Finding the Ideal San Francisco Apartment – Week 4

Brian Moore

1. **Background and Introduction**

   Few changes in life are simultaneously as frightening, uncertain and exciting as moving to a new city for work or school, yet, historically, upwards of 40% of Americans moved to a new city over the course of their lifetime. While moving is often new and exciting, the process of selecting a new place to live is extremely challenging and poses serious financial risks, even in the best of circumstances.

   Primarily, this is because you're not necessarily familiar with your surroundings before you arrive in a new city, yet, you're expected to make critical, long term decisions regarding housing and transportation with limited time and information. As a result, you might unknowingly rent an apartment that's priced fairly and close to work, but, lacks quality community spaces like parks, restaurants, and suffers from high rates of petty crime. Issues such as these are compounded when renters move to complex and competitive markets like the San Francisco apartment market. Renters looking to enter this market face a high stakes uphill battle with two primary challenges:

   Problem #1: Identifying 'Quality' neighborhoods well suited to their individual needs.
   Problem #2**:** Identifying fairly priced, overpriced and underpriced apartments in desirable neighborhoods.

   Given time and resources, renters can overcome the challenges above, however, data science and machine learning tools can provide renters with the answers they need to make informed decisions about where they choose to inside a much narrower time horizon.

2. **Purpose and Audience**

   The purpose of this project is to provide new renters and those unfamiliar with the San Francisco real estate market with advanced analytics and resources, allowing them to:

   1. Determine which neighborhoods are best suited to their individual needs.
   2. Determine if a potential apartment or house is over or underpriced.

3. **Data acquisition and cleaning**

   To provide answers to both problems listed above, I will be utilizing data from the following sources:

   Problem #1 – Apartment Listing Data Sources: Before determining whether apartments are under/overpriced, we must begin with a rich dataset of available apartments located in the city of San Francisco. After speaking with locals, I determined that Craigslist is the go-to resource due to its high rate of quality listings. As a result, all apartment listing

data originates from scraped craigslist listings. Specifically, I built a scraper designed to gather and store over 2,000 current listings containing:

1. Apartment pricing data
2. Geographic and neighborhood data
3. Sq/Ft, number of bedrooms and bathrooms
4. Parking and laundry availability
5. Pet occupancy

The raw apartment data will then be cleaned, analyzed and loaded into a Random Forest ML classification model.

Problem #2 – "Neighborhood Quality" Data: Determinizing neighborhood quality is far more nuanced, however, for the sake of simplicity, I've decided to utilize data from two sources to help me objectively identify desirable and undesirable neighborhood characteristics. First, I'll be utilizing a comprehensive crime dataset provided by the City of San Francisco outlining "police incident activity" segmented by location, type of crime, date, and time. Given the importance of living in a safe neighborhood, I will use "police incident activity" (and lack thereof) as a crude proxy to determine neighborhood desirability. I'm also going to be relying on the assumption that Neighborhoods with less police incidents are safer than those with more police activity. Next, I'll be utilizing location data from the Foursquare API to build a comprehensive dataset of venues around potential candidate apartments. The venue data from Foursquare will be used to identify neighborhoods with high concentrations of 3rd places" or public spaces where I can meet friends and interact with the community. The Foursquare dataset contains over 4,000 venues and outlines:

1. Venue Name
2. Venue Genre
3. Venue Location (lat, lon)
4. Venue Neighborhood

Other data sources: San Francisco Neighborhood GeoJSON data is also utilized throughout this analysis to accurately determine neighborhood names and boundaries.