# FINAL PROJECT

Predicting Conversion based on Search Attributes for a Travel Site

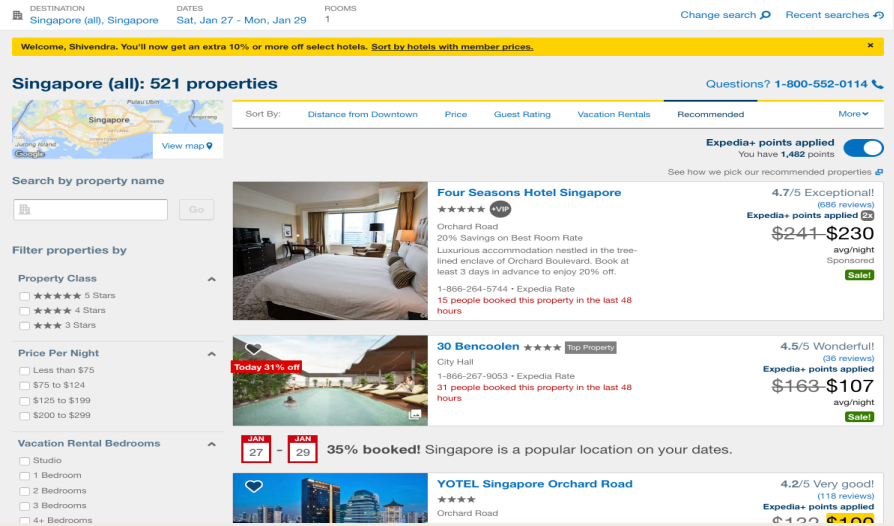
-- Shivendra Kishor (DAT-SF-40)

# Hypothesis:

- Predicting a conversion of a search using search parameters available on the travel website

  OR

- How likely a customer is going to book an hotel based on his search parameters from his/her search session attributes

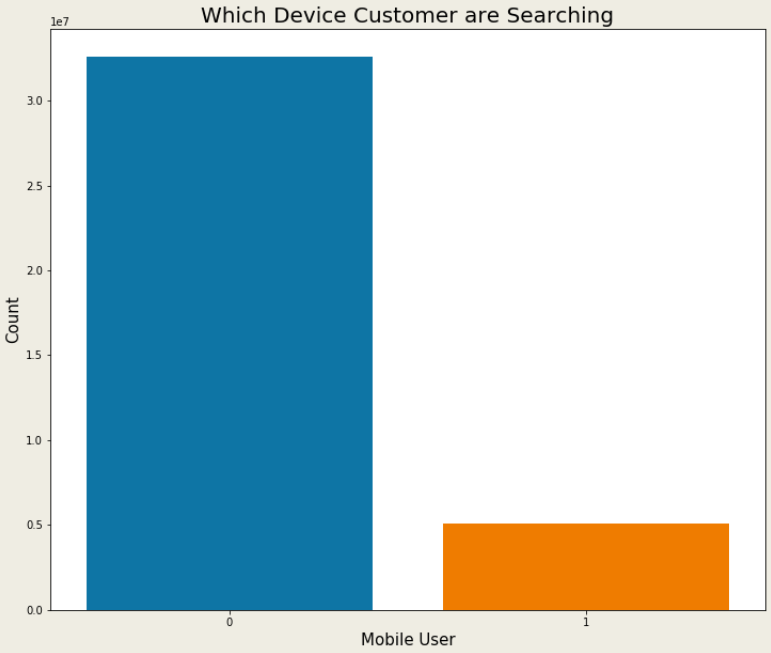


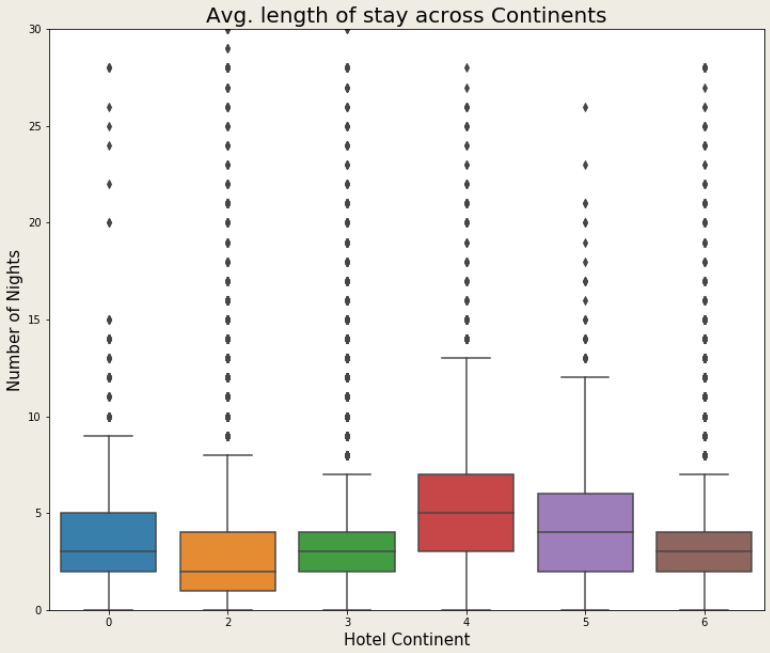
# Data exploration: What Type of data we have

| Column name | Description | Data type |
|---|---|---|
| date_time | Timestamp | string |
| site_name | ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...) | int |
| posa_continent | ID of continent associated with site_name | int |
| user_location_country | The ID of the country the customer is located | int |
| user_location_region | The ID of the region the customer is located | int |
| user_location_city | The ID of the city the customer is located | int |
| orig_destination_distance | Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated | double |
| user_id | ID of user | int |
| is_mobile | 1 when a user connected from a mobile device, 0 otherwise | tinyint |
| is_package | 1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise | int |
| channel | ID of a marketing channel | int |
| srch_ci | Checkin date | string |
| srch_co | Checkout date | string |
| srch_adults_cnt | The number of adults specified in the hotel room | int |
| srch_children_cnt | The number of (extra occupancy) children specified in the hotel room | int |
| srch_rm_cnt | The number of hotel rooms specified in the search | int |
| srch_destination_id | ID of the destination where the hotel search was performed | int |
| srch_destination_type_id | Type of destination | int |
| hotel_continent | Hotel continent | int |
| hotel_country | Hotel country | int |
| hotel_market | Hotel market | int |
| is_booking | 1 if a booking, 0 if a click | tinyint |
| cnt | Numer of similar events in the context of the same user session | bigint |
| hotel_cluster | ID of a hotel cluster | int |

# Data exploration: Characteristics of a dataset

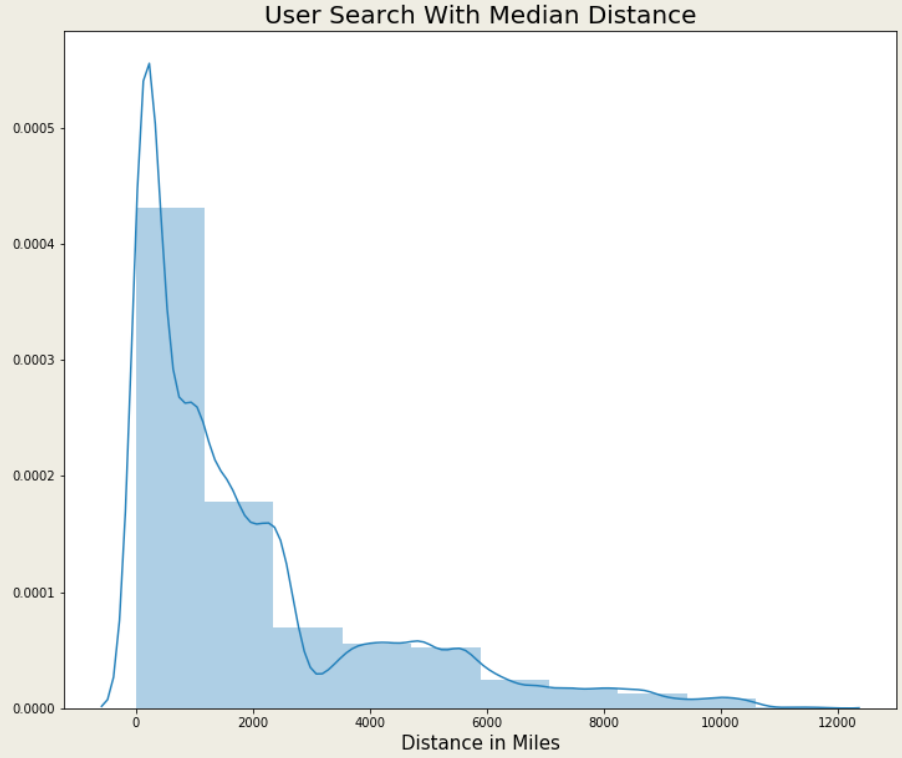
- Users are searching more on Desktop on Expedia site compare to Mobile
- Average length of stay is higher for Continent 4



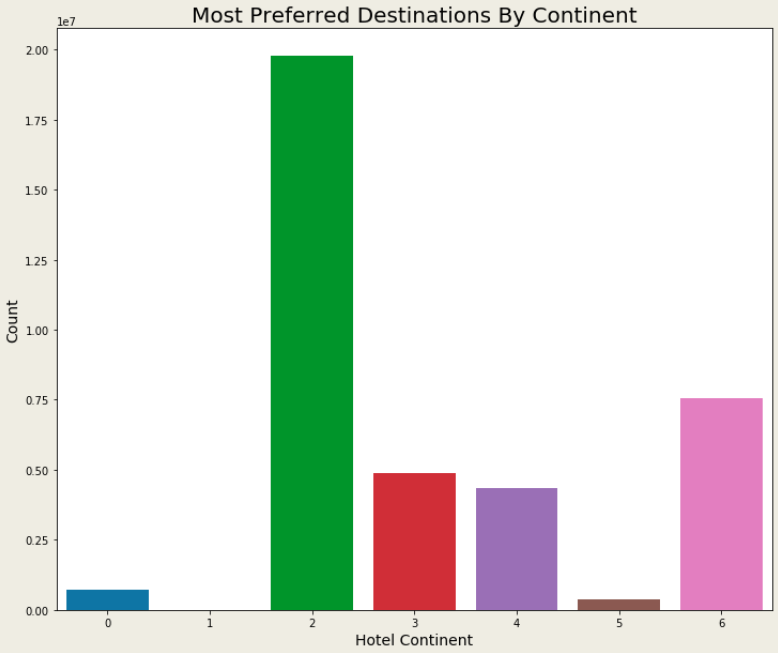
# Data exploration: Characteristics of a dataset

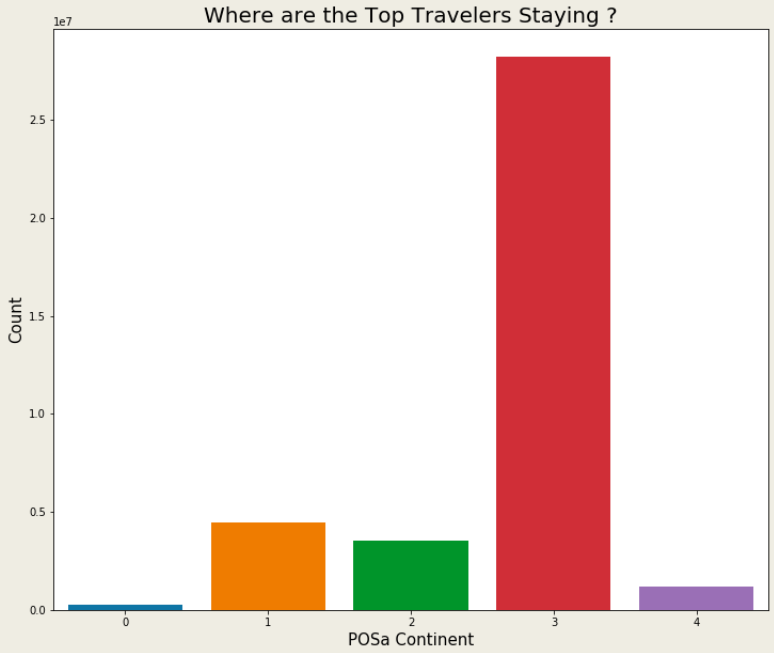
- Most of the users made a search with 2000 miles of their search location



# Data exploration: Characteristics of a dataset

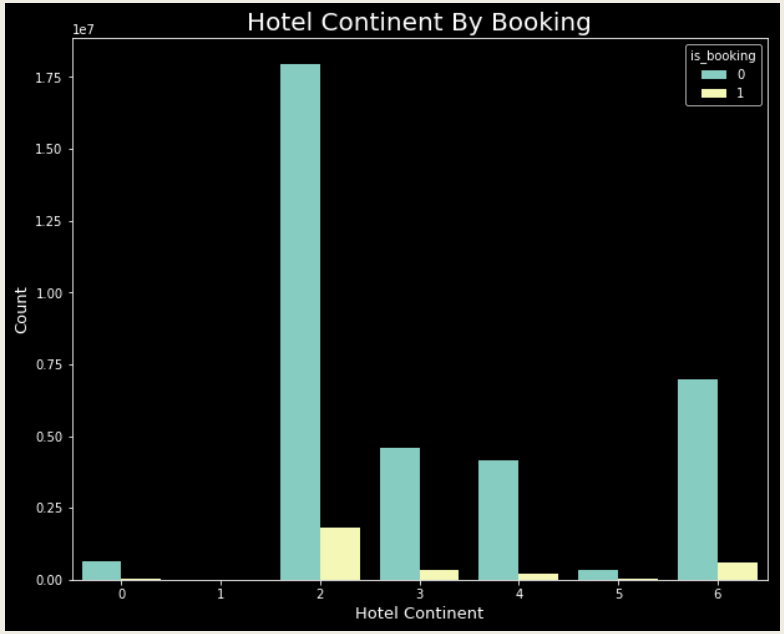
- Most preferred destinations for Customers are from Continent 2
- Point of sale is High for Expedia Site from Continent 3

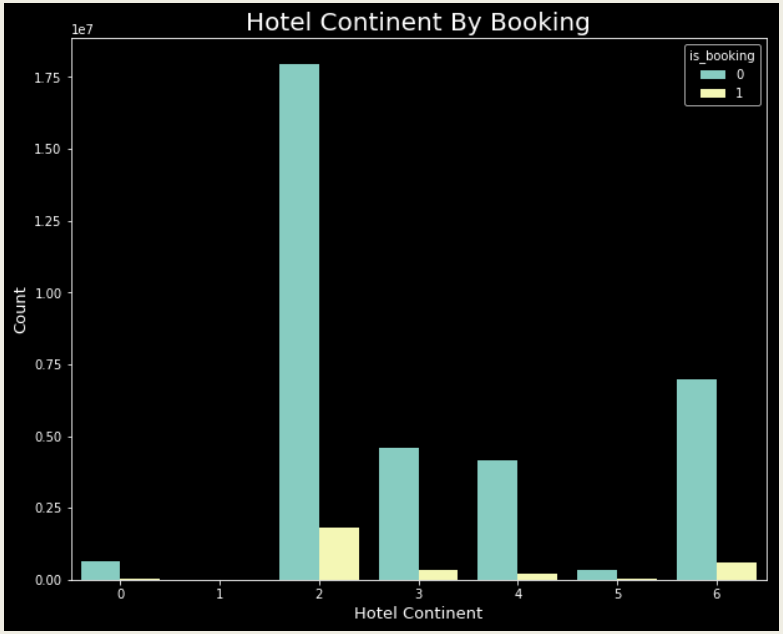


# Data exploration: Characteristics of a dataset

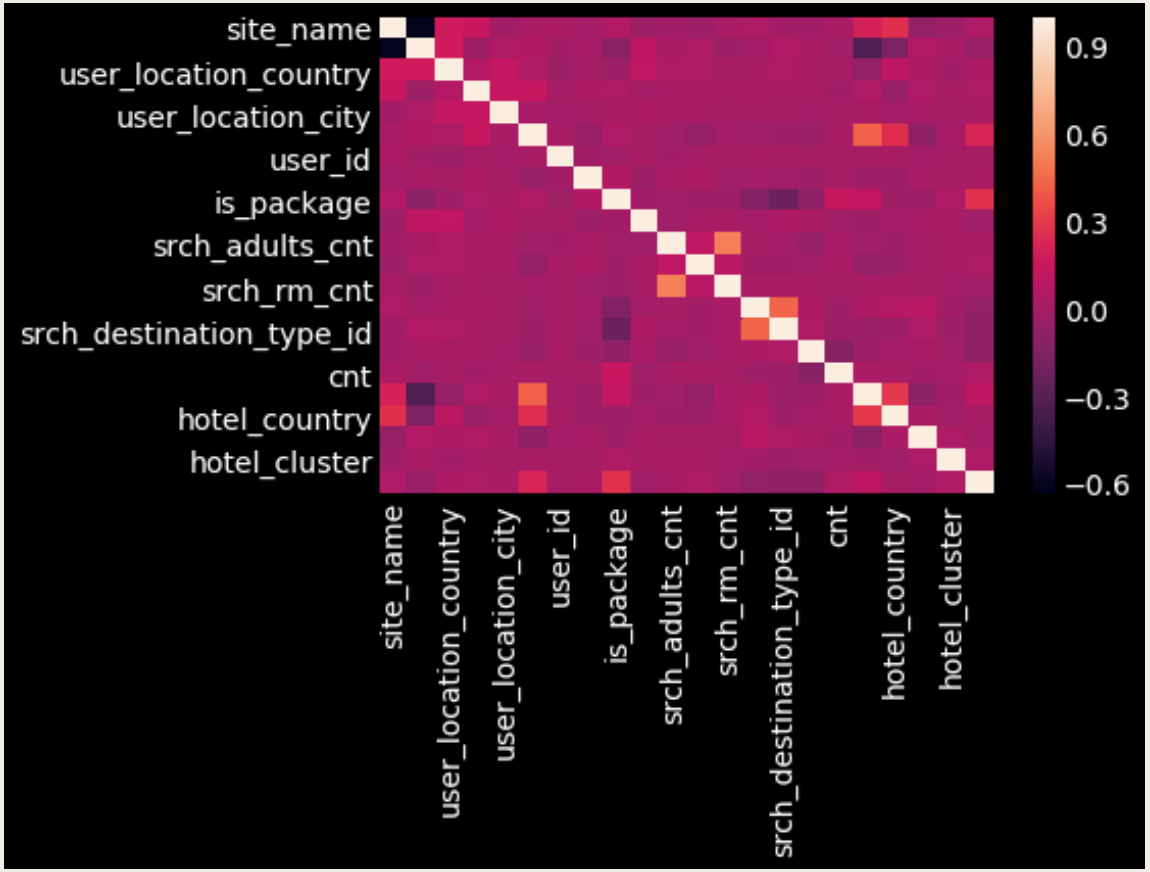
- Maximum Booking is happening Continent 2 and 6
- PoS Continent 3 has High booking event



# Data exploration: Characteristics of a dataset

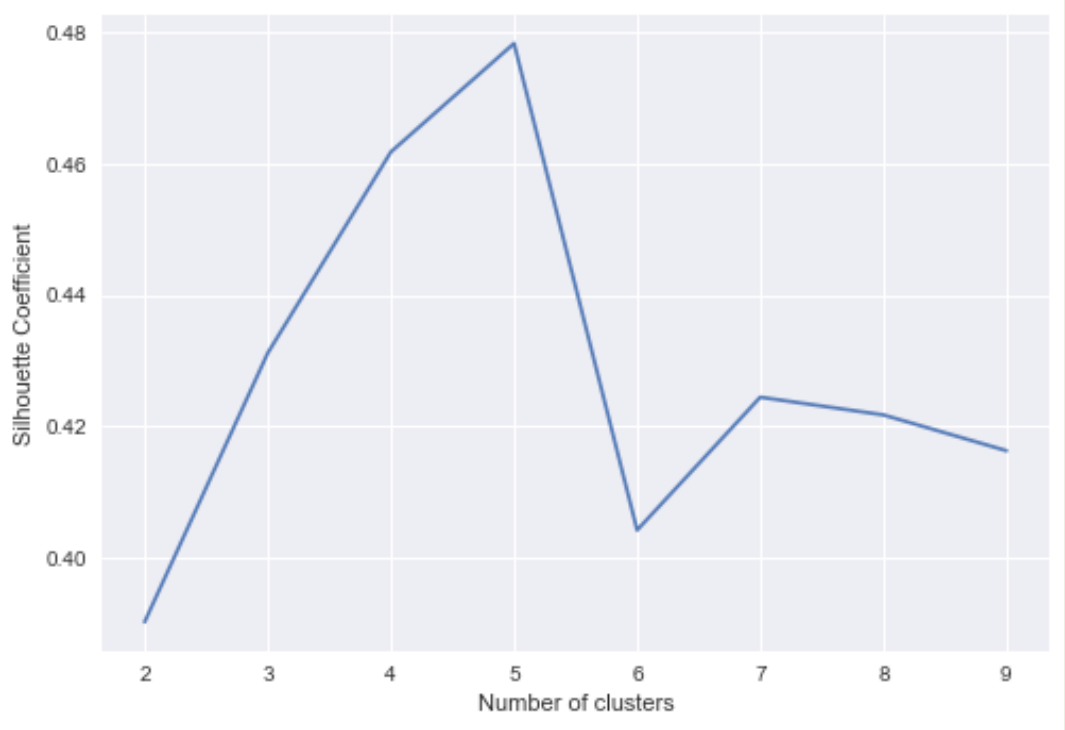
- Correlation :



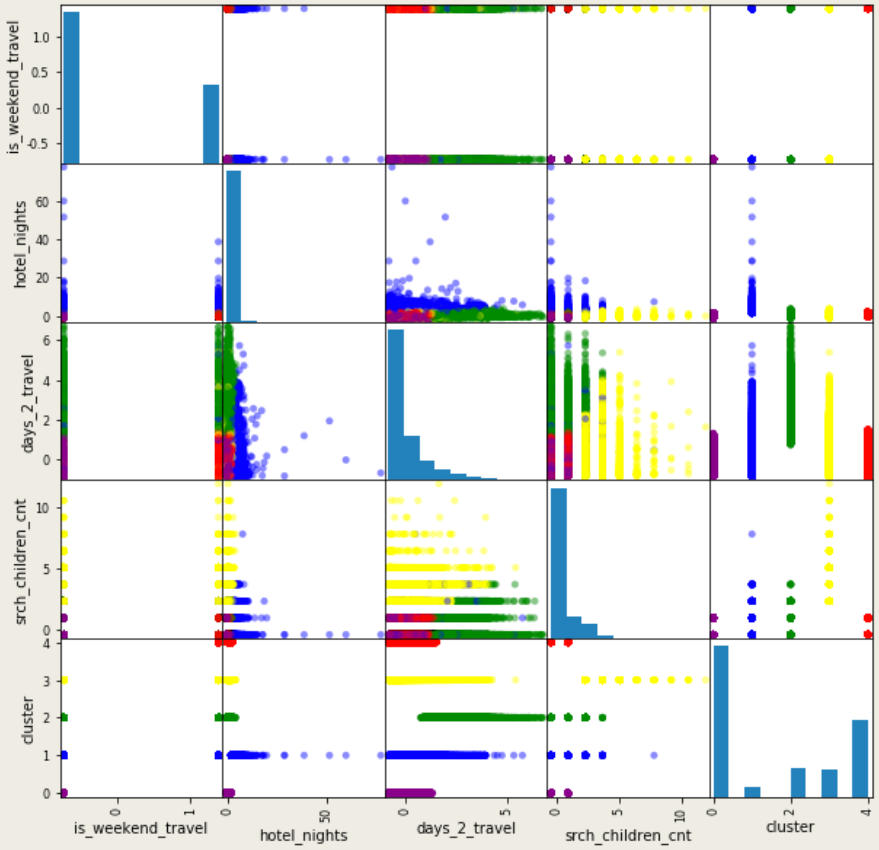
# Data exploration: Characteristics of a dataset

- Updated Null values associated with Origin Distance attributes with Mean value
- Dropped Null values for Model dataset
- Kept only the data with Night Spends at Hotel >0
- Kept data with Day to Travel based on search check-In date >0
- Created Indicator for Weekend travel yes now
- Creating Clusters using different attributes

# **Clustering:** using K means Algorithm



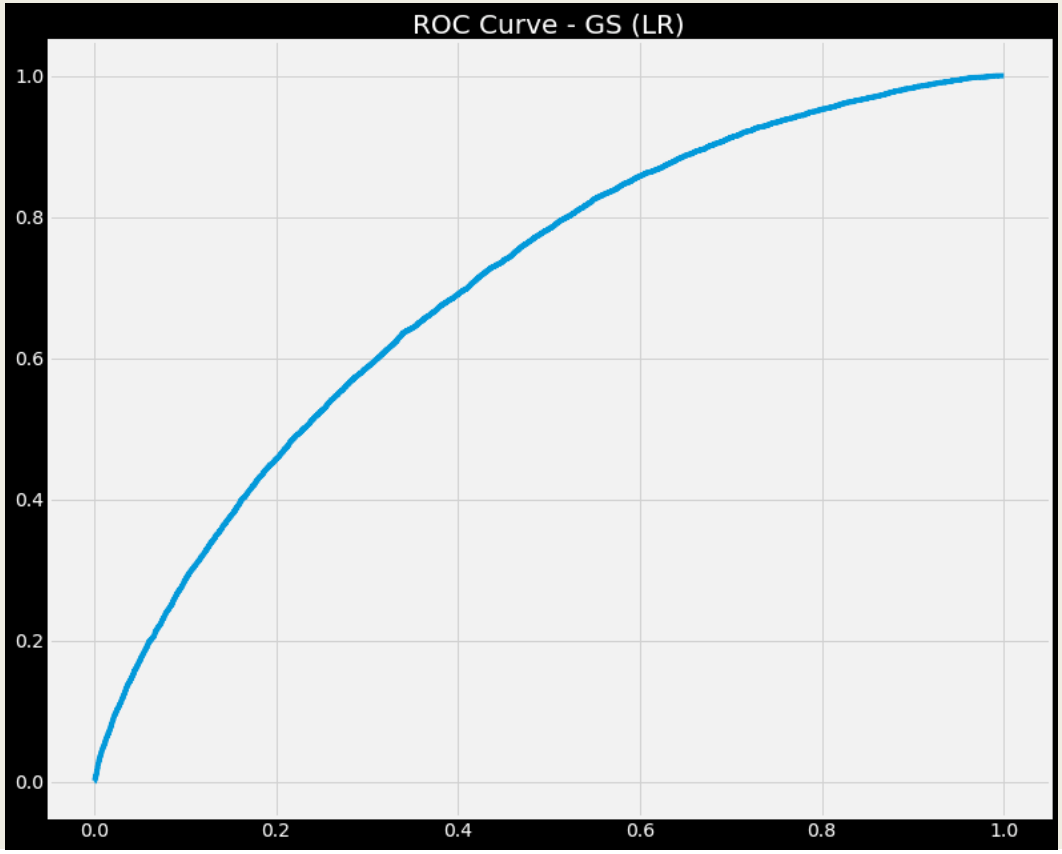
# **Clustering:** Scatter plot for clusters (Kmeans, n= 5)



# Models

- Logistic Regression (0. 0.688318722972)
- Ridge Classifier  (0.664997519063)
- Random Forest (0.632023711055)

# ROC Curve:



# Thank You!!!