Google GStore Customer Transaction Analysis

Team Member:
Duo Zhang
Meixing Dong
Yu Tian
Zishen Li

Agenda

- Key Goal
- Data preparation & Description
- Data exploration
- Modeling Approach
- Conclusion & Further Investigation

What we are trying to do

Goal: predict the revenue of the Gstore in the future

- What features may impact the revenue
- How to use these features to build the model
- What are the possible promotional strategies.

Data Preparation and Description

- Introduction
- Data tidy
- Pre-processing

Introduction of the dataset

- The data
  - Google Analytics Customer Revenue Prediction
  - **12** variables(4 JSON format) and **903,653** observations.

Some of the variables

- **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.
- **hits** - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.

Data source: https://www.kaggle.com/c/ga-customer-revenue-prediction

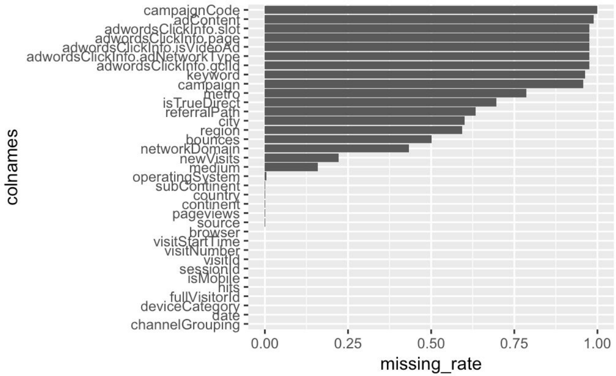
Data tidy

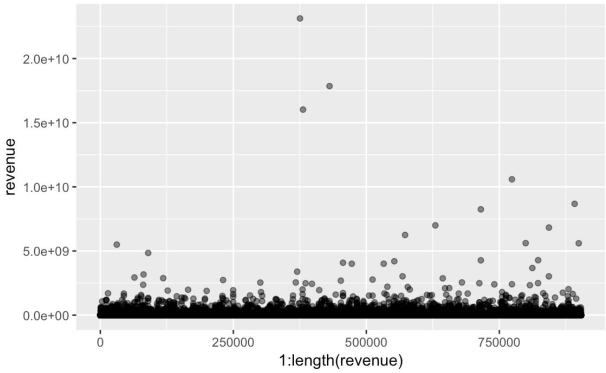
- Parse the JSON format
- Convert variables to their natural representation

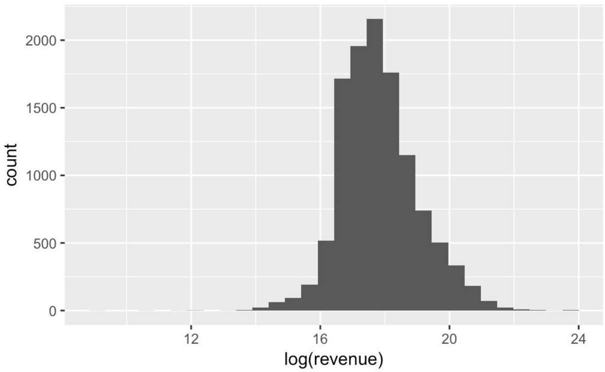
55 variables → 51 categorical variables; 4 continuous variables

Data pre-processing

- Constant columns: 20
- Missing value
  - 15 variables has more than 50% missing value
  - 98.7% of the response variable are missing(not buy)
- Normally distributed response variable(without missing value)







Data exploration

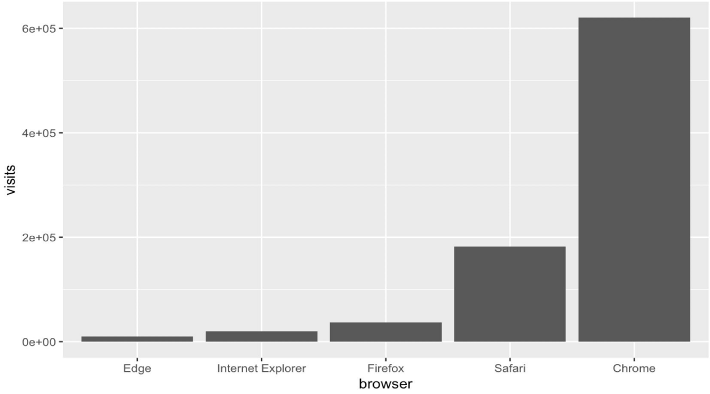
- Channel related features
- Geographical features
- Time related features
- User behaviours

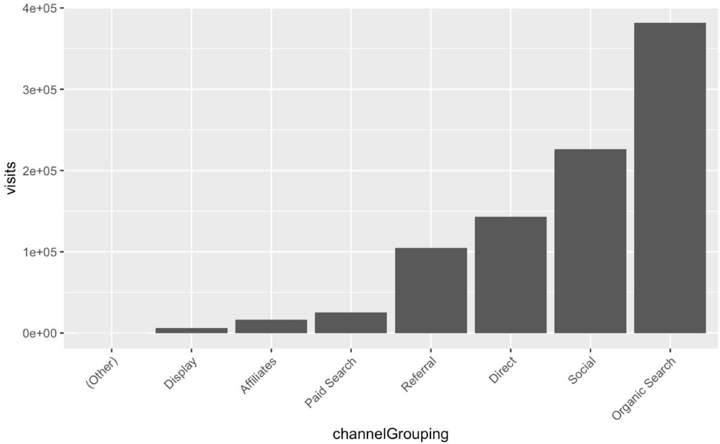
Channel related features

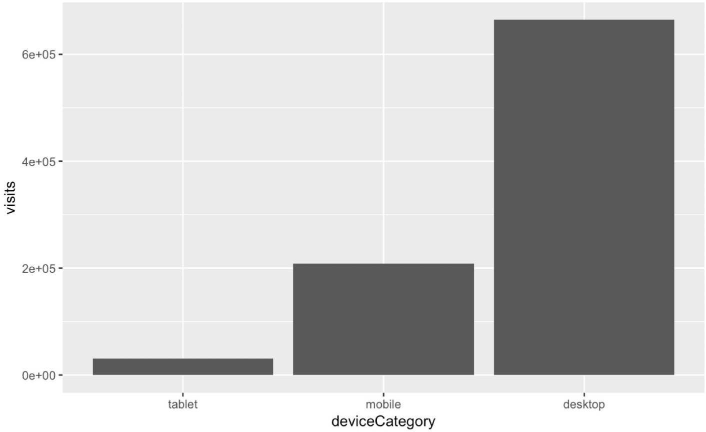
- Usage frequency
- Contributions to the revenue

Channel related features-Usage Frequency

- Organic Search and Social are the two most frequent channels
- Desktop and Mobile are the two most frequent devices
- Chrome and Safari are the most two popular browser

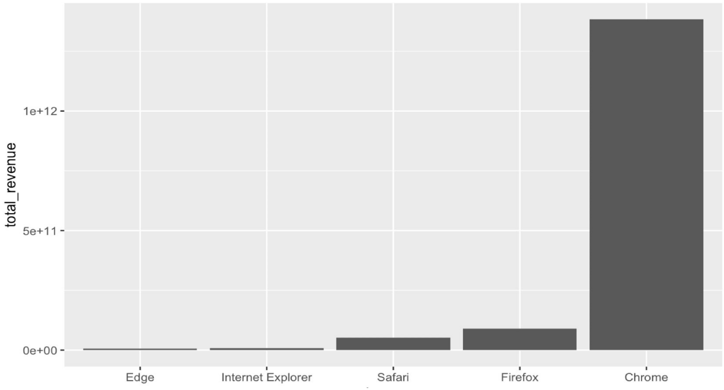


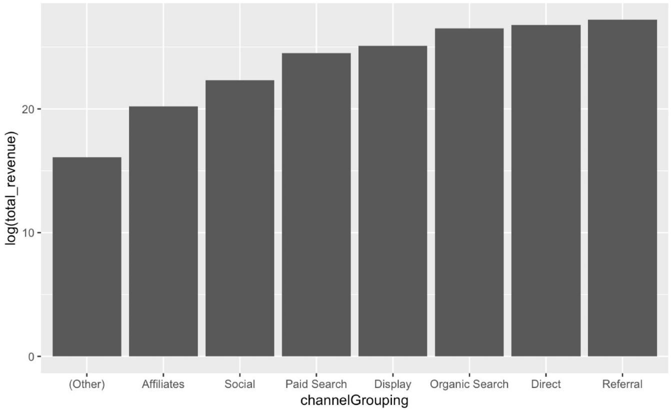


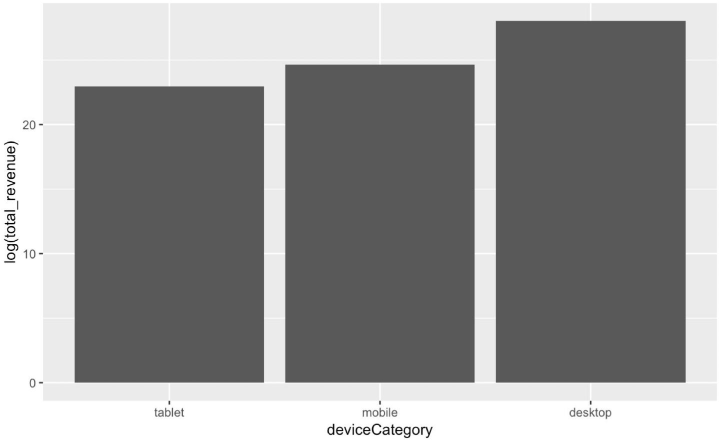


Channel related features-Contribution to the revenue

- Direct and Referral contribute to the revenue most, not the Organic Search and Social
- Desktop and mobile are two devices that contribute to the revenue most.
- Users from Chrome produce the highest total revenue.
- Safari contributes less than Firefox even though it is more popular than it.







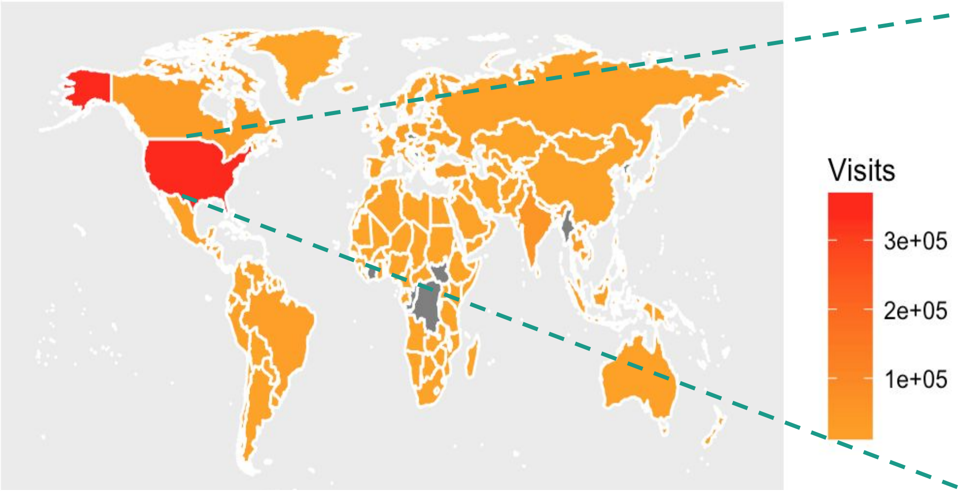
Geographical features

- Visits
- Browser & Device
- Page views & Hits
- Revenue

Where are the visits from?

Top one country: U.S. **Top 5 cities** : Mountain View, New York, San Francisco, Sunnyvale, San Jose

Visits in different countries

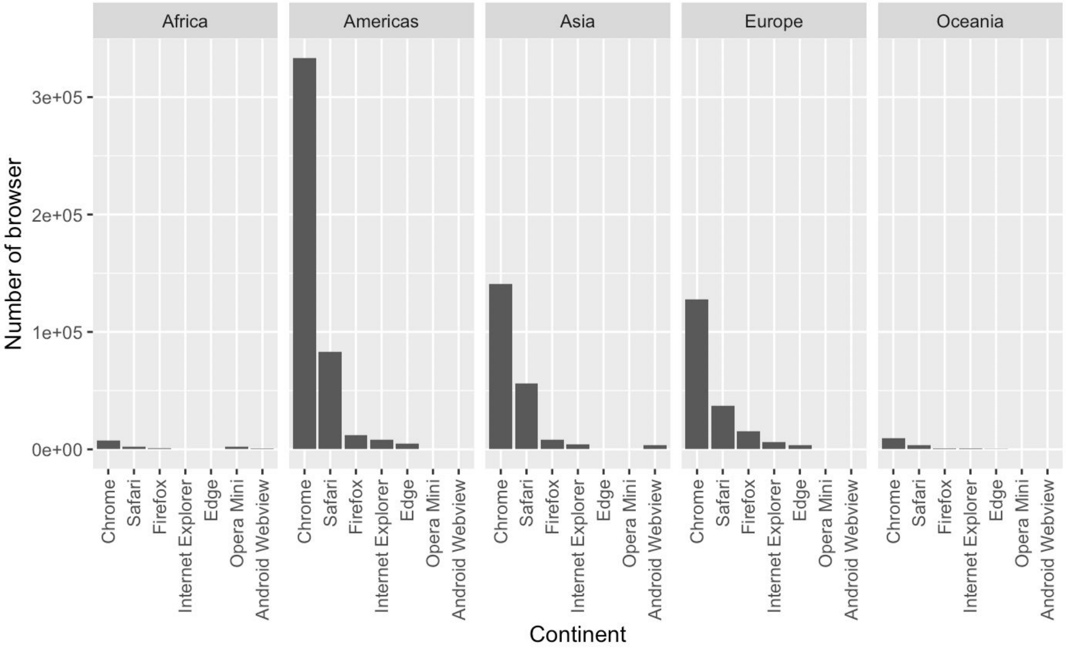


Top 10 Cities with the most visits

Geographical Distribution of Browsers

Top 7 browsers that used:

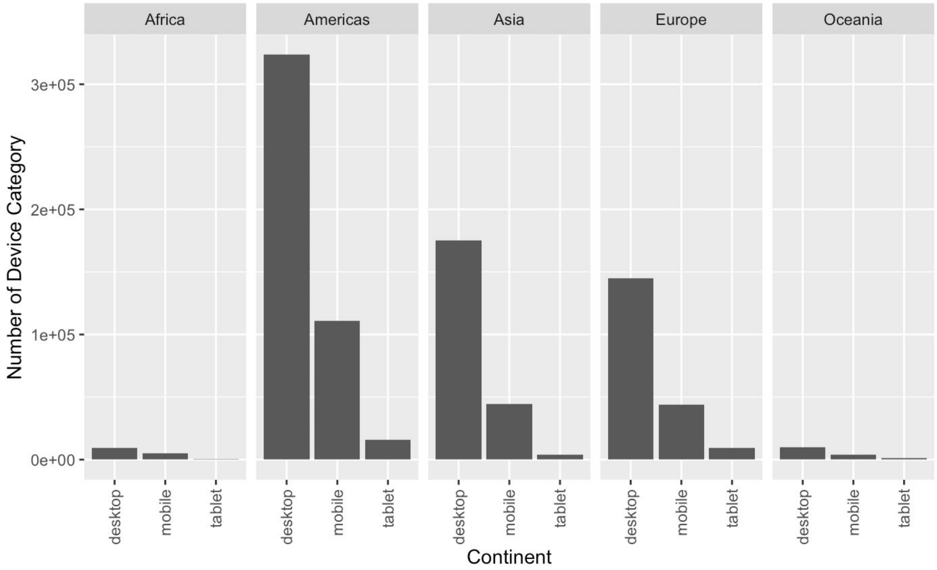
- Chrome
- Safari
- Firefox
- IE
- Edges
- Android Webview



Geographical Distribution of Devices

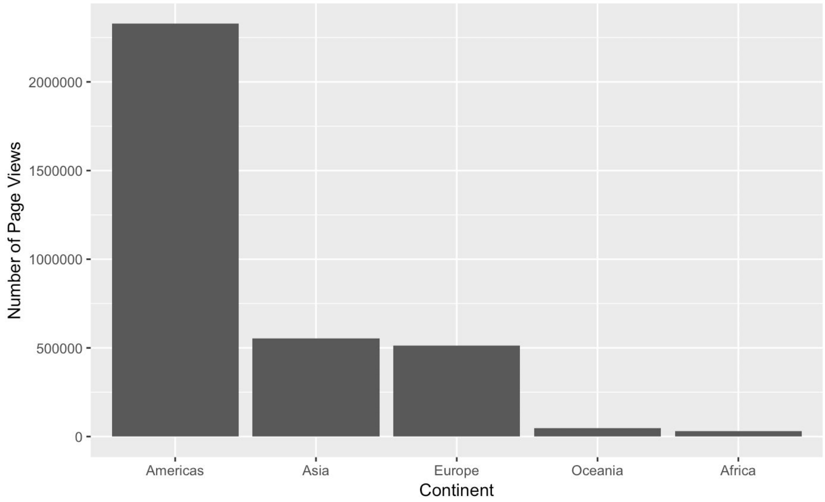
Devices:

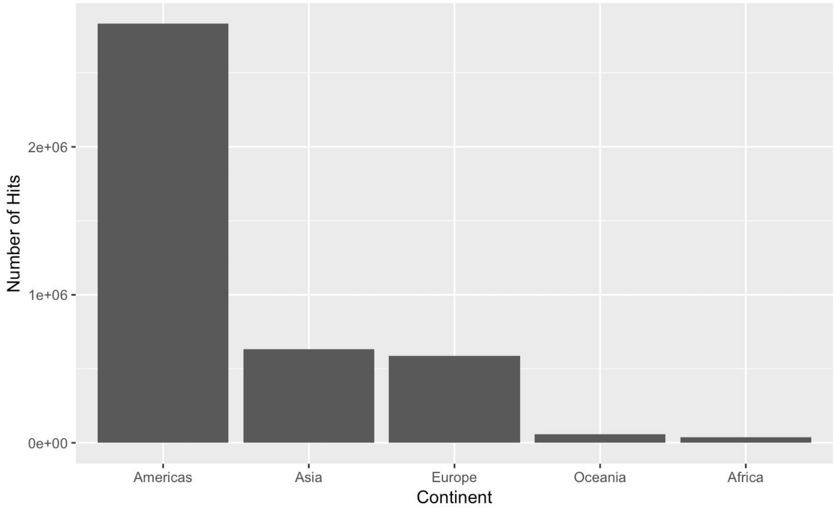
- Desktop
- Mobile
- Tablet



Geographical Distribution of Page Views & Hits

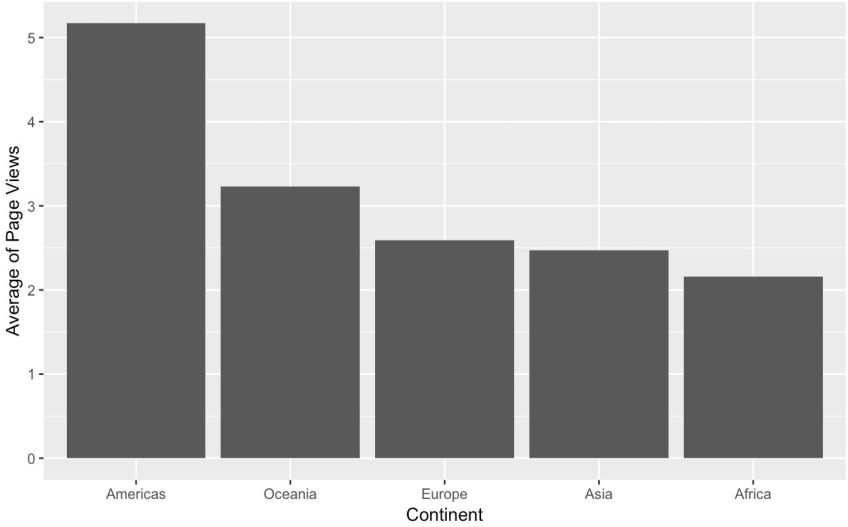
Top one continents: Americas

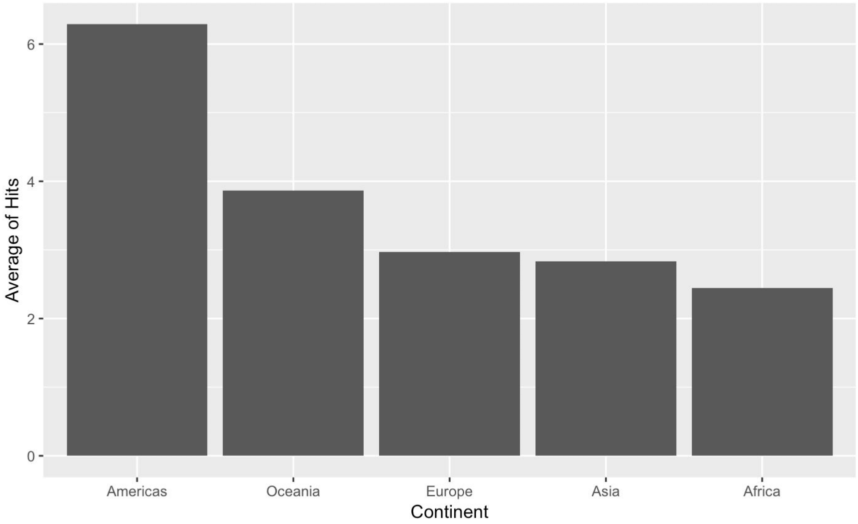




Average Page Views & Hits

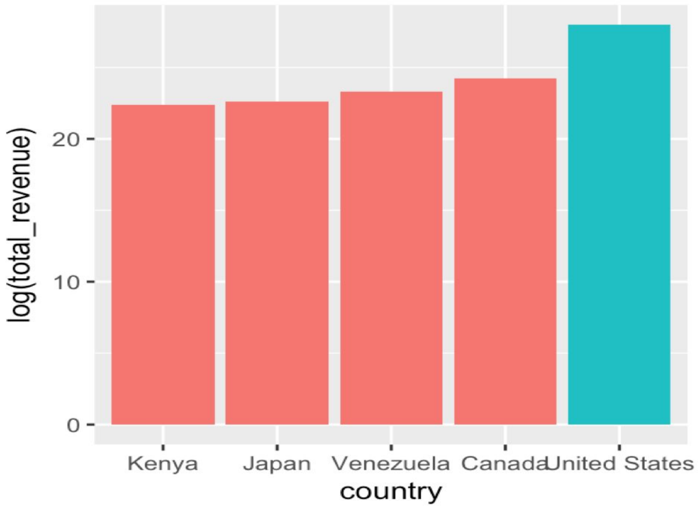
Top one continents: Americas

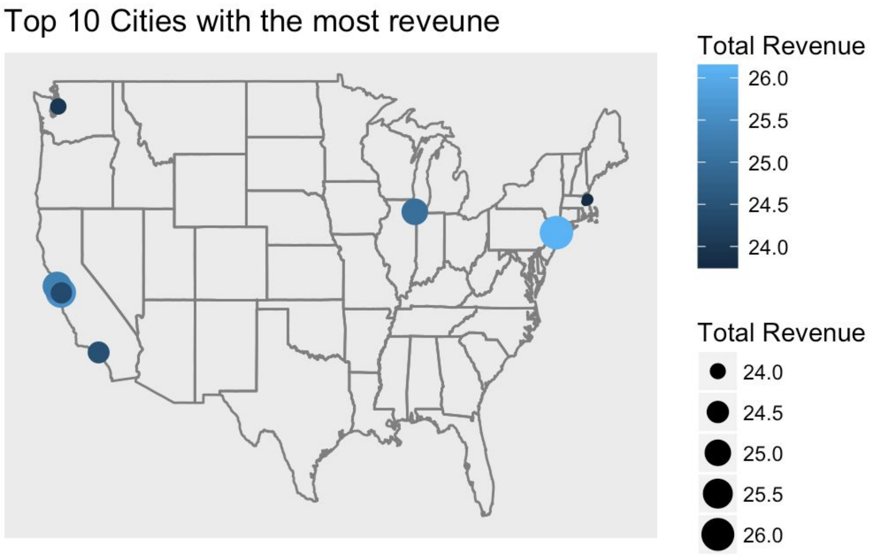




Geographical Distribution of Gstore's sales

Top one country: U.S. **Top 5 cities** : New York, Mountain View, San Francisco, Chicago, Los Angeles



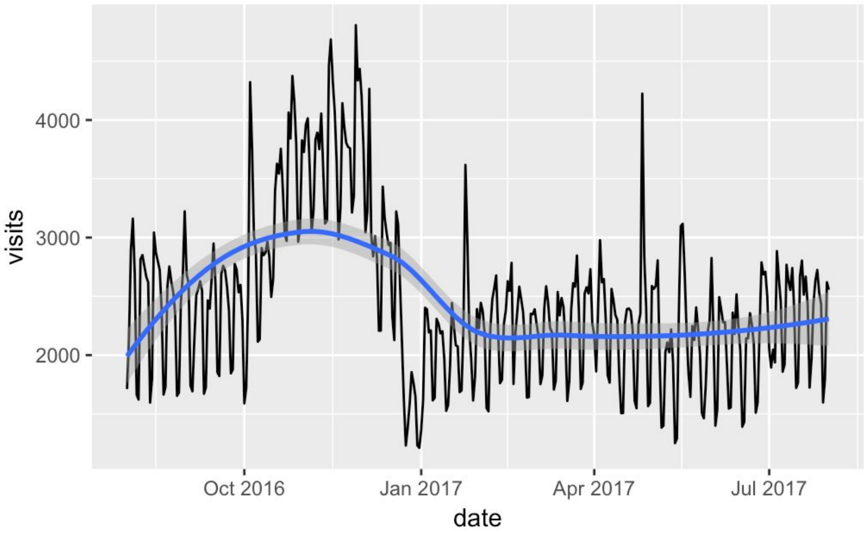


Time related features

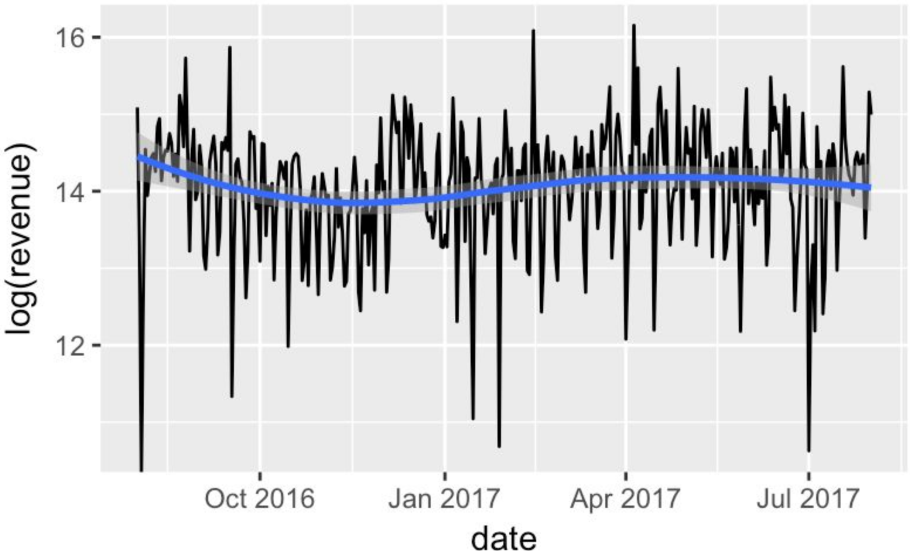
- Trends of user accesses and transactions
- Weekly and monthly behaviours
- Visits of device over periods

Trends of user accesses and transactions

- Visit peak around Nov. 2016

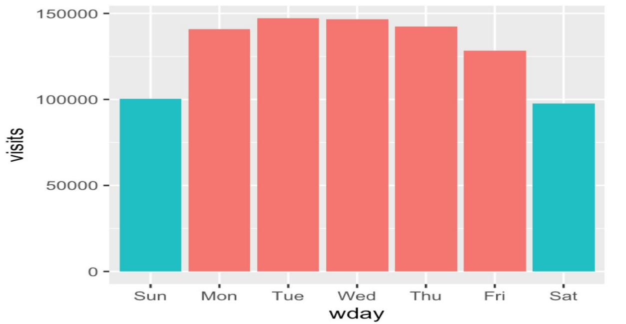


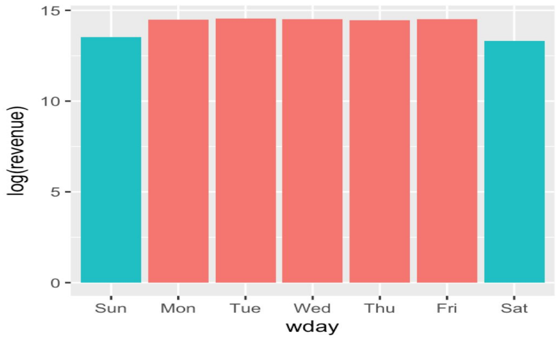
- Opposite pattern

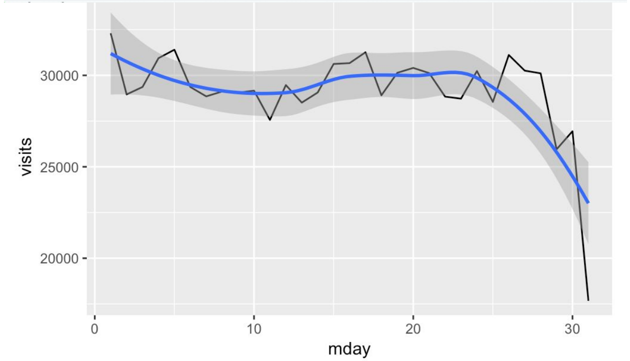


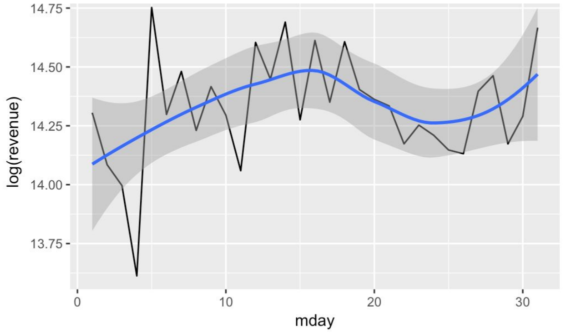
Weekly and monthly behaviours

- Surprisingly, people visit Gstore more often on weekdays.
- The monthly trends of the number of visits and the revenue are opposite





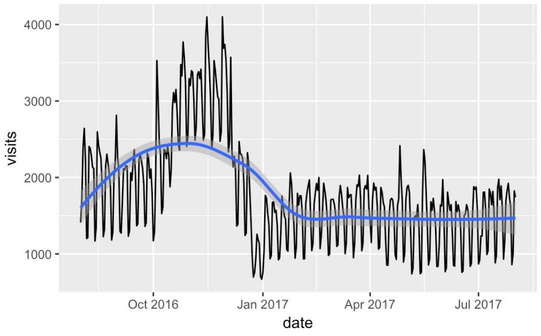




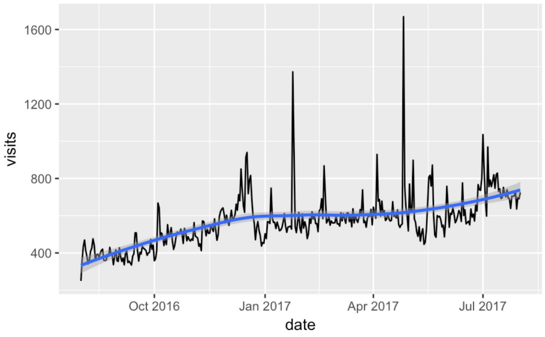
Visits of Device over periods

The visits from mobile and tablet are increasing while that of desktop is decreasing

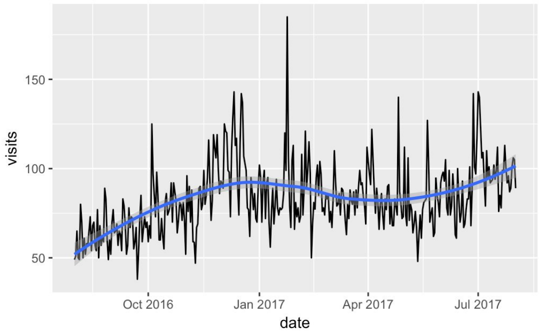
Desktop



Mobile

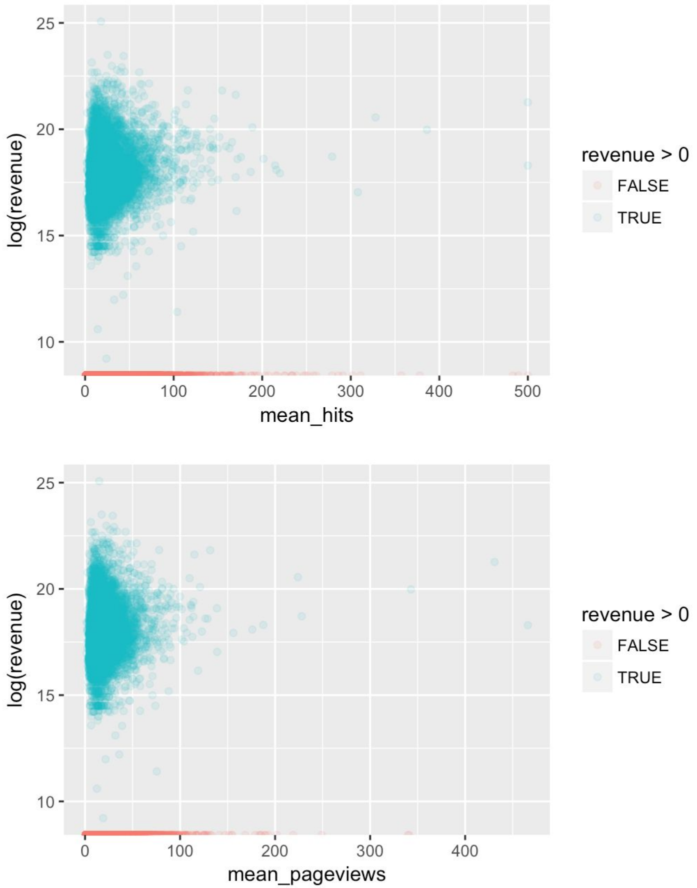


Tablet



User behaviors

- The average page views and hits for users who buy or not are quite different
- Features could be used for classification model in the next step



Model approach

- Use historical data to build up model
- try to predict user transaction revenue
    - based on user behavior and user information.

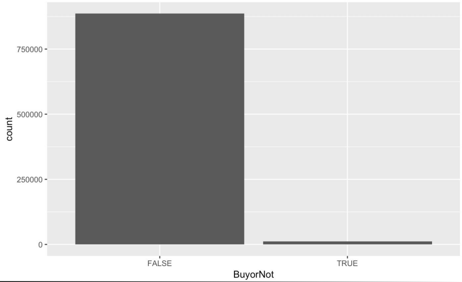
Model - Challenges

High Dimension

- 903653 observation;
- 55 columns
- Integer, json, char, boolean

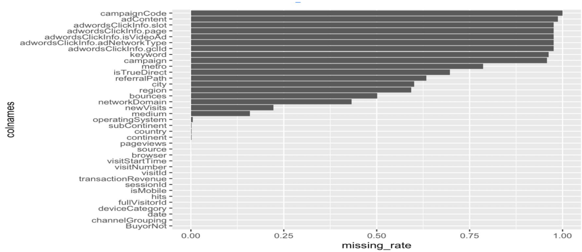
Highly imbalance

- 98.7% observation is without purchase behavior

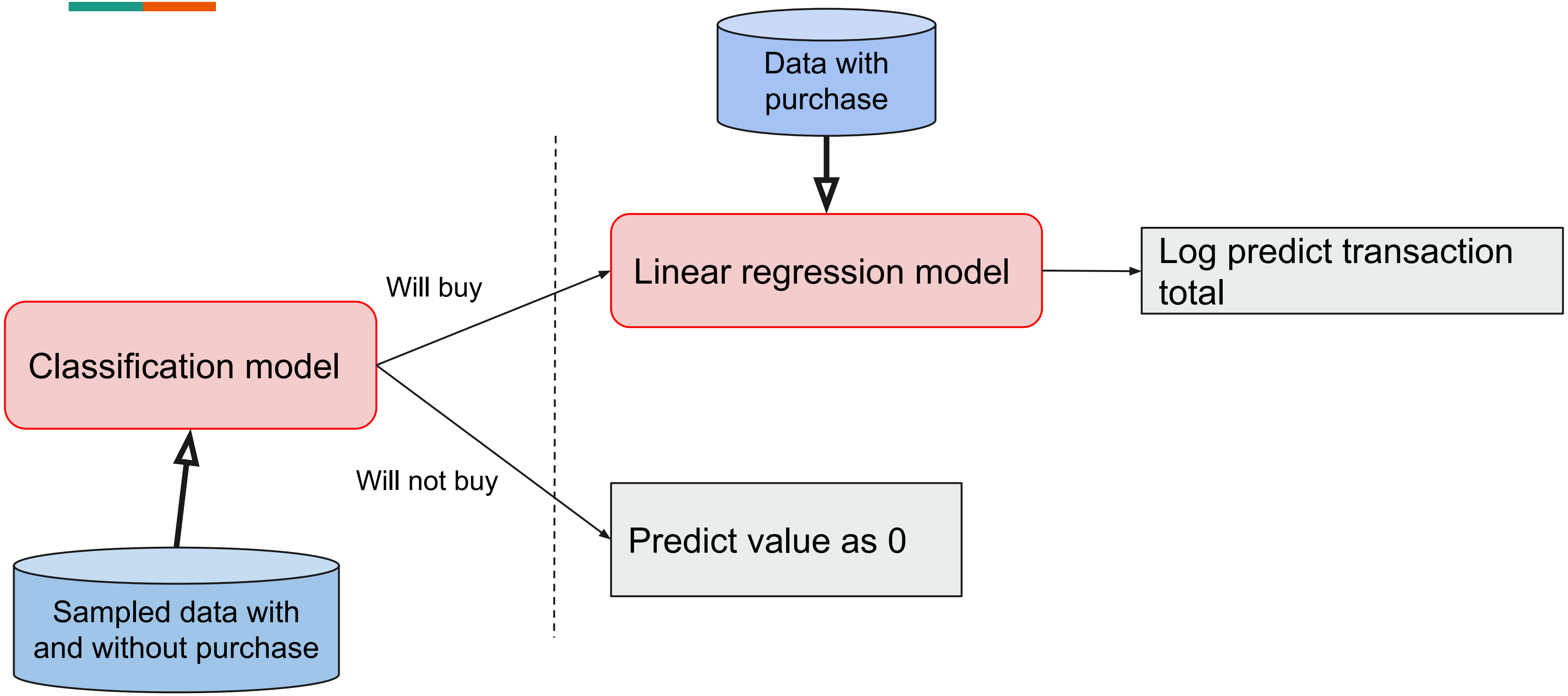


Missing value

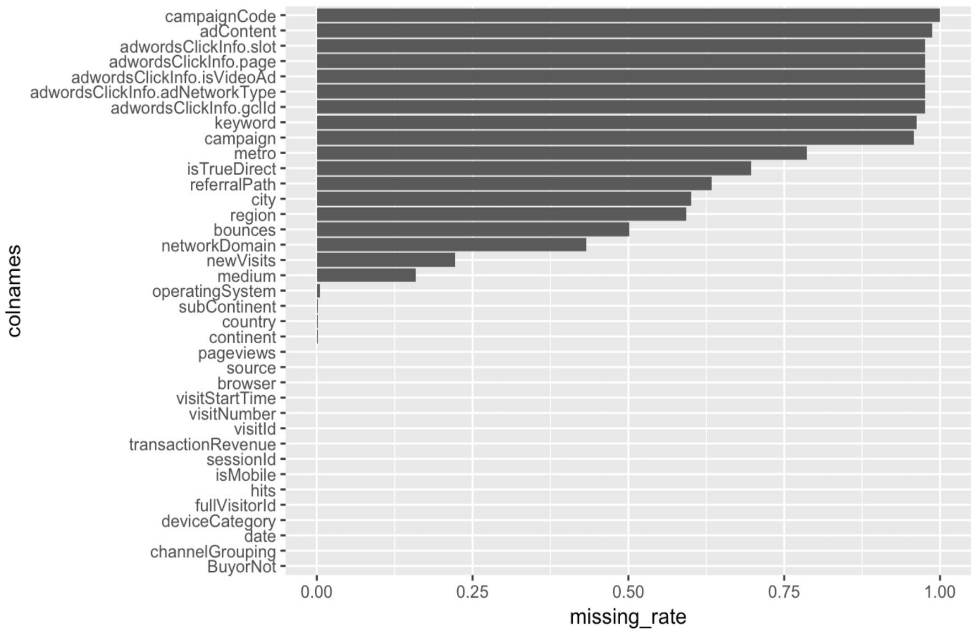
- Over ⅓ features have more than 50% missing value



Problem formulation



Data Prepare for Modeling (missing value)



1. **Delete variable with more than 10% missing value**
   ie. campaignCode, adContent

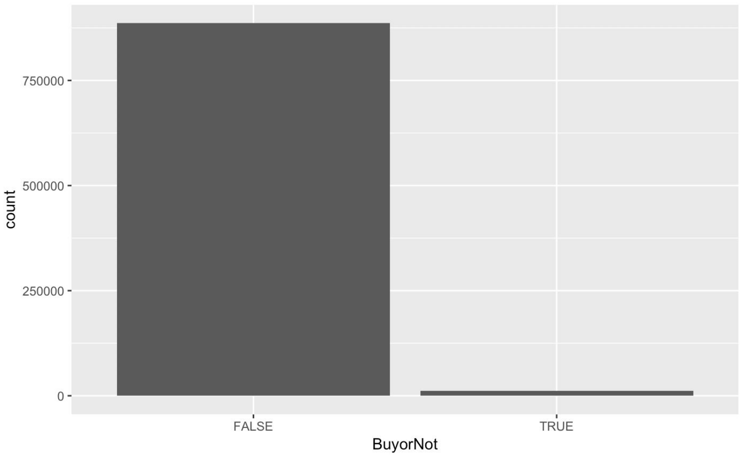
2. **Remove observation with missing value**

3. **Remove column containing no valid information for fitting model**
   ie. userId, sessionId

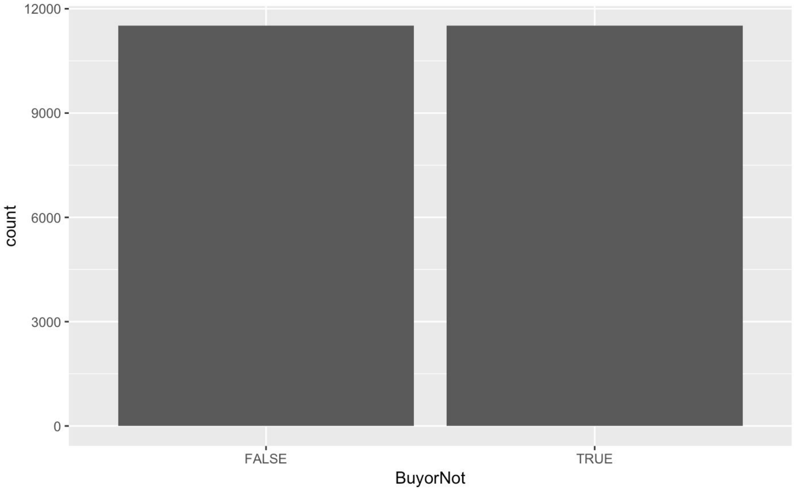
4. **Remove columns with duplicated information**
   ie. continent, sub-continent

Classification

Challenge Imbalance



Solution: Down Sampling



Reason - Enough number of data

Classification - Fitting model

Model setting

1. training 80%
2. test 20%
3. threshold: 0.3
   care more about customer with purchase

Model fitting

```
set.seed(123)
glm_dataset <- resample_partition(sample_dataset_balance_forglm, c(train = 0.8,test = 0.2))
glm_dataset$train <-as.tibble(glm_dataset$train)
glm_dataset$test <-as.tibble(glm_dataset$test)
fit_logit <- glm(BuyorNot ~., family=binomial(link="logit"), data=glm_dataset$train)
```

Accuracy

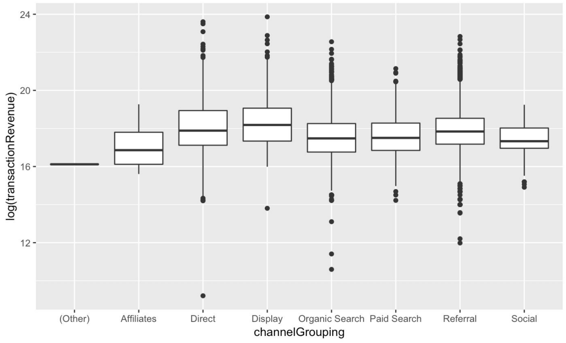
```
mean(pred_fc == glm_dataset$test$BuyorNot, na.rm=TRUE)
```

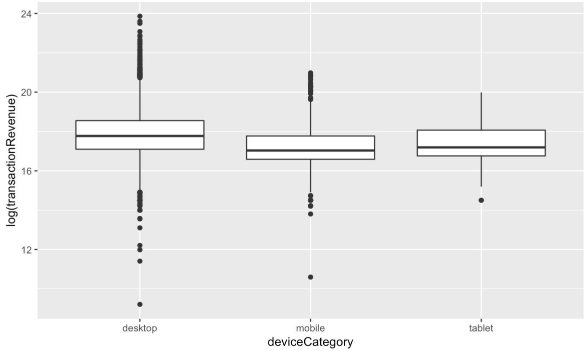
```
[1] 0.9554735
```

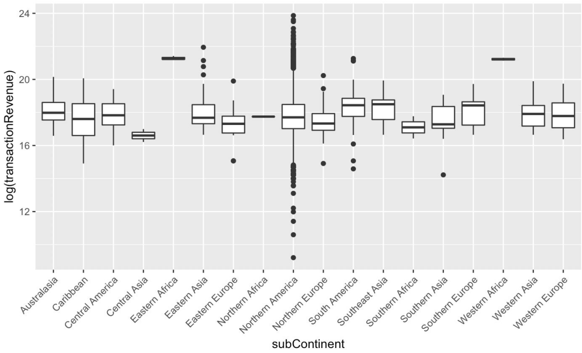
Feature significance

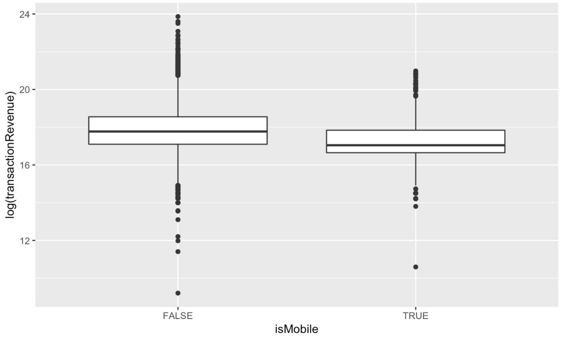
- **subContinentNorthern America**
  3.741e+00 8.555e-01 4.373 1.23e-05 ***
- operatingSystemiOS
  5.507e-01 2.085e-01 2.642 0.00825 **
- subContinentCaribbean
  3.191e+00 1.114e+00 2.865 0.00417 **
- subContinentCentral Asia
  3.855e+00 1.867e+00 2.065 0.03895 *
- subContinentEastern Africa
  3.925e+00 1.497e+00 2.622 0.00873 **
- **subContinentNorthern America**
  3.741e+00 8.555e-01 4.373 1.23e-05 ***
- **hits**
  -1.946e-01 9.405e-03 -20.688 < 2e-16 ***
- **pageviews**
  5.484e-01 1.460e-02 37.552 < 2e-16 ***

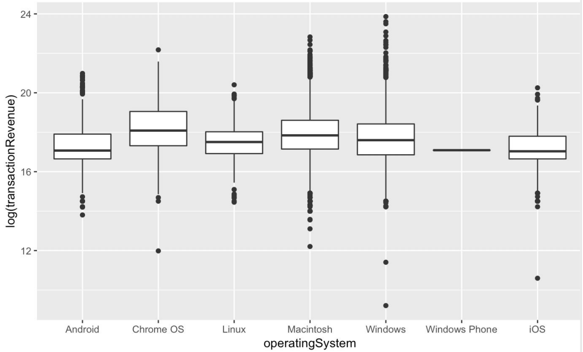
Regression Variable Analysis (Categorical)

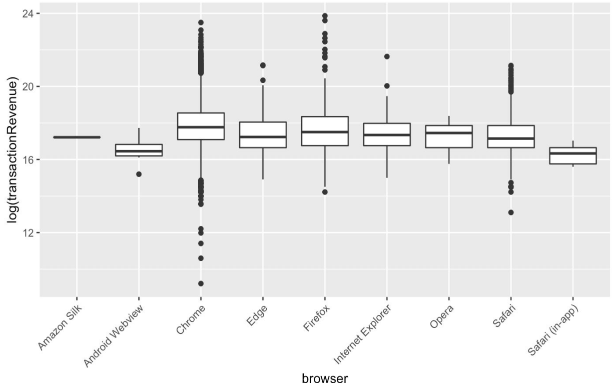




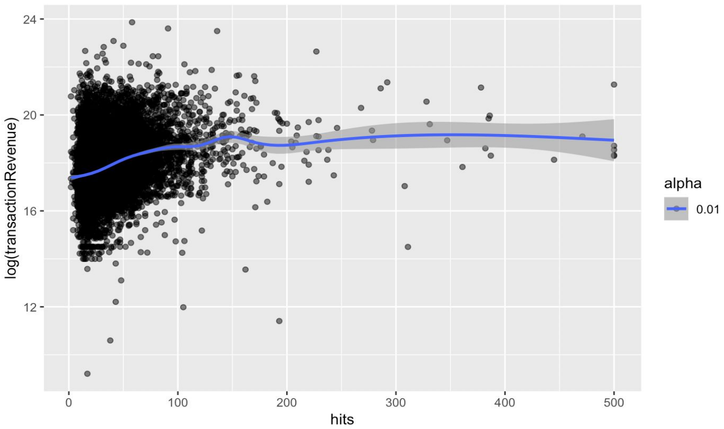


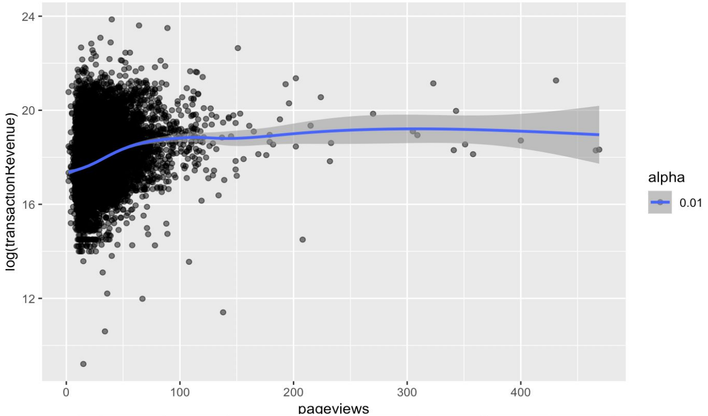


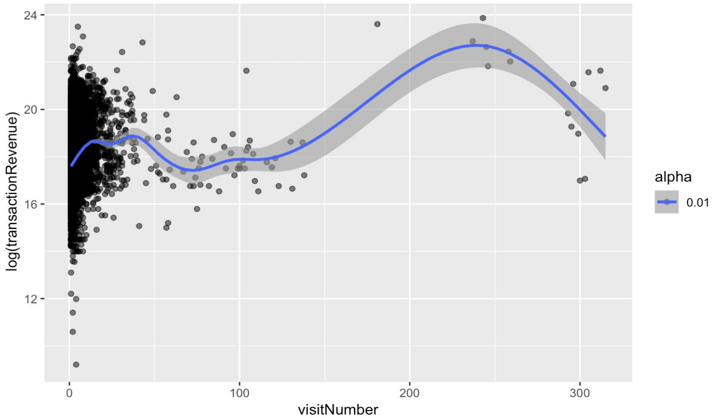


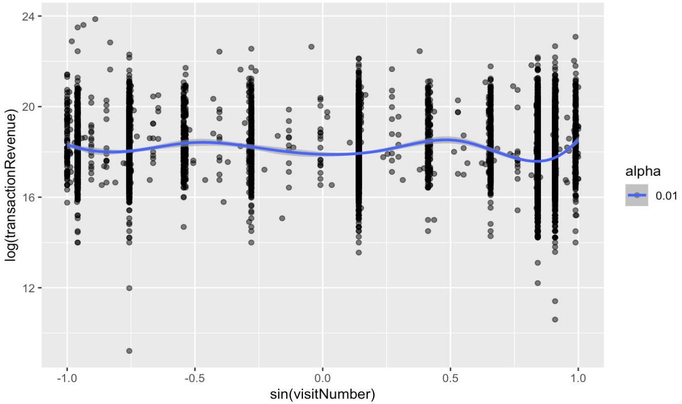


Regression Variable Analysis (Continuous)









Regression - Fitting Model

Model setting

1. training 80%
2. test 20%

Model fitting

```
set.seed(123)
reg_dataset <- resample_partition(sample_dataset_balance_forreg,c(train = 0.8, test = 0.2))
reg_dataset$train <- as_tibble(reg_dataset$train)
fit_reg <- lm(log(transactionRevenue) ~.,data = reg_dataset$train)
```

RMSE Analysis

rmse(fit_reg, reg_dataset$train) → 1.097742
rmse(fit_reg, reg_dataset$test) → 1.112894

Feature significance

hits

0.025828 0.001997 12.935 < 2e-16 ***

pageviews

-0.022655 0.002797 -8.100 6.22e-16 ***

visitNumber

0.012128 0.000998 12.153 < 2e-16 ***

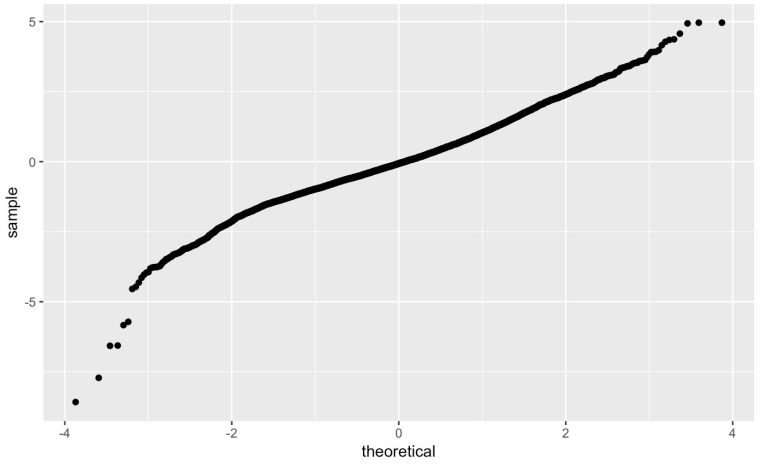
subContinentEastern Africa

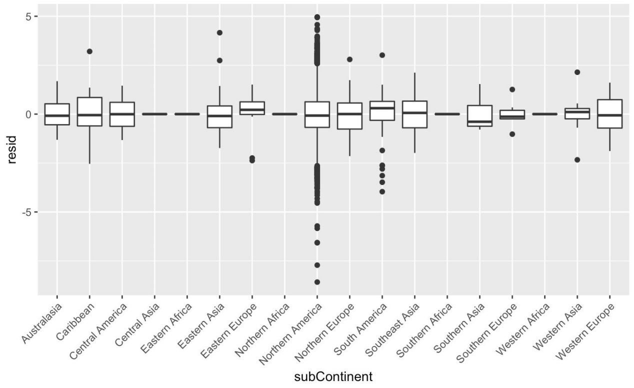
3.125614 1.142387 2.736 0.00623 **

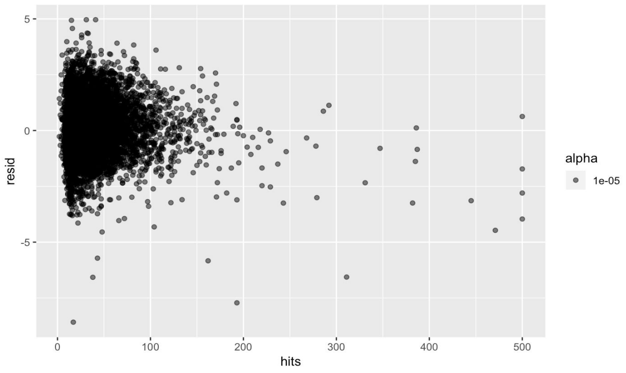
subContinentWestern Africa

2.962579 1.142188 2.594 0.00951 **

Residual Analysis







Conclusion and Future Work

Conclusion:

- More PVs and Hits brings more revenue
- Metropolis generates more PVs and Hits than countryside
- Hits, pageview, and visitnumber are useful on both classification and regression model
- The weekly and monthly pattern of visits and revenue may help create an accurate advertising plan
- Devices could be a potential revenue point
- Gstore still needs to increase its brand awareness

Future work:

1. Check the PVs and Hits in different devices over the period
2. Check if there is a potential to increase the PV and hits from other areas(Not Americas)

Model side:

1. Use regularization to find out the most important feature.
2. Include time as feature and implement time series model