AGENT3: DATA STUDY

AN ANALYSIS INTO 2022 CHANGE WHOLESALE CAMPAIGNS

Table of Contents

1. Business Problem Context

The business problem is to identify the optimal marketing stream for Change Wholesale to invest in for advertising its products to achieve its three key objectives:

- Securing new business with brokers who are registered but not trading with Change.
- Increase engagement with newly registered and prospective brokers.
- Encouraging new broker signups.

To address this problem, the research team has identified four key questions:

- Which paid media campaigns have historically performed well for Change Wholesale?
- Which campaign elements show the biggest impact on website conversions?
- How cost effective are the various ad formats and platforms with respect to the campaign metrics: clicks, impressions and reach?
- In which geographical areas should advertisement effort be concentrated?

2. Project development process

To achieve efficient outcomes throughout the research process, responsibilities have been split between team members. There has been a separate person assigned to each of the main datasets (Google Analytics, Creative Data, Demographics Data) and another person to perform regression analysis. Each person then worked on their own part of the dataset, making sure:

1. The data is cleaned by the mean of identifying and handling missing or inconsistent values, checking for outliers, and removing duplicate entries if necessary.
2. Exploring the data sets using descriptive statistics, such as mean, median, standard deviation, and percentiles, to understand the distributions of the data.
3. Merging and grouping dataset as appropriate.
4. Visualising the data using graphs, plots, and tables. Particular emphasis has been put on clarity and message of the graphs & tables. For example:
    - In demographics data it was decided that tables are a better visualisation tool given multiple variables (count and CTR) with wide ranging spreads (CTR of 0.001 and 1, count any value between 31 and 86525).
    - It's important to note that not all graphs have been used in the final presentation to the client. Some have been put in the appendix and were only used for context or to better familiarise ourselves with the data. An example of such graph is below in figure 2.1: when analysing total completions, it was noted that Broker Logins (one of the key targets for Change Wholesale) forms only a small part of the overall completions.

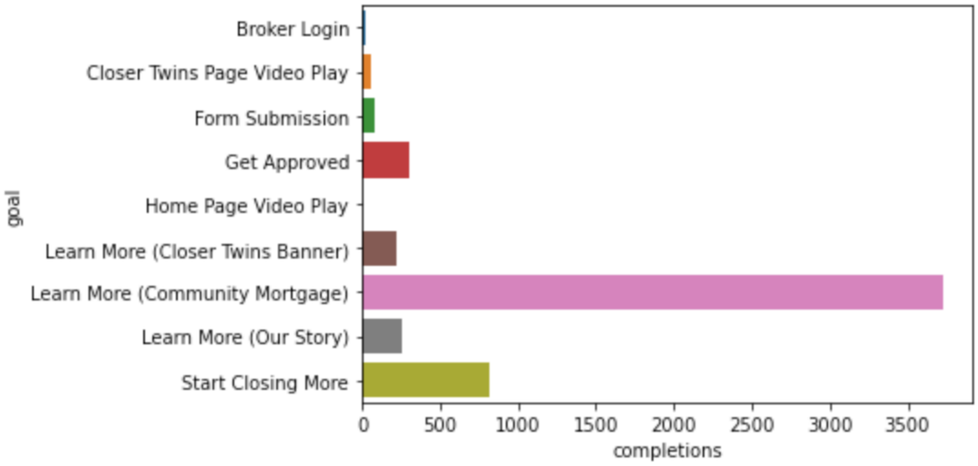


Figure 2.1: total completions per defined goal.

This insight was then used to explore further (figure 2.2): zooming in on only Broker Login completions to understand which platform could be used to improve this metric. Despite a low volume of data, this allowed the careful conclusion that Trade Media and LinkedIn are relatively strong contributors to Broker Logins, showing a different pattern than the general audience (where Google SEM leads most metrics). Therefore, LinkedIn and Trade Media should not be dismissed entirely by the client.

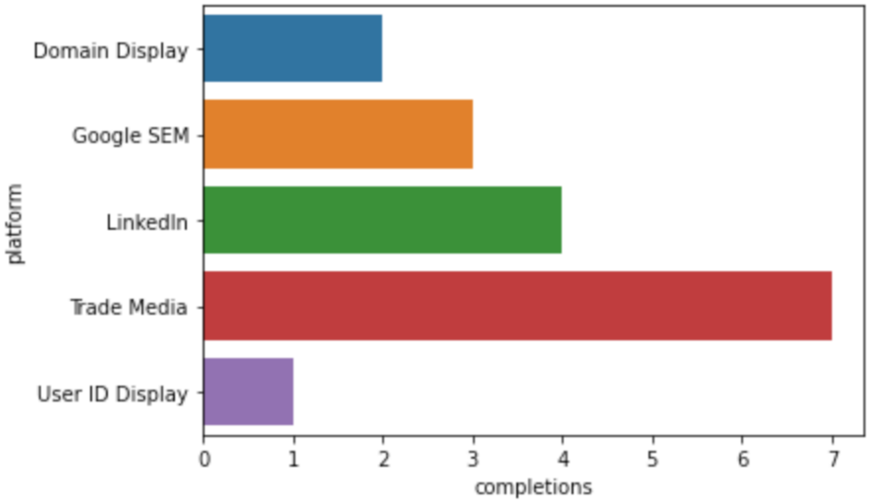


Figure 2.2: total completions per platform for goal broker logins.

5. Finally, the appropriate graphs/tables were selected to create a coherent presentation. We have concentrated on the breakdown of completion by audience/ platform, CTR exploration in Demographics data, Time Series of completions to spot any seasonality and help the client concentrate marketing efforts in more popular months. This was to ensure clarity of the presentation and to avoid duplication of messages.

Furthermore, in addition to investigating the individual source files provided by Agent3, two other types of analysis have been performed:

1. Exploration of US Census 2020 data to help drive geographical recommendation of the ad effort.
2. CTR modelling as described below.

A predictive model for click-through rate (CTR) was created which can efficiently forecast CTR given an ad format, spend, and month. Our analysis showed that ad format and month played a significant role in determining the value of CTR.

Upon reviewing the data, we observed that high spending in certain months, such as November, October, and September, resulted in a higher CTR score. Conversely, high spending in other months, such as May, April, and July, did not yield a high CTR score. This aligned with seasonality effect previously explored in completion data.

3. Technical overview of the code

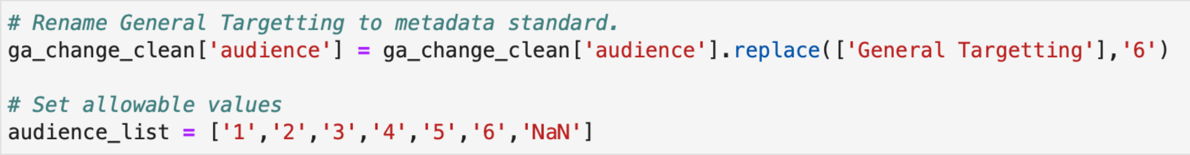
The data has been cleaned and analysed separately for all datasets.

3.1 Google Analytics Data

The Google analytics data consisted of three Excel sheets, which were saved as separate .csv files:

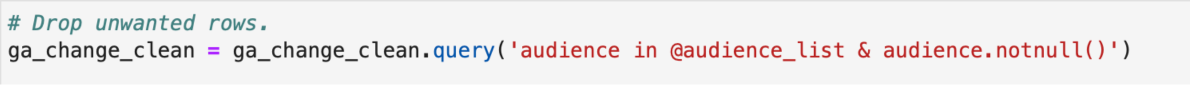
- Change_2022_Google Analytics Ma.csv: “ga_change” notebook.
- General stats - web traffic.csv: “ga_general” notebook.
- Goal stats - web traffic.csv: “ga_goal” notebook.

The three files are very similar in content and structure, so only the unique aspects used for analysis of each file will be discussed. Similar cleansing was used on all three files, dropping unwanted columns, renaming audiences to reflect metadata and focusing in most cases on rows where audience was not null (as these rows represented campaign-related traffic, which is the focus of this study).



```
# Rename General Targetting to metadata standard.
ga_change_clean['audience'] = ga_change_clean['audience'].replace(['General Targetting'],'6')

# Set allowable values
audience_list = ['1','2','3','4','5','6','NaN']
```

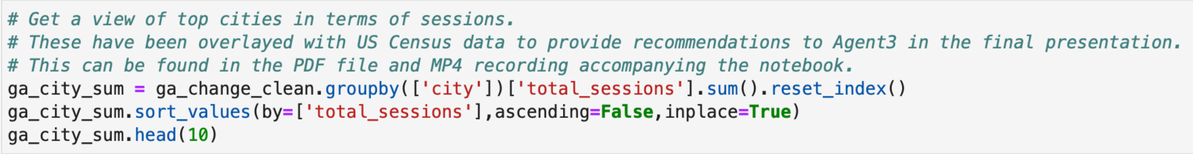


```
# Drop unwanted rows.
ga_change_clean = ga_change_clean.query('audience in @audience_list & audience.notnull()')
```

Specific activity for each file:

ga_change:

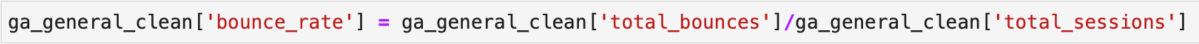
Data uniquely used from this file is around City/Country location per session. This was used in the client presentation to overlay on the 2020 US Census data described in Section 5. The top 10 cities (all US) were identified through below code snippet:



```
# Get a view of top cities in terms of sessions.
# These have been overlayed with US Census data to provide recommendations to Agent3 in the final presentation.
# This can be found in the PDF file and MP4 recording accompanying the notebook.
ga_city_sum = ga_change_clean.groupby(['city'])['total_sessions'].sum().reset_index()
ga_city_sum.sort_values(by=['total_sessions'],ascending=False,inplace=True)
ga_city_sum.head(10)
```

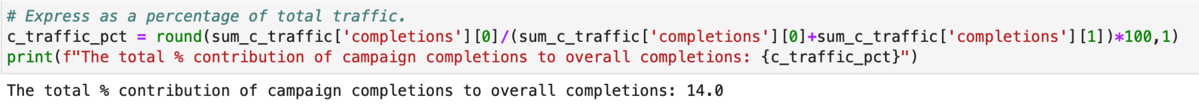
ga_general:

Data uniquely used from this file is around the total bounces per session identified by Google Analytics data. This was used to calculate the bounce rate via the following code:



```
ga_general_clean['bounce_rate'] = ga_general_clean['total_bounces']/ga_general_clean['total_sessions']
```

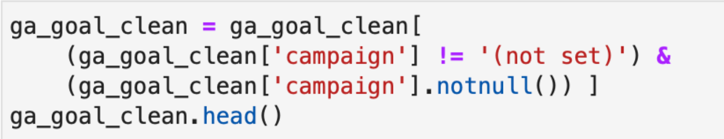
ga_goal:
This dataset was mostly analysed from its unique addition of providing completion data per goal. A check was done to see which part of completions originate from campaigns versus general traffic:



```
# Express as a percentage of total traffic.
c_traffic_pct = round(sum_c_traffic['completions'][0]/(sum_c_traffic['completions'][0]+sum_c_traffic['completions'][1])*100,1)
print(f"The total % contribution of campaign completions to overall completions: {c_traffic_pct}")
```

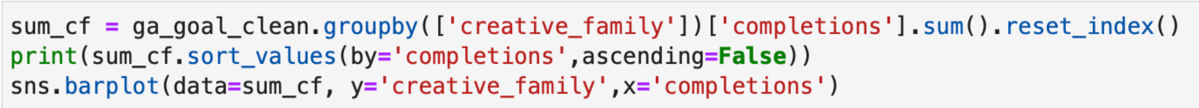
```
The total % contribution of campaign completions to overall completions: 14.0
```

Data was further cleaned to only include campaign-related goal/completions results:



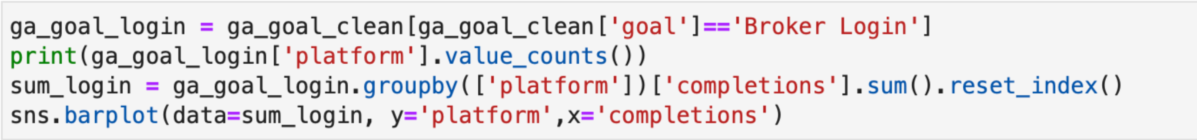
```
ga_goal_clean = ga_goal_clean[
    (ga_goal_clean['campaign'] != '(not set)') &
    (ga_goal_clean['campaign'].notnull()) ]
ga_goal_clean.head()
```

Various ‘groupby’ clauses were then used to get insights into factors leading to high completions against key goals. An example based on grouping by creative family:



```
sum_cf = ga_goal_clean.groupby(['creative_family'])['completions'].sum().reset_index()
print(sum_cf.sort_values(by='completions',ascending=False))
sns.barplot(data=sum_cf, y='creative_family',x='completions')
```

The last section of code zooms in on key subsets of the data related to specific goals and platforms:

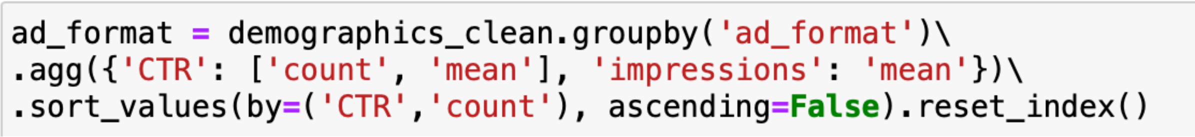


```
ga_goal_login = ga_goal_clean[ga_goal_clean['goal']=='Broker Login']
print(ga_goal_login['platform'].value_counts())
sum_login = ga_goal_login.groupby(['platform'])['completions'].sum().reset_index()
sns.barplot(data=sum_login, y='platform',x='completions')
```

3.2 Demographics data

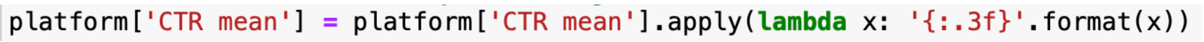
The demographics dataset has been loaded, cleaned (removal of null values rows) and irrelevant columns dropped off. CTR ratios were calculated from impressions and clicks. The performance of various ads was then analysed based on Platform, Audience and Ad formats. Based on Agent3 feedback, we have done this in two ways: either first summing up impressions and clicks and then subsequently calculating CTR ratios from grouped impressions and clicks, or averaging over individual CTRs. Both gave similar results.

In the code we used the ‘groupby’ Pandas function together with ‘agg’ functions (aggregating in different way over different columns):



```
ad_format = demographics_clean.groupby('ad_format')\
.agg({'CTR': ['count', 'mean'], 'impressions': 'mean'})\
.sort_values(by=('CTR','count'), ascending=False).reset_index()
```

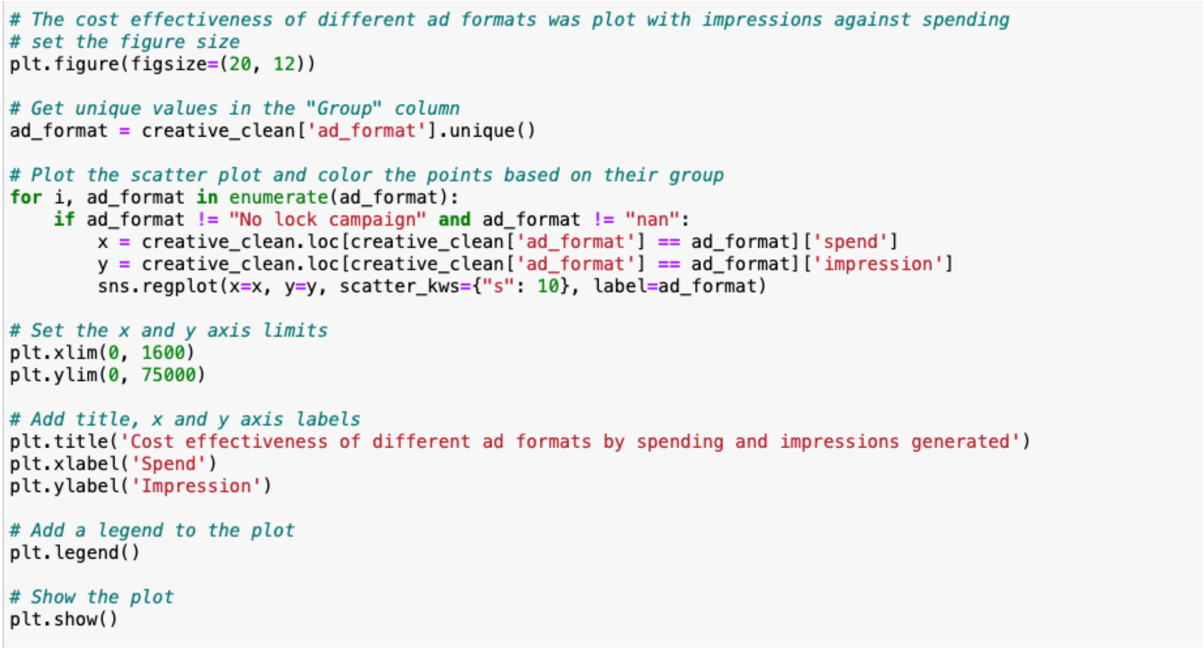
Some Lambda functions were also used to help with formatting:



```
platform['CTR mean'] = platform['CTR mean'].apply(lambda x: '{:.3f}'.format(x))
```

3.3 Creative Data

The creative dataset was initially cleaned in Excel for quick adjustment of inconsistent formatting regarding numeric columns for ease when loading into python. The relevant columns used in the analysis were formatted as appropriate for cleaning. As per the comments Agent3 made about their audience and platform data, the respective fields were tidied up by categorising misnamed objects. The cost-effectiveness of ad formats and platforms was then investigated using a combination of Seaborn and Matplotlib to generate a scatter plot of each metric (clicks, impressions and reach) against spend to visualize the cost-effectiveness of different advertising formats and proceeding that ad platforms.



```
# The cost effectiveness of different ad formats was plot with impressions against spending
# set the figure size
plt.figure(figsize=(20, 12))

# Get unique values in the "Group" column
ad_format = creative_clean['ad_format'].unique()

# Plot the scatter plot and color the points based on their group
for i, ad_format in enumerate(ad_format):
    if ad_format != "No lock campaign" and ad_format != "nan":
        x = creative_clean.loc[creative_clean['ad_format'] == ad_format]['spend']
        y = creative_clean.loc[creative_clean['ad_format'] == ad_format]['impression']
        sns.regplot(x=x, y=y, scatter_kws={"s": 10}, label=ad_format)

# Set the x and y axis limits
plt.xlim(0, 1600)
plt.ylim(0, 75000)

# Add title, x and y axis labels
plt.title('Cost effectiveness of different ad formats by spending and impressions generated')
plt.xlabel('Spend')
plt.ylabel('Impression')

# Add a legend to the plot
plt.legend()

# Show the plot
plt.show()
```

The code utilised several basic for loops and if statements to group the plot by the desired observed categories (ad format and ad platform) for easy visual comparison.

3.4 Predictive CTR model

The CTR predictive model was built mainly on the weighted CTR column, having removed all numerical columns except spend. Subsequently, the categorical variables were one-hot encoded after binning the dates into months.

The biggest improvement occurred when categorical variables were one-hot encoded instead of label encoded (as the data is nominal, not ordinal) and would be the suggested approach for other data sets over longer time periods.

3.5 ANOVA Model

We have created an ANOVA model to further analyse CTR scores. As the outcome of the model were not included in the final presentation, we put technical details of the model in appendix.

4. Insights, Trends and Patterns

The various ad platforms and formats had different impacts on the observed clicks, reach and impressions, summarised by the return on investment relative to their counterparts shown in table 4.1 below.

Ad platform or Ad format	Clicks ROI	Impressions ROI	Reach ROI
Instagram and Facebook	Low	Low	Low
Domain	Moderate	High	Moderate-Low
Trade Media	Low	High	low
Native	Low	High	Moderate
CPC (Google SEM)	High	Low	Low
Display	Moderate	Moderate	Moderate
Single image and video	Low	Low	Low

Table 4.1: ROI scores per at platform/format.

Additionally, the predictive model built on spend, the various ad formats, and advert months an R^2 score of 0.79-0.80. Confirming the dependency of the various factors above to the ad formats and a seasonality to the CTRs.

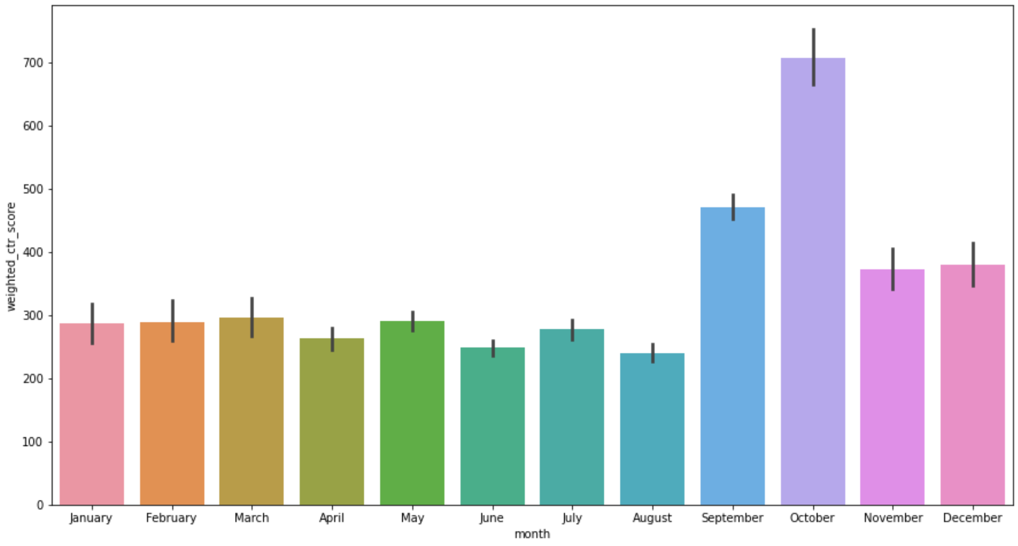


Figure 4.2: weighted CTR per campaign month (2022).

5. Recommendations and conclusions

This section has been divided according to the three key objectives identified in the initial scoping phase of the project:

- Extending reach and existing impressions
- Converting existing users
- Increasing stickiness/retention of existing users

5.1 Extending the reach and increasing impressions

- Instagram and Facebook as platforms did not generate a good ROI hence more focus can be put on a combination of domain display and trade media for increased outreach to new and existing brokers.
- As for ad formats, single image and video-related content can be reduced in favour of native and display ads.
- The advertisement effort should be concentrated in the southern part of the US and some metropolitan areas where the percentage of ethnic minorities (which is one of the target audiences) is the highest. In particular it should be targeted in areas shown in non-orange colour below (figure 5.1): South California, South Texas, New Mexico, “Lower South” (Missisipi, Alabama, Georgia, South Carolina), Washington DC, New York City areas.

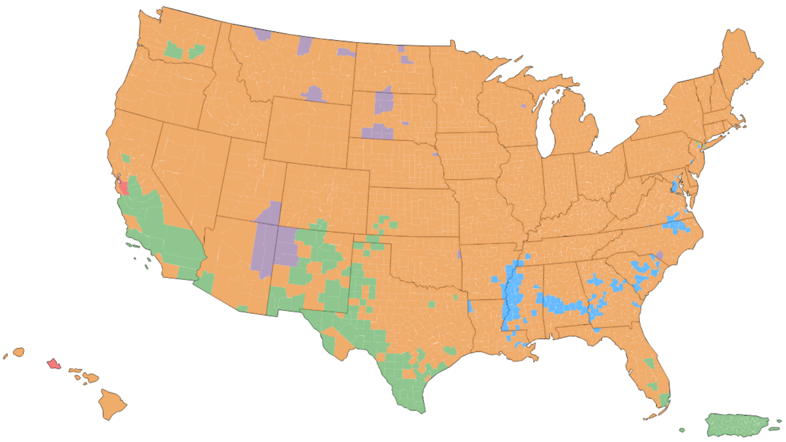


Figure 5.1: US Census data (2020) – non-orange areas likely to be ‘underbanked’.

5.2 Converting existing users

- As the campaign had an overall trend following seasonality, a more seasonal approach could be taken for specific objectives. For example: to maximise clicks, more investment in CPC and Google SEM advertisements can be made when the audience base is more active, aiming for the peak before and after the summer period.
- For customer acquisition, the current funnel and proportion of trade media advertisements can be reviewed and increased, where current content can be reviewed to re-engage existing brokers.
- The current funnel can be reviewed to offer more 'Learn More (Community Mortgage)' CTAs

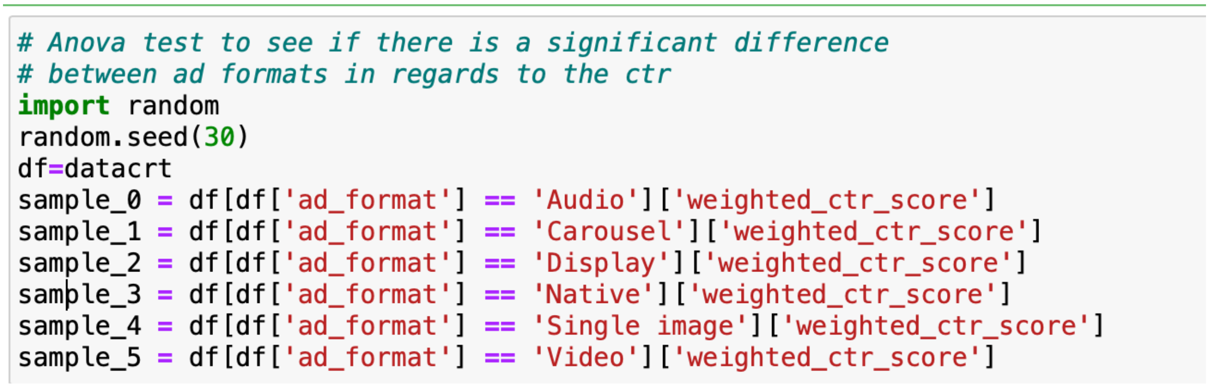
5.3 Increasing stickiness for users

- Higher-quality content can be provided which targets the relevant interests via Google trend analysis. Communications can be personalized with new or prospective brokers. Events can be hosted that reward strong performance for brokers.
- Targeting informative ads for the purpose of informing the customer through high impressions generating platforms and formats such as Domain and Trade media alongside display and native adverts.

6.Appendix

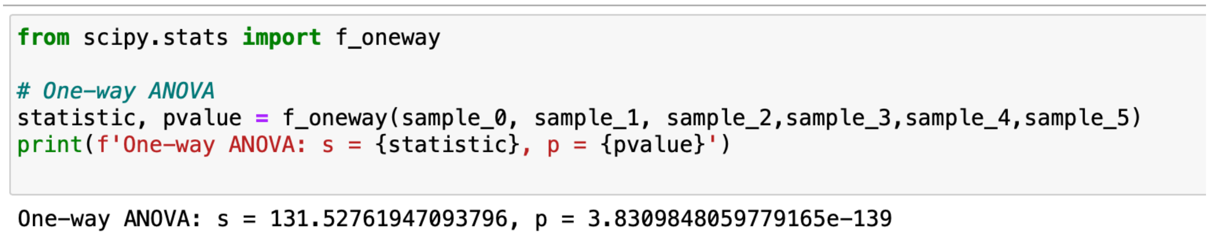
6.1 ANOVA model

- We created a one-way ANOVA model to see if the weighted CTR score across different ad formats differed, effectively telling us if Ad Format had a significant impact on the CTR.
- We checked the main assumption of normality and as this is a large real life dataset we can't always have very good normality. We created Q-Q plots which showed a straight line trend (for majority of the curve, excluding outliers) implying an acceptable amount of normality.
- The p-value was far less than 0.05 thus we rejected the null hypothesis that the mean of the CTR across the different ad formats were the same and accepted the alternative that they were not.
- The code to separate the ad formats was as follows:



```
# Anova test to see if there is a significant difference
# between ad formats in regards to the ctr
import random
random.seed(30)
df=datacrt
sample_0 = df[df['ad_format'] == 'Audio']['weighted_ctr_score']
sample_1 = df[df['ad_format'] == 'Carousel']['weighted_ctr_score']
sample_2 = df[df['ad_format'] == 'Display']['weighted_ctr_score']
sample_3 = df[df['ad_format'] == 'Native']['weighted_ctr_score']
sample_4 = df[df['ad_format'] == 'Single image']['weighted_ctr_score']
sample_5 = df[df['ad_format'] == 'Video']['weighted_ctr_score']
```

- We used the one-way ANOVA from scipy.stats to conclude the p-value as follows:



```
from scipy.stats import f_oneway

# One-way ANOVA
statistic, pvalue = f_oneway(sample_0, sample_1, sample_2,sample_3,sample_4,sample_5)
print(f'One-way ANOVA: s = {statistic}, p = {pvalue}')
```

```
One-way ANOVA: s = 131.52761947093796, p = 3.8309848059779165e-139
```

6.2 Other graphs & tables

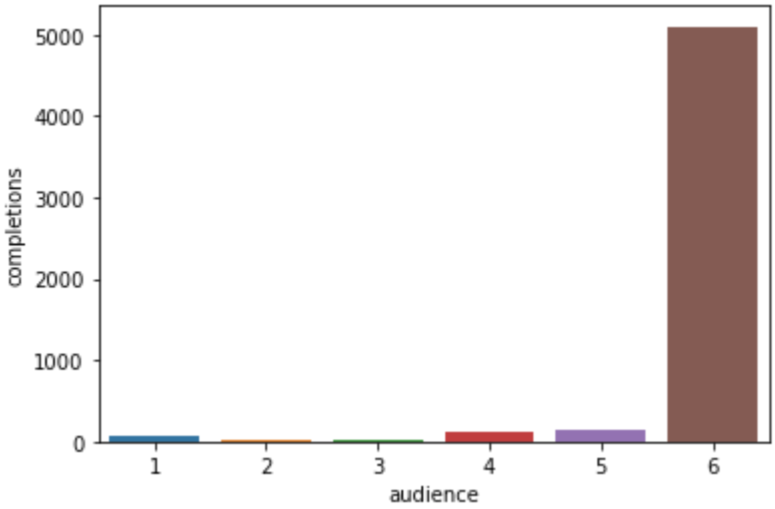


Figure 6.1: Breakdown of completions by audience from the google analytics spreadsheet.

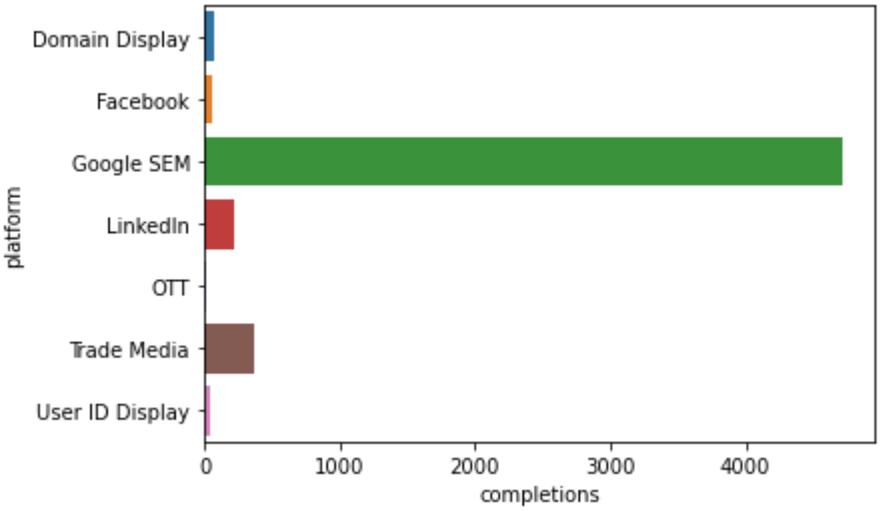


Figure 6.2: Breakdown of completions by ad platform from the google analytics spreadsheet.

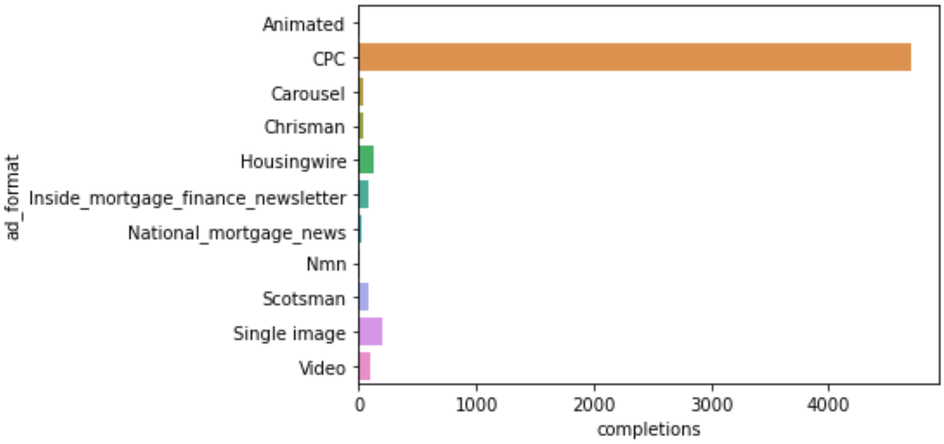
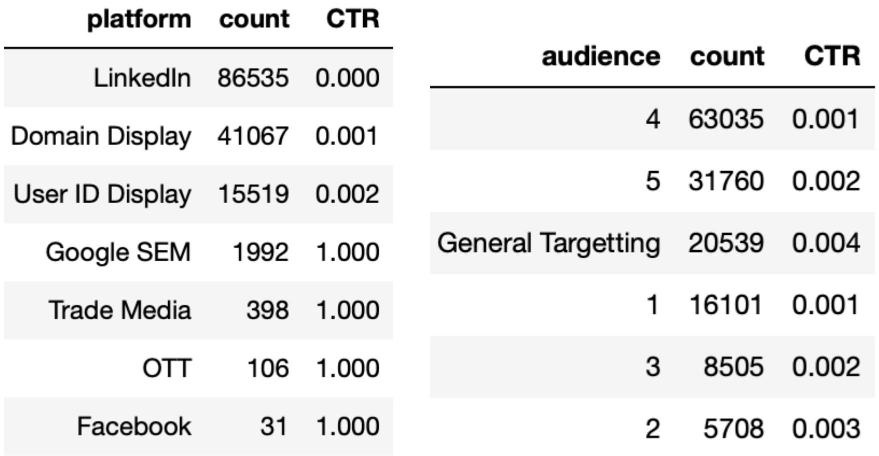


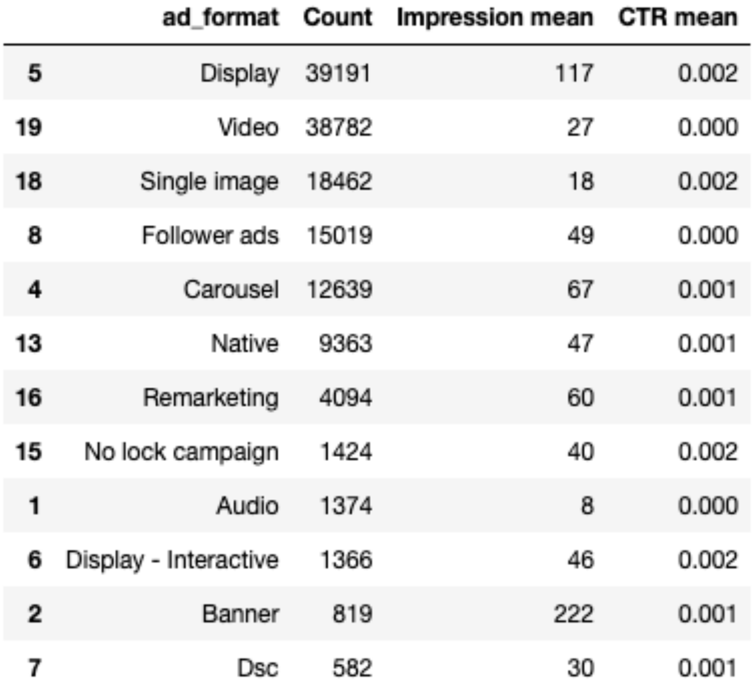
Figure 6.3: Breakdown of completions by ad format from the google analytics spreadsheet.



platform	count	CTR
LinkedIn	86535	0.000
Domain Display	41067	0.001
User ID Display	15519	0.002
Google SEM	1992	1.000
Trade Media	398	1.000
OTT	106	1.000
Facebook	31	1.000

audience	count	CTR
4	63035	0.001
5	31760	0.002
General Targetting	20539	0.004
1	16101	0.001
3	8505	0.002
2	5708	0.003

Figure 6.4: Breakdown of CTR and use cases by ad platform and audience from the demographics spreadsheet.



	ad_format	Count	Impression mean	CTR mean
5	Display	39191	117	0.002
19	Video	38782	27	0.000
18	Single image	18462	18	0.002
8	Follower ads	15019	49	0.000
4	Carousel	12639	67	0.001
13	Native	9363	47	0.001
16	Remarketing	4094	60	0.001
15	No lock campaign	1424	40	0.002
1	Audio	1374	8	0.000
6	Display - Interactive	1366	46	0.002
2	Banner	819	222	0.001
7	Dsc	582	30	0.001

Figure 6.5: Breakdown of CTR, use cases and impressions by ad format from the demographics spreadsheet.

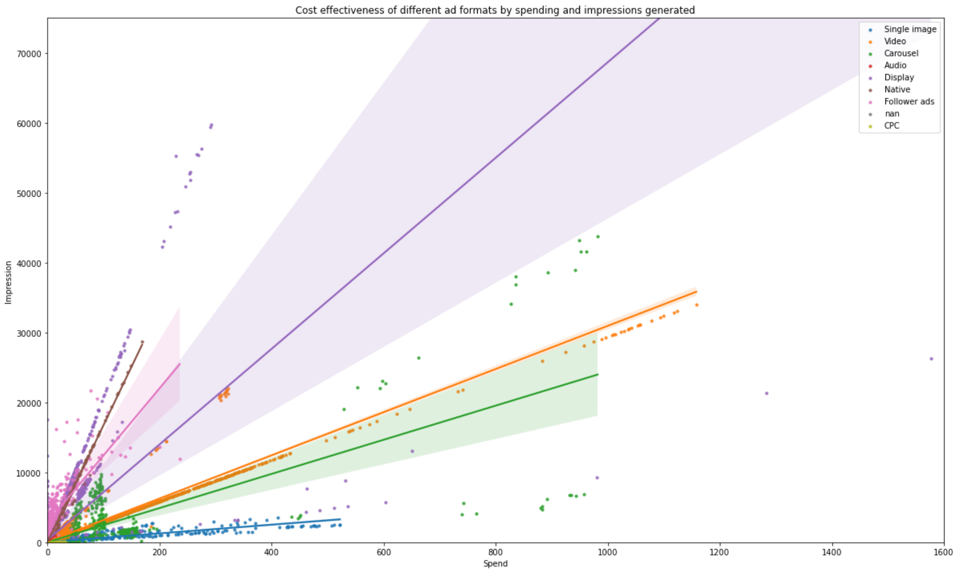


Figure 6.6: Breakdown of the cost effectiveness of various ad formats by impressions generated from the creative spreadsheet.

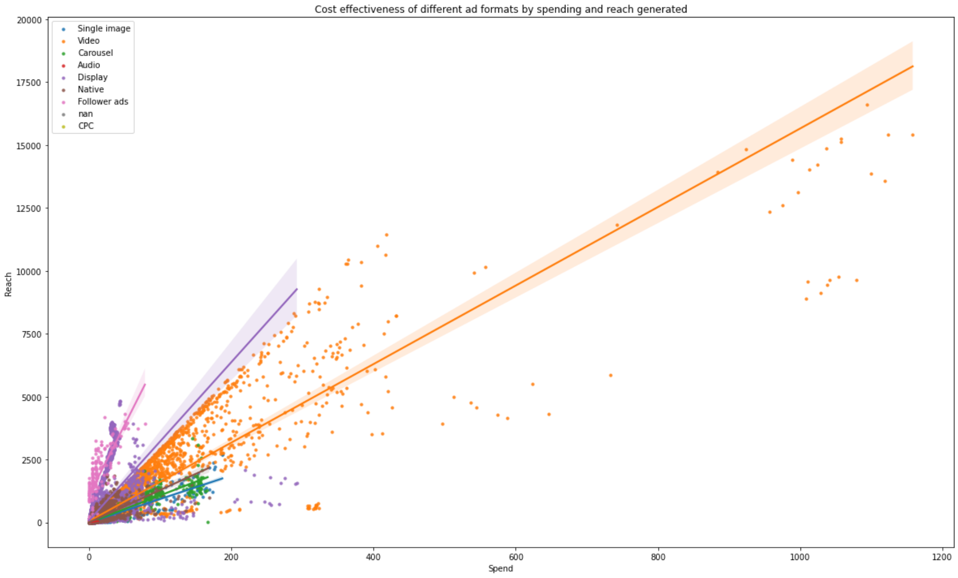


Figure 6.7: Breakdown of the cost effectiveness of various ad formats by reach generated from the creative spreadsheet.

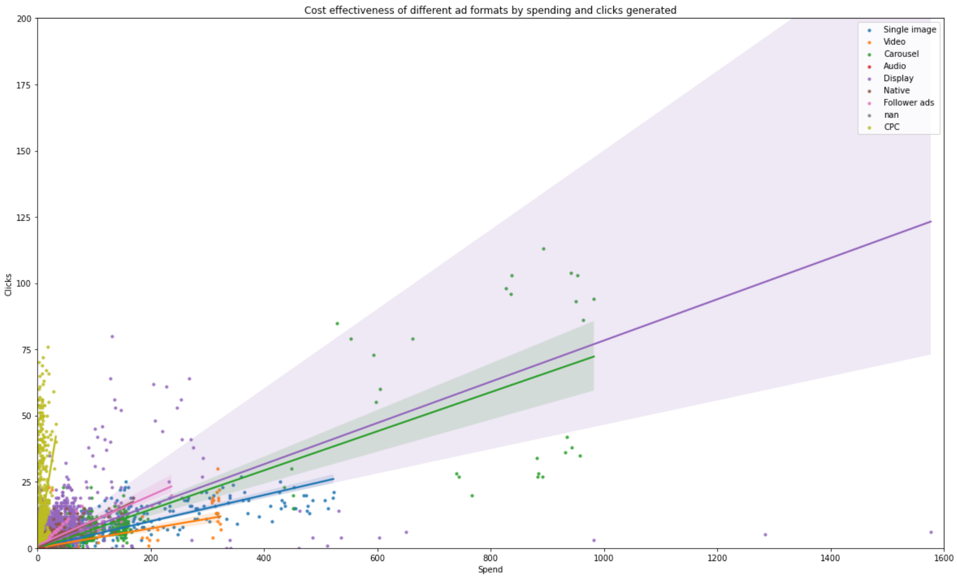


Figure 6.8: Breakdown of the cost effectiveness of various ad formats by the clicks generated from the creative spreadsheet.

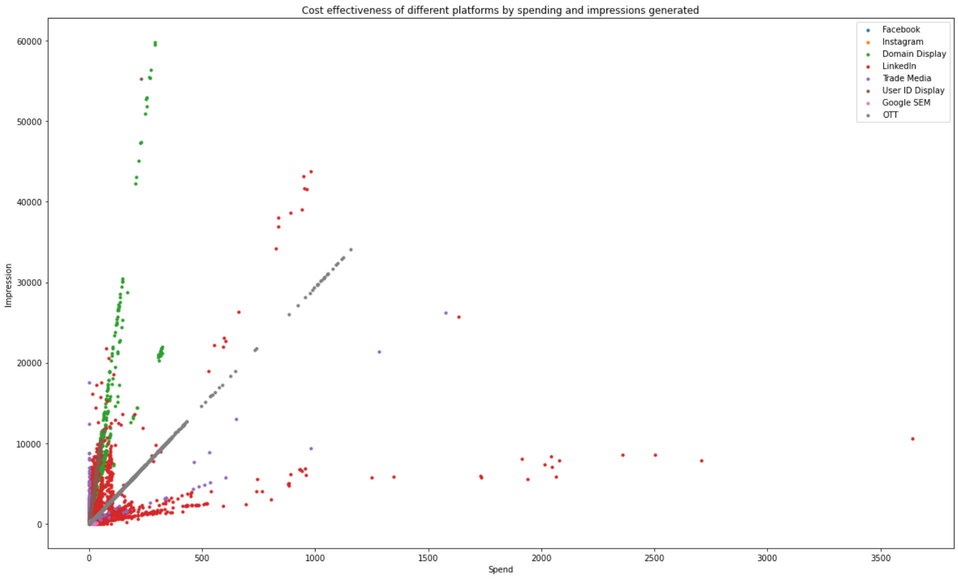


Figure 6.9: Breakdown of the cost effectiveness of various ad platforms by impressions generated from the creative spreadsheet.

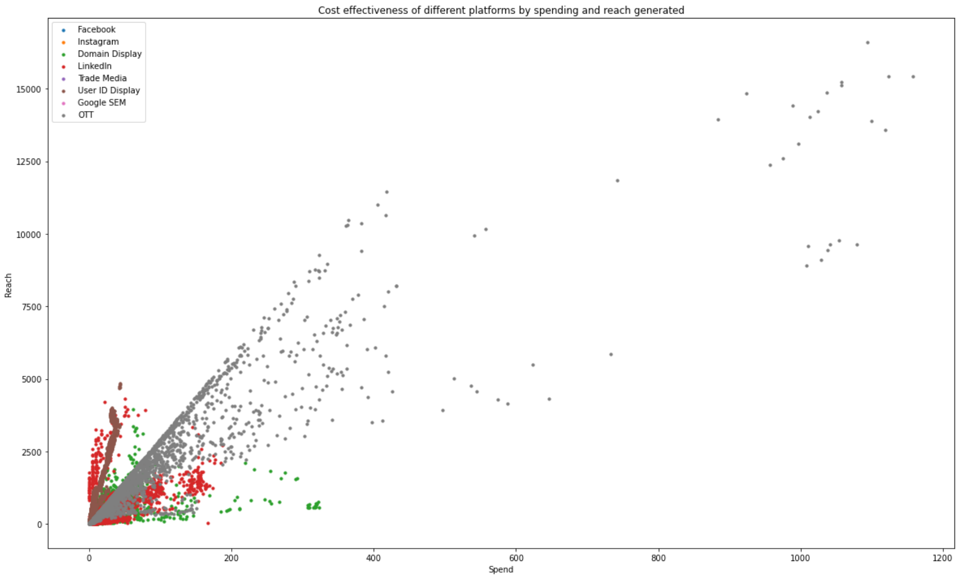


Figure 6.10: Breakdown of the cost effectiveness of various ad platforms by reach generated from the creative spreadsheet.

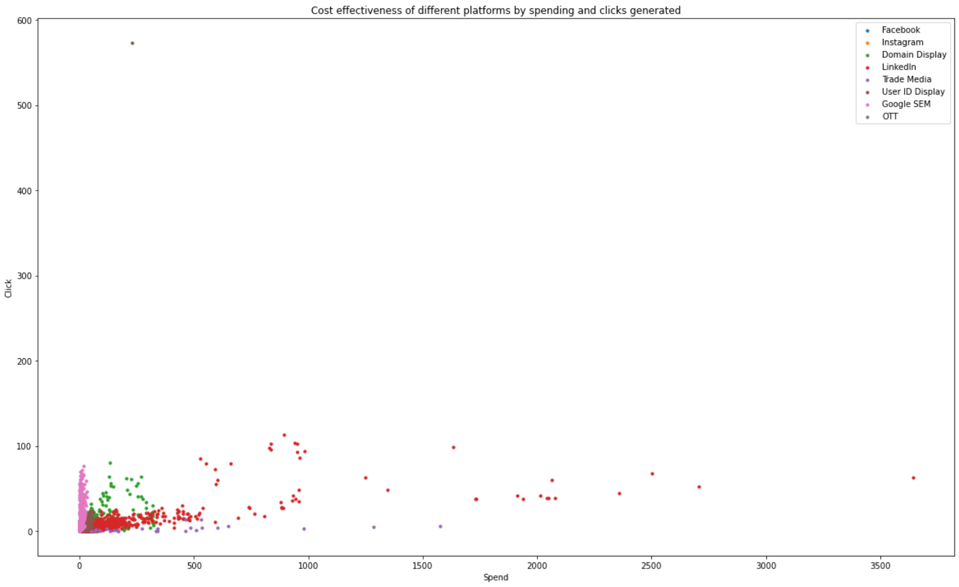


Figure 6.11: Breakdown of the cost effectiveness of various ad platforms by clicks generated from the creative spreadsheet.