End-to-End Project Execution Workflow for AlloyTower Inc.

Phase 1: Problem Definition

Objective: Build predictive models and analytics to support real estate decisions.

Steps:

1. **Define Prediction Goals:**
   - **Goals**:
     - Predict property valuations.
     - Future price forecasting while understanding the market trend.
   - **Tools**: Use **Git** and **GitHub** for version control and collaboration on goal definition and project documentation.
2. **Data Preprocessing (Expected from the DA)**
3. **Feature Selection:**
   - **Feature Identification**: Based on the clean data provided by the DA, the **Data Scientist (DS)** will identify the most relevant features to use for the predictive models. This will include:
     - **Correlation Analysis**: Identifying correlations between input features and target variables.
     - **Statistical Tests**: Using tests like ANOVA, Chi-Square, or Mutual Information to assess the significance of features.
     - **Feature Importance**: Leveraging models (e.g., Random Forest, XGBoost, Mutual Information) to calculate the feature importance.
     - **Dimensionality Reduction**: Using techniques like **PCA (Principal Component Analysis)** or **t-SNE** for reducing high-dimensional data when necessary.
     - **Tools**:

        - **Scikit-learn** for feature selection techniques.
        - **MLflow** for tracking feature selection experiments and results.

Phase 2: Model Training & Validation

Objective: Train models on historical data and validate their performance.

Steps:

1. **Model Training:**
    - **Data Splitting**: Split data into training and validation sets (e.g., 80/20 split).
    - **Training**: The **Data Scientist** will train models using the cleaned and preprocessed data with selected features, validated by the DA. Cross-validation techniques will be used to assess the model's generalization performance.
    - **Tools**:
        - **Scikit-learn** for model training and cross-validation.
        - **MLflow** for tracking experiments, model performance, and hyperparameters.
2. **Hyperparameter Tuning:**
    - Use techniques like **Grid Search** or **Random Search** to optimize model hyperparameters.
    - **Tools**:
        - **Scikit-learn** or **Optuna** for hyperparameter optimization.
        - **MLflow** for tracking hyperparameter search and tuning experiments.
3. **Model Evaluation:**
    - Evaluate models on relevant performance metrics:

        - **Regression models**: **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**.

        - **Classification models**: **Precision, Recall, F1-Score**.

    - **Model Comparison**: Compare different models to determine the best-performing model.
    - **Tools**:

        - **Scikit-learn** for evaluation metrics and model comparisons.
        - **MLflow** for logging model performance metrics.

Phase 3: Model Deployment & Continuous Improvement

Objective: Deploy models for real-time decision-making and continuously improve model accuracy.

Steps:

1. **Model Deployment:**

    - **Deploy the best-performing models** into the production environment, based on the validation results.

    - Integrate models with the real estate platform.

    - **Tools**:

        - **Docker** to containerize models for deployment.
        - **FastAPI** for building an API endpoint for the purpose of inference.
        - **Hugging Face** for serving models in the production environment.
        - **Streamlit/Gradio** to build interactive dashboards for real-time predictions.

2. **Monitor Model Performance:**

    - Continuously monitor the model's performance in production, ensuring it remains accurate as market conditions change.

    - **Tools**:

- **MLflow** for real-time monitoring of model performance.