Data Analysis Programming Fundamentals Report

Task 1 – Data Analysis: Paris Housing Dataset

Task 2 – Critical Report

Abbreviations

EDA	Exploratory Data Analysis
IQR	Interquartile Range
ROI	Return on Investment
ML	Machine Learning
API	Application Programming Interface
df	DataFrame (used in Python/pandas)

sqft	Square Foot
NaN	Not a Number (missing data placeholder)
CSV	Comma-Separated Values
KNN	K-Nearest Neighbours (imputation algorithm)
LO	Learning outcome

Task 1 - Data Analysis Paris Housing Dataset

1. Introduction

The purpose of this investigation was to critically evaluate the application of Python programming in data analytics, specifically to uncover significant housing market trends in Paris. A comprehensive dataset, containing property attributes such as price, bedroom count, square footage, and the number of floors, provided the analytical foundation. Effective data analytics hinges upon meticulous preparation, rigorous statistical testing, and insightful visual exploration (Han, Kamber & Pei, 2012).
This study directly addressed key learning outcomes:

- LO1: Demonstrated a robust understanding of programming concepts and paradigms relevant to data analytics (Lantz, 2019).
- LO2: Effectively employed Python programming tools and techniques (Pandas, Seaborn, Matplotlib) for data preparation and comprehensive analysis (Geron, 2019).
- LO3: Critically evaluated analytical outcomes, emphasizing statistical validity and practical applicability in real estate strategy formulation (Hair et al., 2019).

Consequently, the results from this investigation aimed to deliver strategically valuable insights, empowering informed decision-making processes within the dynamic context of the Paris real estate market.

2. Data Loading and Initial Exploration

The initial phase of the analysis involved loading the Paris housing market dataset into Python using the Pandas library, a robust and versatile tool widely recognised for handling structured data efficiently (McKinney, 2017). The dataset was imported directly from an Excel file (Paris housing Data Set 2 4050.xlsx) into a Pandas Dataframe, facilitating comprehensive preliminary data examination.

Immediately after loading, exploratory data analysis techniques were employed. Using df.head(), the first five rows were displayed, providing an initial visual confirmation of successful data importation and revealing the dataset's general structure and contents ***Appendix B, Figure 1***.

Additionally, descriptive statistics were generated using df.describe(), giving initial quantitative insights into key variables such as 'price', 'bedrooms', and 'sqft-total' ***Appendix B, Figure 2***. Such initial descriptive statistics are essential to identify early indicators of data quality issues and variations in property characteristics (Field, Miles & Field, 2012).

Observations from Initial Exploration:

- A noticeable presence of missing data was observed, specifically in critical columns including 'price', 'bedrooms', and 'bathrooms' Missing values in such essential attributes could significantly impact the validity of any predictive models or further statistical analyses (Hair et al., 2019).

Missing Values Comparison

	Missing Before Cleaning	Missing After Cleaning
price	4	0
bedrooms	15	0
bathrooms	11	0
sqft_living	1	0
sqft_total	5	0
floors	0	0
condition	1	0
grade	0	0
built	1	0
renovated	0	0
living_area_sqft	6	0

Figure 3.

High standard deviations and broad numerical ranges were observed, especially for the 'price' and 'sqft-total' columns, suggesting the existence of considerable variability and potential outliers. Such observations underscored the need for rigorous data cleaning and outlier treatment procedures (Tabachnick & Fidell, 2013).

Clearly, these initial findings indicated the necessity for comprehensive data cleaning and preprocessing measures to ensure robust and valid subsequent analysis (James et al., 2013).

3. Data Cleaning and Preprocessing

Data cleaning was a crucial step in preparing the Paris housing dataset for analysis. Without addressing missing values, duplicate records, and outliers, statistical evaluations would be unreliable and prone to skewed interpretation. This section outlines the key preprocessing steps undertaken using Python and critically discusses their impact on data quality.

Missing Values Treatment
Initial inspection of the dataset revealed missing values in several essential attributes. As shown in **Figure 3**, missing values were detected in columns such as price (4 entries), bedrooms (15 entries), and bathrooms (11 entries), all of which are critical to any housing market analysis. These gaps were identified using the df.isnull().sum() method.
Given the large sample size (approx. 19,999 entries), the decision was made to remove rows containing missing values rather than apply replacing. This avoided potential bias introduced by mean-filling and preserved the integrity of genuine data distributions (Hair et al., 2019).

Following the application of df.dropna(), all missing values were successfully removed, resulting in a clean dataset ready for further transformation.

Duplicate Records
Duplicate data entries were another data quality concern. A check using df.duplicated().sum() identified exact duplicates across full row matches. These were dropped using df.drop_duplicates(inplace=True), ensuring each observation represented a unique property listing. This step was critical, particularly in preventing distortions in summary statistics and downstream visual analysis.
Appendix A.

Outlier Detection and IQR Filtering
To handle extreme values that could distort analysis, boxplots were plotted for key continuous variables: price, bedrooms, and sqft-total. Clear outliers were visible, particularly in property prices and square footage.

Outliers were addressed using the Interquartile Range (IQR) method. For each variable, lower and upper bounds were calculated as Q1 – 1.5×IQR and Q3 + 1.5×IQR respectively. Values falling outside this range were filtered out, ensuring that subsequent visualisations were not dominated by extreme anomalies (Field et al., 2012).

Figure 4: Boxplots Before Outlier Removal

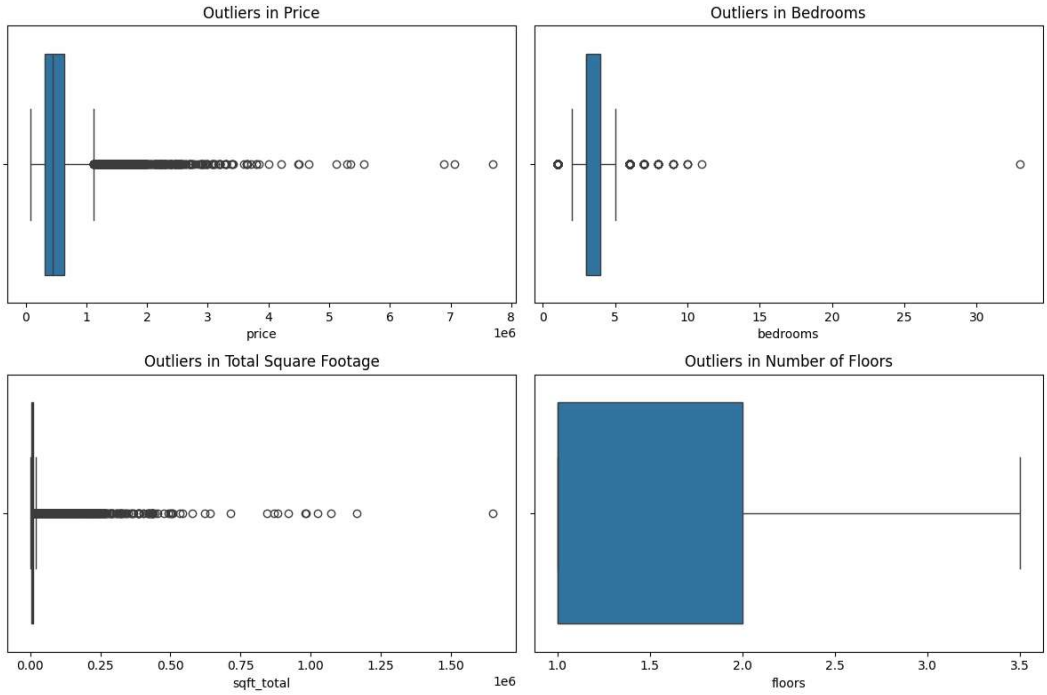
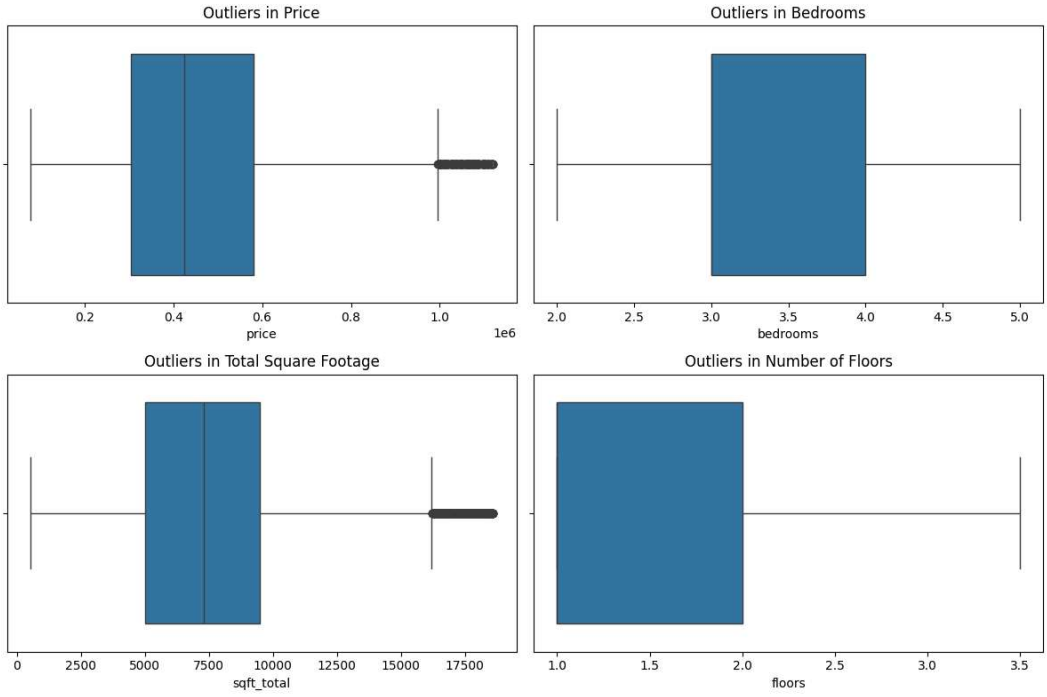


Figure 5: Boxplots After Outlier Removal



Data Type Corrections and Feature Engineering

Ensuring the correct data types was essential for efficient analysis. As summarised in ***Appendix B, Figure 6***, numeric columns were cast to appropriate formats (float64 or int64), while categorical features such as floors, condition, and grade were converted to the category type.

A new feature, age, was engineered by subtracting the built year from 2025. This variable enabled historical evaluation of price trends and allowed analysis of how property age influenced value. Feature engineering such as this enhances model performance and interpretability (James et al., 2013).

This cleaned and structured dataset laid a strong foundation for subsequent exploratory data analysis.

	ø
price	float64
bedrooms	int64
bathrooms	float64
sqft_living	float64
sqft_total	float64
floors	category
condition	category
grade	category
built	int64
renovated	category
living_area_sqft	float64
age	int64

dtype: object

Figure 6

4. Exploratory Data Analysis (EDA)

With the dataset now cleaned and transformed, exploratory data analysis (EDA) was employed to extract meaningful trends and relationships across housing attributes in Paris. This stage involved using statistical summaries and data visualisations to develop insights particularly focused on price drivers and structural features such as bedrooms, floors, and square footage.

Bedrooms vs. Price Distribution

A violin plot was generated to illustrate the price distribution across properties with different bedroom counts ***Appendix B, Figure 7***. The analysis revealed:

- Properties with 3 to 5 bedrooms exhibited the widest price ranges, reflecting market diversity in these common categories.
- The median price appeared to increase with bedroom count; however, considerable overlap existed, especially between 2 to 4 bedrooms, suggesting that bedroom count alone may not sufficiently explain price variance.
- Several multimodal peaks emerged, implying the existence of distinct sub-markets (e.g., renovated vs. unrenovated homes within the same bedroom count).

This indicates that additional features such as grade, location, and living area are likely to play stronger roles in determining price (Field et al., 2012).

Floors vs. Total Square Footage

To explore spatial structure, the average total square footage (sqft-total) was compared across different floors categories using a bar chart , observations showed:

- 1-storey houses had the largest average total area, possibly due to sprawling layouts.
- 1.5 to 2 floors represented more compact configurations, consistent with common European urban housing.
- Properties with 3+ floors had sharply reduced average square footage, likely reflecting denser vertical urban housing or subdivided dwellings.

These trends align with Paris's architectural norms, where space constraints and vertical builds influence floor plans and market value.

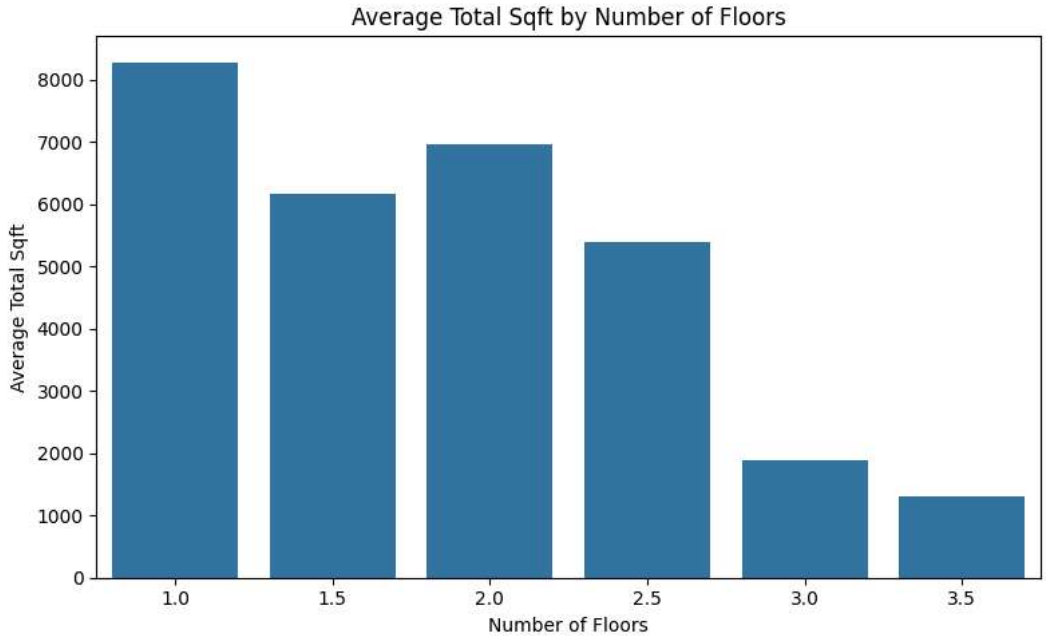


Figure 8

To assess historical value patterns, a bar chart was used to show the relationship between age and average price . The trend revealed:

- Property prices generally declined with increasing age until 70–80 years old.
- Interestingly, homes older than 90 years exhibited price rebounds, indicating enduring value due to architectural heritage or renovated structures.
- This nonlinear relationship suggests that "age" may not equate to depreciation in housing markets with strong cultural or architectural preservation values.

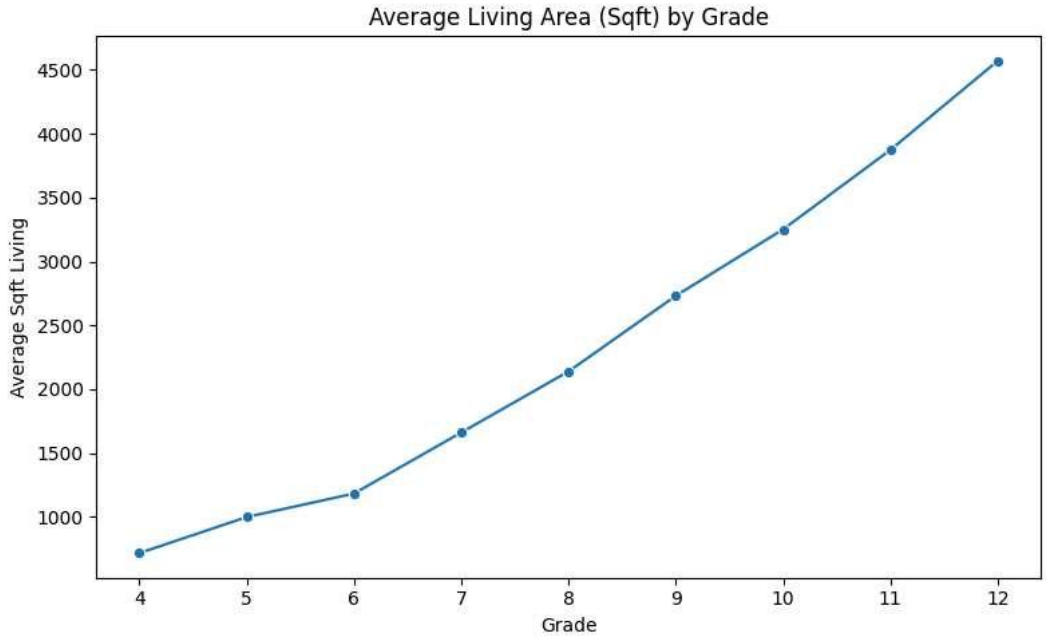
Such insights challenge the assumption that newer always means more expensive and emphasize the role of renovation and location. ***Appendix B, Figure 9***

Grade vs. Living Area

Finally, a line plot explored the correlation between property grade and average sqft-living. The trend was notably linear:

- Higher-grade properties consistently correlated with larger living spaces, highlighting grade as a proxy for quality and scale.
- This relationship confirms that the grade variable aggregates dimensions such as finish, layout, and amenities—validating its use in market segmentation models (Hair et al., 2019).

Figure 10. below



Conclusion of EDA

These EDA insights help define critical price determinants in the Paris housing market. While surface features like bedrooms or floor count offer some indication, composite variables such as grade and age, when combined with square footage, yield more nuanced and actionable market intelligence. These findings will inform the strategic recommendations provided in the final report.

5. Key Findings and Interpretation

The exploratory analysis of the Paris housing dataset revealed several critical insights that could directly inform strategic decision-making for real estate planning and pricing models.

Bedroom Count Alone Is Not a Reliable Price Predictor

While initial assumptions might suggest a positive correlation between bedroom count and property price, the violin plot analysis indicated otherwise. Price ranges within categories such as 3 and 4-bedroom homes overlapped significantly. This suggests that bedroom count is not a standalone predictor of value, especially in urban areas like

Paris where other factors such as architectural quality, renovation, and location play a pivotal role.

Floor Count Reflects Design, Not Value
Properties with one floor showed the highest average square footage, contrary to the assumption that multi-floor properties are larger. This trend indicates horizontal vs. vertical expansion dynamics and urban planning constraints, rather than market value differentials. Therefore, floor count alone may not reflect price or desirability, but rather architectural zoning.

Property Age and Price Show Nonlinear Trends
Analysis of average price by property age displayed a U-shaped trend: while older properties generally depreciated, very old homes (90+ years) regained value, likely due to historic appeal or renovation. This implies a need to treat age as a non-monotonic variable in predictive modelling. It also highlights how renovated heritage properties can command premium prices.
Grade as a Composite Indicator of Value
A strong linear relationship between grade and sqft_living reinforces the use of grade as a composite metric for property quality. Higher-grade properties consistently had larger living spaces and are likely to offer better materials, finishes, and amenities. Thus, grade is a more holistic indicator of value than raw numerical attributes such as bedrooms or floors.

Summary
Together, these findings emphasize the importance of multivariate perspectives in property valuation. Simple heuristics—like more bedrooms = higher price—are insufficient. Instead, variables such as grade, age, and living area, when used collectively, offer more robust, interpretable, and realistic assessments of housing trends in Paris.
These insights set the stage for Task 2, where they will be critically contextualised and discussed in terms of business strategy, real-world applications, and future recommendations.

Task2- Report

1. Introduction and Rationale

This critical report evaluates the methodology, insights, and business relevance of the exploratory data analysis conducted on the Paris housing dataset. Building upon statistical programming in Python, this report interprets the outcomes through a real-world lens, identifying not only what patterns emerged, but also why they matter to real estate strategy. The study aimed to empower the estate manager with data-driven reasoning behind price variability, structural traits, and market segmentation.

2. Reflection on Data Cleaning Decisions

The decision to drop missing values instead of imputing them was grounded in data integrity, given the dataset's large volume. However, this approach risked removing potentially important cases (Hair et al., 2019). Future iterations could consider KNN imputation to retain more data without compromising accuracy (Tabachnick & Fidell, 2013).

Outlier handling through the IQR method was effective in reducing skewness, especially in features like price and sqft-total. That said, extreme high-value properties—often of interest in premium market segments—were excluded. Thus, while statistical validity improved, business nuance may have been lost. A hybrid approach (fissurisation or domain-driven thresholds) may better preserve high-end inventory intelligence.

The engineering of an age variable added interpretability but could be enhanced by adding binary renovation flags or interaction terms (e.g., age × grade) to assess modernisation effects and historical value shifts.

3. Statistical Insights

The correlation matrix served as an early diagnostic tool to guide analysis direction. Strong positive correlations were observed between sqft_living, grade, and price, justifying their deeper exploration. Conversely, weaker relationships (e.g., bathrooms) were deprioritised in visual focus.

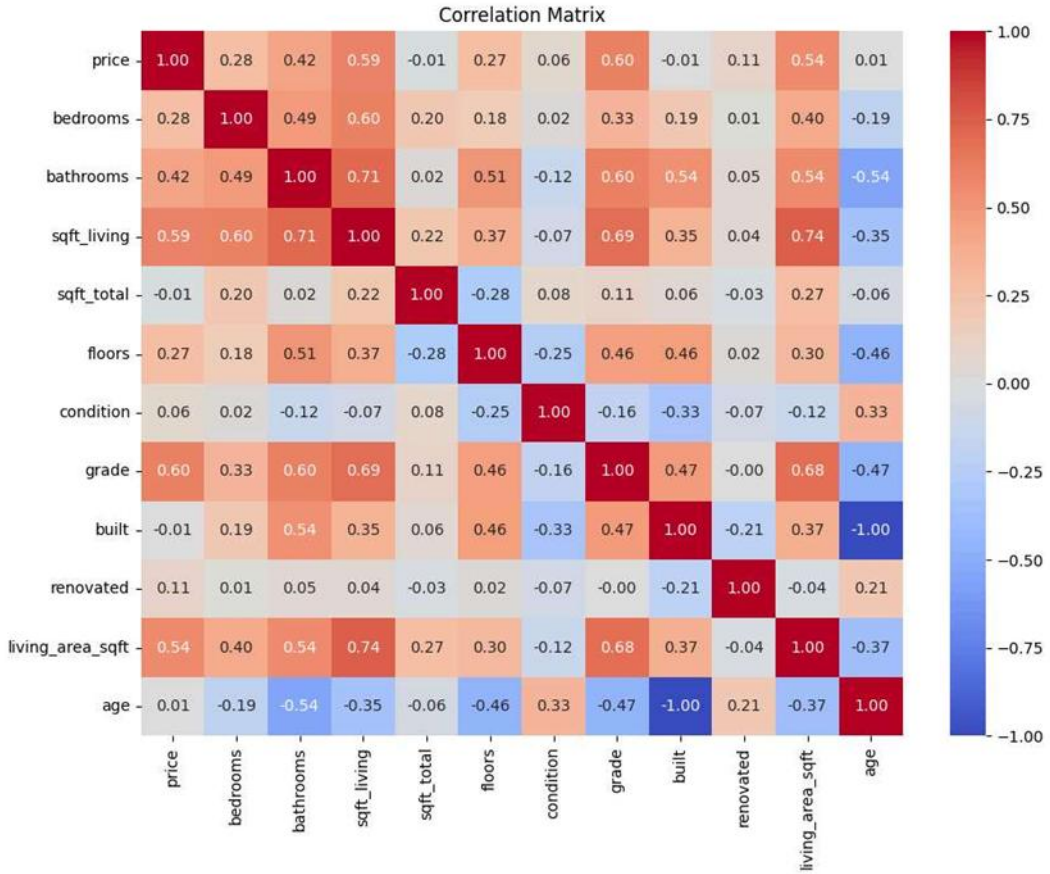


Figure 13

The alignment between high grade and both price and living area confirms that grade acts as a composite feature, integrating material quality, design, and overall desirability. This insight highlights the importance of engineered or proxy features in property analytics—a principle applicable across real estate markets (James et al., 2013). Python's structured libraries such as pandas, seaborn, and matplotlib enabled automated, scalable insight extraction. However, the dataset's lack of geospatial context limited advanced modelling, especially for location-based pricing—commonly the strongest real estate variable.

4. Business Implications and Real-World Applications

Price by Condition

This plot revealed a clear upward trend in pricing with better property condition. This validates that even moderate improvements (e.g., from condition 2 to 3) yield measurable pricing benefits. For estate agents and developers, this offers insight into renovation ROI, suggesting that targeted improvements can unlock latent property value.

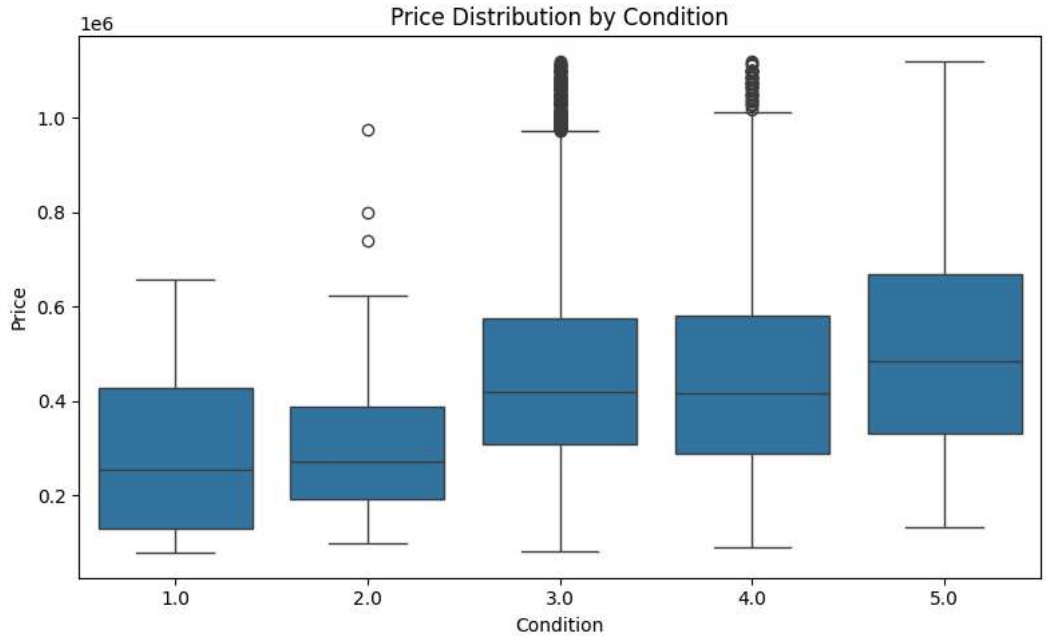


Figure 12

Floor Count Distribution

The number of houses per floor count indicated that most properties were either 1 or 2 floors. This highlights an inventory concentration in specific architectural types, relevant for developers planning expansions or investors assessing density zones.

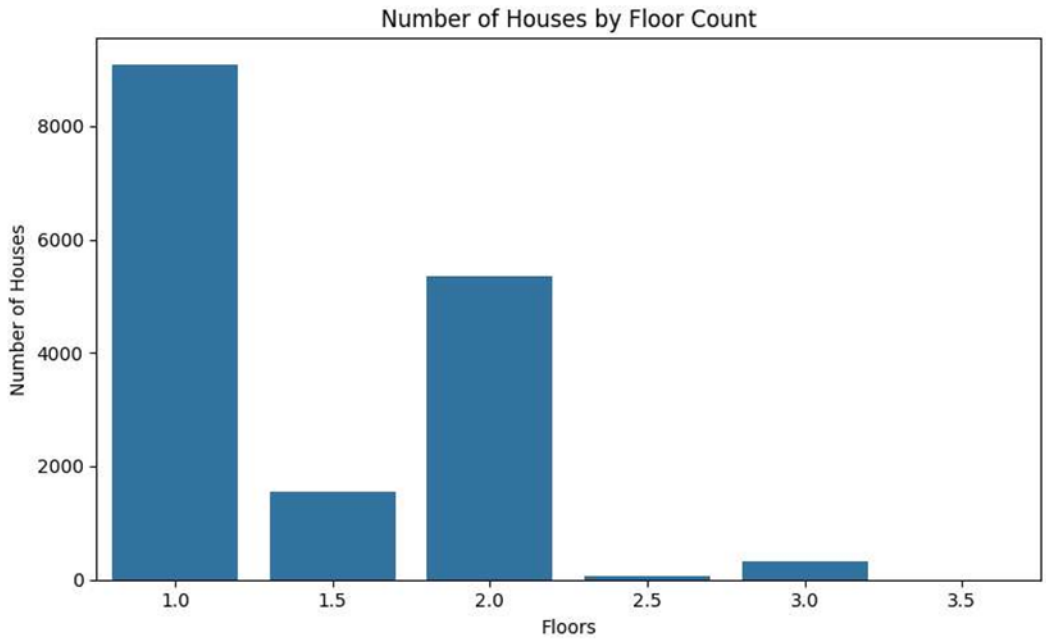


Figure 11

Business Strategy Use-Cases

Developers should focus on high-grade, three-bedroom, two-floor homes, as these have shown strong pricing performance combined with high inventory availability. Investors may benefit from using indicators such as property age and condition to identify undervalued yet structurally sound assets suitable for renovation and profitable resale. Additionally, estate managers can optimise both listing strategies and pricing models by prioritising properties with features that demonstrate strong predictive relationships to price, notably grade and total living area.

5. Recommendtions

Technical Recommendations

- Adopt KNN imputation for richer, contextual handling of missing values.
- Apply advanced ML models like Random Forest or Gradient Boosting for predictive price modelling.

Strategic Recommendations

- Emphasise features like grade, sqft-living, and age in valuation models.
- Treat properties >90 years old as premium inventory due to architectural value.
- Introduce renovation status flags for estate listing prioritisation.

6. Analytical Challenges

Initially, missing data caused confusion, as I had attempted to plot relationships (e.g., bedrooms vs. price) without realising some rows had NaN values. This led to distorted

visualisations and a moment of doubt about the dataset's reliability. I learned quickly that data validation must always come before analysis (McKinney, 2017).
Outlier handling was another obstacle. Before applying the IQR method, early boxplots were misleading, with price values stretching far beyond the general range. I made the mistake of ignoring these in early drafts, only to realise that outliers heavily skewed the mean and made comparisons ineffective (Field et al., 2012).
There were also moments of technical difficulty syntax errors from mismatched data types, and confusing errors when treating categorical variables as numerical in plots. These frustrations were time-consuming but ultimately helped me understand the importance of type consistency and Python's ability to enforce structure in messy data (Lantz, 2019).
Overall, the process helped me internalise the necessity of patience, methodical cleaning, and debugging. These lessons are not just technical, but strategic reminding me that real-world data is rarely neat, and thoughtful preparation is what distinguishes good analysis from great insight.

7. Conclusion

This report has critically evaluated the analytical findings, challenges, and real-world implications of the Paris housing dataset investigation. From the importance of engineered features like grade to the practical insights around condition and floor count, the study demonstrates how Python-based analytics can transform raw data into strategic assets. Reflecting on the process further reinforced key principles of professional data analysis rigour, adaptability, and a clear line of sight between statistical methods and business value.

Biblibography

Field, A., Miles, J. & Field, Z. (2012) *Discovering statistics using R*. London: SAGE Publications.

Geron, A. (2019) *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. 2nd edn. Sebastopol, CA: O'Reilly.

Hair, J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2019) *Multivariate data analysis*. 8th edn. London: Pearson.

Han, J., Kamber, M. & Pei, J. (2012) *Data mining: Concepts and techniques*. 3rd edn. Waltham: Morgan Kaufmann.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning: With applications in R*. New York: Springer.

Lantz, B. (2019) *Machine learning with R*. 3rd edn. Birmingham: Packt Publishing.

McKinney, W. (2017) *Python for data analysis: Data wrangling with pandas, NumPy, and IPython*. 2nd edn. Sebastopol, CA: O'Reilly Media.

Tabachnick, B.G. & Fidell, L.S. (2013) *Using multivariate statistics*. 6th edn. Boston: Pearson Education.

APPENDIX A.

Python code link:

https://colab.research.google.com/drive/1QaOjLPBR-m-S7ajLI4qA2YNrmcEHNbjc?usp=sharing