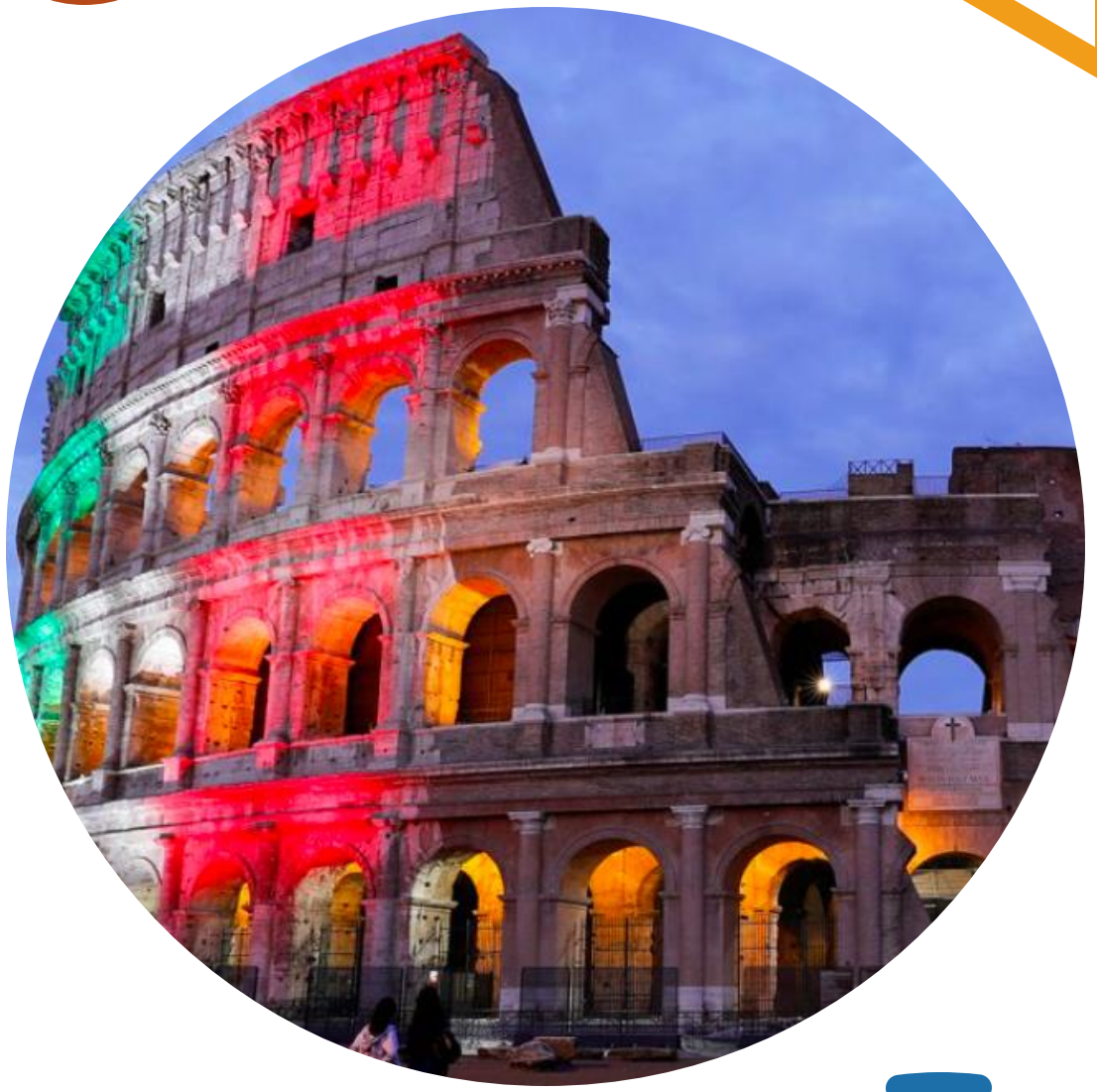


Ratings for Air BnB Stays in Italy

Elisabeth Johnson

Regression Module

Final Submission

Why Air BnB?

Started with the intention of helping those who do not have the means to book a hotel.

Made $4.8 Billion dollars in 2021 with only just over 6K employees

In December 10 2020, Air BnB became a public company raising $3.5 Billion in funds.

Cities Considered

- Venice
- Milan
- Florence
- Rome
- Turin

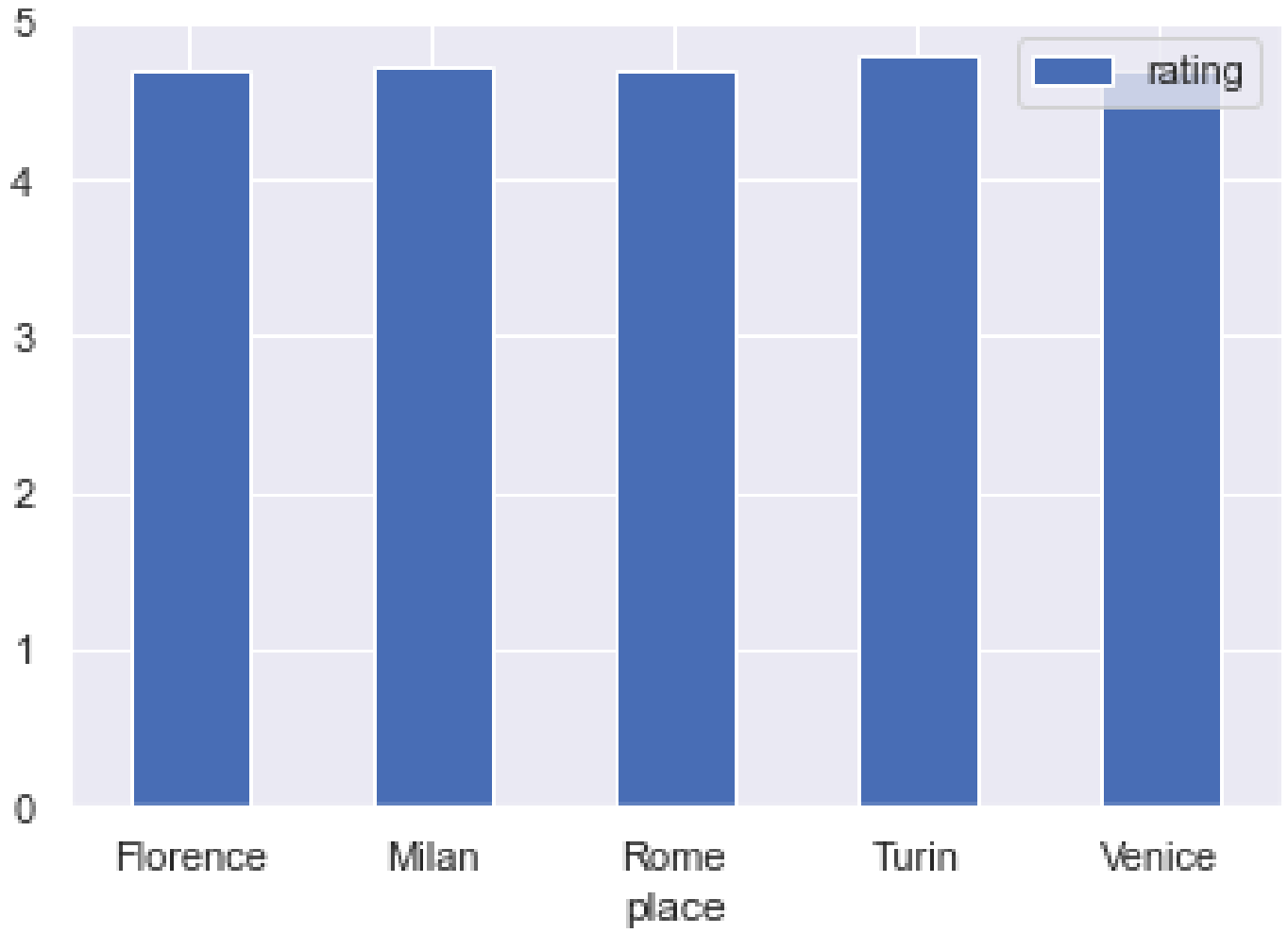
Features Considered

Is/is not a rare find or SuperHost	Is/is not a condo	Is/is not a private residence	Is/is not a shared residence
Is/is not a vacation residence	Is/is not a home	Is/is not an apartment	Is/is not Rome
Is/is not Venice	Is/is not Milan	Is/is not Turin	Is/is not Florence
Is/is not a hotel	Is/is not a villa	Is/is not a loft	Is/is not a guest residence

Target considered

Air BnB Stay Rating!

All cities have similar average ratings.



P- Values

The cities have little to no correlation with the rating predictor so we remove these values.

```
In [8]: import numpy as np
        import statsmodels.api as sm
        model = sm.OLS(y,X)


        # Here the Cities are causing noise in our data. we can drop them
        results = model.fit()

        results.pvalues
```

```
Out[8]: rare_or_superhost    3.269799e-20
        condo                1.519122e-04
        private              7.313963e-04
        shared               1.783857e-01
        vacation             7.530318e-04
        home                 4.989348e-02
        apartment            4.713558e-03
        Rome                 0.000000e+00
        Venice               0.000000e+00
        Milan                0.000000e+00
        Turin                0.000000e+00
        Florence             0.000000e+00
        hotel                6.828621e-01
        villa                3.792998e-02
        loft                 3.218547e-03
        guest                7.821115e-03
        dtype: float64
```

R^2 Before and After City Feature Removal

BEFORE
.13

AFTER
.98

Mean Absolute Error

	Ridge	Linear	Lasso
Repeat K-Fold (10 split, 3 repeat)	.19000	.19000	.20800
K-Fold (10 split, 0 repeat)	.18670	.18672	.20544

Mean Absolute Error

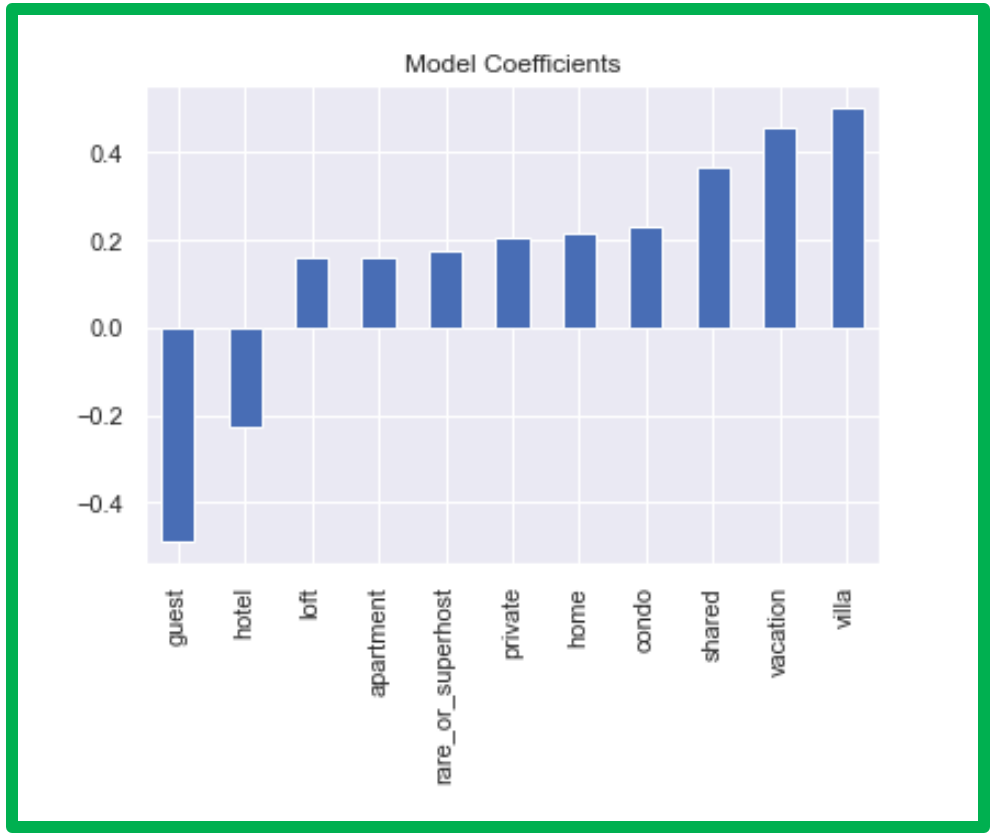
	Ridge	Linear	Lasso
Repeat K-Fold (10 split, 3 repeat)	.19000	.19000	.20800
K-Fold (10 split, 0 repeat)	.18670	.18672	.20544

Actual	Predicted
4.88	4.816182
4.67	4.644558
5.00	4.687606
4.48	4.644558
4.60	4.818912
4.35	4.687606
5.00	4.716090
5.00	4.890444
4.77	4.687606
4.47	4.644558
5.00	4.818912
4.80	4.890444
4.81	4.861960
5.00	4.716090
4.81	4.687606
4.50	4.644558
4.82	4.644558
4.48	4.644558
4.92	4.890444
4.61	4.700608
4.69	4.644558
4.80	4.716090
4.67	4.687606
5.00	4.687606
4.18	4.644558

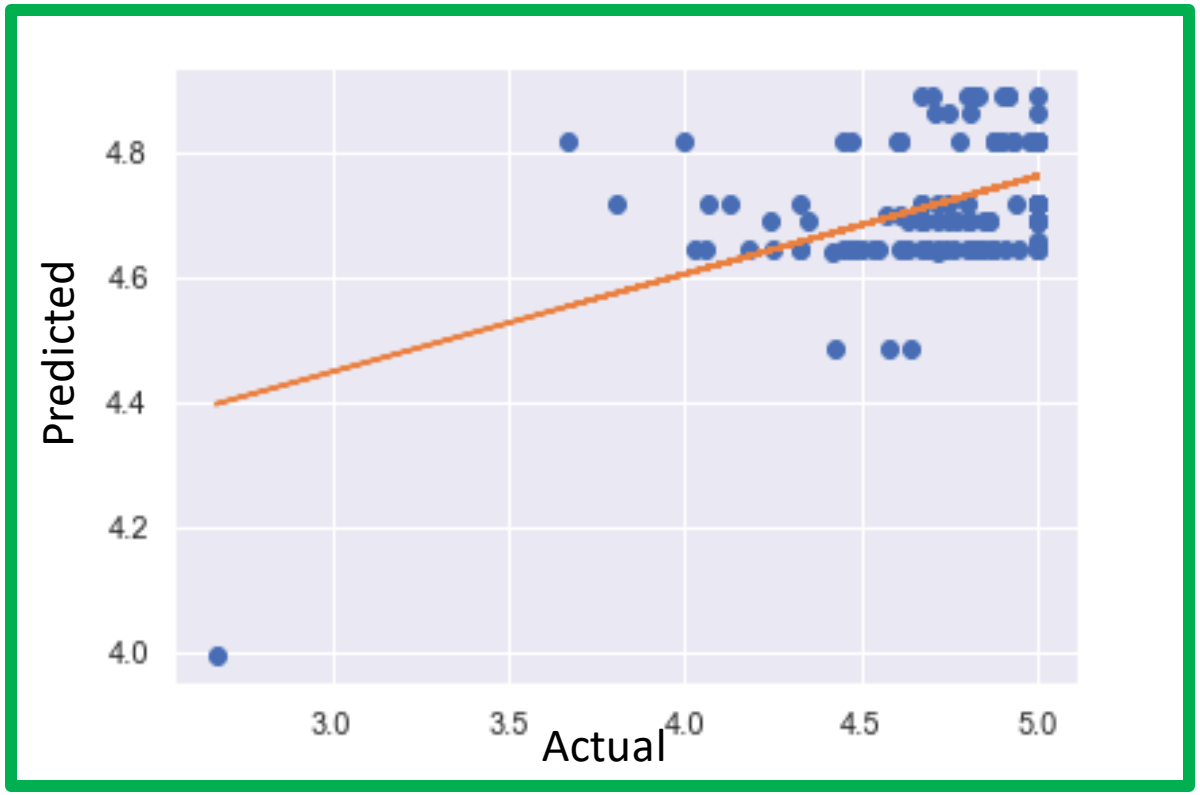
Sample Performance

Ridge Regression Fit with 10 Split K-Fold

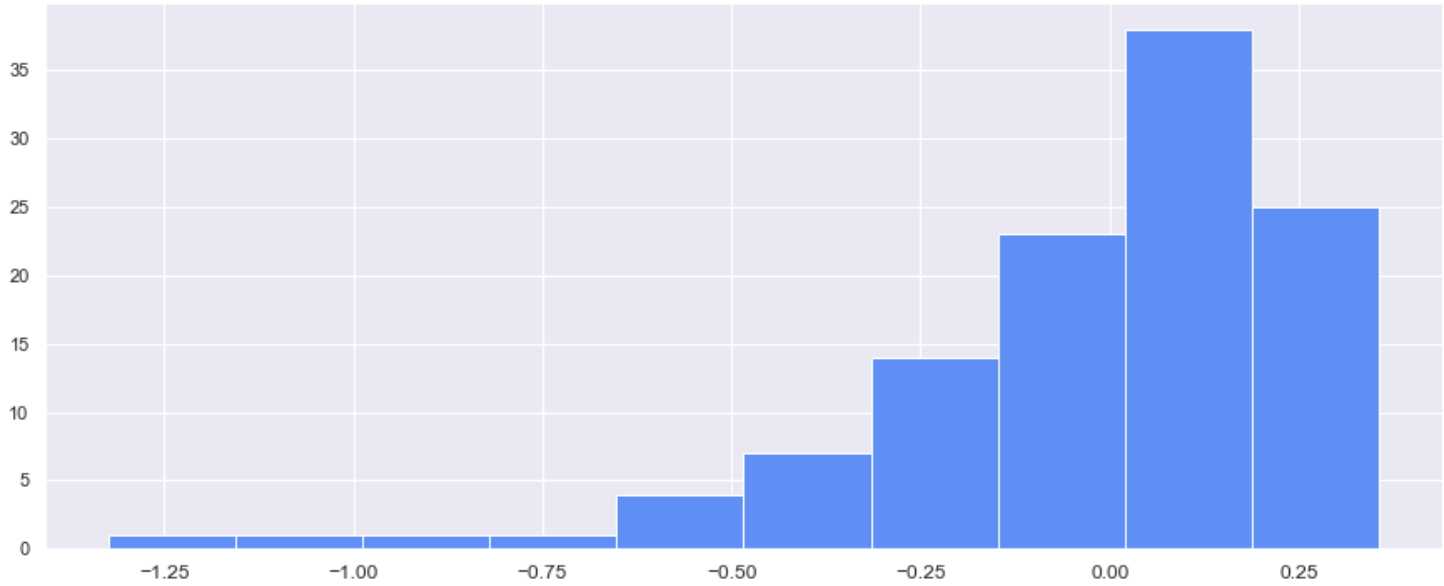
Coefficients



Regression Line



Residual Histogram



References

Thank You!