Real Estate Sales Prediction Analysis

Chuqi(Angel) Jin

Contents

01

INTRODUCTION

02

DATA DESCRIPTION

03

EXPLORATORY DATA ANALYSIS

04

MODEL BUILDING

05

FINAL MODEL & INTERPRETATION

06

CONCLUSION & LIMITATION

Introduction

Dataset: Real Estate Sales (Kaggle)

Goals: Which factors significantly impact property sale price?

- **Hypothesis:** The year of the real estate transaction has a significant on the sale price.

- **Original Dataset:** The dataset initially included real estate transaction data from Connecticut between 2009 till 2022, with a total of 10,000 observations and 12 variables.

- **Cleaned Dataset:** After carefully cleaning and selecting, I got 2574 observations and 12 variables.
  - Remove all 0 and NA values.
  - Filter the observations to the years 2017–2022.
  - Divided the sale price and estimated values by 1,000.

- **Model Building Dataset**: each categorical variables has too many category, subset top 2 of each categorical variables and left dataset with 397 variables for model building

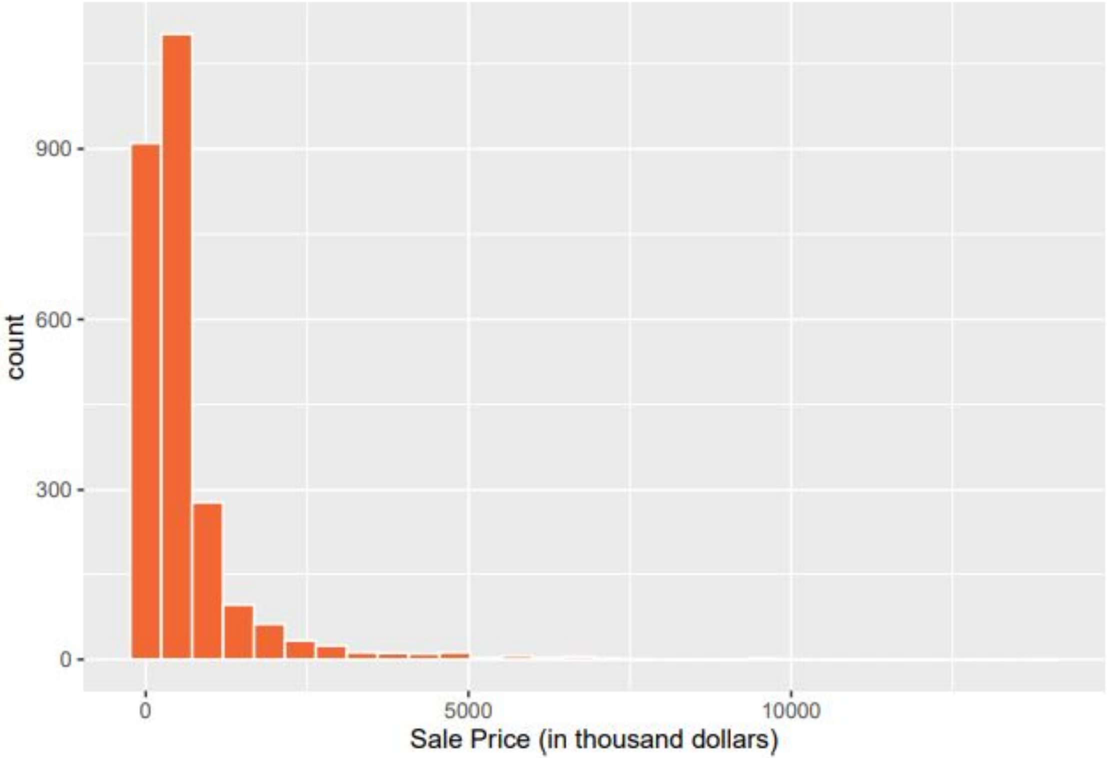
Data Description

Response Variable

Sale.Price: (measured in thousand dollars) – The actual sale price of the property

- **Date**: The property transaction date (Numerical)
- **Year**: The property transaction year (Numerical)
- **Locality**: The property locality/area (Categorical)
- **Estimated.Value**: (measured in thousand dollars) The estimated value of the property (Numerical)
- **Property**: Types of properties suitable for various family sizes (Categorical)
- **Residential**: Indicates whether the property is designated for residential use (Categorical)
- **num_rooms**: The number of rooms in the property (Numerical)
- **num_bathrooms**: The number of bathrooms in the property (Numerical)
- **carpet_area**: (measured in square feet) The carpet area of the property (Numerical)
- **Property_tax_rate**: Tax rate applied to the property's assessed value (Numerical)
- **Face**: Direction the main entrance/facade of the property is oriented towards (Categorical)

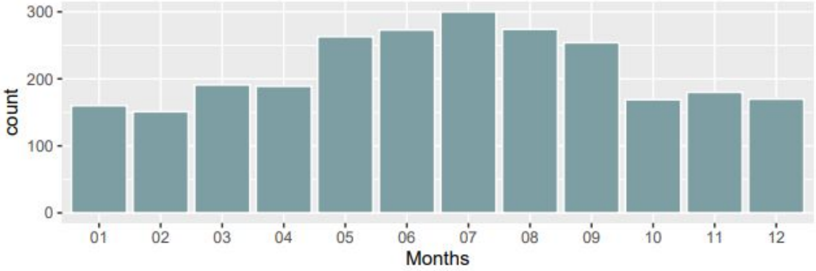
EDA: Response Variable

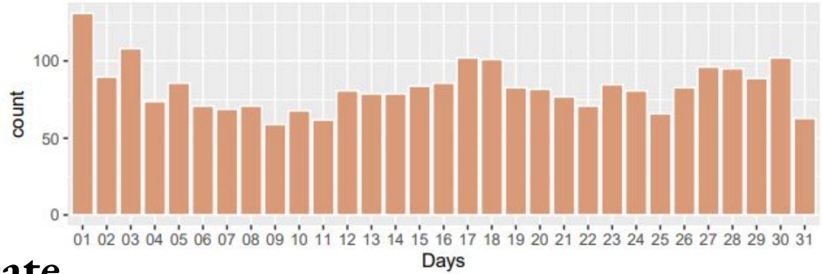


Sale.Price

- The histogram indicates that the response variable is not normally distributed.
- Right skewed

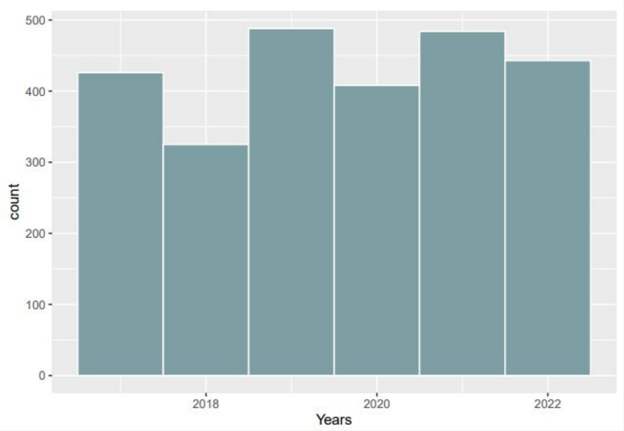
EDA: Predictor Variables

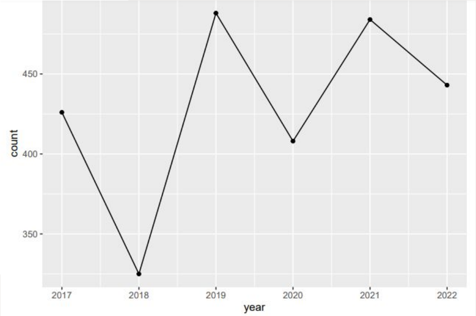




Date

– July is the most popular month for real estate sales.
– The most real estate sales occurred on the first day of the month.

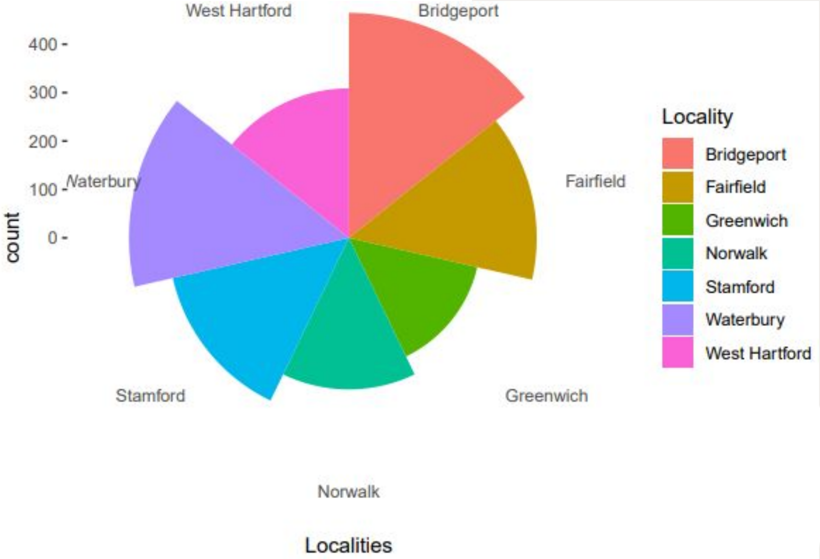




Year

– Real estate sales were highest in 2019 and lowest in 2018.

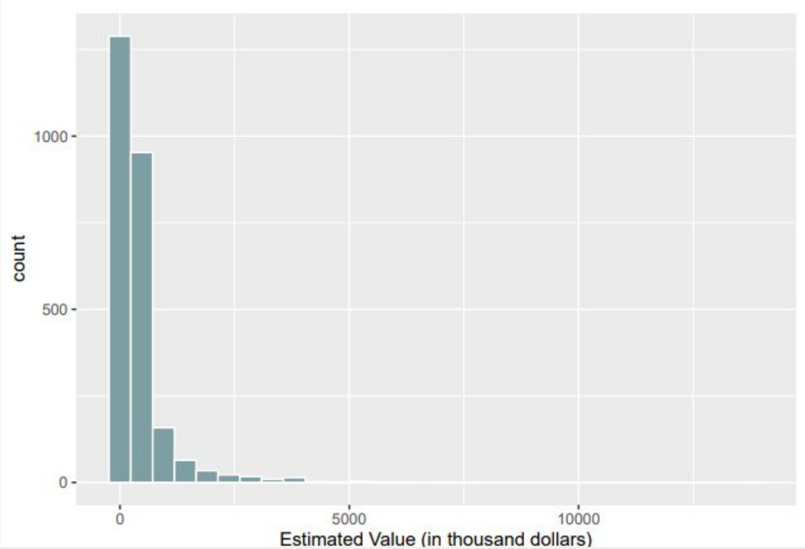
EDA



Locality

– 6 different locations

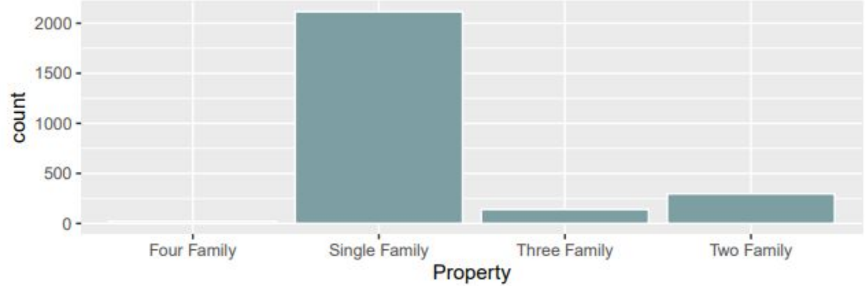
– The top 2 localities are Bridgeport & Waterbury

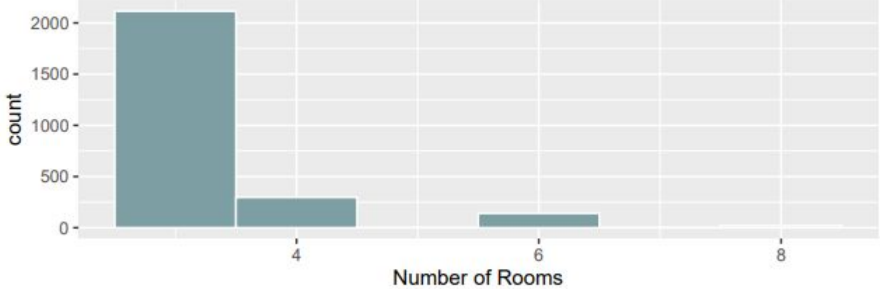


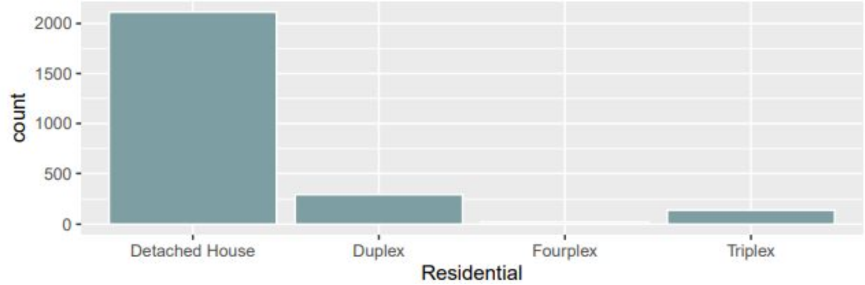
Estimated Value

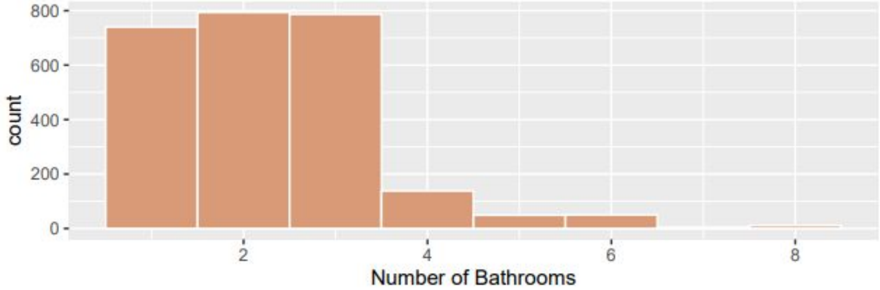
- The histogram indicates that this variable is not normally distributed.
- Right skewed

EDA









Property

– 4 different property types

– The top 2 properties are Single & Two Family

Residential

– 4 different residential types

– The top 2 residence are Detached House & Duplex

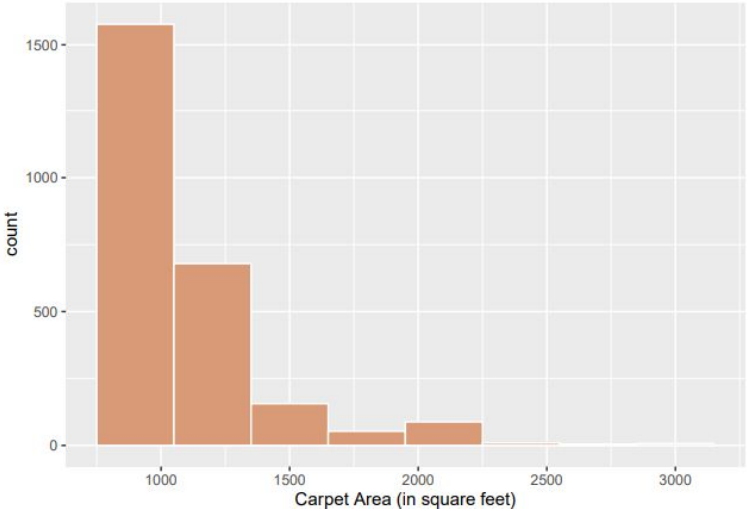
Room numbers

– Between 3 and 8 rooms

Bathroom numbers

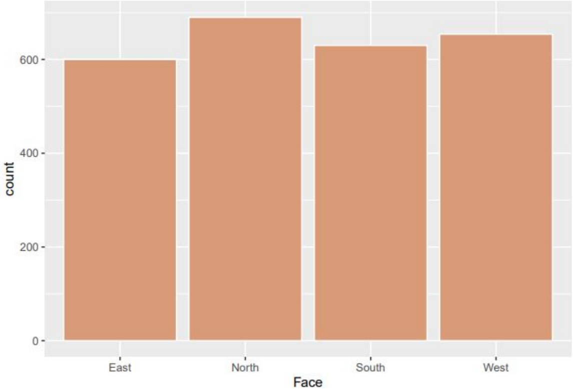
– Between 1-8 bathrooms

EDA



Carpet Area

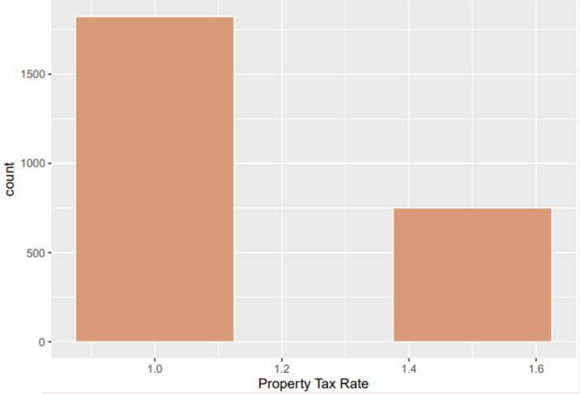
– Range of 900 - 2989 square feets



Face

– 4 different facing directions

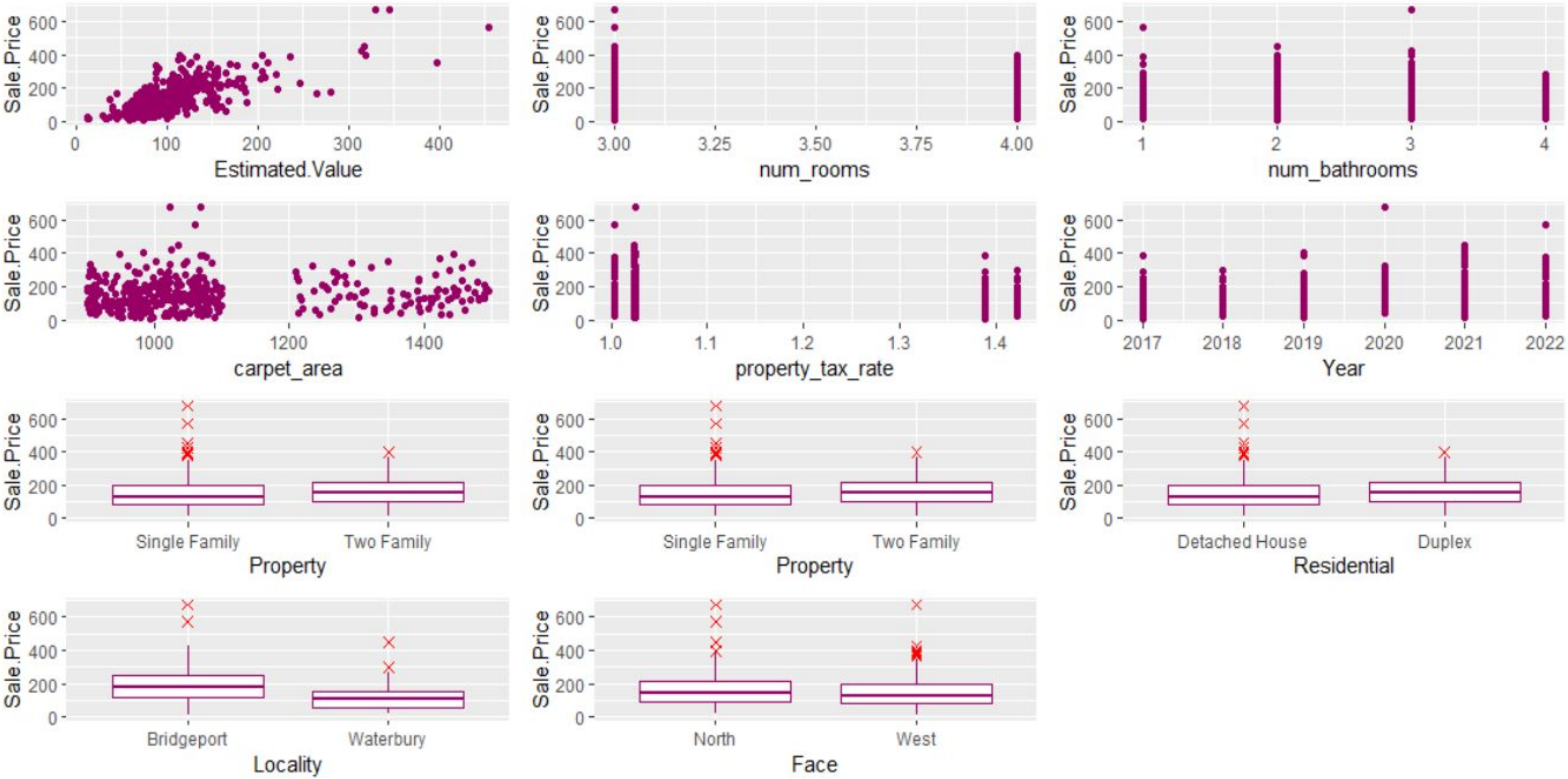
– The top 2 directions are North & West



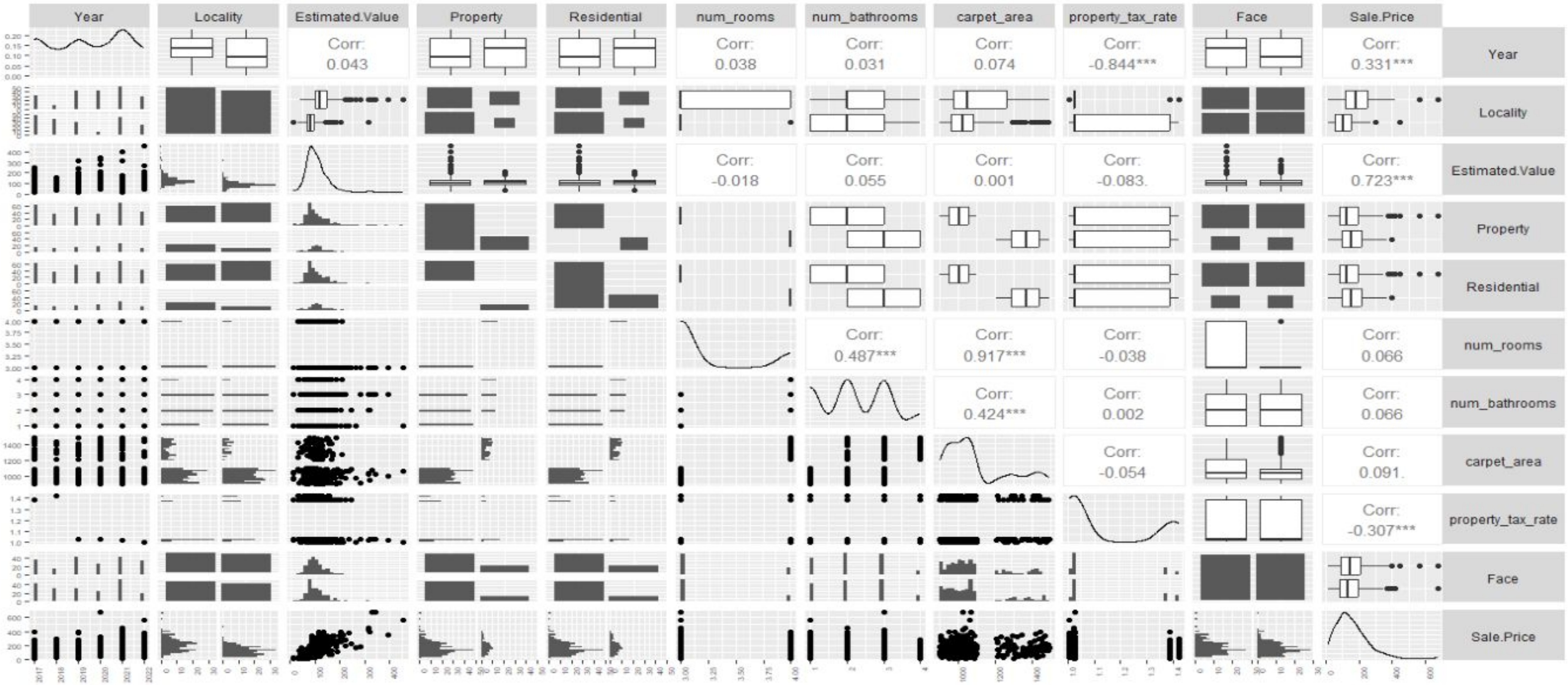
Property Tax Rate

– Range from 1.004 to 1.422

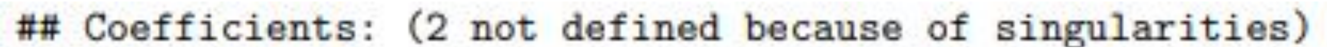
EDA


Sale.Price
0
200
400
600
0
100
200
300
400
Estimated.Value
3.00
3.25
3.50
3.75
4.00
num_rooms
1
2
3
4
num_bathrooms
1000
1200
1400
carpet_area
1.0
1.1
1.2
1.3
1.4
property_tax_rate
2017
2018
2019
2020
2021
2022
Year
Single Family
Two Family
Property
Single Family
Two Family
Property
Detached House
Duplex
Residential
Bridgeport
Waterbury
Locality
North
West
Face

Multicollinearity Analysis

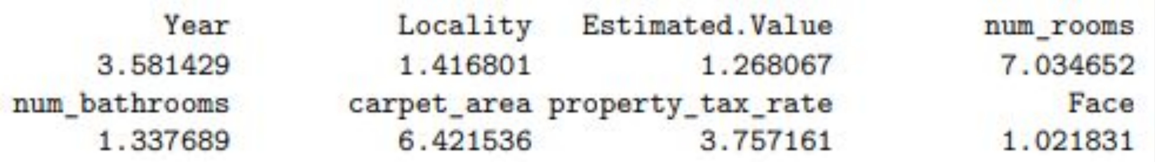


VIF



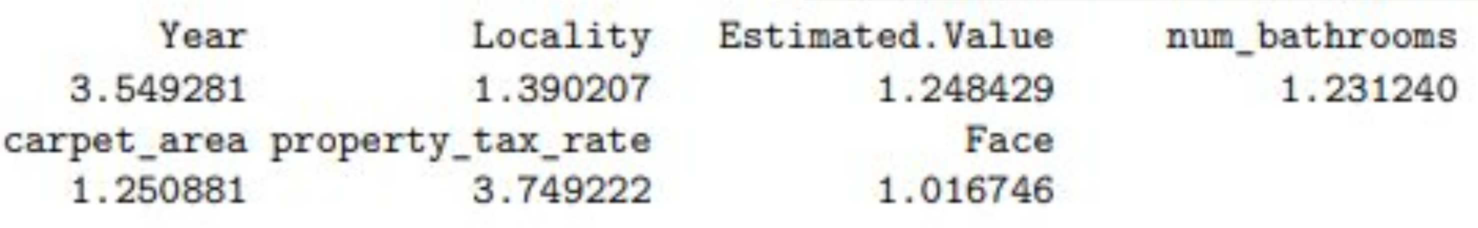
```
## Coefficients: (2 not defined because of singularities)
```

Remove Property & Residential



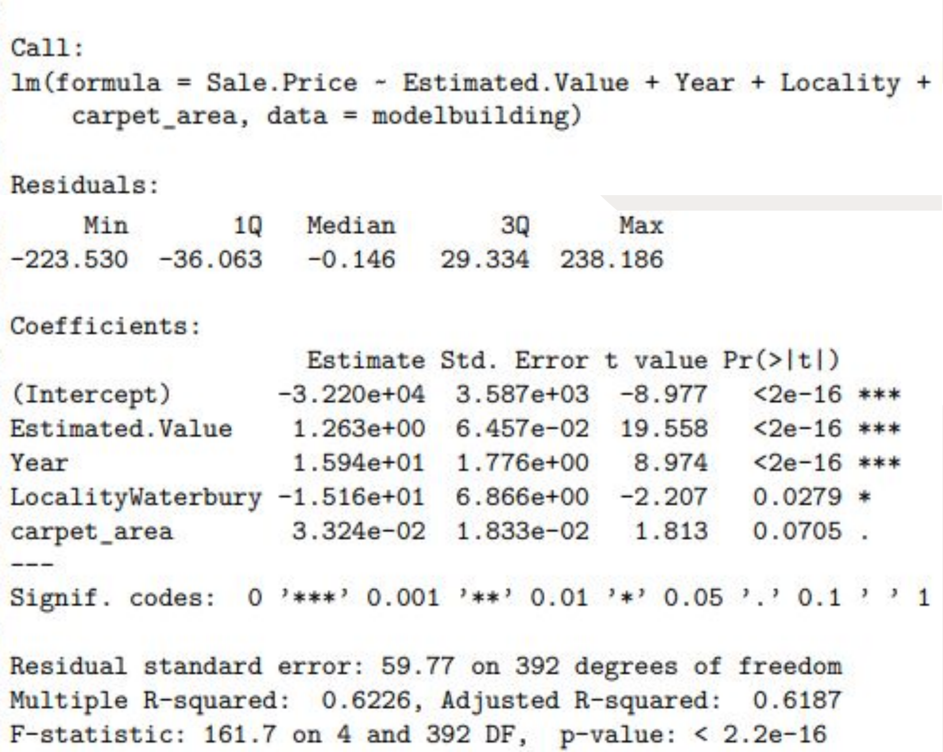
```
          Year         Locality  Estimated.Value        num_rooms
      3.581429         1.416801         1.268067         7.034652
 num_bathrooms      carpet_area property_tax_rate            Face
      1.337689         6.421536         3.757161         1.021831
```

Remove num_rooms

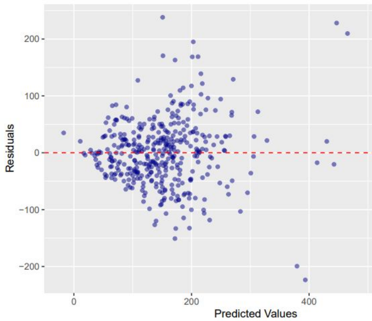


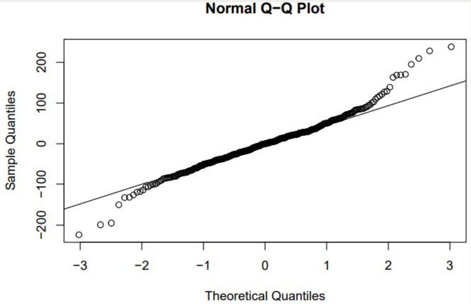
```
         Year          Locality   Estimated.Value     num_bathrooms
     3.549281          1.390207          1.248429          1.231240
  carpet_area property_tax_rate              Face
     1.250881          3.749222          1.016746
```

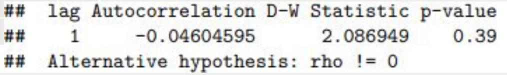
Forward Stepwise Model



```
Call:
lm(formula = Sale.Price ~ Estimated.Value + Year + Locality +
    carpet_area, data = modelbuilding)

Residuals:
     Min       1Q   Median       3Q      Max
-223.530  -36.063   -0.146   29.334  238.186

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.220e+04  3.587e+03  -8.977   <2e-16 ***
Estimated.Value   1.263e+00  6.457e-02  19.558   <2e-16 ***
Year              1.594e+01  1.776e+00   8.974   <2e-16 ***
LocalityWaterbury -1.516e+01  6.866e+00  -2.207   0.0279 *
carpet_area       3.324e-02  1.833e-02   1.813   0.0705 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 59.77 on 392 degrees of freedom
Multiple R-squared:  0.6226, Adjusted R-squared:  0.6187
F-statistic: 161.7 on 4 and 392 DF,  p-value: < 2.2e-16
```





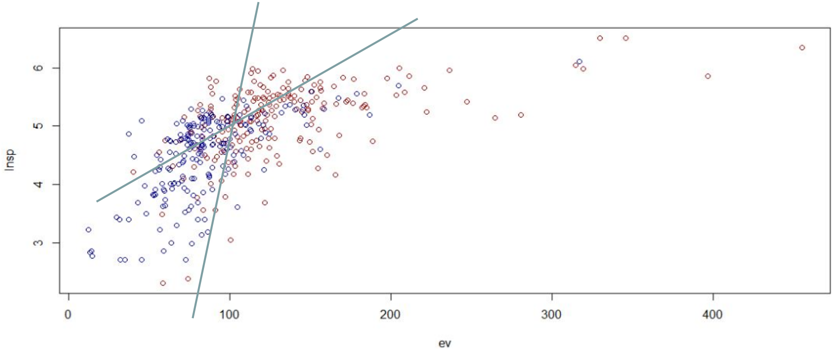


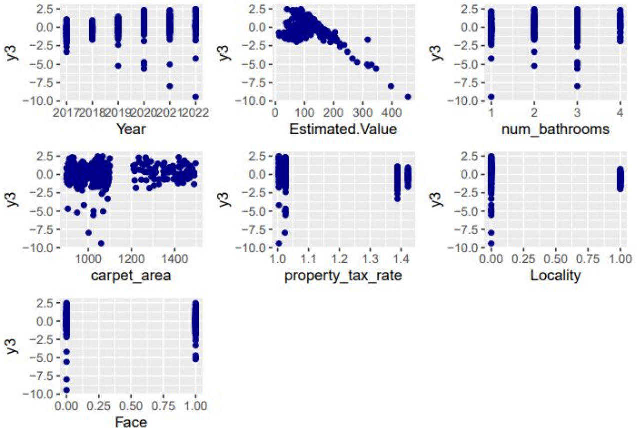
```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04604595      2.086949    0.39
##  Alternative hypothesis: rho != 0
```

- Final predictors: Estimated value, year, locality and carpet area.
- Adjusted R^2: 0.6187
- Not pass Normality, Linearity or Constant Variance
- Pass independence check

Interaction Term

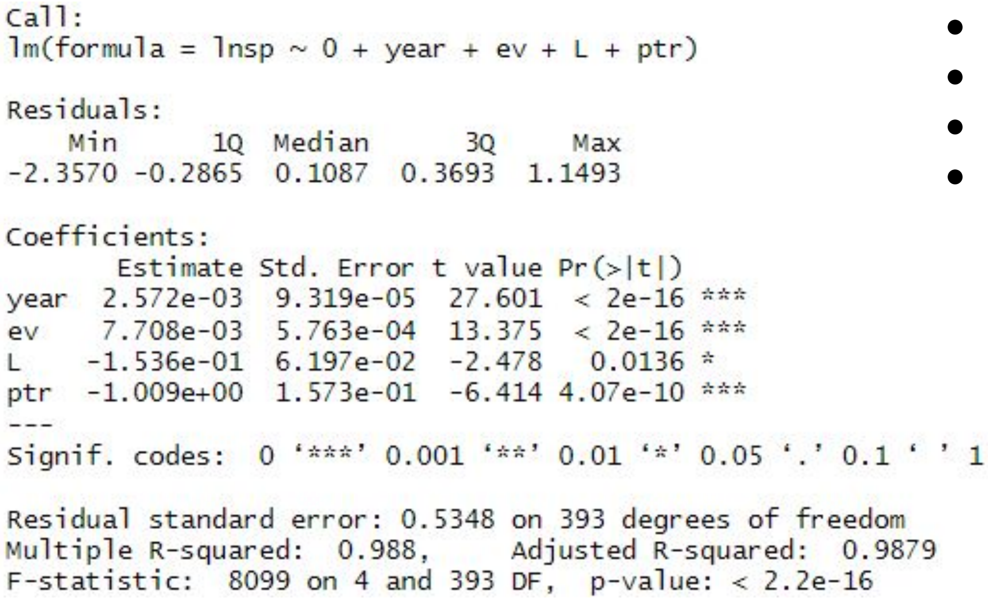
- Convert all categorical variables into binary variables
- Check interaction between Estimated Value & Locality
- Check the effect of Year
    - Present but does not seem to significantly impact the model
- Decides to transform Sale.Price (y) using log transformation
- Recheck the interaction between Estimated Value & Locality
- Recheck the effect of Year
- Decide to remove carpet area, number of bathrooms and facing direction from the predictor variables





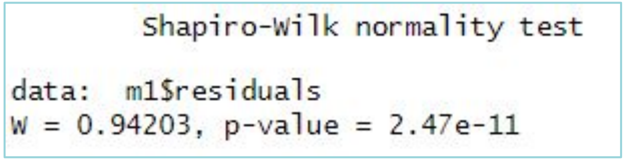
Model 1

ln(Sale Price) = Year + Locality + Estimated Value + Property Tax Rate

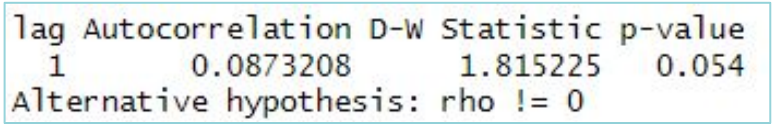


```
Call:
lm(formula = lnsp ~ 0 + year + ev + L + ptr)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3570 -0.2865  0.1087  0.3693  1.1493 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
year  2.572e-03  9.319e-05  27.601  < 2e-16 ***
ev    7.708e-03  5.763e-04  13.375  < 2e-16 ***
L    -1.536e-01  6.197e-02  -2.478   0.0136 *  
ptr  -1.009e+00  1.573e-01  -6.414 4.07e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5348 on 393 degrees of freedom
Multiple R-squared:  0.988,	Adjusted R-squared:  0.9879 
F-statistic:  8099 on 4 and 393 DF,  p-value: < 2.2e-16
```

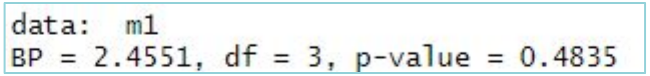
- Adjusted R-squared: 0.9879
- Residuals not normally distributed
- Residuals satisfied independence test
- Residuals satisfied constant variance test



```
        Shapiro-Wilk normality test

data:  m1$residuals
W = 0.94203, p-value = 2.47e-11
```



```
lag Autocorrelation D-W Statistic p-value
  1       0.0873208      1.815225   0.054
Alternative hypothesis: rho != 0
```

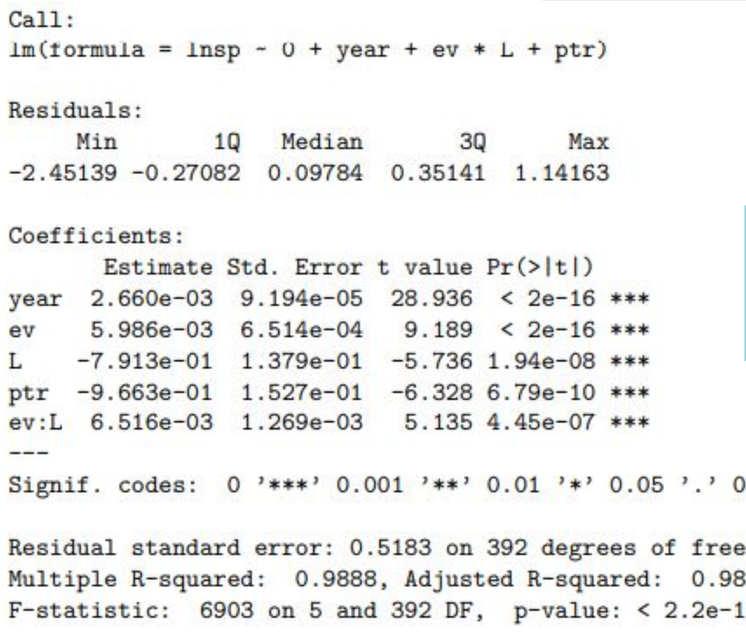


```
data:  m1
BP = 2.4551, df = 3, p-value = 0.4835
```

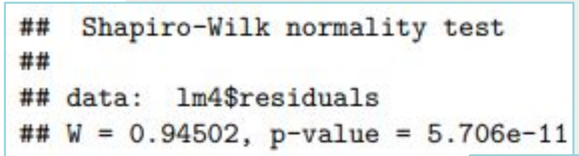
Model 2

ln(Sale Price) = Year + Locality * Estimated Value + Property Tax Rate

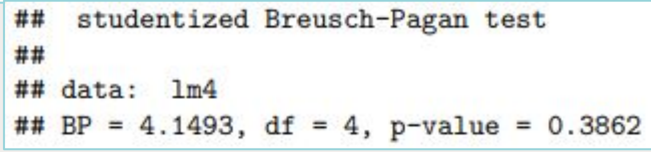
- Adjusted R-squared: 0.9886
- Residuals not normally distributed
- Residuals failed independence test, so it shows autocorrelation
- Residuals satisfied constant variance test



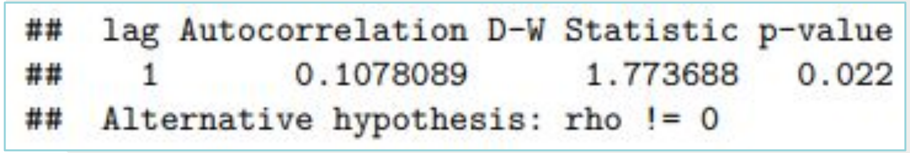
```
Call:
lm(formula = lnsp ~ 0 + year + ev * L + ptr)

Residuals:
     Min       1Q   Median       3Q      Max
-2.45139 -0.27082  0.09784  0.35141  1.14163

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
year  2.660e-03  9.194e-05  28.936  < 2e-16 ***
ev    5.986e-03  6.514e-04   9.189  < 2e-16 ***
L    -7.913e-01  1.379e-01  -5.736 1.94e-08 ***
ptr  -9.663e-01  1.527e-01  -6.328 6.79e-10 ***
ev:L  6.516e-03  1.269e-03   5.135 4.45e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5183 on 392 degrees of freedom
Multiple R-squared:  0.9888, Adjusted R-squared:  0.9886
F-statistic:  6903 on 5 and 392 DF,  p-value: < 2.2e-16
```



```
##  Shapiro-Wilk normality test
##
## data:  lm4$residuals
## W = 0.94502, p-value = 5.706e-11
```

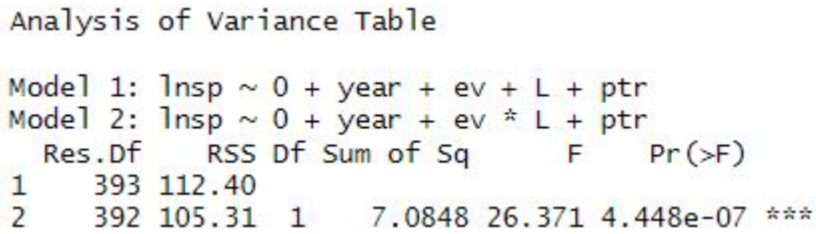


```
##  studentized Breusch-Pagan test
##
## data:  lm4
## BP = 4.1493, df = 4, p-value = 0.3862
```

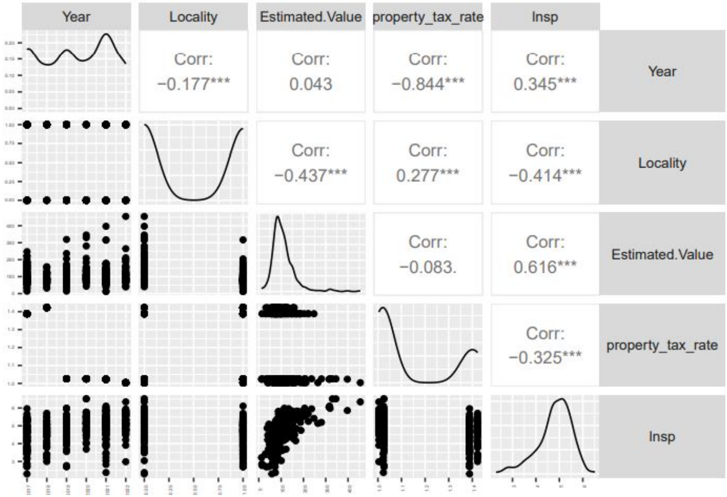


```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1078089      1.773688   0.022
##  Alternative hypothesis: rho != 0
```

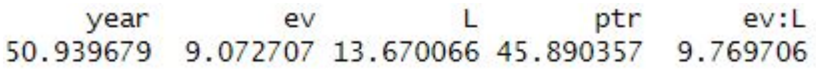
Model Comparison & Model Diagnostics



```
Analysis of Variance Table

Model 1: lnsp ~ 0 + year + ev + L + ptr
Model 2: lnsp ~ 0 + year + ev * L + ptr
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    393 112.40                                  
2    392 105.31  1    7.0848 26.371 4.448e-07 ***
```



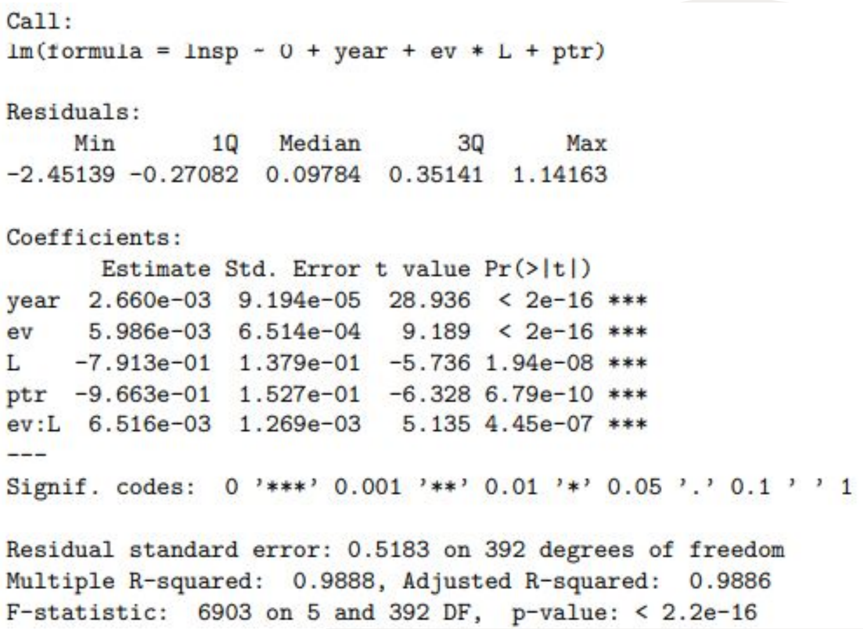
- P-value is extremely small, therefore, model 1 vs. model 2 indicates that there is a huge difference between 2 models.
- Choose Model 2 as our final model



```
     year        ev         L       ptr      ev:L 
50.939679  9.072707 13.670066 45.890357  9.769706 
```

Final Model

ln(Sale Price) = 2.660e^{-3}Year + 5.986e^{-3} Estimated Value - 7.913e^{-1}Locality + 6.516e^{-3} (Locality * Estimated Value) - 9.6636e^{-1}Property Tax Rate



```
Call:
lm(formula = lnsp ~ 0 + year + ev * L + ptr)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.45139 -0.27082  0.09784  0.35141  1.14163 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
year  2.660e-03  9.194e-05  28.936  < 2e-16 ***
ev    5.986e-03  6.514e-04   9.189  < 2e-16 ***
L    -7.913e-01  1.379e-01  -5.736 1.94e-08 ***
ptr  -9.663e-01  1.527e-01  -6.328 6.79e-10 ***
ev:L  6.516e-03  1.269e-03   5.135 4.45e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5183 on 392 degrees of freedom
Multiple R-squared:  0.9888,	Adjusted R-squared:  0.9886 
F-statistic:  6903 on 5 and 392 DF,  p-value: < 2.2e-16
```

Conclusion & Limitation

Conclusion: The final model I obtained has a very high adjusted R-squared. The final model suggests that a real estate sales price is influenced by the year, estimated value, locality, property tax rate, and the interaction of estimated value and location. Based on my hypothesis, I can conclude that the year of the real estate transaction has a significant effect on the sale price.

Limitation: My final model does not satisfy the normality or independence assumptions, and it has a high vif score with several variables, which could be due to my dataset being too small. As a result, for my future work, I'd like to collect additional observations, expand the dataset size, and rebuild the model to see if I can achieve better results for residual assumption checks.

THANK YOU!

Questions?