MONASH University

# FIT5196 Data Wrangling

# Assessment 1:

# Exploratory Data Analysis (EDA)

| | |
|---|---|
| Group Members: | Adrian Leong Tat Wei (27030768) |
| | Jun Yuan (35833645) |
| | Low Xuan Nan (35373849) |
| Group Number: | 35 |
| Semester: | Semester 2 2025 |
| Submission Date: | 15th September 2025 |
| Tutor Name: | Sailaja Rajanala |

# Table of Content

# 1. Introduction

Exploratory Data Analysis helps us look at a dataset and find basic trends, patterns, and possible errors (IBM, n.d.). In this report we do EDA on photo metadata from Flickr. The data has fields like User_ID, Title, Country and others. Our goals are simple. We check the structure of the table. We review data quality, including missing values and types. We look for useful patterns. From these findings we write machine learning questions. Later we can test these questions with simple models for prediction and analysis.

# 2. Design of EDA

The design of the EDA aimed to provide a scientific, systematic and reproducible approach to understand the dataset, evaluate its quality, and uncover meaningful patterns and trends. The objectives of the EDA are to assess the data quality, examine the distribution and relationships between attributes, and generate insights that can foster the formulation of machine learning research questions. The EDA follows a methodical and logical sequence.

First, we load and merge the xml and json files to inspect the data structure and schema. Next, we conduct data cleaning via regular expressions to remove XML and JSON tags, emojis, and non-English letters. We then convert missing values to np.NaN, as well as parse the data into the appropriate data types. In the data quality assessment step, missing, invalid, duplicates are identified and evaluated to determine the reliability of the dataset. The main 3 types of EDA are then carried out: Univariate, to explore single-attribute distributions; Bivariate, to compare relationships between pairs of attributes; and Multivariate, to examine complex patterns. Finally, the results from these analyses were examined to extract findings and insights that empower the machine learning research questions.

# 3. Findings from EDA

## 3.1 Finding 1: Title coverage

Titles appear for about 96% of posts. Descriptions are about 45%. Tags are about 66%. Title is the most reliable text field.

## 3.2 Finding 2: Text length

The average title length is about 27 characters. The average description length is about 176 characters. Many titles are very short. Some descriptions are very long.

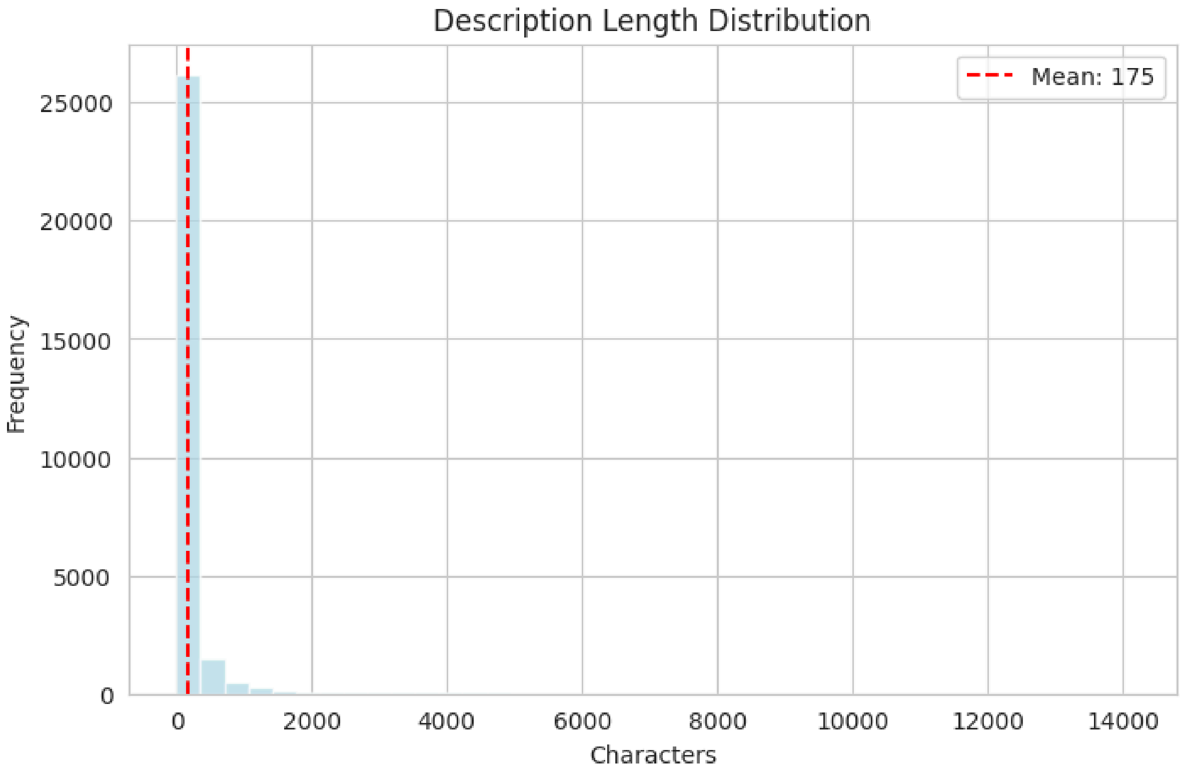


Figure 1: Description Length Distribution.

### 3.3 Finding 3: Common tags

The most frequent tags include australia landscape melbourne nature and nsw. These top tags cover a large share of posts. Many other tags are rare.

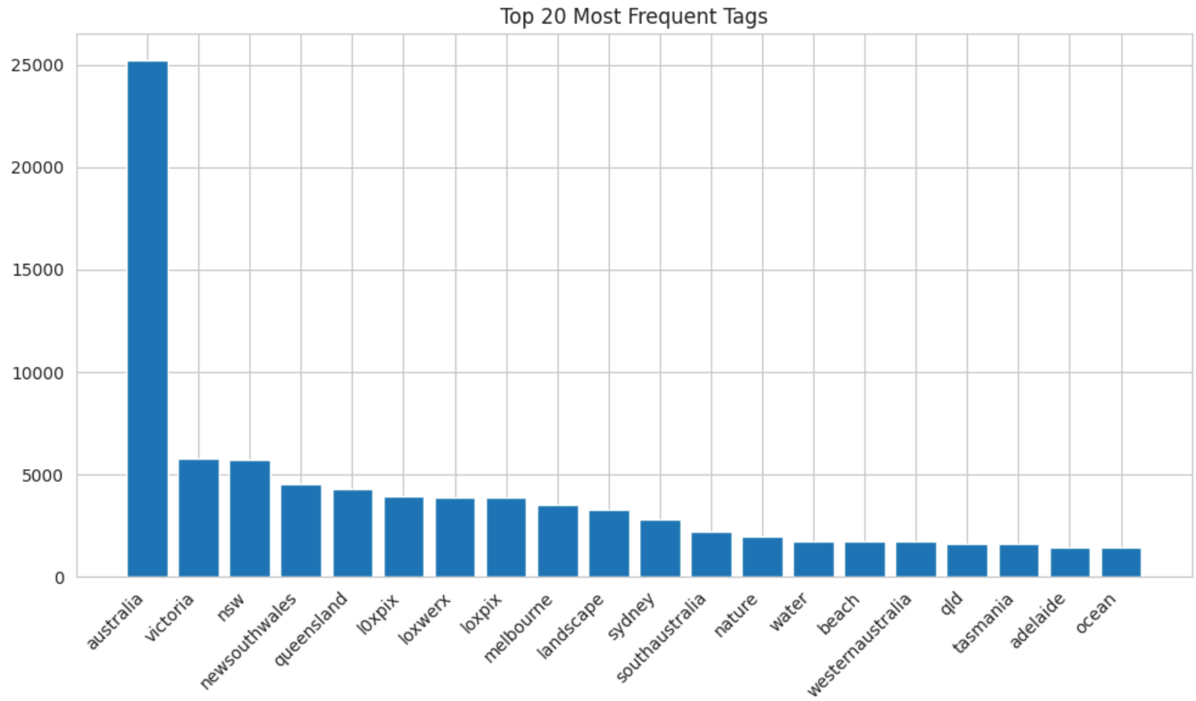
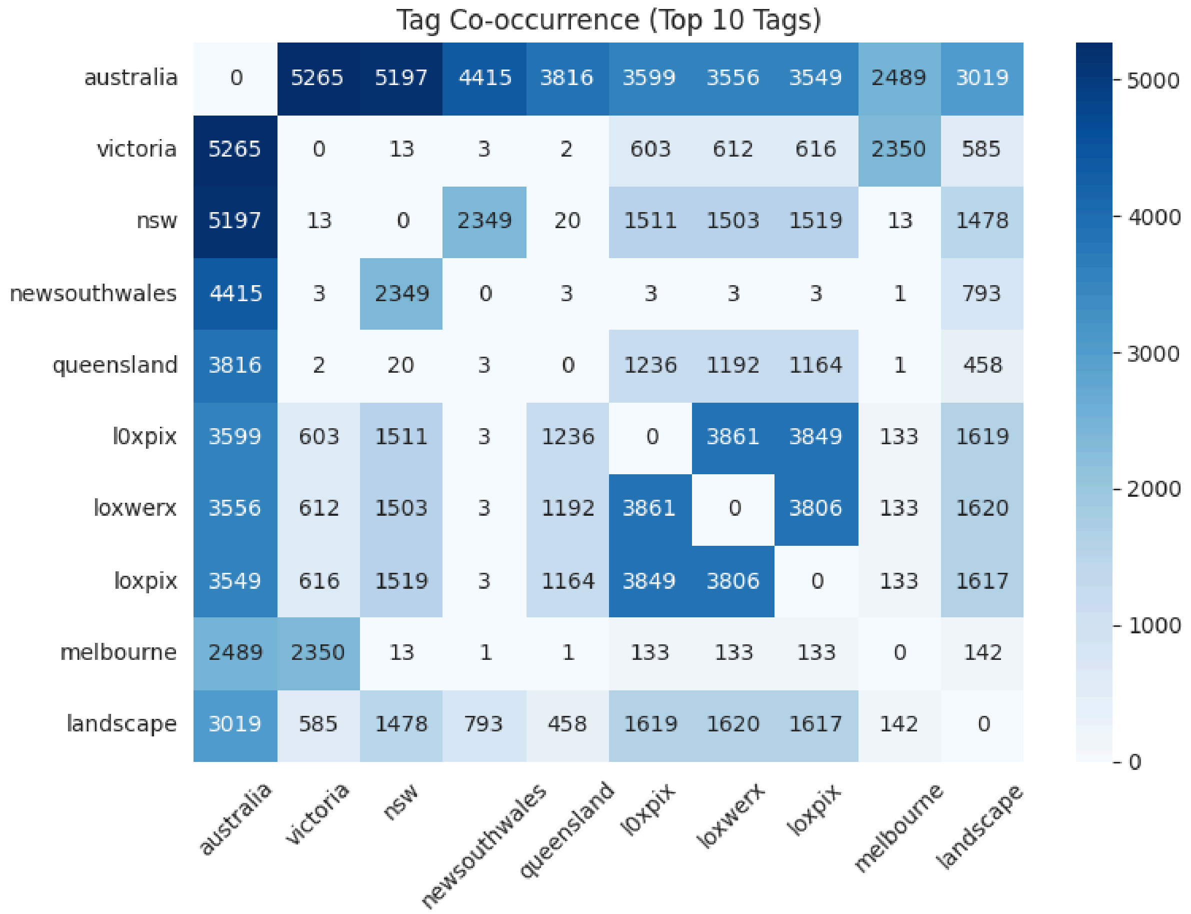


Figure 2: Top 20 Most Frequent Tags.

### 3.4 Finding 4: Tag co occurrence

Place tags appear together with nature tags many times. For example australia or melbourne often come with landscape water or sky. Users mix place and theme when they label photos.



Tag Co-occurrence (Top 10 Tags)

| | australia | victoria | nsw | newsouthwales | queensland | l0xpix | loxwerx | loxpix | melbourne | landscape |
|---|---|---|---|---|---|---|---|---|---|---|
| australia | 0 | 5265 | 5197 | 4415 | 3816 | 3599 | 3556 | 3549 | 2489 | 3019 |
| victoria | 5265 | 0 | 13 | 3 | 2 | 603 | 612 | 616 | 2350 | 585 |
| nsw | 5197 | 13 | 0 | 2349 | 20 | 1511 | 1503 | 1519 | 13 | 1478 |
| newsouthwales | 4415 | 3 | 2349 | 0 | 3 | 3 | 3 | 3 | 1 | 793 |
| queensland | 3816 | 2 | 20 | 3 | 0 | 1236 | 1192 | 1164 | 1 | 458 |
| l0xpix | 3599 | 603 | 1511 | 3 | 1236 | 0 | 3861 | 3849 | 133 | 1619 |
| loxwerx | 3556 | 612 | 1503 | 3 | 1192 | 3861 | 0 | 3806 | 133 | 1620 |
| loxpix | 3549 | 616 | 1519 | 3 | 1164 | 3849 | 3806 | 0 | 133 | 1617 |
| melbourne | 2489 | 2350 | 13 | 1 | 1 | 133 | 133 | 133 | 0 | 142 |
| landscape | 3019 | 585 | 1478 | 793 | 458 | 1619 | 1620 | 1617 | 142 | 0 |

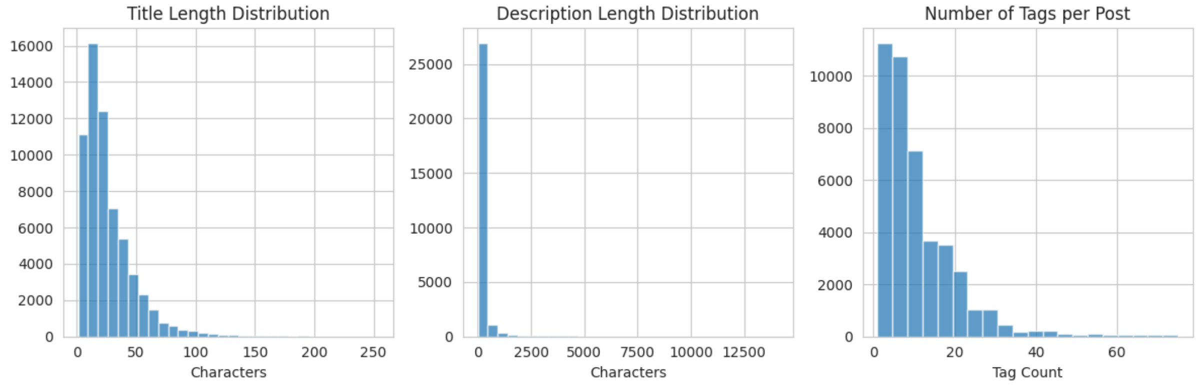
Figure 3: Top 10 Co-occurring Tags.

### 3.5 Finding 5: Cleaning quality

After cleaning there are fewer strange symbols and mixed language parts. Tokenised tags have fewer empty values. Counting and visualising become more stable.

### 3.6 Finding 6: Tags per post

A post has on average about 11 tags. Some posts have many tags and some have none. The distribution is uneven.



Title Length Distribution

Description Length Distribution

Number of Tags per Post

Figure 4: Title Length Distribution, Description Length Distribution, and Number of Tags per Post.

### 3.7 Finding 7: Text and tags overlap

Title plus tags appear together for a large number of posts. This gives a solid base for text based analysis when description is missing.

### 3.8 Finding 8: Valid, invalid and multi country values

Only 50.665% of the values in the Country column contain valid country names, while 49.312% with invalid country names and 0.023% are multivalued.

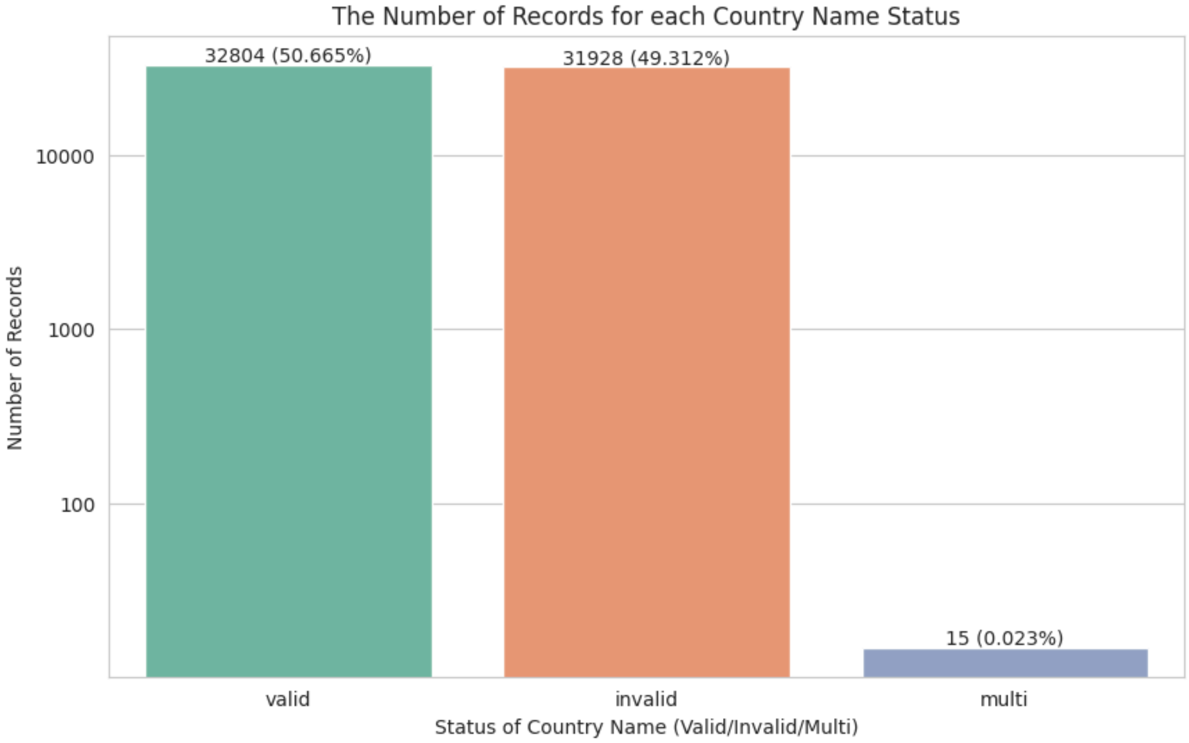


Figure 5: The Number of Records for each Country Name Status (Valid, Invalid, Multi).

### 3.9 Finding 9: Highest country value count

Australia accounts for the largest share of records in the Country column, representing 48.1% of the dataset.

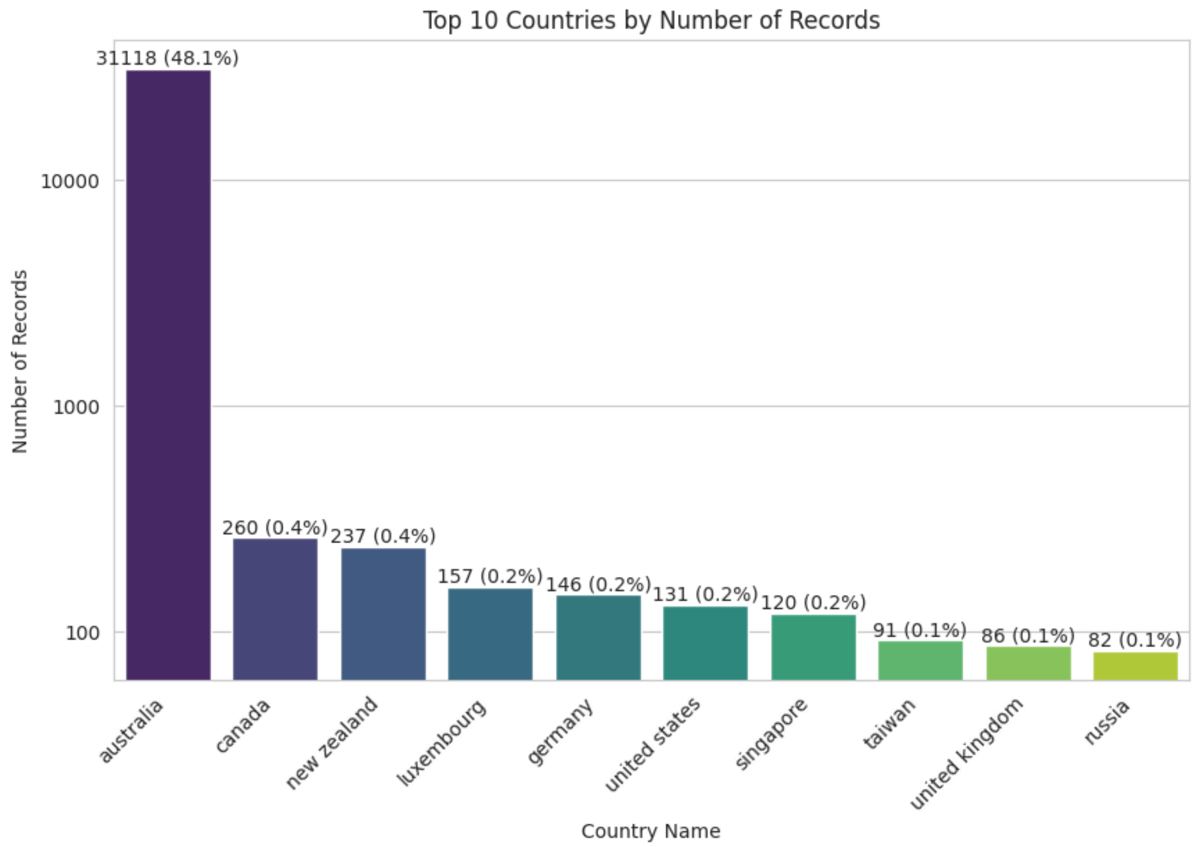


Figure 6: Top 10 Countries by Number of Records.

### 3.10 Finding 10: Coordinates plot with Longitude and Latitude

Scatter map of longitude and latitude shows that the dataset only contains coordinates from Australia.

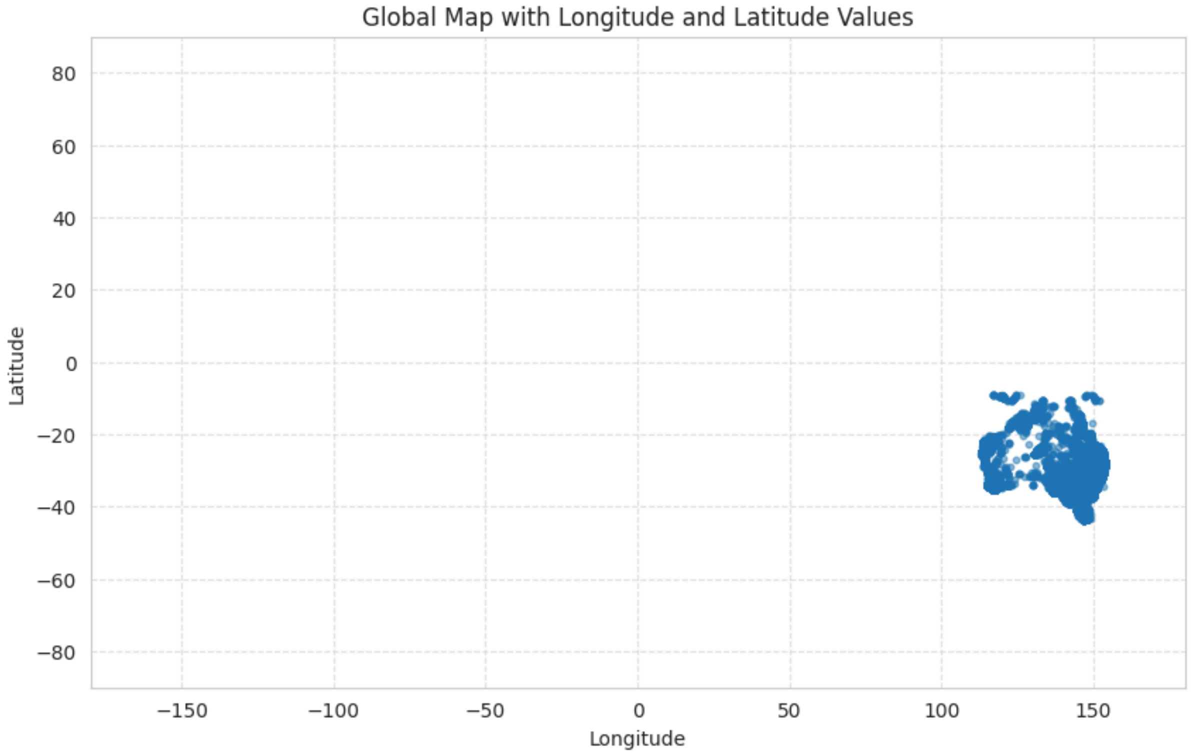


Figure 7: Global Map with Longitude and Latitude Values.

### 3.11 Finding 11: Inconsistent temporal data

The data collection for temporal data is inconsistent as photo posting records exist as early as 2007, while photo-taking records only appear from 2017 onwards.

### 3.12 Finding 12: Highest photo activity count

The number of photos taken, posted, and the minimum photos taken all peaked in 2019, with 14,628, 13,659, and 14,630 photos respectively.

### 3.13 Finding 13: Abnormal photo activity count

Photos taken plummeted to only 6 photos in 2024. Photos posted remained less than 10 photos annually during 2007 to 2017, then surged in 2018 and 2019 with over 12,000 posts annually.

### 3.14 Finding 14: Min_Taken_Date and Taken_Date have similar trend

The Min_Taken_Date distribution mirrors the Taken_Date, with a peak in 2019 and a drop in 2020-2022, followed by recovery in 2023.

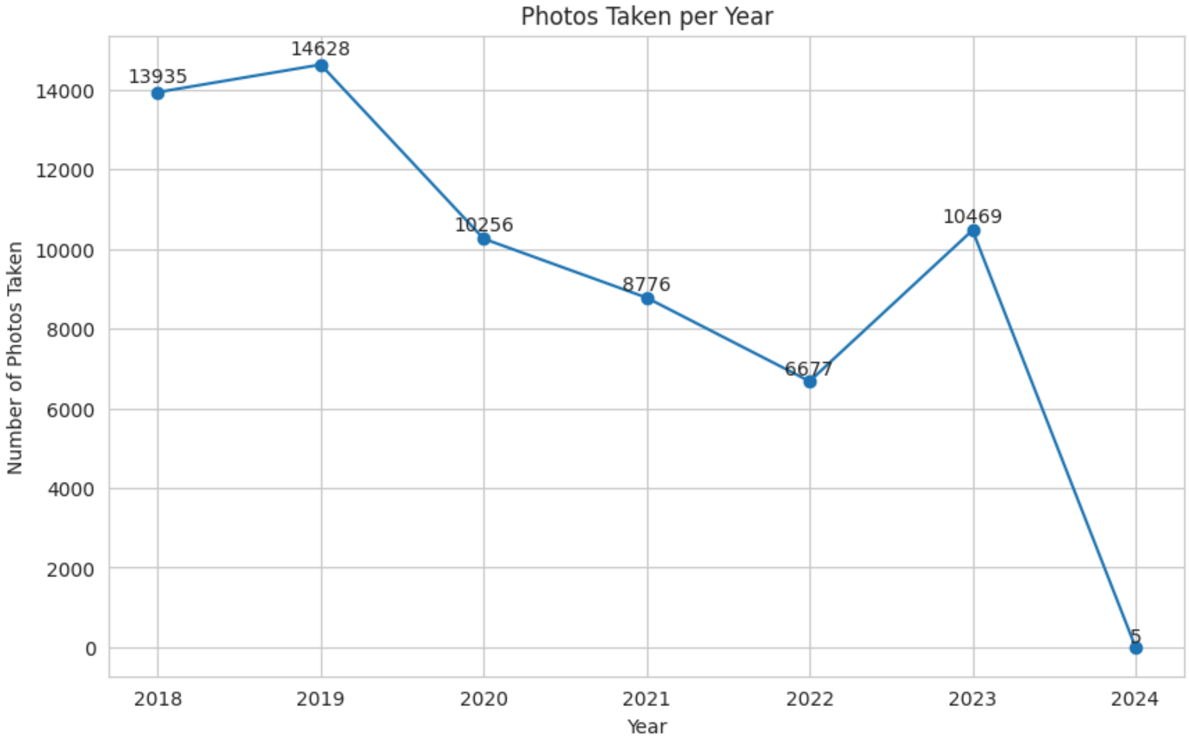


Figure 8: Photos Taken per Year.

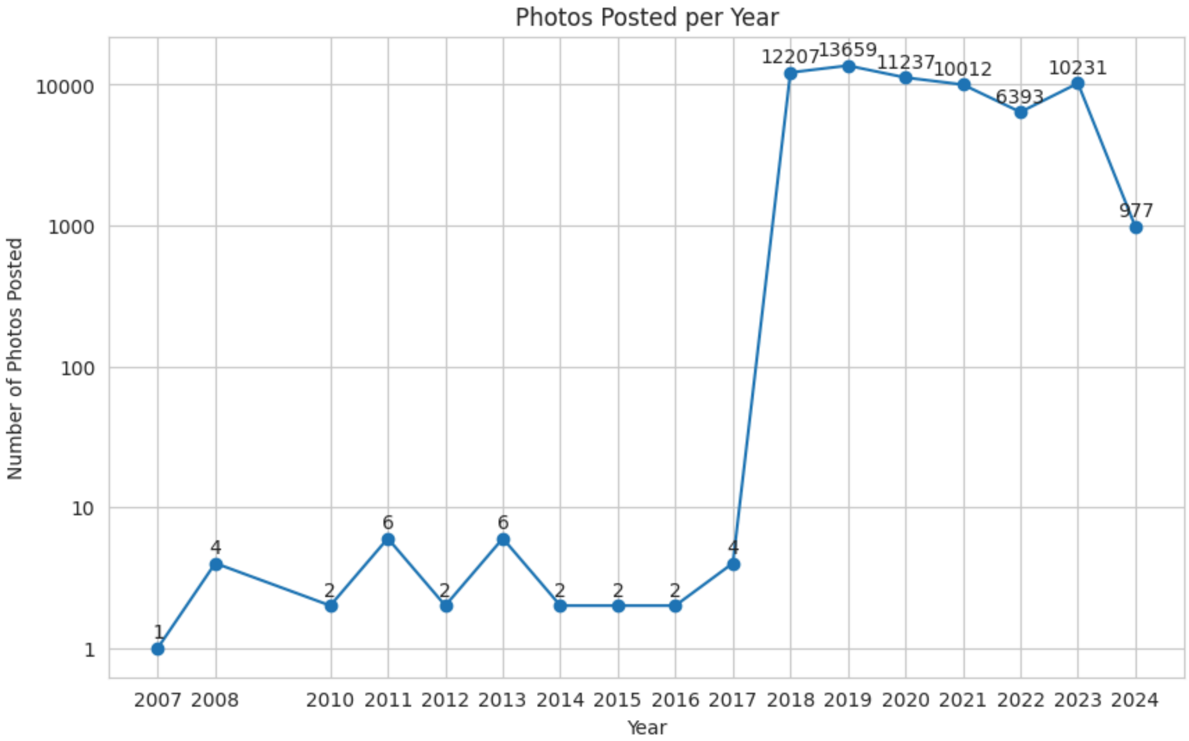


Figure 9: Photos Posted per Year.

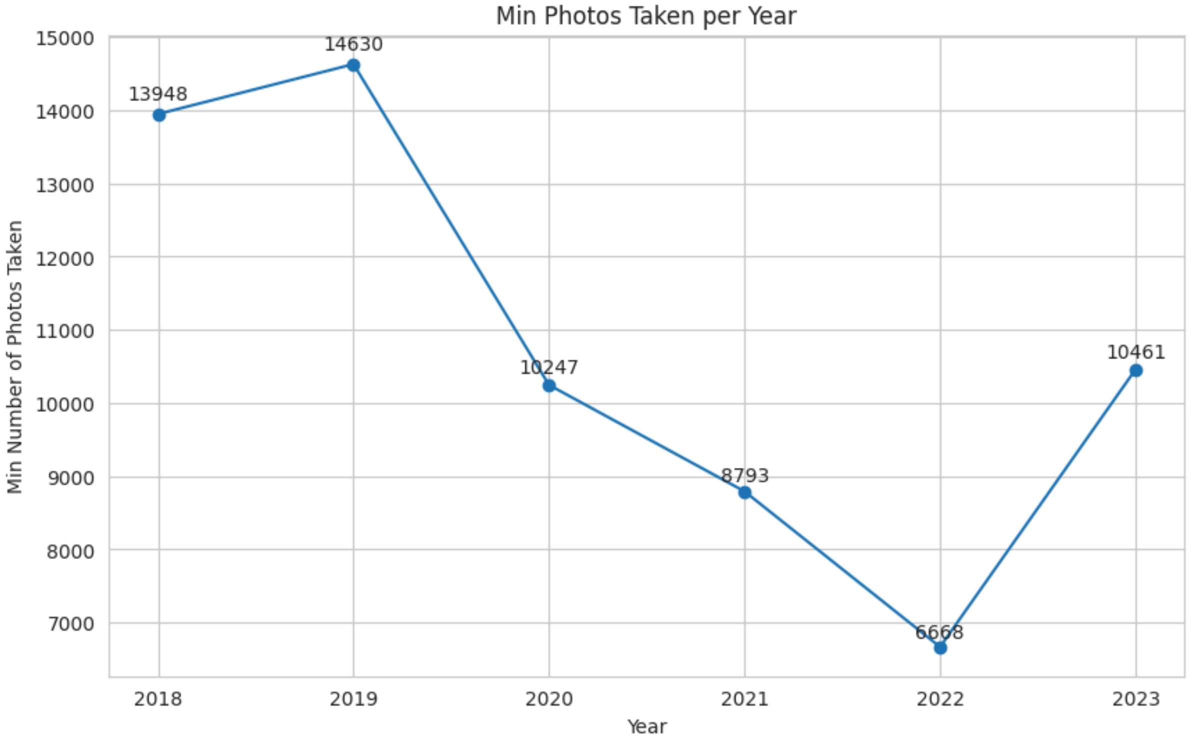


Figure 10: Minimum Photos Taken per Year.

### 3.15 Finding 15: Photo taking activity trend in 2023

Photos taken in 2023 peaked during May (~ 1090) and showed another smaller rise in September (~ 980), while the lowest point was in June (~ 700).

### 3.16 Finding 16: Photo posting activity trend in 2023

Photos posted in 2023 peaked during May (~ 1030) but gradually declined in the second half of the year with a slight increase in October(~ 800 - 930 per month).

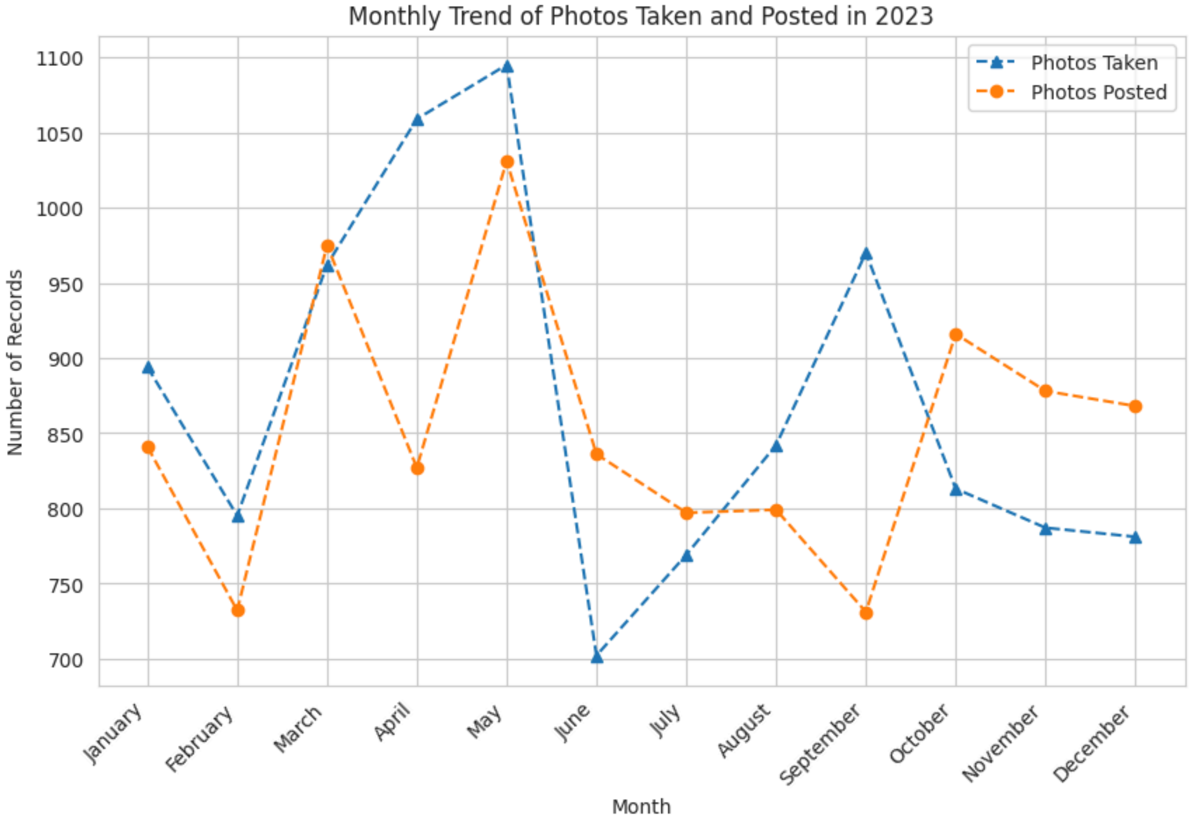


Figure 11: Monthly Trend of Photos Taken and Posted in 2023.

3.17 Finding 17: Hotspots for photo activity in 2023

The highest concentration for photo activity in 2023 is clustered around major cities such as Sydney, Melbourne, and Brisbane, with secondary hotspots visible in Adelaide and Perth.

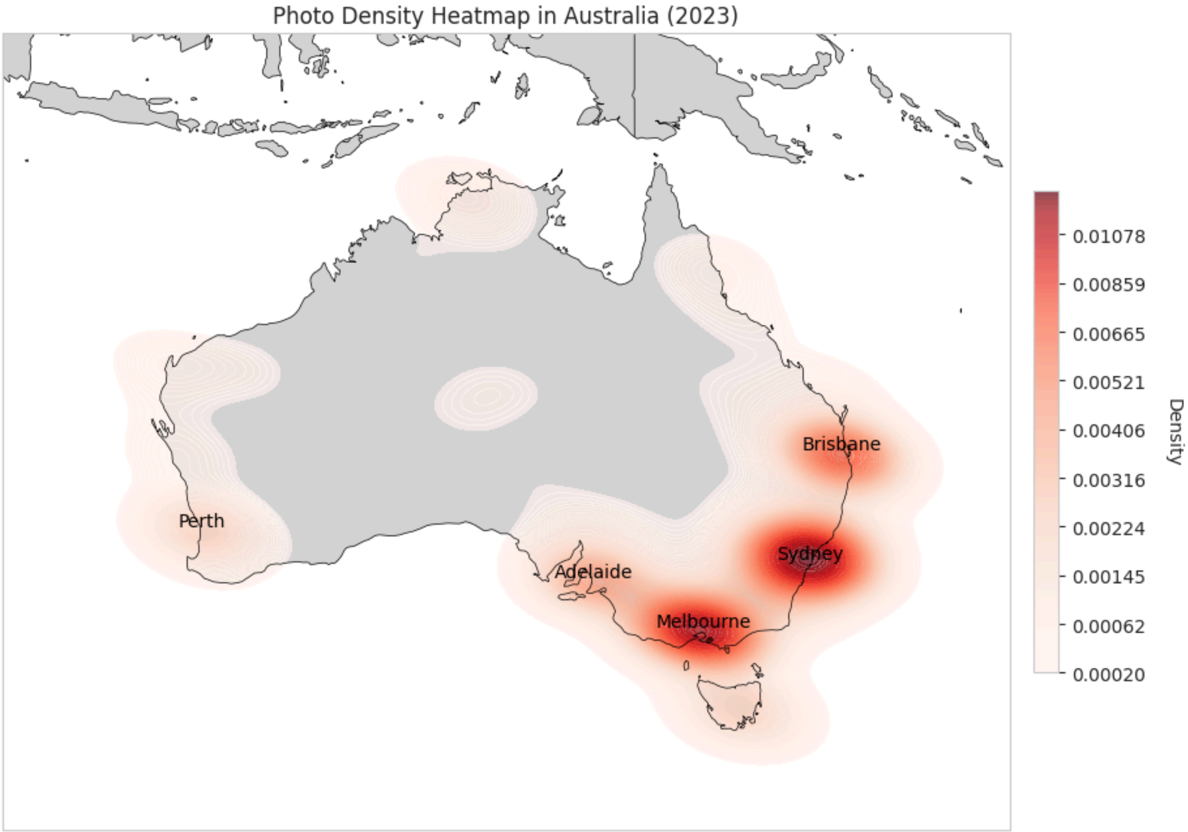


Figure 12: Photo Density Heatmap in Australia (2023).

### 3.18 Finding 18: User post distribution

The distribution of the number of posts users make follows a log distribution. The vast majority of users make very few posts, but the users that do post a lot, post an incredible amount.

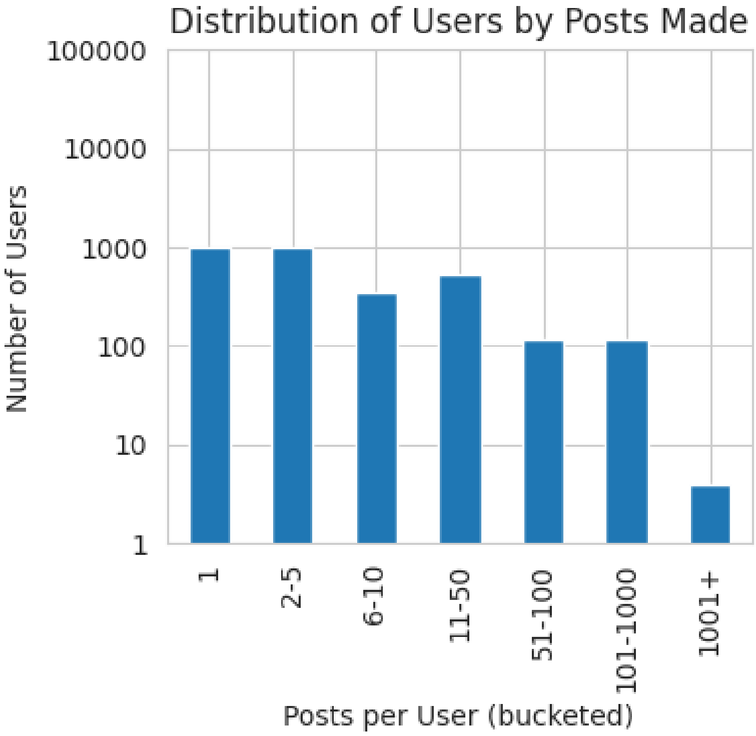


Figure 13: Distribution of Users by Posts Made.

### 3.19 Finding 19: Server distribution

The server distribution is reasonably well distributed, but there are some servers with quite little activity on them. The farm distribution, which presumably is a network of servers, is much less well balanced; almost all of the activity is allocated to farm 66 in particular.

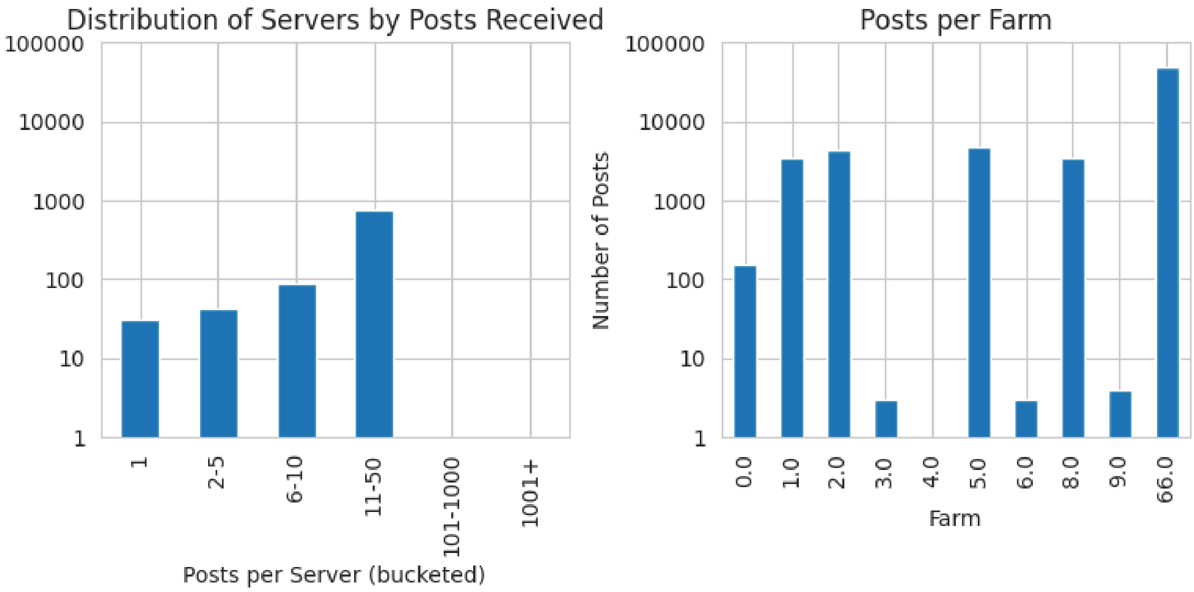


Figure 14: Distribution of Servers by Posts Received and Posts per Farm.

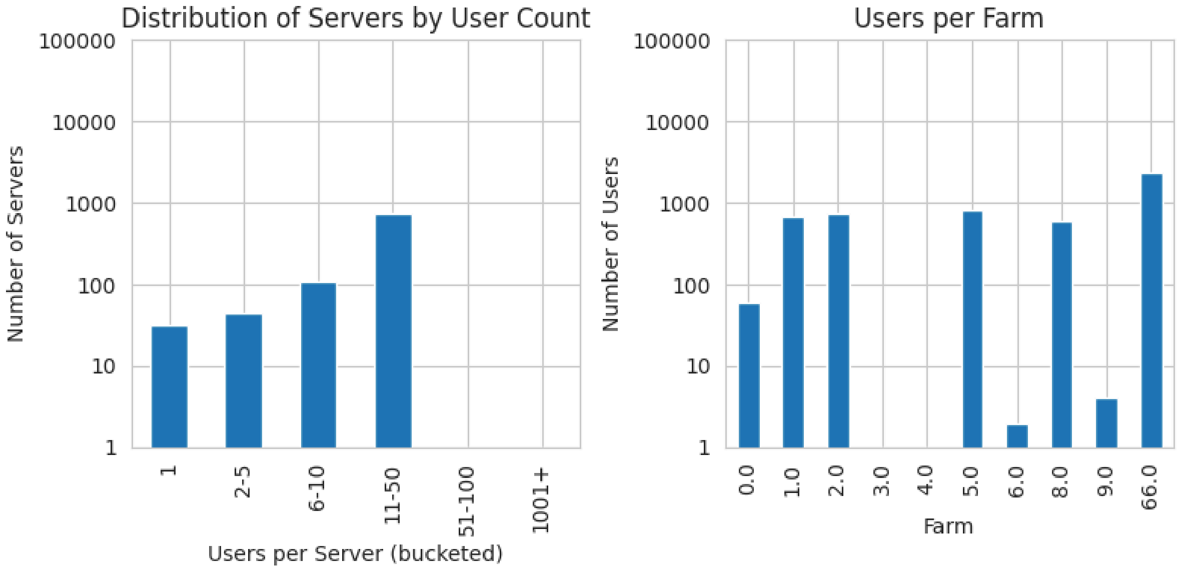


Figure 15: Distribution of Servers by User Count and Users per Farm.

### 3.20 Finding 20: User movement

User movement can be identified by identifying user post location and dates. We find that a majority of movement for holiday is from major cities into the countryside, while a majority of long-term migration is from the countryside into major cities.

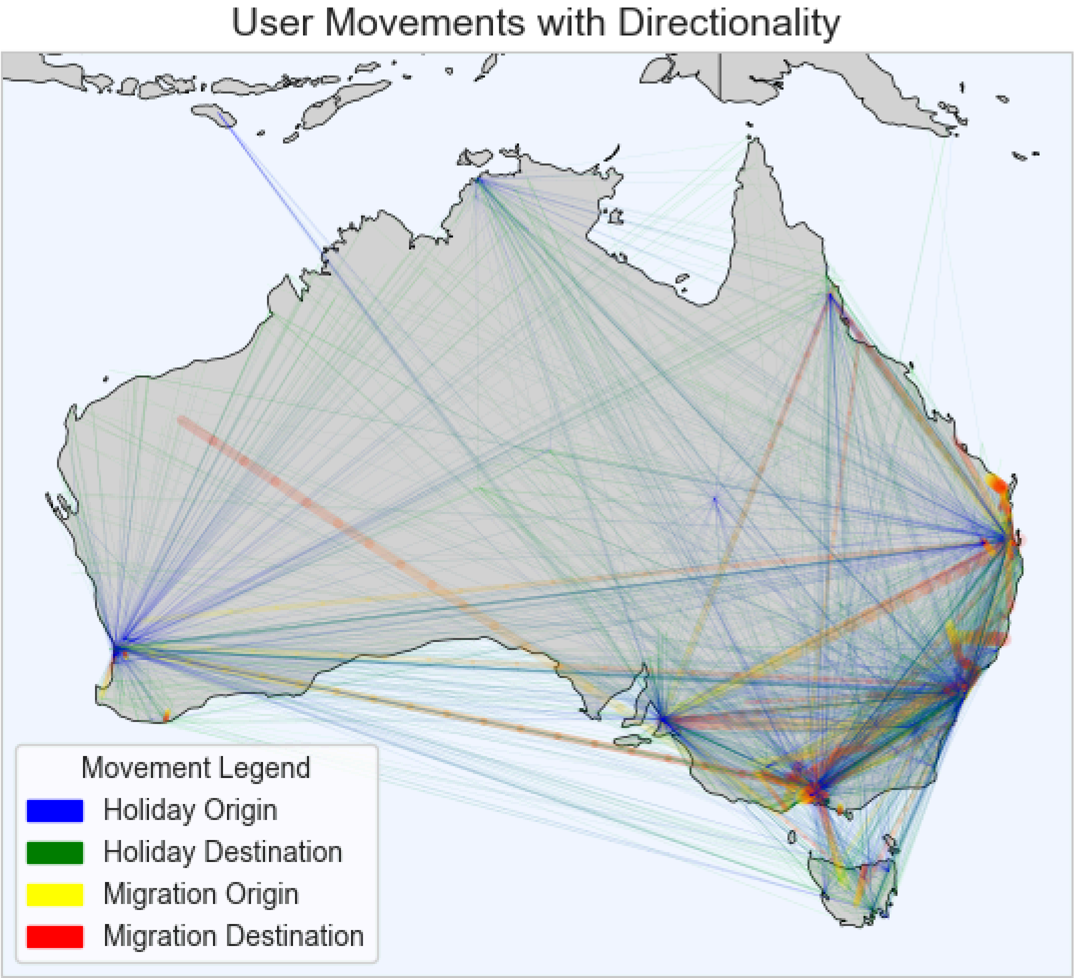


Figure 16: User Movements with Directionality.

3.21 Finding 21: Duplicate posts

Out of 70000 posts, there were 5253 duplicate posts - identical fields in everything except only a minor difference in date taken. There may be some bugs in the app somewhere that causes/allows these duplicate data entries.

3.22 Finding 22: Limitations of data privacy

A full 100% of the 70000 post dataset were only viewable to the public, and not restricted to friends or family. This is understandable as it would violate data privacy, but it also means that no analysis can be done on user behaviour towards data privacy.

# 4. Key Insights and Machine Learning Research Questions

## 4.1 Key Insights

4.1.1 Key Insight 1: Stable title field

The title has the highest coverage and steady quality. Use it first for text features. Linked to Finding 1.

4.1.2 Key Insight 2: Uneven description field

Description is missing more often and length varies a lot. Handle missing and very long text. Linked to Findings 1 and 2.

#### 4.1.3 Key Insight 3: Long tail in tags

A few tags are very common and many are rare. Group or smooth rare tags before modelling. Linked to Finding 3.

#### 4.1.4 Key Insight 4: Place with nature

Place tags often appear with nature tags. Add a light place signal to topic features. Linked to Finding 4.

#### 4.1.5 Key Insight 5: Title plus tags are enough

Many posts have both title and tags. We can run text tasks when the description is missing. Linked to Finding 7.

#### 4.1.6 Key Insight 6: Cleaning improves stability

Cleaning reduces strange symbols and mixed language. Counts and charts are more stable. Linked to Finding 5.

#### 4.1.7 Key Insight 7: Two user writing styles

Some users write very short titles and some write long descriptions. Features should respect this gap. Linked to Finding 2.

#### 4.1.8 Key Insight 8: Uneven tags per post

Some posts have many tags and some have none. Normalise by tag count or use simple weights. Linked to Finding 6.

#### 4.1.9 Key Insight 9: Head tags show place focus

Top tags are often regions and cities. Geo aware features help retrieval and clustering. Linked to Finding 3.

#### 4.1.10 Key Insight 10: Join text with time and place

Combining text with month and location gives clearer themes and better search. Linked to Findings 3 and 4.

#### 4.1.11 Key Insight 11: Unreliable Country data

The mismatch between the Country column with Longitude and Latitude column highlights the need to rely on coordinates over country field.

#### 4.1.12 Key Insight 12: Inconsistent temporal records

The mismatch of year range in temporal records suggested either incomplete metadata or selective collection. Analyses involving long-term trends should therefore focus on the overlapping window from 2018 to 2023.

#### 4.1.13 Key Insight 13: Change in platform usage or data collection

The sharp rise in 2018 for photo posting activity, suggested either change in platform usage or data collection. This aligns with the peak of photo-taking activity, showing consistency between taken and posted patterns.

#### 4.1.14 Key Insight 14: Seasonal effect on photo activities

The sharp rise in photo-taking activity from February to May corresponds with Australia's Autumn season where many major cultural festivals are held (Australia, 2022). In contrast, the steady activity observed from June to September aligns with the Winter season, when landscape photography is popular due to clearer skies and tranquil landmarks (AdamC, 2024).

#### 4.1.15 Key Insight 15: Delayed posting behaviour during vacation

The photo posting taking trend shows that there are some delayed or backlog posting during peak photo taking period in April and September, suggesting that the delayed posting behaviour may occur during trips, festivals, or seasonal events.

4.1.16 Key Insight 16: Population density and tourism activities affect photo taking activity
Photo taking activity may be strongly associated with population density (Australian Bureau of Statistics, 2025) and tourism activities, reflecting user behaviour centered around major cities.

## 4.2 Machine Learning Research Questions

4.2.1 Question 1:  Can we predict post categories from title and description text?
From looking at the tags, I noticed that different types of posts have different tag patterns. Nature photos usually have tags like 'landscape', 'outdoor', 'beach', while city photos have 'city', 'melbourne', 'architecture'. Since most posts have titles (95%) and the tag patterns are quite different, we could probably train a model to predict the category by analyzing the words in titles and descriptions. The TF-IDF method could work well here to find important words for each category.

4.2.2 Question 2: Can we group similar posts together using their tags?
Since each post has about 11 tags on average, there's lots of information to compare posts. When I looked at the data, posts about similar topics tend to share many tags - like landscape photos often have 'nature', 'outdoor', 'australia' together. We could measure how similar two posts are by counting how many tags they share (Jaccard similarity) and then use clustering algorithms like K-means to group them. The co-occurrence analysis already shows that some tags naturally go together, which makes clustering possible.

4.2.3 Question 3: Who are the people posting extraordinary amounts of posts?
When looking at the distribution of posts made by users, I notice that many users don't post a lot, but some users post an incredible volume of posts. Making an incredible volume of posts is a behaviour of interest for a multitude of reasons (For example, a study on students (Li et al., 2024) found that brain rot content, often associated with social media usage, significantly affects student academic anxiety, academic engagement, and mindfulness for the worse), so we want to analyze what type of people they are. Location of these users is an obvious trait to look for, but we can also look at the post categories they post, the language they use, to identify their behaviours.

4.2.4 Question 4: How can we better distribute server load?
The number of users and posts are not very directly correlated with the number of servers and/or farms, some servers and/or farms serve many more people than others. This could be indicative of poor load balancing, which is an infrastructure problem that virtually every company wants to always improve on (if it is feasible to do so). By identifying the noteworthy servers and/or farms, and comparing their locations in proximity to their users, we can identify weak links in the infrastructure where some places may be overburdened, while other places may have too many resources for their value.

4.2.5 Question 5: Can we predict photo taking and posting activity based on season?
The monthly trend showed fluctuations during seasons, highlighting photo taking and posting activities peak in May and dropping significantly in June. This reflects that seasonality influences user behaviour in both photo taking and posting. Therefore, Months input extracted from Post_Date and Taken_Date are the key predictors to predict the seasonal photo taking and posting activities. For instance, regression models such as Random Forest Regressor could be applied to predict the number of photos being taken or posted throughout the season, and using classification models to categorise months into high or low activity levels. The

prediction may benefit platforms in resource planning to forecast server loads and storage needs by predicting when uploads will spike to avoid performance issues during seasonal surges.

4.2.6 Question 6: Can we identify major hotspots of user activity in Australia during a certain year?
From the photo density heatmap for 2023, it showed high concentration of photo activity around Australia's major cities such as Sydney, Melbourne, and Brisbane, while rural areas exhibited minimal activity. By using geospatial data such as Longitude and Latitude, pairing with Post_Date or Taken_Date allow us to identify the major hotspots of user activity in Australia over the years or a certain year with machine learning techniques such as K-means, XGBoost, etc. This can benefit tourism to target hotspots for event promotion and tourism campaigns in high-activity cities.

4.2.7 Question 7: Can we identify user movement through social media data?
Based on the multivariate analysis using user, location and time data, we can identify where users have been, and by extension, where people have moved around to, and whether they are there for a short holiday or have moved there for long-term stay. This is tremendously helpful not just to tourism marketing and planning, but also for long-term government and urban planning.

# 5. Conclusion

In conclusion, the exploratory data analysis provided a comprehensive view of the dataset across textual, user, temporal, and geospatial dimensions. The analysis of text fields showed that titles are generally reliable, descriptions are uneven, and tags follow a long-tail distribution. Place-related tags often co-occur with nature tags, suggesting semantic structure that can be leveraged for text-based models. Cleaning these fields reduced noise, improved stability in charts and counts, and ensured that each finding could be supported by visual evidence in the notebook.
From the perspective of access and user behaviour, the dataset was confirmed to be entirely public, which aligns with its collection process. Duplicate post IDs were detected and cleaned, improving reliability. User-level analysis revealed a skewed posting distribution, where most users contribute few records while a small subset of highly active users dominate activity. In addition, user movement patterns could be observed, with holidays typically involving travel from major cities into the countryside, and long-term migration in the opposite direction.
Temporal and geospatial analysis highlighted inconsistencies in coverage, with posting records available from 2007 but photo-taking records only from 2017. The most reliable overlapping window was 2018-2023, with photo-taking and posting activity peaking in 2019 before declining sharply. Within 2023, seasonal effects were evident where activity rose in autumn months and dipped in winter, while backlog posting suggested delayed uploads during festivals and trips. Spatially, all records were concentrated in Australia, with major hotspots in Sydney, Melbourne, and Brisbane, and secondary activity in Adelaide and Perth.
Together, these findings enabled the generation of sixteen insights and seven machine learning questions, covering topics such as text-to-topic modelling, tag-based clustering, anomaly detection, seasonal activity forecasting, hotspot identification, and server load planning. While the dataset has limitations including missing descriptions, uneven tag counts, non-content tags, inconsistent country fields, and abnormal activity counts in 2024, the analyses demonstrated that systematic cleaning and focused study of the reliable data window can produce stable, reproducible results. Overall, this EDA lays a strong foundation for subsequent machine learning tasks and provides valuable directions for both research and practical applications.

## 6. References

1. AdamC. (2024, April 19). *Is winter a good time to visit Australia? | ULTIMATE*. Ultimate Adventure Travel. https://www.ultimate.travel/our-blog/is-winter-a-good-time-to-visit-australia/
2. Australia, T. (2022, March 16). *Australia's seasons - Tourism Australia*. Www.australia.com. https://www.australia.com/en-my/facts-and-planning/when-to-go/australias-seasons.html
3. Australian Bureau of Statistics. (2025). *Capital cities continue strong growth*. Australian Bureau of Statistics. https://www.abs.gov.au/media-centre/media-releases/capital-cities-continue-strong-growth
4. IBM. (n.d.). *Exploratory Data Analysis*. Ibm.com. https://www.ibm.com/think/topics/exploratory-data-analysis
5. Li, G., Geng, Y., & Wu, T. (2024). Effects of short-form video app addiction on academic anxiety and academic engagement: The mediating role of mindfulness. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1428813