Report Outline

1. Title Page

- Project Title
- Team Members
- Course & Instructor
- Submission Date

2. Executive Summary

- Objectives
- Key Findings
- Recommendations

3. Introduction

3.1 Background & Motivation

- We are investigating: “Which variables are most relevant for influencing a property’s price?”
- Our goal is to guide new real-estate investors and to build a predictive model: “Can we predict house price given input data?”

3.2 Research Questions & Hypotheses

- **Research Question:** What are the most relevant variables affecting property prices, and how effectively can we forecast house prices based on these variables?
- **Hypotheses:**
- House size, number of rooms, and neighborhood location will strongly influence price.
- Unexpected features may also exert significant effects.
- Weak individual variables might combine as powerful predictors.

4. Dataset Description

4.1 Source & Access

- **Original Source:** Kaggle “House Prices: Advanced Regression Techniques” competition
- **Access URL:** https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- **File Used:** `train.csv` (1,460 rows × 81 columns)

4.2 Structure & Content

- **Rows:** 1,460
- **Columns:** 81
- **Numerical:** `LotArea` (int), `YearBuilt` (int), `GrLivArea` (int), `SalePrice` (int) (target), etc.

- **Categorical:** `MSZoning` (object), `Street` (object), `SaleCondition` (object), etc.
- **Ordinal:** `OverallQual` (1–10), `OverallCond` (1–10), `BsmtQual` (Ex, Gd, TA, Fa, Po)

4.3 Data Cleaning & Normalization

- **Missing Values:**
- Imputed `LotFrontage` by median value per `Neighborhood`
- Replaced “NA” in categorical fields (e.g., `Alley`, `FireplaceQu`) with "None"
- Dropped columns with >50% missing values (e.g., `PoolQC`, `Fence`)
- **Type Conversions:**
- Ensured year fields (`YearBuilt`, `YrSold`) are numeric
- Mapped ordinal categories to integer scales (e.g., `BsmtQual`, `ExterQual`)
- **Normalization:**
- Applied min–max scaling to size features (`GrLivArea`, `LotArea`) for consistency in modeling

4.4 Computed Metrics

- **HouseAge:**

```
HouseAge = YrSold - YearBuilt
```

- **TotalBath:**

```
TotalBath = FullBath + 0.5 * HalfBath + BsmtFullBath + 0.5 * BsmtHalfBath
```

- **PricePerSqFt:**

```
PricePerSqFt = SalePrice / GrLivArea
```

- **RemodelAge:**

```
RemodelAge = YrSold - YearRemodAdd
```

5. Exploratory Data Analysis (EDA)

5.1 Univariate Analysis

5.1.1Key Variable Distributions

- Describe the distribution, central tendency, and spread for each key numeric variable (e.g., SalePrice, GrLivArea).

5.1.2 Charts Used (histograms, box plots, etc.)

- List and describe the charts chosen (e.g., histogram of SalePrice, box plot of GrLivArea).

5.2Bivariate & Multivariate Analysis

5.2.1 Relationships Between Variables

- Outline which variable relationships were investigated (e.g., SalePrice vs. OverallQual, PricePerSqFt vs. Neighborhood).

5.2.2 Charts Used (scatter plots, heatmaps, violin plots, etc.)

- List and describe charts used to explore these relationships (e.g., scatter plot of Carat vs. Price by Cut, correlation heatmap).

6. Key Insights

6.1 Insight 1 (description + supporting chart)

- Provide the first major finding and reference the visualization that illustrates it.

6.2 Insight 2 (description + supporting chart)

- Provide the second major finding and reference its supporting chart.

6.3 Insight 3 (description + supporting chart)

- Provide the third major finding and reference its supporting chart.

7. Data Story & Narrative

7.1 Logical Flow of Findings

- Summarize the progression from EDA to insights in a coherent narrative.

7.2 Central Take-Home Message

- State the key message that readers should remember.

8. Recommendations & Next Steps (optional)

8.1 Follow-Up Questions

- List additional questions prompted by the analysis.

8.2 Additional Data Needed

- Describe any further data required to answer those questions.

8.3 Proposed Actions

- Suggest possible actions or decisions based on findings.

9. Conclusion

9.1 Recap of Objectives & Findings

- Restate goals and summarize key findings concisely.

9.2 Limitations & Applicability

- Note dataset limitations and applicability of results to other contexts.

10. References

10.1 Data Sources

- Cite the original dataset URL and any supplementary data sources.

10.2 Tools & Libraries

- List key tools, libraries, and frameworks used (e.g., pandas, seaborn, matplotlib).

11. Appendices

11.1 Jupyter Notebook Link or File

- Provide link or filename for the notebook with complete code.

11.2 Dataset Sample (≤1,000 rows)

- Include a sample extract or link.

11.3 Supplemental Code Listings

- Provide any functions or scripts referenced in the report.