Project Proposal: Data Mining and Text Analytics Framework for Rental Price Prediction and Anomaly Detection in the UK Student Housing Market

1. Research hypothesis and objectives

1.1 Research hypothesis

This project hypothesises that a combined data mining and natural language processing (NLP) approach can effectively predict rental prices and detect anomalies in student housing markets. It further asserts that textual elements, like listing descriptions and mentioned amenities, significantly influence perceived value and price strategies.

It assumed that:

- Structured features of a listing, such as the property type, location, inclusion of bills and weekly rent prices have a significant impact on rental price and can be modelled using supervised machine learning algorithms (Chan, 2024).
- Unstructured data like free-text listing descriptions and amenities information will influence the perceived attractiveness and pricing of properties. These can be quantified using NLP methods such as TF-IDF, keyword extraction, and sentiment analysis (Zhao, 2024).
- Pricing anomalies listings that are significantly overpriced or underpriced compared to market norms can be detected using unsupervised anomaly detection techniques like Isolation Forest or statistical residual analysis (Sultan, 2023).

This hypothesis responds directly to real-world challenges identified in the United Kingdom student housing market, such as price segmentation, affordability concerns, and rent inflation.

1.2 Scientific Ambition and Novelty

This research is novel in its integration of structured and unstructured data within a unified model, applied to an underexplored domain: “The student housing market”. Unlike prior work focusing on general housing data (Chan, 2024), this project explores how listing description affects price and uses both predictive modelling and anomaly detection to promote housing transparency.

By following the CRISP-DM methodology, the project aims to deliver a scalable, data-driven framework for fair rent estimation and anomaly detection in student accommodation listings.

1.3 Overall Aim

The overall aim of the project is to develop a data-driven framework that can predict fair weekly rental prices for student accommodation listings, detect anomalous or potentially misleading pricing, and analyse how listing descriptions and amenities contribute to price formulation.

1.4 Research Objectives

1. Develop predictive models using machine learning algorithms to estimate weekly rental prices based on structured features such as location, property type, and inclusion of bills.
2. Apply anomaly detection methods such as residual-based techniques, to identify listings that significantly diverge from predicted market values.
3. Conduct text analytics on unstructured listing descriptions using TF-IDF, keyword extraction, and sentiment analysis to investigate how language and stated amenities influence rental prices. To enhance the extraction of key linguistic features from property descriptions, corpus analysis tools such as Sketch Engine will also be employed.
4. Validate the proposed project through a pilot study focused on student accommodation in Leeds, using Unipol data as a real-world case to test predictive accuracy and anomaly detection methods.

5. Propose a generalisable framework for extending the approach to student housing markets in other UK cities, with the aim of enhancing market transparency and fairness.

2. Background

The student rental housing market in the United Kingdom has become increasingly fragmented and financially burdensome, particularly in urban centres with large student populations. Rising rents, short-term leases, and limited regulatory oversight have left many students facing affordability challenges without adequate means to assess whether listed prices are justified. In cities like Leeds, where student demand is concentrated in key areas such as Hyde Park and Headingly, rents have risen sharply by over 120% in some zones over the last decade, intensifying the need for more transparent pricing frameworks (Unipol, 2024).

Existing platforms such as Unipol and UniHomes offer listing portals but lack benchmarking or anomaly detection features, which means students must rely on their intuition or informal networks when judging whether a property offers good value. This results in information asymmetry that disadvantages tenant and hinders market efficiency. Addressing this issue aligns with growing academic interest in applying data-driven approaches to social infrastructure problems, especially those affecting vulnerable or transitional populations such as students (NUS, 2024).

Despite the growing availability of rental listing data, academic utilisation remains limited, particularly in the context of student accommodation. Most existing studies focus on general real estate trends or home ownership, with minimal attention to transient or shared housing. For example, (Chan, 2024) used Random Forest and XGBoost to predict rent prices in Texas, using structured features such as square footage and location. While effective, this model excluded unstructured data such as property descriptions or advertised amenities, missing potentially influential factors that shape perceived value.

Similarly, (Zhao, 2024) developed a hybrid model using structured indicators from online Chinese rental platforms and advocated for future research to incorporate textual sentiment and amenity related language. The findings suggested that language used in listings significantly affected consumer behaviour and price elasticity but stopped short of integrating this into predictive or diagnostic systems. This gap has been recognised internationally but remains largely unaddressed in the UK student context.

Recent progress in natural language processing (NLP) offers methodological tools to close this gap. Techniques such as TF-IDF, topic modelling, and sentiment analysis can extract meaningful insights from unstructured text, revealing how amenities and narrative framing influence pricing strategies. Studies like (Sultan, 2023) introduced the Price Anomaly Score (PAS) based on model residuals to identify overpriced or underpriced listings, but this method has yet to be enriched with linguistic signals or applied to university cities.

The University of Leeds is well-positioned to lead this research due to its interdisciplinary strength in data science and social policy. The Leeds Institute for Data Analytics (LIDA) has contributed to a range of urban infrastructure projects, including transport and health inequalities, but has not yet explored student rental markets using machine learning or NLP approaches. This proposal aims to fill that gap by developing a scalable framework that combines structured features (e.g. location, property type, rent inclusions) with unstructured data (e.g. listing descriptions and amenities) to predict fair market rent and flag anomalous listings.

In sum, this research builds on international studies in rental prediction, NLP enhanced modelling, and anomaly detection, while extending them into a highly relevant but under researched domain. By leveraging the University of Leeds expertise and the rich context of online listings, the proposed framework seeks to contribute both academically and socially supporting informed renting decisions and equitable market behaviour in the UK's student accommodation sector.

3. Importance and contribution to knowledge

This project contributes directly to current societal priorities by addressing the growing crisis of rental affordability and information inequality in the UK student housing market. As cost-of-living pressures intensify, students are increasingly unable to meet basic living needs. National figures reveal that 26% of students cannot afford rent in full, 17% rely on foodbanks, and 84% face serious housing issues such as mould or heating failures (NUS, 2024). These conditions not only compromise wellbeing but also restrict educational opportunity, particularly for students from low-income backgrounds, thereby entrenching structural inequality (Lewis, 2023).

The project offers practical tools that translate advanced analytics into actionable insights for diverse stakeholders. Rather than relying on static listings, institutions and platforms can use this framework to dynamically assess the fairness of rental prices and detect early signs of market distortion. This creates an opportunity for more equitable distribution of housing resources and more informed policy responses. By embedding such capabilities into existing digital infrastructure, the system supports platform accountability and fosters trust among end-users like students, universities, and housing providers. Universities could also use the tool to provide students with rent benchmarking dashboards or affordability assessments during the housing search process, helping them make informed choices and avoid exploitative pricing.

Aligned with the development of key emerging industries, particularly the UK's growing prop tech sector, the project demonstrates how artificial intelligence can be applied responsibly to improve housing services. The framework integrates machine learning, text analytics, and explainable AI techniques in a modular design, making it highly adaptable for integration into private student accommodation platforms (e.g. UniHomes, Zoopla) or university accommodation systems. These innovations can support smarter regulation, demand forecasting, and platform accountability contributing to the digital transformation of housing infrastructure and services.

Technologically, the project advances the application of explainable machine learning in complex, high-stakes environments. Unlike black-box models, this approach leverages interpretable linguistic features (e.g. amenities, listing structure) to enhance user confidence. It sets a precedent for transparent AI in other domains where fairness and interpretability are critical, such as public services, education, and healthcare.

This project brings together insights from several disciplines to tackle a complex, real-world problem. It draws on urban analytics to understand how rent levels vary across neighbourhoods, sociology to explore how housing affects social inequality, education policy to consider the impact of rent on access to higher education, and corpus linguistics to analyse the language used in rental listings. By weaving these perspectives together, the research gains both depth and practical relevance, opening conversations across fields from data science and housing economics to student wellbeing and public policy.

Nationally, the framework can be scaled across other UK university cities facing similar pressures, including Manchester, Bristol, and Nottingham. Its compatibility with publicly available data and institutional systems means it can be deployed in partnership with universities, student unions, and local authorities, supporting a national standard for pricing fairness in student accommodation. Internationally, the methodology is transferable to university hubs in other countries experiencing housing strain, such as Australia, the Netherlands, and the U.S., thus offering global research benefit and application.

By following the CRISP-DM methodology, the project ensures transparency, replicability, and long-term usability, making it well suited for future academic expansion and policy engagement. Its outputs ranging from interpretable models to open-source tools will contribute to new knowledge in algorithmic fairness, digital housing governance, and socially responsible AI, making it a valuable resource for national and international research initiatives.

4. Pilot study

4.1 Case Study Challenge

This pilot study evaluated a machine learning framework for predicting weekly rent and detecting pricing anomalies in student accommodation listings from Leeds. Using data scraped from Unipol, the model integrated structured attributes with linguistically features derived from natural language processing and corpus analysis. It aimed to determine whether a machine learning model that integrates structured rental data with text-mined features could effectively predict weekly rent and identify anomalous listings within a single city platform. The case study operationalised the broader project aims (see Chan, 2024; Zhao, 2024), with a particular focus on the contribution of text analytics techniques namely TF-IDF, Sketch Engine-derived phrases, and Word2Vec embeddings in improving predictive performance beyond the use of structured features alone.

4.2 Data Acquisition

To support the development of a rental price prediction and anomaly detection model, a tailor-made data acquisition pipeline was implemented using Playwright, a Python-based browser automation framework. The chosen data source was Unipol, a central platform for student accommodation listings in Leeds. This site was selected due to its high listing volume, broad property type coverage, and relevance to the target user group. To ensure diversity and coverage in the sample, filters were systematically applied across combinations of property type and bedroom count, ensuring a balanced sample across rental segments. Key fields were extracted with timed delays, and the raw data were saved to CSV.

As a result of the systematic filtering, a total of 1,021 listings were collected, covering a broad spectrum of rent prices, property categories (e.g. studios, bedsits, shared houses), and postcode zones (e.g. LS2, LS6, LS7). Visual inspection confirmed the diversity and completeness of the dataset, supporting downstream analysis. Throughout the scraping process, care was taken to ensure that no login or personal information was accessed, and the data was collected in accordance with ethical web scraping practices, with throttled request rates and respect for publicly accessible content.

4.3 Data Cleaning and Preparation

Following data acquisition, the dataset was pre-processed using a custom Python script to ensure consistency and completeness for modelling. In line with CRISP-DM principles, the data were transformed into a structured, analysis-ready format. Initial cleaning addressed missing values and inconsistent types, *weekly_rent* and deposit fields were converted using and rent values were imputed with the median where applicable. Listings with critical omissions were excluded. Postcodes were extracted via regular expressions and missing entries were filled with “Unknown”. Descriptions were cleaned using NLTK, lowercased, tokenised, and filtered for stop words and stored in a *clean_description* field for use in downstream TF-IDF and Word2Vec feature generation.

Property types were classified using rule-based keyword detection, assigning each listing to one of eight standardised categories. This approach improved label precision compared to Unipol’s broader categories. Feature engineering was then conducted in two stages. First, derived metrics such as *desc_length* and the rent-to-deposit ratio were calculated. Second, semantic amenities (e.g. “private bathroom”) were extracted using a combination of spaCy entity recognition and regular expressions. These features were binarized using MultiLabelBinarizer and filtered to include only those present in at least ten listings, ensuring statistical relevance.

Listings with complete data across all required features were retained and exported to a cleaned CSV file. This refined dataset served as the foundation for the modelling pipeline, anomaly detection processes, and the development of the rental fairness checker interface.

4.4 System Architecture and Feature Development

To capture linguistic signals in the listing descriptions, several semantic features were developed in model.py. These were used alongside structured variables to enrich the input space for predictive modelling. Text descriptions were first vectorised using TF-IDF, with the top 100 most informative terms retained. Word2Vec embeddings were also trained on tokenised descriptions to capture word-level semantic similarity. Word2Vec vectors were clustered using KMeans, with binary indicators assigned per cluster. In addition, a curated set of domain-specific phrases was extracted based on Sketch Engine inspired collocations (e.g. “modern kitchen”). These phrases were encoded as binary features to detect marketing and amenity language. The most frequent terms from each Word2Vec cluster are shown in Figure 6A, revealing distinct groups related to amenities, quality descriptors, and location-specific language. These features captured both lexical and semantic patterns.

4.5 Predictive Modelling and Results

The final feature set including structured attributes, engineered binary indicators, TF-IDF vectors, Word2Vec semantic clusters, and Sketch Engine-style phrase flags was processed through a unified modelling pipeline. A ColumnTransformer was employed to apply appropriate transformations to each feature group, enabling seamless integration of numerical, categorical, and text-derived data. A Random Forest Regressor was trained to predict weekly rent, using an 80/20 train-test split for each target. Evaluation used standard metrics: MSE, MAE, and R^2. The weekly rent model achieved an MSE of 693.74, MAE of £19.89, and an R^2 score of 0.70, indicating a strong ability to explain variance in rental prices. As illustrated in Figure 3A, predictions for weekly rent were closely aligned with actual values, while Figure 4A shows that most prediction errors clustered near zero, with a minority of outliers contributing to a long right-tailed distribution. Results confirm that linguistic features improve predictive accuracy, especially for detailed listings.

4.6 Anomaly detection

To identify rental listings with potentially unfair or outlier pricing, an anomaly detection procedure was implemented based on residual analysis. Anomalies were identified from absolute errors between actual and predicted rent. Listings with errors exceeding the 95th percentile threshold (approximately £40) were flagged as anomalies. These included overpriced shared houses and undervalued listings lacking descriptions or amenities.

The ten most extreme anomalies, ranked by absolute error magnitude, are presented in Figure 5A, illustrating the model’s capacity to detect statistically significant deviations from expected market behaviour. All flagged listings were exported, which may serve as a useful reference for students, accommodation officers, or policy groups interested in rental fairness. This interpretable approach supports the project’s goal of enabling informed rental decisions.

4.7 Rental Fairness Assessment Interface

To demonstrate real-world applicability, an interactive command-line interface was developed to provide a rental fairness check based on the trained model. Users can input listing attributes including location, property type, rent inclusions, a brief description, and the proposed weekly rent. The system processes this input through the same pipeline used during model training and returns a data-driven fairness evaluation. If the proposed rent significantly exceeds the predicted value beyond the established anomaly threshold the listing is flagged.

4.8 Conclusion

This pilot study demonstrated the feasibility of applying data mining and text analytics to student rental pricing. By combining structured data with TF-IDF, Word2Vec, and Sketch Engine-style features, the model achieved strong predictive accuracy ($R^2 = 0.70$) and enabled transparent anomaly detection. While limited to Leeds and lacking spatial or user validation data, the framework is modular, interpretable, and scalable. Future work should expand to other cities, incorporate geospatial variables, and engage stakeholders to refine thresholds and support broader deployment across UK student housing platforms.

5. Programme and methodology

5.1 Work Programme Overview

This project follows the CRISP-DM framework, which provides a structured and repeatable methodology for executing data-driven projects. The six phases process contains Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment Planning. This map directly to key milestones in a 24-week schedule. Each stage is adapted to the specific challenges of rental price prediction and anomaly detection in the student housing market.

5.2 Methods and Techniques

A combination of supervised learning, unsupervised clustering, and text analytics techniques will be used to model rent values and detect anomalies. Methods and tools are chosen for compatibility with mixed-type data and stakeholder-facing interpretability.

Data needed will be acquired from the student housing platform using Playwright, a Python-based browser automation tool. Scraped data will include both structured fields and unstructured property descriptions. Pandas will be used for structured data cleaning, while NLTK and spaCy will support text tokenisation, stop word filtering, and lemmatisation. Descriptions will be transformed into a *clean_description* field to support text mining. Feature extraction will combine TF-IDF, Word2Vec embeddings, and Sketch Engine-style collocation mining. TF-IDF will identify frequent terms, Word2Vec will support semantic modelling, and KMeans will cluster descriptions into thematic categories. Sketch Engine will be used to extract bi-grams and trigrams relevant to amenities and pricing, which will be binarized and integrated into the feature set.

A Random Forest Regression model will be trained to predict weekly rent based on the combined feature space. Its ability to handle heterogeneous features and return interpretable feature importance scores makes it well suited for this application. For anomaly detection, residuals will be calculated between predicted and actual rent values. Listings with errors above the 95th percentile will be flagged for review as potential pricing outliers. A CLI-based fairness tool will provide real-time rent assessments and support usability testing through scenario-based simulations.

5.3 Experiments and Evaluation

Model performance will be validated using an 80/20 train-test split stratified by property type and location. Evaluation metrics will include MSE, MAE, and R^2. Feature importance rankings will clarify which variables, structured or textual, most influence pricing. Anomaly detection will be evaluated using residual plots and inspection of outliers. The fairness checker will undergo scenario-based testing with varied inputs to evaluate usability, clarity, and consistency. This will guide refinements before user engagement.

5.4 Stakeholders and User Contribution

Students are the primary end users of this system, as they stand to benefit from increased transparency in rental pricing and deposit expectations. The tool will enable them to assess whether the rent they are being asked to pay aligns with market norms for similar properties, helping to identify cases of potential overpricing or unfair contract terms. This empowers students to make more informed decisions and avoid being overcharged, particularly in cities where affordability is a known concern. The system can also be embedded into university housing guidance or student support portals, providing a trusted resource during the housing search process.

Secondary stakeholders include university accommodation offices and student unions, who can use aggregated model outputs to monitor pricing trends and advocate for fairer housing practices. The tool's anomaly detection capability could flag properties that deviate significantly from predicted norms, supporting campaigns around affordability or regulatory review. Landlords and letting agents may also benefit from the tool's predictive insights. By using rent and deposit predictions as benchmarks, landlords can adjust pricing strategies to stay competitive while maintaining fairness. Deposit prediction models could inform landlords about what is considered reasonable in each rental context, reducing disputes and aligning expectations on both sides.

During evaluation, informal testing will be conducted with students and advisors to refine usability, interface clarity, and how predictions and anomalies are presented. Feedback will ensure the tool delivers relevant and actionable insights.

5.5 Milestones and Deliverables

The research will follow a 24-week timeline aligned with CRISP-DM phases, as illustrated in the Gantt chart. Month 1 covers business and data understanding, including research framing and Unipol data scraping. Month 2 focuses on data preparation, including structured data cleaning, text preprocessing, and feature extraction using TF-IDF, Word2Vec, and Sketch Engine. Feature engineering will be completed by Week 8.

Model development takes place across Months 3 and 4, starting with baseline Random Forest training and anomaly detection, followed by model refinement, error analysis, and feature importance review. By Week 16, the model is expected to be production ready. In Month 5, the CLI-based fairness checker will be built and tested using scenario-based simulations. Final evaluation and potential dataset expansion will occur in the last month, alongside report writing and submission.

Key milestones such as "Feature Engineering Complete", "Final Model Ready", and "Model Evaluation Complete" will be used to track progress and ensure quality delivery, as shown in the Gantt chart.

5.6 Alignment with CRISP-DM Phases

All project phases map closely to CRISP-DM stages, ensuring structured progression:

CRISP-DM Phase	Project Activities
Business Understanding	Define objectives; analyse affordability issues in UK student housing
Data Understanding	Scrape and explore Unipol listings; identify variable structure
Data Preparation	Clean structured fields; preprocess text; engineer semantic features
Modelling	Train Random Forest regressors; apply residual-based anomaly detection
Evaluation	Evaluate model performance (MSE, MAE, R^2); review anomalies and feature importance

Table 1: CRISP-DM Phases and Corresponding Project Activities

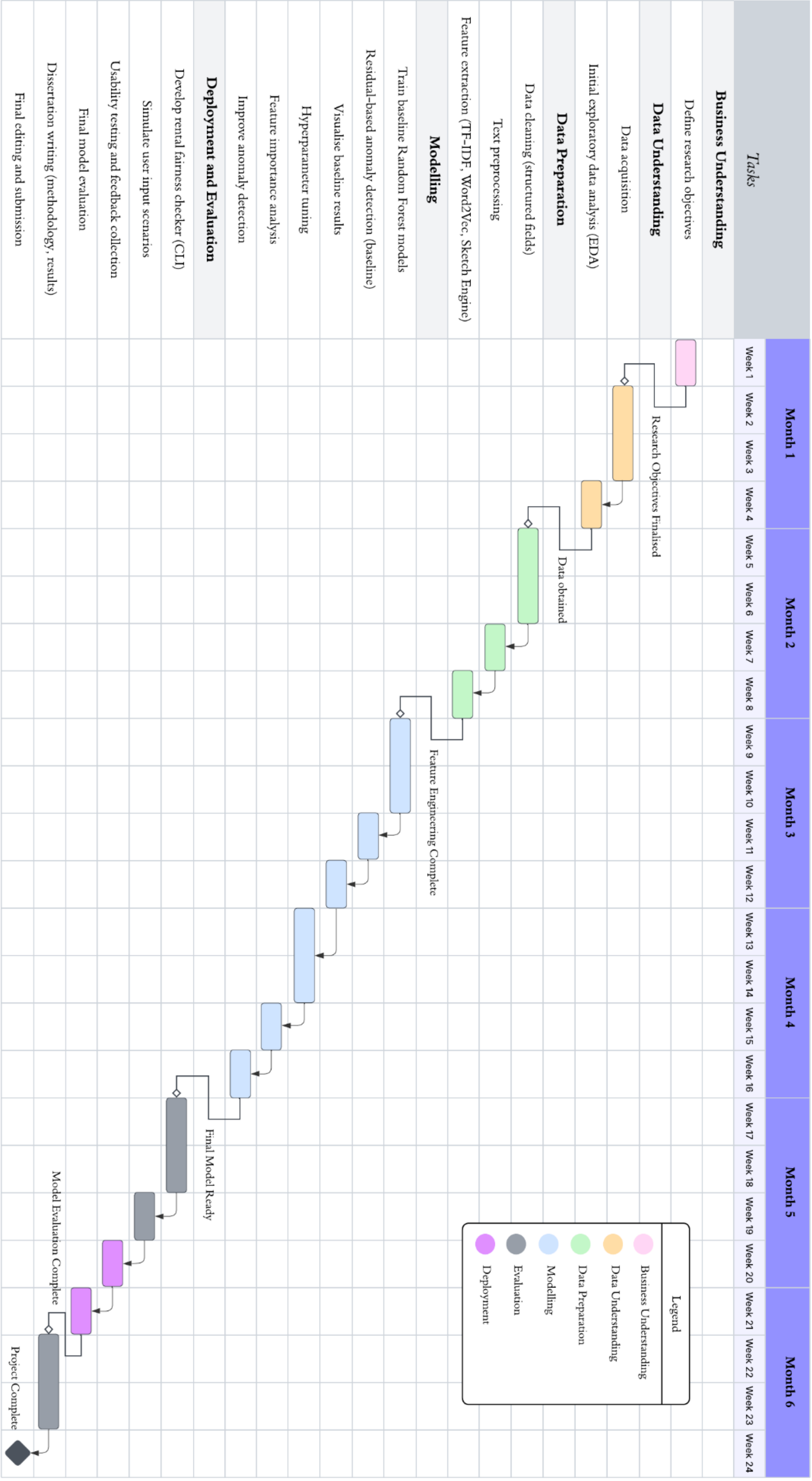
This alignment ensures methodological discipline, modular progress tracking, and compatibility with industry-standard data workflows.

5.7 Project Management Approach

The project will follow an agile, phase-driven approach with weekly checkpoints aligned to CRISP-DM stages. Tasks will be tracked using a visual Gantt chart, and all code, datasets, and results will be version-controlled via GitHub to ensure transparency and reproducibility.

Each phase will conclude with a short retrospective to reflect on outcomes, address technical obstacles, and adjust priorities. This approach allows for continuous improvement while maintaining alignment with the overall timeline and deliverables. The project lead will monitor weekly progress and log milestones to ensure timely completion of all objectives.

5.8 Workplan diagram (Gantt chart)



References

Chan, H. R., 2024. *Rent Price Prediction with Advanced Machine Learning Methods: A Comparison of California and Texas,* s.l.: Department of Statistics and Applied Probability, University of California, Santa Barbara, United States .

Lewis, J., 2023. *Students and the rising cost of living,* s.l.: UK Parliament, House of Commons Library.

NUS, 2024. *New NUS research reveals extent of student housing crisis.* [Online]
Available at: https://www.nus.org.uk/housing-survey-2024

Sultan, Y., 2023. Utilizing Model Residuals to Identify Rental Properties of Interest: The Price Anomaly Score (PAS) and Its Application to Real-time Data in Manhattan.

Unipol, 2024. *Leeds: assessment of the student housing market 2023,* Leeds: Unipol.

Zhao, Y., 2024. House Price Prediction: A Multi-Source Data Fusion Perspective.

Appendix A

A.1 Data Mining and Text Analytics Tools in Pilot Study

Tools used:

- Playwright - used for scraping full listing data from Unipol
- Pandas - for data cleaning and manipulation of structured fields
- NLTK and spaCy - for text preprocessing (tokenisation, stopword removal, lemmatisation)
- TF-IDF and Word2Vec - for extracting lexical and semantic features from descriptions
- KMeans - to cluster semantic word embeddings
- Sketch Engine - for identifying pricing-relevant bi-grams and collocations
- Random Forest Regression - used to train the predictive model
- Residual analysis - applied for anomaly detection based on prediction errors
- Platform – Virtual studio code

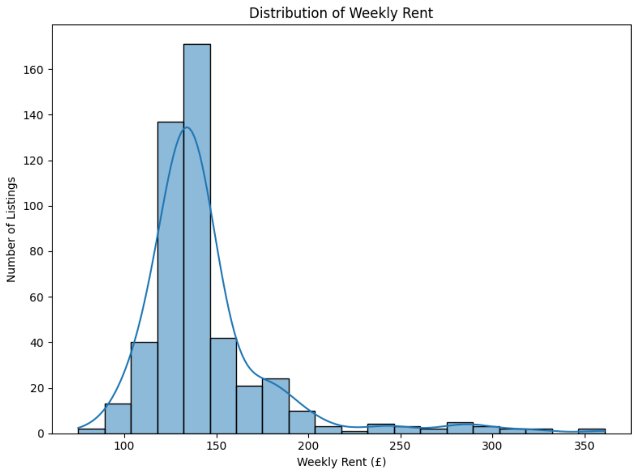


Figure A1. *Rent distribution in Leeds listings*

This chart confirmed non-normal distribution, supporting the decision to use Random Forest over linear regression.

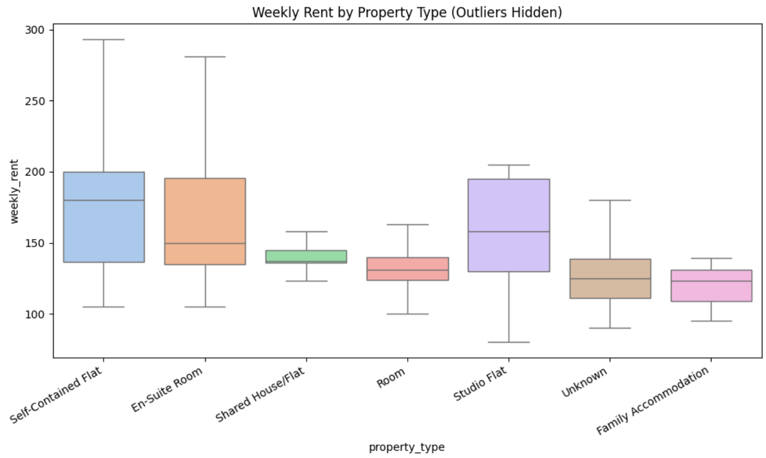


Figure A2. *Average weekly rent grouped by property type*

Helped validate that shared houses, studios, and bedsits follow distinct pricing trends

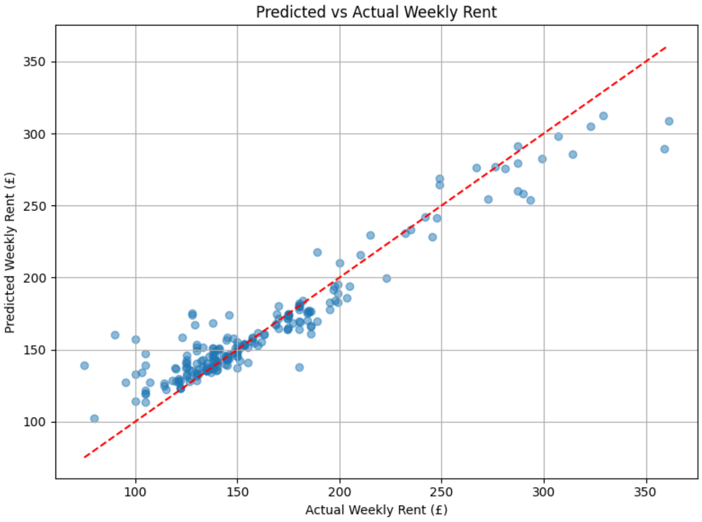


Figure A3. *Predicted vs actual weekly rent values*

Shows a strong fit along the diagonal, indicating model accuracy

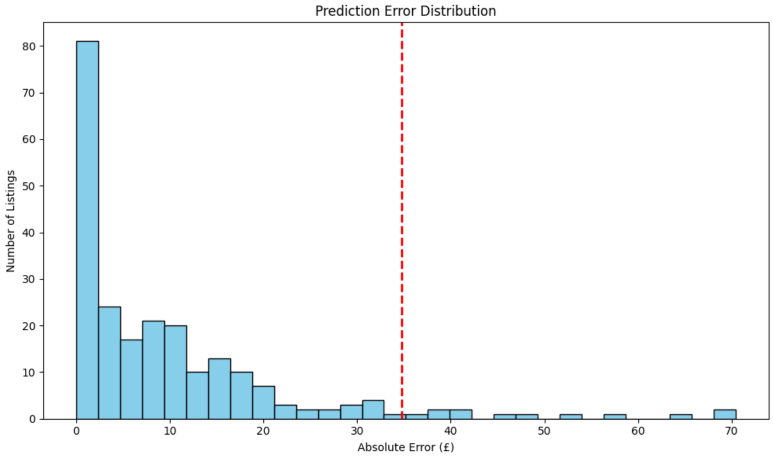


Figure A4. *Prediction error distribution (residuals)*

Most listings had low errors, though outliers required anomaly flagging

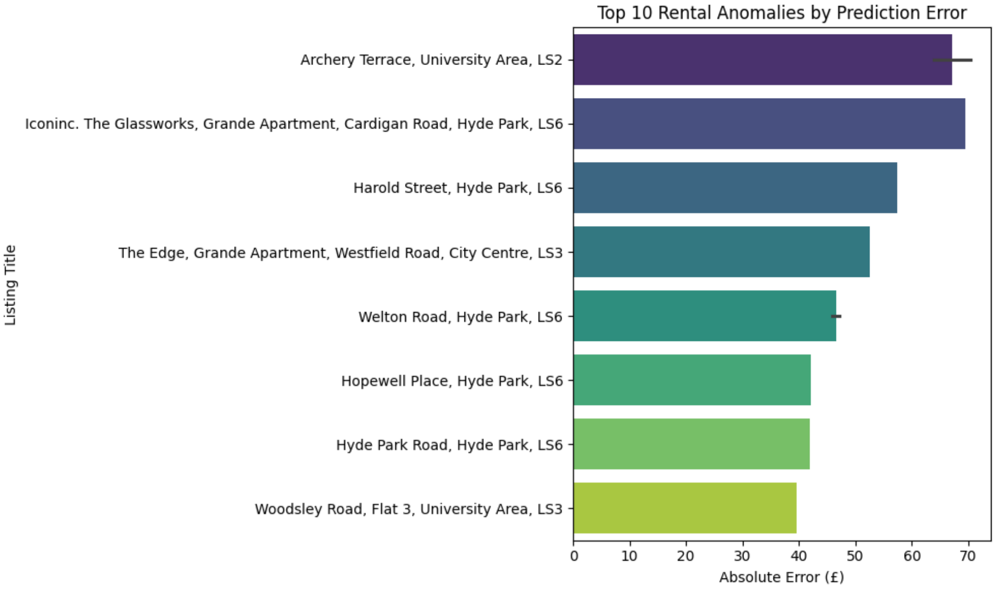


Figure A5. *Listings with the largest residual errors (top 10 anomalies)*

Flagged listings were analysed as potential examples of overpricing or under specification

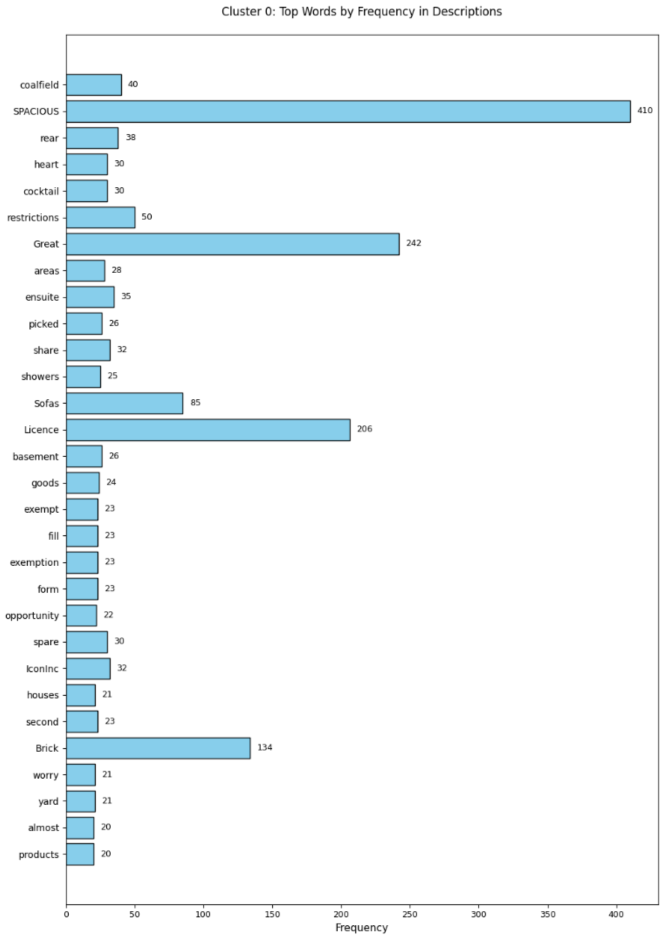
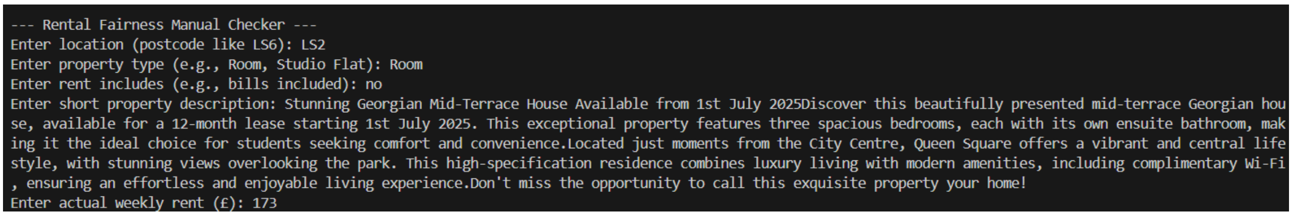


Figure A6. *Word2Vec cluster frequency charts*

Cluster 0: High-frequency terms related to proximity and campus access.



```
--- Rental Fairness Manual Checker ---
Enter location (postcode like LS6): LS2
Enter property type (e.g., Room, Studio Flat): Room
Enter rent includes (e.g., bills included): no
Enter short property description: Stunning Georgian Mid-Terrace House Available from 1st July 2025Discover this beautifully presented mid-terrace Georgian hou
se, available for a 12-month lease starting 1st July 2025. This exceptional property features three spacious bedrooms, each with its own ensuite bathroom, mak
ing it the ideal choice for students seeking comfort and convenience.Located just moments from the City Centre, Queen Square offers a vibrant and central life
style, with stunning views overlooking the park. This high-specification residence combines luxury living with modern amenities, including complimentary Wi-Fi
, ensuring an effortless and enjoyable living experience.Don't miss the opportunity to call this exquisite property your home!
Enter actual weekly rent (£): 173
```

Figure A7. *CLI for anomaly detection*

Rental fairness manual checker

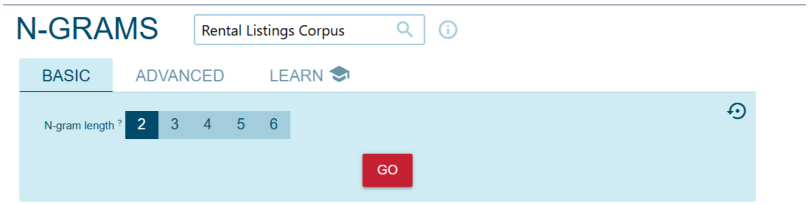


Figure A8. *Sketch engine N-gram*

Obtain sketch engine inspired collocations

Code Repository

All code developed for data collection, cleaning, modelling, anomaly detection, and interface prototyping is available at the following GitHub repository

- scraper_enhance.py – Used to collect full Unipol listings using Playwright
- data_cleaning.py – Preprocessing script for converting raw CSVs into cleaned structured datasets
- model.py – Contains model training logic for TF-IDF, Word2Vec, Random Forest, and anomaly detection

https://github.com/yeeheng12/COMP2121

A.2 Tools in searching for background information

Tools used:

- Google Scholar – prompt (Rent price prediction, Student living cost Leeds)

A.3 Tools to draft the report and improve grammar and style

Tools used:

- ChatGPT – prompt (organize thoughts pasted above)
- Grammarly