IDS 570
STATISTICS FOR MANAGEMENT
FINAL PROJECT REPORT

MANUEL HUERTA ROJAS-660866129
DEVYANI SINGH-654766729
TARUN KHURANA-667683595
MEENAKSHI PARAMESWARAN-673380596
JOSHUA ROSS-674456829

Executive Summary:

The data for this project was provided by Manuel from the company he works in. This company deals in electric heating items and the data was sales data for the years 2012-2015. This report gives a summary of statistical modelling for profit percentage for various items of this company along with the implementation and analysis of this model with linear and multilinear regression techniques.

The initial dataset comprised of 21126 observations with 20 variables.

We merged 3 datasets to add 4 new variables (Average Temperature, Distance from Illinois, Income per Capita and Number of Households) to our dataset. We added three extra calculated variables (customer status, profit and profit percentage) to our dataset and also changed some variables like year from numeric type to factor to better affect the dataset. After merging and cleaning, the final dataset contains 20435 observations and 25 variables. 5 independent variables have been chosen from this final dataset to model the dependent variable ("profit percentage") as per the research question.

The univariate statistics summarizes the distribution of profit percentage and all independent variables across the data.

The bivariate analysis on dependent variable, profit percentage and each independent variables was conducted.

Welch two samples t-test, correlation-test and ANOVA are used to determine the effect of independent variables on the dependent variable i.e. profit percentage. The relationship between the dependent variable and independent variables have been displayed by various plots.

Finally, linear regression and multilinear regression models are proposed for modelling the profit percentage with the independent variables that were hypothesized to possibly have an influence on the profit percentage.

The results agree with the earlier tests (ANOVA, t-test, and correlation-test) and they confirm that the three hypotheses out of 4 are true.

Introduction:

About the company:

We are using the sales data of a retail and wholesale company of electric heating products which is located in the Northwest suburbs of Chicago.

About the dataset:

The sales data of the company for years 2012-2015 was obtained from the workplace of one of our team members (Manuel Huerta Rojas) where it was collected in a couple of ways as listed below:

- The sales representatives entering data over the phone or via the emails from clients.
- The clients creating data entries directly into the database via online shopping portals.

The data had 21126 observations and 20 variables in raw form.

Details of the Variables:

FIELD	CONTENT	DATA TYPE
Company	Company the Data is from	factor
client_id	Unique Identification of a Client	integer
sales_order	Unique Identification of a Sales Order	integer
Item	Unique Identification of an Item	integer
Sale	Total Revenue (Qty per Unit Price)	numeric
Cost	Total Cost of Goods (Qty per Unit Cost of Goods)	numeric
dist_channel	Distribution Channel (Direct Sales, Contractors etc.)	factor
product_line	Product Line (Outdoor , Indoor)	factor
Year	Shipping (Sale) date's year	integer
Per	Shipping (Sale) date's Month	integer
Qtr	Shipping (Sale) date's Quarter	integer
Wk	Shipping (Sale) date's Week	integer
St	State 2 Digit Code where the Order ships	factor
Region	Economic Region (Grouping of states)	factor
rep1_ord	Unique Identification of a sales representative	integer
cust_year	Customer Creation date's Year	integer
cust_per	Customer Creation date's Month	integer
Ctry	Country where the Order ships	factor
Sqft	The volume of the item in square feet	integer

Working on the data:

The following steps were taken before starting to work on the data

Step 1: Cleaning the data

- Removing NA's and blank values
- Dropping unused columns like company name, week, product line etc.

Step 2: Adding the data

- Calculated some variables ,like customer status, profit and profit percentage, based on the existing data for better analysis of our Hypotheses
- Merging different datasets like **average temperature, distance from Illinois, income per capita, number of households and item type** to work on our mentioned hypotheses

Step 3: Tidying the data

- Changing variables types like client_id, sales_order and item from numeric to character and variable types like dist_channel, year, per, qtr, cust_year, cust_per from numeric to factor to better suit our analysis.

Research Question:

What factors affect company's profit percentage for the years 2012-2015?

Hypotheses:

- **Average Temperature** may have an effect on profit percentage as the company acknowledges the high seasonality of the sales during cold weather.
- **Square Feet (size of product)** may affect profit percentage because bigger products or orders usually have a higher discount as per company policy or due to negotiation.
- **Customer Status (New and Old customers)** may affect profit percentage because returning customers are usually provided with a higher discount tier.
- **Distribution channels** may have an effect on profit percentage because different distribution channels have different price listings and discount tiers.
- **Item type** is to have an effect in profit percentage as loose cable is priced less per same area compare to cable attached to a mesh.

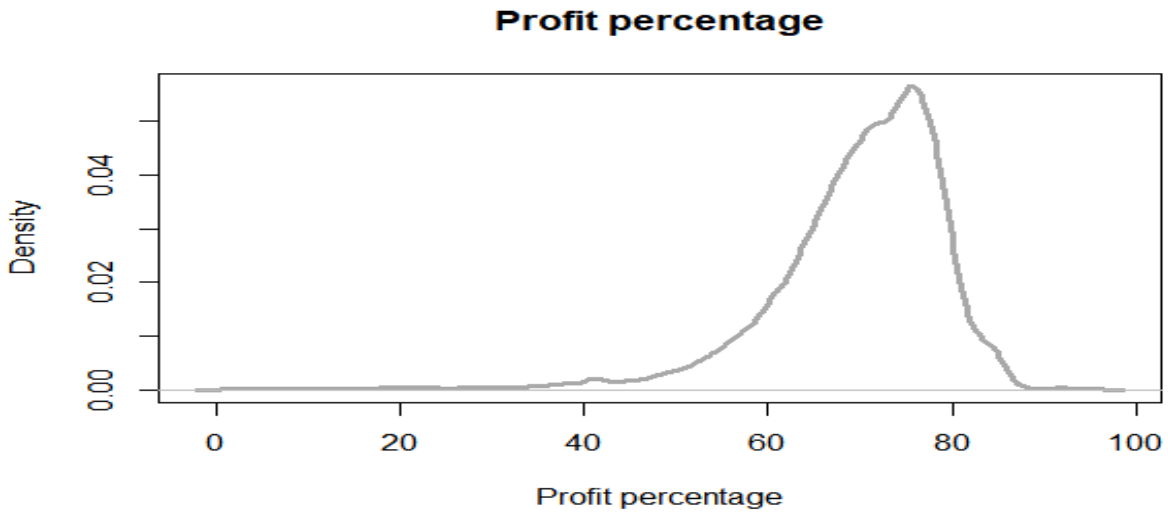
Limitations:

The dataset did not have the all the factors affecting profit percentage as we believe there can be some additional factors like discount percentage which may affect profit percentage of an electric heating company.

Univariate Analysis:

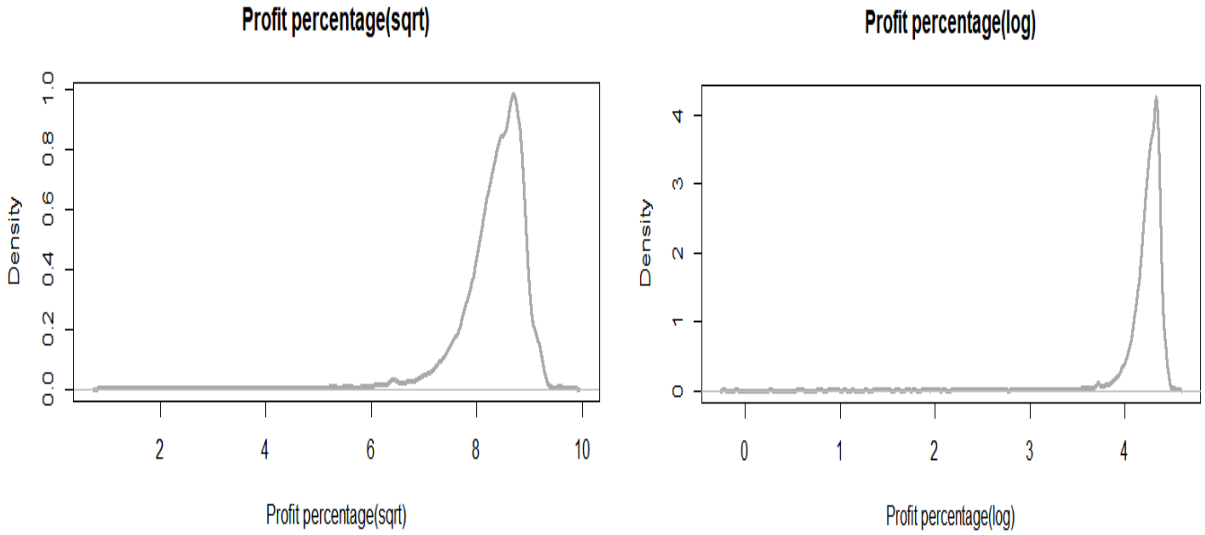
Dependent Variable:

Profit Percentage



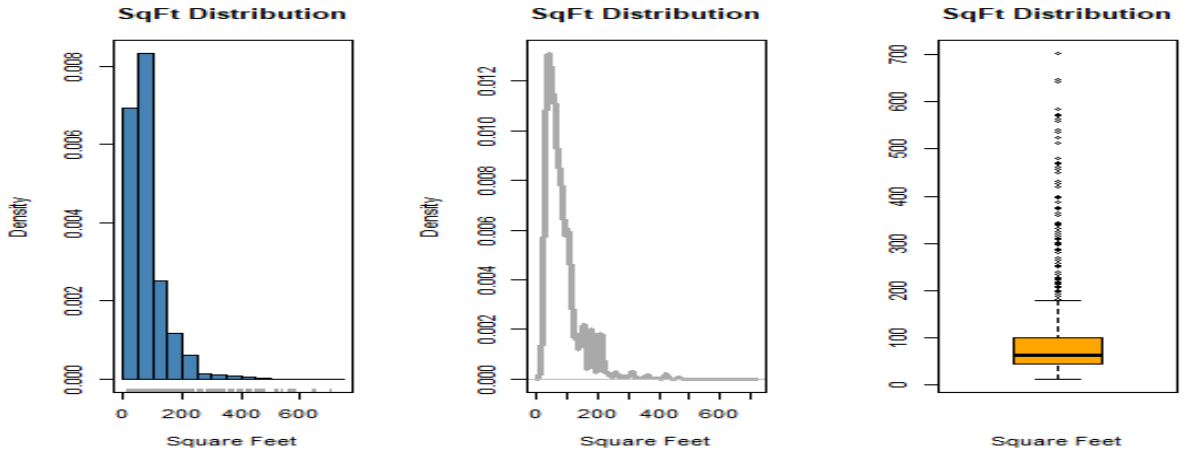
The plot for profit percentage is slightly left skewed as a mean of 69.7 is smaller than the median (71.45) with skew value of -1.98 and kurtosis value of 7.49, which in turn is higher than a normally distributed plot.
Also, the outliers for the dependent variable exists in the range of 0 to 40 but the range of values is 95.01 with maximum values lying in the region 40 to 95.

NOTE: We tried transforming the dependent variable using the sqrt and log functions but found that the original was the best fit, that is, the closest to a normally distributed plot.



Independent Variables:

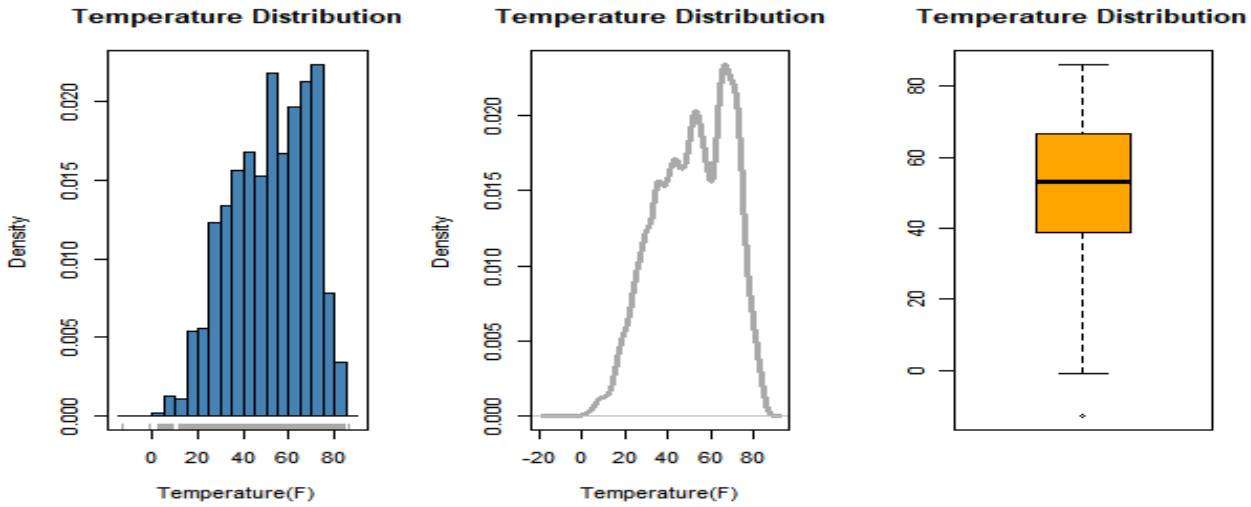
Square feet (sqft)



The median is smaller than the mean, so it is right skewed and the kurtosis is 11.35 which is very far from a normally distributed plot. Also, the skewness is 2.69 which shows that it is heavily skewed.

The range is 690, between 12 and 702 with a lot of outlier as visible in the boxplot.

Average Temperature (avg_temp)



The median is slightly bigger than the mean, so it is left skewed and the kurtosis is 0.78 which is quite closer to a normally distributed plot. Also, the skewness is -0.31 which shows that it is only slightly skewed.

The range is 98.7, between -12.8 and 85.9.

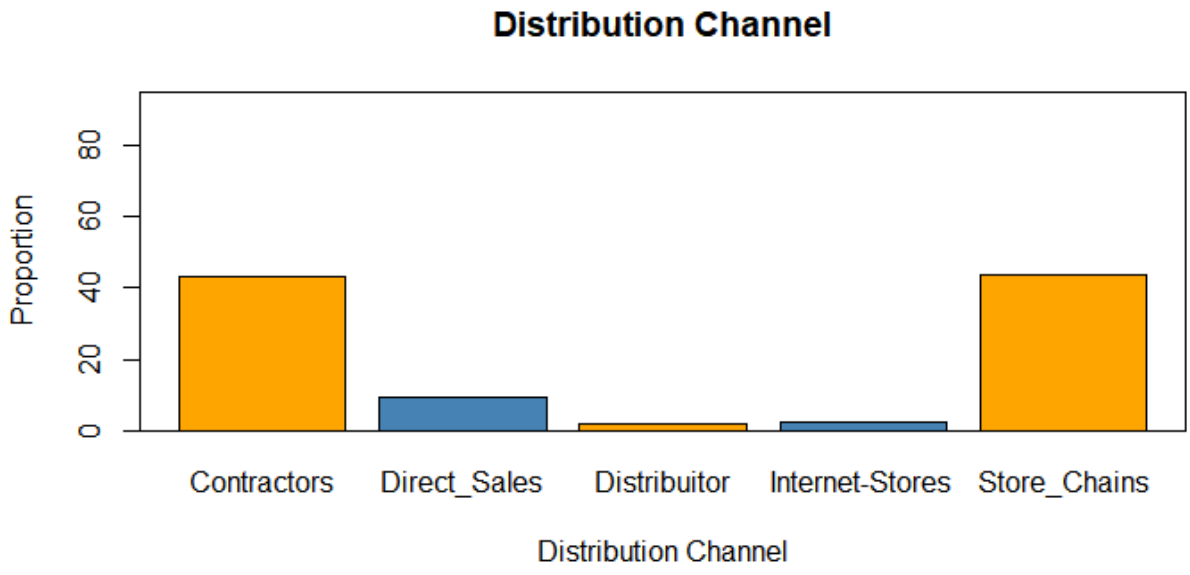
NOTE: As it is multi modal graph, we tried to convert it into a factor of three different temperature levels during the bivariate analysis.

Customer Status (cust_status)

Customer status is identified by comparing item's shipping month and year to customer's creation month and year.

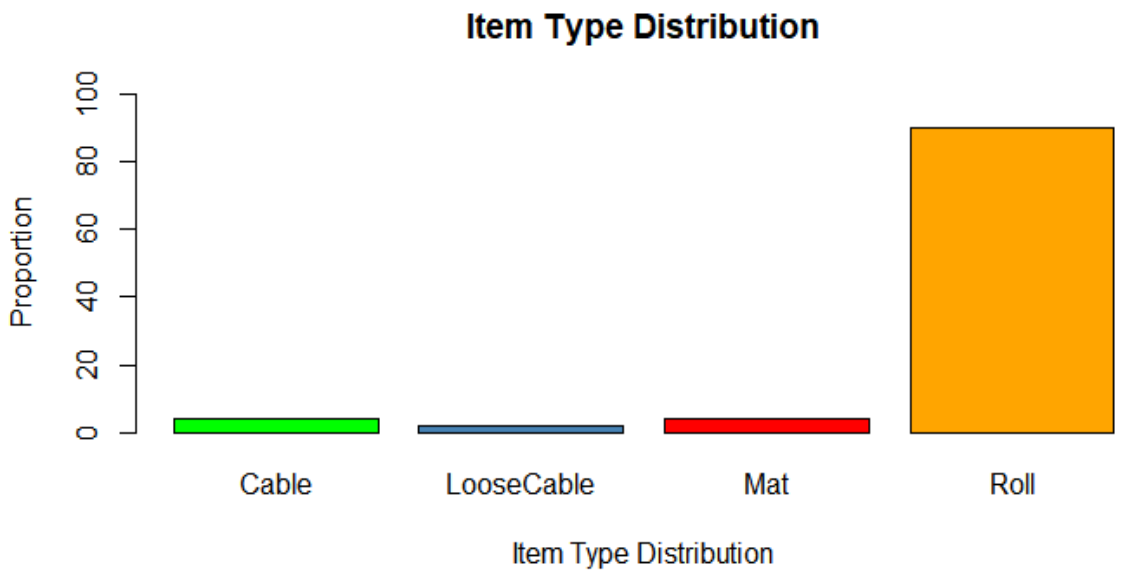
Returning customers hold the majority share which is nearly 9 times the new customer's share.

Distribution Channel (dist_channel)



From the plots we see that Store chains and Contractors have nearly equal share and they hold the majority share.

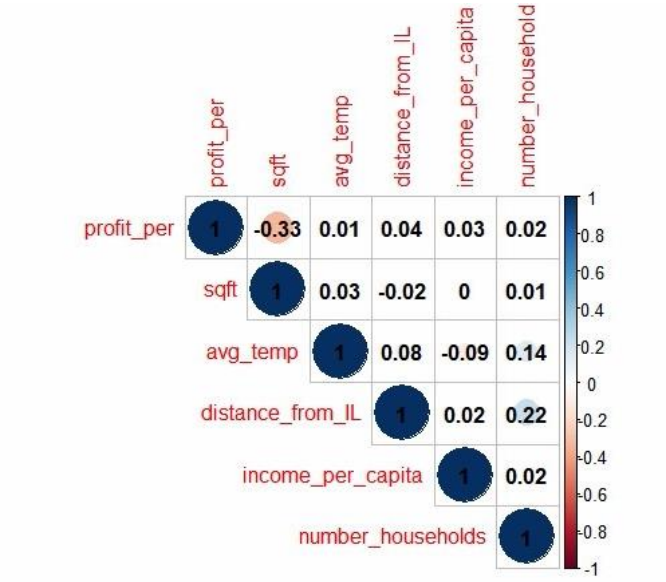
Item Type (item_type)



Roll Item type holds the majority share close to 90% of the total item type distribution. Also Cable, Loose Cable and Mat item types hold close to 2-3% each of the total item type distribution.

Bivariate Analysis and Hypothesis testing

Numerical Variables

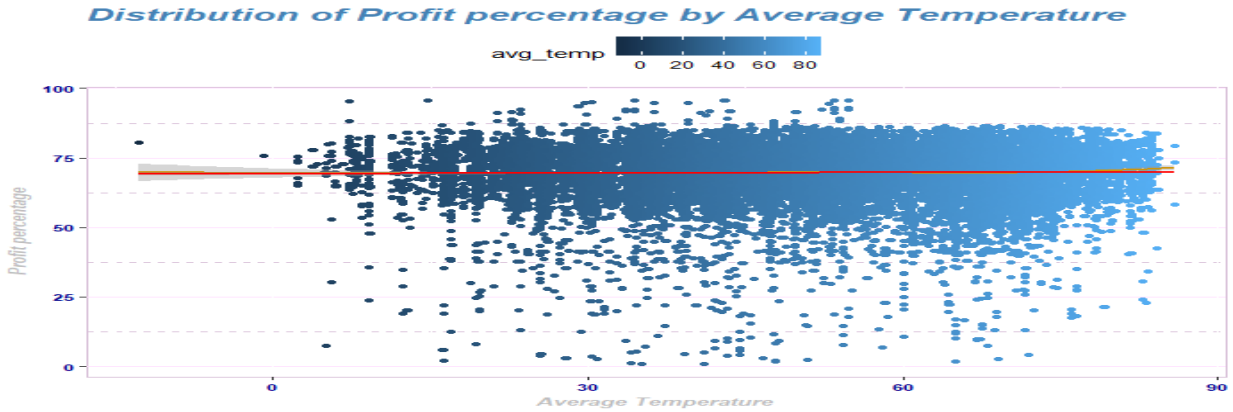


The correlation matrix above shows the correlation of numerical variables in the dataset. It can be found that besides square feet, there is not much correlation of other numerical variables with profit percentage in the data set.

Square feet have the highest correlation of 0.33(33%) and the average temperature has the lowest correlation of 0.01(1%) with profit percentage.

Profit Percentage (Dependent) and Average Temperature (Independent)

ANALYSIS



The graph is a horizontal linear model that fits the data. Profit percentage seems independent of average temperature.

To confirm this analysis, a cor.test was performed on profit percentage and average temperature to see whether average temperature should be included for further analysis.

The null hypothesis is that there is no relationship between the two variables, profit percentage and average temperature.

```
Pearson's product-moment correlation

data:  NewFinalData$profit_per and NewFinalData$avg_temp
t = 1.8306, df = 20433, p-value = 0.06718
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0009058465  0.0265114245
sample estimates:
      cor
0.0128052
```

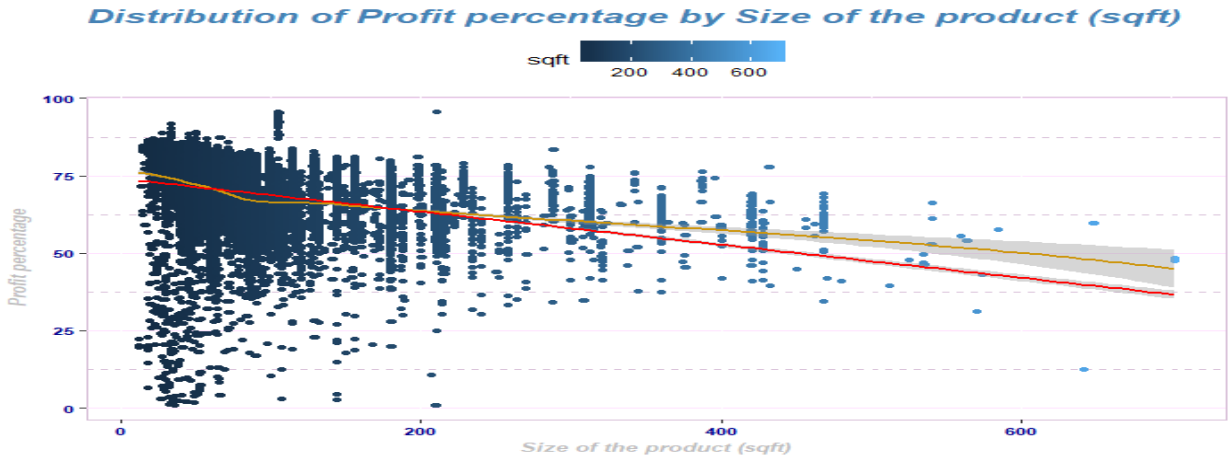
From the test we find that p value is high (>0.05), hence, we fail to reject the null hypothesis and we will not consider this variable for the model.

The correlation between profit percentage and average temperature is 1.28% which is a very low correlation.

Even though we thought that the customer state's average temperature will have the greatest effect on profit percentage, (since it's a heating equipment, we thought more sales will happen when temperature is low) from the analysis and testing we find that it has no effect on the profit percentage of the company. This means that irrespective of the temperature or time of the year, sales are happening for the company.

Profit Percentage (Dependent) and Square Feet (Independent)

ANALYSIS



From the graph, profit percentage and size of the product purchased (sqft) seems negatively correlated to each other. From the graph, we also see that smaller products are sold more in number as data points are concentrated in the range 0-200. The graph depicts a downward trend due to its outliers as size of the product increases, profit percentage decreases.

To confirm this analysis, a cor.test was performed on profit percentage and size of the product purchased (sqft) to see whether square feet should be included for further analysis.

The null hypothesis is that there is no relationship between the two variables, profit percentage and square feet

```
Pearson's product-moment correlation

data:  NewFinalData$profit_per and NewFinalData$sqft
t = -49.293, df = 20433, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3382015 -0.3136936
sample estimates:
       cor
-0.3260023
```

From the test we find that p value is low (<2.2e-16). We reject the null hypothesis and hence, we will consider this variable for the model.

The correlation between profit percentage and square feet is -32.6% which is a good correlation even though the correlation value is negative.

The correlation test proves that there is significant difference in profit percentage with size of the product sold. The test also shows that they are correlated negatively.

This is because when size of the product (sqft) increases (i.e., bigger products or products in bulk), higher discounts are given as per company policies or due to negotiation. The high discount tier causes company's profit percentage to come down. The company gets more profit percentage when they sell small products as the discounts given to them are low.

Categorical Variables

Profit percentage (Dependent) and Customer Status (Independent)

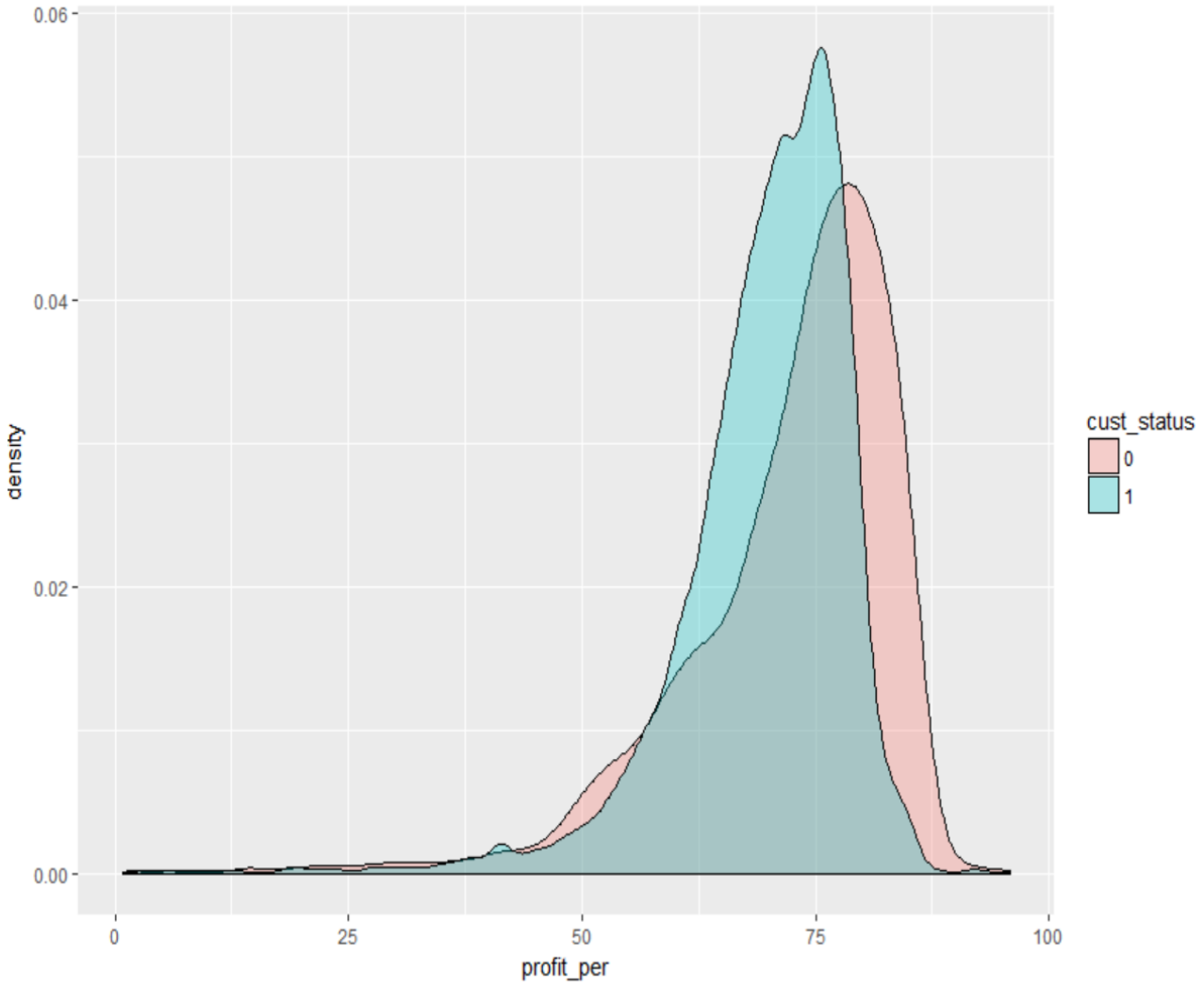
STATISTICAL ANALYSIS

New customers have a higher profit percentage, some customers will be granted a higher discount level after the first purchase, and therefore this variable is likely to have an impact in profit percentage.

Customer Status	Status Indicator	Average Profit %	Median Profit %	Std.Dev. Profit %
New	0	72.0372	75.0155	11.6438
Returning	1	69.4447	71.2174	9.6644

PLOTTING ANALYSIS

This density plot help us visualize the mean of new customers (type=0) to the right of the mean of the recurring customers (type=1)



RESULTS OF T-TEST

We are comparing the median of the new customers against the median of the recurring customers to see if the former is greater than the latter. The P-Value is low, hence we reject the null and accept the alternate hypothesis that the mean of the new customers is greater than the mean of the recurring customers.

We will include this variable in our regression model because we know of the company policy to grant higher discounts for returning clients.

```
One Sample t-test
t = 9.7513, df = 1925, p-value < 0.00000000000000022
alternative hypothesis: true mean is greater than 69.45
95 percent confidence interval:
71.60058      Inf
sample estimates:
mean of x
72.0372
```

Profit percentage (Dependent) and Distribution Channel (Independent)

STATISTICAL ANALYSIS

Direct-Sales generate the highest average and median profit percentage of all the channels. Items sold through the Direct-Sales channel are usually sold to end users at price list.

Internet Stores and Distributors contribute the least average profit percentage, these channels are granted a higher discount level because they in turn need to resale to individuals and other retailers.

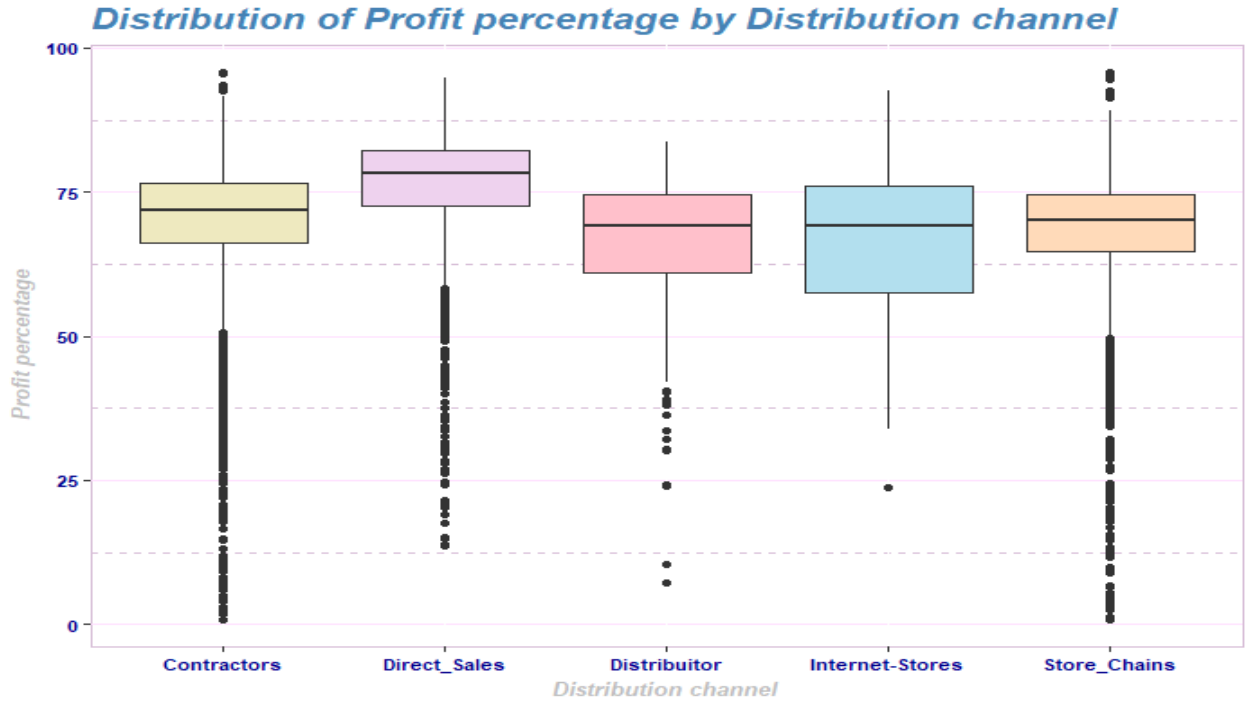
Contractors are in the middle range of average profit percentage, they are usually professionals that buy the product to sell it to end users as part of their contracting work.

Distribution Channel	Average Profit %	Median Profit %	Std.Dev. Profit %
Contractors	69.9824	71.9194	9.8424
Direct Sales	75.6527	78.2659	10.3744
Distributor	66.0446	69.0338	11.8912
Internet Stores	66.3449	69.1152	10.7471
Store Chains	68.4317	69.9977	9.1733

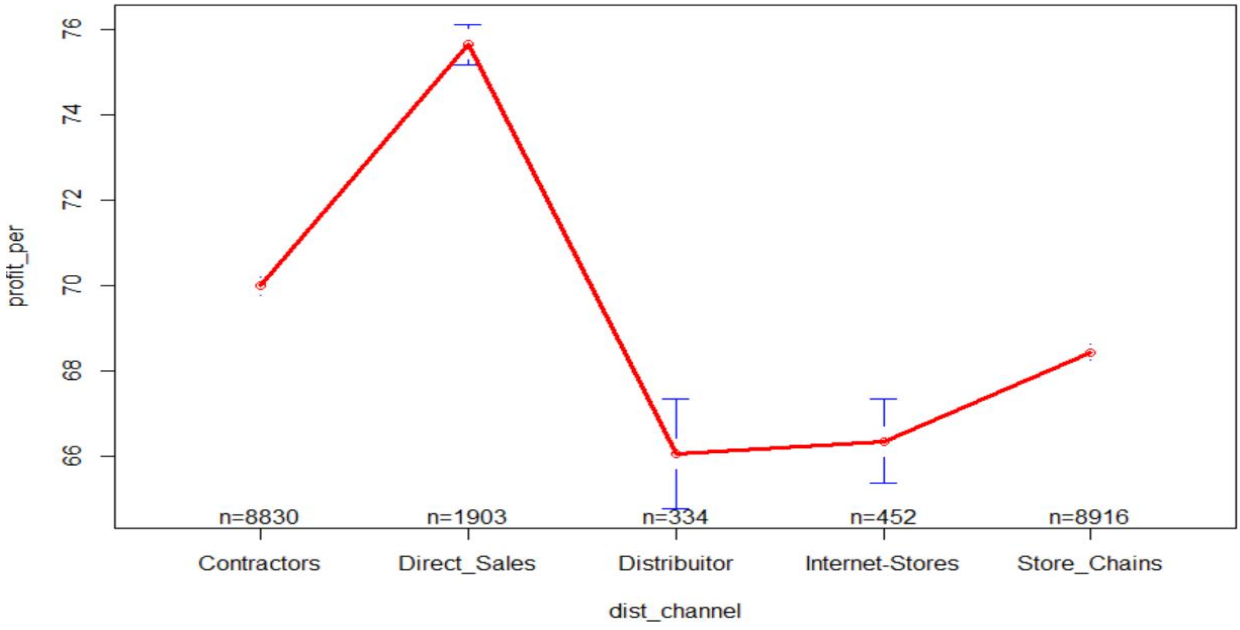
PLOTTING ANALYSIS

The profit percentage IQR of internet stores is the biggest, but at the same time this channel does not have as many outliers as the other channels, suggesting a more even pricing and less additional discounts varying per transaction.

Store chains and Direct sales show the most extreme values of profit percentage.



Plotting the means we can see that there is a visible difference between the means of the distribution channels as only two channels, Distributor and Internet stores, seem to be have close means.



RESULTS OF ANOVA

The test returns a low p-value, hence, we reject the null hypothesis and accept that the means of the distribution channels are different.

The test that compares the groups show that 9 of the 10 comparisons have a p-value adjusted of zero.

Given the difference in the means of the groups we will consider this variable for our regression model.

Anova

```
                Df  Sum Sq Mean Sq F value              Pr(>F)
dist_channel     4   92026   23007   246.2 <0.0000000000000002 ***
Residuals    20430 1909362      93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tukey’s HSD

```
                                diff    lwr   upr p adj
Direct_Sales-Contractors        5.67   5.00  6.34  0.00
Distribuitor-Contractors       -3.94  -5.41 -2.47  0.00
Internet-Stores-Contractors    -3.64  -4.91 -2.37  0.00
Store_Chains-Contractors       -1.55  -1.95 -1.15  0.00
Distribuitor-Direct_Sales      -9.61 -11.17 -8.04  0.00
Internet-Stores-Direct_Sales   -9.31 -10.69 -7.93  0.00
Store_Chains-Direct_Sales      -7.22  -7.89 -6.56  0.00
Internet-Stores-Distribuitor    0.30  -1.60  2.20  0.99
Store_Chains-Distribuitor       2.39   0.92  3.86  0.00
Store_Chains-Internet-Stores    2.09   0.82  3.36  0.00
```

Profit percentage (Dependent) and Item type (Independent)

STATISTICAL ANALYSIS

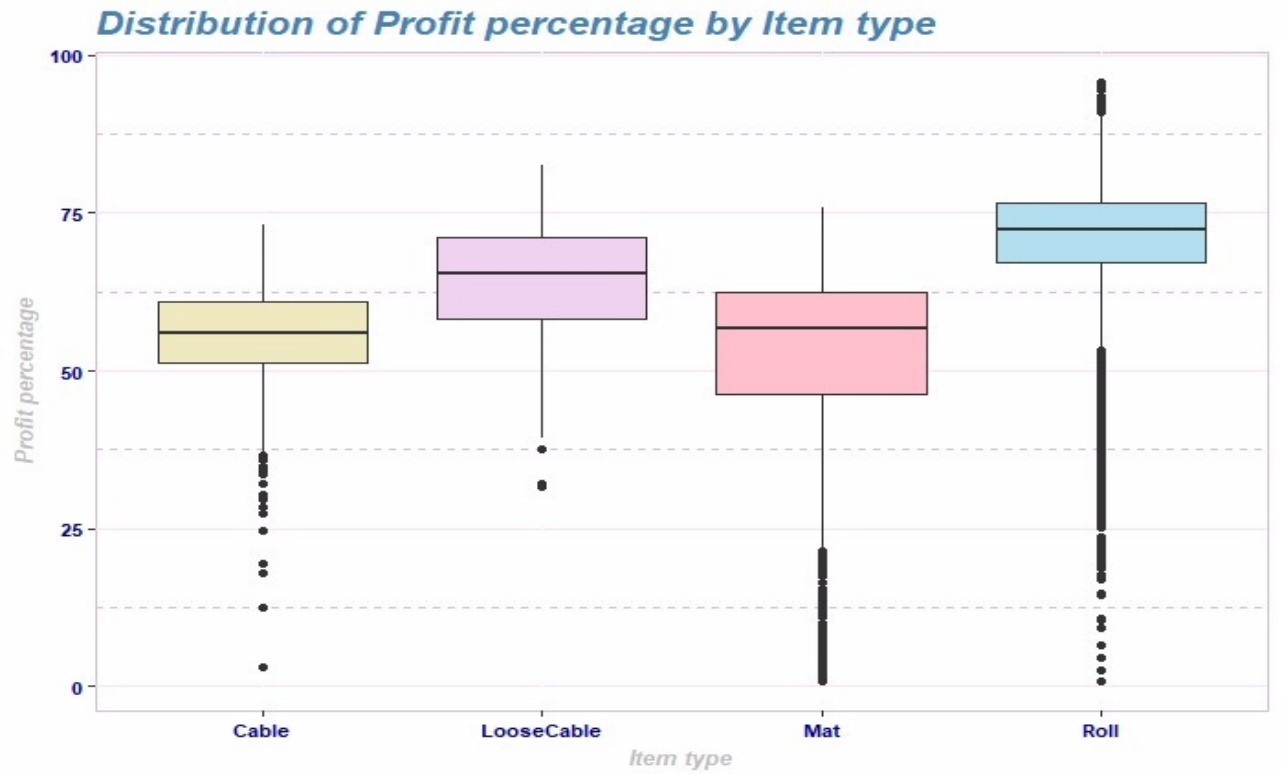
Indoor Heating-Roll has the highest profit percentage mean and median. This product is the same as the loose cable attached to a fiberglass mesh with the purpose to facilitate installation and reduce time. The product is priced more given this added feature.

Outdoor heating that comprise the Mat and Cable types have a marked lower profit percentage than indoor heating. The outdoor projects are usually larger in area. The products are also larger than indoor, but the price list has less profit markup.

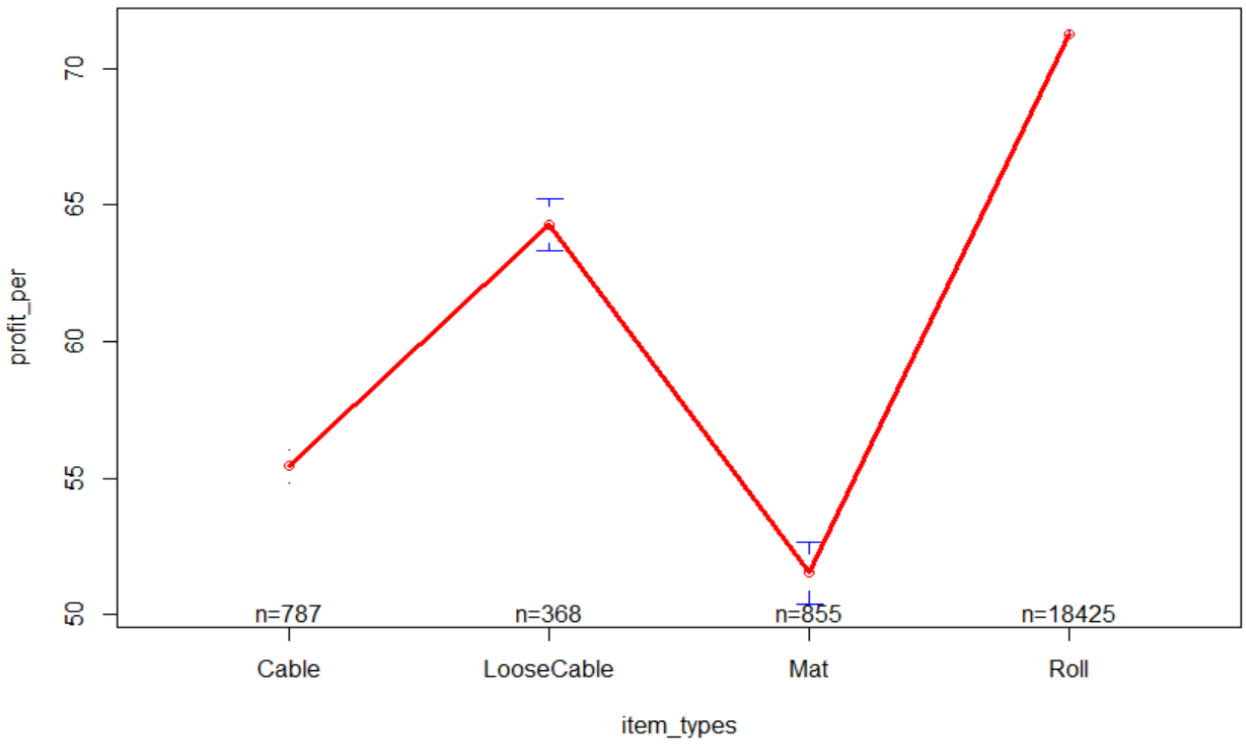
Item Type	Product Line	Average Profit %	Median Profit %	Std.Dev. Profit %
LooseCable	Indoor	64.2840	65.4499	9.2812
Roll	Indoor	71.2487	72.2778	7.9772
Cable	Outdoor	55.4341	55.9079	8.6198
Mat	Outdoor	51.5272	56.7229	16.8139

PLOTTING ANALYSIS

The profit percentage IQR of the Roll is the Mat (outdoor) is the biggest. Of all the item types, the loose cable (indoor) has the least outliers.



Plotting the means we can observe that there is a difference among all item types.



RESULTS OF ANOVA

The test returns a low p-value for this variable, we reject the null hypothesis and accept that the means of the item types are different.

All of the comparisons (6) show a p-value of zero, hence, we will consider this variable for our regression model.

Anova

```
               Df  Sum Sq Mean Sq F value             Pr(>F)
item_type       3  497514  165838    2253 <0.0000000000000002 ***
Residuals   20431 1503874      74
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tukey’s HSD

```
                    diff    lwr    upr p adj
LooseCable-Cable    8.85   7.46  10.24     0
Mat-Cable          -3.91  -5.00  -2.82     0
Roll-Cable         15.81  15.01  16.62     0
Mat-LooseCable    -12.76 -14.13 -11.38     0
Roll-LooseCable     6.96   5.80   8.13     0
Roll-Mat           19.72  18.95  20.49     0
```

We also tried converting the average temperature into 3 different levels to see if any relationship exists.

Temperature Level (3 level factor) and Profit Percentage

STATISTICAL ANALISYS

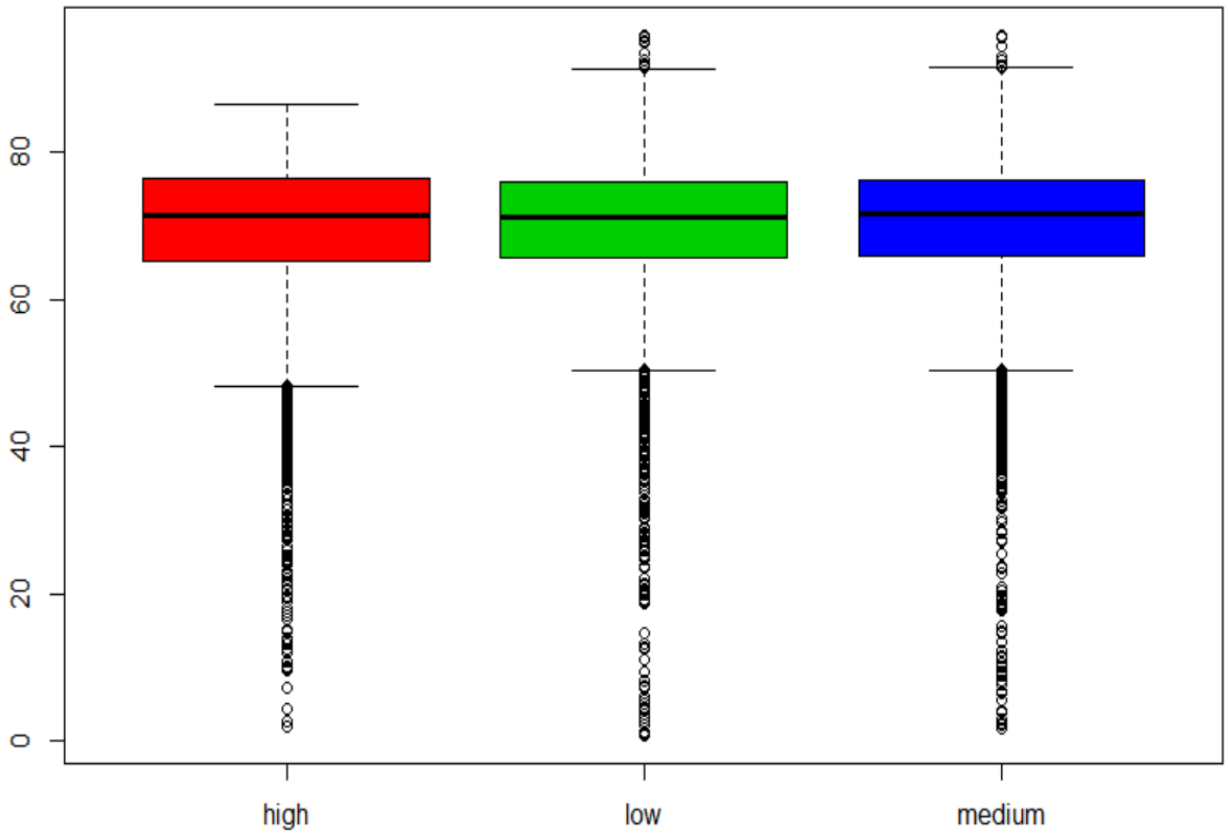
The average temperature variable is a numeric variable. We will create a categorical variable compressing the temperatures in three levels

- Low (less than 42 degrees), outdoor installation is not advisable below it
- Medium (between 42 degrees to 60 degrees)
- High (over 60 degrees), most hot months are during summer.

Temperature Level	Level Contents	Number Observations
Low	Less than 42 degrees	6,135
Medium	Between 42 and 60 degrees	6,677
High	Over 60 degrees	7,623

PLOTTING ANALYSIS

The three temperature levels are very similar, this might indicate low variability between the groups.



RESULTS OF ANOVA

The test returns a high p-value, we cannot reject the null hypothesis, and hence, we will not consider this categorical variable for the regression model.

Anova

```
              Df  Sum Sq Mean Sq F value Pr(>F)
temp_level     2     528  263.93   2.695 0.0676 .
Residuals   20432 2000860   97.93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tukey’s HSD

```
             diff   lwr  upr p adj
low-high    -0.08 -0.47 0.32  0.90
medium-high  0.30 -0.09 0.69  0.16
medium-low   0.38 -0.03 0.79  0.08
```

Linear Regression Models

Model 1 – Predicting Profit percentage with independent variable Square feet

For the first model, we use dependent variable profit percentage and independent variable square feet. From our bivariate analysis of these two variables, we had seen that they had a negative, or inverse relationship. As square feet increases, profit percentage decreases, and this is likely due to the bulk discounts that were offered by the company. Larger the size of the product the customer buys, the better the deal the company offers.

From the analysis we found out that at 95% confidence, the average profit percentage lies between 73.90% and 74.34%, and the average decreases at a rate of 0.056 and 0.051 per additional square foot. Although there was a confirmed correlation between the two variables, the regression model ended up having an adjusted R-squared of 10.62% making this correlation a weak one.

Model 2 - Predicting Profit percentage with independent variables Square feet and Item type

From the previous model we saw that the dependent variable profit percentage had a weak correlation with independent variable square feet. From the bivariate analysis we had seen that a relation exists between profit percentage and independent variable item type. So now we add item type to this existing model.

The item type had 4 item types with "Roll" item showing the highest profit percentage. So we relevel this factor. After releveling Item Type according to the "Roll" item and adding this in the existing model, the new regression model had an adjusted R-squared of 32.38%, i.e. an increase of more than 20%.

Model 3 - Predicting Profit percentage with independent variables Square feet, Item type and Distribution Channel

From the previous model we saw that adding item type in existing model of profit percentage and square feet showed an improvement of 21.7%, making the adjusted R-squared value as 32.38%. The bivariate analysis between Distribution Channel and Profit Percentage showed that the channel with the highest average Profit Percentage was "Direct Sales" followed by "Contractors", which showed us that a relationship existed between Distribution Channel and Profit Percentage. So now we add distribution type to this existing model.

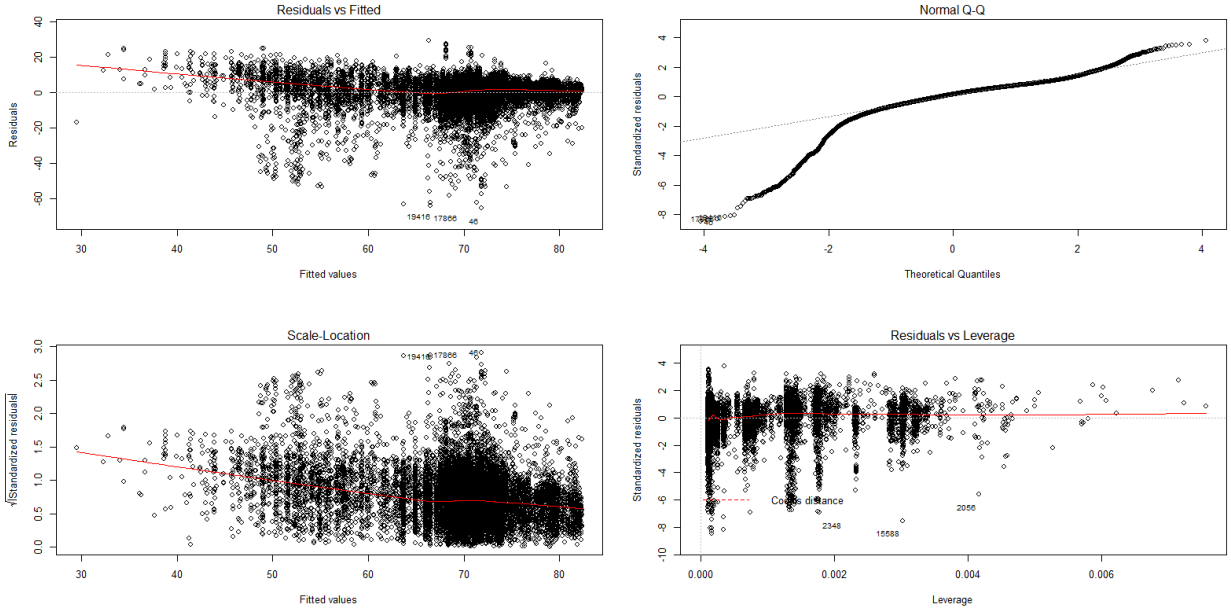
We now relevel the distribution channel factor with "Direct Sales" as it had showed the highest profit percentage. After releveling and adding this in the existing model, the new regression model had an adjusted R-squared of 38.96%, i.e. an increase of more than 6%.

Model 4 - Predicting Profit percentage with independent variables Square feet, Item type, Distribution Channel and Customer status

From the previous model we saw that adding distribution channel in existing model showed an improvement of 6.58%, making the adjusted R-squared value as 38.96%. From the bivariate analysis, Profit Percentage was shown to have a relationship with the level "Returning Customer" from Customer Status.

We now relevel Customer status with "Returning" status. After releveling Customer Status according to "Returning Customer", the variable was added to the previous model and the new adjusted R-squared was 39.37% which was a 0.40% increase.

After running regression diagnostics, we plot this model with the residuals.



Now we remove some of the outliers, and remodel.

Model 5 - Predicting Profit percentage with independent variables Square feet, Item type, Distribution Channel and Customer status – After removing outliers

After removing some of the outliers we try to remodel our previous model. We see that the new adjusted R-square value is 39.63% with an increase of 0.25% from the previous model. Hence, we see a minimal increase in adjusted R-squared even after removing some of the outliers.

We also checked for multicollinearity, but observed that the model was free from the risk of having multicollinearity.

To improve our model further, we tried to transform our variables.

Model 6 - Predicting Profit percentage with independent variables Log of Square feet, Item type, Distribution Channel and Customer status

First we considered transforming the Profit Percentage. However after observing the square-root and the log of the variable Profit Percentage were further skewing the variable, so decided to leave Profit Percentage the same. We then tried transforming the Square Feet variable. The log (Square Feet) gave us a close to normally distributed graph, so we decided to go with log (Square Feet) in our model. We see that the new adjusted R-square value is 41.41% with an increase of 1.79% from the previous model.

Our final regression model was dependent variable Profit Percentage combined with the independent variables, log(Square Feet), Item Type “Roll”, Distribution Channel “Direct Sales”, Customer Status “Returning Customer” with an adjusted R-squared value as 41.41% after removing the outliers.

Conclusion

The most substantial take away we received from this project was that we learned how it is to thoroughly understand your data and how carefully you must choose your dependent variable in order to accurately analyse the data. This became abundantly clear when we had already done a significant amount of analysis using sales and realized that our results were not substantive. After we worked to attain more data to further expand the boundaries of the project and change the dependent variable, we began to see how much of a difference having a larger amount of data can have. It allows you to have more insight into how the variables interact with each other when you have more of them in your dataset. Working on this project overall also helped us develop hands on experience. This experience gave us the tools needed to complete this project with much more significant results than we would have anticipated possible even a few weeks earlier to completing the project.

The conclusions that we reached in terms of our data were that most of our hypotheses that we formed around the effects the data would have on Profit Percentage ended up being true. The only variable that did not have the relationship we expected was Average Temperature. We expected there to be a higher Profit Percentage in states that had a lower Average Temperature; our thought process being that the colder states would be the ones who would buy more heating products. However, there ended up being no relationship between Average Temperature per state and the Profit Percentage. This could likely be due to the way the company performs sales. Sales are likely not built directly around needs of individuals needs based on any particular climate, rather they likely make most of their sales from building relationships with certain clients and then those clients consistently buy product. Then those sales would have a direct effect on Profit Percentage.