Data Analysis using Python-

What is Panda?

It’s an open source library in python for data analysis, data manipulation and data visualisation.

Python Makes it easier to work with data.

Its got a ton of functionality

Its well supported by the community

Its under active development

Its got a lot of documentation

Its plays well with other packages

Its built on top of numpy for numerical python, numerical computing.

It also works with scikit learn for machine learning

How do you get pandas?

How do I read a tabular data file into pandas?

Tabular data is the data with rows & columns like excel sheet. Comman format .CSV

```
import pandas as pd
```

```
x=pd.read_table('http://bit.ly/chiporders')
x.head()
```

```
m=pd.read_table('http://bit.ly/movieusers')
pd.read_table('http://bit.ly/movieusers',sep='|',header=None)
```

```
user_cols=['user_id','age','gender','occupation','Zip_code']
```

```
users=apd.read_table('http://bit.ly/movieusers',sep='|',header=None,
names=user_cols)
```

```
users.head()
```

How do I select a Panda series from a data frame?

As I might need some analysis or manipulation on a series

Each Column in a data frame is known as Panda Series.

```
ufo=pd.read_table('http://bit.ly/uforeports',sep=',')
```

or

```
ufo=pd.read_table('http://bit.ly/uforeports')
```

```
type(ufo)
```

```
ufo.head()
```

```
ufo['City']  ‘or’ ufo.City
```

```
type(ufo['City'])
```

```
ufo['Colors Reported']
```

Why do some pandas commands end with parentheses, and other command don’t?

```
import pandas as pd
movies=pd.read_csv('http://bit.ly/imdbratings')
movies.head()                                #method
movies.describe()
movies.shape                                 #attributes
movies.dtypes

movies.describe(include=['object'])          #object type data desc
```

How do I rename column in pandas dataframe?

```
ufo=pd.read_csv('http://bit.ly/uforeports')
ufo.head()
ufo.columns
ufo.rename(columns={'Colors Reported':'Colors_Reported','Shape Reported':'Shape_Reported'},inplace=True)
ufo.columns
```

‘or’

```
ufo_cols=['city','colors reported','shape reported','state','time']
ufo.columns=ufo_cols
ufo.head()
```

‘or’

```
ufo=pd.read_csv('http://bit.ly/uforeports',names=ufo_cols)
```

#Replace all spaces with underscore

```
ufo.columns=ufo.columns.str.replace(' ','_')
```

How Do I remove Columns from a pandas data frame?

```
import pandas as pd
ufo=pd.read_csv('http://bit.ly/uforeports')
ufo.head()

ufo.drop('Colors Reported', axis=1, inplace=True)          #columns
ufo.head()

ufo.drop(['City','State'], axis=1, inplace=True)
ufo.head()

ufo.drop([0,1], axis=0, inplace=True)                      #rows
ufo.head()
```

How to sort a Pandas DataFrame or Series?

```
movies=pd.read_csv('http://bit.ly/imdbratings')
movies.head()

movies.title.sort_values()
```

```
type(movies.title.sort_values())

movies.title.sort_values(ascending=False)

movies.sort_values('title')
movies.sort_values('duration')
movies.sort_values(['content_rating','duration'])
```

How do I filter rows of a pandas DataFrame by column value?

```
booleans=[]
for length in movies.duration:
    if length >= 200:
        booleans.append(True)
    else:
        booleans.append(False)

booleans[0:5]
len(booleans)
type(booleans)

is_long=pd.Series(booleans)
type(is_long)

is_long.head()
movies[is_long]
```

```
#replace for loop
is_long=movies.duration>=180
is_long.head()
movies[is_long]
```

```
movies[movies.duration>=175]
```

```
#to get any particular column for ex. genre
movies[movies.duration>=175].genre
‘or’
movies[movies.duration>=175][‘genre’]
‘or’
movies.loc[movies.duration>=175,'genre']
```

How do Apply multiple filter criteria to a pandas Data Frame?

```
True or False
True and False
movies[(movies.duration>=175) & (movies.genre=='Drama')]
movies[(movies.duration>=175) | (movies.genre=='Drama')]
```

```
#to replace multiple “or” condition
movies.genre.isin(['Crime','Drama','Action'])
movies[movies.genre.isin(['Crime','Drama','Action'])]
```

```
#To read data from particular columns only
Ufo=pd.read_csv(‘http://bit.ly/uforeports’,usecols[‘city’,’state’])
Ufo.columns
```

```
#fastest way to read from csv file
Ufo=pd.read_csv(‘http://bit.ly/uforeports’,nrows=3)
Ufo
```

```
#How do dataframes and series work with regard to selecting individual entries
and iteration(for X in userdata)?
for c in ufo.City:
        print(c)
```

```
#drop every non-numeric column from a data frame
Drinks=pd.read_csv(‘http:/bit.ly/drinksbycountry’)
Drinks.dtypes

Import numpy as np
Drinks.select_dtypes([include=np.number]).dtypes
```