PERSPECTIVES ON DATA SCIENCE CS746

PROJECT MILESTONE – 2

GROUP 18

GROUP MEMBERS:

1. ABRAAR MOHAMMED (B594M942) – **Team Lead**
2. SYED UMER TARIQ (X785Y683)
3. FNU ABDULLAH (W665G463)
4. FNU SYED ABDUL RAHMAN (U778E584)
5. ABUL KHALIQUE BAIG MIRZA (F884W446)
6. MUDASSIR KAHDER ELAHEE KHADER QURYSHI (W879Q599)

PROJECT AIM: To create a website aimed at guiding international students towards the best housing options available to them.

PROJECT RESOURCES:

1. Questionnaire Form
2. Questionnaire Results
3. Wichita Transit Information
4. Independent Houses
5. Apartments
6. Restaurants and Groceries

STEPS IN DATA ANALYSIS (MILESTONE – 1):

1. DATA CLEANING
2. EXPLORATORY DATA ANALYSIS (EDA)
3. REGRESSION ANALYSIS

INDIVIDUAL CONTRIBUTION

Our team has deliberated and acted on the following elements:

1. ABUL KHALIQUE BAIG MIRZA (F884W446)
2. MUDASSIR KAHDER ELAHEE KHADER QURYSHI (W879Q599)
    - **Identified the necessary datasets and the resources utilized for this purpose.**
3. FNU ABDULLAH (W665G463)
4. FNU SYED ABDUL RAHMAN (U778E584)
    - **Developed strategy for analysis and performed data cleaning and wrangling.**
5. ABRAAR MOHAMMED (B594M942)
6. SYED UMER TARIQ (X785Y683)
    - **Conducted Exploratory Data Analysis (EDA) on the dataset to uncover insights.**

INTRODUCTION

In the evolving landscape of academia, the "Perspectives on Data Science CS 746" project stands out as an innovative approach aimed at enhancing the housing search experience for international students, thanks to the application of data science. This initiative, driven by Group 18, focuses on analyzing housing, transit, and survey data to offer students abroad practical advice on selecting the best living options.

The project unfolds through rigorous data cleaning and exploratory data analysis (EDA), essential steps that ensure the reliability of our findings and support smart housing decisions. Our team's collaborative work emphasizes the critical role of accurate data and deep insights in understanding the complex preferences of students when it comes to housing. This brief report encapsulates our analytical process and the transformative potential of data science in improving the lives of international students, highlighting a path forward where data-driven insights facilitate better housing choices and, ultimately, a better student experience.

Datasets Used:

Our team has examined all the resources supplied by the professor, as mentioned in the PROJECT RESOURCES section above. For implementation, we primarily selected files that match our project objectives, focusing on three key files: Questionnaire Responses, Independent Houses, and Restaurant data files.

Data Cleaning Procedures Used:

1. We loaded all the project-related files into our Jupyter Notebook environment, determining the dimensions of the dataset as (1199, 42) and primarily focusing on the data type, especially for the Questionnaire Responses, before applying similar considerations to the other resources.
2. Given the extensive number of columns - 42 for Questionnaire Responses, 20 for Apartment data, 5 for Independent Houses, 12 for Restaurants, and 11 for Transit Stop data - we developed a code to efficiently retrieve column names from any Excel sheet based on user input.
3. We assessed the presence of missing values in each column of the Questionnaire Responses, categorizing the data into numerical and categorical. We then standardized the missing values to "Unknown", "No", or "NaN", depending on their context, and applied median or mode where appropriate without compromising the dataset's integrity

for further analysis.

4. We refined the data by removing white spaces, converting text strings representing numeric ranges into precise numeric values, and standardizing the format of square footage, distances, room and bathroom counts, and address information for clarity and consistency.
5. We eliminated the column asking if the respondent wished to complete the questionnaire due to a high number of missing responses (1172) and its redundancy.
6. We meticulously renamed columns, performed a preliminary count of missing values, replaced these with "NaN" or "Unknown", and applied median or mode adjustments based on the data classification.
7. For data such as the total monthly rent, we set a cap at $3000, corrected erroneous entries to "NaN", and ensured numeric conversions were accurate, incorporating descriptive statistics for a comprehensive overview where needed.
8. We imputed missing values using appropriate statistics, like the median for numerical columns, to maintain data integrity.
9. We identified and addressed potential outliers, ensured data type accuracy and consistency, and made necessary corrections.
10. We utilized box plots to visualize the distribution of data, providing insights into the dataset's characteristics, summarizing the data types, and reviewing unique column values post-cleanup to catch any discrepancies that could benefit from further refinement.

Exploratory Data Analysis (EDA) Process:

Our exploratory data analysis (EDA) methodology was characterized by a sequential and detailed approach aimed at comprehensively understanding the dataset derived from the Questionnaire Response Form. This rigorous process encompassed several critical steps, outlined below:

1. **Data Importation and Initial Assessment:** The journey commenced with the importation of all pertinent data from the Questionnaire Response Form into our analytical framework. This foundational step, executed in cells 1 and 2, set the stage for a thorough initial assessment. We scrutinized the dataset's fundamental attributes, including data types, dimensions, and the presence of missing values. Utilizing functions such as **head**, **info**, and **describe**, we gained a preliminary insight into the dataset's structure, paving the way for a more granular analysis.

2. **Descriptive Statistical Analysis:** Advancing to the next level of our EDA, we embarked on a comprehensive examination of descriptive statistics for each column within the questionnaire. This endeavor allowed us to delve into the nuances of the dataset, offering a detailed exploration of every facet of the Questionnaire Responses. This step was instrumental in painting a vivid picture of the data at hand.
3. **Focused Variable Analysis:** With a robust understanding of the data's overarching structure and its alignment with our project objectives, we honed in on the variables critical to our hypothesis testing: respondents' satisfaction with their current living situation and the total monthly rent of their units. This phase of our analysis was multifaceted, encompassing several key analyses:
    - **Distribution Analysis:** We meticulously analyzed the distribution of satisfaction levels and monthly rents, seeking to understand the spread and central tendencies of these variables.
    - **Outlier Identification:** Through the creation of boxplots, we pinpointed outliers in satisfaction levels and rent amounts, ensuring our analysis accounted for these anomalies.
    - **Relationship Exploration:** A scatter plot was generated to explore the potential relationship between rent and satisfaction levels, visually inspecting the dynamics between these key variables.
    - **Correlation Analysis:** We calculated the correlation coefficient to quantitatively assess the relationship between rent and satisfaction, supplementing our visual analyses with statistical evidence.
    - **Summary Statistics Compilation:** Summary statistics for both variables were compiled, providing a comprehensive snapshot of their characteristics.
4. **Distribution Analysis of Numerical Data:** Expanding our analysis, we assessed the distribution of all numerical data within the dataset. This step involved visualizing these distributions to achieve a deepened understanding of the data's characteristics, further enriching our analysis.
5. **Insights into Rent and Satisfaction Correlation:** The culmination of our analyses was presented through the correlation coefficient alongside summary statistics for both satisfaction levels and monthly rent. Our findings indicated a correlation coefficient of approximately 0.117, suggesting a "WEAK POSITIVE" link between rent and satisfaction levels. This reveals a "MINOR INCLINATION FOR SATISFACTION TO RISE AS RENT INCREASES", though the correlation's strength is insufficient for

definitive conclusions without further statistical scrutiny.

6. **Advanced Multivariate Analysis:** Venturing beyond basic analyses, we conducted a Multivariate Analysis and developed a correlation matrix. This sophisticated analytical technique allowed us to uncover and understand the interrelationships among multiple variables within our dataset, offering a more nuanced understanding of the data's dynamics.

Visualization Results for EDA

Figure 1: Distribution of Rent Prices Figure 1 presents a box plot that illustrates the distribution of rent prices. The plot highlights the median rent price, depicted by the orange line, as well as the interquartile range which shows where the middle 50% of rent prices fall. The whiskers extend to show the range of the data excluding outliers.

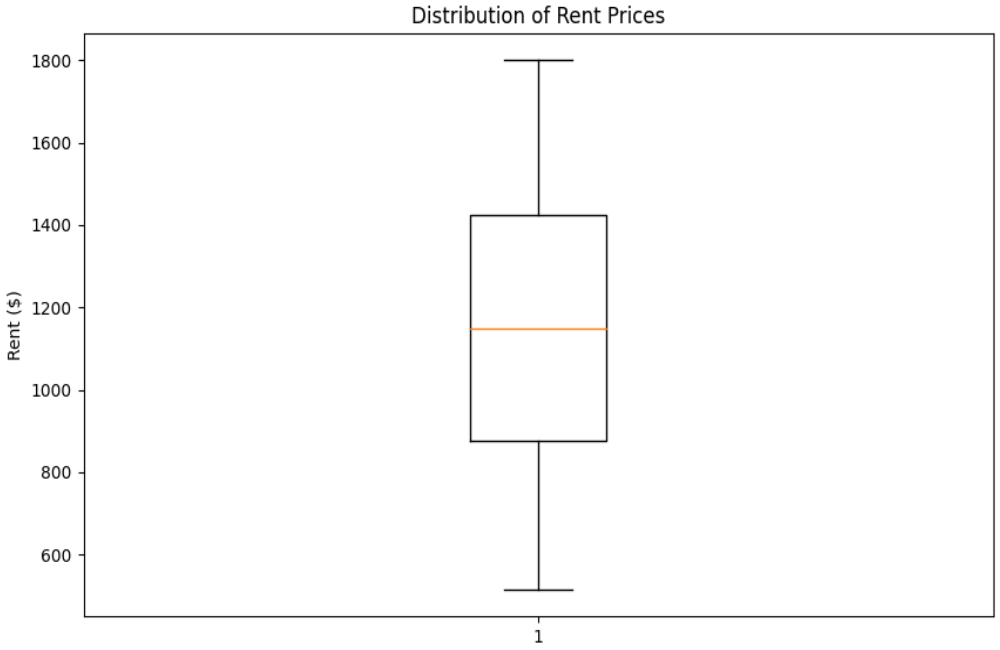


Figure 2 shows a histogram overlaid with a line graph showing the distribution of respondents' overall satisfaction with their current living conditions, on a scale from 1 to 5. The bars represent the count of responses for each satisfaction level, and the line graph emphasizes the distribution pattern, showing peaks at specific satisfaction levels, suggesting clusters of responses around these points.

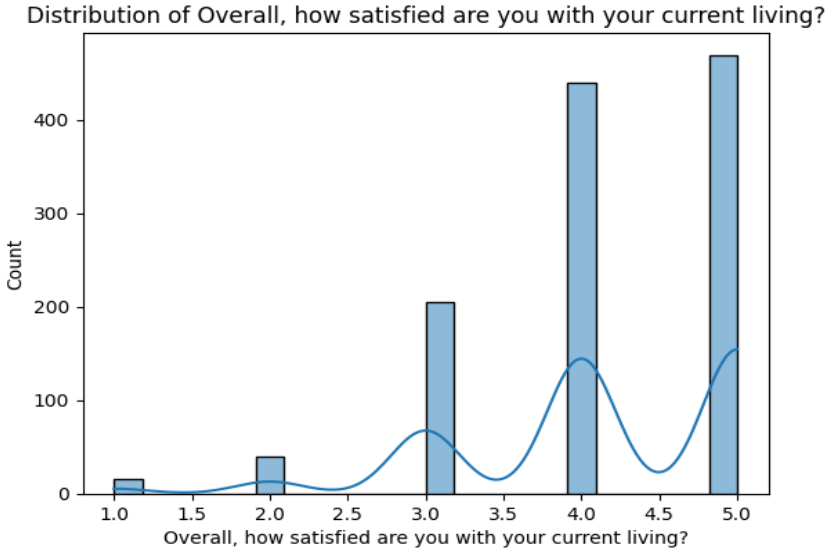


Figure 3 displays a histogram with a line graph detailing the efficiency of management in apartment complexes, rated on a scale of 1 to 5. The histogram shows a distribution of responses, with peaks indicating common ratings for the management's responsiveness and efficiency in maintenance services.

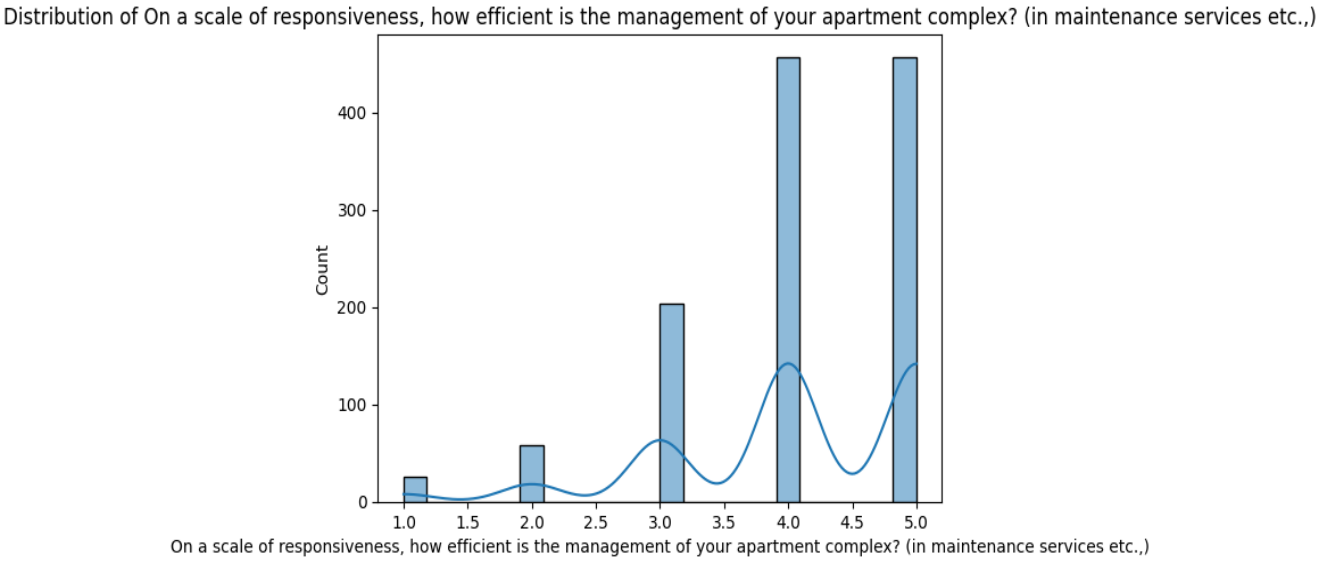


Figure 4 illustrates residents' perceptions of safety in their area, represented on a scale from 1 to 5. The graph highlights the frequency of each rating, with the line graph underscoring where the majority of the responses cluster, particularly noting where residents feel most and least secure.

Figure 5 shows the distribution of total monthly rent for units, revealing a significant concentration of rents around the $750 to $1250 range, with a noticeable peak near $1000.

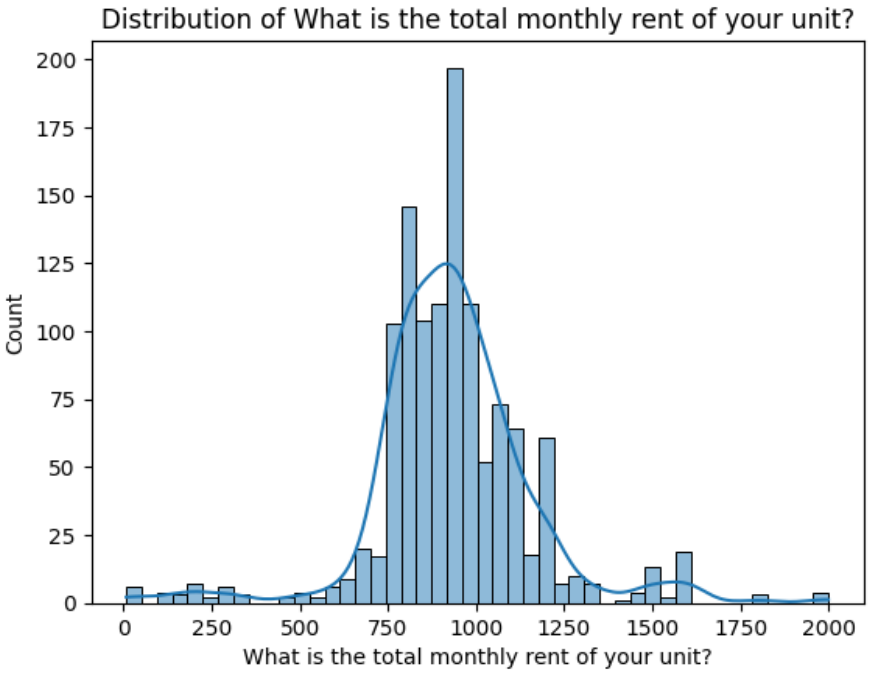


Figure 6 illustrates the distribution of the total amount paid initially by tenants, covering fees such as application fees, deposits, and others. Most initial payments are clustered at the lower end, indicating that most tenants pay less than $5,000 initially.

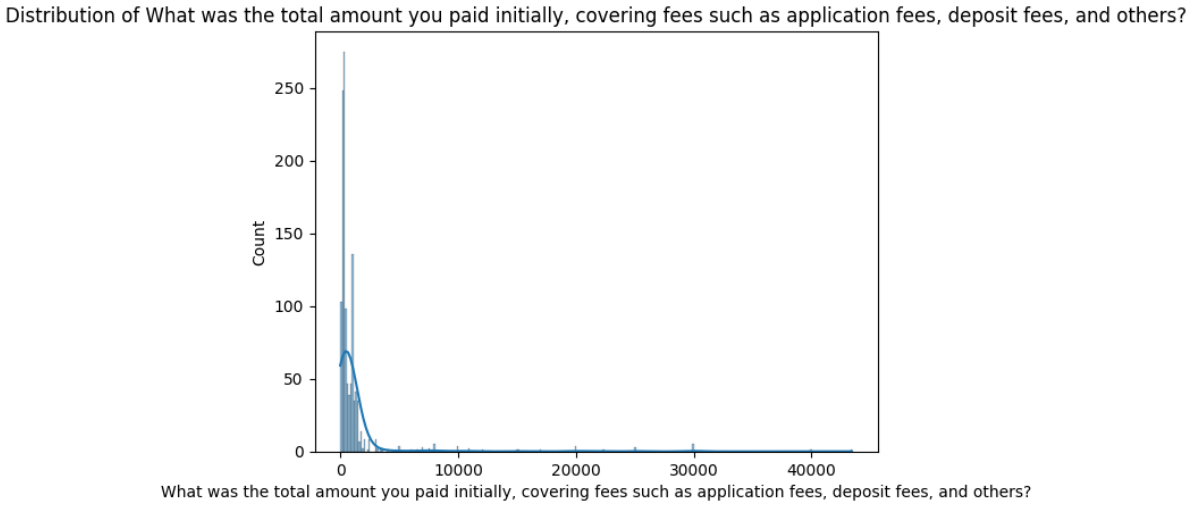


Figure 7 displays the distribution of monthly rental insurance payments for those who have it. The data indicates that the vast majority pay less than $500 per month for rental Insurance.

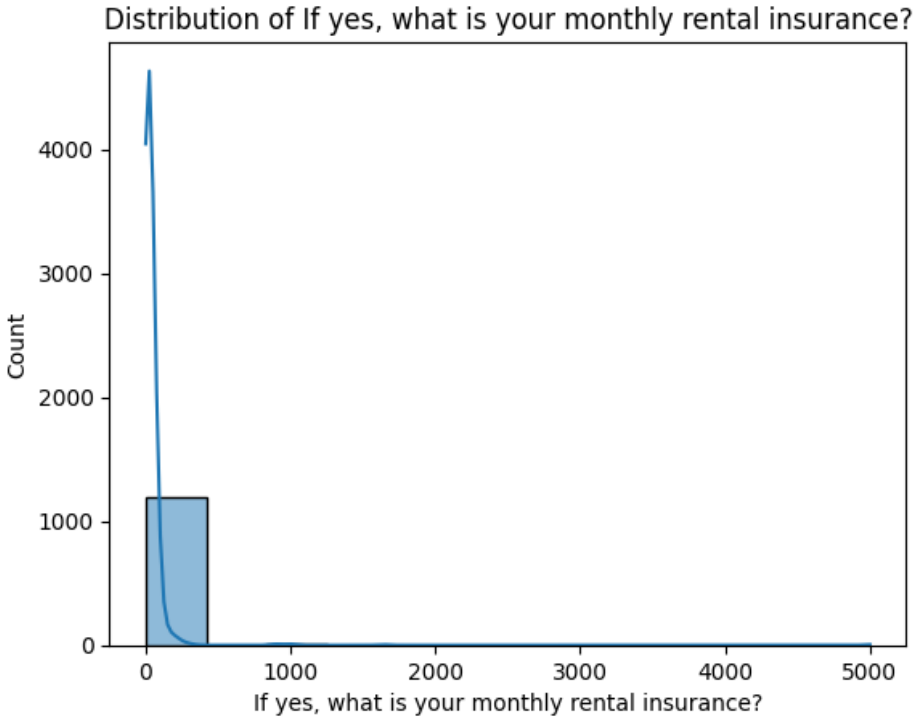


Figure 8 presents the distribution of monthly bills not included in the rent, such as electricity and parking. This histogram shows a steep peak at the lower end, with most respondents reporting extra costs of around $200 or less.

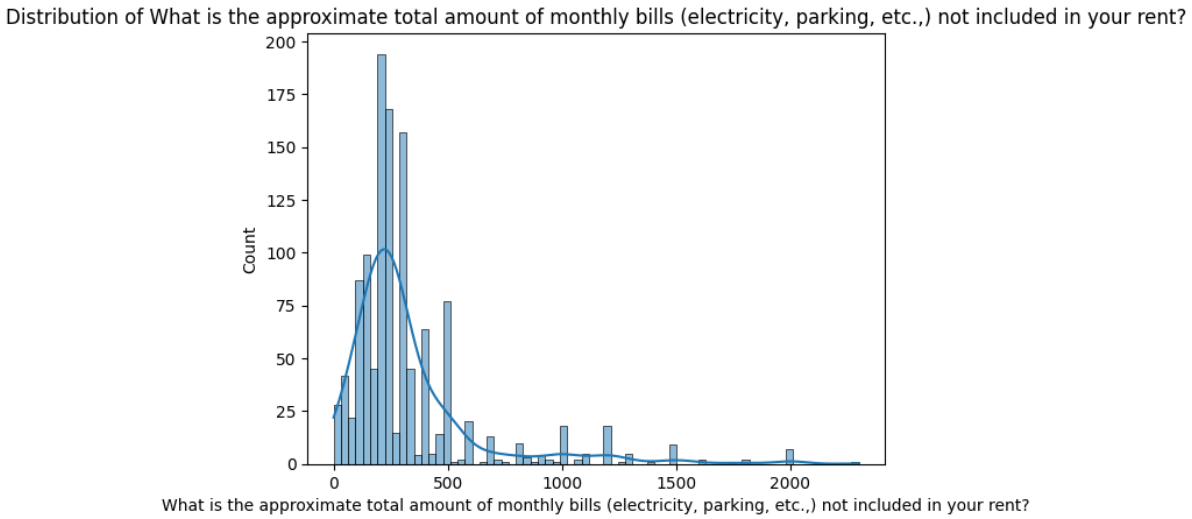


Figure 9 shows how many times residents have relocated within Wichita. The data indicates that a significant number of respondents have not relocated or have moved only once since arriving.

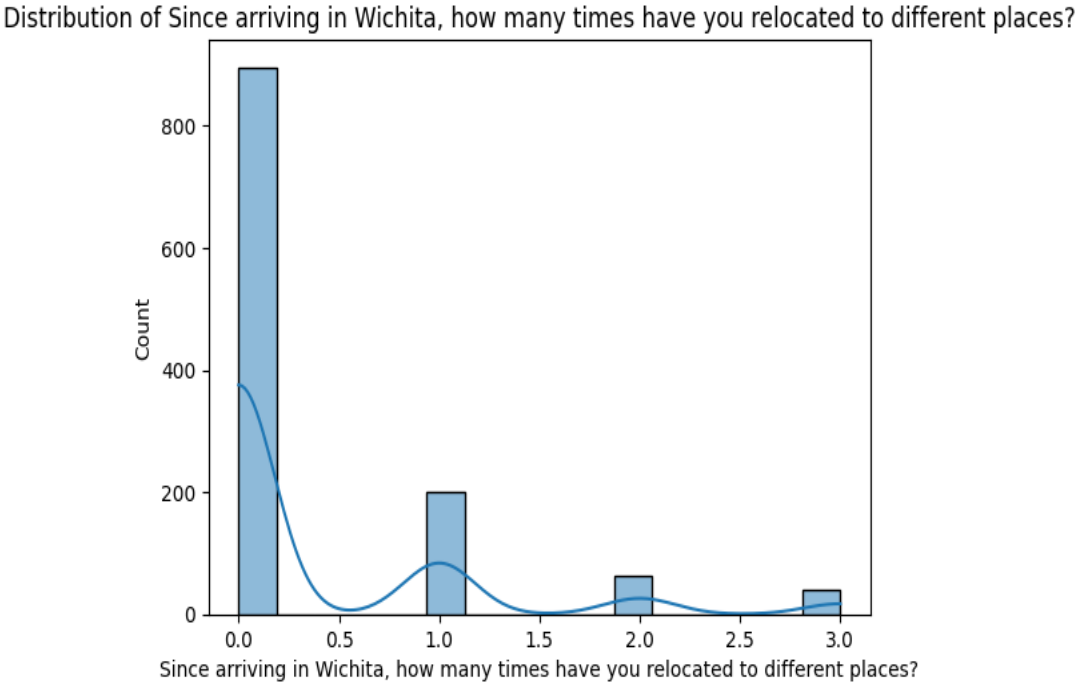


Figure 10 depicts levels of satisfaction among respondents regarding their previous apartment. A prominent peak suggests a high level of satisfaction, rating close to 5 on a scale from 1 to 5.

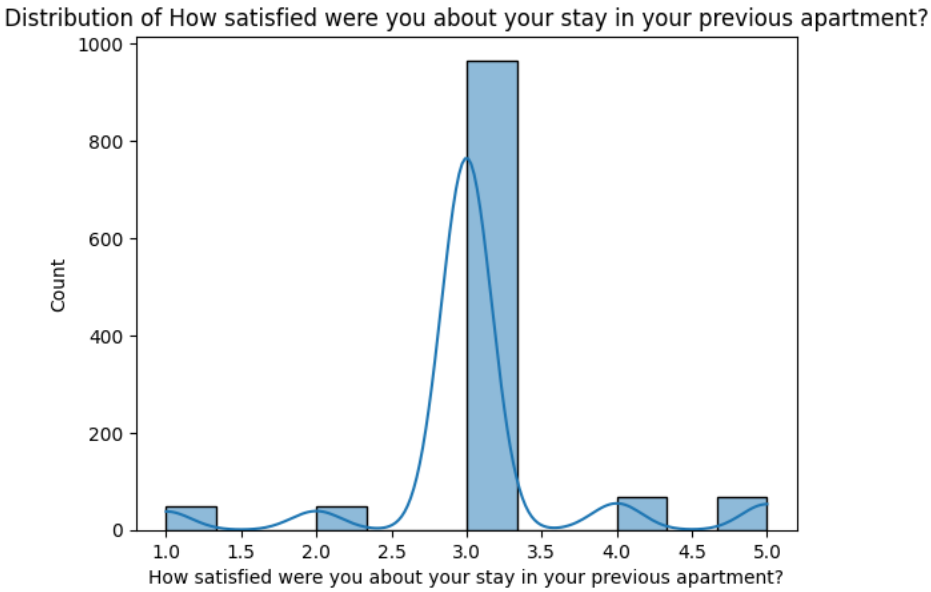


Figure 11 illustrates the duration it takes for respondents to walk to or from Wichita State University (WSU). The histogram shows a significant number of respondents reporting a walking time of less than 10 minutes, indicating proximity to the university.

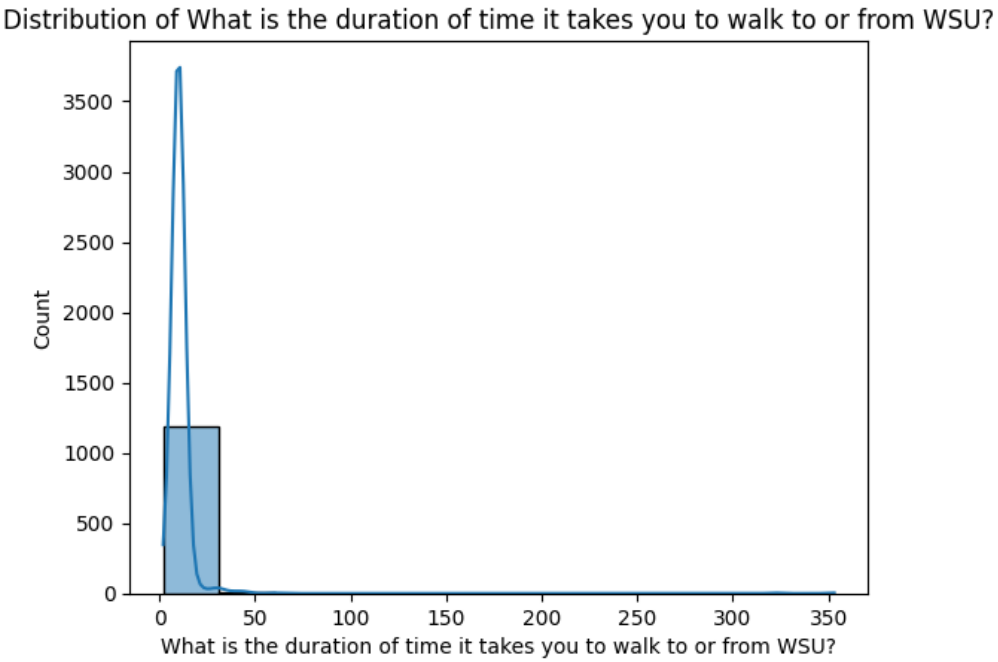


Figure 12 shows the time it takes for respondents to walk from their residence to the nearest transit stop. The majority indicate a walking time of less than 5 minutes, emphasizing close proximity to transit services.

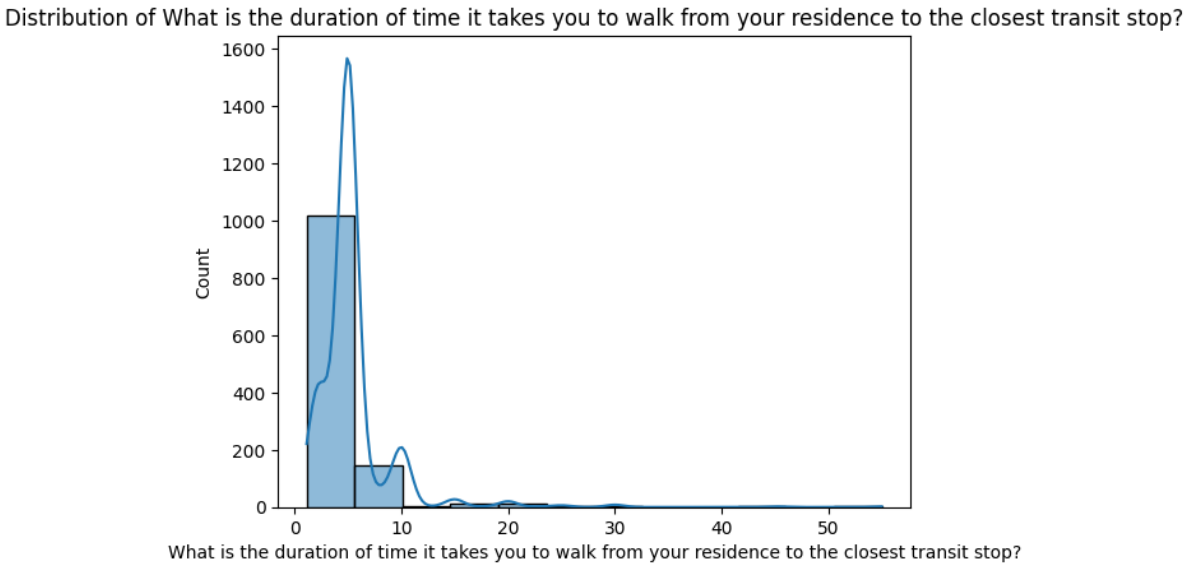


Figure 13 details the time it takes for transit to reach from the nearest stop to the university. Most responses cluster around 10 minutes, suggesting efficient transit connections for the majority of respondents.

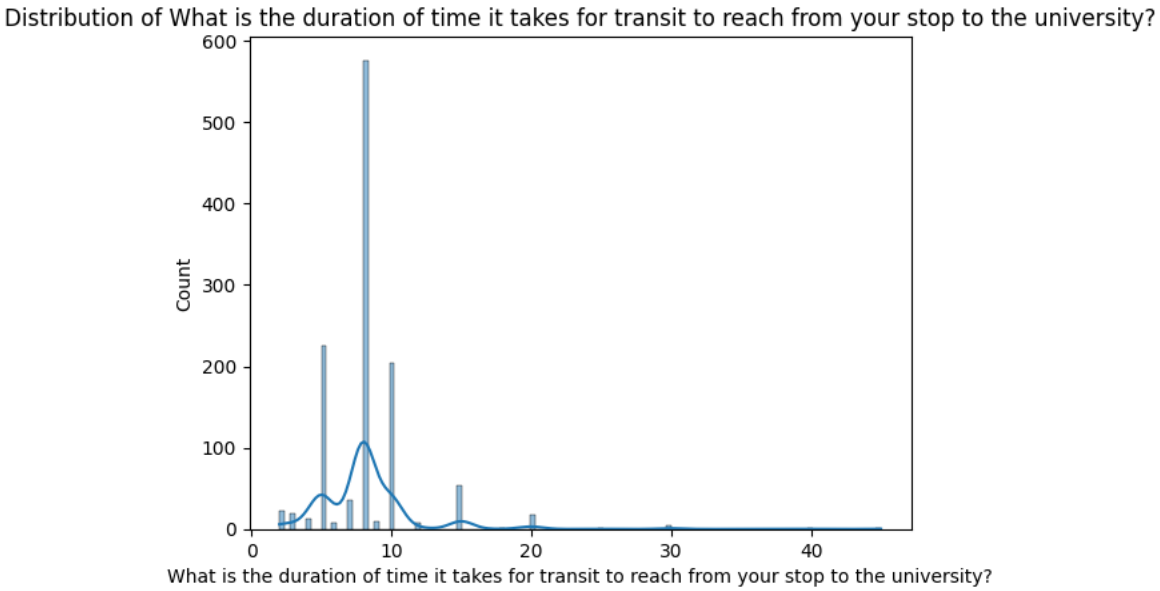


Figure 14 displays the duration it takes to drive from respondents' residences to the university. A significant peak at around 5 minutes suggests that many live close to the university.

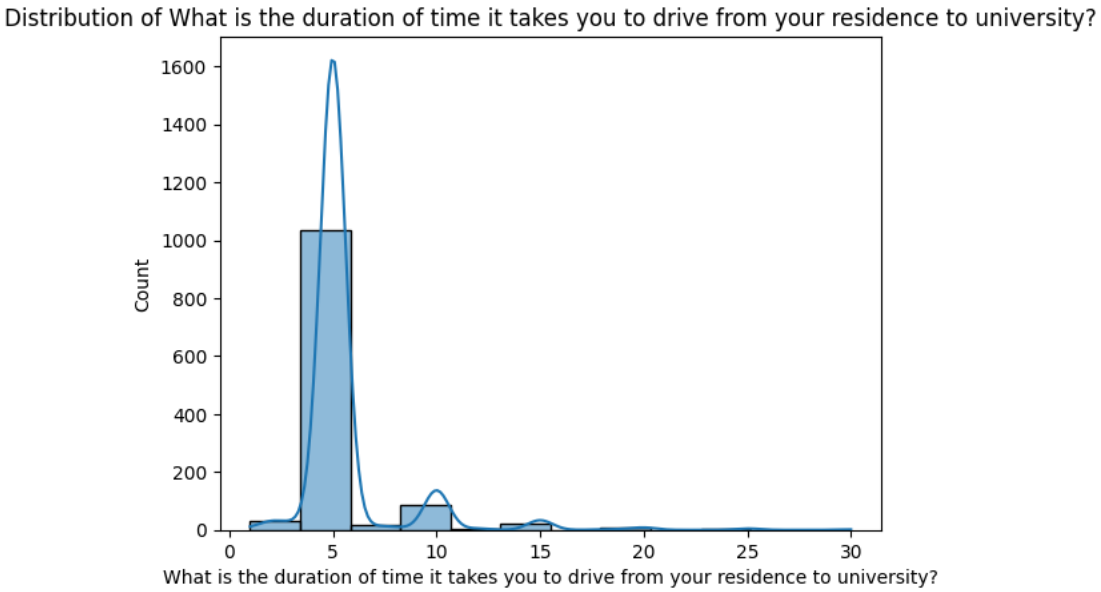


Figure 15 illustrates the cost of travel from residences to the university. It shows that the vast majority of respondents spend very little, with most costs clustering at the lower end of the scale.

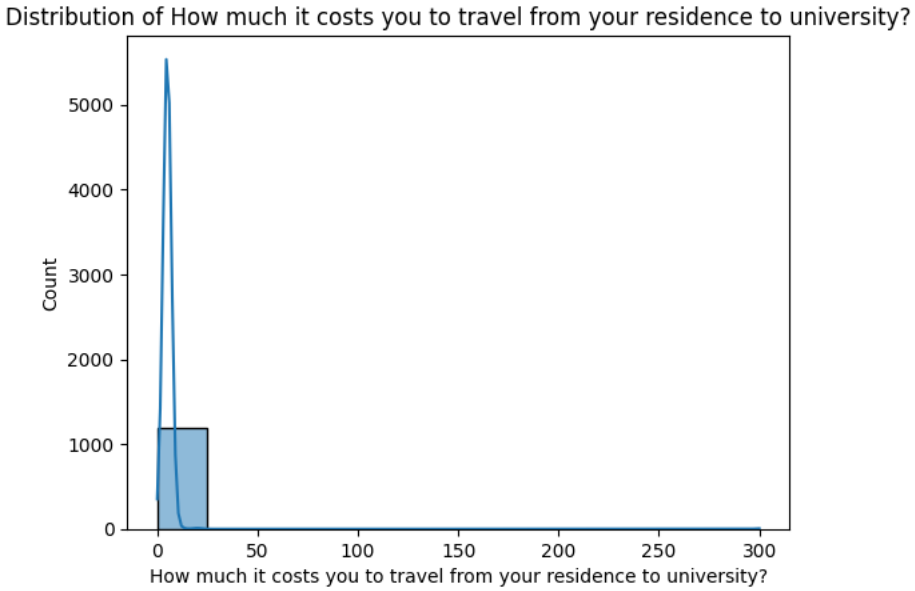


Figure 16 This histogram shows the time it takes respondents to reach their closest grocery store. The data clusters primarily between 5 and 20 minutes, indicating reasonable access to grocery shopping for most.

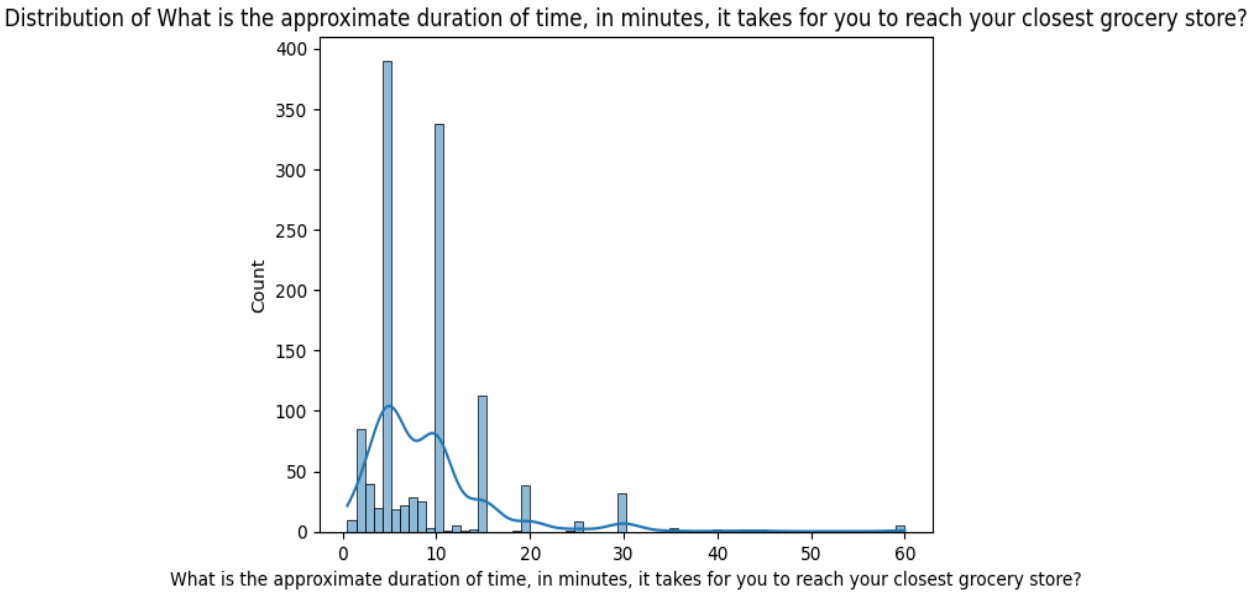


Figure 17 presents the time taken to reach a preferred nearby restaurant. Most respondents report travel times of less than 20 minutes, highlighting convenient access to dining options.

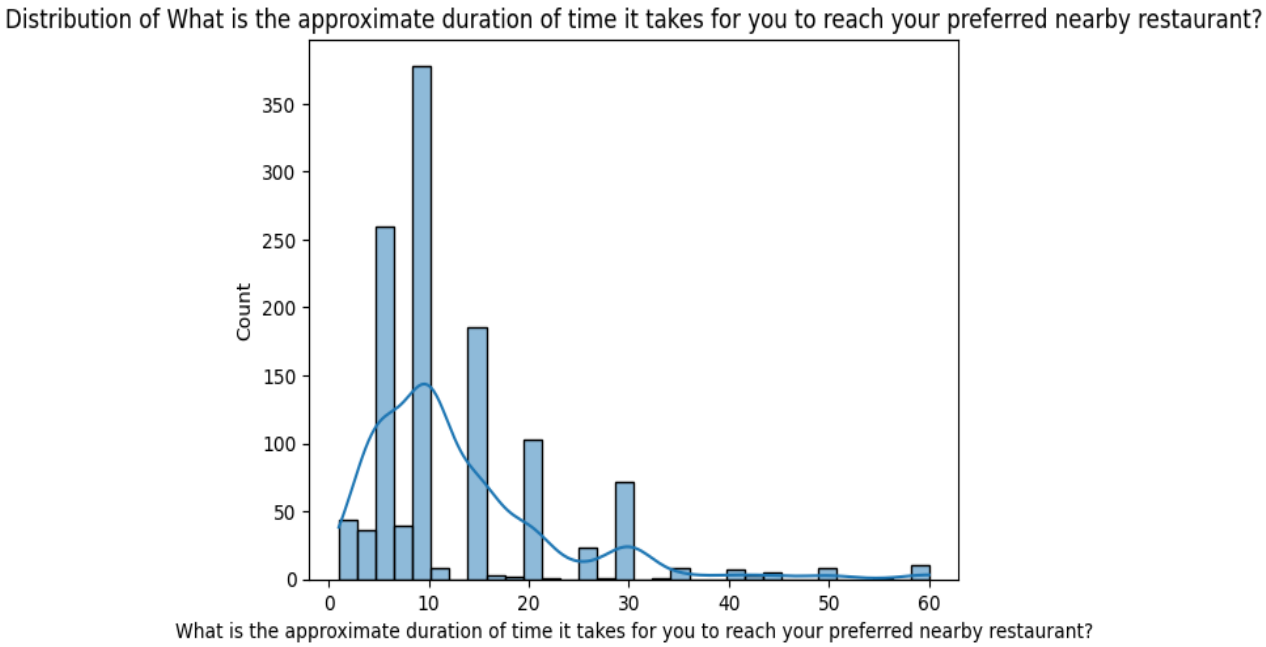


Figure 18 depicts respondents' satisfaction with their current stay. A significant peak at the high end of the scale indicates a high level of satisfaction among the majority of respondents.

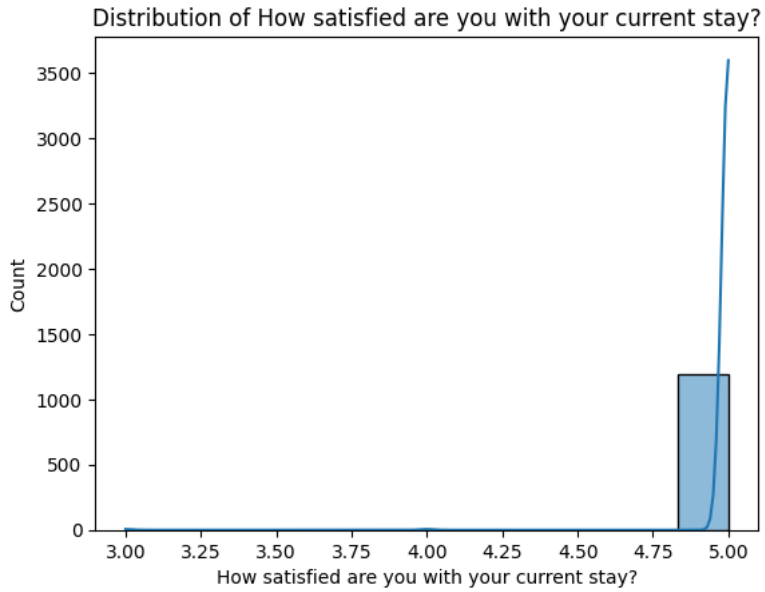


Figure 19 shows the distribution of monthly rent among respondents who answered "Yes" to a prior question. The peak around $1000 to $1250 suggests this is a common rent range.

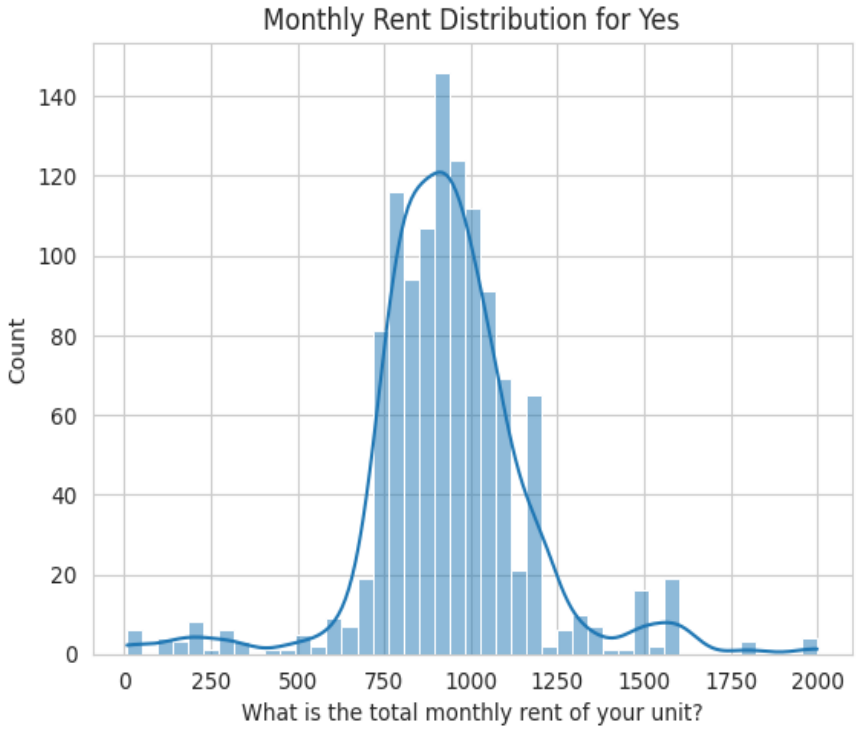


Figure 20 shows the monthly rent distribution for respondents who answered "No" to a prior question, with the majority paying between $900 and $1100.

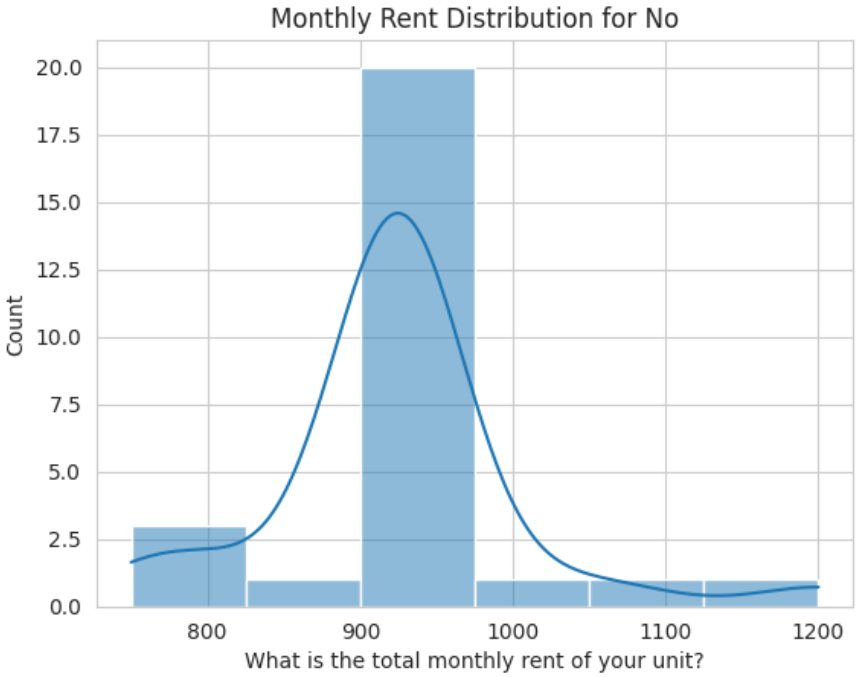


Figure 21 illustrates the proportion of respondents identifying as international students, indicating that the majority are not international students.

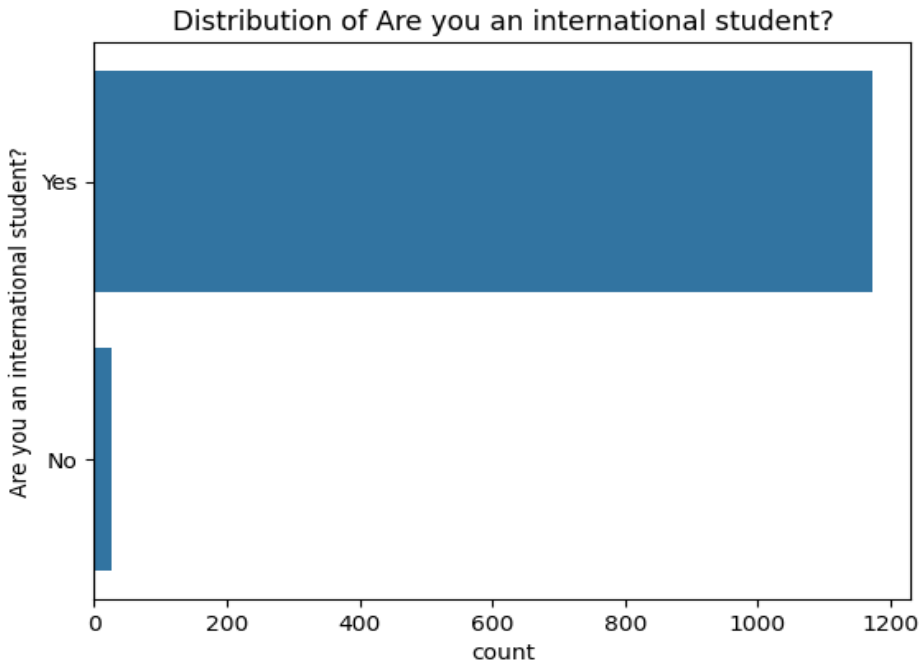


Figure 22 presents the living arrangements of respondents, showing that most live in apartments, followed by rented houses.

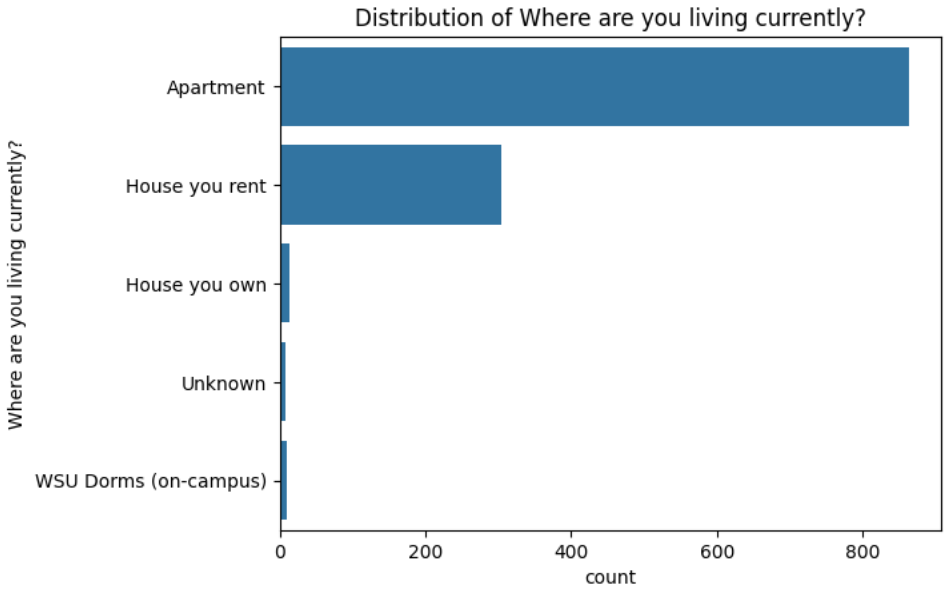


Figure 23 displays the distribution of respondents by apartment or street name, with "21W" and "The Landing" being among the most common.

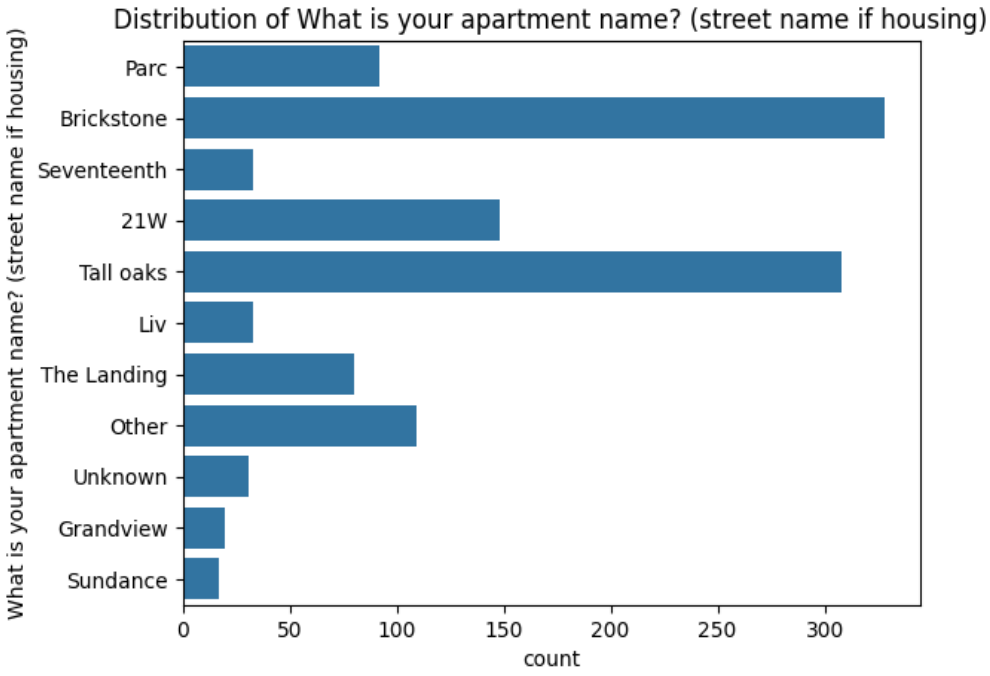


Figure 24 depicts how many people are staying in each housing unit, with one-person units being the most prevalent.

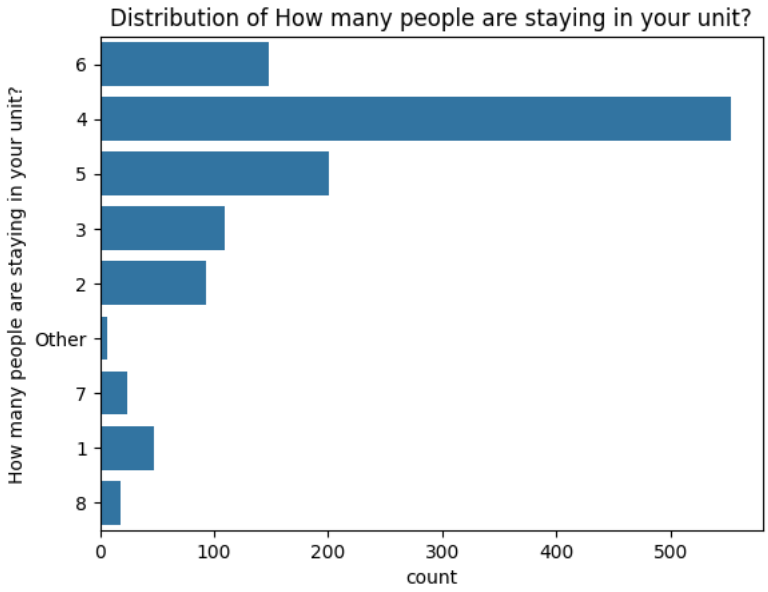


Figure 25 shows the distribution of unit sizes, with one-bedroom one-bathroom units being the most common among respondents.

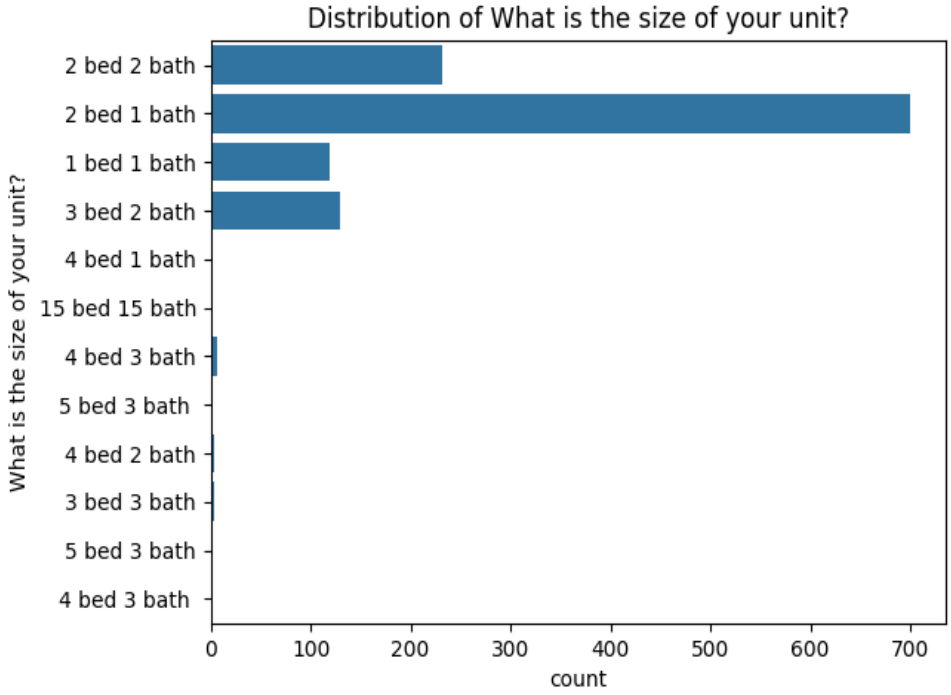


Figure 26 illustrates the proportion of respondents who pay for rental insurance monthly, with a larger portion indicating they do not.

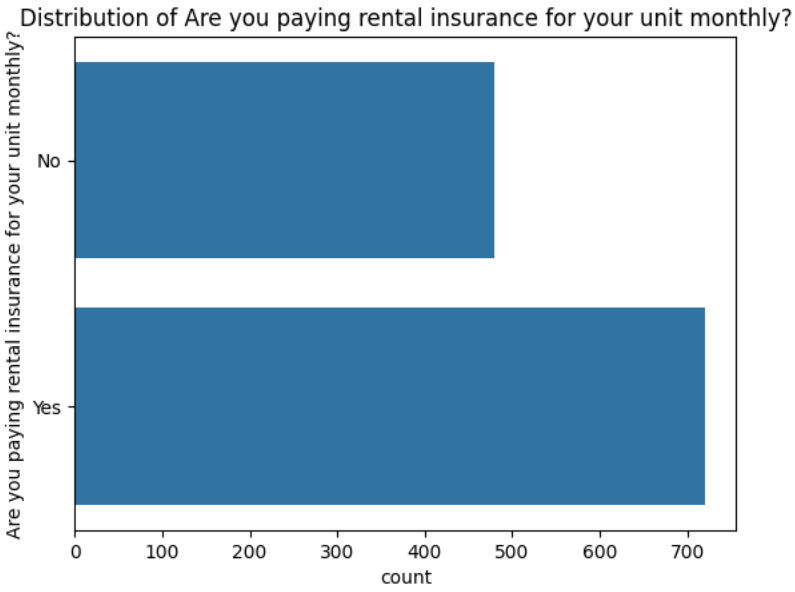


Figure 27 presents data on whether respondents' apartments require a guarantor, with varied responses, though many indicate that a guarantor is not required.

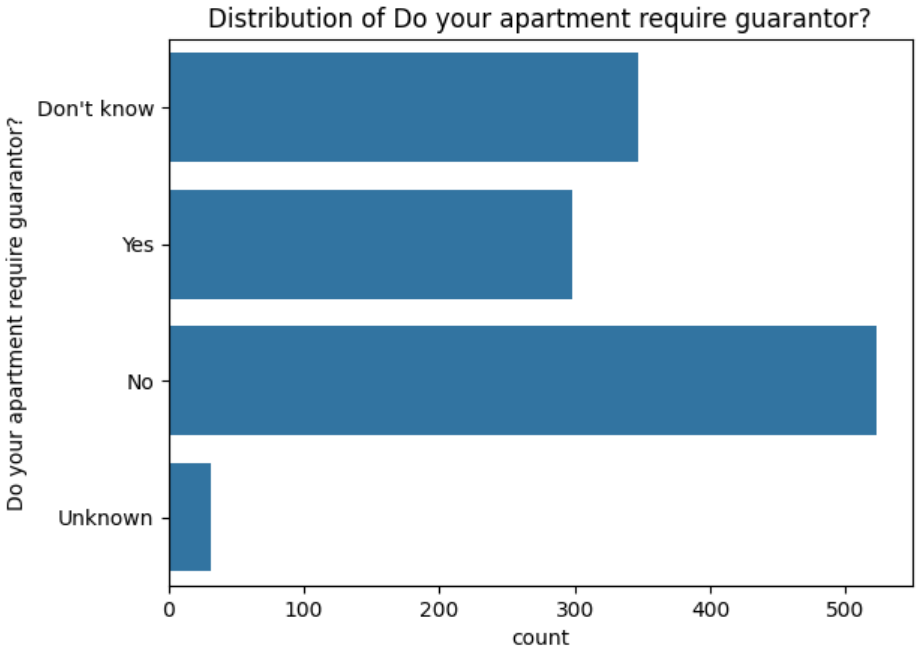


Figure 28 shows whether apartments are pet-friendly, indicating that a significant majority of apartments do allow pets.

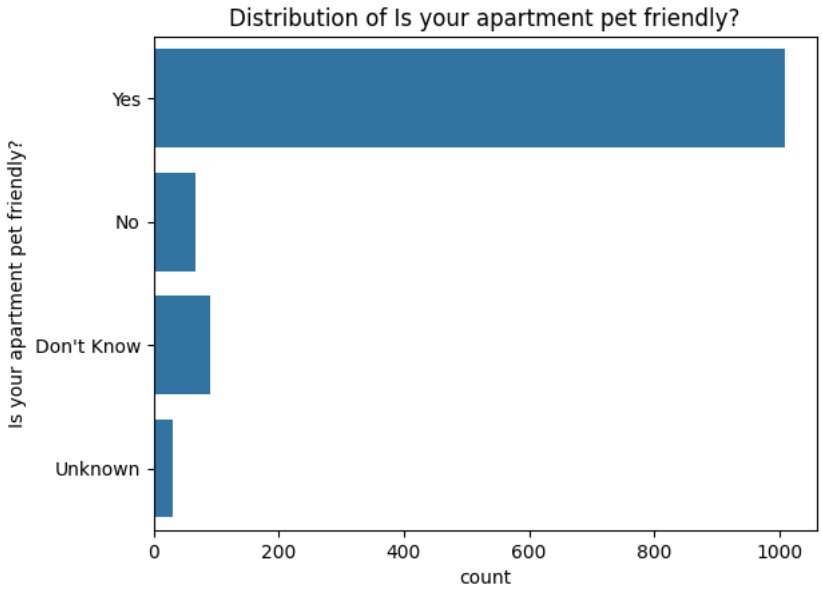


Figure 29 shows the distribution of respondents having a free parking lot allocated to their unit. The overwhelming majority reported not having a free parking lot.

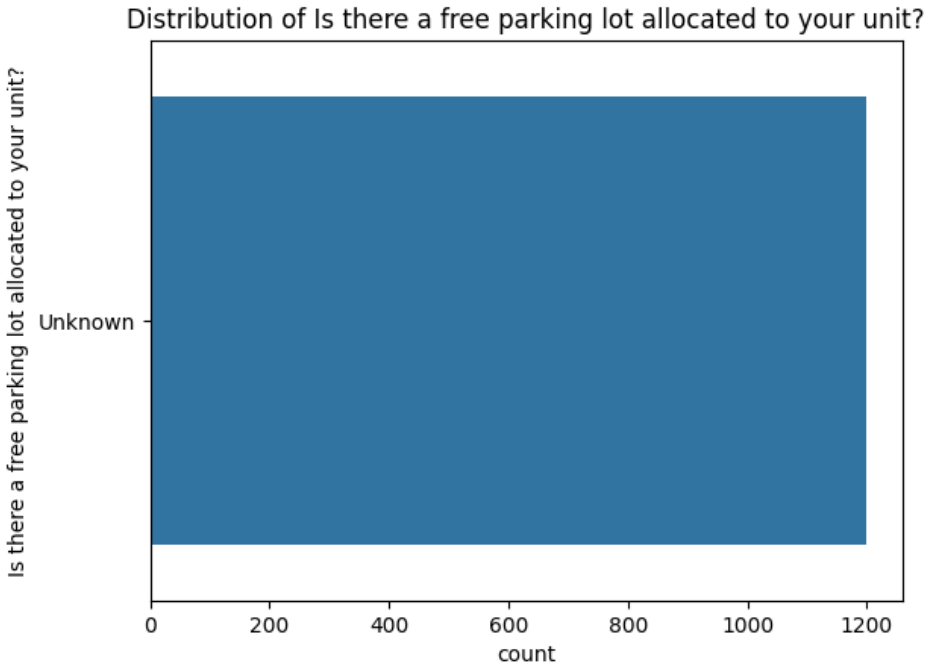


Figure 30 illustrates the preferred modes of transportation when leaving their residences, with driving being the most popular choice, followed by transit.

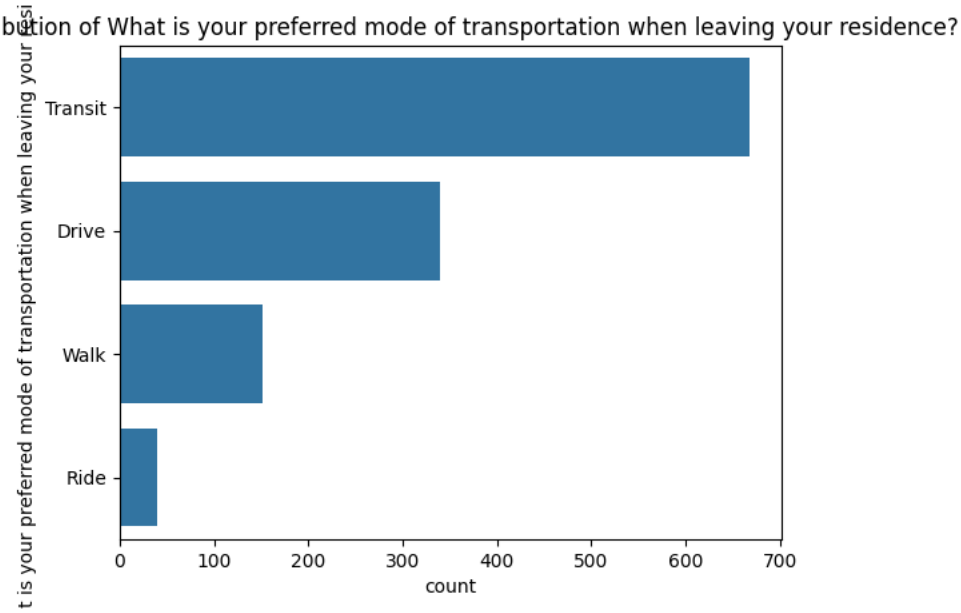


Figure 31 highlights the types of laundry amenities available in respondents' accommodations. The majority have laundry facilities in their unit, with fewer having access within the apartment complex.

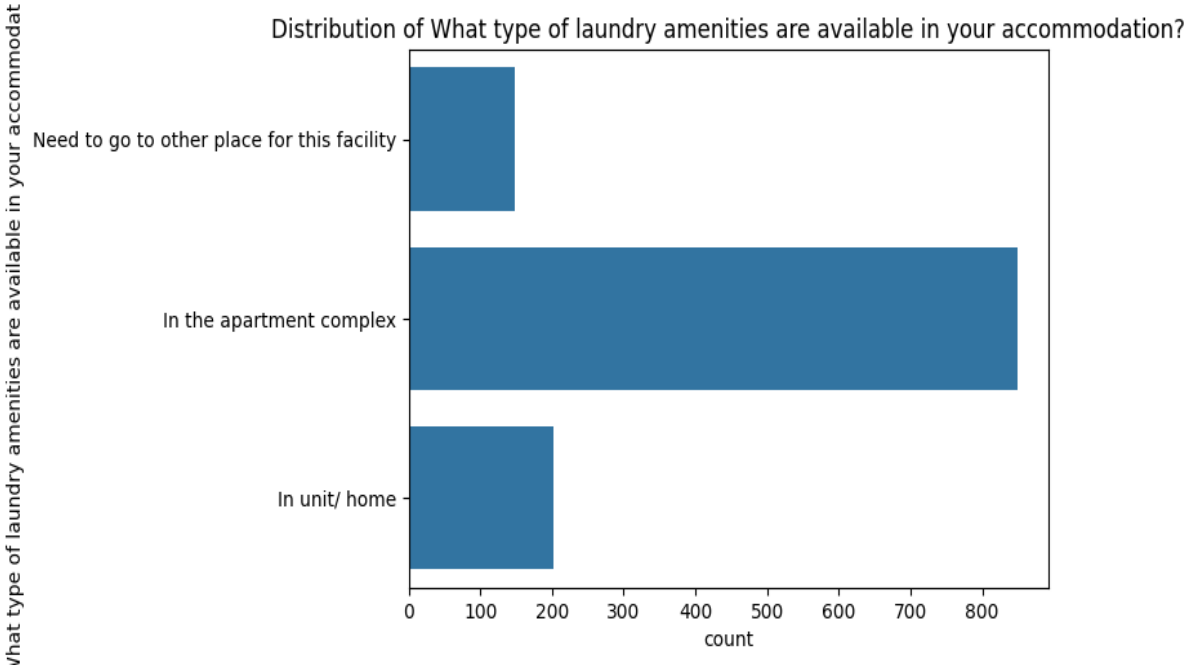


Figure 32 displays the current residence distribution of respondents. A significant number reside in Shocker Hall, followed by The Flats and other locations.

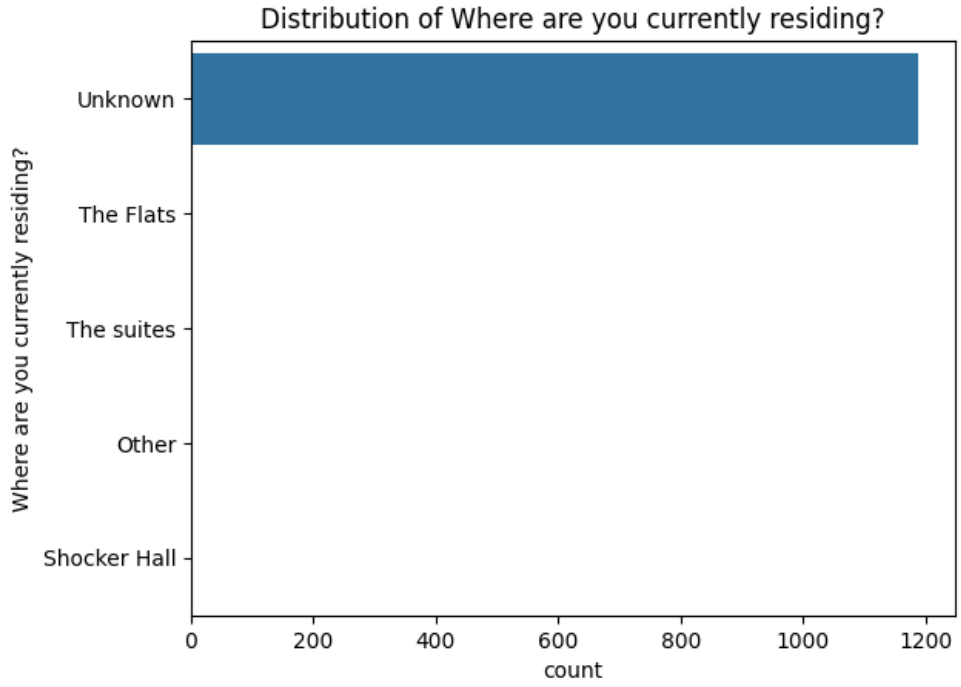


Figure 33 depicts the distribution of satisfaction levels with their current accommodation. A notable number of respondents rated their satisfaction at the highest level (5).

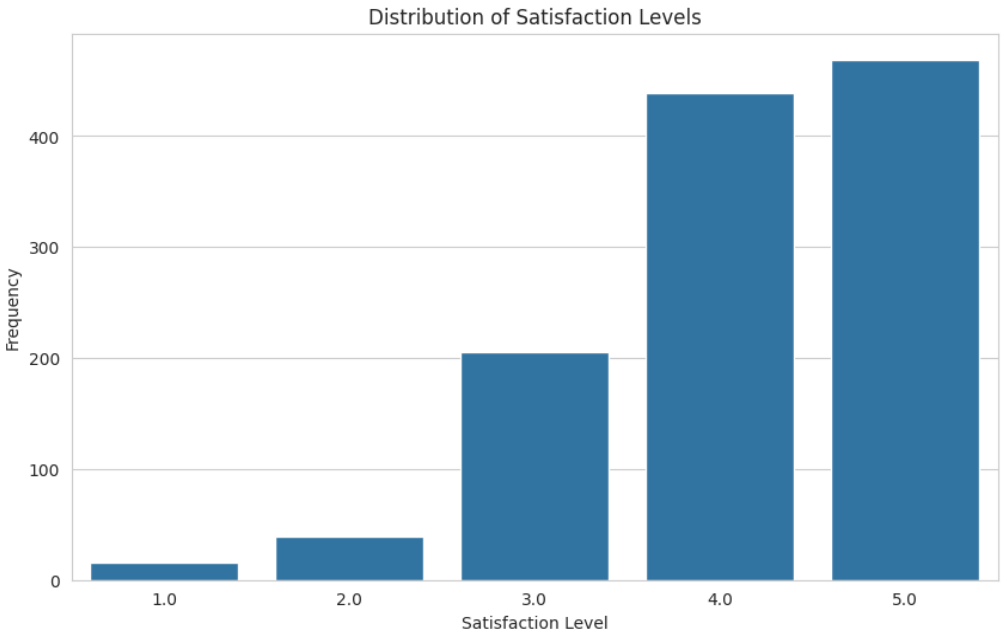


Figure 34 presents a boxplot of satisfaction levels, showing the spread and outliers. The median satisfaction level is around 4, indicating generally high satisfaction.

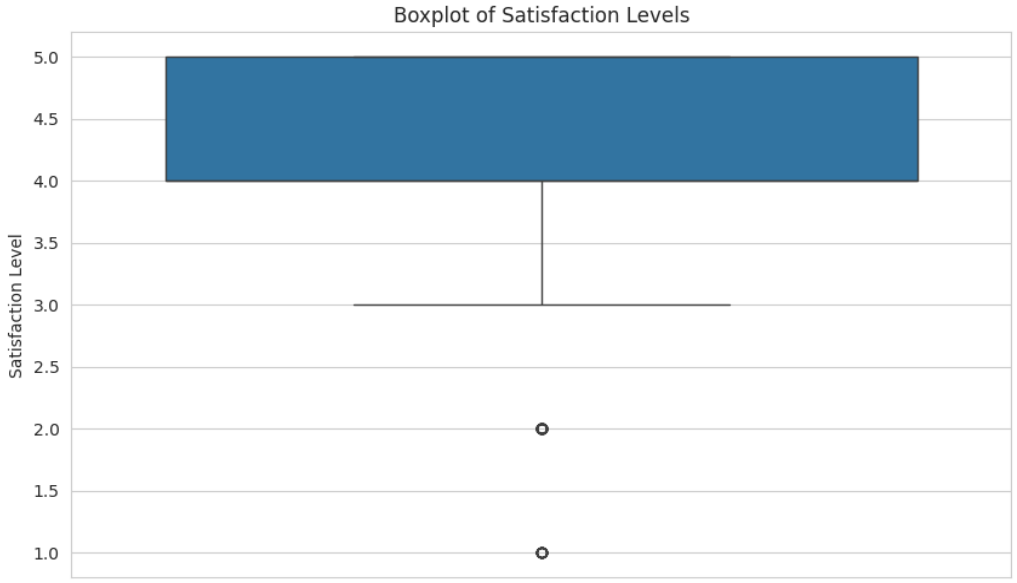


Figure 35 shows a boxplot of monthly rent, highlighting the distribution, median, and outliers. The median rent lies around $1000, with most rents clustered between $750 and $1250.

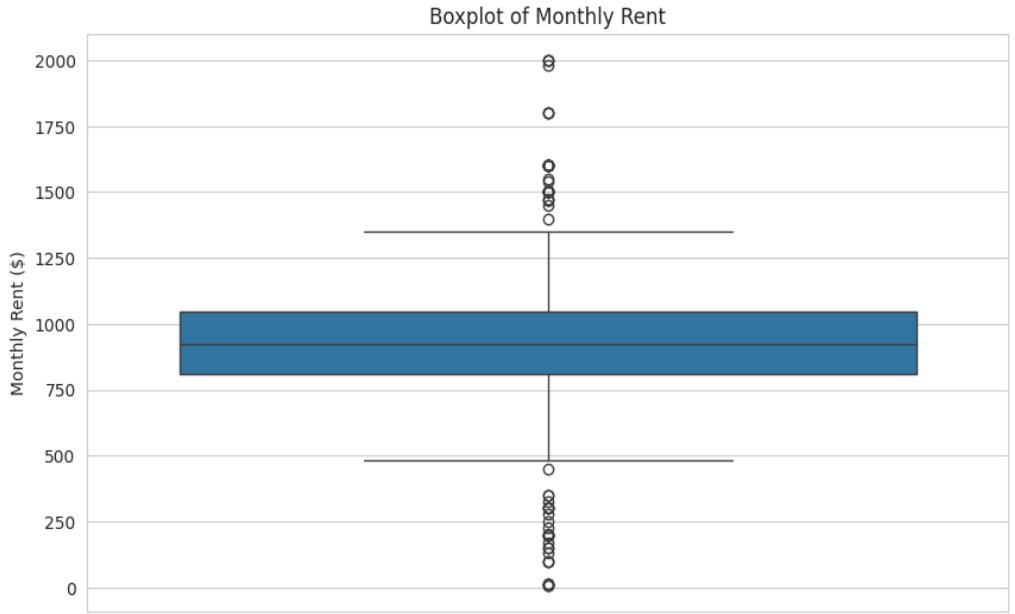


Figure 36 displays a scatter plot of rent versus satisfaction level, showing no clear correlation between the amount of rent paid and satisfaction.

Figure 37 presents histograms and a scatter plot analysis of satisfaction level against monthly rent. It further explores the distribution and relation between how much respondents pay and how satisfied they are with their accommodation.

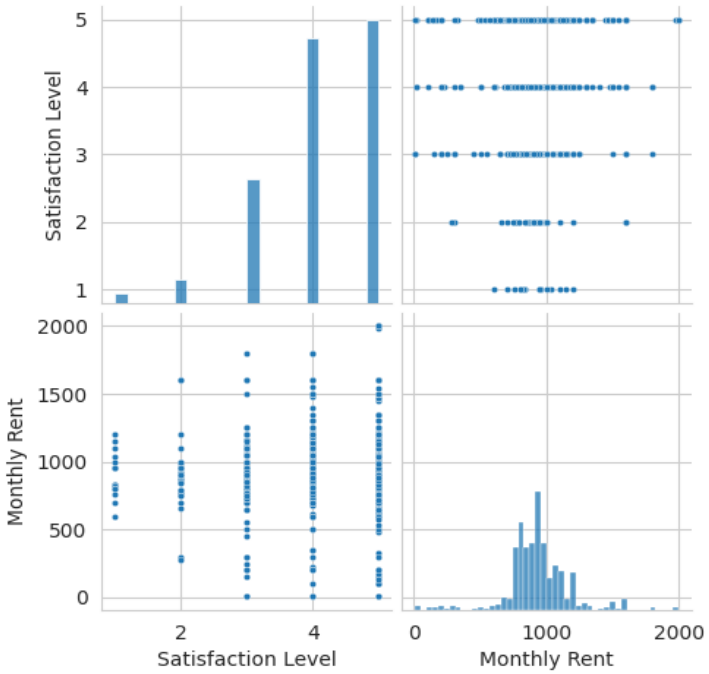


Figure 38 shows a correlation matrix heatmap for various survey variables, indicating how different aspects such as monthly rent, satisfaction levels, and amenities correlate with each other. Significant correlations are clearly visible between monthly rent and unit size, as well as between satisfaction levels and the availability of amenities.

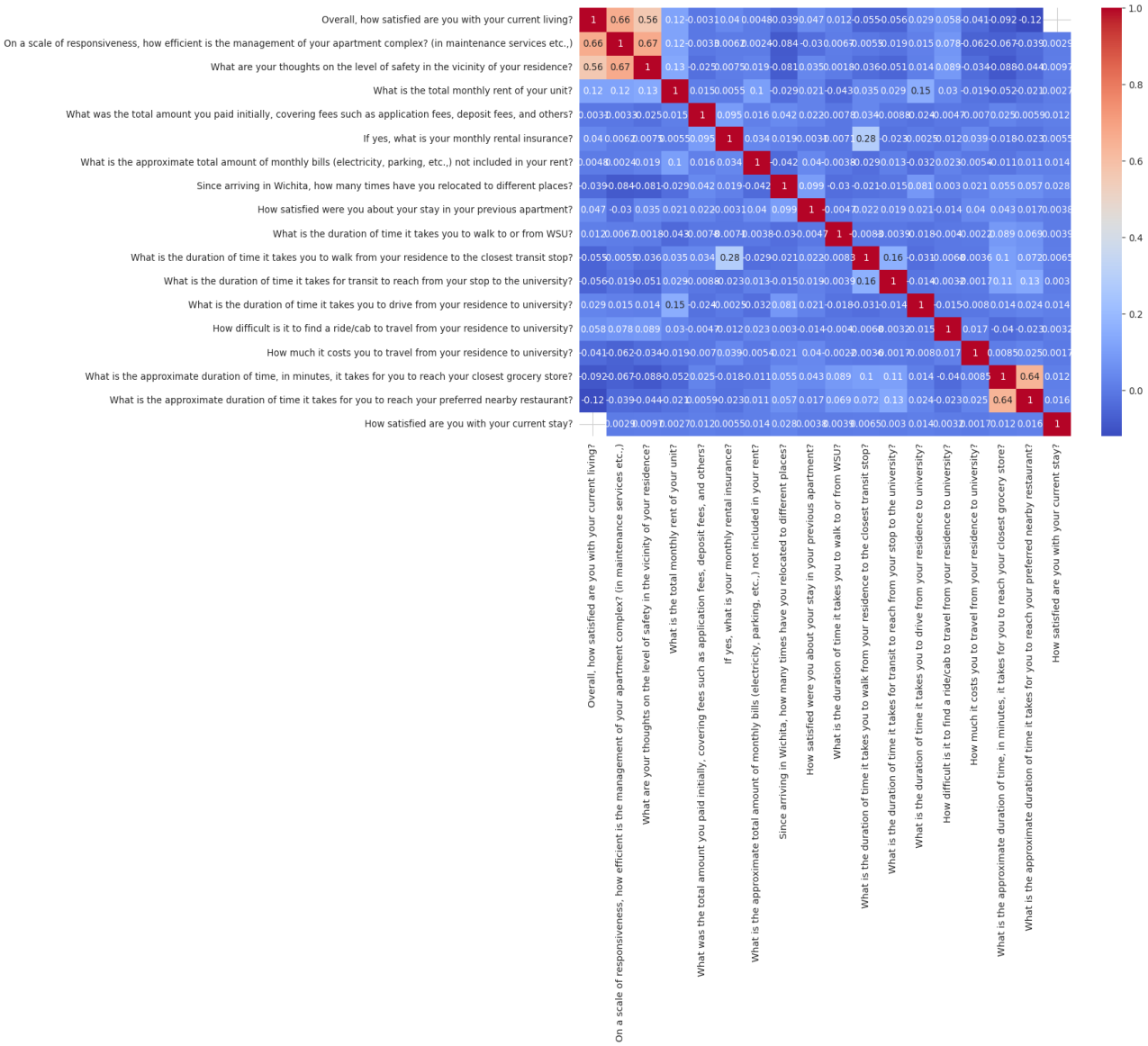


Figure 39 displays a simple heatmap illustrating the correlation between satisfaction level and monthly rent, emphasizing a moderate positive correlation, indicating that higher rent is somewhat associated with higher satisfaction.

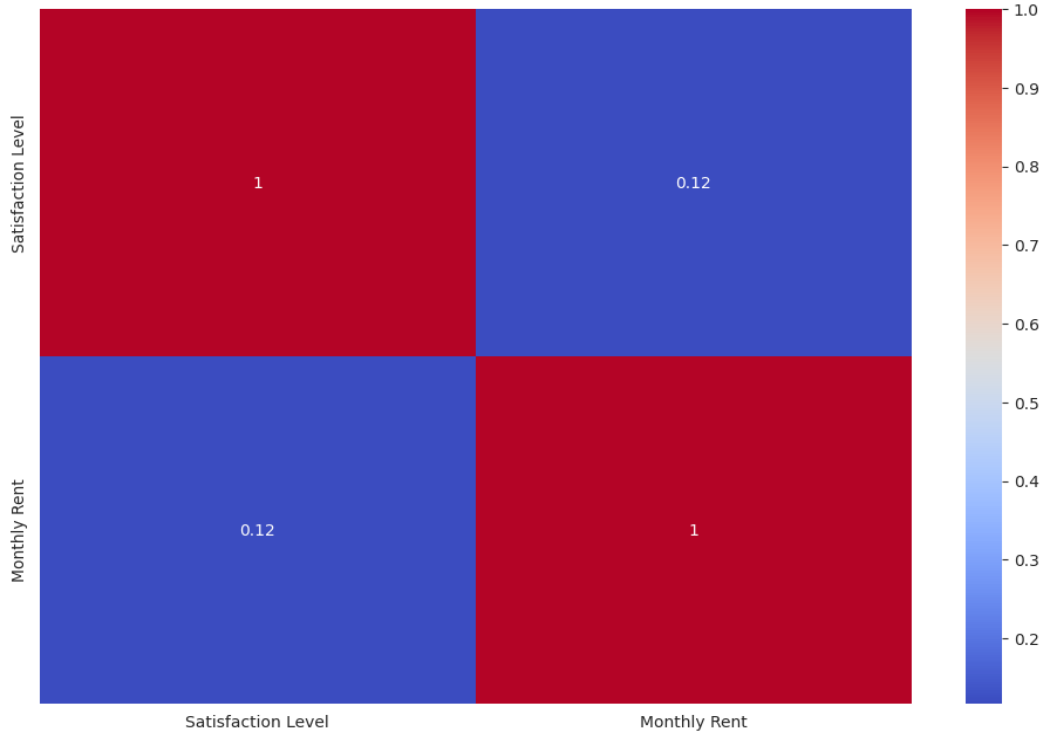
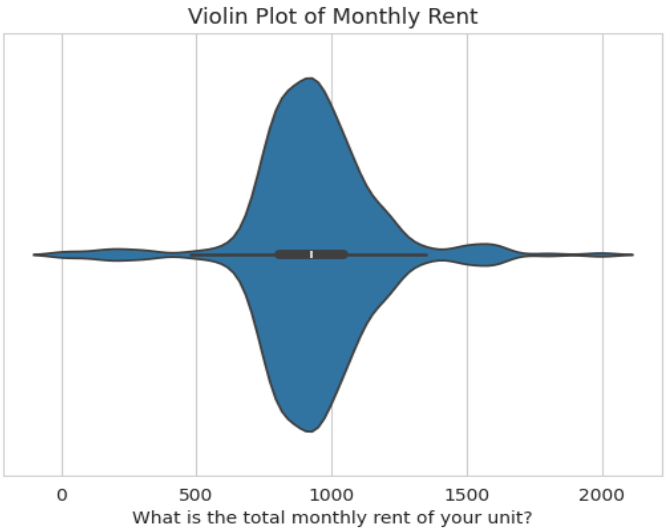


Figure 40 presents a violin plot of monthly rent, which visually combines a box plot and a density plot. The distribution shows a wide range of rents with a dense concentration around $1000, indicating it as a common rental price point.



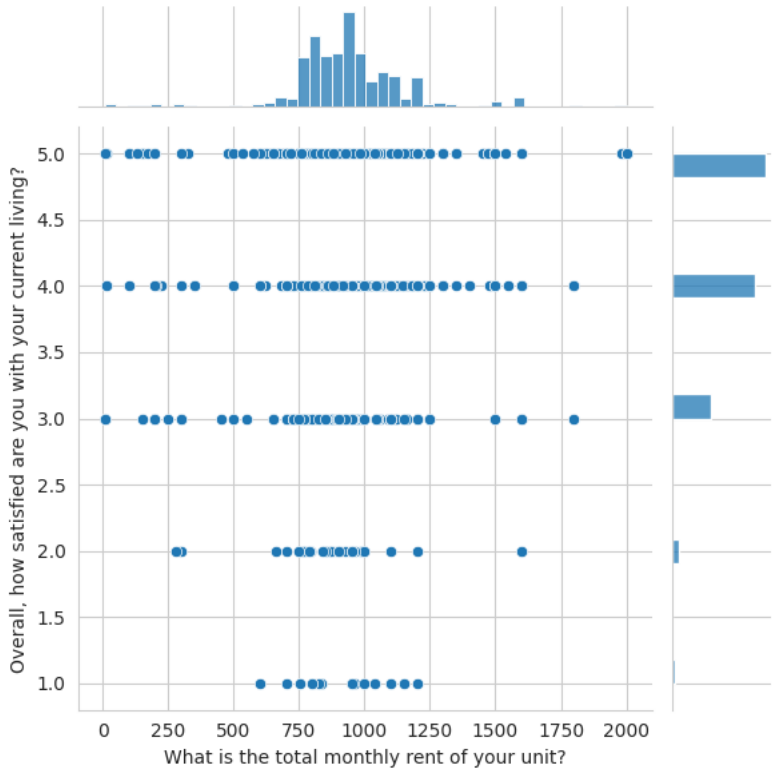
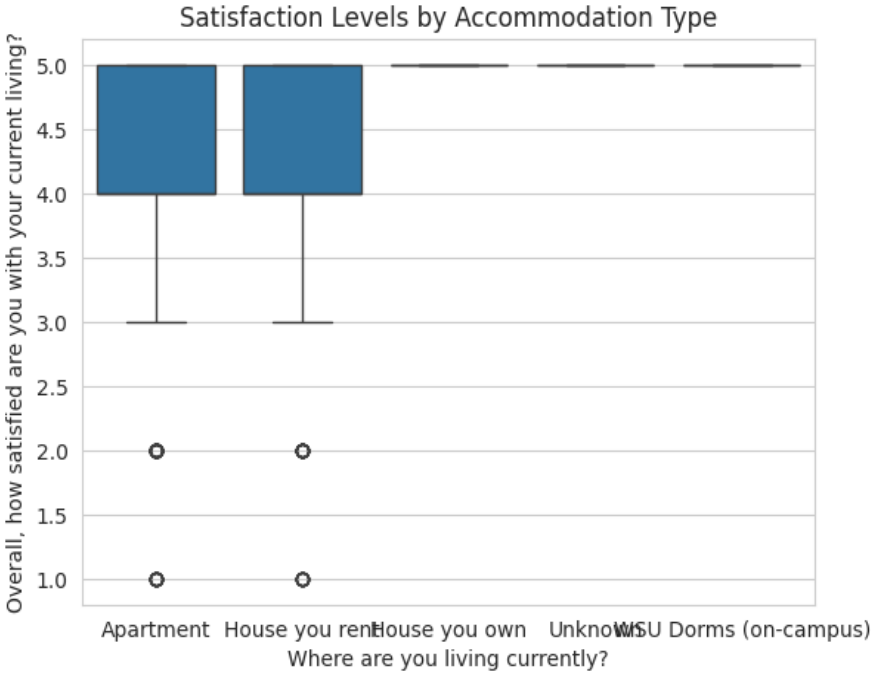


Figure 41 illustrates satisfaction levels by accommodation type using a box plot. It indicates that respondents in apartments tend to have varying satisfaction levels, while those in houses and WSU dorms have more consistent satisfaction scores.



As part of our **Exploratory Data Analysis (EDA)**, we engaged in sophisticated visualizations, detected, and addressed outliers, managed missing values, and conducted segmentation analysis. Additionally, we analyzed textual data and undertook comparative analysis, incorporating these techniques into our comprehensive EDA process.

KEY INSIGHTS

1. **Data Integrity**
    - Our meticulous approach to **data cleaning** ensured the analysis rested on accurate and reliable data. This process was crucial for the integrity of our findings, enabling us to confidently navigate through the complexities of student housing preferences.

2. **Understanding Housing Satisfaction:**
    - The **exploratory data analysis** revealed a **weak positive correlation between rent levels and satisfaction**, challenging common perceptions and illuminating the intricate dynamics at play. This insight underscores the necessity of considering a broad range of factors when evaluating housing options, beyond just the cost.
3. **The Role of Data Science**
    - This project exemplified the transformative potential of data science in addressing real-world challenges. By systematically analyzing diverse datasets, we extracted valuable insights that can significantly improve the decision-making process for international students seeking housing.

REGRESSION ANALYSIS REPORT

Original Analysis

Target Variable: Overall satisfaction with current living conditions

Predictor Variable: Total monthly rent of the unit

Regression Outcomes

- **Model Fit (R-squared):** 0.013

The R-squared value of 0.013 indicates a very weak explanatory power of the model with respect to the variance in the satisfaction levels.

Coefficients

- **Constant:** 3.719862
- **Monthly Rent:** 0.000450

The coefficient for the monthly rent suggests a minimal impact on the overall satisfaction level, implying that for every unit increase in rent, satisfaction increases by a factor of 0.000450.

P-values

- **Constant:** 1.969e-178
- **Monthly Rent:** 0.0000758

The extremely low p-values indicate that the coefficients are statistically significant, despite the model's low explanatory power.

Interpretation

The analysis under Scenario 1 reveals that while the effect of monthly rent on satisfaction is statistically significant, it has a negligible practical impact on the residents' overall satisfaction. This suggests that factors other than rent might be more influential in determining satisfaction levels and should be considered in further analyses.

Analysis Report: Scenario 1

Overview

Target Variable: Overall satisfaction with current living conditions

Predictor Variable: Total monthly rent of the unit

Model Results

R-squared: 0.013

This indicates that only about 1.3% of the variance in satisfaction levels can be explained by changes in monthly rent, highlighting a weak correlation.

Coefficients:

Constant (Intercept): 3.7199

This suggests that the baseline level of satisfaction, independent of rent, is moderately high.

Rent Coefficient: 0.0004

This shows a very slight positive relationship between rent and satisfaction, suggesting that higher rents are associated with marginally higher satisfaction.

P-value for Rent: 7.58e-05

The rent coefficient is statistically significant, as indicated by a p-value much less than the conventional alpha level of 0.05.

Interpretation

The analysis from Scenario 1 reveals a minor positive correlation between monthly rent and satisfaction levels, which is contrary to the initial hypothesis that lower rent might correlate with higher satisfaction. Despite the statistical significance of the rent coefficient, the extremely low R-squared value suggests that monthly rent is not a strong predictor of satisfaction. These findings suggest that other factors likely play a more significant role in influencing residents' satisfaction levels.

Analysis Report: Scenario 2

Overview

Target Variable: Overall satisfaction with current living conditions

New Predictor Variable: Total monthly bills not included in rent (e.g., electricity, parking)

Model Results

R-squared: 0.013

Similar to Scenario 1, this value indicates that only about 1.3% of the variance in satisfaction levels is explained by the monthly bills, suggesting a weak correlation.

Coefficients

Constant (Intercept): 3.7199

This implies a baseline satisfaction level that is independent of the monthly bills.

Bills Coefficient: 0.0004

There is a very slight positive association between the total amount of monthly bills and satisfaction, suggesting that higher bills are marginally correlated with increased satisfaction.

P-value for Bills: 7.58e-05

This coefficient is statistically significant, indicating that while the effect size is small, the relationship between bills and satisfaction is statistically reliable.

Interpretation

The findings from Scenario 2 indicate a minor positive correlation between the total monthly bills and satisfaction levels. Despite the statistical significance of the bills coefficient, the low R-squared value shows that the monthly bills, similar to rent, are not strong predictors of satisfaction. This analysis suggests that further investigation into other potential influencing factors would be beneficial in understanding the determinants of satisfaction levels.

Scenario 2 Analysis Summary

Objective:

Evaluate the impact of monthly bills (excluding rent) on satisfaction levels.

Variables:

- Target Variable: Satisfaction Level ("Overall, how satisfied are you with your current living?")
- Predictor Variable: Approximate total amount of monthly bills (excluding rent)

Model Results:

- **R-squared**: Practically 0, indicating that the predictor variable (monthly bills) does not explain any variability in satisfaction levels.
- **Coefficients**:
    - **Constant (Intercept)**: 4.1432, representing a baseline satisfaction level.
    - **Bills Coefficient**: -7.25e-06, suggesting a very small negative impact of

monthly bills on satisfaction.

- **P-value for Bills**: 0.933, demonstrating that the association is not statistically significant.

Interpretation

The analysis reveals that the amount spent on monthly bills does not significantly affect satisfaction levels.

The negligible R-squared value underscores that the monthly bills do not account for variations in satisfaction.

The hypothesis that lower monthly expenses (either through rent or bills) lead to higher satisfaction is not supported by the data in this scenario.

Scenario 3 Analysis Summary

Objective:

- Explore the impact of apartment management efficiency on the number of residents per unit.

Variables:

- **New Target Variable**: Number of people staying in the unit.
- **New Predictor Variable**: Efficiency of the apartment complex's management, as rated on a scale of responsiveness and maintenance services.

Model Results:

- **R-squared**: Essentially 0, indicating no meaningful relationship between management efficiency and the number of people staying in the unit.
- **Coefficients**:
    - **Constant (Intercept)**: 4.1340, suggesting a base number of residents typically found in units.
    - **Management Efficiency Coefficient**: 0.0093, showing a very minor effect that is statistically insignificant.
- **P-value for Management Efficiency**: 0.818, confirming that the relationship is not statistically significant.

Interpretation:

- The analysis indicates that the perceived efficiency of apartment management does not significantly influence the number of residents per unit.
- The extremely low R-squared value further supports the lack of correlation between these two variables.

General Conclusion Across Three Scenarios:

- Neither rent levels, monthly bills, nor the efficiency of apartment management notably influence resident satisfaction or the number of residents per unit.
- The consistently low R-squared values in all scenarios suggest that factors beyond financial considerations and management efficiency are crucial in determining resident satisfaction.
- This analysis underscores the importance of considering a broader range of lifestyle and personal factors to fully understand housing satisfaction dynamics.

CONCLUSION

The project highlights the indispensable role of data science in enhancing student living experiences. Through dedicated analysis and innovative thinking, we have taken significant steps towards improving the housing selection process for international students. Our project not only addresses an immediate practical need but also contributes to the broader discourse on leveraging data science for social good. As we look to the future, the possibilities for expanding this work are vast and promising, with the potential to make a lasting impact on the lives of international students worldwide.