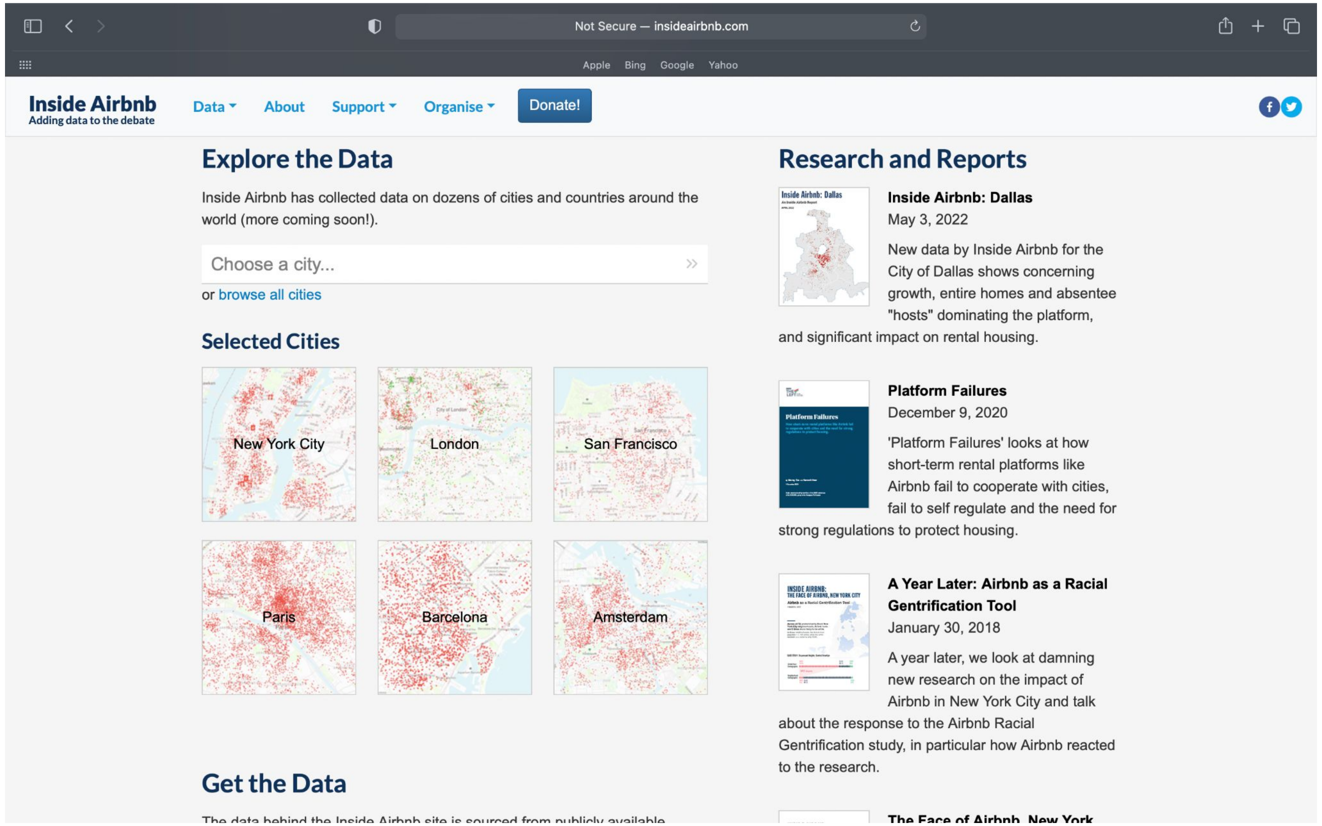
Airbnb Dataset Analysis

Chris Monteleone

About the Data!



- Where the Project Originates
    - Inside Airbnb, a project that provides data about Airbnb's impact on resident communities
    - Launched in 2016 as an investigatory website to scrape and report data on Airbnb
    - Created to reveal illegal renting on the site and gentrification caused by landlords buying properties to rent on Airbnb
    - *Gentrification: Process of changing the character of a neighborhood through the influx of more affluent residents and businesses.*

Why the Airbnb Dataset?

- Question I Found Interesting: Can we use the Inside Airbnb Dataset to make more educated investments into Airbnb?
- Even more interesting: Can we do this for the Clark County area?
- Yes! Data is available through Inside Airbnb

What’s Included in the Data

```
[47]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14808 entries, 0 to 14807
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              14808 non-null  int64
 1   name                            14808 non-null  object
 2   host_id                         14808 non-null  int64
 3   host_name                       14369 non-null  object
 4   neighbourhood_group             0 non-null      float64
 5   neighbourhood                   14808 non-null  object
 6   latitude                        14808 non-null  float64
 7   longitude                       14808 non-null  float64
 8   room_type                       14808 non-null  object
 9   price                           14808 non-null  int64
 10  minimum_nights                  14808 non-null  int64
 11  number_of_reviews               14808 non-null  int64
 12  last_review                     10992 non-null  object
 13  reviews_per_month               10992 non-null  float64
 14  calculated_host_listings_count  14808 non-null  int64
 15  availability_365                14808 non-null  int64
 16  number_of_reviews_ltm           14808 non-null  int64
 17  license                         408 non-null    object
dtypes: float64(4), int64(8), object(6)
memory usage: 2.0+ MB
```

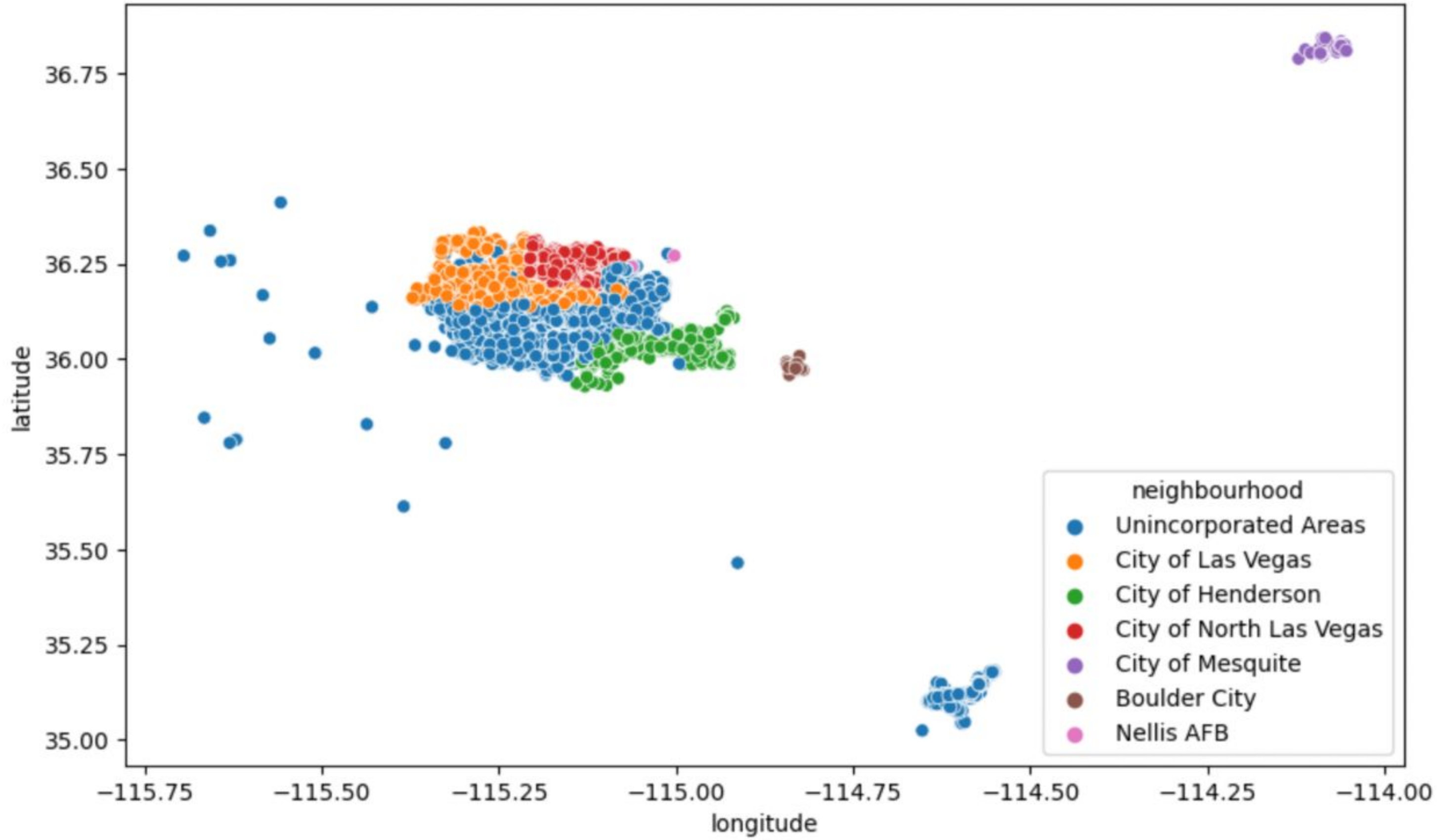
Data That I Found Useful:

- neighbourhood
- latitude / longitude
- room_type
- price
- minimum_nights
- number_of_reviews
- availability_365

Preprocessing:

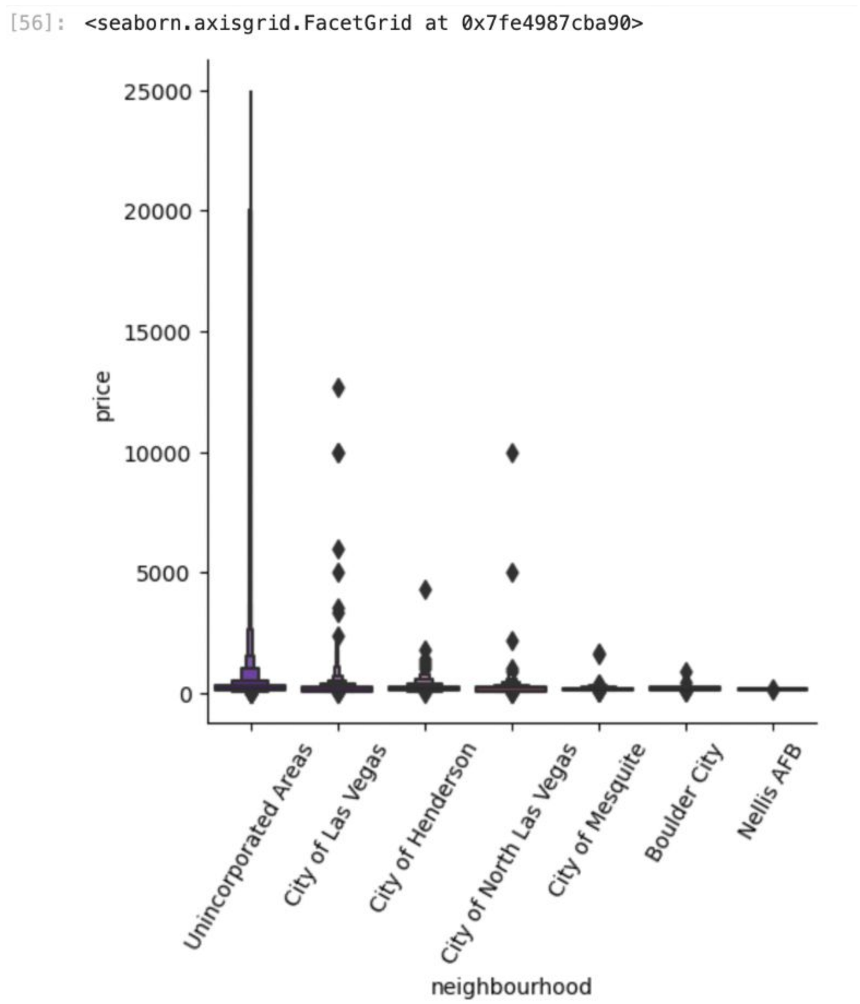
- Remove other columns that were irrelevant

Dataset Distribution By Area



- Clark County, NV Data
  - Includes:
    - Boulder City (Far SE)
    - City of Henderson (SE)
    - City of Las Vegas (NW)
    - City of Mesquite (E)
    - City of North Las Vegas (N)
    - Nellis AFB (Far NE)
    - Unincorporated Areas (Other) [Possibly problematic]
- This is important because…

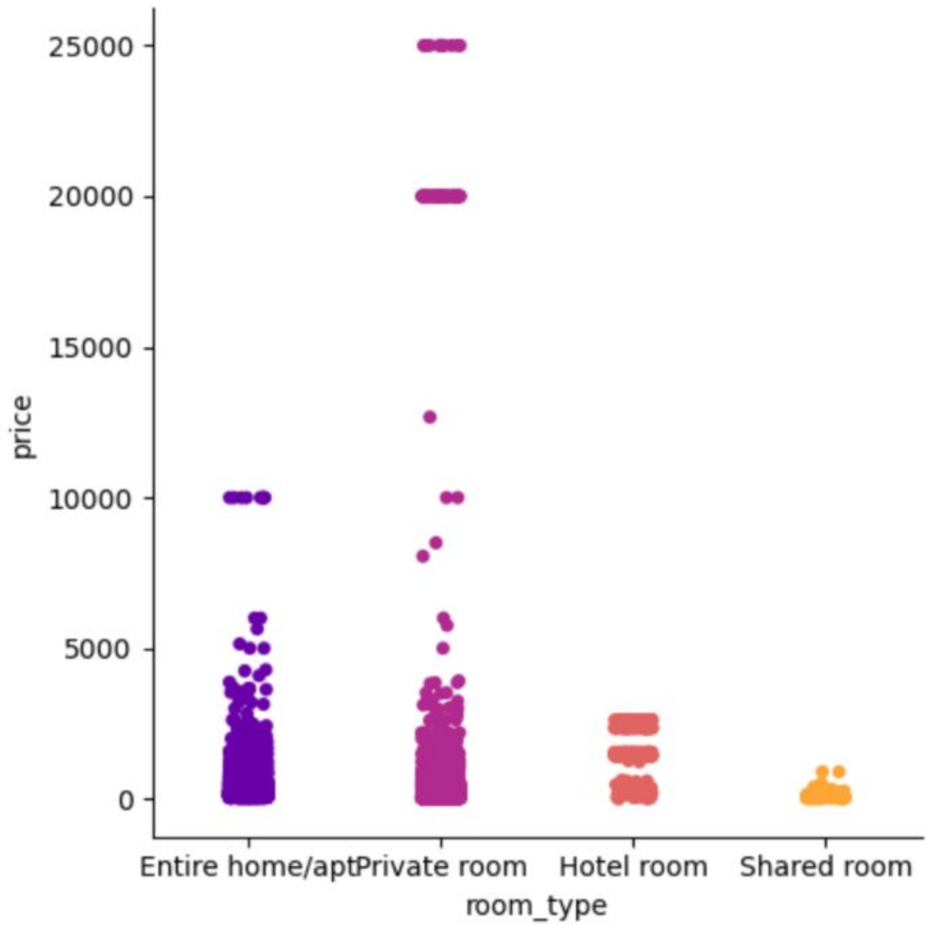
Analysis: Price by Area



- "Unincorporated Areas" Slightly Problematic
  - Not local to one geographic region, so it's difficult to draw conclusions based on the price data
- Still useful to compare other areas
  - Highest Price: City of Las Vegas
  - Lowest Price: Nellis AFB

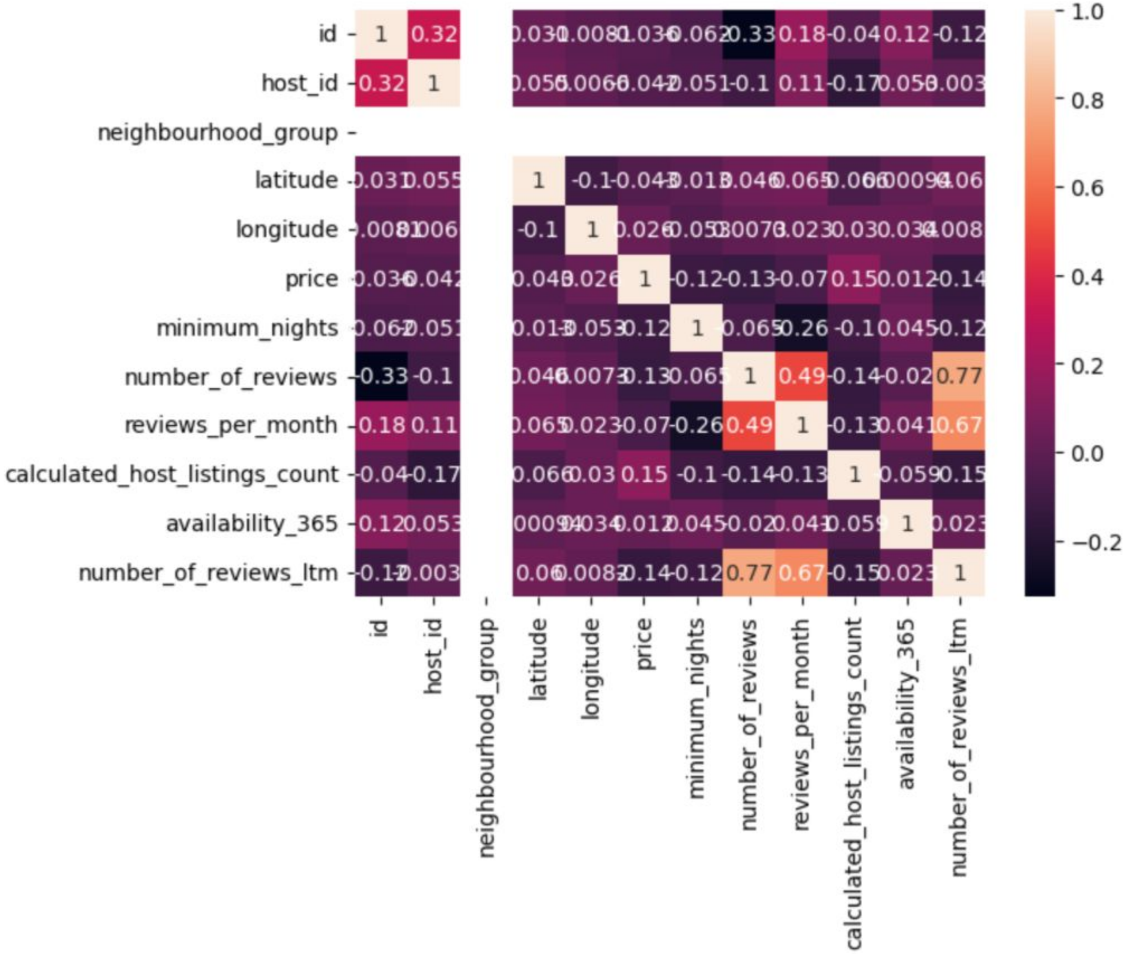
Analysis: Price By Type

[51]: <seaborn.axisgrid.FacetGrid at 0x7fe46b0d37c0>



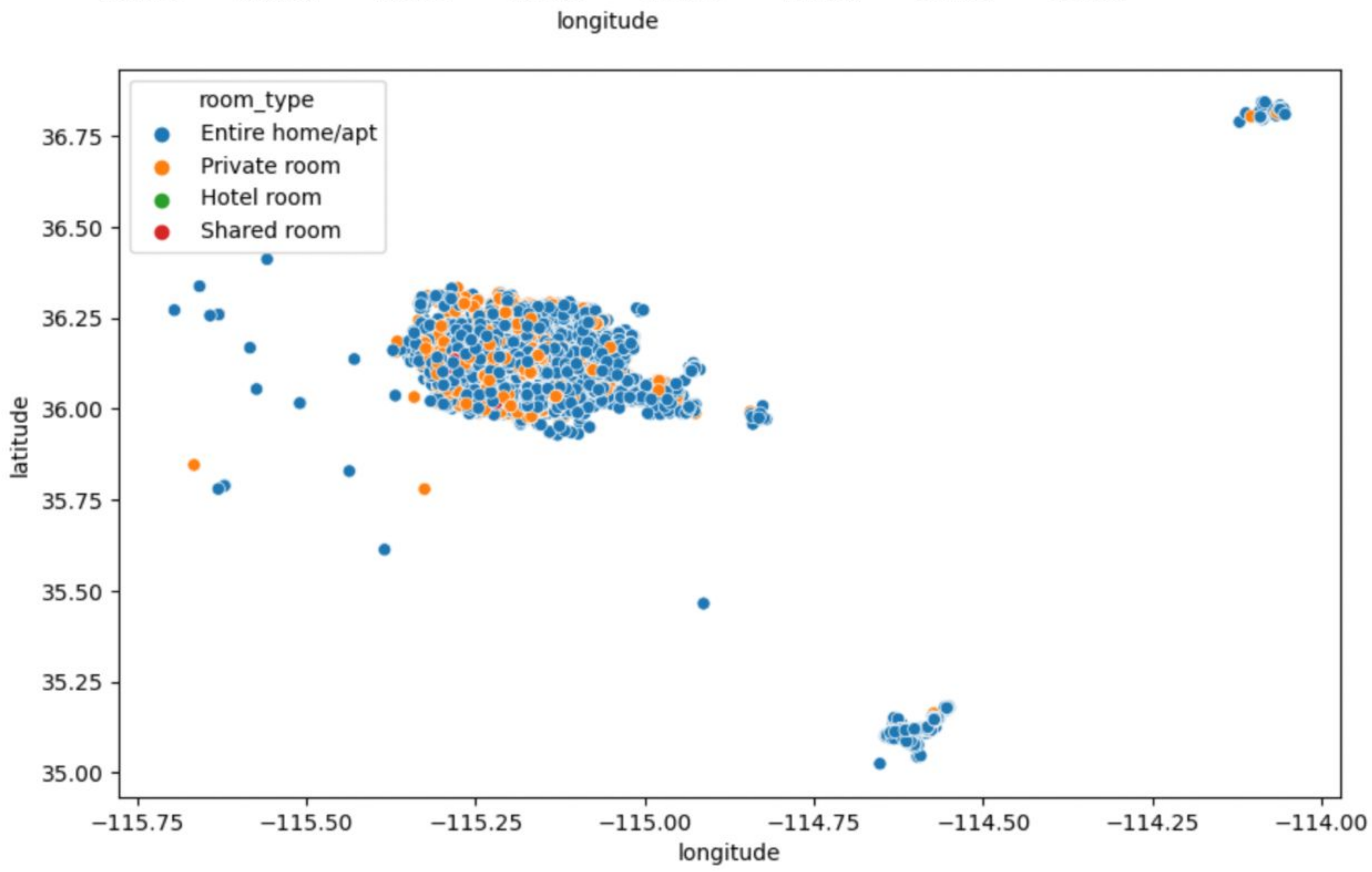
- Entire Home:
  - Most common but caps at a lower price
- Private Room:
  - Highest prices at 25000 per night
- Hotel / Shared Room:
  - Lowest pricing
  - Less common

Correlations



- Difficult to interpret in this way
- Inspires further exploration of correlated variables
- Price category especially important
- Will dive deeper into these correlations with linear regression!

Analysis: Type By Area



- Private room more common on the West side than East
- Entire Home / Apt seems to be more common on the East side
- Hotel room / shared room very uncommon (unable to see any red / green)

Linear Regression: Number of Reviews VS Price (Preparation)

Hypothesis: As number of reviews increases, this implies popularity, meaning the owner can charge more for their listings, increasing price.

[84]:
```
#independent variable
X = np.array(data['number_of_reviews']).reshape(-1,1)
#dependent variable
y = np.array(data['price']).reshape(-1,1)
```

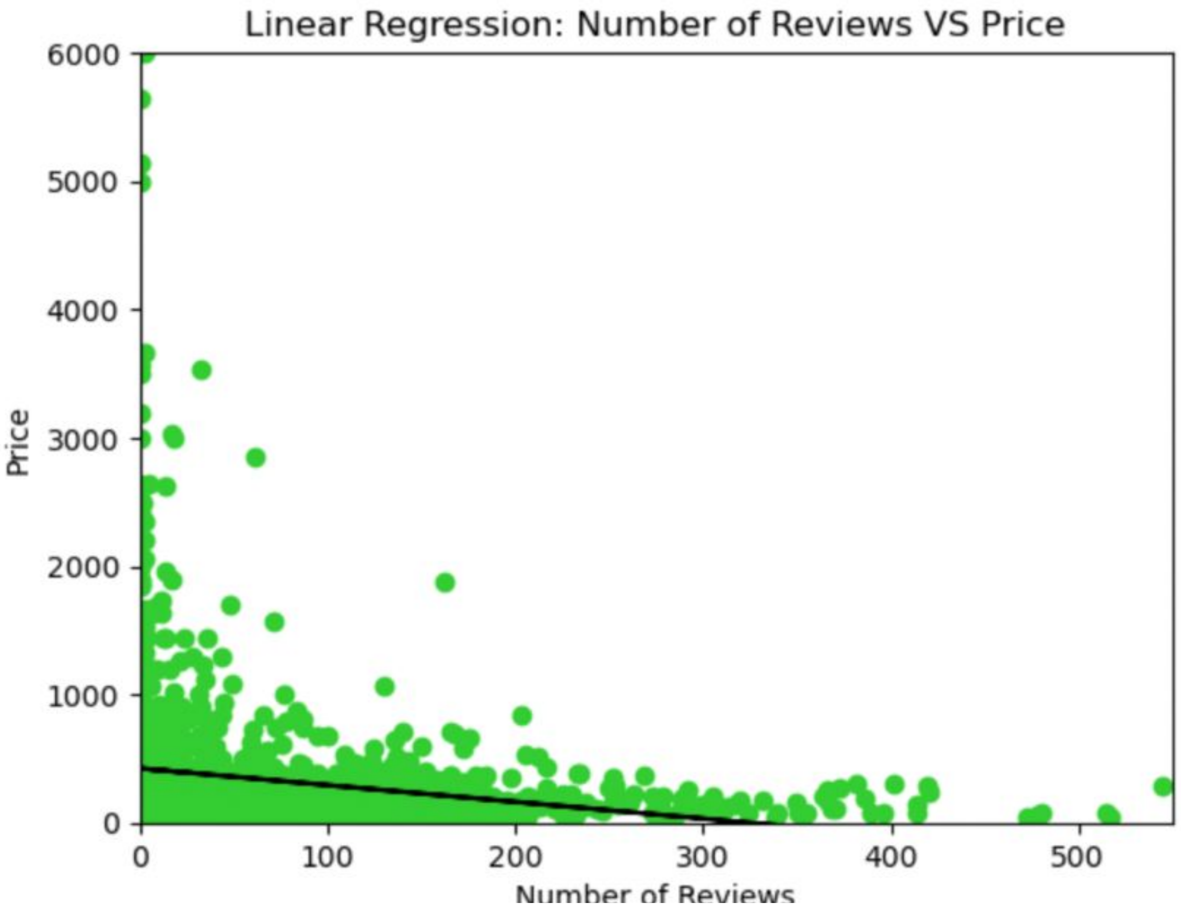
[106…
```
#Split into test and train data
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.25)
#Regression Type
regr = LinearRegression()
```

[114]:
```
#Fit
regr.fit(X_train, y_train)
print(regr.score(X_test,y_test))

#Prediction
y_pred = regr.predict(X_test)

plt.scatter(X_test, y_test, color ='limegreen')
plt.plot(X_test, y_pred, color ='k')
plt.title('Linear Regression: Number of Reviews VS Price')
plt.xlabel('Number of Reviews')
plt.ylabel('Price')
plt.show()
```

- Preparation:
  - Take independent variable as X, dependent variable as Y
  - Split into test and train
  - Fit and predict
- Prediction:
  - As number of reviews increase, price will increase as well.
  - Positive correlation.
- Implications if True:
  - New Airbnb listings might have to charge less since they're less "credible" (less reviews), and can charge more over time.
- What's your guess? (Accept / Reject)

Linear Regression: Number of Reviews VS Price

~~Hypothesis: As number of reviews increases, this implies popularity, meaning the owner can charge more for their listings, increasing price.~~



- Hypothesis Rejected
    - As number of reviews increase, price decreases
- Possible Reasons:
    - Expensive properties are rented less, providing less opportunity for reviews
- Implications:
    - New Airbnb listings that are not highly reviewed will not have to charge less

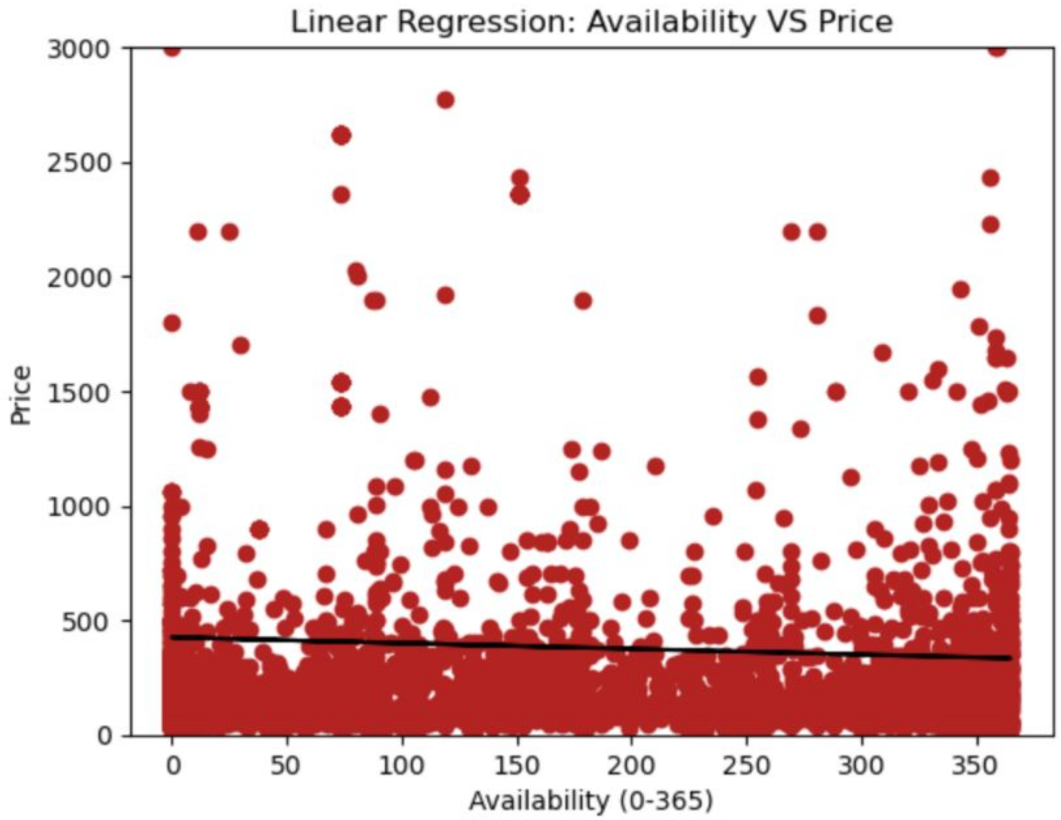
Linear Regression (2): Availability VS Price

Hypothesis: Properties that are available more are more popular and receive more demand through repeat renters, and are therefore able to charge more.

- Independent variable:
  - Availability: The number of days that a property is listed per year from 0 to 365
- Dependent variable:
  - Price
- My prediction: Higher availability means higher price
  - Do you have a prediction?

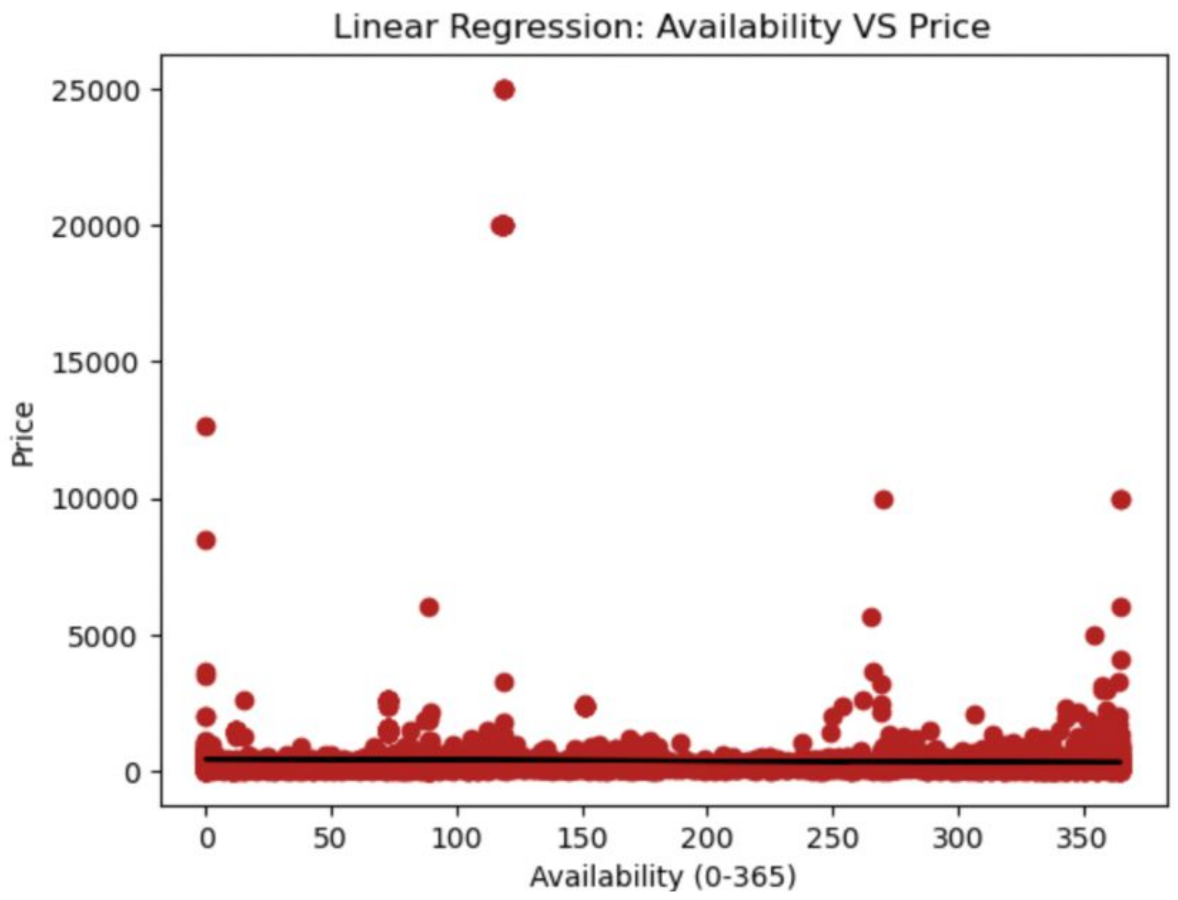
Linear Regression (2): Availability VS Price

~~*Hypothesis: Properties that are available more are more popular and receive more demand through repeat renters, and are therefore able to charge more.*~~



- Hypothesis Rejected
  - As availability increases, price decreases slightly
- Possible Reasons:
  - Properties that are available more often are able to generate more revenue per month, and listers are able to charge less
- Implications:
  - Seasonal Properties
  - If a property is only listed during a certain season or time of year, they won't expect to be able to afford a significant price increase

Sidenote: Scaling!



- Almost didn’t see the slight negative correlation because of scaling of the y-axis
- Initially thought that this had no correlation!

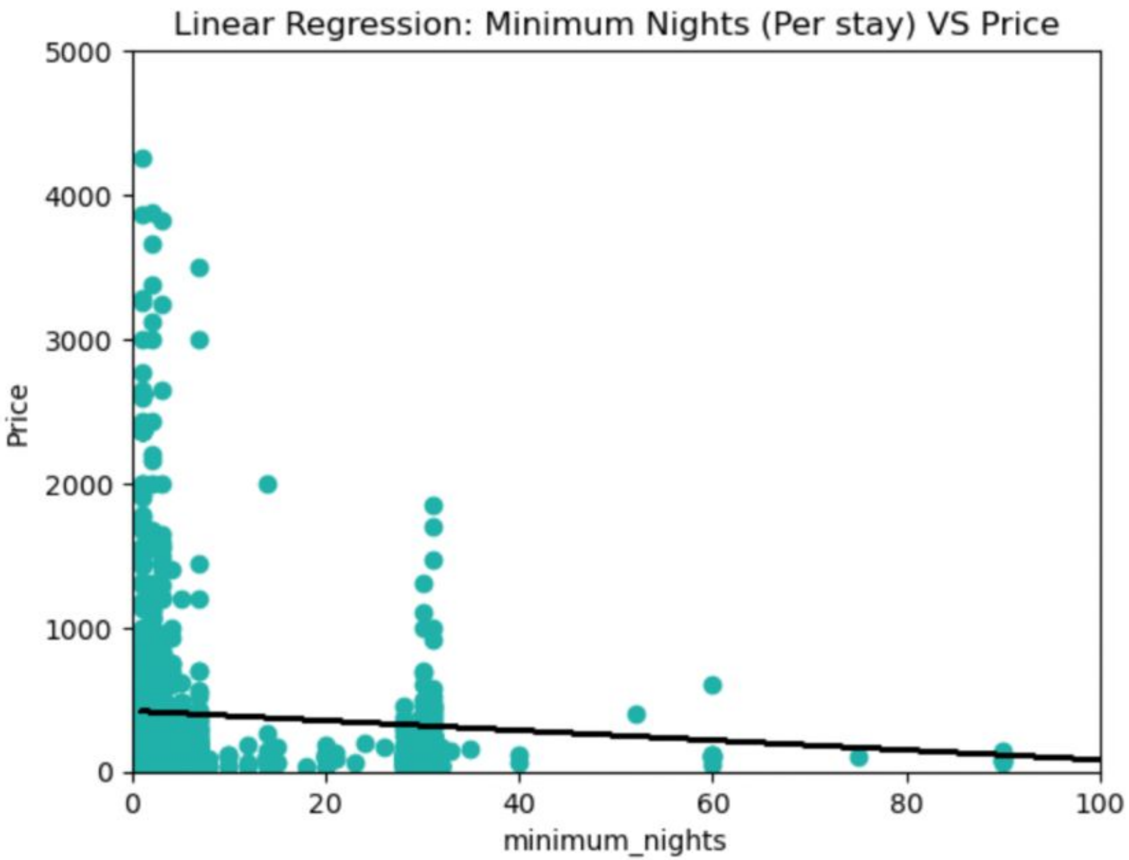
Linear Regression (3): Minimum Nights VS Price

Hypothesis: As minimum nights increases, owners can afford to charge less, since there are less fees that normally occur between guests (cleaning, etc)

- Independent variable:
    - Availability: The number of days that a property is listed per year from 0 to 365
- Dependent variable:
    - Price
- My prediction: Higher availability means higher price
    - Do you have a prediction?

Linear Regression (3): Minimum Nights VS Price

Hypothesis: As minimum nights increases, owners can afford to charge less, since there are less fees that normally occur between guests (cleaning, etc)



- Hypothesis Accepted!
  - As minimum nights increases, price decreases
- Possible Reasons:
  - Properties that rotate guests more often have to hire cleaners or other help that increase price. Without this added cost, owners can charge less.
- Implications:
  - If owners are able looking to increase their prices, they may want to decrease their minimum stay.

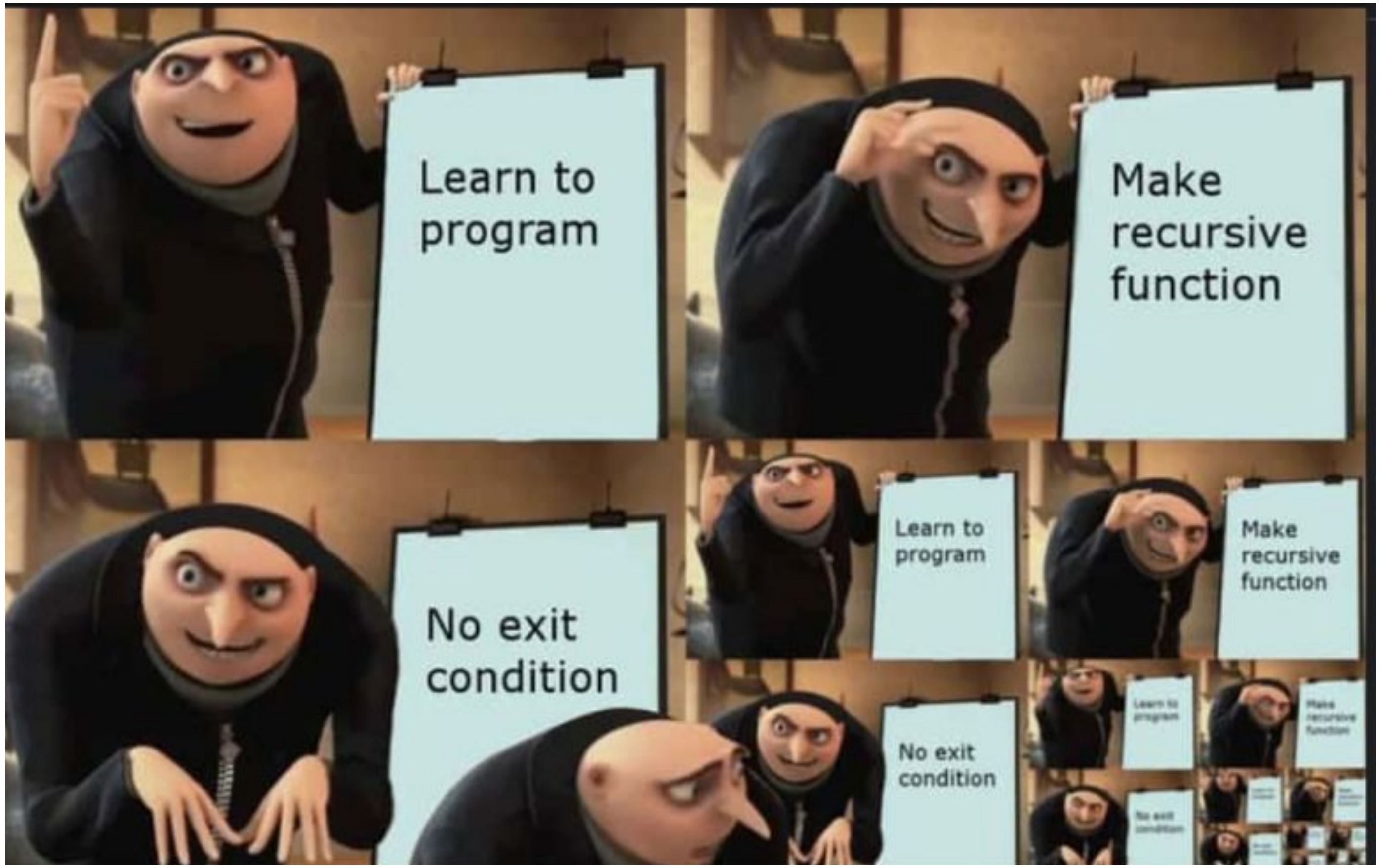
Conclusion Summary

Conclusion: We can make unexpected connections using Data Visualization and the Inside Airbnb Dataset

- What we did:
  - Analysis of Data
    - Important features
    - Type By Area
    - Distribution By Area
  - Correlation
    - To find linked variables
  - Linear Regression
    - Generated Hypothesis
    - Number of Reviews vs Price
    - Availability vs Price
    - Minimum nights vs Price

- Why This is Important:
  - Personally, my guesses were wrong 2 out of 3 times
  - Important to investors that are financially invested to make educated, correct decisions
  - Further exploration can be done in all three areas to ensure correct decision making

Thank You!



Learn to program
Make recursive function
No exit condition
Learn to program
Make recursive function
No exit condition