STAT 456 FINAL PROJECT

SRI MEDICHERLA, MICHELLE HARRIS, HAILEY LEE

BUSINESS PROBLEM

Which variables significantly predict the price of a car?

How well do these variables explain car prices?

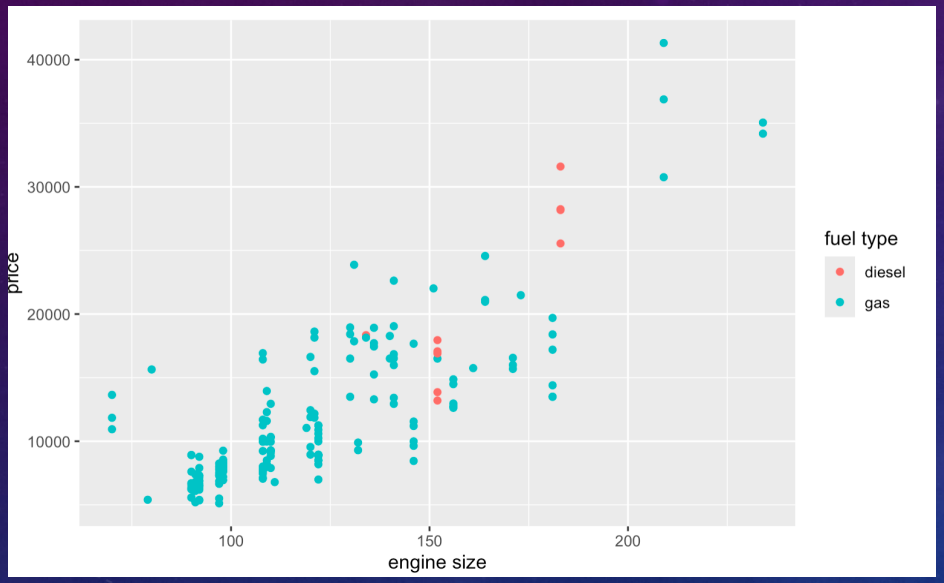
PROPOSED SOLUTION

- Conduct Exploratory Data Analysis (EDA)
- Split dataset into training and testing
- Investigate each variable's effect on price through modeling
- Determine which model is the strongest through testing each model

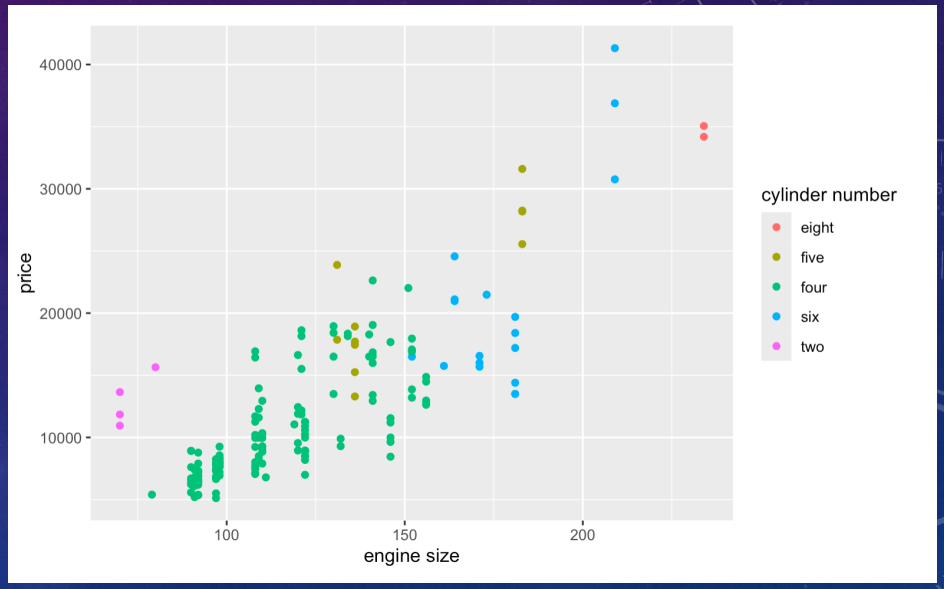
EDA

EXPLORATORY PLOTS: VARIABLE RELATIONSHIPS WITH PRICE

EDA CONTD...



Plot of engine size vs. Price colored by fuel type



Plot of engine size vs. Price colored by number of cylinders

MULTIPLE LINEAR REGRESSION (MLR): MODEL 1

- Started with linear model with ALL variables
- Adjusted R^2 value: 0.9762
- P-value: 9.666e-16

```
Residual standard error: 1233 on 23 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.9762
F-statistic: 47.16 on 181 and 23 DF,  p-value: 9.666e-16
```

MLR MODEL 2

- Next, model with only continuous variables
- Adjusted R^2 value: 0.8589
- RMSE: 3224.741
- P-value: 2.2e-16

```
lm(formula = price ~ ., data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12049.8  -1814.8   -172.7   1408.0  12791.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -48703.162  17230.845  -2.827 0.005343 **
wheelbase           153.510    110.705   1.387 0.167588
carlength           -35.536     60.793  -0.585 0.559727
carwidth            255.082    268.948   0.948 0.344418
carheight           114.007    149.557   0.762 0.447069
curbweight            1.136      1.970   0.577 0.564857
enginesize          123.411     15.679   7.871 6.32e-13 ***
boreratio          -324.821   1468.923  -0.221 0.825291
stroke            -2975.049    843.628  -3.526 0.000558 ***
compressionratio    310.901    102.361   3.037 0.002813 **
horsepower           37.003     19.855   1.864 0.064304 .
peakrpm               3.233      0.911   3.549 0.000517 ***
citympg            -216.214    203.923  -1.060 0.290713
highwaympg          145.037    183.443   0.791 0.430395
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3253 on 151 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8589
F-statistic: 77.79 on 13 and 151 DF,  p-value: < 2.2e-16
```

MLR MODEL 3

- Model with significant codes variables
- Adjusted R^2 Value: 0.844
- RMSE: 3353.8
- P-value: 2.2e-16

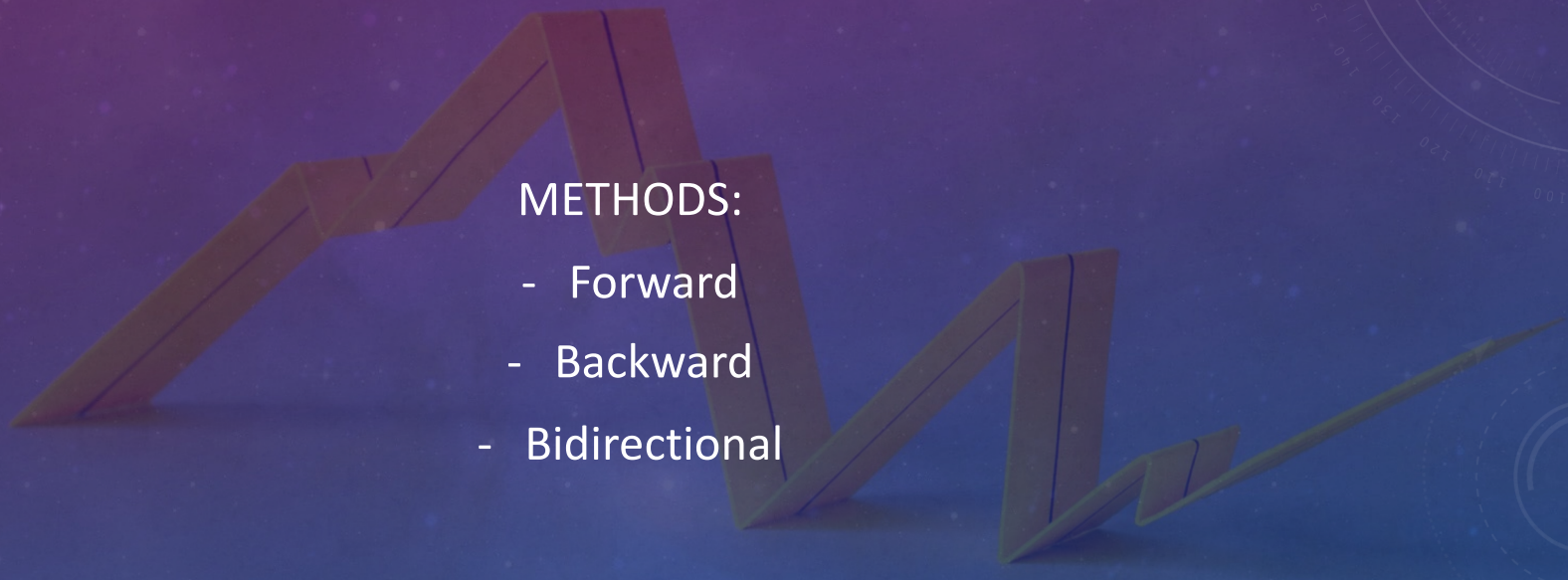
```
Call:
lm(formula = price ~ stroke + compressionratio + peakrpm + enginesize
+
    horsepower, data = train_data)

Residuals:
   Min     1Q Median     3Q    Max
-15377  -1636   -355   1525  12585

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.806e+04  4.897e+03  -3.688 0.000310 ***
stroke          -2.715e+03  8.556e+02  -3.173 0.001811 **
compressionratio  3.641e+02  8.633e+01   4.218 4.13e-05 ***
peakrpm           2.609e+00  8.338e-01   3.129 0.002086 **
enginesize        1.419e+02  1.410e+01  10.061  < 2e-16 ***
horsepower        5.132e+01  1.446e+01   3.550 0.000506 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3420 on 159 degrees of freedom
Multiple R-squared:  0.8488,    Adjusted R-squared:  0.844
F-statistic: 178.5 on 5 and 159 DF,  p-value: < 2.2e-16
```

STEPWISE REGRESSION



METHODS:

- Forward
- Backward
- Bidirectional

```
#Stepwise Regression

```{r}
initial_model <- lm(price ~ ., data = train_data)
#stepwise using aic
stepwise_model_aic <- step(initial_model, direction = "both", trace = 1, k =
log(nrow(train_data)), criterion = "aic")
#stepwise using bic
stepwise_model_bic <- step(initial_model, direction = "both", trace = 1, k =
log(nrow(train_data)), criterion = "bic")
#stepwise using adjr2
stepwise_model_adjr2 <- step(initial_model, direction = "both", trace = 1, k =
log(nrow(train_data)), criterion = "adjr2")
summary(stepwise_model_aic)
summary(stepwise_model_bic)
summary(stepwise_model_adjr2)
```
```

- Stepwise regression code using AIC, BIC, and adjr2 as the criterion

- P-value: nearly 0
- Adjusted R^2: .8621

```
> summary(stepwise_model_aic)

Call:
lm(formula = price ~ wheelbase + enginesize + stroke + compressionratio +
    horsepower + peakrpm, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12790.1  -1726.3   -244.4   1494.6  12381.9

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -41674.100   6842.603  -6.090 8.25e-09 ***
wheelbase           248.872     53.337   4.666 6.51e-06 ***
enginesize          119.839     14.077   8.513 1.24e-14 ***
stroke            -2875.852    805.365  -3.571 0.000472 ***
compressionratio    300.989     82.314   3.657 0.000348 ***
horsepower           57.279     13.655   4.195 4.54e-05 ***
peakrpm               3.096      0.791   3.914 0.000135 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3217 on 158 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8621
F-statistic: 171.8 on 6 and 158 DF,  p-value: < 2.2e-16
```

Notice that the p-values and adjusted R^2 using the 'BIC' and 'adjr2' criterion are the same as using the 'AIC' criterion.

```
> summary(stepwise_model_bic)

Call:
lm(formula = price ~ wheelbase + enginesize + stroke + compressionratio +
    horsepower + peakrpm, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12790.1  -1726.3   -244.4   1494.6  12381.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -41674.100   6842.603  -6.090 8.25e-09 ***
wheelbase           248.872     53.337   4.666 6.51e-06 ***
enginesize          119.839     14.077   8.513 1.24e-14 ***
stroke            -2875.852    805.365  -3.571 0.000472 ***
compressionratio    300.989     82.314   3.657 0.000348 ***
horsepower           57.279     13.655   4.195 4.54e-05 ***
peakrpm               3.096      0.791   3.914 0.000135 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3217 on 158 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8621
F-statistic: 171.8 on 6 and 158 DF,  p-value: < 2.2e-16

> summary(stepwise_model_adjr2)

Call:
lm(formula = price ~ wheelbase + enginesize + stroke + compressionratio +
    horsepower + peakrpm, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-12790.1  -1726.3   -244.4   1494.6  12381.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -41674.100   6842.603  -6.090 8.25e-09 ***
wheelbase           248.872     53.337   4.666 6.51e-06 ***
enginesize          119.839     14.077   8.513 1.24e-14 ***
stroke            -2875.852    805.365  -3.571 0.000472 ***
compressionratio    300.989     82.314   3.657 0.000348 ***
horsepower           57.279     13.655   4.195 4.54e-05 ***
peakrpm               3.096      0.791   3.914 0.000135 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3217 on 158 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.8621
F-statistic: 171.8 on 6 and 158 DF,  p-value: < 2.2e-16
```

LET'S EXPLORE WHY THAT MAY HAPPEN...

- Small dataset
- Overfitting
- Correlated predictors

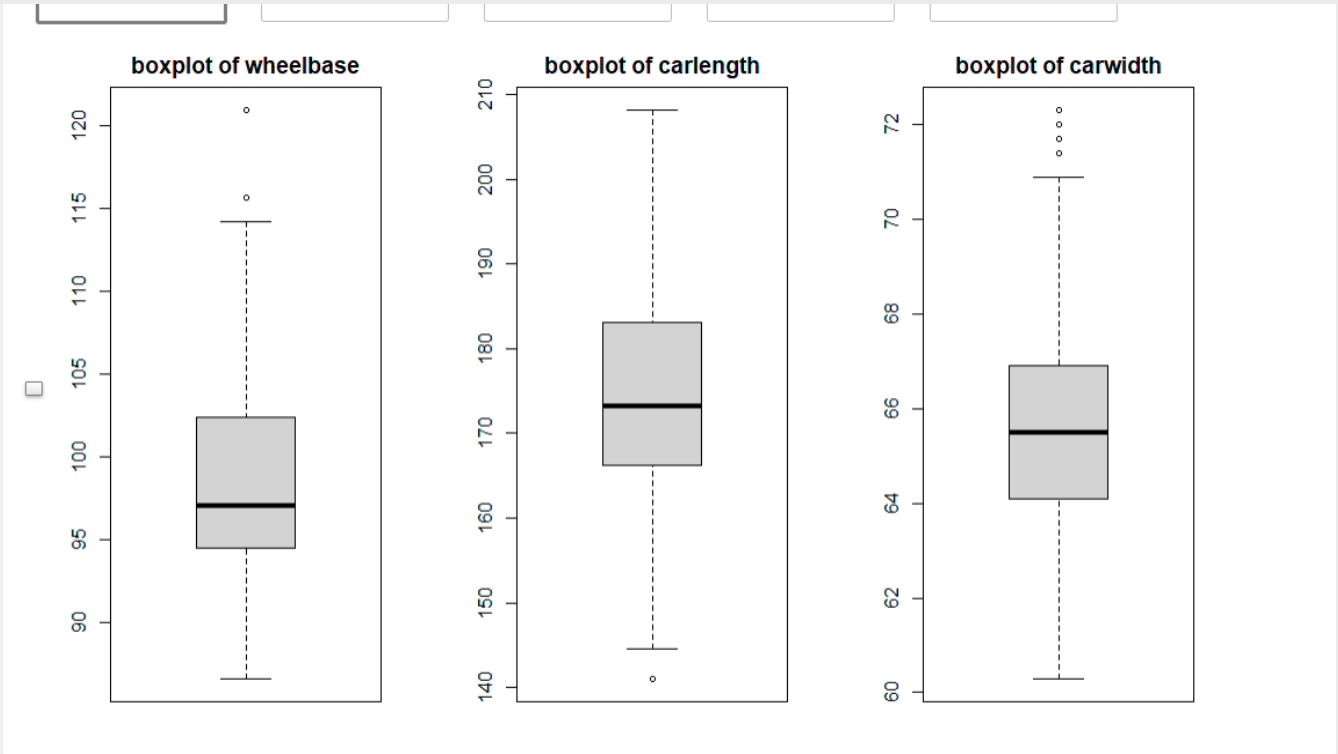
```
131
132 #Testing the stepwise model
133
134 ```{r}
135 stepwise_model_selected <- stepwise_model_aic
136 predictions <- predict(stepwise_model_selected, newdata = test_data)
137 rmse <- sqrt(mean((test_data$price - predictions)^2))
138 rmse
139 ```
```

```
[1] 3493.335
```

- Testing the model
- RMSE: 3493.335

RMSE is a little high... maybe outliers?

EXPLORING OUTLIERS AND REMOVING THEM

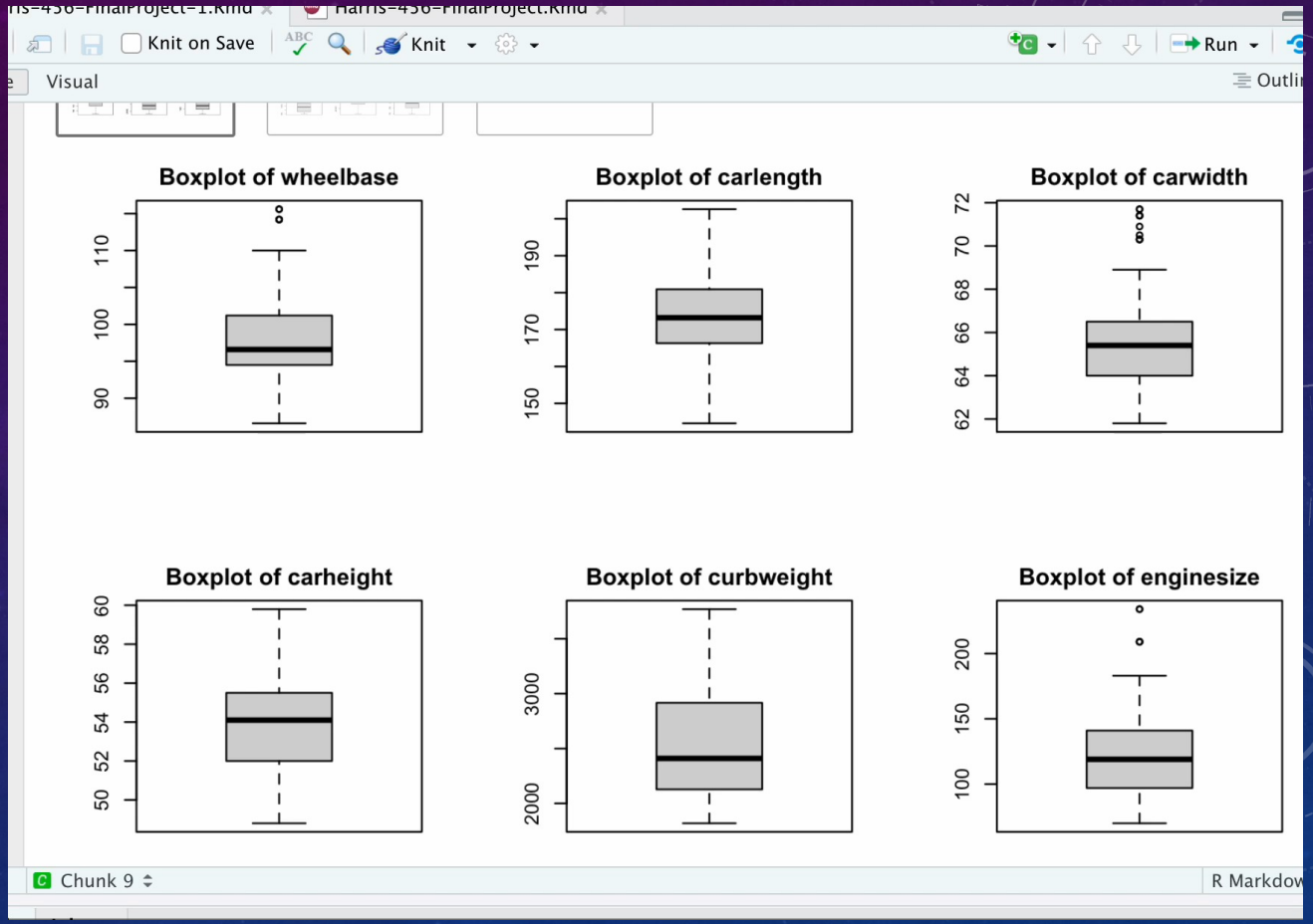


- Boxplots of original continuous variables

INVESTIGATION: REMOVING OUTLIERS?

RMSE before: 3493.335

RMSE after: 6028.778



STEPWISE REGRESSION MODEL VS LINEAR REGRESSION MODEL

STEPWISE MODEL (before removing outliers)
Price = -41674.100 + 248.872 * wheelbase + 119.839 * enginesize – 2875.852 * stroke + 300.989 * compressionratio + 57.279 * horsepower + 3.096 * peakrpm

MLR MODEL 2: (BEST)
Price = -48703.162 + 153.510 * wheelbase – 35.536 * carlength + 255.082 * carwidth + 114.007 * carheight + 1.136 * curbweight + 123.411 * enginesize – 324.821 * boreratio –2975.049 * stroke + 310.901 * compressionratio + 37.003 * horsepower – 3.233 * peakrpm – 216.214 * citympg + 145.037 * highwaympg

FINDINGS AND CONCLUSIONS

•Which variables are most significant?

- Enginesize
- Stroke
- Compressionratio
- Peakrp