Sprint 2 Report DC Rentability Analysis

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

Modeling Strategy and Implementation

This sprint moves the DC Rentability project from descriptive analysis to concrete statistical modeling of rental vacancy across Washington, DC census tracts. The outcome of interest is the tract-level rental vacancy rate (vacancy_rate_rental) from the ACS 5-year housing characteristics. The central question is how vacancy varies across neighborhoods as a function of rent levels, tenure structure, turnover, overcrowding, and rent burden. A longer-term goal is to turn these relationships into a framework that could support decisions such as "how aggressively can a landlord set rent in a given tract while keeping vacancy risk below some threshold." Given this goal and the data structure, I use Ordinary Least Squares (OLS) regression as the primary modeling approach, treating vacancy_rate_rental as a continuous outcome. The unit of analysis is the census tract, and each observation contains the vacancy rate plus a set of tract-level predictors. The core predictors in this sprint are median_rent, renter_share, recent_movers_share, overcrowded_share, rent_burden35_share, and median_rooms. These variables jointly capture price level, tenure composition, household turnover, crowding, housing cost burden, and housing stock size. OLS is appropriate here because it provides an interpretable linear relationship between vacancy and these predictors, which aligns with the QSSR emphasis on understanding and explaining estimates rather than maximizing black-box predictive performance. This framework is consistent with common practice in applied housing and urban economics research that relies on tract-level ACS data and linear models.This modeling strategy reflects a recognized approach in housing research that combines tract-level (or sub-metropolitan) vacancy data with multivariate regression frameworks. In particular, Du, Wang, Zou, and Shi (2018) demonstrate that census-tract housing vacancy rates can be effectively estimated using multivariable linear regression, even augmenting census data with high-resolution night-light and geospatial indicators. Their successful application of regression at the tract scale provides a methodological precedent for analyzing rental vacancy variation across neighborhoods. Additionally, the U.S. Census Bureau's Housing Vacancies and Homeownership Survey (HVS) establishes standardized definitions of vacancy and vacancy-rate calculation that underpin many of the publicly released datasets used for empirical housing-market analysis. Together, they justify using an OLS regression on tract-level ACS (or similar) data to model rental vacancy in a city-level context such as Washington, DC.

To complement the continuous vacancy model, I also estimate a robustness model that treats vacancy as a binary outcome. Specifically, I create a high_vacancy indicator equal to 1 if a tract's rental vacancy rate is at or above the 75th percentile of the DC distribution and 0 otherwise. In the cleaned dataset used for modeling, this 75th percentile threshold is 10.20 percent. I then fit a logistic regression model that predicts the probability of high_vacancy using the same predictors as in the OLS model. This logistic specification connects directly to risk framing: instead of focusing on the exact vacancy rate, it focuses on whether a tract is unusually vacancy-prone relative to other tracts in the city.

Hyperparameter and design decisions are intentionally simple and transparent. For the OLS specification, I include a constant term and the six predictors listed above, with no interaction or polynomial terms in this sprint. Because residual diagnostics (discussed later) suggest some heteroskedasticity, I estimate OLS with heteroskedasticity-robust (HC3) standard errors rather than relying on homoscedasticity. For the logistic model, I use scikit-learn's LogisticRegression with default regularization and a higher maximum number of iterations to ensure convergence. I do not tune hyperparameters extensively because the primary goal at this stage is interpretability and basic robustness, not squeezing out marginal gains in accuracy.

For data partitioning, I use an 80/20 train–test split for both the continuous and binary models, with a fixed random seed (42) for reproducibility. The OLS model uses a straightforward random split of census tracts into training and test sets. For the logistic model, I perform a stratified train–test split on the high_vacancy indicator to ensure that both high- and non-high-vacancy tracts appear in both sets. To further assess generalization for the OLS model, I apply five-fold cross-validation: in each fold, the model is refit on four-fifths of the tracts and evaluated on the remaining fifth, and I record mean absolute error (MAE) and root mean squared error (RMSE). This combination of a holdout test set and cross-validation provides evidence about how stable the model is within the DC tract context and helps detect overfitting.

Overall, the modeling strategy is conservative but well justified: OLS with robust standard errors on tract-level ACS data is a standard empirical choice for this type of problem, and the logistic robustness check adds a second, risk-focused perspective on vacancy while staying within interpretable statistical frameworks.

Model Development and Training

Model development begins with a simple but essential baseline. The baseline model ignores all predictors and simply predicts the same vacancy rate for every tract in the test set, equal to the mean vacancy_rate_rental in the training data. In this dataset, the baseline_mean from baseline_vacancy_metrics.csv is 6.90 percent (6.8987). When this constant prediction is applied to test tracts, the baseline achieves a mean absolute error of 5.49 percentage points and a root mean squared error of 6.63 percentage points (MAE = 5.4922, RMSE = 6.6264). These numbers mean that if I completely ignore heterogeneity and treat every tract as "average," I am typically off by five to seven percentage points on vacancy. Any more complex model must beat this baseline to be considered useful.

The primary model is an OLS regression predicting vacancy_rate_rental from median_rent, renter_share, recent_movers_share, overcrowded_share, rent_burden35_share, and median_rooms. The OLS summary shows an R-squared of 0.210 and an adjusted R-squared of 0.178, with an F-statistic p-value of 0.00296, indicating that the model jointly explains a statistically significant portion of variation in vacancy across tracts. The robust HC3 coefficient results show that, holding other variables constant, median_rent has a positive coefficient (0.0021, $p = 0.018$), renter_share has a negative coefficient (-7.7167, $p = 0.022$), overcrowded_share has a negative coefficient (-21.9602, $p = 0.028$), rent_burden35_share has a positive coefficient (7.6204, $p = 0.021$), and median_rooms has a negative coefficient (-1.8619, $p = 0.001$). recent_movers_share has a negative coefficient (-3.7287) but is not statistically significant ($p \approx 0.582$).

From an error perspective, the OLS model substantially improves on the baseline. According to ols_vacancy_metrics.csv, the training MAE is 4.10 and training RMSE is 5.20 (train_mae = 4.1039, train_rmse = 5.1973). On the held-out test set, the MAE is 4.40 and the RMSE is 5.68 (test_mae = 4.3966, test_rmse = 5.6836). Compared to the baseline MAE of 5.49 and RMSE of 6.63, this represents an improvement of about one percentage point in mean absolute error and about one percentage point in RMSE on the test set. The gap between training and test errors is modest, which suggests that the model is not severely overfitting and that the learned relationships generalize reasonably well to unseen tracts.

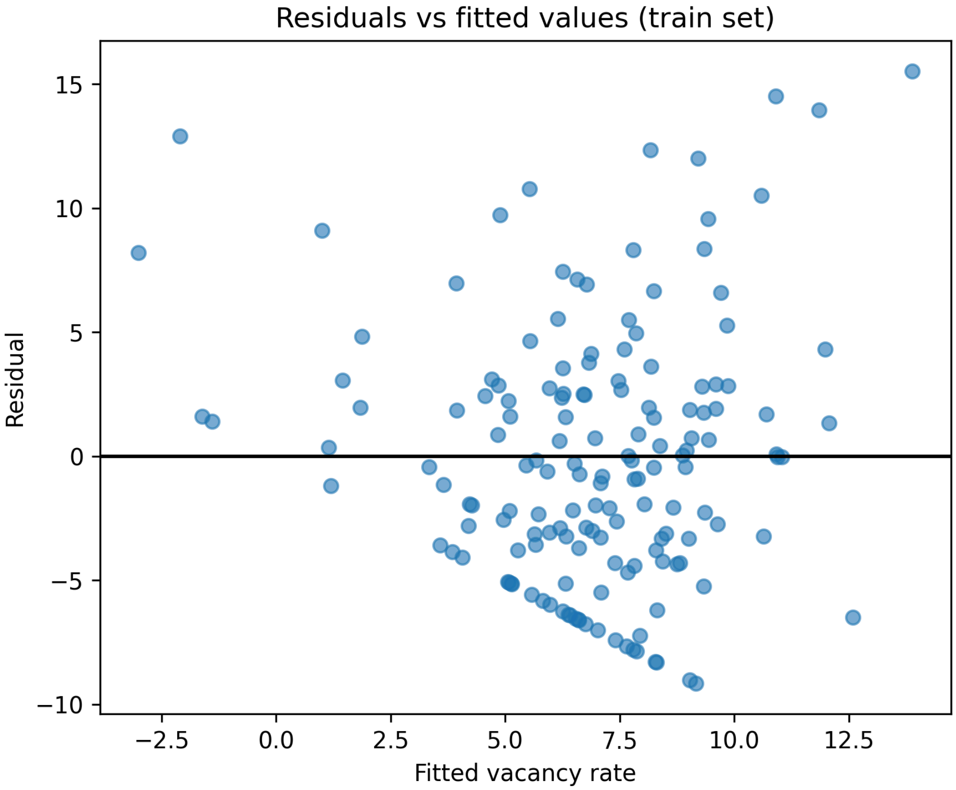
To further evaluate stability, I implement five-fold cross-validation using the same specification. The results in ols_vacancy_cv_metrics.csv show fold-level MAE values of 4.40, 4.09, 4.03, 4.03, and 5.18, and RMSE values of 5.68, 5.28, 5.02, 5.17, and 6.30. The mean MAE across folds is 4.35 (with standard deviation about 0.49), and the mean RMSE is 5.49 (with standard deviation about 0.51). These cross-validated metrics are similar to the single train–test metrics and show moderate variability across folds, indicating that results are fairly robust to which tracts are in the training versus validation sets.

The robustness logistic model uses the binary high_vacancy outcome, where high_vacancy is 1 when vacancy_rate_rental is at or above 10.20 percent, the 75th percentile, and 0 otherwise. Using the same predictors and a stratified 80/20 split, the logistic regression produces the metrics summarized in logit_high_vacancy_metrics.csv. The area under the ROC curve is 0.77, and overall accuracy on the test data is 0.80. The confusion matrix entries are tn = 30, fp = 0, fn = 8, tp = 2, meaning the model correctly identifies all low-vacancy tracts in the test set but only two of ten high-vacancy tracts. The coefficient table in logit_high_vacancy_coefficients.csv shows a negative intercept (-0.3517) and suggests that higher median_rent (coef ≈ 0.00080) and higher rent_burden35_share (coef ≈ 0.6301) are associated with higher log-odds of being in the high-vacancy group, while higher renter_share, recent_movers_share, overcrowded_share, and median_rooms are associated with lower log-odds (negative coefficients). Because this logistic model comes from scikit-learn, I do not have p-values for these coefficients, so I treat this pattern as descriptive rather than formally significant.

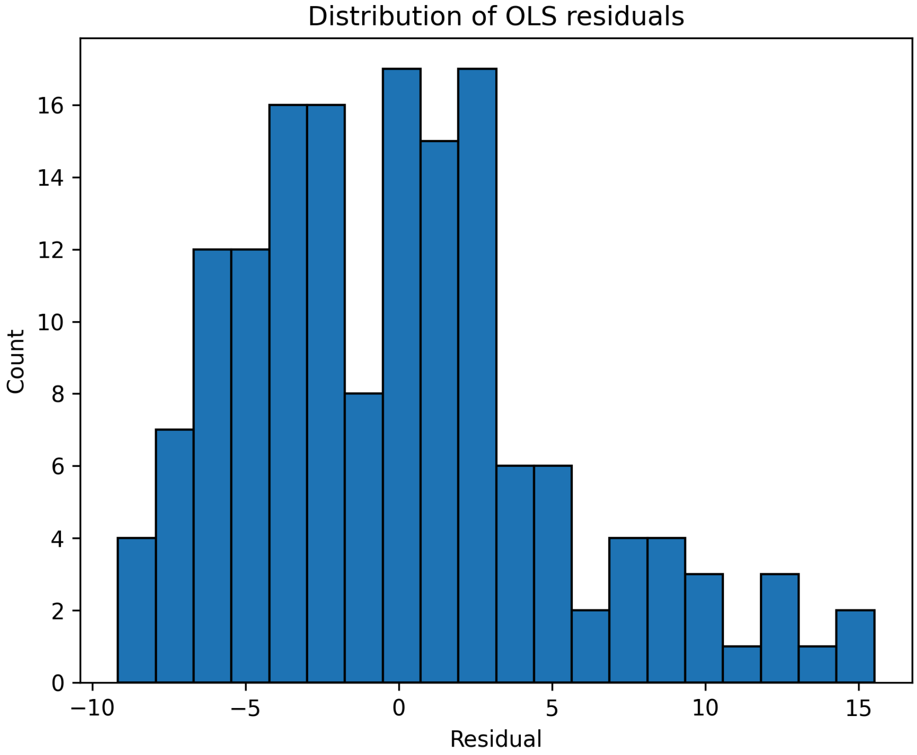
Feature engineering in this sprint stays intentionally focused. The main engineered elements are the binary high_vacancy indicator used in the logistic model and the standardized coefficients for the OLS model. The standardized coefficients, stored in ols_vacancy_standardized_coefficients.csv, express each predictor's effect in standard deviation units and allow direct comparison of their relative importance. In that table, median_rooms has the largest absolute standardized coefficient (-0.527), followed by renter_share (-0.292), median_rent (0.236), rent_burden35_share (0.219), overcrowded_share (-0.116), and recent_movers_share (-0.061, which is also the only non-significant predictor). These standardized coefficients form the basis for feature importance discussions in the next section.

Model Evaluation and Diagnostics

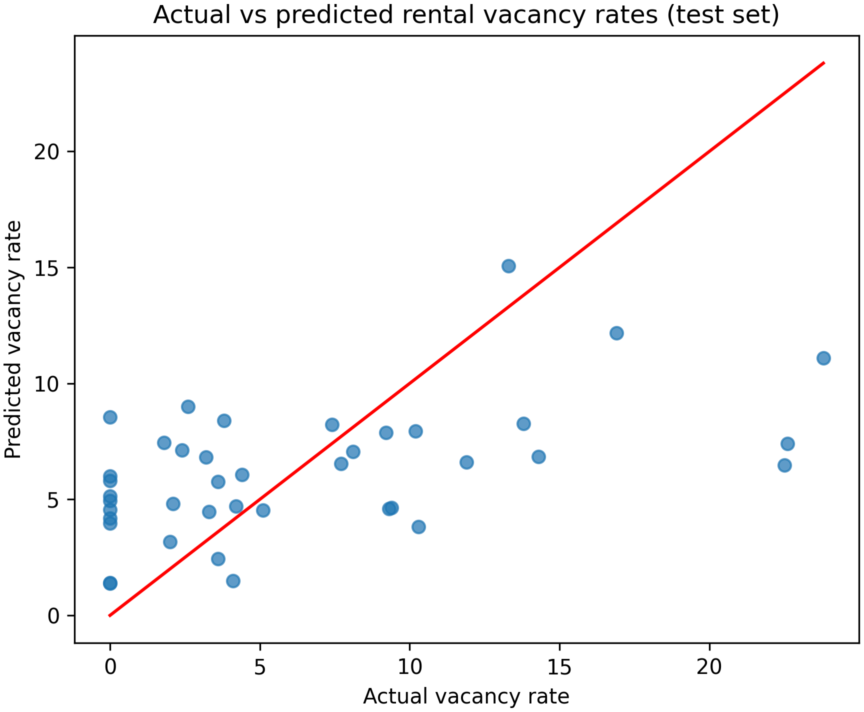
The starting point for evaluation is comparing training and test errors for the OLS model. Training MAE (4.10) and RMSE (5.20) are slightly lower than test MAE (4.40) and RMSE (5.68), as expected when moving from model fitting to out-of-sample evaluation. The fact that the differences are modest suggests that the model is not dramatically overfitting and that its performance is fairly stable across the sample. The cross-validation results reinforce this conclusion: the mean MAE across five folds is 4.35 and the mean RMSE is 5.49, with only moderate variation across folds. These values are close to the single test-set metrics, which supports the idea that the model generalizes reasonably well within the DC tract context. Residual diagnostics provide a deeper look at the OLS assumptions. The residuals-versus-fitted plot for the training data shows residuals scattered around zero across the full range of predicted vacancy values. There is no strong evidence of nonlinear patterns such as pronounced curvature, and while there is some increase in residual spread at higher fitted values, the pattern is not extreme. This mild heteroskedasticity is one reason for using HC3 robust standard errors in the model estimation. Overall, the residual plot supports the linear specification as a reasonable first-order approximation rather than indicating a major misspecification.



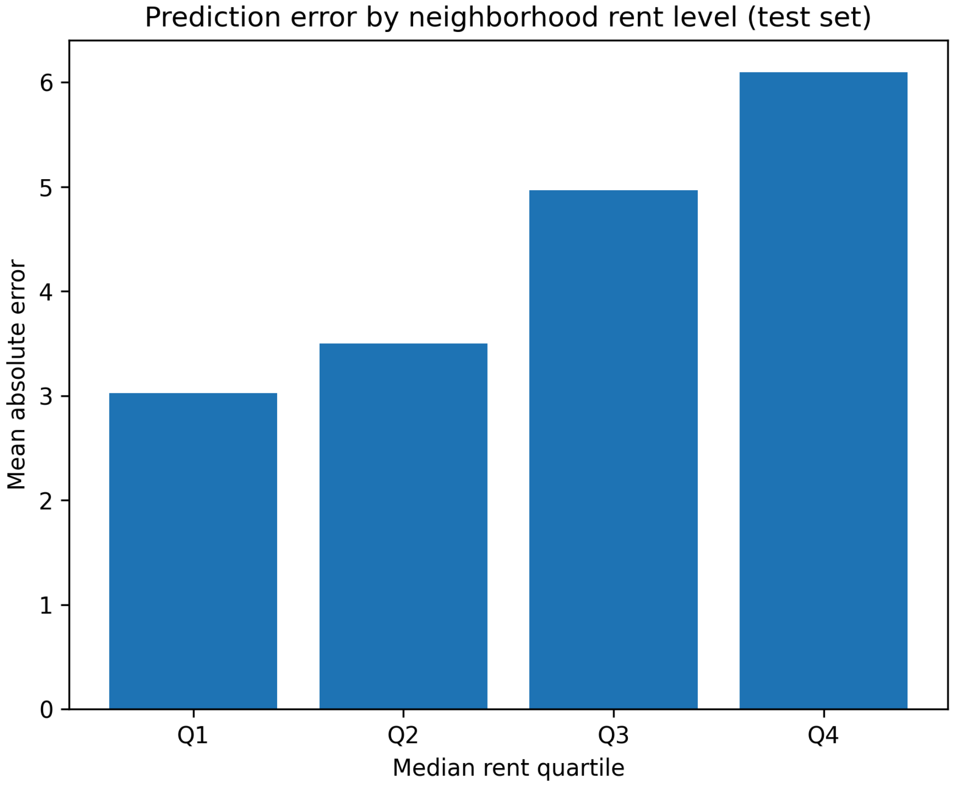
The residual histogram shows that residuals are roughly symmetric around zero, with a moderate concentration near zero and heavier tails that are common in cross-sectional neighborhood data. The Omnibus and Jarque–Bera tests reported in the OLS summary suggest some departure from perfect normality, but given the sample size (156 tracts used in the final regression) and the use of robust standard errors, these departures are not fatal. They indicate that there are some outlier or extreme cases, which is expected when modeling local housing markets.



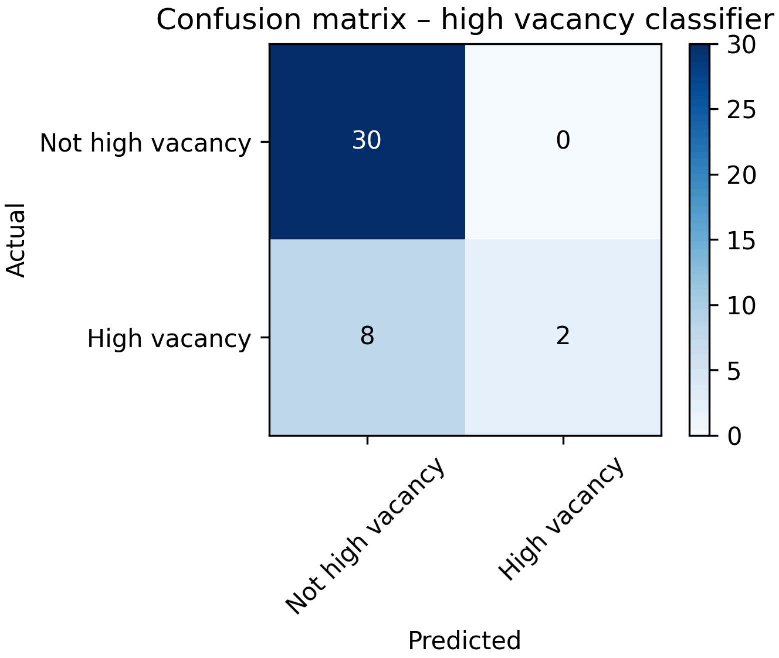
An actual-versus-predicted scatterplot for the test set provides an intuitive sense of predictive performance. Each point corresponds to a tract, with its actual vacancy rate on one axis and the predicted vacancy rate from the OLS model on the other. Most points lie reasonably close to the 45-degree line, especially for vacancy rates in the middle of the distribution. Deviations are larger at very low and very high vacancy rates, which is typical: extreme cases often reflect idiosyncratic local conditions or omitted variables. Taken together with the error metrics, this visualization confirms that the model meaningfully improves over the constant-mean baseline and captures real structure in the data.

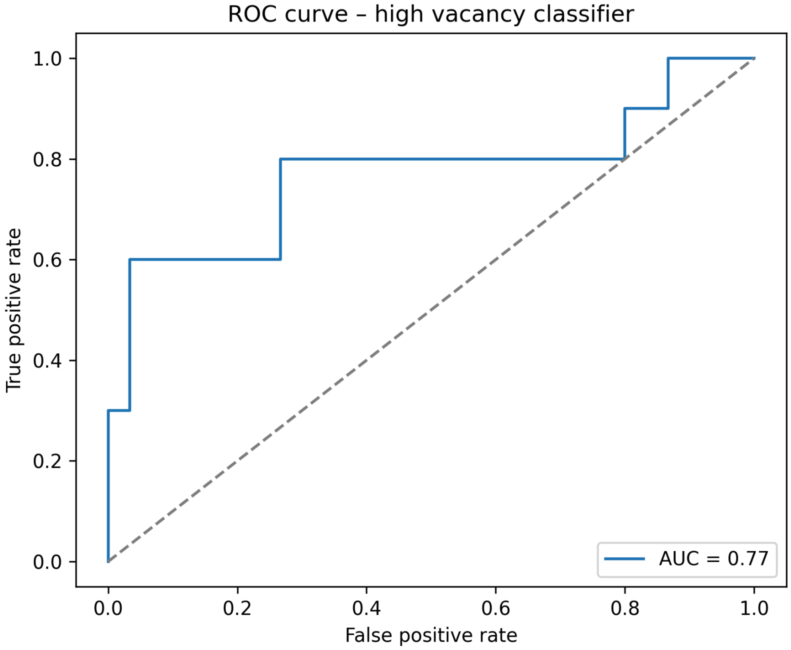


To explore how prediction error varies across different rent environments, I compute mean absolute error by quartiles of median_rent for the test tracts. The resulting bar chart shows the mean absolute error for each rent quartile. In my results, errors are generally smallest in the middle quartiles and somewhat larger in the lowest and highest rent quartiles. This pattern is intuitive. In lower-rent neighborhoods, factors such as building condition, safety, or informal rental arrangements may strongly influence vacancy but are not directly captured by the ACS predictors. In the highest-rent neighborhoods, niche markets and luxury dynamics may make vacancy more volatile and harder to capture with tract-level averages.



The logistic high-vacancy model is evaluated using classification metrics. An AUC of 0.77 means that, when comparing one high-vacancy and one non-high-vacancy tract at random, the model assigns a higher predicted probability of high vacancy to the true high-vacancy tract roughly 77 percent of the time. An overall accuracy of 0.80 is also strong, but the confusion matrix reveals that the default 0.5 probability threshold produces a conservative classifier: there are 30 true negatives and no false positives, but there are also 8 false negatives and only 2 true positives. In other words, the model almost never incorrectly labels a low-vacancy tract as high-vacancy, but it fails to capture many of the actual high-vacancy tracts. In a setting where missing high-vacancy tracts is costly, it might be preferable to lower the threshold so that more tracts are flagged at the cost of some false positives.





Overall, the diagnostics suggest that the OLS model is reasonably well-behaved: it improves meaningfully over the baseline, has consistent performance across folds, and passes basic residual checks without severe violations. The logistic model behaves as a conservative classifier at the default threshold but has decent ranking ability, which is useful for identifying higher-risk tracts if thresholds are adjusted in future work.

Results and Interpretation

The main substantive result from Sprint 3 is that rental vacancy across DC census tracts is structured and meaningfully associated with rent levels and neighborhood housing characteristics. The OLS model's test MAE of 4.40 and RMSE of 5.68 compare favorably to the baseline MAE of 5.49 and RMSE of 6.63, indicating that incorporating tract-level predictors reduces typical prediction errors by around one percentage point. An R-squared of 0.210 implies that the model explains about 21 percent of the variation in vacancy rates across tracts, which is a significant share given the complexity of local housing markets.

The standardized coefficients in ols_vacancy_standardized_coefficients.csv provide a clear ranking of predictor importance. median_rooms has the largest absolute standardized coefficient at about -0.53, suggesting that, in standard-deviation terms, the typical effect of changing median room count is stronger than any other predictor in the model. Tracts with higher median_rooms tend to have lower vacancy, all else equal, which is consistent with the idea that larger, higher-quality housing stock is associated with more stable occupancy. The next most influential predictor is renter_share, with a standardized coefficient of -0.29: tracts with higher shares of renters tend to have lower vacancy rates, holding other factors constant. This might reflect strong demand in renter-dominated neighborhoods or institutional rental markets with more professional property management.

Median_rent and rent_burden35_share both have positive standardized coefficients (0.24 and 0.22 respectively), indicating that higher rents and higher shares of rent-burdened households are associated with higher vacancy, once other variables are controlled for. This aligns with the intuitive idea that aggressive pricing and affordability strain can push some units into longer spells of vacancy. overcrowded_share has a smaller but still meaningful negative standardized coefficient (-0.12), suggesting that tracts with more overcrowding are somewhat less likely to have high vacancy, possibly reflecting tight housing conditions. recent_movers_share has the smallest standardized coefficient (-0.06) and is not statistically significant in the OLS model, indicating that, once other variables are in the model, turnover is not a strong independent predictor of vacancy in this dataset.

These results can be used to frame the “maximum sustainable rent” idea. Because the OLS model is linear, it can be written as an equation connecting vacancy_rate_rental to median_rent and the other predictors. For a given tract with known renter_share, recent_movers_share, overcrowded_share, rent_burden35_share, and median_rooms, the model can be used to determine what value of median_rent would correspond to a chosen target vacancy rate, such as 5 or 10 percent. This inversion does not imply that changing rent will causally change vacancy to that exact level, but it gives landlords or policymakers a structured way to think about how aggressive a rent level is relative to vacancy patterns in similar tracts. It moves the conversation from a purely intuitive “this feels too high” to “the model suggests that at this rent level, a tract with these neighborhood characteristics might expect around X percent vacancy.”

The logistic high-vacancy model offers a complementary perspective centered on risk. With a threshold of 10.20 percent defining high vacancy, the model’s AUC of 0.77 indicates that the same predictors have meaningful ability to distinguish high-vacancy tracts from others. The coefficient signs in logit_high_vacancy_coefficients.csv suggest that, holding other variables constant, higher median_rent and higher rent_burden35_share are associated with increased log-odds of being a high-vacancy tract, while higher renter_share, overcrowded_share, recent_movers_share, and median_rooms are associated with decreased log-odds. In plain language, expensive, highly rent-burdened areas are more likely to be in the high-vacancy group, while renter-dominated, more crowded, and larger-unit tracts are less likely to be in that upper tail of vacancy. Because I do not have p-values for the logistic coefficients, I treat these as descriptive patterns that broadly align with the OLS interpretation rather than as definitive causal claims.

Together, the OLS and logistic models show that vacancy is not random noise. It is systematically related to rent levels, affordability stress, housing stock, and tenure composition. The continuous model provides a baseline mapping between predictors and expected vacancy, while the logistic model highlights which conditions are associated with being in the highest-vacancy quartile. Both models materially improve on the constant-mean baseline, and both support the idea of using tract-level data to build a Rentability Score or rent–vacancy tradeoff tool in the final sprint.

Limitations, Assumptions, and Threats to Validity

Despite the progress in this sprint, there are several important limitations that shape how these results should be interpreted. On the data side, all variables are derived from ACS 5-year estimates. These estimates carry sampling error, particularly for smaller tracts or those with unusual characteristics. In this sprint, I treat the published ACS point estimates as exact values and do not propagate margins of error or sampling variability into the regression analysis. This means that coefficient standard errors and p-values may be somewhat optimistic relative to a fully measurement-error-aware approach.

The dataset is also cross-sectional. It captures vacancy and housing characteristics at essentially one point in time, without explicit information about how vacancy evolves over time in response to changing rents, new construction, or policy interventions. As a result, the models estimate static relationships between vacancy and neighborhood characteristics, not dynamic responses or causal impacts. Additionally, the predictor set is limited to ACS housing variables and a few derived features. Many potentially important drivers of vacancy such as crime, transit access, amenities, school quality, short-term rental activity, or landlord behavior are not included. If these omitted variables are correlated with the included predictors, the estimated coefficients may capture some of their effects, leading to omitted-variable bias.

Methodologically, the OLS model assumes linear relationships between the predictors and vacancy. While residual plots suggest that linearity is a reasonable first-order approximation, the true relationships could be nonlinear, especially at the extremes of rent or vacancy. The condition number reported in the OLS summary is relatively large (around 7.12e4), which signals potential multicollinearity among some predictors. Although the robust standard errors help mitigate some issues, multicollinearity can inflate standard errors and make coefficient estimates more sensitive to small changes in the data. Additionally, I treat tracts as independent observations, but in reality they are spatial units, and vacancy in one tract may be related to vacancy in adjacent tracts through shared amenities, neighborhood reputation, or spillovers. I have not yet performed spatial autocorrelation tests such as Moran’s I on the residuals, so spatial dependence remains a possible threat to validity.

Causality is a central limitation. Both the OLS and logistic models are descriptive and estimate conditional associations. They do not identify causal effects of changing median_rent, renter_share, or any other predictor on vacancy. In reality, rents, vacancies, and neighborhood characteristics evolve jointly in response to underlying demand, supply, and policy. For example, higher rents and lower vacancy might both be driven by high demand in a desirable neighborhood rather than rent changes causing vacancy changes. For this reason, I interpret the "maximum sustainable rent" concept as a scenario analysis: it uses the model as a structured benchmark for thinking about vacancy risk at different rent levels but does not guarantee that raising or lowering rent to a given level will produce the corresponding vacancy predicted by the model.

Generalizability is limited by geography and time. The model is estimated on 196 tracts in Washington, DC, using data from a specific ACS period. Patterns may differ in other cities, suburban areas, or rural regions. Even within DC, future changes in policy, infrastructure, or the broader economy could alter the relationships between rent, neighborhood characteristics, and vacancy. Without retraining the model on updated data, its accuracy and relevance could decline. Finally, there may be selection and measurement issues, such as variation in how vacancy is reported, that are not fully captured in this analysis.

Taken together, these limitations mean that the model is best interpreted as a descriptive tool for understanding current tract-level patterns in DC and for supporting structured, scenario-based reasoning about rent and vacancy, rather than as a causal model or a universally generalizable prediction system.

Sprint 4 Plan and Refinement Strategy

Sprint 4 will focus on refining the modeling work from this sprint and translating it into a more decision-oriented framework, with the goal of making the results actionable for non-technical stakeholders such as small landlords or local policymakers. A central deliverable will be a Rentability Score that uses the OLS model's predicted vacancy along with rent and affordability indicators to rate tracts on how "rentable" they are under different price scenarios. A simple version of this score might reward tracts where predicted vacancy is low or moderate at reasonable rent levels and penalize tracts where high rent and high rent burden coincide with elevated vacancy risk.

On the technical side, I plan to add at least two refinements. First, I will run spatial diagnostics on the OLS residuals, such as calculating a Moran's I statistic, to test for spatial autocorrelation. If residuals show significant spatial clustering, that would suggest that a spatial error or spatial lag model could be a valuable extension. Depending on time and course expectations, I may either implement a simple spatial model or explicitly acknowledge spatial dependence as a limitation and direction for future research. Second, I will explore adding one or two additional predictors that can be merged at the tract level without extensive cleaning, such as a simple transit accessibility measure or a crime indicator, to see whether they meaningfully improve model fit and reduce omitted-variable concerns.

Sprint 4 will also emphasize communication and packaging. I will draft a non-technical summary that describes the key findings in plain language, including how vacancy relates to rent, tenure, housing stock, and affordability. I will select and refine a small set of core visualizations such as the standardized coefficient bar chart, an updated actual-versus-predicted plot, the error-by-rent-quantile plot, and possibly a map of predicted vacancy or Rentability Score by tract to include in the final report and presentation. For the final set of visuals in Sprint 3, I focus on a small group of figures that collectively tell the story of tract-level rental vacancy in DC. The key diagnostic and interpretive plots are sprint3_standardized_coefficients.png (ranking standardized coefficients for all predictors), sprint3_residuals_vs_fitted_train.png and sprint3_residual_histogram_train.png (checking linearity and residual distribution), sprint3_actual_vs_predicted_test.png (showing overall predictive performance on the test set), and sprint3_error_by_rent_quartile_test.png (highlighting where the model performs better or worse across the rent distribution). For the high-vacancy robustness check, I include sprint3_logit_roc_curve.png and sprint3_logit_confusion_matrix.png as summary visuals of classification performance. I do not yet include map-based visualizations of vacancy or prediction error, but those could be added in a later sprint to spatially contextualize which DC tracts are most at risk of sustained vacancy. In parallel, I will ensure that the GitHub repository clearly documents how to reproduce the analysis, including instructions for running src/models/vacancy.py and for generating the figures and outputs used in the report.

The main risks for Sprint 4 are scope creep and time. Adding spatial models and new external datasets can quickly make the project more complex than necessary for the course. To manage this, I will prioritize finalizing the OLS and logistic models, defining and implementing the Rentability Score, cleaning up documentation, and crafting high-quality narrative and visual explanations. Advanced elements such as spatial modeling or richer external variables will be treated as stretch goals. I plan to seek guidance from the instructor or TAs on how much depth they expect on spatial dependence and whether a well-diagnosed OLS model plus clear discussion is sufficient for the QSSR track.

Self-Assessment

At the end of Sprint 3, I am generally on track with my project timeline. Data cleaning and feature preparation from earlier sprints are stable, and I now have a baseline model, a primary OLS model, and a logistic robustness model implemented, evaluated, and saved in the repository. The biggest win of this sprint is that vacancy is no longer just an abstract concept in the proposal; it is represented by concrete models that clearly improve over a naive baseline, have interpretable coefficients, and behave consistently under cross-validation. I also now have standardized coefficients and classification metrics that will make it easier to communicate "what matters most" to non-technical audiences.

The biggest challenge has been choosing an appropriate level of complexity. It is tempting to add interactions, nonlinear transformations, spatial dependence structures, and multiple external datasets all at once. However, each layer adds diagnostic and interpretation overhead. In this sprint, I deliberately kept the model relatively simple and instead focused on getting the basics right: transparent specification, robust standard errors, sensible train–test and cross-validation strategies, and clear error and residual diagnostics. Given that, I would rate my confidence in the current results at around 8 out of 10. The models clearly outperform the baseline and produce coherent, interpretable patterns, but known limitations around omitted variables, spatial dependence, and causal interpretation keep me from rating them higher.

From a documentation standpoint, the GitHub repository is in good shape for this stage. The main modeling script for this sprint lives at src/models/vacancy.py. The cleaned ACS-plus-features dataset is stored under data/processed/dc_acs_cleaned_with_features.csv. The key output files generated in this sprint baseline_vacancy_metrics.csv, ols_vacancy_metrics.csv, ols_vacancy_cv_metrics.csv, ols_vacancy_standardized_coefficients.csv, logit_high_vacancy_metrics.csv, and logit_high_vacancy_coefficients.csv are stored in the outputs/ directory. Figures used in the report are saved in the figures/ directory with descriptive names such as sprint3_standardized_coefficients.png, sprint3_residuals_vs_fitted_train.png, sprint3_residual_histogram_train.png, sprint3_actual_vs_predicted_test.png, and sprint3_error_by_rent_quartile_test.png. The README documents the overall project structure and how to run the scripts, and I will extend it in Sprint 4 to reference the final report and any additional deployment artifacts.The completed Sprint 3 report has also been added to the repository as reports/Sprint_3_Report.pdf, and the README has been updated to reflect the full modeling workflow, new outputs, and instructions for running the vacancy modeling script.

Going into Sprint 4, the main support I would find helpful from instructors or TAs is feedback on two points: first, whether the "maximum sustainable rent" and Rentability Score framing is an appropriate and honest way to interpret the OLS model within the QSSR expectations; and second, how much emphasis they expect on addressing spatial dependence and external data sources versus focusing on a well-explained, thoroughly diagnosed OLS model. With that guidance, I can tune the scope of the final sprint to balance rigor, interpretability, and the practical constraints of time.

References

Du, M., Wang, L., Zou, S., & Shi, C. (2018). *Modeling the census tract level housing vacancy rate with the Jilin-1-03 satellite and other geospatial data*. Remote Sensing, 10(12), 1920. https://doi.org/10.3390/rs10121920 MDPI+1

U.S. Census Bureau. (n.d.). *Housing Vacancies and Homeownership Survey (HVS): Definitions & Explanations*. Retrieved November 28, 2025, from https://www.census.gov/housing/hvs/definitions.pdf Census.gov+1