Advanced Analytics Report
Geospatial Analysis
Airbnb San Diego

Rithwik Chhugani – BS18BDS008
Pratiksha Sharma – BS18BDS010
Gautam Sadarangani – BS18BDS015

Abstract:

In this project, we aim to perform Spatial Exploratory Analysis and Spatial Regression in data from Airbnb San Diego. We find that a spatial implementation with log price of a rental property as the target overtakes the simple regression model by nearly 5%.

Introduction:

Spatial data is any data with direct or indirect reference to a specific location or geographical area. This is often thought of as maps but there is much more spatial data is good for. Things like how we travel, where we live, where do we work are examples of spatial components we deal with on an everyday basis. By performing analysis on these spatial components, we can discover relationships that exist between them. This data can also be fed to spatial regression models to predict values for a dependent variable. For instance, the rent you pay for your accommodation or your office space very much depends on the neighborhood it is situated in, its proximity to say public parks, recreational centers or areas of high accessibility to downtown city regions. This is what makes geospatial data of key importance in solving numerous problems and implementing a vast class of models.

In this project, we aim to look closely at property data from the popular renting and lodging platform – Airbnb for the city of San Diego. This study performs exploratory data analysis (EDA) and spatial regression (SR) with the price of rentals as our dependent variable. We have found that spatial regression provides a significant performance improvement over basic regression.

Motivation:

This is a case of an ongoing temperature study in Antarctica; scientists working at meteorological departments are trying to record the temperature in Antarctica to conduct research on climate change but given the adverse weather condition and geography of the region, they are not able to deploy enough thermo sensors. This leads to the formation of multiple blind spots in their datasets which is effectively solved with the use of SR and clustering to approximate the missing values.

This is just one case study. We came across multiple such creative uses of geospatial data and regression which strongly motivated us to do this project.

Background:

As straightforward as it may look, spatial data is not best suited for regression techniques. In fact, it often violates the assumptions/ requirements of Ordinary least squares (OLS) regression. This makes it important for us to use appropriate diagnostic tools in conjunction with regression tools to make sense of an obvious relation as explained in the examples earlier.

Airbnb is a multi-million dollar renting and lodging platform that allows non-commercial users to access and download their listings data for free. Given their business revolves around geography, they extensively capture, use and publish spatial data in addition to a normal listing information sheet. Effective EDA and SR is possible on such data and useful in providing us with interesting insights.

Dataset Description:

The spatial datasets selected “neighbourhoods.geojson” and “regression_db.geojson” represent the details of Airbnb properties all over the area of San Diego, California that are centered around Balboa Park.

There are a total of 6110 property listings in this data and 108 geometrical records of neighborhoods. Below, we have listed and described some of the features in our datasets:

airbnb_db:

Feature	**Description**
accommodates	the number of people the listing can accommodate
bathrooms	number of bathrooms in the listing
bedrooms	number of bedrooms in the listing
beds	number of beds in the listing
neighborhood	name of the neighborhood
pool	does the listing have a pool or not
d2balboa	this is the distance of the listing from balboa park
coastal	this says whether the listing is near a beach or not
price	nightly prices
log_price	log values of prices
id	airbnb listing id
pg_Aparatment	whether or not the listing is an apartment (land owned by the owner)
pg_Condominium	whether or not the listing is a Condominium (land not owned by the owner)
pg_House	whether or not the listing is a house
pg_Other	any other type of listing except the specified
pg_Townhouse	whether or not the listing is a townhouse
rt_Entire home/apt	does the owner want to rent entire place or not
rt_Private room	does the owner want to rent a private room or not
rt_Shared room	does the owner want to rent a shared room or not
geometry	spatial data of the listing

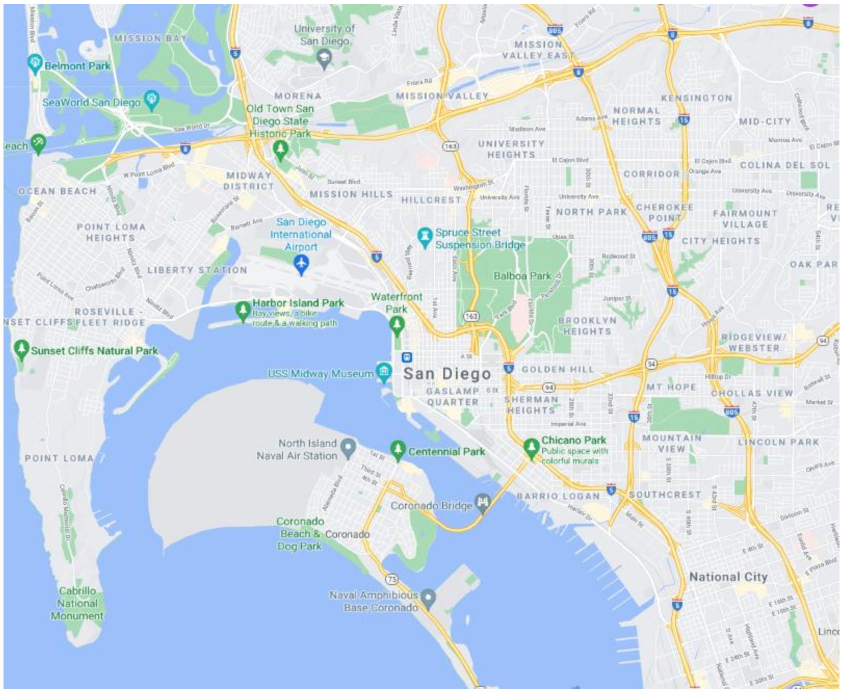
Neighbors:

Feature	Description
Neighborhood	name of the neighborhood
Geometry	spatial information about the neighborhood

Analysis:

Overview:

This dataset highlights San Diego Airbnb rental/lodging rates and contains a mixture of continuous and categorical variables which is evident looking at the data description.



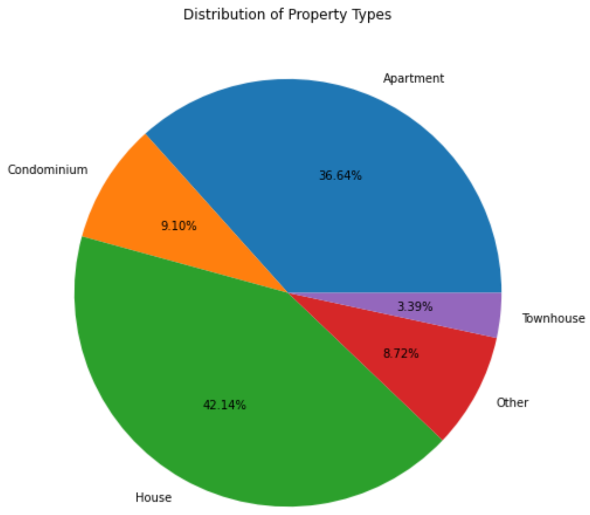
One of the columns in this dataset is proximity to Balboa park. Balboa park, at nearly 1200 acres, is bigger than New York’s Central Park. Besides it’s gigantic footprint, it’s home to our planet’s greatest zoo. This park attracts tourists from across the globe keeping the AirBnBs around it always in high demand. Likewise, there are other geographic features in this dataset that contribute to the nightly prices of the AirBnBs which will be explored in depth in our analysis.

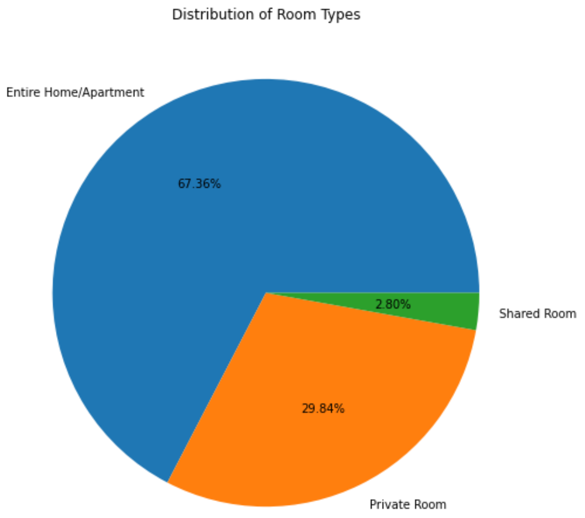
Preprocessing steps:

The quality of the records extracted are clean and didn’t require any preprocessing or conforming. This implies that we didn’t have to combat data imbalance or missing values.

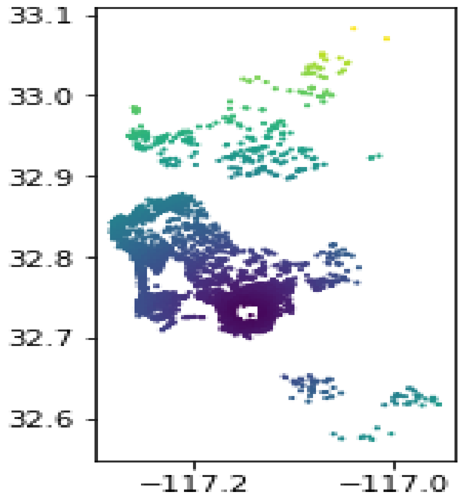
Exploring the data:

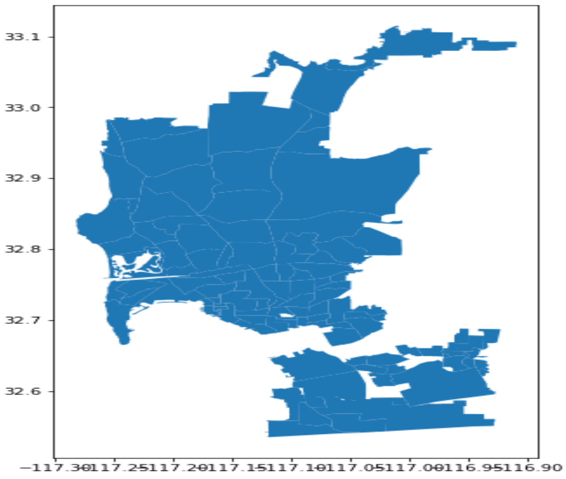
Nearly 40% of the properties are houses followed by 35% for apartments and approximately 70% of rentals are for the entire home or the apartment with nearly 30% for private room rentals. A detailed distribution is shown in the pie chart below:





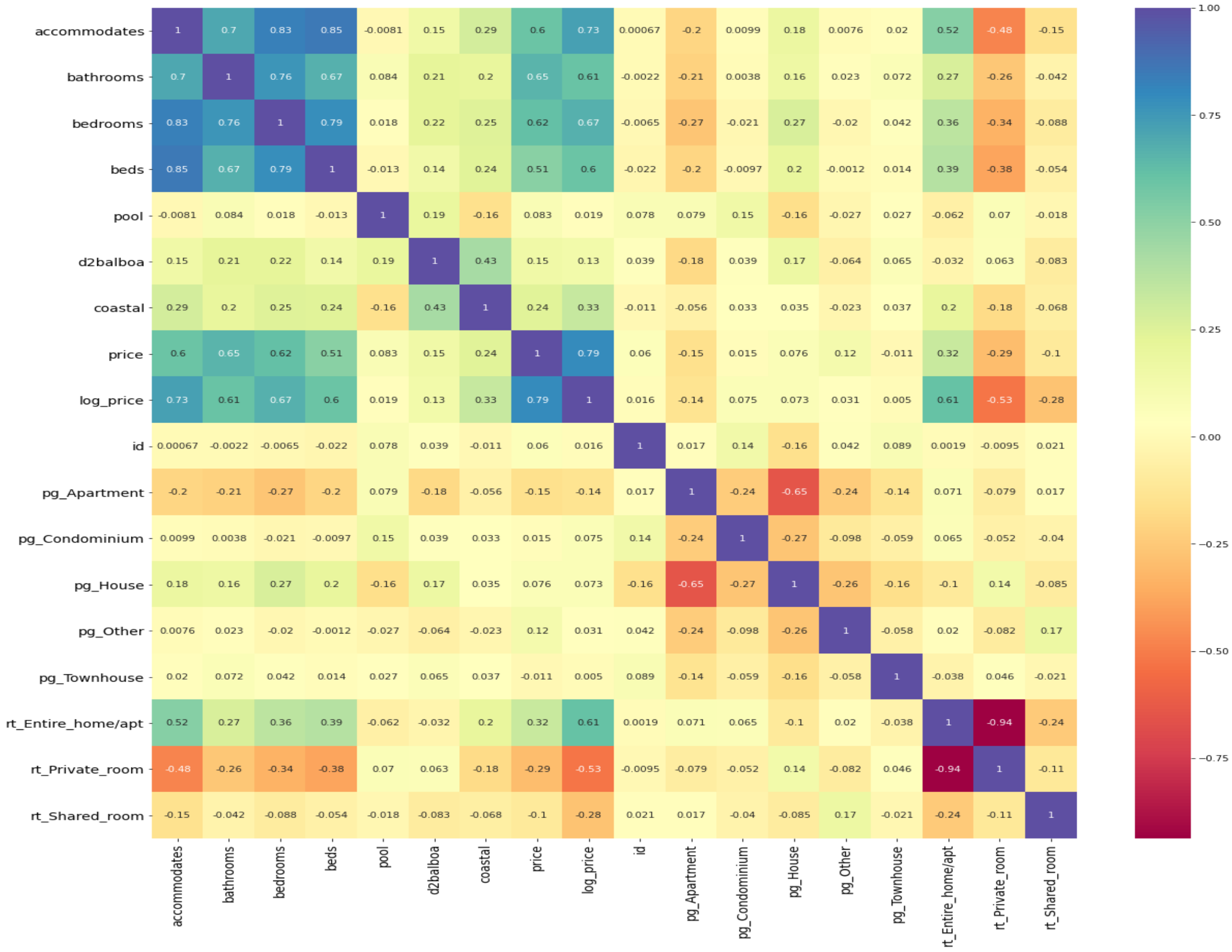
Plotting the geometry of our listings:





Exploring Correlation:

We calculate correlations within our data to determine the variables that we want to include in our regression model. The heatmap showing the same is given below:



For our analysis and regression models, we have selected the following features:

1. Accommodates
2. Bathrooms
3. Bedrooms
4. Beds
5. rt_Private_room
6. rt_Shared_room
7. rt_Condominium
8. pg_House
9. pg_Other

10. pg_Townhouse

The above features are going to be fed to a regression model to perform non-spatial regression first, to be able to predict the nightly log_price of the AirBnBs followed by spatial regression by adding the spatial dimension to the independent variables. The library used for regression analysis is going to be Pysal.

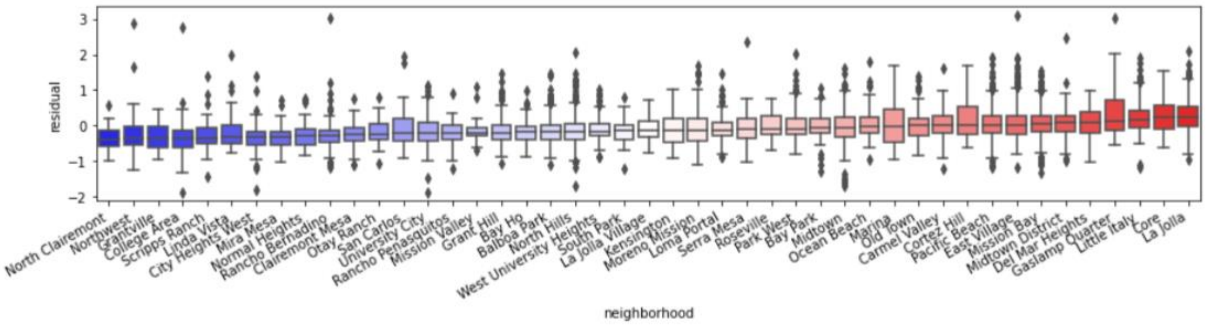
Non-spatial Regression:

After performing non-spatial regression on this data – keeping log_price as the target and the aforementioned features as our predictors, we get an R squared value of 0.67 indicating that our model works well 67% of the time. This is pretty good as we do not have any idea about the latent factors that may have a strong control on the rental prices.

However, an analysis of the errors (or residuals) of each neighborhood divulges more information.

Analysis of Residuals:

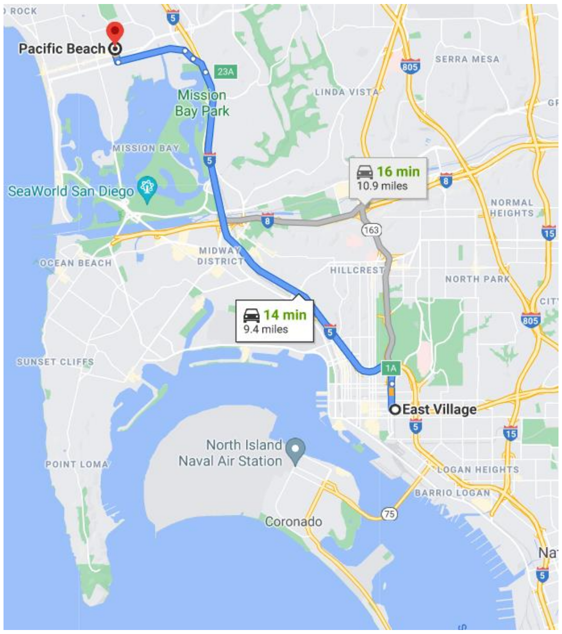
The data and statistics do not give us sufficient information on the desirability of a neighborhood. For instance, living near the beach could be just a personal preference and not a mass opinion or some neighborhoods might have strict rules and restrictions laid around noise levels which might be disliked by youngsters. The list would be never ending but upon further analysis we can find patterns among neighborhoods.



The boxplot of residuals shows us that there are some neighborhoods with extremely low and some with extremely high residuals. And, the ones with lower residuals or the ones with higher residuals tend to be geographically closer / nearby with others in the same categories too. This is an interesting find. Think how some nearby locations might be getting similar kind of predictions even though that might not be correct.

For example, properties in locations near to each other may compete by offering lower or more economical pricing. Of course, this is a latent attribute and is neither captured by an explicitly stated feature in our dataset or by our model. But this can surely be the reason why we might be experiencing clustering of errors in our predictions.

Let’s look at Pacific Beach and East Village. Both of them have a very similar distribution of residuals. Below is the map showing the distance between both the neighborhoods and time to reach from A to B.

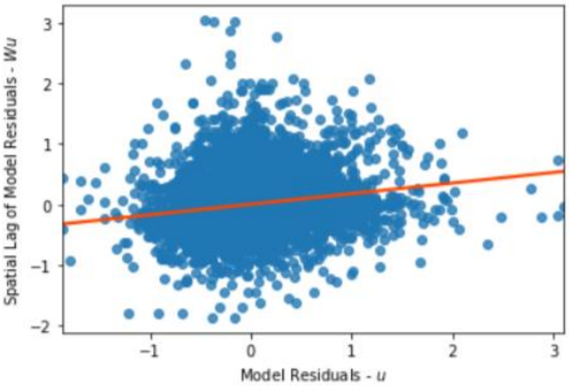


It’s quite evident that these neighborhoods are quite close to one another. This shows the similarity in prediction errors for both the areas indicating spatial spillovers in the nightly prices. Perhaps the hosts of the properties might be adjusting prices looking at one another’s listing to be a more preferred location than the other.

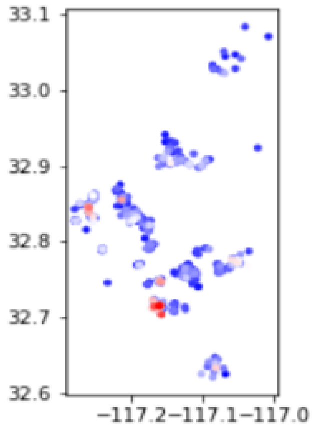
However, since we haven’t utilized geographical or spatial information, our model is unaware of such a phenomena. Due to this we could try to find some clustering in our errors indicating the need for absence of information relevant for predicting the nightly pricing of the listings.

K-Nearest Neighbors to confirm clustering of residuals:

First, we’ll consider a KNN model with only 1 nearest neighbor clocking in the relationship between AirBnB and it’s other closest AirBnB.



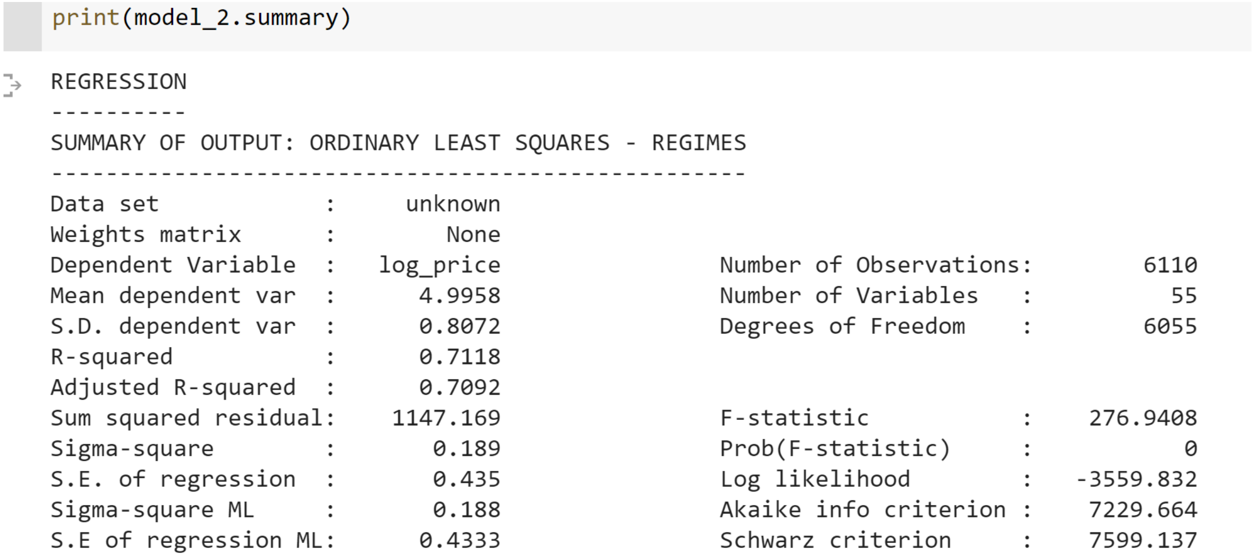
The above plot shows how our errors are clustered. This means that if the price of an AirBnB is overestimated, it’s very likely that the neighboring AirBnB might also be overestimated by the model. However, this is not true if you increase the number of neighborhoods (k in the knn model) which means that this clustering effect seems to die out if proximity of many neighborhoods is concerned. But this does not imply that our residuals are not subject to form clusters. Let’s see what happens if we take k=20.



In this case, we still get smaller clusters that show our model under-predicting prices for listings and their neighboring listings.

Spatial Regression:

We now add a spatial component to our data by indicating the neighborhood of our listing as a predictor variable. The resulting model provides a significant increase in the value of R squared OLS regression. We jump from 67% to 72% which further confirms the value and importance of spatial data.



```
print(model_2.summary)
```

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES - REGIMES
---------------------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  :   log_price                Number of Observations:        6110
Mean dependent var  :      4.9958                Number of Variables   :          55
S.D. dependent var  :      0.8072                Degrees of Freedom    :        6055
R-squared           :      0.7118
Adjusted R-squared  :      0.7092
Sum squared residual:    1147.169                F-statistic           :    276.9408
Sigma-square        :       0.189                Prob(F-statistic)     :           0
S.E. of regression  :       0.435                Log likelihood        :   -3559.832
Sigma-square ML     :       0.188                Akaike info criterion :    7229.664
S.E of regression ML:      0.4333                Schwarz criterion     :    7599.137
```

Conclusion:

After performing the above analysis, it is found that:

- Applying non-spatial regression gives correct results 67% of the time.
- Adding spatial component to the analysis improves the accuracy to 72%

These results indicate that when encountering a geological data, it is better to use spatial regression which can take into account the spatial complexity of the data and help reduce the residual errors that are faced with non-spatial methods.