CS777 O1 Fall 2024

Philip Ai

10/09/2024

Topic: **Analysis and Predictions of Living Expenditures Across Major USA Metropolises**

The use of libraries and frameworks: Pyspark, Statsmodel, and Scikit-learn.

**Introduction**

United States of America governs wide range of territories with the same institutions, currency, and languages. Nevertheless, there are some discrepancies across constituents. The cost of living is one of the conspicuous distinctions which average people are experiencing every single day. There is no denying that major cities are like super magnets assembling people to reside, whereas the degree of expenditures to make a living in each metropolitan region differs. This study is going to showcase overall differences of cost of living among 20 major metropolises in US and address the following questions:

Question 1: Compare the average living expenditures over 20 major US metropolis nationwide.

Question 2: Estimate regional cost of living, based on 4 attributes (Consumer Price Index, household Income, population density, and real estate prices).

**Methodology**

Step 1: Data preparations

There are 5 sources of data loaded as dataframes from individual unique CSV spreadsheets. All of them go through preliminary parsing by Pyspark and end up being aggregated into a single curated CSV dataset, consisting of 6 columns, metropolises, median cost of living, median sales prices of real estates, population densities, median household incomes, and Consumer Price Indexes (CPI). The spreadsheet covers 20 major US metropolis.

Step 2: One-Way ANOVA (Analysis of variance)

At 95% confidence level, compare Real Estate Median Sales Price across major metropolises to diagnose whether at least one pair of cities differ. The same method applies to comparison of Consumer Price Index (CPI).

Step 3: Tukey's Honestly Significant Difference (HSD) Multiple Comparisons Test

At 95% confidence level, in case the ANOVA test concludes there is a difference among groups (cities), Tukey’s test is utilized to unfold root causes of significant differences where at least one pair of metropolises will be identified. Step 2 and Step 3 resolve question 1.

Step 4: Multiple Linear Regression with PySpark MLlib

The consolidated dataset is randomly split into training set and testing set in the ration of 4:1, followed by invoking Pyspark regression library to fit selected rows of training data. The equation of multiple linear regression model is produced with a set of coefficients and an intercept value after the training process.

In the end, the model will be evaluated by the testing set allocated at the early stage. Root Mean Squared Error (RMSE) and R-squared (R2) are calculated to measure effectiveness of the model.

Step 5: Multiple Linear Regression with SKlearn

Redo the step 4 by a different library. The consolidated dataset is randomly split into training set and testing set, at the ration of 4:1, following by scaling/standardizing numeric values (optional in regression models), and then invoke linear\_model.LinearRegression library to fit selected rows of training data. The equation of multiple linear regression model is constructed with a set of coefficients and an intercept value after the training process.

In the end, the model will be evaluated by the testing set allocated at the early stage of this phase. Root Mean Squared Error (RMSE) and R-squared (R2) are calculated to assess effectiveness of the model.

Besides, the use of statsmodels library analyzes each computation of regression coefficient to verify if they are significant contributing factors to the model.

Putting parameters and concrete independent variables into the fitted formula can deliver projected cost of living, which is the answer to question 2.

Step 6: Data Visualizations

Draw comparison bar plots of Median Household Income, Cost of Living, and their delta values between metropolises.

Draw comparison box plots of Real Estate cost and CPI across metropolises.

**Results**

1. One-Way ANOVA (Analysis of variance)

A black and white text on a black background

Description automatically generated

The P-value is close to 0, below the significance threshold 0.05 🡪 There must be house price differences between one pair of cities at least, based on 95% confidence level.

A black and white text

Description automatically generated

The P-value is close to 0, below the significance threshold 0.05 🡪 There must be CPI differences between one pair of cities at least, based on 95% confidence level.

1. Tukey's Honestly Significant Difference (HSD) Multiple Comparisons Test

Among 190 combinations of group-by-group (city-by-city) comparisons, the report suggests 171 cases be significant different in terms of real estate prices, under 95% confidence level. The screenshot below indicates the price Boston houses are significantly different from all major cities in US.

A screenshot of a computer code

Description automatically generated

Among 190 combinations of group-by-group (city-by-city) comparisons, the report suggests 179 cases be significant different in terms of consumer price index (CPI), under 95% confidence level. The screenshot below indicates the CPI of Philadelphia is significant different than all major cities in US, expect Washington DC.

A black text with numbers

Description automatically generated with medium confidence

1. Multiple Linear Regression

All the coefficients play significant roles in the regression model (Especially X1: Real Estate Prices), for the reason all of their P-values do not exceed 0.05 significance level (95% confidence level). The understatement of each coefficient is that every single unit change of one independent variables with constraints for the rest of variables (Household Income, CPI, Population Density, or Real Estates) will result in corresponding increments/decrements of dependent variable (Cost of living). The equation is formulated to be:

CostOfLiving = (17656.830691724073\*Median\_HousePrice) + (3141.0836025201875\*Median\_HouseHold\_Index) + (-671.9980505086305\*Population\_Density) + (90764.79060665362) A close-up of a number

Description automatically generated

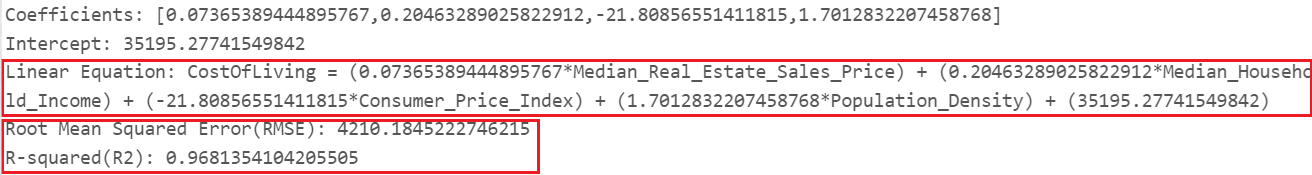
The R2 value holds an excellent caliber position. It indicates that the model is effective, the variability is relatively in line with the model, and that the predicted independent variables are reliable. For this case, the value of 97% of R-squared means that the model explains 97% of the variability in Cost of Living.

Root Mean Square Error (RMSE) is the least squares errors measuring how far apart between predicted values and actual values. Low RMSE values mean that predictions are accurate and fit the model. In this case, the RMSE looks high, but the magnitude of each feature value is relatively higher than normal. As a result, the size of 3587.22 is minimal contrast to hundreds of thousands of dollars of house values. It is still an adequate and acceptable model.

A screenshot of a computer

Description automatically generated

Besides, even if all participating features are not normalized and standardized, the outcome is still promising with evaluations of 96.8% of R2 value and 4210 of RMSE (623 + 3587 = 4210).



**Discussion**

**A graph of a number of people

Description automatically generated** **A graph of a number of people

Description automatically generated**

**A graph of a number of people

Description automatically generated** **A graph with blue squares

Description automatically generated**

**A graph of a number of people

Description automatically generated A graph of blue rectangular bars

Description automatically generated with medium confidence**

A graph of a number of people

Description automatically generated A graph of a comparison of a power

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated A graph of a comparison of power

Description automatically generated with medium confidence

Those bar plots express disparities between median income and expense nationwide. They show the pattern of “the higher amount of the income is, the higher amount of the expenditures is”. Residents of East Coast and West Coast cities generally earn more income. Particularly, there are some metropolises (Los Angeles, Miami, New York, Riverside, San Diego, San Franciso, and Tampa) whose delta between median income and expense is negative. In other words, there are 50% of residents cannot afford livelihood by their meager wage. Their financial health is certainly at risk of poverty.

A graph of a chart

Description automatically generated with medium confidence A graph with blue and black lines

Description automatically generated

A graph of a comparison of consumer price index

Description automatically generatedA graph of a graph showing the price of a real estate

Description automatically generated with medium confidence

A graph of a number of countries/regions

Description automatically generated with medium confidenceA graph with blue and black squares

Description automatically generated

A graph of a comparison of consumer price index

Description automatically generated A graph of a graph showing the average price of a home

Description automatically generated with medium confidence

Those box plots visualize the analysis of the results section above. Apparently, the range of Consumer Price Index (CPI) and real estate prices greatly vary nationwide. However, it is noteworthy that the level of CPI does not necessarily determine corresponding value of real estate’s prices. the level of CPI is positively correlated, instead of positive proportion. For instance, the CPI of Seattle is higher than that of San Francisco, but the house prices in San Francisco is much more exorbitant than Seattle. The identical situation also occurs to Philadelphia VS Phenix, Detroit VS Houston, and Los Angeles VS Miami.

**Conclusion**

This multiple regression model is proven to provide a reliable insight of estimated cost of living based on Household Income, Consumer Price Index, Population Density, and Real Estate Prices. As long as users fill aforementioned variables, it generates suitable prediction. Besides, this model considers the various way of life as well as mixed income levels. The model would adjust the predicted value, despite users enter disproportionately small income. Then, this resident would need to make every effort to search for an inexpensive residence and lead a frugal life to sustain the livelihood. On the other hand, high income residents may be fond of finding a luxurious home and the cost of living is increased accordingly.

It goes without saying the cost of housing is the top foremost factor of living expenditures. Its influence is 5-fold than the runner-up independent variable (population density). That is to say, the metropolis would become badly and ill affordable for residents to settle, once the real estates’ prices surge to the skyrocketed amount. It gives rise to the phenomenon that more and more people emigrate to other satellite regions. Losing residents would make cities fall into a “vicious circle” when prosperity stagnates, criminal rates soar, and other society disorders prevail.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**References**

1. Yahoo Finance: See the Cost of Living in America’s 50 Largest Metro Areas, <https://finance.yahoo.com/news/real-estate-2024-see-cost-140058623.html>
2. Zillow public housing data research: <https://www.zillow.com/research/data>
3. Downloadable Geographic Databases: <https://www.SimpleMaps.com>
4. US Department of Commerce: Household Income in States and Metropolitan Areas (2023)
5. US Bureau of Labor Statistics: <https://www.bls.gov/cpi/regional-resources.htm>