# Table of Contents

[Table of Contents 1](#_Toc72072913)

[Executive Summary 4](#_Toc72072914)

[1.0 Situation Context & Objectives 5](#_Toc72072915)

[1.1 Objectives and Goals 5](#_Toc72072916)

[1.1.1 Overview of Home Ownership in New Zealand 5](#_Toc72072917)

[1.1.2 Housing Affordability in New Zealand 5](#_Toc72072918)

[1.1.3 Problem Area 6](#_Toc72072919)

[1.1.4 Current Solutions 7](#_Toc72072920)

[1.1.5 Objectives of Study 7](#_Toc72072921)

[1.1.6 Success Criteria 7](#_Toc72072922)

[1.2 Assessment of Situation 8](#_Toc72072923)

[1.2.1 Key Data Sources 8](#_Toc72072924)

[1.2.2 Requirements Assumptions and constraints 9](#_Toc72072925)

[1.2.3 Risks and Contingencies 11](#_Toc72072926)

[1.2.4 Terminology 12](#_Toc72072927)

[1.3 Data Mining Objectives 13](#_Toc72072928)

[1.3.1 Objectives 13](#_Toc72072929)

[1.3.2 Success Criteria 13](#_Toc72072930)

[1.4 Project Plan 14](#_Toc72072931)

[1.4.1 Risks and Resources 14](#_Toc72072932)

[1.4.2 Project Plan 16](#_Toc72072933)

[2.0 Data Understanding 17](#_Toc72072934)

[2.1 Collection of Initial Data 17](#_Toc72072935)

[2.1.1 Data Collection, Formats and Potential Challenges 17](#_Toc72072936)

[2.1.2 Analysis and Prediction from Datasets 19](#_Toc72072937)

[2.2 Description of Data 20](#_Toc72072938)

[2.2.1 Data Quantity and Quality 20](#_Toc72072939)

[2.3 Exploration of Data 26](#_Toc72072940)

[2.3.1 Household Labour Force Survey 26](#_Toc72072941)

[2.3.2 Household Economic Survey 30](#_Toc72072942)

[2.3.3 Consumer Price Index 33](#_Toc72072943)

[2.3.4 House Price Index 35](#_Toc72072944)

[2.3.5 Residential Sales Statistics 36](#_Toc72072945)

[2.4 Verification of Data Quality 39](#_Toc72072946)

[2.4.1 Household Labour Force Survey 39](#_Toc72072947)

[2.4.2 Household Economic Survey 40](#_Toc72072948)

[2.4.3 Consumer Price Index Data 41](#_Toc72072949)

[2.4.4 House Price Index Data 42](#_Toc72072950)

[2.4.5 Residential Sales Data 43](#_Toc72072951)

[3.0 Data Preparation 45](#_Toc72072952)

[3.1 Data Selection 45](#_Toc72072953)

[3.2 Data Cleaning 51](#_Toc72072954)

[3.3 Data Construction 56](#_Toc72072955)

[3.4 DataSource Integration 72](#_Toc72072956)

[3.5 Data Formatting 78](#_Toc72072957)

[4.0 Data Transformation 80](#_Toc72072958)

[4.1 Data Reduction 80](#_Toc72072959)

[4.2 Data Projection 83](#_Toc72072960)

[5.0 Data Mining Method Selection 89](#_Toc72072961)

[5.1 Discussion on Data Mining Objectives and Methods 89](#_Toc72072962)

[5.1.1 Classification 89](#_Toc72072963)

[5.1.2 Regression 89](#_Toc72072964)

[5.1.3 Clustering 90](#_Toc72072965)

[5.2 Data Mining Method Selection 91](#_Toc72072966)

[6.0 Data Mining Algorithm Selection 92](#_Toc72072967)

[6.1 Exploratory Analysis and Discussion 92](#_Toc72072968)

[6.1.1 Discussion on Algorithm Theory 92](#_Toc72072969)

[6.1.2 Discussion on Models 97](#_Toc72072970)

[6.2 Data Mining Algorithm Selection 102](#_Toc72072971)

[6.3 Model and Parameters 105](#_Toc72072972)

[7.0 Data Mining 110](#_Toc72072973)

[7.1 Test Designs 110](#_Toc72072974)

[7.2 Data Mining Activity 111](#_Toc72072975)

[7.3 Patterns from Model Output 113](#_Toc72072976)

[8.0 Interpretation 114](#_Toc72072977)

[8.1 Discussion on Mined Patterns 114](#_Toc72072978)

[8.2 Visualisation of Data, Results, Models and Patterns 114](#_Toc72072979)

[8.3 Interpretation of Results, Models and Patterns 128](#_Toc72072980)

[8.4 Assessment and Evaluation of Results, Models and Patterns 129](#_Toc72072981)

[8.5 Iteration of Steps 1-7 132](#_Toc72072982)

[9.0 References 134](#_Toc72072983)

[List of Tables 136](#_Toc72072984)

[List of Figures 136](#_Toc72072985)

# 

# Executive Summary

Around the world, cities are growing at an unprecedented rate to accommodate the growing population. This presents the local economies with the foundation for the growth of business opportunities and job creation. To sustain growing populations and expanding cities, affordable housing is required. Safe, reliable, stable, and comfortable accommodation is essential for social and individual wellbeing. Urban centres worldwide are under stress to house their poorest residents through social or public housing. Still, the low-to-middle income households often rely heavily on market-based mechanisms to access affordable housing. This, in turn, requires efficient regulatory and governance measures to maintain public and investor confidence in the property markets. To enable and sustain an environment for affordable housing, efforts and contribution from the public sector, private sector, and non-profit stakeholders are essential. Infrastructure, investment, and local and central government's economic policies aimed at social and financial inclusion are required (Menon, Hodkinson, Galal, Charles, & Reckford, 2019).

Housing affordability remains a critical challenge for New Zealand in social and economic terms. Declining homeownership rates have made New Zealand one of the most unaffordable countries to live (pwc New Zealand, 2020). This study focuses on housing affordability within New Zealand for younger kiwis to understand the need for alternative solutions to tackle an increasingly worsening situation.

# 1.0 Situation Context & Objectives

## 1.1 Objectives and Goals

### 1.1.1 Overview of Home Ownership in New Zealand

The homeownership rates have been the lowest in New Zealand since the 1950s, according to the data collected from the 2018 Census. The census data shows the homeownership rates at the highest in the 1990s at 74%, and by 2018 the rates have fallen to 65%, the lowest since 1951. Tasman at 75.6% and Marlborough at 72.5% have the highest rates of homeownership, and Auckland and Gisborne have the lowest in 2018, at 59.4%. The proportions of households that own their own homes vary significantly across New Zealand (Dickson et al., 2020).

Unlike many other goods and services, housing expenditure whether it is renting or owning make up a significant portion of a household income. It also accounts for a large share of household wealth and retirement accumulations for many New Zealanders (Dickson et al., 2020). However, the declining homeownership rates may reduce the ability to pass on wealth and exacerbate the gap between older and younger generations.

### 1.1.2 Housing Affordability in New Zealand

The lack of housing affordability varies from country to country and city to city. Still, the most common causes include housing costs increasing faster than income, the supply of houses unable to meet the demand, scarcity of land, and demographic changes such as population growth, aging and changes in household composition (Menon et al., 2019).

Defining affordability goes beyond the cost of purchasing a house. It needs to consider the operating and maintenance costs as well. Furthermore, housing affordability there are several interrelated factors that affect housing affordability in New Zealand.

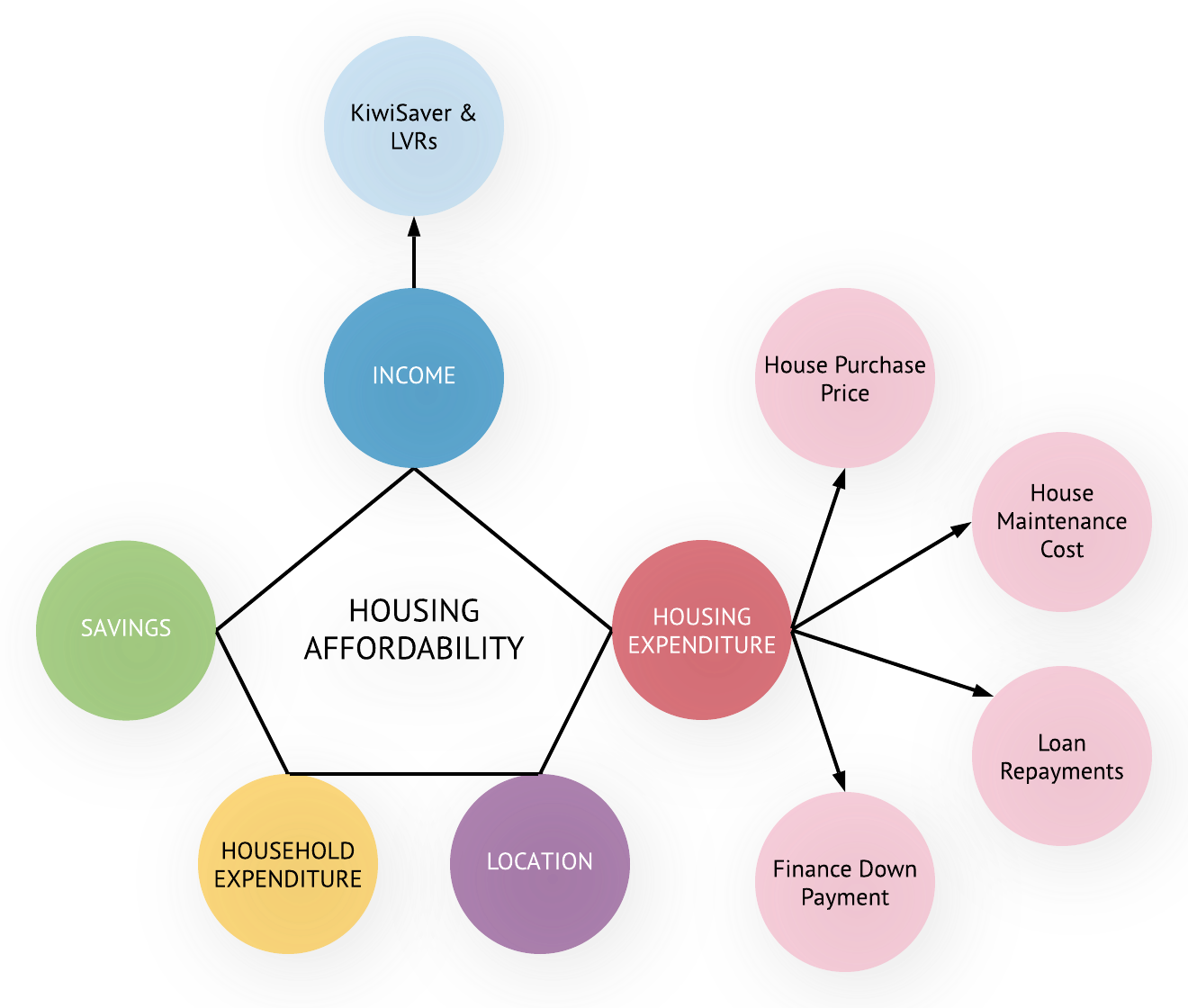


Figure 1 - Basic components of housing affordability

### 1.1.3 Problem Area

According to the census data collected from 1986, homeownership rates have been the lowest in the age group of 20 to 35 years (Dickson et al., 2020). Rising house price inflation, especially in Auckland over the past decade, can potentially be recognised as a risk to financial and economic stability. New Zealand’s house prices, in comparison to incomes or rents, are considered high on an international basis and very high when compared to New Zealand’s historical trend (Dickson et al., 2020).

### 1.1.4 Current Solutions

The Reserve Bank of New Zealand has in the past and continues to use today the official cash rate (OCR) to influence the interest rates as well as use the loan-to-value ratio (LVR) restrictions (since October 2013) to influence and control the house price inflation (Reserve Bank of New Zealand (2015)). The government’s KiwiBuild policy and the ability for first home buyers to use their Kiwisaver funds to contribute towards purchasing a home have not been entirely successful. Regardless of these measures, house price inflation continues to increase. The property market imbalance is more significant in Auckland than in other regions within (New Zealand Reserve Bank of New Zealand (2015)).

The reasons for property market imbalance are primarily linked to the fast growth of the migrant population and the slow supply of housing stock. The supply of housing has been a long-term issue across New Zealand. It is further complicated with other range of topics such as scarcity of suitable land, planning and zoning restrictions, lack of coordination in infrastructure planning, and fragmented and inefficient building industry. The house price inflation is further aggravated with the arrival of permanent migrants prior to border closure due to Covid-19. The declines in interest rates, rebalancing of investment portfolios and temporary removal of LVR on lending have encouraged house price inflation. These factors have contributed to and increased the number of hurdles the younger adults have to pass to afford a house in any region of New Zealand (Reserve Bank of New Zealand (2021)).

### 1.1.5 Objectives of Study

With the rising concerns of house price inflation and affordability of housing for first home buyers in New Zealand, it is vital to observe the problem in detail and explore the viable solutions that can be applied. This study will focus on the understanding of complex factors that affect housing affordability and determine if alternative solutions are required. The main objective of this exercise is to determine:

1. If government intervention is required to support housing affordability for young kiwis aged between 20 to 35 years.
2. If government intervention is required, should it apply uniformly across all of New Zealand, or should it be region specific.

### 1.1.6 Success Criteria

The success of the study is judged by the ability to identify the government and policymakers, the gap between the ability to afford a house and the actual house price inflation. The study should also conclude if the gap between the affordability and actual housing market varies across New Zealand.

## 1.2 Assessment of Situation

### 1.2.1 Key Data Sources

This data mining study brings together housing information from several sources to address the study objectives. The key sources of data sets are listed below:

1. The New Zealand Census of Population and Dwellings provides a snapshot every five years of the official count of people and dwelling in New Zealand. Information related to study from the census (Statistics New Zealand (2018)):

* Population
* Age
* Location

1. The Household Economic Survey (HES) is a survey that collects data on household incomes and expenditures. The survey is carried out typically every year since 1974 (excluding the years from 1998-2007) by the Ministry of Housing and Urban Development. It uses a sample size between 3,000 - 5,000 households each year but has since risen to 20,000 homes since 2019 (Statistics New Zealand (2019)).

* Location
* Average Annual Housing Costs
* Average Annual Household Disposable Income
* Size of Households

1. Quarterly Employment Survey provides data on employment in New Zealand. It includes levels, changes, the sum of total earnings (hourly or weekly), and hours worked. In this case, data on earnings is the key take away from the survey that relates to the study (Only if census has insufficient information) (Statistics New Zealand (2016)).
2. Household Labour Force Survey (HLFS) provides regular and comprehensive information on New Zealand’s working population. It includes various statistics on household incomes, regions, genders, and the data is presented according to years (Statistics New Zealand (2020)).
3. The Consumer Price Index (CPI) measures the changing price of a fixed basket of goods and services purchased by New Zealand households. CPI is published quarterly by Statistics New Zealand (Statistics New Zealand (2020)).
4. Loan-to-Value Ratio data related to residential housing from the Reserve Bank of New Zealand since its introduction on 1st October 2013 (Reserve Bank of New Zealand (2021)).
5. House Price Index sourced from the Reserve Bank of New Zealand and Core Logic. The data provides the changes in house prices from year to year (Reserve Bank of New Zealand (2021).
6. Residential property sales statistics obtained from the University of Auckland Library database, sourced initially from CoreLogic. CoreLogic is the largest provider of property information, analytics, and geospatial location information in Australia and New Zealand. These data files include information on geographic location, sales dates and prices, capital and land values, building floor areas, land areas and building type (CoreLogic (2021)).

Multiple data sources have been identified to be studied more in-depth to understand, which provide the most accurate information. Specific datasets offer a reliable source of information in situations where data gaps occur.

### 1.2.2 Requirements Assumptions and constraints

**Requirements**

1. **Functional**

* An analytical outcome that explains the following:
  + Understand income and expense habits and saving trends.
  + Understand and explain the current housing market and price increases.
  + All of the above at national and regional levels
* Understand and explain various datasets involved and their individual contribution to the study.
* Establish and explain correlations between datasets.
* Draw conclusion on level of aid required at national and regional levels.

1. **Non-Functional**

* The study should abide by New Zealand privacy and copyright laws.
* Publicly available datasets should be used.
* No individual’s data should be disclosed.

**Assumptions**

1. Data mining knowledge and understanding required for the study to be learned through the course and recommended readings.
2. Software and hardware required to carry out the data mining process.
3. Auckland university student login is required to access many of the softwares.
4. There was a seven-year gap between 2006 and 2013 due to postponement caused by the 2011 Christchurch earthquakes.
5. The Covid-19 pandemic has caused disruption in the housing market due to the monetary policies that are implemented by the government to counteract the impacts of the pandemic. These policies may have short-term and long-term impacts on the New Zealand property market.

**Constraints**

1. Statistics New Zealand Microdata includes unit record data and summary data. This microdata may contain identifiable, privacy-sensitive information and records of individual people, households, businesses, and organisations that may be more helpful and accurate for the study (Statistics New Zealand (2020)). Due to the confidentiality and privacy concerns of the data, it is not readily available. It requires a lengthy application process and written authorisation to access it. Therefore, only the publicly released version of this data is used for this exercise.
2. Data available is already aggregated and so may affect the accuracy.
3. For the purposes of this study, affordable houses are defined as buildings designed and intended for private accommodation, such as detached houses, townhouses, units, and apartments.

4. Data available in certain circumstances may be at least a year old.

### 1.2.3 Risks and Contingencies

Table 1 - Risks and contingencies

|  |  |
| --- | --- |
| **RISKS** | **CONTINGENCIES** |
| Data Coverage | The data available is largely aggregated into broad regions in New Zealand. Smaller regions are consolidated into larger; this may affect the overall region results. The result may not be as fine-grained. |
| Data Gaps | Specific datasets have been published multiple times a year, whereas others have been published once in five years. In other circumstances, certain events have caused delays in timely data collection, such as the delay of census due to 2011 Christchurch earthquakes. These kinds of events have affected all data sources mentioned above in some form or other.  These data gaps do limit the accuracy of the results. They can be compensated with other well-established data sources. For this study’s purposes, the data gaps are to be identified, therefore providing the context and clarity to the results of the study. Where possible, the data gaps can be filled in from other government sources and identified where this occurs. The importance is placed on eliminating hidden gaps, which may cause misinterpretation of results. |
| Results are inconclusive in certain age groups or certain regions | The report must clarify the reasons for the inconclusive results for each iteration. It should identify the missing data or factors that may cause this situation. Following the identification of the cause, action can be taken to obtain the missing information from recognised sources, therefore providing more clarity to the situation and further information for the next iteration. |

### 1.2.4 Terminology

Table 2 - Terminology and abbreviations

|  |  |
| --- | --- |
| **ABBREVIATION** | **FULL - FORM** |
| CPI | Consumer Price Index |
| HES | Household Economic Survey |
| HPI | House Price Index |
| HLFS | Household Labour Force Survey |
| LVR | Loan-to-Value Ratio |
| QES | Quarterly Employment Survey |

## 1.3 Data Mining Objectives

### 1.3.1 Objectives

The goal is to cleanse raw data and analyse for trends and patterns to predict if alternative solutions are required for housing affordability.

**General Goals**

1. Collection of sufficient data from reliable sources.
2. Understand the data structure, accuracy, integrity and relationships.
3. Uncover trends and patterns to enable decision making.
4. Are average incomes higher than household expenditures (including housing expenses)?
5. Are the house prices in comparison to incomes and expenditures lower or higher?
6. Identify any outliers and extremes.

**Specific Goals**

1. Identify the saving capacity of young New Zealander’s using the income and expenditure datasets available according to regions.
2. Use savings data to determine if a person can purchase a house within 3 years.
3. Identify the historic residential sale trends across New Zealand regions.
4. Use historic sales data to determine lump sum down payment required to purchase a house in a specific region.
5. Draw comparison between capacity for down payment (savings for down payment) and cost of purchasing the house (down payment value).

### 1.3.2 Success Criteria

**General Goals**

1. Data quality is well understood.
2. Incremental achievement of objectives as well as the process is aligned with overall study objectives and is documented.
3. Documented methodology of data mining for all iterations.
4. Provide a conclusive result for the study objectives by using current information on incomes, expenditures, and the housing market.

**Specific Goals**

1. Annual savings identified per person according to year and region of New Zealand.
2. 3 years accumulated savings identified per person according to year and region of New Zealand.
3. Average house sale prices per year and per region identified.
4. 10 % of average house price calculated to determine the down payment required.
5. Visualise the gap between the capacity to save for deposit and down payment required in order to predict the future trend.

## 1.4 Project Plan

### 1.4.1 Risks and Resources

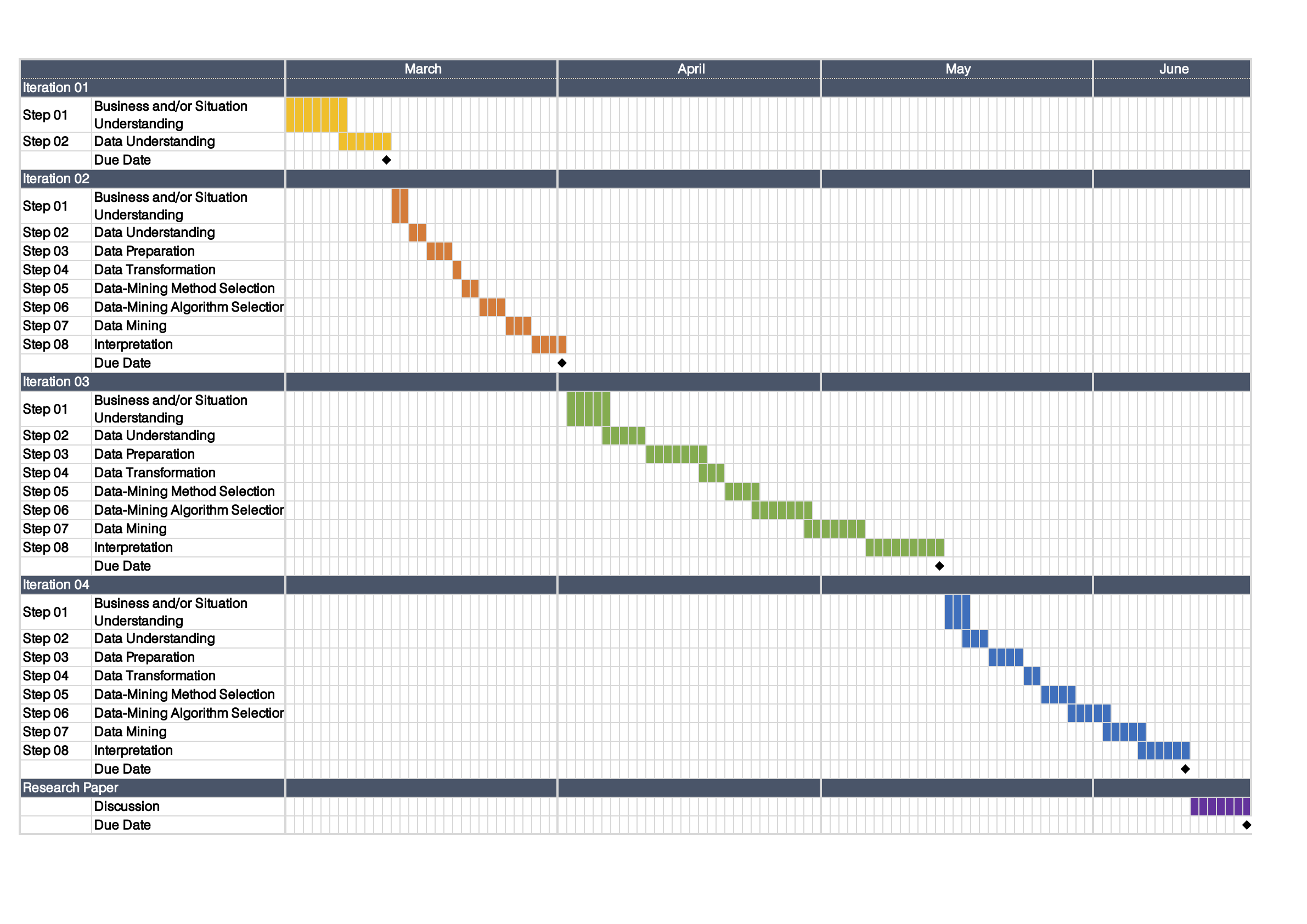
Table 3 - Risks and resources required for each of the data mining steps

|  |  |
| --- | --- |
| RISKS | RESOURCES AND ACTIONS |
| Step 01: Business and/ or Situation Understanding - 10% | |
| * Economic policies or fluctuations which result in drastic changes in expenditure, income, loan to value ratios, or house prices * Integration of background knowledge | * New data may be required to incorporate the changes. * Specific assumptions may be required prior to study to counter the effects of the changes so, clear contextual information is communicated. * Research must be conducted to identify the current solutions and factors contributing to the problem identified. Where information is lacking, it must be explicitly mentioned, to ensure correct interpretation of the problem, contextual knowledge and finally the results. |
| Step 02: Data Understanding - 10% | |
| * Technological issues, for example, unsuitable operating system * Data collection * Data quality * Data complexity * Data privacy | * Technical expertise or guidance may be required in order to resolve the technological issues from the University of Auckland. * New sources of data may need to be identified and used if required information is incomplete or missing. * Other data sources can be used to fill the missing information in the current dataset. * Further simplification and cleaning of data may be required in data preparation phase. * Addresses, locations, names or personal details must be removed during the data preparation phase to ensure privacy and security. |
| Step 03: Data Preparation - 15% | |
| * Technological issues * Problems with integrating data * Problems with aggregating data * Problems with exporting/ extracting data | * Technical expertise or guidance may be required in order to resolve the technological issues from the University of Auckland. * Other python libraries may be required such as pandas to integrate, aggregate, and export data. * Python libraries such as matplotlib may be required for data description and understanding the distribution. |
| Step 04: Data Transformation - 5% | |
| * Technological issues * Problems with visualisation of data | * Technical expertise or guidance may be required in order to resolve the technological issues from the University of Auckland. * Python libraries such as matplotlib may be required for data description and understanding the distribution. |
| Step 05: Data Mining Method Selection - 10% | |
| * Inability to find most suitable data mining method | * Further guidance may be required from tutors for selection of correct data mining method. * Further research must be conducted to identify the solutions and choose the correct method. |
| Step 06: Data Mining Algorithm Selection - 15% | |
| * Inability to find adequate models or algorithms | * Further research must be done in order to find the correct algorithm. Refer to Weka textbook for explanations. * Various different algorithm must be tried and tested in order to find the correct algorithm therefore, further exploratory analysis is required. * Models and parameters to be changed and tested to assess the performance of algorithms |
| Step 07: Data Mining - 15% | |
| * Technological issues * Issues with pattern identification | * Refer to PySpark documentation to resolve issues related to data mining software. * Technical expertise or guidance may be required in order to resolve the technological issues from the University of Auckland. * Further research on analysing and interpreting data mining results must be done to identify patterns. Guidance may be required from tutors. |
| Step 08: Interpretation - 20% | |
| * Economic changes * Generating visualisations | * Conclusions drawn must incorporate the economic changes that may affect the study results therefore correct interpretation of the data mining results is conveyed. * Use of Tableau or other softwares such as PowerPoint or Excel to generate more clear visualisations, if the visualisation generated from data mining software is insufficient in communicating the information. |

### 1.4.2 Project Plan

The following project plan is created based on the percentage of time allocation for each iteration. Given milestones are used to work backwards and allocate activities.

Table 4 - Project plan



# 2.0 Data Understanding

## 2.1 Collection of Initial Data

### 2.1.1 Data Collection, Formats and Potential Challenges

The datasets available for the study are mainly existing and collected via government surveys. They are all collected from the government data catalogue or the Reserve Bank of New Zealand website. The majority of the datasets are available in the excel spreadsheet format. The table below identifies the data sources, useful attributed to the datasets, possible common connections to merge data from multiple sets and sources, and potential issues.

Table 5 - Datasets, data sources, key attributes, and related issues

|  |  |  |  |
| --- | --- | --- | --- |
| DATASET & SOURCE LINKS | KEY ATTRIBUTES FOR ANALYSIS | COMMON CONNECITONS | POTENTIAL ISSUES/ ACTIONS |
| **Household Labour Force Survey**  [http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471#](http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471) | * Income * Year * Age Group * Sex * Location | * Regions * Year | * **Unnecessary Information** - removal of unimportant attributes. * **Consolidation** - data is also split by ethnicity; therefore, it requires consolidation. * **Survey Design** - the survey is focused on collecting the employed and unemployed figures rather than the income figures. Therefore, the focus of the survey is on employment numbers. * **Sampling** - high sampling errors associated with small estimates - this makes many of the smaller estimates unreliable or unusable. * **Survey Design** - HLFS does not measure the quality of people’s jobs, e.g., utilisation of skills, how much they are paid (except in June quarters), whether they get sick leave, etc. * **Survey Design** - issues regarding the definitions used in the HLFS (i.e., to be counted as employed you only have to have worked for one hour or more in a week, or you can even work unpaid in a family business. And to be unemployed you have to be available to start a job and be actively seeking work - not just looking at job advertisements). * **Survey Redesign** - in the June quarter of 2016 a redeveloped version of the HLFS went into the field. The key purpose of the 2016 HLFS redevelopment was to improve the relevance and quality of our labour market statistics. The new content includes more information about the nature of people’s employment conditions and work arrangements. This data can be used to better understand different patterns of employment. |
| **Household Economic Survey**  [http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471#](http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471) | * Average Housing Costs * Average Weekly Disposable Income * Household Composition | * Regions * Year | * **Unnecessary Information** - removal of unimportant attributes. * **Integration** - Linking the common attribute of the region together with the Census dataset. * **Sample Size** - It uses a sample size between 3,000 - 5,000 households each year but has since risen to 20,000 homes in 2019. Therefore, the values do not consider all households of New Zealand. * **Survey Design** - survey population excludes: * New Zealand usual residents temporarily overseas who don't return within the survey period * New Zealand usual residents temporarily staying elsewhere in New Zealand who don't return within the survey period * New Zealand usual residents who live in remote areas that are costly or difficult to access * people residing in non-permanent dwellings (i.e., in tents or caravans not permanently sited) * people residing at a wharf or landing place (i.e., people in ships, boats). |
| **Consumer Price Index**  <https://catalogue.data.govt.nz/dataset/consumers-price-index> | * Year * Index Change | * Year | * **Integration** - connecting to other datasets with CPI published more frequently than other surveys. * **No further issues found** |
| **Loan-to-Value Ratio**  <https://www.rbnz.govt.nz/statistics/c30> | * Year * Percent of Deposit Required | * Year | * **Format** - Not is a spreadsheet or readily available format, therefore information needs to be collected from the timeline provided on the Reverse Bank of New Zealand website. |
| **House Price Index**  <https://www.rbnz.govt.nz/statistics/m10>  M10 Housing Spreadsheet | * Year * Percent of Change | * Year | * **Missing Data** - overlaying with all other information. * **Region** - the data |
| **Residential Sales Statistics**  <https://www-library-auckland-ac-nz.ezproxy.auckland.ac.nz/eproducts/NZ_housedata/residential-property.htm> | * Street Address * Region * Date of Sale * Year of Sale * Sale Price * Land Value * Improvement Value * Building Age * Capital Value * Sale ID * QV ID | * Region * Year of Sale | * **Unnecessary Information** - Many attributes may not be required for the study; therefore, they need to be cleaned before mining. * **Volume** - the size of the data is significant, therefore, download, importing tasks are time consuming. |

Due to the number of multiple datasets required, many attributes within the datasets are not useful, such as ethnicity for the study, therefore they will not be considered in the final result.

### 2.1.2 Analysis and Prediction from Datasets

Although measuring housing affordability is a complex exercise, it is achievable with the datasets mentioned above. These datasets provide a clear picture of the income of New Zealand residents in a specific age group, their housing expenses, annual disposable income, and changes in house prices over time. These factors combined together can provide sufficient information to conclude if the younger generation of New Zealand is able to afford a house. There is more detailed microdata available, but due to the lengthy process of application for written authorisation, it is not feasible to consider for this study.

## 2.2 Description of Data

### 2.2.1 Data Quantity and Quality

The amount of data, value types, and coding schemes are described in the table below for each dataset. The larger datasets will produce better accuracy for the results.

Table 6 – Data sets’ quality and quantities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DATASET & FORMATS | AMOUNT OF DATA | KEY ATTRIBUTES | VALUE TYPES | CODING SCHEMES |
| **Household Labour Force** Survey/ Statistics NZ in (Excel Spreadsheet Format) | 3036 Rows  4 Columns | Year | **Numeric**  (Eight-byte, signed integer) | 1998-2020 |
| Region | **Categorical**  (ANSI/MBCS character string) | Northland, Auckland, Waikato, Bay of Plenty, Gisborne/ Hawkes Bay, Taranaki, Manawatu-Wanganui, Wellington, Nelson/ Tasman/ Marlborough/ West Coast, Canterbury, Otago, Southland, Total |
| Age Group | **Categorical**  (ANSI/MBCS character string) | 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+ |
| Value | **Numeric**  (Double-precision floating-point number) | NA |
| **Text  Description automatically generatedincome\_year** = defines the timeline of the data collected. The survey provides the data from 1998 to 2020.  Figure 2 - Income data description  **income\_regions** = provides the regional zones from which the data was collected. This gives us a regional understanding as well as national understanding of income data.  **income\_age\_groups** = provides the information on the income values according to age groups. This allows us to isolate specific age groups required for the study.  **income\_value** = defines the income value associated with the above attributes. | | | | |
| **Household Economic Survey**/ Statistics NZ in (Excel Spreadsheet Format) | 350 Rows  4 Columns | Year | **Numeric**  (Eight-byte, signed integer) | 2007-2019 |
| Expenditure Category | **Categorical**  (ANSI/MBCS character string) | Total, Food & Drinks with Related Sub-Groups, Clothing & Footwear, Housing (Rent or Own), Property Expenses, Property Taxes, Household Bills with Related Sub-Groups, Medical Expenses, Transport Expenses, Personal Expenses, Contribution to Savings, Insurance, Education, Other Subgroups |
| Region | **Categorical**  (ANSI/MBCS character string) | Auckland, Wellington, Rest of North Island, Canterbury, Rest of South Island |
| Values | **Numeric**  (Double-precision floating-point number) | NA |
| **Text  Description automatically generatedexpenditure\_year** = defines the years in which data was collected since 2007.  Figure 3 - Expenditure data description  **expenditure\_regions** = defines the board regions of New Zealand where the households are located.  **expenditure\_category** = defines the categories associated with typical expenditure of a New Zealand household. It includes the total value, which allows us to filter the expenditure values rather than calculating the total.  **expenditure\_value** = defines the expenditure in New Zealand dollars of households. | | | | |
| **Consumer Price Index**  (Excel Spreadsheet Format) | 33 Rows  6 Columns | Year | **DateTime** | 1988-2020 |
| Index Value | **Numeric**  (Double-precision floating-point number) | NA |
| **Text  Description automatically generatedcpi\_series\_id** = defines the recording of the index as a datetime object.  Figure 4 - Consumer price index data description  **cpi\_year** = the year the index was recorded to show the year-on-year changes.  **cpi\_quarter\_01**, **cpi\_quarter\_02**, **cpi\_quarter\_03** and **cpi\_quarter\_04** = the consumer price index values for each quarter of every year. These may have to be averaged for merging datasets as other data is recorded yearly. | | | | |
| **House Price Index**  (Excel Spreadsheet Format) | 21 Rows  6 Columns | Year | **DateTime** | 2000-2020 |
| House Price Index | **Numeric**  (Double-precision floating-point number) | NA |
| Text  Description automatically generated**hpi\_series\_id** = defines the recording of the index as a datetime object.  Figure 5 - House price index data description  **hpi\_year** = the year the index was recorded to show the year-on-year changes.  **hpi\_quarter\_01**, **hpi\_quarter\_02**, **hpi\_quarter\_03** and **hpi\_quarter\_04** = the house price index values for each quarter of every year. These may have to be averaged similar to CPI dataset for merging datasets as other data is recorded yearly. | | | | |
| **Residential Sales Statistics**  (Excel Spreadsheet Format)  Note: This data source is huge and contains large amounts of information not relevant to study; therefore, only selected key attributes are mentioned. The other attributes such as building age, city, sale ID and many more are required to be cleaned. | 171,279 Rows  13 Columns | Sale Date | **Date** | 1990-2019 |
| Sale Price | **Numeric**  (Eight-byte, signed integer) | NA |
| Region | **Categorical**  (ANSI/MBCS character string) | Auckland (Unitary), Bay of Plenty Region, Canterbury Region, Gisborne (Unitary), Hawkes Bay Region, Manawatu-Whanganui Region, Northland Region, Otago Region, Southland Region, Taranaki Region, Tasman Nelson Marlborough, Waikato Region, Wellington Region, West Coast Region |
| The above image displays the columns within the dataset. Attributes such as sales\_qv\_id, sales\_id, sales\_city, and more are not required therefore, have to be deleted within the data preparation phase. The key attributes are listed below.  Figure 6 - Sales data description  **sales\_region** = the location of the property sold.  **sale\_year** = reference to the year the particular house was sold.  **gross\_sale\_price** = the gross sale price is used as opposed to net sale price as it includes all expenses of purchasing a house such as chattels, GST or tax. | | | | |

Text

Description automatically generatedThe majority of the value types within the datasets are either numeric or categorical. Most of the key attributes can be used, prioritised, and analysed to conclude the study objectives. With specific data clean-up steps and additional data preparation, it is possible to merge these datasets. Below is the screenshot for python commands used for data description.

Figure 7 - Python - Data description

## 2.3 Exploration of Data

There can be some data exploration carried out on the individual datasets. The datasets do need to be processed prior to integration with other data to provide more reliable results. Below is the summary of the findings from individual datasets.

### 2.3.1 Household Labour Force Survey

**Key Findings**

* The data is split according to regions in New Zealand as well as offers the total incomes in all regions allowing the analysis of housing affordability across New Zealand and specific regions.
* The regions containing the main city centres have higher income levels than rural regions of New Zealand.
* The income figures are also recorded in specific age groups; therefore, data of the 20-35-year-old working population can be isolated and studied separately to the remainder.
* Income figures are also available in average as well as median figures for specific regions, allowing multiple interpretations.
* Income figures are also separated according to gender allowing further study to understand if housing affordability differs according to gender.

Chart, treemap chart

Description automatically generated

Figure 8 - Income values by regions

Chart, bubble chart

Description automatically generated

Figure 9 - Income values by age groups

Chart, scatter chart

Description automatically generated

Figure 10 - Weekly income level changes by year

### 2.3.2 Household Economic Survey

**Key Findings**

* Certain data under the expenditure category, such as illicit drugs, are “null”. According to the data source, these figures are suppressed due to their sensitivity.
* Average expenses related to renting and owning are included, reducing the need to rely on data from the consumer price index dataset.
* Expenditure categories are extremely detailed and filtration of these can provide detailed information and customisation for analysis. For instance, patterns of expenditure within a group such as food, can be identified and compared with another group such as transport to compare the changes over the past few years.
* The graph below shows a clear trend in the average weekly expenditure of New Zealanders over the years. Year on x-axis and amount on the y-axis. This trend contributes to the overall hypothesis of the study and highlights the fact that increasing the cost of living is one of the causes of housing unaffordability.

A picture containing table

Description automatically generated

Figure 11 - Expenditure data by categories

Chart, treemap chart

Description automatically generated

Figure 12 - Expenditure data by regions

Shape, rectangle

Description automatically generated

Figure 13 - Expenditure data by year

### 2.3.3 Consumer Price Index

**Key Findings**

* The consumer price index is measured quarterly and shows the changes in the price of goods and services typically purchased by New Zealand households.
* Regular and frequent collection of the data means that the chances of errors and gaps are reduced significantly.
* The index figures are split into groups and sub-groups such as health group is sub categorised with indexes on medical products, appliances, and equipment, hospital care, and out-patient services. Although it is not directly relevant to this study, information can be gained to understand which expenditure group contributes to housing unaffordability.
* Consumer price index information is also split across the major regions of New Zealand, providing a more accurate understanding of price changes.
* Separate consumer price index data is available with seasonal adjustments which allow for the market fluctuations to be considered.
* In comparison to the Household Economic Survey, the consumer price index shows the percentage of change over the years in the typical expenditure of New Zealand households. On the other hand, the Household Economic Survey provides changes in dollar figures. The data-mining exercise must consider both and assess the results to determine, which provides a more accurate understanding. Using both datasets individually in specific iterations and together is required to further analyse the information. It can also serve as a reference data for a deeper interpretation of the data mining results.

Chart

Description automatically generated

Figure 14 - Consumer Price Index change by year

### 2.3.4 House Price Index

**Key Findings**

* The house price index measures the changes in house prices across New Zealand. This is used to understand the valuation of the housing stock, which can then be used against household expenditures to understand affordability.

Chart

Description automatically generated

Figure 15 - House price index changes per year

### 2.3.5 Residential Sales Statistics

**Key Findings**

* The dataset includes the sales data of houses per region. This requires some consolidation as the data is recorded for each property, therefore either the sales data should either be used as average or median.
* There are numerous attributes associated with each sale record. Although not relevant for this specific study; many avenues of research can be sourced from this dataset. Correlations between the sale price and regions, sale price and floor areas, sale prices and capital values and many more factors can be studied.
* Data consolidation and clean-up is required prior to understanding the clear relationship between sale date and prices.

Chart, treemap chart

Description automatically generated

Figure 16 - Average house sale prices per region

Chart, bar chart

Description automatically generated

Figure 17 - Average sale price changes over years and regions

## 2.4 Verification of Data Quality

### 2.4.1 Household Labour Force Survey

Table 7 - Verification of data quality - Household Labour Force Survey

|  |  |
| --- | --- |
| **Missing Data** | Certain income values are missing from the dataset as shown on the count value in the screenshot below. These need to be dealt with during the data cleansing stage. |
| **Data Errors** | No data errors observed. |
| **Measurement Errors** | Values are correct but inconsistent capitalisation is used across the dataset. |
| **Coding Inconsistencies** | No coding inconsistencies observed. |
| **Bad Metadata** | Metadata consistent with the key attributes required. Additional metadata to be excluded - in form of an additional spreadsheet. |

Text

Description automatically generated

Figure 18 - Null values in income data

A picture containing text, scoreboard, plaque

Description automatically generated

Figure 19 - describe () function on income values

### 2.4.2 Household Economic Survey

Table 8 - Verification of data quality - Household Economic Survey

|  |  |
| --- | --- |
| **Missing Data** | No missing data |
| **Data Errors** | 17 outliers observed in the expenditure value attribute and need to be investigated. |
| **Measurement Errors** | The dataset includes total weekly expenditure as well as individual categories in the same attributes. These need to be separated for correct calculations. |
| **Coding Inconsistencies** | No coding inconsistencies found. |
| **Bad Metadata** | No bad metadata found. |

Text

Description automatically generated with medium confidenceText

Description automatically generated

Figure 20 - Null values in expenditure data

Figure 21 - describe () function on expenditure data

### 2.4.3 Consumer Price Index Data

As a key economic performance indicator for New Zealand, this dataset is regularly used by multiple agencies, therefore it has very good data quality.

Table 9 - Verification of data quality - Consumer Price Index Data

|  |  |
| --- | --- |
| **Missing Data** | No missing data found. |
| **Data Errors** | No data errors found. |
| **Measurement Errors** | No data measurement found. |
| **Coding Inconsistencies** | No coding inconsistencies observed. |
| **Bad Metadata** | Additional metadata to be excluded - in form of an additional spreadsheet. Spreadsheet on Gross Domestic Product and related data is not required for the study. |

Text

Description automatically generated with medium confidenceText

Description automatically generated

Figure 22 - describe () function on consumer price index data

Figure 23 - Null values in consumer price index data

### 2.4.4 House Price Index Data

Table 10 - Verification of data quality - House Price Index Data

|  |  |
| --- | --- |
| **Missing Data** | The House Price index value for the final quarter of 2020 is missing.  No data available prior to 2000, whereas consumer price index data is available from 1988. |
| **Data Errors** | No data errors found. |
| **Measurement Errors** | No data measurement found. |
| **Coding Inconsistencies** | No coding inconsistencies observed. |
| **Bad Metadata** | No bad metadata observed. |

Text

Description automatically generated

Figure 24 - Null values in house price index data

Text

Description automatically generated with medium confidence

Figure 25 - describe () function on house price index data

### 2.4.5 Residential Sales Data

Table 11 - Verification of data quality - Residential Sales Data

|  |  |
| --- | --- |
| **Missing Data** | No missing data. |
| **Data Errors** | Outliers and extremes found in many attributes and need to be addressed. |
| **Measurement Errors** | Houses with indeterminate age have been noted as “mixed”. This affects the completeness of the data and may need to be discarded. Although, this is not a key attribute for the study. |
| **Coding Inconsistencies** | Regions of each sale have an associated ID. Inconsistent coding has been observed when data sample was obtained with python sample commands. Therefore, the region IDs should be corrected or should not be used. |
| **Bad Metadata** | Additional metadata to be excluded - in form of an additional spreadsheet containing the data dictionary. |

Text

Description automatically generated

Text

Description automatically generated

Figure 27 - describe () function on residential sales data

Figure 26 - Null values on residential sales data

Text

Description automatically generated

Figure 28 - Python commands for data verification

# 3.0 Data Preparation

## 3.1 Data Selection

Five datasets have been selected from the above-mentioned list. These have been added to the python file and filtered as required.

Text

Description automatically generated

Figure 29 - Data importing, reading, verification and description in python

Below table summarises the key attributes included and excluded in the selection process and the reasoning

Table 12 - Data selection

|  |  |  |  |
| --- | --- | --- | --- |
| DATASETS | KEY ATTRIBUTES TO BE USED | EXCLUDED OR FILTERED ATTRIBUTES | COMMENTS |
| **Income Data** | income\_regions  income\_age\_groups  income\_year  income\_value | income\_age\_groups | The objective of the study is to assess housing affordability for the younger generations. Therefore, only three age groups are selected, and the remainder are excluded.   * 20 to 24 * 25 to 29 * 30 to 34 |
| Text  Description automatically generated  Figure 30 - Condition for filtering age groups in the income data | | A picture containing table  Description automatically generated  Figure 31 - Sample results of filter condition on age groups | |
| **Expenditure Data** | expenditure\_regions  expenditure\_year  expenditure\_value | expenditure\_category | The expenditure category includes totals as well as the individual expense categories. For the purposes of this study, only the typical expense of a household is required. Hence, the python command was used to filter total expense records and then delete the attribute. |
| Text  Description automatically generated  Figure 32 - Condition for filtering categories in expenditure data | | A picture containing table  Description automatically generated  Figure 33 - Random sample results of filter condition on expenditure categories | |
| **House Sales Data** | sale\_region  sale\_year  gross\_sale \_price | sales\_qv\_id  sales\_id  sales\_city  sale\_region\_id  sale\_region  sale\_date  sale\_year  net\_sale\_price  gross\_sale \_price  capital\_value  land\_value  improvement\_value  house\_age | Sale QV and Sale IDs are removed as they do not offer any specific link to other datasets and are not useful as we do not have to identify individual records.  The objective of the study is to assess the regions, not individual cities, hence the city attribute is removed.  Sale region ID is removed as inconsistencies were found, also the region column provides more detailed information than the region ID. Therefore, to avoid duplication, the attribute is removed.  Sale date and year column provides duplicate information. As the study is focused on changes occurring year by year, the day and month from the datetime field is not required. Therefore, the sale date is removed.  Net sale price does not include the cost of GST and chattels paid while house purchase occurred, whereas the gross sale price considers all aspects of purchase. Hence net sale price is removed.  Capital, land and improvement values although important are not relevant to this study as the objective is to look for historic sales data. Therefore, these three fields are removed. These can be helpful in understanding the changes across the years, but the easier and more accessible information is provided from the house price index dataset, which takes these factors into account.  House age is also not required in the sales analysis therefore it is excluded. |
| Text  Description automatically generated  Figure 34 - Python statement to eliminate unnecessary attributes | | Text  Description automatically generated  Figure 35 - Remaining attributes and the data types | |
| **House Price Index Data** | hpi\_year  hpi\_quarter\_01  hpi\_quarter\_02  hpi\_quarter\_03  hpi\_quarter\_04 | series\_id | The series ID contains a specific date for recording the house price index for the quarter. Since the study is focused on changes occurring year-by-year, the series ID creates duplication with the HPI year attribute. Therefore, it is excluded. |
| Text  Description automatically generated  Figure 36 - Python statement to remove unnecessary attributes | | Text  Description automatically generated  Figure 37 - Remaining attributes and the data types | |
| **Consumer Price Index Data** | cpi\_year  cpi\_quarter\_01  cpi\_quarter\_02  cpi\_quarter\_03  cpi\_quarter\_04 | series\_id | Similar to above, the series ID contains a specific date for recording the consumer price index for the quarter. Since the study is focused on changes occurring year-by-year, the series ID creates duplication with the CPI year attribute. Therefore, it is excluded. |
| Text  Description automatically generated  Figure 38 - Python statement to remove unnecessary attributes | | Text  Description automatically generated  Figure 39 - Remaining attributes and the data types | |

## 3.2 Data Cleaning

Outliers, extremes, blank, and null values are removed or treated using logical process generated from the data audit process.

Table 13 - Data cleaning

|  |  |
| --- | --- |
| DATASETS | COMMENTS |
| **Income Data** | The attribute income\_value has null values. These null values have been removed |
| Text  Description automatically generated  Figure 40 - Python statement to check and remove null values  Text  Description automatically generated  Figure 41 - Results before and after removing null values |
| **Expenditure Data** | Some outliers to occur within the dataset. The outliers fall within a specific expenditure category called “exchanges, returns or trade-ins” if they are negative. Although this category increases a household’s disposable income as it is money coming back in, taking scale into account and the direct impact on expenditure, these are relevant, therefore are to remain. The outliers that are positive (above $1027) are relevant and remain in the dataset as they strongly correlate to the high expenditure in some of the major cities in New Zealand. Therefore, removal of these can have an impact on the data mining results by completely removing the Wellington and Auckland regions. |
| Figure 42 - Python statement for minimum and maximum values  Text  Description automatically generated  Figure 43 - Results for quantile, minimum and maximum values |
| **Sales Data** | Majority of the outliers, extremes, and null values appear in attributes that can be removed through the selection process. Gross sale price attribute, sale region and sale year attributes do not contain any outliers therefore, no action is required.  The issue with the measurement errors regarding house age as mentioned in the data quality verification step is no longer relevant as the field is not important for the study and has been removed in the data cleaning step. |
| Text  Description automatically generated  Figure 44 - Python statements for outlier detection in sales data  Text  Description automatically generated with low confidence  Figure 45 - Results for outliers in sales data |
| **House Price Index Data** | The house price index has one null value for the final quarter of 2020. This means that this value must be added to give an accurate result for the year 2020. Since the majority of the data is present for 2020, and the data between the previous years is drastically different, using the mean or random value is not appropriate. The previous quarters value is used to replace the null value. |
| Text  Description automatically generated  Figure 46 - Python statement for replacing null data in house price index  A picture containing text  Description automatically generated  Figure 47 - Results for replacing null data in house price index |
| **Consumer Price Index Data** | There are no outliers, extremes, blank or null values in the consumer price index dataset. Therefore, it does not require any corrective action. As mentioned above, this is a key economic indicator to assess the country’s performance, used by many agencies, therefore the data obtained is valid and carefully aggregated already. It shows a steady and gradual increase over the years. |
| Figure 48 - Python statement for detection of null values in consumer price index  Text  Description automatically generated  Figure 49 - Results for null values in consumer price index  Text  Description automatically generated  Figure 50 - Python statement for outlier detection |

## 3.3 Data Construction

New attributes have been created using python to either facilitate integration of the data (explained in next section) or to aggregate the existing data as explained in the table below.

Table 14 - Data construction

|  |  |
| --- | --- |
| DATASETS | COMMENTS |
| **Income Data - Age Group** | Cleaned and filtered, there are now only three age groups. In order to differentiate them easily, unique code has been added to assist further analysis and readability. |
| Text  Description automatically generated  Figure 51 - Python code for assigning age group code  A picture containing diagram  Description automatically generated  Figure 52 - Results for assigning age group code |
| **Income Data - Regions** | There are many smaller regions included in this dataset in comparison to the expenditure dataset. In order to integrate this data according to regions, they need to have a common attribute. Therefore, assigning a separate region code that can be associated with the income dataset as well as expenditure dataset was required using if else conditions as shown below. |
| Text  Description automatically generated  Figure 53 - Python code for assigning region codes in income data  Text  Description automatically generated  Figure 54 - Results for assigning region codes for income data |
| **Expenditure Data** | Similar to income data, to integrate the expenditure data according to regions with the income data, a common attribute is required. Therefore, the same region codes are used. This is required to meet the second objective of the study (assessing housing affordability in regions). |
| Text  Description automatically generated  Figure 55 - Python code for assigning region code for expenditure data  Text  Description automatically generated  Figure 56 - Results for assigning region code for expenditure data |
| **Sales Data - Regions** | To integrate historic sales data with income and expenditure common region code is required. The regions within the residential sales dataset are much finer grained when compared to income or expenditure dataset. |
| Text  Description automatically generated  Figure 57 - Python statement for assigning region codes  Text  Description automatically generated  Figure 58 - Sample results of assigning region codes to sales data |
| **House Price Index** | All house price index data was obtained for each quarter of the year. Since we are analysing all data by year, this data needed to be aggregated and averaged. Therefore, another attribute was created to calculate the mean value for a year. |
| Figure 59 - Python code for calculating annual house price index  A picture containing text, scoreboard  Description automatically generated  Figure 60 - Results after calculating annual house price index |
| **Consumer Price Index** | All consumer price index data was obtained for each quarter of the year. Similar to the house price index, we are analysing all data by year, this data needed to be aggregated and averaged. Therefore, another attribute was created to calculate the mean value for a year. |
| Figure 61 - Python code for calculating the annual consumer price index  A picture containing text, electronics  Description automatically generated  Figure 62 - Results after calculating annual consumer price index |

Additionally, further attributes are created as required after the data integration using the python. Listed below. Refer to data construction stream for location within the flow of data.

Table 15 - Additional fields

|  |  |
| --- | --- |
| ADDITIONAL ATTRIBUTES | COMMENTS |
| **income\_value\_mean** | Average income value to obtain a single figure for per year, per age group, and per region. |
| Figure 63 - Python code for calculating the average income  A picture containing table  Description automatically generated  Figure 64 - Results after averaging the income figures |
| **expenditure\_per\_person** | The data available on expenditure is based on households not individuals, therefore, to obtain individual expenditure, the figure is divided by 3. According to Statistics New Zealand, the average New Zealand household is made up of three people. |
| Figure 65 - Python code for calculating expenditure per person  A picture containing graphical user interface  Description automatically generated  Figure 66 - Sample results for calculating expenditure per person |
| **gross\_sale\_price\_mean** | Average of historic sale price per year, and per region. |
| Figure 67 - Python code for calculating average sale prices across regions and by year  A picture containing graphical user interface  Description automatically generated  Figure 68 – Sample results from calculating average sale prices |
| **annual\_savings** | Annual income subtracted from annual expenditure. |
| Figure 69 - Python code for calculating annual savings  A picture containing table  Description automatically generated  Figure 70 - Random samples of results from annual savings calculations |
| **downpayment\_capacity** | It is assumed that a typical down payment required to purchase a house is equal to approximately five years of savings. |
| Figure 71 - Python code for calculating down payment capacity  Text  Description automatically generated with low confidence  Figure 72 - Sample of results from down payment calculations |
| **loan\_to\_valuation\_ratio** | It is assumed that this is 10 % of the purchase/ sale price of the house for a first home buyer. |
| Figure 73 - Python code for calculating loan to value ratio  Graphical user interface, text  Description automatically generated  Figure 74 - Sample of results for loan to value ratio calculations |
| **affordability** | The categories below show the levels of savings required to pay for a deposit based on house sales price data. The calculation is based on down payment capacity (3 years of savings) and loan-to-value ratio (10% of house sale price) to determine is the person is able to afford the deposit.   * 0 – 20 = Greater or at least 20% of savings towards the deposit * 21 – 40 = Greater than 20% but less than 40% of savings required for a deposit. * 41 – 60 = Greater than 40% but less than 60% of savings required for a deposit. * 61 – 80 = Greater than 60% but less than 80% of savings required for a deposit. * 81 – 100 = Greater than 80% but less than 100% of savings required for a deposit. * 100+ = Greater than 100% savings compared to the amount required for a deposit. |
| Text  Description automatically generated  Figure 75 - Python code for assigning affordability category  A picture containing graphical user interface  Description automatically generated  Figure 76 - Sample of results for affordability attribute |

## 3.4 DataSource Integration

The datasets used in this study originate from five different sources. In order to perform data mining exercises, these need to be integrated. During the data exploration phase, region and year were identified as the common attributes. The previous phase consolidated the various different regions by region code. This next step is to merge the datasets.

Table 16 - Data integration

|  |  |
| --- | --- |
| MERGED DATASETS | COMMENTS |
| **Income and Expenditure Data** | The income and expenditure is merged using the attributes year and the region codes. This results in the table which contains records on common regions, years, income values, and expenditure values. |
| Graphical user interface, text  Description automatically generated  Figure 77 - Python code for merging income and expenditure data  A picture containing graphical user interface  Description automatically generated  Figure 78 - Results for merging income and expenditure data |
| **Prior to Merging Sales Data** | Combining income, expenditure, and sales data would create data duplication. There are over 150,000 records for the sales data and when merged, the income and expenditure values would be duplicated multiple times for the same year and region but different sales data. To prevent this and achieve a cleaner dataset sales data requires aggregation. Python code is used to derive the mean gross sale price for each year and each region. |
| Figure 79 - Aggregate sales data  A picture containing graphical user interface  Description automatically generated  Figure 80 - Results from aggregating the data |
| **Income, Expenditure and Sales Data** | Sales data can now be merged with the previously integrated income and expenditure data. Same region codes are added in the previous steps of data construction. The table generated includes records on regions, years, income values, expenditure values, and sales figures. |
| Text  Description automatically generated  Figure 81 - Python code for merging income, expenditure and sales data    Graphical user interface, text  Description automatically generated with medium confidence  Figure 82 - Results from merging income, expenditure and sales data |
| **House Price Index and Consumer Price Index Data** | The house price index and consumer price index are reference datasets. These can be merged by year to provide a table containing the common years, mean house price index value, and mean consumer price index value. |
| Figure 83 - Python code for merging house and consumer price index  Graphical user interface  Description automatically generated with medium confidence  Figure 84 - Results from merging house and consumer price index |
| **Income, Expenditure, Sales, House Price Index and Consumer Price Index** | All datasets can now be merged on the year attributes to create a table with region, year, income values, expenditure values, mean house price index, and mean consumer price index. |
| Text  Description automatically generated  Figure 85 - Python code for merging all datasets  Table  Description automatically generated with low confidence  Figure 86 - Results from merging all datasets |

## 3.5 Data Formatting

Python code was used to clean up the dataset for easier readability. Sort order of fields is changed to ascending order and fields are reorganised and renamed to begin with year and regions. Fields such as income, expenditure or gross sale price values along with duplicated columns are removed as they have been aggregated into other fields, therefore no longer required.



Figure 87 - Drop unnecessary columns

Text

Description automatically generated

Figure 88 - Sort all columns

Graphical user interface, text

Description automatically generated

Figure 89 - Renaming all columns

Text

Description automatically generated

Figure 90 - Results from sorting and reformatting the data

# 4.0 Data Transformation

## 4.1 Data Reduction

Following data cleaning and integration, all fields were analysed using distribution graphs. The distribution graph was used to analyse categorical data and histogram graph was used to analyse continuous data. The screenshot below summaries all fields involved. The distribution of year, region, and age groups is even across the respective categories. This is as a result of consolidation, aggregation, and integration of data in the data preparation phase. Therefore, no action is required.

Table 17 - Distribution Graphs for Year, Region and Age Groups

|  |  |
| --- | --- |
| KEY ATTRIBUTES | DISTRIBUTIONS |
| **Year** | Graphical user interface  Description automatically generated  Figure 91 - Data distribution by year |
| **Region** | A screenshot of a computer  Description automatically generated with medium confidence  Figure 92 - Data distribution by region |
| **Age Group** | A screenshot of a computer  Description automatically generated with medium confidence  Figure 93 - Data distribution by age groups |

## 4.2 Data Projection

Using the balancing or boosting the data to remedy the skewed data would be an extreme measure for fields with continuous data. The continuous data is relatively well distributed as shown in the histograms below.

Table 18 - Data projection

|  |  |
| --- | --- |
| KEY ATTRIBUTES | DISTRIBUTION |
| **annual\_savings** | Chart, histogram  Description automatically generated  Figure 94 - Data distribution for annual savings |
| **downpayment\_capacity** | Chart, histogram  Description automatically generated  Figure 95 - Data distribution for down payment capacity  Since the annual savings and down payment capacity is calculated from the same income and expenditure datasets, their results are similar. They are both evenly distributed as shown by the resemblance of the bell-shape on the histogram. |
| **loan\_to\_value\_ratio** | Chart, histogram  Description automatically generated  Figure 96 - Data distribution for loan to value ratio  The data in the top image is positively skewed, meaning there are a large number of occurrences where the loan to value ratio is lower. Additionally, the lower values indicate the more common residential sale prices and hence the loan to value ratio. More expensive houses are less frequently purchased in comparison to the typical kiwi home. |
| **average\_sale\_price** | Chart, histogram  Description automatically generated  Figure 97 - Data distribution of average residential sale prices  The average sale price and loan to valuation ratio are calculated from the same data therefore resulting in similar distribution. |
| **hpi\_value** | Chart, box and whisker chart  Description automatically generated  Figure 98 - Data distribution for house price index |
| **cpi\_value** | Chart, histogram, box and whisker chart  Description automatically generated  Figure 99 - Data distribution for consumer price index |

# 5.0 Data Mining Method Selection

## 5.1 Discussion on Data Mining Objectives and Methods

### 5.1.1 Classification

Classification method belongs to supervised learning technique in which labelled data is used to make predictions in a non-continuous form. Classification algorithms used in machine learning use the input training data for the purpose of predicting the likelihood or probability that the new data will fall into one of the predetermined categories. For example, determining the likelihood of heart disease. This is a binary classification since there are only two classes, one with heart disease and the other without. The algorithm requires training data to understand how the input data is related to these two classes, and once the model is trained with higher level of accuracy, it can be used to detect whether a particular patient has heart disease or not.

Classification is the most suitable data mining method for this study as the objective is to predict if the younger generation of New Zealanders are able to afford a home within three years. Data such as the regions, age groups, expenditure, income, historical sale prices, annual savings, loan to value ratio, and down payment capacity are independent variables that can help determine the dependent variable which are the affordability classes. In this method, the output variable (affordability) is split into multiple classes and is dependent on the various different independent variables, which are either categorical or numeric. Similar to the above-mentioned example of binary classes (heart disease or no heart disease), the affordability classes assist in understanding the data in a summarized and categorical manner.

### 5.1.2 Regression

Regression also falls under the supervised learning technique in which labelled data is used to make predictions in a continuous form. It is used to model the relationship between a dependent variable (target) and independent variables (predictors) with one or more independent variables. Regression analysis helps in understanding how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. Unlike classification, regression is used to output continuous numbers. A typical example of this data mining method is predicting the house prices based on input parameters such as distance to city centre, section size, house size, or number of bedrooms. A regression data mining method is used for predicting the quantity compared to classification, which includes predicting the class label.

Although a very useful method, this study requires a categorical target variable to classify affordability levels. These classifications of affordability provide more aggregated information on the situation compared to numeric data, thus giving a clear indication for government intervention, which, in turn is the study objective. A continuous or numeric target can also be used to meet the business objectives, but it would require further aggregation of data to determine if government intervention is required. For example, it is much easier and clear to conclude that a person requires support to pay for a deposit for house purchase when they fall in the 60-80% category. Compared to this, a numeric value which may refer to the person having $75,850 for a deposit of $100,000 can be difficult to interpret. The numeric target may identify an overall trend such as the house prices are increasing therefore, more government intervention is required but, when this scenario is scaled up, it is very difficult to arrive to a clear conclusion from a numeric target regarding the level of government intervention required. For this reason, regression data mining method will not be suitable for the study.

### 5.1.3 Clustering

Clustering method falls under un-supervised learning technique that is used to identify patterns within the datasets. The un-supervised learning technique is used to draw references from datasets consisting of input data without associated labels. Typically used as a process to identify meaningful structure, explanatory underlying processes, generative features, and groupings present within the data. Clustering itself is the task of dividing the data into a number of groups such that the data points within the same group are more similar compared to the data points within other groups. Therefore, it determines the intrinsic grouping among the un-labelled data present. The aim is to segregate groups with similar traits and assign them into clusters. Clustering is often used for market segmentation, social network analysis, search result grouping, medical imaging, image segmentation, anomaly detection, and many more applications in a variety of industries.

While clustering is an important data mining method, it is not suitable for this study. Clustering data mining method falls under the un-supervised learning technique; therefore, it requires un-labelled data. For example, if an image containing dogs and cats is presented as an input to the machine learning model without labelling it as an image of dogs and cats. The machine itself analyses the similarities, differences and the patterns between the animals to group them separately without our input. It allows the model to work on its own to discover patterns and information that was previously undetected. Whereas the data involved in this study is already labelled. For example, the income data is already grouped by regions and age groups or the overall labels given to the datasets such as expenditure, income, savings etc. Since, the data is already labelled, the clustering data mining method unsuitable for the study.

## 5.2 Data Mining Method Selection

Classification is selected as the data mining method for this study for the reasons listed below.

1. Dependent variable (target) has been identified and labelled.
2. Independent variables (predictors) have been identified and labelled.
3. All data is labelled therefore, only methods associated with supervised learning technique are relevant, thus narrowing the choice to regression and classification (more in-depth explanation given under 5.1.3 Clustering).
4. In the classification method, there is a requirement to quantify the impact of variables (also referred to as independent or X variables) on the dependent variables (Y variable). However, in classification the Y variable is categorical. Hence, the model is trained on X or independent variables and predicts observations into predefined classes. The independent variables region, age, and year are categorical, whereas others are numeric.
5. The data mining goals, and the business objectives require the output to be in a categorical form. Therefore, there is a need to quantify the impact of numerous variables on a categorical entity (also referred to as the dependent of Y variable). Thus, making classification the most suitable data mining method.
6. The requirement is also to determine the level of intervention required, such as a person with 20-40% saving for a deposit requires more support than a person with 80-100% savings. Therefore, the use of categorical target is critical to achieving the business objectives.

# 6.0 Data Mining Algorithm Selection

## 6.1 Exploratory Analysis and Discussion

### 6.1.1 Discussion on Algorithm Theory

Weka data mining software contains over 76 algorithms for classification and regression data mining methods alone. The classifiers within Weka are designed to be trained for predicting a single attribute, typically the target. Depending on the data, some classifiers can only learn from nominal attributes, others can only learn from numeric attributes (primarily for regression method), while the remaining from nominal and numeric. Out of the 76 classifiers available, some disabled once the data is loaded. This is due to their ability or limitations of predicting a categorial or numeric target or their limitations of analysing categorical or numeric independent (input or Y) variables (Ian Whitten, 2016).

The table below describes the breakdown of the relevant algorithms related to the classification data mining method from Weka. To understand their suitability with the data mining objectives, the advantages and disadvantages are studied along with their key definition to assist in decision making.

Table 19 - Weka Algorithms

|  |  |
| --- | --- |
| MODELS | COMMENTS |
| Bayes Classifiers (Priyankur, 2019) | Naïve Bayes techniques are a set of supervised learning algorithms based on applying Bayes’ Theorem with the “naïve” assumption that all data is conditionally independent. For example, in a dataset it might appear that there is a correlation between the occurrences of income and age. However, if it can be assumed that income and age are actually mutually independent and the correlation can be attributed to the existence of an external factor, such as education, then we can apply Naive Bayes.  **Advantages:**   * Easily scalable for larger datasets. * Works quickly and uses less time. * Requires less training data compared to other algorithms. * Better suited for categorical input variables than numerical.   **Disadvantages:**   * Zero probability problem – when the conditional probability is zero for a particular attribute, it fails to provide a valid prediction. * The assumption that all predictors are independent is rare in real life, therefore limiting the use of this algorithm in real-life cases.   Versions of Bayes Classifiers explained below. |
| Bayes Net | Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier (Ian Whitten, 2016). |
| Naïve Bayes | Class for a Naive Bayes classifier using estimator classes (Ian Whitten, 2016). |
| Naïve Bayes Multinominal Text | Multinominal Naïve Bayes implements naïve Bayes algorithm for multinominal distributed data. Multinomial naive bayes for text data (Ian Whitten, 2016). |
| Naïve Bayes Updateable | Class for a Naïve Bayes classifier using estimator classes (Ian Whitten, 2016). |
| Functions | Group of algorithms that can be written down as mathematical equations. Other methods such as decision trees and rules cannot therefore, these algorithms are categorised together in Weka. |
| Logistic Regression | Logistic regression is a classification algorithm used to find the probability of event success or failure. It is used when the dependent variable is binary in nature but, can be extended to suit multi-class classifications. It supports categorising data into discrete classes by studying the relationship from the given set of labelled data (Priyankur, What is Logistic Regression in Machine Learning, 2019).  **Advantages:**   * Easy to implement, interpret and efficient to train. * This algorithm can be easily extended for multi-class classification (multinominal logistic regression). * Provides good accuracy for simple datasets. * The predicted parameters give an inference about the importance of each feature. The direction of association is also given (positive and negative). Therefore, identifying the relationships between features is easy. * Allows models to be updated easily to reflect new data unlike, decision trees or support vector machines.   **Disadvantages:**   * If the number of observations is less than number of features, logistic regression is not suitable as it may lead to overfitting. * The assumption of linearity between the dependent variable and independent variables. Therefore, non-linear problems cannot be solved with this algorithm. * If two independent variables have a strong correlation, then only one should be used. Repetition of information can lead to wrong training of parameters. * Only import and relevant features should be used to build the model otherwise, the probabilistic predictions made may be incorrect and model’s predictive value may degrade. * The algorithm is sensitive to outliers. |
| Simple Logistic Regression | Classifier for building linear logistic regression models. Simple Logistic uses LogitBoost whereas Logistic uses a ridge estimator. SimpleLogistic builds logistic regression models, fitting them using LogitBoost with simple regression functions as base learners (Ian Whitten, 2016). |
| SMO  (Sequential Minimal Optimisation) | SMO refers to the specific efficient optimization algorithm used inside the Support Vector Machines (SVM) implementation, which stands for Sequential Minimal Optimization.  Support Vector Machines are developed for binary classification problems, although extensions to the technique have been made to support multi-class classification and regression problems.  SVM was developed for numerical input variables, although will automatically convert nominal values to numerical values. Input data is also normalized before being used. SVM work by finding a line that best separates the data into the two groups. This is done using an optimization process that only considers those data instances in the training dataset that are closest to the line that best separates the classes. The instances are called support vectors. In almost all problems of interest, a line cannot be drawn to neatly separate the classes, therefore a margin is added around the line to relax the constraint, allowing some instances to be misclassified but allowing a better result overall (Brownlee, 2019).  **Advantages:**   * Works well when there is a clear separation between classes. * Memory efficient. * Very effective with high dimensional data (where number of features may exceed number of observations). * Can work well with image data.   **Disadvantages:**   * Unsuitable for large datasets due to long training time. * Difficult to understand and interpret the final model, variable weights and individual impact. * Does not perform well with noisy datasets or when targets may overlap. * Choosing the optimal kernel function for SVM is a difficult task. |
| Lazy | Lazy learning methods simply store the data and the generalising beyond these data is postponed until an explicit request or a query is made. Explained further below. |
| IBk | K-Nearest Neighbours (KNN) classifier. It is a lazy learning, non-parametric algorithm and uses data with several classes to predict the classification of the new sample point. Neighbour-based classification is a type of instance-based learning or non-generalised learning. It does not attempt to construct a general internal model, rather simply stores the instances of the training data.  Classification is computed from a single majority vote of the nearest neighbour of each point. A query point is assigned the data class which has the most representatives within the nearest neighbours of the point. KNN is non-parametric since it does not make any assumptions on the data being studied. It can be used in both regression and classification problems (DN, 2020).  **Advantages:**   * Quick calculation time. * Simple algorithm – to interpret. * Versatile – useful for regression and classification. * High accuracy – you do not need to compare with better-supervised learning models. * No assumptions about data – no need to make additional assumptions, tune several parameters, or build a model. This makes it crucial in nonlinear data case.   **Disadvantages:**   * Accuracy depends on the quality of the data. * With large data, the prediction stage might be slow. * Sensitive to the scale of the data and irrelevant features. * Require high memory – need to store all of the training data. * Given that it stores all of the training, it can be computationally expensive. |
| KStar | K\* is an instance-based classifier. The class of a new instance is based upon the class of those training instances similar to it, as determined by some similarity function. The underlying assumption of instance-based classifiers such as K\* is that similar instances will have similar classes (Ian Whitten, 2016). |
| Rules | Rule based algorithms in Weka. |
| PART | Part obtains rules from partial decision trees. It builds the tree using C4.5’s heuristics with the same user-defined parameters as J4.8 (Ian Whitten, 2016). |
| Decision Table | It evaluates feature subsets using best-first search and can use cross-validation for evaluation. An option uses the nearest-neighbour method to determine the class for each instance that is not covered by a decision table entry, instead of the table’s global majority, based on the same set of features (Ian Whitten, 2016). |
| Trees | Decision Trees are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-regression). They work by creating a tree to evaluate an instance of data, start at the root of the tree and moving down to the leaves (roots) until a prediction can be made. The process of creating a decision tree works by selecting the best split point in order to make predictions and repeating the process until the tree is a fixed depth. After the tree is constructed, it is pruned in order to improve the model’s ability to generalize to new data (Poojari, 2019).  **Advantages:**   * Less effort required during pre-processing. * It forces consideration of all possible outcomes of a decision and traces each path to a conclusion. * Decision trees do not require normalisation of data. * Missing values in the data also not affect the process of building a decision tree to a considerable extent. * The decision tree is intuitive and easy to interpret for communication. * Very helpful in decision related problems.   **Disadvantages:**   * A small change in data can lead to a large change in the structure of the decision tree causing instability. * Model training time is relatively more as complexity is high. * Not applicable to regression problems. * Single decision tree is often a weak learner; therefore, a bunch of decision trees (random forest formula) is better for prediction. |
| Random Forest | Class for constructing a forest of random trees. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result (Ian Whitten, 2016). |
| J48 | Class for generating a pruned or unpruned C4.5 decision tree (Ian Whitten, 2016). |
| REPTree | Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e., as in C4.5) (Ian Whitten, 2016). |
| Meta Learning Classifiers | Algorithms that use or combine multiple algorithms, like Ensembles. Multiple methods combine the predictions from multiple models in order to make more robust predictions. These classifiers can be of two types:  **Homogenous:** All the algorithms are the same but are trained on different versions of the train data so that the chances of overfitting can be reduced.  **Heterogeneous:** Different algorithms work in tandem, and results from different models are combined to provide a single result.  The **advantage** of ensemble methods is that they are highly accurate and solve the overfitting issue. The **disadvantage** includes dealing with large number of parameters that, if not set appropriately, can severely compromise the functioning of the model (Rocca, 2019).  Although this would be beneficial to the study, limited knowledge and experience of these algorithms and fine tuning the parameters may be a challenge. |

### 6.1.2 Discussion on Models

The section above discusses the limitations and advantages of various models that need to be considered in selection of the data mining algorithm. The section below shows the exploration from results of the algorithms above in Weka. This practical exploratory analysis provides information on accuracy, which can support the decision making for the algorithm.

**Parameters and settings used for exploration:**

* Attributes:
  + index = Numeric = Predictor/ Independent Variable
  + year = Nominal = Predictor/ Independent Variable
  + region\_code = Nominal = Predictor/ Independent Variable
  + region = Nominal = Predictor/ Independent Variable
  + age\_group\_code = Nominal = Predictor/ Independent Variable
  + income\_weekly\_mean\_value = Numeric = Predictor/ Independent Variable
  + expenditure\_weekly\_per\_person= Numeric = Predictor/ Independent Variable
  + average\_sale\_price = Numeric = Predictor/ Independent Variable
  + hpi = Numeric = Predictor/ Independent Variable
  + cpi = Numeric = Predictor/ Independent Variable
  + annual\_savings = Numeric = Predictor/ Independent Variable
  + downpayment\_capacity = Numeric = Predictor/ Independent Variable
  + loan\_to\_value\_ratio = Numeric = Predictor/ Independent Variable
  + **affordability = Nominal = Target**
* Table 20 below shows the results for **70% training data and 30% test** **data** used for analysis (explained below in section 7.1 Test Designs).
* Table 21 below shows the results for **80% training data and 20% test data** used for analysis (further explained below).
* Weka Filter – NumerictoNominal for conversion of attributes (year, region\_code, region, and age\_group\_code)

Graphical user interface

Description automatically generated

Figure 100 - Exploration of algorithms in Weka

Table 20 - Exploratory analysis of algorithms with 70/30 split

|  |  |  |  |
| --- | --- | --- | --- |
| CATEGORY | ALGORITHM | RESULTS | |
| Bayes Classifier | **BayesNet** | Correctly Classified Instances: 15 – 68.1818% | Incorrectly Classified Instances: 7 – 31.8182% |
| **NaiveBayes** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| **NaiveBayesMultinominalText** | Correctly Classified Instances: 12 – 54.5455% | Incorrectly Classified Instances: 10 – 45.4545% |
| **NaiveBayesUpdateable** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| Functions | **Logistic** | Correctly Classified Instances: 12 – 54.5455% | Incorrectly Classified Instances: 10 – 45.4545% |
| **MultilayerPerception** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |
| **SimpleLogistic** | **Correctly Classified Instances: 18 – 81.8182%** | **Incorrectly Classified Instances: 4 – 18.1818%** |
| **SMO** | Correctly Classified Instances: 16 – 72.7273% | Incorrectly Classified Instances: 6 – 27.2727% |
| Lazy | **IBk** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |
| **KStar** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |
| Meta | **AdaBoostM1** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| **Bagging** | Correctly Classified Instances: 12 – 54.5455% | Incorrectly Classified Instances: 10 – 45.4545% |
| **LogitBoost** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |
| **MultiClassClassifier** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| **Vote** | Correctly Classified Instances: 12 – 54.5455% | Incorrectly Classified Instances: 10 – 45.4545% |
| Rules | **DecisionTable** | **Correctly Classified Instances: 18 – 81.8182%** | **Incorrectly Classified Instances: 4 – 18.1818%** |
| **OneR** | Correctly Classified Instances: 13 – 59.0909% | Incorrectly Classified Instances: 9 – 40.9091% |
| **PART** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| Trees | **ZeroR** | Correctly Classified Instances: 12 – 54.5455% | Incorrectly Classified Instances: 10 – 45.4545% |
| **J48** | Correctly Classified Instances: 14 – 63.6364% | Incorrectly Classified Instances: 8 – 36.3636% |
| **RandomForest** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |
| **RandomTree** | Correctly Classified Instances: 17 – 77.2727% | Incorrectly Classified Instances: 5 – 22.7273% |

**Iterating backwards from section 7.1 Test Designs**. Once the Simple Logistic algorithm was selected (refer section 6.2 Data Mining Algorithm Selection for reasoning), it was tested with various different training and test data splits to understand how the model performs with various different sizes of training data. This exercise resulted in a much better accuracy for an 80/20 spit rather than a 70/30 split. Therefore, it was important to understand how the accuracy level affects the other algorithms. The table below shows the results for the same algorithms as above, but for an 80% training data and 20% test data.

Table 21 - Results from exploratory analysis with 80/20 split

|  |  |  |  |
| --- | --- | --- | --- |
| CATEGORY | ALGORITHM | RESULTS | |
| Bayes Classifier | BayesNet | Correctly Classified Instances: 14 – 93.3333% | Incorrectly Classified Instances: 1 – 6.6667% |
| **NaiveBayes** | Correctly Classified Instances: 10 – 66.6667% | Incorrectly Classified Instances: 5 – 33.3333% |
| **NaiveBayesMultinominalText** | Correctly Classified Instances: 8 – 53.3333% | Incorrectly Classified Instances: 7 – 46.6667% |
| **NaiveBayesUpdateable** | Correctly Classified Instances: 10 – 66.6667% | Incorrectly Classified Instances: 5 – 33.3333% |
| Functions | **Logistic** | **Correctly Classified Instances: 13 – 86.6667%** | **Incorrectly Classified Instances: 2 – 13.3333%** |
| **MultilayerPerception** | **Correctly Classified Instances: 13 – 86.6667%** | **Incorrectly Classified Instances: 2 – 13.3333%** |
| **SimpleLogistic** | **Correctly Classified Instances: 13 – 86.6667%** | **Incorrectly Classified Instances: 2 – 13.3333%** |
| **SMO** | **Correctly Classified Instances: 13 – 86.6667%** | **Incorrectly Classified Instances: 2 – 13.3333%** |
| Lazy | IBk | Correctly Classified Instances: 14 – 93.3333% | Incorrectly Classified Instances: 1 – 6.6667% |
| **KStar** | Correctly Classified Instances: 12 – 80.0000% | Incorrectly Classified Instances: 3 – 20.0000% |
| Meta | **AdaBoostM1** | Correctly Classified Instances: 10 – 66.6667% | Incorrectly Classified Instances: 5 – 33.3333% |
| **Bagging** | Correctly Classified Instances: 10 – 66.6667% | Incorrectly Classified Instances: 5 – 33.3333% |
| **LogitBoost** | Correctly Classified Instances: 12 – 80.0000% | Incorrectly Classified Instances: 3 – 20.0000% |
| **MultiClassClassifier** | Correctly Classified Instances: 11 – 73.3333% | Incorrectly Classified Instances: 4 – 26.6667% |
| **Vote** | Correctly Classified Instances: 8 – 53.3333% | Incorrectly Classified Instances: 7 – 46.6667% |
| Rules | **DecisionTable** | **Correctly Classified Instances: 13 – 86.6667%** | **Incorrectly Classified Instances: 2 – 13.3333%** |
| **OneR** | Correctly Classified Instances: 10 – 66.6667% | Incorrectly Classified Instances: 5 – 33.3333% |
| **PART** | Correctly Classified Instances: 11 – 73.3333% | Incorrectly Classified Instances: 4 – 26.6667% |
| **ZeroR** | Correctly Classified Instances: 8 – 53.3333% | Incorrectly Classified Instances: 7 – 46.6667% |
| Trees | **J48** | Correctly Classified Instances: 11 – 73.3333% | Incorrectly Classified Instances: 4 – 26.6667% |
| RandomForest | Correctly Classified Instances: 14 – 93.3333% | Incorrectly Classified Instances: 1 – 6.6667% |
| **RandomTree** | Correctly Classified Instances: 11 – 73.3333% | Incorrectly Classified Instances: 4 – 26.6667% |

These results are significantly different to the ones for the 70/30 split. Therefore, to investigate these results further, top three best performing models were selected. These three models (bayes net, IBk and random forest) were tested with 60/40, 70/30, 80/20, and 90/10 training and test data splits as shown in the screenshot below. When tested, majority of them either had lower accuracy or the models were not able to determine the precision, F-measure, or MCC accuracy measures for some of the classes (shown in screenshot below). Therefore, the decision was made to revert back to the 70% training and 30% test data split (results from Table 20) and using the SimpleLogistic algorithm (further explained below) as these showed the most stable results in terms of accuracy as well as detailed accuracy by class for all classes.

Graphical user interface, text

Description automatically generated

Figure 101 - Test design for three models

## 6.2 Data Mining Algorithm Selection

The current information about the dataset and requirements for mining:

* Less training data (not much data available overall)
* Easy to interpret and less complexity due to lack of in-depth knowledge and experience with data mining.
* Categorical and numeric predictor variables
* Categorical target variable
* Categorical target variable has multiple classes (not binary)

Based on the analysis above of various algorithms, their limitations, advantages and disadvantages, accuracy of the models, and the given dataset the best suited algorithm for the study is SimpleLogistic Algorithm.

Broadly, all the algorithms under Bayes Classifiers and Meta Classifiers performed poorly and returned lower levels of accuracy. This may be caused by model parameter settings, especially for the Meta Classifiers as they are highly sensitive to these changes (refer disadvantages section 6.1.1 Discussion on Algorithm Theory – under Meta Learning Classifiers). Issues with model training time is not considered in the decision making as the dataset is relatively small therefore, the time difference between the models is negligible.

The remaining algorithms within the Lazy, Trees, Rules, and Function categories performed much better in terms of accuracy. Many factors could have contributed to this, as their limitations and capabilities vary significantly. As mentioned earlier, the test design also has an effect on the accuracy. Although the overall accuracy of 80/20 split in many of the other models was better, their accuracy for individual classes with indeterminant therefore, unreliable. Out of all the algorithms from the 70/30 split, two algorithms performed best, which were SimpleLogistic (under functions) and DecisionTable (under rules) gave the best accuracy.

The screenshots below show the detailed results of both. Although the DecisionTable model is able to achieve good accuracy, it displays lack of information under the Detailed Accuracy By Class section. This means that it is unable to determine the precision, F-Measure and MCC. Precision answers the question, how many of the selected items are relevant? And recall answers the question, how many relevant items are selected? Precision can be seen as a measure of quality and recall as a measure of quantity. Higher precision means that an algorithm returns more relevant results than irrelevant ones, and high recall means that an algorithm returns most of the relevant results. The F-Measure column refers to the mean of precision and recall. SimpleLogistic algorithm performs much better for all of the accuracy measures concerned in comparison to DecisionTable with regards to all classes, therefore it is best suited for the study.

A picture containing text, receipt, screenshot

Description automatically generated

Figure 102 - Results from DecisionTable Algorithm

A picture containing text, receipt

Description automatically generated

Figure 103 - Results from SimpleLogistic Algorithm

## 6.3 Model and Parameters

**Parameters and settings used for data mining:**

* Input Attributes:
  + index = Numeric = Predictor/ Independent Variable/ X Variable
  + year = Nominal = Predictor/ Independent Variable/ X Variable
  + region\_code = Nominal = Predictor/ Independent Variable/ X Variable
  + region = Nominal = Predictor/ Independent Variable/ X Variable
  + age\_group\_code = Nominal = Predictor/ Independent Variable/ X Variable
  + income\_weekly\_mean\_value = Numeric = Predictor/ Independent Variable/ X Variable
  + expenditure\_weekly\_per\_person= Numeric = Predictor/ Independent Variable/ X Variable
  + average\_sale\_price = Numeric = Predictor/ Independent Variable/ X Variable
  + hpi = Numeric = Predictor/ Independent Variable/ X Variable
  + cpi = Numeric = Predictor/ Independent Variable/ X Variable
  + annual\_savings = Numeric = Predictor/ Independent Variable/ X Variable
  + downpayment\_capacity = Numeric = Predictor/ Independent Variable/ X Variable
  + loan\_to\_value\_ratio = Numeric = Predictor/ Independent Variable/ X Variable
* Output Attribute:
  + **affordability = Nominal = Target/ Dependent Variable/ Y Variable**
  + Classes within affordability:
    - 0 – 20
    - 21 – 40
    - 41 – 60
    - 61 – 80
    - 81 – 100
    - 100
* 70% Training Data and 30% Test Data

Graphical user interface

Description automatically generated

Figure 104 - Loading data from a csv file



Figure 105 - Attribute’s visualisation

Graphical user interface

Description automatically generatedGraphical user interface, text, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Figure 106 - Training and test data split

Figure 107 - Classifier evaluation selections

Figure 108 - Filter selection

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Figure 109 - Evaluation metrics for model output

Figure 110 - Model parameters used for data mining

# 7.0 Data Mining

## 7.1 Test Designs

Overfitting in data mining occurs when a model learns the detail and noise from the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise and small fluctuations within the data is picked up as learned concepts by the model. This creates a problem in which these concepts do not apply to new data and negatively impact the model’s ability to generalise.

On the other hand, underfitting refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will display poor performance on the training data. Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. It is often a result of an excessively simple model (Koehrsen, 2018).

Models that are to be built, must be executed on a separate dataset to the test dataset. By default, Weka splits the data by 66% for training and 34% for testing. A more common, industry standard partition can be applied to the data so that the slip is 70% training and 30% test data. The reason for 70% of data is designated for training, is to give models a larger amount of data to be able to accurately understand the trends and patterns from the training data. An 90% training data to 10%, or the 80/20 split although good for modelling, is insufficient for testing and may lead to overfitting. A 60/40 split returns a higher cost of losing training data for modelling. To prevent the problems of overfitting or at the other spectrum, underfitting, the 70/30 split provides the strongest results in terms of correlation when compared to 60/40, 80/20 and 90/10. The 70/30 split provides a balance and reduces the chances of overfitting or underfitting. The table below displays the results of accuracy when different percentages are used. The 70/30 split returns the best results with the least amount of gap between the performance of training data and the test data, ensuring the model is not overfitting or underfitting. Whereas the 80/20 split provides better overall accuracy but, not individual class detailed accuracy.

Table 22 - Test Design splits on SimpleLogistic Algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| PERCENTAGE OF TRAINING DATA | PERCENTAGE OF TEST DATA | TEST DATA RESULTS | |
| Correctly Classified Instances | Incorrectly Classified Instances |
| 60 | 40 | 23 – 76.6667% | 7 – 23.3333% |
| **70** | **30** | **18 – 81.8182%** | **4 – 18.1818%** |
| **80** | **20** | **13 – 86.6667%** | **2 – 13.3333%** |
| 90 | 10 | 5 – 71.4286% | 2 – 28.5714% |

## Text Description automatically generated with medium confidence7.2 Data Mining Activity

The SimpleLogistic model ran successfully. The final model output shown below.

Text

Description automatically generated with medium confidence

Figure 111 - Part one of the SimpleLogistic algorithm output

A picture containing text, receipt, screenshot

Description automatically generated

Figure 112 - Part two of the Simple Logistic algorithm output

## 7.3 Patterns from Model Output

Below are the patterns identified with data mining goals in perspective. Refer to section 8.2 Visualisation of Data, Results, Models and Patterns for more graphical representation of trends.

**Basic Trends Observed**

1. Weekly income values and expenditure values per person have been increasing over the years.
2. Income has been increasing faster than the weekly expenditure, therefore annual savings have increased over the years.
3. House price index and consumer price index has been increasing over the years.
4. Average sale prices have been increasing over the years, which in turn has increase loan to value ratio figures.
5. With the increase in annual savings over the years, the down payment capacity has also increased.
6. The down payment capacity increase is more gradual than the loan to value ratio.

**Underlying Trends Observed**

1. Individuals in age group 30 to 34 are closest to affording a house. Whereas ages 20 – 29 are further away in purchasing a house due to high amount of deposit required and lower annual savings.
2. Auckland, followed by Wellington and then the South Island have the greatest levels of unaffordability, this coincides with the higher house prices in Auckland and Wellington.
3. Greatest down payment capacity is greatest in Wellington, which also has highest incomes.
4. Better affordability is achieved with higher income and lower expenditure.
5. Better affordability is achieved with higher income and higher annual savings.
6. Better affordability is achieved with higher annual savings and lower expenditure.
7. Better affordability is achieved with lower loan to value ratio and higher annual savings.
8. Better affordability is achieved with lower loan to value ratio and higher down payment.
9. Unable to make a corelation between year and affordability due to lack to data for a clear conclusion.

# 8.0 Interpretation

## 8.1 Discussion on Mined Patterns

While the basic patterns are apparent in the raw data as well as the patterns identified after the data mining exercise, the underlying patterns such as the affordability levels within age groups and regions are less apparent. Following the analysis in sections 6.2 Data Mining Algorithm Selection to 7.1 Test Designs the model gives 81% correctly classified instances. Although a reliable measure, the detailed accuracy by class statistics reinforces the patterns and model results. It provides a more detailed class breakdown of the classifier’s prediction accuracy. An "optimal" classifier will have ROC area values approaching a value of 1, and a value of 0.5 can be compared to random guessing. The ROC area for all classes is relatively high which indicates the classifier has a good performance.

The time taken by the model to produce an output was negligible. Although with more data, it would consume more time. There are no missing values or null values within the dataset therefore the model did not experience any issues as the data was cleansed and prepared prior to modelling.

## 8.2 Visualisation of Data, Results, Models and Patterns

Shape, rectangle

Description automatically generated

Figure 113 - Affordability levels

**Basic Trends Observed**

1. Weekly income values and expenditure values per person have been increasing over the years.

Chart, line chart

Description automatically generated

Figure 114 - Income and expenditure trend

1. Income has been increasing faster than the weekly expenditure, therefore annual savings have increased over the years.

Chart, line chart

Description automatically generated

Figure 115 - Annual savings and down payment capacity trends

1. House price index and consumer price index has been increasing over the years.

Chart, line chart

Description automatically generated

Figure 116 - Consumer price index and house price index trends

1. Average sale prices have been increasing over the years, which in turn has increase loan to value ratio figures.

Shape, rectangle

Description automatically generated

Figure 117 - Average house sale prices and loan to value ratio trend

1. With the increase in annual savings over the years, the down payment capacity has also increased.
2. The down payment capacity increase is more gradual than the loan to value ratio.

Chart, line chart

Description automatically generated

Figure 118 - Annual savings, down payment capacity and loan to value ratio trends

**Underlying Trends Observed**

1. Individuals in age group 30 to 34 are closest to affording a house. Whereas ages 20 – 29 are further away in purchasing a house due to high amount of deposit required and lower annual savings.

Chart

Description automatically generated

Figure 119 - Affordability according to age groups

1. Auckland, followed by Wellington and then the South Island have the greatest levels of unaffordability, this coincides with the higher house prices in Auckland and Wellington.

Chart, bar chart

Description automatically generated

Figure 120 - Affordability according to region and loan to value ratio

1. Greatest down payment capacity is greatest in Wellington, which also has highest incomes.

Chart, bar chart

Description automatically generated

Figure 121 - Affordability according to down payment capacity and region

1. Better affordability is achieved with higher income and lower expenditure.

Chart, scatter chart

Description automatically generated

Figure 122 - Effect of income and expenditure on affordability

1. Better affordability is achieved with higher income and higher annual savings.

Chart, scatter chart

Description automatically generated

Figure 123 - Effect of income and annual savings on affordability

1. Better affordability is achieved with higher annual savings and lower expenditure.

Chart, scatter chart

Description automatically generated

Figure 124 - Effect of expenditure and annual savings on affordability

1. Better affordability is achieved with lower loan to value ratio and higher annual savings.

Chart, scatter chart

Description automatically generated

Figure 125 - Effect of loan to value ratio and annual savings on affordability

1. Better affordability is achieved with lower loan to value ratio and higher down payment.

Chart, scatter chart

Description automatically generated

Figure 126 - Effect of down payment capacity and loan to value ratio on affordability

## 8.3 Interpretation of Results, Models and Patterns

The datasets are collected from government sources to ensure consistency. The same data is used for by the government for policymaking therefore, it is ideal for understanding affordability within New Zealand as the business objective is to determine if government intervention is required. Sufficient data was available to understand and analyse the data mining objectives and business objectives. The general data mining goals mentioned in section 1.3 Data Mining Objectives related to collecting, understanding the accuracy, data structure, integrity and the relationships within the datasets. It also included understanding the outliers, missing values, and null data. The data understanding, preparation and transformation process assisted in achieving this objective. The raw data showed patterns that could be explored to understand the inherent relationships within and in between the datasets. This exploration of raw data was then used to further integrate and aggregate the data for data mining. Outliers and missing values are also located and investigated through the same steps.

The data clearly showed a trend of income and expenditure increasing over the years but, income is increasing at a faster rate as showed in the line graph above. The house prices have also increased over the years. This is reflected within the historical sales data as well as the house price index data. Previously, the ability to save for down payment within three years was possible. In the recent years, the house prices have increased at a rapid rate therefore, the loan to value ratio, which determines the amount of deposit paid, has also increased. When the figures of loan to value ratio (deposit required) is compared with ability to save for a down payment (3 years of annual savings), it shows a clear pattern of house prices increasing at a faster rate than the incomes.

The specific data mining goals also referred to the calculation of an individual’s savings capacity. This was achieved through the use of income and expenditure data per person to obtain the annual savings for age groups of 20 to 24, 25 to 20 and 30 to 34. The data was also region specific therefore, annual savings could be calculated according to age as well as regions. The annual savings data was then used to calculate the ability to save for down payment within three years which was also according to age groups and regions. Aggregating the data according to age groups and regions provided more fine-grained information.

Historical trends of residential sales were identified through the sales dataset. Average house price per region and per year was calculated. This was then used to identify the typical deposit (10% loan to value ratio of the average house price) required per year and per region. The calculated down payment capacity from annual savings and the deposit required, calculated from the average house prices, was used to determine the affordability levels. The affordability measure was spilt into multiple classes rather than a binary yes or no (can afford to cannot afford to buy a house) to give a better understanding. All of the data mining objectives have been achieved through the steps above.

## 8.4 Assessment and Evaluation of Results, Models and Patterns

**Business Objective 1: If government intervention is required to support housing affordability for young kiwis aged between 20 to 35 years.**

**Affordability Levels**

0 – 20 = indicates 3 years of annual savings only amounts to maximum of 20% for deposit required for the house.

21 – 40 = indicates 3 years of annual savings only amounts to maximum of 40% for deposit required for the house.

41 – 60 = indicates 3 years of annual savings amounts to maximum of 60% for deposit required for the house. Therefore, another few years of savings is required to meet the amount of deposit required.

61 – 80 = indicates 3 years of annual savings amounts to maximum of 80% for deposit required for the house. Hence, another one to two years of savings may be able to bridge the gap in order to pay for a deposit.

81 – 100 = indicates 3 years of annual savings amounts to maximum of 100% for deposit required for the house. This means that the individuals may be able to afford a deposit in the near future.

100+ = indicates 3 years of annual savings is greater than the deposit required for a house purchase. Therefore, they are able to purchase a house anytime.

There was a clear pattern emerging from the analysis that showed age groups 20 to 24 and 25 to 29 had the highest levels of difficulty in purchasing a house with three years of savings and regardless of their location in New Zealand. This was a result of lower income levels within both of these groups. This shows clear evidence that government assistance or action is required in order increase home ownership within younger age groups. The third age group showed higher levels of affordability due to their increase in income over the years.

Diagram

Description automatically generated with low confidence

30 to 34

25 to 29

20 to 24

Figure 127 - Affordability by age groups and region

Graphical user interface

Description automatically generated

Figure 128 - Income levels by age groups

**Business Objective 2: If government intervention is required, should it apply uniformly across all of New Zealand, or should it be region specific.**

The results for the business objective one identified the need for government intervention for age groups 20 to 24 and 25 to 29. The data mining also showed the significant differences in affordability between the various regions. Auckland, followed by Wellington are the two main regions that require more concentrated government assistance as indicated in the graphs above.

Graphical user interface, text, application

Description automatically generatedFigure 129 - Attribute Correlations

The house price index and consumer price index datasets are used to cross check and reinforce the trends observed with the historical residential sales data and the expenditure data. This means that they do not have a significant impact on the affordability and results in least amount of correlation. The correlation results confirm the trends observed and identify annual savings, down payment capacity, age, and income as the major drivers for affordability of houses.

**Limitations**

Since the expenditure dataset included limited number of regions compared to other datasets, the overall study had to be restricted to broader regions of New Zealand. Therefore, more fine-grained analysis on smaller regions is not achievable until the raw data becomes available.

## 8.5 Iteration of Steps 1-7

Many iterations have been carried out from data understanding onwards. Below table summarises the steps repeated and the associated reason. All of the steps identified below, show the locations that required revisiting the python code and re-importing the data into Weka above and iterating through the steps in order to generate the correct information.

Table 23 - Iterations carried out through the data mining exercise

|  |  |
| --- | --- |
| **Data Understanding** | |
| Visualisation of Raw Data | Once the datasets are loaded to the python code, the functions such as describe(), unique(), dtypes were used to check the measurement values. This was followed by the use of matplotlib to visualise the data. Tableau was also used to interpret the raw data. Depending on the type of graph used, the measurement values had to be adjusted in order to produce the graph. For instance, the field year in the income dataset was read as continuous initially but had to be changed to categorical in order to visualise the dataset. This step was repeated across all datasets in order to get a clear understanding. |
| **Data Preparation** | |
| Filtering Age Groups Code | During the assignment of age groups, presence of null values was noticed, therefore python code was added to detect and remove the null values. |
| Expenditure per Person | While calculating the annual savings for an individual, it was noticed that the expenditure values are per household whereas the income data is per person. Therefore, another field was added in order to calculate expenditure value per person and used later to derive the annual savings. |
| Region Codes Added | While attempting to merge from the income and expenditure datasets based on regions and year, it was observed that the income dataset has more fine grain detail regarding regions and the expenditure dataset had more broader regions. Hence, region codes had to be added to create a common field in order to merge the datasets. |
| Aggregating the Income Data | Duplication of data was noticed following the merging of income and expenditure datasets; therefore, mean income values were calculated in order remove duplication |
| Merging Data | Since the datasets were merged step by step, they created duplicate fields. For example, when the house price index and consumer price index were merged, the table included the year from HPI dataset as well as CPI dataset. Therefore, python code had to be added to remove this duplication. This change is noted after a few steps, requiring the return to above to remove duplicate fields. |
| Additional Fields | As new fields were added and aggregated, python code had to be added to reduce (drop) the unnecessary fields. |
| **Data Transformation and Data Mining** | |
| Algorithm Selection | Various different algorithms are tested with the data to select the most suitable. This required significant number of iterations. |
| Partitioning | As mentioned above under test designs, various partitions with different percentage splits between training, test and validation data were attempted and compared with selected model output. This required various iterations. |
| Test Designs | Following the selection of appropriate algorithm, various different test/ training data partitions were tried. This resulted in a significant difference in the model output, therefore algorithm selection was re-evaluated and re-run to ensure correct selection. This was also followed by retesting the partitions created for test and training data. |
| **Visualisations** | |
| Communication of Results | Various types of graphs and specific colours for classes of affordability had to be tried to best communicate the results |

# 9.0 References

Brownlee, J. (2019, August 22). How To Use Classification Machine Learning Algorithms in Weka. Retrieved from: https://machinelearningmastery.com: <https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>

CoreLogic (2021). About Us. Retrieved from: <https://www.corelogic.co.nz/about-us>

DN, S. (2020, August 6). K-Nearest Neighbours Algorithm. Retrieved from: https://medium.com: https://medium.com/analytics-vidhya/k-nearest-neighbors-algorithm-7952234c69a4

Koehrsen, Will (2018). Overfitting vs Underfitting: A complete Example. Retrieved from: <https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765>

Ian Whitten, E. F. (2016). The Weka Workbench. Retrieved from: <https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf>

Poojari, D. (2019, August 2). Machine Learning Basics: Decision Tree From Scratch. Retrieved from: https://towardsdatascience.com: https://towardsdatascience.com/machine-learning-basics-descision-tree-from-scratch-part-i-4251bfa1b45c

Priyankur, S. (2019, September 23). What is Logistic Regression in Machine Learning. Retrieved from: https://www.knowledgehut.com: https://www.knowledgehut.com/blog/data-science/logistic-regression-for-machine-learning

Priyankur, S. (2019, September 23). What is Naive Bayes in Machine Learning. Retrieved from: https://www.knowledgehut.com: <https://www.knowledgehut.com/blog/data-science/naive-bayes-in-machine-learning>

Rocca, J. (2019, April 23). Ensemble methods: bagging, boosting and stacking. Retrieved from: https://towardsdatascience.com: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Statistics New Zealand (2018). *Census of Population and Dwelling*. Retrieved from: <https://www.stats.govt.nz/tools/2018-census-place-summaries/>

Statistics New Zealand (2016). *Quarterly Employment Survey*. Retrieved from: <https://catalogue.data.govt.nz/dataset/quarterly-employment-survey>

Statistics New Zealand (2019). *Household Economic Survey*. Retrieved from: <https://catalogue.data.govt.nz/dataset/household-economic-survey-17-18-warm-and-dry>

Statistics New Zealand (2020). *Household Labour Force Survey*. Retrieved from: <https://catalogue.data.govt.nz/dataset/household-labour-force-survey>

Statistics New Zealand (2020). *Consumer Price Index*. Retrieved from: <https://catalogue.data.govt.nz/dataset/consumers-price-index>

Statistics New Zealand (2020). *Apply to use microdata for research*. Retrieved from: <https://www.stats.govt.nz/integrated-data/apply-to-use-microdata-for-research/>

PwC New Zealand (2020). *The next step in addressing housing affordability*. Retrieved from: <https://www.pwc.co.nz/insights-and-publications/2020-publications/the-next-step-in-addressing-housing-affordability.html>

Reserve Bank of New Zealand (2015). *Action needed to reduce housing imbalances*. Retrieved from: <https://www.rbnz.govt.nz/research-and-publications/speeches/2015/speech2015-04-15>

Reserve Bank of New Zealand (2021). *Monetary Policy Statement*. Retrieved from: <https://www.rbnz.govt.nz/-/media/ReserveBank/Files/Publications/Monetary%20policy%20statements/2021/mpsfeb21.pdf?revision=d3022919-4b30-4663-ab8a-e3be4ea2b453>

Reserve Bank of New Zealand (2021). *Loan-to-valuation ratio restrictions*. Retrieved from: <https://www.rbnz.govt.nz/regulation-and-supervision/banks/macro-prudential-policy/loan-to-valuation-ratio-restrictions>

Reserve Bank of New Zealand (2021). *Housing - M10*. Retrieved from: <https://www.rbnz.govt.nz/statistics/m10>

Statistics New Zealand (2020) *Housing in Aotearoa: 2020*. Retrieved from: <https://www.stats.govt.nz/reports/housing-in-aotearoa-2020>

World Economic Forum (2019). *Making Affordable Housing a Reality in Cities.* Retrieved from: <http://www3.weforum.org/docs/WEF_Making_Affordable_Housing_A_Reality_In_Cities_report.pdf>

# List of Tables

[Table 1 - Risks and contingencies 11](#_Toc72073238)

[Table 2 - Terminology and abbreviations 12](#_Toc72073239)

[Table 3 - Risks and resources required for each of the data mining steps 14](#_Toc72073240)

[Table 4 - Project plan 16](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72073241)

[Table 5 - Datasets, data sources, key attributes, and related issues 17](#_Toc72073242)

[Table 6 – Data sets’ quality and quantities 20](#_Toc72073243)

[Table 7 - Verification of data quality - Household Labour Force Survey 39](#_Toc72073244)

[Table 8 - Verification of data quality - Household Economic Survey 40](#_Toc72073245)

[Table 9 - Verification of data quality - Consumer Price Index Data 41](#_Toc72073246)

[Table 10 - Verification of data quality - House Price Index Data 42](#_Toc72073247)

[Table 11 - Verification of data quality - Residential Sales Data 43](#_Toc72073248)

[Table 12 - Data selection 46](#_Toc72073249)

[Table 13 - Data cleaning 51](#_Toc72073250)

[Table 14 - Data construction 56](#_Toc72073251)

[Table 15 - Additional fields 64](#_Toc72073252)

[Table 16 - Data integration 72](#_Toc72073253)

[Table 17 - Distribution Graphs for Year, Region and Age Groups 80](#_Toc72073254)

[Table 18 - Data projection 83](#_Toc72073255)

[Table 19 - Weka Algorithms 92](#_Toc72073256)

[Table 20 - Exploratory analysis of algorithms with 70/30 split 99](#_Toc72073257)

[Table 21 - Results from exploratory analysis with 80/20 split 100](#_Toc72073258)

[Table 22 - Test Design splits on SimpleLogistic Algorithm 110](#_Toc72073259)

[Table 23 - Iterations carried out through the data mining exercise 132](#_Toc72073260)

# List of Figures

[Figure 1 - Basic components of housing affordability 6](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135696)

[Figure 2 - Income data description 21](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135697)

[Figure 3 - Expenditure data description 22](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135698)

[Figure 4 - Consumer price index data description 23](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135699)

[Figure 5 - House price index data description 24](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135700)

[Figure 6 - Sales data description 25](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135701)

[Figure 7 - Python - Data description 25](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135702)

[Figure 8 - Income values by regions 27](#_Toc72135703)

[Figure 9 - Income values by age groups 28](#_Toc72135704)

[Figure 10 - Weekly income level changes by year 29](#_Toc72135705)

[Figure 11 - Expenditure data by categories 30](#_Toc72135706)

[Figure 12 - Expenditure data by regions 31](#_Toc72135707)

[Figure 13 - Expenditure data by year 32](#_Toc72135708)

[Figure 14 - Consumer Price Index change by year 34](#_Toc72135709)

[Figure 15 - House price index changes per year 35](#_Toc72135710)

[Figure 16 - Average house sale prices per region 37](#_Toc72135711)

[Figure 17 - Average sale price changes over years and regions 38](#_Toc72135712)

[Figure 18 - Null values in income data 39](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135713)

[Figure 19 - describe () function on income values 39](#_Toc72135714)

[Figure 20 - Null values in expenditure data 40](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135715)

[Figure 21 - describe () function on expenditure data 40](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135716)

[Figure 22 - describe () function on consumer price index data 41](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135717)

[Figure 23 - Null values in consumer price index data 41](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135718)

[Figure 24 - Null values in house price index data 42](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135719)

[Figure 25 - describe () function on house price index data 42](#_Toc72135720)

[Figure 26 - Null values on residential sales data 43](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135721)

[Figure 27 - describe () function on residential sales data 43](#_Toc72135722)

[Figure 28 - Python commands for data verification 44](#_Toc72135723)

[Figure 29 - Data importing, reading, verification and description in python 45](#_Toc72135724)

[Figure 30 - Condition for filtering age groups in the income data 46](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135725)

[Figure 31 - Sample results of filter condition on age groups 46](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135726)

[Figure 32 - Condition for filtering categories in expenditure data 47](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135727)

[Figure 33 - Random sample results of filter condition on expenditure categories 47](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135728)

[Figure 34 - Python statement to eliminate unnecessary attributes 49](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135729)

[Figure 35 - Remaining attributes and the data types 49](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135730)

[Figure 36 - Python statement to remove unnecessary attributes 50](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135731)

[Figure 37 - Remaining attributes and the data types 50](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135732)

[Figure 38 - Python statement to remove unnecessary attributes 50](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135733)

[Figure 39 - Remaining attributes and the data types 50](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135734)

[Figure 40 - Python statement to check and remove null values 51](#_Toc72135735)

[Figure 41 - Results before and after removing null values 51](#_Toc72135736)

[Figure 42 - Python statement for minimum and maximum values 52](#_Toc72135737)

[Figure 43 - Results for quantile, minimum and maximum values 52](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135738)

[Figure 44 - Python statements for outlier detection in sales data 53](#_Toc72135739)

[Figure 45 - Results for outliers in sales data 53](#_Toc72135740)

[Figure 46 - Python statement for replacing null data in house price index 54](#_Toc72135741)

[Figure 47 - Results for replacing null data in house price index 54](#_Toc72135742)

[Figure 48 - Python statement for detection of null values in consumer price index 55](#_Toc72135743)

[Figure 49 - Results for null values in consumer price index 55](#_Toc72135744)

[Figure 50 - Python statement for outlier detection 55](#_Toc72135745)

[Figure 51 - Python code for assigning age group code 56](#_Toc72135746)

[Figure 52 - Results for assigning age group code 56](#_Toc72135747)

[Figure 53 - Python code for assigning region codes in income data 57](#_Toc72135748)

[Figure 54 - Results for assigning region codes for income data 58](#_Toc72135749)

[Figure 55 - Python code for assigning region code for expenditure data 59](#_Toc72135750)

[Figure 56 - Results for assigning region code for expenditure data 59](#_Toc72135751)

[Figure 57 - Python statement for assigning region codes 60](#_Toc72135752)

[Figure 58 - Sample results of assigning region codes to sales data 61](#_Toc72135753)

[Figure 59 - Python code for calculating annual house price index 62](#_Toc72135754)

[Figure 60 - Results after calculating annual house price index 62](#_Toc72135755)

[Figure 61 - Python code for calculating the annual consumer price index 63](#_Toc72135756)

[Figure 62 - Results after calculating annual consumer price index 63](#_Toc72135757)

[Figure 63 - Python code for calculating the average income 64](#_Toc72135758)

[Figure 64 - Results after averaging the income figures 64](#_Toc72135759)

[Figure 65 - Python code for calculating expenditure per person 65](#_Toc72135760)

[Figure 66 - Sample results for calculating expenditure per person 65](#_Toc72135761)

[Figure 67 - Python code for calculating average sale prices across regions and by year 66](#_Toc72135762)

[Figure 68 – Sample results from calculating average sale prices 66](#_Toc72135763)

[Figure 69 - Python code for calculating annual savings 67](#_Toc72135764)

[Figure 70 - Random samples of results from annual savings calculations 67](#_Toc72135765)

[Figure 71 - Python code for calculating down payment capacity 68](#_Toc72135766)

[Figure 72 - Sample of results from down payment calculations 68](#_Toc72135767)

[Figure 73 - Python code for calculating loan to value ratio 69](#_Toc72135768)

[Figure 74 - Sample of results for loan to value ratio calculations 69](#_Toc72135769)

[Figure 75 - Python code for assigning affordability category 70](#_Toc72135770)

[Figure 76 - Sample of results for affordability attribute 71](#_Toc72135771)

[Figure 77 - Python code for merging income and expenditure data 72](#_Toc72135772)

[Figure 78 - Results for merging income and expenditure data 73](#_Toc72135773)

[Figure 79 - Aggregate sales data 74](#_Toc72135774)

[Figure 80 - Results from aggregating the data 74](#_Toc72135775)

[Figure 81 - Python code for merging income, expenditure and sales data 75](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135776)

[Figure 82 - Results from merging income, expenditure and sales data 75](#_Toc72135777)

[Figure 83 - Python code for merging house and consumer price index 76](#_Toc72135778)

[Figure 84 - Results from merging house and consumer price index 76](#_Toc72135779)

[Figure 85 - Python code for merging all datasets 77](#_Toc72135780)

[Figure 86 - Results from merging all datasets 77](#_Toc72135781)

[Figure 87 - Drop unnecessary columns 78](#_Toc72135782)

[Figure 88 - Sort all columns 78](#_Toc72135783)

[Figure 89 - Renaming all columns 79](#_Toc72135784)

[Figure 90 - Results from sorting and reformatting the data 79](#_Toc72135785)

[Figure 91 - Data distribution by year 80](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135786)

[Figure 92 - Data distribution by region 81](#_Toc72135787)

[Figure 93 - Data distribution by age groups 82](#_Toc72135788)

[Figure 94 - Data distribution for annual savings 83](#_Toc72135789)

[Figure 95 - Data distribution for down payment capacity 84](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135790)

[Figure 96 - Data distribution for loan to value ratio 85](#_Toc72135791)

[Figure 97 - Data distribution of average residential sale prices 86](#_Toc72135792)

[Figure 98 - Data distribution for house price index 87](#_Toc72135793)

[Figure 99 - Data distribution for consumer price index 88](#_Toc72135794)

[Figure 100 - Exploration of algorithms in Weka 98](#_Toc72135795)

[Figure 101 - Test design for three models 101](#_Toc72135796)

[Figure 102 - Results from DecisionTable Algorithm 103](#_Toc72135797)

[Figure 103 - Results from SimpleLogistic Algorithm 104](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135798)

[Figure 104 - Loading data from a csv file 106](#_Toc72135799)

[Figure 105 - Attribute’s visualisation 107](#_Toc72135800)

[Figure 106 - Training and test data split 108](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135801)

[Figure 107 - Classifier evaluation selections 108](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135802)

[Figure 108 - Filter selection 108](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135803)

[Figure 109 - Evaluation metrics for model output 109](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135804)

[Figure 110 - Model parameters used for data mining 109](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135805)

[Figure 111 - Part one of the SimpleLogistic algorithm output 111](file:////Users/suchiundevia/Documents/infosys_722_assignment/iteration_03/Iteration_03_Report.docx#_Toc72135806)

[Figure 112 - Part two of the Simple Logistic algorithm output 112](#_Toc72135807)

[Figure 113 - Affordability levels 114](#_Toc72135808)

[Figure 114 - Income and expenditure trend 115](#_Toc72135809)

[Figure 115 - Annual savings and down payment capacity trends 116](#_Toc72135810)

[Figure 116 - Consumer price index and house price index trends 117](#_Toc72135811)

[Figure 117 - Average house sale prices and loan to value ratio trend 118](#_Toc72135812)

[Figure 118 - Annual savings, down payment capacity and loan to value ratio trends 119](#_Toc72135813)

[Figure 119 - Affordability according to age groups 120](#_Toc72135814)

[Figure 120 - Affordability according to region and loan to value ratio 121](#_Toc72135815)

[Figure 121 - Affordability according to down payment capacity and region 122](#_Toc72135816)

[Figure 122 - Effect of income and expenditure on affordability 123](#_Toc72135817)

[Figure 123 - Effect of income and annual savings on affordability 124](#_Toc72135818)

[Figure 124 - Effect of expenditure and annual savings on affordability 125](#_Toc72135819)

[Figure 125 - Effect of loan to value ratio and annual savings on affordability 126](#_Toc72135820)

[Figure 126 - Effect of down payment capacity and loan to value ratio on affordability 127](#_Toc72135821)

[Figure 127 - Affordability by age groups and region 130](#_Toc72135822)

[Figure 128 - Income levels by age groups 130](#_Toc72135823)

[Figure 129 - Attribute Correlations 131](#_Toc72135824)