

# ***Phase 2 Project Launch***

Linear Regression Modeling  
of King County Real Estate Sale Prices

May 27, 2021

// **FLATIRON SCHOOL**

# Predicting Real Estate Sale Prices in King County, WA

## Inferential Linear Regression Project

- Build a model aligned to **business questions** that explains as much of the **total variance** of real estate sale prices in King County while **reducing violations** of the assumptions of linear regression.

## Project Assignment

- Note: Rubric is similar, with a few changes.

# CRISP-DM



# Business Understanding: Stakeholder

- Choose a stakeholder
- Example: a real estate agency that helps homeowners buy and/or sell homes.



## Business Understanding

- Brainstorm with partner about what features of a house drive sale price
- Link to stakeholder priorities
- Add features to model according to these priorities



# Business Understanding: Recommendations

- Include **at least 2** important parameter estimates in your final recommendations
- Interpret effect for stakeholder

>=



# Exploratory Data Analysis

## - Pandas

- Shape of the data
  - How many records?
  - How many features?
  - NA inspection
- Distinguish between continuous/categorical variables

## - Maps

- Folium/Geopandas

## - Visualization (MPL/Seaborn)

- Histograms
  - Bar charts
  - Boxplots for outliers
  - Scatterplots
  - Pairplots
  - Heatmaps
- 
- **Your final notebook must include 3 high quality visualizations of relevant findings**

# Data Preparation/Feature Engineering

## General Data Prep

- NAs/outliers/duplicates
- Scaling
- Log-transformations

## Feature Engineering

- One-Hot-Encoding/Label Binarizing
- Polynomial features and interactions
- Custom transformations



# Iterative Model Building: First Simple Model

- Poor performance model to check your process (i.e. you are able to generate valid prediction)
- Baseline  $R^2$
- Ex: Phase 2 FSM: Simple linear regression with 1 highly correlated feature.



# Iterative Model Building: Beyond the FSM

- Add **complexity**
- Add features aligned to **business questions**
- **Document** your progress through the iterative process



# Assumptions of Linear Regression

- With each iteration of your model, check whether your model violates the [assumptions of linear regression](#).
- It will be difficult to create a model with high  $R^2$  without violation. If your final model violates some assumptions, that is ok. However, you have to demonstrate that you have checked the assumptions, and made attempts to adhere to them

# Linear Regression: Valid Inferential Models

## The Assumptions of Linear Regression

Linear Relationship

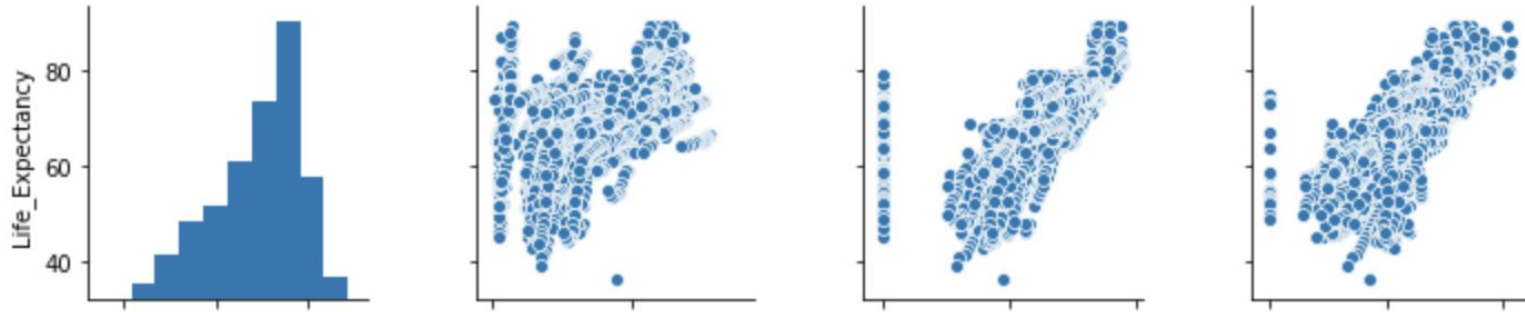
No Multicollinearity

Homoscedasticity

Normal Distribution of Errors

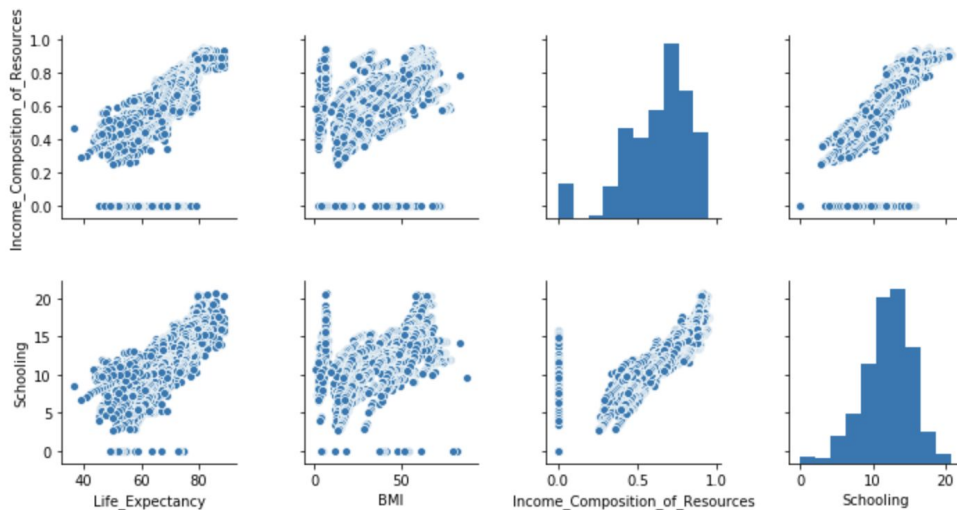
# Assumption #1: Linear Relationship between Dependent and Independent Variables

- Scatterplots that show positive or negative correlation
- Numerical correlation between target and predictors **.corr()**



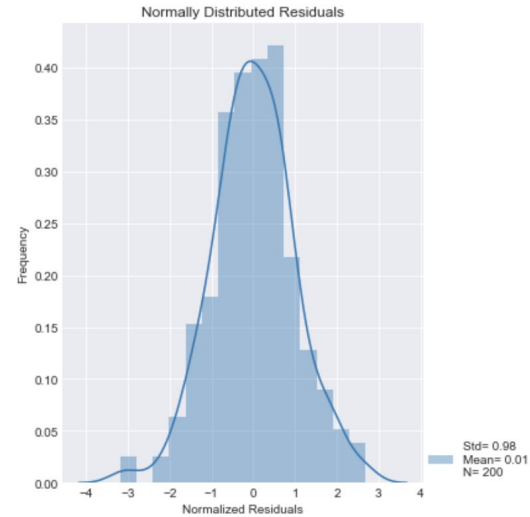
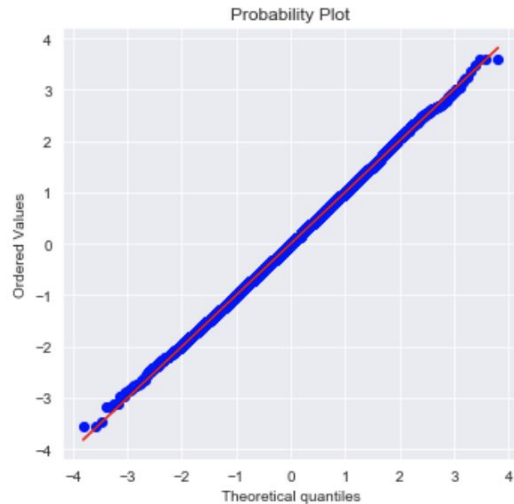
## Assumption #2: Low Multicollinearity Between Independent Features

- Scatter plots visualizing the correlation of independent features.
- Heatmap of correlation matrix
- `.corr()`



## Assumption #3: Normal Distribution of Errors

- Q-Q plot and histogram of residuals



## Assumption #3: Normal Distribution of Errors

### Quantitative Diagnostics

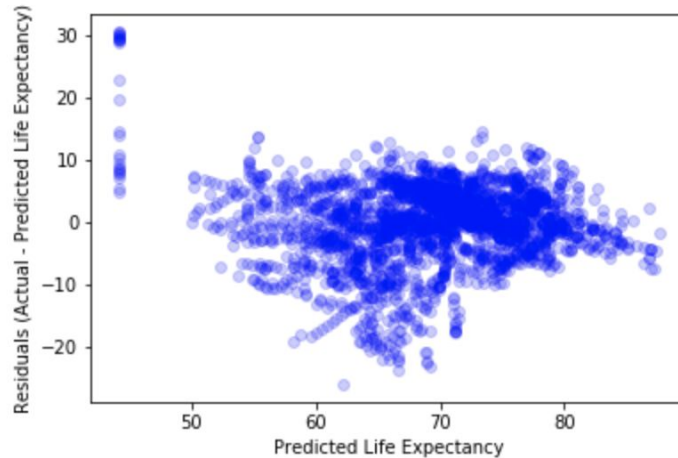
- The **Jarque-Bera**: The lower the score, the more normal the distribution of errors.

Omnibus:	283.391	Durbin-Watson:	0.267
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1122.013
Skew:	-0.445	Prob(JB):	2.28e-244
Kurtosis:	5.989	Cond. No.	46.7



## Assumption #4: Homoscedasticity of the Errors

- Linear regression assumes that the variance of the dependent variable is homogeneous across different value of the independent variable(s). We can visualize this by plotting the predicted values vs. the residuals.



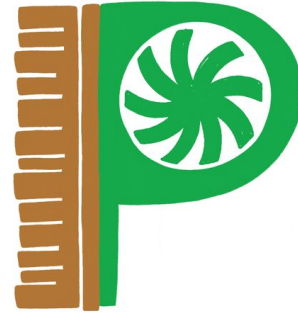
# Assumptions of Linear Regression

- [Linear Regression Cumulative Lab](#)



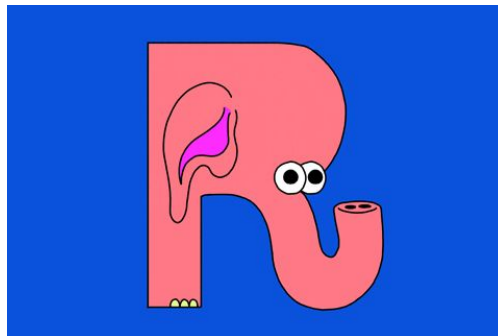
## Calculated Coefficients' P-Values

- Don't forget those P-Values!
- Check whether your betas are statistically significant.
- If a coefficient's p-value  $> .05$ , you should not include it in your final model.

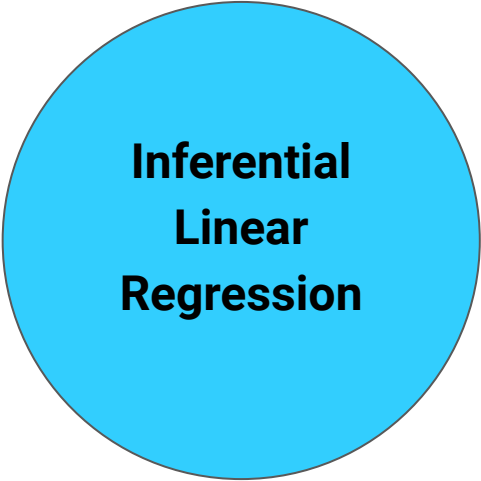


# Maximize R-Squared

- Don't forget about  $R^2$
- Create a model that explains a large amount of the total variance
- A model that does not violate any assumptions, but that as a very low  $R^2$ , has little value



# Remember This?



**Inferential  
Linear  
Regression**



**Predictive**

# **Bake Off Branch**



# ***Group Assignments and Schedule***

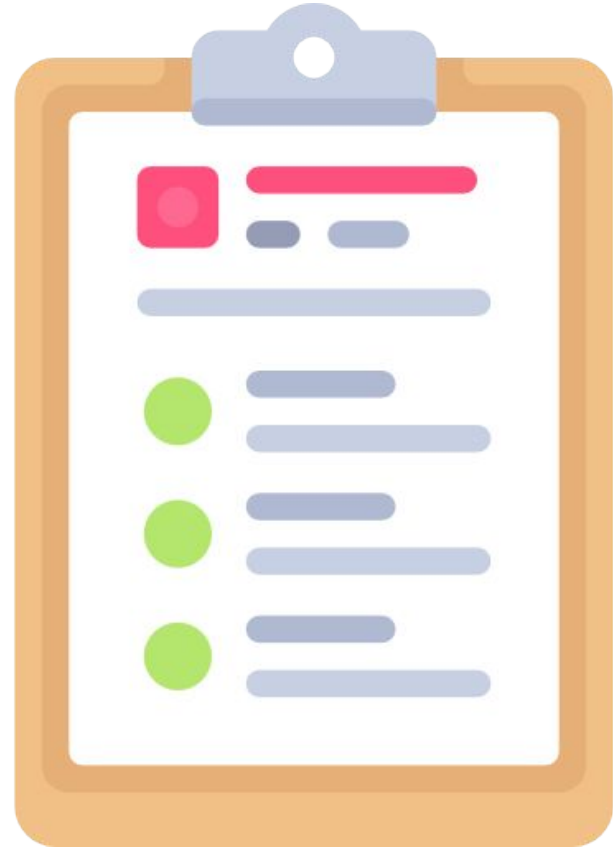


# ***Group Project Best Practices***

1. Get to Know Your Partner
2. Define Individual Project Contributions
3. Meet Regularly
4. Communicate Actively, Clearly, and Transparently

Credit to Dr. Grace Oh, University of Illinois College of Education ([website](#))

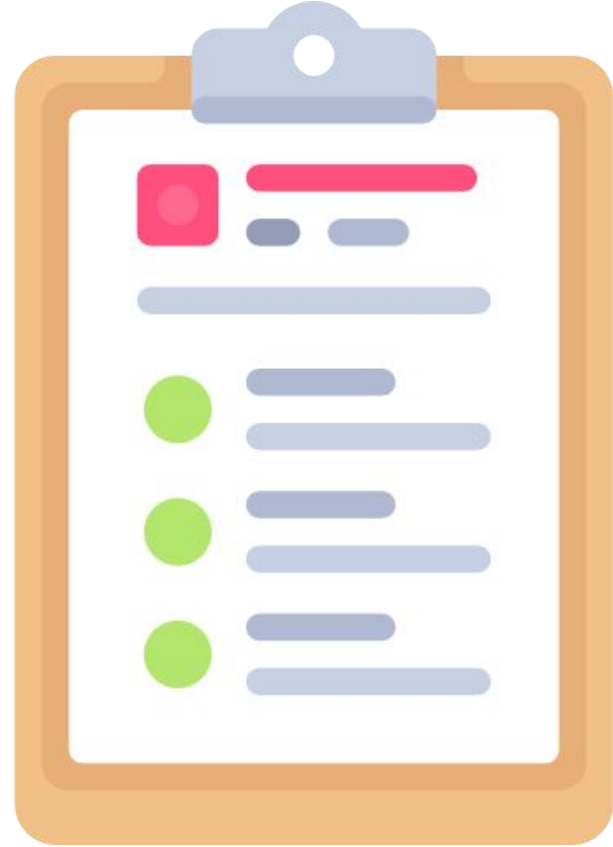
Icons made by [Freepik](#) from [Flaticon](#)





# ***Project Contract***

1. Make a copy of this [Google Doc](#)
2. Slack to instructor by EOD Friday



## Project Groups

- Group 1: Ryan, Aaron
- Group 2: Brian, Erin
- Group 3: Victor, Brad
- Group 4: Kyle, Kevin
- Group 5: George, Ramil

## Project Events

- **Friday (today):** Kickoff, contracts, project work
- **Monday:** Flatiron Closed for Memorial Day
- **Tuesday:** Project Work
- **Wednesday AM:** Group Check-ins (sign-up slots on calendar)
- **Thursday AM:** Non-Technical Dress Rehearsals
- **Friday AM:** Non-Technical Presentations
- **Friday 5 PM PST**
  - Submit link to repo and completed deliverables
  - Submit Bake Off Predictions to Slack thread

# Deliverables

- **There are 3 Deliverables for this Project**
  - **Github repository**
    - README.md
    - Clear history of commits
    - Work organized in clear file structure
  - **Final Notebook**
    - One clean notebook which consolidates individual team members' work
  - **Non-technical Presentation**
    - Aim for 5 minutes total

***Questions?***