Question 1

a. A point at (0,1.5) is an example, anything far top right will likely work
b. 1 Nearest Neighbor: Triangle, 3 Nearest Neighbor: Square
c. It would be classified as square. In this graph, 0 is the average. If a new feature has both attributes below average, then x and y in the graph would be negative, and the only thing that something falling in that range can be classified as is square since there are only squares in that area.

Question 2

a. The 11th bar is a larger width than the 10th bar, so it has more income and hence, it likely to contain a larger percentage as well. The 10th bar only yas 90k-99,999, while the 11th bar represents all incomes from 100k to 129,999

Question 3

a. (y_estimate - 6) / .4 = .75 * (24 - 23) / .5 (Regression line equation in standard units), y_estimate = 33/5
b. The true line in the regression line. We already have an estimate for our regression line given our sample, we calculated it in part a. The goal for bootstrapping is to find the height of the **true line** which we can't tell.

Question 4

a. 25 students. There are 500 students, and we know that, the the P-Value cut-off is the proportion of times we expect to incorrectly reject the null. There are 500 tests with a p-value cut off of .05, so 500 * .05 is 25.

Question 5

a. There are two ways of classifying correctly. Drawing a class A, and getting it correct, or Drawing a class B, and getting it correct.
   For the class A, you have a .85 chance of drawing it, and given you draw it, you have a .9 chance of getting it right. So, .85 * .9 is the chance drawing an A and getting it correct. Similarly for Class B, we have a .15 chance of drawing a Class B, and we have a .98 chance of getting it right given it's class B, so we have a .15 * .98 chance of both.
   Add the two up and get .85*.9 + .15*.98 for the chance of classifying correctly, since there's only two ways of classifying correctly.
b. **Not in Scope for Summer 2018:** (.15*.98)/((.85*.9)+(.15*.98))

Question 6

a. For x values (-3,-2,-1,0,1,2,3), your y values should be (4,4,3,2,3,4,4)
b. 0. minimize(my_function) will return the argument to my_function which returns the smallest value. Given our graph above, we know that the argument 0 will return 2, which is the smallest possible return value.

Question 7

a.
```
r_values = make_array()
for i in np.arange(10000):
      resample = bp.sample()
      new_r = corr(resample.column(0), resample.column(1))
      r_values = np.append(r_values, new_r)
left_end - percentile(5, r_values)
right_end = percentile(95, r_values)
make_array(left_end, right_end)
```

b. Null: r = .6, alternative: r not equal to .6, using a P-Value cut-off of 10% (since we made a 90% CI), we would fail to reject the null if .6 is in our interval, and reject the null hypothesis if 0.6 is outside of the interval.

Question 8

a. Graph B. First bar is shorter than the second bar because they have the same amount of cars, but the width of the first bar is way bigger. To compensate, the height must be less. Since the areas are the same of the first two bars, and the widths are off by roughly half, the heights should be off by roughly half as well.
b. 19.08. len(prices) = 152, so the 10th percentile of prices is the 15.2th element. Round up to the 16th element, and this is 19.08.

Question 9

a.
```
insured.where('Insured', 1).num_rows / insured.num_rows
```

b.
```
combined = insured.join('Zip Code',states).select('State',
'Insured')
combined.group('State', sum).sort(1,
descending=true).column(0).item(0)
```

c.
```
combined = insured.join('Zip Code',states).select('State',
'Insured')
combined.group('State', np.mean).sort(1, descending =
True).column(0).item(0)
```

Question 10

a. 2 pounds, we know that 3 SD on each side should contain most of the data, 2 SD on each side should contain 95%, and 1 SD should contain around 68%, so 2 pounds seems the most reasonable here
b. 400. The width of histogram 2 is half the width of histogram one. To get this by only changing sample size, you want sqrt(sample size) to be 2 times as much as sqrt(100) or the sqrt(original sample size), which is 10. 400 achieves this.

Question 11

a. First plot. To begin with, actual values are over the residual line, then they fall below, then they fall above again in the original graph.
b. Less than 0.995. There will be more points and the graph will become fuzzier and show less clustering around a straight line. This is an example of ecological correlation

Question 12

a. False. It is a confidence interval for all households in the **population**. We already know the median income for the households in the sample.
b. Can not be approximated. This interval is trying to find the median annual income for the population, but can not tell anything about how the population is distributed.
c. More than 54,000. Incomes are generally skewed right (some outliers make way more money), the mean of the interval will be larger than the median.

Question 13

a. No, it does not say the observed distributions are the same. it says that the difference in our observed distributions are due to chance, and it actually compares the population distributes from handedness in males and females and claims that they are the same.
b. The distributions of handedness among men in the population is different than the distributions of handedness among females in the population.
c. TVD because we are comparing two categorical distributions and we want to see if they are the same.
d. Option 3. If the null hypothesis is true, it shouldn't matter which values correspond to what labels of male and female, since the population distributions are the same.

Question 14

a. x values of (-3,-2,-1,0) have y values of (5/3, 0, 5/3, 20/3)
b. Option 2, slope of -2 gives the smallest possible error for MSE (0)

Question 15

a. The results achieved were from random guessing, so the chance of getting a question right is ⅕. Any difference in our sample is due to random chance.
b. The chance of getting a question right is less than ⅕.
c. Normal distribution, center of 200, and 180 is to the left. We expect, under the null, to see 200 correct answers, and since we are simulating sampling proportions, we expect it to be normally distributed.
d. Shade everything to the left of 180 since smaller values of the test statistic point towards the alternative, and the p-value calculation requires us to find the chance of getting values equal to the observed test statistic or further in the direction of the alternative.

Question 16

a. We are looking at bootstrapping, so sampling with replacement. To begin with, if we draw four times, the chance that the first person is unique is 4/4. Now, the chance that the second person is unique is ¾ (pick anyone but the first person). For the third choice, we need to pick one of the other two people that we have not picked yet, which happens with chance 2/4. Lastly, in the 4th pick, we need to pick the unique person, which is ¼. So, the total probability is 4/4 * ¾ * 2/4 * ¼.

b.
```
def same(N):
      prod = 1
      for i in np.arange(N):
            prod = prod * ((N-i)/N)
```