ALY 6015: Intermediate analytics
Final Project Report

Melbourne Housing Data analysis

Jose De Leon
Sanna K Baldeh
Darshil Dinesh Mistry
MD Maniur Rahman

Contents

1 Introduction

1.1 The Data

The data we selected for the final project is the Melbourne Housing Dataset, sourced from kaggle.come. It includes records of significant features and transactions related to seller houses in Melbourne, Australia. The dataset contains information about the property's location, number of rooms, type of property, land dimension, floor area, time of sale, sale price, sale method (e.g., auction or private), and the estate agent involved in the transaction. It provides important information for understanding trends in the Melbourne housing market and modeling property prices, as well as examining how various local characteristics contribute to changing real estate values. The following is a list of all the columns included.

- **Suburb:** The suburb house is located in.
- **Address:** Property address.
- **Rooms:** Number of rooms.
- **Type:** Type of property.
- **Method:** Is the property sold or is still available.
- **SellerG:** Realtor the house is listed with.
- **Date:** Date the property was sold.(if sold)
- **Distance:** Distance from the nearest Central Business District.
- **PostCode:** Postal code of the house.
- **Bedroom:** Number of bedrooms.
- **Bathroom:** Number of bathrooms.
- **Car:** Number of cars that can be parked.
- **Landsize:** Size of the land.
- **BuildingArea:** The carpet area.
- **YearBuilt:** Year when the house was built.
- **CouncilArea:** The Council jurisdiction the house falls under.
- **Latitude:** Geographic latitude.
- **Longitude:** Geographic longitude.
- **RegionName:** Region the house is located in.
- **PropertyCount:** The number of houses available at a location.
- **ParkingArea:** Area of private parking.
- **Price:** Selling price of the property.

1.2 Cleaning

The dataset was collected from real world sources. Hence the data was not very clean. It had many null values and outliers. Thus, before starting our analysis, we analyze the data for missing values and outliers and apply fixes where needed.

```
##
## Total number of rows:  34857
```

```
##
## Number of rows with at least one missing value:  24398
```

```
##        Suburb       Address         Rooms          Type        Method
##             0             0             0             0             0
##       SellerG          Date      Distance      Postcode       Bedroom
##             0             0             0             0          8217
##      Bathroom           Car      Landsize  BuildingArea     YearBuilt
##          8226          8728         11810             0         19306
##   CouncilArea      Latitude    Longtitude    Regionname Propertycount
##             0          7976          7976             0             0
##   ParkingArea         Price
##             0          7610
```

Focusing on the null values, we impute data where possible and remove those which can not be imputed. We also fixed the data types where needed.

```
# Convert BuildingArea to numeric and handle 'Inf'
df$BuildingArea <- suppressWarnings(as.numeric(df$BuildingArea))
df$BuildingArea[is.infinite(df$BuildingArea)] <- NA

# Replace 0 in Landsize with NA
df$Landsize[df$Landsize == 0] <- NA

# Remove any commas, spaces, or non-numeric characters if needed
df$Distance <- gsub(",", "", df$Distance)
df$Distance <- trimws(df$Distance)

# Convert to numeric the variable distance
df$Distance <- as.numeric(df$Distance)
```

```
## Warning: NAs introduced by coercion
```

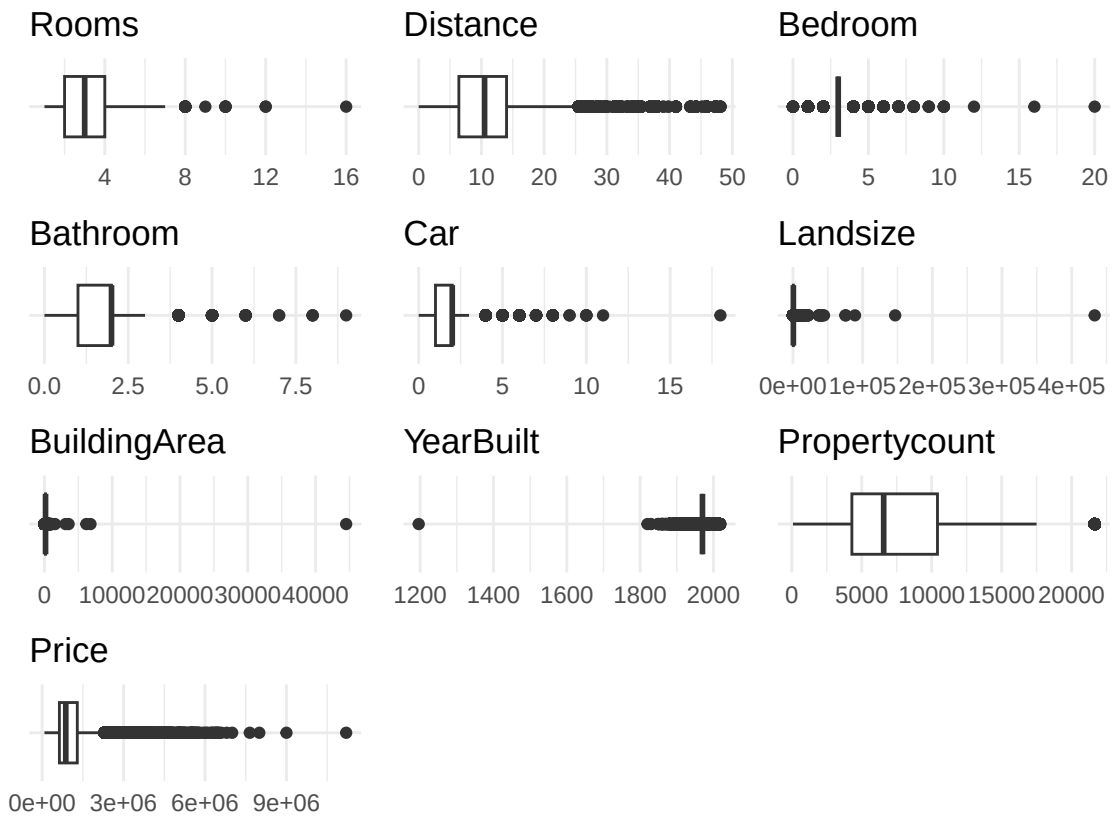
```
#inputing distance missing value
df$Distance[is.na(df$Distance)] <- median(df$Distance, na.rm = TRUE)
# Drop unnecessary columns (keep CouncilArea, ParkingArea, Propertycount)
df <- dplyr::select(df, -Latitude, -Longtitude, -Address, -SellerG, -Date, -Method,
-Postcode)
# Impute with median: Landsize, Car, Bathroom, Bedroom, Propertycount
for (col in c("Landsize", "Car", "Bathroom", "Bedroom", "Propertycount")) {
df[[col]][is.na(df[[col]])] <- median(df[[col]], na.rm = TRUE)
}

# Impute BuildingArea with mean
df$BuildingArea[is.na(df$BuildingArea)] <- mean(df$BuildingArea, na.rm =
TRUE)

# Impute YearBuilt with mode
df$YearBuilt[is.na(df$YearBuilt)] <- mfv(df$YearBuilt, na_rm = TRUE)
```

```r
# Save rows with missing Price to a separate table
missing_price_rows <- df %>%
filter(is.na(Price))
df <- df %>%
filter(!is.na(Price), !is.na(Distance), !is.na(Suburb))
df$Suburb <- droplevels(as.factor(df$Suburb))

# Clean Regionname in both datasets (remove "Metropolitan")
df$Regionname <- gsub(" Metropolitan", "", df$Regionname)
missing_price_rows$Regionname <- gsub(" Metropolitan", "",
missing_price_rows$Regionname)

# Filter out rows with missing price
df <- df %>%
filter(!is.na(Price) & !is.na(Type) & !is.na(Regionname))


df$Propertycount <- as.integer(df$Propertycount)
```

```
## Warning: NAs introduced by coercion
```

```r
df <- na.omit(df)
```

With the missing values dealt with, we can focus on outliers. We check the summary and box plots to understand the outliers.




```
##
```

```
##
##             Suburb          Rooms            Type              Distance
##  Reservoir      :  727   Min.   : 1.000   Length:27244       Min.   : 0.00
##  Bentleigh East:  493   1st Qu.: 2.000   Class :character   1st Qu.: 6.40
##  Richmond       :  439   Median : 3.000   Mode  :character   Median :10.50
##  Preston        :  415   Mean   : 2.992                      Mean   :11.28
##  Brunswick      :  387   3rd Qu.: 4.000                      3rd Qu.:14.00
##  Essendon       :  361   Max.   :16.000                      Max.   :48.10
##  (Other)        :24422
##     Bedroom          Bathroom          Car             Landsize
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :     1.0
##  1st Qu.: 3.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:   494.0
##  Median : 3.000   Median :2.000   Median : 2.000   Median :   557.0
##  Mean   : 3.035   Mean   :1.688   Mean   : 1.787   Mean   :   620.8
##  3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.:   592.0
##  Max.   :20.000   Max.   :9.000   Max.   :18.000   Max.   :433014.0
##
##   BuildingArea       YearBuilt    CouncilArea         Regionname
##  Min.   :    0.0   Min.   :1196   Length:27244       Length:27244
##  1st Qu.:  156.0   1st Qu.:1970   Class :character   Class :character
##  Median :  160.3   Median :1970   Mode  :character   Mode  :character
##  Mean   :  158.9   Mean   :1968
##  3rd Qu.:  160.3   3rd Qu.:1970
##  Max.   :44515.0   Max.   :2019
##
##  Propertycount   ParkingArea           Price
##  Min.   :   83   Length:27244       Min.   :   85000
##  1st Qu.: 4294   Class :character   1st Qu.:  635000
##  Median : 6567   Mode  :character   Median :  870000
##  Mean   : 7567                      Mean   : 1050210
##  3rd Qu.:10412                      3rd Qu.: 1295000
##  Max.   :21650                      Max.   :11200000
##
```

Based on the boxplot and summary; Car, Landsize, BuildingArea, Yearbuilt, Property count and price are the columns that have significant outliers, that can affect our analysis and cause distortions. They can be removed as follows

```r
remove_outliers_iqr <- function(df, cols) {
  keep <- rep(TRUE, nrow(df))

  for (col in cols) {
    if (!is.numeric(df[[col]])) {
      warning(paste("Skipping non-numeric column:", col))
      next
    }

    Q1 <- quantile(df[[col]], 0.25, na.rm = TRUE)
    Q3 <- quantile(df[[col]], 0.75, na.rm = TRUE)
    IQR_val <- Q3 - Q1
    lower <- Q1 - 1.5 * IQR_val
    upper <- Q3 + 1.5 * IQR_val

    keep <- keep & df[[col]] >= lower & df[[col]] <= upper
  }
```

```
  return(df[keep, ])
}

cols_to_trim <- c("Car", "Landsize", "BuildingArea", "YearBuilt", "Propertycount", "Price")
df <- remove_outliers_iqr(df, cols_to_trim)
```

2 Exploratory Data Analysis

With the data cleaned, we can focus on exploratory data analysis, creating visualizations and exploring the data.

3 Research Questions

3.1 How does the location (suburb and distance to CBD) affect housing prices?

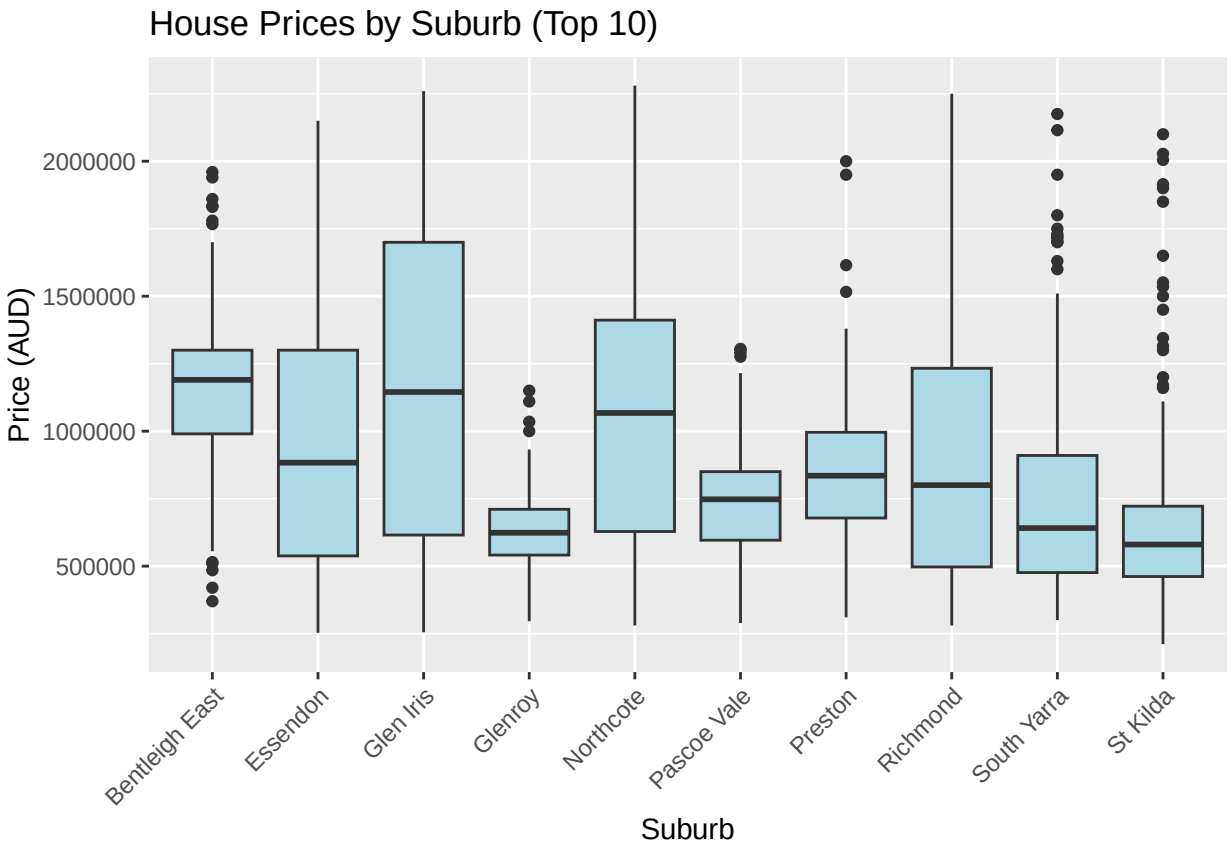
Based on the preliminary exploratory data analysis, an initial regression model (model_region) was developed to assess how location, specifically Distance to the Central Business District (CBD) and Regionname , influenced housing prices. This baseline model revealed that Distance had a significant negative effect on price, and certain regions (e.g., Southern, Western) exhibited substantial variation. However, the model explained only 18% of the variance in price (adjusted $R^2 = 0.1805$), and diagnostic checks revealed non-linearity and

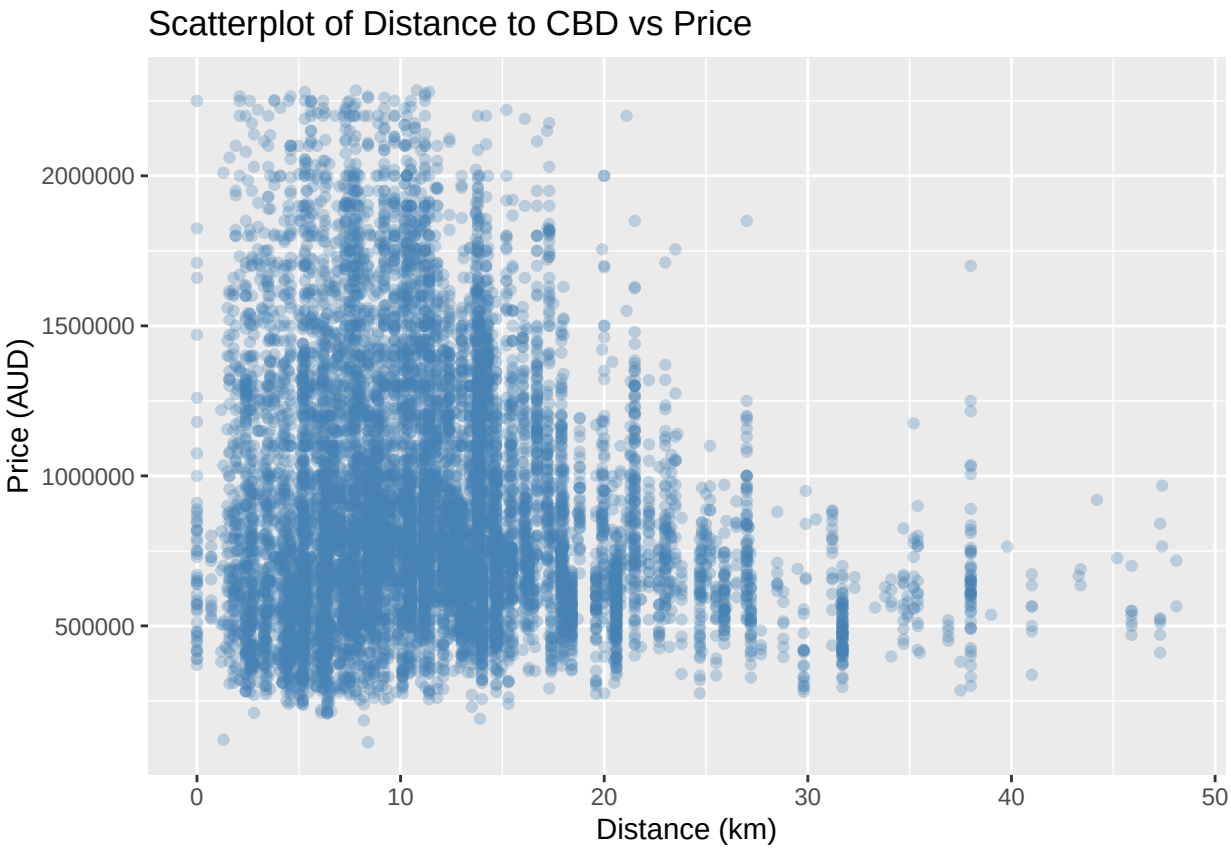
heteroscedasticity. To improve linearity, a second model (model_logdist) was introduced, applying a log transformation to Distance while retaining Regionname. Although this modified the coefficient interpretation, it did not improve overall performance (adjusted $R^2 = 0.1742$), and both AIC and BIC remained high relative to the baseline. These results suggested that while location mattered, it alone was insufficient to explain housing price variation in Melbourne.

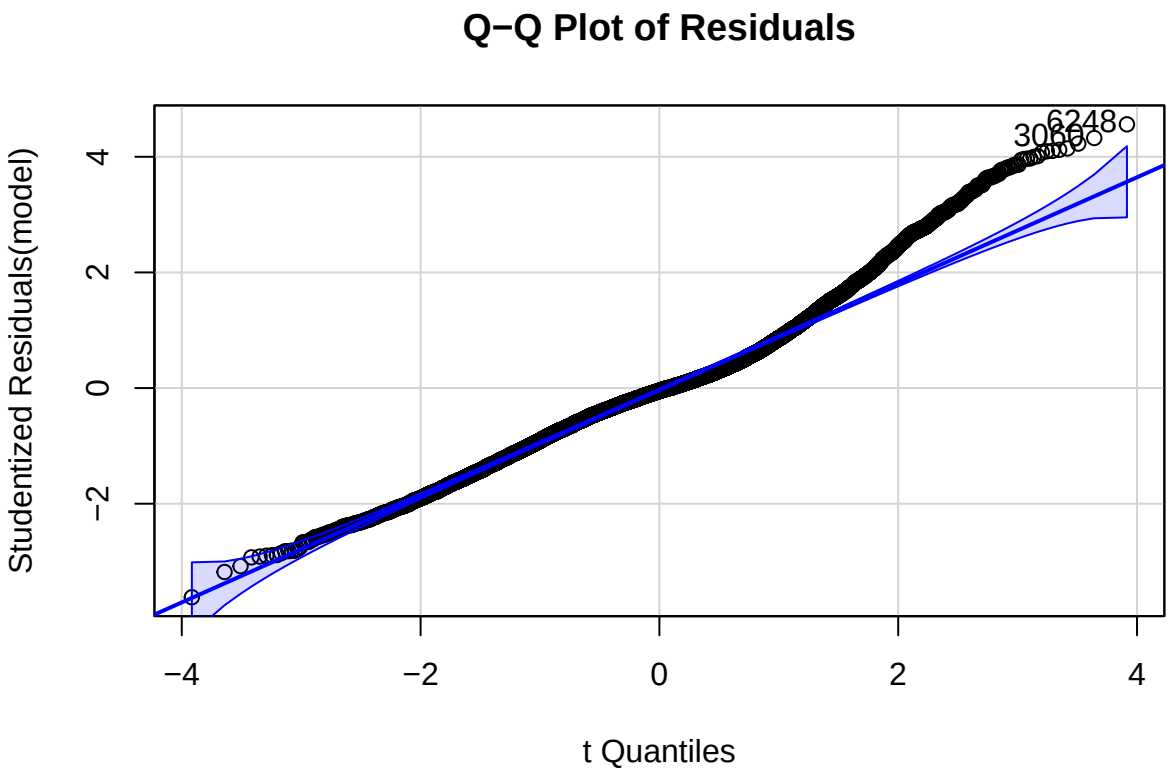
This led to the development of a more comprehensive model, lastmodel_distance, which added key property-level features alongside Distance and Regionname. These included Rooms, Bathroom, Landsize, BuildingArea, YearBuilt, Type, and Car, all of which were identified during EDA (table 2 in the appendix) as being significantly correlated with Price. The final model achieved substantial improvements across all evaluation metrics: adjusted R^2 increased to 0.5677, residual standard error dropped to 421,800, and both AIC and BIC values were significantly lower than in previous models. Additionally, all predictors were statistically significant ($p < 0.001$), with the exception of RegionnameWestern Victoria, and multicollinearity was not a concern (all GVIFs < 2.1). These improvements validated the decision to integrate both geographic and physical property characteristics for more accurate modeling. The summary of the calculations can be found in table 1 in the appendix.

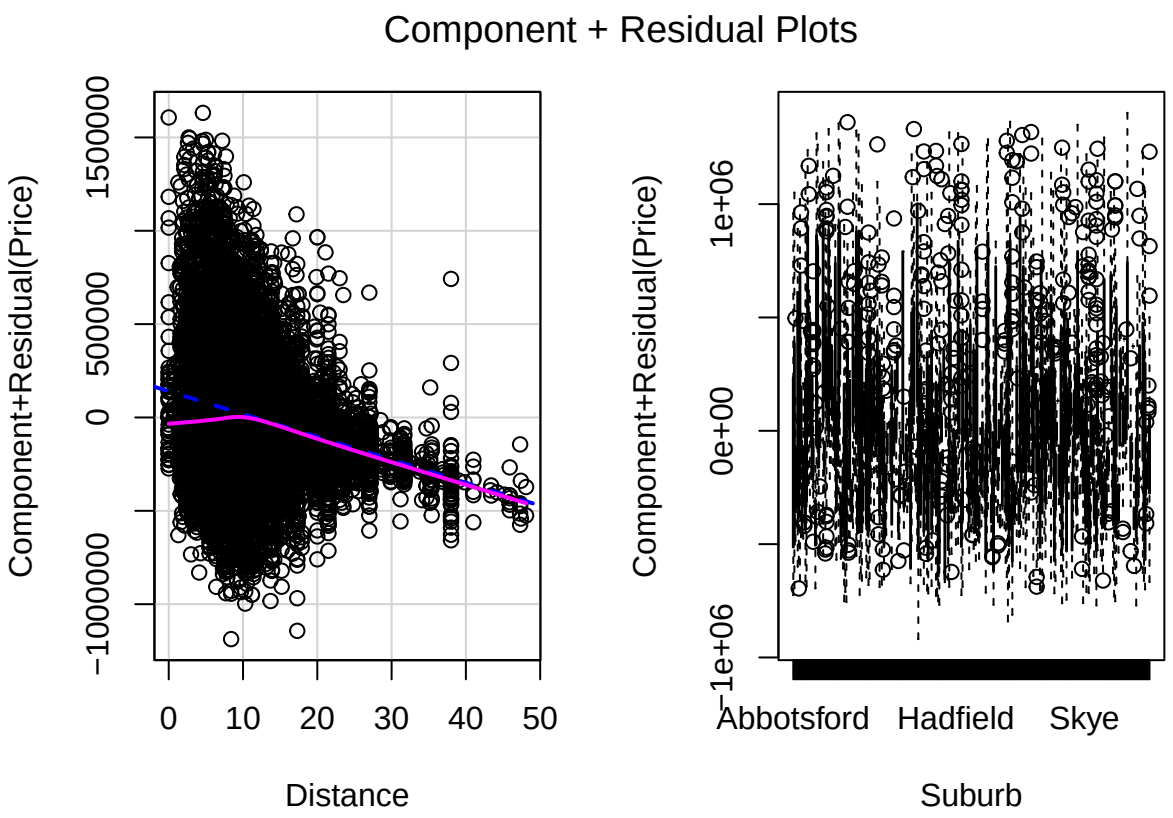
In answering Research Question 1 — How does location (suburb and distance to CBD) affect housing prices? — the results show that while location is indeed a key factor, not all location-based variables contribute equally. Initially, Suburb was considered a predictor due to notable price variation across neighborhoods; however, incorporating Suburb introduced severe multicollinearity, reducing model clarity. Broader location variables like Regionname and Distance proved to be more effective and stable predictors. In the final model, Distance had a strong negative association with price, and some regions (e.g., Southern, Eastern Victoria) consistently saw higher prices than others (e.g., Western, Northern). Still, location alone was not sufficient, the model's improvement came primarily from incorporating structural property features such as Rooms, Bathroom, Landsize, BuildingArea, YearBuilt, Car, and Type, all of which added significant predictive value.

For future research, to further enhance model performance, a second, more focused round of exploratory data analysis could be conducted. This will specifically examine whether key predictors like Distance, Landsize, BuildingArea, and Rooms have nonlinear relationships with housing prices. Using diagnostic plots and smoothing techniques, the goal is to identify where transformations or flexible modeling (e.g., splines or GAMs) can improve predictive accuracy and model fit.

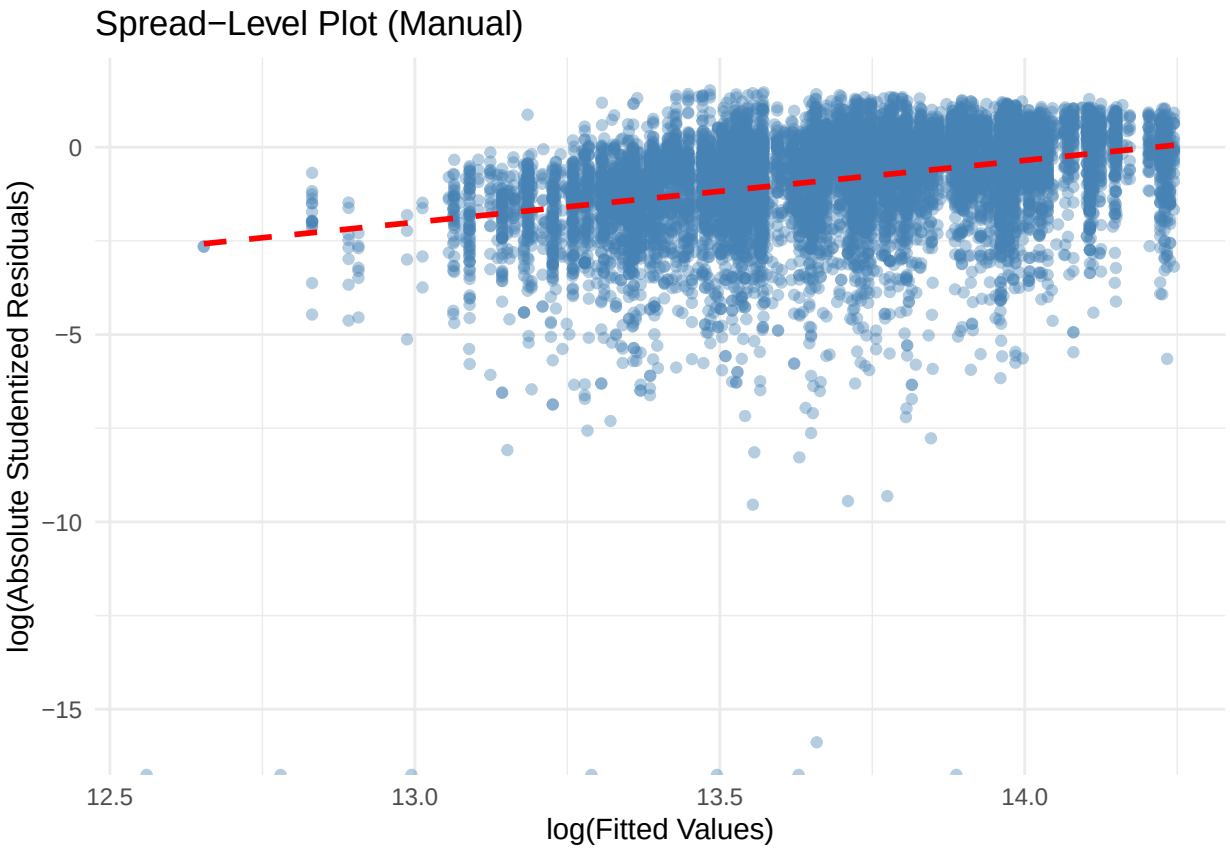

House Prices by Suburb (Top 10)
Price (AUD)
2000000
1500000
1000000
500000
Bentleigh East
Essendon
Glen Iris
Glenroy
Northcote
Pascoe Vale
Preston
Richmond
South Yarra
St Kilda
Suburb

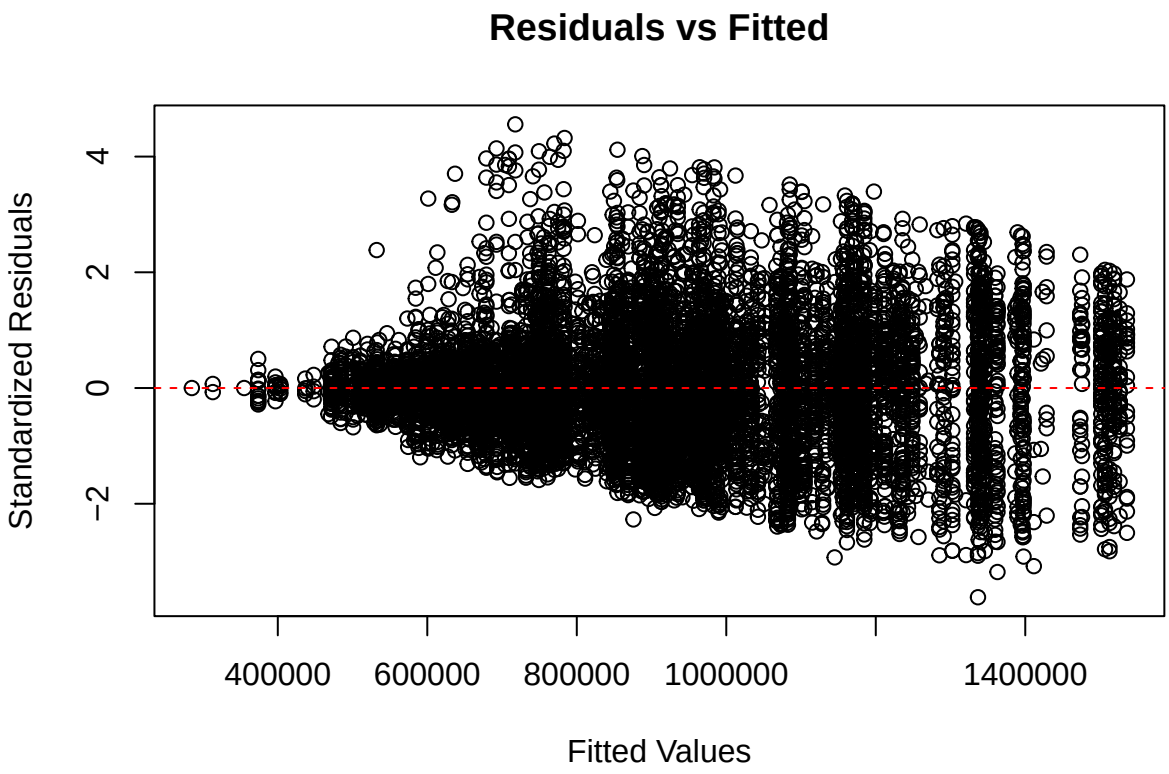

Scatterplot of Distance to CBD vs Price
Price (AUD)
2000000
1500000
1000000
500000
0
10
20
30
40
50
Distance (km)


Q−Q Plot of Residuals
Studentized Residuals(model)
t Quantiles
6248
−4
−2
0
2
4




```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 24 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 17 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


Spread−Level Plot (Manual)
log(Absolute Studentized Residuals)
0
−5
−10
−15
12.5
13.0
13.5
14.0
log(Fitted Values)




```
## Warning in cor(numeric_clean, use = "complete.obs"): the standard deviation is
## zero
```

3.2 Is there a significant difference in housing prices based on number of rooms and property type?

To examine whether house prices in Melbourne differ significantly based on the number of rooms and property type, we performed two separate one-way ANOVAs. This method is appropriate because both predictors—Rooms and Type—are categorical variables, and our goal was to compare the mean house prices across multiple independent groups.

3.2.1 Effect of Room Count on Price

A one-way ANOVA was conducted to assess whether average house prices varied by the number of rooms. The test yielded a statistically significant result:

$F(11, 27{,}235) = 707.3$, $p < .001$, indicating that at least one group mean differs from the others.

Boxplot visualizations supported this finding by showing a consistent increase in price as the number of rooms increased. In particular, properties with six or more rooms exhibited both higher average prices and greater price variability. These patterns likely reflect the inclusion of luxury or high-end homes in these larger room categories.

3.2.2 Effect of Property Type on Price

We also conducted a one-way ANOVA to examine price differences among property types (house, townhouse, unit). This test also revealed a highly significant result:

$F(2, 27{,}244) = 2152$, $p < .001$, suggesting that mean prices differ significantly by property type.

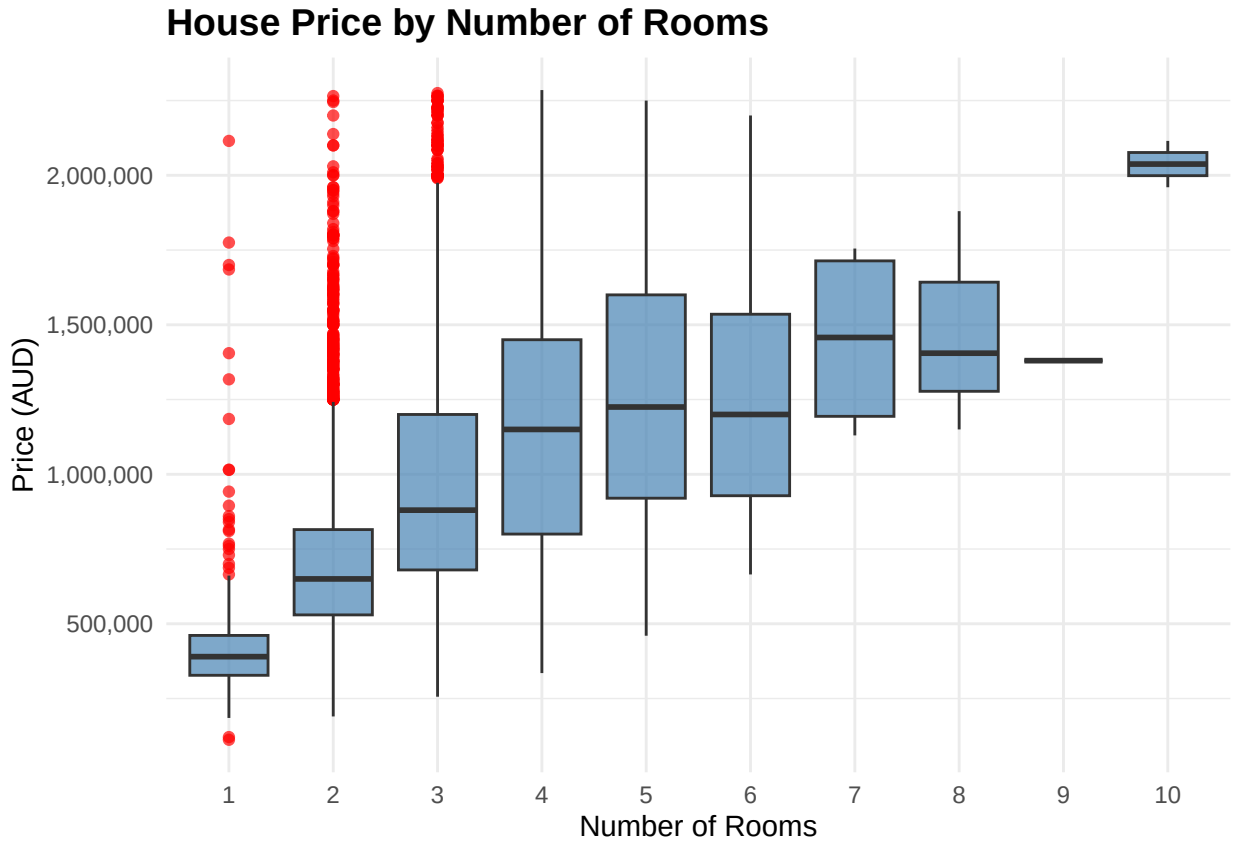
The boxplots showed that:

- Houses had the highest average prices and widest range,
- Townhouses followed with moderate pricing, and
- Units were the least expensive.

To identify specific group differences, we performed a Tukey HSD post-hoc test, which confirmed that all pairwise comparisons were statistically significant ($p < .001$). The mean differences were:

- Townhouse vs. House: -$272,641
- Unit vs. House: -$575,775
- Unit vs. Townhouse: -$303,134

The results demonstrate that both room count and property type significantly influence housing prices in Melbourne. These variables are thus important factors to consider in housing market evaluations and predictive modeling of real estate prices.



Tukey multiple comparisons of means 95% family-wise confidence level

Fit: aov(formula = Price ~ Type, data = df)

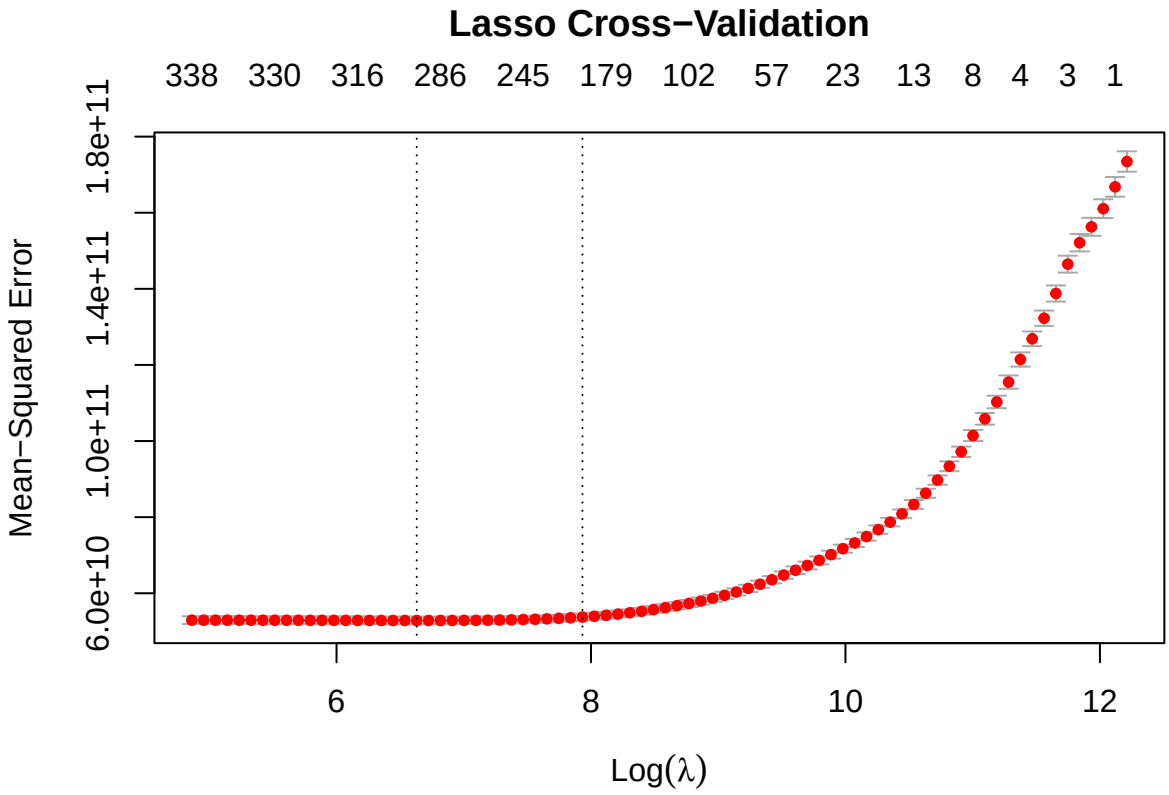
$Type diff lwr upr p adj t-h -108178.7 -134096.3 -82261.04 0 u-h -411043.5 -430178.2 -391908.85 0 u-t -302864.9 -331128.1 -274601.64 0

3.3 Can we accurately predict house prices using a regression model?

To accurately predict house prices based on the available data, selecting the best predictors is very important to get a food fit on the data. The dataset picked has a complex set of features with both numerical and categorical columns. To pick the best predictors among numerical columns, a correlation coefficient matrix heatmap was used, as shown in Figure 7 in Appendix. An interesting fact that can be observed from the plot is that the Price is highly related with Year Built, Bathrooms, Rooms and Bedrooms. However, it does not show a strong relation with Land size or Building area.

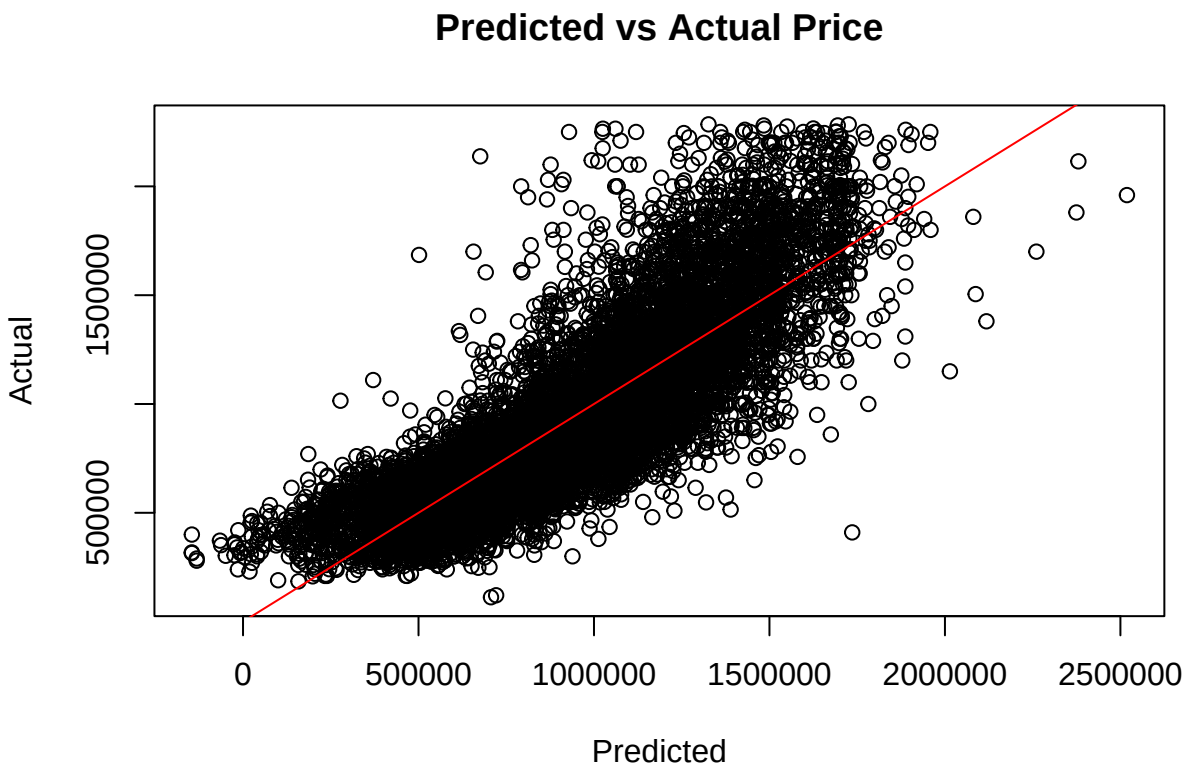
Among Categorical columns, we created dummies for columns that did not have too many categories. This includes Type, Region name and parking area. We also detected and deleted outliers, so that they did not mess with the training process. Followed by this, we trained a LASSO regression model to complete both feature selection and get a good fit simultaneously. The R2 score of the model increased from 0.632 to 0.66 with a mean absolute error being about 235,000. This is a significant update from the base model submitted with module 4.

The following figure shows the actual vs predicted values.




```
## MAE: 166307.1
## R-squared: 0.7112
```

```
plot(preds, y, main = "Predicted vs Actual Price", xlab = "Predicted", ylab = "Actual")
abline(0, 1, col = "red")
```



3.4 Is there a significant difference in median house prices between different regions in Melbourne?

To evaluate whether house prices differ significantly across Melbourne's regions, we conducted a one-way ANOVA. This statistical test was appropriate because Regionname is a categorical variable, and our goal was to compare average prices across multiple independent groups (regions).

ANOVA Results

The one-way ANOVA revealed a statistically significant effect of region on house prices:

$F(7, 27{,}239) = 695.2$, $p < .001$

This result indicates that mean house prices are not equal across the different regions of Melbourne.

Visualization and Interpretation Boxplot analysis reinforced the statistical finding. Regions such as Southern, Eastern, and South-Eastern exhibited higher average house prices and greater price variability. In contrast, outer regions like Western Victoria and Eastern Victoria showed lower median prices and narrower spreads.

This pattern suggests that central and southern regions, likely due to better infrastructure, amenities, and proximity to business districts, command higher property values and attract premium housing.

Post-Hoc Analysis (Tukey HSD) To pinpoint which regional differences were statistically significant, we conducted a Tukey HSD post-hoc test. Most pairwise comparisons showed significant differences in average price ($p < .05$). Key examples include:

Southern vs. Western Victoria: +$963,321

Southern vs. Eastern Victoria: +$681,600

Northern vs. Eastern: -$247,294

Only a few comparisons—such as Northern Victoria vs. Eastern Victoria—did not yield significant differences, indicating some similarity in average pricing across these more distant regions.

The analysis confirms that location plays a critical role in determining house prices in Melbourne. Properties in central and southern areas tend to be significantly more expensive, highlighting the strong influence of region as a driver of real estate value.

3.5 Is there a significant relationship between property type (e.g., house, unit, townhouse) and the selling price?

One-Way ANOVA on Log-Transformed Price

To statistically test whether property type affects price, we conducted a one-way ANOVA.

After adjusting for covariates, both townhouses and units still had significantly lower selling prices compared to houses. Other predictors like Rooms and Distance were also significant, validating that property type independently influences price.

A unit is about 92% less likely to be in the high-price category than a house. Townhouses are about 57% less likely. Property type is a strong classification predictor.

Even under regularization, property type remains among the most influential features affecting price, confirming its robustness in prediction. This unsupervised learning technique segmented the housing market into 3 clusters, visually confirming distinct groupings related to price and size.

The analysis strongly supports the conclusion that property type has a significant and consistent influence on housing prices in the Melbourne real estate market. This conclusion is reinforced through multiple analytical techniques:

- Descriptive statistics and boxplots show that houses generally have higher median prices than townhouses and units.
- A one-way ANOVA confirmed that these differences are statistically significant ($p < .001$), indicating that the average log-transformed prices vary by property type.
- The Generalized Linear Model (GLM) further validated that even after controlling for other influential variables like number of rooms, bathrooms, and distance to the CBD, property type remains a significant independent predictor. Units and townhouses consistently sell for less than houses.
- Logistic regression added practical insight, showing that units are about 92
- The LASSO regression model, which penalizes less influential predictors, still retained property type as a major contributor to price prediction, highlighting its robustness in modeling.
- Finally, K-means clustering identified distinct market segments aligned with property value and size. This supports strategic segmentation and targeted marketing or development strategies based on property characteristics.

4 Conclusion

There is clear and consistent evidence that property type significantly impacts selling price in the Melbourne housing market. Houses command the highest prices, while units and townhouses fall into lower price categories. These findings are actionable for stakeholders such as real estate investors, policymakers, and developers, as they emphasize the importance of property classification when assessing market value or planning housing strategies.

5 Recommendation

Investors and developers should prioritize house-type properties when aiming for high-value returns, while units and townhouses may be more suitable for affordable housing initiatives or first-time buyers. Future research could expand this model by incorporating additional variables like location-specific trends, property condition, and market seasonality.

6 Justification for methods used

1. **Multiple linear regression:** It is used to analyze how location affects housing prices because it allows us to assess the impact of both numeric (Distance) and categorical (Suburb) variables on a continuous outcome (Price). MLR helps isolate the effect of each variable while controlling for others, making it ideal for understanding complex, location-driven variation. It also enables us to perform essential diagnostic checks to evaluate assumptions like linearity, homoscedasticity, and multicollinearity.
2. **One-Way ANOVA (Analysis of Variance):** ANOVA is appropriate because:
    - The independent variable Type is categorical (with three groups: house (h), unit (u), townhouse (t)).
    - The dependent variable Price is continuous.
    - ANOVA determines whether the mean prices are significantly different across these types.
    - It is also appropriate for research question 5 because Regionname is a categorical variable with multiple levels, and we aimed to test whether average housing prices differ across regions. The Tukey HSD test followed to determine which regional comparisons were statistically significant.

3. **Tukey HSD Post-Hoc Test:** If ANOVA is significant, post-hoc comparison is necessary to identify which groups are significantly different.

Two one-way ANOVAs: Its appropriate for research question 2 because both room count and property type are categorical variables, and our goal was to compare mean prices across multiple groups. ANOVA is appropriate for detecting whether group means differ significantly, and Tukey HSD was used to identify specific pairwise differences.

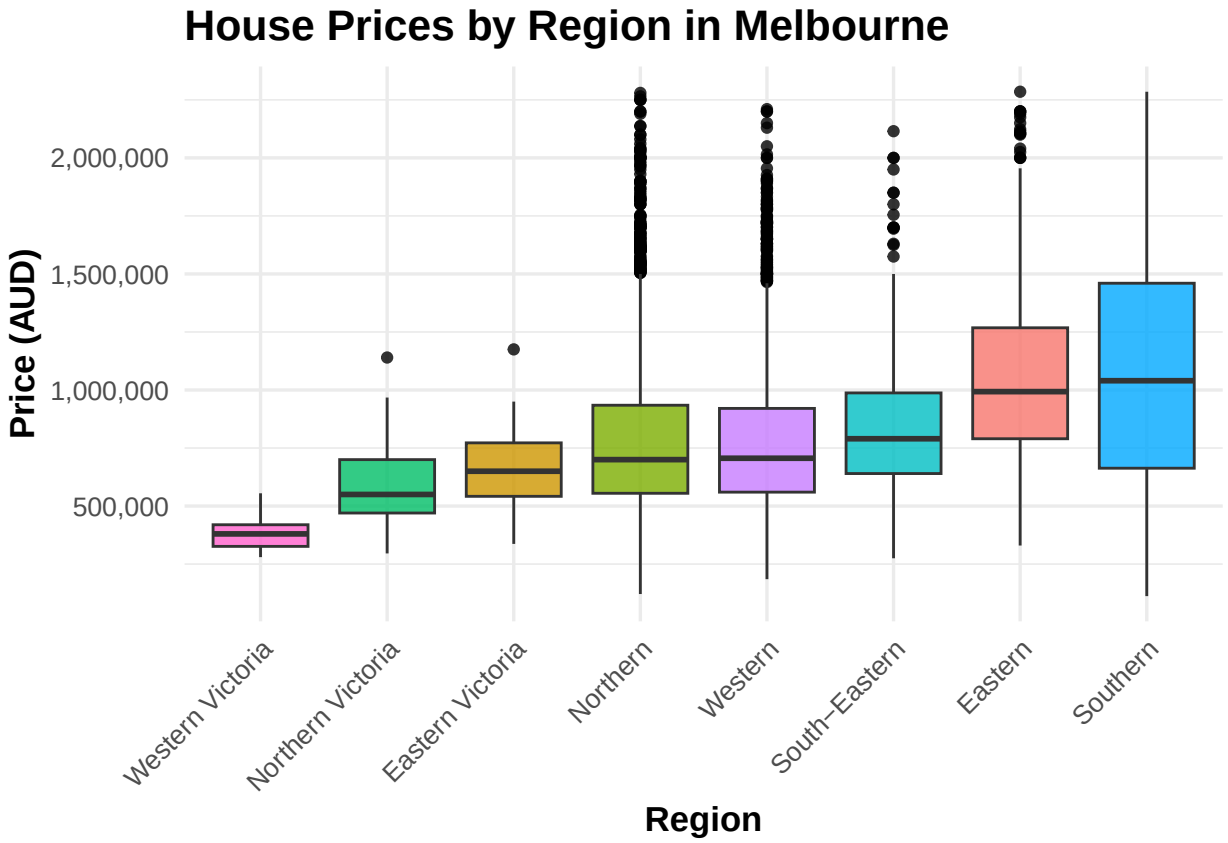


Figure 1: House Prices by Region in Melbourne

```
# One-way ANOVA: Does region affect price?
anova_region <- aov(Price ~ Regionname, data = df)

# Tukey HSD: Post-hoc test to compare regions
TukeyHSD(anova_region)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Price ~ Regionname, data = df)
##
## $Regionname
##                                       diff          lwr           upr
## Eastern Victoria-Eastern          -383332.65 -547057.0236 -219608.28393
## Northern-Eastern                  -257107.23 -295079.0085 -219135.45734
## Northern Victoria-Eastern         -455703.02 -634510.5044 -276895.54382
## South-Eastern-Eastern             -206589.58 -262412.5732 -150766.59535
```

```
## Southern-Eastern                     46520.50    9367.2277   83673.78196
## Western-Eastern                    -264513.72 -303863.9016 -225163.54326
## Western Victoria-Eastern           -661638.21 -891067.7147 -432208.70386
## Northern-Eastern Victoria           126225.42  -35949.3116  288400.15332
## Northern Victoria-Eastern Victoria  -72370.37 -310762.7006  166021.95988
## South-Eastern-Eastern Victoria      176743.07    9485.7999  344000.33898
## Southern-Eastern Victoria           429853.16  267868.1157  591838.20145
## Western-Eastern Victoria            118818.93  -43684.0678  281321.93041
## Western Victoria-Eastern Victoria  -278305.56 -556687.8217      76.71054
## Northern Victoria-Northern         -198595.79 -375985.4459  -21206.13652
## South-Eastern-Northern               50517.65    -582.0644  101617.36169
## Southern-Northern                   303627.74  274041.5825  333213.89301
## Western-Northern                     -7406.49  -39708.3654   24895.38632
## Western Victoria-Northern          -404530.98 -632857.2179 -176204.73492
## South-Eastern-Northern Victoria     249113.44   67065.5342  431161.34547
## Southern-Northern Victoria          502223.53  325007.2773  679439.78061
## Western-Northern Victoria           191189.30   13499.4864  368879.11696
## Western Victoria-Northern Victoria -205935.19 -493447.1533   81576.78294
## Southern-South-Eastern              253110.09  202615.6247  303604.55352
## Western-South-Eastern               -57924.14 -110056.2927   -5791.98362
## Western Victoria-South-Eastern     -455048.63 -687012.4647 -223084.78538
## Western-Southern                   -311034.23 -342369.8531 -279698.60145
## Western Victoria-Southern          -708158.71 -936350.2626 -479967.16563
## Western Victoria-Western           -397124.49 -625684.0052 -168564.96857
##                                        p adj
## Eastern Victoria-Eastern           0.0000000
## Northern-Eastern                   0.0000000
## Northern Victoria-Eastern          0.0000000
## South-Eastern-Eastern              0.0000000
## Southern-Eastern                   0.0036968
## Western-Eastern                    0.0000000
## Western Victoria-Eastern           0.0000000
## Northern-Eastern Victoria          0.2616748
## Northern Victoria-Eastern Victoria 0.9842089
## South-Eastern-Eastern Victoria     0.0296439
## Southern-Eastern Victoria          0.0000000
## Western-Eastern Victoria           0.3416565
## Western Victoria-Eastern Victoria  0.0501224
## Northern Victoria-Northern         0.0158984
## South-Eastern-Northern             0.0552771
## Southern-Northern                  0.0000000
## Western-Northern                   0.9971459
## Western Victoria-Northern          0.0000022
## South-Eastern-Northern Victoria    0.0008830
## Southern-Northern Victoria         0.0000000
## Western-Northern Victoria          0.0246091
## Western Victoria-Northern Victoria 0.3693016
## Southern-South-Eastern             0.0000000
## Western-South-Eastern              0.0173322
## Western Victoria-South-Eastern     0.0000001
## Western-Southern                   0.0000000
## Western Victoria-Southern          0.0000000
## Western Victoria-Western           0.0000039
```

7 References

Bluman, A. G. (2012). Elementary statistics (10th ed.). McGraw-Hill Education.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.

Kabacoff, R. I. (2015). R in action: Data analysis and graphics with R (2nd ed.). Manning Publications.

Provost, J.-S. (2025). Regression diagnostics with R [Video]. Canvas. https://northeastern.instructure.com/courses/221013/assignments/2683591

Provost, J.-S. (2025). Chi square and ANOVA [Video]. Canvas. https://northeastern.instructure.com/courses/221013/assignments/2683593

Provost, J.-S. (2025). Logistic regression and classification [Video]. Canvas@Northeastern. https://northeastern.instructure.com/courses/221013/assignments/2683596

Provost, J.-S. (2025). Module 3 – Recap: GLM & logistic regression [PDF]. Department of College of Professional Studies, Northeastern University.