Exploring the Impact of Property Features, Crime Rates, Demographic Changes, and Educational Proximity on Housing Prices Northern Territory, Australia.

Name: Shruthi Shashidhara Shastry,

Student ID: 33684669,

Teaching Associate: Farah Kabir & Ashwini Priya Narasimhan , Applied 11.

Contents

INTRODUCTION

Problem Description: This project undertakes a detailed analysis of the dynamics influencing housing prices across, Northern Territory, Australia. It focuses on examining how various factors such as property attributes, crime statistics, demographic trends, and proximity to educational facilities shape real estate values. This investigation aims to provide comprehensive insights that could facilitate effective decision-making in urban development and property management**.**

Questions

1. How do property characteristics (such as the number of bedrooms, bathrooms, and parking availability) and local crime rates collectively influence housing prices in the Northern Territory's cities and suburbs?
2. How does the changing population within the Northern Territory's government regions affect the property values in those areas?
3. How does the proximity to educational institutions affect housing prices in the Northern Territory?

Motivation: Fueled by a keen interest in property value determinants, this study investigates the Northern Territory's property market, examining how crime, demographics, property features, and education proximity influence real estate values, aiming to guide development and urban strategy.

DATA WRANGLING AND DATA CHECKING

This is critical phase of the project involves cleaning , checking and transforming our datasets into formats that optimize them for analysis. We perform data wrangling and data checking specifically tailored to each question, which enhances the utility and relevance of the data. Since the Northern Territory dataset applies to all questions, let's begin by wrangling it first.

NT Australian Dataset (Australian Housing Prices (kaggle.com))- Contains data of approximately 1,000 property listings in Northern Territory, Australia. It includes data types such as categorical (e.g., Property Type), numerical (e.g., Price), textual (e.g., Address), and spatial (e.g., Latitude, Longitude) attributes, (latitude and longitude will be derived through geocoding).

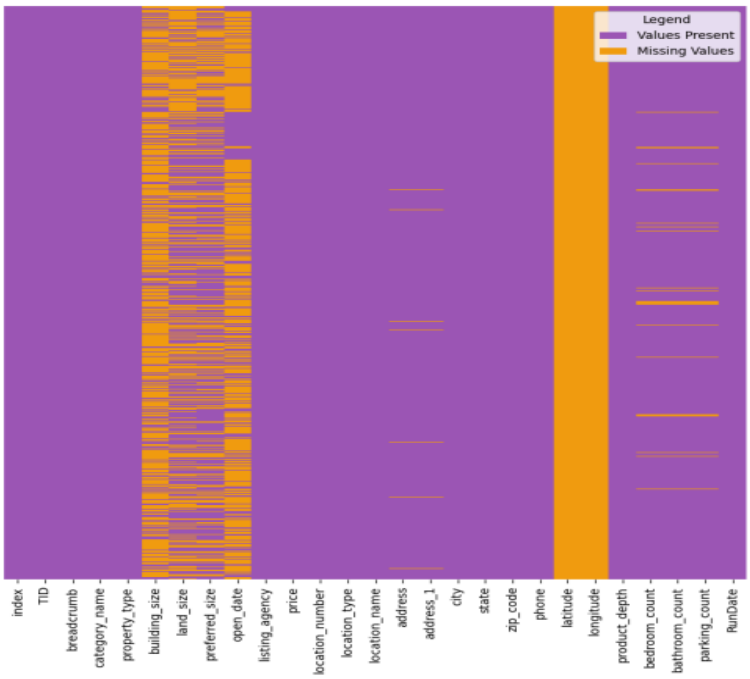

Figure1 : Heatmap visualisation to identify null values

Heatmap visualization (Figure1) is done to graphically represent the presence of missing data (null values) in a dataset, providing a quick and intuitive way to identify patterns and areas where data might be incomplete.

The visualization reveals that several columns, including open date(698 missing), building size(720 missing), land size(467), and preferred size(391), exhibit a significant number of missing values. Given this high incidence of missing data, it is better to remove these columns to avoid potential inaccuracies in the analysis. So I have removed these columns as its not necessary.

Although longitude and latitude also show substantial missing entries, these will not be removed as the plan is to get these using geocoding, which will be crucial for the spatial visualizations relevant to the research questions.

Columns such as TID, breadcrumb, category_name, location_type, phone, and run date are not required for our analysis and is be dropped to streamline the dataset.

Since there are only 33 missing entries out of 1000 for bedroom_count, bathroom_count, and parking_count since its only 3.3% of data is missing its best to remove those rows. (Bhumitdevni, 2023)

Upon further analysis of the NT Australia dataset, initially presumed to encompass the entirety of the Northern Territory, it became evident through latitude and longitude plotting that the data predominantly pertains to the suburban region of Darwin as seen in figure 2 and 3. Considering this geographical focus and the common region in all the remaining 3 data sources, the scope of my project has been refined. This localized dataset provides a valuable opportunity to explore and answer key research questions within a specific urban context. The narrowed focus will concentrate on understanding the interplay between property characteristics, crime rates, demographic shifts, and the proximity to educational institutions, particularly within Darwin city, which has emerged as the central area common to all datasets and most pertinent to my research objectives.So, I have changed my questions focusing on Darwin city:

1. How do property characteristics (such as the number of bedrooms, bathrooms, and parking availability) and local crime rates collectively influence housing prices in the Darwin city, Northern Territory?
2. How does the changing population within the Darwin city, Northern Territory affect the property values?
3. How does the proximity to educational institutions affect housing prices in the Darwin city, Northern Territory?

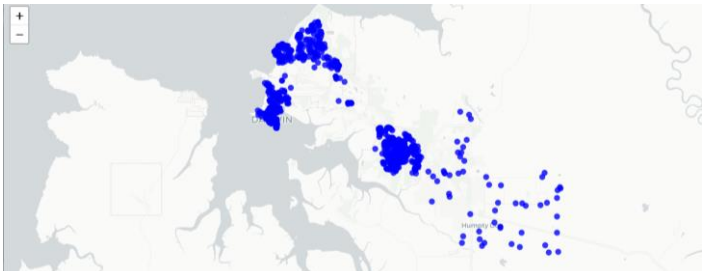

Figure 2: Geographical Distribution of Property Locations in Darwin Northern Territory

Figure 3: Concentration of Property Listings in the Darwin Northern Territory

Upon reviewing the remaining data, several observations are noted:

- To ensure the analysis specifically targets Darwin, the dataset was filtered to include only properties within the Darwin city limits, aligning with the geographical focus of the three additional datasets utilized in this study.
- Anomalies in price and location_name suggest they might mistakenly contain **similar data**, upon investigation in Python it was found that all the data in price and location_name is matching hence it's duplicate and we can just drop location_name column.
- It is observed that there are **no duplicate entries** in this dataset.
- I utilized the **geocode add-on** within **Google Sheets** to obtain precise latitude and longitude coordinates for the listed addresses in a sample of 1,000 entries from the NT Australian dataset. By applying this tool, I successfully pinpointed the geographical locations. I extracted the addresses from the dataset, inputted them into Google Sheets, and then employed the geocode extension to acquire the necessary coordinates. Initially, I attempted to generate latitude and longitude using Python code, but the mapping accuracy fell short when compared to the results from Google Sheets' geocoding. Consequently, I opted to leverage the Google Sheets geocode add-on for more reliable data visualization and to address my research questions effectively.I merged Australian dataset for the Northern Territory, which lacked latitude and longitude information, with an csv file generated from the geocode extension that provided these coordinates as shown in the images below.
- Address and address_1 overlap, with address_1 being more specific, thus the address column is dropped.
- Upon review of the dataset, the price column displays irregular entries as depicted in Figures 4 and 5, with variations such as "$220,000-$250,000" and non-numeric strings including "Auction" and "$450,000 - negotiable". This inconsistency necessitates data cleansing. Due to complexity of this column , all commas were removed to streamline the figures, and only rows with numeric values were retained, thereby discarding certain entries.

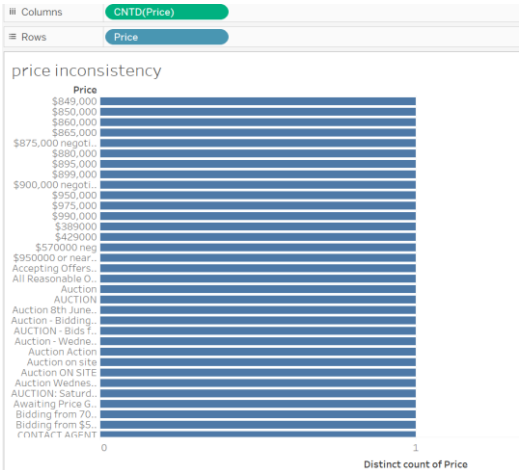

Figure 4 : Visualization of Price inconsistency in Dataset

Figure 5 : Price Category Distribution

Outlier Detection:

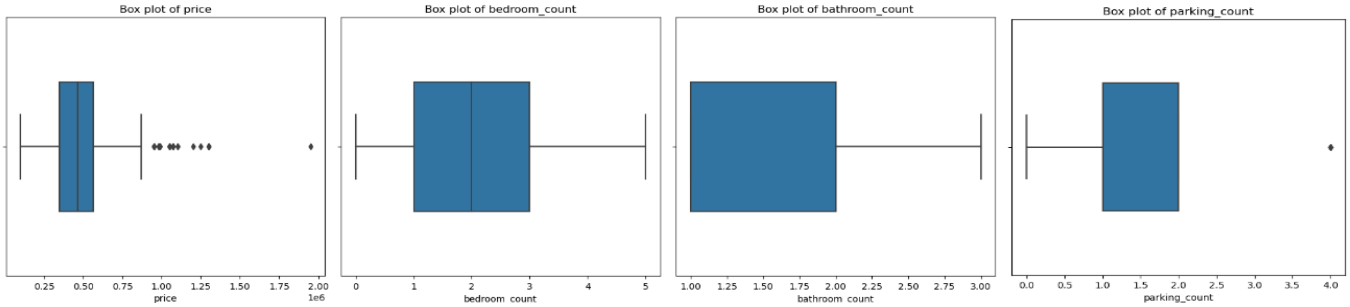

Figure 6: Outlier Detection across Property Features

This set of box plots displays the distribution of prices, bedroom count, bathroom count, and parking count, highlighting potential outliers beyond the whiskers that could impact the analysis.

The box plot analysis(figure 6) revealed the presence of outliers—data points significantly differing from other observations. To manage these outliers, the following approach was applied:
Z-Score Method: Z-scores were calculated for the remaining 'price' data, with values over 3 (indicating a deviation of more than three standard deviations from the mean) identified as extreme outliers and removed.

The figure 7 box plots reflect data post-outlier removal, providing a more accurate representation of the core dataset's distribution. These visualizations offer insights into the median, spread, and overall distribution for the 'price', as well as the counts for bedrooms, bathrooms, and parking spaces, now with reduced influence from atypical and potentially misleading values.

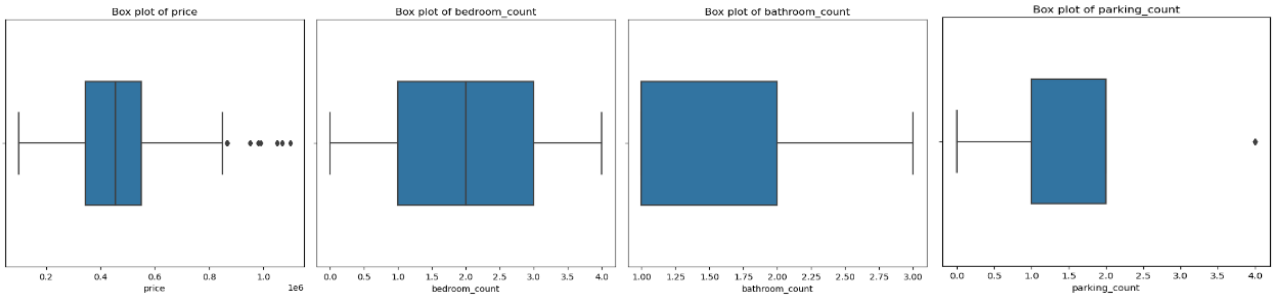

Figure 7: Post-Outlier Removal Distribution of Property Features — These box plots illustrate the refined distribution of property prices and features after applying outlier removal techniques, resulting in a dataset better suited for accurate statistical analysis.

Data wrangling and checking for the aforementioned dataset were conducted utilizing Python (via Jupyter Notebook) and Google Sheets for geocoding purposes. The mapping was accomplished with R, while identification of price inconsistencies was performed using Tableau.

1) **Question 1** :
How do property characteristics (such as the number of bedrooms, bathrooms, and parking availability) and local crime rates collectively influence housing prices in the Darwin city, Northern Territory?

To address Question 1 we will first need to prepare the relevant datasets :**NT Australian Dataset** as mentioned above and **NT Crime Statistics** (This tabular CSV dataset comprises approximately 45,000 entries spanning from 2008 to 2022, detailing crime incidents throughout the Northern Territory. It encompasses diverse data types, including categorical (e.g., Offence Category), nominal (e.g., Reporting Region), and temporal (e.g., Year of Crime) variables, providing an extensive snapshot of crime patterns (https://data.nt.gov.au/dataset/current-northern-territory-crime-statistics-may-2022)).

The data wrangling for the NT Australian Dataset is already complete as described earlier. Next, we will focus on refining the NT Crime Statistics dataset:

- **Unwanted Column Removal** : Due to all entries in the 'Statistical Area 2' column being '–' (null) for the required data(darwin), we will remove this column. Similarly, the 'As At' column, which is not relevant to our analysis, will also be dropped.
- There are no **duplicate entries** in this dataset which was observed using python.
- **Handling Null Values**: In the crime dataset, all entries are complete except for those related to 'Alcohol Involvement' in assault cases and 'DV Involvement' (domestic violence). Assuming that when data isn't recorded, it indicates a negative response to those particular questions in this case, hence I am replacing null or '–' values or text with '***Alcohol Involvement unknown'*** (figure 8) in Alcohol Involvement column with 'No alcohol involved' and '–' in 'DV Involvement' column with 'Non-DV'.
- **Filtering by Region :**will filter the data to include only entries from Darwin, aligning it with NT Australian Dataset.
- Data wrangling and validation for the crime dataset was conducted utilizing Python within a Jupyter Notebook environment and Excel for analytical visualization.

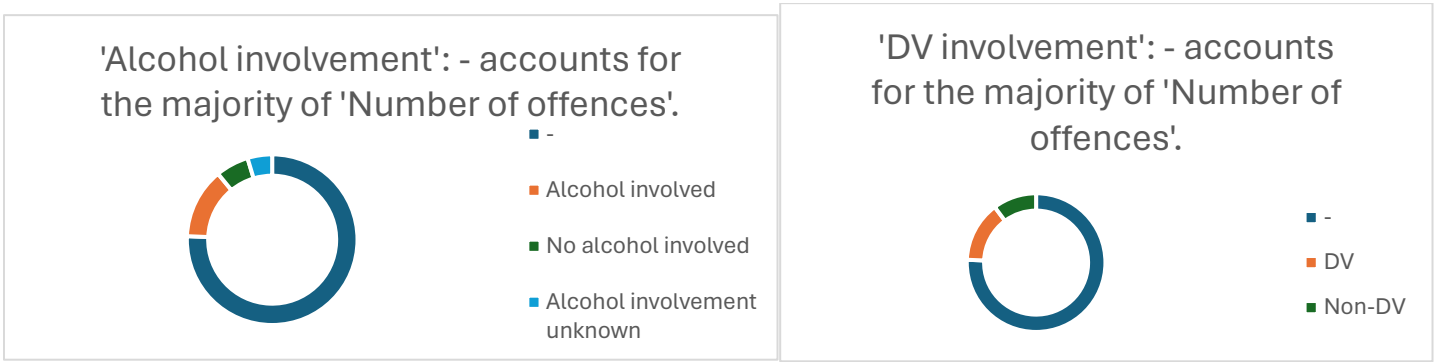

Figure 8: Doughnut chart representation of majority values in alcohol involvement and domestic violence (DV) columns.

2) **Question 2**: How does the changing population within the Darwin city, Northern Territory affect the property values?

In order to answer this question 2 we need the following datasets :**NT Australian Dataset** as mentioned above and **NT Government Regions population (1986-2023):** (This dataset provides comprehensive population reports from 1986 to 2023 across the Northern Territory, featuring both spatial (e.g., region) and temporal (e.g., population data over time) attributes, disaggregated by age, sex and Indigenous status. (https://data.nt.gov.au/dataset/nt-government-regions-1986-2023/resource/754fb4e0-33d8-4641-b06e-4e867407a171)).

The data wrangling for the NT Australian Dataset is already complete as shown above. Next, we will focus on refining the **NT Government Regions population** dataset:

- For **this dataset** there are no null values and no duplicate entries.
- **Filtering by Region :**will filter the data to include only entries from Darwin, aligning it with NT Australian Dataset.
- As **Status** column is not significant for analysis, it was dropped.
- The data wrangling for the NT Government Regions Population dataset was done using **python**(Jupyter Notebook).

3) **Question 3**: How does the proximity to educational institutions affect housing prices in the Darwin city, Northern Territory?

In order to answer this question we will make use of the following datasets: **NT Australian Dataset** and **NT School List** dataset consisting of 271 columns x 16 columns such as school names, addresses, levels (Pre, Primary, Middle, Senior), sectors, and remote categories.(https://data.nt.gov.au/dataset/school-list/resource/3d8f5e87-ccff-45d8- 96be-dbbd4db5070)

The data wrangling for the NT Australian Dataset is already complete as shown above. Next, we will focus on refining the **NT School List** dataset:

- Upon examination using Python, the NT school list dataset was found to be **free of duplicates**.
- The data was refined to feature entries exclusively from Darwin to correspond with the NT Australian Dataset.
- Columns, such as Postal Address, URI, Email, Principal, Telephone Number, Fax Number, and NTG Remote Definition, were omitted to simplify analysis.
- Latitude and longitude coordinates for the schools' physical addresses were acquired through the **geocode add-on** in Google Sheets.
- This latitude and longitude was merged with NT school list dataset.
- Using R script ,real estate and school data were read, was converted into spatial objects, and then distance from each property to the nearest school was calculated. It adds this distance information to the real estate dataset. The script ensures that the data includes only properties with both a recorded distance to the nearest school and a listed price, omitting any missing values. The final output is a refined dataset prepared for further analysis, specifically to examine the impact of a property's proximity to schools on its market price. This process sets the stage for potential regression or other statistical analyses to understand the influence of school proximity on property values.
- The data wrangling for the NT school list dataset was done using **python**(Jupyter Notebook) and **R**.

DATA EXPLORATION

1. **Question 1**: How do property characteristics (such as the number of bedrooms, bathrooms, and parking availability) and local crime rates collectively influence housing prices in the Darwin city, Northern Territory?

To begin investigating the first question, let's methodically examine each of the datasets relevant to this inquiry.

Property Price Variation within Darwin City(Proportional Symbol Map):

This map uses proportional symbols (circles) to represent property prices across Darwin City. The size of each circle corresponds to the price of the property, allowing for a quick visual assessment of pricing across different neighbourhoods. The use of colour intensity along with size adds an additional layer of information, making it easier to identify high-price and low-price areas. Such a map is particularly useful for visual spatial patterns in housing prices, which could be influenced by factors like proximity to the city centre, amenities, or waterfront views. It's clear that there are clusters of higher-value properties in certain areas, and this type of map is particularly useful for real estate analysis, urban planning, and economic studies, providing a clear visual of how property prices vary across different parts of the city.(figure 9)

Tableau was used to visualise this Proportional symbol map.

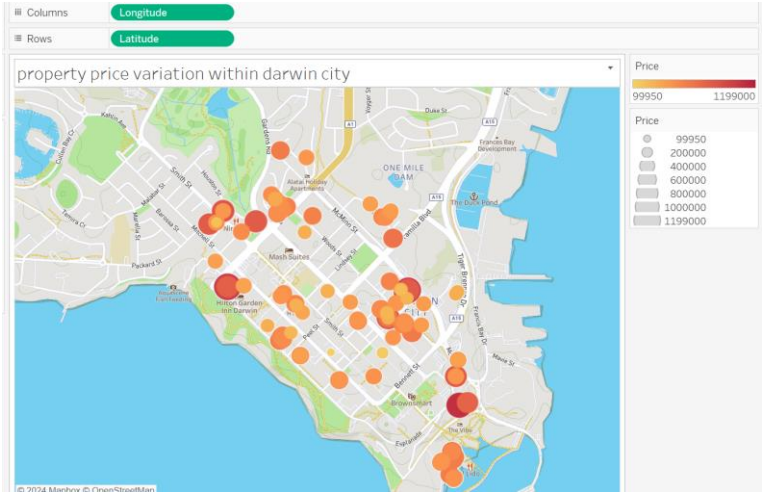

Figure 9: Proportional symbol map showing property price variation within Darwin city

Horizontal bar chart representing total number of offenses categorised by type within darwin

- The y-axis lists types of offenses in descending order, with the most serious crimes at the top and less serious offenses towards the bottom.
- The x-axis represents the number of offenses recorded for each type, extending from 0 to over 1000 offenses.
- Each bar's length corresponds to the total number of offenses recorded for each type, providing a visual measure of frequency.
- The darker bars, potentially indicating a higher count, suggest that "Theft and related offenses (other than MV)" and "Property damage offenses" are the most frequent, with counts nearing or surpassing 1000.
- Serious crimes such as "Murder," "Manslaughter," and "Driving causing death" have very few occurrences in comparison to other types of offenses.
- From this chart, it can be inferred that property-related offenses and theft are the most common, while violent crimes and offenses against the person occur less frequently. This visualization can be crucial for law enforcement and public policy makers to allocate resources effectively and address crime rates within a community. (figure 10)
- R was used to visualise this bar chart.

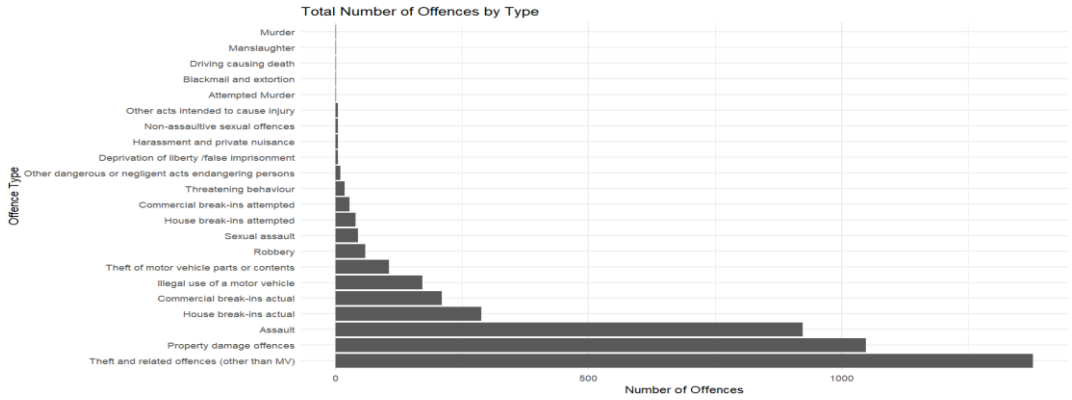

Figure 10: Horizontal bar chart representing total number of offenses categorised by type within darwin

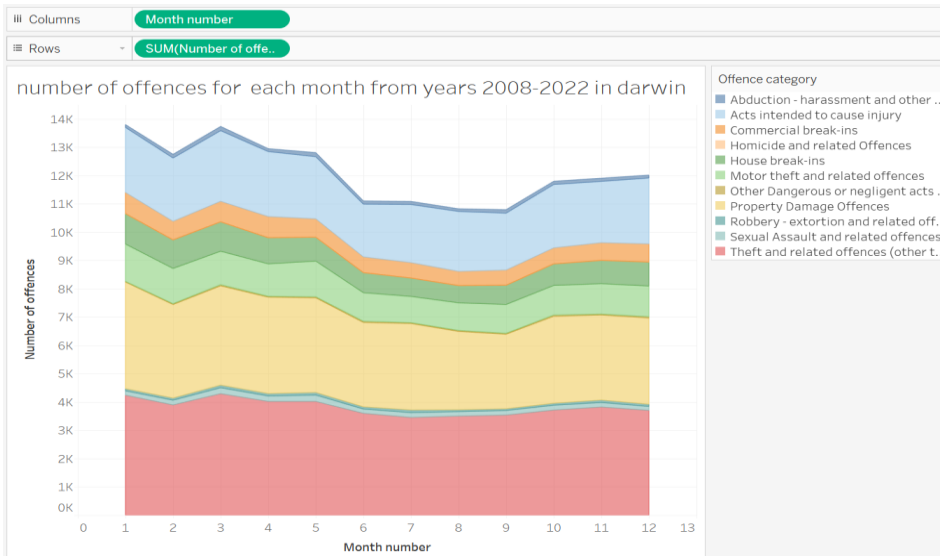

Figure 11 : Area chart that represents the total number of offenses for each month from the years 2008-2022 in Darwin.

Area chart that represents the total number of offenses for each month from the years 2008-2022 in Darwin.

The x-axis displays the month numbers, from 1 to 12, corresponding to January through December. The y-axis shows the number of offenses, stacked by category as indicated by the colour legend.(figure 11)

- The color-coding and stacking allow us to see the proportion of each offense category to the total offenses each month.
- The layering of offenses shows some categories consistently constitute a larger proportion of the total, possibly suggesting patterns that could inform preventive measures or resource allocation.
- Seasonal fluctuations are evident, with a higher count of offenses noted at the start of the year, followed by a gradual decline in offenses as the year progresses. This pattern suggests a trend where crime rates peak during the earlier months and then diminish over the course of the year. However, without specific month labels or year-on-year data, this observation is speculative based on the chart provided.
- Tableau was used to visualise this area chart.

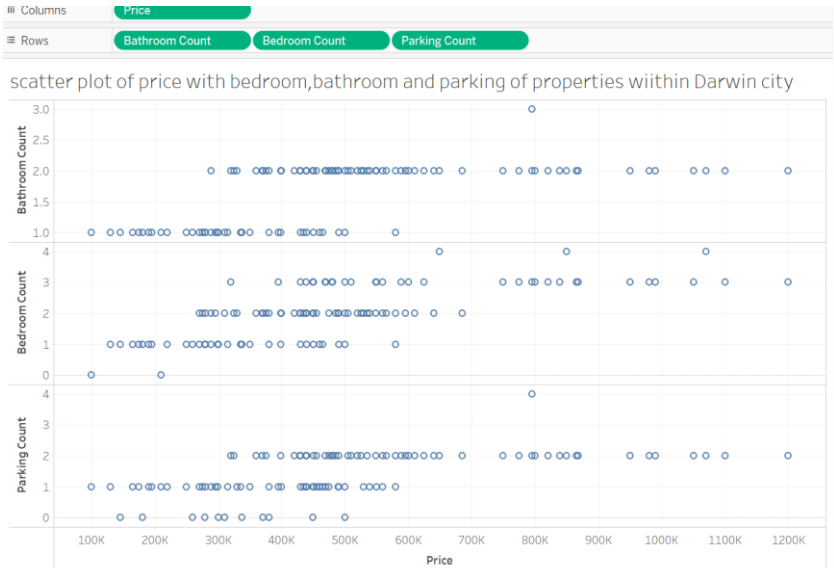

Figure 12 : The scatter plot shows the relationship between property prices and the number of bathrooms, bedrooms, and parking spaces

The scatter plot shows the relationship between property prices and the number of bathrooms, bedrooms, and parking spaces of properties in Darwin city.

- The plots align data points along the price axis to reflect each property's characteristics.
- A denser clustering of points at the lower end of the price spectrum indicates a higher number of properties with fewer bathrooms and bedrooms.
- Some properties with higher bedroom and bathroom counts are associated with higher prices, indicating a positive correlation between these features and price.
- The parking count has a more scattered distribution across the price range, suggesting that the number of parking spaces might not be as strongly correlated with the price as the number of bedrooms or bathrooms are.
- This visualization suggests that while bathroom and bedroom counts are important factors in property valuation, the number of parking spaces may not significantly influence the price. Tableau is used for this visualisation.(figure 12)
- Tableau was used to visualise this scatterplot.

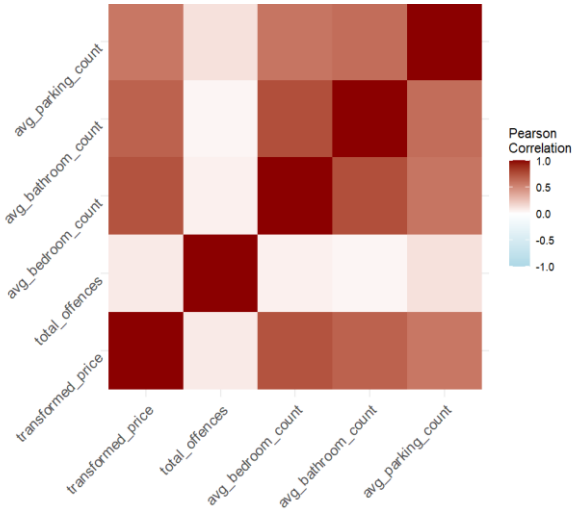

Figure 13: This heatmap shows the correlation between different numeric variables such as price , bedroom_count etc in the dataset.

Correlation Matrix (Heatmap):

This heatmap(figure 13) shows the correlation between different numeric variables in my dataset. A correlation coefficient can range from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

- The colours represent the strength and direction of the correlation, with dark red indicating positive correlation and light blue indicating negative correlation.
- From the matrix we can observe that avg_bedroom_count and avg_price show a strong red color, it means that as the number of bedrooms increases, the price tends to increase as well.
- This is a powerful exploratory tool to identify potential predictors for housing prices and to understand the relationships between different features of the properties.
- The correlation between total offenses and property features or prices is likely meant to be shown here. However, it's either not present or not visible, which may be due to lack of granularity in crime data. If crime data were more detailed, one might expect to see either a positive or negative correlation indicated by the color shading.

The heatmap(figure 13) provides a visual representation of the correlation between various property features and housing prices. However, it lacks the necessary detail to analyze the impact of local crime rates on these prices due to the absence of suburb-level crime data. For a comprehensive analysis, the crime data would need to be as granular as the property data, which includes specific information down to the suburb level.The missing piece is the suburb-level crime statistics which would allow us to correlate the frequency or severity of crimes in each suburb with the housing prices in that same suburb. Without this level of detail, any attempt to analyze the correlation between local crime rates and property values would be imprecise. For instance, crime rates might vary significantly within a city from one suburb to another, and this variation could have a corresponding impact on housing prices, which cannot be captured with broader city-level crime data. The existing dataset likely aggregates crime data at a city-wide level, which masks the local variations that are crucial for determining the safety of a neighborhood and, by extension, its property values. The inclusion of suburb-level crime data would provide the granularity needed to uncover patterns and correlations that are pertinent to homebuyers, real estate investors, and policymakers. R was used to visualise this correlation matrix.

Data Aggregation : The data is grouped by address, and several summary statistics like correlation matrix as shown above and regression analysis as shown below are computed for each group. This includes average price, total offenses at that address, and average counts of bedrooms, bathrooms, and parking spaces.

```
Call:
lm(formula = avg_price ~ avg_bedroom_count + avg_bathroom_count +
    avg_parking_count + total_offences, data = combined_data_aggregated)

Residuals:
    Min      1Q  Median      3Q     Max
-351763 -112154  -10068   76011  512513

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         80581.963  54367.062   1.482  0.14089
avg_bedroom_count  145568.777  25271.232   5.760 6.51e-08 ***
avg_bathroom_count  12799.567  45191.133   0.283  0.77748
avg_parking_count   74253.662  25995.247   2.856  0.00504 **
total_offences         -1.130      2.134  -0.529  0.59747
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 152400 on 121 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.5266
F-statistic: 35.76 on 4 and 121 DF,  p-value: < 2.2e-16
```

Figure 14 : Regression summary or output

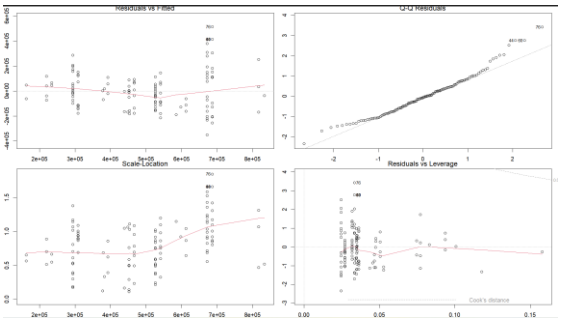

Figure 15: Residuals plots

The regression output and residual plots are used together to evaluate the impact of property features and crime rates on housing prices in Darwin City.

Regression Output Interpretation: (figure 14)

- The model indicates that the number of bedrooms has a strong positive effect on housing prices (significant at $p < 0.001$).
- Parking availability also shows a significant positive correlation with price ($p < 0.05$).
- Total offenses do not significantly influence housing prices in this model.
- The R-squared value suggests that approximately 54.17% of the variance in housing prices is explained by the variables in the model.

Residual Plots Interpretation: (figure 15)

- The "Residuals vs Fitted" plot should ideally show no discernible pattern; however, the presence of a pattern or outliers (760, 860) may suggest non-linearity or heteroscedasticity.
- The "Q-Q Residuals" plot indicates how closely the residuals follow a normal distribution. Deviations from the line, especially in the tails, suggest potential issues with normality.
- The "Scale-Location" plot (not mentioned in the text) is used to check for homoscedasticity — the spread of residuals should be consistent across all levels of fitted values.
- The "Residuals vs Leverage" plot helps to identify influential cases (points far from the center of the leverage). Cook's distance lines can indicate points that have a large influence on the calculation of the regression coefficients.

The regression analysis shows that certain property characteristics, namely the number of bedrooms and parking availability, are significant predictors of housing prices in Darwin City. The model's diagnostic plots suggest that while the model is generally significant and explains a substantial proportion of variance in housing prices, there may be issues such as potential outliers or non-constant variance in the residuals that could affect the model's assumptions. The lack of significant impact from total offenses might be due to the mentioned data granularity issues, as the model cannot account for the localized effects of crime rates on property values without more detailed crime data at the suburb level. These findings illustrate the complexities in modeling real estate prices and the importance of thorough diagnostic checking in regression analysis.

R was used to visualise Regression and residual plots.

Further exploration for property dataset is shown in Appendix part 1 and for crime dataset is shown in Appendix part 2.

Question 2 : How does the changing population within the Darwin city, Northern Territory affect the property values?

Pie chart illustrates the population proportion by Aboriginal status in the year 2022 in Darwin

The chart (figure 16) is divided into two segments:

- The larger segment, covering the majority of the chart (86.9%), represents the Non-Aboriginal population.
- The smaller segment (13.1%) represents the Aboriginal population.

This visualization clearly indicates that the Non-Aboriginal population makes up a substantial majority in the specified year, while the Aboriginal population constitutes a significant minority. The data could be valuable for understanding demographic composition, informing policy decisions, and ensuring representation and resources are appropriately allocated. R was used to visualise this Pie chart.

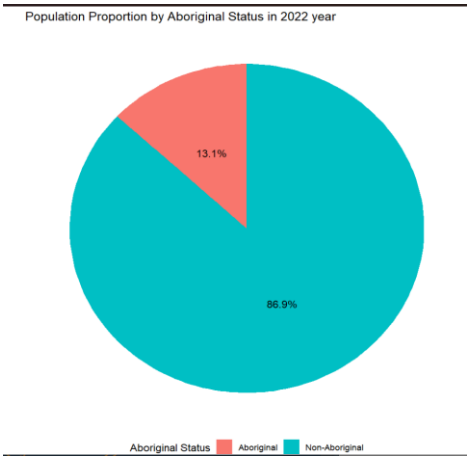

Figure 16: Pie chart illustrates the population proportion by Aboriginal status in the year 2022 Darwin.

Line graph depicts the population trends for male and female residents in Darwin city over a span of 37 years. It shows two lines, each representing one sex, charting population numbers against time from 1968 to 2023.(figure 17)

- Both male and female populations have been increasing over the years.
- The male population consistently remains higher than the female population throughout the period, suggesting a gender imbalance in the city's demographics.
- There are points where the growth rate appears to change, such as around 2002 where both lines show a steeper increase, indicating a period of accelerated population growth.
- The trend lines show a slowing of population growth for both males and females around 2018, as indicated by the lines flattening out.

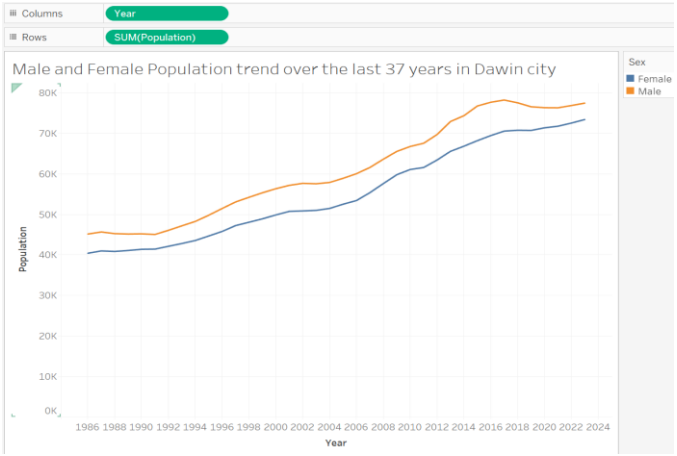

Figure 17: Line graph depicting the population trends for male and female residents in Darwin over a span of 37 years.

Overall, the population of Darwin city has been on an upward trajectory for both genders. The reasons behind these trends would require additional context such as migration, birth and death rates, or economic factors. Tableau was used to visualise this graph.

The stacked bar graph visualizes the total population distributed across different age groups from 1986 to 2023, categorized by Aboriginal status.(figure 18)

Each age group is represented by a pair of bars:

- The orange bars represent the Non-Aboriginal population within each age group.
- The blue bars represent the Aboriginal population within the same age groups.

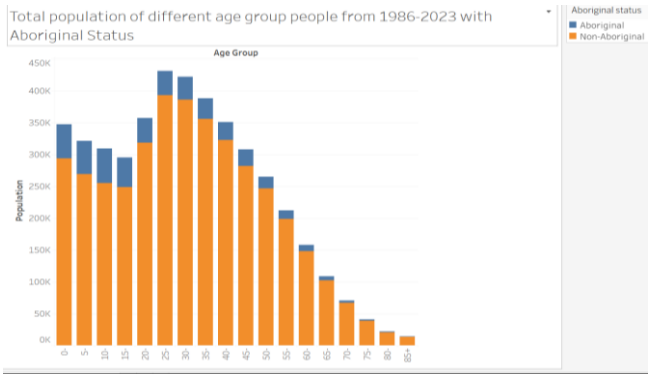

Figure 18: The stacked bar graph visualizes the total population distributed across different age groups from 1986 to 2023, categorized by Aboriginal status.

From the younger to older age groups, the population first increases, reaches a peak, and then declines, which is typical due to lower numbers in the oldest segments of a population. The graph shows that the Non-Aboriginal population is consistently higher in number across all age groups compared to the Aboriginal population. The largest populations for both categories appear to be in the middle age groups, possibly indicating the most economically active segments of the

This type of visualization can help policymakers understand demographic structures and provide insights for planning services such as education, healthcare, and elderly care. It also helps in recognizing the age distribution differences between Aboriginal and Non-Aboriginal populations, which can be crucial for targeted social programs. Tableau was used to visualise this graph.

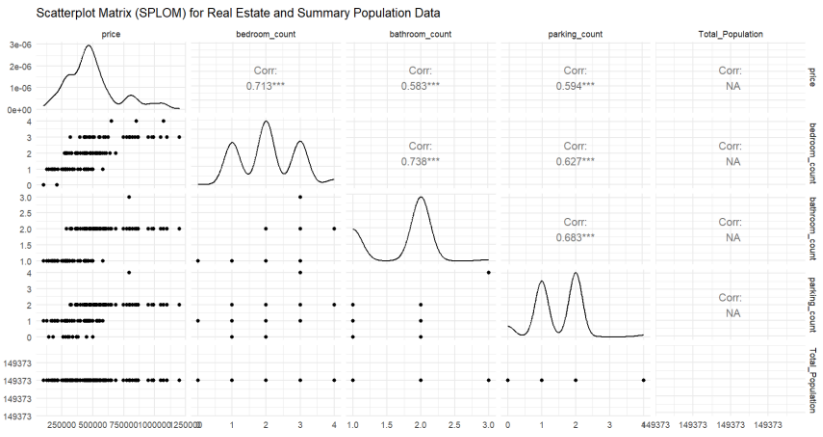

Figure 19 : The scatterplot matrix (SPLOM) displayed explores the relationship between property prices and various property characteristics such as bedroom count, bathroom count, and parking count, alongside total population data.

The scatterplot matrix (SPLOM) displayed explores the relationship between property prices and various property characteristics such as bedroom count, bathroom count, and parking count, alongside total population data.(figure 19)

- The diagonal contains histograms and density plots showing the distribution of each variable: property prices, bedroom count, bathroom count, and parking count. The shape of these distributions can provide insights into the spread and central tendencies of each feature.
- Below the diagonal are scatterplots showing the pairwise relationships between variables. Correlation coefficients are provided to quantify these relationships, with asterisks indicating the significance level. For example, a high positive correlation between bedroom count and price suggests that as the number of bedrooms increases, so does the price.

Above the diagonal, the corresponding correlation coefficients are again displayed, with their significance levels marked by asterisks.

The total population data doesn't seem to be plotted, likely due to it not being at the same granularity as the property features. This aligns with my observation that the lack of address-level population data makes it difficult to directly assess how population changes at a more local level may affect property values. For a comprehensive analysis, population data would need to be aligned with property locations at the suburb level, allowing for a detailed study of how demographic shifts impact real estate prices within specific neighborhoods. R was used to visualise this scatterplot matrix.

Further exploration for population is shown in Appendix part 3.

Question 3 : How does the proximity to educational institutions affect housing prices in the Darwin Northern Territory?

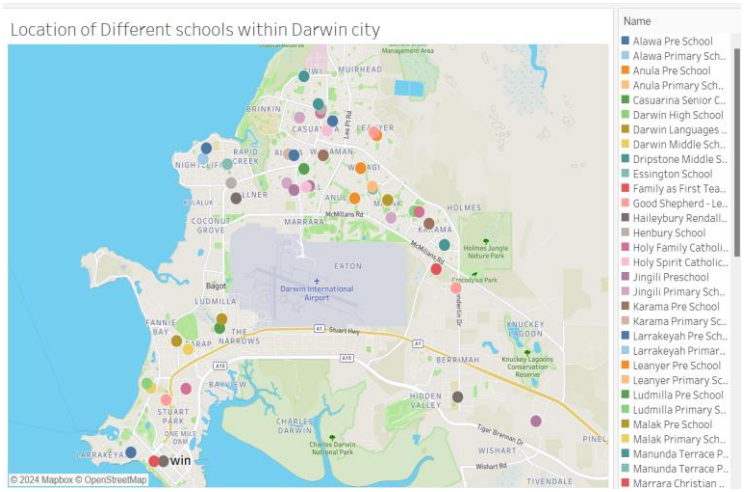

Figure 20 : Map showing the location of different schools within Darwin city .

Map showing the location of different schools within Darwin city : This map provides a geographic visualization of the locations of various educational institutions within Darwin City. Each dot represents a school, and the color indicates different types of schools.(figure 20)

The list on the right gives names of the institutions, which includes a range from pre-schools to high schools. The concentration of schools within certain areas could suggest neighborhood educational hubs, which might be attractive to families with children, potentially affecting real estate demand and prices in those areas.

For property value analysis, this map could be overlaid with property locations and prices to analyze patterns or correlations between school proximity and housing prices. It serves as a foundational layer for spatial analysis, which could be enhanced by additional data such as property characteristics, demographic information, and transportation accessibility to provide a more nuanced understanding of the factors driving real estate values in Darwin City.Tableau was used to visualise this map.

The scatter plot visualizes the relationship between the proximity of properties to the nearest school (x-axis) and their housing prices (y-axis), segmented into three clusters represented by different colors.(figure 22)

Cluster 1 (Red): Consisting of fewer properties, this cluster has higher-priced homes that are relatively close to schools. The regression line suggests a slight negative trend, indicating that within this high-value market segment, closer proximity to schools might not be as significant a factor in housing price.

Cluster 2 (Green): These properties are moderately priced and tend to be located very close to schools. The regression line shows a flat trend, suggesting that within this cluster, the distance to schools does not have a major impact on price variations.

Cluster 3 (Blue): This is the largest cluster by count and includes properties with a wide range of distances to schools. The majority are lower-priced, and the regression line indicates a slight negative trend, which could mean that for this cluster, properties closer to schools may command a slight premium. The regression line for this cluster shows a negative trend, which typically would suggest that as the distance to the nearest school decreases, the price increases.

The clusters and regression lines imply that while proximity to schools is a factor in property pricing, its impact varies across different market segments. For higher-priced properties, proximity to schools might be less important compared to other luxury features. For mid-range properties, the effect of school proximity on price is minimal. Meanwhile, for lower-priced properties, there seems to be a minor correlation where closer proximity could influence prices positively, likely appealing to families prioritizing access to education. R was used to visualise this scatterplot regression analysis.

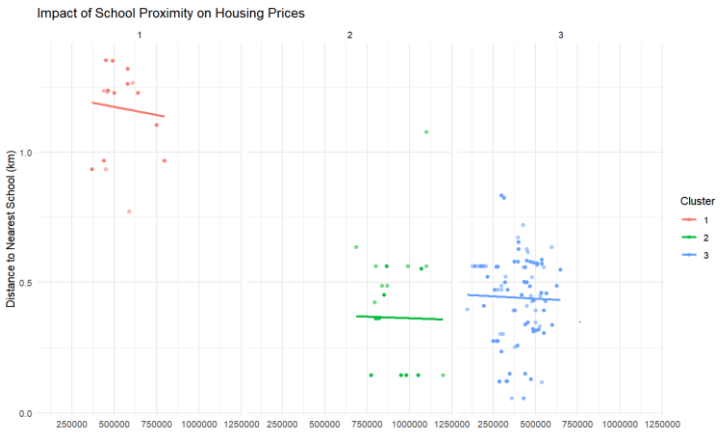

Figure 22 : The scatter plot visualizes the relationship between the proximity of properties to the nearest school (x-axis) and their housing prices (y-axis), segmented into three clusters represented by different colors.

Conclusion

This project provided an in-depth exploration of how property characteristics, crime rates, demographic changes, and proximity to educational institutions influence housing prices in Darwin City, Northern Territory, Australia. Through rigorous data analysis, it was found that property features such as the number of bedrooms and parking availability have a significant positive impact on housing prices, while crime rates did not show a significant correlation with housing values in this specific model. Demographic factors like the changing population composition were also examined to understand their influence on property values.

Demographic Changes: The study also looked into how shifts in the population composition within Darwin City influence housing prices. It was found that demographic factors, such as age distribution and household composition, play a crucial role in shaping housing market dynamics. These changes affect consumer behavior and housing needs, thereby influencing property values.

Educational Institutions: Proximity to educational institutions was another key factor analyzed. The study noted that proximity to well-regarded schools tends to increase property values due to the high demand among families for access to quality education for their children. However, the magnitude of this effect varied across different market segments, potentially influenced by other factors like the type of housing or local amenities

The data-driven approach underscored the importance of detailed, suburb-level crime data to provide more precise insights into the relationship between local crime rates and property values, population and property values. The analysis further highlighted the influence of demographic shifts and proximity to educational institutions on housing prices, although the impact varied across different market segments**.** The exploration of these elements provided a comprehensive understanding of the factors driving housing prices in Darwin City.

Reflection

This project underscored the importance of data granularity and quality in statistical modeling. The lack of detailed crime data at the suburb level posed a significant challenge, emphasizing the need for more refined data in future analyses. The study also demonstrated the value of performing thorough diagnostic checks during regression analysis to identify potential model assumption violations, such as non-linearity or heteroscedasticity.

Looking back, it would have been beneficial to incorporate more diverse data sources to fill existing gaps, particularly in obtaining more detailed crime statistics. This experience has also illuminated the importance of considering various demographic factors and their dynamic shifts over time in urban studies. For future projects, employing alternative modeling techniques or more sophisticated data imputation methods might prove beneficial in managing incomplete data more effectively. This project not only provided valuable insights into the housing market of Darwin City but also offered critical lessons on the complexities of data-driven real estate analysis

Bibliography

[1] Bhumitdevni. (2023, September 19). EDA of Australia Housing. Kaggle. https://www.kaggle.com/code/bhumitdevni/eda-of-australia-housing

APPENDIX

Part 1: More Exploration done for Property Dataset

Housing Prices by Bedroom Count (Boxplot):

The boxplot breaks down average housing prices based on the count of bedrooms. Each box represents the interquartile range (IQR) of prices for properties with a specific number of bedrooms, with the median price indicated by the line within the box. The x-axis represents the average bedroom count, with categories likely ranging from 0 (studio) to 4 bedrooms. The y-axis represents the housing prices.(figure 1)

- The median price increases with the number of bedrooms, which is expected since more bedrooms usually mean a larger and potentially more expensive house.
- There is a wide range in prices for houses with 3 and 4 bedrooms, as shown by the long whiskers.
- The price distribution for houses with 0 or 1 bedroom is quite narrow, indicating less variability in the prices of smaller properties.
- There are outliers in several categories, especially for 3 and 4 bedrooms, suggesting some houses are priced much higher than the average for their bedroom count.

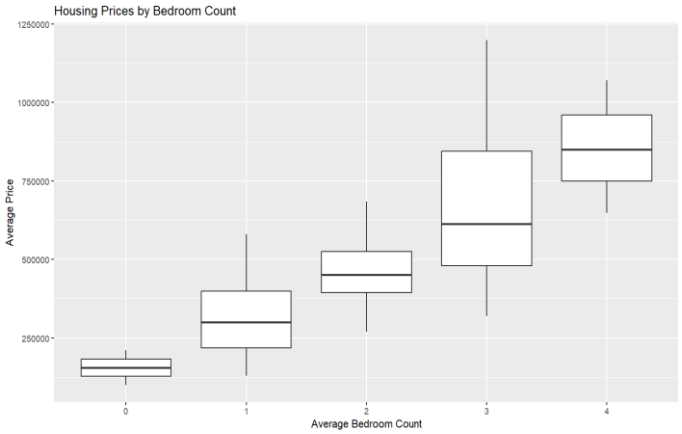

Figure 1: Boxplot showing housing price by bedroom count

This box plot is useful for understanding how the number of bedrooms relates to housing prices and the variation within each category. It can also be helpful for identifying trends and anomalies in the data.

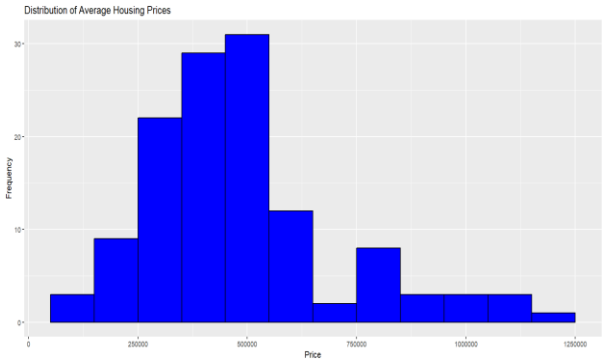

Figure 2 : Distribution of average housing price

Distribution of Average Housing Prices (Histogram):

This histogram represents the distribution of average housing prices. The x-axis indicates the price of the houses. The y-axis shows the frequency, which refers to the number of occurrences within each price range. (figure 2). This shows the distribution of housing prices, which is skewed to the right. This suggests most properties fall within a mid-range price bracket, with fewer properties at the high-end.

From this histogram, we can infer that: The distribution is not symmetrical; it has a peak in the middle and is skewed to the right, meaning there are a few houses with very high prices. The most frequent average price range for houses falls between approximately $250,000 and $500,000, as indicated by the tallest bar. There are significantly fewer houses in the lowest (<$100,000) and highest (>$800,000) price ranges.

This kind of visualization clarifies the overall spread of housing prices in given dataset and can be indicative of the affordability of housing in the area under study.

Part 2: More Exploration done for Crime Dataset

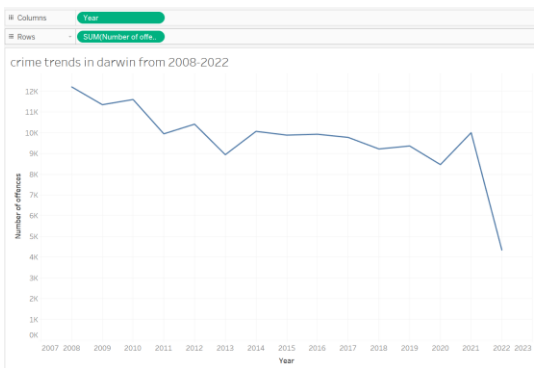

Figure 3 : Line chart of crime trends over 15 years

Line Chart - Crime Trends Over the Past 15 Years:

The chart is designed to display the total number of offenses year by year.(figure 3)

- The y-axis quantifies the number of offenses, while the x-axis represents the year.
- A notable feature of the line chart is the trend line's movement over time downward trend in the total number of crimes.
- The sharp decrease in offenses in the last year could indicate effective crime reduction strategies or other external factors, like a change in data collection methods or an actual reduction in crime incidents.