Quiz 4 - Data Preparation

1. Which of the following is NOT a data quality issue?

- Inconsistent data
- **Scaled data**
- Missing values
- Duplicate data

2. Imputing missing data means to

- **replace missing values with something reasonable.**
- drop samples with missing values.
- replace missing values with outliers.
- merge samples with missing values.

3. A data sample with values that are considerably different than the rest of the other data samples in the dataset is called an/a _____.

- **Outlier**
- Invalid data
- Noise
- Inconsistent data

4. Which one of the following examples illustrates the use of domain knowledge to address a data quality issue?

- Simply discard the samples that lie significantly outside the distribution of your data
- Drop samples with missing values
- **Merge duplicate records while retaining relevant data**
- None of these

5. Which of the following is NOT an example of feature selection?

- Adding an in-state feature based on an applicant's home state.
- Re-formatting an address field into separate street address, city, state, and zip code fields.
- Removing a feature with a lot of missing values.
- **Replacing a missing value with the variable mean.**

6. Which one of the following is the best feature set for your analysis?

- **Feature set with the smallest set of features that best capture the characteristics of the data for the intended application**
- Feature set with the smallest number of features
- Feature set with the largest number of features
- Feature set that contains exclusively re-coded features

7. The mean value and the standard deviation of a zero-normalized feature are

- mean = 0 and standard deviation = 0
- mean = 1 and standard deviation = 0
- **mean = 0 and standard deviation = 1**
- mean = 1 and standard deviation = 1

8. Which of the following is NOT true about PCA?

- PCA stands for principal component analysis
- PC1 and PC2, the first and second principal components, respectively, are always orthogonal to each other.
- PC1, the first principal component , captures the largest amount of variance in the data along a single dimension.
- **PCA is a dimensionality reduction technique that removes a feature that is very correlated with another feature.**