Student Names: TODO
Collaboration Statement:

Turning in this assignment indicates you have abided by the course Collaboration Policy:

`www.cs.tufts.edu/comp/136/2022s/index.html#collaboration-policy`

Total hours spent: TODO

We consulted the following resources:

- TODO
- TODO
- ...

These are the official instructions for checkpoint 1. You can find instructions on how to submit at `www.cs.tufts.edu/comp/136/2022s/checkpoint1.html`

Please consult the full project description at `https://www.cs.tufts.edu/comp/136/2022s/project.html` in addition to this document when working on this checkpoint. It gives details on what we expect when choosing a model and coming up with performance hypotheses.

Exploratory Data Analysis

In this section, you should describe the procedure and results of an exploratory analysis to better understand the properties of the dataset you are working with. You should include any relevant properties you described in checkpoint 0 in addition to your analysis for this checkpoint. (This way we don't have to keep refering back to your previous submission to understand the content of this one.)

At the minimum, your analysis should include the marginal distributions of each feature in your dataset, the distribution of the outcome you will predict (where relevant), and the amount and types of missing data. Other types of analyses you may want to consider include looking at relationships between different features and between features and outcomes.

You should make sure that your analysis covers any properties of the dataset that you want to use when forming your performance hypotheses (for example, if your hypothesis relies on the relationship between the features and label being linear, you should provide some evidence to support that from your exploratory data analysis).

Describing your analysis

In this subsection, describe the exploratory data analysis you will conduct. For each distinct type of analysis you describe (for example examining missing data, or looking at feature correlations), you should clearly state what the goal of this analysis is, and what procedure you followed for conducting this analysis. You should also describe and justify any design choices you made.

This subsection should be 1/2-2/3 of a page.

Analyzing the results

In this subsection, you should describe the results of your exploratory data analysis. This should include both a general summary of your findings (for example, you do not need to

show us the distribution of every single feature, but you should tell us about the general trends), and 3-4 graphs that you analyze in more detail. Your description of general trends should be no more than 2 paragraphs.

Each graph should have 1-2 paragraph of analysis. These graphs should include all of the dataset properties that you use in your hypotheses (except where graphs wouldn't be relevant, for example if your hypothesis depends on the size of the dataset, you do not need a graph to describe that). At least 2 of your hypotheses should relate back to your graphs.

Each graph should include a title, a legend (where applicable), and clear labels on the axes. Your analysis of each graph should include a description of what you see on the graph (for example, line A is higher than line B in the left half of the graph), and a description of the implications of your findings (for example, this means that A outperforms B in the low data regime).

Model and Learning Method Properties

In this section, you should describe your chosen model and learning method in detail. You should also be sure to list any properties of your model and learning method that you will use in your performance hypotheses.

Please consult the project overview page for instructions on choosing a model: `https://www.cs.tufts.edu/comp/136/2022s/project.html`

This section should be no more than 1 page total.

Model

This subsection should include a description of the prior and the likelihood (in terms of equations and words). You should describe in words any assumptions that your model makes. You should pay particular attention to describing any properties that you will reference in your performance hypotheses.

Learning method (inference)

This subsection should include a description of which parameters you will learn from data, and how you will do so. If your learning method makes any approximations (for example, the laplace approximation for the logistic regression posterior predictive), but sure to describe these. You should pay particular attention to describing any properties that you will reference in your performance hypotheses.

Performance Hypotheses

In this section, you should state 3 (or more) performance hypotheses describing how you believe your model will perform on your dataset. These should reference specific model/learning method properties and dataset properties described in the previous sections. They should also be tied to a measurable performance outcome, and you should describe how you will measure this outcome (you need to have a way to figure out whether your hypothesis is true or not!). For example, if you hypothesize that your model will train slowly because of X property of your dataset and Y property of your model, you should state that you will compute runtime of your model training.

Please consult the project overview page for instructions on how we want you to specify performance hypotheses: `https://www.cs.tufts.edu/comp/136/2022s/project.html`

This section should 1/2-2/3 of a page long.