In [11]:

```
import os
import pandas as pd
pd.options.display.float_format = '{:20,.2f}'.format
pd.set_option('display.max_rows', 5000)
pd.set_option('display.max_columns', 5000)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', -1)
```

In [12]:

```
yellow_pages_yelp_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Dat
                          r'/Unprocessed Data/customer_samples/yellow_pages_yelp_
```

## Yellow Pages data

In [13]:

```
yp_fields = ['id', 'name', 'address', 'city', 'zipcode', 'phone']
```

In [14]:

```
yellow_pages_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'yellow_pages.csv'),
    sep = ',',
    quotechar = '"',
    usecols  = yp_fields,
    dtype={'zipcode': 'str'}
)[yp_fields]
```

In [15]:

```
yellow_pages_data.rename(
    columns = {
        'zipcode': 'postalcode',
    },
    inplace=True
)
```

In [16]:

```
yellow_pages_data.head()
```

Out[16]:

| | id | name | address | city | postalcode | phone |
|---|---|---|---|---|---|---|
| **0** | 1 | Tao Tao Restaurant | 175 S Murphy Ave | Sunnyvale | 94086 | (408) 736-3731 |
| **1** | 2 | Dinner at Buca di Beppo | NaN | NaN | NaN | NaN |
| **2** | 3 | Sneha South & North Indian Restaurant | 1214 Apollo Way # 404-B | Sunnyvale | 94085 | (408) 481-0700 |
| **3** | 4 | The Armenian Gourmet | 929 E Duane Ave | Sunnyvale | 94085 | (408) 732-3910 |
| **4** | 5 | Round Table Pizza | 415 N Mary Ave | Sunnyvale | 94085 | (408) 733-1365 |

## Yelp data

In [18]:

```
yelp_fields = ['id', 'name', 'address', 'city', 'zipcode', 'phone']
```

In [19]:

```
yelp_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'yelp.csv'),
    sep = ',',
    quotechar = '"',
    usecols  = yelp_fields,
    dtype={'zipcode': 'str'}
)[yelp_fields]
```

In [20]:

```
yelp_data.rename(
    columns = {
        'zipcode': 'postalcode',
    },
    inplace=True
)
```

In [21]:

```
yelp_data.head()
```

Out[21]:

| | id | name | address | city | postalcode | phone |
|---|---|---|---|---|---|---|
| **0** | 1 | The Little Cakes | Sunnyvale, CA 94085 | Phone number | NaN | NaN |
| **1** | 2 | Sunnyvale Cafe | 223 E Maude Ave | Sunnyvale | 94085 | (408) 530-8191 |
| **2** | 3 | Obed Mediterranean Cuisine | 911 E Duane Ave | Sunnyvale | 94085 | (408) 685-2269 |
| **3** | 4 | DishDash | 190 S Murphy Ave | Sunnyvale | 94086 | (408) 774-1889 |
| **4** | 5 | Toofu | S Murphy Ave | Sunnyvale | 94085 | NaN |

## Labeled data comparisons

In [30]:

```
lb_fields = ['ltable.id', 'rtable.id', 'gold']
```

In [31]:

```
labeled_data = pd.read_csv(
    os.path.join(yellow_pages_yelp_path, 'labeled_data.csv'),
    sep = ',',
    quotechar = '"',
    comment = '#',
    usecols  = lb_fields,
    #dtype = {'gold': 'str'}
)[lb_fields]
```

In [32]:

```
labeled_data.rename(
    columns =
    {
        'ltable.id' : 'yellow_pages_id',
        'rtable.id' : 'yelp_id',
        'gold'      : 'duplicate'
    },
    inplace = True
)
```

## Cands set

```
In [25]: cands_data = pd.read_csv(
             os.path.join(yellow_pages_yelp_path, 'candset.csv'),
             sep = ',',
             quotechar = '"',
             comment = '#',
             #usecols  = lb_fields,
             #dtype = {'gold': 'str'}
         )#[lb_fields]
```

```
In [26]: cands_data.head()
```

Out[26]:

|  | _id | ltable.id | rtable.id | ltable.address | ltable.city | ltable.name | ltable.phone | ltable.state | ltable |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 5 | 1 | 54 | 175 S Murphy Ave | Sunnyvale | Tao Tao Restaurant | (408) 736-3731 | CA | 9 |
| **1** | 8 | 1 | 72 | 175 S Murphy Ave | Sunnyvale | Tao Tao Restaurant | (408) 736-3731 | CA | 9 |
| **2** | 12 | 1 | 134 | 175 S Murphy Ave | Sunnyvale | Tao Tao Restaurant | (408) 736-3731 | CA | 9 |
| **3** | 16 | 3 | 141 | 1214 Apollo Way | NaN | NaN | NaN | NaN |  |
| **4** | 22 | 7 | 46 | 528 Lawrence Expy | Sunnyvale | Tasty Indian Pizza | (408) 738-8761 | CA | 9 |

```
In [82]: cands_data = cands_data(~(cands_dta[]))
```

```
Out[82]: Index(['_id', 'ltable.id', 'rtable.id', 'ltable.address', 'ltable.cit
         y', 'ltable.name', 'ltable.phone', 'ltable.state', 'ltable.zipcode',
         'rtable.address', 'rtable.city', 'rtable.name', 'rtable.phone', 'rtabl
         e.state', 'rtable.zipcode'], dtype='object')
```

## Remove invalid nan entries from all data: yelp, yellow pages, cand sets and labeled data

In [33]:
```
labeled_data[labeled_data['duplicate'].isnull()]
```

Out[33]:

| | yellow_pages_id | yelp_id | duplicate |
|---|---|---|---|
| **6** | 63 | 141 | nan |
| **219** | 1834 | 1995 | nan |
| **220** | 1834 | 2528 | nan |
| **225** | 1868 | 2454 | nan |
| **227** | 1894 | 1913 | nan |
| **324** | 3670 | 3794 | nan |
| **331** | 4107 | 4145 | nan |
| **332** | 4131 | 4162 | nan |
| **357** | 4555 | 4443 | nan |

In [34]:
```
yellow_pages_to_remove = labeled_data[labeled_data['duplicate'].isnull(
yelp_to_remove = labeled_data[labeled_data['duplicate'].isnull()]['yelp_
```

In [36]:
```
yellow_pages_data = yellow_pages_data[~(yellow_pages_data['id'].isin(ye
```

In [37]:
```
yelp_data = yelp_data[~(yelp_data['id'].isin(yelp_to_remove))]
```

In [38]:
```
labeled_data = labeled_data[ ~(labeled_data['duplicate'].isnull())]
```

In [39]:
```
labeled_data['duplicate'] = labeled_data['duplicate'].astype('int64')
```

In [52]:
```
labeled_data[labeled_data['duplicate'] == 1].head()
```

Out[52]:

| | yellow_pages_id | yelp_id | duplicate |
|---|---|---|---|
| **0** | 20 | 71 | 1 |
| **5** | 48 | 1473 | 1 |
| **7** | 68 | 76 | 1 |
| **8** | 80 | 71 | 1 |
| **9** | 139 | 3 | 1 |

In [ ]:
```

```

In [64]:

```
yellow_pages_data[yellow_pages_data['id'] == 1831]
```

Out[64]:

| | id | name | address | city | postalcode | phone | source |
|---|---|---|---|---|---|---|---|
| **1830** | 1831 | Seasons 52 | 10250 Santa Monica Blvd | Los Angeles | 90067 | (310) 277-5252 | yellow_pages |

In [65]:

```
yelp_data[yelp_data['id'] == 1831]
```

Out[65]:

| | id | name | address | city | postalcode | phone | source |
|---|---|---|---|---|---|---|---|
| **1611** | 1831 | The Oasis | 241 El Camino Real | Menlo Park | 94025 | (650) 326-8896 | yelp |

In [44]:

```
customer_all_path = (r'/home/ubuntu/jupyter/ServerX/1_Standard Data Int
                     r'/Processed Data/customer_samples/')
```

In [45]:

```
yellow_pages_data['source'] = 'yellow_pages'
yelp_data['source'] = 'yelp'
```

In [46]:

```
yellow_yelp_all = pd.concat([yellow_pages_data, yelp_data])
#yellow_yelp_all.rename(columns={'record_id': 'Id'}, inplace=True)
```

In [47]:

```
yellow_yelp_all.to_csv(customer_all_path + 'yellow_yelp_all.csv', sep='
```

In [48]:

```
labeled_data.to_csv(customer_all_path + 'yellow_yelp_label.csv', sep=',
```