Install Anaconda (Python, Jupyter) and data (DC workshop titanic.csv or larger titanic.xls from email)

Anaconda -> Jupyter -> Browser -> DC workshop 4 or location from email

Note: ctl+enter to execute, ctl+v to paste, watch three types of single quote … logout and x

.ipynb file remembers commands but need to rerun

ls

import pandas as pd

ti = pd.read\_csv(‘titanic.csv’)

ti2 = read\_excel(“titanic.xls”) => note lowercase col names

==

ti.shape => (887, 8) or (1309, 14)

ti.columns => 8 or 14 names

ti.info() => range, columns, non-null and type

ti.describe() => stats for numerical

ti.head(10) => first 5 or specified

ti.isnull().sum()

ti.Age[33]=None and redo above (was 66.0) or ti2

type(ti) => dataframe

type(ti.iloc[0]) => series or [[0]] dataframe e.g. [[0,1]]

==

ti.iloc[0:9] => first

ti[:3] v [::3]

ti.iloc[-1] => last 886 or 1308

ti.iloc[887] => index error

ti[ti.Age>62].sort\_values([‘Age’]) => 18 oldest: 16 male, only oldest survived

ti.where(ti.Age>62) => mask

==

ti.Age.duplicated().head() v ti.Age.head() => 5th

ti.Age.head() v ti.Age.rank().head() => value or position

ti.Age.mean() or median or mode or sum or count etc…

ti.Age.corr(ti.Survived) v Pclass.corr => -0.06, -0.34

ti.Pclass.unique() => [3,1,2]

ti[['Pclass', 'Survived']].groupby(['Pclass']).mean().sort\_values(by='Survived', ascending=False)

ti[['Sex', 'Survived']].groupby(['Sex']).mean().sort\_values(by='Survived', ascending=False)

ti[['Siblings/Spouses Aboard', 'Survived']].groupby(['Siblings/Spouses Aboard']).mean().sort\_values(by='Survived', ascending=False)

ti[[' Parents/Children Aboard', 'Survived']].groupby(['Parents/Children Aboard']).mean().sort\_values(by='Survived', ascending=False)

ti.Name.str.split().str[0].unique() => 17 titles

ti.Name.str.split().str[0].value\_counts() => frequency

ti[ti.Name.str.split().str[0]=='the'] => the Countess of …

tot=ti2.isnull().sum().sort\_values(ascending=False)

perc=(round(ti2.isnull().sum()/ti2.isnull().count()\*100,1)).sort\_values(ascending=False)

md=pd.concat([tot,perc], axis=1, keys=['Total','%'])

md => 7 x mv% and 7 x 0% … fill, drop or tricky cleaning

==

ti.hist() v boxplot() v plot()

%matplotlib inline

import matplotlib.pyplot as plt

plt.bar(x=range(3),height=ti.groupby('Pclass')['Survived'].mean())

ti2.groupby('embarked').mean()

plt.bar([‘S’,’C’,’Q’],ti2['embarked'].value\_counts())