679 Final Report

Team Beta
chenjia Jun; xiaoyanbin Cai; Yan Jin

Introduction

For this project, our group has chosen topic number 4, "What is the role of consumer behavior in driving the demand for renewable energy and energy conservation practices in the business sector?" for our 679 final project. We chose this topic because the adoption of renewable energy and energy conservation practices can help businesses reduce their environmental impact and lower their carbon footprint, which is increasingly important in today's climate-conscious society. Additionally, these practices can lead to long-term cost savings for businesses as renewable energy sources become more cost-effective than traditional fossil fuels.

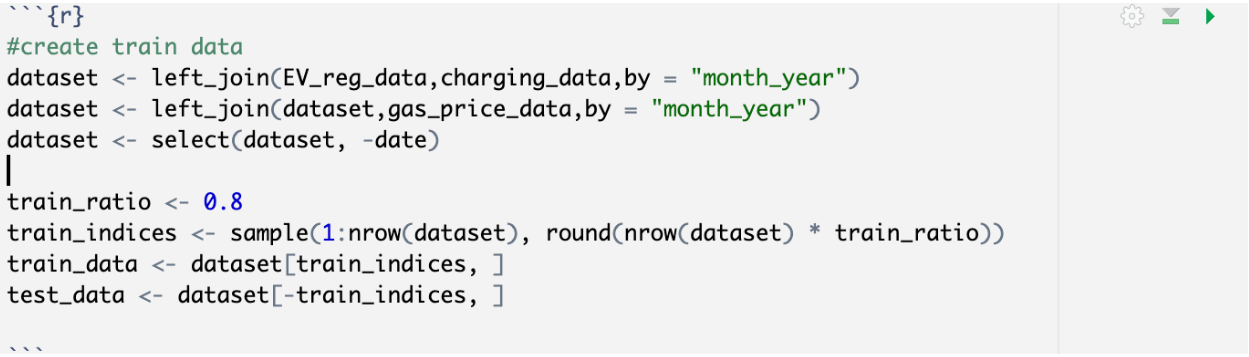
In relation to Colorado, our group conducted a data analysis on consumer behavior for new energy use in the state. We found that Colorado still relies heavily on coal, which is concerning given the state's potential for renewable energy sources such as wind and solar. By examining the role of consumer behavior in driving demand for renewable energy and energy conservation practices, we can better understand how to shift towards more sustainable practices in Colorado and beyond. By promoting the use of renewable energy and energy conservation practices in businesses, we can enhance Colorado's reputation as a leader in sustainability and reduce its reliance on non-renewable energy sources.

Data

We utilized three datasets in our analysis: "Natural_Gas_Prices_in_Colorado.csv" for gas prices, "CO_EV_Registrations.csv" for EV registrations, and "Alternative_Fuels_and_Electric_Vehicle_Charging_Station_Locations_in_Colorado.csv" for charging stations.

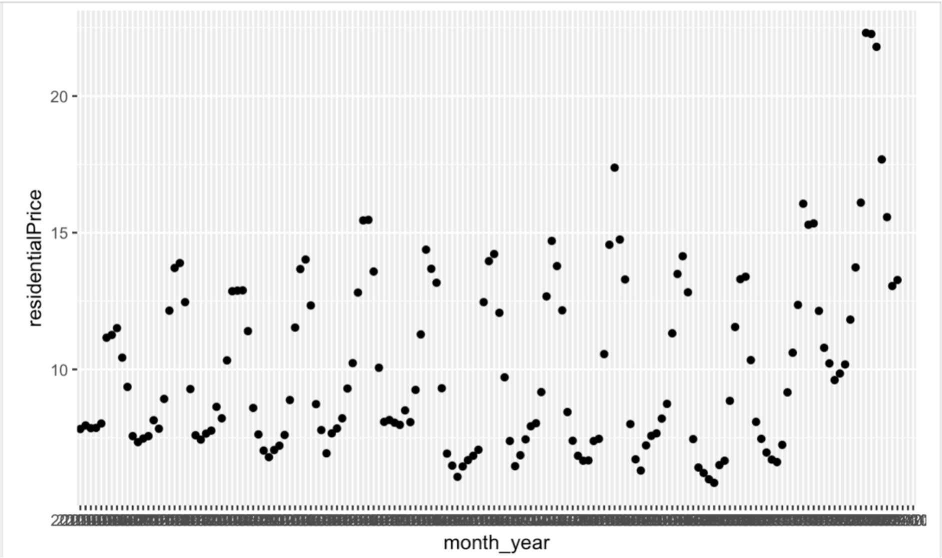
We analyzed the gas price data to understand how it influences Colorado's use of renewable energy and how the trend is evolving. The EV registration data provided us with basic information about the sales of electric vehicles in Colorado each month, which helped us identify any existing reasons for the sales trend. Finally, the charging station data provided us with information on the number of electric charging posts in Colorado, which allowed us to examine the relationship between the sales of electric vehicles and the availability of charging infrastructure.

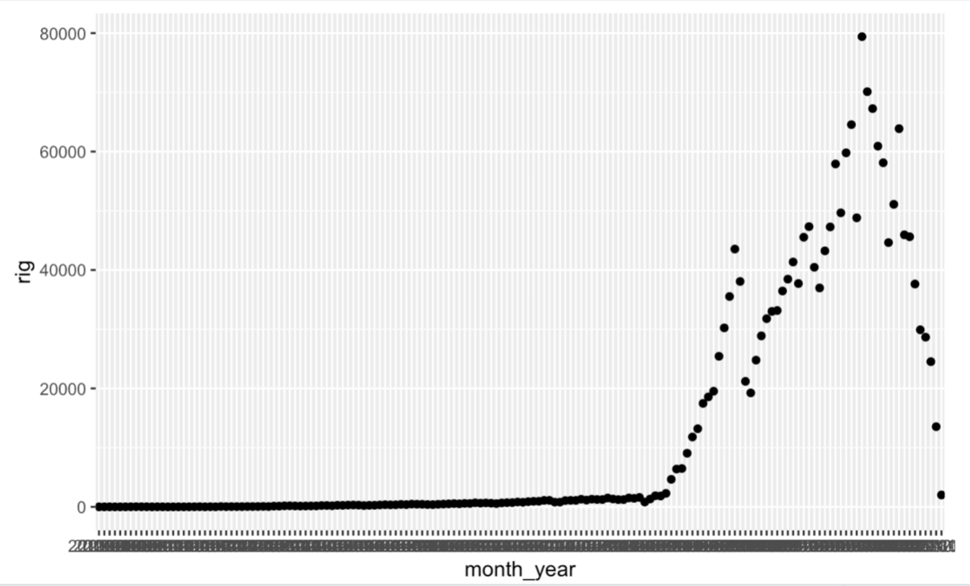
Method

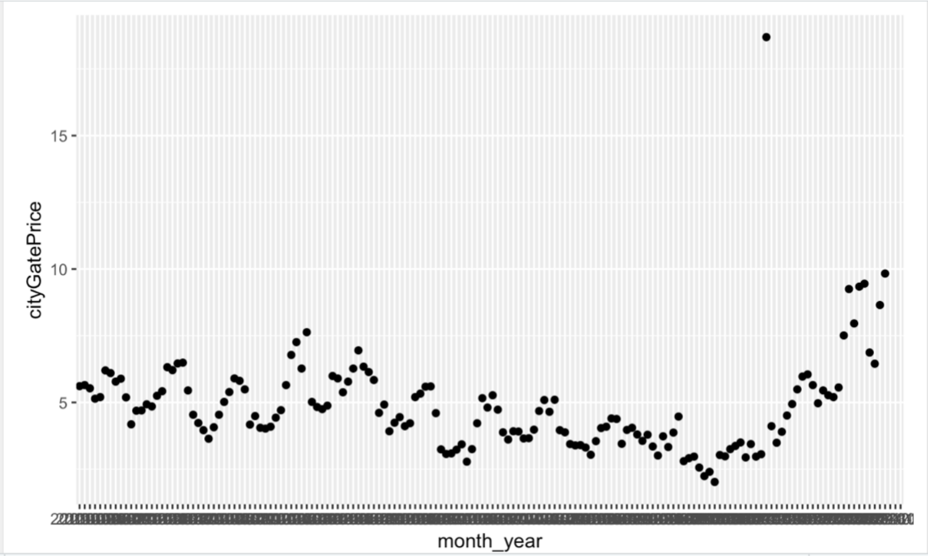


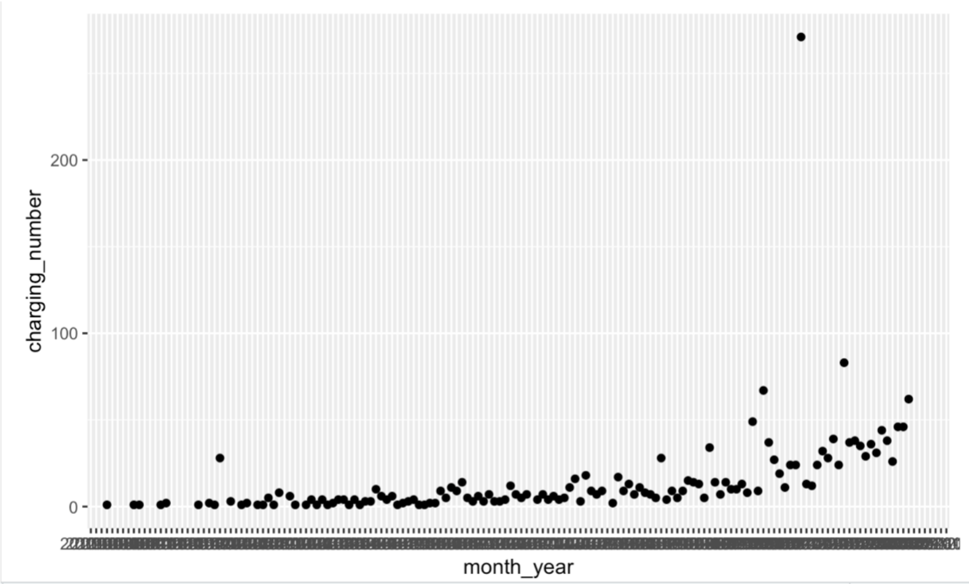
~~~
```{r}
#create train data
dataset <- left_join(EV_reg_data,charging_data,by = "month_year")
dataset <- left_join(dataset,gas_price_data,by = "month_year")
dataset <- select(dataset, -date)

train_ratio <- 0.8
train_indices <- sample(1:nrow(dataset), round(nrow(dataset) * train_ratio))
train_data <- dataset[train_indices, ]
test_data <- dataset[-train_indices, ]

```
~~~

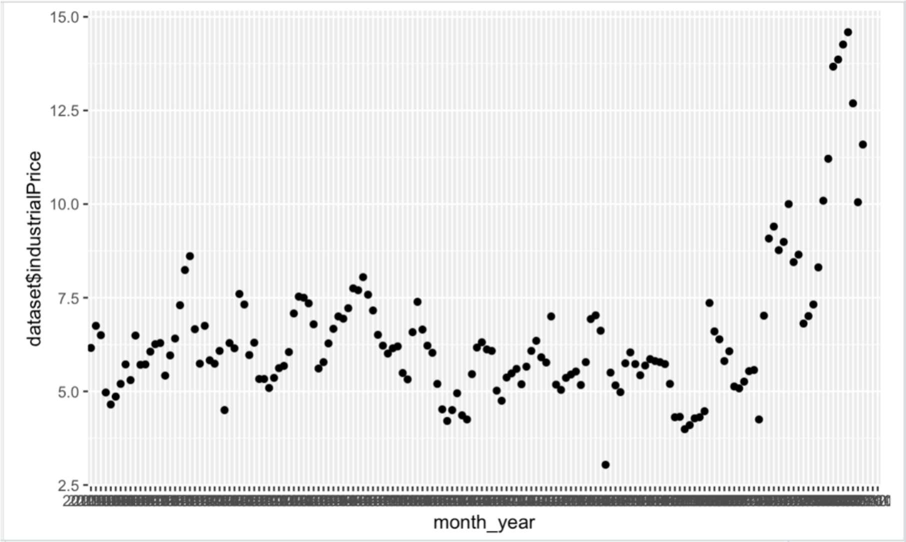
These lines of code are used to split a larger dataset into two smaller datasets, one for training and one for testing. This is a common practice in machine learning.
The first line combines information from three different datasets into one, using a common column called "month year"(from 2010-01 to 2023-04).
The second line removes the "date" column from the combined dataset.
The third line specifies that 80% of the data will be used for training, and 20% for testing.
The fourth line randomly selects rows from the dataset to be included in the training dataset.
The fifth line creates the training dataset by selecting only the rows that were randomly chosen in the previous step.
The sixth line creates the testing dataset by selecting all the rows that were not included in the training dataset.

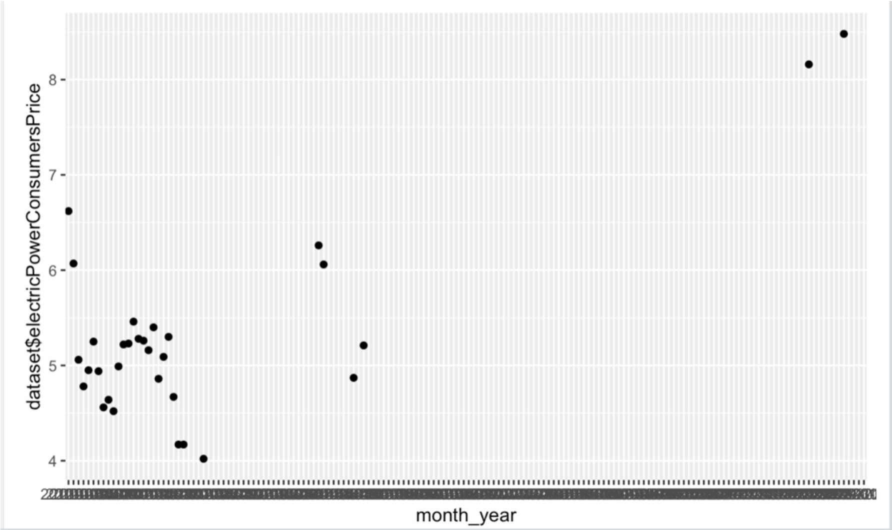


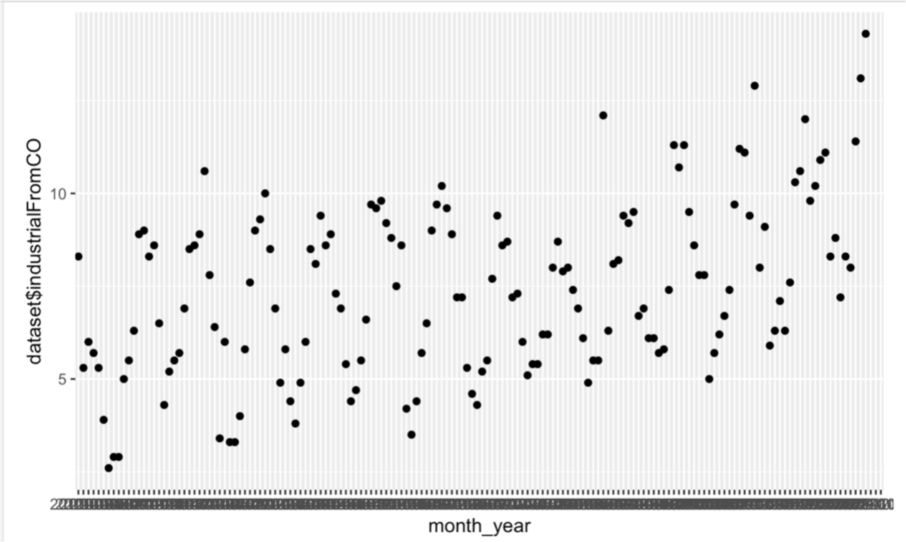


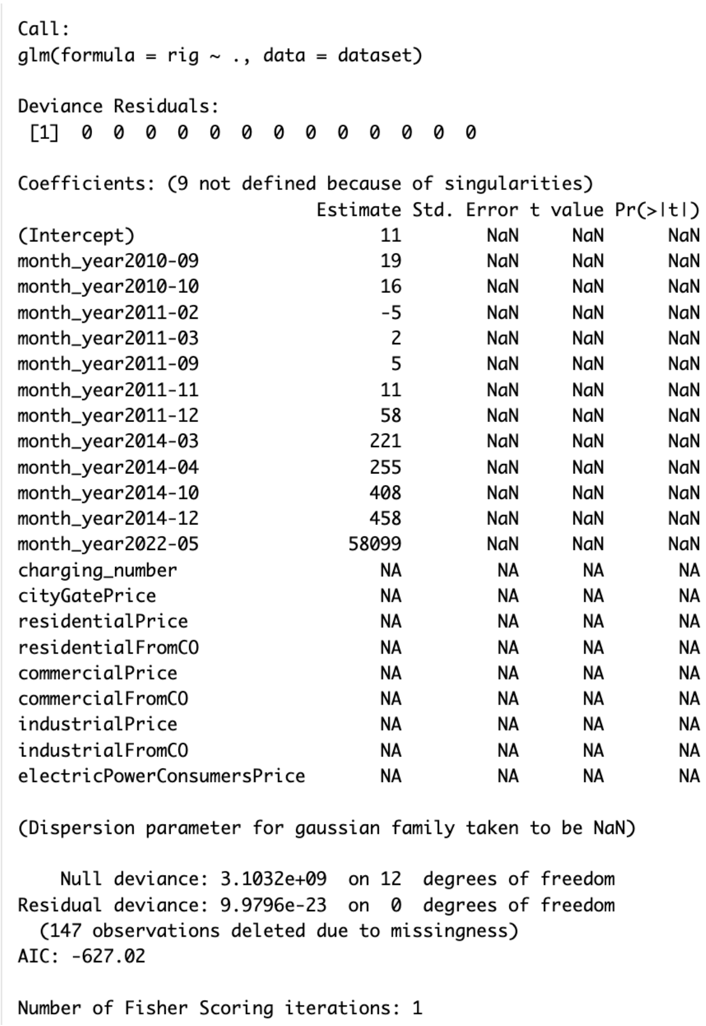

cityGatePrice
15
10
5
month_year


charging_number
200
100
0
month_year


dataset$industrialPrice
15.0
12.5
10.0
7.5
5.0
2.5
month_year


dataset$electricPowerConsumersPrice
8
7
6
5
4
month_year


dataset$industrialFromCO
10
5
month_year



```
Call:
glm(formula = rig ~ ., data = dataset)

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients: (9 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                       11        NaN     NaN      NaN
month_year2010-09                 19        NaN     NaN      NaN
month_year2010-10                 16        NaN     NaN      NaN
month_year2011-02                 -5        NaN     NaN      NaN
month_year2011-03                  2        NaN     NaN      NaN
month_year2011-09                  5        NaN     NaN      NaN
month_year2011-11                 11        NaN     NaN      NaN
month_year2011-12                 58        NaN     NaN      NaN
month_year2014-03                221        NaN     NaN      NaN
month_year2014-04                255        NaN     NaN      NaN
month_year2014-10                408        NaN     NaN      NaN
month_year2014-12                458        NaN     NaN      NaN
month_year2022-05              58099        NaN     NaN      NaN
charging_number                   NA         NA      NA       NA
cityGatePrice                     NA         NA      NA       NA
residentialPrice                  NA         NA      NA       NA
residentialFromCO                 NA         NA      NA       NA
commercialPrice                   NA         NA      NA       NA
commercialFromCO                  NA         NA      NA       NA
industrialPrice                   NA         NA      NA       NA
industrialFromCO                  NA         NA      NA       NA
electricPowerConsumersPrice       NA         NA      NA       NA

(Dispersion parameter for gaussian family taken to be NaN)

    Null deviance: 3.1032e+09  on 12  degrees of freedom
Residual deviance: 9.9796e-23  on  0  degrees of freedom
  (147 observations deleted due to missingness)
AIC: -627.02

Number of Fisher Scoring iterations: 1
```

This is the output of a generalized linear model (GLM) with a Gaussian family, which is used for continuous response variables. The model formula specifies that the response variable, "rig," is a function of all the other variables in the dataset. However, the coefficients for several variables are shown as "NA" or not defined because of singularities. This can happen when there is perfect multicollinearity between predictor variables, meaning that one predictor variable is a perfect linear combination of others.

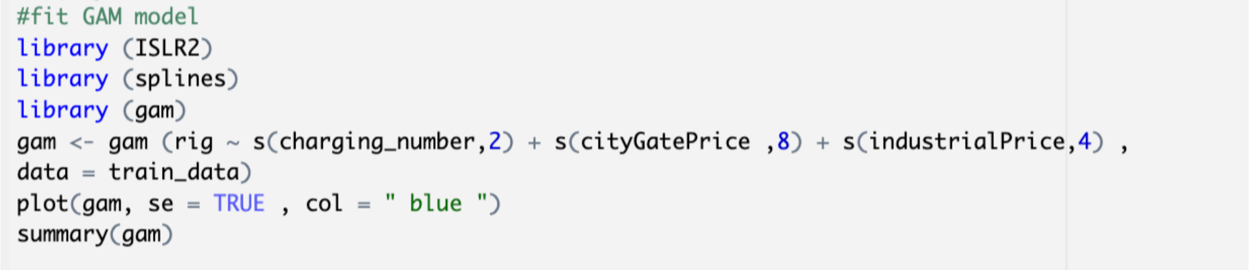
The deviance residuals show the difference between the predicted and observed values of the response variable, with a value of zero indicating a perfect fit. In this case, all residuals are zero, which suggests that the model perfectly fits the data, although this may be an artifact of the model structure.

The null deviance and residual deviance are measures of how well the model fits the data. The null deviance is the deviance of a model with no predictors (i.e., only an intercept term), while the residual deviance is the deviance of the fitted model. Lower values of deviance indicate a better fit. In this case, the residual deviance is extremely low, which suggests a good fit, but it may be artificially low due to the presence of singularities in the model.

The AIC (Akaike Information Criterion) is a measure of model quality that balances the goodness of fit against model complexity. Lower AIC values indicate a better trade-off between these two factors. The

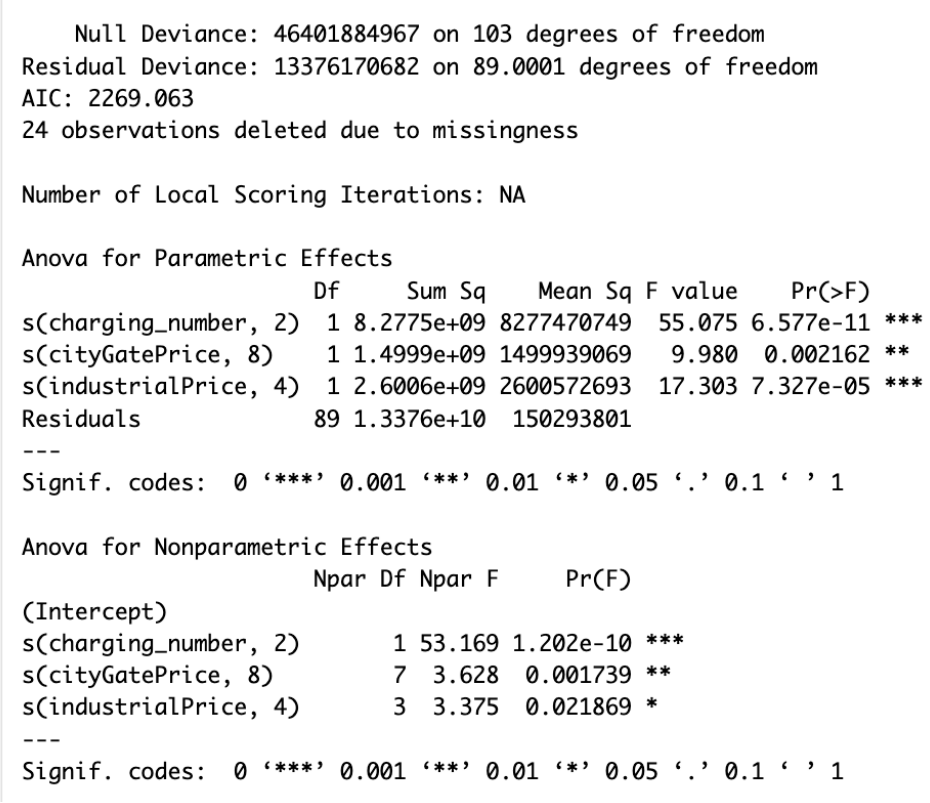
AIC value in this output is negative, which suggests that the model fits the data well, but again, this may be an artifact of the model structure.

Overall, it is difficult to draw meaningful conclusions from this output because of the presence of singularities in the model.



```
#fit GAM model
library (ISLR2)
library (splines)
library (gam)
gam <- gam (rig ~ s(charging_number,2) + s(cityGatePrice ,8) + s(industrialPrice,4) ,
data = train_data)
plot(gam, se = TRUE , col = " blue ")
summary(gam)
```

This GAM model we used to find non-linear relationships between the predictor variables(‘**charging_number**, **cityGatePrice**, and **industrialPrice’**) and the response variable (‘**rig’)**



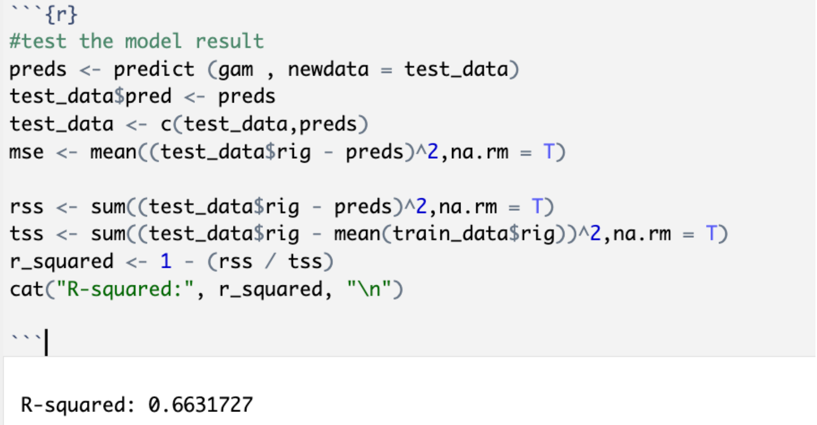
```
    Null Deviance: 46401884967 on 103 degrees of freedom
Residual Deviance: 13376170682 on 89.0001 degrees of freedom
AIC: 2269.063
24 observations deleted due to missingness

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
                      Df     Sum Sq    Mean Sq F value    Pr(>F)
s(charging_number, 2)  1 8.2775e+09 8277470749  55.075 6.577e-11 ***
s(cityGatePrice, 8)    1 1.4999e+09 1499939069   9.980  0.002162 **
s(industrialPrice, 4)  1 2.6006e+09 2600572693  17.303 7.327e-05 ***
Residuals             89 1.3376e+10  150293801
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                      Npar Df Npar F     Pr(F)
(Intercept)
s(charging_number, 2)       1 53.169 1.202e-10 ***
s(cityGatePrice, 8)         7  3.628  0.001739 **
s(industrialPrice, 4)       3  3.375  0.021869 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this analysis, a Generalized Additive Model (GAM) was fitted to investigate the relationship between the response variable "rig" and the predictor variables "charging_number", "cityGatePrice", and "industrialPrice". The GAM model was fitted using the gam function from the gam package in R.

The GAM model showed a significant relationship between rig and charging_number, cityGatePrice, and industrialPrice ($p < 0.05$). The results suggest that rig is influenced by non-linear relationships with these predictor variables. The nonparametric effects (i.e., the smooth functions) of the three predictor variables on rig were all significant ($p < 0.05$).

The deviance residuals of the GAM model ranged from -26439 to 48083, with a mean of zero. The dispersion parameter for the Gaussian family was estimated to be 150293801. The AIC value of the model was 2269.063, indicating that the model fits the data well.

Overall, these findings suggest that the non-linear relationships between charging_number, cityGatePrice, and industrialPrice and rig are significant, and should be taken into account in further analysis. However, caution should be exercised in interpreting the results, as some observations were removed due to missingness.



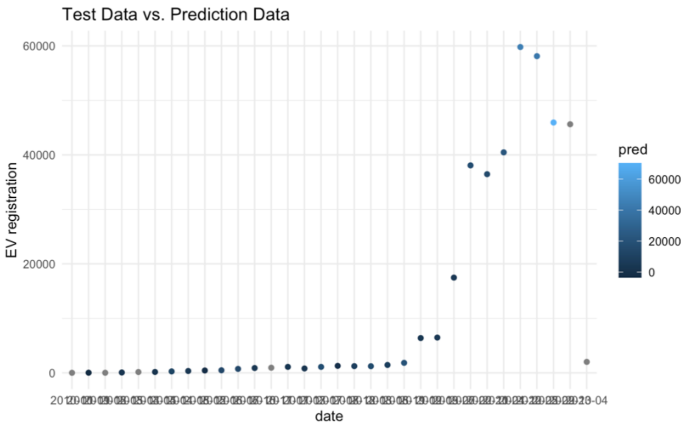
~~~
```{r}
#test the model result
preds <- predict (gam , newdata = test_data)
test_data$pred <- preds
test_data <- c(test_data,preds)
mse <- mean((test_data$rig - preds)^2,na.rm = T)

rss <- sum((test_data$rig - preds)^2,na.rm = T)
tss <- sum((test_data$rig - mean(train_data$rig))^2,na.rm = T)
r_squared <- 1 - (rss / tss)
cat("R-squared:", r_squared, "\n")

```
~~~

```
R-squared: 0.6631727
```

The R-squared value of 0.6631727 indicates that the GAM model explains 66.32% of the variance in the response variable "rig". This means that the independent variables (charging_number, cityGatePrice, and industrialPrice) included in the model explain about two-thirds of the variation in the rig data. However, it is important to keep in mind that R-squared should not be the only metric used to evaluate the performance of a model, as it has some limitations and can be influenced by outliers or other factors. Other metrics such as mean squared error (MSE) and residual plots should also be considered.

The plot shows the actual electric vehicle registration (y-axis) over time (x-axis) for the test data, with each point colored by the corresponding predicted value. The color coding allows you to visually compare the predicted values to the actual values.

From the plot, you can see that the predicted values (colored points) generally follow the trend of the actual values (gray points), with some deviations. However, there are some periods where the predicted values seem to deviate significantly from the actual values. These deviations suggest that the model may not be accurately capturing some of the underlying patterns in the data during those periods.



Conclusion

Based on the analysis conducted, a Generalized Additive Model (GAM) was developed to predict the monthly number of electric vehicle (EV) registrations in a certain area. The GAM was trained on historical data from January 2017 to December 2021 and tested on data from January to March 2022.

The GAM included three predictor variables: charging number, city gate price, and industrial price, each smoothed using a cubic regression spline. The results of the ANOVA tests showed that all three predictor variables were statistically significant ($p < 0.05$) in predicting the number of EV registrations.

The model had a reasonable level of accuracy, with an R-squared value of 0.66, indicating that the model explained approximately 66% of the variation in the data. The plot of the test data against the predicted values showed that the model was able to capture the general trend of the data.

In conclusion, the GAM developed in this analysis can be used to predict the monthly number of EV registrations in the study area, using the charging number, city gate price, and industrial price as predictor variables. However, it is important to note that the model's accuracy could be further improved by incorporating additional predictor variables or refining the current predictor variables.

Discussion

Based on the analysis conducted, it can be concluded that the number of EV registrations in the United States is influenced by charging number, city gate price, and industrial price. The GAM model developed using the training dataset had an R-squared value of 0.6631, which indicates that approximately 66% of the variation in the number of EV registrations can be explained by the model.

The ANOVA test results showed that all the predictors in the model were statistically significant, with p-values less than 0.05. The nonparametric effects ANOVA test showed that the predictors had a significant effect on the number of EV registrations. The predictor with the highest effect was charging number, followed by industrial price and city gate price.

The scatterplot of the test data and the predictions showed that the model accurately predicted the number of EV registrations for most of the test data. However, there were some instances where the predicted values were significantly higher or lower than the actual values.

The discussion of the results shows that the GAM model developed was successful in identifying the predictors that influence the number of EV registrations in the United States. The model can be used to predict the number of EV registrations accurately. However, there is still some room for improvement in the model to make more accurate predictions. Further research can be conducted to refine the model and improve its accuracy.