Below are the summary statistics I plan on developing for the spatial data like coordinates and uncertainty in meters etc.

1. Mean nearest neighbour distance is exactly what the name says!

2. G function is the cumulative frequency distribution of the nearest neighbour distance. It gives the probability for a specified distance, that the nearest neighbour distance to another event in the pattern will be less than the specified

3. F function is the cumulative frequency distribution of the distance to the nearest event in the pattern from random locations not in the pattern.

4. K function is based on all inter-event distances, not simply nearest neighbour distances. Interpretation of the K function is tricky for the raw figures and makes more sense when statistical analysis is carried out as discussed in a later section.

5. Pair correlation function is a more recently developed method, which like the K function is based on all inter-event distances, but which is non-cumulative, so that it focuses on how many pairs of events are separated by any particular given distance.

Thus, it describes how likely it is that two events chosen at random will be at some particular separation.

The relevant links for above source are:

<https://www.e-education.psu.edu/geog586/book/export/html/1734>

<http://gispopsci.org/point-pattern-analysis/>

More over basic statistics like mean, max, min, range and standard-deviation will be common.

I also checked the inter-quartile range of the data, but the results were biased. If possible I will remove the biasness.

For grouping of the data I used cluster analysis and K-means clustering.

For the summary table for string fields below are the points which I found relevant:-

1. In the scientific name and name field we can get the frequency of each data entry and therefore get the most abundant species and the least.

2. If the area wise or country wise data is available we can also get a table to show user the most abundant species in particular area.

3. The number of NA values in particular field.

4. For class, phylum, kingdom, family, genus etc above functions can be extended.

5. The summary statistics will also depend on the user, like if a user requests data for all species in a particular area we can extend the summary by giving details of most abundant species in that area. Currently I am studying this area to get some more important features.

For the date, year and other temporal fields we can use measures of central tendency where ever applicable and also, we can get density of species with time i.e. between which time period most species were spotted.

Currently I have written some code regarding the above ideas. Link:

<https://github.com/AshwinAgrawal16/BiodiversitydatacleaningGSoC17/blob/master/FurtherProjectWork/DwCSummary/DwCSummary-RawCode>