Hyunju Shim, 3035345693

ELECTRONIC CAR CHARGER SALE ANALYSIS

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 contribute to the increase/decrease of the electricity sale?
- **Objective of Analysis**: Guide the customer to pick up locations for new chargers

The question and objective of the project is vague. It is important to firstly identify and illustrate the question again in statistical language. Therefore, the question becomes:

- How does each covariate affect the outcome? And is the correlation significantly different from 0 in 0.05 significance level?
- Is it necessary to include certain covariates to improve the model fitting?

Overview of Methodology:

1. We will start with exploring the dataset such as visualizing the target variable and using heatmap to get a general overview of correlation between variables.
2. We will then examine each variable in detail to logically question whether the feature is useful in answering our question.
3. After finding potentially valuable features, 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.
4. Extra analysis of other variables will also be carried out. They may help the customer have insightful data on *how* to station the chargers.
5. Interpretation of our results will be carried out.

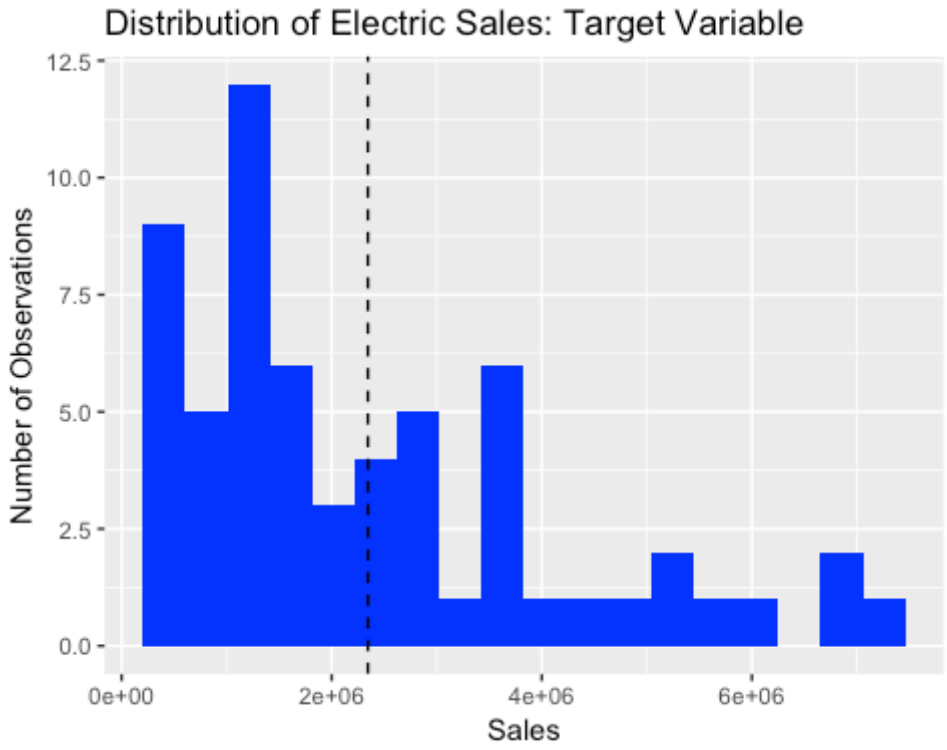
Statistical language, R, will be used to carry out the analysis and to produce meaningful graphics.

2. EXPLORING THE DATASET

2.1 ANALYSIS OF TARGET VARIABLES

Before starting 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.

```
library(ggplot2)
library(dplyr)
library(ggthemes)
library(reshape2)
library(GGally)
data<- read.csv('data.csv')
p1 <- data %>% ggplot(aes(sale)) + geom_histogram(bins = 18, fill = "blue") + labs(x = "Sales", y ="Number of Obs
ervations") + ggtitle("Distribution of Electric Sales: Target Variable") + geom_vline(xintercept = mean(data$sale),
lty=2)
```



* Note that the dotted line represents the mean of Electricity Sales.

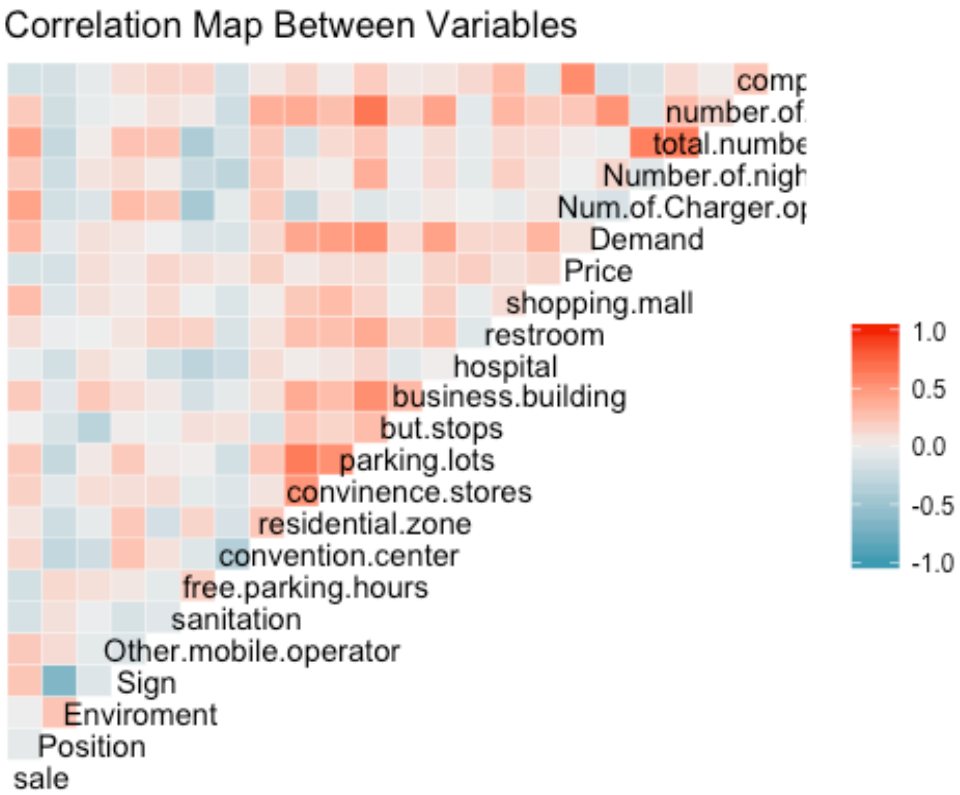
- From the distribution we find that:

A. The target distribution is right skewed.

B. There are notable high-count peaks at several specific price values.

2.2 CORRELATION MAP OF VARIABLES

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



Insights:

- We are interested in the left-most column of the Heatmap which shows an overall correlation between the variables. Colors that are close to white indicate close-to-zero correlation, whereas red and blue colors indicate positive and negative correlation respectively.
- We are able to see that many features are correlated to each other. For example, Residential zone and parking lot features are close to bright red. It logically makes sense because places residential towns will need more parking lots.
- By looking at the 'Sale' column (left-most column), we get a general overview of possible features we may explore in our analysis.

3. ANALYSIS OF EACH VARIABLE

3.1 REMOVING UNRELATED FEATURES

a. Features that have same value for all observations

- There are 5 features that have same values for all observations. They are: “Convenience”, “Nighttime”, “China Mobile”, “China United Network”, “China Telecommunications Corporation”.

- We can eliminate these variables during our analysis because they cannot give us meaningful information. All sites have the same feature (e.g. all sites have Chine Mobile signal)

b. Features that have too few observations

- Feature “Daytime” has only one observation that is different (i.e. only one observation is 0; all others are 1). This means that all other chargers open day time, and accordingly have higher electricity sale. Logically, opening longer time (whether it be day or night) will lead to higher sale.

- Moreover, since we only have one observation that is different, including this variable may lead to overfitting to this dataset. It would be better to drop this feature.

- Also, it does not help in solving the question, ‘choosing locations for new chargers’ because sale will increase if it opens in daytime even if it is the same location.

- This case is similar in “Theater” feature. It also has only one observation which is greater than 0. Thus, we eliminate features “daytime” and “theater”.

c. Features that are not related to the “Location” of the charger

- We always need to consider the objective of our analysis. The objective is to find meaningful features when deciding on the ‘location’ of the next charger for our customer. However, there are features that are not related to the location of the charger. These features are those that can be altered even after the location is changed, and they are just the ‘environment’ which the charger is in.

- Position, Environment, Sign, Sanitation, Number of Charger open in day time, Number of night-only charger, Total number of chargers: These features can be changed even if the location is chosen, and thus should not affect our analysis of ‘charger location’. For example, if putting up a sign for the charger turns out to significantly affect the sale, the customer can put up a sign after choosing the location. And due to this fact, we will add an additional analysis of these variables in the back.

- We will focus on the features that **cannot be altered** and is dependent on the location.

3.2 ANALYSIS OF EACH FEATURE

We now have 11 Features that may potentially be useful. Steps for statistical analysis is as follows:

- 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 reject the null hypothesis

For each variable, we will be:

1) Plotting a simple scatterplot with its corresponding regression line

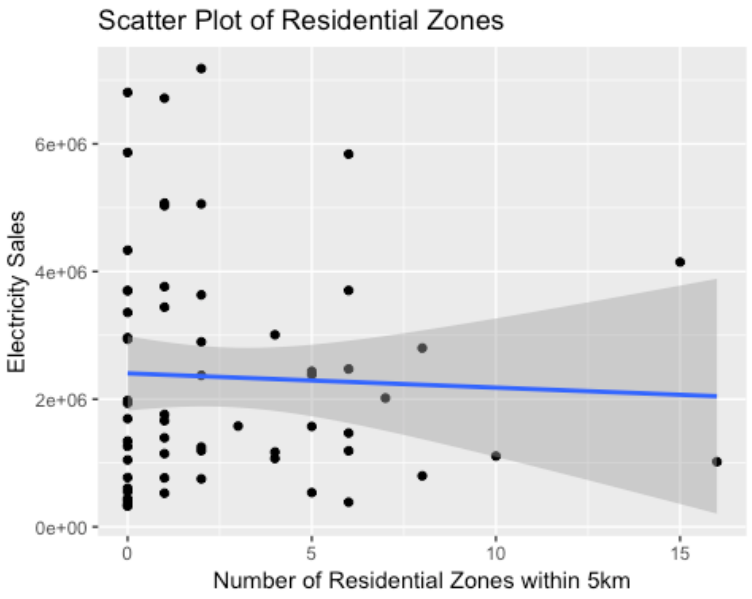
2) Constructing ANOVA table and its summary

3) State whether the feature is significant on 0.05 level

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

3.2.1. 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")
```



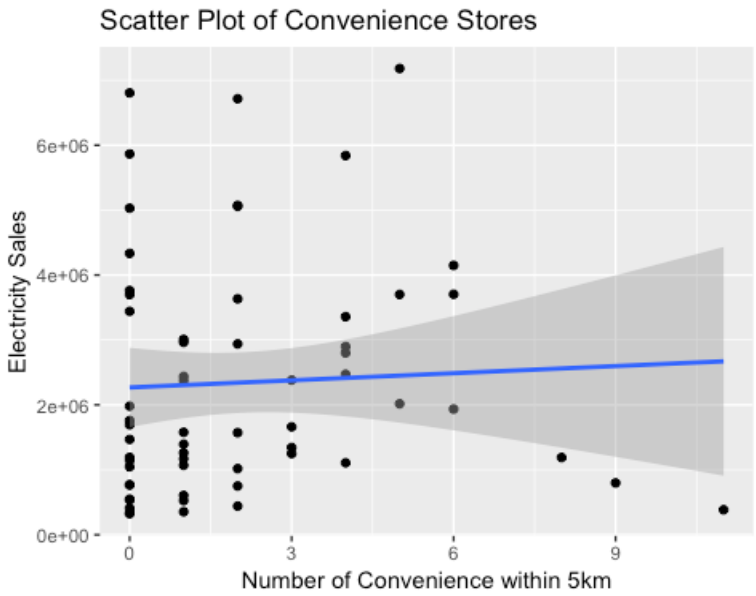

```
results = lm(sale ~ residential.zone, data=data)
summary(results)
## Call:
## lm(formula = sale ~ residential.zone, data = data)
##
## Residuals:
```

```
##      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
```

- 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 does not have correlations with sale.

3.2.2. **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")
```

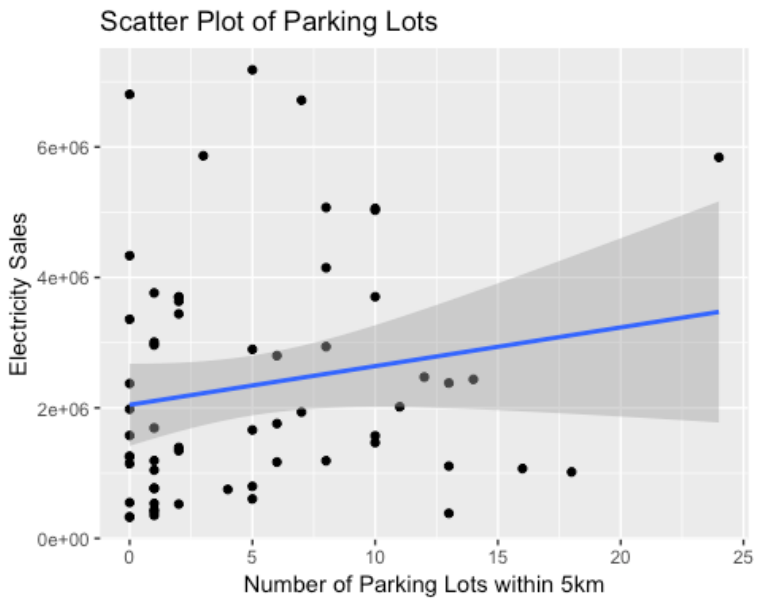



```
results = lm(sale ~ convinence.stores, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##                      Df    Sum Sq       Mean Sq      F value    Pr(>F)
## convinence.stores  1    4.8856e+11  4.8856e+11  0.1479    0.7019
## Residuals            59   1.9484e+14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

- P-Value is 0.7019, and thus we do not reject the null hypothesis. The feature Convenience Store does not have correlations with sale.

3.2.3. 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")
```

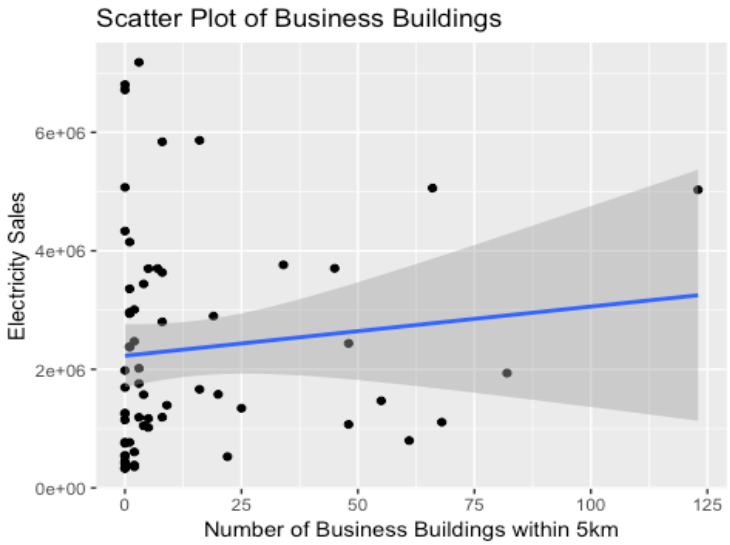



```
results = lm(sale ~ parking.lots, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##                Df    Sum Sq         Mean Sq       F value    Pr(>F)
## parking.lots  1    6.1279e+12     6.1279e+12    1.9109    0.1721
## Residuals    59   1.8920e+14      3.2068e+12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

- P-Value is 0.1721, and thus we do not reject the null hypothesis. The feature Parking Lots does not have correlations with sale.

3.2.4. 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")
```

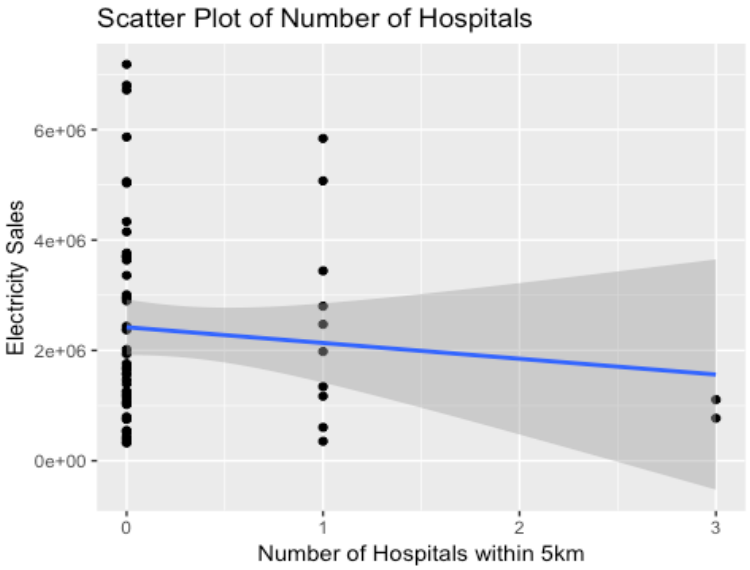



```
results = lm(sale ~ business.building, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##                   Df    Sum Sq       Mean Sq      F value   Pr(>F)
## business.building  1   2.5108e+12  2.5108e+12    0.7683    0.3843
## Residuals         59   1.9282e+14  3.2681e+12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

- P-Value is 0.3843, and thus we do not reject the null hypothesis. The feature Business Building does not have correlations with sale.

3.2.5. **Number of Hospitals**

```
ggplot(data, aes(x=data$hospital, y=data$sale)) + geom_point() + ggtitle("Add geom_point with coloring") + geo
m_smooth(method='lm') + ggtitle("Scatter Plot of Number of Hospitals") + ylab("Electricity Sales")+ xlab("Numbe
r of Hospitals within 5km")
```

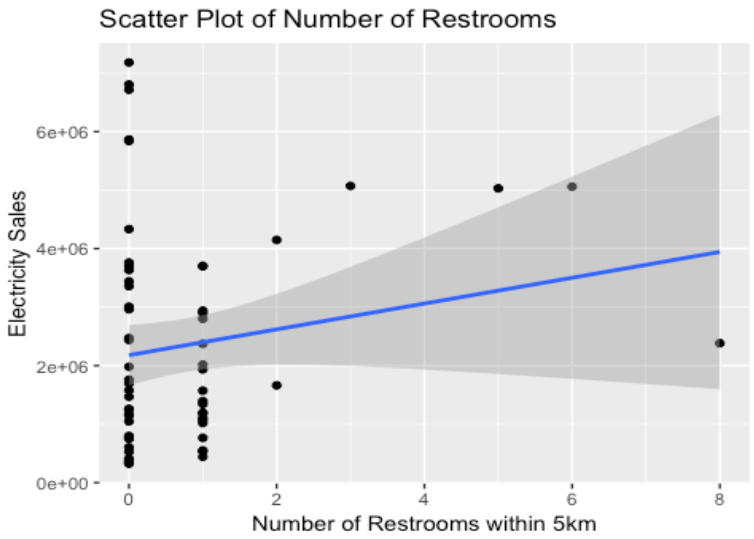


```
results = lm(sale ~ hospital, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##              Df   Sum Sq         Mean Sq       F value   Pr(>F)
## hospital     1    1.9426e+12    1.9426e+12    0.5927    0.4445
## Residuals    59   1.9339e+14    3.2777e+12
```

- P-Value is 0.4445, and thus we do not reject the null hypothesis. The feature Hospital does not have correlations with sale.

3.2.6. **Number of Restrooms**

```
ggplot(data, aes(x=data$restroom, y=data$sale)) + geom_point() + ggtitle("Add geom_point with coloring") + geo
m_smooth(method='lm') + ggtitle("Scatter Plot of Number of Restrooms") + ylab("Electricity Sales")+ xlab("Num
ber of Restrooms within 5km")
```

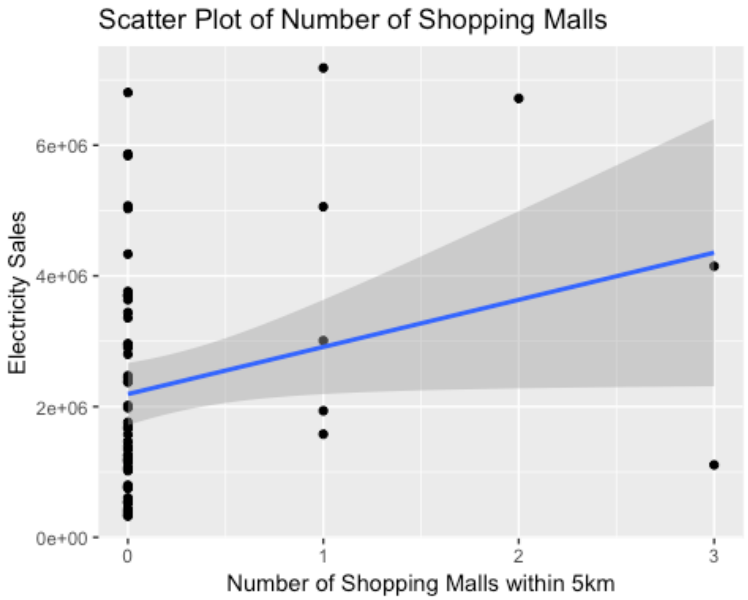



```
results = lm(sale ~ restroom, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##              Df   Sum Sq         Mean Sq       F value   Pr(>F)
## restroom     1    6.1942e+12    6.1942e+12    1.9323    0.1697
## Residuals    59   1.8914e+14    3.2057e+12
```

- P-Value is 0.1697, and thus we do not reject the null hypothesis. The feature Restroom does not have correlations with sale.

3.2.7. Number of Shopping Malls

```
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") + x
lab("Number of Shopping Malls within 5km")
```

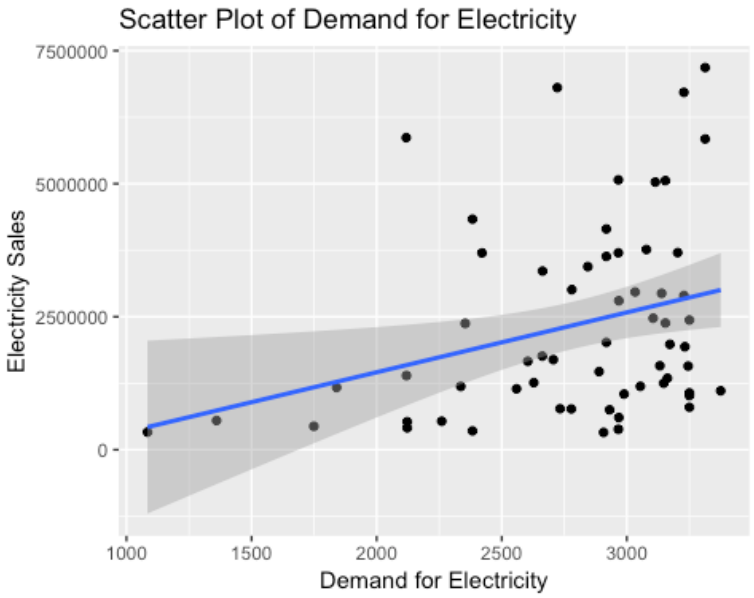



```
results = lm(sale ~ shopping.mall, data=data)
summary(results)
##
## Call:
## lm(formula = sale ~ shopping.mall, data = data)
##
## Residuals:
##     Min      1Q          Median      3Q     Max
## -3247502 -1334656  -528780  1167305  4614509
##
## Coefficients:
##                  Estimate  Std. 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.

3.2.8. **Demand for Electricity**

```
ggplot(data, aes(x=data$Demand, y=data$sale)) + geom_point() + ggtitle("Add geom_point with coloring") + geo
m_smooth(method='lm') + ggtitle("Scatter Plot of Demand for Electricity") + ylab("Electricity Sales") + xlab("Dem
and for Electricity")
```

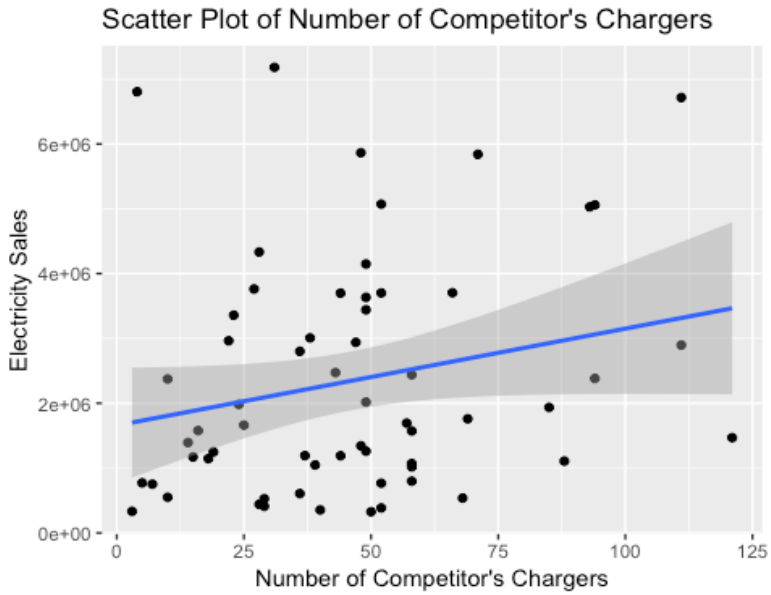



```
results = lm(sale ~ Demand, data=data)
summary(results)
##
## Call:
## lm(formula = sale ~ Demand, data = data)
##
## Residuals:
##     Min      1Q      Median      3Q     Max
## -2157062 -1283834  -440801  1036117  4537692
##
## Coefficients:
##                 Estimate   Std. Error      t value   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.

3.2.9. Number of Competitor's 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")
```

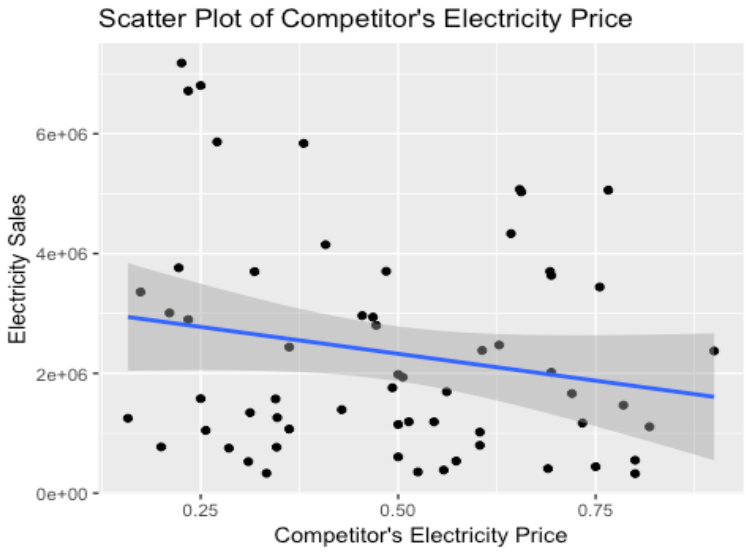



```
results = lm(sale ~ number.of.competitors..charger, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##                                  Df   Sum Sq        Mean Sq      F value   Pr(>F)
## number.of.competitors..charger  1    1.0122e+13  1.0122e+13   3.2246    0.07766 .
## 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 have correlations with sale.

3.2.10. Competitor's Electricity 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")
```

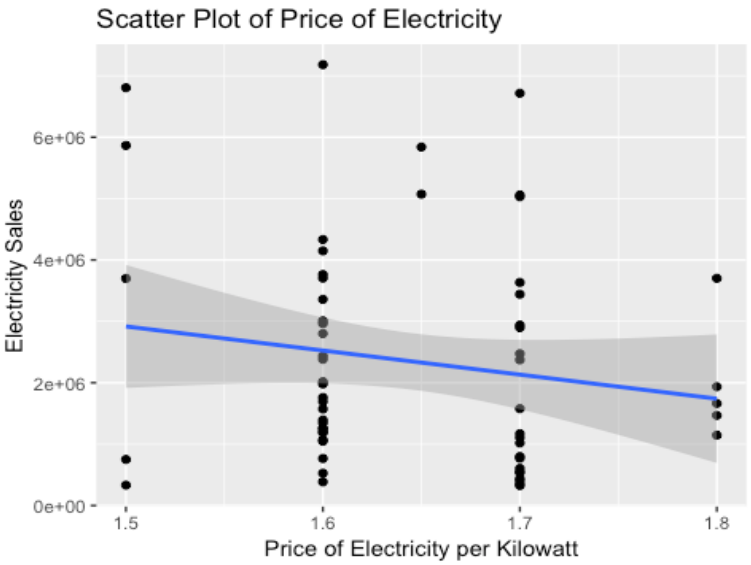



```
results = lm(sale ~ competitors..price, data=data)
anova(results)
## Analysis of Variance Table
##
## Response: sale
##                      Df  Sum Sq         Mean Sq      F value    Pr(>F)
## competitors..price   1   7.5631e+12     7.5631e+12   2.3765     0.1285
## Residuals            59  1.8777e+14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

- P-Value is 0.1285, and thus we do not reject the null hypothesis. The feature Competitor's Price does not have correlations with sale.

3.2.11. **Price of Electricity**

```
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 does not have correlations with sale.

4. EXTRA ANALYSIS OF OTHER FEATURES

- Now, we carry out extra analysis on some of the features we left out as mentioned above. 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.
- 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 and consider after choosing the location
- 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

a. 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 have correlations with sale.

b. 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 have correlations with sale.

c. 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 have correlations with sale.

d. 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 have correlations with sale.

e. 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 have correlations with sale.

f. 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
```

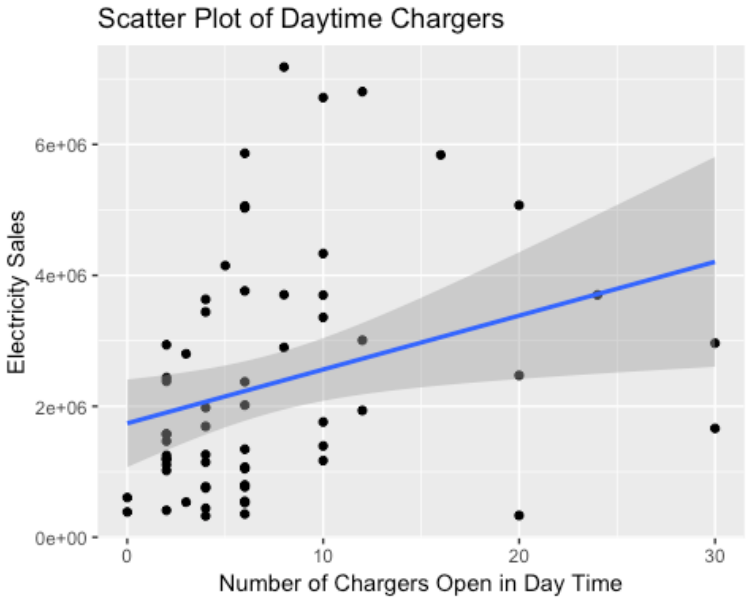
- P-Value is 0.7234, and thus we do not reject the null hypothesis. Position does not have correlations with sale.

g. 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")
```



h. 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
```

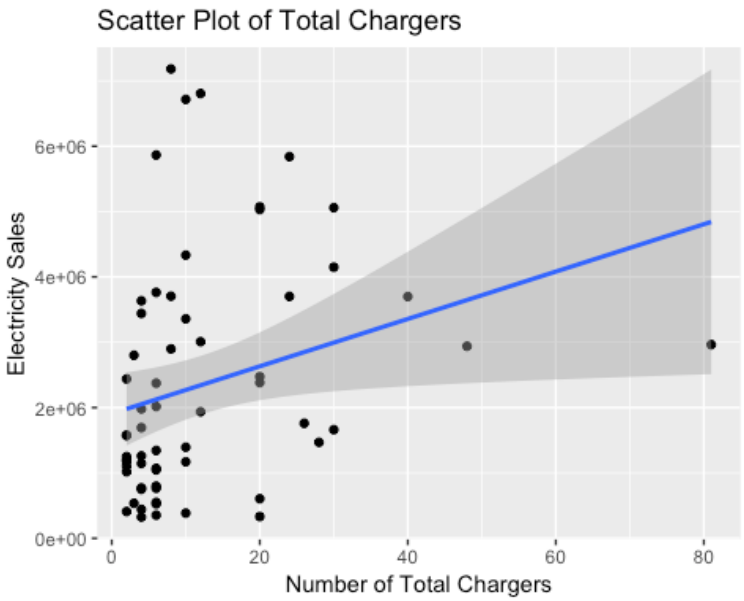
- P-Value is 0.2276, and thus we do not reject the null hypothesis. Number of night only charger does not have correlations with sale.

i. 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")
```



5. INTERPRETATION OF RESULTS

- Through our statistical analysis, we conclude that only the features “Number of Shopping Malls within 5 km” and “Demand of Electricity” have positive correlations with electricity sales.

- The result is interesting in some ways. It is logically expected that if there is more demand of electricity in that region, people will buy more. However, it is interesting that number of shopping malls in the region affect the sale. It may be due to the fact that people usually take their cars when they are shopping, and they take that chance to charge their cars.

- It was logically expected that competitor’s price may have some significant correlation, but the result was not significant. Number of competitor’s chargers (p-value of 0.07766) could have concluded to be significant if our level of significance was 0.1.

- Moreover, we also found that “Number of total chargers” and “Number of Chargers open in Day Time” has positive correlations with sale. This will give extra insight to our customers. If possible, it is better to increase the number of total chargers and number of chargers open in day time in their station.