# Real Estate Investment Analysis: Identifying Optimal Suburbs for Apartment Investment in Victoria, Australia

## DATA SUMMARY:

The three provided data sets were used in the study to determine which Victoria, Australia, suburb would be the greatest place to invest in real estate:

- The data in the Apartment_prices.csv shows the median price of houses in various suburbs in 2023.
- Historical_demographics.csv contains data from the previous year's priority growth areas, median income, unemployment rate, and population growth rate.
- Data on the unemployment rate, population growth rate, median income, and priority growth area for the upcoming year are provided by projected_demographics.csv.

```
> summary(merged_data)
 Suburb_name        Median_price_2023  Historical_population_growth Historical_median_income
 Length:458         Length:458         Min.   :2.050               Min.   : 67160
 Class :character   Class :character   1st Qu.:3.812               1st Qu.:102134
 Mode  :character   Mode  :character   Median :4.415               Median :134817
                                       Mean   :4.576               Mean   :134322
                                       3rd Qu.:5.197               3rd Qu.:164900
                                       Max.   :8.690               Max.   :203191
                                                                   NA's   :1
 Historical_unemployment_rate Historical_priority_growth_area Projected_population_growth Projected_median_income
 Min.   :-3.050               Min.   :0.0000                  Min.   :2.116              Min.   : 69167
 1st Qu.: 3.197               1st Qu.:0.0000                  1st Qu.:3.925              1st Qu.:106170
 Median : 5.320               Median :0.0000                  Median :4.534              Median :139586
 Mean   : 5.551               Mean   :0.1878                  Mean   :4.717              Mean   :139478
 3rd Qu.: 7.728               3rd Qu.:0.0000                  3rd Qu.:5.326              3rd Qu.:171284
 Max.   :17.330               Max.   :1.0000                  Max.   :8.857              Max.   :209992

 Projected_unemployment_rate Projected_priority_growth_area
 Min.   :-3.133              Min.   :0.0000
 1st Qu.: 3.172              1st Qu.:0.0000
 Median : 5.285              Median :0.0000
 Mean   : 5.553              Mean   :0.1878
 3rd Qu.: 7.787              3rd Qu.:0.0000
 Max.   :17.080              Max.   :1.0000

>
```

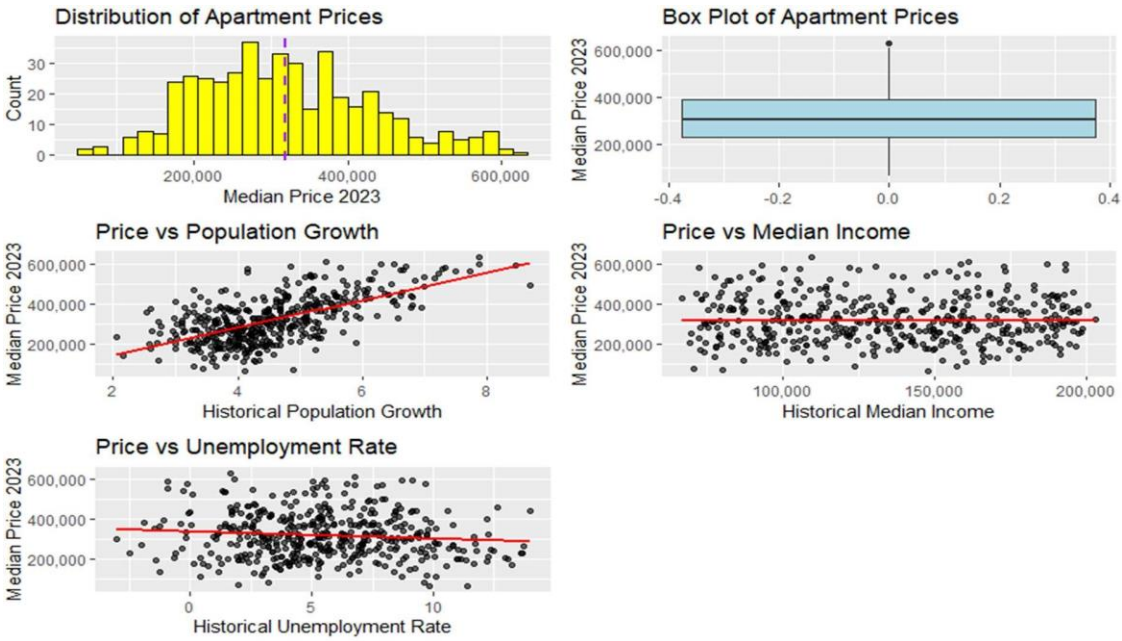
The three data sets were first combined on "Suburb_name," after which the data was cleaned and processed.

- Historical_median_income had one missing value, and Median_price_2023 had one incorrect value, which was changed to the column mean.

```
> na_summary <- sapply(merged_data, function(x) sum(is.na(x)))
> print(na_summary)
                  Suburb_name              Median_price_2023    Historical_population_growth
                            0                             0                               0
      Historical_median_income    Historical_unemployment_rate Historical_priority_growth_area
                            0                             0                               0
   Projected_population_growth         Projected_median_income     Projected_unemployment_rate
                            0                             0                               0
Projected_priority_growth_area
                            0
```

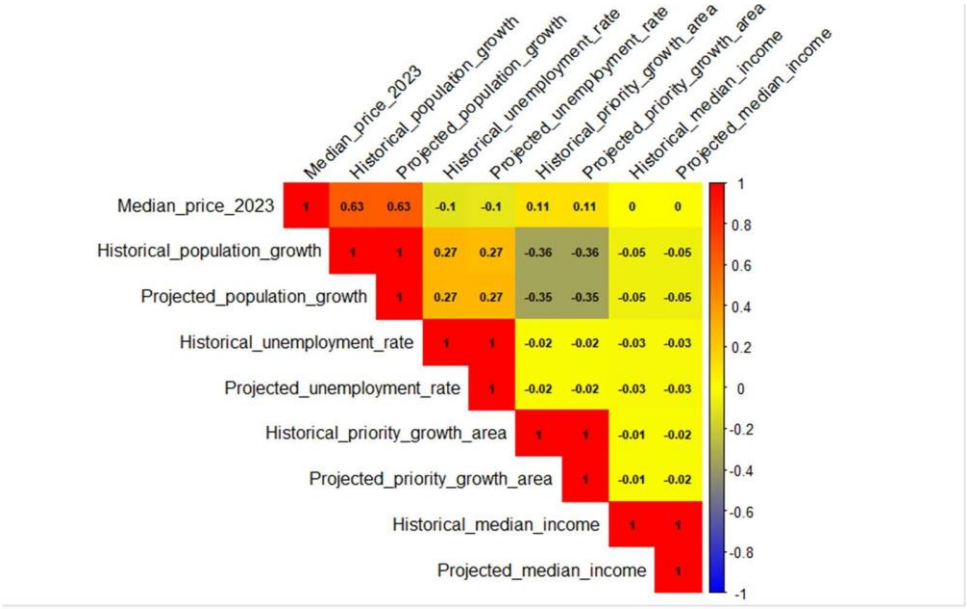
- Outliers were identified using boxplots and handled by excluding the outliers using IQR method.

After cleaning the data, it is found that the majority of the suburbs have median prices between 200k and 400k, furthermore, the median price of the apartments has no correlation between Historical median income and very slight negative correlation (-0.1) with Historical unemployment rate.




## MODEL ESTIMATION:

Using the cleaned data, the correlation between the variables was analysed and it was noted that 'Median_price_2023' had the highest positive correlation of 0.63 between 'Historical_population_growth' and 'Projected_population_growth' among other variables.



A linear regression model was selected for its simplicity and interpretability. The independent variables selected were:

- Historical population growth
- Historical unemployment rate
- Historical priority growth area

Historical median income has been excluded from the model as it has no correlation with median prices.

These variables were selected because these are some of the key factors which influence the ROI of an apartment.

The formula for the regression model is :

$$Median\ price = \beta0 + \beta1 * \text{Historical\_population\_growth} + \beta2 * \text{Historical\_unemployment\_rate} + \beta3 * \text{Historical\_priority\_growth\_area}$$

Where:

β0 = -67585 (Intercept)

β1 = 93795

β2 = -11063

β3 = 122622

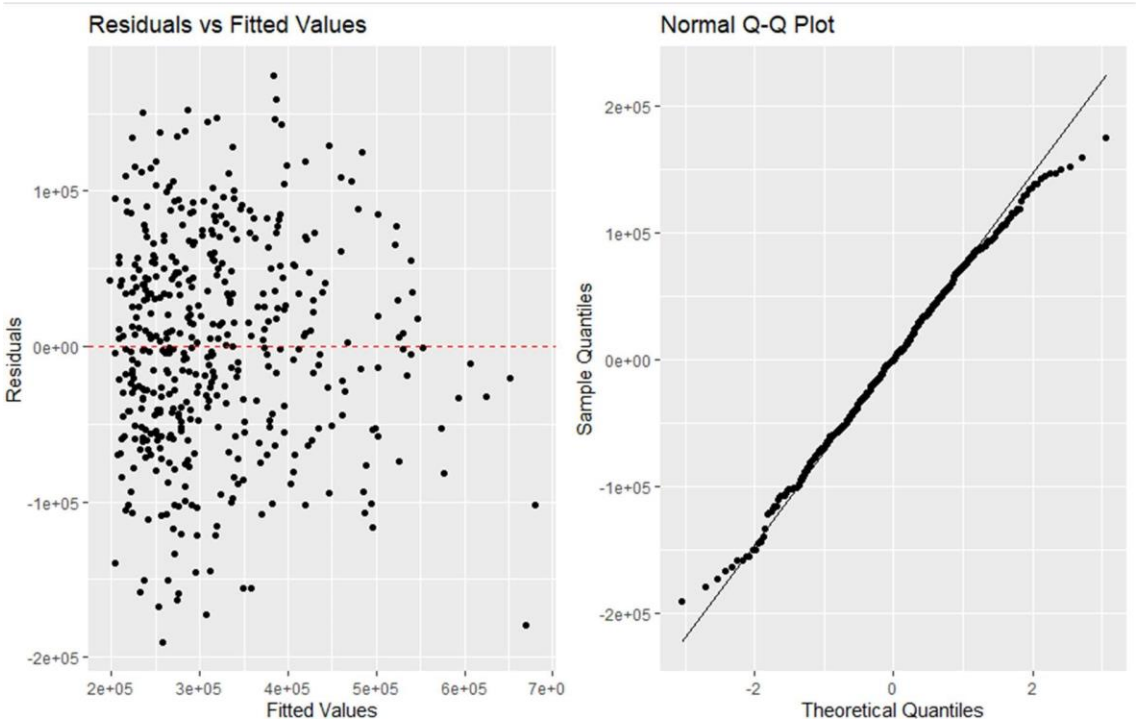
## MODEL INTERPRETATION:

Based on the model summary, it is evident that Historical_population_growth, Historical_unemployment_rate and Historical_priority_growth_area are significant independent variables.

```
> # Print model summary
> summary(model)

Call:
lm(formula = Median_price_2023 ~ Historical_population_growth +
    Historical_unemployment_rate + Historical_priority_growth_area,
    data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max
-190941  -49229    -779   49661  174624

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -67585      16062  -4.208 3.12e-05 ***
Historical_population_growth    93795       3442  27.247  < 2e-16 ***
Historical_unemployment_rate   -11603       1053 -11.014  < 2e-16 ***
Historical_priority_growth_area 122622      9008  13.613  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69200 on 444 degrees of freedom
Multiple R-squared:  0.6342,    Adjusted R-squared:  0.6317
F-statistic: 256.6 on 3 and 444 DF,  p-value: < 2.2e-16

> # Assuming you've run the model, interpret the results
> coef_summary <- summary(model)$coefficients
> print(coef_summary)
                               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)                   -67585.15  16061.941  -4.207783 3.122016e-05
Historical_population_growth   93795.40   3442.395  27.247134 8.287280e-97
Historical_unemployment_rate  -11602.64   1053.421 -11.014248 4.163475e-25
Historical_priority_growth_area 122621.62  9007.828  13.612784 1.631469e-35
> |
```

The suburbs with higher historical rates of population growth and median income are expected to have higher median apartment prices, according to the positive coefficients for both variables. On the other hand, the negative correlation for unemployment rate suggests that lower apartment prices are related to higher unemployment rates.

The multiple R- squared value 0.6342 suggests that this model can interpret 63.42% changes in the median prices based on the used independent variables.



- The lack of a clear pattern suggests that the linear model adequately explains the relationship between the independent variables and the dependent variable. The residuals are evenly scattered along the horizontal axis suggesting that the variance of the residuals is consistent across all levels of fitted values.

- The Q—Q plot depicts that majority of the residual follow the $45^0$ degree line suggesting a near normal distribution, validating the assumption required for linear regression.

## RECOMMENDATIONS:

This model is used to find the predicted median price of the apartments in the next year, which are:

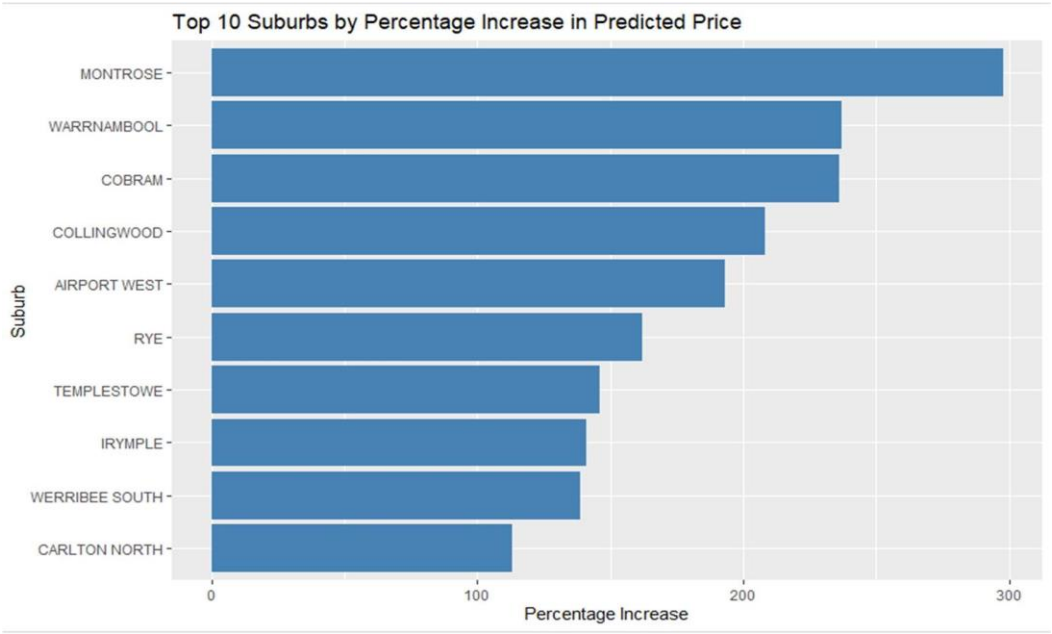
```
> print(head(predictions, 10))
              Suburb Predicted_Price
164      SANDRINGHAM        699687.5
85           SEAFORD        686373.2
190        ASHBURTON        683128.7
200    BACCHUS MARSH        640635.2
93           MACLEOD        631011.7
250           HALLAM        613252.1
196        SWAN HILL        590133.7
16  CRANBOURNE WEST         588791.0
45     KANGAROO FLAT        586684.0
75            BALWYN        576932.8
>
```


Then the percentage increase in the median prices is calculate using the formula

$$percentage\ increase = \frac{\qquad\qquad\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad} * 100$$

As an investor, the company should invest in the suburb which had the highest percentage increase in the median price

```
> # Display top 10 suburbs by percentage increase
> print(top_suburbs)
       Suburb_name Percentage_Increase
105       MONTROSE            297.6612
213    WARRNAMBOOL            237.1826
413         COBRAM            236.0978
410    COLLINGWOOD            208.3508
308   AIRPORT WEST            193.2448
256            RYE            161.8279
197    TEMPLESTOWE            145.9734
399        IRYMPLE            140.8023
434 WERRIBEE SOUTH            138.5129
315  CARLTON NORTH            113.2275
>
```




Based on these observations, it is evident MONTROSE is the suburb which had the highest percentage increase in the median price of 297.66%. If this trend continues, investing in Montrose will give the highest ROI for the company.