DSCC/CSC/TCS 462 Assignment 1

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

Daxiang Na

09.22.2022

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

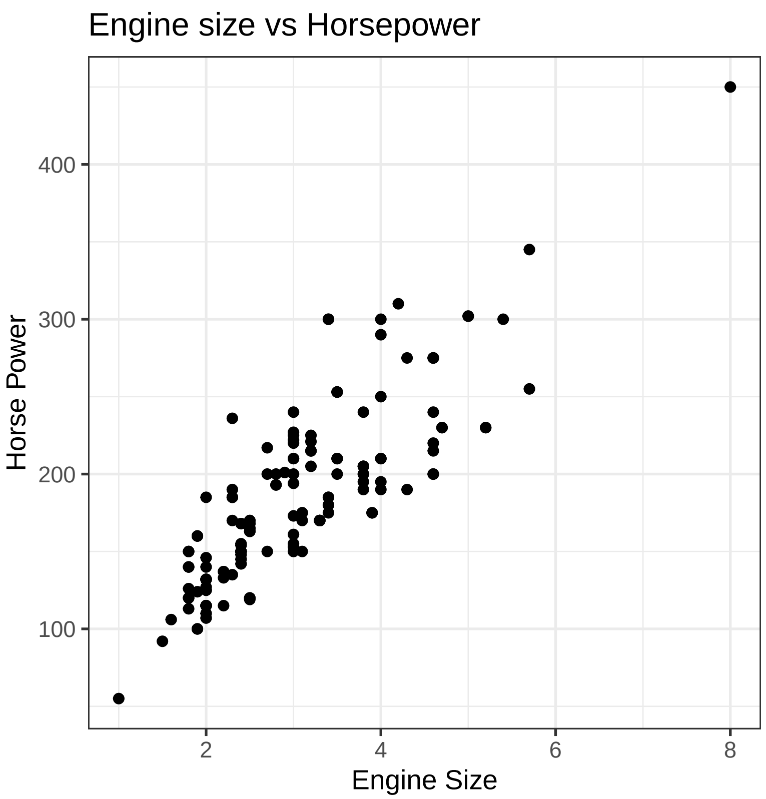
Feedback from TA: 1)a) comment about form,strength and direction is missing, 2) d) the test is effective as in case of disease False negative is important. 6) a) wrong

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.

```
p <- ggplot(car_sales, aes(x = Engine_size,
    y = Horsepower))
p + geom_point() + theme_bw() + theme(aspect.ratio = 1) +
    labs(title = "Engine size vs Horsepower",
        x = "Engine Size", y = "Horse Power")
```



There is a potential outlier in the upright corner.

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

```
r <- cor(x = car_sales$Engine_size, y = car_sales$Horsepower,
    method = "pearson")
print(r)
```

```
## [1] 0.8366494
```

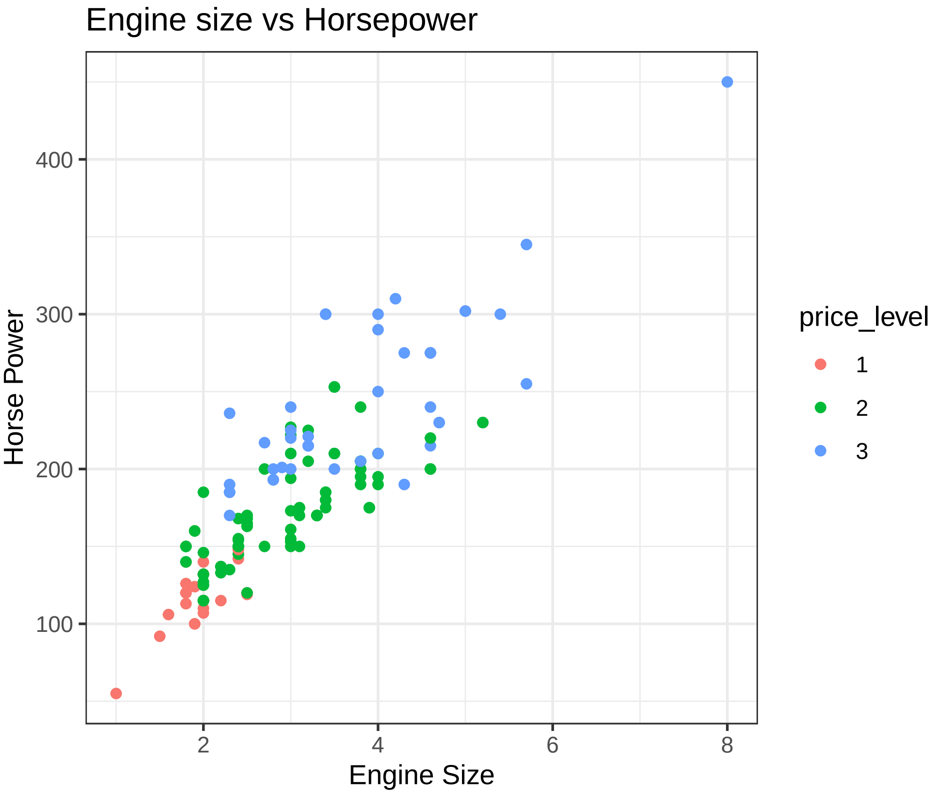
The correlation coefficient between horsepower and engine size is 0.84, indicating that those two variables are highly correlated, and it is possible due to the outlier in the upright corner.

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$price_level <- cut(car_sales$price,
    breaks = c(0, 15000, 30000, 1e+05), labels = c(1,
        2, 3))
```

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.

```
p <- ggplot(car_sales, aes(x = Engine_size,
    y = Horsepower))
p + geom_point(aes(color = price_level)) +
    theme_bw() + theme(aspect.ratio = 1) +
    labs(title = "Engine size vs Horsepower",
        x = "Engine Size", y = "Horse Power")
```



This plot shows that cheaper cars tend to have smaller engize size and lower horsepower, and expensive cars tend to have bigger engize size and high horsepower.

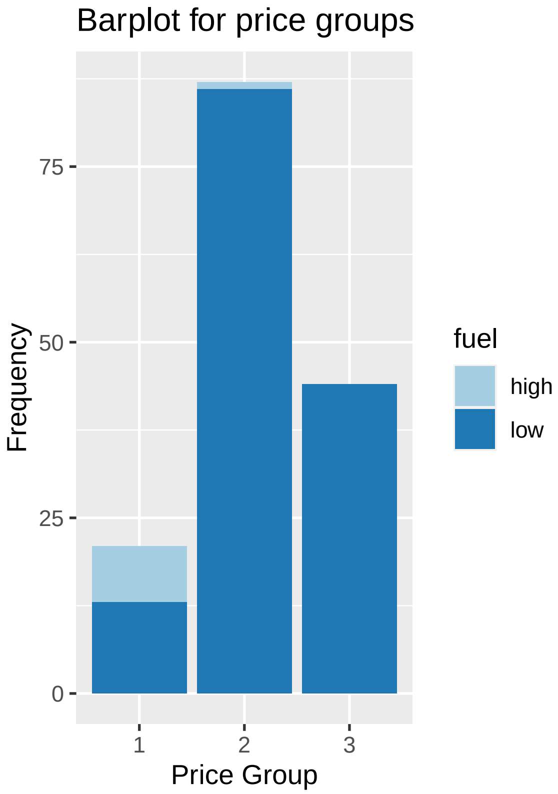
e. Create a new categorical variable that indicates whether the fuel efficiency is greater than 30. Use the following example code as a template:

```
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.

```
p <- ggplot(car_sales, aes(x = price_level,
    fill = fuel))
```

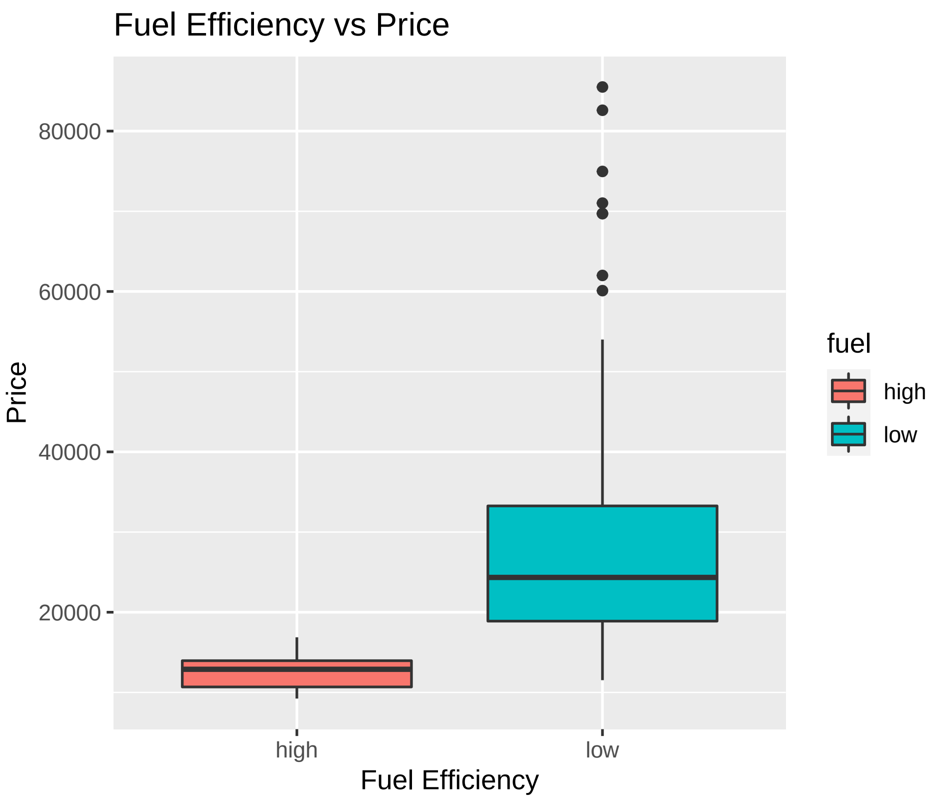
```r
p + geom_bar() + scale_fill_brewer(palette = "Paired") +
    theme(aspect.ratio = 2) + labs(title = "Barplot for price groups",
    x = "Price Group", y = "Frequency")
```



Cars with high prices tend to show low fuel efficiency, and cars with low prices may have high 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:

```r
p <- ggplot(car_sales, aes(y = price, x = fuel))
p + geom_boxplot(aes(fill = fuel)) + theme(aspect.ratio = 1) +
    labs(title = "Fuel Efficiency vs Price",
        x = "Fuel Efficiency", y = "Price")
```



This plot shows that cars with high fuel efficiency show low prices, but cars with low fuel efficiency show higher average price, and the variation is much larger.

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.

```
sens = 59/84
spec = 416/428
print(paste0("sensitivity is ", sens))
```

```
## [1] "sensitivity is 0.702380952380952"
```

```
print(paste0("specificity is ", spec))
```

```
## [1] "specificity is 0.97196261682243"
```

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

```
# NPV: negative predictive value =
# P(D2|T-) PPV: positive predictive
# value = P(D1|T+)
prev = 0.027
nonp = 1 - prev
PPV = (sens * prev)/((sens * prev) + (1 -
    spec) * nonp)
NPV = (spec * nonp)/((spec * nonp) + (1 -
    sens) * prev)
print(paste0("PPV is ", PPV))
```

```
## [1] "PPV is 0.410085962367105"
```

```
print(paste0("NPV is ", NPV))
```

```
## [1] "NPV is 0.991574658673062"
```

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

This test is not very effective, its sensitiviy is pretty low and its PPV is even lower than 50%, which means that when test yields positive result, the chance the patient actually has the disease is lower than 50%. However, its NPV is pretty high, which means that when it gives negative result, it is very likely that this patient does not have trouble.

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.)

```
A = 0.2
B = 0.4
AB = 0.1
C = 0.5
DA = 0.25
DB = 0.6
D_AB = 0.4
DC = 0.2
D = (DA * A + DB * B - D_AB * AB) + DC *
    C
```

```r
print(paste0("Pr(D) is ", D))
```

```
## [1] "Pr(D) is 0.35"
```

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

```r
BD = (DB * B)/D
print(paste0("Pr(B|D) is ", BD))
```

```
## [1] "Pr(B|D) is 0.685714285714286"
```

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 $\Pr(A \cup B|D^c)$?

```r
A_AND_B = 1 - C
C_Dc = ((1 - DC) * C)/(1 - D)
A_AND_B_Dc = 1 - C_Dc
print(paste0("Pr(A and B|Dc) is ", A_AND_B_Dc))
```

```
## [1] "Pr(A and B|Dc) is 0.384615384615385"
```

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).

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

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?

```r
PA1 = 0.2
PA2 = 0.142
PA3 = 0.076
PA1A2 = 0.028
PA1A3 = 0.015
PA2A3 = 0.01
PA1A2A3 = 0.002
P = PA1 + PA2 + PA3 - PA1A2 - PA1A3 - PA2A3 +
    PA1A2A3
print(paste0("The probability that the integer is divisible by 5, 7, or 13 is ",
    P))
```

```
## [1] "The probability that the integer is divisible by 5, 7, or 13 is 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?

```
a <- choose(2 + 3 + 5, 3)
print(paste("There are", a, "ways of forming an unordered 3-person committee."))
```

```
## [1] "There are 120 ways of forming an unordered 3-person committee."
```

   b. How many different party breakdowns (e.g., ABC, CCC, etc.) are possible when forming an unordered 3-person comittee?

0A: BBB, BBC, BCC, CCC

1A: ABB, ABC, ACC

2A: AAB, AAC

Answer: 9 breakdowns in totals.

   c. How many ways are there of forming an unordered 3-person committee if at least one member must be from party A?

```
ABB = choose(2, 1) * choose(3, 2)
ABC = choose(2, 1) * choose(3, 1) * choose(5,
    1)
ACC = choose(2, 1) * choose(5, 2)
AAB = choose(3, 1)
AAC = choose(5, 1)
total = ABB + ABC + ACC + AAB + AAC
print(paste("There are", total, "ways of forming an unordered 3-person committee if at lea
```

```
## [1] "There are 64 ways of forming an unordered 3-person committee if at least one membe
```

There are 64 ways of forming an unordered 3-person committee if at least one member must be from party A.

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?

The following is the wrong way: “least_one” calculates the possibilities without specifying any bin. I will just discard this direction.

```
least_one <- 6 * choose(19 - 1, 5 - 1)  # collection number that at least one bin has one
least_two <- choose(20 - 1, 6 - 1) - least_one  # collection number that none of bins has
print(paste("There are", least_two, "distinct collections of ice cream cones are there wh
```

```
## [1] "There are -6732 distinct collections of ice cream cones are there where at least t
```

The right way: consider each bin/flavor has got 2 children, then what you are going to do is to select the rest of them (20 - 2*6 = 8) for those 6 flavors, and it is possible that there is one flavor that none of them would choose, then the possibilities will be "choose(8 + 6 - 1, 6 - 1)".

```r
least_two <- choose(8 + 6 - 1, 6 - 1)
print(least_two)
```

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

Answer: There are 1287 distinct collections of ice cream cones are there where at least two children must order each flavor.

b. There are five cats and five dogs, all distinguishable from one another. How many distinct ways are there of seating them at a round table such that every cat is adjacent to two dogs and every dog is adjacent to two cats? Note that here two orderings are not considered distinct if it is possible to rotate one and achieve the other. For instance, if there are only four seats at the table, the order Cat 1 - Dog 1 - Cat 2 - Dog 2 is the same as Cat 2 - Dog 2 - Cat 1 - Dog 1.

```r
n = 1 * 5 * 4 * 4 * 3 * 3 * 2 * 2 * 1 * 1
print(paste("There are", n, "ways for seating them."))
```

```
## [1] "There are 2880 ways for seating them."
```

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?

It took me about 5 to 6 hours to finish it. I feel it was a bit long.

- Who, if anyone, did you work with on this assignment?

Myself

- What questions do you have relating to any of the material we have covered so far in class?

Independence vs mutually exclusive. How can two events have non-empty intersection but P(A|B) = P(A)?