Gallery Furniture Sales Data Analysis

Huiching Kang, Caitlyn King, Jacob Ruiz

Lone Star College - North Harris
HONRH 2033

July 8, 2022

Abstract

Gallery Furniture, a Houston-based retail furniture store, tasked our team with providing insights about the effectiveness of their Home For the Holidays (HFTH) sales campaign. With the completion of an online form, potential customers can receive a coupon for $100 off a purchase of at least $1000. We were provided with deidentified data that included the coupon submission date, submission platform, zip code, sales order date, and sales total. We utilized data visualizations and a machine learning model to explore various trends in the provided data set (e.g. geography) and also to evaluate aspects of the HFTH campaign effectiveness. While Harris County had the most coupon form submissions, we found that the suburban counties outside of Houston performed better in both coupons per resident and coupon to purchase conversion rate. Recommendations for Gallery Furniture include input validation on the coupon forms, collecting additional data such as store location, and geographically targeted advertising.

Introduction

Gallery Furniture is a family-owned furniture store based in Houston, Texas. Founded by Jim “Mattress Mack” McIngvale, the company takes pride in selling affordable, made in America products. They have three locations; their original location is still the largest storefront, and offers customers a unique shopping experience with exotic animals and fresh desserts. Mattress Mack and Gallery Furniture have collaborated with many charitable organizations to provide community outreach programs, and launched their own program for the education of young people. A large amount of the store’s proceeds are invested into the betterment of the City of Houston.

Each year, Gallery Furniture runs over 100 advertising campaigns across the three store locations in an effort to provide customers with the best prices and promote their product. They advertise their campaigns on their website, cable television, social media platforms (e.g., Facebook, Instagram, Twitter), and send emails and text messages. With each campaign, a coupon of $100 off a purchase of $1000 or more is offered to customers who fill out a coupon voucher. The voucher form asks for their name, phone number, zip code, and email.

Gallery Furniture asked our team to provide them with insights into the effectiveness of their seasonal advertisements and recommendations for improved use of the voucher forms. Specifically, our team was tasked with providing insights to the success of the Home For The Holidays sales campaign.

Data and Data Preparation

Our team was provided a deidentified data set from the Home For The Holidays campaign voucher form. Before our team received the data set, the customers' names, phone numbers and email addresses had been removed. Instead, we received a unique number labeled as "LeadID" to identify the customer. Other variables included in the data set were the coupon submission date, zip code, and the media source from which the voucher was submitted. Each row in the dataset represented one form submission from one person. For customers who made a purchase, the dataset also included the Customer Code (an identifier used by the store), the sales order date, and the sales total. For those who did not complete a purchase, these cells were empty.

Data Cleaning

Before we could begin our analysis, some data cleaning was necessary. We cleaned and analyzed our data within the Jupyter Notebook environment. To manipulate the data, we used the Pandas library in Python.

The biggest problem in our data set was duplicated data. For some customers, there were over 20 rows of data, indicating that they had repetitively submitted the same form. Based on the time stamps, submissions were only seconds apart. It appeared that the customer may have been unsure whether the submission went through, and submitted again. To tackle our duplicated data problem, we first sorted by the form submission date. We then highlighted rows which had identical values for LeadID, Sales Total, and Sales Order Date. For each group, we kept only the last one because it would be the closest time to the purchase date.

Even after duplicate form submissions were removed, some customers still had multiple purchases. We were able to determine that they were separate submissions because the Sales Totals were different, and occasionally negative, indicating a return. In our analysis, we wanted each customer to only be counted once. So we grouped by the customer code and summed the purchases for each customer. After removing all the duplicates, our dataset had 1133 rows, representing 1133 unique people who submitted a form.

The next data cleaning task involved the zip codes. Some of the zip code submissions had four-digit extensions, and some were invalid because they included letters or the wrong number of digits. We deleted the four-digit extensions, and we changed the invalid zip codes to zeros, so they would be recognized as invalid in our analysis.

We also converted the form submission dates and sales dates to Python datetime objects, so they could be subtracted and sorted. Also, some of the values for utm_source were combined. For example, "mm_facebook" and "facebook" were all classified as "Facebook."

Data Analysis

After the data had been cleaned and prepared for analysis, we decided to direct our focus to the aspects of customer zip code, utm source, and time difference. We wanted to evaluate the effectiveness of the platforms Gallery Furniture had chosen for advertisements. We also wanted to learn the geographic distribution of Gallery Furniture customers. In our analysis, the Python packages we used included Seaborn, Geopandas, Pandas, Numpy, MatPlotLib, and Folium.

Time Difference

First, we calculated the time difference between coupon submission and sales order date to show how long it takes most customers to complete their purchase. In Figure 1, we see that a large proportion of customers purchase their furniture 0 days after they submitted the online form. Due to limitations in the data (we had both date and clock time for the form submission but only date for the furniture sales) the time difference is not exact. A time delta of 0 between form submission and purchase means that the customer completed their purchase within 24 hours of their coupon submission.

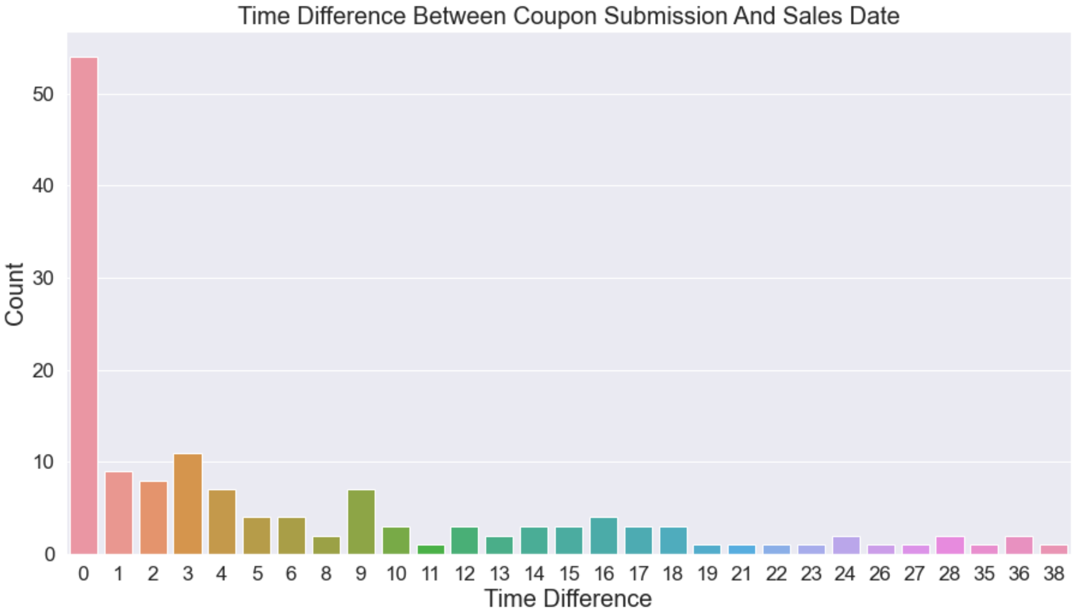

Figure 1: Frequency count for the time difference (in days) between online form submission date and sales date, for customers who submitted a form and purchased furniture.

The heat map in Figure 2 shows the relative frequencies of different sales totals compared to the time difference between coupon submission and completed purchase time. The x-axis shows the number of days between form submission and purchase. The y-axis represents the sales total amount for each customer. Some of these totals include a negative transaction which represents a return. The darker blue areas represent a larger number of customers; the lighter blue areas represent smaller numbers of customers.

The plots on this graph show that most customers who participated in this campaign made their purchase within one day of submitting their coupon voucher. Also, most customers made a purchase of $5000 or less. All the customers with purchases over $15,000 waited more than five days after submitting their coupon voucher.

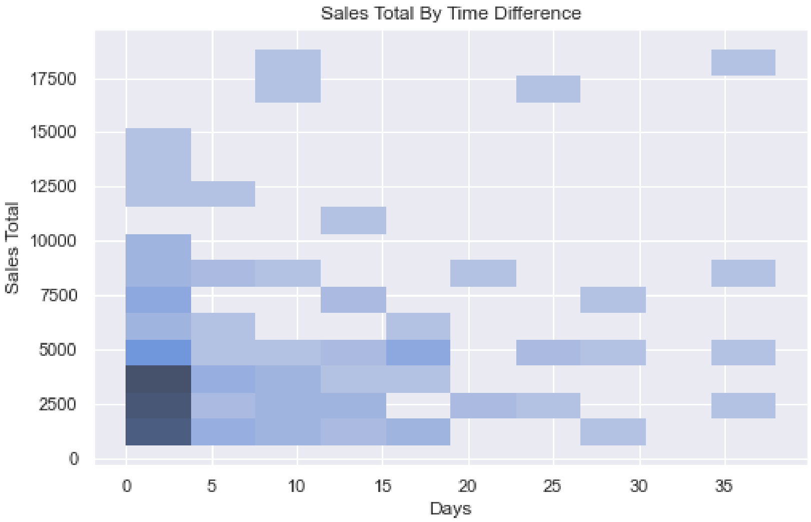

Figure 2: Frequency count for the time difference between online form submission date and sales date, for people who submitted forms and also bought furniture.

UTM Source

Next, we looked at the success of various coupon platforms for attracting potential customers. The bar plot in Figure 3 shows the number of submissions from each platform where advertisements for the campaign were available. These platforms include the company website, email, text messaging, and Facebook. The company's website received nearly 600 coupon form submissions, while Facebook received less than 50 submissions.

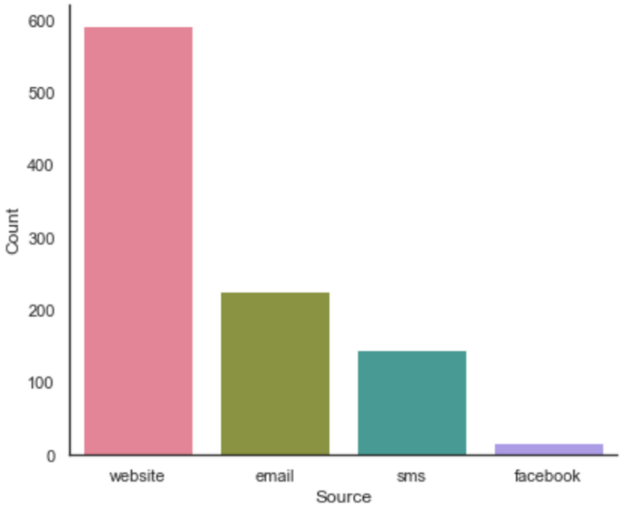

Figure 3: Frequency count for form submissions on various platforms.

Geographic Distribution of Home for the Holidays Campaign Participants

Next, our team examined the geographical distribution of the form submissions. Our team observed that other regions outside the Greater Houston Area participated in the Home for the Holidays campaign. The U.S. map in Figure 4 shows blue markings on the zip codes where at least one customer submitted a HFTH form. While the campaign participants were mostly concentrated in the Greater Houston area, there were also form submissions from California, Washington, Oklahoma, Nevada, Louisiana, Missouri, Florida, Georgia, Tennessee, Kentucky, Illinois, Indiana, Ohio, New York, and most surprisingly Alaska.

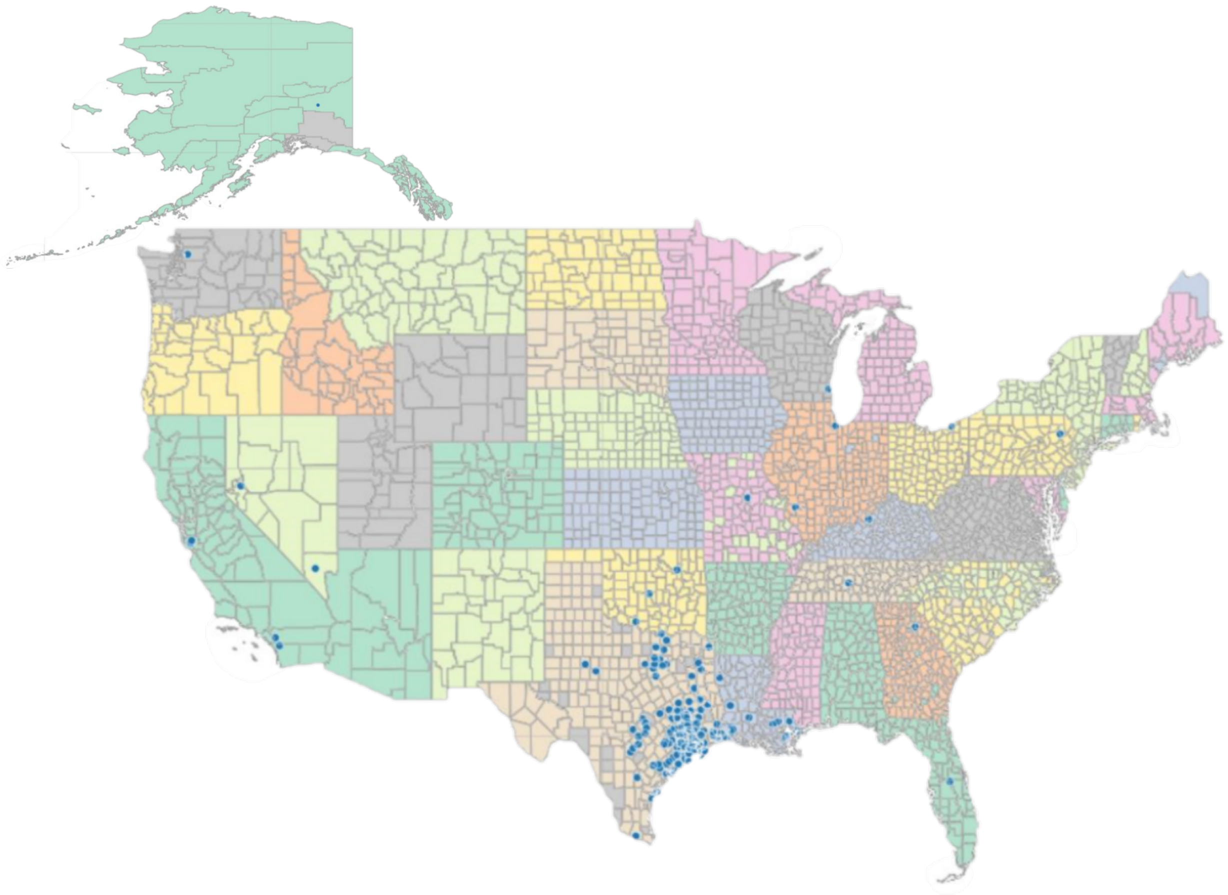

Figure 4: Geographic distribution of zip codes from coupon form submissions. Each dot indicates a zip code with at least one form submission.

Interactive Maps of Form Submissions

We wanted to interact with the data in more detail and examine closely the geographic distribution within Texas. We were interested in locating clusters of participants relative to geographic anchors such as roads and cities.

For this plot, we used Folium, a tool package in Python that enables information from a data set to be combined with a map. The user can interact with the open street map and see various geographic details along with the data.

Both Figures 5 and 6 show the counts of form submissions in different geographical regions.

A blue pin represents a single submission; a circle with a number represents multiple submissions. Each circle is associated with a polygon; the number in the circle is the count of form submissions within that polygon. The polygons can be seen by hovering over the circle in the interactive version of the map (link is available in Figure 6 caption).

Figure 6 shows the counts for Central and East Texas, including the Greater Houston region where Gallery Furniture is located. If we zoom in using the interactive version, we can see how these counts break down by subregion. The Houston area has the greatest counts among Central and East Texas.

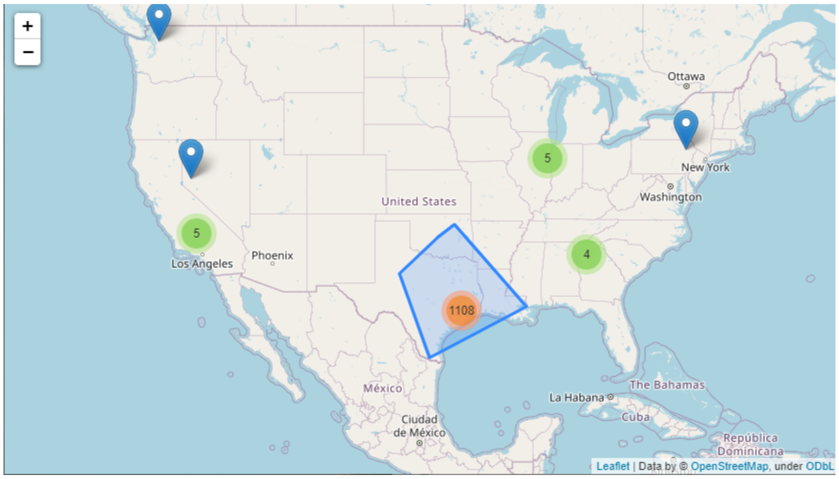

Figure 5: Count of form submissions by region in the U.S.

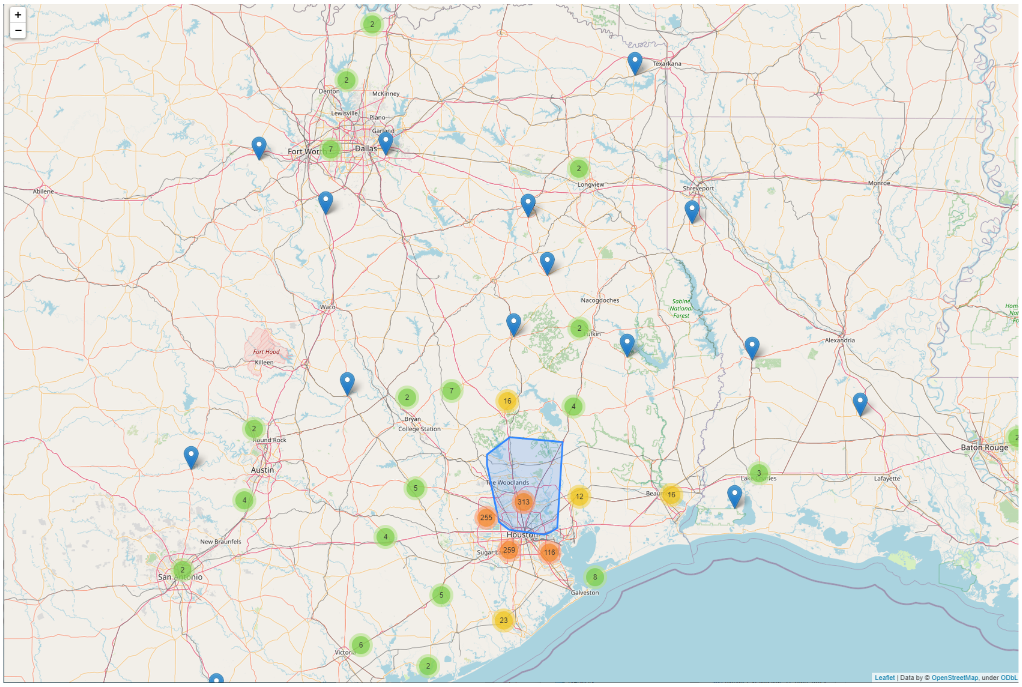

Figure 6: Count of form submissions by region for Central and East Texas.
Interactive version:
`http://nhmath.lonestar.edu/Faculty/TravisJ/PICMath/forms-by-region-HFTH.html`

Form Submissions and Sales for Texas Counties

Next, we looked into the frequency of form submissions from different counties in Texas. The Home for the Holidays campaign had 1133 submissions total. We found that the majority of people who submitted the coupon forms and completed a purchase were located in Harris County, where Houston is located. Other counties with large numbers of submissions were Fort Bend, Montgomery, and Brazoria, which are near Houston.

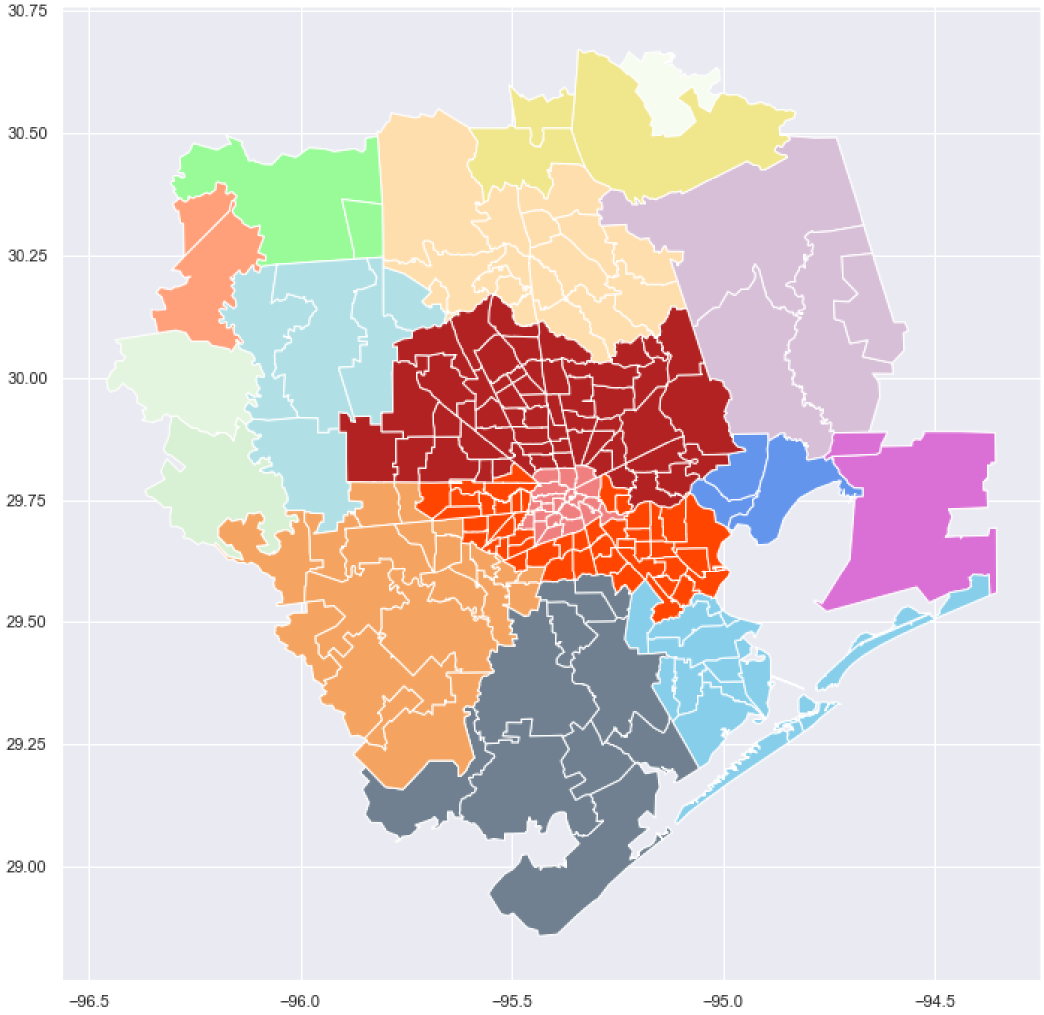

Figure 7: Map of the Greater Houston area divided into regions. Regions are composed of zip codes grouped as closely as possible by county. Harris County is subdivided into three regions.

Next, we analyzed the Greater Houston area more closely. We divided the Houston area into regions as shown in the map in Figure 7. The regions in the map are composed of zip codes, aligned as closely as possible to county boundaries. We assigned each zip code to a county based on publicly available maps. Some zip codes crossed county boundaries. For these we chose the county that contained the most zip code area. When necessary, we consulted with zip code databases that listed the primary and secondary counties for the zip codes. Due to the size of Harris County, we divided the county into three regions: inner loop (pinkish red), outer lower loop (red-orange), and outer upper loop (red). All the regions on the map have coupon submissions except the two faint tea green regions on the left most region, where no coupon submissions were made.

Once we determined the regions, we examined how the form submissions were distributed among the regions. Figure 8 shows the total form submissions per region; Figure 9 shows the counts of form submissions that were converted to a purchase. Overall, 13% percent of customers who submitted the coupon form actually made a purchase.

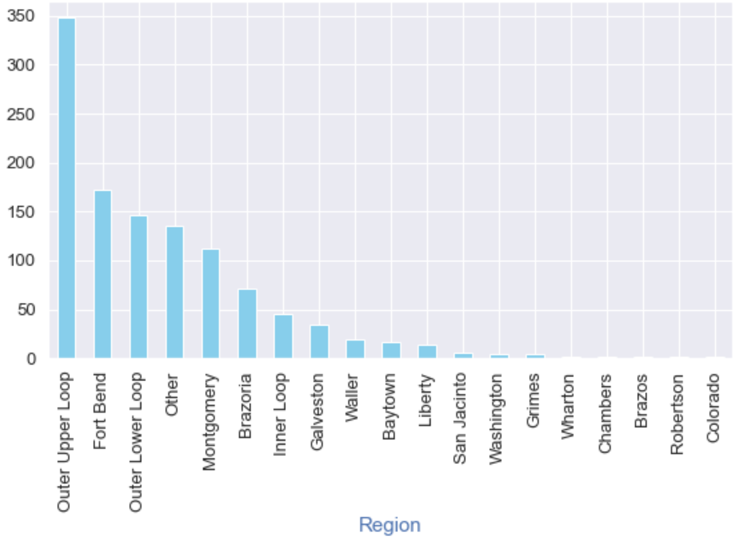

Figure 8: Frequency count of form submissions by region for Greater Houston area.

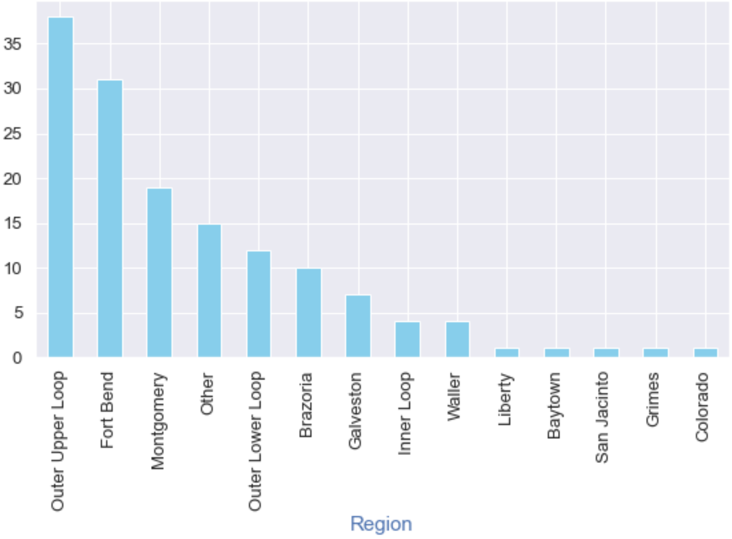

Figure 9: Frequency count of form submissions converted to purchase by region for Greater Houston area.

Recall that Harris County is composed of three subregions (Outer Upper Loop, Outer Lower Loop, Inner Loop). Adding the counts for these subregions in Figure 9 results in 54 total customers. Additionally, from both figures, it is clear that Harris and Fort Bend Counties submitted the majority of coupon forms for this ad campaign. Other counties with a large number of form submissions and purchases are Montgomery and Brazoria.

The large count of customers from the Harris, Fort Bend, and Montgomery Counties can be understood by the store locations. There are two stores in Harris County and one in Fort Bend County. The North Freeway store, the original and largest Gallery Furniture location, is in the Outer Upper Loop region of Harris County. Montgomery County is adjacent to Harris County on the north, so this location in the north portion of Harris County is likely to be convenient for residents of Montgomery County.

Figure 10 displays the ratios of the count of customers who submitted the form and purchased to the count of all form submissions. In other words, the y-axis represents the percentage of form submissions converted to purchases.

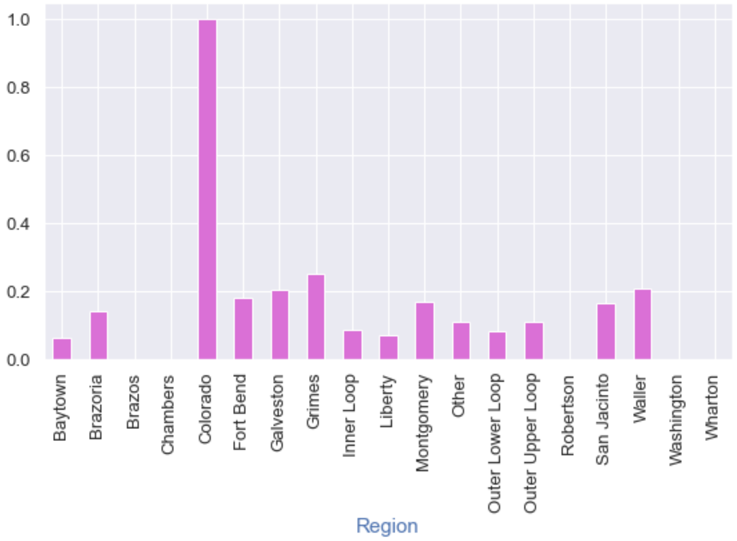

Figure 10: Ratio of count of customers who submitted the form and purchased to count of all form submissions, by region in the Greater Houston area.

You may notice the outlier values, which either have a value of 1 or 0, with a value of 1 corresponding to 100% and 0 corresponding to 0%. These outliers are Colorado, Washington, Robertson, Brazos, and Chambers Counties. The sample size for each of the outliers is substantially low compared to other counties (e.g. Harris). For example, Colorado County only has 1 person who submitted the form; that person purchased furniture resulting in a ratio of 100%. For regions with relatively low frequency counts, percentages should be taken with caution.

Next, we looked at the zip codes with the highest counts of form submissions. The bar graph in Figure 11 includes all customers who submitted the coupon form in the 10 zip codes with the most form submissions.

Form Sumissions and Sales for Zip Codes

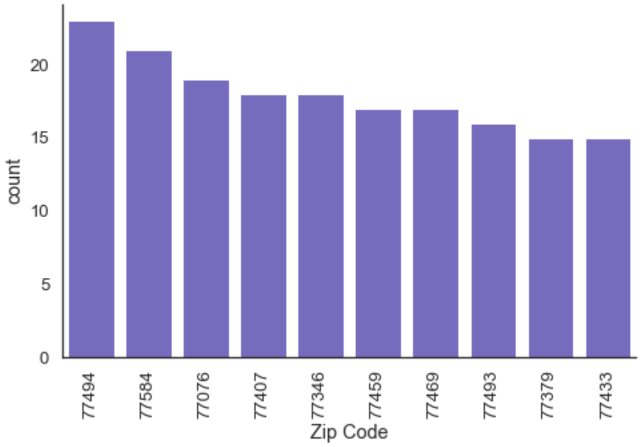

Figure 11: Zip codes with the highest number of form submissions

Interestingly, the North Freeway store location is in 77076, which has the third highest count of form submissions. Also the Grand Parkway store location is in 77407, the fourth highest zip code for form submissions. Lastly, the Post Oak location is in 77056. This is not included in this figure, as it is not among the top 10 zip codes. All the zip codes in this figure are in Harris, Brazoria, or Fort Bend County.

Next, we looked at the zip codes with the highest count of purchasing customers. Figure 12 includes the 10 zip codes with the highest counts of form submissions converted to purchase. There were only 145 people who submitted the Home For the Holidays form and also purchased furniture. Therefore the frequency count for any individual zip code is expected to be small.

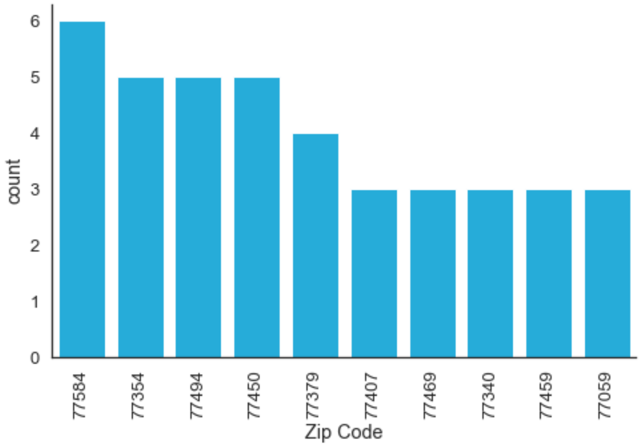

Figure 12: Zip codes with the highest frequency count for purchases

The top 5 zip codes for purchases were all in in Brazoria, Montgomery, Ft. Bend, and Harris Counties. While the zip codes for the North Freeway and Grand Parkway Store locations were the third and fourth highest for submissions, this was not the case for purchases. The zip code of the

Grand Parkway store location, 77407, was ranked 6th for number of forms converted to purchase. The zip codes of the North Freeway and Post Oak stores were not in the top 10.

Figure 13 includes the 10 zip codes with the highest total sales amount. As you can see, sales for the top zip code are over twice the amount for the second highest zip code. This zip code, 77584, is in Brazoria County. Zip code 77566 is also in Brazoria, in 6th place with sales of nearly $20,000. In second place is zip code 77354, with a sales total of just over $20,000. This zip code is in Montgomery County, along with 77356, which was 8th in sales. Three of the zip codes in this list (77494, 77450, 77407) are located in Fort Bend County with a combined sales amount of nearly $60,000. Zip code 77650, in 7th place, is in Galveston County. Zip codes 77433 and 77370 are in Harris County, in 4th and 9th place, respectively.

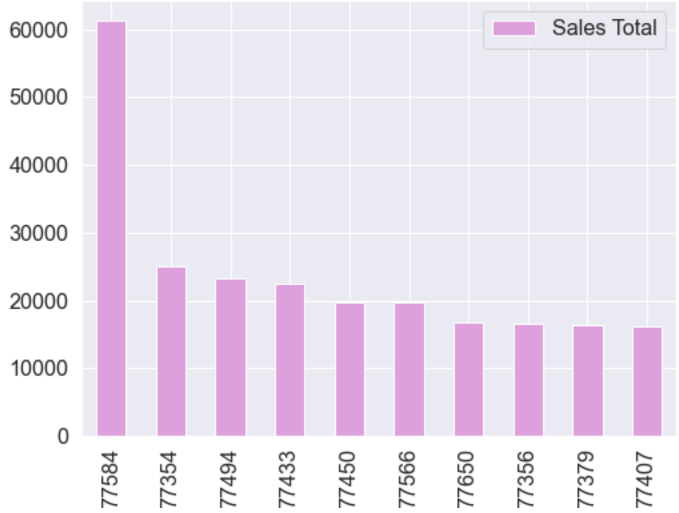

Figure 13: Top 10 purchasing zip codes

Distance from Gallery Furniture vs. Sales Total

Considering more geographic features, we examined the relationship between sales amount and distance from the store. An initial scatter plot of distance and sales total for the entire data set showed that most purchasers were concentrated within 150 km away from the Gallery Furniture North Freeway location. For that reason, we restricted the plot to only those customers within 150 km of the store.

On the scatter plot in Figure 14, the x-axis shows the distance from the zip code in each form submission to the Gallery Furniture North Freeway location. We see from the graph that the people who had highest sales total amount were living approximately 50 km (31 miles) away from Gallery Furniture North Freeway location. The majority of purchasers had a sales total below $5,000 and were less than 75 km from the N. Freeway store.

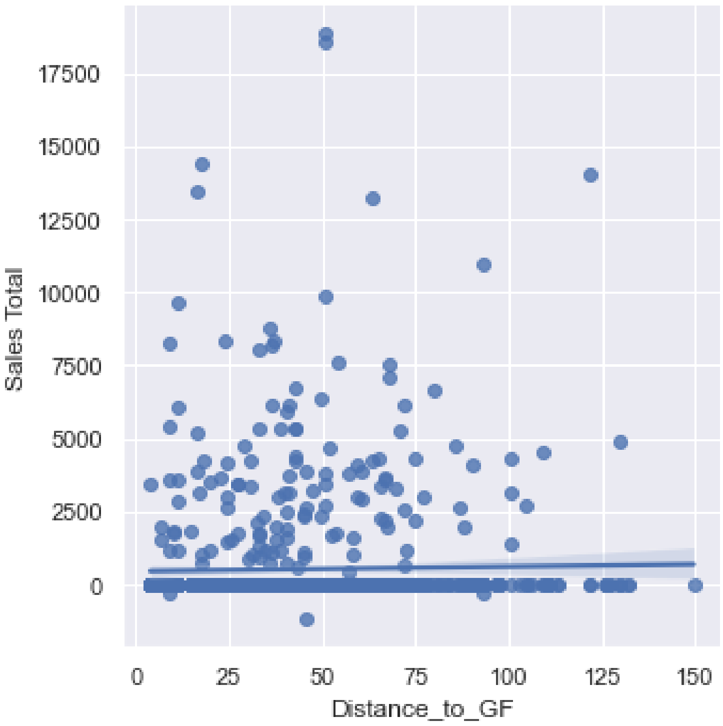

Figure 14: Scatter plot of sales total and distance (in km) to Gallery Furniture N. Freeway location. (Distances greater than 150 km are excluded.)

Relationship of Zip Code Features to Purchase Status

Next we examined the characteristics of the zip codes to see if they were related to whether people are likely to purchase. We used the USZipcode Python package to obtain the geographic and demographic data for the zip codes in the data set. Variables from USZipcode we used were latitude, longitude, median household income, and median home value. Using the latitude and longitude, we also calculated the distance from the zip code to Gallery Furniture.

We wanted to use a supervised machine learning algorithm to see if these variables could be used to predict whether a person submitting the coupon form actually purchases furniture. In addition to the zip code characteristics, we also added the utm_source variable as one of the predictor variables.

Correlations of Variables

When implementing a prediction model, the results are better if variables are not highly correlated. The Pearson correlation plot in Figure 15 shows that there are not too many features strongly correlated with one another.

This is good from a point of view of feeding these features into the algorithm because this means that there isn't much redundant or superfluous data in our training set. Therefore, each feature carries unique information. All correlations were below 0.5 in magnitude except Median Home Value and Median Household Income, which had a pairwise correlation of 0.69.

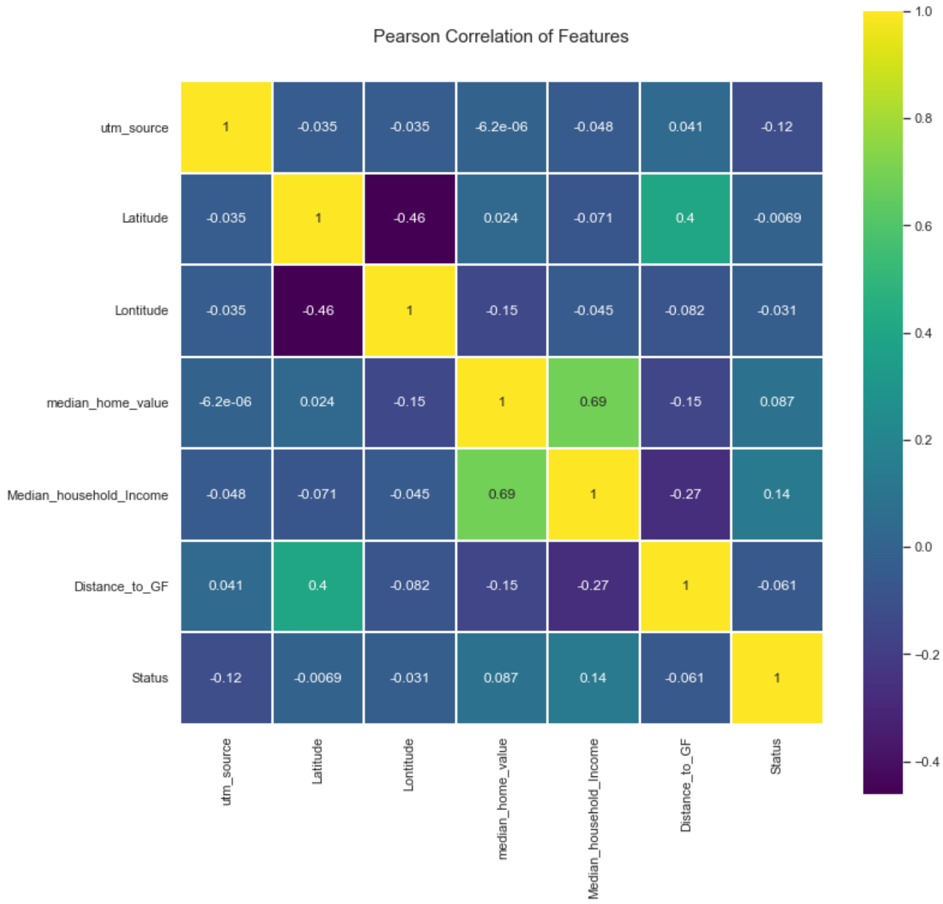

Figure 15: Pearson correlations for variables used in model.

For our predictive model, the outcome variable was Status, which represented whether a person submitting the form purchased furniture. A Status of 1 indicates furniture was purchased; a Status of 0 indicates the person submitting the form did not purchase. When determining which features to use in the predictive model, we chose only those for which the magnitude of correlation with the outcome variable was at least 0.1.

Only two feature variables met the criteria, UTM Source and Median Household Income. The correlations are shown in Table 1. UTM Source was a categorical variable with labels as values (e.g., website, facebook). We recoded this as a quantitative variable.

Variable	Correlation with Status
Utm Source	-0.119111
Median Household Income	0.137330

Table 1: Correlations for predictor variables variables used in our algorithm model.

Application of the Decision Tree Classifier Algorithm

We chose to use a decision tree model. We applied the decision tree algorithm from the Python Sci-kit Learn package. We split our data set into a training set and a test set, with 70% of the data points used for training the model and 30% used to test how well the model predicts the outcome. The split set was carried out by using an algorithm that used randomness to choose a test set that was representative.

The underlying logic of the decision tree is trying to find the right questions to ask based on the predictive variables so that the tree leads to good prediction of the outcome. The decision tree generated by the algorithm for this data had three layers, as seen in Figure 16. Each node represents a decision point. The first layer was based on a cutoff value of $60,805.293 for Median Household Income. The second layer involved values for UTM Source. The third value involved both UTM Source and Median Household income, with a cutoff of $38,499.988 for one of the nodes.

The "gini" value serves as a measure of the value of information coming from that node. Higher values (closer to 1) indicate the node was useful for prediction; lower values (close to 0) indicate the node was not very useful for prediction.

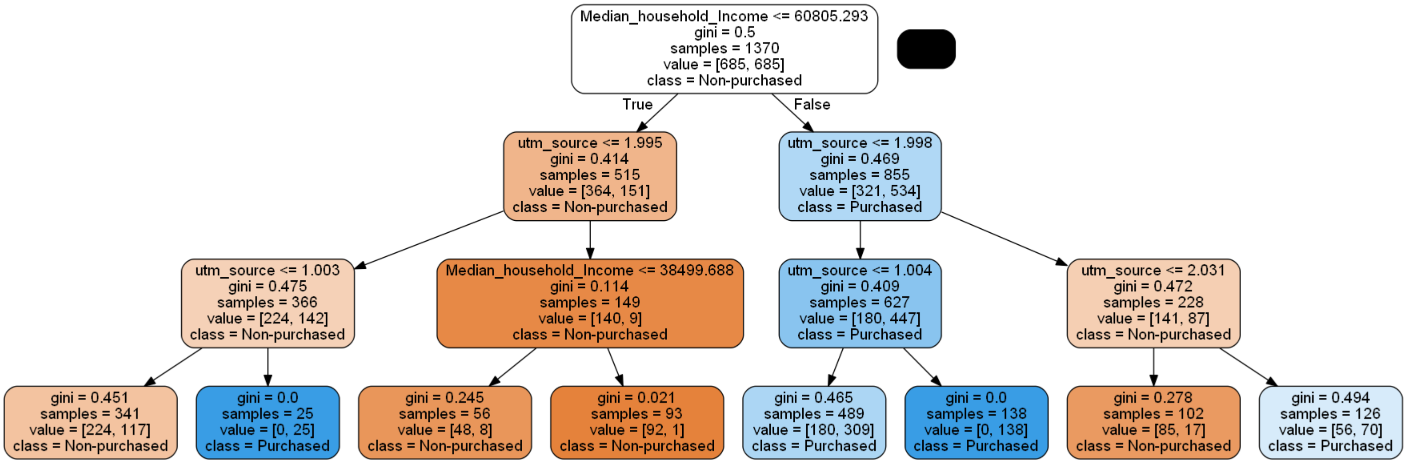

Figure 16: Decision tree graph resulting from using SMOTE for handling imbalanced data.

Predictions Generated by the Decision Tree Model

After the decision tree model was generated, we applied it to our test data set, which included 338 data points (approximately 30% of the data points). The predictions resulting from our first iteration of the decision tree model are shown in Table 2. There were a total of 41 people who actually purchased, or 12.13% of the 338 people. However, the model only predicted a purchase for 12 of these. The model predicted nearly everyone to be a non-purchaser. We can see the data set is imbalanced, with the number of purchasers being far less than the number of non-purchasers.

	Predicted Non-Purchase (Status=0)	Predicted Purchase (Status = 1)
Actual Non-Purchase (Status = 0)	287	10
Actual Purchase (Status = 1)	39	2

Table 2: Actual and predicted values from applying initial decision tree model (without SMOTE) to test data.

Due to the imbalance in the data set, we implemented the decision tree again, using synthetic minority oversampling technique (SMOTE), an algorithm for handling imbalanced data. In the training phase, the SMOTE algorithm samples an equal number of people who purchased and people who did not purchase, reducing the effect of the imbalance on the result.

The final decision tree shown in Figure 16 is from the iteration using the SMOTE technique. The actual and predicted values of the outcome variable are shown in Table 3. Compared to the initial version, more people are predicted to purchase.

	Predicted Non-Purchase (Status=0)	Predicted Purchase (Status = 1)
Actual Non-Purchasers (Status = 0)	236	61
Actual Purchasers (Status = 1)	33	8

Table 3: Actual and predicted values from applying final decision tree model (with SMOTE) to test data.

The results from the original decision tree (not using SMOTE, Table 2) produce a prediction accuracy of 85.5%. Of the 338 data points in the test set, 289 were correctly predicted (2 purchasers and 287 non-purchasers). The results from the final decision tree (using SMOTE, Table 3) produce a prediction accuracy of 72.2%, with 244 of the 338 data points correctly predicted (8 purchasers and 236 non-purchasers). If we had simply predicted every single person to be a non-purchaser, we would have an accuracy of 88.9%. So basically neither of these models are very good for prediction, but the lower accuracy in the SMOTE version is a better representation of how the model is actually doing. The 85.5% accuracy from he first version is inflated due to the imbalance.

Recommendations

We suggest that Gallery Furniture modify their voucher forms for future sales campaigns in order to eliminate problems we experienced during the data cleaning process. Two major problems our team faced were duplicates and unusable data through invalid entries in the voucher form. The data set contained a large number of duplicates, mostly due to multiple form submissions by the same person. Some zip code entries contained information that did not follow the standard 5-digit format (with or without the standard 4-digit extension), which made the data invalid for geographical analysis.

Because of these issues, our team recommends that Gallery Furniture should use input validation for future voucher forms. For example, if a customer inputs an invalid zip code, the form should prompt them to fix it. Also, if the customer received an immediate popup confirming that their form was submitted successfully, then most likely they would not resubmit it. These improvements would reduce data troubles and provide higher quality data for future analysis.

Additional information such as age, gender, and store location could allow for deeper and a more useful analysis of the effectiveness of future sales campaigns. However, our team recognizes that customer privacy is of utmost concern, so we suggest that these questions be optional for future voucher forms.

Our team also believes collecting data after customer purchases could improve the effectiveness for future sales campaigns. We recommend that Gallery Furniture develops a follow up form for purchases through the sales campaigns to explore views and criticisms from nonpurchasers and customer interest in competing furniture stores.

If Gallery Furniture is not already doing so, an additional customer satisfaction form could help track customer experience and service for each campaign. Also, website tracking data such as the amount of clicks on products and pages could help analyze the relationship between customer interest and purchases for both purchasers and nonpurchasers. Additionally, targeting specific geographical areas through social media could help Gallery Furniture reach more customers in zip codes with lower rates of form submissions and sales, such as parts of Harris County.

Discussion

Our team's primary objective was to analyze the effectiveness of the Home For The Holidays (HFTH) sales campaign. Our main focus was on the geographic distribution of sales and coupon form submissions across the Greater Houston area. When we divided the Greater Houston area into regions mostly based on counties, a few counties stood out.

Fort Bend, Montgomery, and Brazoria Counties contributed sales of $115,014, $101,400, and $87,700, respectively. Based on the populations of these counties [1], this meant they had sales of $0.14, $0.16, and $0.24, per resident, respectively. One zip code in Brazoria County, 77584, generated nearly $60,000 in sales. The coupon to purchase conversion ratios for these counties were 18%, 17%, and 14%, all larger than the overall ratio of 13% for the whole dataset. Galveston County was also noteworthy with $37,002 in total sales and a coupon to purchase conversion ratio of 21% and a $0.11 per resident based on population.

While Harris County had the highest total sales at $191,200, it underperformed the previously mentioned counties when population is considered. Harris County has a population of 4,731,135, more than twice the population of Brazoria, Fort Bend, Montgomery, and Galveston combined. Harris County's sales per resident was only $0.04 and the coupon to purchase conversion ratio was only 10%.

Overall, our analysis showed that the Gallery Furniture HFTH was more successful in several counties surrounding Houston (Fort Bend, Montgomery, Brazoria, Galveston) than in Harris County, the primary county for Houston. Some counties near Houston had only a small number of form submissions, so the coupon to purchase conversion rate and the sales per resident are not very meaningful. However, Harris, Fort Bend, and Montgomery Counties had the highest form submission counts of 539, 172, and 112 respectively. With these large sample sizes and the higher coupon to purchase conversion ratios and sales per resident ratios for Fort Bend and Montgomery compared to Harris County, we are confident in asserting that Fort Bend and Montgomery County were especially effective contributors for the HFTH sales campaign.

The higher coupon to purchase conversion ratios of suburban counties (Fort Bend, Montgomery, Brazoria, Galveston) compared to Harris County are consistent with the results of our decision tree analysis of the zip codes. When examining the zip code characteristics, we observed that the median home value and median household income for the zip codes of purchasers were slightly higher than for people who submitted a coupon form but did not make a purchase. However, it was not a large difference and it was not enough of a difference to help the decision tree model accurately predict purchase status based on zip code characteristics. This indicates that Gallery Furniture is doing a good job of converting form purchases to sales, even in less affluent zip codes.

Acknowledgements

The project would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We highly appreciate Gallery Furniture, especially Edward Raupp, Data/Attributions Director, and James McIngvale, Digital Director, for providing a great opportunity to explore data analysis in a real world professional setting.

We would also like to show our deepest gratitude to the Lone Star College-North Harris administration and the Honors College for offering this class and empowering students to learn and grow through industry experience. We would also like to thank Professor Jennifer Travis for her constant guidance and supervision to get the most out of this project.

We give our special thanks to PIC Math for helping to get this class started at the Lone Star College-North Harris campus. PIC Math is a program of the Mathematical Association of America (MAA) and the Society for Industrial and Applied Mathematics (SIAM). Support is provided by the National Science Foundation (NSF grant DMS-1722275).

References

[1] The County Information Project, Texas Association of Counties.
`https://txcip.org/tac/census/morecountyinfo.php?MORE=1086`