My title*

My subtitle if needed

First author Another author

December 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Child care subsidies play a critical role in making high-quality early childhood education and care accessible to families and communities. In Toronto, licensed child care centres serve as essential providers, offering regulated and professional environments that support child development. Subsidies for licensed child care centres help close the affordability gap by offsetting the high costs of quality care, making it accessible to more families (Cleveland and Krashinsky 2009). These subsidies not only ease the financial burden on families but also ensure that children have access to nurturing environments that foster cognitive, social, and emotional growth during their formative years (Vines 2020). Understanding the factors that determine which licensed centres receive subsidies is crucial for policymakers and stakeholders to promote equitable resource allocation and maximize the benefits of early childhood programs. Licensed child care centres in urban settings like Toronto play a pivotal role in delivering high-quality, structured child care. These centres adhere to stringent regulations, ensuring compliance with standards for safety, staffing, and curriculum. Research highlights that children attending licensed centres, particularly those supported by subsidies, experience better developmental outcomes (Adams et al., 2013). Subsidized centres provide professional environments with trained educators, comprehensive programming, and age-appropriate resources, offering children a strong foundation for lifelong learning and success (Herbst, 2014). Subsidies are a cornerstone of this ecosystem, enabling licensed centres to cover operational costs, retain qualified staff, and maintain compliance with regulatory standards, all of which enhance the quality of care provided (Ryan et al. 2011). Despite their importance, disparities in the allocation of subsidies remain a significant concern. Research indicates that centres in certain neighborhoods or serving specific populations may receive fewer subsidies, even when demand is

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

high (Johnson, Ryan, and Brooks-Gunn 2012). Additionally, characteristics such as enrollment capacity, accreditation status, and program focus (e.g., infant care versus pre-kindergarten) often influence a centre's eligibility and prioritization for funding (**vines2020accessin?**). These discrepancies highlight the need for a data-driven approach to understanding and improving the distribution of subsidies among licensed child care centres, ensuring that resources are allocated equitably to maximize their impact. This study utilizes the Toronto Open Data: Licensed Child Care Centres dataset to explore the factors that affect subsidy allocation to licensed centres. By analyzing variables such as ward, operating auspice (Commercial, Non Profit or Public), CWELCC, type of building, and total space, this research aims to uncover the question - "What factors influence the allocation of subsidies to licensed child care centres in Toronto?". The findings will contribute to informing policies that promote equity and efficiency in subsidy allocation, ultimately supporting the goals of accessible and high-quality child care in Toronto. Additionally, these insights can serve as a valuable tool for child care centres to self-assess and enhance their eligibility for subsidies.

Telegraphing paragraph: The remainder of this paper is structured as follows. Section **??**….

2 Data

2.1 Data Overview

The dataset used in this study was sourced from the (**Childrens_Services?**) made publicly available by the City of Toronto. This original raw dataset provides detailed information about licensed child care centres, including 20 variables capturing aspects of their location, operating auspice (e.g., non-profit, public, or commercial), space usage, building type, participation in government programs such as the Canada-Wide Early Learning and Child Care (CWELCC) system, and other operational details. These data offer valuable insights into the factors influencing subsidy allocation, a key policy tool for improving access to early childhood education and care. By translating real-world phenomena into structured data entries, this dataset enables a comprehensive exploration of equity and efficiency in child care funding. Detailed data collection analysis is in appendix

```
data <- read.csv("../data/02-analysis_data/cleaned_data.csv")
```

2.2 Method

The dataset used for this study is the **`Licensed Child Care Centres dataset`**, sourced from the (Gelfand 2022) It provides detailed information about licensed child care centres in Toronto, capturing aspects such as their governance, capacity, infrastructure, and subsidy status. The original dataset consisted of 1071 records for all licensed child care centres within the city. For

Table 1: Sample of Cleaned Data

Overview of Cleaned Data
Displaying the first 10 rows

ward	AUSPICE	bldg_type	cwelcc_flag	TOTSPACE	subsidy
3	Non Profit Agency	Public Elementary School	1	164	1
8	Non Profit Agency	Public Elementary School	1	83	1
25	Non Profit Agency	Catholic Elementary School	1	102	1
10	Non Profit Agency	Other	1	65	1
20	Non Profit Agency	High Rise Apartment	1	26	1
24	Non Profit Agency	Community College/University	1	62	1
6	Non Profit Agency	Public High School	1	49	1
24	Commercial Agency	High Rise Apartment	1	46	1
19	Non Profit Agency	Public Elementary School	1	51	1
8	Non Profit Agency	Public Elementary School	1	153	1

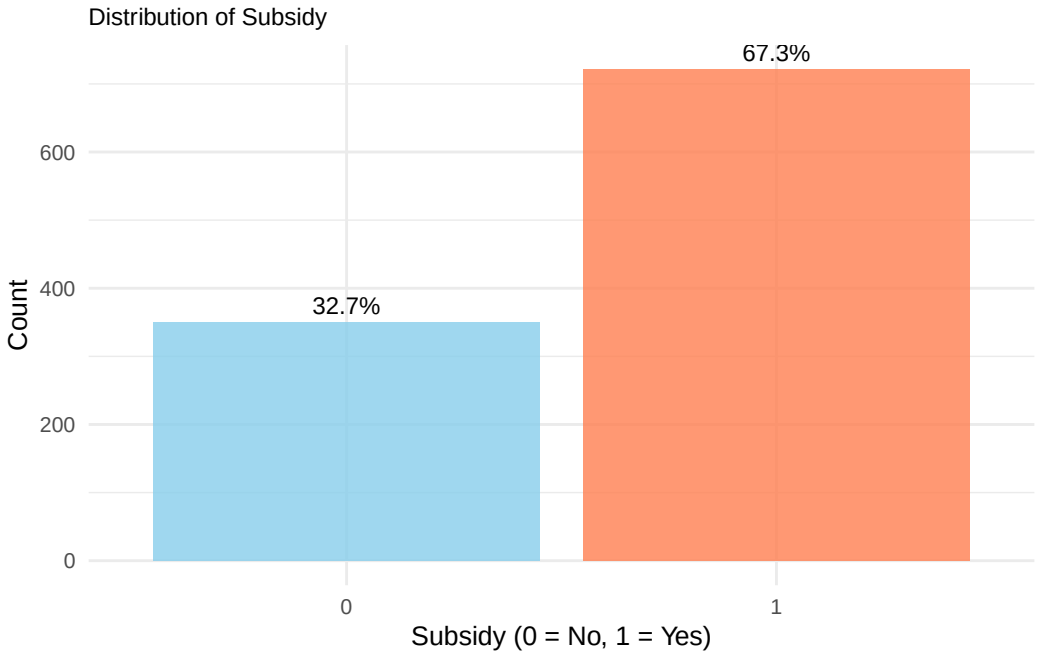

Figure 1: Subsidy Allocation Distribution

this analysis, the data underwent preprocessing to focus on variables relevant to the study, such as subsidy status, building type, CWELCC participation, total space, and operating auspice. These variables were retained to examine how different factors influence subsidy allocation. The dependent variable **subsidy**, originally recorded as "Yes"/"No," was encoded as 1 for subsidized and 0 for non-subsidized centres. Similarly, the **cwelcc_flag** variable was converted into a binary format (1 =Y, 0 =N). Categorical variables such as **bldg_type** and **AUSPICE** were consolidated to simplify analysis and address sparse categories. Additionally, missing and irrelevant records were removed to ensure the dataset was both accurate and meaningful for the research objectives.

The data for this study was systematically downloaded, cleaned, analyzed, and visualized using **R** (R Core Team 2023), a statistical programming language. The following are major packages used for this study: **opendatatoronto**(Gelfand 2022): Used to access and retrieve the Licensed Child Care Centres dataset directly from the City of Toronto's open data portal. **readr** (Wickham, Hester, and Bryan 2024): Simplified the import and parsing of raw data into R. **tidyverse** (Wickham et al. 2019): Streamlined data manipulation, cleaning, and visualization processes. **dplyr** (Wickham et al. 2023): Provided tools for filtering, transforming, and summarizing the dataset effectively. **ggplot2** (Wickham 2016): Created powerful and flexible visualizations tailored to the analysis needs. **car** (Fox and Weisberg 2019): Used for diagnostic tools, including Variance Inflation Factor (VIF) tests, to assess multicollinearity. **caret** (Kuhn and Max 2008): Enabled the development, validation, and evaluation of machine learning models, including training-test splits and performance metrics. **glmnet** (Friedman et al., 2010): Applied for fitting regularized regression models and feature selection. **stargazer** (Hlavac 2022): Generated formatted regression tables for outputs. **knitr** (Xie 2021): Dynamically integrated code, results, and plots into the final document for seamless reporting. ## Measurement This analysis focuses on the following variables, with a specific emphasis on subsidy as the dependent variable: `subsidy`: The binary dependent variable indicating whether a licensed child care centre receives a government subsidy. 1: The centre is subsidized. 0: The centre is not subsidized. `ward`: `A numeric variable representing the ward number for child care centres.`AUSPICE: `The operating auspice of the child care centre, describing its governance and operational model. Possible values include: Non-Profit: Centres operated by non-profit organizations, often reinvesting surplus revenues into quality improvements. Commercial: For-profit centres operated by private organizations. Public: Centres run by public agencies or school boards.`bldg_type: `The type of building where the child care centre operates, reflecting its infrastructure. Examples include: bldg_typeCommercial Building bldg_typeCommunity College/University bldg_typeCommunity Health Centre bldg_typeCommunity Rec/Centre - Board Run bldg_typeCommunity/Rec Centre - City bldg_typeCommunity/Recreation Centre bldg_typeHigh Rise Apartment bldg_typeHospital/Health Centre bldg_typeHouse bldg_typeIndustrial Building bldg_typeLow Rise Apartment bldg_typeOffice Building bldg_typeOther bldg_typePlace of Worship bldg_typePrivate Elementary School bldg_typePublic (school closed) bldg_typePublic Elementary Special`

`bldg_typePublic High School bldg_typePublic Middle School bldg_typePurpose Built bldg_typeSynagoguecwelcc_flag: A binary variable indicating participation in the Canada-Wide Early Learning and Child Care (CWELCC) program: 1: The centre participates in CWELCC, enabling reduced child care fees. 0: The centre does not participate in CWELCC.`TOTSPACE‘: A numerical variable representing the total licensed capacity (spaces available) for all age groups at a child care centre. Detailed information about these variables' information and data structure is presented in Table **??**.

The variables were carefully selected based on literature-supported relevance to subsidy allocation and their representation of real-world phenomena

Figure **??** illustrates the distribution of subsidy status (1 = Subsidized, 0 = Not Subsidized) across various building types housing licensed child care centres. Notably, Public Schools, Purpose-Built Facilities, and Community Recreation Centres exhibit higher proportions of subsidized centres. These facilities are often designed to meet regulatory requirements for child care, including adequate space, safety standards, and accessibility, aligning closely with subsidy allocation policies (Cleveland and Krashinsky 2009). Conversely, building types such as Industrial Buildings, Private Elementary Schools, and Office Buildings show lower proportions of subsidized centres, likely due to infrastructure challenges or misalignment with subsidy eligibility criteria, such as limited accessibility or higher operational costs (**yan2011impact?**). These patterns suggest that building type significantly influences subsidy distribution. Given the numerous categories of building types, a de tailed analysis is warranted to fully understand these trends.

Figure **??** illustrates the proportional distribution of subsidy status (1 = Subsidized, 0 = Not Subsidized) across the different operating auspices of licensed child care centres: Commercial Agency, Non-Profit Agency, and Other. Non-Profit Agencies and "Other" entities are predominantly subsidized, while Commercial Agencies display a more balanced distribution. This aligns with research indicating that non-profits rely heavily on subsidies to deliver public goods and services, as they often operate in markets with limited profitability (Hansmann 1979). Conversely, commercial entities are less reliant on subsidies due to their revenue-driven models.The dominance of subsidies in the "Other" category suggests this group may include hybrid or public-private organizations aligned with specific government initiatives (Anheier 2014). Such reliance reflects the strategic use of subsidies to support services underserved by the private market (Weisbrod 2000).

Figure **??** a heatmap illustrates the correlations between three variables: TOTSPACE (likely representing total space or capacity), subsidy (indicating subsidy status or amount), and cwelcc_flag (potentially denoting eligibility for a specific program or funding). The positive correlation between TOTSPACE and subsidy (0.25) suggests that larger facilities are modestly more likely to receive subsidies, reflecting their capacity to serve larger populations or provide greater public benefits. This aligns with research indicating that larger organizations often have the resources and visibility to secure subsidies (Hansmann, 1980; Salamon, 2002). Additionally, the stronger correlation between subsidy and cwelcc_flag (0.48) highlights that

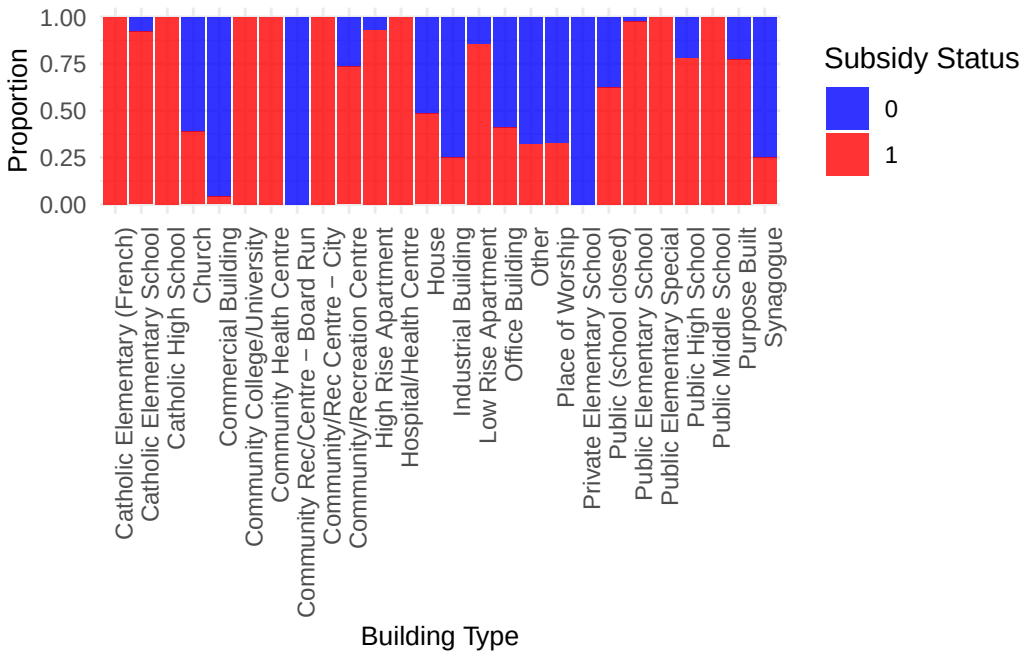

Figure 2: Proportional Distribution of Subsidy Status Across Building Types

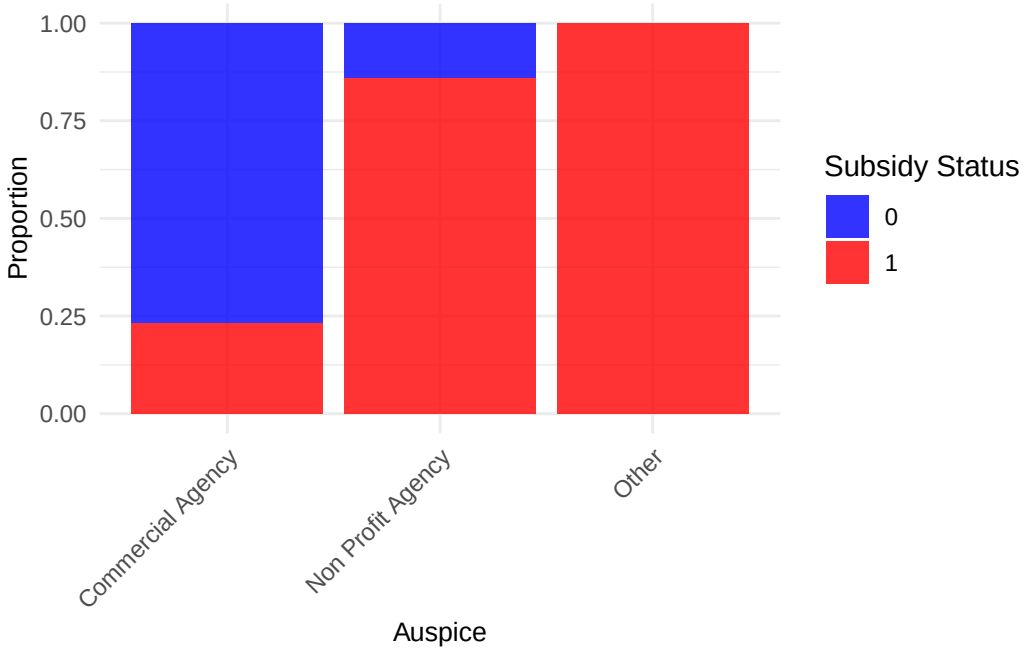

Figure 3: Proportional Distribution of Subsidy Status by Auspice

subsidy allocation may target entities meeting specific programmatic or policy criteria, consistent with the literature emphasizing strategic targeting of subsidies to maximize societal impact (Weisbrod, 1998). The weaker correlation between TOTSPACE and cwelcc_flag (0.17) suggests that program eligibility is less dependent on size and more on qualitative factors like service type or demographic focus, which is supported by Anheier's (2005) analysis of nonprofit funding models. Together, these correlations emphasize the nuanced role of subsidies in balancing operational scale and policy alignment, underscoring the importance of strategic allocation in public funding (Salamon, 1995).

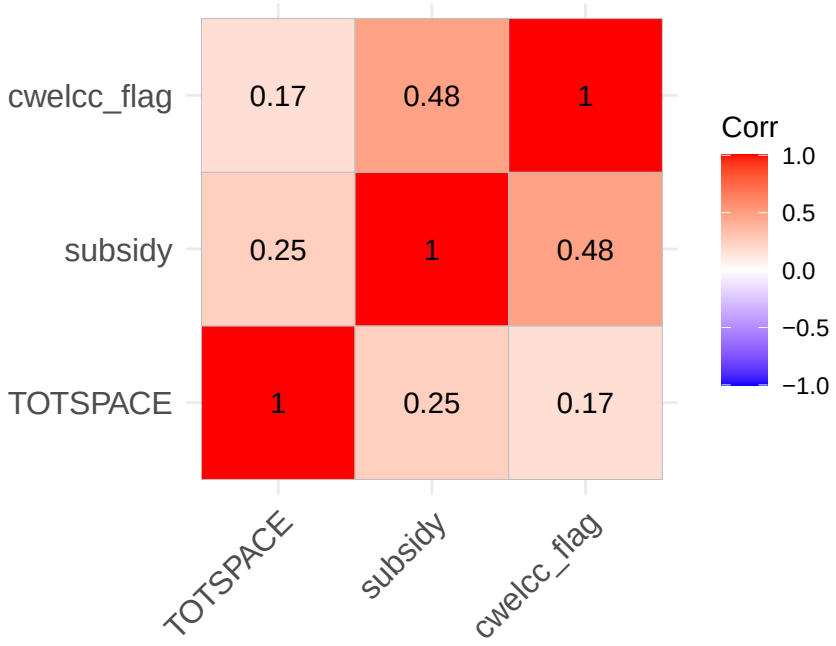

Figure 4: Correlation Heatmap of Key Variables: TOTSPACE, Subsidy, and CWELCC Flag

3 Model

3.1 Model Specification

To investigate the factors influencing subsidy allocation to licensed child care centres, a reduced logistic regression model was specified. The dependent variable,`subsidy`, is a binary indicator representing whether a child care centre receives government subsidy (1 = subsidized, 0 = not subsidized). The model includes three key predictors: `Operating Auspice (AUSPICE)`: A categorical variable indicating the governance model of the child care centre (e.g., Non-Profit Agency, Other). Non-profit agencies are hypothesized to be positively associated with subsidy allocation, consistent with previous research emphasizing their prioritization in funding schemes (Cleveland & Krashinsky, 2005). `Participation in the CWELCC`

`Program (cwelcc_flag)`: A binary variable capturing whether the centre participates in the Canada-Wide Early Learning and Child Care program (1 = participates, 0 = does not participate). Centres participating in this initiative are expected to have higher odds of receiving subsidies due to their alignment with government objectives of affordability and accessibility (Friendly & Ballantyne, 2022). `Total Space (TOTSPACE)`: A continuous variable representing the number of licensed spaces available in the centre. Larger centres are hypothesized to have higher odds of subsidy allocation, as they can accommodate more families and align with policy goals of maximizing access (Forry et al., 2010). The logistic regression model can be expressed mathematically as follows:

3.2 Model Justification

The logistic regression model was chosen for this analysis because it is well-suited to the binary nature of the dependent variable, subsidy status (1 = Subsidized, 0 = Not Subsidized). A logistic regression model using the binomial family is specifically designed to model dichotomous outcomes by estimating the log-odds of the event occurring as a linear function of predictor variables (Hosmer Jr, Lemeshow, and Sturdivant 2013). The binomial family is appropriate here because it assumes that the dependent variable follows a Bernoulli distribution, where each observation represents a binary outcome (subsidized or not subsidized). This ensures that the predicted probabilities remain between 0 and 1, aligning with the real-world constraints of the problem. Additionally, logistic regression provides interpretable coefficients, which indicate the direction and magnitude of the relationship between each predictor and the log-odds of subsidy allocation. This makes it particularly useful for guiding policy decisions, as coefficients can be directly converted into odds ratios for actionable insights (Peng, Lee, and Ingersoll 2002). The ward variable was excluded because it lacked statistical significance and added redundancy, as its effects were captured by other predictors like CWELCC participation and total space. Additionally, ward had no strong theoretical justification as a direct determinant of subsidy allocation. The building type variable was removed due to high dimensionality, sparse representation in many categories, and statistical insignificance. Its effects are likely mediated by other variables, such as total space and auspice. Excluding it improved parsimony and interpretability. The final model retained operating auspice, CWELCC participation, and total space, as these predictors are strongly supported by theory and data. This approach balances simplicity and accuracy, ensuring the model remains relevant for informing equitable subsidy allocation policies.

3.3 Model Assumptions

To ensure the validity of the logistic regression model, several key assumptions were assessed, including independence of observations, the appropriateness of a binary outcome, the linearity of predictors with the logit and absence of multicollinearity. The analysis integrates results from visual diagnostics, multicollinearity tests, and statistical measures. 1. Independence of

Observations The logistic regression model assumes that the observations are independent of each other. In this analysis, each data point corresponds to an individual child care center, ensuring independence. There is no clustering or repeated measures within the dataset, which validates this assumption. 2. Binary Outcome The logistic regression model assumes a binary dependent variable. In this case, the outcome variable, `subsidy`, is binary, indicating whether a child care center receives a subsidy (1 = Subsidized, 0 = Not Subsidized). This aligns with the model's requirement, ensuring the suitability of the binomial family for fitting the data. 3. Linearity of Predictors with the Logit The Figure **??** The component + residual plots evaluate the linearity of the continuous variables and the relationship between categorical predictors and the logit transformation. graph

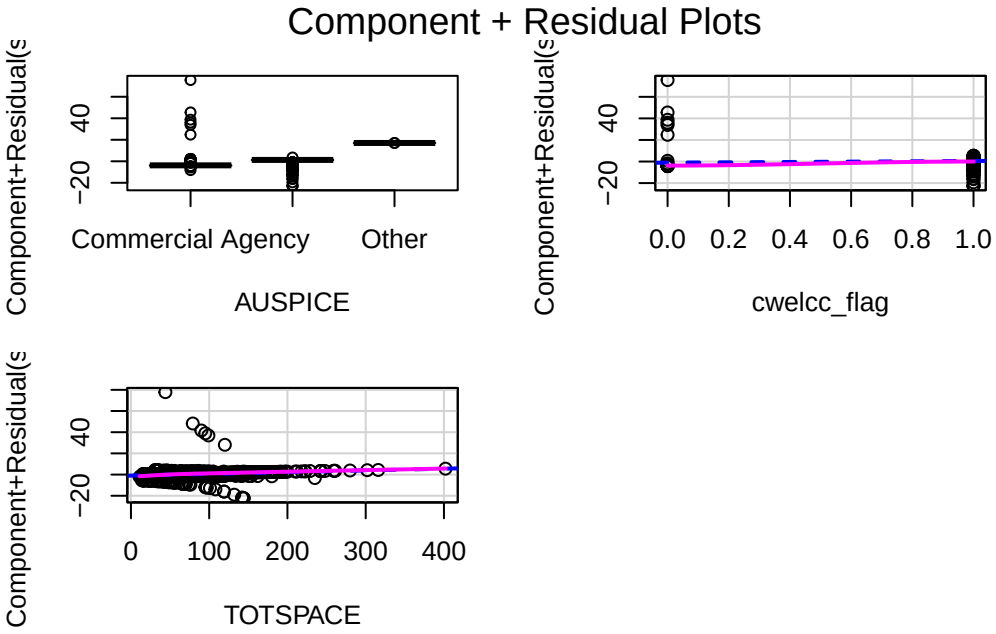

Figure 5: CR Plot for Linearity Chekck

`TOTSPACE`: The relationship between the total space (TOTSPACE) and the logit appears approximately linear, as indicated by the flat, consistent pattern of residuals around the horizontal axis. This supports the assumption of linearity. `CWELCC Flag`: The plot for CWELCC participation shows a horizontal trend, suggesting no significant deviation from linearity. The data support the inclusion of this variable as a binary predictor. `AUSPICE`: The residual patterns for categorical levels of AUSPICE (e.g., "Commercial Agency" and "Other") indicate distinct clusters with consistent variability. This suggests the categorical nature of this variable does not violate linearity assumptions. Overall, the visual inspection suggests that the linearity assumption for the predictors with the logit is met.

4. Multicollinearity Assessment Variance Inflation Factor (VIF): To assess multicollinearity among the predictors, the VIF values for the reduced model variables were calculated.