Understanding International Visitor Trends in New Zealand

Group #5

2024-10-14

Group 5:

- Hui Xue,
- Margarita Grishechkina,
- Samuel Stewart,
- Yufei Zhu,
- Zhikai Yao

Introduction

The tourism sector is not a standalone industry but includes components from various sectors such as accommodation, food services, retail, arts, recreation, and transport. In 2023, tourism contributed **NZD 13.3 billion** to New Zealand's GDP, which accounted for **3.5%** of the country's economic output. This is lower than the **4.2%** contribution seen in 2020 before the COVID-19 pandemic but higher than the **2.7%** during the pandemic. This highlights tourism plays important role in New Zealand's economy. ***Tourism GPD***. Thus, the project aim was to analyze the trends in international visitor arrivals to New Zealand over the past five years with understanding how these changes impact various aspects of the tourism sector, such as marketing strategies, resource allocation, and overall competitive positioning.

Research question

How have international visitor trends changed over the past five years, and what impact do these changes have on New Zealand's tourism industry?

Pre-processing

The main dataset contains 1,657,122 observations and 15 variables. We cleaned the data by removing discrepancies, checking for missing values, and normalizing it. We also split the data into distinct periods, combined daily data into monthly summaries, and applied log transformations to make patterns more visible.

```r
# Boxplot for outliers by year
boxplot(total_movements ~ year, data = monthly_data)
```

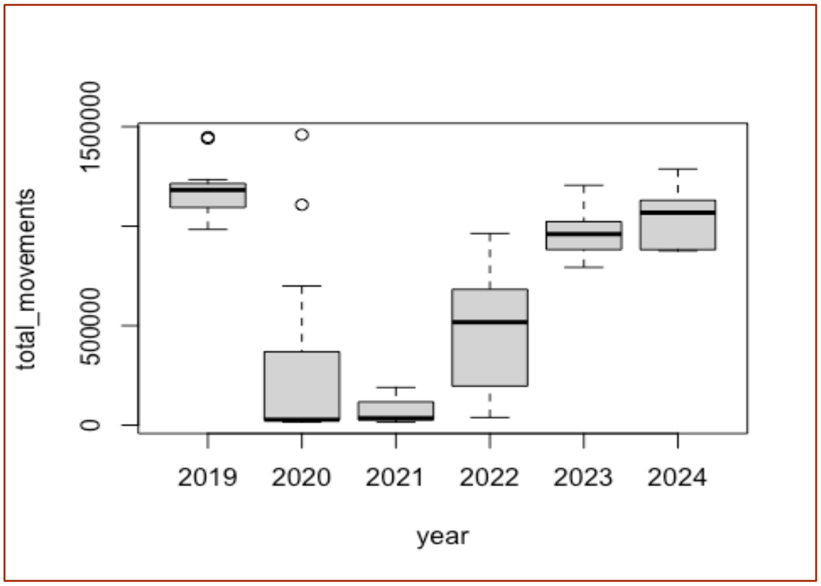

Figure 1, Boxplot for outliers by year

The boxplot shows the total movements by year from 2019 to 2024. In 2019: The total movements are relatively stable with few outliers. There is a significant decrease in total movements due to COVID-19 in 2020, with a wider spread indicating more variability. The lowest point shows in 2021, reflecting heavy restrictions. 2022 year shows some recovery but still below pre-COVID levels, representing a wider range of values. Finally, total movements are increasing in 2023-2024, showing a gradual recovery close to 2019 levels.

```r
# Combine year, month, and day into a single date column
workdf$date <- as.Date(with(workdf, paste(year, month, day,
                                          sep = "-")), "%Y-%m-%d")

# Check for missing values in the entire dataframe
if (all(colSums(is.na(workdf)) == 0)) {
  print("There are no missing values in the dataset.")
} else {
  print("There are missing values in the dataset.")
}
```

```
## [1] "There are no missing values in the dataset."
```

We use R Markdown to prepare this report, combining code, visualizations, and textual explanations in a single document. This allows us to generate dynamic and reproducible reports that include both the analysis and the narrative, enabling seamless integration of R code and output into well-structured documents.

Packages Used:

1. *dplyr: For data manipulation tasks like filtering, grouping, and summarizing.*
2. *ggplot2: For creating static data visualizations.*
3. *kableExtra: Enhances table formatting in reports.*
4. *knitr: Used to generate dynamic reports in R Markdown.*
5. *leaflet: Used for creating interactive web maps.*
6. *lubridate: For working with date and time data easily.*
7. *maps: Provides geographical maps for visualization.*
8. *randomForest: Implements the Random Forest algorithm for prediction tasks.*
9. *readr: For reading and writing data from flat files (e.g., CSV).*
10. *readxl: For reading Excel files into R.*
11. *sf: For handling geospatial data.*

Analysis

1. Yearly Visitor Number Fluctuations from 2003 to June 2024

There are two key dates indicating the official COVID-19 period in New Zealand:

- **March 19, 2020** - New Zealand closed its borders to nearly all non-residents and non-citizens to contain the spread of COVID-19
- **July 31, 2022** - The government fully reopened its borders to all international visitors and visa categories, marking the final step in reconnecting New Zealand with the world. ***(NZ Immigration)***

First, the plot was created to visualize time series over the longer period from 2003.

```r
# Create a plot to visualise time series
plot(ts_visitor, col = "blue", ylim = range(c(ts_visitor, ts_total)),
     ylab = "Number of Visitors", xlab = "Year",
     main = "Time Series of Foreign Visitors to New Zealand",
     xaxt = "n", yaxt = "n",
     cex.main = 0.8,
     cex.lab = 0.5,
     cex.axis = 0.4)

axis(2, at = pretty(range(c(ts_visitor, ts_total))),
     labels = format(pretty(range(c(ts_visitor, ts_total))), big.mark = ","),
     las = 1, cex.axis = 0.6)

years <- seq(start_year, 2024, by = 1)
axis(1, at = years, labels = years, cex.axis = 0.6)

lines(ts_total, col = "red")
legend("topleft", legend = c("Visitors Visa", "Total Visitors"),
       col = c("blue", "red"), lty = 1, cex = 0.7)
```

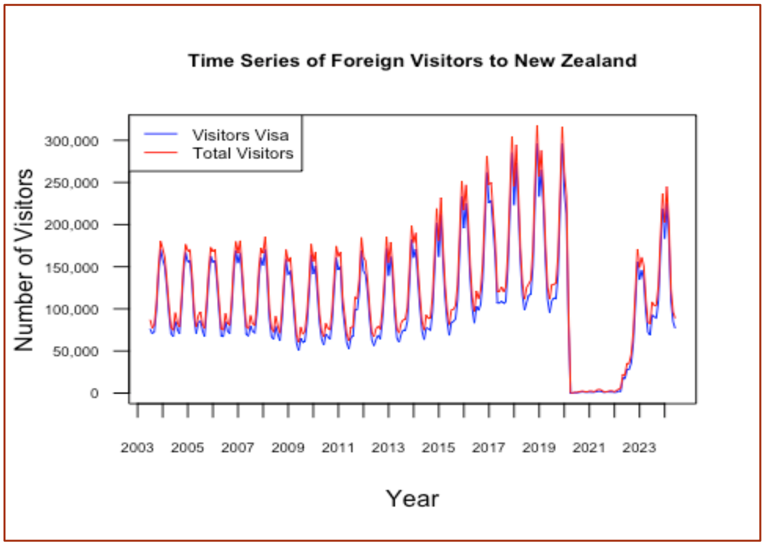

Figure 2, Time series of foreign people to NZ

Before the pandemic, from 2003 to 2019, visitor numbers to New Zealand steadily increased, with a small drop around 2009, likely due to the global financial crisis. However, in 2020, visitor numbers sharply declined due to COVID-19, with border closures and travel restrictions almost completely stopping international tourism.

After 2022, as restrictions eased, visitor numbers started to recover but have not yet reached the pre-pandemic peaks seen in 2016–2017. Most foreign visitors come on visitor visas, although Australian citizens are excluded from this analysis due to visa waiver policies, allowing for consistency across the report.

2. *Travel Purpose*

```r
# Plot lines
ended_year_july <- read_excel(new_file, sheet = "Year Ended July")
long_data <- pivot_longer(ended_year_july, cols = c("2020", "2021", "2022",
                                                     "2023", "2024"),
                          names_to = "Year", values_to = "Value")
ggplot(long_data, aes(x = Year, y = Value, color = `Travel purpose`,
                      group = `Travel purpose`)) +
  geom_line(size = 0.8) +
  geom_point() +
  labs(title = "Comparison of Travel Purposes Across Years 2020-2024",
       x = "Year",
       y = "Number of Visitors",
       color = "Travel Purpose") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 20))
    ) +
  scale_y_continuous(labels = comma)
```

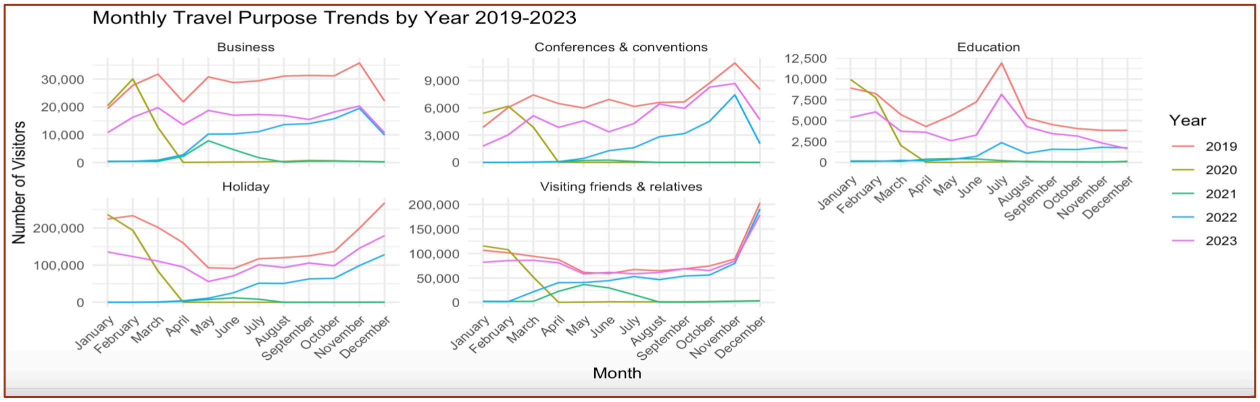

Figure 3, Monthly Travel Purpose by year, 2019-2023

The travel trends are primarily driven by holiday and visiting friends and relatives, which see significant peaks in July and December. Business and conference-related travel stays consistent but at comparatively lower levels. Education-related travel spikes during the July-August period, which aligns with school holidays

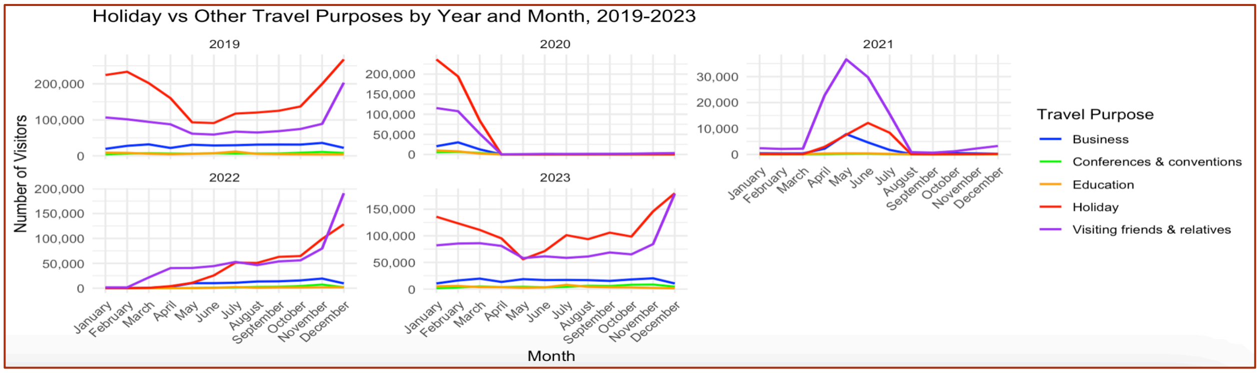

Figure 4, Holidays vs other travel purposes by year, 2019-2023

Holiday travel leads across most months, with a pronounced peak in 2023. The data shows a significant decline in 2020 because of the Covid-19 pandemic, followed by a recovery in 2021, which is most evident in the resurgence of holiday travel.

Next to compare seasonal travel purpose trends.

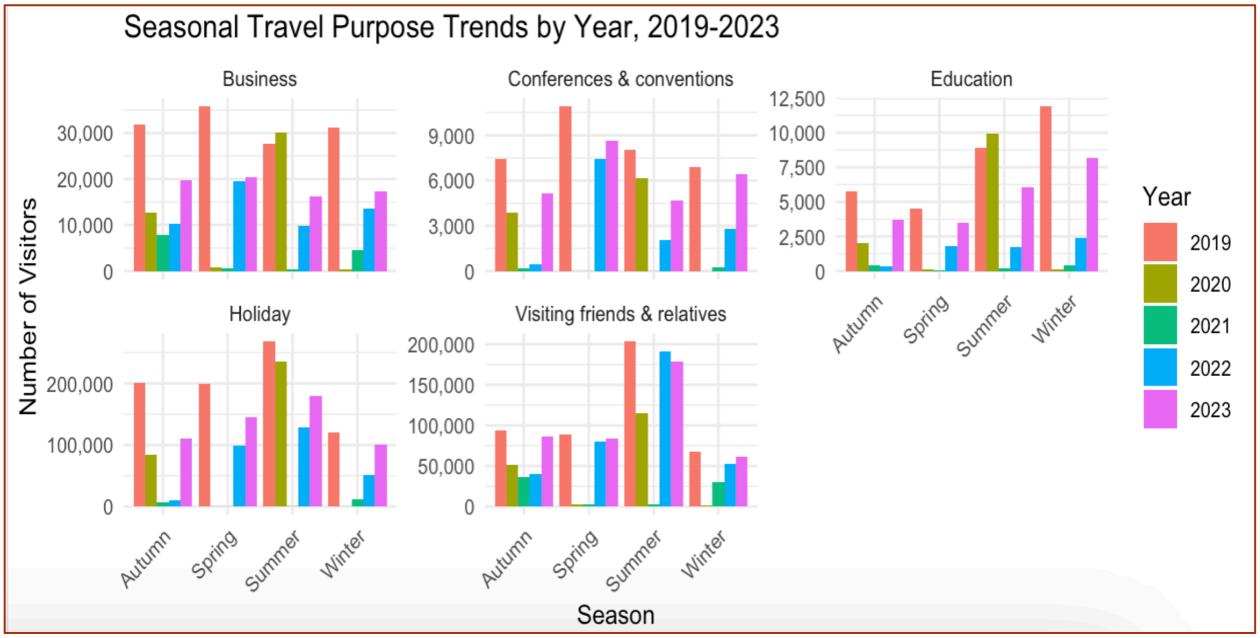

Figure 5, Seasonal Travel Purpose Trends by year 2019-2023

Winter and summer experience the highest travel volumes for holidays, while business travel remains relatively stable across spring and autumn. Travel to visit friends and relatives' peaks in both winter and summer, likely due to holiday seasons and school breaks.

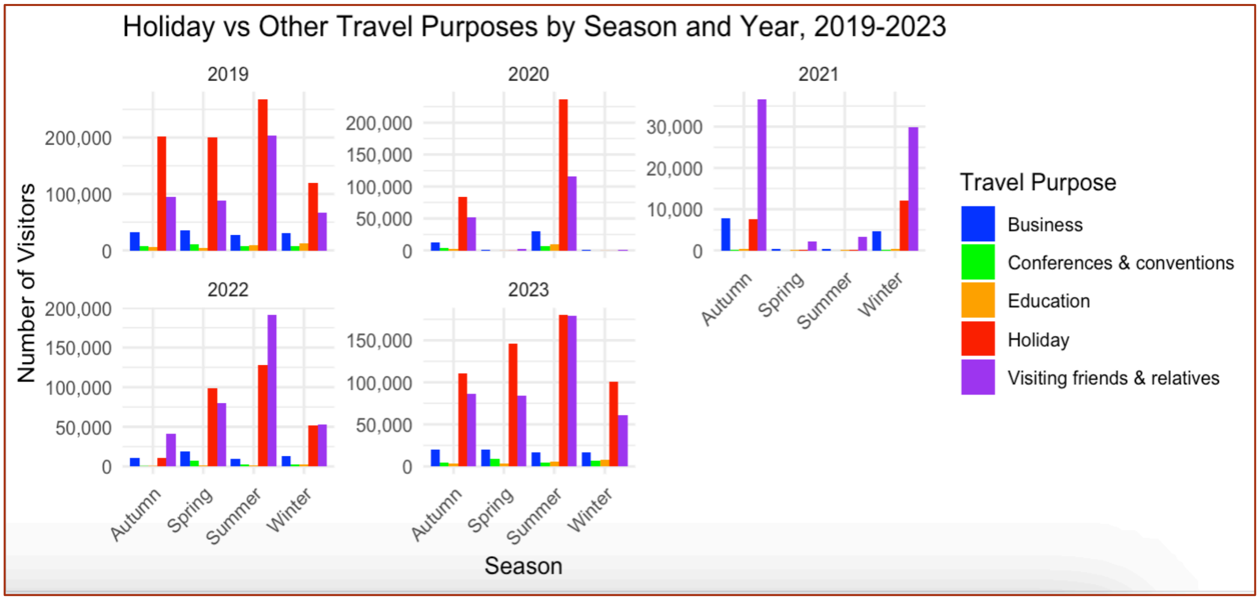

Figure 6, Holidays vs other travel purpose by season, 2019-2023

Holiday travel consistently surpasses other travel purposes, especially during the winter and summer seasons. While business, conference, and education-related travel started to recover after 2022, they still trail behind holiday travel in terms of visitor numbers.

3. *Processing of arrivals by NZ airports movements*

The heat map below shows the total movements across airports in New Zealand, with larger circles indicating airports with higher movement frequencies. Airports with over 500,000 movements are marked in red, while those with fewer are marked in blue. By using heat map, it helps quickly highlights airports with the most traffic and provides a spatial understanding of where air traffic is concentrated across the country.

```
# Code sample
leaflet(data = work_with_coords) %>%
  addTiles() %>%
  addCircleMarkers(
    ~longitude, ~latitude,
    radius = ~sqrt(total_movements) / 200,
    color = ~ifelse(total_movements > 500000, "red", "blue"),
    fillOpacity = 0.5,
    popup = ~paste("Airport:", customs_port_code, "<br>Total Movements:",
                   total_movements)
  ) %>%
  setView(lng = 174.7850, lat = -41.2865, zoom = 5) %>%
  # Add a title
  addControl("<strong>NZ Total Movements</strong>", position = "topright") %>%
  # Add a legend
  addLegend(
    position = "bottomright",
    colors = c("red", "blue"),
    labels = c("> 500,000", "<= 500,000"),
    title = "Frequency",
    opacity = 1
  )
```

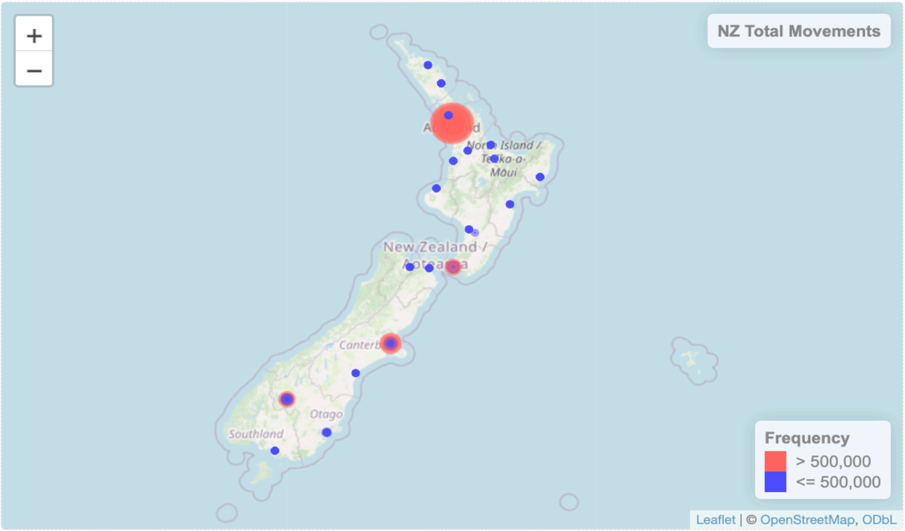

Figure 7, Plot airport movements on the New Zealand map

The table 1 displays the Top 5 airports in New Zealand based on their total movements from 2019 to 2024. It provides the movement data for each year, showing how airport traffic fluctuated, especially before, during, and after the COVID-19 pandemic.

```
# Create a table to show top 5 airports over the time
# Get the top 5 airports per year
top5_airports_per_year <- work_with_coords %>%
  group_by(year) %>%
  arrange(year, desc(total_movements)) %>%
  mutate(rank = row_number()) %>%
  filter(rank <= 5) %>%
  select(year, airport_name, total_movements)

# Reshape the data so that each year is a column
top5_wide <- top5_airports_per_year %>%
  pivot_wider(names_from = year, values_from = total_movements)

# Create a table
top5_wide %>%
  kable(col.names = c("Airport Name", "2019", "2020", "2021", "2022", "2023",
                      "2024"),
        caption = "Top 5 Airports by Year (Total Movements)") %>%
  kable_styling(full_width = FALSE, position = "center", font_size = 12)

# Compare years table
compare_years %>%
  kable(col.names = c("Year", "Total Arrivals", "Percentage"),
        caption = "Year comparison of arrivals") %>%
  kable_styling(full_width = FALSE, position = "center", font_size = 12)
```

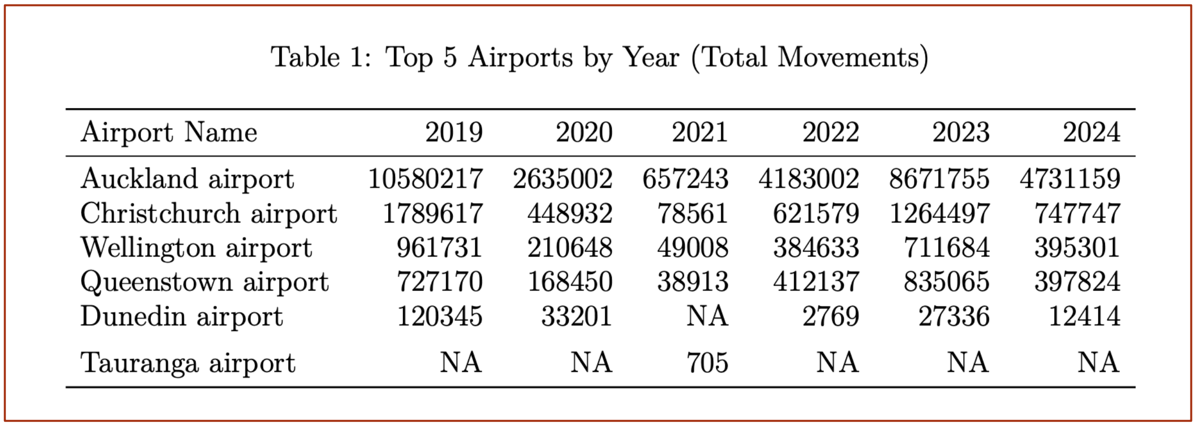

Table 1: Top 5 Airports by Year (Total Movements)

Airport Name	2019	2020	2021	2022	2023	2024
Auckland airport	10580217	2635002	657243	4183002	8671755	4731159
Christchurch airport	1789617	448932	78561	621579	1264497	747747
Wellington airport	961731	210648	49008	384633	711684	395301
Queenstown airport	727170	168450	38913	412137	835065	397824
Dunedin airport	120345	33201	NA	2769	27336	12414
Tauranga airport	NA	NA	705	NA	NA	NA

Auckland Airport consistently had the highest movements, despite a significant drop in 2020 (COVID-19 lockdown) and then recovery in 2023 and 2024. The other airports, like Christchurch and Queenstown, follow a similar pattern with dips in 2020 and gradual recovery. This table allows for easy comparison of how each airport has been impacted by global events (like COVID-19) and how movements are recovering over time. It also highlights which airports are the most significant in the country, making it useful for understanding the overall aviation recovery trend.

The table 2 highlights the percentage of total arrivals relative to 2019, the last full pre-COVID time.

```
# Compare arrivals
# Summarize total movements by year
yearly_totals <- workdf %>%
  group_by(year) %>%
  summarise(yearly_movements = sum(total_movements, na.rm = TRUE))

# Filter for years
base_year <- yearly_totals %>% filter(year == 2019) %>% pull(yearly_movements)
compare_years <- yearly_totals %>%
  filter(year %in% c(2019, 2020, 2021, 2022, 2023, 2024)) %>%
  mutate(ratio_to_2019 = yearly_movements / base_year * 100) %>%
  mutate(ratio_to_2019 = round(ratio_to_2019, 2))

# View the result
compare_years %>%
  kable(col.names = c("Year", "Total Arrivals", "Percentage"),
        caption = "Table 2, Year comparison of arrivals") %>%
  kable_styling(full_width = FALSE, position = "center", font_size = 12)
```

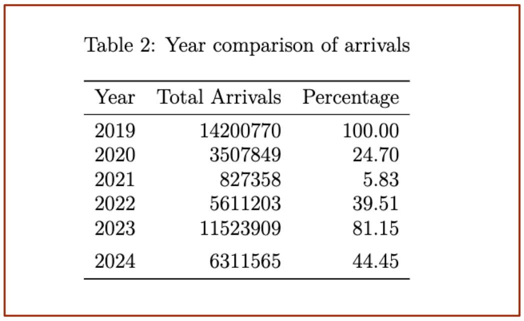

Table 2: Year comparison of arrivals

Year	Total Arrivals	Percentage
2019	14200770	100.00
2020	3507849	24.70
2021	827358	5.83
2022	5611203	39.51
2023	11523909	81.15
2024	6311565	44.45

In 2020, only 24.7% of 2019's traffic was recorded due to border closures in March. In 2021, this dropped further to 5.83% during the peak of restrictions. By 2022, with borders partially reopening, arrivals recovered to 39.51% of 2019 levels. In 2023, a strong recovery saw arrivals reaching 81.15%. By mid-2024, arrivals are at 44.45%, showing growth in the first half of 2024 compared to previous years, but still not back to pre-pandemic levels.

4. *Visitor Number Changes by Country*

```
ggplot(top10_long, aes(x = Year, y = Arrivals, color = `Home Countries`, group = `Home Countries`)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(title = "Trends of Top 10 Home Countries of Visitors (2020~2024) ",
       x = "Year", y = "Number of Visitors",
       color = "Home Countries") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 20))
    )
```

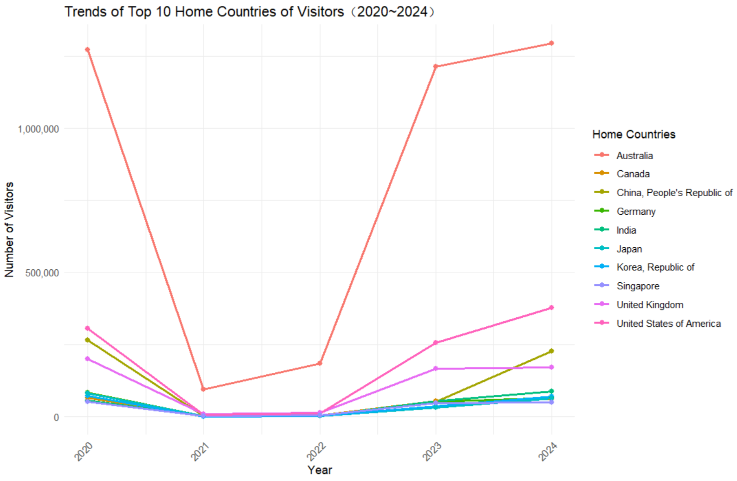

Figure 8, Top 10 home countries

Based on this part of analysis we have identified that Australia remains the leading source of visitors to New Zealand. Although visitor numbers dropped significantly in 2021 due to the pandemic, they rebounded quickly and remained strong in 2023 and 2024. The United States and the United Kingdom also experienced a rapid recovery after 2022, with steady growth through 2023 and 2024. China and Japan showed continuous growth in visitor numbers, particularly China, which experienced a sharp rebound from the low point in 2021. European countries like Germany and France also showed slow but steady growth in tourist numbers.

5. Visitor characteristics

5.1. Age processing

```
ggplot(age_summary, aes(x = age_at_travel_range, y = total_movements, fill = period)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.5)) +
  labs(title = "Impact of Age Groups on Tourism Before, During, and After COVID-19",
       x = "Age Group",
       y = "Total Movements") +
  theme_minimal() +
  scale_fill_manual(values = c("Before COVID-19" = "#00AFBB", "During COVID-19" = "#E7B800", "After COVID-
19" = "#FC4E07")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

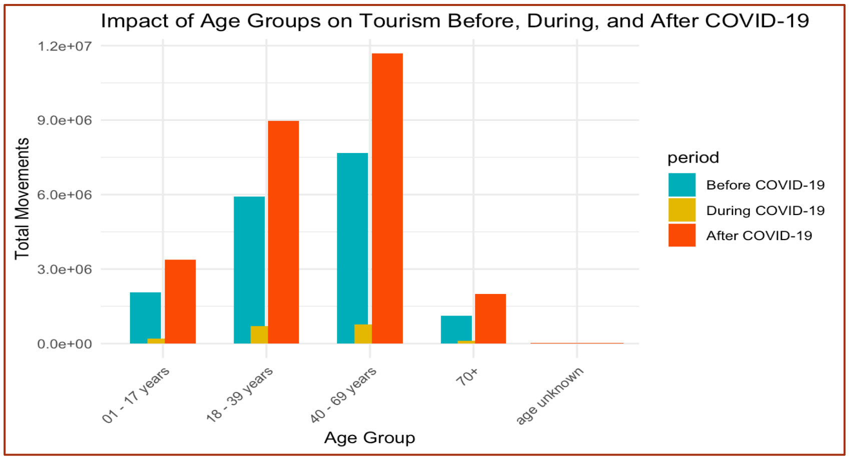

Figure 9, Impact of age groups on three periods of pandemic

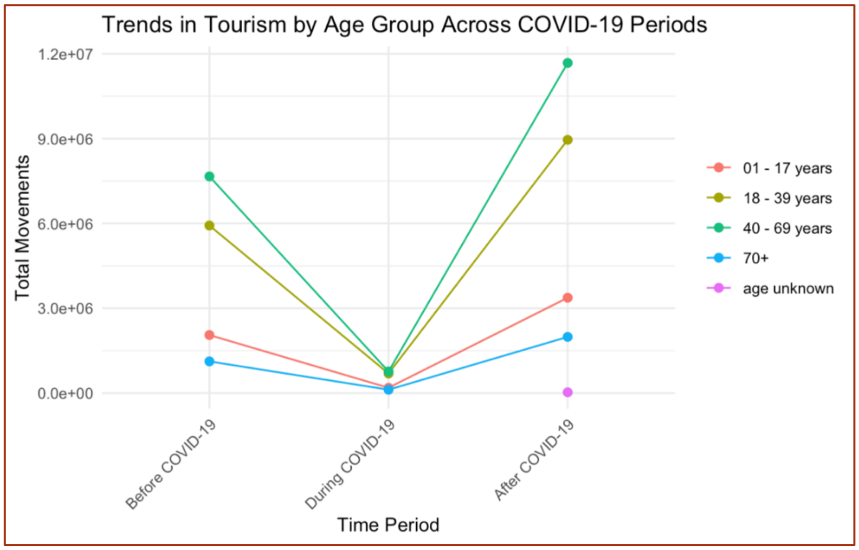

Figure 10, Trends in tourism by age group

These visualizations compare visitor movements by age group across three distinct periods. The bar plot shows total movements for each group, while the line chart highlights trends over time, illustrating how COVID-19 impacted tourism in different age demographics.

Before COVID-19, tourism in New Zealand was strong across all age groups, with the 18-39 and 40-69 age groups being the most active. Older travelers were also frequent, though in slightly lower numbers. During the pandemic, all age groups saw a sharp decline in travel, especially younger and older travelers, reflecting global travel bans and uncertainty. Older travelers were likely more cautious due to health risks. After COVID-19, there are signs of recovery, but the pace varies. Younger travelers have rebounded quickly, while older travelers have been slower to return, possibly due to lingering safety concerns and changing preferences.

5.2. Gender Analysis

As shown in the bar chart here we can see a balanced distribution of travel between genders with no significant gap.

```
# Gender distributional plot

ggplot(gender_distribution, aes(x = sex_code,
                                y = Total_Movements, fill = sex_code)) +
  geom_bar(stat = "identity", color = "black") +
  labs(x = "Gender", y = "Total Movements",
       title = "Gender Distribution of Travelers to New Zealand",
       fill = "Gender") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1") +
  theme(
    axis.title.x = element_text(margin = margin(t = 10)),
    axis.title.y = element_text(margin = margin(r = 10))
  )
```

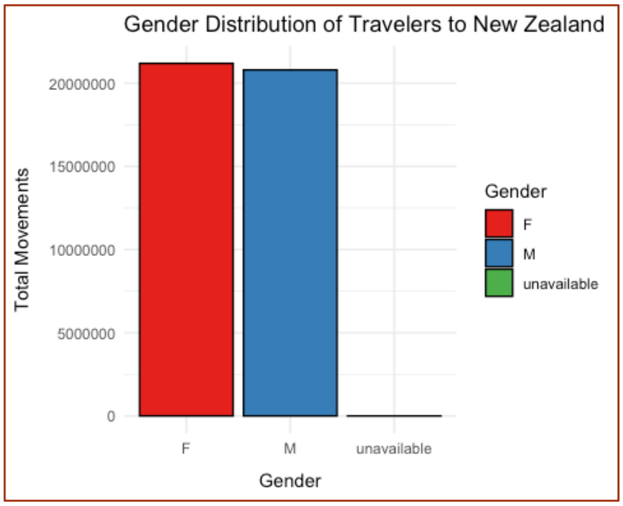

Figure 11, Gender distribution of travels in New Zealand

This graph below shows that both genders follow similar seasonal travel patterns, with peaks during the summer. This indicates that gender does not significantly influence the choice of when to travel, which is more affected by universal factors like climate or holiday seasons. For tourism strategists, this similarity suggests that seasonal promotions and packages can be universally targeted to both genders without the need for significant customization based on gender.

```
# Seasonal gender distribution
ggplot(seasonal_data, aes(x = season, y = Total_Movements, color = sex_code, group = sex_code)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(x = "Season", y = "Total Movements",
       title = "Seasonal Gender Distribution of Travelers") +
  theme_minimal() +
  theme(
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text(margin = margin(r = 15))
  ) +
  scale_color_brewer(palette = "Set1")
```

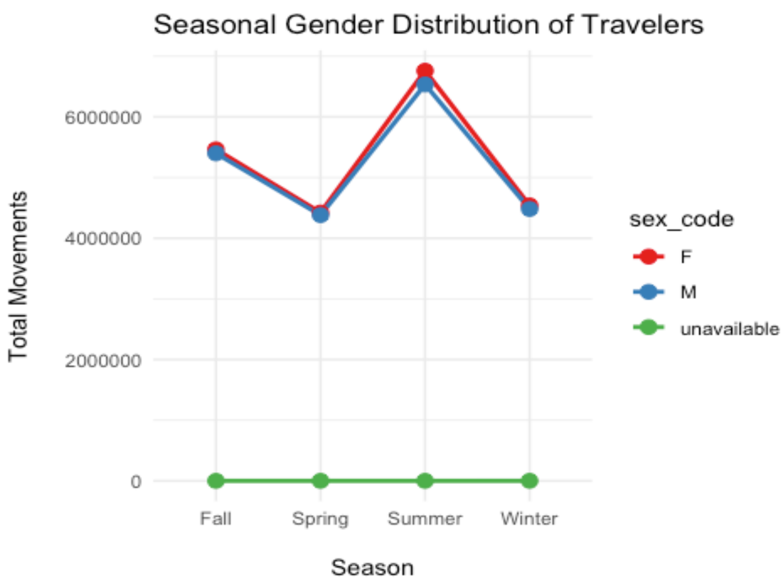

Figure 12, Seasonal gender distribution of travelers

Across all visualisations the balance between male and female remains consistent meaning that travel behaviours are not skewed towards one gender, and there is no need to target one group.

6. Residual analysis

This code analyses visitor movements before, during, and after COVID-19, focusing on age groups and seasons. It prepares the data, applies a Random Forest model to predict visitor movements, and examines the differences between predicted and actual values (residuals). The code also visualizes these residuals by age group and season to understand how well the model performs.

```
# Aggregate monthly data by age group and other relevant features
monthly_data_by_age <- workdf %>%
  group_by(year, month, age_at_travel_range, season) %>%
  summarise(total_movements = sum(total_movements,
                                  na.rm = TRUE), .groups = "drop")

# Log-transform total movements for modeling purposes
monthly_data_by_age$log_total_movements <-
  log1p(monthly_data_by_age$total_movements)

# Fit a Random Forest model on the log-transformed response
fit_rf_log_age <- randomForest(log_total_movements ~ year +
                                 month + age_at_travel_range + season, data = monthly_data_by_age, ntree =
 100)

# Extract predicted values and residuals for the log-transformed model
monthly_data_by_age$predicted_log_rf <- predict(fit_rf_log_age,
                                                monthly_data_by_age)
monthly_data_by_age$residuals_log_rf <-
  monthly_data_by_age$log_total_movements - monthly_data_by_age$predicted_log_rf
```

```r
# Create a scatter plot

ggplot(monthly_data_by_age, aes(x = predicted_log_rf, y = residuals_log_rf)) +
  geom_point(color = "blue", size = 1.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residual Analysis of Monthly Total Movements with Age Groups",
       x = "Predicted Log Total Movements",
       y = "Residuals") +
  theme_minimal()
```

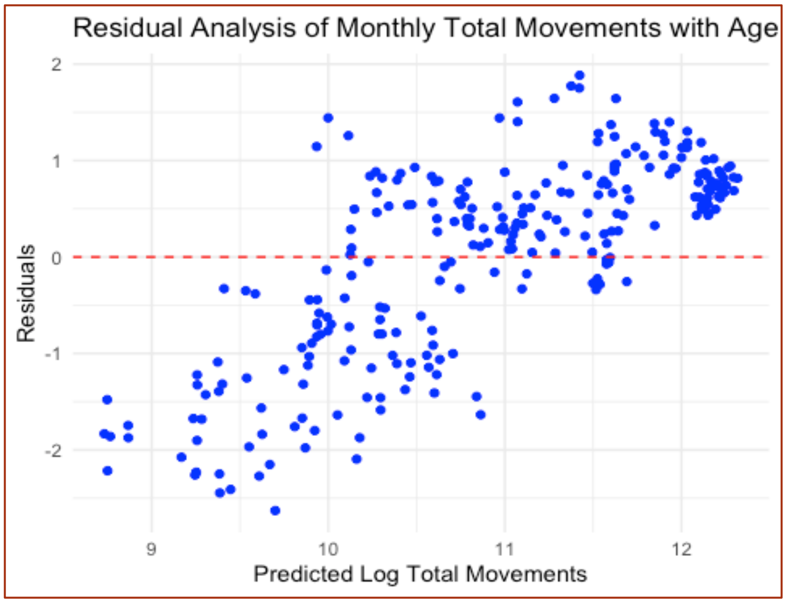

Figure 13, Resudual analysis of monthly total movements with age

The scatter plot shows that the model underestimates larger values, leading to more prediction errors when travel numbers are higher. The red dashed line serves as a reference, highlighting how far off the predictions are across different levels. The trend in residuals suggests that certain features, like seasonality, affect the model's accuracy and may need further adjustments.

```r
# Create a boxplot

ggplot(monthly_data_by_age, aes(x = season, y = residuals_log_rf)) +
  geom_boxplot() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed", size = 1) +
  labs(title = "Residuals by Season",
       x = "Season",
       y = "Residuals") +
  theme_minimal()
```

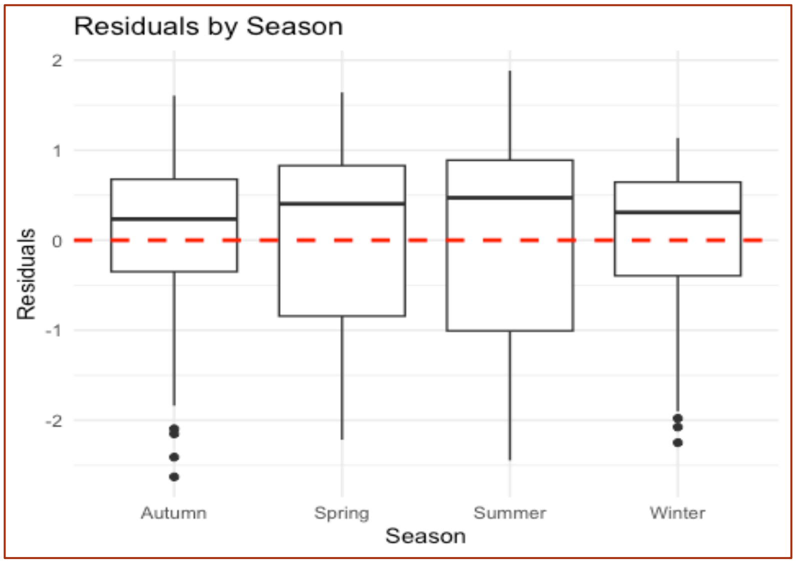

Figure 14, Residual analysis of monthly total movements with age

Summer and Spring show greater variability in prediction accuracy, as reflected in the wide range of residuals. This suggests that travel patterns in these seasons are more unpredictable and harder for the model to capture accurately. On the other hand, Winter and Autumn have negative outliers, indicating the model tends to overestimate travel movements during these periods.

Travel patterns before and during the pandemic were highly irregular due to restrictions and uncertainties. Since the world is transitioning into more stable travel behavior post-pandemic, focusing on data from the "After COVID-19" period (e.g., after January 2022) may provide a clearer and more relevant basis for predicting future movements.

Key Findings

The pre-pandemic growth in visitor numbers highlights New Zealand's strong global appeal as a tourist destination. However, the severe disruptions caused by COVID-19 led to a sharp decline in travel. The recovery observed in 2022 and beyond shows a resilient tourism sector, which bounced back to 81.15% of pre-pandemic levels in 2023, with 44.45% growth seen by mid-2024. This recovery can be attributed to strategic responses and adaptive measures across the industry, positioning the sector for continued growth.

Strategic Insights from Data

Our analysis, using tools like R and Tableau, uncovered key insights into demographic shifts and seasonal travel patterns. Summer and winter have the highest travel volumes, but also the greatest variability in predictions, indicating that tourism during these seasons is harder to forecast. Travel projects could be introduced to boost tourism in off-seasons like spring and autumn, optimizing resource allocation and marketing strategies.

Future-Focused Strategies

Moving forward, the tourism industry should focus on seasonality trends, investing more effort in promoting travel during high-demand periods like summer and winter. Additionally, creating projects to attract visitors during off-peak seasons will help stabilize travel flows year-round and maximize tourism potential.

Collaboration for Continuous Improvement

Collaboration between tourism operators, government agencies, and researchers is key to ensuring the industry's sustained recovery and growth. By working together and embracing innovation, New Zealand can adapt to the evolving preferences of international visitors and continue to thrive as a leading travel destination.

Final Thought

By embracing change and innovation, New Zealand can further enhance its reputation as a premier global destination, attract a diverse range of visitors, and achieve a faster recovery from the pandemic-induced decline in tourism.