House Prices in Argentina

JonPaul Ferzacca, Joey Musholt

Project Description

In the Argentine real estate landscape, our dataset offers a robust .csv compilation of property listings including over 1 million data points. Through this data, we endeavor to discern key trends such as price prediction based on property attributes like bedroom count and size, the correlation between listing duration and price, and the interplay between price and location. This exploration can give insights for local governance and real estate investors, guiding data-informed housing decisions.

Prior Work

- Housing Price Predictor
  - Mapping of Data
  - Measuring Feature Importance
  - https://www.kaggle.com/code/msorondo/a-housing-price-predictor-for-buenos-aires-c

Data Source

- 2 million rows
- 25 potential features
- Accessed via Kaggle:

  https://www.kaggle.com/datasets/msorondo/argentina-venta-de-propiedades/data

- Originally sourced from Properati (property listing platform):

  https://www.properati.com.ar/

- Downloaded on both PCs

Tools

- Python
- NumPy
- Pandas
- Sklearn
- MatPlotLib
- Seaborn
- Github
- Geopy

Techniques

Data Cleaning

- Cleaning of data types
- Replacing NaNs with imputed medians
- Removing data objects with too many missing features

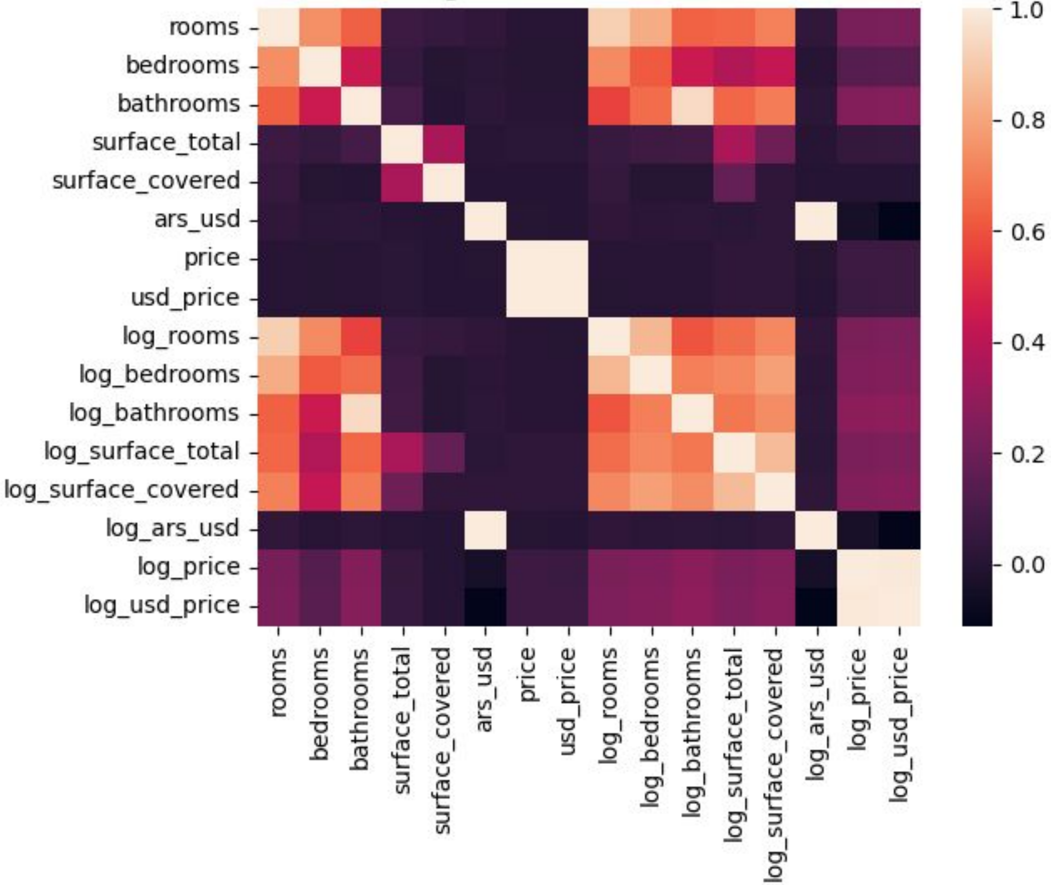
Data Preprocessing

- Creation of Log Features, including Price
- Feature Engineering
  - Using start/end/posting dates to calculate listing age and duration
  - Using lat/long to calculate distance from Buenos Aires city center

Model Building

- Linear Regression with one-hot encoding for property_type and operation_type
- Settled on predicting log_price to encompass non-linearities

Fig 3: Correlation Matrix



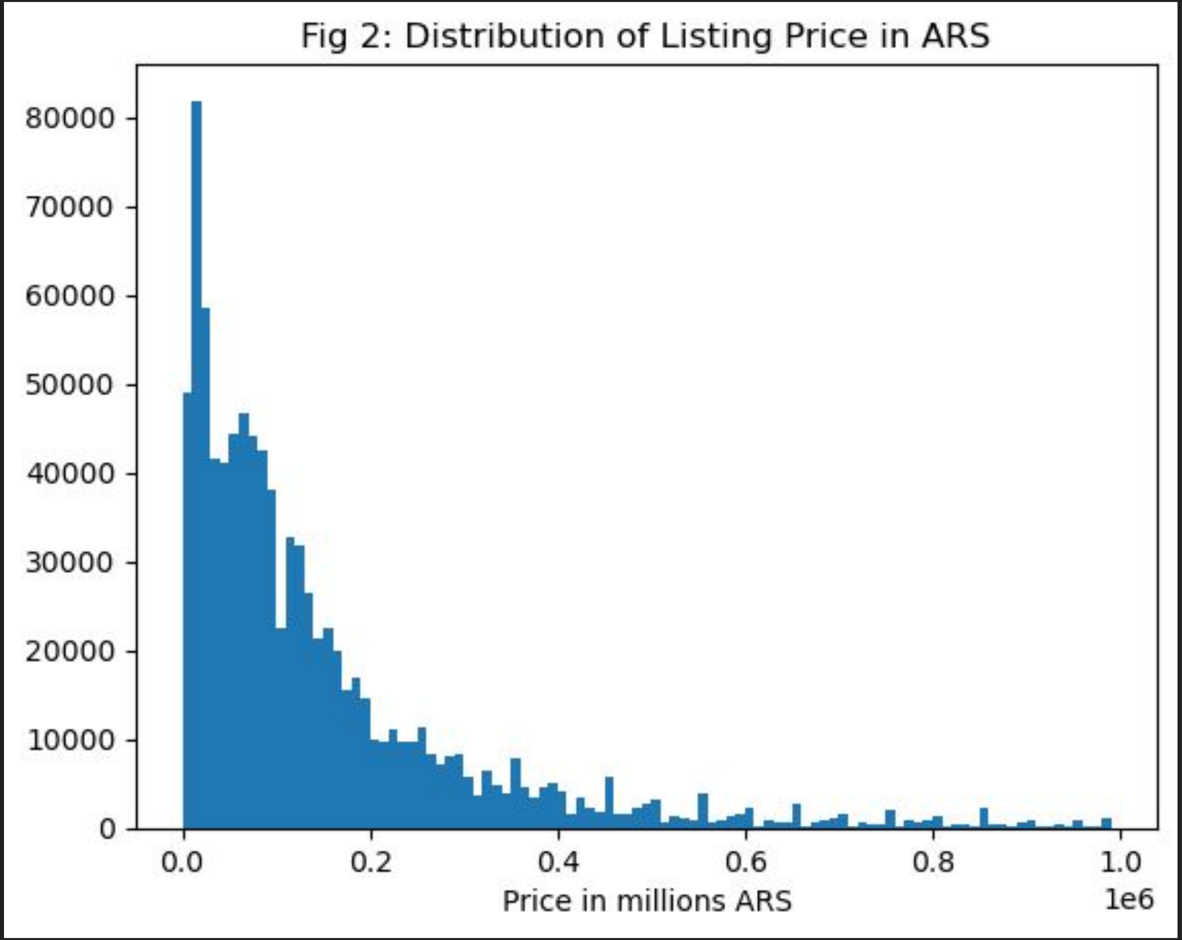
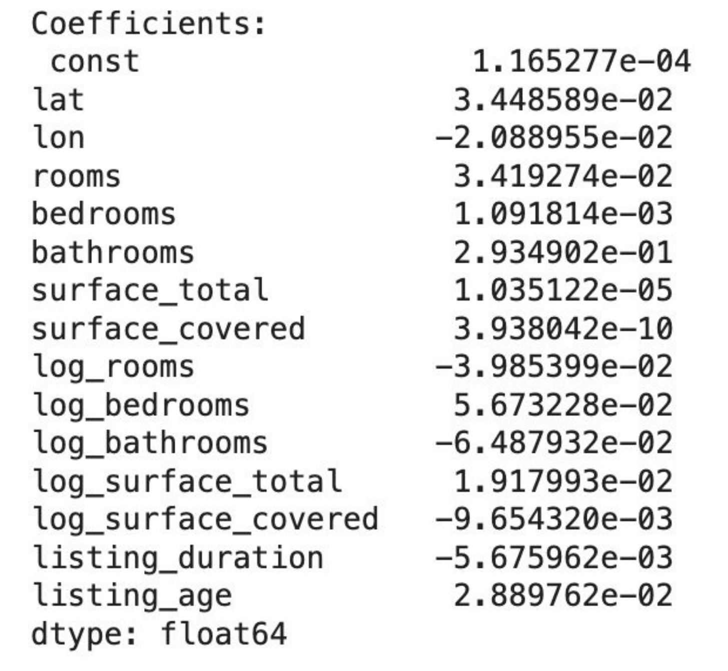
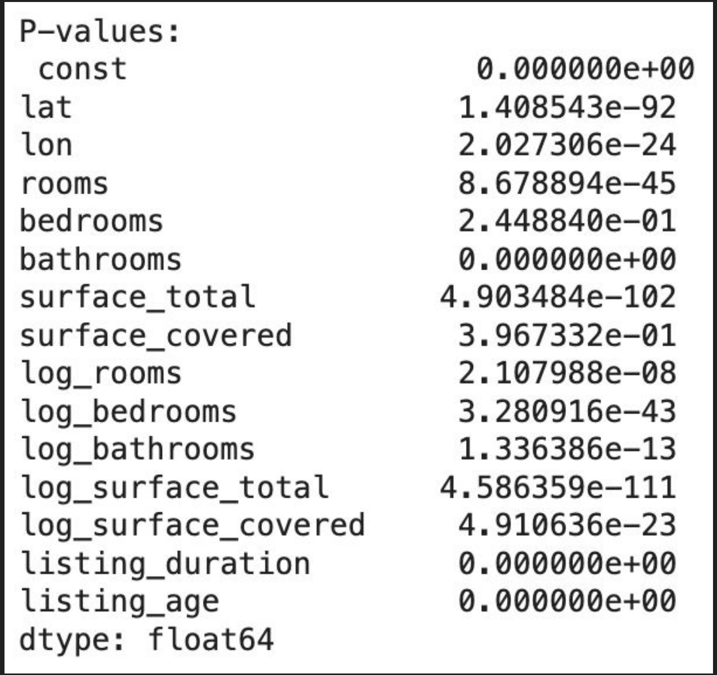


Fig 2: Distribution of Listing Price in ARS
80000
70000
60000
50000
40000
30000
20000
10000
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Price in millions ARS

Model Evaluation



```
Coefficients:
 const                   1.165277e-04
lat                     3.448589e-02
lon                    -2.088955e-02
rooms                   3.419274e-02
bedrooms                1.091814e-03
bathrooms               2.934902e-01
surface_total           1.035122e-05
surface_covered         3.938042e-10
log_rooms              -3.985399e-02
log_bedrooms            5.673228e-02
log_bathrooms          -6.487932e-02
log_surface_total       1.917993e-02
log_surface_covered    -9.654320e-03
listing_duration       -5.675962e-03
listing_age             2.889762e-02
dtype: float64
```



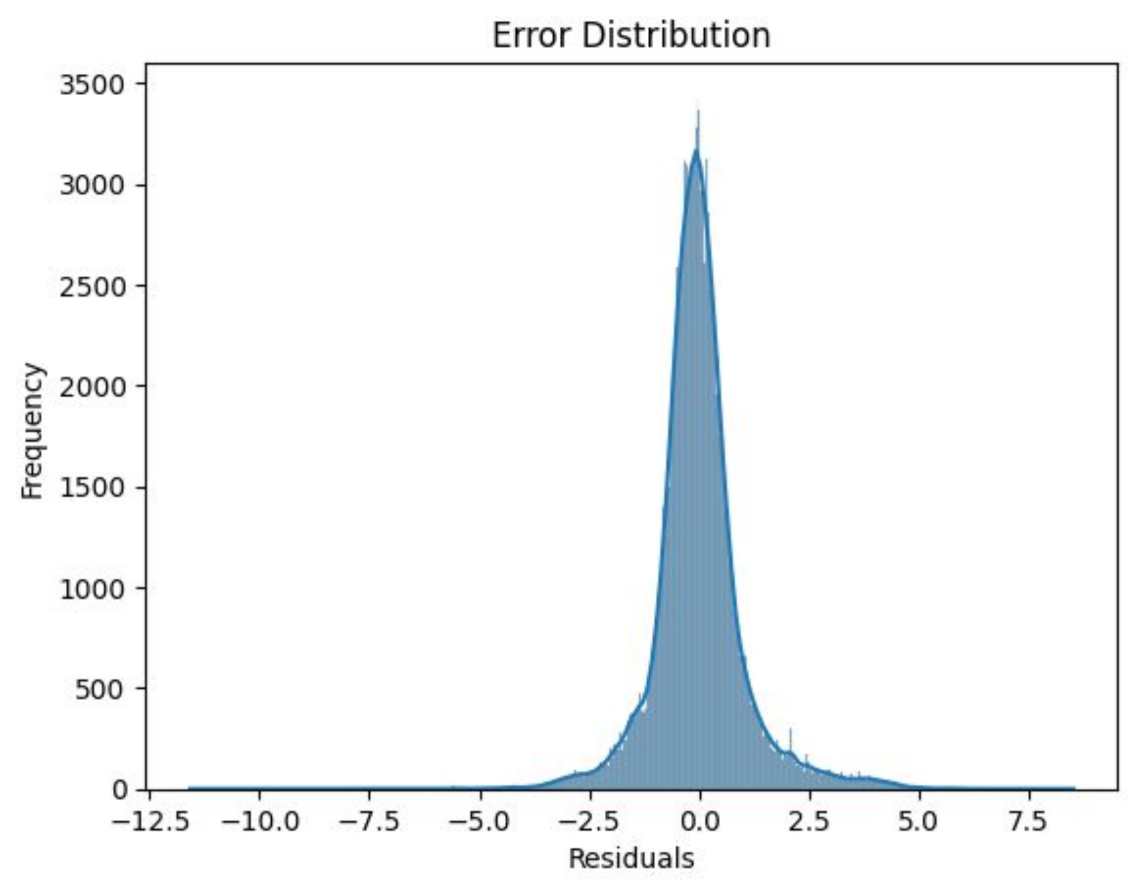
```
P-values:
 const                    0.000000e+00
lat                      1.408543e-92
lon                      2.027306e-24
rooms                    8.678894e-45
bedrooms                 2.448840e-01
bathrooms                0.000000e+00
surface_total           4.903484e-102
surface_covered          3.967332e-01
log_rooms                2.107988e-08
log_bedrooms             3.280916e-43
log_bathrooms            1.336386e-13
log_surface_total       4.586359e-111
log_surface_covered      4.910636e-23
listing_duration         0.000000e+00
listing_age              0.000000e+00
dtype: float64
```

Model Evaluation

Mean Absolute Error (MAE): 0.6768588295130806

Mean Squared Error (MSE): 1.0080951295549976

R-squared (R^2): 0.5062517425548707


Error Distribution
Frequency
3500
3000
2500
2000
1500
1000
500
0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
Residuals