Seattle Airbnb Listings

Airbnb Price Prediction in and around Seattle Area

Saisri Potluru – *saisri.m.potluru-1@ou.edu*| Prasuna Mitikiri - *prasuna.mitikiri@ou.edu*

DSA 5103 Intelligent Data Analytics

The University of Oklahoma, Norman

Executive Summary:

Airbnb is a popular and fastest growing online platform through which the landlord can rent their properties to tenants for a short-term stay. Though it does not own any real estate listings, it acts as a dealer to receive commissions for each booking. It provides a short-term rental platform for people across the world. The people who want to get benefit from the use of underutilized resources can list their homes or apartments and allow others to rent their property. In cities like Seattle, where Airbnb is one of the options, travelers can find more affordable accommodations through rental platforms to choose among the best. Hosts must decide to price for their property on a daily basis based on its features with the help of Airbnb dealers.

The dataset used for this project has been extracted from the Kaggle repository which covers 7,576 Seattle listings. It provides the details of home features, reviews, latitude, longitude, address, and customer satisfaction until the end of 2018. The datatypes of the attributes range from categorical, numeric, double, factor, etc. Some features are found skewed towards the right. As the goal is to predict price and there lies a huge difference between the value of the observations, some features have been transformed to avoid getting some unusual results. Latitude and longitudes are the variables which helps to identify the number of Airbnb listings available in Seattle and its neighboring places. By spotting the locations on the map, it is found that the competition among Airbnb stays is in its peak especially in the city of Seattle when compared to its surrounding areas.

The collected data is relatively small, so random sampling has been performed to split the data as test and train sets. As the predictor variable is a continuous datatype, the initial algorithm that is preferred to train the model is a popular and flexible multiple linear regression model. The model worked well when it is applied to the test dataset with an error rate of 0.42. However, the non-parametric supervised regression technique “Multivariate adaptive regression spline” yields best results over multiple linear regression, Lasso, Ridge regression techniques. The key to success is to identify salient features that have a high impact on the target variable. For this dataset, out of eighteen features, around four are found useful.

The models that are built use important features. For future work, combining attributes like latitude, longitude with the preprocessed precise area to create new features that may assist in gaining useful information that could potentially help landlords in gaining profits as well as competing with competitors. Collecting contemporary features along with updated details may give additional insights on listings. Also, including features like restaurants, recreational places that utilize latitude and longitude to derive distance information can be an add-on to the dataset as well as beneficial for the owners to have insight on the price to be decided.

Problem Description and Background:

Travelling and recreation are one of the thriving areas where people from all over the world have been traveling to many places for their own benefit no matter their economic status. With this new Airbnb staying keeps on growing. Especially in large cities like Seattle, house owners would like to get benefit from their own properties. This may affect the existed owners. As it is mentioned that, hosts must decide the pricings, it would be difficult for them to come up with a precise pricing structure. Keeping more price may lead to losing the customers and setting less price may increase the customers but that may not give any profits for the owners as they need to pay taxes, commissions, etc. Moreover, several years ago, Airbnb used to suggest listings to the homeowners for free. But now they started taking commissions for their suggestions. This leads them to recommend the property with accurate value is crucial.

In contrast to the conventional problem, price optimization helps homeowners who share their property on Airbnb to set the optimum price for their listings. The entire work has been designed and implemented from the supply-side perspective. This dataset has been used in the Kaggle competition to predict the price. As the dataset is not enough to predict the values, there are not many models that have been built. Considering this as an opportunity to develop a model, this dataset has been chosen. The main goal of this project is to develop a model using a machine algorithm to aid the property owners by utilizing the minimal data that is available about the property.

Exploratory Data Analysis:

Understanding the structure of data is crucial steps to begin with. As it is mentioned the dataset has eighteen variables and around eight thousand observations. The enlisted variables below help to have an idea about the available data.

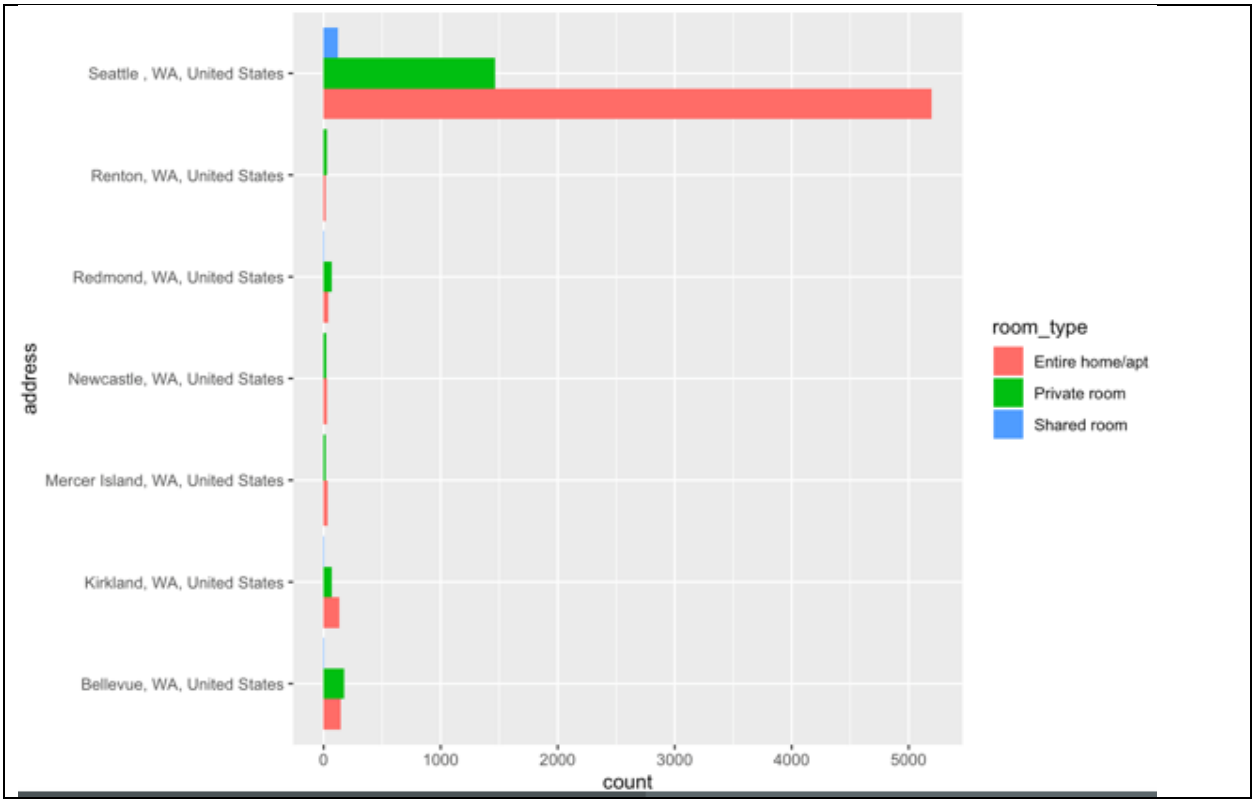
Attribute	Description
Room_ID	Unique room IDs
Host_ID	Owner IDs
Room_type	Type of the room (Entire home/ Apt, Private Room, Shared Room)
Address	City with State (WA)
Reviews	Number of reviews by customers
Cust_Sat:	Ratings given by the customers based on their satisfaction levels
Accommodates	Room capacity
Bedrooms:	Number of bedrooms
Bathrooms:	Number of bathrooms

Price	Price per night
Last_Modified	Time and date when the data is last updated
Latitude & Longitude	Latitude and Longitude of the Airbnb area
Name	Apartment / Building names
Currency	Currency in USD
Rate_Type	Nightly stay

While doing exploratory data analysis, several attributes are found to be not useful for building a model. Therefore, those trivial attributes have been removed. The initial step that performed is locating the listings on the map to see the distribution and crowd of the Airbnb listings. The below map shows the Airbnb lodgings.

The points on the above plots have been retrieved using latitude and longitude. The maximum number of observations taken from Seattle city. However, some other neighboring areas have also been considered for price prediction. By observing the cities under “address” variable, it can be found that the same cities have been considered as different ones because of the delimiters. This is handled later in the data preparation.

Room_type is one of the considerable attributes. It is found that the most popular room_type in the Seattle area is “Entire home/ apartment. Private and shared rooms are not that popular. The plot also provides several insights such as setting high rents to “entire home/ apartment” type rooms that may potentially decrease the number of customers. The predicted prices may help owners to avoid such conditions.



The dataset has some missing values especially in cust_satisfaction variable. The proportion of missing values in that variable is almost 19%. “Bathroom” has almost negligible amount of missingness. These missing values have been treated in the data preprocessing step. As almost every dataset has outliers, the Airbnb price listing has also outliers. Representation of the outliers and treatment can be found in the further discussion.

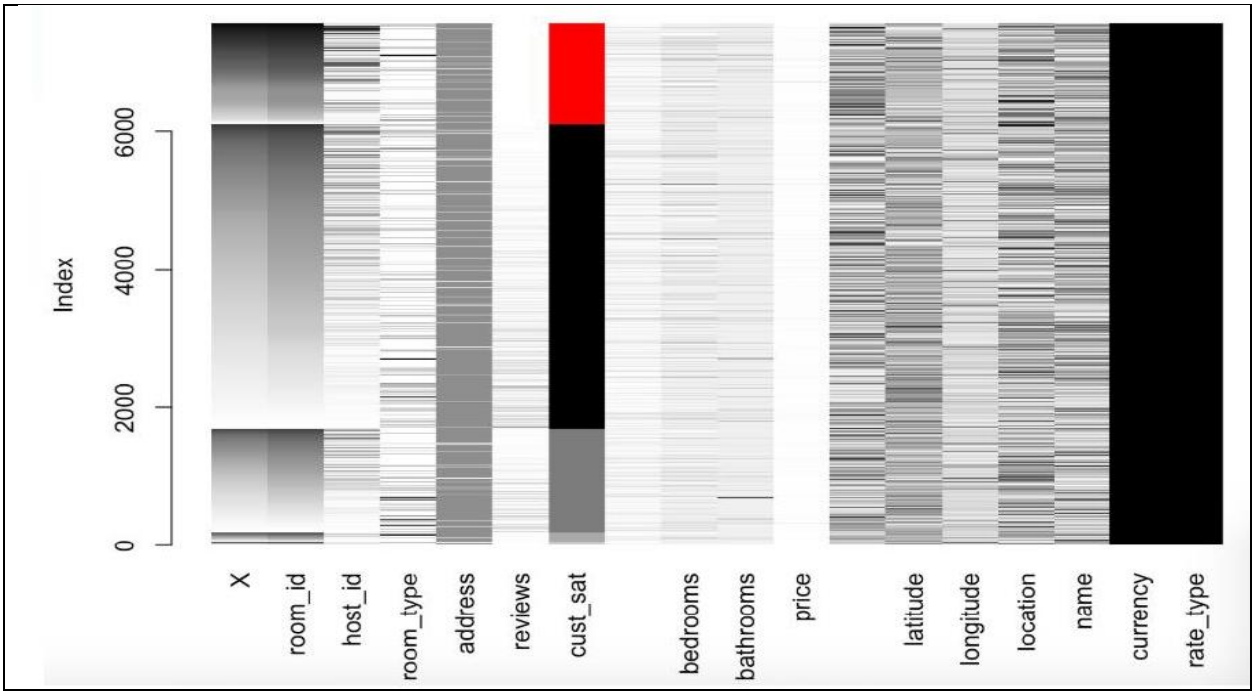
Analysis Plan:

Regression Techniques that have been chosen to build the model are multiple linear regression, lasso, ridge, and MARS. Since the dataset consist of many continuous variables, the above-mentioned supervised learning techniques work better in training the model. Since the model is small, it may suffer from overfitting when applying the built model on the validation dataset. However, the optimum model that is built neither overfitted nor underfitted when testing on the validation data. The strength of the model is choosing linear “regression models”. They may handle the available dataset with respect to datatypes of the features.

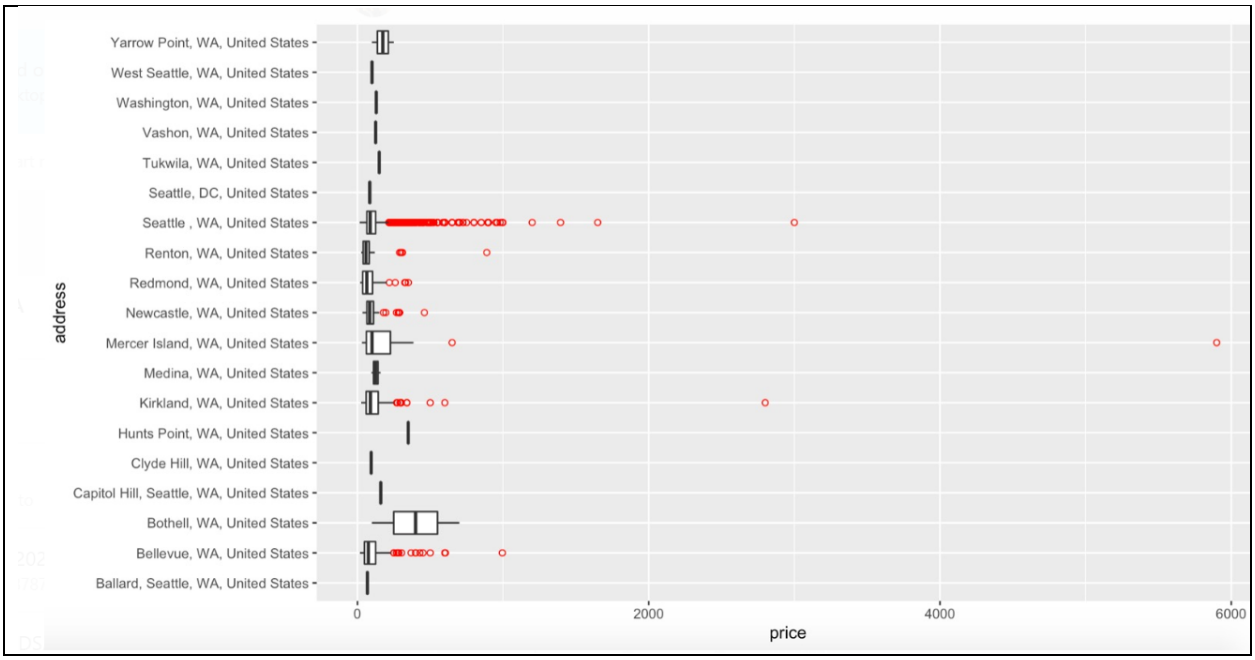
Feature Selection, Missingness, Treating Outliers:

Among eighteen available features, few of the features are found helpful. Unfortunately, many features have been removed as they provide trivial and redundant information such as Currency, location (Consists of sequence digits which does not mean anything) , latitude & longitude (the dataset already have equivalent locations in the address variable), name (name of the building which helps positively when working with text analysis) , last_modified (time, date and year of the data being updated lately which is same for the entire dataset), rate_type ("nightly" for all observations).

Cust_sat variable has missing values up to 19% and they are imputed with new factor level. Because, imputing with mean or mode may mislead the model since the behavior of the customer can not be predicted as there is no other supporting data available in the dataset. "Bathroom" has only one missing value that is imputed with mode.



Having outliers with respect to small dataset is a big drawback. The predictor variable has several outliers. Most unusual outliers have been removed from the dataset. Those observations can be found below.



The presence of delimiters, special characters, language (Chinese) and spacing in the address variable considered two same cities as different ones. of as different cities. To avoid such conditions, those have been assigned with the appropriate cities.

For modeling, among available dataset four variables are chosen as an independent variable for the model formulation based on their correlation within the variables. Among all, cust_satisfaction and accommodation are the ones which are having highly correlated with the predictor variable. Therefore, the model has been formulated with these significant attributes.

Modeling:

To build the model, the dataset is divided into test and train with 30% and 70% using random sampling. As the separation done based on the random sampling, the presence of several unique factor levels in the data threw error. This has been handled by removing the unique level. Among chosen modeling techniques, MARS performed better on the test data with 0.41 error rate. MARS is A non-parametric regression model that evaluates all potential groups across the range of predictor value by partitioning the data. An efficient algorithm which detects the interaction between features.

Number of built models and their respective error rates are given below.

MODELS	TRAIN DATA	TEST DATA
Linear Model	0.4237017	0.4141882
Lasso Model	0.4245293	0.4141882
Ridge Model	0.425754	0.4164112
MARS Model	0.4191478	0.4131458

In the data preprocessing step, while checking for the distribution, some variables did not show proper distribution. Including them for building model leads to a greater error rate. So, applied logarithmic transformation while building the model.

Mars model has been with the cross-validation function, which takes several iterations and run the model based on the provided value and produces optimal values that works better among all. For this built model, the optimal values of parameters degree and prune are 3 and 23 respectively. For the performance evaluation, Root Mean Squared Error has been considered. The performance of the built models is relative similar and varies with negligible value.

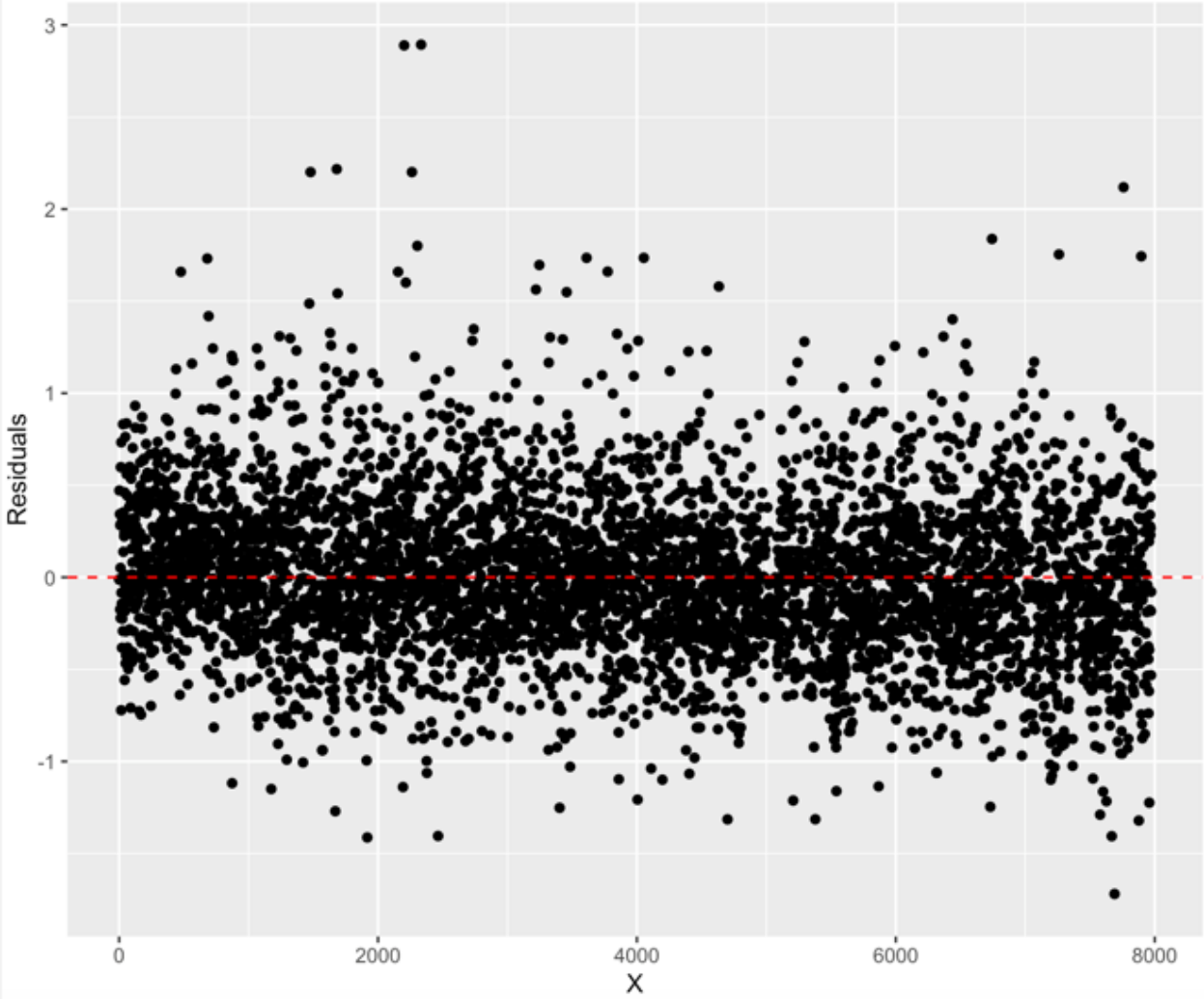
Performance Evaluation:

```
> rmse(log(testing$price),predictM8)
[1] 0.4141882
```

Below table provides the list of models have been built and their corresponding values.

Model	Method	Package	Hyperparameter	Selection	R squared	RMSE
Linear	lm	stats	NA	NA	0.493	0.42
Lasso	glmnet	glmnet	lambda	0.001	0.494	0.426
Ridge	Glmnet	Glmnet	lambda	0.001	0.493	0.426
MARS	earth	earth	degree	3	0.478	0.414

Residuals:



Plot shows the distribution of the residuals with respect to actual values. The residuals stays closer to the horizontal red line (represents the actual values) indicated the better built model. In the above case, some residuals spread away from the actual line, which mean the built model still have some error rate which is shown earlier.

Conclusion:

The predicted model helps the hosts in setting the optimal price value for their listings to make a maximum profit by overcoming a conventional challenge of setting price every day. Also, considering the predicted values with respected to area, for the new owners who want to start investing in the real estate may help them to have some insights on their future profits. However, it seems, there is some important information may be overlooked during the modeling process. Future works on the removed data may help in doing additional feature engineering.

Future Work:

Since the dataset have text datatype for instance “name” attribute, using text analytics would convert the data into factor levels would be a good add-on in building the precise model.

References:

[1] Kaggle Repository

https://www.kaggle.com/shanelev/seattle-airbnb-listings

[2] Wikipedia

https://en.wikipedia.org/wiki/Airbnb

[3] Towards Data science

https://towardsdatascience.com/

Appendix:

Cross Validation:

Linear Model:

```
Linear Regression

5305 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 4548, 4546, 4546, 4549, 4547, 4547, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.4261284  0.4937235  0.3221334

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Lasso:

```
> lasso
glmnet

5305 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 4547, 4547, 4548, 4547, 4548, 4547, ...
Resampling results across tuning parameters:

  lambda    RMSE       Rsquared   MAE
  0.000100  0.4264426  0.4945412  0.3224391
  0.050075  0.4388009  0.4790430  0.3298630
  0.100050  0.4621293  0.4495995  0.3468485
  0.150025  0.4843547  0.4407105  0.3645010
  0.200000  0.5138596  0.4118067  0.3879860

Tuning parameter 'alpha' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 1e-04.
```

Ridge:

```
> ridge
glmnet

5305 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 4547, 4546, 4546, 4548, 4548, 4547, ...
Resampling results across tuning parameters:

  lambda  RMSE       Rsquared   MAE
  0.0001  0.4269714  0.4931513  0.3227769
  0.1112  0.4289863  0.4909642  0.3239891
  0.2223  0.4330317  0.4878331  0.3266229
  0.3334  0.4375868  0.4851022  0.3295085
  0.4445  0.4423150  0.4827145  0.3325799
  0.5556  0.4470489  0.4806093  0.3358923
  0.6667  0.4516911  0.4787454  0.3392171
  0.7778  0.4561935  0.4770840  0.3425542
  0.8889  0.4605313  0.4755954  0.3458411
  1.0000  0.4647023  0.4742446  0.3490892

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 1e-04.
```

MARS:

```
5305 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (3 fold)
Summary of sample sizes: 3536, 3538, 3536
Resampling results across tuning parameters:

  degree  nprune  RMSE       Rsquared   MAE
  1         2     0.4764950  0.3679816  0.3633844
  1        12     0.4243277  0.4988779  0.3194741
  1        23     0.4241127  0.4993855  0.3193461
  1        34     0.4241127  0.4993855  0.3193461
  1        45     0.4241127  0.4993855  0.3193461
  1        56     0.4241127  0.4993855  0.3193461
  1        67     0.4241127  0.4993855  0.3193461
  1        78     0.4241127  0.4993855  0.3193461
  1        89     0.4241127  0.4993855  0.3193461
  1       100     0.4241127  0.4993855  0.3193461
  2         2     0.4780355  0.3639390  0.3637177
  2        12     0.4237184  0.5004643  0.3191866
  2        23     0.4236388  0.5007591  0.3188749
  2        34     0.4236388  0.5007591  0.3188749
  2        45     0.4236388  0.5007591  0.3188749
  2        56     0.4236388  0.5007591  0.3188749
  2        67     0.4236388  0.5007591  0.3188749
  2        78     0.4236388  0.5007591  0.3188749
  2        89     0.4236388  0.5007591  0.3188749
  2       100     0.4236388  0.5007591  0.3188749
  3         2     0.4764950  0.3679816  0.3633844
  3        12     0.4262777  0.4946936  0.3206172
  3        23     0.4243053  0.4993010  0.3195885
  3        34     0.4243053  0.4993010  0.3195885
  3        45     0.4243053  0.4993010  0.3195885
  3        56     0.4243053  0.4993010  0.3195885
  3        67     0.4243053  0.4993010  0.3195885
  3        78     0.4243053  0.4993010  0.3195885
  3        89     0.4243053  0.4993010  0.3195885
  3       100     0.4243053  0.4993010  0.3195885
```