Assignment 5

NYC Real Estate Analysis using R Language

Tuba Anwar

16 April ,2025

AD 571 A1

Boston University, Metropolitan College

TABLE OF CONTENT

EXECUTIVE SUMMARY

This report analyzes residential real estate sales in the Elmhurst neighborhood of NYC from 2009 to 2023 and provides forecasts through 2025. The primary objective was to evaluate sales trends and identify the key drivers of property value using statistical modeling in R.

Two forecasting models were evaluated—ARIMA and ETS. The ARIMA(0,1,0) model predicted flat future sales but showed poor reliability due to wide and unrealistic confidence intervals. In contrast, the ETS (M,A,M) model better captured seasonality and trends, especially quarterly peaks, and offered a more stable and interpretable forecast. It emerged as the preferred model for neighborhood-level projections, estimating that sales would peak in Q2 and dip in Q2 each year, albeit with some uncertainty due to historical outliers.

To analyze individual property pricing, a multiple linear regression model was developed using factors such as property age, building size (gross square feet), number of units, building type, and sale date. The results revealed that property size, age, and unit count significantly influence prices. Certain building types (e.g., H9 is High rise residential, D1 -institutional building , and RR is railway related properties) were associated with premium pricing, while others (like R9) which is High density residential building were correlated with lower prices. The model explained 27% of the variation in log-transformed sale prices. VIF diagnostics indicated moderate multicollinearity between gross square feet and unit count, but no issues with property age or building type.

Residual analysis identified bargains (properties sold below predicted value) and overpriced sales, providing actionable insights for potential investment opportunities. Visualizations clearly showed how well the model captured market dynamics and flagged pricing anomalies.

INTRODUCTION

This report presents the findings from an in-depth analysis of residential property sales in the Elmhurst neighborhood from 2009 to 2023, with forecasts extending through 2025. The objective was to evaluate historical sales trends and develop predictive models that support strategic pricing and investment decisions.

TASK 1

Based on past patterns of total sales from the year 2009-2023 the future sales of residential properties are determined in Elmhurst neighborhood for next eight quarters that means a forecast for next two years are created as a quarter are 4 in a year so 8 quarter /4 quarter /year =2 years.So,for the year 2024 and 2025 the forecast of residential properties are created.

The total sales in the **figure 1** shows the total residential sales in quarters for year 2009 to 2023.For the **figure 1**,ts function was used from package timeseries to convert the data to a time series object using the quarters of total residential sales from year 2009 to 2023 in which it can be seen that year 2020 had the highest sales in quarter 1 111,207,318 , quarter 2 is 38,430,453 ,quarter 3 is 1,636,720,181 , quarter 4 is 72,277,474 This was to specify a starting time and the number of periods in one year.Then auto.arima function was used to forecast the model .With the help of its summary,in **figure 2** it can be interpreted that ARIM (0,1,0) is ARIMA stands for: AutoRegressive,Integrated (which means differencing),MA = Moving Average

And ARIMA(0,1,0) means:

1. AR (0) : No Auto-Regressive part

No use of past values to predict the current one. It doesn't say: "Sales this quarter depend on last quarter's sales."

2. I (1) : One round of differencing

data was not stable , so it subtract each value from the one before it to make the trend disappear. This just means: instead of looking at actual sales, it indicates how much residential sales are changing from quarter to quarter."

3. MA (0): No Moving Average

Its not using the past forecasting errors to improve current forecast.

In simpler words,it can be interpreted that the best prediction for the next quarter's sales is just the sales from last quarter which is adjusted for average change.

Then the metric ME which is Mean Error which is 1,485,674 means that on average,forecast overestimated actual sales by 145674,RMSE(Root Mean Squared Error) which is 298,610910 means the large value which is indicating huge deviations between forecast ,actual ,MAE (Mean Absolute Error) is 101765985 that is the average abosulte difference between actual and forecast.MPE(Mean Percentage Error) is -47.25% that means on average forecast is underestimated by 47%.The negative sign shows underestimate. MAPE (Mean Absolute Percentage Error) is 81.00% that means on average the forecast is 81 % off from actual values which is quite high .MASE (Mean Absolute Scaled Error) is 0.91 which is below 1 which is good.ACF1 is -0.526 which is negative autocorrelation in residual which is overfitting .Then the **figure 3** shows where the forecasted values of year 2024 and 2025 which is eight quarter and 80% and 95% confidence intervals which is the range where the values are likely to fall. The model predicted that flat real estate sales of about **$**134.1 million for every quarter for the next

8 quarters.Then visualization of forecast for quarter 8 in Elmhurst residential neighborhood is provided in figure 4 where the blue line is for the forecasting of year 2024 and 2025 which is constant which is around $134 million for every quarter and it has wide confidence intervals which is shown by the shaded blue area especially in 95% interval which shows an extreme wide and negative values .

TASK 2

Since the model is not giving good results ,the multiple linear regression was used to determine if there is any relationship to residential sales in Elmhurst from 2009-2023 for which another forecast is created using ets to get better capture of trends and seasonality in the data and also along with it 8 quarters are also created .Then with help of lm function a multiple regression model was created and its summary in **figure 4** revealed that the intercept estimate is 31.3 million and its p value is 0.6625 which means the baseline for quarter 4 at time is 0,time estimate is 2.91 million with p value of 0.00599 which means that sale increasing at 2.91 million per quarter and is marginal significant and quarter 1 estimate is 4 million with p value of 0.9572 and it has no significant effect and quarter 2 estimate is – 30.5 million with p value of 0.6826 and has slight non-significant decrease and quarter 3 has estimate of 96.7 million which id increase in sales but is not statistically significant.

R square explains only 11.5% of the variation of sales which is not a strong fit. Overall, The model suggests sales may increase slightly over time (by about $2.91M per quarter), but the relationship is only marginally significant. The model doesn't explain sales very well overall (low R^2 which is high error), and the seasonal (quarterly) effects are not statistically significant.

In order to use the multiple regression model for a forecast of next 8 quarter, the code of Time = (nrow(regression_data)+1):(nrow(regression_data)+8) was used to forecast next 8 quarters after the historical data of year 2009 -2023.

Then quarter column was turned into dummy variables.This entire code was done to predict future sales, regression model needs the same columns as historical data .Then using ets function the confidence interval of 80 and 95 was done which showed that model forecasts real estate sales with noticeable seasonal variation, peaking in Q3 and dipping in Q2 each year. However, the wide confidence intervals,especially in Q3 indicate significant uncertainty in the predictions as seen in **figure 5**

Then the ETS(M,A,M) model capture the sales data with a slight downward trend and strong seasonal effects that scale with the sales level which was high especially high in Q2. The low smoothing parameters (alpha, beta, gamma) indicated that the model updates slowly with new data, suggesting stable patterns over time. The model provides a slightly better fit than ARIMA, as indicated by its lower AIC value (2468.43 vs. 2473.16), making it a more suitable choice for forecasting in this case as seen in **figure 6.**

Then there were two visulaizations created one for regression model of 95% for actual and predicted value with seasonality and other visualization was created using ets function which show both historical years from 2009-2023 years and predicted years 2024,2025.

The visualization of multiple regression model based on seasonality along with confidence interval for year 2009 -2023 and predictable year forecast for sales are 2024,2025 which xasix and sale prices are in y-axis.The line with olum color shows actual historical years from 2009 -

2023 and dashed red line is for forecasted sales for next 8 quarters.Blue shaded area is 95% confidence interval showing expected range of uncertainty as displayed in **figure 7.**

Figure 8 shows the visualization of n ETS model forecast plot for total residential real estate sales from 2009 through 2025. In which the black line shows the historical sales data. blue line is the ETS model forecast for the next 8 quarters and dark and light blue shaded areas represent the 80% and 95% confidence intervals, showing the range where future sales are expected to fall.

Interpretation

The ETS model captures strong seasonal peaks in the forecast, reflecting patterns seen in past data. However, the wide confidence intervals especially after sharp historical outliers indicated high uncertainty, suggesting that while the ETS model adapts to past seasonal behavior, it is sensitive to extreme values and less confident in future accuracy.

TASK 2

2)

New colomns were made in R language for sale price where log was used to avoid outliers,year built which was renamed as property age ,gross sqaure feet,total units in which the residential units were taken and building type which are categorical ,sale date was taken to determine if they have any relationship with sale price of residential property in Elmhurst neighborhood.In order to establish the multiple regression model lm function was used where sale price was the dependent variable and property age,gross sqaure feet,total units of residential properties,sale date was taken .Summary as provided in **figure 9** shows that building size which is gross sqaure feet ,property age, and number of units significantly impact real estate prices (in log form). Some building types, like H9 is High Rise Residential and RR which is Railway related properties, strongly increase prices, while others like R9 significantly reduce them. The model explains about 27% of the variation in prices.

TASK 3

1)

With help of summary function, the coefficient ‘ s p value showed most useful predictors of sale price were GROSS_SQUARE_FEET, PROPERTY_AGE, and TOTAL_UNITS, as they were statistically significant and had strong effects on price .And least predictors of sale price of residential properties were   variables of any of the BUILDING_TYPE for instance ., A1, B2, R4, etc.) had high p-values (above 0.05), indicating no statistically significant impact on the sale price.These variables add complexity without contributing meaningful predictive power, and could potentially be removed or grouped for a more parsimonious model.

The redundant independent variables were checked using vif function as seen in **figure 10** which showed that   GROSS_SQUARE_FEET and TOTAL_UNITS show high multicollinearity adjusted VIF was > 5 and close to 10.which means these two variables are strongly correlated they may be capturing similar information (e.g., bigger buildings often have more units). PROPERTY_AGE and BUILDING_TYPE show no significant multicollinearity.as for BUILDING_TYPE, the GVIF is adjusted for its many factor levels, and the corrected value is low (1.04), which is acceptable.

- **2)**

- Residual were calculated as actual price and predicted price where a positive redidual would means the property is sold for less which is the bargain and a negative residual means it sold for more than predicted (overpriced).

These were visualized in a scatter plot, with color intensity showing the size of the residual, and ranked using arrange().A scatter plot as seen **in figure 11**

shows the how **predicted sale prices** compare to **actual sale prices** for each residential property in Elmhurst neighborhood , based on multiple regression model. .The x -axis shows actual sale prices and y-axis: Predicted sale prices and the red dashed line: Ideal line where Predicted = Actual which is a (perfect prediction) whereas each point is a property, colored by its residual (prediction error) .The colors purple and blue points (positive residuals which are the bargains as the predicted price was higher than actual and other hand the color Red and orange points (negative residuals) are showing overpriced as Actual price was higher than predicted

TASK 4

Historical quarterly sales data from 2009 to 2023 was used to project future sales across the next eight quarters (2024 and 2025). Initial forecasting using the ARIMA model (AutoRegressive Integrated Moving Average) yielded ARIMA(0,1,0) as the best fit. This model assumed no autoregressive or moving average terms and one level of differencing to achieve stationarity.

However, the ARIMA model produced flat quarterly sales forecasts (~$134.1M) with wide confidence intervals, including negative values. This indicated high uncertainty and weak reliability.

To enhance accuracy, exponential smoothing (ETS) modeling was used. ETS(M,A,M), which accounts for multiplicative error, additive trend, and multiplicative seasonality, captured significant seasonal variation in the data. The model predicted strong seasonal peaks (especially in Q2) and demonstrated more stability with a lower AIC (2468.43) compared to ARIMA. Still, confidence intervals remained wide due to past outliers.

The ETS model emerged as the preferred approach for neighborhood-level forecasts due to its ability to capture seasonality and long-term patterns more effectively than ARIMA forecasting model.

To better understand individual property values, a multiple regression model was developed using features such as property age, gross square footage, number of units, and building type (categorical).

Key findings

- Gross square footage: Strong, positive impact on sale price (statistically significant)
- Year built was renamed as property age : Small, negative effect (significant).
- Number of units: Negative and significant, suggesting larger unit counts reduce per-unit price.Here the number of units aare for residential units in Elmhurst neighborhood.
- Building types: H9 (High-Rise), D1 (Institutional), and RR (Railroad-related) had large, positive impacts, while others like R9 negatively influenced price

The model explained approximately 27% of the variation in sale prices. This suggests moderate explanatory power and room for future model enhancements.

Multicollinearity was assessed using Variance Inflation Factor (VIF):

- Gross square feet and total units had high VIF values (>8), indicating potential multicollinearity
- Building type and property age showed no redundancy

Pricing Disparities and Investment Opportunities

Residual analysis (difference between actual and predicted prices) identified pricing outliers:

- Bargains: Properties sold below predicted value (positive residuals)
- Overpriced: Properties sold above predicted value (negative residuals)

A scatter plot comparing actual vs. predicted prices visually highlighted these disparities. Properties with the largest positive residuals represent potential undervalued investments, while negative residuals suggest overvaluation.

These disparities may stem from:

- Renovations or building conditions not captured in data
- Location premiums or market timing

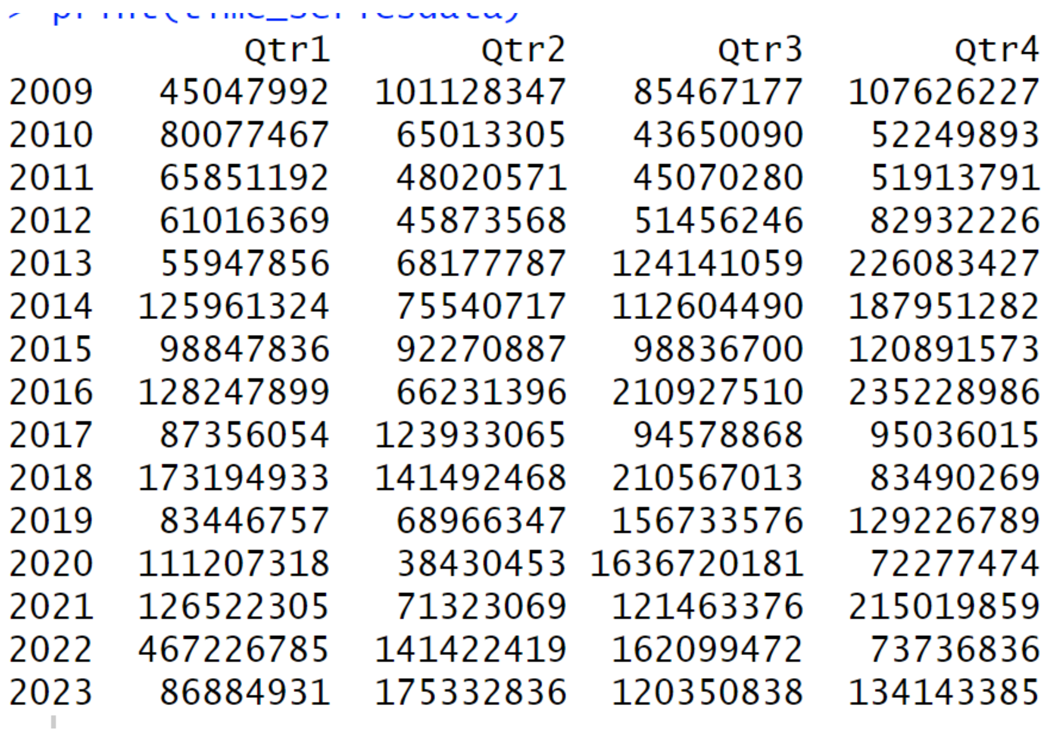
RECOMMENDATION

- Use Regression Modeling for Individual Property Pricing- The multiple linear regression analysis revealed that gross square footage, property age, and number of units are significant predictors of sale price. The brokerage can leverage these factors to refine listing prices, provide more accurate property valuations, and better inform client investment strategies.
- • Be Cautious with Multicollinearity
  The model showed high multicollinearity between total units and square footage. To avoid overestimating the impact of size-related variables, brokerage should consider simplifying or consolidating overlapping metrics in future pricing tools.

APPENDIX

Figure 1

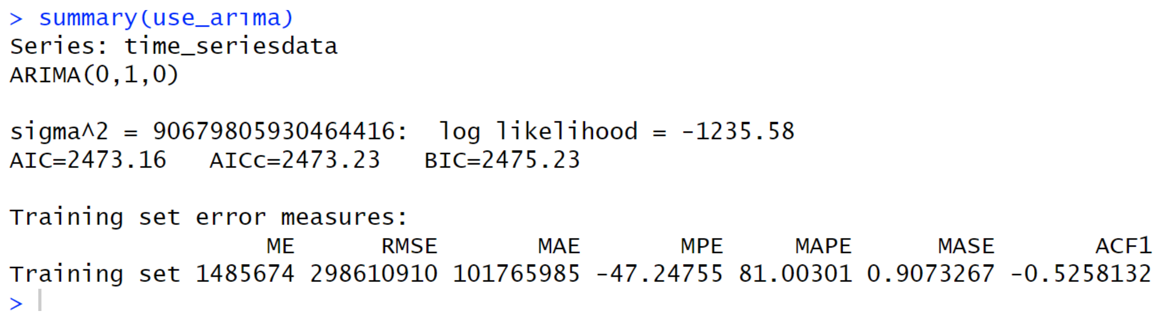
Shows total residential sales in quarters from year 2009 to 2023 in Elmhurst neighborhood



```
> print(time_seriesdata)
          Qtr1       Qtr2       Qtr3       Qtr4
2009  45047992  101128347   85467177  107626227
2010  80077467   65013305   43650090   52249893
2011  65851192   48020571   45070280   51913791
2012  61016369   45873568   51456246   82932226
2013  55947856   68177787  124141059  226083427
2014 125961324   75540717  112604490  187951282
2015  98847836   92270887   98836700  120891573
2016 128247899   66231396  210927510  235228986
2017  87356054  123933065   94578868   95036015
2018 173194933  141492468  210567013   83490269
2019  83446757   68966347  156733576  129226789
2020 111207318   38430453 1636720181   72277474
2021 126522305   71323069  121463376  215019859
2022 467226785  141422419  162099472   73736836
2023  86884931  175332836  120350838  134143385
```

Figure 2

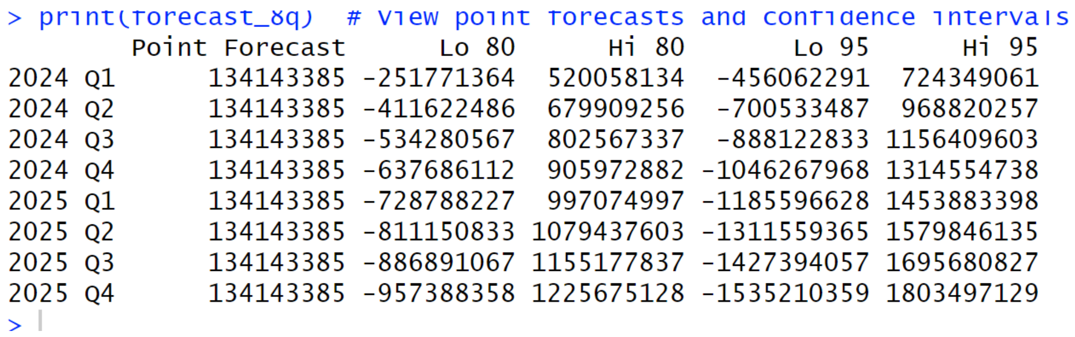
Summary of forecast using auto arima model for time series



```
> summary(use_arima)
Series: time_seriesdata 
ARIMA(0,1,0) 

sigma^2 = 90679805930464416:  log likelihood = -1235.58
AIC=2473.16   AICc=2473.23   BIC=2475.23

Training set error measures:
                  ME      RMSE       MAE       MPE     MAPE      MASE       ACF1
Training set 1485674 298610910 101765985 -47.24755 81.00301 0.9073267 -0.5258132
> 
```

Figure 3

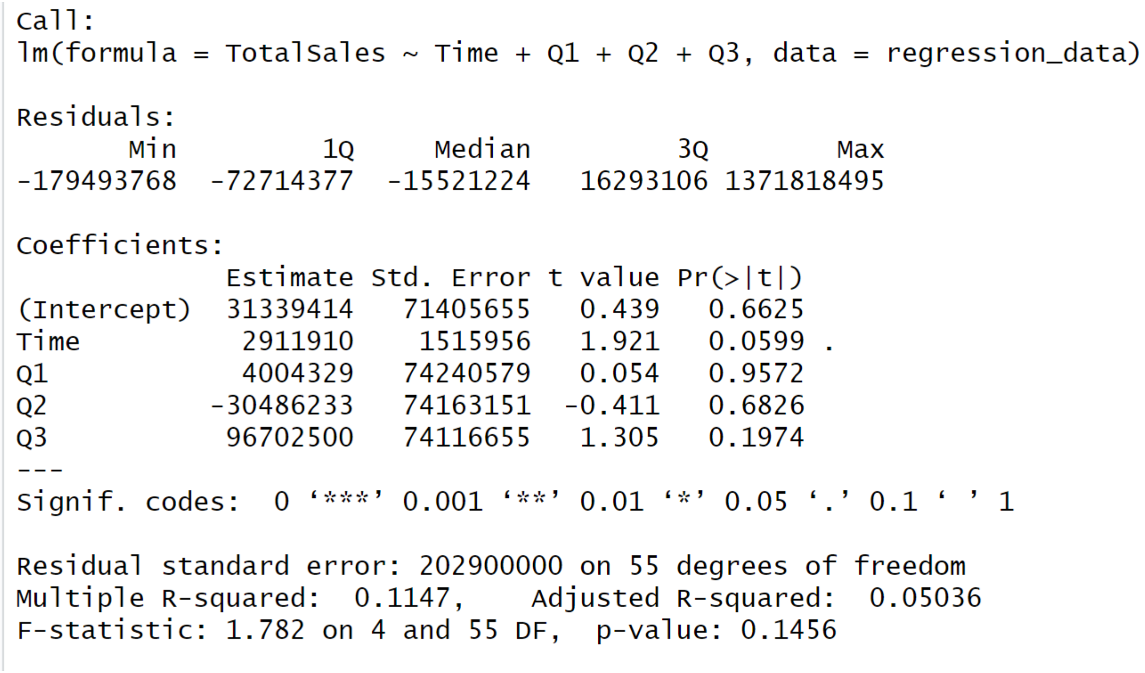
Shows the 80 and 95 confidence interval for forecasted years 2024-2025



```
> print(forecast_8q)  # View point forecasts and confidence intervals
        Point Forecast      Lo 80      Hi 80       Lo 95      Hi 95
2024 Q1      134143385 -251771364  520058134  -456062291  724349061
2024 Q2      134143385 -411622486  679909256  -700533487  968820257
2024 Q3      134143385 -534280567  802567337  -888122833 1156409603
2024 Q4      134143385 -637686112  905972882 -1046267968 1314554738
2025 Q1      134143385 -728788227  997074997 -1185596628 1453883398
2025 Q2      134143385 -811150833 1079437603 -1311559365 1579846135
2025 Q3      134143385 -886891067 1155177837 -1427394057 1695680827
2025 Q4      134143385 -957388358 1225675128 -1535210359 1803497129
> 
```

Figure 4

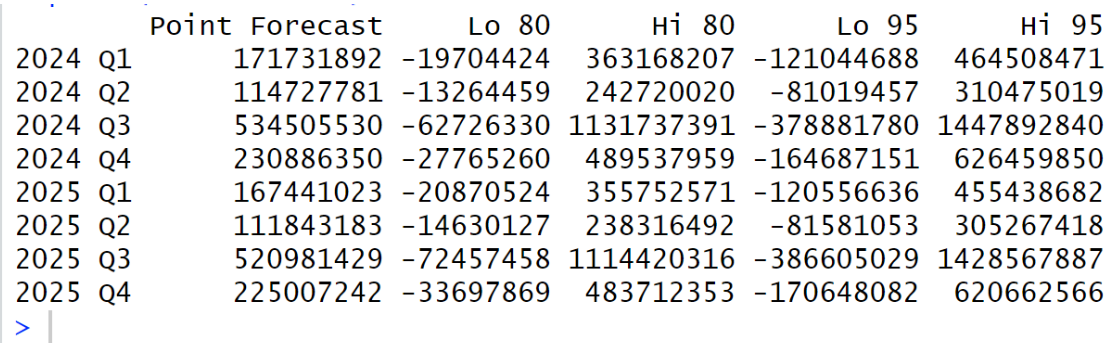
Shows the summary of multiple regression model for Elmhurst neighborhood



```
Call:
lm(formula = TotalSales ~ Time + Q1 + Q2 + Q3, data = regression_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-179493768  -72714377  -15521224   16293106 1371818495 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  31339414   71405655   0.439   0.6625  
Time          2911910    1515956   1.921   0.0599 .
Q1            4004329   74240579   0.054   0.9572  
Q2          -30486233   74163151  -0.411   0.6826  
Q3           96702500   74116655   1.305   0.1974  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202900000 on 55 degrees of freedom
Multiple R-squared:  0.1147,	Adjusted R-squared:  0.05036 
F-statistic: 1.782 on 4 and 55 DF,  p-value: 0.1456
```

Figure 5

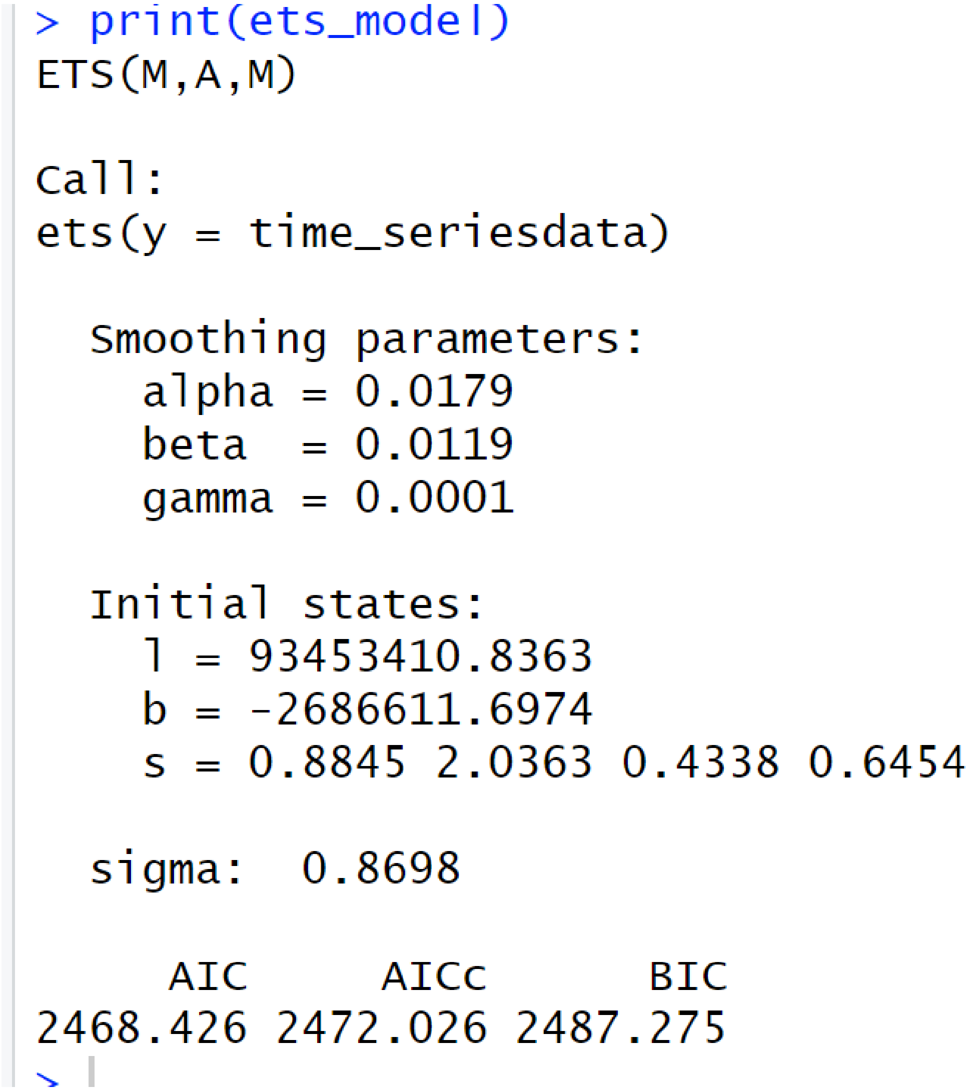
Show the confidence interval of 80 and 95 for forecast of 8 quarter i.e 2024-2025 year using ets function



```
        Point Forecast     Lo 80      Hi 80      Lo 95      Hi 95
2024 Q1      171731892 -19704424  363168207 -121044688  464508471
2024 Q2      114727781 -13264459  242720020  -81019457  310475019
2024 Q3      534505530 -62726330 1131737391 -378881780 1447892840
2024 Q4      230886350 -27765260  489537959 -164687151  626459850
2025 Q1      167441023 -20870524  355752571 -120556636  455438682
2025 Q2      111843183 -14630127  238316492  -81581053  305267418
2025 Q3      520981429 -72457458 1114420316 -386605029 1428567887
2025 Q4      225007242 -33697869  483712353 -170648082  620662566
> 
```

Figure 6

Shows the summary of ets model



```
> print(ets_model)
ETS(M,A,M)

Call:
ets(y = time_seriesdata)

  Smoothing parameters:
    alpha = 0.0179
    beta  = 0.0119
    gamma = 0.0001

  Initial states:
    l = 93453410.8363
    b = -2686611.6974
    s = 0.8845 2.0363 0.4338 0.6454

  sigma:  0.8698

     AIC     AICc      BIC
2468.426 2472.026 2487.275
>
```

Figure 7

Visualization of multiple regression model with actual and predicted sales of 95% confidence interval

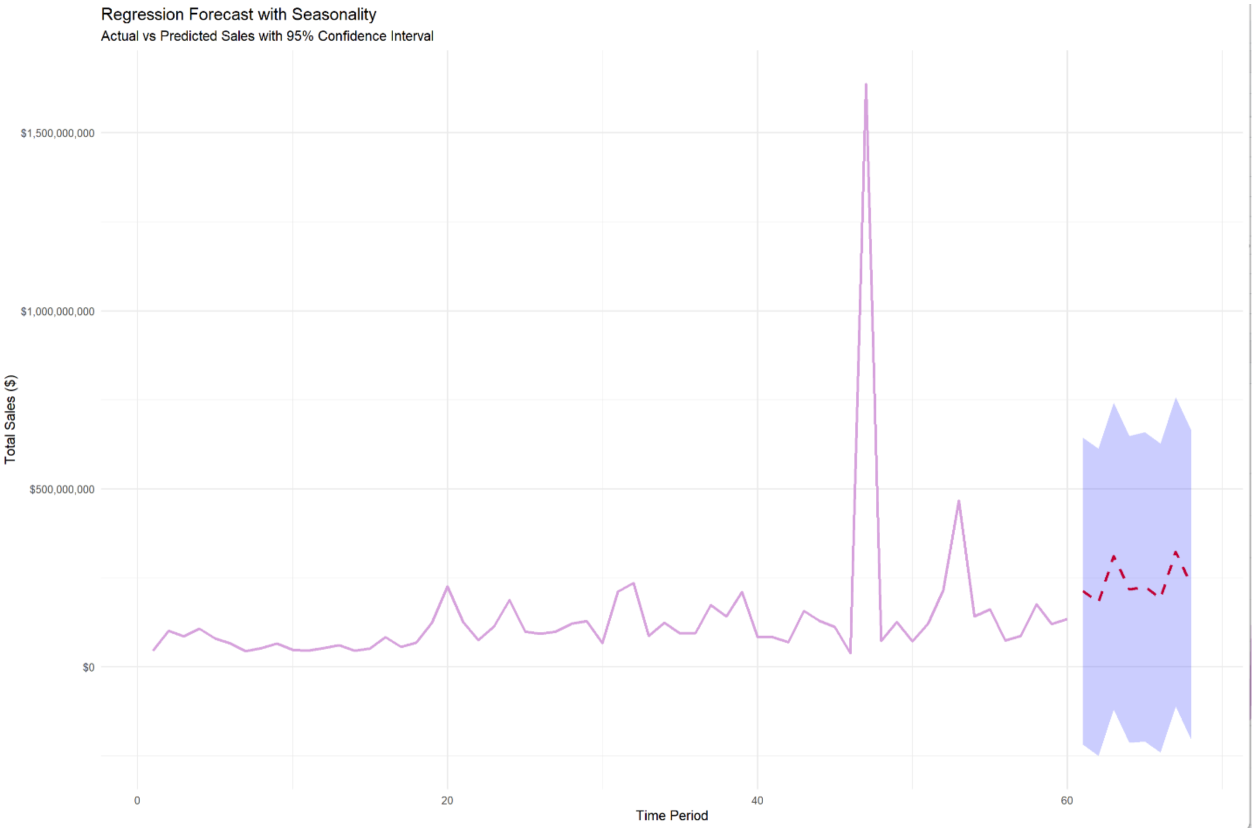


Figure 8

Shows the ets forecast model

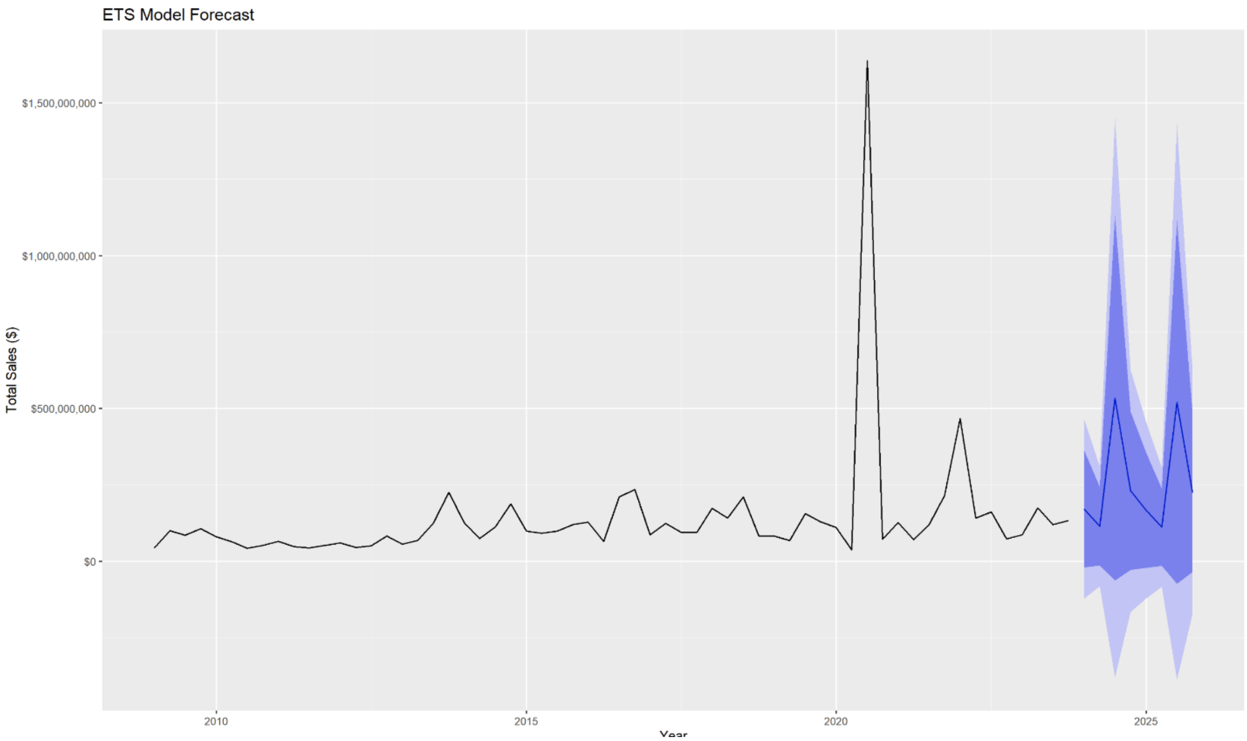
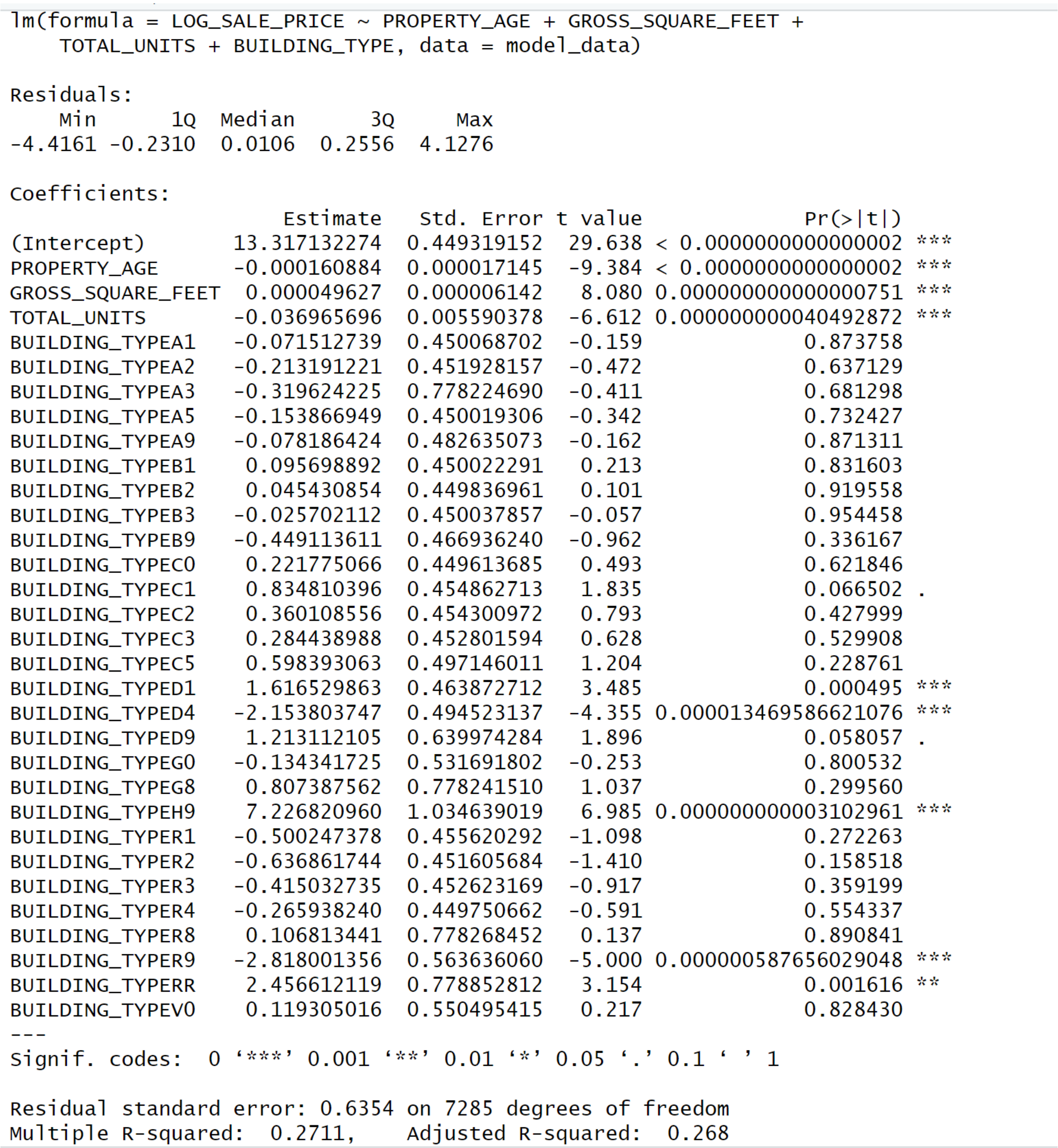


Figure 9

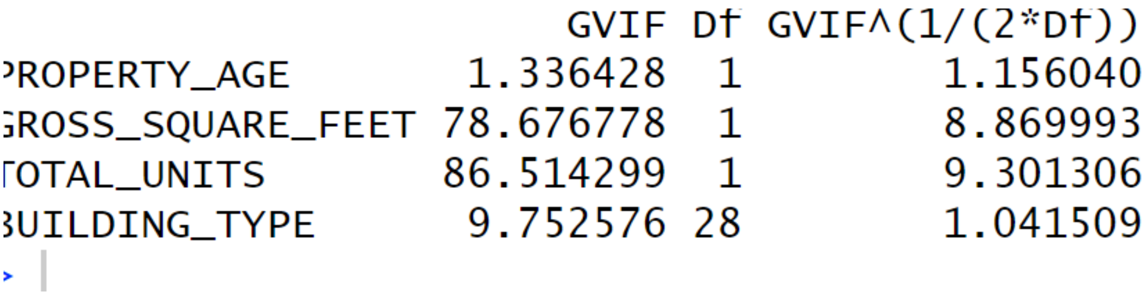
Shows the summary of multiple regression model for year built,total units,gross sqaure feet,building type categorical and sale date with sale price as dependent variable



```
lm(formula = LOG_SALE_PRICE ~ PROPERTY_AGE + GROSS_SQUARE_FEET + 
    TOTAL_UNITS + BUILDING_TYPE, data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4161 -0.2310  0.0106  0.2556  4.1276 

Coefficients:
                      Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)       13.317132274  0.449319152  29.638 < 0.0000000000000002 ***
PROPERTY_AGE      -0.000160884  0.000017145  -9.384 < 0.0000000000000002 ***
GROSS_SQUARE_FEET  0.000049627  0.000006142   8.080 0.000000000000000751 ***
TOTAL_UNITS       -0.036965696  0.005590378  -6.612 0.000000000040492872 ***
BUILDING_TYPEA1   -0.071512739  0.450068702  -0.159             0.873758    
BUILDING_TYPEA2   -0.213191221  0.451928157  -0.472             0.637129    
BUILDING_TYPEA3   -0.319624225  0.778224690  -0.411             0.681298    
BUILDING_TYPEA5   -0.153866949  0.450019306  -0.342             0.732427    
BUILDING_TYPEA9   -0.078186424  0.482635073  -0.162             0.871311    
BUILDING_TYPEB1    0.095698892  0.450022291   0.213             0.831603    
BUILDING_TYPEB2    0.045430854  0.449836961   0.101             0.919558    
BUILDING_TYPEB3   -0.025702112  0.450037857  -0.057             0.954458    
BUILDING_TYPEB9   -0.449113611  0.466936240  -0.962             0.336167    
BUILDING_TYPEC0    0.221775066  0.449613685   0.493             0.621846    
BUILDING_TYPEC1    0.834810396  0.454862713   1.835             0.066502 .  
BUILDING_TYPEC2    0.360108556  0.454300972   0.793             0.427999    
BUILDING_TYPEC3    0.284438988  0.452801594   0.628             0.529908    
BUILDING_TYPEC5    0.598393063  0.497146011   1.204             0.228761    
BUILDING_TYPED1    1.616529863  0.463872712   3.485             0.000495 ***
BUILDING_TYPED4   -2.153803747  0.494523137  -4.355 0.000013469586621076 ***
BUILDING_TYPED9    1.213112105  0.639974284   1.896             0.058057 .  
BUILDING_TYPEG0   -0.134341725  0.531691802  -0.253             0.800532    
BUILDING_TYPEG8    0.807387562  0.778241510   1.037             0.299560    
BUILDING_TYPEH9    7.226820960  1.034639019   6.985 0.000000000003102961 ***
BUILDING_TYPER1   -0.500247378  0.455620292  -1.098             0.272263    
BUILDING_TYPER2   -0.636861744  0.451605684  -1.410             0.158518    
BUILDING_TYPER3   -0.415032735  0.452623169  -0.917             0.359199    
BUILDING_TYPER4   -0.265938240  0.449750662  -0.591             0.554337    
BUILDING_TYPER8    0.106813441  0.778268452   0.137             0.890841    
BUILDING_TYPER9   -2.818001356  0.563636060  -5.000 0.000000587656029048 ***
BUILDING_TYPERR    2.456612119  0.778852812   3.154             0.001616 ** 
BUILDING_TYPEV0    0.119305016  0.550495415   0.217             0.828430    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6354 on 7285 degrees of freedom
Multiple R-squared:  0.2711,	Adjusted R-squared:  0.268 
```

Figure 10

Use of Vif function to see redundancy in independent variable



```
                      GVIF Df GVIF^(1/(2*Df))
PROPERTY_AGE        1.336428  1        1.156040
GROSS_SQUARE_FEET  78.676778  1        8.869993
TOTAL_UNITS        86.514299  1        9.301306
BUILDING_TYPE       9.752576 28        1.041509
> 
```