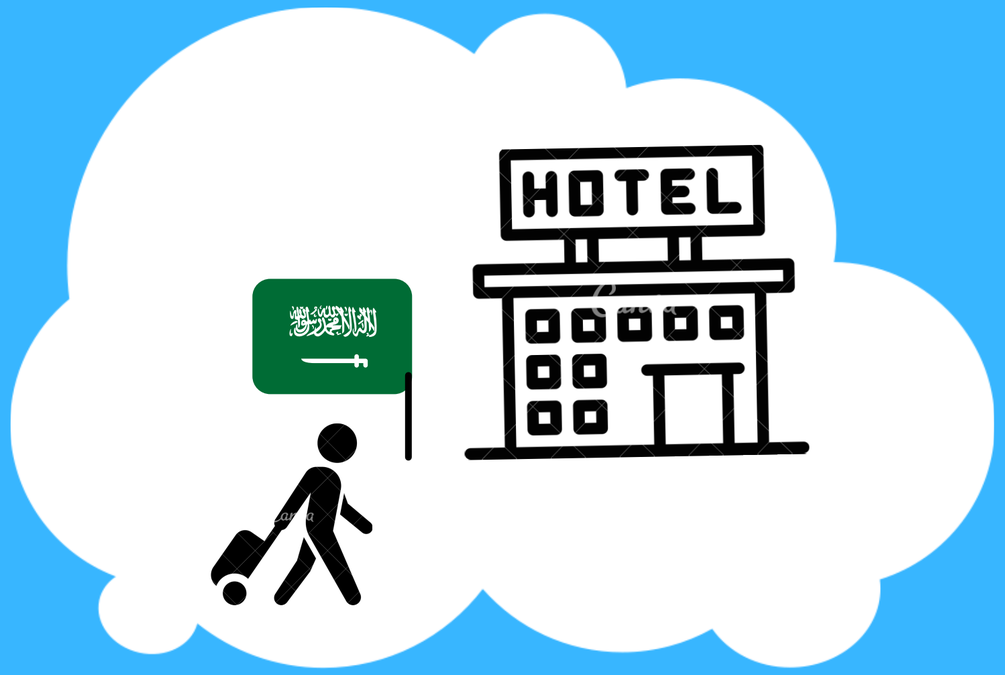
Hotel Room Price Predictor Booking.com

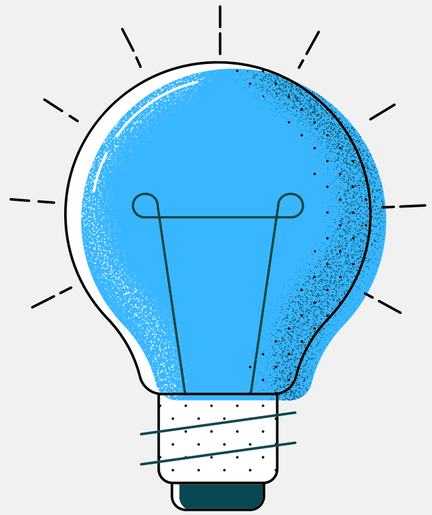
Prepared by:
Abdultawwab Safarji

Story Time



Motivation!

The Ministry of Tourism needs to understand the fluctuation in hotel room prices on and off.



Why?

To apply **price control** and make them competitive with the top countries attracting tourists around the world.

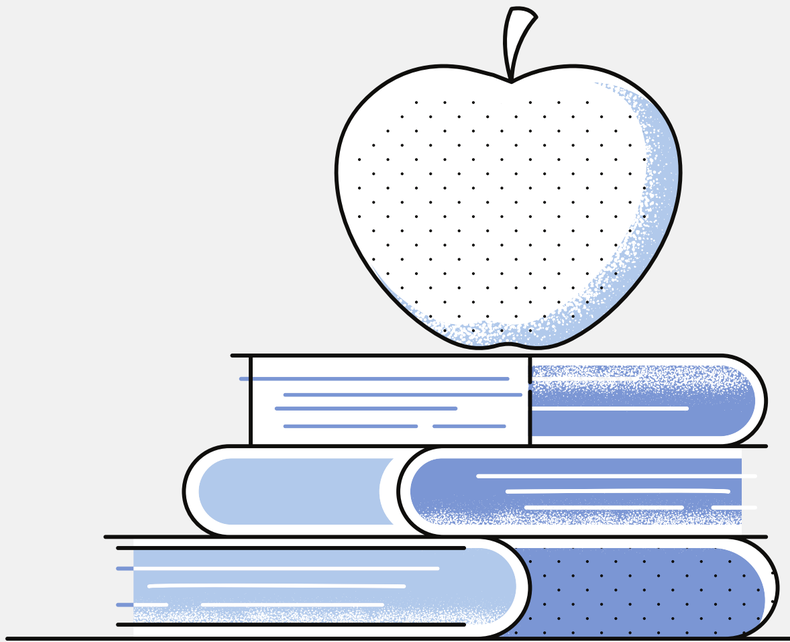
Problem Understanding!

Why Booking.com Prediction?

- Help tourists and gov to know hotel prices in KSA.
- Regulate prices and make them competitive.
- Provide price predictions for hotels in most cities.
- Discovering data error entries and outliers.
- Unstable search results in some cities.

Who will benefit from Booking.com analysis?

- Ministry of Tourism in Saudi Arabia
- Booking.com//sa
- Hotels Agencies
- Travel Agencies
- Customers



Plan of attack

1 ⟷ 2 ⟷ 3 ⟷ 4 ⟷ 5 ⟷ 6

Data Cleaning & Prepatation

Handle nulls, zero and duplicate values.

Data Exploration

Visualize data and apply descriptive statistics.

Dimensionality Reduction

Reduce number of input

Preprocess Data

Feature engineering & selection by applying transformations, scaling, and partitioning data to train, validate and test.

Build Models and test

fit several regression modeling algorithms with different combos of independent features, validate and test all models to get the best fit.

Develop The Product

develop a user interface to predict the price given relevant input.

Data Preparation

Booking.com

Using Booking.com to scrape the following features:

- Due to the time constraint as well as the unavailability of all types of data to collect, only the scraped dataset will be used.

1. Hotel Name
2. Location
3. Room Type
4. Price
5. Price for
6. Number of Beds
7. Hotel Rating
8. Rating Title
9. Number of Reviews
10. Room Size

location
Riyadh
Jeddah
Makkah
Madinah
Taif
Hail
King Abdullah Economic City
Abha
Tabuk
Buraydah
Dammam
Al Ula
Yanbu
Al Jubail
Turayf
Najran
Al Khobar
Afif
Jazan
Al Ahsa
Al Khafji
Al Baha

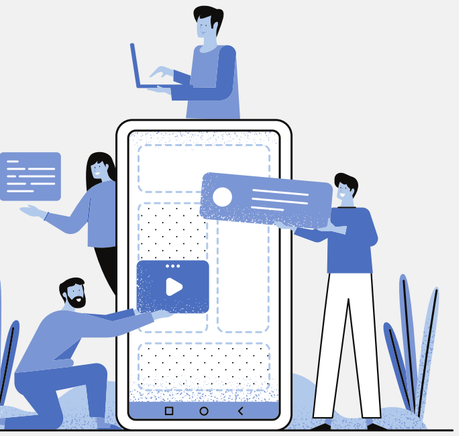
DATA VALIDATION

Validate for

Missing Values

Duplicate Data

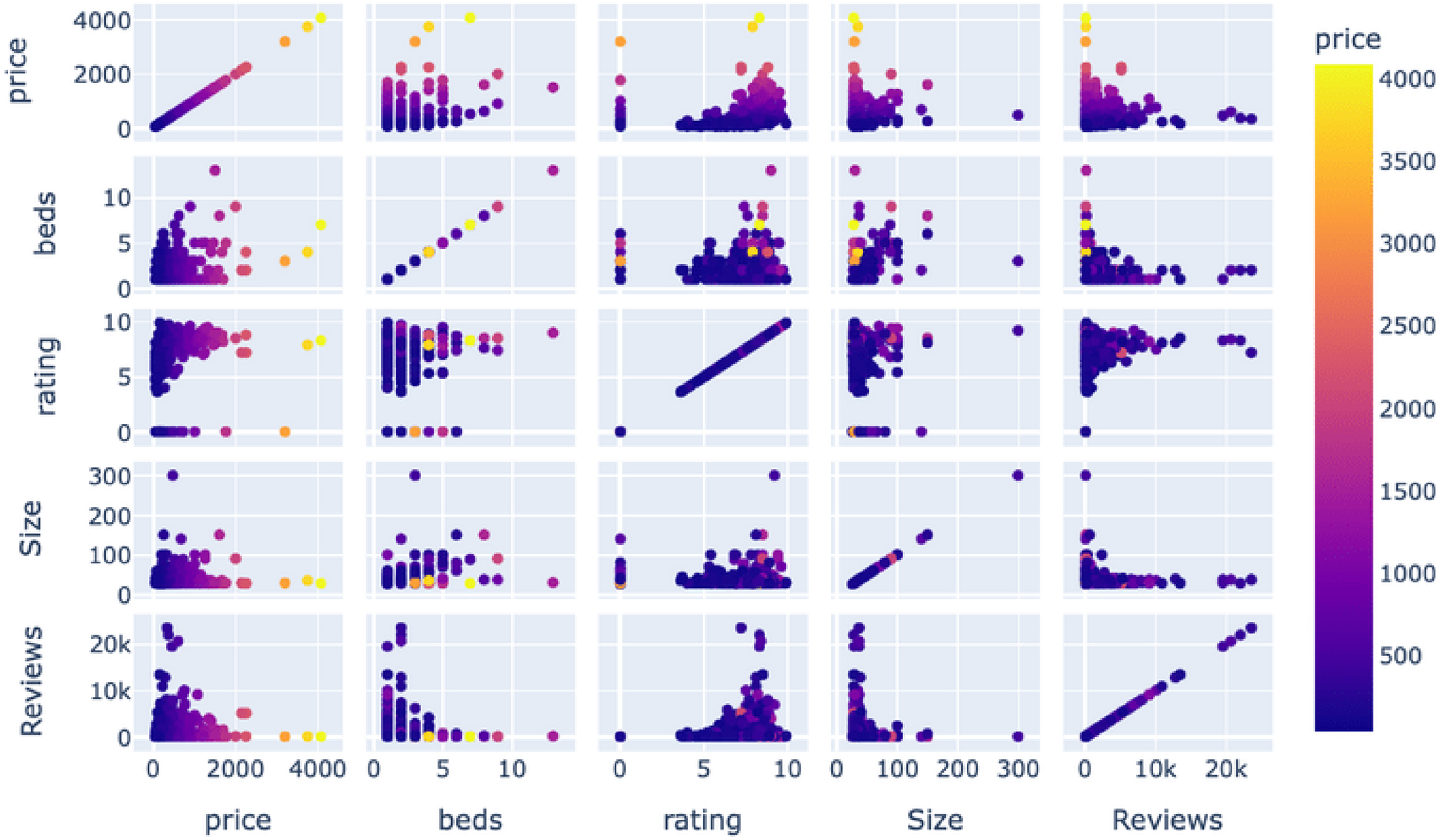
outlier and Illogical Entries



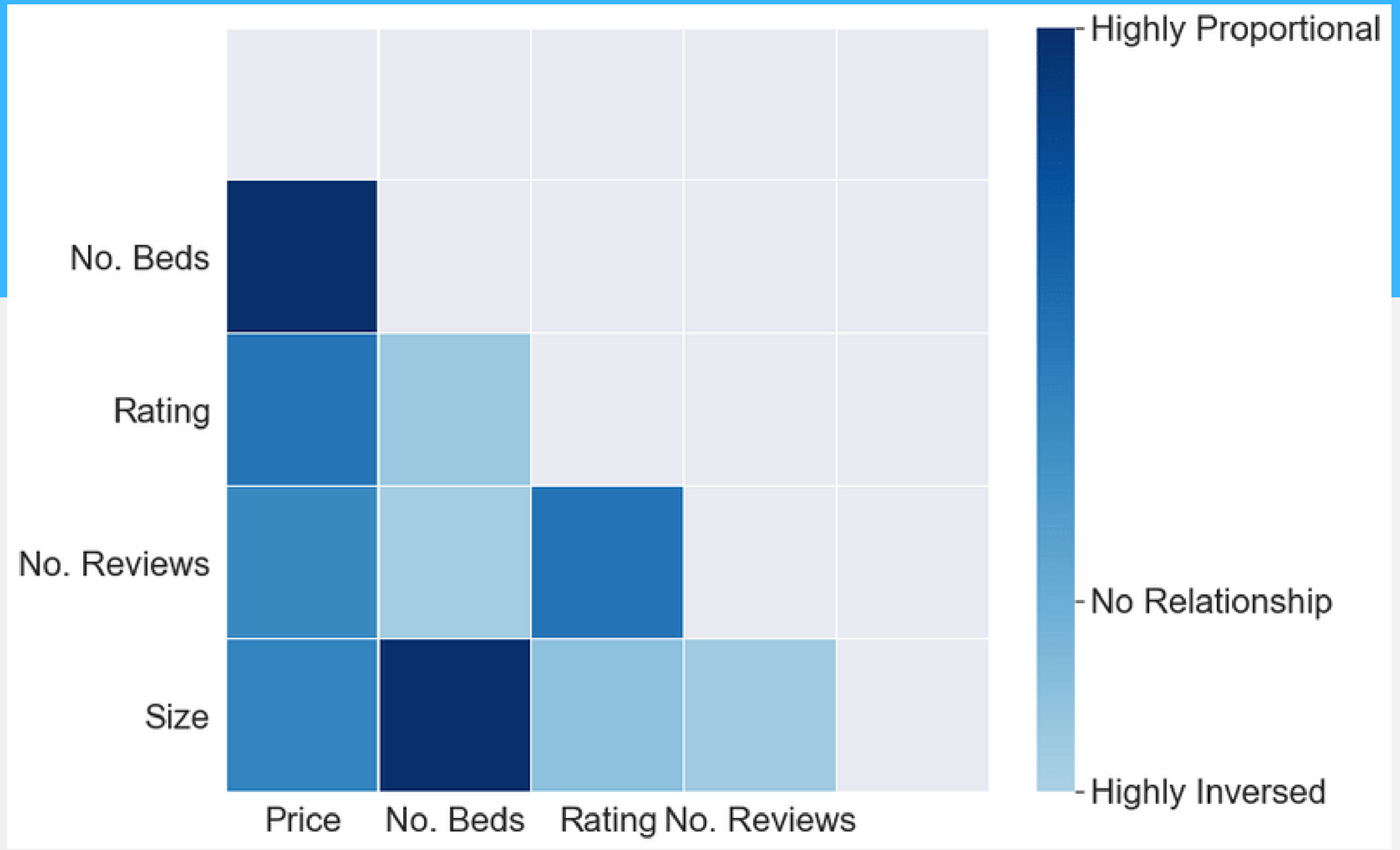
	hotel_name	location	price	room_type	beds	rating	rating_title	number_of_ratings	per_night
756	Mnazel Jawraq	Riyadh	1665	Villa with Private Pool	4	8.8	Fabulous	58	1
757	Alreef Diamond Villas	Taif	1500	Standard Villa	13	9.0	Superb	101	1
795	Faifa Hotel	Fayfa	1210	Superior Suite	1	7.0	Good	302	1
867	Bay La Sun Hotel and Marina - KAEC	King Abdullah Economic City	1074	Deluxe Double Room with City View	1	8.2	Very good	9054	1
88	مزرعة بإطلالة على الجبل	Madain	1603	Family Room	8	8.5	Very good	15	1

Data Exploration

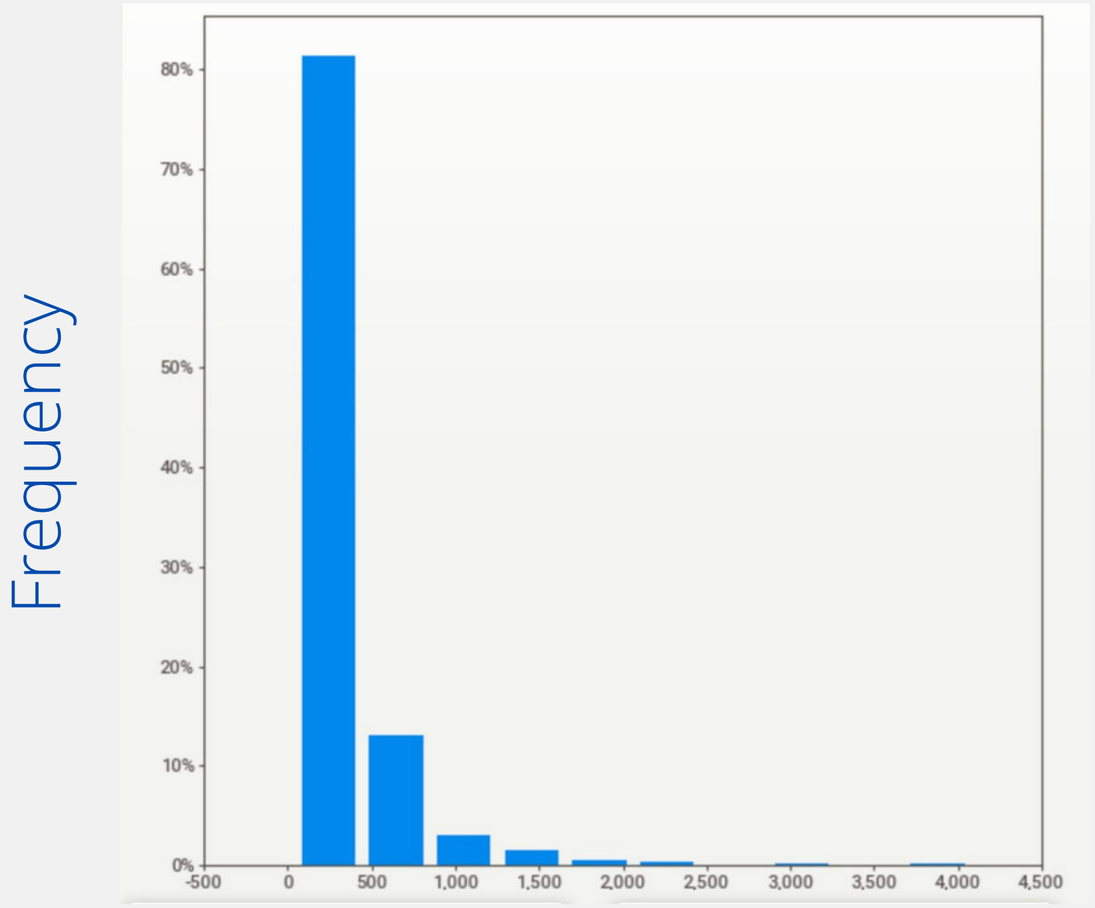
Distribution of all of the features.

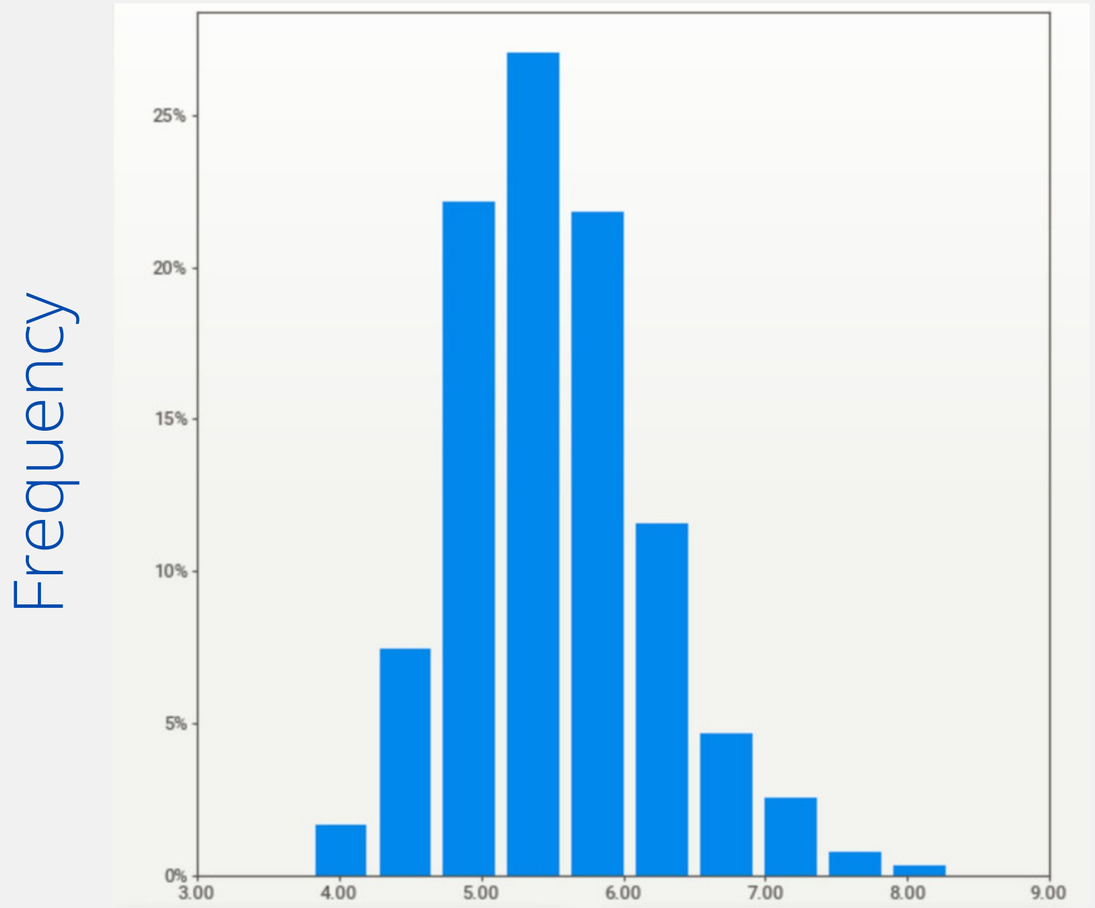


Relationships between features



Normalizing the Price Feature





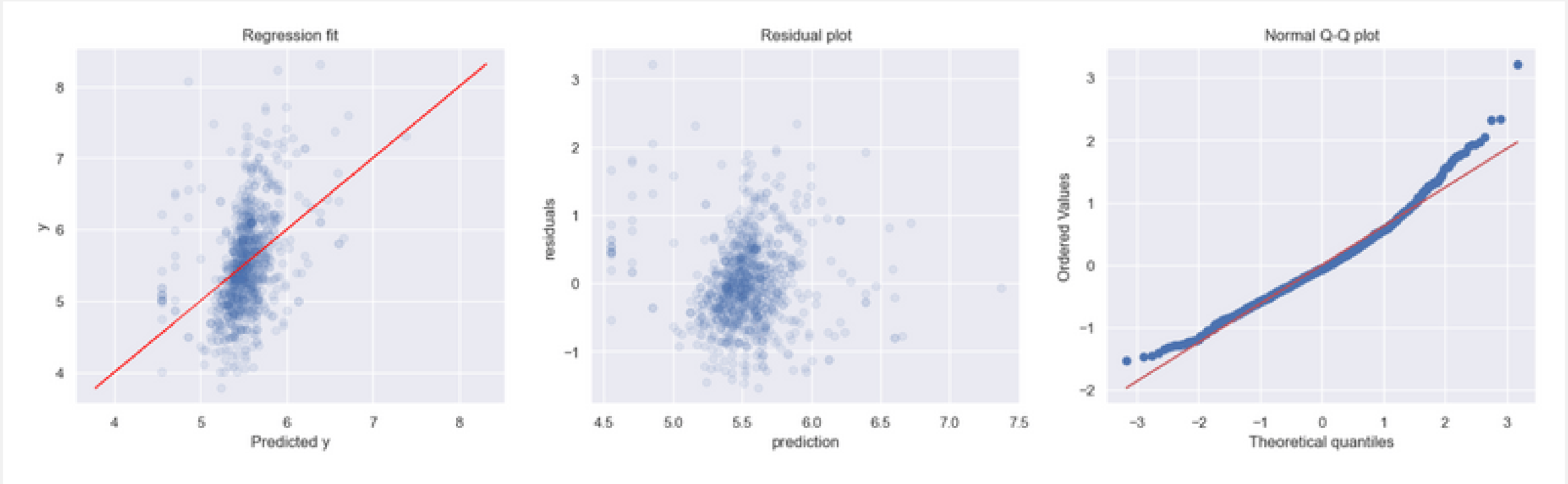
Linear Regression Analysis

What is linear regression?

- Type of predictive analysis.
- Building the regression model required understanding of how the dependent variable (Room Price) and independent variables are correlated with each other and how they are distributed.

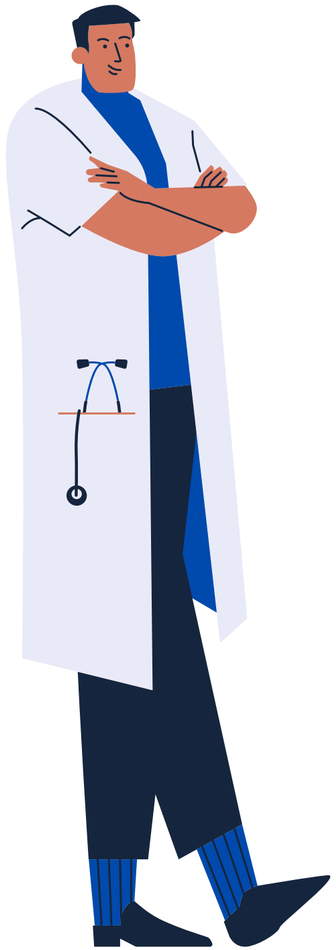
Model issue:

First Model Testing (R2= 0.2014)



CERTIFICATE OF

COMPLETION?



Results of Models

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lasso	Lasso Regression	147.1776	101164.6683	273.2976	0.2818	0.5531	0.4925	0.0220
omp	Orthogonal Matching Pursuit	142.3016	111329.8838	285.4695	0.2046	0.5493	0.4084	0.1390
catboost	CatBoost Regressor	138.6498	105141.9414	277.6973	0.1849	0.5079	0.4367	0.3000
huber	Huber Regressor	147.5762	117600.1012	299.7046	0.1589	0.5744	0.4306	0.1820
rf	Random Forest Regressor	134.5654	110373.6016	287.5146	0.1349	0.5049	0.3661	0.1520
llar	Lasso Least Angle Regression	166.2657	115070.1757	297.6114	0.0515	0.6174	0.6148	0.2470
en	Elastic Net	180.7168	125455.8107	313.8303	0.0457	0.6717	0.7235	0.0260
xgboost	Extreme Gradient Boosting	140.8743	130497.9691	307.6976	0.0312	0.5072	0.4244	0.4260
lightgbm	Light Gradient Boosting Machine	173.2290	112082.6970	303.6613	-0.0004	0.6397	0.6377	0.3180
lr	Linear Regression	160.5944	102867.9918	288.5394	-0.0077	0.7189	0.5864	0.0310
et	Extra Trees Regressor	138.2051	114154.2029	295.9174	-0.0428	0.5278	0.3743	0.2780
br	Bayesian Ridge	194.3929	132685.1899	325.6796	-0.0455	0.7268	0.8229	0.2000
ridge	Ridge Regression	191.3623	133600.8346	325.6445	-0.0512	0.7272	0.7706	0.0230
gbr	Gradient Boosting Regressor	152.7281	114364.2158	296.9347	-0.0523	0.5685	0.5240	0.0680
knn	K Neighbors Regressor	203.8325	147923.6585	352.2775	-0.3372	0.7390	0.7959	0.1540
dt	Decision Tree Regressor	146.5721	141271.4982	332.7113	-0.9782	0.6032	0.4166	0.0220
ada	AdaBoost Regressor	351.1982	216653.2283	445.9806	-1.5849	1.0342	1.8010	0.0550
par	Passive Aggressive Regressor	1348.7571	15399700.9417	2532.0828	-339.5597	1.4569	5.2795	0.0250

Model Testing

Regularization:
- Ridge Regularization
- Lasso Regularization
- ElasticNet Regularization

Parameter Tuning:
- Train-Validate-Test (split)
- Cross-validation
- K-fold
- Grid Search for Hyper-Parameters
- Randomized Search for Hyper-Parameters

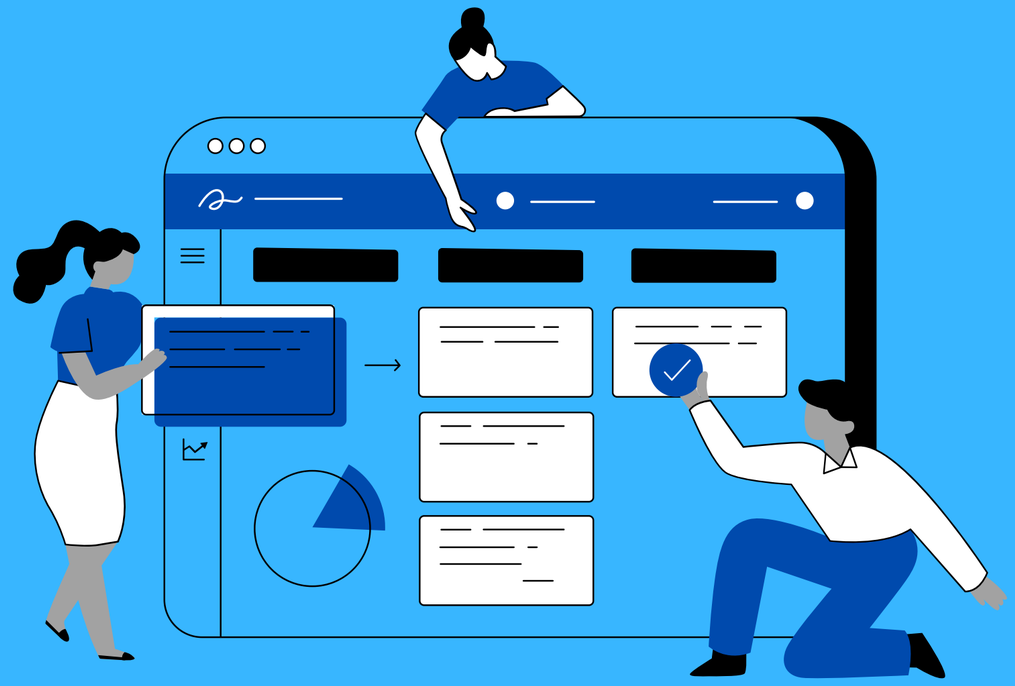
Scaling:
- RobustScaler
- QuantileTransformer
- PowerTransformer
- Normalizer
- MinMaxScaler

Guassian Transformations:
- Log
- Box-Cox
- Polynomial

***Then checked the outlier influence and multicollinearity using Variance Inflation Factor (VIF).**

	feature	VIF
0	Log_price	24.098882
1	Reviews	1.311794
2	rating	19.750120
3	beds	3.384291

Results of Best Models



Linear

91.61% Accurate

Polynomial

93.40% Accurate

Gradient Booster

92.75% Accurate

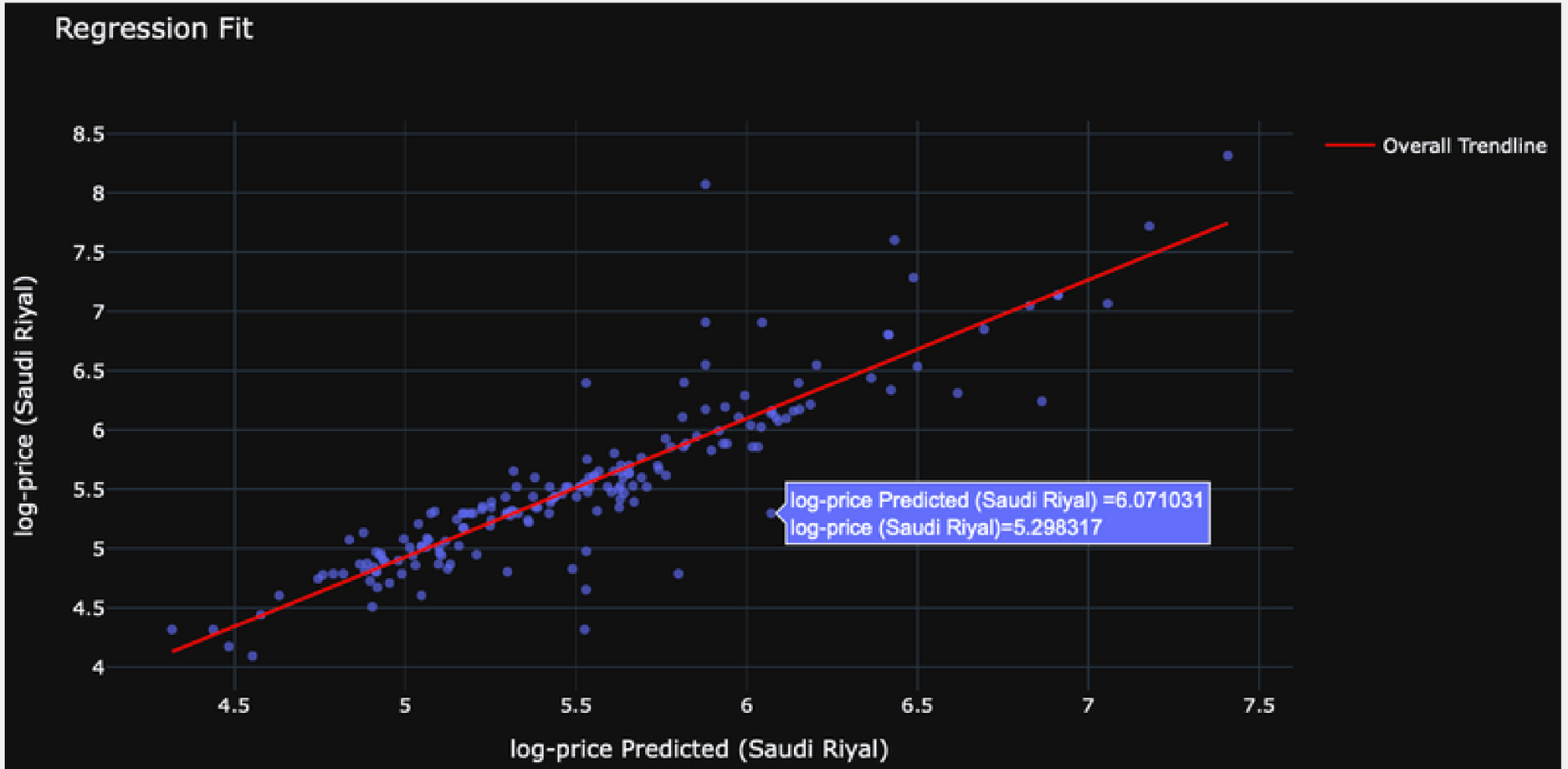
Random Forest

96.46% Accurate

Best Model

Random Forest Regressor:

- Evaluate model performance.
- Achieving 96.46% accuracy.
- The mean absolute error (MAE) is 0.1974 degrees.



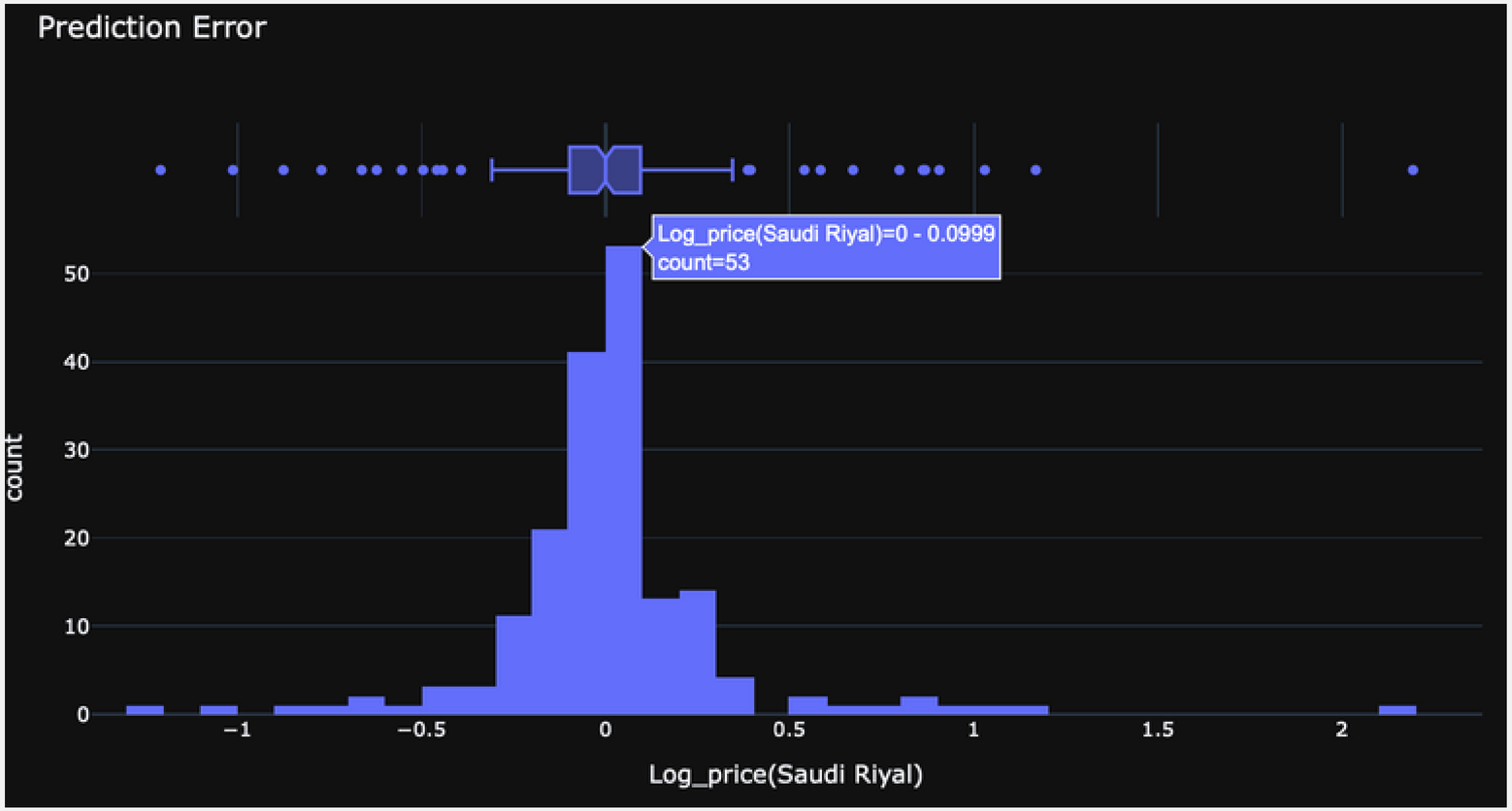
Test set

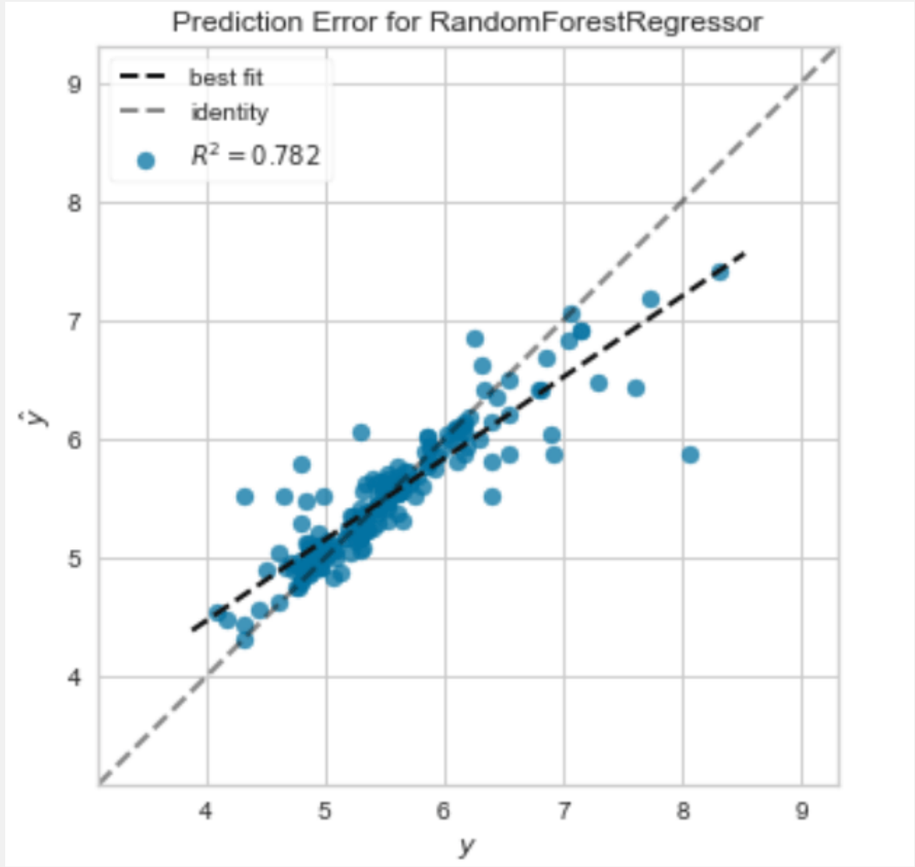
	Real Values	Predicted Values
536	420.0	368.712835
538	150.0	155.856314
852	162.0	153.590575
741	299.0	325.383444
537	360.0	380.760178
...	...	...
358	284.0	278.906786
371	275.0	318.566326
318	225.0	228.011934
283	565.0	591.331989
805	315.0	254.209970

Best Model

Random Forest Regressor Residuals:

- Prediction of errors.
- Explore the residuals to make sure everything was fine with the data.
- RF should have gotten a good fit.

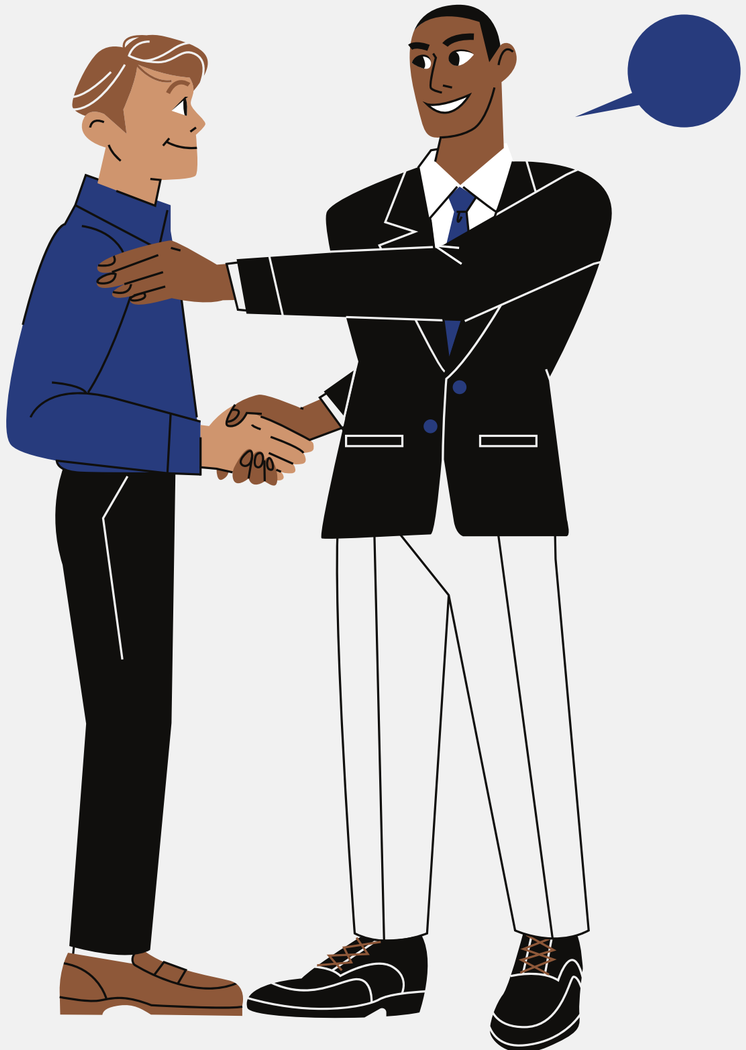




The Demo

SCAN ME

Conclusion & Recommmendation



- Model building is iterative over time, to increase the accuracy it's a must.
- Booking.com should handle false information and data entry about hotels in KSA.
- Ministry of Tourism should monitor and control hotels prices and increase tourists' trust in online booking platforms.
- Ensure data quality.

Future Work

- Increase the accuracy of the model by adding more relevant variables.
- Develop a better user experience.
- Seasonality and inflation factor.
- Adding more data from other online travel agencies platforms.

Thank You For Listening.

Do you have any questions?