Data Mining
Home work 01

aqeel labash

13 February 2016

First Question

I read it :)

Second Question

The research contain the following data visualization:

1. Plot
2. Bar chart
3. Histogram
4. Stream graph
5. Treemap
6. Scatter Plot

Here is the explanation for what and how he used each one of them.

1. **Plot:** the most used one, he used to compare : Brooklyn Monthly Taxi Pickups, Uber vs Taxi pickups in Brooklyn,Manhattan monthly taxi pickups,etc...
2. **Bar chart :**Used it to compare the number of trips with different level of snow (other one with rain)
3. **Histogram:**used in the research to view (how many trips , time duration) to travel from 72nd & Broadway to Wall Street.
4. **Stream graph:** Used to view travel time from Midtown, Manhatten to different airports.It shows how long the trip take during day hours.It also clarify various probabilities for the trip duration.
5. **Treemap:** Used in the research to show the active areas in NYC during night hours (depending on the location and number of pickups)
6. **Scatter Plot:** am not pretty sure but I believe he used this way of visualization at the beginning of the research to draw the map of NYC depending on GPS for pickups and drop offs.

Third Question

Many business decisions could be built upon this data.Here is a list of what came to my mind:

1. Where to invest in what: decide in which places to invest depending on must drop off places.
2. Help taxis companies to well distribute there cars.
3. Identify main problems and bottle nicks for city mapping and suggest better organization over city buildings, parks , etc...

For Uber : I believe targeting the most crowded pickups areas with there advertisements (get more people from that area to be working with them).To predict a value I think is pretty hard from this data and nothing come to my mind that can lead to specific or round number.

Fourth Question

1. **Increasing the number** of rolling dice gradually drive the probabilities for all events to get closer to each other.

2. **Company Problem both discs defective:** the probability that Bob buy two defective is equal to

$$0.025 * 0.025 = 0.000625$$

(Probability First disc is defective * probability second disc is defective).

3. **Only one disc defective:** Here we have two cases lead to get one defective and one working. (d,c) + (c,d) (d mean defective , c mean correct).It's calculated as following:

$$Pr(x = d) * Pr(x = c) + Pr(x = c) + Pr(x = d) = 0.025 * 0.975 + 0.975 * 0.025 = 0.48$$

4. **Company Found 4 defective disks** for the company to find 4 defective discs it need to test for each one 1/0.025 = 40 disk has to be tested to get one defective (dividing the probability over one to get the dataset size).For total of 4 defective we need 40*4 = 160 which is the expected number of disks to be tested to get 4 defective disks.

Fifth Question

The probability of success is (Student prepared and succeed + student not prepared and succeed)

$$Pr(S \cap P) + Pr(S \cap NP)$$

where S = Succeed , NP = Not prepared , P = Prepared , Pr = Probability.
depending on conditional probability rules we can write

$$Pr(S \cap P) = Pr(P) * Pr(S|P) = 0.8 * 0.7 = 0.56$$

$$Pr(S \cap NP) = Pr(NP) * Pr(S|NP) = 0.2 * 0.4 = 0.08$$

$$Pr(S \cap P) + Pr(S \cap NP) = 0.56 + 0.08 = 0.64$$

Sixth Question

To simulate the **exercise 4** I wrote the following code :

```
import random
import numpy as np
def GetDisk():
  return random.randint(0,1000)
n=0
totaltries =0
lst = []
for i in range(0,100):
  while (n<4):
    if (GetDisk()<=25):
      n+=1
    totaltries+=1
    lst.append(totaltries)
  totaltries=0
  n=0
print lst
print np.mean(lst)

```

Function GetDisk get random integer between 1-1000.Which simulate picking disks randomly.If the value was under or equal 25 that means the disk is defective other wise it's working.
I repeated the operation for 100 times for each time I stopped when I get 4 defective disks and then calculated the average which was around 160. Values varied between (140-170).
The simulation for **exercise 5** in the following code :

```
#Simulation for PickSuccess Student
def PickSuccessStudent():
#Student is Prepared(0.8 probability)
if (random.randint(0,10)<=8):
#Student Succeed(0.7 probability)
if (random.randint(0,10)<=7):
return True
#Student faild (0.3 Probability)
else:
return False
#Student is not prepared (0.2 probability)
else:
#Student Succeed (0.4 probability)
if(random.randint(0,10)<=4):
return True
#Student failed (0.6 probability)
else:
return False
succeed=[]
for j in range (0,100):
tmp =0
for i in range (0,100):
if PickSuccessStudent():
tmp+=1
succeed.append(tmp)
print float(np.mean(succeed))/float(100)
```

The previous code showed that the probability is around (67-69) where function PickSuccessStudent() simulate if student succeed the exam or not.

Please Note: The code for this question is attached with the file under name **Q6.py**.

End of File