Customer Ad Campaign

The dataset is a leading Fashion retailer in Australia. MiQ runs a display advertising campaign for this brand, where it shows ads to users leading them to make a purchase on the brand's website.

Ankita Mathur

Steps

01
Data Preprocessing

02
Feature Engineering

03
EDA

04
Further EDA: Product Analysis and Segmentation

Data Preprocessing

Based on quick analysis, I observed that the data was inherently messy and had a lot of noise. Also, we saw few of the database assumptions such as primary key, redundancy and consistency were violated. Let’s see what we observed and how we addressed them.

Data

Size of the data set : (49999, 15)

NA ::

timestamp	0.000000
user_id	0.000000
ip_address	0.000000
product_name	0.000000
product_id	0.000000
is_first_order	0.000000
user_gender	2.348047
payment_type	0.000000
number_of_products	0.000000
order_coupon_code	76.211524
city	0.674013
country_province	8.360167
user_birthday	54.551091
country	0.000000
revenue	0.000000

Issue and Fixes

- **NA** values were present as empty fields as well as **undefined** as a text in many columns.
  - **replaced undefined** with **NA**.
  - **dropped the columns** such as **birthday** (>50% data is missing) and **country province** (not sure how to interpret).
  - **dropped the rows** that has **missing gender** and **city**.
  - **imputed** missing **coupons** with **na** as text.

- I observed that the text fields such as **product description** and **city** has same records with **varying casing**.
  - normalized them to avoid data redundancy and consistency issues.

- I detected **inconsistency** with the **product ids** and the **product names**.
  - Examples will include **a transaction** has **20 product ids** but **10 product names.** We dropped all the rows which had this issue.
  - Further analysis revealed that the assumption of **product id** as **Primary Key** is **flawed**. For example, certain product names are trimmed such as **sammy scallop hem top** was named **sammy scall** and both has the **same id**.
  - We addressed this issue by **normalizing** all the **product names** with the **longest available name under the same id** after creating a **SKU master**. *It is to be noted that two products can have same name but different ids.*

lost ~3.6% of data because of the cleaning so far.

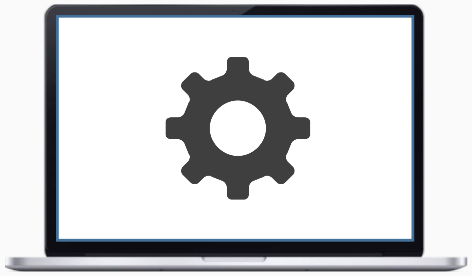
Issue and Fixes

- Next, moved on the validation of user demographics. The **assumption** was that **a particular user** will have an **unique gender** and **a single country**.
    - found a number of cases where this was violated.
    - got rid of the observations.

- Found that **a city belongs to two countries** (AU and NZ). Now, while this can be **perfectly normal**, did some research on some of the observations on Google Maps. failed to get a hit for them in both AU and NZ.
    - So, decided to **normalize** the **country for a city** with the **mode** (whichever country has the highest frequency for such a city, will be chosen and for a tie it will be random).

- Next, came up with **nopayment** as a payment type and each of those transaction had a revenue of 0.
    - Most probably they are **returned items**.
    - ignored and **removed** the occurrences.

- As a result of all the data cleaning steps, have lost ~6.2% data, which is quite acceptable.



Feature Engineering

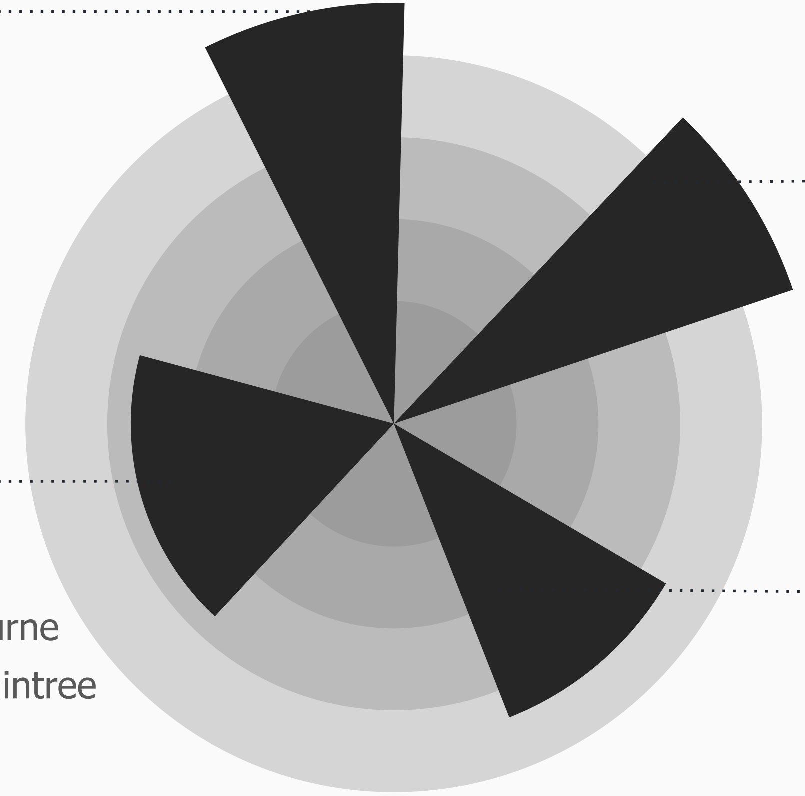
We created some of the features such as **Total Product Count (basket size)**, **Weekday Name**, **Week Number**, **Hours** to facilitate the analysis and finding trends.

Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and generate quantitative analysis.

Frequency

- **Single Purchasers**: 90.58%
- **Multiple Purchasers**: 9.28%
- **First-time Transactions:** 76.36%
- **Frequency of Purchase**: 3.82 days



Overall

- **Total Revenue**: 6310086.73
- **Total SKU (ids) Sold**: 25269
- **Total units sold**: 93671

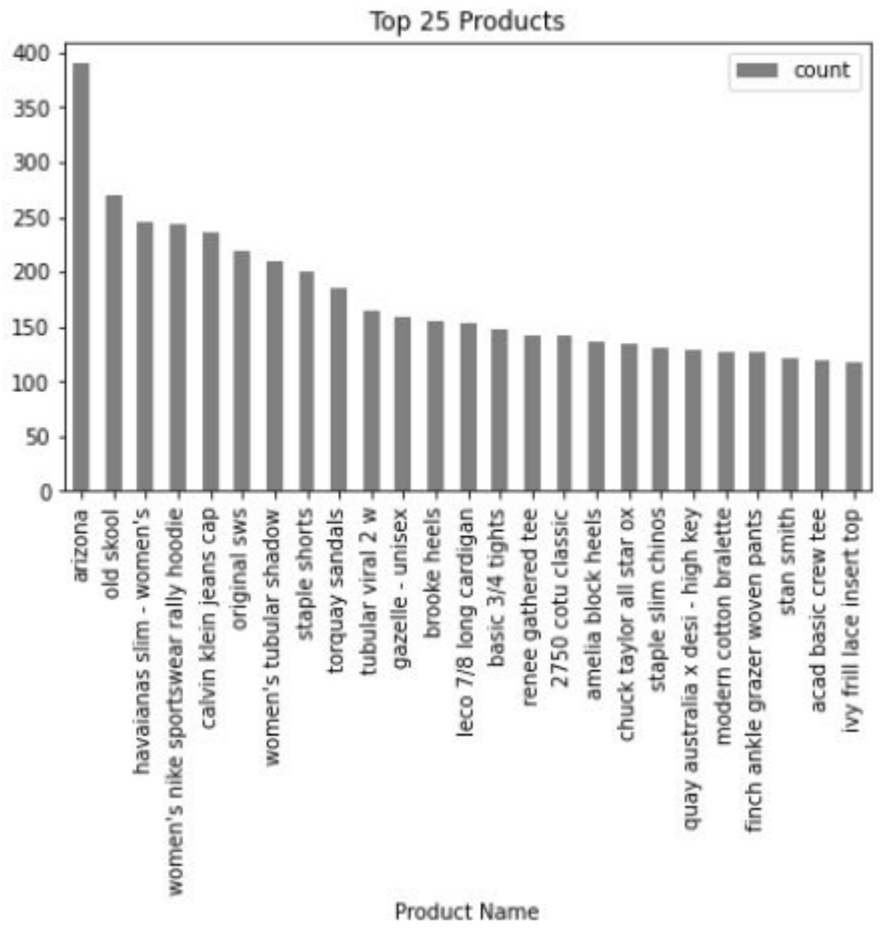
Attributes

- **Payment Type**: cc@braintree
- **Geography**: Sydney, Auckland, Melbourne
- **Payment when coupon applied**: braintree
- **Time of Day**: 8-10 am / 8pm-2am
- **Day of Week** (Quality): Weekends
- **Day of Week** (Quantity): Tuesday/Wednesday

Basket

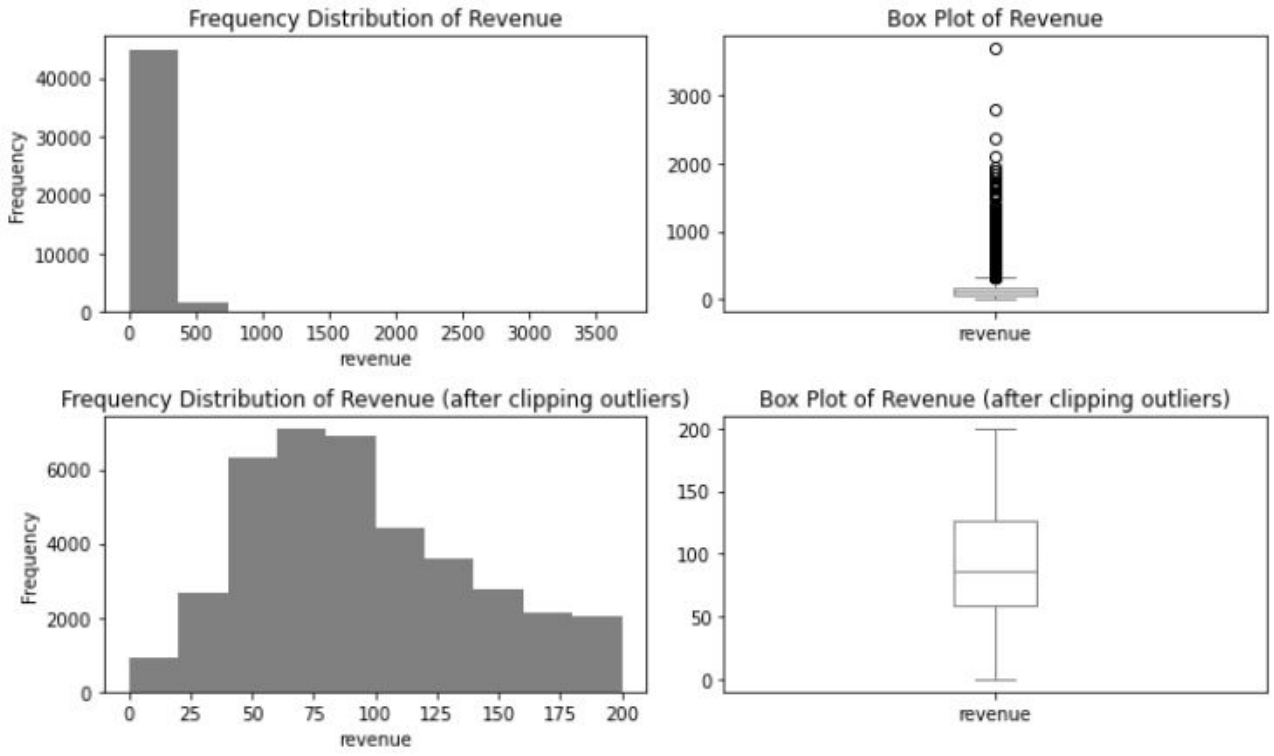
- **Average Unique Quantity**: 2
- **Revenue per Order**: 134.53
- **Revenue per Unit SKU**: 78.24

Products



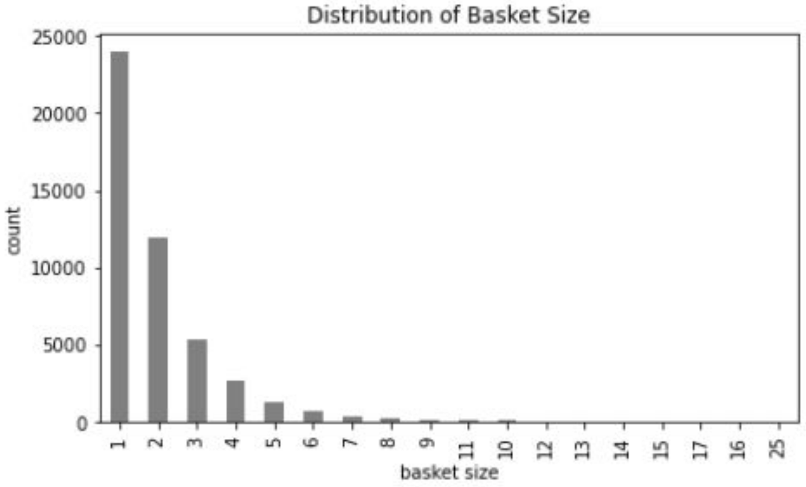
- can see the **top 25 products** sold on the **left**.
- On the **top**, we have the **times** the **same product** was **bought across the weeks**. We see that in **very rare cases**, the **same product was bought more than once**. This makes the **product dimension very sparse** and pose **challenges** in finding the **Product Affinity**.

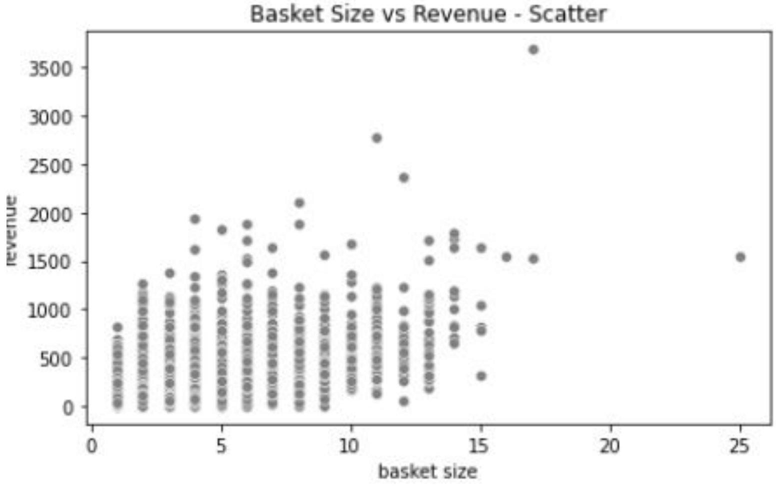
Revenue

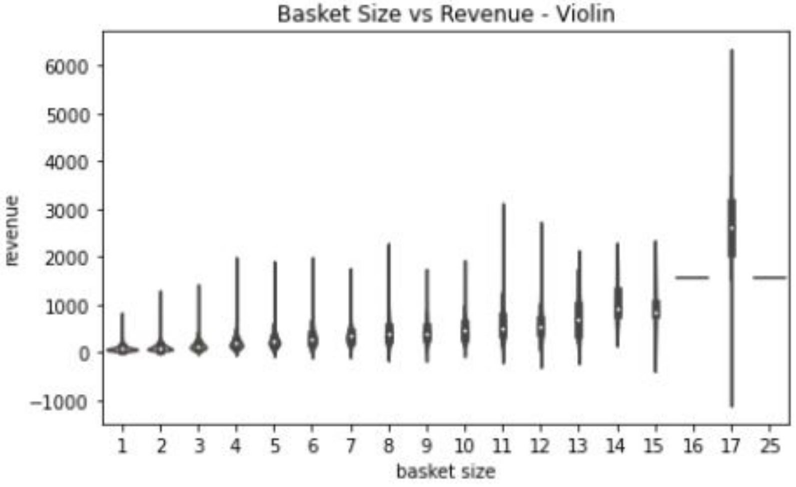


- can see a **skewed distribution** of revenue with lots of **outliers**.
- **won’t treat outliers** because in some cases the users have bought a **large number of SKUs** and it is **acceptable**.
- However, **clipped the revenue below 200** to see how the distribution looks like and it’s kind of **normal distribution**. So, we can loosely say that **sales above 200** is occurring **less probable**, but not unlikely.

Basket Size





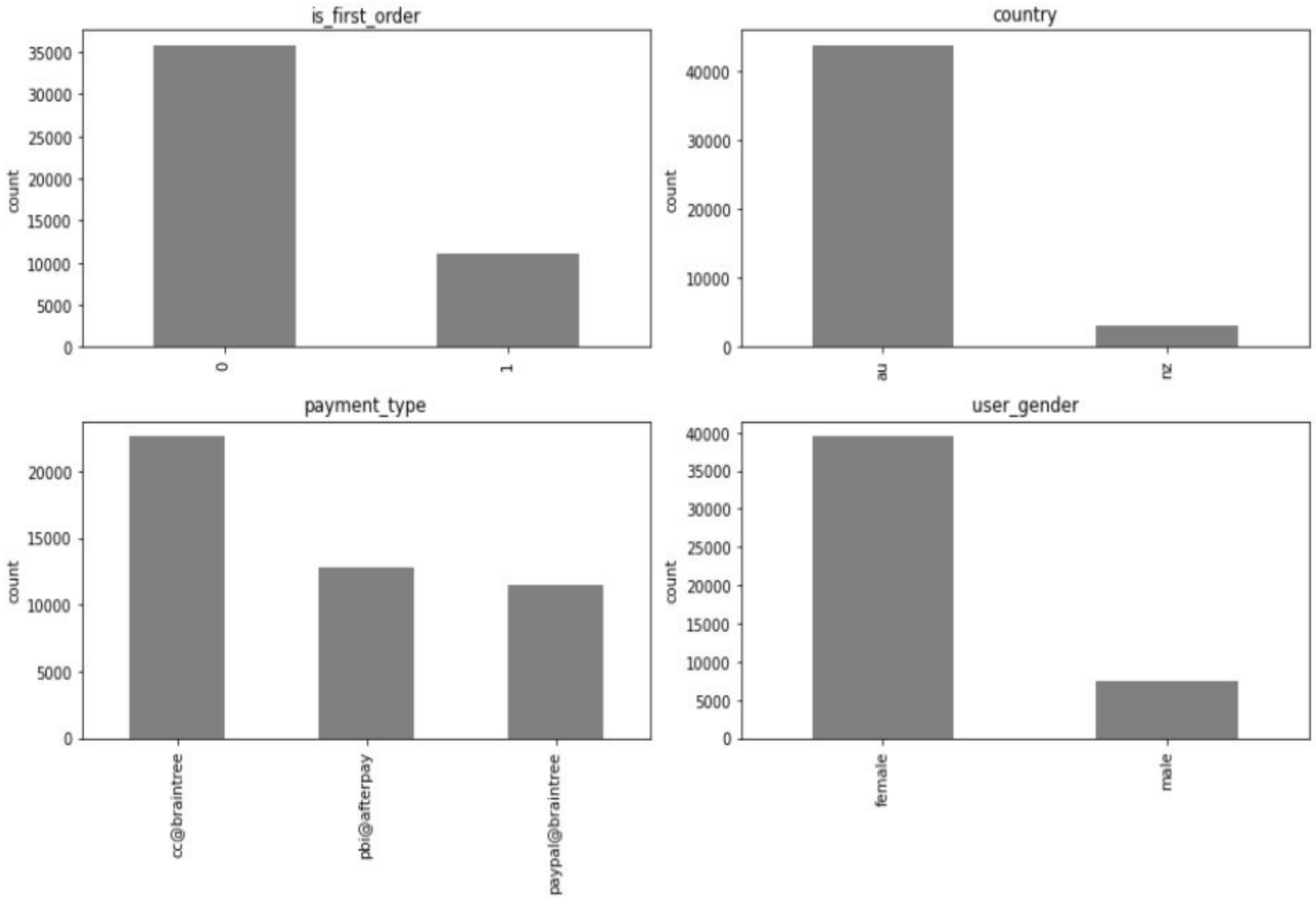


We can observe that the distribution is skewed with the **majority** of the **basket size** limiting to **<6 items**. Thus as discussed before, this will again be a problem for the sparsity of the matrix for finding Product Affinity.

We can see a **loose correlation** that with **increasing basket size**, the actual **revenue is also increasing**. This kind of proves our initial skewed behaviour of revenue and this is acceptable.

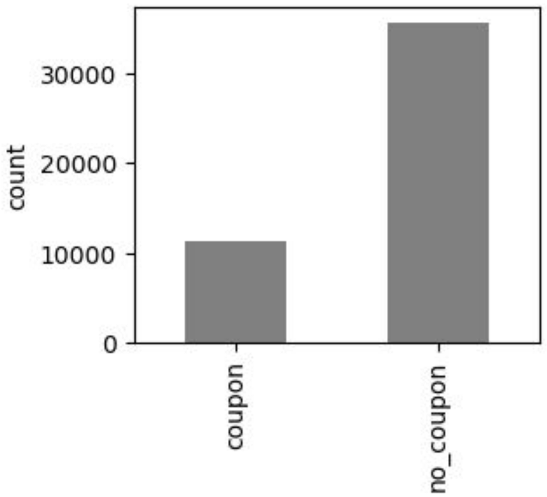
We also wanted to check the actual density of the revenue and the peaks prove that with increasing basket size, the probability of having a greater revenue is more.

Frequencies

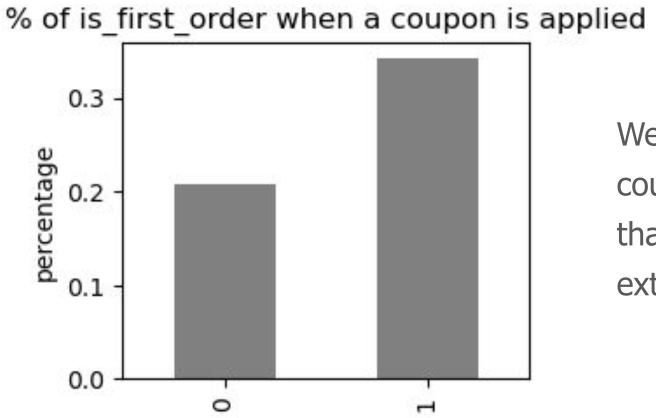


- We see that many of the **customers** have made a **purchase second time**.
  - What are the driving factors for a second purchase?
- Mostly the orders are from **Australia** than **New Zealand**.
  - Does this necessary extend to the **average revenue** earned from both countries?
- **CC Braintree** is the most used payment gateway.
  - Can the brand get into a **partnership** with them and give discounts if items are bought using that platform?
- The products purchased are mainly by women.
  - Does the inventory has less **male items**? How can that be **boosted**?

Coupons



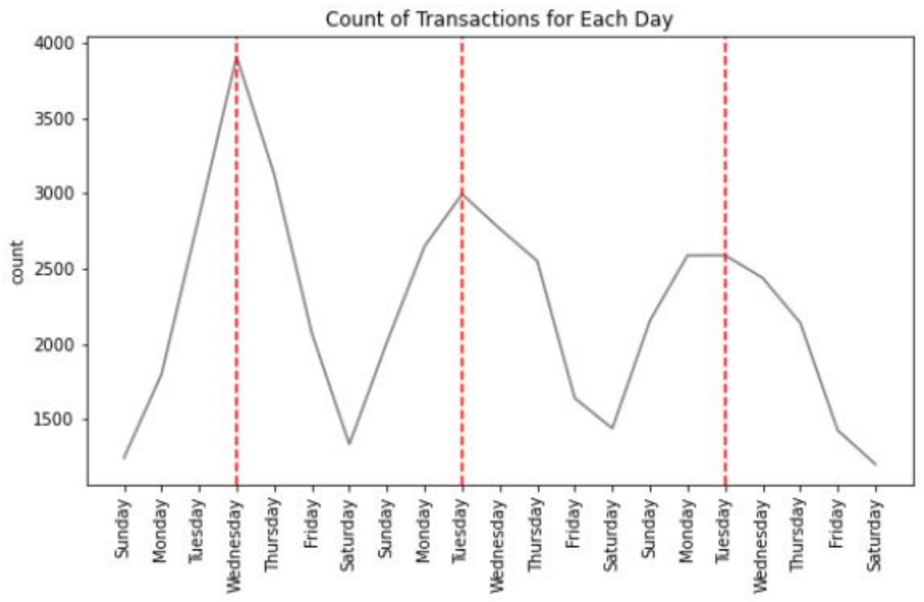
Around **~25%** of the purchases are made using **coupons**

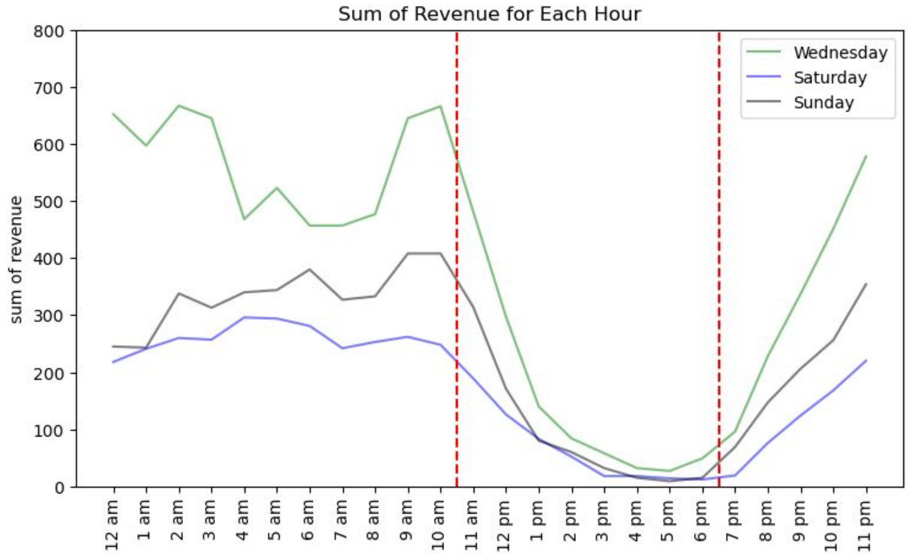


We observe that the ~35% of the repeat customers have applied a coupon code whereas only ~20% of the first timers. This proves that definitely **coupons** are **retaining the customers** and to some extent also playing a role to **attract customers**.

We observe that around ~25% of the transactions made using braintree is done with coupons. So, we can **boost** the relationship with the **payment gateway** - **braintree**.

Number of Hits



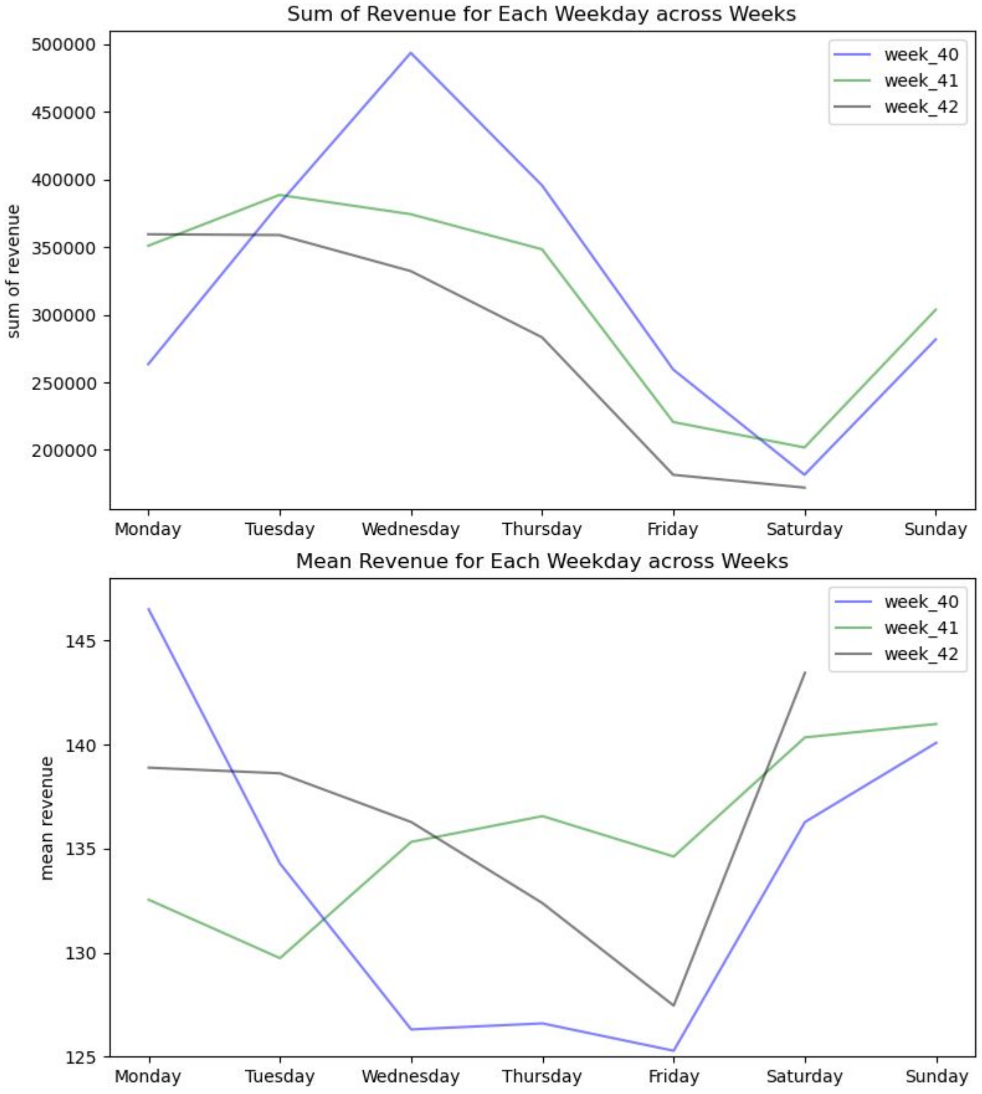


We can see that the **number of hits to the website** and also the **revenue earned** is **maximum** during the **weekdays** (Wednesdays). This can mean some of the followings:

- The **items sold** are **mainly used** by the customers on the **weekends**. So they pile up or buy during the week till Wednesday and then plan for the holidays.
- The **buying pattern of the Australians** is mostly on the weekdays.
- From **10 am to 6 pm** there's **not much sales** in weekdays or weekends. The best time to advertise is in the morning **(8-10 am)** and in the nights **(8-2 am)**.

The key takeaway is, we can do **more promotions** and **advertisements** on **Sundays** and **Mondays** because that will build up their mindset and definitely on the peaks i.e. **Tuesdays** and **Wednesdays** by showing what people in their neighbourhood are buying and what is the hot product that week. Also, the hours should be between what we discussed.

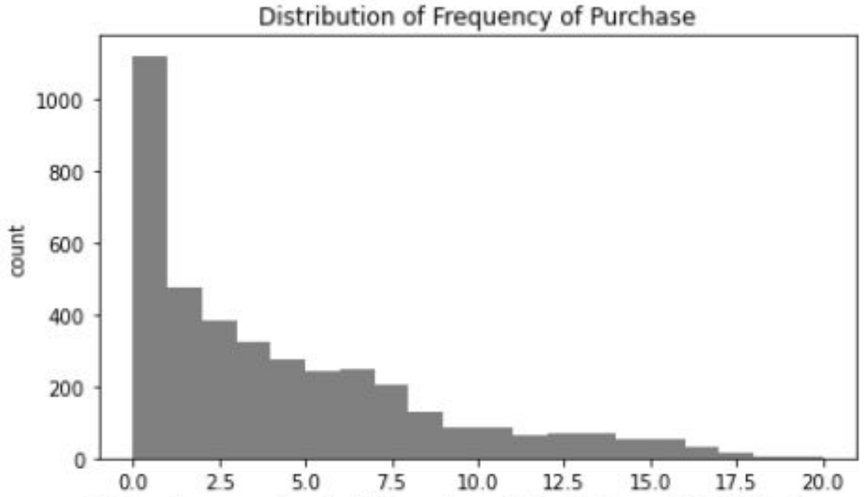
Quality of Revenue

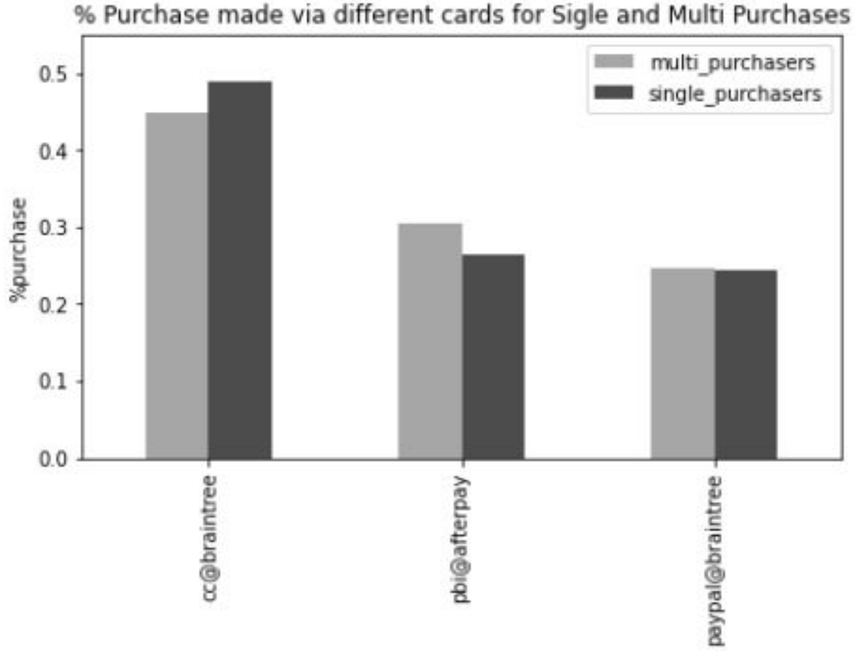


As compared to the sum of revenue which happens to peak on mid-week (because of greater hits/visits), the **mean revenue** is better on the **weekends**.

- This maybe is an indication that the **items bought during the weekdays** are of **quick**/**cheap** in nature as compared to the items bought during the weekends.
- For marketing strategy, we would recommend that for the items mostly sold during the weekends can be traced by **user behaviour** and **site visits**. We can **promote those items** well **in advance** in the **weekdays for several weeks** to lure and make them buy it.

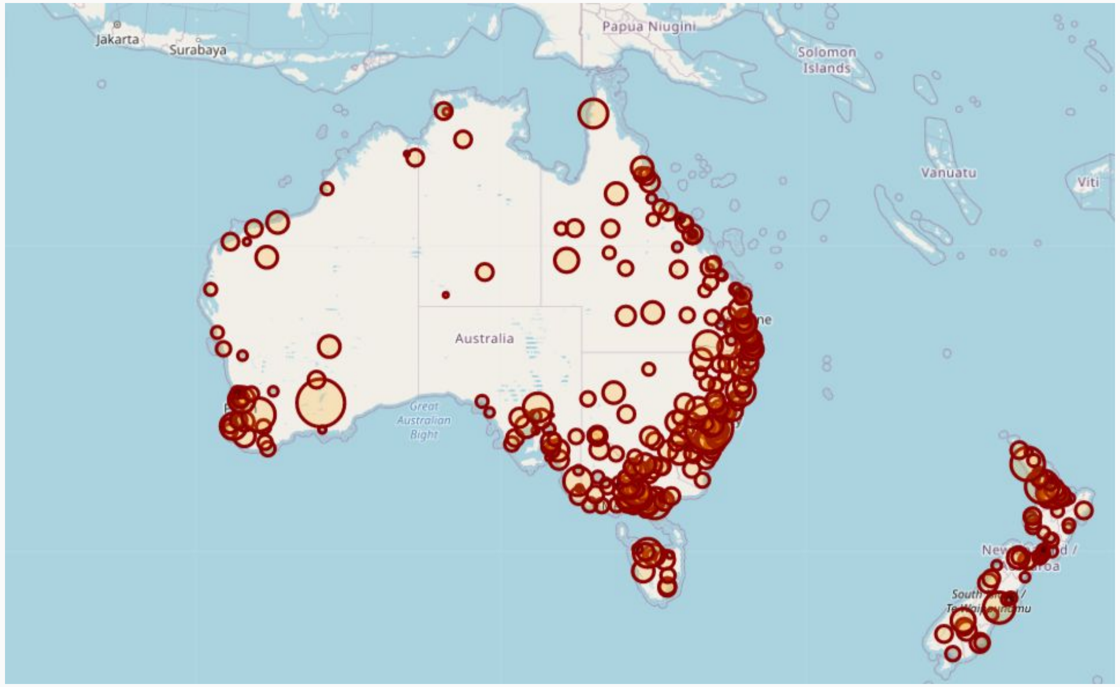
Purchases



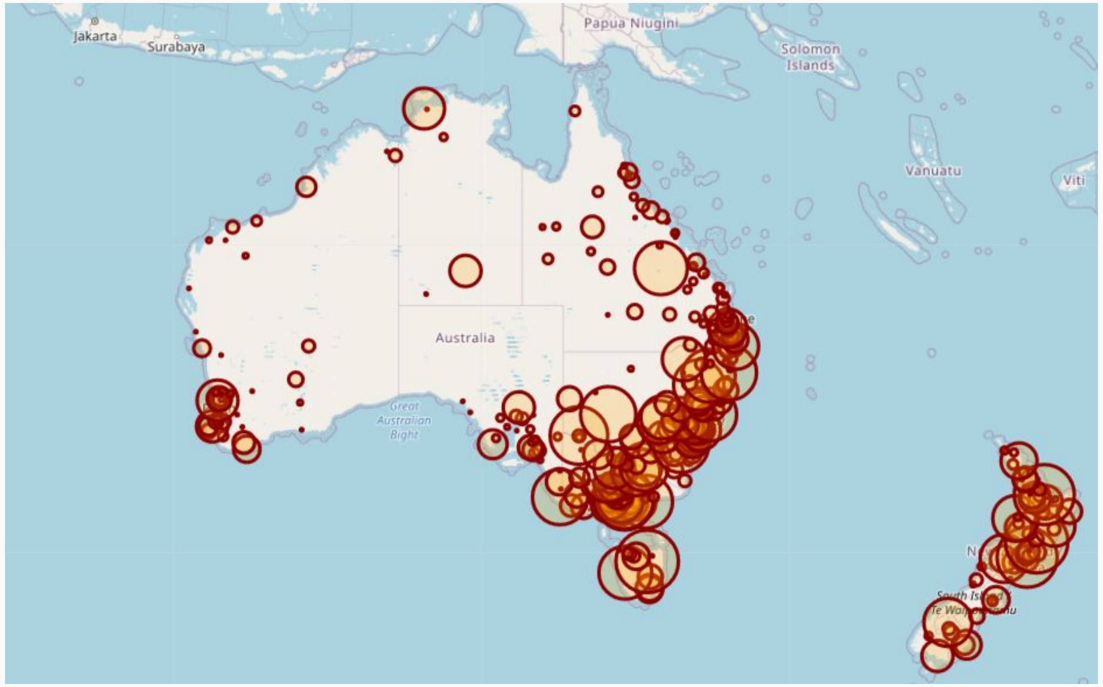


- Average Purchase Frequency is 3.82 days.
- We can see here that the **Purchase Frequency** is again **skewed**.
- Mostly the users have a frequency of 1 day. We have to some extent of multi-purchases in the region of <10 days.
- So, for the **users** who are doing **multiple purchases**, we can keep them more active by giving more **offers** and better **recommendations**.
- For, the **customers** who are doing **single-purchases**, we can send out **reminders**, give them **more offers** and see if they **react to promotional sales**.
- We can also see that **cc@braintree** remains the **favourite payment mechanism**. We can make deal with them, so that if a **customer pays via that platform, they can get some discounts**.

Where are the Sales?



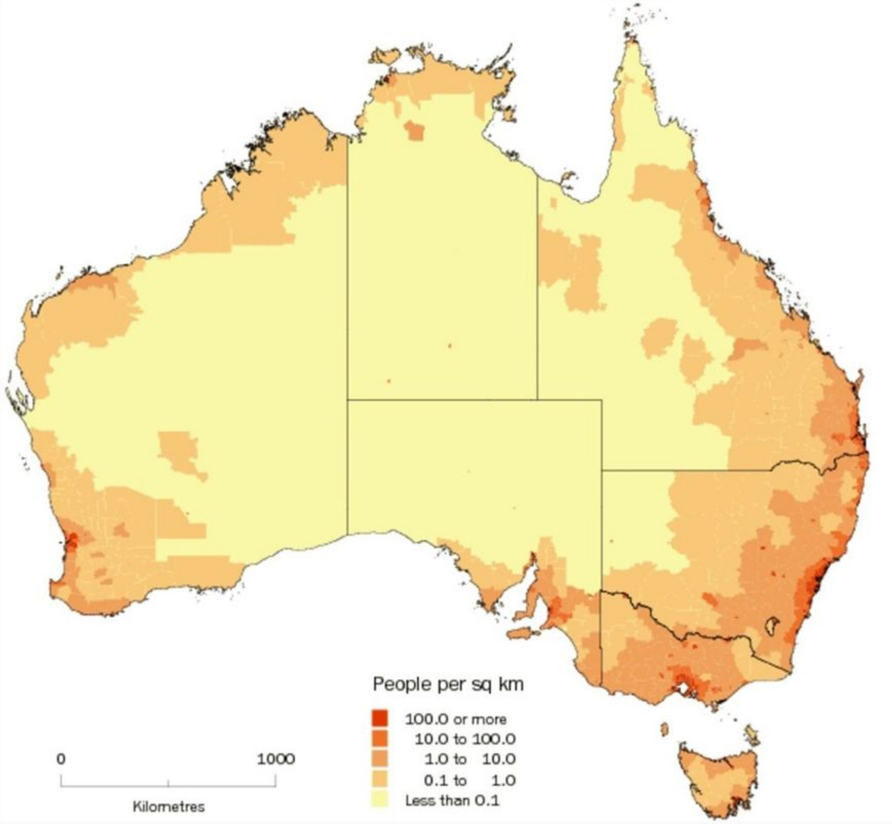
average revenue

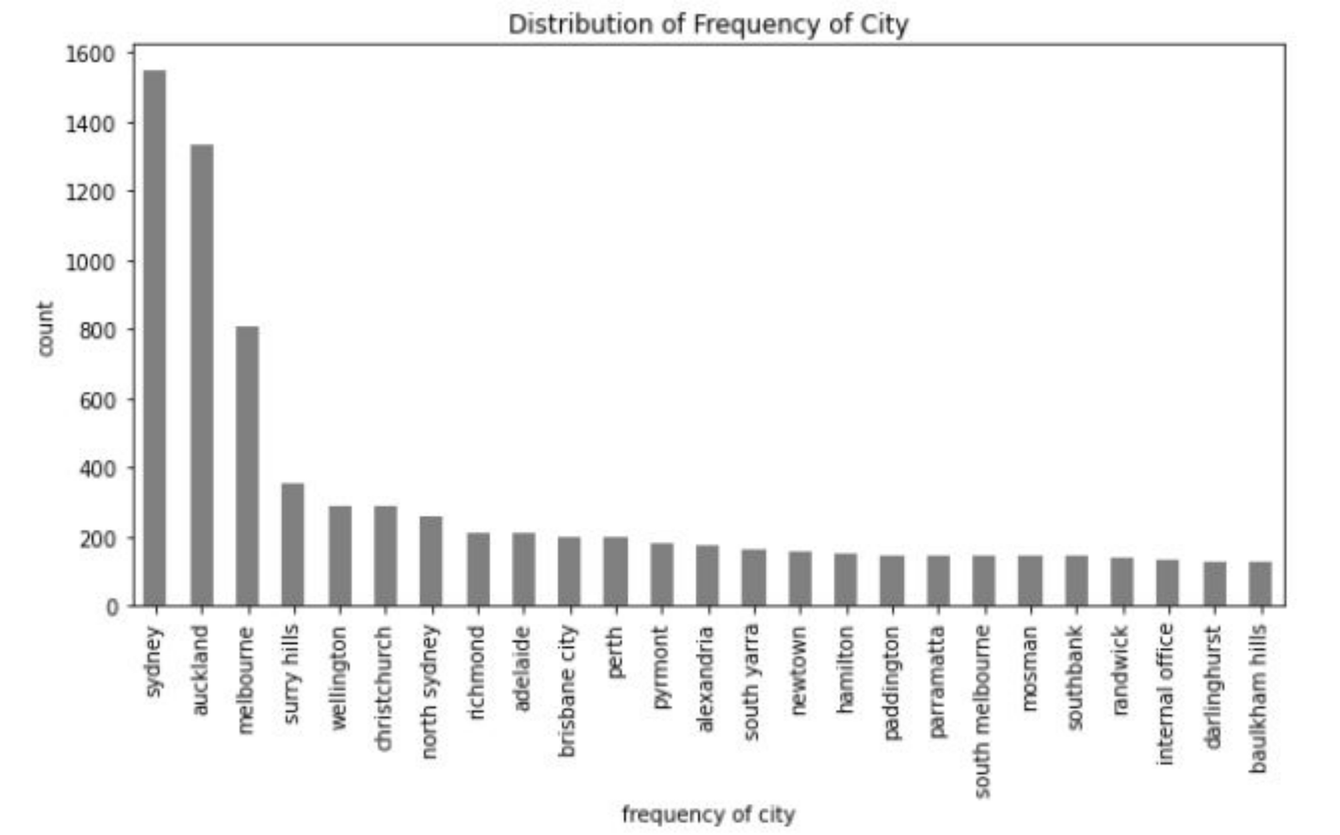


total revenue

Since the sales seems mostly on the coastal region, this naturally leave us with the question that what kind of items does the brand sell?

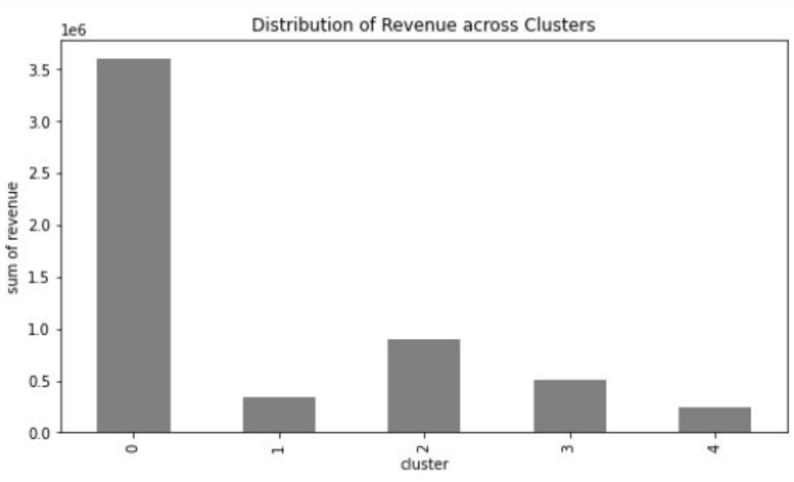
Where are the Sales?

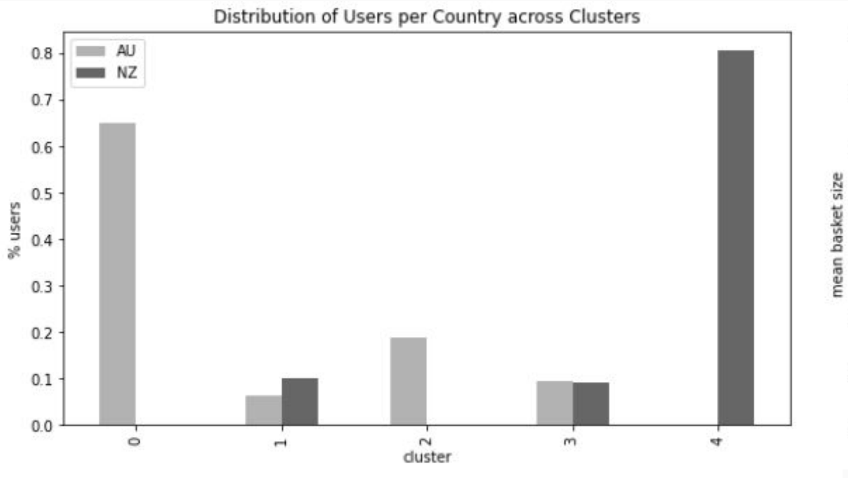


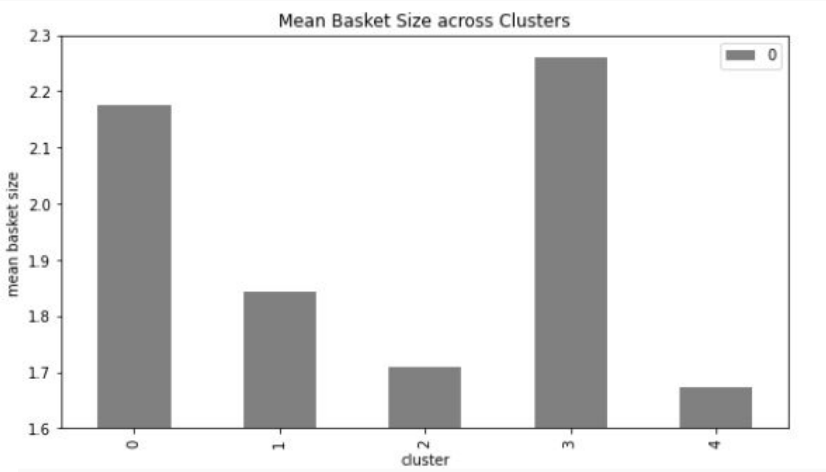


Based on the population density, we can see that Australia is mainly populated in the coastal region and hence this proves the point why our sales were also in the same region. In the next graph, we can see that Sydney, Auckland and Melbourne has the highest transactions. So, we can also do geography based campaigning.

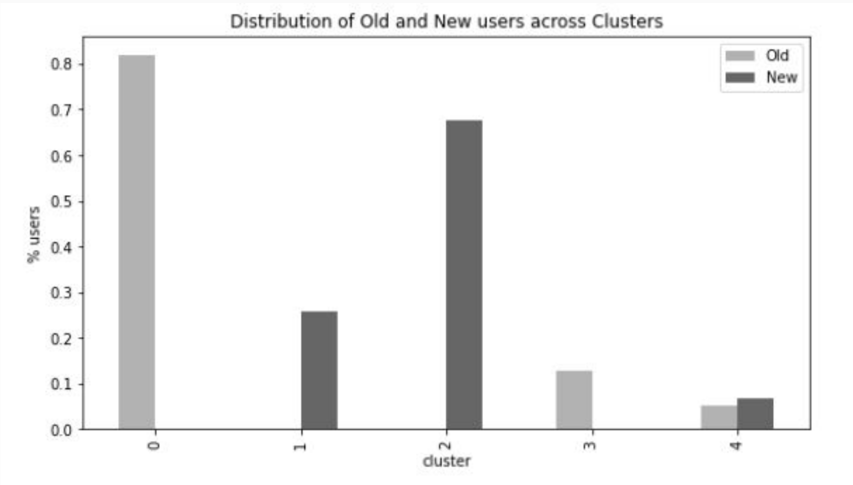
Segmentation

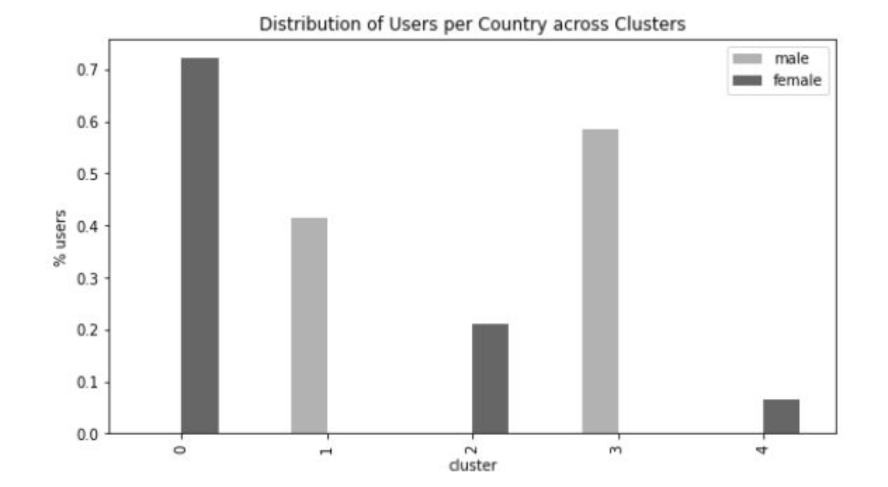






Is_first_order	Max()
Revenue	Mean()
Total_products	Mean()
Country	Max()
Gender	Max()





0 has the Australian, male, old-users who have produce best quality and the highest revenue.

1 has the first-time, female buyers.

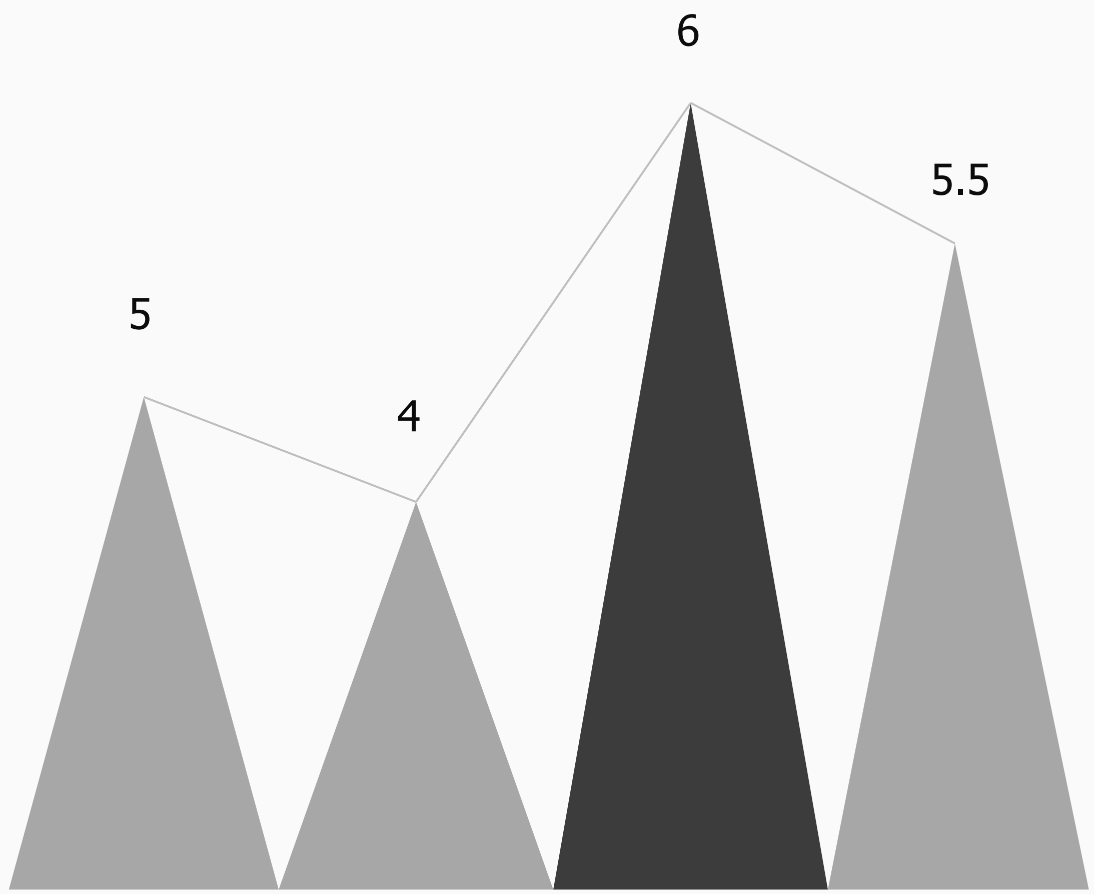
2 has the first-time, male buyers.

3 has the old female users who have a big basket size.

4 has the New Zealand users who produce least quality revenue and neither have a big basket size.

Strategy

Once we have identified some of the interesting patterns, let’s conclude by stating some strategy to enhance the sales and user experience.



proposed a number of strategies on the way. However let's see the most important decisions.

- Target promotions on weekends for quick sales in the weekdays.
- Study user behaviour and page visits to recommend flagship products in the weekdays over weeks/months.
- Use the timing recommendation - (8-10 am)/(8-2 am) to remind/app-banner campaigns to attract customers.
- Can open a partnership with braintree to provide discounts who use that as a mode of payment.
- Use a better recommended (not MBA) for generating product affinity.
- Study the purchase pattern of users in cluster 0 to know their needs and recommend/remind products/alternatives in the right time.
- Give promo to cluster 1 and 2 so that they are attracted to the website.
- Recommend related products that for cluster 3 to make them purchase more.
- Study the behaviour of cluster 4 and try to engage them more by campaigning the brand in New Zealand and with promotional discounts.

Thank You