

Expedia Hotel Recommendation

Website: <http://www.expediahotelrecommendation.com/>

Team 10

Vishwa Shah

Avinita Shetty

Jun Wu

Table of Contents

[Background 3](#_Toc449739764)

[**Challenge:** 3](#_Toc449739765)

[**Goals** 3](#_Toc449739766)

[**Dataset** 3](#_Toc449739767)

[Project Plan 4](#_Toc449739768)

[Project Report 5](#_Toc449739769)

[**Data Exploration** 5](#_Toc449739770)

[**Data Preparation** 8](#_Toc449739771)

[**Data Analysis by Visualizations** 9](#_Toc449739772)

[**Modeling** 21](#_Toc449739773)

[**Deployment** 24](#_Toc449739774)

# Background

## **Challenge:**

Which hotel type will an Expedia customer book?

Expedia wants to optimize hotel search by providing personalized hotel recommendations to their users. This is no small task for a site with hundreds of millions of visitors every month!

Currently, Expedia uses search parameters to adjust their hotel recommendations, but there aren't enough customer specific data to personalize them for each user. In the competition, Expedia is expecting Kagglers to contextualize customer data and predict the likelihood a user will stay at 100 different hotel groups.

## **Goals**

Based on the dataset provided by Expedia, we are planning to achieve the following goals by performing multiple ways of data analytics, and comparing then applying data mining models.

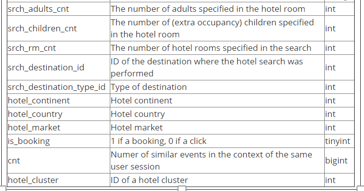
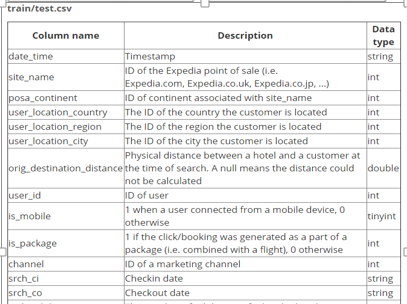
* Perform data exploration and extract meaningful information for business users
* Predict which group of hotels the customer performing the search would likely to book.
* Provide business users/ customers a user-friendly web access to observe data on web pages

## **Dataset**

Data Description:

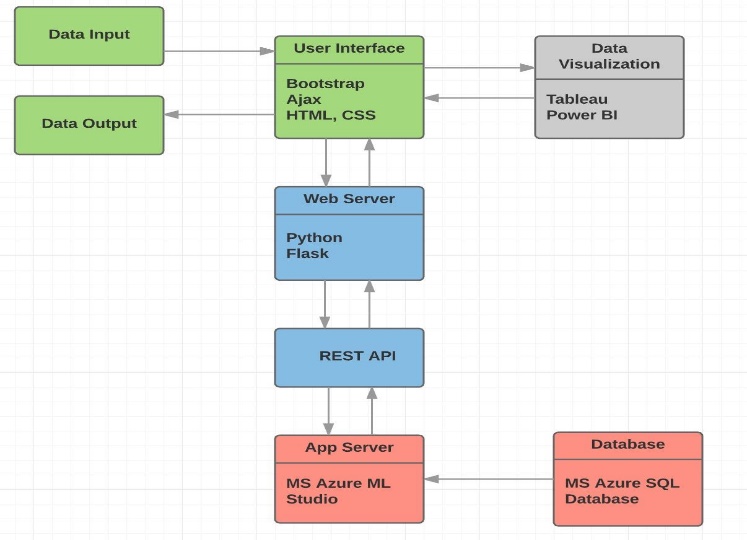
Logs of customer behavior (separate training and testing dataset)

Include customer profile(User ID, Geo location, search site, device), search event information (date, check-in/check-out day, destination info).



# Project Plan

The project has been developed through multiple phases. The major steps are listed below.

1. Data Exploration
2. Data Preparation
3. Data Visualization
4. Modeling
5. Web Deployment

Also, we followed our product architecture and achieved most of our functionalities when we finished our project. In the following sessions, detailed project breakdown which covers the actions and deliverables for each developing phase will be reported.

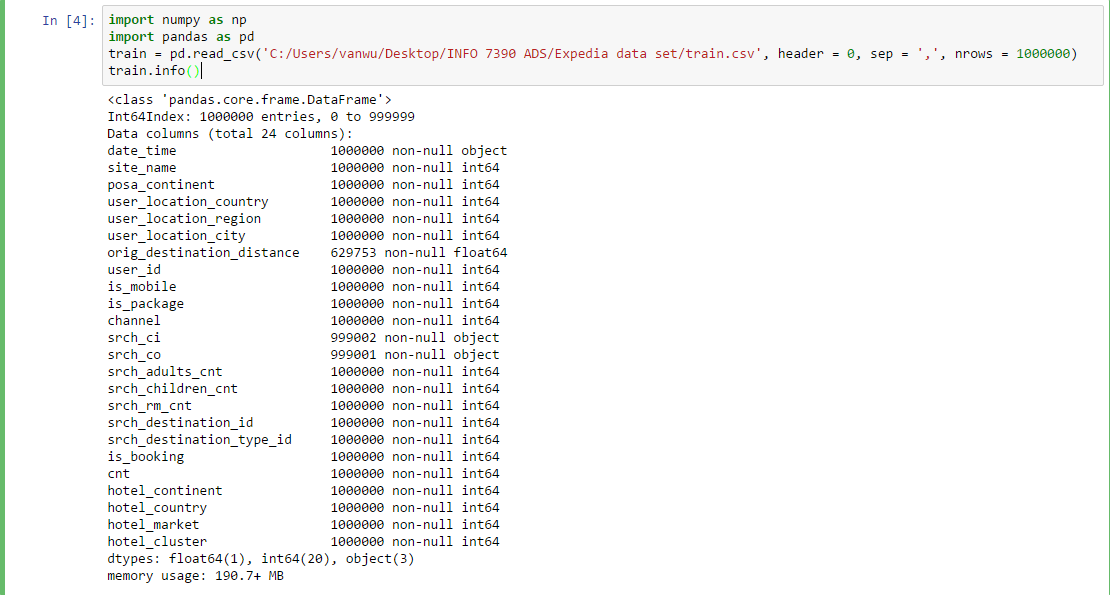
# Project Report

## **Data Exploration**

Tool: Python

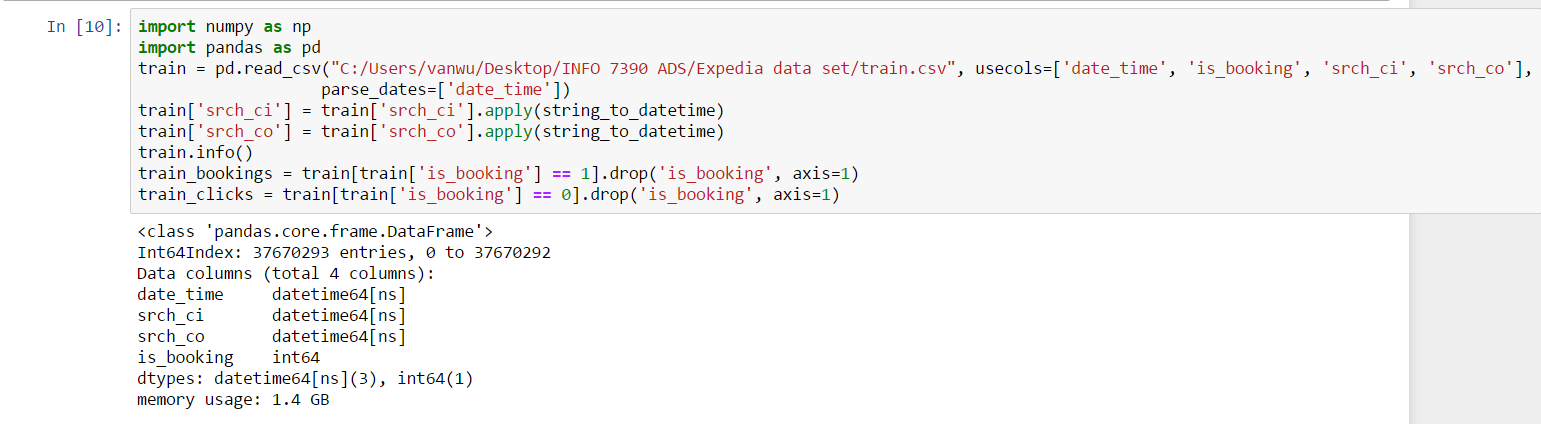
By using Python and leveraging libraries such as Numpy, Pandas, Matplot and Seaborn for data summarization and visualization, we managed to evaluate the data quality and understand our business problems in different aspects.

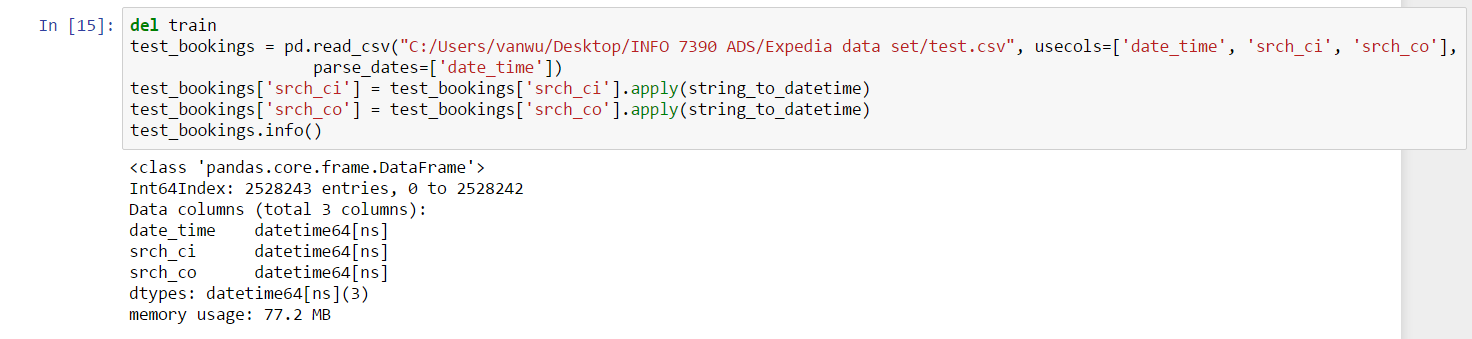
The dataset is sized over 3.5 gb, over 36 millions of rows. We extracted the first 1 million rows for and inspect the data quality using python for the purpose of efficiency. According to the result, 21 out of 24 columns has no missing values (regardless of improper values). For column srch\_ci and srch\_co about 0.1% data are missing. And for column orig\_destination\_distance, almost 40% percent of data are missing.

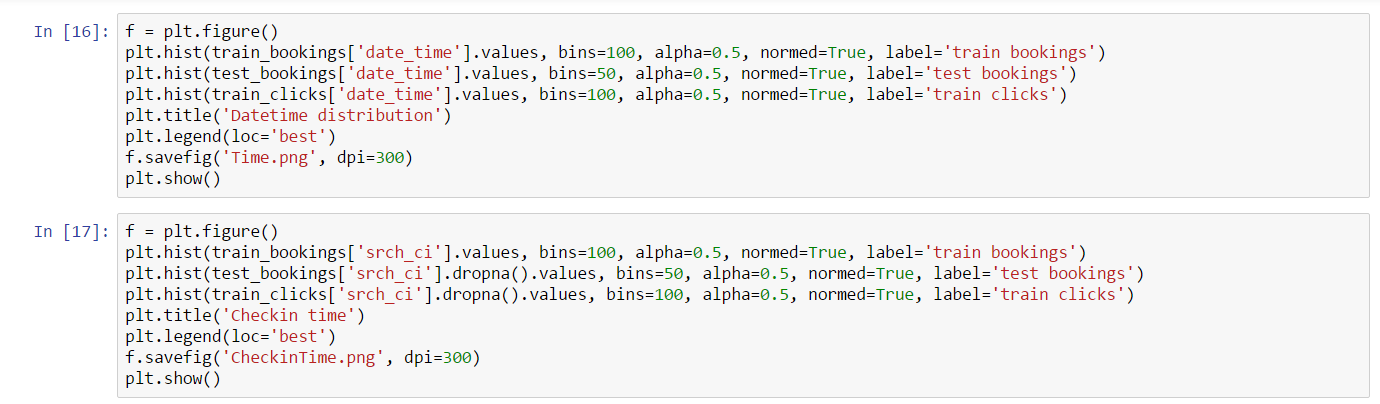


According to the data description, our data includes time-based customer events which may lead to a sales event (booking through Expedia website). In order to have a better understanding of their business situation, two charts has been developed for the convenience of observing the booking/clicking trends from 2013 to 2015.

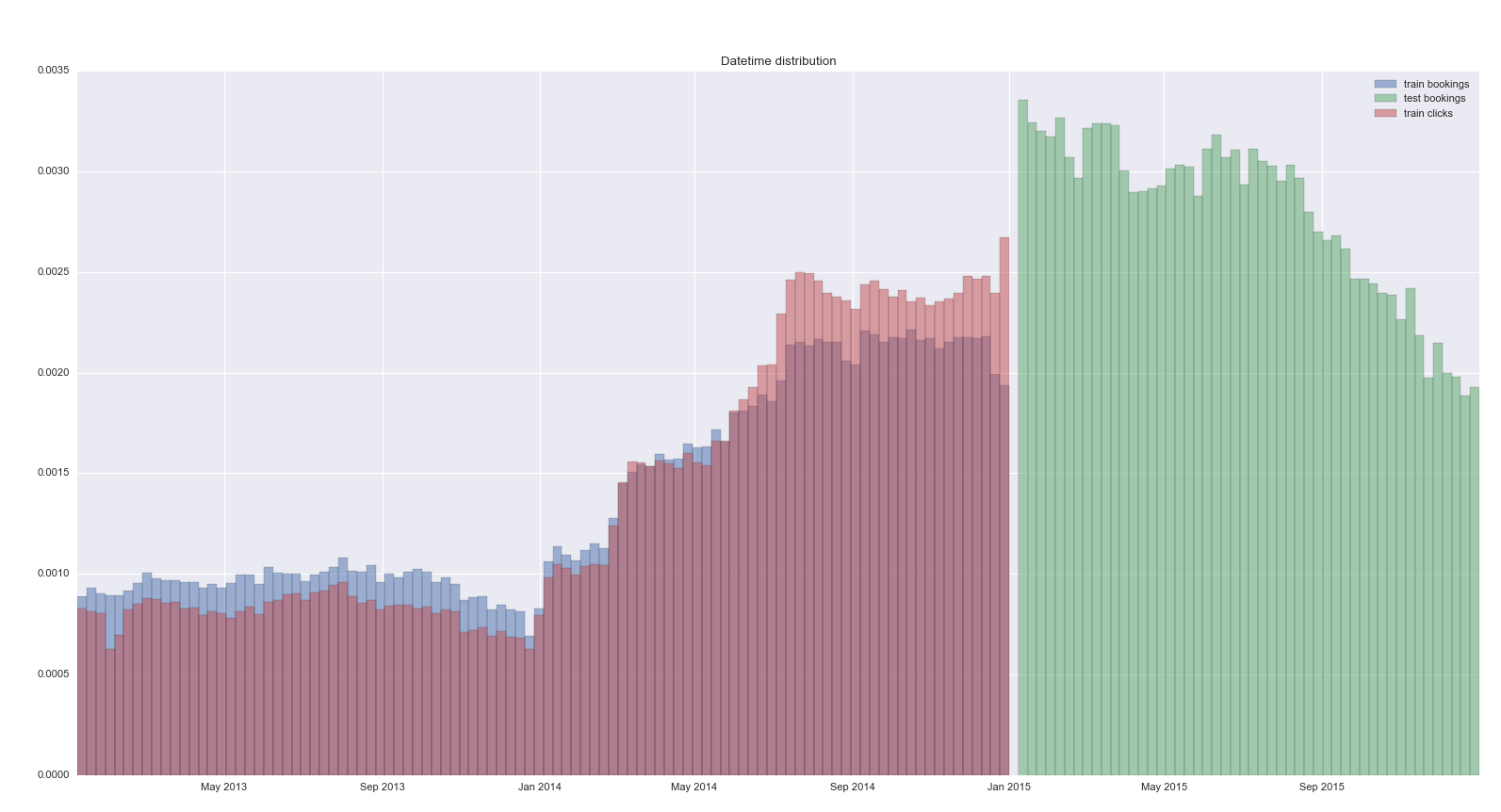


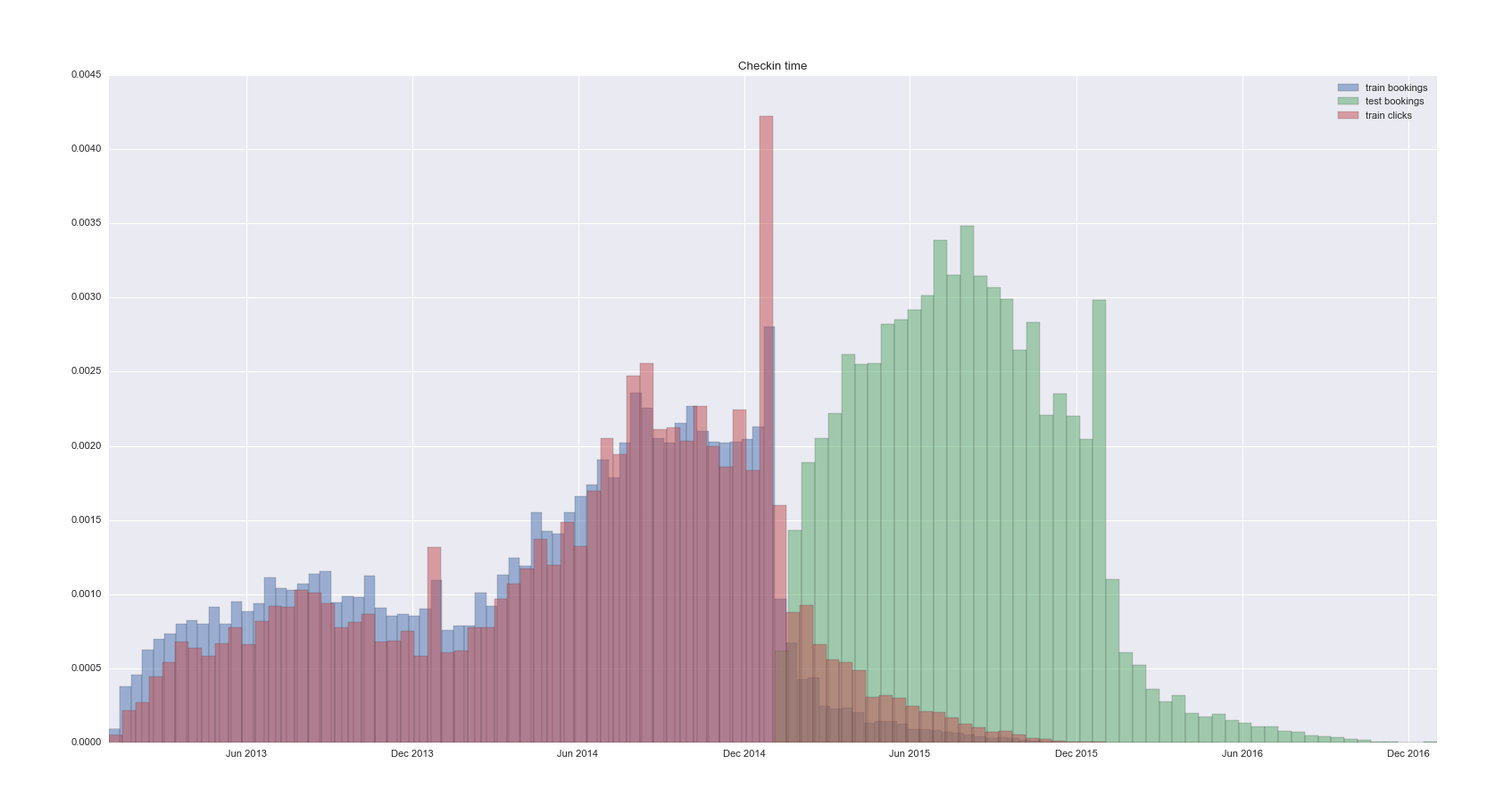


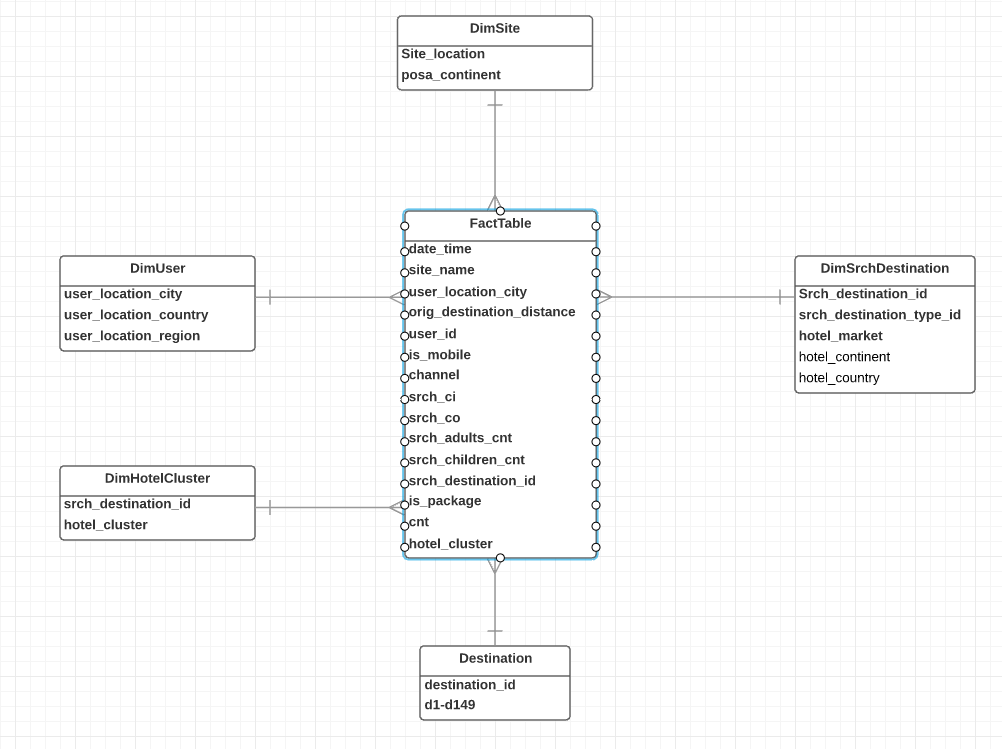




The chart below shows the booking number changes by time. The time frame was the date time when the customer events happen. Also, we observed that expedia business had a fast growth in 2014 and 2015.



The chart below shows customers’ travelling trends over time. The time frame is the hotel check-in dates the customers booked or searched. We found that the majority of travelling, which is indicated by hotel check-in happens during summer and December, which fits the reality of hotel industry.

Also, by reviewing each column of the first dataset, we noticed the hierarchical relationships of five groups of variables. We decided to build a dimensional data model for two reasons: First, our dataset has a large number of rows, by reducing the columns containing redundant values, data size can be optimized, which will improve the efficiency of data storage and visualization. Second, with the variable which represents the lowest granularity of each hierarchy, we will be able to build our classification model with less features but same accuracy.

By the end of data preparation, we managed to gain a good understanding of our dataset. Also, we planned out our actions in second phase - Data Preparation.

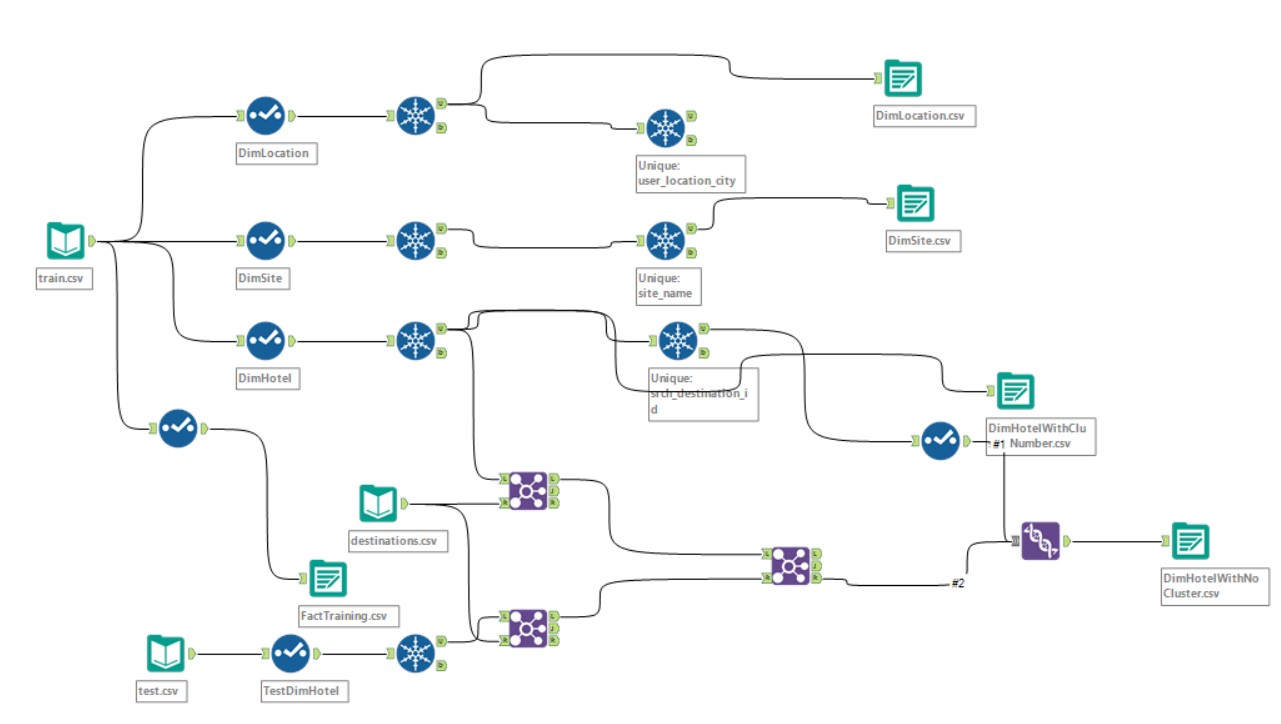
## **Data Preparation**

Tool: Alteryx

In this step we managed to perform data cleaning and transformation to convert our dataset into a star schema for the convenience of visualization and analysis.

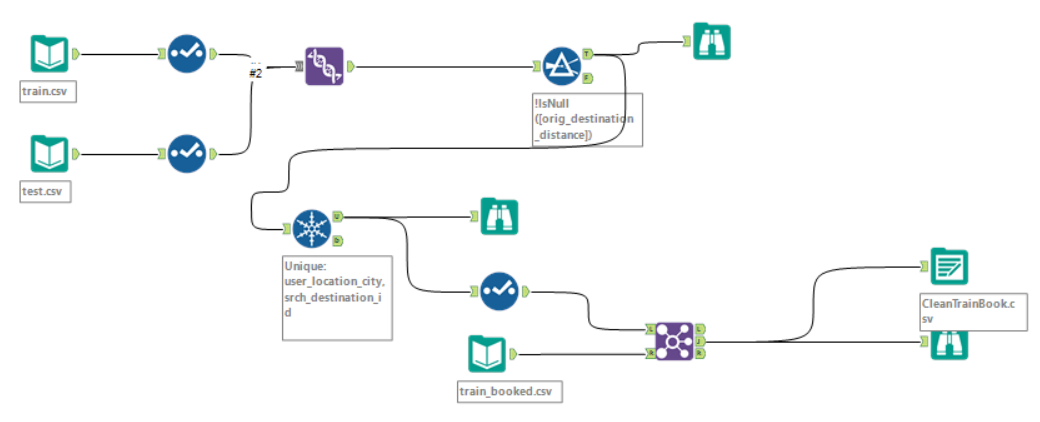
Using the data preparation tool Alteryx, we were able to build up pipelines to perform multiple data transformation, manipulation process step by step, and by the end, generate our data output.

There are a variety of widgets used in the pipeline. To build each dimension table, we first select the columns belonging to the hierarchy, followed with filtering out duplicated values. For the fact table which still contains every individual customer event, we kept the columns which act as a foreign key and dropped the columns that can be reconnected using these foreign keys. In the special use case of handling the new search\_destination which cannot be found in the reference table, we managed to filter out these rows by joining the result of unique search\_destination list from both training and testing dataset with their reference table.



About missing value handling, we deleted the rows with missing values in search\_ci and search\_co columns, the row number that has been dropped was less than 0.1% of the entire dataset, which is acceptable.

For Orig\_dest\_distance due to a large amount of missing values (about 40%), as well as the lack of geographical information, it was difficult to estimate missing value by using statistical methods. As a result, we decide to only fill in the values whose starting point (user\_location\_city) and destination point (search\_dest\_id) are known from our existing datasets.



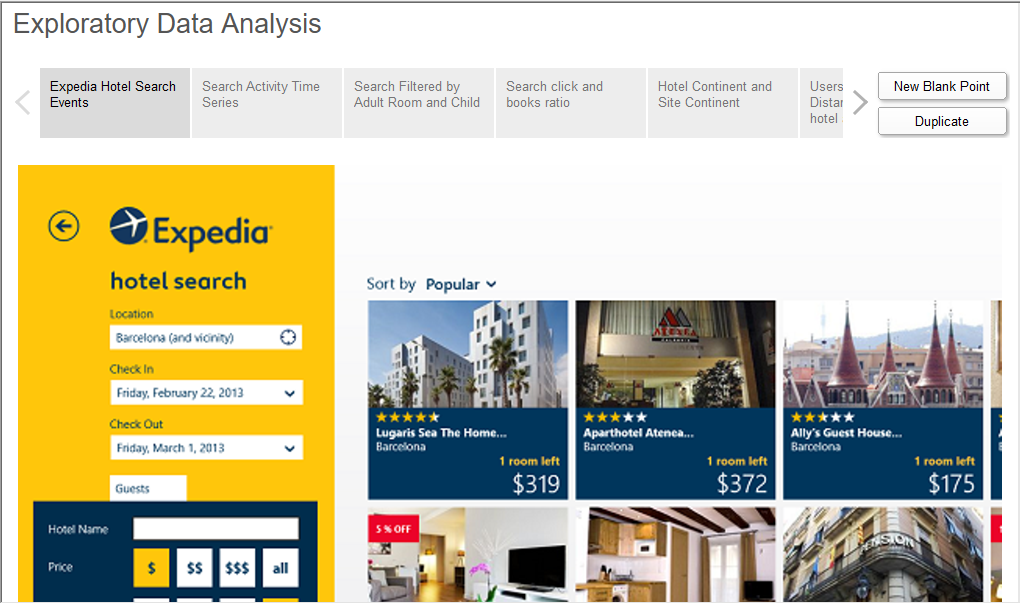
By the end, we were able to fill in about 20 percent of the missing data. For the rest of the missing values, since we have a large dataset and we expect the rows with no missing values are enough for modeling purpose, we decided to drop the rows with missing values.

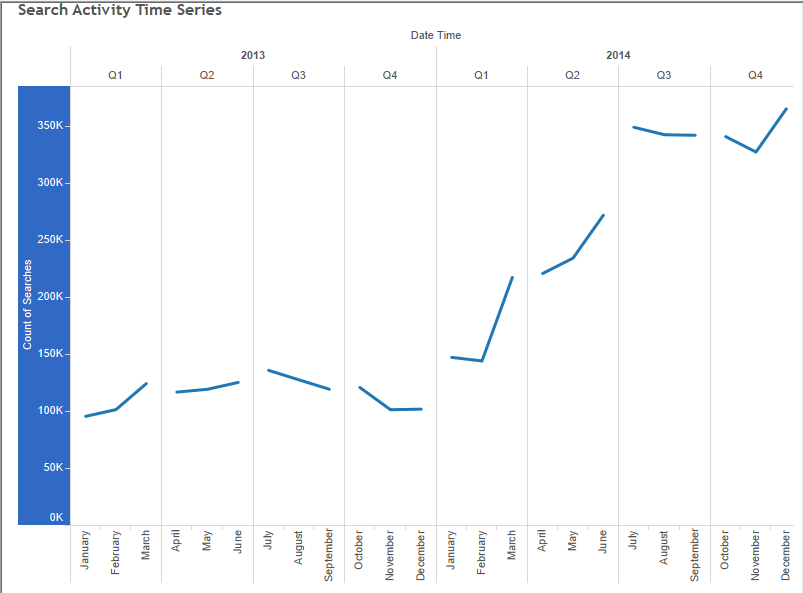
## **Data Analysis by Visualizations**

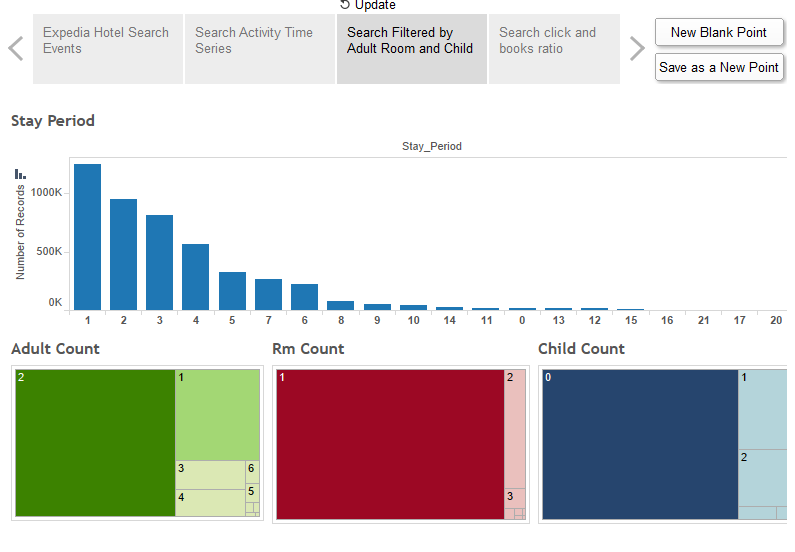
Tools: Tableau, Power BI

For the purpose of understanding and making sense if the Data we have done visualizations using Tableau and Power BI. In Tableau, we covers the following user stories.

* Exploratory Data Analysis
* User Analysis
* Hotel Analysis





This above time series is for the overall user searches over the years 2013-14. It shows how the overall searches increased for the year 2014 and mostly toward the Q3 and Q4 period.

The above dashboard shows maximum no of searches according to no. of Rooms, Adult Count and Children Count. It also shows what is the most probable period of stay the user would enter(Stay Period is calculated using the check-in and check-out days the user entered while searching)

On the above analysis we find out that most users commonly recorded searches are for:

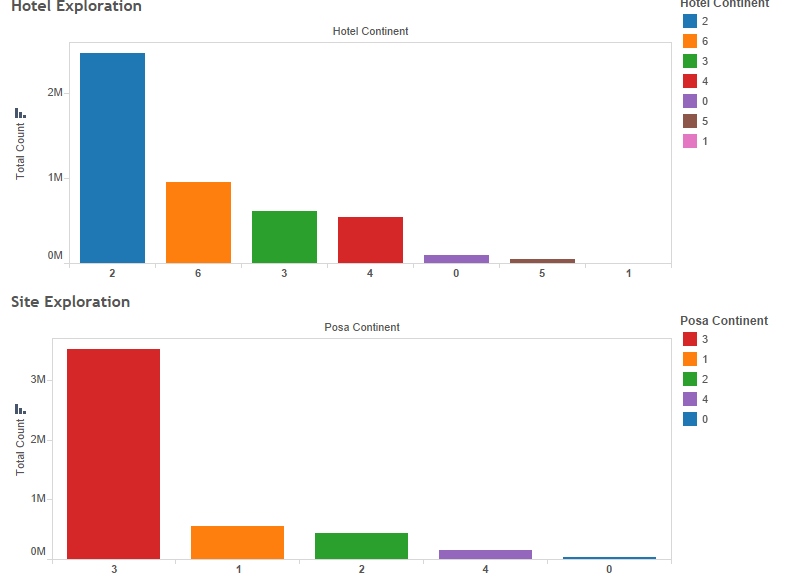
Adult count: 2 or 1 adult/adults

Room count: 1 or 2 rooms

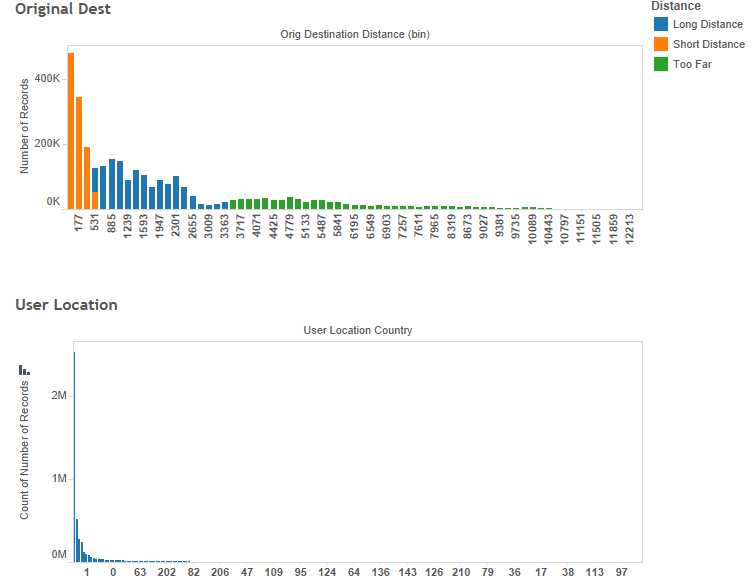
Child count: 0 and Max no. of stay is 1-2 day/days



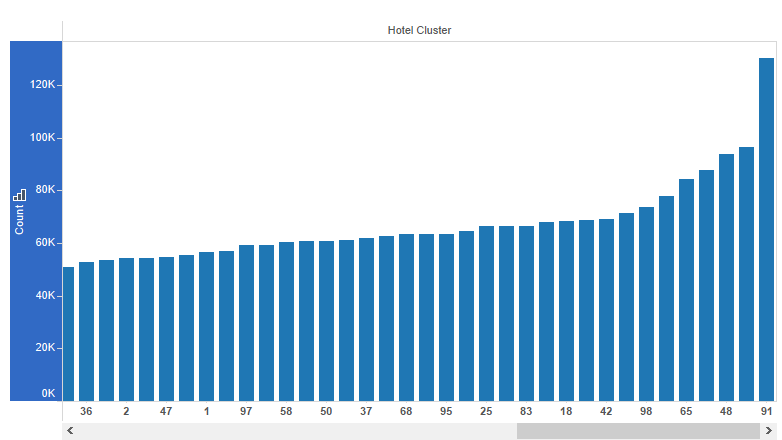
The above dashboard shows the click versus the book searches, which means which searches were made to actually book a hotel versus which were used to just browse through the hotel. Users may use Mobile vs other devices for their searches. Also the users may look for packaged deals as opposed to just hotel reservation. The above dashboard describes exactly how many users checked out packaged deals and how many actually booked it. Also the overall clicks are far greater than the bookings made.



The above dashboard shows no. of records in each continent. The Hotel continent stand for the searches made for hotel in those continents. Posa Continent stands for continents for which there are different site names of Expedia via which the search has been done.

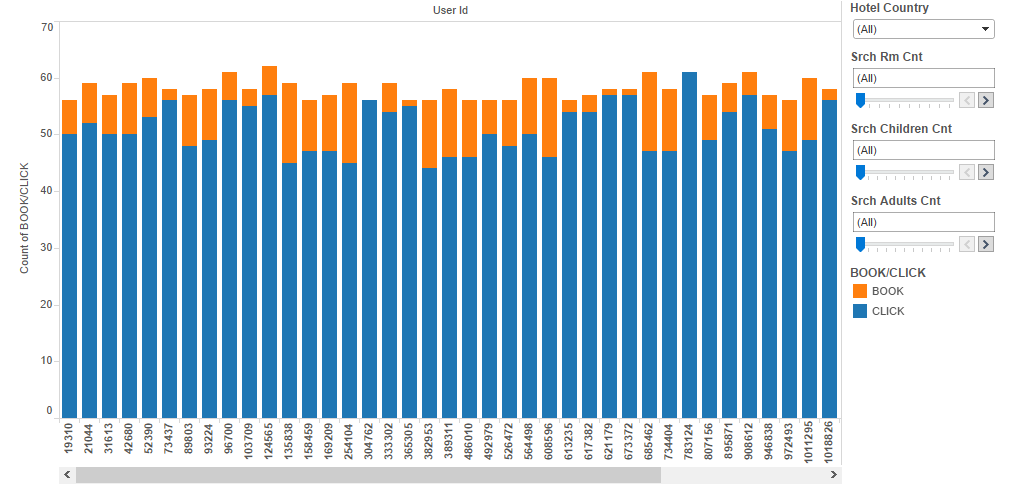


This dashboard shows the distance of the user from the searched destination. We have categorized it to 4 categories for simplicity purpose: Short Distance, Long Distance, Too far and not computable. We can see maximum no of searches by that particular user made as per the distance type and also in the second dashboard we can see which countries have maximum no. of users.

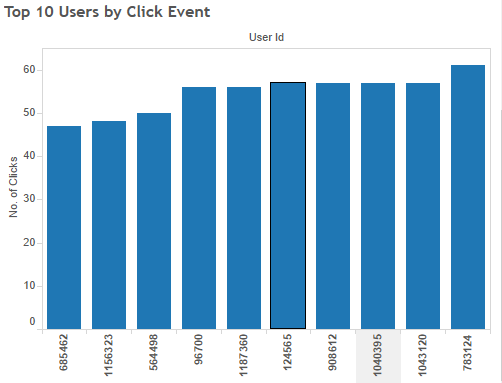
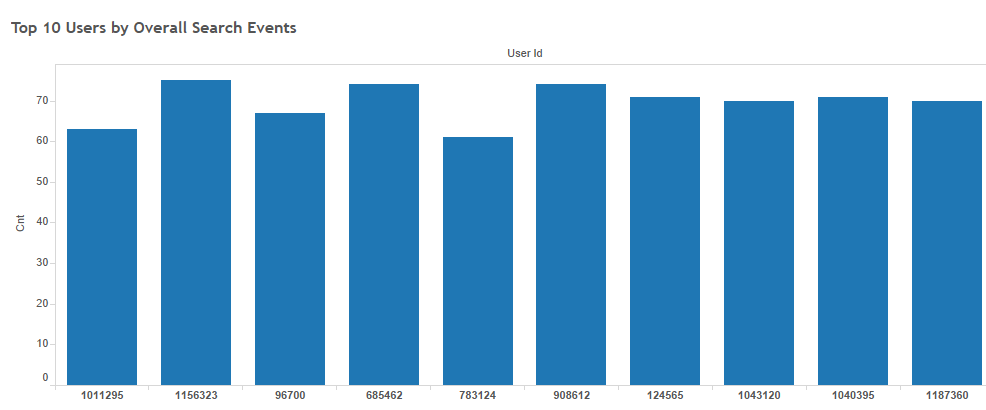
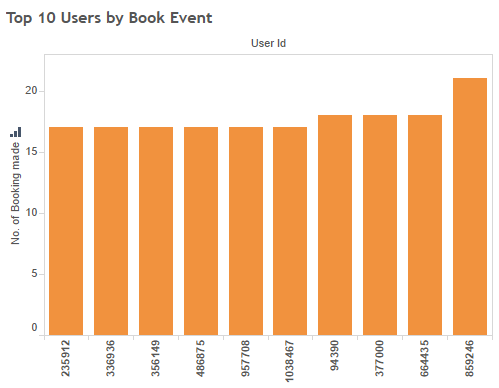


The dashboard above shows the hotel clusters by no. of searches for each.

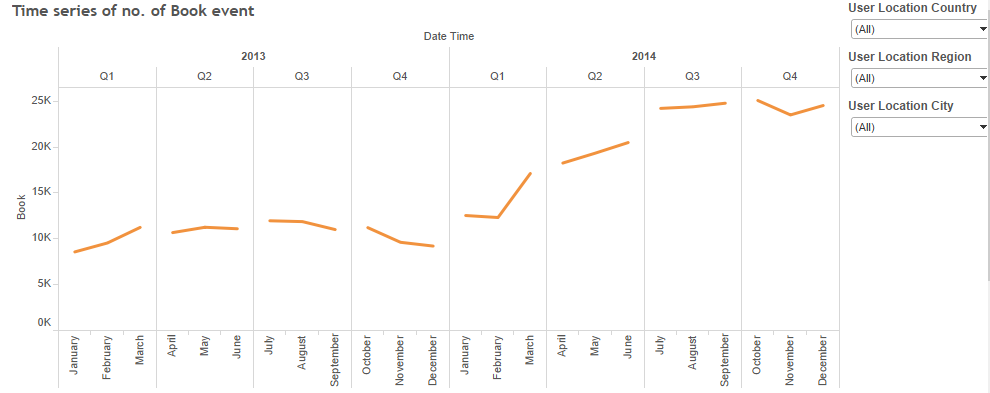
**User Analysis**



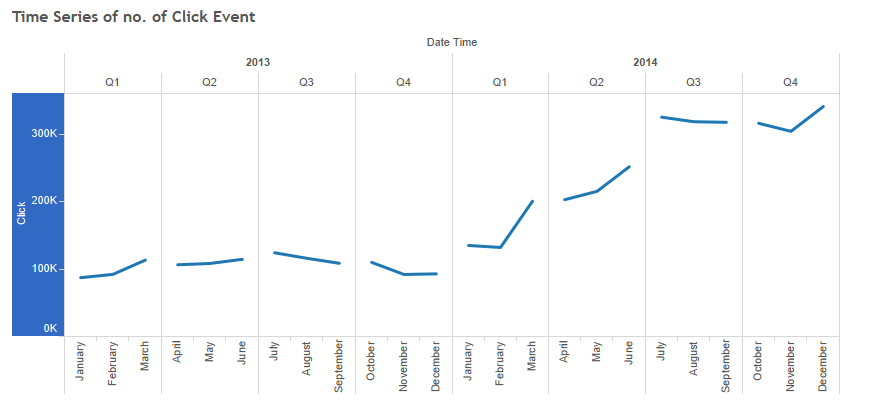
The above dashboard shows the Users activity (click/book) by various filter like the no. of Rooms they selected, the child count and the adult count.



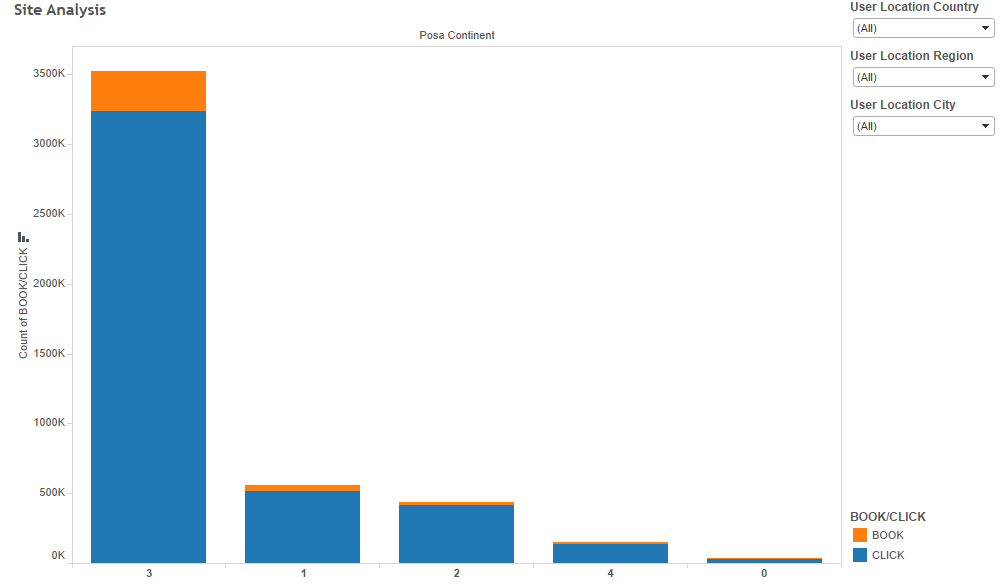
The above 3 dashboards show the top 10 users by overall searches, booking searches and click searches respectively.



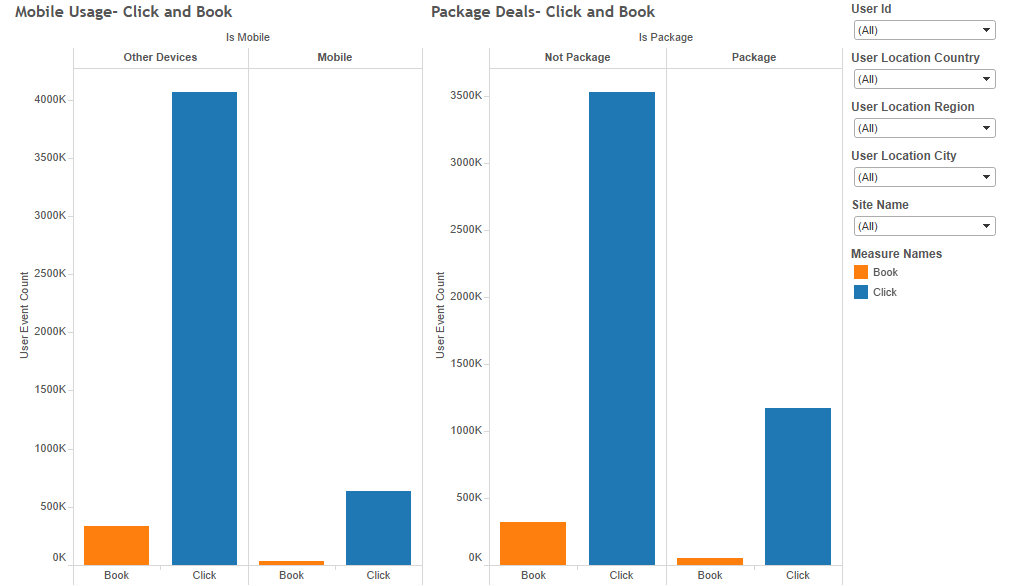
The above dashboard shows the time series for the booking event and also you can drill it by users location.



The above dashboard shows the time series for the click event and also you can drill it by users location.



The above dashboard shows which site continent has highest searches and this can be drilled down to see which sites have highest searches and also you can filter it by the users location.

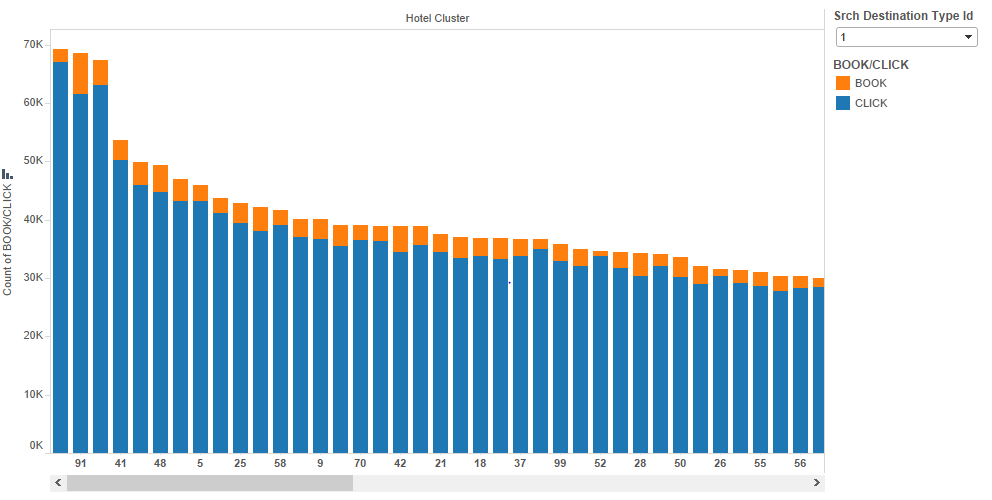


The above dashboard shows the distribution of search event by package deals or mobile usage by filtering them on basis of the site, user and user location.

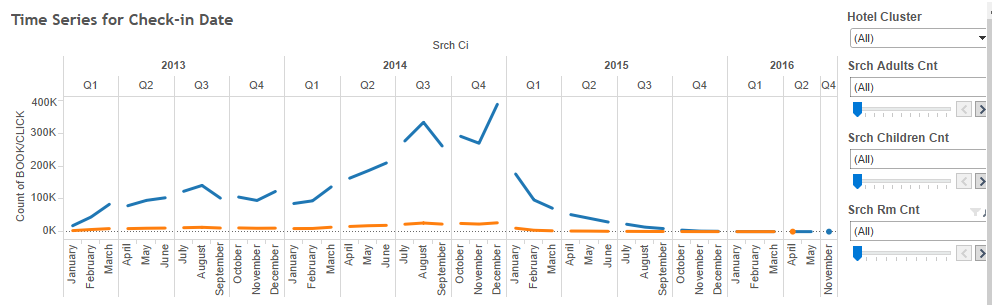
**Hotel Analysis**

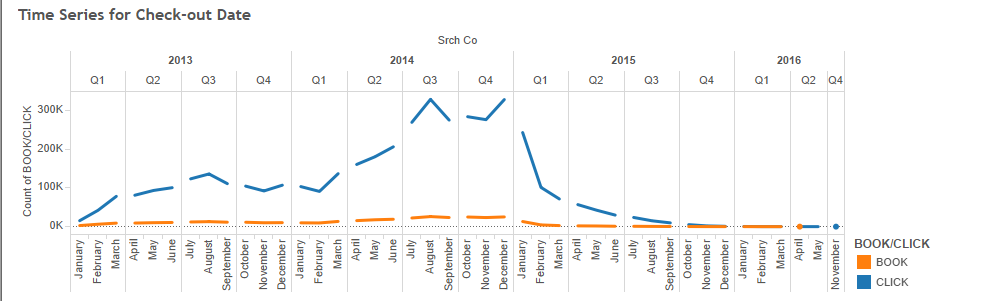


The above dashboard show the time series for event filtered by Hotel market and the various count other parameters. Also it shows the top 10 destination by book and click event which can be filtered by the filters shown.

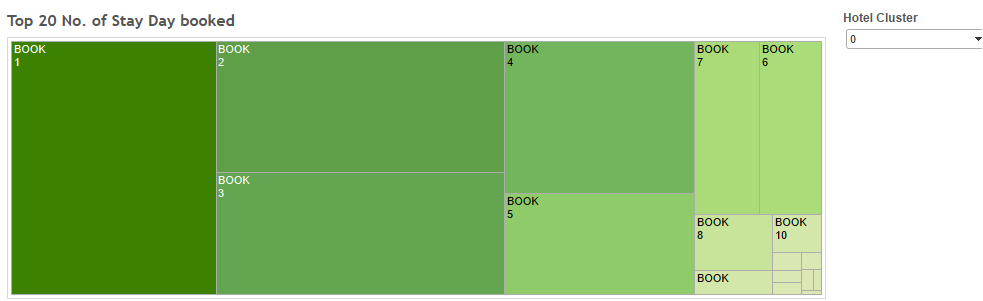


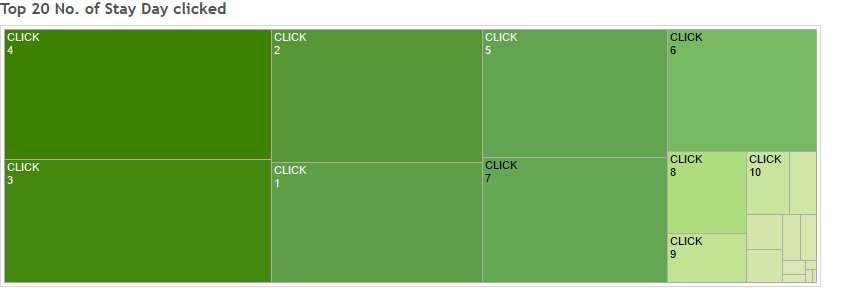
This shows the most popular Hotel Clusters by search event and can be filtered by the Destination Type Id.



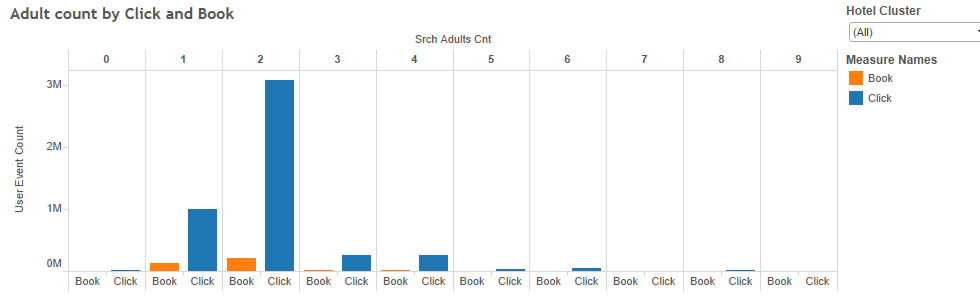


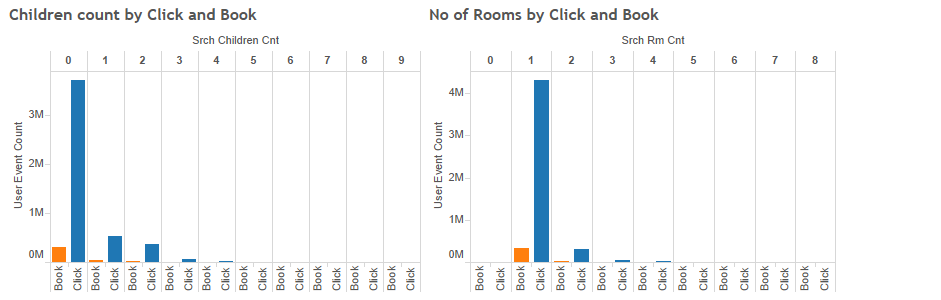
The above dashboard shows the Time Series of the Check-in and Checkout dates. We can see that the trend is similar for both. We can filter to see what kind of event it was with respect to parameter like child count, adult count, room count and hotel cluster.





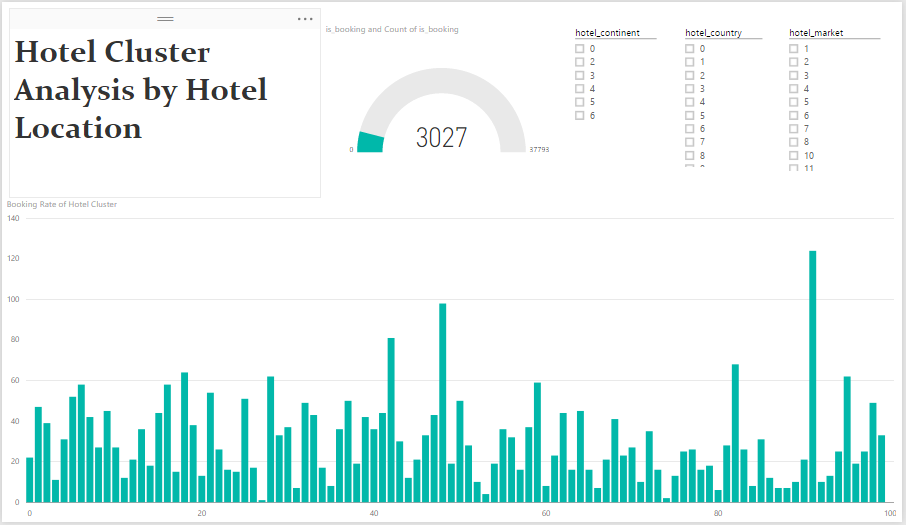
The dashboard above shows the no. of days stayed by event type which can be filtered by the cluster to see the stay maximum stay period in each cluster.



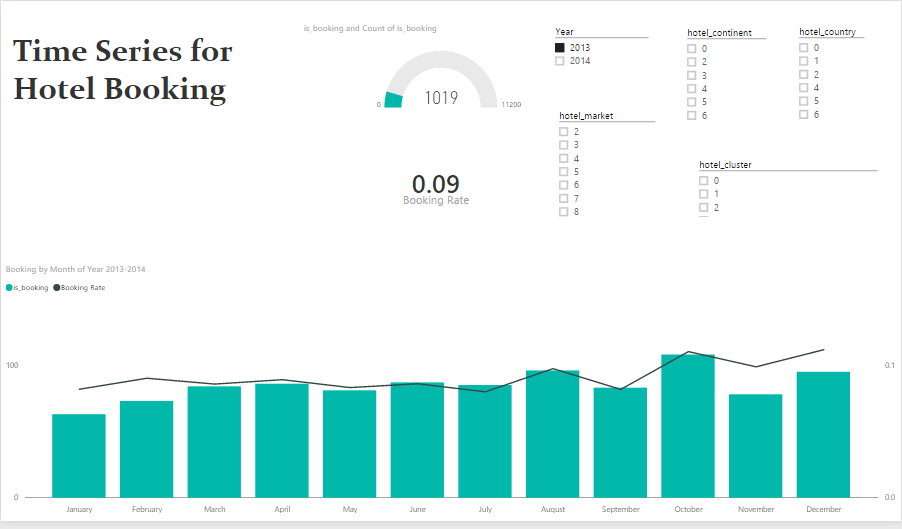


The dashboard above shows adult count, child count and room count by hotel cluster.

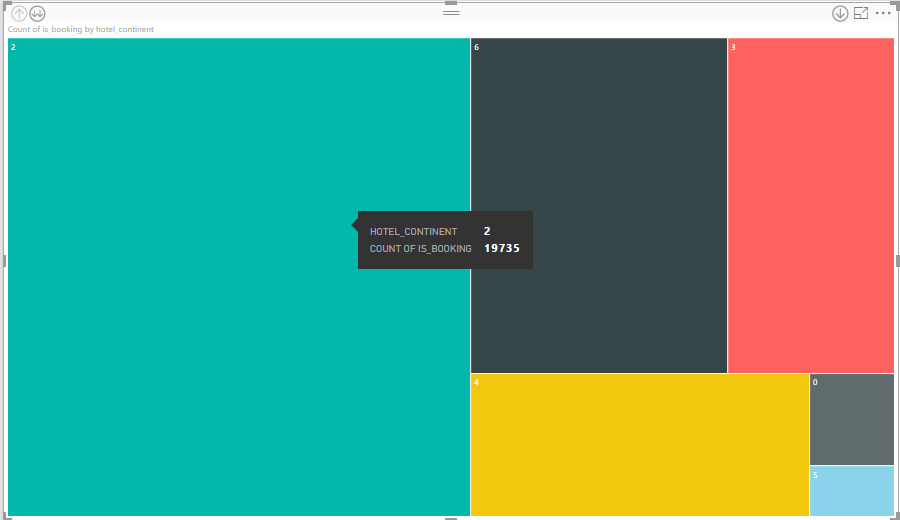
Power BI Dashboards



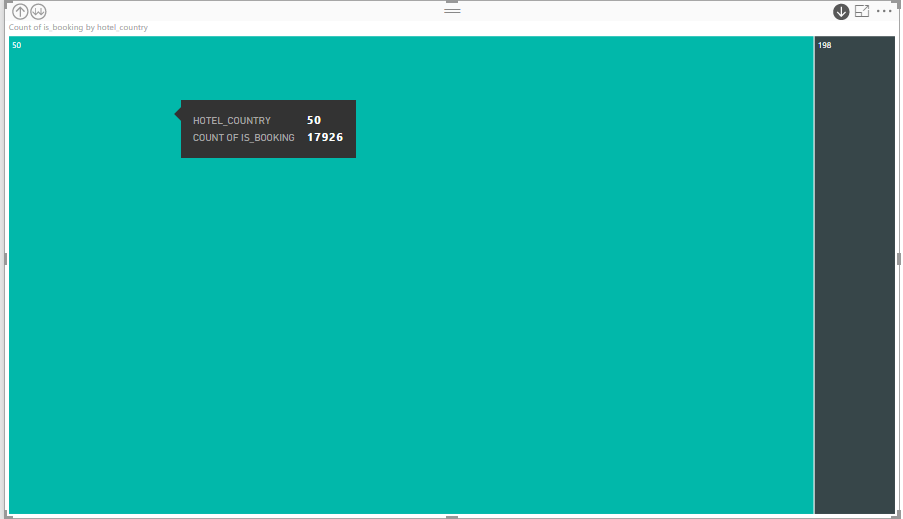
The above dashboard shows which continent/country/market have how many bookings for the hotl cluster in them.



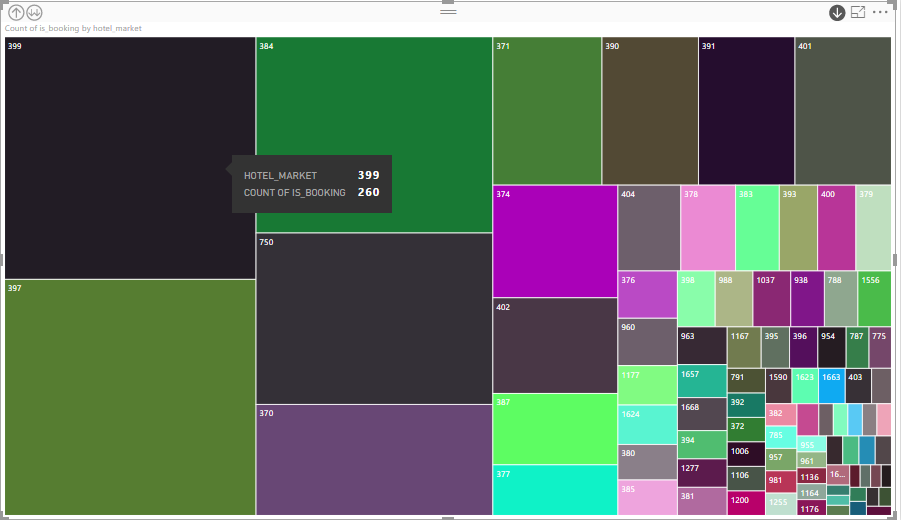
The dashboard above helps to visualize the time series for each cluster by filtering it by cluster no.

Hotel Hierarchy drilled by

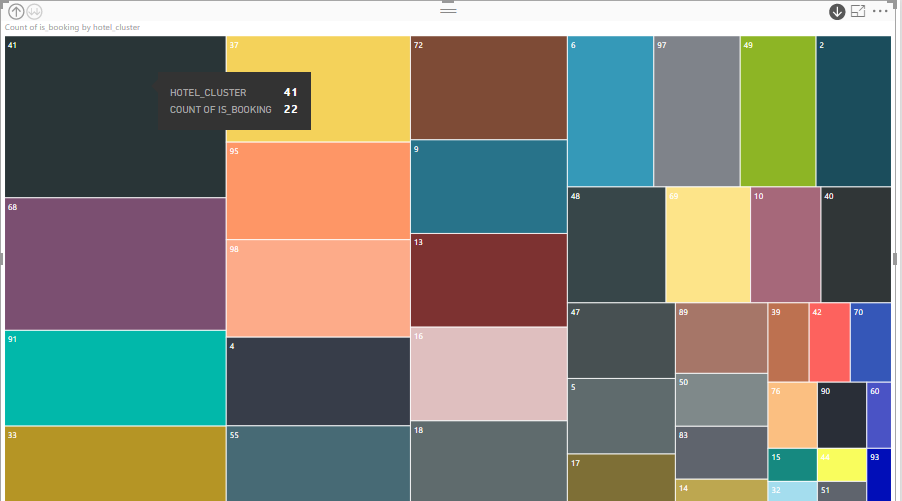
Hotel Continent



Hotel Country



Hotel Market

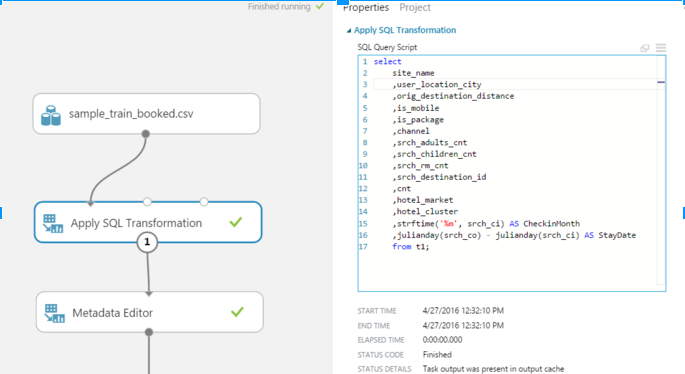
Hotel Cluster

The above tree map is representation if the Hotel Hierarchy, it shows which clusters belong to what market which market belongs to which country and finally which country belongs to which continent.

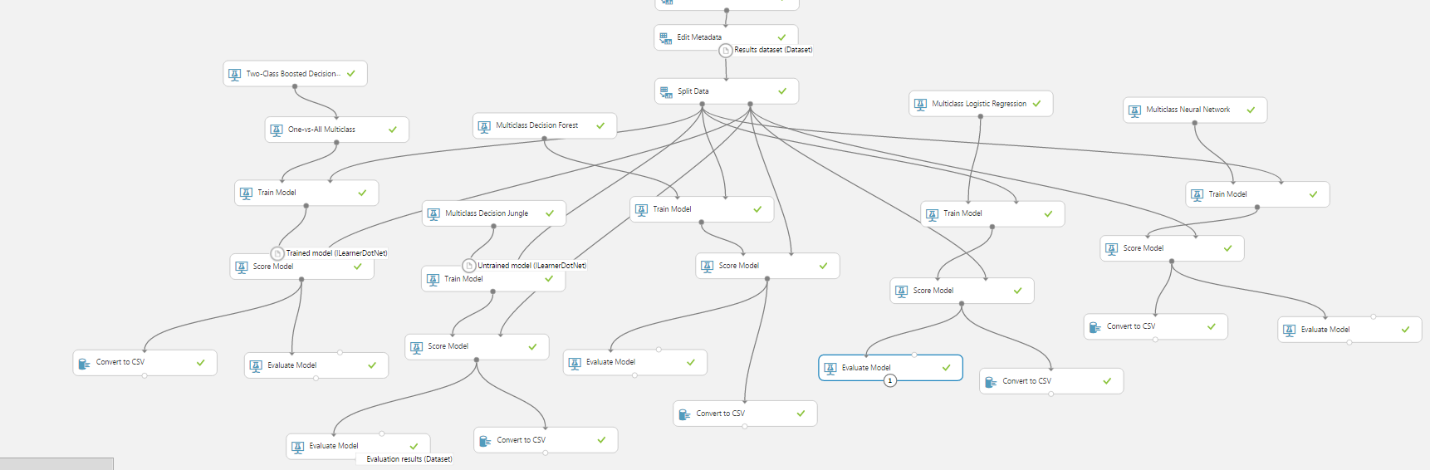
## **Modeling**

Tools: Microsoft Azure ML Studio, Python, Alteryx

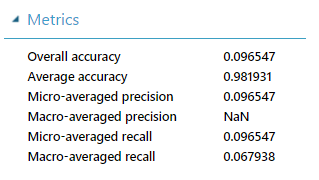
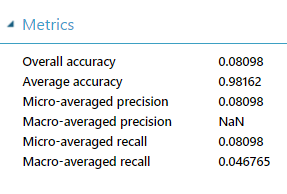
Multiple tables has been uploaded into Microsoft Azure ML Studio, for modeling and data blending purposes.

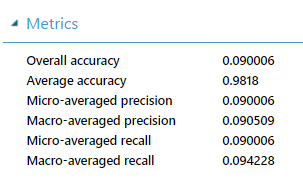
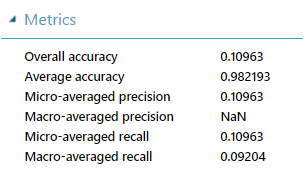
By using the widget “Apply SQL Transformation”, variables has been selected using a sample dataset randomly selected from the original dataset. Derived column CheckinMonth and StayDay (number of days between check in and check out dates).

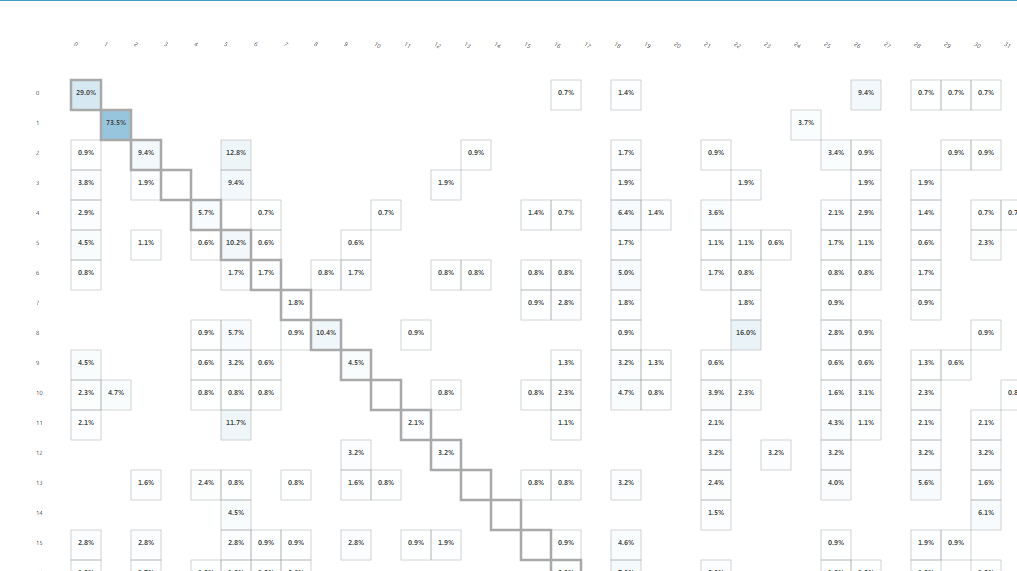
By using Metadata Editor, data has been casted into the right format. Categorical values also have been generated.

With the options available in Azure ML Studio, we applied 5 different models and compared their performance.

Model evaluation matrix are listed below from left to right the same order as they are in the graph. Multiclass Logistic regression has the highest accuracy rate.

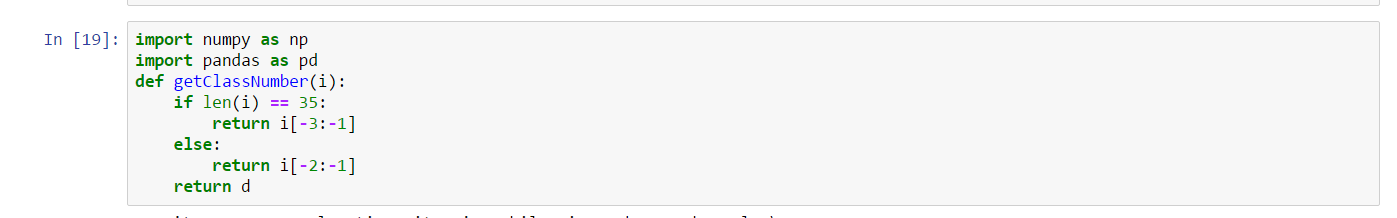


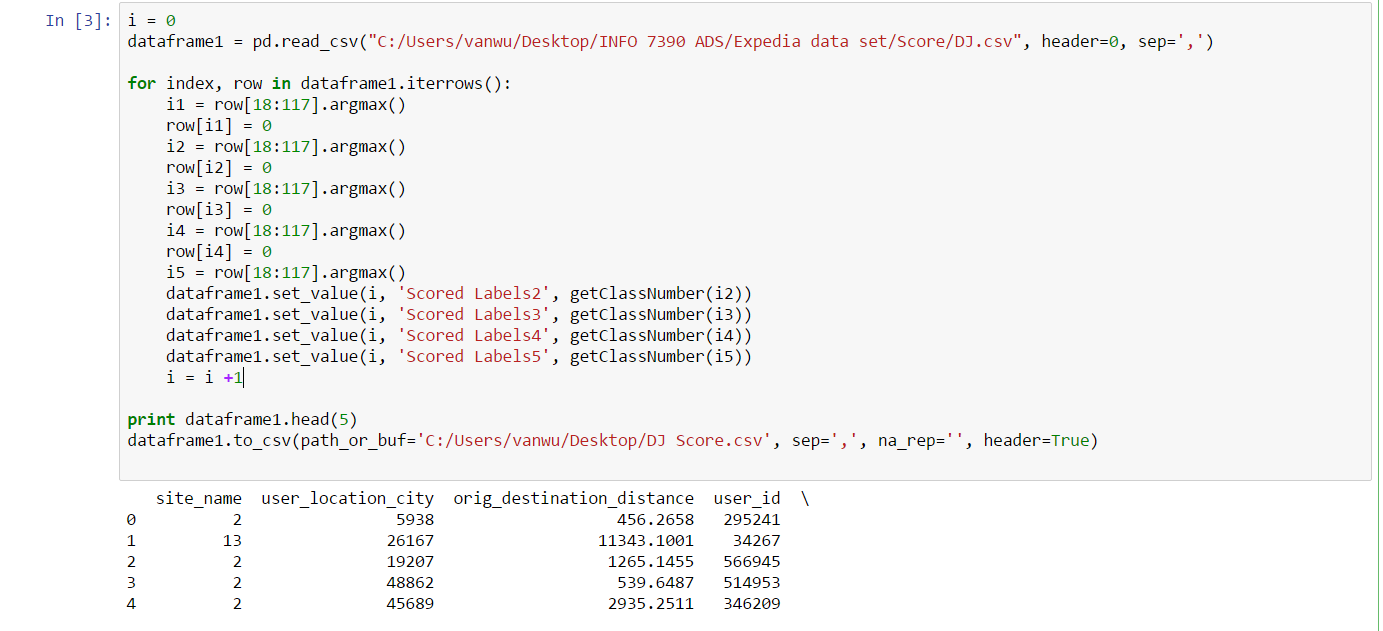


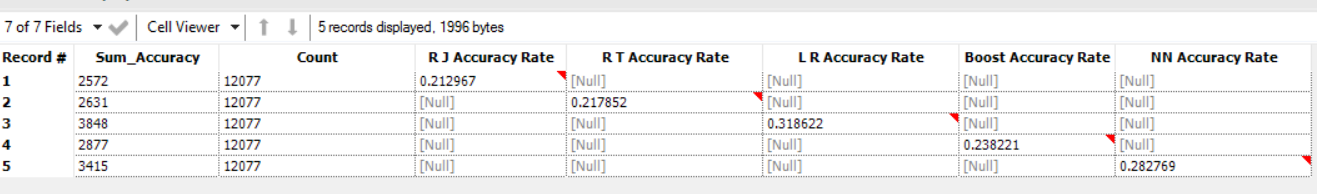
According to the data definition, hotel\_clusters act as labels for each search\_destination, and one search\_destination may falls into multiple hotel\_clusters. And each hotel\_cluster represents certain features of the search\_destinations in its group. Due to limited information from the training dataset and special business scenario, we performed a second model evaluation by taking the top five hotel\_clusters with highest probabilities in the classification models.

First, we downloaded the scored dataset and generated the top five cluster labels for each record using Python. Second, we loaded all five dataset into Alteryx and created the new evaluation matrix. According to our result, Multiclass Logistic Regression had the highest accuracy rate - 31.86%.

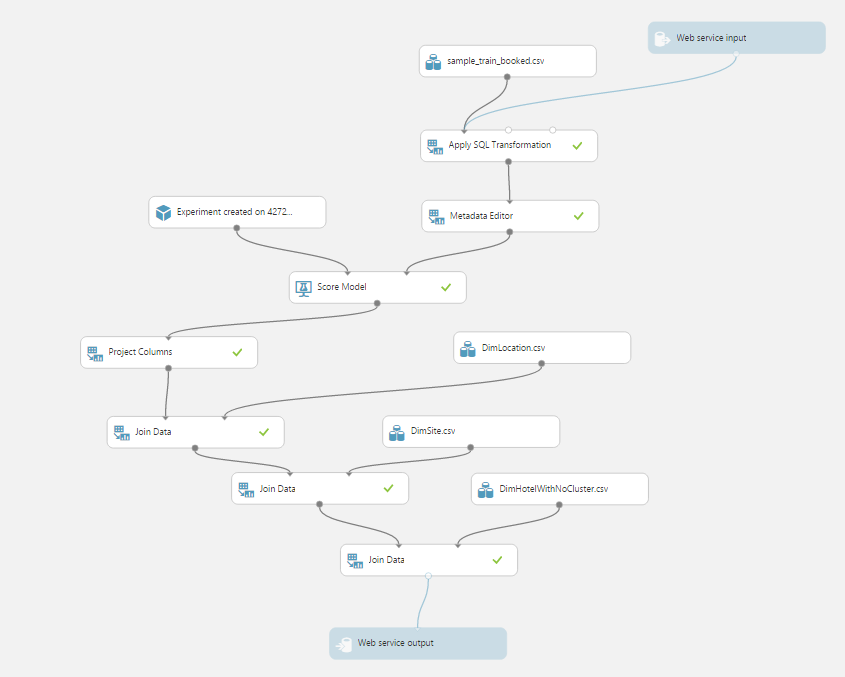
Based on our modeling results, we deployed a web service with Logistic Regression model.







## **Deployment**

After deploying our model as a web service, we modified our pipeline by joining our scored output table with three dimensions. By doing this, we will be able to retrieve complete information and present them on our web pages. 

**Tools**: Python server, Flask framework, REST API

In this step we deployed the web service and integrate it in front-end. We captured input data from the user and passed into the web service.

We called the API in front-end for the result using Javascript Ajax function.

We used Flask Framework for redirecting to the pages.

The website includes all the visualizations and brief description of the project

The Web Service input includes the Search Destination ID, Check-in Date, Check-out Date, Number of Adults while travelling, Number of Children and Number of Rooms searched

The output includes the Predicted Hotel Cluster ID and its related information like User Location Country, Hotel Continent, City, etc.

**Front-End:**

**Technologies**: Bootstrap, JQuery, Javascript, JQuery-UI and various javascript libraries



