# Visualization of data from e-commerce company .

## **Data facts**

Number of rows of data available: 400,000

Number of columns of data available : 13

### **Column Fields:**

* Ip – User IP Address. Duplicates present. Numeric field. Int field
* login\_or\_not - Login activity for the session. Binary field
* #page\_view  - Total No of pages visited on the website. Numeric field. Int
* page\_view\_time\_seconds\_avg – Average time spent in each page in seconds. Float field
* #page\_product\_view - Total No of pages that are products (will be generally less than page\_view)
* page\_product\_time\_avg - Average time spent on products page
* number\_searches – No of searches done in the website. Int field
* #times\_check\_shoppingcart – No of time the shopping cart has been viewed. Int field
* read\_Q&A – User read Q&A for product. Binary field
* read\_product\_definition – User read product def. Binary field
* previous\_site –
  + e-commerce
  + search engine
  + email
  + advertisement
  + other
* time\_of\_day –
  + morning
  + noon (11:30 - 2pm)
  + afternoon
  + evening (after 6pm)
* purchase – Purchase decision

## **Sampling the data**

Since the data is too huge to analyze, we take sample of the data to drill down to 10000 rows.

## Data Visualization using python in Jupyter notebook

We plot the purchase behavior vs the predictors and do the statistical analysis of the data.

### Purchase or no purchase vs time spent in viewing the landing page



Findings:

We can see that maximum number of purchases are highest within 120 seconds / 2 minutes of landing page views.



Findings: Whenever there is no purchase, the customer does not view the landing page for more than 1.5 minutes generally.

**Targeted visualization to select the fields for data models**

**Technique**

To understand the role of each field and whether or not they should be considered for the model the following steps have been performed:

Step 1: Separate the dataset into purchase and non-purchase entries

Step 2: Pick-up a field for each dataset.

Step 3: Study the distribution of the field over each data-set. Percentage of positive entries to total entries in case of binary fields and histogram in case of continuous numeric fields.

Step 4: Look for differences between the percentage/graph of both the datasets. If they are similar/close, ignore the field as it is not an indicator of purchase/non-purchase decision. If not, pick-up.

**Observation**

1. **IP Address:** Not in scope right now. Please refer to suggestions for the further steps for this particular field.
2. **Login or not:** **Should be considered.** Difference of 25% observed between purchase and non-purchase datasets.
3. **Page View & Page View time seconds avg:** Instead of individual consideration, the product of these fields should be **stored in a new field and considered**.
4. **Page Product view & Page product time avg:** The product of these fields into a new field has been studied for the datasets and it shows close to no variation. Hence, both the **fields can be ignored.**
5. **Number Searches:** Similar trend for both datasets. **Can be ignored.**
6. **Times check shopping cart:** Similar trend for both datasets. **Can be ignored.**
7. **Read Q&A:** **Should be considered.** Difference of 34% observed between purchase and non-purchase datasets.
8. **Read Prod Definition:** **Should be considered.** Difference of 26% observed between purchase and non-purchase datasets.
9. **Previous site:** Similar trend for both datasets. **Can be ignored.**
10. **Time of the day:** **Should be considered.** Considerable difference in trend between the two models.

**Suggestion**

1. **IP Address:**

An algorithm for rating an IP can be introduced based on the traffic from it. This will be useful to quantify the IP and incorporate it into the model. A method where the input is the IP address and the output is its rating can be a good approach. Should be picked up as our next objective.

1. **Classification fields:**

Fields like time of the day can be made into separate binary fields. Since by default one of them will always be one, only 3 fields indicating 3 times of day will be the most efficient way to move forward.

1. **Fields to be considered for Model:**
   1. IP address (post rating system implementation)
   2. Login or not
   3. Total time (page view X page view time avg sec)
   4. Read Q&A
   5. Read Product definition
   6. Time of the day (Post classification)