CUSTOMER REVENUE PREDICTION
GOOGLE STORE

AUTHOR - NEHA

NOV, 2018

WHAT'S THE PROBLEM?

- 80/20 rule has proven true for Google merchandise store or GStore
- Small percentage of customers produce most of the revenue
- Marketing teams need prediction of revenue per customer to make appropriate investments in promotional strategies

DATA

- Data comes from Google Analytics in the form of CSV file
- CSV file includes json blobs with many fields
- 900K+ rows each corresponding to a visit to the GStore
- 40 columns or attributes related to the visits
- Transaction revenue hidden in of the json field is the target variable

DATA – SOME IMP FIELDS

- fullVisitorId - unique identifier for each visit
- visitId - identifier for the session
- sessionId - combination of visitor and visit IDs

fullVisitorId	sessionId	visitId
1131660440785968503	1131660440785968503_1472830385	1472830385
377306020877927890	377306020877927890_1472880147	1472880147
3895546263509774583	3895546263509774583_1472865386	1472865386
4763447161404445595	4763447161404445595_1472881213	1472881213
27294437909732085	27294437909732085_1472822600	1472822600
2938943183656635653	2938943183656635653_1472807194	1472807194

- date - unique identifier for each visit
- visitStartTime - identifier for the session
- visitNumber - combination of visitor and visit IDs

date	visitStartTime	visitNumber
20160902	1472830385	1
20160902	1472880147	1
20160902	1472865386	1
20160902	1472881213	1
20160902	1472822600	2
20160902	1472807194	1

DATA – SOME IMP FIELDS

- device - device used to access the Store, group field
- totals - aggregate values across session, group field
- geoNetwork - geography of user, group field

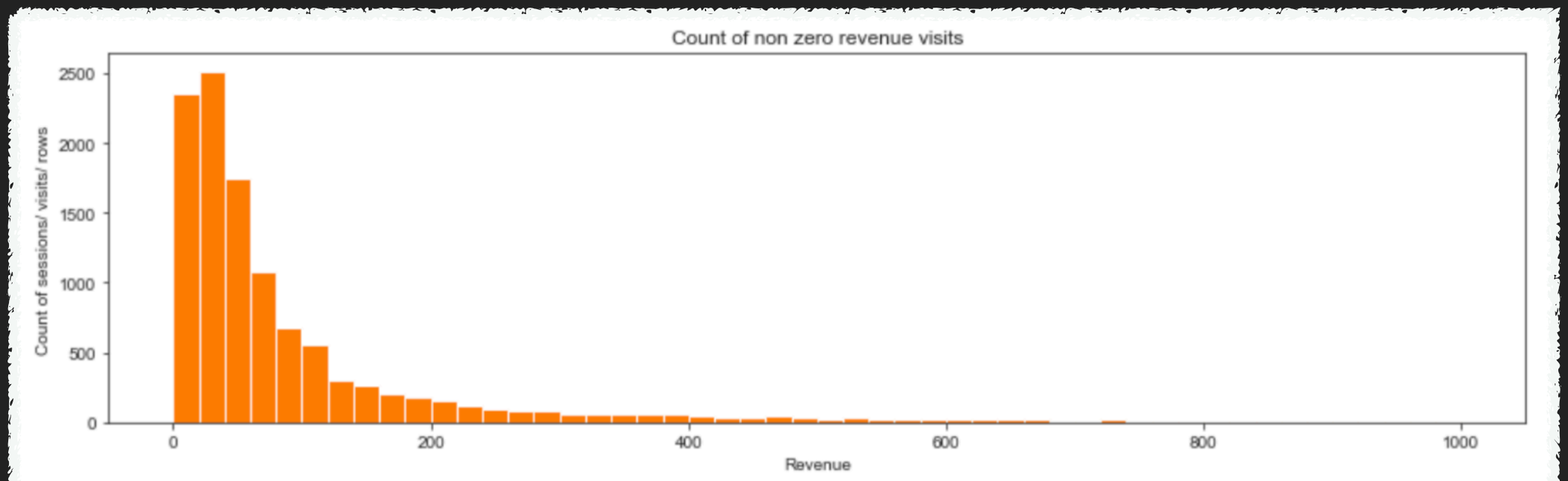
device.browser	device.deviceCategory	device.isMobile	device.operatingSystem
Chrome	desktop	False	Windows
Firefox	desktop	False	Macintosh
Chrome	desktop	False	Windows
UC Browser	desktop	False	Linux
Chrome	mobile	True	Android
Chrome	desktop	False	Windows

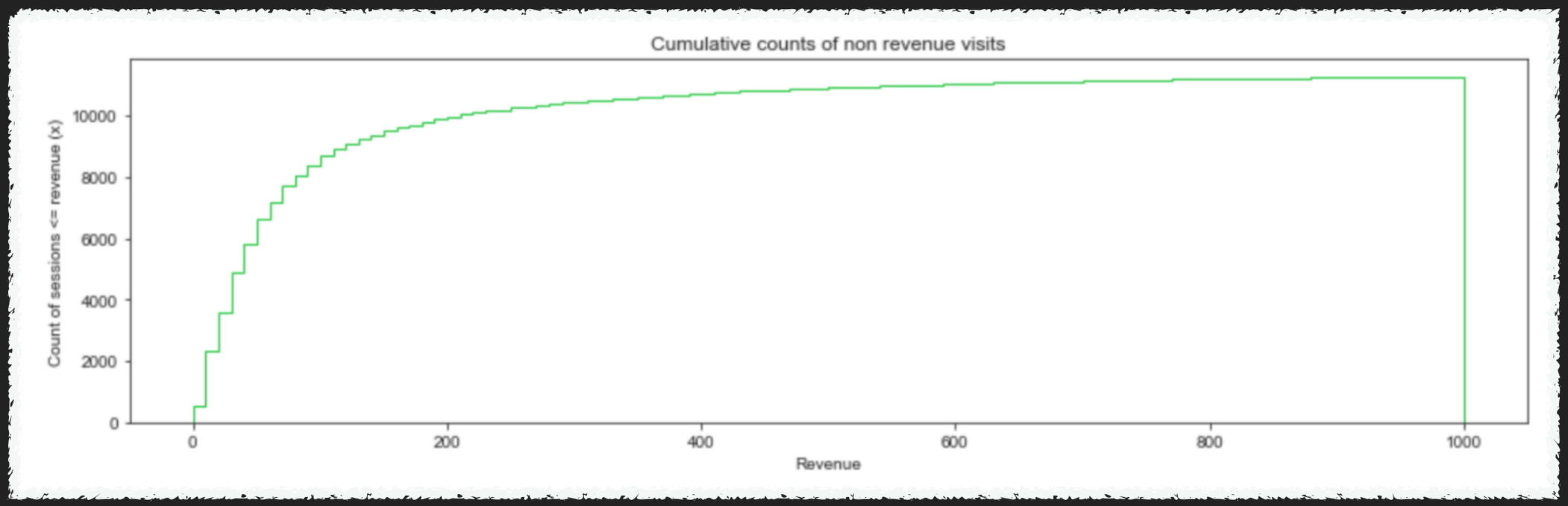
totals.bounces	totals.hits	totals.newVisits	totals.pageviews	totals.transactionRevenue
1	1	1	1	NaN
1	1	1	1	NaN
1	1	1	1	NaN
1	1	1	1	NaN
1	1	0	1	NaN
1	1	1	1	NaN

geoNetwork.city	geoNetwork.continent	geoNetwork.country	geoNetwork.metro	geoNetwork.networkDomain	geoNetwork.region	geoNetwork.subContinent
Izmir	Asia	Turkey	(not set)	ttnet.com.tr	Izmir	Western Asia
not available in demo dataset	Oceania	Australia	not available in demo dataset	dodo.net.au	not available in demo dataset	Australasia
Madrid	Europe	Spain	(not set)	unknown.unknown	Community of Madrid	Southern Europe
not available in demo dataset	Asia	Indonesia	not available in demo dataset	unknown.unknown	not available in demo dataset	Southeast Asia
not available in demo dataset	Europe	United Kingdom	not available in demo dataset	unknown.unknown	not available in demo dataset	Northern Europe
not available in demo dataset	Europe	Italy	not available in demo dataset	fastwebnet.it	not available in demo dataset	Southern Europe

VISUALIZE & EXPLORE

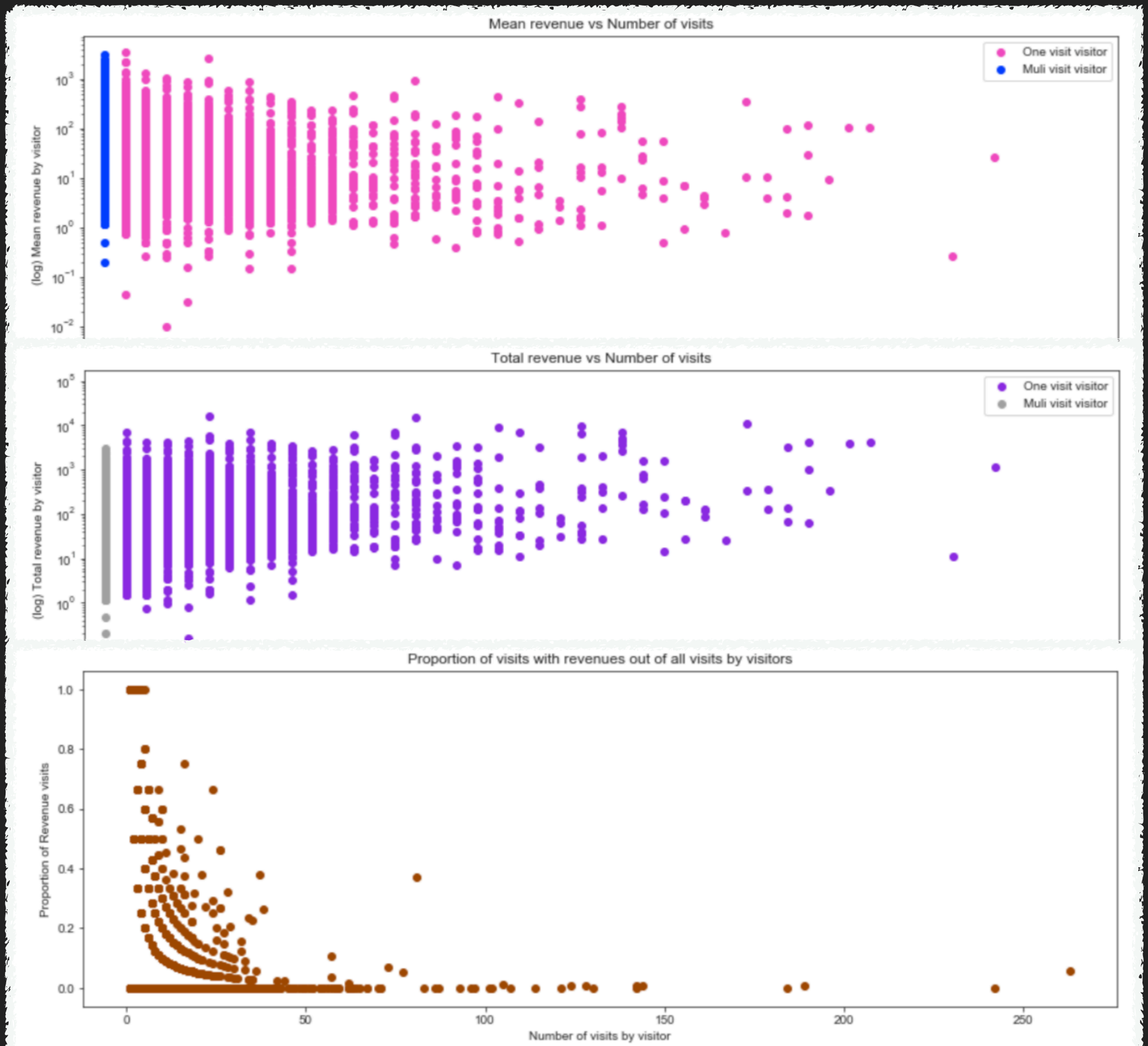
TARGET VARIABLE – TRANSACTION REVENUE





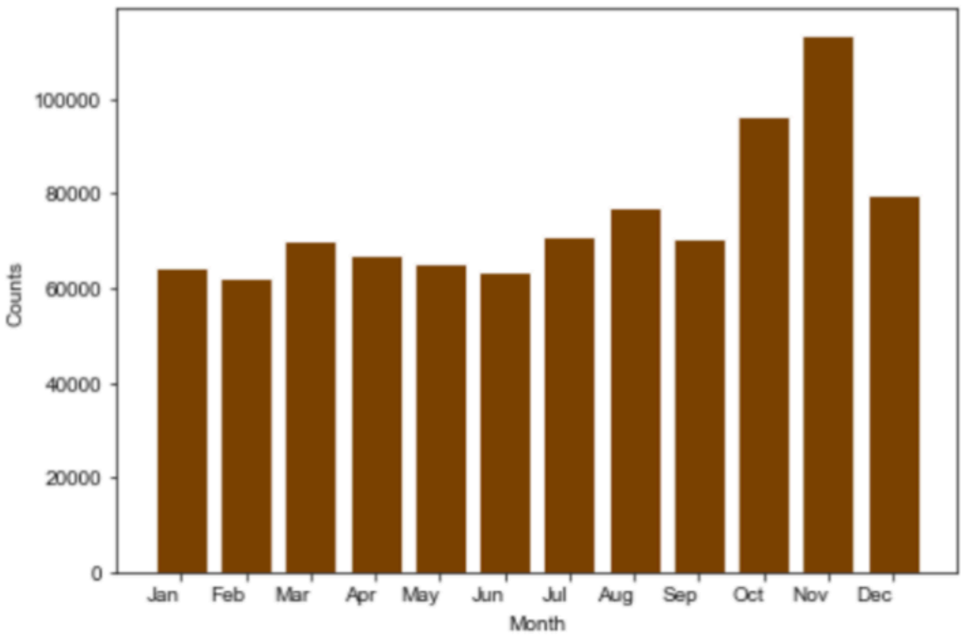
VISUALIZE & EXPLORE

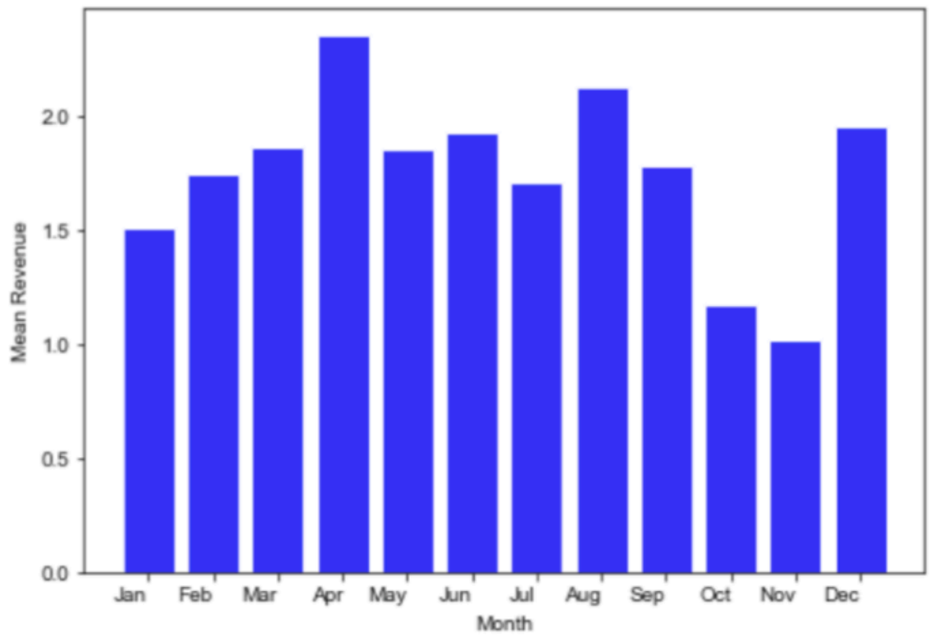
NUMBER OF VISITS

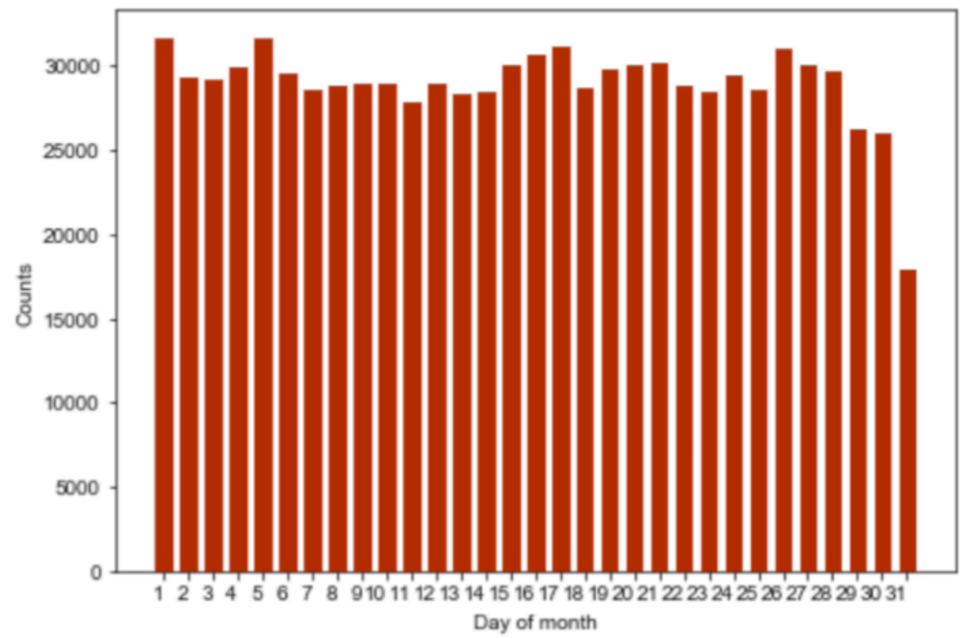


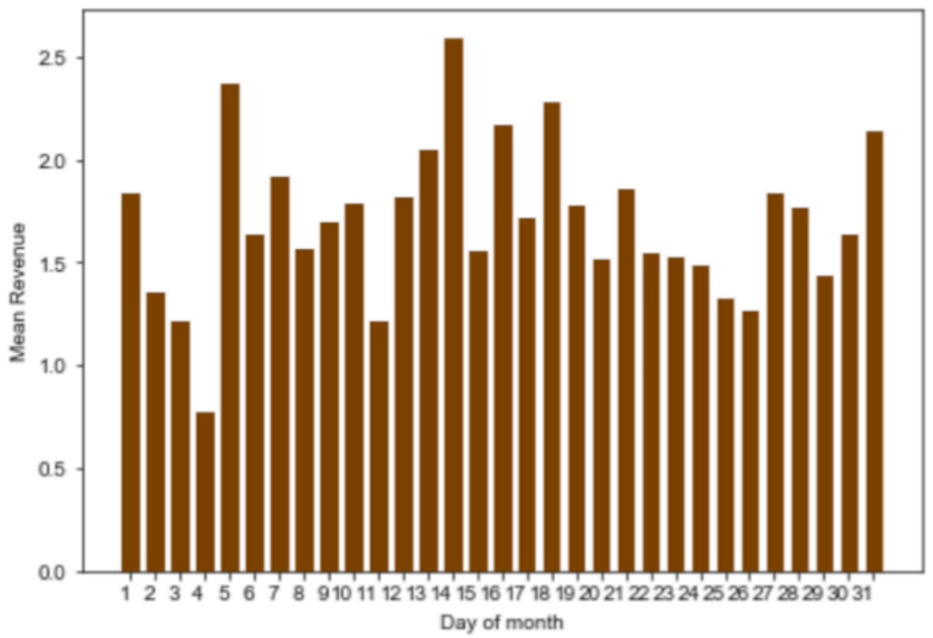
VISUALIZE & EXPLORE

MONTH AND DAY OF MONTH



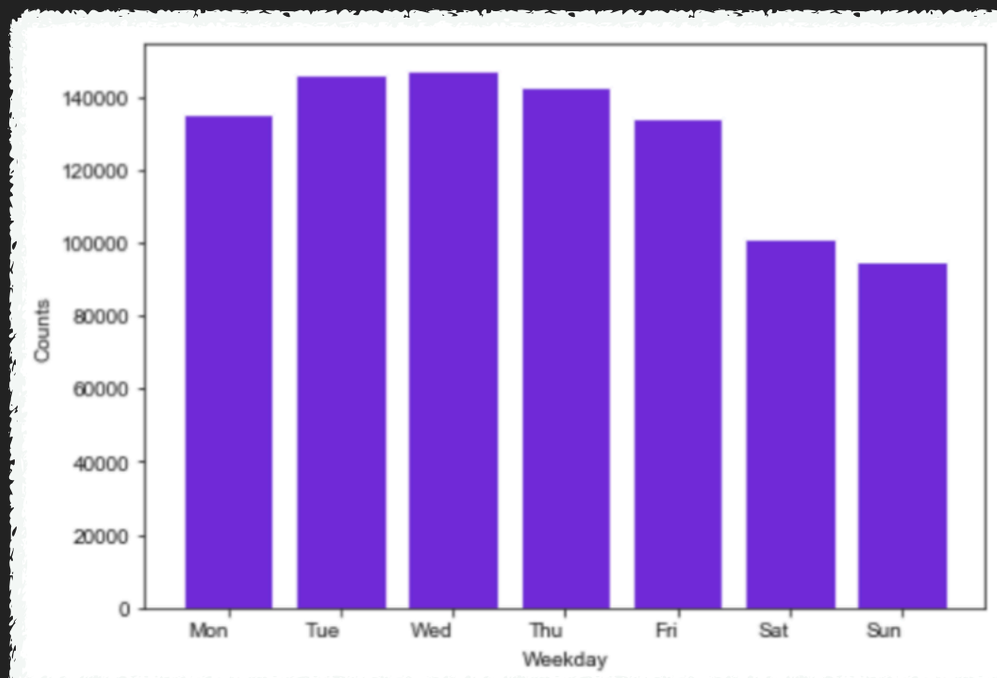


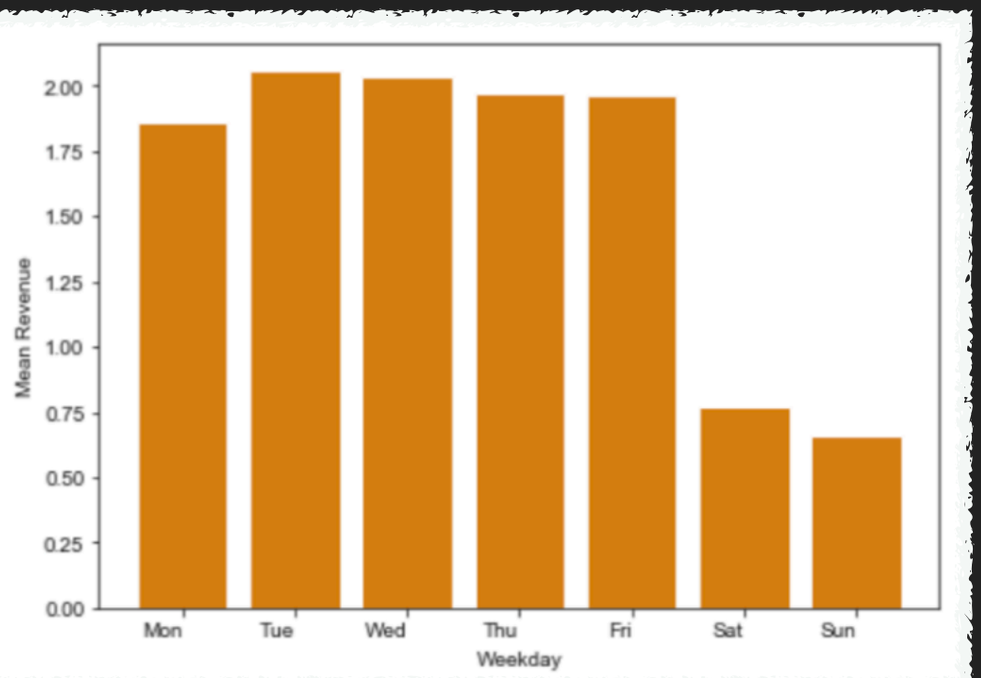


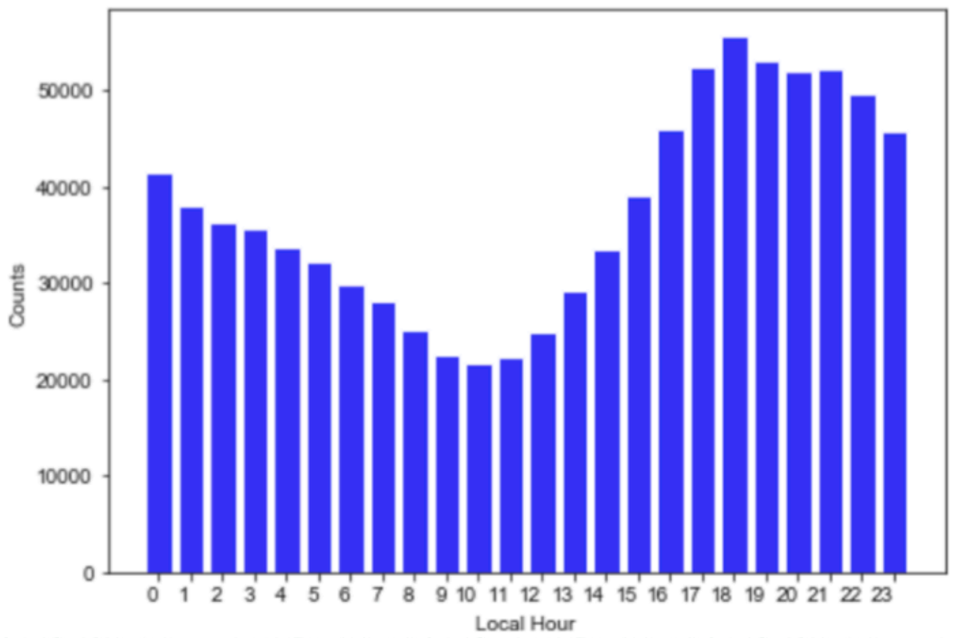


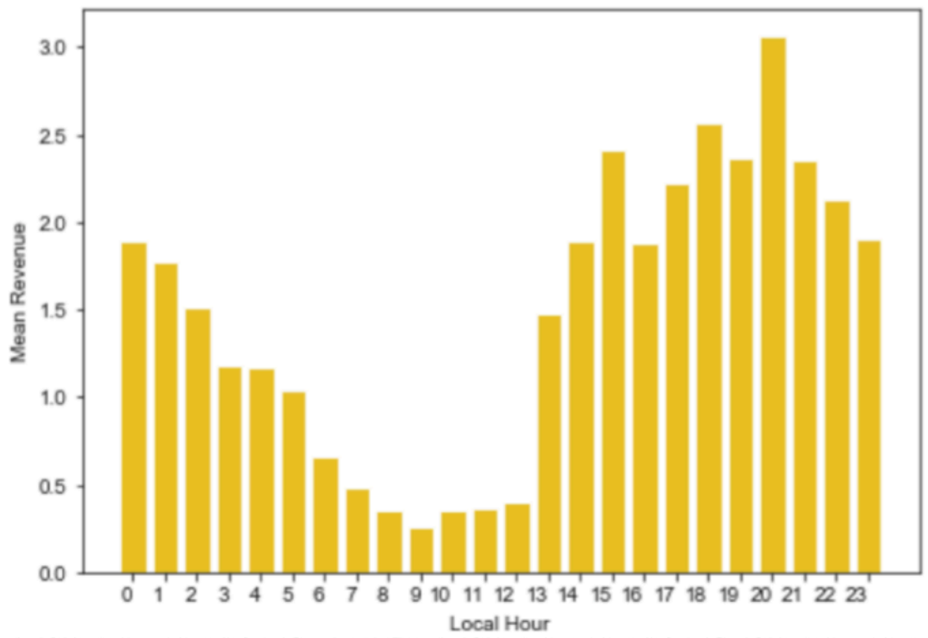
VISUALIZE & EXPLORE

WEEKDAY AND LOCAL HOUR



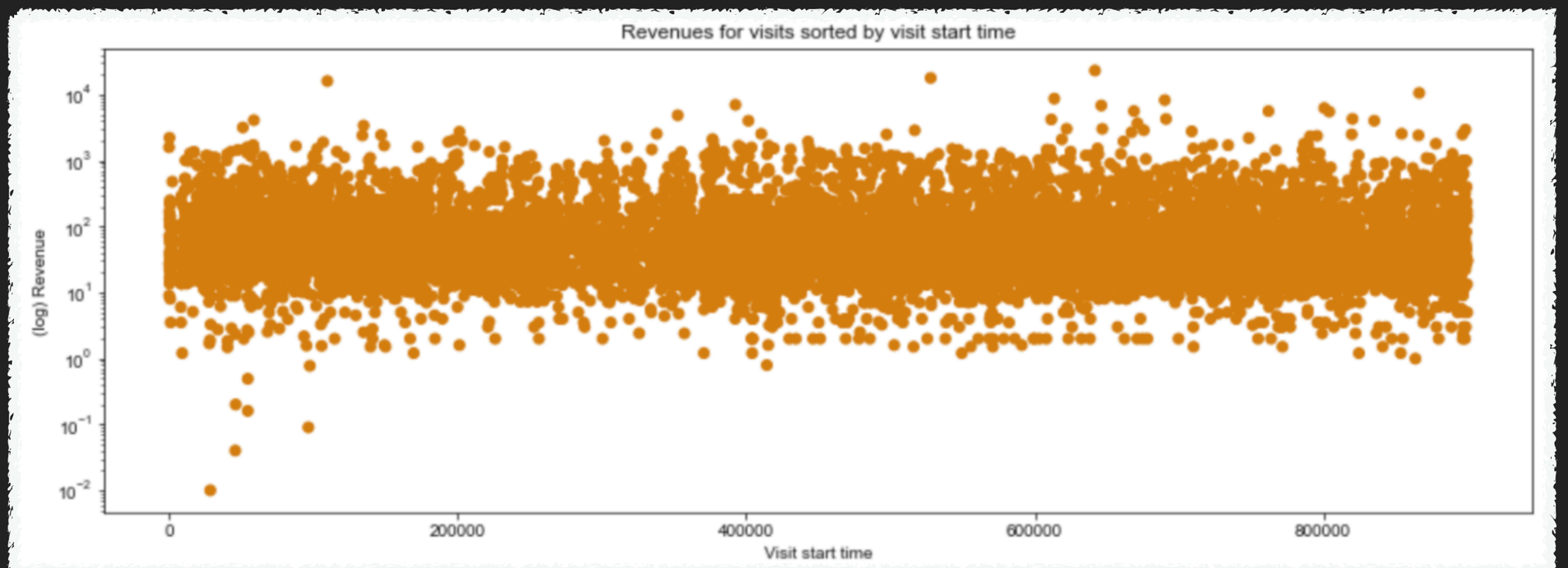


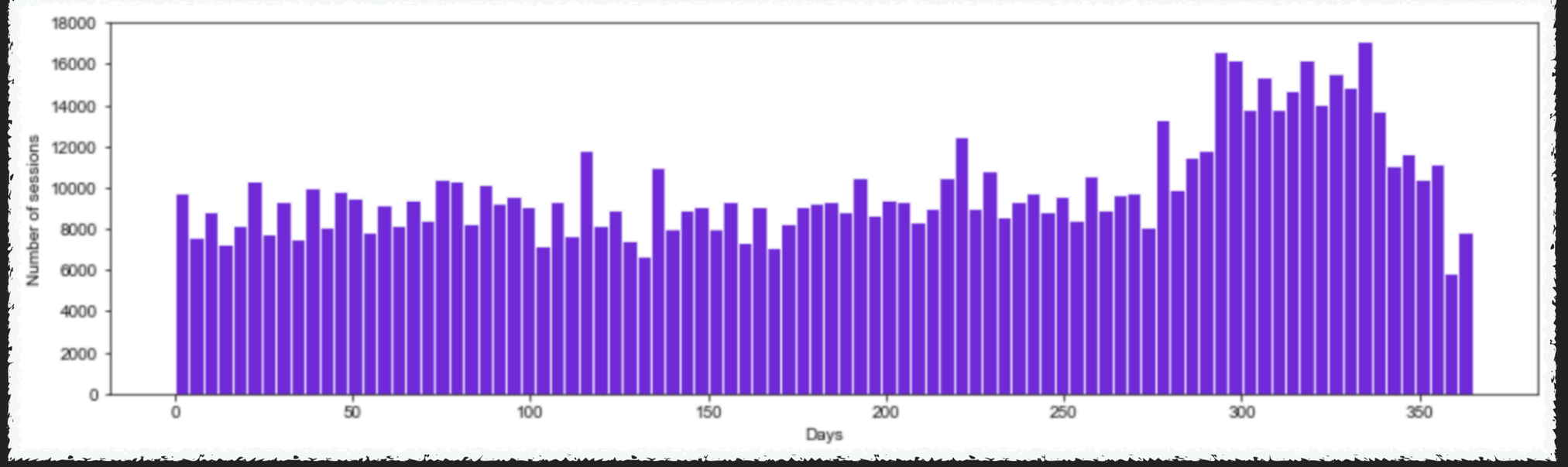




VISUALIZE & EXPLORE

VISIT START TIME AND DAYS OF YEAR

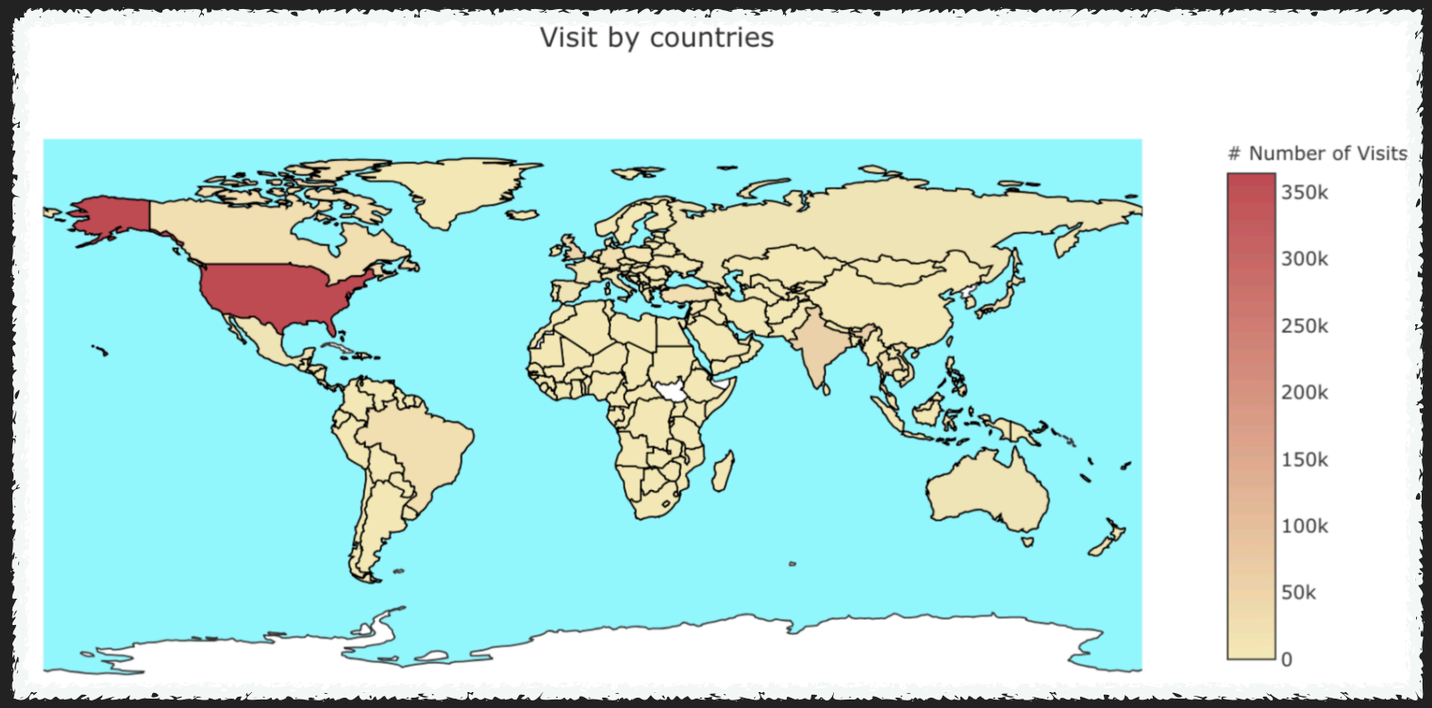


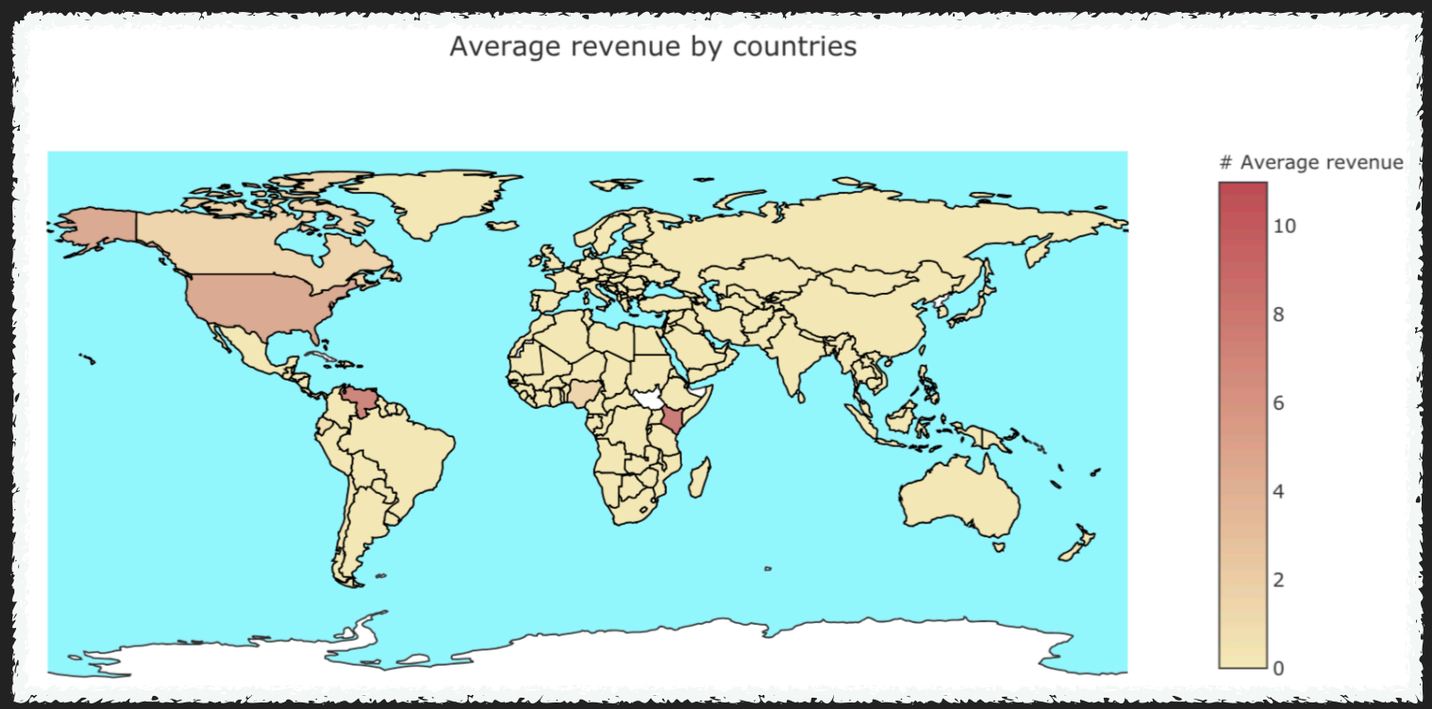


VISUALIZE & EXPLORE

VISIT START TIME AND DAYS OF YEAR

- continents are redundant
- sub-continent represent continent as well
- inconsistent rows for city, country, continent combinations were found and deleted





ML MODELING

- Overfitting and Underfitting
- Model Evaluation Criteria - Root mean squared error (RMSE)

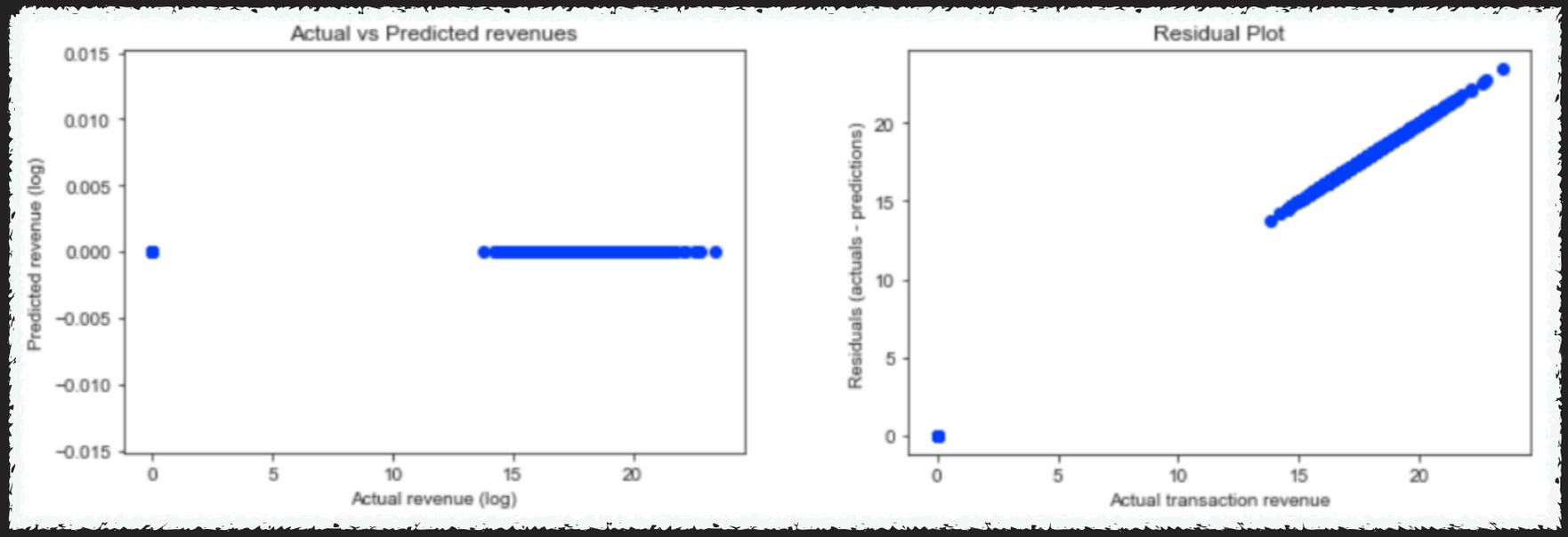
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

- Cross Validation, Held out test set
- DRY principle - using Python functions
- Cardinality reduction and hot encoding for categorical variables
- Basic features selection

BASELINE RESULTS

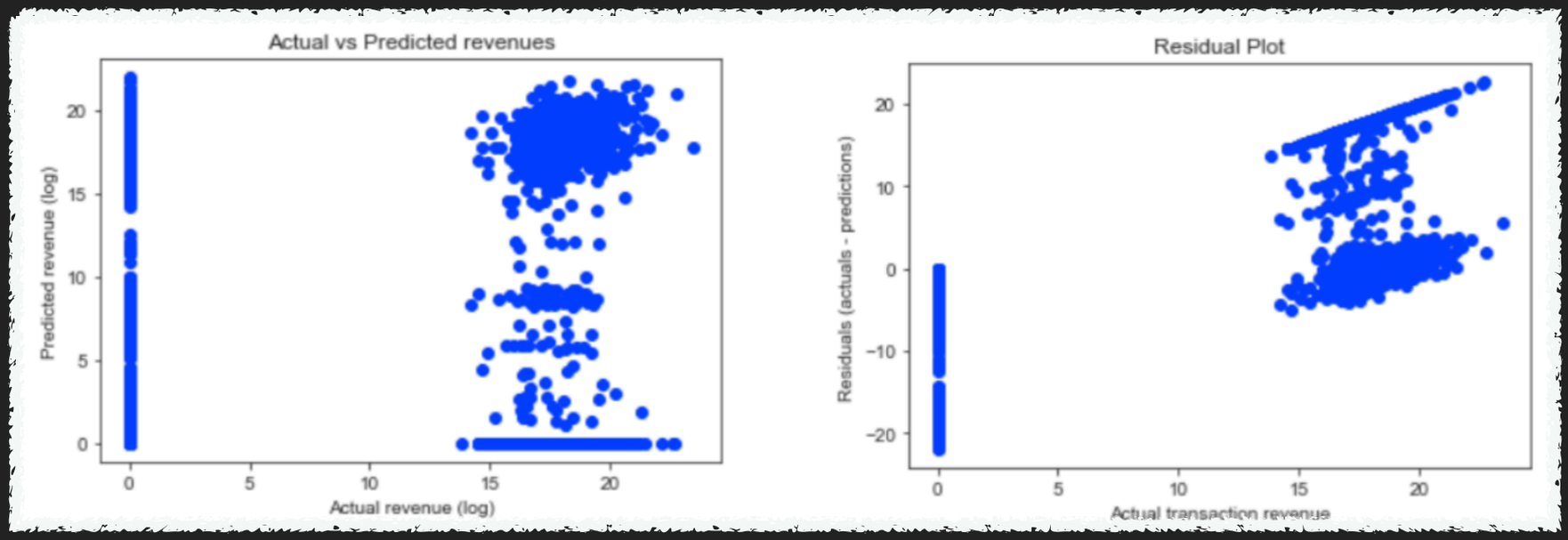
- Predictions at human level - always predict revenue as zero. The results will serve as comparison point for all our future predictions.

RMSE: 0.290



- Decision Tree - Quick and dirty to see ML based baseline.

RMSE: 0.298



FEATURE ENGINEERING

hour - Extracted this feature from visitStartTime. The login hour might serve some business meaning.

local_hour - This is hour but calculated keeping in mind the location of the visitor. This is more meaningful than just UTC hour.

day - We saw some patterns within the month. This feature can capture that pattern.

day_of_year - Again based on EDA there are variations in the data for all the days of the year. weekday - Captures the weekly patterns of the data.

month - Captures the monthly pattern of the data.

quarter - The whole year can be divided into quarters. This feature might capture holiday season or financial year start etc.

FEATURE ENGINEERING

week_day_end - As seen from EDA there are more logins or weekends but less sales. It might indicate that people have leisure time and they browse and not necessarily interested in buying.

holidays - At the end of the year there is first rise in the number of the logins and then some days later there is rise in revenue transactions. We can define that period as holiday period.

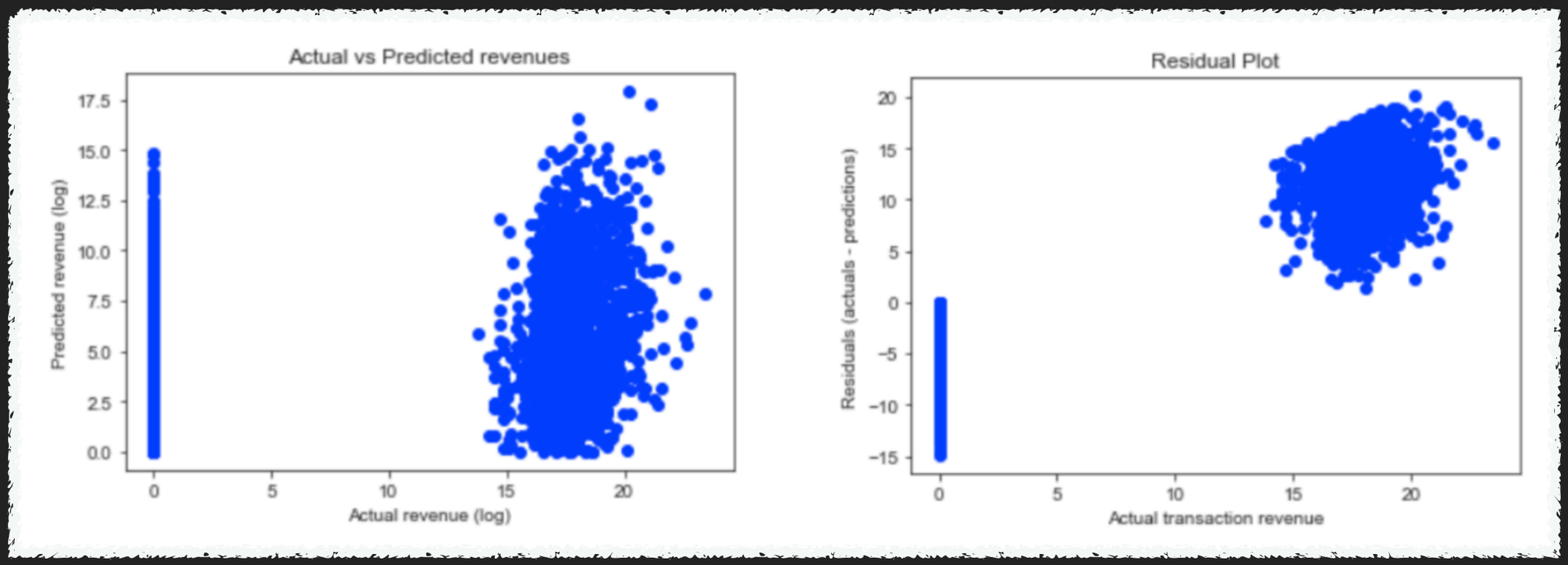
month_period - Whole month can be divided into start, mid and end. Start might indicate the arrival of salaries etc. Month end might be related to more expenditure.

active_hours - As seen during the EDA that morning hours are generally low on revenue transactions. This makes sense because people are generally busy getting ready for the work day ahead.

FINAL MODEL

The best model is the Light GBM. It is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm.

The **RMSE** of 1.756 is better than all other RMSEs. The residual plot and actual vs predicted plot below show that this is indeed better model.



RECOMMENDATIONS

- Model can predict the revenues per visitor. Target the advertisement towards individual visitor based on their potential expenditure.
- We see more visits during the holiday time period but not much increase in revenue. The advertisement data can be used to better advertise in that season.
- Local active hours show that the morning hours are least active and we could target particular segment of customers like old or non working people.
- Video Ads seem to be wasted. Might spent the advertising budget elsewhere.
- Traffic source shows that "Google search" brings lot of revenue.
- Loyal visitors can be identified and rewarded.

CONCLUSIONS

- Problem can be modified to serve even a greater business purpose. We can actually predict revenues for future dates, for which we have no data collected whatsoever.
- Above can be done in next phase of the project. It will involve Time Series modeling and the model will be based on per visitor. The goal can be summarized as: Predict for the current visitors, if they will generate any revenue for future month (or some period) and if yes, predict the amount too.
- The current problem was hard because of the huge amount of data (1 million rows approx.). Due to the lack of available resources we could not try out XGB or NeuralNets on the full data (full cardinality categorical features) and with cross validation. Given more resources we can try those models.
- We can try many more new features like: Cross features like device_browser, browser_networkDomain, hits_to_pageviews_ratio
- We can try assigning ranks to unique values of categorical variables. The ranks can be based on mean revenue for that value of category or count of visits. For e.g. the "Northern America" region has the highest mean revenue And so it can be given rank 1st and so on.