Project: Electronic car charger sale analysis

Hyunju Shim, 3035345693

10/21/2018

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggthemes)  
library(reshape2)  
library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

setwd("/Users/Hyunjulie")  
data<- read.csv('modified.csv')  
data <- subset(data, select = -c(X...Site))

# 1. Introduction

This project will aim to carry out an analysis of Electronic Car Charger Sales. A brief introduction of the customer is as follows:

??? Customer: Electronic Car Charger company in Guangzhou ??? Data: Daily Electricity consumption at each station (26 columns & 61 rows) ??? \* Question \* : Which Covariates (features, factors) contribute to the increase/decrease of the electricity sale?

??? \* Our Objective of Analysis\* : Guide to pick up locations for new chargers

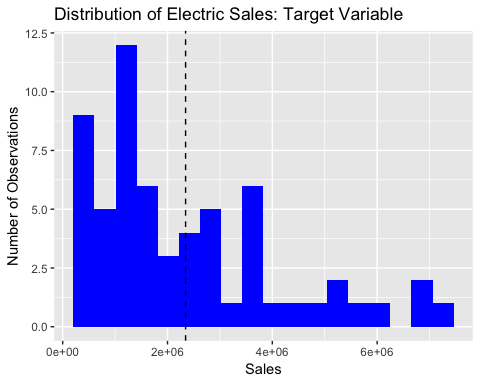
The question and objective of the project is vague. It is important to objectively idenfity and illustrate the question again in statistical language. When stating it as a statistical question, the question becomes: - How does each covariate affect the outcome? - Are they significantly different from 0 in 0.05 significance level? - Is it necessary to include certain covariates to improve the model fitting?

Throughout the analysis, we will be using ANOVA (analysis of variance). ANOVA is a statistical method in which the variation in a set of observations is divided into distinct components. It enables us to idenfity whether a feature significantly affects the target.

## 1.1 Analysis of Target Variable

Before starting our the analysis of features and variables, it is important to know how our target variable (Sale of Electricity) is distributed. General overview of ‘Sale Distribution’ will help us to investigate deeper.

p1 <- data %>%  
 ggplot(aes(sale)) +  
 geom\_histogram(bins = 18, fill = "blue") +  
 labs(x = "Sales", y ="Number of Observations") +  
 ggtitle("Distribution of Electric Sales: Target Variable") +   
 geom\_vline(xintercept = mean(data$sale), lty=2)  
p1

 \* Note that the dotted line is the mean of Sales.

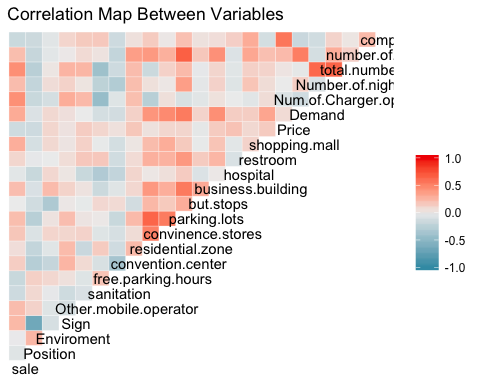
From the Distribution we find: A. The target distribution is right skewed. B. There are notable high-count peaks at several specific price values.

-> There are some sites that excel prominently relative to other places. We may need to look for reasons on those outliers.

1. Correlation Map

data[seq(0,23)] %>% ggcorr(method = c("pairwise","spearman"), label = FALSE, angle = -0, hjust = 0.2) + coord\_flip() + ggtitle("Correlation Map Between Variables")

## Coordinate system already present. Adding new coordinate system, which will replace the existing one.



We now explore correlation with each variable

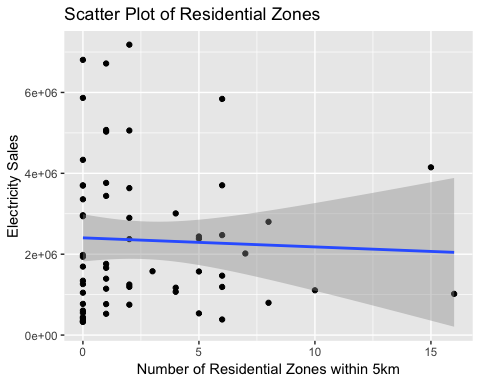
* Constructing F-Test for each probable variable Null Hypothesis: The feature does not affect the sale of electricity Alternative Hypothesis: Feature affects the sale of electricity
* Reject at 0.05 level: if p < 0.05 we should reject the null hypothesis

For each of 11 variables, we will be 1) Plotting a simple scatterplot with its corresponding regression line 2) Constructing ANOVA table and its summary 3) State whether the feature is significant on 0.05 level

Variables: Residential zone, convenience stores, parking lots, business buildings, hospital, restroom, shopping mall, Demand, number of competitors’ charger, competitor’s price, price

## 1. Residential Zone

ggplot(data, aes(x=data$residential.zone, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Residential Zones") + ylab("Electricity Sales") + xlab("Number of Residential Zones within 5km")



results = lm(sale ~ residential.zone, data=data)  
summary(results)

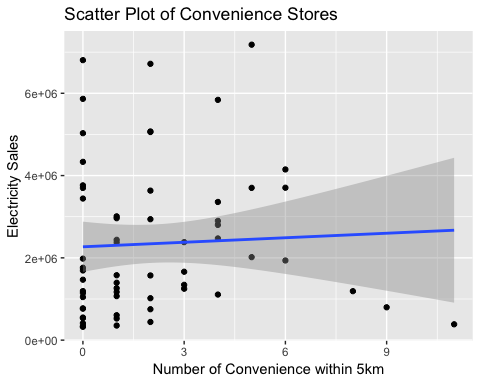
##   
## Call:  
## lm(formula = sale ~ residential.zone, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2078609 -1244580 -711227 1059288 4821634   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2404520 292410 8.223 2.3e-11 \*\*\*  
## residential.zone -22402 66654 -0.336 0.738   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1818000 on 59 degrees of freedom  
## Multiple R-squared: 0.001911, Adjusted R-squared: -0.01501   
## F-statistic: 0.113 on 1 and 59 DF, p-value: 0.738

#anova(results)

F-statistic for residential zone is 0.113 on 1 and 59 degrees of freedom. P-Value is 0.738, and thus we do not reject the null hypothesis. The feature Residential zone statistically does not affect sales.

## 2. Number of Convenience Stores

ggplot(data, aes(x=data$convinence.stores, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Convenience Stores") + ylab("Electricity Sales") + xlab("Number of Convenience within 5km")



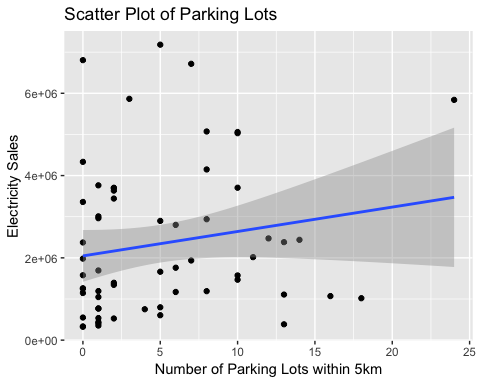
results = lm(sale ~ convinence.stores, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## convinence.stores 1 4.8856e+11 4.8856e+11 0.1479 0.7019  
## Residuals 59 1.9484e+14 3.3024e+12

P-Value is 0.7019, and thus we do not reject the null hypothesis. The feature Convenience Store statistically does not affect sales.

## 3. Number of Parking Lots

ggplot(data, aes(x=data$parking.lots, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Parking Lots") + ylab("Electricity Sales") + xlab("Number of Parking Lots within 5km")



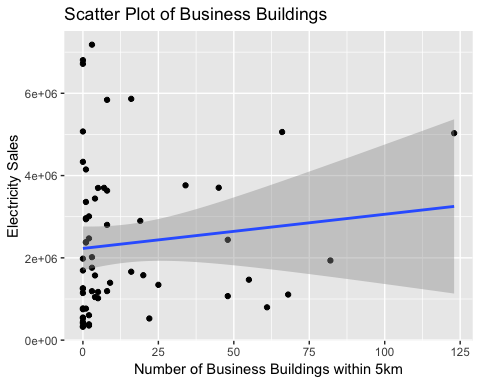
results = lm(sale ~ parking.lots, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## parking.lots 1 6.1279e+12 6.1279e+12 1.9109 0.1721  
## Residuals 59 1.8920e+14 3.2068e+12

P-Value is 0.1721, and thus we do not reject the null hypothesis. The feature Parking Lots statistically does not affect sales.

## 4. Number of Business Buildings

ggplot(data, aes(x=data$business.building, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Business Buildings") + ylab("Electricity Sales") + xlab("Number of Business Buildings within 5km")



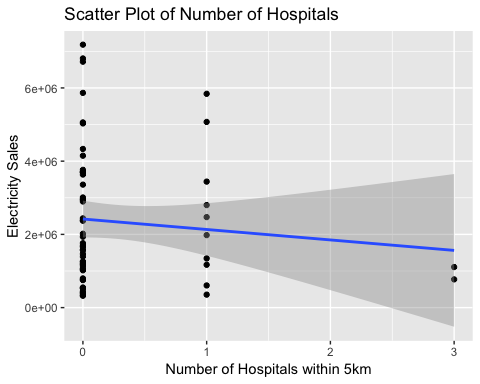
results = lm(sale ~ business.building, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## business.building 1 2.5108e+12 2.5108e+12 0.7683 0.3843  
## Residuals 59 1.9282e+14 3.2681e+12

P-Value is 0.3843, and thus we do not reject the null hypothesis. The feature Business Building statistically does not affect sales.

## 5. Hospital

ggplot(data, aes(x=data$hospital, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Number of Hospitals") + ylab("Electricity Sales")+ xlab("Number of Hospitals within 5km")



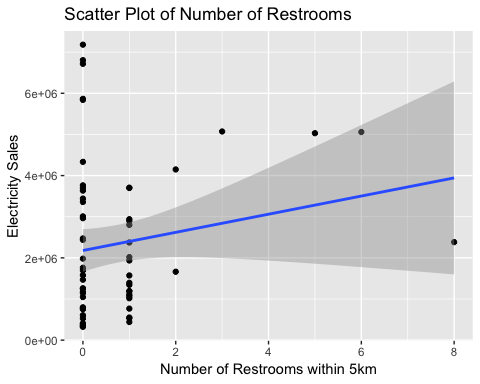
results = lm(sale ~ hospital, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## hospital 1 1.9426e+12 1.9426e+12 0.5927 0.4445  
## Residuals 59 1.9339e+14 3.2777e+12

P-Value is 0.4445, and thus we do not reject the null hypothesis. The feature Hospital statistically does not affect sales.

## 6. Restroom

ggplot(data, aes(x=data$restroom, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Number of Restrooms") + ylab("Electricity Sales")+ xlab("Number of Restrooms within 5km")



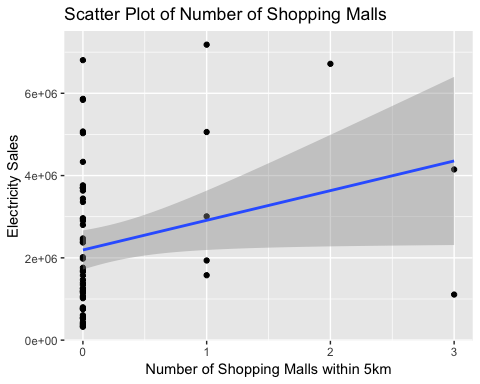
results = lm(sale ~ restroom, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## restroom 1 6.1942e+12 6.1942e+12 1.9323 0.1697  
## Residuals 59 1.8914e+14 3.2057e+12

P-Value is 0.1697, and thus we do not reject the null hypothesis. The feature Restroom statistically does not affect sales.

## 7. Shopping Mall

ggplot(data, aes(x=data$shopping.mall, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Number of Shopping Malls") + ylab("Electricity Sales") + xlab("Number of Shopping Malls within 5km")



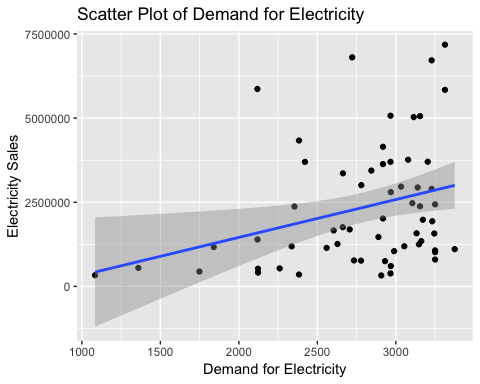
results = lm(sale ~ shopping.mall, data=data)  
summary(results)

##   
## Call:  
## lm(formula = sale ~ shopping.mall, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3247502 -1334656 -528780 1167305 4614509   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2191284 237855 9.213 5.06e-13 \*\*\*  
## shopping.mall 721411 357515 2.018 0.0482 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1760000 on 59 degrees of freedom  
## Multiple R-squared: 0.06456, Adjusted R-squared: 0.0487   
## F-statistic: 4.072 on 1 and 59 DF, p-value: 0.04816

Because the p-value is 0.04816, we reject our null hypothesis. The feature shopping mall has statistically significant correlation with electric sales. It has positive correlation, meaning that sales will increase if number of shopping malls within 5km increases.

## 8. Demand

ggplot(data, aes(x=data$Demand, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Demand for Electricity") + ylab("Electricity Sales") + xlab("Demand for Electricity")



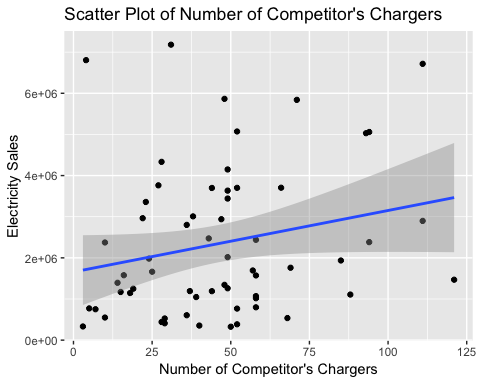
results = lm(sale ~ Demand, data=data)  
summary(results)

##   
## Call:  
## lm(formula = sale ~ Demand, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2157062 -1283834 -440801 1036117 4537692   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -792777.1 1295309.0 -0.612 0.5429   
## Demand 1124.5 457.3 2.459 0.0169 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1733000 on 59 degrees of freedom  
## Multiple R-squared: 0.09294, Adjusted R-squared: 0.07757   
## F-statistic: 6.046 on 1 and 59 DF, p-value: 0.01689

Because the p-value is 0.01689, we reject our null hypothesis. The feature Demand has significant correlation with electric sales. It has positive correlation, meaning that sales will increase for increased demand.

## 9. Number of Competitors’ Charger

ggplot(data, aes(x=data$number.of.competitors..charger, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Number of Competitor's Chargers") + ylab("Electricity Sales")+ xlab("Number of Competitor's Chargers")



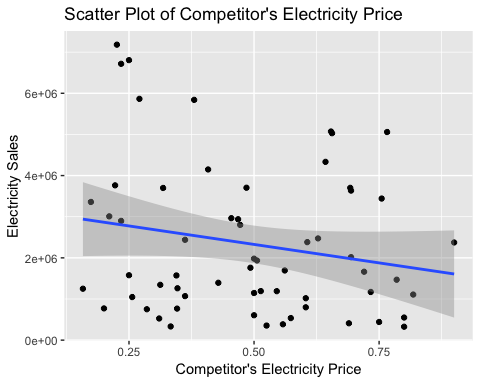
results = lm(sale ~ number.of.competitors..charger, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)   
## number.of.competitors..charger 1 1.0122e+13 1.0122e+13 3.2246 0.07766 .  
## Residuals 59 1.8521e+14 3.1391e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-Value is 0.07766, and thus we do not reject the null hypothesis. The feature Number of Competitors’ Charger does not affect sales.

## 10. Competitor’s Price

ggplot(data, aes(x=data$competitors..price, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Competitor's Electricity Price") + ylab("Electricity Sales")+ xlab("Competitor's Electricity Price")



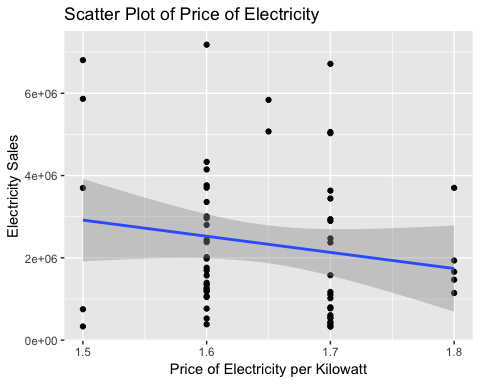
results = lm(sale ~ competitors..price, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## competitors..price 1 7.5631e+12 7.5631e+12 2.3765 0.1285  
## Residuals 59 1.8777e+14 3.1825e+12

P-Value is 0.1285, and thus we do not reject the null hypothesis. The feature Competitor’s Price statistically does not affect sales.

## 11. Price

ggplot(data, aes(x=data$Price, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Price of Electricity") + ylab("Electricity Sales")+ xlab("Price of Electricity per Kilowatt")



results = lm(sale ~ Price, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Price 1 5.3358e+12 5.3358e+12 1.6569 0.203  
## Residuals 59 1.8999e+14 3.2202e+12

P-Value is 0.203, and thus we do not reject the null hypothesis. The feature Price statistically does not affect sales.

### Extra analysis

Now, we carry out extra analysis on some of the features we left out during our main analysis. Those features include: - Position, Environment, Sign, Sanitation, Other mobile Carries, Convention center, Number of Chargers open in day time, Number of night-only chargers, and Total number of Chargers.

Features that are mostly single values (all zeros/ones, little observations) are left out in the analysis. Even if these features are statistically significant in its effect in electricity sales, it is not useful when choosing the location of the charging station. As stated above, this is because these factors can be ‘modified’ even after choosing the location. It will offer the customer some extra information about what kind of things to do/consider after choosing the location (e.g. putting up signs)

Similarly, we will construct F-Test for each variable - Null Hypothesis: The feature does not affect the sale of electricity - Alternative Hypothesis: Feature affects the sale of electricity - Reject at 0.05 level: if p < 0.05 we should reject the null hypothesis

1. Position of the station

results = lm(sale ~ Position, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Position 1 4.1779e+11 4.1779e+11 0.1265 0.7234  
## Residuals 59 1.9491e+14 3.3036e+12

P-Value is 0.7234, and thus we do not reject the null hypothesis. Position does not affect sales.

1. Environment

results = lm(sale ~ Enviroment, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Enviroment 1 1.1887e+12 1.1887e+12 0.3613 0.5501  
## Residuals 59 1.9414e+14 3.2905e+12

P-Value is 0.5501, and thus we do not reject the null hypothesis. Environment does not affect sales.

1. Sign

results = lm(sale ~ Sign, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Sign 1 8.4301e+12 8.4301e+12 2.6612 0.1082  
## Residuals 59 1.8690e+14 3.1678e+12

P-Value is 0.1082, and thus we do not reject the null hypothesis. Sign does not affect sales.

1. Sanitation

results = lm(sale ~ sanitation, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## sanitation 1 2.0721e+12 2.0721e+12 0.6326 0.4296  
## Residuals 59 1.9326e+14 3.2756e+12

P-Value is 0.4296, and thus we do not reject the null hypothesis. Sanitation does not affect sales.

1. Other mobile Operator

results = lm(sale ~ Other.mobile.operator, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Other.mobile.operator 1 5.1428e+12 5.1428e+12 1.5954 0.2115  
## Residuals 59 1.9019e+14 3.2235e+12

P-Value is 0.2115, and thus we do not reject the null hypothesis. Other mobile Operator does not affect sales.

1. Convention center

results = lm(sale ~ convention.center, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## convention.center 1 1.2019e+12 1.2019e+12 0.3653 0.5479  
## Residuals 59 1.9413e+14 3.2903e+12

P-Value is 0.7234, and thus we do not reject the null hypothesis. Position does not affect sales.

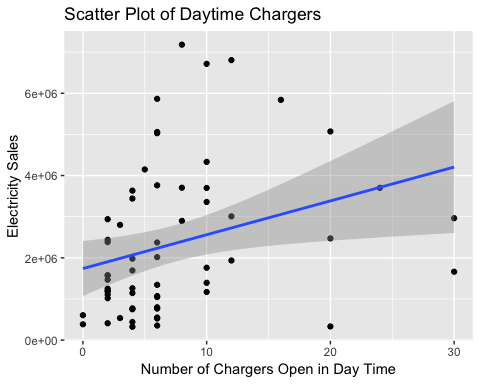
1. Number of Charger open in day time

results = lm(sale ~ Num.of.Charger.open.in.day.time, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Num.of.Charger.open.in.day.time 1 1.7622e+13 1.7622e+13 5.8506 0.01868 \*  
## Residuals 59 1.7771e+14 3.0120e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-Value is 0.01868. We reject our null hypothesis that the feature does not affect sales. Number of Charger open in day time has significant positive effect in electricity sales.

ggplot(data, aes(x=data$Num.of.Charger.open.in.day.time, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Daytime Chargers") + ylab("Electricity Sales") + xlab("Number of Chargers Open in Day Time")



1. Number of night only charger

results = lm(sale ~ Number.of.night.only.charger, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Number.of.night.only.charger 1 4.8004e+12 4.8004e+12 1.4865 0.2276  
## Residuals 59 1.9053e+14 3.2293e+12

P-Value is 0.2276, and thus we do not reject the null hypothesis. Number of night only charger does not affect sales.

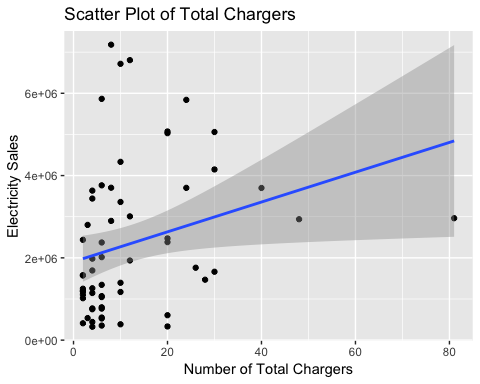
1. Total number of chargers

results = lm(sale ~ total.number.of.charger, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)   
## total.number.of.charger 1 1.4627e+13 1.4627e+13 4.7759 0.03284 \*  
## Residuals 59 1.8070e+14 3.0628e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

P-Value is 0.03284. We reject our null hypothesis that the feature does not affect sales. Total number of chargers has significant positive effect in electricity sales.

ggplot(data, aes(x=data$total.number.of.charger, y=data$sale)) + geom\_point() + ggtitle("Add geom\_point with coloring") + geom\_smooth(method='lm') + ggtitle("Scatter Plot of Total Chargers") + ylab("Electricity Sales") + xlab("Number of Total Chargers")



results = lm(sale ~ shopping.mall + Demand, data=data)  
anova(results)

## Analysis of Variance Table  
##   
## Response: sale  
## Df Sum Sq Mean Sq F value Pr(>F)   
## shopping.mall 1 1.2610e+13 1.2610e+13 4.2873 0.04286 \*  
## Demand 1 1.2128e+13 1.2128e+13 4.1235 0.04689 \*  
## Residuals 58 1.7059e+14 2.9412e+12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1