THE TORONTO HOUSE PRICE PREDICTION MODEL

Executive Summary

The Toronto House Price Prediction Model (THPPM) is a comprehensive project aimed at developing a model to forecast the price of residential properties in Toronto accurately. This project offers invaluable insights into the factors that influence the pricing of properties in Toronto, which benefits a wide range of stakeholders, including real estate agents, homeowners, and prospective buyers.

To create the THPPM, the team used data from two sources: Zillow API and Toronto open data. The final database included ten variables, and a linear regression model was developed to predict the property's final price. The team also utilized data preparation techniques such as hot encoding and outlier detection to ensure data quality.

The final model summary showed an R-squared of 50.2%, indicating that the model can explain 50.2% of the final sold price variability based on the features considered. While the model was tested and found well-fitted, continuous refinement and updates are recommended to improve its accuracy.

To enhance the THPPM's accuracy, external factors such as economic indicators, demographic trends, and local amenities should be incorporated into the model. Doing so will allow the THPPM to account for additional variables that may influence property prices in Toronto.

In conclusion, the THPPM is a valuable tool for predicting the prices of residential properties in Toronto. Its development and implementation offer numerous benefits to stakeholders, including real estate agents, homeowners, and prospective buyers, providing them with valuable insights into the factors that influence property prices in Toronto.

Introduction

The real estate market in Toronto has grown significantly in recent years, creating a highly competitive landscape for buyers and sellers. In recent times, the real estate market in Toronto has grown significantly, creating a highly competitive landscape for buyers and sellers. Many factors have contributed to this, such as the increase in population, inflation, and the recession. Although we cannot bring down housing prices, we can predict how much housing prices are increasing so residents to plan accordingly.

The Toronto House Price Prediction Model (THPPM) aims to develop a model that can accurately forecast the price of residential properties based on their physical characteristics, such as their square footage, number of bedrooms and bathrooms, lot size, and location. While economic conditions, market trends, and local amenities also impact housing prices, the physical attributes of a house remain among the most significant determinants of its market value.

The model will use regression analysis to examine the relationships between housing features and their prices. The methodology involves gathering a diverse dataset of housing prices and physical characteristics from various sources, including public databases and real estate listings. Different techniques are then employed to analyze the data and identify patterns that can be used to predict housing prices with a high degree of accuracy.

This project can potentially benefit a wide range of stakeholders, including real estate agents, homeowners, and prospective buyers. By accurately predicting housing prices based on physical attributes, we can help buyers make more informed decisions and help sellers price their homes more competitively. Additionally, this project can provide valuable insights into the factors that influence housing prices and contribute to a better understanding of the real estate market as a whole.

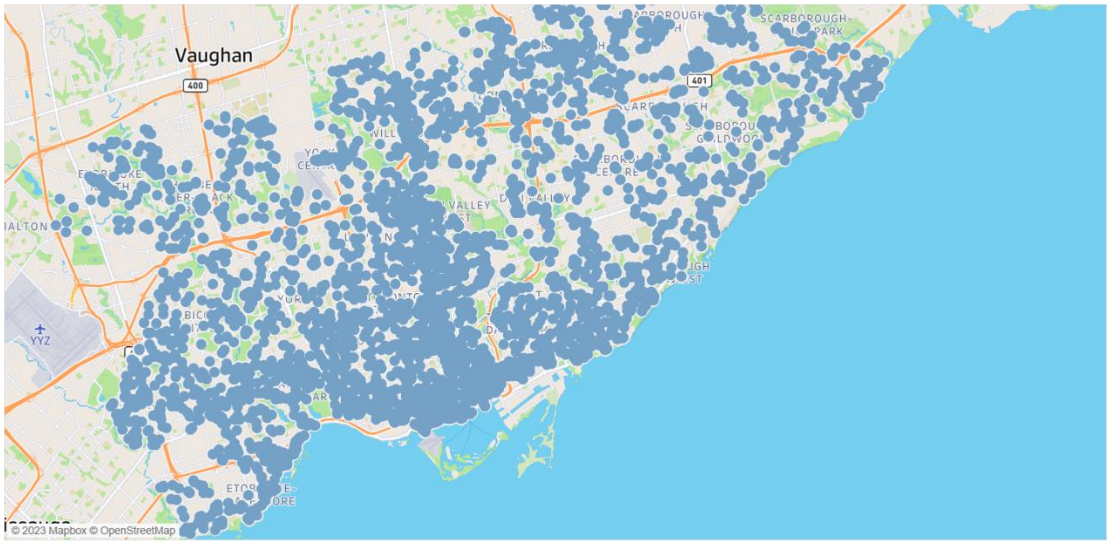
Methodology

A. Data Collection

Zillow API and Toronto open data were used for the model creation. The dataset from Zillow API includes information on house types, room sizes, and addresses of more than 10,000 Toronto properties. Additional information on the locations, including the latitude and longitude of schools and calculation of the distances between the housing properties and schools, were referred from the Toronto open data.

Combining all the information from the two data sources, the final data set used includes variables: number of bedrooms, bathrooms, dens, parking, sizes, list price, the mean income of the neighborhood, property type, addresses of property, distance from schools and housing sold price.

A linear regression model was developed to predict the house sold price based on the other variables included in the dataset.



© 2023 Mapbox © OpenStreetMap

Figure 1: Housing Locations

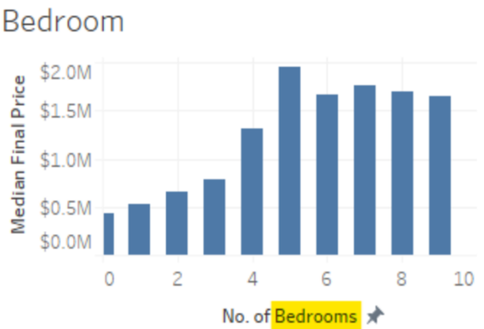


***Figure 6**: Median Final Price of House Based on Number of Bedrooms*

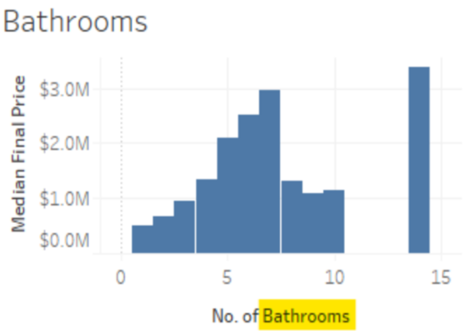


***Figure 6**: Median Final Price of House Based on Number of Bathrooms*

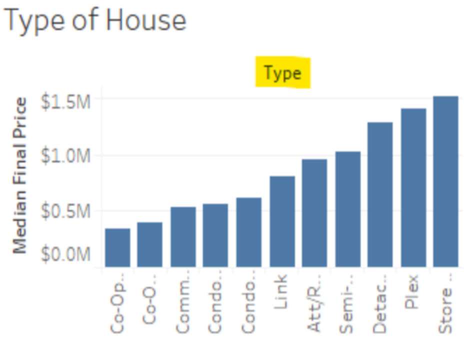


***Figure 6**: Median Final Price of House Based on the House Type*

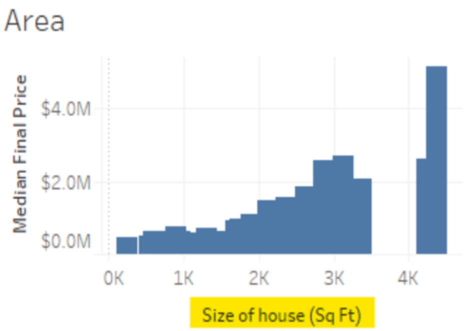


***Figure 6**: Median Final Price of House Based on Housing Area*

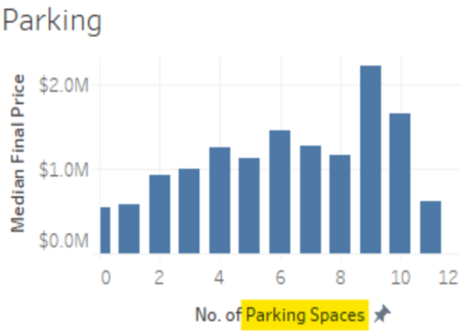


***Figure 6**: Median Final Price of House Based on Number of Parking*

Figure 2 and Figure 3 closely follow a normal distribution where a 5-Bedroom house is relatively pricier compared to others, and houses with more than 7 bathrooms tend to be less expensive compared to those with lesser bathrooms.

The median final price of a house is very low for a co-op apartment and very high for a house with a store with an apartment/office, as shown in Figure 4.

Figure 5 shows that the median housing price follows a consensus in terms of the house size, it has an upward trend, and as the area increases, the house price increases.

On the other hand, the median final price of a house doesn't follow a uniform trend regarding the parking spaces, as shown in Figure 6. The priciest house has 9 parking spaces.

B. Data preparation

Before performing the regression model, two steps were taken to prepare the data.

1. One-hot encoding: to perform the regression model, the house data was converted from string to integer data type: 1 – Co-Ownership Apartment, 2 – Co-Op Apartment, 3- Commercial Element Condo, 4- Condo Apartment, 5 – Condo Townhouse, 6- Semi-Detached, 7- Detached

2. Detection and Removal of Outliers: The boxplot was used to detect outliers, and the interquartile range (IQR) method was used to filter them out.

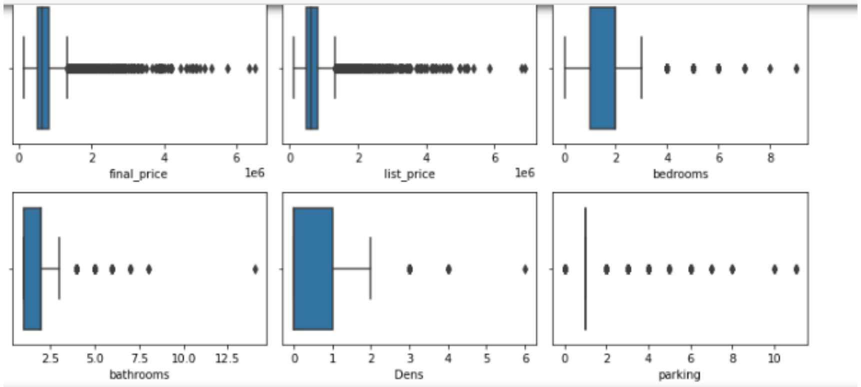


***Figure** 7 - Boxplot and IQR Method*

Results after removing the outliers:

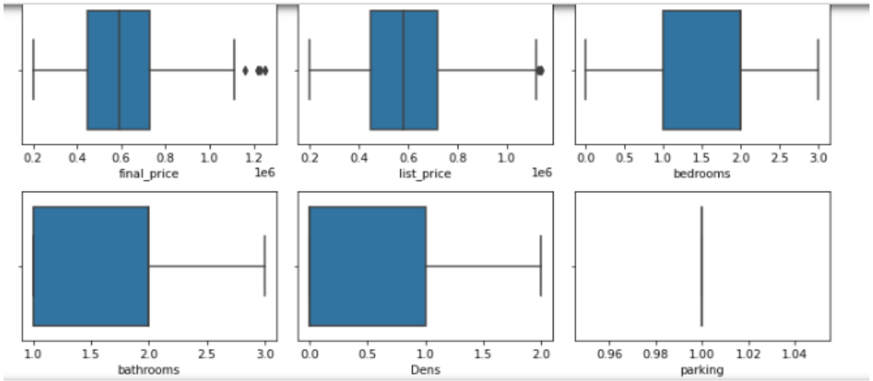
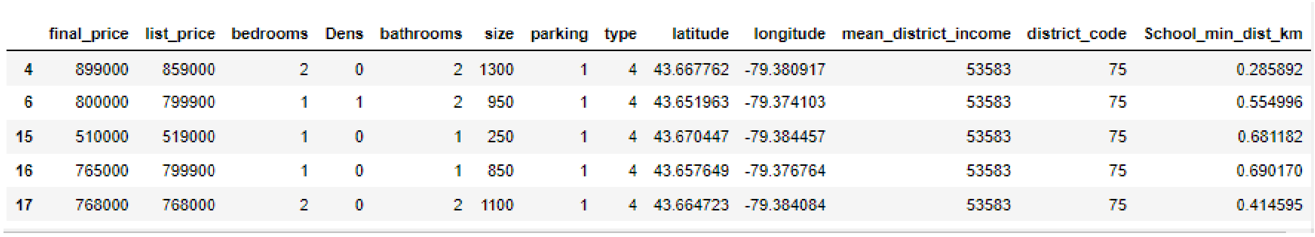


Figure 8: Boxplot after Outlier Removal

Final dataset overview:



	final_price	list_price	bedrooms	Dens	bathrooms	size	parking	type	latitude	longitude	mean_district_income	district_code	School_min_dist_km
4	899000	859000	2	0	2	1300	1	4	43.667762	-79.380917	53583	75	0.285892
6	800000	799900	1	1	2	950	1	4	43.651963	-79.374103	53583	75	0.554996
15	510000	519000	1	0	1	250	1	4	43.670447	-79.384457	53583	75	0.681182
16	765000	799900	1	0	1	850	1	4	43.657649	-79.376764	53583	75	0.690170
17	768000	768000	2	0	2	1100	1	4	43.664723	-79.384084	53583	75	0.414595

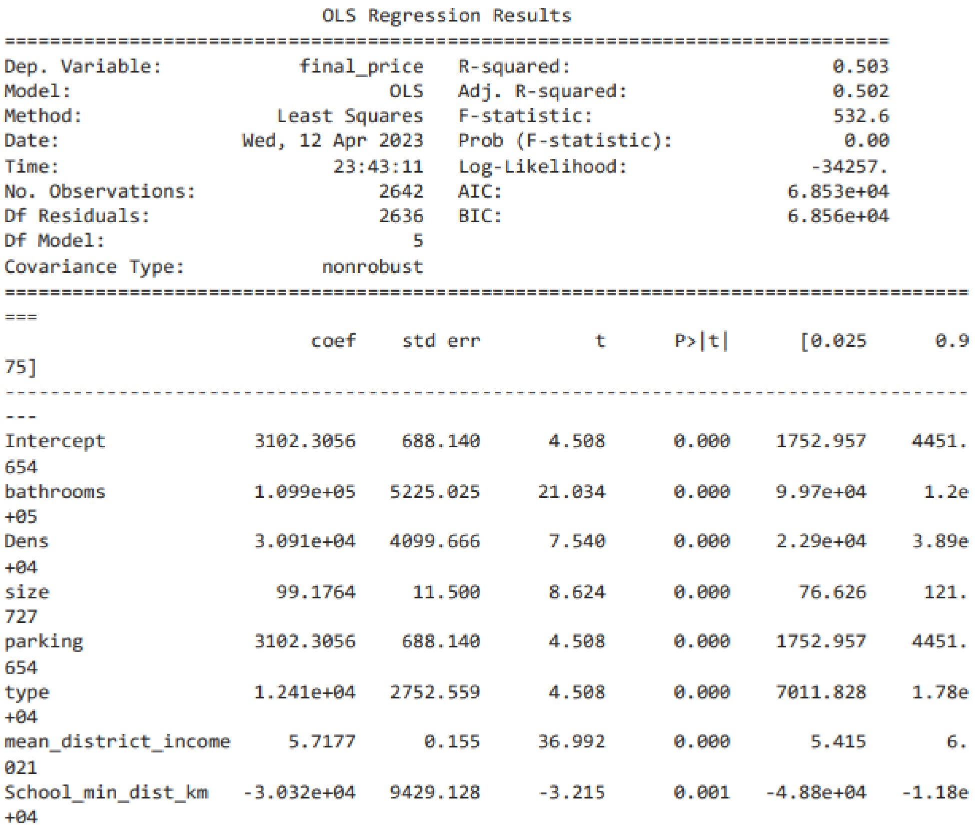
C. Model Building and Training

In this study, a linear regression model was utilized to predict house prices based on independent variables. Linear regression is a commonly used statistical technique for exploring the relationship between dependent and independent variables, and for predicting values based on these variables.

Because the p-value for the individual test of bedrooms is greater than 0.05 and the coefficients for list_price, latitude, and longitude largely influence the predictions to result, we remove these features in the final model.

Based on the final model summary, Figure 9 shows that the R-squared is 50.2%, which indicates that the model can explain 50.2% of the final sold price variability according to the features.

The final model uses the details related to the number of bedrooms, bathrooms, housing size, housing type, parking, district income, and distance from schools.



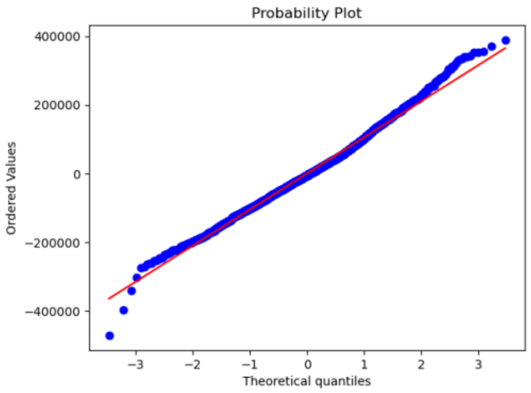
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:            final_price   R-squared:                       0.503
Model:                            OLS   Adj. R-squared:                  0.502
Method:                 Least Squares   F-statistic:                     532.6
Date:                Wed, 12 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:43:11   Log-Likelihood:                -34257.
No. Observations:                2642   AIC:                         6.853e+04
Df Residuals:                    2636   BIC:                         6.856e+04
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
=====================================================================================
===
                           coef    std err          t      P>|t|      [0.025      0.9
75]
-------------------------------------------------------------------------------------
---
Intercept             3102.3056    688.140      4.508      0.000    1752.957    4451.
654
bathrooms             1.099e+05   5225.025     21.034      0.000    9.97e+04     1.2e
+05
Dens                  3.091e+04   4099.666      7.540      0.000    2.29e+04    3.89e
+04
size                    99.1764     11.500      8.624      0.000      76.626     121.
727
parking               3102.3056    688.140      4.508      0.000    1752.957    4451.
654
type                  1.241e+04   2752.559      4.508      0.000    7011.828    1.78e
+04
mean_district_income     5.7177      0.155     36.992      0.000       5.415       6.
021
School_min_dist_km   -3.032e+04   9429.128     -3.215      0.001   -4.88e+04   -1.18e
+04
```

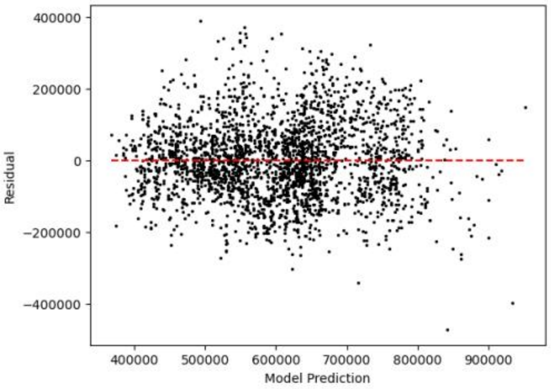
Figure 9: OLS Regression Results

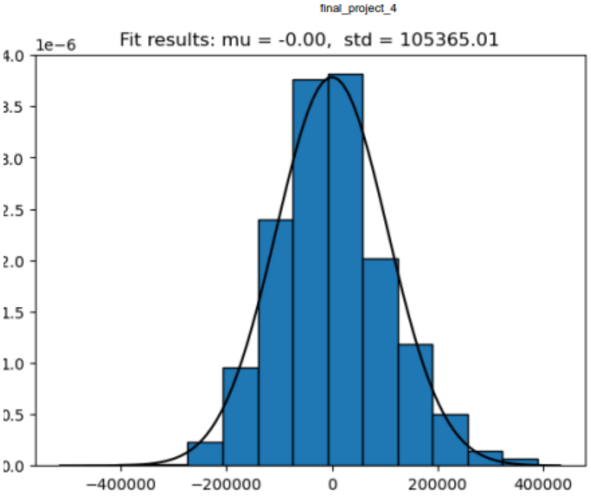
From the final model summary, all features have a p-value less than 0.05, significantly influencing the house price prediction.

D. Model Evaluation

From the residuals vs fitted value and normal Q-Q plot, we can see this model is fitted.

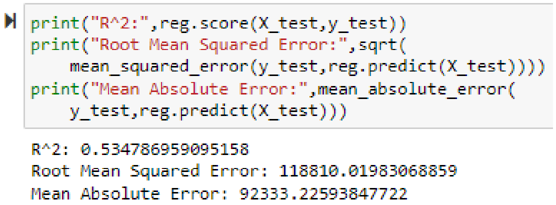






E. Model Test

Test Data

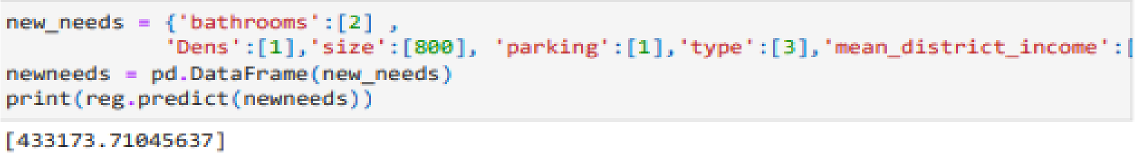


```
print("R^2:",reg.score(X_test,y_test))
print("Root Mean Squared Error:",sqrt(
    mean_squared_error(y_test,reg.predict(X_test))))
print("Mean Absolute Error:",mean_absolute_error(
    y_test,reg.predict(X_test)))
```

```
R^2: 0.534786959095158
Root Mean Squared Error: 118810.01983068859
Mean Absolute Error: 92333.22593847722
```

From the Test Data summary, we can see the stats of the Testing data sets. The R^2 is higher than the Train data sets.

To test the accuracy of the model, we also used information from the needs of a possible client to check and predict the price of the house that the client desires.



```
new_needs = {'bathrooms':[2] ,
             'Dens':[1],'size':[800], 'parking':[1],'type':[3],'mean_district_income':[
newneeds = pd.DataFrame(new_needs)
print(reg.predict(newneeds))
```

```
[433173.71045637]
```

# of Bathroom = 2

# of Bedrooms = 1

House Size = 800 sq ft

# of Parking =1

House Type = commercial element condominium

Predicted Price = ~$433,174

Results and Discussions

The R-squared value of the final model was found to be 50.2%, indicating that 50.2% of the variability in the final sold price can be explained by the features included in the model. Furthermore, it was observed from the final model summary that all features had a p-value less than 0.05, indicating that they significantly influenced the prediction of house prices.

- The number of bedrooms, bathrooms, and parking spaces have a significant positive impact on the house price, while the distance from schools has a significant negative impact.
- The size of the house and the property type also have a significant impact on the house price.
- The mean income of the neighborhood does not appear to have a significant impact on the house price.

In summary, the linear regression model employed in this study was effective in predicting house prices based on the selected independent variables. The findings suggest that the included features significantly influenced the prediction outcomes, and provide insights into the relative importance of these features in the housing market. However, the model's limitations and the potential for unmeasured confounding variables should also be acknowledged. Further research could explore the applicability of these findings to other regions or time periods. It is also recommended to dig deeper into the details of the outliers that were initially taken out from the data preparation process.

Conclusions and Recommendations

Based on the analysis, it is recommended that property buyers prioritize properties with more bedrooms, bathrooms, and parking spaces to maximize the value of their investment. It is also recommended to consider the distance from schools when choosing a property, as this can significantly impact the house price.

Real estate agents and property developers could also use the insights from this analysis to better understand the factors that drive housing prices in Toronto and make informed decisions about where to invest and develop properties.

Furthermore, policymakers could use this analysis to inform decisions about school zoning and infrastructure planning to support the housing market and ensure equitable access to quality education.

The real estate market is constantly evolving, and it is essential to continuously refine and update the model to reflect the latest trends and developments. This can include incorporating new data sources, adjusting the model parameters, and retraining the model periodically to ensure its accuracy and effectiveness. It is essential to evaluate the performance of the model on a regular basis to identify any areas for improvement and fine-tune it to achieve better results. Since the R-Squared of our model is not satisfactory, continuous refinement and updates on the model are recommended.

And while the physical attributes of a house are significant determinants of its market value, other external factors can also impact housing prices. It is recommended to incorporate relevant external factors, such as economic indicators, demographic trends, and local amenities, into the model to improve its accuracy and predictive power.

To enhance the THPPM's accuracy, external factors such as economic indicators, demographic trends, and local amenities should be incorporated into the model. Doing so will allow the THPPM to account for additional variables that may influence property prices in Toronto.

Finally, the analysis provides a starting point for future research on the housing market in Toronto. Additional research could explore the impact of other variables, such as proximity to public transportation, crime rates, or environmental factors, on housing prices. Additionally, further research could explore the applicability of the findings to other regions or time periods