**PROJECT PROPOSAL: PREDICTIVE MODELING AND SPATIAL ANALYSIS OF HOUSING PRICES IN THE SAN FRANSICO BAY AREA**

* **Dataset overview:** The dataset uses a subset of housing sales data from Zillow’s Research platform, focusing specifically on properties in the San Francisco Bay Area. The dataset includes property-level details such as location (city, zip, coordinates), size (lot and built square footage), number of bedrooms, year built, date of sale, and sale price.
* **Context and motivation:**

The San Francisco Bay Area is one of the most competitive and expensive real estate markets in the world, shaped by tech industry growth, urban development, and socioeconomic disparities. While traditional models predict prices based on property features, this project goes further by:

* **Incorporating spatial trends** (proximity to tech hubs, public transit, or high-crime areas).
* **Exploring ethical implications** of algorithmic pricing (gentrification risks, bias in valuation).
* **Leveraging advanced techniques** like **geospatial clustering** and **ensemble modeling** for higher accuracy.
* **Initial Observations:**   
   The dataset contains 8,000+ rows with several key features: Numerical: price, br (bedrooms), lsqft (lot size), bsqft (built area), year (built year), and geographic coordinates (lat, long). Categorical: county, city, zip, street. Missing data exists in several fields like bedrooms (br), lot size (lsqft), built square footage (bsqft), and year. Sale prices range widely, from below $100K to over $1M+, reflecting the region's diversity. A histogram of housing prices shows a right-skewed distribution, indicating many mid-priced homes and fewer high-end sales.
* **Objective of the Study** :

The primary goal is to analyze and predict housing prices using property attributes and location. This analysis will help uncover the key factors that drive home prices in the Bay Area and build a model that can estimate price based on those features.

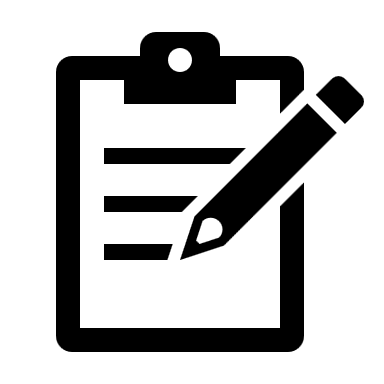
* **Key Research Questions:**

**Feature Importance:** Which structural (bedrooms, sqft) and locational (city, zip, proximity to key areas) factors most influence pricing?

**Spatial Trends:** Can we identify "hidden gem" neighborhoods where prices are undervalued relative to amenities?

**Model Robustness:** How do different imputation strategies (mean vs. predictive modeling) affect price predictions?

**Ethical Impact:** Does the model inadvertently favor certain demographics or neighborhoods, reinforcing housing inequality?

* **Hypotheses to Test:** 

✅ **H1:** Built square footage (bsqft) and lot size (lsqft) have a nonlinear relationship with price (diminishing returns after a threshold).  
✅ **H2:** Proximity to tech hubs (e.g., Palo Alto, SF) increases prices, but crime rates and school quality modulate this effect.  
✅ **H3:** Tree-based models (Random Forest) outperform linear regression due to complex feature interactions.  
✅ **H4:** Imputing missing values via **k-NN or MICE** yields better results than simple mean/median imputation.

* **Methodology:**

**📊 Data Preprocessing**

* Missing Data Handling**:** Compare traditional imputation (mean/median) vs. machine learning-based imputation (MICE, k-NN).
* Outlier Treatment: Winsorize extreme prices to reduce skewness.

**🌍 Geospatial Feature Engineering**

* **Proximity Metrics:** Calculate distances to key landmarks (tech campuses, parks etc.,).
* **Neighborhood Clustering:** Apply **DBSCAN or K-means** on coordinates to detect housing price hotspots.
* **Interactive Maps:** Use **Power BI or Tableau** to visualize price distributions and trends.

**🤖 Modeling Approach:**

| **Model** | **Use Case** |
| --- | --- |
| **Linear Regression (Baseline)** | Interpretability, feature coefficients |
| **Lasso/Ridge Regression** | Handling multicollinearity |
| **Random Forest /** | Capturing nonlinear patterns |
|  |  |

**📈 Evaluation Metrics**

* **Primary:** R², RMSE, MAE
* **Secondary:** SHAP values (explainability), spatial autocorrelation (Moran’s I)
* **Expected Outcomes & Impact**

✔ **A predictive model** with **>85% R²** (if data permits).  
✔ **Interactive dashboards** showing price trends and undervalued areas.  
✔ **Policy insights** on affordability gaps and potential bias in automated valuations.

* **Ethical & Societal Considerations**
* **Bias Detection:** Check if the model systematically underprices homes in lower-income areas.
* **Fair Housing Implications:** Discuss how AI-driven pricing could affect housing accessibility.
* **Deliverables**

1. **Jupyter Notebook** with full analysis (EDA, modeling, visualizations).
2. **Interactive Web Dashboard:**(Power BI/ Tableau) for real estate trend exploration.
3. **Technical Report:**summarizing findings, model limitations, and ethical concerns.