In [18]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn
import re
```

Import the data and get a high-level picture

In [19]:

```
df = pd.read_csv('sales.csv')
df.head()
```

Out[19]:

	order_id	name	ordered_at	price	quantity	line_total
0	10000	"ICE CREAM" Peanut Fudge	2018-01-01 11:30:00	$3.50	3	$10.50
1	10000	"ICE CREAM" Peanut Fudge	2018-01-01 11:30:00	$3.50	1	$3.50
2	10001	"SORBET" Raspberry	2018-01-01 12:14:54	$2.50	2	$5.00
3	10001	NaN	2018-01-01 12:14:54	$1.50	1	$1.50
4	10001	"CONE" Dipped Waffle Cone	2018-01-01 12:14:54	$3.50	1	$3.50

In [20]:

```
df.shape
```

Out[20]:

```
(29922, 6)
```

In [21]:

```
df.dtypes
```

Out[21]:

```
order_id        int64
name           object
ordered_at     object
price          object
quantity        int64
line_total     object
dtype: object
```

TODO: Fix column datatypes

Change ordered_at to datetime

Change price and line_total to float

In [51]:

```
df['ordered_at'] = pd.to_datetime(df['ordered_at'])
```

In [47]:

```
df['line_total']=pd.to_numeric(df['line_total'].apply(lambda x:x[1:]))
```

In [52]:

```
df.dtypes
```

Out[52]:

```
order_id                int64
name                   object
ordered_at     datetime64[ns]
price                  object
quantity                int64
line_total            float64
dtype: object
```

TODO: drop if duplicated or null

In [34]:

```
df[df.duplicated()].shape[0]
```

Out[34]:

```
538
```

In [56]:

```
df = df.drop(list(df[df.duplicated()].index))
df
```

Out[56]:

	order_id	name	ordered_at	price	quantity	line_total
139	10049	"CONE" Dipped Waffle Cone	2018-01-02 13:47:55	$3.50	1	3.5
178	10063	"ICE CREAM" Earl Gray	2018-01-02 21:28:36	$0.50	3	1.5
207	10073	"ICE CREAM" Mint Chip	2018-01-03 01:45:00	$1.50	3	4.5
273	10091	"SORBET" Blood Orange	2018-01-03 09:14:48	$2.50	2	5.0
278	10092	"CONE" Sugar Cone	2018-01-03 09:44:29	$1.00	2	2.0
351	10118	"SORBET" Blood Orange	2018-01-03 22:56:09	$2.50	3	7.5
500	10167	"ICE CREAM" Matcha	2018-01-04 18:18:03	$1.50	1	1.5
768	10258	"ICE CREAM" Candied Bacon	2018-01-06 12:15:38	$0.50	1	0.5
776	10260	"ICE CREAM" Wildberry	2018-01-06 13:20:51	$1.50	2	3.0
814	10271	"MISC" Ice Cream Cake	2018-01-06 16:46:11	$2.00	3	6.0
879	10294	"ICE CREAM" Maple Brown Sugar	2018-01-07 03:47:41	$2.00	3	6.0
889	10298	"ICE CREAM" Wildberry	2018-01-07 05:49:11	$1.50	3	4.5
924	10308	"ICE CREAM" Double Fudge Chunk	2018-01-07 10:47:11	$3.50	1	3.5
944	10316	"CONE" Sugar Cone	2018-01-07 13:48:19	$1.00	3	3.0
1021	10345	"ICE CREAM" Dulce De Leche	2018-01-08 03:20:38	$1.50	3	4.5
1075	10361	"ICE CREAM" Vanilla Bean	2018-01-08 11:46:54	$1.50	1	1.5
1115	10371	"CONE" Brownie Cone	2018-01-08 16:47:53	$3.00	3	9.0
1140	10380	"ICE CREAM" Earl Gray	2018-01-08 20:13:03	$0.50	2	1.0
1163	10387	"MISC" Ice Cream Cake	2018-01-08 23:47:45	$2.00	3	6.0
1180	10393	"SORBET" Raspberry	2018-01-09 02:32:23	$2.50	3	7.5
1234	10414	"BEVERAGE" Espresso	2018-01-09 13:42:09	$2.50	1	2.5
1334	10444	"ICE CREAM" Matcha	2018-01-10 04:30:53	$1.50	3	4.5

	order_id	name	ordered_at	price	quantity	line_total
1448	10482	"CONE" Brownie Cone	2018-01-10 22:40:59	$3.00	2	6.0
1519	10502	"ICE CREAM" Strawberry	2018-01-11 08:48:07	$3.50	1	3.5
1650	10546	"ICE CREAM" Peanut Fudge	2018-01-12 06:36:06	$3.50	1	3.5
1727	10577	"BEVERAGE" Iced Coffee	2018-01-13 00:04:31	$2.50	1	2.5
1747	10586	"CONE" Dipped Waffle Cone	2018-01-13 05:28:03	$3.50	2	7.0
1849	10619	"ICE CREAM" Vanilla Bean	2018-01-13 21:52:08	$1.50	1	1.5
2168	10717	"ICE CREAM" Dulce De Leche	2018-01-15 23:35:28	$1.50	3	4.5
2240	10741	"SORBET" Lemon	2018-01-16 12:03:32	$2.50	2	5.0
...	...	...	...	...	...	...
29892	17082	"ICE CREAM" Maple Brown Sugar	2018-05-29 06:13:56	$2.00	2	4.0
29893	15532	"CONE" Waffle Cone	2018-04-25 22:08:26	$4.00	1	4.0
29894	12462	"SORBET" Blood Orange	2018-02-22 13:27:40	$2.50	3	7.5
29895	19620	"ICE CREAM" Double Fudge Chunk	2018-07-20 23:10:36	$3.50	1	3.5
29896	19221	"CONE" Sugar Cone	2018-07-12 03:48:50	$1.00	2	2.0
29897	12474	"ICE CREAM" Earl Gray	2018-02-22 17:23:57	$0.50	1	0.5
29898	17161	"ICE CREAM" Wildberry	2018-05-30 22:54:10	$1.50	2	3.0
29899	15838	"ICE CREAM" Maple Brown Sugar	2018-05-02 22:15:45	$2.00	2	4.0
29900	18082	"SORBET" Raspberry	2018-06-18 13:36:16	$2.50	2	5.0
29901	12331	"ICE CREAM" Dark Chocolate	2018-02-19 19:04:01	$4.00	1	4.0
29902	15373	"CONE" Dipped Waffle Cone	2018-04-22 22:17:45	$3.50	3	10.5
29903	12456	"CONE" Cookie Cone	2018-02-22 10:20:27	$4.00	2	8.0
29904	11079	"ICE CREAM" Vanilla Bean	2018-01-23 17:15:32	$1.50	1	1.5
29905	15969	"CONE" Brownie Cone	2018-05-05 17:50:35	$3.00	2	6.0
29906	19748	"MISC" Ice Cream Cake	2018-07-23 14:18:20	$2.00	1	2.0
29907	11601	"ICE CREAM" Rocky Road	2018-02-04 14:36:06	$3.50	1	3.5

	order_id	name	ordered_at	price	quantity	line_total
29908	13140	"CONE" Brownie Cone	2018-03-08 06:00:34	$3.00	1	2.0
29909	15209	"CONE" Waffle Cone	2018-04-19 07:28:18	$4.00	1	4.0
29910	16118	"ICE CREAM" Wildberry	2018-05-08 17:28:03	$1.50	1	1.5
29911	12568	"ICE CREAM" Maple Brown Sugar	2018-02-24 15:57:59	$2.00	2	4.0
29912	12842	"SORBET" Lemon	2018-03-02 09:19:02	$2.50	3	7.5
29913	19802	"ICE CREAM" Vanilla Bean	2018-07-24 19:09:38	$1.50	3	4.5
29914	19901	"ICE CREAM" Dark Chocolate	2018-07-26 20:46:29	$4.00	3	12.0
29915	17090	"ICE CREAM" Maple Brown Sugar	2018-05-29 10:04:32	$2.00	1	2.0
29916	12807	"BEVERAGE" Iced Coffee	2018-03-01 15:00:58	$2.50	3	7.5
29917	18452	"ICE CREAM" Dulce De Leche	2018-06-26 03:56:13	$-1.50	2	-3.0
29918	12889	"ICE CREAM" Dark Chocolate	2018-03-03 10:06:21	$4.00	3	12.0
29919	14526	"ICE CREAM" Peanut Fudge	2018-04-05 17:33:24	$3.50	3	10.5
29920	19589	"CONE" Dipped Waffle Cone	2018-07-20 09:10:01	$3.50	2	7.0
29921	19270	"ICE CREAM" Earl Gray	2018-07-13 09:20:21	$0.50	2	1.0

529 rows × 6 columns

In [58]:

```
df.isnull()
```

```
...
```

In [25]:

```
df[df['name'].isnull()].head()
```

Out[25]:

	order_id	name	ordered_at	price	quantity	line_total
3	10001	NaN	2018-01-01 12:14:54	$1.50	1	$1.50
6	10002	NaN	2018-01-01 12:23:09	$3.00	3	$9.00
27	10007	NaN	2018-01-01 15:03:17	$2.50	1	$2.50
77	10026	NaN	2018-01-02 03:25:40	$0.50	2	$1.00
88	10031	NaN	2018-01-02 05:45:48	$3.50	3	$10.50

Sanity check for value ranges and to check assumptions

In [26]:

```
df[(df['price'] * df['quantity']) != df['line_total']].shape[0]
```

Out[26]:

```
19924
```

In [67]:

```
df[df['line_total'] < 0].shape[0]
df.dtypes
```

Out[67]:

```
order_id                int64
name                   object
ordered_at     datetime64[ns]
price                 float64
quantity                int64
line_total            float64
dtype: object
```

TODO:

Set line_total = price * quantity if different Remove if line total < 0

In [71]:

```
df['line_total']=df['price']*df['quantity']
df
```

...

In [118]:

```
df.describe()
```

Out[118]:

	order_id	price	quantity	line_total
count	529.000000	529.00000	529.000000	529.000000
mean	14894.508507	2.42344	2.041588	4.935728
std	2884.990683	1.13372	0.824673	3.191731
min	10029.000000	-2.50000	1.000000	-7.500000
25%	12496.000000	1.50000	1.000000	2.500000
50%	14666.000000	2.50000	2.000000	4.000000
75%	17284.000000	3.50000	3.000000	7.500000
max	19996.000000	4.00000	3.000000	12.000000

TODO: Get value between "" in name and put it in category column

In [116]:

```
df['catogory'] = df['name'].apply(lambda x: re.search(r'["].+["]',str(x)).group(0) if  re.search(r'
```

In [117]:

```
df.head()
```

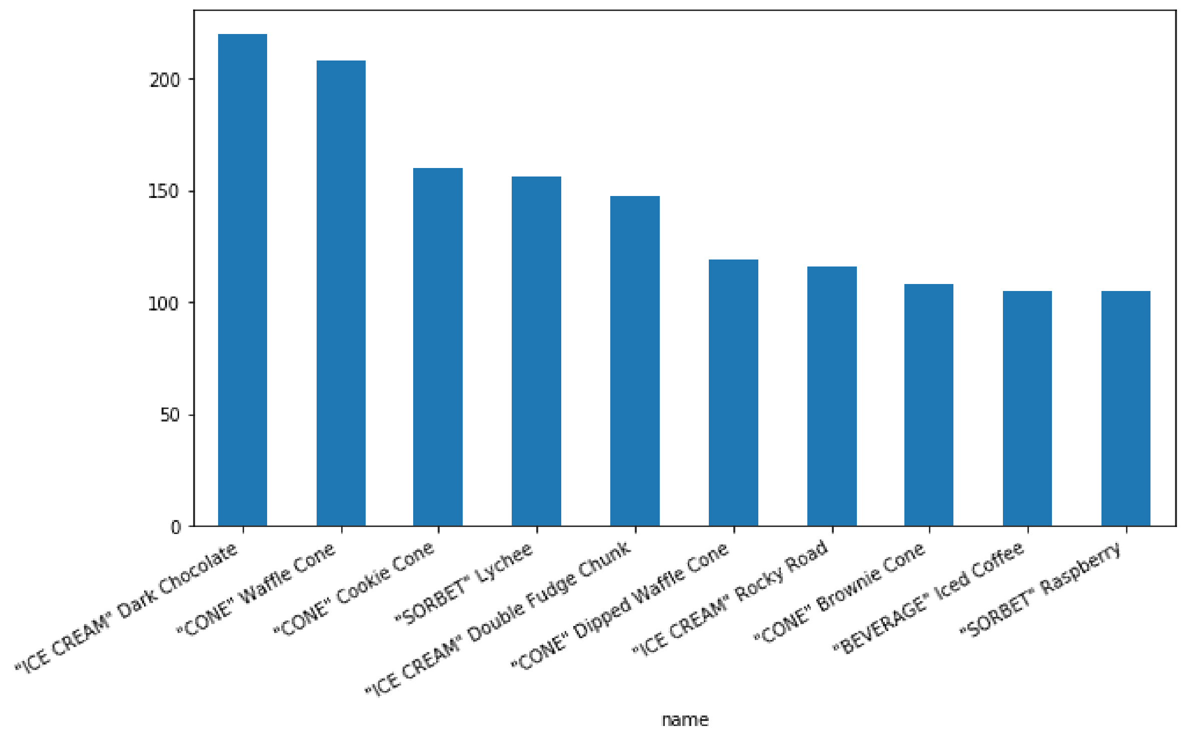
Out[117]:

	order_id	name	ordered_at	price	quantity	line_total	catogory
139	10049	"CONE" Dipped Waffle Cone	2018-01-02 13:47:55	3.5	1	3.5	"CONE"
178	10063	"ICE CREAM" Earl Gray	2018-01-02 21:28:36	0.5	3	1.5	"ICE CREAM"
207	10073	"ICE CREAM" Mint Chip	2018-01-03 01:45:00	1.5	3	4.5	"ICE CREAM"
273	10091	"SORBET" Blood Orange	2018-01-03 09:14:48	2.5	2	5.0	"SORBET"
278	10092	"CONE" Sugar Cone	2018-01-03 09:44:29	1.0	2	2.0	"CONE"

Analysis, finally!

In [121]:

```
f, ax = plt.subplots(figsize=(10, 6))
df.groupby('name')['line_total'].sum().sort_values(ascending=False).head(10).plot(kind='bar')
f.autofmt_xdate()
plt.show()
```


"ICE CREAM" Dark Chocolate
"CONE" Waffle Cone
"CONE" Cookie Cone
"SORBET" Lychee
"ICE CREAM" Double Fudge Chunk
"CONE" Dipped Waffle Cone
"ICE CREAM" Rocky Road
"CONE" Brownie Cone
"BEVERAGE" Iced Coffee
"SORBET" Raspberry
name