Airbnb Price Predictive model
CIS 512
Prof. Sumanlata Ghosh

Presented by: **Anand Vekariya**

Data Cleaning

Drop column

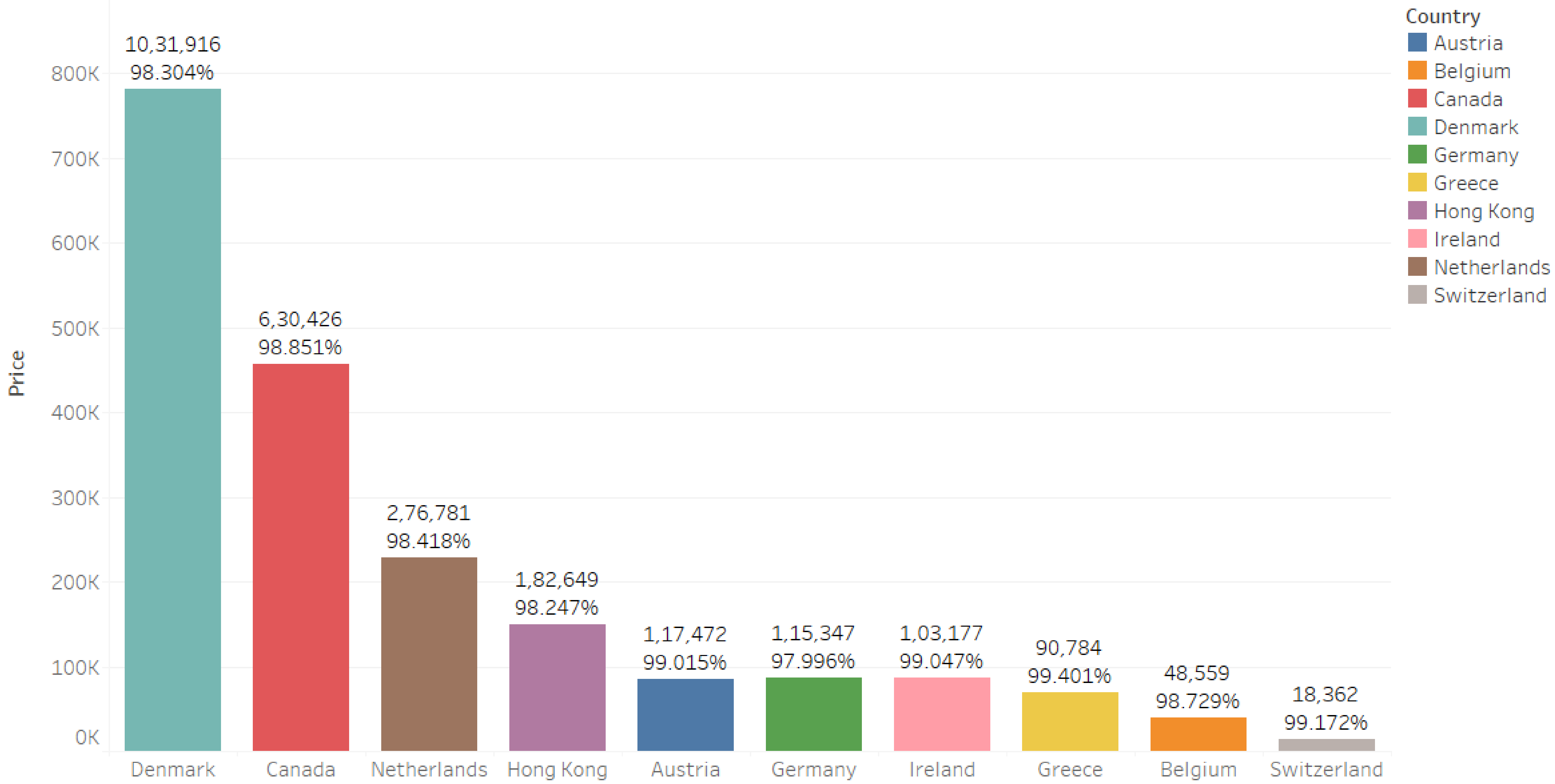
Listing Url, Scrape ID, Name, Summary, Space, Description, Experiences Offered, Neighborhood Overview, Notes, Access, Interaction, House Rules, Thumbnail Url, Medium Url, Picture Url, XL Picture Url, Host URL,	Host Name, Host About, Host Acceptance Rate, Host Thumbnail Url, Host Picture Url, Host Neighbourhood, Host Listings Count, Host Total Listings Count, Host Verifications, Neighbourhood, Neighbourhood Cleansed, Neighbourhood Group Cleansed, State, Zipcode, Smart Location, Country Code, Geolocation, Cancellation Policy	Latitude, Longitude, Square Feet, Has Availability, Availability 30, Availability 60, Availability 90, Availability 365, Calendar last Scraped, First Review, Review Scores Accuracy, Review Scores Cleanliness, Review Scores Checkin, Review Scores Communication, Review Scores Location, Review Scores Value, Jurisdiction Names

Columns	Data Description	Action Taken
Amenities	Multiple facilities and requirements	Split into categories; new column = Yes/No.
Price	Contains NA values.	Check weekly/monthly values; calculate mean price based on country.
Bedrooms, Beds, Bathrooms	Contains NA values	Use 'Accommodates' column to find actual numbers; otherwise, set to 0.
Host Location Match	Host location matches property location	Host Location (split) = "Street" or "City."
Last Rented (in months)	Findout Last time rented.	'Last Rented (in months)' = 'Last Scraped' - 'Last Review.'
Host Age	Determine host age.	'Host Age' = 'Last Scraped' - 'Host Since.'

Columns	Data Description	Action Taken
Property Type	Unorganized property types.	Categorize based on house types.
Features	Long strings with essential details.	Split features; create new column with Yes/No.
Transit, Market, Weekly Price, Monthly Price, License	Long strings with mostly null values.	If empty, set to 'No'; else, set to 'Yes.'
Host Response Time, Beds, Security Deposit, Cleaning Fee, Review Scores Rating, Extra People, Reviews per Month	Contains NA values.	Fill with 0 for int/float; else, set to "Not Mention."

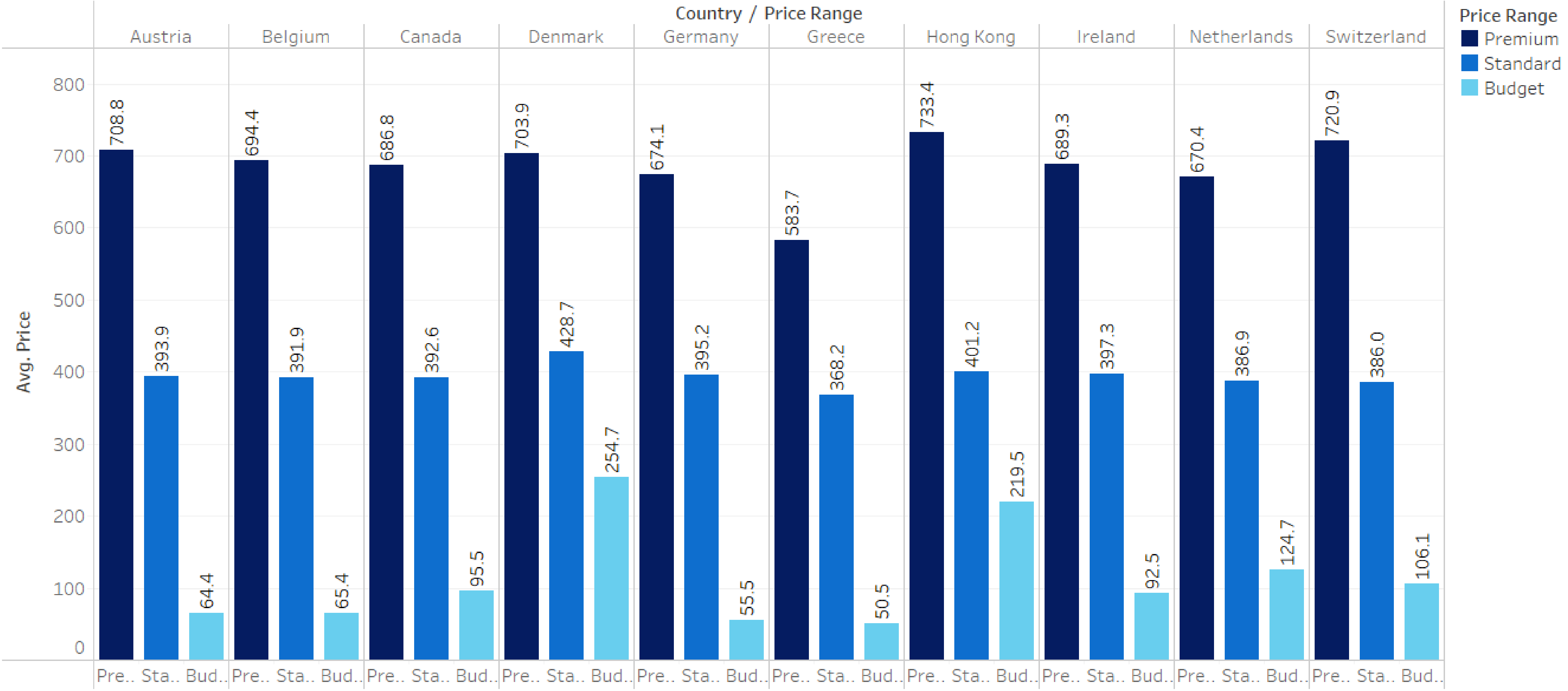
Data Analysis

Total Revenue by Country



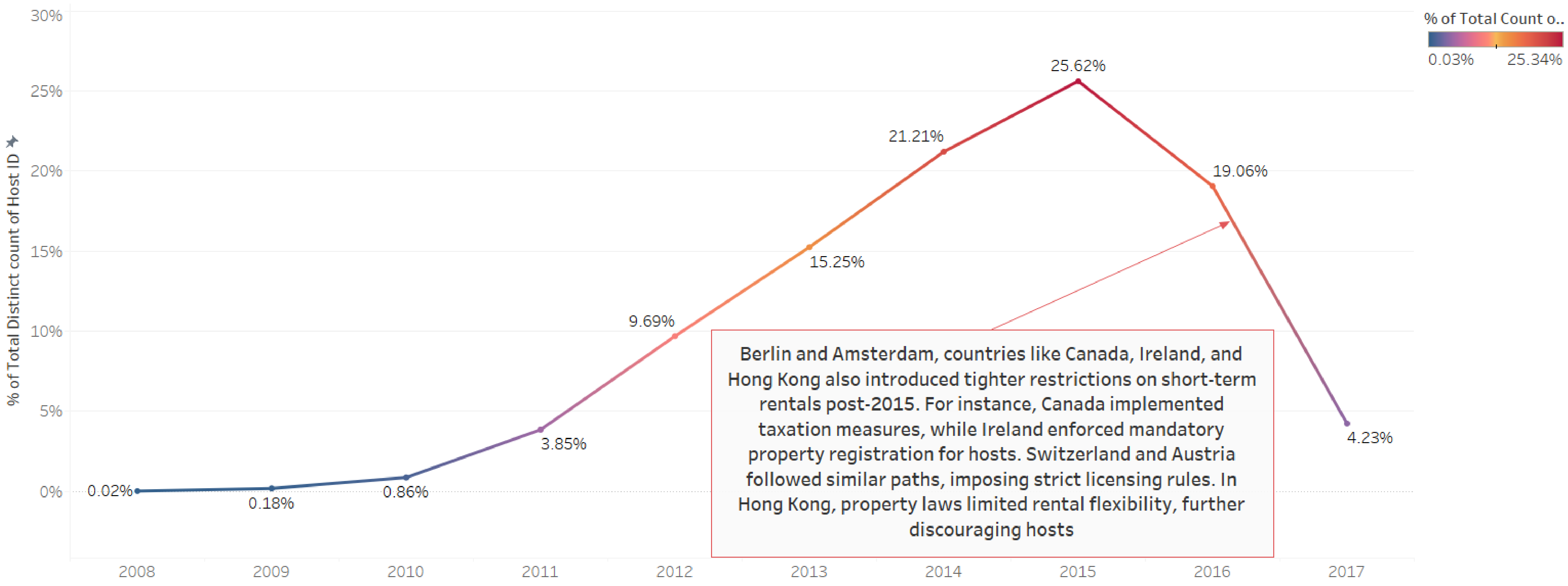
Shows which countries generate the highest revenue from Airbnb rentals, emphasizing quality hosts (Superhosts).

Average Airbnb Prices Across Countries



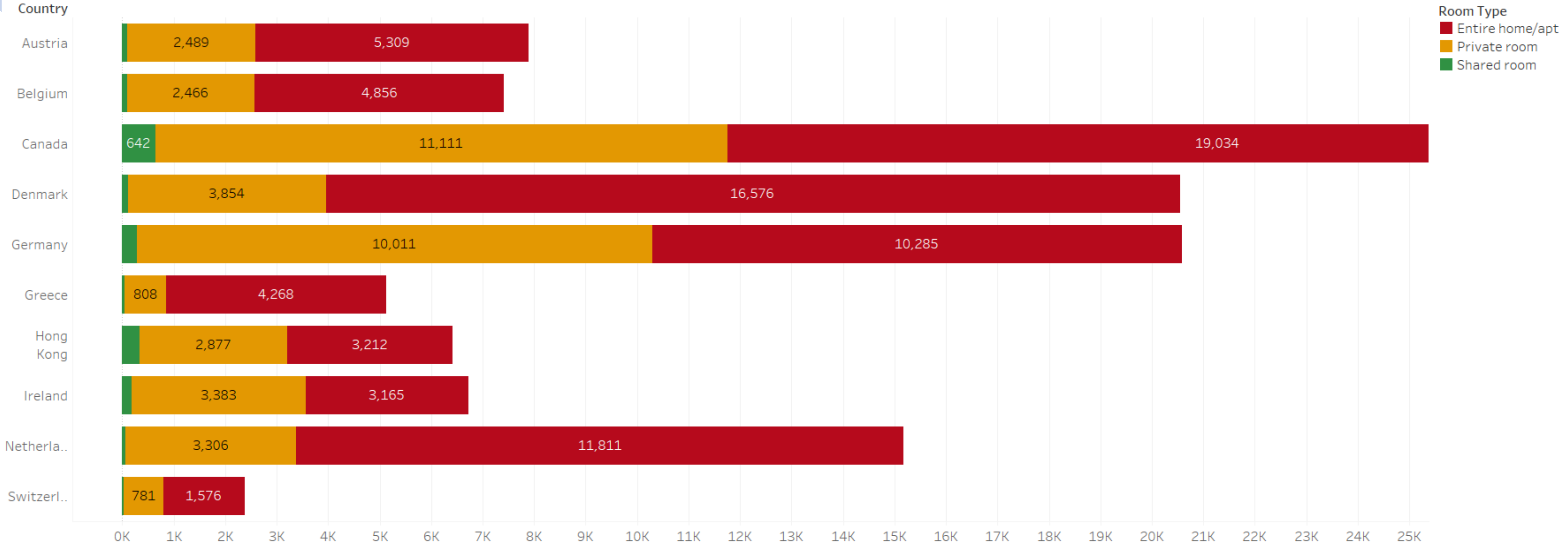
It compares average Airbnb prices across countries by Budget, Standard, and Premium categories, highlighting Denmark and Hong Kong as the most expensive, while Greece and Germany offer more affordable options.

Yearly Host Registration Rates on Airbnb



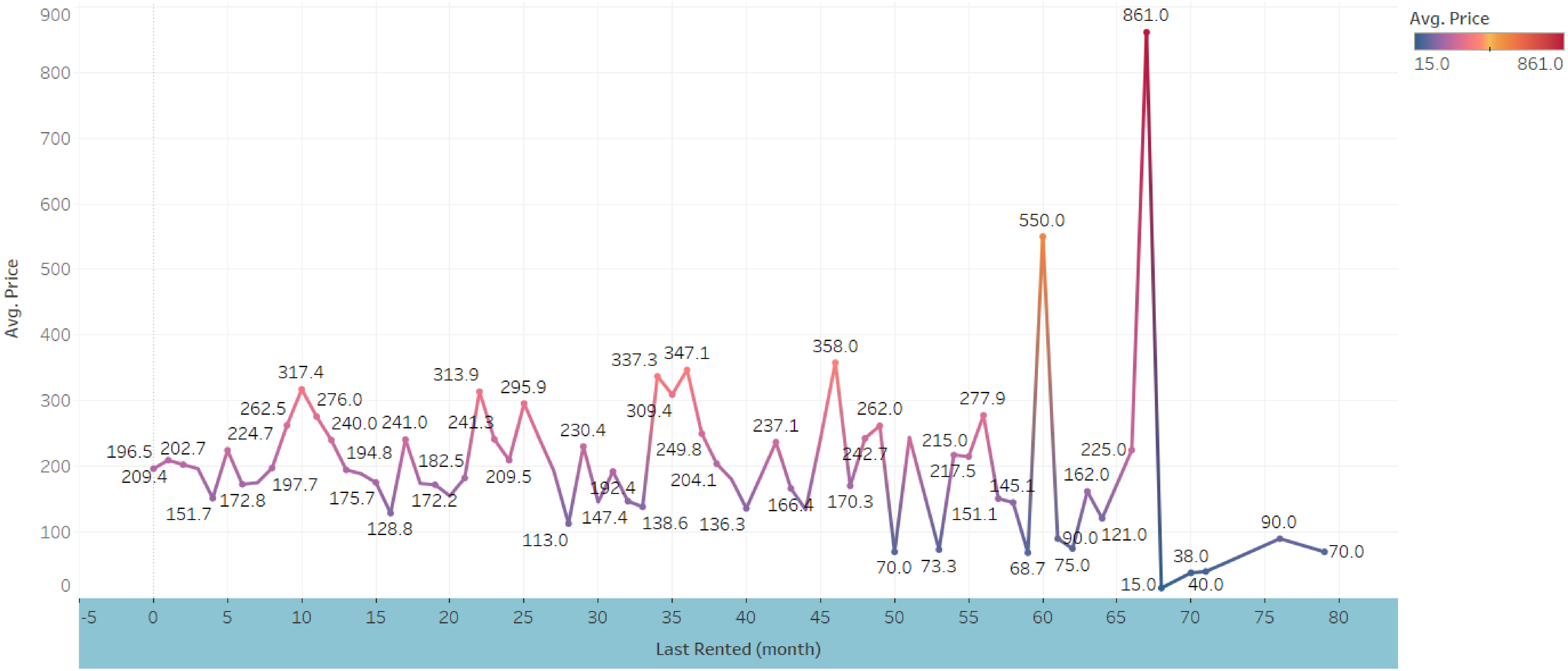
Host registrations began at 0.02% in 2008, gradually increasing to 25.63% by 2015. Following this peak, a significant drop occurred, with registrations falling to 4.23% by 2017. The rapid growth between 2011 and 2015 contrasts sharply with the steep decline in the subsequent years.

Room Type Distribution by Country



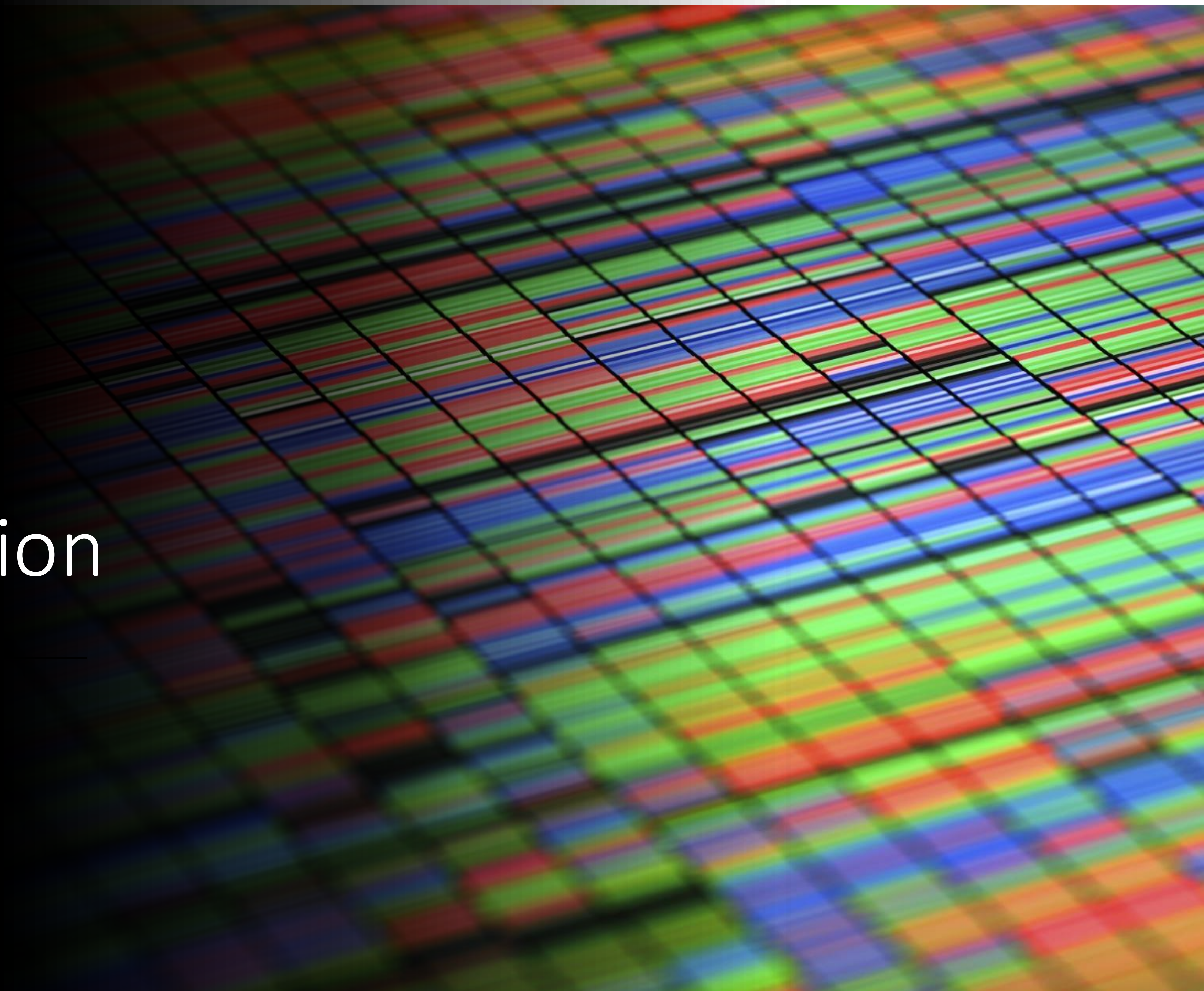
Shows the popularity of different room types in each country, helping investors understand which types of properties are more in demand.

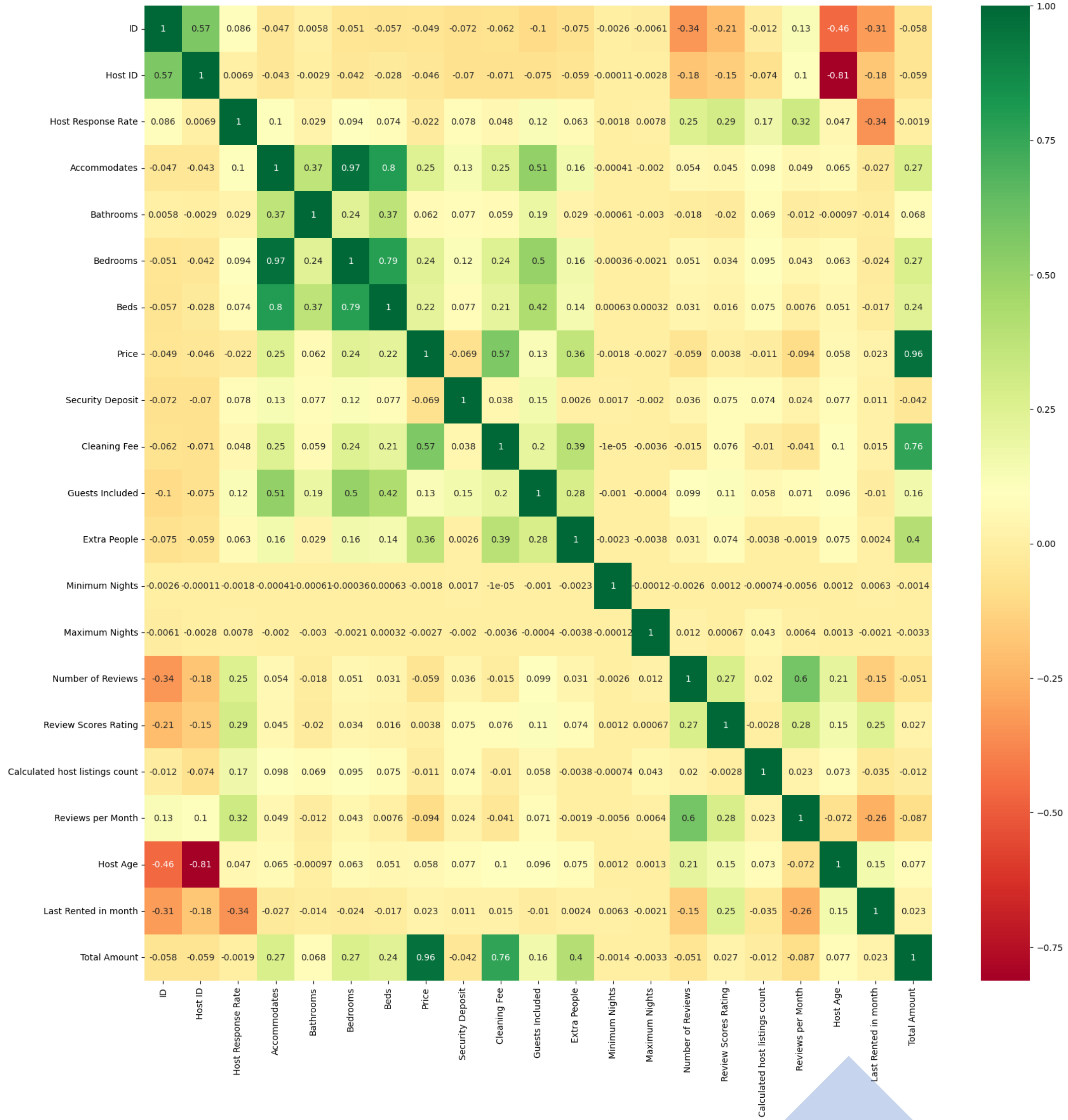
Current Average Price X Number of Months Since Last Rental

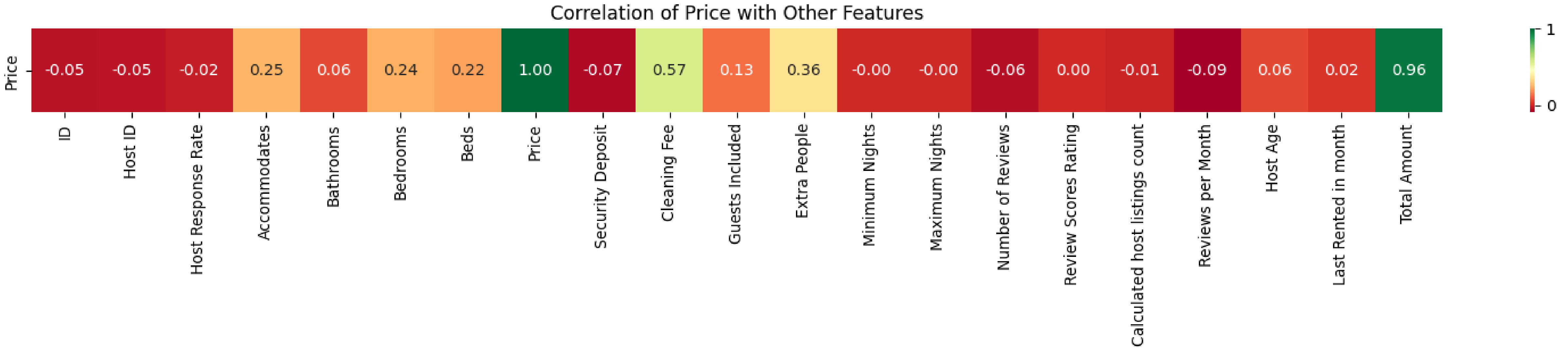


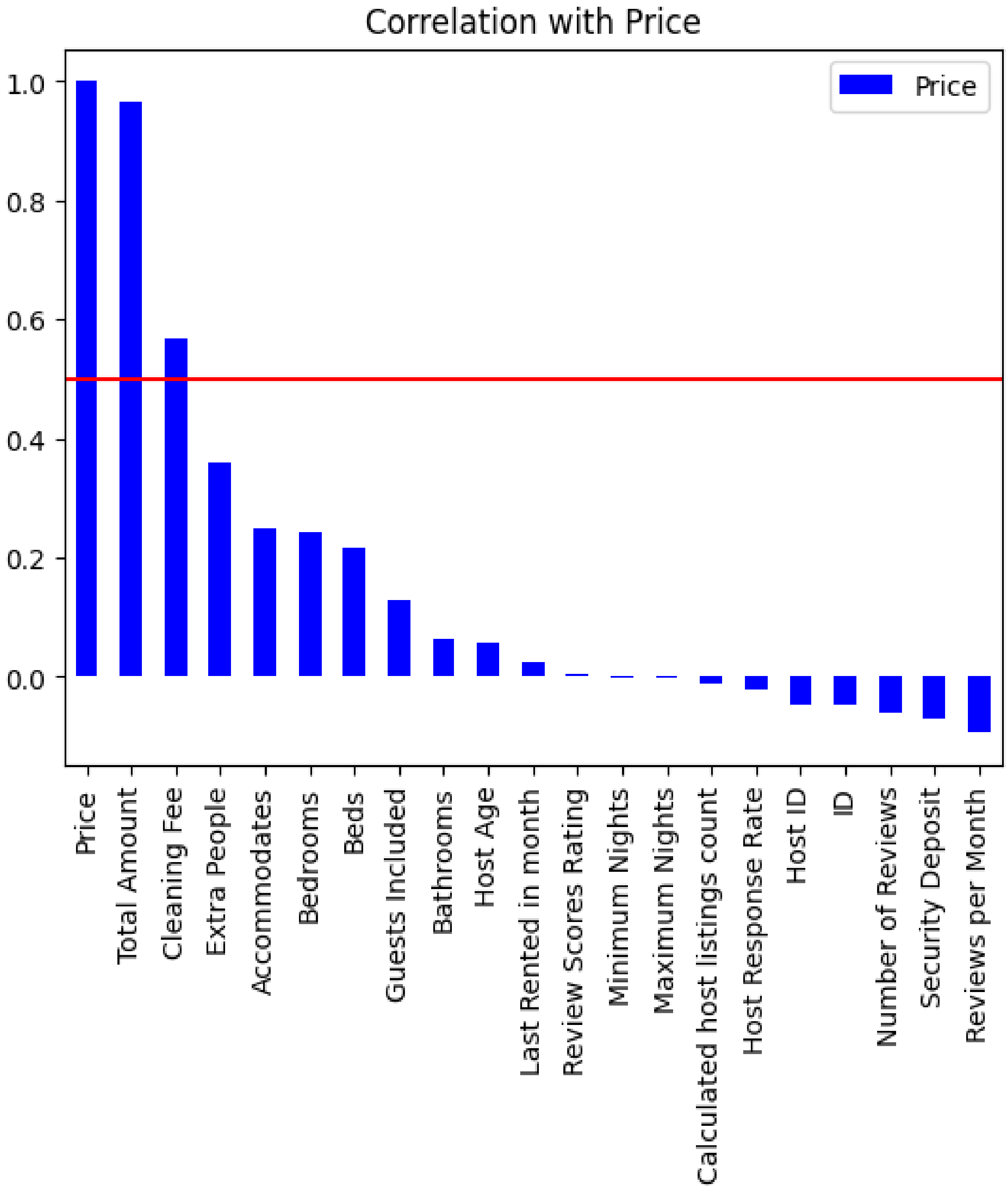
This chart displays the current average price of Airbnb Property X, indicating the number of months since its last rental. It provides insights into how rental availability may influence pricing.

Data Correlation




ID 1 0.57 0.086 -0.047 0.0058 -0.051 -0.057 -0.049 -0.072 -0.062 -0.1 -0.075 -0.0026 -0.0061 -0.34 -0.21 -0.012 0.13 -0.46 -0.31 -0.058
Host ID 0.57 1 0.0069 -0.043 -0.0029 -0.042 -0.028 -0.046 -0.07 -0.071 -0.075 -0.059 -0.00011 -0.0028 -0.18 -0.15 -0.074 0.1 -0.81 -0.18 -0.059
Host Response Rate 0.086 0.0069 1 0.1 0.029 0.094 0.074 -0.022 0.078 0.048 0.12 0.063 -0.0018 0.0078 0.25 0.29 0.17 0.32 0.047 -0.34 -0.0019
Accommodates -0.047 -0.043 0.1 1 0.37 0.97 0.8 0.25 0.13 0.25 0.51 0.16 -0.00041 -0.002 0.054 0.045 0.098 0.049 0.065 -0.027 0.27
Bathrooms 0.0058 -0.0029 0.029 0.37 1 0.24 0.37 0.062 0.077 0.059 0.19 0.029 -0.00061 -0.003 -0.018 -0.02 0.069 -0.012 -0.00097 -0.014 0.068
Bedrooms -0.051 -0.042 0.094 0.97 0.24 1 0.79 0.24 0.12 0.24 0.5 0.16 -0.00036 -0.0021 0.051 0.034 0.095 0.043 0.063 -0.024 0.27
Beds -0.057 -0.028 0.074 0.8 0.37 0.79 1 0.22 0.077 0.21 0.42 0.14 0.00063 0.00032 0.031 0.016 0.075 0.0076 0.051 -0.017 0.24
Price -0.049 -0.046 -0.022 0.25 0.062 0.24 0.22 1 -0.069 0.57 0.13 0.36 -0.0018 -0.0027 -0.059 0.0038 -0.011 -0.094 0.058 0.023 0.96
Security Deposit -0.072 -0.07 0.078 0.13 0.077 0.12 0.077 -0.069 1 0.038 0.15 0.0026 0.0017 -0.002 0.036 0.075 0.074 0.024 0.077 0.011 -0.042
Cleaning Fee -0.062 -0.071 0.048 0.25 0.059 0.24 0.21 0.57 0.038 1 0.2 0.39 -1e-05 -0.0036 -0.015 0.076 -0.01 -0.041 0.1 0.015 0.76
Guests Included -0.1 -0.075 0.12 0.51 0.19 0.5 0.42 0.13 0.15 0.2 1 0.28 -0.001 -0.0004 0.099 0.11 0.058 0.071 0.096 -0.01 0.16
Extra People -0.075 -0.059 0.063 0.16 0.029 0.16 0.14 0.36 0.0026 0.39 0.28 1 -0.0023 -0.0038 0.031 0.074 -0.0038 -0.0019 0.075 0.0024 0.4
Minimum Nights -0.0026 -0.00011 -0.0018 -0.00041 -0.00061 -0.00036 0.00063 -0.0018 0.0017 -1e-05 -0.001 -0.0023 1 -0.00012 -0.0026 0.0012 -0.00074 -0.0056 0.0012 0.0063 -0.0014
Maximum Nights -0.0061 -0.0028 0.0078 -0.002 -0.003 -0.0021 0.00032 -0.0027 -0.002 -0.0036 -0.0004 -0.0038 -0.00012 1 0.012 0.00067 0.043 0.0064 0.0013 -0.0021 -0.0033
Number of Reviews -0.34 -0.18 0.25 0.054 -0.018 0.051 0.031 -0.059 0.036 -0.015 0.099 0.031 -0.0026 0.012 1 0.27 0.02 0.6 0.21 -0.15 -0.051
Review Scores Rating -0.21 -0.15 0.29 0.045 -0.02 0.034 0.016 0.0038 0.075 0.076 0.11 0.074 0.0012 0.00067 0.27 1 -0.0028 0.28 0.15 0.25 0.027
Calculated host listings count -0.012 -0.074 0.17 0.098 0.069 0.095 0.075 -0.011 0.074 -0.01 0.058 -0.0038 -0.00074 0.043 0.02 -0.0028 1 0.023 0.073 -0.035 -0.012
Reviews per Month 0.13 0.1 0.32 0.049 -0.012 0.043 0.0076 -0.094 0.024 -0.041 0.071 -0.0019 -0.0056 0.0064 0.6 0.28 0.023 1 -0.072 -0.26 -0.087
Host Age -0.46 -0.81 0.047 0.065 -0.00097 0.063 0.051 0.058 0.077 0.1 0.096 0.075 0.0012 0.0013 0.21 0.15 0.073 -0.072 1 0.15 0.077
Last Rented in month -0.31 -0.18 -0.34 -0.027 -0.014 -0.024 -0.017 0.023 0.011 0.015 -0.01 0.0024 0.0063 -0.0021 -0.15 0.25 -0.035 -0.26 0.15 1 0.023
Total Amount -0.058 -0.059 -0.0019 0.27 0.068 0.27 0.24 0.96 -0.042 0.76 0.16 0.4 -0.0014 -0.0033 -0.051 0.027 -0.012 -0.087 0.077 0.023 1
ID
Host ID
Host Response Rate
Accommodates
Bathrooms
Bedrooms
Beds
Price
Security Deposit
Cleaning Fee
Guests Included
Extra People
Minimum Nights
Maximum Nights
Number of Reviews
Review Scores Rating
Calculated host listings count
Reviews per Month
Host Age
Last Rented in month
Total Amount
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Correlation of Price with Other Features
Price
-0.05
-0.05
-0.02
0.25
0.06
0.24
0.22
1.00
-0.07
0.57
0.13
0.36
-0.00
-0.00
-0.06
0.00
-0.01
-0.09
0.06
0.02
0.96
ID
Host ID
Host Response Rate
Accommodates
Bathrooms
Bedrooms
Beds
Price
Security Deposit
Cleaning Fee
Guests Included
Extra People
Minimum Nights
Maximum Nights
Number of Reviews
Review Scores Rating
Calculated host listings count
Reviews per Month
Host Age
Last Rented in month
Total Amount
1
0


Correlation with Price
Price
1.0
0.8
0.6
0.4
0.2
0.0
Price
Total Amount
Cleaning Fee
Extra People
Accommodates
Bedrooms
Beds
Guests Included
Bathrooms
Host Age
Last Rented in month
Review Scores Rating
Minimum Nights
Maximum Nights
Calculated host listings count
Host Response Rate
Host ID
ID
Number of Reviews
Security Deposit
Reviews per Month

Feature Engineering

Feature Engineering Process:

Feature Engineering Process:

- Converted categorical variables to numerical using one-hot encoding
- Created new columns for each category:
- 1 indicates presence 0 indicates absence
- Original categorical columns dropped

Categorical Variables Encoded:

- **Host Response Time**
- **Country**
- **Property Type**
- **Room Type**
- **Bed Type**
- **Cancellation Policy**

Example:

- Before: Room Type = "Entire home/apt"
- After: Room Type_Entire home/apt = 1, others = 0

Resulting Dataframe Shape:

- Rows: 123,061
- Columns: 91

Preparing Data for Modeling

Correlation Filtering:

- Removed columns with correlation < 0.02 to price

Data Splitting:

- Training Data: 75% (~92,295 rows)
- Testing/Validation Data: 25% (~30,766 rows)

Data Standardization:

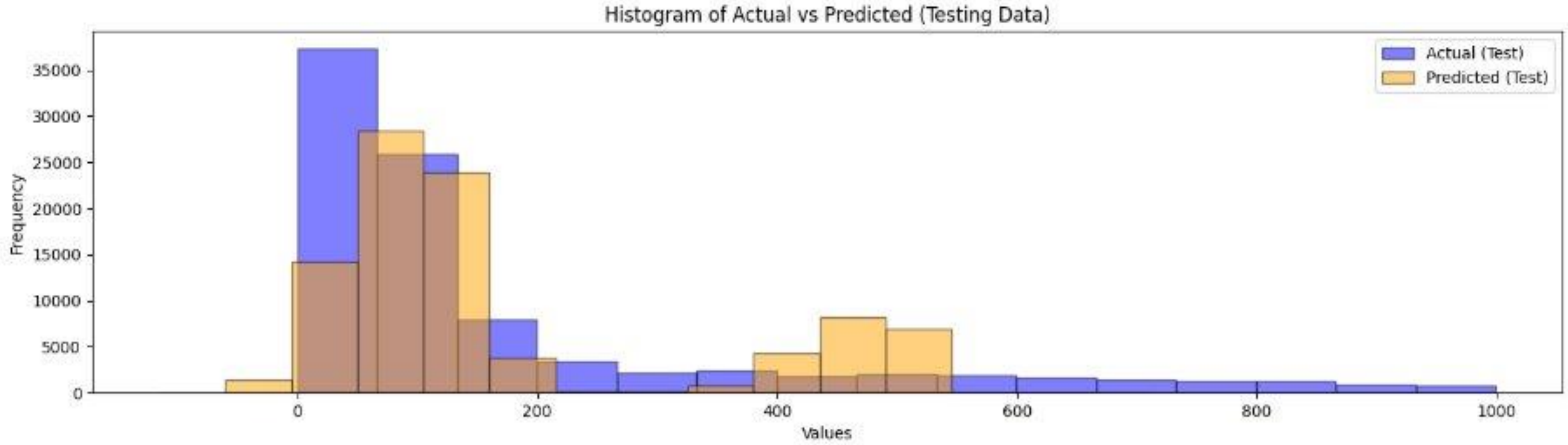
- Applied StandardScaler to features
- Standardizes features by removing the mean and scaling to unit variance

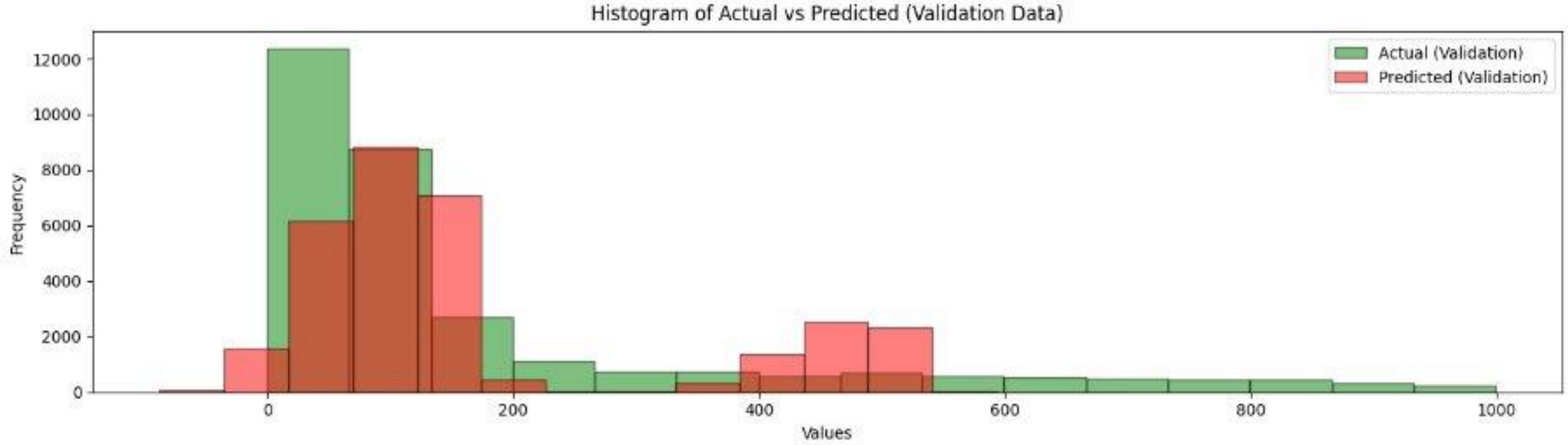
Dimensionality Reduction:

- Performed Principal Component Analysis (PCA)
- Reduced feature set to 57 columns for modeling

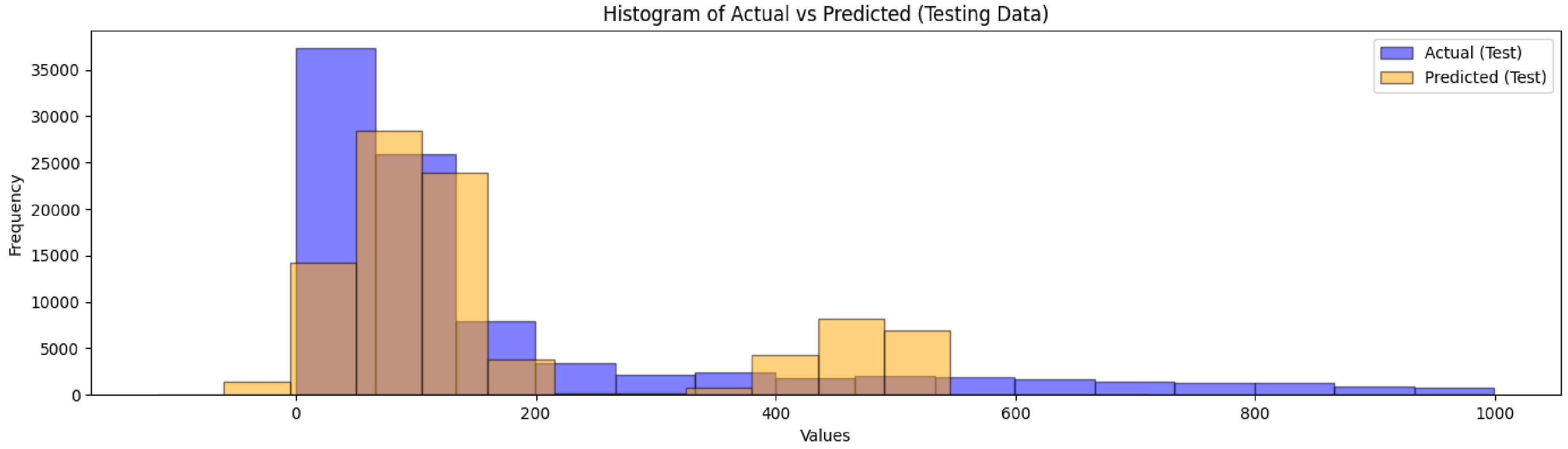
Predictive Modeling

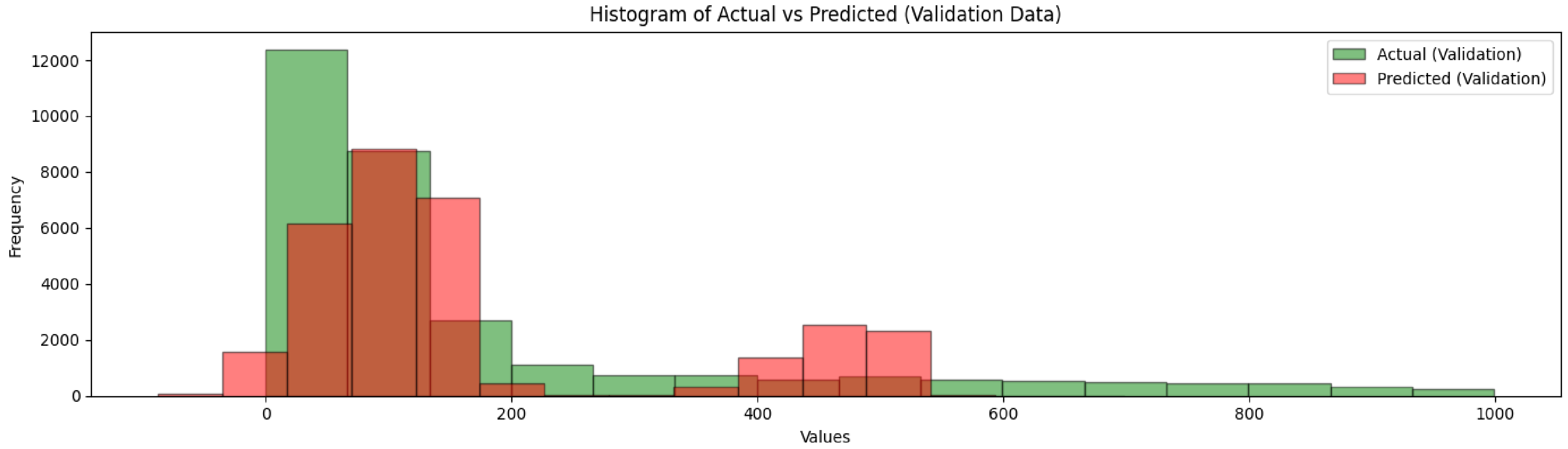
Linear Regression



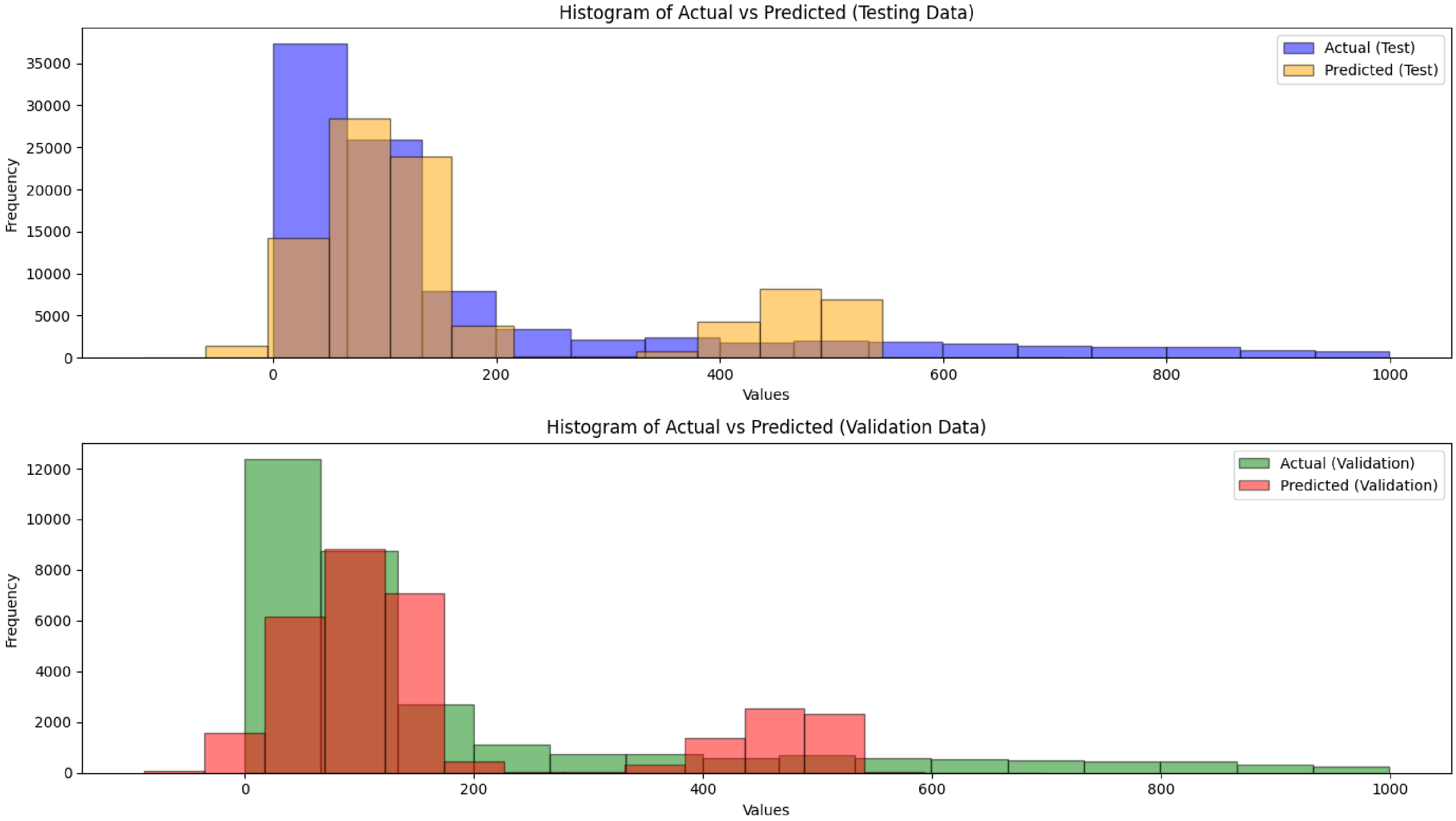


Random Forest Regression

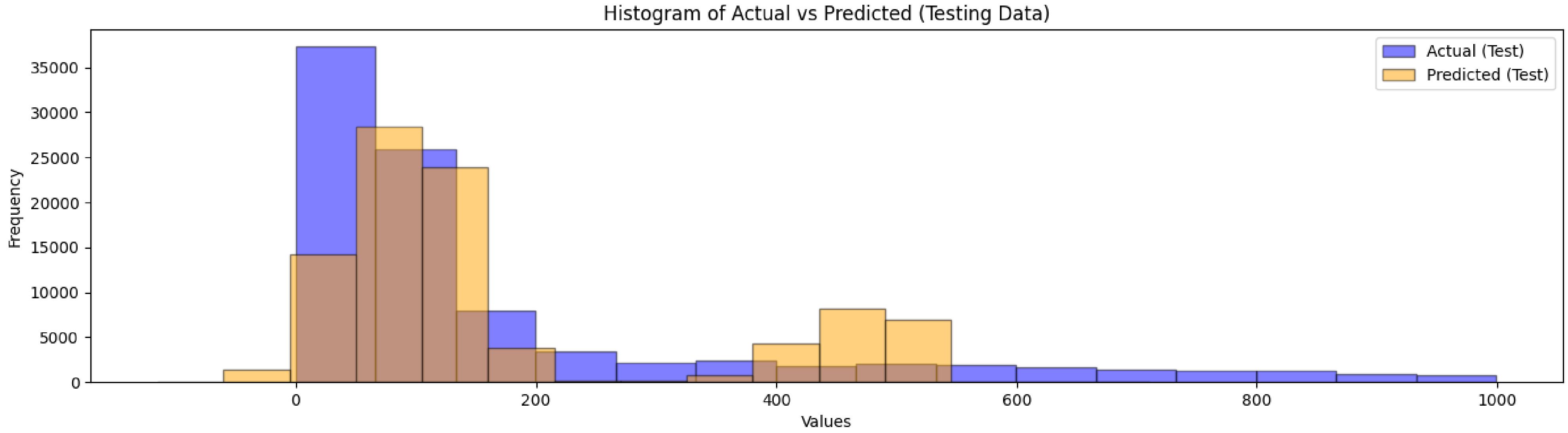


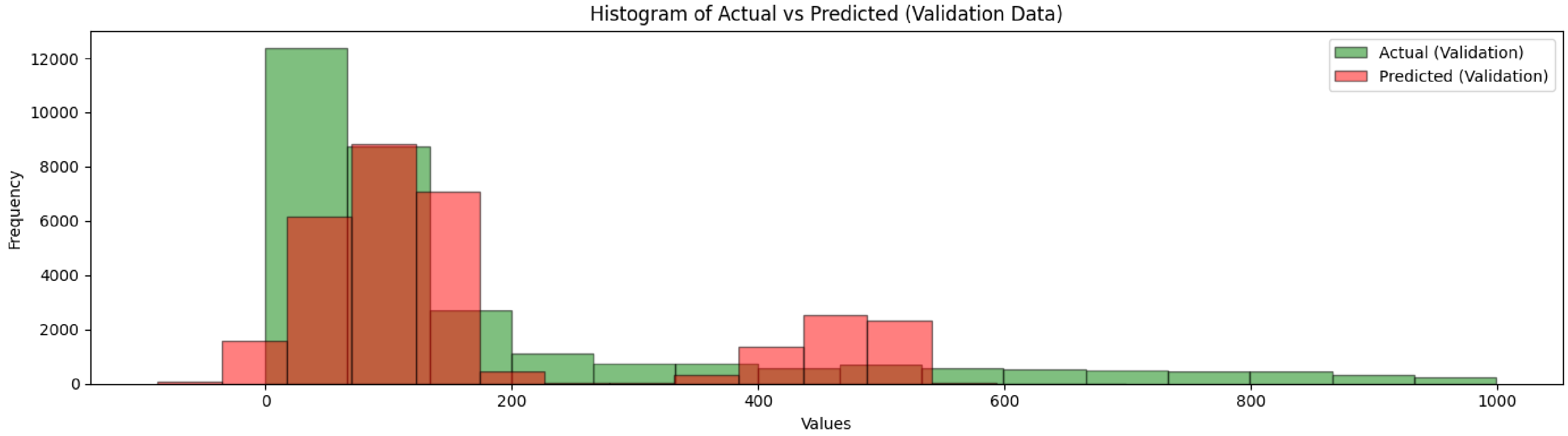


Lasso Regression



Ridge Regression





Model Performance Comparison

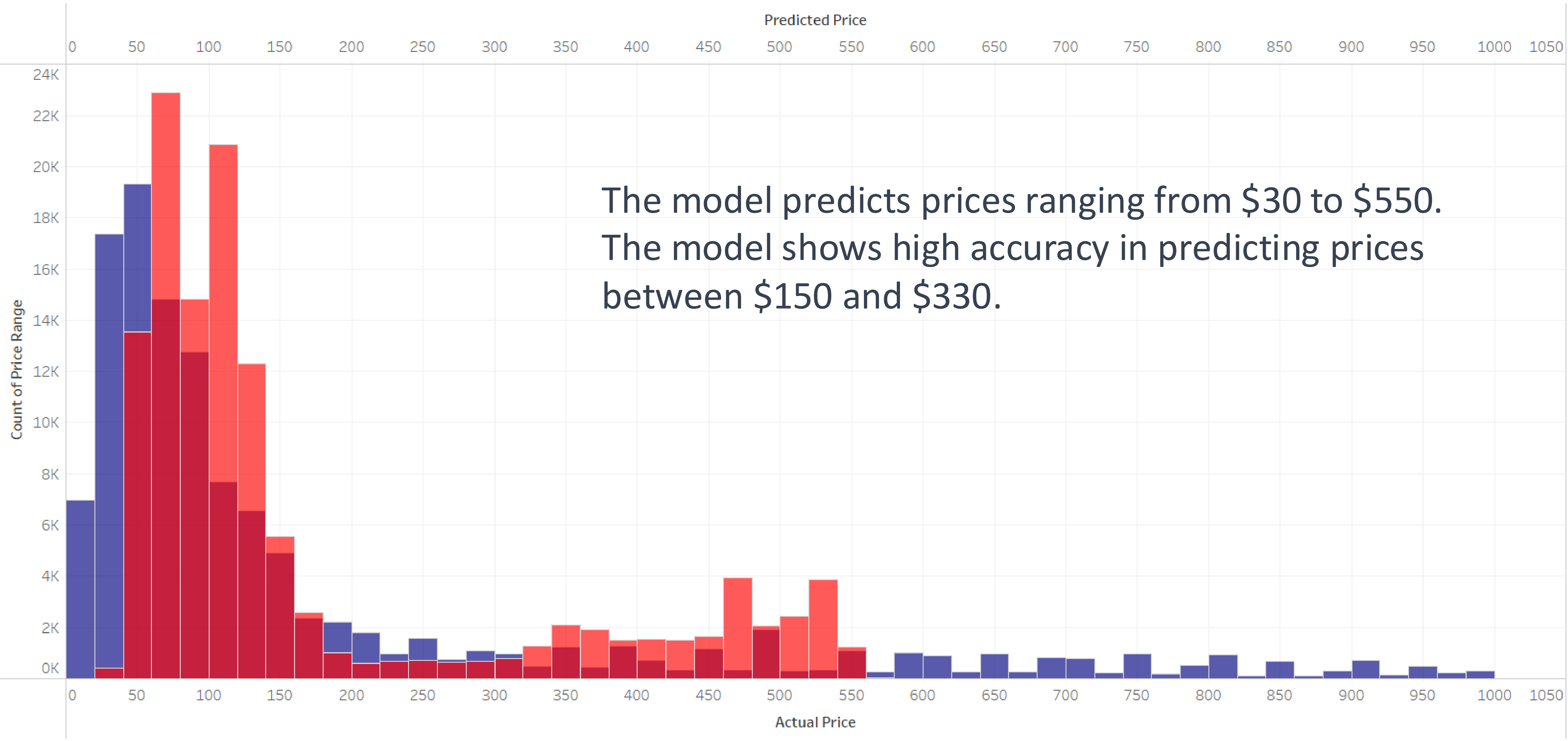
	Model	R^2	Adjusted R^2	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)
Train	Linear Regression	0.55	0.55	21416.28	146.34
	Random Forest Regression	0.61	0.61	18478.35	135.94
	Lasso Regression	0.55	0.55	21416.22	146.34
	Ridge Ridge Regression	0.55	0.55	21416.2	146.34
Validation	**Model**	**R^2**	**Adjusted R^2**	**Mean Squared Error (MSE)**	**Root Mean Square Error (RMSE)**
	Linear Regression	0.56	0.56	20997.67	144.91
	Random Forest Regression	0.6	0.6	18809.03	137.15
	Lasso Regression	0.56	0.56	20997.88	144.91
	Ridge Ridge Regression	0.56	0.56	20997.76	144.91

Winning Model: Random Forest Regression

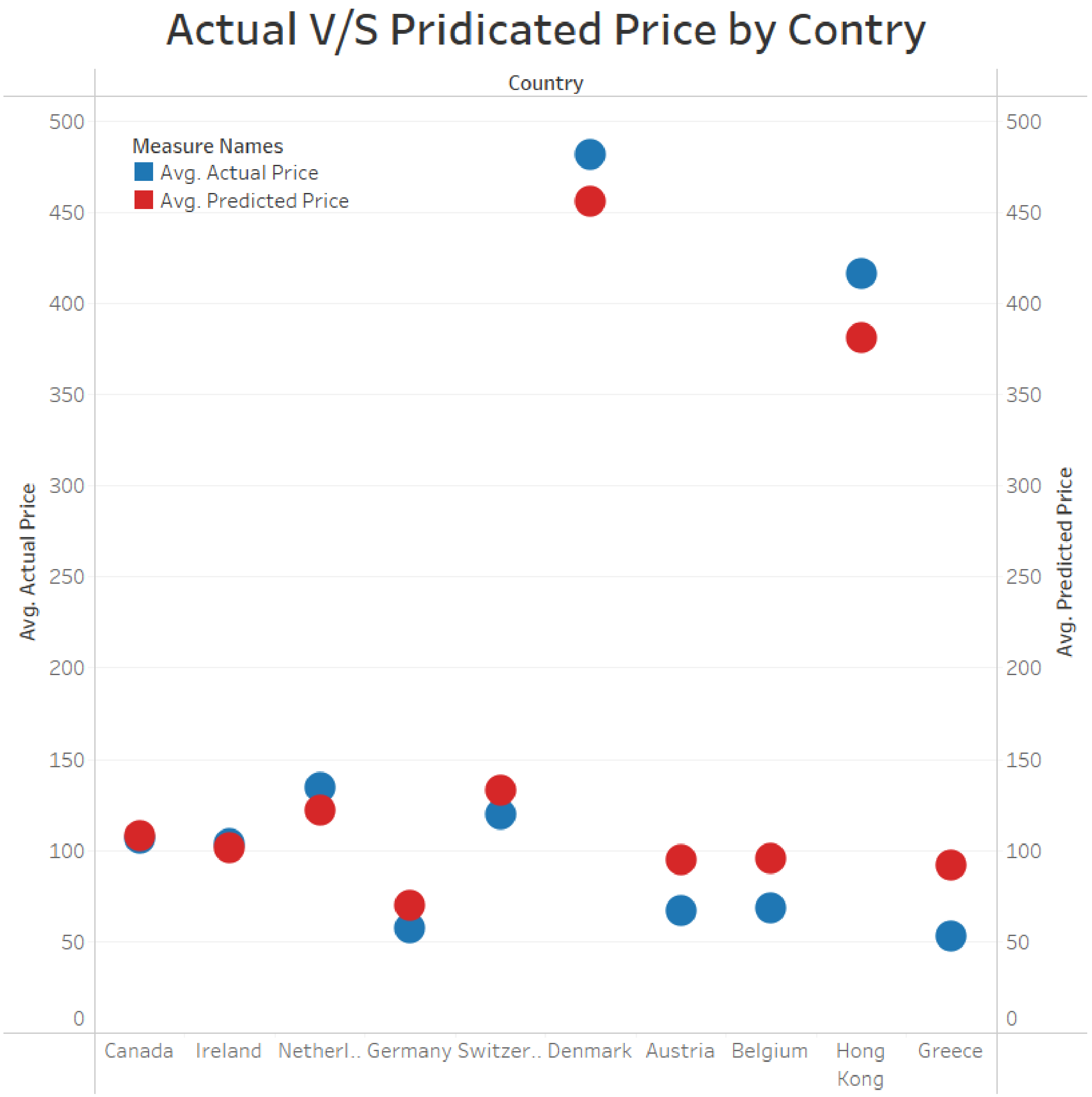
- Key Insights:
  - **Superior Accuracy**: Outperforms other models in both training and validation sets
  - **Higher Explanatory Power**: R^2 of **0.60-0.61 vs 0.55-0.56** for other models
  - **Lower Prediction Errors: ~12% lower MSE and ~6% lower RMSE** compared to other models
  - **Consistent Performance:** Minimal difference between training and validation metrics, indicating good generalization
- Conclusion:
  - Random Forest Regression demonstrates superior performance in:
    - Capturing complex data patterns
    - Making accurate predictions
    - Generalizing well to unseen data

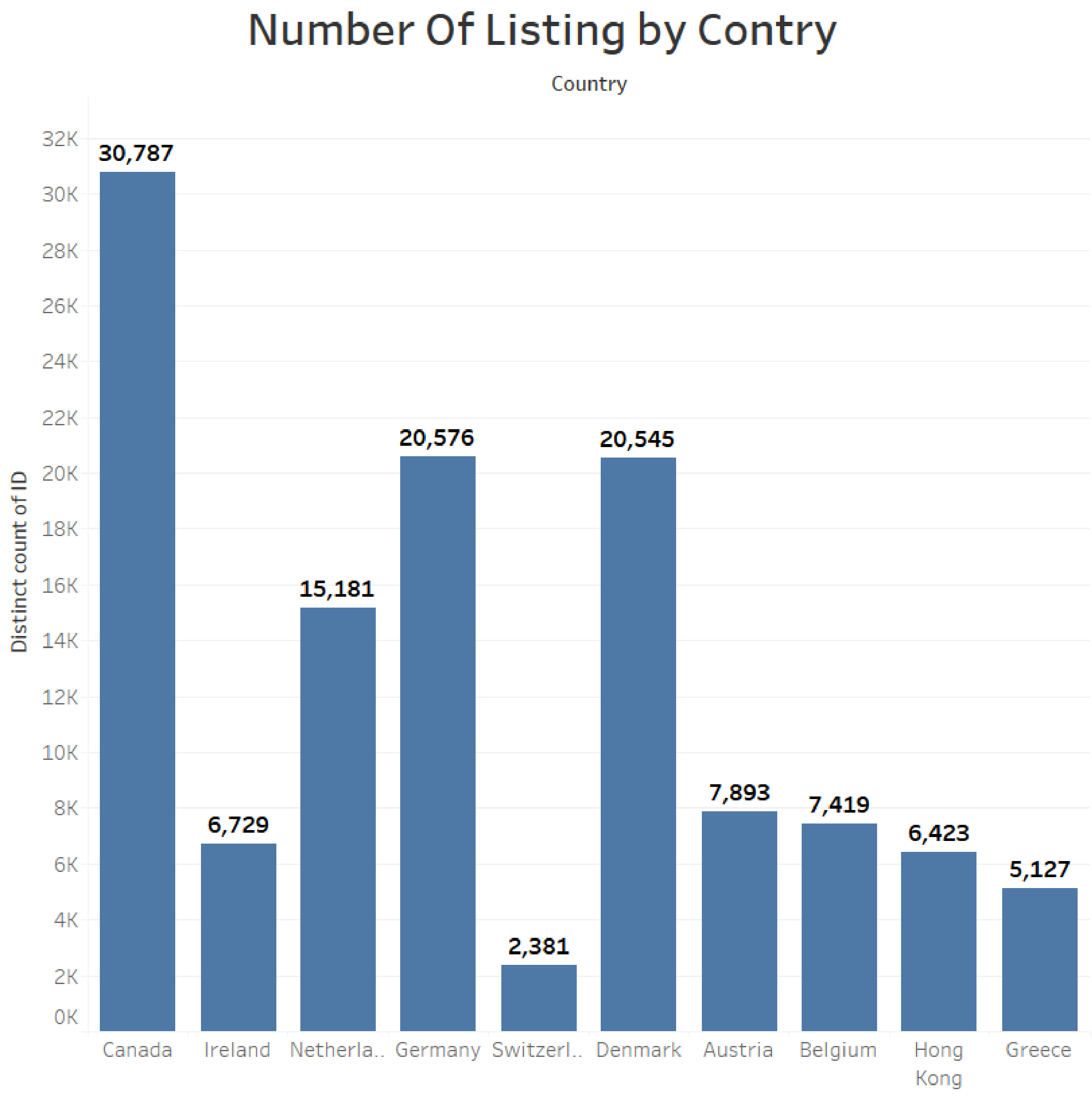
Model-Based Outcome Data Analysis

Actual V/S Pridicated Price Histogram

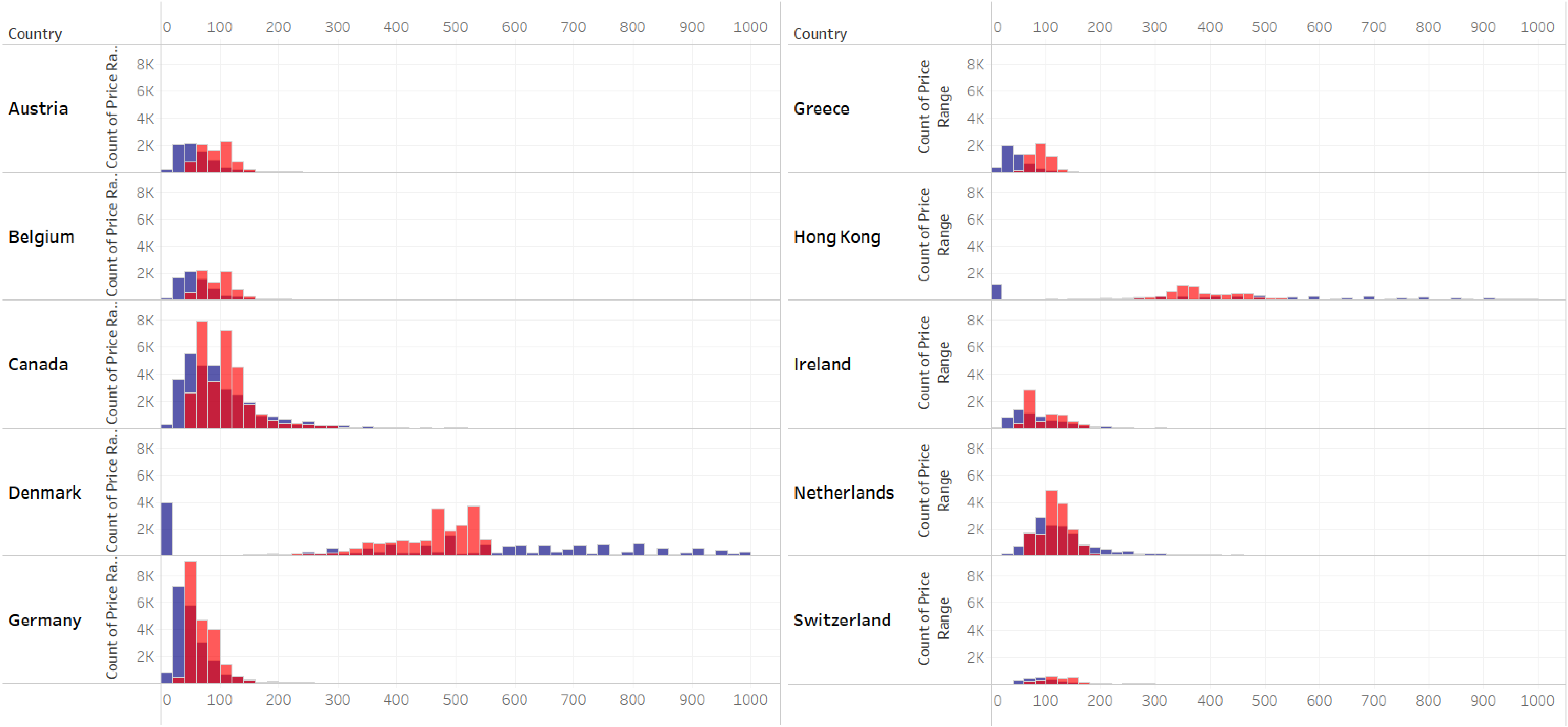


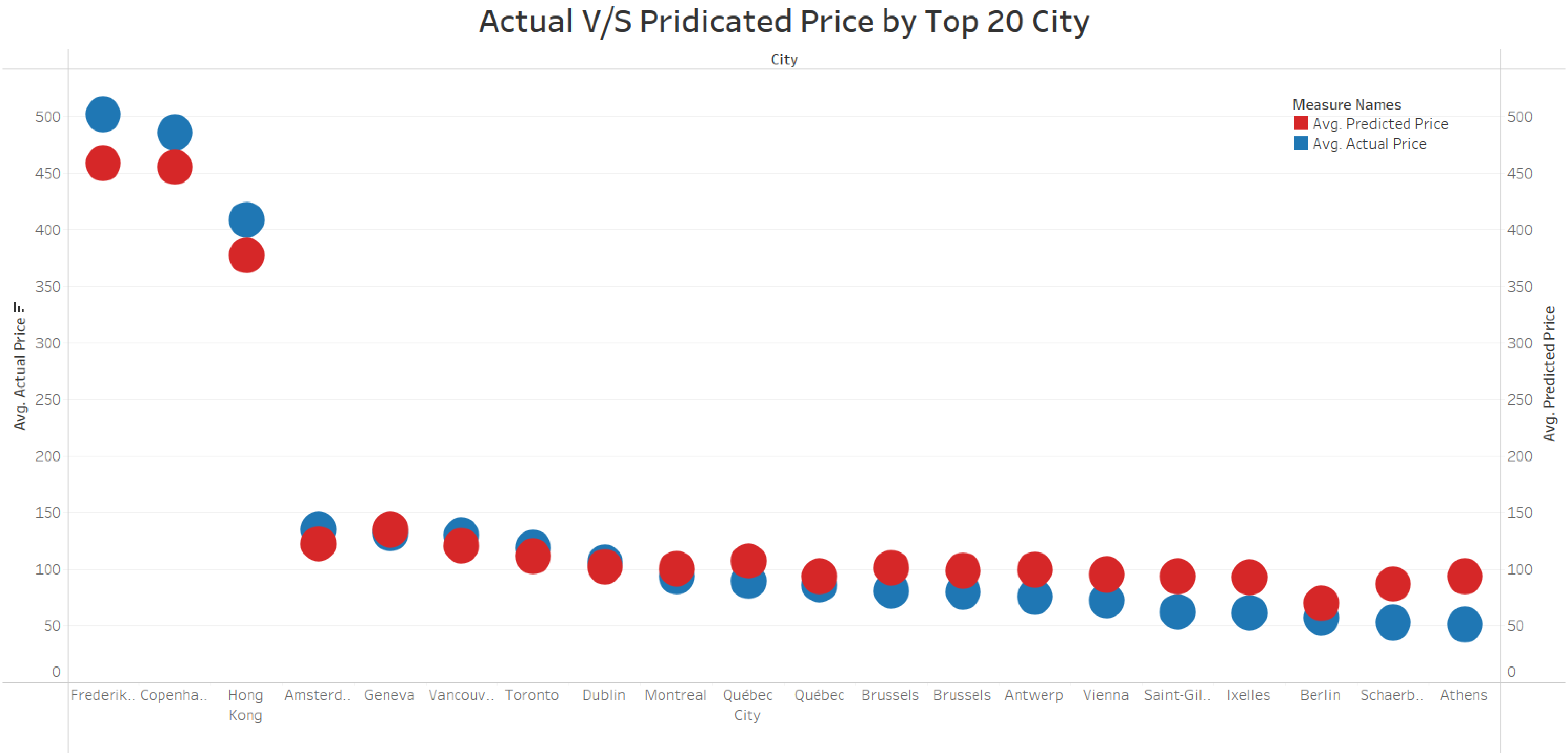
The model predicts prices ranging from $30 to $550. The model shows high accuracy in predicting prices between $150 and $330.


Actual V/S Pridicated Price by Contry
Country
Measure Names
Avg. Actual Price
Avg. Predicted Price
Avg. Actual Price
Avg. Predicted Price
500
450
400
350
300
250
200
150
100
50
0
Canada
Ireland
Netherl..
Germany
Switzer..
Denmark
Austria
Belgium
Hong Kong
Greece


Number Of Listing by Contry
Country
Distinct count of ID
32K
30K
28K
26K
24K
22K
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
30,787
6,729
15,181
20,576
2,381
20,545
7,893
7,419
6,423
5,127
Canada
Ireland
Netherla..
Germany
Switzerl..
Denmark
Austria
Belgium
Hong Kong
Greece

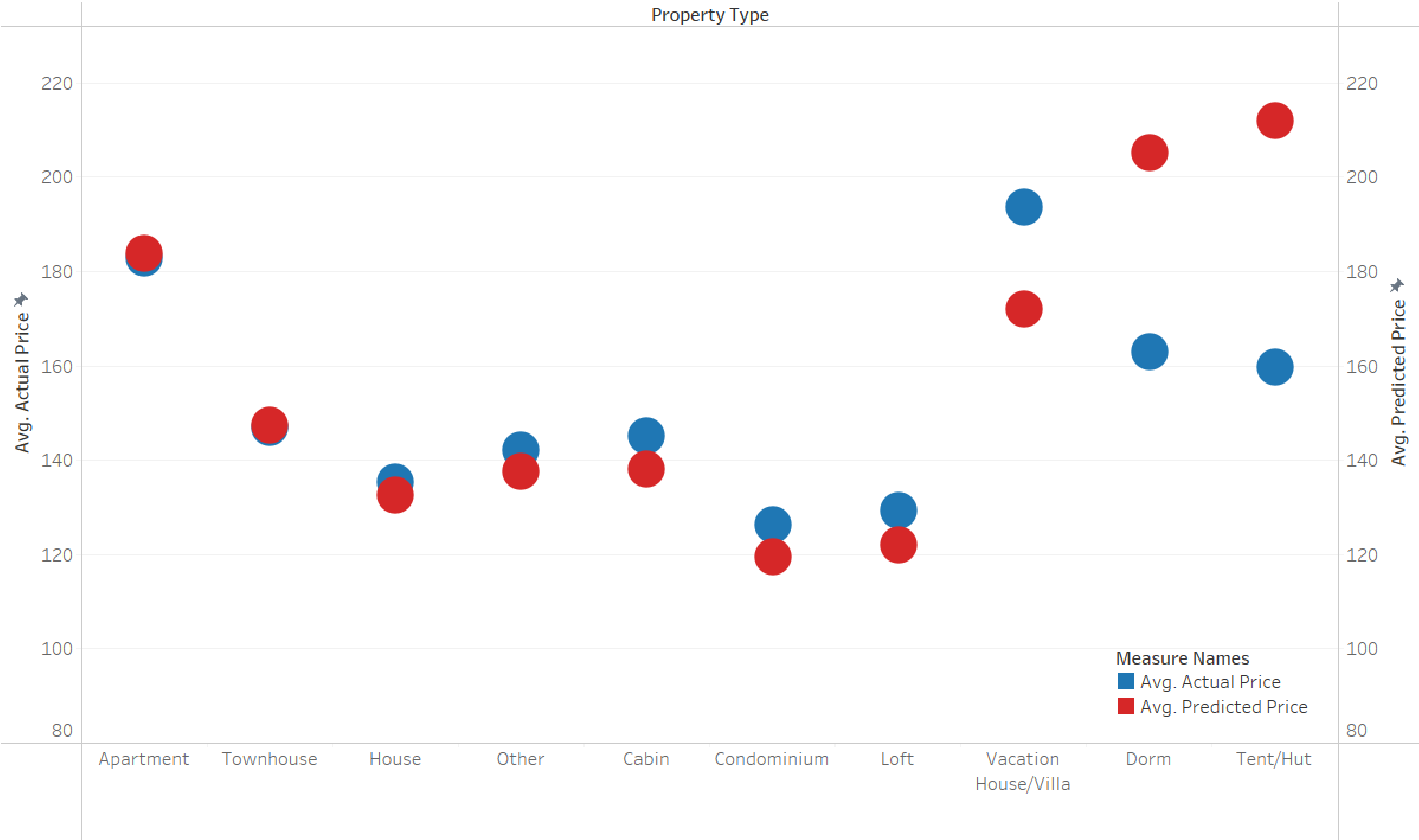
Actual V/S Pridicated Price Histogram by Contry in Detail





In the graph, cities like **Geneva, Toronto, Dublin, Montreal, Québec** and **Berlin** show minimal differences between the actual and predicted average prices, indicating accurate predictions for these locations.

Actual V/S Pridicated Price by Property Type



Number Of Listing by Property Type

Property Type	
Apartment	97,592
House	16,210
Condominium	3,879
Loft	1,603
Townhouse	1,219
Vacation House/Villa	954
Other	649
Dorm	544
Cabin	388
Tent/Hut	23

In the chart, the **predicted prices** for **Apartment** and **Townhouse** closely match their **actual prices**, indicating accurate predictions for these property types. For **House**, the predicted price is also quite close to the actual price, showing a high level of prediction accuracy.

Conclusion

- **Canada**, **Netherlands**, and **Germany** are the top choices for investment, offering reliable price predictions and stable markets, particularly in **Toronto**, **Amsterdam**, and **Berlin**.
- Focus on major cities like **Toronto**, **Vancouver**, **Amsterdam**, **Rotterdam**, **Berlin**, and **Munich** where demand and price accuracy are high.
- Invest in **apartments** and **townhouses** for reliable returns. **Houses** are viable, but may have slight discrepancies in price predictions.

Any Questions?