Technical Assessment
Machine Learning Developer
Intern

1.1. Overview

In this assessment, you will perform Exploratory Data Analysis (EDA) on a Housing Prices dataset to uncover insights about the data, identify trends, and prepare it for potential modelling.
The goal is to demonstrate your understanding of data analysis, visualization, and reporting in Python using tools such as pandas, NumPy, matplotlib, and seaborn.
You are expected to complete the analysis in a Jupyter Notebook (.ipynb).
The dataset can be downloaded from here

1.2. Core Tasks (Required)

1. Dataset Acquisition
   - Download the dataset from the provided Kaggle link.
   - Load it into your notebook and provide a brief description of the dataset and its features.
2. Data Understanding
   - Display dataset structure (shape, data types, sample records).
   - Identify numerical and categorical features.
   - Report missing values and duplicate records.
3. Data Cleaning & Preparation
   - Handle missing or incorrect data appropriately
   - Encode categorical variables if necessary.
   - Ensure data types are consistent and meaningful.
4. Exploratory Data Analysis (EDA)
   - Generate descriptive statistics (mean, median, mode, variance, etc.).
   - Visualize distributions of key variables
   - Explore relationships between features and the target variable (**SalePrice**)
   - Identify correlations
5. Insights & Interpretation
   - Summarize your findings — which factors appear to influence house prices the most
   - Discuss potential issues or patterns
   - Provide recommendations for future modelling or feature engineering
6. Notebook Organization & Presentation.
   - Use clear **section headings** and **markdown explanations** throughout the notebook.
   - Include **comments** in your code for clarity.
   - Ensure the notebook runs from top to bottom without errors.

1.3. Stretch Tasks (Optional)

1. Feature Engineering (Optional)
   - Create new features that could help improve predictive modelling.
2. Preliminary Modelling
   - Train a simple regression model (e.g., Linear Regression) to predict SalePrice using cleaned data.
   - Report model performance using metrics like RMSE or R^2.

1.4. Deliverables

1. A single Jupyter Notebook file (.ipynb) containing:
   - All EDA steps
   - Visualization and summaries
   - Conclusions and insights
2. README.md with brief setup instructions and overview.

1.5. What Will Be Assessed

1. Data Understanding – Ability to describe dataset structure and identify key issues.
2. Data Cleaning – Handling of missing values, types, and preprocessing quality.
3. EDA & Visualization – Use of plots, descriptive statistics, and exploration depth.
4. Insights & Interpretation – Quality and clarity of conclusions.
5. Notebook Quality – Code readability, comments, and narrative flow.

1.6. Grading Process

1. The examiner will review your Jupyter Notebook by running it from start to finish.
2. Evaluation will be based on:
   - Data Understanding & Cleaning (30%)
   - EDA & Visualization (30%)
   - Insights & Interpretation (20%)
   - Notebook Quality (20%)

1.7. Important Notes

1. Expected time: ~6-8 hours total
2. It's acceptable not to complete every optional task. Completing Tasks 1–6 is sufficient.
3. Clearly document all steps and assumptions in markdown cells.
4. Ensure the notebook runs without errors from start to finish.
5. Save and submit your work as a **.ipynb** file.