Housing Prices EDA Project

Machine Learning Developer Intern

Comprehensive Project Specification

Version 1.0

*Sybrin - Confidential*

# 1. Executive Summary

This document provides a comprehensive specification for the Machine Learning Developer Intern technical assessment. The project focuses on conducting Exploratory Data Analysis (EDA) on the Housing Prices dataset to demonstrate proficiency in data analysis, visualization, and Python programming.

Candidates will analyze real-world housing data to uncover meaningful insights, identify patterns and correlations, and prepare the data for potential machine learning applications. This assessment evaluates both technical competencies and professional documentation practices essential for a data science role.

# 2. Project Overview

## 2.1 Background

Understanding housing price dynamics is a fundamental problem in real estate analytics and machine learning. This project uses a comprehensive housing dataset containing numerous features that influence property values, including structural characteristics, location attributes, and quality metrics.

## 2.2 Project Scope

The scope of this assessment includes:

* Acquisition and loading of the Housing Prices dataset
* Comprehensive data understanding and profiling
* Data cleaning and preprocessing
* Statistical analysis and visualization
* Insight generation and interpretation
* Professional documentation in Jupyter Notebook format
* (Optional) Feature engineering and preliminary predictive modeling

## 2.3 Expected Duration

The total estimated time for completion is 6-8 hours. This includes data exploration, analysis, visualization, and documentation. Candidates are encouraged to manage their time effectively and focus on delivering quality insights rather than rushing through all optional components.

# 3. Project Objectives

## 3.1 Primary Objectives

1. **Data Proficiency:** Demonstrate ability to load, inspect, and understand complex datasets
2. **Data Quality Management:** Identify and appropriately handle missing values, duplicates, and data type inconsistencies
3. **Statistical Analysis:** Generate and interpret descriptive statistics to understand data distributions
4. **Data Visualization:** Create meaningful visualizations that effectively communicate patterns and relationships
5. **Analytical Thinking:** Identify key factors influencing house prices and formulate data-driven insights
6. **Professional Documentation:** Produce clear, well-organized, and reproducible analysis

## 3.2 Learning Outcomes

Upon completion, candidates will have demonstrated:

* Practical application of Python data science libraries
* Understanding of EDA best practices and methodologies
* Ability to extract actionable insights from raw data
* Skills in preparing data for machine learning pipelines
* Competency in technical communication and reporting

# 4. Technical Requirements

## 4.1 Required Tools and Libraries

**Programming Language:**

* Python 3.7 or higher

**Core Libraries:**

* **pandas:** Data manipulation and analysis
* **NumPy:** Numerical computing and array operations
* **matplotlib:** Basic plotting and visualization
* **seaborn:** Statistical data visualization

**Development Environment:**

* Jupyter Notebook or JupyterLab

**Optional Libraries (for stretch tasks):**

* **scikit-learn:** Machine learning and model evaluation
* **plotly:** Interactive visualizations (optional enhancement)

## 4.2 Dataset Information

**Source:** Kaggle Housing Prices Dataset

**Access Link:** Provided in assessment document (refer to "here" link)

**Format:** CSV (Comma-Separated Values)

**Target Variable:** SalePrice (continuous numeric value representing house sale price)

**Features:** Multiple numerical and categorical variables describing house characteristics

## 4.3 System Requirements

* Minimum 4GB RAM (8GB recommended)
* Python environment with package management (pip or conda)
* Internet connection for dataset download and library installation

# 5. Detailed Task Breakdown

## 5.1 Core Tasks (Required)

### Task 1: Dataset Acquisition and Loading

**Objective:** Successfully obtain and load the housing dataset into your analysis environment.

**Requirements:**

1. Download the dataset from the provided Kaggle link
2. Load the dataset using pandas (pd.read\_csv)
3. Provide a brief written description of the dataset in markdown
4. Document the features/columns present in the dataset

**Expected Output:** Loaded DataFrame with clear documentation of dataset structure and purpose.

### Task 2: Data Understanding

**Objective:** Thoroughly understand the dataset's structure, content, and quality.

**Requirements:**

1. Display dataset shape (number of rows and columns)
2. Show data types for each column (df.dtypes, df.info())
3. Display sample records (df.head(), df.sample())
4. Identify and list numerical features
5. Identify and list categorical features
6. Report missing values (df.isnull().sum())
7. Check for and report duplicate records (df.duplicated().sum())

**Expected Output:** Comprehensive overview tables and summaries showing dataset characteristics and data quality issues.

### Task 3: Data Cleaning and Preparation

**Objective:** Prepare the dataset for analysis by handling quality issues.

**Requirements:**

* **Missing Data Handling:**
  + Assess the nature and extent of missing data
  + Choose appropriate strategies (deletion, imputation, flagging)
  + Document and justify your chosen approach
* **Data Type Correction:**
  + Ensure numerical features are numeric types
  + Ensure categorical features are appropriate types
* **Categorical Encoding:**
  + Apply encoding if necessary for analysis (label encoding, one-hot encoding)

**Expected Output:** Clean dataset ready for analysis, with documented preprocessing steps and rationale.

### Task 4: Exploratory Data Analysis (EDA)

**Objective:** Uncover patterns, relationships, and insights through statistical analysis and visualization.

**Requirements:**

* **Descriptive Statistics:**
  + Generate summary statistics (mean, median, mode, std, variance)
  + Identify outliers and extreme values
  + Analyze value ranges and distributions
* **Distribution Analysis:**
  + Create histograms for key numerical variables
  + Generate box plots to identify outliers
  + Examine distribution of the target variable (SalePrice)
* **Relationship Analysis:**
  + Create scatter plots showing features vs SalePrice
  + Generate pair plots for highly correlated features
  + Analyze categorical variable impacts using bar plots or box plots
* **Correlation Analysis:**
  + Calculate correlation matrix for numerical features
  + Create correlation heatmap visualization
  + Identify features most strongly correlated with SalePrice

**Expected Output:** Rich collection of visualizations and statistical summaries revealing data patterns and relationships.

### Task 5: Insights and Interpretation

**Objective:** Synthesize analysis findings into actionable insights and recommendations.

**Requirements:**

* Identify the top 5-10 factors that most influence house prices
* Discuss any surprising or counterintuitive findings
* Highlight potential data quality issues or anomalies discovered
* Identify patterns or trends in the data
* Provide recommendations for feature engineering
* Suggest next steps for predictive modeling

**Expected Output:** Clear, well-reasoned summary of findings with business implications and technical recommendations.

### Task 6: Notebook Organization and Presentation

**Objective:** Ensure the notebook is professional, readable, and reproducible.

**Requirements:**

* Use clear, descriptive section headings (Markdown headers)
* Provide markdown explanations before and after code cells
* Include comments within code for clarity
* Ensure all visualizations have titles, axis labels, and legends
* Follow a logical flow from data loading through insights
* Test that the notebook runs completely without errors

**Expected Output:** Professional, publication-ready Jupyter Notebook that tells a clear data story.

## 5.2 Stretch Tasks (Optional - Bonus Points)

### Optional Task 1: Feature Engineering

**Objective:** Create new features that could enhance predictive model performance.

**Suggested Approaches:**

* Create interaction features (e.g., TotalSF = GrLivArea + TotalBsmtSF)
* Generate polynomial features for non-linear relationships
* Bin continuous variables into categories
* Create age-related features (e.g., HouseAge = YrSold - YearBuilt)
* Extract temporal features from date columns

**Expected Output:** Documented new features with rationale for their creation and potential impact.

### Optional Task 2: Preliminary Modeling

**Objective:** Build a baseline predictive model to validate the prepared data.

**Requirements:**

* Split data into training and testing sets
* Train a Linear Regression model as a baseline
* Calculate performance metrics (RMSE, MAE, R²)
* Visualize predicted vs actual values
* Discuss model performance and potential improvements

**Expected Output:** Functional baseline model with performance metrics and analysis.

# 6. Project Deliverables

## 6.1 Primary Deliverable: Jupyter Notebook

**File Format:** .ipynb (Jupyter Notebook)

**Naming Convention:** housing\_eda\_[YourName].ipynb

**Content Requirements:**

* Title cell with project name and author information
* All required EDA tasks (Tasks 1-6)
* Visualizations and statistical outputs
* Markdown documentation and code comments
* Conclusions and insights section
* Optional: Feature engineering and modeling sections

## 6.2 Supporting Deliverable: README File

**File Format:** README.md (Markdown)

**Content Requirements:**

* **Project Title and Overview:** Brief description of the project
* **Setup Instructions:** How to install required libraries
* **Execution Instructions:** How to run the notebook
* **Dataset Information:** Where to obtain the dataset
* **Key Findings Summary:** 3-5 bullet points of main insights
* **Author Information:** Your name and contact details

## 6.3 Deliverable Quality Standards

* **Completeness:** All required tasks must be addressed
* **Reproducibility:** Notebook must run from start to finish without errors
* **Clarity:** Code and explanations should be clear and understandable
* **Professionalism:** Proper formatting, no spelling errors, organized structure
* **Insights:** Demonstrates analytical thinking and domain understanding

# 7. Assessment Criteria and Grading

## 7.1 Evaluation Dimensions

The submission will be evaluated across five key dimensions, each contributing to the overall assessment score:

| **Evaluation Criterion** | **Weight** | **Key Evaluation Points** |
| --- | --- | --- |
| **Data Understanding & Cleaning** | **30%** | * Comprehensive dataset profiling * Appropriate handling of missing values * Correct data type management * Proper preprocessing methodology |
| **EDA & Visualization** | **30%** | * Quality and relevance of visualizations * Appropriate statistical analysis * Exploration depth and thoroughness * Correlation and relationship analysis |
| **Insights & Interpretation** | **20%** | * Quality of conclusions drawn * Business relevance of insights * Recommendations for modeling * Critical thinking and analysis |
| **Notebook Quality** | **20%** | * Code readability and structure * Documentation completeness * Narrative flow and organization * Professional presentation |

## 7.2 Grading Process

1. **Initial Review:** The examiner will run the notebook from top to bottom to verify it executes without errors
2. **Content Assessment:** Each section will be evaluated against the specific criteria listed above
3. **Quality Review:** Overall presentation, professionalism, and attention to detail
4. **Bonus Consideration:** Optional tasks may provide additional points above the base score
5. **Final Scoring:** Weighted scores from each dimension are combined for a total assessment score

## 7.3 Success Criteria

A successful submission will demonstrate:

* Technical competence in Python data analysis libraries
* Strong understanding of EDA methodologies
* Ability to extract meaningful insights from data
* Professional documentation and communication skills
* Attention to detail and code quality

# 8. Timeline and Milestones

## 8.1 Recommended Work Schedule

While candidates may organize their time as they prefer, the following schedule is recommended:

| **Phase** | **Time** | **Activities** |
| --- | --- | --- |
| **Setup & Acquisition** | 0.5 - 1 hour | Download dataset, setup environment, install libraries, initial data loading |
| **Data Understanding** | 1 - 1.5 hours | Explore dataset structure, identify data types, check for missing values, understand features |
| **Data Cleaning** | 1 - 1.5 hours | Handle missing values, correct data types, encode categories, remove duplicates |
| **EDA & Visualization** | 2 - 3 hours | Generate statistics, create visualizations, analyze distributions, explore correlations |
| **Insights & Documentation** | 1 - 1.5 hours | Synthesize findings, write insights, create recommendations, add markdown documentation |
| **Review & Polish** | 0.5 - 1 hour | Test notebook execution, add comments, improve formatting, create README |

## 8.2 Submission Deadline

Please refer to your assessment invitation email for the specific submission deadline. Late submissions may incur penalties or may not be accepted, depending on circumstances.

# 9. Best Practices and Tips

## 9.1 Code Quality Guidelines

* **Use descriptive variable names:** Choose meaningful names (e.g., 'avg\_sale\_price' not 'x')
* **Comment your code:** Explain what complex operations do and why
* **Follow PEP 8 style:** Use consistent formatting and spacing
* **Keep cells focused:** Each cell should perform one logical task
* **Avoid code repetition:** Use functions for repeated operations

## 9.2 Visualization Guidelines

* Always include titles and labels on all plots
* Choose appropriate plot types for your data
* Use color purposefully to highlight important information
* Ensure plots are readable (proper sizing, font sizes)
* Add brief interpretations after each visualization

## 9.3 Documentation Tips

* Use markdown headers to organize sections logically
* Write explanations before showing code or results
* Explain your reasoning for analytical decisions
* Summarize key findings at the end of each major section
* Include a final conclusions section tying everything together

## 9.4 Common Pitfalls to Avoid

* **Insufficient data exploration:** Don't rush through EDA; take time to truly understand the data
* **Over-complicated visualizations:** Keep plots simple and focused on one message
* **Missing data rationale:** Always explain why you chose specific handling strategies
* **Lack of insights:** Don't just show numbers and plots; interpret what they mean
* **Notebook doesn't run:** Always test from top to bottom before submitting
* **Poor organization:** Ensure logical flow from data loading to conclusions

# 10. Submission Guidelines

## 10.1 Pre-Submission Checklist

Before submitting, verify the following:

| **✓** | **Checklist Item** |
| --- | --- |
| ☐ | Notebook runs from start to finish without errors |
| ☐ | All six core tasks are completed |
| ☐ | Markdown cells provide clear explanations throughout |
| ☐ | All visualizations have titles, labels, and legends |
| ☐ | Code includes appropriate comments |
| ☐ | Insights and conclusions section is complete |
| ☐ | README.md file is included with setup instructions |
| ☐ | File names follow the specified naming convention |
| ☐ | No spelling or grammatical errors in documentation |
| ☐ | Author information is included in the notebook and README |

## 10.2 Submission Format

**Package Contents:**

* housing\_eda\_[YourName].ipynb (required)
* README.md (required)

**Submission Method:**

Follow the submission instructions provided in your assessment invitation email. Typically, this will involve uploading files to a designated portal or sending them to a specified email address.

## 10.3 Support and Questions

If you encounter any issues or have questions about the assessment:

* Review this specification document thoroughly
* Consult Python and library documentation
* Contact the assessment coordinator using the email provided in your invitation

# 11. Appendices

## 11.1 Sample README Template

Below is a template you can use for your README.md file:

---

**# Housing Prices EDA Project**

**## Overview**

Brief description of the project and its objectives.

**## Dataset**

Information about the Housing Prices dataset and source.

**## Setup Instructions**

1. Install required libraries: pip install pandas numpy matplotlib seaborn

2. Download the dataset from [link]

3. Open housing\_eda\_[YourName].ipynb in Jupyter

**## Key Findings**

- Finding 1

- Finding 2

- Finding 3

**## Author**

[Your Name] - [Contact Email]

---

## 11.2 Useful Python Code Snippets

**Data Loading:**

import pandas as pd

df = pd.read\_csv('housing\_data.csv')

**Missing Values Analysis:**

missing = df.isnull().sum()

missing\_percent = (missing / len(df)) \* 100

**Correlation Heatmap:**

import seaborn as sns

correlation = df.corr()

sns.heatmap(correlation, annot=True)

## 11.3 Recommended Resources

* **Pandas Documentation:** https://pandas.pydata.org/docs/
* **Seaborn Tutorial:** https://seaborn.pydata.org/tutorial.html
* **Matplotlib Gallery:** https://matplotlib.org/stable/gallery/
* **Scikit-learn Documentation:** https://scikit-learn.org/stable/

**Good luck with your assessment!**

*We look forward to reviewing your work.*