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.

```
Apartment Prices Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Suburb_name        458 non-null    object
 1   Median_price_2023  458 non-null    object
dtypes: object(2)
memory usage: 7.3+ KB
None

Historical Demographic Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 5 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Suburb_name                     458 non-null    object 
 1   Historical_population_growth    458 non-null    float64
 2   Historical_median_income        457 non-null    float64
 3   Historical_unemployment_rate    458 non-null    float64
 4   Historical_priority_growth_area  458 non-null    int64  
dtypes: float64(3), int64(1), object(1)
memory usage: 18.0+ KB
None
```

```
Projected Demographic Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 5 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Suburb_name                     458 non-null    object
 1   Projected_population_growth     458 non-null    float64
 2   Projected_median_income         458 non-null    float64
 3   Projected_unemployment_rate     458 non-null    float64
 4   Projected_priority_growth_area  458 non-null    float64
dtypes: float64(4), object(1)
memory usage: 18.0+ KB
None
```

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

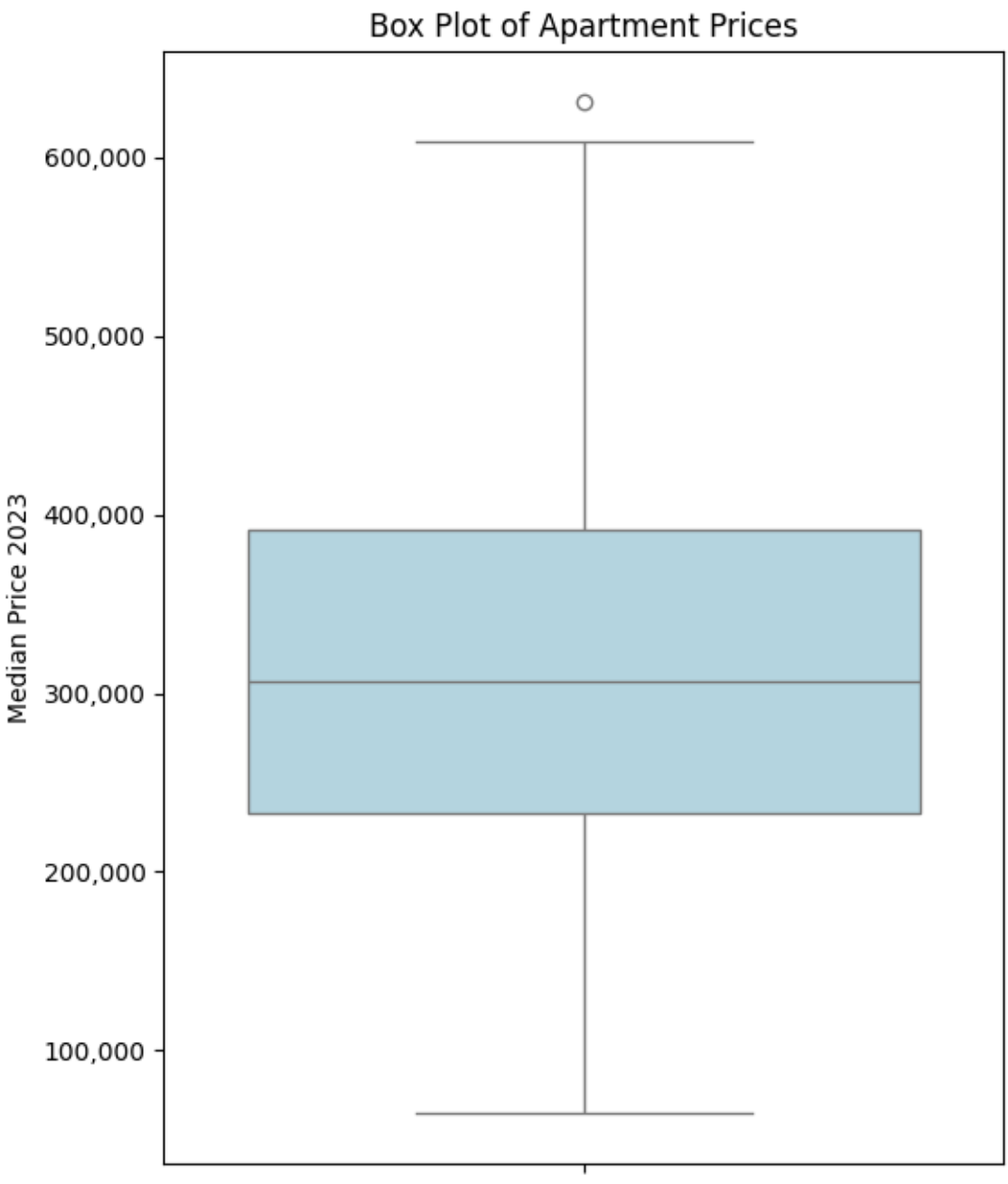
• **Historical_median_income** had one missing value, and Median_price_2023 had one incorrect value, which was changed to the column mean.

```
Summary of merged data after conversion:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Median_price_2023               457 non-null    float64
 1   Historical_population_growth    458 non-null    float64
 2   Historical_median_income        457 non-null    float64
 3   Historical_unemployment_rate    458 non-null    float64
 4   Historical_priority_growth_area 458 non-null    int64
 5   Projected_population_growth     458 non-null    float64
 6   Projected_median_income         458 non-null    float64
 7   Projected_unemployment_rate     458 non-null    float64
 8   Projected_priority_growth_area  458 non-null    float64
dtypes: float64(8), int64(1)
memory usage: 32.3 KB
None
```

• 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.


Distribution of Apartment Prices
Count
0
5
10
15
20
25
30
100,000
200,000
300,000
400,000
500,000
600,000
Median Price 2023


Box Plot of Apartment Prices
Median Price 2023
600,000
500,000
400,000
300,000
200,000
100,000


Price vs Population Growth
Median Price 2023
600,000
500,000
400,000
300,000
200,000
100,000
2
3
4
5
6
7
8
9
Historical Population Growth


Price vs Median Income
Median Price 2023
600,000
500,000
400,000
300,000
200,000
100,000
80,000
100,000
120,000
140,000
160,000
180,000
200,000
Historical Median Income


Price vs Unemployment Rate
Median Price 2023
600,000
500,000
400,000
300,000
200,000
100,000
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
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 :
$Median\ price$ = β0 + β1 ∗ Historical_population_growth + β2
∗ Historical_unemployment_rate + β3 ∗ Historical_priority_growth_area

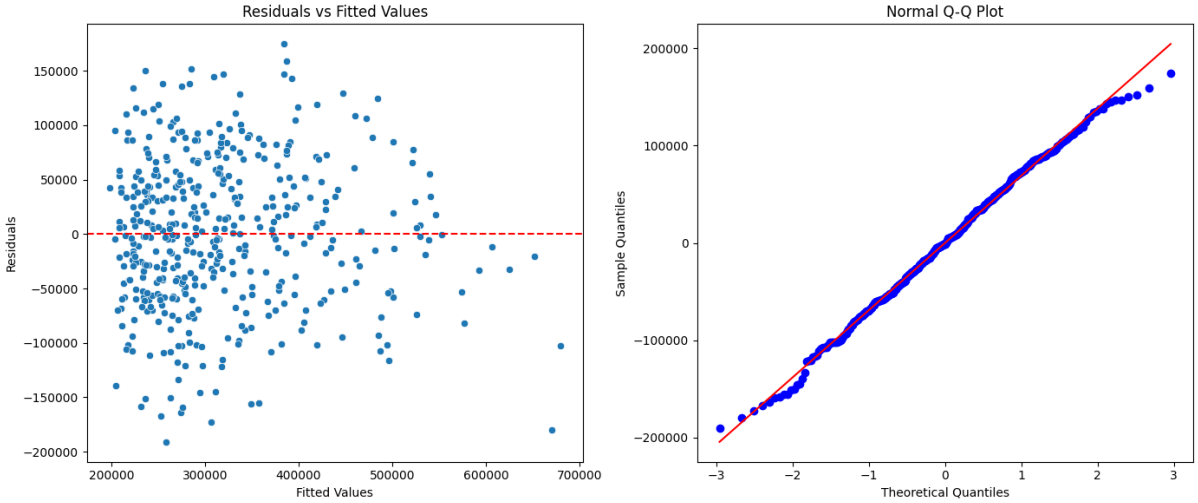
Where:
β0 = -67585 (Intercept)
β1 = 93795
β2 = -11063
β3 = 122622

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.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Median_price_2023   R-squared:                       0.634
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     256.6
Date:                Fri, 24 Jan 2025   Prob (F-statistic):           1.50e-96
Time:                        07:15:06   Log-Likelihood:                -5626.5
No. Observations:                 448   AIC:                         1.126e+04
Df Residuals:                     444   BIC:                         1.128e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    -6.759e+04   1.61e+04     -4.208      0.000   -9.92e+04    -3.6e+04
Historical_population_growth  9.38e+04   3442.395     27.247      0.000    8.7e+04    1.01e+05
Historical_unemployment_rate -1.16e+04   1053.421    -11.014      0.000   -1.37e+04   -9532.326
Historical_priority_growth_area 1.226e+05  9007.828     13.613      0.000    1.05e+05     1.4e+05
==============================================================================
Omnibus:                        2.979   Durbin-Watson:                   2.189
Prob(Omnibus):                  0.225   Jarque-Bera (JB):                2.648
Skew:                          -0.105   Prob(JB):                        0.266
Kurtosis:                       2.688   Cond. No.                         40.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Model Coefficients:
Intercept                   -67585.153711
Historical_population_growth    93795.396014
Historical_unemployment_rate   -11602.635977
Historical_priority_growth_area  122621.622913
dtype: float64
```

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 degree line suggesting a near normal distribution, validating the assumption required for linear regression.

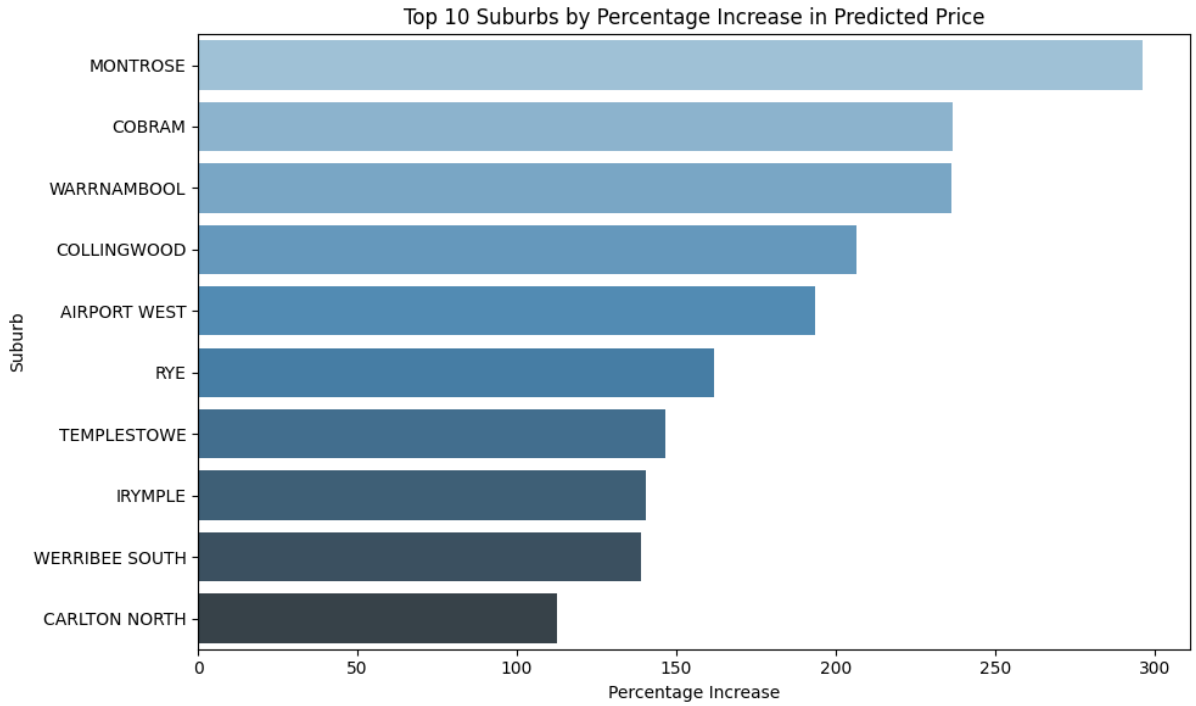
RECOMMENDATIONS:
This model is used to find the predicted median price of the apartments in the next year, which are:

	Suburb	Predicted_Price
170	SANDRINGHAM	699179.155663
88	SEAFORD	686589.891925
206	BACCHUS MARSH	640941.793576
96	MACLEOD	630878.128464
256	HALLAM	613581.087475
202	SWAN HILL	589274.661633
16	CRANBOURNE WEST	588286.585151
46	KANGAROO FLAT	587042.162546
77	BALWYN	576950.897546
177	BRIGHTON	569402.817496

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

$$percentage\ increase = \left(\frac{Predicted\ median\ price - Original\ medianprice}{Original\ median\ price}\right) * 100$$

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



Based on these observations, it is evident **Montrose** is the suburb which had the highest percentage increase in the median price of 296.41%. If this trend continues, investing in Montrose will give the highest ROI for the company.