CSSE 386 – Data Mining with Programming
Rose-Hulman Institute of Technology

## Worksheet 04

Name (Print):_____            Section:_____

1. How null/missing values are typically coded in the data?

    1.
    2.

2. What are the most common reasons for missing data?

3. Types and Origins of Missing Data

| Term | Definition |
|------|------------|
| MCAR | |
| MAR | |
| MNAR | |

4. What is the difference between Person-Level and Item Level Missingness?

5. How much missing data can be ignored (typically)?

    - 50%
    - 10%
    - 20%

6. Missing data can be deleted (typically)?

    - 50%
    - 15%
    - 20%

7. What are strategies for deletion? Name at least one

    -
    -

8. Imputation types

| Type | Method |
|---|---|
|  | Logistic regression, NA as a level |
|  | Mean, Median, Linear Regression |

9. Identify python functions for missing data

```
1    # Check for missing or NaN values
2
3 ------ -------
4
5 # Removes rows, columns with missing data
6
7 ------------
8
9 # Fills missing values with a specified value
10
11 -------------
```

10. Deleting using axis

```
1    # Delete rows
2
3 data.dropna( _____)
4
5 # Delete columns
6
7 data.dropna(_____)
```

11. Imputation

```
1    # Median
2
3 data.fillna( _____)
4
5 # Mean
6
7 data.fillna(_____)
```