Python_Class_Project

March 6, 2023

[1]:

```
import itertools
import requests
import html5lib
from bs4 import BeautifulSoup
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import scipy.stats as sps
import statsmodels.formula.api as sm
mpl.style.use('seaborn')
mpl.rcParams['font.family'] = 'serif'
%matplotlib inline
```

1 Part I: Simulation

1.1 Approximation of π using Monte Carlo method

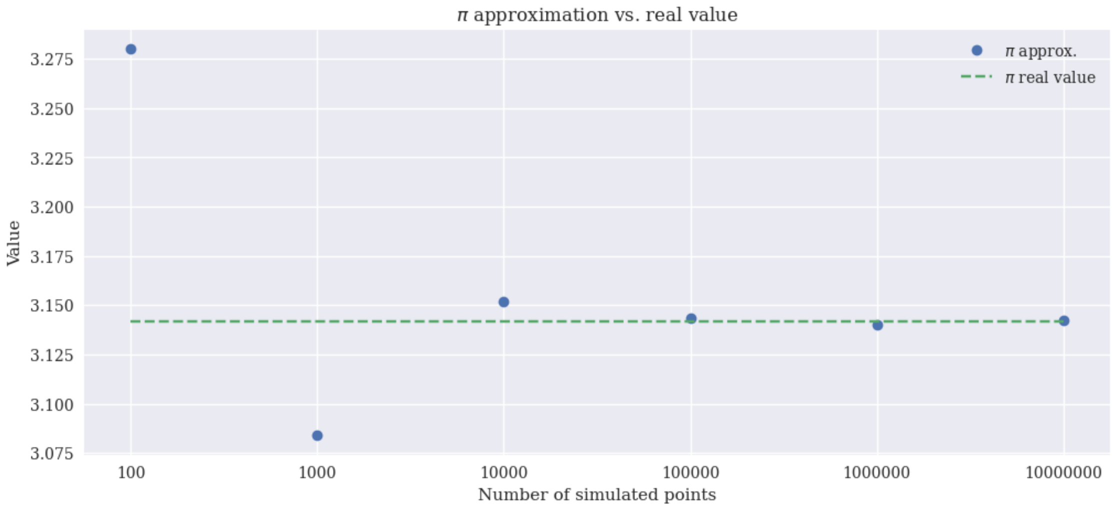
The goal is to approximate π by randomly drawing points on the square $[0, 1] \times [0, 1]$ with a uniform distribution, and then counting the proportion of these points falling at a Euclidean distance smaller than 1 from the point (0.0). This proportion must tend towards the area of the corresponding quarter of a circle, which corresponds to $\frac{\pi}{4}$.

1. Write a function taking an integer n as input, performing the procedure described above by simulating n points and returning the resulting approximate value of π.

2. Calculate approximate values of π by applying your function to $n = 10^k$ for $k \in \{2, 3, 4, 5, 6, 7\}$. Load these values in a container, then draw a figure comparing the approximations to the real value (as shown below).

[5]:

```
######### ######### ######### ############## ######### ######### #########
######### ######### ######### Your code here ######### ######### #########
######### ######### ######### ############## ######### ######### #########
```



1.2 Algorithm simulating Gaussian distribution

Here we are going to apply a classic method to generate i.i.d. standard normal random variables. Mathematically, it can be shown that if R follows exponential distribution of parameter $1\backslash 2$, and if θ is uniformly distributed on the interval $[0, 2\pi]$, then the variables

$$X = R\cos(\theta) \qquad \text{and} \qquad Y = R\sin(\theta)$$

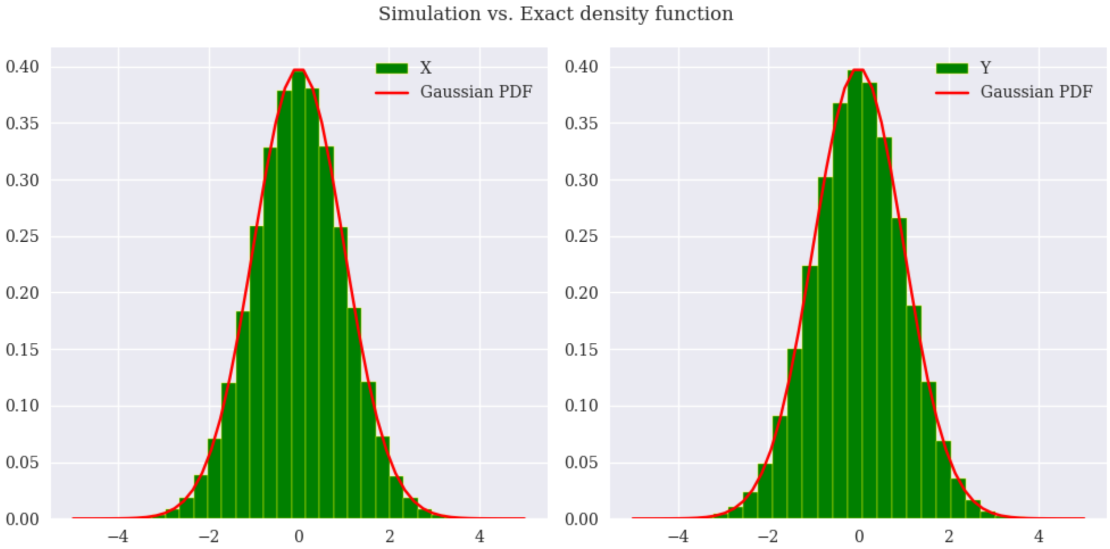
are independent and follow normal standard distribution.

3. Write a function taking an integer n as input, and returning X and Y containing n variables simulated according to the following algorithm:
   - Generate $U1 \sim Uniform(0, 1)$ and $U2 \sim Uniform(0, 1)$ where $U1 \perp U2$
   - Simulate R using inverse of the cumulative distribution method: $R = \sqrt{-2\log(U1)}$
   - Set $\theta = 2\pi U2$
   - Set $X = R\cos(\theta)$ and $Y = R\sin(\theta)$

4. Simulate X and Y n=10^6 times. Draw a figure with two horizontal subplots, displaying a histogram with simulated values of X compared to actual Gaussian density function on the left, and a histogram with simulated values of Y compared to the Gaussian density function on the right. *Hint*: set the attribute `density` to `True` to normalize the histogram and make it comparable to the density.

[7]:

```
######### ######### ######### ############## ######### ######### #########
######### ######### ######### Your code here ######### ######### #########
######### ######### ######### ############## ######### ######### #########
```



2 Part II: Data Scraping, Analysis & Visualisation

2.1 Scraping:

In this section you are going to scrape the parisian hotel listings from the wesite www.booking.com, focusing on such characteristics as price, location, etc., and then process and visualise certain aspects associated with this data.

1. Scrape this page. Load the content into a BeautifulSoup object. Inspect the page, then extract all apartment listings.
2. How many hotel listings have you retrieved? Check the link to the second page. Find a way to modify the link in question 1 in order to access other pages, then write a program that loops over the listings' pages (say, first 20 pages), scrapes them, retrieves all the individual listings and stacks them in a list. You should get a list of about 600 hotel listings.

2.2 Extracting and saving information:

We will start with a couple of functions you will need to process the data scraped from the webpage.

3. Write a function `extract_first_number` that takes a string as input and extracts *all digits that appear before the first letter* in a form of an integer.
4. Write a function `extract_value_before_word` that takes a full string and its substring (word) as input and extracts *a sequence of digits that appears before the given word and after any preceding non-numeric characters* in a form of an integer.
5. Write a function `extract_distance` that takes a full string containing distance and units (m ou km, see examples) as input and extracts the value corresponding to the distance, returning in as a float in km.

[149]:
```
### Examples:
extract_distance('4.5 km from the center')
```

[149]:
```
4.5
```

[150]:
```
extract_distance('500 m from the center')
```

[150]:
```
0.5
```

6. Go through all the extracted hotel listings and load in lists the following information (separate list for each information type, within every list maintain the same order as that in which listings appear on the webpage):
   - Names: load them as strings.
   - Links: also as strings.
   - Districts: use the function `extract_first_number` to extract them from the location strings and load them as floats.
   - Distances from the center: use the function `extract_distance` to load them as floats.
   - Number of stars: integers from 0 to 5.
   - Ratings: load them as floats.
   - Prices per night: use the function `extract_value_before_word` to load them in a numerical format. Take the discounted value if there is a discount. Pay attention to the fact that the prices on the page are given for two nights.
   - Free cancellation: True if present and False otherwise.
   - Breakfast included: True if present and False otherwise.
7. Create a hotel listings dictionary containing labels for previously mentioned characteristics as keys and the corresponding lists as values. Using Pandas.DataFrame, write the extracted information in an excel file, every row corresponding to one hotel listing, and every column to a certain characteristic.

[139]:
```
hotel_listings_df.head()
```

[139]:
```
                               Name  \
0                Hôtel La Sanguine
1                Hôtel de Bordeaux
2           Hôtel Avenir Jonquière
3                       Hotel Anya
4  UCPA SPORT STATION HOSTEL PARIS

                                                Link  District  Distance  \
0  https://www.booking.com/hotel/fr/la-sanguine.e…       8.0       2.7
1  https://www.booking.com/hotel/fr/de-bordeaux-p…      10.0       2.0
2  https://www.booking.com/hotel/fr/avenir-jonqui…      17.0       4.5
3  https://www.booking.com/hotel/fr/anya.en-gb.ht…      11.0       2.2
4  https://www.booking.com/hotel/fr/ucpa-sport-st…      19.0       4.8

   Nb. stars  Rating  Price per night (euro)  Free cancellation  \
0          2     7.8                   156.5              False
```

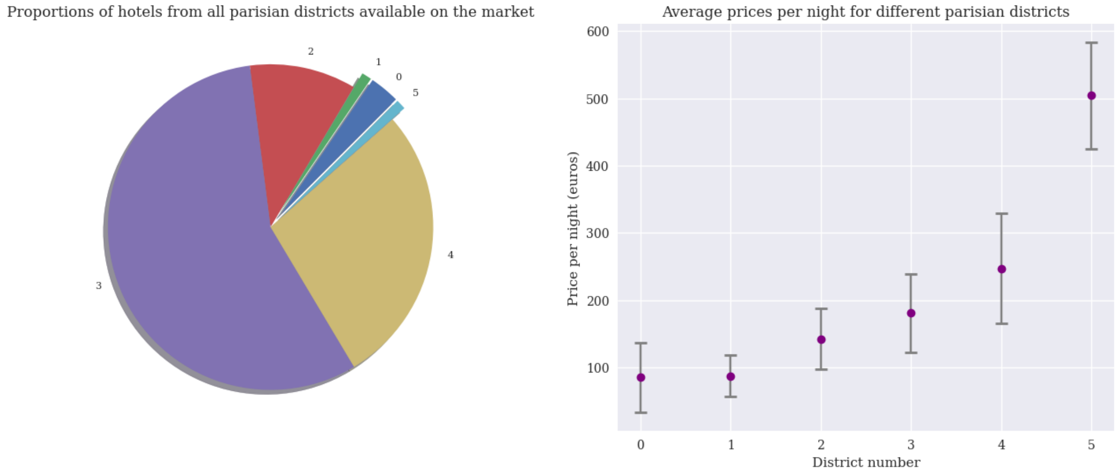
```
1           2     7.9                          121.0               False
2           1     7.9                           66.0               False
3           0     7.4                           85.0                True
4           0     8.0                           43.0                True

   Breakfast included
0               False
1               False
2               False
3               False
4               False
```

2.3 Paris districts & Prices per night

8. Make a plot with two horizontal subplots. Plot a pie chart showing proportions of hotel from every category with respect to the number of stars on the left, and a figure showing mean values of prices per night in each of these categories with standard deviation (you can use function `matplotlib.pyplot.errorbar`). Try to make it understandable: add axes labels, subplot titles, etc. Comment on any interesting trends that you may observe from the graphs.

[172]:

```
######### ######### ######### ############## ######### ######### #########
######### ######### ######### Your code here ######### ######### #########
######### ######### ######### ############## ######### ######### #########
```



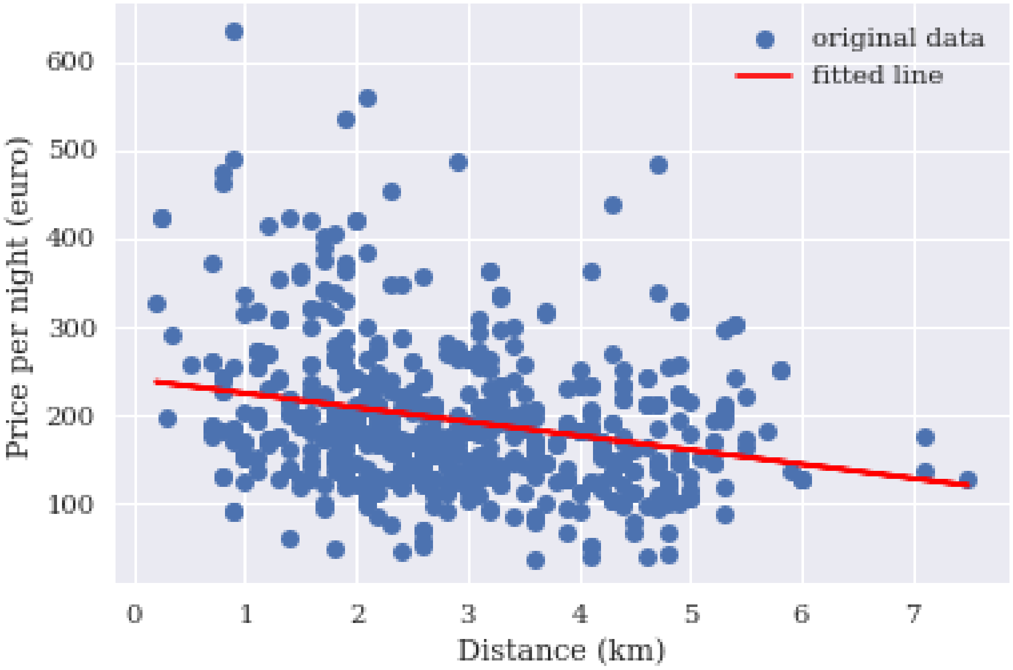
2.4 Outliers detection

9. Use the function `zscore` of the module `scipy.stats` in order to determine the outliers in the data with respect to prices per night. Calculate the z-score within every star category and identify the hotels with z-score larger than 3 (i.e. extraordinarily expensive hotels). Calculate the average rating for these hotels. How does it compare to the average rating of all hotels?

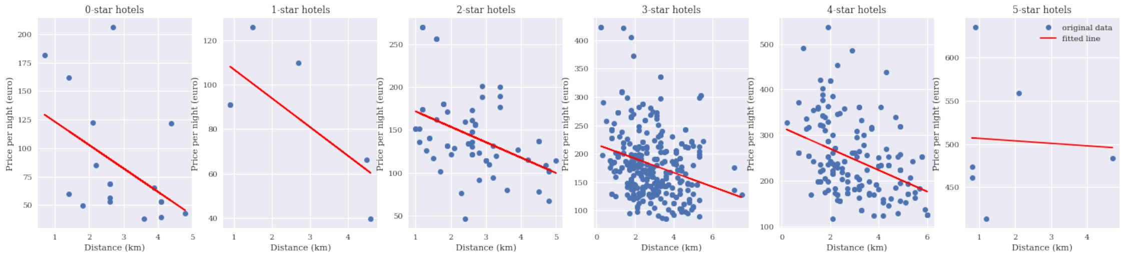
2.5 Regression

10. Using function `linregress` of module `scipy.stats` in order to calculate a linear least-squares regression predicting the prices (euro) based on the distance from center (km). Extract the slope and intercept, then plot the obtained regression fitted line along with the data points. Does linear model appear to represent well the dependency of prices on surface? Then do the same within each star category, plot all regressions on one figure using subplots. Compare the obtained results.

[68]:
```
######### ######### ######### ############## ######### ######### #########
######### ######### ######### Your code here ######### ######### #########
######### ######### ######### ############## ######### ######### #########
```



[96]:
```
######### ######### ######### ############## ######### ######### #########
######### ######### ######### Your code here ######### ######### #########
######### ######### ######### ############## ######### ######### #########
```



2.6 Some statistical tests

11. Perform a Student's t-test in order to determine whether there is a significant difference in prices per night depending on the presence or the absence of the free cancellation option by applying the function `ttest_ind` from the module `scipy.stats`. Do the same for the breakfast included option and comment.

12. Using a for loop over all parisian districts, perform a series of t-tests in order to determine whether there is a significant difference in prices per night between different districts. Print the obtained statistiscts.