Final Project Phase 1 : Project Strategy Planning

Ridha Paramesh

Objective:

The project aims to conduct a thorough analysis of household-level responses to the American Community Survey for households in Oregon, using a subset of variables from the 2015 1-year survey. The objective is to gain insights into demographic patterns, socio-economic factors, and electricity payment behaviors within the specified criteria, ensuring clarity in the analysis process and justifying any modifications made.

We are then asked to create a model to predict electricity costs for a typical Oregon household. Lets break it into smaller, more manageable questions to aid in our analysis as follows:

- ☐ Is there a need to clean the data?
- ☐ What is the average cost of electricity for apartments?
- ☐ What is the average cost of electricity for houses?
- ☐ Is there a statistically significant difference between the electricity costs in apartments and houses?
- ☐ Which predictive model is most effective in estimating electricity costs?
- ☐ What is the distribution of monthly electricity costs (ELEP) across different types of dwellings in Oregon?
- ☐ Does the number of bedrooms (BDSP) correlate with electricity costs (ELEP) across dwellings?
- ☐ How do electricity costs (ELEP) vary with the tenure of the household (TEN)?
- ☐ Is there a significant difference in electricity costs (ELEP) between households with and without persons under 18 (R18)?
- ☐ Can we predict electricity costs (ELEP) using dwelling characteristics (BLD, BDSP, NP) and household demographics (R18, R60)?

```
# Load the dataset
data <- read.csv("OR_acs_house_occ.csv")
# Explore the dataset
summary(data)
```

```
##     SERIALNO             NP              TYPE       ACR
##  Min.   :     70   Min.   : 1.000   Min.   :1   Length:15166
##  1st Qu.: 368628   1st Qu.: 1.000   1st Qu.:1   Class :character
##  Median : 748326   Median : 2.000   Median :1   Mode  :character
##  Mean   : 749620   Mean   : 2.403   Mean   :1
##  3rd Qu.:1126788   3rd Qu.: 3.000   3rd Qu.:1
##  Max.   :1513284   Max.   :13.000   Max.   :1
##
##       BDSP           BLD                ELEP            FULP
##  Min.   :0.000   Length:15166       Min.   :  4.0   Min.   :   1.00
##  1st Qu.:2.000   Class :character   1st Qu.: 70.0   1st Qu.:   2.00
##  Median :3.000   Mode  :character   Median :100.0   Median :   2.00
##  Mean   :2.789                      Mean   :116.2   Mean   :  85.25
##  3rd Qu.:3.000                      3rd Qu.:150.0   3rd Qu.:   2.00
```

```
##  Max.   :7.000                      Max.   :540.0   Max.   :2500.00
##
##       GASP            HFL                 RMSP           TEN
##  Min.   :  3.00   Length:15166       Min.   : 1.00   Length:15166
##  1st Qu.:  3.00   Class :character   1st Qu.: 4.00   Class :character
##  Median :  3.00   Mode  :character   Median : 6.00   Mode  :character
##  Mean   : 35.68                      Mean   : 6.01
##  3rd Qu.: 50.00                      3rd Qu.: 7.00
##  Max.   :350.00                      Max.   :16.00
##
##       VALP             YBL                R18                R60
##  Min.   :   1000   Length:15166       Length:15166       Length:15166
##  1st Qu.: 160000   Class :character   Class :character   Class :character
##  Median : 250000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 301966
##  3rd Qu.: 360000
##  Max.   :2476000
##  NA's   :4632
```

```
str(data)
```

```
## 'data.frame':    15166 obs. of  16 variables:
##  $ SERIALNO: int  70 163 178 243 300 603 743 803 847 882 ...
##  $ NP      : int  4 2 1 2 1 3 2 2 2 2 ...
##  $ TYPE    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ACR     : chr  "House on less than one acre" "House on less than one acre" "House on less than one
##  $ BDSP    : int  2 2 3 4 2 3 3 2 2 2 ...
##  $ BLD     : chr  "One-family house detached" "One-family house detached" "One-family house detached"
##  $ ELEP    : int  70 100 60 80 150 200 120 60 70 70 ...
##  $ FULP    : int  2 600 2 2 2 2 2 2 2 2 ...
##  $ GASP    : int  3 3 110 20 3 20 3 3 3 50 ...
##  $ HFL     : chr  "Wood" "Fuel oil, kerosene, etc." "Utility gas" "Utility gas" ...
##  $ RMSP    : int  4 7 8 8 3 7 4 4 4 6 ...
##  $ TEN     : chr  "Rented" "Owned with mortgage or loan" "Owned free and clear" "Owned free and clear
##  $ VALP    : int  NA 225000 315000 200000 95000 90000 200000 NA 225000 370000 ...
##  $ YBL     : chr  "1939 or earlier" "1939 or earlier" "1939 or earlier" "1950 to 1959" ...
##  $ R18     : chr  "1 or more" "none" "none" "none" ...
##  $ R60     : chr  "none" "none" "1 or more" "1 or more" ...
```

```
head(data)
```

```
##   SERIALNO NP TYPE                        ACR BDSP                        BLD
## 1       70  4    1 House on less than one acre    2 One-family house detached
## 2      163  2    1 House on less than one acre    2 One-family house detached
## 3      178  1    1 House on less than one acre    3 One-family house detached
## 4      243  2    1 House on less than one acre    4 One-family house detached
## 5      300  1    1 House on less than one acre    2    Mobile home or trailer
## 6      603  3    1 House on less than one acre    3 One-family house detached
##   ELEP FULP GASP                      HFL RMSP                         TEN
## 1   70    2    3                     Wood    4                      Rented
## 2  100  600    3 Fuel oil, kerosene, etc.    7 Owned with mortgage or loan
## 3   60    2  110             Utility gas    8        Owned free and clear
## 4   80    2   20             Utility gas    8        Owned free and clear
```

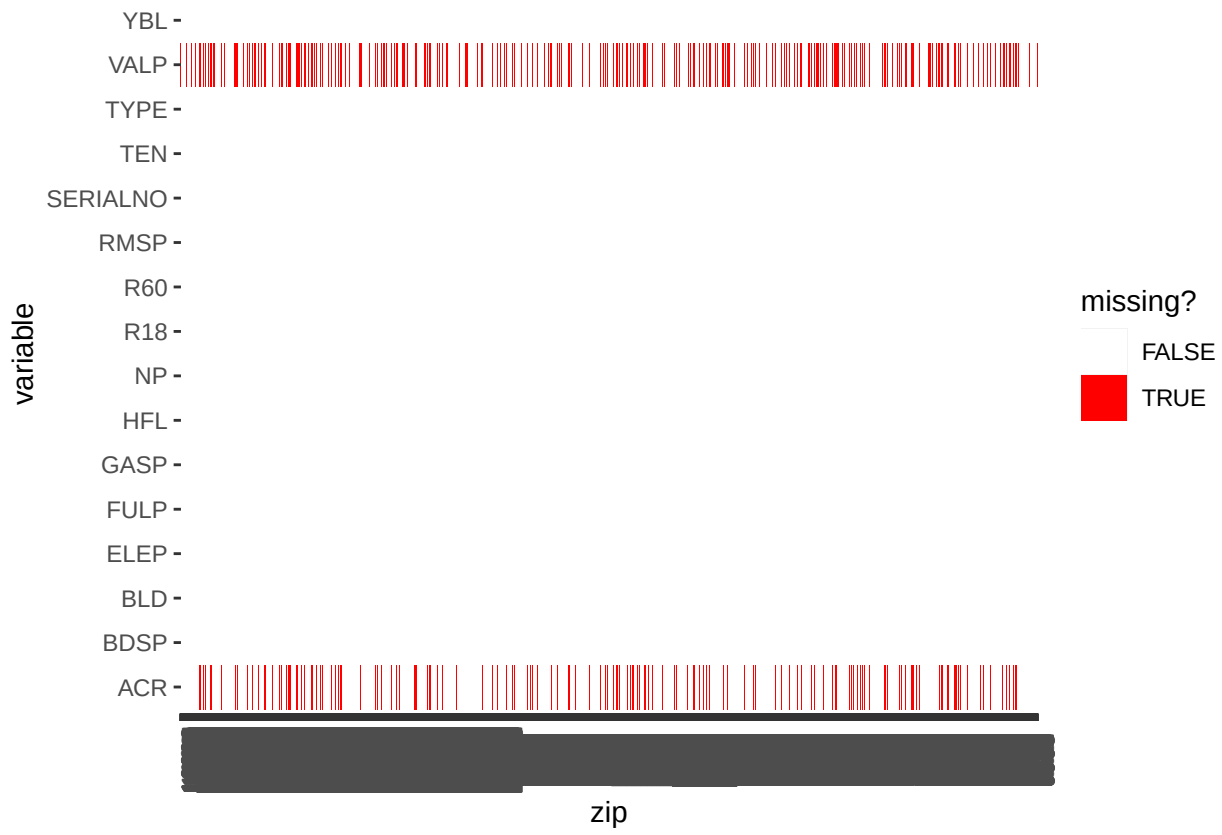
```
## 5  150    2    3             Electricity    3        Owned free and clear
## 6  200    2   20             Electricity    7 Owned with mortgage or loan
##     VALP             YBL       R18       R60
## 1     NA 1939 or earlier 1 or more      none
## 2 225000 1939 or earlier      none      none
## 3 315000 1939 or earlier      none 1 or more
## 4 200000    1950 to 1959      none 1 or more
## 5  95000    1990 to 1999      none 1 or more
## 6  90000 1939 or earlier 1 or more      none
```

Check for missing values

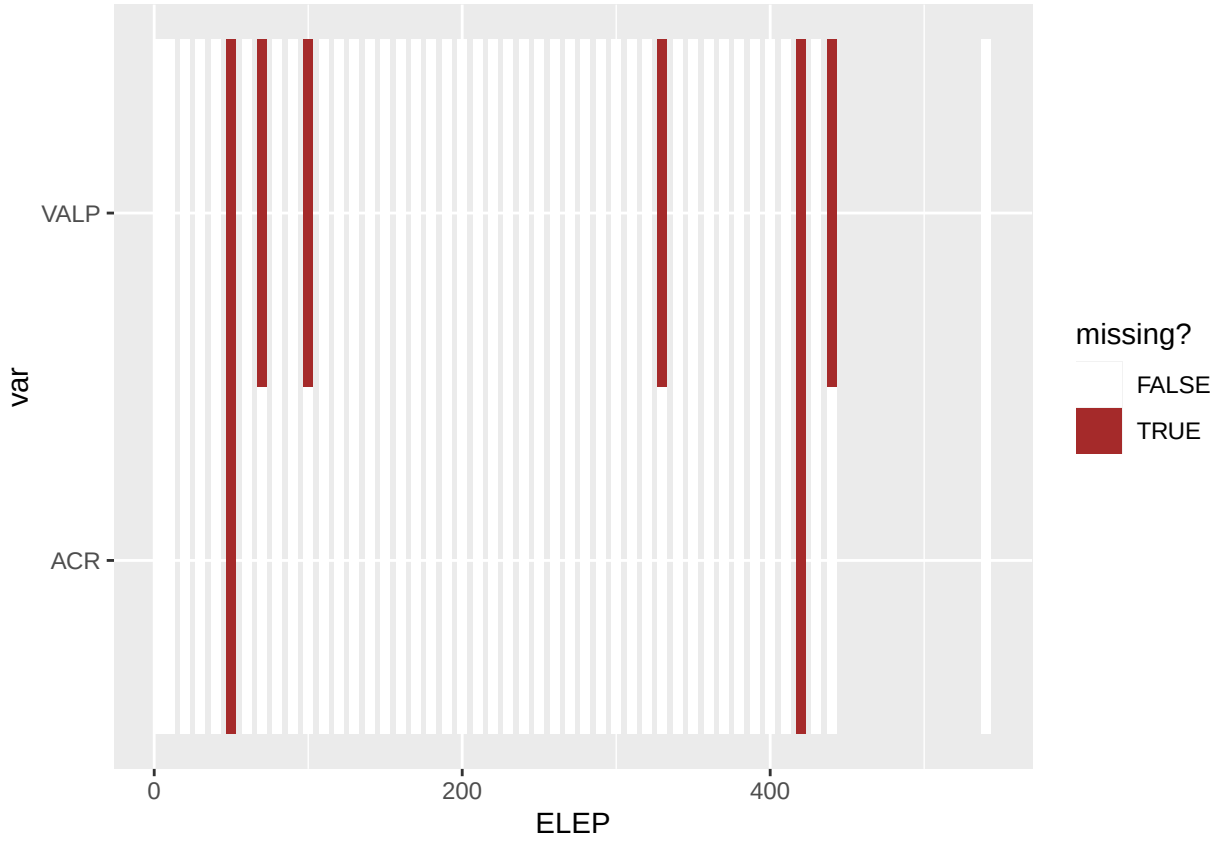
```
colSums(is.na(data))
```

```
## SERIALNO       NP     TYPE      ACR     BDSP      BLD     ELEP     FULP
##        0        0        0     2586        0        0        0        0
##     GASP      HFL     RMSP      TEN     VALP      YBL      R18      R60
##        0        0        0        0     4632        0        0        0
```

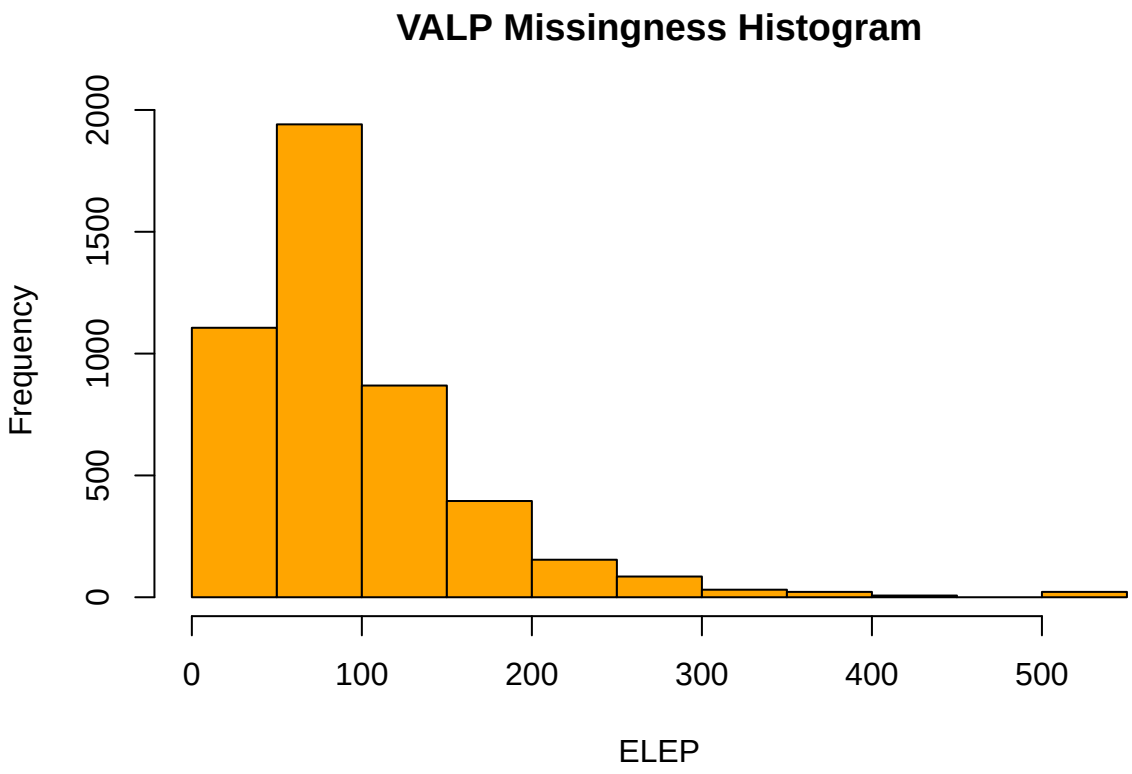
```
data$zip <- rownames(data)
df_long <- gather(data, variable, value, -zip)

qplot(zip, variable, data = df_long, geom= "tile",
      fill = is.na(value)) +
  scale_fill_manual("missing?" , values = c('TRUE'="red", 'FALSE' = "white")) +
  theme (axis.text.x = element_text(angle=90))
```



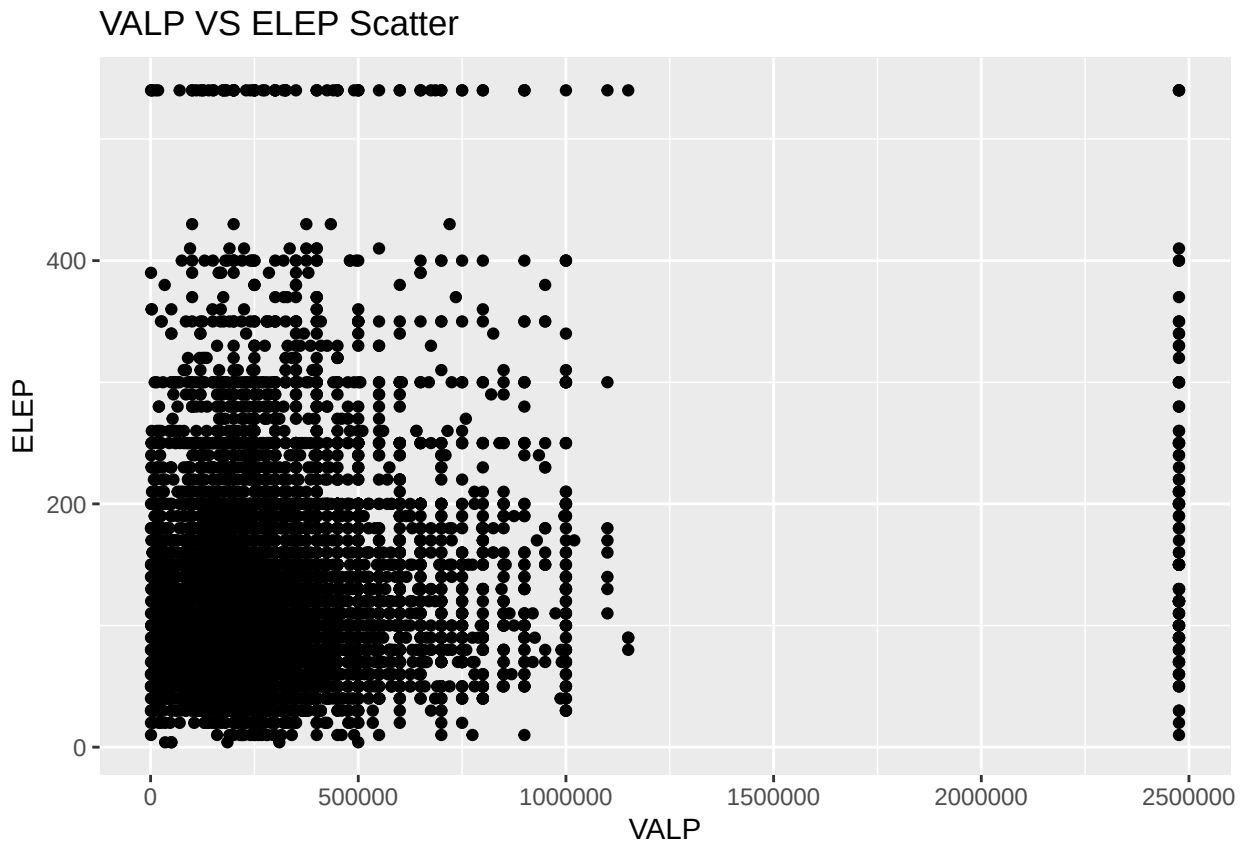

```
data$zip <- rownames(data)
dfN <- data[, c('ACR', 'ELEP', 'VALP')]
dfN_long <- gather(dfN, var, val, -ELEP)
qplot(ELEP, var, data = dfN_long, geom = "tile",
      fill = is.na(val)) +
  scale_fill_manual("missing?",
                    values = c('TRUE' = "brown", 'FALSE' = "white")) +
  theme(axis.text.x = element_text(angle = 0))
```



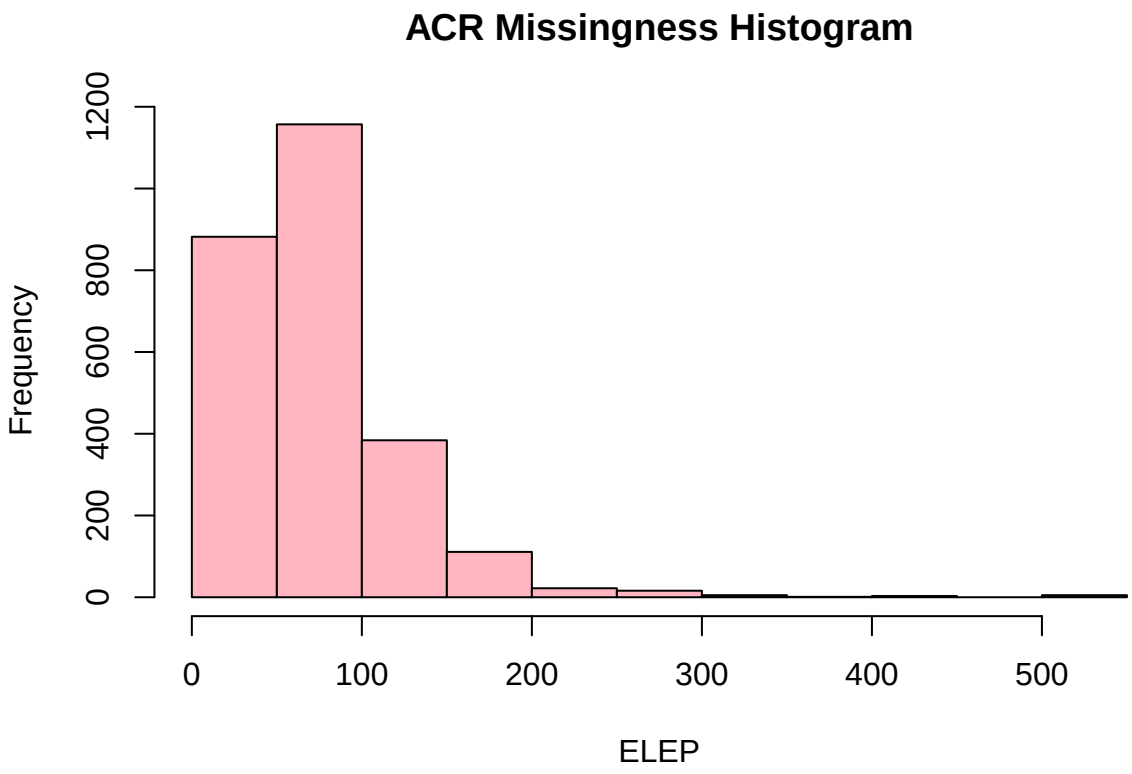

```
valpMissingHist_df <- subset(data[, c('VALP', 'ELEP')], is.na(VALP) == TRUE)
hist(valpMissingHist_df[, 'ELEP'], main = 'VALP Missingness Histogram', xlab = 'ELEP', col = "orange")
```



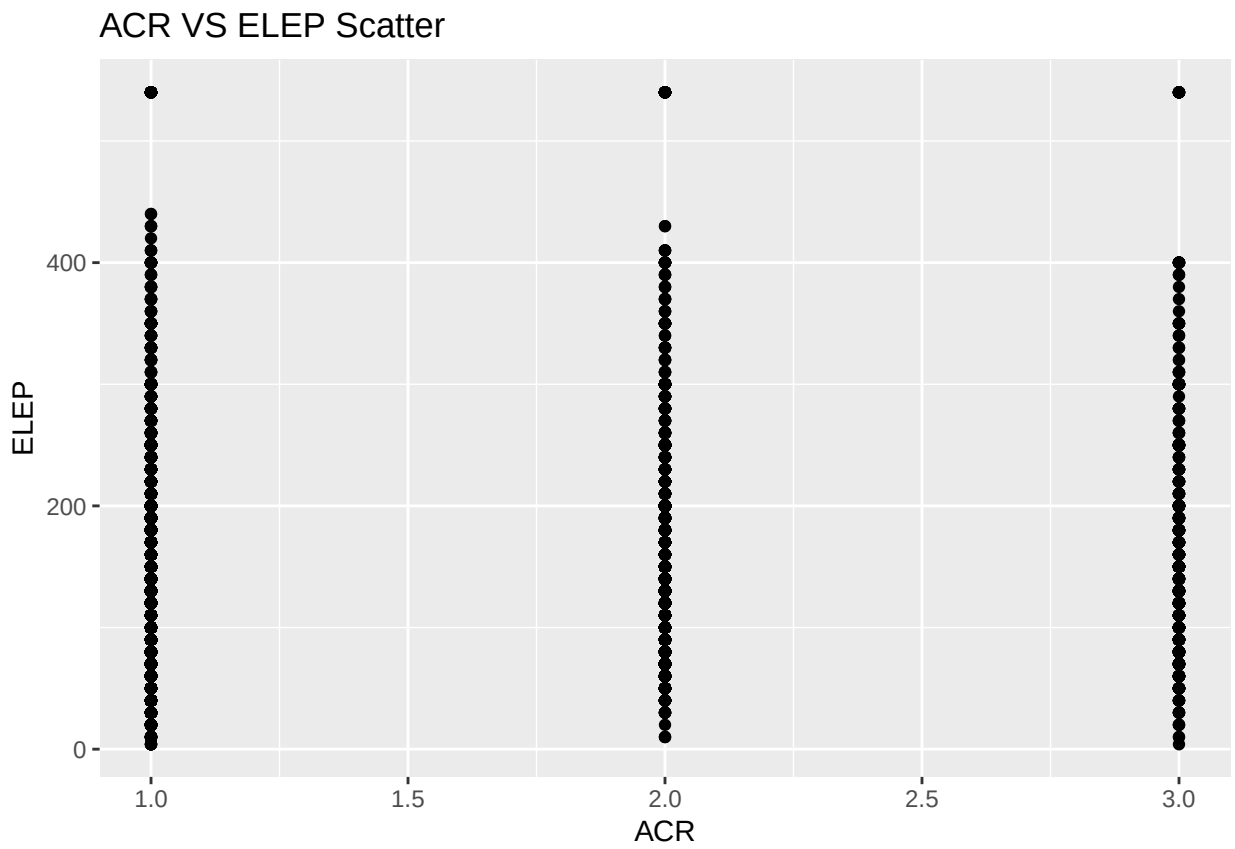

```
valpNoMissingHist_df <- subset(data[, c('VALP', 'ELEP')], is.na(VALP) == FALSE)
qplot(valpNoMissingHist_df[, 'VALP'], valpNoMissingHist_df[, 'ELEP'], main = 'VALP VS ELEP Scatter', yla
```




```
acrMissingHist_df <- subset(data[, c('ACR', 'ELEP')], is.na(ACR) == TRUE)
hist(acrMissingHist_df[, 'ELEP'], main = 'ACR Missingness Histogram', xlab = 'ELEP', col = "lightpink")
```




```
acrNoMissingHist_df <- subset(data[, c('ACR', 'ELEP')], is.na(ACR) == FALSE)
qplot(as.numeric(factor(acrNoMissingHist_df[, 'ACR'])), acrNoMissingHist_df[, 'ELEP'], main = 'ACR VS E
```

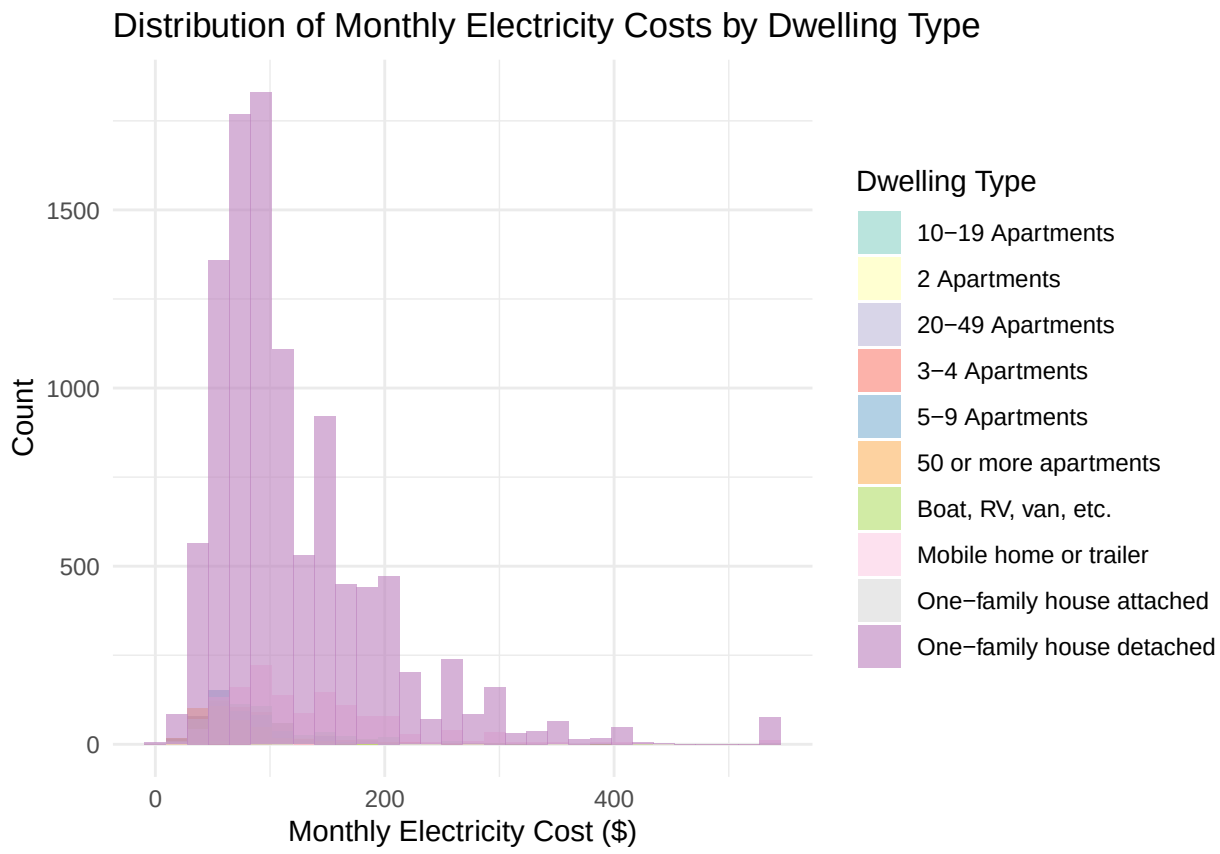


Data cleaning was necessary for this dataset, primarily addressing missing values. Specifically, variables ACR and VALP exhibited a notable number of null values, with 2586 and 4632 missing entries, respectively. Analysis of the missingness distribution histograms indicated that both VALP and ACR were missing-not-at-random, with a higher prevalence of missing values at the lower end of the ELEP histograms.

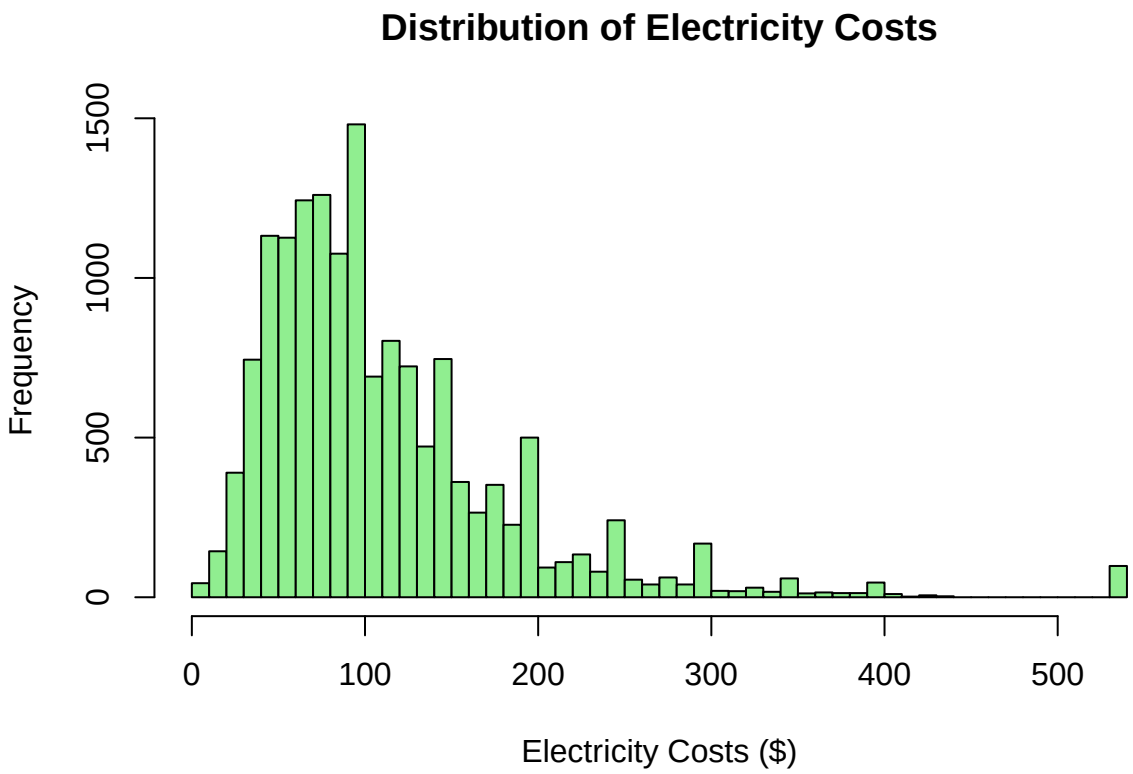
Upon inspection of scatter plots depicting the relationship between ACR/VALP and ELEP, minimal correlation was observed. Given these findings, it is recommended to exclude ACR and VALP from the analysis. These variables represent property value and lot size, and their removal is unlikely to adversely impact the study, as suggested by the scatter plots.

Few plots to understand the Dataset deeper

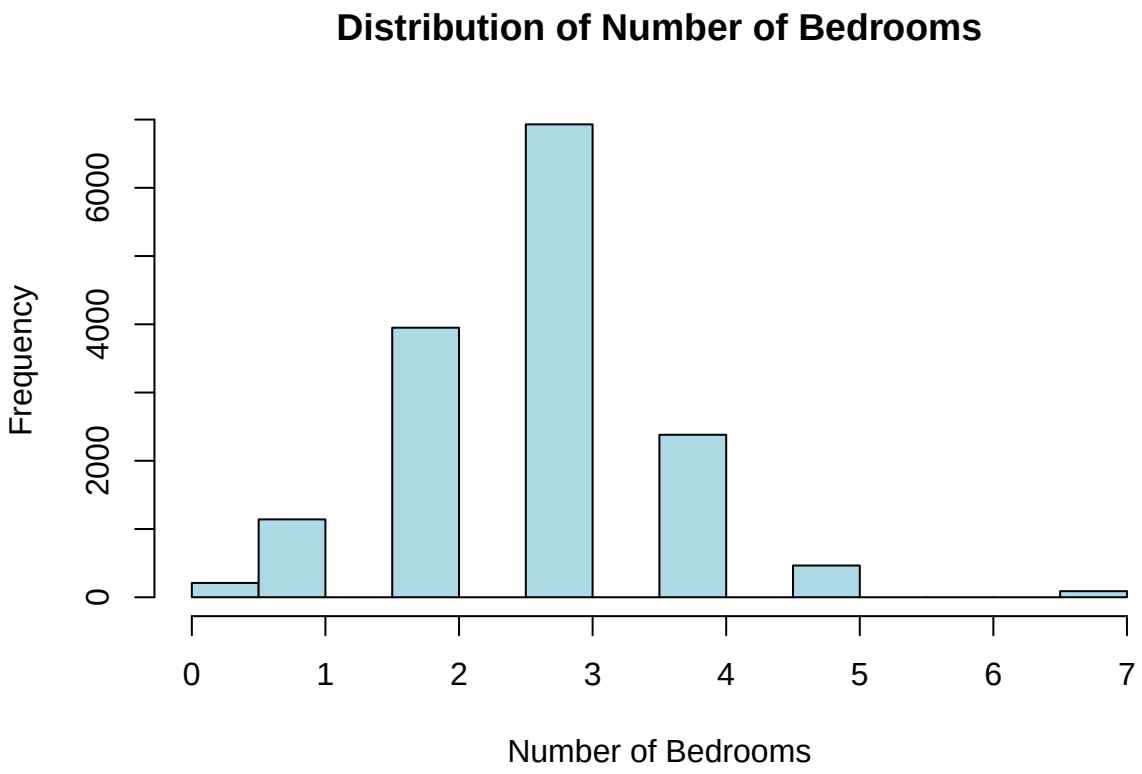
```
# Data Cleaning

#Distribution of Monthly Electricity Costs Across Different Types of Dwellings
data %>%
  ggplot(aes(x = ELEP, fill = as.factor(BLD))) +
  geom_histogram(position = "identity", alpha = 0.6, bins = 30) +
  scale_fill_brewer(palette = "Set3") +
  labs(x = "Monthly Electricity Cost ($)",
       y = "Count",
       fill = "Dwelling Type",
       title = "Distribution of Monthly Electricity Costs by Dwelling Type") +
  theme_minimal()
```



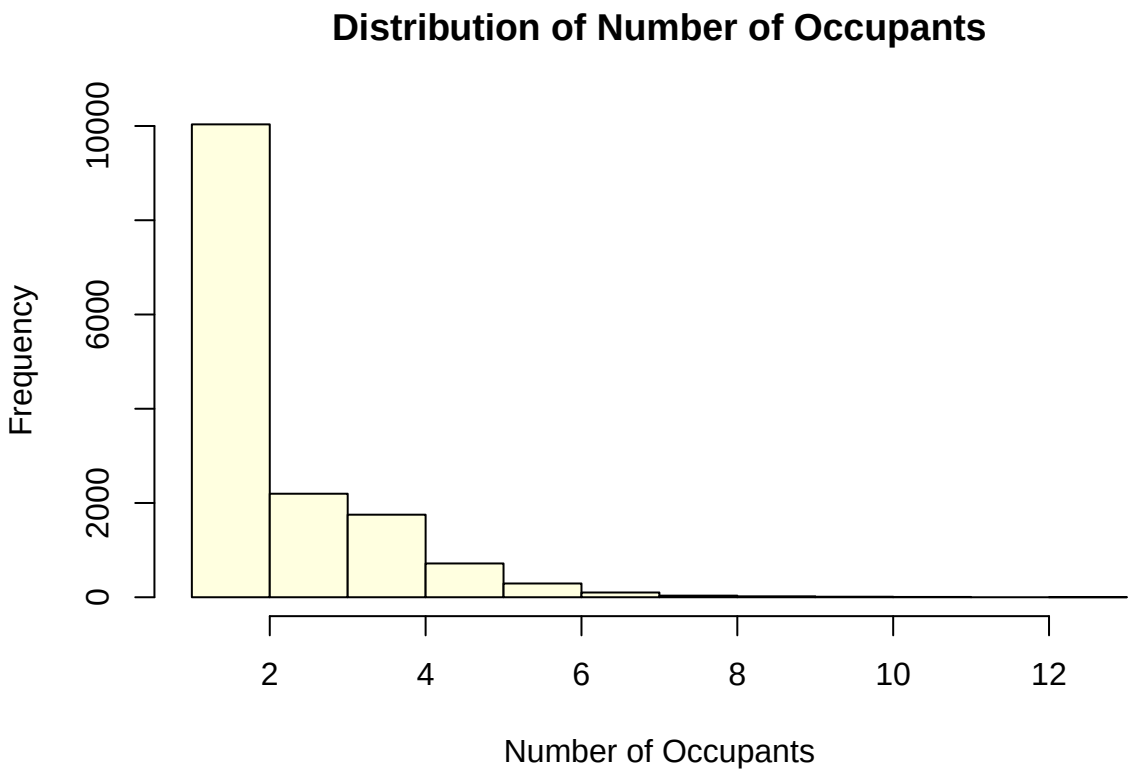

```r
# Convert relevant variables to appropriate data types
data$ELEP <- as.numeric(as.character(data$ELEP))
data$BDSP <- as.numeric(as.character(data$BDSP))
data$NP <- as.numeric(as.character(data$NP))


# Univariate Analysis
# Analyze distribution of electricity costs
hist(data$ELEP, main = "Distribution of Electricity Costs", xlab = "Electricity Costs ($)", breaks = 50
```



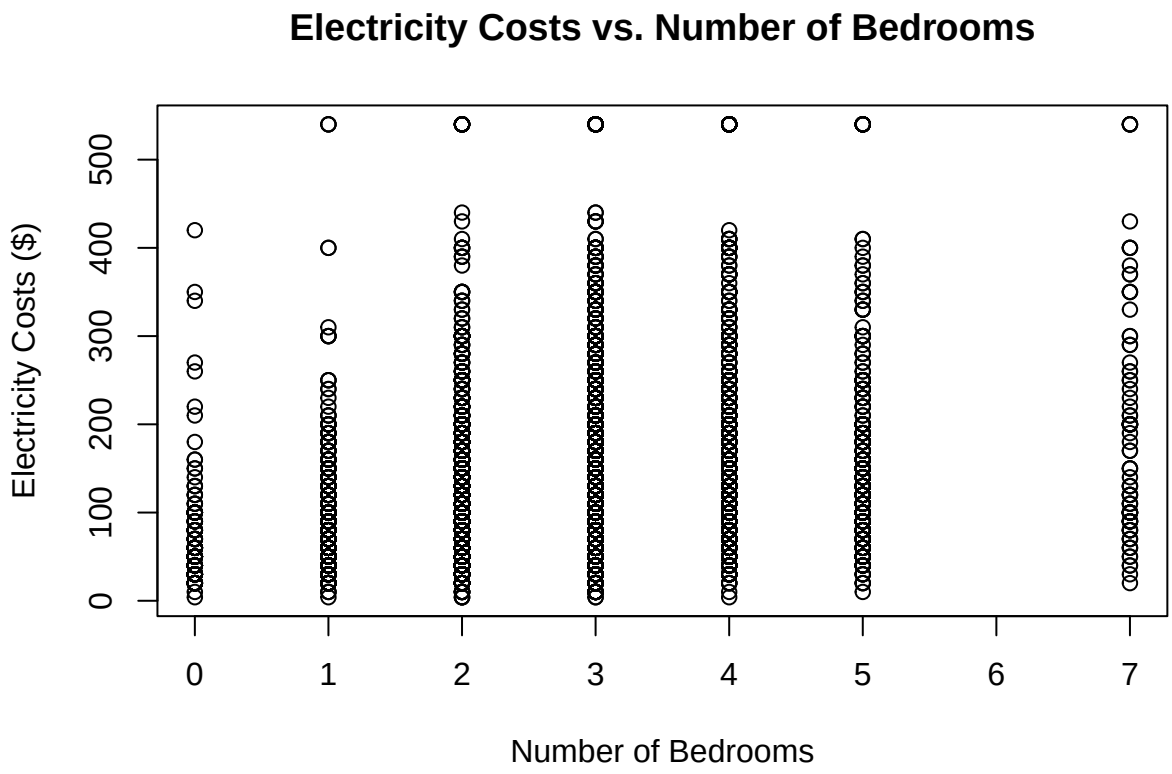

```r
# Bedrooms and Occupants
hist(data$BDSP, main = "Distribution of Number of Bedrooms", xlab = "Number of Bedrooms", breaks = 10, c
```



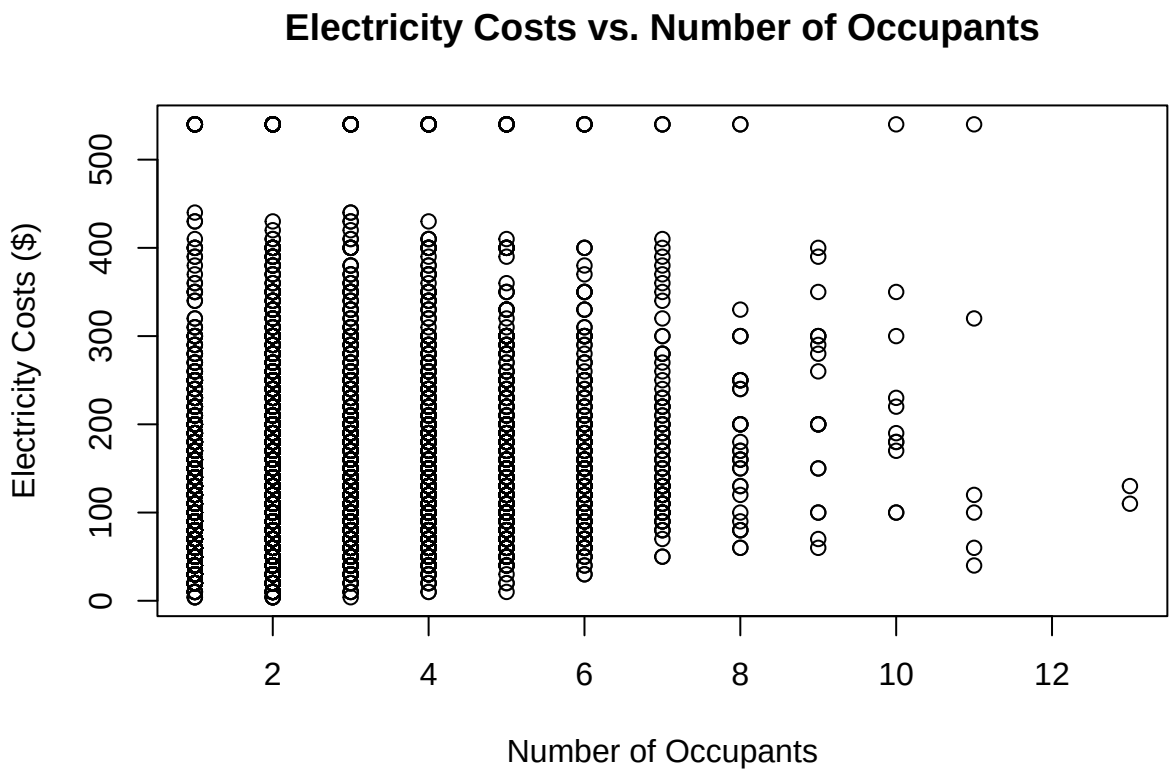

```
hist(data$NP, main = "Distribution of Number of Occupants", xlab = "Number of Occupants", breaks = 10,c
```



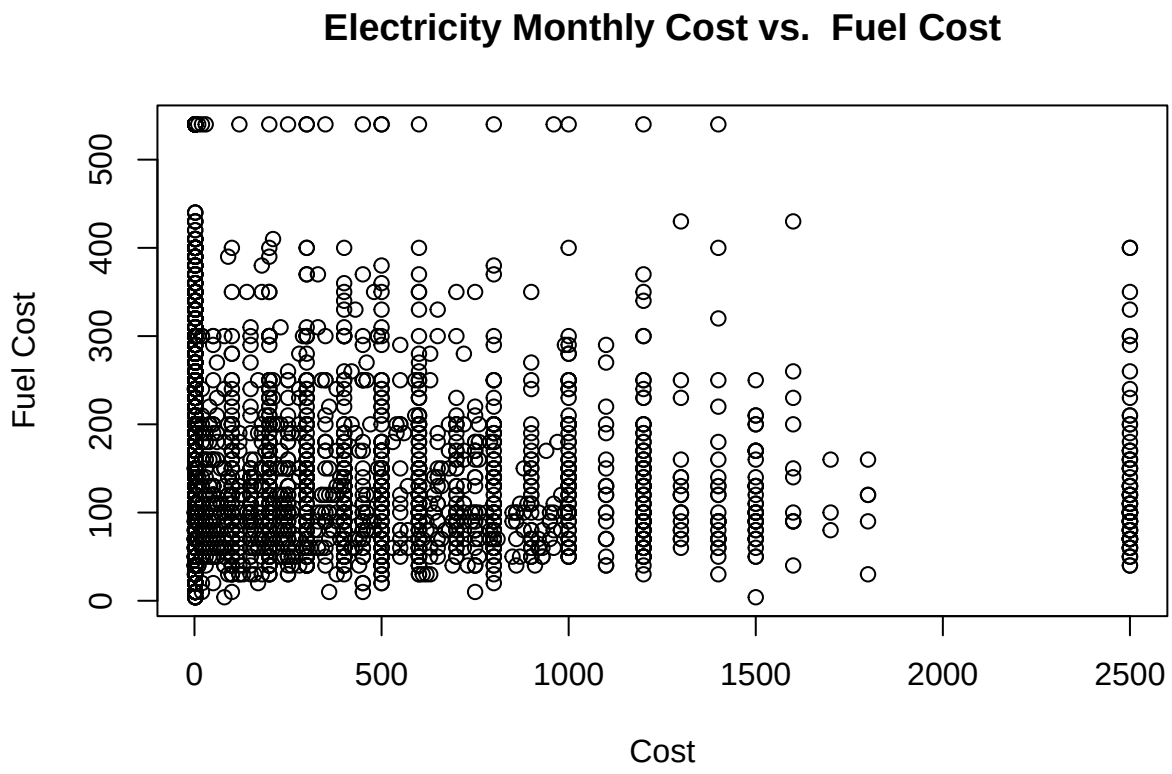

```
# Bivariate Analysis
# Explore relationships between electricity costs and number of bedrooms/occupants
plot(data$BDSP, data$ELEP, main = "Electricity Costs vs. Number of Bedrooms", xlab = "Number of Bedrooms
```



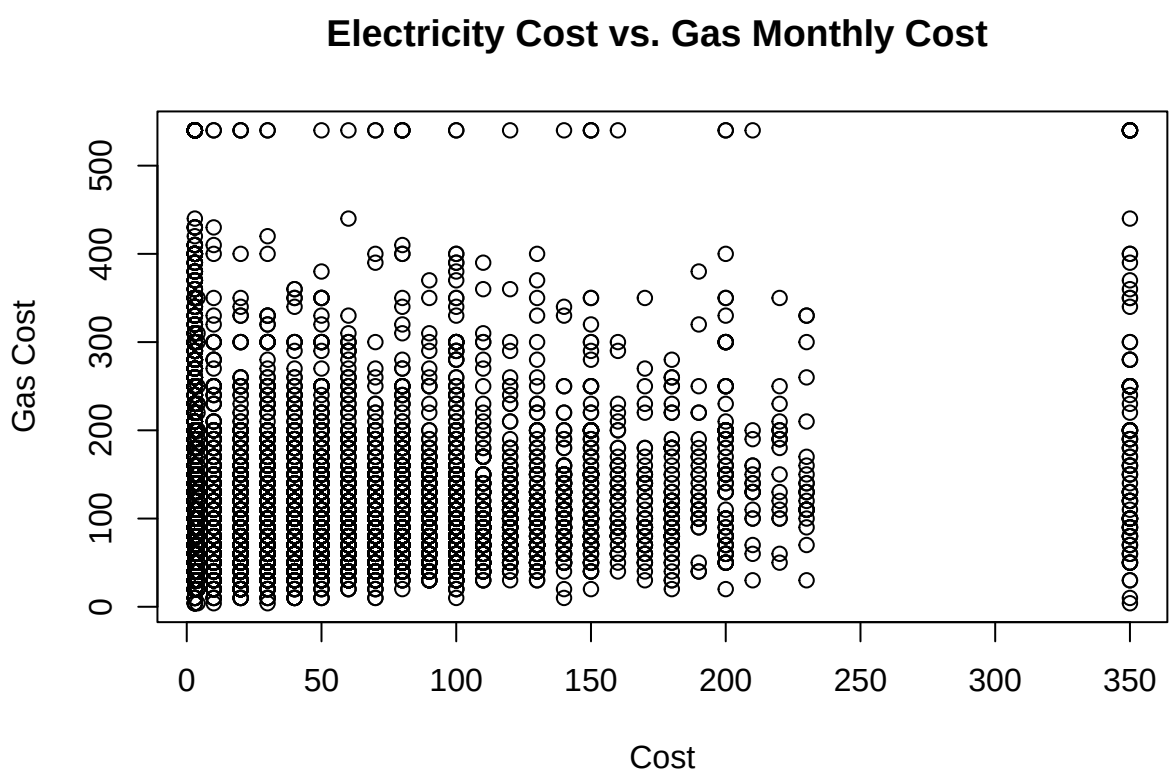

```
plot(data$NP, data$ELEP, main = "Electricity Costs vs. Number of Occupants", xlab = "Number of Occupants
```



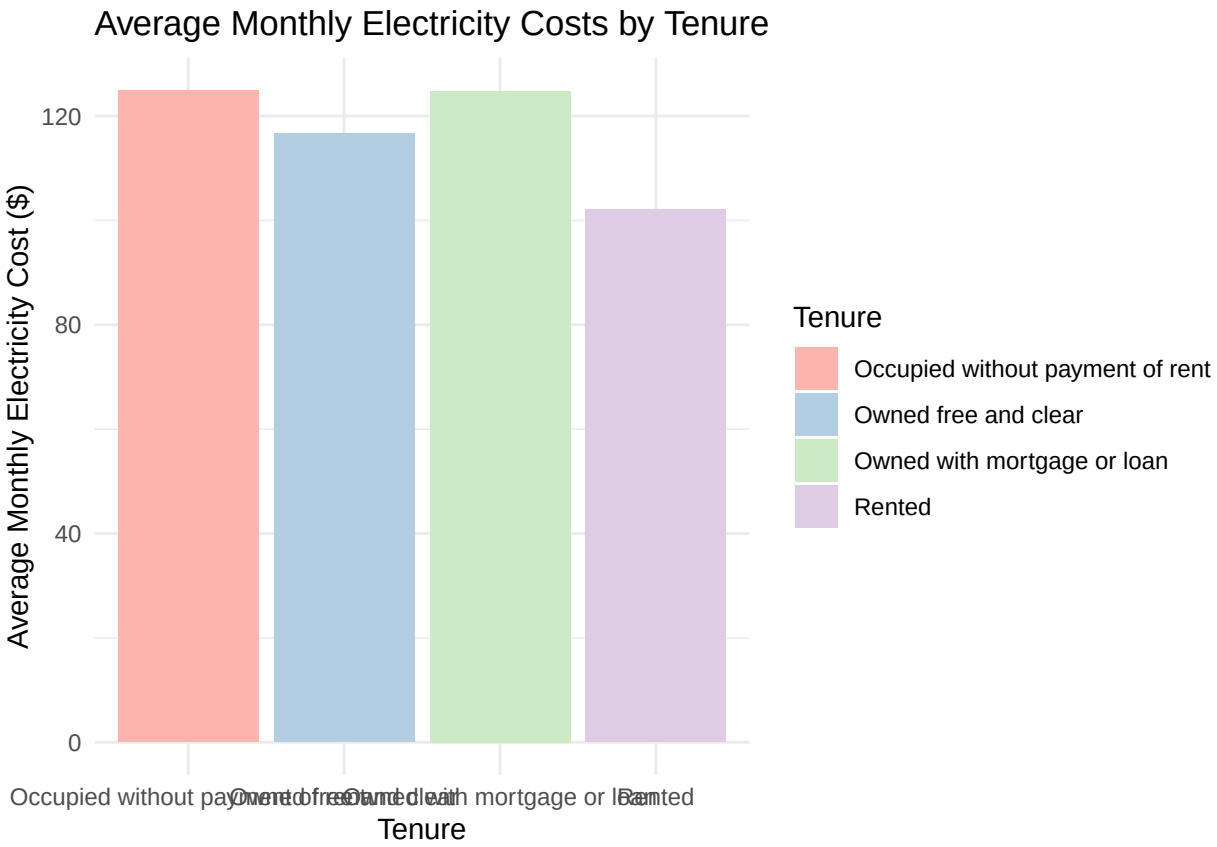

```
plot(data$FULP, data$ELEP, main = 'Electricity Monthly Cost vs.  Fuel Cost', xlab ='Cost'  , ylab = 'Fu
```




```
plot(data$GASP, data$ELEP, main = 'Electricity Cost vs. Gas Monthly Cost ', xlab =  'Cost', ylab = 'Gas
```




```r
# Identify and compare electricity costs between apartments and houses
# Assuming BLD variable indicates the type of housing with codes for apartments and houses
apartments <- data %>% filter(BLD >= 4 & BLD <= 9)  # Assuming codes 4-9 are apartment types
houses <- data %>% filter(BLD == 2 | BLD == 3)      # Assuming codes 2 and 3 are house types

#Electricity Costs by Tenure of the Household
data %>%
  group_by(TEN) %>%
  summarise(Average_ELEP = mean(ELEP, na.rm = TRUE)) %>%
  ggplot(aes(x = as.factor(TEN), y = Average_ELEP, fill = as.factor(TEN))) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Pastel1") +
  labs(x = "Tenure",
       y = "Average Monthly Electricity Cost ($)",
       fill = "Tenure",
       title = "Average Monthly Electricity Costs by Tenure") +
  theme_minimal()
```



From the plots we can say that there is :

NO SIGNIFICANT VISUAL CORRELATION BETWEEN COST AND NUMBER OF OCCUPANTS

NO SIGNIFICANT VISUAL CORRELATION BETWEEN COST AND GAS PRICS

MAYBE CORRELATION BETWEEN COST AND FUELCOST

MAYBE CORRELATION BETWEEN BETWEEN COST AND NUMBER OF BEDROOMS

Broader Predicting Electricity Costs Using Dwelling Characteristics and Household Demographics

```r
# Assuming ELEP is numeric and other predictors are appropriately coded
model <- lm(ELEP ~ BDSP + NP + as.factor(BLD) + as.factor(R18) + as.factor(R60), data = data)
summary(model)
```

```
##
## Call:
## lm(formula = ELEP ~ BDSP + NP + as.factor(BLD) + as.factor(R18) +
##     as.factor(R60), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205.59  -43.14  -15.50   24.76  467.60
##
```

```
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             33.2344     4.5388   7.322 2.56e-13
## BDSP                                     9.5482     0.7089  13.469  < 2e-16
## NP                                      14.7768     0.6259  23.610  < 2e-16
## as.factor(BLD)2 Apartments              15.0040     5.0525   2.970 0.002986
## as.factor(BLD)20-49 Apartments           2.9533     5.1375   0.575 0.565404
## as.factor(BLD)3-4 Apartments             2.5488     4.5612   0.559 0.576312
## as.factor(BLD)5-9 Apartments            -1.9157     4.6195  -0.415 0.678371
## as.factor(BLD)50 or more apartments     -1.4444     4.8957  -0.295 0.767977
## as.factor(BLD)Boat, RV, van, etc.       14.1724    14.1610   1.001 0.316937
## as.factor(BLD)Mobile home or trailer    43.8528     3.9945  10.978  < 2e-16
## as.factor(BLD)One-family house attached  8.8573     4.4568   1.987 0.046898
## as.factor(BLD)One-family house detached 21.1660     3.6476   5.803 6.65e-09
## as.factor(R18)none                       6.7402     1.9416   3.472 0.000519
## as.factor(R60)none                      -6.6491     1.2995  -5.117 3.15e-07
##
## (Intercept)                            ***
## BDSP                                   ***
## NP                                     ***
## as.factor(BLD)2 Apartments             **
## as.factor(BLD)20-49 Apartments
## as.factor(BLD)3-4 Apartments
## as.factor(BLD)5-9 Apartments
## as.factor(BLD)50 or more apartments
## as.factor(BLD)Boat, RV, van, etc.
## as.factor(BLD)Mobile home or trailer   ***
## as.factor(BLD)One-family house attached *
## as.factor(BLD)One-family house detached ***
## as.factor(R18)none                     ***
## as.factor(R60)none                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.98 on 15152 degrees of freedom
## Multiple R-squared:  0.1338, Adjusted R-squared:  0.1331
## F-statistic: 180.1 on 13 and 15152 DF,  p-value: < 2.2e-16
```

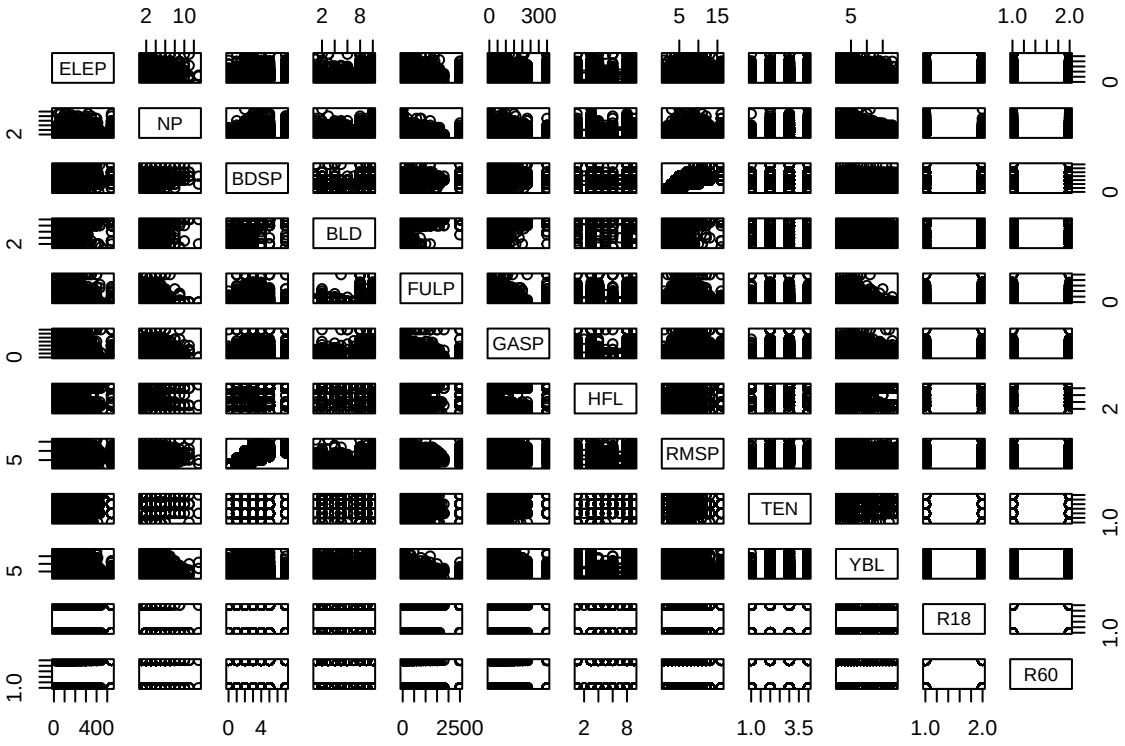
The predictive model described serves as an initial approach to understand the relationship between electricity costs and various household and dwelling characteristics. It provides a broad overview rather than a deep analytical prediction, primarily because it relies on linear relationships and assumes a simplistic interaction between variables. While useful for identifying potential predictors and their general direction of influence, this model does not account for more complex dynamics, non-linear relationships, or interactions between predictors that could significantly affect accuracy. It's a starting point for exploratory analysis, meant to highlight areas for further, more sophisticated modeling efforts, such as incorporating polynomial terms, interaction effects, or using advanced techniques.

```
panel.hist <- function(x, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks
  nB <- length(breaks)
```

```
  y <- h$counts
  y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "white", ...)
}


panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y, use = "complete.obs"))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste(prefix, txt, sep = "")
  if (missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex =  cex.cor * (1 + r) / 2)
}
```

Correlation check

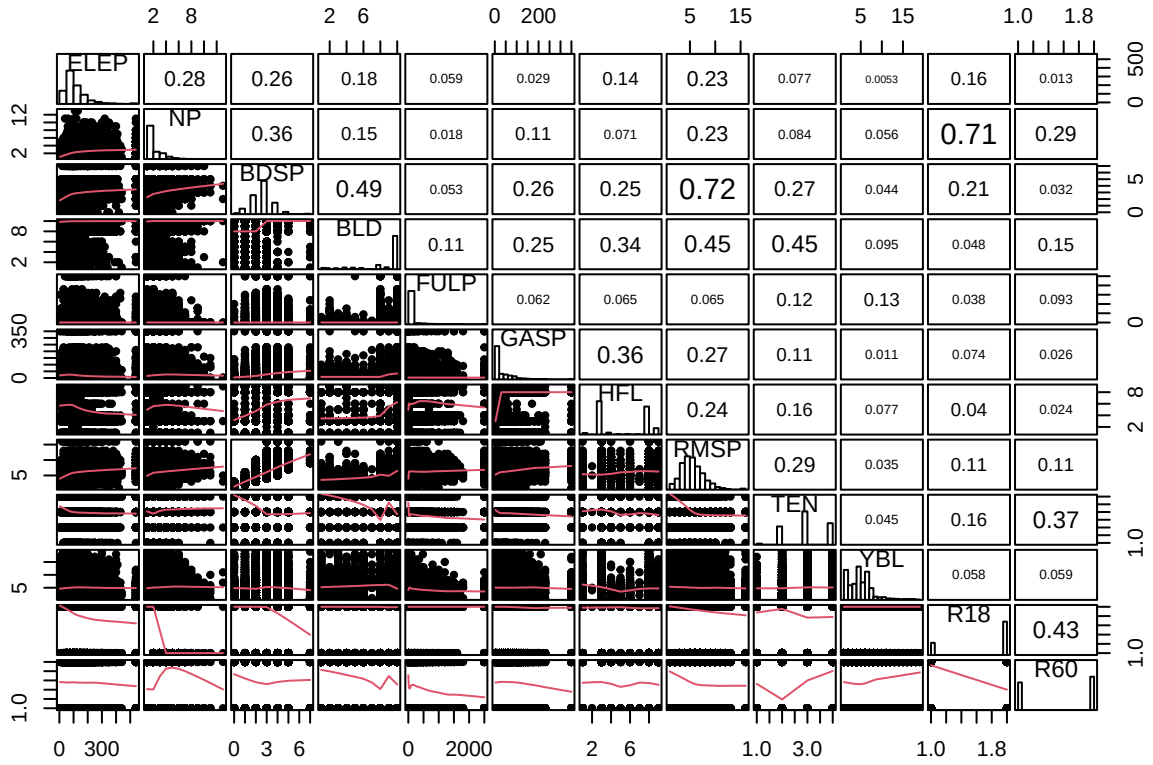
```
# Correlation check
data.corr <- data[, c('ELEP', 'NP', 'BDSP', 'BLD', 'FULP', 'GASP', 'HFL', 'RMSP', 'TEN', 'YBL', 'R18',
data.corr[c('BLD', 'HFL', 'TEN', 'YBL', 'R18', 'R60')] <- lapply(data.corr[c('BLD', 'HFL', 'TEN', 'YBL'

plot(data.corr)
```



```
summary(data.corr)
```

```
##       ELEP              NP              BDSP             BLD
##  Min.   :  4.0   Min.   : 1.000   Min.   :0.000   Min.   : 1.000
##  1st Qu.: 70.0   1st Qu.: 1.000   1st Qu.:2.000   1st Qu.: 8.000
##  Median :100.0   Median : 2.000   Median :3.000   Median :10.000
##  Mean   :116.2   Mean   : 2.403   Mean   :2.789   Mean   : 8.695
##  3rd Qu.:150.0   3rd Qu.: 3.000   3rd Qu.:3.000   3rd Qu.:10.000
##  Max.   :540.0   Max.   :13.000   Max.   :7.000   Max.   :10.000
##       FULP              GASP             HFL              RMSP
##  Min.   :   1.00   Min.   :  3.00   Min.   :1.000   Min.   : 1.00
##  1st Qu.:   2.00   1st Qu.:  3.00   1st Qu.:3.000   1st Qu.: 4.00
##  Median :   2.00   Median :  3.00   Median :4.000   Median : 6.00
##  Mean   :  85.25   Mean   : 35.68   Mean   :5.488   Mean   : 6.01
##  3rd Qu.:   2.00   3rd Qu.: 50.00   3rd Qu.:8.000   3rd Qu.: 7.00
##  Max.   :2500.00   Max.   :350.00   Max.   :9.000   Max.   :16.00
##       TEN              YBL              R18              R60
##  Min.   :1.000   Min.   : 1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.: 3.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median : 5.00   Median :2.000   Median :2.000
##  Mean   :3.013   Mean   : 5.37   Mean   :1.733   Mean   :1.547
##  3rd Qu.:4.000   3rd Qu.: 7.00   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :19.00   Max.   :2.000   Max.   :2.000
```

```
pairs(data.corr, upper.panel = panel.cor, diag.panel = panel.hist, lower.panel = panel.smooth, gap = 1/5
```



Once the analysis of data missingness was completed, steps were taken to rank relevant fields and potentially eliminate any practically irrelevant ones from the dataset. Following identification, these fields underwent a correlation check with ELEP before considering complete removal.

For several categorical fields (BLD, HFL, TEN, YBL, R18, R60), conversion to numeric was necessary for correlation assessment. The complete scatter matrix is provided in Appendix X, and correlation metrics are detailed below. Upon examination, none of the practically relevant factors exhibited noteworthy correlations, confirming their insignificance in the model. Notably, RMSP and BDSP displayed a high correlation (0.72), leading to the recommendation of excluding RMSP to avoid redundancy.

While FULP, GASP, and YBL did not exhibit significant correlations, their relevance is confirmed to be minimal. Therefore, it is suggested to retain BLD, HFL, BDSP, NP, and ELEP for continued consideration.

Field	Description	Relevance to ELEP	Correlation to ELEP
SERIALNO TYPE	Serial Number Type of Unit	No No (always 1) Yes	N/A N/A 0.28 - 0.26
NP ACR BDSP BLD	Number of Persons in the house	(requires	0.18 1 0.06 0.029 0.14
ELEP FULP GASP	Lot size Number of Bedrooms	modification) - Yes	0.23 0.077 - 0.01 0.16
HFL RMSP TEN	Units in Structure Electricity	Yes (requires	0.013
VALP YBL R18 R60	(monthly Cost) Yearly Fuel Cost	modification) Yes	
	(excluding gas & electricity) Gas	Maybe Maybe Yes	
	(monthly cost) House Heating	(requires	
	Fuel Number of Rooms Tenure	modification) Yes	
	Property value When structure	(Redundant) No -	
	first built Presence of persons	Maybe No No	
	under 18 Presence of persons		
	over 59		

Cleaning the Columns

```
# DECLARE FINAL CUSTOM DATAFRAAME
df_custom <- data[, c('ELEP', 'FULP', 'GASP', 'YBL', 'BLD', 'HFL', 'BDSP', 'NP')]
head(df_custom)
```

```
##   ELEP FULP GASP             YBL                        BLD
## 1   70    2    3 1939 or earlier One-family house detached
## 2  100  600    3 1939 or earlier One-family house detached
## 3   60    2  110 1939 or earlier One-family house detached
## 4   80    2   20    1950 to 1959 One-family house detached
## 5  150    2    3    1990 to 1999    Mobile home or trailer
## 6  200    2   20 1939 or earlier One-family house detached
##                         HFL BDSP NP
## 1                      Wood    2  4
## 2 Fuel oil, kerosene, etc.    2  2
## 3              Utility gas    3  1
## 4              Utility gas    4  2
## 5              Electricity    2  1
## 6              Electricity    3  3
```

```
new_df <- df_custom %>%
  filter(!(BLD %in% c('Mobile home or trailer', 'Boat, RV, van, etc.'))) %>%
  mutate(
    BLD_Adjusted = case_when(
```

```
      grepl('house', BLD, fixed = TRUE) ~ 'House',
      TRUE ~ 'Apartment'
    ),
    HFL_Adjusted = case_when(
      grepl('Electricity', HFL, fixed = TRUE) ~ 'Electricity',
      TRUE ~ 'Not Electricity'
    ),
    YBL_Adjusted = ifelse(YBL >= 2005, '2005 to 2015', as.character(YBL))
  )

head(new_df)
```

```
##   ELEP FULP GASP             YBL                       BLD
## 1   70    2    3 1939 or earlier One-family house detached
## 2  100  600    3 1939 or earlier One-family house detached
## 3   60    2  110 1939 or earlier One-family house detached
## 4   80    2   20    1950 to 1959 One-family house detached
## 5  200    2   20 1939 or earlier One-family house detached
## 6  120    2    3    1960 to 1969 One-family house detached
##                         HFL BDSP NP BLD_Adjusted    HFL_Adjusted    YBL_Adjusted
## 1                      Wood    2  4        House Not Electricity 1939 or earlier
## 2 Fuel oil, kerosene, etc.    2  2        House Not Electricity 1939 or earlier
## 3              Utility gas    3  1        House Not Electricity 1939 or earlier
## 4              Utility gas    4  2        House Not Electricity    1950 to 1959
## 5              Electricity    3  3        House     Electricity 1939 or earlier
## 6              Electricity    3  2        House     Electricity    1960 to 1969
```

```
summary(new_df)
```

```
##       ELEP            FULP              GASP            YBL
##  Min.   :  4.0   Min.   :   1.00   Min.   :  3.00   Length:13774
##  1st Qu.: 70.0   1st Qu.:   2.00   1st Qu.:  3.00   Class :character
##  Median :100.0   Median :   2.00   Median : 10.00   Mode  :character
##  Mean   :114.1   Mean   :  82.38   Mean   : 37.87
##  3rd Qu.:140.0   3rd Qu.:   2.00   3rd Qu.: 60.00
##  Max.   :540.0   Max.   :2500.00   Max.   :350.00
##      BLD                HFL                 BDSP             NP
##  Length:13774       Length:13774       Min.   :0.000   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Median :3.000   Median : 2.000
##                                        Mean   :2.811   Mean   : 2.417
##                                        3rd Qu.:3.000   3rd Qu.: 3.000
##                                        Max.   :7.000   Max.   :13.000
##  BLD_Adjusted       HFL_Adjusted       YBL_Adjusted
##  Length:13774       Length:13774       Length:13774
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Explanatory Model for Multiple Regression

Proposed Method

This study employs regression models to estimate the monthly electricity bill difference between apartments and houses in Oregon. We designate BLD as 0 for apartments and 1 for houses. The full model with interactions, represented as $(ELEP|BLD, BDSP, NP) = \beta_0 + \beta_1 BLD + \beta_2 BDSP + \beta_3 NP + \beta_4(BLD * NP) + \beta_5(BLD * BDSP) + \beta_6(BLD * BDSP * NP)$, is compared to a reduced model lacking interaction terms, given by $(ELEP|BLD, BDSP, NP) = \beta_0 + \beta_1 BLD + \beta_2 BDSP + \beta_3 NP$. The comparison favors the full model as the more suitable option (Extra SS F-test, p-value=0.0008967).

To refine the model, insignificant interaction terms and three-variable interaction terms are removed to mitigate noise and over-fitting. The resulting model, $\mu(ELEP|BLD, BDSP, NP) = \beta_0 + \beta_1 BLD + \beta_2 BDSP + \beta_3 NP + \beta_4(BDSP * NP) + \beta_5(BLD * BDSP)$, is considered post-training. However, a comparison of AIC and BIC values suggests that the reduced model, without interaction terms, is more appropriate for this study.

```
new_df$HA <- "bbbbb"
new_df$HA[which(grepl("house",new_df$BLD))] <- "house"
new_df$HA[which(grepl("apartment",new_df$BLD,ignore.case=TRUE))] <- "apt"

df_final <- subset(new_df, HA!="bbbbb" )
summary(df_final)
```

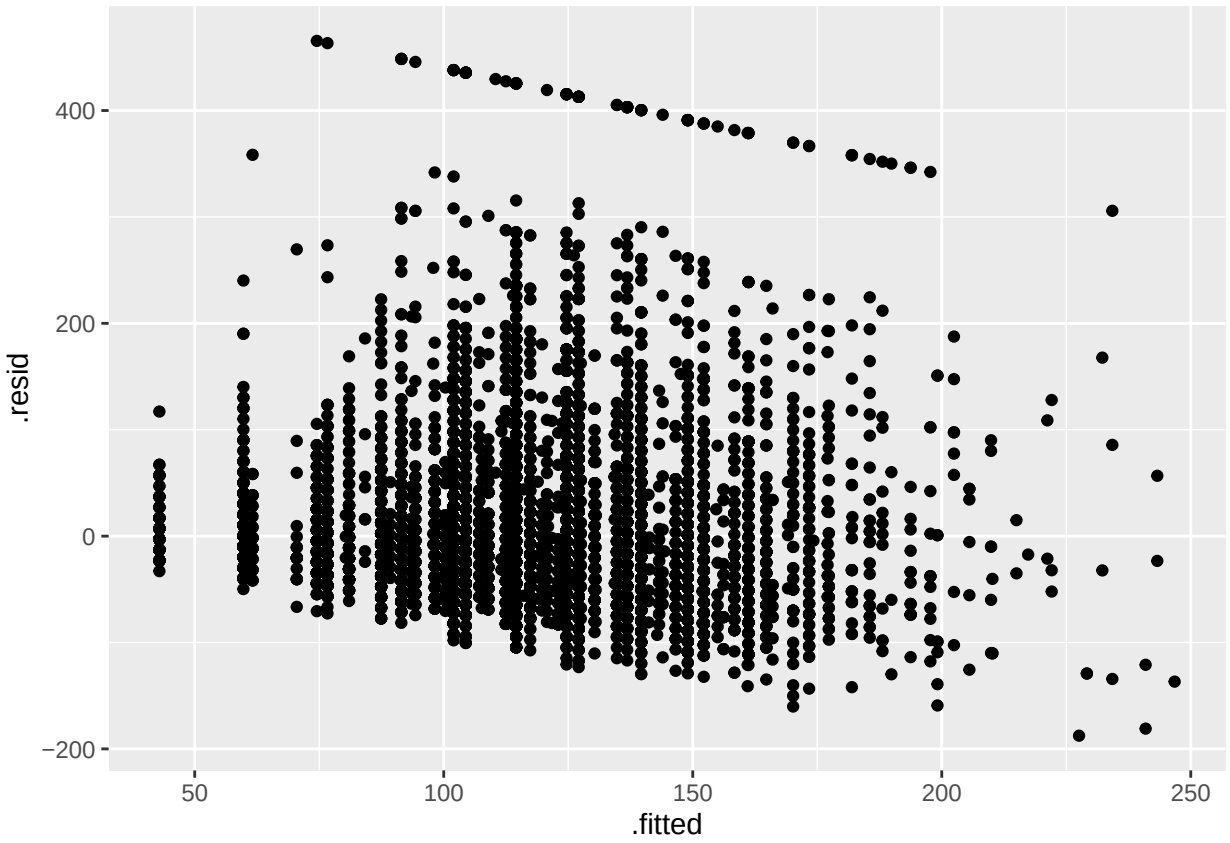
```
##       ELEP             FULP              GASP            YBL
##  Min.   :  4.0   Min.   :   1.00   Min.   :  3.00   Length:13774
##  1st Qu.: 70.0   1st Qu.:   2.00   1st Qu.:  3.00   Class :character
##  Median :100.0   Median :   2.00   Median : 10.00   Mode  :character
##  Mean   :114.1   Mean   :  82.38   Mean   : 37.87
##  3rd Qu.:140.0   3rd Qu.:   2.00   3rd Qu.: 60.00
##  Max.   :540.0   Max.   :2500.00   Max.   :350.00
##      BLD                HFL                BDSP             NP
##  Length:13774       Length:13774       Min.   :0.000   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:2.000   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Median :3.000   Median : 2.000
##                                        Mean   :2.811   Mean   : 2.417
##                                        3rd Qu.:3.000   3rd Qu.: 3.000
##                                        Max.   :7.000   Max.   :13.000
##  BLD_Adjusted       HFL_Adjusted       YBL_Adjusted            HA
##  Length:13774       Length:13774       Length:13774       Length:13774
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
# Fitting rich model with interactions
full_model <- lm(ELEP ~ HA * BDSP * NP, data = df_final)
summary(full_model)
```

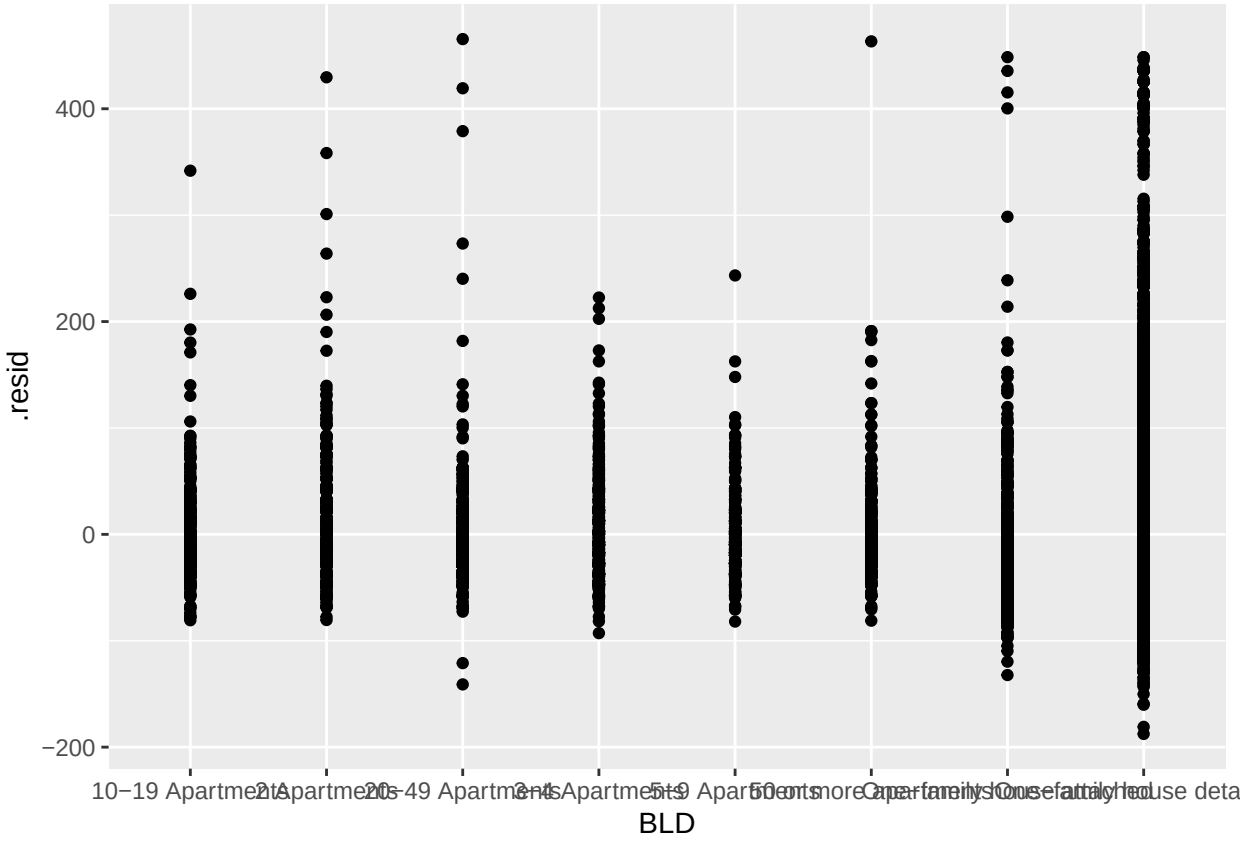
```
##
## Call:
## lm(formula = ELEP ~ HA * BDSP * NP, data = df_final)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -187.57  -44.31  -14.54   25.23  465.47
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       24.192      5.496   4.402 1.08e-05 ***
## HAhouse           32.600      7.151   4.559 5.20e-06 ***
## BDSP              20.848      2.791   7.468 8.60e-14 ***
## NP                18.715      2.939   6.367 1.98e-10 ***
## HAhouse:BDSP      -9.970      3.139  -3.176 0.001495 **
## HAhouse:NP        -5.013      3.369  -1.488 0.136722
## BDSP:NP           -3.972      1.134  -3.503 0.000461 ***
## HAhouse:BDSP:NP    3.591      1.226   2.929 0.003402 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.62 on 13766 degrees of freedom
## Multiple R-squared:  0.1249, Adjusted R-squared:  0.1245
## F-statistic: 280.8 on 7 and 13766 DF,  p-value: < 2.2e-16
```

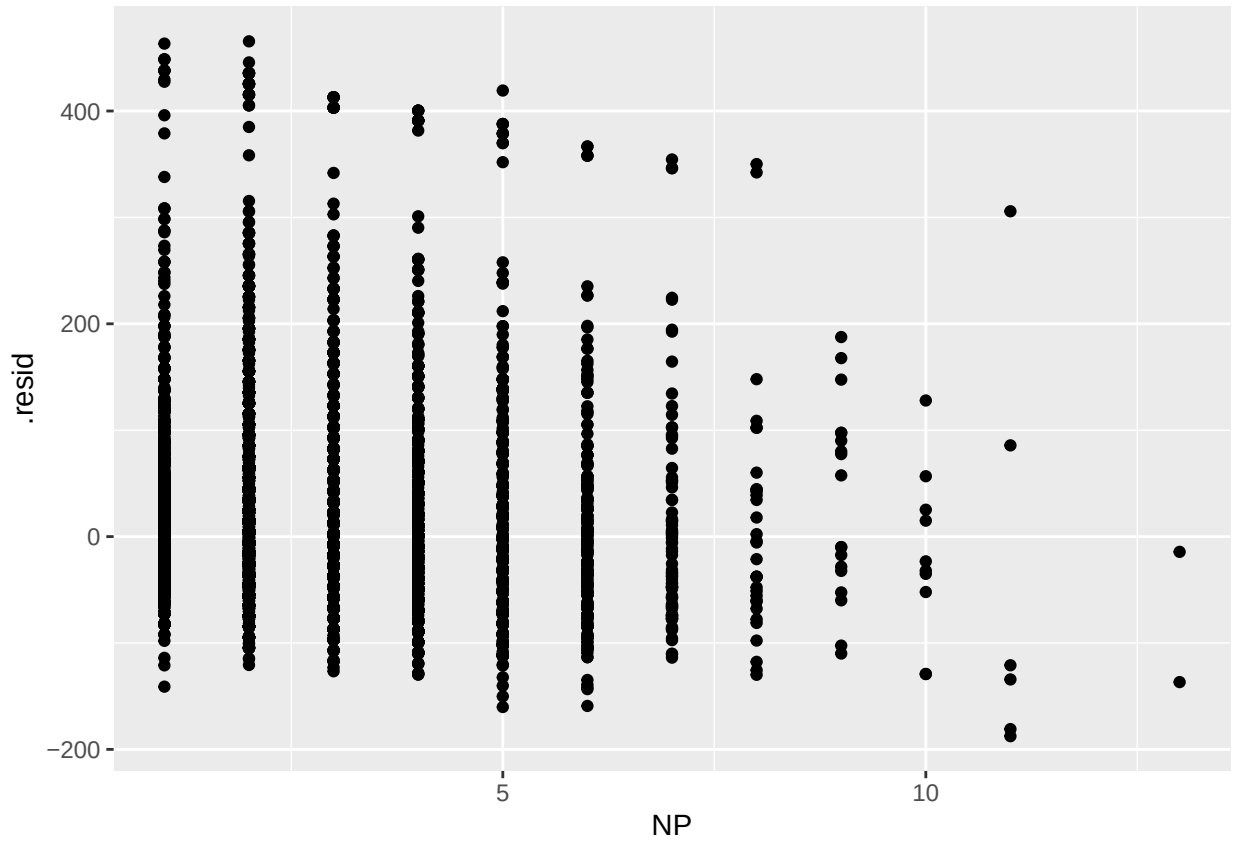
```r
# Residual Plots for full model with interactions
residual_plots <- broom::augment(full_model, data = df_final)
qplot(.fitted, .resid, data = residual_plots)
```



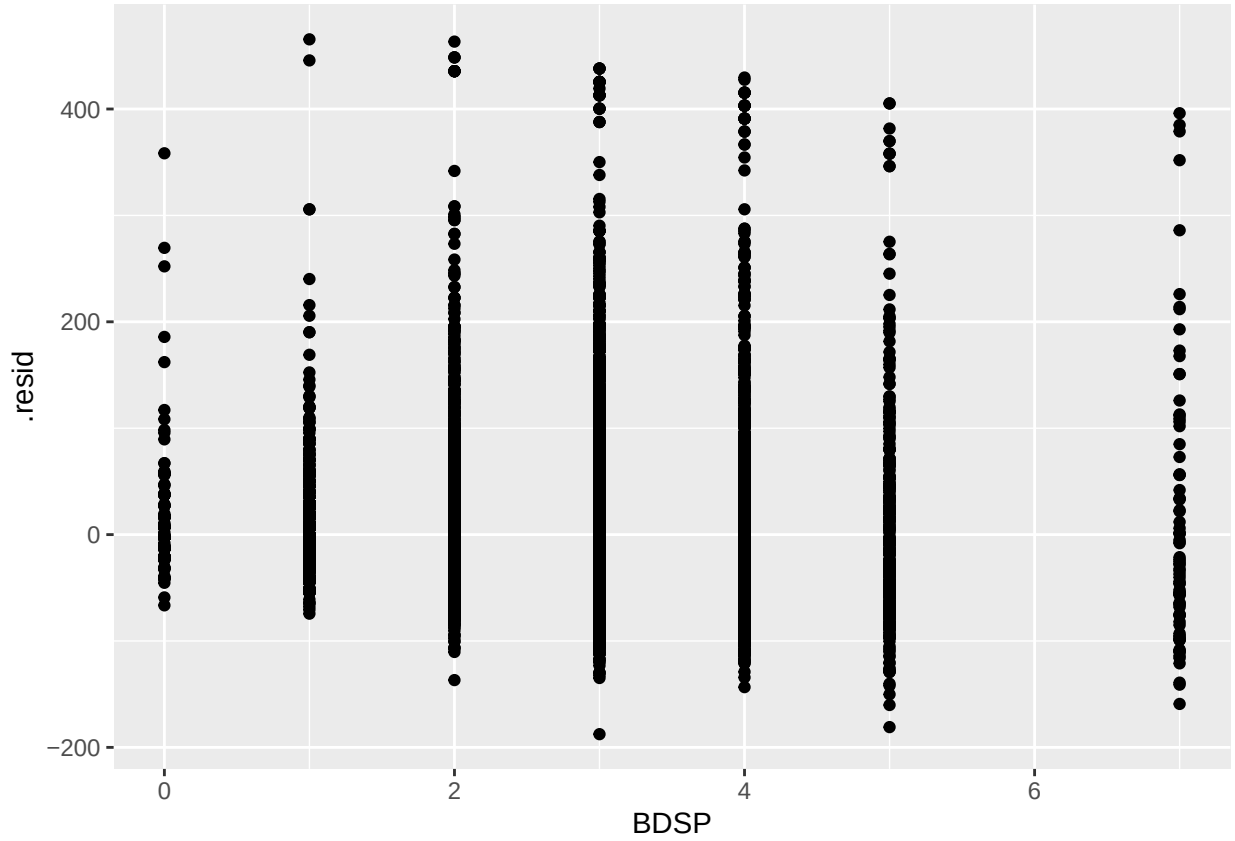
```
qplot(BLD, .resid, data = residual_plots)
```




```
qplot(NP, .resid, data = residual_plots)
```



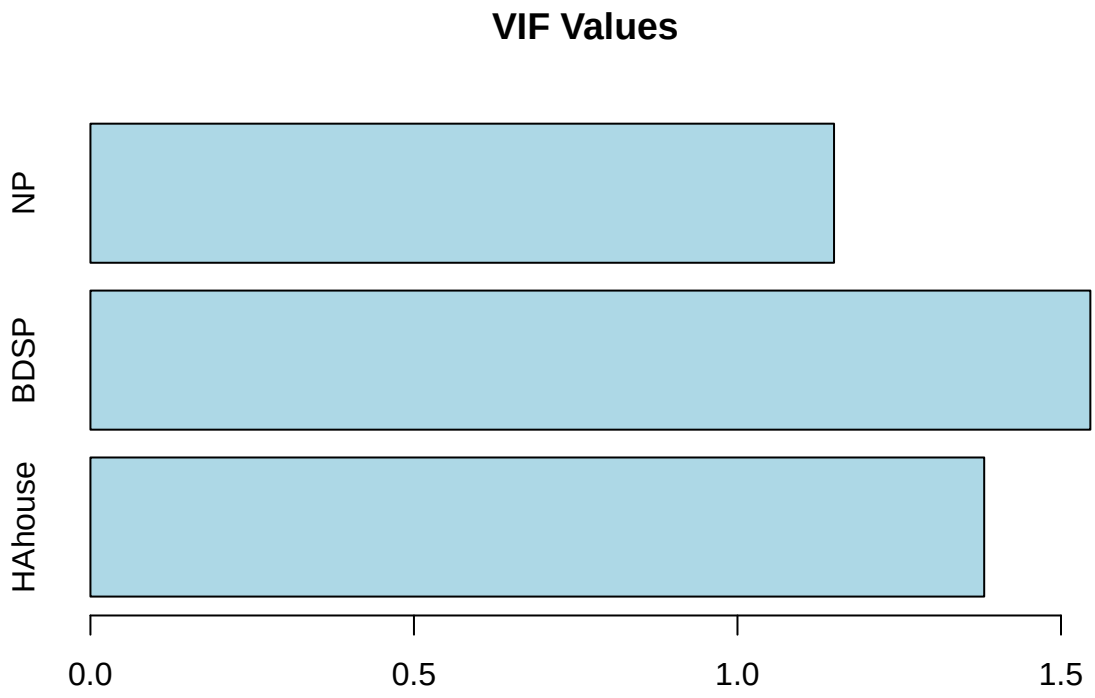

```
qplot(BDSP, .resid, data = residual_plots)
```




```
# Fitting Model without interactions
reduced_model <- lm(ELEP ~ HA + BDSP + NP, data = df_final)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = ELEP ~ HA + BDSP + NP, data = df_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -183.70  -42.98  -14.64   25.13  467.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.9979     1.8687   21.40   <2e-16 ***
## HAhouse      19.4731     1.7932   10.86   <2e-16 ***
## BDSP         10.3674     0.7214   14.37   <2e-16 ***
## NP           12.0356     0.4683   25.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.66 on 13770 degrees of freedom
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.1236
## F-statistic: 648.2 on 3 and 13770 DF,  p-value: < 2.2e-16
```

```r
# Checking for multicollinearity
vif_values <- vif(reduced_model)
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "lightblue")
abline(v = 5, lwd = 3, lty = 2)
```




```r
# Comparing model with and without interactions
anova(reduced_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ HA + BDSP + NP
## Model 2: ELEP ~ HA * BDSP * NP
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1  13770 66813068
## 2  13766 66722335  4     90733 4.68 0.0008967 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Fitting regression model with some interactions
selected_model <- lm(ELEP ~ HA + BDSP * NP + HA:BDSP, data = df_final)
summary(selected_model)
```

```
##
## Call:
```

```
## lm(formula = ELEP ~ HA + BDSP * NP + HA:BDSP, data = df_final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -183.97  -43.65  -14.81   24.76  467.25
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.0291     3.6922   9.487  < 2e-16 ***
## HAhouse       20.9916     4.2202   4.974 6.64e-07 ***
## BDSP          12.5276     1.7828   7.027 2.21e-12 ***
## NP            13.7411     1.4029   9.795  < 2e-16 ***
## BDSP:NP       -0.5373     0.4102  -1.310    0.190
## HAhouse:BDSP  -1.0188     1.9377  -0.526    0.599
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.66 on 13768 degrees of freedom
## Multiple R-squared:  0.1239, Adjusted R-squared:  0.1236
## F-statistic: 389.5 on 5 and 13768 DF,  p-value: < 2.2e-16
```

```
# Comparing model without interaction and considered model above.
anova(reduced_model, selected_model)
```

```
## Analysis of Variance Table
##
## Model 1: ELEP ~ HA + BDSP + NP
## Model 2: ELEP ~ HA + BDSP * NP + HA:BDSP
##   Res.Df      RSS Df Sum of Sq     F Pr(>F)
## 1  13770 66813068
## 2  13768 66800123  2     12945 1.334 0.2634
```

```
AIC(reduced_model, selected_model)
```

```
##                df      AIC
## reduced_model   5 155997.1
## selected_model  7 155998.4
```

```
BIC(reduced_model, selected_model)
```

```
##                df      BIC
## reduced_model   5 156034.7
## selected_model  7 156051.1
```

```
# Extracting estimates and CI of model without interactions.
summary(reduced_model)$coefficients
```

```
##             Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 39.99788  1.8686695 21.40447 5.134693e-100
## HAhouse     19.47313  1.7932360 10.85921  2.327360e-27
## BDSP        10.36737  0.7213878 14.37143  1.696103e-46
## NP          12.03562  0.4683361 25.69868 2.664254e-142
```

```r
confint(reduced_model)
```

```
##                 2.5 %   97.5 %
## (Intercept) 36.335033 43.66073
## HAhouse     15.958141 22.98812
## BDSP         8.953356 11.78139
## NP          11.117616 12.95362
```

Utilizing data from the American Community Survey on households, a multiple linear regression approach was employed to assess the average disparity in monthly electricity bills between individuals residing in apartments and houses in Oregon. The established regression model takes the form: Let BLD = 0 for apartments and 1 for houses, $\mu(ELEP|BLD, BDSP, NP) = \beta_0 + \beta_1 BLD + \beta_2 BDSP + \beta_3 NP$.

The findings indicate that, on average, individuals in apartments experience a monthly electricity bill approximately \$19.47 lower than those residing in houses in Oregon, accounting for the number of occupants and bedrooms. With a 95% confidence level, the estimated average difference in monthly electricity bills for apartments and houses, considering a fixed number of occupants and bedrooms, falls within the range of \$15.96 to \$22.99, respectively.

Prediction Problem Strategy

Proposed Meathod

For the prediction problem strategy, the goal is to develop a model capable of accurately forecasting electricity costs for households in Oregon. This endeavor demands meticulous data preparation, which includes cleaning and feature engineering, alongside the evaluation of the model.

In the analysis of the Oregon household dataset from the American Community Survey, multiple regression is employed to forecast electricity costs. Upon scrutinizing the data, it is observed that the ACR and VALR variables contain missing values, prompting their exclusion from the dataset due to their minimal contribution to predicting electricity costs. The variable SERIALNO, serving as a unique identifier, is also omitted from the model as it imparts no supplementary information. Additionally, the exclusion of the TYPE variable from the predictive model is justified by its collinearity with other predictor variables, offering negligible additional insights to the model.

The methodology I want to follow integrates forward validation set approaches, further strengthened by a 10-fold k-means cross-validation to ensure robustness and reliability in model comparisons. Key evaluation metrics, such as root mean squared error(RMSE), Adjusted R-Squared, BIC, and CP values, will be examined across all models.

Following this thorough analysis, the decision on whether to proceed with the exhaustive or forward validation methods will hinge on the performance metrics' comparative outcomes. This structured approach is designed to ensure that the predictive model is not only precise but also generalizable to new data, providing insightful observations into the dynamics affecting electricity costs in Oregon households.