Quiz 11 - Cluster Analysis in Spark

1. What percentage of samples have 0 for rain_accumulation?

- **157812 / 158726 = 99.4%**
- 157237 / 158726 = 99.1%
- There is not enough information to determine this

2. Why is it necessary to scale the data (Step 4)?

- Since the values of the features are on different scales, all features need to be scaled so that all values will be positive.
- **Since the values of the features are on different scales, all features need to be scaled so that no one feature dominates the clustering results.**
- Since the values of the features are on different scales, all features need to be scaled so that the cluster centers can be displayed on the same plot for easier analysis.

3. If we wanted to create a data subset by taking every 5th sample instead of every 10th sample, how many samples would be in that subset?

- **317,452**
- 1,587,257
- 158,726

4. This line of code creates a k-means model with 12 clusters:

```
kmeans = KMeans (k=12, seed=1)
```

What is the significance of "seed=1"?

- **This sets the seed to a specific value, which is necessary to reproduce the k-means results**
- This means that this is the first iteration of k-means. The seed value is incremented by 1 every time k-means is executed
- This specifies that the first cluster centroid is set to sample #1

5. Just by looking at the values for the cluster centers, which cluster contains samples with the lowest relative humidity?

- **Cluster 4**
- Cluster 3
- Cluster 9

6. What do clusters 7, 8, and 11 have in common?

- **They capture weather patterns associated with warm and dry days**
- They capture weather patterns associated with high air pressure
- They capture weather patterns associated with very strong winds

7. If we perform clustering with 20 clusters (and seed = 1), which cluster appears to identify Santa Ana conditions (lowest humidity and highest wind speeds)?

- **Cluster 12**
- Cluster 1
- Cluster 16

8. We did not include the minimum wind measurements in the analysis since they are highly correlated with the average wind measurements. What is the correlation between min_wind_speed and avg_wind_speed (to two decimals)? (Compute this using one-tenth of the original dataset, and dropping all rows with missing values.)

- **0.97**
- -0.12
- 0.62