



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud

Executive Summary

Qussay Al-Qirim

Predicting Regional Growth

For 5G Development

Connecting New Zealand

This report aims to predict New Zealand's growth areas for developing 5G infrastructure. The methods used to reach and support the findings are divided into Linear regression, Cluster and Forecast Analysis.



The Linear regression model highlights the factors contributing most to industrial GDP. Our analysis shows that these are the following factors that play a significant role in industrial GDP and can indicate the economic potential of that area:

* The average number of active bonds
* Number of people aged 40-64 years old
* Total government spending in that area
* Total number of the following businesses in that area
  + Financial and Insurance services
  + Professional, Scientific, and Technical services
  + Rental hiring and Real Estate Services

The cluster analysis shows areas with a similar growth pattern to Auckland's. This analysis highlights Tauranga City as an area with the highest industrial economic growth and the potential to remain a key growth area. This conclusion is further supported by a forecast analysis for the next ten years, which continued to show Tauranga City as a main growing area in the country.

As a result of this analysis, it is recommended for Spark to focus on Tauranga City and explore the economic profile of areas in Tauranga City using the economic factors noted in the linear regression results.

Contents

[1. Introduction 4](#_Toc136622905)

[2. Data 4](#_Toc136622906)

[2.1 Dataset Overview 4](#_Toc136622907)

[2.2 Data Collection 5](#_Toc136622908)

[2.3 Data Construction 5](#_Toc136622909)

[2.4 Data Selection 6](#_Toc136622910)

[2.5 Data Imputation 6](#_Toc136622911)

[2.6 Main Dataset 6](#_Toc136622912)

[3. Methodology and Analysis 7](#_Toc136622913)

[3.1 Data Normalization 7](#_Toc136622914)

[3.2 Data Leakage 7](#_Toc136622915)

[3.3 Methodology Overview 9](#_Toc136622916)

[3.4 Linear Regression Assumptions 9](#_Toc136622917)

[3.4.1 Linearity 9](#_Toc136622918)

[3.4.2 Principal Component Analysis 11](#_Toc136622919)

[3.4.3 Feature Engineering 12](#_Toc136622920)

[3.4.4 Normality 12](#_Toc136622921)

[3.4.5 Modelling 12](#_Toc136622922)

[3.4.5.1 Train & Test Split 12](#_Toc136622923)

[3.4.5.2 Linear Regression Model 13](#_Toc136622924)

[3.4.6 Testing for Multicollinearity 13](#_Toc136622925)

[3.4.7 Lasso Regression 13](#_Toc136622926)

[3.4.8 Performance Measure 13](#_Toc136622927)

[3.5 Time Series Cluster 14](#_Toc136622928)

[3.5.1 Distance Measure 14](#_Toc136622929)

[3.5.2 Clustering Model 14](#_Toc136622930)

[3.5.3 Identify Territorial Authorities with large growth rate. 14](#_Toc136622931)

[3.6 Forecasting 15](#_Toc136622932)

[4. Findings & Discussion 15](#_Toc136622933)

[4.1 Linear Regression Model 15](#_Toc136622934)

[4.1.1 Multicollinearity findings 15](#_Toc136622935)

[4.1.2 Performance measure findings and model of choice 17](#_Toc136622936)

[4.1.3 Lasso Regression Findings 18](#_Toc136622937)

[4.1.4 Linear Regression Coefficients 19](#_Toc136622938)

[4.2 Time Series Clustering 22](#_Toc136622939)

[4.2.1 Areas clustered with Auckland. 22](#_Toc136622940)

[4.2.2 Areas with highest growth rate 23](#_Toc136622941)

[4.3 Forecasting 28](#_Toc136622942)

[5. Limitations 28](#_Toc136622943)

[6. Recommendation and Conclusion 29](#_Toc136622944)

[References: 30](#_Toc136622945)

[Appendix A 35](#_Toc136622946)

[Appendix B 40](#_Toc136622947)

[Appendix C 41](#_Toc136622948)

# Introduction

Spark is New Zealand’s largest telecommunications and digital services provider (Spark, n.d.). It was founded in 1987, and its primary goal is to help all of New Zealand win big in a digital world. In order to continuously achieve that goal, Spark’s corporate strategy for the next three years will focus on two areas; This includes investment in data centres and new technologies to drive growth in Spark’s core markets and building a high-performance culture through an agile framework to deliver better customer experiences (Spark, 2023).

According to the chief executive of Spark, Jolie Hodson, Spark's dual brands, Spark and Skinny, for the mobile market serve different ends of the price spectrum with the ability to target customers with the appropriate product. This ability is expected to improve with the intensification of 5G coverage (2023). Furthermore, Spark’s success requires them to meet the increasing need of customers to communicate and do business, and 5G technology will be a crucial element in that success, and to achieve the corporate strategy, Spark plans to invest $40 - $60 million dollars in 5G infrastructure and $250 - $300 million dollars in the data center market.

As a result, there is great interest for Spark in identifying areas within New Zealand that are experiencing high growth and potential future growth to install Spark’s 5G infrastructure. Therefore, this analysis aims to identify areas of high growth in New Zealand for the next 5 to 10 years. Considering the result, the analysis will construct a single dataset using external sources. Second, a linear regression model is built to understand the effect of the different variables on the Annual Industry GDP. Third, a time series cluster highlights the areas that share a similar annual industry GDP growth pattern to Auckland. Finally, forecasting is performed on the highlighted areas to understand their growth for the next ten years.

# Data

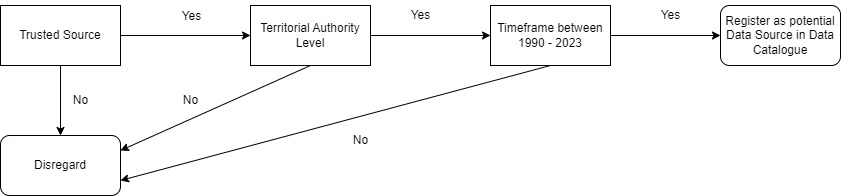
## Dataset Overview

The data for this analysis is collected from various external sources focusing on the Territorial Authorities of New Zealand. The critical growth identifier for these areas is the Regional Annual Industry GDP. Many factors influence industry GDP and are included in the dataset to understand the interaction of those variables with the industry GDP.

## Data Collection

**Figure 2.1**

Data Collection Logic Diagram



The data sources collected follow a specific data quality process before they are identified as a potential data source. The data quality process is broken down into three steps: data source, area parameter, and time frame, as displayed in Figure 2.1. Firstly, the data set source is reviewed to ensure it comes from a trusted location. The data is collected from the following sources: StatsNZ, MBIE (Ministry of Business, Innovation, and Employment), CoreLogic, Infometrics, and Infoshare.

Secondly, the area parameter is on a territorial authority to align with the objective of the analysis. Koordinates define the Territorial Authority as a second tier of local government in New Zealand, a level below the regional council. The sixty-seven territorial authorities comprise 13 city councils, including Auckland Council, 53 district councils, and the Chatham Islands Territory (2021).

Lastly, the time frame of the data set is between (1990 – 2023). Once a data set complies with the three data quality steps, the data set is registered as a potential data set for further analysis.

## Data Construction

Due to sourcing multiple different data sets from different sources, the data sets collected are transformed to follow a uniform structure with two primary keys. Each territorial authority has its unique area code identifier, which will act as a primary key, along with the year, to merge the different data sets into a single dataset. The combined single dataset is composed of (88) variables and (2,701) observations, ranging from the years (1991 – 2048).

## Data Selection

An analysis of the percentage of missing values for each column in the dataset reveals eight variables with a missing values percentage of seventy percent or greater. These variables were deemed unnecessary due to the lack of information they present and removed from the dataset. Furthermore, the percentage of missing values for rows with the territorial authorities Chatham Island and Area Outside Territory were fifty-two and fifty-four percent for the latter. As such, rows with both territorial authorities provided scarce information and were removed from the dataset. Due to the missing value analysis, eight variables and sixty-four observations were removed.

Furthermore, most variables revealed missing data before the year (2000) and after the year (2022) and were removed from the dataset. Additionally, due to the presence of COVID during the years (2020) to (2022), values from these variables are skewed and removed from the dataset. As a result, the time frame of the dataset ranges from (2000 – 2019). The resulting dataset is eighty (80) variables and (1,320) observations.

## Data Imputation

The remaining missing values were imputed using Kalman smoothing and Backcasting techniques. With Kalman smoothing, it is assumed that not all values are available, and therefore the goal of smoothing is to approximate the missing values given the known measurements (Barratt & Boyd, 2020). This technique is used for columns with missing values in the middle of the time series. Furthermore, backcasting is the estimation of x(t) for t ∈ {−M,−M + 1,−M + 2, ... − 2,−1} and is assumed that the available data is time-series related without missing values in the middle. The technique reverses the time series and fits an ARIMA model to produce forecasts for the past years (Caporin & Sartore, 2006). These techniques were used interchangeably based on the location of missing values within a column.

# Methodology and Analysis

## Data Normalization

Feature scaling is essential in machine learning and statistical analysis, such as correlation and principal component analysis. Feature scaling is applied to the primary dataset to make the data clean, noise-free, and consistent (Mohamad & Usman, 2013). In addition, as explained by Alshaher, feature scaling makes all the variables play a role in statistical and machine learning models so that no single variable impacts the model due to its large value (2021).

Data normalization standardizes the values of the primary dataset by converting them into a specific range. The data normalization techniques used for the analysis are Min-Max and Log transformation. The Min-Max method provides an easy way to compare variables with different scales and is used to normalize the primary dataset to perform a correlation and principal component analysis. In comparison, log transformation is used for the linear regression modeling in compliance with the normality assumption of a linear regression model.

Min-Max normalization transforms the raw data into a value between 0 and 1 (Mohamad & Usman, 2013). The lowest value is set to 0, and the highest is set to 1. The normalized value is computed in (3.1) (Google, 2022).

|  |  |
| --- | --- |
|  | 3.1 |

Log scaling computes the log of the values in the main dataset and is useful when outliers exist in the dataset. Log scaling compresses a wide range of values to a narrow range of values and is seen in (3.2) (Google, 2022), transforming the distribution of the values to a normal distribution.

|  |  |
| --- | --- |
|  | 3.2 |

## Data Leakage

As explained by Singh, data leakage occurs when the data used in the training process contains information about what the model is trying to predict (Singh, 2022). This could lead to overly optimistic predictions from a linear regression model along with poor results when the model is implemented on the test data. This produces completely invalid predictive models (Brownlee, 2016). In the case of this analysis, the objective is to understand the underlying effects of different variables on the variable to predict, which is the Annual Industry GDP. In turn, within the dataset, nineteen industry GPD variables explain the contribution of different business sectors to the total annual industry GDP, seen in Table 3.1, that are highly linked to Annual Industry GDP and are removed from the dataset. The number of variables in the data set is reduced from eighty (80) to sixty-one (61).

**Table 3.1:**

Removed Variables to avoid Data Leakage

|  |  |
| --- | --- |
| **No.** | **Variable:** |
| 1 | Regional Industry GDP-Wholesale Trade ($ M) |
| 2 | Regional Industry GDP-Transport, Postal And Warehousing ($ M) |
| 3 | Regional Industry GDP-Retail Trade ($ M) |
| 4 | Regional Industry GDP-Rental, Hiring And Real Estate Services ($ M) |
| 5 | Regional Industry GDP-Public Administration And Safety ($ M) |
| 6 | Regional Industry GDP-Professional, Scientific And Technical Services ($ M) |
| 7 | Regional Industry GDP-Owner Occupied Property Operation ($ M) |
| 8 | Regional Industry GDP-Manufacturing ($ M) |
| 9 | Regional Industry GDP-Information Media, Telecommunications And Other Services ($ M) |
| 10 | Regional Industry GDP-Health Care And Social Assistance ($ M) |
| 11 | Regional Industry GDP-GST On Production, Import Duties and Other Taxes ($ M) |
| 12 | Regional Industry GDP-Forestry, Fishing, Mining, Electricity, Gas, Water & Waste Services ($M) |
| 13 | Regional Industry GDP-Food And Beverage Services ($ M) |
| 14 | Regional Industry GDP-Financial And Insurance Services ($ M) |
| 15 | Regional Industry GDP-Education And Training ($ M) |
| 16 | Regional Industry GDP-Construction ($ M) |
| 17 | Regional Industry GDP-Agriculture ($ M) |
| 18 | Regional Industry GDP-Administrative And Support Services ($ M) |
| 19 | Regional Industry GDP-Accommodation ($ M) |

## Methodology Overview

The analysis is divided into three components: linear regression, time series clustering, forecasting. Linear regression is performed to predict annual industry GDP and understand the underlying impact of different variables on the annual industry GDP. Secondly, time series clustering is utilized to cluster regions with similar industry GDP growth patterns as Auckland and highlight the higher growing regions than Auckland. Lastly, the forecasting method is used to understand the growth of the regions highlighted in the time series clustering for the next ten years.

## Linear Regression Assumptions

As the Boston University School of Public Health (2016) highlighted, four assumptions are associated with a linear regression model to ensure a good predictive model. First, the relationship between the dependent and independent variables is linear. Second, the variance of the residuals is the same for any value of the dependent variable. Third, observations are independent of each other. Fourth, the independent variables are normally distributed.

Annual Industry GDP is the dependent variable used and is the focus of the regression analysis. With the data set containing eighty-one (81) variables, a series of dimension reduction analyses are performed to test the assumptions of linear regression and select variables with the most significant impact on annual industry GDP. These dimension reduction analyses are broken down into four tests: correlation analysis, principal component analysis, variance inflation factor, and lasso regression.

### Linearity

To understand the linear relationship between the annual industry GDP and the rest of the variables in the dataset, a correlation analysis is performed with a focus on the Pearson correlation coefficient. The Pearson correlation coefficient is a statistical metric that measures the strength and direction of a linear relationship between two variables (Benesty et al., 2008). The value in Pearson's correlation coefficient is between a negative one (-1) and a positive one (+1). As Saputra Rangkuti et al. explained, a value of -1 means a perfect negative relationship, a +1 means a perfect positive correlation, and 0 means no linear correlation between both variables (2018).

For the analysis, a min-max normalization process was performed on the dataset to prevent variables with large values from skewing the analysis. After this, a Pearson correlation analysis was performed. Variables with a correlation coefficient greater than or equal to 0.5 and less than 0.9 and statistically significant are selected from the dataset for further analysis. The 0.5 coefficient threshold was used due to the large number of variables in the dataset and interest in features that strongly and positively influenced annual industry GDP; in addition, variables greater than 0.9 were not selected as this presented a near-perfect correlation with annual industry GDP, the dependent variable in the linear regression and would skew the prediction of the regression model. From the sixty-one variables in the analysed dataset, twenty-two variables met the correlation threshold, as seen in Table 3.2.

**Table 3.2:**

Correlation coefficient of different variables with Annual Industry GDP

|  |  |  |
| --- | --- | --- |
| **Variable** | **Pearson Correlation** | |
| Housing-Average Number of Active Bonds | | 0.8646\* |
| Business Demography - Financial and Insurance Services - Enterprise | | 0.8602\* |
| Property-Regional Median Home Price Index ($) | | 0.8298\* |
| Population Estimation - 40-64 Years Old | | 0.8234\* |
| Business Demography – Rental Hiring and Real Estate Services - Enterprise | | 0.8106\* |
| Business Demography – Professional, Scientific and Technical Services - Enterprise | | 0.8039\* |
| Local Authority Financial Stat- Total Government Spending | | 0.7803\* |
| Population Estimation - Total | | 0.7312\* |
| Business Demography – Other Services - Enterprise | | 0.7251\* |
| Business Demography – Administrative and Support Services - Enterprise | | 0.7171\* |
| Business Demography-Annual Total Employee | | 0.704\* |
| Business Demography - Construction - Enterprise | | 0.6953\* |
| Business Demography – Construction - Employee | | 0.6797\* |
| Business Demography- Accommodation and Food Services - Enterprise | | 0.6684\* |
| Business Demography-Annual Total Enterprise | | 0.6652\* |
| Business Demography – Health Care and Social Assistance - Enterprise | | 0.6072\* |
| Business Demography – Accommodation and Food Services - Employee | | 0.6062\* |
| Business Demography – Other Services - Employee | | 0.564\* |
| Business Demography – Health Care and Social Assistance - Employee | | 0.563\* |
| Business Demography - Professional, Scientific and Technical Services - Employee | | 0.536\* |
| Business Demography – Education and Training - Employee | | 0.5234\* |
| Business Demography – Public Administration and Safety - Employee | | 0.5154\* |
| \* P < 0.05 | | |

### Principal Component Analysis

Principal component analysis (PCA) is a factor reduction technique used to reduce the number of features from a dataset that consists of interrelated variables. The goal of PCA is to compress the size of the data set by keeping only the critical information and identifying the contribution of observation to a component (Abdi & Williams, 2010). For PCA to work correctly, there must be some degree of correlation among the variables (Kolekar, 2019), and the data must be standardized (Ringnér, 2008).

Therefore, PCA will be performed on the highly correlated variables highlighted in Table 3.2 and standardized using the min-max method. This analysis's key objective is understanding each variable's contribution to the component and removing variables that fall below the average contribution.

As seen in Appendix C, the chart shows the average contribution of each variable. The red dotted line represents the contribution if the variables were uniform. A helpful method is to use the dotted line as a baseline and focus on variables with contributions above the baseline (Abdi & Williams, 2010). As a result, the following variables in Table 3.3 show a contribution greater than the threshold and are used for the linear regression analysis.

**Table 3.3:**

Variables with high contribution in PCA and selected for Linear Regression

|  |  |
| --- | --- |
| **No.** | **Variable** |
| 1 | Business Demography – Financial and Insurance Services - Enterprise |
| 2 | Housing – Average Number of Active Bonds |
| 3 | Business Demography – Rental Hiring and Real Estate Services - Enterprise |
| 4 | Population Estimation – 40-64 Years Old |
| 5 | Business Demography – Construction - Enterprise |
| 6 | Property-Regional Median Home Price Index ($) |
| 7 | Business Demography – Professional, Scientific and Technical Services - Enterprise |
| 8 | Business Demography – Annual Total Enterprise |
| 9 | Local Authority Financial Stat – Total Government Spending |
| 10 | Business Demography – Construction - Employee |
| 11 | Business Demography – Administrative and Support Services - Enterprise |
| 12 | Business Demography – Annual Total Employee |

### Feature Engineering

Multicollinearity occurs when two or more independent variables in a data frame highly correlate in a regression model (Bhandari, 2020). A series of feature engineering steps can assist in reducing the level of multicollinearity in the data set. Feature engineering combines two variables into one variable, allowing us to capture information in both variables (Bhandari, 2020).

The growth percentage of our variables of interest in Tables 3.2 and 3.3 is calculated. This feature engineering is done by dividing the value of that variable by the total for its industry. For example, Business Demography – Construction - Enterprise growth rate equals Business Demography – Construction - Enterprise Count divided by Business Demography-Annual Total Enterprise Count.

### Normality

Normality is the assumption that the values in the data set follow a normal distribution. The purpose of a QQ (quantile-quantile) plot is to visualize the distribution of the values. When looking at a QQ plot, a normally distributed dataset will see points match up a long straight line (Wang et al., 2016).

A QQ plot analysis on the primary dataset shows that values do not follow a normal distribution, as seen in Appendix A. As a result, log transformation is applied to the values in the dataset to follow a normal distribution.

### Modelling

#### Train & Test Split

Having logged the data set and according to best modelling practice, the data set is split into a training and test data set. The training data is used to train the model before exposing the model to unseen data, which is the test data. The practice follows an 80% to 20% ratio split between the training and test data (Google, 2019).

#### Linear Regression Model

Two linear regression models are built to compare and proceed with the better model. A pre-PCA linear regression model with annual industry GDP as the dependent variable includes the independent variables before they are removed, as seen in Table 3.2, and a post-PCA linear regression model where several variables have been removed. The post-PCA linear regression model has the same dependent variable, annual industry GDP, and with the independent variables seen in Table 3.3.

### Testing for Multicollinearity

The variance inflation factor (VIF) is a method used to measure the induced collinearity between the independent variables in the regression model (Craney & Surles, 2002). This measure is used once a linear regression model is built to understand the interaction of independent variables among each other. As highlighted by Craney & Surles, there is no formal cut-off value to determine when a VIF is too large, but there are suggestions for a cut-off value of 5 or 10 (2002). For this analysis, a cut-off value of ten will be assigned when analyzing the VIF values.

### Lasso Regression

Lasso regression is a technique to reduce model complexity and prevent overfitting (Saptashwa Bhattacharyya, 2018). It is a type of regularization that shrink coefficients and can become zero. Those coefficients with a value of zero can then be removed. In other words, it is another technique utilized to reduce the multicollinearity effect in the linear regression model and acts as another feature selection method.

### Performance Measure

In order to determine the better linear regression model to proceed with, several measures are used to compare the two linear regression models, Pre and Post PCA linear regression models. The first measure is to compare the R squared of both models. As defined by Miles, R squared measures the model’s goodness of fit (2005) and how well the regression line approximates the actual data. Second, the mean absolute error (MAE) is another widely used measure in model evaluations (Chai & Draxler, 2014). Mean absolute error is the average of all absolute errors, which is the difference between predicted and actual values (Glen, 2016). Therefore, a lower MAE value indicates a better model fit. Lastly, RMSE values will be compared between both models to indicate a better model fit and, similarly, the lower the RMSE value, the better the model fit.

## Time Series Cluster

Time series clustering is a method of clustering that works with dynamic data. The distance measure and clustering model are the most important elements to consider for time series clusters (Sardá-Espinosa, 2019).

### Distance Measure

Euclidean distance is the most common approach when selecting a distance to classify data in data clustering models. However, with time series clustering, this poses a few problems. Euclidean distance is very sensitive to scaled and missing data (Heka.ai, 2022), and the time frame must be consistent across the different groups due to Euclidean distance matching on a one-to-one basis (Denyse, 2021). A proposed solution to this problem is Dynamic Time Warping (DTW) as a distance measure. DTW is an elastic measurement that can handle different time frames and finds the best matching path between two-time series (Luo et al., 2023). DTW focuses on matching similar patterns between time series regardless of time.

### Clustering Model

In this report, Hierarchical clustering is used along with DTW as a distance measure. Hierarchical clustering does not require a specific number of clusters to be specified in the algorithm and is powerful in visualization (Luo et al., 2023). This analysis aims to cluster the different territorial authorities in New Zealand based on the annual industry GDP growth rate and observe the areas that share a similar growth pattern with Auckland.

### Identify Territorial Authorities with large growth rate.

After identifying the areas grouped with Auckland, the total average growth rate from 2000 – 2019 is calculated for each area closely related to Auckland’s growth rate pattern and to the areas that experienced the highest growth rate, within that cluster, towards the end of 2019.

## Forecasting

Forecasting is performed on the highlighted clusters to predict future trends of annual industry GDP for the next ten years. The Autoregressive integrated moving average (ARIMA) approach uses a statistical model that uses time series to predict future trends (Hayes, 2022).

# Findings & Discussion

## Linear Regression Model

### Multicollinearity findings

**Table 4.1:**

Multicollinearity VIF values for Pre-PCA Linear Regression model

|  |  |  |
| --- | --- | --- |
| **Pre-PCA Linear Regression Model** | | |
| **Variable** | **VIF Value** | |
| Housing - Average Number of Active Bonds - Count | | 39.65 |
| Population – Estimation Total - Count | | 28.87 |
| Percent In Total Enterprise Count – Professional, Scientific and Technical Services | | 12.73 |
| Percent In Total Enterprise Count – Other Services | | 12.51 |
| Percent In Total Enterprise Count – Health Care and Social Assistance | | 12.42 |
| Percent In Total Enterprise Count – Administrative and Support Services | | 9.69 |
| Percent In Total Enterprise Count – Accommodation and Food Services | | 8.25 |
| Percent In Total Enterprise Count – Construction | | 7.61 |
| Local Authority Financial Stat – Total Government Spending ($k) | | 6.16 |
| Property-Regional Median Home Price Index ($) | | 4.6 |
| Percent In Total Employee Count – Accommodation and Food Services | | 4.48 |
| Percent In Total Employee Count – Health Care and Social Assistance | | 3.67 |
| Percent In Total Employee Count – Professional, Scientific and Technical Services | | 3.6 |
| Percent In Total Enterprise Count – Financial and Insurance Services | | 3.54 |
| Percent In Total Employee Count – Other Services | | 3.38 |
| Percent In Total Population – (40 – 64) Years Old | | 3.02 |
| Percent In Total Enterprise Count – Rental Hiring and Real Estate Services | | 2.65 |
| Percent In Total Employee Count – Construction | | 1.95 |
| Percent In Total Employee Count – Education and Training | | 1.88 |
| Percent In Total Employee Count – Public Administration and Safety | | 1.61 |
| Percent in Annual Enterprise Increase | | 1.12 |
| Percent in Annual Employee Increase | | 1.1 |
|  | | |

**Table 4.2**

Multicollinearity VIF values for Post-PCA linear regression model

|  |  |  |
| --- | --- | --- |
| **Post PCA Regression Model** | | |
| **Variable** | **VIF Value** | |
| Housing – Average Number of Active Bonds – Count | | 8.82 |
| Percent In Total Enterprise Count – Professional, Scientific and Technical Services | | 7.76 |
| Percent In Total Enterprise Count – Administrative and Support Services | | 7.22 |
| Percent In Total Enterprise Count – Construction | | 5.93 |
| Local Authority Financial Stat – Total Government Spending ($k) | | 4.97 |
| Percent In Total Enterprise Count – Financial and Insurance Services | | 2.85 |
| Property-Regional Median Home Price Index ($) | | 2.74 |
| Percent In Total Population – (40 – 64) Years Old | | 2.23 |
| Percent In Total Enterprise Count – Rental Hiring and Real Estate Services | | 2.1 |
| Percent In Total Employee Count – Construction | | 1.66 |
| Percent in Annual Enterprise Increase | | 1.1 |
| Percent in Annual Employee Increase | | 1.09 |

Provided the cut-off VIF value of ten, as explained in the methodology section, a comparison is made between both regression models. As seen in Table 4.1, the pre-PCA linear regression model contains several variables that suffer from high multicollinearity values above 10. In contrast, including only the variables highlighted from the PCA solves the multicollinearity issue, as seen in Table 4.2, where the highest VIF value is 8.82, below the cut-off value.

### Performance measure findings and model of choice

**Table 4.3**

Performance measures for post and pre-PCA linear regression models

|  |  |  |  |
| --- | --- | --- | --- |
|  | **R squared** | **RMSE** | **MAE** |
| Pre-PCA Linear Regression Model | 0.97 | 0.22 | 0.18 |
| Post-PCA Linear Regression Model | 0.94 | 0.31 | 0.26 |

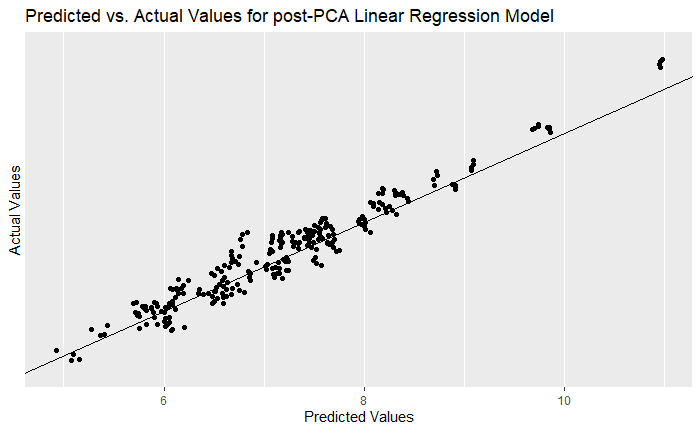
As seen in Table 4.3, the pre-PCA linear regression model has higher R-squared, lower RMSE and MAE values than the post-PCA linear regression model. Furthermore, plotting the actual vs. predicted values on a chart, as displayed in Figures 4.1 and 4.2, shows all the points closely grouped around the regressed diagonal line. This means that the predicted values are reasonably close to the actual values. For example, if the actual value is eight, the predicted value is close to eight. Although pre-PCA model measures showed slightly better values than the post-PCA model, the post-PCA model solves the issue of multicollinearity that plays a significant role in overfitting the model and producing inaccurate predictions. Therefore, characterized by its high R-squared value, good model fit as seen in Figure 4.2, and low multicollinearity, the post-PCA linear regression model will be the preferred model for further analysis.

**Figure 4.1**

A picture containing text, screenshot, line, plot

Description automatically generated

**Figure 4.2**



### Lasso Regression Findings

After narrowing down to the regression model of choice, Lasso regression is performed on the post-PCA model to review the coefficients and decide whether further variable reduction is needed. As seen in Table 4.4, there are no variables whose coefficient is reduced to zero, indicating that all variables play a major role in the regression model. As such, removing any further variables from the linear regression model of choice is unnecessary.

**Table 4.4**

Lasso Regression Coefficient Results

|  |  |
| --- | --- |
| **Post PCA Model** | |
| **Variable** | **Coefficient** |
| Housing – Average Number of Active Bonds – Count | 0.735 |
| Percent In Total Population – (40 – 64) Years Old | 0.210 |
| Local Authority Financial Stat – Total Government Spending ($k) | 0.180 |
| Percent In Total Enterprise Count – Financial and Insurance Services | 0.131 |
| Percent In Total Enterprise Count – Professional, Scientific and Technical Services | 0.095 |
| Percent In Total Enterprise Count – Rental Hiring and Real Estate Services | 0.087 |
| Percent In Total Enterprise Count – Administrative and Support Services | 0.007 |
| Percent in Annual Employee Increase | -0.002 |
| Percent in Annual Enterprise Increase | -0.012 |
| Percent In Total Employee Count – Construction | -0.023 |
| Property – Regional Median Home Price Index ($) | -0.112 |
| Percent In Total Enterprise Count – Construction | -0.498 |
|  |  |

### Linear Regression Coefficients

**Table 4.5**

Post-PCA linear regression coefficients

|  |  |  |
| --- | --- | --- |
| **Post PCA Model** | |  |
| **Variable** | **Coefficient** | **Std. Error** |
| Housing – Average Number of Active Bonds – Count | 0.74\*\*\* | 0.020 |
| Percent In Total Population – (40 – 64) Years Old | 0.32\* | 0.160 |
| Local Authority Financial Stat – Total Government Spending ($k) | 0.18\*\*\* | 0.019 |
| Percent In Total Enterprise Count – Financial and Insurance Services | 0.13\*\*\* | 0.021 |
| Percent In Total Enterprise Count – Professional, Scientific and Technical Services | 0.1\*\* | 0.036 |
| Percent In Total Enterprise Count – Rental Hiring and Real Estate Services | 0.1\*\* | 0.038 |
| Percent In Total Enterprise Count – Administrative and Support Services | 0.03 | 0.041 |
| Percent in Annual Employee Increase | -0.002 | 0.005 |
| Percent in Annual Enterprise Increase | -0.01\*\* | 0.004 |
| Percent In Total Employee Count – Construction | -0.02 | 0.036 |
| Property – Regional Median Home Price Index ($) | -0.13\*\*\* | 0.030 |
| Percent In Total Enterprise Count – Construction | -0.53\*\*\* | 0.049 |
| \*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05 | |  |

Displayed in table 4.5, of the 12 variables in the model, nine are highlighted as statistically significant and impact annual industry GDP. From those deemed as statistically significant, average number of active bonds played an important role in contributing to the annual GDP. Active bonds are a form of debt securities issued by the local governments to raise money for affordable housing development projects (Chen, 2022). The phenomena behind such analysis could be attributed to a few reasons as explained by Asian Development Bank. Housing bonds act as an additional funding source for mortgages increasing the number of home buyers. Additionally, an increase supply in housing leads to an increased demand for construction and labour, increases the population, in turn increasing income and boosting economic growth (2019).

Furthermore, percent of total population aged 40 – 64 years old carried a large weight in the model. Jonathan Law defines that the expansion of economic growth is expressed in terms of the increase of national income (2009). As seen in figure 4.3, data collected with regards to the average weekly income by age from statsNZ (2019) shows that the age group 40 to 64 years old have the largest amount of average total weekly income from years 2000 – 2022, further supporting the regression results.

**Figure 4.3**

Total Average Weekly Income by Age Structure in New Zealand

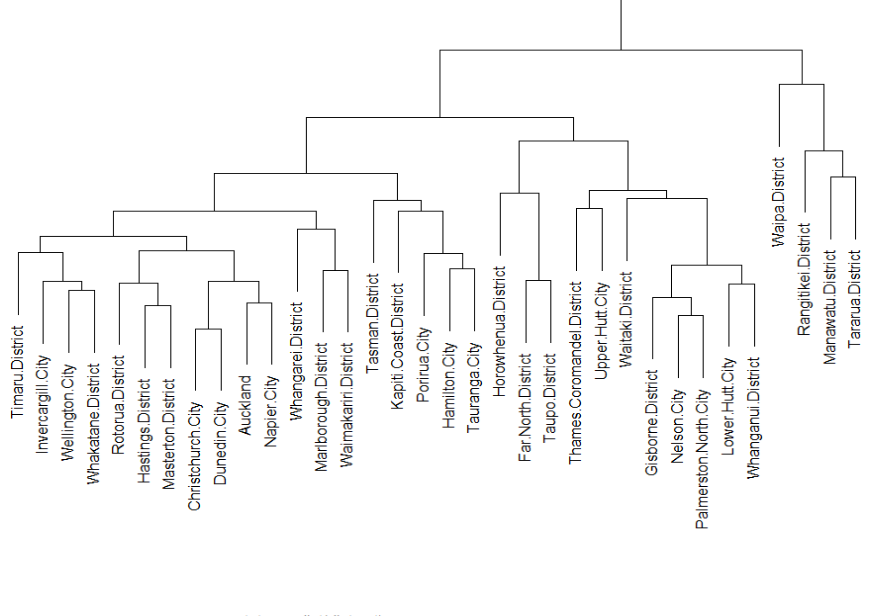
Moreover, government spending plays a major role in increasing the annual GDP. This finding is backed through an insight into the NZ economy by Robertson in which he mentioned, an increased government spending induces expenditure across the economy which in turn will drive economic growth (Robertson, 2019).

## Time Series Clustering

### Areas clustered with Auckland.

**Figure 4.4**

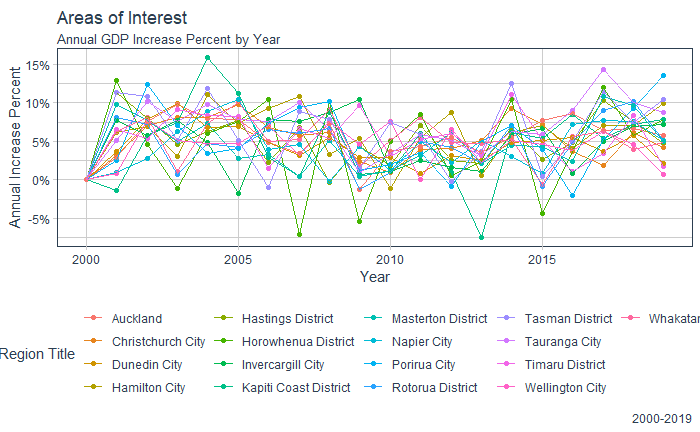
Dendrogram of Territorial Authorities Clustered with Auckland



After performing a time series clustering, several areas are clustered with Auckland, as shown in Figure 4.4. Areas that are located within the red box, as displayed in Figure 4.4, are the areas that display the most similar industry annual GDP growth pattern as Auckland. Additionally, as seen in Figure 4.5, three more areas (Porirua City, Tasman District, Hamilton City and Tauranga City) are within the same cluster as Auckland and do not lie within the red box but show the highest industry GDP growth towards the end of 2019. As a result, these areas will be the focus when calculating the total average annual industry growth rate per area.

**Figure 4.5**

Time series analysis areas clustered with Auckland based on annual industry GDP growth Rate.



### Areas with highest growth rate

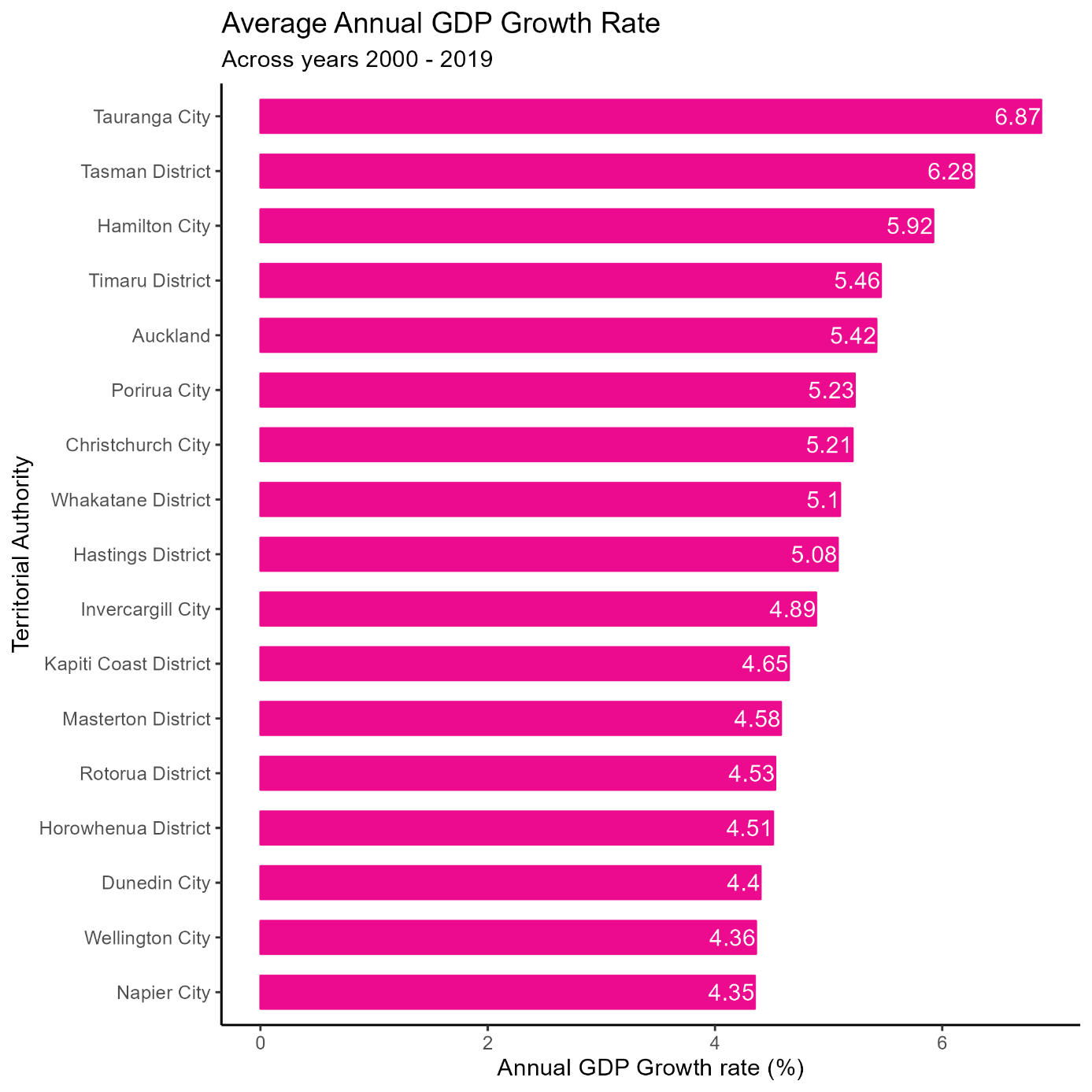
**Table 4.6**

Average annual GDP growth rate per Territorial Authority

|  |  |
| --- | --- |
| **Region title** | **Average annual industry GDP growth rate** |
| Tauranga City | 6.87% |
| Tasman District | 6.28% |
| Hamilton City | 5.92% |
| Timaru District | 5.46% |
| Auckland | 5.42% |
| Porirua City | 5.23% |
| Christchurch City | 5.21% |
| Whakatane District | 5.10% |
| Hastings District | 5.08% |
| Invercargill City | 4.89% |
| Kapiti Coast District | 4.65% |
| Masterton District | 4.58% |
| Rotorua District | 4.53% |
| Horowhenua District | 4.51% |
| Dunedin City | 4.40% |
| Wellington City | 4.36% |
| Napier City | 4.35% |
| Note. The average annual GDP growth rate column is sorted in descending order. | |

**Figure 4.6**

Average annual GDP growth rate from years 2000 - 2019

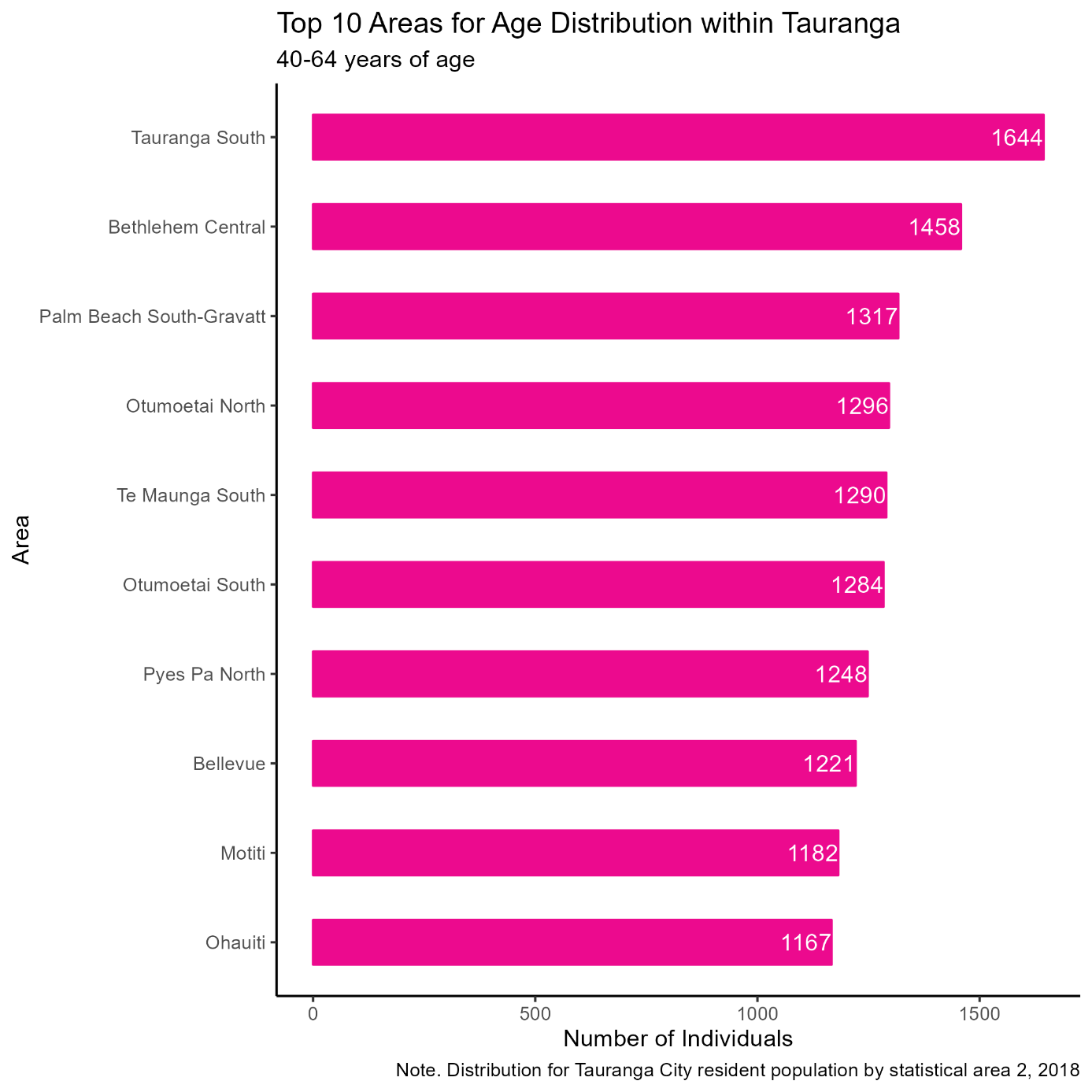


After identifying the regions of interest from the cluster analysis, further investigation is performed to understand the highest-performing region. The average of the annual industry GDP growth rate across nineteen years from 2000 – 2019 for those areas is calculated. As seen from Table 4.6 and visualized in figure 4.6 , Tauranga City appears at the top and above Auckland, with the largest average annual industry GDP growth rate of 6.87%. Interestingly, this analysis aligns with Infometrics’s regional economic profile of New Zealand, in which it was stated that Tauranga City’s GDP grew by 8.1% in 2022, with New Zealand's GDP growing by 5.3% during the same year (Infometrics, 2023).

Furthermore, a deeper dive into Tauranga city on a SA2 presents further opportunities to capitalise. Utilizing the variables of heavy weight observed from the regression model earlier, several areas within Tauranga city appear that could be given priority such as Tauranga South and Tauranga Central, as seen in figures 4.7 and 4.8. Additionally, new community projects are in development to accommodate the projected increase in population growth (Tauranga City Council, n.d.). Communities such as Te Tumu and Tauriko West are the largest community projects which present opportunities to invest 5G infrastructure.

**Figure 4.7**

Age Distribution within Tauranga City



**Figure 4.8**

Professional, Scientific and Technical Services distribution within Tauranga City

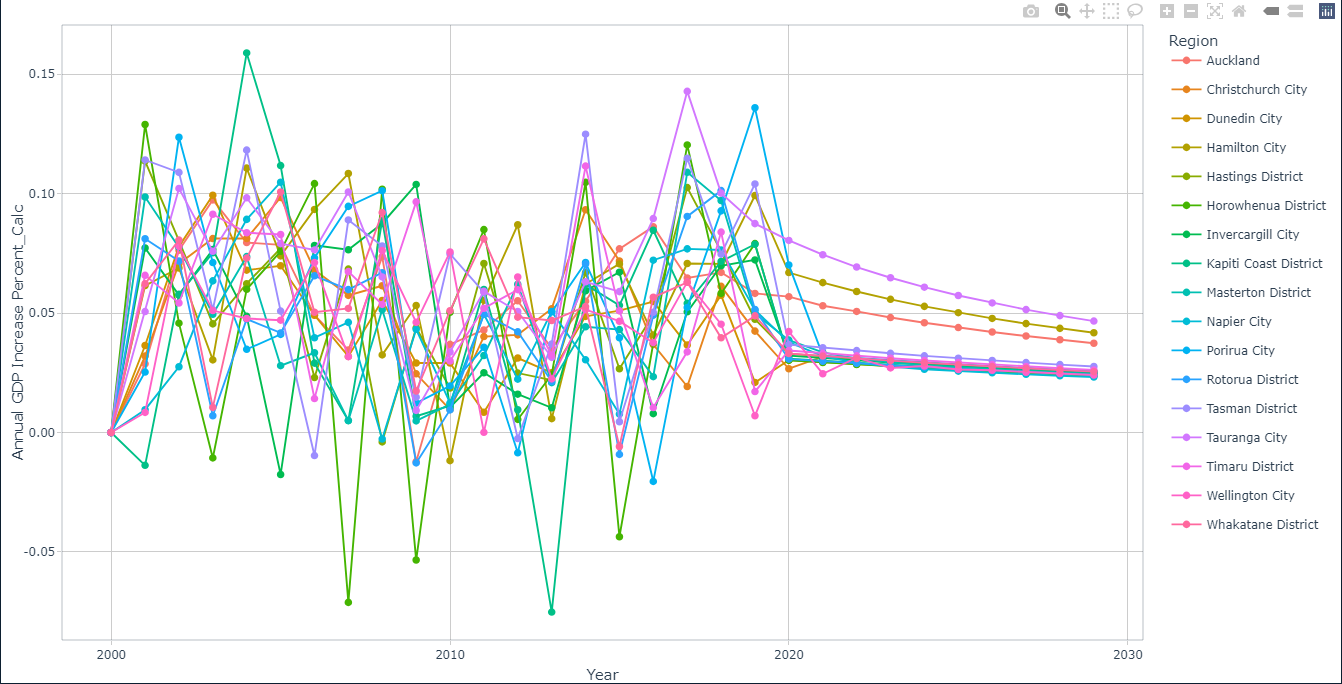
A picture containing text, screenshot, number, plot

Description automatically generated

## Forecasting

**Figure 4.9**

Forecasting Annual Industry GDP growth rate per area



As shown in Figure 4.9, the Annual industry GDP growth rate for areas mentioned in Table 4.6 is forecasted for the next ten years. It can be observed that Tauranga City remains the area to experience high economic growth.

# Limitations

The dataset contains many variables that influence industry GDP, but these variables are a small representation of factors that could influence the industry annual GDP, narrowing the scope of the analysis. Furthermore, the analysis covers the territorial authorities in New Zealand and will require a more granular approach to understand the economic profile of that area, for example, the use of SA2 levels.

Lastly, the dataset does not include values during the COVID period (2020 – 2022) which introduced major economic and technological disruptions, and therefore might not fully reflect the complete current reality when forecasting for future years.

# Recommendation and Conclusion

From the analysis, it is recommended that Spark capitalise on the following opportunities during their 5G development. Spark should target Tauranga city as the territorial authority of focus for further 5G development infrastructure. According to Infometrics, economic growth in Tauranga City averaged 4.8% per year over the last 10 years compared with an average of 3.0% per year nationally (2022).

Furthermore, Infometrics highlights professional, scientific & tech services, a variable of interest from the regression model, as the largest contributor to overall growth in Tauranga City, growing 17.9% over the year 2022 (2022). To further back the analysis, results determined from data collected by Tauranga city council, Spark should prioritise Tauranga Central and Tauranga South as areas of focus within Tauranga city as seen in figures 4.6 and 4.7, due to their significant weight in contributing to annual GDP.

Lastly, Tauriko West is highlighted by Tauranga City council as a large urban development with an estimated 3,000 to 4,000 new homes starting in 2025. Additionally, Te Tumu will provide housing for an additional 15,500 people in 2025 (Tauranga City Council, n.d.). As a result, it is recommended that Spark reach out to Tauranga City Council and government officials to gain additional insight into the development plans for Te Tumu and Tauriko West, this will assist in efficiently rolling out the 5G infrastructure into those areas.

In this report, we analysed the factors that contribute the most to industry GDP and followed by discovering territorial authorities that share a similar growth pattern to Auckland. Lastly, the analysis was followed with a 10-year forecast on the Areas of interest to notice any changes in the future. All the above findings from the cluster and forecast analysis pointed towards Tauranga City as a territorial authority to experience the highest industry economic growth.

With these findings, Spark should focus on the potential that Tauranga City could deliver and prioritise its 5G investment fund towards building 5G infrastructure in that area. Having highlighted the factors that contribute the most to industry GDP, it would provide useful insight to understand the profile of Tauranga City using those factors and get an idea of which areas within Tauranga City experience the highest growth for those economic factors.

In conclusion, the approach's feasibility has been supported by the report’s findings and backed by findings from external statistical parties such as Infometrics and by industry professionals from Spark. With these findings, we hope to provide Spark with a direction to pursue its 5G investment backed by a data-driven approach.

# References:

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. <https://doi.org/10.1002/wics.101>

Alshaher, H. (2021). *Studying the Effects of Feature Scaling in Machine Learning*. Www.proquest.com. <https://www.proquest.com/docview/2621053145?pq-origsite=gscholar&fromopenview=true>

Asian Development Bank. (2019). Developing the Housing Bond Market. In *Asia Bonds Online*. https://asianbondsonline.adb.org/documents/abm/abm\_jun\_2019\_developing\_housing\_bond\_market.pdf

Barratt, S., & Boyd, S. (2020). *Fitting a Kalman Smoother to Data*. <https://doi.org/10.23919/acc45564.2020.9147485>

Benesty, J., Chen, J., & Huang, Y. (2008). On the Importance of the Pearson Correlation Coefficient in Noise Reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(4), 757–765. <https://doi.org/10.1109/tasl.2008.919072>

Bhandari, A. (2020, March 19). *Multicollinearity | Causes, Effects and Detection Using VIF (Updated 2023)*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#What_Is_Multicollinearity>?

Boston University School of Public Health. (2016, January 6). *Simple Linear Regression*. Sphweb.bumc.bu.edu. <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html>

Brownlee, J. (2016, August 1). *Data Leakage in Machine Learning*. Machine Learning Mastery. <https://machinelearningmastery.com/data-leakage-machine-learning/>

Caporin, M., & Sartore, D. (2006). Methodological Aspects of Time Series Back-Calculation. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.950923>

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions*, *7*(1), 1525–1534. <https://doi.org/10.5194/gmdd-7-1525-2014>

Chen, J. (2022, April 26). *Housing Bonds Definition*. Investopedia. <https://www.investopedia.com/terms/h/housingbonds.asp>

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cut-off Values. *Quality Engineering*, *14*(3), 391–403. <https://doi.org/10.1081/qen-120001878>

Denyse. (2021, August 9). *Time Series Clustering — Deriving Trends and Archetypes from Sequential Data*. Medium. <https://towardsdatascience.com/time-series-clustering-deriving-trends-and-archetypes-from-sequential-data-bb87783312b4>

Google. (2019). *Training and Test Sets: Splitting Data  |  Machine Learning Crash Course*. Google Developers. <https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data>

Google. (2022, July 18). *Normalization*. Google Developers. <https://developers.google.com/machine-learning/data-prep/transform/normalization>

Hayes, A. (2022). *Autoregressive Integrated Moving Average (ARIMA)*. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical

Heka.ai. (2022, June 9). *Time series clustering*. Medium. <https://heka-ai.medium.com/time-series-clustering-b84bcaaa63ac>

Infometrics. (2022). *Tauranga City Annual Economic Profile 2022*. Infometrics.co.nz. <https://ecoprofile.infometrics.co.nz/Tauranga%2BCity/PDFProfile>

Infometrics. (2023). *Regional Economic Profile of Tauranga City 2022*. Infometrics.co.nz. https://ecoprofile.infometrics.co.nz/Tauranga%20City/PDFProfile

Kolekar, B. P. P. and P. (2019). Chapter 12 PCA | BioSakshat - Free Study Materials. In *biosakshat.github.io*. <https://biosakshat.github.io/pca.html>

*Koordinates*. (2021, November 30). Datafinder.stats.govt.nz; StatsNZ. <https://datafinder.stats.govt.nz/layer/106668-territorial-authority-2022-generalised/>

Law, J. (2009). *A Dictionary of Business and Management*. Oxford University Press. <https://doi.org/10.1093/acref/9780199234899.001.0001>

Luo, Z., Zhang, L., Liu, N., & Wu, Y. (2023). *Time series clustering of COVID-19 pandemic-related data*. <https://doi.org/10.1016/j.dsm.2023.03.003>

Miles, J. (2005). R-Squared, AdjustedR-Squared. *Encyclopedia of Statistics in Behavioral Science*. <https://doi.org/10.1002/0470013192.bsa526>

Mohamad, I. B., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. <https://doi.org/10.19026/rjaset.6.3638>

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304. <https://doi.org/10.1038/nbt0308-303>

Robertson, N. (2019). *Increased government spending to drive economic growth | BERL*. Berl.co.nz. <https://berl.co.nz/economic-insights/increased-government-spending-drive-economic-growth>

Saptashwa Bhattacharyya. (2018, September 26). *Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn*. Medium; Towards Data Science. <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>

Saputra Rangkuti, F. R., Fauzi, M. A., Sari, Y. A., & Sari, E. D. L. (2018). Sentiment Analysis on Movie Reviews Using Ensemble Features and Pearson Correlation Based Feature Selection. *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*. <https://doi.org/10.1109/siet.2018.8693211>

Sardá-Espinosa, A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal*, *11*(1), 22. <https://doi.org/10.32614/rj-2019-023>

Singh, P. (2022, January 10). *Data Leakage in Machine Learning: How it can be detected and minimize the risk*. Medium. <https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562>

Spark. (2023, April). *Spark New Zealand releases three-year strategy*. Www.sparknz.co.nz. <https://www.sparknz.co.nz/news/Spark_New_Zealand_releases_FY24-26_three-year_strategy/>

Spark. (n.d.). *Who we are*. Www.sparknz.co.nz. Retrieved June 2, 2023, from <https://www.sparknz.co.nz/about/about-spark/>

StatsNZ. (2019). *Income by region, sex, age groups and ethnic groups*. Govt.nz. [https://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471#](https://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7471)

Stephanie Glen. (2016, October 25). *Absolute Error & Mean Absolute Error (MAE)*. Statistics How To. <https://www.statisticshowto.com/absolute-error/>

*Tauranga City Council > Our Future > Enabling growth*. (n.d.). Www.tauranga.govt.nz. <https://www.tauranga.govt.nz/our-future/enabling-growth>

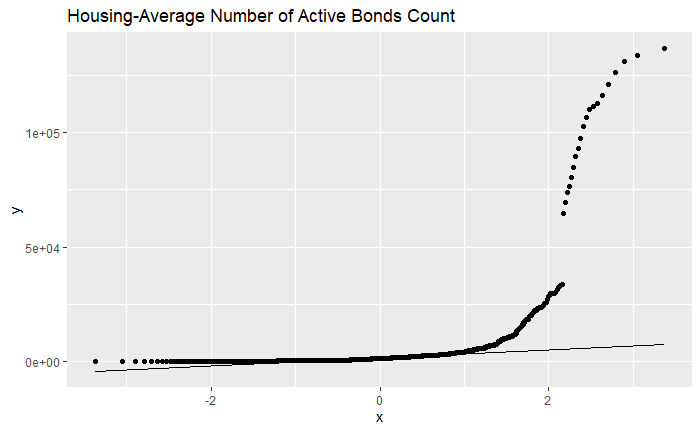
Tauranga City. (n.d.). *Tauranga City Council > Our Future > Enabling growth > New community: Tauriko West*. Www.tauranga.govt.nz. <https://www.tauranga.govt.nz/our-future/enabling-growth/new-community-tauriko-west>

Tauranga City. (n.d.-a). *New community: Te Tumu - Tauranga City Council*. Www.tauranga.govt.nz. https://www.tauranga.govt.nz/our-future/enabling-growth/new-community-te-tumu

Wang, Y., Steele, T., & Zhang, E. (2016). *QQ PLOT QQ PLOT INTERPRETATION: Quantiles*. <https://math.illinois.edu/system/files/inline-files/Proj9AY1516-report2.pdf>

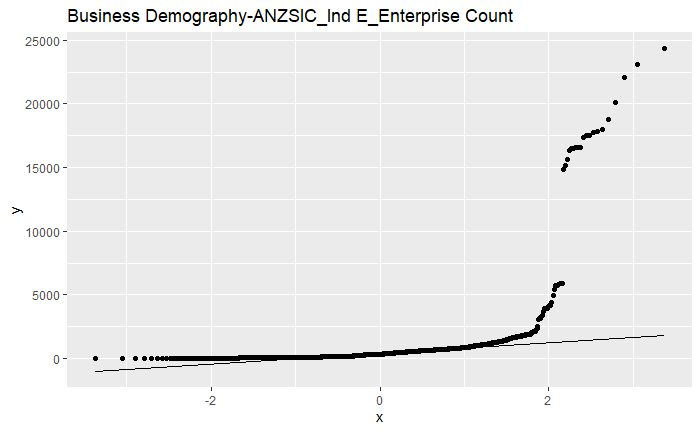
# Appendix A

QQ Plots of main dataset



A picture containing text, screenshot, plot, diagram

Description automatically generated



A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, diagram, plot, screenshot

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, plot, line

Description automatically generated

# Appendix B

ANZSIC 2006 DIVISION, SUBDIVISION, GROUP AND CLASS CODES AND TITLES

|  |  |
| --- | --- |
| **Code** | **Industry** |
| A | AGRICULTURE, FORESTRY AND FISHING |
| B | MINING |
| C | MANUFACTURING |
| D | ELECTRICITY, GAS, WATER AND WASTE SERVICES |
| E | CONSTRUCTION |
| F | WHOLESALE TRADE |
| G | RETAIL TRADE |
| H | ACCOMMODATION AND FOOD SERVICES |
| I | TRANSPORT, POSTAL AND WAREHOUSING |
| J | INFORMATION MEDIA AND TELECOMMUNICATIONS |
| K | FINANCIAL AND INSURANCE SERVICES |
| L | RENTAL HIRING AND REAL ESTATE SERVICES |
| M | PROFESSIONAL, SCIENTIFIC AND TECHNICAL SERVICES |
| N | ADMINISTRATIVE AND SUPPORT SERVICES |
| O | PUBLIC ADMINISTRATION AND SAFETY |
| P | EDUCATION AND TRAINING |
| Q | HEALTH CARE AND SOCIAL ASSISTANCE |
| R | ARTS AND RECREATION SERVICES |
| S | OTHER SERVICES |

# Appendix C

A picture containing text, line, font, screenshot

Description automatically generatedPrincipal Component Analysis