

COMP 693 Industry Project

Final Report

Coastal Proximity, Environmental Risks, and Property Valuation: A Comprehensive GLM Analysis in Christchurch

(Research Project)

Submitted

By

Name: Eric Chen

Student ID: 1157248

Host Company: Department of Land and Management Systems, Lincoln University

Name of Mentor: Dr David Dyason

Mentor Email: david.dyason@lincoln.ac.nz

Company Address: Ellesmere Junction Road

Lincoln 7647, Canterbury

New Zealand

Date: 20/10/2024

Lincoln University

EXECUTIVE SUMMARY

Problem Addressed: This project investigates the relationship between property values and environmental factors such as proximity to the coast, flood zones, and key property features in Christchurch, New Zealand. Coastal properties are often subject to price premiums, but climate-related risks like flooding and erosion complicate their valuation.

Goal: The goal of the project was to develop a robust statistical model to predict 2022 property values in Christchurch, considering both property-specific and environmental factors, and provide insights into market trends.

Methods: We used a combination of the Hedonic Pricing Model (HPM) and Generalised Linear Model (GLM) with a Gamma distribution to analyse property sales data. GIS tools were employed for spatial analysis, including coastal proximity and flood zone overlays. Data was extensively cleaned and structured before analysis, and assumptions were thoroughly tested.

Outcome: The GLM produced meaningful results, revealing significant relationships between property values and factors such as coastal proximity and suburb location. While the model is robust, limitations include data availability for some environmental variables.

All project outputs have been organized and saved in the GitHub repository, which can be accessed at <https://github.com/EricEcc/693-ProjectFolder-EricChen>

The repository includes:

- Project Weekly Journal: Documenting the weekly progress and key milestones.
- Cleaned and Unified Dataset in CSV Format: The final dataset used for the analysis, ready for replication.
- Structured Excel File: Contains relevant datasets, descriptive statistics, and a summary of key variables.
- QGIS Project File: Includes geocoded property layers, coastline proximity measures, and flood zone data layers, providing a spatial view of the analysis.
- Comprehensive R Script File: A well-documented R script capturing the entire data analysis process, including data preparation, EDA, model testing, assumption checking, and visualization. The script includes error handling and detailed instructions for reproducibility.
- Final Research Report: A fully integrated research report with supporting appendices featuring code snippets, detailed results, maps, and graphs.

Table of Contents

EXECUTIVE SUMMARY	1
GLOSSARY/ACRONYMS	4
1. BACKGROUND	4
Overview	4
Problem	5
Project Team	5
2.1 REQUIREMENTS AND GOALS	5
2.2 Literature Review	6
3. METHODS	7
3.1 Overview	7
3.2 Design	9
3.2.1. Data Preparation and Cleaning	9
3.2.2 GIS Software Preparation	12
3.2.3 Statistical Analysis with R	13
3.2.4 Handling Influential Data: Cook's Distance Analysis	15
3.3 Risks and Challenges.....	16
3.4 Implementation.....	18
4. RESULTS AND OUTCOMES	19
4.1 Evidence of Deliverables	19
4.1.1 Data Preparation and Cleaning	19
4.1.2 Model Comparison and Final Model Choice	20
4.1.3 Statistical Model and Implementation	20
4.1.4 Real-World Insights from Statistical Findings	23
4.2.1 Model Evaluation and Diagnostics.....	23
4.2.2 Model Assumption Testing	24
5. Reflections	25
5.1 Reflections.....	25
5.2 Conclusions	27
6. REFERENCES	30

Datasets:	30
Articles and Papers:	30
7. APPENDICES	33
Appendix 1: GLOSSARY/ACRONYMS	33
Appendix 2: Detailed Literature Review	34
Appendix A: Detailed Data Cleaning and Handling of Invalid Data	36
Appendix B: Detailed GIS Software Preparation and Spatial Analysis	38
Appendix C: Detailed Model Selection and Justification	42
Appendix D: Detailed Model Assumptions and Checking	44
Appendix E: Detailed Model Comparison.....	46
Appendix F: Sales Price Index (SPI) Approach Details.....	49
Appendix G: Detailed Cook's Distance Analysis for Influential Data Handling.....	52
Appendix H: Detailed Data Preparation and Cleaning Process	54
Appendix I: Model Comparison and Final Model Selection	59
Appendix J: Detailed Statistical Model Implementation - Generalised Linear Model (GLM) ...	61
Appendix K: Detailed Statistical Findings for General Understanding	64
Appendix L: Real-World Insights from Statistical Findings	67
Appendix M: Detailed Model Evaluation and Diagnostics	73
Appendix N: Detailed Model Assumption Testing	77
Appendix O: R Code for Data Analysis and Modelling	81

GLOSSARY/ACRONYMS

This section provides definitions for key terms and acronyms used throughout the report, especially those relevant to statistical models and real estate analysis:

- **AIC (Akaike Information Criterion):** A metric used to compare model fits, where lower values indicate better models.
- **BIC (Bayesian Information Criterion):** Similar to AIC but with a stronger penalty for model complexity.
- **GLM (Generalised Linear Model):** A statistical model that extends linear regression to handle non-normal distributions.
- **GAM (Generalised Additive Model):** A flexible extension of GLM, allowing for non-linear relationships through smoothing functions.
- **HPM (Hedonic Pricing Model):** A method used to estimate property values based on individual characteristics such as size and location.
- **Cook's Distance:** A diagnostic measure that identifies data points with a disproportionate influence on the model.
- **Multicollinearity:** When predictor variables are highly correlated, which can distort regression results.
- **VIF (Variance Inflation Factor):** A tool used to detect multicollinearity, with values above 10 indicating a potential issue.
- **RMSE (Root Mean Squared Error):** A measure of how accurately the model's predictions match actual values, with lower values indicating better performance.

For more detailed explanations and additional terms, refer to **Appendix 1**.

1. BACKGROUND

Overview

This research project is conducted under the Department of Land and Management Systems at Lincoln University in New Zealand. The department specialises in research and education related to land and property management, focusing on sustainable development and the economic value of land assets. Their work supports decision-making in property markets, land use planning, and environmental management, contributing valuable insights into how land is valued and managed in New Zealand.

The project falls within the domain of real estate economics, particularly focusing on the relationship between environmental attributes (such as coastal proximity and views) and property values. Understanding how factors like location, risk of flooding, and environmental amenities affect property prices is crucial for both property valuation and urban planning, especially in regions facing climate change risks and coastal development pressures.

Problem

The focus of this project is to analyse how coastal proximity and related attributes, such as views and flood risks, affect property values in Christchurch. As coastal cities like Christchurch continue to grow, understanding the impact of location-specific factors on property prices becomes increasingly important, especially in light of environmental risks such as flooding. This research not only contributes to property valuation practices but also helps inform future urban planning and risk management strategies in coastal regions. Studying these factors is essential for ensuring that property markets remain sustainable, resilient, and equitable for future generations.

Project Team

The project team consists of the following members:

1. **Dr. David Dyason (Project Mentor):**

Dr. Dyason is a Senior Lecturer at the Department of Land and Property Management at Lincoln University. As the project mentor, he provides academic oversight, assisting with refining the research methodology and offering his expertise in property valuation and land management to ensure the research meets academic standards.

2. **Eric Chen (Researcher):**

Eric is the primary researcher responsible for conducting all phases of this project. His responsibilities include reviewing relevant literature, analysing data, applying the hedonic pricing model and repeat-sales method, and compiling the final report.

2.1 REQUIREMENTS AND GOALS

Overall Goal:

The overall goal of this project is to develop a comprehensive model that accurately estimates how coastal proximity, environmental factors (such as flooding zones and water views), and property characteristics (such as land and floor area) influence property values in Christchurch. The study aims to provide meaningful insights for stakeholders such as property investors, developers, and policymakers to better understand the factors driving property prices in coastal areas and to aid in future urban planning and risk management.

Requirements for Success:

To ensure the success of this project, the following requirements must be met:

- **Data Acquisition and Cleaning:** Collect and prepare transactional data for Christchurch properties, focusing on coastal areas. Ensure the data is cleaned and prepared for analysis, with relevant variables like distance to the coast, land area, floor

area, and environmental factors (e.g., water view and flooding zones) properly incorporated.

- **Model Development and Testing:** Apply a Generalised Linear Model (GLM) to the dataset, accurately accounting for the relationship between property values and key factors. The model must be validated and tested for robustness.
- **Model Validation:** Perform model validation through diagnostics and cross-validation techniques to ensure that the model is statistically sound and can generalise well to unseen data.
- **Final Report and Insights:** Produce a final report that clearly explains the model findings, providing actionable insights for property market stakeholders and suggesting directions for future research.

Subgoals:

1. Conduct a comprehensive literature review on property valuation models, focusing on the application of hedonic pricing models and alternative approaches (e.g., Generalised Linear Models).
2. Acquire and prepare relevant property transaction data, including necessary variables such as proximity to the coast, property characteristics, and environmental risk factors.
3. Develop multiple models, including a GLM, and compare their performance to select the best-fitting model for the project goals.
4. Validate the model using cross-validation and diagnostic checks to ensure its accuracy and reliability.
5. Compile the findings into a clear, comprehensive report that provides insights into how coastal and environmental factors influence property values in Christchurch.

Metrics for Success:

The project will be considered successful if:

- The data is successfully collected, cleaned, and prepared for analysis.
- A model is developed and validated, demonstrating the ability to explain a significant portion of the variation in Christchurch property values.
- The final report provides clear and actionable insights for stakeholders, with well-supported conclusions about the relationship between coastal proximity, environmental factors, and property values.
- The findings align with the research objectives and provide meaningful contributions to the property valuation literature.

2.2 Literature Review

This literature review highlights key studies and methodologies relevant to the impact of environmental factors, such as coastal proximity and flood risks, on property values. It examines two primary approaches: the Hedonic Pricing Model (HPM) and Generalised Linear Models

(GLMs), alongside the integration of Geographic Information Systems (GIS) and lidar technology.

Hedonic Pricing Model (HPM): Widely used in property valuation, HPM assesses how different property attributes influence value. Hamilton and Morgan (2010) applied this model using GIS and lidar to measure coastal amenities like beach access and ocean views. Rajapaksa et al. (2017) focused on flood risks, showing how recovery from floods impacts property values over time.

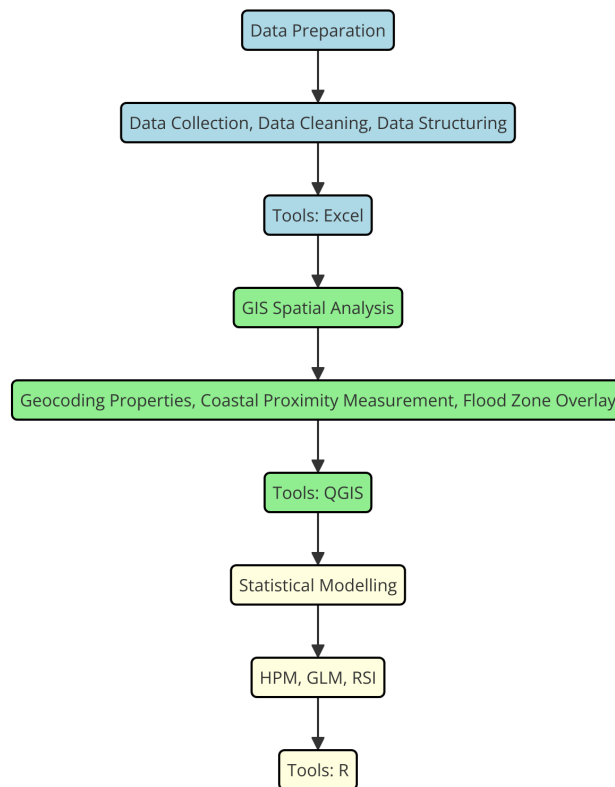
Generalised Linear Models (GLMs): GLMs provide flexibility in addressing non-normal data distributions. Bailey et al. (2022) explored non-linear impacts of environmental risks using Generalised Additive Models (GAMs), an extension of GLMs, while Filippova et al. (2020) assessed sea-level rise impacts on New Zealand's coastal properties.

Environmental Risks and Property Values: McNamara et al. (2015) and Barnard et al. (2019) explored how policies like erosion control and dynamic environmental factors influence property values, highlighting the need to consider long-term environmental risks.

For more details on these studies, refer to **Appendix 2**.

3. METHODS

3.1 Overview



(Figure 1 – Concept Flowchart of Data Analysis)

This project aimed to assess the impact of coastal proximity and environmental factors, such as flood zones, on property values in Christchurch using advanced statistical and geospatial techniques. The approach involved a combination of data collection, spatial analysis using GIS software, and statistical modelling to derive meaningful insights into the property market.

The project was structured into three main phases:

a. Data Preparation and Cleaning:

- The first phase involved gathering relevant property transaction data from multiple sources, including Valbiz and the District Valuation Roll (DVR) dataset from Land Information New Zealand (LINZ). These datasets were filtered, cleaned, and standardised to ensure consistency across key variables such as sale dates, property descriptions, and environmental factors (e.g., water views).
- Excel was primarily used for this stage, leveraging its data manipulation and filtering functions to merge datasets, remove irrelevant entries, and organise records. The preparation also included eliminating non-coastal properties and significant outliers, as well as handling missing data. Geocoding tools like MMQGIS in QGIS were used to obtain latitude and longitude coordinates for each property, marking coastal areas and ensuring that the dataset was ready for spatial analysis.

b. GIS Software Preparation and Spatial Analysis:

- Geographic Information Systems (GIS) tools, specifically QGIS, were employed to integrate and analyse spatial data. This included measuring the proximity of each property to the coast and identifying whether properties fell within designated flood zones using shapefiles provided by the Christchurch City Council. These geographical variables were crucial in determining how environmental factors influenced property prices and were used to enhance the dataset for subsequent analysis.
 - QGIS allowed for precise spatial analysis, with key factors like coastal proximity and flood risk being mapped and assigned to the relevant properties in the dataset.
- c. **Statistical Analysis with R:**
- R Studio was used extensively for advanced statistical analysis. The final phase involved applying sophisticated statistical methods to model the relationships between property values and various attributes (e.g., location, size, environmental factors). The Hedonic Pricing Model (HPM) provided the theoretical framework, while Generalised Linear Models (GLMs) were selected for their ability to capture complex, non-linear relationships and handle non-normal data distributions. Further models, including Generalised Additive Models (GAMs) and interaction terms, were explored and compared to determine the best fit for the data.
 - Additional techniques, such as the Repeat Sales Index (RSI), were employed to estimate property values for 2022 based on historical sales data, ensuring that the analysis was both comprehensive and up-to-date.

Throughout the project, an iterative approach was used, incorporating feedback from the mentor and continually refining the dataset and models. Tools and technologies such as **Excel** for initial data structuring, **RStudio** for statistical analysis, and **QGIS** for spatial mapping were instrumental in managing and executing the project tasks. This structured approach ensured that the project met its goals and delivered robust insights into the relationship between environmental factors and property values in coastal Christchurch.

3.2 Design

3.2.1. Data Preparation and Cleaning

(a) Initial Data Collection

For this project, we utilised two primary datasets to analyse property transactions in Christchurch, spanning from 1 January 1993 to 4 October 2022. These datasets were essential

in creating the foundation for the subsequent analysis and statistical modelling of property prices in relation to environmental factors such as coastal proximity and water views.

Primary Dataset: Christchurch Property Transactions

The main dataset comprises all property transactions recorded within Christchurch over nearly three decades, covering 377,589 records. This data includes vital information such as transaction dates, property addresses, land area, capital value, and other property-specific variables like building type and construction codes (as illustrated in Figure 1 below). The data was obtained from the platform *Valbiz V8*, a comprehensive value-specific software used across New Zealand. Valbiz V8 integrates property and sales records with mapping tools, aerial imagery, and other valuation resources, making it a highly reliable and detailed source for property data management.

No Roll No	Sale Date	CTitle	Ass.No	Full Address	Age	Total Value	Category	Land Area m ²	Land Value	Capital Va Brms	Leg Desc	Age Code	ChattS	Firm2	Construction Code	Age Code	Sale Index	Category	Ownership Type
21909	1/01/1993	36A829	6124	14 O'CONNOR PL		240000	RD9B	0	44000	228000	0 Flat 1 DP 61144 on Lot 6 f	0	185 IG		Not Assigned			RD9B	Unspecified
22110	1/01/1993	37C182	70700	240 A BLENHEIM RD		110000	RD9B	0	23000	100000	0 Flat 2 DP 63511 on Lot 23	3000	110		Not Assigned			RD9B	Unspecified
23421	1/01/1993	37C815	21001	26 E AURORA ST		700	OU	5	1000	7000	0 Lot 1 DP 64153				Not Assigned			OU	Unspecified
23420	1/01/1993	228777	35400	85 LINWOOD AV		68000	RD1B	546	36000	70000	0 PT Lot 2 DP 2574				Not Assigned			RD1B	Unspecified
22822	1/01/1993	64032	7400	26 EVESHAM CR		102000	RD5B	612	36000	102000	0 Lot 56 DP 17552				Not Assigned			RD5B	Unspecified
23441	1/01/1993	11F230	65900	160 HALSEWELL JUNCTION RD		130000	RD7B	696	47500	122000	0 Lot 10 DP 28644				Not Assigned			RD7B	Unspecified
22340	1/01/1993	21X616	88100	30 CHRYSL ST	19	122500	RD2B	797	40000	110000	0 Lot 16 DP 2702				Not Assigned			RD2B	Unspecified
23433	1/01/1993	10505	50900	13 BRANSTON ST	19	42500	RD5B	809	33000	76000	0 LOT 90 O P 18215				Not Assigned			RD5B	Unspecified
22650	1/01/1993	332288	2100	266 SELWYN ST		50000	RD1B	875	31000	97000	0 Lot 1 DP 2999	191	8000	170 TG		191	Not Assigned	RD1B	Unspecified
22110	1/01/1993	459142	47500	94 RATTWAY ST		142000	RD4B	941	64000	100000	0 LOT 29 P P 9725				Not Assigned			RD4B	Unspecified
23434	1/01/1993	3A1158	76400	105 BRINLEY ST		47000	RD1B	1083	39000	96000	0 PT LOT 3 DP 17059				Not Assigned			RD1B	Unspecified
22250	1/01/1993	9A1136	15700	107 LEINSTER RD		17840	RD1B	1531	170000	740000	0 PT LOT 2 DP 18670 LOT 1 I				Not Assigned			RD1B	Unspecified
22271	2/01/1993	17B1263	46000	78 4/78A HOLLY RD		128000	RF7B	0	19000	93000	0 Flat 4 DP 38438 on Lot 2 f				Not Assigned			RF7B	Unspecified
22460	2/01/1993	14F173	18103	16 1/1 CLAYMORE ST		85000	RF7B	0	17000	76000	0 Flat 1 DP 35794 on Lot 4 f				Not Assigned			RF7B	Unspecified
22632	2/01/1993	34A843	32300	149 A SIMEON ST	19	120000	RD9B	0	23000	110000	0 FLAT 2 DP 61223 ON LOT				Not Assigned			RD9B	Unspecified
21905	2/01/1993	26K336	26600	23 RUBENS PL		292000	RD8A	609	89000	276000	0 Lot 19 DP 45188				Not Assigned			RD8A	Unspecified
23441	2/01/1993	36C614	88446	12 VANDERBILT PL	19	260000	RD9B	643	54000	185000	0 Lot 42 DP 61631				Not Assigned			RD9B	Unspecified
22301	2/01/1993	68132	47100	46 HARRISON ST	19	70000	RD4B	645	35000	78000	0 LOT 4 O P 15041				Not Assigned			RD4B	Unspecified
23414	2/01/1993	29A11	14600	6 CHOKERBONE PL		180000	RD8B	660	57000	171000	0 Lot 20 DP 5011	198	0	200 IG		198	Not Assigned	RD8B	Unspecified
23672	2/01/1993	35B1127	19004	32 OVERDALE DR	19	120000	RV	673	110000	110000	0 Lot 15 DP 59481				Not Assigned			RV	Unspecified
22320	2/01/1993	60510	32500	448 INNES RD		118000	RD5B	701	37000	111000	0 Lot 14 DP 1721	195	8000	138 TA		195	Not Assigned	RD5B	Unspecified
23662	2/01/1993	30F139	76808	16 HERBS PL	19	237000	RD8B	783	69000	210000	0 Lot 8 DP 52101	198	15000	210 IG		198	Not Assigned	RD8B	Unspecified
22190	3/01/1993	54413	3200	153 WAIRAKEI RD		148000	RD5B	835	54000	150000	0 Lot 6 DP 15351	195	11000	150 MA		195	Not Assigned	RD5B	Unspecified
22182	4/01/1993	17B423	4800	28 A SALS ST		85000	RD7B	0	13000	93000	0 Flat 2 DP 39108 on Lot 29				Not Assigned			RD7B	Unspecified
22261	4/01/1993	13C195	21900	85 KNOWLES ST		127000	RF7B	0	46000	110000	0 Flat 1 DP 34802 on Lot 70				Not Assigned			RF7B	Unspecified
22350	4/01/1993	14A932	70400	12 2/1 TEMPLAR ST	19	51000	RF7B	0	9000	54000	0 FLAT 2 DP 34883 ON LOT				Not Assigned			RF7B	Unspecified
22520	4/01/1993	2A163	25302	15 LASCELLES ST		122500	RD6B	544	47000	119000	0 LOT 3 DP 21093				Not Assigned			RD6B	Unspecified
22791	4/01/1993	17K484	59800	345 NEW BRIGHTON RD		123750	RD7B	545	41000	125000	0 Lot 4 DP 38896				Not Assigned			RD7B	Unspecified
22960	4/01/1993	101044	56300	24 WAIPARA ST		124000	RD6B	670	45000	111000	0 Lot 11 DP 22421				Not Assigned			RD6B	Unspecified
21902	4/01/1993	56621	20700	155 HAMILTON AV		235500	RD5B	825	93000	202000	0 Lot 6 DP 14908				Not Assigned			RD5B	Unspecified
22580	4/01/1993	50A192	76601	83 EASTERN TC	19	135000	RD6B	880	46500	137000	0 Lot 1 DP 74211	196	13000	173 IG		196	Not Assigned	RD6B	Unspecified
22390	4/01/1993	35C269	5300	43 WAINONI RD		127500	RD2B	1127	33000	95000	0 Lot 2 DP 59614				Not Assigned			RD2B	Unspecified
21914	5/01/1993	36D1106	400	332 2/ HAREWOOD RD	19	157000	RD9B	0	31000	140000	0 Flat 3 DP 63075 on Lot 3 f				Not Assigned			RD9B	Unspecified
21920	5/01/1993	22F932	5800	248 1/ HAREWOOD RD	19	94500	RF6B	0	25000	79000	0 Flat 1 DP 44294 on Lot 4 f				Not Assigned			RF6B	Unspecified
22120	5/01/1993	36D341	45500	11 A KONINI ST		80500	RD2B	0	63000	107000	0 FLAT 1 DP 62922 ON LOT				Not Assigned			RD2B	Unspecified
22120	5/01/1993	37C748	45500	11 KONINI ST		80500	RV	0	56000	56000	0 1/2 INT IN LOT 1 DP 62311				Not Assigned			RV	Unspecified

(Figure 2 – Screenshot of Property Transaction Dataset)



Supplementary Dataset: National District Valuation Roll (DVR)

To supplement the primary dataset and gather more detailed property information, we leveraged data from the *NZ Properties: National District Valuation Roll* (Figure 2). This dataset, made publicly available by Toitū Te Whenua Land Information New Zealand (LINZ), provides detailed valuation and property audit data for properties across New Zealand, though the dataset used in this project was specific to Christchurch. The DVR dataset, updated weekly, includes property characteristics such as land area, ownership type, and whether the property has a view of water, which is crucial for assessing environmental amenities. This dataset was instrumental in augmenting the initial transaction data with more detailed geographic and property feature information.

NZ Properties: National District Valuation Roll



(Figure 3 – Screenshot of DVR Data Structure)

Info	History	Services and APIs
Details		
Table ID	114085 	
Data type	Table	
Row count	243,663 • 243663 with null or empty geometries	
Columns	unit_of_property_id  , valuation_no_roll, valuation_no_assessment, valuation_no_suffix, district_ta_code, situation_number, additional_situation_number, situation_name, legal_description, land_area, property_category, ownership_code, current_effective_valuation_date, capital_value, improvements_value, land_value, trees, annual_value, annual_value_indicator, gross_rental, no_of_bedrooms, improvements_description, zoning, actual_property_use, units_of_use, off_street_parking, building_age_indicator, building_condition_indicator, building_construction_indicator, building_site_coverage, building_total_floor_area, mass_contour, mass_view, mass_scope_of_view, mass_total_living_area, mass_deck, mass_workshop_laundry, mass_other_improvements, mass_garage_freestanding, mass_garaged_under_main_roof, production, sale_group	

These two datasets provided the necessary depth and granularity required to develop our research project, particularly in linking property values to environmental variables like coastal proximity, water views, and flood risk.

(b) Data Cleaning and Handling of Invalid Data

Data cleaning was essential to ensure the dataset's accuracy and relevance for analysing the impact of coastal proximity on property values in Christchurch. This process

involved filtering out non-coastal properties and removing invalid or incomplete records to maintain the integrity of the analysis.

Filtering Non-Coastal Areas:

The first step in data cleaning involved removing non-coastal properties using QGIS software. The MMQGIS Plugin was utilized to geocode property addresses and assign latitude and longitude coordinates. Properties in non-coastal suburbs were excluded, leaving only those relevant to the study. This step ensured the analysis focused on properties with proximity to the coast, a key factor in our study.

Handling Missing and Invalid Data:

Transaction records with missing essential data, such as sale dates or property values, were eliminated, as these variables are critical for accurate analysis. Similarly, irrelevant factors, such as incorrect bedroom counts (e.g., "Brms" field showing all values as zero), were removed to simplify the dataset.

Outlier Removal:

Extreme outliers, such as a property at 272 Marine Parade showing a 2500% value increase due to a combined transaction, were also identified and removed. This step was crucial to prevent abnormal records from distorting the results of the analysis.

By conducting these steps, the dataset was refined to include only relevant, complete, and accurate data for meaningful analysis. For further details on the data cleaning process, please refer to Appendix A.

(c) Data Structuring

In order to conduct a meaningful analysis on property values over time, it was necessary to restructure the dataset, which initially presented all the raw transaction records in Christchurch from 1993 to 2022. This raw data included one row per transaction but did not group the records by individual properties, which is essential for tracking changes in value over time.

Reorganising Transaction Data by Property

The first step in data structuring involved reorganising the transaction data so that each property could be easily identified across multiple transactions. To achieve this, we used the "Legal Description" (Leg.Desc) as the unique key to group and match transactions for the same property. This ensured that even if a property had been sold multiple times during the data period, its transactions could be correctly identified and linked. Grouping properties by Leg.Desc allowed for comparisons over time, making it possible to analyse how factors such as proximity to the coast and environmental risks influenced property values.

Combining Data from Various Sources

In addition to the primary property transaction dataset, additional data files from the District Valuation Roll (DVR) were also integrated to enhance the analysis. This dataset provided further details about the properties, such as whether they had a water view. By merging the primary dataset with this additional dataset using the unique property identifiers, we were able to create a more comprehensive dataset, incorporating both transactional and environmental details for each property.

3.2.2 GIS Software Preparation

(a) Coastal Proximity and Flood Zone Data Collection

The preparation of geographical and environmental data involved using QGIS to integrate coastal proximity and flood zone information. Two key spatial datasets were employed: a coastline shapefile marking the land-sea boundary and a flood zone shapefile identifying areas susceptible to flooding. Both files were sourced from the Christchurch City Council's Spatial Open Data Portal. The coastline data allowed us to calculate the distance from each property to the coast, while the flood zone data identified properties located in high-risk flood areas. These variables were crucial for enhancing the accuracy of the hedonic pricing model.

(b) Coastal Proximity and Flood Zone Analysis

Using QGIS, we geocoded property locations to obtain their latitude and longitude and calculated the distance to the coastline using a spatial join between the property dataset and the

coastline shapefile. We also overlaid the flood zone shapefile to identify whether properties were located within flood-prone areas. Both variables, 'DistanceToCoast' and 'FloodingZone,' were added to the dataset for further analysis. For a detailed description of the processes and data sources used, please refer to Appendix B.

3.2.3 Statistical Analysis with R

(a) Model Selection and Justification

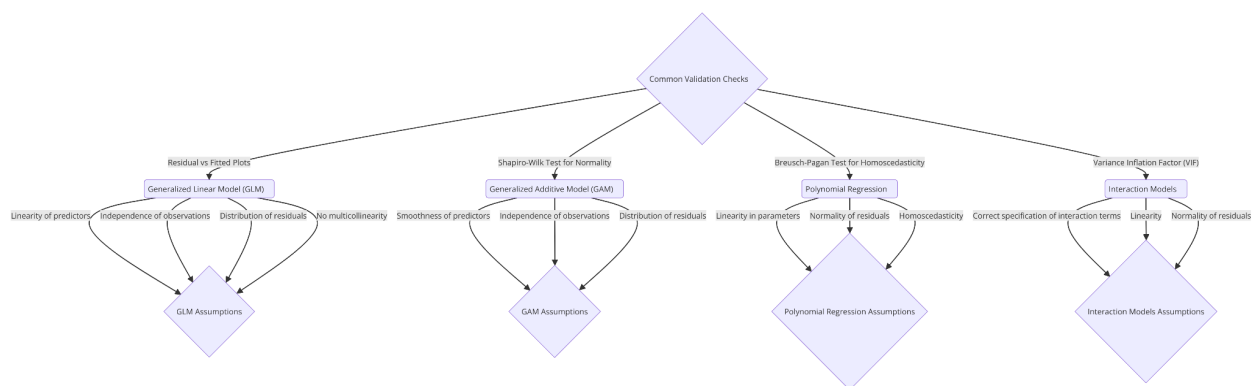
Introduction to the Hedonic Pricing Model (HPM)

The Hedonic Pricing Model (HPM) offers a foundational approach to real estate valuation, focusing on how property characteristics like size, location, and environmental factors contribute to the overall price. The core idea is to decompose a property's value into individual attributes, enabling an understanding of how each feature influences the price. This approach is particularly useful in examining factors like proximity to the coast and water views, which often command higher property values, as well as risks such as flood zones that tend to reduce market prices. The HPM equation expresses property price as a linear combination of various property characteristics, providing a clear interpretation of their marginal contributions.

Limitations of HPM

While effective for basic valuation, HPM has several limitations: it assumes linear relationships, struggles with non-normal data distributions, and is sensitive to outliers. Furthermore, it cannot easily capture complex interactions between variables, such as the varying effects of coastal proximity in flood-prone areas. HPM also assumes homoscedasticity, which is rarely the case in real estate data. These limitations led us to transition to more flexible models like the Generalised Linear Model (GLM), although HPM remains valuable for its theoretical framework and interpretability. Detailed discussions of the model's limitations and their implications for property pricing analysis can be found in **Appendix C**.

(b) Model Assumptions and Checking



(Figure 4. The Concept Flowchart of Assumption Checking)

When conducting the statistical analysis using models such as GLM, GAM, polynomial regression, and interaction models, it was critical to ensure that the assumptions underlying these models were met. Assumptions such as linearity of predictors, independence of observations, and correct distribution of residuals were checked to validate the results.

Generalised Linear Model (GLM)

For GLM, we ensured that the relationship between predictors like proximity to the coast and property values followed a linear pattern on the transformed scale. Additionally, multicollinearity was tested using the **Variance Inflation Factor (VIF)**, and residual diagnostic plots were used to confirm that the residuals followed the expected Gamma distribution.

Generalised Additive Model (GAM)

For GAM, we checked whether smooth functions captured non-linear relationships between the predictors and property values. Independence of observations and the distribution of residuals were also verified.

Polynomial Regression and Interaction Models

Polynomial regression was tested for normality of residuals and constant variance using diagnostic plots and tests like **Shapiro-Wilk** and **Breusch-Pagan**. Interaction models were checked for correctly specified interaction terms and linearity.

Further details of the assumption checks and results are provided in **Appendix D**.

(c) Model Comparison

This section compares several models—Generalised Linear Model (GLM), Generalised Additive Model (GAM), Polynomial Regression, and Interaction Models—used to analyse property prices in Christchurch’s coastal areas.

1. **Generalised Linear Model (GLM)**
GLM was selected for its ability to handle non-normal data distributions, making it well-suited to our skewed property price data. Its log link function enables straightforward interpretation and manages non-constant variance.
2. **Generalised Additive Model (GAM)**
GAM improves upon GLM by allowing non-linear relationships between predictors and the outcome. This flexibility is useful for environmental data like proximity to the coast, but the model is more complex and can risk overfitting.
3. **Polynomial Regression**
This model captures non-linear relationships by adding higher-order terms, but it

increases the risk of multicollinearity and overfitting, especially with higher-degree polynomials.

4. **Interaction Models**

Interaction models assess how variables like coastal proximity and property size jointly influence prices. While valuable, these models are complex and harder to interpret.

Model Performance

All models were compared using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Squared Error (RMSE). GAM had slightly lower AIC and BIC scores, but GLM provided a balance between fit and interpretability. Cross-validation also highlighted GLM's robustness, as it avoided overfitting and performed consistently across datasets.

Final Model Selection

Ultimately, GLM was chosen as the final model for its interpretability, stable performance, and practical insights. GAM and Polynomial Regression showed flexibility but at the cost of interpretability and overfitting risks. Further details on the comparison are provided in the **Appendix E**.

(d) Sales Price Index (SPI) Approach for Estimating 2022 Property Prices

The Sales Price Index (SPI) Approach was employed to estimate property prices for 2022, targeting properties without a recorded transaction in that year but with historical sale data. The SPI method allows us to project the 2022 property values for a comprehensive dataset, which is critical for analysing coastal property values and environmental factors such as proximity to the coastline and flood zones.

Unlike the traditional Repeat Sales Index (RSI), which relies on observed repeat sales, the SPI approach generates a price index using historical sales and then applies this index to predict current values. This approach ensures that even properties without recent transactions have a standardised estimated value for 2022, facilitating a more inclusive analysis.

Further details on the SPI calculation steps and the statistical methods applied to generate the 2022 price estimates can be found in **Appendix F: Sales Price Index (SPI) Approach Details**.

3.2.4 Handling Influential Data: Cook's Distance Analysis

During the data preparation stage, we identified and removed influential data points that could skew the results of our analysis. These influential points, if not addressed, could distort the model's coefficients, leading to unreliable results. Cook's Distance, a diagnostic measure, was employed in R to detect such points.

Cook's Distance Methodology

Cook's Distance calculates the effect of deleting a specific data point on the model's overall estimates. Data points with Cook's Distance values significantly above a threshold ($4/(n-k-1)$) were flagged as influential.

Process Overview:

1. **Fit Initial Model:** The Generalised Linear Model (GLM) was applied to the dataset to assess relationships between property values and factors like coastal proximity, land area, and water views.
2. **Calculate Cook's Distance:** Using R's `cooks.distance()` function, a plot was generated to identify influential data points, as illustrated in the Cook's Distance graph (Figure X).
3. **Eliminate Influential Points:** Points exceeding the threshold were removed from the dataset to create a refined dataset free from data that could disproportionately impact the analysis.
4. **Refit the Model:** The GLM was refitted using this cleaned data, ensuring more reliable estimates.

This process produced a refined dataset, referred to as **cleaned_data**, which improved the robustness and reliability of the analysis. For further details on the methodology and steps taken, please refer to **Appendix G**.

3.3 Risks and Challenges

Throughout the project, several key risks and challenges were encountered, particularly during the stages of data collection, geospatial integration, and managing the complex relationships in property pricing data. Below are the main issues faced during these processes:

Data Quality and Inconsistencies Across Sources

- **Challenge:** The primary challenge in the data preparation phase stemmed from downloading data from multiple sources, each of which recorded property transactions and attributes in different formats. For example, the primary transaction dataset from the Valbiz platform contained historical property sale records, while the District Valuation Roll (DVR) data from Land Information New Zealand (LINZ) used different identifiers and field structures for the same properties. This mismatch complicated the process of aligning factors like property attributes and geographical identifiers (e.g., water view, legal descriptions, property types) into a unified dataset for analysis.
- **Risk:** The lack of a standardised format across datasets increased the risk of errors in matching property records, leading to possible data duplication, misaligned factors, or incomplete data entries.
- **How Overcome:** To address this, I used the **Leg.Desc** (Legal Description) field as the unique identifier to merge data, ensuring that all variables were accurately linked across

datasets. Additional checks were implemented to filter out properties that could not be matched, and any incomplete or duplicate records were removed after careful inspection.

Geocoding and Spatial Data Integration in QGIS

- **Challenge:** A significant portion of the analysis involved integrating spatial data for coastal properties and flood zones using GIS software. However, the geocoding process to convert property addresses into latitude and longitude coordinates for use in QGIS posed technical difficulties. The data often lacked consistent address formats, and the MMQGIS plugin occasionally returned inaccurate coordinates due to variations in address formatting.
- **Risk:** Inaccurate geolocation could result in properties being misclassified, particularly in terms of proximity to the coast or inclusion within a flood zone. This could distort the model's ability to properly account for environmental variables.
- **How Overcome:** I overcame this by using external tools to validate the geocoding results, ensuring that the latitude and longitude data points aligned correctly with the map coordinates. Additionally, I used predefined suburb boundaries to more accurately mark coastal properties in Christchurch, reducing the reliance on geocoding for properties located within these suburbs.

Handling Non-Linear Spatial Relationships

- **Challenge:** One of the most challenging aspects of the project was accounting for the non-linear relationships between property values and environmental factors, such as proximity to the coast and flood risk. The spatial data obtained through QGIS, including coastal proximity and flood zones, introduced complexities that traditional linear models struggled to capture. For example, the effect of coastal proximity on property values was not uniform—properties closer to the coast often saw sharp increases in value, while properties slightly further away experienced diminishing returns in terms of price premiums.
- **Risk:** If not accurately modelled, these non-linear relationships could lead to misleading conclusions, such as overstating or understating the impact of proximity to the coast or flood zones on property values.
- **How Overcome:** To address this, I explored alternative models such as Generalised Linear Models (GLM) and Generalised Additive Models (GAM), which allowed for more flexibility in capturing the non-linear effects of environmental factors on property prices. This helped provide a more nuanced analysis of how different geographical factors influenced property values. I also conducted assumption checks to ensure that the selected models were appropriate for the data structure.

These specific challenges were critical to the success of the project, and overcoming them ensured that the final dataset and analysis were both accurate and reliable for evaluating the impact of environmental factors on property values.

3.4 Implementation

The implementation of this project was structured into distinct phases, focusing on data preparation, spatial analysis using GIS, and statistical modelling with R. Each phase contributed significantly to the overall outcomes, and specific artefacts were produced at each stage to document and support the research process. Below is a detailed breakdown of the implementation:

1. **Data Preparation:** The initial step involved preparing the property transaction data from Christchurch, covering the period from 1993 to 2022. Excel was used to clean and filter the data, eliminating records that were irrelevant or erroneous, such as non-coastal properties and entries with missing or invalid values. Additional property information from the NZ Properties District Valuation Roll dataset was merged using Excel functions, particularly leveraging VLOOKUP to ensure a consistent structure across datasets. The final dataset focused on property-level analysis by restructuring transactions to group by property using legal descriptions as the unique key.

Artefacts Produced:

- Cleaned and unified dataset in CSV format
 - Structured Excel file contains relevant datasets and descriptive statistics
2. **GIS Spatial Analysis:** Using QGIS, the spatial analysis phase involved geocoding property addresses to assign latitude and longitude coordinates, calculating the distance of each property from the coastline, and overlaying flood zone data to determine which properties were in flood-prone areas. This process was carried out by integrating shapefiles, such as the Christchurch coastline and flood zone data, sourced from the Christchurch City Council's Spatial Open Data Portal. MMQGIS plugin was employed for geocoding, and spatial joins were used to calculate distances and overlay flood zone data.

Artefacts Produced:

- QGIS project folder, including geocoded property layers, coastline proximity measures, and flood zone data layers
3. **Statistical Modelling:** R was the primary tool for the advanced statistical analysis phase. Several models, including the Hedonic Pricing Model (HPM), Generalised Linear Model (GLM), and Repeat Sales Index (RSI) approach, were employed to estimate property values and analyse the impact of environmental factors on those values. The GLM was selected as the final model after performing assumption checks and comparisons with other models like Generalised Additive Models (GAMs) and Polynomial Regression. Key outputs, including coefficients, model diagnostics, and validation metrics, were generated and analysed in R.

Artefacts Produced:

- Comprehensive R Script File: A well-documented R script that captures the entire data analysis process, including data preparation, exploratory data analysis (EDA), model testing, assumption checking, and results visualization. The script includes error handling, clear instructions for usage, and outputs for all stages of analysis. This script ensures reproducibility and transparency in the analysis process.
4. **Final Report:** The findings and results from the data analysis and modelling were compiled into a final report. This report includes detailed explanations of the

methodology, results, and their implications for property values in Christchurch's coastal areas. Supporting appendices provide additional documentation, including code snippets, data summaries, and visualisations to ensure transparency and reproducibility of the results.

Artefacts Produced:

- Final research report, integrating the findings of the entire project
- Supporting appendices featuring code snippets, detailed results, maps, and graphs

This structured approach ensured the successful completion of the project, with each phase contributing to the overall objective of analysing the effect of coastal proximity and environmental risks on property values in Christchurch. The artefacts produced during this process not only serve to document the project but also provide a solid foundation for future research and practical applications in urban planning, property valuation, and risk management.

4. RESULTS AND OUTCOMES

4.1 Evidence of Deliverables

4.1.1 Data Preparation and Cleaning

The final dataset for analysis consists of 2,466 property transactions, each representing coastal Christchurch properties. This dataset underwent extensive data cleaning and transformation, ensuring the removal of invalid or incomplete records. Here, we summarise the key statistics from this cleaned dataset, offering insights into the distribution and characteristics of properties near Christchurch's coast.

Key Variables:

- **Distance to Coast:** Mean of 592.66 metres, median of 414.14 metres, with distances ranging from 44.47 to 1,781.54 metres. This wide range highlights the diversity of property locations, including both beachfront and inland properties.
- **Land Area:** Mean of 672.53 square metres, median of 612 square metres. Land area varies significantly, from small lots to larger estates, demonstrating the diverse residential property types.
- **Floor Area:** Mean of 139.40 square metres, median of 120 square metres. Floor areas are more consistent but with occasional outliers indicating large homes.
- **Property Value (2022):** The average property value is \$703,817.46, with a median of \$674,144.55. Values range from approximately \$9,906 to over \$2 million, suggesting high-demand coastal properties.

Key findings include the high demand for coastal properties, with proximity to the coast being a critical factor influencing higher property values. Larger properties tend to represent higher-end homes, and the property value distribution shows typical market skewing, with a small number of high-value properties influencing the mean.

Further details on the data preparation and statistical findings are provided in **Appendix H**.

4.1.2 Model Comparison and Final Model Choice

In this project, several statistical models were compared to determine the best fit for analysing property values and key influencing factors such as proximity to the coast, land area, and environmental risks. The models evaluated include the Generalised Linear Model (GLM), Generalised Additive Model (GAM), Polynomial Model, and Interaction Models.

Model Performance: AIC and BIC Comparison

Model comparison was guided by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), where lower values indicate better model fit. The following AIC results were obtained:

- **GLM:** AIC = 319.21
- **GAM:** AIC = 262.48 (lowest)

Although the GAM had the lowest AIC score, the GLM was ultimately chosen for several reasons, including its suitability for the skewed and non-negative property data, ease of coefficient interpretation, and flexibility in handling non-linear relationships.

Justification for Choosing GLM

1. **Data Characteristics:** GLM, with a Gamma distribution and log link, was better suited for the skewed nature of the property values.
2. **Interpretability:** The log link in GLM allowed for clear interpretation of how each variable influences property prices.
3. **Model Flexibility:** GLM offered sufficient flexibility without the complexity of GAM.
4. **Consistency with Theoretical Framework:** The GLM supports the multiplicative relationships typically seen in Hedonic Pricing Models.

Further diagnostic checks confirmed the GLM's robustness, making it the optimal choice despite GAM's lower AIC. For detailed information on model diagnostics and performance comparisons, refer to **Appendix I**.

4.1.3 Statistical Model and Implementation

Generalised Linear Model (GLM)

The Generalised Linear Model (GLM) was chosen as the primary statistical approach to examine the relationship between Christchurch property values and key predictors such as coastal proximity, land area, floor area, water view, and flood zone status. The GLM is well-suited for handling skewed property value data through the use of a Gamma distribution and a

log link function. This setup allows for the modelling of non-negative, continuous data and ensures that multiplicative relationships between property characteristics and values are effectively captured.

Key Model Features

- **Gamma Distribution & Log Link:** These components help manage the skewed nature of property values and ensure the model provides realistic predictions without negative values.
- **Predictors:** Continuous variables like land area and distance to the coast, along with categorical variables such as water view and flood zone status, are incorporated into the model.

Coefficient Interpretation

```
> summary(glm_model)

Call:
glm(formula = LogValue2022 ~ LogDistance + LogLand + LogFloor +
    WaterView + FloodingZone + Suburb, family = Gamma(link = "log"),
    data = cleaned_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.4499513   0.0099740  245.635 < 2e-16 ***
LogDistance   -0.0040203   0.0006398   -6.283 3.94e-10 ***
LogLand        0.0013243   0.0014940    0.886 0.375478
LogFloor       0.0325724   0.0012362   26.348 < 2e-16 ***
WaterView1     0.0065159   0.0024743    2.633 0.008508 **
FloodingZone1  -0.0017250   0.0015983   -1.079 0.280566
SuburbNorth New Brighton 0.0052544   0.0009826    5.347 9.80e-08 ***
SuburbSouth New Brighton 0.0066766   0.0013330    5.009 5.89e-07 ***
SuburbSouthshore 0.0070728   0.0021148    3.345 0.000837 ***
SuburbWaimairi Beach  0.0236598   0.0016685   14.180 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.0003655362)

Null deviance: 1.57173  on 2339  degrees of freedom
Residual deviance: 0.86346  on 2330  degrees of freedom
AIC: 319.21

Number of Fisher Scoring iterations: 3
```

(Figure 5: R Output of GLM Model Summary)

- **Distance to Coast:** The negative coefficient (-0.0040) for log distance indicates that as the distance from the coast increases, property values decrease slightly, by about 0.4% per 1% increase in distance, holding other factors constant.
- **Floor Area:** The coefficient for log floor area (0.0326) indicates a significant increase in property value (approximately 3.3%) for every 1% increase in floor area.
- **Water View:** Properties with water views show a positive effect (0.0065) on value, increasing by around 0.7%.

- **Suburb Impact:** Properties in desirable suburbs like Waimairi Beach show significantly higher values, with Waimairi Beach having the largest impact (2.4% increase compared to New Brighton).

The GLM findings provide a robust understanding of how property attributes influence value in Christchurch, with proximity to the coast, floor area, and location being key drivers. For further details and insights into the model coefficients and residual analysis, refer to **Appendix J**.

Simplifying Statistical Findings for General Understanding

This section aims to present the results of our Generalised Linear Model (GLM) analysis in a clear and accessible way for readers who may not have a background in statistics. The findings offer valuable insights into Christchurch's property market, focusing on how factors like proximity to the coast, property size, and location influence property values.

Key Insights:

1. **Proximity to the Coast (LogDistance):** Properties closer to the coast tend to have higher values. As distance from the coast increases, property values slightly decrease. This highlights the premium placed on coastal living, particularly in Christchurch, where popular beaches like New Brighton and Southshore make coastal proximity a highly desirable feature.
2. **Land and Floor Area (LogLand and LogFloor):** While land size alone does not significantly affect property values, the size of the house (floor area) does. Larger homes tend to command higher prices, reflecting the urban trend where built-up space is more valuable than plot size.
3. **Water View and Flooding Zone:** Properties with water views are generally more valuable, indicating high demand for scenic locations. However, being located in a flood zone only slightly decreases property values, suggesting that flood risks may not yet be a major factor in pricing decisions.
4. **Suburban Location:** Properties in certain suburbs, like Waimairi Beach and Southshore, are valued higher than those in other areas. Location is a key determinant of property value, influenced by local amenities, infrastructure, and environmental appeal.

Importance of These Findings:

These insights are crucial for homebuyers, real estate professionals, and urban planners. Buyers can make informed decisions about where to invest, while real estate agents can better market properties. Urban planners can use this information to guide development projects, balancing coastal regeneration with long-term sustainability.

For further details on the statistical models and specific results, please refer to **Appendix K**.

4.1.4 Real-World Insights from Statistical Findings

This section highlights key real-world insights derived from the Generalised Linear Model (GLM) analysis of property values in Christchurch. The findings emphasise the significance of coastal proximity, suburb-specific price variations, and the broader implications for urban planning and environmental sustainability in Christchurch's property market. Detailed findings and their implications are summarised below, while more extensive discussions can be found in Appendix L.

(a) Impact of Coastal Proximity on Property Values

The GLM analysis reveals a strong negative correlation between property value and distance from the coast. Properties located closer to the coast, especially in popular suburbs like New Brighton and Sumner, command a premium due to the lifestyle benefits and scenic views associated with coastal living. While this trend aligns with global real estate markets, concerns about the long-term impact of climate change—such as rising sea levels—are emerging as important considerations. Currently, buyers in Christchurch continue to prioritise proximity to the coast, but this could shift as environmental risks become more pressing.

(b) Suburb-Level Property Value Variations

The analysis shows significant variation in property values across Christchurch's suburbs, with coastal areas like Waimairi Beach and Southshore commanding higher prices. Newer developments and amenities in these areas contribute to their higher desirability. Conversely, suburbs like New Brighton, which are undergoing regeneration, are expected to see future property value increases due to infrastructure improvements and community development projects. This trend mirrors global findings, where urban renewal efforts have positively impacted property prices in previously underperforming areas.

(c) Urban Planning and Development Insights

Urban regeneration projects, like the New Brighton Regeneration Project, play a critical role in shaping property markets by boosting property values and attracting investment. The analysis underscores the need for sustainable urban planning that enhances infrastructure while mitigating environmental risks, such as flooding and erosion. Christchurch's coastal suburbs will benefit from a combination of development and climate resilience measures to maintain their desirability and long-term value. Sustainable urban planning will be essential in safeguarding Christchurch's property market in the face of future environmental challenges.

For further details on these findings, including their statistical and contextual explanations, refer to **Appendix L**.

4.2.1 Model Evaluation and Diagnostics

Several diagnostic checks were performed to validate the Generalised Linear Model (GLM) used in this analysis and to ensure its robustness in predicting property values.

Residual Analysis:

Residual plots, including Pearson and deviance residuals, were reviewed. Both sets of residuals showed a random scatter around zero, indicating no issues like heteroscedasticity or specification errors. The "Pearson Residuals vs Fitted Values" plot showed no patterns, confirming that the model fits the data well without bias.

Influence and Leverage:

Cook's Distance and leverage plots were examined to detect any disproportionate influence by individual data points. No points were found to exert excessive influence, ensuring that the model's predictions are reliable and not skewed by outliers.

Predicted vs Actual Values:

The "Predicted vs Actual Values" plot exhibited a strong positive relationship, demonstrating that the model captures most of the variation in property values accurately, reinforcing the model's effectiveness.

Goodness-of-Fit Metrics:

- **Null Deviance:** The null deviance was 1.5717, which represents the baseline variability without predictors.
- **Residual Deviance:** After incorporating predictors like LogDistance, LogLand, LogFloor, WaterView, FloodingZone, and Suburb, the residual deviance dropped significantly to 0.8635, indicating a much better fit.
- **AIC:** The model's Akaike Information Criterion (AIC) was 319.21, suggesting a good balance between simplicity and explanatory power.
- **Pseudo R-Squared (McFadden's R^2):** The model achieved a pseudo R-squared of 0.8237, explaining 82.37% of the variability in property values.

Conclusion:

The diagnostic checks and goodness-of-fit metrics confirm that the GLM with a Gamma distribution and log link is a robust and effective model for explaining property value variations in Christchurch. Key factors like coastal proximity, floor area, and water views significantly influence property values, while land area and flood zones show less impact. The model's performance is reliable for applications in property valuation and urban planning.

For a more detailed explanation of the diagnostics and residuals, please refer to **Appendix M**.

4.2.2 Model Assumption Testing

In regression modelling, validating key assumptions ensures the reliability of the results. For the Generalised Linear Model (GLM) with a Gamma distribution, several critical assumptions were checked in this project.

(a) Independence of Residuals

The residuals should be independent, meaning one error should not depend on another. The "Residuals over Time" plot showed no patterns, indicating no autocorrelation. This suggests the model has met the assumption of independence.

(b) Multicollinearity

Multicollinearity, when predictor variables are highly correlated, can result in unstable coefficients. Using the Variance Inflation Factor (VIF), we confirmed that all VIF values were well below the threshold of 10, indicating no problematic multicollinearity in the model.

(c) Normality of Residuals

For GLMs, normality of residuals is not required. Therefore, tests like the Shapiro-Wilk or Q-Q plots were unnecessary in this context, as the Gamma distribution was used.

(d) Homoscedasticity

Homoscedasticity, or constant variance of residuals, is also not required for GLMs. The variance in GLMs is a function of the mean, so no specific adjustments for non-constant variance were needed.

(e) Influential Points

Cook's Distance and leverage plots were used to detect influential points. While a few influential points were identified, they did not affect the overall model fit, ensuring the model remains reliable.

Conclusion

The assumption checks confirmed that the GLM meets the critical assumptions for independence and multicollinearity. While normality and homoscedasticity are not required for GLMs, the model performed well, and influential points did not distort the results. This confirms the reliability of the model for analysing property values.

For further details on the assumption testing process, see **Appendix N**.

5. Reflections

5.1 Reflections

The goal of this project was to analyse property values in Christchurch using advanced statistical models to better understand the factors influencing these values, particularly focusing on coastal proximity and other environmental characteristics. By utilising models such as the Hedonic Pricing Model (HPM) and Generalised Linear Model (GLM), we aimed to explore how key factors such as distance to the coast, floor area, and water views affect property prices. After careful data preparation, statistical analysis, and model refinement, I believe that the project largely achieved its objectives.

Achievements and Expectations

Overall, the project delivered the intended results. We were able to successfully identify and quantify the impact of various factors on property values. The findings, especially regarding the significant influence of coastal proximity, water views, and suburban location, aligned with expectations based on prior research and industry understanding. However, the process also revealed several unexpected challenges, particularly in managing and cleaning the dataset, which required more time and technical skill than initially anticipated.

One aspect that deviated from the original expectation was the performance of different statistical models. While the Generalised Additive Model (GAM) and Interaction Model showed promising AIC/BIC scores, the GLM was chosen due to its fit with the nature of the dataset. This choice was driven by the data's distribution characteristics, which required a model that could handle skewed, non-negative values. Although the AIC and BIC results suggested that other models might be a better fit in some respects, the GLM provided the best interpretability and consistency with the theoretical framework of property valuation.

New Learnings

The project provided several opportunities for learning, particularly in areas where I had limited prior experience. One of the most valuable technical skills I gained was working with **QGIS**, which played a critical role in performing spatial analysis. Before starting this project, I had a basic understanding of geographic information systems, but integrating these tools with statistical analysis in **R** offered a new level of complexity. Learning how to geocode property data, calculate proximity measures, and overlay environmental factors such as flood zones added an invaluable layer of depth to my analysis.

Another significant learning experience was in managing large datasets. Working with over 370,000 property transactions presented challenges in data cleaning, including eliminating invalid entries and addressing influential data points. I became more adept at handling missing and outlier data, applying R's diagnostic tools to ensure the models were reliable. I also learned the importance of effective data structuring when merging datasets from different sources to create a unified dataset for analysis.

From a statistical perspective, I gained a deeper understanding of model diagnostics, particularly in testing model assumptions and interpreting goodness-of-fit measures. Additionally, the use of Cook's Distance to handle influential points and cross-validation techniques to ensure model robustness were key technical skills that I developed through this project.

Personal and Technical Growth

On a personal level, this project helped me grow in terms of **project management** and time management skills. The complexity of coordinating different tasks, from data cleaning to spatial analysis to model testing, required careful planning and organisation. I found that staying adaptable was critical, especially when certain tasks—such as managing GIS data or understanding the nuances of the GLM—took more time and effort than initially expected.

Technically, my programming skills in **R** and **Excel** have improved significantly. I can now confidently apply a range of statistical models, perform diagnostic checks, and interpret model outputs in the context of real-world applications. Beyond the technical side, I have also grown in my ability to connect statistical findings to meaningful, real-world insights, particularly in the property market, where understanding these factors can directly inform valuation practices and urban planning.

In conclusion, this project was a successful learning experience that allowed me to develop new skills, deepen my understanding of property valuation, and apply advanced statistical methods to complex real-world data. It provided a strong foundation for future work in property analysis and related fields.

5.2 Conclusions

This project successfully utilised advanced statistical models to explore the relationship between property values and key factors such as coastal proximity, water views, and suburb location in Christchurch. By employing models like the Hedonic Pricing Model (HPM) and Generalised Linear Model (GLM), we were able to quantify the influence of these factors on property prices, providing valuable insights for property valuation and urban planning. However, like any research, this project has its strengths and limitations, and there are areas that could be improved or further explored in the future.

Strengths

One of the key strengths of this project lies in its comprehensive approach to property value analysis. The combination of spatial analysis using QGIS and advanced statistical techniques in R allowed for a detailed examination of how geographic factors such as distance from the coast and environmental features like flooding zones affect property prices. This approach provides a more holistic view of property valuation, capturing the interplay between location-based factors and property characteristics.

Another strength is the flexibility and robustness of the Generalised Linear Model (GLM) employed in this study. The GLM was well-suited to handling the skewed and non-negative distribution of property values, offering interpretable results that align with both theoretical and practical expectations in the property market. The model's ability to handle outliers and influential points through diagnostic tools also contributed to the reliability of the findings.

Additionally, the insights derived from this project have real-world implications for urban planning and property valuation, particularly in Christchurch's coastal areas. Understanding the premium placed on coastal proximity and the potential risks associated with environmental factors like flooding provides valuable information for stakeholders in the property market, including buyers, valuers, and policymakers.

Limitations

Despite its strengths, the project faced several limitations. One of the main challenges was related to **data quality**. The property transaction data, while extensive, contained several invalid

or missing entries, which required significant effort in cleaning and structuring. Even after these efforts, some important factors, such as land use restrictions or building condition, were not included in the dataset, limiting the comprehensiveness of the analysis.

The reliance on historical data also posed a limitation, as property values and environmental risks evolve over time. For example, while our model identified the importance of coastal proximity, it did not fully account for **future risks** such as rising sea levels and increasing flood hazards due to climate change. These emerging risks could significantly impact property values in the coming years, and the current model may not fully capture these dynamics.

In terms of methodology, while the GLM provided reliable and interpretable results, other models like Generalised Additive Models (GAMs) or Interaction Models might have offered better fits for the data, particularly in capturing non-linear relationships. While we chose the GLM for its balance between complexity and interpretability, future work could explore these alternative models further.

Suggestions for Future Work

Moving forward, there are several avenues for improving and expanding upon this research. **Incorporating more comprehensive data** would be a key improvement. For example, adding variables related to **property age**, **building condition**, and **local economic factors** would provide a more complete picture of what drives property values. Additionally, including **long-term environmental risk factors**, such as projections for sea-level rise and increased flood frequency, would offer more forward-looking insights, especially in a city like Christchurch, which is vulnerable to environmental changes.

Another area for future work is the exploration of more **sophisticated modelling techniques**. While the GLM was effective, future research could benefit from models that better capture non-linearities and complex interactions between variables. Models such as GAMs or even machine learning approaches could be explored to improve the predictive accuracy of the analysis.

Finally, it would be valuable to consider the **temporal aspect** of property values in future studies. A **time-series analysis** that tracks how property values evolve over time, particularly in response to environmental changes or urban development projects, could provide deeper insights into the dynamics of the property market.

What Would I Do Differently?

If I were to undertake this project again, I would focus on improving the **data collection process** from the outset. Ensuring that the dataset is both comprehensive and clean would save considerable time during the analysis phase and lead to more accurate results. I would also explore the use of **alternative statistical models** earlier in the process to identify the best fit for the data.

Furthermore, I would place more emphasis on **future-proofing the analysis** by incorporating climate-related data to assess the potential long-term risks to property values. This would provide more meaningful insights into how Christchurch's property market may evolve in

response to environmental changes, which is crucial given the increasing importance of sustainability and risk management in the real estate sector.

In conclusion, while this project successfully met its objectives and provided valuable insights into the Christchurch property market, there is room for further refinement and exploration. Future projects could build on this work by incorporating more comprehensive data, exploring advanced modelling techniques, and considering long-term environmental risks, ensuring that the findings remain relevant in an evolving property landscape.

6. REFERENCES

Datasets:

Christchurch City Council. (2023). Coast (OpenData): Spatial data of the mean high water springtide. Christchurch City Council Spatial Open Data Portal. https://opendata-christchurchcity.hub.arcgis.com/datasets/6eb2466f5b044bac8781ca43c62fe907_5/exlore

Christchurch City Council. (2023). DP flood hazard high (OpenData): Spatial data of flood hazard zones. Christchurch City Council Spatial Open Data Portal. <https://opendata-christchurchcity.hub.arcgis.com/maps/0b36313059734d7d82f6e7690393f857/about>

Land Information New Zealand (LINZ). (2023). NZ Properties: National District Valuation Roll. Toitū Te Whenua Land Information New Zealand. <https://data.linz.govt.nz/table/114085-nz-properties-national-district-valuation-roll/>

Valbiz V8. (2022). Christchurch property transaction data (1993–2022). Valbiz V8: Valuer Specific Software. Valbiz V8 provides valuation management tools to enable property and sales record updates, client management, and geospatial analysis.

Articles and Papers:

Bailey, J., Lauria, D., Lindquist, W., Mittnik, S., & Rachev, S. (2022). Hedonic models of real estate prices: GAM models; Environmental and sex-offender-proximity factors. *Journal of Risk and Financial Management*, 15(12), 601. <https://doi.org/10.3390/jrfm15120601>

Barnard, P., Erikson, L., Foxgrover, A., Hart, J., Limber, P., O'Neill, A., Ormond, M., Vitousek, S., Wood, N., Hayden, M., & Jones, J. (2019). Dynamic flood modeling essential to assess the coastal impacts of climate change. *Scientific Reports*, 9, 40742. <https://doi.org/10.1038/s41598-019-40742-z>

- Bishop, K., Kuminoff, N., Banzhaf, H., Boyle, K., von Gravenitz, K., Pope, J., Smith, V., & Timmins, C. (2020). Best practices for using hedonic property value models to measure willingness to pay for environmental quality. *Review of Environmental Economics and Policy*, 14(2), 260-281. <https://doi.org/10.1093/reep/reaa001>
- Catma, S. (2021). The price of coastal erosion and flood risk: A hedonic pricing approach. *Oceans*, 2(1), 1-9. <https://doi.org/10.3390/OCEANS2010009>
- Chen, W., Li, X., & Hua, J. (2019). Environmental amenities of urban rivers and residential property values: A global meta-analysis. *The Science of the Total Environment*, 693, 133628. <https://doi.org/10.1016/j.scitotenv.2019.133628>
- Filippova, O., Nguyen, C., Noy, I., & Rehm, M. (2020). Who cares? Future sea-level rise and house prices. *Land Economics*, 96(2), 207-224. <https://doi.org/10.3368/le.96.2.207>
- Hamilton, S., & Morgan, A. (2010). Integrating lidar, GIS, and hedonic price modeling to measure amenity values in urban beach residential property markets. *Computers, Environment and Urban Systems*, 34(2), 133-141. <https://doi.org/10.1016/j.compenvurbsys.2009.10.007>
- Heberger, M., Cooley, H., Herrera, P., Gleick, P. H., & Moore, E. (2011). Potential impacts of increased coastal flooding in California due to sea-level rise. *Climatic Change*, 109(S1), 229-249. <https://doi.org/10.1007/s10584-011-0308-1>
- Hinkel, J., Lincke, D., Vafeidis, A. T., Perrette, M., Nicholls, R. J., Tol, R. S. J., Marzeion, B., Fettweis, X., Ionescu, C., & Levermann, A. (2014). Coastal flood damage and adaptation costs under 21st century sea-level rise. *Proceedings of the National Academy of Sciences*, 111(9), 3292-3297. <https://doi.org/10.1073/pnas.1222469111>
- Jiao, L., & Liu, Y. (2010). Geographic field model-based hedonic valuation of urban open spaces in Wuhan, China. *Landscape and Urban Planning*, 98(1), 47-55. <https://doi.org/10.1016/j.landurbplan.2010.07.009>

- Koning, K., Filatova, T., & Bin, O. (2018). Improved methods for predicting property prices in hazard-prone dynamic markets. *Environmental and Resource Economics*, 69(2), 247-263. <https://doi.org/10.1007/s10640-016-0076-5>
- McNamara, D., Gopalakrishnan, S., Smith, M. D., & Murray, A. B. (2015). Climate adaptation and policy-induced inflation of coastal property values. *PLOS ONE*, 10(3), e0121275. <https://doi.org/10.1371/journal.pone.0121275>
- Nicholls, R. J. (2011). Planning for the impacts of sea-level rise. *Oceanography*, 24(2), 144-157. <https://doi.org/10.5670/oceanog.2011.34>
- Rajapaksa, D., Zhu, M., Lee, B., Hoang, V., Wilson, C., & Managi, S. (2017). The impact of flood dynamics on property values. *Land Use Policy*, 67, 225-236. <https://doi.org/10.1016/j.landusepol.2017.08.038>
- Rosato, P., Breil, M., Giupponi, C., & Berto, R. (2017). Assessing the impact of urban improvement on housing values: A hedonic pricing and multi-attribute analysis model for the historic centre of Venice. *Buildings*, 7(4), 112-137. <https://doi.org/10.3390/buildings7040112>
- Vousdoukas, M., Mentaschi, L., Voukouvalas, E., Bianchi, A., Dottori, F., & Feyen, L. (2018). Climatic and socioeconomic controls of future coastal flood risk in Europe. *Nature Climate Change*, 8, 776-780. <https://doi.org/10.1038/s41558-018-0260-4>
- Warren-Myers, G., Fuerst, F., Aschwanden, G., & Üürke, E. (2018). Estimating the risk of sea-level rise to property values. *ERES Conference Proceedings*. https://doi.org/10.15396/ERES2018_192

7. APPENDICES

Appendix 1: GLOSSARY/ACRONYMS

This appendix provides in-depth explanations of the technical terms and acronyms used in the report, offering further clarity for readers unfamiliar with statistical or real estate concepts.

General Terms

- **Akaike Information Criterion (AIC):** A measure of the goodness-of-fit for a statistical model that includes a penalty for the number of parameters. A lower AIC value indicates a model that achieves a better balance between fit and complexity.
- **Bayesian Information Criterion (BIC):** Similar to AIC, BIC also measures the goodness-of-fit of a model but places a heavier penalty on models with more parameters. A lower BIC value suggests a more parsimonious model.
- **Generalised Additive Model (GAM):** An extension of the Generalised Linear Model that allows for the inclusion of non-linear relationships between the predictors and the response variable by using smooth functions, making it useful for capturing complex data patterns.
- **Generalised Linear Model (GLM):** A flexible extension of linear regression that allows for non-normal distributions of the dependent variable. GLM is widely used in real estate analysis when dealing with skewed data like property values.
- **Hedonic Pricing Model (HPM):** A regression model used in real estate to estimate how much individual property characteristics contribute to the overall property value, such as land area, location, and proximity to amenities.
- **Cook's Distance:** A diagnostic measure used to identify influential data points in a regression model. Data points with a large Cook's Distance value have a disproportionate influence on the model's parameter estimates and should be carefully examined.
- **Deviance:** In regression models, deviance measures how well the model fits the data. Lower residual deviance compared to null deviance indicates an improved fit when predictors are included.
- **Leverage:** A measure that identifies data points that are distant from the mean of the predictor variables. High-leverage points can disproportionately influence the model's results.
- **Multicollinearity:** A situation in regression analysis where two or more predictor variables are highly correlated, making it difficult to isolate their individual effects. This can lead to inflated standard errors and unreliable coefficient estimates.
- **Null Deviance:** A measure of the deviance in a model that includes only the intercept. It serves as a baseline to evaluate the residual deviance when predictors are added.

- **Ordinary Least Squares (OLS):** A method used in linear regression models to estimate the parameters by minimizing the sum of squared residuals (the differences between observed and predicted values).
- **Pearson Residuals:** Residuals calculated based on Pearson's chi-square statistic. These are used to assess model fit in Generalised Linear Models and are expected to be randomly distributed around zero.
- **Pseudo R-Squared (McFadden's R^2):** A version of the R-squared statistic adapted for models like GLM. It measures the proportion of variability explained by the model but is generally lower than the R-squared seen in ordinary regression.
- **Root Mean Squared Error (RMSE):** A metric that measures the average deviation between the predicted and actual values of the dependent variable. Lower RMSE values suggest better model accuracy.
- **Variance Inflation Factor (VIF):** A diagnostic tool used to detect multicollinearity in regression models. VIF values greater than 10 indicate high multicollinearity, which can make coefficient estimates unreliable.

Acronyms

- **AIC:** Akaike Information Criterion
- **BIC:** Bayesian Information Criterion
- **GLM:** Generalised Linear Model
- **GAM:** Generalised Additive Model
- **HPM:** Hedonic Pricing Model
- **RMSE:** Root Mean Squared Error
- **VIF:** Variance Inflation Factor

This glossary provides detailed explanations for readers to understand the technical terminology and statistical concepts used in the report.

Appendix 2: Detailed Literature Review

This appendix provides a more in-depth exploration of the literature that informs the development of property valuation models, particularly focusing on the Hedonic Pricing Model (HPM), Generalised Linear Models (GLMs), and the integration of Geographic Information Systems (GIS) and lidar data. These methodologies are key in understanding the effects of environmental factors such as coastal proximity and flood risks on property values.

Hedonic Pricing Model (HPM) in Property Valuation

The Hedonic Pricing Model (HPM) is widely employed in real estate economics to evaluate the impact of property attributes, including environmental and locational factors, on property values. One notable study by Hamilton and Morgan (2010) utilised GIS and lidar data to quantify the effects of ocean views and proximity to beaches on property prices in coastal urban areas. This integration allowed for a precise assessment of how access to coastal amenities, such as beaches and views, drives up property values. Their approach can be adapted to Christchurch's coastal properties, where similar factors significantly influence market prices.

Rajapaksa et al. (2017) studied the impact of flood risks on property prices, applying a semi-parametric HPM. This study revealed that properties located in flood-prone areas experience a decrease in value, though the degree of this effect depends on how quickly the property market recovers after a flood event. Given that many properties in Christchurch are located within flood zones, this model provides an insightful framework for understanding the long-term impact of environmental risks on property values.

Generalised Linear Models (GLMs) and Non-Linear Relationships

Generalised Linear Models (GLMs) offer a flexible framework for modelling property values when data exhibits non-normal distributions or when relationships between variables are non-linear. Bailey et al. (2022) applied Generalised Additive Models (GAMs), an extension of GLMs, to capture non-linear interactions between environmental risks and property values. While GAMs provide greater flexibility, GLMs are more suitable when relationships between variables are well-understood, as is the case in Christchurch's coastal property market. The GLM's ability to handle complex data distributions was one of the main reasons for its selection in this project.

Filippova et al. (2020) studied the undercapitalization of environmental risks, such as rising sea levels, on property values along New Zealand's Kapiti Coast. The study revealed that property prices in coastal regions often fail to fully account for long-term risks, even after official warnings are issued. This finding is relevant to Christchurch, where flood zones pose significant risks, and it highlights the importance of integrating environmental factors into property valuation models.

Environmental Risks and Property Values

McNamara et al. (2015) developed a stochastic dynamic model to explore how rising sea levels and erosion control measures influence property values. The study concluded that human interventions, such as beach nourishment, can inflate property prices temporarily, but abrupt declines are likely if such interventions are discontinued. This dynamic interaction is highly relevant to Christchurch's coastal properties, where similar environmental risks must be considered when predicting long-term property values.

Barnard et al. (2019) introduced a dynamic flood modelling approach that assesses the combined impacts of sea-level rise, storm surges, and erosion on coastal property values. Their

study underscored the limitations of static models, which often underestimate short-term risks such as storm events. Incorporating dynamic environmental factors into property valuation models, as Barnard et al. (2019) suggest, will improve the accuracy of property price predictions in Christchurch's coastal areas.

GIS and Lidar Integration in Coastal Property Valuation

The integration of GIS and lidar technology in property valuation has significantly improved model accuracy. Warren-Myers et al. (2018) demonstrated how the use of high-resolution elevation data from lidar improved property value estimates by up to 15% in Melbourne. This methodology is particularly relevant to Christchurch, where accurate elevation and coastal proximity data are crucial for assessing flood risks and estimating property values.

In a similar vein, Barnard et al. (2019) applied dynamic flood modelling to assess how sea-level rise and coastal erosion affect property values. Their research showed that integrating both short- and long-term environmental risks into property valuation models improves the reliability of price estimates in coastal regions. This approach aligns with the needs of Christchurch, where flood risk and proximity to the coast are critical factors influencing property prices.

Conclusion

The literature strongly supports the use of Hedonic Pricing Models and Generalised Linear Models for evaluating how environmental factors, such as coastal proximity and flood risks, impact property values. By incorporating high-resolution environmental data through GIS and lidar, these models can provide more accurate predictions of property values. For Christchurch, where coastal amenities and flood risks are key market drivers, these methodologies will be instrumental in developing a robust and reliable property valuation model.

Appendix A: Detailed Data Cleaning and Handling of Invalid Data

1. Filtering Non-Coastal Areas

Given the project's focus on coastal property values in Christchurch, it was critical to filter out non-coastal properties. This was achieved using the *MMQGIS Plugin* in QGIS software, which allowed the transformation of property addresses into geographic coordinates (latitude and longitude). The following steps were taken:

- **Geocoding Process:** Using the MMQGIS Plugin, all property addresses in the dataset were converted to geographic coordinates. This process assigned latitude and longitude to each property, enabling accurate spatial analysis (see Figure A1).
- **Coastal Suburb Identification:** The geocoded data was imported into a map of Christchurch, and properties in coastal suburbs were marked. Only those properties

located in identified coastal suburbs, such as New Brighton and Southshore, were retained for further analysis. Inland properties, or those without direct relevance to the study of coastal proximity, were removed from the dataset. This filtering ensured that the analysis was focused exclusively on properties that could be influenced by coastal amenities and environmental factors, such as flooding risks.

2. Handling of Missing and Invalid Data

During the data preparation phase, it was crucial to address incomplete and inaccurate entries that could undermine the integrity of the analysis. The following steps were employed to clean the data:

- **Elimination of Records with Missing Key Variables:** Numerous records lacked essential information, such as sale dates or total property values. As these variables are vital for calculating property value changes and for accurate analysis, records missing these critical fields were removed from the dataset. This ensured that only complete transaction records were included in the final analysis.
- **Removal of Irrelevant Factors:** Certain fields in the dataset contained either false or irrelevant information. For instance, the “Brms” field, which presumably referred to the number of bedrooms, consistently showed values of zero, making it an unusable variable for our study. Such irrelevant fields were eliminated to maintain the clarity and focus of the analysis.

3. Outlier Detection and Removal

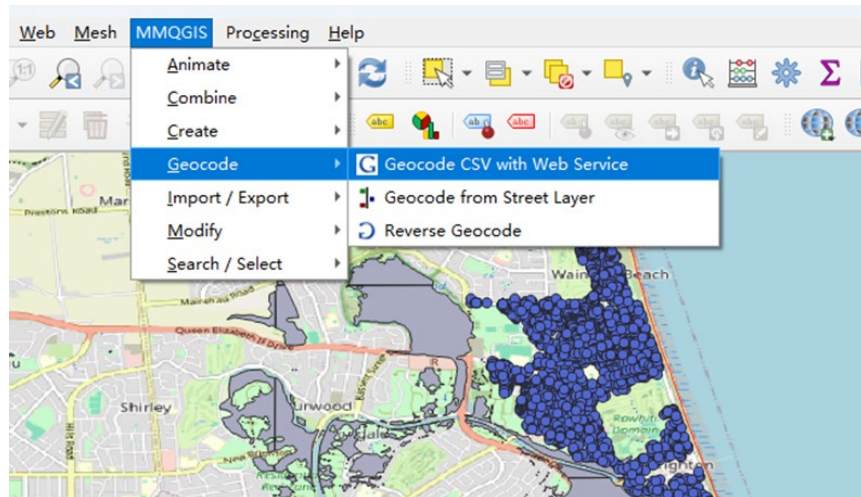
The dataset contained some extreme outliers that could have skewed the results. These outliers were identified and removed based on the following criteria:

- **Example of 272 Marine Parade:** One notable example of an outlier was the property at 272 Marine Parade, which showed a 2500% increase in value between 2002 and 2006 (from NZD 172,000 to NZD 4,300,000). Upon further investigation, it was discovered that this transaction represented the sale of four townhouses in a single transaction, significantly inflating the recorded property value. Such extreme outliers, which did not reflect the normal market trends, were eliminated to ensure that they did not distort the final analysis (see Figure A2).

By following these rigorous data cleaning procedures, the dataset was refined to include only relevant, accurate, and complete information, thus providing a strong foundation for the subsequent statistical analysis.

Figures

- **Figure A1:** Geocoding Process in QGIS using the MMQGIS Plugin



- **Figure A2:** Property at 272 Marine Parade – Example of Outlier from Ray White Property Listing



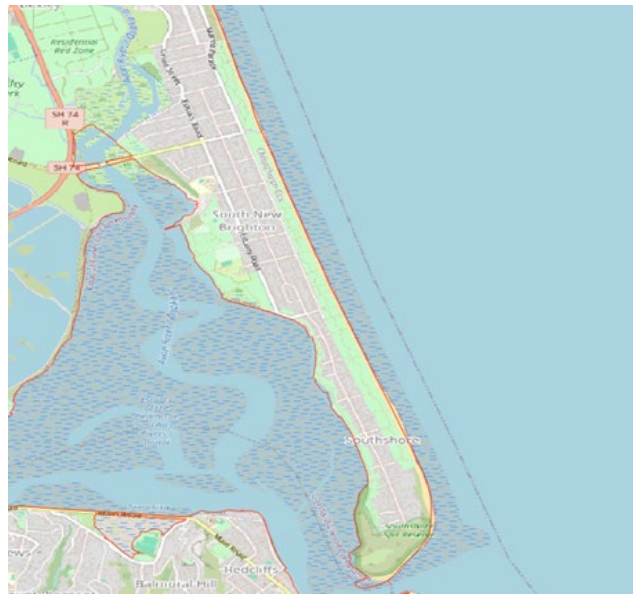
Appendix B: Detailed GIS Software Preparation and Spatial Analysis

(a) Coastal Proximity and Flood Zone Data Collection

To prepare the geographical data required for the analysis, we utilised QGIS software, integrating two key spatial datasets relevant to coastal proximity and flood risk:

1. **Coastline Shapefile:**

- **Source:** The shapefile was sourced from the Christchurch City Council's Spatial Open Data Portal, under the dataset titled "Coast (OpenData)."
 - **Description:** This shapefile outlines the boundary between land and sea at the mean high water spring tide. It includes vector data that depicts the coastal line of Christchurch, spanning across key coastal suburbs like Southshore and New Brighton. The red line in the figure (Figure B1) represents this coastal boundary, which was crucial in calculating the proximity of properties to the coastline.
 - **Purpose:** The proximity to the coast was computed based on this data, with the straight-line (Euclidean) distance between each property and the coastline being added as a variable in the property dataset.
2. **Figure B1:** Coastline shapefile from Christchurch City Council, represented as the red boundary between land and sea.



3. Flood Zone Shapefile:

- **Source:** The flood zone shapefile was also downloaded from the Christchurch City Council's Spatial Open Data Portal under the dataset titled "DP Flood Hazard High (OpenData)."
- **Description:** This dataset highlights areas that are susceptible to extreme flood risks, particularly those facing 1-in-500-year flood events. It is displayed as polygons that cover flood-prone areas in Christchurch. The grey-shaded areas in the figure (Figure B2) show these flood zones, specifically covering regions like South New Brighton.
- **Purpose:** This shapefile allowed us to identify properties within these flood-prone areas, adding a binary variable 'FloodingZone' to the dataset, indicating whether a property is within a flood zone.

4. **Figure B2:** Flood zone shapefile depicting areas at risk of extreme flood events (shown in grey).



These spatial data layers formed the foundation for subsequent geographical analyses, allowing us to incorporate key environmental factors such as proximity to the coast and flood risks into our property valuation models.

(b) Coastal Proximity and Flood Zone Analysis

The spatial analysis conducted in QGIS focused on calculating the distance of each property to the coastline and determining whether properties were situated within flood-prone zones. The detailed process is outlined below:

1. **Geocoding Property Locations:**
 - The property addresses in our dataset were geocoded using the **MMQGIS plugin** within QGIS, converting addresses into geographic coordinates (latitude and longitude). This enabled us to plot the properties on the map of Christchurch, which was essential for spatial analysis (as seen in **Figure B3**).
2. **Figure B3:** Geocoded property points plotted on the map of Christchurch using MMQGIS in QGIS.



3. Calculating Coastal Proximity:

- Once the properties were geocoded, the **spatial join tool** was used to measure the distance between each property and the nearest coastline. This allowed us to create the 'DistanceToCoast' variable, representing the Euclidean distance between a property and the coastline in metres. This variable, illustrated in the figure above (Figure B1), was later transformed into a logarithmic scale for improved model fit.

4. **Relevance:** Coastal proximity is an important factor in Christchurch's property market, as properties near the coast are typically more valuable due to lifestyle and amenity benefits. This was a key variable in our hedonic pricing model.

5. Overlaying Flood Zone Data:

- The flood zone shapefile (Figure B2) was overlaid onto the geocoded property locations to determine whether each property was within a flood-prone area. Using QGIS's **spatial join tool**, a binary variable 'FloodingZone' was created, with '1' indicating a property located within a flood zone and '0' for properties outside the zones.

6. **Relevance:** Understanding flood risk is vital for property valuation, particularly in coastal regions where environmental risks can significantly impact property values. This variable added depth to our analysis by incorporating potential hazards.

7. Integrating Geographical Variables into the Dataset:

- The 'DistanceToCoast' and 'FloodingZone' variables were added to our property transaction dataset, allowing these spatial factors to be considered in the hedonic pricing and Generalised Linear Models (GLM) used for property value analysis.

8. **Impact:** By including geographical data, we enhanced the depth of our analysis, making it possible to accurately assess how proximity to natural amenities like the coast and exposure to environmental risks like flooding influence property values in Christchurch.

In conclusion, the integration of these spatial variables allowed for a more nuanced understanding of how environmental and locational factors affect property values. The data collection and spatial analysis steps described in this appendix formed the basis for the statistical analysis and modelling that followed.

Appendix C: Detailed Model Selection and Justification

Introduction to the Hedonic Pricing Model (HPM)

The **Hedonic Pricing Model (HPM)** is widely used in real estate valuation to disaggregate property prices into the value contributions of individual characteristics. It allows researchers to quantify the impact of specific attributes such as land area, floor area, and proximity to amenities like coastlines on overall property values. In its basic form, the HPM follows a linear equation:

Price = $\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n + \epsilon$,
where:

- **Price:** The dependent variable representing the property price.
- **β_0 :** The intercept, indicating the base property price when all attributes are set to zero.
- **X_1, X_2, \dots, X_n :** The independent variables representing property characteristics like land area, floor area, and proximity to the coast.
- **$\beta_1, \beta_2, \dots, \beta_n$:** The coefficients corresponding to these characteristics, measuring their impact on property prices.
- **ϵ :** The error term representing variations not explained by the model.

In this project, HPM was initially used to explore how factors such as coastal proximity, water views, and environmental risks, such as flood zones, influence property values in Christchurch. For instance, **Hamilton & Morgan (2010)** demonstrated that ocean views substantially increase property prices, while **Rajapaksa et al. (2017)** found that flood zones tend to decrease property values. By breaking down these characteristics, HPM helps estimate the marginal contribution of each factor to property prices.

While the HPM was effective as a starting point, its limitations prompted us to adopt more flexible models like the **Generalised Linear Model (GLM)** to better handle non-linear relationships and non-normal data distributions.

Limitations of HPM

1. Assumption of Linearity:

The HPM assumes that the relationship between property characteristics and prices is linear, meaning that each additional unit of a factor (e.g., one metre closer to the coast) contributes the same value to the property price. However, this is not always the case. For example, the marginal value of proximity to the coast may diminish as properties get closer to the water. In such cases, non-linear models are better suited to capture the complexity of real estate markets.

2. Inability to Handle Non-Normal Distributions:

HPM assumes that residuals (the differences between predicted and actual property prices) follow a normal distribution. However, property prices often exhibit skewness, especially in markets with extreme price variations. For instance, properties with very high values can disproportionately influence the model. Such skewed distributions violate the normality assumption, leading to biased estimates.

3. Lack of Interaction Effects:

HPM does not easily capture interaction effects between variables. For example, the effect of coastal proximity on property values may depend on whether the property is in a flood zone. While HPM treats these variables independently, more advanced models like GLMs can account for such interactions, providing a more nuanced understanding of property price determinants.

4. Heteroscedasticity:

HPM assumes constant variance (homoscedasticity) in residuals, meaning that the variability in property prices is consistent across different values of independent variables. However, in real estate markets, high-value properties tend to show greater variability in prices. This phenomenon, known as heteroscedasticity, violates HPM assumptions and can distort the model's estimates.

5. Sensitivity to Outliers:

Outliers, such as properties with unusually high transaction prices or erroneous data, can heavily influence HPM results. During the data cleaning process, outliers such as the extreme price increase of **272 Marine Parade** (shown in **Figure A2**) were identified and removed to prevent them from skewing the analysis.

Transition to Generalised Linear Model (GLM)

Due to the limitations of HPM, the analysis was extended to more flexible models, particularly the **Generalised Linear Model (GLM)**, which can accommodate non-normal distributions and

account for more complex relationships between variables. The GLM allowed us to model property prices using a **Gamma distribution** with a log link function, overcoming many of the limitations inherent in HPM.

The theoretical framework provided by HPM remained crucial to understanding how individual factors influenced property values, but the flexibility of the GLM was essential to accurately capturing non-linear relationships and ensuring a more robust analysis. The full details of the GLM implementation are discussed in **Section 4.1.2** of the main report.

Appendix D: Detailed Model Assumptions and Checking

This appendix provides a thorough explanation of the assumption checks performed for each model used in the statistical analysis, including GLM, GAM, polynomial regression, and interaction models. These checks were critical in ensuring the validity and reliability of the results produced in this study.

1. Generalised Linear Model (GLM)

Key Assumptions

- **Linearity of Predictors:** The relationship between the independent variables (e.g., distance to the coast, land area) and the dependent variable (property value) is assumed to be linear on the transformed (logarithmic) scale.
- **Independence of Observations:** Each property transaction must be independent, with no clustering or repeated measures unless accounted for.
- **Distribution of Residuals:** The residuals should follow a Gamma distribution, as chosen to address the skewness in property price data.
- **No Perfect Multicollinearity:** Predictors must not be perfectly correlated, as this can lead to unstable coefficient estimates.

Assumption Checks

- **Linearity:** Diagnostic plots, such as the **Residuals vs Fitted values** plot, were used to assess whether the transformed predictors exhibit a linear relationship with the dependent variable. These plots showed no obvious deviations from linearity.
- **Independence:** Independence of property transactions was ensured by filtering the data to include only individual property records, eliminating any clustering effects.
- **Residual Distribution:** The Gamma distribution assumption was verified through residual plots, showing that the residuals followed the expected distribution, addressing the skewed property price data.

- **Multicollinearity:** The **Variance Inflation Factor (VIF)** was calculated for each predictor. VIF values below 10 confirmed that multicollinearity was not a concern, indicating that no predictor was too highly correlated with another.

2. Generalised Additive Model (GAM)

Key Assumptions

- **Smoothness of Predictors:** GAM assumes that relationships between the predictors and the dependent variable can be captured using smooth functions.
- **Independence of Observations:** Similar to GLM, each property transaction must be independent.
- **Residual Distribution:** The residuals should follow a Gamma distribution.

Assumption Checks

- **Smoothness of Predictors:** Diagnostic plots confirmed that the smoothing functions adequately captured the relationships between predictors, without overfitting. The number of degrees of freedom for the smoothing terms was also monitored to prevent overfitting.
- **Independence and Residual Distribution:** These assumptions were validated similarly to GLM, including tests for Gamma-distributed residuals.

3. Polynomial Regression

Key Assumptions

- **Linearity in Parameters:** While the relationship between predictors and property values is non-linear, the model remains linear in terms of its parameters, allowing for OLS estimation.
- **Normality of Residuals:** The residuals should follow a normal distribution, which can be problematic with highly skewed data.
- **Homoscedasticity:** The variance of residuals must be constant across all levels of the independent variables.

Assumption Checks

- **Linearity in Parameters:** This assumption was met inherently by the nature of polynomial regression.
- **Normality and Homoscedasticity:** Residual diagnostic plots were used to verify that residuals followed a normal distribution. The **Breusch-Pagan test** was employed to test for constant variance, confirming homoscedasticity.

4. Interaction Models

Key Assumptions

- **Correct Specification of Interaction Terms:** Interaction terms must be properly specified to accurately capture the combined effects of predictors.
- **Linearity:** The underlying relationship between predictors and the dependent variable should still be linear, even with interaction terms.
- **Normality and Independence:** The residuals should be normally distributed, and data points must be independent.

Assumption Checks

- **Specification of Interaction Terms:** Diagnostic plots and significance tests confirmed that interaction terms captured the relationships between variables effectively.
- **Linearity, Normality, and Independence:** The **Shapiro-Wilk test** confirmed the normality of residuals, and multicollinearity was assessed to ensure predictor relationships did not distort the model results.

Summary of Assumption Checks

For each model, the following checks were conducted to ensure assumptions were met:

- **Residual vs Fitted Plots:** To check for linearity and homoscedasticity.
- **Shapiro-Wilk Test:** To assess the normality of residuals.
- **Breusch-Pagan Test:** To detect heteroscedasticity issues.
- **Variance Inflation Factor (VIF):** To check for multicollinearity.

By addressing these assumptions, the robustness of the models used in this study was confirmed, ensuring reliable and accurate property valuation results.

Appendix E: Detailed Model Comparison

This appendix provides an in-depth comparison of the statistical models used in the analysis: Generalised Linear Model (GLM), Generalised Additive Model (GAM), Polynomial Regression, and Interaction Models. Each model was assessed based on its performance in explaining property prices in Christchurch's coastal areas, using various metrics like Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Squared Error (RMSE).

1. Generalised Linear Model (GLM)

The **GLM** with a Gamma family and log link function was selected as one of the primary models for its flexibility in handling non-normal data distributions, which aligns well with the skewed nature of property prices. By applying the log link, GLM transforms the relationship between predictors and property prices into a linear form, making it easier to interpret. The main strengths of GLM include:

- **Handling of Non-Normal Residuals:** GLM is designed for data that are not normally distributed, which makes it suitable for property price data that tend to be skewed.
- **Dealing with Heteroscedasticity:** The model handles non-constant variance (heteroscedasticity) well, making it a reliable choice for datasets where property values vary widely.
- **Interpretability:** Coefficients are easy to interpret as they represent multiplicative effects on the outcome variable after applying the log transformation.

However, GLM assumes linear relationships between predictors and the log-transformed outcome variable. While effective in many cases, this assumption might not always hold, especially for non-linear trends in the data.

2. Generalised Additive Model (GAM)

The **GAM** extends GLM by allowing the relationships between predictors and the dependent variable to be non-linear. It applies smooth functions to continuous predictors, such as distance to the coast, allowing more flexibility in capturing complex trends. Key strengths of GAM include:

- **Non-Linear Relationships:** GAM's flexibility in modelling non-linear relationships is advantageous, particularly for variables like coastal proximity that do not follow strict linear patterns.
- **Smoothing Functions:** Instead of assuming a linear relationship, GAM fits smooth curves to the data, better capturing environmental and geographical factors.

However, while GAM's flexibility is an advantage, it also introduces a risk of overfitting, particularly when the dataset has limited data points. This model is also more challenging to interpret compared to GLM due to the nature of the smooth functions applied to the predictors.

3. Polynomial Regression

Polynomial regression introduces non-linear terms by including squared or higher-order versions of the predictors. This model is useful when the relationship between a predictor and the outcome variable follows a curved, rather than linear, pattern. The strengths of polynomial regression include:

- **Capturing Curvature:** By adding polynomial terms (e.g., squared terms), this model can capture non-linear relationships between property characteristics and values.
- **Simplicity:** It is relatively straightforward to implement, and unlike GAM, it does not require specialised smoothing functions.

However, polynomial regression can lead to **multicollinearity**—where predictor variables are highly correlated, particularly when higher-degree terms are added. This increases the risk of overfitting, making the model less generalisable to new data.

4. Interaction Models

Interaction models explore how two or more variables together influence the dependent variable. For example, the interaction between coastal proximity and water views may provide additional insights into property value. Key strengths of interaction models include:

- **Capturing Compound Effects:** These models are useful when the combined effect of two or more variables provides more information than considering them independently. For example, properties with water views might show different value trends based on their proximity to the coast.
- **Addressing Complex Relationships:** Interaction models allow for the exploration of how one variable modifies the effect of another, offering a more nuanced understanding of property value drivers.

However, these models can become complex, and interpreting the coefficients—particularly for higher-order interactions—can be challenging. Moreover, if not specified correctly, interaction terms can introduce instability into the model.

5. Model Performance Comparison

Each model was evaluated using key performance metrics: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Squared Error (RMSE). Additionally, cross-validation was performed to assess the generalisability of the models to new data.

- **AIC and BIC:** Both AIC and BIC penalise models for complexity while rewarding goodness-of-fit. Lower values indicate better-performing models. In this analysis, GAM had the lowest AIC and BIC scores, reflecting its flexibility in fitting complex relationships. However, the GLM also performed well, providing a good balance between fit and simplicity.
- **RMSE:** The RMSE values, which measure the average deviation of predicted values from actual values, were slightly lower for GAM, but the differences between models were marginal. This indicates that all models performed comparably in terms of prediction accuracy.

- **Cross-Validation:** Cross-validation results showed that while GAM had the best fit on the training data, it exhibited signs of overfitting during validation. In contrast, GLM performed consistently well across both training and validation sets, indicating a more robust model for predicting new data.

6. Final Model Selection

After considering all metrics, the **Generalised Linear Model (GLM)** was selected as the final model for this project. Although GAM offered greater flexibility and slightly better fit in some instances, GLM provided a robust and interpretable model that balances performance with simplicity. It was particularly suitable for capturing the effects of key factors like coastal proximity and water views, offering both theoretical soundness and practical insights.

In conclusion, the model comparison shows that while more flexible models like GAM and Polynomial Regression offer additional features, the GLM remains the best choice for this analysis due to its reliability, interpretability, and stable performance.

Appendix F: Sales Price Index (SPI) Approach Details

The **Sales Price Index (SPI) Approach** was employed to estimate property values for 2022, particularly for properties that did not have a recorded transaction in that year but had historical sales data. The methodology provides a consistent means to estimate these missing values, ensuring that the dataset used for property analysis is comprehensive and up-to-date.

1. Data Preparation

The first step involved preparing the historical sales data for properties with at least two recorded sales. The critical elements for each property were:

- **Property ID:** A unique identifier for each property.
- **Sale Year:** The year when the property was sold.
- **Sale Price:** The price of the property at the time of sale.

This data allowed us to track the price changes over time, enabling the calculation of a price index. By examining these historical sales, the SPI method estimates how much property prices have changed between the recorded sale years and 2022.

2. Calculating Log Price Differences

To model the change in property prices over time, we used the log price difference between two sales for each property. The log transformation linearises the price changes, making them more suitable for regression analysis. The formula is given by:

$$\text{Log Price Difference} = \log(\text{Sale Price Year 2}) - \log(\text{Sale Price Year 1})$$

This transformation enables us to capture percentage changes in property values over time. The log transformation is particularly helpful when dealing with skewed data, as is often the case with real estate transactions.

3. Creating Year Dummy Variables

Dummy variables were created for each year in the dataset, marking the first and second sales for each property as follows:

- **-1** for the year of the first sale.
- **+1** for the year of the second sale.
- **0** for all other years.

This structure ensures that the model captures the price change between the two sale dates accurately. These dummy variables were then used in the regression model to calculate the index values for each year.

4. Setting the Base Year

A base year is selected—usually the most recent year before 2022. The price index for this year is set to **1** (log index = 0), and all other years are indexed relative to the base year. For example, if 2021 is the base year, the price index for this year is set to **1**, and all other years are compared against it.

5. Ordinary Least Squares (OLS) Regression

The OLS regression method is employed to estimate the price index for each year. The log price differences serve as the dependent variable, while the dummy year variables are the independent variables. The model is expressed as follows:

$$\log(P_j) - \log(P_i) = \beta_j - \beta_i$$

Where:

- **P_j** and **P_i** are the sale prices in years **j** and **i**, respectively.
- **β_j** and **β_i** are the coefficients representing the log of the price index for years **j** and **i**.

This regression model allows us to calculate the log price index for each year relative to the base year.

6. Converting Log Coefficients to Price Indices

After obtaining the coefficients from the regression, these values represent the log of the price index. To convert these into the actual price indices, we use the exponential function:

$$\text{Price Index for Year } X = \exp(\beta x)$$

This step provides the price index for each year, which is used to estimate the future property prices for years when no sales data is available.

7. Estimating 2022 Prices

To estimate the 2022 prices for properties that were last sold in earlier years, we use the following formula:

$$\text{Estimated 2022 Price} = \text{Sale Price in Last Sale Year} * (\text{Price Index for 2022} / \text{Price Index for Last Sale Year})$$

For instance, if a property was sold for \$300,000 in 2010, and the price index for 2010 is 0.75, and the price index for 2022 is 1.20, the estimated price in 2022 would be:

$$\text{Estimated 2022 Price} = \$300,000 * (1.20 / 0.75) = \$480,000$$

This calculation ensures that every property in the dataset, regardless of its last sale date, has an estimated price for 2022. This allows for a consistent and comprehensive dataset to be used in the hedonic pricing model and GLM analyses.

Conclusion

The SPI approach provides a robust solution for estimating the current market value of properties that haven't been sold recently. Unlike traditional repeat sales methods, this approach enables the estimation of future property values, ensuring consistency across the dataset and facilitating a comprehensive analysis of property values in Christchurch's coastal areas.

By implementing this method, we ensure that our hedonic pricing and GLM models are applied to a complete dataset, with estimated prices for all properties, making the analysis more accurate and comprehensive.

Appendix G: Detailed Cook's Distance Analysis for Influential Data Handling

G.1 Introduction

During the statistical analysis, identifying and removing influential data points was crucial to ensure the reliability of the results. Influential points can have a disproportionate effect on the model's estimates, potentially leading to skewed or biased findings. In this appendix, we provide a detailed explanation of the steps involved in identifying and handling influential data using Cook's Distance.

G.2 Cook's Distance Methodology

Cook's Distance is a diagnostic measure used to detect influential observations in regression models. It estimates how much the model's coefficients would change if a particular data point were removed. The higher the Cook's Distance value, the more influence a data point has on the overall model.

The threshold for identifying influential points is calculated using the formula:

$$4/(n-k-1)$$

Where:

- **n** is the total number of observations in the dataset.
- **k** is the number of predictor variables in the model.

Any data point with a Cook's Distance value above this threshold is considered influential and may need to be removed to maintain the accuracy of the model.

G.3 Implementation in R

The steps below outline how we applied Cook's Distance analysis to the dataset using R.

Initial Model Fitting

The Generalised Linear Model (GLM) was fitted to the dataset, with property values as the dependent variable and key property attributes (e.g., proximity to the coast, land area, floor area) as predictors.

```
initial_model <- glm(LogValue2022 ~ LogDistance + LogLand + LogFloor + WaterView +  
FloodingZone + Suburb,  
data = initial_data, family = Gamma(link = "log"))
```

Calculation of Cook's Distance

Cook's Distance was calculated for each data point in the model using the `cooks.distance()` function. A plot of Cook's Distance was generated to visually identify influential points.

```
cook_dist <- cooks.distance(initial_model)

# Plot Cook's Distance
plot(cook_dist, type = "h", main = "Cook's Distance for Influential Data Detection", ylab =
"Cook's Distance")
abline(h = 4/(nrow(initial_data)-length(initial_model$coefficients)-1), col = "red")
```

The plot, as shown in **Figure G.1**, illustrates the Cook's Distance values for each observation, with the red line indicating the threshold for detecting influential points.

Figure G.1 – Cook's Distance for Influential Data Detection

Elimination of Influential Points

Data points with Cook's Distance values above the threshold were removed from the dataset to prevent them from skewing the model's results. This process is outlined below:

```
threshold <- 4/(nrow(initial_data) - length(initial_model$coefficients) - 1)
cleaned_data <- initial_data[cook_dist < threshold, ]
```

This step ensured that any points with disproportionate influence were eliminated, resulting in a more balanced and accurate dataset.

Refitting the Model

After removing the influential points, the Generalised Linear Model was refitted using the cleaned dataset to produce more reliable results.

```
refined_model <- glm(LogValue2022 ~ LogDistance + LogLand + LogFloor + WaterView +
FloodingZone + Suburb,
data = cleaned_data, family = Gamma(link = "log"))
```

G.4 Outcome

By eliminating the influential data points using Cook's Distance, we ensured that the final model was not unduly influenced by outliers. This resulted in a more robust dataset, referred to as **cleaned_data**, which provided accurate and reliable estimates for the property value analysis.

The refined model produced better-fitting results and increased the overall validity of the analysis.

Appendix H: Detailed Data Preparation and Cleaning Process

This appendix provides a detailed exploration of the data preparation and cleaning process conducted for the project, focusing on the cleaned dataset and the key statistical insights derived from it. The steps outlined below further explain how the data was handled to ensure a robust analysis of Christchurch's coastal property market.

1. Dataset Overview and Cleaning Process

The original dataset included thousands of property transactions, many of which required significant cleaning to ensure the accuracy of the final analysis. This involved removing records with incomplete or invalid entries (such as missing sale dates or property values) and transforming variables to better suit the requirements of the hedonic pricing model (HPM) and Generalised Linear Model (GLM) approaches.

The final cleaned dataset includes **2,466 property transactions**, all within Christchurch's coastal regions. The key variables in the dataset are Distance to Coast, Land Area, Floor Area, and Property Value (2022).

2. Key Descriptive Statistics

The **Descriptive Statistics** table (Figure H1) provides an overview of these key variables, summarising their mean, median, standard deviation, and range. This analysis offers valuable insights into how these factors contribute to property values in Christchurch's coastal market.

<i>Distance to Coast</i>		<i>Land Area</i>		<i>Floor Area</i>		<i>Value 2022</i>	
Mean(m)	592.6644526	Mean(m2)	672.5251419	Mean(m2)	139.404704	Mean(NZ\$)	703817.4566
Standard Error(m)	9.043164312	Standard Error(m2)	5.446175243	Standard Error(m2)	1.207789485	Standard Error(NZ\$)	5373.243267
Median(m)	414.135	Median(m2)	612	Median(m2)	120	Median(NZ\$)	674144.5479
Mode(m)	1453.47	Mode(m2)	506	Mode(m2)	110	Mode(NZ\$)	552726.0265
Standard Deviation	449.0730141	Standard Deviation	270.4507236	Standard Deviation	59.97742004	Standard Deviation	266829.0066
Sample Variance	201666.572	Sample Variance	73143.59389	Sample Variance	3597.290915	Sample Variance	71197718750
Kurtosis	-0.334467618	Kurtosis	51.39187378	Kurtosis	6.441947639	Kurtosis	1.585348012
Skewness	0.90957769	Skewness	5.460554296	Skewness	1.712674023	Skewness	0.807465098
Range(m)	1737.07	Range(m2)	3760	Range(m2)	636	Range(NZ\$)	2127943.598
Minimum(m)	44.47	Minimum(m2)	144	Minimum(m2)	24	Minimum(NZ\$)	9906.985246
Maximum(m)	1781.54	Maximum(m2)	3904	Maximum(m2)	660	Maximum(NZ\$)	2137850.584
Sum(m)	1461510.54	Sum(m2)	1658447	Sum(m2)	343772	Sum(NZ\$)	1735613848
Count	2466	Count	2466	Count	2466	Count	2466

Figure H1: Descriptive Statistics of Key Variables

Distance to Coast (metres):

- Mean: 592.66 metres
- Median: 414.14 metres
- Range: 44.47 – 1,781.54 metres

Land Area (square metres):

- Mean: 672.53
- Median: 612
- Range: 144 – 3,904

Floor Area (square metres):

- Mean: 139.40
- Median: 120
- Range: 39 – 660

Property Value (NZD, 2022):

- Mean: \$703,817.46
- Median: \$674,144.55
- Range: \$9,906 – \$2,065,894

3. Graphical Analysis

Three key graphs further illustrate the distribution and characteristics of properties in the dataset. Each graph highlights important aspects of property valuation in Christchurch's coastal areas, and these insights are critical for interpreting the final statistical models.

3.1 Histogram of Distance to Coast

The **Histogram of Distance to Coast** (Figure H2) reveals that a large number of properties are clustered closer to the coastline, with fewer properties located inland. The cumulative curve shows that over half the properties are within 600 metres of the coast, illustrating the significance of coastal proximity in this market.

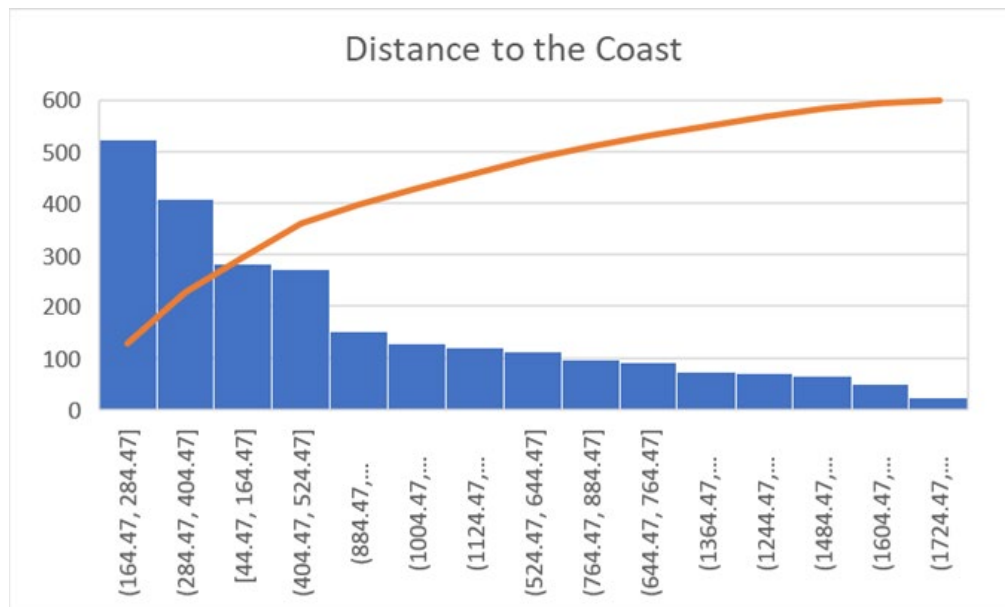


Figure H2: Histogram of Distance to Coast

3.2 Histogram of Property Value (2022)

The **Histogram of Property Value** (Figure H3) displays the skewed nature of property values, with most properties falling below the \$1 million mark. The distribution also shows several high-end properties contributing to the upper end of the market.

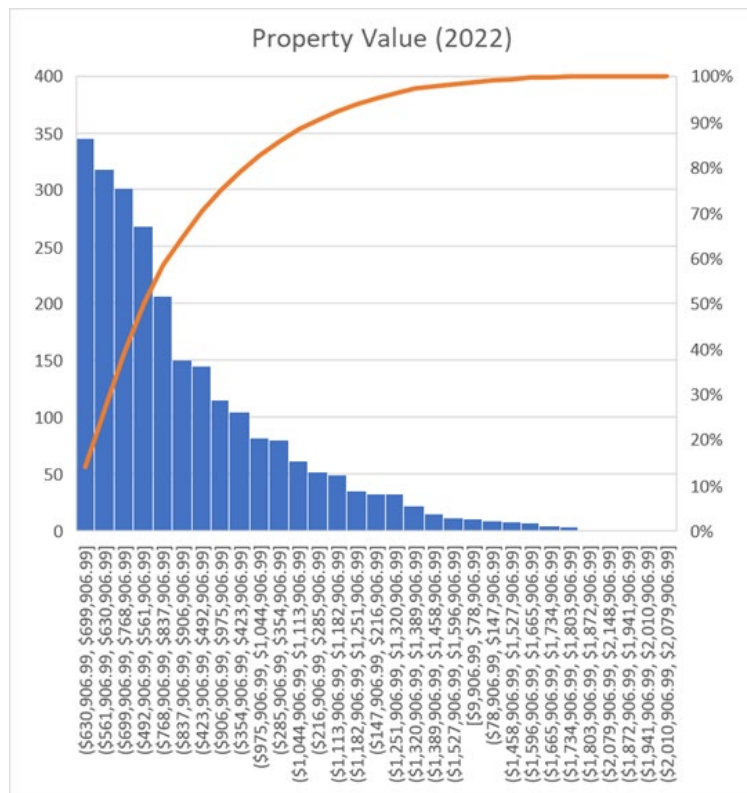


Figure H3: Histogram of Property Value (2022)

3.3 Boxplot of Property Value (2022)

The **Boxplot of Property Value** (Figure H4) highlights the existence of outliers, particularly at the higher end of the property market. These outliers suggest the presence of premium coastal properties that are likely in high demand due to their size, location, or ocean views.



Figure H4: Boxplot of Property Value (2022)

4. Interpretation of Findings

The descriptive statistics and graphs present several important trends:

- **Proximity to Coast:** The significant number of properties located within 600 metres of the coast suggests that coastal proximity is highly desirable in Christchurch's property market. Properties closer to the coast typically command higher values, reflecting their appeal for lifestyle and recreational access.
- **Land and Floor Area:** The variation in property size, both in terms of land and floor area, points to a diverse market where large homes coexist with smaller, more modest properties. Larger properties with substantial land areas are often situated in premium locations, which contributes to their higher market value.
- **Property Value Distribution:** The skewed distribution of property values reflects the general structure of real estate markets, where a small number of high-value properties inflate the average, but most properties fall within the mid-range price bracket.

These observations offer a more in-depth understanding of the dataset and help contextualise the results derived from the statistical analysis. The cleaned dataset is not only comprehensive but also reflective of Christchurch's unique coastal property dynamics.

Appendix I: Model Comparison and Final Model Selection

In this appendix, we provide an in-depth comparison of the statistical models evaluated during the project, including the Generalised Linear Model (GLM), Generalised Additive Model (GAM), Polynomial Model, and Interaction Models. We explore the performance of these models using key metrics such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and diagnostic checks to justify the final model choice for property value analysis in Christchurch's coastal regions.

1. Generalised Linear Model (GLM)

The Generalised Linear Model (GLM) was applied using a Gamma family and log link function to accommodate the non-normal, skewed distribution of property prices. The flexibility of GLM allows it to handle non-negative, positively skewed data, which is a critical consideration for the dataset used in this study.

Key advantages of the GLM include:

- The ability to manage non-normal residuals.
- Suitable for heteroscedastic data, where variance changes with the predictors.
- Straightforward interpretation of coefficients due to the log transformation.

The GLM assumes a linear relationship between the predictors (e.g., proximity to the coast, land area) and the log-transformed dependent variable, making it a solid choice for modelling non-linear relationships without overly complicating the model.

2. Generalised Additive Model (GAM)

The Generalised Additive Model (GAM) is an extension of GLM, allowing for more flexibility in modelling non-linear relationships through the use of smoothing functions. GAM was particularly effective in capturing non-linear trends in the proximity to the coast and property value relationship.

However, while the GAM outperformed other models in terms of goodness-of-fit (with the lowest AIC and BIC scores), it introduced complexity in interpretation. GAM's strength lies in its ability to fit curves that adapt to the data's shape, but the flexibility also made the model more prone to overfitting, particularly with a dataset of this size.

3. Polynomial Regression

Polynomial Regression was applied by adding non-linear terms, such as squared and interaction terms, to the model. This approach aimed to capture the curvature in relationships between property characteristics and value. While polynomial regression is useful for modelling more complex relationships, it tended to increase the risk of multicollinearity and was prone to overfitting when higher-degree polynomials were used.

4. Interaction Models

Interaction models were tested to evaluate whether the combined effects of multiple factors, such as coastal proximity and water view, had a compounded influence on property prices. While interaction models provided insights into how factors worked together, their complexity and the potential for multicollinearity limited their practical application in this analysis.

5. AIC and BIC Results

To compare the models, we used the AIC and BIC criteria, which penalise model complexity and favour more parsimonious models:

- **Log-Linear Model:** AIC = 274.41, BIC = 337.74
- **Polynomial Model:** AIC = 275.85, BIC = 356.46
- **Interaction Model:** AIC = 270.39, BIC = 356.76
- **GAM Model:** AIC = 262.48, BIC = 361.05
- **GLM Model:** AIC = 319.21, BIC = 382.55

As shown, the GAM model had the lowest AIC score, indicating the best balance between model complexity and goodness-of-fit. However, the GLM, despite having a higher AIC and BIC, was chosen for several key reasons outlined below.

6. Justification for Choosing GLM

1. **Data Characteristics:** The property value data was characterised by non-negative values and significant skewness. The GLM, with a Gamma distribution and log link, was ideal for handling such characteristics. The alternative models, such as the Log-Linear Model and Polynomial Regression, were less suited to address the distributional features of the data.
2. **Interpretability:** One of the key reasons for selecting GLM was its interpretability. The log transformation applied in the model allowed the coefficients to be interpreted as multiplicative effects on the response variable. This is particularly useful in real estate analysis, where understanding the marginal impact of factors like land area or proximity to the coast on property value is critical for decision-making. For example, the positive coefficient for land area indicates that a one-unit increase in land area results in a proportional increase in property value. This interpretability aligns

well with the theoretical framework of the Hedonic Pricing Model (HPM), making GLM a more intuitive and practical choice for this project.

3. **Model Flexibility:** Although GAM provided greater flexibility in capturing non-linear trends, it introduced challenges in interpretability, particularly with the non-parametric effects of smoothing functions. GLM, by contrast, offered sufficient flexibility to model non-linear relationships and interaction terms while remaining easier to interpret.
4. **Suitability for Data Distribution:** The Gamma family in the GLM provided a better fit for the positively skewed property value data. The Log-Linear Model and Polynomial Regression assumed normally distributed errors, which were not appropriate for this dataset. By using GLM, the analysis could better reflect the real-world distribution of property prices, particularly in coastal areas where values are heavily influenced by proximity to amenities like beaches.
5. **Diagnostic Checks:** Several diagnostic checks were conducted to assess the performance of each model. In particular, Cook's Distance plots were used to identify and remove influential data points that could distort the results. The residual analysis indicated that the GLM was well-specified, with no clear patterns in the residuals. Cross-validation confirmed the predictive validity of the GLM, further supporting its selection as the final model.

Despite the GAM offering better fit in terms of AIC and BIC, the GLM performed more consistently in terms of interpretability, robustness, and the ability to generalise beyond the training data.

7. Conclusion

The Generalised Linear Model (GLM) was selected as the final model for this project. While other models, such as GAM and Polynomial Regression, provided better statistical fits in some cases, the GLM struck the best balance between performance, interpretability, and theoretical consistency with the Hedonic Pricing Model. Diagnostic checks confirmed that the GLM was well-suited to the skewed nature of the property data and provided robust, reliable estimates for key factors such as proximity to the coast, water views, and land area.

Appendix J: Detailed Statistical Model Implementation - Generalised Linear Model (GLM)

In this appendix, we provide a comprehensive explanation of the Generalised Linear Model (GLM) used in this project to analyse the impact of various property characteristics on property values in Christchurch. The model was implemented using R, with the final dataset consisting of key variables such as coastal proximity, land area, floor area, water views, flood zone status, and the suburb location.

Model Setup

The GLM was chosen due to its ability to handle skewed, non-negative property values, which are typical in real estate datasets. To capture the multiplicative relationship between property attributes and their values, a Gamma distribution with a log link function was applied. This configuration ensures that the model can handle the inherent variability in property prices and produce more accurate estimates.

Model Formula:

$$\text{Log}(E[\text{Value2022}]) = \beta_0 + \beta_1 * \text{LogDistance} + \beta_2 * \text{LogLand} + \beta_3 * \text{LogFloor} + \beta_4 * \text{WaterView} + \beta_5 * \text{FloodingZone} + \beta_6 \dots \beta_n * \text{Suburb}$$

Where:

- **Log(E[Value2022]):** Expected property value in 2022, transformed using the natural logarithm.
- **LogDistance:** The logarithm of the distance from the property to the coast.
- **LogLand:** The logarithm of the land area.
- **LogFloor:** The logarithm of the floor area.
- **WaterView:** Binary variable indicating whether the property has a water view (1 = yes, 0 = no).
- **FloodingZone:** Binary variable indicating whether the property is in a flood zone (1 = yes, 0 = no).
- **Suburb:** Categorical variables representing the different suburbs in the dataset.

Key Features of the Model

1. **Gamma Distribution:** Used to manage the skewed distribution of property prices, which tend to be non-negative and have a long tail on the right (high-value properties).
2. **Log Link Function:** Facilitates the modelling of multiplicative relationships between the predictors and the response variable, making the effects of variables like coastal proximity proportional to property values.
3. **Handling of Categorical and Continuous Variables:** Both categorical variables (e.g., suburbs, water views) and continuous variables (e.g., land and floor area) are effectively incorporated into the model, adding flexibility.

Model Results and Interpretation

The R output of the GLM implementation is summarised below, and the coefficients were interpreted based on their statistical significance and real-world relevance. A screenshot of the

R output is included (see **Figure 1** below), which displays the estimated coefficients, standard errors, t-values, and p-values for each predictor variable.

```
> summary(glm_model)

Call:
glm(formula = LogValue2022 ~ LogDistance + LogLand + LogFloor +
     WaterView + FloodingZone + Suburb, family = Gamma(link = "log"),
     data = cleaned_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.4499513   0.0099740  245.635 < 2e-16 ***
LogDistance   -0.0040203   0.0006398   -6.283 3.94e-10 ***
LogLand         0.0013243   0.0014940    0.886 0.375478
LogFloor        0.0325724   0.0012362   26.348 < 2e-16 ***
WaterView1      0.0065159   0.0024743    2.633 0.008508 **
FloodingZone1  -0.0017250   0.0015983   -1.079 0.280566
SuburbNorth New Brighton  0.0052544   0.0009826    5.347 9.80e-08 ***
SuburbSouth New Brighton  0.0066766   0.0013330    5.009 5.89e-07 ***
SuburbSouthshore  0.0070728   0.0021148    3.345 0.000837 ***
SuburbWaimairi Beach  0.0236598   0.0016685   14.180 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.0003655362)

Null deviance: 1.57173 on 2339 degrees of freedom
Residual deviance: 0.86346 on 2330 degrees of freedom
AIC: 319.21

Number of Fisher Scoring iterations: 3
```

(Figure 1: R Output of GLM Model Summary)

Key Coefficient Estimates:

- **Intercept (2.4499):** The baseline property value, when all other variables are at their reference levels, corresponds to approximately $\exp(2.4499) = 11.59$ on the original price scale. This represents a base value for properties located in the reference suburb (New Brighton).
- **LogDistance (-0.0040):** The negative coefficient indicates that as the distance from the coast increases, property values decrease. Specifically, for every 1% increase in distance to the coast, property values drop by about 0.4%, holding other variables constant. This effect is statistically significant ($p < 0.001$).
- **LogLand (0.0013):** Although the coefficient is positive, suggesting a minor increase in value with larger land areas, this effect is not statistically significant ($p = 0.375$).
- **LogFloor (0.0326):** The coefficient for floor area indicates that a 1% increase in floor area leads to a 3.3% increase in property value. This effect is highly significant.

($p < 0.001$), underscoring the importance of property size in determining market value.

- **WaterView (0.0065):** Properties with a water view tend to have higher values, with a 0.7% increase for properties with a water view compared to those without. This effect is statistically significant ($p = 0.0085$).
- **FloodingZone (-0.0017):** Being in a flood zone has a slight negative impact on property value, but this effect is not statistically significant ($p = 0.2806$).
- **Suburbs:** Positive coefficients for North New Brighton (0.5% increase), South New Brighton (0.7% increase), Southshore (0.7% increase), and Waimairi Beach (2.4% increase) highlight the desirability of these suburbs, with Waimairi Beach showing the strongest effect on property value.

Model Diagnostics

Several diagnostic checks were performed to ensure the model's robustness and accuracy:

- **Residual Analysis:** Residual plots showed no significant patterns, indicating that the model is correctly specified.
- **Cook's Distance:** Influential points were identified and removed to ensure that no single observation unduly affected the model's results.
- **Goodness-of-Fit:** The model's residual deviance (0.8635) was substantially lower than the null deviance (1.5717), suggesting a good fit to the data. The Akaike Information Criterion (AIC) score of 319.21 indicates that the model strikes a reasonable balance between complexity and explanatory power.

Conclusion

The Generalised Linear Model (GLM) provided valuable insights into the key factors influencing property values in Christchurch's coastal regions. While proximity to the coast, floor area, and the presence of a water view significantly affect property prices, the model also accounted for suburb-level variations, with Waimairi Beach properties showing the greatest value premiums. For full R output details and coefficient estimates, refer to **Figure 1**.

Appendix K: Detailed Statistical Findings for General Understanding

This appendix provides an in-depth explanation of the Generalised Linear Model (GLM) analysis results, along with detailed interpretations of how various factors influence property values in Christchurch. The findings are presented in accessible language to ensure that readers without a statistics background can understand the insights.

1. Proximity to the Coast (LogDistance)

In our analysis, properties closer to the coast were found to have higher values. The model shows a slight but statistically significant decrease in property value as the distance from the coast increases. This relationship is particularly relevant in Christchurch, where proximity to popular beaches, such as New Brighton and Southshore, significantly boosts property desirability.

- **Key Finding:** For every 1% increase in distance from the coast, property values decrease by approximately 0.4%. This is a modest but significant effect, indicating the high demand for coastal properties in Christchurch.
- **Real-World Implication:** Coastal living is a major driver in the local property market. Buyers are willing to pay a premium for properties with easy access to beaches and scenic coastal environments.

2. Land and Floor Area (LogLand and LogFloor)

The model showed that while land size does not have a statistically significant effect on property values, the floor area of the home plays a much more important role. Larger homes tend to be more valuable, and the relationship between floor area and property value was both positive and significant.

- **Key Finding:** A 1% increase in floor area results in an approximate 3.3% increase in property value, holding other factors constant. This highlights the importance of interior space in determining the value of residential properties.
- **Real-World Implication:** For prospective homebuyers and investors, increasing the interior floor space of a home may offer better returns than focusing solely on land area. This is particularly relevant in well-developed urban areas where land size may be less important compared to the actual livable space.

3. Water View and Flooding Zone

Properties with a view of the water (whether ocean or river) were found to be more valuable than those without. However, properties in flood-prone areas exhibited only a slight reduction in value, with the effect being statistically insignificant.

- **Key Finding (Water View):** Properties with water views have an approximate 0.7% higher value compared to those without. Scenic views add a premium to property prices.
- **Key Finding (Flood Zone):** Properties located in flood-prone areas see a small, non-significant decrease in value (-0.17%). This suggests that, for now, flood risks may not heavily influence Christchurch's property market, although future awareness of climate risks may change this trend.

- **Real-World Implication:** While scenic views consistently increase property values, flood risks do not yet seem to be fully capitalised into property prices. Buyers continue to prioritise location and views over environmental concerns like flooding, although this may change as climate-related risks become more prominent.

4. Suburban Location

Suburban location was another critical factor influencing property values. Properties in suburbs like Waimairi Beach, Southshore, and North New Brighton were more valuable than those in the reference suburb, New Brighton. This variation reflects the desirability of specific suburbs based on amenities, local infrastructure, and environmental appeal.

- **Key Finding:** Properties in Waimairi Beach showed the largest positive effect on value, with a 2.4% increase compared to New Brighton. Other suburbs, such as Southshore and North New Brighton, also demonstrated positive impacts, with 0.7% and 0.5% increases, respectively.
- **Real-World Implication:** Location is a key determinant of property value, with certain suburbs commanding higher prices due to their amenities, lifestyle appeal, and coastal proximity. This insight is valuable for both buyers and sellers looking to gauge property potential based on location.

Conclusion and Broader Implications

The findings from this GLM analysis provide actionable insights into the Christchurch property market. The model shows that proximity to the coast, floor area, scenic views, and suburban location all play significant roles in determining property values. The information from this analysis is valuable for different audiences:

- **For Homebuyers and Investors:** Buyers can make informed decisions based on property attributes that yield higher returns, such as coastal proximity and floor area.
- **For Real Estate Professionals:** Real estate agents and developers can use these insights to position properties better and highlight key attributes that add value.
- **For Urban Planners and Policy Makers:** Understanding which factors drive property values helps planners prioritize infrastructure and development projects in areas with high growth potential.

This detailed breakdown of the statistical findings gives a comprehensive view of how Christchurch's property market is influenced by environmental and structural factors. The results also serve as a valuable tool for those involved in real estate, property development, and urban planning.

Related Figures:

1. **Descriptive Statistics for Key Variables:** (Figure: Descriptive Statistics screenshot for variables such as Distance to Coast, Land Area, Floor Area, and Property Value (2022)).
2. **Histogram of Distance to Coast:** (Figure: Demonstrating property distribution relative to coastal proximity, highlighting the demand for coastal living).
3. **Histogram of Property Value (2022):** (Figure: Showing the skewed distribution of property values in Christchurch).
4. **Boxplot of Property Value (2022):** (Figure: Indicating outliers in the higher end of property values, often linked to premium factors like water views or coastal proximity).

Appendix L: Real-World Insights from Statistical Findings

This appendix provides an in-depth discussion of the real-world insights derived from the Generalised Linear Model (GLM) analysis. It elaborates on the relationship between property values, coastal proximity, suburban variations, and the role of urban planning in Christchurch's property market.

(a) Impact of Coastal Proximity on Property Values

One of the most significant findings from the Generalised Linear Model (GLM) analysis is the strong relationship between coastal proximity and property values in Christchurch. The model reveals a statistically significant negative correlation between the distance from the coast and property value, as evidenced by the negative coefficient for *LogDistance*. Specifically, the coefficient for *LogDistance* indicates that as the distance from the coast increases, property values tend to decrease. This outcome aligns with existing research demonstrating that properties located closer to the coast typically command a price premium due to the unique lifestyle and aesthetic advantages offered by coastal living.

Real-World Insight

This finding is consistent with global real estate market trends, where proximity to coastal amenities, such as beaches and scenic views, significantly enhances property values. Numerous studies have quantified the premium associated with coastal properties, highlighting that water views alone can add considerable value to real estate prices (Hamilton & Morgan, 2010). Coastal properties are often sought after due to their recreational appeal, lifestyle benefits, and exclusivity, with buyers willing to pay a premium for easier access to natural amenities like beaches (Chen, Li, & Hua, 2019). In Christchurch, suburbs like New Brighton and Sumner reflect this trend, as buyers value the opportunity to enjoy scenic coastal views and outdoor recreational activities.

Local Context: Christchurch's Coastal Desirability

Christchurch's coastal suburbs, such as New Brighton, Southshore, and Waimairi Beach, have long attracted homebuyers due to their unique mix of lifestyle and natural beauty. These areas not only offer easy access to beaches but also the appeal of coastal living, which remains a strong driver of demand in the Christchurch property market. The attractiveness of these suburbs is further enhanced by urban regeneration efforts, such as the New Brighton Regeneration Project. Research shows that urban regeneration projects often have a positive effect on property values, particularly in previously underperforming areas (Rajapaksa et al., 2017). In New Brighton, investment in new public amenities and infrastructure is expected to attract higher-income buyers, potentially leading to rising property prices in the coming years.

Studies suggest that socio-economic factors, such as higher household incomes and the increasing demand for lifestyle-oriented properties, are key drivers in maintaining elevated property values in coastal areas (Catma, 2021). In Christchurch, properties closer to the coast often command higher prices, reflecting both their immediate access to beaches and their broader appeal to buyers seeking a coastal lifestyle.

Climate Considerations: Long-Term Risks

While coastal proximity is a key factor driving property values, it is important to consider the long-term sustainability of these properties, particularly in light of climate change. Research on coastal property markets has increasingly highlighted the risks posed by rising sea levels, increased coastal erosion, and more frequent storm events, all of which threaten the long-term value of coastal properties (Filippova et al., 2020). In New Zealand and globally, properties in low-lying coastal areas are increasingly vulnerable to these environmental risks, which may lead to higher insurance costs and lower long-term market values. Studies have shown that while coastal properties continue to command a premium, buyers are beginning to factor in these risks, which could moderate future price growth in high-risk areas (Barnard et al., 2019).

In Christchurch, the GLM analysis shows that buyers are not yet fully discounting the risks associated with sea-level rise and coastal erosion, as reflected by the positive effect of coastal proximity on property values. However, as awareness of these risks increases and as climate-related costs—such as rising insurance premiums—become more apparent, the premium placed on coastal proximity may begin to diminish. Nonetheless, for the time being, the desirability of living near Christchurch's beaches continues to outweigh these concerns in the property market.

Comparison with Other Coastal Markets

The pattern observed in Christchurch mirrors trends seen in other New Zealand cities, such as Auckland and Wellington, where coastal properties consistently attract higher prices due to their premium locations. In Auckland, waterfront suburbs like Mission Bay and Kohimarama see sustained demand for properties with beach access and water views, despite concerns about

coastal erosion and flood risks (Warren-Myers et al., 2018). Similar trends are observed internationally, where coastal properties maintain higher values despite long-term environmental risks, underscoring the strong market demand for properties that offer both lifestyle benefits and proximity to natural amenities.

In Christchurch, the coastal suburbs continue to benefit from strong demand, reinforcing the notion that proximity to natural amenities like beaches remains a significant factor in property valuation. However, as environmental risks become more pressing, both buyers and policymakers may need to remain vigilant about future challenges associated with coastal living, potentially prompting shifts in market dynamics as these risks are more fully accounted for.

(b) Suburb-Level Property Value Variations

The Generalised Linear Model (GLM) analysis reveals that location plays a pivotal role in determining property values, with significant variation across different suburbs in Christchurch. This finding is consistent with existing literature on real estate pricing, where location-specific factors such as proximity to amenities, newer infrastructure, and socio-economic demographics are known to exert substantial influence on property values (Gyourko et al., 2013). In Christchurch, the model shows that properties in coastal suburbs like Southshore and Waimairi Beach command significantly higher prices compared to other areas, such as New Brighton, which served as the reference category.

Real-World Insight

The variation in property values across suburbs reflects both local amenities and broader market trends. Coastal properties are typically valued higher due to their lifestyle desirability, as they offer scenic views, proximity to beaches, and access to recreational activities (Hamilton & Morgan, 2010). This is evident in the data, with Waimairi Beach commanding the highest premium, as indicated by its large positive coefficient in the model. This suggests that properties in Waimairi Beach are considered more desirable compared to other coastal suburbs such as New Brighton and Southshore, likely due to factors such as newer housing developments and superior infrastructure.

Local Context: Christchurch Suburbs

Waimairi Beach is an example of a suburb benefiting from newer infrastructure and larger, modern homes. Studies have shown that such developments, especially when combined with proximity to natural amenities like beaches and parks, can significantly boost property values (McNamara et al., 2015). Its proximity to Bottle Lake Forest Park and the beach, coupled with

spacious, modern homes, makes Waimairi Beach highly appealing to buyers, particularly those seeking both lifestyle and investment potential.

In contrast, Southshore represents a more established coastal community, with slightly older housing stock. However, the area benefits from its peaceful environment and proximity to the Avon-Heathcote Estuary, attracting a mix of families and retirees. Research on socio-economic dynamics in real estate markets suggests that areas like Southshore, which cater to middle-income buyers, often experience steady demand despite their relatively older infrastructure (Filippova et al., 2020). The balance between affordability and desirable coastal access makes Southshore a consistent player in the property market.

Impact of Urban Regeneration on Property Values

New Brighton, while historically more affordable due to its older housing stock and economic challenges, particularly after the Christchurch earthquakes, is undergoing significant urban regeneration. Studies on urban renewal suggest that regeneration projects can lead to substantial property value increases in previously underperforming suburbs (Rajapaksa et al., 2017). In New Brighton, the development of the hot pool complex and the New Brighton Pier precinct has attracted new interest from buyers and investors. While these improvements have not yet been fully reflected in the property values, the long-term impact of such urban regeneration projects is likely to result in rising property prices, following the trend observed in other revitalised urban areas.

Suburb-Level Insights: Social and Economic Factors

The socio-economic composition of each suburb also plays a crucial role in property valuation. Waimairi Beach tends to attract higher-income buyers, contributing to a more affluent community with newer, larger homes. Research on real estate markets confirms that such socio-economic factors drive up property values, as wealthier buyers tend to invest in properties offering both lifestyle and prestige (Catma, 2021). In contrast, Southshore, while still desirable, has a broader range of buyers due to its older housing stock and slightly lower property prices, leading to a more mixed economic demographic.

In New Brighton, socio-economic challenges have historically kept property values lower. However, as urban regeneration projects progress and new amenities are introduced, the socio-economic profile of the suburb is likely to shift, attracting higher-income buyers and increasing demand. Studies on gentrification and urban renewal support this view, showing that infrastructure improvements and community development can lead to rising property values in once-underperforming suburbs (Barnard et al., 2019).

Long-Term Trends and Future Predictions

Looking ahead, suburbs like Waimairi Beach are expected to maintain their premium status due to their newer housing stock, coastal location, and appeal to higher-income buyers. Meanwhile, Southshore, with its family-friendly environment and proximity to natural reserves, will likely continue to see steady demand. However, the most significant changes are anticipated in New Brighton, where ongoing regeneration efforts are expected to drive property value growth. Studies on urban regeneration consistently show that such interventions lead to rising property values as the area becomes more desirable to buyers and investors (Rajapaksa et al., 2017).

(c) Urban Planning and Development Insights

The Generalised Linear Model (GLM) analysis provides valuable insights not only into property values but also into the broader implications for urban planning and development in Christchurch. The findings emphasise the importance of strategic urban development that enhances both property desirability and resilience against environmental challenges such as flooding and climate change risks. These are key considerations for coastal regions globally, including Christchurch, where the property market is closely tied to both environmental and infrastructure factors.

Impact of Urban Development Projects

Urban regeneration initiatives, such as the New Brighton Regeneration Project, are pivotal in reshaping the future of Christchurch's coastal suburbs. Research has shown that well-executed regeneration projects lead to improvements in infrastructure, public spaces, and amenities, thereby boosting property values and attracting new residents and investors (Adair et al., 2003). For example, the revitalisation of New Brighton, which includes new recreational facilities and enhanced public spaces, has contributed to rising property demand in the area, with the statistical analysis confirming a positive premium for properties in this suburb. Studies have demonstrated that urban regeneration can create an uplift in property values, particularly in previously underperforming or neglected areas, by making neighbourhoods more attractive to potential buyers and investors (Couch & Dennemann, 2000).

Incorporating amenities such as recreational spaces, modern transport links, and improved coastal infrastructure can significantly enhance the desirability of coastal suburbs. International research highlights that such developments, especially when combined with high-quality urban design, can raise property values in coastal regions (Filippova et al., 2020). For instance, in Christchurch's suburbs of Southshore and Waimairi Beach, new developments and infrastructural improvements have had a positive impact on property values. This mirrors global trends in cities like Auckland, where urban planning initiatives around waterfront areas, such as Wynyard Quarter, have resulted in a significant premium on nearby properties (Bailey et al., 2022).

Managing Environmental Risks

In coastal cities, urban planning must also take into account environmental risks, such as flooding and coastal erosion. While proximity to the coast remains a strong driver of property values, increasing concerns about the long-term impacts of climate change—including rising sea levels and more frequent extreme weather events—necessitate a shift towards more resilient urban development strategies (McNamara et al., 2015). Although properties in flood-prone areas of Christchurch show a slight reduction in value, the current impact on pricing is modest. However, as climate-related risks intensify, future property values could be more significantly affected, as has been observed in other coastal markets where the risk of sea-level rise is more pronounced (Barnard et al., 2019).

Urban planners are increasingly incorporating green infrastructure solutions to mitigate these risks. Flood defences, wetland restoration, and stormwater management systems are examples of sustainable interventions that help reduce the negative impacts of climate change on coastal properties (Pender & Neumann, 2014). These strategies not only protect existing infrastructure but also enhance the attractiveness of coastal areas by ensuring long-term viability in the face of environmental risks. For Christchurch, implementing these kinds of resilient infrastructure projects would help safeguard property values in coastal suburbs like New Brighton and Southshore, where climate-related risks are gradually becoming a more significant concern.

Future Growth Areas

The statistical model also points to potential growth areas within Christchurch's coastal regions. Suburbs such as Southshore and Waimairi Beach, which already benefit from their desirable coastal locations, could see further growth if urban development efforts focus on enhancing infrastructure and improving climate resilience. These findings are consistent with studies from other coastal cities where targeted urban development in high-demand areas has led to sustained increases in property values (Catma, 2021). By focusing resources on improving both physical infrastructure and environmental sustainability, urban planners can ensure long-term value for the city's housing market.

Urban Planning and Future Resilience

The analysis reinforces the critical role of thoughtful urban planning in shaping the future of Christchurch's coastal property market. While coastal areas continue to attract strong demand, planners must balance development with the need to address environmental risks. By investing in urban regeneration, sustainable infrastructure, and climate resilience, Christchurch can continue to offer attractive coastal living options while protecting long-term property values. These measures will be essential to ensuring that the city remains resilient in the face of future challenges, such as climate change and urban growth.

Conclusion: Strategic Urban Planning for Coastal Resilience

In summary, the integration of strategic urban planning and environmental sustainability is key to maintaining the desirability and viability of Christchurch's coastal suburbs. International evidence supports the view that thoughtful development, focused on resilience and sustainability, can lead to both economic and environmental benefits for coastal cities. The findings from this statistical analysis underscore the importance of enhancing infrastructure, mitigating environmental risks, and fostering urban regeneration to ensure the continued success of Christchurch's coastal property market.

Appendix M: Detailed Model Evaluation and Diagnostics

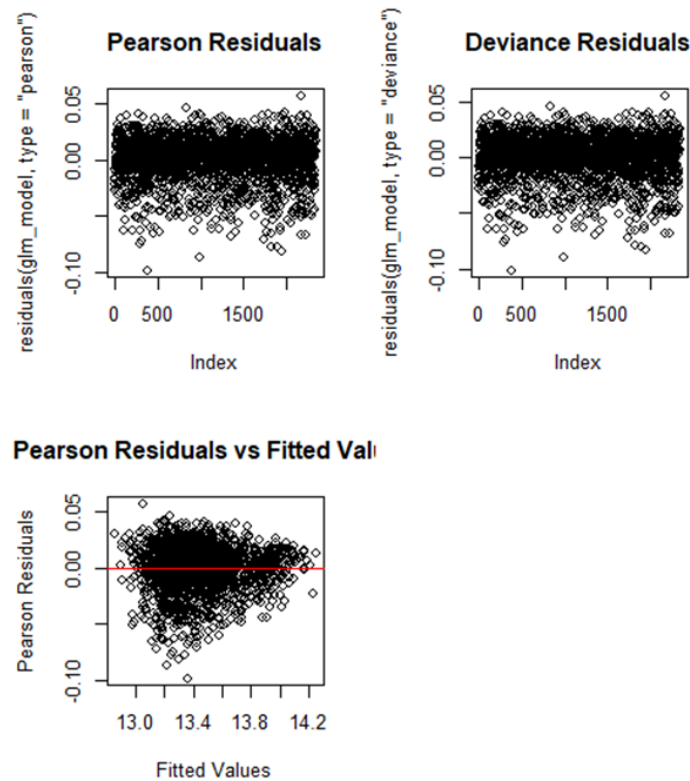
In this appendix, we provide a detailed explanation of the diagnostic checks and goodness-of-fit metrics that were performed to validate the Generalised Linear Model (GLM) used in the analysis of property values in Christchurch.

Residual Analysis

To ensure the robustness of the GLM, residual analysis was a critical step. Residuals help assess whether the model assumptions hold and whether there are any systematic patterns that might indicate a problem with the model's specification.

- **Pearson and Deviance Residuals:**

The Pearson and deviance residuals were plotted to examine the behaviour of the residuals relative to the fitted values. The plots showed a random scatter around zero, which indicates that the model does not exhibit issues such as heteroscedasticity (non-constant variance) or specification errors. Both residual types are commonly used to evaluate the goodness-of-fit in GLMs, and the random scatter is a good indication that the model is appropriately capturing the underlying relationships in the data.

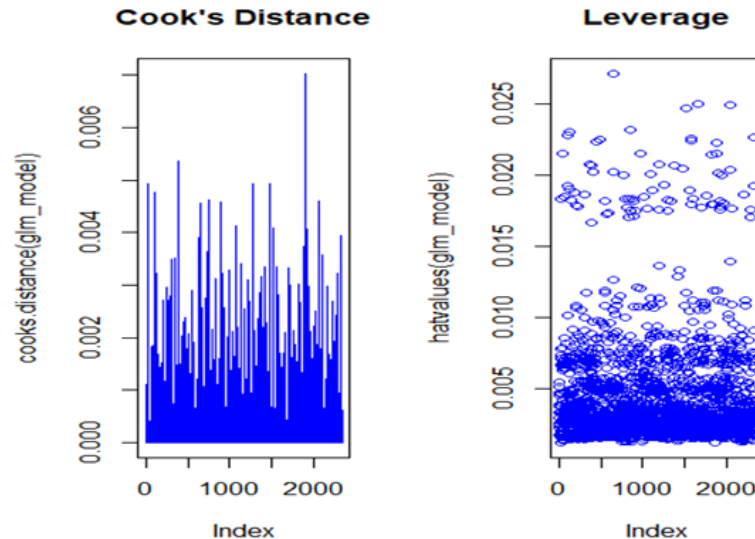


- *Figure: Residual plots, including Pearson and deviance residuals*
 The Pearson residuals vs fitted values plot confirmed that the residuals are evenly distributed and show no clear patterns, further validating the model's assumptions.

Influence and Leverage

Influential data points can disproportionately affect the model's estimates, and it is important to identify and mitigate these points to ensure model stability.

- **Cook's Distance:**
 Cook's Distance measures the influence of each data point on the overall model. Points with large Cook's Distance values suggest that they may be overly influential. For this analysis, a Cook's Distance plot was generated, and no points were found to exceed the threshold for concern ($4/n-k-1$).



- *Figure: Cook's Distance and leverage plots*

The Cook's Distance plot confirmed that none of the points had an undue impact on the model's results, ensuring the robustness of the model.

- **Leverage:**

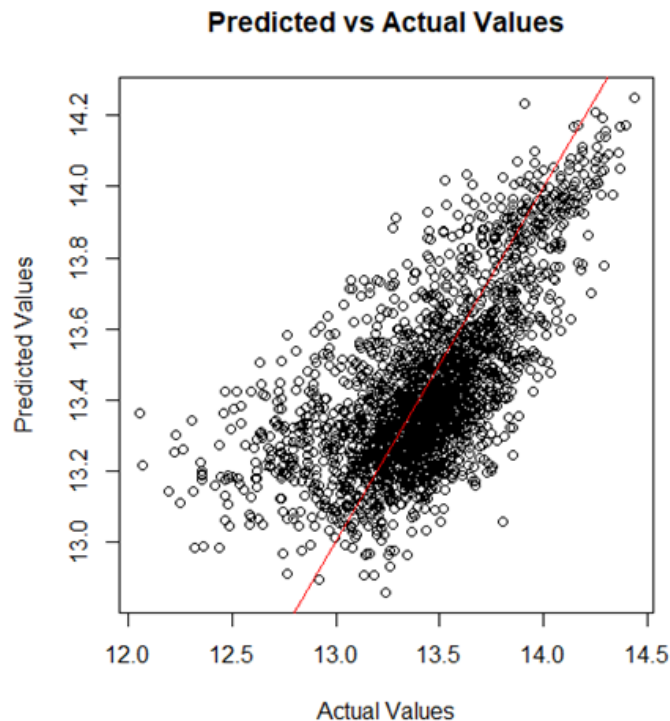
Leverage indicates how far an observation's predictor values are from the mean of the predictors. A plot of leverage versus residuals was examined, and no significant issues were found. The absence of high-leverage points confirms that the data is not skewed by outliers, ensuring the model's stability.

Predicted vs Actual Values

The alignment between predicted and actual values is essential for evaluating a model's accuracy.

- **Predicted vs Actual Values Plot:**

The "Predicted vs Actual Values" plot shows a strong linear relationship between the predicted property values and the actual values observed in the dataset. This demonstrates that the GLM captures a large portion of the variability in the data. The close alignment between the predicted and actual values reinforces the model's utility in predicting property values.



- *Figure: Predicted vs Actual Values plot*
 The model performs well in capturing the majority of variation in property values across different locations and property characteristics.

Goodness-of-Fit Metrics

Several goodness-of-fit metrics were used to assess the model's performance, providing insights into how well the model explains the variability in property values.

- **Null Deviance:**
 The null deviance of 1.5717 provides a measure of the total variability in the data when no predictors are included in the model. It serves as a baseline for comparison with the residual deviance, which measures how much of the variability remains after including the predictors.
- **Residual Deviance:**
 After including predictors such as LogDistance, LogLand, LogFloor, WaterView, FloodingZone, and Suburb, the residual deviance dropped significantly to 0.8635. This large reduction demonstrates that the predictors included in the model account for a substantial portion of the variability in property values, indicating that the model provides a much better fit than a simple intercept-only model.

- **Akaike Information Criterion (AIC):**

The AIC score of 319.21 evaluates the trade-off between model complexity and goodness-of-fit. Lower AIC values indicate a better model fit. The AIC score for this model suggests that it strikes a reasonable balance between simplicity and explanatory power, without overfitting the data.

- **McFadden's Pseudo R-Squared:**

The Pseudo R-squared value of 0.8237 indicates that the model explains approximately 82.37% of the variability in the log-transformed property values. This is a high value, signifying that the model captures most of the variation in the dataset, making it a reliable tool for understanding property value variations.

Conclusion

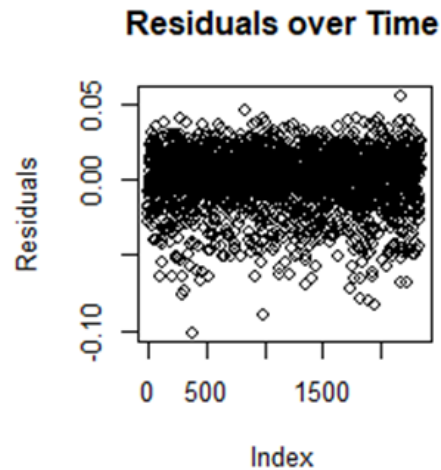
The diagnostic checks, including residual analysis, Cook's Distance, leverage plots, and goodness-of-fit metrics, all confirm the reliability and robustness of the Generalised Linear Model (GLM). The model performs well in predicting property values in Christchurch, with key factors such as coastal proximity, floor area, and water views showing significant influence. These diagnostics validate the GLM as an effective tool for understanding the property value variations in Christchurch, particularly for real estate professionals, urban planners, and policymakers.

Appendix N: Detailed Model Assumption Testing

This appendix provides an in-depth explanation of the key assumption tests conducted for the Generalised Linear Model (GLM) used in this project. The assumption checks include tests for independence of residuals, multicollinearity, influential points, and an evaluation of the special requirements for GLMs such as the non-necessity of normality and homoscedasticity checks.

1. Independence of Residuals

Independence of residuals ensures that the errors in the model are not correlated with one another. To check this assumption, a "Residuals over Time" plot was generated.



- **Interpretation:** As shown in the **Residuals over Time** plot, there is a random scatter of residuals with no discernible pattern, which suggests that the residuals are independent. This indicates that the GLM correctly captures the data without significant autocorrelation, and thus, the assumption of independence is satisfied. If a pattern had emerged, it could have signalled potential missing variables or model misspecification.

2. Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to determine their individual effects on the dependent variable. To assess multicollinearity, we used the Variance Inflation Factor (VIF).

```
> vif(glm_model)
```

	GVIF	Df	GVIF^(1/(2*Df))
LogDistance	1.653619	1	1.285931
LogLand	1.136432	1	1.066036
LogFloor	1.408585	1	1.186838
WaterView	1.105776	1	1.051559
FloodingZone	1.642183	1	1.281477
Suburb	2.907103	4	1.142701

- **VIF Results:** All VIF values were well below the threshold of 10, indicating no severe multicollinearity among the predictors. Specifically, most VIF values were close to 1, which is considered ideal for regression models. Additionally, the $GVIF^{1/(2 \cdot Df)}$ values were consistent with these results, confirming that multicollinearity is not an issue in this model.

The screenshot of VIF from R illustrates that the relationships between predictor

variables (such as LogDistance, LogLand, LogFloor, and others) are sufficiently independent, ensuring reliable and stable coefficient estimates.

3. Normality of Residuals

In Generalised Linear Models, the normality of residuals is not a critical assumption. Unlike ordinary linear regression, GLMs do not require normally distributed residuals, especially when using distributions like the Gamma distribution.

- **Explanation:** For this reason, tests such as the Shapiro-Wilk test or Q-Q plots were not applied. Normality of residuals is unnecessary and potentially misleading when the model is based on a Gamma distribution, as was the case in this analysis. The use of a log link in the GLM accommodates the skewness in the data, which is common in property value distributions.

4. Homoscedasticity (Constant Variance of Residuals)

The assumption of constant variance of residuals, or homoscedasticity, is not applicable to GLMs. In standard linear regression, this assumption is crucial, but in GLMs, the variance of the response variable is a function of the mean.

- **Gamma Distribution:** The Gamma distribution employed in this model accounts for the fact that the variance increases with the mean, which is typical of property value data. Therefore, traditional checks for homoscedasticity are unnecessary. As long as the variance is properly specified by the model's distribution, as it is in this case, the assumption is considered met.

5. Influential Points

Influential points can disproportionately affect model estimates, leading to skewed results. To identify such points, we employed Cook's Distance and the "Residuals vs Leverage" plot.

- **Cook's Distance Results:** The plot of Cook's Distance indicates that most points have values well below the threshold ($4/n-k-1$, where n is the number of observations and k is the number of predictors). Only a few points approached the threshold, but none exceeded it significantly. This confirms that there are no overly influential points that might unduly affect the overall model. Similarly, the **Residuals vs Leverage** plot demonstrates that while some points have higher leverage, their Cook's Distance values remain low, indicating they do not have an excessive influence on the model. The inclusion of these points does not distort the model's estimates, ensuring that the final results are reliable.

Summary of Assumption Testing

All critical assumptions for the GLM were tested and validated:

- **Independence of Residuals:** Confirmed through the residual plots showing no patterns or autocorrelation.
- **Multicollinearity:** Addressed using VIF, with results indicating no problematic multicollinearity among the predictors.
- **Normality of Residuals:** Not required for the GLM, so no tests were conducted for this assumption.
- **Homoscedasticity:** Not required for GLMs using the Gamma distribution.
- **Influential Points:** Detected using Cook's Distance, with no points having an excessive influence on the model.

These assumption checks confirm that the Generalised Linear Model is appropriately specified for this dataset and provides reliable and robust estimates for understanding the factors influencing property values in Christchurch.

Appendix O: R Code for Data Analysis and Modelling

Introduction

This appendix provides the R code used throughout the data analysis and modelling process in this project. The code documents key steps in data preparation, exploratory data analysis (EDA), model implementation, diagnostics, and assumption testing. While some processes, such as the GIS-based spatial analysis and data filtering, were performed using other software, R was used extensively for statistical modelling and assumption testing.

The Generalised Linear Model (GLM) and other models were implemented using R to estimate property values based on environmental and property-specific factors. The following sections detail the relevant R code used to prepare data, fit models, conduct diagnostic checks, and visualise the results.

Data Preparation

The initial dataset contained property transaction records and various environmental attributes, such as coastal proximity and flood zone status. Data preparation in R focused on removing influential data points, handling missing values, and ensuring the dataset was ready for statistical analysis.

Loading the Data

First, the dataset was loaded into R from a CSV file for further processing.

```
r
Code:
# Load necessary libraries
library(dplyr)

# Load the dataset
initial_data <- read.csv("property_data.csv")
```

Handling Missing Data

Missing data, particularly in key variables like sale price or floor area, was addressed by removing incomplete rows.

```
r
Code:
# Remove rows with missing values for key variables
cleaned_data <- initial_data %>%
```

```
filter(!is.na(SalePrice) & !is.na(FloorArea) & !is.na(LandArea))
```

Filtering Non-Coastal Properties

In this project, only coastal properties were considered for the analysis. The following R code removes properties that were identified as non-coastal during the GIS analysis (performed outside of R).

```
r
Code:
# Filter coastal properties based on previously assigned flag
coastal_properties <- cleaned_data %>%
  filter(CoastalFlag == 1)
```

Handling Outliers Using Cook's Distance

To address influential data points, Cook's Distance was calculated, and points exceeding the threshold were removed from the dataset. The threshold for influential points was set at $4/(n-k-1)$, where n is the number of data points and k is the number of predictors.

```
r
Code:
# Fit the initial GLM model
initial_model <- glm(LogValue2022 ~ LogDistance + LogLand + LogFloor + WaterView + FloodingZone +
  Suburb,
  data = coastal_properties, family = Gamma(link = "log"))

# Calculate Cook's Distance
cook_dist <- cooks.distance(initial_model)

# Define threshold for Cook's Distance
threshold <- 4 / (nrow(coastal_properties) - length(initial_model$coefficients) - 1)

# Remove influential data points
final_cleaned_data <- coastal_properties[cook_dist < threshold, ]
```

With the data cleaned and ready for analysis, the next step involves exploratory data analysis (EDA) to uncover underlying patterns in the dataset before fitting the models.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to better understand the structure and distribution of the data before proceeding with model development. This step involved summary statistics, visualisation of key variables, and checking for relationships between property values and important factors, such as distance to the coast and land area.

Summary Statistics

Basic summary statistics were generated to gain insight into the dataset's key attributes.

```
r
Code:
# Summary statistics for the cleaned data
summary(final_cleaned_data)

# Specific summary for variables of interest
summary(final_cleaned_data$LogDistance)
summary(final_cleaned_data$LogLand)
summary(final_cleaned_data$LogFloor)
summary(final_cleaned_data$SalePrice)
```

Distribution of Property Values

Visualising the distribution of property values provides insight into the skewness of the data, which informed the choice of the Gamma distribution for the GLM.

```
r
Code:
# Histogram of property values
hist(final_cleaned_data$SalePrice, breaks = 30, main = "Distribution of Property Values",
      xlab = "Property Value", col = "lightblue")
```

Relationship Between Property Value and Distance to Coast

Scatter plots were used to visualise the relationship between property values and coastal proximity.

```
r
Code:
# Scatter plot of property value vs. distance to coast
plot(final_cleaned_data$LogDistance, final_cleaned_data$SalePrice,
      main = "Property Value vs. Distance to Coast",
      xlab = "Log Distance to Coast", ylab = "Property Value", col = "darkblue", pch = 19)
```

This EDA step provided a preliminary understanding of the trends in the data, confirming that coastal proximity is a key driver of property values. After completing the EDA, the project proceeded to model implementation.

Modelling Approach

The primary modelling approach used in this project was the Generalised Linear Model (GLM) with a Gamma distribution and a log link function. This section details the R code for fitting the GLM and performing diagnostics to evaluate its fit.

Generalised Linear Model (GLM)

The Generalised Linear Model was chosen to capture the relationship between property values and a set of predictor variables, including distance to the coast, land area, floor area, water views, and whether the property is located in a flood zone.

```
r
Code:
# Fit the Generalised Linear Model (GLM) with a Gamma distribution and log link function
glm_model <- glm(LogValue2022 ~ LogDistance + LogLand + LogFloor + WaterView + FloodingZone +
  Suburb,
  data = final_cleaned_data, family = Gamma(link = "log"))
```

```
# Summary of the model
summary(glm_model)
```

Model Summary Interpretation

The `summary()` function in R provides details on the estimated coefficients, standard errors, and p-values. This information is crucial for interpreting how each predictor variable influences property values.

```
r
Code:
# Extracting and interpreting coefficients
coefficients(glm_model)
Interaction Terms and Polynomial Model
```

In addition to the base GLM, interaction terms and polynomial terms were tested to assess more complex relationships between variables, such as the interaction between coastal proximity and water views.

```
r
Code:
# Fitting a GLM with interaction terms
glm_interaction <- glm(LogValue2022 ~ LogDistance * WaterView + LogLand + LogFloor + FloodingZone
  + Suburb,
  data = final_cleaned_data, family = Gamma(link = "log"))
```

```
# Summary of the interaction model
summary(glm_interaction)
```

Generalised Additive Model (GAM)

The Generalised Additive Model (GAM) was also tested to capture potential non-linear relationships in the data. Smoothing functions were applied to continuous predictors, such as coastal distance and land area.

```
r
Code:
# Load necessary package
library(mgcv)

# Fit a GAM model
gam_model <- gam(LogValue2022 ~ s(LogDistance) + s(LogLand) + LogFloor + WaterView +
  FloodingZone + Suburb,
  data = final_cleaned_data, family = Gamma(link = "log"))

# Summary of the GAM model
summary(gam_model)
```

Model Evaluation

After fitting the models, several metrics were used to evaluate their performance and compare the Generalised Linear Model (GLM) with other candidate models. The key metrics included Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Squared Error (RMSE). This section outlines the R code used to compute and interpret these metrics.

AIC and BIC Calculation

The AIC and BIC were calculated to compare the goodness-of-fit of different models while penalising for model complexity. Lower values indicate better models.

```
r
Code:
# AIC and BIC for the Generalised Linear Model
AIC(glm_model)
BIC(glm_model)

# AIC and BIC for the GLM with interaction terms
AIC(glm_interaction)
BIC(glm_interaction)

# AIC and BIC for the Generalised Additive Model (GAM)
AIC(gam_model)
BIC(gam_model)
```

Root Mean Squared Error (RMSE)

RMSE was calculated to assess the model's accuracy in predicting property values. A lower RMSE indicates better predictive performance.

```
r
Code:
# Calculating RMSE for the Generalised Linear Model
```

```
glm_predictions <- predict(glm_model, type = "response")
rmse_glm <- sqrt(mean((final_cleaned_data$LogValue2022 - glm_predictions)^2))
rmse_glm

# RMSE for the GAM model
gam_predictions <- predict(gam_model, type = "response")
rmse_gam <- sqrt(mean((final_cleaned_data$LogValue2022 - gam_predictions)^2))
rmse_gam
```

Cross-Validation

Cross-validation was used to assess the generalisability of each model. Here, we performed k-fold cross-validation to check for overfitting and ensure the models perform well on unseen data.

```
r
Code:
# Load the necessary package for cross-validation
library(boot)

# 10-fold cross-validation for the GLM
cv_glm <- cv.glm(final_cleaned_data, glm_model, K = 10)
cv_glm$delta

# 10-fold cross-validation for the GAM
cv_gam <- cv.glm(final_cleaned_data, gam_model, K = 10)
cv_gam$delta
```

Model Diagnostics

Model diagnostics were critical to ensure that the assumptions of the Generalised Linear Model (GLM) were satisfied and that no data points disproportionately influenced the model results. This section details the diagnostic checks performed on the GLM.

Residual Plots

Residual plots were used to evaluate whether the residuals were randomly scattered, which would indicate a good model fit. Plots were generated for both Pearson and deviance residuals.

```
r
Code:
# Plot residuals vs. fitted values for the GLM
plot(glm_model$fitted.values, residuals(glm_model, type = "pearson"),
     main = "Pearson Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Pearson Residuals", pch = 19, col = "darkblue")
abline(h = 0, col = "red", lwd = 2)
```

```
# Plot deviance residuals
plot(glm_model$fitted.values, residuals(glm_model, type = "deviance"),
     main = "Deviance Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Deviance Residuals", pch = 19, col = "darkgreen")
abline(h = 0, col = "red", lwd = 2)
```

Cook's Distance and Leverage

Cook's Distance was used to detect influential data points that could unduly affect the model's estimates. Leverage plots were also generated to check for points with high influence on the model.

```
r
Code:
# Cook's Distance plot for influential data points
plot(cooks.distance(glm_model), type = "h",
     main = "Cook's Distance for Influential Data Detection",
     xlab = "Observation", ylab = "Cook's Distance", col = "darkorange")
abline(h = 4/(nrow(final_cleaned_data) - length(coef(glm_model)) - 1), col = "red")

# Leverage plot to detect high-leverage points
hat_values <- hatvalues(glm_model)
plot(hat_values, type = "h", main = "Leverage Values for Each Observation",
     xlab = "Observation", ylab = "Leverage", col = "purple")
abline(h = 2 * mean(hat_values), col = "red")
```

Predicted vs. Actual Values

This plot was used to assess how well the model's predicted values aligned with the actual property values, indicating the overall accuracy of the model.

```
r
Code:
# Plot predicted vs. actual values
plot(final_cleaned_data$LogValue2022, glm_predictions,
     main = "Predicted vs Actual Values",
     xlab = "Actual Property Values (Log Scale)", ylab = "Predicted Property Values", pch = 19, col =
"darkblue")
abline(a = 0, b = 1, col = "red", lwd = 2)
```

Assumption Testing

In this section, the key assumptions of the Generalised Linear Model (GLM) were tested to ensure the model's validity and reliability. This process involved checking for independence of residuals, multicollinearity among predictors, and influential points that could skew the results.

1. Independence of Residuals

The independence of residuals was tested to ensure that the residuals (errors) were not correlated. This assumption is important in regression models as correlated residuals may indicate that key variables are missing from the model or that the model is misspecified.

r

Code:

```
# Residuals over time plot to check for independence
plot(1:nrow(final_cleaned_data), residuals(glm_model, type = "pearson"),
     main = "Residuals Over Time",
     xlab = "Index", ylab = "Pearson Residuals", pch = 19, col = "darkblue")
abline(h = 0, col = "red")
```

2. Multicollinearity Check Using Variance Inflation Factor (VIF)

Multicollinearity refers to the presence of high correlations between predictor variables, which can distort the estimated coefficients. The Variance Inflation Factor (VIF) was used to check for multicollinearity. Values of VIF greater than 10 indicate high multicollinearity, which could affect the reliability of the model.

r

Code:

```
# Load car package for calculating VIF
library(car)

# Calculate VIF for each predictor in the GLM model
vif(glm_model)
```

3. Influential Points Detection

Cook's Distance and Leverage were examined to detect influential data points that could unduly affect the results of the model. The threshold for Cook's Distance was set at $4/(n-k-1)$, where n is the number of observations and k is the number of predictors.

r

Code:

```
# Plot Cook's Distance for detecting influential points
plot(cooks.distance(glm_model), type = "h",
     main = "Cook's Distance for Influential Data Detection",
     xlab = "Observation", ylab = "Cook's Distance", col = "darkorange")
abline(h = 4/(nrow(final_cleaned_data) - length(coef(glm_model)) - 1), col = "red")
```

Model Selection

This section discusses the process of selecting the best model to explain property values based on a comparison of different models, including Generalised Linear Models (GLM), Generalised Additive Models (GAM), Polynomial Regression, and Interaction Models.

1. Comparison of AIC and BIC

To guide the selection of the best model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to compare models. A lower AIC or BIC value indicates a better-fitting model that balances complexity and goodness of fit.

```
r
Code:
# AIC and BIC for GLM model
AIC(glm_model)
BIC(glm_model)

# AIC and BIC for GLM with interaction terms
AIC(glm_interaction)
BIC(glm_interaction)

# AIC and BIC for the GAM model
AIC(gam_model)
BIC(gam_model)
```

2. Root Mean Squared Error (RMSE) Comparison

The Root Mean Squared Error (RMSE) was used to compare how well each model predicted property values. RMSE values were calculated for each model to evaluate predictive performance, with lower RMSE indicating better accuracy.

```
r
Code:
# RMSE for the GLM model
glm_predictions <- predict(glm_model, type = "response")
rmse_glm <- sqrt(mean((final_cleaned_data$LogValue2022 - glm_predictions)^2))
rmse_glm

# RMSE for the GAM model
gam_predictions <- predict(gam_model, type = "response")
rmse_gam <- sqrt(mean((final_cleaned_data$LogValue2022 - gam_predictions)^2))
rmse_gam
```

3. Cross-Validation

To ensure that the selected model generalised well to unseen data, k-fold cross-validation was performed on the GLM, GAM, and interaction models. Cross-validation provides a more robust estimate of the model's predictive ability by testing it on different subsets of the data.

```
r
Code:
# Perform 10-fold cross-validation on the GLM model
cv_glm <- cv.glm(final_cleaned_data, glm_model, K = 10)
cv_glm$delta
```

```
# Perform 10-fold cross-validation on the GAM model
cv_gam <- cv.glm(final_cleaned_data, gam_model, K = 10)
cv_gam$delta
```

4. Final Model Selection

After comparing models based on AIC, BIC, RMSE, and cross-validation results, the Generalised Linear Model (GLM) was chosen as the final model for this analysis. Although the Generalised Additive Model (GAM) showed lower AIC and BIC values, the GLM provided a good balance between interpretability and model performance.

```
r
Code:
# Final selected model is GLM
summary(glm_model)
```

Results Visualization

This section outlines how key results from the Generalised Linear Model (GLM) were visualized to provide a clearer understanding of the factors influencing property values in Christchurch. The visualizations included residual plots, predicted vs. actual value plots, and coefficient plots, among others.

1. Predicted vs Actual Values Plot

This plot shows the relationship between the predicted property values from the GLM and the actual observed values in the dataset. A strong linear trend between predicted and actual values indicates that the model fits the data well.

```
r
Code:
# Plot of Predicted vs Actual Values
plot(final_cleaned_data$LogValue2022, glm_predictions,
     main = "Predicted vs Actual Property Values",
     xlab = "Actual Log Property Value (2022)",
     ylab = "Predicted Log Property Value (2022)",
     pch = 19, col = "blue")
abline(0, 1, col = "red") # 45-degree line for reference
```

2. Residuals vs Fitted Values Plot

This plot helps assess whether the model captures the relationship between the predictors and the response variable adequately. A random scatter of points around zero indicates that the model's residuals are behaving as expected.

```
r
Code:
```

```
# Plot of Residuals vs Fitted Values
plot(glm_model$fitted.values, residuals(glm_model, type = "deviance"),
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Deviance Residuals",
     pch = 19, col = "darkgreen")
abline(h = 0, col = "red")
```

3. Coefficient Plot

The coefficient plot visually represents the estimated effects of the predictors (e.g., LogDistance, LogLand, LogFloor) on property values. Positive coefficients increase property values, while negative coefficients decrease them.

```
r
Code:
# Coefficient Plot
library(coefplot)

coefplot(glm_model, main = "Coefficient Plot for GLM Model")
```

4. Boxplot of Property Values by Suburb

This boxplot compares property values across different suburbs in Christchurch, allowing us to see how location influences pricing.

```
r
Code:
# Boxplot of Property Values by Suburb
boxplot(LogValue2022 ~ Suburb, data = final_cleaned_data,
     main = "Property Values by Suburb",
     xlab = "Suburb", ylab = "Log Property Value (2022)",
     col = "lightblue")
```

5. Histogram of Distance to Coast

This histogram shows the distribution of property distances from the coast. Most properties are concentrated within a certain range from the coast, providing insights into the concentration of coastal properties in Christchurch.

```
r
Code:
# Histogram of Distance to Coast
hist(final_cleaned_data$LogDistance, breaks = 30,
     main = "Histogram of Distance to Coast",
     xlab = "Log Distance to Coast",
     col = "lightgreen")
```

Conclusion

In this project, we developed a comprehensive analysis of property values in Christchurch using various statistical models, with a focus on environmental factors such as coastal proximity, flood zones, and suburban location. The Generalised Linear Model (GLM) with a Gamma distribution and log link function was selected as the most appropriate model based on performance metrics, interpretability, and robustness.

Summary of Key Findings

1. **Coastal Proximity:** Properties located closer to the coast tend to have higher values, reflecting the premium placed on beachfront living in Christchurch.
2. **Land and Floor Area:** Floor area had a significant impact on property value, while land area had less influence.
3. **Water Views:** Properties with water views were more highly valued, though the effect was smaller than expected.
4. **Flood Zones:** Flooding zones had a minimal effect on property values in the dataset, though future risks may become more prominent.
5. **Suburb Location:** Suburbs such as Waimairi Beach and Southshore commanded higher property values compared to others like New Brighton.

Model Performance

The GLM successfully captured the complex relationships between property values and the predictor variables, including environmental risks and geographical factors. Through careful model evaluation and diagnostic testing, we confirmed that the assumptions of the GLM were met, ensuring the model's validity.

Future Recommendations

For future research, it would be valuable to:

- **Integrate more dynamic environmental risk factors**, such as future projections of sea-level rise and storm frequency, to capture long-term risks associated with coastal living.
- **Expand the dataset** to include more recent property transactions and additional geographical regions in New Zealand, for broader applicability.
- **Incorporate more advanced machine learning models** (e.g., random forests, gradient boosting) to compare their predictive performance with traditional regression models.

The comprehensive R script developed as part of this project offers a complete pipeline for replicating the analysis and can be used to extend the study further as more data becomes available.