Revenue Forecasting for Google GStore

# Abstract

To understand customer behaviour and generate revenue projection is one of the most fundamental and imperative problem in business analytics. Companies relay on it to make appropriate marketing strategies and forecast key financials. This capstone project will explore and analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset. The goal is to predict the natural log of the revenue per customer visit:

The project is adapted from Kaggle competition “Google Analytics Customer Revenue Prediction” and utilizes the training data set provided for the participants. The data contains customer visits from August 1st, 2016 to August 1st, 2017. It also includes other features related to the visits such as channel, device and geography of the users which may reveal insights to inform actionable operational changes and a better use of marketing budgets.

Revenue will be predicted using Auto Regressive Integrated Moving Average model (ARIMA), linear regression models, K-Nearest Neighbours (KNN) and XGBoost model using RStudio. Models will be evaluated based on the root mean squared error (RMSE).

Data Source: <https://www.kaggle.com/c/ga-customer-revenue-prediction/data>

GitHub Repository: <https://github.com/dn317/CKME136-Capstone>

# Introduction

To understand customer behaviour and generate revenue projection is one of the most fundamental and imperative problem in business analytics. Companies relay on it to make appropriate marketing strategies and forecast key financials. This capstone project will analyze a Google Merchandise Store customer dataset to predict revenue per visit. Revenue will be predicted using Auto Regressive Integrated Moving Average model (ARIMA), linear regression models, K-Nearest Neighbours (KNN) and XGBoost model using RStudio. Models will be evaluated based on the root mean squared error (RMSE).

# Literature Review

Previous works have attempted to approach sales prediction using both traditional and novel techniques. Traditional statistical techniques, such as time-series models, exponential smoothing, regression models are often applied. Newer approaches introduce hybrid models and advanced techniques such as neural networks and support vector machine (SVM).

An example of the application of traditional statistical techniques is the research by Winters (1960). Winters used only the historical sales (time-series) as the input to predict future sales under the exponential forecasting models. This early research sacrificed the information gain from other types of inputs to achieve a model that can make predictions quickly, cheaply, easily and routinely. In a recent study to predict sales in Rossmann Stores, Pavlyshenko (2019) considered different machine-learning approaches for time series forecasting. Random Forest algorithm was applied first and the prediction under this method revealed that when the machine-learning algorithm is applied on non-stationary data we may observe a bias which is a constant (stable) under- or over-valuation of sales and thus requires correction. The study then introduced historical sales values into the data to demonstrates that machine-learning generalization can improve the accuracy of prediction even with a small number of historical data.

Hybrid approaches apply different techniques in different stages of the prediction. In Chen and Lu’s (2017) study, a clustering-based model was proposed for computer retailing sales forecasting. This method applied the clustering technique to divide training data into groups having consistent characteristics before constructing forecasting models for each group. The appropriate model for the test data was then selected by matching the test data to the cluster with similar features. This approach increases the relevance between training data and the data to be forecasted. Lu and Chang (2014) proposed a hybrid sales forecasting scheme for information technology product sales. The scheme combined independent component analysis (ICA) with K-means clustering and support vector regression (SVR). ICA was first used to extract features from the data. K-means algorithm was then applied for clustering the sales data into several groups based on selected featured. Finally, the SVR forecasting models were applied to generate forecasting results for each group. Some hybrid approaches created a more complete framework for forecasting from preprocessing to prediction. Research by Guo and Wong (2013) proposed a multivariate intelligent decision-making (MID) model for retail sales forecasting. The model integrates a data preparation and preprocessing module, a variable selection module and a prediction module to forecast the sales volumes. Besides, the proposed MID model does not rely on the time series of historical sales data of products to be forecasted, which can thus provide overall sales forecasts for both old and new retail products.

Novel techniques such as artificial neural networks (ANN) have been applied successfully to problems concerning sales of food products. A study in the grocery industry by Penpece and Elam (2014) forecasted the sales revenue of the grocery retailing industry in Turkey by using artificial neural networks with marketing costs, gross profits, and competitors’ gross profits as inputs. Forecasts were very strong for major industry players thus the research concluded that ANN is a suitable tool for forecasting sales revenue at the grocery retailing industry. Doganis et al. (2005) applied a time-series methodology that combined two artificial intelligence technologies to forecast the daily sales volumes of short shelf-life food products, such as fresh milk.

Originally, SVMs were developed for pattern recognition and classification problems (Cortes, 1995) but have been adapted to solve regression problems in recent years (Wu, 2008; Du, 2011). In a study by Wen et al. (2014), the researchers developed an algorithm based on an SVM to predict the daily sales of grapes with deficient data. The study concluded that SVRs’ predictions are closer to real data than the artificial neural network and decision tree in this case.

With the variety of models mentioned above, one would naturally wonder what the best and most appropriate model is to use given a forecasting target. Comparative studies have been conducted to evaluate these contenders. Chu and Zhang (2003) presented a comparative study between linear and nonlinear models for retail sales forecasting. Traditional linear models, such as the ARIMA model, were compared with their nonlinear counterparts implemented via neural networks. The results suggest that the nonlinear method is the preferred approach and prior seasonal adjustment of the data can significantly improve the forecasting performance of the neural networks. Sharma and Sharma (2012) compared several machine learning techniques in sales forecasting including moving average, exponential moving average (EMA), ANN, K-Nearest Neighbour (KNN) and an approach combining both moving average and ANN. The study concluded that the combined approach gives more accurate results than all other techniques.

# Dataset

**Initial Raw Dataset**

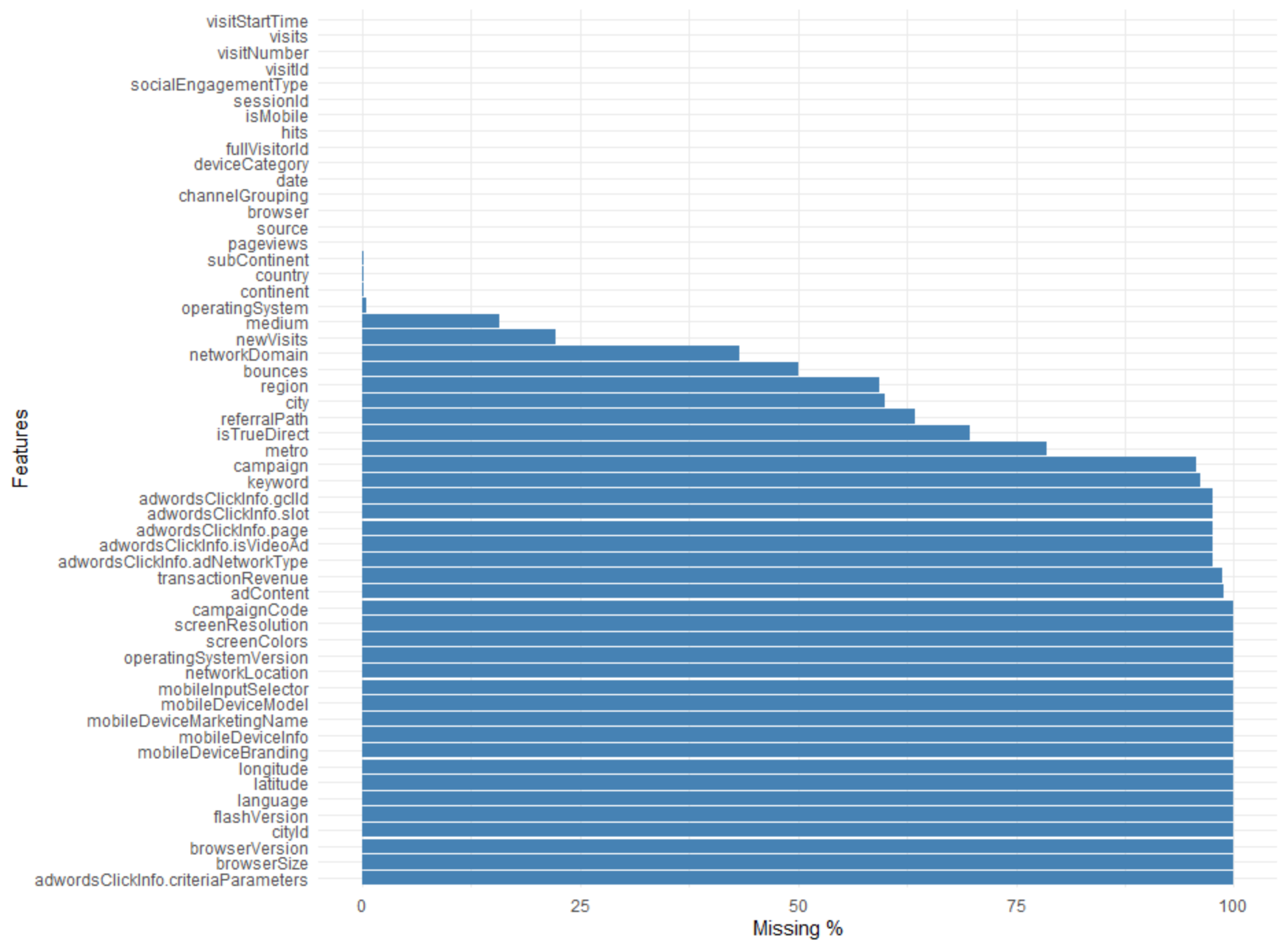
The dataset used for the project is the training dataset for the original Kaggle competition and will be split into training and test set in the modelling process. It contains information related to customer visits at a Google Merchandise Store between August 1st, 2016 to August 1st, 2017. Each row represents one visit to the store. The raw dataset includes 903,653 observations and 12 columns:

* fullVisitorId- A unique identifier for each user of the Google Merchandise Store.
* channelGrouping - The channel via which the user came to the Store.
* date - The date on which the user visited the Store.
* device - The specifications for the device used to access the Store.
* geoNetwork - This section contains information about the geography of the user.
* socialEngagementType - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
* totals - This section contains aggregate values across the session.
* trafficSource - This section contains information about the traffic source from which the session originated.
* visitId - An identifier for this session. This is part of the value usually stored as the \_utmb cookie. This is only unique to the user. For a completely unique ID, use a combination of fullVisitorId and visitId.
* visitNumber - The session number for this user. If this is the first session, then this is set to 1.
* visitStartTime - The timestamp (expressed as POSIX time).
* hits - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.

Four columns, including device, geoNetwork, trafficSource and totals, contain data in JSON format which need to be parsed to extract additional features. In one of those JSON columns, totals, the sub-column transactionRevenue is the revenue information we are trying to predict. The parsed dataset has 55 columns.

**Missing Values**

We observe that there are many missing values in the dataset. Values such as “not available in demo dataset”, “(not set)”, “unknown.unknown”, “(not provided)” should also be treated as NA. Upon examination 17 columns containing 100% missing values are removed from the dataset in this step.



**Data Type**

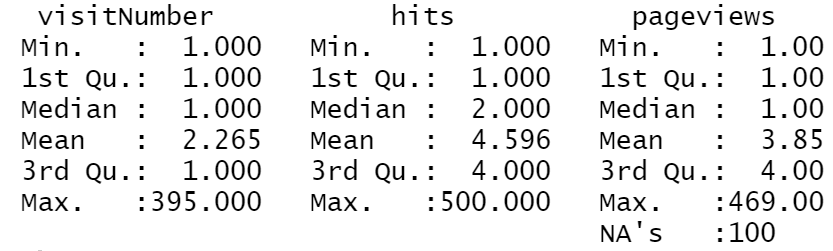
Several columns have incorrect data types, for example, date and visitStartTime are read as integer, some numeric columns are read as character, and all columns converted form JSON are of type character. Also, as indicated in the original data description, fullVisitorID must be loaded as character for all IDs to be properly unique. These columns are assigned the correct data types in this step.

**Constant Columns**

Examining the dataset, seven variables have only one unique value , i.e. zero variance. As they do not provide additional information in revenue prediction, they are removed in this step. The resulted dataset has 31 columns remaining.

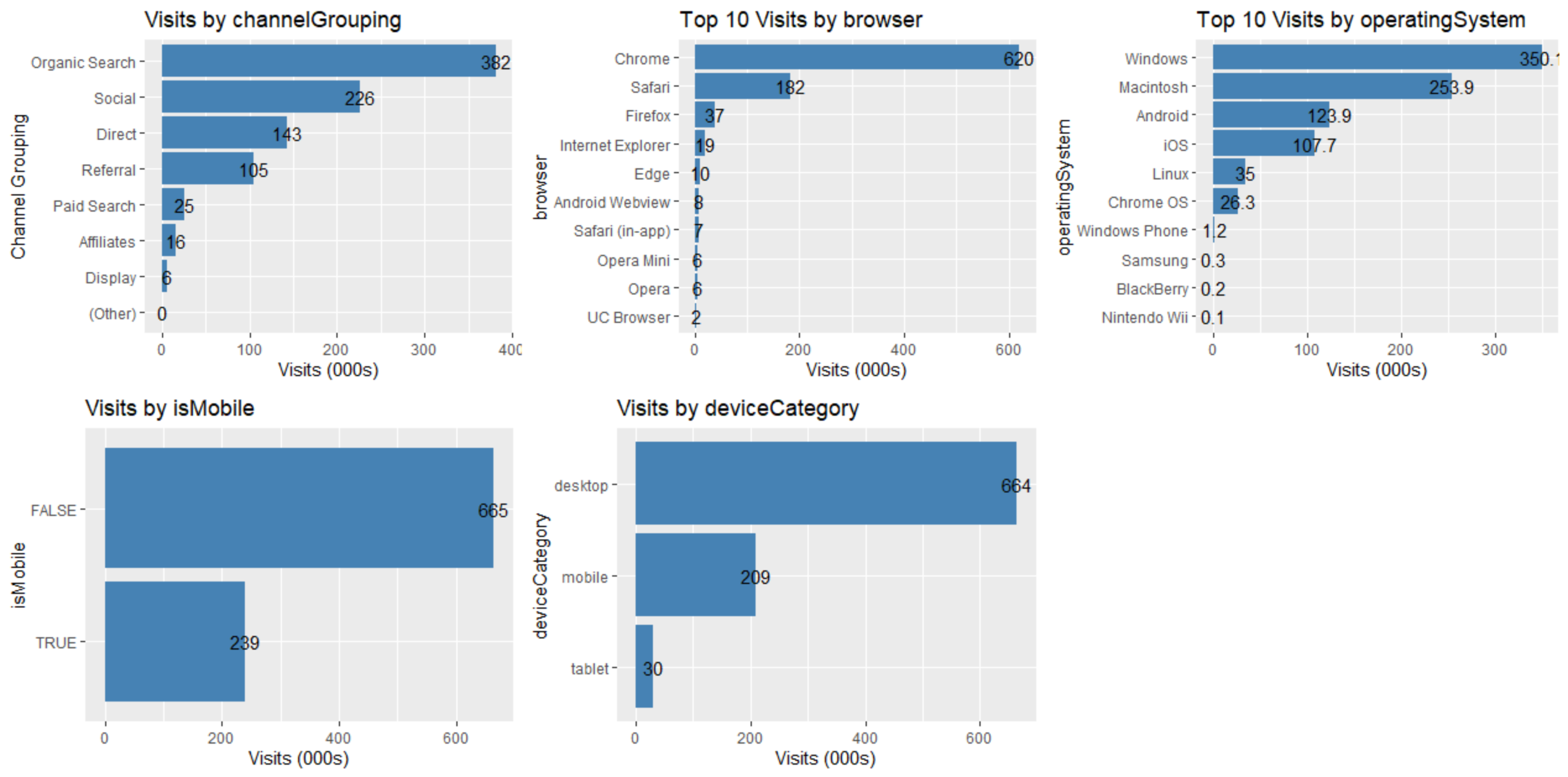
**Numeric Columns**

There are three numeric input columns remaining in the dataset, visitNumber, hits and pageviews. Based on the descriptive summary, on average, a customer has 2 sessions, 4 hits and 4 pages views in the store. All three variables have extreme outliers.

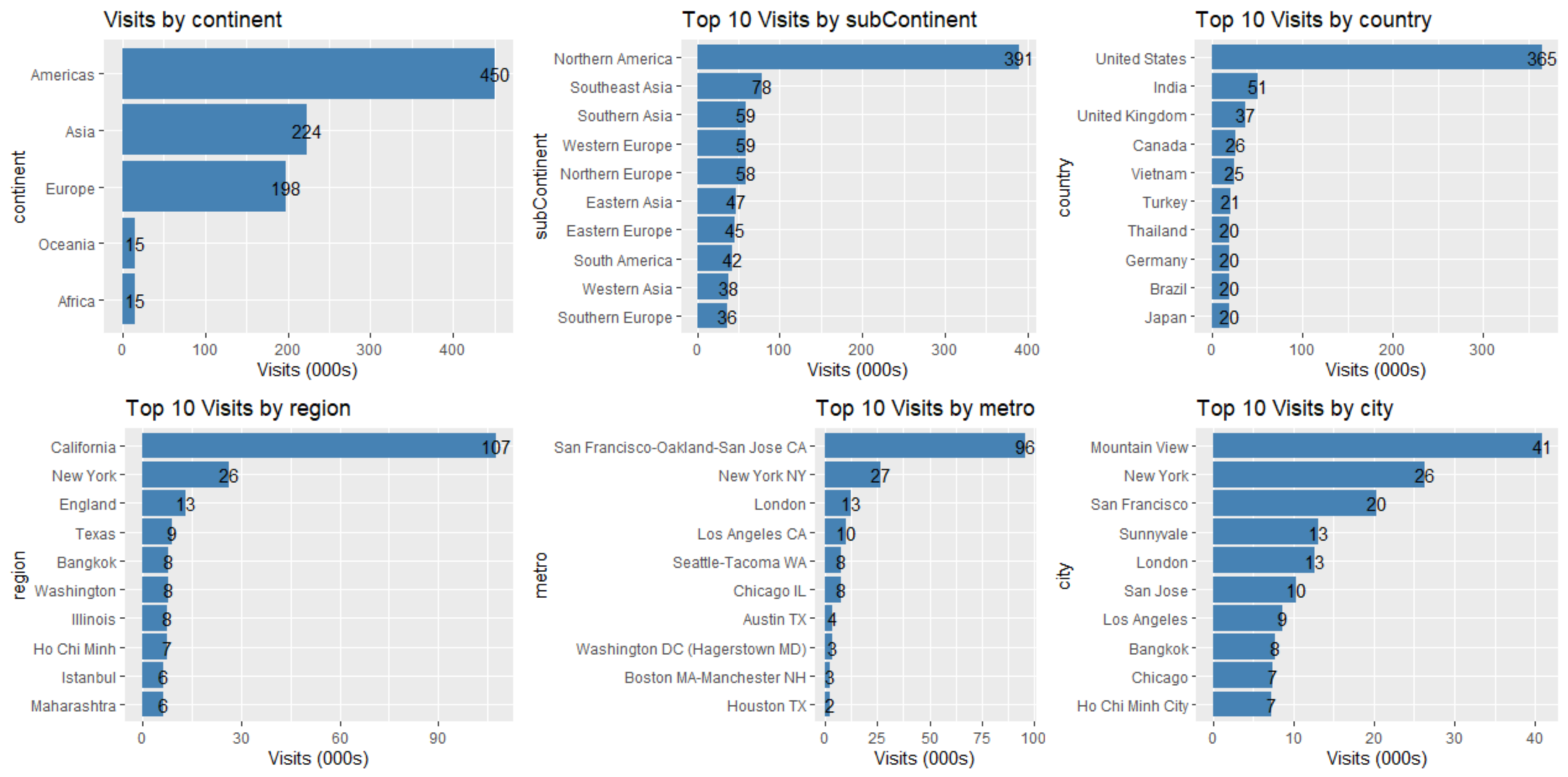


**Nominal Columns**

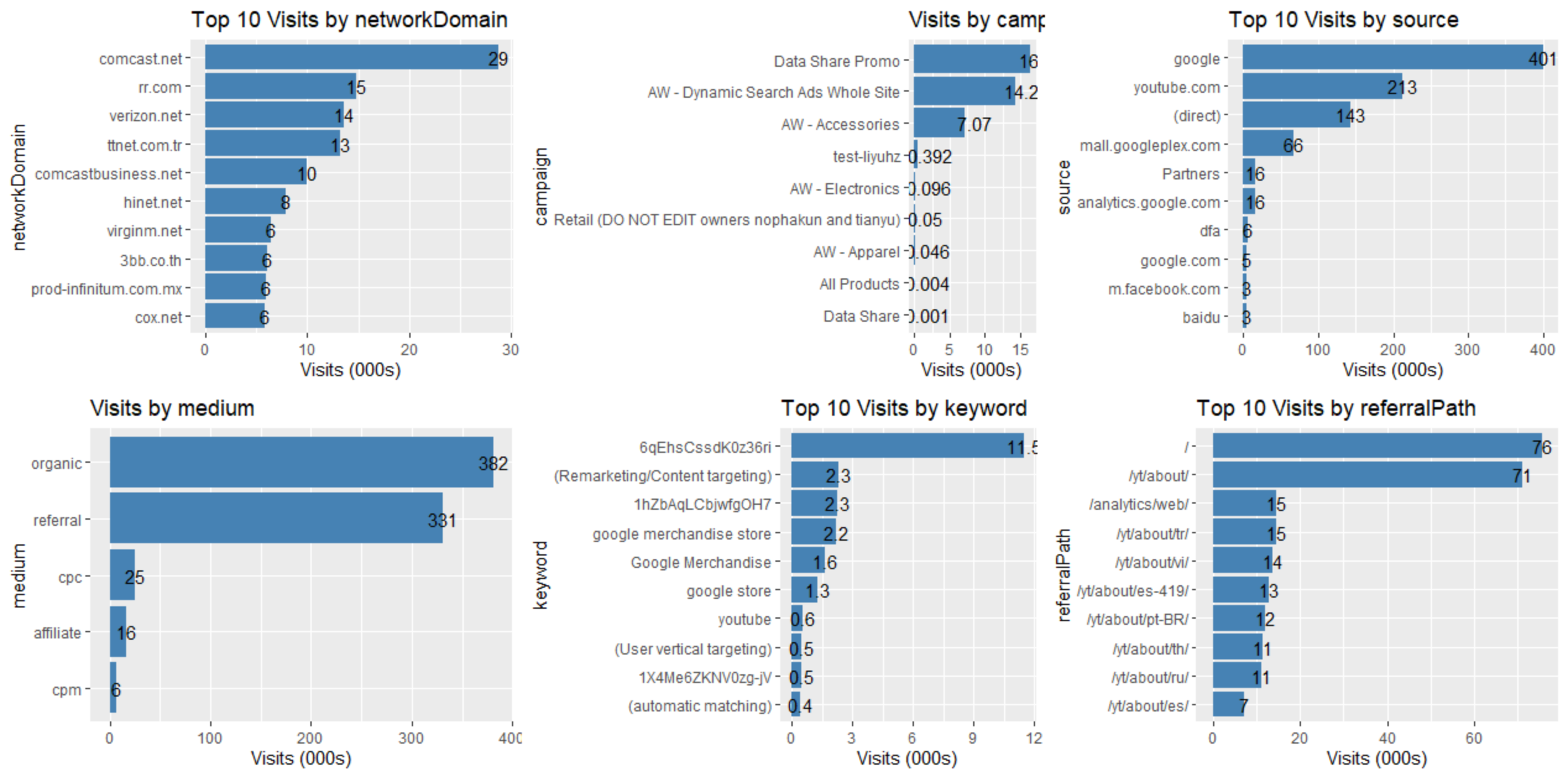
The nominal columns can be coarsely split into four groups based on the information they capture: access channel and platform, visitor geography, source of traffic, and advertisement. The first group provides information on the channel and platforms through which customers access the store. Most customers come to the store via organic search, followed by social network and directly entering the website. The data also reveals that referral is an effective channel of getting customer visits whereas paid search and affiliates are not as effective. Chrome is the dominant browser used by the customers to access GStore and only 26% the visits are initiated on a mobile devise.



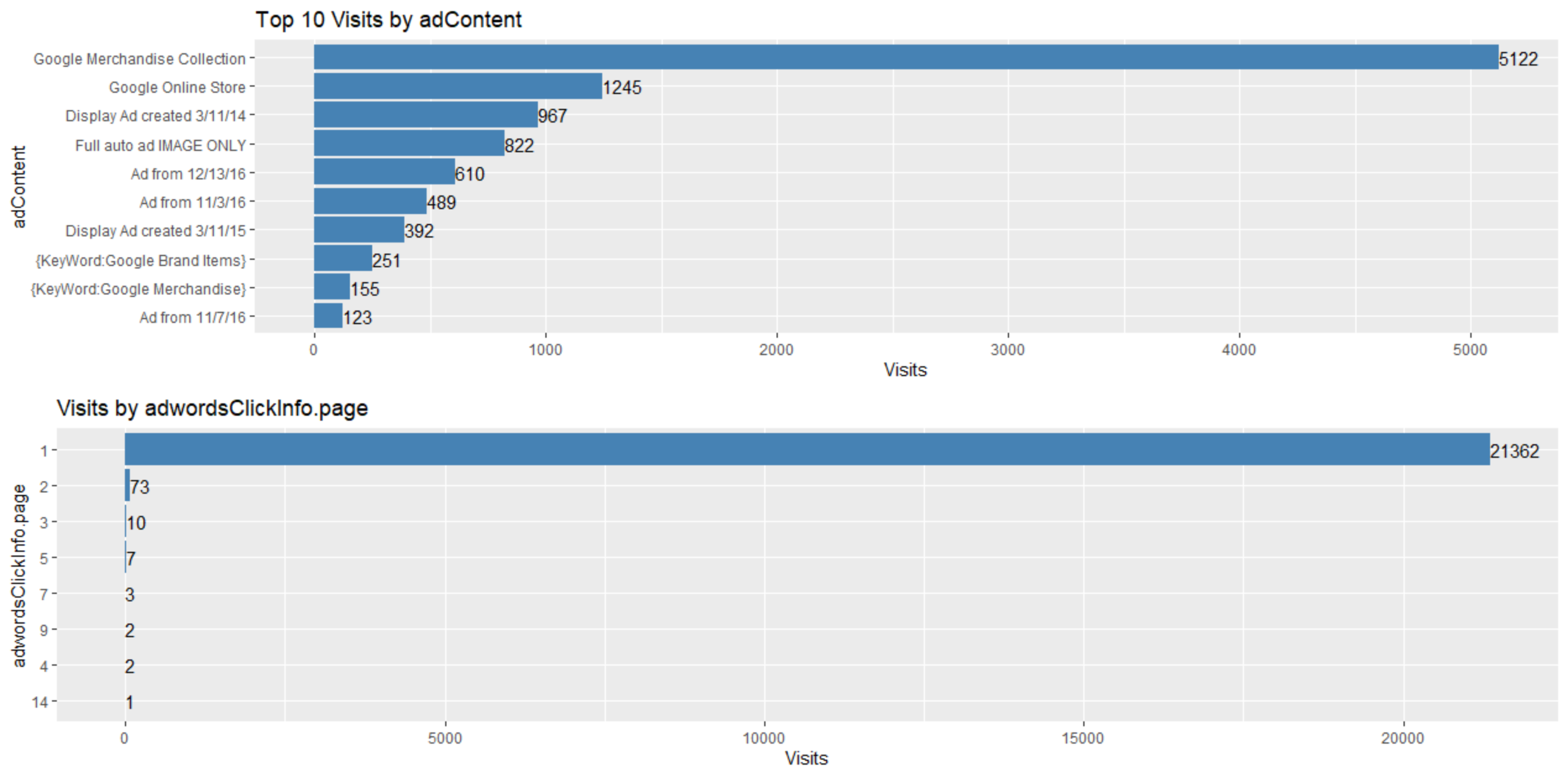
Geographically, the visits in North America darwfs the rest of the world with over 45% of the traffic coming from the United States alone. Region, metro and city all have over 50% missing values thus may not present a complete picture of the distribution.

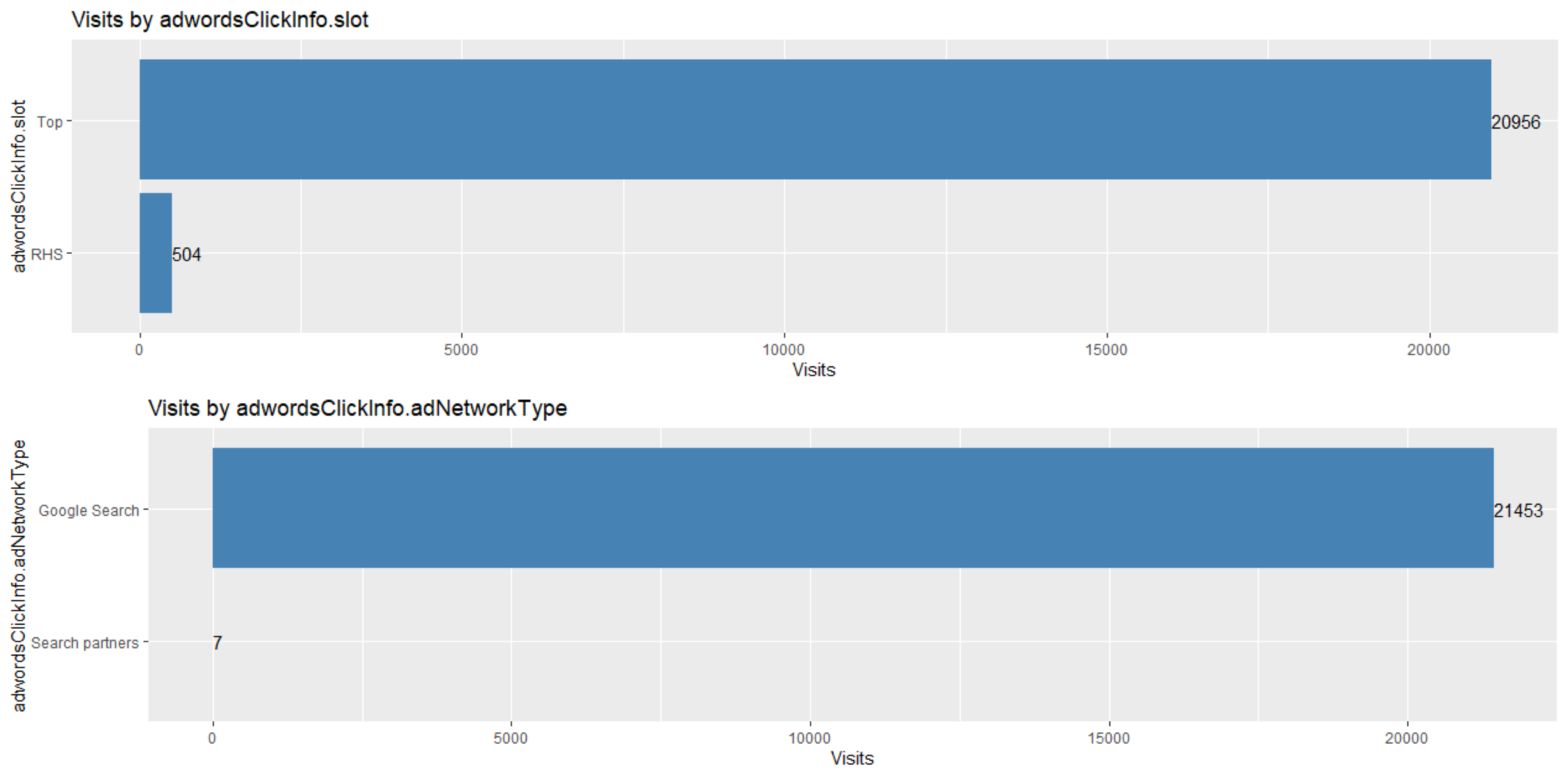


In terms of the source of traffic, google is the leading portal bringing almost twice as much visits as that generated by YouTube, which comes in second place. Organic search and referral are the dominate medium for traffic whereas cpc (cost per click), affiliates and cpm (cost per mille) entice little customer visits.



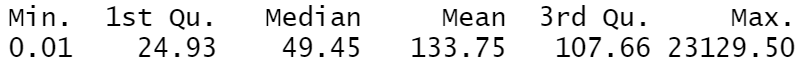
The last group of columns provide information on advertisements. Ad content featuring google merchandise collection appears to be the most effective amongst all. adwordsClickinfor.page tracks the page number in search results where the ad was shown and adwordsClickinfor.slot tracks the position of the ad. From the data, it is clear that customers seldom click beyond the first page and ad appears on the top of the page attracts more traffic compared to placement on the right hand side (RHS).



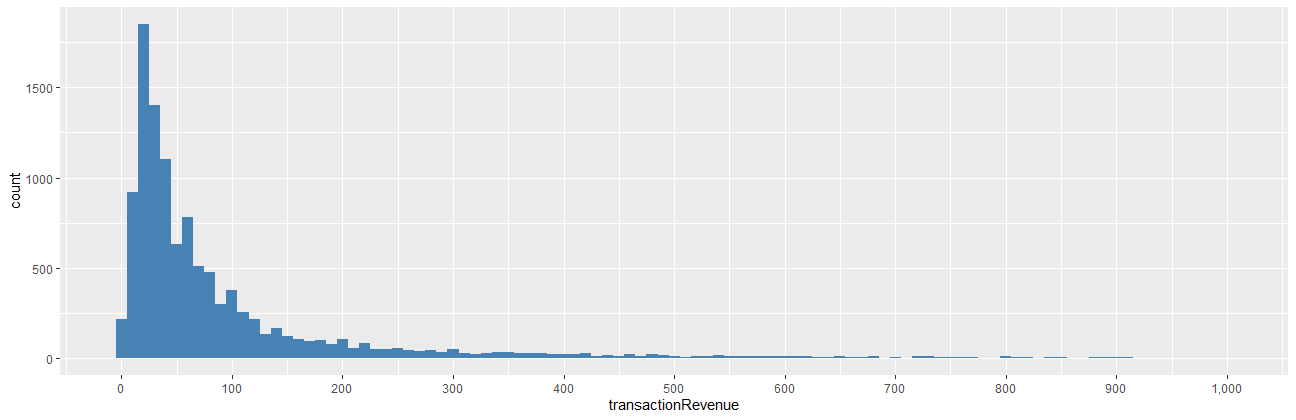


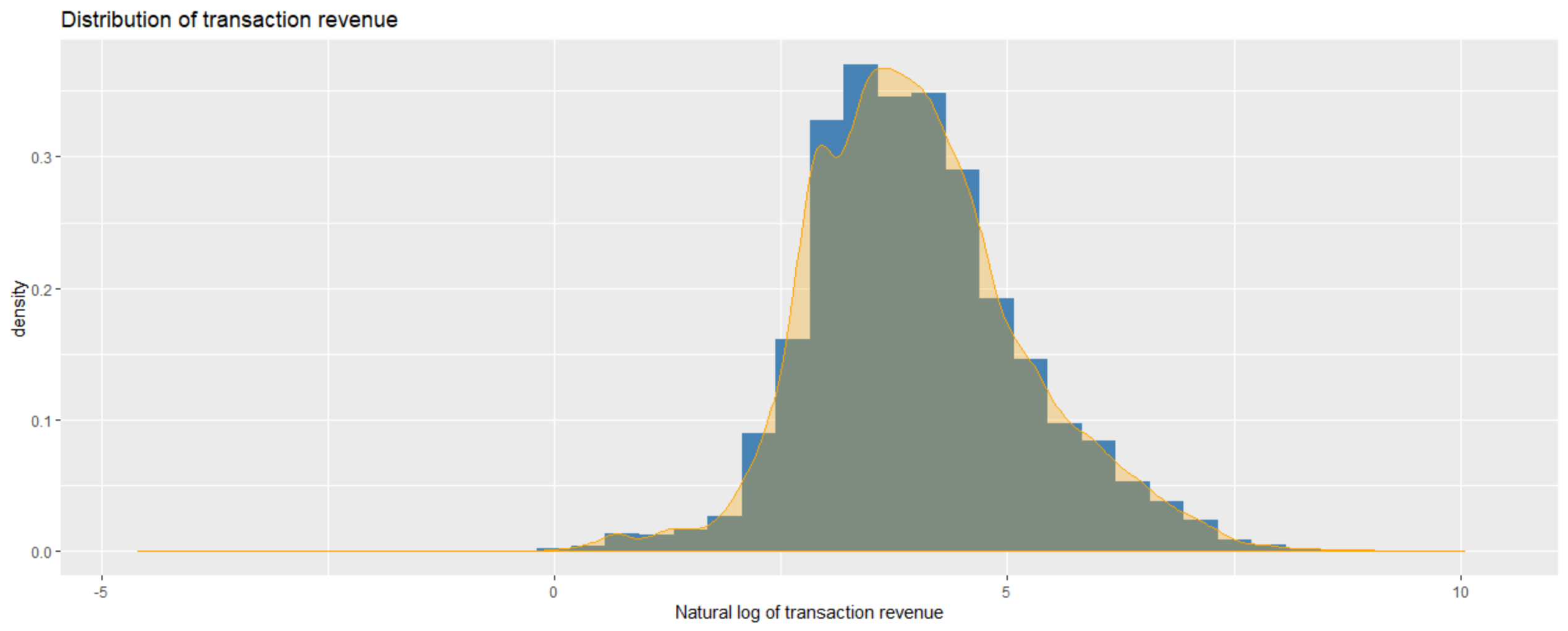
**Target Variable**

11,515 records have non-zero transactionRevenue which is only 1.27% of the total visits. This translates into a total revenue of 1.5 million USD over the 12-month period. For visits with purchase, the revenue ranges from one dollar cent to over 23,000 USD, half of which is below 50 dollars. Note that the transaction revenue column is multiplied by 10^6 in the raw data, not in unit dollar. It is converted back to unit dollars to represent actual transaction revenue for the rest of the analysis and prediction.



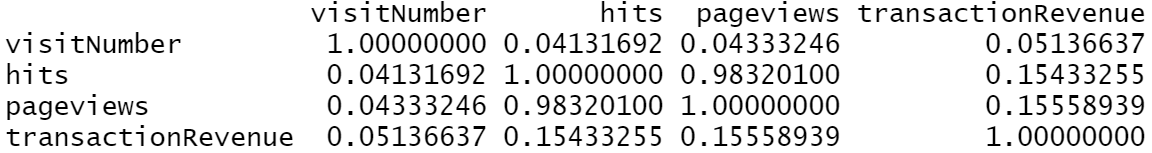
As the distribution of the revenue is very right skewed, the below histogram only displays count of transactions with revenue below 1,000 USD.



The natural log of revenue in unit dollar, however, appears to be approximately normally distributed with a mean around 4. 

**Correlation**

Examining the correlation between the numeric columns, i.e. visitNumber, hits, pageviews and transactionRevenue, we observe that hits and pageviews have a very strong correlation coefficient of 0.98. A hit is a request to a web server for a file, such as a web page, image, JavaScript, or Cascading Style Sheet. There may be many hits per pageview since an HTML page can contain multiple files, such as images. A pageview is each time a visitor views a page on the website regardless of the hits. Therefore, these two columns are correlated by definition, more pageview results in more hits. Both have a weak correlation with target variable, transaction revenue.



# Approach

The project is conducted using the approach depicted in the following charts.





## Step 1: Download Data

The data is downloaded from Kaggle website for the original GStore competition which can be found in this link: <https://www.kaggle.com/c/ga-customer-revenue-prediction/data>. There are 3 datasets for the competition, training, testing and the list of visitor ID of which the revenue is to be predicted. Only the training set is downloaded for this project as the test set posted does not contain the target variable, i.e. transactionRevenue.

## Step 2: Load Data

This step loads the downloaded dataset into RStudio. Four columns in the dataset are in JSON format therefore are parsed to extract additional features.

## Step 3: Identify Missing Values

This step identifies missing values in the dataset. Values such as “not available in demo dataset”, “(not set)”, “unknown.unknown”, “(not provided)” are also replaced with value “NA”. Columns containing 100% missing values are removed from the dataset.

## Step 4: Correct Data Type

This step assigns the correct data types to each column, including the ones parsed from JSON columns.

## Step 5: Identify Constant Columns

This step identifies and removes columns with constant values, i.e. zero variance.

## Step 6: Univariate Analysis

This step analyzes the individual column in the training set. The columns are examined in two separate groups based on data types, which are numeric and nominal. Within the nominal group, the columns are further divided into four sub-groups based on the information captured: access channel and platform, visitor geography, source of traffic, and advertisement. Five-number statistics summary are calculated for the numeric columns whereas frequency charts are plotted for nominal columns.

The target variable, transactionRevenue, is also examined in this step. The transaction revenue column in the raw data is multiplied by 10^6, not in unit dollar. It is converted back to unit dollars in this step for the rest of the analysis and prediction. Frequency and density of log of target variable are plotted to show the distribution of revenue.

## Step 7: Bivariate Analysis

This step examines the Pearson correlation between numeric columns.

## Step 8: Exploratory Analysis

This step explores the relationship between the target variable and input columns. Total visits, total revenue and mean revenue are plotted for key input columns to draw meaningful insights. Heatmap is used to show the concentration of revenue by geographic regions.

Revenue is also examined as timeseries to show any patterns over time. Simple ARIMA and ARIMA plus pageviews as an additional regression are run to predict log of daily revenue .

## Step 9: Linear Regression

This step attempts to forecast log revenue using both simple and multiple regressions. Feature selections and further dimension reduction is performed, and stepwise AIC method is used to select the best feature combination.

## Step 10: KNN

This step attempts to forecast log of revenue using KNN regression.

## Step 11: XGBoost

This step attempts to forecast log revenue using XGBoost.

## Step 12: Result Comparison

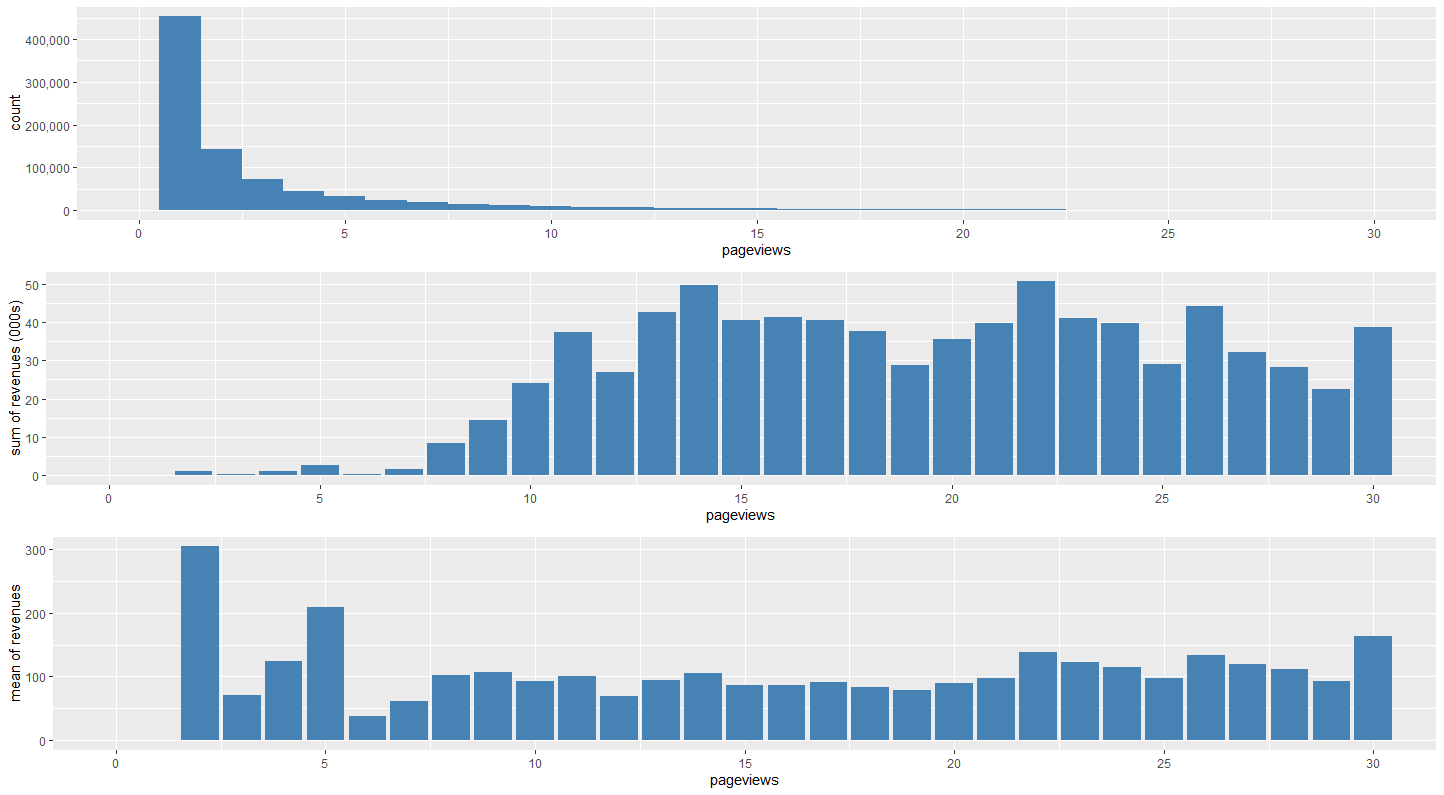
This step compares the forecast results from ARIMA, linear regression models, KNN and XGBoost. Models are evaluated based on the root mean squared error (RMSE).

# Exploratory Data Analysis

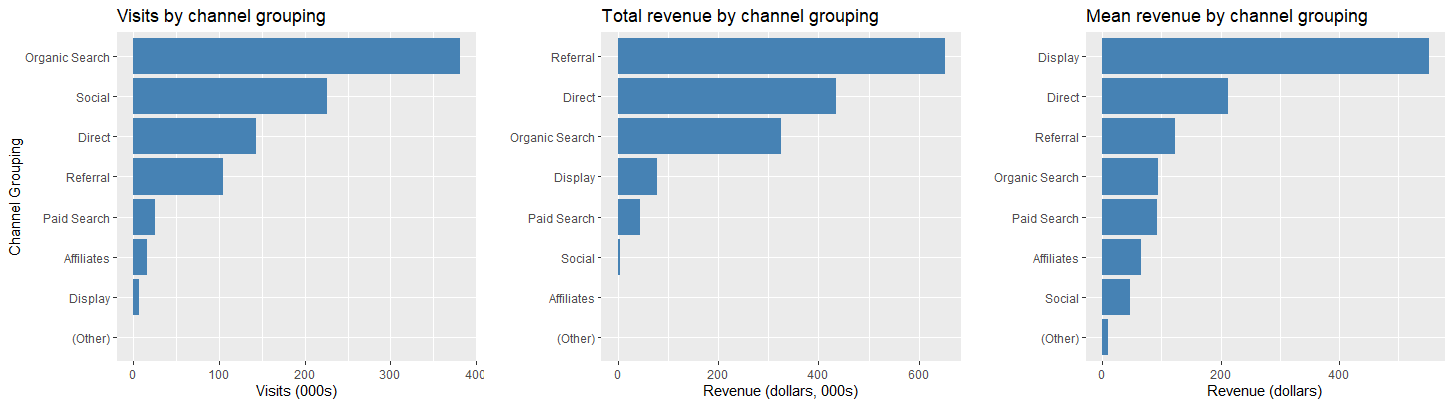
**Pageviews & Revenue**

A pageview is each time a visitor views a page on the website regardless of the hits. The distribution of pageviews is extreme right skewed as show in the first plot below. Sessions with more than 30 pageviews are hardly visible with frequencies of less than 1,000, therefore are excluded from the plot.

The second plot shows that visits with one pageviews, although frequent, do not generate any revenue. Most revenue comes from relatively small numbers of sessions with over more pageviews .

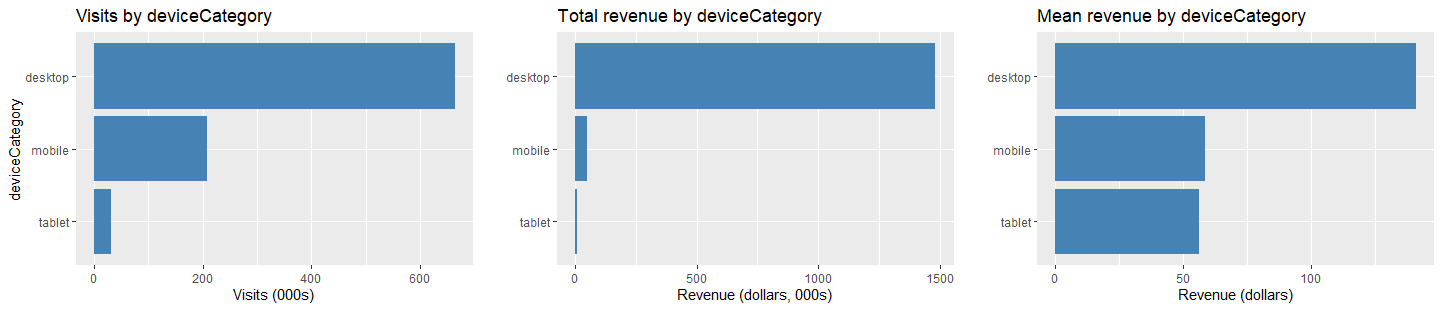


**Channel & Revenue**

Channel defines how users come to the website. Organic Search leads to the most visits but only moderate revenue. Social generates the second most visits but the visits hardly translate into revenue. On the other hand, Direct and Referral, especially Referral, delivers most revenue with a relatively small number of visits. 

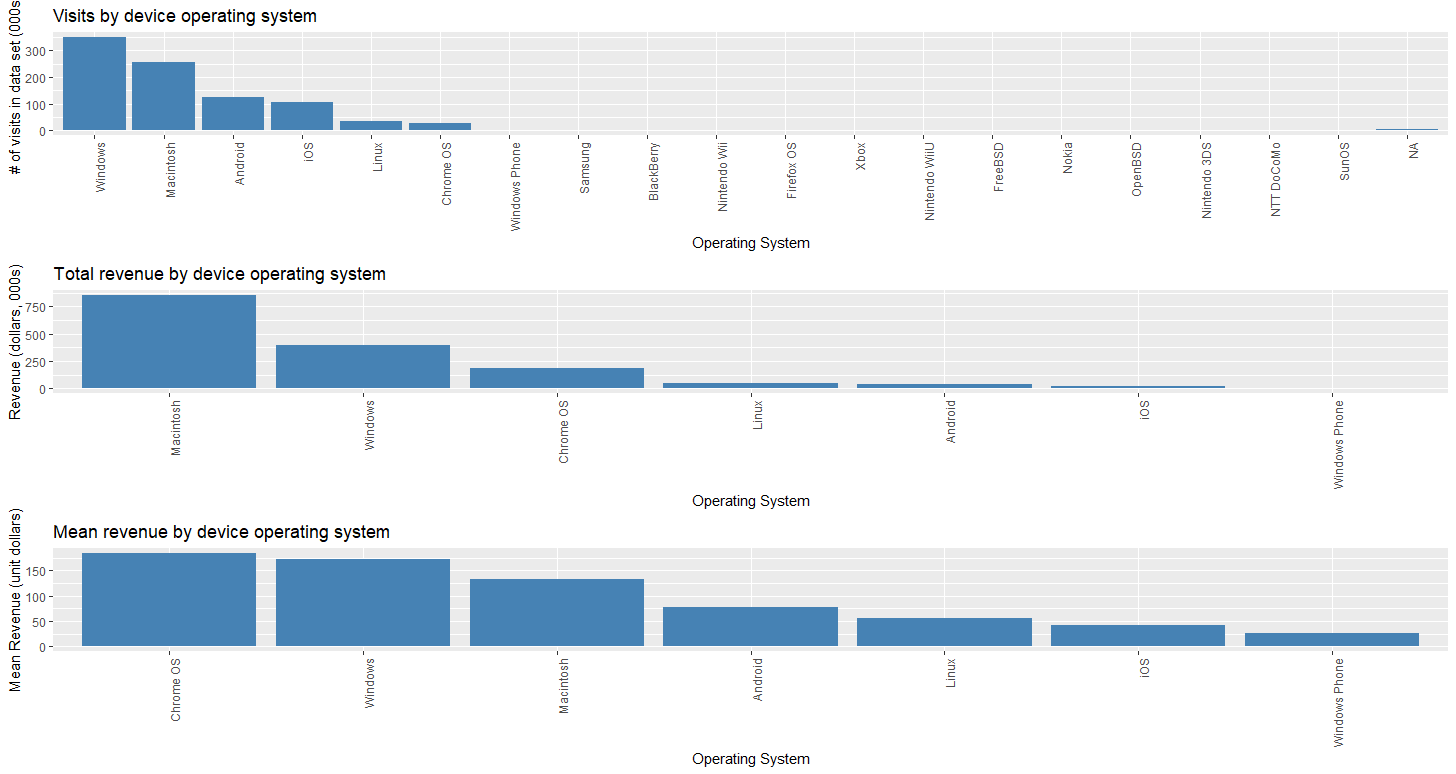
**Mobile Device & Revenue**

In the time period examined, majority of the visits and revenue is generated by desktop customers. Mobile and tablet have considerably fewer sessions and lower revenues per visit.



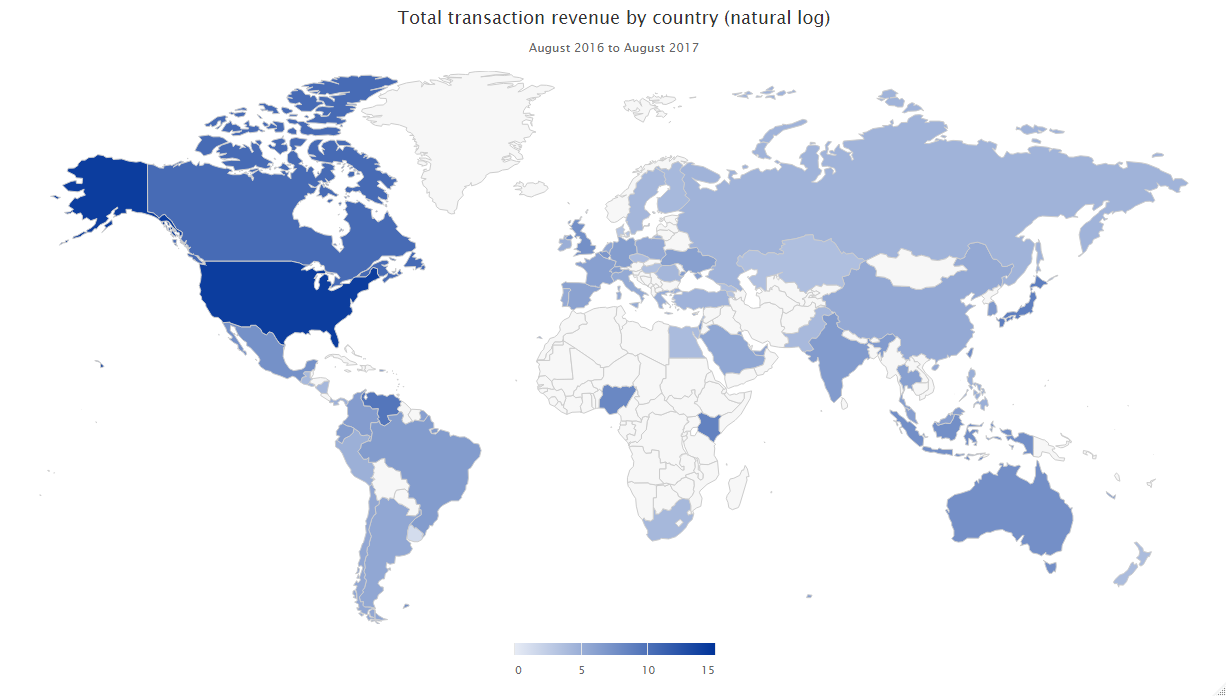
**Operating System & Revenue**

In terms of operating systems, what stands out is that Macintosh generates more than twice the revenue than Windows with fewer sessions. In addition, it confirms again that the mobile visits, i.e. Android, iOS and Windows Phone, lead to little purchase compared to the number of visits.



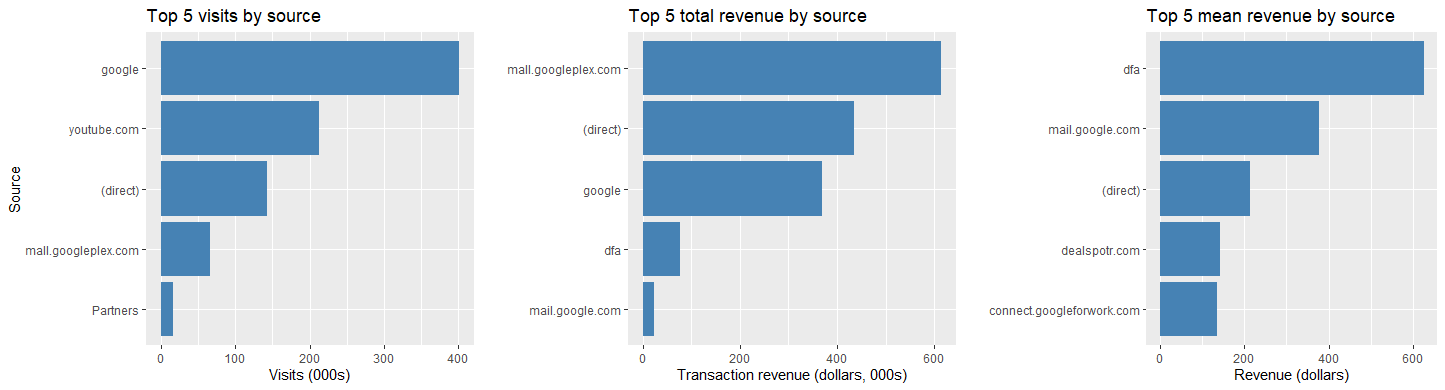
**Country & Revenue**

The heatmap below shows the concentration of revenue by country. Log of total transaction revenue rather than raw transaction revenue is used to get better dispersion for the choropleth palette. The United States has the highest concentration which is consistent with the number of visit plotted in the data description section.



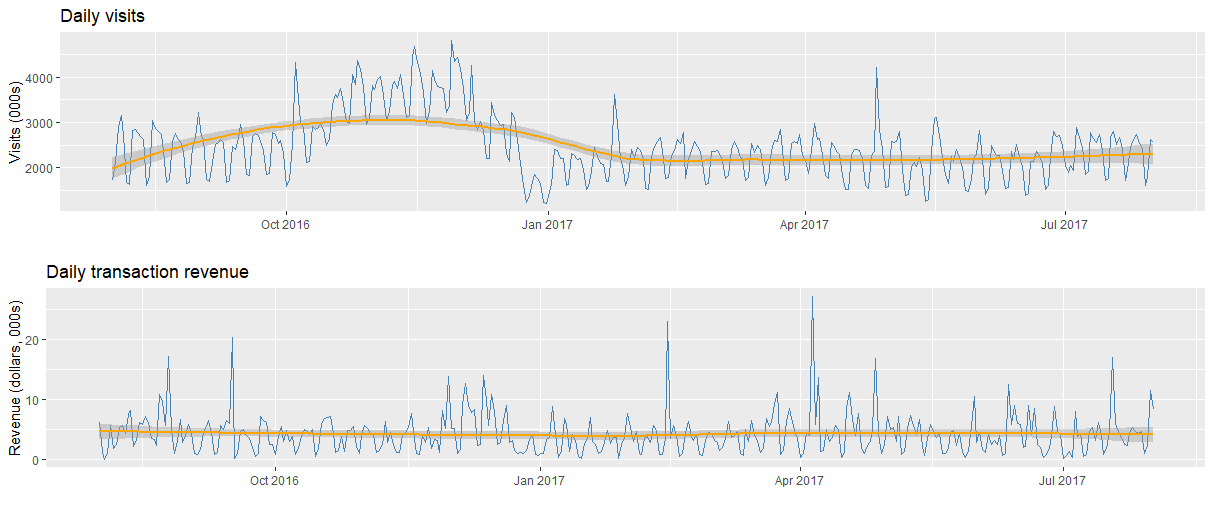
**Source & Revenue**

The Source plots reveal a surprising relationship between visits and revenue. Although google and YouTube generate a lot of traffic, they do not deliver nearly as much revenue, especially YouTube. On the other hand, direct and mall.googleplex.com generate more revenue with a modest amount of sessions.

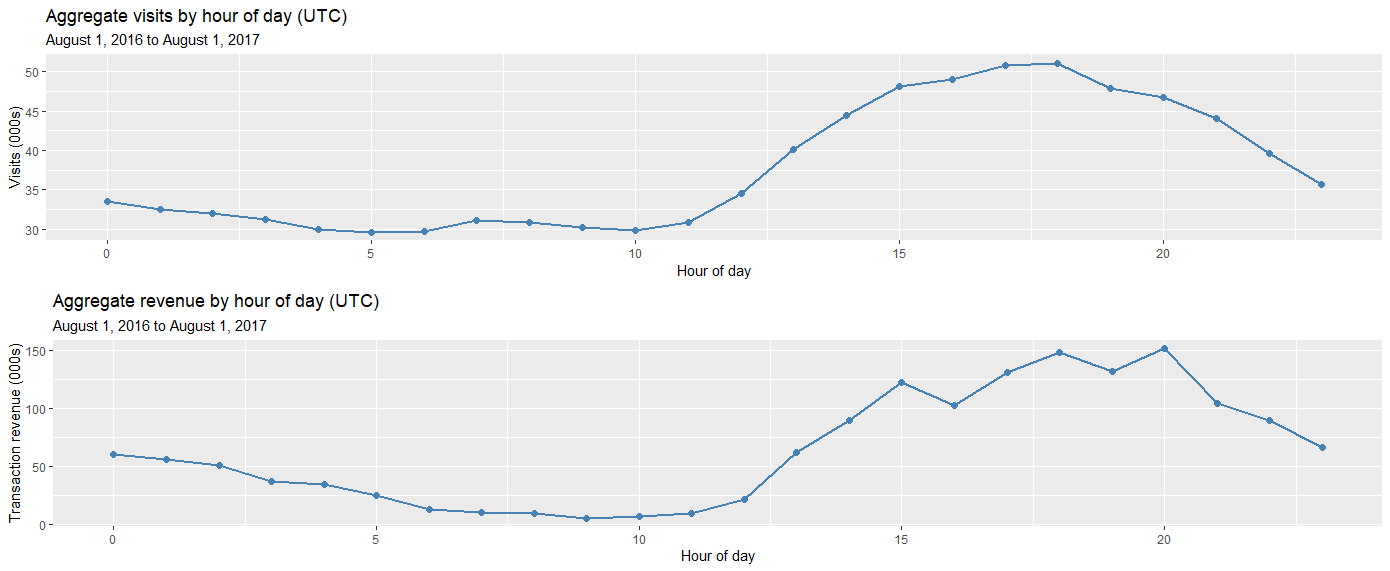


**Time & Revenue**

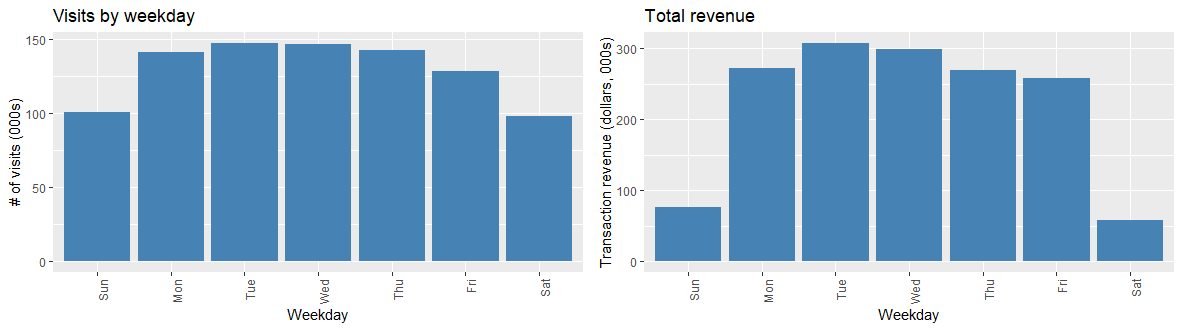
The number of daily visits peaks in November and December 2016, but this does not result in higher daily revenue. In fact, there is no visible trend in daily revenue, nor does it appear to have any correlation with daily visits. There are several peaks in daily revenue, such as end of February and beginning of April 2017, which could be a result of promotional events and worth further investigation.



There is a clear trend in visits and revenue during different hours of the day. Both seem to pick up around noon, peak at 6pm and then gradually level off for the rest of the day.

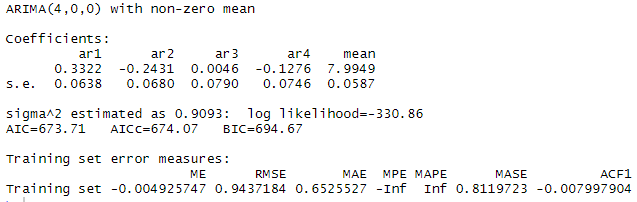


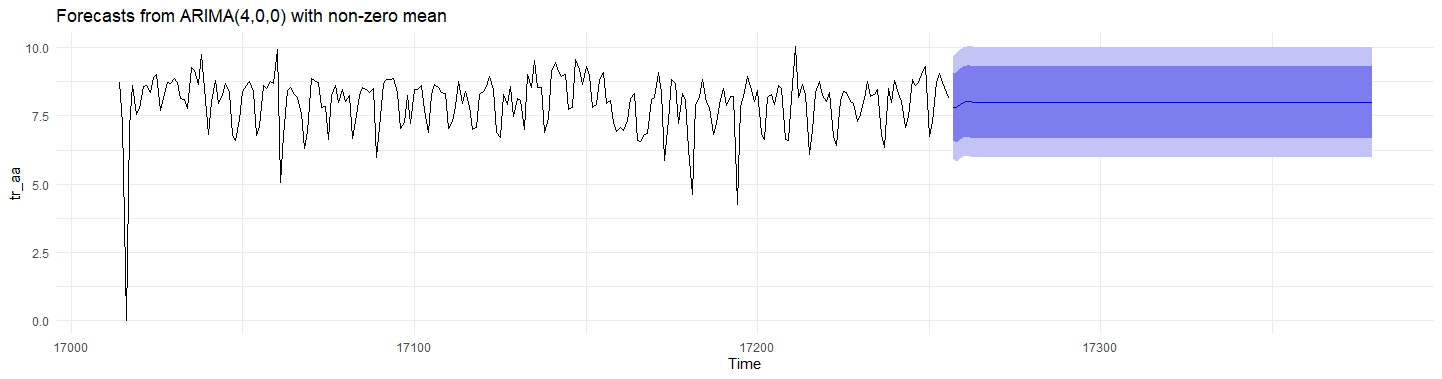
Weekends have less visits and significantly lower purchase compared to weekdays. A logical explanation could be that Google swag is mostly bought by businesses rather than private individuals, therefore are concentrated during the weekdays, especially mid of the week.



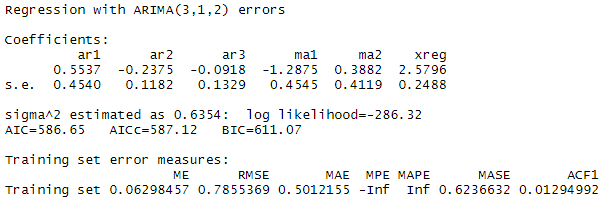
**Timeseries ARIMA**

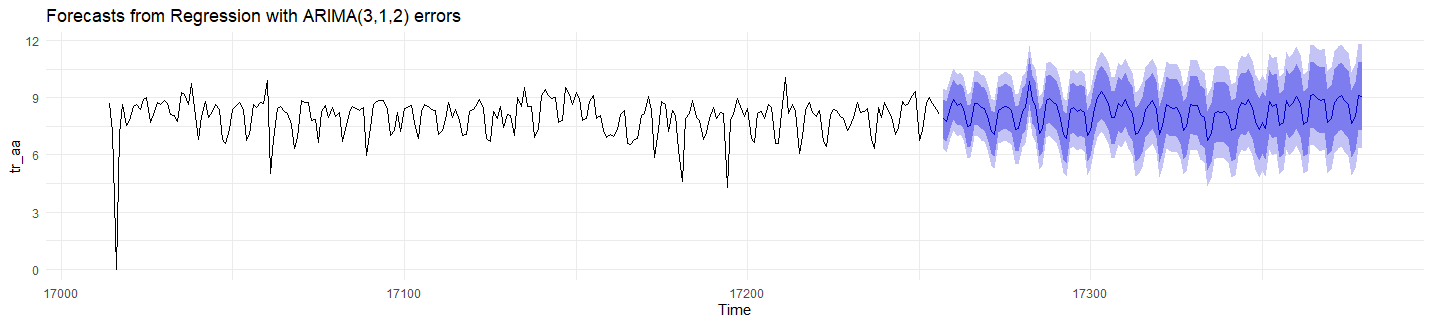
Given that revenue is a timeseries, it is intuitive to attempt to predict it using a timeseries model. Here we use ARIMA, which is a widely used model to predict timeseries data. The data is split into training and test set based on date. Training set includes all visits from August 1st, 2016 to March 31st, 2017 (approximately 70% of the data) and test set includes April 1st, to August 1, 2017. Revenue is firstly aggregated by date and then fed into the model as the only input. The table below shows the output from the auto.arima function in R. It is clear from the line chart that this forecast model has little predictive power beyond a very short period of time. The RMSE on the test set is 0.9596.





To improve the ARIMA model, daily pageview is added as a regressor. The updated line chart displays a reasonable fluctuation of daily revenue as opposed to a flat line. The RMSE on the test set is 0.7248, an improvement from the simple ARIMA model.



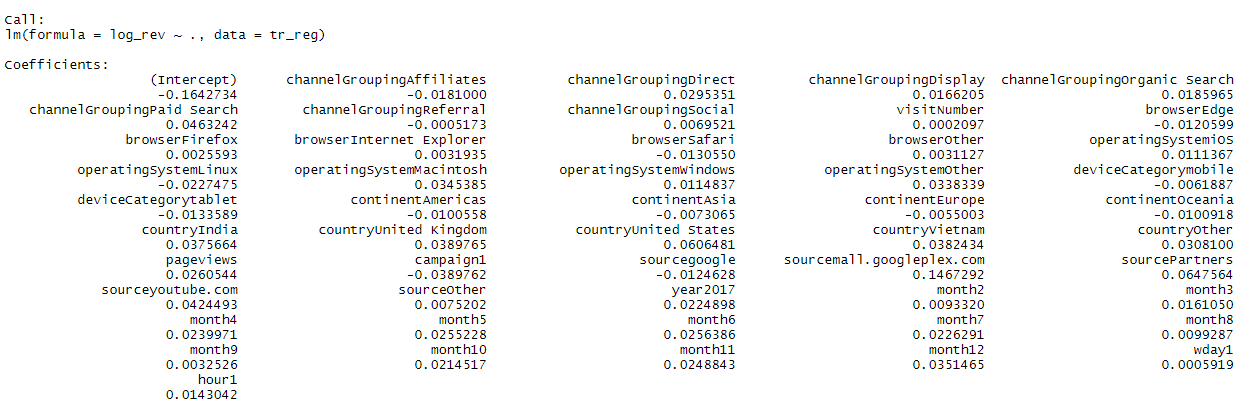


# Results

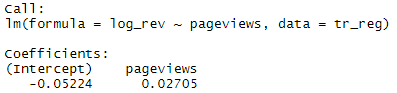
**Linear Models**

To predict log of revenue, we first start with linear regression models. Both simple and multiple regression models are considered for comparison. The following 16 features are selected based on observations from data description section and exploratory data analysis: date, channelGrouping, visitNumber, browser, operatingSystem, deviceCategory, continent, country, pageviews, campaign, source, log\_rev, year, month, wday (i.e. day of the week), hour. Most of these features are categorical (class “factor” in R) and level reduction is performed to reduce the dimension in the model. For example, all levels except for the top five levels of browser are lumped into level “other”; day of the week has only two levels, i.e. weekday and weekend; hour of the day is also reduced to two levels, i.e. 0 to 12, 13 to 23. The data is split into training and test set using 70/30 rule with random sampling.

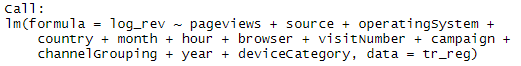
The first model includes all 16 selected features, The RMSE on the test set is 0.4253. Coefficients are as below:



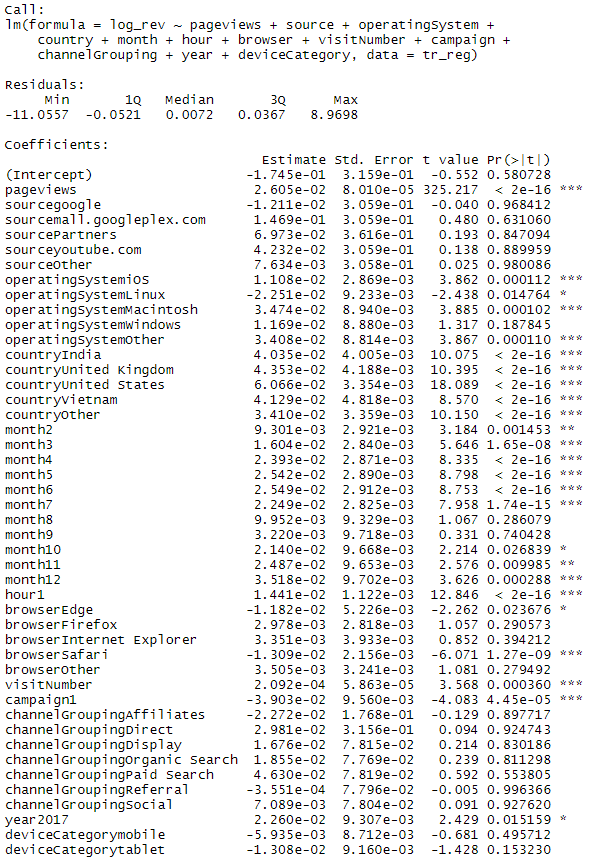
In the data description section, we observed that there is a weak correlation between pageviews and revenue. regsubsets function in R, which automatically selects best features for regression models, also shows pageviews as the best single predictor out of all 16 selected features. Therefore, the second model uses pageviews as the single input to predict revenue. The RMSE on test set is 0.4279, slightly higher than the full model . Coefficient is as below:



The last linear model is constructed using features selected by stepwise AIC. stepAIC function in R is used here. The final model includes 12 out of 16 features:



Coefficients are shown in the table below. For most input variables that are significant at 0.05 level, there is positive relationship between them and the log revenue. It seems that customers with a Linux operating system or with an Edge or Safari browser generate less revenue per session. Another interesting finding is that campaign seems to reduce the revenue per session. This is rather counter-intuitive and may warrant further investigation.

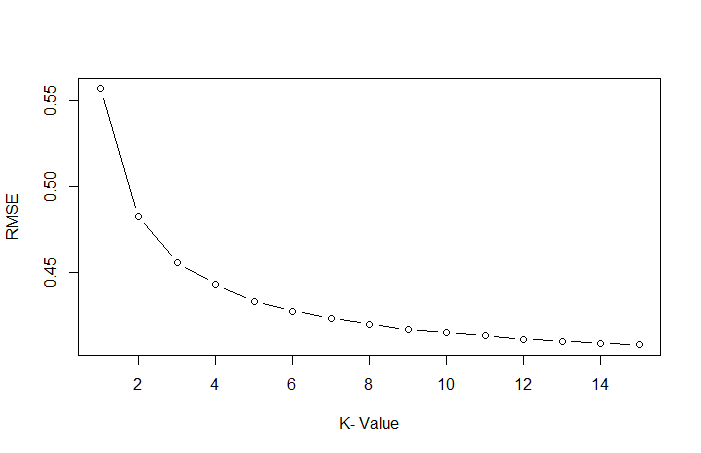
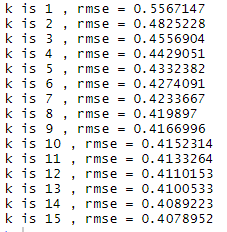




This model achieved a RMSE of 0.4253 on the test set, same as the full model. However, the R-squared is only 0.17, meaning only 17% of the variance in log revenue is explained by the model.

**KNN**

KNN, a supervised machine learning algorithm, is also considered for prediction. When k is 7, KNN achieves a level of RMSE comparable to the multiple linear model generated by stepwise AIC. As k increases, RMSE continues to decrease and approaches 0.40.



**XGBoost**

The final model is constructed based on XGBoost, a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. The chart below shows the importance of features based on XGBoost. Pageviews is clearly the top feature which is consistent with the output from the linear regression. The RMSE on the test set is 0.4024, the lowest amongst all models considered.

|  |
| --- |
|  |

**Result Comparison**

Based on RMSE, XGBoost is the best model out of five models considered. It is worth noticing that with RMSE around 0.4, the models only explain about 17% of the variance in log revenue.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Multiple Linear Regression | Simple Linear Regression (pageviews) | Optimal Multiple Linear Regression | KNN (k=15) | XGBoost |
| RMSE | 0.4253 | 0.4279 | 0.4253 | 0.4079 | 0.4024 |

# Conclusions

The project attempts to predict the revenue per visit for Goggle GStore using data provided for a similar Kaggle competition. To understand customer behaviour and generate revenue projection is one of the most fundamental and imperative problem in business analytics. The analysis shows that number of pageviews is an important predictor of customer visit and more pageviews leads to higher revenue during that visit. Among the models explored, XGBoot performed the best according to RMSE measure.

The data also reveal a number of interesting insights which could inform managerial decisions. Although Google is a company with global footprint, the visits to and revenue generated in GStore is heavily concentrated in the Americas, especially the United States. Customers who are referred to the store or access the store by directly entering the website are the top revenue regenerators. On the other hand, customers who are directed to the store by organic searches or Youtube hardly make actual purchases despite the high volumes. Majority of customers access and make purchases on desktop and Macintosh users generate more than twice the revenue than Windows with fewer sessions. Purchase volume is higher during the weekdays and concentrated in the afternoon throughout early evening hours.

# References

|  |
| --- |
| Chen, I., Chen, I., Lu, C., & Lu, C. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing.*Neural Computing and Applications, 28*(9), 2633-2647. doi:10.1007/s00521-016-2215-x |
| Chu, C., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting.*International Journal of Production Economics, 86*(3), 217-231. doi:10.1016/S0925-5273(03)00068-9 |
| Cortes, C. (1995). Support-vector networks. *Machine Learning*, 20(3), 273. doi:10.1023/A:1022627411411 |
| Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing.*Journal of Food Engineering, 75*(2), 196-204. doi:10.1016/j.jfoodeng.2005.03.056 |
| Du, X. F., Leung, S. C. H., Zhang, J. L., & Lai, K. K. (2013). Demand forecasting of perishable farm products using support vector machine.*International Journal of Systems Science, 44*(3), 556-567. doi:10.1080/00207721.2011.617888 |
| Guo, Z. X., Wong, W. K., & Li, M. (2013). A multivariate intelligent decision-making model for retail sales forecasting.*Decision Support Systems, 55*(1), 247-255. doi:10.1016/j.dss.2013.01.026 |
| KumarSharma, S., & Sharma, V. (2012). Comparative analysis of machine learning techniques in sale forecasting.*International Journal of Computer Applications, 53*(6), 51-54. doi:10.5120/8429-2198 |
| Lu, C., & Chang, C. (2014). A hybrid sales forecasting scheme by combining independent component analysis with K-means clustering and support vector regression.*The Scientific World Journal, 2014*, 624017-8. doi:10.1155/2014/624017 |
| Pavlyshenko, B. (2019). Machine-Learning Models for Sales Time Series Forecasting. *Data*, *4*(1), 15. doi:10.3390/data4010015 |
| Penpece, D., & Elma, O. E. (2014, October). Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey. *International Journal of Trade, Economics and Finance*, 5. |
| Wen Q., Mu W., Sun L., Hua S., Zhou Z. (2014) Daily Sales Forecasting for Grapes by Support Vector Machine. In: Li D., Chen Y. (eds) Computer and Computing Technologies in Agriculture VII. CCTA 2013. *IFIP Advances in Information and Communication Technology*, vol 420. Springer, Berlin, Heidelberg |
| Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages.*Management Science, 6*(3), 324-342. doi:10.1287/mnsc.6.3.324 |
| Wu, Q., Yan, H., & Yang, H. (2008). A forecasting model based support vector machine and particle swarm optimization. Paper presented at the 218-222. doi:10.1109/PEITS.2008.37 |