Assignment 9: How much for that car?

FirstName LastName

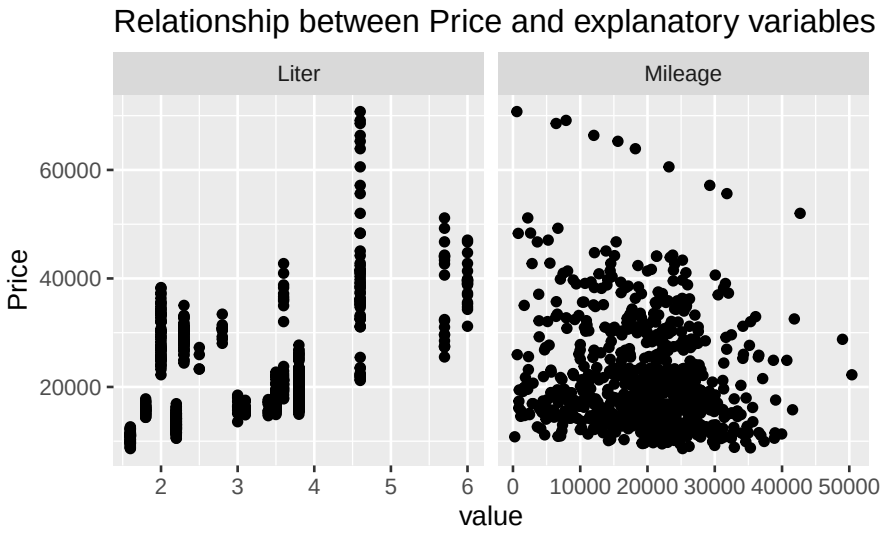
2022-10-28

Exercise 1

i. The other continuous variable in this dataset is Mileage.

ii.

```
car_prices %>%
  pivot_longer(cols = c('Liter', 'Mileage'),
               names_to = "names",
               values_to = "value") %>%
  ggplot() +
  geom_point(mapping = aes(x = value, y = Price)) +
  facet_wrap(~ names, scales = "free_x") +
  labs(title="Relationship between Price and explanatory variables")
```



Exercise 2

```
continuous_model <- lm(Price ~ Liter + Mileage, data = car_prices)
```

```
continuous_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	9426.6014688	1095.0777745	8.608157	0.0e+00
Liter	4968.2781155	258.8011436	19.197280	0.0e+00
Mileage	-0.1600285	0.0349084	-4.584237	5.3e-06

```
continuous_model %>%
  glance() %>%
  select(r.squared:statistic)
```

r.squared	adj.r.squared	sigma	statistic
0.3291279	0.3274528	8106.466	196.4841

- R^2 is a measure of how correlated the explanatory and response variables are. If $R^2 = 1$, then all the points fall on a straight line. If $R^2 = 0$, then there is no correlation between the variables. In our case, $R^2 = 0.3291279$, which is closer to 0 than to 1.

Exercise 3

```
# predict model plane over values
lit <- unique(car_prices$Liter)
mil <- unique(car_prices$Mileage)
grid <- with(car_prices, expand.grid(lit, mil))
d <- setNames(data.frame(grid), c("Liter", "Mileage"))
vals <- predict(continuous_model, newdata = d)

# form surface matrix and give to plotly
m <- matrix(vals, nrow = length(unique(d$Liter)), ncol = length(unique(d$Mileage)))
p <- plot_ly() %>%
  add_markers(
    x = ~car_prices$Mileage,
    y = ~car_prices$Liter,
    z = ~car_prices$Price,
    marker = list(size = 1)
    ) %>%
  add_trace(
    x = ~mil, y = ~lit, z = ~m, type="surface",
    colorscale=list(c(0,1), c("yellow","yellow")),
    showscale = FALSE
    ) %>%
  layout(
    scene = list(
      xaxis = list(title = "mileage"),
      yaxis = list(title = "liters"),
      zaxis = list(title = "price")
    )
```

```
  )
if (!is_pdf) {p}
```

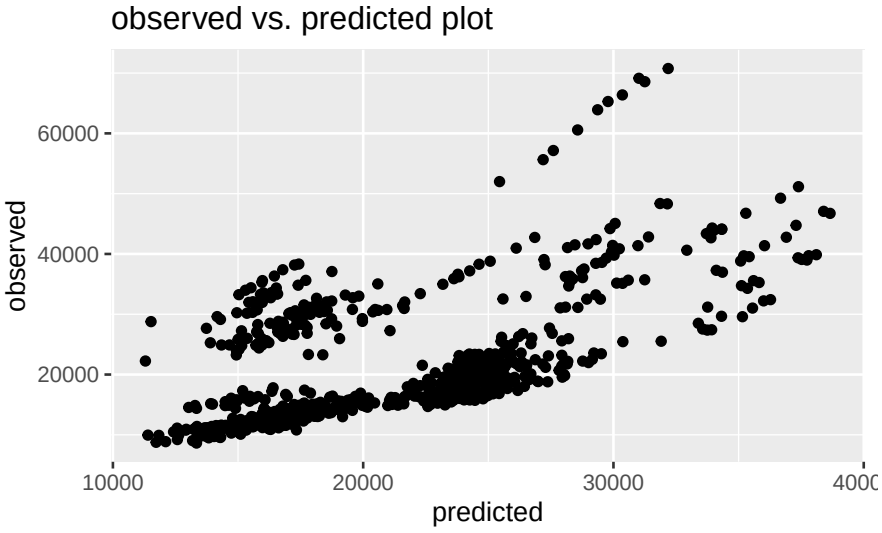
- By examining and rotating the 3D plot, the model can describe the focus direction of the data, but it cannot accurately fit the entire data because the points are scattered relative to the location of the model. From the 3D graph, it seems like the model meets the assumption of linearity because there is no obvious curve in the relationship. The variability of the residuals (above and below the model surface) seems to be reasonably constant, and so this condition is met. To see if the Nearly normal residuals condition is met or not, a residuals distribution should be created. It is easier to see what's going on with a 2D univariate model than a 3D multivariate model since a 3D multivariate model requires to be seen from various angles.

Exercise 4

```
continuous_df <- car_prices %>%
  add_predictions(continuous_model) %>%
  add_residuals(continuous_model)
```

Exercise 5

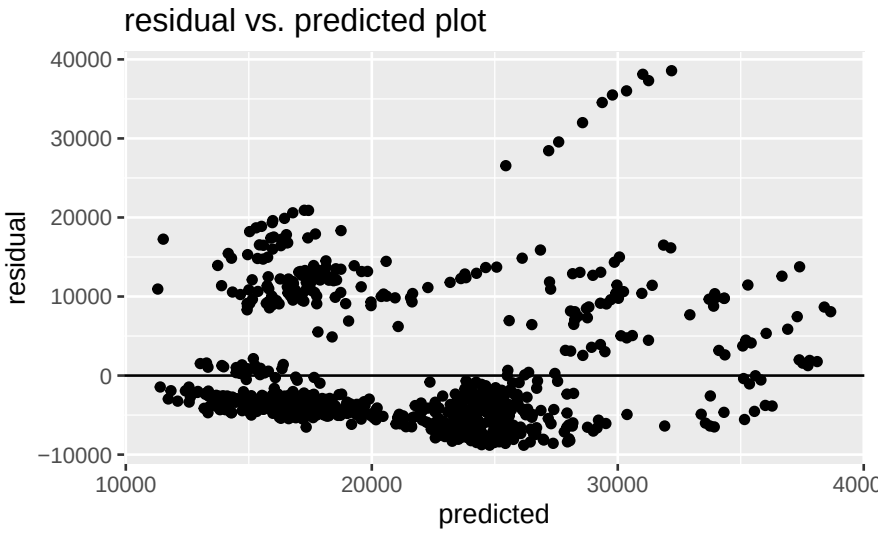
```
continuous_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = Price)) +
  geom_abline(slope = continuous_model$coefficients[2],
              intercept = continuous_model$coefficients[1]) +
  labs(title = "observed vs. predicted plot",
       x = "predicted",
       y = "observed")
```



- There is not a obvious curve in the graph so this is a linear relationship. Hence, the linear model's assumption of linearity is met.

Exercise 6

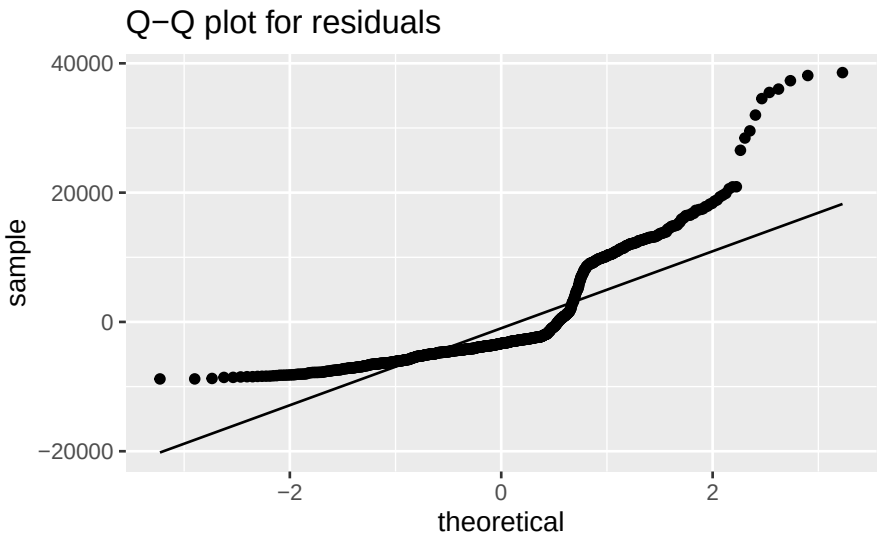
```
continuous_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(yintercept = 0) +
  labs(title = "residual vs. predicted plot",
       x = "predicted",
       y = "residual")
```



- It looks like the variability is reasonably constant all the way along the line except for a few outliers. This means that this model meets the 3rd condition of Constant Variability.

Exercise 7

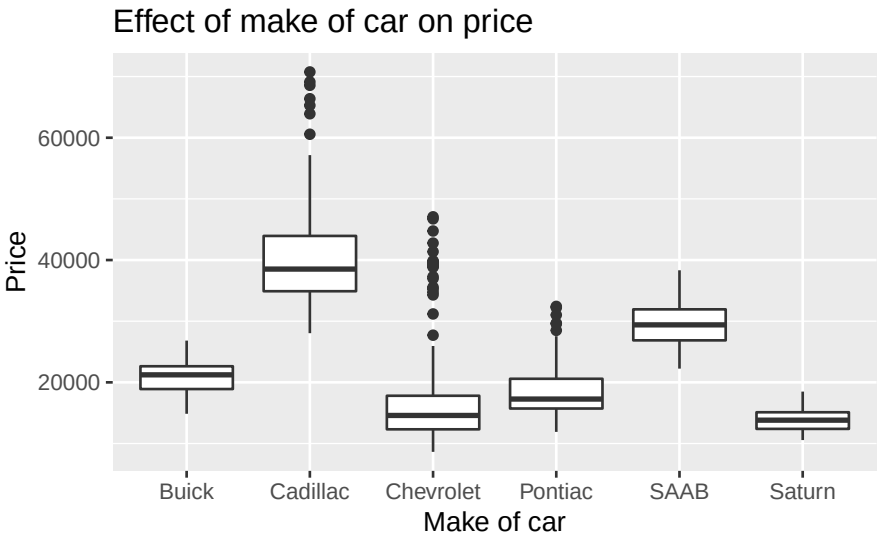
```
continuous_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid)) +
  labs(title = "Q-Q plot for residuals")
```



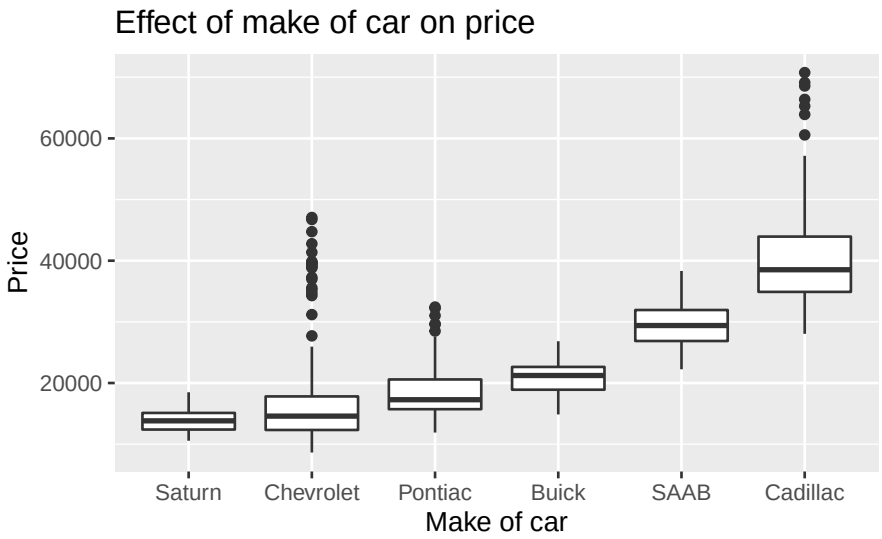
- The nearly-normal residuals condition appears to be violated since the data points are constantly deviating from the line.

Exercise 8

```
car_prices %>%
  ggplot() +
  geom_boxplot(aes(x = Make, y = Price)) +
  labs(x = "Make of car", title = "Effect of make of car on price")
```




```
car_prices %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(Make, Price, FUN=median), y = Price)) +
  labs(x = "Make of car", title = "Effect of make of car on price")
```



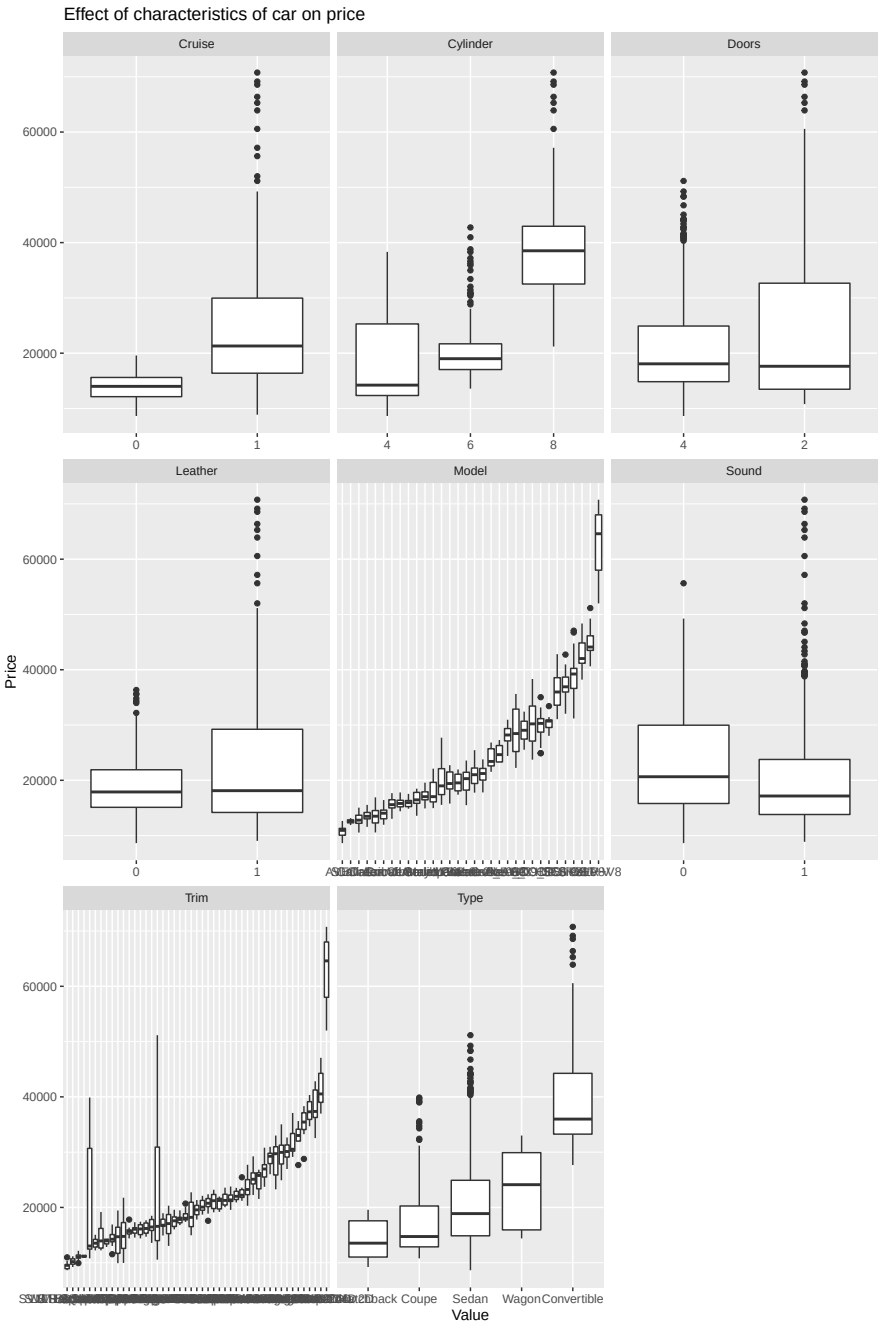
i. Saturn has the lowest median price.

ii. Cadillac has the greatest interquartile range of prices.

iii. Chevrolet, Pontiac, and Cadillac have outliers.

Exercise 9

```
car_prices %>%
  pivot_longer(
    cols = -c('Price':'Make','Liter'),
    names_to="name",
    values_to="value",
    values_transform = list(value = 'factor')
    ) %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(value, Price, FUN=median), y = Price)) +
  facet_wrap(~name, scales = "free_x") +
  labs(x = "Value", title = "Effect of characteristics of car on price")
```



Exercise 10

Exercise 11

Exercise 12