**Name: Greg Wells**

**Project Name: National housing price determinants in the U.S**

1. **Introduction**

One of the most commonly used data sources for residential real estate information is Zillow.com. Both consumers and real estate agents as well as investors commonly use this site to search for properties of interest. It contains similar information as to what is available in the MLS (multiple listing service) used by real estate professionals. This project will focus on a common set of parameters to create a subset of this data: Type: single family, Year Built: 1998-2010, Size: 2250-2750 sq ft, Bedrooms: 3+, Baths: 3+, Lot size: ¼- 1/2 acre, with garage. This subset of data will be gathered via an API and augmented with additional datasets, such as airport proximity, to create new features for analysis. This white paper is intended to provide an overview of data mining on this assembled housing dataset, including data preparation, exploratory data analysis, modeling, and evaluation.

1. **Dataset:**

The Zillow data is freely available only via the website, Zillow.com. However, there is a subscription-based API that can facilitate downloads of all columns on a record-by-record basis. For this analysis, a web scraper will be used to gather addresses from a manual browsing session to gather addresses that conform with the selection criteria. A python program will leverage the address list to query the Zillow API retrieving all available data for each address and subsequently store in a pandas dataframe. This data will be augmented with additional data such as distance from a major airport which will be calculated using a dataset from https://data.humdata.org/dataset/ourairports-usa

1. **Data Preparation:**

The first step in data mining on the housing dataset is to prepare the data. This involves cleaning and transforming the data to ensure that it is in a format that can be easily analyzed. Once the dataset has been downloaded, it is important to check for missing values, outliers, and inconsistencies. This dataset will likely contain duplicates due to the data collection method; duplicates will be dropped. The Zillow dataset contains approximately 300 columns, many of which are empty or not relevant to this analysis, so those columns will be dropped to decrease the dataset storage size and processing time. Records with missing values of significance will be dropped, and outliers will be removed or treated appropriately.

1. **Exploratory Data Analysis**:

After the data has been prepared, the next step is to perform exploratory data analysis (EDA). EDA involves visualizing and summarizing the data to gain insights into the relationships between the variables. For this housing dataset, this may involve plotting the

distribution of each variable, creating scatterplots to explore pairwise relationships, and calculating summary statistics for each variable. EDA can help identify any patterns or relationships that may exist in the data and guide the selection of appropriate modeling techniques.

1. **Modeling**

The next step is to select a modeling technique. I will use three algorithms for the housing dataset, including decision trees, logistic regression, and k-nearest neighbors. The goal of modeling here is to predict the asking price based on the characteristics of the property. The dataset will contain datapoints from across the continental U.S. based on the selection criteria. The modeling process involves splitting the data into training and testing sets, training the model on the training set, and then evaluating the model's performance on the testing set.

1. **Evaluation**

Model evaluation is an important step in data mining on the housing dataset. This involves calculating performance metrics, such as accuracy, precision, recall, and F1 score, to determine how well the model performs. The metrics will depend on the modeling technique used, but in general, a higher accuracy indicates a better-performing model.