ELITE DATA SCIENCE

PYTHON CHEATSHEET: DATA SCIENCE BASICS

In this cheat sheet, we summarize common and useful functionality from Pandas, NumPy, and Scikit-Learn. To see the most up-to-date full version, visit the online cheatsheet at elitedatascience.com.

SETUP

First, make sure you have the following installed on your computer:

- Python 2.7+ or Python 3
- Pandas
- Jupyter Notebook (optional, but recommended)

*note: We strongly recommend installing the Anaconda Distribution, which comes with all of those packages.

IMPORTING DATA

```
pd.read_csv(filename)
pd.read_table(filename)
pd.read_excel(filename)
pd.read_sql(query, connection_object)
pd.read_json(json_string)
pd.read_html(url)
pd.read_clipboard()
pd.DataFrame(dict)
```

EXPLORING DATA

```
df.shape()
df.head(n)
df.tail(n)
df.info()
df.describe()
s.value_counts(dropna=False)
df.apply(pd.Series.value_counts)
df.describe()
df.mean()
df.corr()
df.count()
df.max()
df.min()
df.median()
df.std()
```

SELECTING

```
df[col]
df[[col1, col2]]
s.iloc[0]
s.loc[0]
df.iloc[0,:]
df.iloc[0,0]
```

DATA CLEANING

```
df.columns = ['a','b','c']
pd.isnull()
pd.notnull()
df.dropna()
df.dropna(axis=1)
df.dropna(axis=1,thresh=n)
df.fillna(x)
s.fillna(s.mean())
s.astype(float)
s.replace(1,'one')
s.replace([1,3],['one','three'])
df.rename(columns=lambda x: x + 1)
df.rename(columns={'old_name': 'new_ name'})
df.set_index('column_one')
df.rename(index=lambda x: x + 1)
```

FILTER, SORT AND GROUP BY

```
df[df[col] > 0.5]
df[(df[col] > 0.5) & (df[col] < 0.7)]
df.sort_values(col1)
df.sort_values(col2,ascending=False)
df.sort_values([col1,col2], ascending=[True,False])
df.groupby(col)
df.groupby([col1,col2])
df.groupby(col1)[col2].mean()
df.pivot_table(index=col1, values= col2,col3], aggfunc=mean)
df.groupby(col1).agg(np.mean)
df.apply(np.mean)
df.apply(np.max, axis=1)
```

JOINING AND COMBINING

```
df1.append(df2)
pd.concat([df1, df2],axis=1)
df1.join(df2,on=col1,how='inner')
```

WRITING DATA

```
df.to_csv(filename)
df.to_excel(filename)
df.to_sql(table_name, connection_object)
df.to_json(filename)
df.to_html(filename)
df.to_clipboard()
```