Data Driven Decision Making

Google Merchandise Store Analysis and Prediction

Jovan Trajceski

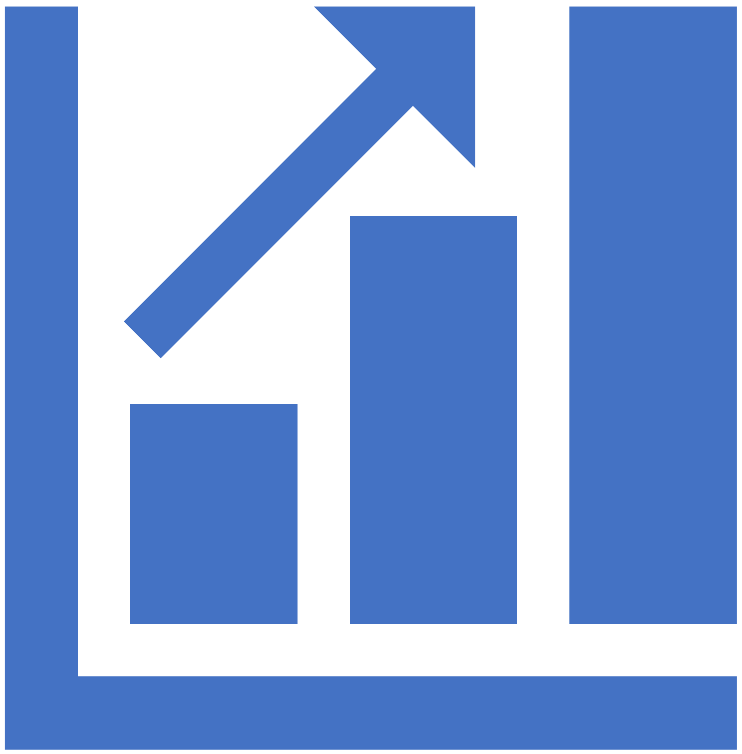
About

We offer a solution which will help you connect with your customers.

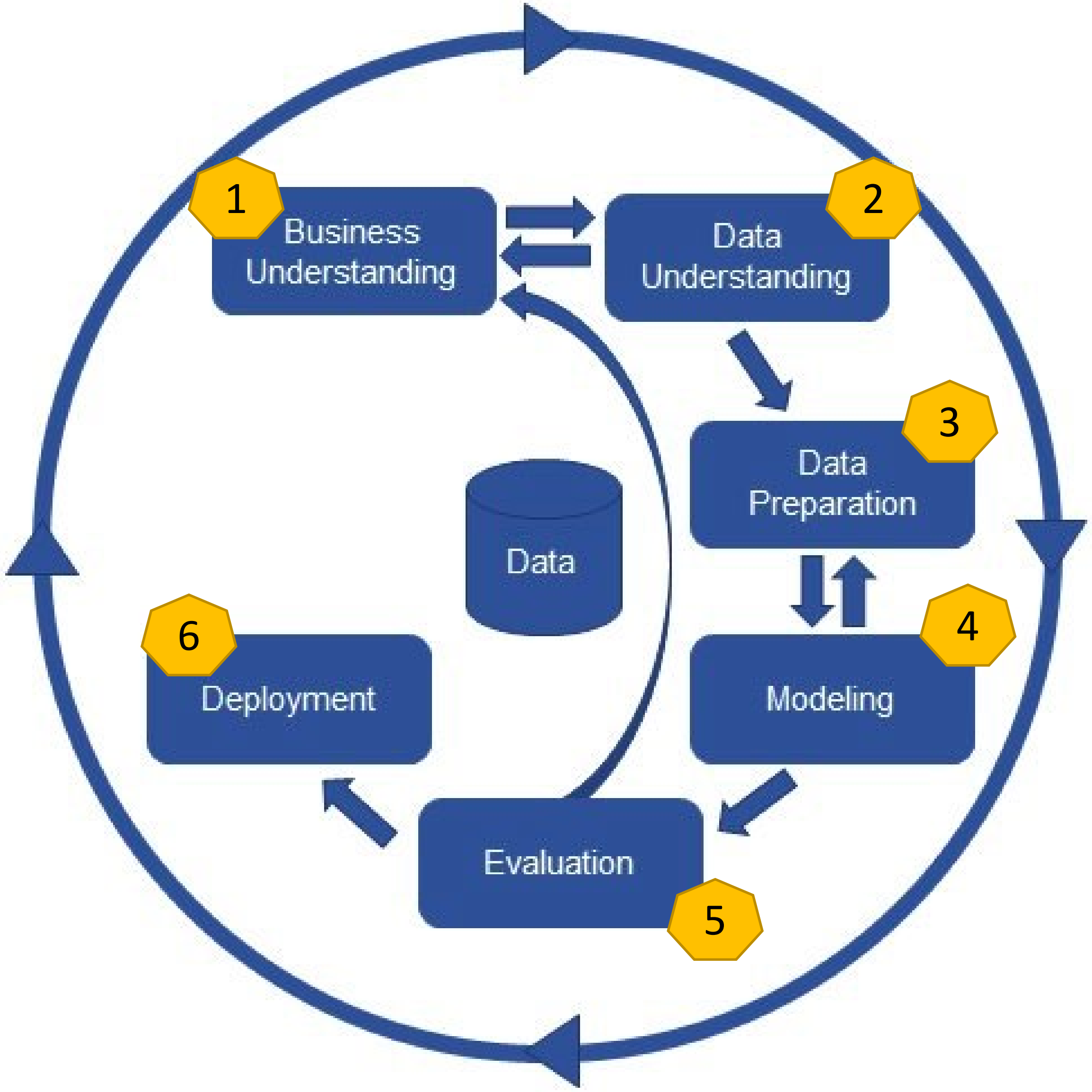
Our solution provides comprehensive understanding of the consumers and will enable you to discover important opportunities and accelerate your growth.

Problem Definition

Increase sales revenue of Google Merchandise Store by converting visitors into buyers at a higher rate by utilizing Machine Learning and Modeling.



Action Plan



1
Business Understanding
2
Data Understanding
3
Data Preparation
4
Modeling
5
Evaluation
6
Deployment
Data

E-Commerce Business – Business Understanding

- What other e-commerce companies are doing in the market to convert more customers?
  - Live commerce by KOL –TikTok
  - Simplify the payment process –Amazon, eBay, Taobao
  - Detailed product description (3D plot, video, VR, etc.) – Amazon, Taobao
  - Clearly understand and identify target client – PDD
  - Online customer service by real person – Taobao
  - Easy return policy and free shipping – Amazon
  - Excellent website UI design, mobile app and web, accurate delivery – Amazon, Taobao

- What is the reason for e-commerce not generating enough revenue?
  - Lack of product description
  - Lack of mobile platform application.
  - Complex payment process
  - Confusing website design and ambiguous price before final checkout

Save Time!

Data Description

Target Variable: Transaction Revenue

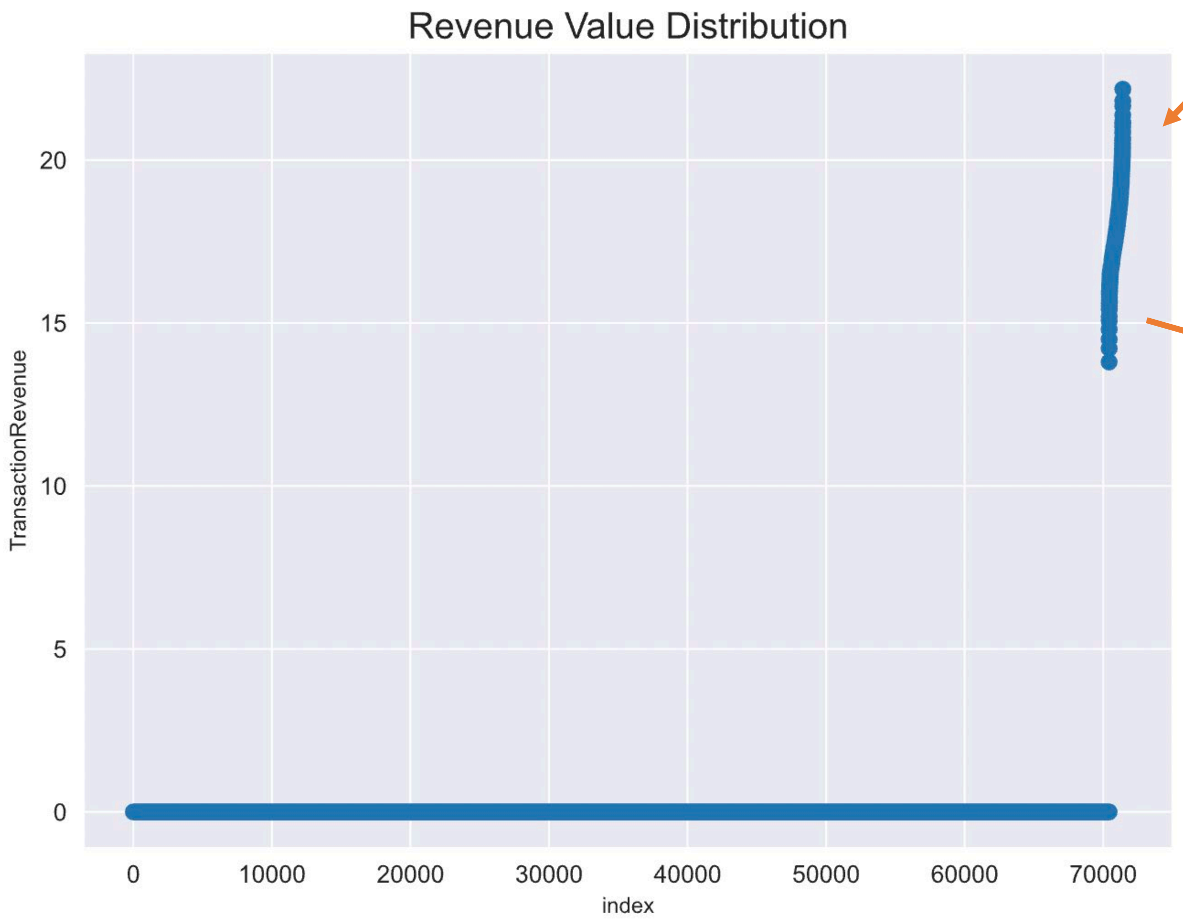
Feature Examples:

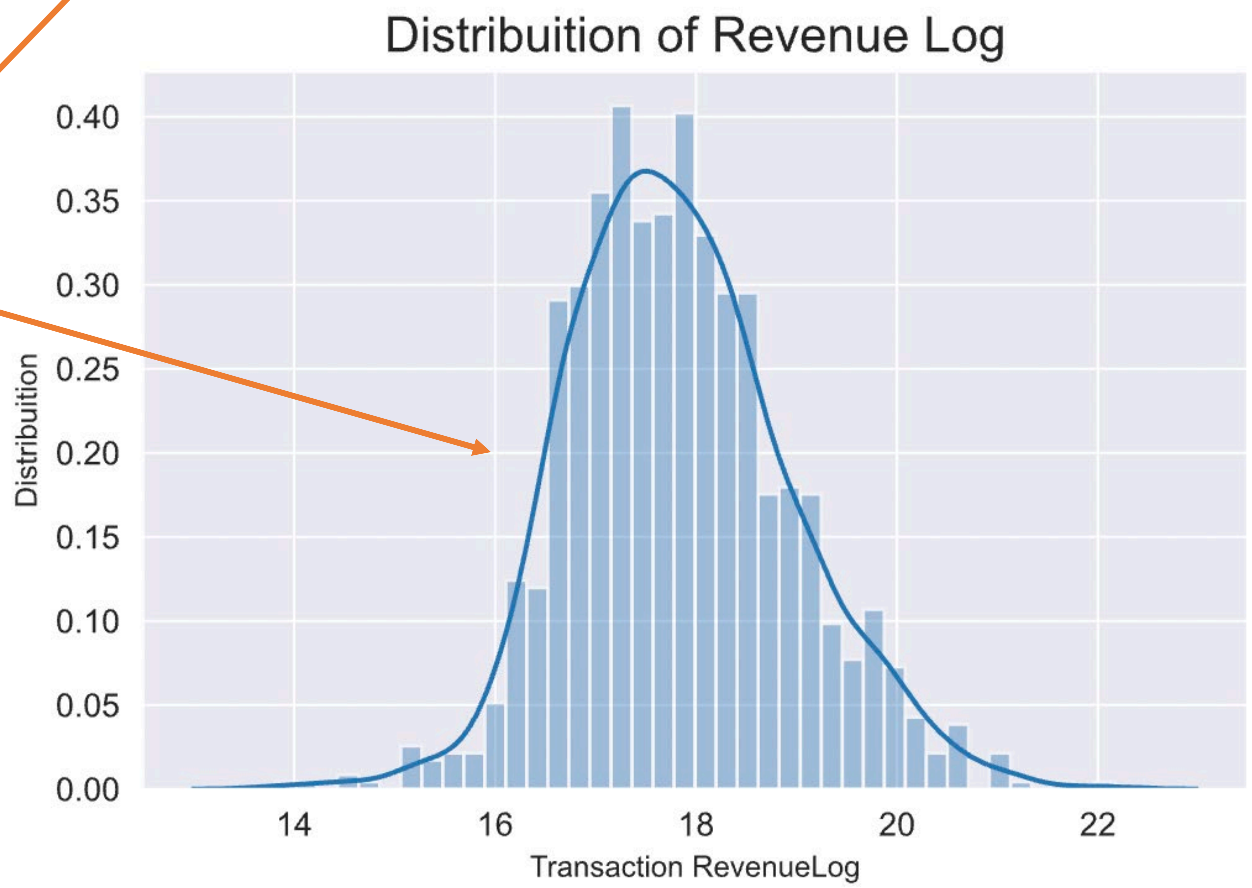
- totals.hits: Total number of hits within the session
- totals.visits: The number of sessions. This value is 1 for sessions with interaction events. The value is null if there are no interaction events in the session.
- channelGroupping: The Default Channel Group associated with an end user's session for this View.
- visitStartTime: The timestamp (expressed as POSIX time). We defined it as 'VisitHour' to extract only the 'Hour'
- date: The date of the session in YYYYMMDD format. We divided it into separate features: Year, months, day, weekday
- device.browser: the browser used (e.g., "Chrome" or "Firefox")
- device: This section contains information about the user devices.
- device.operatingSystem: The operating system of the device (e.g. "Macintosh" or "Windows")
- device.OperatingSystemVersion: The version of the operating system.

Findings Summary

Target Variable: Total Revenue

Less than 20% of customers generate revenue

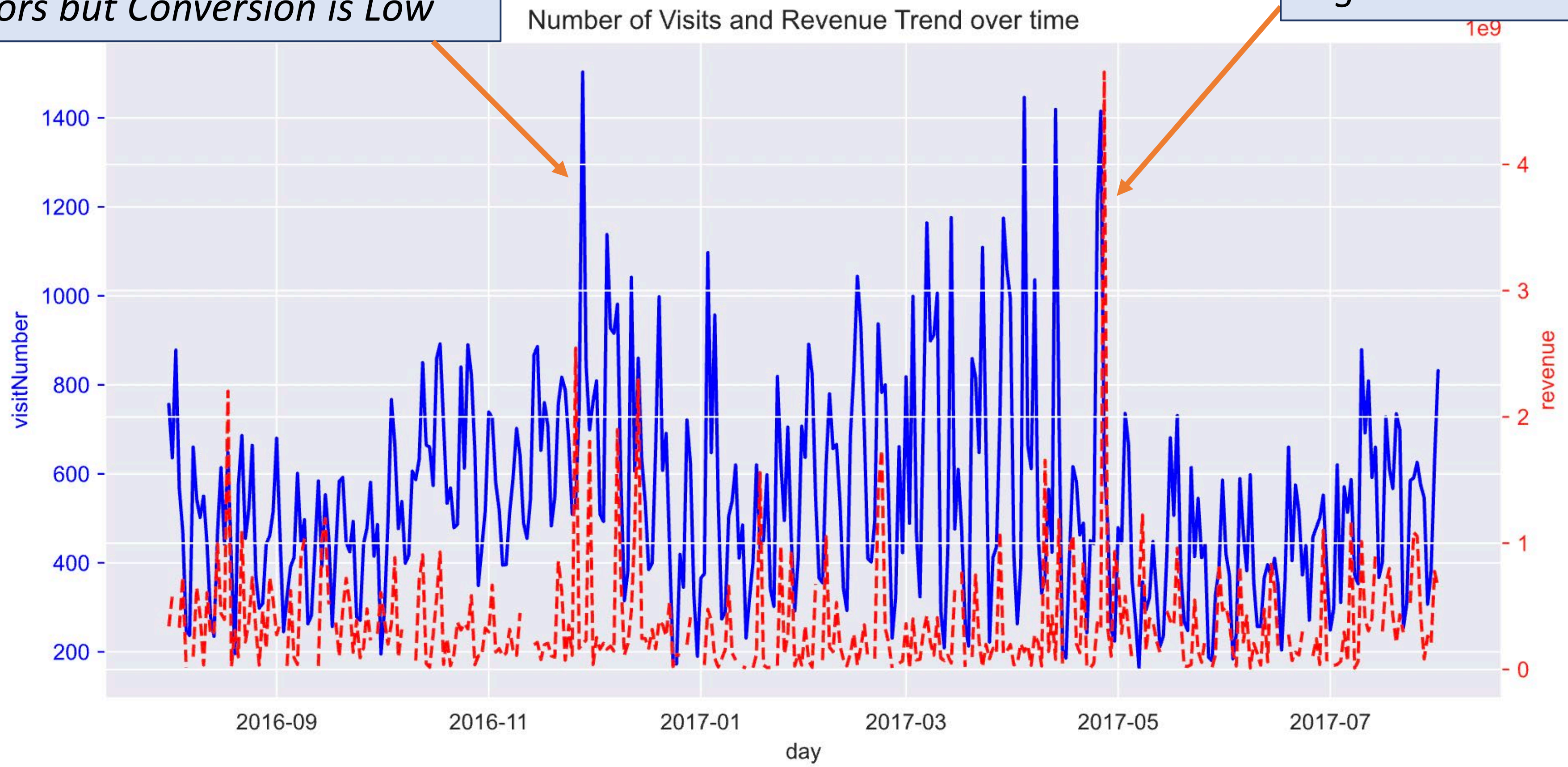




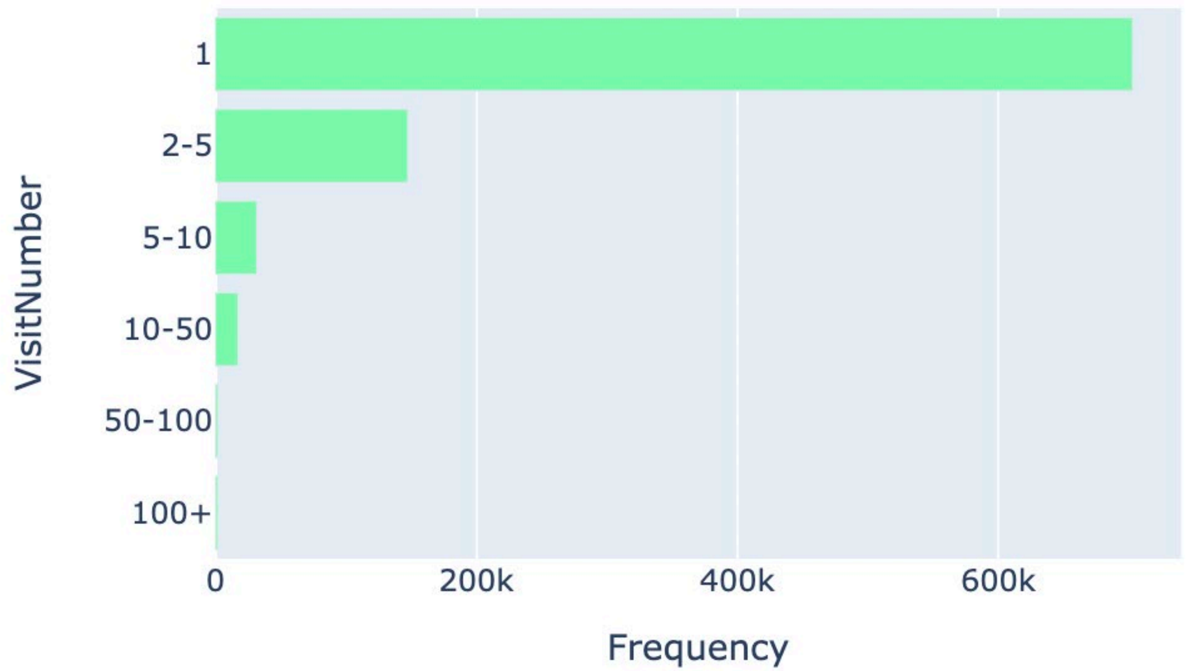
Number of Visits vs Generated Revenue

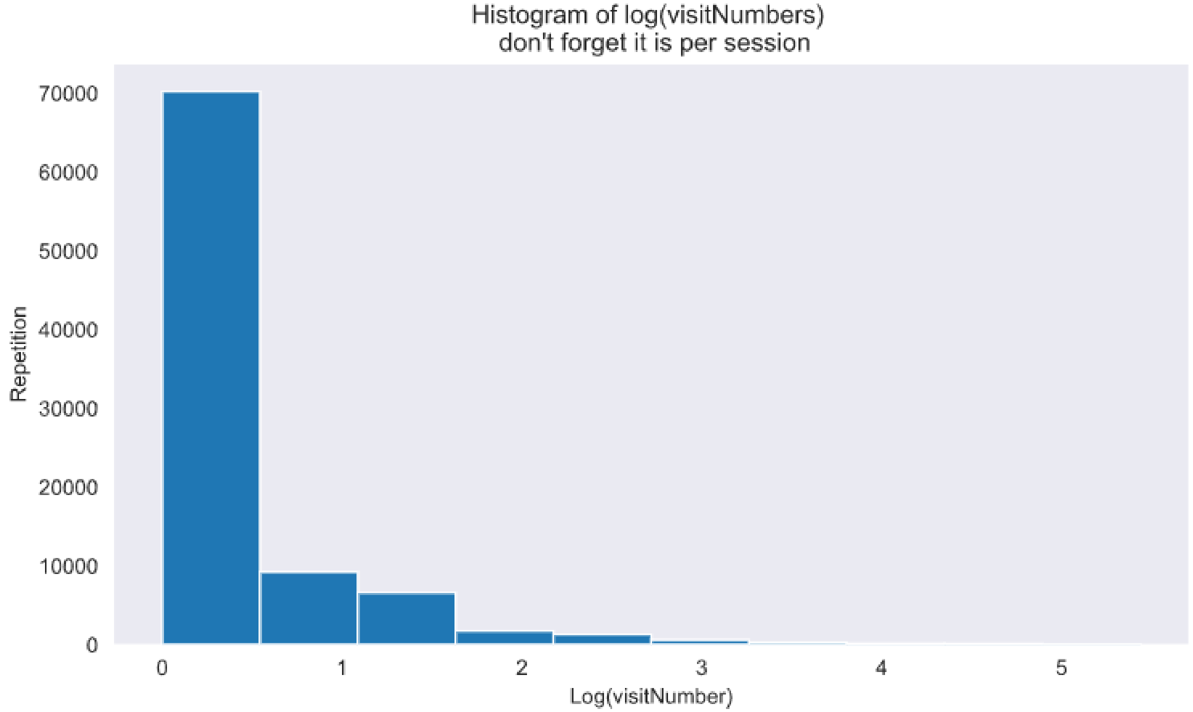
A lot of Visitors but Conversion is Low

Highest Revenue Generated



Visit Frequency per User



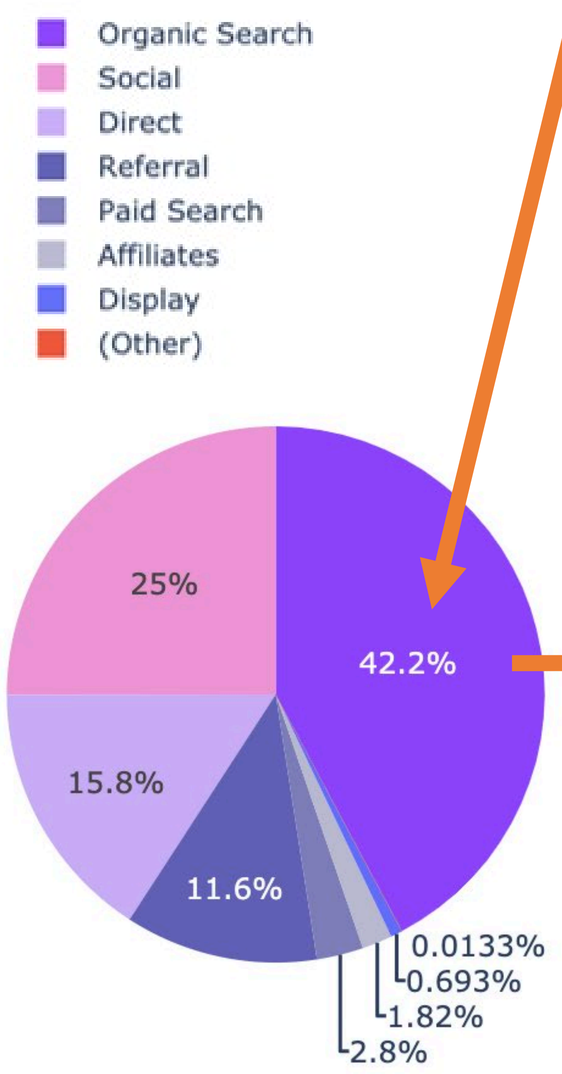


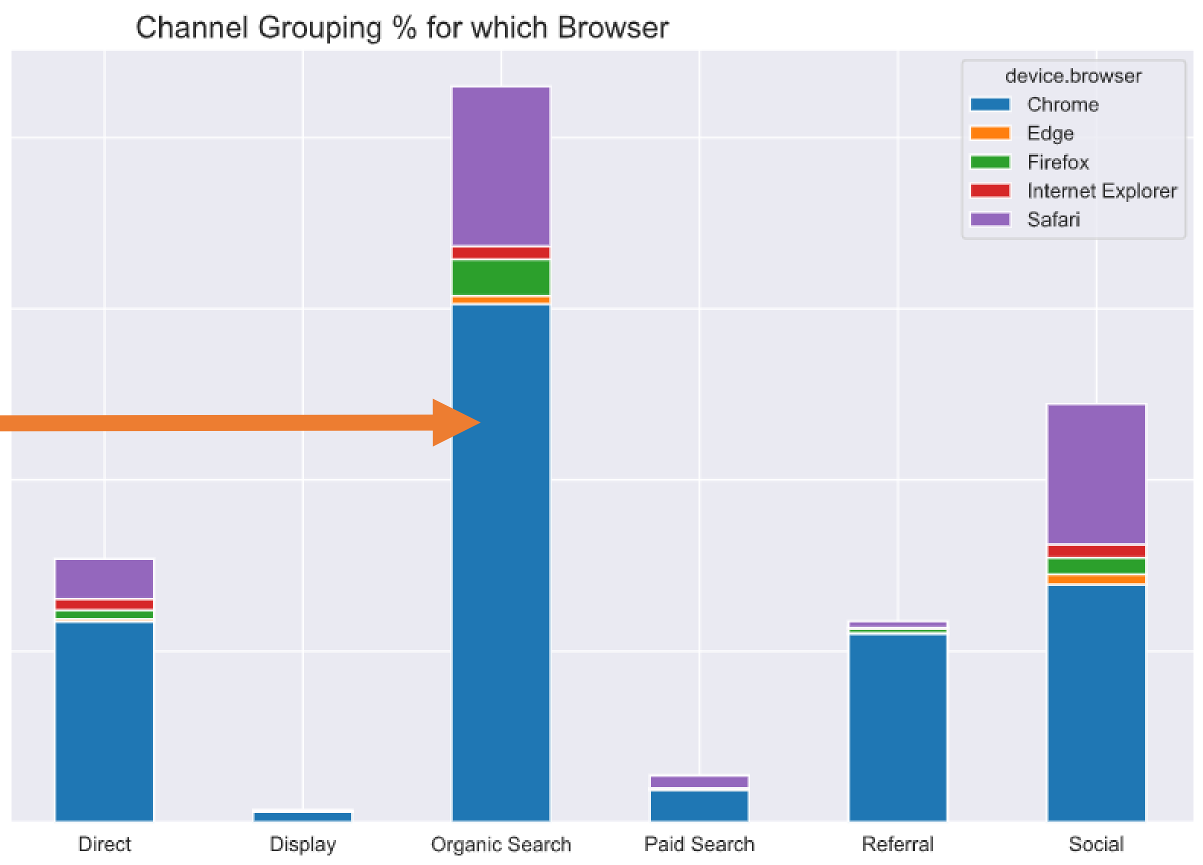
Only 12 % of users are repetitive users

80 percent of sessions have visitNumber lower than 2.0 times

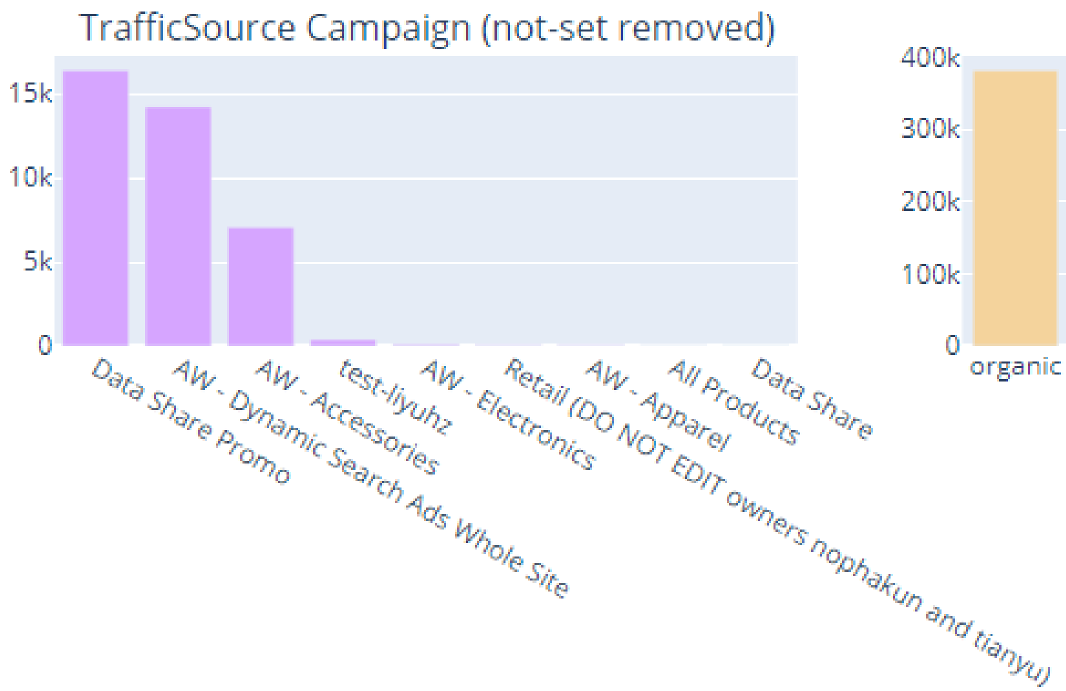
Channel Grouping

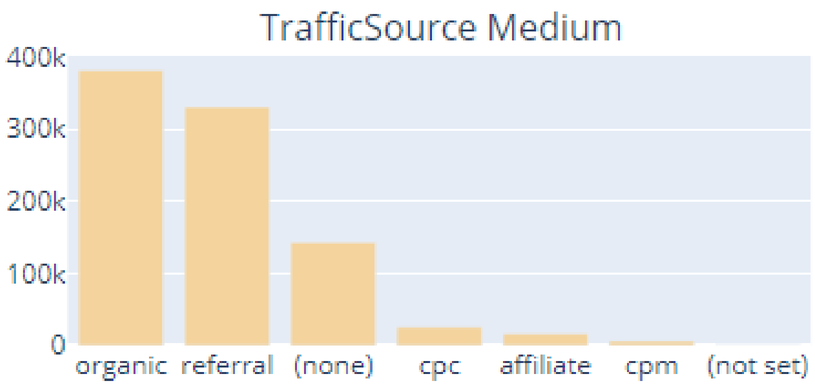
Most of our visitors find us through Organic Search

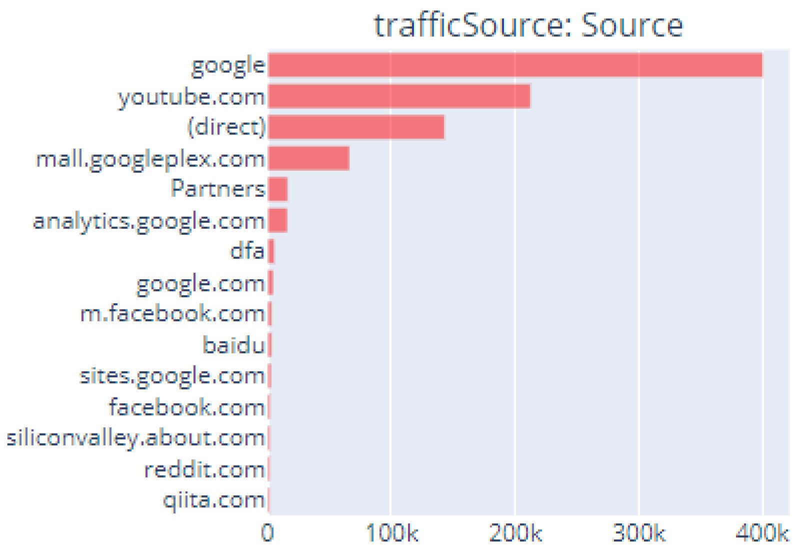




Traffic Source



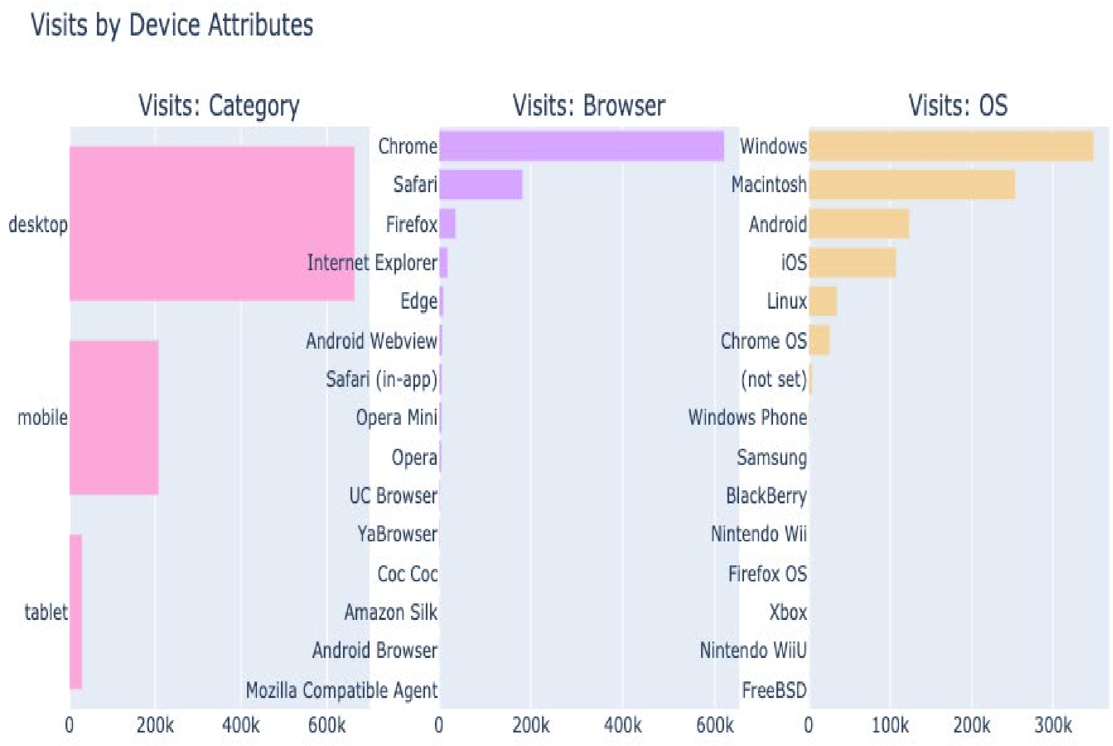


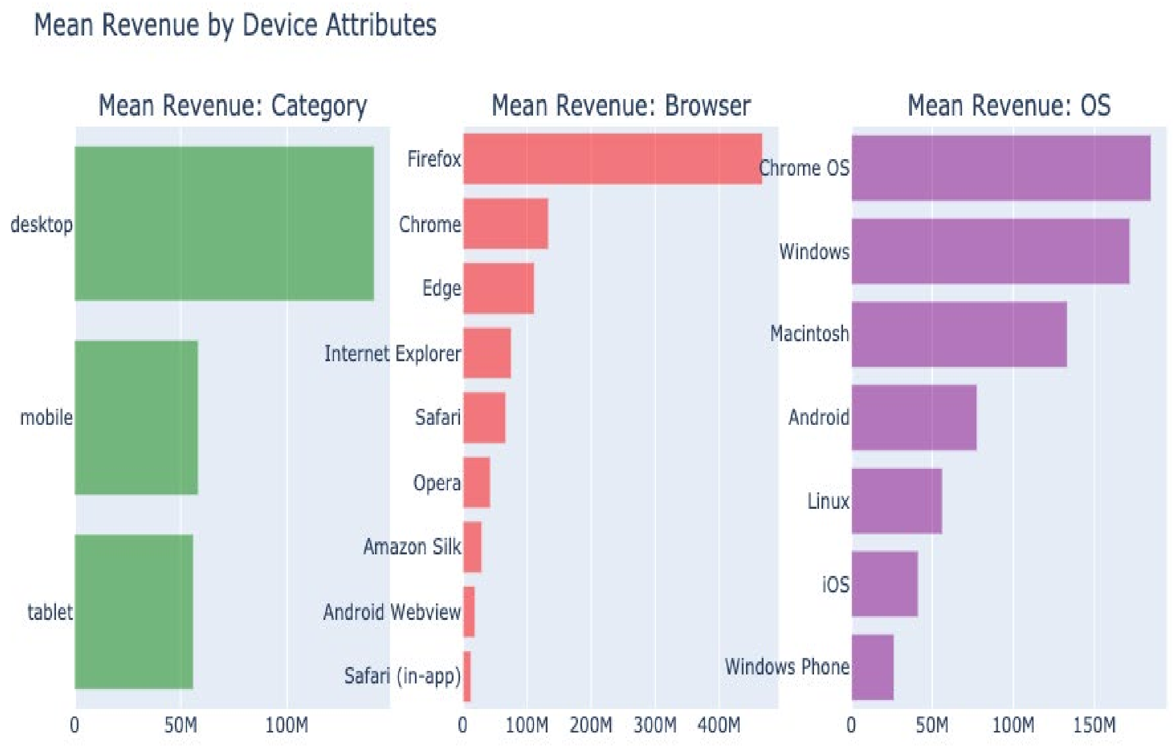


Revenue by Device Attributes

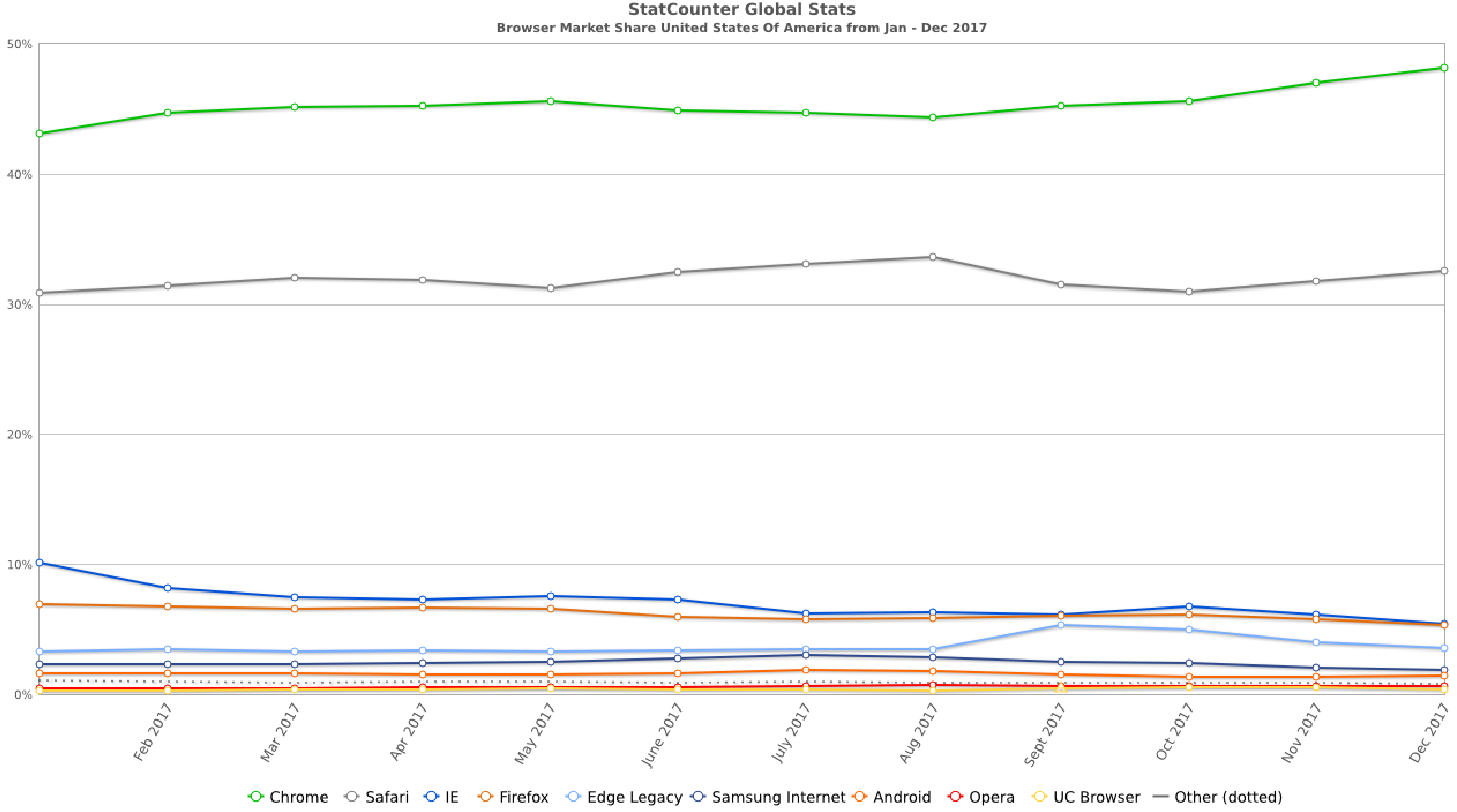
Most users navigate the store via desktop

Highest Mean Revenue from Firefox Users





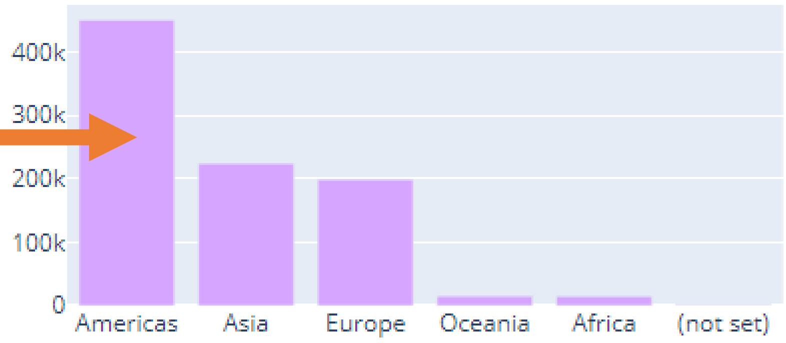
Popularity of the Browsers over years



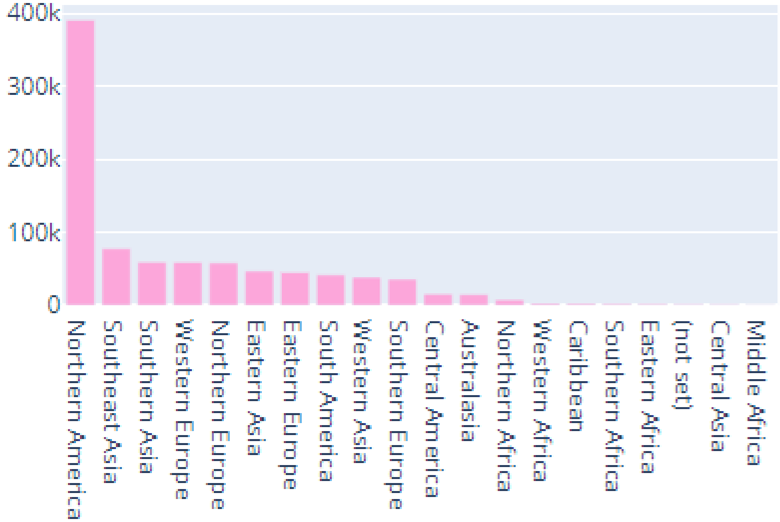
Revenue and Visits by Continent

Most Visitors are coming from America

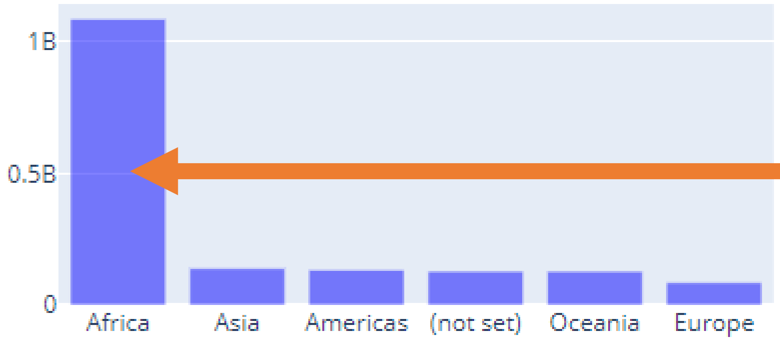
Visits : GeoNetwork Continent



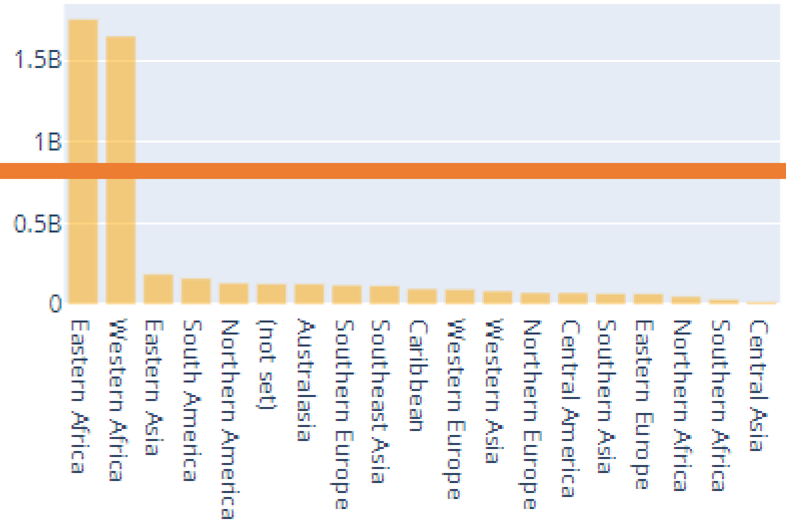
Visits : GeoNetwork subContinent



Mean Revenue: Continent



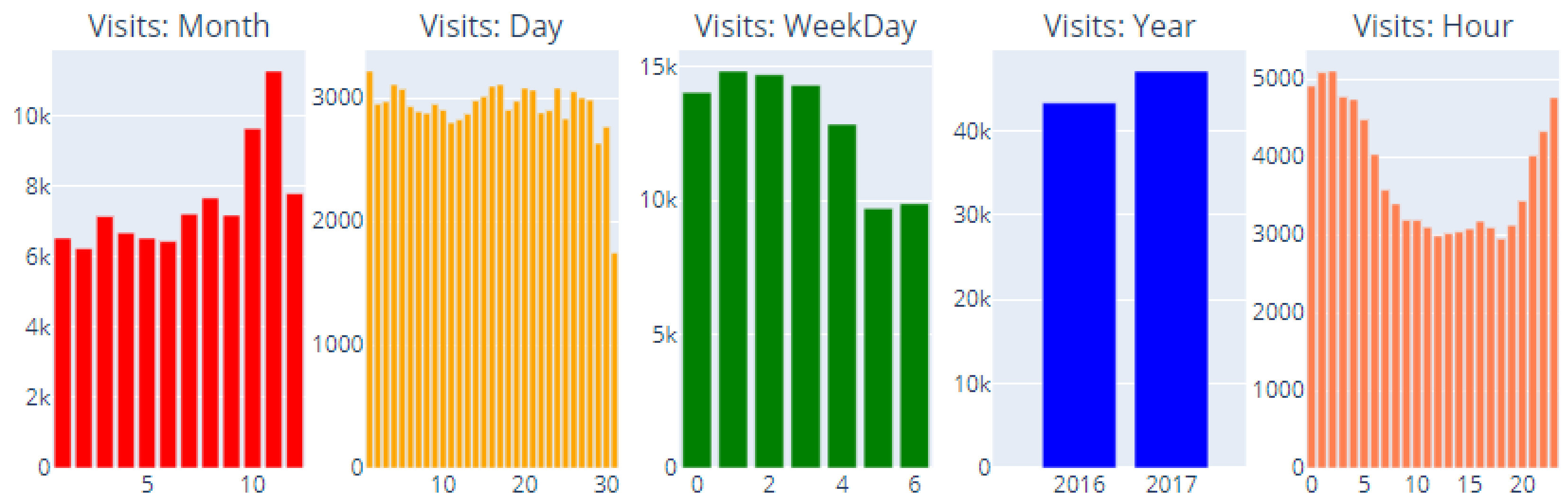
Mean Revenue: SubContinent



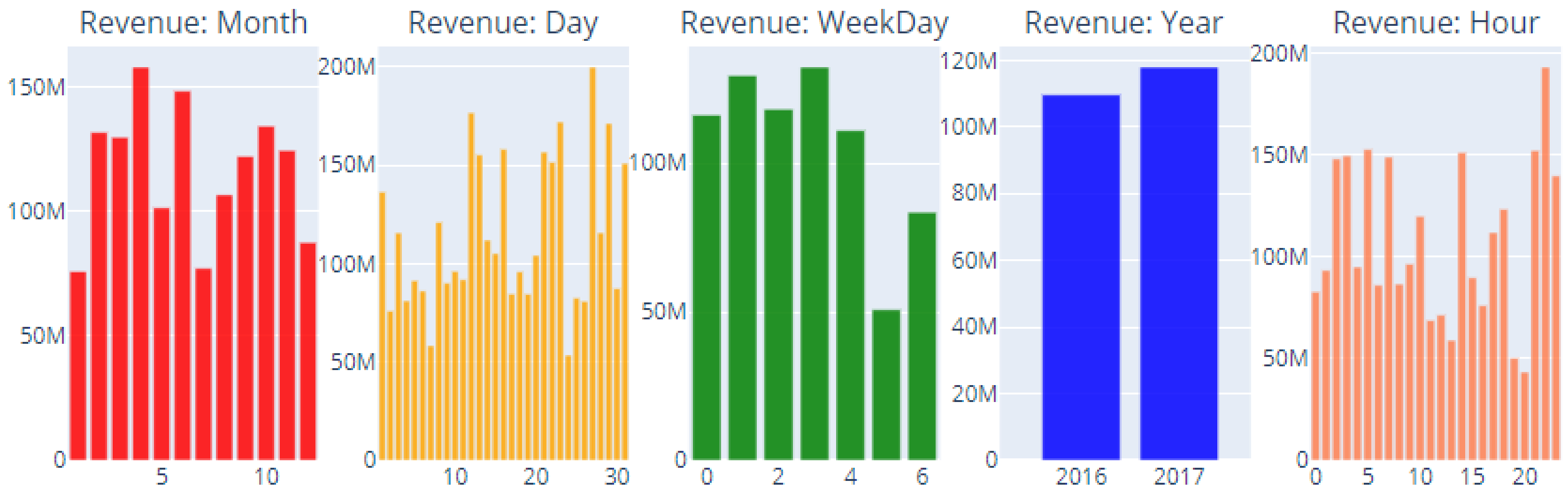
Highest mean revenue comes from Africa (why?)

Visits and Revenue by Month, Weekday, Day

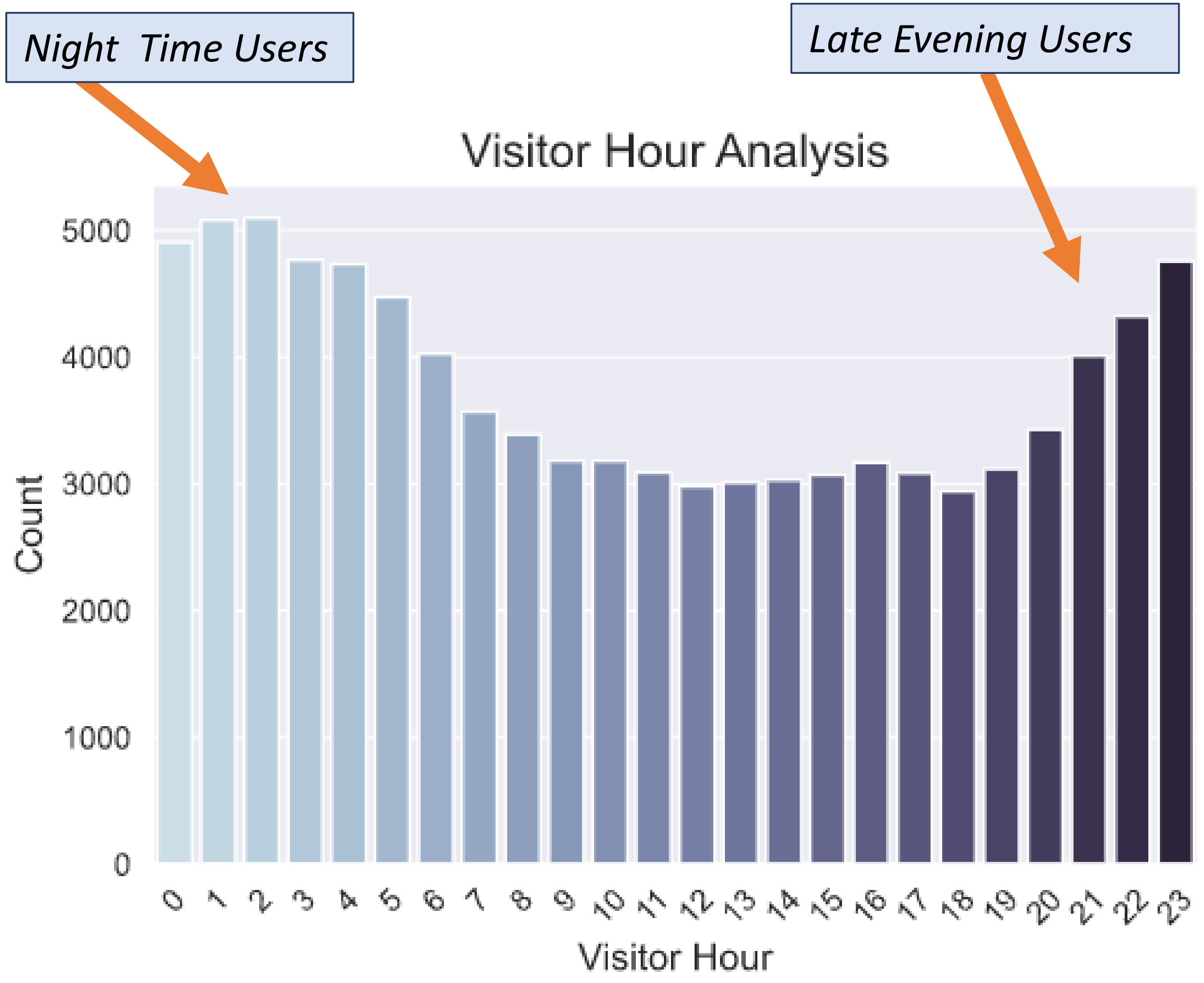
Highest Number of Visits: November and Tuesday

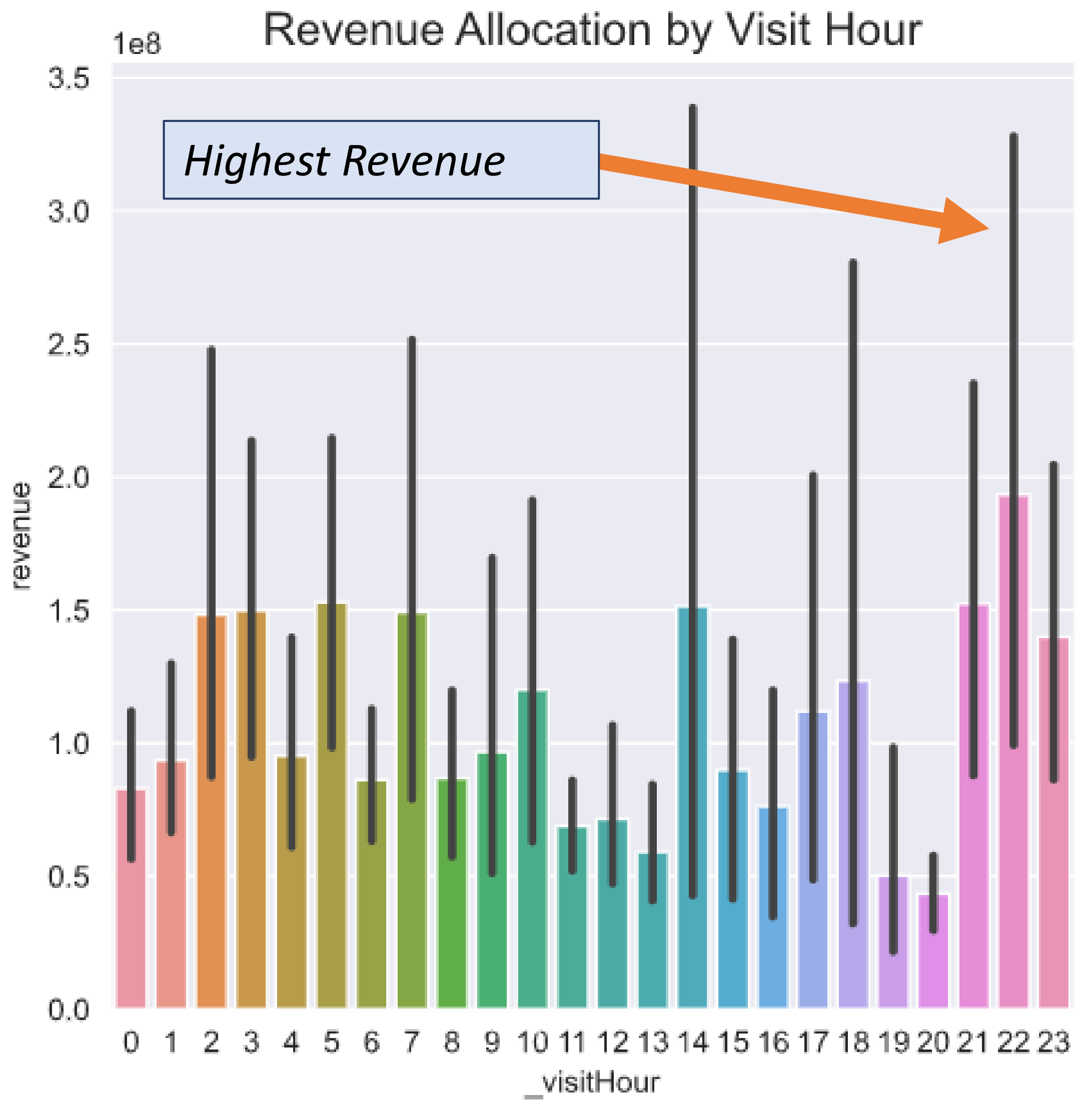


Highest Mean Revenue: April, End of the Month, Thursdays, 10PM

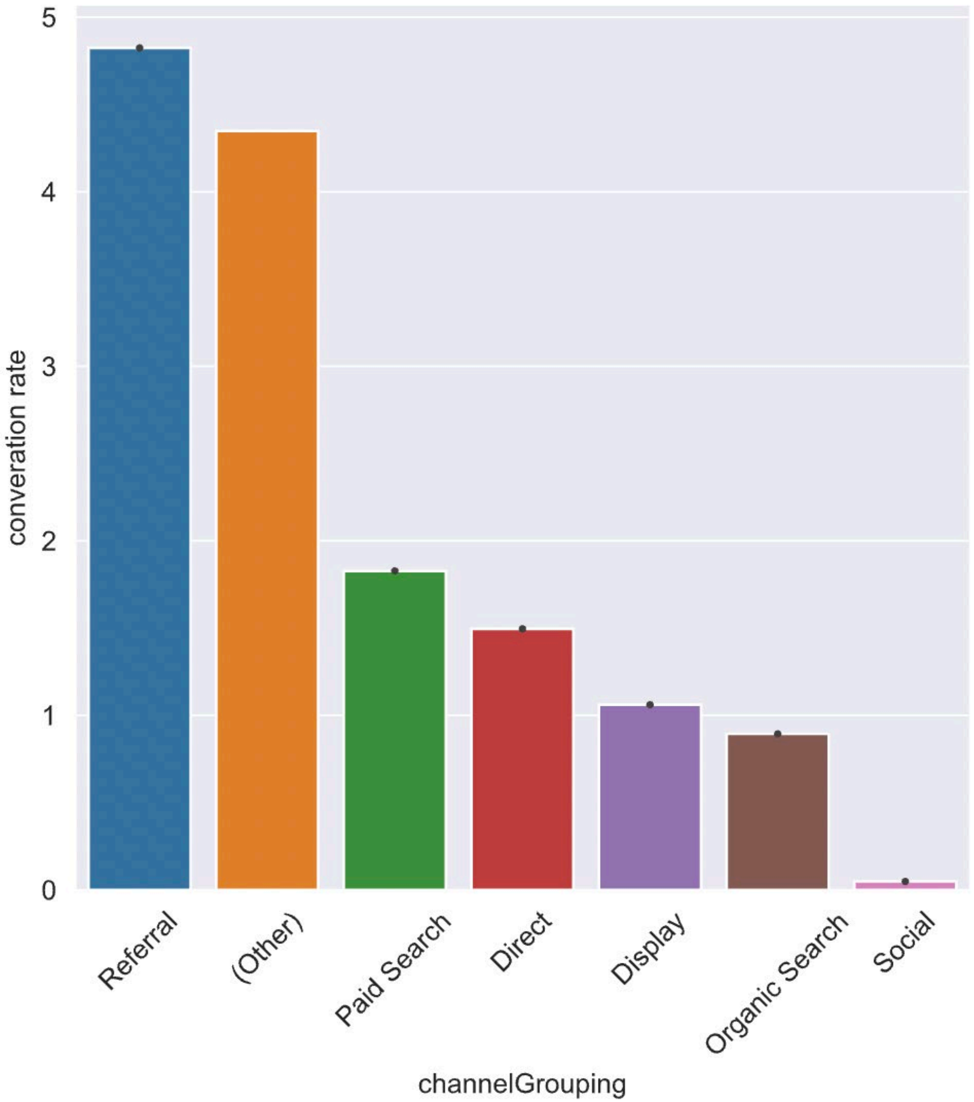


Visit Hour and Revenue





Conversion Rate



Referral has the highest conversation rate

Channel	Conversion Rate
Referral	5.1
Paid Search	1.9
Direct	1.32
Display	1.56
Organic Search	0.89
Social	0.04

Data Cleaning and Handling NA's

NAs

```
totals.visits                                          0
totals.hits                                            0
totals.pageviews                                     100
totals.bounces                                    453023
totals.newVisits                                  200593
totals.transactionRevenue                         892138
trafficSource.campaign                                 0
trafficSource.source                                   0
trafficSource.medium                                   0
trafficSource.keyword                             502929
trafficSource.adwordsClickInfo.criteriaParameters      0
trafficSource.isTrueDirect                        629648
trafficSource.referralPath                        572712
trafficSource.adwordsClickInfo.page               882193
trafficSource.adwordsClickInfo.slot               882193
trafficSource.adwordsClickInfo.gclId              882092
trafficSource.adwordsClickInfo.adNetworkType      882193
trafficSource.adwordsClickInfo.isVideoAd          882193
trafficSource.adContent                           892707
trafficSource.campaignCode                        903652
dtype: int64
```

totals.pageviews: replace N/A with 0

totals.transactionRevenue: replace N/A with 0

totals.bounces: replace N/A with 0

totals.newVisits: replace N/A with 0

trafficSource.keyword: drop it

trafficSource.isTrueDirect: enconding first, replace N/A with 0

trafficSource.referralPath: we might drop it

trafficSource.adwordsClickInfo.page: we might drop it

trafficSource.adwordsClickInfo.slot: replace N/A with blank

trafficSource.adwordsClickInfo.gclId: we might drop it

trafficSource.adwordsClickInfo.adNetworkType: we might drop it

trafficSource.adwordsClickInfo.isVideoAd: we might drop it

trafficSource.adContent: we might drop it

trafficSource.campaignCode: we might drop it

Modeling

Base Model: Multiple Linear Regression

- Easy to interpretate when there are few variables present
- Not good at handling multicollinearity
- Cannot handle second-hand data well

Final Model: Light GBM

- Faster training speed and higher efficiency
- Better accuracy than any other boosting algorithm
- Compatibility with Large Datasets
- Parallel learning supported

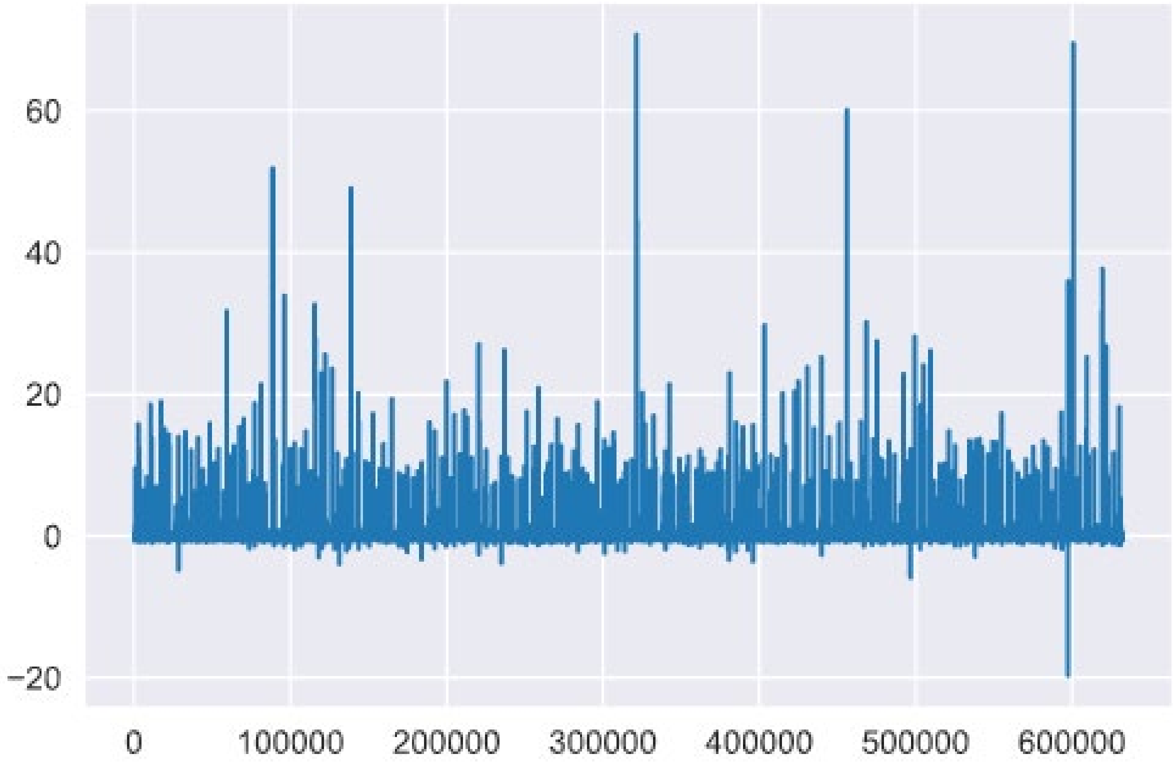
Base Model Evaluation

OLS Regression Results

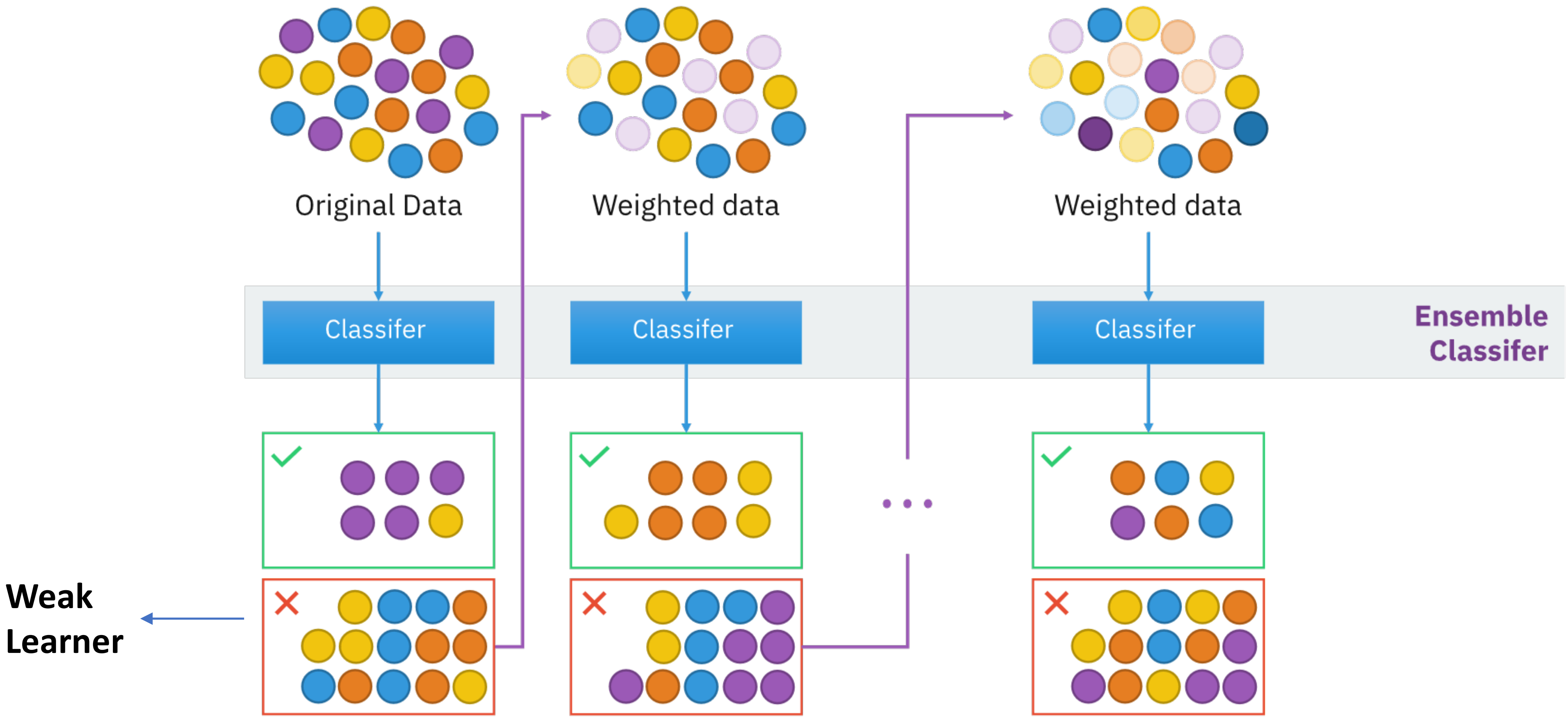
Dep. Variable:	0	R-squared (uncentered):	0.192
Model:	OLS	Adj. R-squared (uncentered):	0.192
Method:	Least Squares	F-statistic:	5797.
Date:	Thu, 15 Apr 2021	Prob (F-statistic):	0.00
Time:	19:48:52	Log-Likelihood:	-1.2715e+06
No. Observations:	632557	AIC:	2.543e+06
Df Residuals:	632531	BIC:	2.543e+06
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
device.isMobile	-0.1492	0.016	-9.119	0.000	-0.181	-0.117
totals.hits	-0.1093	0.001	-83.833	0.000	-0.112	-0.107
totals.pageviews	0.2674	0.002	146.845	0.000	0.264	0.271
totals.bounces	0.3238	0.005	63.435	0.000	0.314	0.334
totals.newVisits	-0.1876	0.009	-20.925	0.000	-0.205	-0.170
trafficSource.isTrueDirect	0.1518	0.010	14.475	0.000	0.131	0.172
_weekday	-0.0032	0.001	-2.617	0.009	-0.006	-0.001
_day	-0.0003	0.000	-1.136	0.256	-0.001	0.000
_month	-0.0042	0.001	-6.214	0.000	-0.006	-0.003
_year	-0.0002	1.23e-05	-15.476	0.000	-0.000	-0.000
_visitHour	-0.0019	0.000	-6.082	0.000	-0.003	-0.001

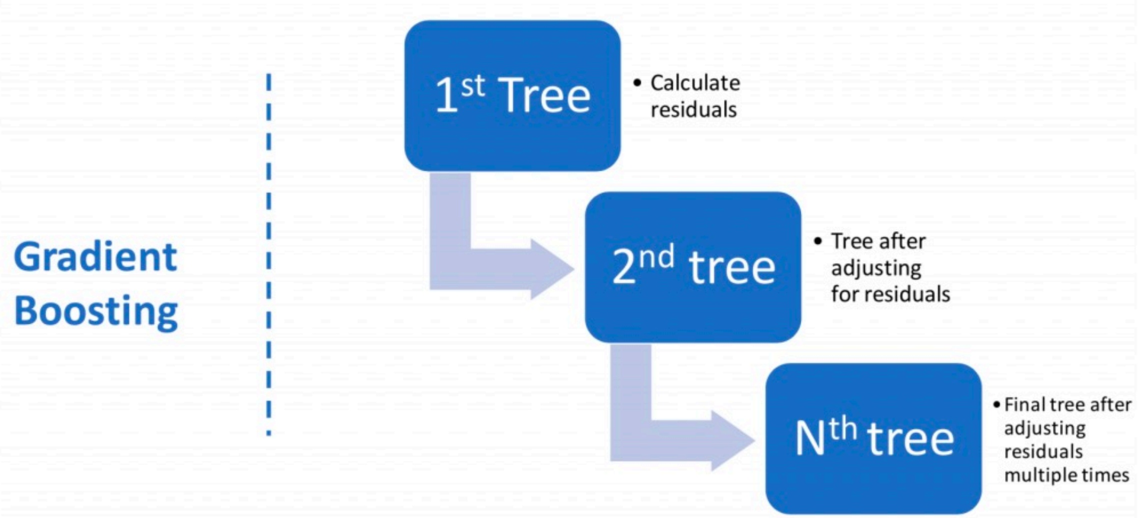
High Variance between Prediction and Actual

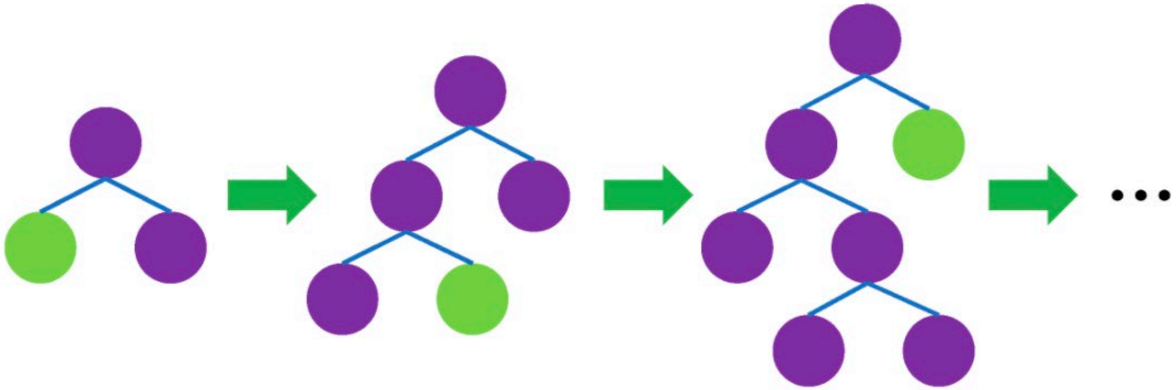


Boosting Algorithm



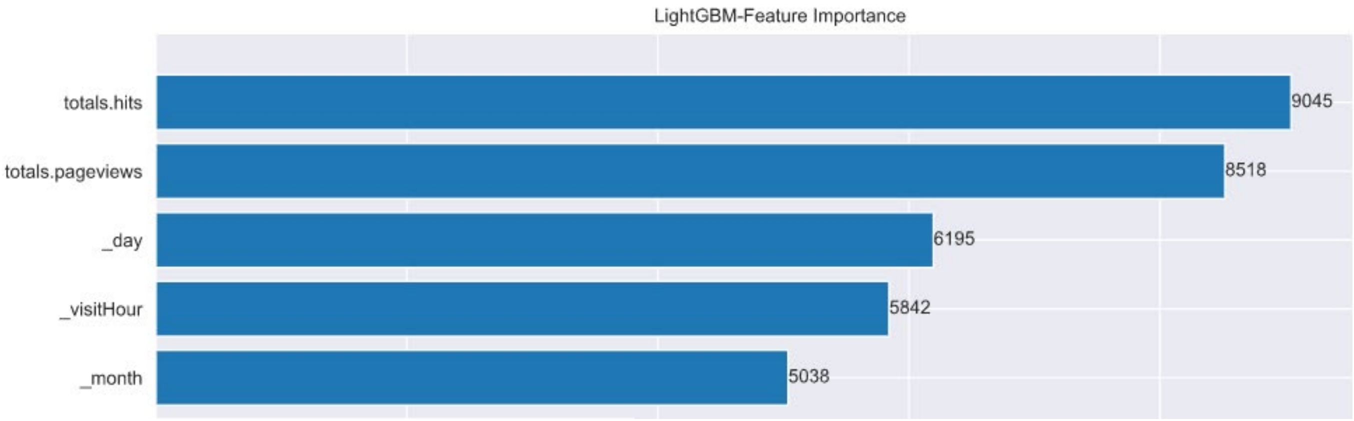
Light GBM





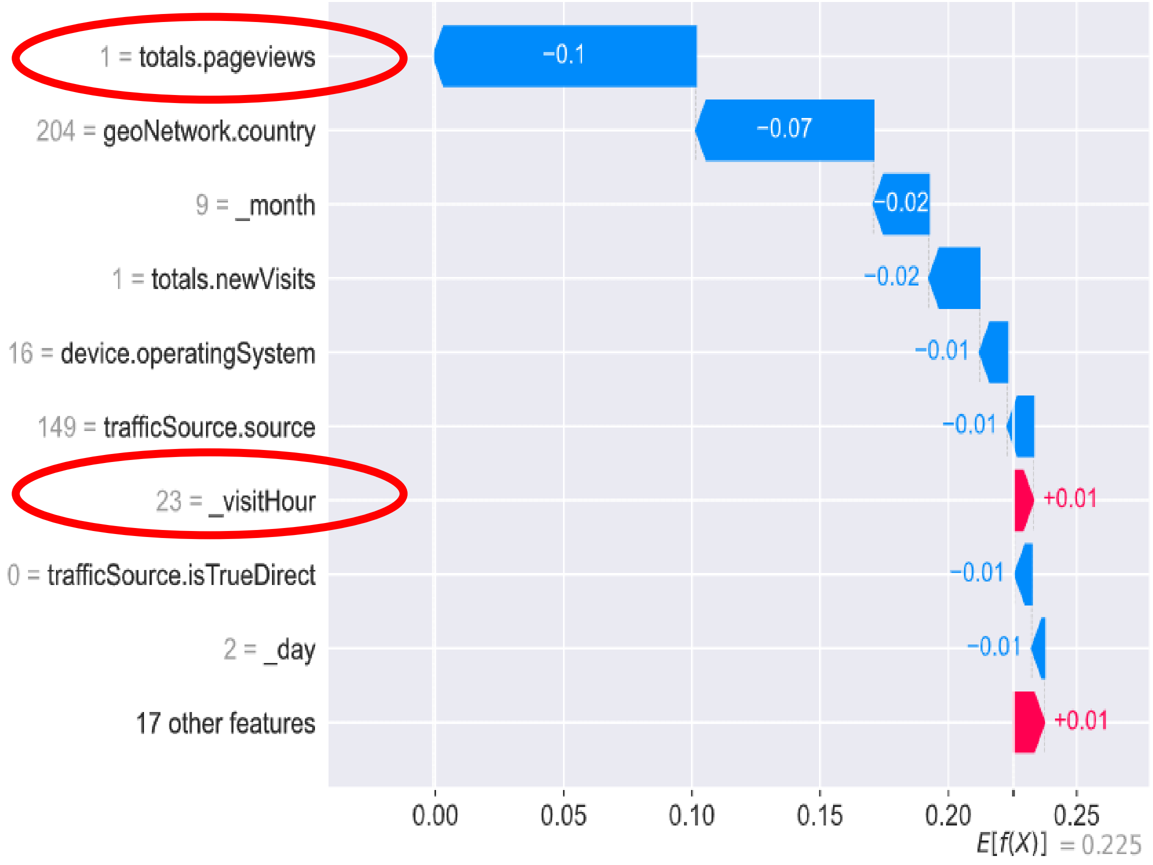
Leaf-wise tree growth

Evaluation of Light GBM

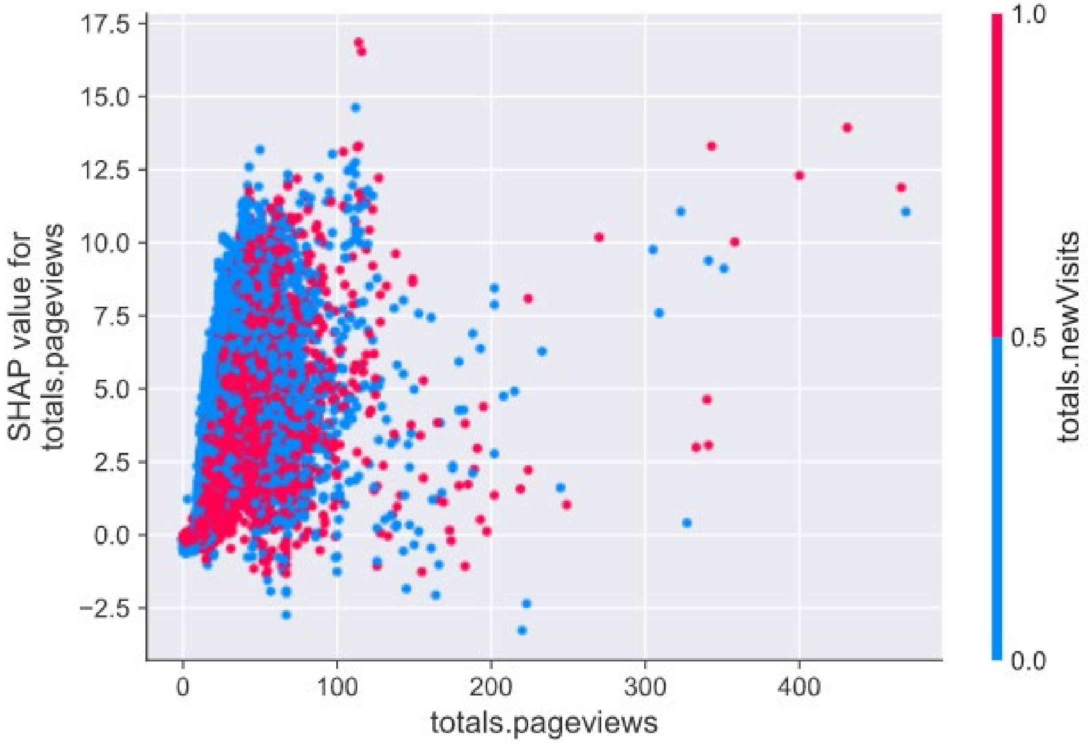


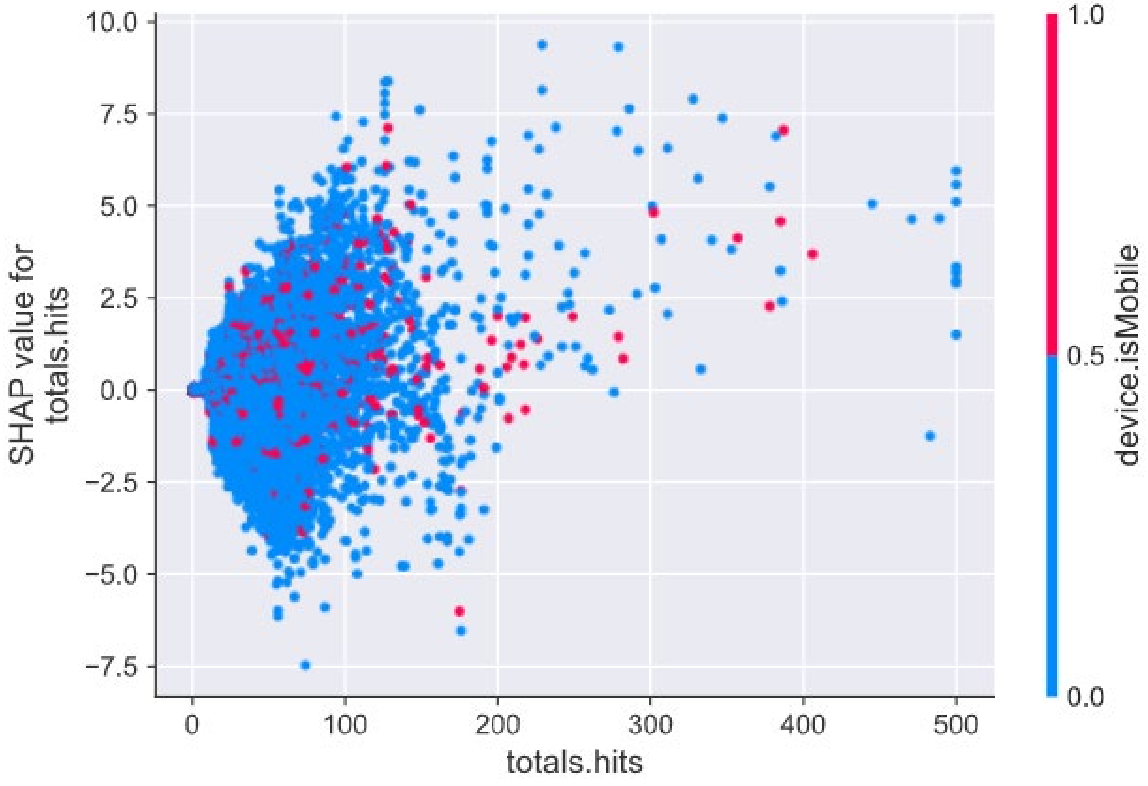

```
Training until validation scores don't improve for 150 rounds
[50]	valid_0's rmse: 1.72155
[100]	valid_0's rmse: 1.66003
[150]	valid_0's rmse: 1.6435
[200]	valid_0's rmse: 1.63709
[250]	valid_0's rmse: 1.63479
[300]	valid_0's rmse: 1.63336
[350]	valid_0's rmse: 1.63257
[400]	valid_0's rmse: 1.63188
[450]	valid_0's rmse: 1.63166
[500]	valid_0's rmse: 1.63119
[550]	valid_0's rmse: 1.63114
[600]	valid_0's rmse: 1.63098
[650]	valid_0's rmse: 1.63114
[700]	valid_0's rmse: 1.6314
Did not meet early stopping. Best iteration is:
[597]	valid_0's rmse: 1.63095
```

Feature importance on transaction level



Feature Dependence





Evaluation metrics:

1. Technical metric: RMSE/MSE

- Places higher weight on larger errors than smaller errors
- It is sensitive to outliers

2. Business metric: YoY Revenue Growth

- Increase revenue on a channel level

Recommendation

A/B landing page for Different Geographic Areas

Ensure all visitors are contactable (email or cell phone number)

Drive engagement and user action through gamifying the site experience

Introduce social proof messages

Personalized exit popups and overlays

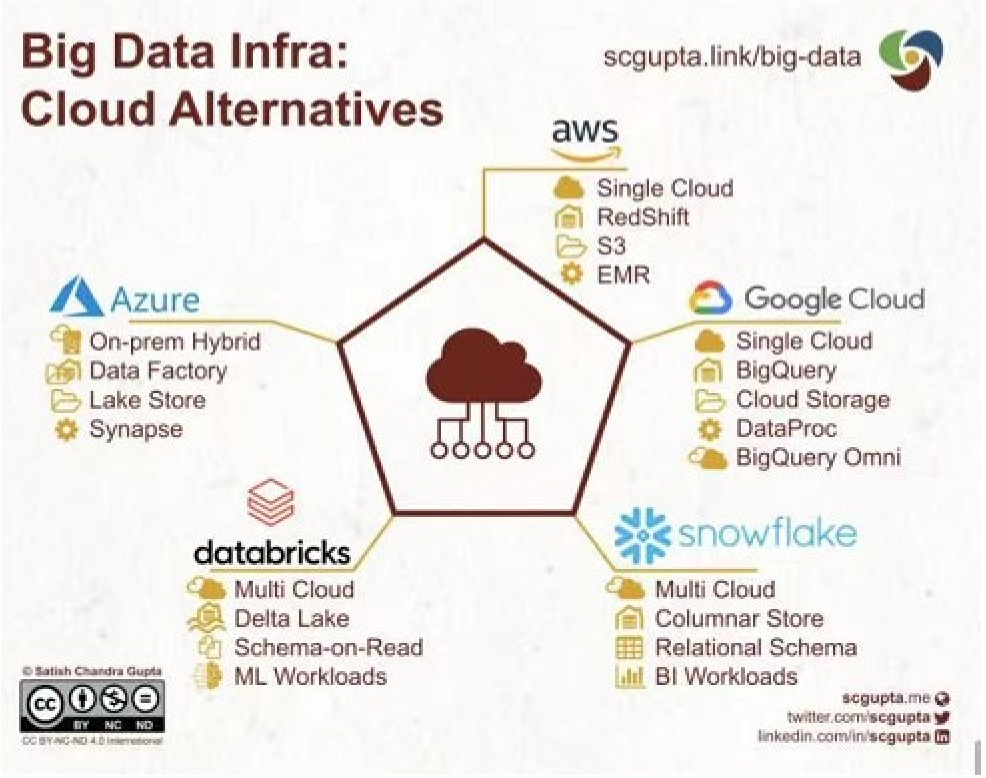
Include Push notifications

Allocate marketing budget for the end of the month

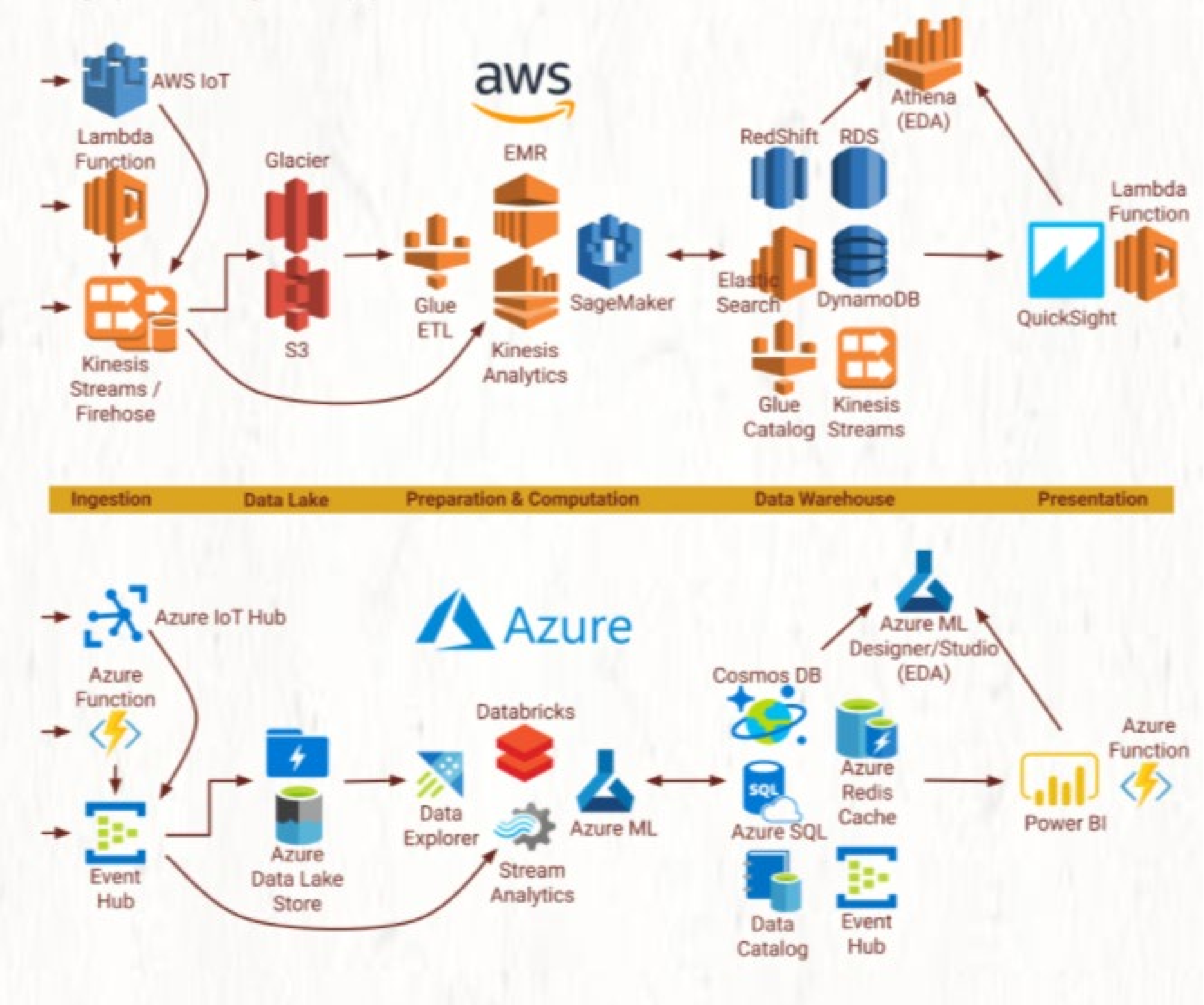
Visit Hours: mostly late evening 9-10PM, followed by 2PM, and 5-7AM

Target potential customers during evening and night hours, week days and closer to month end

Deployment

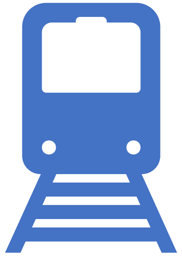


Big Data Infra:
Cloud Alternatives
scgupta.link/big-data
aws
Single Cloud
RedShift
S3
EMR
Azure
On-prem Hybrid
Data Factory
Lake Store
Synapse
Google Cloud
Single Cloud
BigQuery
Cloud Storage
DataProc
BigQuery Omni
databricks
Multi Cloud
Delta Lake
Schema-on-Read
ML Workloads
snowflake
Multi Cloud
Columnar Store
Relational Schema
BI Workloads
© Satish Chandra Gupta
scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta



AWS IoT
Lambda Function
Kinesis Streams / Firehose
Glacier
S3
Glue ETL
aws
EMR
Kinesis Analytics
SageMaker
RedShift
RDS
Athena (EDA)
Elastic Search
DynamoDB
Glue Catalog
Kinesis Streams
Lambda Function
QuickSight
Ingestion
Data Lake
Preparation & Computation
Data Warehouse
Presentation
Azure IoT Hub
Azure Function
Event Hub
Azure Data Lake Store
Data Explorer
Azure
Databricks
Stream Analytics
Azure ML
Cosmos DB
Azure SQL
Azure Redis Cache
Data Catalog
Event Hub
Azure ML Designer/Studio (EDA)
Azure Function
Power BI

Deployment Mode:



Train: Batch train in a certain frequency

Predict: Batch predict in a certain frequency

Batch deployment usually uses a deployment platform to schedule and monitor jobs

Appendix

Table 1

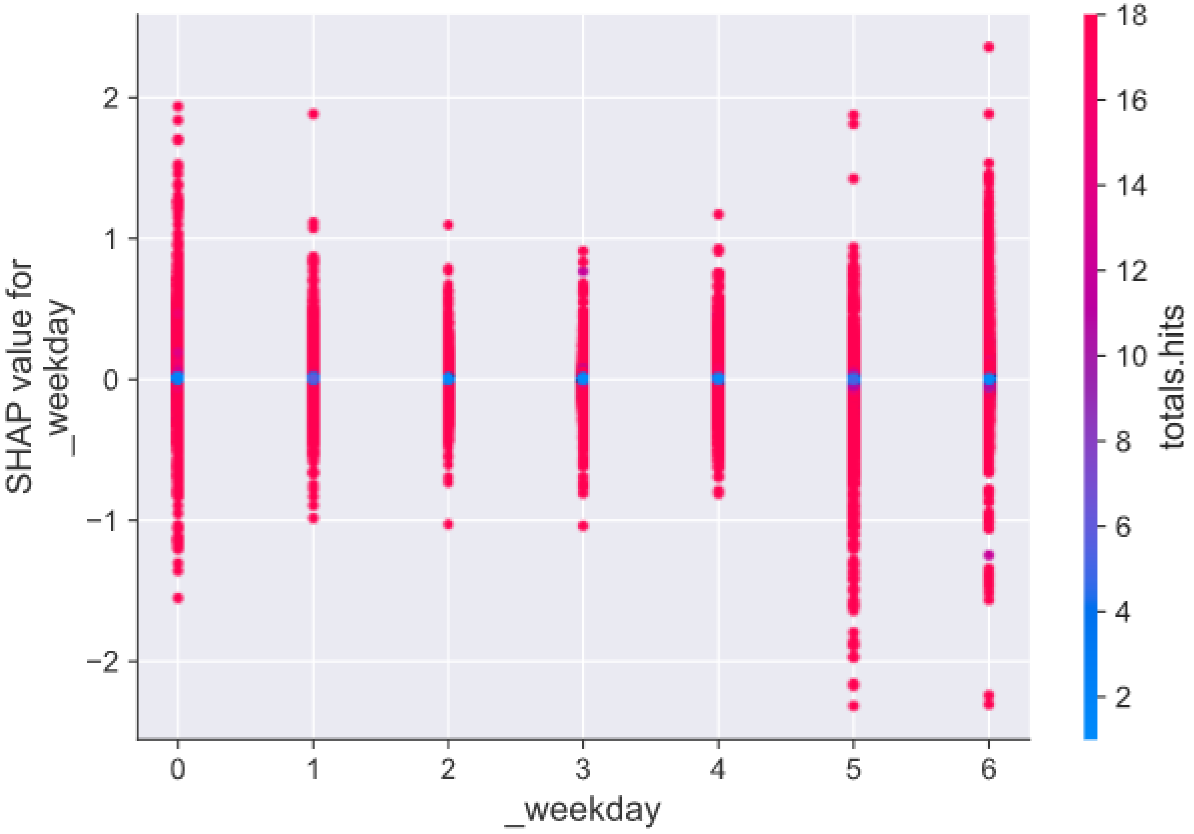
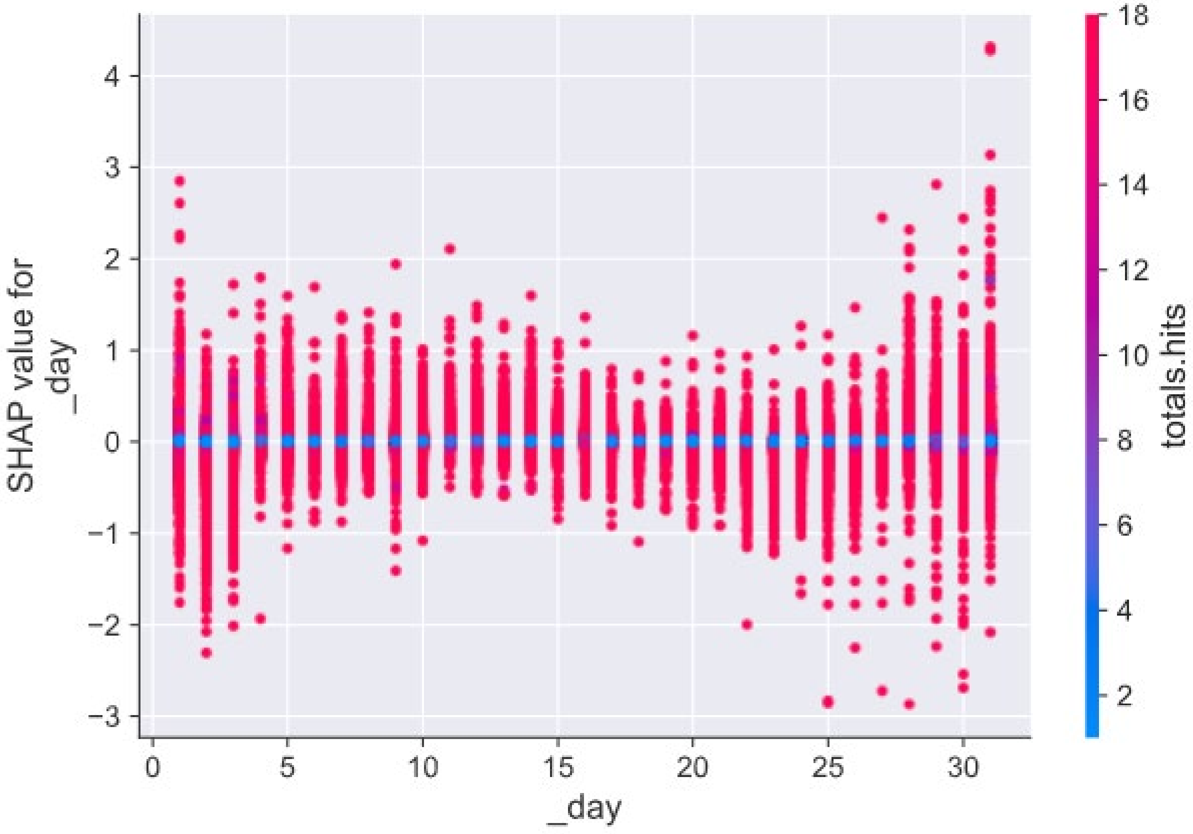


Table 2



Appendix: Data Quality

Microsoft Word
Document