**Understanding the Spatial Patterns of House Prices in King County, WA**

**Map the prices in King County, WA. Provide a screenshot of the map and qualitatively describe any patterns you observe in data.**

1. **Seek and summarize spatial patterns in house prices in King County, WA. Use a local spatial summary metric and explain your results. Are there spatial patterns identifiable in house prices.**

I chose the tool ‘Cluster and Outlier Analysis (Anselin Moran’s I)’ to identify potential patterns of house prices. The ‘price’ is the input field, and I used the KNN to quantify the spatial relationship between each data point. The closest k features are included in the analysis. The number of neighbors (k) is specified by the Number of Neighbors parameter. The number of neighbors is the default value of 8. The number of permutations is set to the biggest value of 9999. The higher it is, the more significant the statistics.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 1 parameters of the tool

According to Figure 2, on this map, blue points are house prices which are classified as "Low-low cluster". "Low-low cluster” refers to a clustering phenomenon where an area and its surrounding areas all have relatively low attribute values. Pink points are ‘’High- high cluster’ which indicates high house prices in surrounding areas. Grey points refer to areas where spatial autocorrelation is not quite significant. Red and blue points are outliers. Now we can conclude that the areas surrounding Bellevue are characterized by high housing prices. Apart from Bellevue, the areas from downtown Seattle to Renton, Burien, and Kent are considered low-cost housing areas. Occasionally some places have a higher or lower house price than its neighbors are outliers.

**地图

描述已自动生成**

**Figure 2 The output of running local Moran’s I**

1. **Display relationships between other variables and house price.**

I chose ‘Scatter plot matrix’ to display relationships between and house price. Multivariate analysis allows for the simultaneous examination of relationships between multiple variables, rather than analyzing each variable pair separately. In this graph, not only the relationships between price and other variables are displayed, but the relationship among other variables is shown as well (Figure 3).

散点图

中度可信度描述已自动生成

Figure 3 Scatter Plot Matrix of house prices and other variables

1. **Describe the relationships between variables that have relatively high linear relationships with house price.**

The relationship between sqft\_living and price is relatively high. It’s more likely that they are positively correlated that the more sqft\_living, the more price (Figure 4). And sqft\_living15 and bathrooms both have a positive R2.

图形用户界面, 图表, 应用程序

描述已自动生成

Figure 4 Scatter Plot Matrix

图表, 散点图

描述已自动生成

Figure 5 sqft\_living and price

Also, the sqft\_above and price has the same linear relationship (Figure 5).

1. **From your previous analysis (1a) select a portion of the data that corresponds to a cluster of high prices. Copy the feature class to a new feature class and display relationships between different variables. Do you observe considerable changes to R2 compared to 1c? Explain.**

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 6 select data by attributes ‘COType KNN’

I use ‘Select By Attributes’ in the attributes table. After selecting data that are in the ‘High – Hight cluster’ and exporting them to a new feature class (Figure 7), now the the data that corresponds to a cluster of high prices are shown as below (Figure 8)

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure 7 Export selected features into a new feature class

地图

描述已自动生成

Figure 8 High price of houses near Seattle.

Then, I used Spatial join to join other attributes from the original dataset into this new feature class. Because running Moran’s I analysis, newly created attributes of the output don’t contain these attributes such as sqft\_living (Figure 8).

图形用户界面, 应用程序

描述已自动生成  
Figure 8 Spatial join features based on ‘Price’

After running Spatial Join, these generated new attributes table is displayed below (Figure 9). This attributes table contains not only the outcome of KNN but also attributes associated with house prices.

图形用户界面, 表格

描述已自动生成

Figure 9 new table contains outcomes of KNN and original attributes

The layout of this graph is according to the R2 (Figure 10). 图形用户界面, 图表, 散点图

描述已自动生成

Figure 10 Scatter plot matrix of selected data in ‘High – High’ cluster

Compared to the previous scatter plot matrix, we can observe the decrease of R2 of each regression of a pair of variables.

When analyzing high-priced clusters, the chosen data range becomes more limited. This suggests that while these variables may have strong predictive power for housing prices in the overall dataset, their explanatory power for price variability may decrease within the more homogeneous high-priced clusters. Increased homogeneity leads to reduced explanatory power of model predictors for prices in this specific subset, potentially resulting in a decrease in the R² value.

**Fitting A Linear Regression Model to Your Data**

1. **Firstly, fit a Generalized Linear Regression (GLR) model to describe house price using some of the predictors you delineated from previous section.**

Firstly, I used ‘Delete Identical’ to delete the same rows to prepare for the GLR model. Then I used the tool ‘Generalized Linear Regressions’ to run a regression between the dependent variable price and independent variables bathrooms, sqft\_living, sqft\_above and grade (Figure 11).

地图

描述已自动生成

Figure 11 Outcomes of GLR

1. **Describe the regression diagnostics that the tool returns.**

The regression analysis results include model coefficients (Coef), standard errors (StdError), t-statistics (t\_Stat), and probability values (Prob) to estimate the relationship between house price and several predictors. Based on the provided results, we can conclude that:

1. ‘bathrooms’:

- Coefficient (Coef): -10496.651881, which suggests that with each additional bathroom, the house price is expected to decrease by approximately $10,496.65.

- t-Statistic (t\_Stat): -2.16736, which indicates the magnitude of the coefficient relative to its standard error.

- Probability value (Prob): 0.030227, which is the two-tailed p-value corresponding to the t-statistic, indicating that the variable's coefficient is statistically significant (P-value < 0.05).

2. ‘sqft\_living’ (living area):

- Coefficient: 238.870918, meaning for every additional unit of living area (usually square feet), the house price is expected to increase by about $238.87.

- t-Statistic: 39.577474, a very high t-value, indicating a strong positive impact of living area on house price.

- Probability value: 0, a p-value close to zero indicates a very significant impact of this variable.

3. ‘grade’ (housing grade):

- Coefficient: 123811.357434, for each increase in housing grade, the house price is expected to increase by about $123,811.35.

- t-Statistic: 37.836124, indicating that housing grade is a strong predictor.

- Probability value: 0, also indicating a very significant positive impact of housing grade on house price.

4. ‘sqft\_above’ (above ground area):

- Coefficient: -51.626553, for every additional unit of above ground area, the house price is expected to decrease by about $51.63.

- t-Statistic: -8.37598, indicating that this variable is statistically significant but has a negative impact on house price.

- Probability value: close to zero, indicating that this effect is statistically significant.

5. ‘Intercept’:

- Coefficient: -790517.695748, which is the estimated value of the house price when all predictors are zero.

- t-Statistic: -40.587855, indicating that the intercept is very significant.

- Probability value: close to zero, indicating that the intercept is statistically significant.

1. **Is GLR a good model of the house price? Why or why not?**

**图表

描述已自动生成**

Figure 12 Scatter plot matrix of GLR with Person-R on the upper left

I think it performs OK, because it tells us the relationship between price, grade, and sqft\_living. The R2 between price and these variables is relatively high (Figure 12). However, it cannot interpret the negative relationship between the number of bathrooms and price. The negative relationship is very counterintuitive. Above all, it is not a good model.

1. **Define a Geographically Weighted Regression (GWR) model for house price. Describe the regression diagnostics for GWR. Is GWR a better regression model? Explain using regression diagnostics.**

I choose sqft\_living and grade as independent variables and price as dependent variable. The outcome of running a GWR is shown below (Figure 13).

地图

描述已自动生成

Figure 13 The outcome of GWR analysis

表格

描述已自动生成

Figure 14 The diagnostic of GWR

Model diagnostics:

1) R2: 0.7763

- The coefficient of determination indicates that approximately 77.63% of the variance in the dependent variable (house price) can be explained by the GWR model.

2) AdjR2 (Adjusted R2): 0.7756

- This is the R2 value adjusted for the number of explanatory variables included in the model, which is similar to R2, suggesting a good fit that accounts for the number of variables.

3) AIC (Akaike Information Criterion): 334628.9254

- A lower AIC value indicates a better model. This criterion helps compare different models. Without another model's AIC for comparison, it’s difficult to assess this value in isolation.

4) Sigma-Squared: 42430784511.3552

- The sigma-squared value typically represents the variability in the model's errors or residuals. A lower value indicates that the residuals are, on average, closer to zero, which is desirable.

5) Sigma-Squared MLE (Maximum Likelihood Estimation): 42300936546.0507

- This is another measure of the variance of the model's errors, estimated using maximum likelihood. Like sigma-squared, lower values are preferable.

6) Effective Degrees of Freedom: 12241.5061

- This metric adjusts the degrees of freedom for the model based on the spatial structure and data used. In GWR, this can differ from the traditional degrees of freedom due to the local nature of the model.

7) Adjusted Critical Value of Pseudo-t Statistics: 2.8751

- This value is used to assess the statistical significance of local parameter estimates. A higher value indicates a stricter threshold for significance.

The GWR model demonstrates a relatively high R2 and adjusted R2 values, suggesting a good fit to the data. It has utilized a significant number of features and includes key variables such as living space and grade, which appear to explain a large proportion of the variability in housing prices within the given distance band.

**Fitting A Tree-Based Regression Model to Data**

1. **Next, apply the Forest-Based Classification and Regression tool in ArcGIS Pro. Can you use all predictors, why or why not? Describe if this is a good model given diagnostic results.**

**图形用户界面, 应用程序

描述已自动生成**

1. **Display and discuss the meaning of variable importance.**

**表格

描述已自动生成**

Figure 15 The importance table of variables

According to the table, we can see the "Importance" value next to each variable, which may be based on some measure of the variable's influence on the response variable in local regressions, such as standardized coefficients or their contribution to the model's goodness of fit. Here's an explanation of the importance of each variable:

1. grade: The importance value is 0.350638, the highest in the table, indicating that the grade of the house has the greatest impact on housing prices. The grade of the house may represent its overall quality and characteristics.
2. sqft\_living: The importance value is 0.288116, also very high, indicating that the living area size has a significant impact on housing prices. Generally, larger living areas lead to higher prices.
3. sqft\_living15: The importance value is 0.086903, representing that the average living area of the 15 nearest houses also has a relatively large impact on housing prices.
4. bathrooms: The importance value is 0.065834, indicating that the number of bathrooms is also an important predictor of housing prices.
5. view: The importance value is 0.046593, suggesting that the quality of the view also has some impact on housing prices.
6. waterfront: The importance value is 0.048291, indicating that waterfront properties are also an important factor in housing prices.
7. **Document how you set up the training and testing data. Change the test data percentage and explain its effect on the model performance.**

Firstly, the default percentage of the training dataset is 10%.

文本

描述已自动生成

Figure 15 The diagnostics of regression with 10% training dataset

The prediction type I set is ‘Train and predict’, the percentage of the training dataset is 50%, and the diagnostics are shown below (Figure 16).

文本

描述已自动生成

Figure 16 The diagnostics of regression with 50% training dataset

Compared to these different diagnostics, we can see the R2 doesn’t change after the rearrangement of size. However, the importance of variables has slightly raised which indicates the improvement of model.

1. **How can you prune the random forest model? Perform pruning and comment on its impact. Display your result.**

图形用户界面, 应用程序

描述已自动生成

Figure 17 the parameters of how to prune a decision tree

I use the original data to train and test, and the tool automatically to divide the dataset into several parts for training and testing. The number of trees is set to the default value. However, I changed the Minimum Tree Depth to 4 and the Minimum Leaf Size to 6 because I manually chose 6 important variables to run this decision tree model. Pruning the branches (Set depth as 4) can reduce the probability of overfitting.

**The Best Supervised Learning Model**

1. **Compare all regression models with respect to the coefficient of determination. Which one is the most accurate?**

After comparing all regression models, I would like to choose the Geographically Weighted Regression (GWR) model is the most accurate in displaying and predicting the house price in King County, WA, for it offers the greatest R2  (0.78) compared with others (Generalized Linear Regression (GLR) model = 0.55, Forest-Based Classification and Regression model = 0.64). Therefore, with the highest R2, the GWR performs best.

1. **What are the strengths and weaknesses of the methods you used.**

To talk about the GWR, GWR allows for the modeling of spatial variation in data, which traditional regression models that produce global coefficients may not capture. However, we cannot avoid multicollinearity. For example, the variable ‘bathroom’ has a potential relationship with ‘sqft\_living’ which is with the square of living space increase, the house should have more bathrooms. But in these models, we set ‘bathroom’ as one variable and we ignore the endogeneity of variables which results in bias. The important variable ‘grade’ is similar to the ‘bathroom’ that has a potential relationship with ‘sqft\_living’. The issue of endogeneity is a critical concern in many statistical models, including GWR, and can impact the accuracy of the model's estimations.

1. **Which method gave you the most insight for drivers behind the house price in King County, WA. Elaborate your reasoning.**

The Moran’s I gave me a direct insight into how those data should be clustered and how the house price is distributed near Seattle and its surrounding counties. In the outcome, ‘High-high clusters’ could represent areas with higher property values, higher income levels, or any other variable of interest that exhibits a high value. These might be affluent neighborhoods or areas with desirable amenities that boost property values. ‘Low-low’ clusters represent areas where the variable of interest—like property values, for instance—is consistently low, possibly due to factors like lack of amenities, lower income levels, or other socio-economic indicators. Compared to those regression models, though this spatial correlation analysis may not predict house price, it works like running a density estimation to display hot spots.