Graduate Energy Informatics Project Group 2

Charlie McCollough

2024-04-15

Vehicle Emissions G/H Project

Energy Informatics

Group 2

```
suppressMessages(suppressWarnings({
  library(ggplot2)
  library(dplyr)
  library(rpart)
  library(rpart.plot)
  library(readxl)
  library(corrplot)
  library(fastDummies)
}))
```

Install Packages

```
# Load the data
fuel_data <- read_excel("FuelConsumption - IMPROVED.xlsx")
fuel_data <- subset(fuel_data, select = -Year)

fuel_data$MAKE <- factor(fuel_data$MAKE)
fuel_data$MODEL <- factor(fuel_data$MODEL)
fuel_data$`VEHICLE CLASS` <- factor(fuel_data$`VEHICLE CLASS`)
fuel_data$SIZE <- factor(fuel_data$SIZE)
fuel_data$TRANSMISSION <- factor(fuel_data$TRANSMISSION)
fuel_data$FUEL <- factor(fuel_data$FUEL)
```

```
paste(colnames(fuel_data), collapse = ", ")
```

Load Data

```
## [1] "MAKE, MODEL, VEHICLE CLASS, SIZE, ENGINE SIZE, CYLINDERS, TRANSMISSION, NUMBER OF GEARS, FUEL, I
```

```
#Im taking out Fuel as there isnt much of a distribution in the data set
#fuel_data_dummies <- dummy_cols(fuel_data, select_columns = c("VEHICLE CLASS", "SIZE", "FUEL"))
fuel_data_dummies <- dummy_cols(fuel_data, select_columns = c("VEHICLE CLASS", "SIZE"))
fuel_data_dummies <- subset(fuel_data_dummies, select = -c(MAKE, MODEL, `VEHICLE CLASS`, SIZE, TRANSMISS

#paste(colnames(fuel_data_dummies), collapse = ", ")
```

```
head(fuel_data)
```

Data Transform

```
## # A tibble: 6 x 13
##   MAKE  MODEL   `VEHICLE CLASS` SIZE   `ENGINE SIZE` CYLINDERS TRANSMISSION
##   <fct> <fct>   <fct>           <fct>          <dbl>     <dbl> <fct>
## 1 ACURA 1.6EL   COMPACT         Small            1.6         4 Automatic
## 2 ACURA 1.6EL   COMPACT         Small            1.6         4 Manual
## 3 ACURA 3.2TL   SEDAN           Medium           3.2         6 Automatic
## 4 ACURA 3.5RL   SEDAN           Medium           3.5         6 Automatic
## 5 ACURA INTEGRA COMPACT         Small            1.8         4 Automatic
## 6 ACURA INTEGRA COMPACT         Small            1.8         4 Manual
## # i 6 more variables: `NUMBER OF GEARS` <dbl>, FUEL <fct>,
## #   `FUEL CONSUMPTION` <dbl>, COEMISSIONS <dbl>, TURBO <dbl>,
## #   SUPERCHARGED <dbl>
```

```
# Summary statistics for numerical variables
numerical_summary <- sapply(fuel_data[, sapply(fuel_data, is.numeric)], function(x) {
    c(mean=mean(x, na.rm=TRUE), sd=sd(x, na.rm=TRUE),
      min=min(x, na.rm=TRUE), median=median(x, na.rm=TRUE), max=max(x, na.rm=TRUE))
})
numerical_summary
```

Summary Statistics

```
##        ENGINE SIZE CYLINDERS NUMBER OF GEARS FUEL CONSUMPTION COEMISSIONS
## mean      3.265728  5.805947        4.480438        14.713615   296.80908
## sd        1.231012  1.625588        0.597153         3.307044    65.50418
## min       1.000000  3.000000        3.000000         4.900000   104.00000
## median    3.000000  6.000000        4.000000        14.400000   288.00000
## max       8.000000 12.000000        6.000000        30.200000   582.00000
##             TURBO SUPERCHARGED
## mean   0.03755869  0.004694836
```

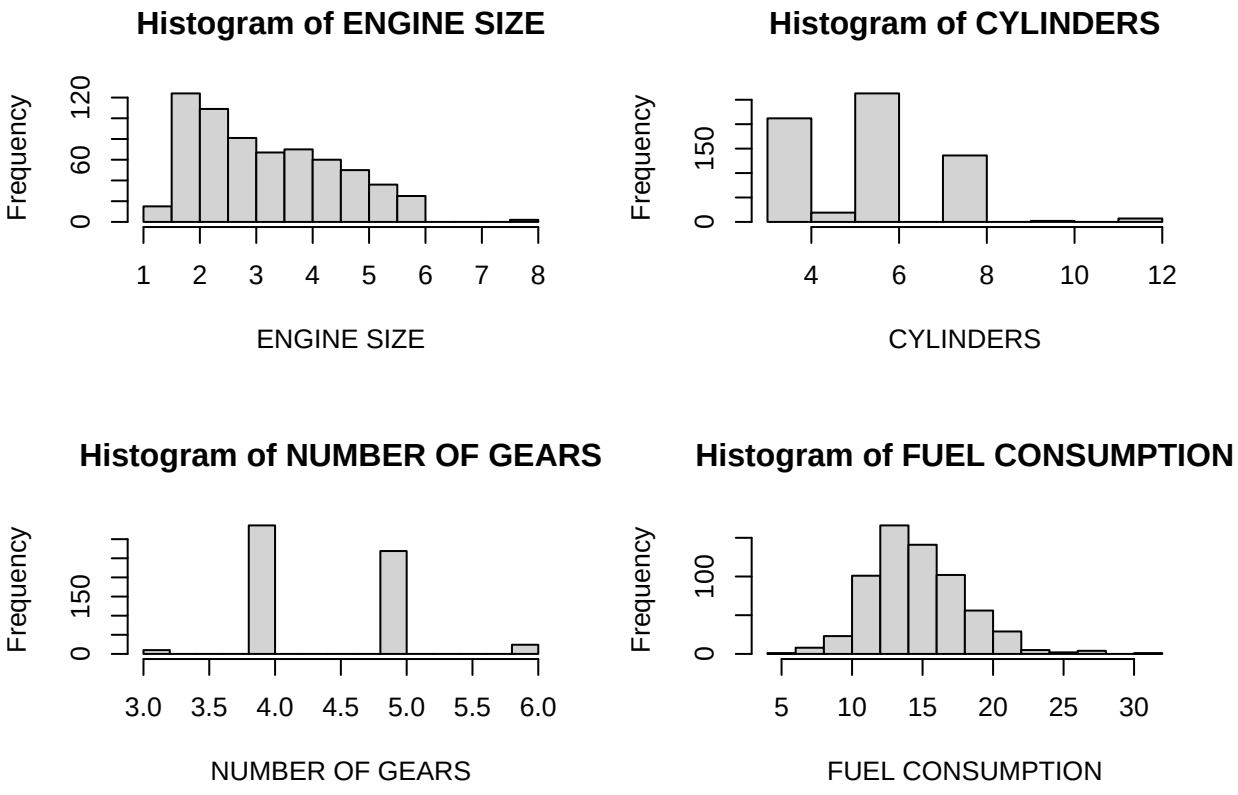
```
## sd     0.19027530  0.068411390
## min    0.00000000  0.000000000
## median 0.00000000  0.000000000
## max    1.00000000  1.000000000
```

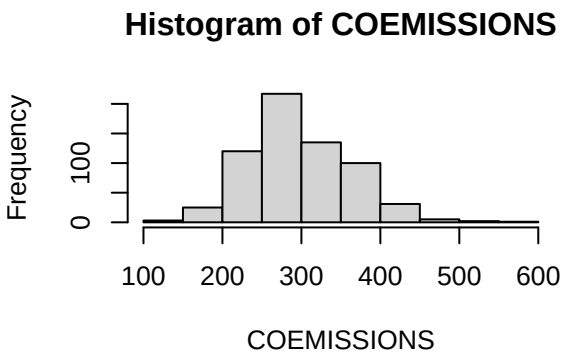
```
categorical_variables <- sapply(fuel_data, is.factor)
categorical_variables["MODEL"] <- FALSE


categorical_summary <- sapply(fuel_data[, categorical_variables], table)
categorical_summary
```

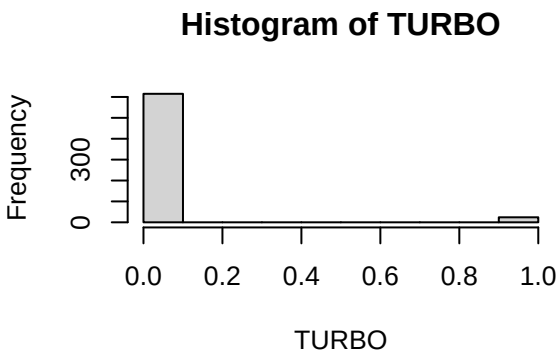
```
## $MAKE
##
##         ACURA          AUDI           BMW         BUICK      CADILLAC
##             9            22            29             6             5
##     CHEVROLET      CHRYSLER        DAEWOO         DODGE       FERRARI
##            63            19            10            46             7
##          FORD           GMC         HONDA       HYUNDAI      INFINITI
##            62            38            17            10             4
##         ISUZU        JAGUAR          JEEP           KIA    LAND ROVER
##             8            10             8             4             4
##         LEXUS       LINCOLN         MAZDA MERCEDES-BENZ        NISSAN
##             7             6            20            22            19
##    OLDSMOBILE      PLYMOUTH       PONTIAC       PORSCHE          SAAB
##             5             1            21             8             9
##        SATURN        SUBARU        SUZUKI        TOYOTA    VOLKSWAGEN
##            15            12            13            43            36
##         VOLVO
##            21
##
## $`VEHICLE CLASS`
##
## COMPACT   SEDAN   SPORT     SUV   TRUCK     VAN
##     220     105      10     161     103      40
##
## $SIZE
##
##  Large Medium  Small
##    283    102    254
##
## $TRANSMISSION
##
## Automatic    Manual
##       425       214
##
## $FUEL
##
##    CNG DIESEL    GAS
##      1      6    632
```

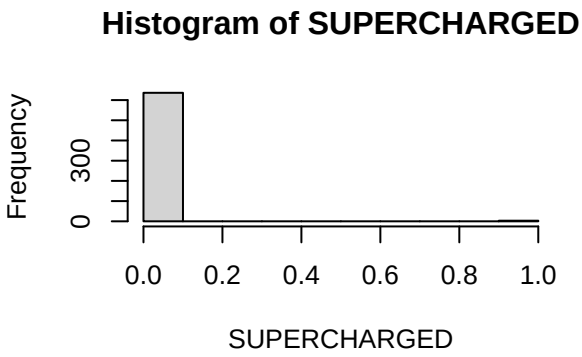
```
# Descriptive Statistics Visualization
# Histograms for numerical variables
```

```r
num_vars <- names(fuel_data[, sapply(fuel_data, is.numeric)])
par(mfrow=c(2, 2)) # Adjust the grid layout based on the number of variables
for (var in num_vars) {
  hist(fuel_data[[var]], main=paste("Histogram of", var), xlab=var)
}
```

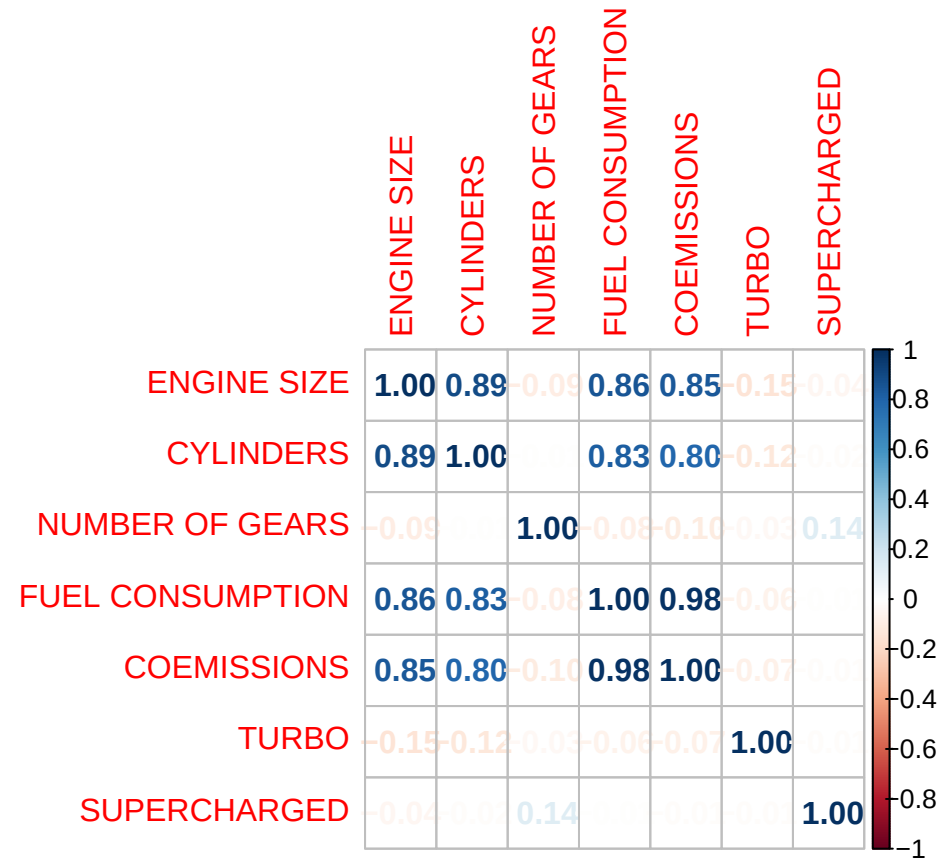









```r
# correlation Matrix
numeric_columns <- sapply(fuel_data, is.numeric)
cor_matrix <- cor(fuel_data[, numeric_columns & !names(fuel_data) %in% c("Year")])

corrplot(cor_matrix, method='number')
```



Correlation Matrix

Linear Regressions Simple Linear Regression

```
fuel_lm_model1 <- lm(COEMISSIONS ~ `ENGINE SIZE` + `NUMBER OF GEARS`, data=fuel_data)
summary(fuel_lm_model1)
```

```
##
## Call:
## lm(formula = COEMISSIONS ~ `ENGINE SIZE` + `NUMBER OF GEARS`,
##     data = fuel_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.546  -18.321   -0.648   18.871  188.473
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        161.437     11.338  14.238   <2e-16 ***
## `ENGINE SIZE`       45.093      1.116  40.395   <2e-16 ***
## `NUMBER OF GEARS`   -2.654      2.301  -1.153    0.249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 34.55 on 636 degrees of freedom
## Multiple R-squared:  0.7226, Adjusted R-squared:  0.7217
## F-statistic: 828.3 on 2 and 636 DF,  p-value: < 2.2e-16
```

Linear Regressions with Dummy Vars

I'm going to take out the 'NUMBER OF GEARS' variable as it is not significant, and add the categorical variables that I made.

```
fuel_lm_model2 <- lm(COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS`, data=fuel_data)
summary(fuel_lm_model2)
```

```
##
## Call:
## lm(formula = COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS`, data = fuel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.407 -17.798  -1.848  16.516 213.494
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            151.239      3.637  41.582  < 2e-16 ***
## `ENGINE SIZE`           39.503      1.147  34.444  < 2e-16 ***
## `VEHICLE CLASS`SEDAN    -1.339      3.691  -0.363    0.717
## `VEHICLE CLASS`SPORT    59.086      9.901   5.968 4.01e-09 ***
## `VEHICLE CLASS`SUV      30.109      3.268   9.213  < 2e-16 ***
## `VEHICLE CLASS`TRUCK    35.874      4.028   8.907  < 2e-16 ***
## `VEHICLE CLASS`VAN      39.780      5.808   6.849 1.76e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.58 on 632 degrees of freedom
## Multiple R-squared:  0.7841, Adjusted R-squared:  0.7821
## F-statistic: 382.6 on 6 and 632 DF,  p-value: < 2.2e-16
```

Now adding 'SIZE' also

```
fuel_lm_model2 <- lm(COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` + SIZE, data=fuel_data)
summary(fuel_lm_model2)
```

```
##
## Call:
## lm(formula = COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` +
##     SIZE, data = fuel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.484 -16.374  -1.489  16.749 217.162
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            187.026      8.086  23.129  < 2e-16 ***
```

```
## `ENGINE SIZE`          38.233      1.175  32.548  < 2e-16 ***
## `VEHICLE CLASS`SEDAN  -29.977      7.945  -3.773 0.000177 ***
## `VEHICLE CLASS`SPORT   59.667      9.732   6.131 1.54e-09 ***
## `VEHICLE CLASS`SUV      1.788      6.704   0.267 0.789780
## `VEHICLE CLASS`TRUCK    8.759      6.811   1.286 0.198924
## `VEHICLE CLASS`VAN     10.050      8.349   1.204 0.229159
## SIZEMedium             -4.184      4.776  -0.876 0.381320
## SIZESmall             -32.467      6.596  -4.922 1.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.05 on 630 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7895
## F-statistic: 300.2 on 8 and 630 DF,  p-value: < 2.2e-16
```

Adding 'FUEL'

```
fuel_lm_model3 <- lm(COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` + SIZE + FUEL, data=fuel_data)
summary(fuel_lm_model3)
```

```
##
## Call:
## lm(formula = COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` +
##     SIZE + FUEL, data = fuel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.931 -16.266  -2.021  15.505 216.656
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          145.6747    31.1614   4.675 3.60e-06 ***
## `ENGINE SIZE`         37.9330     1.1622  32.639  < 2e-16 ***
## `VEHICLE CLASS`SEDAN -31.3256     7.8479  -3.992 7.34e-05 ***
## `VEHICLE CLASS`SPORT  58.4328     9.6087   6.081 2.07e-09 ***
## `VEHICLE CLASS`SUV     0.3442     6.6260   0.052    0.959
## `VEHICLE CLASS`TRUCK   7.5216     6.7283   1.118    0.264
## `VEHICLE CLASS`VAN    10.0736     8.2763   1.217    0.224
## SIZEMedium            -4.4438     4.7143  -0.943    0.346
## SIZESmall            -32.8143     6.5105  -5.040 6.09e-07 ***
## FUELDIESEL            -6.4332    32.4762  -0.198    0.843
## FUELGAS               43.8523    30.0399   1.460    0.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.66 on 628 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.795
## F-statistic: 248.4 on 10 and 628 DF,  p-value: < 2.2e-16
```

Taking out FUEL as it was not statistically significant, doing what I want now

```r
fuel_lm_model4 <- lm(COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` + SIZE + TRANSMISSION, data=fuel_data
summary(fuel_lm_model4)
```

```
##
## Call:
## lm(formula = COEMISSIONS ~ `ENGINE SIZE` + `VEHICLE CLASS` +
##     SIZE + TRANSMISSION, data = fuel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.444 -17.078  -1.292  18.073 220.139
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            189.499      8.257  22.950  < 2e-16 ***
## `ENGINE SIZE`           37.918      1.194  31.770  < 2e-16 ***
## `VEHICLE CLASS`SEDAN   -30.880      7.963  -3.878 0.000116 ***
## `VEHICLE CLASS`SPORT    60.703      9.749   6.226 8.73e-10 ***
## `VEHICLE CLASS`SUV       1.394      6.704   0.208 0.835381
## `VEHICLE CLASS`TRUCK     9.153      6.810   1.344 0.179448
## `VEHICLE CLASS`VAN       9.077      8.369   1.085 0.278511
## SIZEMedium              -4.080      4.773  -0.855 0.392901
## SIZESmall              -32.263      6.592  -4.894 1.26e-06 ***
## TRANSMISSIONManual      -3.924      2.707  -1.450 0.147617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.03 on 629 degrees of freedom
## Multiple R-squared:  0.7929, Adjusted R-squared:  0.7899
## F-statistic: 267.5 on 9 and 629 DF,  p-value: < 2.2e-16
```

```r
library(modelsummary)
```

```
## Warning: package 'modelsummary' was built under R version 4.3.3
```

```
## Version 2.0.0 of `modelsummary`, to be released soon, will introduce a
##   breaking change: The default table-drawing package will be `tinytable`
##   instead of `kableExtra`. All currently supported table-drawing packages
##   will continue to be supported for the foreseeable future, including
##   `kableExtra`, `gt`, `huxtable`, `flextable, and `DT`.
##
##   You can always call the `config_modelsummary()` function to change the
##   default table-drawing package in persistent fashion. To try `tinytable`
##   now:
##
##   config_modelsummary(factory_default = 'tinytable')
##
##   To set the default back to `kableExtra`:
##
##   config_modelsummary(factory_default = 'kableExtra')
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	161.437	187.026	145.675	189.499
	(11.338)	(8.086)	(31.161)	(8.257)
ENGINE SIZE	45.093	38.233	37.933	37.918
	(1.116)	(1.175)	(1.162)	(1.194)
NUMBER OF GEARS	−2.654			
	(2.301)			
VEHICLE CLASSSEDAN		−29.977	−31.326	−30.880
		(7.945)	(7.848)	(7.963)
VEHICLE CLASSSPORT		59.667	58.433	60.703
		(9.732)	(9.609)	(9.749)
VEHICLE CLASSSUV		1.788	0.344	1.394
		(6.704)	(6.626)	(6.704)
VEHICLE CLASSTRUCK		8.759	7.522	9.153
		(6.811)	(6.728)	(6.810)
VEHICLE CLASSVAN		10.050	10.074	9.077
		(8.349)	(8.276)	(8.369)
SIZEMedium		−4.184	−4.444	−4.080
		(4.776)	(4.714)	(4.773)
SIZESmall		−32.467	−32.814	−32.263
		(6.596)	(6.511)	(6.592)
FUELDIESEL			−6.433	
			(32.476)	
FUELGAS			43.852	
			(30.040)	
TRANSMISSIONManual				−3.924
				(2.707)
Num.Obs.	639	639	639	639
R2	0.723	0.792	0.798	0.793
R2 Adj.	0.722	0.790	0.795	0.790
AIC	6345.8	6173.3	6158.4	6173.1
BIC	6363.6	6217.9	6211.9	6222.2
Log.Lik.	−3168.886	−3076.631	−3067.185	−3075.565
F	828.333	300.160	248.432	267.509
RMSE	34.47	29.84	29.40	29.79

```
model_list <- list(
  "Model 1" = fuel_lm_model1,
  "Model 2" = fuel_lm_model2,
  "Model 3" = fuel_lm_model3,
  "Model 4" = fuel_lm_model4
)

summary_table <- modelsummary(model_list)

summary_table
```

```
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
```

```
stargazer(fuel_lm_model1, fuel_lm_model2, fuel_lm_model3, fuel_lm_model4,
          type = "text", title = "Model Summary",
          align = TRUE, header = FALSE)
```

```
##
## Model Summary
## ================================================================================================
##                                                  Dependent variable:
##                     ---------------------------------------------------------------------------
##                                                       COEMISSIONS
##                               (1)                      (2)                      (3)
## ------------------------------------------------------------------------------------------------
## `ENGINE SIZE`              45.093***                38.233***                37.933***
##                             (1.116)                  (1.175)                  (1.162)
##
## `NUMBER OF GEARS`           -2.654
##                             (2.301)
##
## `VEHICLE CLASS`SEDAN                                -29.977***               -31.326***
##                                                      (7.945)                  (7.848)
##
## `VEHICLE CLASS`SPORT                                59.667***                58.433***
##                                                      (9.732)                  (9.609)
##
## `VEHICLE CLASS`SUV                                   1.788                    0.344
##                                                      (6.704)                  (6.626)
##
## `VEHICLE CLASS`TRUCK                                 8.759                    7.522
##                                                      (6.811)                  (6.728)
##
## `VEHICLE CLASS`VAN                                   10.050                   10.074
##                                                      (8.349)                  (8.276)
##
## SIZEMedium                                           -4.184                   -4.444
##                                                      (4.776)                  (4.714)
##
## SIZESmall                                           -32.467***               -32.814***
##                                                      (6.596)                  (6.511)
##
## FUELDIESEL                                                                    -6.433
##                                                                              (32.476)
##
## FUELGAS                                                                       43.852
##                                                                              (30.040)
##
## TRANSMISSIONManual
##
##
## Constant                   161.437***               187.026***               145.675***
```

```
##                                (11.338)                (8.086)                (31.161)
##
## ---------------------------------------------------------------------------------------------------
## Observations                     639                     639                     639
## R2                              0.723                   0.792                   0.798
## Adjusted R2                     0.722                   0.790                   0.795
## Residual Std. Error      34.555 (df = 636)       30.052 (df = 630)       29.658 (df = 628)
## F Statistic           828.333*** (df = 2; 636) 300.160*** (df = 8; 630) 248.432*** (df = 10; 628) 267
## ===================================================================================================
## Note:                                                                                 *p<0.1
```

Ridge Regression

```
#str(fuel_data_dummies)
```

```
#install.packages("glmnet")
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.3.3
```

```
## Loaded glmnet 4.1-8
```

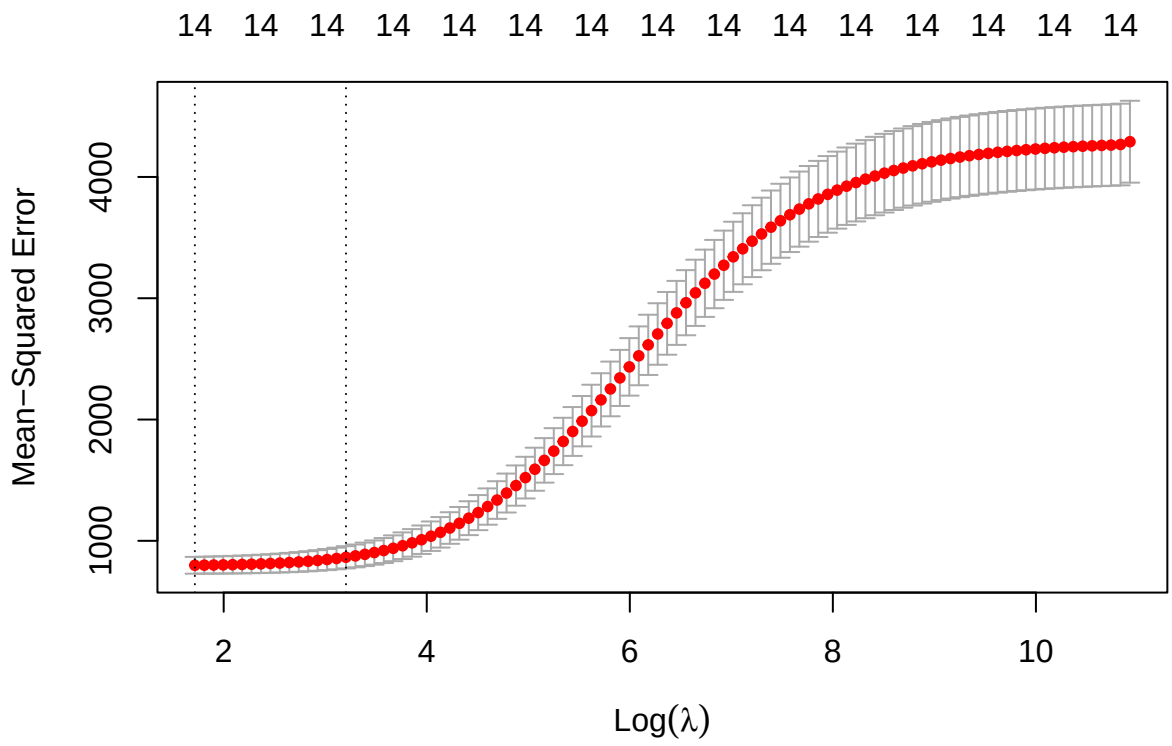
```
y <- fuel_data_dummies$COEMISSIONS
lambda_values <- 10^seq(10, -2, length = 100)

ridge_model <- glmnet(fuel_data_dummies, y, alpha = 0, lambda = lambda_values)
```

```
#This will spit out all of the coefficients

#coef(ridge_model, s = 0.01)
```

```
coef_matrix <- as.matrix(coef(ridge_model, s = 0.01))

coef_df <- as.data.frame(coef_matrix, stringsAsFactors = FALSE)
names(coef_df) <- "Coefficient"

coef_df$Variable <- rownames(coef_df)

coef_df <- coef_df[order(-abs(coef_df$Coefficient)), ]

top_10_coef <- head(coef_df, 10)
print(top_10_coef)
```

```
##                         Coefficient              Variable
## COEMISSIONS               0.9985245           COEMISSIONS
## SIZE_Small               -0.8193850            SIZE_Small
## SIZE_Medium              -0.7640605           SIZE_Medium
## SIZE_Large               -0.7468186            SIZE_Large
## VEHICLE CLASS_SPORT       0.5543246   VEHICLE CLASS_SPORT
## VEHICLE CLASS_VAN         0.5062250     VEHICLE CLASS_VAN
## VEHICLE CLASS_TRUCK       0.4997911   VEHICLE CLASS_TRUCK
## (Intercept)               0.4907870           (Intercept)
## VEHICLE CLASS_COMPACT     0.4845282 VEHICLE CLASS_COMPACT
## VEHICLE CLASS_SUV         0.4731350     VEHICLE CLASS_SUV
```

```r
x <- as.matrix(fuel_data_dummies %>% dplyr::select(-c(COEMISSIONS)))
y <- fuel_data_dummies$COEMISSIONS

cv_ridge <- cv.glmnet(x, y, alpha = 0)

plot(cv_ridge)
```




```r
best_lambda <- cv_ridge$lambda.min

final_model <- glmnet(fuel_data_dummies, y, alpha = 0, lambda = best_lambda)

final_coef <- coef(final_model, s = best_lambda)
final_coef_matrix <- as.matrix(final_coef)
```

```r
final_coef_df <- as.data.frame(final_coef_matrix, stringsAsFactors = FALSE)
names(final_coef_df) <- "Coefficient"
final_coef_df$Variable <- rownames(final_coef_df)

final_coef_df <- final_coef_df[order(-abs(final_coef_df$Coefficient)), ]
top_10_final_coef <- head(final_coef_df, 10)
print(top_10_final_coef)
```

```
##                     Coefficient            Variable
## (Intercept)           44.339824         (Intercept)
## VEHICLE CLASS_SPORT   12.301270 VEHICLE CLASS_SPORT
## VEHICLE CLASS_SEDAN   -7.636590 VEHICLE CLASS_SEDAN
## VEHICLE CLASS_VAN      6.942059   VEHICLE CLASS_VAN
## TURBO                  5.719083               TURBO
## VEHICLE CLASS_TRUCK    5.101337 VEHICLE CLASS_TRUCK
## ENGINE SIZE            4.925938         ENGINE SIZE
## CYLINDERS              4.609454           CYLINDERS
## SIZE_Large             4.247873          SIZE_Large
## SIZE_Small            -4.020796          SIZE_Small
```