**Data Analysis of the Real Estate Market in Perth**

**The author: Elena Haworth**

**Part A.**

* 1. **Problem Statement and Background:**

The real estate market is experiencing a boom in 2020-2021, experts say that that’s an impact of COVID-19 where people are spending less on traveling, coming back to the country from overseas and they have to buy a property to live in or just invest money in a property and that’s the main reason why the prices started going up. Here comes a question as to what is a real price of the property? An investor (or a family) is planning to buy a house in Perth, WA and he needs an independent assessment and valuation of the property he is planning to invest into. The investor wants to get an understanding whether the price offered by the real estate agency is overestimated or underestimated in order to make a weighted investment decision. The investor’s budget is $800 000.

The real estate has the following portfolio of the properties available for sale at the moment.

|  |  |  |
| --- | --- | --- |
| **DESCRIPTION** | **ADDRESS** | **PRICE** |
| 4bed x 2bath x 1garage, Applecross, floor size 200 m2 | 11A Macrae Road | 1500000 |
| 3bed x 2bath x 2garage, Applecross, floor size 72 m2 | 2/11a Collier Street | 895000 |
| 4bed x 3bath x 2garage, Burns Beach, floor size 320 m2 | 1 Piermont Avenue | 950000 |
| 4bed x 1bath x 3garage, Burns Beach, floor size 120 m2 | 16 Second Avenue | 650000 |
| 4bed x 2bath x 2garage, Karrinyup, floor size 185 m2 | 11 Trusley Way | 960000 |
| 4bed x 2bath x 2garage, Victoria Park, floor size 161 m2 | 101 Teague Street | 720000 |

* 1. **Resources.**

The type of data source is secondary, it was obtained as a structured dataset in csv format from the website <https://www.kaggle.com/syuzai/perth-house-prices>. [1]

**Acknowledgements.** According to the dataset description [1] this data was originally scrapped from the website <http://house.speakingsame.com/>. It includes data about 322 Perth suburbs and about 100 rows per suburb in average. Longitude and Latitude data was acquired from the website [www.data.gov.au](http://www.data.gov.au). School ranking data was acquired from the website <https://bettereducation.com.au/Default.aspx>. [1]

The subject of the research is individual residential properties in Perth, WA.

The data set represents a portfolio of 33656 properties that were sold in Perth, WA between 1988 and 2020. The data set consists of 33656 rows and 19 columns.

The value of a residential property is viewed as a combination of characteristics (such as location, size, construction, etc.) that contribute to the price of the property in some measurable way. The property has such characteristics as address, postcode, suburb, number of bedrooms, number of bathrooms, garage size, the size of the land, the size of the living area, the year the property was built. Neighboring characteristics that are relevant to the property: distance to CBD, the nearest to the property station, distance from the property to the nearest station, latitude, longitude, the nearest school, the distance to the nearest school and school’s rank. The selling characteristics: the price of the property, the date when the property was sold.

**Format**

The data frame contains the following columns:

|  |  |
| --- | --- |
| Name of the column | Description |
| 1. ADDRESS | the address of the property |
| 1. SUBURB | the suburb where the property is located |
| 1. PRICE | the price the property was sold |
| 1. BEDROOMS | number of bedrooms in the property |
| 1. BATHROOMS | number of bathrooms in the property |
| 1. GARAGE | number of car spaces in the property |
| 1. LAND\_AREA | the size of the land |
| 1. FLOOR\_AREA | the size of the living area of the property |
| 1. BUILD\_YEAR | the year when the property was built |
| 1. CBD\_DIST | the distance to CBD |
| 1. NEAREST\_STN | the nearest train station |
| 1. NEAREST\_STN\_DIST | the distance to the nearest station |
| 1. DATE\_SOLD | the date when the property was sold |
| 1. POSTCODE | post code |
| 1. LATITUDE | the angular distance of a place north or south of the earth's equator |
| 1. LONGITUDE | the angular distance of a place east or west of the Greenwich meridian |
| 1. NEAREST\_SCH | the nearest school |
| 1. NEAREST\_SCH\_DIST | distance to the nearest school |
| 1. NEAREST\_SCH\_RANK | the nearest school’s rank |

The snapshot of the data base from csv file.

Graphical user interface, application, table, Excel

Description automatically generated

In this research I’m planning to use Excel and Jupyter notebook using Python language.

To perform the data analysis and machine learning in my research I’m planning to use an open-source interactive web tool Jupyter and *Python* language. Python Packages that are going to be used: Pandas, Numpy, Matplotlib, Pydotplus, Sklearn, Seaborn.

* 1. **Business requirement**

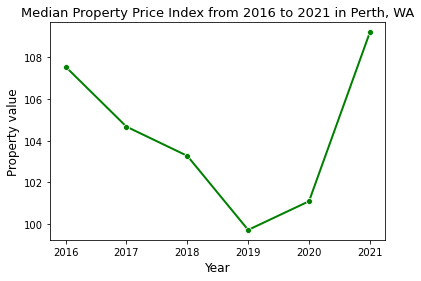
My research will help to understand what the trends of the property market over the last 30 years are, what characteristics of the property affect the price and contribute to the value of the property.

By looking at the graph of the median property price from 1988 to 2020 in Perth, Western Australia, an upward trend is distinctly observed in the housing market up until 2007. In 2008 there was a global crisis and the housing market crashed, following by a rocket rise in 2009 and 2010. Between 2012 and 2020 the market is showing more or less stable trend, though a slight downward trend is noticeable.

Chart, line chart

Description automatically generated

The year of 2020 became a year of change due to COVID-19. According to Australian Bureau of Statistics the prices for houses in Perth “rose 9.9% over the last twelve months as of March 2021. This was the largest quarterly rise in the Residential Property Price Index for Perth since the December quarter 2009.” [2]



As the price is going up people are becoming more interested in investing into the property, the need of evaluating a property has become in the focus. What does the price depend on? How to understand if the price of a property that is in the market is overestimated or underestimated?

* 1. **Business Model**

The current research is sitting in an area of the real estate market and will help to estimate the value of property and give an indicator to an investor which dwelling units are overpriced or underpriced. The advantage of this area is that the data is often available online and can be acquired from real estate web sites. But the real challenge is that when the data is scaped from a web site, it is in disparate format, very inconsistent and it takes a lot of time and efforts to bring this data into a structured format so that it is ready for data analysis. In order to answer the question raised in the Problem statement section of this assignment I’m planning to perform the following steps.

1. Find a suitable dataset in csv format (found on web site [www.kaggle.com](http://www.kaggle.com) to save time on web scrapping and formatting the data).
2. Import the dataset into Jupyter notebook using the Python language
3. Perform Data wrangling and cleaning
4. Perform Exploratory data analysis and feature selection
5. Model development: Decision tree regression and Linear regression, and model evaluation
6. Interpret and implement the analysis results and give advice

This research will show the characteristics that affect the property value, will build a model for estimating a property price and will give advice if a property is overpriced by comparing the model price and the market price in order to make a decision to buy or not to buy.

* 1. **Data Analysis**

After loading the dataset in Jupiter notebook, I started exploring the data.

There were missing values in three columns: GARAGE, BUILD\_YEAR and NEAREST\_SCH\_RANK.

The values in the columns BUILD\_YEAR and GARAGE can’t be restored with any sort of computation or interpolation methods as it won’t make sense. For my research I decided to drop the rows with NaN values.

As for the column NEAREST\_SCH\_RANK, initially I decided to drop the entire column from the dataset because the total amount of missing values is about 32% of the values in this column, and this is considered as significant and as a result can affect the analysis. But thinking that this might be quite a strong factor when a family is buying a house, I decided to leave this column and drop NaN value rows.

I decided to drop the columns ADDRESS, POSTCODE, NEAREST\_STN, NEAREST\_SCH as they are not useful features in the analysis.

While examining the description of the dataset, there are some suspicious values that can be either the outliers or errors are noticed.

Graphical user interface, application, table, Excel

Description automatically generated

Column GARAGE contains max value 50, which is most likely an outlier or an error. I decided to select the rows with values bigger than 10 in the column GARAGE. Looking at the selected data I can assume that these values in the GARAGE column are errors and should be removed from the dataset as 3-bedroom house with the price of $360000 and land size 692 m2 doubly will have a garage with a capacity for 12 cars.

I applied the similar logic to the column LAND\_AREA with a maximum value of 999999.

The other way to check the outliers was to plot the data. The boxplot between the number of bedrooms and the price showed that some of the price values are outside the range of data. But I assume that this can be, because the price could be affected by other features like bigger land area, luxury suburb or the nearest top ranked school.

Graphical user interface, application

Description automatically generated

Grouping the data by number of bedrooms and plotting the result in a bar plot using seaborn library shows that 3- and 4-bedroom houses are the most selling houses.

Graphical user interface

Description automatically generated

The target variable is the column PRICE, it represents the price at which the property was sold. The dependent variables are the rest.

My next step was performing a correlation analysis in order to understand the relationships between the targeting variable and the rest of the variables and also, I used the correlation coefficients for a feature selection. I build a heatmap for all features and calculated the correlation with the targeting variable PRICE by using the Pandas method corr() and sorted the result.

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

It can be noticed that the Price of the property is most positively correlated with the size of the floor area (corr = 0.57), i.e., the bigger the floor area the higher the price. The other variables that affect the price are the school rank, the number of bathrooms, bedrooms and garage size, LONGITUDE (can be interpreted as a measurement of distance to the ocean given that a longitude is the angular distance of a place east or west of the meridian) as well as distance to CBD and to the nearest school.

Based on this analysis I selected features for the further analysis (|corr| > 0.1) and drop the others.

Selected numerical columns: FLOOR\_AREA, BATHROOMS, BEDROOMS, GARAGE, BUILD\_YEAR, CBD\_DIST, LONGITUDE, NEAREST\_SCH\_RANK.

At this point the dataset is cleaned and ready for the further analysis.

**Modelling.**

As the target variable (Price) is numerical and continuous in nature, a regression method will be used in order to build a machine learning model. I’m going to evaluate the price of the property through Decision Tree Regression and Linear Regression in Python using scikit-learn machine learning library.

I split the data into training and testing set with a ratio 80:20 and I created a regressor object using method DecisionTreeRegressor (max\_depth=4, random\_state=1234), after which I fitted regression model with training data. The R-square of the Decision Tree regression model is: 0.6078980352430245 as a measure of accuracy of the model. The visualization of this model is in the picture below.

Graphical user interface, text, application, Word

Description automatically generated

The decision tree regression model did not work accurate enough due to big variation of data. Another model that I’m building is a Linear Regression model for the same training data. The accuracy of the linear regression model appeared to be much better with R-square= 0.7993302101268622. It can be seen on the graph below that the predicted values and tested values are distributed more accurate.

Graphical user interface, chart, scatter chart

Description automatically generated

**Implementation of the machine learning model.** I have a trained model representing a valuation of a property. By inputting characteristics of a property that an investor wants to evaluate into this model, he can receive predicted value of the property, compare it with the market price and make a decision whether this property is undervalued, and it would be a good investment or overpriced and it makes sense to negotiate the price or reject this option.

The result of the modeling is presented on a picture below.

Graphical user interface, chart

Description automatically generated

It became obvious that the price is strongly dependent on the size of the floor area, number of bathrooms, bedrooms and the size of the garage in terms of the property attributes, and also the location plays an important role in forming the price: the proximity to CBD and to the ocean. Another important feature that affects the price is the nearest school ranking, as the schooling system works in Australia in a way that a student should live in the catchment area in order to be able to study at a certain school, thus a good school plays an important role when a family is choosing a house to buy.

An investor was offered 6 options of the property available in the market. A short description of the properties can be seen in the picture above. The evaluation of the price and recommendations are presented in the table below.

Graphical user interface

Description automatically generated

Please see the link to my presentation:

<https://drive.google.com/file/d/1NaaXMgdCzL8mXV6zfpnIq6TSAqsmGB23/view?usp=sharing>

# References

|  |  |
| --- | --- |
| [1] | [Online]. Available: https://www.kaggle.com/syuzai/perth-house-prices. |
| [2] | ABS. [Online]. Available: https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/residential-property-price-indexes-eight-capital-cities/latest-release. [Accessed August 2021]. |