Once you've identified a Use Case and Data Set it is time to get familiar with data. In the process model this task is called Initial Data Exploration. Please take a minute or two to (re)visit the following lecture

https://www.coursera.org/learn/data-science-methodology Module 2 - Data Understanding

Please also revisit http://coursera.org/learn/ds/ Module 3 - Mathematical Foundations and Module 4 - Visualizations

Given the lectures above, please create statistics and visualization on your Data Set to identify good columns for modeling, potential data quality issues and anticipate potential feature transformations necessary.

Create a jupyter notebook where you document your code and include visualizations as first deliverable. Please also stick to the naming conventions explained in the the process model manual.

So, the most important reasons / steps are:

- Identify quality issues (e.g. missing values, wrong measurements, …)

- Assess feature quality – how relevant is a certain measurement (e.g. use correlation matrix)

- Get an idea on the value distribution of your data using statistical measures and visualizations

✓ Complete