Team 5 Employer Project Technical Report

Assignment 3

Alex Cox, Alia S Torreadrado, Michael Gray, Mutale Kondolo

18th September 2023

Contents Page

Final Pitch Recorded Presentation
Linked scripts and documents
GitHub Repository
Word Count: 1625

Background/Context

Since the height of the COVID pandemic, non-qualified mortgages (non-QM) are recovering their share of the USA market[1], with the Government increasing their focus on developing affordable products for consumers. Change Wholesale[2] (CW), the USA's number one non-QM lender, is aiming to increase their financing of underbanked borrowers, through being on the panel for a wider range of mortgage brokers.
This analysis will look to assess how CW can secure new business, increase engagement and reach a broader audience of brokers by answering:

I. How can CW optimise their digital media budget?
II. What role does the type of broker company *and* location have with campaign effectiveness?

Project development process

Project roadmap was created. Datasets were shared and allocated to team members for exploratory, basic cleaning and statistics. Members documented and shared their exploratory insights from their datasets. Data was then harmonised for merging and aggregation purposes and shared. Business questions were allocated to team members and datasets were used accordingly. We created a cleaning list to determine how the team would approach cleaning. Our priority was non-destructive cleaning and to stay true to the data. It was agreed that only obvious duplicates and fully null observations were removed. Decimals and NA values were left as true, as aggregation and visualisation coding can work around these. No variables were removed, separate data frames were created, if/as required. Outliers were studied, but not removed.

Exploratory analysis was carried out to determine variable correlations and relationships. The data was also studied to determine its distribution and normality and suitability for modelling.

Key aggregations and merges

The three datasets ggoals_final, ggeneral_final & creative_final were merged through developing a new column in each that was a concatenation of their common attributes (date, audience, creative_family, creative_version, and platform). Prior to merging, each dataset needed its data aggregated to include only those attributes. Both ggoals_final and ggeneral_final had to remove

[1] https://www.fdic.gov/analysis/household-survey/2021report.pdf
[2] https://changewholesale.com/

general traffic data, as creative_final only had campaign data. Once these datasets were merged, it gave the capacity for integrated visualisations, exploring relationships further and their potential for modelling.

To compare the benefits of one campaign over another and quantify improvements, accounting for any impact across audiences individually (through spend, completions, clicks, reach, impressions and CTR values) and gauge differences in success markers, variables were aggregated from both original datasets creative and google goals. This generated the actual cost of completions/impressions/clicks for each audience by campaign. If we had a conservative revenue figure per completion, we could determine return on investment (related references).

Visualisations

The colour palettes used were often standard palettes in R and Jupyter Notebook or Lab, which do take into account accessibility for visual impairment. Some palettes were changed to match across Jupyter and R programming for presentation purposes. Accessibility was taken into account in terms of title and annotation visibility and using contrasting colours, bold formatting, adequate font size and labelling. Visualisation scales and axes were adapted as required. Where data was at extremes of the scales, separate visualisations were created to allow focusing on both extremes of the data. For example, the underbanked opportunity matrix was created because of limitations faced with the bubble chart created and the visibility of states. The matrix was designed to give greater clarity for CW on the prioritisation of states to target across the USA.

Initial pitch to final pitch process

A presentation was given to the client where initial feedback was provided to the team. This feedback included; quantifying the recommendations, detailing the impact these recommendations have, changes to behaviour over time and general formatting of the visualisations allowed the team to develop the presentation further. To reduce the busy appearance of slides, titles for each business question were added onto separation/introduction slides. As a result of this the final presentation provided clearer insights and recommendations to the client.

Technical overview of the code

R

It is inherently designed for statistical analysis, modelling, customisable and publication quality graphics and visualisations. Although in many ways more intuitive than Python, challenges were that Plotly (interactive visualisations) crashed and slowed the system and required re-coding to Ggplot. Matching aesthetics and palettes across Python and R visualisations for presentations was time consuming and imperfect. Treemaps were more straightforward in Python. Particularly useful libraries were psych(correlation matrices), moments(kurtosis and skew), tidyverse and dplyr(mutations, filters, calculations within code through 'summarize' and grouping of data), corrplot(exploring relationships between data), car(regression models - unfortunately the data was not normally distributed enough for linear regression modelling).

Python

Python was used to work on the Google Analytics CSV's that were provided. This is because Python can handle large datasets and is a versatile language that allows dataframes to be created and modified with the Pandas library. Python allows the data to be sense checked with .info(), .shape () and many other commands to ensure that the data is correct to work with. After the data had been sense-checked new dataframes were created by using an array of commands to answer the objectives such as calculating the sum, count and grouping the data based on certain criteria. These new dataframes were able to be plotted by using the Matplotlib and Seaborn libraries to give a visualisation of the results. These visualisations helped answer some of the key objectives and were used during the presentation stage. Further analysis of the dataframes to provide insight into the statistics was also done to help the team understand the data at a granular level . The new dataframes were exported as CSV's, which were then later used in Tableau and Excel for further data analysis and visualisations. Python also helped merge some of the data that was provided to the team but there were certain limitations as to what could be merged.

In the demographic CSV (*dem_final*), a series of conditions through 'if' and 'elif' statements, enabled categorisation of companies based upon their audiences profile. Companies were broken into groups; six groups where only one of audience 1-6 was in their profile, and a seventh group with multiple audiences in their profile. A series of group by dataframes were created across company groups to analyse clicks and impressions, understanding which companies engaged greatest with the campaign. Exporting these data frames to CSVs allowed for more agile visualisations built in excel (*Company_Audience_Analysis.xlsx*) demonstrating to CW the companies to focus future campaign efforts on.

A decision tree was generated, in the search for a non parametric model option for the data. As the data was not balanced, it would require further smoting. Hot encoding posed a challenge in terms of labelling because there were a large number of categorical variables. Further work on a decision tree might be helpful in better understanding the behavioural side of successful elements. It is considered to be a visually intuitive model for stakeholders also.

Tableau

The newly created CSV file (*Cities_Unique.csv*) merged geolocation data from python libraries to the google analytics CSV, was then imported into Tableau for visualisation of campaign session performance across national, state, and city level. The data was primarily broken down into sessions by state, and further broken down by key variables/headings in the dataset (e.g. Creative family, Audience, etc.) to better understand the levels of activity by state.

A dataset from the Federal Deposit Insurance Corporation website[3], providing the level of underbanked households by state was merged with the number of campaign sessions by state *(Sessions_USA_Population.csv*). Once imported into tableau this allowed for development of a list of states for Change Wholesale to prioritise their future mortgage broker targeting, by developing a Underbanked Opportunity Matrix. Using the index calculations from the dataset, tableau allowed the ability to visualise the comparison between sessions and the population of the underbanked population in order to accurately map out and create the matrix.

Patterns, trends, and insights

In conclusion our top insights and recommendations for Change Wholesale would be as follows:

Spend Distribution

- When analysing audience 6, SEM ads was a standout performer. CW may look to consider utilising it as the sole platform to reach audience 6.
- This left room to investigate the large amount of funds directed at audience 6 through Closer Twins. Redirection of funds from Closer Twins into audience 1-5, who are less responsive to the current marketing strategy may offer better returns.
- Targeting Close Faster for audiences 1- 5, as overall it has a lower cost per completion, reach and click, as well as the best average CTR score across the three main campaigns.

[3] https://www.fdic.gov/analysis/household-survey/data-downloads/index.html

CW could also look to increase advertising spend on Facebook, specifically spend from User ID display due to its high bounce rate, to secure new business with brokers who are already registered (audiences 1-3).

Improve engagement with underbanked population

- CW could target USA States with a higher-than-expected proportion of underbanked households, but with a low level of campaign presence.
- By CW increasing their focus on these states, it has the potential to raise interest with brokers who serve a larger underbanked community.

Enhance company media distribution

- Paid media can be more effectively used to increase CTR for an identified group of ten broker companies.
- The most engaged brokers were Movement Mortgage, EXP Realty, and Guaranteed Rate, contributing roughly 10% of all campaign impressions and clicks.
- CW could more effectively use paid media to increase CTR amongst the identified group of ten broker companies.
- By increasing CTR of these ten brokers to the average of their industry peers, this could increase the total clicks of the campaign by 7%.

Further work

- Complete non parametric modelling to better understand the behavioural dimensions of the marketing campaign effectiveness.
- Breaking down changes in variables (e.g. clicks, campaign by completions) throughout the marketing period in more granular and measured detail in determining accuracy of trends.

Appendix

Data cleaning discussion list

Action	Considerations / detail
Duplicates	Ensure exact duplicates in rows across all columns, not just within one. If you are going to drop columns, it may be better to remove duplicates first and then drop them. Ideally remove duplicates into a separate table that you can explore to see if there are any patterns that might explain what happened in the process or if there are any culprits or associations. There are functions that count how many copies or duplicates there are for each row.
Missing data	Decisions about approach to NA data. This is different to zero results which do carry a meaning. Consider other versions of NA such as '-999'. Beware that removing NAs in many cases will remove other potentially useful data and NAs can be very widespread (in my data every row has NAs). Removing NAs therefore could introduce bias. With a view to limit the introduction of bias and behaving non-destructively with the data, there may be an advantage to working around the missing data – many functions ignore NA data and sometimes you can include syntax to ignore NAs. Some forms of summary statistics can be mischaracterised due to the number of NAs. Some plotting libraries can successfully ignore NAs, others not. Many machine learning logarithms cannot handle NAs. Correlation information can also be skewed. We could consider using logarithms and functions that can handle NAs wherever possible. We can study and incorporate this into technical info and rationale for decisions made. Forecasting can be significantly influenced by NAs. Judicial choosing of variable columns and adapted dataframes for particular studies may be another option.
Data types	Ensure correct data types – sometimes numeric data is presented as characters or string/categorical. Note that some data that appears numeric is actually categorical, like the numbers assigned to audience.
Remove white spaces	Minimise errors due to inconsistency. Checking for unique values will help with this.
Standardise	Standardise syntax in terms of capitals, spaces or underscore for each of coding and minimising errors in coding. Ideally remove spaces in column headings but make sure you technical data shows the change in nomenclature and the necessary references to it. Where there are observations within columns (use unique function) that appear as two different versions of the same thing or are non standardised, standardise – e.g. in my data some observations that referred to groups 5 and 6 appeared as string instead of the majority which appeared as string numbers.
Consistent metrics	Check numeric data and consider the units. If the same observations appear elsewhere – it may be helpful to standardise this also. Hints may be decimal places and the number of digits. This will allow us to compare these results.
Scaling for comparison / normalisaton	Variables with larger scales can dominate or bias modelling and clustering. Scaling can help interpret the importance of features. Allegedly scaling can help improve

	visualising the relationship between variables. The choice of scaling methods depends on the characteristics of the data and how you want to use it. Decision trees and random forests apparently aren't particularly sensitive in this regard.
Outliers	Outliers, skew and kurtosis can affect analysis and modelling. Especially linear regression. Removing them can also introduce bias. Discussion to be held with team – this may be something we address on a case by case basis or across the board.
Drop columns	We have been advised that some columns are not necessary by A3. However, some may prefer to keep them as they have found them useful in the past. In code, it is easy to create a dataframe with all columns and another with dropped columns to meet all needs.
Basic statistics	Including skew, kurtosis, standard deviation and error, range, etc. this data can be very helpful in understanding data spread and frequency of observations within columns or variables. Some are better represented as histograms or bar charts.
Always consider unique values	Check for unique values within variables
Decimals	This is a tricky matter. Some data will essentially be zero with a very large number of decimal places. How important the decimal places needs to be considered, because when you remove decimal places, you lose differences between data. In the spend and CTR columns in creative data, there are up to 6 and over 9 decimal places of data and this data often has three decimal places of zero before it starts having numbers. Remember that CTR is number of clicks divided by the number of impressions. We don't want to end up with no data. An option may be to multiply all the numbers by 1000 or 10,000 and then remove all or minimise decimal places to two. Very large numbers of decimal places also cause problems in terms of visualisation.

Insights

Data Set	feature	insight/observation
	Goal_Stats	***Not including duplicated data and NAs (graphs added to word document in Alex data file)***
	Goals	We need to review each of the goal options and map against the target groups
Goal_Stats	Audience	Audience 6 is the largest group
Goal_Stats	Completions	Audience 6 completed the most number of completions compared to other groups
Goal_Stats	Mean	Audience 6 had the highest average of completions per group
Goal_Stats	Goal/completions	Learn More(Community Mortgage) goal had the most number of completions
Goal_Stats	Platform/completions	The Google SEM platform had the most number of completions
Goal_Stats	Ad format/completions	The CPC ad_format had the most number of completions
Goal_Stats	Campaign/completions	The Brand_Exact campaign had the most number of completions
Goal_Stats	Creative Family/comple	The SEM Ads creative family had the most number of completions
Goal_Stats	Creative version/compl	The Change Wholesale creative_version had the most number of completions
Goal_Stats	Campaign traffic/ comp	General traffic had a lot more completions (24822) than those completed through the campaign (3894)
Goal_Stats	Dates	June had the most number of completions in total (4642completions) and July with the least number (2934)
Goal_Stats	Goal	Learn more(community mortgage) and start closing more completed the most goals in both June & September, with July being the least
Goal_Stats	Campaign	Brand_exact campaign was most completed campaign in June and least in September
Goal_Stats	Ad format	The most successful ad_format CPC completed the most in June and least in April, although there was a sharp increase in completions after april
Goal_Stats	Platform	April had the least number of completions via platforms whereas May and June had the most. Trade media kept reducing from April to 0 completions in august
Goal_Stats	Campaign traffic	July seen a significant drop in completions through general traffic, whereas June had the most.
Goal_Stats	NA	**Further analysis and subsetting depends on decision on duplicates & NA value decision**
Google_Analytics	Audience	Audience 6 accounts for 29% of total sessions, with Audience 4 (27%) then 5 (24%) having the most sessions from April - October
Google_Analytics	Sessions	September and October have the greatest volume of total sessions (51% of total). Potentially as a result of consistent impressions on the audience over seven months? Audience 4 and 5 drive this Sept-October impression growth
Google_Analytics	Campaign	FY23_broker_campaign contributed 55% of total sessions, followed by brand_exact (20%) and FY22_broker_campaign_ph2 (16%). FY23 and ph22 drove nearly all session growth in Sept and October
Goal_Stats	Platform	April had the least number of completions via platforms whereas May and June had the most. Trade media kept reducing from April to 0 completions in august
Goal_Stats	Campaign traffic	July seen a significant drop in completions through general traffic, whereas June had the most.
Goal_Stats	NA	**Further analysis and subsetting depends on decision on duplicates & NA value decision**
Google_Analytics	Audience	Audience 6 accounts for 29% of total sessions, with Audience 4 (27%) then 5 (24%) having the most sessions from April - October
Google_Analytics	Sessions	September and October have the greatest volume of total sessions (51% of total). Potentially as a result of consistent impressions on the audience over seven months? Audience 4 and 5 drive this Sept-October impression growth
Google_Analytics	Campaign	FY23_broker_campaign contributed 55% of total sessions, followed by brand_exact (20%) and FY22_broker_campaign_ph2 (16%). FY23 and ph22 drove nearly all session growth in Sept and October
Google_Analytics	Platform	Domain_Display, Usier_ID display and Google SEM all have a similar volume of total sessions, with LinkedIn at ~ half the level. Domain, User ID, and LinkeIn drive the session volume increases in Sept/October
Google_Analytics	Creative Family	Unfair advantage accounts for 40% of total sessions, with Sem Ads at 26%. CloserTwins & CloseFaster have relatively same volume of sessions. Unfair advantage drives nearly all of session volume gains in Sept/October
Google_Analytics	Ad Format	(Blank) Ad Format has 40% of total sessions, with CPC and Single Image the next biggest with volume of sessions. Single image and (blank) contribute largest increase in impressions in Sept-Oct
Google_Analytics	Audience x Date/month	Audience 1-4 have Sept & October as biggest months with impressions. Audience 5 level of May impressions same as Sept, while Audience 6 impressions highested in May and June
Google_Analytics	Audience x Campaign	Audience 1-5 is exlusively FY22_broker_campaign_ph2 & FY23_broker_campaigns, with ~80% of all impressions from FY23 campaigns. Audience 6 impressions is 70% Brand Exact, with rest mostly through Brand Phase and FY23_Change digital
Google_Analytics	Audience x Platform	User ID is the platform that drives greatest level of impressions form Audience 1-3. With Audience 4 it is split relatively evently between User ID, Domain Display and LinkedIn. Audience 5 is majority Domain, followed by LinkedIn. With Aud6, almost 90% of impressions through Google SEM, with Trade accounting for most of the rest
Google_Analytics	Audience x Ad Format	(Blank) and Single image, the biggest Ad formats in Aud1-5 in terms of impressions delivered. Videos important for audience 1 and 2, carousel & animated for audience 5, with CPC account for almost 90% of audience 6 impressions
Google_Analytics	Audience x Creative Fa	Creative Family portfolio for Audience 1-3 is very similar in terms of impressions, with Unfair Advantages accoutning for 45%, CloserTwins 30% and CloseFaster 25%. Unfair Advantages accounts for 60% of impressions in Aud4-5, while SEM Ads is 90% of impressions for Audience 6.

Google_Analytics	Audience x Campaign	Audience 1-5 is exlusively FY22_broker_campaign_ph2 & FY23_broker_campaigns, with ~80% of all impressions from FY23 campaigns. Audience 6 impressions is 70% Brand Exact, with rest mostly through Brand Phase and FY23_Change digital	Mike
Google_Analytics	Audience x Platform	User ID is the platform that drives greatest level of impressions form Audience 1-3. With Audience 4 it is split relatively evently between User ID, Domain Display and LinkedIn. Audience 5 is majority Domain, followed by LinkedIn. With Aud6, almost 90% of impressions through Google SEM, with Trade accounting for most of the rest	Mike
Google_Analytics	Audience x Ad Format	(Blank) and Single image, the biggest Ad formats in Aud1-5 in terms of impressions delivered. Videos important for audience 1 and 2, carousel & animated for audience 5, with CPC account for almost 90% of audience 6 impressions	Mike
Google_Analytics	Audience x Creative Fa	Creative Family portfolio for Audience 1-3 is very similar in terms of impressions, with Unfair Advantages accoutning for 45%, CloserTwins 30% and CloseFaster 25%. Unfair Advantages accounts for 60% of impressions in Aud4-5, while SEM Ads is 90% of impressions for Audience 6.	Mike
Google_Analytics	Audiene x Creative Ver	There are so many creative versions across the dataset. 39 in total, Need to review further to see if this is of any value, or addtional clustering can be done... May be useful when looking at completions stat only	Mike
Genral Stats	Canmpaign Traffic	Data may be skewed as General traffic accounts for majority of the activity on the success metrics (Total Sessions, bounces and duration)	
General Stats	Creative Family	SEM ads seem have the most total sessions per month. Key to note values that were (not set) or did niot have a specific creative family marked had the highest number of sessions dues all these values being considered under general traffic	Mutale
General Stats	Total Bounces	The most bounces occur in the mojth of June (this includes both campaign and general traffic)	Mutale
General Stats	Total Sessions x Duratic	Audiences 1-5 have unfairadvantage accountign for majority of their sessions, as well as duration spent	Mutale
General Stats	Creative Family x Creat	All enegagament on SEM Ads comes from Genral targetting/ Audience 6, with Change Wholesale creat	Mutale
General Stats	bounces x Audience	Audience 5 had the highest number of bounces . All bounces alo come from the CloseFaster Family on	Mutale
General Stats	Sessions	High across the board in April with a steep decline in May based on Platform and ad_format. There is a	Mutale
General Stats	Bounces	Total bounces are very low on average across audiences with majority of them being condensed into a	Mutale
General Stats	Campaign Traffic	Spike in general traffic in May (potential to explore suggesting a push in ad campaign promotion in Ma	Mutale
General Stats	Platform x Ad_format	Google Sem was the platform with the longest total duration spent on it. Google Sem only has 1 ad for	Mutale
General Stats	Audience x Creative Far	SEM ADS were exclusively exposed to Audience 6	

feature	insight
campaigns, audiences	Not all audiences were explosed to all campaigns
audiences	Some audiences appear to have been more specifically tageted
success (clicks / impressions)	for Audiences 1-3 only a handful of campaigns had any success at all
campaigns / formats	Most of the campaigns (terminology) were only delivered in one format.
impressions	Audiences in order of most impressions (total): 4, **1, 3,** 5, 6, **2**
clicks per audience	Audience in order of clicks: 4,6, **3, 1**, 5, **2**
successs rate (converting impressions to clicks)	Audience in order of success_rate: 6, 5, 4, **1, 3, 2**
successful formats (overall)	If you remove the lowest proproty (ie Audience 6) then there are only three stand-out formats measured by success rate: interactive and animated, with banner far behind
in order of success, by audience:	Audience 1: single image, native, carousel, display, and video
	Audience 2: single image, display, video
	Audience 3: animated
	Audience 4: display - interactive, display, (with middling success from: banner, native; and minor success from: single image, carousel, Dsc)
	Audience 5: interactive, native, animated, banner (with some minor success from other formats)
	Audience 6: native, interactive, Nmn, National Mortgage News, inside mortgage finance newsletter, video, Housing Wire, CPC, Scotsman (and minor success from Single Image)
Campaigns: clicks	Audience 1, 5 & 6:Between August and September there was a drop in clicks, this being a drastic drop for Audience 6
	Audience 2: the effect of Aug - September change was either positive or neutral depending on how click outliers are treated (limited to 2 or 5)
clicks per audience: three plases	There have been three phases: May - Jul, Jul - Sep, Sep -Oct
	Audience 1: succesively FEWER clicks over the whole period, but especially the last phase
	Audience 2: very up and down until September - much more sustained after September
	Audience 3: considerable decrease since August / September
	Audience 4: considerable rise since September (up and up)
	Audience 5: VERY definite drop in clicks, successively, over the three phases
	Audience 6: First two phases quit successful: sudden drop for all of last phase/
Phases: A (May - Jul), B (Jul - Aug), C(Sept- Oct)	

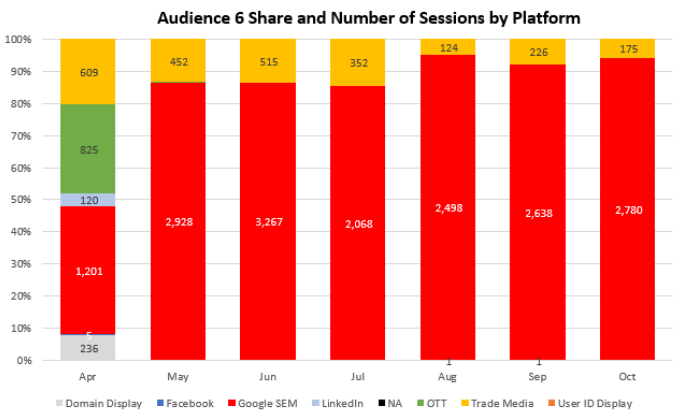
	Audience 2: the effect of Aug - September change was either positive or neutral depending on how click outliers are treated (limited to 2 or 5)
clicks per audience: three plases	There have been three phases: May - Jul, Jul - Sep, Sep -Oct
	Audience 1: succesively FEWER clicks over the whole period, but especially the last phase
	Audience 2: very up and down until September - much more sustained after September
	Audience 3: considerable decrease since August / September
	Audience 4: considerable rise since September (up and up)
	Audience 5: VERY definite drop in clicks, successively, over the three phases
	Audience 6: First two phases quit successful: sudden drop for all of last phase/
Phases: A (May - Jul), B (Jul - Aug), C(Sept- Oct)	
Most significant phases: Up to August, and from September	
Companies	Most successful (non zero) success_rates are very low:
	In audience 1, 334 out of 1487 companies have a success rate above 0. In audience 2, 132 out of 736 companies have a success rate above 0. In audience 3, 338 out of 883 companies have a success rate above 0. In audience 4, 510 out of 4906 companies have a success rate above 0. In audience 5, 1437 out of 4059 companies have a success rate above 0. In audience 6, 1437 out of 5419 companies have a success rate above 0.
	The company definitely makes a difference.
	In Audiences 1-3, The 'a...,' series is the 'slow and steady' campaign series: it rarely has a success rate over about .2, but it is by far the one that is most successful across a much wider range of companies (ie without the a series, very few campaigns would have success, and very few companies would be successfully reached).
	StackAdapt (Audience 1) and the F series (Audiences 1-3) are the campaigns that reach very high success_rates, but in only a very few companies.
	For audience 1, only about 25 companies experience a 1:1 success_rate, and in audiences 2 and 3 it's closer to 10.
	Audiences 4 and 5 have very few companies reaching a high click rate (i.e. 1:1 or thereabouts)
	But audience 6 has relatively very many.

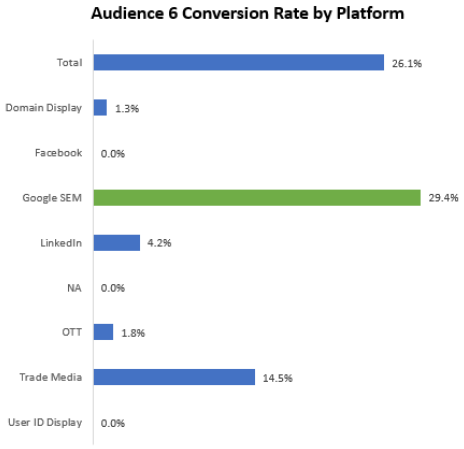
feature	insight	obj1	obj2	obj3	q1	q2	q3
Important flaw	campaign data flawed in this data set - I'm not ready to give up, I need to cross reference with correct values						
clicks and spend	strong positive correlation at 77%. The more you spend, the more clicks. Bear in mind that proportionately some variables do better than others though	yes	yes	yes	yes	yes	maybe
clicks and reach	moderate positive relationship at 60%. The more unique individuals engaged with ad, the more clicks, obviously	yes	yes	yes	yes	yes	maybe
clicks and weighted_CTR_score	71% positive association. As might be expected. Suggests they can be used as proxies for each other.	yes	yes	yes	yes	yes	maybe
reach and spend	58% middling positive relationship - spend does not equate strongly with unqiue number of engagements with ads	maybe	maybe	maybe	yes	maybe	maybe
spend and weighted_CTR_score	65% positive relationship. The more you spend the better your conversion success rate	yes	yes	yes	yes	yes	maybe
impressions and reach	90% positive association. Good proxies for each other. So the more viewing opportunities, the more unique engagements (reach).	yes	yes	yes	yes	yes	maybe
follows and impressions	reach and follows show a 60% positive correlation. Impressions help, but do not predict follows, at least not on their own	yes	yes	yes	yes	yes	maybe
impressions and weighed_CTR_score	91% positive correlation. Impressions are a good predictor or proxy for wieghted_CTR_score. Good for modelling. The more exposure to adverts, the better the weighted_engagement	yes	yes	yes	yes	yes	maybe
follows and reach	60% positive correlation. General estimations of engagement don't necessarily result in follows.	yes	yes	yes	yes	yes	maybe
reach and weighted_CTR_score	92% positive correlation. These are good proxies or indicators for each other.	yes	yes	yes	yes	yes	maybe
follows and weighted_CTR_score	middling 53% positive correlation. Could become part of predictive model for particular audiences and campaigns	maybe	maybe	maybe	maybe	maybe	maybe
negative variable associations	0.21 (negative) association between CTR_score and full_video_views. To a smaller negative degree: a mixture of CTR and CTR_score with impressions, reach, full_video_views, follows. This may be because they are markers of very different marketing options, on differet platforms.	no	no	no	no	no	no

negative variable associations	0.21 (negative) association between CTR_score and full_video_views. To a smaller negative degree: a mixture of CTR and CTR_score with impressions, reach, full_video_views, follows. This may be because they are markers of very different marketing options, on differet platforms.	no	no	no	no	no	no
ad_format and audience	Advert exposure by audience type differs depending on the ad_format. Some ad_formats are very specific to an audience – (audience 6 exclusively targeted through TV, Tablet, mobile, follower ads, desktop and CPC, while the remaining platforms address a mixture of audience groups). Generally, audience 6 have their own exclusive ad_format exposures, except Display, which they are also addressed through with others. It looks like the No Lock Campaign has snuck into ad_platform, so that has been removed now but was only 28 values.	yes	yes	yes	yes	yes	maybe
audience and impressions by platform	Audience 6 has its own exclusive platforms with Google SEM, OTT and Trade Media, although they are also exposed through Domain Display and LinkedIn. The largest amount of exposure as defined by impressions is through Domain Display across the board, followed on by User ID Display. Audience 2 seems to be reached out to the least across audiences. 1 row was automatically removed due to missing values.	yes	yes	yes	yes	yes	maybe
average spend by audience	Highest spend to unaccounted groups due to NAs. Second highest advertising spend on audience 6, followed in descending order by audience 5, audience 4, audience 1, audience 3, audience 2. Lowest advertising spend on audience 2.	yes	yes	yes	yes	yes	maybe
total spend by audience group	3970 rows were removed due to missing values. The highest spend was on the audience 6 group, followed in descending order by audience 4, 5, 1, 3, 2.	yes	yes	yes	yes	yes	maybe
peculiarities audience 1		yes	yes	yes	yes	yes	maybe
peculiarities audience 2	Avg lowest weighted_CTR by audience group. Avg lowest impressions by audience group. Avg lowest reach by audience group.	yes	yes	yes	yes	yes	maybe
peculiarities audience 3		yes	yes	yes	yes	yes	maybe
peculiarities audience 4	average lowest clicks. Avg lowest CTR. Avg highest impressions.	yes	yes	yes	yes	yes	maybe
peculiarities audience 5	on average highest full_video_views.	yes	yes	yes	yes	yes	maybe
peculiarities audience 6	on average NO full_video_views. Average highest clicks. Avg highest CTR. On avg only audience group that follows at all. Highest av weighted_CTR by audience group. Avg highest reach by audience group.	yes	yes	yes	yes	yes	maybe
Top 3 avg campaign cost (spend)	Highest by far compared to other campaign spend and are No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe
Top 3 avg clicks by campaign	No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe
Top 3 avg impressions by campaign	No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe

total spend by audience group	3970 rows were removed due to missing values. The highest spend was on the audience 6 group, followed in descending order by audience 4, 5, 1, 3, 2.	yes	yes	yes	yes	yes	maybe
peculiarities audience 1		yes	yes	yes	yes	yes	maybe
peculiarities audience 2	Avg lowest weighted_CTR by audience group. Avg lowest impressions by audience group. Avg lowest reach by audience group.	yes	yes	yes	yes	yes	maybe
peculiarities audience 3		yes	yes	yes	yes	yes	maybe
peculiarities audience 4	average lowest clicks. Avg lowest CTR. Avg highest impressions.	yes	yes	yes	yes	yes	maybe
peculiarities audience 5	on average highest full_video_views.	yes	yes	yes	yes	yes	maybe
peculiarities audience 6	on average NO full_video_views. Average highest clicks. Avg highest CTR. On avg only audience group that follows at all. Highest av weighted_CTR by audience group. Avg highest reach by audience group.	yes	yes	yes	yes	yes	maybe
Top 3 avg campaign cost (spend)	Highest by far compared to other campaign spend and are No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe
Top 3 avg clicks by campaign	No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe
Top 3 avg impressions by campaign	No Lock Campaigns that are Domain targetted.	yes	yes	yes	yes	yes	maybe
Top 3 avg reach by campaign	HGTV, Pluto, Discovery.	yes	yes	yes	yes	yes	maybe
Top 3 avg full_video_views campaigns	5 - StackAdapt - remarketing - video, audience 4 video, audience 5 video.	yes	yes	yes	yes	yes	maybe
Top avg campaign follows	only two campaigns gerenrated follows: follower ads - demographic targeting, follower ads - domain list	yes	yes	yes	yes	yes	maybe
Top 3 avg CTR by campaigns	Brand_Exact, Brand_phrase, audience 2 CRM - No Lock Campaign - Carousel (updated)	yes	yes	yes	yes	yes	maybe
Top 3 avg_weighted CTR campaigns	Domain targeting NO Lock campaigns	yes	yes	yes	yes	yes	maybe
categorical correlations	96% for campaign and audience, 45% for ad_format and audience, 50% for creative_size and audience, 47% for platform and audience, 32% for creative_family and audience, 47% for creative_version and audience. There is only one strong association with categorical value of audience and that is campaign.	yes	yes	yes	yes	yes	maybe

Audience 6 Conversion Rate and Sessions by Platform

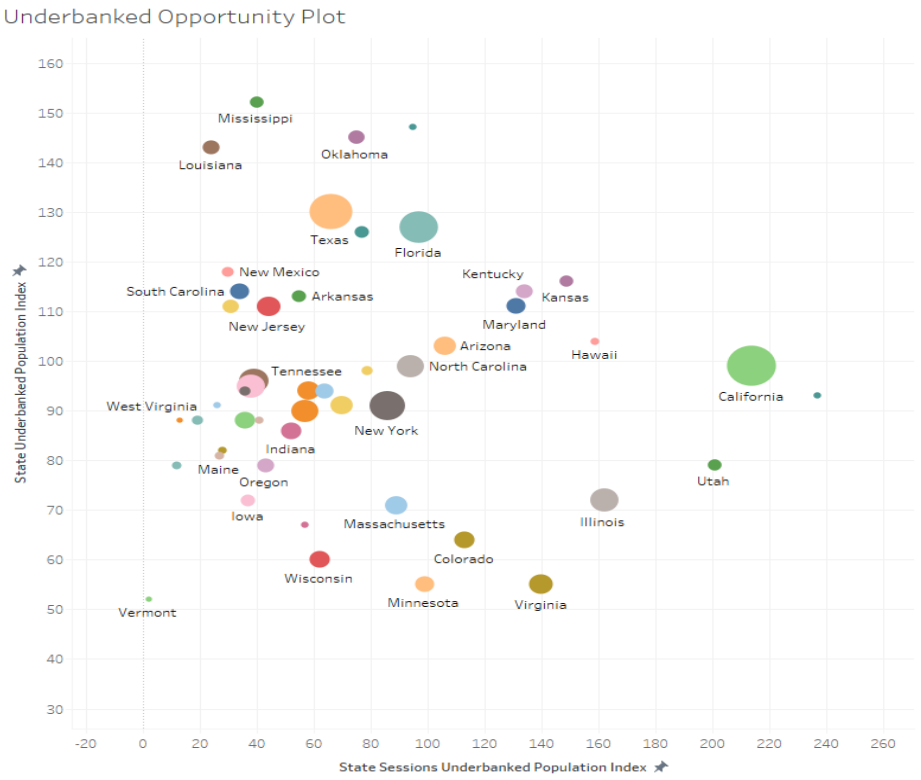




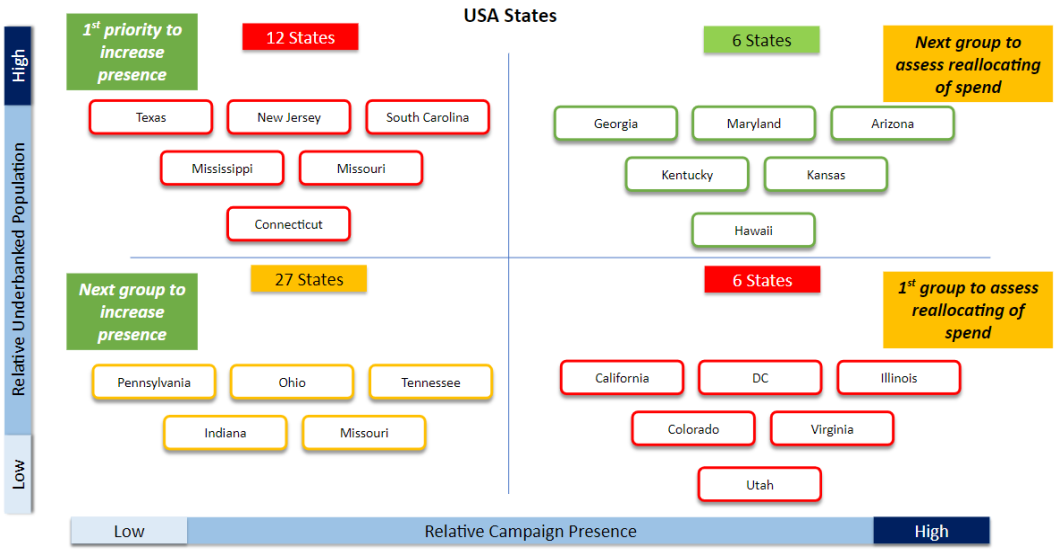
Target Broker Companies

Exp Realty
Guaranteed Rate
Movement Mortgage
American Pacific Mortgage
Nexa Mortgage
Real Estate Ebroker
Crosscountry Mortgage
Finance Of America Companies
American Financial Network
Vip Mortgage

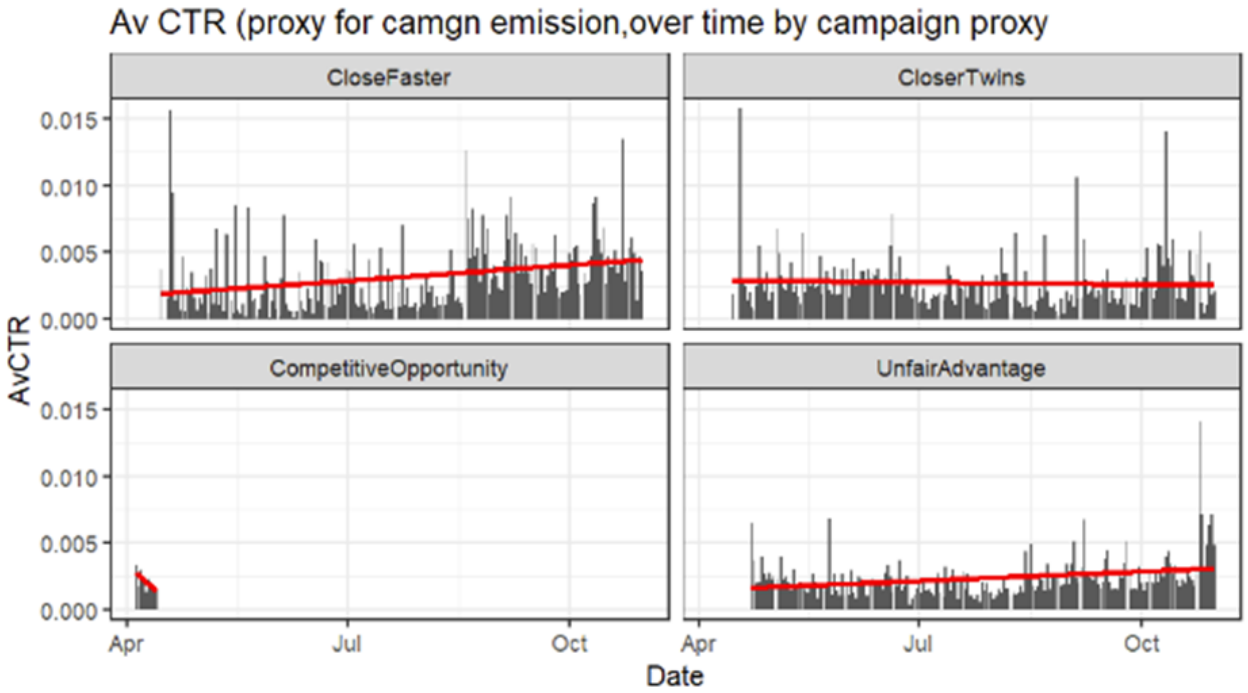
Underbanked Opportunity Plot



Underbanked Opportunity Matrix



Campaign Proxy Over Time



Harmonisation report

Please see 'harmonising_datasets_technical_report' (pdf) included in Github for a full breakdown of the changes made. The script is available as 'harmonising.r'.

Some key harmonisation measures included:

- Checking all variable unique values and if they are obviously different versions of the same thing, matching the syntax and observation entry
- Variables were compared across datasets to determine which datasets could be matched to use together as aggregates, merges and visualisations. Here the decision was made to use creative_family variable as a proxy for campaign
- Changing blank values to NA, which we learnt does not change the csv (for example when used in other programmes)
- Variable and observation data was matched in nomenclature for ease of cross dataset working (all lower case and words separated using underscores)
- Changing data types and formatting of dates (which we learnt does not change the dataset csv, as problems persist when imported in different programs

Related scripts and documents

CSVs	Description
almagamated_spend_campaign.csv	Aggregated data from goals and creative data sets to determine the cost and relationship between variables. Applied to spend per completion, clicks and reaches analysis for business question 1.
Company_Aud_Groups.csv	The csv merged between Company and Company_Audience dataframes that identified how audiences were segmented across companies
Creative_Goal_Merged_2ndSept.csv	The csv developed after the merging of the creaitive_final, ggoals_final, and ggeneral_final datasets
US_City_States.csv	Dataframe developed from geonamescache python library, with city name, latitude, longitude, city population, state name and state code.
GAnalytics_States.csv	Imported csv and dataframe merging between US_City_States.csv and ganalytics_final.csv
Cities_Duplicates_Final.csv	A follow_up from GAnalytics_States.csv, which highlights the cities that have been duplicated to the city name being in multiple states (completed through excel analysis)
Cities_Unique.csv	Identical to Cities_Duplicates_Final, but with all city duplicates removed. This is the datasource for the tableau Cities_Unique_Solution file
Sessions_USA_Population.csv	After creating groupby of total sessions by USA state from Cities_Unique.csv, this was merged with a datasource from the Federal Deposit Insurance Corporation website, which gave number of households and number of underbanked households by state from 2021. This is the datasource for the tableau Underbanked_Population_Solution.twb
creative_duplicates.csv	These are the duplicates exported from the creative dataset being cleaned, for a separate check.

ggoals_final.csv	This is the post basic cleaning (by team members), then harmonised dataset for sharing and business questions. It would have originally been a sheet from Change 2022_GA writeback.xlsx
dem_final.csv	This is the post basic cleaning (by team members), then harmonised dataset for sharing and business questions. It would have originally been Change 2022 Demographic Data Writeback_091122
creative_final.csv	This is the post basic cleaning (by team members), then harmonised dataset for sharing and business questions. It would have originally been Change Creative Data writeback_091122
ganalytics_final.csv	This is the post basic cleaning (by team members), then harmonised dataset for sharing and business questions. It would have originally been a sheet from Change 2022_GA writeback.xlsx
ggeneral_final.csv	This is the post basic cleaning (by team members), then harmonised dataset for sharing and business questions. It would have originally been a sheet from Change 2022_GA writeback.xlsx
creative1.csv	Input dataset for harmonisation
googlegeneral.csv	Input dataset for harmonisation
googlegoals.csv	Input data set for harmonisation
analytics.csv	Input data for harmonisation
dem.csv	Input data for harmonisation

Python Notebooks	Description
Creative, GGoal, GGeneral df Merge_EDA.ipynb	A script that was able to merge the creative_final, ggoal_final and ggeneral_final dataframes through their common attributes. This was to give the potential for integrated visualisations across multiple tools, explore variable relationships further and their potential for modelling.

Q2. States and City Analysis _EDA.ipynb	A script that was able to merge the ganalytics_final dataframe along with the python library geonamescache to evaluate which states and cities had the great level of engagement throughout the campaign
Q2.Company Broker Analysis_EDA.ipynb	A script was developed to look at the segmentation of companies by different audiences, helping to identify which companies delivered the greatest level of campaign engagement and prioritise where CW should focus efforts in the future.
Objective 1-3 Solution.ipynb	A python script providing key insights into objectives 1-3
question_1_solution_python.ipynb	These are the additional tree maps I supplemented r script (questions_1_solution.r) with.
Objective 2_EDA.ipynb	A python script looking at initial insights for objective 2
a3_model_eda.ipnyb	In progress. Decision Tree modelling started. Was intended for eventual SMOTING and completion.

Tableau & Excel Analysis	Description
Company_Audience_Solution.xlsx	The Company_Aud_Group.csv was then developed into excel, with pivots being created to assess share of total companies and clicks by audience profile, plus identifying the biggest opportunity for CW with future campaigns.
G_Goals_final_Solution.xlsx	Analysis from ggoals_final csv, that was able to identify that share of goals completed by audience, and how this evolved over time, and across platforms and creative families.
Goal_General_Merge_Solution.xlsx	Analysis from the Creative_Goal_Merged_2ndSept.csv, that was able to identify completion rate across audiences, platform and creative family, and also compare share of campaign spend by completion level. This

	could identify where value for money was delivered, and where CW could consider rationalising in the future.
USA_States_ Sessions_EDA.xlsx	This excel file could help clearly identify the necessary allocation of states in the underbanked matrix previously described.
amalgamated_spend_campaign_extrapolated.xlsx	This is further aggregation analysis layered over the granular data on amalgamated_spend_campaign.csv
Change 2022_GA writeback_091122.xlsx	Original dataset provided
Change 2022 Demographics Data writeback_091122.xlsx	Original dataset provided
Change 2022 Creative Data writeback_091122.xlsx	Original dataset provided
Objectives 1-3 solution.twb	Several worksheets showing completions by goal, completion by platform, reach by audience and impressions by audience.
Cities_Unique_Solution.twb	Tableau file providing visualisations for the location data. Consists of worksheets looking at sessions by state according to various variables.
Underbanked_Population_Solution.twb	Tableau file providing visualisations looking at the underbanked population by state. These visualisations also compare the sessions by underbanked population.

R script	Description
harmoninsing.r	R script for harmonising datasets before
Goal Stats - Web traffic EDA.r	R script for the initial exploratory data analysis of the provided file
creative_cleaning_eda.r	R script for basic cleaning and exploratory data analysis of creative dataset
question_1_solution.r	Solution to question 1 business question (script) only - does not include the wealth of analysis generated to answer question 1.
question_1_prep_eda.r	This is the preparatory analysis for business question 1 (all the analysis used to determine how to answer business question 1) and question_1_solution.r
question1_eda.r	R script exploratory/prep work to answer business question 2

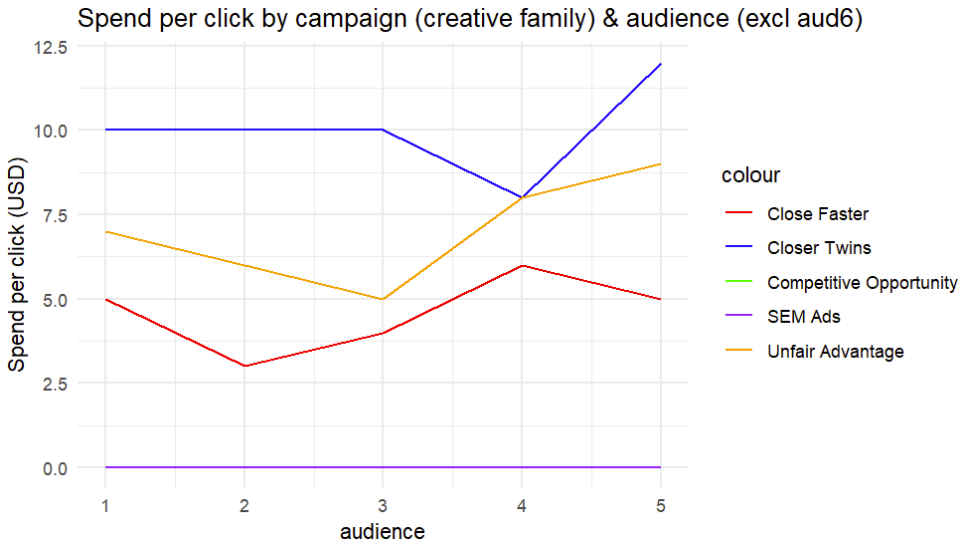
Word or PDF	Description
harmoninsing_datasets_technical_report	technical report of harmonisation process
Mortgage and Media Industry Research	Provides an overview of Change Wholesale, the USA mortgage and underbanked market, as well as insights on marketing effectiveness with advertising
creativedata_cleaning	Summary report EDA and basic cleaning for creative dataset before harmonisation phase

Adaptations to visualisations (examples)

Filtering out data with extreme values that obscure others and manual scale colours for consistency of message in visualisation:

```r
# Spend by click for each audience and campaign (excl closer twins aud 6).
filtered_data <- spend_amalgamated %>%
  filter(audience != "6")

ggplot(filtered_data, aes(x = audience))+
  geom_line(aes(y = ct_spend_by_click,
                color = "Closer Twins"), group = 1) +
  geom_line(aes(y = cf_spend_by_click,
                color = "Close Faster"), group = 1) +
  geom_line(aes(y = ua_spend_by_click,
                color = "Unfair Advantage"), group = 1) +
  geom_line(aes(y = co_spend_by_click,
                color = "Competitive Opportunity"), group = 1) +
  geom_line(aes(y = sem_spend_by_click, color = "SEM Ads"), group = 1)+
  labs(title =
         "Spend per click by campaign (creative family) & audience (excl aud6)",
       x = "audience", y = "Spend per click (USD)") +
  theme_minimal() +
  scale_color_manual(values = c("Closer Twins" = "blue",
                                "Close Faster" = "red",
                                "Unfair Advantage" = "orange",
                                "Competitive Opportunity" = "green",
                                "SEM Ads" = "purple"))
```



Adjusting axis by reformulating very large values (avoid e numbers)

```r
# Re-do facet wrap impressions with better scale axis, so divide by 1000 to
# represent K:
# Impressions over time by campaign proxy.
imp_camp_time <- creative_final %>%
  drop_na(impressions, date, creative_family) %>%
  filter(creative_family %in% c("CloseFaster", "CloserTwins",
                                "CompetitiveOpportunity",
                                "UnfairAdvantage")) %>%
  group_by(creative_family, date) %>%
  summarise(total_impressions = sum(impressions/1000)) %>%
  ggplot(aes(date, total_impressions, group = creative_family)) +
  geom_bar(stat = "identity") +
  theme_bw() +
  labs(title = "Impressions over time by campaign proxy",
       x = "Date",
       y = "Total impressions in thousands (k)") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  facet_wrap(~creative_family)
imp_camp_time
# Care must be taken with this trend lines.  Although they look correct, further
# analysis could be done to confirm by periodic aggregations and by audience
# response within campaigns.
```

Roadmap