Our group consists of three members: Allie Craddock, Esther Kim, and Haley Fore. While creating the model and gathering data, each member had our own set of tasks to complete. For the Housing Price Prediction model, we split the tasks evenly with each member's strengths taken into consideration. We also met accordingly to work together to solve issues with the model and dataset. Thus, each member played a critical role in contributing to both the model and the team's productivity.

Allie reached out to the dataset author to clarify the team's understanding of the dataset. Allie hosted the repository for the model and downloaded all the datasets, visual figures, and models for ease of access and reproduction. She also did the overarching formatting of the repository. Allie also performed an analysis of different aspects of the data throughout the notebooks to accompany the coding. Allie also performed a meta-analysis of the dataset regarding the social ramifications and the underlying ethical questions.

Esther formatted the Jupyter notebooks to analyze the data. Esther also sorted the dataset by programming the comparison of models while working with machine learning and performing prediction calculations. Esther contributed to the brainstorming aspect of the dataset, as well as the organizational features, and her main contribution to the group was constructing the main portion of the model itself.

Haley drafted, created, and wrote the Model Card and Data Sheet with information provided by Allie and Esther. Haley researched secondary sources to provide an additional perspective from an ethical stance. While writing the Model Card, Haley connected previous course readings to the Housing Price Prediction model, *Social Justice* in particular. Haley also gathered visual representations created by other group members. Haley's main contributions include the creation of the Model Card and Data Sheet, as well as secondary research and associating course readings to the model and dataset.

While the team began using exploratory data analytic techniques (EDA), we began to notice that the data was almost *too* simplistic. After reaching out to the author of the dataset for further clarification, we realized that the data was not only synthetic, the price values were also randomly generated. Although some inspiration for the range of values was taken from real-world examples, the only feature of the houses that had a significant correlation with the price was the square footage.

This was only further proven when we began to train our machine-learning model. We created three different models, with each model adding another house feature to help improve the model's prediction accuracy. However, the first model (which was only trained on square footage) and the last iteration of the model (trained on square footage, neighborhood type (rural, suburban, and urban), and number of bedrooms) were only incrementally more accurate. This only solidified the fact that the dataset the model was based on was flawed and simplistic.

This further reduced the practicality of the dataset, which was frustrating as it also made us question the importance of our model. We had to step back and question how simply performing the technical work only *revealed* how flawed those actions were. After examining our correspondence with the author, we began to formulate our own technical analysis of the purpose and the underlying ethical concerns of the project. Sure, the technical work was fine, but how could we, as a team, demonstrate why our work was still important— even if it didn't accomplish what we originally intended?

The dataset itself was created for computer scientists to practice with, according to the author. However, even a practice dataset should have more transparency, and those lacking complexity should be treated with caution. For example, the neighborhood type (rural, suburban, and urban) did not correlate to the housing prices (contrary to real life). However, the dataset was also missing key components. It failed to bring up relevant socioeconomic factors that could determine the pricing of homes as well. Black and Latino homes tend to be appraised for much less than homes in white neighborhoods, for example.

Additionally, factors in the US regarding redlining and inaccessibility to housing loans by minority groups also contribute to the complexity of the housing market. The housing market has a rich history that was sidestepped in the generation of this dataset. Touching upon the readings of *Social Justice,* the people who would be most impacted by this model would be the minority groups often unseen.

This is all to say that the biggest hurdle of our project was grappling with a clearly flawed dataset. It was frustrating to be knee-deep into EDA and modeling work, only to be dissatisfied with the results. However, it helped contribute to a new understanding of how a flawed dataset can contribute to misperceptions and inaccurate biases— and how we can prevent it, by recognizing it and acknowledging it.

One of the insights about data, ML, and digital communication provided by the project is the importance of features and their impact on the dataset. For this project, different features such as square feet, neighborhood, number of bedrooms, and bathrooms contribute to and influence the prices of homes. The EDA phase allowed us to visually explore the relationships between the features and ML helped us to dig deeper into the data and learn more about each factor's contribution to the price. We were able to identify patterns and relationships between the combination of features and their impact on the data analysis.

For example, we identified that the number of bathrooms has a greater correlation to house prices than the type of neighborhood. This was done by comparing the MSE and R^2 values for the three models that we created. By starting with one feature and adding more features to later models, we were able to see and understand the key features in data analysis. In digital communication, providing the process and steps along with the accurate numbers to represent the steps enhances analysis comprehension and may be used to influence future questions and explorations.

As we worked with synthetic data, we recognized the simplicity and privacy benefits of synthetic data but also identified ethical challenges. We found out that the price values were randomly generated which explained the closeness of our relationship values. We also recognized that the use of synthetic data does not accurately represent the relationship between house features and their price. This brings up questions about whether or not the data generation accounts for socioeconomic factors when making predictions. For the dataset that we worked with, neighborhood type had little correlation with housing prices. This information with price predictions raises questions about the accuracy of the predictions in the real world. We recognized how important it is to recognize the limitations of machine learning such as potential bias behind numbers. This insight helped us understand the importance of transparent communication, bias, and ethical implications of machine learning predictions.

Communication is another insight that we better understood through this project. When we first saw the dataset, we recognized that there were many different analyses that we could do given the numerous rows of data. Evaluating the models that we created and interpreting them using MSE and R-squared values helped to quantify the accuracy of our models. The R-squared values for our three models were 57.21, 57.22, and 57.77 percent respectively. Although the change in values is noticeable, explaining the process of the numbers and their relationships helps improve the understanding of analysis and enhance the model's predictive power. Interpreting the R-squared values and the gradual increase is an important step in communicating the predictions that allowed us to understand the model performance beyond the statistics. Being clear in digital communication helps convey the insights of our numbers that identify patterns, improve transparency, and inform decision-making.