Pandas

In this exercise will you be introduced to a lot of different pandas dataframe function which can help you in your work with data.

Exercise 1

Read the Titanic.xlsx file into a dataframe called titanic. Use the read_excel function.

Exercise 2

Examine the data.
Use the following two methods to examine the data:
`titanic.head()`
`titanic[:5]`

Exercise 3

Example the last part of the data
Use the following two methods to examine the data:
`titanic.tail()`
`titanic[-5:]`

Exercise 4

Calculate the mean on the Age column using the mean function

Exercise 5

Calculate the mean of the first five rows of the Age column.

Exercise 6

There is a lot of NaN in the Age column. Try to use the isnull method to find the number of rows which contain a NaN. Use the size property to get the number. The correct number is 557.

Tip: You can use the fillna method to fill-in some other value than NaN.

Exercise 7

Calculate mean of Age column for females and males, and dead and alive.

Results:
Male mean age: 31.01
Female mean age: 29.40

Dead mean age: 31.13
Survived mean age: 29.36

Exercise 8

You can use std() calculate standard deviation.
Calculate the std on the Age column. It should be 14.26.

Exercise 9

Create a sample of 100 passengers. Count surviving and not-surviving. Use the groupby method to group by survived.
Example:

```
titanic.sample(100).groupby('Survived').size()
Survived
0 68
1 32
dtype: int64
```

Your result will differ.

Exercise 10

Count the surviving numbers based on sex. Use goupby.
Result:

```
SexCode Survived
0 0 709
1 142
1 0 154
1 308
dtype: int64
```

Exercise 11

Count the number of surviving children (Age less than 18)

Result: 348

Exercise 12

Create a new column which will hold the surname of the passengers. Fill the column with the surname of each passenger.
You can iterate through the dataframe using for loop, or you can use the str.split function like this:

```
titanic['Surname'] = titanic['Name'].str.split(',',expand=True)[0]
```

This split the Name column and expand such that it is a dataframe that is returned.

Get the count of each name. Use the value_counts method. Get the top 5 names.

The list should be this:

```
Sage 11
Andersson 10
Goodwin 8
Asplund 7
Panula 6
Name: Surname, dtype: int64
```

Exercise 13

Plot histogram over the titanic dataset. Use the hist() function.
If you use something else than spyder or jupyter you must execute the following code to get the plots rendered on the screen:

```
import matplotlib.pyplot as plt
plt.show()
```