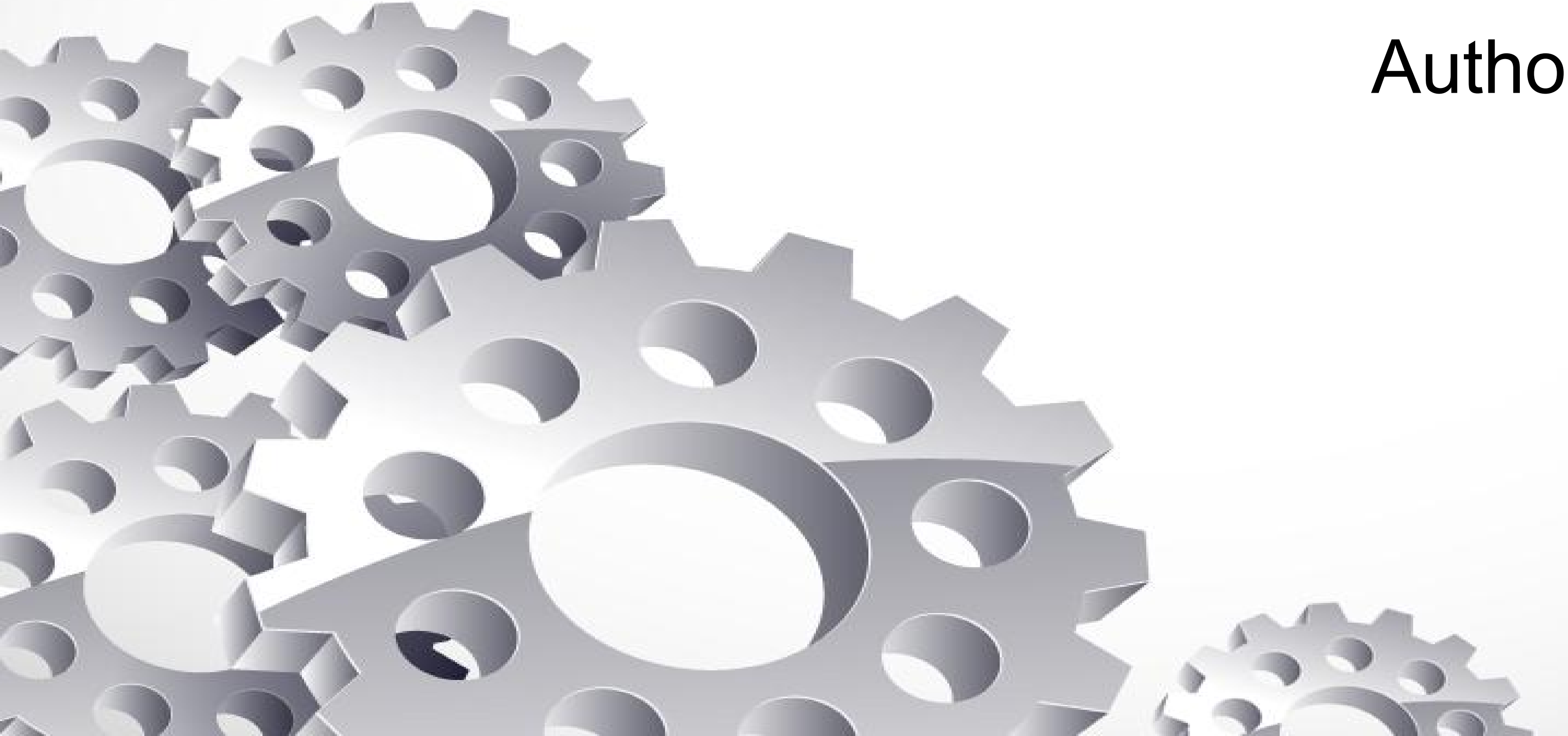
Data science Phase 2 Project

Author: Nobert Akwir



Business Understanding

- The project is going to be to develop a pricing algorithm to help set a target price for houses that are to be bought and/or sold by homeowners through the real estate agency. The goal is to save the real estate agency and prospective/current homeowners some time and to help ensure consistency in pricing of houses in the area.

Data Understanding

- This project uses the King County House Sales dataset, which can be found in kc_house_data.csv in the data folder in this repo. The description of the column names can be found in column_names.md in the same folder. As with most real world data sets, the column names are not perfectly described, so we'll have to do some research or use our best judgment if we have questions about what the data means.

Methods.

- We are trying to find the best multiple linear regression model to assist us in predictive purposes of house prices in King County.

Results

- 1. Our Y-intercept, indicates that the base price for a house in King county, with all other dependent attributes at 0, should be 6765900.901904495.

- 2. For every unit increase in bathrooms, sqft_living, floors, condition and grade, the price of the houses increase by 50109.359950, 182.028980, 20835.586811, 18406.181028, 128614.320549 respectively.

- 3. Having a waterfront at the house, increases the price of the house by 697976.729420, which shows this is the most expensive aspect of the houses.

- 4. For every unit increase in bedrooms, footage of the lot(sqft_lot) an year_built, the price of the houses decrease by 42144.849922, 0.257080, 3879.881453 respectively

Conclusion and Recomendation

- Basically, from our little eperiment, we shouldnt really rely on our final model as the final say in pricing the house as we have seen we are violating most of the assumptions of linear regression. This was also seen from our Mean Squared Error value earlier. Therefore, our model should be purposed for predictive purposes only. We should also work on looking on the outliers in our variable since they are really affecting our model. Though purposely leaving them on since in the real world, this are truly the prices of the houses, but it seems that handling them in a better way may definitely lead to better model.