**Discrete Questions**

* for all questions below provide the code you used to obtain your answer -within which place any clarifying code comments you would like to make.
  + How many distinct ads were shown for the 10 day period in total and per day?
  + **Answer:**
    - Code in answer\_ad\_count.py
    - #rows in month=[4122995, 5337126, 3870752, 3335302]
    - len dicts 141021 4122995
    - len dicts 141022 5337126
    - len dicts 141023 3870752
    - len dicts 141024 3335302
    - #rows parsed=40428967
    - #rows in month=[3363122, 3835892, 3225010, 5287222]
    - len dicts 141025 3363122
    - len dicts 141026 3835892
    - len dicts 141027 3225010
    - len dicts 141028 5287222
    - #rows parsed=40428967
    - #rows in month=[3832608, 4218938, 0, 0]
    - len dicts 141029 3832608
    - len dicts 141030 4218938
    - In total we have 40428967 distinct ads.
  + Which app\_domain is the most successful for the 10 day period?
  + **Answer:**
  + Code in answer\_success.py. Note: colName variable was changed to obtain results for the different questions. The calcBin function was used from the web.
  + In this result set there was no minimum #trials applied in the ranking. Results are sorted based on the HitRate99%lowerBound (refer to the success criteria question for details.)
  + Answer format: #rank=app\_domain: [#hits, #trials, hitRate, HitRate99%lowerBound, HitRate99%upperBound]
    - #distinct members=559
    - #rows parsed=40428967
    - # 1 = 99b4c806 : [6434, 10041, 0.6407728313912957, (0.6282587209039416, 0.6530196283519374)]
    - # 2 = f5a7c834 : [103, 166, 0.6204819277108434, (0.51460416873772, 0.7120982644086825)]
    - # 3 = 73fc6786 : [66, 105, 0.6285714285714286, (0.49328397354989684, 0.7406783193912148)]
    - # 4 = 5576cc84 : [17, 21, 0.8095238095238095, (0.4798977591774681, 0.9473814530806107)]
    - # 5 = eaaf6d12 : [33, 53, 0.6226415094339622, (0.4274355570475261, 0.7747196621365018)]
  + **In this second result set there is a minimum #trials of 100 applied in the ranking. Results are sorted based on the HitRate99%lowerBound (refer to the success criteria question for details.)**
  + **Answer format: #rank=app\_domain: [#hits, #trials, hitRate, HitRate99%lowerBound, HitRate99%upperBound]**
    - # 1 = 99b4c806 : [6434, 10041, 0.6407728313912957, (0.6282587209039416, 0.6530196283519374)]
    - # 2 = f5a7c834 : [103, 166, 0.6204819277108434, (0.51460416873772, 0.7120982644086825)]
    - # 3 = 73fc6786 : [66, 105, 0.6285714285714286, (0.49328397354989684, 0.7406783193912148)]
    - # 4 = 5bcedd7d : [60, 125, 0.48, (0.36040169852120535, 0.5938310169038319)]
    - # 5 = b97def0d : [52, 133, 0.39097744360902253, (0.28159730825851215, 0.5026911265814482)]
  + Which banner position proved to be the most successful for the 10 day period?
  + **Answer:**
    - # 1 = 7 : [29628, 13949, 0.3201000527801363]
    - # 2 = 4 : [6276, 1428, 0.1853582554517134]
    - # 3= 1 : [9182118, 2065164, 0.18361449459522755]
    - # 4 = 3 : [1663, 372, 0.1828009828009828]
    - # 5 = 0 : [24327689, 4781901, 0.16427235835338114]
    - # 6 = 5 : [5076, 702, 0.12149532710280374]
    - # 7 = 2 : [11451, 1550, 0.11922159833858934]
  + Provide the top 10 most successful site\_categories for the 10 day

period?

* + **Answer:**
  + #distinct members=26
  + #rows parsed=40428967
    1. dedf689d : [11907, 12593, 0.514]
    2. 3e814130 : [2187061, 863245, 0.2830027544777475]
    3. 42a36e14 : [1934, 581, 0.23101391650099404]
    4. 28905ebd : [5842608, 1534600, 0.20801907713595713]
    5. f028772b : [10384133, 2272940, 0.1795786434983823]
    6. 74073276 : [12, 2, 0.14285714285714285]
    7. 70fb0e29 : [20926, 3298, 0.1361459709379128]
    8. 50e219e0 : [14410884, 2126350, 0.1285795435923565]
    9. c0dd3be3 : [37376, 4714, 0.11199809931100024]
    10. 75fa27f6 : [143070, 17915, 0.11128365996832003**]**
* What was the criteria for success that was employed for the questions above?
  + **Answer:**

Per category: HitRate=#hits/#records

Next we find the 99% binomial test confidence bound over the above HitRate measure i.e. P(HitRate>bound)=0.99. This bound will take the number of trials into consideration and hence will be always smaller than the measured HitRate. When there are higher number of records in the data set (actual trials of the ad) the bound will be closer to the sample HitRate, while in cases where fewer trials of ads are available, the bound will be further smaller than the measured HitRate (e.g. rank#4 vs. rank#1).

One might also prefer to apply a threshold for the minimum number of trials per category. In the app\_domain success analysis section, results from both methods are reported. As can be seen the top three app\_domains are identical in both methods.

A next direction to further improve the success analysis can involve the breaking down of each app\_domain into relevant subcategories e.g. site\_category etc., apply the above analysis on each subcategory and finally consolidate the measures to obtain the app\_domain success index. This method will make sure that the success of an app\_domain is measured across various subareas and hence is balanced rather than a single general metric.

**Data insights**

* This section of the assignment is really asking the candidate to provide insights as to:
  + Any trends discovered.
  + Any links or relationships between the fields that can be gathered.
* There is no clear expected answer for this component.
* Whatever insights are discovered or conclusions reached should be articulated clearly and concisely.
* Visualisations of data should be provided
  + though no specific tools are required, the ggplot library either within R or python is easy to use and has much in the way of online support
  + Alternatively matplotlib is also a library that has been explored within Mediative.

For the analysis of data in this section, we have taken a 10% random sample of rows in the original csv file (answer\_sampler.py) and stored the resulting data set in sample.csv file. This way we can run the analytical code and plotting (answer\_explore.r) on a data set that fits into memory but still preserves the patterns in the original data set.

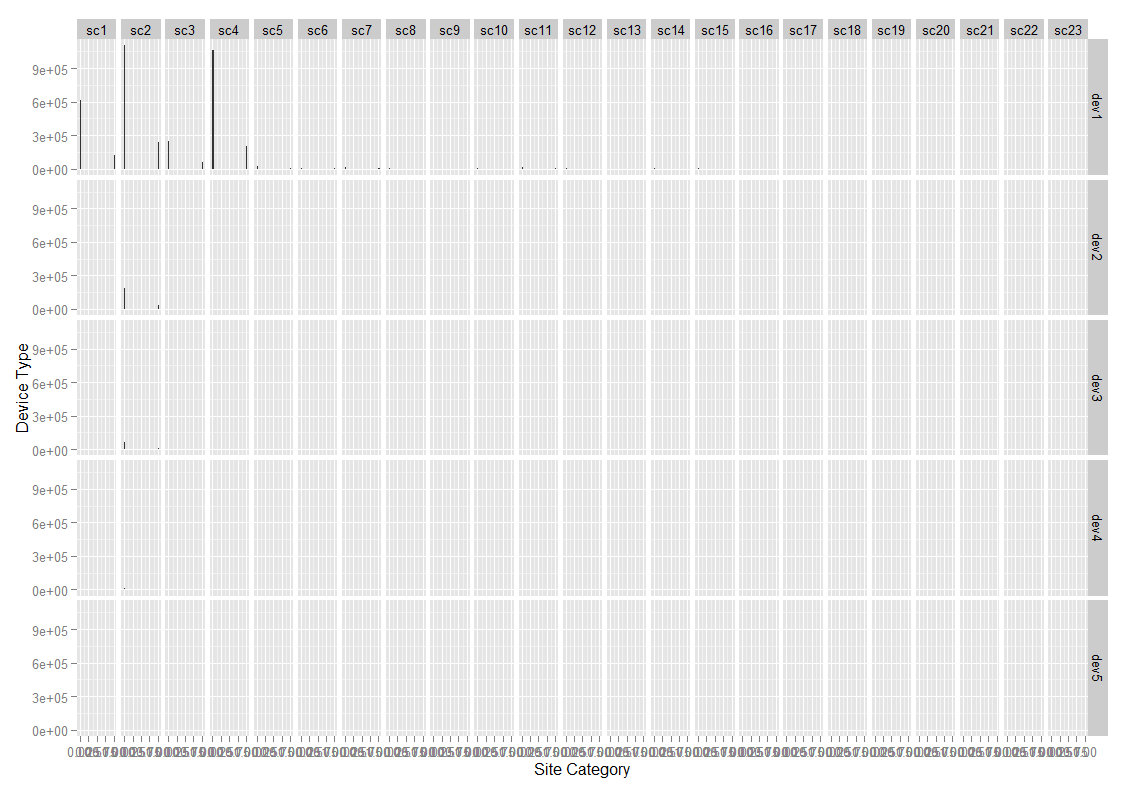


Figure .Click histograms per site category and device type

Shows that most of our data is from device dev1 and site categories sc1 through sc4. dev2 and dev3 are the next frequent devices which are almost only used with sc2.

No apparent difference in hit rate is observable.

Based on unique device\_id analysis, there is about 5036000 unique devices in our database of about 40,428,000

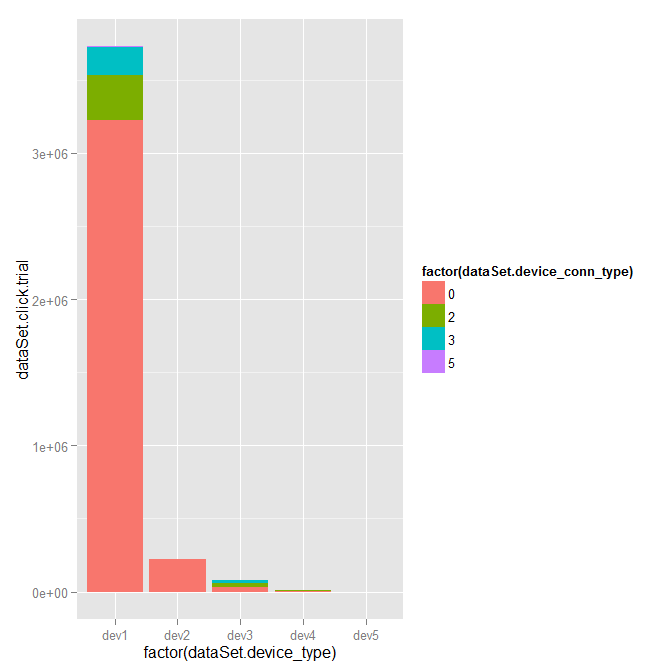


Figure number of trials per device and device connection type.

Shows that most of our users use device 1 and connection type 0.

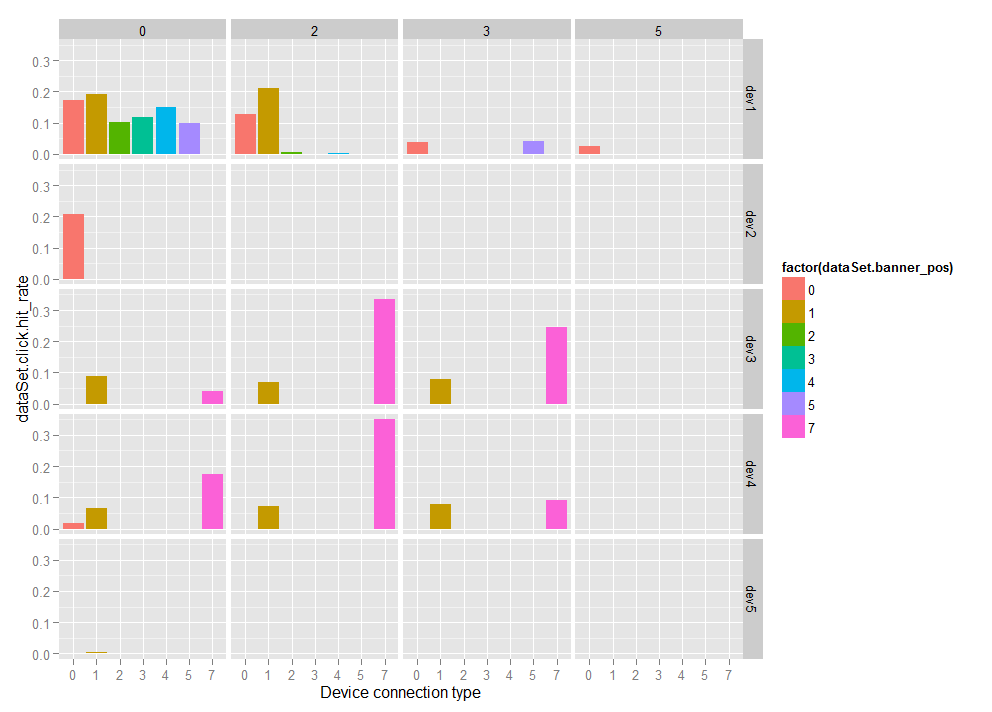


Figure The 99% confidence lower bound on the hit rate for various positions in different devices and connection types

This visualizes the superior success rate of the position 7 broken down by the device and connection attributes. For the far top left bar plot where we have a lot of records, positions 0 and 1 seem to be more successful.

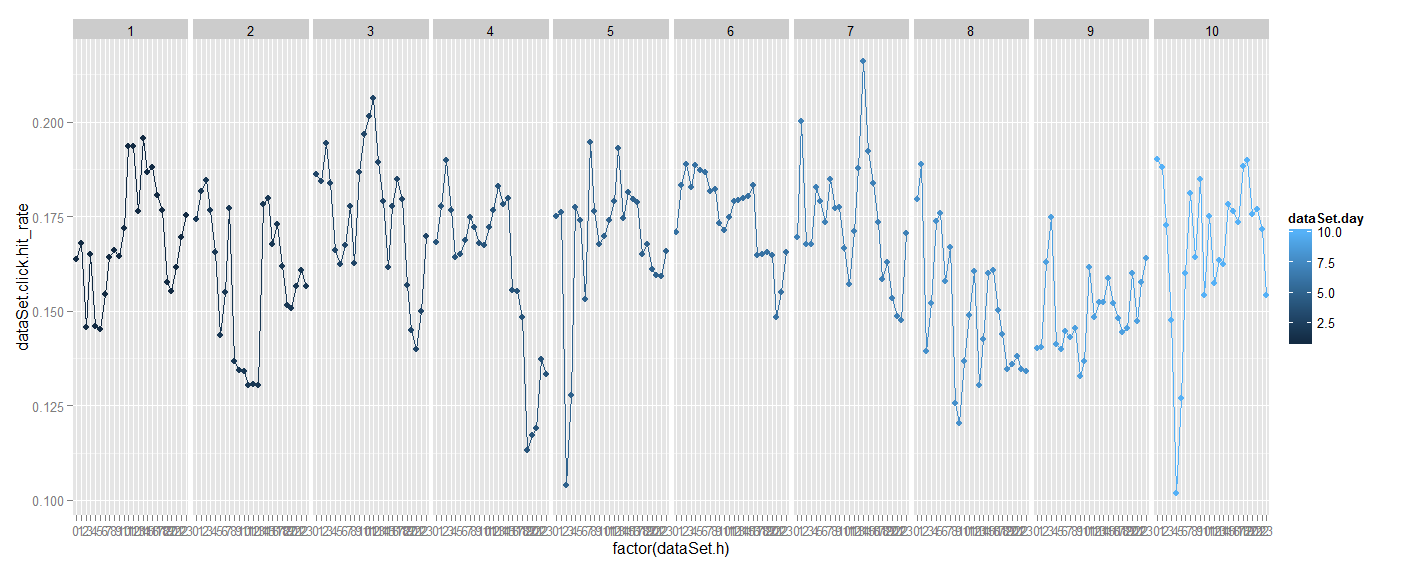


Figure hit rate change through consecutive hours within the 10 days of data.

This graph shows how the hit rate has changed through the 10 days and through the hours in each day. It can be noticed that there is significant autocorrelation in the hit rate values (i.e. the hit rate is highly correlated with the previous hour hit rate). To get a general idea about the dependency of hit rate on hour of day the following figure shows that we have our highest success rates hours 1 and 15.

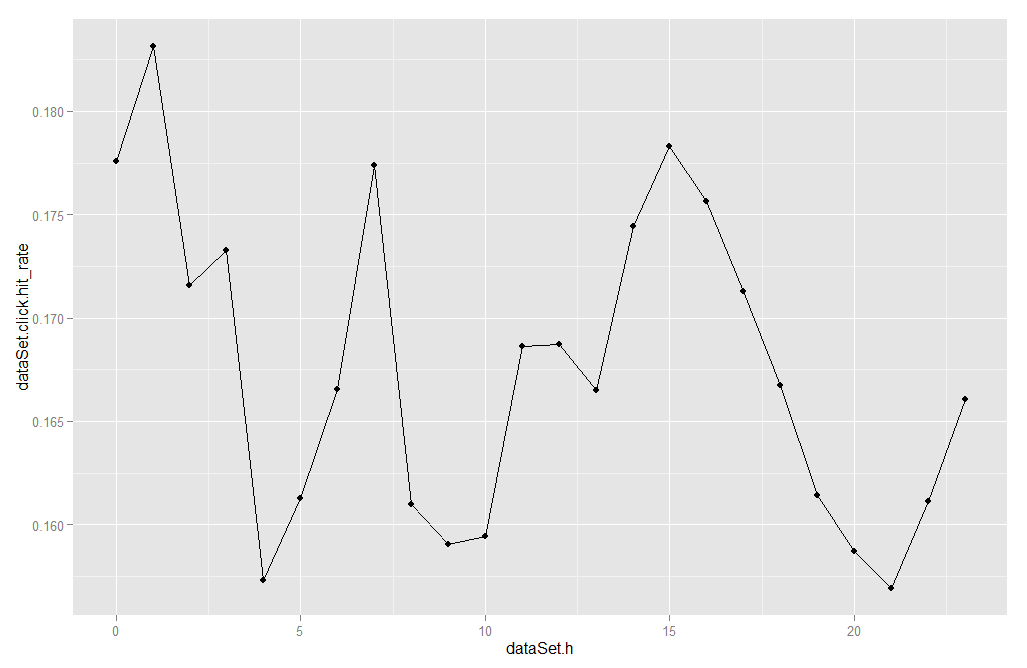


Figure overall hourly hit rate trend

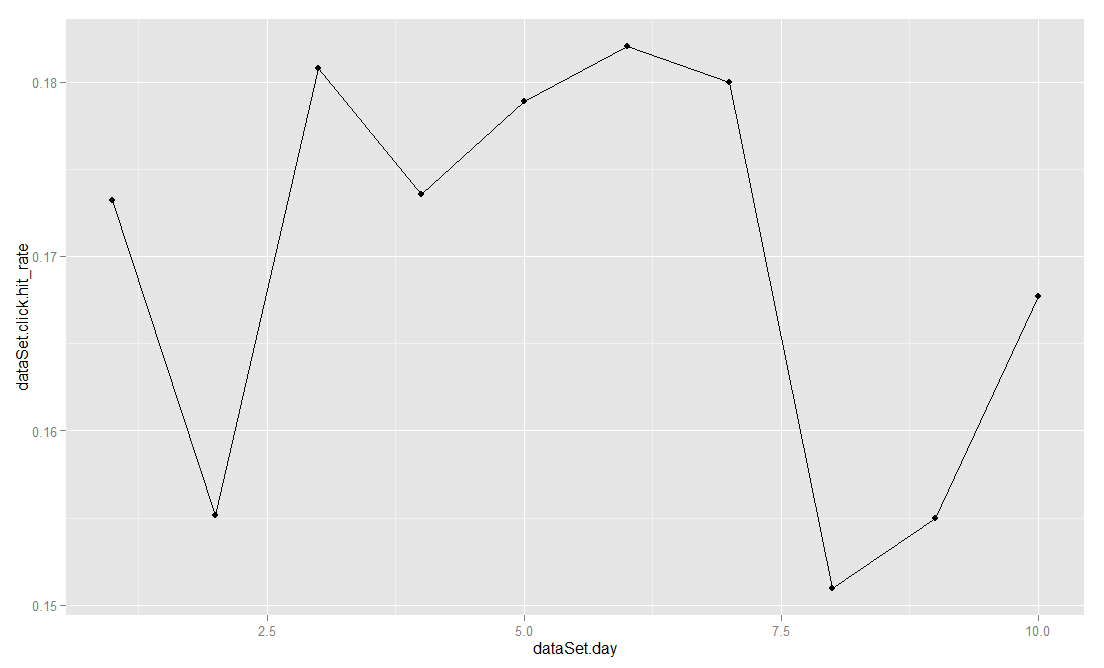


Figure overall daily hit rate trend

This graph shows that we had best hit rates on the days 3 through 7 and had a sharp drop on the 8th day.

Of course for further analysis each graph and data point in it (e.g. day 8 in the above graph) can possibly be drilled down to see the role of other attributes e.g. site\_category etc.

|  |  |  |
| --- | --- | --- |
| Id | 1. #distinct in sample data | 1. column |
|  | 1. - | 1. id: ad identifier |
|  | 2 | 1. click: 0/1 for non-click/click |
|  | 240 | 1. hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC. |
|  | 7 | 1. C1 -- anonymized categorical variable |
|  | 7 | 1. banner\_pos |
|  | 3396 | 1. site\_id |
|  | 4234 | 1. site\_domain |
|  | 23 | 1. site\_category |
|  | 4852 | 1. app\_id |
|  | 319 | 1. app\_domain |
|  | 26 | 1. app\_category |
|  | 503664 | 1. device\_id |
|  | 1634499 | 1. device\_ip |
|  | 6363 | 1. device\_model |
|  | 5 | 1. device\_type |
|  | 4 | 1. device\_conn\_type |
|  | 2466 | 1. C14 -- anonymized categorical variables |
|  | 8 | 1. C15 |
|  | 9 | 1. C16 |
|  | 431 | C17 |
|  | 4 | 1. C18 |
|  | 67 | 1. C19 |
|  | 166 | C20 |
|  | 60 | 1. C21 |