**DECIPHERING DATA TO GET BUSINESS INSIGHTS**

**– USING R Vs PYTHON NUMPY LIBRARY**

Purpose of this post is to encourage people to learn more programming languages for Data analysis. By a sample analysis I have tried to demonstrate this point. Being already versed with R, my experience learning Numpy suggests that if a person is having knowledge on any of the language (R or Python) for Data Analysis then picking up concepts of the other become really easy.

If concepts about data and its containers are clear, then learning tool/language to analyze it just requires getting acquainted with keywords. And when it comes to which one to choose for data analysis, it is generally said that Python is better for data manipulation and repeated tasks, while R is good for ad hoc analysis and exploring datasets.

I have experience in doing data analysis using R and recently I started practicing Data analysis using Python. Following three libraries in Python are most useful for analyzing data.

* **Numpy**
* **Pandas**
* **MatPlotLib**

In this post, I will demonstrate Data analysis on a sample Data File to answer some of the business questions using Numpy Library for Python vs R.

Numpy helps in data analysis using Numpy arrays. It is the fundamental package for scientific computing with Python. A good understanding of Numpy will help using tools like Pandas effectively.

The dataset I have chosen to demonstrate this adhoc data analysis is from:

<http://insideairbnb.com/get-the-data.html>



* **FILE FORMAT**: CSV FILE
* **NO. OF RECORDS**: 7789
* **NO. OF ATTRIBUTES**: 14

This dataset contains data about various AirBnBs in Berlin:

* room types they offer
* total reviews received by each
* no. of bedrooms each one offers
* no. of people each one can accommodate
* minimum no. days required for stay and most importantly
* price for each

**Business questions we are trying to answers by our data analysis:**

1. **Total number of AirBnb in Berlin** – How many AirBnBs are there in Berlin?
2. **Different Room Types available** – What are the different types room type each one offers
3. **Areas where AirBnb is available** – Different areas of Berlin where these AirBnbs are located
4. **AirBnb with highest number of reviews** – AirBnb which has received most reviews
5. **List of AirBnBs which can accommodate maximum number of people**
6. **Costliest AirBnb** – AirBnb which is most expensive
7. **Top 10 Area in Berlin where AirBnb cost more**

**ANSWERING THE ABOVE QUESTIONS USING R**

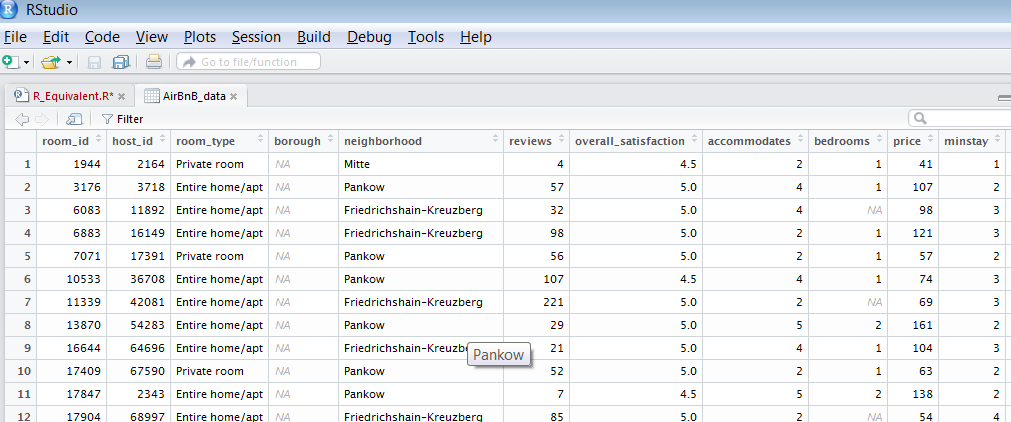
**<Berlin\_AirBnB.csv is placed in my local system>**

***#Setting directory path to the location where file has been placed using setwd() –set working directory command***

setwd("C:\\XXXX\\YYYY\\ZZZZ\\AAAA")

#Import CSV file data into an array/matrix by reading the CSV file

AirBnB\_data <- read.csv(file="Berlin\_AirBnB.csv", header=FALSE, sep=",")



Now that we have data imported from the CSV, we can start answering our business questions:

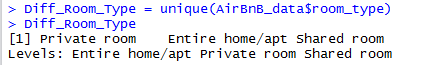
**#Total number of AirBnb in Berlin**

Total\_AirBnb = length (unique(AirBnB\_data$room\_id))



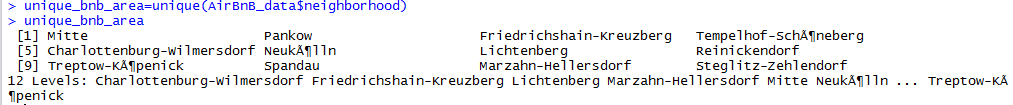
**#Different Room Types available**

Diff\_Room\_Type = unique(AirBnB\_data$room\_type)



**#Areas where AirBnb is available**

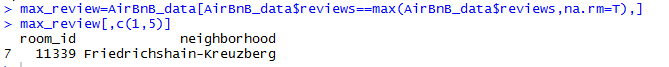
unique\_bnb\_area=unique(AirBnB\_data$neighborhood)



**# AirBnb with highest number of reviews**

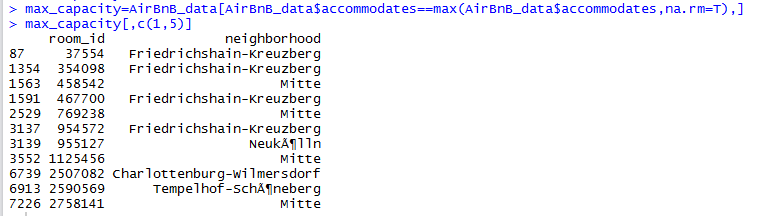
max\_review=AirBnB\_data[AirBnB\_data$reviews==max(AirBnB\_data$reviews,na.rm=T),]

max\_review[,c(1,5)]



**# List of AirBnBs which can accommodate maximum number of people**

max\_capacity=AirBnB\_data[AirBnB\_data$accommodates==max(AirBnB\_data$accommodates,na.rm=T),]max\_capacity[,c(1,5)]



**#Costliest AirBnb**

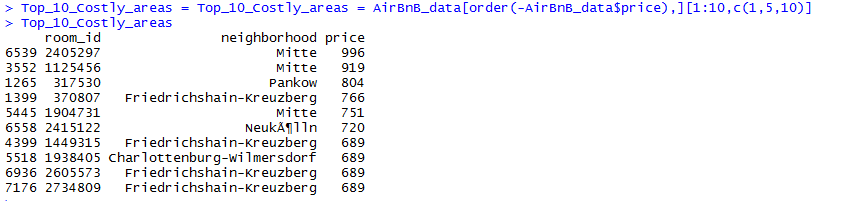
max\_price=AirBnB\_data[AirBnB\_data$price==max(AirBnB\_data$price,na.rm=T),]

max\_price[,c(1,5,10)]



**#Top 10 Area in Berlin where AirBnb cost more**

Top\_10\_Costly\_areas = AirBnB\_data[order(-AirBnB\_data$price),][1:10,c(1,5,10)]



**ANSWERING THE ABOVE QUESTIONS USING PYTHON-NUMPY LIBRARY**

**# Import Numpy as np**

import numpy as np

import os

**#Command to check current working directory**

os.getcwd()

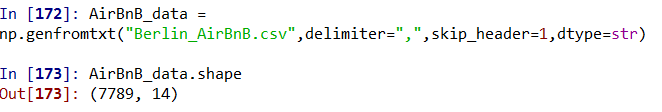
#Command to change working directory path

os.chdir('C:\\XXXX\YYYY\ZZZZ\AAAA')

**#Import CSV file data into an array using Numpy**

AirBnB\_data = np.genfromtxt("Berlin\_AirBnB.csv",delimiter=",",skip\_header=1,dtype=str)

AirBnB\_data.shape



**Now that we have data imported from the CSV, we can start answering our business questions:**

**#Total number of AirBnb in Berlin**

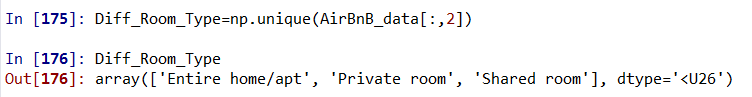
np.size(np.unique(AirBnB\_data[:,0].astype(int)))



**# Different Room Types available**

Diff\_Room\_Type=np.unique(AirBnB\_data[:,2])

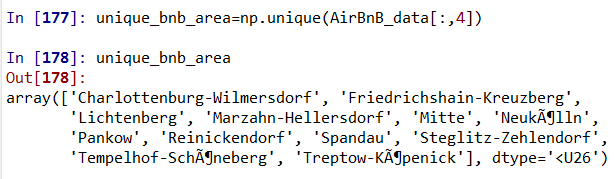
Diff\_Room\_Type



**# Areas where AirBnb is available**

unique\_bnb\_area=np.unique(AirBnB\_data[:,4])

unique\_bnb\_area



**# AirBnb with highest number of reviews**

max\_review=np.amax(AirBnB\_data[:,5].astype(int))

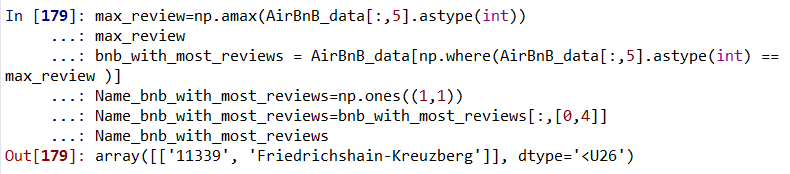
max\_review

bnb\_with\_most\_reviews = AirBnB\_data[np.where(AirBnB\_data[:,5].astype(int) == max\_review )]

Name\_bnb\_with\_most\_reviews=np.ones((1,1))

Name\_bnb\_with\_most\_reviews=bnb\_with\_most\_reviews[:,[0,4]]

Name\_bnb\_with\_most\_reviews



**# List of AirBnBs which can accommodate maximum number of people**

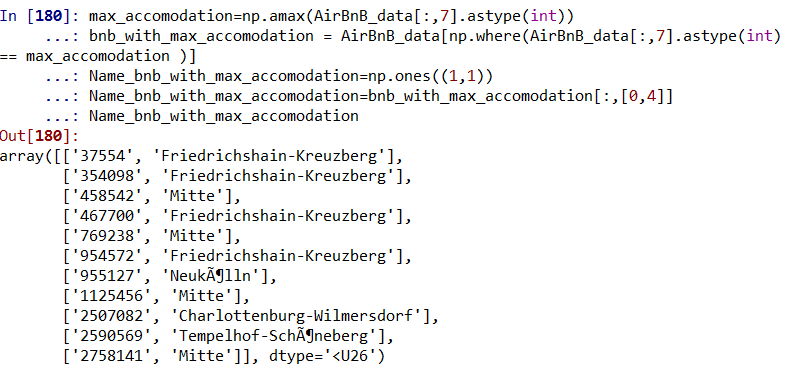
max\_accomodation=np.amax(AirBnB\_data[:,7].astype(int))

bnb\_with\_max\_accomodation = AirBnB\_data[np.where(AirBnB\_data[:,7].astype(int) == max\_accomodation )]

Name\_bnb\_with\_max\_accomodation=np.ones((1,1))

Name\_bnb\_with\_max\_accomodation=bnb\_with\_max\_accomodation[:,[0,4]]

Name\_bnb\_with\_max\_accomodation



**# Costliest AirBnB**

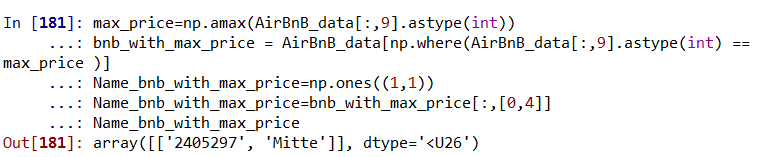
max\_price=np.amax(AirBnB\_data[:,9].astype(int))

bnb\_with\_max\_price = AirBnB\_data[np.where(AirBnB\_data[:,9].astype(int) == max\_price )]

Name\_bnb\_with\_max\_price=np.ones((1,1))

Name\_bnb\_with\_max\_price=bnb\_with\_max\_price[:,[0,4]]

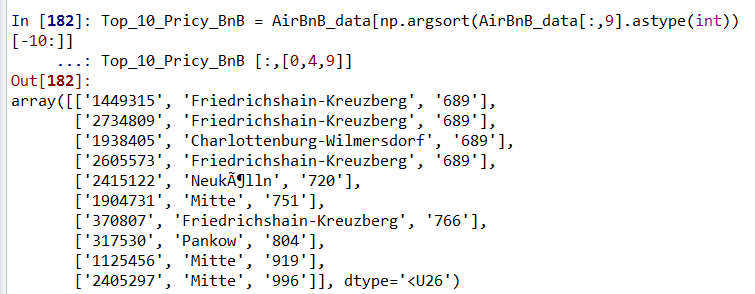
Name\_bnb\_with\_max\_price



**# Top 10 Area in Berlin where AirBnb cost more**

Top\_10\_Pricy\_BnB = AirBnB\_data[np.argsort(AirBnB\_data[:,9].astype(int))[-10:]]

Top\_10\_Pricy\_BnB [:,[0,4,9]]



Hope this practical example of doing data analysis to answer business critical questions using two different languages helps those who are new to the world of Data analytics.

In my next post, I will come up with some more Data analysis using Python-Pandas Library.

Happy Learning!