# Revenue to Plan Prediction – Milestone- 1 Report

1. **Introduction**
   1. *Problem Statement*

Optimization of business decisions across the company are made based on the revenue forecast. Since revenue forecasts are the backbone to business decisions across the entire organization (how much inventory should I order, how many people do I need in department x to fulfill expected demand, how much money do I need to invest in advertising, etc.), failing to reach or exceed expectations can be detrimental to retailers. Creating an accurate revenue forecast is challenging for any retailer, and then followed by the additional challenge of understanding on a day to day basis if you over or under-forecasted. Retailers are making decisions daily regarding merchandising, marketing plans and marketing spend. As a digital marketing agency, our clients enlisting our services to produce and execute on strategies to reach and exceed these goals. Our job is to make accurate assessments of performance, set appropriate expectations, as well as to maximize digital marketing campaigns for our clients to achieve their goals.

* 1. *Client: For confidentiality purposes, the name of this client will remain anonymous. The client will be referred to as “the Client” throughout.*

The client for this project is a women’s fashion retailer selling a variety of product categories such as clothing, shoes, and jewelry with their primary sales around their handbag product line. While the brand owns and operates a handful of full-price and outlet retail stores as well as selling merchandise through third party retailers, this project focuses on predicting whether the client will achieve their goal for their eCommerce business.

This client is publicly owned and operated. At the beginning of each year, the Executive Level Management team releases their annual forecast projections for the eCommerce division of the company broken down by month. The marketing team for the eCommerce division utilizes this information to create their own set of daily goals and their marketing plan for each month. They then work with an agency partner to help strategize and execute these marketing plans to reach these goals. Should the team not achieve their goals then they risk consequences such as budget cuts and potential job loss. Should the agency not achieve these goals then the agency risks losing the client.

* 1. *Dataset*

The final dataset consists of historical data around the client’s eCommerce website, their daily eCommerce goals (provided directly from the client) and their daily advertising spend for digital marketing channels. There are 393 observations (each observation is data from a unique date ranging from 12/4/2016 through 1/1/2018). December 4, 2016 was the start date for the client for fiscal December 2016.

Feature Details:

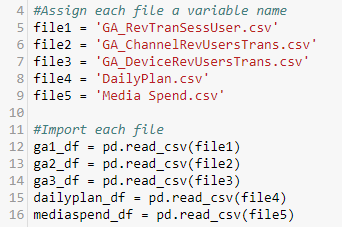
* Date
* Revenue
* Users
* Sessions
* Transactions
* Quantity
* Bounces
* Pages / Session
* Avg. Session Duration
* Session Duration
* Revenue, Users and Transactions, Each Broken Down for the Following Marketing Channels by Date (39 columns):
  + Affiliate
  + CSE
  + Direct
  + Display
  + Email
  + Organic Search
  + Organic Social
  + Other
  + Paid Search
  + Paid Social
  + Partnerships
  + Referral
  + Social
* Revenue, Users and Transactions, Each Broken Down for the Following Devices by Date (9 columns):
  + Desktop
  + Mobile
  + Tablet
* Marketing Event (this text column highlights if a marketing event happened on the date that could have had a potential impact on website behavior)
* 9 Binary columns were derived from the Marketing Event column. If a “1” exists in the column then it represents that the specific type of marketing event was executed on that date, if 0 exists in the column then that type of event was not executed on that date.
  + Special Sale (conducted on a separate website)
  + Mark Downs
  + Gift with Purchase
  + Friends and Family Sale
  + Deal of the Day Sale
  + New Arrivals added on Site
  + Sale with Additional Purchase
  + Sale on Sale
  + Employee Appreciation Event
* Email Event (this text column highlights if a marketing email was delivered on this date)
* Email Day (binary column derived from Email Event column, 1 if email was executed on the date, 0 if no email was executed)
* Revenue Plan
* Traffic Plan
* Net Shipped Revenue Plan
* Orders Plan
* Units Plan
* Margin Dollar Plan
* Marketing Advertising Spend by Channel for the following marketing channels (6 columns)
  + Affiliate
  + CSE
  + Display
  + PLA
  + Paid Search
  + Paid Social
  1. *Other Potential Datasets:* In addition to the three data sources utilized, the following additional datasets could provide additional insight into the ability to reach a revenue goal:
* Inventory data
* Shopping cart details (number of carts created, number abandoned)
* Website outage data
* Site speed and loading time data

1. **Data Wrangling**

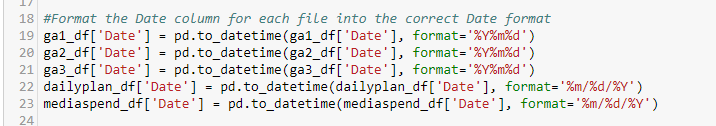
*Goal:* Import each of the five datasets into Python, combine and prepare into a final dataset for analysis and modeling.

*Overview of Tasks Performed:*

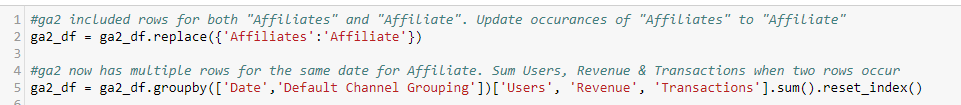
* Imported all five datasets
* Formatted the date column for each of the five files
* Updated Marketing Channel Names to single name per day
* Created a single row for each date
* Filled missing values
* Created the binary column for ‘Email Day’
* Created a new column ‘Promo Type’ and additional binary columns
* Created ‘Avg. Session Duration Seconds’ column
* Joined all five files to create a single file.
  1. *Imported all five files:*



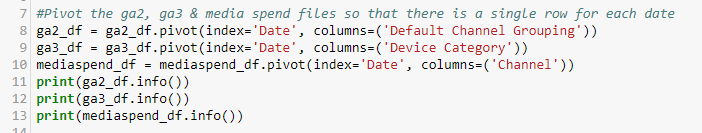
* 1. *Formatted the date column for each of the five files:*



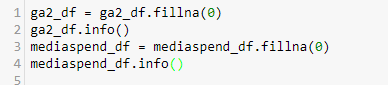
* 1. *Updated Marketing Channel Names to single name per day:* The dataset for Marketing Channels contained both “Affiliates” and “Affiliate”.
* Updated the name “Affiliates” to “Affiliate” to create a single label.
* Summed the ‘Users’, ‘Revenue’ and ‘Transactions’ columns to create a single row for each day for the marketing channel.



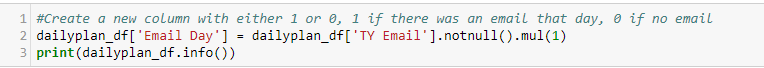
* 1. *Created a single row for each date:* Three of the datasets contained multiple rows for each date. To combine by date with the other two data files, and perform analysis on the data a pivot was applied to create a single row for each date.



* 1. *Filled missing values:* After pivoting, there were NaNs in two of the three DataFrames that were pivoted. Replaced NaN with “0”.

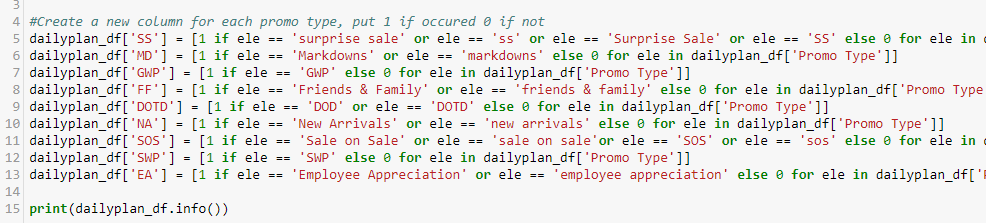


* 1. *Created the binary column for ‘Email Day’:* The existence of text in the ‘Email Event’ column indicates that at least one email campaign was executed on that day.

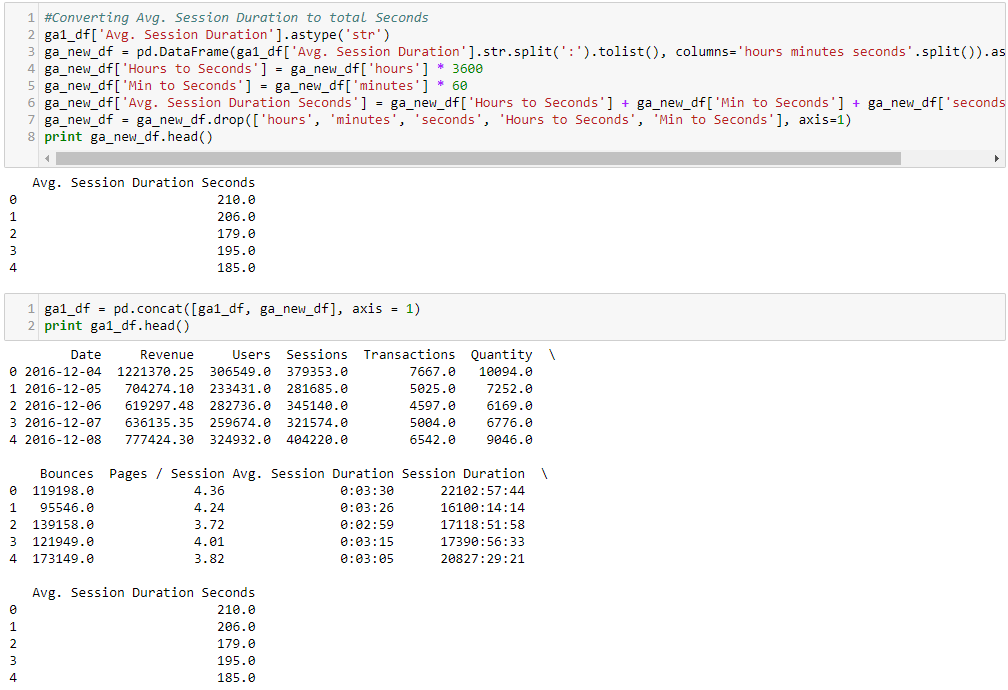


* 1. *Created a new column ‘Promo Type’ and additional binary columns:*
* Based on knowledge of the client’s promotional calendar, identified key words in the ‘Marketing Event’ column that identify if one of nine key marketing initiatives was live of that day.
* Extracted key words/phrases from the ‘Marketing Event’ column and placed into the new ‘Promo Type’ column.
* The ‘Promo Type’ column was then utilized to create nine marketing initiative binary columns that could potentially be impacting revenue (and that are taken into consideration during goal planning). Filled each of the new columns with 1 if the initiative occurred on that date and 0 if not.

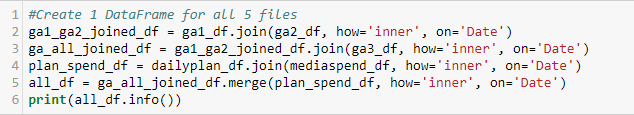




* 1. *Created ‘Avg. Session Duration Seconds’ column:*
* Created a new DataFrame and filled with three new columns created by splitting the ‘Average Session Duration’ column (0:00:00 format) into ‘hours’, ‘minutes’ and ‘seconds’.
* Created a new column for ‘Hours to Seconds’ by multiplying the ‘hours’ column by 3600.
* Created a new column for ‘Min to Seconds’ by multiplying the ‘minutes’ column by 60.
* Added the ‘Hours to Seconds’, ‘Min to Seconds’ and ‘seconds’ columns together into a new ‘Avg. Session Duration Seconds’ column.
* Dropped all of the columns from the new DataFrame except the ‘Avg. Session Duration Seconds’ column.
* Concatenated the new DataFrame to the original DataFrame.



* 1. *Joined all five files to create a single file:*



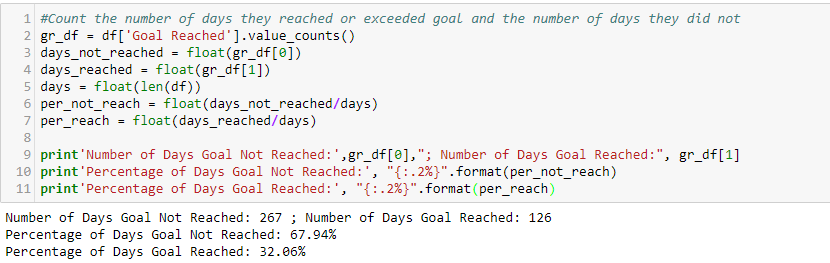
1. **EDA**

*Goal:* Gain insight into the final dataset to begin to understand the following:

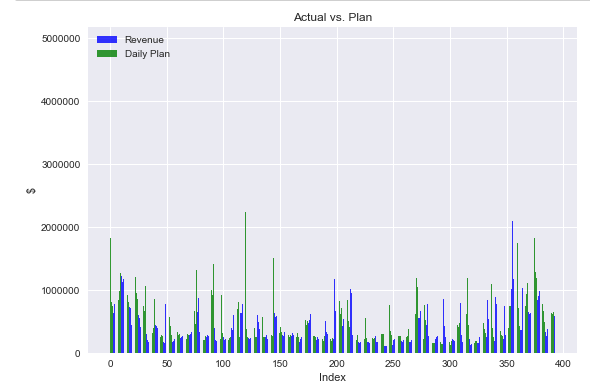
* Uncover the underlying structure of the data
* The percentage of the dates the client reached or exceeded their goal
* Identify if any outliers or anomalies exist in the data and why
* Begin to understand the type of impact (if any) that various marketing events and email campaigns may have on revenue and traffic.

*Overview of Tasks Performed:*

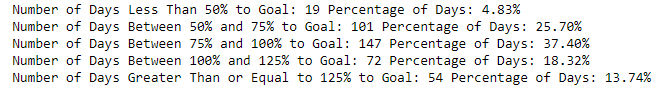
* Identified the percentage of days the client reached or exceeded goal
* Analyzed date to identify trends (highs, lows, increases, decreases, anomalies)
* Removed outlier from the final dataset
* Computed statistical analysis on the final dataset
* Explored the impact of email and marketing events on revenue
  1. *Identified the percentage of days the client reached or exceeded goal:*
* Created a new binary column with 1 if they reached or exceeded their goal and 0 if they did not reach their goal
* Created two new variables, one that counts when the new column is ‘0’ for ‘days\_not\_reached’ and the other that counts when the new column is ‘1’ for ‘days\_reached’.
* Divided each of the two variables by the total number of days for the percentages.



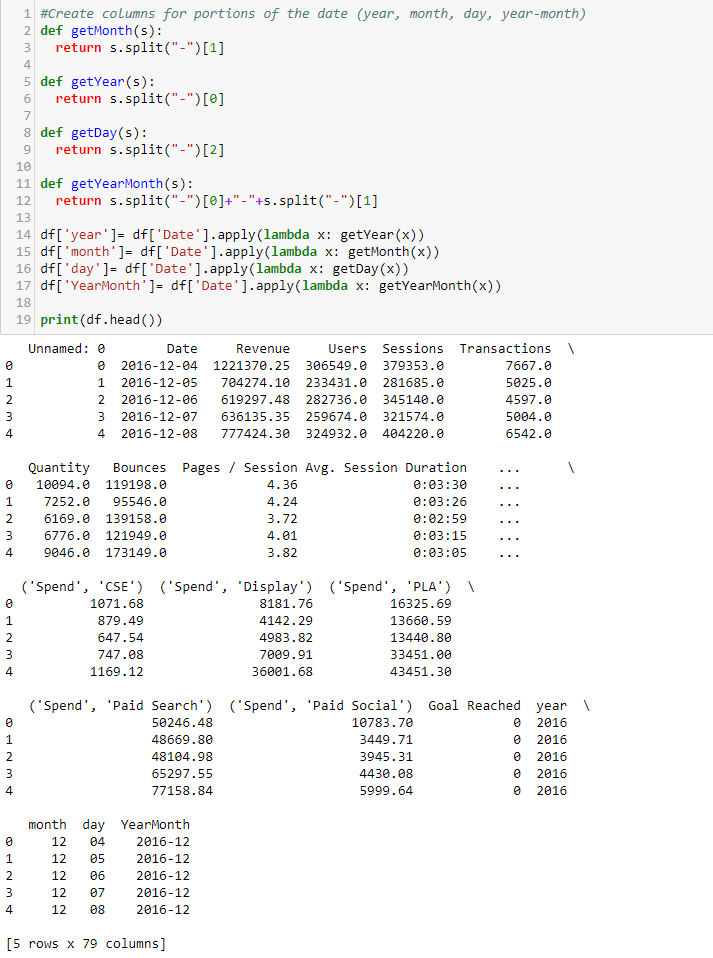
* Graph revenue compared to plan daily:

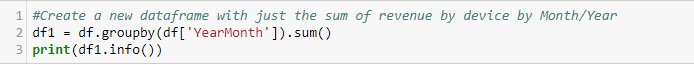


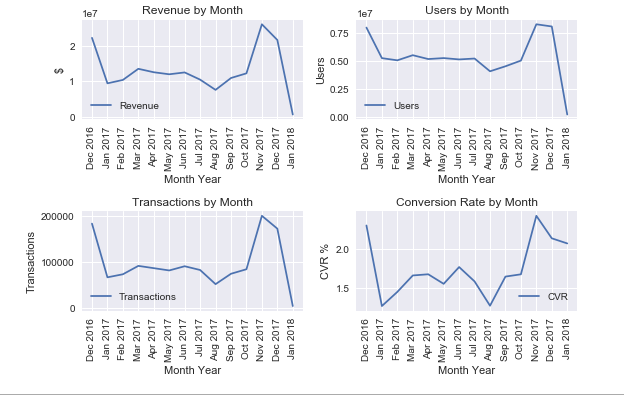
* Identify how far off the goal is from revenue on average:



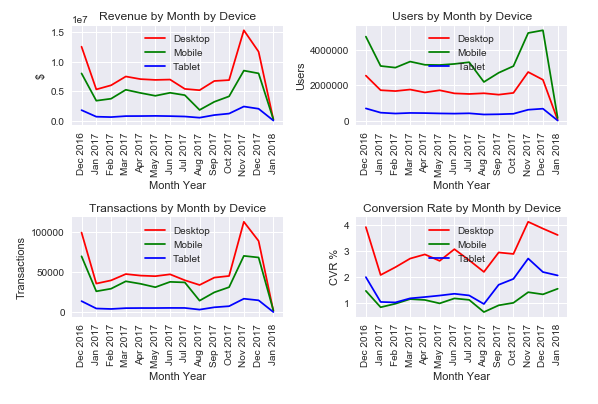
* 1. *Analyzed date to identify trends (highs, lows, increases, decreases, anomalies):*
* Split the date column into four new columns, ‘year’, ‘month’, ‘day’ and ‘YearMonth’.



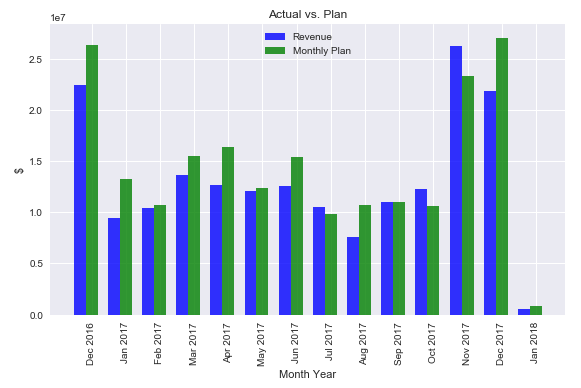
* Group and sum the data by ‘YearMonth’ column to have monthly totals for visualization by month.
* Graphed KPIs by month to identify highs, lows, possible trends.



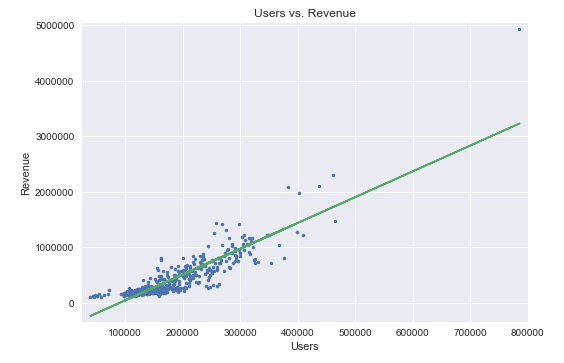
* Graphed KPIs by Device Type to understand how each device type is contributing.

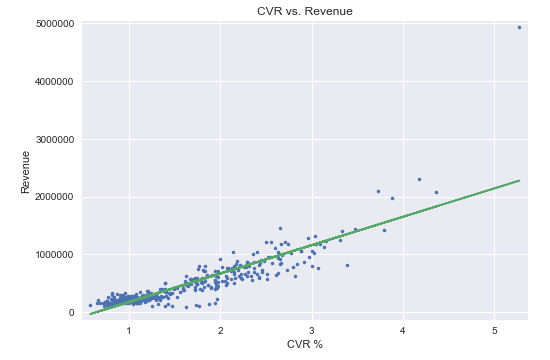


* Identify if there are any months when total monthly revenue versus total monthly planned revenue are drastically different.

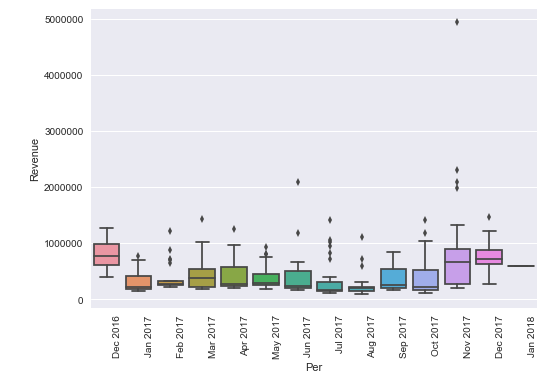


* Compare the relationship between KPIs





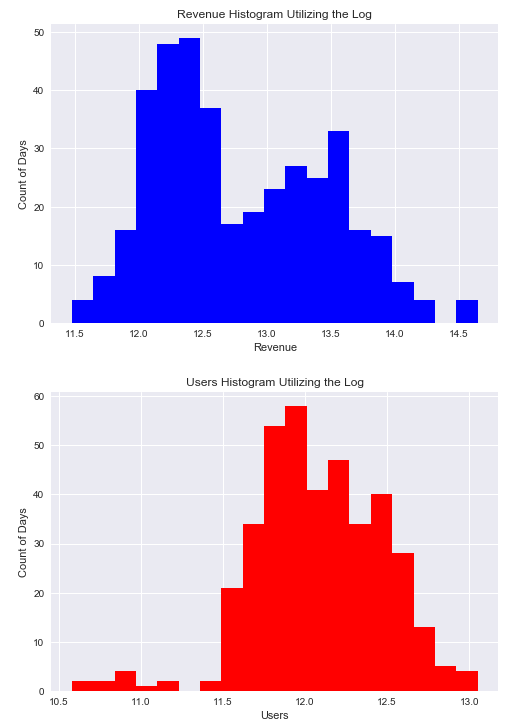
* Identify typical revenue by month and identify where outliers exist:



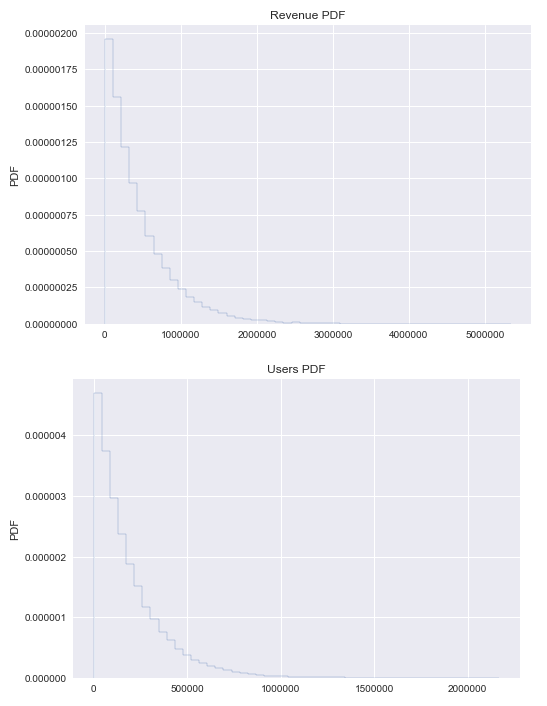
* 1. *Removed outlier from final dataset:* Dropped the outlier (Cyber Monday) from the dataset



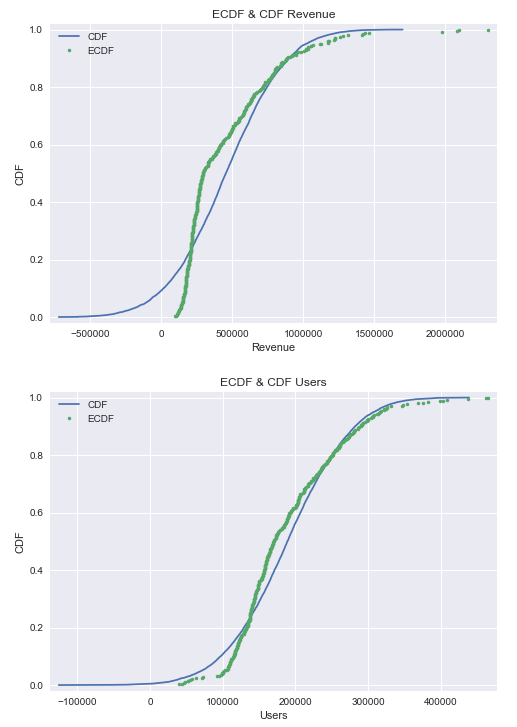
* 1. *Conducted statistical analysis on final dataset*
* Generated histograms of the log for each KPI:



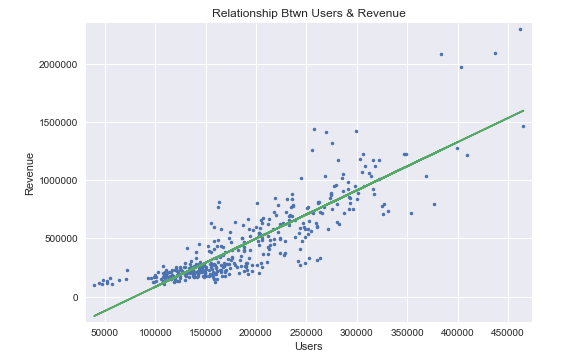
* Graphed the PDF of each KPI:

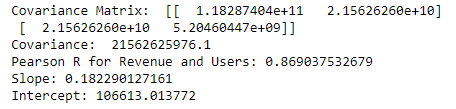


* Compute and graph the ECDF of each:

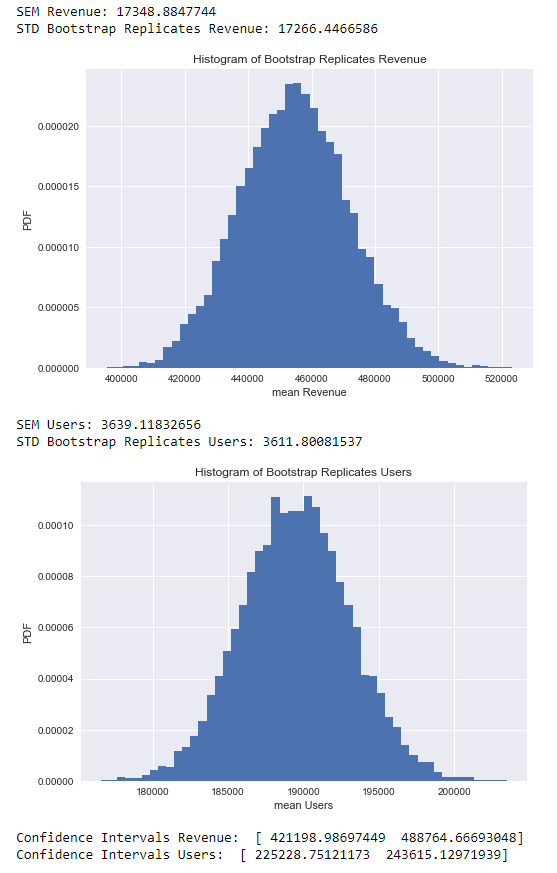


* Re-Graph the Relationship between Revenue and Users with the outlier removed and Compute the Pearson R:

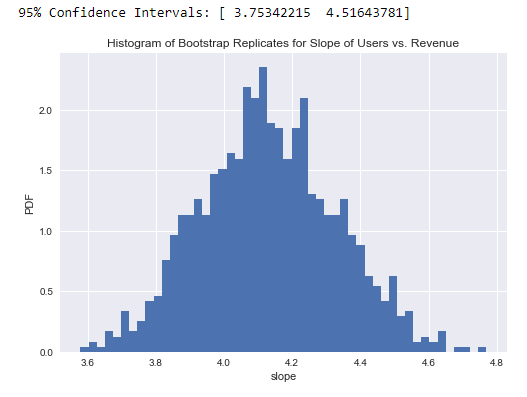


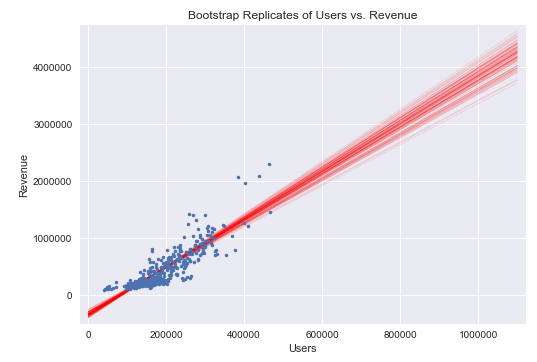


* Create Bootstrap Replicates for Revenue and Users:

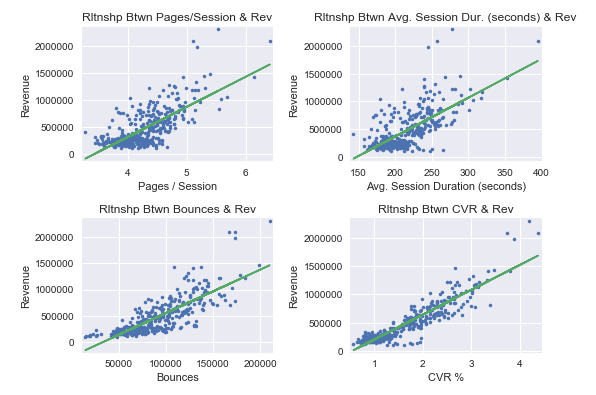


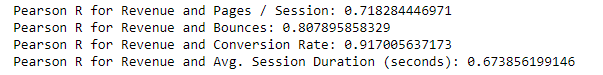
* Compare the slope of Users and Revenue using Bootstrap Replicates:



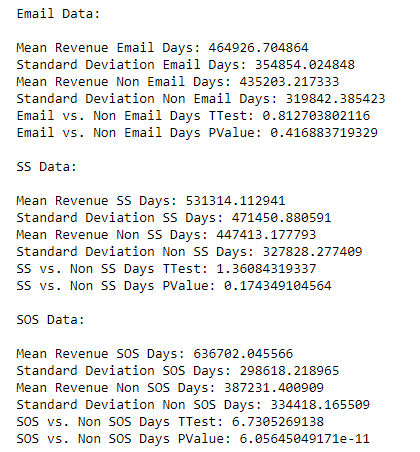


* Explored the correlation between Revenue and other KPIs (Pages/Session, Bounces, Conversion Rate, Average Session Duration (seconds)).





* 1. *Explored the impact of email and marketing events on revenue*
* Selected the top three of the ten binary columns based on the quantity of the event occurring (SS, SOS and Email Days). The rest of the binary events had 32 or fewer occurrences.



1. **Summary of Initial Findings**

During the initial analysis, it was uncovered that only 32% of the 393 days in the dataset have successfully reached or exceeded their daily plan. 5% of days reached less than 50% of the goal, 25% achieved between 50-75% of goal and 37% came closer to reaching goal (between 75%-100%). This confirms the need for more accurate forecasting, as well as indicating a need for knowing further in advance if the client is not predicted to reach goal so that they can adjust marketing efforts accordingly. In terms of reaching a monthly (looking at the calendar, not fiscal month), December and January appear to struggle the most to achieve the goal.

Initial findings indicate that there is a strong correlation between traffic, pages/session, bounces and not surprisingly conversion rate and revenue. Not as strongly correlated was the average session duration in seconds. So far, the only marketing event that had a PValue less than 0.05 was the SOS event, this indicates that this feature may be important when building the model.

1. **Notes for Further Exploration**

Additional investigation should be done regarding the channel driving the traffic to the website. This information will become important when determining how to impact a day that’s not predicted to reach its revenue goal.