Boosting Brand Recognition and Sales by Exploring Facebook Advertising Data

Prepared by:
Andrew Cho
Byoung Chul Hwang
Abhi Opesh
Chethan Shivaram

May 6, 2021

Contents

List of tables

List of figures

Executive summary

This report provides a comprehensive analysis of data from three of our past social media marketing campaigns with the intent of identifying the factors which most strongly contributed to our past marketing success, as well as opportunities for future growth. Data is cleaned, prepared and explored before being used to train and build a suite of statistical predictive models. The exploratory findings indicate that our investment in online marketing activities has ramped up over time, suggesting that the firm is keeping up with a growing trend of digital service adoption. Additionally, landing page conversion rates are found to be commensurate with industry standards, indicating steady performance.

The regression analysis results indicate that the number of impressions generated over the course of an advertisement's deployment is the most significant predictor of greater sales conversion. In terms of demographic characteristics, the results indicate that target audience age is the most significant predictor of greater sales conversion. On the other hand, target audience gender is shown to be of relatively low importance in this regard.

Recommendations are discussed in more detail towards the end of the report. They include: focusing strategic efforts on targeting the right age groups, discounting the importance of targeting specific gender groups, conducting further analyses of CPC and CPM pricing options, and conducting further research into the defining characteristics of Campaign 2.

1 Introduction

A continuing pattern of rapid technological development and the widespread adoption of digital services over the past decade has led to a meteoric rise and expansion of electronic commerce and online business. These disruptions continue to shake the foundations underlying traditional brick-and-mortar stores which have historically been unmoving cornerstones of the retail industry. The ongoing COVID-19 pandemic has only made these transformations more glaringly apparent. Retail giants such as Best Buy and Nordstrom have managed to retain over 60-70% of their sales volume during the past year by leveraging their online sales platforms (Walton 2020). Our firm has sensibly positioned itself thus far to stay ahead of these shifting commercial tides through the early adoption of digital marketing practices. However, we still stand to gain much as an organization by leveraging business analytics to illuminate new opportunities and avenues for growth.

1.1 Business problem

The objective of this analytical project will be to explore data collected from our organization's past Facebook marketing campaigns to identify the factors that most contributed to our past success, as well as areas of improvement for future growth. Identifying the key performance indicators and demographic characteristics that are most consistently associated with higher sales conversion should be instrumental in the development of data driven strategies that will guide our future marketing efforts.

Furthermore, ensuring that we are targeting the appropriate demographics should allow us to build and maintain stronger customer relationships. The importance of preserving customer loyalty has only risen alongside the proliferation of digital marketing channels, as the process of browsing through store to store has quickly been reduced to a simple matter of a few clicks (RetailNext 2014). We will need to be aware of the challenge this poses as we work towards expanded sales and sustained growth.

1.2 Intended audience

The details covered in this project are likely to be of interest to firms across diverse sectors. Consumers of all kinds have long made evident their demand for e-commerce. The COVID-19 pandemic has only further accelerated the need for online marketing practices as social distancing measures, business shutdowns, and stay-at-home orders have led to an increase in contactless shopping and on-demand fulfillment. By mid-April 2020 alone, online orders grew 68%, surpassing 40% of total retail sales, with meaningful gains in categories where digital commerce penetration had been historically low, such as grocery products (Deloitte 2020).

Of course, the results from our analysis will still be specifically tailored to our firm's marketing strategies. As such, it will be difficult for other firms to simply take our findings at face value and apply them with any real degree of success. After all, firms are differentiated by numerous characteristics and typically compete on varying market segments. The analytical procedures we carry out, on the other hand, can probably be tuned, adjusted, and adapted for external use.

2 Data

The dataset we intend to use for this project pertains to three of our past social media marketing campaigns, and contains 1,143 observations across eleven variables. Of these eleven variables, two have been selected and removed for reasons which will be discussed further below. Each row of data can be thought of as a unique advertisement tied to one of three online marketing campaigns, and is further composed of demographic data on the audience to which the ad was shown, and various advertisement performance metrics. Facebook defines an audience as a group of users that matches a set of demographic characteristics specified by the advertising firm (Facebook n.d.).

The first two columns of data within our dataset contain identifiers associated with each individual advertisement and the specific past marketing campaign each advertisement was part of. One column of demographic information corresponds to the average value of the age bracket associated with a particular audience. For ease of future analyses, audiences in their early thirties are assigned the value 32, while audiences in their late thirties are assigned the value 37, and so on. The other column of demographic information contains binary values corresponding to the gender makeup of a particular audience, with a 1 representing a male audience and a 0 representing a female audience. The dataset does not appear to account for non-binary gender identities, perhaps due to its age.

The following are descriptions of the advertisement performance measures included in the dataset. One such column contains information on the number of times a particular advertisement was shown to a particular audience, whereas another contains information on the number of times a particular audience clicked on a particular advertisement. A subsequent column of data pertains to the dollar amount our firm paid Facebook in order to have a particular advertisement delivered to a particular audience. The last two columns correspond to the number of total and approved conversions ultimately generated by each advertisement deployed by our firm.

2.1 Data collection

This dataset was obtained through one of our firm's data exports onto an online data repository (Gokagglers 2017). The combination of qualitative and quantitative information contained within the dataset should be quite suitable for the analytical processes our team plans to carry out. The subsequent analysis should lead to the discovery of valuable insights into effective advertising and marketing strategies. Furthermore, the dataset appears to be of high quality, containing data that is clean and easy to interpret while also being free of any missing information. The expectation is that the data should allow us to infer which key performance indicators most strongly contribute to the performance of a marketing campaign.

2.2 Data preparation

Our team tackled the task of preparing the dataset by first checking whether it contained any missing information that we would need to address. This is important because the presence of missing data can pose a number of serious problems for our data analysis going forward (Kang 2013). Fortunately, we were able to confirm that our dataset was complete and free of any missing data. We then proceeded to convert the data within the columns into formats that would be easier to manipulate and interpret through the use of statistical software. In the case of target audience age, audiences in their early thirties (30-34) are assigned the value 32, while audiences in their late thirties (35-39) are assigned the value 37, and so on. Target audience gender was also converted into a binary number format with a value of 1 for male and 0 for female.

Next, we removed two columns containing redundant and meaningless data that were not useful to us. A column containing Facebook's advertisement identification numbers was removed because the first column in our dataset already provides a unique per-advertisement identification number. A column containing numeric codes which Facebook assigns to its users based on information available on their public profiles was removed because the codes were impossible to interpret in the context of our dataset. Without any information on the products being advertised in the anonymized marketing campaigns and the interest groups associated with these codes, the values proved essentially meaningless. We encountered no further considerations regarding our data.

3 Descriptive analytics

The objective of this analytical project is to explore different KPIs and identify which ones contribute most strongly to the success of a marketing campaign. However, not all marketing campaigns are deployed in an identical manner, and it may not be appropriate to run a single model on data aggregated from all three of the campaigns if their characteristics are substantially different. We used a series of box plots to explore some quantitative characteristics associated with each marketing campaign in order to identify any potential differences therein. We also transformed some of the variables for the sake of the visualizations. This was deemed worthwhile as the visualizations would otherwise have been heavily skewed due to some irregular data in Campaign 3.

Figure 1 below indicates that there appear to be substantial differences in the characteristics of each of the three campaigns. Campaign 3 appears to have consistently enjoyed the most user engagement in terms of clicks, impressions, and conversion, while also having the highest level of per-ad spending. This seems to set Campaign 3 apart from Campaigns 1 and 2, which seem more similar to each other in comparison.

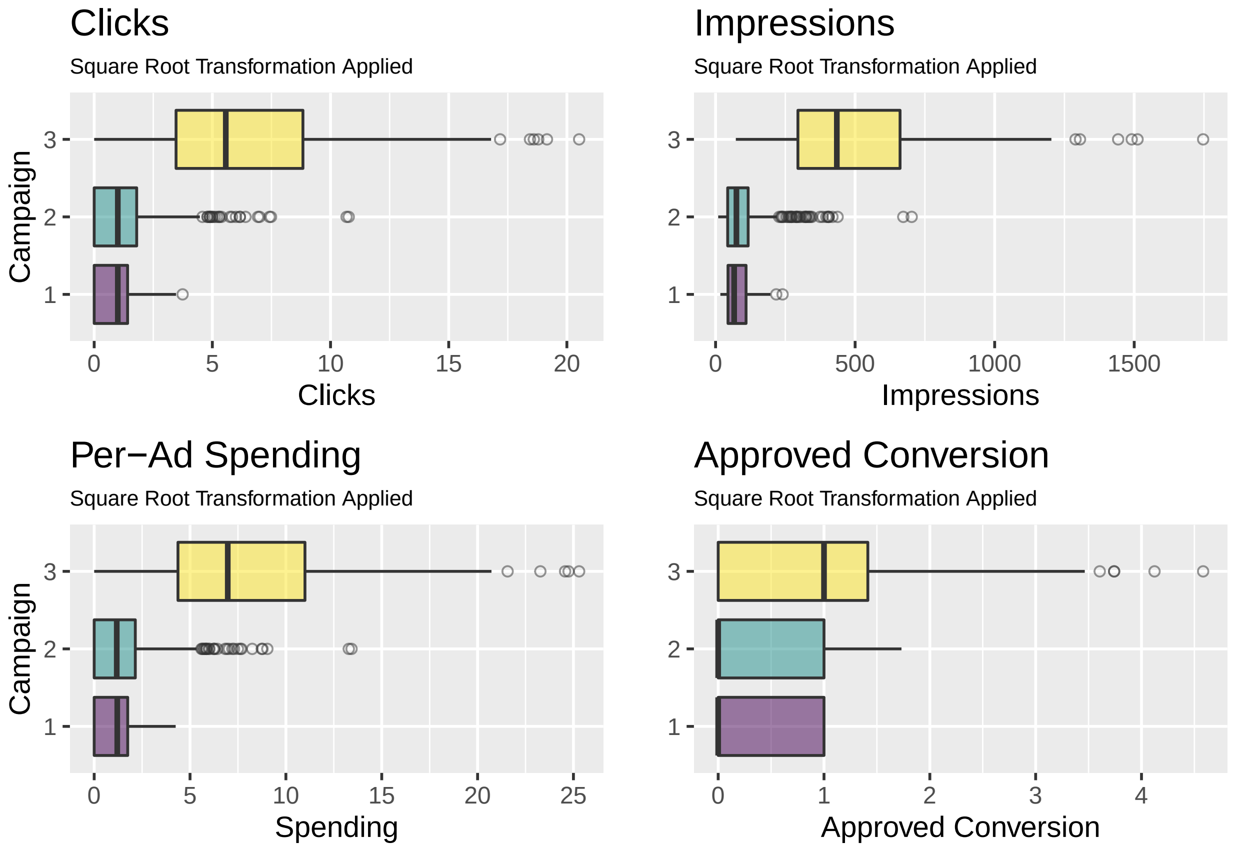


Figure 1: Box Plots of Selected Campaign Characteristics

In addition to the boxplots above, a table detailing the number of advertisements per marketing campaign is included below to further highlight a potential difference between the three marketing campaigns.

Table 1: Advertisement Counts Per Campaign

Campaign	Count
1	54
2	464
3	625

Table 1 indicates a substantial disparity in advertisement count between the three marketing campaigns. Campaigns 2 and 3 have counts in the hundreds, whereas there are only 54 advertisements in Campaign 1. We will need to keep this in consideration when interpreting our results, as this disparity may introduce undesirable bias and variability.

Next, we used a scatter plot of the various quantitative variables in our data to explore the relationships between the variables. Here, the variables are not differentiated by campaign as we wanted to explore the overarching interactions between the different KPIs, which will aid us in later analyses.

Figure 2, displayed below, indicates a positive relationship between all of our quantitative variables. The strongest relationships appear to be between the amount paid to have a particular advertisement delivered and the number of clicks and impressions generated by the particular advertisement. The weakest relationship seems to be between the number of times a user within a particular audience clicked on a particular ad and the resulting number of conversions, or users who ultimately bought the product after seeing the ad.

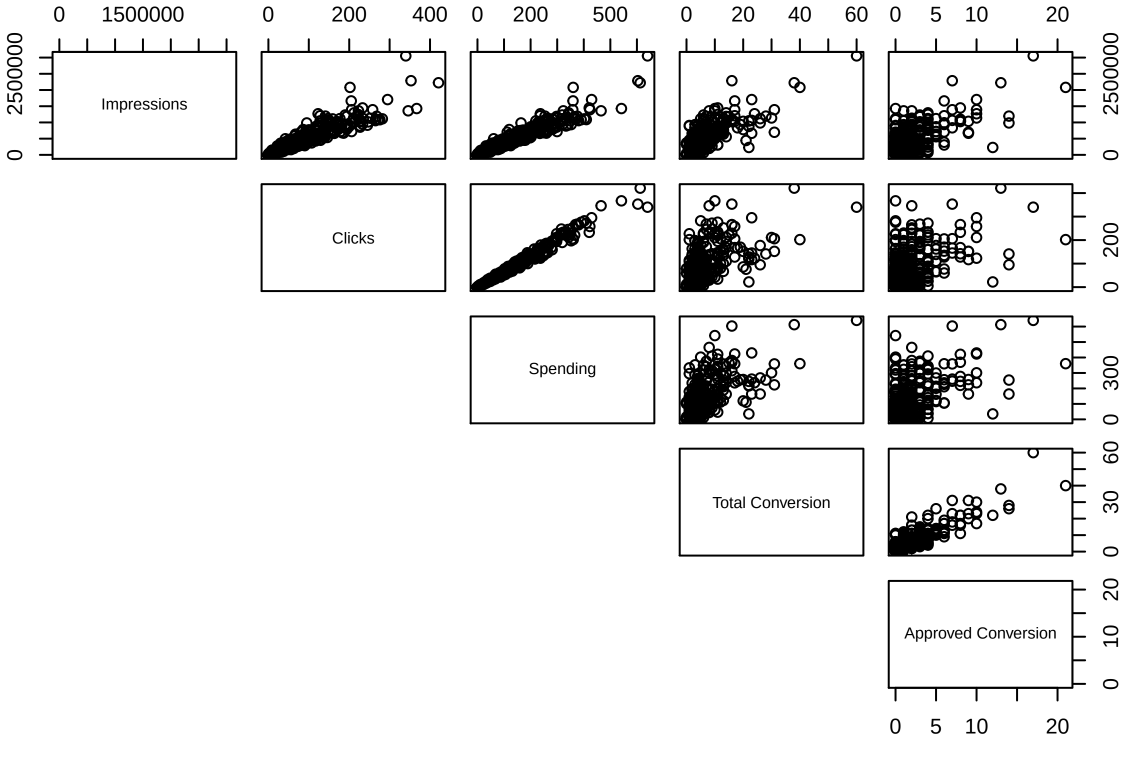


Figure 2: Correlation Plots of Quantitative Predictors

In addition to the scatter plots above, a correlation matrix is included below to augment our exploration of the quantitative variables.

Table 2: Correlation Matrix of Quantitative Predictors

	Impressions	Clicks	Spending	Total Conversion	Approved Conversion
Impressions	1.000	0.949	0.970	0.813	0.684
Clicks	0.949	1.000	0.993	0.695	0.560
Spending	0.970	0.993	1.000	0.725	0.593
Total Conversion	0.813	0.695	0.725	1.000	0.864
Approved Conversion	0.684	0.560	0.593	0.864	1.000

Table 2 provides a matrix of correlation coefficients associated with the variable relationships depicted in the preceding scatter plots. This allows us to more confidently compare the relative strengths and weaknesses of the variable relationships. For instance, we can now assert that the strongest positive correlation between our quantitative variables can be found between per-ad spending and the number of user clicks generated by a particular advertisement.

Lastly, we used a series of bar plots to explore the demographic distribution of the various audiences the ads were presented to. Figure 3 indicates that the 30-34 age bracket occurs most frequently across all audiences, which suggests that it may be the firm's primary target demographic. At a total of 426 instances, this demographic occurs almost twice as often as any of the other age groups. The 45-49 age bracket comes second, with a total of 259 instances. It is also worth noting that the firm tends to target more male audiences than female audiences.

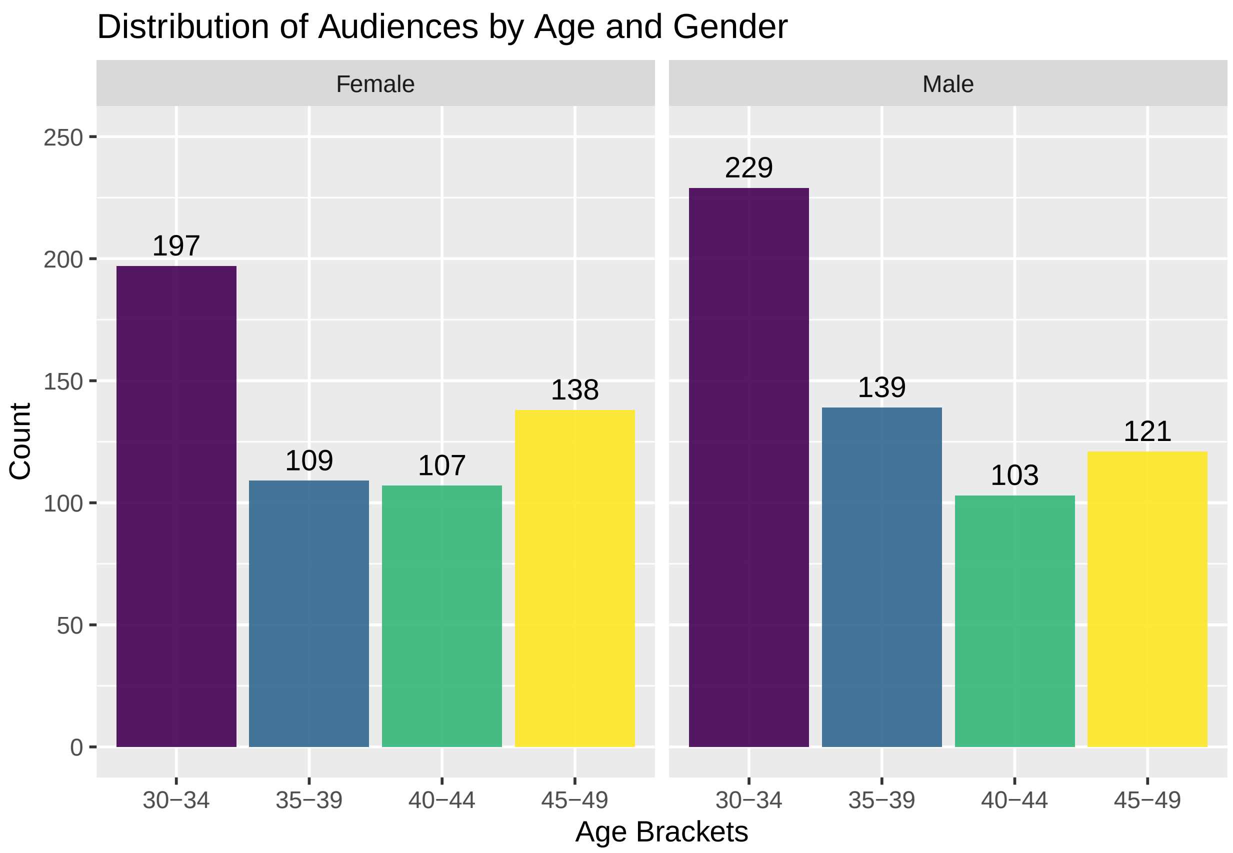


Figure 3: Bar Plots of Age and Gender Distributions

3.1 Insight summary

As previously stated, Figure 1 and Table 1 indicate that the characteristics of the three campaigns appear to be substantially different. Campaign 3 demonstrates the highest median value and variability across most of the selected quantitative performance metrics. It also appears as though the firm invested the most money into Campaign 3 given that the deployed advertisements tended to have higher levels of per-ad spending. As a result, Campaign 3 seems to have achieved a higher level of user engagement, though it would be too early to draw any causal conclusions at this point. It is worth noting that there is a large degree of disparity between the number of advertisements deployed within each campaign. As previously mentioned, this may have interesting statistical implications for our predictive results going forward.

Figure 2 and Table 2 show positive relationships between each of the quantitative variables within this data set. The very strongest positive relationship is found between the amount spent on a particular advertisement and the number of clicks generated, with a correlation coefficient of 0.993. The next strongest positive relationship can be found between the amount spent on a particular ad and the number of impressions generated, with a correlation coefficient of 0.970. These findings appear to accord with Facebook's cost per link click (CPC) pricing model, whereby advertisers are charged solely based on the number of times their ad was clicked on (WebFX 2021). Additionally, a further inspection of our data showed that there were no advertisements with zero clicks generated for which spending was nonzero. This supports the theory that the firm chose to pursue a CPC pricing model because it suggests the firm did not pay for advertisements that did not generate any clicks. As such, it would make sense that the highest correlation coefficient is associated with the relationship between per-ad spending and number of clicks generated.

This investigation is relevant because there are a wide variety of pricing models that Facebook offers to advertisers. One alternative pricing model is based on cost per thousand impressions (CPM), whereby advertisers are instead charged based on the number of times their advertisement is shown on a user's screen (Facebook 2020). However, we know now that the firm did not elect to choose a CPM pricing model, which indicates that the relatively high correlation coefficient between per-ad spending and impressions is not directly related to the firm's marketing campaign design. Rather, it may simply be the case that increasing spending on an advertisement with the goal to generate more clicks also increases its visibility to users, which might intuitively lead to a greater number of impressions. Since our team is interested in uncovering less intuitive insights from our firm's data, it may be beneficial to control for per-ad spending in further analyses. This might allow us to identify other variables that might lead to higher values of impressions, clicks, and other KPI's.

On the other hand, the weakest positive relationship is noted between the number of times a user within a particular audience clicked on a particular ad and the number of resultant conversions. This finding seems to be consistent with landing page conversion rates across a majority of industries. Studies have found that these conversion rates tend to be quite low, ranging from 2.6% to 6.1% (Unbounce n.d.). As such, we might expect a relatively weaker link between advertisement clicks and conversions towards actual sales.

The distributions depicted in Figure 3 provide a few interesting insights into what the firm's preferred target demographics may be. When gender is factored in, the most frequently occurring demographic is associated with audiences consisting of males aged 30 to 34. Audiences consisting of females aged 30 to 34 come at a close second. The three remaining age demographics occur with relatively similar frequency, with roughly 200 to 250 audiences associated with each. It may prove interesting to investigate the effectiveness of the firm's supposed targeting strategy. This could conceivably be accomplished by comparing the level of spending devoted to each demographic group as well as their respective conversion rates.

4 Predictive analytics

The specific goal of our predictive analytics is to observe how a variety of factors contribute to the performance of online marketing campaigns. In particular, we are focused on determining which factors have the greatest impact on campaign success. It is worth noting that firms that are interested in boosting their brand awareness may find total conversion to be the most suitable measurement of ad performance. On the other hand, firms that are more interested in boosting sales may find approved conversion to be the more appropriate metric. Therefore, we intend to explore the impact of the various features in our dataset on two separate response metrics — total conversion and approved conversion. We will build out linear regression and random forest models for the two distinct response metrics to achieve these goals.

4.1 Process

The predictive analytics process began with the splitting of our full dataset into smaller representative subsets for the purpose of training and evaluating the predictive models that we developed. After carrying out this step, we considered our methodological options and decided to implement multiple linear regression and random forest models in our analysis. We decided to proceed with these methodologies because they allow for the use of all types of variables to predict the outcome of the given quantitative response variables.

For both methods, an initial set of two separate models — one for each response variable — containing all the primary variables was built. The initial models included the variables of target audience's age and gender, a firm's per-advertisement spending, the specific marketing campaign associated with the ad, and the number of clicks and impressions generated over the course of an advertisement's deployment. One model was built to predict approved conversion and another was built to predict total conversion.

Next, variables that may not provide useful information for accurate prediction were removed to create different models that were later compared to the originals. These models were evaluated through various assessment methods that measure model prediction error. Finally, the models with the lowest prediction errors were chosen as the final models in our analysis.

4.2 Assessments

We first utilized the linear regression method in our analysis. Figure 1 and Figure 2 below depict moderate to strong linear relationships between the different key performance indicators and our two response variables — total conversion and approved conversion, respectively. Additionally, we did not encounter any grievous violations of the relevant model building assumptions, so we decided that it would be appropriate to proceed with our analysis.

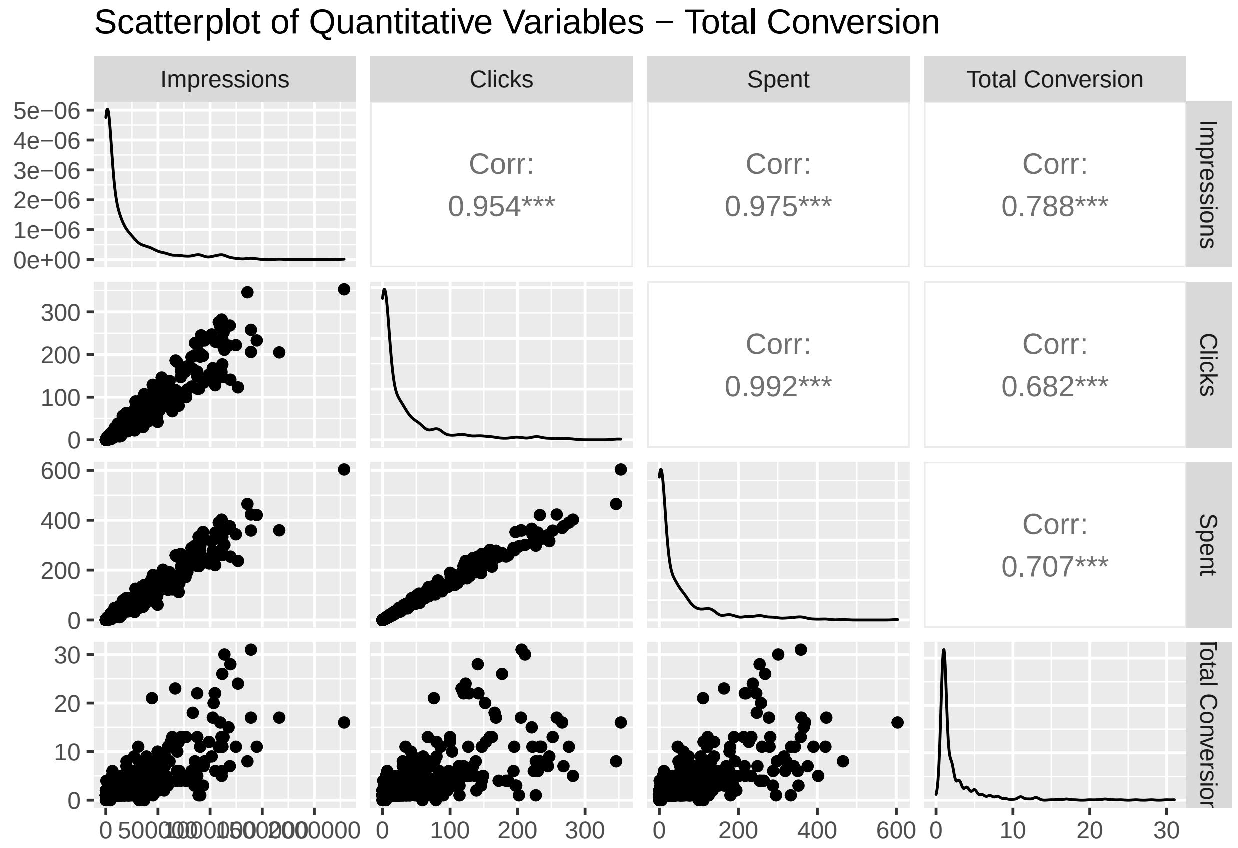


Figure 4: Correlation Plot of Quantitative Predictors and Total Conversion

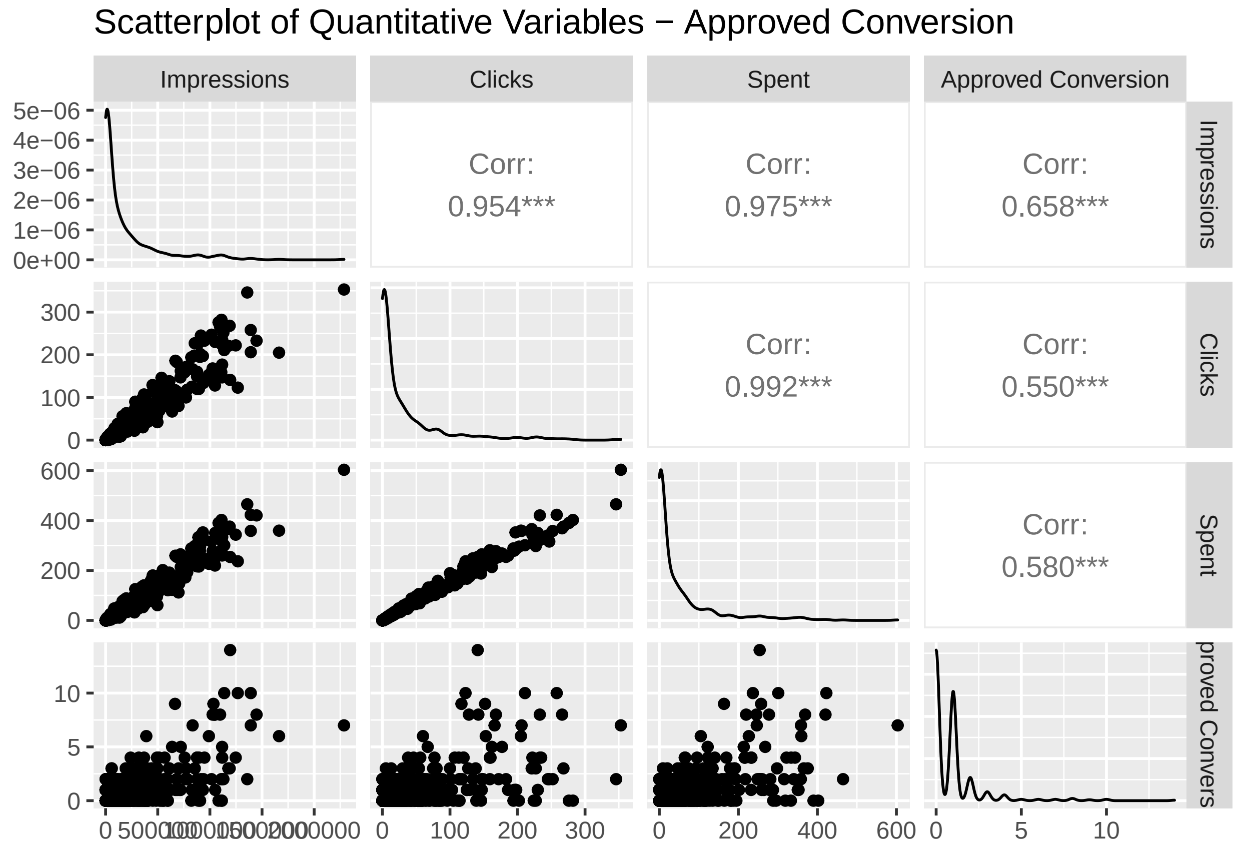


Figure 5: Correlation Plot of Quantitative Predictors and Approved Conversion

Upon conducting a regression analysis on our initial model, we saw that only a handful of predictors were significant. We hence iteratively removed the least significant variables from the model and conducted overall significant tests to approach our final models for assessment. We were left with a pair of models for both of our response variables.

The model that had the highest proportion of variance explained in predicting total conversion, also known as R-Squared, had high degrees of multicollinearity. We decided to remove the variable that most contributed to this undesirable intercorrelation within the model. This resulted in a final model with more manageable levels of multicollinearity, though this came at the cost of a slightly lower R-Squared value.

Similarly, the model predicting approved conversion that had the highest R-squared value also had high multicollinearity. We once again compared this model to one without the variable that contributed to the high multicollinearity to result in a model with little to no multicollinearity but a far lower R-squared value.

The full and reduced models for each response variable were then assessed based on their respective prediction error, measured by the root mean squared error values, or RMSE. A lower RMSE is indicative of a model with better fit to the data. While our goal does not necessarily lie in predicting some value, these assessment measures are important because they give us valuable information regarding the soundness of our models. Since we are interested in making inferences about the relationship between our two response variables and the selected predictors, we want to be using the best possible models as a basis for drawing these conclusions.

Table 3: Linear Regression Model Assessment - Validation Set

Model Name	RMSE
Initial - Approved Conversion	4.60
Final - Approved Conversion	5.39
Initial - Total Conversion	2.14
Final - Total Conversion	2.21

The table above summarizes the linear regression model assessment results based on the validation data. The initial models for total conversion and approved conversion result in lower prediction errors and thus seem to better fit our data. We hence continue with these models as our final models of choice.

Table 4: Linear Regression Model Assessment - Testing Set

Model Name	RMSE
Initial - Approved Conversion	4.32
Initial - Total Conversion	2.86

The table above summarizes the linear regression model assessment results based on the testing data.

Next, our team conducted a random forest analysis. We began by building an initial pair of models containing all of our primary predictor variables. Next, we built a set of simpler models by sequentially removing variables with very low variable importance. The goal of this procedure is to discover an alternative model that demonstrates comparable performance in assessments while also being simpler and easier to interpret. In our case, this involved removing the predictor variables associated with the target audience's gender, as well as the variable associated with the specific marketing campaign each ad belonged to.

While these simpler models may be slightly easier to interpret, our team was also careful about their performance relative to the full, initial models. We found that any further variable removals resulted in drastic reductions in various model assessment metrics. Additionally, all of the remaining predictors indicated high variable importance during our exploratory assessments. Therefore, we retained these remaining variables in our final models and made no further changes. Each of these models were once again assessed based on their respective prediction error by looking at RMSE values.

Table 5: Random Forest Model Assessment - Validation Set

Model Name	RMSE
Initial - Approved Conversion	1.656308
Reduced - Approved Conversion	1.659218
Initial - Total Conversion	3.776401
Reduced - Total Conversion	3.859157

The table above summarizes the random forest model assessment results based on the validation data. The initial approved conversion and initial total conversion models demonstrated lower prediction errors. As such, our group decided to select these two models as our final random forest models of choice.

Table 6: Random Forest Model Assessment - Testing Set

Model Name	RMSE
Initial - Approved Conversion	1.430102
Initial - Total Conversion	3.023581

The table above summarizes the final random forest model assessment results based on the testing data. Interestingly enough, the two final models demonstrate lower prediction error here relative to the earlier assessment results.

4.3 Results

As seen in the tables below, our final linear regression model indicates that the audience's age, and the number of clicks and impressions generated over the course of an advertisement's deployment, a firm's per-advertisement spending, and the specific marketing campaign associated with the ad are important predictors for total conversion. The number of impressions generated over the course of an advertisement's deployment, a firm's per-advertisement spending, and the specific marketing campaign associated are important variables in predicting approved conversion. While a few of these variables turned out to be individually insignificant, we found that the models that excluded these relevant variables were not as adequate.

Table 7: Total Conversion Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1955752	0.6138945	3.576470	0.0003730
Age	-0.0462674	0.0154999	-2.985007	0.0029376
Impressions	0.0000274	0.0000015	17.861137	0.0000000
Clicks	0.0346849	0.0133889	2.590566	0.0097875
Spent	-0.0802075	0.0119990	-6.684540	0.0000000
Campaign 1	0.5142357	0.3959088	1.298874	0.1944281
Campaign 2	0.5670039	0.1998993	2.836447	0.0046978

R2	Adj..R2	P.Value
0.7556	0.7535	< 2.2e-16

Table 8: Approved Conversion Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0154164	0.0755958	0.2039312	0.8384682
Impressions	0.0000102	0.0000006	16.1993751	0.0000000
Spent	-0.0229321	0.0021301	-10.7659380	0.0000000
Campaign 1	0.4717083	0.1938542	2.4333148	0.0152175
Campaign 2	0.3822973	0.0977004	3.9129551	0.0001003

R2	Adj..R2	P.Value
0.6098	0.6075	< 2.2e-16

Our final model for total conversion results in a R-squared value of 0.7556, meaning that about 75.56% of the variation in total conversion is attributed to our model. Similarly, the final model for approved conversion results in a R-squared value of 0.6098, meaning that about 60.98% of the variation in approved conversion is explained by our model.

The results of our random forest model analysis are summarized in the table below. The leftmost column contains our final set of predictor variables, listed in order of decreasing importance. The values in the middle and rightmost columns are measures of each variable's relative importance in each of our two random forest models. Higher values indicate greater variable importance, as these values represent a percent increase in prediction error that would result from the removal of each respective variable.

Table 9: Random Forest Variable Importance - Higher Values for Higher Importance

Variables	Importance - Approved Conversion	Importance - Total Conversion
Impressions	21.667	23.868
Spent	15.536	17.291
Clicks	15.235	15.348
Age	13.526	18.842
Campaign 2	9.747	12.147
Campaign 1	-0.153	5.464
Male	4.458	2.760

We find that the number of clicks and impressions generated over the course of an advertisement's deployment, and a firm's per-advertisement spending all appear to be important variables in both the approved conversion and total conversion models, as one might readily intuit. The less intuitive result lies within the relative importance of the demographic variables, as well as the campaign variables. The age of an advertisement's target audience appears to be a relatively important factor in generating greater conversion, whereas target audience gender does not. Furthermore, it seems as though information about an advertisement belonging to the first marketing campaign is far less important than information about an advertisement belonging to the second marketing campaign.

4.4 Insight summary

Based on the assessments we ran, we have information regarding the variables that affect marketing campaign success the most.

The number of times an ad was shown to the particular audience, i.e. the impressions generated over the course of an advertisement's deployment is the most significant predictor of both total conversion and approved conversion as indicated by linear regression. Impressions also consistently demonstrated the highest level of variable importance across the two final random forest models. This suggests that a marketing campaign's success as measured by total conversion and approved conversion largely hinges on the number of times the ads appear on user screens. This is somewhat intuitive, since ads that are shown more frequently have greater visibility and are therefore likely to induce more users to inquire about and ultimately purchase a product.

Per-advertisement spending is the second most important predictor of both total and approved conversion as concluded by both of the methodologies we employed. It is interesting to note, however, that an increase in the amount of money paid to deliver an advertisement actually decreases both total and approved conversion as per linear regression. Therefore, minimizing overall advertising costs, or at least making sure these funds are being used efficiently, seems to be beneficial for the success of a marketing campaign.

We found the number of times a user clicked on a particular ad to be a significant variable in our total conversion linear regression models. It was also a variable of relatively high importance in both of our random forest models. We were surprised to see that it was not individually significant in predicting approved conversion in our linear regression models, especially when considering our firm's utilization of a CPC pricing model. We believe that this may be due to the fact that the advertisements with the most conversions also tended to generate a large number of impressions and a relatively lower number of clicks.

The age associated with the audience of a particular advertisement is a significant variable in our total conversion linear regression model. We find a negative relationship between age and total conversion, which suggests that younger audiences tend to inquire about our products more often. It is also a variable of importance in both of our random forest models. On the other hand, it is not a significant predictor of approved conversion in our linear regression model.

Our findings regarding the predictor variables associated with the specific marketing campaigns each ad belonged to were a mixed bag. Information regarding an advertisement's association with Campaign 1 turned out to be insignificant, whereas the same for Campaign 2 turned out to be highly significant. We find it likely that this was caused by the disparity in scale between the three marketing campaigns. As previously discussed, Campaign 1 only contained 54 advertisements, whereas Campaign 2 and Campaign 3 contained 464 and 625 advertisements respectively. It seems clear enough from the assessment results that this variable provides our models with valuable information. However, the firm's marketing department may be better equipped to delve into this result given our lack of information regarding the specific particulars of each campaign.

Finally, we found that the target audience's gender was of relatively low importance for approved conversion, and of little to no importance for total conversion in our random forest models. It was also completely insignificant in both of our linear regression models.

5 Conclusions

Our analysis indicates that there are a few KPIs that are strongly associated with higher sales conversion. Based on our predictive analytics, a marketing campaign's success as measured by total and approved conversion is largely dependent on the number of times an ad was shown to target audiences. In addition to the number of impressions an advertisement manages to generate, the social media marketing performance also appears to depend heavily on the amount of per-advertisement spending.

Initial descriptive analytics showed the strongest relationships between per-advertisement spending and impressions, as well as between clicks and impressions. Our data exploration also indicated that there were no advertisements with zero clicks and nonzero spending, which is consistent with the CPC pricing model our firm employs on the Facebook platform. Interestingly, we noted that the relationship between the number of audience clicks and the number of resultant conversions was one of the weakest among the different variables in our dataset. This initial finding appears to be incongruent with the results of our predictive analysis, which indicate that the number of audience clicks has a significant effect on the number of realized conversions.

Our models indicate that target audience age is a significant predictor of total conversion. As per our descriptive analysis, we note that the most frequently occurring age demographic among our target audiences consists of males aged 30 to 34, followed by females aged 30 to 34. The three remaining age ranges consist of older individuals and trails slightly behind in terms of frequency of occurrence. These findings seem to indicate that younger audiences tend to inquire about our products more frequently than those that are older. We do not observe the same effect when it comes to actually converting these website inquiries into sales, however. It is quite possible that this has little do do with the efficacy of our Facebook marketing strategy and more to do with the design of our online sales platform itself. Additionally, we that target audience gender does not have a significant effect on conversion.

Finally, an exploration of the characteristics of the three marketing campaigns described in the dataset indicate substantial differences between the three. Campaign 3 demonstrates the greatest variability and median measurements across most of the aforementioned metrics in our data. It is worth noting that Campaign 3 consisted of 625 advertisements, while Campaign 2 consisted of 464 and Campaign 1 consisted of only 54. Accordingly, Campaign 3 also appeared to have enjoyed the highest level of engagement among target audiences. This disparity in observation counts appears to have been borne out in our models as well. Information regarding an advertisement's association with Campaign 1 was insignificant while the same for Campaign 2 was highly significant.

6 Recommendations

1. **Focus strategic efforts on targeting the right age groups**

Our results indicate that target audience age is the single most important demographic characteristic with respect to maximizing sales conversion. Findings also show that younger audiences tended to inquire about the product more often. Further market research into audiences aged 20-24 and 25-29 could prove especially valuable.

2. **Discount importance of targeting specific gender groups**

The findings point to target audience gender being one of the least important factors with respect to maximizing sales conversion. This suggests that demand for our product is similar across different gender groups, and that it may not be worth expending limited resources to identify and target specific gender groups.

3. **Conduct further analyses of CPC and CPM pricing models**

Our models show that the number of impressions generated by an advertisement is a stronger predictor of sales conversion than the number of audience clicks. However, it would be difficult to make a case for a switch from a CPC to a CPM pricing model without a deeper understanding of the marginal costs and benefits of either option. It is entirely possible that such a transition could lead to higher advertising costs without any significant performance improvements.

4. **Conduct further research into the defining characteristics of Campaign 2**

Our models suggest that an advertisement's association with Campaign 2 may have implications for both total and approved conversion. More research should be done into the defining characteristics of Campaign 2 to gain more insight into how the firm should construct future campaigns.

References

Deloitte. 2020. “Seven Trends Impacting the Retail and Consumer Products Industries Amid a Global Pandemic and Beyond.” *Deloitte Press Release*, Accessed February 20, 2021. https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/retail-and-consumer-products-industries.html.

Facebook. n.d. “Four Tips for Selecting Your Ad Audience.” Accessed March 12, 2021. https://www.facebook.com/business/learn/lessons/facebook-ad-audience-reach-tips.

———. 2020. “How You Are Charged If Your Ad Receives No Impression or Link Click.” Accessed April 3, 2021. https://www.facebook.com/business/help/189320824449558?id=1792465934137726.

Gokagglers. 2017. “Sales Conversion Optimization: How to Cluster Customer Data for Campaign Marketing.” Accessed March 4, 2021. https://www.kaggle.com/loveall/clicks-conversion-tracking.

Kang, Hyun. 2013. “The Prevention and Handling of the Missing Data.” *Korean Journal of Anesthesiology*, Accessed March 6, 2021. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/.

RetailNext. 2014. “Transforming Physical Retail: The Power of e-Commerce-Style Analytics for Brick-and-Mortar Stores.” *RetailNext White Paper*, Accessed February 21, 2021. http://retailnext.net/wp-content/uploads/2014/01/RetailNext-Transforming-Physical-Retail-Whitepaper-Jan2014.pdf.

Unbounce. n.d. “What Is the Average Conversion Rate for a Landing Page?” Accessed April 2, 2021. https://unbounce.com/average-conversion-rates-landing-pages/.

Walton, Chris. 2020. “E-Commerce Data Suggests Some Physical Retail Stores May Be a Luxury Consumers No Longer Need.” *Forbes Editor’s Picks*, Accessed February 20, 2021. https://www.forbes.com/sites/christopherwalton/2020/06/12/retails-recovery-will-at-best-be-a-w-disguised-as-a-v/?sh=7087b7d138f3.

WebFX. 2021. “How Much Does Facebook Advertising Cost in 2021?” Accessed April 3, 2021. https://www.webfx.com/social-media/how-much-does-facebook-advertising-cost.html.

Appendix A: Data

							Total	Approved
Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Conversion	Conversion
708746	campaign_1	32	1	7350	1	1.43	2	1
708749	campaign_1	32	1	17861	2	1.82	2	0
708771	campaign_1	32	1	693	0	0.00	1	0
708815	campaign_1	32	1	4259	1	1.25	1	0
708818	campaign_1	32	1	4133	1	1.29	1	1
708820	campaign_1	32	1	1915	0	0.00	1	1
708889	campaign_1	32	1	15615	3	4.77	1	0
708895	campaign_1	32	1	10951	1	1.27	1	1
708953	campaign_1	32	1	2355	1	1.50	1	0
708958	campaign_1	32	1	9502	3	3.16	1	0
708979	campaign_1	32	1	1224	0	0.00	1	0
709023	campaign_1	32	1	735	0	0.00	1	0
709038	campaign_1	32	1	5117	0	0.00	1	0
709040	campaign_1	32	1	5120	0	0.00	1	0
709059	campaign_1	32	1	14669	7	10.28	1	1
709105	campaign_1	32	1	1241	0	0.00	1	1
709115	campaign_1	32	1	2305	1	0.57	1	0
709124	campaign_1	32	1	1024	0	0.00	1	1
709179	campaign_1	37	1	4627	1	1.69	1	0
709183	campaign_1	37	1	21026	4	4.63	2	1
709320	campaign_1	37	1	1422	0	0.00	1	1
709323	campaign_1	37	1	7132	2	2.61	1	0
709326	campaign_1	37	1	12190	2	3.05	1	0
709327	campaign_1	37	1	12193	2	3.06	1	1
709328	campaign_1	37	1	3332	0	0.00	1	1
709455	campaign_1	37	1	559	0	0.00	1	0
709544	campaign_1	37	1	7440	2	2.98	1	1
709614	campaign_1	42	1	19113	4	5.52	1	0
709756	campaign_1	42	1	10976	2	1.69	1	1
709761	campaign_1	42	1	2861	0	0.00	1	0
709899	campaign_1	42	1	1398	0	0.00	1	1
709901	campaign_1	42	1	23817	7	8.47	1	1

Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Total Conversion	Approved Conversion
710045	campaign_1	47	1	47224	12	15.82	1	0
710088	campaign_1	47	1	2283	1	1.47	1	0
710360	campaign_1	47	1	2182	1	1.53	1	1
710477	campaign_1	32	0	2654	0	0.00	1	1
710480	campaign_1	32	0	57665	14	18.07	1	1
710571	campaign_1	32	0	3091	1	1.61	1	1
710617	campaign_1	32	0	5014	1	1.19	1	0
710623	campaign_1	32	0	38726	7	9.22	1	0
710628	campaign_1	32	0	1473	0	0.00	1	0
710682	campaign_1	32	0	1186	0	0.00	1	0
710763	campaign_1	32	0	5369	1	1.51	1	0
710836	campaign_1	32	0	22221	7	9.43	1	1
710867	campaign_1	32	0	1185	0	0.00	1	0
710880	campaign_1	32	0	13019	5	6.96	1	0
710961	campaign_1	37	0	2508	1	1.22	1	0
710968	campaign_1	37	0	5864	2	2.80	1	1
711217	campaign_1	37	0	2783	1	1.60	1	0
711623	campaign_1	42	0	3812	1	1.13	2	1
711764	campaign_1	47	0	11199	4	5.73	1	1
711785	campaign_1	47	0	292	0	0.00	1	0
711877	campaign_1	47	0	17572	7	9.38	1	0
712052	campaign_1	47	0	1448	0	0.00	1	1
734209	campaign_2	32	1	1772	0	0.00	1	1
734210	campaign_2	32	1	13329	4	5.63	1	1
734215	campaign_2	32	1	13659	3	3.84	1	0
734243	campaign_2	32	1	739	0	0.00	1	1
734266	campaign_2	32	1	605	0	0.00	1	0
734272	campaign_2	32	1	1030	0	0.00	1	0
734290	campaign_2	32	1	5374	1	1.04	4	0
734313	campaign_2	32	1	790	0	0.00	1	1
734314	campaign_2	32	1	962	0	0.00	1	0
734352	campaign_2	37	1	4423	1	1.46	1	1
734361	campaign_2	37	1	12382	2	2.84	1	1
734381	campaign_2	37	1	2938	1	1.35	1	1

							Total	Approved
Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Conversion	Conversion
734399	campaign_2	37	1	239	0	0.00	1	0
734418	campaign_2	37	1	591	0	0.00	1	0
734421	campaign_2	37	1	10332	4	5.75	1	0
734427	campaign_2	37	1	8259	3	3.98	1	0
734433	campaign_2	37	1	12158	3	4.45	1	0
734582	campaign_2	42	1	7709	2	1.32	2	0
734605	campaign_2	42	1	834	0	0.00	1	0
734660	campaign_2	47	1	1299	0	0.00	2	0
734666	campaign_2	47	1	371	0	0.00	1	0
734726	campaign_2	47	1	10466	3	4.09	1	0
734737	campaign_2	47	1	839	0	0.00	1	0
734785	campaign_2	32	0	5576	1	1.53	1	1
734794	campaign_2	32	0	4010	0	0.00	1	0
734796	campaign_2	32	0	39337	7	10.03	1	1
734800	campaign_2	32	0	1635	0	0.00	1	0
734803	campaign_2	32	0	1631	0	0.00	1	0
734852	campaign_2	32	0	13479	3	4.25	1	0
734854	campaign_2	32	0	57022	13	20.29	3	3
734856	campaign_2	32	0	5453	1	1.39	1	1
734866	campaign_2	32	0	11803	3	4.44	1	0
734881	campaign_2	32	0	4259	1	1.57	1	1
734901	campaign_2	32	0	1554	0	0.00	1	0
734903	campaign_2	32	0	5323	1	1.29	1	1
734925	campaign_2	37	0	5024	1	1.41	1	1
734939	campaign_2	37	0	104648	24	33.33	4	2
734968	campaign_2	37	0	8504	3	3.34	1	1
734999	campaign_2	37	0	20277	6	8.05	1	0
735014	campaign_2	37	0	12403	4	5.21	1	1
735032	campaign_2	37	0	498	0	0.00	1	1
735033	campaign_2	37	0	652	0	0.00	0	0
735043	campaign_2	37	0	1357	0	0.00	1	1
735048	campaign_2	37	0	1393	0	0.00	1	0
735065	campaign_2	42	0	648	0	0.00	1	0
735109	campaign_2	42	0	708	0	0.00	1	1

							Total	Approved
Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Conversion	Conversion
735140	campaign_2	42	0	6907	2	2.35	1	0
735143	campaign_2	42	0	39035	13	19.33	1	0
735151	campaign_2	42	0	926	0	0.00	1	0
735184	campaign_2	42	0	4412	1	1.45	1	0
735189	campaign_2	42	0	9965	3	4.05	1	0
735213	campaign_2	47	0	73634	23	32.98	1	0
735220	campaign_2	47	0	69708	20	31.29	1	0
735242	campaign_2	47	0	530	0	0.00	1	0
735247	campaign_2	47	0	14257	6	8.79	1	0
735289	campaign_2	47	0	20362	5	9.12	1	1
735290	campaign_2	47	0	12215	4	6.26	1	0
735298	campaign_2	47	0	85412	28	38.64	2	1
736869	campaign_2	32	1	2338	1	0.24	1	0
736890	campaign_2	32	1	2522	0	0.00	1	0
736893	campaign_2	32	1	3587	0	0.00	1	0
736977	campaign_2	32	1	1273	0	0.00	1	0
736988	campaign_2	32	1	3891	1	1.09	1	0
736995	campaign_2	32	1	1888	0	0.00	1	0
736997	campaign_2	32	1	1895	0	0.00	1	0
737097	campaign_2	37	1	715	0	0.00	1	0
737130	campaign_2	37	1	11199	2	2.68	1	0
737320	campaign_2	37	1	5676	2	3.01	1	0
737375	campaign_2	42	1	1415	0	0.00	1	0
737524	campaign_2	42	1	2148	1	1.58	1	1
737644	campaign_2	47	1	45401	10	14.06	1	0
737657	campaign_2	47	1	7478	2	2.90	1	1
737658	campaign_2	47	1	4919	1	1.59	1	0
737674	campaign_2	47	1	533	0	0.00	1	1
737766	campaign_2	47	1	1447	0	0.00	1	1
737896	campaign_2	32	0	17553	3	4.59	1	0
737931	campaign_2	32	0	3343	1	0.54	1	0
737961	campaign_2	32	0	523	0	0.00	1	0
737995	campaign_2	32	0	1873	0	0.00	1	0
738006	campaign_2	32	0	34740	7	13.41	1	1

							Total	Approved
Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Conversion	Conversion
738067	campaign_2	32	0	658	0	0.00	1	0
738098	campaign_2	32	0	1539	0	0.00	1	0
738307	campaign_2	37	0	3010	1	0.86	1	1
738389	campaign_2	42	0	27081	9	10.77	1	1
738408	campaign_2	42	0	20233	4	5.59	3	0
738413	campaign_2	42	0	147159	36	58.16	3	1
738423	campaign_2	42	0	21664	7	10.62	1	1
738436	campaign_2	42	0	9112	4	5.46	1	1
738463	campaign_2	42	0	542	0	0.00	1	0
738528	campaign_2	42	0	402	0	0.00	1	1
738560	campaign_2	42	0	1338	0	0.00	1	0
738582	campaign_2	47	0	46150	15	20.18	1	1
738592	campaign_2	47	0	493821	116	176.38	4	1
738593	campaign_2	47	0	92011	27	34.39	2	1
738598	campaign_2	47	0	12956	4	5.49	1	1
738606	campaign_2	47	0	529	0	0.00	1	0
738637	campaign_2	47	0	944	1	1.42	1	0
738648	campaign_2	47	0	111090	38	51.97	5	1
747212	campaign_2	32	1	7208	2	3.19	1	0
747213	campaign_2	32	1	1746	0	0.00	1	0
747220	campaign_2	32	1	2474	0	0.00	2	2
747222	campaign_2	32	1	12489	2	1.96	1	0
747223	campaign_2	32	1	8032	1	0.60	2	0
747248	campaign_2	32	1	472	0	0.00	1	1
747332	campaign_2	32	1	792	0	0.00	1	1
747362	campaign_2	37	1	4607	1	1.15	1	1
747369	campaign_2	37	1	13355	2	3.18	1	1
747370	campaign_2	37	1	2936	0	0.00	1	0
747401	campaign_2	37	1	2793	1	0.98	1	1
747435	campaign_2	37	1	1032	0	0.00	1	0
747439	campaign_2	37	1	1662	0	0.00	1	1
747489	campaign_2	42	1	4016	2	1.48	1	1
747514	campaign_2	42	1	14843	3	2.94	1	1
747645	campaign_2	47	1	9674	3	4.60	1	1

Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Total Conversion	Approved Conversion
747659	campaign_2	47	1	12186	2	2.67	1	0
747675	campaign_2	47	1	673	0	0.00	1	0
747678	campaign_2	47	1	370	0	0.00	1	1
747712	campaign_2	47	1	450	0	0.00	1	1
747790	campaign_2	32	0	2077	0	0.00	1	1
747791	campaign_2	32	0	31393	8	10.96	1	1
747795	campaign_2	32	0	8410	2	2.36	1	1
747798	campaign_2	32	0	25884	5	7.35	1	0
747824	campaign_2	32	0	608	0	0.00	1	1
747828	campaign_2	32	0	28488	10	9.34	1	0
747852	campaign_2	32	0	10126	3	4.62	1	0
747859	campaign_2	32	0	22572	5	8.50	1	0
747863	campaign_2	32	0	1955	0	0.00	1	1
747879	campaign_2	32	0	493	0	0.00	1	0
747903	campaign_2	32	0	1491	0	0.00	1	1
747911	campaign_2	32	0	1495	0	0.00	1	1
747968	campaign_2	37	0	512	0	0.00	0	0
747991	campaign_2	37	0	4868	2	2.42	1	0
748000	campaign_2	37	0	6585	2	2.95	1	0
748007	campaign_2	37	0	10164	2	3.72	1	1
748014	campaign_2	37	0	11182	4	4.45	1	0
748045	campaign_2	37	0	1238	0	0.00	1	0
748086	campaign_2	42	0	34127	8	13.07	1	0
748087	campaign_2	42	0	29466	7	10.85	2	0
748089	campaign_2	42	0	38759	9	10.85	1	0
748091	campaign_2	42	0	41720	10	12.06	1	1
748225	campaign_2	47	0	18602	5	8.86	1	0
748230	campaign_2	47	0	83929	21	27.73	4	1
748231	campaign_2	47	0	25194	6	7.35	1	0
748233	campaign_2	47	0	78627	19	26.53	1	0
748235	campaign_2	47	0	102695	25	39.43	3	0
748294	campaign_2	47	0	82827	24	47.93	3	0
748295	campaign_2	47	0	9240	3	6.04	1	0
748303	campaign_2	47	0	7706	2	2.37	1	0

							Total	Approved
Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Conversion	Conversion
748314	campaign_2	47	0	7821	4	6.34	1	1
748341	campaign_2	47	0	1363	0	0.00	1	1
776318	campaign_2	47	0	3569	0	0.00	1	1
776322	campaign_2	47	0	119063	34	53.22	1	0
776323	campaign_2	47	0	99078	23	35.80	2	0
776325	campaign_2	47	0	452398	114	180.22	1	0
776334	campaign_2	47	0	191223	48	76.41	1	0
776336	campaign_2	47	0	22216	6	9.55	1	0
776338	campaign_2	47	0	48291	11	18.02	1	0
776353	campaign_2	42	0	27559	8	13.37	1	0
776373	campaign_2	47	0	10194	4	4.59	2	1
776383	campaign_2	47	0	1168	0	0.00	1	1
776405	campaign_2	47	0	40126	16	25.86	1	0
776416	campaign_2	47	0	3659	1	0.49	1	1
776430	campaign_2	32	1	3200	0	0.00	1	0
776464	campaign_2	47	0	7550	1	1.68	1	1
776469	campaign_2	47	0	45397	15	25.42	1	1
776473	campaign_2	32	1	23086	2	3.31	1	1
776475	campaign_2	32	1	16425	1	1.55	1	0
776476	campaign_2	32	1	43756	5	5.44	0	0
776477	campaign_2	32	1	9982	0	0.00	1	0
776489	campaign_2	47	0	175389	55	81.61	1	0
776494	campaign_2	32	1	7015	0	0.00	1	0
776515	campaign_2	47	0	12706	3	4.99	1	1
776519	campaign_2	47	0	70702	20	31.71	1	0
776533	campaign_2	47	0	63927	16	25.52	2	0
776534	campaign_2	47	0	15105	3	4.26	1	0
776538	campaign_2	32	0	8774	1	1.83	1	0
776551	campaign_2	32	0	14459	1	1.39	1	0
776552	campaign_2	32	0	21596	2	2.81	1	0
776553	campaign_2	32	0	66765	8	11.05	1	0
776563	campaign_2	32	0	1369	0	0.00	1	1
776579	campaign_2	32	0	26910	5	7.23	1	0
776603	campaign_2	32	0	506	0	0.00	1	0

Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Total Conversion	Approved Conversion
776615	campaign_2	47	1	11988	3	4.27	1	0
776623	campaign_2	47	1	19353	6	9.48	1	1
776631	campaign_2	42	1	10960	2	2.89	1	0
776643	campaign_2	32	1	33491	6	10.57	2	1
776644	campaign_2	32	1	20083	2	3.20	2	1
776659	campaign_2	32	1	8817	0	0.00	1	1
776661	campaign_2	32	1	15466	1	0.97	1	0
776662	campaign_2	32	1	27072	3	4.37	1	0
776663	campaign_2	32	1	15753	1	0.57	1	1
776668	campaign_2	42	1	3523	1	1.81	1	1
776685	campaign_2	42	1	7745	0	0.00	1	0
776686	campaign_2	42	1	18709	2	3.32	1	0
776687	campaign_2	42	1	8022	0	0.00	2	1
776696	campaign_2	32	1	7966	1	1.18	1	1
776697	campaign_2	32	1	4132	0	0.00	1	1
776698	campaign_2	32	1	12785	3	4.73	2	1
776699	campaign_2	32	1	8213	1	1.38	1	1
776722	campaign_2	32	1	545	0	0.00	1	1
776725	campaign_2	42	1	2479	1	1.26	1	0
776780	campaign_2	42	1	3812	2	3.05	1	0
776793	campaign_2	47	1	1609	0	0.00	1	0
776799	campaign_2	47	1	10257	3	3.58	1	1
776817	campaign_2	42	1	12356	4	6.28	1	0
776825	campaign_2	47	1	7410	1	1.21	1	0
776829	campaign_2	47	1	140098	28	46.63	1	0
776831	campaign_2	47	1	107021	20	34.44	1	0
776840	campaign_2	37	1	2797	1	1.29	1	0
776861	campaign_2	47	1	16461	6	9.22	1	0
776892	campaign_2	42	1	17488	5	7.72	1	0
776928	campaign_2	37	1	9750	2	1.50	1	1
776935	campaign_2	47	1	1136	0	0.00	1	1
777105	campaign_2	47	1	4333	1	0.18	1	1
777130	campaign_2	37	1	6260	0	0.00	1	0
777131	campaign_2	37	1	6359	0	0.00	1	0

Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Total Conversion	Approved Conversion
777166	campaign_2	32	1	2383	0	0.00	1	1
777187	campaign_2	42	1	11292	3	5.39	1	1
777198	campaign_2	32	1	12729	4	5.78	1	0
777200	campaign_2	32	1	1898	0	0.00	1	1
777201	campaign_2	32	1	1882	0	0.00	1	1
777235	campaign_2	32	1	2883	1	0.99	1	1
777248	campaign_2	32	0	3989	1	1.28	1	0
777261	campaign_2	42	1	19603	4	5.28	1	1
777382	campaign_2	42	1	3047	1	1.38	1	0
777398	campaign_2	37	1	3029	1	1.05	1	1
777410	campaign_2	47	1	3490	1	1.34	1	1
777482	campaign_2	47	1	2479	0	0.00	1	0
777495	campaign_2	42	1	19581	7	10.43	2	0
777519	campaign_2	47	1	19537	5	6.10	1	0
777625	campaign_2	47	1	59433	12	19.66	3	0
777627	campaign_2	47	1	157534	33	56.19	2	0
777638	campaign_2	42	1	1781	0	0.00	1	1
777670	campaign_2	42	1	23769	4	6.03	1	0
777673	campaign_2	42	1	7101	0	0.00	1	0
777742	campaign_2	37	1	4726	1	1.83	1	1
777758	campaign_2	32	1	5209	1	0.96	2	0
777794	campaign_2	32	1	13473	3	2.62	3	0
777816	campaign_2	42	1	500	0	0.00	1	1
777871	campaign_2	32	1	4616	1	1.36	1	0
777904	campaign_2	32	1	3279	0	0.00	1	0
777905	campaign_2	32	1	3288	0	0.00	1	0
778037	campaign_2	37	1	14615	4	6.05	1	0
778048	campaign_2	32	1	56615	12	19.88	2	0
778085	campaign_2	32	1	11735	3	4.53	1	1
778087	campaign_2	32	1	15910	5	6.78	1	0
778112	campaign_2	37	1	11446	2	3.09	1	1
778113	campaign_2	37	1	4595	0	0.00	1	0
778124	campaign_2	32	1	4871	0	0.00	1	0
778148	campaign_2	37	1	3199	0	0.00	1	0

Ad ID	Campaign	Age	Gender	Impressions	Clicks	Spent	Total Conversion	Approved Conversion
778156	campaign_2	32	1	9388	2	3.14	1	0
778161	campaign_2	32	1	17954	6	7.54	2	1
778208	campaign_2	42	0	2755	0	0.00	1	0
778264	campaign_2	42	0	8152	1	0.99	1	0
778266	campaign_2	42	0	74542	19	34.15	1	0
778421	campaign_2	42	0	6699	2	3.09	1	0
778422	campaign_2	42	0	11911	4	3.96	1	0
778461	campaign_2	42	1	10090	2	2.65	1	1
778471	campaign_2	32	1	1273	0	0.00	1	1
778483	campaign_2	42	0	24188	5	8.18	1	0
778529	campaign_2	32	1	2214	0	0.00	1	0
778556	campaign_2	42	1	9735	4	4.13	1	1
778590	campaign_2	32	1	1371	0	0.00	1	1
778600	campaign_2	42	0	10750	4	5.39	1	0
778626	campaign_2	32	1	7629	1	0.72	1	1
778628	campaign_2	32	1	4608	0	0.00	1	0
778674	campaign_2	37	1	3732	0	0.00	1	0
778689	campaign_2	32	1	7453	1	1.68	1	1
778722	campaign_2	37	0	41785	14	19.10	1	0
778737	campaign_2	37	1	8077	2	3.58	1	1
778756	campaign_2	37	0	5602	1	1.58	1	0
778804	campaign_2	32	1	6184	2	2.75	1	1
778808	campaign_2	32	1	1738	0	0.00	1	0
778964	campaign_2	37	0	112460	25	41.29	1	0
779057	campaign_2	42	1	4414	0	0.00	1	0
779106	campaign_2	37	0	14670	7	9.41	1	0
779438	campaign_2	32	0	33144	9	13.41	1	0
779453	campaign_2	47	1	4397	1	0.95	1	0
779488	campaign_2	47	1	1006	0	0.00	1	0
779573	campaign_2	37	0	89527	24	32.29	1	0
779608	campaign_2	37	0	2459	0	0.00	1	0

Appendix B: Data preparation details

R

We found that the first column named ad_id already provided a unique per-advertisement identification number, so we dropped the facebook campaign ID variable to avoid redundancy in the data. We also found that the interest column is not useful for analysis as it is impossible to interpret in context of the data, so we removed this variable as well. We also checked to see whether there were any missing values in the entire dataset and found none across all the variables.

```
## drop facebook campaign ID and interest variables
mydata$fb_campaign_id <- NULL
mydata$interest <- NULL

## check for NA values in data frame
apply(mydata, 2, function(x) any(is.na(x))) #FALSE for all variables

##       ad_id    campaign         age        male impressions      clicks
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##       spent  total_conv    app_conv       is_c1       is_c2
##       FALSE       FALSE       FALSE       FALSE       FALSE
```

We then set the names of our variables to lowercase letters and renamed certain columns for ease of understanding.

```
## rename columns and set to
names(mydata) <- tolower(names(mydata))
names(mydata)[names(mydata) == "gender"] <- "male"
names(mydata)[names(mydata) == "xyz_campaign_id"] <- "campaign"
names(mydata)[names(mydata) == "total_conversion"] <- "total_conv"
names(mydata)[names(mydata) == "approved_conversion"] <- "app_conv"
```

Next, we re-coded certain variables. We changed the gender variable from a character of M and F to boolean values of 1 and 0, respectively, where 1 represents TRUE for whether an individual identifies as male and 0 represents FALSE for whether an individual identifies as male. We also re-coded the age variable from a string to an integer value as it would be easier for analysis to represent the data as a single value rather than a range of ages. We lastly changed the campaign id variable into a factor that is more easily interpreted as campaign 1, 2, and 3 rather than as an unique campaign identifier that does not necessarily have any meaning in the context of our analysis.

```
## recode gender variable
mydata$male[mydata$male == 'M'] <- 1
mydata$male[mydata$male == 'F'] <- 0
mydata$male <- as.integer(mydata$male)

## recode age variable
mydata$age[mydata$age == '30-34'] <- 32
mydata$age[mydata$age == '35-39'] <- 37
mydata$age[mydata$age == '40-44'] <- 42
mydata$age[mydata$age == '45-49'] <- 47
mydata$age <- as.integer(mydata$age)

## recode campaign id variable
mydata$campaign[mydata$campaign == 916] <- "campaign_1"
mydata$campaign[mydata$campaign == 936] <- "campaign_2"
mydata$campaign[mydata$campaign == 1178] <- "campaign_3"
mydata$campaign <- as.factor(mydata$campaign)
```

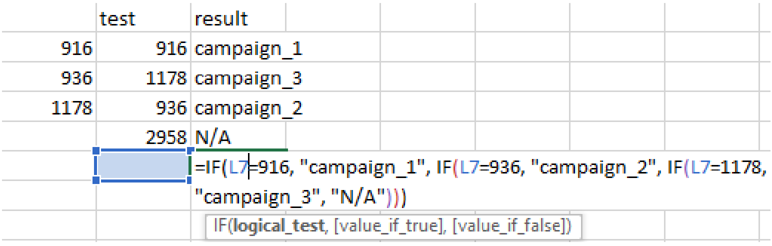
We also created two dummy variables associated with the 3 minus 1 levels of our campaign variable. This preliminary step was necessary because we intended to include campaign as a categorical variable within our linear regression models.

```
## create 2 dummy variables for the 3 levels of the campaign id variable
mydata <- mydata %>%
  mutate(is_c1 = ifelse(campaign=="campaign_1",1,0),
         is_c2 = ifelse(campaign=="campaign_2",1,0))
```

Excel

Given that the first column of data already provided a unique per-advertisement identification number, we decided to stow away the column of Facebook advertisement identification numbers away in order to avoid unnecessary redundancy. We also removed the column of interest codes because the values were impossible to interpret in the context of our data set. Finally, we confirmed that there were no missing or blank values in our data and proceeded to carry out four primary data preparation tasks.

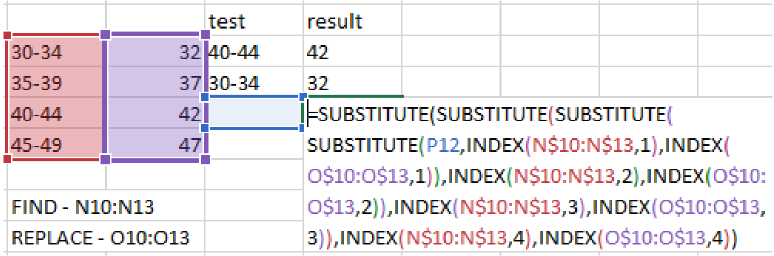
First, we decided to assign each of the three marketing campaigns with labels that were more descriptive than their abstract identification numbers. We accomplished this through the use of a nested if() call which we applied to the column values using the fill function.



	test	result
916	916	campaign_1
936	1178	campaign_3
1178	936	campaign_2
	2958	N/A

```
=IF(L7=916, "campaign_1", IF(L7=936, "campaign_2", IF(L7=1178,
"campaign_3", "N/A")))
IF(logical_test, [value_if_true], [value_if_false])
```

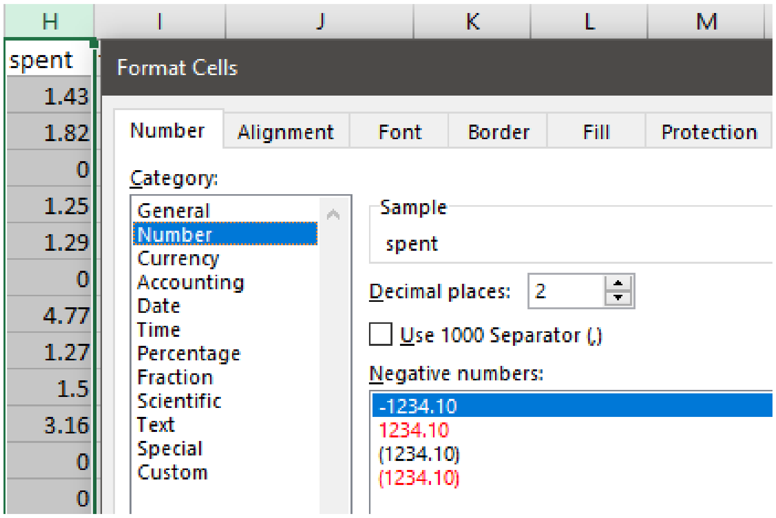
Second, we looked at the age column and noted that it would be more convenient to have a single integer than a string representing an age range, especially for modelling and other analytical purposes. We accomplished this through the use of a series of nested substitute() and index() calls. We used absolute row number references in order to avoid having the indices shift vertically as we substituted column values with the fill function.



		test	result
30-34	32	40-44	42
35-39	37	30-34	32
40-44	42		
45-49	47		
FIND - N10:N13			
REPLACE - O10:O13			

```
=SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(
SUBSTITUTE(P12,INDEX(N$10:N$13,1),INDEX(
O$10:O$13,1)),INDEX(N$10:N$13,2),INDEX(O$10:
O$13,2)),INDEX(N$10:N$13,3),INDEX(O$10:O$13,
3)),INDEX(N$10:N$13,4),INDEX(O$10:O$13,4))
```

Third, we rounded the long decimal data in the spent column to two decimal places in order to have them reflect typical dollar values. We accomplished this by highlighting the column and adjusting the formatting of the cells appropriately. We chose not to include any currency symbols in order to avoid complications in later analytical and modeling processes.



spent
1.43
1.82
0
1.25
1.29
0
4.77
1.27
1.5
3.16
0
0

Category: Number

Sample: spent

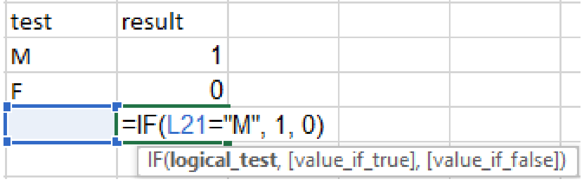
Decimal places: 2

□ Use 1000 Separator (,)

Negative numbers:

-1234.10
1234.10
(1234.10)
(1234.10)

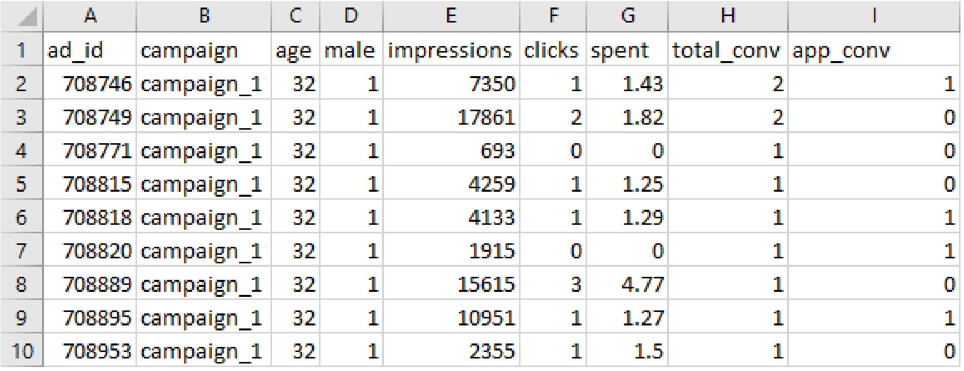
Fourth, we converted the character data in the gender column into a boolean with the value "1" for male and "0" for female. We accomplished this through the use of an if() call which we applied to the column values using the fill function.



test	result
M	1
F	0
	=IF(L21="M", 1, 0)

IF(**logical_test**, [value_if_true], [value_if_false])

Finally, we changed the column names to better reflect their content as well as for the sake of simplicity. For instance, we renamed the gender column as "male" in order to reflect the fact that our boolean associates the value "1" with the male gender. Furthermore, the names were changed to lowercase and also abbreviated where appropriate. A small screenshot of the resulting spreadsheet is provided below.



	A	B	C	D	E	F	G	H	I
1	ad_id	campaign	age	male	impressions	clicks	spent	total_conv	app_conv
2	708746	campaign_1	32	1	7350	1	1.43	2	1
3	708749	campaign_1	32	1	17861	2	1.82	2	0
4	708771	campaign_1	32	1	693	0	0	1	0
5	708815	campaign_1	32	1	4259	1	1.25	1	0
6	708818	campaign_1	32	1	4133	1	1.29	1	1
7	708820	campaign_1	32	1	1915	0	0	1	1
8	708889	campaign_1	32	1	15615	3	4.77	1	0
9	708895	campaign_1	32	1	10951	1	1.27	1	1
10	708953	campaign_1	32	1	2355	1	1.5	1	0

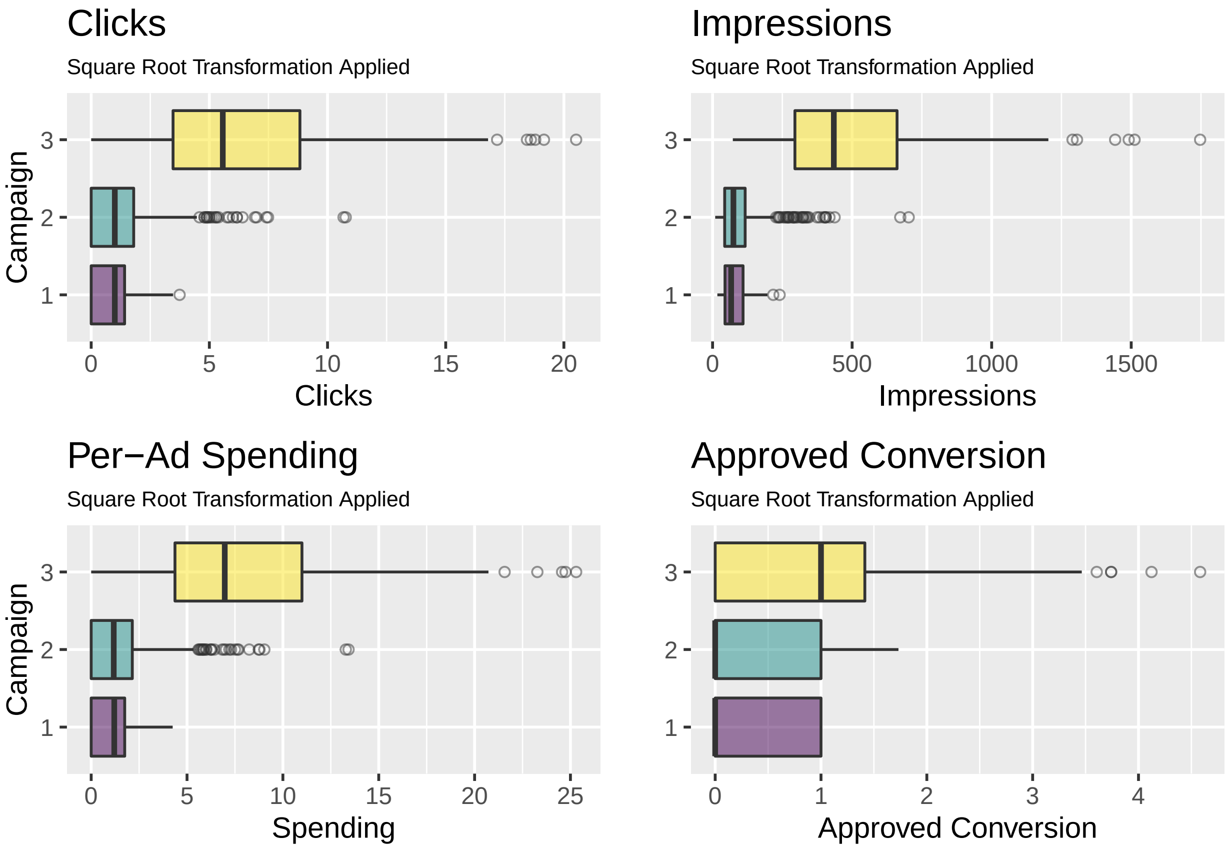
Appendix C: Analytics details

Descriptive analytics

First, we used the ggplot graphical environment to generate four separate box plots, each depicting a quantitative ad performance metric across the three levels of our campaign variable. A square root transformation was applied to normalize the data, which was initially quite skewed as a result of a number of outliers in the Campaign 3 data. This allowed us to create graphics which were slightly more interesting to look at. We then used a helpful function from the gridExtra package to help us arrange the four plots in a single graphic. The underlying data was temporarily manipulated to improve the clarity of our visualizations.

```
## temporarily recode data for clarity
mydata$campaign <- as.factor(recode(mydata$campaign,
                                    "campaign_1" = 1,
                                    "campaign_2" = 2,
                                    "campaign_3" = 3))

## boxplot of per-ad clicks in each campaign
# square-root transformation applied within the initial ggplot function call
p1 <- ggplot(mydata, aes(campaign, sqrt(clicks), fill=campaign)) +
  geom_boxplot(outlier.shape=1, alpha=0.5) +
  scale_fill_viridis(discrete="true", alpha=0.9) +
  labs(x="Campaign",
       y="Clicks",
       title="Clicks",
       subtitle="Square Root Transformation Applied") +
  theme(legend.position="none",
        plot.title=element_text(size=14),
        plot.subtitle=element_text(size=8)) +
  coord_flip()

## boxplot of per-ad impressions in each campaign
# square-root transformation applied within the initial ggplot function call
p2 <- ggplot(mydata, aes(campaign, sqrt(impressions), fill=campaign)) +
  geom_boxplot(outlier.shape=1, alpha=0.5) +
  scale_fill_viridis(discrete="true", alpha=0.9) +
  labs(x="",
       y="Impressions",
       title="Impressions",
       subtitle="Square Root Transformation Applied") +
  theme(legend.position="none",
        plot.title=element_text(size=14),
        plot.subtitle=element_text(size=8)) +
  coord_flip()

## boxplot of per-ad spending in each campaign
# square-root transformation applied within the initial ggplot function call
p3 <- ggplot(mydata, aes(campaign, sqrt(spent), fill=campaign)) +
  geom_boxplot(outlier.shape=1, alpha=0.5) +
  scale_fill_viridis(discrete="true", alpha=0.9) +
  labs(x="Campaign",
       y="Spending",
       title="Per-Ad Spending",
       subtitle="Square Root Transformation Applied") +
  theme(legend.position="none",
        plot.title=element_text(size=14),
        plot.subtitle=element_text(size=8)) +
  coord_flip()

## boxplot of per-ad approved conversions in each campaign
# square-root transformation applied within the initial ggplot function call
p4 <- ggplot(mydata, aes(campaign, sqrt(app_conv), fill=campaign)) +
  geom_boxplot(outlier.shape=1, alpha=0.5) +
  scale_fill_viridis(discrete="true", alpha=0.9) +
  labs(x="",
       y="Approved Conversion",
       title="Approved Conversion",
       subtitle="Square Root Transformation Applied") +
  theme(legend.position="none",
        plot.title=element_text(size=14),
        plot.subtitle=element_text(size=8)) +
  coord_flip()

## revert temporary data changes
mydata$campaign <- recode(mydata$campaign,
                          "1" = "campaign_1",
                          "2" = "campaign_2",
                          "3" = "campaign_3")
```

```
## arranging the four plots in a 2x2 grid
grid.arrange(p1,p2,p3,p4, ncol=2)
```



Second, we created a kable in order to depict the number of observations per campaign. This was a relatively simple matter of passing a list containing counts of observations in each campaign into the kable function and specifying new column names. The underlying data was temporarily manipulated to improve the clarity of our visualization.

```
## temporarily recode data for clarity
```

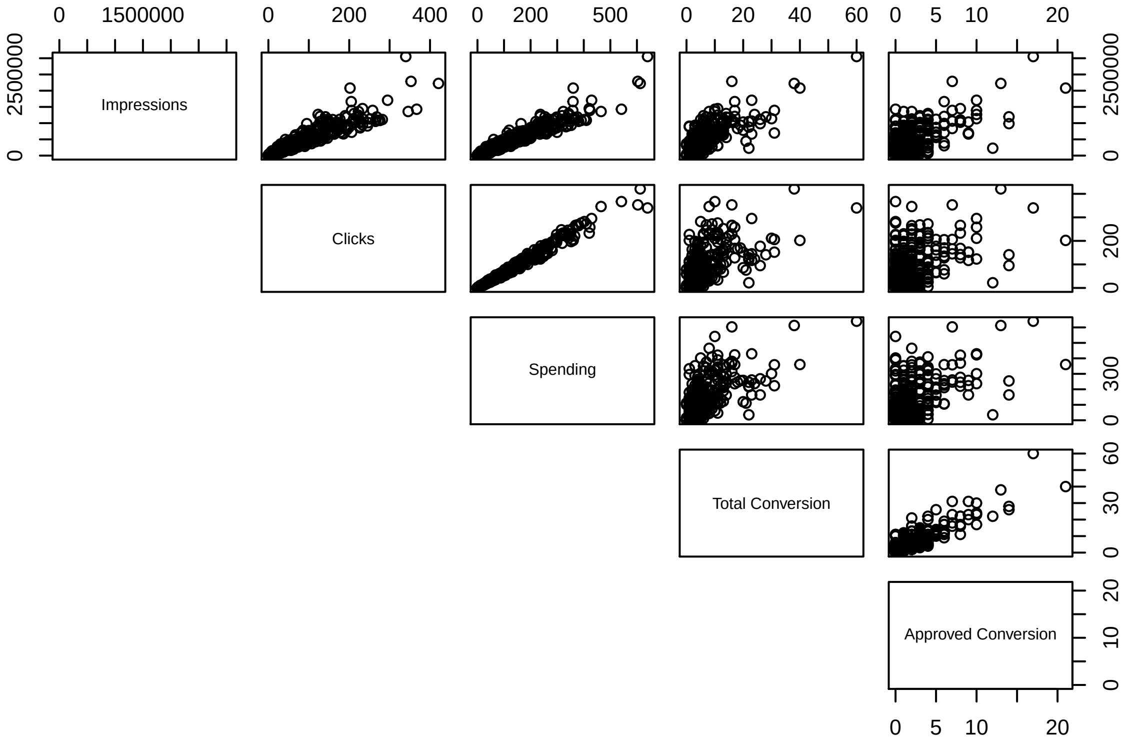
```
mydata$campaign <- as.factor(recode(mydata$campaign,
                                    "campaign_1" = 1,
                                    "campaign_2" = 2,
                                    "campaign_3" = 3))

## table of counts of observations in each campaign
mydata %>%
  count(campaign) %>%
  kable(col.names = c("Campaign","Count"))
```

Campaign	Count
1	54
2	464
3	625

```
## revert temporary data changes
mydata$campaign <- recode(mydata$campaign,
                          "1" = "campaign_1",
                          "2" = "campaign_2",
                          "3" = "campaign_3")
```

Third, we used the base-R graphical environment to generate a correlation plot in order to visualize the relationship between the quantitative variables in our data set. We accomplished this by using the base-R pairs function and passing in a selection of columns containing data on our quantitative variables. We manipulated the labels option in order to display variable names which were more descriptive and easier to interpret.

```
## generate scatter plots of all quantitative variables
pairs(mydata[5:9],
      labels = c("Impressions","Clicks","Spending",
                 "Total Conversion","Approved Conversion"),
      lower.panel = NULL)
```



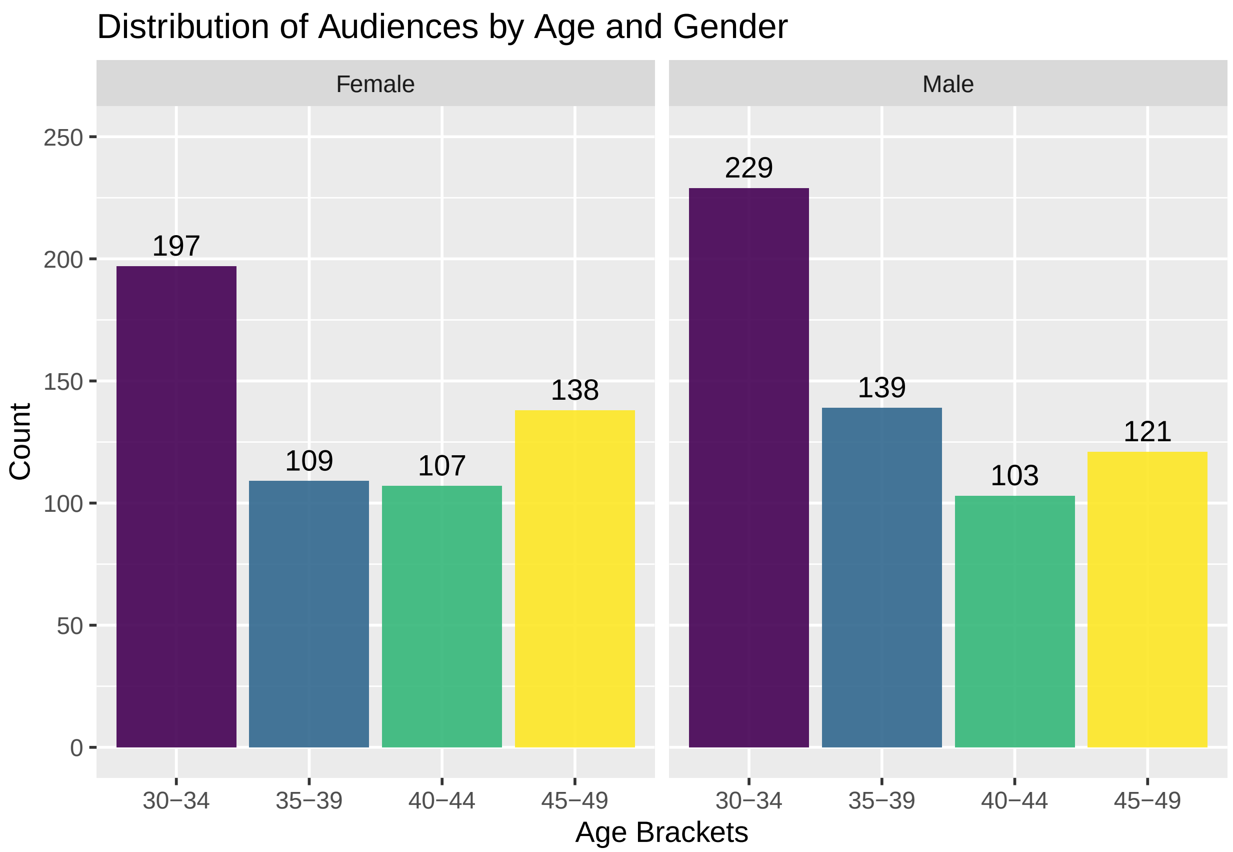
Fourth, we created a correlation matrix to better understand the relationship between the quantitative variables using the base-R function cor and saved this as a data frame. We also rounded the values within the matrix to the thousandths place using the round function. We adjusted the row and column names of the data frame containing the correlation matrix in order to display variable names which were more descriptive and easier to interpret. Finally, we created a kable in order to present these correlation values.

```
## table of correlation coefficients of all quantitative variables
correlation = as.data.frame(row = c("Impressions","Clicks","Spending",
                                    "Total Conversion","Approved Conversion"),
                            round(cor(mydata[5:9]), 3))
names(correlation) <- c("Impressions","Clicks","Spending",
                        "Total Conversion","Approved Conversion")
kable(correlation)
```

	Impressions	Clicks	Spending	Total Conversion	Approved Conversion
Impressions	1.000	0.949	0.970	0.813	0.684
Clicks	0.949	1.000	0.993	0.695	0.560
Spending	0.970	0.993	1.000	0.725	0.593
Total Conversion	0.813	0.695	0.725	1.000	0.864
Approved Conversion	0.684	0.560	0.593	0.864	1.000

Lastly, we used the ggplot graphical environment to generate two bar plots in order to visualize the distribution of observations by age and gender. We affixed labels for observation counts to each bar in order to make the figure easier to interpret. The underlying data was temporarily manipulated to improve the clarity of our visualizations.

```
## temporarily recode data for clarity
mydata$male <- recode(mydata$male,
                      "0"="Female",
                      "1"="Male")
mydata$age <- recode(mydata$age,
                     "32" = "30-34",
                     "37" = "35-39",
                     "42" = "40-44",
                     "47" = "45-49")

## bar plots of audience distribution by age and gender
ggplot(mydata, aes(x=age, fill=as.factor(age))) +
  geom_bar(show.legend = FALSE) +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) +
  facet_wrap(~male) +
  scale_fill_viridis(discrete="true", alpha=0.9) +
  labs(x = "Age Brackets",
       y = "Count",
       title = "Distribution of Audiences by Age and Gender") +
  lims(y=c(0,250))
```



```
## revert temporary data changes
mydata$male <- recode(mydata$male,
                      "Female" = 0,
                      "Male" = 1)
mydata$age <- recode(mydata$age,
                     "30-34" = 32,
                     "35-39" = 37,
                     "40-44" = 42,
                     "45-49" = 47)
```

Predictive analytics

We began by setting a seed for reproducibility and then split our primary dataset into training, validation, and testing subsets for the sake of our predictive analytics process.

```
## Set seed for data splitting
set.seed(4220)
```

```
## All-Campaign training/validation/testing sets
fb.div <- mydata %>%
  initial_split(prop = 0.6, strata = campaign)
fb.div1 <- fb.div %>%
  testing() %>%
  initial_split(prop = 0.5, strata = campaign)
fb.train <- training(fb.div)
fb.val <- training(fb.div1)
fb.test <- testing(fb.div1)
```

Then, we trained a number of random forest models using the training subset and assessed them based on their RMSE values by using the testing and validation sets. The pruned models were built with the goal of finding a simpler model with similar or comparable model utility that was easier to interpret. We arrived at our reduced models by sequentially removing variables with very low variable importance.

```
## Random Forest Analysis - Initial Models
rf.app <- randomForest(app_conv ~ age + male + impressions +
                         clicks + spent + is_c1 + is_c2,
                       data = fb.train, mtry = 2, importance = TRUE)
rf.tot <- randomForest(total_conv ~ age + male + impressions +
                         clicks + spent + is_c1 + is_c2,
                       data = fb.train, mtry = 2, importance = TRUE)

## Random Forest Analysis - Reduced Models
rf.app.1 <- randomForest(app_conv ~ age + impressions + clicks + spent,
                           data = fb.train, mtry = 2, importance = TRUE)
rf.tot.1 <- randomForest(total_conv ~ age + impressions + clicks + spent,
                           data = fb.train, mtry = 2, importance = TRUE)
```

First, the pair of campaign dummy variables were removed due to is_c1 having very low or negative variable importance in both of our full models. The gender variable was the next to be removed, as we found that it had a very low variable importance in the reduced model. The four remaining variables all had high variable importance measures and were therefore retained in our final reduced models. Summaries of the model assessment results are organized and presented in a table.

```
## Random Forest Model Assessment - Training Data
# Create a vector of model names
mods.tr <- c("Initial - Approved Conversion",
             "Reduced - Approved Conversion",
             "Initial - Total Conversion",
             "Reduced - Total Conversion")
# Create a vector of RMSE values for each model
rmse.tr <- c(sqrt(rf.app$mse[length(rf.app$mse)]),
             sqrt(rf.app.1$mse[length(rf.app.1$mse)]),
             sqrt(rf.tot$mse[length(rf.tot$mse)]),
             sqrt(rf.tot.1$mse[length(rf.tot.1$mse)]))
# Add vectors to a data frame, output a summary table
sum.tr <- data.frame(mods.tr, rmse.tr)
sum.tr %>%
  kable(col.names = c("Model Name", "RMSE"))
```

Model Name	RMSE
Initial - Approved Conversion	1.134338
Reduced - Approved Conversion	1.143724
Initial - Total Conversion	2.298172
Reduced - Total Conversion	2.306920

```
## Random Forest Model Assessment - Validation Data
# Generate predictions using validation set
pred.init.app <- predict(rf.app, fb.val)
pred.red.app <- predict(rf.app.1, fb.val)
pred.init.tot <- predict(rf.tot, fb.val)
pred.red.tot <- predict(rf.tot.1, fb.val)
# Create a vector of model names
mods.v <- c("Initial - Approved Conversion",
            "Reduced - Approved Conversion",
            "Initial - Total Conversion",
            "Reduced - Total Conversion")
# Create a vector of RMSE values for each model
rmse.v <- c(rmse(fb.val$app_conv, pred.init.app),
            rmse(fb.val$app_conv, pred.red.app),
            rmse(fb.val$total_conv, pred.init.tot),
            rmse(fb.val$total_conv, pred.red.tot))
# Add vectors to a data frame, output a summary table
sum.v <- data.frame(mods.v, rmse.v)
sum.v %>%
  kable(col.names = c("Model Name", "RMSE"))
```

Model Name	RMSE
Initial - Approved Conversion	1.656308
Reduced - Approved Conversion	1.659218
Initial - Total Conversion	3.776401
Reduced - Total Conversion	3.859157

```
## Random Forest Model Assessment - Testing Data
# Generate predictions using testing set
pred.test.app <- predict(rf.app, fb.test)
pred.test.tot <- predict(rf.tot, fb.test)
# Create a vector of model names
mods.t <- c("Initial - Approved Conversion",
            "Initial - Total Conversion")
# Create a vector of RMSE values for each model
rmse.t <- c(rmse(fb.test$app_conv, pred.test.app),
            rmse(fb.test$total_conv, pred.test.tot))
# Add vectors to a data frame, output a summary table
sum.t <- data.frame(mods.t, rmse.t)
sum.t %>%
  kable(col.names = c("Model Name", "RMSE"))
```

Model Name	RMSE
Initial - Approved Conversion	1.430102
Initial - Total Conversion	3.023581

```
## Random Forest Variable Importance
```

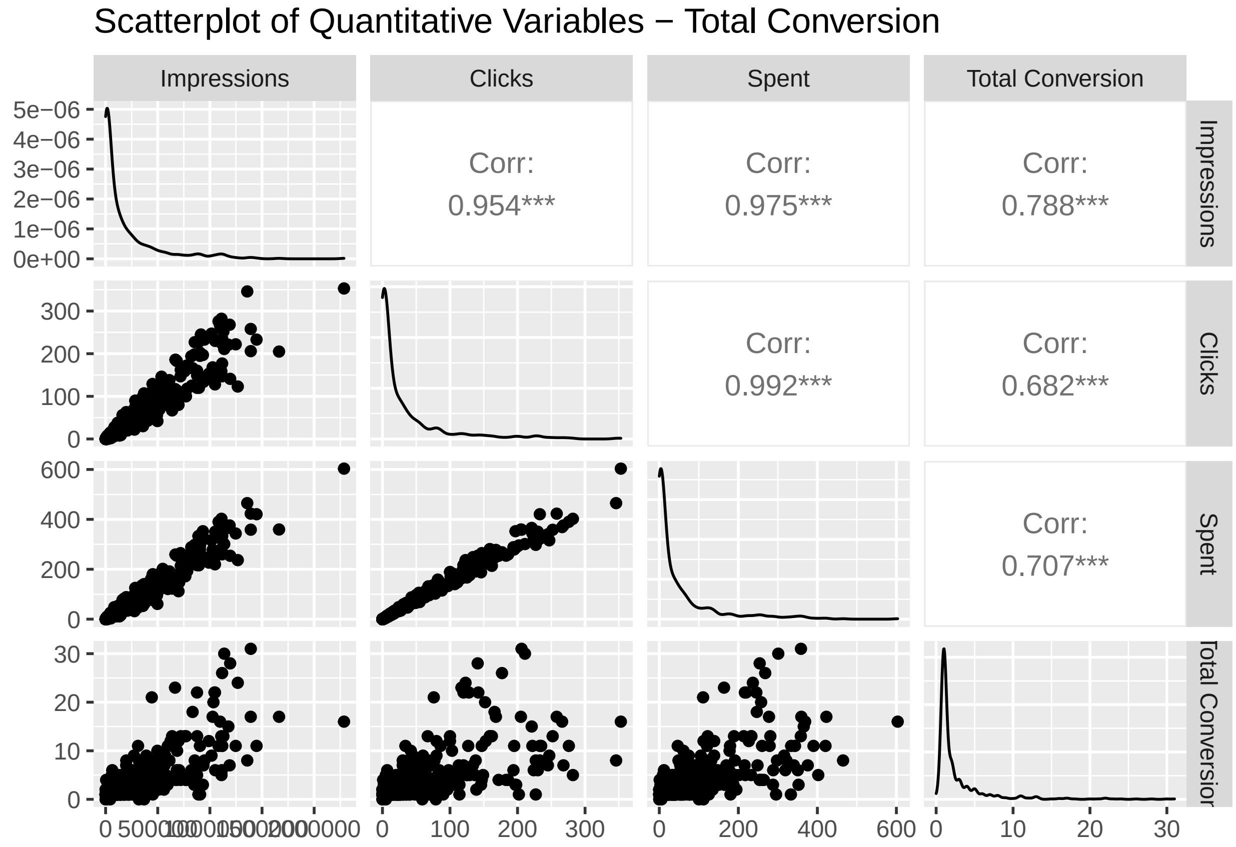
```
# Create a vector of variable names
varnam <- c("Impressions", "Spent", "Clicks", "Age",
            "Campaign 2", "Campaign 1", "Male")
# Create a vector of variable importance measures for each model
# Based on %IncMSE - higher values indicate higher importance
# values obtained by passing each model through an importance() function call
app.imp <- c(21.667, 15.536, 15.235, 13.526, 9.747, -0.153, 4.458)
tot.imp <- c(23.868, 17.291, 15.348, 18.842, 12.147, 5.464, 2.760)
# Add vectors to a data frame, output a summary table
sum.imp <- data.frame(varnam, app.imp, tot.imp)
sum.imp %>%
  kable(col.names = c("Variables",
                      "Importance - Approved Conversion",
                      "Importance - Total Conversion"))
```

Variables	Importance - Approved Conversion	Importance - Total Conversion
Impressions	21.667	23.868
Spent	15.536	17.291
Clicks	15.235	15.348
Age	13.526	18.842
Campaign 2	9.747	12.147
Campaign 1	-0.153	5.464
Male	4.458	2.760

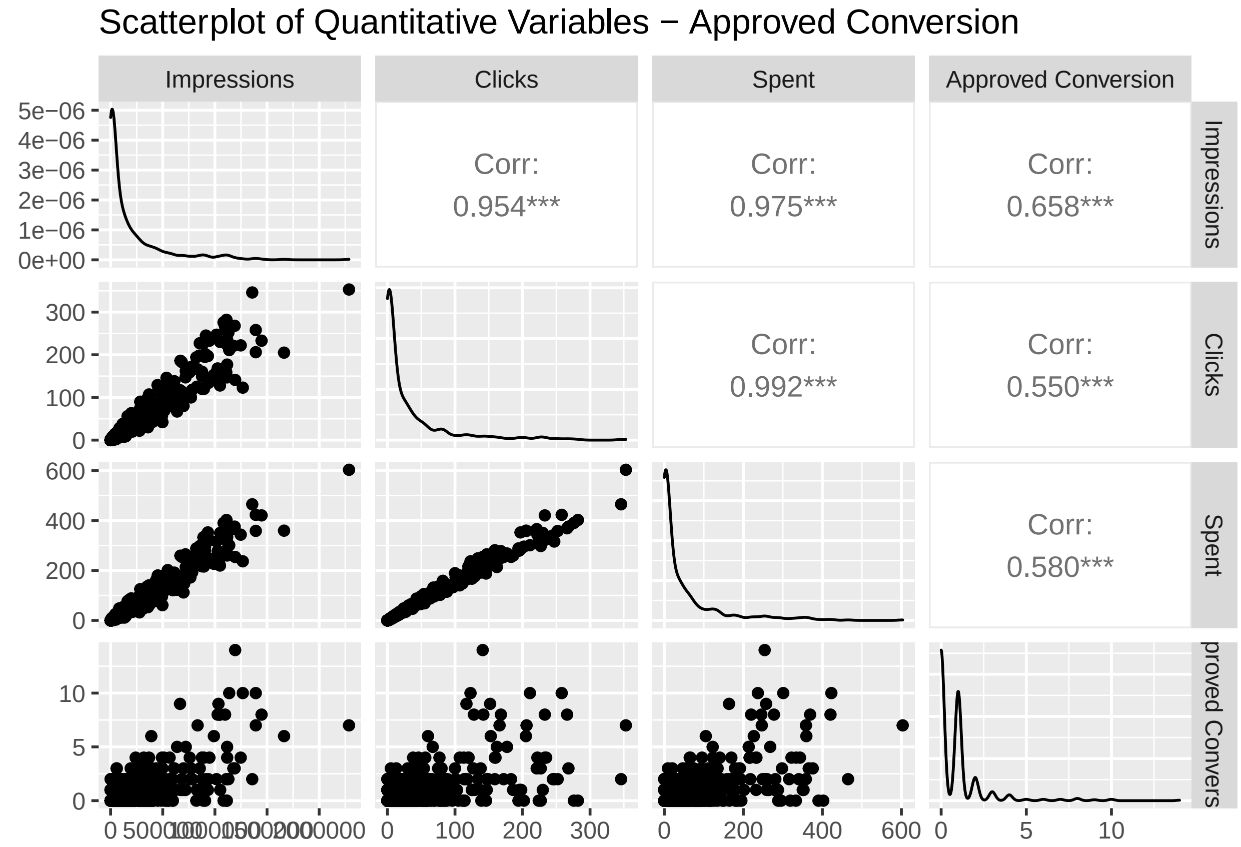
While our goal does not necessarily lie in predicting some value, these assessment measures still give us valuable information regarding the soundness of these predictive models. Since we are interested in making inferences about the relationship between our two response variables and the selected predictors, it is important that we use the best possible models as a basis for drawing these conclusions.

Finally, we ran a number of linear regression models based on the training subset and assessed them based on their r-squared values. We include some elements of our exploratory process, as well as our process for checking the regression assumptions. Summaries of the model assessment results at each stage are organized and presented in tables.

```
## Linear Regression Model - Correlation
#Total Conversion
fb.train %>%
  select(impressions, clicks, spent, total_conv) %>%
  ggpairs(title = "Scatterplot of Quantitative Variables - Total Conversion",
      columnLabels = c("Impressions", "Clicks", "Spent", "Total Conversion"))
```



```
#Approved Conversion
fb.train %>%
  select(impressions, clicks, spent, app_conv) %>%
  ggpairs(title = "Scatterplot of Quantitative Variables - Approved Conversion",
      columnLabels = c("Impressions", "Clicks", "Spent", "Approved Conversion"))
```



Having explored the strength of correlation between the quantitative variables and either of our response variables, we move on to building our total conversion models.

```
## Linear Regression - Total Conversion Model building
# Fit total conversion linear regression model using all primary predictors
# Use training data
LinReg.mod <- lm(total_conv ~ age + male + impressions + clicks + spent +
                   is_c1 + is_c2,
                 data = fb.train)

# Fit reduced model by removing male
LinReg.mod1 = lm(total_conv ~ age + impressions + clicks + spent +
                   is_c1 + is_c2,
                 data = fb.train)
# anova(LinReg.mod1, LinReg.mod)
# reduced model is better
# Fit reduced model by removing campaign dummy variables
LinReg.mod2 = lm(total_conv ~ age + impressions + clicks + spent,
                data = fb.train)
# anova(LinReg.mod2, LinReg.mod1)
# full model is better

# Check multicollinearity for the best reduced model so far
# vif(LinReg.mod1)
# highest multicollinearity with spent, then clicks and impressions
# remove spent

# Fit reduced model by removing spent
LinReg.mod3 = lm(total_conv ~ age + impressions + clicks + is_c1 + is_c2,
                data = fb.train)
# vif(LinReg.mod3) #no multicollinearity
# anova(LinReg.mod3, LinReg.mod1)
# full model is still better
```

At this point, we proceed with total conversion models mod1 and mod3 and perform intermediate assessments using the training data. The full model mod1 is chosen because it has a higher R-Squared, despite having some degree of multicollinearity. The reduced model mod3 was chosen because it is simpler and has a far lower degree of multicollinearity.

```
## Linear Regression -  Total Conversion Model Assessment - Training Data
# Add predicted values and residuals to training data
LinReg.add1 <- fb.train %>%
  add_predictions(LinReg.mod1) %>%
  add_residuals(LinReg.mod1) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Initial Total Conversion")
LinReg.add2 <- fb.train %>%
  add_predictions(LinReg.mod3) %>%
  add_residuals(LinReg.mod3) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Final Total Conversion")

# Combine prediction information
LinReg.add <- LinReg.add1 %>%
  bind_rows(LinReg.add2) %>%
  group_by(method)

rmse <- yardstick::rmse

# Residual standard error
# Use training data
LinReg.add %>%
  rmse(truth = total_conv, estimate = pred_total)

## # A tibble: 2 x 4
##   method                          .metric .estimator .estimate
##   <chr>                           <chr>   <chr>          <dbl>
## 1 LinReg Final Total Conversion   rmse    standard        2.21
## 2 LinReg Initial Total Conversion rmse    standard        2.14
```

We conclude our initial assessments of our total conversion models here. We note that the full total conversion model appears to perform better, at least based on the training data. We will reassess these models later with the validation set before making a final decision. Next, we move on to building our approved conversion models.

```
## Linear Regression - Approved Conversion Model building
# Fit approved conversion linear regression model using all primary predictors
# Use training data
LinReg.mod.a <- lm(app_conv ~ age + male + impressions +
                     clicks + spent + is_c1 + is_c2,
                   data = fb.train)
summary(LinReg.mod.a)

##
## Call:
## lm(formula = app_conv ~ age + male + impressions + clicks + spent +
##     is_c1 + is_c2, data = fb.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9607 -0.4984 -0.2485  0.5243  7.6640
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.115e-01  3.041e-01   1.682 0.093017 .
## age         -1.357e-02  7.626e-03  -1.780 0.075533 .
## male         4.039e-02  9.036e-02   0.447 0.654993
## impressions  9.571e-06  7.524e-07  12.721  < 2e-16 ***
## clicks      -2.125e-03  6.991e-03  -0.304 0.761219
## spent       -1.935e-02  6.045e-03  -3.202 0.001430 **
## is_c1        4.441e-01  1.944e-01   2.285 0.022624 *
## is_c2        3.765e-01  9.972e-02   3.776 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 678 degrees of freedom
## Multiple R-squared:  0.5239, Adjusted R-squared:  0.519
## F-statistic: 106.6 on 7 and 678 DF,  p-value: < 2.2e-16

# Fit reduced model by removing clicks
LinReg.mod.a1 <- lm(app_conv ~ age + male + impressions +
                      spent + is_c1 + is_c2,
                    data = fb.train)
# anova(LinReg.mod.a, LinReg.mod.a1)
# reduced model is better

# Fit reduced model by removing male
LinReg.mod.a2 <- lm(app_conv ~ age + impressions +
                      spent + is_c1 + is_c2,
                    data = fb.train)
# anova(LinReg.mod.a2, LinReg.mod.a1)
# reduced model is better
# Fit reduced model by removing age
LinReg.mod.a3 <- lm(app_conv ~ impressions +
                      spent + is_c1 + is_c2,
                    data = fb.train)
# anova(LinReg.mod.a3, LinReg.mod.a2)
# reduced model is better
# no more insignificant variables to remove

# Check multicollinearity for the best reduced model so far
# vif(LinReg.mod.a3)
# high multicollinearity between impressions and spent

# Fit reduced model by removing impressions
LinReg.mod.a4 <- lm(app_conv ~ spent + is_c1 + is_c2,
                    data = fb.train)
# summary(LinReg.mod.a4)
# removing impressions drastically lowers Adj. R-Squared
# vif(LinReg.mod.a4)
# very low vif values across the board
```

At this point, we proceed with the approved conversion models mod.a3 and mod.a4 and perform intermediate assessments using the training data. The reduced model mod.a3 is chosen because it is simpler than the full model with comparable R-Squared while also having manageable multicollinearity. The other reduced model mod.a4 is chosen because it has very a very low degree of multicollinearity, but also a much lower R-Squared. We want to compare these and see which one performs better.

```
## Linear Regression -  Approved Conversion Model Assessment - Training Data
# Add predicted values and residuals to Training data
LinReg.add.a1 <- fb.train %>%
  add_predictions(LinReg.mod.a3) %>%
  add_residuals(LinReg.mod.a3) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Initial Approved Conversion")
```

```
LinReg.add.a2 <- fb.train %>%
  add_predictions(LinReg.mod.a4) %>%
  add_residuals(LinReg.mod.a4) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Final Approved Conversion")

# Combine prediction information
LinReg.add.a <- LinReg.add.a1 %>%
  bind_rows(LinReg.add.a2) %>%
  group_by(method)

# Residual standard error
# Use training data
LinReg.add.a %>%
  rmse(truth = total_conv, estimate = pred_total)
```

```
## # A tibble: 2 x 4
##   method                             .metric .estimator .estimate
##   <chr>                              <chr>   <chr>          <dbl>
## 1 LinReg Final Approved Conversion   rmse    standard        3.87
## 2 LinReg Initial Approved Conversion rmse    standard        3.63
```

```
# initial performs better based on training data - in terms of RMSE
```

We conclude our initial assessments of our approved conversion models here. We note that the slightly more complex model mod.a3 appears to perform better, at least based on the training data. We reassess these models, as well as our total conversion models in the next section with the validation data.

```
## Linear Regression Model Assessment - Validation Data
# Total Conversion
LinReg.add3 <- fb.val %>%
  add_predictions(LinReg.mod1) %>%
  add_residuals(LinReg.mod1) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Initial Total Conversion")

LinReg.add4 <- fb.val %>%
  add_predictions(LinReg.mod3) %>%
  add_residuals(LinReg.mod3) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Final Total Conversion")

# Combine prediction information
LinReg.add.v <- LinReg.add3 %>%
  bind_rows(LinReg.add4) %>%
  group_by(method)

# Residual standard error
LinReg.add.v %>%
  rmse(truth = total_conv, estimate = pred_total)

## # A tibble: 2 x 4
##   method                          .metric .estimator .estimate
##   <chr>                           <chr>   <chr>          <dbl>
## 1 LinReg Final Total Conversion   rmse    standard        2.21
## 2 LinReg Initial Total Conversion rmse    standard        2.14

# Approved Conversion
LinReg.add.a3 <- fb.val %>%
  add_predictions(LinReg.mod.a3) %>%
  add_residuals(LinReg.mod.a3) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Initial Approved Conversion")

LinReg.add.a4 <- fb.val %>%
  add_predictions(LinReg.mod.a4) %>%
  add_residuals(LinReg.mod.a4) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Final Approved Conversion")

# Combine prediction information
LinReg.add.a.v <- LinReg.add.a3 %>%
  bind_rows(LinReg.add.a4) %>%
  group_by(method)

# Residual standard errors
LinReg.add.a.v %>%
  rmse(truth = total_conv, estimate = pred_total)

## # A tibble: 2 x 4
##   method                             .metric .estimator .estimate
##   <chr>                              <chr>   <chr>          <dbl>
## 1 LinReg Final Approved Conversion   rmse    standard        5.39
## 2 LinReg Initial Approved Conversion rmse    standard        4.60
```

We conclude our validation set assessments here and decide on our final total and approved conversion models. We decided that the total conversion model mod1 with the age, impressions, clicks, spent, and campaign variables performed better than its counterpart. We also decided that the approved conversion model mod.a3 with the impressions, spent, and campaign variables performed best in its group. We perform our final assessments for these models in the next section with the testing data.

```
## Linear Regression Model Assessment - Testing Data
# Total Conversion
LinReg.add5 <- fb.test %>%
  add_predictions(LinReg.mod1) %>%
  add_residuals(LinReg.mod1) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Initial Total Conversion")

# Residual standard error
```

```
LinReg.add5 %>%
  rmse(truth = total_conv, estimate = pred_total)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        2.86
```

```
# Approved Conversion
LinReg.add.a5 <- fb.test %>%
  add_predictions(LinReg.mod.a3) %>%
  add_residuals(LinReg.mod.a3) %>%
  rename(pred_total = pred,
         residuals = resid) %>%
  mutate(method = "LinReg Final Approved Conversion")

# Residual standard errors
LinReg.add.a5 %>%
  rmse(truth = total_conv, estimate = pred_total)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        4.32
```

The code in the following section corresponds to the linear regression model summary tables used in the body of the report.

```
## Linear Regression Model - Summary Tables
mods.lr <- c("Initial - Approved Conversion",
             "Final - Approved Conversion",
             "Initial - Total Conversion",
             "Final - Total Conversion")
mods.lr1 <- c("Initial - Approved Conversion",
             "Final - Approved Conversion",
             "Initial - Total Conversion",
             "Final - Total Conversion")
rmse.lr.tr <- c(3.63, 3.87, 2.14, 2.21)
rmse.lr.v <- c(4.60, 5.39, 2.14, 2.21)
rmse.lr.te <- c(4.32, 2.86)

sum.lr.tr <- data.frame(mods.lr, rmse.lr.tr)
sum.lr.v <- data.frame(mods.lr, rmse.lr.v)
sum.lr.te <- data.frame(mods.lr1, rmse.lr.te)

sum.lr.tr %>%
  kable(col.names = c("Model Name", "RMSE"))
```

Model Name	RMSE
Initial - Approved Conversion	3.63
Final - Approved Conversion	3.87
Initial - Total Conversion	2.14
Final - Total Conversion	2.21

```
sum.lr.v %>%
  kable(col.names = c("Model Name", "RMSE"))
```

Model Name	RMSE
Initial - Approved Conversion	4.60
Final - Approved Conversion	5.39
Initial - Total Conversion	2.14
Final - Total Conversion	2.21

```
sum.lr.te %>%
  kable(col.names = c("Model Name", "RMSE"))
```

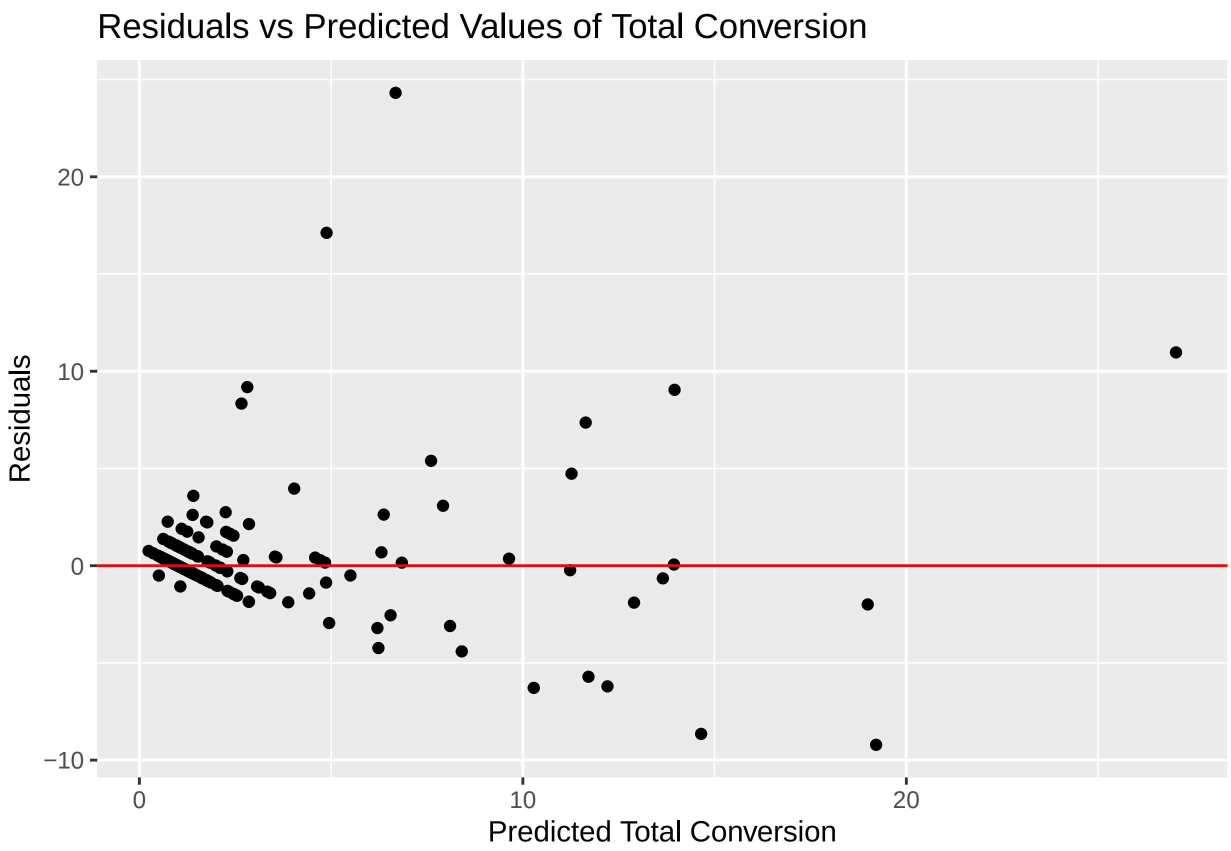
Model Name	RMSE
Initial - Approved Conversion	4.32
Final - Approved Conversion	2.86

Model Name	RMSE
Initial - Total Conversion	4.32
Final - Total Conversion	2.86

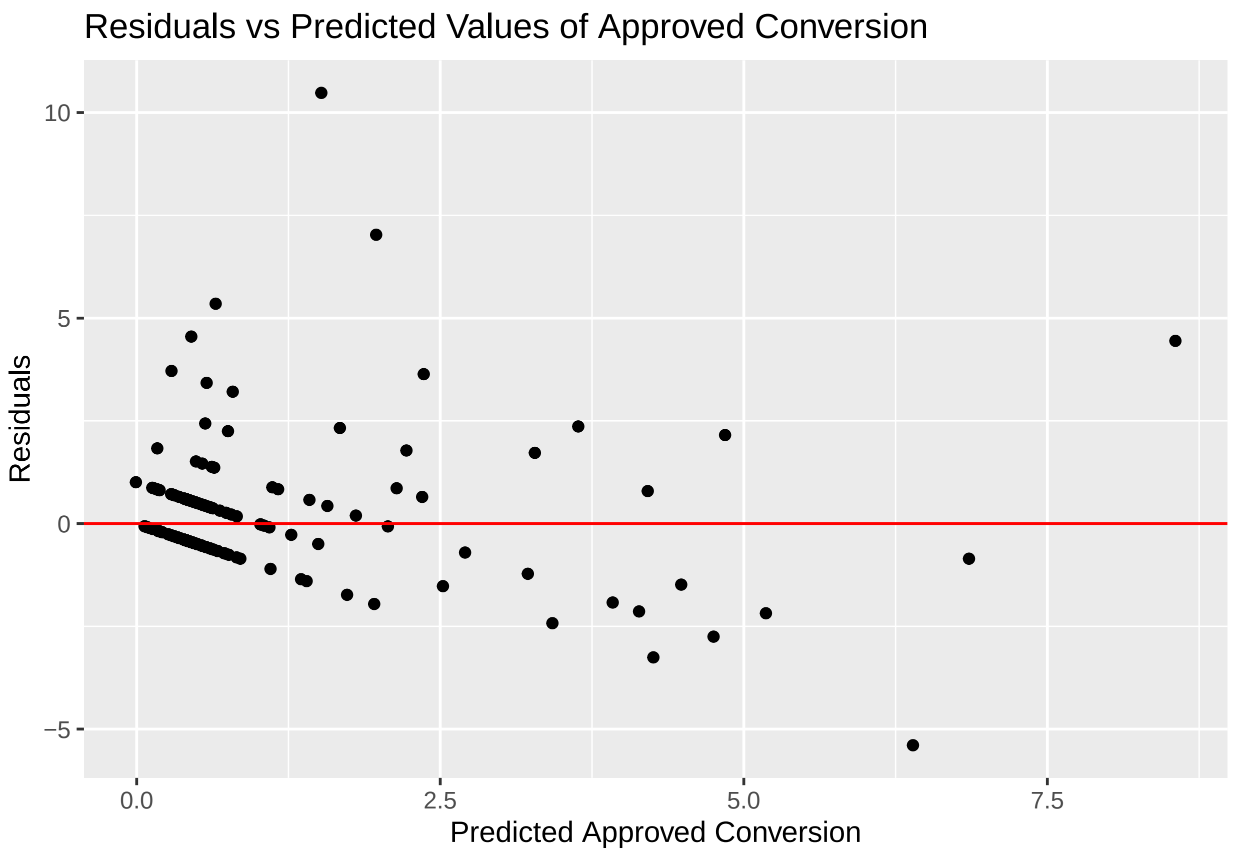
```
Final_Total = coef(summary(LinReg.mod1))

row.names(Final_Total) <- c("(Intercept)", "Age", "Impressions", "Clicks",
                            "Spent", "Campaign 1", "Campaign 2")

Final_TotalR = data.frame("R2" = 0.7556,
                          "Adj. R2" = 0.7535,
                          "P-Value" = "< 2.2e-16")

Final_Approved = coef(summary(LinReg.mod.a3))

row.names(Final_Approved) <- c("(Intercept)", "Impressions", "Spent",
                               "Campaign 1", "Campaign 2")

Final_ApprovedR = data.frame("R2" = 0.6098,
                             "Adj. R2" = 0.6075,
                             "P-Value" = "< 2.2e-16")
```

Additionally, we checked the linear regression assumptions for our final predictive models. While the residual plots below seem to indicate some initial fanning, the patterns appear to revert as the predicted values become larger. This seems to suggest that we might not have a very serious violation of constant variance in our residuals. As such, we decided that the linear regression methodology was appropriate to continue with this analysis.

```
## Linear Regression Model - Assumptions
# Total Conversion
LinReg.add5 %>%
  ggplot(aes(x = pred_total, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red")  +
  xlab('Predicted Total Conversion') + ylab('Residuals') +
  ggtitle('Residuals vs Predicted Values of Total Conversion')
```



```
# Approved Conversion
LinReg.add.a5 %>%
  ggplot(aes(x = pred_total, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  xlab('Predicted Approved Conversion') + ylab('Residuals') +
  ggtitle('Residuals vs Predicted Values of Approved Conversion')
```


Residuals vs Predicted Values of Approved Conversion
Residuals
10
5
0
−5
0.0
2.5
5.0
7.5
Predicted Approved Conversion

Appendix D: Comment incorporation

Our team received a number of helpful comments regarding our first submission which helped us rein in the scope of our analytical project and decide what our focus was going to be. A number of factors led to us completely overhauling the concept for our project, which made the context of many of these comments obsolete. Nevertheless, we acknowledge the impact they made on the second version of our first deliverable here.

The Introduction section of our first deliverable described the context of our business problem and built up a narrative that incorporated the challenges retailers faced during the COVID-19 pandemic. At the time, we wanted to highlight the importance of maintaining customer loyalty amidst the crashing wave of digital commerce, and how these traditional retail businesses could benefit from adopting technology and data-driven strategies. One reviewer suggested that we include an in-depth explanation of CRM database systems in this section. Our team decided against this, in favor of keeping the section concise and proceeding under the assumption that our readers would already be familiar with this technology. Other comments suggested that we write less about the growing importance of technology and instead focus more on how brick-and-mortar retailers should be operating during the pandemic. This was a fair point, and we incorporated this feedback by removing some extra content on the growth of e-commerce and highlighting the struggles of traditional retailers. We also received quite a bit of positive feedback regarding the way we set up the background and context for our project, which was great to see.

The Business Problem section of our first deliverable began to introduce a few ways that traditional retailers would benefit from leveraging data analytics. These included the ability to deploy more effective online marketing campaigns by identifying key demographic groups and reflecting on past marketing efforts by analyzing key performance indicators. As mentioned earlier, we also wrote about the importance of building and maintaining customer loyalty during the pandemic times. We received a number of comments here that alerted us to the fact that we weren't conveying our business problem in a clear enough manner. We attempted to rectify this by restructuring the section and stating our business problem up front. However, it doesn't seem as though we did a good enough job with the second version of this submission, as we received additional feedback regarding this same point. We hope that our goals come across more clearly now that we've made further adjustments to the section. Additionally, we received a lot of feedback regarding our decision to pursue solutions to the issue of customer loyalty maintenance. Our group agreed that this topic was slightly too vague, and that we would need to shift our focus to something else. This led us to commit to the aforementioned overhaul, which involved us focusing more on the performance of the firm that our dataset was based upon. Finally, we received a suggestion that said we should mention how e-commerce was becoming more common and accepted in the modern day. We decided that this would be unnecessary and redundant given that we provided similar context in the introduction section.

The Intended Audience section of our deliverable described the rapid growth of digital commerce across multiple industries during the COVID-19 pandemic, as well as the idea that traditional retailers had much to gain by building out an online platform. We received a few comments that raised concerns about the scope of our analysis. One acknowledged that our analysis would indeed benefit retail businesses across a wide variety of industries, and advised us to instead focus on just a few industries. The other suggested that we differentiate large brick-and-mortar retailers from smaller family-owned businesses. I believe that our team addressed both of these suggestions in our concept overhaul by choosing to focus on the firm our dataset was based upon. We received a comment regarding the verbosity of this section, and we incorporated the feedback by making efforts to shorten the section. The length of the final version of this section should reflect these changes. One comment suggested that we provide a citation to support one of the claims we made about how building an online presence could boost a retailer's sales performance. This was information we gathered from a Deloitte report that was cited earlier in the document, so we inserted an extra citation here to make this clearer. Ultimately, however, this section was removed in order to tighten the focus of our final deliverable.

We also received general formatting feedback regarding the spacing between paragraphs and the content of our reference entries. We made sure to rectify the spacing issues in the later deliverables. Our team also updated the reference entries to accord with the new guidelines that were posted later into the semester.

We received quite a lot of good feedback on the Data section of the report which were relatively easy to incorporate. We responded to most of Professor Martinet's comments by implementing a wide range of stylistic changes. Many of these changes involved replacing references to raw variable names with descriptive information regarding the content of the variables in question. We modified our figures by adjusting column labels where appropriate and appending captions where needed.

The Professor's comments regarding our content allowed us to revise the section and make our writing more clear and cohesive. Most importantly, we were able to internalize some feedback and develop a clearer sense of direction regarding our analytic process. Additionally, this section of the report initially had a data dictionary that listed out each of our variables. We incorporated the Professor's feedback here by reorganizing the variable descriptions into paragraph form.

We made changes to the data preparation details section by adding more text explanations to accompany our commented code chunks. We hope that the changes help future readers to better understand our data preparation process. Furthermore, we took some helpful advice here and moved some of our exploratory findings in Excel into the Descriptive Analytics section of the report.

Finally, we received a number of helpful comments from Professor Martinet regarding our in-text citations and the format of our bibliography. We were able to clarify some points of confusion and adjust the way we documented our sources both in the body of our text and in our references file. We also removed a number of unnecessary citations pertaining to the data manipulation steps we carried out in Excel and R.

We also received some feedback from the TA and our peers, which were slightly less helpful. For the data section, we were advised to explain how key variables like conversion, interest, and spent could be useful. This was valuable feedback and was easily incorporated into our deliverable, as we were able to simply just give a rationale for why we were interested in these variables. In addition, we were told that repeating the ad_id variable in the second row of kable output was useful. We eventually reorganized our kable output entirely, so this feedback didn't end up being very useful.

For the data collection section, we primarily focused on adding more in-text citations into our deliverable to properly give credit for the information we were borrowing. We were also advised to consider adding some context by discussing data ethics or some background on Google Adwords. This ultimately led to us doing more background research on these topics and doing our best to incorporate them into the section. We discovered that the dataset actually pertained to Facebook ads, which was helpful for our later deliverables as well.

For the data preparation section, we were asked to consider transforming some of the values to be easier to interpret. We thought this made sense, as values that were easier to interpret would be more meaningful and useful. These were relatively simple changes to make in both R and Excel, and were documented in the relevant code walkthrough sections.

For the data preparation details in the R section, we were told that transforming the age to an integer was useful for analysis, but decreased the interpretability of the data, so we should leave it as is. We understood why the reviewer thought this was the case, but we ultimately decided against incorporating this feedback because we knew we would eventually transform this variable for our linear regression anyway.

For the data preparation details in Excel section, we were told that the images were showing up in weird places and were too small. These were two things that were fairly simple to address in the R-markdown file with some changes to our R code that ultimately ended up in a more aesthetically pleasing output file.

For the Descriptive Analytics section, we incorporated Professor Martinet's comment by changing some of the wording of the introduction. We definitely had lost sight of our audience and resorted to using statistical language throughout this section. Her comments helped make this section more understandable to a statistically illiterate reader. The TA's comment mentioned that we forgot to include insights, but we had included these further down in the report. However, we realized that we didn't provide any intermediate explanations after each figure. As such, we incorporated this feedback by including brief explanations after each graphic. Commenter 2 for this section was especially helpful as they suggested that we visualize gender and age disparities using a histogram. This made our visualizations more understandable to a non-statistician, as histograms are much more easily interpretable than box plots. Commenter 3 would have been helpful if we were writing this report for a statistical audience but since we explain the diagram further on in the report, we felt no need to get into the details of the transformation. Commenter 4's comment about proportions were largely incoherent so we elected to ignore it.

Throughout this project it was easy to feel lost as none of us had a real business analytics background. Professor Martinet's comment let us know we were on the right track and gave us the confidence to extend this type of analysis to further deliverables. We incorporated the

TA's comments by separating our insights by our variables rather than by our figures. This forced us to think about how the figures related to each other and our business problem. Wordiness and statistical language were definitely a problem throughout this report and we appreciated commenter 3's suggestion. We changed the wording here to make the insights section more readable. We appreciated that commenter 4 could understand our insights even if others found them wordy. Their suggestion of comparing our rates to the general industry rates was especially helpful for garnering insights from the data, and we included this in our revised submission.

Professor Martinet's comments suggested minor code changes that we incorporated in our revisions. The TA's comments helped us a lot as we were previously unaware of the facet_wrap function. This improved our visualizations to make them slightly easier to interpret. Commenter 3 brought our attention to the fact that we were missing comments on our R code. We incorporated this feedback by adding more comments to our coding details section. We felt that this allowed the reader to follow our process with more ease. Commenter 4 made us realize that we were missing captions for certain plots. We quickly fixed this for our revised deliverable.

From all the feedback and comments on our deliverable 4 — both on the draft and on the final deliverable — we noticed a few consistent suggestions. At various points in our deliverable, our wording and explanations became a little too technical. For example, we used words such as RMSE, ANOVA, and so on. We replaced these words with phrases that were less technical, such as "prediction error" to replace RMSE. We also received a couple of comments regarding some confusing wording or redundant sentences in our deliverable. We attempted to go through our deliverables and edit phrases that were confusing to the readers. We also removed any redundant sentences we were able to find. Finally, there were some formatting issues throughout the deliverable in terms of how our tables and graphs were structured. Many of our tables used raw variable names such as "total_conv" that wouldn't be very presentable to any executives reading our report. There were also some formatting issues in our analytics detail section which was easily fixable as well. For example, we were advised to add titles and axis labels for our residual plots and remove the captions of tables in our appendices.

Our draft for deliverable 4 contained repetitive information in the process and predictive analytics sections. We were advised to remove the repetitive information from one section, which we did. Additionally, our process section described the details of our linear regression and random forest methodologies rather than focusing on the actual process we followed to build our models. Similarly, we had described a lot of our modeling process in the assessments section, rather than the process section. We received various TA and Professor comments advising us to edit these sections. These comments were very helpful because they clarified the expectations regarding which content was to be included in each section. We edited our predictive analytics section to summarize our goals for our analysis, i.e. to determine what variables are significant to our two conversion rates. We then edited the process section to include a high level explanation of how we obtained our model. To ensure we did not make this section technical, we simply mentioned that we started off with a model with all of our variables and then created different models to compare after removing insignificant variables. A lot of this information was moved from our assessments section to our process section, as suggested.

In our assessments section, we were advised to remove the training set assessments as well as to move our testing set assessments from the results section to the assessments section. We removed the training set assessments because we realized that the assessment measures generated through this process were only used to gain information about how our model was being trained. However, they did not add to our analysis of finding the best model to conclude which variables are the most important predictors for our two conversion rates. We also removed the testing set assessment results from the results section and placed them within the assessments section. The comments helped us realize that it would be better for us to use the results section to discuss our final models and the variables contained therein. With this, the Professor suggested that we needed to make clear what exactly RMSE (prediction error) meant for our analysis since it did not actually give us information about how well our model predicts. We responded to this suggestion by explaining the importance of RMSE in a couple of sentences. In essence, we explained that we are interested in making inferences about the relationship between our two response variables and the selected predictors and that assessment measures like RMSE are relevant because they give us valuable information regarding the soundness of our models.

One peer also suggested we add information on linear regression assumptions through plots and graphs in our assessments section. We considered this, however, we thought it was too technical to add to this section and instead decided to elaborate on our assumptions check in the analytics detail section. Nevertheless, we did our best to keep our explanation of the regression assumptions as non-technical as possible. However, we did end up including a pair of scatter plots in the body of the report in order to demonstrate the linear relationships between our variables. We decided to only keep this in the section as the plot is easily digestible, even for someone who is not well versed in statistics. Initially, we included only a single plot showing all of the quantitative variables together. However, the Professor commented to make two separate graphs to make clear that the conversion rates were not being used in each others' models. We made the necessary changes as we realized that it would make our analysis more comprehensible as it reiterates that there are two separate response variables being analyzed. Additionally, separating it into two plots made the individual relationships between the variables more clear.

Another comment spoke to the clarity of the random forest analysis, claiming that it was unclear whether the gender variable was important in our final model or if it was removed. We clarified this in the report by explaining that the gender and campaign dummy variables were removed for the pruned and reduced models because they had lower variable importance. However, our validation set model assessments indicated that the full models containing these variables had a better fit to our data. Thus, our final models do include the gender and campaign variables.

One peer suggested adding the summary outputs for the regression models as well as the variable importance table for our random forest analysis in the results section. This was a very useful suggestion that we immediately incorporated into our deliverable as these two outputs summarize what our model finally gives us and answers our question of what variables are the most important.

Finally, we were advised by the Professor to rewrite the insights summary to explain the important variables separately instead of explaining the methods separately. This made a lot of sense to incorporate because a lot of the takeaways from both of the methodologies we employed were the same. By structuring this section around the variables rather than each method, we did not have to repeat the same takeaways multiple times. We also believe that these changes made our conclusions slightly easier to follow.