Python for Librarians - Week 3 Workalong Answers

In case you need them, here are the completed answers for the Week 3 Workalong notebook.

In [ ]:
```
import pandas
#We `import as plt` just like we did with Pandas to make it so we have to type less
import matplotlib.pyplot as plt
%matplotlib inline
```

In [ ]:
```
sci_hub_data = pandas.read_csv("https://raw.githubusercontent.com/elibtronic/lja_datasets/master/week_3_sci_hub_worksheet.csv")
```

Question 1

Can you **describe** the dataframe?

In [ ]:
```
sci_hub_data.describe()
```

Question 2

How many unique publishers are in this dataset? (Hint: each unique prefix will indicate another Publisher in this data)

In [ ]:
```
# How mandy unique publishers are in the data?
sci_hub_data["doi_prefix"].nunique()
```

Question 3

How many unique users are in this dataset?

In [ ]:
```
# unique users
sci_hub_data["user_id"].nunique()
```

Question 4

See if you can now replicate this for the Montreal data in the next cell. You just need to modify lines 6 & 8

In [ ]:
```
montreal_top_publishers = sci_hub_data[sci_hub_data["city"] == "Montreal"].grou
pby("doi_prefix")\
                                .count()\
                                .sort_values(by='user_id',ascending=False)\
                                .head(10)["user_id"]

plt.pie(montreal_top_publishers, labels=montreal_top_publishers.index)

plt.title("Top Ten Publisher downloaded in Montreal")
plt.show()
```

Question 5

Let's try making a histogram for downloads per person in Montreal. Most of the code is written for you already. You just need to fill in lines: 7,9,12,13,16

In [ ]:
```
montreal_downloads_by_user = montreal_users.groupby('user_id')\
                                .count()\
                                .sort_values(by = "doi_whole", ascending=False)["do
i_whole"].head(100)


#how many different buckets we'll put values into (try experimenting with diffe
rent values)
bins = 25

plt.hist(montreal_downloads_by_user, bins)

#When we have a graph with X and Y axis we can add labels to them
plt.ylabel("Number of Users")
plt.xlabel("Number of Downloads")

#Descriptive title
plt.title("Downloads per user in Montreal")

#Generate our Graph
plt.show()
```