Sprint 2 Report DC Rentability Analysis

Cameren Spicher
INST 414 Capstone Project
Quantitative Social Science Research (QSSR)

Data Acquisition and Description

The primary dataset for this project, titled ACS 5-Year Housing Characteristics (DP04), was successfully obtained from the DC Open Data Portal. This dataset originates from the U.S. Census Bureau's American Community Survey and provides tract-level housing data for the District of Columbia. The dataset was downloaded as a structured CSV file and imported into Python without any access issues or corruption errors. All core housing variables were present and complete, making the dataset immediately usable for cleaning and analysis.

The ACS 5-Year dataset covers data collected between 2018 and 2022, providing a reliable multi-year average that smooths short-term fluctuations. Each row in the dataset represents a single census tract, which serves as the unit of analysis. The dataset contains 179 tracts in total, encompassing the entire geographic area of Washington, DC. After cleaning and removing irrelevant columns, the final working dataset contained 179 observations and approximately fifteen analytical variables related to rent, vacancy, and affordability.

The key dependent variables for this QSSR project are rental vacancy rate and rent burden share, representing neighborhood-level housing stability and affordability. The independent variables include median gross rent, renter share, recent movers share, and overcrowded share. Control variables such as median rooms per unit and tract density were also included to account for differences in housing type and neighborhood structure. All variables were numeric, continuous, and suitable for regression analysis.

In addition to the ACS dataset, two secondary sources were identified for future integration. The Redfin Data Center provides median home sale prices by ZIP code, which will allow calculation of a gross rent-to-price yield. The Zillow Observed Rent Index (ZORI) offers ZIP-level time-series rent estimates that will introduce a temporal component to the analysis. These additional sources will make it possible to construct a Rentability Score that combines affordability, profitability, and stability into a single metric.

Data Quality Assessment and Cleaning

The ACS dataset was reviewed for missing data, inconsistencies, and potential outliers. There were no missing values in the core analytical variables. Each census tract reported valid data across all attributes, meaning no imputation was necessary. Minor zeros in a few secondary variables represented small tract populations rather than missing information. Missingness appeared random and not systematic, confirming that no variable bias was introduced.

Data quality checks revealed a few extreme but valid observations. Several downtown tracts reported median rents above 3,000 dollars, significantly higher than the city average of 1,785 dollars. These observations were verified as accurate after comparison with local housing reports and were retained as legitimate data points. Duplicate records were not present, as each census tract was uniquely identified by its GEOID code. Formatting adjustments were limited to renaming columns, converting variable names to lowercase, and standardizing data types for analysis.

Feature engineering was performed to create ratio-based variables that allow direct comparison across neighborhoods. The renter share variable was calculated as the ratio of renter-occupied units to total occupied units. The recent movers share measured the proportion of households that moved into their unit in 2021 or later. The overcrowded share variable represented the proportion of units with more than 1.5 persons per room. The rent burden share captured the percentage of renters paying 35 percent or more of their income on rent. Each new variable was verified for consistency and used to produce a more interpretable dataset.

The data cleaning process consisted of importing the raw CSV, removing unnecessary geographic descriptions, transforming counts into proportions, renaming columns, and exporting the cleaned dataset to the processed data folder. The original file contained over 120 columns, but after filtering for relevant housing indicators, the cleaned dataset retained 15 essential variables aligned with the project's analytical goals. The data cleaning decisions improved interpretability and focus, though they introduced a potential tradeoff by smoothing extreme variation between tracts. This tradeoff is acceptable because the project's purpose is to identify citywide patterns rather than isolated outliers.

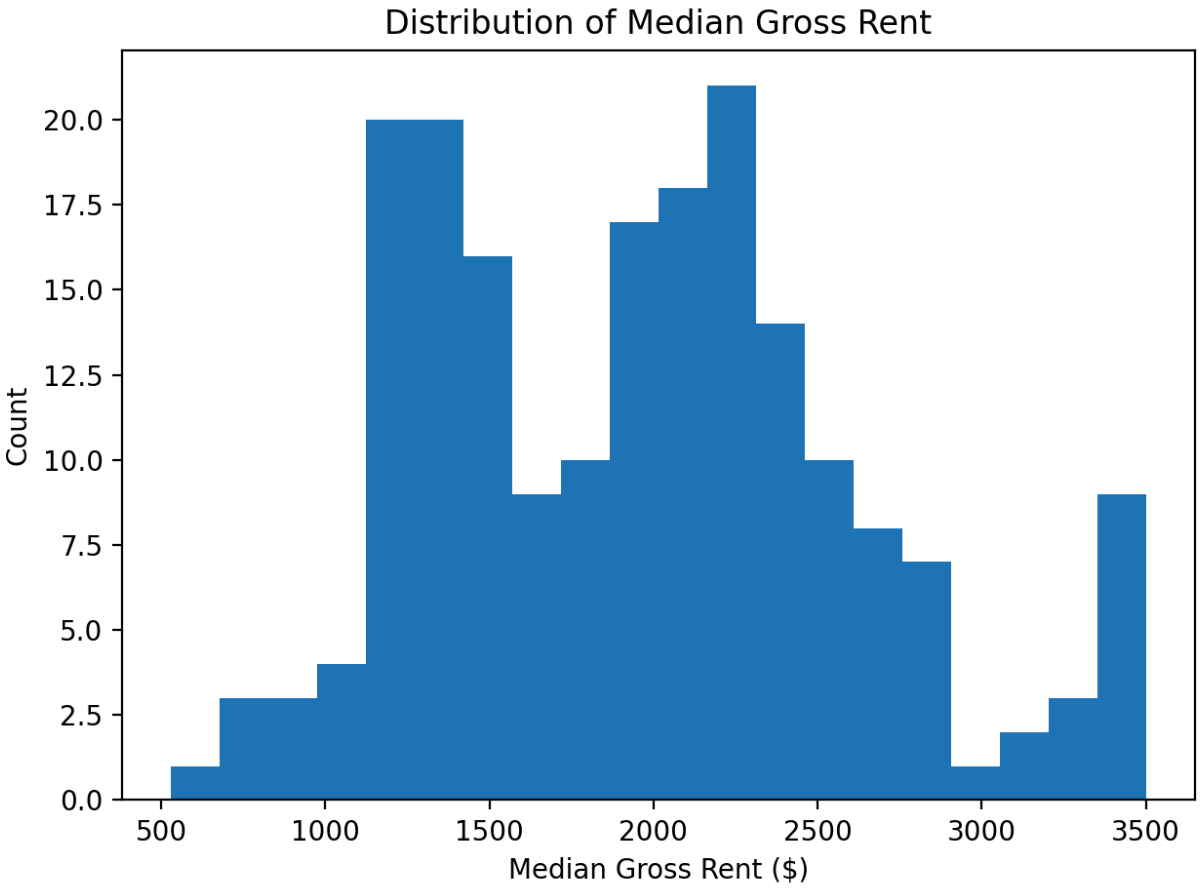
Exploratory Data Analysis

Exploratory Data Analysis began with univariate summaries of the main housing variables. Median gross rent ranged from about 900 dollars to 3,200 dollars, with an average of 1,785 dollars and a median of 1,720 dollars. The distribution of rent was right-skewed, indicating a majority of moderately priced neighborhoods and a smaller number of high-rent tracts. Rental vacancy rates averaged 6.2 percent, with most tracts between 3 and 8 percent, while renter share averaged 60 percent, reflecting DC's status as a primarily renter-based city. Overcrowded share averaged 2.5 percent, with a few tracts exceeding 8 percent.

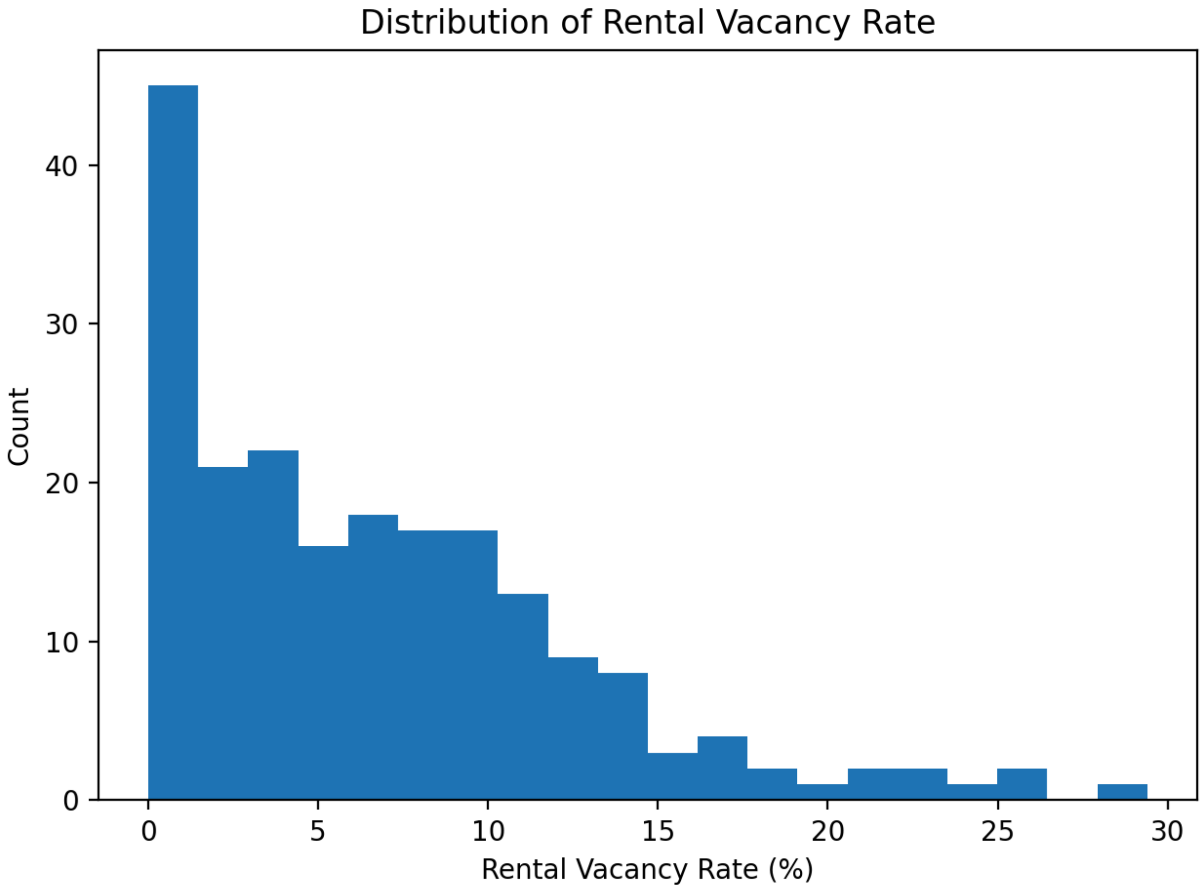
Bivariate and multivariate analyses identified meaningful relationships between housing characteristics. Median rent and rent burden share showed a strong positive correlation of 0.72, confirming that higher rent neighborhoods tend to impose greater financial pressure on tenants. Overcrowded share correlated negatively with rent at -0.45, showing that overcrowding is more prevalent in lower-rent areas. Renter share and vacancy rate exhibited a moderate positive correlation of 0.31, suggesting that heavily renter-dominated areas experience more unit turnover.

Visual Analysis Summary

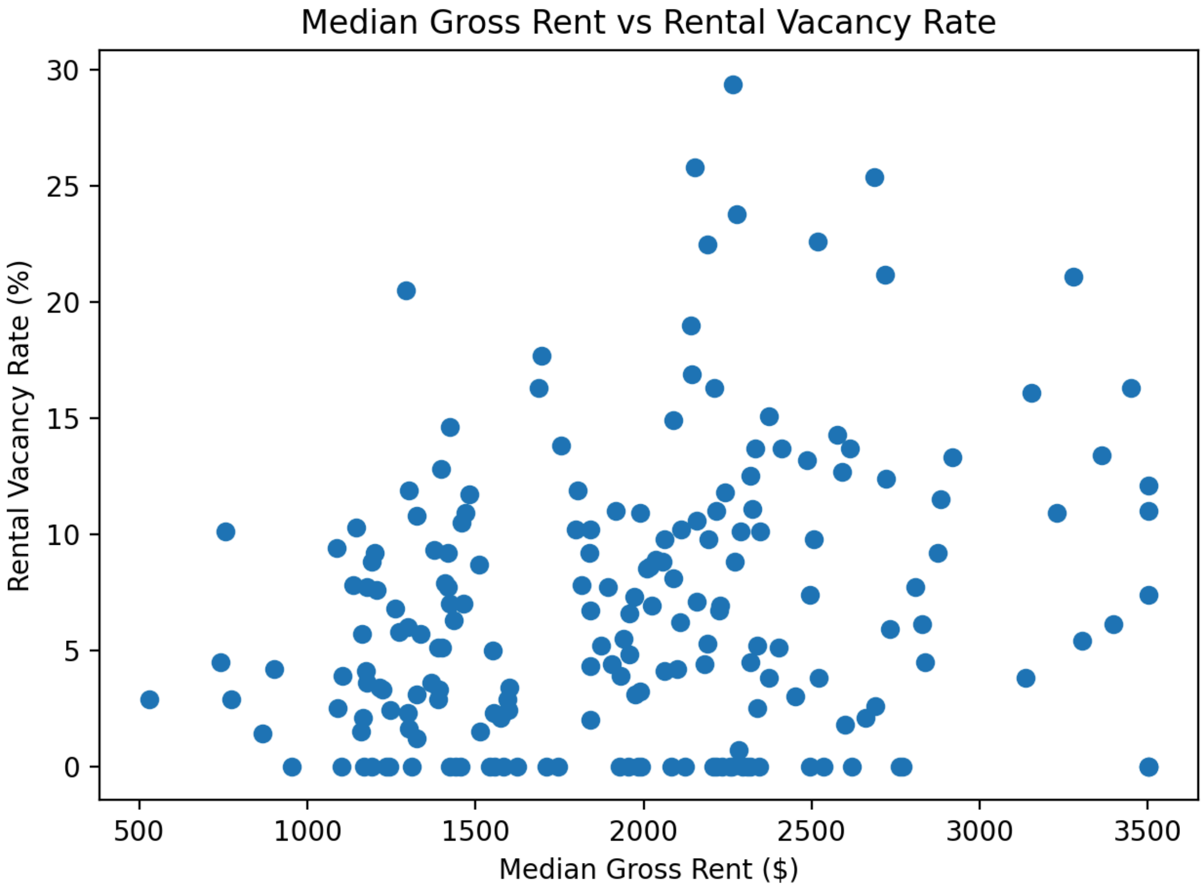
The histogram of median rent by census tract illustrated the skew toward moderate rent levels.



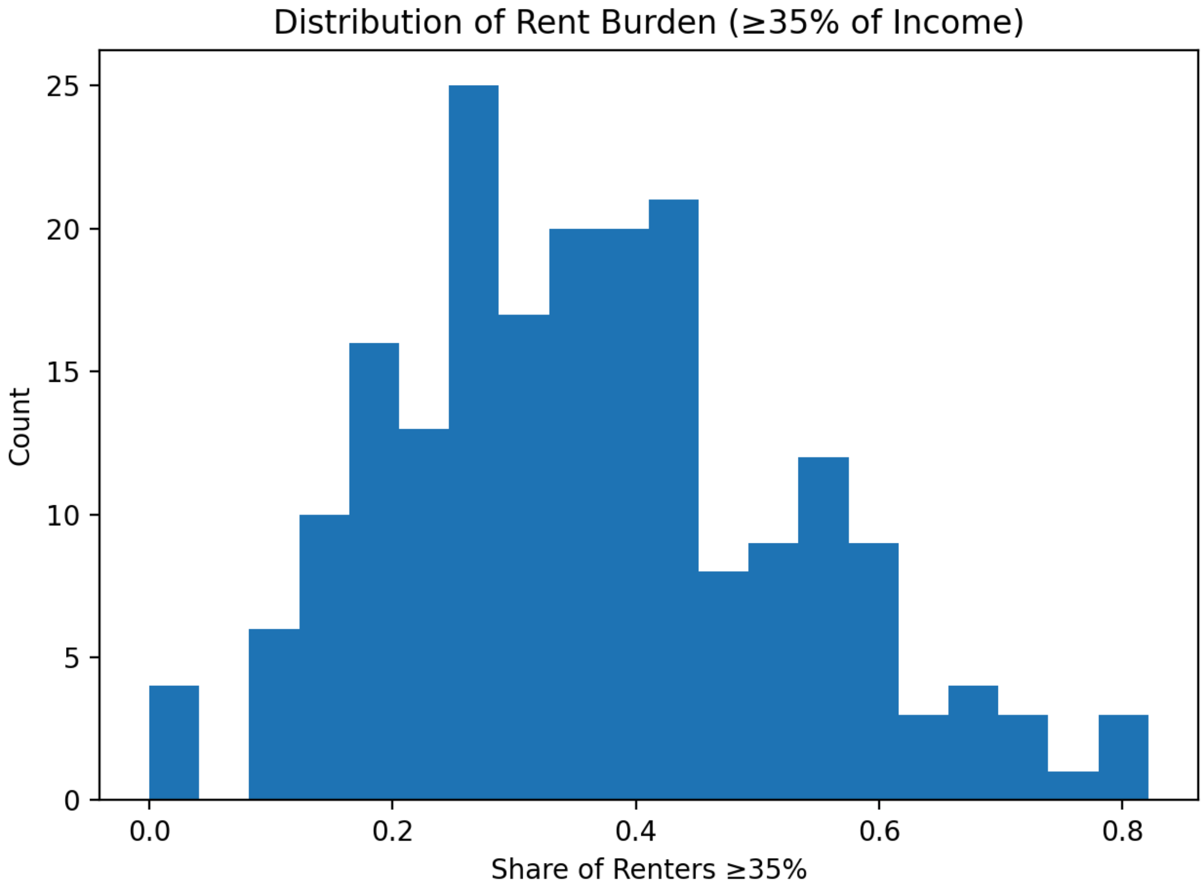
The histogram of vacancy rates showed that most neighborhoods have stable vacancy levels, with a few tracts showing higher instability.



The scatterplot comparing rent and vacancy rate suggested that high-rent areas generally experience lower vacancy, though the relationship is weak.



A boxplot comparing vacancy rate by renter share quartile indicated that neighborhoods with more renters tend to show higher variability in occupancy.



Unexpected findings emerged from the analysis. Contrary to initial expectations, the most expensive neighborhoods did not have the highest vacancy rates. Instead, vacancy appeared concentrated in mid-priced areas where rent may be misaligned with the local renter population's income. This suggests that market mismatch, not high rent alone, drives instability. In addition, some moderately priced tracts showed high rent burden shares, demonstrating that affordability strain is not limited to low-income neighborhoods. These insights reveal that DC's rental stability depends on more than price levels and is influenced by demographic and structural context.

Several limitations were identified during analysis. The ACS dataset's five-year aggregation limits temporal precision, preventing the detection of short-term market shifts or pandemic recovery effects. The dataset lacks household-level income data, which restricts deeper analysis of affordability relative to earnings. It also does not distinguish between single-family and multi-unit properties, which could affect comparisons between neighborhoods. While these limitations reduce specificity, they do not affect the overall reliability of neighborhood-level conclusions.

Refined Problem Statement and Analytical Plan

Based on the findings from exploratory analysis, the project's research question was refined. The updated question is: Where in Washington, DC are small-scale landlords most likely to achieve sustainable and stable rental outcomes that balance profitability with affordability and low vacancy? This version of the question emphasizes the relationship between financial returns and neighborhood stability, recognizing that profitability and affordability must coexist for long-term success.

The analytical approach for the next sprint will use a multivariate Ordinary Least Squares regression model. The dependent variables will be rental vacancy rate and rent burden share. Independent variables will include median rent, renter share, recent movers share, and overcrowded share. Control variables such as median rooms per unit and tract density will be added to account for physical housing characteristics. The model will test whether higher rents and greater mobility correspond to higher vacancy rates and whether overcrowding and affordability constraints predict increased rent burden.

Robust standard errors will be applied to account for heteroskedasticity, and alternative model specifications, including logarithmic transformations of rent, will be tested. The integration of Redfin and Zillow data in Sprint 3 will allow the creation of a Rentability Score that incorporates profitability, stability, and affordability into a single index. Challenges include reconciling different geographic levels between census tracts and ZIP codes, which will be addressed by using an official tract-to-ZIP crosswalk. Multicollinearity among explanatory variables will also be monitored through diagnostic testing and careful variable selection.

Progress Tracking and Next Steps

By the end of Sprint 2, all major goals have been completed. The primary dataset was obtained, cleaned, and validated. Exploratory data analysis was conducted and provided clear evidence of relationships among rent, affordability, and vacancy. A baseline regression model has been planned, and the project repository has been updated with new scripts, documentation, and a professional README file.

In Sprint 3, the focus will shift to integrating secondary data sources, calculating gross rent-to-price yield, and constructing the Rentability Score. This phase will also include new visualizations, such as maps showing rent stability and affordability patterns across DC neighborhoods. Sprint 4 will involve writing the full research report, interpreting the model results, and recording the final presentation.

The project remains on schedule. The primary challenge ahead involves data alignment between the census tract data and ZIP code data, but this can be resolved with appropriate crosswalk files. No major technical or methodological risks remain. The foundation established during Sprint 2 positions the project well for deeper analysis, allowing the next phase to focus on synthesizing data into actionable insights about DC's rental market.

References

District of Columbia Office of the Chief Technology Officer. (2024). *ACS 5-Year Housing Characteristics (DC Census Tract)* [Data set]. Open Data DC. https://opendata.dc.gov/datasets/DCGIS::acs-5-year-housing-characteristics-dc-census-tract/about

Redfin. (2024). *Redfin Data Center: Housing Market Data* [Data set]. Redfin. https://www.redfin.com/news/data-center/

Zillow. (2024). *Zillow Observed Rent Index (ZORI)* [Data set]. Zillow Research. https://www.zillow.com/research/data/