Exploratory analysis report

Introduction

The objective of the exploratory analysis is to gather insights into the data provided by Spark and flare. This also ensures that the data has been cleaned effectively to ensure a reliable result as we progress throughout the project. The insights specifically focus on the various factors that influence the profit margin of each booking. This will help optimize financial strategies in the company.

There were also 4 key questions identified that can be answered by the exploratory analysis, further analysis can also provide further evidence.

(a) What is the most popular performer booking?

(b) What is the most frequently booked booking type (standard, community, corporate)?

(c) What is the average profit margin for each booking type

(d) What month has the highest profit margin?

The data contains 57 variables and 116 observations. The key variables identified are Income, Total expenses, and actual profit and the various performer types. Income is the total revenue generated, total expenses are the total costs incurred and actual profit is the net profit of each booking.

Outliers

There were several outliers removed depending on the test that was performed. These outliers were determined using z-scores, this represents the number of standard deviations a data point is from the mean. Any observations with a z-score greater than 3 or less than 3 were flagged as outliers. This method was used because we believed the mean and standard deviations are good descriptors of the data. It was also discovered that it positively affected the results of the tests when removing based on the z-score vs the Interquartile range.

Descriptive statistics

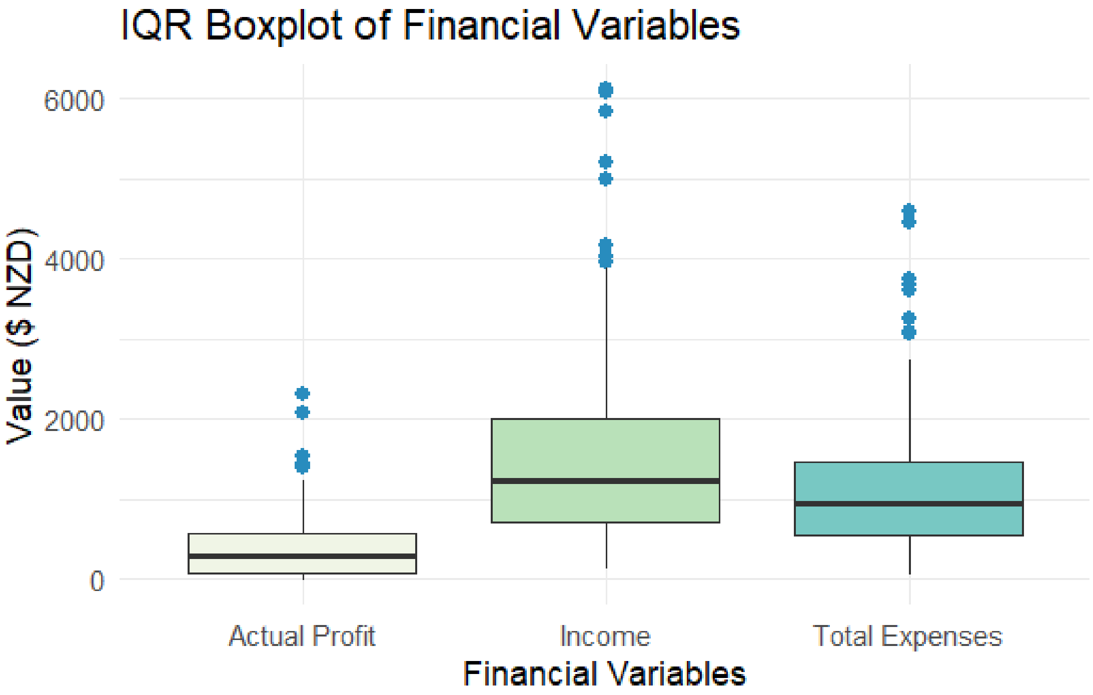
A series of descriptive statistics were calculated on the key variables and recalculated after removing outliers. The distribution of the income after removing the outliers shows there is a high variability with a Standard deviation ($1,320.67) significantly high compared to the mean ($1.620.98). The interquartile range suggests that the middle 50% of the data is widespread, meaning there is a lot of variation between the 25th and 75th percentiles. This high variability shows there are differences in event sizes and contracts, this suggests that some bookings are substantial, and others are smaller, leading to a wide range of incomes.

The Standard deviation ($969.37) for total expenses (mean=$1,214.66) is also high but lower than income meaning there is a moderate spread in expense values. The interquartile range is also large for total expenses indicating that the middle 50% are varied. This could be due to varying costs for different events. This needs to be managed and monitored carefully within the business as large fluctuations in expenses can affect the financial stability of the business.

Once the outliers are removed the mean decreased significantly for Actual profit. This indicates that the outliers inflated the profit values ($627.12 down to $406.30). The Standard deviation is higher than the mean, this indicates that the profit values vary greatly from the average, this shows there are high and low profit values. There is a large Interquartile range, this indicates that the profit outcomes vary depending on a variety of factors, this could be event size, booking type or other factors. The variability in the profit highlights the need for better cost control and pricing strategies. Further analysis should be done to understand what causes high profit margins or losses to help optimize the performance of the business.

In summary the descriptive statistics shows that further analysis needs to be done to understand which factors affect profitability to assist in stabilizing the profitability. Further analysis could be done to determine which bookings yield high or lower income to assist in target marketing and that the business needs to monitor expenses of the bookings to identify major cost drivers.

Figure 1 Box plot of key financial variables



The yearly income was also calculated to determine which year had the highest income. 2023 has the highest income ($95,347.25), followed by 2022 ($63,117.16). This shows that there is an upward trend over the years. 2024 is the current year and has done $36,061.43 as of July 2024.

Figure 2 Box plot of Yearly income

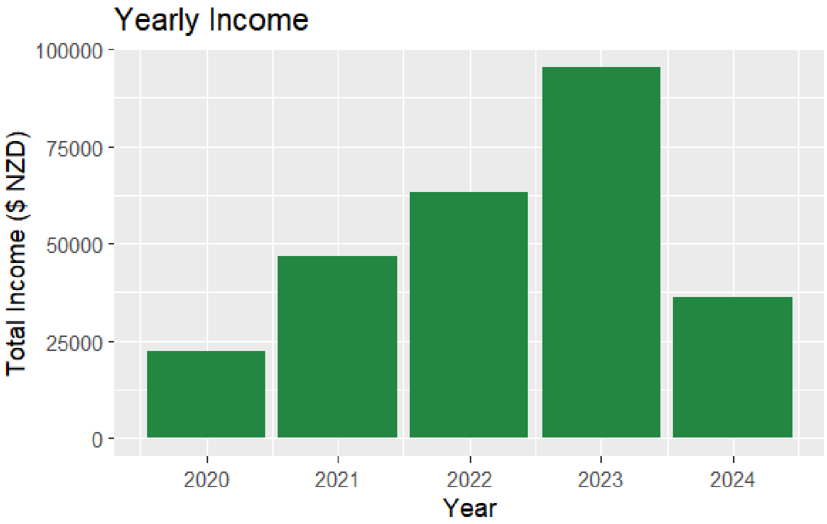
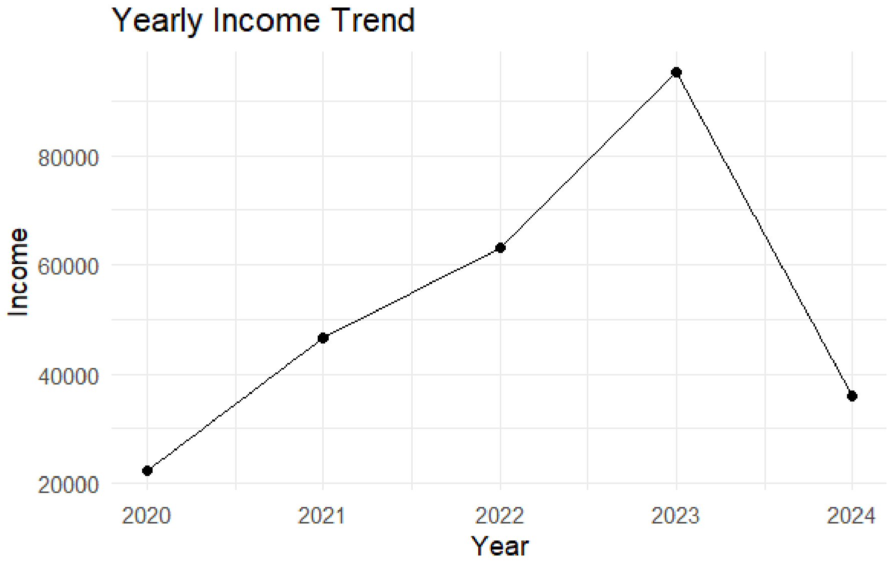


Figure 3 Trend of yearly income



We then analysed the data by breaking the year into quarters. Q1 is month 1-3, Q2 is month 4-6, Q3 is month 7-9 and Q4 is month 10-12. This showed that Q2 is 2023 was the highest income followed by Q4 2022. Q3 in 2024 was not complete year but Q1 in 2024 was lower than previous years ad Q2 is higher than the previous year.

Figure 4 Box plot of quarterly income

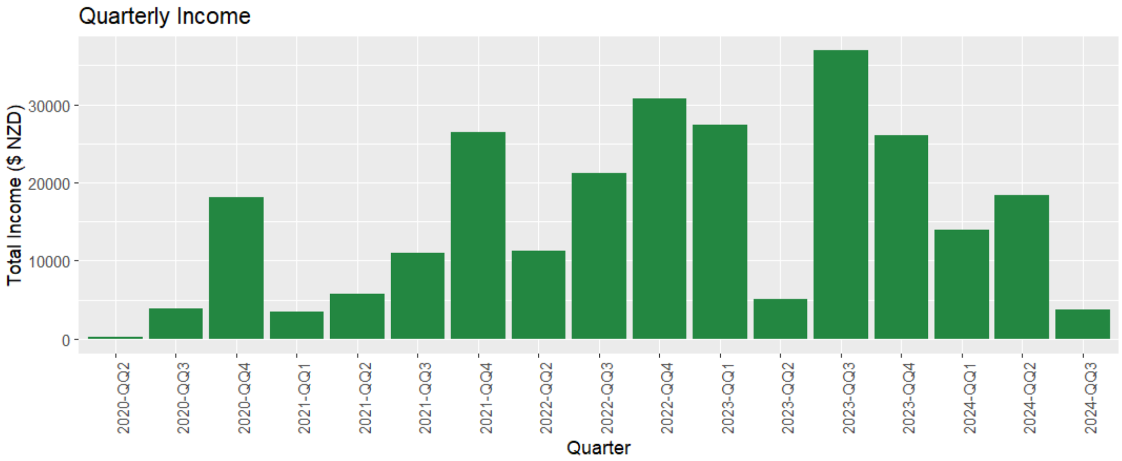
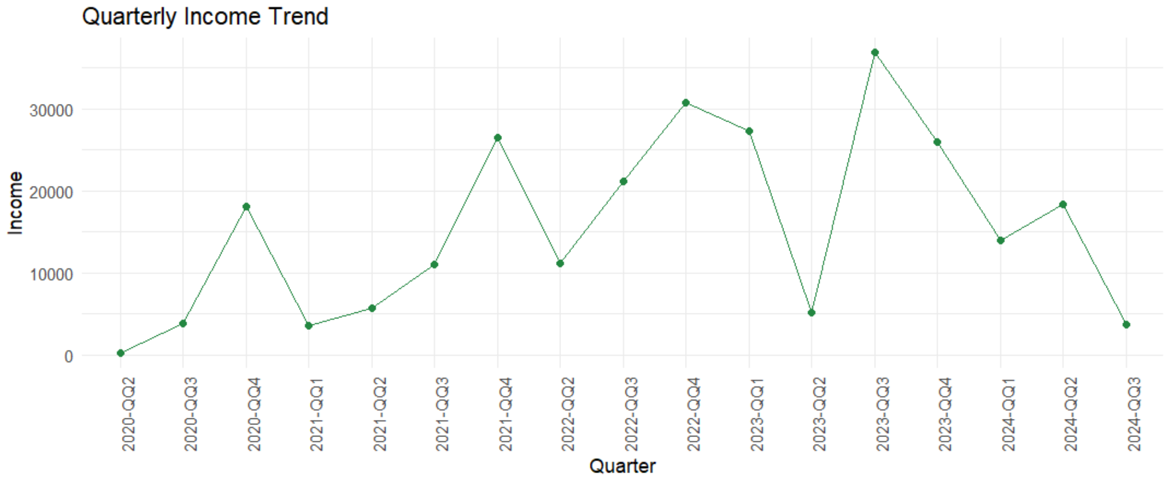


Figure 5 Quarterly income trend



What is the most popular performer booking?

Based on a summary of the bookings Ushers is the most popular booking with 1,215.45 hours. This is followed by stilts with 55 hours booked and roving performers booked for 48.52 hours. There is a wide range of performers across the bookings with a significant drop off in the number of bookings beyond the top few performers. This suggests a strong preference for certain performers over others. The lowest booked performers are fire show, magician, and drummers.

Figure 6 Total bookings for all performers

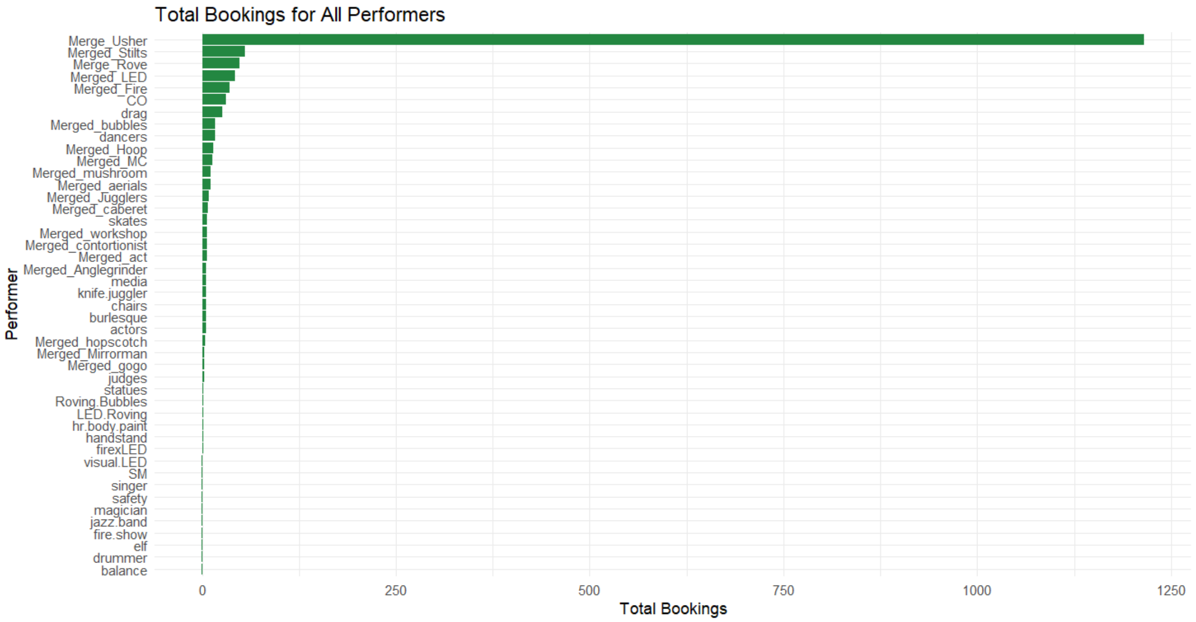


Figure 7 Stacked bar plot Total bookings for all performers

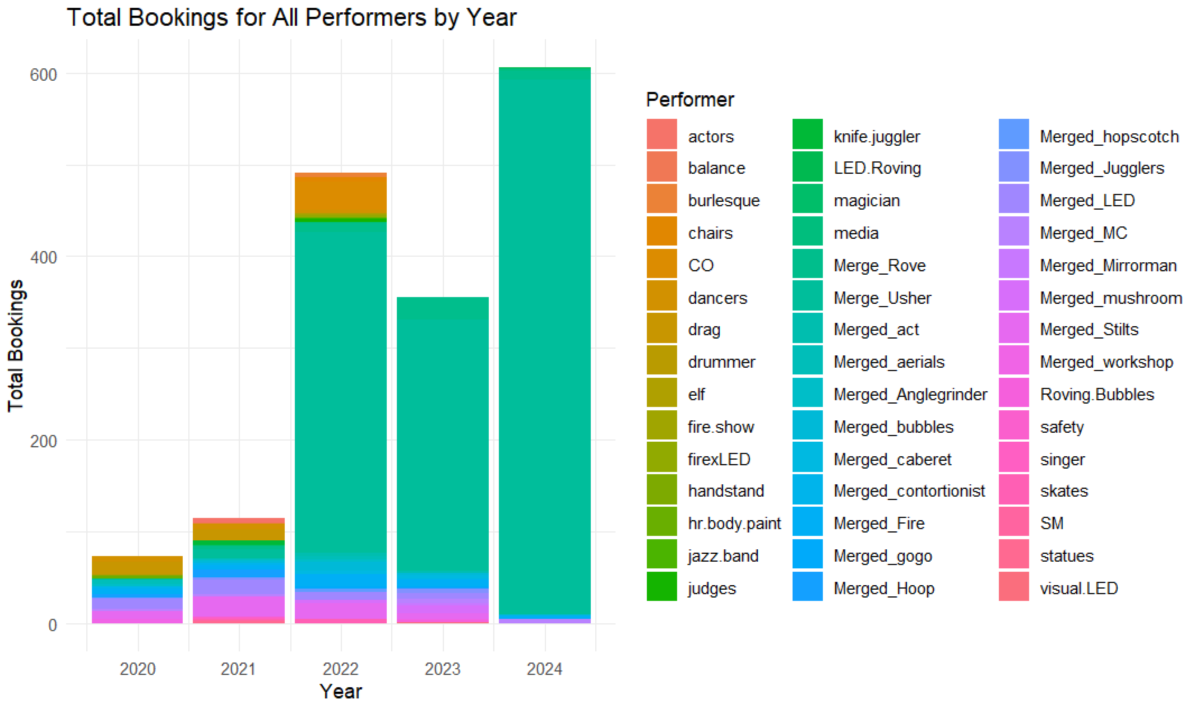


Figure 8 Total bookings for performers excluding Ushers

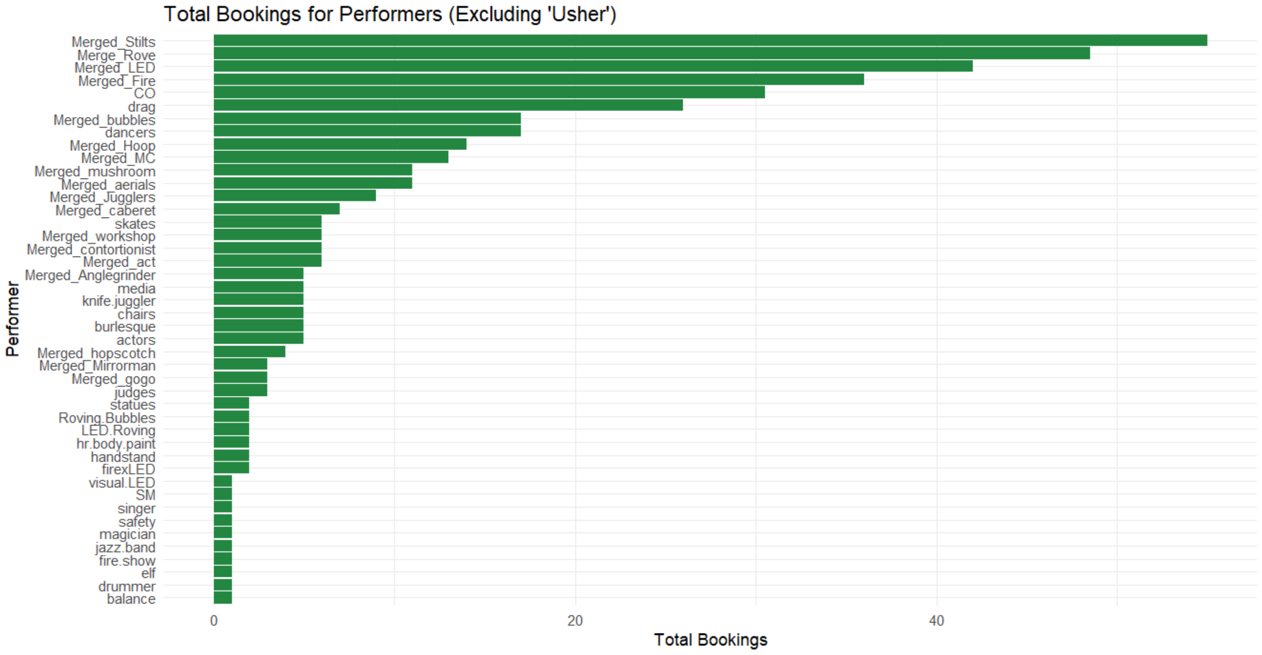
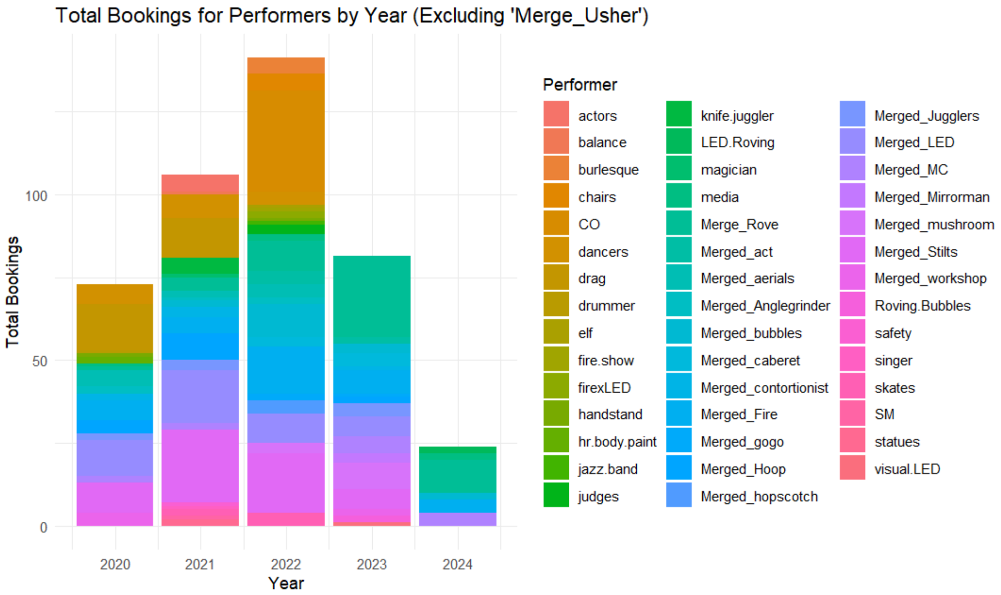


Figure 9 Stacked bar plot of Total bookings for performers excluding ushers



What is the average profit margin for each booking type?

The booking type with the highest average profit margin is the community booking type ($11147) followed by standard ($359) and cooperate ($340). This trend is also the same for income and total expenses. This may be due to a lack of data, there is more information and bookings for community (40) and standard (55) and corporate (21).

Figure 10 Average profit by booking type

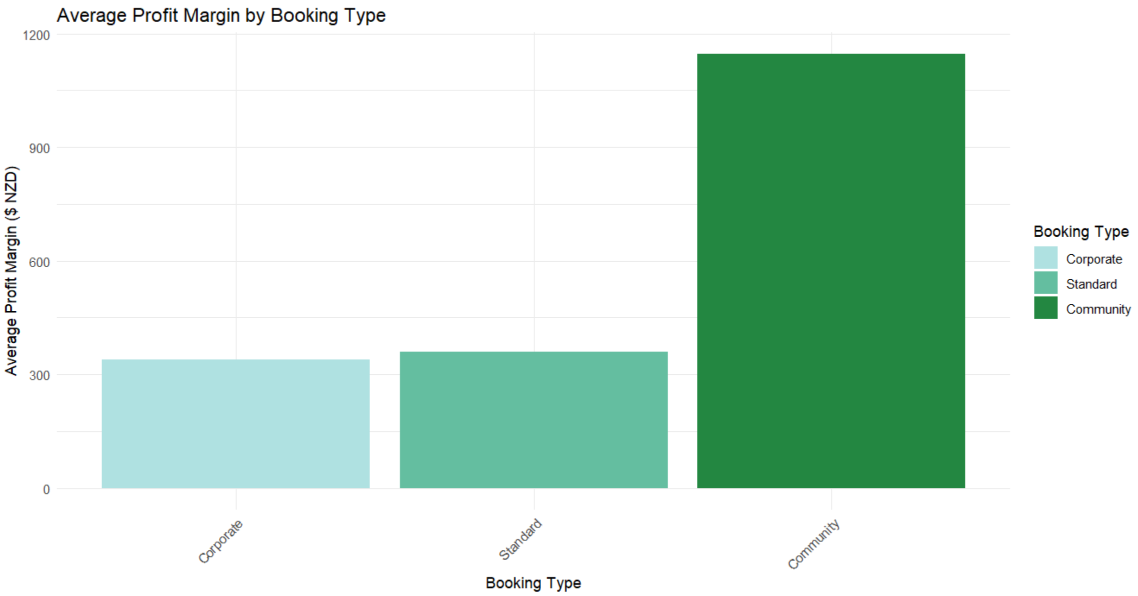


Figure 11 Average profit by booking type and year

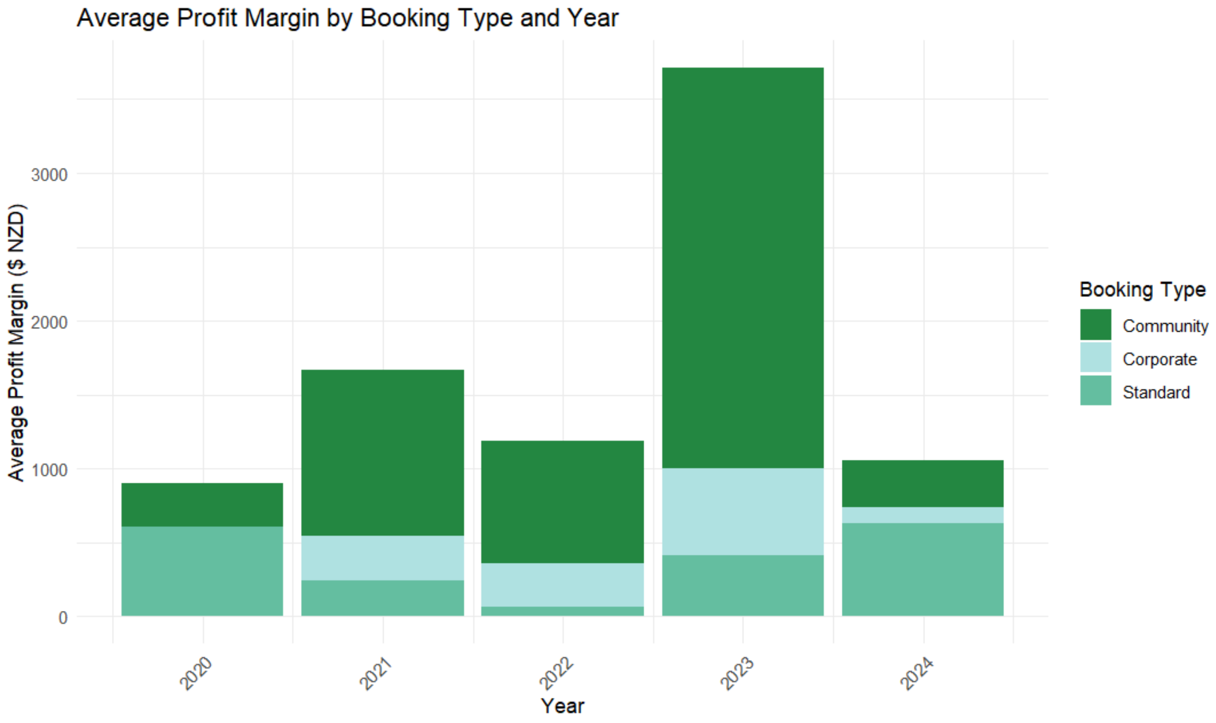


Figure 12 Average income by booking type

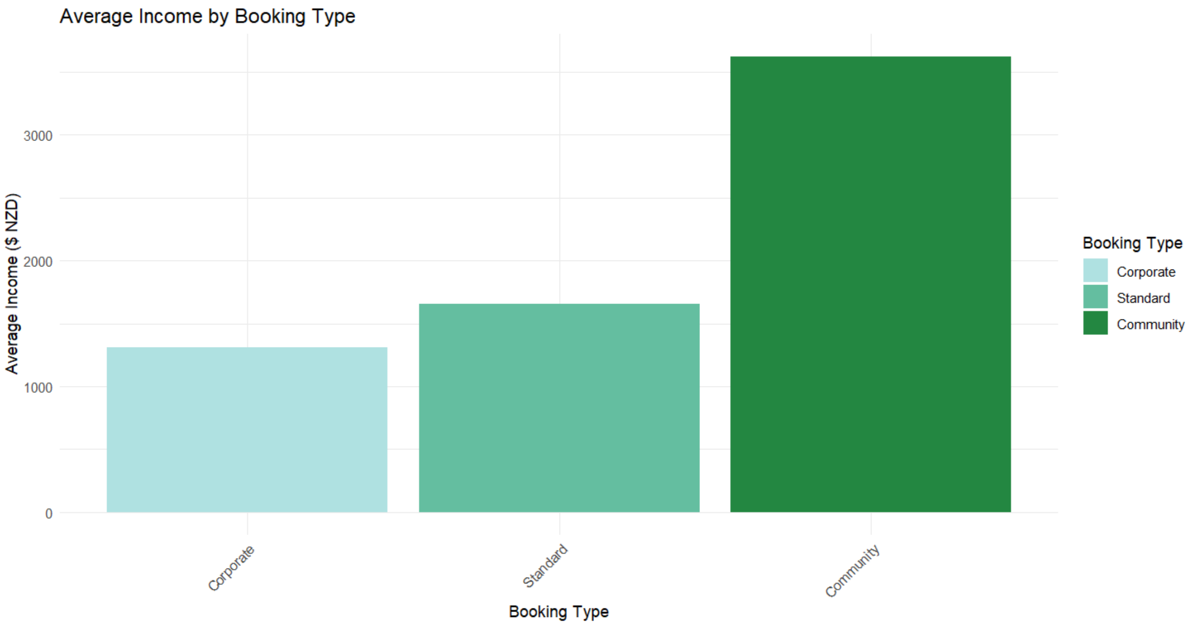


Figure 13 Average income by booking type and year

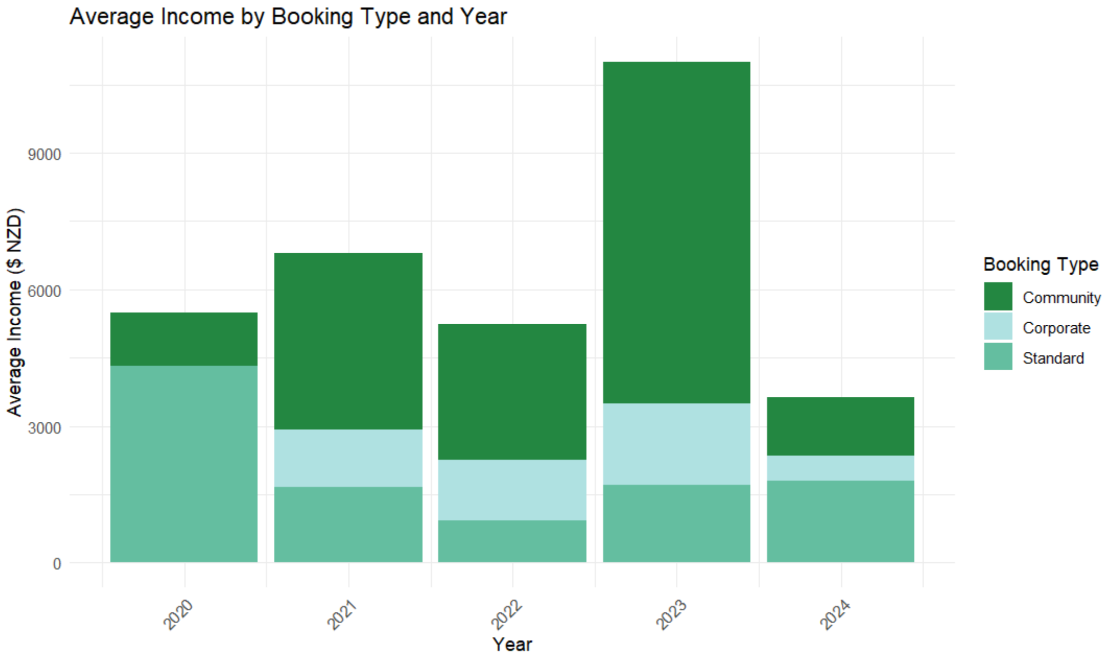


Figure 14 Average expenses by boking type

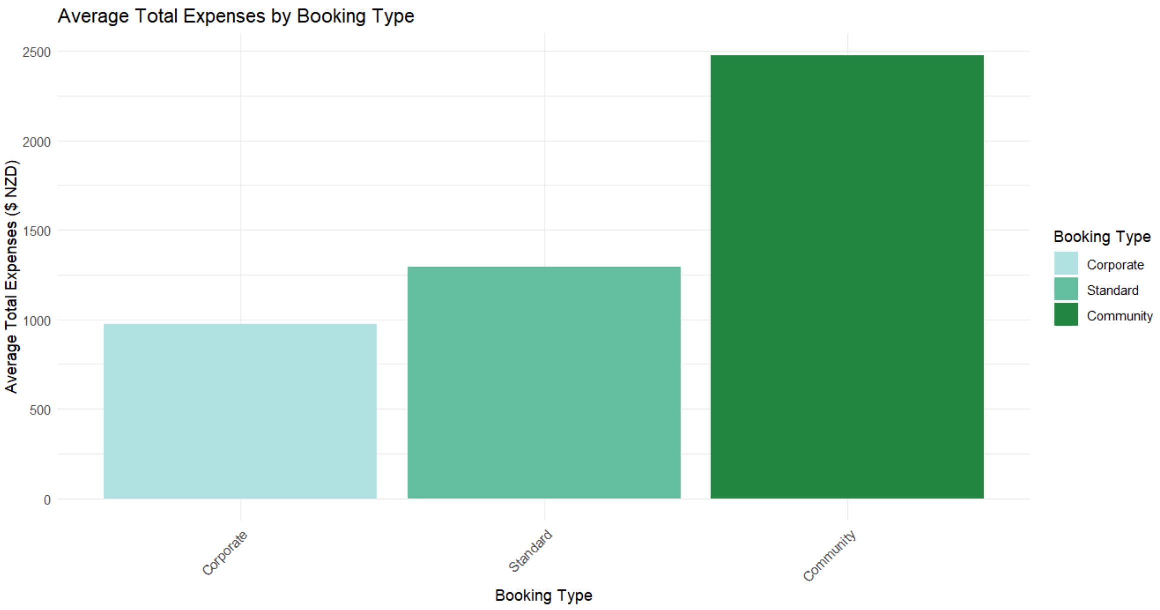
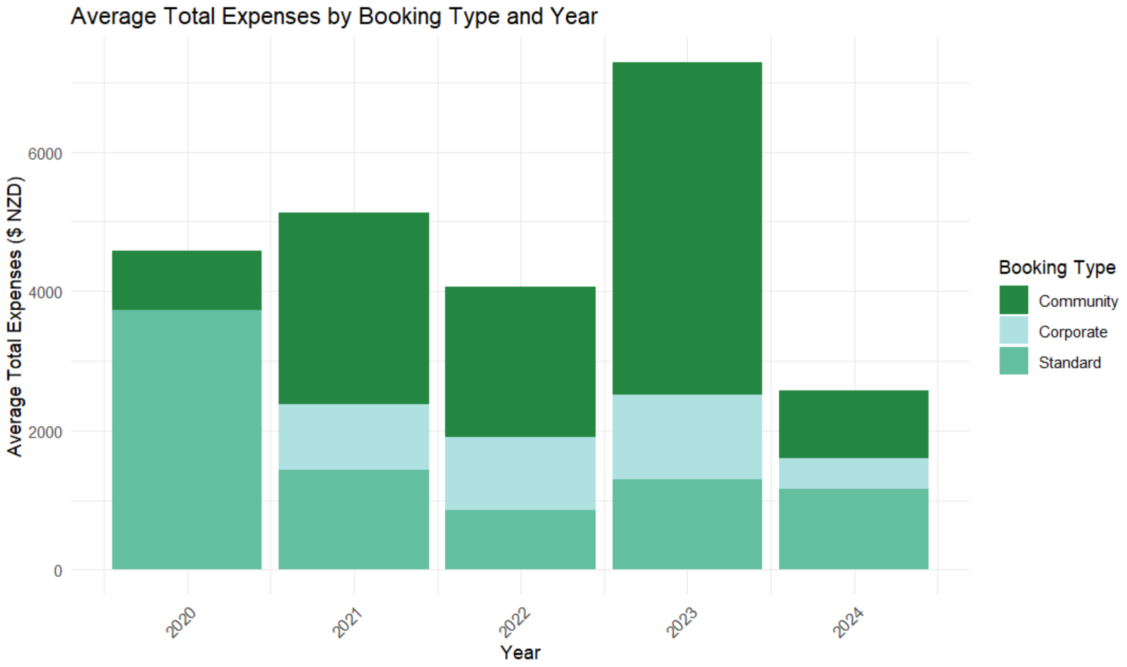


Figure 15 Average total expenses by booking type and year



What is the most frequently booked booking type (standard, community, corporate)?

The most frequently booked booking type over all is Standard (55), followed by Community (40).

This is also true for 2022-2023 however there were more standard bookings in 2021 and 2024.

Figure 16 Overall distribution of booking type

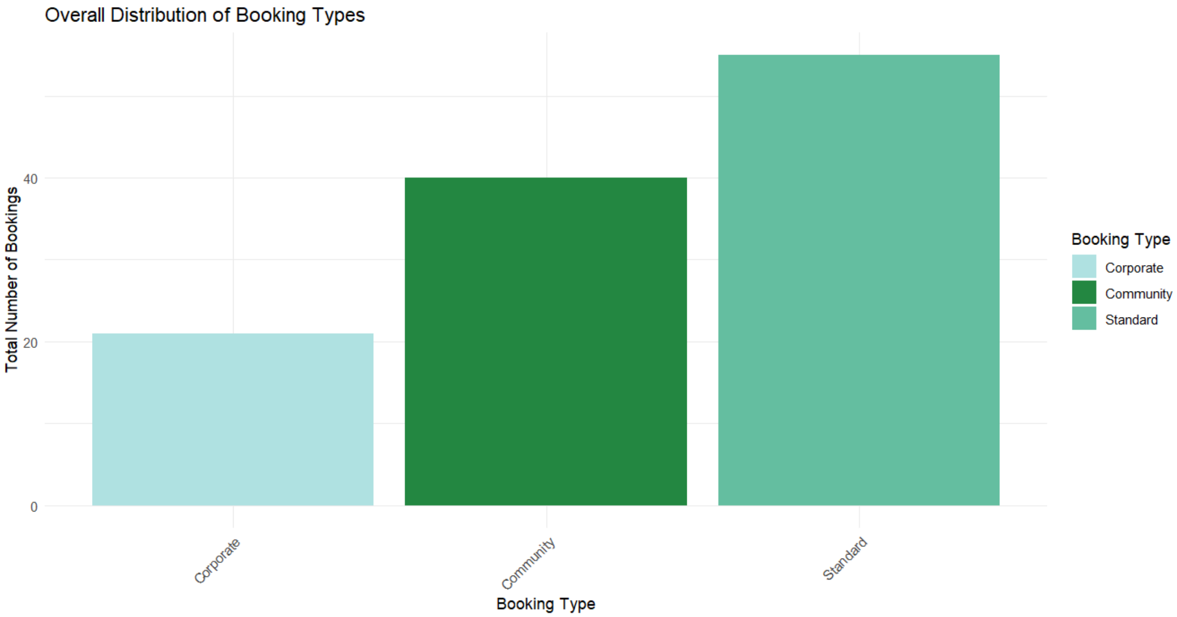
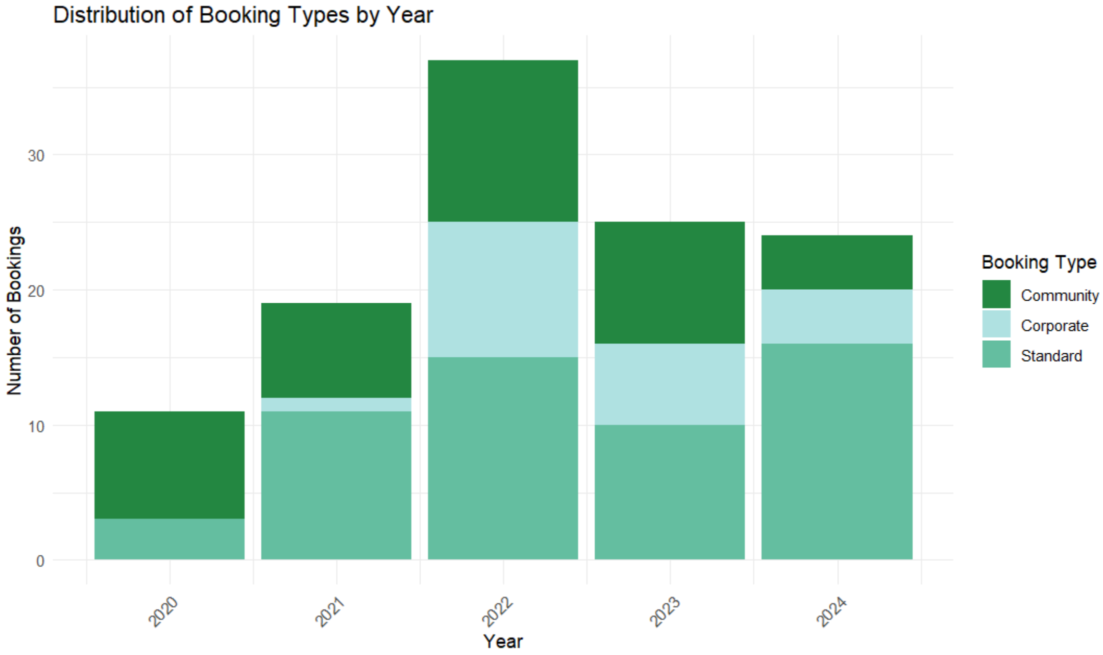


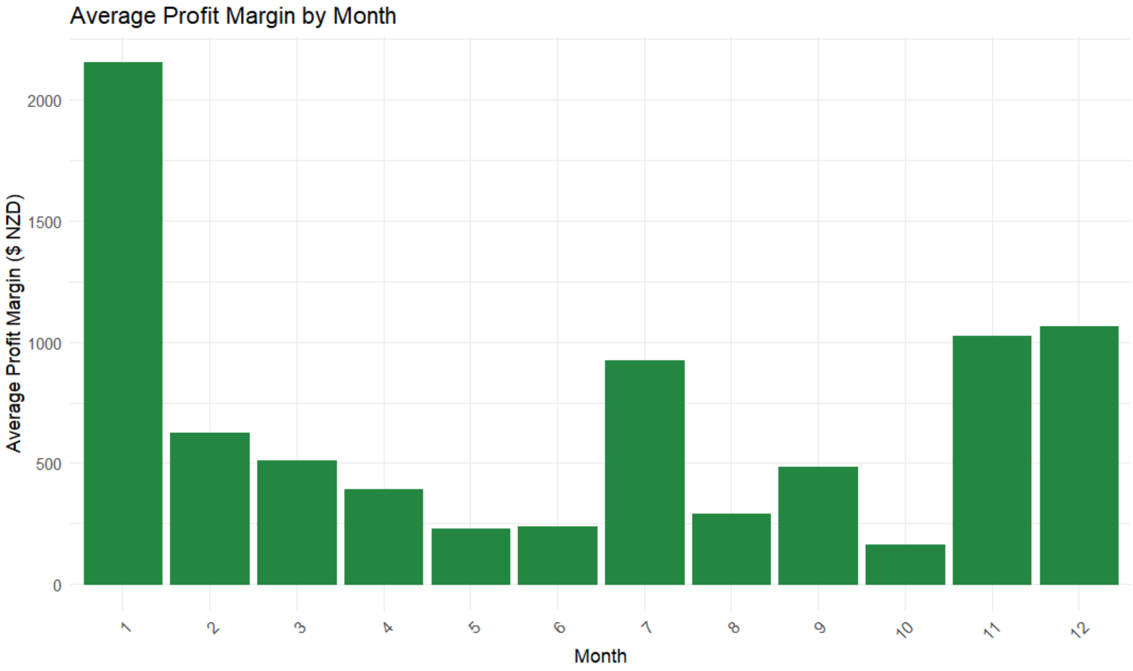
Figure 17 distribution of booking type by year



What month has the highest profit margin?

January has the highest average profit margin ($2153), and October is the least profitable ($165). The second highest average profit margin is December ($1,063). This can be used to design a strategy and plan for seasonal variations. A lot more events tend to happen over summer months when people are outside more. This can be further investigated with a trend analysis.

Figure 18 Average profit margin by month



Correlation and regression analysis

A Pearson correlation test was performed and determined that income was closely related to profit (correlation= 0.9484) and total expenses (correlation= 0.9836). This means that as income increases profit, and expenses also go up. The relationship between income and profit is not as perfectly linear as income is with total expenses. A scatter plot was created to visualise the relationship between income and total expenses as well as income with profit. Figure 10 shows that there is a near perfect alignment of data points along the regression line. This shows that there is a strong positive relationship. Figure 11 shows a strong positive relationship between income and profit.

Figure 19 scatter plot of income vs total expenses

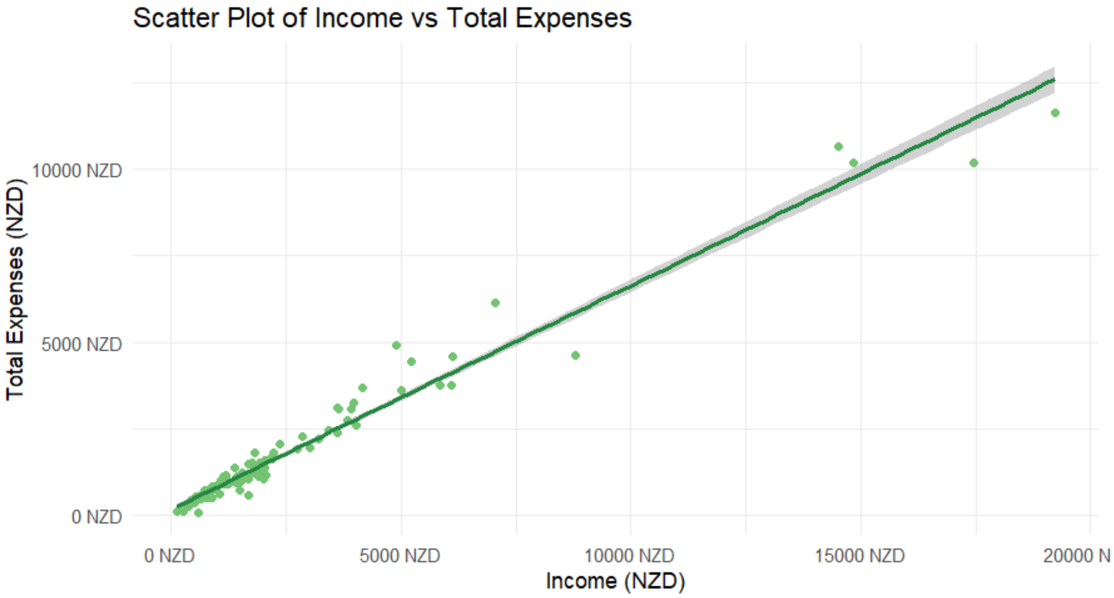
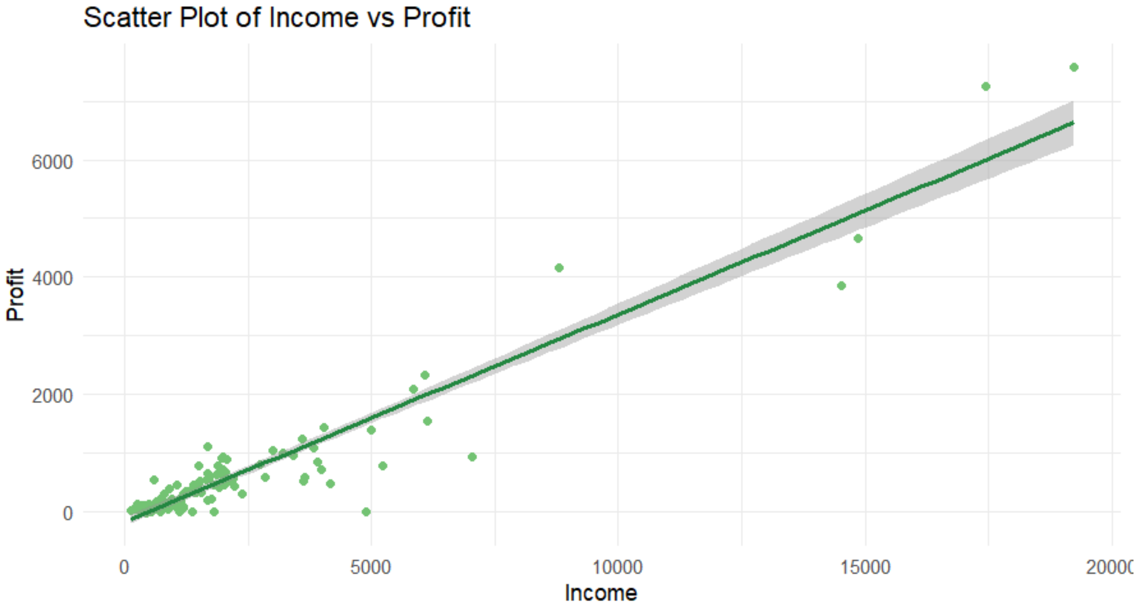


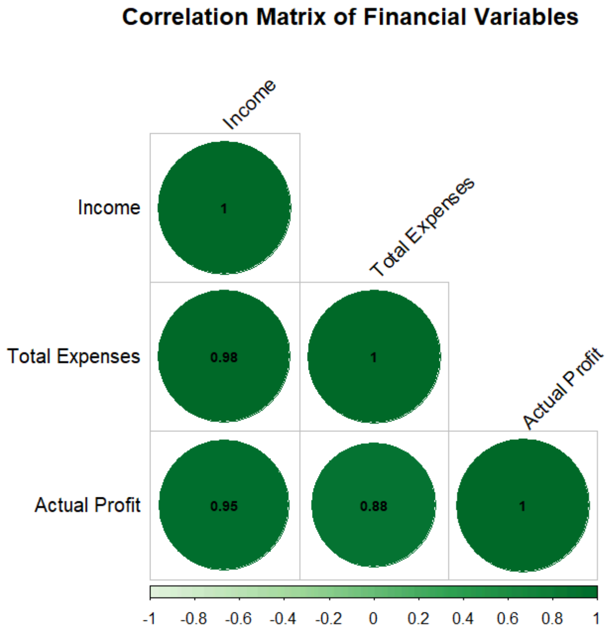
Figure 20 scatter plot of income vs profit



A multiple regression model confirms that income and total expenses are significant predictors of profit. This is shown by a coefficient of 1 and -1 for income and total expenses respectively. This means that each dollar directly affects profit. There is no unexplained variability.

The multiple regression model also indicates that booking types are not statistically significant, suggesting booking type does not have a significant impact on profit.

Figure 21 Correlation Matrix



Extended analysis-R-square, MAPE

The R-squared value of 0.8995 suggests that 89.95% of the variance in profit can be explained by the income; this provides further evidence that there is a strong relationship between income and profit. After removing outliers for profit using the z-score method the R-squared value decreases to 76.6%. This seems to indicate that the outliers play a significant role in the data's variance.

A MAPE test (mean absolute percentage error) was run, initially it was extremely high (3268.5%) which meant that the models' predictions are far from the actual values. The data was then cleaned further removing very low and small values, consequently the MAPE dropped significantly to 92.24%. This improved the model's predictive accuracy, but the error is still high. Following this outlier was removed and the MAPE dropped further to 65.5%. This is better but still shows a considerable error in prediction. This could be due to the data not being normally distributed which was confirmed by a Shapiro-Wilk normality test. Further investigation should be conducted on this during the prediction modelling stage.

The Shapiro-Wilks normality test showed that all three key variables (profit, income, expenses) are not normally distributed with p-values less than 2.2e-16. This will assist us when picking which prediction models to run as the reliability of linear regression models can be questionable.

To determine whether there are statistically significant differences between the means of income and booking type, profit across booking type and income across months an ANOVA test was conducted. There is no statistically significant difference between the months when compared to income with a p-value greater than 0.05 (0.198).

There is a statistical significance for income across types. the P-value is less than 0.05 (0.00261). A Tukey post-hoc test was run and determined that corporate and standard bookings have significantly different incomes compared to community types but there is no difference between corporate and standard bookings.

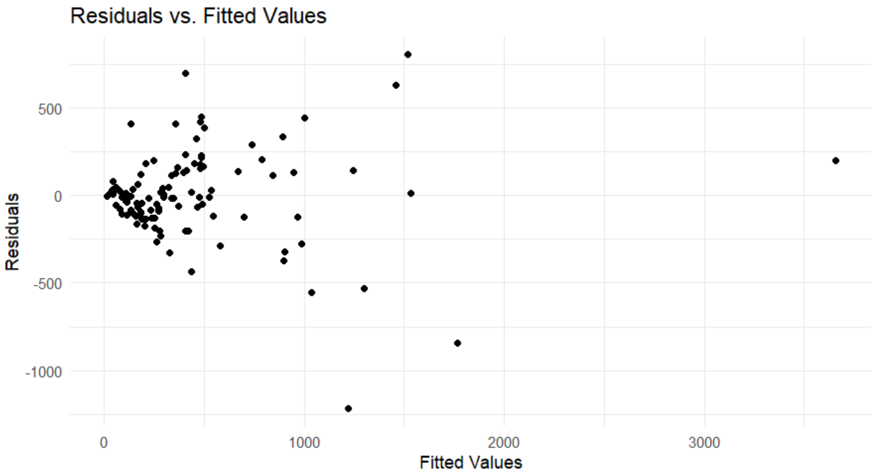
There is a statistically significant difference for profit across the booking types. The P-value is less than 0.05 (0.023). Tukey's post hoc test showed that there is a significant difference between profit and corporate and community; as well as standard and community. However, there is no significant difference for profit between standard and corporate types.

Initial modelling

To have a better understanding of the data when the prediction modelling is started an initial linear regression model was fitted. The profit was set as the dependent variable with income and total expenses as independent variables. It was run on the original data and a version with outliers removed.

A Residual vs fitted values plot was done to see if the residuals exhibit any patterns or non-linearity. The plot was very difficult to interpret so outliers were removed to confirm if there are any patterns.

Figure 22 Residuals vs Fitted values



Once the outliers were removed you can see that there is non-linearity of the data. The Q-Q plot of the data showed that the residuals deviate significantly from the normal line indicating that the data is not normally distributed. If it where you would see more data points on the red line

Figure 23 Q-Q plot of residuals

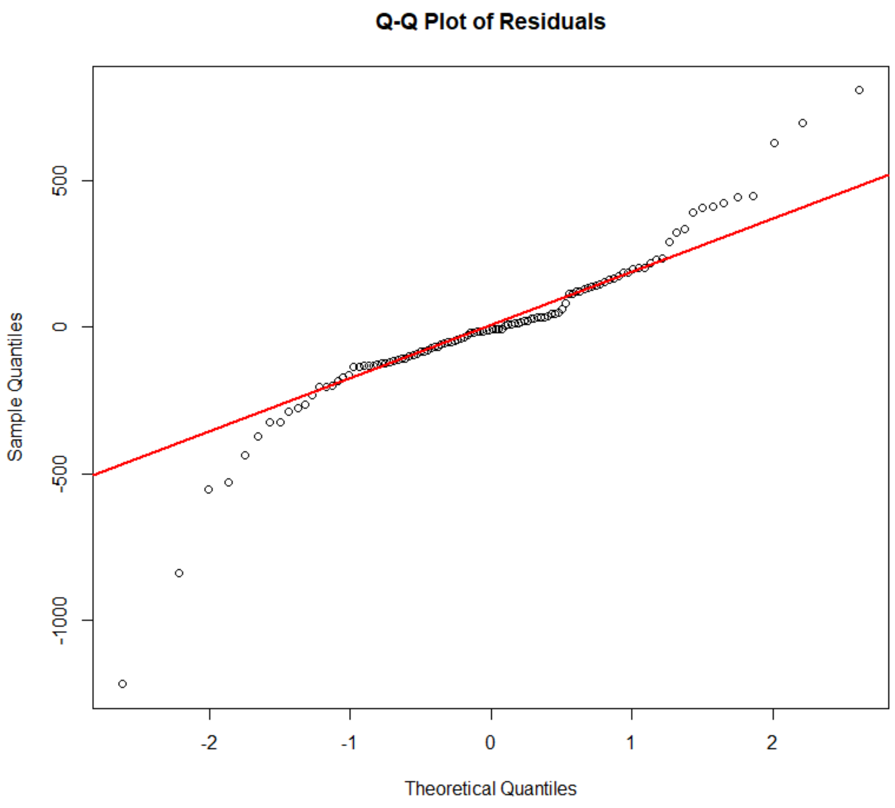
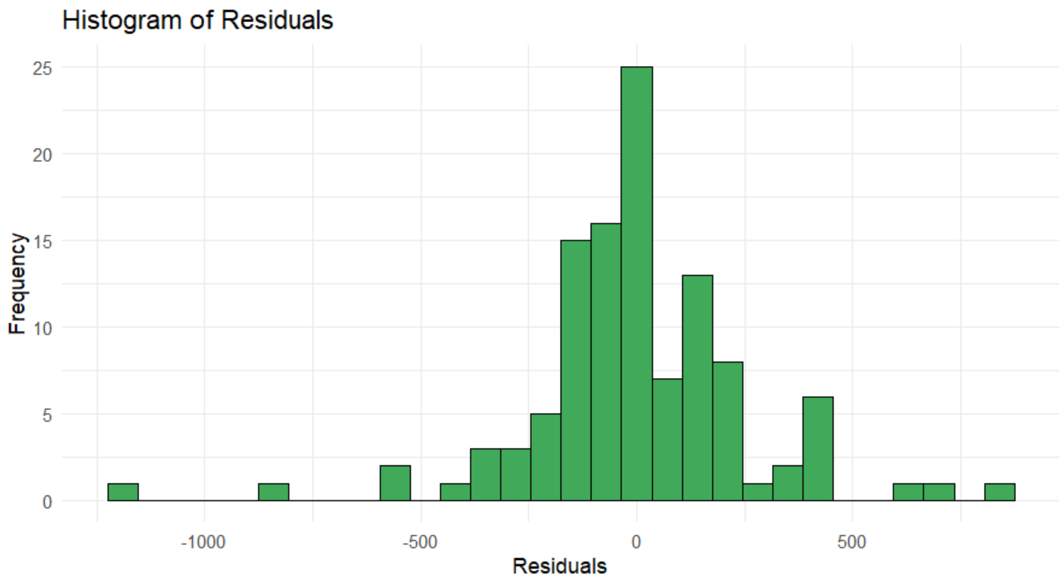
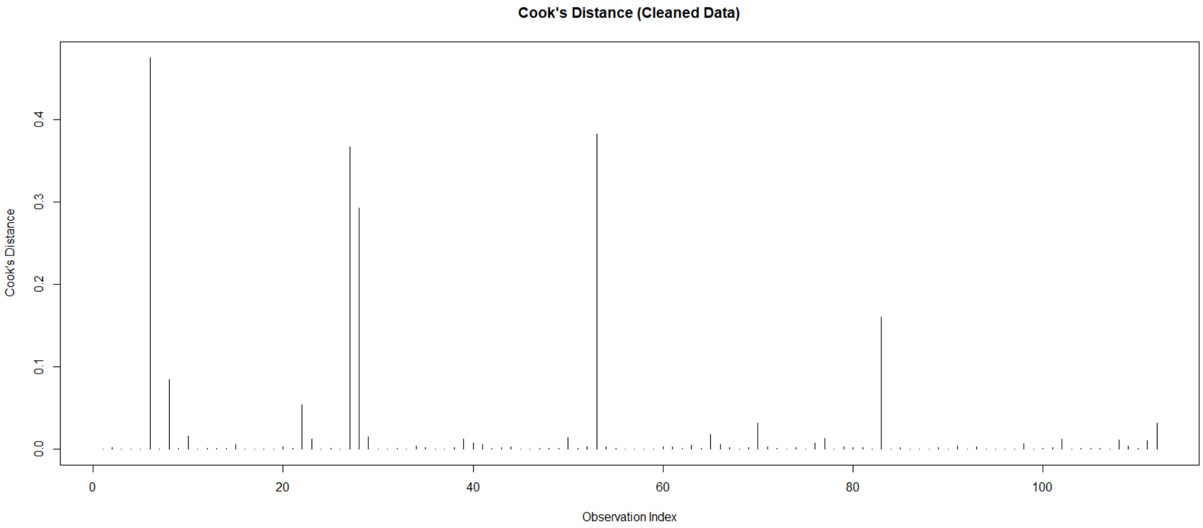


Figure 24 Histogram of residuals



A cook's distance plot was done to identify influential observations that might affect the model, once the outliers were removed the plot showed fewer influential points meaning the data fitted better and were more stable.

Figure 25 Cook's Distance plot



The Shapiro wilks test done on the residuals confirmed that the data is still not normally distributed.

There are several issues with this model that will help the team when we do the prediction modelling stages, these issues are normality, heteroscedasticity and influential points. These will need to be further investigated to ensure the predictive model used fits the data well.