YOU CAN START YOUR PROJECT FROM HERE

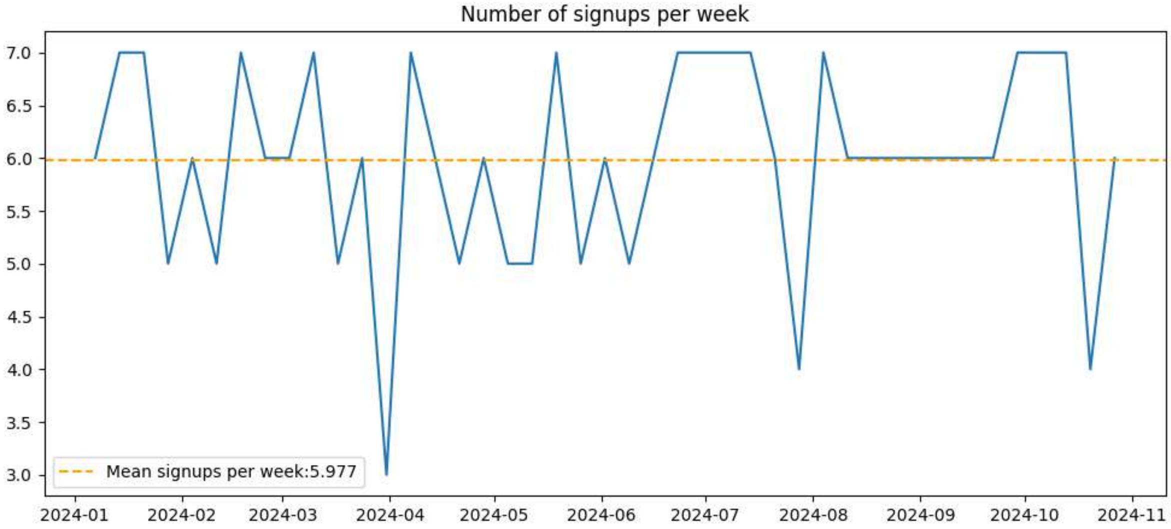
Introduction

This is a report analysing the recent customer signups dataset. It contains a list of 300 customers and 10 associated details. The details include the acquisition source, whether they opted-in to receive marketing materials, which plan they selected, as well as their name, email, customer ID, signup date, age, gender, and region. The aim of this analysis is to conduct a data quality audit and derive insights into user acquisition trends which will help the Marketing and Onboarding team improve acquisition and engagement. This analysis also aims to answer 4 specific business questions as requested.

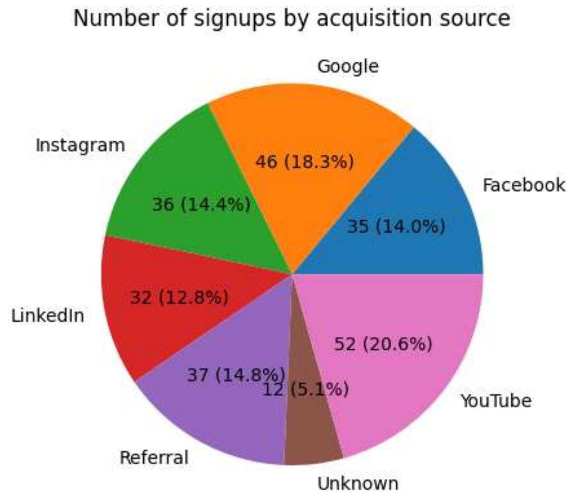
Data Cleaning Summary

Overall, about 85.95% of the original data remained after cleaning. There was one duplicate customer ID which was removed. Customers with missing or invalid ages or signup dates were removed. Customers with invalid entries for “gender” and “marketing opt-in” were also removed. Most other invalid non-empty entries were dropped for efficiency, with the exception of ‘??’ changed to ‘unknown’ in the ‘source’ column. All other missing entries were filled with “unknown” to keep them in the analysis. All word-based entries were standardised to be lowercase only.

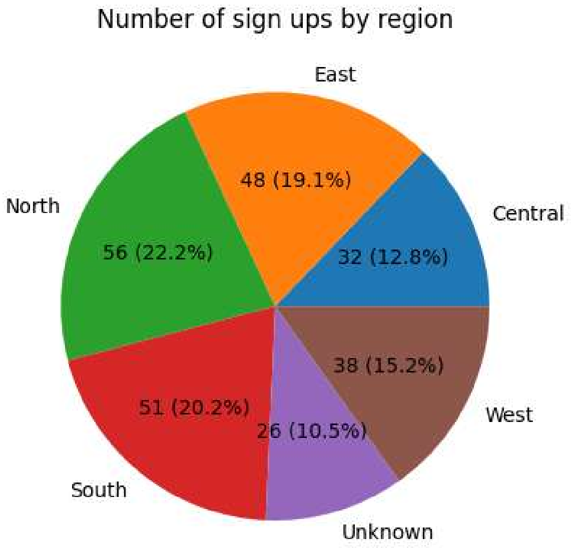
Key Findings & Trends



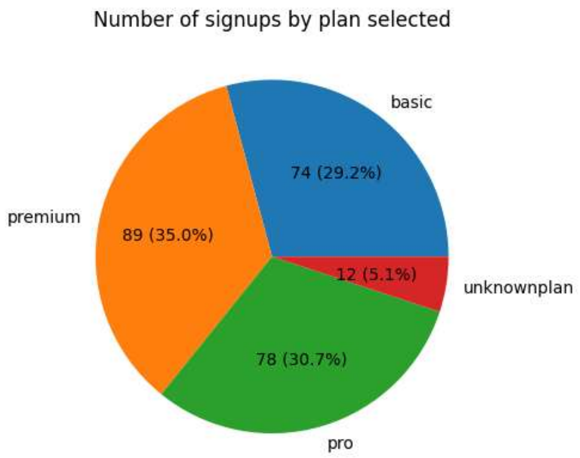
From the above figure, we can see that the average number of signups per week was about 5.977. In particular, we see 3 large dips which immediately spike up in the next week. These could represent where marketing campaigns had ended for a week, then restarted the following week, but we would have to check with Marketing if this was the case. Notwithstanding this, the period from late June to mid-October had mostly average or above-average signup numbers, while the period before that had mostly below-average or average signup numbers.



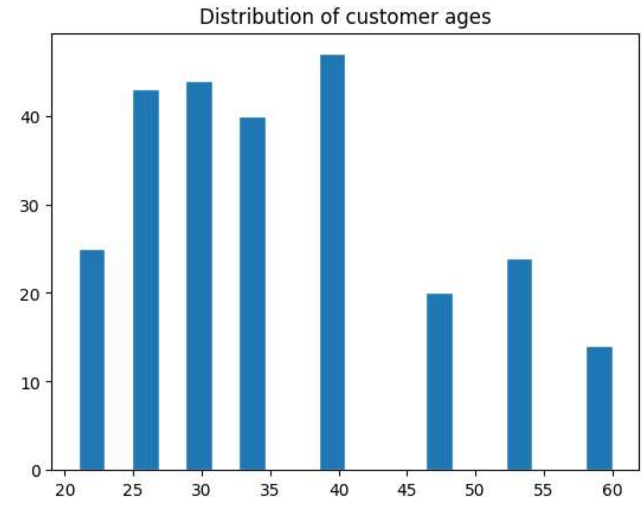
From the figures on the left, we see that most of the signups come through YouTube, followed by Google and direct referral. About 5% come from unknown sources. These could be other sources outside of the named 6, or simply missing data.



We see that most of the signups come from the North, followed by the South and the East. Of concern is the number of signups for whom a region could not be determined, making up just over 10% of the dataset.



We see that the Premium plan was most commonly selected, followed by the Pro and Basic plans. We could not determine the plan closen for 5% of signups.



The figure on the left shows the distribution of the ages of the signups. The average signup age was 35.6, and the median was 34. The youngest was 21 and the oldest was 60. We can see that the majority of signups come from people aged 25-40. There are also noticable ‘gaps’ in the distribution, indicating potential marketing opportunities for new users.

Table 1: Opt-in Statistics by gender

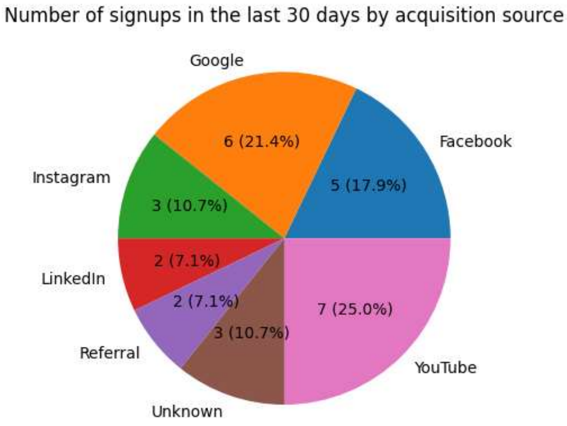
Gender	Total no. of users	No. of users who opted-in	Opt-in Rate
Female	83	38	45.78
male	78	33	42.31
Non-binary	37	17	45.95
Other	52	21	40.38

Finally, the table above shows that the opt-in rate is around 3-5% higher for females and non-binary users compared to males and users of other genders.

Business Questions Answers

Below are answers to the 4 requested business questions.

1. Which acquisition source brought in the most users last month?



Defining “last month” as the final 30 days of the dataset, we can see from the figure on the left that YouTube brought in the most users last month. This is consistent with the overall trend that YouTube brought in the most users. This was followed by Google and Facebook.

2. Which region shows signs of missing or incomplete data?

Table 2: Missing entries by region

Region	North	East	West	South	Central	Unknown
No. of missing entries	18	14	11	10	9	6

From the table above, we can see that the North and East regions had the most missing data.

3. Are older users more or less likely to opt in to marketing?

Table 3: Opt-in Statistics by age group

Age Group	Total No. of Customers	No. who opted-in	Opt-in Rate (%)
20–24	25	7	28.00
25–29	87	41	47.13
30–34	40	17	42.50
35–39	0	0	N/A
40–44	47	22	46.81
45–49	20	10	50.00
50–54	24	10	41.67
55–59	0	0	N/A
60–64	14	6	42.86

From the table above, it is not obvious whether age has any effect on the likelihood of opting in to marketing, although it seems to have no effect. Further statistical testing is required to confirm this.

4. Which plan is most commonly selected, and by which age group?

Table 4: Number of each plan selected by age group. Note that there were no users aged 35-39 or 55-59. Largest number for each group has been bolded.

Age Group	Plan Selected			
	Basic	Premium	Pro	Unknown
20–24	8	4	**12**	1
25–29	**30**	**30**	23	4
30–34	9	**15**	14	2
35–39	N/A	N/A	N/A	N/A
40–44	10	**22**	13	2
45–49	4	**9**	5	2
50–54	**9**	6	8	1
55–59	N/A	N/A	N/A	N/A
60–64	**5**	4	4	1

From the table above, we see that the most commonly selected plan across the majority of users is the Premium plan. The exceptions are that young users aged 20-24 mostly selected the Pro plan, and older users aged 50-60 preferred the Basic plan.

Recommendations

Based on the above analysis, we recommend the following actions:

1. Targeted marketing towards the 35-39 year old age group

This is a clear gap in the age distribution, especially since we have many users aged 30-34 and 40-44. More should be done to investigate why people aged 35-39 did not sign up while people slightly older or younger did, with the findings then used to improve the product and/or craft more targeted campaigns toward this untapped segment.

2. Targeted marketing towards females and non-binary people

Earlier analysis showed that these segments had higher opt-in rates than other genders, though more statistical analysis is needed to say if they are indeed more likely to opt-in than other genders. If so, a targeted marketing campaign towards new females and non-binary people could result in a larger proportion of our user base being opted-in to marketing, potentially increasing customer retention in the future since more existing users will receive tailored marketing.

3.Improved data reporting, especially in the North, East, and West regions

Most columns had about 3 – 4% of their values missing, with 2 reaching 10% or more. This led to lots of entries being filled with ‘unknown’, and hence being unable to draw useful insights from them. More should be done across all regions to improve data reporting, but especially in the North, East and West, since as discussed earlier in this report, customers in these regions had the most number of missing values.

Data Issues

One data quality issue found was the entries with invalid values. This includes an age of ‘thirty’ instead of ‘30’, an acquisition source of ‘??’, and a gender of ‘123’ among various other invalid entries. Such entries make the cleaning process more complicated than necessary, as the analyst must decide whether to drop all such rows entirely, or manually editing the entry to what the analyst interprets the entry to be e.g. replacing ‘thirty’ with ‘30’. Often times the analyst will choose to simply remove the row for efficiency, leading to loss of data that could have been useful if only it was recorded properly.

One way to fix this in the future is to implement checks during the data input stage. For entries that should only contain numbers, ensure that data entered is numeric. For entries that should only have a limited number of choices, make sure the entry is one of those choices (e.g. only ‘Yes’ or ‘No’, not ‘Nil’). These checks will increase efficiencies for future analysts as they spend less time cleaning the data, and more time analysing the data, and also lead to more usable data points.