YELP DATASET CODE

In [169]:
```
#Import necessary libraries and packages
import pandas as pd
import numpy as np
from pandas.io.json import json_normalize
import matplotlib.pyplot as plt
import seaborn as sns
```

In [170]:
```
#Import primary CSV data file
business_original = pd.read_csv('business.csv')
```

In [171]:
```
#First view of data
business_original.head(15)
```

Out[171]:

	Unnamed: 0	business_id	name	address	city	state	postal_co
0	0	f9NumwFMBDn751xgFiRbNA	The Range At Lake Norman	10913 Bailey Rd	Cornelius	NC	280
1	1	Yzvjg0SayhoZgCljUJRF9Q	Carlos Santo, NMD	8880 E Via Linda, Ste 107	Scottsdale	AZ	852
2	2	XNoUzKckATkOD1hP6vghZg	Felinus	3554 Rue Notre-Dame O	Montreal	QC	H4C 1
3	3	6OAZjbxqM5ol29BuHsil3w	Nevada House of Hose	1015 Sharp Cir	North Las Vegas	NV	890
4	4	51M2Kk903DFYI6gnB5I6SQ	USE MY GUY SERVICES LLC	4827 E Downing Cir	Mesa	AZ	852
5	5	cKyLV5oWZJ2NudWgqs8VZw	Oasis Auto Center - Gilbert	1720 W Elliot Rd, Ste 105	Gilbert	AZ	852
6	6	oiAlXZPIFm2nBCt0DHLu_Q	Green World Cleaners	6870 S Rainbow Blvd, Ste 117	Las Vegas	NV	891
7	7	ScYkbYNkDgCneBrD9vqhCQ	Junction Tire & Auto Service	6910 E Southern Ave	Mesa	AZ	852
8	8	pQeaRpvuhoEqudo3uymHIQ	The Empanadas House	404 E Green St	Champaign	IL	618
9	9	EosRKXIGeSWFYWwpkbhNnA	Xtreme Couture	700 Kipling Avenue Etobicoke	Toronto	ON	M8Z 5

	Unnamed: 0	business_id	name	address	city	state	postal_co
10	10	MbZMmwo-eL0Jnm_Yb9KJrA	Chinook Landscaping and Design	NaN	Calgary	AB	T2J 2
11	11	7Dv4_HAxsxvadEsT5fxQBg	Dependable Brakes & Exhaust	1110 Saw Mill Run Blvd	Pittsburgh	PA	152
12	12	M_guz7Dj7hX0evS672wIwA	Chocolate Shoppe Ice Cream	2831 Parmenter St	Middleton	WI	535
13	13	JjJs3o60uQCfctDjs45cmA	Convertabath	116 N Roosevelt Ave, Bldg B, Ste 124	Chandler	AZ	852
14	14	kOICO53wbOiOJcKuCgOQ3A	Tan Las Vegas	5465 Simmons St	North Las Vegas	NV	890

In [172]:
```
#Determine the actual span of our data
business_original.shape
```

Out[172]: (209393, 15)

In [173]:
```
#Unique cities in dataset
business_original['city'].nunique()
```

Out[173]: 1250

In [174]:
```
#Assess the data types for ease of analysis
business_original.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209393 entries, 0 to 209392
Data columns (total 15 columns):
Unnamed: 0      209393 non-null int64
business_id     209393 non-null object
name            209392 non-null object
address         200714 non-null object
city            209391 non-null object
state           209393 non-null object
postal_code     208884 non-null object
latitude        209393 non-null float64
longitude       209393 non-null float64
stars           209393 non-null float64
review_count    209393 non-null int64
is_open         209393 non-null int64
attributes      180348 non-null object
categories      208869 non-null object
hours           164550 non-null object
dtypes: float64(3), int64(3), object(9)
memory usage: 24.0+ MB
```

In [175]:
```
#Check for duplicate records in the dataset
business_original.duplicated().sum()
```

Out[175]:
```
0
```

In [176]:
```
#Check for Null values
business_original.isnull().sum()
```

Out[176]:
```
Unnamed: 0          0
business_id         0
name                1
address          8679
city                2
state               0
postal_code       509
latitude            0
longitude           0
stars               0
review_count        0
is_open             0
attributes      29045
categories        524
hours           44843
dtype: int64
```

In [177]:
```
#check what percentage of hours is null
(business_original.hours.isnull().sum()/len(business_original))*100
```

Out[177]:
```
21.415711126924013
```

In [178]:
```
#hours is not really useful and has a bad format so we drop it
business_original.drop('hours', axis=1, inplace=True)
```

In [179]:
```
#updated span/size of dataset
business_original.shape
```

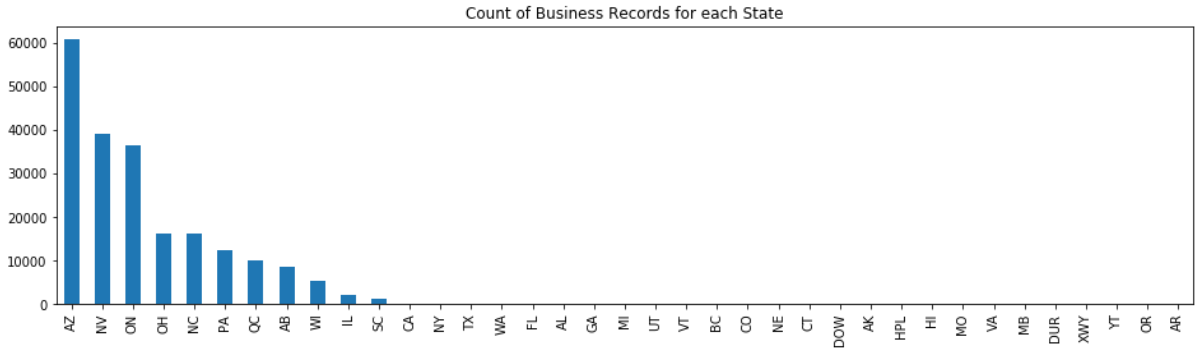
Out[179]: (209393, 14)

In [180]:
```
#Analyze the state column to determine which states will be of use
business_original['state'].value_counts()
```

Out[180]:
```
AZ     60803
NV     39084
ON     36627
OH     16392
NC     16218
PA     12376
QC     10233
AB      8682
WI      5525
IL      2034
SC      1328
CA        23
NY        22
TX         6
WA         5
FL         3
AL         3
GA         3
MI         2
UT         2
VT         2
BC         2
CO         2
NE         2
CT         2
DOW        1
AK         1
HPL        1
HI         1
MO         1
VA         1
MB         1
DUR        1
XWY        1
YT         1
OR         1
AR         1
Name: state, dtype: int64
```

In [181]:
```
#Visulaize the distribution above
ax = business_original['state'].value_counts()
ax.plot.bar(figsize = (16,4), title="Count of Business Records for each
 State")
```

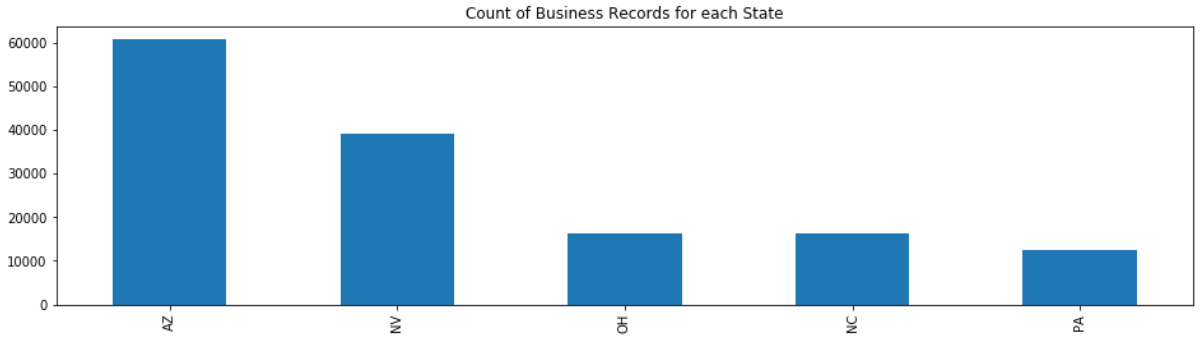
Out[181]: <matplotlib.axes._subplots.AxesSubplot at 0x1a4be3e250>



In [182]:
```
#Graph new order of states
filt = ['AZ','NV','NC','OH','PA']
state_filt= business_original['state'].isin(filt)
graph=business_original[state_filt]
```

In [183]:
```
ax_1 = graph['state'].value_counts()
ax_1.plot.bar(figsize = (16,4), title="Count of Business Records for eac
h State")
```

Out[183]: <matplotlib.axes._subplots.AxesSubplot at 0x1a26eb4d90>



In [184]:
```
#Hence, filter needs only relevant states
filt1 = ['AZ','NV','NC','OH','PA']
state_filt1= business_original['state'].isin(filt1)
business = business_original[state_filt1]
business.head()
```

Out[184]:

	Unnamed: 0	business_id	name	address	city	state	postal_code	
0	0	f9NumwFMBDn751xgFiRbNA	The Range At Lake Norman	10913 Bailey Rd	Cornelius	NC	28031	35
1	1	Yzvjg0SayhoZgCljUJRF9Q	Carlos Santo, NMD	8880 E Via Linda, Ste 107	Scottsdale	AZ	85258	33
3	3	6OAZjbxqM5ol29BuHsil3w	Nevada House of Hose	1015 Sharp Cir	North Las Vegas	NV	89030	36
4	4	51M2Kk903DFYI6gnB5I6SQ	USE MY GUY SERVICES LLC	4827 E Downing Cir	Mesa	AZ	85205	33
5	5	cKyLV5oWZJ2NudWgqs8VZw	Oasis Auto Center - Gilbert	1720 W Elliot Rd, Ste 105	Gilbert	AZ	85233	33

In [185]:
```
business['state'].value_counts()
```

Out[185]:
```
AZ    60803
NV    39084
OH    16392
NC    16218
PA    12376
Name: state, dtype: int64
```

In [186]:
```
#How many records do we have left to work with?
business.shape
```

Out[186]:
```
(144873, 14)
```

In [187]:
```
#Begin exploration of categories
#Check for null values
business['categories'].isnull().sum()
```

Out[187]:
```
376
```

In [188]:
```
#Replace null values
business["categories"].fillna("",inplace=True)
```

In [189]:
```
#Reset index and drop unneccessry columns
business=business.reset_index().drop(columns=['Unnamed: 0','index'])
```

In [190]:
```
#Filter out only records that fall into important categories
targets = ['Restaurants', 'Fast Food','Shopping','Beauty','Spa','Nightli
fe','Auto', 'Arts','Entertainment','Active Life']
business=business[business.categories.str.contains('|'.join(targets))]
```

In [191]:
```
#What do we have left?
business.shape
```

Out[191]: (95051, 13)

In [192]:
```
#CREATE FUNCTION TO SINGLE OUT AREA OF PRIMARY INTEREST FOR ANALYSIS
```

In [193]:
```
def Restaurant(x):
    if ('restaurants' in x.lower()) or ('fast food' in x.lower()) or (
'restaurant' in x.lower()):
        return 1
    else:
        return 0
```

In [194]:
```
business["Restaurant"] = business["categories"].apply(Restaurant)
business[["categories","Restaurant"]].head(10)
```

Out[194]:

	categories	Restaurant
0	Active Life, Gun/Rifle Ranges, Guns & Ammo, Sh...	0
1	Health & Medical, Fitness & Instruction, Yoga,...	0
2	Hardware Stores, Home Services, Building Suppl...	0
4	Auto Repair, Automotive, Oil Change Stations, ...	0
6	Auto Repair, Oil Change Stations, Automotive, ...	0
7	Automotive, Auto Repair	0
9	Beauty & Spas, Tanning	0
11	Shopping, Shoe Stores, Fashion	0
12	Event Planning & Services, Wedding Planning, F...	0
13	Weight Loss Centers, Fitness & Instruction, Bo...	0

In [195]:
```
business["Restaurant"].sum()
```

Out[195]: 35305

Extracting Attributes

In [196]:
```
#Expand attributes columns by splitting and create dummy variables
business["attributes"]=business["attributes"].str.replace("{","")
business["attributes"]=business["attributes"].str.replace("}","")
business["attributes"]=business["attributes"].str.replace("'","")
business["attributes"]=business["attributes"].str.replace('"',"")
business["attributes"]=business["attributes"].astype(str)
pd.set_option('display.max_columns', 50)
business.head()
```

Out[196]:

	business_id	name	address	city	state	postal_code	latitude	lo
0	f9NumwFMBDn751xgFiRbNA	The Range At Lake Norman	10913 Bailey Rd	Cornelius	NC	28031	35.462724	-80
1	Yzvjg0SayhoZgCljUJRF9Q	Carlos Santo, NMD	8880 E Via Linda, Ste 107	Scottsdale	AZ	85258	33.569404	-111
2	6OAZjbxqM5ol29BuHsil3w	Nevada House of Hose	1015 Sharp Cir	North Las Vegas	NV	89030	36.219728	-115
4	cKyLV5oWZJ2NudWgqs8VZw	Oasis Auto Center - Gilbert	1720 W Elliot Rd, Ste 105	Gilbert	AZ	85233	33.350399	-111
6	ScYkbYNkDgCneBrD9vqhCQ	Junction Tire & Auto Service	6910 E Southern Ave	Mesa	AZ	85209	33.393885	-111

In [197]:
```
#Create Parking variable
def Parking(x):
    if ('valet: True' in x) or ('garage: True' in x) or ('lot: True' in
x):
        return 1
    else:
        return 0
```

In [198]:
```
business['Parking']=business['attributes'].apply(Parking)
```

In [199]:
```
#Create Kid_friendly variable
def Kid_friendly(x):
    if 'GoodForKids: True' in x:
        return 1
    else:
        return 0
```

In [200]:
```
business['Kid_friendly']=business['attributes'].apply(Kid_friendly)
```

In [201]:
```
#Create Reservations variable
def Reservations(x):
    if 'RestaurantsReservations: True' in x:
        return 1
    else:
        return 0
```

In [202]:
```
business['Reservations'] = business['attributes'].apply(Reservations)
```

In [203]:
```
#Create Price range variable
def Price_Range(x):
    if 'RestaurantsPriceRange2: 1' in x:
        return 1
    elif 'RestaurantsPriceRange2: 2' in x:
        return 2
    elif 'RestaurantsPriceRange2: 3' in x:
        return 3
    else:
        return 4
```

In [204]:
```
business['Price_Range'] = business['attributes'].apply(Price_Range)
```

In [205]:
```
#Create creditcard variable
def Credit_card(x):
    if "BusinessAcceptsCreditCards: True" in x:
        return 1
    else:
        return 0
```

In [206]:
```
business['Credit_card'] = business['attributes'].apply(Credit_card)
```

In [207]:
```
#Create wheelchair access variable
def wheelchair_access(x):
    if 'WheelchairAccessible: True' in x:
        return 1
    else:
        return 0
```

In [208]:
```
business['wheelchair_access'] = business['attributes'].apply(wheelchair_
access)
```

In [209]:
```
#Create breakfast variable
def good_for_breakfast (x):
    if 'breakfast: True' in x:
        return 1
    else:
        return 0
```

In [210]:
```
business['good_for_breakfast'] = business['attributes'].apply(good_for_b
reakfast)
```

In [211]:
```
#Create lunch variable
def good_for_lunch (x):
    if 'lunch: True' in x:
        return 1
    else:
        return 0
```

In [212]:
```
business['good_for_lunch'] = business['attributes'].apply(good_for_lunch
)
```

In [213]:
```
#Create dinner variable
def good_for_dinner (x):
    if 'dinner: True' in x:
        return 1
    else:
        return 0
```

In [214]:
```
business['good_for_dinner'] = business['attributes'].apply(good_for_dinn
er)
```

In [215]:
```
#Create alcohol variable
def alcohol (x):
    if ('Alcohol: ufull_bar' in x) or ('Alcohol: ubeer_and_wine' in x):
        return 1
    else:
        return 0
```

In [216]:
```
business['alcohol'] = business['attributes'].apply(alcohol)
```

In [217]:
```
#Create happyhour variable
def happyhour (x):
    if 'HappyHour: True' in x :
        return 1
    else:
        return 0
```

In [218]:
```
business['happyhour'] = business['attributes'].apply(happyhour)
```

In [219]:
```
#Create wifi variable
def wifi (x):
    if ('WiFi: ufree' in x) or ('WiFi: free' in x) or ('WiFi: yes' in x)
or ('WiFi: uyes' in x) or ('WiFi: True' in x) or ('WiFi: uTrue' in x):
        return 1
    else:
        return 0
```

In [220]:
```
business['wifi'] = business['attributes'].apply(wifi)
```

In [221]:
```
#Create table service variable
def table_service (x):
    if 'RestaurantsTableService: True' in x :
        return 1
    else:
        return 0
```

In [222]:
```
business['table_service'] = business['attributes'].apply(table_service)
```

In [223]:
```
#Create Entertainment
def Entertainment (x):
    if ('HasTV: True' in x) or ('dj: True' in x) or ('background_music:
 True' in x) or ('jukebox: True' in x) or ('live: True' in x) or ('vide
o: True' in x) or ('karaoke: True' in x):
        return 1
    else:
        return 0
```

In [224]:
```
business['Entertainment'] = business['attributes'].apply(Entertainment)
```

In [225]:
```
#Create takeout variable
def takeout (x):
    if 'RestaurantsTakeOut: True' in x :
        return 1
    else:
        return 0
```

In [226]:
```
business['Takeout'] = business['attributes'].apply(takeout)
```

In [227]:
```
#Create Noise_Level variable

def Noise_Level(x):
    if ('NoiseLevel: uquiet' in x) or ('NoiseLevel: quiet' in x):
        return 1
    elif ('NoiseLevel: uaverage' in x) or ('NoiseLevel: average' in x):
        return 2
    elif ('NoiseLevel: uloud' in x) or ('NoiseLevel: loud' in x):
        return 3
    else:
        return 4
```

In [228]:
```
business['Noise_Level'] = business['attributes'].apply(Noise_Level)
```

In [229]:
```
#Create Reservations variable

def Reservations (x):
    if 'RestaurantsReservations: True' in x :
        return 1
    else:
        return 0
```

In [230]:
```
business['Reservations'] = business['attributes'].apply(Reservations)
```

In [231]:
```
#Create Delivery variable

def Delivery (x):
    if 'RestaurantsDelivery: True' in x :
        return 1
    else:
        return 0
```

In [232]:
```
business['Delivery'] = business['attributes'].apply(Delivery)
```

Extracting Categories

In [233]:
```
#Create FastFood variable
def FastFood (x):
    if 'Fast Food' in x :
        return 1
    else:
        return 0
```

In [234]:
```
business['FastFood'] = business['categories'].apply(FastFood)
```

In [235]:
```
#Create Ethnicity variable
def ethnicity (x):
    if ('american' in x.lower()) or ('burgers' in x.lower()):
        return 'American'
    elif 'chinese' in x.lower():
        return 'Chinese'
    elif ('mexican' in x.lower()) or ("tex-mex"in x.lower()):
        return 'Mexican'
    elif 'italian' in x.lower():
        return 'Italian'
    elif ('japanese' in x.lower()) or ('sushi' in x.lower()):
        return 'Japanese'
   # elif 'thai' in x.lower():
      #  return 'Thai'
   # elif 'indian' in x.lower():
      #  return 'Indian'
   # elif 'korean' in x.lower():
      #  return 'Korean'
    else:
        return 'other'
```

In [236]:
```
business['Ethnicity'] = business['categories'].apply(ethnicity)
```

In [237]:
```
#Remove foreign symbols from name to allow for counting chains
business["name"]=business["name"].str.replace(' ','')
business["name"]=business["name"].str.replace("'","")
business["name"]=business["name"].str.replace(',',"")
business["name"]=business["name"].str.replace('.',"")

business["name"]=business["name"].astype(str)
business["name"]=business["name"].str.lower()
```

In [238]:
```
#Select only restaurants for data analysis before chain is counted
Rest_filt= business["Restaurant"]==1
Restaurant=business[Rest_filt]
Restaurant.head(10)
```

Out[238]:

	business_id	name	address	city	state	posta
14	CsLQLiRoafpJPJSkNX2h5Q	middleeastdeli	4508 E Independence Blvd	Charlotte	NC	
21	vjTVxnsQEZ34XjYNS-XUpA	wetzelspretzels	4550 East Cactus Rd, #KSFC-4	Phoenix	AZ	
24	fnZrZlqW1Z8iWgTVDfv_MA	carlsjr	9595 W Tropicana Ave	Las Vegas	NV	
28	98hyK2QEUel8v2y0AghfZA	pholeesvietnameserestaurant	1541 E 38th St, Ste 101	Cleveland	OH	
29	fhNf_sg-XzZ3e7HEVGuOZg	meatchixandwieners	6530 S Decatur Blvd	Las Vegas	NV	
30	Ga2Bt7xfqoggTypWD5VpoQ	amandosbros	2602 W Southern Ave	Tempe	AZ	
31	xFc50drSPxXkcLvX5ygqrg	boomerssweethomechicago	5932 W Bell Rd, Ste D-109	Glendale	AZ	
33	tLpkSwdtqqoXwU0JAGnApw	wendys	4602 Northfield Road	Cleveland	OH	
34	Sd75ucXKoZUM2BEfBHFUOg	chinagourmet	3460 E Southern Ave, Ste 109	Mesa	AZ	
37	lK-wuiq8b1TuU7bfbQZgsg	hingetown	NaN	Cleveland	OH	

In [239]:
```
#Create chain counts column by counting occurence of names
Restaurant['Chain_Counts'] = Restaurant.groupby(['name'])['name'].transf
orm('count')
```

```
/Users/abidemiolaoye/opt/anaconda3/lib/python3.7/site-packages/ipykerne
l_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

In [240]:
```
#Declare chain if chain counts is 4 or more.
def Chain (x):
    if x >= 4 :
        return 1
    else:
        return 0
```

In [241]:
```
#Create Is_Chain column
Restaurant['Is_Chain'] = Restaurant['Chain_Counts'].apply(Chain)
```

```
/Users/abidemiolaoye/opt/anaconda3/lib/python3.7/site-packages/ipykerne
l_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

In [242]:
```
#Drop longitude and latitude since they're not needed
Restaurant.drop(columns=['longitude','latitude'], inplace=True)
```

In [243]:
```
#Confirm shape of DF
Restaurant.shape
```

Out[243]:
```
(35305, 33)
```

In [244]:
```
#Check for number of Open restaurants
Restaurant['is_open'].sum()
```

Out[244]:
```
23867
```

In [245]:
```
#Check for number of Closed restaurants
len(Restaurant['is_open'])-(Restaurant['is_open'].sum())
```

Out[245]:
```
11438
```

In [246]:
```
#Check again for null values
Restaurant.isnull().sum()
```

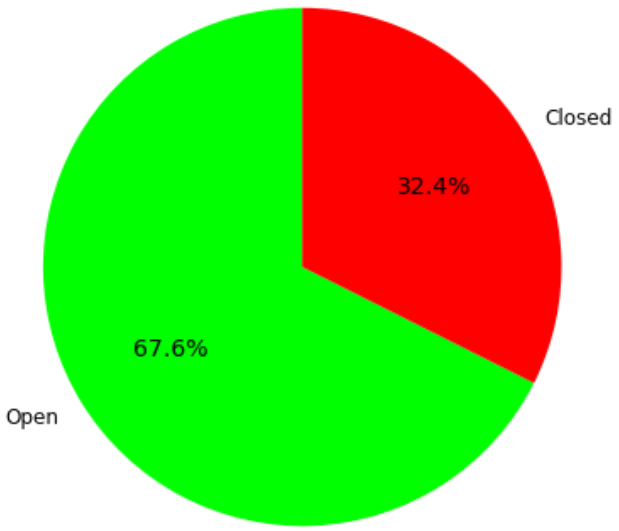
Out[246]:
```
business_id             0
name                    0
address               395
city                    0
state                   0
postal_code            39
stars                   0
review_count            0
is_open                 0
attributes              0
categories              0
Restaurant              0
Parking                 0
Kid_friendly            0
Reservations            0
Price_Range             0
Credit_card             0
wheelchair_access       0
good_for_breakfast      0
good_for_lunch          0
good_for_dinner         0
alcohol                 0
happyhour               0
wifi                    0
table_service           0
Entertainment           0
Takeout                 0
Noise_Level             0
Delivery                0
FastFood                0
Ethnicity               0
Chain_Counts            0
Is_Chain                0
dtype: int64
```

In [247]:
```
#Make pie chart to show distribution of open and closed businesses'

# Pie chart
labels = ["Open", 'Closed']
sizes = [23867, 11438]
#colors
colors = ['Lime','Red']

fig1, ax1 = plt.subplots(figsize=(10,5))
fig1.subplots_adjust(0.3,0,1,1)
patches, texts, autotexts = ax1.pie(sizes, colors = colors, labels=label
s, autopct='%1.1f%%', startangle=90)
for text in texts:
    text.set_color('black')
    text.set_size(12)
for autotext in autotexts:
    autotext.set_color('black')
    autotext.set_size(14)

# Equal aspect ratio ensures that pie is drawn as a circle
ax1.axis('equal')
plt.tight_layout()
plt.show()
```



In [248]:
```
Restaurant.state.value_counts()
```

Out[248]:
```
AZ    12130
NV     8345
OH     5914
NC     4656
PA     4260
Name: state, dtype: int64
```

In [249]:
```
Restaurant.postal_code.value_counts() #Reject
```

Out[249]:
```
89109    1022
85281     557
89119     499
85251     458
89102     433
         ...
15301       1
44034       1
28130       1
15038       1
44096       1
Name: postal_code, Length: 548, dtype: int64
```

In [250]:
```
#Check for ethnicity distribution
#Looks very skewed so it may not be used. There are 600 levels. This doe
s not seem feasible for analysis within this time frame.
Restaurant.Ethnicity.value_counts()
```

Out[250]:
```
other       14430
American    11503
Mexican      3565
Italian      2405
Chinese      2074
Japanese     1328
Name: Ethnicity, dtype: int64
```

In [251]:
```
Restaurant.head()
```

Out[251]:

	business_id	name	address	city	state	postal_c
14	CsLQLiRoafpJPJSkNX2h5Q	middleeastdeli	4508 E Independence Blvd	Charlotte	NC	28
21	vjTVxnsQEZ34XjYNS-XUpA	wetzelspretzels	4550 East Cactus Rd, #KSFC-4	Phoenix	AZ	85
24	fnZrZlqW1Z8iWgTVDfv_MA	carlsjr	9595 W Tropicana Ave	Las Vegas	NV	89
28	98hyK2QEUel8v2y0AghfZA	pholeesvietnameserestaurant	1541 E 38th St, Ste 101	Cleveland	OH	44
29	fhNf_sg-XzZ3e7HEVGuOZg	meatchixandwieners	6530 S Decatur Blvd	Las Vegas	NV	89

In [252]:
```
Restaurant.shape
```

Out[252]:
```
(35305, 33)
```

In [253]:
```

```