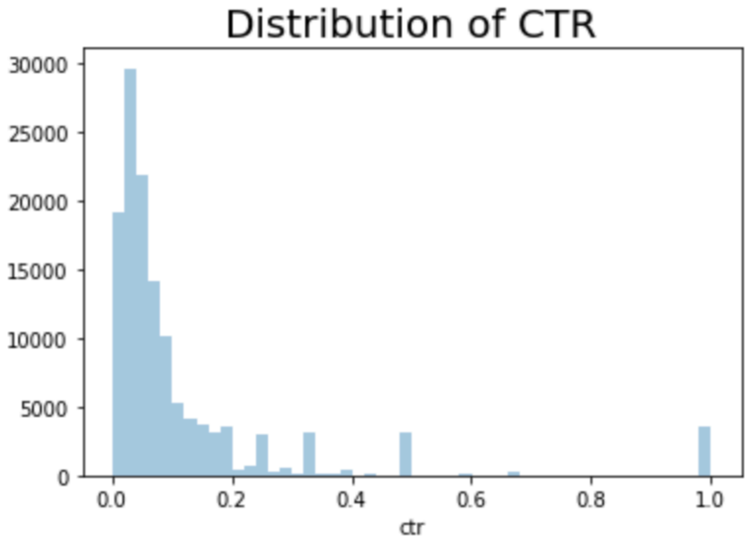
Analyse Click-through rates for hotels

Tanya Yi-En Chen
Outlining all the answers to the tasks and explain your thought process.

Part 1 – Clicked items investigation

The dataset contains information on advertised items online with the number of impressions, clicks and information on displayed positions for each item.

1. Calculate the CTR of each item. What is the overall avg CTR?



By dividing the number of clicks by numbers of impression, click-through rate for each item is calculated. The distribution of the frequency of CTR is as shown in the histogram above.

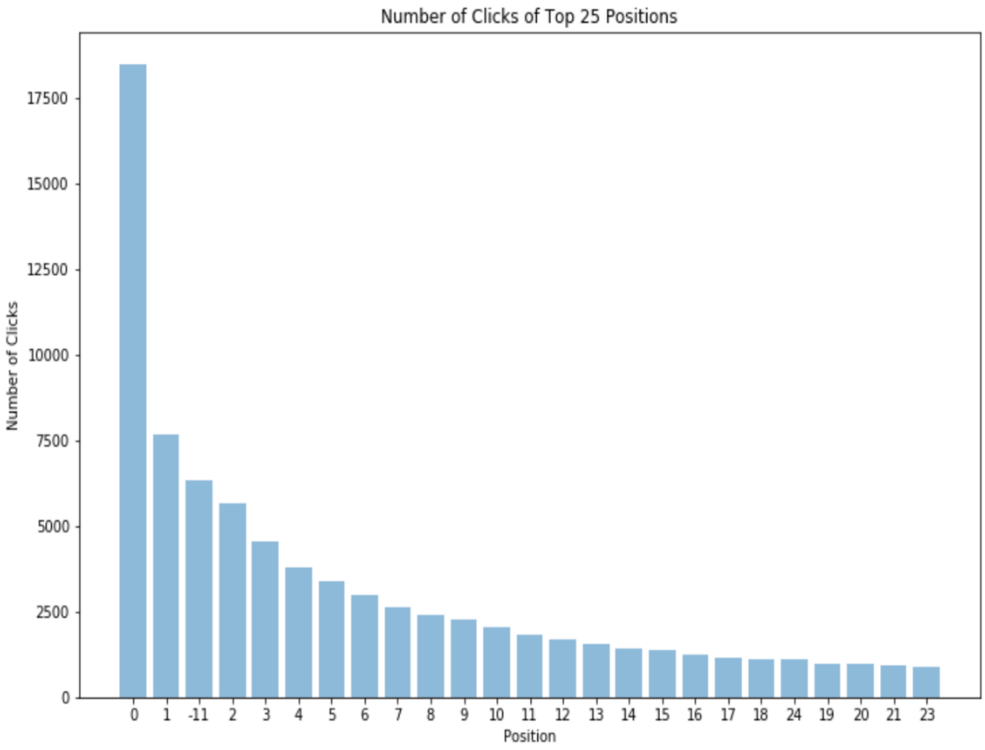
It can be observed that the distribution of CTR is highly positive skewed with most of distribution spread between 0 and 0.2. The average CTR is 11.46%.

It is also worth investigating what drive the extreme values. For example, influencer effect.

2. What is the distribution of clicks among the top 25 positions? What is the share of the first positions? On how many positions are approx. Half of the click-outs made?

The following table shows the top 25 positions with the most clicks. The distribution is shown as the figure below. It is observed that the pages in the front got more clicks, which makes sense as users start browsing from the first page and would not necessarily click through all pages. The first position is the first page with 18,505 clicks, taking up 23.28% of total clicks. On the top 5 most clicked positions are approximately half of the click-outs made, which takes up 49.9% of all clicks, meaning that on average around half of users click an item without browsing through later than first few pages.

```
Positions   count
  0        18505
  1         7693
  -11       6352
  2         5672
  3         4555
  4         3805
  5         3402
  6         3001
  7         2629
  8         2426
  9         2248
  10        2068
  11        1836
  12        1691
  13        1552
  14        1433
  15        1364
  16        1255
  17        1174
  18        1127
  24        1092
  19         993
  20         981
  21         912
  23         867
```



3.Describe the relationship between the average displayed position and the clicked displayed position. What are your thoughts about the variance between the two?

The Pearson’s correlation coefficient is 0.45 with a p-value smaller than 0.

The average displayed position and the clicked displayed position are significantly positively related to each other. This makes sense as if a position is displayed more, there is more chance that the position would get a click.

4. In the dataset, we provided you with the average displayed position. What can be wrong with using averages?

If an outlier exists in the dataset, it would affect the average and may be biased in terms of presenting a truthful representation of the majority. In the case with outliers that can largely affect the analysis, median would be better than mean.

Part 2 – Session investigation

1. ***Describe the data set that you have received. Calculate the 5 most frequent values per column (with frequency). Can you find any suspicious results? If so, what are they? And how would you fix these for the analysis?***

The dataset provides data on each session with a click on an item. Before starting to analyse the data, it can be observed that there is one missing value in the dataset. Compared to the sample size, dropping the missing value would not lead to a big impact on the analysis and therefore the missing value is excluded.

```
Departure_days  count
 4.0           59506
  1.0           52780
 -1000000.0     52530
  5.0           51982
  6.0           48879
```

By examining the top 5 most frequent values for each column, it can be observed that in variable "departure_days" the number -1000000 does not make sense. Since within -1000000, there are 52530 samples which is too big to be eliminated without affecting the overall insight. The distribution of the departure days is then examined.

```
Departure_days  count
  -1000000.0     52530
   0.0             265
   1.0           52780
   2.0           44105
   3.0           43034
   4.0           59506
   5.0           51982
   6.0           48879
   7.0           44781
   8.0           41252
   9.0           36211
   10.0          32743
   11.0          29313
   12.0          28089
   13.0          26433
   14.0          27037
```

```
15.0         26364
16.0         22884
17.0         20486
18.0         19033
19.0         31727
20.0         32092
21.0         32107
22.0         35603
23.0         31811
24.0         28078
25.0         26211
26.0         14077
27.0         14366
28.0         14319
```

It is observed that most of the distribution are within 10 days after the search. Whilst the first 15 days are all included in the data, logically, day 0 with 265 instances is too small as it is not uncommon for people to decide on searching for a hotel on a very short notice. Therefore, my judgement would be combined the -100000 data with day 0 data. Of course, the most accurate measure would be to confirm with the data provider.

2. Which search type has the lowest average displayed position? What is the best sorting order for this search type? Which search type should be excluded for a statistical reason?

By sorting the data by its search type and calculate the average displayed positions, the search types from the lowest to the highest average displayed position are as the following table shows:

```
search_type  avg displayed pos
   2116.0    1.372424
   2100.0    2.500000
   2114.0    3.805842
   2113.0    4.890490
   2115.0    5.346740
   2111.0    5.819748
```

Type 2116 has the lowest average displayed position. Next its best sorting order is calculated. Finally, within search type 2116 the frequency of each sorting order is calculated.

```
Sorting_order  count
     12.0     278270
     312.0     64087
     21.0      17729
     212.0      7449
     412.0      4970
     41.0       1679
     112.0       988
     32.0        224
     0.0         187
```

It shows that the sort order 12 is the best sorting order for the search type 2116 with the frequency of 74% out of all occurrences.

```
Search_type  count
   2113.0    928598
   2116.0    375583
   2111.0    230316
   2115.0    194310
   2114.0    174986
   2100.0         2
```

Looking at other search types, the item "2100" only has two observations which are too small compared to the overall sample. Therefore, it can be dropped to provide a better insight of the dataset.

3. What are the top 10 "best" and "worst" performing items? Explain what metric you have chosen to evaluate the performance and why.

Click-through rate (CTR) is the metric used here to evaluate each item's quality, as the main objective of the analysis is to find out how many clicks are generated when the items are displayed.

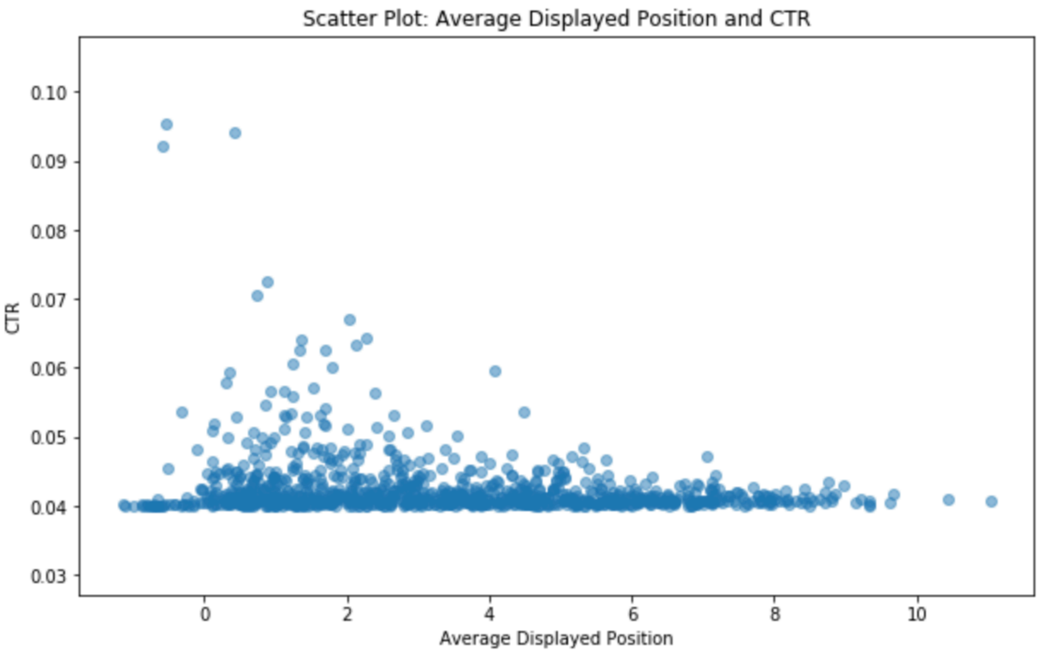
The 10 items with the best CTRs and 10 items with the worst CTRs are shown as follows:

```
Best CTR items:  CTR
    509976.0     1.0
    3850512.0    1.0
    4196870.0    1.0
    4055884.0    1.0
    2678192.0    1.0
    6690574.0    1.0
    1076380.0    1.0
    7094708.0    1.0
    3935158.0    1.0
    2381760.0    1.0
```

```
Worst CTR items  CTR
   2089262.0    0.04
   2089320.0    0.04
   2089486.0    0.04
   2089666.0    0.04
   2089916.0    0.04
   2090234.0    0.04
```

```
2090478.0    0.04
2090572.0    0.04
2090882.0    0.04
5001.0       0.04
```

4. Describe and visualise the relationship between the average displayed position and the CTR among the top 1000 most clicked items.



Above is the scatter chart demonstrating the relationship between average displayed position and each item’s CTR. It is observed that average displayed position and CTR are largely uncorrelated with each other with statistical significance. The displayed position of each item does not largely affect the item's performance, and therefore other indicators that may interact with CTR are worth investigating in order to derive the most optimal and effective search for the users.