**Deliverable 1: Project Proposal**

*The Business Case*

To better understand the fluctuations in the housing market, it is first necessary to analyze the factors that determine a property’s value. Understanding how these factors play into the cost of real estate property, and which factors are significant in raising or lowering a property value, will allow for better estimations on real estate properties. The real estate market is highly dynamic, influenced by economic policies, interest rates, and supply-demand factors. By leveraging historical data, this project aims to: provide insights into market trends; help investors, policymakers, and homebuyers make informed decisions; and forecast future home prices with interpretable predictive models.

*Approach*

In this project we will follow the CRISP-DM framework including the following phases and explained in greater detail below:

* **Business Understanding** – formulate framing business and analytics questions
* **Data Understanding** – collect data from Zillow, Federal Reserve, and IRS
* **Data Preparation** – handle missing values, standardize, transform data
* **Modeling** – test and compare models – OLS Regression, Random Forest, Lasso Regression
* **Evaluation** – evaluate models based on Root Mean Squared Error (RMSE) and Coefficient of Determination (R2)
* **Deployment** – provide completed model to client for application

*The Business Question*

Our focus is on the question – how can we accurately predict the price of a single-family home?

*The Analytics Question*

In our study, we aim to understand the effect that historical home values, new construction rates, interest rates, and tax policies have on the value of a property. Our analytics question revolves around a quantitative outcome (price) and can be analyzed using quantitative methods, as the data will be numeric.

*The Datasets and Key Variables*

1. Zillow Housing Data – Home Value, New Constructions, Home Sales, Median Home Price
2. Federal Reserve Economic Data – Fed Interest Rate Levels
3. IRS Historical Data – Personal Income Tax Rates, Effective Tax Rates

*Methodology*

During our data cleaning and preparation, we will address missing values using median value imputation for numeric variables (e.g. Home Value). Any missing categorical values (e.g. City) will be removed. We will test and compare three different predictive models: OLS Regression, Random Forests, and Lasso Regression. For cross validation testing we will use a 70/30 split between training and testing data, respectively. Model performance will be evaluated using RMSE to capture prediction error and R² to evaluate model fit to determine the best predictive approach.