1) K-means is guaranteed to converge within X iterations, where X equals:
Let N be size of data, K is number of centers
None of the above. K-means is guaranteed to converge eventually, but we don't have a bound on how many iterations it will take.

2) The best value of K should always be at least 10.
False - best value of K depends on the data.

3) K Means++ is K-Means with a different way of initializing the centers.
True, K Means++ tries to maximize the distance between the initial clusters.

4) Suppose we've already run K-means with k=3, and have the following cluster centers: {Red=(1, 6), Green=(5, 3), Blue=(2, 2)}. We then receive a new point (4, 5), which cluster do you predict it belongs to?
We predict based on the cluster center which has the shortest distance to the point.
Distance((4,5), (1,6)) = 3.16
Distance((4,5), (5,3)) = 2.2
Distance((4,5), (2,2)) = 3.6

So the closest cluster is Green.

5) The Reservoir Sampling algorithm has an approximate runtime of:
K = number of centers, N = size of dataset, d = number of dimensions

O(N) – this answer implicitly assumes we're only sampling one center, like we did in HW5. If you wanted to sample *x* items from the dataset, you'd also need to keep a heap of items sorted by the weight key (u_i^(1/w_i)).

6) K-Means++ initialization will not work if our dataset is too large to fit in memory

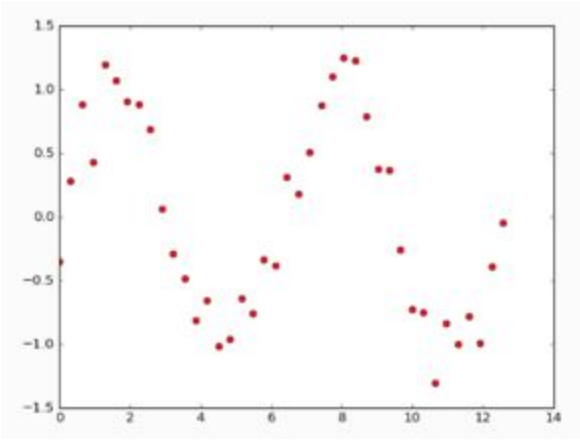
False because we can use a streaming algorithm to sample our initial points -- this is what you did in HW5!

7) We have two data points (2, 3), (4, 5) and we've computed our regression line as y = ½x + 1. What is our mean squared error?

From the picture you can see the first data point is 1 unit from the red regression line, and the second data point is 2 units from the regression line.

$(1^2 + 2^2) / 2 = 2.5$

You can also plug the points into the equation and arrive at the same answer.



8) Suppose we wanted to apply linear regression to this data. Which of the following features would you include to better fit the data?

Higher degree polynomials can help better fit the non-linear data. Choose 1, x, x^2, x^3.

9) Increasing the number of features will guarantee your model to perform better.

False because your model may overfit on the training data

10) Mark the following true statements regarding regularization:

- Regularization is a way of mitigating overfitting - True

- Increasing the regularization parameter will decrease bias
- Increasing the regularization parameter will increase bias - True
- Decreasing the regularization parameter will increase variance - True
- Decreasing the regularization parameter will decrease variance