```
# This Python 3 environment comes with many helpful analytics libraries installed



import numpy as np # linear algebra
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from matplotlib.pylab import rcParams
rcParams['figure.figsize'] = 12, 4
#import warnings
#warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_columns', 100)
# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files
 in the input directory

from subprocess import check_output
print(check_output(["ls", "../input"]).decode("utf8"))

# Any results you write to the current directory are saved as output.
```

```
Dataset_Challenge_Dataset_Agreement.pdf
yelp_business.csv
yelp_business_attributes.csv
yelp_business_hours.csv
yelp_checkin.csv
yelp_review.csv
yelp_tip.csv
yelp_user.csv
```

In [2]:

```
business = pd.read_csv('../input/yelp_business.csv')
```

In [3]:
```
business.head(5)
```

Out[3]:

	business_id	name	neighborhood	address	city	state	postal_code
0	FYWN1wneV18bWNgQjJ2GNg	"Dental by Design"	NaN	"4855 E Warner Rd, Ste B9"	Ahwatukee	AZ	85044
1	He-G7vWjzVUysIKrfNbPUQ	"Stephen Szabo Salon"	NaN	"3101 Washington Rd"	McMurray	PA	15317
2	KQPW8lFf1y5BT2MxiSZ3QA	"Western Motor Vehicle"	NaN	"6025 N 27th Ave, Ste 1"	Phoenix	AZ	85017
3	8DShNS-LuFqpEWIp0HxijA	"Sports Authority"	NaN	"5000 Arizona Mills Cr, Ste 435"	Tempe	AZ	85282
4	PfOCPjBrlQAnz__NXj9h_w	"Brick House Tavern + Tap"	NaN	"581 Howe Ave"	Cuyahoga Falls	OH	44221

In [4]:
```
business_hours = pd.read_csv("../input/yelp_business_hours.csv")
```

In [5]:
```
business_hours.head()
```

Out[5]:

	business_id	monday	tuesday	wednesday	thursday	friday	saturday	sunday
0	FYWN1wneV18bWNgQjJ2GNg	7:30-17:0	7:30-17:0	7:30-17:0	7:30-17:0	7:30-17:0	None	None
1	He-G7vWjzVUysIKrfNbPUQ	9:0-20:0	9:0-20:0	9:0-20:0	9:0-20:0	9:0-16:0	8:0-16:0	None
2	KQPW8lFf1y5BT2MxiSZ3QA	None	None	None	None	None	None	None
3	8DShNS-LuFqpEWIp0HxijA	10:0-21:0	10:0-21:0	10:0-21:0	10:0-21:0	10:0-21:0	10:0-21:0	11:0-19:0
4	PfOCPjBrlQAnz__NXj9h_w	11:0-1:0	11:0-1:0	11:0-1:0	11:0-1:0	11:0-1:0	11:0-2:0	11:0-0:0

In [6]:
```
business.columns
```

Out[6]:
```
Index(['business_id', 'name', 'neighborhood', 'address', 'city', 'state',
       'postal_code', 'latitude', 'longitude', 'stars', 'review_count',
       'is_open', 'categories'],
      dtype='object')
```

In [7]:
```
business.shape
```

Out[7]:
```
(174567, 13)
```

In [8]:
```
#Null Values...
business.isnull().sum().sort_values(ascending=False)
```

Out[8]:
```
neighborhood    106552
postal_code        623
longitude            1
latitude             1
state                1
city                 1
categories           0
is_open              0
review_count         0
stars                0
address              0
name                 0
business_id          0
dtype: int64
```

In [9]:
```
#are all business Id's unique?
business.business_id.is_unique #business_id is all unique
```

Out[9]:
```
True
```

In [10]:

```
business.city.value_counts()
```

Out[10]:

Las Vegas	26775
Phoenix	17213
Toronto	17206
Charlotte	8553
Scottsdale	8228
Pittsburgh	6355
Mesa	5760
Montréal	5709
Henderson	4465
Tempe	4263
Chandler	3994
Edinburgh	3868
Cleveland	3322
Madison	3213
Glendale	3206
Gilbert	3128
Mississauga	2726
Stuttgart	2000
Peoria	1706
Markham	1564
North Las Vegas	1393
Champaign	1195
Scarborough	1095
North York	1092
Surprise	1018
Richmond Hill	888
Concord	864
Brampton	839
Goodyear	772
Vaughan	768
	...
M7	1
Pincourt	1
Dollard-des Ormeaux	1
East Gwilimbury	1
Canonsburd	1
Chateau	1
Lake Park	1
Middleburg Hts.	1
Chertsey	1
Currie	1
Allegheny	1

```
Bedford Hts.                    1
Hemmingford                     1
N W Las Vegas                   1
Chester Township                1
Shaker Hts                      1
Cleveland Hghts.                1
Lübeck                          1
Ben Avon                        1
Northfield Center Township      1
Baie-D'urfe                     1
Mesa Arizona                    1
Shandwick                       1
Plan                            1
Median                          1
Monreoville                     1
Leaside                         1
Henderson and Las vegas         1
Côte-Saint-Luc                  1
Vaughn City                     1
Name: city, Length: 1093, dtype: int64
```

Top 50 most reviewed businesses

In [11]:

```
business[['name', 'review_count', 'city', 'stars']].sort_values(ascending=False,
by="review_count")[0:50]
```

Out[11]:

	name	review_count	city	stars
97944	"Mon Ami Gabi"	7361	Las Vegas	4.0
119907	"Bacchanal Buffet"	7009	Las Vegas	4.0
69993	"Wicked Spoon"	5950	Las Vegas	3.5
81212	"Gordon Ramsay BurGR"	5447	Las Vegas	4.0
139699	"Earl of Sandwich"	4869	Las Vegas	4.5
19191	"Hash House A Go Go"	4774	Las Vegas	4.0
80590	"The Buffet"	4018	Las Vegas	3.5
124412	"Lotus of Siam"	3964	Las Vegas	4.0
21006	"Serendipity 3"	3910	Las Vegas	3.0
93038	"The Buffet at Bellagio"	3838	Las Vegas	3.5
26748	"ARIA Resort & Casino"	3794	Las Vegas	3.5
80626	"The Cosmopolitan of Las Vegas"	3772	Las Vegas	4.0
25096	"Secret Pizza"	3741	Las Vegas	4.0
6670	"Luxor Hotel and Casino Las Vegas"	3621	Las Vegas	2.5
6782	"Bouchon at the Venezia Tower"	3570	Las Vegas	4.0
10567	"MGM Grand Hotel"	3444	Las Vegas	3.0
169223	"McCarran International Airport"	3284	Las Vegas	3.5
170798	"Gangnam Asian BBQ Dining"	3262	Las Vegas	4.5
112523	"The Venetian Las Vegas"	3101	Las Vegas	4.0
116002	"Bachi Burger"	3065	Las Vegas	4.0
50087	"Hash House A Go Go"	3050	Las Vegas	4.0
43069	"Mesa Grill"	3012	Las Vegas	4.0
106267	"Flamingo Las Vegas Hotel & Casino"	2938	Las Vegas	2.5
138776	"Gordon Ramsay Steak"	2935	Las Vegas	4.0
60845	"XS Nightclub"	2884	Las Vegas	4.0
126918	"Bellagio Hotel"	2780	Las Vegas	3.5
124334	"Holsteins Shakes and Buns"	2771	Las Vegas	4.0
24586	"The Peppermill Restaurant & Fireside Lounge"	2703	Las Vegas	4.0
69105	"Mandalay Bay Resort & Casino"	2687	Las Vegas	3.5
8709	"Planet Hollywood Las Vegas Resort & Casino"	2681	Las Vegas	3.0
36120	"Guy Fieri's Vegas Kitchen & Bar"	2674	Las Vegas	3.5
128975	"Egg & I"	2595	Las Vegas	4.5
166920	"Pho Kim Long"	2594	Las Vegas	3.5
27862	"Shake Shack"	2549	Las Vegas	4.0
2803	"Monte Carlo Hotel And Casino"	2507	Las Vegas	2.5
800	"Excalibur Hotel"	2504	Las Vegas	2.5
118712	"Gordon Ramsay Pub & Grill"	2502	Las Vegas	3.5
108433	"Grand Lux Cafe"	2490	Las Vegas	4.0
100272	"Tacos El Gordo"	2448	Las Vegas	4.0
154617	"Wynn Las Vegas"	2441	Las Vegas	4.0

	name	review_count	city	stars
161630	"Burger Bar"	2440	Las Vegas	4.0
155142	"Caesars Palace Las Vegas Hotel & Casino"	2393	Las Vegas	3.0
73007	"Yardbird Southern Table & Bar"	2360	Las Vegas	4.5
5068	"Giada"	2349	Las Vegas	3.5
89974	"Rollin Smoke Barbeque"	2320	Las Vegas	4.5
13125	"Vdara Hotel"	2315	Las Vegas	4.0
20329	"Monta Ramen"	2291	Las Vegas	4.0
73708	"The Palazzo Las Vegas"	2248	Las Vegas	4.0
102148	"Treasure Island"	2237	Las Vegas	3.0
4137	"Phoenix Sky Harbor International Airport"	2215	Phoenix	3.0

Number of businesses listed in different cities

In [12]:
```
city_business_counts = business[['city', 'business_id']].groupby(['city'])\
['business_id'].agg('count').sort_values(ascending=False)
```

In [13]:
```
city_business_counts = pd.DataFrame(data=city_business_counts)
```

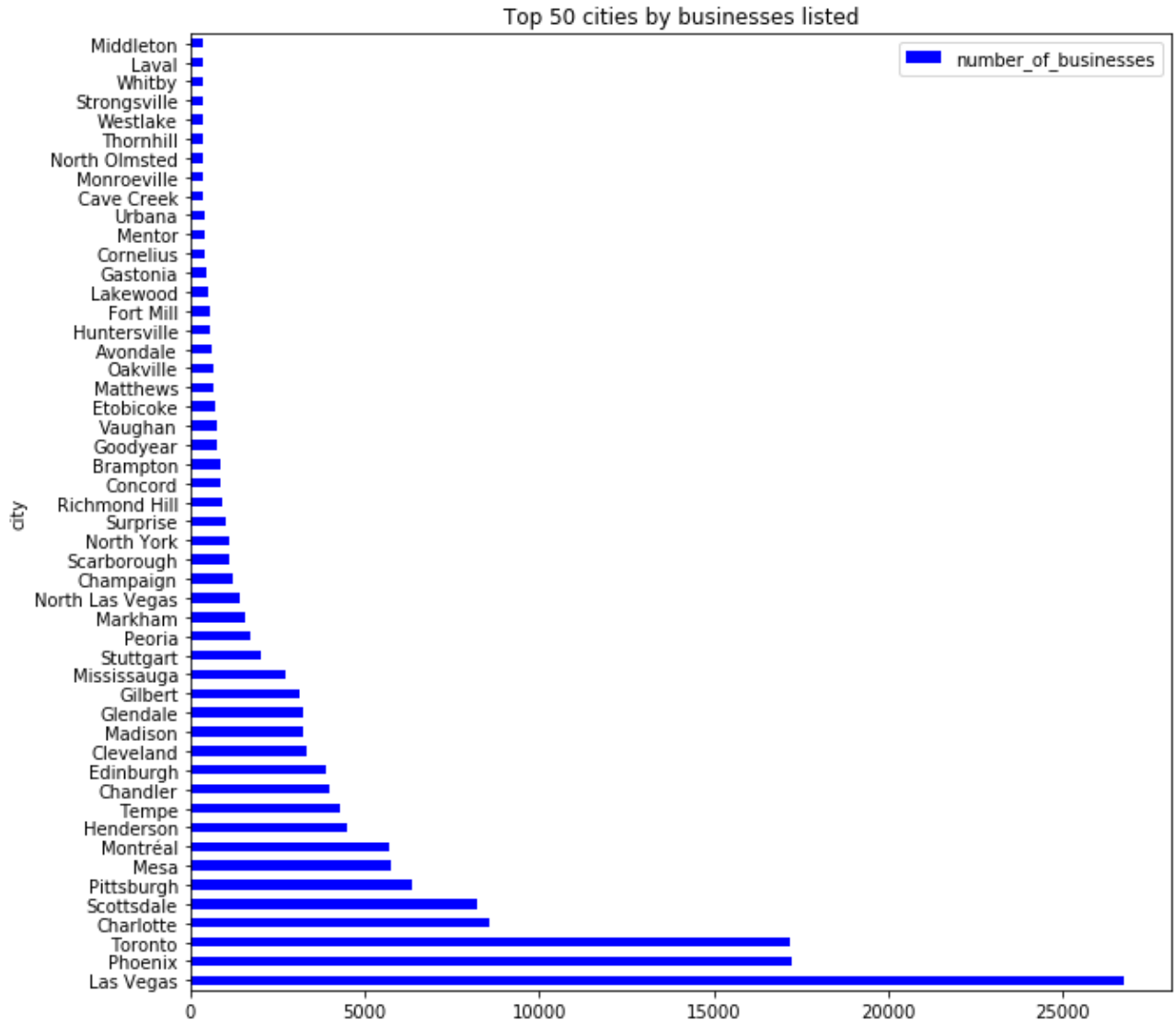
In [14]:
```
city_business_counts.rename(columns={'business_id' : 'number_of_businesses'}, inp
lace=True)
```

In [15]:

```
city_business_counts[0:50].sort_values(ascending=False, by="number_of_businesses"
)\
.plot(kind='barh', stacked=False, figsize=[10,10], colormap='winter')
plt.title('Top 50 cities by businesses listed')
```

Out[15]:

```
Text(0.5,1,'Top 50 cities by businesses listed')
```



Cities with most reviews and best ratings for their businesses

In [16]:

```
city_business_reviews = business[['city', 'review_count', 'stars']].groupby(['city']).\
agg({'review_count': 'sum', 'stars': 'mean'}).sort_values(by='review_count', ascending=False)
city_business_reviews.head(10)
```

Out[16]:

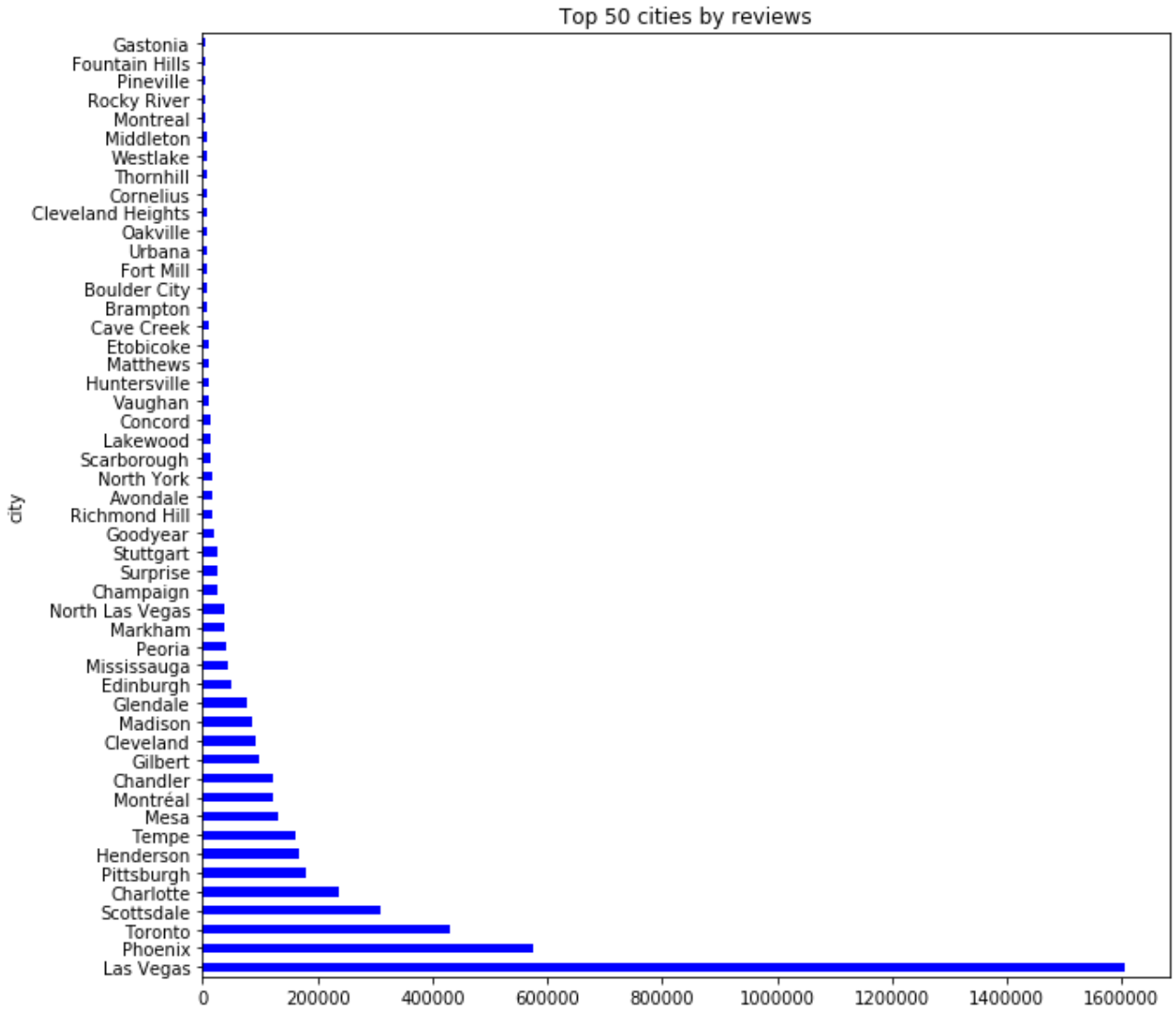
	review_count	stars
city		
Las Vegas	1604173	3.709916
Phoenix	576709	3.673793
Toronto	430923	3.487272
Scottsdale	308529	3.948529
Charlotte	237115	3.571554
Pittsburgh	179471	3.629819
Henderson	166884	3.789362
Tempe	162772	3.729885
Mesa	130883	3.636024
Montréal	122620	3.706604

In [17]:

```
city_business_reviews['review_count'][0:50].plot(kind='barh', stacked=False, figsize=[10,10], \
                                                 colormap='winter')
plt.title('Top 50 cities by reviews')
```

Out[17]:

```
Text(0.5,1,'Top 50 cities by reviews')
```

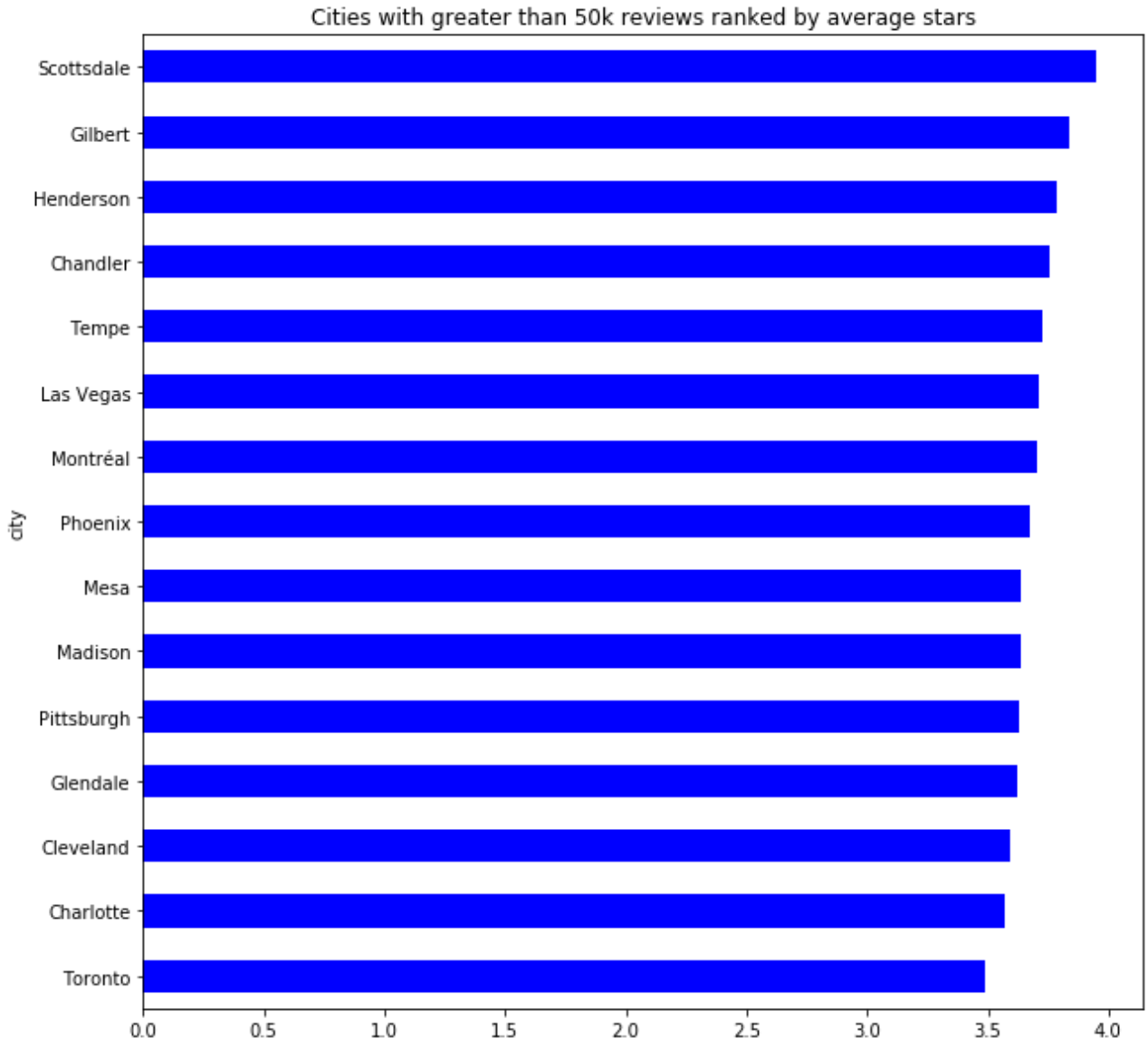


In [18]:

```
city_business_reviews[city_business_reviews.review_count > 50000]['stars'].sort_v
alues()\
.plot(kind='barh', stacked=False, figsize=[10,10], colormap='winter')
plt.title('Cities with greater than 50k reviews ranked by average stars')
```

Out[18]:

```
Text(0.5,1,'Cities with greater than 50k reviews ranked by average stars')
```



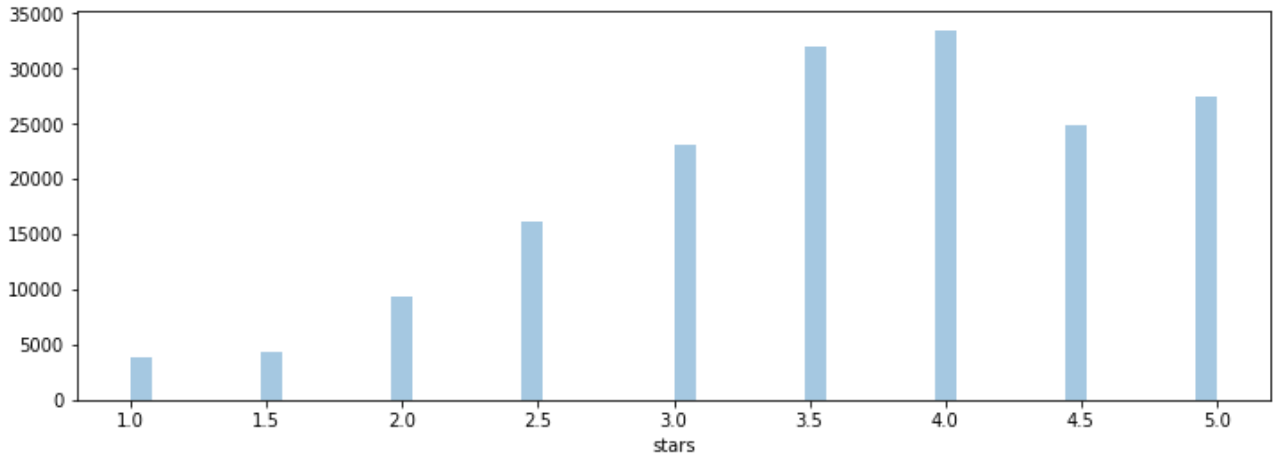
Distribution of stars

In [19]:
```
business['stars'].value_counts()
```

Out[19]:
```
4.0    33492
3.5    32038
5.0    27540
4.5    24796
3.0    23142
2.5    16148
2.0     9320
1.5     4303
1.0     3788
Name: stars, dtype: int64
```

In [20]:
```
sns.distplot(business.stars, kde=False)
```

Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fddf08c5358>
```



How many are open and how many closed?

In [21]:
```
business['is_open'].value_counts()
```

Out[21]:
```
1    146702
0     27865
Name: is_open, dtype: int64
```

Lets look into user tips on businesses before looking at reviews

In [22]:
```
tip = pd.read_csv('../input/yelp_tip.csv')
```

In [23]:
```
tip.head(10)
```

Out[23]:

	text	date	likes	business_id	user_id
0	Great breakfast large portions and friendly wa...	2015-08-12	0	jH19V2I9fIslnNhDzPmdkA	ZcLKXikTHYOnYt5VYRO5sg
1	Nice place. Great staff. A fixture in the tow...	2014-06-20	0	dAa0hB2yrnHzVmsCkN4YvQ	oaYhjqBbh18ZhU0bpyzSuw
2	Happy hour 5-7 Monday - Friday	2016-10-12	0	dAa0hB2yrnHzVmsCkN4YvQ	ulQ8Nyj7jCUR8M83SUMoRQ
3	Parking is a premium, keep circling, you will ...	2017-01-28	0	ESzO3Av0b1_TzKOiqzbQYQ	ulQ8Nyj7jCUR8M83SUMoRQ
4	Homemade pasta is the best in the area	2017-02-25	0	k7WRPbDd7rztjHcGGkEjlw	ulQ8Nyj7jCUR8M83SUMoRQ
5	Excellent service, staff is dressed profession...	2017-04-08	0	k7WRPbDd7rztjHcGGkEjlw	ulQ8Nyj7jCUR8M83SUMoRQ
6	Come early on Sunday's to avoid the rush	2016-07-03	0	SqW3igh1_Png336VIb5DUA	ulQ8Nyj7jCUR8M83SUMoRQ
7	Love their soup!	2016-01-07	0	KNpcPGqDORDdvtekXd348w	ulQ8Nyj7jCUR8M83SUMoRQ
8	Soups are fantastic!	2016-05-22	0	KNpcPGqDORDdvtekXd348w	ulQ8Nyj7jCUR8M83SUMoRQ
9	Thursday night is $5 burger night	2016-06-09	0	KNpcPGqDORDdvtekXd348w	ulQ8Nyj7jCUR8M83SUMoRQ

In [24]:
```
tip.shape
```

Out[24]:
```
(1098324, 5)
```

How many of the selected words are used in the user tips?

In [25]:
```
selected_words = ['awesome', 'great', 'fantastic', 'amazing', 'love', 'horrible',
'bad', 'terrible',
                  'awful', 'wow', 'hate']
selected_words
```

Out[25]:
```
['awesome',
 'great',
 'fantastic',
 'amazing',
 'love',
 'horrible',
 'bad',
 'terrible',
 'awful',
 'wow',
 'hate']
```

In [26]:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(vocabulary=selected_words, lowercase=False)
#corpus = ['This is the first document.','This is the second second document.']
#print corpus
selected_word_count = vectorizer.fit_transform(tip['text'].values.astype('U'))
vectorizer.get_feature_names()
```

Out[26]:

```
['awesome',
 'great',
 'fantastic',
 'amazing',
 'love',
 'horrible',
 'bad',
 'terrible',
 'awful',
 'wow',
 'hate']
```

In [27]:

```
word_count_array = selected_word_count.toarray()
word_count_array.shape
```

Out[27]:

```
(1098324, 11)
```

In [28]:

```
word_count_array.sum(axis=0)
```

Out[28]:

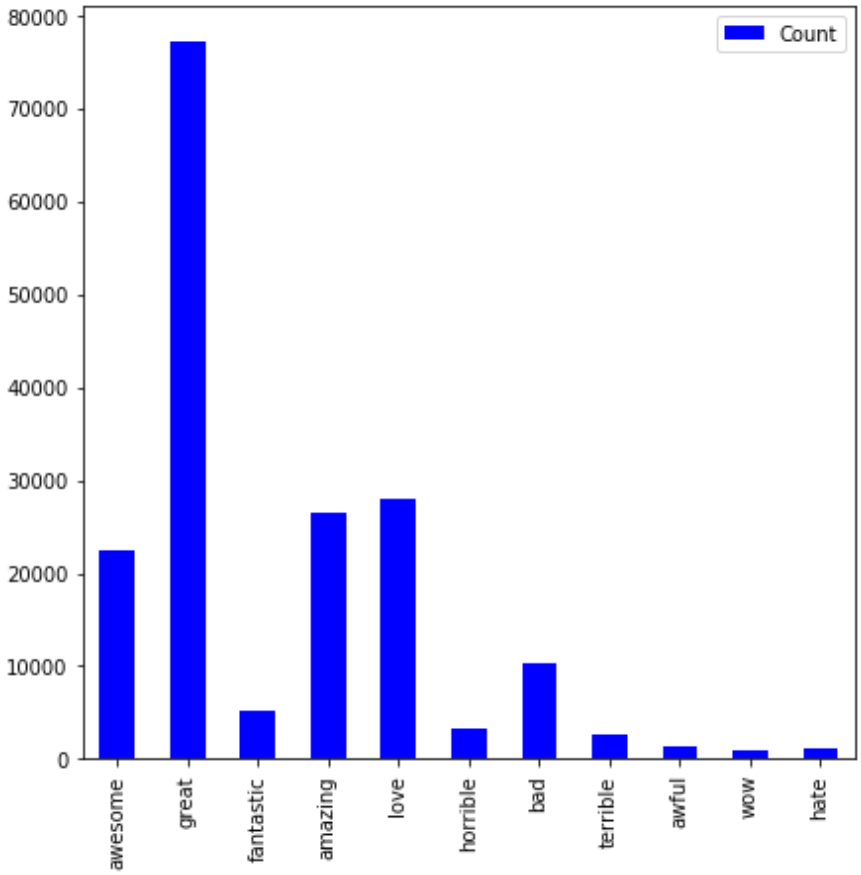
```
array([22354, 77169,  5168, 26547, 27972,  3233, 10207,  2589,  1338,
         862,  1214])
```

In [29]:

```
temp = pd.DataFrame(index=vectorizer.get_feature_names(), \
                    data=word_count_array.sum(axis=0)).rename(columns={0: 'Count'
})
```

In [30]:
```
temp.plot(kind='bar', stacked=False, figsize=[7,7], colormap='winter')
```

Out[30]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdd11c57470>
```



We see that most of the tips are positive rather than negative!

Lets look at one restaurant with high star rating and one with low star rating and see what the user tips look like

Lets look at "Earl of Sandwich" restaurant in Las Vegas which has 4.5 rating

In [31]:

```
business[(business['city'] == 'Las Vegas') & (business['stars'] == 4.5)]
```

Out[31]:

	business_id	name	neighborhood	address	city	state	postal_
26	VBHEsoXQb2AQ76J9l8h1uQ	"Alfredo's Jewelry"	Southeast	"5775 S Eastern, Ste 103"	Las Vegas	NV	89119
41	1Jp_hmPNUZArNqzpbm7B0g	"Task Electric"	Spring Valley	"7260 Cimarron Rd, Ste 130"	Las Vegas	NV	89113
60	v2GJWvZqEAjUc22hZUYzYw	"John Armond Actor's Studio"	Westside	"8125 W Sahara Ave, Ste 210"	Las Vegas	NV	89117
82	bOOgAB_CEWWsxalAthnRSw	"Tenors of Rock"	The Strip	"3475 S Las Vegas Blvd"	Las Vegas	NV	89109
110	ZmMCgM4RCqCXJ0Lswu6yxw	"A Professional Appliance Repair"	NaN	""	Las Vegas	NV	89108
153	PJ-VbAtIOso1dqd2frQqqg	"Donut Tyme"	Sunrise	"4268 E Charleston Blvd"	Las Vegas	NV	89104
176	cehTmoCXPi0a3FwCE3Tq2Q	"Red Wing Shoes"	Southeast	"2370 E Serene Ave"	Las Vegas	NV	89123
177	Uy3_5nLo3sYkAuSX6mjdmg	"Geebee's Bar & Grill"	Southeast	"8560 Las Vegas Blvd S"	Las Vegas	NV	89123
193	P5TLch0Fu9p3o6W2hRSz0g	"Terrible Herbst"	Southeast	"1785 E Sunset Rd"	Las Vegas	NV	89119
225	xiSEUnaX77EhNz-l3ag7RA	"LV Nail Lounge"	Spring Valley	"8530 W Warm Spring Rd, Ste 105"	Las Vegas	NV	89113
274	dPxZI9lrKTl5dvFfnb1_Ig	"Trattoria Italia"	Anthem	"9905 S Eastern Ave, Ste 140"	Las Vegas	NV	89183
303	WfB_SsYeKy83QQsqAAyGVQ	"Cancun Bar & Grill"	Southeast	"5006 S Maryland Pkwy, Ste 17"	Las Vegas	NV	89119
333	oDMiR7xWFWNSG4zOXFajdg	"Coral Academy of Science"	Southeast	"8185 Tamarus St"	Las Vegas	NV	89123
366	u29lf2yPd-qK5ThAS9FRQQ	"Kinthai"	Westside	"4105 W Sahara Ave"	Las Vegas	NV	89102
457	1e2ZRUm9lpX3vrmraRx-yQ	"Tacos Colima"	NaN	""	Las Vegas	NV	89032
505	bs5ZQW4z83ml6kuZWd-Q1A	"Enliven Skin and Beauty"	Westside	"8751 W Charleston Blvd"	Las Vegas	NV	89117
518	lil28runuoryt7uMQLXSRQ	"Flipperspiel Underground Arcade Club"	Southeast	"6000 S Eastern Ave, Ste 4D"	Las Vegas	NV	89119
574	0Yeb_P24sj6MwG2qmuehkA	"Till's Bar"	NaN	"344 E Desert Inn Rd"	Las Vegas	NV	89109
596	hTzcHtk4-0QJnFUbkKpd5Q	"Citi Trends"	NaN	"703 N Rancho Dr"	Las Vegas	NV	89106

	business_id	name	neighborhood	address	city	state	postal_
691	_ewxwEwJM-IYfIYnKpQOZw	"Mexicali Raspados"	Eastside	"4865 S Pecos Rd"	Las Vegas	NV	89121
778	v0byOL8VL6v6muGa1anxFA	"The Hummus Factory"	Westside	"7875 W Sahara Ave, Ste 101"	Las Vegas	NV	89117
783	i57cZRoLUU9QUPCI0ErWGQ	"Dad's Bail Bonds"	Downtown	"600 S 3rd St"	Las Vegas	NV	89101
839	3MBON3dW2a1NKjgo9H780Q	"Anson Edwards & Higgins Plastic Surgery Assoc...	Spring Valley	"8530 W Sunset Rd, Ste 130"	Las Vegas	NV	89113
883	VOHbo_5g1rLwu65AkMHacQ	"Hair'z Melinda"	Summerlin	"10300 W Charleston Blvd"	Las Vegas	NV	89135
918	crGdKSRKi25R2vorZ7skzg	"Tortilleria San Diego"	Downtown	"235 N Eastern Ave"	Las Vegas	NV	89101
952	kvRJjMN1XZtfgLxhVP_BPw	"Modern Landscape, LLC"	Centennial	"7733 Silver Mallard Ave"	Las Vegas	NV	89131
974	4Zqv7NyeiuqMOV8wWlhB4A	"ReVamp Extensions"	Spring Valley	"9640 W Tropicana Ave, Ste 125, Studio102"	Las Vegas	NV	89147
990	zcVZc4SadqLgUHKWL2ZilQ	"ChefKas Catering"	Spring Valley	""	Las Vegas	NV	89118
1070	USMFeacfapj3IXvZahIj-A	"Vape Kraze Vapors"	Downtown	"618 S Las Vegas Blvd"	Las Vegas	NV	89101
1077	Njydzc1qePniw9hdkikYbg	"Gettinger Chiropractic"	Downtown	"1229 S Eastern Ave"	Las Vegas	NV	89104
...	...	...	...	...	...	...	...
173097	gKj0wkJhQ0YT-Aw65e7hQQ	"Stephen K Montoya, MD"	Eastside	"3196 S Maryland Pkwy"	Las Vegas	NV	89109
173149	1La-mfFF5RTYOhujtSUVaQ	"911 Keys & Sound"	Eastside	""	Las Vegas	NV	89121
173184	N0YNLwSNejIpJ1E3qQorhg	"FiveStar Mobile iPhone Repair"	The Strip	""	Las Vegas	NV	89109
173206	pvASvwMlwe7-zY5rN9oBAQ	"Fast-Fix Jewelry & Watch Repairs"	South Summerlin	"1980 Festival Plaza Dr, Ste 120"	Las Vegas	NV	89135
173225	VIW-4GDAKGaXUdLtTVSFqQ	"Khoury's Fine Wine & Spirits"	Anthem	"9915 S Eastern Ave, Ste 110"	Las Vegas	NV	89183
173343	maUGKOLNdvYKUVQoyh0-gg	"The Law Office of Joseph P Reiff"	Downtown	"3001 E Charleston Blvd, Ste A"	Las Vegas	NV	89104
173393	fnfh9LuxfmLw59vC_kEBCA	"Madewell"	The Strip	"3200 Las Vegas Blvd S, Ste 2470"	Las Vegas	NV	89109

	business_id	name	neighborhood	address	city	state	postal_
173461	1_RnfPCfPKAhEolhM5yeUw	"Scott Biggs, DDS - Micro Endodontics"	Northwest	"4450 N Tenaya Way, Ste 240"	Las Vegas	NV	89129
173500	ad7NXTxHr2BmjyQvGnfotQ	"Specialty Surgery Center"	NaN	"7250 Cathedral Rock Dr"	Las Vegas	NV	89128
173518	LIU7IcJtD9VieoIo__wd9Q	"Thai Pan Cuisine"	Westside	"463 S Decatur Blvd"	Las Vegas	NV	89107
173580	QMozb3XreozGjEMrabs__A	"The Joint Chiropractic - Blue Diamond"	Southwest	"4150 Blue Diamond Rd, Ste 107"	Las Vegas	NV	89139
173612	EwN1LCoJXB0z_a-LxLFKyQ	"Paleteria Y Neveria Mexicana"	Sunrise	"865 N Lamb Blvd, Ste 10"	Las Vegas	NV	89110
173781	BD18SKv935HDmlKrLPkhLA	"Chocolate Swan"	NaN	"3930 Las Vegas Blvd S., Ste 121B, Mandalay Ba...	Las Vegas	NV	89162
173825	kd96_x4saxuoe_eJHbS87A	"Martin Garage Doors of Nevada"	NaN	"6667 S Schuster St"	Las Vegas	NV	89118
173915	lmpJvZI_F9iBmNoHi4jW1A	"Bestcuts & More"	NaN	"5255 S Decatur Blvd"	Las Vegas	NV	89118
174011	ekTUBCcsRTheOvYa7fPQpQ	"Sweet Bubble Bath Confections in Excalibur"	The Strip	"3850 S Las Vegas Blvd"	Las Vegas	NV	89109
174020	1dl7wV-zwqliz0xOrzLBUQ	"Universal Solar Direct of Las Vegas"	NaN	"4775 W Teco Ave, Ste 115"	Las Vegas	NV	89118
174098	Tefx_N6A6nrsdj4jHHnbYg	"Le Petit Café & Bakery"	Spring Valley	"6496 Medical Center St, Ste 100"	Las Vegas	NV	89148
174131	ZEouZiCVwjla4ePWboaVkg	"Raysco"	NaN	"3995 W Quail Ave, Ste D"	Las Vegas	NV	89118
174184	_rQb4DXr4i-XOb3c_LOKdg	"Bikram Yoga Westside"	South Summerlin	"3700 S Hualapai Way, Ste 203-204"	Las Vegas	NV	89147
174195	WBdgcjOt9qJfaeGcdzaMLA	"Chic Cache"	Westside	"5191 W Charleston Blvd, Ste 175"	Las Vegas	NV	89146
174220	muriGdv1pnJaNZTQfZq9CQ	"Kim Layson Beauty"	Spring Valley	"8680 W Warmsprings Rd"	Las Vegas	NV	89148
174321	m3_NFDiJ8ib2fUzoqwzm5Q	"Buy Buy Baby"	South Summerlin	"2315 Summa Dr, Ste 120"	Las Vegas	NV	89135

	business_id	name	neighborhood	address	city	state	postal_
174332	RgLA2YwJ53xoeMMgc7L7oA	"Scooter Nation"	Southwest	"7060 W Warm Springs Rd, Ste 130"	Las Vegas	NV	89113
174355	Nu-SSGx_BFb9eMOm8qO4Mg	"Paul's Auto Service"	Downtown	"1754 E Charleston Blvd"	Las Vegas	NV	89104
174380	l3l3RvS7IXogVpanFu6QlA	"Nulook Floor"	NaN	"5277 Cameron St, Ste 120"	Las Vegas	NV	89118
174386	EZ0pK8z6jG8uv4DNZhrRuA	"11th Street Records"	Downtown	"1023 Fremont St"	Las Vegas	NV	89101
174417	MKrvEEejLBeUsjZRBtVxrQ	"Green Valley Shoe Repair"	Southwest	"7835 S Rainbow Blvd, Ste 21"	Las Vegas	NV	89139
174455	Fv4EXwV30rwGD2NzN1ekgA	"Gorilla Sushi"	Eastside	"1801 E Tropicana Ave, Ste 2"	Las Vegas	NV	89119
174539	swjz4q8gI79Ndg4APuHEUA	"Stonegate Real Estate Services"	Westside	"3030 S Jones Blvd, Ste 105"	Las Vegas	NV	89146

4006 rows × 13 columns

In [32]:

```
business[business.name=='"Earl of Sandwich"']
```

Out[32]:

	business_id	name	neighborhood	address	city	state	postal_code
107416	Ffhe2cmRyloz3CCdRGvHtA	"Earl of Sandwich"	NaN	"4321 W Flamingo Rd"	Las Vegas	NV	89103
131049	3fT1kcQ-MVEImGHa3hll5w	"Earl of Sandwich"	South Summerlin	"2010 Festival Plaza Dr, Ste 180"	Las Vegas	NV	89135
139699	DkYS3arLOhA8si5uUEmHOw	"Earl of Sandwich"	The Strip	"3667 Las Vegas Blvd S"	Las Vegas	NV	89109
166792	fE7x3Ui2mzdwdfJnd7r_1g	"Earl of Sandwich"	The Strip	"3570 Las Vegas Blvd S"	Las Vegas	NV	89109

- Points to remember:
  1. There are 4 branches
  2. Two of them are on the strip
  3. Since there are multiple, lets pick by index

In [33]:
```
# This is where  have been to :)
business.loc[139699,:]
```

Out[33]:
```
business_id                                  DkYS3arLOhA8si5uUEmHOw
name                                             "Earl of Sandwich"
neighborhood                                                The Strip
address                                       "3667 Las Vegas Blvd S"
city                                                       Las Vegas
state                                                             NV
postal_code                                                    89109
latitude                                                     36.1082
longitude                                                   -115.172
stars                                                            4.5
review_count                                                    4869
is_open                                                            1
categories      Caterers;Sandwiches;Restaurants;Food Delivery ...
Name: 139699, dtype: object
```

In [34]:
```
earl_of_sandwich = tip[tip.business_id==business.loc[139699,:].business_id]
```

In [35]:
```
earl_of_sandwich_selected_word_count = \
vectorizer.fit_transform(earl_of_sandwich['text'].values.astype('U'))
```

In [36]:

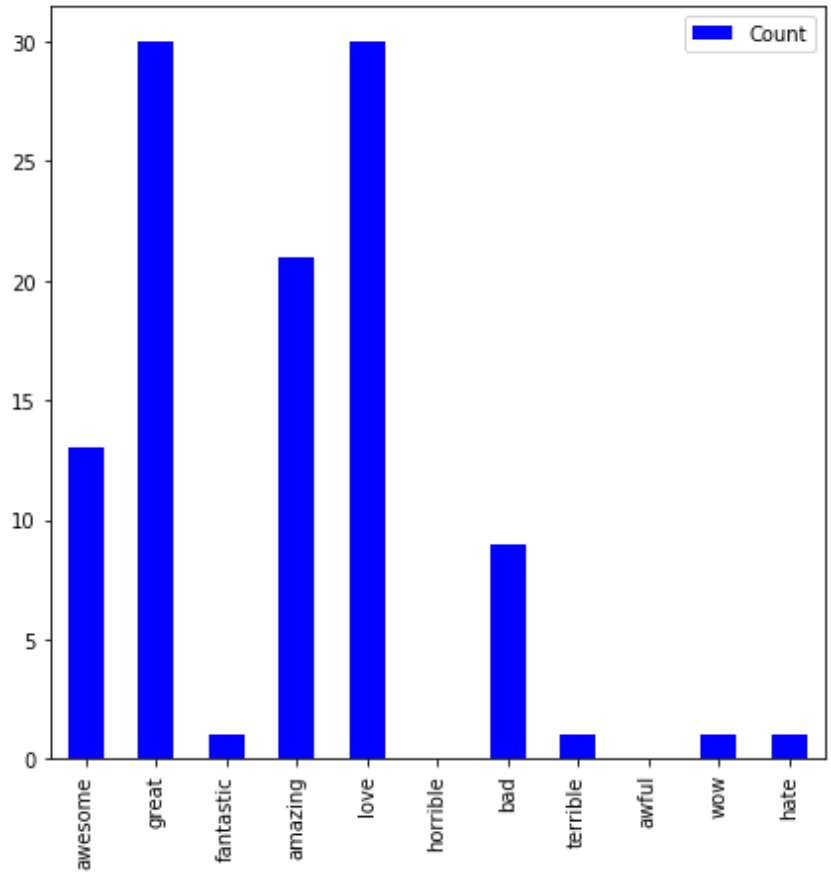
```
word_count_array = earl_of_sandwich_selected_word_count.toarray()
temp = pd.DataFrame(index=vectorizer.get_feature_names(), \
                    data=word_count_array.sum(axis=0)).rename(columns={0: 'Count'
})
temp
```

Out[36]:

	Count
awesome	13
great	30
fantastic	1
amazing	21
love	30
horrible	0
bad	9
terrible	1
awful	0
wow	1
hate	1

In [37]:
```
temp.plot(kind='bar', stacked=False, figsize=[7,7], colormap='winter')
```

Out[37]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fddf7ca12b0>
```



We can see that the tips are mostly positive!

In [38]:

```
business[['name', 'review_count', 'city', 'stars']]\
[business.review_count>1000].sort_values(ascending=True, by="stars")[0:15]
```

Out[38]:

	name	review_count	city	stars
26796	"Fox Rent A Car"	1036	Las Vegas	1.5
135687	"Westgate Las Vegas Resort & Casino"	1288	Las Vegas	2.0
800	"Excalibur Hotel"	2504	Las Vegas	2.5
106267	"Flamingo Las Vegas Hotel & Casino"	2938	Las Vegas	2.5
108652	"Stratosphere"	1662	Las Vegas	2.5
122546	"Hooters Casino Hotel Las Vegas"	1334	Las Vegas	2.5
75338	"Harrah's Las Vegas Hotel & Casino"	1413	Las Vegas	2.5
138691	"Circus Circus Las Vegas Hotel and Casino"	2171	Las Vegas	2.5
74551	"Rio All Suites Hotel & Casino"	2080	Las Vegas	2.5
65954	"Hakkasan Nightclub"	1607	Las Vegas	2.5
88571	"MGM Grand Buffet"	1005	Las Vegas	2.5
5494	"The LINQ Hotel & Casino"	1469	Las Vegas	2.5
6670	"Luxor Hotel and Casino Las Vegas"	3621	Las Vegas	2.5
2803	"Monte Carlo Hotel And Casino"	2507	Las Vegas	2.5
155142	"Caesars Palace Las Vegas Hotel & Casino"	2393	Las Vegas	3.0

Lets look into Luxor Hotel and Casino Las Vegas which has a 2.5 star

In [39]:

```
business[business['name'] == '"Luxor Hotel and Casino Las Vegas"']
```

Out[39]:

	business_id	name	neighborhood	address	city	state	postal_code	latit
6670	AV6weBrZFFBfRGCbcRGO4g	"Luxor Hotel and Casino Las Vegas"	The Strip	"3900 Las Vegas Blvd S"	Las Vegas	NV	89109	36.095

In [40]:

```
luxor_hotel = tip[tip.business_id==business.loc[6670,:].business_id]
luxor_hotel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 662 entries, 14086 to 718609
Data columns (total 5 columns):
text           662 non-null object
date           662 non-null object
likes          662 non-null int64
business_id    662 non-null object
user_id        662 non-null object
dtypes: int64(1), object(4)
memory usage: 31.0+ KB
```

In [41]:

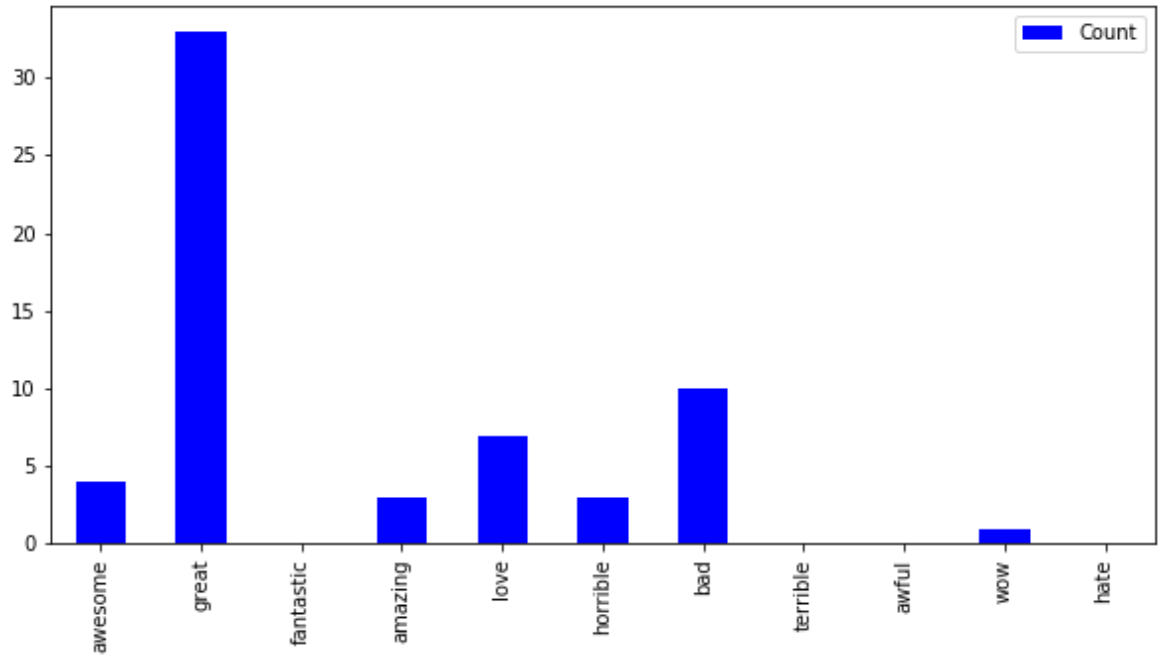
```
luxor_hotel_selected_word_count = vectorizer.fit_transform(luxor_hotel['text'].va
lues.astype('U'))
```

In [42]:

```
word_count_array = luxor_hotel_selected_word_count.toarray()
temp = pd.DataFrame(index=vectorizer.get_feature_names(), \
                    data=word_count_array.sum(axis=0)).rename(columns={0: 'Count'
})
```

In [43]:
```
temp.plot(kind='bar', stacked=False, figsize=[10,5], colormap='winter')
```

Out[43]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdda0abc2b0>
```



This has more positive words than negative, so the user tips for this restaurant are not very predictive of its star! This might make sense because while users write good and bad reviews, tips are naturally like to be what they liked and therefore positive!

Lets look into user reviews

In [44]:
```
reviews = pd.read_csv('../input/yelp_review.csv')
```

In [45]:
```
reviews.shape, tip.shape #there are 5.26 million reviews! 1 million tips
```

Out[45]:
```
((5261668, 9), (1098324, 5))
```

In [46]:
```
reviews.head(5)
```

Out[46]:

	review_id	user_id	business_id	stars	date	
0	vkVSCC7xljjrAI4UGfnKEQ	bv2nCi5Qv5vroFiqKGopiw	AEx2SYEUJmTxVVB18LlCwA	5	2016-05-28	Su pla am nor It...
1	n6QzIUObkYshz4dz2QRJTw	bv2nCi5Qv5vroFiqKGopiw	VR6GpWIda3SfvPC-lg9H3w	5	2016-05-28	Sm una pla cha me
2	MV3CcKScW05u5LVfF6ok0g	bv2nCi5Qv5vroFiqKGopiw	CKC0-MOWMqoeWf6s-szl8g	5	2016-05-28	Les loc bea nei
3	IXvOzsEMYtiJI0CARmj77Q	bv2nCi5Qv5vroFiqKGopiw	ACFtxLv8pGrrxMm6EgjreA	4	2016-05-28	Lov her pla nee
4	L_9BTb55X0GDtThi6GlZ6w	bv2nCi5Qv5vroFiqKGopiw	s2I_Ni76bjJNK9yG60iD-Q	4	2016-05-28	Ha cho alm cro it w

In [47]:
```
tip.head()
```

Out[47]:

	text	date	likes	business_id	user_id
0	Great breakfast large portions and friendly wa...	2015-08-12	0	jH19V2I9fIslnNhDzPmdkA	ZcLKXikTHYOnYt5VYRO5sg
1	Nice place. Great staff. A fixture in the tow...	2014-06-20	0	dAa0hB2yrnHzVmsCkN4YvQ	oaYhjqBbh18ZhU0bpyzSuw
2	Happy hour 5-7 Monday - Friday	2016-10-12	0	dAa0hB2yrnHzVmsCkN4YvQ	ulQ8Nyj7jCUR8M83SUMoRQ
3	Parking is a premium, keep circling, you will ...	2017-01-28	0	ESzO3Av0b1_TzKOiqzbQYQ	ulQ8Nyj7jCUR8M83SUMoRQ
4	Homemade pasta is the best in the area	2017-02-25	0	k7WRPbDd7rztjHcGGkEjlw	ulQ8Nyj7jCUR8M83SUMoRQ

How many of these restaurants serve Japanese food? Lets find out based on reviews!

In [48]:
```
selected_words = ['sushi', 'miso', 'teriyaki', 'tempura', 'udon', \
                  'soba', 'ramen', 'yakitori', 'izakaya']
```

Lets take subset of reviews since there are so many

More data analysis and data science to follow in this and other notebooks!! :)