How to Hack your Dating App

Group members: Alison Picerno, Simone Ritchenson, Arielle Greenberg, Vivian Weigel

Data Source: https://www.kaggle.com/datasets/jmmvutu/dating-app-lovoo-user-profiles

For our data analysis, we decided to look through the columns to understand what they were referring to and then if necessary, change their types. We also then looked at if columns had null values, and found that while many of the columns didn't, additional cleaning steps were needed. For some columns, we decided the best idea was to drop them because they had close to 100% of null values. With the other columns, we decided to replace the null values.

By exploring the data, we also discovered this dataset is focused on female users and their intended partners. This is important to know when making conclusions. We also found that interactions with profiles such as profile visits or 'kisses' are pretty rare. 0 was the most common value for profile interactions, meaning that establishing what makes a good profile is that much more of an important task. Interactions have very significant meanings when they do take place.

Below, we show our entire EDA.

How to Hack your Dating App

Group members: Alison Picerno, Simone Ritchenson, Arielle Greenberg, Vivian Weigel

In [1]:

```
import pandas as pd

# read in data and look at head to get a sense of data
df = pd.read_csv('lovoo_users.csv')
df.head()
```

Out[1]:

	gender	genderLooking	age	name	counts_details	counts_pictures	counts_profileVisits	coun
0	F	M	25	daeni	1.00	4	8279	
1	F	M	22	italiana 92	0.85	5	663	
2	F	M	21	Lauraaa	0.00	4	1369	
3	F	none	20	Qqkwmdowlo	0.12	3	22187	
4	F	M	21	schaessie {3	0.15	12	35262	

5 rows × 42 columns

In [15]:

```
# Look for patterns in data
df['gender'].unique()
```

Out[15]:

```
array(['F'], dtype=object)
```

In [16]:

```
df['genderLooking'].unique()
```

Out[16]:

```
array(['M', 'none', 'both', 'F'], dtype=object)
```

In [17]:

```
df['counts_profileVisits'].value_counts()
```

Out[17]:

```
0         40
1         19
3         10
4          9
18         9
          ..
```

```
5553      1
2637      1
5560      1
10293     1
6890      1
Name: counts_profileVisits, Length: 2676, dtype: int64
```

In [18]:

```python
df['counts_kisses'].value_counts()
```

Out[18]:

```
0       212
1       117
2       102
4        84
3        83
       ... 
1346      1
1425      1
889       1
1025      1
563       1
Name: counts_kisses, Length: 666, dtype: int64
```

In [2]:

```python
# Look at types of data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3992 entries, 0 to 3991
Data columns (total 42 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   gender                  3992 non-null   object 
 1   genderLooking           3992 non-null   object 
 2   age                     3992 non-null   int64  
 3   name                    3992 non-null   object 
 4   counts_details          3992 non-null   float64
 5   counts_pictures         3992 non-null   int64  
 6   counts_profileVisits    3992 non-null   int64  
 7   counts_kisses           3992 non-null   int64  
 8   counts_fans             3992 non-null   int64  
 9   counts_g                3992 non-null   int64  
 10  flirtInterests_chat     3992 non-null   bool   
 11  flirtInterests_friends  3992 non-null   bool   
 12  flirtInterests_date     3992 non-null   bool   
 13  country                 3992 non-null   object 
 14  city                    3706 non-null   object 
 15  location                3979 non-null   object 
 16  distance                3946 non-null   float64
 17  isFlirtstar             3992 non-null   int64  
 18  isHighlighted           3992 non-null   int64  
 19  isInfluencer            3992 non-null   int64  
 20  isMobile                3992 non-null   int64  
 21  isNew                   3992 non-null   int64  
 22  isOnline                3992 non-null   int64  
 23  isVip                   3992 non-null   int64  
 24  lang_count              3992 non-null   int64  
 25  lang_fr                 3992 non-null   bool   
 26  lang_en                 3992 non-null   bool   
 27  lang_de                 3992 non-null   bool   
 28  lang_it                 3992 non-null   bool   
 29  lang_es                3992 non-null   bool   
 30  lang_pt                3992 non-null   bool   
 31  verified               3992 non-null   int64  
 32  shareProfileEnabled    3992 non-null   int64  
 33  lastOnlineDate         3991 non-null   object 
 34  lastOnlineTime         3991 non-null   float64
 35  birthd                 3992 non-null   int64  
 36  crypt                  46 non-null     float64
 37  freetext               113 non-null    object 
 38  whazzup                2399 non-null   object 
 39  userId                 3992 non-null   object 
 40  pictureId              3901 non-null   object 
 41  isSystemProfile        2 non-null      float64
dtypes: bool(9), float64(5), int64(17), object(11)
memory usage: 1.0+ MB
```

In [3]:

```
# change invalid types
df['gender']  = df['gender'].astype(str)
df['genderLooking']  = df['genderLooking'].astype(str)
df['country']  = df['country'].astype(str)
df['city']  = df['city'].astype(str)
df['location']  = df['location'].astype(str)
df['name']  = df['name'].astype(str)
df['freetext']  = df['freetext'].astype(str)
df['whazzup']  = df['whazzup'].astype(str)
df['userId']  = df['userId'].astype(str)
df['pictureId']  = df['pictureId'].astype(str)
```

In [4]:

```
# find the percentage of null values each colum
df.isnull().mean()*100
```

Out[4]:

```
gender                     0.000000
genderLooking              0.000000
age                        0.000000
name                       0.000000
counts_details             0.000000
counts_pictures            0.000000
counts_profileVisits       0.000000
counts_kisses              0.000000
counts_fans                0.000000
counts_g                   0.000000
flirtInterests_chat        0.000000
flirtInterests_friends     0.000000
flirtInterests_date        0.000000
country                    0.000000
city                       0.000000
location                   0.000000
distance                   1.152305
isFlirtstar                0.000000
isHighlighted              0.000000
isInfluencer               0.000000
isMobile                   0.000000
isNew                      0.000000
isOnline                   0.000000
isVip                      0.000000
lang_count                 0.000000
lang_fr                    0.000000
lang_en                    0.000000
```

```
lang_de                    0.000000
lang_it                    0.000000
lang_es                    0.000000
lang_pt                    0.000000
verified                   0.000000
shareProfileEnabled        0.000000
lastOnlineDate             0.025050
lastOnlineTime             0.025050
birthd                     0.000000
crypt                     98.847695
freetext                   0.000000
whazzup                    0.000000
userId                     0.000000
pictureId                  0.000000
isSystemProfile           99.949900
dtype: float64
```

In [5]:

```
# drop columns with very high percentages of missing data
df = df.drop(['isSystemProfile'], axis=1)
df = df.drop(['crypt'], axis=1)
```

In [6]:

```
# find mean distance
df['distance'].mean()
```

Out[6]:

```
207.2300050684237
```

In [7]:

```
# replace null values with mean
df.loc[df["distance"].isnull(), "distance"] = df['distance'].mean()
```

In [8]:

```
# find most common dates
df['lastOnlineDate'].value_counts()
```

Out[8]:

```
2015-04-07T00:08:59Z    7
2015-04-06T14:23:52Z    7
2015-04-19T08:37:52Z    6
2015-04-05T07:13:49Z    6
2015-04-06T16:02:55Z    5
                       ..
2015-04-06T16:03:19Z    1
2015-04-26T09:37:25Z    1
2015-04-26T11:41:36Z    1
2015-04-19T23:59:22Z    1
2015-04-19T11:00:59Z    1
Name: lastOnlineDate, Length: 3470, dtype: int64
```

In [9]:

```
# replace null values with common date
df.loc[df["lastOnlineDate"].isnull(), "lastOnlineDate"] = '2015-04-07T00:08:59Z'
```

In [10]:

```
# find most common time
df['lastOnlineTime'].value_counts()
```

Out[10]:

```
1.428365e+09    7
1.428330e+09    7
```

```
1.429433e+09    6
1.428218e+09    6
1.428336e+09    5
               ..
1.428336e+09    1
1.430041e+09    1
1.430048e+09    1
1.429488e+09    1
1.429441e+09    1
Name: lastOnlineTime, Length: 3470, dtype: int64
```

In [11]:

```
# replace null values with common time
df.loc[df["lastOnlineTime"].isnull(), "lastOnlineTime"] = 1.428365e+09
```

In [13]:

```
# we can see that all null values are gone
df.isnull().mean()*100
```

```
Out[13]: gender                    0.0
         genderLooking             0.0
         age                       0.0
         name                      0.0
         counts_details            0.0
         counts_pictures           0.0
         counts_profileVisits      0.0
         counts_kisses             0.0
         counts_fans               0.0
         counts_g                  0.0
         flirtInterests_chat       0.0
         flirtInterests_friends    0.0
         flirtInterests_date       0.0
         country                   0.0
         city                      0.0
         location                  0.0
         distance                  0.0
         isFlirtstar               0.0
         isHighlighted             0.0
         isInfluencer              0.0
         isMobile                  0.0
         isNew                     0.0
         isOnline                  0.0
         isVip                     0.0
         lang_count                0.0
         lang_fr                   0.0
         lang_en                   0.0
         lang_de                   0.0
         lang_it                   0.0
         lang_es                   0.0
         lang_pt                   0.0
         verified                  0.0
         shareProfileEnabled       0.0
         lastOnlineDate            0.0
         lastOnlineTime            0.0
         birthd                    0.0
         freetext                  0.0
         whazzup                   0.0
         userId                    0.0
         pictureId                 0.0
         dtype: float64
```