Cluster Analysis

**Problem Statement:** In this project, we will mainly concentrate on clustering customers from the YPedia homepage search data, find out various insights

and patterns of customer behavior and solve few business related queries.

* + - * **Section 1 :** Handling Missing Data
      * **Section 2 :** Exploratory Data Ananlysis and Feature Engineering
      * **Section 3 : 3.1** Find out Under-performing and Out-performing Marketing Channels.

**3.2** Perform A/B Test on the Out-performing Channels.

* + - * **Section 4 :** K**-**meansClustering:
  1. What is the optimal number of Clusters for this data?
  2. Descriptive Analysis of all the Clusters in terms of booking rate.
  3. What are the important features that best describes 95% of the variance for each cluster?
     + - **Section 5 :** What lead to a higher chance of booking for individuals in each Cluster?

**About the Dataset:**

The YPedia homepage search data consists of 100,000 observations and 25 features. Since, some of the features were not used in this project, therefore those features will not be mentioned below.

The features set for this project are as follows:

|  |  |
| --- | --- |
| **Features** | **Description** |
| Userid | user id of each customer |
| Date\_time | date and time of the search |
| Srch\_ci | check-in date searched |
| Srch\_co | check-out date searched |
| Channel | marketing channel used [total 11 channels available] |
| Srch\_adults\_cnt | number of adults |
| Srch\_children\_cnt | number of children |
| Srch\_room\_cnt | number of rooms |
| Orig\_destination\_distance | distance from home to destination |
| Is\_mobile | searched from mobile phone? [No: 0, Yes: 1] |
| Is\_package | any package used? [No:0, Yes: 1] |
| Is\_booking | was booking done successfully? [No: 0 , Yes: 1] |

**SECTION 1:**

**Handling missing data:**

There are few features with missing data. They are:

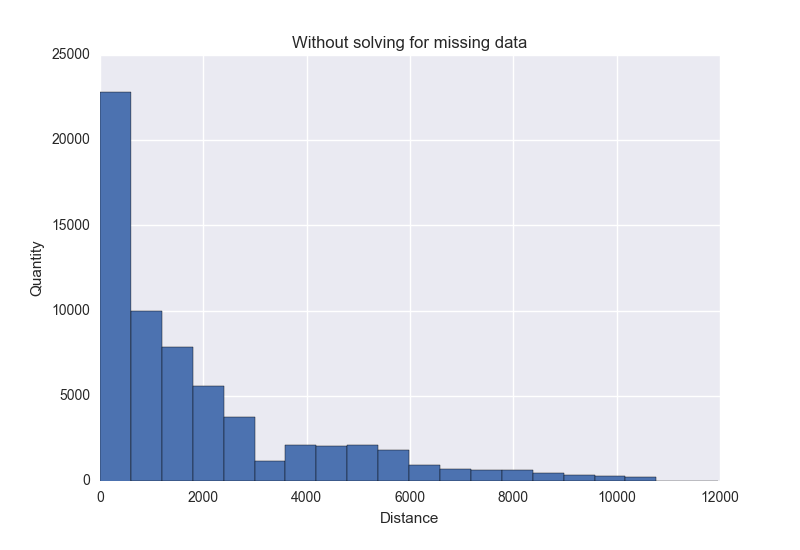
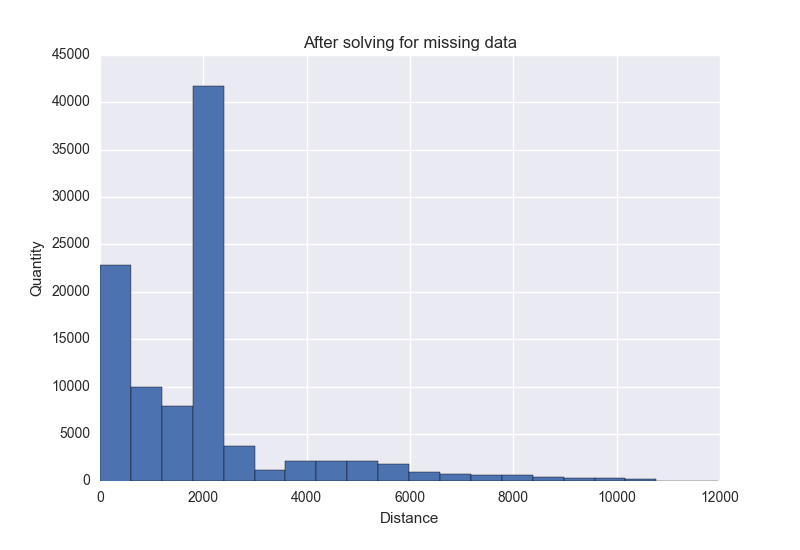
|  |  |
| --- | --- |
| Features | Number of missing data |
| Orig\_destination\_distance | 36085 |
| Srch\_ci | 122 |
| Srch\_co | 122 |

* **Orig\_destination\_distance:** There are a lot of data points missing for this feature (~36%). With a mean of **1960.662** and a median of **1131.835,** this distribution is positively

skewed.

So, it makes sense to impute the mean of this data to the missing values, so as to minimize the difference between the mean and the median.

New mean and median are same, i.e. **1960.662.**



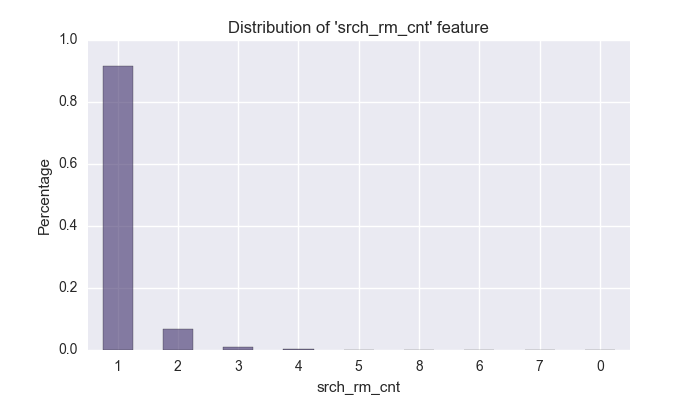
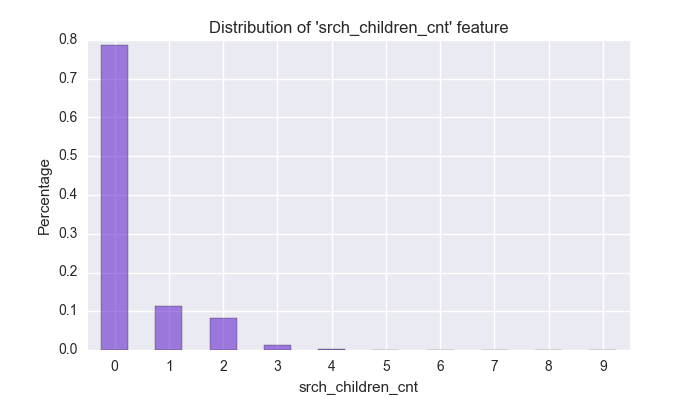
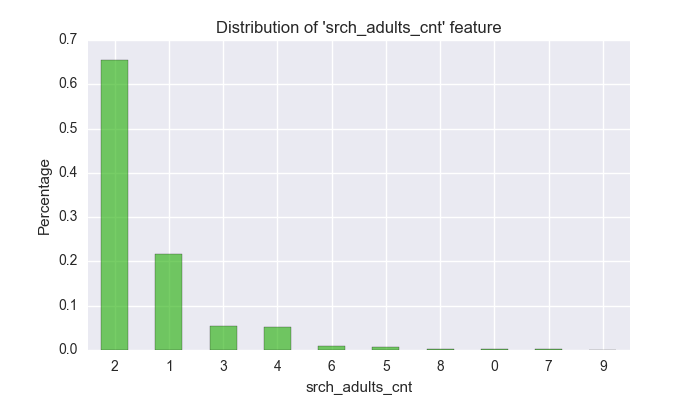
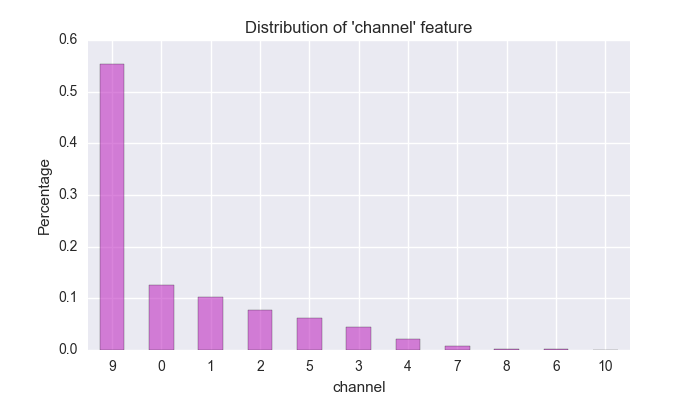
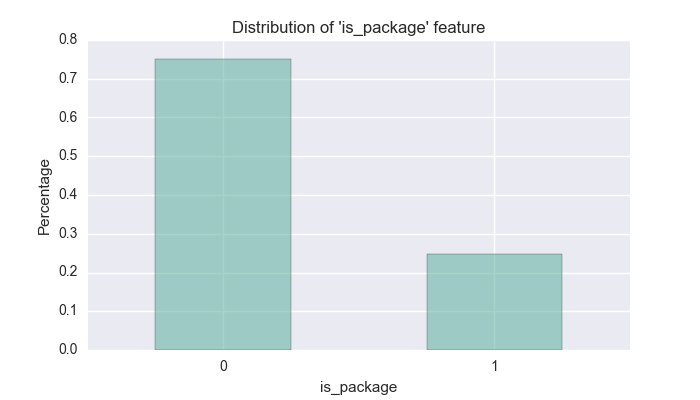
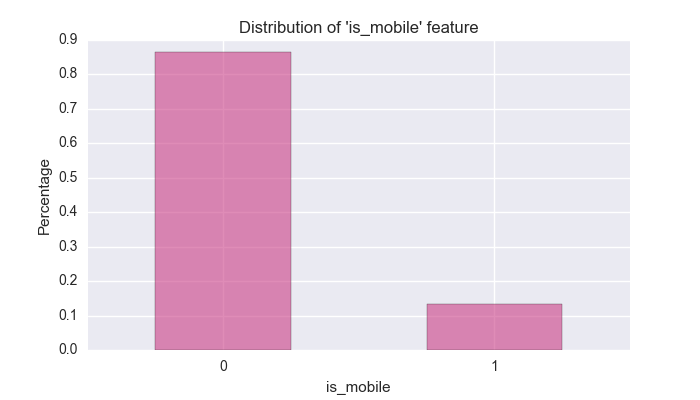
* **Srch\_ci and Srch\_co:** These two features represents dates, lets just remove them.

**NOTE:** After solving for missing data, our dataset now contains 99,878 rows and 12 features.

**SECTION 2:**

**Exploratory Data Analysis:**

1. **Categorical Values:**



* **Is\_booking:** only **8%** of the total searched data was converted.
* **Is\_mobile :** only **13.3%** of the searched data were done by mobile apps.
* **Is\_package:** only **24.8%** of the time a package was used in the search.
* **Channel :** Marketing channel 9 was used the maximum number of times with **55.4%,** followed by channel 0 and 1 with only **12.4%** and **10.2%**

respectively.

* **Srch\_adults\_cnts:** most of the time the adults count per searched results accounted for not more than 2, with **65.5%** and **21.5%** for 2 adults and

1 adult respectively. Followed by for 3 adults which is around **5.4%** only.

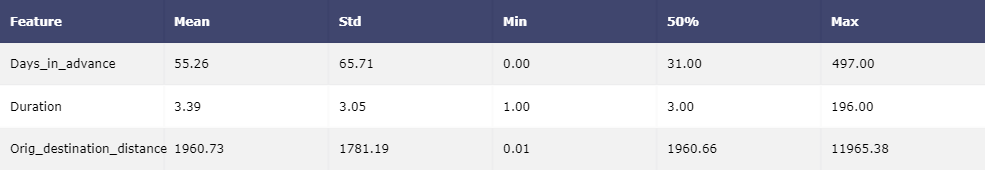
* **Srch\_children\_cnt:** No-children = **78.7%,** 1 children = **11.2%,** well this makes sense from the results we got from the adults count. They are mostly

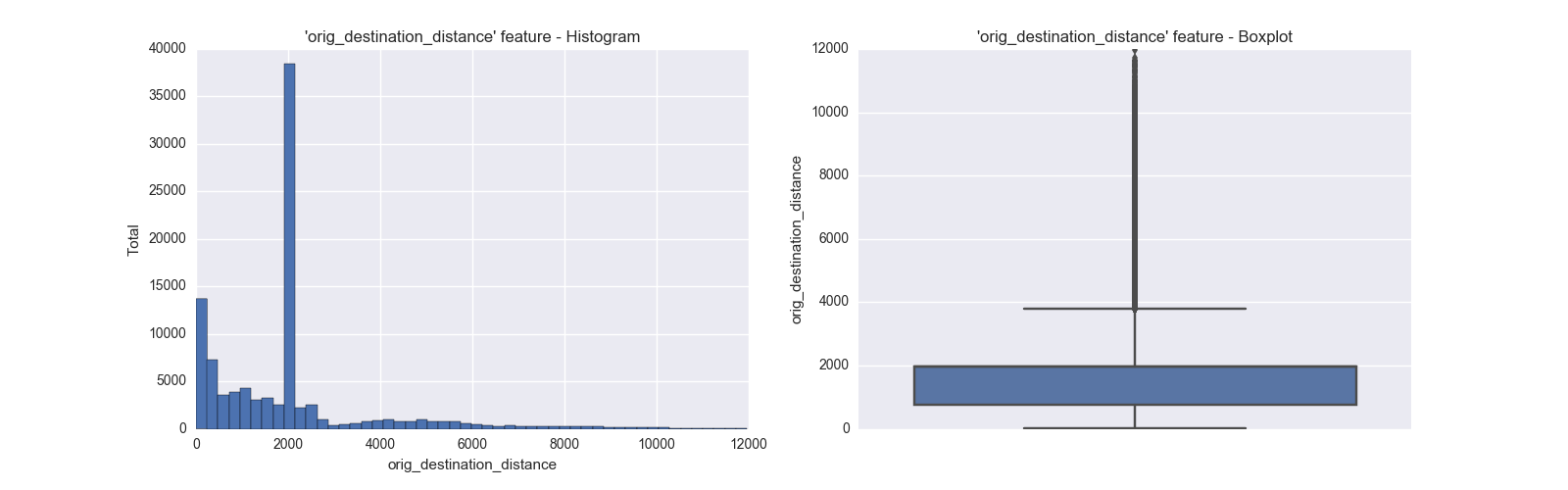
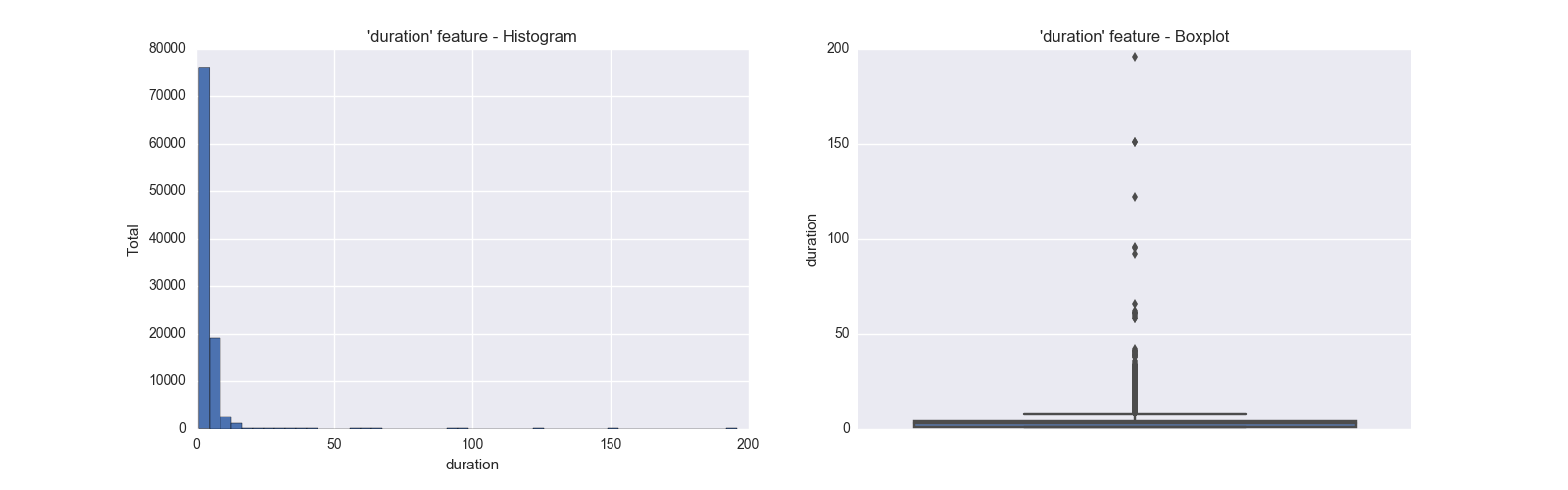
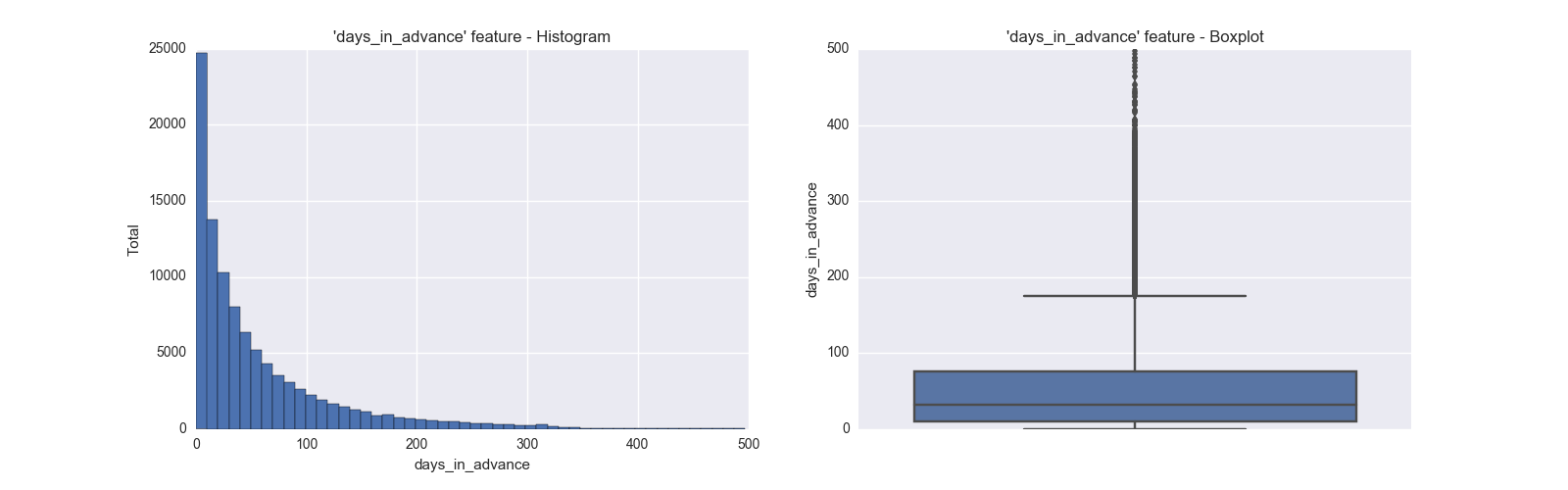
young couples with 0 or 1 children.

* **Srch\_rm\_cnt: 91.6%** of the time the room count was just 1, followed by 2 rooms for just **6.6%**.

1. **Numerical features:** Apart from Orig\_destination\_distance, two more features were created.

* Duration : duration of stay [from the features “srch\_ci” and “srch\_co”]
* Days\_in\_advance : number of days room searched in advance from the booking date [from features “srch\_ci” and “Date\_time”]

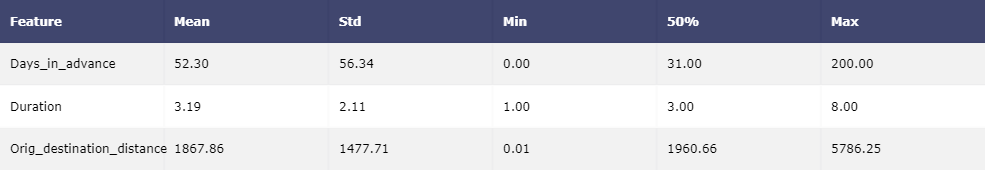


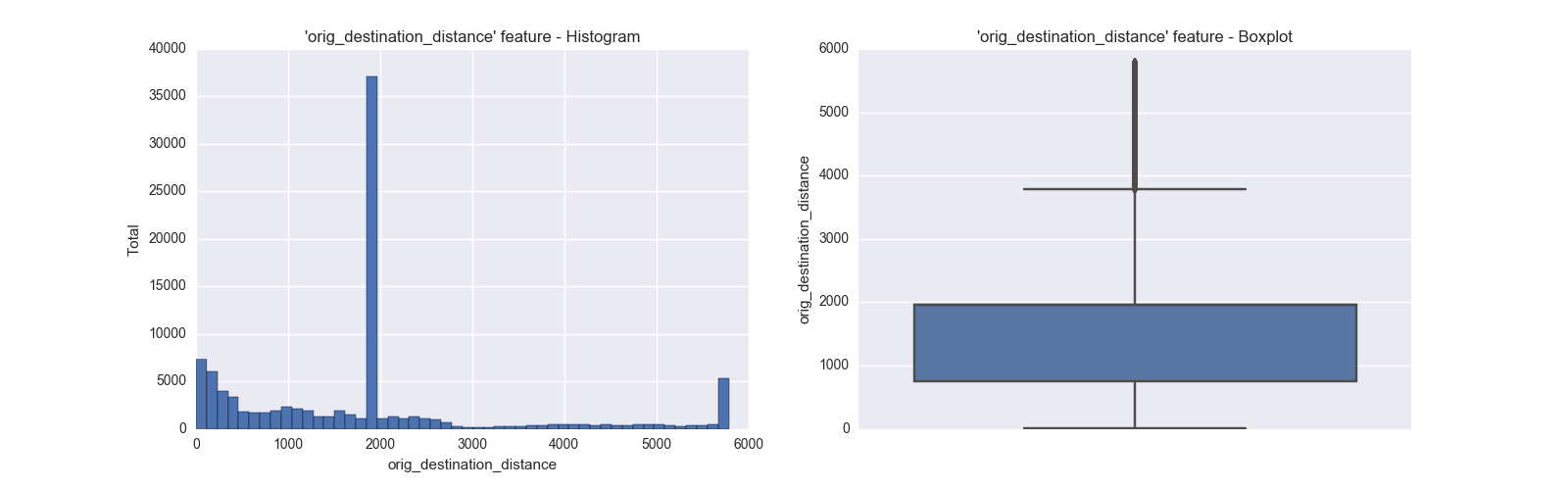
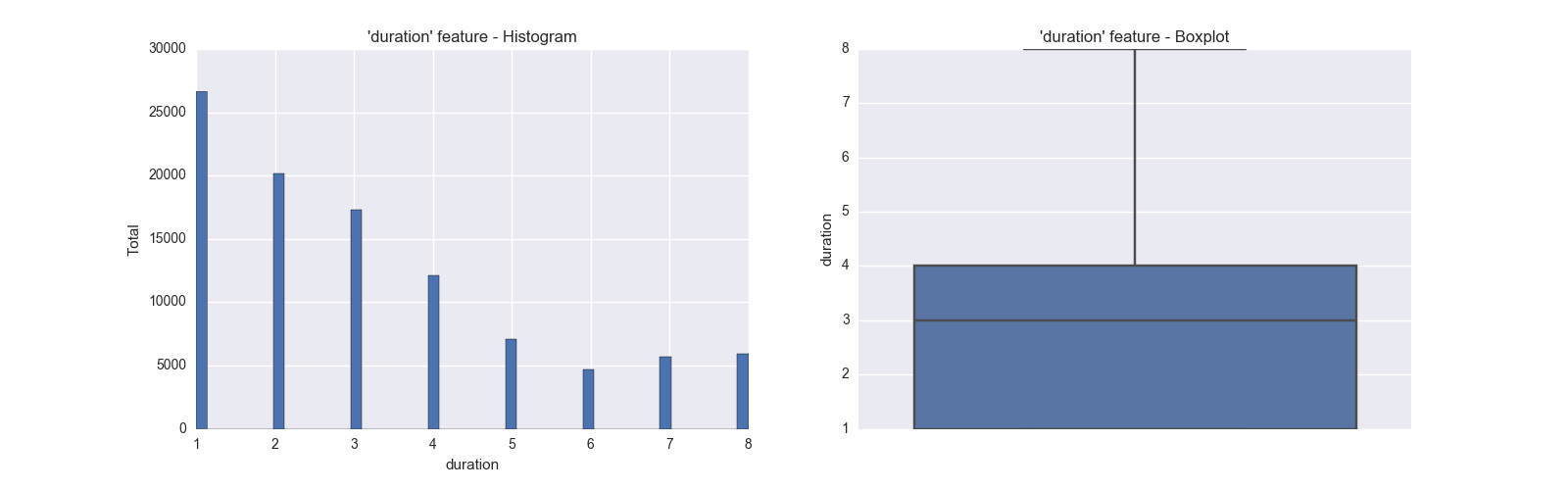
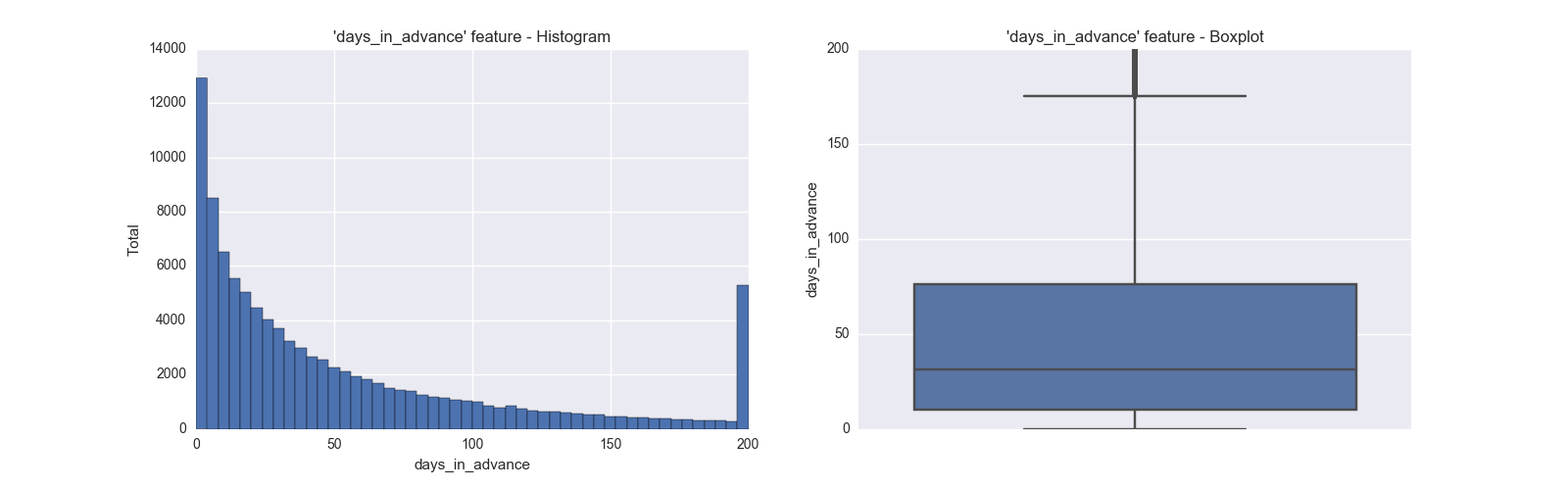


**Feature Transformations:**  The numerical features have a lot of extreme/outlier values, as can be seen from the boxplots. Since these outliers are only present to the top half, lets

impute the values which exceeds 95th percentile, i.e. keep values between [0th – 95th percentile] and impute the exceeding values to the value in the 95th

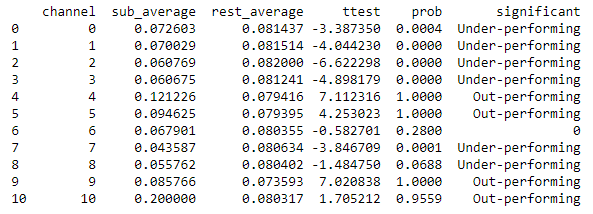
percentile.





**SECTION 3:**

**3.1 “Find out Under-performing and Out-performing Marketing Channels.”**



**Description:**

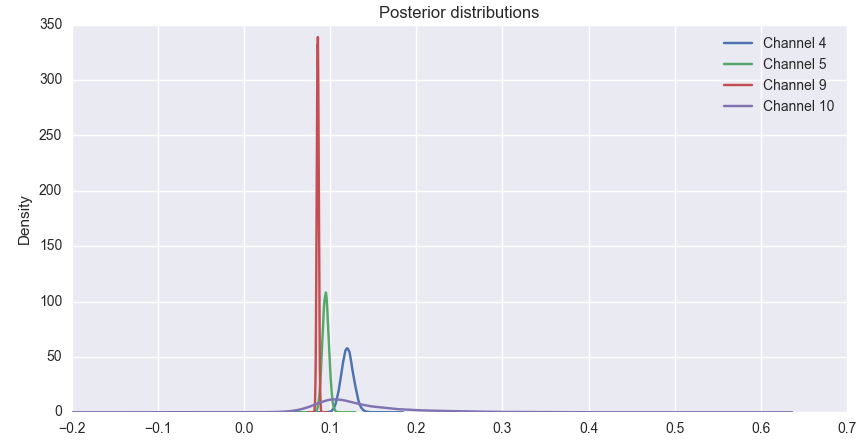
1. **Sub\_average** : booking rate of that particular channel.
2. **Rest\_average** : booking rate of all other channels combined.
3. **Ttest** : two sample t-test value
4. **Prob** : probability obtained from Cumulative Distribution Function
5. **Significant** : If the **probability > 0.9**, we can conclude that this particular channel outperforms other channels, i.e. we are more than 90% confident.

If the **probability < 0.1**, we can conclude that this channel underperforms than the other channels, i.e. again we are 90% confident that it

underperforms.

If **0.1 <= probability <= 0.9**, we cannot conclude anything statistically.

**3.2** **“Perform A/B Test on the Out-performing Channels”**

****

Hierarchical Modelling

In this section, we modeled the channel bookings using a **Binomial Distribution.**

The intuition behind **Binomial(n,p)** is that if we flip a coin with probability p of landing heads n times, how likely is it that we see k heads for some k between 0 and n.

Now, **the number of bookings, ki,** is modeled by **Binomial(ni,pi),** and the **true booking rate for each marketing channel, pi,** is drawn from **Beta(a,b).**

Sampling **a** and **b** from the distribution with the function:  
**f(a**,**b) ~ (a+b)-5/2 , where a**,**b > 0**

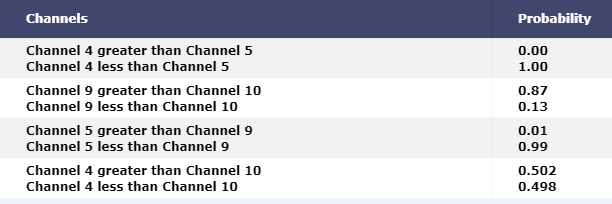
Finally we computed the difference between any two of them, at a time.

From the probabilities, we get the following:

**Channel 9 > Channel 5 > Channel 4 > Channel 10**

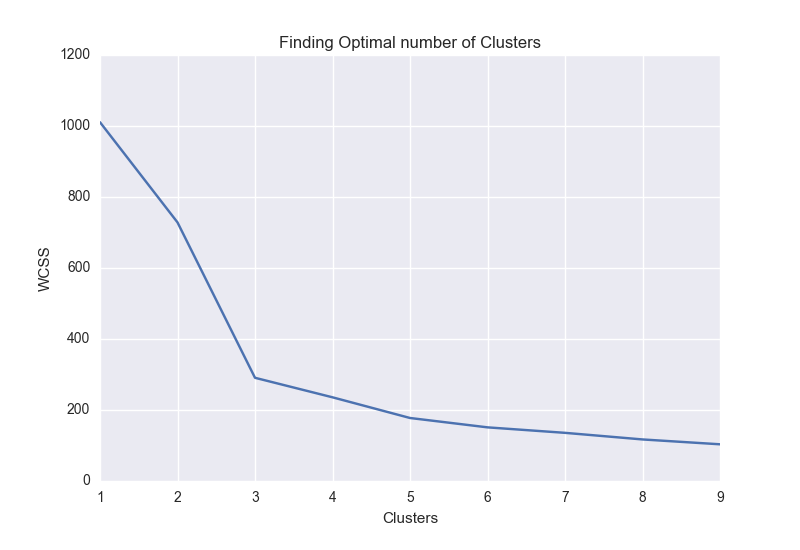
**Conclusion:**

Channel 9 appears to perform better than any other channels by a greater margin, followed by Channel 5.

****

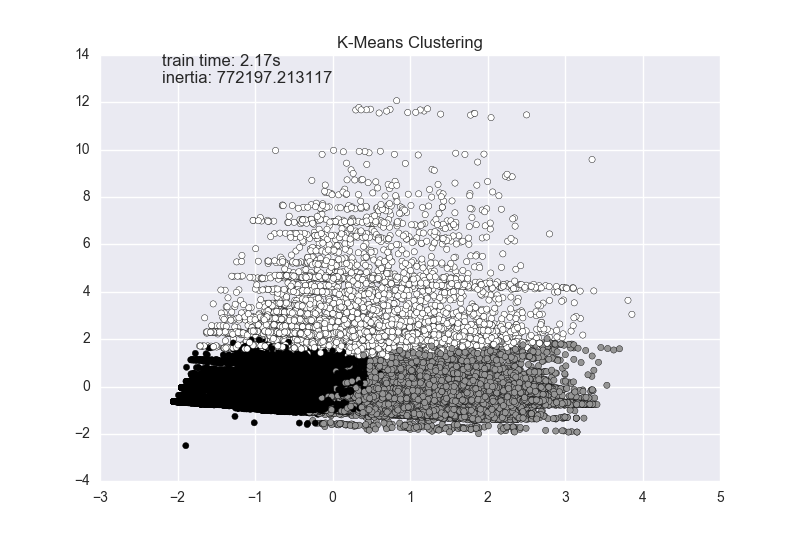
**SECTION 4:**

* 1. **“What is the optimal number of Clusters for this data?”**



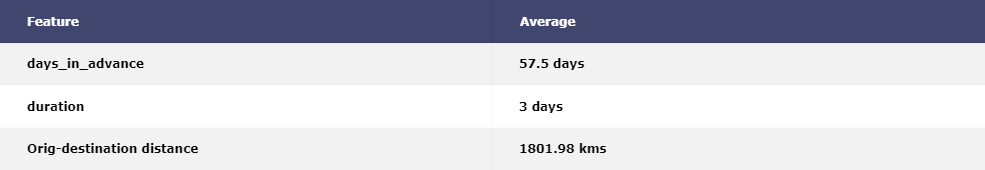
From the WCSS (Within Cluster Sum of Squares), the first 3 clusters have large distances between them and as we

move forward, the value gets minimized. So for this data, the optimal number of clusters is set to 3.



* 1. **“Descriptive Analysis of all the Clusters.”**

|  |
| --- |
| **Cluster - 0** |



Cluster 0 has **6,903** observations.

This cluster represents customers who travels a distance **less than** **1960 kms for 75% of the time,** **1802 kms** **being the average** and stays for an average duration of **3 days.**

They tend to search for hotel **less than** **3 months** **for 75% of the time in advance, 2 months being the average.**

Probability of **mobile apps and package usage are low** as compared to the overall mean.

This cluster best represents **families** who travels in groups, as can be seen from the number of adults, children and room counts.

Out-performing Channels:

**Channel 6 and Channel 9**

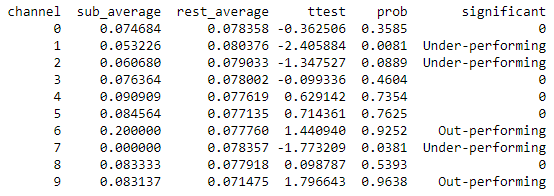
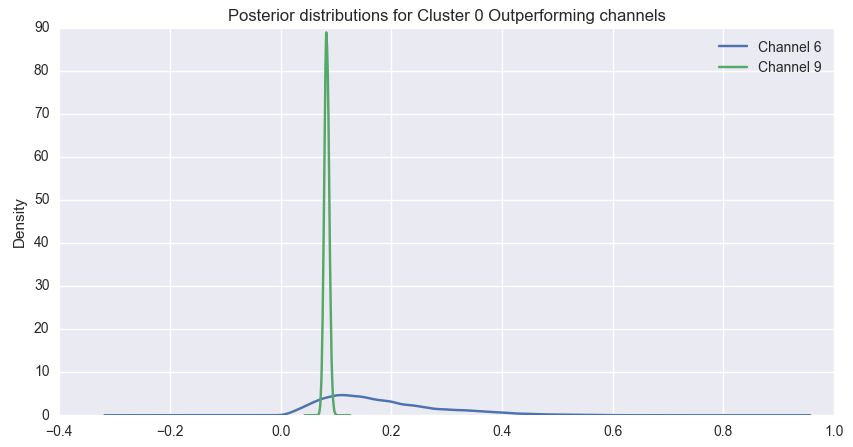
From the hierarchical modelling of all out-performing channels:

Probability that Channel 6 > Channel 9: **0.167**

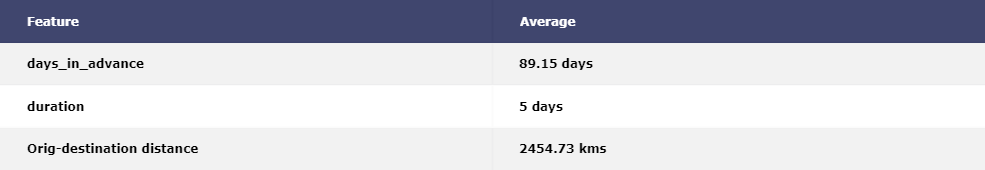
Probability that Channel 6 < Channel 9: **0.833**

**i.e. Channel 9 > Channel 6**

|  |  |  |
| --- | --- | --- |
| **Features** | **Distribution of Users** | **Probability of booking** |
| **Is\_mobile** | **No 88.3%**  **Yes 11.6%** | **0.072** |
| **Is\_package** | **No 82.7%**  **Yes 17.2%** | **0.037** |
| **Srch\_adults\_cnt** | **1 27.9% 6 12.2%**  **2 30.4% 7 2.04%**  **3 18.2% 8 3.2%**  **4 47.7% 9 0.5%**  **5 10.0%** | **9 adults – 0.153**  **8 adults – 0.106**  **3 adults - 0.092**  **7 adults - 0.092** |
| **Srch\_children\_cnt** | **0 68.1% 5 0.6% 1 13.7% 6 0.3%**  **2 10.1% 7 0.8%**   1. **3.8% 8 0.05%**   **4 2.9% 9 0.04%** | **9 children – 0.33**  **7 children – 0.33**  **5 children - 0.09**  **0 children - 0.08** |
| **Srch\_rm\_cnt** | **1 4.1% 5 1.6%**  **2 71.9% 6 0.09%**  **3 15.5% 7 0.04%**  **4 4.2% 8 0.09%** | **4 rooms – 0.14**  **3 rooms - 0.09**  **6 rooms - 0.08**  **5 rooms – 0.07** |

|  |
| --- |
| **Cluster - 1** |



Cluster 1 has **34,547** observations.

This cluster represents customers travels a distance **less than** **2605 kms for 75% of the time,** **2455 kms** **being the average** and stays for an average duration of **5 days.**

They tend to search for hotel **less than** **5 months** **for 75% of the time in advance, 3 months being the average.**

Probability of **mobile apps and package usage are very low** as compared to the overall mean.

This cluster best represents **individual people or couples** having **no children or very less children [1-3].**

Out-performing Channels:

**Channel 4, Channel 5 and Channel 10**

From the hierarchical modelling of all out-performing channels:

Probability that Channel 4 > Channel 5: **0.167**

Probability that Channel 4 < Channel 5: **0.833**

Probability that Channel 5 > Channel 10: **0.818**

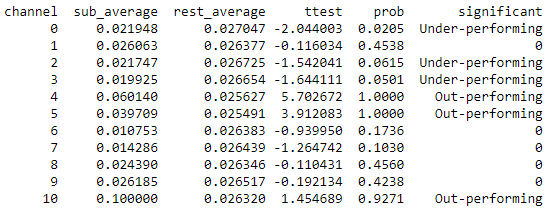
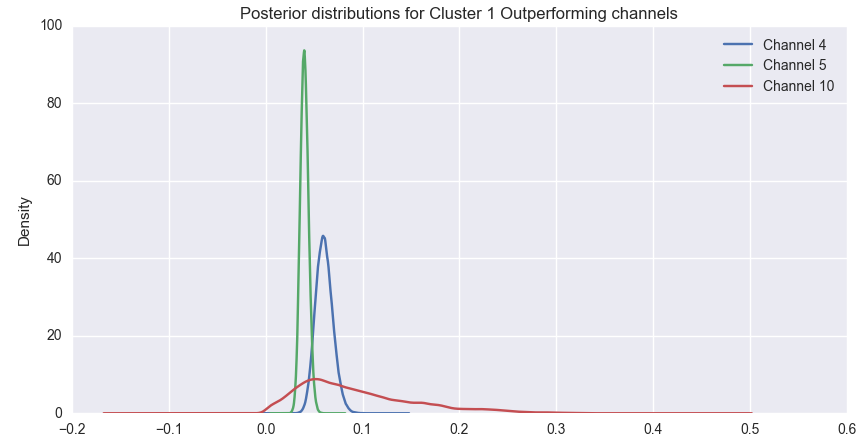
Probability that Channel 5 < Channel 10: **0.182**

Probability that Channel 4 > Channel 10: **0.647**

Probability that Channel 4 < Channel 10: **0.353**

**i.e. Channel 5 > Channel 4 > Channel 10**

|  |  |  |
| --- | --- | --- |
| **Features** | **Distribution of Users** | **Probability of booking** |
| **Is\_mobile** | **No 84.8%**  **Yes 15.1%** | **0.02** |
| **Is\_package** | **No 36.9%**  **Yes 63.0%** | **0.028** |
| **Srch\_adults\_cnt** | **1 17.6% 4 2.0%**  **2 74.7% 5 0.1%**  **3 4.9% 6 0.02%** | **1 adult – 0.032**  **2 adults – 0.022**  **3 adults - 0.021** |
| **Srch\_children\_cnt** | **0 79.1% 4 0.2%**  **1 10.5% 5 0.005%**  **2 8.5% 6 0.01%**  **3 1.4%** | **1 children - 0.03**  **0 children - 0.02**  **2 children - 0.02**  **3 children - 0.02** |
| **Srch\_rm\_cnt** | **1 97.9%**  **2 2.02%**  **3 0.002%** | **2 rooms – 0.03**  **1 rooms - 0.02** |

Cluster 2 has **58,021** observations.

This cluster represents customers who comes from nearby places, with distance **less than** **1960 kms for 75% of the time,** **1525 kms** **being the average** and stays for an average duration of **2 days.**

They tend to search for hotel **less than** **1 and half months** **for 75% of the time in advance, 1 month being the average.**

Probability of **mobile apps usage is little better as compared to other clusters but package usage is still very low**.

This cluster best represents **individual people/couples as well as small families with not more than 3 – 4 adults and 6 children**.

Out-performing Channels:

**Channel 4, Channel 5, Channel 9 and Channel 10**

From the hierarchical modelling of all out-performing channels:

Probability that Channel 4 gets > Channel 5: **0.01**

Probability that Channel 4 gets < Channel 5: **0.99**

Probability that Channel 9 gets > Channel 10: **0.86**

Probability that Channel 9 gets < Channel 10: **0.13**

Probability that Channel 4 gets > Channel 10: **0.58**

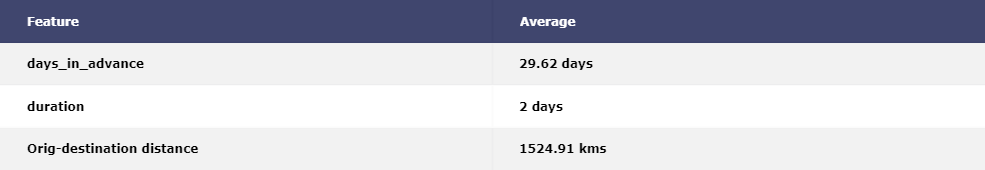
Probability that Channel 4 gets < Channel 10: **0.42**

Probability that Channel 5 gets > Channel 9: **0.02**

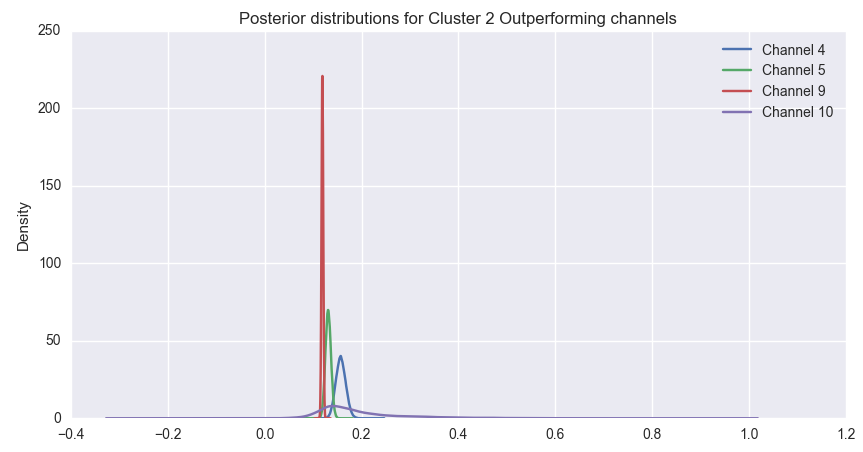
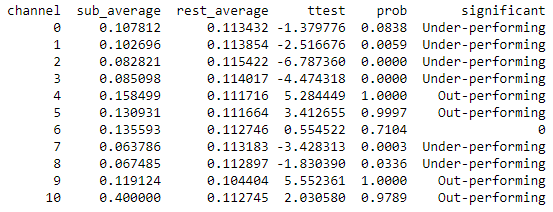
Probability that Channel 5 gets < Channel 9: **0.98**

**i.e. Channel 9 > Channel 5 > Channel 4 > Channel 10**

|  |
| --- |
| **Cluster - 2** |



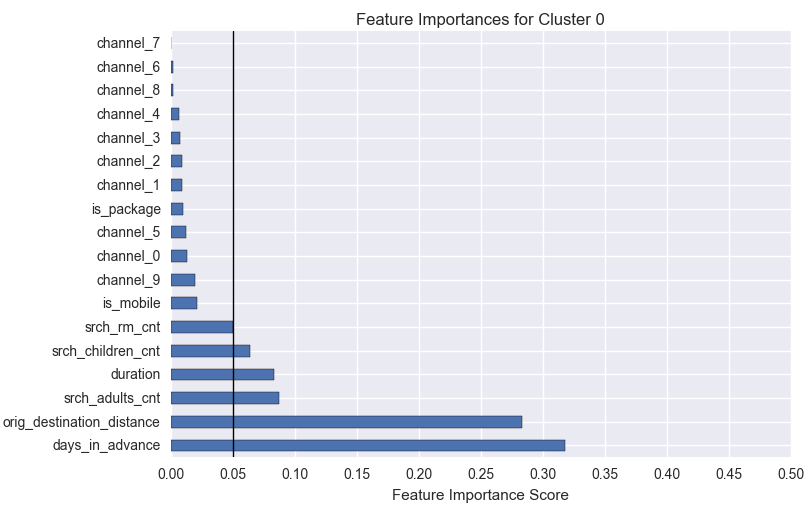
|  |  |  |
| --- | --- | --- |
| **Features** | **Distribution of Users** | **Probability of booking** |
| **Is\_mobile** | **No 87.4%**  **Yes 12.5%** | **0.088** |
| **Is\_package** | **No 97.5%**  **Yes 2.4%** | **0.025** |
| **Srch\_adults\_cnt** | **1 26.2% 4 2.1%**  **2 67.4% 5 0.01%**  **3 4.1%** | **1 adult – 0.15**  **2 adults – 0.09**  **3 adults - 0.08** |
| **Srch\_children\_cnt** | **0 79.7% 4 0.1%**  **1 11.4% 5 0.008%**  **2 7.6% 6 0.01%**  **3 0.9%** | **6 children - 0.142**  **0 children - 0.116**  **1 children - 0.116**  **4 children - 0.112** |
| **Srch\_rm\_cnt** | **1 98.3%**  **2 1.6%** | **2 rooms – 0.19**  **1 rooms - 0.11** |



**4.3 “What are the important features that best describes 95% of variance for each cluster?”**

**NOTE:** The Feature Importances scores were computed using Random Forest. The black vertical line represents 95% variance in scores.

**Cluster - 0**



There are 6 features that explains almost 95% of the variance.

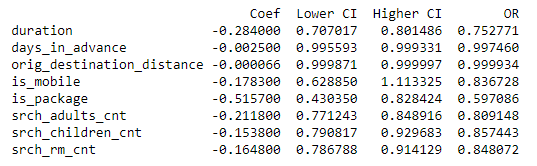
From Section 4.2, we found out that the customers of this cluster travels an average distance of 1800 kms, books their rooms around 2 months in advance and stays 3 days in average.

Also, the probabilities of adults, children and room counts with respect to booking were good. So it makes sense that these features contributes abut 95% variance.

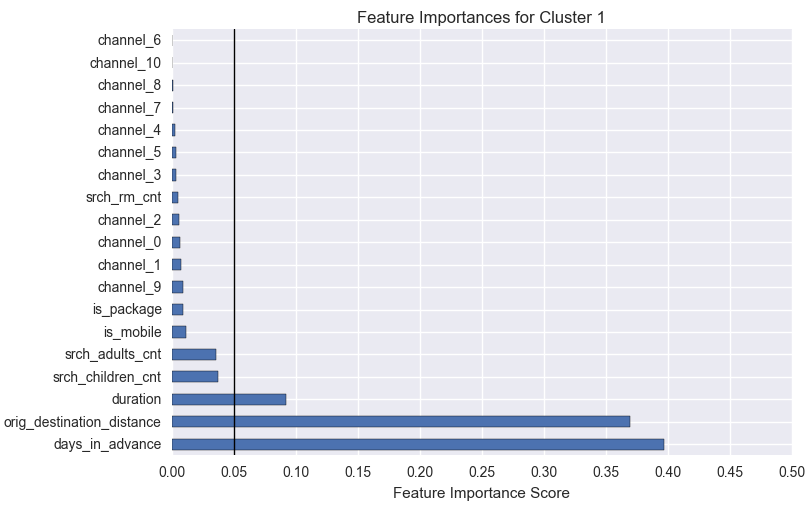
Whereas, the probabilities of mobile and package usage were very low, so they contribute very less.

Lets look at the coefficients, they are all in negative, i.e. booking rate tends to decrease for a each variable when others are held constant. That can be verified from the Odds Ratio.

For a unit increase in ‘X’ feature, booking rate tends to decrease by ‘x’ factor mentioned in OR column.



**Cluster - 1**



There are only 3 features that explains 95% of the variance.

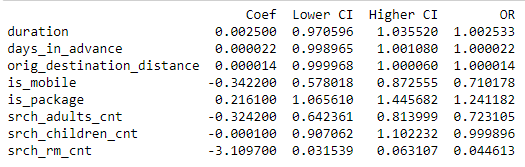
From Section 4.2, we found out that the customers of this cluster travels an average distance of 2454 kms, books their rooms around 3 months in advance and stays 5 days in average, i.e. they travel a long distance, books earlier and stays longer.

But the probabilities of booking for other features are very less.

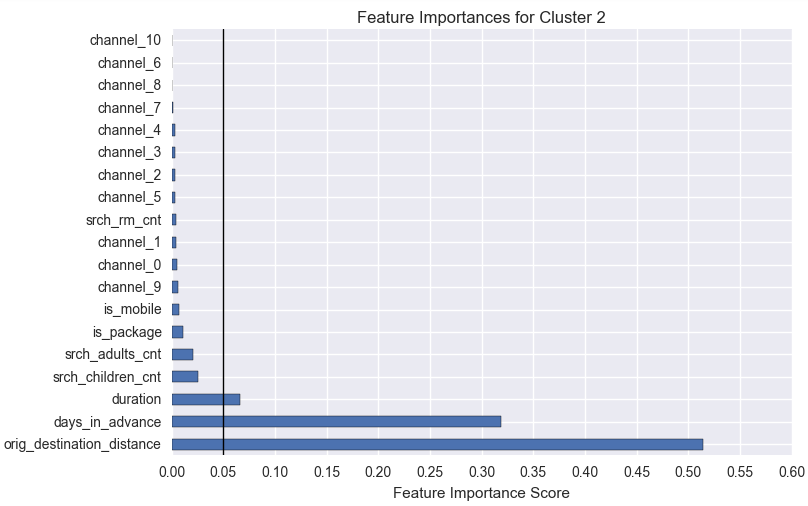
From adults and children count features, we came to know that this cluster consists of individuals or couples with zero or 1-2 children. These 2 features contributes about 6-7% in total.

Taking only the top 3 features, they have an positive impact on booking rate, and for an unit increase in that feature, booking rate tends to increase very slightly. Since, Odds ratios are very little over 1, we can say that booking rate remains same.

Although ‘is\_package’ contributes only about 1%, it can increase booking rate by a factor of 1.24 for an unit increase.



**Cluster - 2**



There are only 3 features that explains 95% of the variance.

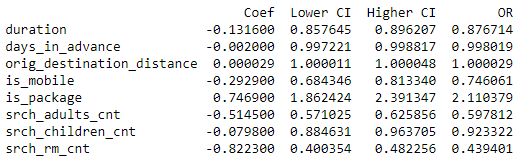
From Section 4.2, we found out that the customers of this cluster travels an average distance of 1525 kms, books their rooms around 1 and half months in advance and stays 2 days in average, i.e. they travel a short distance, books late and stays for a shorter number of days.

But the probabilities of booking for other features are very less.

From adults and children count features, we came to know that this cluster consists of individual people/couples as well as small families with not more than 3-4 adults and 6 children.

These 2 features contributes about 5-6% in total.

Features ‘orig\_destination\_distance’ and ‘is\_package’ have positive impact on booking rate. An unit increase in those features can increase booking\_rate by a factor of 1.000029 (remains same) and 2.11 respectively.



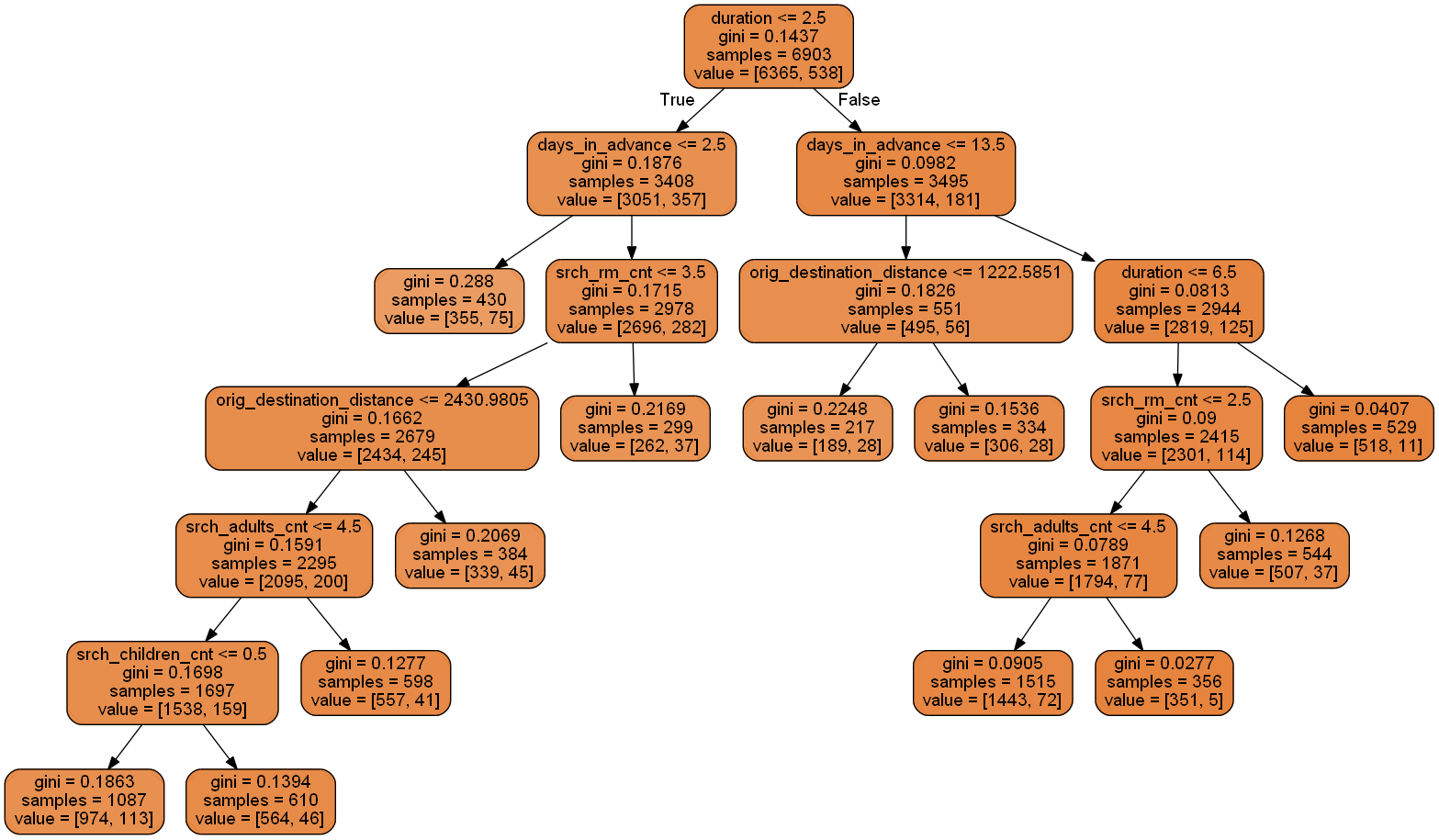
**Section 5: “What lead to a higher chance of booking for individuals in each Cluster?”**

**Cluster - 0**

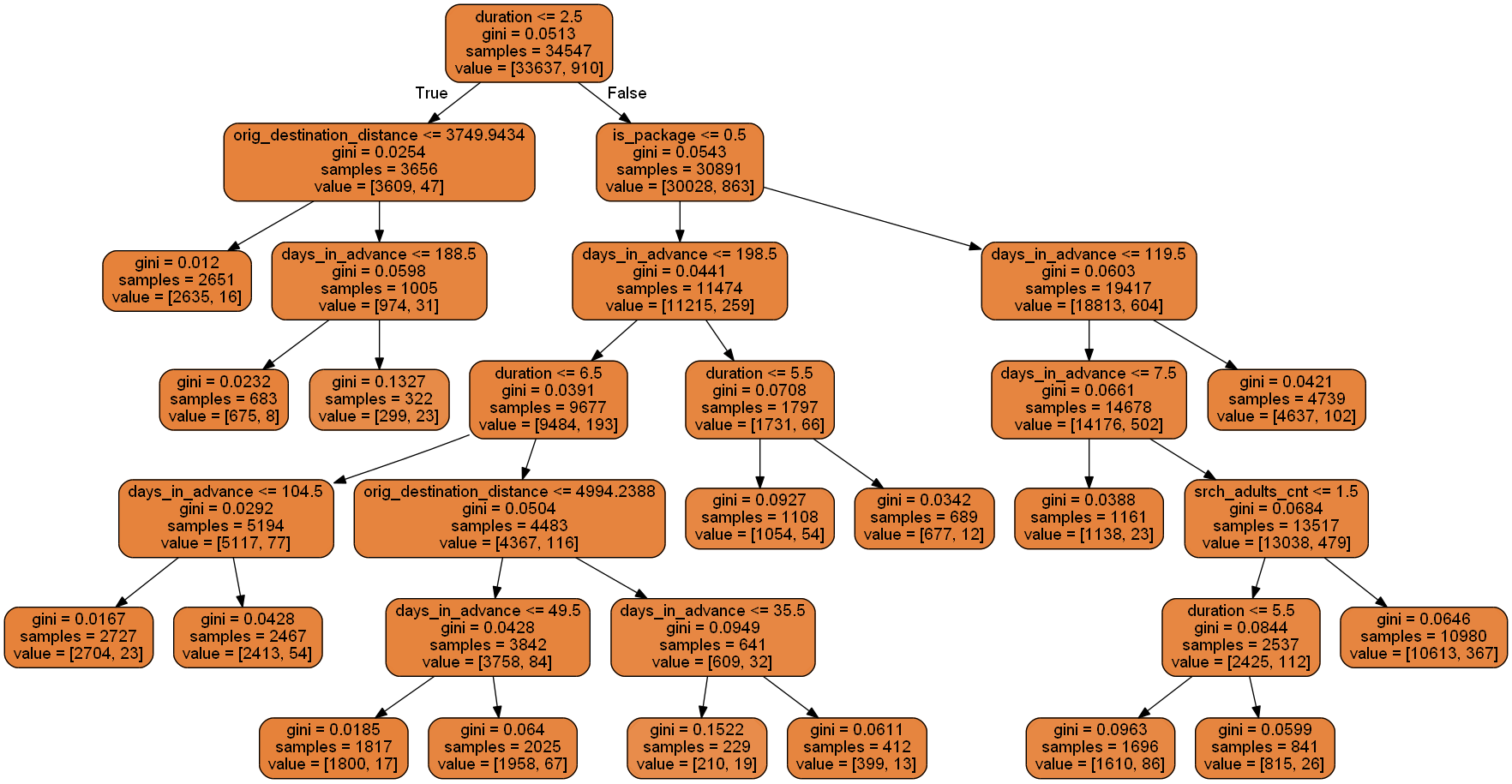
From Section 4.2, we saw that the percentage of Individual customers (i.e. Adults = 1) is about 30%, whereas the probability of booking is very low.

Since, Cluster 0 mainly consists of families who travels in groups, as can be seen from the number of adults, children and room counts, we do not get

any kind of information for individuals in our tree.



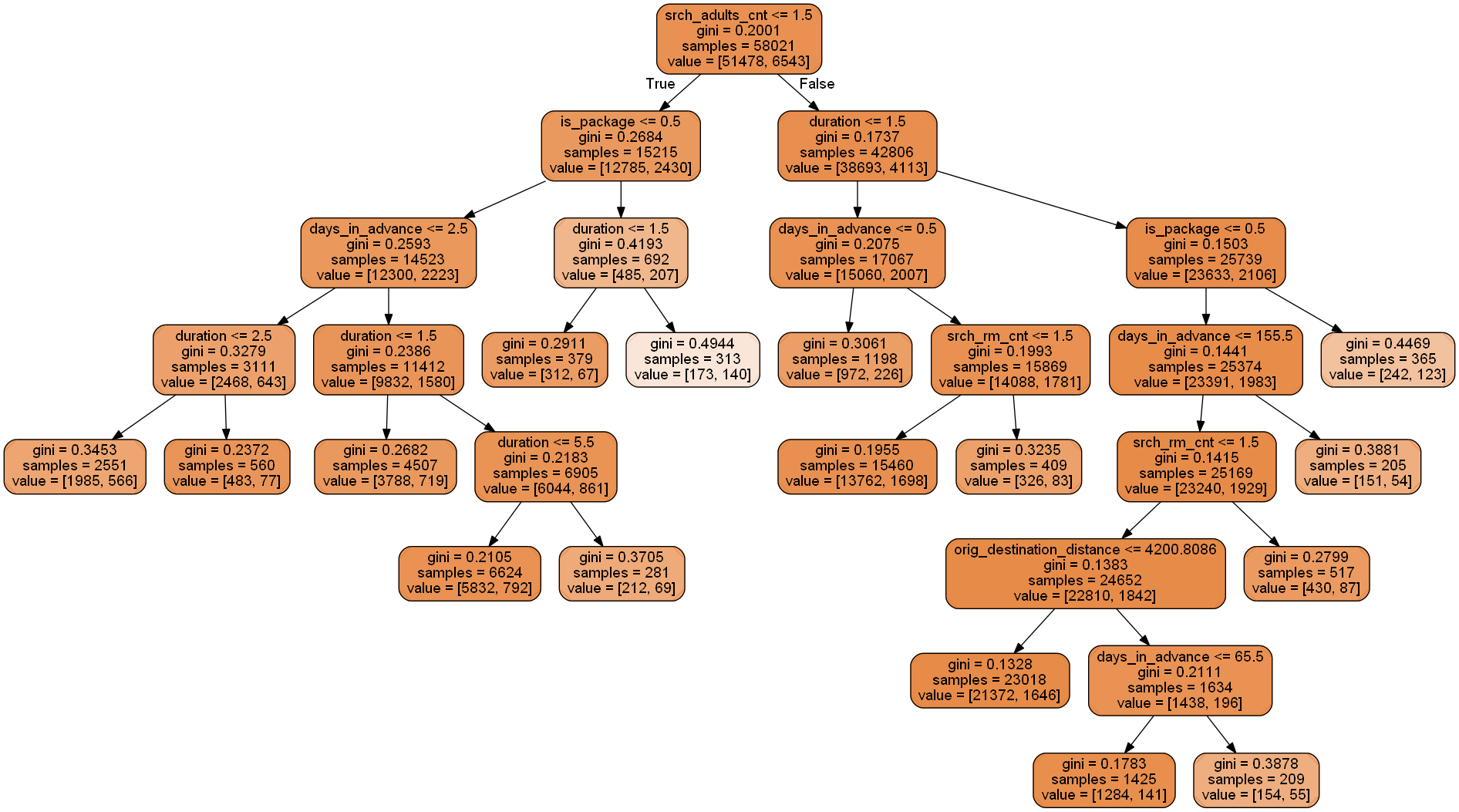
**Cluster - 1**



**Results:**

1. **Duration:** [2 – 5 days], **days\_in\_advance:** [8 – 119 days], **uses Package:** [YES] 🡪 **Probability of booking:** 5.07%
2. **Duration:** [above 5 days], **days\_in\_advance:** [8 – 119 days], **uses Package:** [YES] 🡪 **Probability of booking:** 3.09%

**Cluster - 2**



**Results:**

**Part 1:**

1. **Duration**: [ <= 2 days], **days\_in\_advance**: [<= 2 days], **uses Package**: [NO] 🡪 **Probability of booking**: 22.1%
2. **Duration**: [ > 2 days], **days\_in\_advance**: [<= 2 days], **uses Package**: [NO] 🡪 **Probability of booking**: 13.7%
3. **Duration**: [ 1 day], **days\_in\_advance**: [ <= 2 days], **uses Package**: [NO] 🡪 **Probability of booking**: 15.9%
4. **Duration**: [ 2 - 5 days], **days\_in\_advance**: [ <= 2 days], **uses Package**: [NO] 🡪 **Probability of booking**: 11.9%
5. **Duration**: [ > 5 days], **days\_in\_advance**: [ <= 2 days], **uses Package**: [NO] 🡪 **Probability of booking**: 24.5%

**Part 2:**

1. **Duration:** [ 1 day ], **uses Package:** [YES] 🡪 **Probability of booking:** 17.6%
2. **Duration:** [ > 1 day ], **uses Package:** [YES] 🡪 **Probability of booking:** 44.7%

**Conclusion:**

1. In Section 1, we handled missing values in our dataset.
2. In Section 2, we did some Exploratory Data Analysis and Feature Engineering on our Categorical and Numerical features. We saw how each features are distributed among customers.
3. In Section 3, we performed a Statistical test to find out Under-performing and Out-performing Marketing Channels from our data.

Later, we did A/B test with Hierarchical Modelling on our Out-performing Marketing Channels. We sampled each channel from our distribution to find out the probabilities of booking.

1. In Section 4, we used Within Cluster Sum of Squares (WCSS) to find the optimal number of clusters for this data and used KMeans clustering from SCIKIT-Learn module to cluster the data.

Then, we described each cluster in terms of booking rate and found out how the features of each cluster affects the booking rate.

Lastly, using RandomForestClassifier() from SCIKIT-learn module, we calculated the feature importances of each cluster. And then, using LogisticRegression, we calculated the coefficients and Odds-Ratios of each features.

1. In Section 5, using DecisionTreeClassifier() from SCIKIT-learn module, we generated trees for each cluster and examined the factors that leads to higher booking rate for Individual customers.

**References:**

1. A/B Test : **https://blog.dominodatalab.com/ab-testing-with-hierarchical-models-in-python/**