DSCC/CSC/TCS 462 Assignment 1_Ruoqiao Wang

Due Thursday, September 22, 2022 by 4:00 p.m.

```
library(ggplot2)
car_sales <- read.csv("car_sales.csv")
```

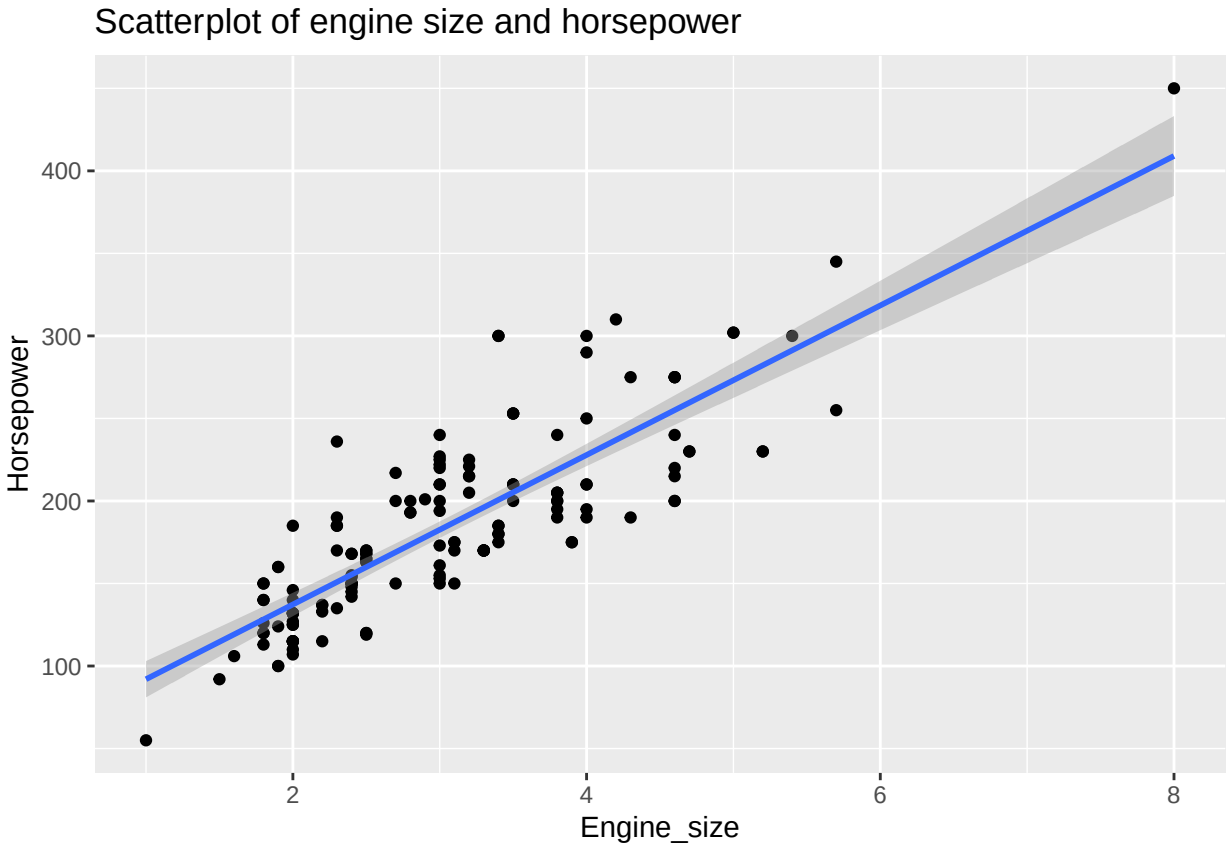
This assignment will cover material from Lectures 3, 4, and 5.

1. For the first part of this assignment, we will explore the relationships between variables using the same "car_sales.csv" dataset as HW0. In particular, we will explore the relationships between multiple variables.

    a. Plot horsepower (y axis) against engine size (x axis). Make sure to label your axes. Comment on the form, strength, and direction of the plot. Note if there are any potential outliers.

```
dat1 <- data.frame(Horsepower = car_sales$Horsepower,
    Engine_size = car_sales$Engine_size)
ggplot(data = dat1, aes(x = Engine_size, y = Horsepower)) +
    geom_point() + stat_smooth(method = "lm") +
    labs(title = "Scatterplot of engine size and horsepower")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Form:linear; Strength:strong;
# Direction:positive; There is no potential
# outlier.
```

b. Calculate the correlation between horsepower and engine size. Comment on this value in relation to your scatterplot

```
cor.test(car_sales$Horsepower, car_sales$Engine_size,
    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  car_sales$Horsepower and car_sales$Engine_size
## t = 18.707, df = 150, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7815469 0.8787984
## sample estimates:
##       cor
## 0.8366494
```

```
# Pearson's Correlation Coefficient =
# 0.8366494; Strong, positive, linear
# relationship between horsepower and
# engine_size with p-value of the test is <<
# 0.05;
```

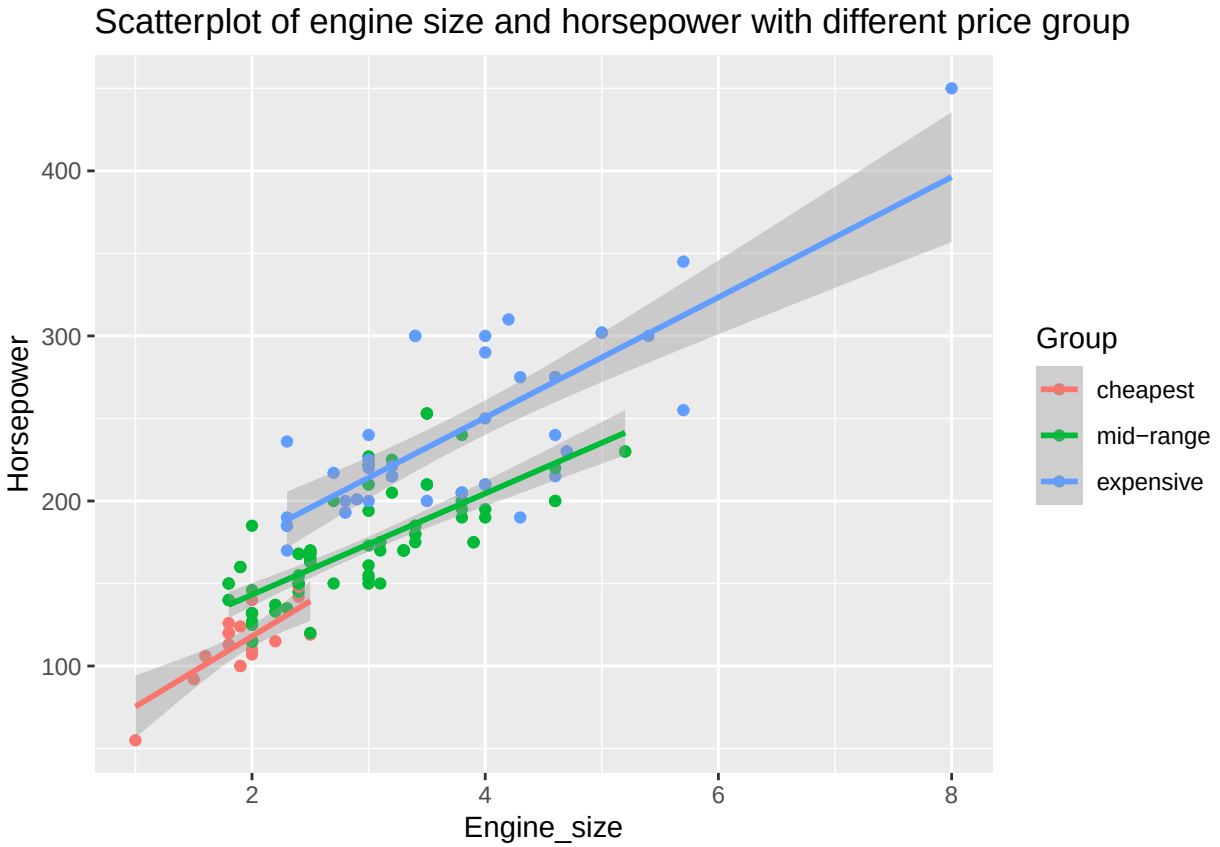
c. Let's break down prices into three groups: the cheapest cars being between 0 and $15000, and mid-range cars being between $15000 and $30000, and the expensive cars costing over $30000. You can use sample code such as this to break price into these three categories.

```
car_sales$new_var <- cut(car_sales$price, breaks = c(0,
    15000, 30000, 90000), labels = c("cheapest",
    "mid-range", "expensive"))
```

d. Plot total horsepower (y axis) against engine size (x axis), but now color points based on which price group they fall into. You can do this by specifying the `col=new_var` option in the `plot()` function. Comment on the results.

```
dat2 <- data.frame(Horsepower = car_sales$Horsepower,
    Engine_size = car_sales$Engine_size, Group = car_sales$new_var)
ggplot(data = dat2, aes(x = Engine_size, y = Horsepower,
    color = Group)) + geom_point() + stat_smooth(method = "lm") +
    labs(title = "Scatterplot of engine size and horsepower with different price group")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



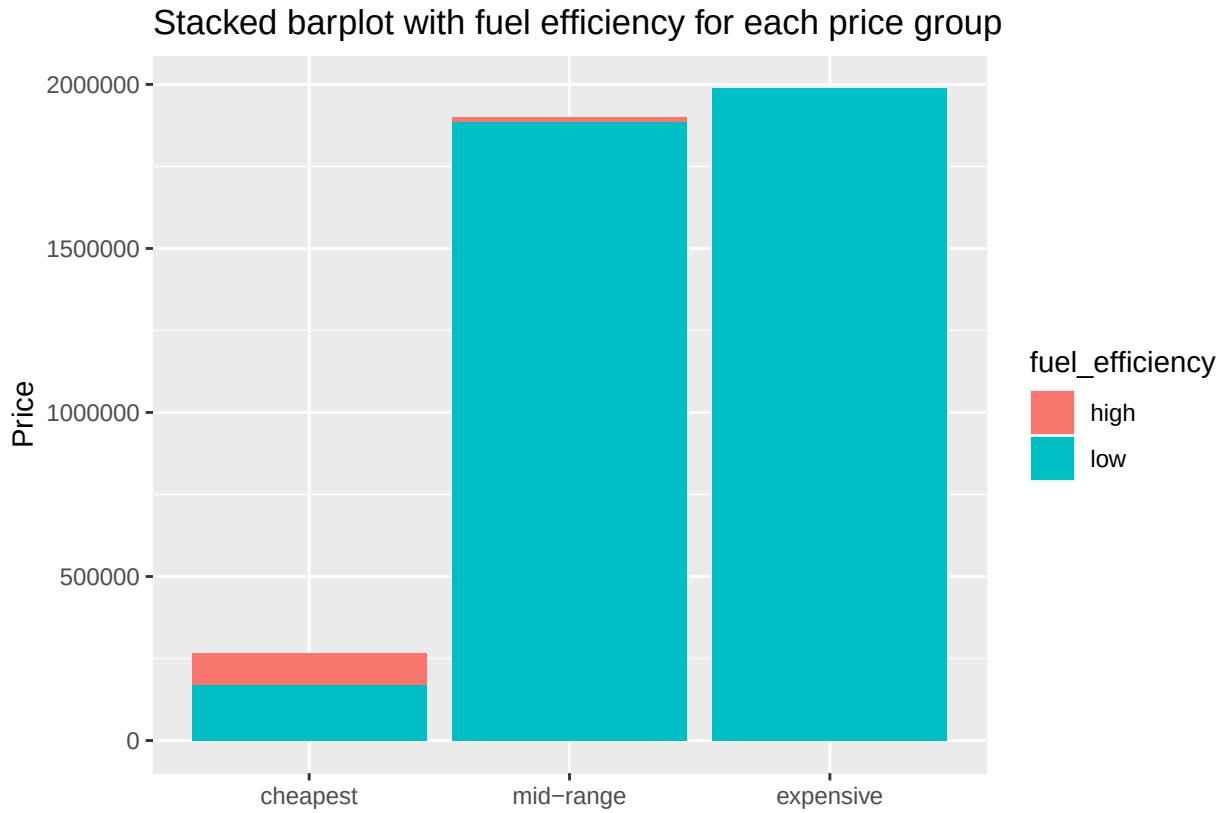

```
# Form:all the groups are linear;
# Strength:all the groups are strong;
# Direction:all the groups are positive;
# There is no potential outlier for all the
# group; The cheapest car group has smaller
# engize size and lower horsepower while
# expensive car group tends has larger
# engize size and higher horsepower.
```

e. Create a new categorical variable that indicates whether the fuel efficiency is greater than 30. Use the following example code as a template:

```
car_sales$fuel <- ifelse(car_sales$Fuel_efficiency >
    30, "high", "low")
```

f. Create a stacked barplot with a bar for each price group (i.e. use `new_var` from above). Each bar should be broken up into two pieces: one for high fuel efficiency and one for low fuel efficiency. Make sure to label your axes and add a legend. Comment on the results.

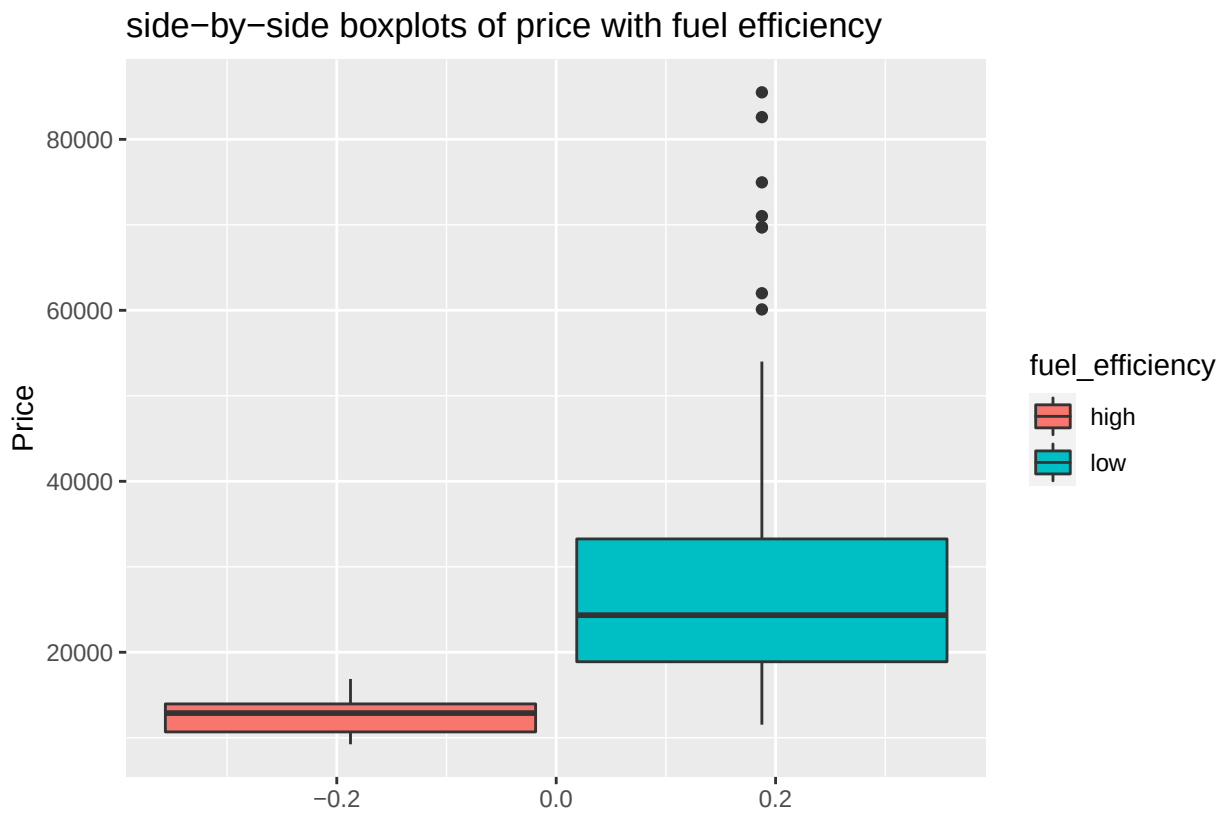
```
dat3 <- data.frame(Price = car_sales$price, Group = car_sales$new_var,
    fuel_efficiency = car_sales$fuel)
ggplot(dat3, aes(x = Group, y = Price, fill = fuel_efficiency)) +
    geom_bar(position = "stack", stat = "identity") +
    ggtitle("Stacked barplot with fuel efficiency for each price group") +
    xlab("")
```




```
# The frequency of high fuel efficiency in
# cheapest car group is higher than other
# two groups, and the expensive car group
# has all low fuel efficiency.
```

g. Make side-by-side boxplots of `price` (not price groups), broken down by fuel efficiency group (low vs. high). Comment on the result:

```
dat4 <- data.frame(Price = car_sales$price, fuel_efficiency = car_sales$fuel)
ggplot(dat4, aes(y = Price, fill = fuel_efficiency)) +
    geom_boxplot() + ggtitle("side-by-side boxplots of price with fuel efficiency") +
    xlab("")
```




```
# The high fuel efficiency group has lower
# price; The low fuel efficiency group has
# higher price.
```

2. Probability: PPV and NPV. A test is created to help detect a disease The test is administered to a group of 84 subjects known to have the disease. Of this group, 59 test positive. The test is also administered to a group of 428 subjects known to not have the disease. Of this group, 12 test positive.

   a. Present this data in a tabular form similar to the following:

Test	Have disease	Do not have disease	Total
Positive	59	12	71
Negative	25	416	441
Total	84	428	512

   b. Calculate the sensitivity and specificity of this test directly from the data.

```
# sensitivity = 0.702381
59/84
```

```
## [1] 0.702381
```

```
# specificity = 0.9719626
416/428
```

```
## [1] 0.9719626
```

c. Assume that the prevalence of the disease is 2.7%. Calculate the NPV and PPV with this prevalence.

```
# NPV=0.9915747
(0.9719626 * (1 - 2.7/100))/(0.9719626 * (1 -
    2.7/100) + (1 - 0.702381) * 2.7/100)
```

```
## [1] 0.9915747
```

```
# PPV=0.4100858
(2.7/100 * 0.702381)/((2.7/100 * 0.702381) + ((1 -
    0.9719626) * (1 - 2.7/100)))
```

```
## [1] 0.4100858
```

d. What conclusions can be drawn regarding the effectiveness of this test?

```
# The effectiveness of test has a high NPV,
# however the PPV is low (<50%), so you may
# want to have more than one test from a
# different brand.
```

3. Probability: Widget production. Consider a factory that produces widgets. These widgets can have one (or more) of three different types: A, B, and C. Suppose that 20% of these widgets have type A, 40% have type B, 10% have both type A and B, and 50% have type C. Any widget of type C only has one type (i.e., there are no widgets of types A and C, B and C, or A, B, and C). Widgets can either be defective (D) or functional (D^c). Denote by $\Pr(D|X)$ the probability that a widget that has type X is defective. The factory knows that $\Pr(D|A) = 0.25$, $\Pr(D|B) = 0.6$, $\Pr(D|A \cap B) = 0.4$, and $\Pr(D|C) = 0.2$.

    a. What is the probability that a widget is defective, $\Pr(D)$? (Hint: Recall the Law of Total Probability.) $\Pr(D) = 0.35$

```
0.25 * 0.2 + 0.6 * 0.4 + 0.2 * 0.5 - 0.4 * 0.1
```

```
## [1] 0.35
```

b. What is the probability that a defective widget is of type B, or $\Pr(B|D)$?
$\Pr(B|D)$ = 0.6857143

```
(0.6 * 0.4)/0.35
```

```
## [1] 0.6857143
```

c. What is the probability that a non-defective (i.e., functional) widget is either type A or type B (or both), i.e., what is $\Pr(A \cup B|D^c)$? $\Pr(A \cup B|D^c) = 0.3846154$

```
0.5 * (1 - (0.25 * 0.2 + 0.6 * 0.4 - 0.4 * 0.1)/(0.2 +
    0.4 - 0.1))/(1 - 0.35)
```

```
## [1] 0.3846154
```

4. Probability: Inclusion-exclusion. Recall that the additive rule tells us for events A and B that are not mutually exclusive that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can extend this additive rule to more than two events, which gives us the general inclusion-exclusion identity as follows:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + ... + (-1)^{n+1} P(A_1 \cap A_2 \cap ... \cap A_n)$$

a. Explicitly write the inclusion-exclusion identity for $n = 3$ events, A_1, A_2, A_3 (i.e., reduce down so that there aren't summations).

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$$

b. Suppose an integer from 1 to 1000 (inclusive) is chosen uniformly at random (i.e., with equal probability). What is the probability that the integer is divisible by 5, 7, or 13?

```
floor(1000/5)/1000 + floor(1000/7)/1000 + floor(1000/13)/1000 -
    floor(1000/5/7)/1000 - floor(1000/7/13)/1000 -
    floor(1000/5/13)/1000 + floor(1000/5/7/13)/1000
```

```
## [1] 0.367
```

5. Combinatorics: Consider a political setting where there are three political parties, A, B, and C vying for seats on a 3-person committee. Party A has 2 members, B has 3 members, and C has 5 members. Members of parties are distinguishable from each other, but positions on the committee are indistinguishable from each other.

a. How many ways are there of forming an unordered 3-person committee?

```
choose(10, 3)
```

```
## [1] 120
```

b. How many different party breakdowns (e.g., ABC, CCC, etc.) are possible when form

```
1 + 1 + choose(2, 1) + choose(3, 1) + choose(2,
    1)
```

```
## [1] 9
```

c. How many ways are there of forming an unordered 3-person committee if at least one me

```
choose(2, 1) * choose(8, 2) + choose(2, 2) * choose(8,
    1)
```

```
## [1] 64
```

6. Combinatorics: Miscellaneous counting.

    a. There are 20 indistinguishable children who would like to have one ice cream cone each. There are 6 distinct flavors of ice cream. How many distinct collections of ice cream cones are there where at least two children must order each flavor?

```
choose(8 + 6 - 1, 6 - 1)
```

```
## [1] 1287
```

b. There are five cats and five dogs, all distinguishable from one another. How many dis

```
factorial(5) * factorial(5)/5
```

```
## [1] 2880
```

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable? 2 days; and it took longer time than I expect.
- Who, if anyone, did you work with on this assignment? Alone.
- What questions do you have relating to any of the material we have covered so far in class?
  NA.