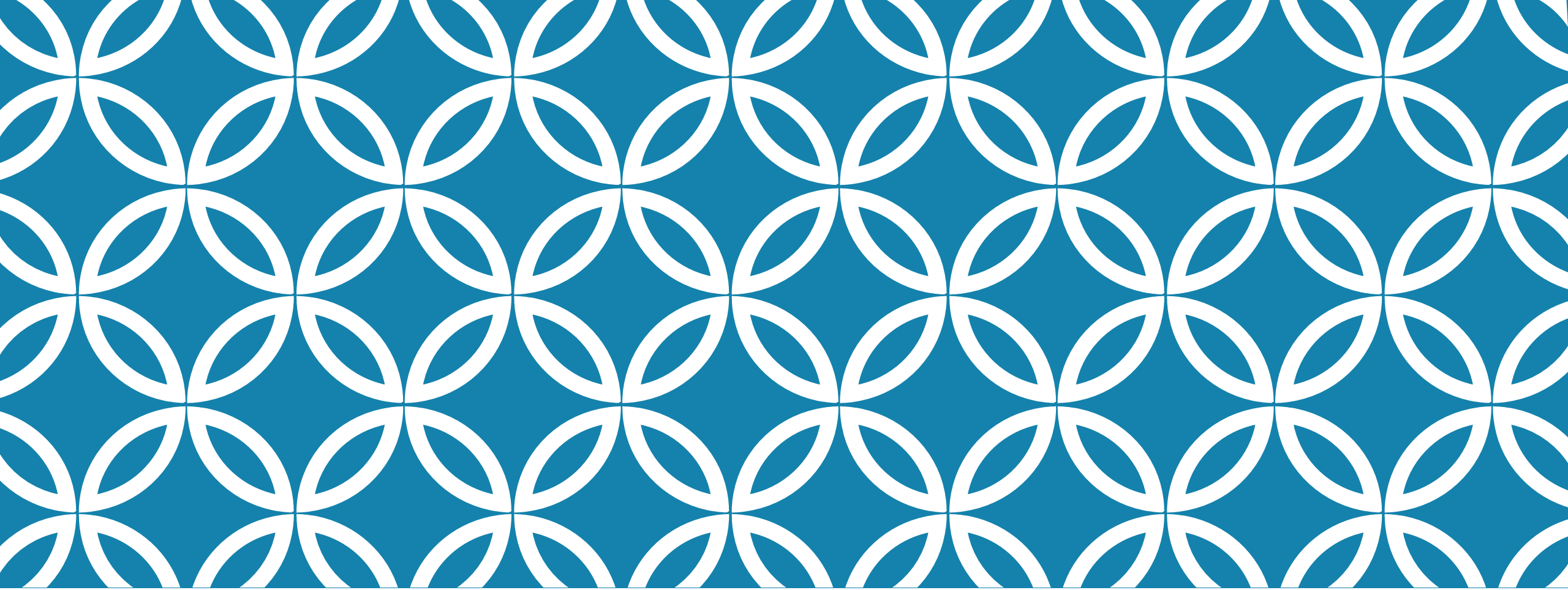

PROJECT 1

MaryJo Zaborowski

DS 04222019

WHAT WAS THE PROJECT?

Data Science Project 1 involves the assessment of a data set from King's County, Washington, containing house sales data. The objective of this study is to develop a predictive model to support house pricing, using the techniques of linear regression.

My philosophy was to rigorously clean the data up front, to enable accelerated analytics.

My plan:

1. Inspect the data
    a. review statistics (missing data,duplicates,extraneous values)
2. Clean data and review content
3. Check the data distribution
    i. plot histograms with KDE
    ii. view predictors vs. target with scatter plots
4. Run a regression to establish a model
5. Adjust the model to improve the predictive result by adding other independent variables and/or transforming the existing variables to improve upon their heteroscedascity and skew

WHAT WAS THE RESULT?

A linear regression model was developed using sqft_living to predict the sales price, however as a predictor of sales price, sqft_living performs inadequately.

The variable histogram showed skew and after running the model, R squared is 0.492, which means 49% of variance in house price(target variable) can be explained using the living square footage.

The residuals were conical in the plots, indicating heteroscedasticity, which breaks the assumption of normality in the data. This is confirmed by residuals in the QQ plot where the points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal($X \sim N(0,1)$).

A second multivariate model was developed using log-transformed values for sqft_living and adding a new variable of bedrooms per bathroom. This model returned an R squared value of 0.361, which performs less predictively than the prior model.

Conclusion: These models are not predictive. Further transformation of the data and use of additional variables would be required to improve on these

DATA IN DETAIL

Model 1 vs. Model 2 OLS Regression Results

Model 1

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.492
Model:	OLS	**Adj. R-squared:**	0.492
Method:	Least Squares	**F-statistic:**	2.087e+04
Date:	Tue, 07 May 2019	**Prob (F-statistic):**	0.00
Time:	16:43:09	**Log-Likelihood:**	-2.9911e+05
No. Observations:	21533	**AIC:**	5.982e+05
Df Residuals:	21531	**BIC:**	5.982e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.218e+04	4404.621	-9.575	0.000	-5.08e+04	-3.35e+04
sqft_living	279.9379	1.938	144.475	0.000	276.140	283.736

Omnibus:	14581.827	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	516123.571
Skew:	2.781	**Prob(JB):**	0.00
Kurtosis:	26.331	**Cond. No.**	5.63e+03

Model 2

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.361
Model:	OLS	**Adj. R-squared:**	0.361
Method:	Least Squares	**F-statistic:**	6082.
Date:	Wed, 08 May 2019	**Prob (F-statistic):**	0.00
Time:	08:45:17	**Log-Likelihood:**	-3.0158e+05
No. Observations:	21533	**AIC:**	6.032e+05
Df Residuals:	21530	**BIC:**	6.032e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.981e+06	7.8e+04	-89.508	0.000	-7.13e+06	-6.83e+06
sqft_living	3.752e+06	3.76e+04	99.902	0.000	3.68e+06	3.83e+06
bed_per_bath	-3.289e+04	3244.179	-10.140	0.000	-3.93e+04	-2.65e+04

Omnibus:	19438.802	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1680636.367
Skew:	4.012	**Prob(JB):**	0.00
Kurtosis:	45.530	**Cond. No.**	125.

WHAT DID I LEARN?

I was disappointed to not be further along in the exploration and analytics of this data. I feel I understand how to improve upon my result, but did not have the time (or skill yet) to execute as I planned. However, I gained some important learnings:

1. Creating a plan first: This was a good step that I took, allowing me to remain on task to the needed steps
2. Rethink time allocation: Time allocated (time management) between data cleaning and model refinement was not well balanced.
3. Build competency in dataframe manipulation: A majority of time and failure in the project was due to my inability to create the right code to transform data as desired. This was true in early data exploration (I attempted to use and transform the view, and location columns (zipcode, lat, and long). After unproductive hours, I moved on.)
4. Study problems solidify learning: This exercise was helpful to integrate learning about each of the course lessons, particularly python, list comprehension, linear regression and multivariate regression.