

STATS 60 Summer 2020: HW2

Conceptual Questions

- 1) Compute the standard deviation of the set $\{3, -1, 4\}$. Do not assume that this is a sample and show your work.

Note that $\mu = \frac{3-1+4}{3} = 2$. Using our usual formula for standard deviation (with a denominator of 3), we have

$$\sigma = \sqrt{\frac{1}{3} \sum (X_i - \bar{X})^2} = \sqrt{\frac{1}{3}(1^2 + 2^2 + (-3)^2)} \approx 2.16$$

- 2) Suppose large company XYZ is a multinational company with over 100,000 employees. This company reported that the average employee makes \$82,000 a year. Do we expect the number of employees at this company that make over \$82,000 to be greater than, about equal, or less than half of the total number of employees?

Note that for a large multinational company, we expect the data to be heavily right skewed. That is, a few top level employees make a lot of money compared to many lower level employees that do not make as much. Hence, we expect the median to be lower than the mean, so the number of employees that make over \$82,000 should be less than half the total number.

- 3) Consider the following histogram of annual family income in a certain city. Suppose the class intervals include the right endpoint but not the left. Suppose the data is uniformly distributed in each bar of the histogram. What percentage of families had an income over \$25,000 but not greater than \$40,000?



Since the data is uniformly distributed per bar, we have

$$0.5 * 30\% + 10\% = 25\%$$

Thus, 25% of the families have such an income. Note if you answered 2.5%, you would get full credit (since the labelling was unclear).

- 4) The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than the former smokers.

a) Why did they study men and women and the different age groups separately?

Sex and age may be confounding variables, so we want to investigate the effect of smoking on health independently of the other factors.

- b) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.

It could be that the former smokers who quit were ones that quit because of poor health. Thus, many former smokers could have had poor health to start with. In this case, we cannot directly draw the conclusion that quitting smoking causes worst health.

For the following questions, refer to the Normal table attached below (do not use R for calculations)

- 5) Suppose we have a classroom in which heights of students are normally distributed. Suppose the average height is 66 inches and the standard deviation is 3 inches.

a) What percentage of students are taller than 70 inches?

Note that with a standard deviation of 3, we get that $70 - 66 = 4$ is about 1.33 standard deviations above the mean, or $z = 1.33$. Checking the table we get between 80.64% and 82.3% of the data in a normal distribution live in within 1.3 to 1.35 standard deviations of the mean. For simplicity, let us linearly interpolate between these numbers and say that

$$80.64 \cdot \frac{2}{5} + 82.3 \cdot \frac{3}{5} \approx 81.6$$

81.6% of the data are within 1.33 standard deviations of the mean. Thus, 18.4% of the data are outside of this range. Since the normal distribution is symmetric, we expect about $\frac{18.4}{2} = 9.2\%$ of the students to have height greater than or equal to 70 inches.

b) What height marks the 30th percentile?

If we want to find the 30th percentile using the following table, it suffices to find the z such that 40% of the data are within $\pm z$ of the mean. This is because by symmetry, 30% of the data should lie below the lower bound of this range. From this table, somewhere about $z \approx 0.52$ gives us the desired coverage. Thus, the lower bound is $66 - 0.52 \cdot 3 = 64.44$, which approximately marks the 30th percentile for height.

- 6) A useful measure of spread is the **Interquartile Range (IQR)**, which measures the distance between the 75th percentile and the 25th percentile (which you may have seen in boxplots). If the data is normally distributed, how many standard deviations wide is the IQR?

Once again, using symmetry, we simply have to find z such that 50% of the data live within $\pm z$. From the table, this seems to be approximately 0.67. Thus, the IQ range is about $0.67 \cdot 2 = 1.34$ standard deviations wide.

As a sanity check, we can run

```
qnorm(0.75) - qnorm(0.25)
```

```
## [1] 1.34898
```

and check that we get a very good approximation using our normal table.



A NORMAL TABLE

<i>z</i>	<i>Height</i>	<i>Area</i>	<i>z</i>	<i>Height</i>	<i>Area</i>	<i>z</i>	<i>Height</i>	<i>Area</i>
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863
0.25	38.67	19.74	1.75	8.63	91.99	3.25	0.203	99.885
0.30	38.14	23.58	1.80	7.90	92.81	3.30	0.172	99.903
0.35	37.52	27.37	1.85	7.21	93.57	3.35	0.146	99.919
0.40	36.83	31.08	1.90	6.56	94.26	3.40	0.123	99.933
0.45	36.05	34.73	1.95	5.96	94.88	3.45	0.104	99.944
0.50	35.21	38.29	2.00	5.40	95.45	3.50	0.087	99.953
0.55	34.29	41.77	2.05	4.88	95.96	3.55	0.073	99.961
0.60	33.32	45.15	2.10	4.40	96.43	3.60	0.061	99.968
0.65	32.30	48.43	2.15	3.96	96.84	3.65	0.051	99.974
0.70	31.23	51.61	2.20	3.55	97.22	3.70	0.042	99.978
0.75	30.11	54.67	2.25	3.17	97.56	3.75	0.035	99.982
0.80	28.97	57.63	2.30	2.83	97.86	3.80	0.029	99.986
0.85	27.80	60.47	2.35	2.52	98.12	3.85	0.024	99.988
0.90	26.61	63.19	2.40	2.24	98.36	3.90	0.020	99.990
0.95	25.41	65.79	2.45	1.98	98.57	3.95	0.016	99.992
1.00	24.20	68.27	2.50	1.75	98.76	4.00	0.013	99.9937
1.05	22.99	70.63	2.55	1.54	98.92	4.05	0.011	99.9949
1.10	21.79	72.87	2.60	1.36	99.07	4.10	0.009	99.9959
1.15	20.59	74.99	2.65	1.19	99.20	4.15	0.007	99.9967
1.20	19.42	76.99	2.70	1.04	99.31	4.20	0.006	99.9973
1.25	18.26	78.87	2.75	0.91	99.40	4.25	0.005	99.9979
1.30	17.14	80.64	2.80	0.79	99.49	4.30	0.004	99.9983
1.35	16.04	82.30	2.85	0.69	99.56	4.35	0.003	99.9986
1.40	14.97	83.85	2.90	0.60	99.63	4.40	0.002	99.9989
1.45	13.94	85.29	2.95	0.51	99.68	4.45	0.002	99.9991

Data Visualization

1)

Call the following to load the `vehicles` data set. If you do not have `fueleconomy` installed, run `install.packages` on it first:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(fueleconomy)
```

```
## Warning: package 'fueleconomy' was built under R version 3.6.3
```

```
data(vehicles)
```

- a) What are the dimensions of the `vehicles` dataset? What are the column names? How many of each value are there in the `cyl` column? What is the mean, median and standard deviation of the `cty` column?

The following code blocks give all the desired attributes.

```
dim(vehicles)
```

```
## [1] 33442    12
```

```
colnames(vehicles)
```

```
## [1] "id"      "make"    "model"   "year"    "class"   "trans"   "drive"   "cyl"     "displ"
```

```
## [10] "fuel"    "hwy"     "cty"
```

```
table(vehicles$cyl)
```

```
##
```

```
##      2      3      4      5      6      8     10     12     16
```

```
##     45    182 12381    718 11885   7550    138    478     7
```

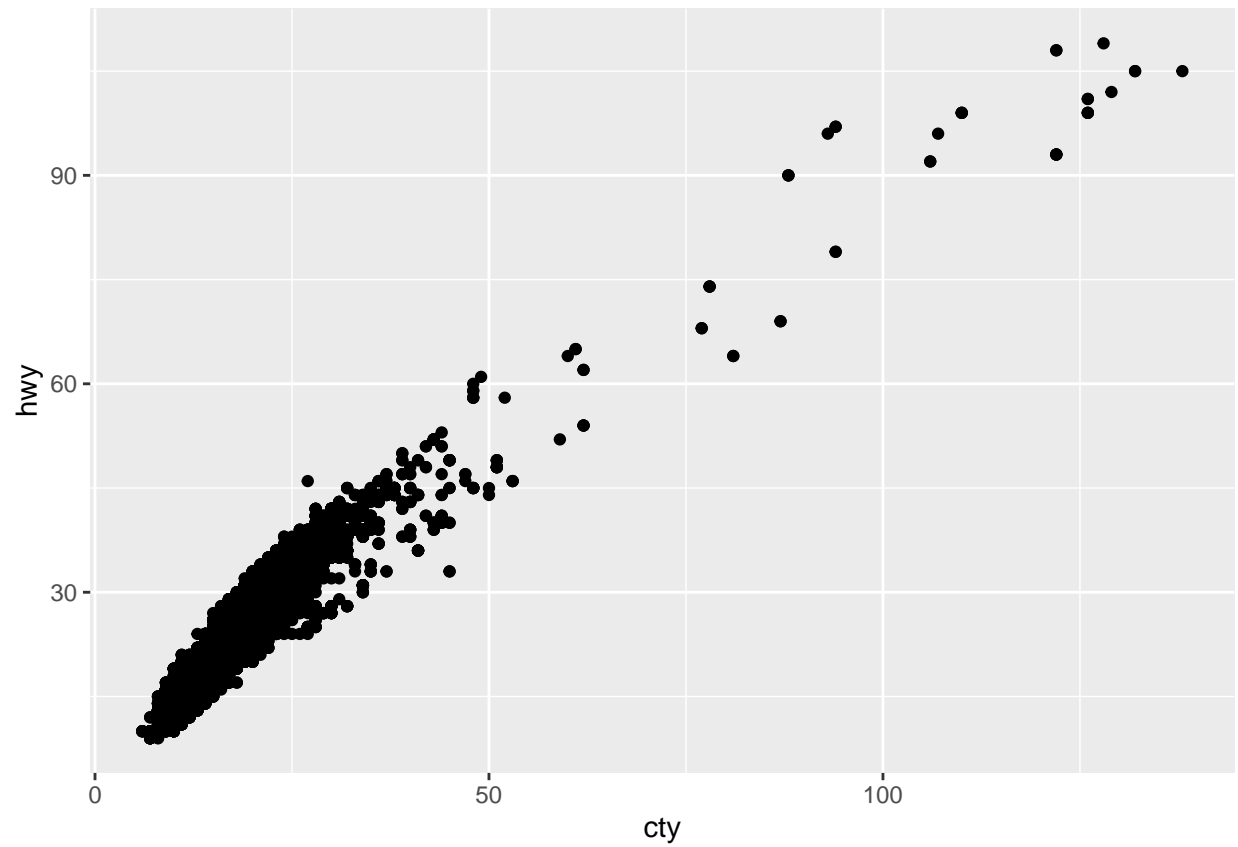
```
summary(vehicles$cty)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##      6.00   15.00   17.00   17.49   20.00   138.00
```

- b) Using `ggplot`, create a scatterplot of `hwy` vs `cty`. Then, using `cyl` as a factor, recreate the above scatterplot coloring by `cyl`. Because there are so many datapoints, the graph is hard to interpret (that is, it is difficult to identify which areas are “denser” than others). Adjust your `alpha` parameter to prevent overplotting and show a more interpretable graph.

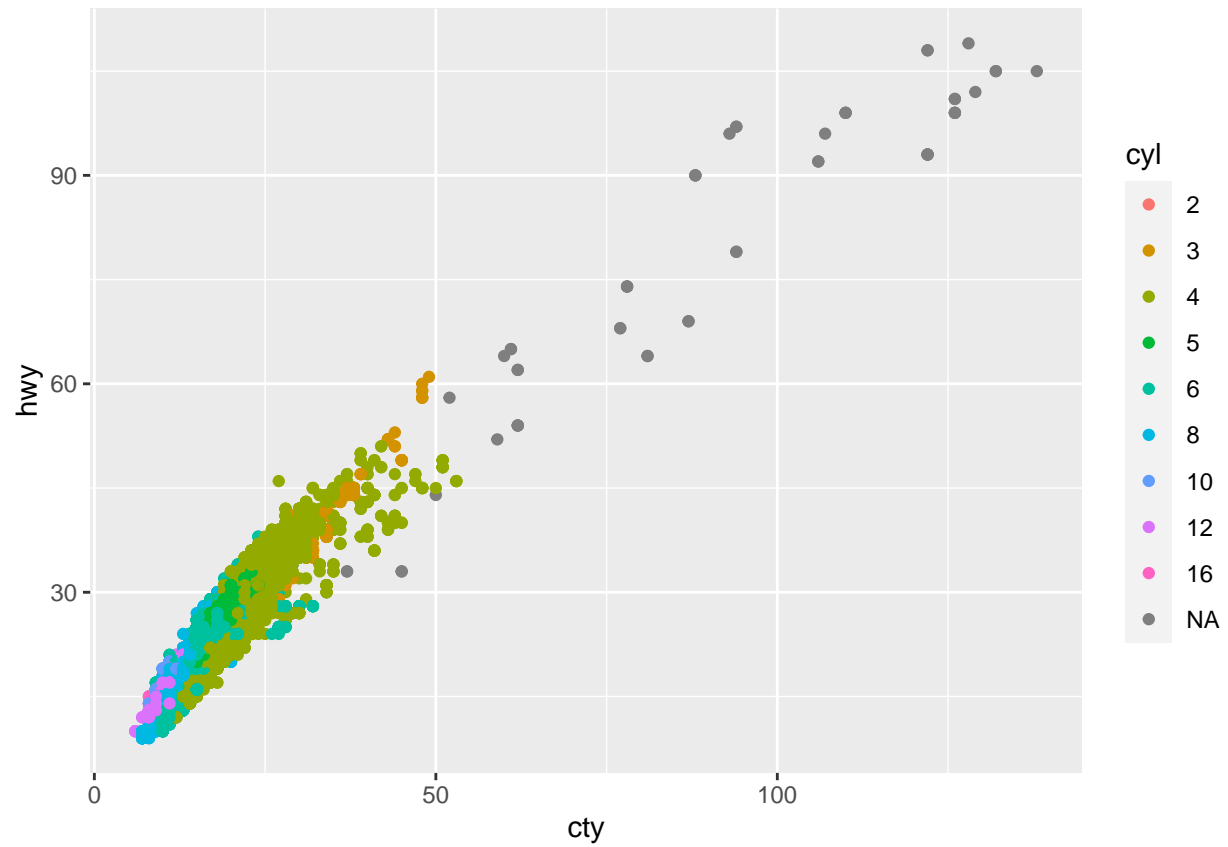
```
ggplot(vehicles, aes(x = cty, y = hwy)) +  
  geom_point()
```



Now adding `cyl` as a factor, we have

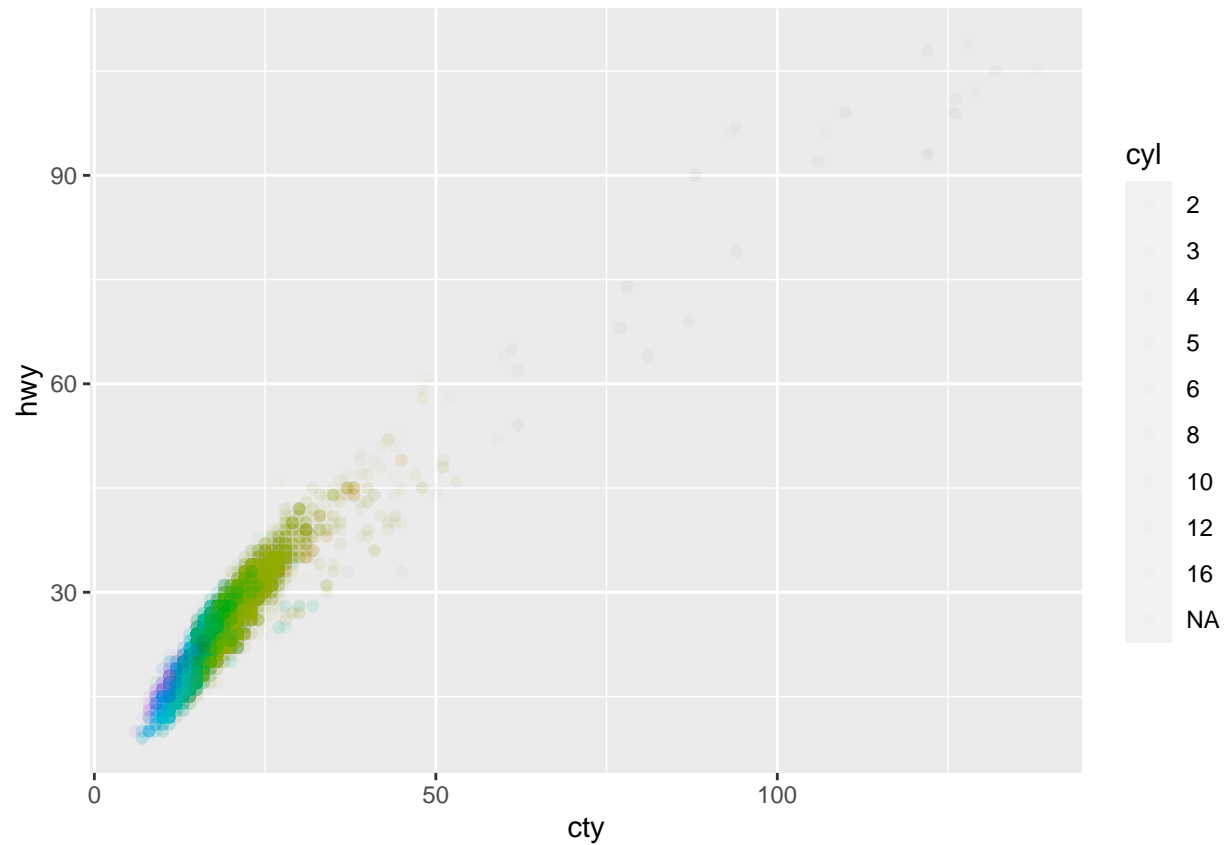
```
vehicles$cyl <- factor(vehicles$cyl)

ggplot(vehicles, aes(x = cty, y = hwy, col = cyl)) +
  geom_point()
```



There is a lot of overplotting here. Let us try adjusting the `alpha` parameter

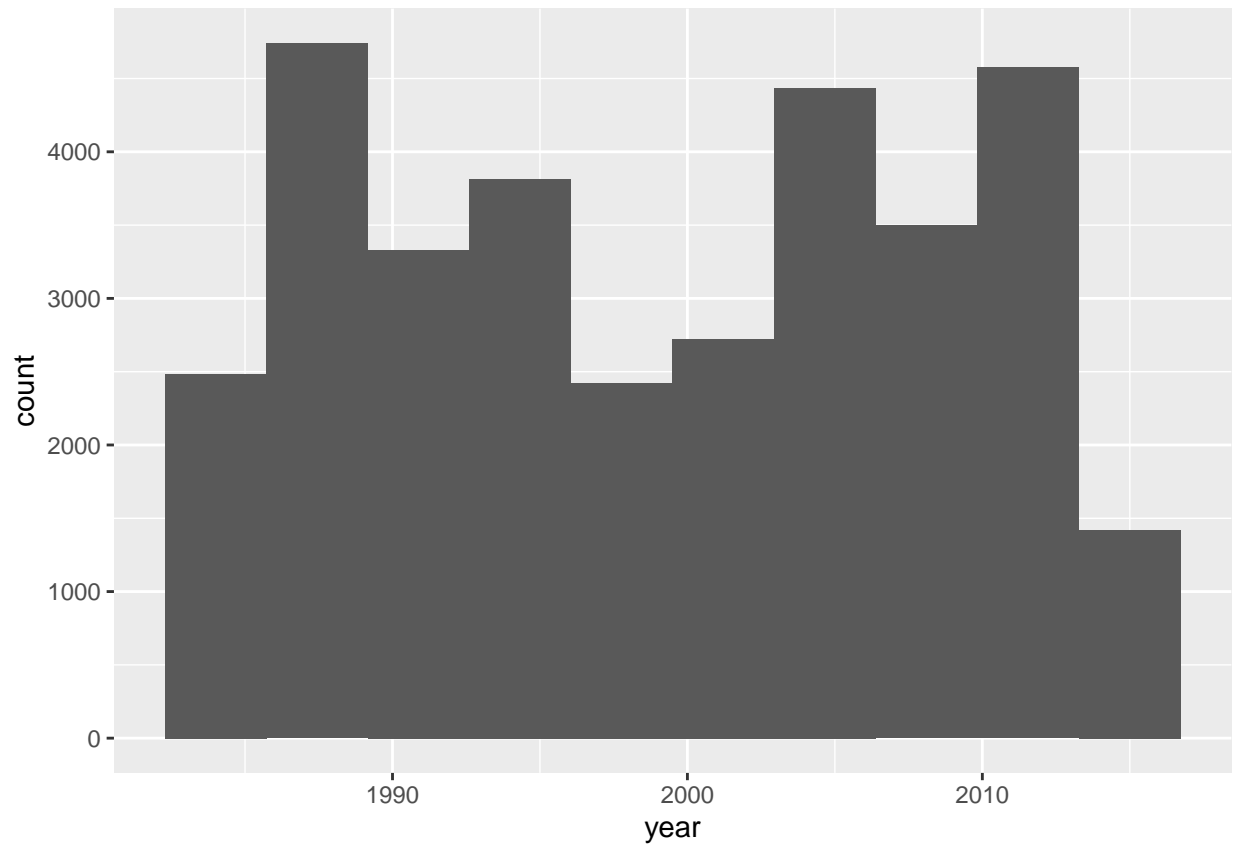
```
ggplot(vehicles, aes(x = cty, y = hwy, col = cyl)) +  
  geom_point(alpha = 0.02)
```



With a very small alpha, we can see that most of the data is concentrated in this ellipse in the lower left corner. The data is much sparse on the edges and away from this mass.

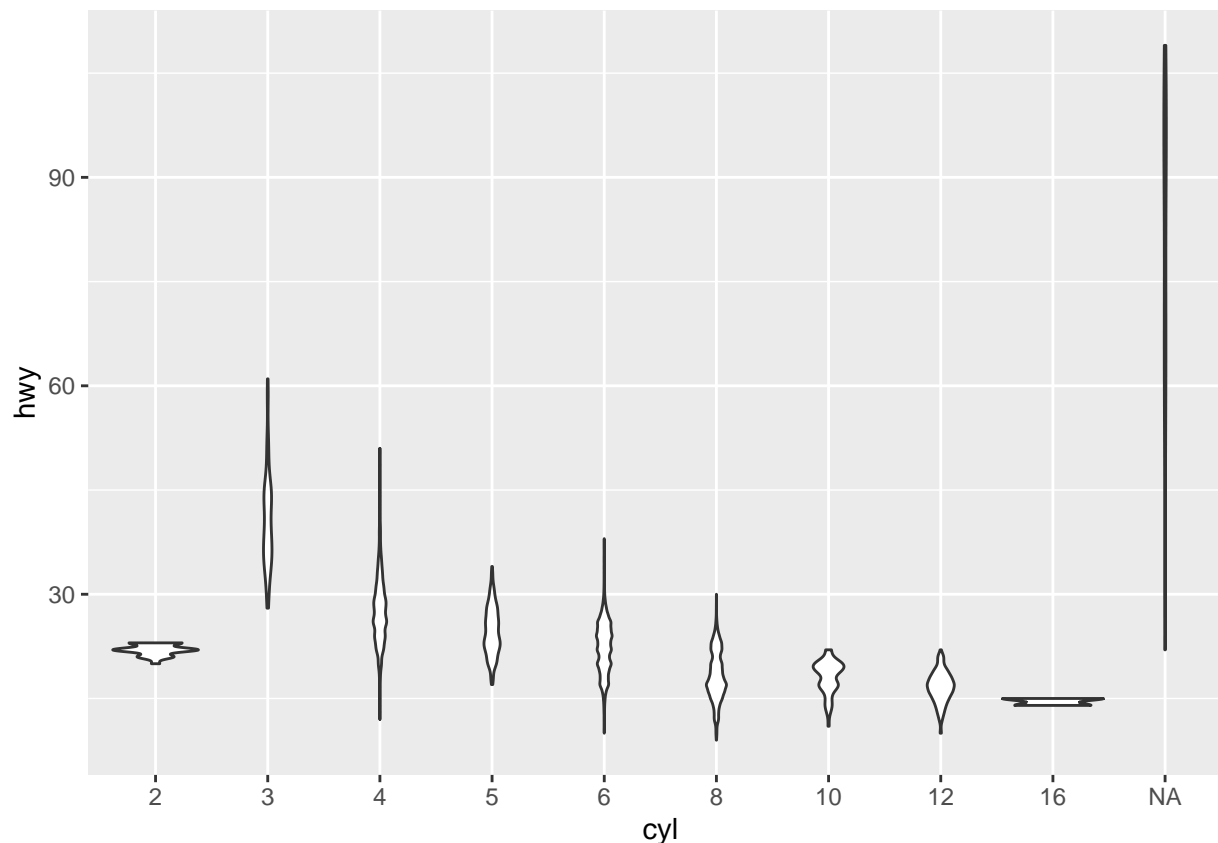
c) Make a histogram of `year`, setting bins to 10.

```
ggplot(vehicles, aes(x = year)) +  
  geom_histogram(bins = 10)
```



d) Create a violin plot and a box plot of `hwy` for each value of `cyl`.

```
ggplot(vehicles, aes(x = cyl, y = hwy)) +  
  geom_violin()
```

2)

The file IsotopeData.xlsx contains pairs of values (stable isotopes of carbon and nitrogen – see https://en.wikipedia.org/wiki/Isotope_analysis) for 45 samples of krill, an important marine species and the source of food for many marine organisms. The samples are classified according to five different species. Devise a plot that aims to show the differences between the five species.

```
library(readxl)

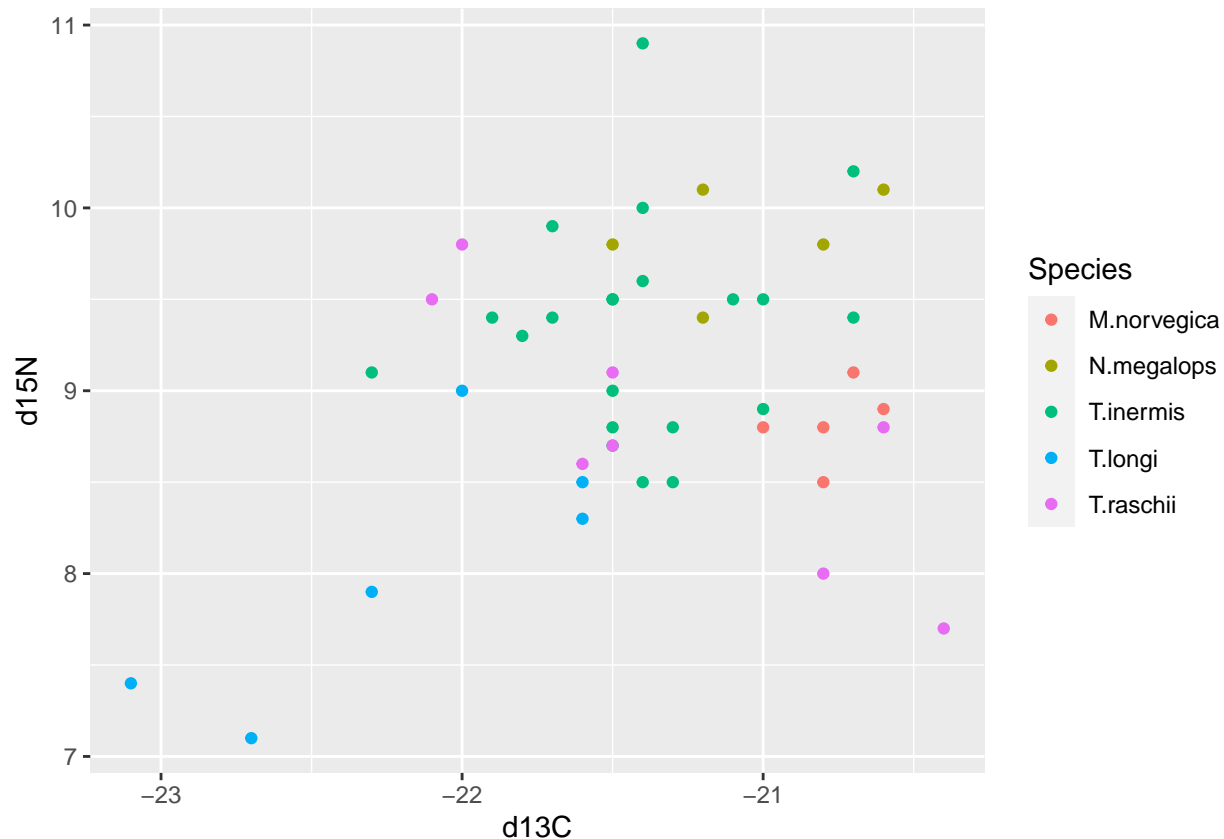
isotope <- read_excel("~/IsotopeData.xlsx") #the columns we care about are 2-4

## New names:
## * `` -> ...5

species_names <- c("T.inermis", "T.raschii", "T.longi", "M.norvegica", "N.megalops")
isotope <- isotope[2:4]

#remap species to name
isotope$Species = as.factor(species_names[isotope$Species])
colnames(isotope) <- c("Species", "d13C", "d15N") #rename the columns

ggplot(data = isotope) +
  geom_point(aes(y = d15N, x = d13C, col = Species))
```



3)

Load the data `mtcars` as we did in the R session.

a) What are the dimensions of this dataset? How many of each value are there in the `cyl` column?

```
data(mtcars)
dim(mtcars)

## [1] 32 11
table(mtcars$cyl)

##
##  4  6  8
## 11  7 14
```

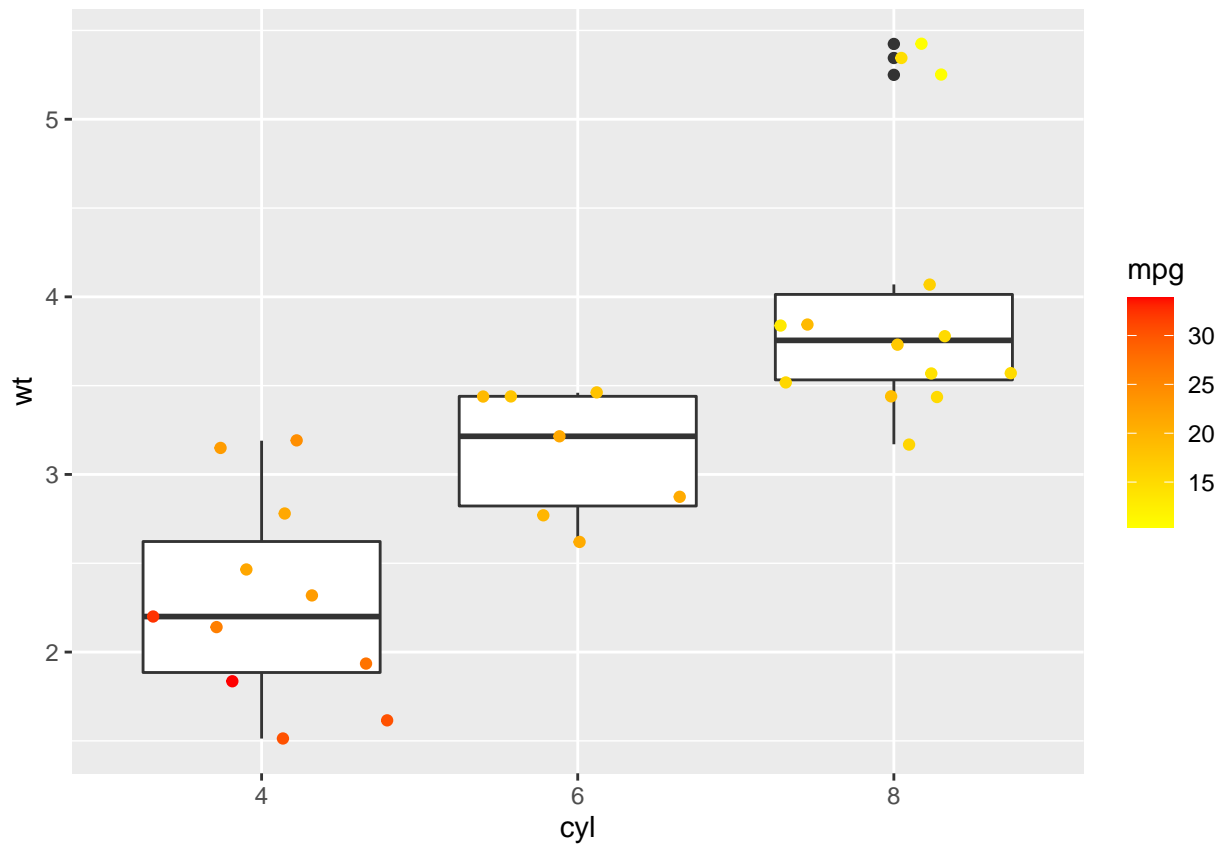
The dimensions are 32 by 11. We can once again look at the counts of each value for `cyl` using the `table` function.

b) Graph a violin plot of `wt` vs `cylinders`. Layer the points onto this plot with jitter. Color the points according to a gradient based on the values of `mpg`. What kind of trends do you notice among the 3 attributes?

```
mtcars$cyl <- as.factor(mtcars$cyl) #set cyl as a factor

ggplot() +
  geom_boxplot(data = mtcars, mapping = aes(x = cyl, y = wt)) +
  geom_point(data = mtcars, mapping = aes(x = cyl, y = wt, col = mpg),
```

```
position = "jitter") +  
scale_colour_gradient(low = "yellow", high = "red")
```



It seems as if though cars with lower cylinders tend to be lighter in weight, and it seems like lighter weight cars typically have higher mpg. In this sense, we expect `cyl` and `wt` to be positively correlated, whereas `mpg` is negatively correlated with both of the above predictors.