2022 Sem2

Project Stage 1 (Notes)

Due: 11:59pm on Sunday at the end of week 6

Value: 5% of the unit

Note: In order not to have a long and complicated instructions, some of the details are included in a separated companion note, The project may sound elaborate, but the work you need to do is not actually very much. It should be easy to fit into the provided two weeks of your time, as long as you interact frequently and apply any feedback from the tutors. Don't wait till near the due date to start! If anything in the instructions is unclear or confusing, please ask about it on Edstem, using the category "Group Report", and sub-category either "Group Formation" or "Stage 1" as needed.

TASK 1: IDENTIFY TOPIC:

For example, it is not a good to ask just "which country has the highest level of wealth?" but it is a good choice of question to ask "what influences the level of wealth in a community?". You might look at datasets that relate to the economy, climate, education, type of government etc.

TASK 2: OBTAIN DATASETS AND METADATA:

While you can choose datasets as you wish, there are some extra requirements if you aim for higher marks, not just Pass level.

- If the group is hoping to score above Pass level,
  - The different datasets (from the different members) should all "have independent origin".
  - Note that this refers to the origin or primary source of the datasets; it is ok for the data to have been obtained from the same data-providing website, as long as the origins are different.
    - For example, data.gov.au offers the possibility to download many datasets, similarly, the various competitions at kaggle.com often have datasets from quite different origins.
  - You also should ensure that each dataset has "medium volume of data", so that automation of processing becomes crucial.
    - For defining volume, we will consider the number of "values": for the most common case, rectangular data e.g., CSV. The contents of a field for an item would be a value. A dataset is considered as medium volume if it contains at least 500 values.
      - So, if you have 100 rows of data, each with 5 attributes, that would be 500 values. For JSON data, the keys don't count, and the values count based on

their atomic (string, number etc) components: so, if one attribute's value somewhere is a list of 5 numbers, that counts as 5 values; if it is a dictionary with 7 keys, each associated to a string, that counts as 7 values.

- Finally, it isn't graded in this stage, but to score well in later stages, your integrated dataset needs to contain both the following:
  - at least one attribute whose value is either a count or a measurement (a number in terms of some unit), and
  - at least one attribute whose value is a string (or numeric identifier where the values are not meaningful as numbers, like the student id).

TASK 3: ENSURE DATA QUALITY:

For example, the work needed may be removing instances that have corrupted or missing values or filling in those missing values in some sensible way; you may be correcting obvious spelling mistakes, or bringing different date formats to a common standard; maybe you need to remove duplicate rows, or deal with inconsistent information (e.g., two different values for population of the same country).

TASK 4: PRODUCE SOME DATA SUMMARIES:

For example, you might calculate (and show the code in your report) the highest value of wealth among all the countries, or the number of countries in the Asia region. This is not intended to be a detailed exploration of the data (that will come in Stage Two), but it is simply a demonstration that the dataset is now in a form where you can work with it, and that you have the required skills in Python coding.

TASK 5: INTEGRATE THE DATA SETS:

For example, we might want to integrate climate data from different locations. It is easiest when the datasets share a structure (for example, they may represent climate data in different states, all with the same schema such as "city, date, maxtemp, mintemp, rainfall (mm)"), so that integration is nothing more than combining the rows one after another (perhaps adding an extra attribute to distinguish which dataset each row came from). However, if the datasets do not share a structure, then there is a lot of decisions needed to find a common structure into which all can be placed (and how to deal with eg missing values etc).

Another case is where the datasets contain different attributes for the same entities; for example, one might have climate data for cities on dates, and another dataset has transport data for the same cities and dates. In that case, the integration can simply make a longer row for each entity, with all the different attributes from the datasets (eg city and date, followed by climate attributes and then by transport attributes). However, care is needed, if the entities are described with different formats etc, so that transformation is needed to allow them to be matched up across data sets.

MARKING

IDENTIFYING THE TOPIC [1 POINT]

This component of assessment is based on the corresponding section of the report.

Full marks: a clear *and exciting* statement of a topic, question, or problem, along with a clear account of *several distinct* stakeholder groups who would be impacted in different ways by improved understanding or solutions, and also a *convincing* explanation of how the datasets can be useful in resolving the issue

Distinction: a clear statement of a topic, question, or problem, along with a clear account of at least one *stakeholder group* who would be impacted by improved understanding or solutions, and also a clear explanation of *how all of the datasets can be useful* in resolving the issue

Pass: a clear statement of a topic, question, or problem, to which all of the provided datasets are potentially relevant.

Flawed: A solid attempt to describe the topic

DATASETS AND METADATA [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report

Full marks: one different dataset for each member of the group, each having a *detailed and thorough* statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance, including evidence that you are authorized to use the data as you have done. There must also be an *insightful discussion* of the strengths and limitations of the dataset. The datasets must have independent origin, and each must be at least medium volume.

Distinction: one different dataset for each member of the group, each having a statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance, including *evidence that you are authorized to use the data* as you have done. The datasets must have *independent origin*, and each must be at least *medium volume*.

Pass: one different dataset for each member of the group, each having a statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance

Flawed: Some dataset

ENSURING DATA QUALITY [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks: there is one different dataset for each member of the group, and in each dataset, several distinct aspects of data quality have been checked in an automated way by Python code, and if any problems were found, they have been all handled in an automated way by Python code (that is, the "clean" dataset from this member does not suffer from any of these particular quality problems). The datasets must have independent origin, and each must be at least medium volume.

Distinction: there is one different dataset for each member of the group, and in each dataset, some aspect of data quality has been checked *in an automated way by Python code* and if any problems were found, they have been handled (that is, the "clean" dataset from this member does not suffer from this particular quality problem). The datasets must have *independent origin*, and each must be at least *medium volume*.

Pass: there is one different dataset for each member of the group, and in each dataset, some aspect of data quality has been checked and if any problems were found, they have been handled (that is, the "clean" dataset from this member does not suffer from this particular quality problem)

Flawed: Some reasonable attempts to improve or check data quality

PRODUCE SOME SIMPLE DATA SUMMARIES [1 POINT]

This component is assessed based on the corresponding subsections of all the separate dataset sections of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks: there is one different dataset for each member of the group, and for each dataset, you have written *well-documented and clear* Python code which runs on the dataset, and correctly reports on *at least three* suitable (and different) aggregate summary statistics (such as the highest value, or the number of different values) from the dataset. At least one of the summaries *in each dataset* must be a grouped-aggregate (for example, if that dataset contains a state attribute, it might report the summaries for some other attribute from each state separately). The datasets must have independent origin, and each must be at least medium volume.

Distinction: there is one different dataset for each member of the group, and for each dataset, you have written Python code which runs on the dataset, and correctly reports on at least one suitable aggregate summary statistic (such as the highest value, or the number of different values) for one attribute of the dataset. At least one of the summaries must be a *grouped-aggregate* (for example, if that dataset contains a state attribute, it might report the summaries for some other attribute from each state separately). The datasets must have *independent origin*, and each must be at least *medium volume*.

Pass: there is one different dataset for each member of the group, and for each dataset, you have written Python code which runs on the dataset, and correctly reports on at least one suitable aggregate summary statistic (such as the highest value, or the number of different values) for one attribute of the dataset

Flawed: Some data summary is produced

INTEGRATE THE DATASETS [1 POINT]

This component is assessed based on the corresponding section of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks: there is one different dataset for each member of the group, and you have created an integrated dataset that has a *well-chosen schema* that contains all of the attributes from all of the members' cleaned datasets; this integrated dataset has a statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance. You have provided *well-documented and clear Python code that automatically creates the integrated dataset* from the separate ones. You have described the challenges involved when integrating the datasets (such as incompatible formats for some data types, or inconsistent naming of some attribute-values such as suburb names). The datasets must have independent origin, and each must be at least medium volume.

Distinction: there is one different dataset for each member of the group, and you have created an integrated dataset that contains *all of the attributes* from *all* of the members' cleaned datasets; this integrated dataset has a statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance. You have described how the integration was performed, either step-by-step, or by providing Python code that creates the integrated dataset from the separate ones. *You have described the challenges involved when integrating the datasets (such as incompatible formats for some data types, or inconsistent naming of some attribute-values such as suburb names).* The original datasets must have *independent origin*, and each must be at least *medium volume*.

Pass: there is one different dataset for each member of the group, and you have created an integrated dataset that contains some of the information from at least two of the cleaned datasets; this integrated dataset has a statement of appropriate metadata that describes the data structure, the data meaning, and the data provenance. You have described how the integration was performed, either step-by-step, or by providing Python code that creates the integrated dataset from the separate ones.

Flawed: either a reasonable attempt at integrating the datasets, or a description of difficulties faced in integrating them