Jeff Slone 3.16.17

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

# run plots in the notebook
%matplotlib inline

url = "http://pbpython.com/extras/sample-salesv2.csv"

sales = pd.read_csv(url)
```

Remove spaces from column names

```
sales.columns = ['acct_num', 'name', 'sku', 'category', 'quantity', 'unit_price', 'ext_price', 'date']
sales.head()
```

	acct_num	name	sku	category	quantity	unit_price	ext_price	date
0	296809	Carroll PLC	QN-82852	Belt	13	44.48	578.24	2014-09-27 07:13:03
1	98022	Heidenreich-Bosco	MJ-21460	Shoes	19	53.62	1018.78	2014-07-29 02:10:44
2	563905	Kerluke, Reilly and Bechtelar	AS-93055	Shirt	12	24.16	289.92	2014-03-01 10:51:24
3	93356	Waters-Walker	AS-93055	Shirt	5	82.68	413.40	2013-11-17 20:41:11
4	659366	Waelchi-Fahey	AS-93055	Shirt	18	99.64	1793.52	2014-01-03 08:14:27

Subset the dataframe to contain only the name, category, quantity and unit price columns

```
subset_df = sales[['name', 'category', 'quantity', 'unit_price']]
subset_df.head()
```

	name	category	quantity	unit_price
0	Carroll PLC	Belt	13	44.48
1	Heidenreich-Bosco	Shoes	19	53.62
2	Kerluke, Reilly and Bechtelar	Shirt	12	24.16
3	Waters-Walker	Shirt	5	82.68
4	Waelchi-Fahey	Shirt	18	99.64

Subset the dataframe to contain only shirt sales

```
shirt_df = subset_df[subset_df['category'] == "Shirt"]
shirt_df.head()
```

	name	category	quantity	unit_price
2	Kerluke, Reilly and Bechtelar	Shirt	12	24.16
3	Waters-Walker	Shirt	5	82.68
4	Waelchi-Fahey	Shirt	18	99.64
5	Kerluke, Reilly and Bechtelar	Shirt	17	52.82
9	Kerluke, Reilly and Bechtelar	Shirt	12	26.98

Calculate the total cost per shirt sale

```
shirt_df['total_cost'] = shirt_df.quantity * shirt_df.unit_price
shirt_df.head()
```

```
/usr/local/lib/python2.7/dist-packages/ipykernel/__main__.py:1: 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-docs/stable/indexing.html#in
dexing-view-versus-copy
  if __name__ == '__main__':
```

	name	**category**	**quantity**	**unit_price**	**total_cost**
2	Kerluke, Reilly and Bechtelar	Shirt	12	24.16	289.92
3	Waters-Walker	Shirt	5	82.68	413.40
4	Waelchi-Fahey	Shirt	18	99.64	1793.52
5	Kerluke, Reilly and Bechtelar	Shirt	17	52.82	897.94
9	Kerluke, Reilly and Bechtelar	Shirt	12	26.98	323.76

Group the shirt sales by company name

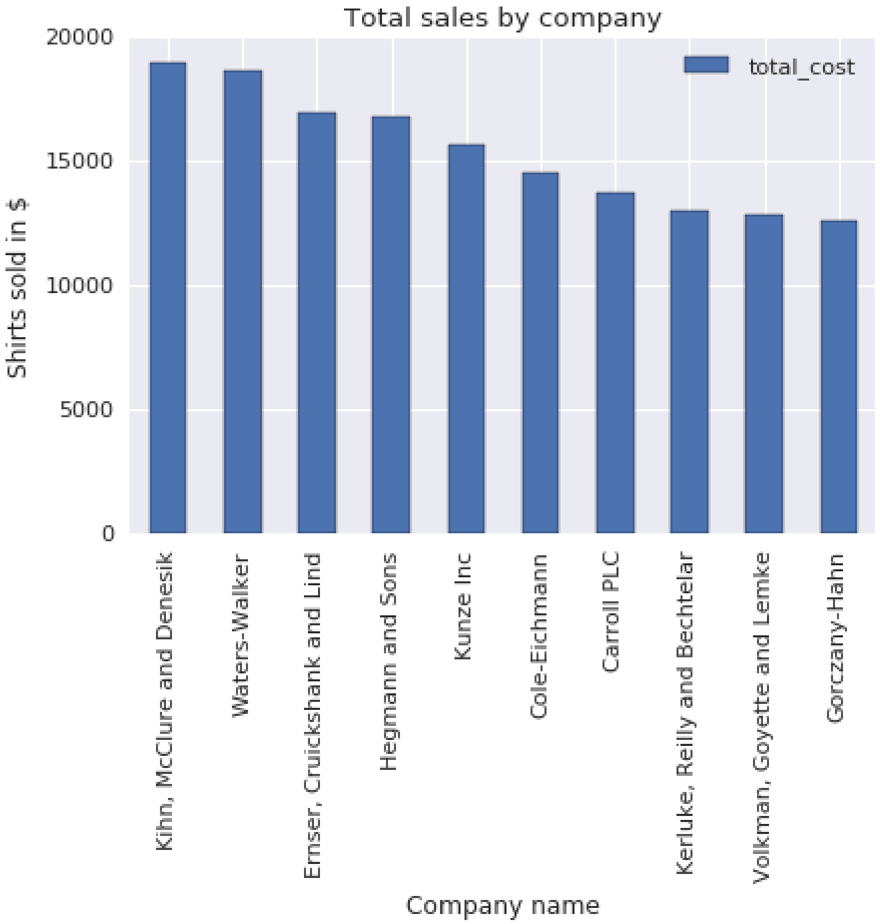
```
grouped = shirt_df.groupby('name', as_index=False).sum()
grouped
```

	name	quantity	unit_price	total_cost
0	Berge LLC	166	1226.54	9670.24
1	Carroll PLC	257	1098.93	13717.61
2	Cole-Eichmann	236	1226.75	14528.01
3	Davis, Kshlerin and Reilly	161	828.51	7533.03
4	Ernser, Cruickshank and Lind	262	1500.25	16944.19
5	Gorczany-Hahn	237	1132.22	12576.83
6	Hamill-Hackett	148	1091.55	8880.04
7	Hegmann and Sons	278	1528.84	16774.47
8	Heidenreich-Bosco	92	582.24	5965.25
9	Huel-Haag	200	1146.17	11944.01
10	Kerluke, Reilly and Bechtelar	269	1038.53	12958.23
11	Kihn, McClure and Denesik	288	1653.58	18956.35
12	Kilback-Gerlach	163	1052.53	9904.85
13	Koelpin PLC	132	786.07	7908.28
14	Kunze Inc	260	1439.92	15638.87
15	Kuphal, Zieme and Kub	252	1167.28	12101.14
16	Senger, Upton and Breitenberg	144	939.38	7659.70
17	Volkman, Goyette and Lemke	220	1136.25	12791.27
18	Waelchi-Fahey	201	1057.67	11689.05
19	Waters-Walker	288	1603.36	18633.71

Graph the top 10 shirt sales

```
top_10 = grouped.sort_values(by='total_cost', ascending=False).head(10)
top_10
top10_plot = top_10.plot(kind="bar",
                         title="Total sales by company",
                         x="name",
                         y="total_cost")
top10_plot.set_xlabel("Company name")
top10_plot.set_ylabel("Shirts sold in $")
```

```
<matplotlib.text.Text at 0x7fcf515c8d10>
```


Total sales by company
total_cost
Shirts sold in $
Kihn, McClure and Denesik
Waters-Walker
Ernser, Cruickshank and Lind
Hegmann and Sons
Kunze Inc
Cole-Eichmann
Carroll PLC
Kerluke, Reilly and Bechtelar
Volkman, Goyette and Lemke
Gorczany-Hahn
Company name