IS415 Report

CHEN HAO XIAN, TAN WEN YANG, and PIERRE JEAN MICHEL, Singapore Management University, Singapore

A clear and well-documented LaTeX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Additional Key Words and Phrases: HDB Prices, Linear Regression, Geographically Weighted Regression, Exploratory Data Analysis

ACM Reference Format:
Chen Hao Xian, Tan Wen Yang, and Pierre Jean Michel. . IS415 Report. In . ACM, New York, NY, USA, 14 pages.

1 MOTIVATION

With the price of HDB resale flats seeing tremendous growth over the years, and news such as "**HDB resale prices accelerate in Jan as million-dollar deals surge by 42%: SRX, 99.co"** [**?**] or **"HDB resale prices rise 2.3% in Q4, slowest increase in 2022"** [**?**], Singaporeans face the ever-growing concern as to whether or not they are paying a fair price for their flats. Given the lack of accessibility to geographically weighted models, users may find it difficult to assess what factors impact the resale price of an HDB flat; indeed, current estimators often only consider linear relationships between dependent and independent variables and fail to include geographical predictors in their models, which may limit the ability of one regression to explain HDB resale prices; however, geographically weighted regressions provide a more sophisticated way to model spatial heterogeneity by accounting for the unique characteristics of different neighborhoods.

Despite the advantages of geographically weighted regressions, they can be difficult for casual users without specialized skills to use effectively. Thus, our research comes in handy to give the right tools to Singaporeans as we aim to:

1. Identify the most significant location-specific variables that affect the resale price of HDB flats in Singapore and quantify their impact on pricing using geographically weighted regression models. By analyzing the relationship between different amenities, such as rail stations, hawker centers, preschools, malls, and mosquito hotspots, we aim to determine which factors have the most significant explanatory power on HDB resale prices and which do not. It will provide valuable insights into the most important factors that homebuyers and sellers should consider when transacting in the HDB resale market.
2. Develop a user-friendly web application that leverages geographically weighted regression models to estimate the resale value of HDB flats in Singapore for a given area. By inputting location-specific variables such as

proximity to rail stations, hawker centers, preschools, malls, and mosquito hotspots, users can receive an estimated resale value for their property. It will provide users with a more accurate estimation of the value of their property, which can help them make better decisions when selling or buying an HDB flat.

3. Promote transparency and reduce information asymmetry in the HDB resale market by providing an accessible and user-friendly tool for estimating resale values. By making geographically weighted regression models more accessible to the public, we hope to empower homeowners, buyers, and policymakers to make more informed decisions about the HDB resale market. It could ultimately lead to better outcomes for buyers and sellers and help policymakers make more informed decisions about housing affordability, urban planning, and education policy.

2 RELEVANT RELATED WORK

We will now discuss past research in the fields of hedonic housing price models and the importance of geography in these models. We shall review three pieces of work in the following paragraphs.

Our first article [**?**], "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria." by M. Helbich, C. Leitner, and A. Kapusta, investigates the spatial heterogeneity of housing prices in Austria by using a hedonic pricing model. This statistical model estimates the value of a property based on its attributes, size, location, and other factors, such as proximity to different amenities (e.g., shopping malls and parks). In this paper, the writers apply geographically weighted regression (GWR) analysis to identify local spatial patterns of the housing market and examine the spatial variability of the hedonic price model coefficients. In their research, they deduce that the GWR approach provides more accurate predictions of housing prices than traditional regression methods, which assume spatial homogeneity of coefficients. Their study revealed that housing prices are significantly influenced by local characteristics of neighborhoods as accessibility, environment, and social amenities, to name a few. Their paper suggests that the spatial heterogeneity of housing prices is crucial to accurately model and analyze the housing market, especially in geographically diverse regions like Austria. In our case, we consider the paper's methodology and insights valuable for our research as we plan to investigate the spatial heterogeneity of housing prices in Singapore and identify local factors that affect house prices.

The second work we will analyze is the HDB Resale Flat Prices [**?**] application developed by the Housing Development Board of Singapore. It provides a platform for users to access information on past resale transactions of Housing and Development Board (HDB) flats in Singapore. It allows users to search for resale transactions based on various criteria, such as location, flat type, and transaction period. This web application is the perfect example of what Singaporeans are missing, a model that explains the relevancy of locational factors on HDB resale price. Nonetheless, this web application still serves as a relevant solution that offers insights into the Singapore housing market and the attributes of HDB flats that may affect transaction prices. We believe there is room for improvement in the functionalities the platform offers. Through our research, we aim to explore ways to improve the current system, perhaps by incorporating additional data sources and applying advanced machine learning algorithms to provide an accurate and valuable explanatory model of housing prices.

The last piece of reading we shall discuss is the Medium article [**?**] "Predict the Selling Price of HDB Resale Flats" by Jim Meng Kok. It uses machine learning to predict the selling price of HDB resale flats in Singapore. The author uses a dataset of HDB resale transactions and applies a random forest regression model to predict the resale price of HDB flats based on various features, including location, size, age, and nearby amenities. The model achieves a high accuracy rate, indicating that machine learning methods can effectively predict housing prices. The article is relevant

to the current research on housing prices in Singapore as it highlights the importance of utilizing machine learning techniques for predicting prices accurately. While the author uses Python in his analysis, we still consider that the insights provided in the article can be valuable for any researcher planning to use other programming languages, as R. The findings suggest that location and amenities play a significant role in determining housing prices, which is consistent with previous research on the topic.

In conclusion, this section has reviewed three pieces of research related to hedonic housing price models and the importance of geography in these models. Overall, these studies emphasize the need to account for location and amenities when modeling and analyzing housing prices and highlight the potential of advanced techniques like geographically weighted regression and machine learning to provide more accurate predictions.

3 DESIGN FRAMEWORK

The design of our Shiny App tries to follow the data visualization general guidelines suggested by Schneiderman's mantra [**?**]: Overview First, Zoom and Filter, Details on Demand. It consists of Eight Different Tabs: About Us, EDA, CorrPlot, Multiple Linear Regression, Geographically Weighted Regression, Predicting, Interactive Map and Upload File [Figure 1] Figure 1. This Navigation Bar is featured throughout all the Shiny App at all time. Each of the Features are place in the order where the user is expected to navigate in. Each of the Tab Features is own unique set of features, however there are 3 major views, map, graphic and tables.

About Us Tab Figure 1 is an overview of application and features no interactivity.

The EDA Tab Figure 2 features an overview of the entire data set that is loaded in the Shiny Model. A Histogram is given to give an overview of selected variable of the data set. Key Summaries is also provide important information that the user might need, such as the summary of the data and the Top 10 towns. This information is given in a form of a table. Finally, the entire data is given in the form of the table for the user to view the entire data. The Shiny Application provide a sidebar, where all the drop downs form are located. We namely allow the user to modify the Start Date, End Date and the Room Type. For the Histogram, the user is able to select the number of bins, the color of the Fill and the data to be displayed.

The Corr Plot Tab Figure 3 features a Graphic Plot of Correlation Plot of the variables. The Correlation Plot is calculated based on the columns of the data set.

The Multiple Linear Regression Tab Figure 4 features graphics of the Linear Regression Model. It also features all the various test to determine the validity of the Linear Regression Model, such as Linearity and Moran I for the Spatial Model. All the outputs are featured as a graphic with explanation on what to look out for. The Side Bar Features a multi select drop down of all the dependent variables of the data set to build the Linear Regression Model. Changing the values in the Multi Select will update all the graphics with the exception of Moran I as it is computationally intensive. As weight matrix is needed to calculate Moran I, as many variables of dnearneight() and nb2listw() function is exposed [**?**].

The Geographically Weighted Regression Tab Figure 5 features graphics of the Geographically Weighted Regression and the explanation needed. The current graphics of the GWR Model is not computed re-actively as the the computation of the bandwidth and the model is computationally intensive. The side bar features as many variables of bw.gwr() and gwr.basic() function is exposed [**?**]. Two buttons named 'Generate GWR' and 'Use Computed BandWidth' can be used to render the graphic re-actively but it will take a long time to load.

The Predict Tab Figure 6 features a simple graphic of the predict values. The side bar provides the ability to toggle between the GWR Model or the Linear Regression Model. The GWR Model requires a Spatial Model to be uploaded, while the Regression Model requires input on a text box, separated by comma.

The Interactive Map Tab Figure 7 features an interactive Map of all the Points in the data set. The user can filter out the Start Date, End Date and Room Type of the data set to display. Any columns in the data set can be displayed on the interactive map and some options of the well know colorbrewer [**?**] is given for the user to select. The user can select the classification methods of the points from a drop down, based on attributes available from tmap package [**?**]. The tab is place towards the end because the interactive map is not the main objective of this project but provided the user to gain a better understanding of the data.

The Upload File Tab Figure 8 features a Table that is only displayed when the user upload a file. The user can upload a file from the sidebar. This tab is placed at the end, as the user should gain a better understanding of the ability of our application attempting to load their own data.

A combination of these three views along with the ability to build their own model based on their own judgement should give the user a great control over the data and more power to analyze and explore.

4 CASE STUDY OF HDB RESALE FLATS

Singapore, an emerging economy with a thriving real estate market, boasts a unique housing landscape that has drawn attention worldwide. Indeed, Singapore is one of the few city-states in the world, making it a compact and densely populated urban environment. This unique characteristic could already be the foundation of a fascinating use case for our web app. Yet, what makes Singapore housing so special is the government's active role in shaping the market, which has led to the development of the Housing and Development Board (HDB) flats – an essential feature of the Singaporean housing landscape.

Since the 1960s, the Singaporean government has heavily involved itself in public housing, and its policies have gained widespread recognition for their success. The country's housing policies have been instrumental in promoting homeownership and providing affordable housing for its citizens. These flats, built and managed by the government, house over 80% of the population [**?**], making Singapore's public housing program one of the most successful in the world. The HDB flats, however, are subject to unique rules and regulations that distinguish them from other housing solutions (e.g., private housing), such as restrictions on resale and renovation. Thus, understanding the factors that drive the prices of HDB resale flats is of great interest. In this context, geographically weighted regression analysis is an ideal method for examining the spatial heterogeneity of HDB resale prices and identifying the local factors that influence them. By considering the specific characteristics of each HDB estate and the surrounding neighborhoods, we can better understand how various factors, such as distance to amenities, transport links, and schools, affect the prices of HDB resale flats. With our web app, policymakers, investors, and residents can use these insights to make more informed decisions about the HDB market.

To analyze the spatial heterogeneity of HDB prices in Singapore, we will utilize a variety of variables that capture different aspects of the flats and their surrounding areas. These variables include AREA_SQM, LEASE_YRS, STOREY_ORDER, and proximity to various amenities such as CBD, childcare, eldercare, hawker centers, MRT stations, parks, top primary schools, malls, supermarkets, clinics, pharmacies, tourist attractions, and libraries. We have sourced this data from websites such as data.gov.sg, Wikipedia, LTA Data Mall, and the MOE website, and additionally used the OneMap API to retrieve and geocode data. Some data preprocessing was necessary as geocoding aspatial fields to convert addresses

into geographic coordinates. Using geographically weighted regression analysis on this rich dataset, we can identify local factors that influence the prices of HDB resale flats in different regions of Singapore.

For the sake of limiting the computing power required to run our analysis, we shall only focus on HDB resale flat transactions from 2021 to 2022. This can be done at the EDA Tab, where the user will filter out the data into the correct time frame. It would appears that the PRICE variable is left-skewed base on the Histogram. Altering the display variable to LEASE_YRS will give a different picture, where the data is right-skewed. Key Summaries of the variable such as the median and mean will be displayed. The mean of the variable based on towns will be displayed as well. From the PRICE variable, the mean is found to be 526124 and the Central Area has the highest mean PRICE.

The Multiple Linear Regression (MLR) model reveals that most variables have a statistically significant effect on housing prices in Singapore [Figure 9]. The intercept is estimated to be 1.607e+05, indicating that a house with all predictor variables set to zero should cost at least **S$160,700**. The variables that positively impact housing prices are AREA_SQM, LEASE_YRS, STOREY_ORDER, PROX_PARK, NUM_KNDRGTN, NUM_ISP_CLIN, and NUM_REGISTERED_PHARM, while the variables with a negative impact are PROX_CBD, PROX_HAWKER, PROX_MRT, PROX_MALL, PROX_SPRMKT, PROX_PHARMACY, and PROX_TOURISM. The variable PROX_CLINIC has a p-value of 0.47737 and is not statistically significant at a 5% error level. Meanwhile AGE and PROX_TOPPRISCH since to return a null result. The R-squared value of 0.7381 indicates that the model's predictor variables can explain 73.81% of the variance in housing prices. The F-statistic of 3171 with a p-value of < 2.2e-16 suggests that the model's overall fit is significant.

We must however understand that Linear Regression Model has a number of key assumptions namely: *Linearity, Normality, Homoscedasticity and Independence* [**?**]. We have conducted all of this test and have determine that our model satisfy this criteria.

Independence Test is already performed at in the Corr Plot Tab Figure 3. All variables that are correlated should be remove already. We have kept all the variables in our default model for demonstration purposes however.

Linearity Test Figure 10 performed have shown that LEASE_YRS, AGE, PROX_MRT and PROX_TOPPRISCH have a VIF more than 10

Homoscedasticity Test Figure 11 performed have shown that most of the residual is scattered around the line, however there are points that deviate significantly from the line.

Normality Test Figure 12 performed have shown that most of the residual resembles a Normal Distribution Curve.

Furthermore, in this study, we conducted a global Moran's I test to assess the presence of spatial autocorrelation in housing prices in Singapore. Our results showed a Moran I statistic standard deviate of 1122.3 and a p-value smaller than 2.2e-16 Figure 13, indicating strong evidence of spatial autocorrelation in resale HDB flats. Specifically, our observed Moran I was 0.217. It is significantly higher than the expected value under spatial randomness. It demonstrates the significance of spatial autocorrelation in housing prices in Singapore, and we should consider it for several reasons.

First, it suggests that housing prices in some regions of the city-state are not independent and that the prices of nearby properties may influence resale value. It means that traditional regression techniques, which assume independence between observations, may need to be revised for modeling housing prices in Singapore.

Moreover, spatial autocorrelation can reveal the underlying spatial structure of housing prices in Singapore, including spatial patterns and trends. Understanding these patterns can provide valuable insights into the factors that drive housing prices in different parts of the country. Empirical evidence also supports the relevance of spatial autocorrelation in housing prices. For example, a study by Wong and Leung [**?**] found significant spatial autocorrelation in housing prices in Hong Kong and suggested that spatial models, such as geographically weighted regression, could improve the accuracy of price predictions.

While the MLR model provides valuable insights into the relationship between the predictor variables and housing prices, it does not consider the spatial variation in the data. Therefore, it may need to capture the heterogeneity observed in different regions of Singapore fully. To address this issue, we will employ a geographically weighted regression (GWR) model, allowing us to examine how the relationship between the predictor variables and housing prices varies spatially across different regions of Singapore.

A GWR Model is build using the exact same data set which result in a R-Squared value of -22174.95 Figure 14. This is likely due to the fact that the model used was not fine tuned at all. When the same data set is used by Megan and Aisyah, such model produced a R-squared value of 0.9 and greater [**?**] [**?**]. This shows that the model has the potential to be significantly better than the Regression Model when proper tuning is performed.

This case study have shown that with the available data, the user is able to gain insights on the data through the EDA. The user is able to make use of the interactive nature of the Shiny Application to focus on a specific variable of interest and make a judgement on whether the variable should be included in the model. The user can reference the CorrPlot to gain insights of the relationship between the variable to modify the variables used in the Linear Regression Model, which the user can see the changes in real time. Test for the Model will also be updated in real time. If the user chose, a model based on the selected variable can be built in real time as well. This will definitely encourage the user to explore the data and fine tune the Model for more in depth exploration and Analysis.

5 CONCLUSION

In conclusion, we have demonstrated the ability to use Shiny in development an application to create an explanatory model for HDB Resale Price. More importantly, we have shown that building a Geographically Weighted Linear Regression Model for HDB Resale Prices is possible with proper fine tuning.

6 POTENTIAL FUTURE WORK

Our Model can be expanded in number of key areas as the potential for the application to grow is large.

Firstly, we can expand the scope of the project to encompass the Private Properties and the introduction of new geographical variable as well. The addition of Private Properties prices will give a more complete view of the housing situation in Singapore. The data for private properties are not publicly available and the inclusion can result in a much better models. The addition of geographical data as well can also be critical as geographical data of interesting variables such as Hospital, Tourist Attraction has not be included in the loaded model. The inclusion of such geographical data into the data set can give more insights to the user as it may review surprising relationship.

Lastly, we can expand the number of models that the project uses as well. In this project only Linear Regression Model and Geographically Weighted Regression is used. Models such as Geographically Weighted Regression Trees can be introduced for the user to further compare and contrast. Some of these possible enhancements could be taken up in the near future.

7 APPENDICES AND FIGURES

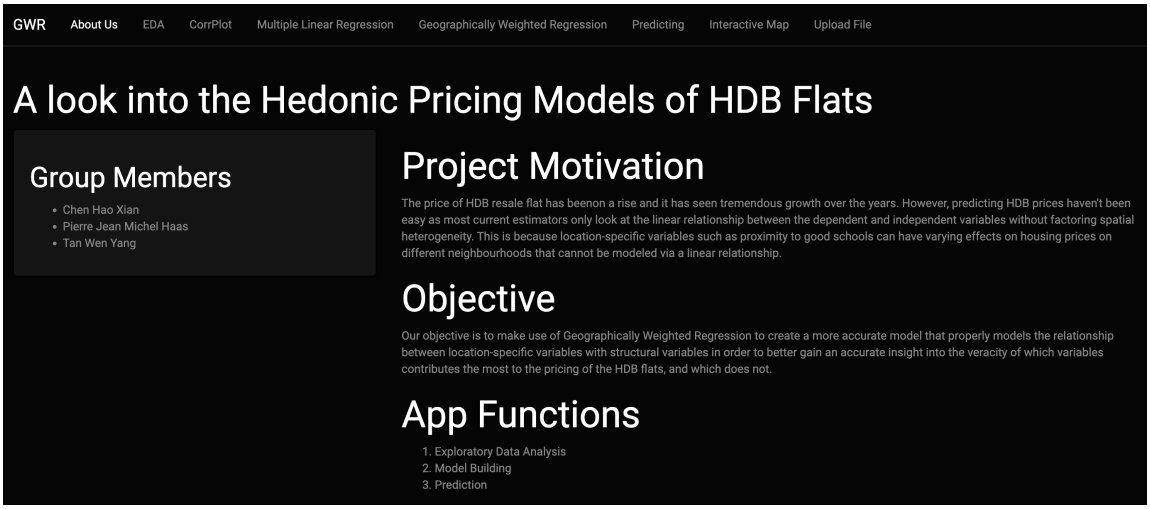

Fig. 1. About Us Tab

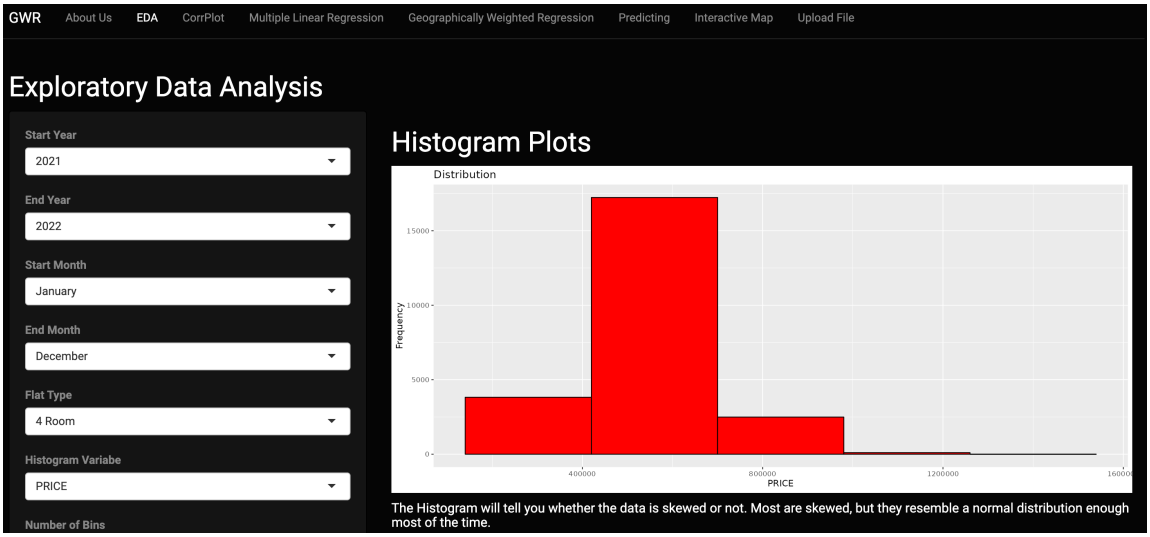

Fig. 2. EDA Tab

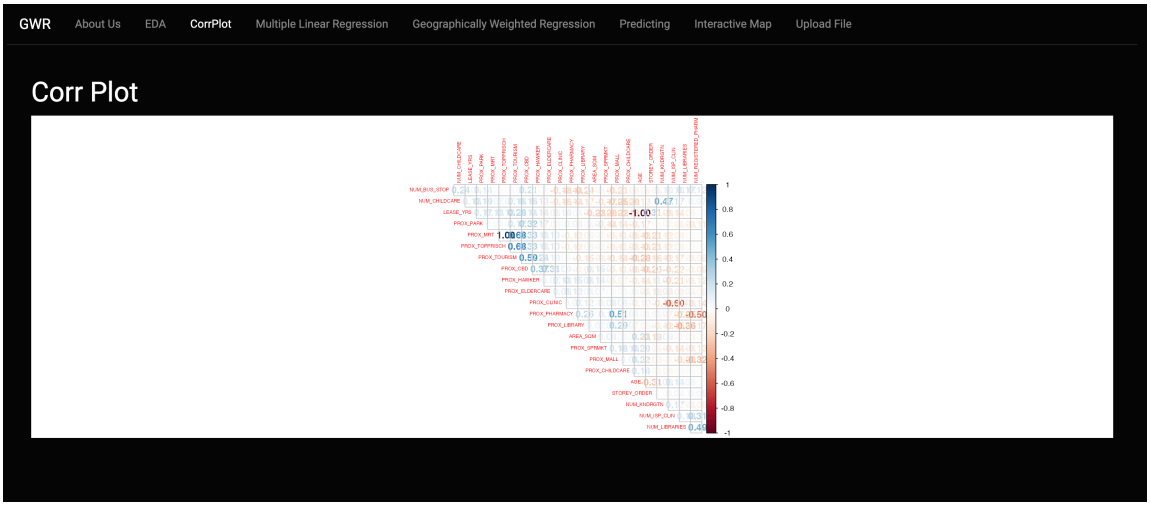

Fig. 3. CorrPlot Tab

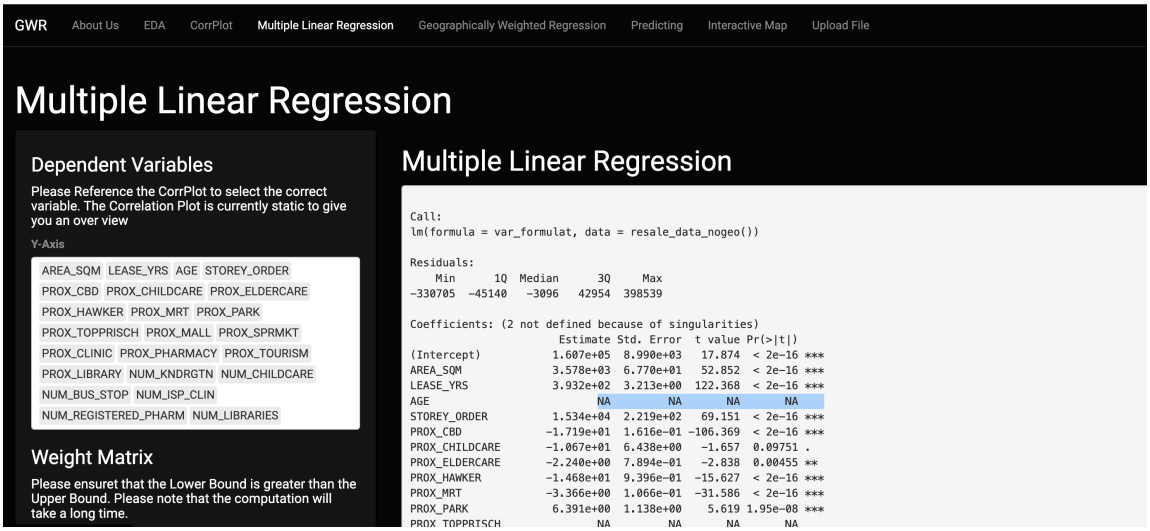

Fig. 4. MLR Tab

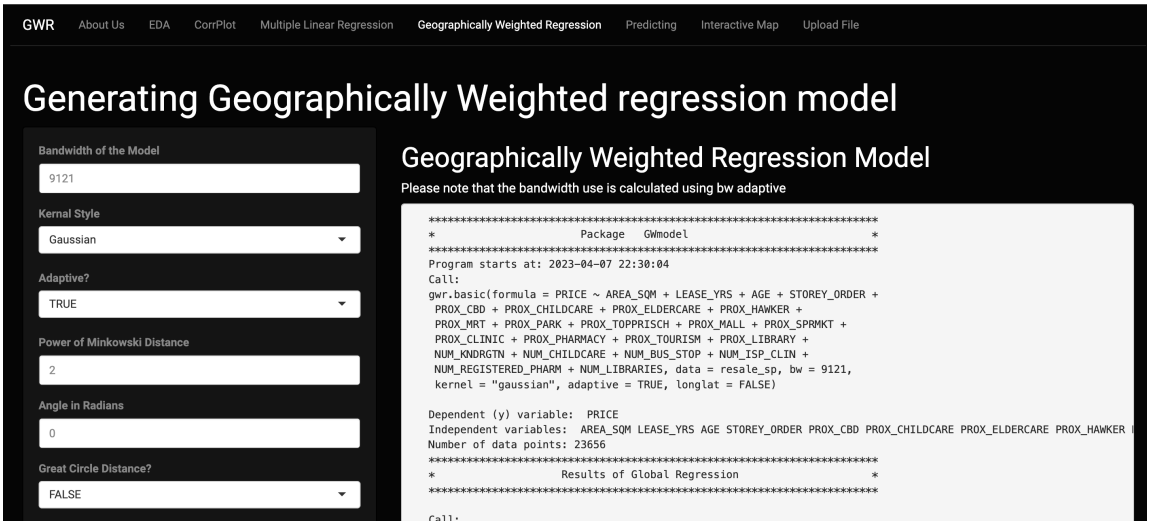

Fig. 5. GWR Tab

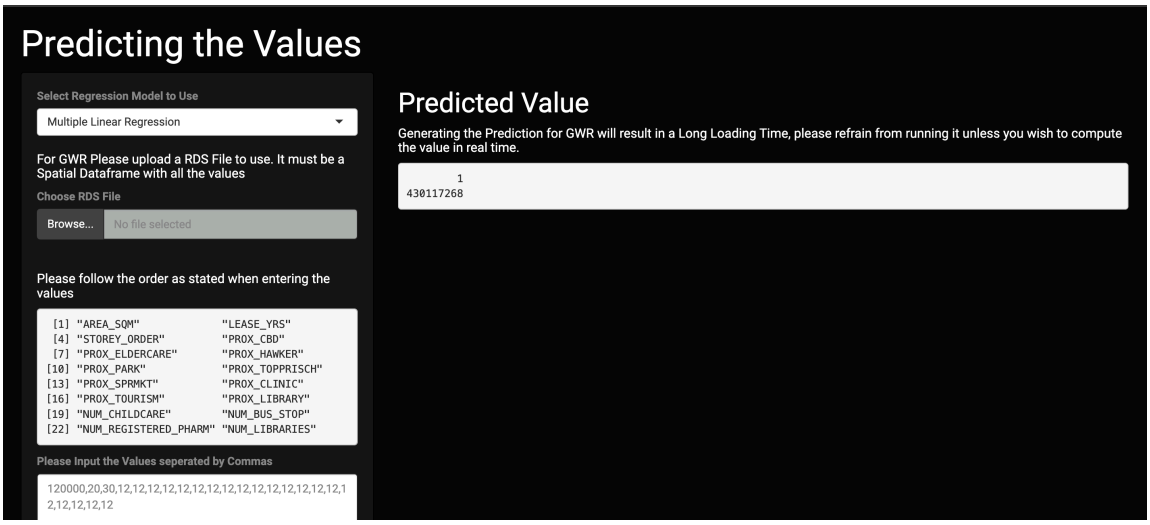

Fig. 6. Predicting Tab

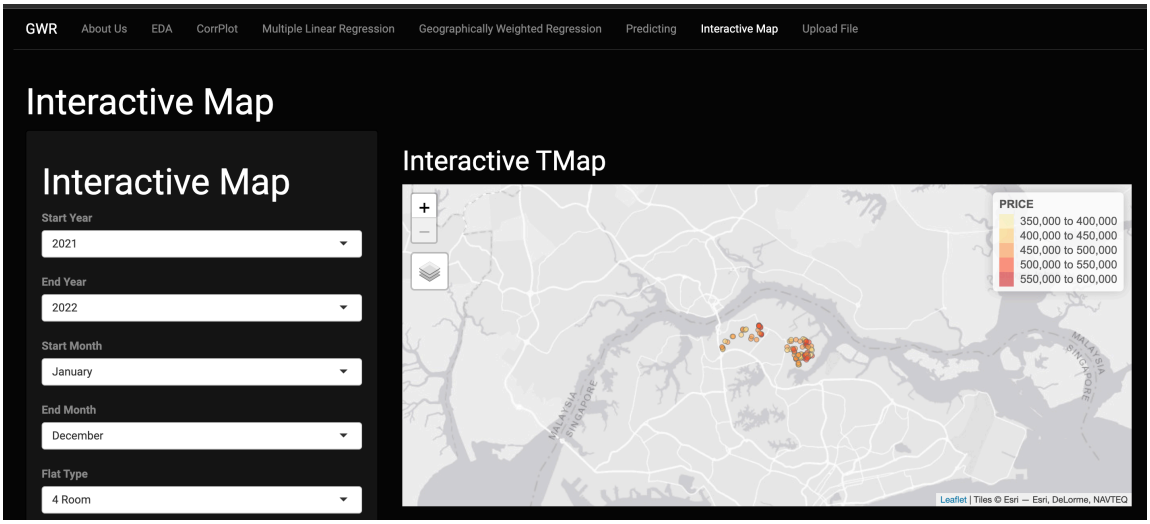

Fig. 7. Interactive Map

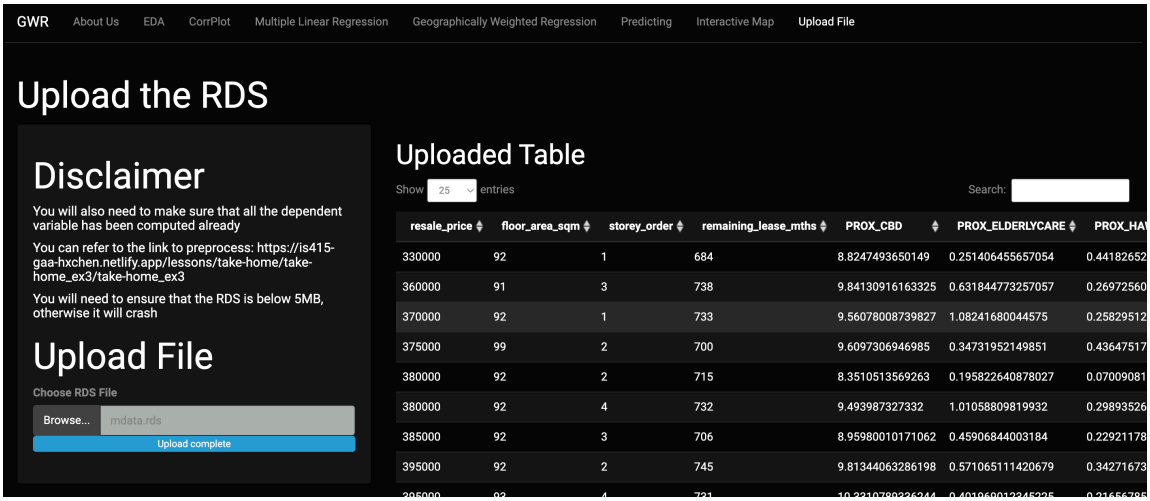

Fig. 8. Upload Tab

```
Coefficients: (2 not defined because of singularities)
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)           1.607e+05  8.990e+03   17.874  < 2e-16 ***
AREA_SQM              3.578e+03  6.770e+01   52.852  < 2e-16 ***
LEASE_YRS             3.932e+02  3.213e+00  122.368  < 2e-16 ***
AGE                         NA         NA       NA       NA
STOREY_ORDER          1.534e+04  2.219e+02   69.151  < 2e-16 ***
PROX_CBD             -1.719e+01  1.616e-01 -106.369  < 2e-16 ***
PROX_CHILDCARE       -1.067e+01  6.438e+00   -1.657  0.09751 .
PROX_ELDERCARE       -2.240e+00  7.894e-01   -2.838  0.00455 **
PROX_HAWKER          -1.468e+01  9.396e-01  -15.627  < 2e-16 ***
PROX_MRT             -3.366e+00  1.066e-01  -31.586  < 2e-16 ***
PROX_PARK             6.391e+00  1.138e+00    5.619 1.95e-08 ***
PROX_TOPPRISCH              NA         NA       NA       NA
PROX_MALL             7.025e+00  1.456e+00    4.825 1.41e-06 ***
PROX_SPRMKT           2.017e+01  3.074e+00    6.560 5.48e-11 ***
PROX_CLINIC           1.953e+00  2.748e+00    0.711  0.47737
PROX_PHARMACY        -1.746e+01  1.733e+00  -10.078  < 2e-16 ***
PROX_TOURISM          3.565e+00  4.798e-01    7.430 1.12e-13 ***
PROX_LIBRARY         -1.802e+01  9.674e-01  -18.633  < 2e-16 ***
NUM_KNDRGTN           7.922e+03  5.113e+02   15.495  < 2e-16 ***
NUM_CHILDCARE        -3.822e+03  2.415e+02  -15.823  < 2e-16 ***
NUM_BUS_STOP         -3.142e+02  1.661e+02   -1.891  0.05864 .
NUM_ISP_CLIN          5.953e+03  5.079e+02   11.721  < 2e-16 ***
NUM_REGISTERED_PHARM  5.849e+03  6.486e+02    9.018  < 2e-16 ***
NUM_LIBRARIES         1.886e+04  2.597e+03    7.262 3.92e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66310 on 23634 degrees of freedom
Multiple R-squared:  0.7381,    Adjusted R-squared:  0.7378
F-statistic:  3171 on 21 and 23634 DF,  p-value: < 2.2e-16
```

Fig. 9. Linear Regression Score

	Variables	Tolerance	VIF
1	AREA_SQM	8.600190e-01	1.162765e+00
2	LEASE_YRS	0.000000e+00	Inf
3	AGE	0.000000e+00	Inf
4	STOREY_ORDER	7.761687e-01	1.288380e+00
5	PROX_CBD	4.038875e-01	2.475937e+00
6	PROX_CHILDCARE	9.221182e-01	1.084460e+00
7	PROX_ELDERCARE	7.905480e-01	1.264945e+00
8	PROX_HAWKER	7.646479e-01	1.307791e+00
9	PROX_MRT	1.030287e-13	9.706034e+12
10	PROX_PARK	7.769189e-01	1.287136e+00
11	PROX_TOPPRISCH	1.030287e-13	9.706034e+12
12	PROX_MALL	6.075860e-01	1.645858e+00
13	PROX_SPRMKT	8.468529e-01	1.180843e+00
14	PROX_CLINIC	6.106914e-01	1.637488e+00
15	PROX_PHARMACY	5.225084e-01	1.913845e+00
16	PROX_TOURISM	3.050333e-01	3.278331e+00
17	PROX_LIBRARY	6.989255e-01	1.430768e+00
18	NUM_KNDRGTN	6.676401e-01	1.497813e+00
19	NUM_CHILDCARE	6.011454e-01	1.663491e+00
20	NUM_BUS_STOP	8.055685e-01	1.241359e+00
21	NUM_ISP_CLIN	5.279619e-01	1.894076e+00
22	NUM_REGISTERED_PHARM	5.355527e-01	1.867230e+00
23	NUM_LIBRARIES	6.565340e-01	1.523150e+00

Fig. 10. Linearity Test

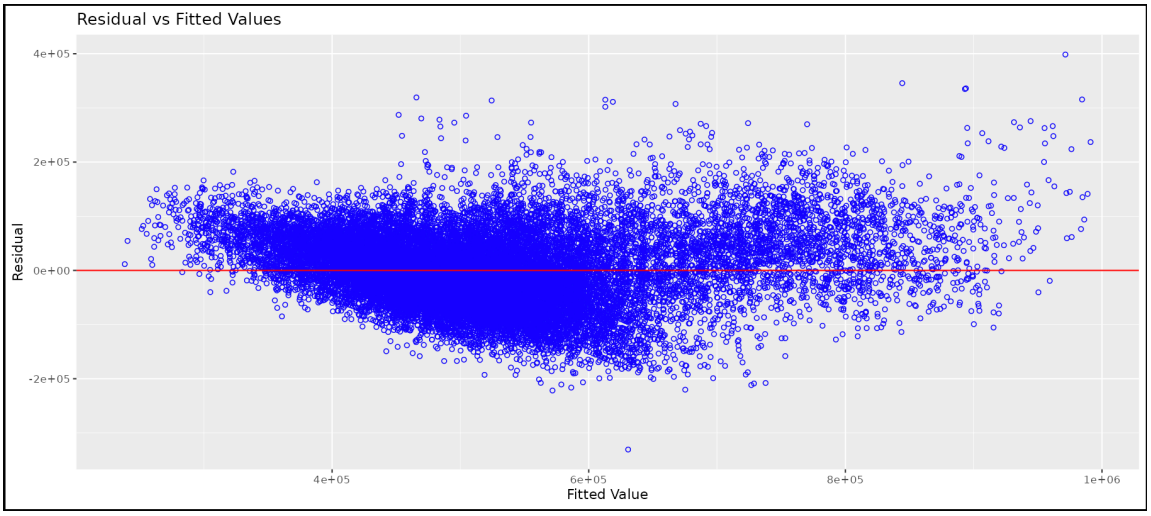

Fig. 11. Homoscedasticity (Non Linearity)

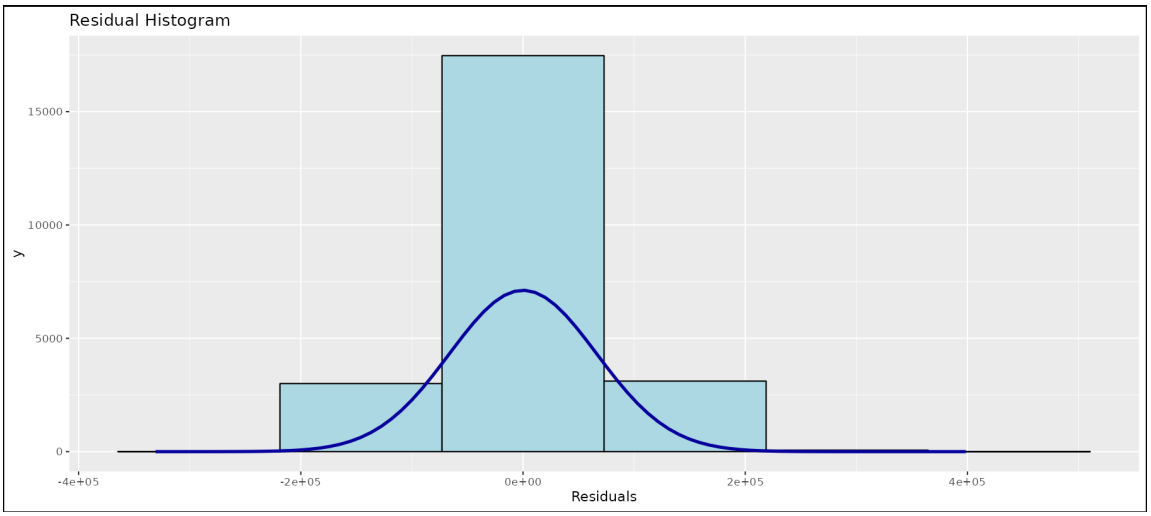

Fig. 12. Normal Distribution of LM Model

```
        Global Moran I for regression residuals

data:
model: lm(formula = PRICE ~ AREA_SQM + LEASE_YRS + AGE + STOREY_ORDER + PROX_CBD +
PROX_CHILDCARE + PROX_ELDERCARE + PROX_HAWKER + PROX_MRT + PROX_PARK + PROX_TOPPRISCH +
PROX_MALL + PROX_SPRMKT + PROX_CLINIC + PROX_PHARMACY + PROX_TOURISM + PROX_LIBRARY +
NUM_KNDRGTN + NUM_CHILDCARE + NUM_BUS_STOP + NUM_ISP_CLIN + NUM_REGISTERED_PHARM +
NUM_LIBRARIES, data = resale_flat_full_nogeo)
weights: nb_lw

Moran I statistic standard deviate = 1122.3, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
    2.168256e-01    -2.391462e-04     3.740997e-08
```

Fig. 13. Moran I Test

```
************************Diagnostic information**************************
Number of data points: 23656
Effective number of parameters (2trace(S) - trace(S'S)): -2172473
Effective degrees of freedom (n-2trace(S) + trace(S'S)): 2196129
AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 860571.4
AIC (GWR book, Fotheringham, et al. 2002,GWR p. 96, eq. 4.22): 860676
BIC (GWR book, Fotheringham, et al. 2002,GWR p. 61, eq. 2.34): 836043.9
Residual sum of squares: 8.797804e+18
R-square value:  -22174.95
Adjusted R-square value:  -237.8623
```

Fig. 14. GWModel