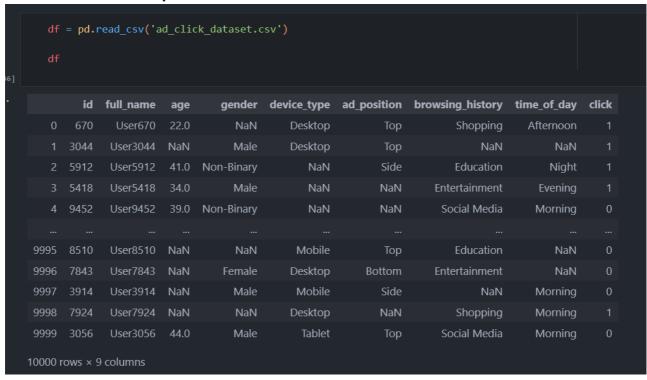
STTAI - Lab Assignment 10

Name	Roll Number			
Romit Mohane	23110279			
Rudra Pratap Singh	23110281			

Using real-world data, this assignment will introduces us to key concepts in A/B testing and Covariate Shift Detection. We performed hypothesis testing using the scipy library and identified distributional shifts in datasets using classification-based techniques.

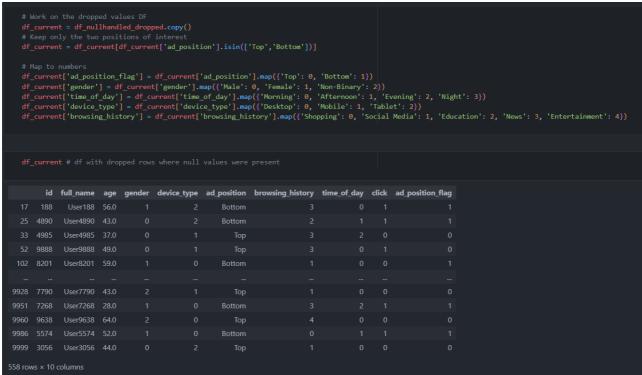
Part 1: A/B Testing using Ad Click Prediction

1. Load the Dataset into a pandas df

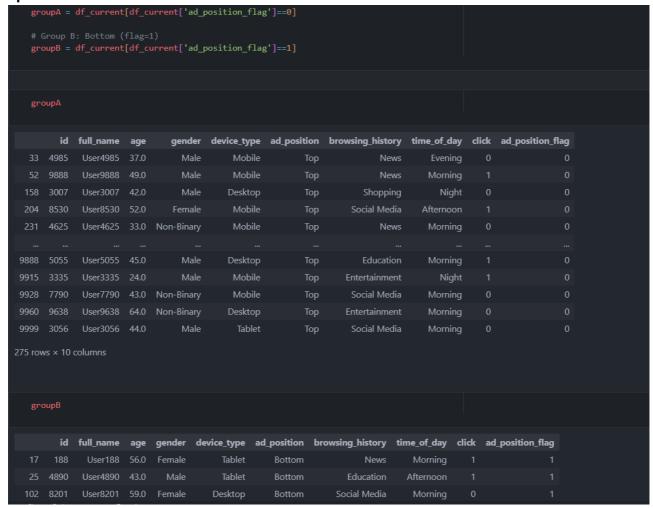


2. Perform necessary data cleaning and preprocessing:

Perform necessary data cleaning and preprocessing:										
df_nullhandled_dropped										
	id	full_name	age	gender	device_type	ad_position	browsing_history	time_of_day	click	
17	188	User188	56.0	Female	Tablet	Bottom	News	Morning	1	
25	4890	User4890	43.0	Male	Tablet	Bottom	Education	Afternoon	1	
33	4985	User4985	37.0	Male	Mobile	Тор	News	Evening	0	
52	9888	User9888	49.0	Male	Mobile	Тор	News	Morning	1	
102	8201	User8201	59.0	Female	Desktop	Bottom	Social Media	Morning	0	
9951	7268	User7268	28.0	Female	Desktop	Bottom	News	Evening	1	
9952	5912	User5912	41.0	Non-Binary	Mobile	Side	Education	Night		
9960	9638	User9638	64.0	Non-Binary	Desktop	Тор	Entertainment	Morning	0	
9986	5574	User5574	52.0	Female	Desktop	Bottom	Shopping	Afternoon	1	
9999	3056	User3056	44.0	Male	Tablet	Тор	Social Media	Morning	0	
816 rov	816 rows × 9 columns									
<pre># Work on the dropped values DF df_current = df_nullhandled_dropped.copy() # Keep only the two positions of interest df_current = df_current[df_current['ad_position'].isin(['Top','Bottom'])]</pre>										
	<pre># Map to numbers df_current['ad_position_flag'] = df_current['ad_position'].map({'Top': 0, 'Bottom': 1}) df_current['gender'] = df_current['gender'].map({'Male': 0, 'Female': 1, 'Non-Binary': 2})</pre>									



3. Split the dataset



1.__ Use the statsmodel's proportions_ztest function to perform an independent two-sample z-test

between Group A and Group B._

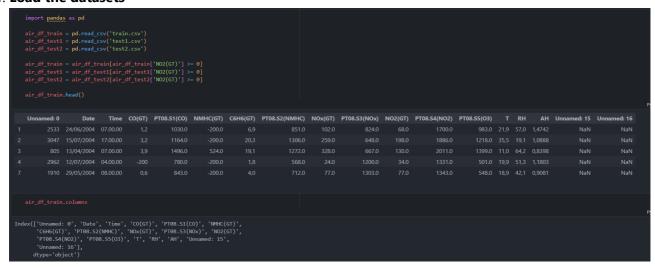
```
clicks = [ groupA['click'].sum(),
              groupB['click'].sum() ]
   # number of users in each group
   nobs = [ len(groupA),
              len(groupB) ]
   print(clicks[0]/nobs[0], clicks[1]/nobs[1], "\n")
   z_score, p_value = proportions_ztest(count=clicks, nobs=nobs)
   print("z-score:", z_score)
   print("p-value:", p_value)
   if p_value < 0.05:
       print("Reject null hypothesis: the two CTRs are not equal.")
   else:
       print("Fail to reject null hypothesis: the two CTRs are equal.")
0.6327272727272727 0.6784452296819788
z-score: -1.1365075404030447
p-value: 0.2557442115851094
Fail to reject null hypothesis: the two CTRs are equal.
```

4. Interpret the result: Is there a statistically significant difference in click-through rates between the two groups? Justify your answer. In our A/B test on 10,000 users (with missing values dropped), we compared click-through rates (CTRs) for ads shown at the Top vs. Bottom positions. Using a two-sample z-test, we obtained z = -1.137 and p = 0.256 (> 0.05), so we accept H₀ that the two CTRs are equal. The negative z-score indicates Bottom-positioned ads achieved a lower CTR than Toppositioned ads.

But, this difference is statistically insignificant.

Part 2: Covariate Shift Detection Using Air Quality Data

1. Load the datasets



2. For each test dataset (test1.csv and test2.csv), compare it with train.csv using the Kolmogorov–Smirnov test (scipy.stats.ks_2samp). Perform the KS test on the NO2(GT) column to identify whether there are any distributional differences

```
ks_statistic_test1, p_value_test1 = ks_2samp(air_df_train['NO2(GT)'], air_df_test1['NO2(GT)'])
print("No*ken for test set 1:", air_df_test1['NO2(GT)'].mean())
print("KS Test for test1.csv")
print("KS Statistic: {ks_statistic_test1}")
print("Fv-value: {p_value_test1}")

# Perform KS test for test2.csv
ks_statistic_test2, p_value_test2 = ks_2samp(air_df_train['NO2(GT)'], air_df_test2['NO2(GT)'])
print("No*ken for test set 2:", air_df_test2['NO2(GT)'].mean())
print("KS Statistic: {ks_statistic_test2,")
print("FS Statistic: {ks_statistic_test2}")
print("Fv-value: {p_value_test2}")

if p_value_test1 < 0.05:
    print("No*keiget the null hypothesis for test1.csv")
else:
    print("No*keiget the null hypothesis for test1.csv")

if p_value_test2 < 0.05:
    print("Reject the null hypothesis for test2.csv")

### Mean for test set 1: 94.53262518968134

***S Test for test1.csv"

Mean for test set 2: 134.7308456852792

***S Test for test2.csv:

KS Statistic: 0.017062220028073977
P-value: 0.9971370322852736

Mean for test set 2: 134.70308456852792

KS Test for test2.csv:

KS Statistic: 0.308853642438679
P-value: 2.53172387531317e-74

#### Fail to reject the null hypothesis for test1.csv
Reject the null hypothesis for test2.csv
```

Therefore, there is a distributional difference in the values of NO2(GT) between the test sets.

- 3. Determine which of the two test datasets (test1.csv or test2.csv) exhibits a covariate shift relative to the training dataset (train.csv). Use the results of the Kolmogorov–Smirnov test to support your answer. The 2nd test set exhibits a covariate shift relative to the training set, since:
- The p-value for test set 1 and train set is 0.99714
- The p-value for test set 2 and train set is around 0

This rejects the Null Hypothesis for Test2 and shows strong covariate shift in test2 dataset with respect to the train set.