# **COMP6200 - Data Science**

Project Proposal

*Submitted by:*

|  |  |
| --- | --- |
| Student Name | Student ID |
| Aarushi Chaudhary | 46113908 |
| Ameer Ali Mavani | 46812342 |
| Arnav Bajaj | 46806423 |
| Kaveri Bhatia Chanana | 46282920 |

## **Data Analysis on House Sales and Rent Data in NSW**

The property market in New South Wales, particularly in Sydney, has been rising in recent years. The cost has increased by nearly 15% in the last year. As the number of individuals coming to NSW to study or work grows, everyone should be able to find a suitable home to live that is both affordable and safe. We will utilize statistics on income, crime, housing sales and rent to characterize all of NSW’s Local Government Areas (LGAs) and see if there is a relationship between them.

## 

## **Project Goal**

* To predict the house sale prices for all LGA’s based on the crime rate and income data,
* To find similar LGAs based on certain characteristics,
* To find the average rental cost per Suburb,
* To find the average sales prices for all LGAs, and
* To find the most and the least popular suburbs/LGA’s for renting as well as sales prices.

## **Data Source and Background**

The data for this project is gathered from the website of [NSW Department of Family and Community Service](https://www.facs.nsw.gov.au/resources/statistics/rent-and-sales/dashboard) , which has published this data quarterly since 2008. For this project, we will analyze the rent data for the June 2021 quarter and sales data for the March 2021 quarter.

Using the table builder, we got the income data from the [Australian Bureau of Statistics Census](https://www.abs.gov.au/statistics/labour/earnings-and-work-hours/personal-income-australia/latest-release#data-download). However, since the census data for year 2021 is not yet available, we will be using the data from 2017-2018.

We gathered the crime data from the [NSW Bureau of Crime Statistics and Research](https://www.bocsar.nsw.gov.au).

## **Format**

The sales data includes sales information for each LGA whereas the rent data includes rent information all LGAs in Sydney. The income data contains information on median income for all LGAs, and we will be using the data for NSW in our project. The crime data from the NSW Bureau of Crime Statistics and Research entails monthly offenses committed across various LGAs in NSW from January 1995 to June 2021. However, for this project, we will use the data for year 2021. The sales, rent, income, and crime data are all stored in CSV format.

## **Work Requirements**

### Data Cleaning and Manipulation

***Missing values***: The rent and sales data have some columns with “-“ and “s” values in them. “-“ signifies 10 or fewer bonds lodged for the rent data and 10 or fewer properties sold for the sales data and “s” signifies 30 or fewer bonds lodged for the rent data and 30 or fewer properties sold for the sales data.

We will replace these values with 0 in this case.

***Dropping columns*:** To achieve the project’s goal and meet its requirements, we will drop columns from the chosen dataset.

We will also merge the datasets to create a data frame, which will allow us to complete our project’s analysis.

## **Techniques expected to be utilized in the project**

Based on income and crime data, we expect to create a linear regression model in order to predict house sales. This would also imply the application of various techniques (or metrics) for evaluation such as the use of Mean Square Error (MSE) and R-squared (r2) in the case of a regression problem. We would start with a simple algorithm (or model) as a foundation for our project and gradually incorporate more complex techniques.

Secondly, we will use the logistic model to group suburbs together based on similar characteristics. However, as we continue with project, the idea will be further developed, and we will expand on our approach. We will use the RFE technique to determine what features affect the classification.

The data analysis done will be viewed using visualization tools.

## **Project Plan**

Week 10: Complete data cleaning and manipulation

Week 11: Complete model building with at least two analytical techniques

Week 12: Run tests and evaluate model