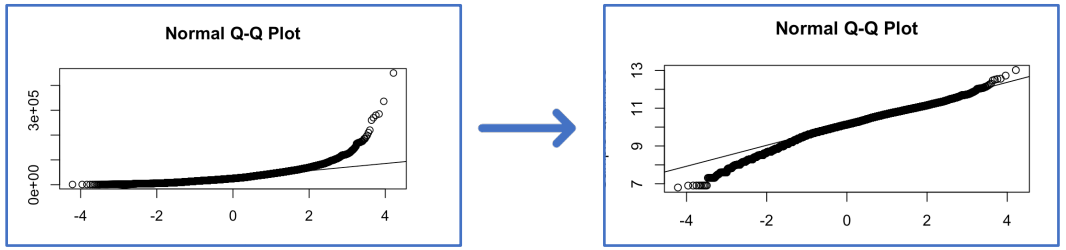
Data Processing and Analysis for Car Price Prediction

1. Exploratory Data Analysis (EDA)

- ✓ Handled missing values based on the missing proportion
  - **Numerical Variables:** median & linear regression imputation
  - **Categorical Variables:** mode & decision tree imputation
- ✓ Dropped non-essential variables **(e.g., id, description, wheel_system_display)**
- ✓ Log-transformation of skewed data **(e.g., price)**



2. Feature Engineering and Selection

Categorical Variables :

- ✗ Removed variables with large number of missing values
- ✓ Transformed categorical variables into factors
- ✓ Consolidated similar types into 'Other' for model simplicity

Numerical Variables:

- ✗ Removed variables with low correlations to the response variable
- ✓ Removed variables that are highly correlated with other predictors

3. Model Building

RMSE: **1,293.21313** Private leader board: **14**

4. Future Improvement

- Refine the feature elimination strategy to avoid removing potentially predictive variables
- Employ methods such as interaction terms in regression models and feature importance in tree-based models for deeper insights