# CS145 - High-Level Project 3 Rubric

This document lays out some high-level expectations for what is a good open-ended portion of the third project. More specific details of grading (i.e. how many points or percentage points correspond to each bullet point) are omitted.

- **Project overview (5%)**
  - +
    - Clearly describes a central question to be explored
  - -
    - Does not provide a central question around which explorations are focused
- **Analysis of your dataset (10%)**
  - +
    - Students show that they are using a meaningful dataset in terms of size and complexity. The overall dataset should be at least 250 MB.
    - Students clearly describe the information captured in the table.
    - It is clear that students understand the structure of their datasets such as data sizes and high-level relationships between tables.
    - Students list the keys and foreign keys between tables that will be used for exploration or describe connections between tables in some other way.
  - -
    - Students use a very simple or very small dataset (e.g. only one table with few columns or a dataset with very few tuples overall).
    - Little to no effort in explaining the dataset. It is not clear to the grader that the student(s) behind the project understands the structure of the data they are working with.

- **Exploring your questions, with appropriate visualizations (55%)**
  - +
    - Considers a wide range of features that cover many facets of the data. Features are insightful and well-tuned to the question asked.
    - In addition to features easily queried from the tables, students engineer additional features based on their knowledge and insights. Engineered features may include combinations of existing features (such as distance metrics), bucketing existing features (creating categorical features from continuous ones), etc. but should require more thought than simply computing a count.
    - Charts are well-crafted and have a meaningful title, labels, and axis. The choice of the plot works well with the motivation behind it (the question it seeks to answer). Charts are neat and display the data clearly.
    - Students provide analysis for patterns and trends in the data as well as hypotheses for some of the more interesting relationships observed.

        - Students do not ignore things that they cannot explain in their visualizations; instead, they identify and give thought to why something unexpected or anomalous may be present.
    - -
        - Little to no analysis is done for explorations. Overall low effort is shown.
        - Visualizations are unclear and hard to read.
        - Few features are explored.
        - Mainly features that are not related to the central questions are explored.
        - Students do not engineer additional features for explorations.
        - Students oversimplify the dataset and attempt to make conclusions from just a few simple queries.

- **Predictions based on your explorations (20%)**
    - +
        - We are not looking for any particular performance benchmark here, as long as the model is reasonable and the prediction problem is well-framed. That said, the model should not completely fail or be unusable for predictions.
        - Students use separate training, validation, and test sets.
        - Metrics used to judge the quality of predictions are clearly explained and are appropriate to the problem.
        - Predictions are reasonably framed. A student does not seek to answer, e.g., a difficult NLP or computer vision using the models available to them on BigQuery.
        - Predictions use features that were shown in explorations to potentially have an influence on the prediction task.
    - -
        - Does not have separate training/validation/test sets.
        - Inappropriate metrics are chosen.
        - The features are haphazardly chosen and don't make too much sense for the prediction task at hand.
        - Students use features that they found to not matter in their explorations.
        - The model fails heavily (e.g. achieves less than 1% accuracy or less than .55 ROC_AUC), contains bugs, or is unable to generate predictions

- **Conclusion (10%)**
    - +
        - Gives good, insightful thoughts into analyzing what has been explored and visualized in the previous sections.
        - Has a keen eye for limitations in analysis, and is careful about not making overly definitive claims.
        - Identifies and attempts to offer explanations for aberrations in the data, visualizations, or predictions.

        - If something seems to be true, explains why with data-backed arguments. Conversely, if something seems to not hold, explains with the help of data or visualizations.
    - -
        - Makes highly simplifying assumptions about the data or visualizations, e.g., "this one SQL query shows this correlation, therefore x must be true"
        - Makes very strong "definitive" conclusions (e.g., "x happens exactly because of y")
        - Clearly does not demonstrate enough effort in reasoning about the data and plots obtained in the previous sections

**FAQ**

- Is there an upper bound on the dataset size?
    - We suggest you stick to datasets that are under 10 GB, since larger datasets will take longer to train. You also have the option to sample a subset of data from larger datasets.
- How many models should I train?
    - It's fine for you to train a single model. We want you to spend time instead of querying, visualizing, and reasoning about your data.
- What kind of features should I explore?
    - We expect you to explore diverse features. By diverse, we mean that the features are not overly related to one another. These features should have some theorized relationship to the dependent variable of your research question, but whether or not they are actually to have an influence does not matter.
        - Example of not diverse: If predicting house prices, then sq ft of the house and sq ft of the land of the surrounding property are not diverse. It's effectively the same feature (you have explored the same dimension of information)
        - Something that would be diverse from sq. ft of a house is, say, the elevation of the house, or the demographics of the neighborhood, or the average income of the neighborhood.
    - We would also like at least a couple of features to be things you compute or construct (more than selecting a column out of your dataset).
        - Example of selecting a column: for predicting house prices, there's a column called "sq. ft".
        - Example of not selecting a column (something you have to compute or construct): You compute the distance to the nearest school.