Struggling with Normality and Independence:

An OLS Multivariate Linear Regression Model

By Jonathan Newcomb
10/2/2022

OVERVIEW

Evidence suggests that the Relationships between real estate Prices and many of the Independent Variables are not Linear.

This is a Problem!

Outline

- Business Problem
- Data Understanding
- Methodology
  - Measurements
  - Process
- Modeling
  - Results
  - Techniques used
- Recommendations
- Next Steps

Business Problem

What factors most influence Home Sales in King County?

YOUR TEAM'S RESPONSES!!!!!

Number of Bedrooms
Number of Bathrooms
Number of Floors
Square Footage of Home
Square Footage of Lot
Basement Size
Patio Size
Garage Size
Property is in a Greenbelt
Property has a nuisance issue
Property View
Home Condition
Home Grade
Sewer system type
Heat source
Property on waterfront
Year the property was built
Year the property was renovated
Location

Data Understanding

TIMEFRAME: Homes sold in King County between 6-10-21 and 6-9-22

DEPENDENT: VARIABLE: Price - based on home sales in timeframe

Most Independent variable data obtained from King County Assessor open records page

Importantly, **Location** Independent variable was constructed using a Geocoding API

- Latitude
- Longitude
- Address

The Iterative Process

Clean, visualize, and understand data

Prepare data for model

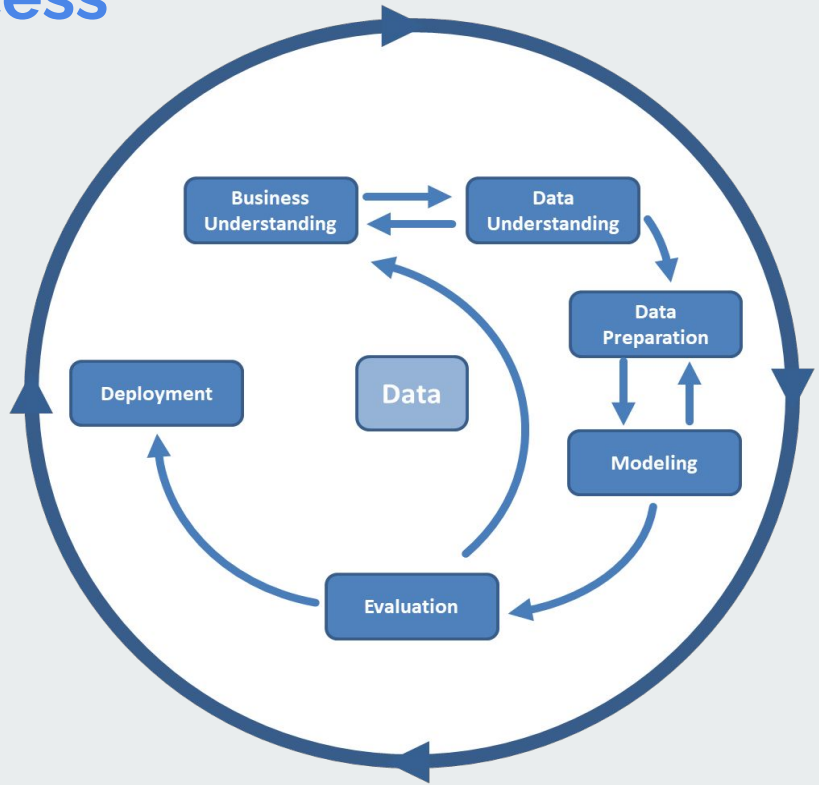
Model relationship between variables

Evaluate model

Revise and refine techniques

Engineer features/pull in more data to help model perform better

Start process over again



HOLD ON!

WHY ARE WE DOING THIS?

WHY IS THIS IMPORTANT?

The OLS model learns from the data we put into the model

The more quality information we give the model the more accurately the model will be able to predict home prices in King County

Measuring our Model

What makes a good linear regression model?

1. **R-squared:**

   should be a high percentage %

2. **Residual Errors:**

   should be a low number

   TESTS FOR RESIDUALS: RMSE/ MAE

Very Important:
Model must be

1. **Linear**
2. **Normal**
3. **Independent**
4. **Homoscedastic**

BASELINE RESULTS 1st Model

Model Performance: this is really low.

Dep. Variable:	price	**R-squared:**	0.413
Model:	OLS	**Adj. R-squared:**	0.412
Method:	Least Squares	**F-statistic:**	1922.
Date:	Sun, 02 Oct 2022	**Prob (F-statistic):**	0.00
Time:	13:11:32	**Log-Likelihood:**	-4.4743e+05
No. Observations:	30111	**AIC:**	8.949e+05
Df Residuals:	30099	**BIC:**	8.950e+05
Df Model:	11		
Covariance Type:	nonrobust		

MAE = 381082.49

RMSE = 687128.08

At least the model is statistically significant overall.

These errors are really high!

Modeling

Primary testing method: A/B testing

A: experiment group
B: control group

again
and again
and again
and again……

MODEL SUCCESS FACTOR

USED **LOCATION** VARIABLES
TO CREATE MORE USEFUL VARIABLES

Address → **zip code**

Latitude, Longitude

Imported data from King County GIS downloads

Elementary Schools
Business Centers
Medical Clinics

Homes Sold

distance between

FINAL RESULTS

This is really good!!!!

Dep. Variable:	price	**R-squared:**	0.779
Model:	OLS	**Adj. R-squared:**	0.778
Method:	Least Squares	**F-statistic:**	808.4
Date:	Sun, 02 Oct 2022	**Prob (F-statistic):**	0.00
Time:	13:13:03	**Log-Likelihood:**	-4.0955e+05
No. Observations:	29200	**AIC:**	8.194e+05
Df Residuals:	29072	**BIC:**	8.204e+05
Df Model:	127		

MAE = 198964.29

RMSE = 298567.17

Much lower errors!

Lots of factors though

This is All great stuff, but.....

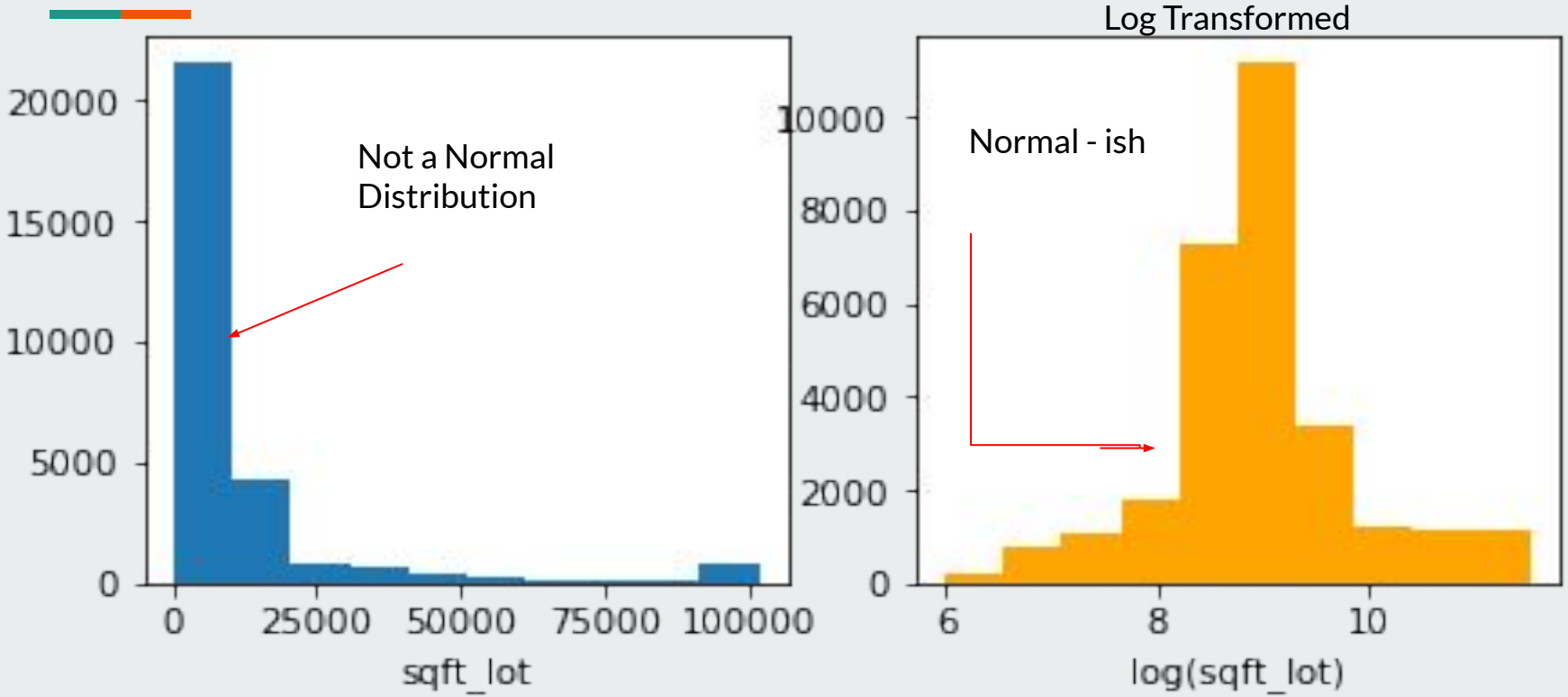
the Model did not pass tests for:

- Normality
- Independence
- Homoscedasticity
- Linearity

Solution

Log Transformation

Sqft Lot - Log Transformation



Now our Data Residuals are both Linear and Homoscedastic

Unfortunately, the residuals are still do not pass tests for normality and Independence.

Recommendations

1. Keep developing the OLS model
    a. Refine the model, try to obtain independence of features
    b. control for relationships that are not linear.

2. Consider developing alternative models
    a. Gradient Descent
    b. Decision tree based model

3. Consider developing a model more appropriate for evaluating non-linear relationships.

GOOD LUCK!
THANK YOU FOR THIS OPPORTUNITY

JONATHAN P. NEWCOMB, ESQ.
10/2/2022
Email: JNEWCOMB 54@GMAIL.COM
Linkedin:
Github: https://github.com/newcojon