12-Week Intern Game Plan for California Price Prediction

Week 1 — Orientation & Setup

- Read the Task Prompt doc and understand the goal.
- Install Python, Git, and IDE.
- Establish FTP connection (FileZilla or Python ftplib) to download the latest 7 to 12 months of CRMLSSold files from /raw/California.
- Review Trestle Property MetaData.pdf in /resources to understand feature definitions.
- Deliverable: Confirm dataset access; write a short note on what each key column means.

Week 2 — Data Exploration

- Load the 7 to 12 months of dataset into pandas.
- Explore distributions of ClosePrice, LivingArea, Bedrooms, Bathrooms, LotSize.
- Restrict analysis to PropertyType = Residential and PropertySubType = SingleFamilyResidence (per task doc).
- Deliverable: Jupyter notebook 01_exploration.ipynb with basic EDA plots.

Week 3 — Data Preprocessing

- Handle missing values (decide whether to drop, impute, or flag).
- Convert categorical fields to numeric (encoding).
- Normalize numerical features if needed.
- Create train/test split (most recent month = test set, first 5 months = train).
- Deliverable: 02_preprocessing.ipynb + cleaned CSV.

Week 4 — Baseline Model

- Train a Linear Regression as the first model.
- Evaluate using R^2 on the test set.
- Record baseline results.
- Deliverable: 03_baseline_model.ipynb.

Week 5 — Additional Models

- Try Decision Tree and Random Forest regressors.
- Compare their test R^2 against baseline.
- Document model behavior (strengths/weaknesses).

- Deliverable: 04_model_comparison.ipynb.

Week 6 — Feature Engineering (v1)

- Add new features: PPSF (ClosePrice / LivingArea), Bed/Bath ratios, property age.
- Re-train models with these features.
- Deliverable: Updated notebook + table comparing old vs new feature sets.

Week 7 — Advanced Models

- Try Gradient Boosting (e.g., XGBoost or LightGBM).
- Perform light hyperparameter tuning (depth, learning rate, n_estimators).
- Deliverable: 05_advanced_models.ipynb with test metrics.

Week 8 — Evaluation Expansion

- Compute metrics beyond R^2: MAPE and MdAPE.
- Summarize insights (e.g., which price bands perform better).
- Deliverable: 06_evaluation.ipynb + metrics_summary.csv.

Week 9 — OPTIONAL Simple Prediction App (Streamlit)

- Build a Streamlit app: user inputs LivingArea, Beds, Baths, LotSize → output predicted price.
- Load trained model with joblib/pickle.
- Deliverable: app.py (streamlit app).

Week 10 — Documentation

- Write a README describing: dataset source, preprocessing, models tested, best results.
- Document instructions to re-run the code and launch the app.
- Deliverable: README.md.

Week 11 — Practice Presentation

- Prepare a 5–8 slide deck summarizing: data, methods, models, evaluation, Streamlit demo.
- Rehearse with peers or mentors.
- Deliverable: Slide draft

Week 12 — Final Presentation & Handoff

- Deliver final Zoom presentation to stakeholders (arrange presentation time with Aidan).

- Submit final repo with: Python scripts (preprocessing, training, evaluation, prediction), Documentation (README, metadata notes), App (app.py).
- Deliverable: Final repo + slides + live demo.

Recurring Sample Weekly Rhythm

- Mon check-in: goals + blockers.
- Thu PR deadline: one meaningful PR per week.
- Fri review: give feedback.