HW8

Homework 8¶

Fill out the following information (each category below should be on a separate line):¶

Name: Eric Adams¶

Date Submitted: Mar 5, 2017¶

Use the following data set to answer the questions for your homework:¶

In [2]:

```
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)
```

In [4]:

```
sales.head()
```

Out[4]:

	account number	name	sku	category	quantity	unit price	ext pric
0	296809	Carroll PLC	QN-82852	Belt	13	44.48	578.24
1	98022	Heidenreich-Bosco	MJ-21460	Shoes	19	53.62	1018.78
2	563905	Kerluke, Reilly and Bechtelar	AS-93055	Shirt	12	24.16	289.92
3	93356	Waters-Walker	AS-93055	Shirt	5	82.68	413.40

	account number	name	sku	category	quantity	unit price	ext pric
4	659366	Waelchi-Fahey	AS-93055	Shirt	18	99.64	1793.52

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

In [ ]:

```
# get rid of spaces in column names
```

In [4]:

```
sales.columns = ['acct_num', 'name', 'sku', 'category', 'quantity', 'unit_price', 'ext_price
```

In [ ]:

```
# let's see what this file has in it
```

In [5]:

```
sales.head()
```

Out[5]:

	acct_num	name	sku	category	quantity	unit_price	ext_price
0	296809	Carroll PLC	QN-82852	Belt	13	44.48	578.24
1	98022	Heidenreich-Bosco	MJ-21460	Shoes	19	53.62	1018.78
2	563905	Kerluke, Reilly and Bechtelar	AS-93055	Shirt	12	24.16	289.92
3	93356	Waters-Walker	AS-93055	Shirt	5	82.68	413.40
4	659366	Waelchi-Fahey	AS-93055	Shirt	18	99.64	1793.52

In [ ]:

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

In [6]:

```
df = sales[['name', 'category', 'quantity', 'unit_price']]
```

In [ ]:

```
# what does the subset look like
```

In [7]:

```
df.head()
```

Out[7]:

	name	category	quantity	unit_price
0	Carroll PLC	Belt	13	44.48

	name	category	quantity	unit_price
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¶

In [18]:

```
shirt_df = df[df['category'] == 'Shirt']
```

In [ ]:

```
# and shirt_df looks like...
```

In [19]:

```
shirt_df.head()
```

Out[19]:

	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

In [ ]:

```
# probably don't need to do the following two statements but did it just because
```

In [20]:

```
shirt_df.describe()
```

Out[20]:

	quantity	unit_price
count	404.000000	404.000000
mean	10.529703	57.516262
std	5.774214	25.203303
min	1.000000	10.380000
25%	6.000000	36.725000
50%	11.000000	60.130000
75%	16.000000	78.772500

	quantity	unit_price
max	20.000000	99.970000

In [21]:

```
shirt_df.dtypes
```

Out[21]:

```
name           object
category       object
quantity        int64
unit_price    float64
dtype: object
```

Calculate the total cost per shirt sale¶

In [22]:

```
shirt_df['shirt_sales'] = shirt_df.quantity * shirt_df.unit_price
```

In [ ]:

```
# does it look right? (yup)
```

In [28]:

```
shirt_df.head()
```

Out[28]:

	name	category	quantity	unit_price	shirt_sales
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¶

In [40]:

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

In [ ]:

```
# and the result looks good
```

In [41]:

```
shirts_by_company.head()
```

Out[41]:

	name	quantity	unit_price	shirt_sales
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

Graph the top 10 shirt sales¶

In [42]:

```
top_sellers = shirts_by_company.sort_values(by='shirt_sales', ascending=False).head(10)
```

In [46]:

```
top_sellers.head(10)
```

Out[46]:

	name	quantity	unit_price	shirt_sales
11	Kihn, McClure and Denesik	288	1653.58	18956.35
19	Waters-Walker	288	1603.36	18633.71
4	Ernser, Cruickshank and Lind	262	1500.25	16944.19
7	Hegmann and Sons	278	1528.84	16774.47
14	Kunze Inc	260	1439.92	15638.87
2	Cole-Eichmann	236	1226.75	14528.01
1	Carroll PLC	257	1098.93	13717.61
10	Kerluke, Reilly and Bechtelar	269	1038.53	12958.23
17	Volkman, Goyette and Lemke	220	1136.25	12791.27
5	Gorczany-Hahn	237	1132.22	12576.83

In [ ]:

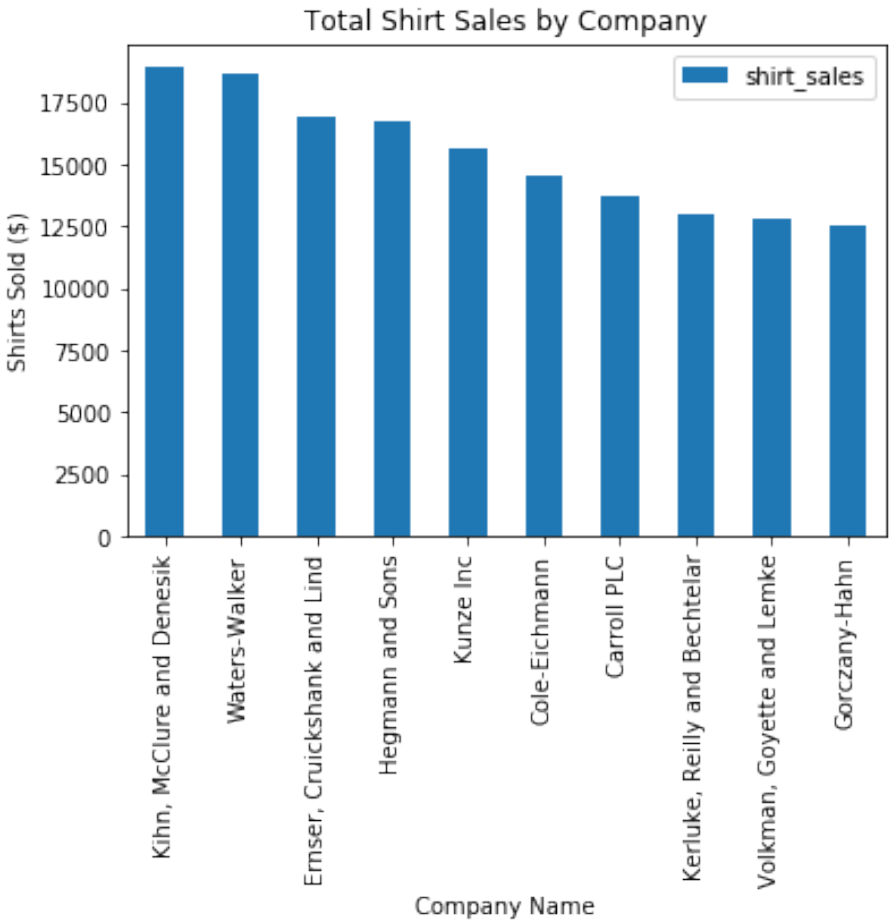
```
# using matplotlib
```

In [44]:

```
shirt_plot = top_sellers.plot(kind="bar",
                              title="Total Shirt Sales by Company",
                              x="name",
                              y="shirt_sales")
shirt_plot.set_xlabel("Company Name")
shirt_plot.set_ylabel("Shirts Sold ($)")
```

Out[44]:

```
<matplotlib.text.Text at 0x109972128>
```



To turn in your homework:¶

- Save this notebook as a PDF (or html if you can't get PDF output working)
- Upload the file to GitHub
- Provide the URL to this file on your GitHub repo in canvas

In [ ]: