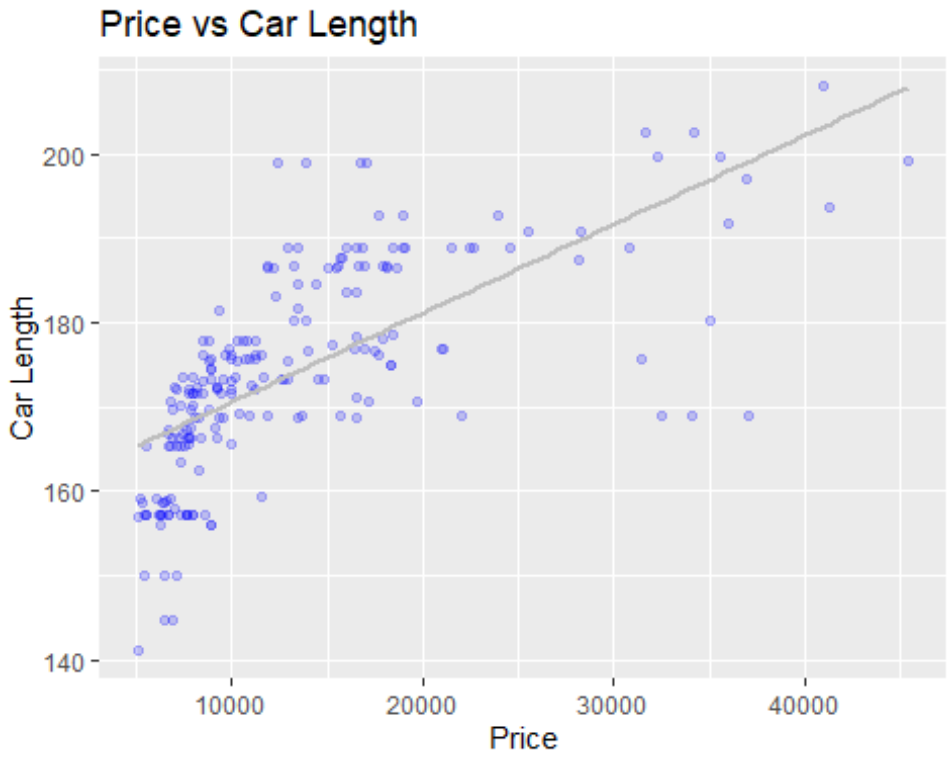
HW2

Background

Question 1: Exploratory Data Analysis [12 points]

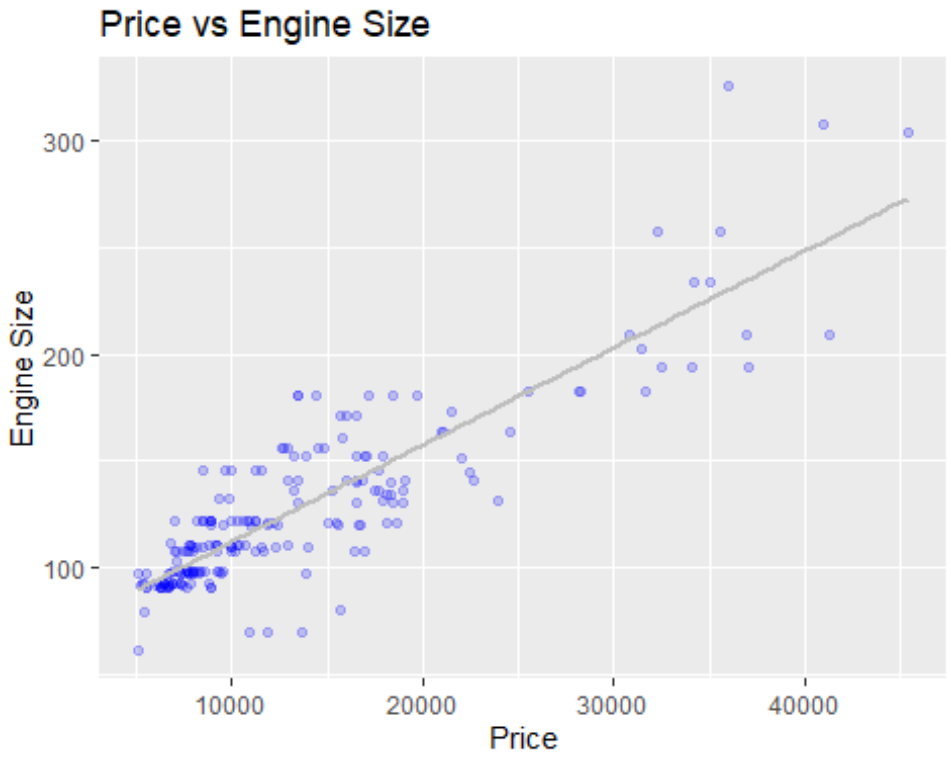
a. **3 pts** Create plots of the response, *price*, against three quantitative predictors (for simplicity) *carlength*, *enginesize*, and *horsepower*. Describe the general trend (direction and form) of each plot.

```
#Plot Price against car length
ggplot(data=CP, aes(x=CP$price, y=CP$carlength)) +
geom_point(alpha=I(0.2),color='blue') +
xlab('Price') +
ylab('Car Length') +
ggtitle('Price vs Car Length') +
geom_smooth(method= "lm",color='gray', se=FALSE)
```

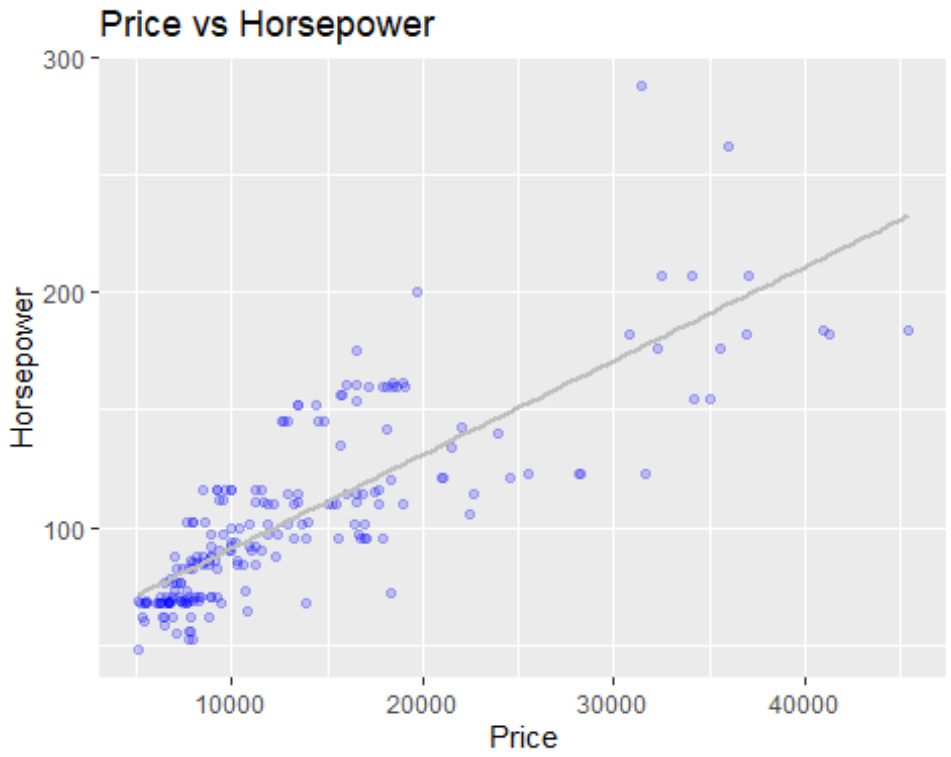



```
#Plot price against engine size
ggplot(data=CP, aes(x=CP$price, y=CP$enginesize)) +
geom_point(alpha=I(0.2),color='blue') +
xlab('Price') +
ylab('Engine Size') +
```

```
ggtitle('Price vs Engine Size') +
geom_smooth(method= "lm",color='gray', se=FALSE)
```




```
#plot price against horse power
ggplot(data=CP, aes(x=CP$price, y=CP$horsepower)) +
geom_point(alpha=I(0.2),color='blue') +
xlab('Price') +
ylab('Horsepower') +
ggtitle('Price vs Horsepower') +
geom_smooth(method= "lm",color='gray', se=FALSE)
```



All three plots have a positive direction, indicating there is an increasing trend in price as each of the three predictors increase. As each predictor increases, the variance of the price appears to increase.

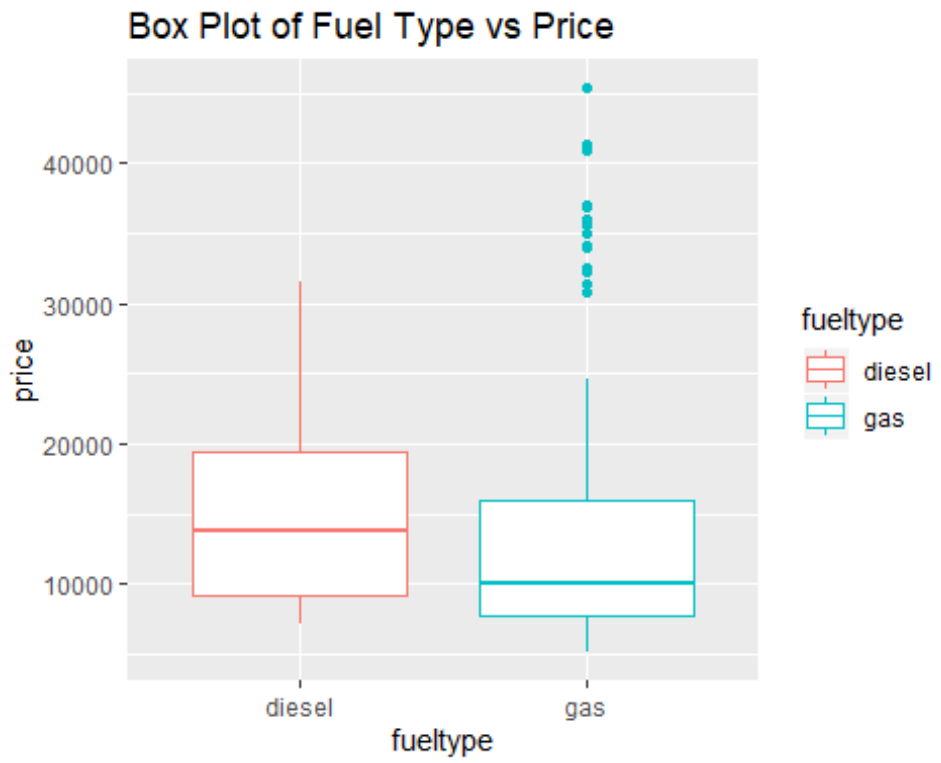
b. **3 pts** What is the value of the correlation coefficient for each of the above pair of response and predictor variables? What does it tell you about your comments in part (a).

```
#Correlation between price and car length
cor(CP$price, CP$carlength)

## [1] 0.68292

#Correlation between price and engine size
cor(CP$price, CP$enginesize)

## [1] 0.8741448

#Correlation between price and horsepower
cor(CP$price, CP$horsepower)

## [1] 0.8081388
```

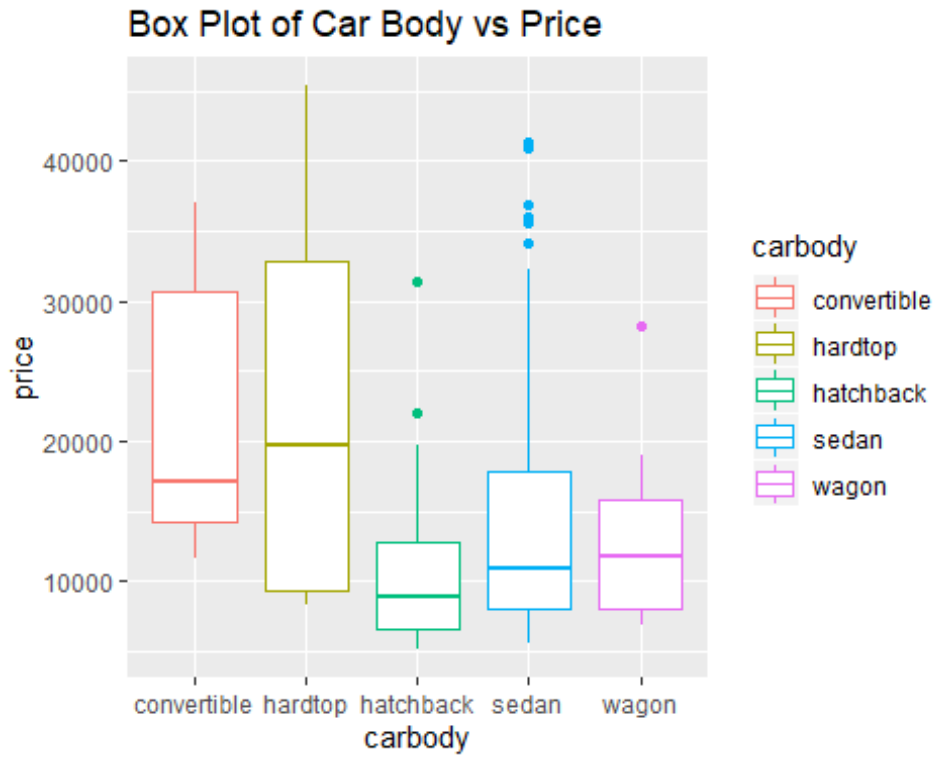
The correlation coefficient between price and car length is 0.68292, indicating a moderate positive relationship. The correlation coefficient between price and engine size is 0.8741448, indicating a strong positive relationship. The correlation coefficient for price and horsepower is 0.8081388 indiciating a strong positive relationship.

c. **3 pts** Create box plots of the response, *price*, and the two qualitative predictors *fueltype*, and *carbody*. Based on these box plots, does there appear to be a relationship between these qualitative predictors and the response?

```
ggplot(CP, aes(x=fueltype, y=price, color=fueltype)) +
geom_boxplot() +
ggtitle("Box Plot of Fuel Type vs Price")
```




```
ggplot(CP, aes(x=carbody, y=price, color=carbody)) +
geom_boxplot() +
ggtitle("Box Plot of Car Body vs Price")
```



In the first plot, we don't see a strong relationship between fuel type and price. The medians of the two types are similar, with diesel's median price slightly under 15,000 and gas right around$10,000, with diesel car prices appearing slightly higher than gas cars. Cars with gas fuel types seem to have more price variance, with multiple outliers on the higher end of price range.

The second plot shows a stronger relationship between car body and price. Convertible and hard top cars appear to have higher prices than hatchback, sedan, and wagon cars, with hatchbacks appearing to have the lowest prices.

d. **3 pts** Based on the analysis above, does it make sense to run a multiple linear regression with all of the predictors?

Yes, all predictors seem to have a moderate, if not strong relationship to the response variable.

Note: Please work on non-transformed data for all of the following questions.

Question 2: Fitting the Multiple Linear Regression Model [10 points]

Build a multiple linear regression model, named *model1*, using the response, *price*, and all 7 predictors, and then answer the questions that follow:

a. **5 pts** Report the coefficient of determination for the model and give a concise interpretation of this value.

```
model1 = lm(price ~.,data=CP)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ ., data = CP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9012.6 -1848.1   -48.1  1658.0 13011.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.235e+04  9.325e+03  -2.397 0.017502 *
## fueltypegas      -3.810e+03  9.596e+02  -3.970 0.000101 ***
## carbodyhardtop   -2.904e+03  1.790e+03  -1.622 0.106401
## carbodyhatchback -5.128e+03  1.436e+03  -3.571 0.000449 ***
## carbodysedan     -4.305e+03  1.477e+03  -2.914 0.003985 **
## carbodywagon     -5.504e+03  1.618e+03  -3.402 0.000811 ***
## carlength         9.564e+01  3.915e+01   2.443 0.015471 *
## enginesize        1.032e+02  1.277e+01   8.082 6.69e-14 ***
## horsepower        4.703e+01  1.383e+01   3.400 0.000818 ***
## peakrpm           2.126e+00  6.605e-01   3.218 0.001512 **
## highwaympg       -6.235e+01  6.868e+01  -0.908 0.365114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3272 on 194 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8323
## F-statistic: 102.2 on 10 and 194 DF,  p-value: < 2.2e-16
```

The coefficient of determination, or the multiple R-squared value is 0.8405, indicating that around 84% of the total variability in the price variable can be explained by this regression model.

b. **5 pts** Is the model of any use in predicting price? Conduct a test of overall adequacy of the model, using $\alpha = 0.05$. Provide the following elements of the test: null hypothesis H_0, alternative hypothesis H_a, F- statistic or p-value, and conclusion.

```
which(summary(model1)$coeff[,4]>0.05)
```

```
## carbodyhardtop     highwaympg
##              3              11
```

H_0: All regression coefficients except the intercept are 0 H_a: At least one of the predictor coefficients is different from 0 at the alpha significance level. F-Statistic: 102.2 p-value: approximately 0 Conclusion: At least one of the predictors have predictive power.

Question 3: Model Comparison [12 points]

a. **4 pts** Assuming a marginal relationship between the car's body type and its price, perform an ANOVA F-test on the means of the car's body types. Using an $\alpha - level$ of

0.05, can we reject the null hypothesis that the means of the car body types are equal? Please interpret.

```
#H_0: all means are equal
#H_a: some means are different
aov_mod <- aov(price~carbody,data=CP)
summary(aov_mod)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## carbody       4 1.802e+09 450499206   8.032 5.03e-06 ***
## Residuals   200 1.122e+10  56088213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the F-test is 0.00000503 which is less than 0.05, so we can reject the null hypothesis that the means of the car body types are equal and that the mean price of at least car body type is different.

b. **4 pts** Now, build a second multiple linear regression model, called *model2*, using *price* as the response variable, and all variables except *carbody* as the predictors. Conduct a partial F-test comparing *model2* with *model1*. What is the partial-F test p-value? Can we reject the null hypothesis that the regression coefficients for *carbody* are zero at $\alpha - level$ of 0.05?

```
model2 = lm(price ~., CP[,-2])

anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ fueltype + carbody + carlength + enginesize + horsepower
+
##     peakrpm + highwaympg
## Model 2: price ~ fueltype + carlength + enginesize + horsepower + peakrpm
+
##     highwaympg
##   Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
## 1    194 2076673688
## 2    198 2254940295 -4 -178266607 4.1634 0.002941 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#H_0: Regression coefficients for carbody are zero at alpha level of 0.05
```

The output provides the partial F-value (4.1634) and p-value (0.002941) which is less than an alpha of 0.05, thus we reject the null hypothesis that regression coefficients for carbody are zero at an alpha level of 0.05

c. **4 pts** What can you conclude from a and b? Do they provide the exact same results?

The tests do not provide the exact same results since they test for different hypotheses, but are both indicate that there is a relationship between car body type and price.

Question 4: Coefficient Interpretation [6 points]

a. **3 pts** Interpret the estimated coefficient of *fueltypegas* in the context of the problem. *Mention any assumption you make about other predictors clearly when stating the interpretation.*

Qualitative variables impact the intercept of a model. The intercept is -22,350 and the estimated coefficient of fueltype gas is -3,810 which suggests that when fueltypegas is zero and the other fuel type (diesel) is one, then the intercept is -26,160.

b. **3 pts** If the value of the *enginesize* in the above model is increased by 0.01 keeping other predictors constant, what change in the response would be expected?

The coefficient for engine size is 103.2. Assuming all else in the model is fixed, the price is expected to change by $1.032 with a .01 increase in engine size.

Question 5: Confidence and Prediction Intervals [10 points]

a. **5 pts** Compute 90% and 95% confidence intervals (CIs) for the parameter associated with *carlength* for the model in Question 2. What observations can you make about the width of these intervals?

```
confint(model1,"carlength",level=0.90)
```

```
##                 5 %    95 %
## carlength 30.93222 160.356
```

```
confint(model1,"carlength",level=0.95)
```

```
##               2.5 %   97.5 %
## carlength 18.42162 172.8666
```

Since zero is not within the range of either of these intervals, we can assume that the parameter is statistically significant. THe width of the interveals increase as the confidence intervals increase from 90% to 95%

b. **2.5 pts** Using *model1*, estimate the average price for all cars with the same characteristics as the first data point in the sample. What is the 95% confidence interval for this estimation? Provide an interpretation of your results.

```
## Design Matrix
X = model.matrix(model1)

## First datapoint
xstar = X[1,] # first row
newdata = CP[1,-8] #remove last column

## Confidence Interval
predict(model1, newdata, interval="confidence")
```

```
##        fit      lwr      upr
## 1 17565.19 14885.37 20245.02
```

THe average estimated price for all cars with the same characteristics as the first data point is $17,565.19. The 95% confidence interval is (14885.37,20245.02)

c. **2.5 pts** Suppose that the *carlenght* value for the first data point is increased to 200, while all other values are kept fixed. Using *model1*, predict the price of a car with these characteristics. What is the 95% prediction interval for this prediction? Provide an interpretation of your results.

```
newnew <- newdata
newnew$carlength = 200
predict(model1, newnew, interval="confidence")

##        fit      lwr      upr
## 1 20549.29 16894.85 24203.73
```

The average price of cars for the same values of the first data point with the exception of a new carlength of 200 is $20,549.29. The 95% prediction interval is (16894.85,24203.73). From these results, it would seem that all other factors holding, an increase in car length resulted in an increase in price. The confidence interval for the increased car length did widen a bit, indicating more slightly more variability in the data for car lengths about this new size.