A close up of a flower

Description automatically generatedASSS

**DATA ANALYSIS AND VISUALISATION OF k2171099 ASSESSMENT DATASET**

**Data Analytics and Visualisation Coursework 1**

By Lum Apeh Temukum, Kingston University, 16 December 16, 2024

# **INTRODUCTION**

In this report, I will analyse a dataset containing information about staff members across various departments of a local government. The dataset includes **366 staff members** and provides details such as the department they work in, their length of employment, the number of sick days taken over the last two years, and their **response target**—the percentage of times they have responded to inquiries within the 48-hour target.

The objective of this analysis is to explore patterns within each variable and identify key factors that influence response target performance. Through visualizations and data-driven insights, the goal is to help the organization make informed decisions to enhance work efficiency and overall staff performance.

To achieve these goals, I will carry out the following tasks.

1. **A single graph for each variable, showing the distribution of the data. The graph format should be appropriate to the data type, and it should be clearly labelled.**
   * 1. I **used** a **bar chart** to show the number of staff members in each department. This was the most suitable way to present the information because each department is a separate category, and it would not make sense to combine or average them.

A grid of white lines on a black background

Description automatically generated

**Observations/Insights:**

* The Environment department employs the largest number of staff (127), followed by the Education department (114), while the Housing (61), Welfare (45), and Heritage (19) departments have smaller staff numbers, indicating varying levels of operational focus and resource allocation among departments.
  + 1. A **bar chart** was used to display the distribution of staff members by length of employment because it effectively compares values across distinct categories. It provides a clear visual representation of trends, such as whether most staff are newly employed, mid-term, or long-term employees.

A graph of green bars

Description automatically generated

**Observations/Insights**

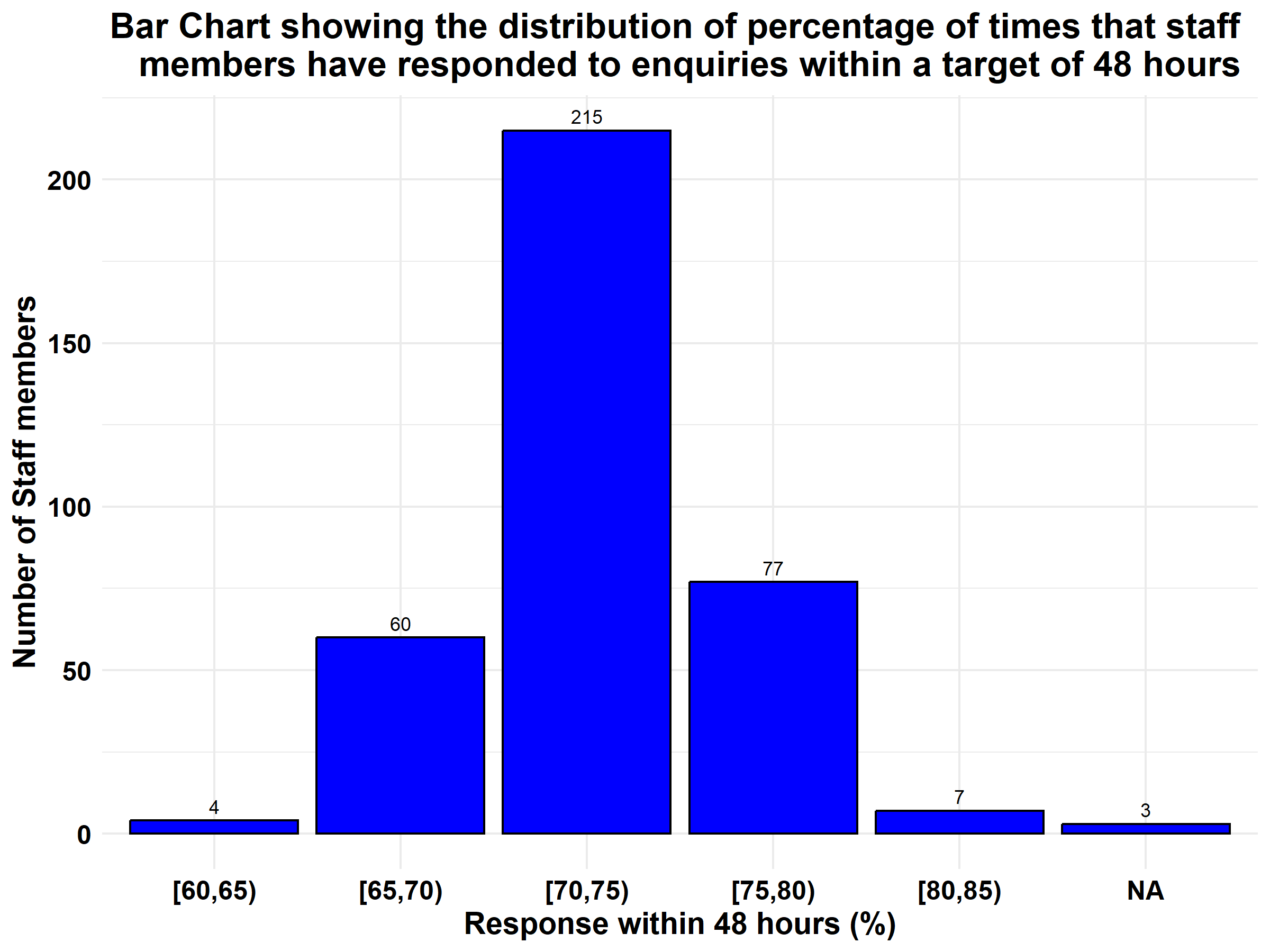
* The **3-5 years** group has the most employees (93), followed by **1 year** (81) and **less than 1 year** (66), suggesting recent hiring.
* The **2-year range** (64 employees) shows slight consistency, but numbers drop significantly for those with **6-10 years** (44) and **more than 10 years** (18), indicating long-term retention issues.
  + 1. I used a histogram to display the distribution of sick days because it provides a clear visualization of the continuous variable, enabling one to identify the patterns, trends, and outliers that can guide decisions related to employee health and attendance management.

A graph with red bars

Description automatically generated

**Observations/Insights:**

* The distribution of sick days indicates that most employees fall within the **10-30 sick days range**, with the highest numbers in the **10-15 days (38 staff)** and **20-25 days (57 staff)** categories. This highlights moderate absenteeism as typical over two years. Beyond **30 days**, the frequency of sick leave decreases significantly (e.g., **30-35 days: 17 staff**), with extended absences being less common.
* However, a small number of employees recorded **40+ sick days**, and a few outliers, such as those with **85-90 days (3 staff)** or **over 115 days (1 staff)**, could indicate long-term health issues, workplace-related stress, or inconsistencies in reporting. These groups may require further attention and support.
  + 1. I used a bar chart because Target response is a continuous quantitative variable. A **bar chart** is ideal for showing the distribution of staff response rates as it compares the number of staff across percentage categories, highlights trends in performance, and provides a clear, intuitive visualization for both technical and non-technical audiences.



**Obervation/Insights:**

* Most staff members (215) fall within the **70-75% response rate**, with fewer performing below **70%** or exceeding **75%**, and only 7 reaching rates of **80-85%**. This indicates a gap in higher response performance and the need for targeted support for those in the **60-65%** range.
* Additionally, missing data for 3 staff members should be reviewed to ensure data completeness.
  + 1. I used a bar chart to show staff distribution by full-time duration because it clearly compares the number of staff in each time category. It highlights patterns, is easy to read, and works well for non-technical audiences.

A graph with a black background

Description automatically generated

**Observation/Insights:**

* Many staff members are in the 3-5 years category (93), reflecting a period of stability and retention. The 1-year group (81) is the second largest, which may indicate recent hiring efforts. A considerable number of staff are relatively new, with 66 employed for less than 1 year and 64 employed for 2 years.
* However, the numbers decrease for longer employment durations, with 44 in the 6-10 years range and only 18 in more than 10 years, suggesting challenges in retaining employees over the long term.

1. **A table (or tables) of appropriate descriptive statistics for each of the four variables, taking into account the distribution of the data and the data type. A brief description of the information contained in the table.**
   * 1. **Descriptive Statistics for Department (Nominal Data)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Category | Count | Percentage |
| 1. | Environment | 127 | 34.7 |
| 2. | Education | 114 | 31.1 |
| 3. | Housing | 61 | 16.7 |
| 4. | Welfare | 45 | 12.3 |
| 5. | Heritage | 19 | 5.2 |

**Description**:

* The **Environment department** employs **127 staff (34.7%)**, and **Education** employs **114 staff (31.1%)**, together making up over **65%** of the workforce. **Housing** has **61 employees (16.7%)**, while the smallest departments, **Welfare (45 employees, 12.3%)** and **Heritage (19 employees, 5.2%)**, together account for just **17.5%**.
  + 1. **Descriptive Statistics for Employment Duration**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Category | Count | Percentage |
| 1. | 3-5 years | 93 | 25.4 |
| 2. | 1 year | 81 | 22.1 |
| 3. | Less than 1 year | 66 | 18.0 |
| 4. | 2 years | 64 | 17.5 |
| 5. | 6-10 years | 44 | 12.0 |
| 6. | More than 10 years | 18 | 5.0 |

**Description:**

* Many employees fall into the **3-5 years** category, with **93 staff (25.4%)**, followed by **1 year** of employment, representing **81 staff (22.1%)**, and **Less than 1 year**, with **66 staff (18.0%)**. Mid-range tenures, such as **2 years (64 staff, 17.5%)** and **6-10 years (44 staff, 12.0%)**, account for a smaller proportion.
* The smallest group consists of employees with **More than 10 years** of service, comprising only **18 staff (5.0%)**. This distribution highlights a workforce predominantly composed of newer and mid-tenure employees.
  + 1. **Descriptive statistics of Quantitative Variables (Sick days and Response Target)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Metric | Sick Days | Response Target |
| 1. | Mean | 29.1 | 72.7 |
| 2. | Median | 21.5 | 72.6 |
| 3. | Standard Deviation | 23.7 | 3.2 |
| 4. | Minimum | 0 | 62.8 |
| 5. | Maximum | 130 | 81.4 |
| 6. | Range | 130 | 18.6 |
| 7. | Skewness | 1.2 | 0.1 |
| 8. | Kurtosis | 1.5 | 0.2 |

**Description:**

**Sick Days**

* The average number of sick days taken by employees is **29.1**, with a median of **21.5 days**, indicating a skewed distribution towards higher values. Additionally, the mean is higher than the median, further emphasizing the skewness. The variability is high (**SD = 23.7**) with a wide range from **0 to 130 days**, showing significant differences in sick leave usage among employees.
* The positive skew (**1.2**) and kurtosis (**1.5**) highlight the impact of a few extreme cases and a heavy-tailed distribution.

**Response Target**

* On average, employees meet the 48-hour response target for **72.7%** of inquiries, with a median of **72.6%**, suggesting a nearly symmetrical distribution.
* The low variability (**SD = 3.2**) and narrow range (**62.8% to 81.4%**) indicate consistent performance across the workforce. Minimal skewness (**0.1**) and kurtosis (**0.2**) show a balanced and evenly distributed data set.

1. **A table of appropriate statistics for response\_target, comparing the groups in department and in time, including a brief summary of what you can interpret from the graph.**

**Response Target Summary by Department and Time**

| department | time | count | mean | std | min | 25% | median | 75% | max |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Education | 1 year | 30 | 70.54483 | 2.5900339 | 65.2 | 68.900 | 70.60 | 72.200 | 75.9 |
| Education | 2 years | 17 | 74.65294 | 3.3391836 | 68.6 | 72.200 | 74.20 | 77.800 | 79.9 |
| Education | 3-5 years | 28 | 73.61071 | 3.1678173 | 67.4 | 71.450 | 73.20 | 75.850 | 80.6 |
| Education | 6-10 years | 12 | 74.97500 | 1.4410381 | 72.9 | 74.000 | 74.50 | 76.000 | 77.3 |
| Education | Less than 1 year | 23 | 70.61739 | 3.0224391 | 63.9 | 69.150 | 71.20 | 72.500 | 75.4 |
| Education | More than 10 years | 4 | 76.92500 | 1.2606215 | 75.6 | 76.125 | 76.80 | 77.600 | 78.5 |
| Environment | 1 year | 31 | 71.07000 | 1.7374476 | 67.2 | 70.100 | 70.75 | 72.500 | 74.5 |
| Environment | 2 years | 26 | 73.16923 | 2.3699400 | 70.3 | 71.025 | 72.40 | 75.000 | 79.3 |
| Environment | 3-5 years | 31 | 72.96667 | 2.4929325 | 67.7 | 71.175 | 73.10 | 74.100 | 78.3 |
| Environment | 6-10 years | 12 | 75.01667 | 2.9091809 | 71.0 | 73.175 | 74.35 | 76.250 | 81.1 |
| Environment | Less than 1 year | 25 | 70.94000 | 2.2137826 | 66.6 | 68.900 | 71.20 | 72.800 | 74.0 |
| Environment | More than 10 years | 2 | 76.30000 | 0.0000000 | 76.3 | 76.300 | 76.30 | 76.300 | 76.3 |
| Heritage | 1 year | 6 | 69.48333 | 2.9280824 | 65.9 | 67.025 | 69.85 | 71.625 | 73.0 |
| Heritage | 2 years | 2 | 72.60000 | 0.7071068 | 72.1 | 72.350 | 72.60 | 72.850 | 73.1 |
| Heritage | 3-5 years | 6 | 72.70000 | 3.2887688 | 69.0 | 69.925 | 72.70 | 75.100 | 76.9 |
| Heritage | 6-10 years | 2 | 72.05000 | 3.8890873 | 69.3 | 70.675 | 72.05 | 73.425 | 74.8 |
| Heritage | Less than 1 year | 2 | 73.95000 | 0.6363961 | 73.5 | 73.725 | 73.95 | 74.175 | 74.4 |
| Heritage | More than 10 years | 1 | 79.90000 | NA | 79.9 | 79.900 | 79.90 | 79.900 | 79.9 |
| Housing | 1 year | 10 | 69.66000 | 3.5731094 | 62.8 | 67.575 | 69.65 | 72.125 | 74.3 |
| Housing | 2 years | 13 | 73.00000 | 2.1256372 | 67.9 | 72.900 | 73.20 | 74.100 | 75.5 |
| Housing | 3-5 years | 17 | 73.61176 | 2.8857153 | 68.5 | 71.300 | 73.40 | 74.900 | 79.7 |
| Housing | 6-10 years | 8 | 74.95000 | 1.7361698 | 72.1 | 73.600 | 75.45 | 76.200 | 77.2 |
| Housing | Less than 1 year | 7 | 69.75714 | 2.6986769 | 65.1 | 68.750 | 69.40 | 71.550 | 73.2 |
| Housing | More than 10 years | 6 | 79.30000 | 2.0307634 | 75.9 | 78.400 | 80.00 | 80.475 | 81.4 |
| Welfare | 1 year | 4 | 69.45000 | 3.2501282 | 64.6 | 69.100 | 70.95 | 71.300 | 71.3 |
| Welfare | 2 years | 6 | 73.45000 | 2.3586013 | 69.7 | 72.800 | 73.35 | 74.500 | 76.8 |
| Welfare | 3-5 years | 11 | 74.68182 | 2.9905913 | 71.9 | 72.400 | 73.10 | 76.500 | 81.4 |
| Welfare | 6-10 years | 10 | 73.29000 | 3.0856118 | 69.1 | 71.000 | 72.75 | 75.575 | 78.6 |
| Welfare | Less than 1 year | 9 | 70.91111 | 2.5467845 | 67.2 | 69.400 | 71.10 | 71.300 | 76.5 |
| Welfare | More than 10 years | 5 | 76.94000 | 1.0784248 | 75.3 | 76.500 | 77.20 | 77.700 | 78.0 |

**Summary of Insights:**

* Employees in the Education department with Less than 1 year of experience have a moderate response rate (70.62%) with some variability. As employees gain experience, response rates improve, peaking at 2-6 years, with More than 10 years employees showing the highest mean response rate (76.93%) and the lowest variability.
* In the Environment department, employees with Less than 1 year of experience have a response rate of 70.94%, with some variability. As employees gain experience, response rates improve, reaching 75.02% for those with 6-10 years, and 76.30% for those with More than 10 years, showing the highest mean and no variability in the latter group.
* In the **Heritage** department, employees with **Less than 1 year** of experience have a response rate of **73.95%** with low variability. As experience increases, response rates improve, reaching **79.90%** for those with **More than 10 years**, the highest rate, though there is no variability data available for this group.
* In the **Housing** department, employees with **Less than 1 year** of experience have a response rate of **69.76%** with moderate variability. As employees gain experience, response rates improve, reaching **79.30%** for those with **More than 10 years**, showing the highest mean with some variability.
* In the **Welfare** department, employees with **3-5 years** of experience have a high mean response rate (**74.68%**) and good consistency. Response rates improve with experience, with **More than 10 years** employees showing the highest rate (**76.94%**) and the lowest variability. Employees with **1 year** of experience show the lowest mean response rate (**69.45%**) and more variability.
* **Housing** and **Environment** departments show the highest response rates for employees with **More than 10 years** of experience (**79.30%** and **76.30%**, respectively).
* **New employees** (Less than 1 year) across all departments tend to show the lowest response rates and more variability, with **Housing** showing the lowest among them (**69.76%**).
* In both the **Education** and **Environment** departments, there is a trend where employees with **2 years of experience** perform at their best. After that, performance tends to stabilize or slightly decline, but overall, **experience is positively correlated with better response rates** and more consistent performance.
* The data suggests that as employees gain more experience in their roles, their ability to meet the 48-hour response target improves, both in terms of the **average performance** and **consistency**. Newer employees tend to have lower performance on average and show more variability, likely due to a learning period or adjustment phase.
* Interestingly, **3-5 years** of experience doesn't show a large jump in performance over **2 years**, but still outperforms employees with less than 1 year or 1 year of experience. This could be an indication that after the initial learning phase, employees settle into a steady rhythm, with only incremental improvements after the first few years.

Overall, this comparison highlights the trend that longer experience generally leads to higher and more consistent response rates across all departments.

1. **An appropriate graph format, showing response\_target statistics for each group in time, and presenting the groups of time in ascending order, including a brief summary of what the graph shows, and why your choice of graph is appropriate.**

* Here, I choose to use a box plot because it efficiently displays a variable’s distribution across several categories, summarizing median, spread, and outliers at a glance. By placing these summaries side-by-side, it enables quick, direct comparisons between groups.

A black background with blue squares

Description automatically generated

A screenshot of a graph

Description automatically generated

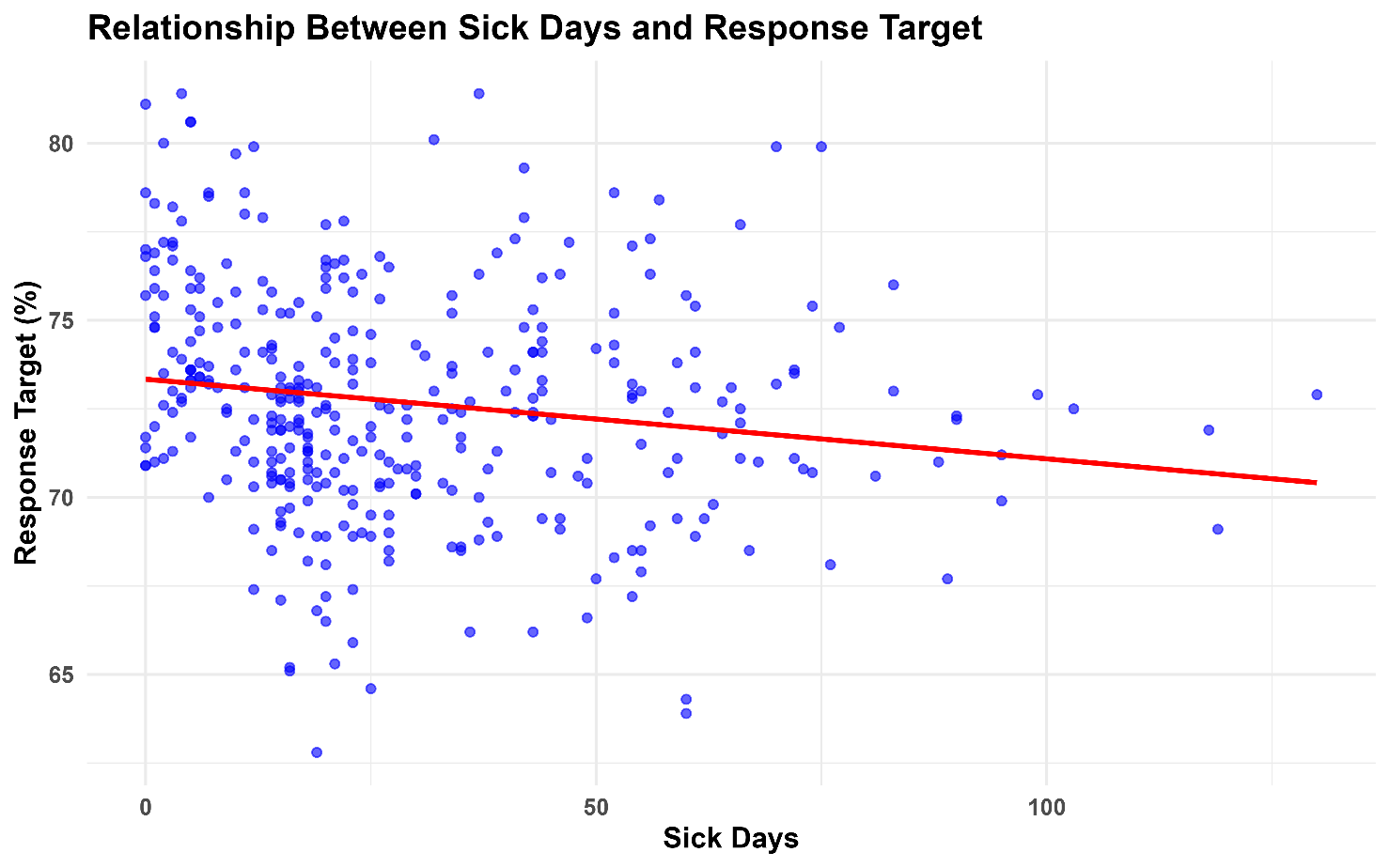
**Summary:**

This boxplot visually compares the distribution of the response target percentage across different lengths of employment. Each category along the x-axis represents a different time range, and the boxes display how the values are spread for that group. The horizontal line within each box indicates the median, the box boundaries show the middle 50% of the data, and the “whiskers” represent the broader spread. Outliers are marked as red dots.

* The boxplot indicates that as employment duration increases, both the median and overall range of response targets generally move upward.
* Categories with shorter employment durations show values clustering more closely around the median, while groups with longer tenure have more varied distributions.
* A few outliers appear, highlighted by red dots, indicating that some individual observations differ significantly from their peers.
* Overall, the graph suggests that employees who have been with the company longer tend to have higher response target percentages, although their results may also be more dispersed.

1. **An appropriate graph format, showing the relationship between sick\_days and response\_target, including a brief summary of what the graph shows, and why your choice of graph is appropriate.**

* I utilized a **scatter plot** to show the relationship between sick days and response target because it is the most appropriate graph for examining the relationship between two numerical variables. This type of graph allows us to visually inspect for any correlation or pattern, such as positive or negative trends, clusters, or outliers.



**Summary:**

* The scatter plot illustrates the relationship between the number of sick days employees have taken in the last two years and the percentage of enquiries they responded to within the 48-hour target.
* Each blue dot represents an individual data point, plotting the person’s sick days against their response target percentage. The red regression line trends slightly downward, suggesting a modest negative relationship: as sick days increase, average response target percentages tend to decrease slightly. However, the data points are widely spread, indicating a weak correlation. In other words, while there’s a hint that more sick days might be associated with a lower response target, the relationship doesn’t appear to be strong or consistent.
* Overall, there does not appear to be a clear or strong trend between the two variables, suggesting that the number of sick days taken by employees may not significantly impact their response rate.