Report Business and Data Understanding

Chapter 1: Introduction

In today's highly competitive real estate market, determining the optimal rental rate for properties is a complex task that significantly impacts revenue and vacancy rates. By leveraging machine learning, property management companies can accurately predict rental rates, thus maximizing occupancy and revenue. Equity Residential, a leading real estate investment trust (REIT), seeks to implement a predictive system to enhance its rental rate setting process. This report outlines the development of a machine learning model to achieve this goal, providing insights into the business problem, data understanding, and the proposed solution.

Chapter 2: Business and Data Understanding

Business Problem and Current Situation

Equity Residential owns and operates a large portfolio of rental properties across the United States. The primary business challenge is to determine optimal rental rates that balance maximizing revenue and minimizing vacancy rates. Currently, rental rates are set based on historical data, market conditions, and manual adjustments by the property management team. This approach is time-consuming and may not capture all relevant factors, leading to suboptimal pricing decisions.

Section 2.1: Terminology

Section 2.1.1: Business Terminology

Term	Description
Key Performance Indicators (KPIs)	Quantitative metrics used to measure the effectiveness of a business or

	organization. Examples in this project include the accuracy of rent forecasts.
Vacancy rate	Percentage of all available rental units that are currently empty. This metric is used to assess the balance of supply and demand in the rental market.
Amenities	Features or services provided in a rental property that may affect its attractiveness and price. Examples include laundries, air conditioning, gyms and swimming pools. (can we take this into account?)
Date of entry	The date when the tenant plans to start renting. This can affect rental rates, especially if demand varies seasonally.
Web scraping	Technology for obtaining web data by extracting it from web resource pages. Web scraping can be done manually by a computer user, however, the term usually refers to automated processes implemented using code that executes GET requests to the target site.

Section 2.1.2: ML Terminology

Term	Description
Regression	A type of machine learning task in which the goal is to predict a continuous value based on input data. In this project, regression models will be used to predict rents.
Features	Independent variables or input data used in the machine learning model for forecasting. Examples include the number of bedrooms, bathrooms, amenities, and vacancy levels.
Target variable	Dependent variable or output that the machine learning model seeks to predict. In this project, the target variable is rent.
Lack of education	When the machine learning model is too simple to capture the basic patterns in the

	data, which leads to poor performance on both training and test data.
Retraining	When a machine learning model studies training data too well, including noise and outliers, which leads to poor generalizability to new, invisible data.

Section 2.2: Scope of the ML Project

Section 2.2.1: Background

Equity Residential is a prominent player in the real estate market, managing thousands of rental units. The company aims to leverage data-driven strategies to enhance operational efficiency and profitability. This project focuses on developing a machine learning model to predict optimal rental rates, leveraging internal and external data sources to make informed pricing decisions.

Section 2.2.2: Business Problem

Equity Residential aims to predict future rental rates in order to improve the decision-making process for both landlords. Accurate price forecasts can help landlords optimize their rental rates to maximize revenue and reduce vacancy rates. The goal is also to reduce the time spent on estimating the cost of renting an apartment.

Section 2.2.3: Business Objectives

1. **Main objective**: Forecasting future rental rates for Equity Residential properties.
    a. **Improving efficiency:** Providing accurate rental price forecasts will help landlords make informed decisions, which in turn will increase the efficiency of the rental market.
    b. **Increase in income:** For landlords, accurate price forecasts can help optimize prices and maximize revenue.
    c. **Increase the evaluation speed:** The landlord wants to determine the cost of renting an apartment that needs to be exposed cheaper and faster.
2. **Related Questions**:
    - How do different amenities affect the rent?

- How does the vacancy rate affect the rent?
- How does seasonality affect rental demand and pricing?

Section 2.2.4: ML Objectives

Develop a regression model to predict rents based on various characteristics such as the number of bedrooms, bathrooms, amenities, and vacancy levels.

Secondary ML Goals:

- Identify key features influencing rental rates.
- Optimize the model to minimize prediction errors and improve accuracy.

Section 2.3: Success Criteria

Section 2.3.1: Business Success Criteria

1. Increase the accuracy of rent forecasts to reduce pricing errors by 10%.
2. Speed up the process of determining the price of this apartment by 70%
3. Reduce the cost of determining the price of this apartment by 50%

Section 2.3.2: ML Success Criteria

1. Achieve an average absolute deviation (MAE) of less than $50 in rent projections.
2. Achieve an R-squared value of more than 0.85 for the regression model.

Section 2.3.3: Economic Success Criteria

1. Increase the accuracy of rental price forecasts to reduce pricing errors by 10%.
2. Speed up the process of determining the price for a given apartment by 70%.
3. Reduce the cost of determining the price for a given apartment by 50%.

Section 2.5. Data collection

Section 2.5.1 Data collection report:

- **Data source**: We took the data from an open source - kaggle, but the author who posted the dataset collected data from Equity Residential websites by web scraping.
- **Data type**: Data includes numeric (rental rates, vacancy levels), categorical (amenities, location) and temporary (date of placement, date of entry) characteristics.
- **Data size**: The dataset consists of 62,800 records with 32 characteristics each.
- **Data collection method**: The data was collected using web scraping tools and manually verified.

Section 2.5.2: Data Version Control Report

- **Data Version**: The current data version is v1.0, collected between June 25, 2021 and July 17, 2021.
- **Data Change Log**: The initial data set with gaps due to changes in the site design. Future versions will seek to fill in these gaps.
- **Data Backup**: Backup for any data changes.
- **Data archiving**: Historical data is archived in local (in the future cloud) storage for long-term storage.
- **Data access control**: Only developers and product owners have access.

Section 2.6: Data Quality Verification

Section 2.6.1: Data Description

The dataset includes 62,800 records and 32 characteristics. Below is a table with a description of all the characteristics obtained during web scraping.

```
|           Index          |                           Descript
|:-------------------------:|:--------------------------------
|           Price          | Price (recorded daily)
|           Beds           | # of Beds in an apartment. 0 mea
|           Baths          | # of baths in an apartment.
|           sq.ft          | # of square feet.
|           Floor          | Floor
|       Move_in_date       | Date the apartment was available
```

building_id	for apartments that had multiple
unit_id	The apartment unit number.
URL	URL that was scraped.
Day_Recorded	Day the row of data was scraped.
Amenity	Text field describing different
Apartment Name	Name of the apartment complex.
Address	Address of apartment complex.
City	City the apartment is in.
Units	Number of units the apartment co
Northern_Exposure	1 if apartment has northern expo
Southern_Exposure	1 if apartment has southern expo
Eastern_Exposure	1 if apartment has eastern expos
Western_Exposure	1 if apartment has western expos
Balcony	1 if apartment has balcony, 0 ot
Walk_In_Closet	1 if apartment has walk in close
Fireplace	1 if apartment has fireplace, 0
City_Skyline	1 if apartment has city skyline,
Kitchen_Island	1 if apartment has kitchen islan
Stainless_Appliances	1 if apartment has stainless ste
Renovated	1 if apartment has been rennovat
Office_Space	1 if apartment has office space,
Days_Till_Available	Days until you could move in.
Day_of_the_week_recorded	What day of the week was the dat
Unique_ID	Unique ID to identify the same a
Estiamted_Vacancy	# of obvs that day/ total units

Section 2.6.2: Data Exploration

First Findings and Initial Hypothesis:

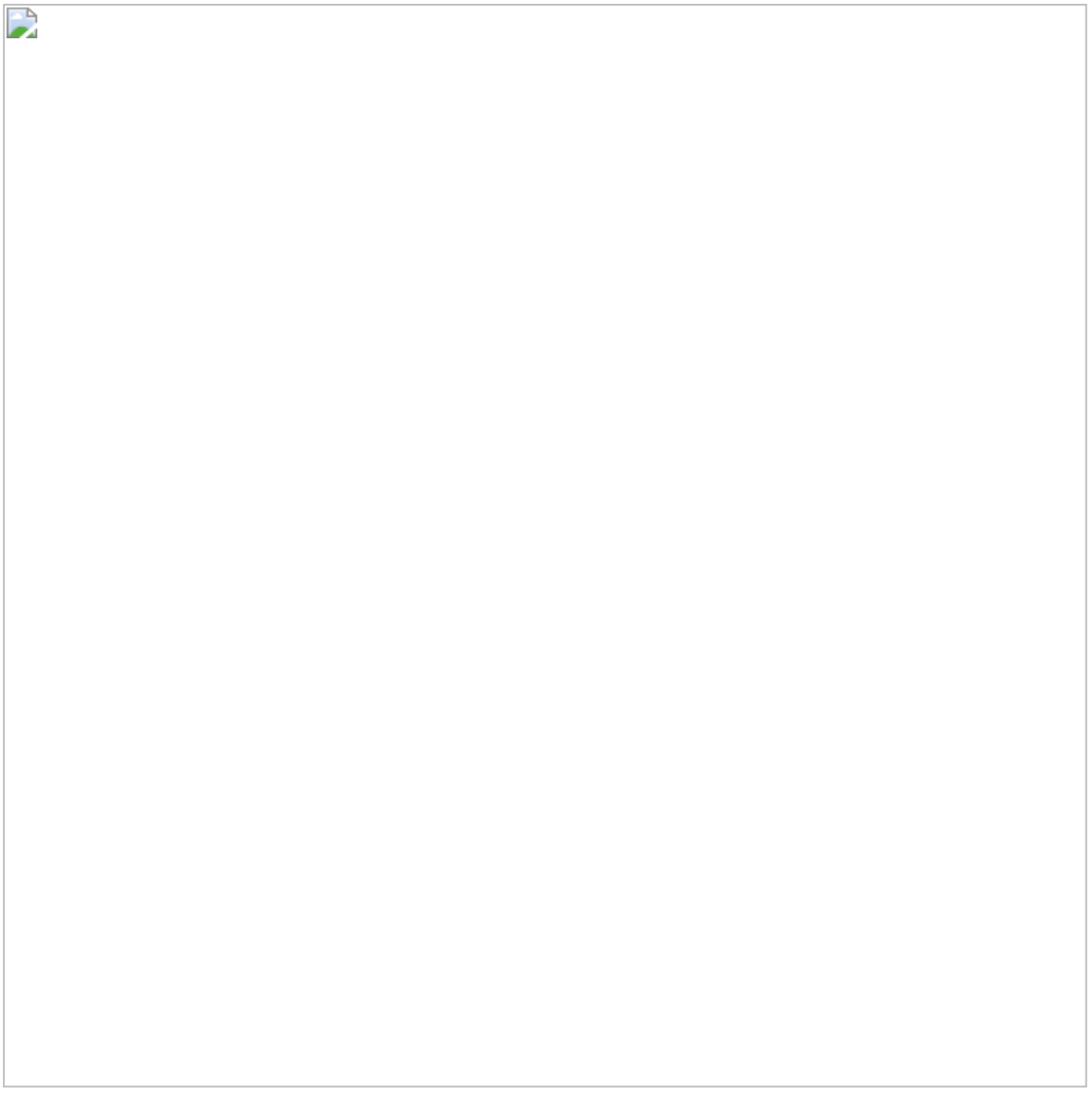
During the initial analysis of the data, we found the following characteristics:

1. **Data structure:**
   - - The data contains the columns: `unit_id`, `Amenity`, `Apartment Name`, `Address`, `Unique_ID`, and others.
   - We have deleted the column `Unnamed: 0`, as it does not carry useful information.
2. Clearing the data:

- Gaps in string values have been eliminated using methods `str.strip()` and `str.lower()`.
- Addresses and other text data have been cleared of line breaks and excessive spaces.

3. Processing of text data:
   - The description of amenities (the `Amenity` column) has been cleaned up using regular expressions to remove unnecessary spaces and characters.
4. Vectorization:
   - Some text data has been converted into vectors for subsequent analysis.

Based on these initial steps, the following graphs and tables have been built for a deeper understanding of the data:



Section 2.6.3: Data Requirements

- **Price** - the price cannot be negative (checking values in the range between expect_column_values_to_be_between)

  Completeness: Expect Price to have no missing values:

  - expect_column_values_to_not_be_null,
  - expect_column_values_to_be_of_type(column="Price", type_="int")

- **Beds** - the number of bedrooms cannot be negative (checking values in the range between expect_column_values_to_be_between)

  Completeness: Expect Beds to have no missing values:

  - expect_column_values_to_not_be_null,
  - expect_column_values_to_match_regex(column="Beds", regex="^\d+$")

  Consistency: Expect Beds values to be whole numbers
- **Baths** - the number of bathrooms cannot be negative (checking values in the range between expect_column_values_to_be_between)

  Completeness: Expect Baths to have no missing values:

  - expect_column_values_to_not_be_null,
  - expect_column_values_to_match_regex(column="Baths", regex="^\d+$")

  Consistency: Expect Baths values to be whole numbers
- **sq.ft** - square meters cannot be negative (checking values in the range between expect_column_values_to_be_between)

  Completeness: Expect sq.ft to have no missing values: expect_column_values_to_not_be_null,

  Consistency: Expect sq.ft values to be either whole or half numbers: expect_column_values_to_match_regex(column="sq.ft", regex="^\d+(\.\d{1})?$"))
- **Floor** - floors cannot be negative (checking values in the range between expect_column_values_to_be_between)

  Completeness: Expect Floor to have no missing values.

  - expect_column_values_to_not_be_null,
  - expect_column_values_to_match_regex(column="Floor", regex="^\d+$")

  Consistency: Expect Floor values to be whole numbers
- **Move_in_date** - expect_column_values_to_match_strftime_format(column="Move_in_date",

strftime_format="%Y-%m-%d")
Validity: Expect Move_in_date to be in a valid date format

Completeness: Expect Move_in_date to have no missing values: expect_column_values_to_not_be_null,

Timeliness: expect_column_values_to_be_greater_than(column="Move_in_date", value="Day_Recorded")(Expect Move_in_date to be in the future relative to Day_Recorded.)

- **unit_id** - expect_column_values_to_match_regex(column="unit_id", regex="^[a-zA-Z0-9]+$")

  Consistency: Expect unit_id to be in a valid format (e.g., alphanumeric)),

  Completeness: Expect unit_id to have no missing values. expect_column_values_to_not_be_null,

  Uniqueness: Expect unit_id to have unique values. expect_column_values_to_be_unique(column="unit_id")

- **URL** - Completeness: Expect URL to have no missing values. expect_column_values_to_not_be_null(column="URL")

  Validity: Expect URL values to be valid URLs. expect_column_values_to_match_regex(column="URL", regex="^(http|https)://[^\s/$.?#].[^\s]*$")

  Uniqueness: Expect URL values to be unique. expect_column_values_to_be_unique(column="URL")

- **Day_of_the_week_recorded** - you can check that the days of the week are in the range of days of the week expect_column_values_to_be_in_set, Completeness: Expect Day_of_the_week_recorded to have no missing values.
  expect_column_values_to_not_be_null,

Section 2.6.4: Data Quality Verification Report

Data quality control

To check the quality of the data, we have performed the following steps:

Completeness of the data:

- Checking for all necessary cases and records.

- Example: The number of unique values in `unit_id` should correspond to the total number of records.

Data accuracy:

- Identification and correction of errors in the data.
- Example: Deleting invalid characters and extra spaces.

Missing values:

- Checking for missing values and their representation.
- Example: Missing values in the 'Price` column.

Statistical analysis:

- Basic data statistics to identify outliers and anomalies.

Conclusion

Based on the data analysis and verification, we can conclude that the data quality is sufficient for further model construction. The main problems, such as omissions and inaccuracies, have been identified and handled.

Section 2.7: Project Feasibility

Section 2.7.1: Inventory of resources

- **Staff**: Data scientists, business analysts, software engineers, domain experts.
- **Data**: Access to historical rental data and real-time updates.
- **Computing resources**: High-performance servers and cloud computing platforms.
- **Software**: Python, Jupyter Notebook, scikit-learn, pandas, SQL databases.

Section 2.7.2: Requirements, assumptions and limitations

- **Requirements**:
  - Timely completion within 6 weeks.
  - High accuracy and interpretability of the results.
  - Compliance with data security requirements.
- **Assumptions**:

    - Historical data are representative of future trends.
    - The absence of significant changes in the market.
- **Limitations**:
    - Limited data due to the update of the company's website with ads, which caused scraping failures.
    - We consider only the summer period and cannot track seasonal changes in the market.
    - We can't update the model often due to the short period of ad tracking.

Section 2.7.3: Risks and contingent measures

- **Risks**:
    - Data quality problems due to scraping failures.
    - Changes in the market that affect the accuracy of models.
- **Contingent measures**:
    - Implementation of reliable data cleaning and validation processes.
    - Regular updating of models with new data to account for changes in the market.

Section 2.7.4: Costs and benefits

- **Costs**:
    - Data collection and cleaning.
    - Development, evaluation and deployment of models.
    - Continuous monitoring and updates.
- **Benefits**:
    - Increased revenue by optimizing pricing.
    - Increasing tenant satisfaction and reducing vacancy rates.
    - Improved market understanding and strategic decision-making.

Section 2.7.5: Feasibility Report

A preliminary machine learning model has been developed to test the feasibility of the project. The model showed promising results, indicating that it is feasible

to develop an accurate predictive system for rental rate determination.

We were able to collect the data, analyze it, clean it up, convert it to load into the model and get the prediction results, which indicates that from a technical point of view, the problem can be solved. The results of the analyses: Inventory of resources, Requirements, assumptions and limitations, Risks and contingent measures, Costs and benefits indicate that the project is being implemented from a business point of view. Additional validation and testing will be carried out in the next phases of the project.

Section 2.8: Project Plan

The project plan includes the following phases:

1. **Understanding Business and Data**: Week 2
    - Inputs: Business Goals, Data Overview
    - Outputs: Defined Goals, Data Understanding
2. **Data preparation**: 1 week
    - Inputs: Initial data
    - Outputs: Cleaned and prepared data
3. **Model Development**: 1 week
    - Inputs: Prepared data
    - Outputs: Trained models
4. **Evaluation of models**: 1 week
    - Inputs: Trained models, test data
    - Outputs: Performance metrics of models
5. **Deployment of models**: 1 week
    - Inputs: Proven models
    - Outputs: Deployed models

Gantt chart and the entire project page at link.