**Customer Analytics Final Report**

**Data Set: Cookie-level web data for an online store**

**Team 12**

Rachana Chetan, Wei Huang, Junoh Jung, Hezekiah Shanu, Chelsea Zheng

**Executive Summary**

IDK, Music Store? is an online music downloading website that is interested in maximizing its revenue using web cookie data to target marketing efforts towards specific market segments.  
  
In our report we first outline the data examination and cleaning of our dataset. We then discuss the best modelling approach we adopted to answer our business problem, ie., K-Means Clustering. We then move on to discussing the limitations and business recommendations based on our analysis.

**Business Problem**

Our dataset does not have a prescribed company or industry, so we assumed the data source is from an online music store. We made this assumption because the website has a widespread international customer base with sales as cheap as a penny in some instances, far too cheap a price to account for shipping. So, we assumed this store sells digital downloads of songs and albums. The business problem for this music store is to determine where to invest their marketing efforts. They want to invest in marketing techniques that generate the highest revenue. So, the goal of this analysis is to evaluate the different site visitors and determine which source URL, city, and customer segment produced the highest revenue, and invest our marketing funds accordingly.

**Methodology**

**Data Cleaning**

1. **Removing Excess Variables**

Our dataset originally contained 20 variables but for an analysis we only included 12 of the original variables. Here are the following steps we took to clean the dataset:

1. We excluded variables that had a uniform value (I.e. all values in the field were equal to a 0 or 1.).
2. We excluded variables that had a dependent relationship to others because these relationships inhibited our regression analysis.
3. We removed customers whose numbers of site entrances and exits were not equal.
4. **Excluding Heterogeneous Variables**

The region, city and date variables all had unique values for each respondent, making them unsuitable for regression analysis. Our team pursued the possibility of creating categories for the regions but were discouraged from doing so because of the large amount of city and region names that were misspelled. This made it difficult to determine what country many of the cities came from.

Although we excluded these variables from our analysis, we included charts depicting the highest grossing regions and a time series analysis of sales in our appendix.

1. **Categorizing Variables**

To prepare the dataset for regression analysis we categorized string and heterogeneous variables into numerical variables.

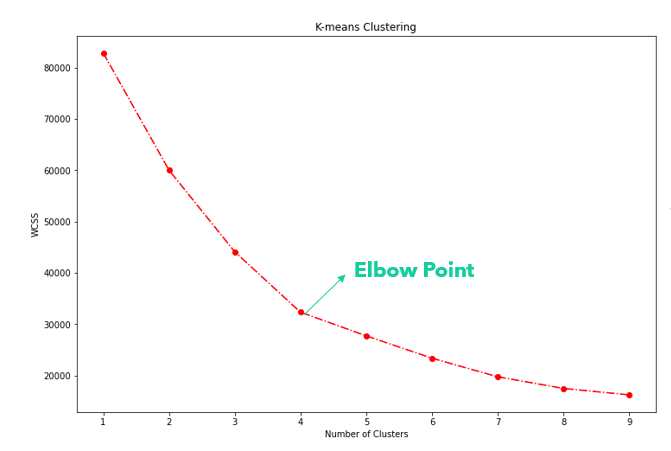
1. Medium:
   1. The medium was either organic, (none) or referral. So, we changed these variables to 1, 2 and 3, respectively.
2. Source:
   1. There were 321 source URLs.
   2. To categorize these, we counted the number of site visits from each URL and grouped them from there. URL sources with more than 400 distinct visits were given their own categories, while the remainder were grouped into “other search engines” or “other websites.”

**Model Decision**

**K-Means Clustering**

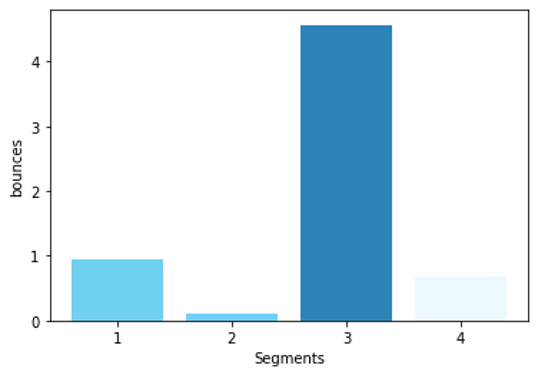
To solve the problems of segmenting the market, several methods are performed, including K-means clustering, decision tree, and KNN. The result of both decision tree analysis and KNN are very unideal, likely to be reflected since there is limited information about the dataset, and the dataset has too many no transactions, compared to transactions, making it not only difficult to make assumptions about how to categorize the customers, but also hard to train the model to perform accurate analysis. As a result, K-means clustering, an unsupervised learning method is chosen for segmenting the market. Due to the nature of K-means clustering, in which only Euclidean distances amongst variables are considered, our model only includes continuous variables when segmenting the market. In the end, we chose to include bounces, numbers of pages viewed, time spent, transaction revenue, and number of visits in the K-means clustering analysis.

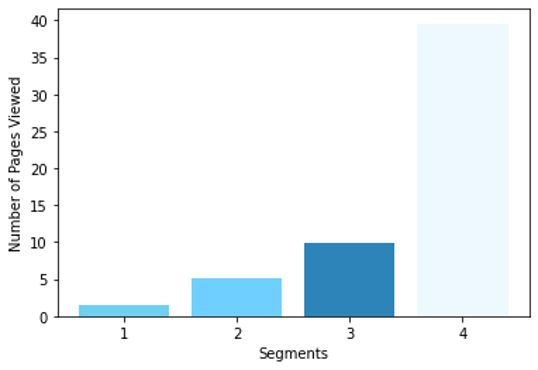
The procedure of K-means clustering involves trying different K, or number of segments, and calculating the within cluster sum of squares. A low within cluster sum of squares indicates that the data in the cluster, are more closely related to each other. To find determine the best K, or number of segments, we are interested in finding the K that not only has a relatively low within cluster sum of squares, but also leads to a big decrease in within cluster sum of squares from K-1 to K. This is because that, as we increase K, the within cluster sum of square would always decrease, but we lose our purpose of segmentation if choose a big number of K, that is almost as big as the amount of data in the dataset. Therefore, an elbow point method is used to find the K. An elbow point is the point where the within cluster sum of squares stops from decreasing quickly. As indicated in the figure below, the optimal number of segments given the current dataset is four.

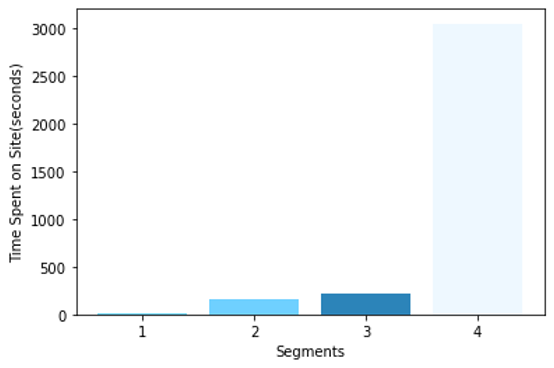


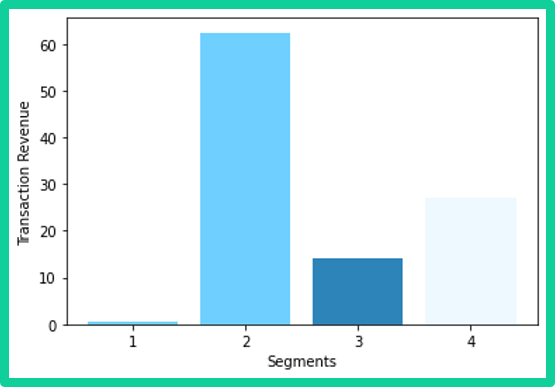
**Finding the Segment of Interest**

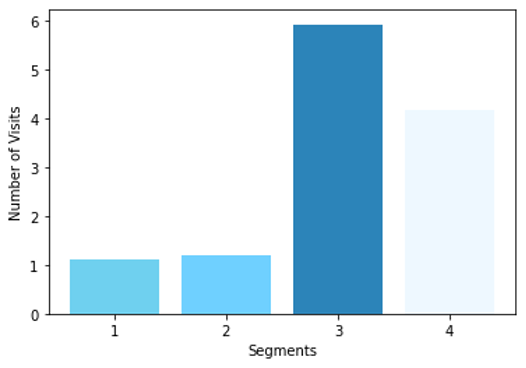
After performing the segmentation to all the data, it would be useful to see which segment is segment of interest for the online store, so more attention can be paid to these specific customers. As a result, we plot out the means of all the continuous variables, based on the segments into bar graphs, and compare the results. As demonstrated from the bar graphs, segment 2 has the highest transaction revenue, while segment 4 has the second highest transaction revenue, but the highest number of visits. As a result, we determine that segment 2 is the segment of interest, as it not only has the highest transaction revenue, but also the lowest bounces, and relatively high number of visits and pages viewed.





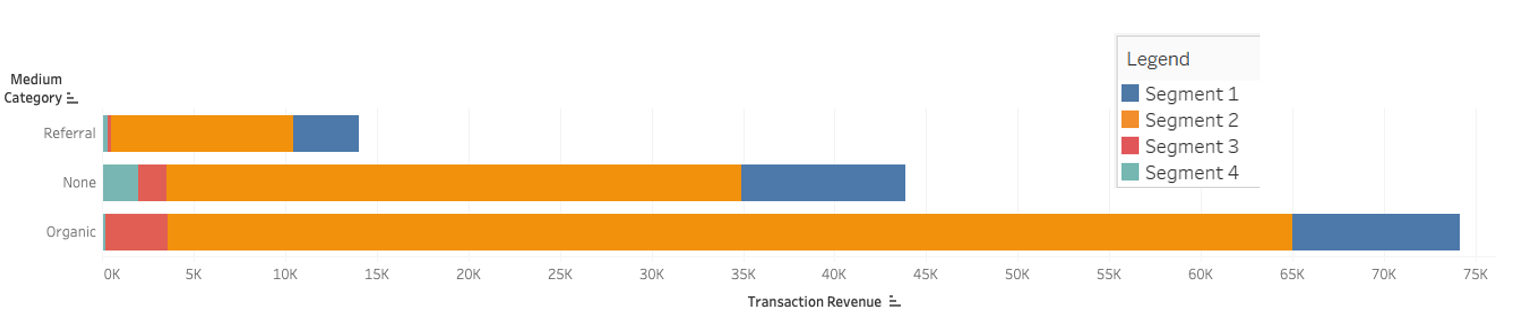






**Data Summary/Visualization**

**Transaction Revenue by Medium**

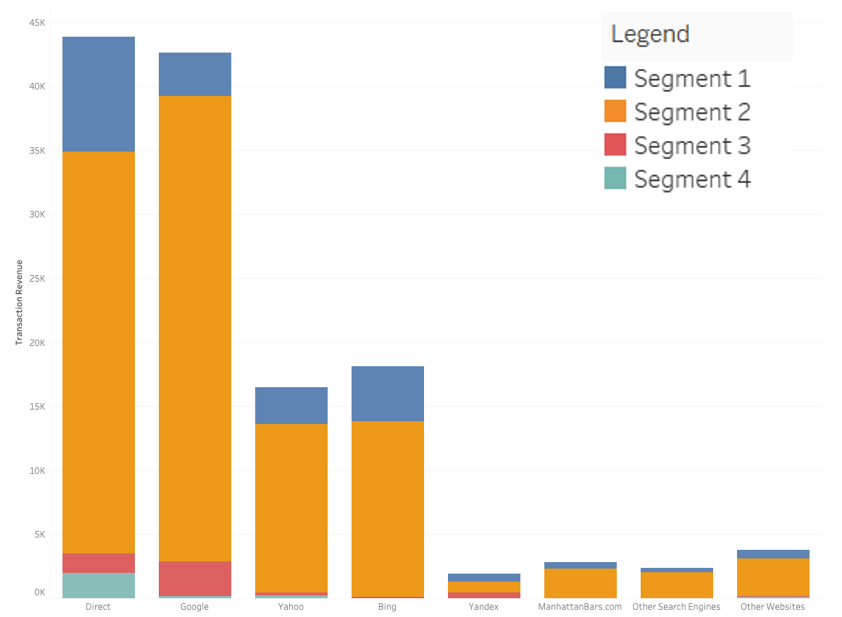


The bar graph above illustrates transaction revenue by medium, separated into four segments. None indicates direct URL addresses and bookmark links. Organic search signifies searches through a search engine such as Google or Yahoo. Referral indicates access through a web sponsor link or by clicking on a banner advertisement.

Customers who entered the website through organic search, aggregating all search engines, were the most profitable. Examining the medium category, it appears that segment 2, labeled as orange, generates the highest revenue across all categories.

**Transaction Revenue by Source Category**

The below bar chart represents transaction revenue by source category, including direct URL, Google, Yahoo, Bing, Yandex, ManhattanBars.com, other search engines, and other websites. The source category is divided into four segments.



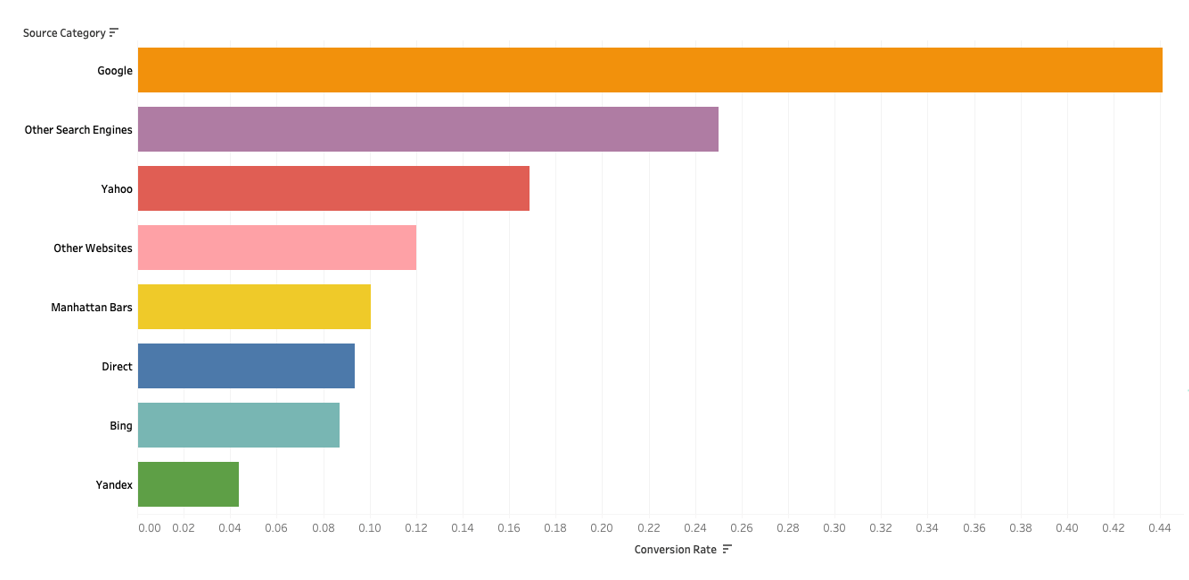
Customers who entered the website through direct URL and Google proved to be the most profitable. In source category segment 2, labeled as orange, the highest transaction revenue is generated in each category.

Conversion Rate

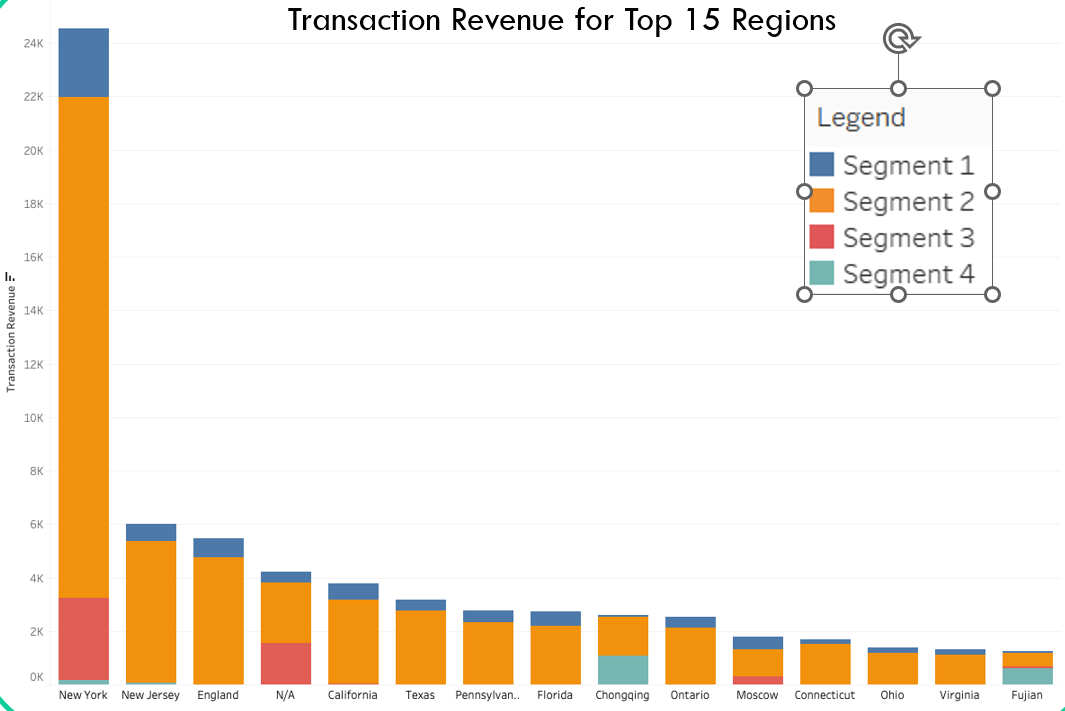
The conversion rate, which refers to the rate of users who end up making a transaction on the website is given by:

Conversion Rate = # of Transactions/ entrance

The graph below shows which websites have the highest conversion rates for our online store. Customers who entered from Google search results have the highest conversion rate at 44%.



**Transaction Revenue by Region**



The above bar chart represents transaction revenue by region for the top 15 regions, including New York, New Jersey, England, California and some more. The source category is divided into four segments.

Customers located in New York are the most profitable. Major cities in North America, and London, generated the most revenue. We can also infer from this graph that China (Chongquing, Fujian) is a market of interest. In segment 2, labeled as orange, the highest transaction revenue is generated in each category.

**Business Recommendations**

1. Given that approximately 98% of the customers spend less than 20 minutes on the website, there's an opportunity to enhance the overall user experience (UI/UX). A more attractive and intuitive design not only can better capture customers' attention but can also encourage them to explore the site further, discover more products or services, and potentially make additional purchases. After all, A positive and seamless user experience has the potential to increase user retention, satisfaction, and overall loyalty.
2. New York and New Jersey are identified as the locations with the most profitable customer base. We should further investigate the demographic information of the customers in these regions and identify what are the key factors leading to high conversion rates and transaction revenue. Next, based on the findings, the company can spend their marketing budget more efficiently on potential high-value customers.
3. We found that customers who find the website through Google searches have generated higher revenue. By allocating more resources to Google paid search, the company would be able to position themselves to attract and engage with a customer segment that has a higher likelihood of contributing more transaction revenue.
4. By understanding the demographics, behaviors, and preferences of the most profitable customers, we can tailor promotional links and referral programs to resonate more effectively with the target market. Therefore, a data-driven approach is critical in this revision process. The company should regularly analyze the performance metrics of each promotional link and referral programs, such as click-through rates, conversion rates, and customer lifetime value. It would allow them to continue optimize the sponsorship strategies to increase return of investment.

**Limitations**

1. Excluding categorical variables from the customer clustering analysis leads to limitation in our methodology since it restricts our scope of understanding the dataset and may have resulted in an incomplete representation of the customer segmentation. Categorical variables are extremely important because they represent qualitative attributes such as customer demographics, preferences, or geographic locations. These aspects play a crucial role in providing a holistic view of customer behavior and characteristics. In addition, we did not exclude outliers in our customer clustering analysis while outliers can impact measures of central tendency (mean, median), variability (standard deviation), and relationships between variables.
2. Our dataset only includes limited background data. We would be able to make more insightful interpretations about each segment if we had more information such as customer IDs, other items in the cart, etc. First, customer IDs are essential for tracking and differentiating each unique customer over time. It allows us to create personalized profiles and accurately analyze customer behaviors, preferences, and engagement patterns. Second, knowing what specific products are in customers’ carts provides valuable context for segmentation, which enables more targeted marketing strategies and personalized recommendations. Lastly, it would be very helpful if we knew what product the store is actually selling so that we could make explicit business recommendations.

**Conclusion**

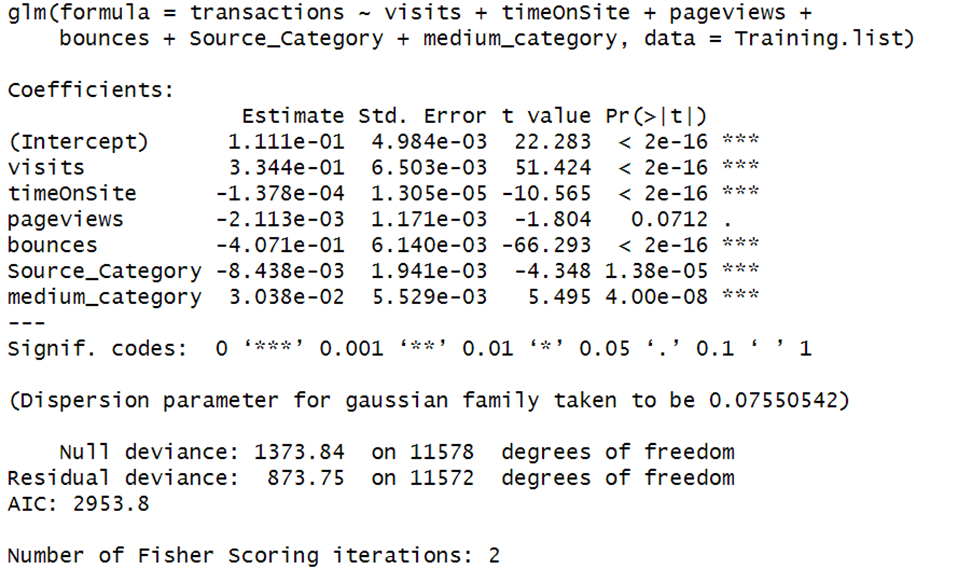
In conclusion, future analysis could be made after adding more comprehensive information about our customers, including customer IDs and other demographic details to understand customer behavior, preferences, and engagement patterns better. We suggest collecting more comprehensive customer information, use them to improve segmentation methods or explore new analytical approaches, consider additional items in the shopping cart for efficient recommendations, and re-evaluate transaction revenue predictions through new regression analysis. These steps together would lead to a more sophisticated solution to the business problems.

**Appendix A – Variable Definition**

|  |  |
| --- | --- |
| latitude | Latitude of the location the request was made from |
| longitude | Longitude of the location the request was made from |
| medium | organic = organic search; referral: web link, e.g., clicking on a banner ad; (none) |
| medium\_source | 0= none, 1 = organic, 2 = referral |
| region | Region the request was made from |
| source | Referring website. Direct is either type-in of URL or bookmarked URL; all other give the referring website, e.g., Google |
| Categorized\_Source | 0 = direct, 1 = google, 2 = yahoo, 3 = rambler, 4 = bing, 5 = yandex, 6 = manhattanbars, 7 = Other Search Websites, 8 = Other Websites |
| Entry Bounce Rate | % Visitors who leave immediately after entering the site. |
| Transaction Completion Rate | 0 for all |
| PVs Per Visit | Page views per visit |
| Goal Completion Rate | 0 for all |
| bounces | # bounces |
| entrances | # entrances |
| exits | # exits |
| goalCompletionsAll | 0 for all |
| newVisits | New visitor to the site (based on tracking) |
| Number of Records | 1 for all |
| pageviews | Total # of page views |
| timeOnSite | time on site in seconds |
| transactionRevenue | Revenue |
| transactions | #transactions |
| visits | Total visits to the site (based on tracking) |

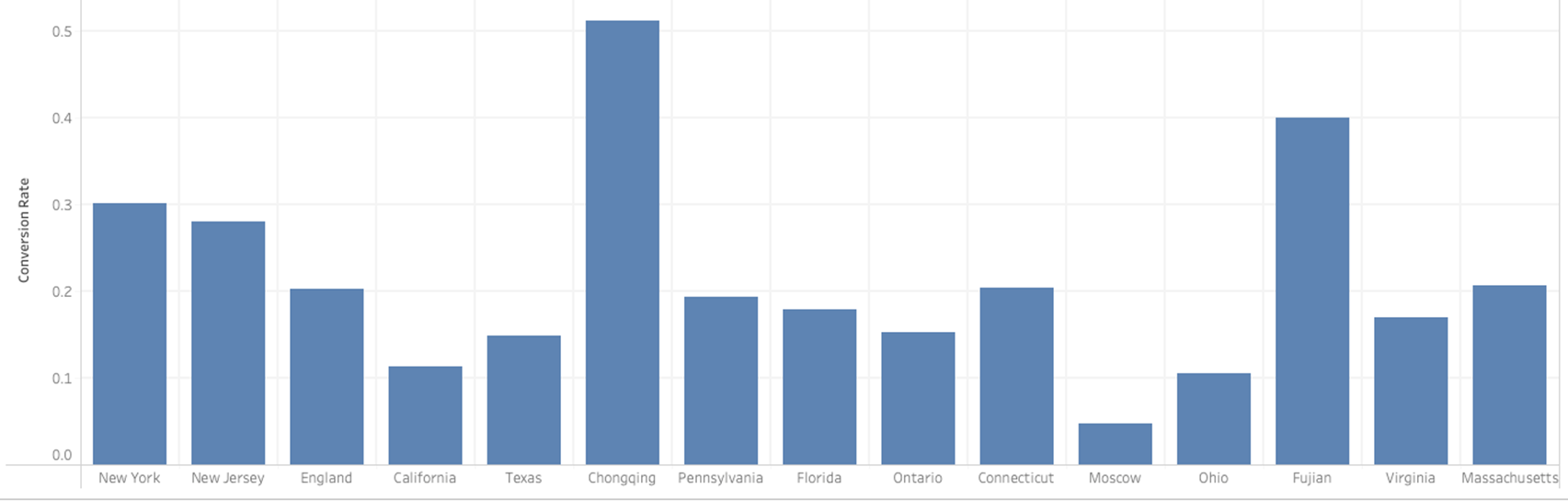
**Appendix B – Regression**

Binary Logit Regression using variables: visits+ timeOnSite+ pageviews+ bounces+ Source\_Category+ medium\_category over number of transactions was run in order to predict how many transactions the online store could expect if they could track the aforementioned variables. Given below is the output of the regression.



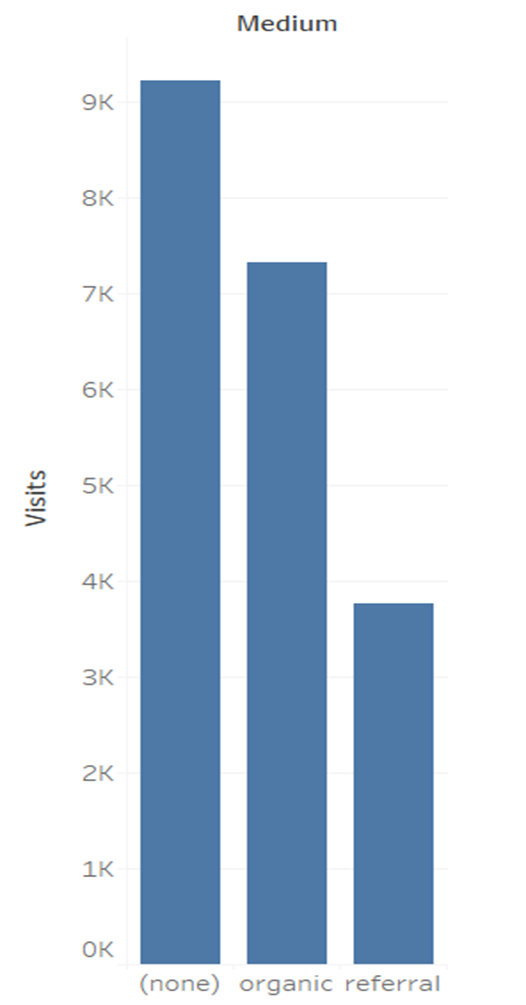
This regression produced a very good result with many highly statistically significant variables but that was only because of the fact that the data was highly skewed. The dataset had just about 85% values that were 0 and 15% 1’s.   
Hence the output was considered unreliable.

**Appendix C – Conversion Rate for Top 15 Grossing Regions**



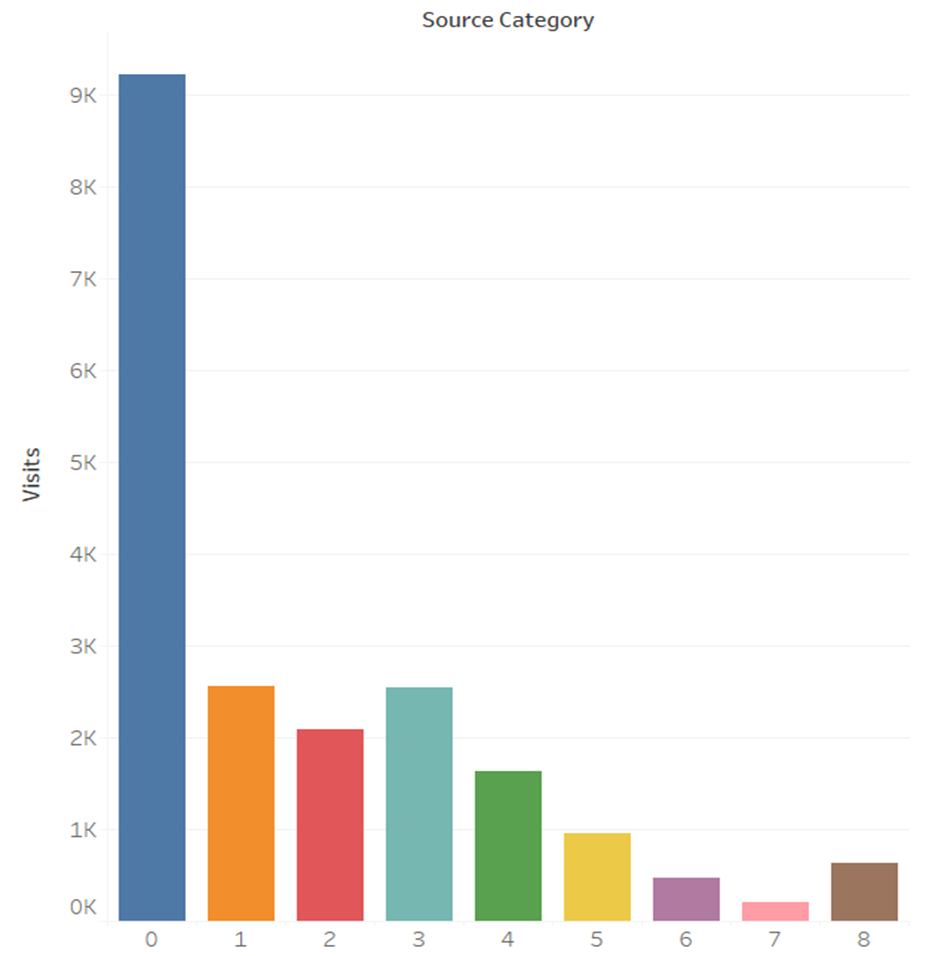
* The top 3 regions with conversion rate are Chongqing (China), Fujian (China), and New York (United States).

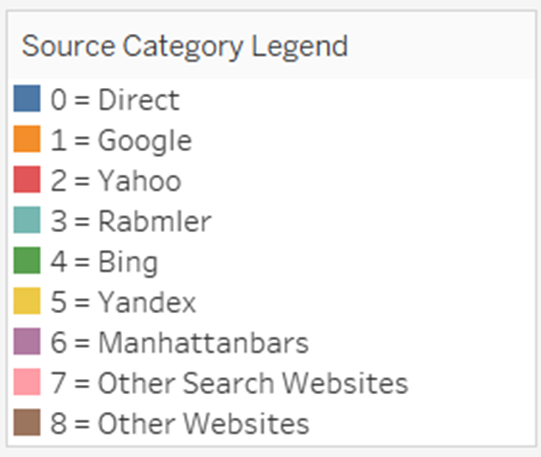
**Appendix D – Distribution of Visits by Medium**



* The graph above gives the total number of visits to the site by medium. The none and organic have most visits within the medium category.

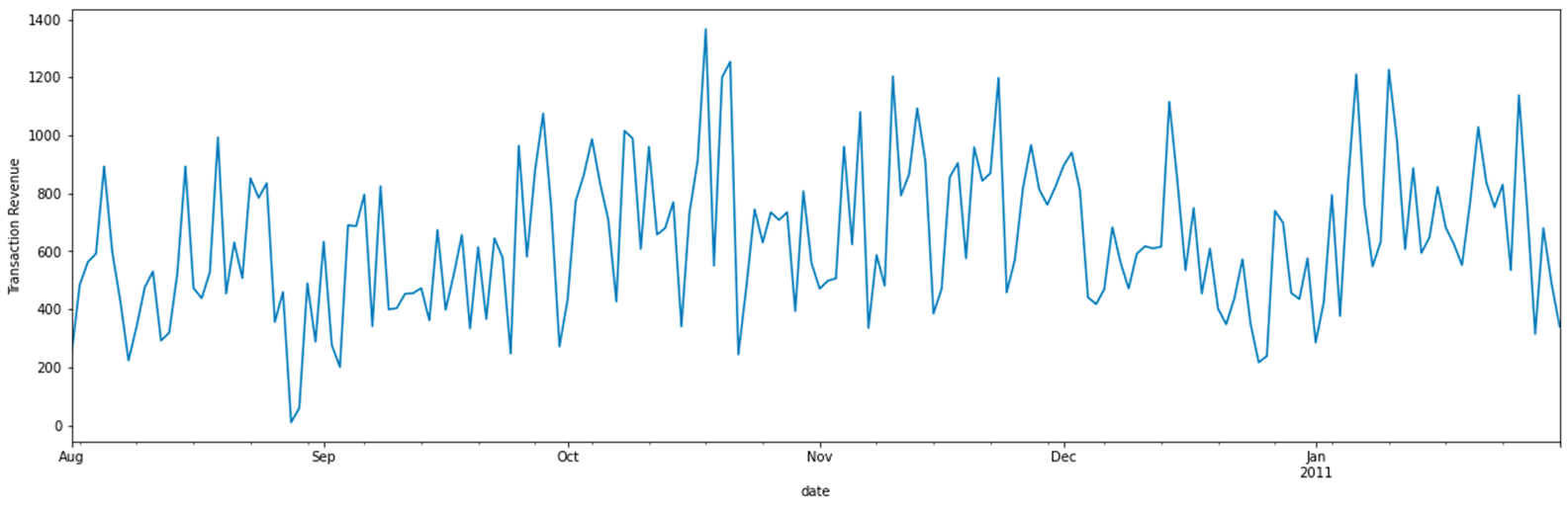
**Appendix E – Visits by Source Category**





* The total number of visits to the site by source category. The direct URL, Google, and Rambler have most visits within the source category.

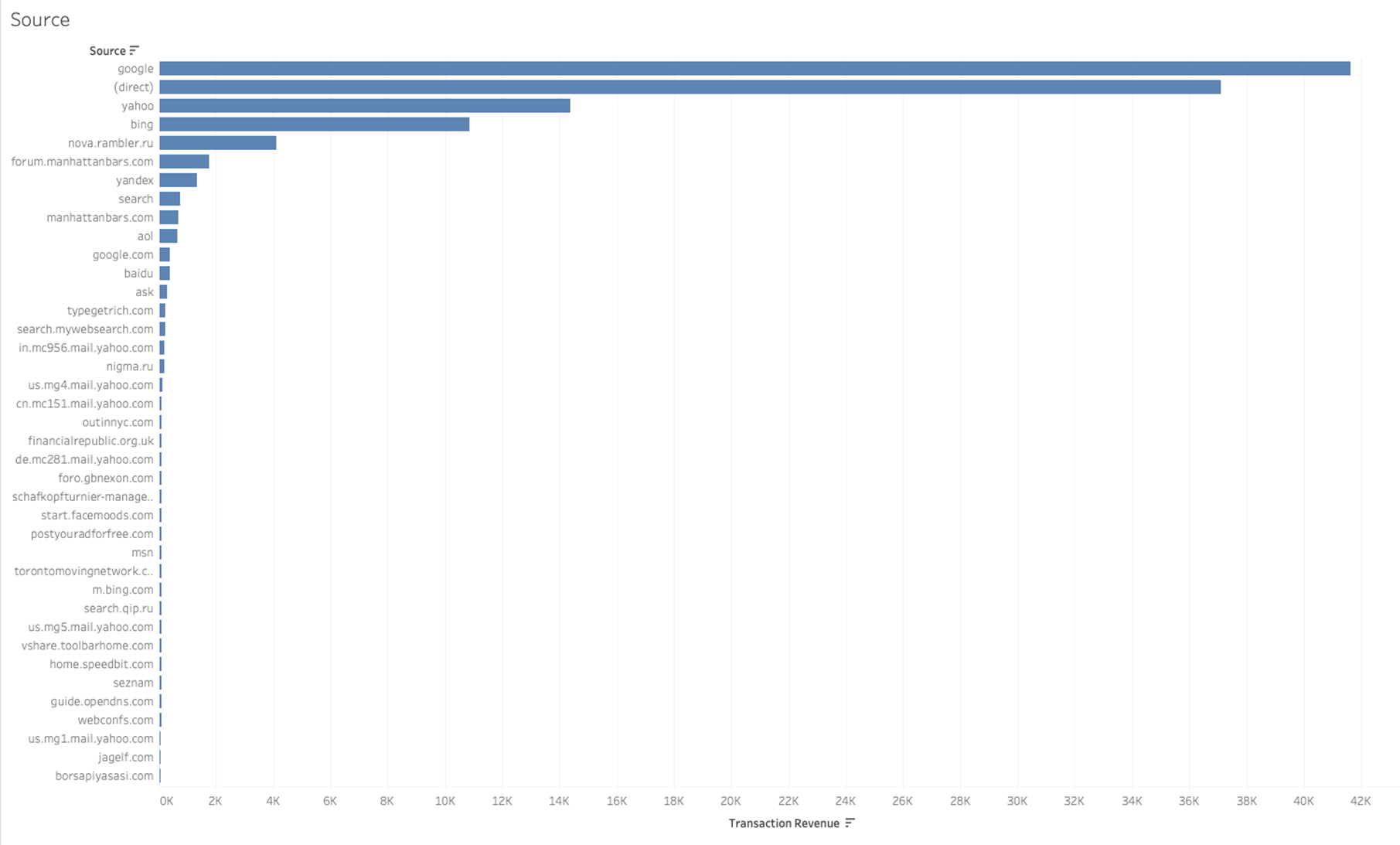
**Appendix F – Time Series Plot of Transaction Revenue**



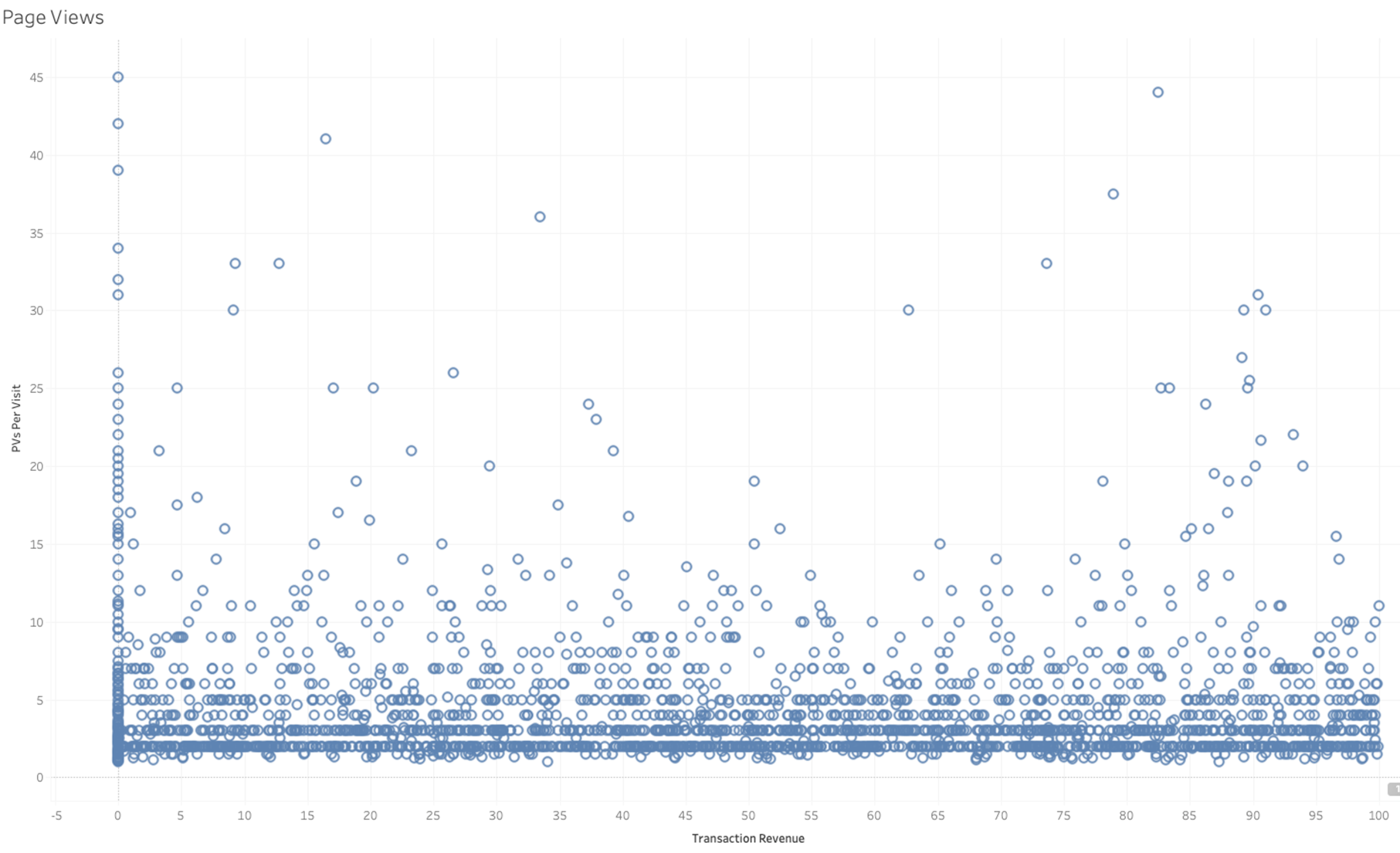
* Transaction Revenue goes up and down throughout the year. The peak happened during the second half of October, and the lowest drop happened during the end of August.

**Appendix G – Transaction Revenue by Source**

* The transaction revenue by source. Google and Direct URL generate the most revenue from the source.



**Appendix H – Page Views per Visit by Transaction Revenue**



**Appendix I – Visits by Transaction Revenue**

