DSCC/CSC/TCS 462 Assignment 1

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

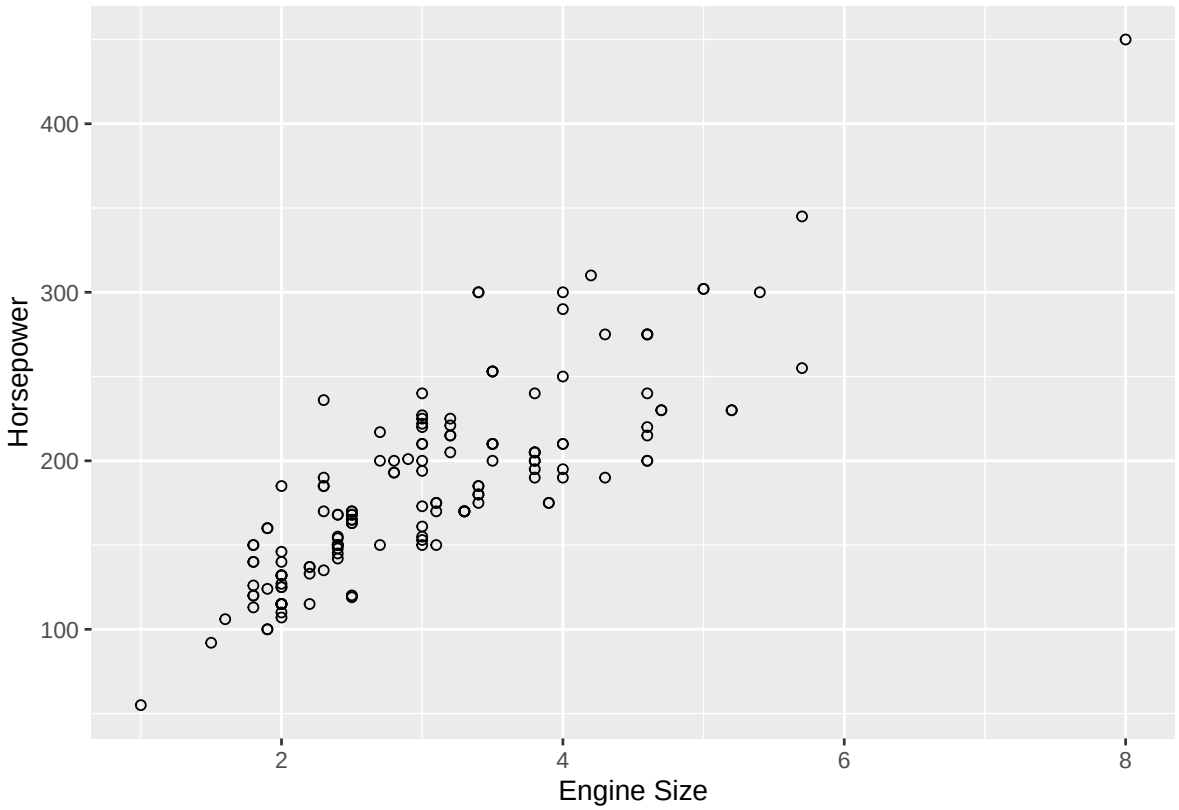
```
library(ggplot2)
car_sales <- read.csv("car_sales.csv")
```

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.

```
plot1 <- ggplot(car_sales, aes(x = Engine_size, y = Horsepower)) +
    geom_point(shape = 1)
plot1 <- plot1 + xlab("Engine Size") + ylab("Horsepower")
plot1
```



There is a relatively strong, positive, linear relationship, with some potential outliers.

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

```
cor(car_sales$Engine_size, car_sales$Horsepower)
```

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

This value of 0.8366 makes sense for the dataset. The larger the engine, the greater the horsepower.

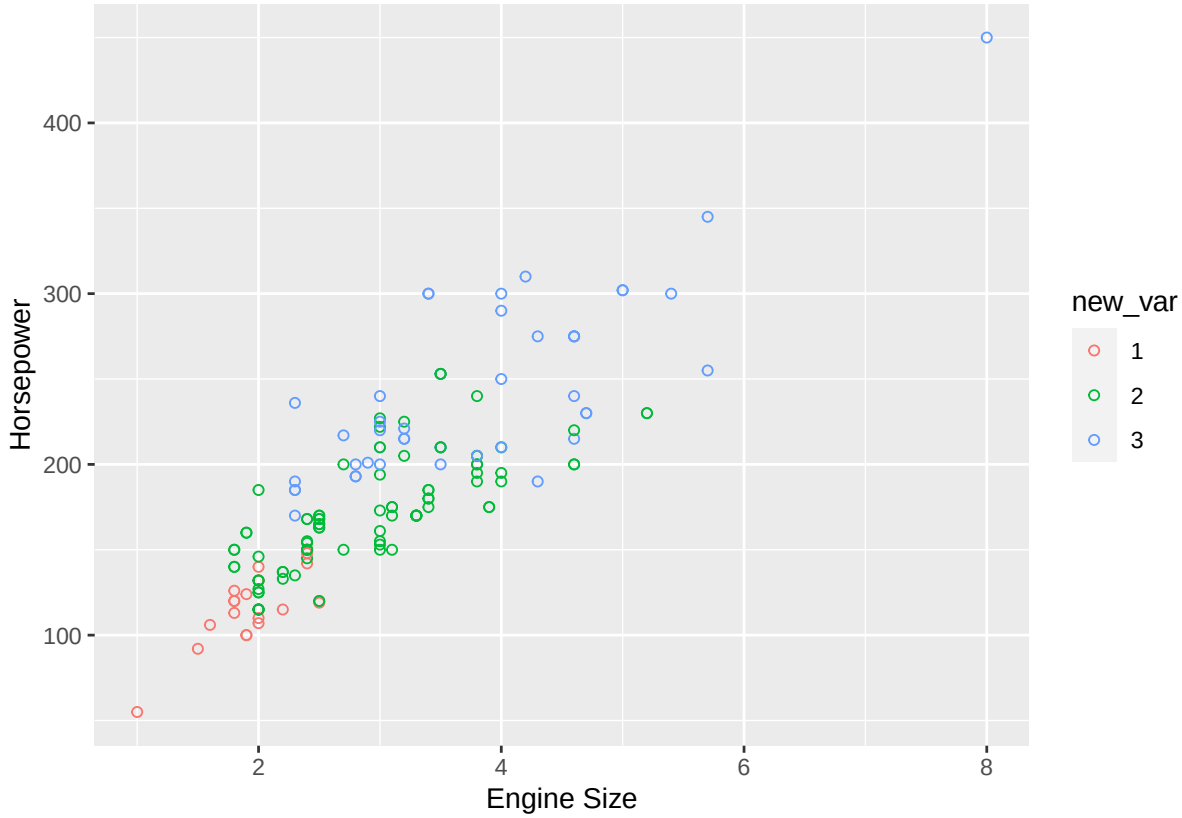
c. Let's break down prices into three groups: the cheapest cars being between 0 and $15000, and mid-range cars being between $15000 and $30000, and the expensive cars costing over $30000. You can use sample code such as this to break price into these three categories.

```
car_sales$new_var <- cut(car_sales$price, breaks = c(0, 15000,
    30000, 90000), labels = c(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.

```
plot1 <- ggplot(car_sales, aes(x = Engine_size, y = Horsepower,
    color = new_var)) + geom_point(shape = 1)
plot1 <- plot1 + xlab("Engine Size") + ylab("Horsepower")
```

```
plot1
```



We can see that the three groups are relatively separated from one another. The most expensive cars tend to have the larger engines and greater horsepower. The cheapest cars are the ones with the slowest engines and lowest horsepower.

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

```
car_sales$fuel <- ifelse(car_sales$Fuel_efficiency > 25, "high",
    "low")
```

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

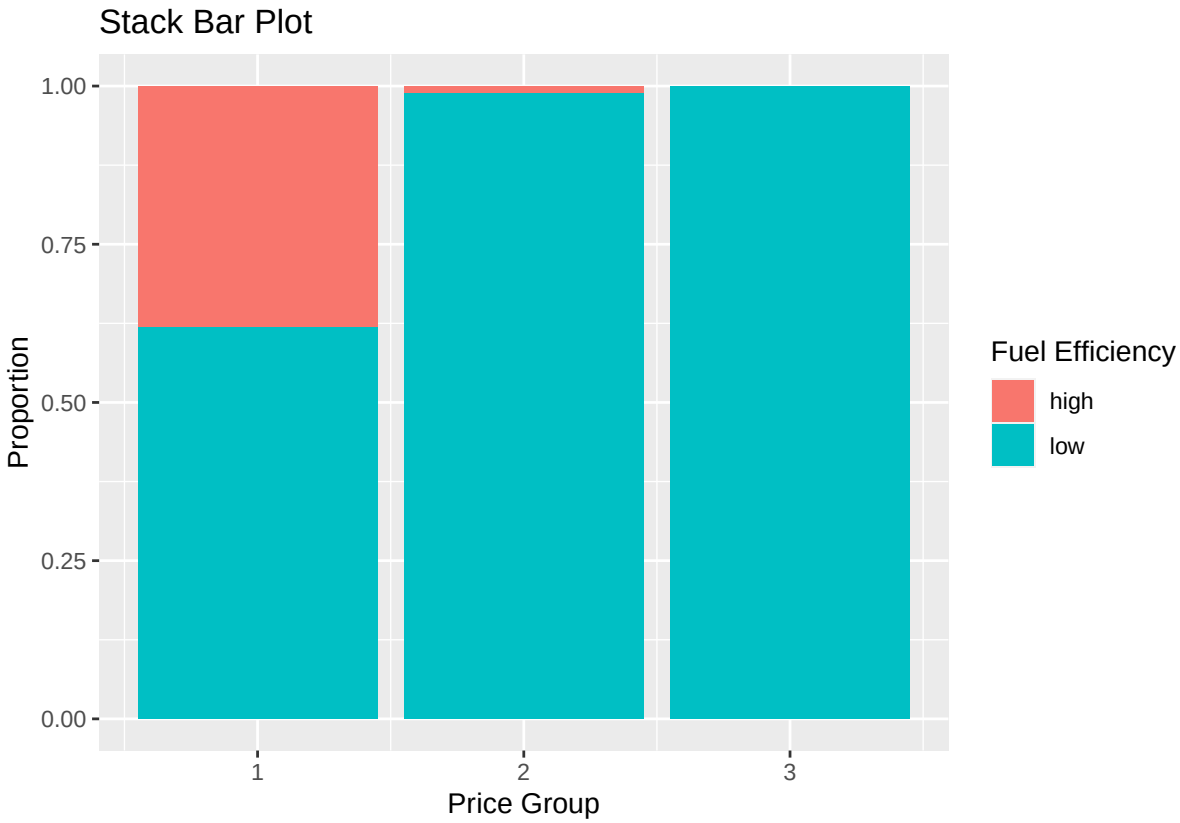
f. Create a stacked barplot with a bar for each price group (i.e. use `new_var` from above). Each bar should be broken up into two pieces: one for high fuel efficiency and one for low fuel efficiency. Make sure to label your axes and add a legend. Comment on the results.

```
library(reshape)
tab1 <- table(car_sales$fuel, car_sales$new_var)
new <- melt(tab1)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```

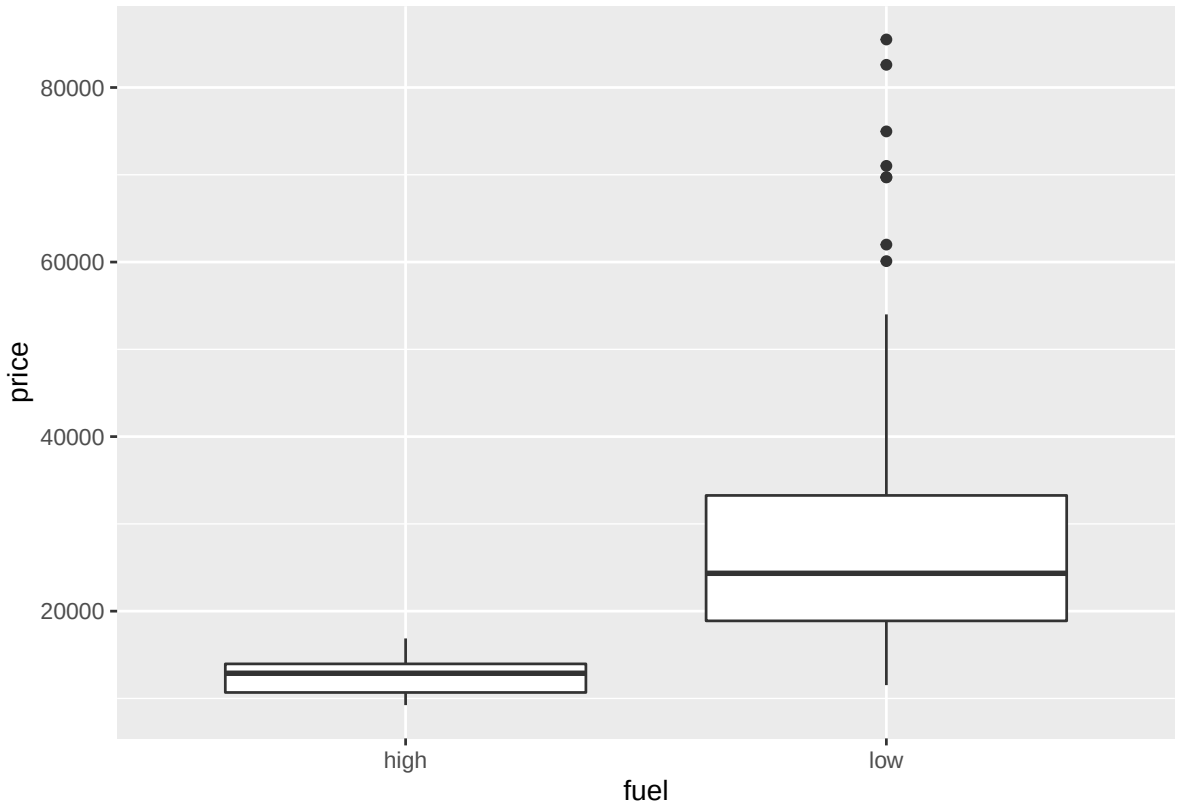
```
ggplot(data = new, aes(x = Var.2, y = value, fill = Var.1)) +
    geom_bar(position = "fill", stat = "identity") + labs(x = "Price Group",
    y = "Proportion", fill = "Fuel Efficiency", title = "Stack Bar Plot")
```



Generally, more expensive cars are likely to have low fuel efficiency, thus costing more in gas as well.

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

```
ggplot(car_sales, aes(x = fuel, y = price)) + geom_boxplot()
```



Cars with high fuel efficiency all tend to be on the lower end of the spectrum in terms of price, and prices tend to be fairly similar among these cars. For cars with low fuel efficiency, there is a wide spread of prices, ranging from very low to very high.

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.

$$sens : P(T^+|D^+) = \frac{59}{84} = 0.702381$$
$$spec : P(T^-|D-) = \frac{416}{428} = 0.97196$$

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

with this prevalence.

$$\begin{aligned} NPV : P(D^-|T^-) &= \frac{P(T^-|D^-)P(D^-)}{P(T^-|D^-)P(D^-) + P(T^-|D^+)P(D^+)} \\ &= \frac{0.97196 \times 0.973}{0.97196 \times 0.973 + (1 - 0.702381) \times 0.027} = 0.9803 \\ PPV : P(D^+|T^+) &= \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \\ &= \frac{0.702381 \times 0.027}{0.702381 \times 0.027 + (1 - 0.97196) \times 0.973} = 0.4101 \end{aligned}$$

d. What conclusions can be drawn regarding the effectiveness of this test?
Overall, the test is quite effective at correctly detecting when someone does not have the disease, but since the disease is so rare and the sensitivity of the test is not greater, the PPV is low.

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, $\Pr(D)$?

From the problem statement, we can see that because 10% of the widgets are type A and B, that means that 10% of the widgets must be only type A (denoted $A \setminus B$) and 30% of the widgets must be only type B (denoted $B \setminus A$). This lets us determine $\Pr(D|A \setminus B)$ and $\Pr(D|B \setminus A)$: using the fact that $\Pr(D|A \cap B) = 0.4$, we have that $\Pr(D|A \setminus B) = 0.1$ and $\Pr(D|B \setminus A) = 2/3$.

By the law of total probability, we have

$$\begin{aligned} \Pr(D) &= \Pr(D|A \setminus B)\Pr(A \setminus B) + \Pr(D|B \setminus A)\Pr(B \setminus A) \\ &\quad + \Pr(D|A \cap B)\Pr(A \cap B) + \Pr(D|C)\Pr(C) \\ &= 0.1 \cdot 0.1 + (2/3) \cdot 0.3 + 0.4 \cdot 0.1 + 0.2 \cdot 0.5 \\ &= 0.01 + 0.2 + 0.04 + 0.1 = 0.35. \end{aligned}$$

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

Use Bayes' Rule:

$$\begin{aligned}\Pr(B|D) &= \frac{\Pr(D|B)\Pr(B)}{\Pr(D)} \\ &= \frac{0.6 \cdot 0.4}{0.35} = \frac{0.24}{0.35} \approx 0.686.\end{aligned}$$

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

Solution 1: Use $\Pr(A \cup B|D^c) = 1 - \Pr(C|D^c)$ and Bayes' Rule.

$$\begin{aligned}\Pr(A \cup B|D^c) &= 1 - \Pr(C|D^c) \\ &= 1 - \frac{\Pr(D^c|C)\Pr(C)}{\Pr(D^c)} \\ &= 1 - \frac{0.8 \cdot 0.5}{0.65} = \frac{0.25}{0.65} \approx 0.385.\end{aligned}$$

Solution 2: Do it directly using Bayes' Rule.

$$\begin{aligned}\Pr(A \cup B|D^c) &= \frac{\Pr(D^c|A \cup B)\Pr(A \cup B)}{\Pr(D^c)} \\ &= \frac{\frac{\Pr(D^c|A\setminus B)\Pr(A\setminus B)+\Pr(D^c|B\setminus A)\Pr(B\setminus A)+\Pr(D^c|A\cap B)\Pr(A\cap B)}{\Pr(A\cup B)} \cdot \Pr(A \cup B)}{\Pr(D^c)} \\ &= \frac{0.9 \cdot 0.1 + (1/3) \cdot 0.3 + 0.6 \cdot 0.1}{0.65} = \frac{0.25}{0.65} \approx 0.385.\end{aligned}$$

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:

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

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?

Let D_x be the event that the integer is divisible by x. We have $\Pr(D_5) = 200/1000 = 1/5$, $\Pr(D_7) = 142/1000$, and $\Pr(D_{13}) = 76/1000$. We also have that $\Pr(D_5 \cap D_7) = \Pr(D_{35}) = 28/1000$, $\Pr(D_5 \cap D_{13}) = \Pr(D_{65}) = 15/1000$, and $\Pr(D_7 \cap D_{13}) = \Pr(D_{91}) = 10/1000$. Lastly, we have that $\Pr(D_5 \cap D_7 \cap D_{13}) = \Pr(D_{455}) = 2/1000$. Putting it all together with the principle of inclusion-exclusion, we have

$$\begin{aligned}\Pr(D_5 \cup D_7 \cup D_{13}) &= \Pr(D_5) + \Pr(D_7) + \Pr(D_{13}) - \Pr(D_5 \cap D_7) - \Pr(D_5 \cap D_{13}) \\ &\quad - \Pr(D_7 \cap D_{13}) + \Pr(D_5 \cap D_7 \cap D_{13}) \\ &= \frac{1}{1000}(200 + 142 + 76 - 28 - 15 - 10 + 2) = \frac{367}{1000}.\end{aligned}$$

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?

$\binom{10}{3} = 120$.

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

9: AAB, AAC, ABB, ABC, ACC, BBB, BBC, BCC, CCC.

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

We can enumerate the possibilities. $AAB : \binom{2}{1} \cdot \binom{3}{1} = 3$, $AAC : \binom{2}{1} \cdot \binom{5}{1} = 5$, $ABB : \binom{2}{1} \cdot \binom{3}{2} = 6$, $ABC : \binom{2}{1} \cdot \binom{3}{1} \cdot \binom{5}{1} = 30$, $ACC : \binom{2}{1} \cdot \binom{5}{2} = 20$. Total: 64.

Alternatively, this is just the complement of forming an unordered 3-person committee with only members from B and C, which is $120 - \binom{8}{3} = 64$.

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?

This is equivalent to fixing 12 childrens' ice cream flavors (2 children per flavor) and then running normal stars and bars for $n = 8$ and $k = 6$. Therefore, we have $\binom{n+k-1}{k-1} = \binom{13}{5} = 1287$ combinations.

   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.

Fix one animal, without loss of generality Cat 1, as the reference point so we don't have to worry about rotations. There are 4! ways of arranging the other dogs and 5! ways of arranging the cats between the dogs. Therefore, there are $5!4! = 2880$ distinct seatings.

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?
- Who, if anyone, did you work with on this assignment?
- What questions do you have relating to any of the material we have covered so far in class?