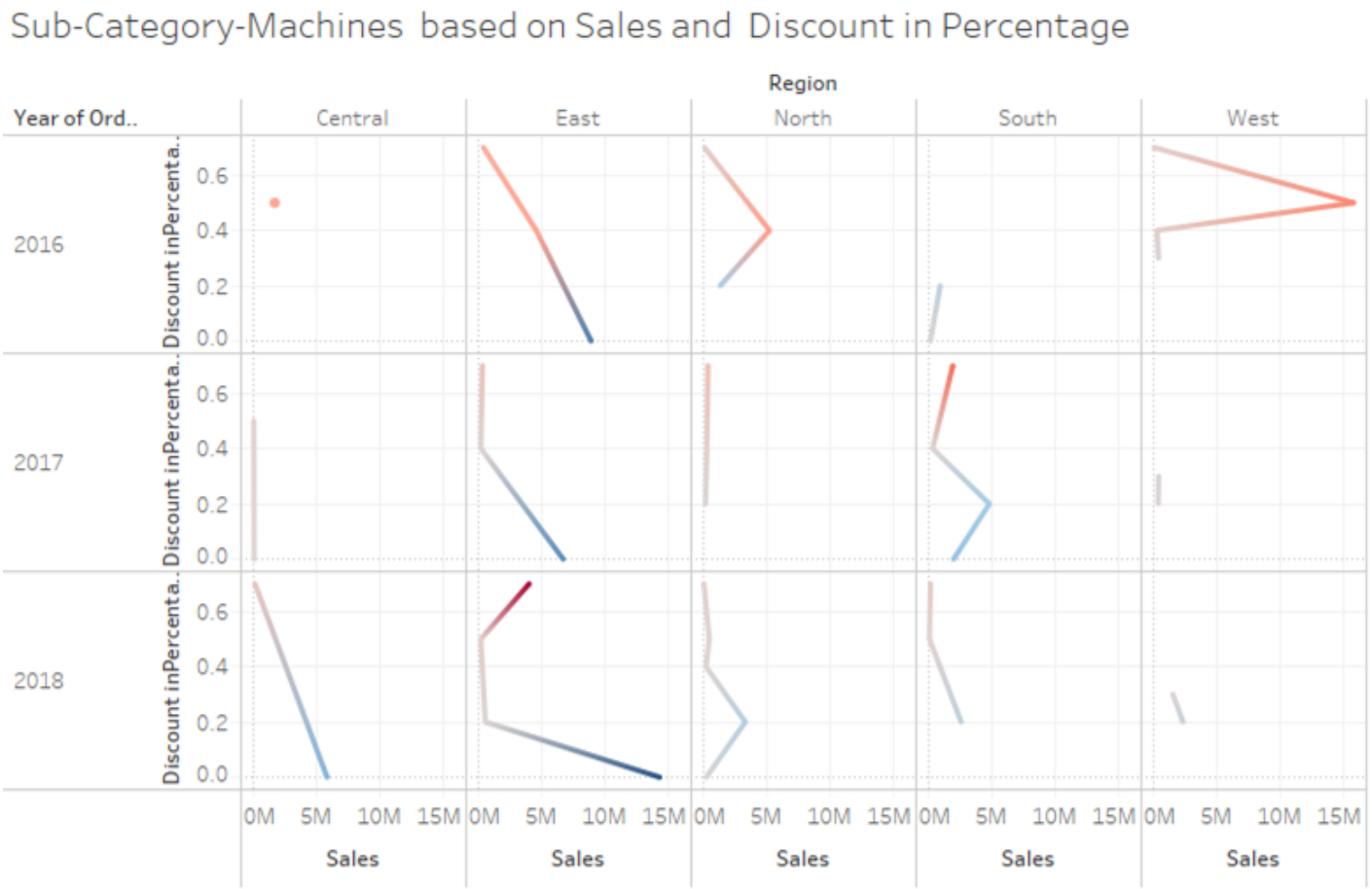
Above graph is about the Sum of Profit for each Order Date Year broken down by Region. Colour shows sum of Profit.

The marks are labelled by sum of Profit. The data is filtered on Sub-Category, which keeps Machines.



The graph above is about Profit vs. Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Machines.

Here, from above both the graphs, we see that as we increase Discount leads to minimizing Profit (making losses).



Above graph is about the trend of sum of Sales for Discount in Percentage broken down by Region vs. Order Date Year. Colour shows sum of Profit. The data is filtered on Sub-Category, which keeps Machines.

From above graph we see that,

For the East Region: without discount our sales are more and making more profit, increasing discount our sales is decreasing and minimizing profit. In 2018 as we increase the discount from 50% to 70% our Sales will increase and make more losses.

For the Central Region: without a discount our sales are more and making more profit. Hence in 2018 we are getting more profit.

Objective 4. What is the focus area for growth for the company in the next 3 years?

Company gains a good amount of profit from the region East, around 33% of the overall profit. And over the past 3 years the region south and east has had a good amount of profit gain but region east is dominating in one year.

Hence company should focus on these two regions. And for the customers and Products we have seen as above.

11. Market Basket Analysis

Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identifies patterns of co-occurrence. A co-occurrence is when two or more things take place together.

Market Basket Analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.

For Full data:

Market Basket Analysis

Sub-Category1

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	41															
Art	57	40														
Binders	101	82	111													
Bookcases	18	9	23	39												
Chairs	37	27	52	76	15											
Copiers	1	5	3	10	1	5										
Envelopes	27	14	20	41	7	23	3									
Fasteners	17	12	22	45	6	19	1	8								
Furnishings	71	52	65	137	21	62	10	20	28							
Labels	27	13	28	56	13	23	4	8	12	43						
Machines	17	8	11	21	4	11	1	5	8	21	7					
Paper	106	70	106	187	31	93	11	41	49	121	58	20				
Phones	74	48	75	125	27	55	7	25	24	97	37	16	114			
Storage	65	36	75	130	26	68	5	28	21	91	44	18	122	72		
Supplies	21	16	16	32	5	13	1	4	8	21	15	1	31	14	20	
Tables	30	17	29	47	7	19		9	5	24	11	5	34	32	34	6

Distinct count of Order ID broken down by Sub-Category. Color shows sum of Quantity. The marks are labelled by a distinct count of Order ID. The view is filtered on Sub-Category.

Above plot is for the subcategories for the full data. It is created in Tableau. We are interested in which subcategories the products have been sold more often. We have put the quantities in the color that is dark color will indicate a greater number of items sold So, we will focus on them and want to recommend to put that items together or near or to recommend for other subcategory products to the customer who is going to buy product from the one subcategory.

The numbers show how many times a particular item from those two subcategories are sold. It is a kind of lower triangular matrix of the full square matrix. The upper triangular values are not shown here for better visual understanding. The upper triangular will be symmetric to the lower triangular.

For e.g., the first cell of the plot has value 41. That indicates that 41 times order has been placed for items of sub category Appliance and Accessories together.

Now we can see that the maximum amount 187 is from the subcategory's papers and binders. That is understandable as customer buying paper will mostly need binders too. But from the market basket analysis we want to find some unusual patterns of buying so, let's focus on that.

We can see that Phones are bought 125 ,114 times with Binders and papers respectively. Now this hasn't happened maximum time but it's a fairly large number.

These numbers are for full data so let's see what is happening every year that the buying patterns are changing with time or place.

The data is filtered on Order Date Year, which keeps 2016:

Sub-Category1

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	16															
Art	20	11														
Binders	29	24	29													
Bookcases	4	2	4	11												
Chairs	9	8	24	22	2											
Copiers		3				1										
Envelopes	9	5	5	16	2	8	2									
Fasteners	7	4	10	13		6		3								
Furnishings	26	23	15	35	4	19	3	4	11							
Labels	8	4	8	18	4	6			3	18						
Machines	3	3	4	5	1	2		2	4	6	1					
Paper	34	30	32	51	7	32	1	11	15	32	19	8				
Phones	20	14	28	37	8	16	3	10	6	28	14	7	35			
Storage	21	11	27	38	8	21		7	5	27	14	7	33	26		
Supplies	12	4	6	10	3	3		1	2	6	3		13	6	5	
Tables	9	7	11	17	2	7		3		9	2	2	8	8	15	2

The data is filtered on Order Date Year, which keeps 2017:

Sub-Category1

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage
Appliances	13														
Art	18	13													
Binders	31	22	40												
Bookcases	10	2	12	16											
Chairs	8	5	15	22	9										
Copiers	1	2	1	6	1	1									
Envelopes	8	2	9	11	2	8									
Fasteners	5	3	6	17	2	5	1	3							
Furnishings	24	10	22	54	10	19	5	8	7						
Labels	6	5	12	17	5	6	3	6	5	11					
Machines	6	1	6	6	2	3	1	1	3	9	3				
Paper	36	13	29	53	16	23	7	16	10	44	19	2			
Phones	27	16	25	48	13	25	2	8	7	37	8	7	33		
Storage	18	10	19	42	10	18	3	12	7	29	11	6	38	21	
Supplies	4	3	3	9	1	3	1	1	3	4	3		8	2	7
Tables	11		6	6	2	4		3	2	5	2		11	9	7

The data is filtered on Order Date Year, which keeps 2018:

Sub-Category (Sales-Data1)	Sub-Category1															
	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	12															
Art	19	16														
Binders	41	36	42													
Bookcases	4	5	7	12												
Chairs	20	14	13	32	4											
Copiers			2	4		3										
Envelopes	10	7	6	14	3	7	1									
Fasteners	5	5	6	15	4	8		2								
Furnishings	21	19	28	48	7	24	2	8	10							
Labels	13	4	8	21	4	11	1	2	4	14						
Machines	8	4	1	10	1	6		2	1	6	3					
Paper	36	27	45	83	8	38	3	14	24	45	20	10				
Phones	27	18	22	40	6	14	2	7	11	32	15	2	46			
Storage	26	15	29	50	8	29	2	9	9	35	19	5	51	25		
Supplies	5	9	7	13	1	7		2	3	11	9	1	10	6	8	
Tables	10	10	12	24	3	8		3	3	10	7	3	15	15	12	4

Above plots are for three years individually. We can see that it also shows the same pattern as the full data. Maximum sale is of Papers and Binders only. But in these also we can see that customers who buy Paper or Binders are buying Phones too.

Now let's check the behaviour of sales in particular Regions.

The data is filtered on Region, which keeps East:

Sub-Category (Sales-Data1)	Sub-Category1															
	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	10															
Art	22	6														
Binders	43	25	41													
Bookcases	8	4	12	18												
Chairs	15	4	20	29	9											
Copiers	1	1		3		2										
Envelopes	6	3	8	20	5	8	1									
Fasteners	3	2	9	19	3	6		4								
Furnishings	22	11	22	58	7	22	3	8	10							
Labels	7	3	11	18	5	8	2	4	5	14						
Machines	7	3	4	10	1	6		2	2	9	2					
Paper	39	20	41	64	17	30	3	16	18	49	19	5				
Phones	26	14	23	57	10	26	1	9	9	31	10	5	42			
Storage	26	9	31	51	8	26	2	14	12	33	16	8	47	28		
Supplies	9	4	9	15	2	4	1	1	2	9	4		9	5	5	
Tables	9	3	6	12	3	5		4	3	6	2	3	12	10	10	3

The data is filtered on Region, which keeps West:

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage
	Sub-Category1														
Appliances	9														
Art	8	4													
Binders	19	18	18												
Bookcases	2	2		6											
Chairs	8	5	11	11	2										
Copiers					1										
Envelopes	3	1	2	4		6									
Fasteners	6	2	2	6	1	3									
Furnishings	12	11	14	26	3	11		3	6						
Labels	9	4	4	12	2	4	1		3	7					
Machines	2	1	3	3	1	2		1	2	2	1				
Paper	19	17	13	39	3	17	1	6	9	20	11	7			
Phones	12	6	11	16	3	9		2	3	23	5	2	19		
Storage	14	9	7	27	3	13	1	2		19	12	2	25	11	
Supplies	8	6	1	5		3		2	2	2	4	1	9	4	5
Tables	7	5	4	10	2	4		1		9	1		6	5	8

The data is filtered on Region, which keeps South:

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
	Sub-Category1															
Appliances	7															
Art	7	13														
Binders	16	17	24													
Bookcases	2	1	3	6												
Chairs	8	9	8	11	1											
Copiers		1	2	3		2										
Envelopes	5	4	2	8	1	4	2									
Fasteners	4	3	8	8	1	5		1								
Furnishings	13	14	13	23	4	6	1	5	5							
Labels	7	3	7	9	2	3	1	1	2	6						
Machines	6	1	1	2	1			1	3	4	3					
Paper	18	10	18	33	3	15	4	7	12	15	10	5				
Phones	15	13	17	23	7	7	2	6	7	23	10	6	21			
Storage	12	5	15	26	5	9		6	5	18	6	4	22	17		
Supplies	2	3	2	6	2	3			1	5	2		6	2	4	
Tables	4	3	6	13		2		1	1	5	4	1	7	11	3	1

The data is filtered on Region, which keeps North:

Sub-Category1

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	8															
Art	12	10														
Binders	17	15	17													
Bookcases	5	1	7	8												
Chairs	4	9	10	20	2											
Copiers		2		1		1										
Envelopes	8	4	7	6	1	3										
Fasteners	3	2	1	10	1	2	1	2								
Furnishings	17	9	12	21	5	19	5	3	5							
Labels	4	1	4	12	4	4		3	1	11						
Machines	1	2	2	4	1	3	1	1	1	5	1					
Paper	21	13	22	35	6	20	2	8	5	23	13	3				
Phones	18	14	21	22	5	9	3	4	3	17	8	3	26			
Storage	11	11	15	18	8	14		6	3	14	10	3	20	15		
Supplies	1	2	1	5	1	1			1	5	4		3	2	4	
Tables	8	4	9	8	1	7		3	1	4	4	1	9	5	8	1

The data is filtered on Region, which keeps Central:

Sub-Category1

Sub-Category (Sales-Data1)	Accessories	Appliances	Art	Binders	Bookcases	Chairs	Copiers	Envelopes	Fasteners	Furnishings	Labels	Machines	Paper	Phones	Storage	Supplies
Appliances	7															
Art	8	7														
Binders	6	7	11													
Bookcases	1	1	1	1												
Chairs	2		3	5	1											
Copiers		1	1	3												
Envelopes	5	2	1	3		2										
Fasteners	1	3	2	2		3		1								
Furnishings	7	7	4	9	2	4	1	1	2							
Labels		2	2	5		4			1	5						
Machines	1	1	1	2						1						
Paper	9	10	12	16	2	11	1	4	5	14	5					
Phones	3	1	3	7	2	4	1	4	2	3	4		6			
Storage	2	2	7	8	2	6	2		1	7		1	8	1		
Supplies	1	1	3	1		2		1	2		1		4	1	2	
Tables	2	2	4	4	1	1								1	5	1

In these, too Paper and Binders are dominant Sub-Categories. In Region East around 30 times Phones are sold with Furnishing Products. Thing to be noticed is sales of copiers is too low compared to other categories. Only 161 copiers are sold in the entire 3 years and sale of copiers with any other subcategories is too low. We can say that it's an item to be sold singularly and hasn't been ever sold with tables.

Trial Models

First of all, when we get the data, we try to fit the regression models on it to answer the last three objectives. But the assumptions were violated. Now as a remedial measure we try to transform the data but it didn't work as well.

Then in the first objective while trying to forecast sales for 2019 firstly we tried to go for multivariate time series. In the Multivariate time series model, there are more than two variables. One is an independent variable which contains time points and other independent variables. Another is a dependent variable which depends on the time and other independent variables, and we want to predict.

- Vector Auto-Regression (VAR).

In a VAR model first, we predict the independent variable. then with the help of it we predict the Sales.

The VAR model predicts the value in floating points. But as our data contain Categorical variables which are integer values, Therefore the predicted floated value is not appropriate for the further forecasting of Sales.

- Random Forest Regression

For this model we take, Sales as dependent variable and independent variable are Quantity, Discount_inPercentage, Customer_ID_coded, Region_coded, Segment_coded, Category_coded, Unit_Price, Sales_Person_ID_coded.

we get R-Square Error associated with Random Forest Regression is: 0.80251
That is 80% variation of dependent variables is explained by independent variables in our model.

As Random Forest Regression gives the model, in which we have to pass the value of the independent variable. But we don't have the values of the independent variable for 2019.
So, our model is best fitted but it does not fully fill our objective. Because our objective is to predict the Sales of 2019.

12. Appendix

[1] Data fields

Attributes		*Columns Names*	*Brief Description*
Date	<->	**Order Date**	*Date of Order*
		Ship Date	*Date when shipped*
		Payment Date	*Date of Payment*
Customer	<->	**Customer ID**	*Customer Identification*
		Customer Name	*Customer Name*
Country	<->	**City**	*City where customer shopped*
		State	*State where customer shopped*
		Region	*Region of City/State*
		Country	*Country of City/State*
Product	<->	**Product ID**	*Product id of item shopped*
		Product Name	*Product name*
Product Group / Business Line	<->	**Segment**	*Segment to which the customer belongs to*
		Category	*Category to which the product belongs to*
		Sub-Category	*Sub Category to which the product belongs to*
Sales Amount	<->	**Sales**	*Sales amount or order*
Sales Margin	<->	**Profit**	*Profit or margin at product level*

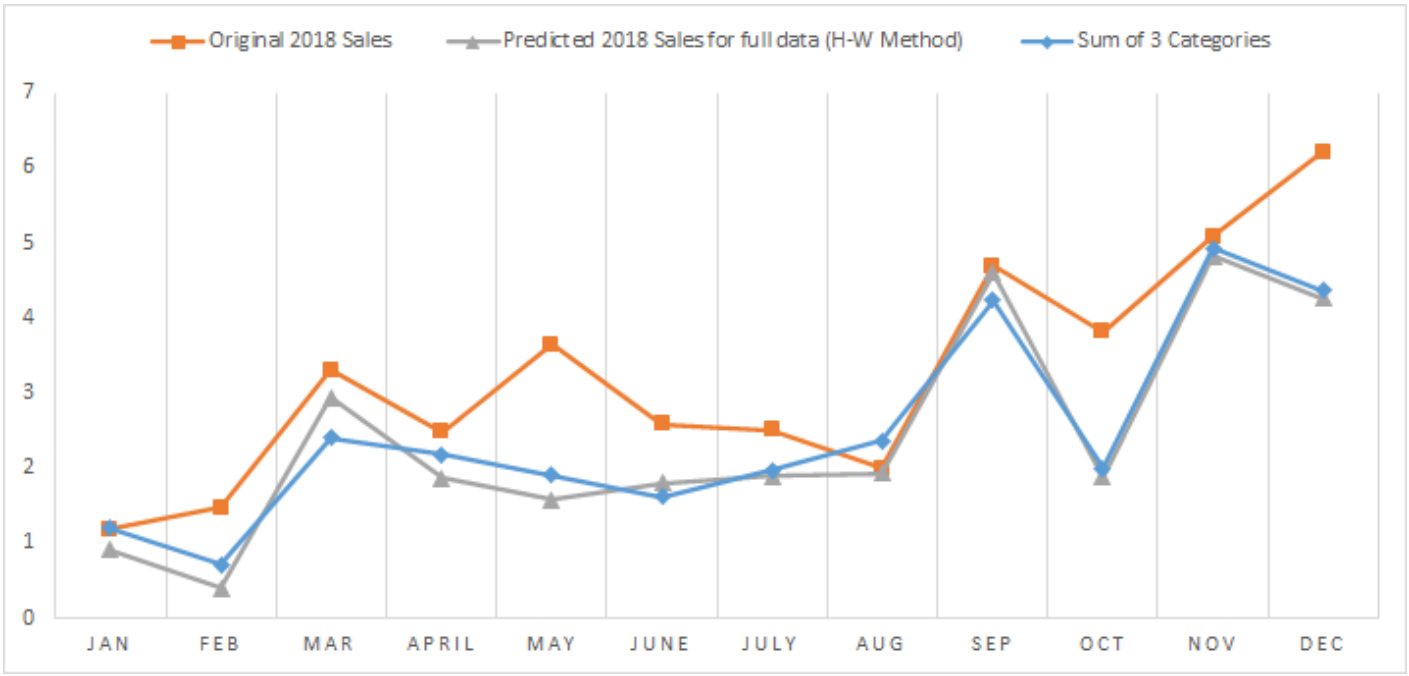
Sales Quantity	<->	**Quantity**	*Quantity of item sold*
Discount	<->	**Discount (in %)**	*Discount in percentage*
Sales Person Id	<->	**Sales Person ID**	*Sales person id*
Sales Person	<->	**Sales Person Name**	*Sales person name*
Credit Days	<->	**Credit Days**	*Payment terms of extending credit*
Order Id	<->	**Order ID**	*Order Id*
Ship Mode	<->	**Ship Mode**	*Mode of Shipment*

[2] HTML file of python code for the first objective:

File Name: objective1_Forecasting Sales for 2019

[https://drive.google.com/file/d/1PwuMzrJlwTDXzg8xbhLENV4Eel6P7A1f/view?usp=sharing]

[3] Comparison graph for the two forecasted sales of 2019.



[4] Excel file of ABC inventory analysis

File Name: data_Profit - Category and Region

[https://drive.google.com/file/d/1S5x4arcHqnVl0TxiYnaiaJl3COL0Xi8Z/view?usp=sharing]

13. References

1. [Smoothing Techniques for time series data | by Sourav Dash | Medium]

2. Basic Econometric by Damodar N. Gujarati (Fifth Edition)
3. [An overview of time series forecasting models | by Davide Burba | Towards Data Science]
4. [Akaike Information Criterion | When & How to Use It (scribbr.com)]
5. [https://machinelearningmastery.com/introduction-to-time-series-forecasting-with-python/]
6. [https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared]
7. [https://anomaly.io/seasonal-trend-decomposition-in-r/index.html]
8. [https://stats.stackexchange.com/questions/333092/why-i-get-the-same-predict-value-in-arima-model]
9. [What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared? - Studytonight]