Final Project Individual Report: Jessa Henderson

Group 5 analyzed a dataset for Airbnb rentals in Florence, Italy. The data was from 2019 and included a large variety of data from the Airbnb website. The group aimed to investigate which Airbnb features impacted Airbnb price. As a team, a plan was created to identify a dataset, define the research question, preprocess the dataset, select features using random forest techniques, and run linear regression to answer the research question. These roles were divided amongst team members and the team communicated regularly outside of class to ensure all group participants were moving towards a common goal.

I assisted in the formation of the initial research question with the group and assisted in talking through appropriate models to use based on our research question. Jyoti and Tanvi took the lead on preprocessing, feature selection, and modeling. I was tasked with building the GUI shell and manipulating the preprocessing, feature selection, and modeling code into the GUI. In order to build an adequate GUI, I completed the tutorial provided to the class in Github, referenced the demo code and GUI as an example, and read other online forums to assist with ideas on GUI design and using Matplotlib within a GUI.

I completed the majority of the coding for the GUI. When I was stuck with integrating the preprocessing code into the exploratory data analysis section of the GUI, Jyoti and Tanvi volunteered to help and completed the distribution graph. I coded the remainder of the GUI using the resources provided for setup and the code created by my team for preprocessing, feature selection, and modeling. I sought out resources to adjust the GUI design so that the menu bar color could match the official logo color of Airbnb and so that the Airbnb logo could serve as the backdrop for the GUI. I used vertical, horizontal, and grid layouts in the GUI, changing the

layout based on the purpose of each menu item's canvas. Additionally, I created the slide show design and flow for the presentation using Google Slides. The presentation theme was designed from scratch. It includes Airbnb official colors and fonts that match the official Airbnb fonts. Content in the slide deck was contributed by all members of the team. In the final group report, I wrote the introduction and copy edited our final version for submission. All of us worked together to finalize all of the pieces before final submission.

Our group aimed to investigate which features impacted the prices of Airbnbs in Florence, Italy. Extensive data preprocessing was complete by Jyoti and Tanvi that explored the data set. It removed variables that would require natural language processing and those that had a large proportion of missing values. Features that remained and had missing values were subject to mean or median imputation. Outliers for Airbnb price were removed as they could have skewed the results. Feature selection was completed via random forest as principal component analysis did not work appropriately with the Florence Airbnb dataset. Random forest pulled the top 22 features to be targeted for use in linear regression. This step allowed us to narrow the variable field in hopes of producing a stronger model. See Figure 4 in the group report for a graph of the top 22 features. Finally, two linear regression models were run. The first model used features selected from a heat map and a second model used the features selected from the random forest selection process. The linear regression model that used features based on the heatmap had a higher R^2 than the one using the features from the random forest selection.

The GUI was designed to highlight the most important findings from the process described above. I designed the GUI canvas for feature selection using random forest and the GUI canvas for linear regression with a design that allowed for comparison between graphs. For

example, the random forest canvas shows graphs that used two different coding features to get to the top features that impact price from the dataset. These can now visually be compared side-by-side by the audience. The linear regression canvas in the GUI has a similar design in that it shows the scatterplot for both models side-by-side for easy comparison.

In conclusion, I found building the GUI to be very challenging but ultimately rewarding. I have a greater understanding of the coding involved in the design of the GUI, beyond statistical calculations that I was more comfortable with prior to this project. My team was also wonderful to work with as we were able to communicate and brainstorm easily about modeling techniques and helped each other gain a deeper understanding of the concepts overall in relation to our research question. In the future, I would love to compare the top features in the Florence Airbnb dataset with other Airbnb datasets from around the world to see if the same trends are world-wide or location-based.

I'm not sure if I fully understand the appropriate way to calculate the code that was created. The GUI code is 715 lines long, including spaces and comments. The majority of the organization and code structure came from the GUI tutorial provided to students as a resource; however, a large portion of these had to be manipulated based on our specific needs. For example, I had to adjust the canvas design based on the model graphs we planned to include. This required manipulating the code to try horizontal, vertical, or grid layouts and adjust specific sizes. While some of the feature selection and modeling code could be directly added to the GUI, I had to reformat the plots to ensure they were properly visualized within the GUI. Finally, I added some completely new lines of code to assist in the design of the GUI. Approximately 10% of my GUI code was copied from the internet without any manipulation.