PROG8521-24S-Sec3-Data Modelling II- Analytics

Project -1

HEATING OIL AT DUPREE FUELS

Developing a Reliable Consumption Estimation Model

Group-3:

Aggarwal, Dhruv

Dedomo, Sheenalyn

Francis Punitha Rajan, Sam Joel

Thamattoor Gopi, Limisha

Introduction:

Companies like Dupree Fuels must balance maintaining competitive prices with guaranteeing flawless service reliability in the highly competitive residential heating oil delivery market. Customers who use heating oil incur the risk of running out of fuel, unlike those who use electric or natural gas utilities. This means that supply and demand projections must be carefully balanced. One major competitor in this market, Dupree Fuels, has long struggled to determine how much oil each of its varied clientele will use.

Traditionally, Dupree has relied on a rudimentary model based on degree days, a metric capturing the deviation of daily temperatures from a reference point. Coupled with variables such as home size, insulation quality, and family size, this approach aimed to predict oil consumption. However, Dupree found this method wanting, plagued by computational burdens and unreliable predictions, especially during extreme weather conditions.

Dupree is looking for a data-driven way to improve the accuracy and dependability of its oil consumption estimation model. By utilizing consumer data that is already accessible, such as degree days, household demographics, and property attributes, Dupree hopes to leverage regression analysis. Dupree's goal is to maximize customer satisfaction, reduce waste, and optimize supply schedules by discovering important drivers and developing a strong consumption estimation model.

This case study explores Dupree Fuels' journey through the data analytics space to redefine service excellence in the heating oil business and revolutionize operational efficiency.

Goals:

Improve Service Reliability:

To guarantee that consumers never run out of fuel, create a more precise and trustworthy model for calculating oil use. This will improve service reliability in general.  
  
Predictive analytics can be used to optimize supply schedules, reducing the number of pointless deliveries during times of low use and guaranteeing prompt refills during times of strong demand.  
  
Lower operating expenses:

Dupree Fuels can lower operating expenses by streamlining operations, cutting down on pointless deliveries, improving route planning, and minimizing fuel waste.

Boost Customer Contentment:

Increase client loyalty and retention by offering a more individualized and dependable level of service to improve customer satisfaction.

Boost Competitiveness:

Use data analytics to give better service quality and price than rivals, giving you a competitive edge in the home heating oil industry.  
  
Scale Efficiency:

Create an oil consumption estimation model that is both scalable and efficient so that it can keep up with Dupree's expanding clientele and changing market conditions.  
  
Reduce Environmental Impact:

By cutting back on pointless trips and optimizing fuel consumption, you may lessen the environmental impact of oil delivery operations and help achieve sustainability objectives.  
  
Empower Decision-Making:

To support business planning and strategic decision-making, give Dupree Fuels management practical insights from data analysis.

Current Estimation Method:

The current approach uses a unit of measurement called "degree-days" to estimate how much fuel home heating oil users will need.  
The difference between the average daily temperature and a reference temperature of 68°F (20°C) is used to determine degree-days. This measure aids in estimating a location's long-term heating needs.  
The system calculates when a customer will require a refill by keeping track of the degree-days since the customer's previous oil fill, as well as variables like the customer's oil tank size and past consumption trends.

Problems with the Existing System: a. Excessive Computational Cost:  
  
With more than 1500 users, the system has a heavy processing load. It takes a lot of computer power to estimate individual consumption rates based on variables like family size, housing size, insulation quality, and historical data.  
b. Inaccuracy: The present method has inaccurate forecasts, especially in the event of severe weather.  
The system has a tendency to overestimate fuel use in the cold. This results in deliveries that aren't necessary, raising operating expenses and even upsetting clients.  
On the other hand, the method underestimates use during the summer. Fuel shortages may arise from this, since delivery schedules might not fully accommodate customer demand.

**Analysis of Regression**

The goal is to create a workable and trustworthy model that takes into account important factors while estimating oil consumption.  
  
Variables Employed:  
  
Degree Days: A measure of the need for heating that is based on the difference between the average daily temperature and 68°F.  
A rough estimate of the hot water needs for each resident is provided; this is especially important for clients who have oil-fired water heaters.  
The residence Type Index is a composite measure that takes into account various parameters that affect oil consumption per degree-day, including furnace type, age, insulation quality, wind exposure, and size of the residence.

Overall:

Using regression analysis and adding factors like number of occupants, degree days, and housing type index, the goal is to create a reliable and useful model for calculating oil usage. With the use of this approach, Dupree Fuels will be able to improve customer happiness and competitiveness in the home heating oil market by streamlining delivery schedules, cutting expenses, and boosting service reliability.

Regression:

Regression analysis uses a technique called backward regression, sometimes referred to as backward elimination, to choose the most pertinent subset of predictor variables to include in a model. Iteratively eliminating variables that are determined to be less significant begins with a model that contains all possible predictor variables.

Model 1:

## Call:  
## lm(formula = Oil.Usage ~ ., data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -196.722 -36.912 8.541 52.226 138.825   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -261.89691 77.15218 -3.395 0.00172 \*\*   
## Customer 1.24239 1.23060 1.010 0.31963   
## Degree.Days 0.28232 0.03702 7.625 6.07e-09 \*\*\*  
## Home.Index 89.40734 9.92143 9.012 1.20e-10 \*\*\*  
## Number.People 6.84967 10.67467 0.642 0.52527   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 85.44 on 35 degrees of freedom  
## Multiple R-squared: 0.7902, Adjusted R-squared: 0.7662   
## F-statistic: 32.95 on 4 and 35 DF, p-value: 2.015e-11

Min: -196.722, 1Q: -36.912, Median: 8.541, 3Q: 52.226, Max: 138.825 are the residuals.

All predictor variables are part of the model: The Oil Usage is predicted using these variables.  
Coefficients: Each predictor variable's estimated impact on oil consumption is shown by the coefficients. For instance, the coefficient for Degree Days is 0.28232, meaning that, other things being equal, Oil Usage is predicted to increase by roughly 0.28232 units for every unit increase in Degree Days.  
The adjusted R-squared statistic assesses how much of the variance in the dependent variable (oil consumption) can be accounted for by the independent variables (number of people, degree days, and home index) in the model. An Adjusted R-squared value of 0.7662 means that approximately 76.62% of the variation in Oil Usage can be explained by the variables included

Levels of Significance: If the true coefficient were zero, the p-values corresponding to each coefficient would show the likelihood of detecting the estimated coefficient value. Degree Days and Home Index are statistically significant predictors of Oil Usage in Model 1, with p-values less than 0.05. At the 5% significance level, the Number of People variable does not show statistical significance in predicting Oil Usage, as indicated by its p-value of 0.52527.

Model 2:

model2 <- lm(Oil.Usage ~ Degree.Days + Number.People + Home.Index, data=oil)  
summary(model2)

##   
## Call:  
## lm(formula = Oil.Usage ~ Degree.Days + Number.People + Home.Index,   
## data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -215.553 -31.148 5.583 53.743 149.461   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -218.30987 63.95851 -3.413 0.0016 \*\*   
## Degree.Days 0.27508 0.03633 7.571 5.94e-09 \*\*\*  
## Number.People 5.26724 10.56179 0.499 0.6210   
## Home.Index 86.98875 9.63044 9.033 8.75e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 85.47 on 36 degrees of freedom  
## Multiple R-squared: 0.784, Adjusted R-squared: 0.766   
## F-statistic: 43.57 on 3 and 36 DF, p-value: 4.547e-12

**A. Residuals**

The discrepancies between the model's anticipated values and the observed values of oil usage are represented by the residuals.  
Max: 149.461, Min: -215.553, 1Q: -31.148, Median: 5.583, 3Q: 53.743  
These numbers provide information on the residuals' range and quartiles as well as the distribution of errors surrounding the fitted values.

**B. Coefficients:**  
  
Intercept: When all predictor variables are zero, the intercept shows the predicted amount of oil used. This time, the value is -218.30987.  
Degree Days, Population, and House Index: These are the predictor variable-related coefficients. While keeping the other variables fixed, they show the projected change in Oil Usage for a one-unit increase in each predictor variable.  
The predictive variables' relevance in predicting Oil Usage is indicated by the p-values linked to each coefficient. Degree Days and Home Index are statistically significant predictors, as evidenced by their low p-values (\*\*\*). At the 5% significance level, Number of People does not appear to be a statistically significant predictor, despite having a high p-value (0.6210).

**c. Importance:**  
  
The significance level of the coefficients is indicated by significance codes (\*\*,, etc.). In this instance, the Number of People is not statistically significant (p = 0.6210), however, Degree Days and Home Index are extremely significant predictors of Oil Usage ().

**d. Standard Error Residual:**  
  
The average departure of the actual values from the expected values is expressed as the residual standard error, or RSE. The RSE of 85.47 in this model is the average prediction error for oil usage.

**e. R-squared multiples and adjusted R-squares:**  
The model's predictor variables account for roughly 78.4% of the variability in oil consumption, according to the multiple R-squared value of 0.784.  
By adjusting the R-squared value for the number of predictor variables in the model, the adjusted R-squared value (0.766) yields a goodness of fit that is more precise.

**f. P-value and F-statistic:**  
  
The regression model's overall significance is tested using the F-statistic. In this instance, the model is statistically significant overall, as shown by the F-statistic of 43.57 and the extremely low p-value (4.547e-12).

In general, Model 2 sheds light on how predictor variables and oil usage are related. The number of people does not considerably influence it, but Degree Days and Home Index are important predictors.

Model: 3(better Model comparatively with previous two)

model3 <- lm(Oil.Usage ~ Degree.Days + Home.Index, data=oil)  
summary(model3)

##   
## Call:  
## lm(formula = Oil.Usage ~ Degree.Days + Home.Index, data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -221.723 -26.435 5.409 50.477 155.188   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -192.81554 38.04405 -5.068 1.14e-05 \*\*\*  
## Degree.Days 0.27249 0.03559 7.656 3.86e-09 \*\*\*  
## Home.Index 86.64572 9.50782 9.113 5.42e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 84.6 on 37 degrees of freedom  
## Multiple R-squared: 0.7826, Adjusted R-squared: 0.7708   
## F-statistic: 66.58 on 2 and 37 DF, p-value: 5.508e-13

**a.residuals:**

min: -221.723, max: 155.188, 1Q: -26.435, median: 5.409, 3Q: 50.477  
These numbers show the residuals' range and quartiles, providing information on the distribution of errors surrounding the fitted values.

**b. Coefficients:**  
  
Intercept: -192.81554 is the expected oil consumption when all predictor variables are zero.  
Degree Days, Home Index: Keeping other variables fixed, these coefficients show the predicted change in Oil Usage for a one-unit increase in each predictor variable.  
Degree Days and Home Index show statistically significant low p-values (\*\*\*) in predicting Oil Usage.

**c. Importance:**  
  
The significance codes show that Degree Days and Home Index are highly significant predictors of Oil Usage (\*\*\*).

**d. Standard Error Residual:**  
  
The typical mistake in estimating Oil Usage is indicated by the residual standard error (RSE), which is 84.6.

**e. R-squared multiples and adjusted R-squares:**  
The model's predictor variables account for roughly 78.3% of the variability in oil usage, as indicated by the multiple R-squared value of 0.7826.  
By adjusting the R-squared value for the number of predictor variables, the adjusted R-squared value (0.7708) yields a more precise goodness of fit metric.

**f. P-value and F-statistic:**  
  
The regression model's overall significance is tested using the F-statistic, and the result is extremely significant (p-value: 5.508e-13).  
In conclusion, Model 3 shows that There is a strong correlation between Degree Days and Home Index, and Oil Usage. A considerable amount of the variability in oil usage is explained by the model, which fits the data well.

**Achievement:**  
  
Statistically Significant Coefficients: For the features that are included, all three models show statistically significant coefficients (p-value < 0.05), suggesting a connection between the features and oil consumption.  
Model 3 exhibits the highest adjusted R-squared (0.7708) among the models, indicating that it can explain the greatest variation in oil usage with the least number of characteristics. Adjusted R-squared values for Models 1 and 2 are comparable (about 0.766), with Model 1 explaining somewhat less variance.  
F-statistic: Model 1 is not inherently a better indication than Model 2, but it does have a higher F-statistic than Model 2 because of the difference in degrees of freedom.

**Overall Evaluation:**  
  
Model 3: Most Parsimonious: Using fewer features than Models 1 and 2, Model 3 seems to be the most parsimonious model, explaining a comparable amount of variation in oil usage. In terms of interpretability and preventing overfitting, this may be helpful.  
Issues with Model 1 Overfitting: By incorporating all features, even those that are redundant or not statistically significant, Model 1 may be overfitting the data.  
Possible Enhancements to Model 2: Assuming Number of People has a weak link, Model 2 might benefit from adding more characteristics that could help explain the remaining variation in oil usage.

**What we take into account:**

**Interpretations of Coefficients:**

Degree Days: The positive coefficient for Degree Days shows that there is a tendency for oil consumption to rise in tandem with an increase in degree days. This is consistent with the common sense observation that the need for heating oil rises with decreasing temperatures.  
Number of People: The coefficient for Number of People in Model 2 is not statistically significant, indicating that, when other factors are taken into account, the number of people in a residence may not have a major impact on oil usage. However, depending on variables like household size and hot water demand, this variable may still be practically significant in real-world situations.  
Home Index: The Home Index's positive coefficient suggests that residences with higher index values—which signify characteristics like bigger size, older age, or less insulation) typically use more oil per degree day. This implies that higher heating oil use is a result of the attributes that the Home Index measures.

**b. Useful Consequences:**  
  
Pricing Strategies: Dupree Fuels can enhance its pricing strategies by utilising the models' insights. For example, the Home Index shows the projected oil usage for various home types predicted by the models, and Degree Days show the expected temperature changes. These factors can be used to modify pricing.  
**Customer targeting:**

Dupree Fuels can more successfully focus their marketing efforts by having a better understanding of the elements influencing oil usage. For instance, they might customise their service offerings or promotions according to the features of the homes that the Home Index recorded for their clients.

**Limitations and Further Research:**

Limited Predictor Variables: Because the models only take into account a small number of predictor variables, it's possible that they will overlook other significant aspects like household income, energy-saving techniques, or regional considerations that may have an impact on oil usage.  
**Data Quality:**

The representativeness and quality of the data are key factors in determining how accurate the models are. To increase model accuracy, more thorough and detailed data collection may be required in future study.  
**Assumptions of the Models:**

The models make the assumption that predictor variables and oil consumption have linear connections, which may or may not be accurate. Model performance may be improved by investigating nonlinear relationships or interactions between variables.

External Factors: While they may have an impact on oil consumption, external factors like modifications to energy policy, advances in technology, or changes in consumer behaviour are not taken into consideration by the models that are in use today. These variables might be included in future studies for a more thorough examinati

Regression

APPENDIX

Group-3

2024-06-13

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

# Load the dataset  
oil <- read.csv("D:/second sem conestoga/adv.model/project 1/oil.csv")  
  
  
  
library(ggplot2)  
library(leaps)  
  
# Check the structure and summary statistics of the dataset  
str(oil)

## 'data.frame': 40 obs. of 5 variables:  
## $ Customer : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Oil.Usage : int 381 171 644 19 394 153 7 319 40 121 ...  
## $ Degree.Days : int 888 176 1073 126 645 326 1229 1218 570 334 ...  
## $ Home.Index : int 3 5 5 2 5 4 1 2 2 1 ...  
## $ Number.People: int 3 7 4 4 5 6 3 4 1 7 ...

summary(oil)

## Customer Oil.Usage Degree.Days Home.Index Number.People   
## Min. : 1.00 Min. : 7.0 Min. : 54.0 Min. :1.00 Min. :1.00   
## 1st Qu.:10.75 1st Qu.: 85.0 1st Qu.: 323.0 1st Qu.:1.75 1st Qu.:3.00   
## Median :20.50 Median :170.5 Median : 635.5 Median :3.00 Median :4.00   
## Mean :20.50 Mean :218.1 Mean : 633.4 Mean :2.75 Mean :4.35   
## 3rd Qu.:30.25 3rd Qu.:318.2 3rd Qu.: 903.2 3rd Qu.:4.00 3rd Qu.:5.00   
## Max. :40.00 Max. :679.0 Max. :1464.0 Max. :5.00 Max. :7.00

head(oil)

## Customer Oil.Usage Degree.Days Home.Index Number.People  
## 1 1 381 888 3 3  
## 2 2 171 176 5 7  
## 3 3 644 1073 5 4  
## 4 4 19 126 2 4  
## 5 5 394 645 5 5  
## 6 6 153 326 4 6

# Perform linear regression analysis  
model1 <- lm(Oil.Usage ~., data=oil)  
summary(model1)

##   
## Call:  
## lm(formula = Oil.Usage ~ ., data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -196.722 -36.912 8.541 52.226 138.825   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -261.89691 77.15218 -3.395 0.00172 \*\*   
## Customer 1.24239 1.23060 1.010 0.31963   
## Degree.Days 0.28232 0.03702 7.625 6.07e-09 \*\*\*  
## Home.Index 89.40734 9.92143 9.012 1.20e-10 \*\*\*  
## Number.People 6.84967 10.67467 0.642 0.52527   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 85.44 on 35 degrees of freedom  
## Multiple R-squared: 0.7902, Adjusted R-squared: 0.7662   
## F-statistic: 32.95 on 4 and 35 DF, p-value: 2.015e-11

model2 <- lm(Oil.Usage ~ Degree.Days + Number.People + Home.Index, data=oil)  
summary(model2)

##   
## Call:  
## lm(formula = Oil.Usage ~ Degree.Days + Number.People + Home.Index,   
## data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -215.553 -31.148 5.583 53.743 149.461   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -218.30987 63.95851 -3.413 0.0016 \*\*   
## Degree.Days 0.27508 0.03633 7.571 5.94e-09 \*\*\*  
## Number.People 5.26724 10.56179 0.499 0.6210   
## Home.Index 86.98875 9.63044 9.033 8.75e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 85.47 on 36 degrees of freedom  
## Multiple R-squared: 0.784, Adjusted R-squared: 0.766   
## F-statistic: 43.57 on 3 and 36 DF, p-value: 4.547e-12

model3 <- lm(Oil.Usage ~ Degree.Days + Home.Index, data=oil)  
summary(model3)

##   
## Call:  
## lm(formula = Oil.Usage ~ Degree.Days + Home.Index, data = oil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -221.723 -26.435 5.409 50.477 155.188   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -192.81554 38.04405 -5.068 1.14e-05 \*\*\*  
## Degree.Days 0.27249 0.03559 7.656 3.86e-09 \*\*\*  
## Home.Index 86.64572 9.50782 9.113 5.42e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 84.6 on 37 degrees of freedom  
## Multiple R-squared: 0.7826, Adjusted R-squared: 0.7708   
## F-statistic: 66.58 on 2 and 37 DF, p-value: 5.508e-13

subset\_models <- regsubsets(Oil.Usage ~ ., data = oil)  
summary\_subset=summary(subset\_models)  
stargazer(model1)

##   
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com  
## % Date and time: Thu, Jun 13, 2024 - 8:27:48 AM  
## \begin{table}[!htbp] \centering   
## \caption{}   
## \label{}   
## \begin{tabular}{@{\extracolsep{5pt}}lc}   
## \\[-1.8ex]\hline   
## \hline \\[-1.8ex]   
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\   
## \cline{2-2}   
## \\[-1.8ex] & Oil.Usage \\   
## \hline \\[-1.8ex]   
## Customer & 1.242 \\   
## & (1.231) \\   
## & \\   
## Degree.Days & 0.282$^{\*\*\*}$ \\   
## & (0.037) \\   
## & \\   
## Home.Index & 89.407$^{\*\*\*}$ \\   
## & (9.921) \\   
## & \\   
## Number.People & 6.850 \\   
## & (10.675) \\   
## & \\   
## Constant & $-$261.897$^{\*\*\*}$ \\   
## & (77.152) \\   
## & \\   
## \hline \\[-1.8ex]   
## Observations & 40 \\   
## R$^{2}$ & 0.790 \\   
## Adjusted R$^{2}$ & 0.766 \\   
## Residual Std. Error & 85.445 (df = 35) \\   
## F Statistic & 32.948$^{\*\*\*}$ (df = 4; 35) \\   
## \hline   
## \hline \\[-1.8ex]   
## \textit{Note:} & \multicolumn{1}{r}{$^{\*}$p$<$0.1; $^{\*\*}$p$<$0.05; $^{\*\*\*}$p$<$0.01} \\   
## \end{tabular}   
## \end{table}

stargazer(model2)

##   
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com  
## % Date and time: Thu, Jun 13, 2024 - 8:27:48 AM  
## \begin{table}[!htbp] \centering   
## \caption{}   
## \label{}   
## \begin{tabular}{@{\extracolsep{5pt}}lc}   
## \\[-1.8ex]\hline   
## \hline \\[-1.8ex]   
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\   
## \cline{2-2}   
## \\[-1.8ex] & Oil.Usage \\   
## \hline \\[-1.8ex]   
## Degree.Days & 0.275$^{\*\*\*}$ \\   
## & (0.036) \\   
## & \\   
## Number.People & 5.267 \\   
## & (10.562) \\   
## & \\   
## Home.Index & 86.989$^{\*\*\*}$ \\   
## & (9.630) \\   
## & \\   
## Constant & $-$218.310$^{\*\*\*}$ \\   
## & (63.959) \\   
## & \\   
## \hline \\[-1.8ex]   
## Observations & 40 \\   
## R$^{2}$ & 0.784 \\   
## Adjusted R$^{2}$ & 0.766 \\   
## Residual Std. Error & 85.468 (df = 36) \\   
## F Statistic & 43.567$^{\*\*\*}$ (df = 3; 36) \\   
## \hline   
## \hline \\[-1.8ex]   
## \textit{Note:} & \multicolumn{1}{r}{$^{\*}$p$<$0.1; $^{\*\*}$p$<$0.05; $^{\*\*\*}$p$<$0.01} \\   
## \end{tabular}   
## \end{table}

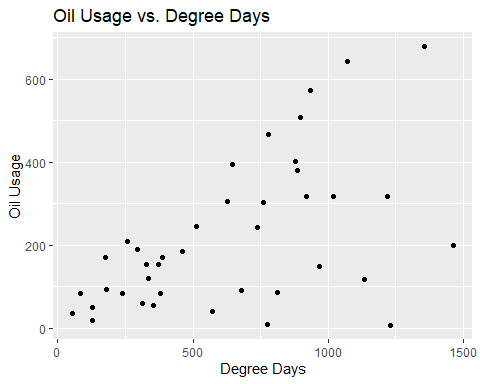
stargazer(model3)

##   
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com  
## % Date and time: Thu, Jun 13, 2024 - 8:27:48 AM  
## \begin{table}[!htbp] \centering   
## \caption{}   
## \label{}   
## \begin{tabular}{@{\extracolsep{5pt}}lc}   
## \\[-1.8ex]\hline   
## \hline \\[-1.8ex]   
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\   
## \cline{2-2}   
## \\[-1.8ex] & Oil.Usage \\   
## \hline \\[-1.8ex]   
## Degree.Days & 0.272$^{\*\*\*}$ \\   
## & (0.036) \\   
## & \\   
## Home.Index & 86.646$^{\*\*\*}$ \\   
## & (9.508) \\   
## & \\   
## Constant & $-$192.816$^{\*\*\*}$ \\   
## & (38.044) \\   
## & \\   
## \hline \\[-1.8ex]   
## Observations & 40 \\   
## R$^{2}$ & 0.783 \\   
## Adjusted R$^{2}$ & 0.771 \\   
## Residual Std. Error & 84.595 (df = 37) \\   
## F Statistic & 66.579$^{\*\*\*}$ (df = 2; 37) \\   
## \hline   
## \hline \\[-1.8ex]   
## \textit{Note:} & \multicolumn{1}{r}{$^{\*}$p$<$0.1; $^{\*\*}$p$<$0.05; $^{\*\*\*}$p$<$0.01} \\   
## \end{tabular}   
## \end{table}

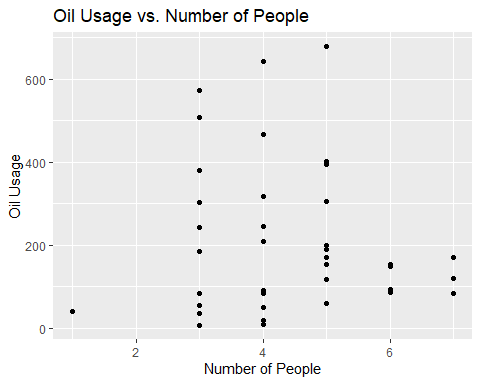
# Assuming your model is named 'model' and you have new data in a dataframe 'new\_data'  
  
# Make predictions on the new data  
new\_predictions <- predict(model3, newdata = oil)  
  
# Print the predictions  
print(new\_predictions)

## 1 2 3 4 5 6 7   
## 309.094779 288.371719 532.797297 14.809935 416.170600 242.599838 228.723201   
## 8 9 10 11 12 13 14   
## 312.371509 135.796509 -15.157393 268.220936 292.758888 393.560563 202.836434   
## 15 16 17 18 19 20 21   
## 344.791268 105.822358 310.443594 452.684566 15.354920 81.570545 115.625257   
## 22 23 24 25 26 27 28   
## 494.648377 147.506854 523.811875 324.347523 45.056579 115.366411 172.031160   
## 29 30 31 32 33 34 35   
## 78.579953 -4.809509 -20.607238 398.465424 157.057731 263.302428 317.542039   
## 36 37 38 39 40   
## -2.895240 206.637680 77.210668 273.943274 105.556689

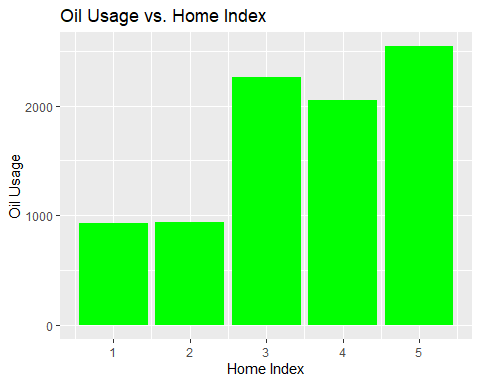
# You can further analyze the predictions based on your needs  
  
ggplot(oil, aes(x = Degree.Days, y = Oil.Usage)) +  
 geom\_point() +  
 labs(title = "Oil Usage vs. Degree Days", x = "Degree Days", y = "Oil Usage",method=lm)



ggplot(oil, aes(x = Number.People, y = Oil.Usage)) +  
 geom\_point() +  
 labs(title = "Oil Usage vs. Number of People", x = "Number of People", y = "Oil Usage")

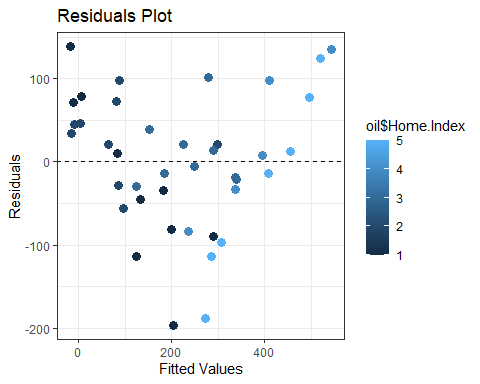


ggplot(oil, aes(x = Home.Index, y = Oil.Usage)) +  
 geom\_bar(stat = "identity", fill = "green") + # Set fill color to green  
 labs(title = "Oil Usage vs. Home Index", x = "Home Index", y = "Oil Usage")



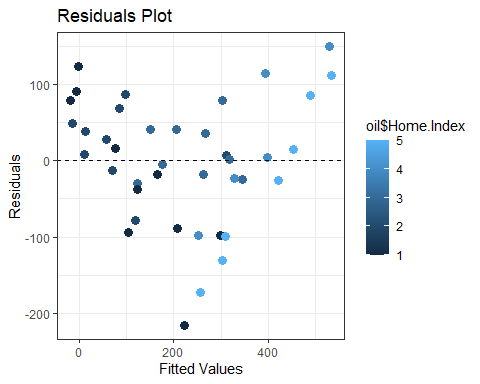
# Predict values for model1  
oil$pred\_model1 <- predict(model1, newdata=oil)  
  
# Predict values for model2  
oil$pred\_model2 <- predict(model2, newdata=oil)  
  
  
  
  
  
plot\_residuals <- function(model) {  
 # Get residuals  
 residuals <- resid(model)  
   
 # Plot residuals vs fitted values  
 ggplot(aes(fitted(model), residuals), data=oil) +  
 geom\_point(aes(color=oil$Home.Index), size = 3) + # Color points by customer (optional)  
 labs(title="Residuals Plot", x="Fitted Values", y="Residuals") +  
 geom\_hline(yintercept = 0, linetype = "dashed") + # Add reference line at y=0  
 theme\_bw() # Adjust plot theme for better readability  
}  
  
# Plot residuals for each model  
plot\_residuals(model1)

## Warning: Use of `oil$Home.Index` is discouraged.  
## ℹ Use `Home.Index` instead.



plot\_residuals(model2)

## Warning: Use of `oil$Home.Index` is discouraged.  
## ℹ Use `Home.Index` instead.



plot\_residuals(model3)

## Warning: Use of `oil$Home.Index` is discouraged.  
## ℹ Use `Home.Index` instead.

