Problem Description

In modern times where every business is highly dependent on its data to make better decisions for growing business, time series analysis plays an important role in helping different business entities to get an idea about how good their sales are by implementing sales forecasting on the historic data.
Given a Retail dataset of a global superstore for 4 years, we have utilized it to perform exploratory data analysis to gain valuable insights and further apply time series analysis to get a forecast of sales of 7 days from the last date of the Training dataset.

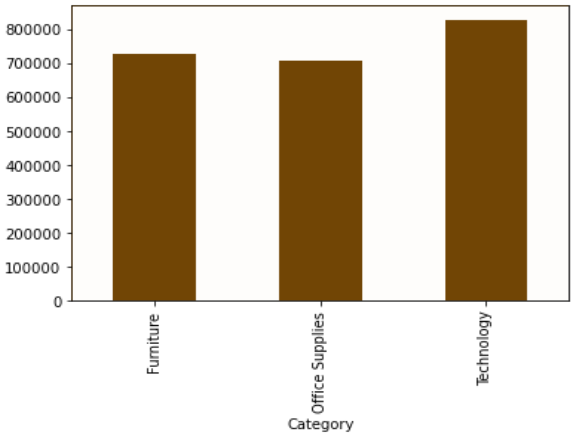
Data Pre-processing

For the pre-processing part, we first looked for the null values present in the data. We found that only the variable named *Postal Code* was having 11 null values, all corresponding to the Burlington city in Vermont. So, we replaced the null values with the actual postal code. Then we combined the *Ship Date* and *Order Date* variables to a new variable named *Month* corresponding to which the order is placed.

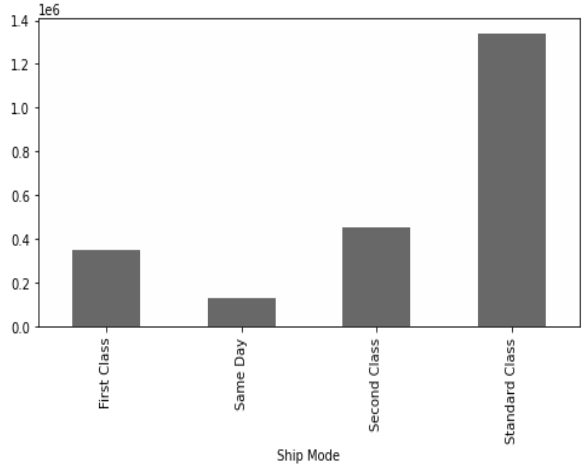
Feature	Unique Values count	Type	Description
Order ID	4922	Nominal	ID of the placed order
Order Date	1230	Nominal	Date of placing order
Ship Date	1326	Nominal	Date when order got shipped
Ship Mode	4	Nominal	The class of shipment mode
Customer ID	793	Nominal	ID of the customer ordering
Customer Name	793	Nominal	Name of the customer
Segment	3	Nominal	The segment to which the good belongs
Country	1	Nominal	Location of Store (Country)
City	529	Nominal	Location of store (City)
State	49	Nominal	Location (State)
Postal Code	626	Nominal	Postal code of the area shipped to
Region	4	Nominal	North, South, East or West
Product ID	1861	Nominal	ID of product ordered
Category	3	Nominal	Category of the product
Sub-Category	17	Nominal	Subcategory of the product
Product Name	1849	Nominal	Name of the product
Sales	5757	**Numeric**	Price of the order

The Given Data consists of 9800 rows. We can see that all the 16 features are Nominal and the target variable (Sales) which we have to predict is Numeric. The *Order ID, Customer Name, Customer ID, Postal Code, Product ID* and *Product Name* features are of no use for EDA/prediction as they are discrete random values which have no relation with the target variable. Also, the company's customers are in only one Country - United States and hence the Country column can also be removed. So now we are left with 7 variables to visualize and get trends from the data.

Data Understanding and Visualizations



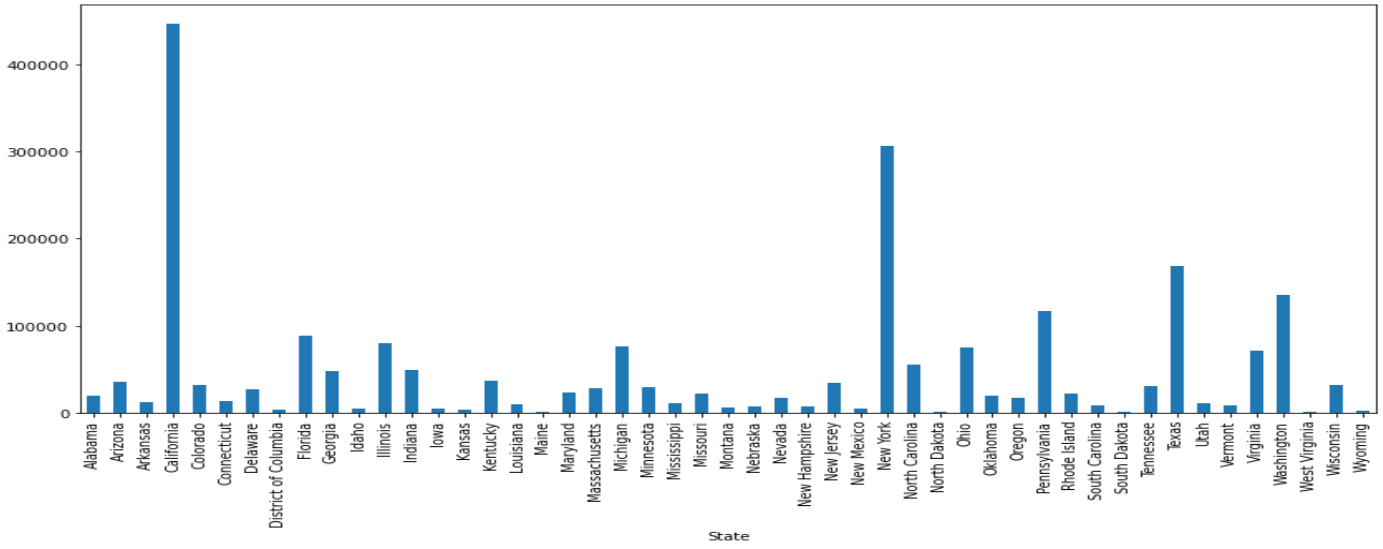
The superstore offers a full range of products to meet the needs of its customers. We see that the products are divided into three different categories: *Furniture*, Office *supplies* and *Technology*. We observe that the net sales in all the categories are the same just the Technology category has slightly more sales than that in Furniture and Office Supplies.



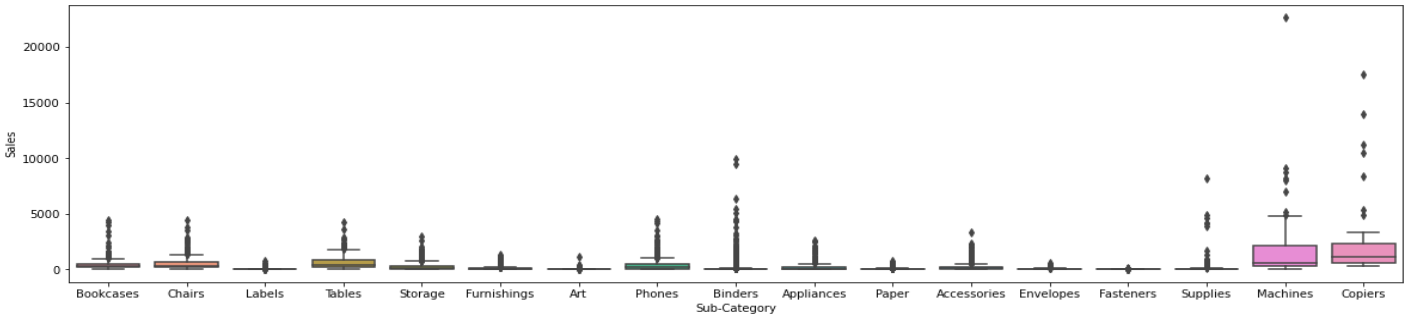
The company offers four possible levels of delivery to its customers, either *Same day*, *First class, Second class* or *Standard class*. We observe here that for most of the customers, *Standard class* is the preferred mode of delivery.

Among the subcategories in the database are accessories, appliances, art, binders, bookcases, chairs, photocopiers, envelopes, fasteners, furniture, labels, machines, paper, telephones, storage tools and tables.

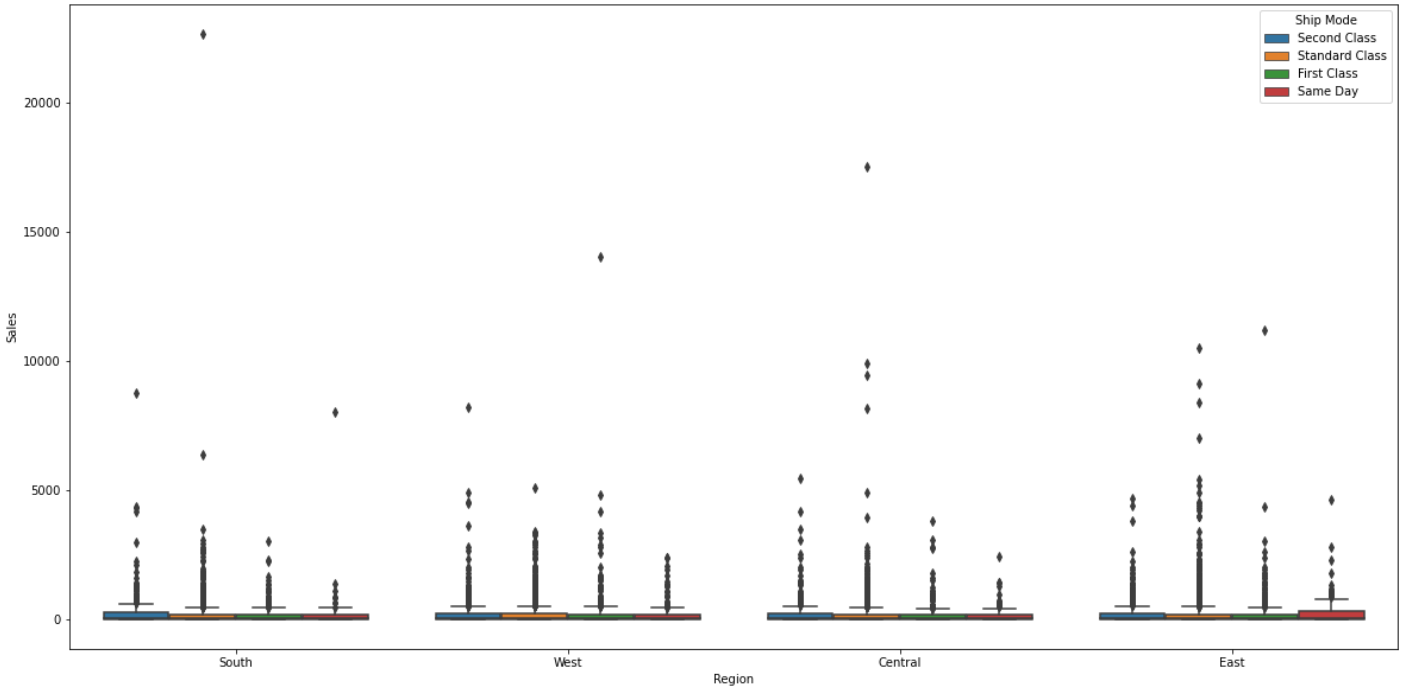
California has the largest volume of sales compared to other states and only 5 states have net sales of more than $ 100,000. We also observe that the

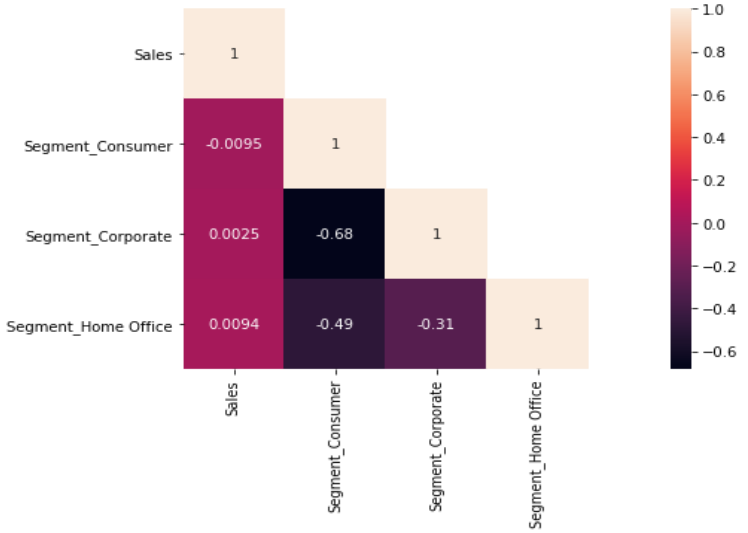


most expensive items sold by the superstore belong to either machines or copiers.

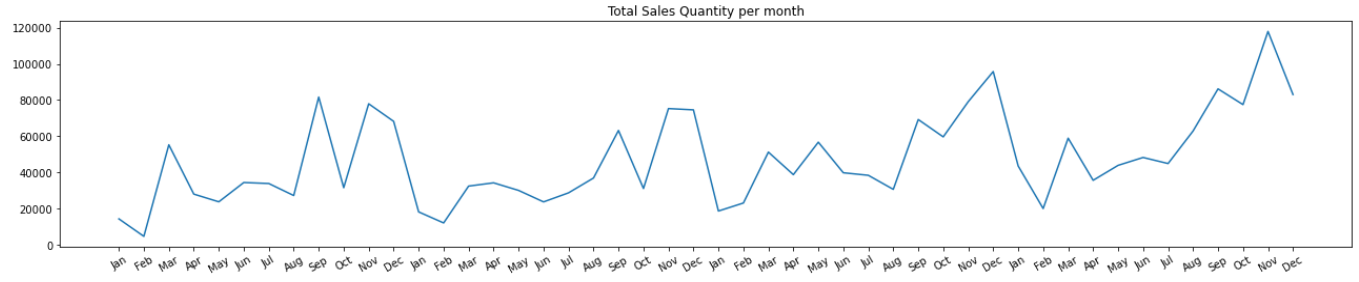


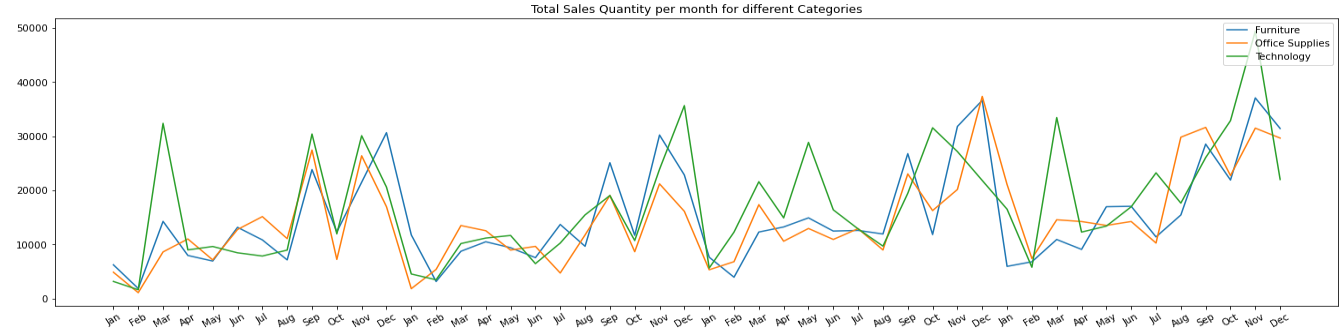
We also observe that the most expensive items sold by the superstore belong to Technology for all four regions since it's median is higher and also it contains the most number of outliers, also for office supplies the median is on the lower side which indicates that it's items are relatively cheaper.

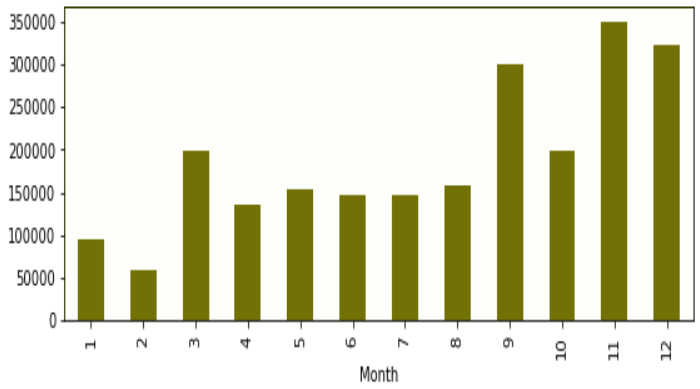


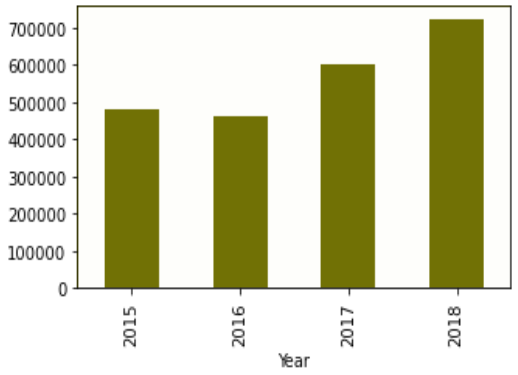


After looking at this heat map plot for sales vs different attributes we can conclude that the covariance value is either tending to zero or it is negative, thus the variables are not strongly correlated.









The sales are highest in the month of November while lowest in the month of February. The sales are lowest in the initial months of the year and grow and reach maximum towards the last months. This is the general trend which repeats itself year on year. This knowledge would be very helpful for the stores to meet the demands of the customer and never fell short of supply and indicate when the stores should reduce stock. We also observe that the net sales increase year on year.

The sales in one Category is somewhat related to the sales in another category as is observed from the above image. The similar trend is seen for different *Segments*, *Ship mode* and other features.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Year												
2015	14205.7070	4519.8920	55205.7970	27906.8550	23644.3030	34322.9356	33781.543	27117.5365	81623.5268	31453.3930	77907.6607	68167.0585
2016	18066.9576	11951.4110	32339.3184	34154.4685	29959.5305	23599.3740	28608.259	36818.3422	63133.6060	31011.7375	75249.3995	74543.6012
2017	18542.4910	22978.8150	51165.0590	38679.7670	56656.9080	39724.4860	38320.783	30542.2003	69193.3909	59583.0330	79066.4958	95739.1210
2018	43476.4740	19920.9974	58863.4128	35541.9101	43825.9822	48190.7277	44825.104	62837.8480	86152.8880	77448.1312	117938.1550	83030.3888

The above table gives sales across all categories for each month of every year.

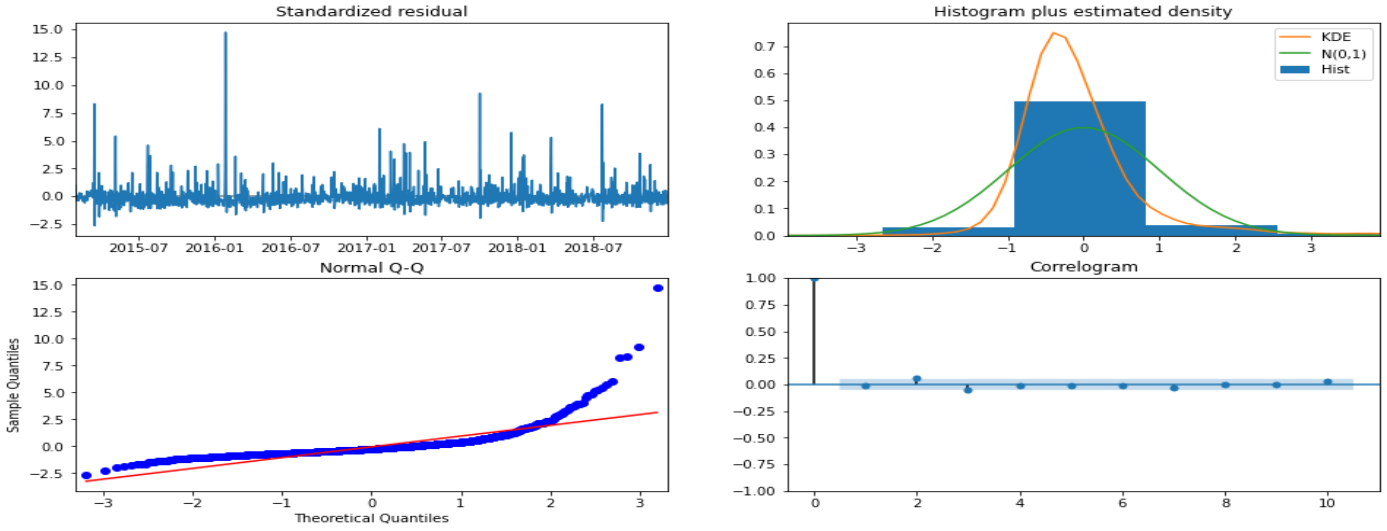
Modelling

For the modelling part, we started by checking whether the time series was stationary or not by performing two types of tests. The first consisted of calculating mean and variance of the data over different periods of time and comparing if the values differed much or not. We saw that the values almost remained constant and hence, the time series was stationary. But to confirm our result, we used another test known as Augmented Dicky Fuller (ADF) test. The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary. We got a p-value <= 0.05 and were able to reject the null hypothesis (H0), the data does not have a unit root and is stationary.
After confirming that the time series was stationary, we decomposed it into trend and seasonality.

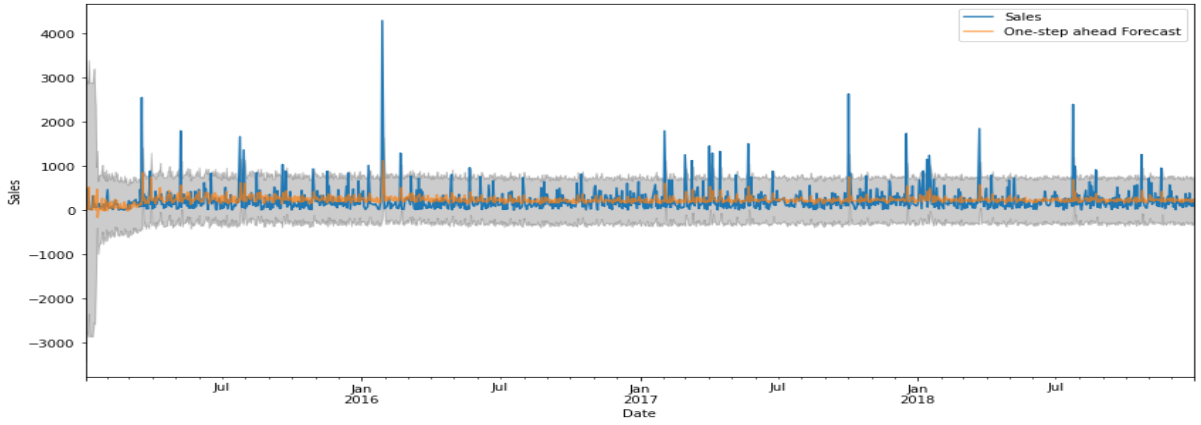
We have chosen the SARIMA model to forecast the sales.
Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that supports univariate time series data with a seasonal component.
SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.
Trend Elements:
There are three trend elements that require configuration.
p: Trend autoregression order.
d: Trend difference order.
q: Trend moving average order.
Seasonal Elements:

There are four seasonal elements that require configuration:
P: Seasonal autoregressive order.
D: Seasonal difference order.
Q: Seasonal moving average order.
m: The number of time steps for a single seasonal period.

The notation for a SARIMA model is specified as: SARIMA(p,d,q)(P,D,Q)m
We use The Akaike information criterion (AIC) for determining the best combination of seasonal parameters for SARIMA. AIC estimates the relative amount of information lost by a given model. The less information a model loses, the higher the quality of that model. After choosing the combination of seasonal parameters with least AIC value, we train the SARIMA model.
After training the model, we get the following results:



Top left: The residual errors seem to fluctuate around a mean of zero and have a uniform variance.
Top Right: The density plot suggests normal distribution with mean zero.
Bottom left: All the dots should fall perfectly in line with the red line. Any significant deviations would imply the distribution is skewed.
Bottom Right: The Correlogram, aka, ACF plot shows the residual errors are not autocorrelated. Any autocorrelation would imply that there is some pattern in the residual errors



From the above plots and the MSE of `262.58` we can conclude that our model is performing fine and is giving good results. So, we now use this to forecast the sales of the next 7 days.

Results and their Interpretation

We get the following results after forecasting the sales:

Prediction Date	Predicted Sales
2018-12-31	191
2019-01-01	223
2019-01-02	223
2019-01-03	221
2019-01-04	226
2019-01-05	217
2019-01-06	227

After observing globally, the database to which we had access, we achieved an interesting goal for this project would be to create a descriptive analysis model allowing us to assess the sales situation of the company during the four years represented in the database. Subsequently using this model, company executives can better orient their strategic decisions regarding the location of their future distribution centres. To do this, various aspects are evaluated with the help of this analysis model, namely the levels of customer orders, sales and shipments according to regions and the level of customer loyalty over the years.