Statistical Learning Project

Dandan Zhao, Marco Furlan

7/14/2022

NOTE: this pdf does not contain the entirety of the code and outputs, it just contains the code and the outputs that were considered relevant, for clarity sake. The full code is sent separately.

Obtaining Data

The dataset we chose to analyse is called the Sample Superstore Dataset. It is a popular artificially generated dataset describing the sellings of a company. Each sample is relative to a purchase, and there are 21 features, 15 categorical and 6 numerical. To see how the dataframe is formatted and its variables we print below the first sample of the dataframe.

```
t(head(data, n=1))
```

```
##                [,1]
## Row ID         "1"
## Order ID       "CA-2020-152156"
## Order Date     "2020-11-08"
## Ship Date      "2020-11-11"
## Ship Mode      "Second Class"
## Customer ID    "CG-12520"
## Customer Name  "Claire Gute"
## Segment        "Consumer"
## Country/Region "United States"
## City           "Henderson"
## State          "Kentucky"
## Postal Code    "42420"
## Region         "South"
## Product ID     "FUR-BO-10001798"
## Category       "Furniture"
## Sub-Category   "Bookcases"
## Product Name   "Bush Somerset Collection Bookcase"
## Sales          "261.96"
## Quantity       "2"
## Discount       "0"
## Profit         "41.9136"
```

Our main goal with this project is going to be to analyse in depth the dataset to create a model that can predict the sales based on the other variables.

Clean and filter data

First things first, we check for NA values:

```
sum(is.na(data))
```

```
## [1] 11
```

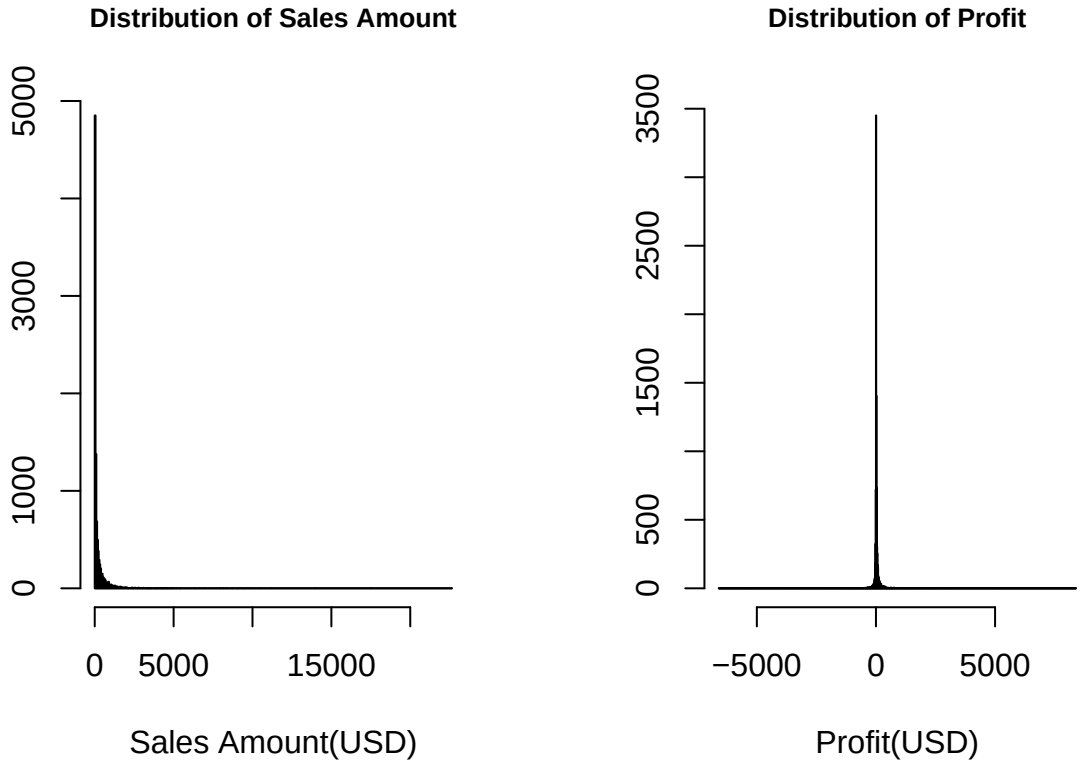
There are some NA values but a further analysis reveals that they all belong to the Postal.Code variable, which we will remove; in fact, we notice that some columns contain useless information (like Row.ID) or redundant information (like Customer.Name and Customer.ID, or Region and Postal.Code whose information is already contained in State for our purposes), so we get rid of those features:

```
data <- subset(data, select=-c(Row.ID,Country.Region,Customer.Name,Postal.Code,Product.Name))
```

We then proceed by cleaning the data relative to the dates, using the library lubridate to split the Order.Date into year, month and day. We also add a Discount.Level column which will contain a categorical variable with value "No Discount", "Low Discount", "Median Discount" and "High discount" based on the discount value; this is done for visualization purposes in future plots.

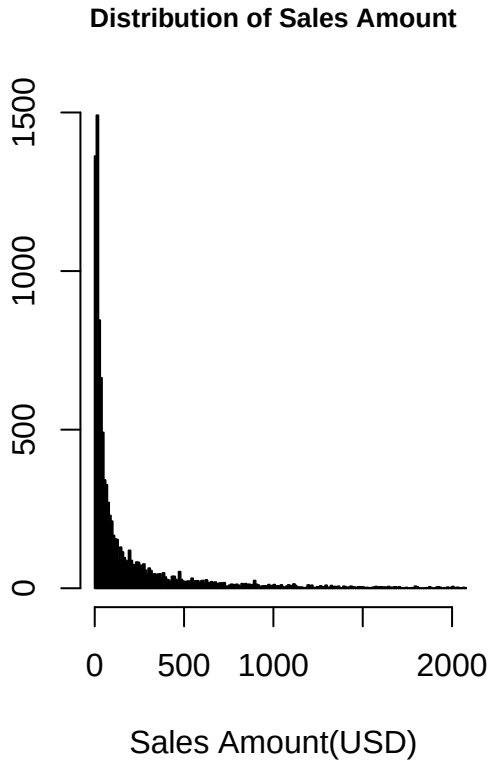
Explore data

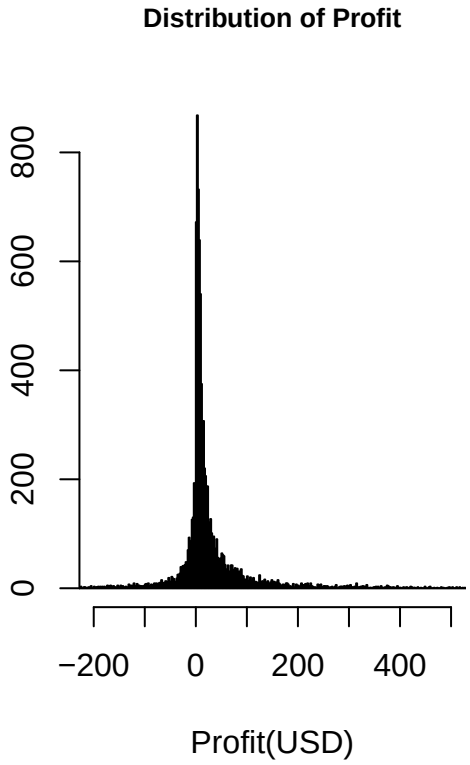
Let's see the histograms of Sales and Profit:



After looking at the plots above we decided to get rid of some extreme cases (or outliars) with respect to Sales and Profit: we set the threshold for the Sales to at most 5000 and for the Profit to between -2000 and 2000; the samples outside these intervals are labeled as outliars not considered relevant for our analysis.

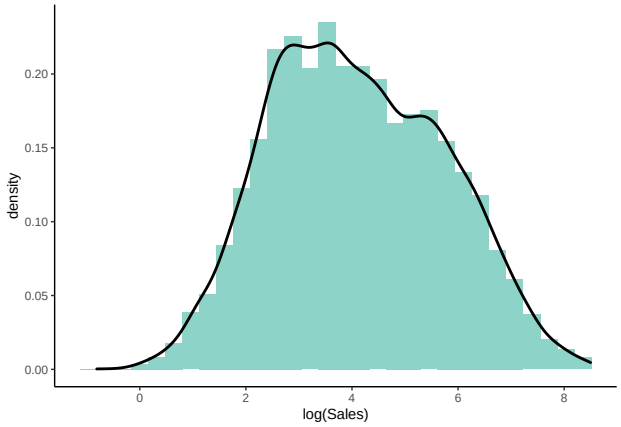
We then plot again the Sales and the Profit:





The histograms are now easier to see and interpret. In particular, we can notice that the Profit mean is slightly to the right of zero, which means our superstore has an overall positive profit.

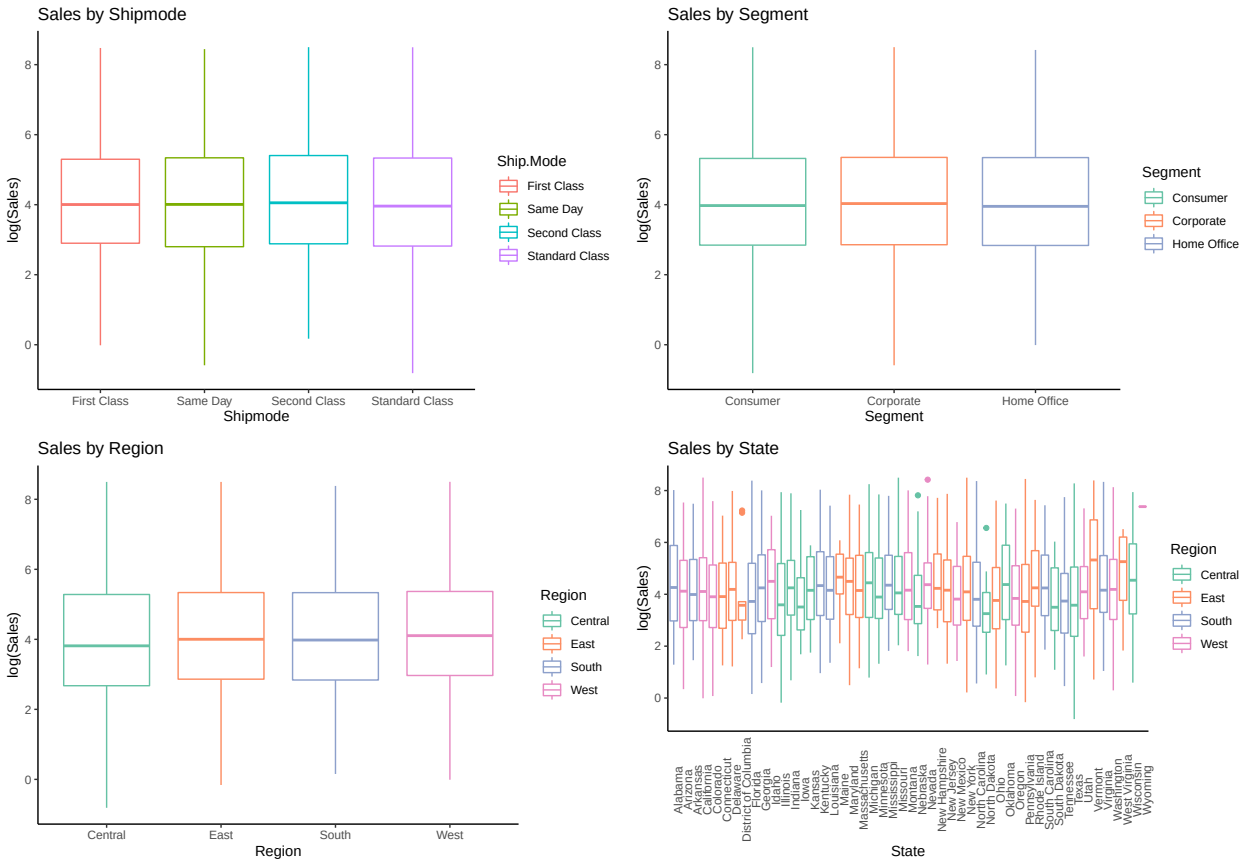
The Sales seem to have an exponential decay, so we plot the histogram of the logarithm of the Sales:



We can see the logarithm of the Sales behaves in a similar fashion as a normal distribution.

Single variable analysis

This section is dedicated to see how the single variables (features) relate to the Sales.



The information given by Shipmode, Segment and Region is not relevant with respect to Sales. The information given by State instead may be relevant - it is to be noted that some states count very few samples so their statistics may not be relevant to consider.

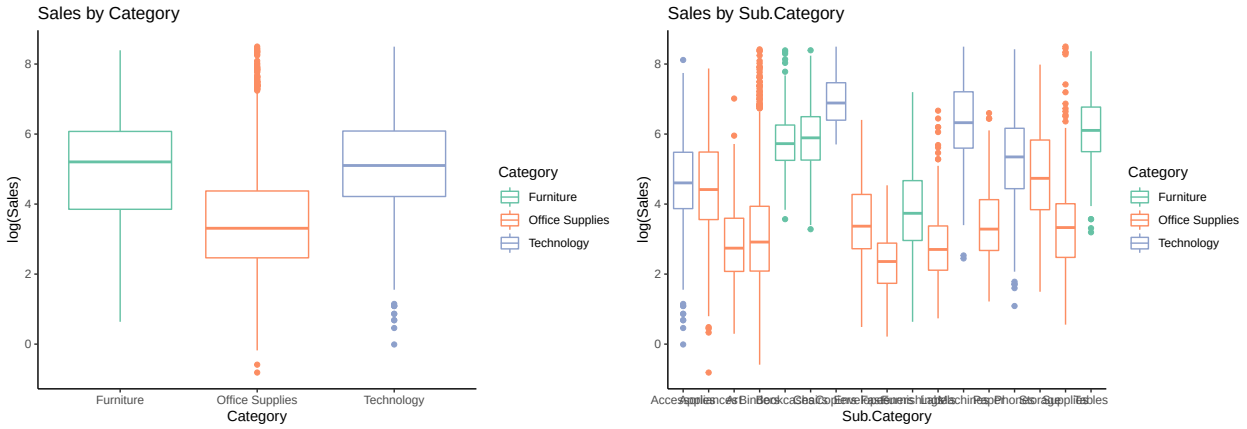
The categorical variable City contains too many categories to be used effectively:

```
length(unique(City))
```

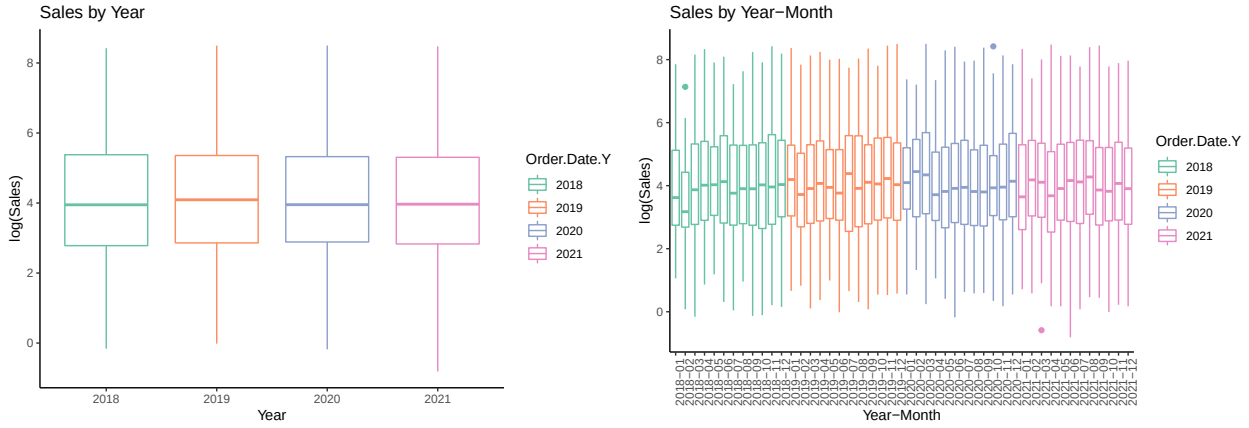
```
## [1] 531
```

Since its information is contained (at least partially) in the States variable, we choose to ignore the City variable.

Let's see the Sales with respect to Category and Sub-Category:



Category and Sub-Category definitely contain information about the Sales, so we will keep them in our model.



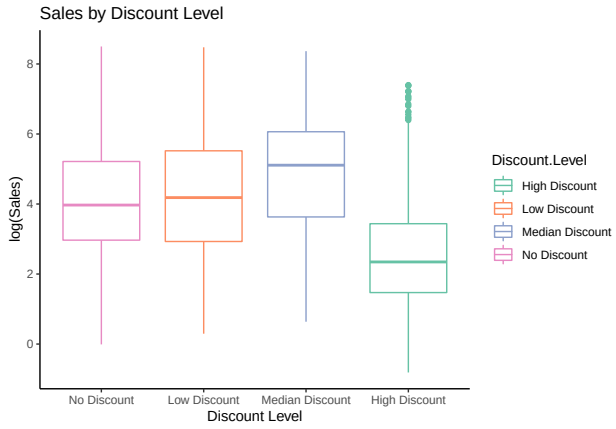
The year doesn't give much information. The month, on the other hand, does, and we will analyse it in more in detail later.

So far we've compared the (logarithm of the) Sales with all the categorical variables, we will now see the relationship with the numerical variables. Let's start from the correlation matrix:

```
cor(sub.data[,c(13,14,15,16)])
```

```
##                Sales    Quantity     Discount      Profit
## Sales      1.00000000 0.257690445 -0.046351970  0.47813900
## Quantity   0.25769045 1.000000000  0.007894283  0.09688136
## Discount -0.04635197 0.007894283  1.000000000 -0.29763658
## Profit    0.47813900 0.096881356 -0.297636580  1.00000000
```

From the matrix we can see that Sales has positive linear relationship with Profit which is consistent with business meaning: more sales means it's possible to get higher profit. The correlation coefficient between Sales and Discount is around 0, which indicates there may not be a linear relationship, so we should look for a non-linear relationship:

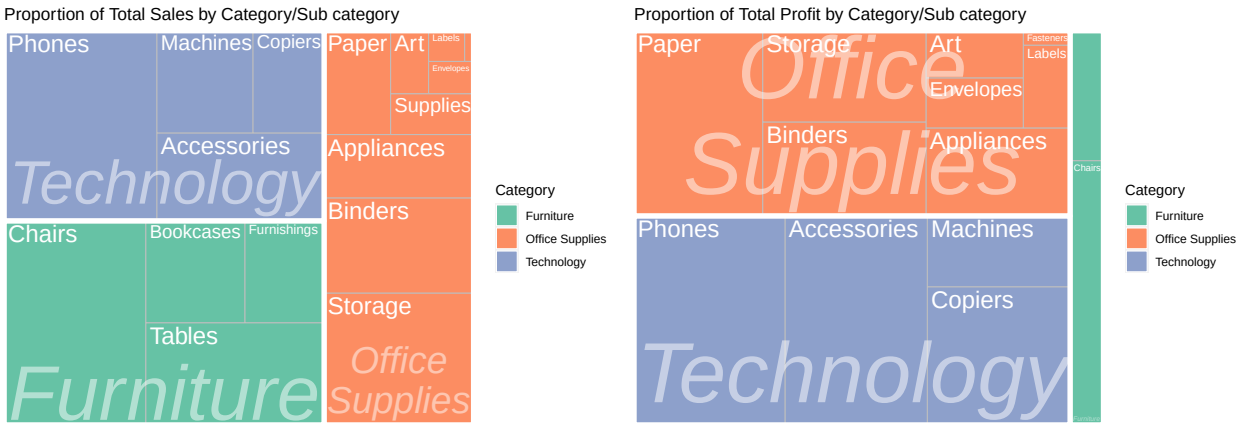


There seems to be a relationship between Sales and Discount. In particular, the higher the discount the higher the sales (except for the high discount case), which makes sense from a business point of view. The high discount case refers to discounts of over 60% and is probably applied in extreme cases where the store needs to get rid of unsold items, hence the low Sales.

Multiple variable analysis

Segment, Region and Shipmode showed no evident relation with Sales. We will start by comparing the combined effect of Category or Subcategory and other variables on the Sales.

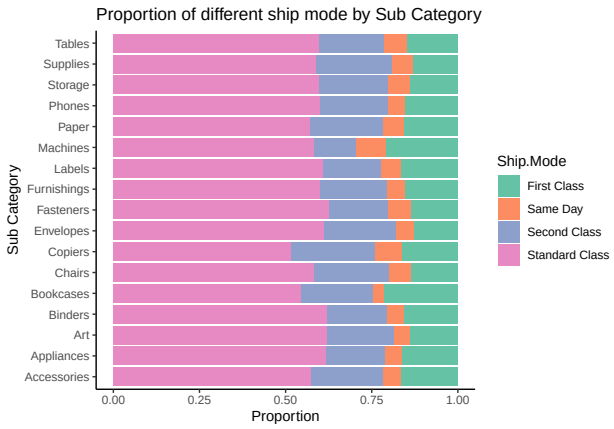
Let's start by comparing Category and Sub-Category with Sales and Profit:



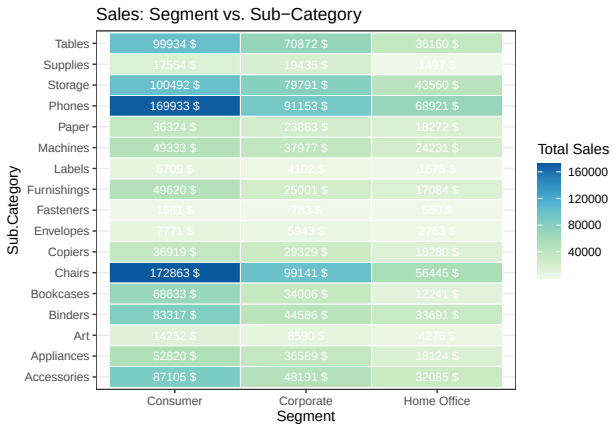
The above graph on the left show how the different Sub-Categories (grouped by Category) contribute to the Sales: the area of the rectangles is proportional to the Sales for each (Sub-)Category. The graph on the right is the same but relative to Profit instead of Sales.

We can notice how the furniture category contributes extensively on the sales, but less to the profit. Furthermore we get an idea of the best selling items: phones and chairs, as we will see, are the carriers of our sales.

Let's now compare Sales and Ship.Mode:

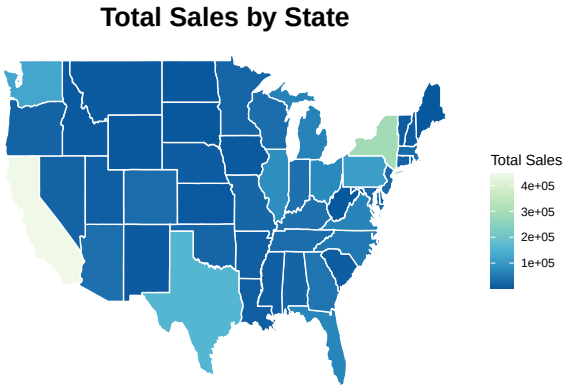


There is no preferred ship mode for different product category. We now compare Sales and Segment:



Once again, no clear differences of Sales between different Segments w.r.t. Sub.Category. It may look like the consumers tend to invest mainly on phones and chairs, but that is no different from corporates or home offices: the difference in colors lies in the difference in numbers. So no new information there.

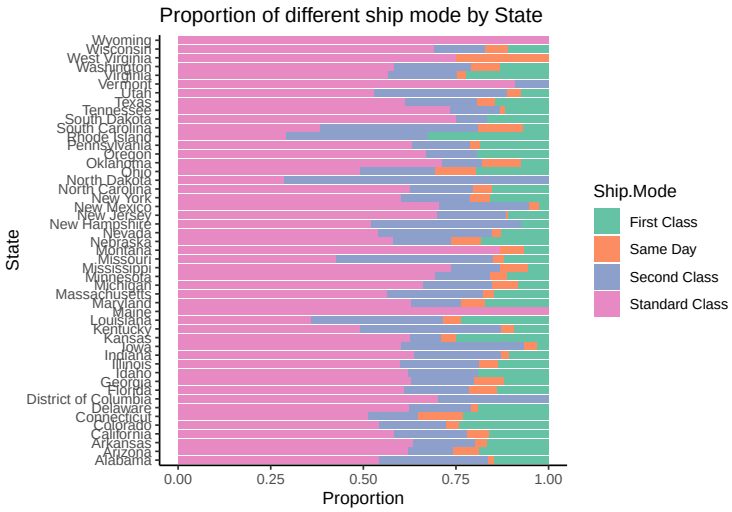
Let us focus on the State variable. Let's see how the Sales change depending on the state:



We can see that the state of California is the main contributor to our Sales, followed by the state of New York. Here is a list of the top 5 states by Sales:

```
## # A tibble: 5 x 2
##   State        S.Sales
##   <fct>          <dbl>
## 1 California   444416.
## 2 New York     287476.
## 3 Texas        158325.
## 4 Washington   124641.
## 5 Pennsylvania 108112.
```

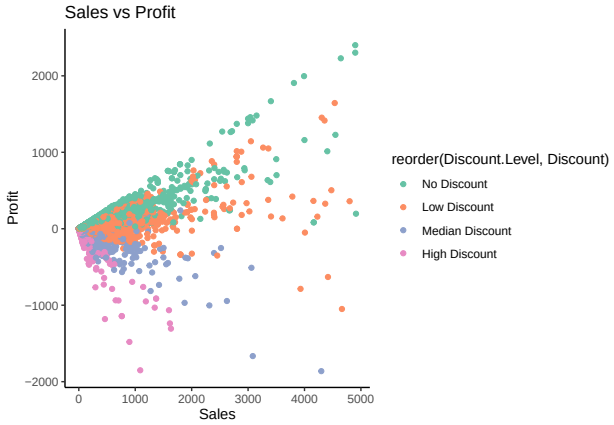
Let's see if there is any difference in the Ship.Mode between different states:

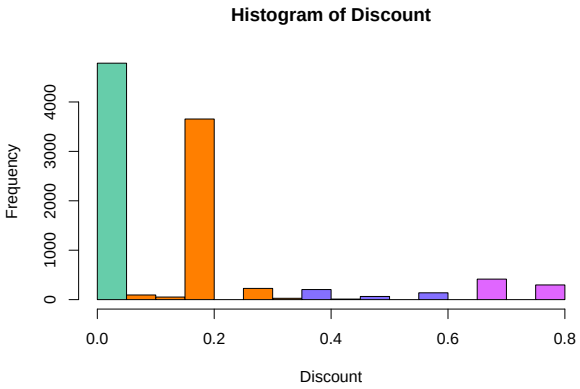


It looks like there are different preferences regarding ship mode between states. We recall that some states count very little samples, for example the state of Wyoming counts 22 samples and they happen to be all with Standard ship mode; states with more samples like California show a more balanced situation, but there still is a relevant difference, so we will take this into account in our model.

Now we will show arguably one of the most interesting plots, the Sales with respect to time:

A very interesting fact about this graph is that there seems to be a repeating pattern through different years. We can see that the blue curve has a repeating trend with period 4 (since we have 4 years, from 2018 to 2021), with a subtle increase through the years. The red curve is the Sales summed per month over the 4 years and it highlights the yearly pattern of the Sales. This curve, appropriately rescaled, can be used to make predictions about how many Sales are expected in the following months, so the superstore knows when to stock up on goods, and when fewer customers are expected.

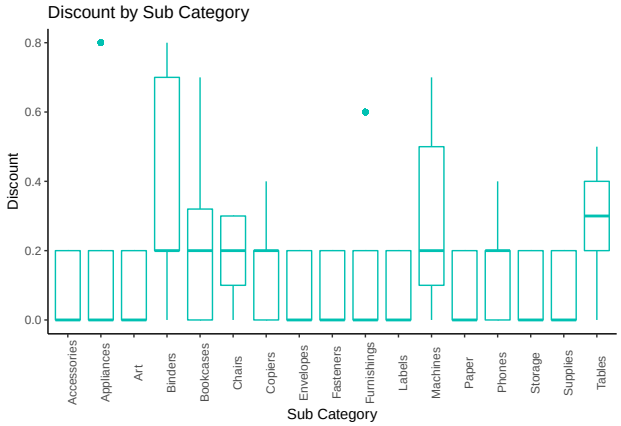


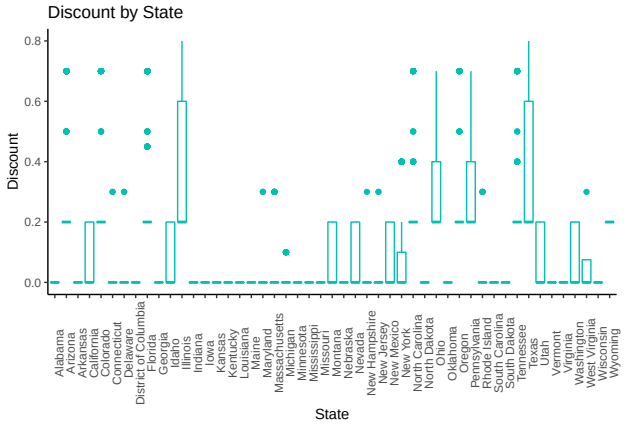


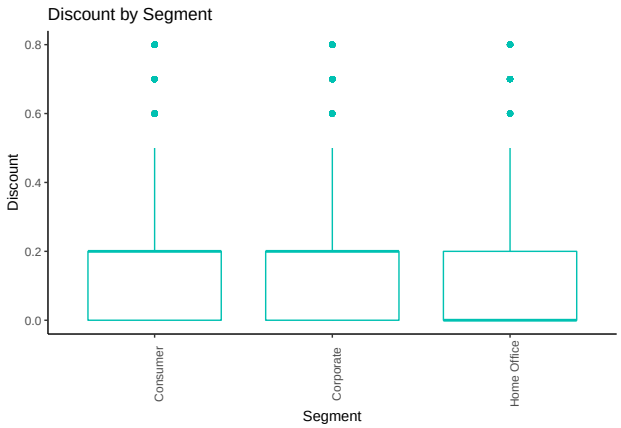
On the graph on the left we plotted Profit vs Sales, coloring the points according to the discount. What we expected was that the higher the discount, the lower the profit, which is in fact what we got, but an interesting phenomenon emerged: in some cases a low discount leads to more profit than no discount at all. This can be due to 2 factors: either the discounted products are popular and frequently bought (which would question the purpose of the discount in the first place), or more likely the low discounts are made in a smart way that makes people more prone to buying the products, maybe in bigger quantities.

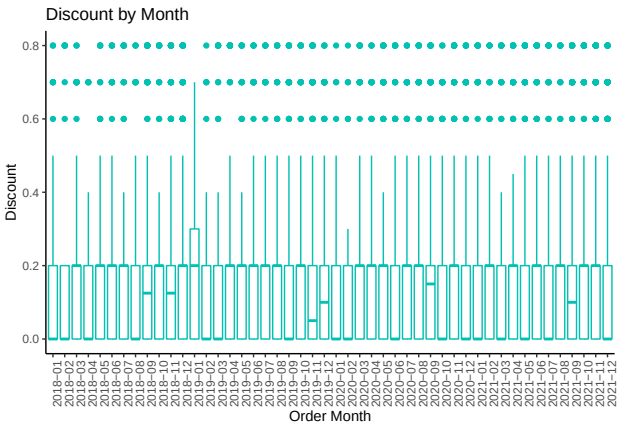
In conclusion, median and high discounts are not recommended because they lead to negative profit (they may have been chosen in the first place for some reason, for example to get rid of unsold products that were occupying place in the store, but that's information we don't have). Low discounts on the other hand can be effective without penalizing critically the profit.

The following boxplots are used to study possible links between Discount and other variables. There are no relevant results.

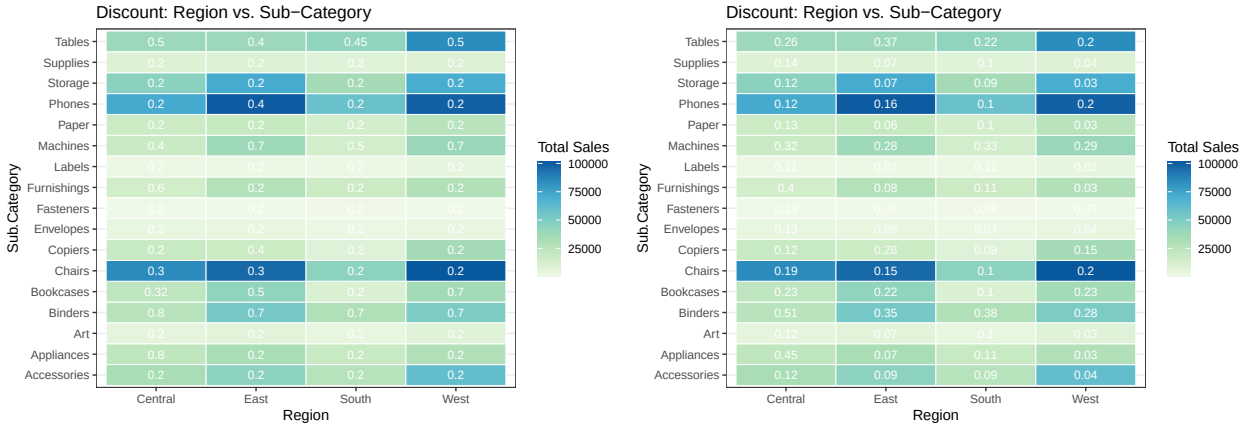




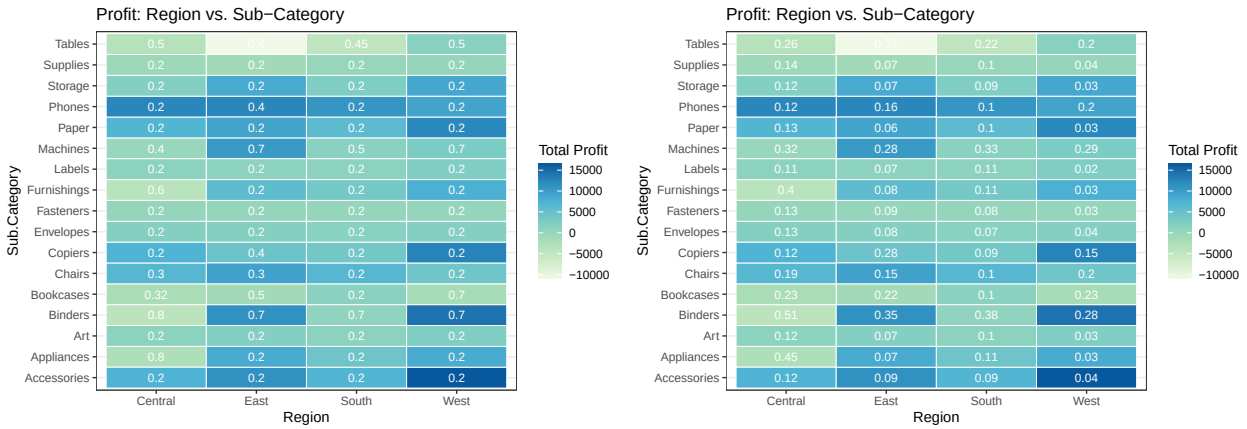


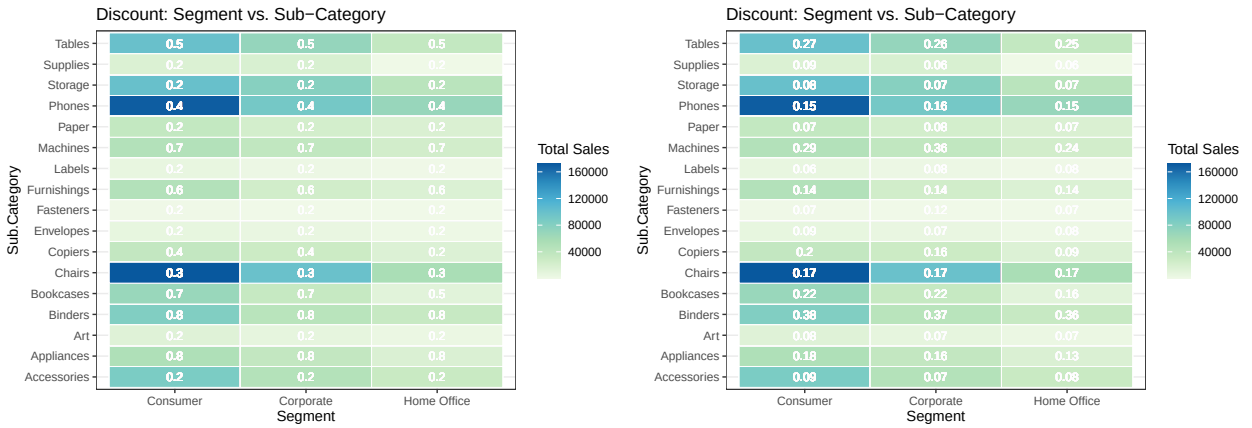


Below we show Region and Sub.Category on the x and y axis, Sales with the color gradient and discount with the number (max discount on the left, average discount on the right). We didn't get new information from this representation, but we chose to include it as it is a nice compact visualization of multiple different phenomenons - for example, east and west count more samples than north or south, chairs and phones are the main incomes of our sales, and items with low discount sell more successfully than items with no discount.



Below the exact same graphs but with Profit as color gradient instead of Sales.





Modelling

We recall that our objective is predicting the Sales with respect to the other variables. In this section we will attempt to build such model.

Model 1: based on previous results

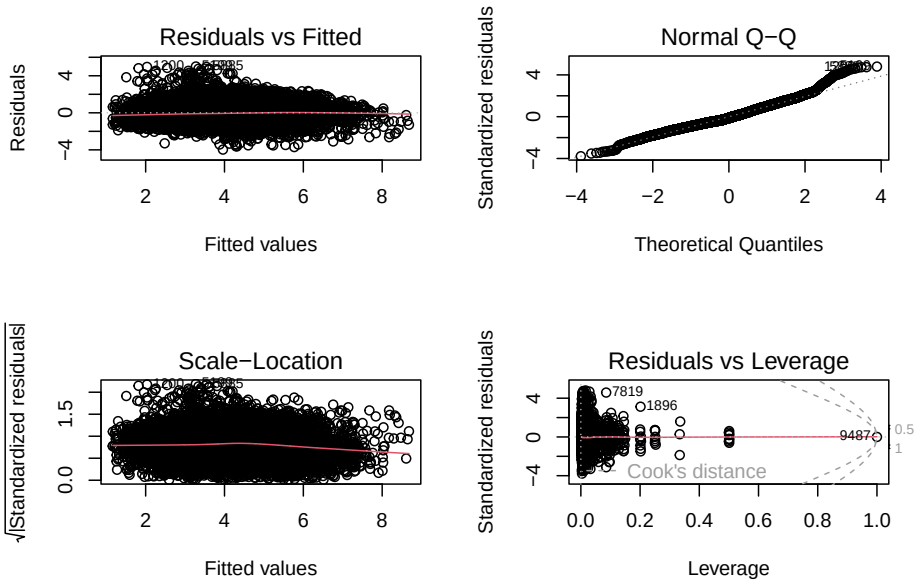
The relevant information for the Sales, from our provious analysis, comes from the variables State, Sub.Category, Order.Date.M, Discount.Level, Quantity, in addition we add interaction variables with Discount.Level:Sub.Category and Discount.Level:State.

```
lm.mod1 <- lm(log(Sales)~State+Sub.Category+Order.Date.M+Discount.Level+Quantity
              +Discount.Level:Sub.Category+Discount.Level:State
              ,sub.data)
```

Our dataset is composed mainly of categorical variables, which isn't ideal for fitting a linear model since the vectorized categorical variables get turned into thousands of (binary) vectors. Nontheless we gave it a try; the summary() function returns the following results:

```
## Residual standard error: 1.052 on 9830 degrees of freedom
## Multiple R-squared: 0.5887, Adjusted R-squared: 0.5829
## F-statistic: 102 on 138 and 9830 DF, p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm.mod1)
```



Model 2: constraint backward elimination

We apply the backward elimination process, and we check the F-value given by the anova function at each step.

We start by defining the model:

```
mod.F <- lm(log(Sales)~State+Sub.Category+Order.Date.M+Discount.Level+Quantity
            +Discount.Level:Sub.Category+Discount.Level:State
            ,sub.data)
```

Then we iterate the following process until the anova function shows evidence of a relevant change in the model:

```
mod.R <- update(mod.F, .~.-Order.Date.M)
anova(mod.R, mod.F)
```

```
## Analysis of Variance Table
##
## Model 1: log(Sales) ~ State + Sub.Category + Discount.Level + Quantity +
##     Sub.Category:Discount.Level + State:Discount.Level
## Model 2: log(Sales) ~ State + Sub.Category + Order.Date.M + Discount.Level +
##     Quantity + Discount.Level:Sub.Category + Discount.Level:State
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1   9841 10889
## 2   9830 10876 11    12.429 1.0212  0.424
```

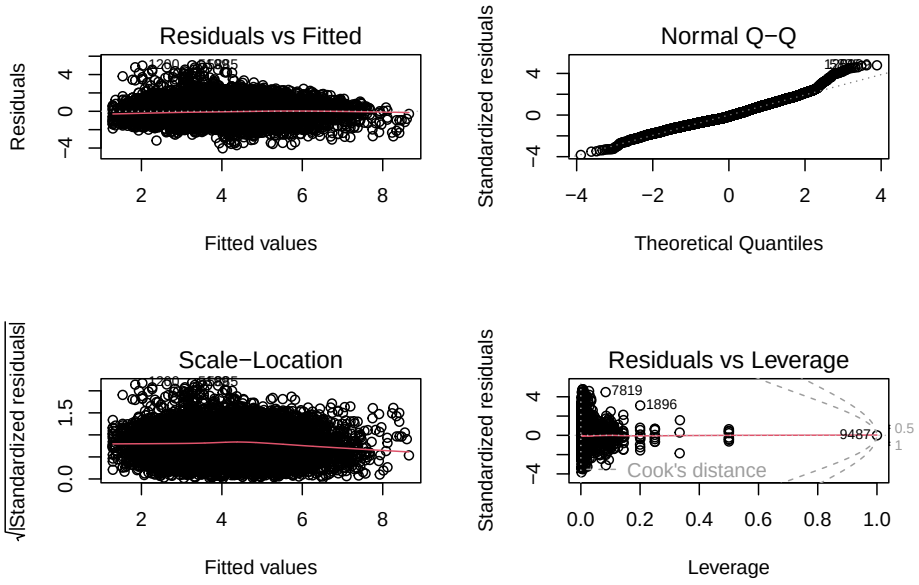
At the end of the process, we get rid of the variables Order.Date.M and Discount.Level:Sub.Category, and end up with the following model:

```
con.best <- lm(log(Sales)~State+Sub.Category+Discount.Level+Quantity
               +Discount.Level:State,sub.data)
```

Output of summary():

```
## Residual standard error: 1.052 on 9865 degrees of freedom
## Multiple R-squared:  0.5871, Adjusted R-squared:  0.5828
## F-statistic: 136.2 on 103 and 9865 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(con.best)
```



Model 3: Lasso Regression

Since our model has way too many variables, we chose to try a Lasso regression to enforce sparsity in the model. To choose the hyperparameter of the Lasso we use cross-validation.

Firstly we design a matrix:

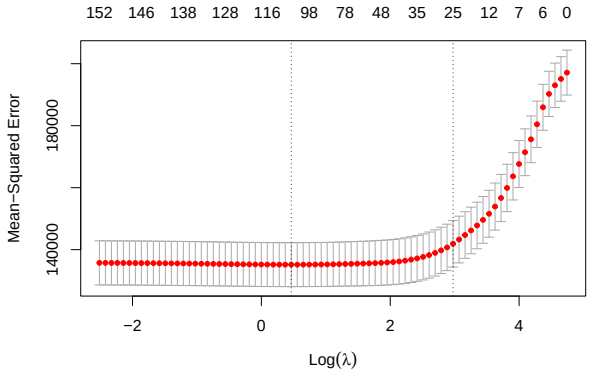
```
library(glmnet)
# design matrix
X <- model.matrix(log(Sales)~State+Sub.Category+Order.Date.M+Discount.Level+Quantity
                  +Discount.Level:Sub.Category+Discount.Level:State, sub.data)

# remove the first column relative to the intercept
X <- X[,-1]

# vector of responses
y <- sub.data$Sales
```

We then proceed by splitting the dataset into training and validation set. Once that is done, we use the function glmnet with alpha=1 to apply the LASSO regression.

```
# use 10 folds cross-validation to choose the value of lambda
cv.out <- cv.glmnet(X[train, ], y[train], alpha = 1, nfold=10)
plot(cv.out)
```



We then identify the best lambda and calculate the MSE:

```
# identify the best lambda value estimated test MSE
bestlam <- cv.out$lambda.min
bestlam
```

```
## [1] 1.58976
```

Finally, we fit the LASSO with the best found coefficient on the full data (code not shown here) to then compare it with the model 2.

Comparing model 2 and model 3

Here we will use the MSE for comparing the last two models:

Model 2:

```
con.best <- lm(log(Sales)~State+Sub.Category+Discount.Level+Quantity
               +Discount.Level:State,sub.data)
con.pred <- predict(con.best, newdata=sub.data)
mean((con.pred-y)^2)
```

```
## [1] 228727.8
```

Model 3:

```
# estimate the test MSE with the best lambda
lasso.pred <- predict(lasso.mod, s=bestlam, newx=X[test,])
mean((lasso.pred-y.test)^2)
```

```
## [1] 104648.3
```

The MSE shows a lower value for the LASSO model. This doesn't necessarily mean that the LASSO model is the best one between the two; in this case, the nonzero coefficient variables contain the interactive terms Sub.Category:Discount.Level and State:Discount.Level. If we keep consistent with the hierarchy principle, the model should contain the same variables as our model 1. Thus, the result of the LASSO is not faithful to our situation.

Model 4: A simpler model

Lastly, we tried a simple model with only 3 variables: Sub.Category, Discount.Level and Quantity. This reveals to have competitive performances with respect to the other models:

```
mod <- lm(log(Sales)~Sub.Category+Discount.Level+Quantity
              ,sub.data[train,])
```

```
mod.pred <- predict(mod, newdata=sub.data[test,])
mean((mod.pred-y.test)^2)
```

```
## [1] 182739.8
```

Conclusions

The dataset we chose, as we saw, is composed of mainly categorical variables, with little numerical variables. For this reason it proved to be well suited for qualitative analysis with graphs, boxplots, histograms and not as suited for numerical analysis and modelling. Nontheless, we still managed to find a decent model for the Sales given the remaining informations, which revealed to be a very hard task since we are using mainly categorical variables to explain a numerical one.

The most important results were found in the graphical part, here we will sum up the main points:

- mid-high discounts (30% and more) do not favour profit and should not be applied by the superstore, unless for necessity reasons (e.g. lack of space in the store).
- low discounts (10% - 20%) instead proves to be effective in increasing the sales without damaging the profit, which may also improve the popolarity of the superstore, so it is highly recommended to keep a low discount on some products, especially the most competitive ones.
- As we saw from the monthly sales graph, there is a pattern that repeats each year, showing a drop in sales in January-February, and peaks in the months of September, November and December. We have some hypothesis on why this is the case, but the main point is that the superstore knows in advance when there's going to be an increase in sellings and can stock up accordingly. A further analysis could give more details on the kind of products that are needed based on the time period.
- The main states which are buying from the superstore are California and New York. We don't know the location of the superstore, but let us assume the main store is in California. Then it would be recommended to focus advertisement in California and neighbouring states to cut on transportation costs and have more competitive prices. Creating branches in the other best selling states, such as New York, Texas and Washington may also be a valuable investment.