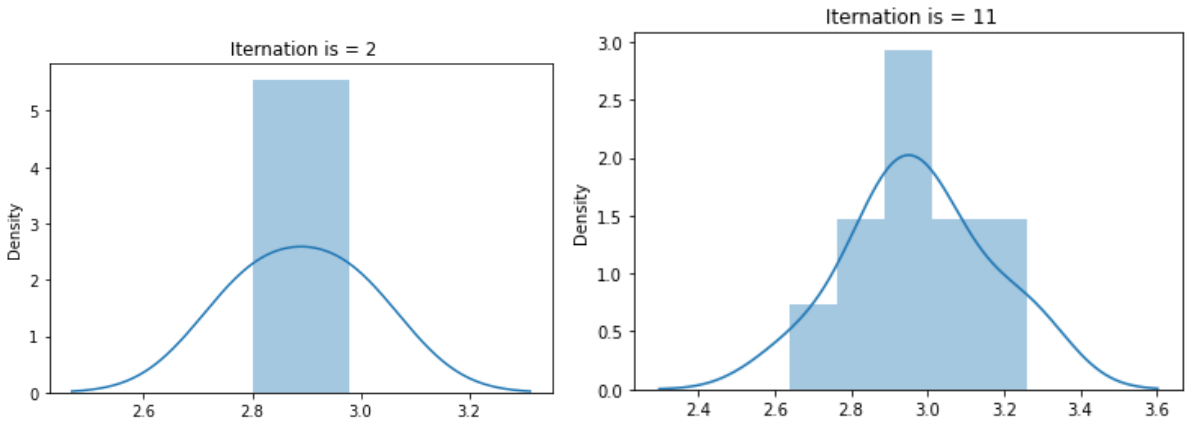
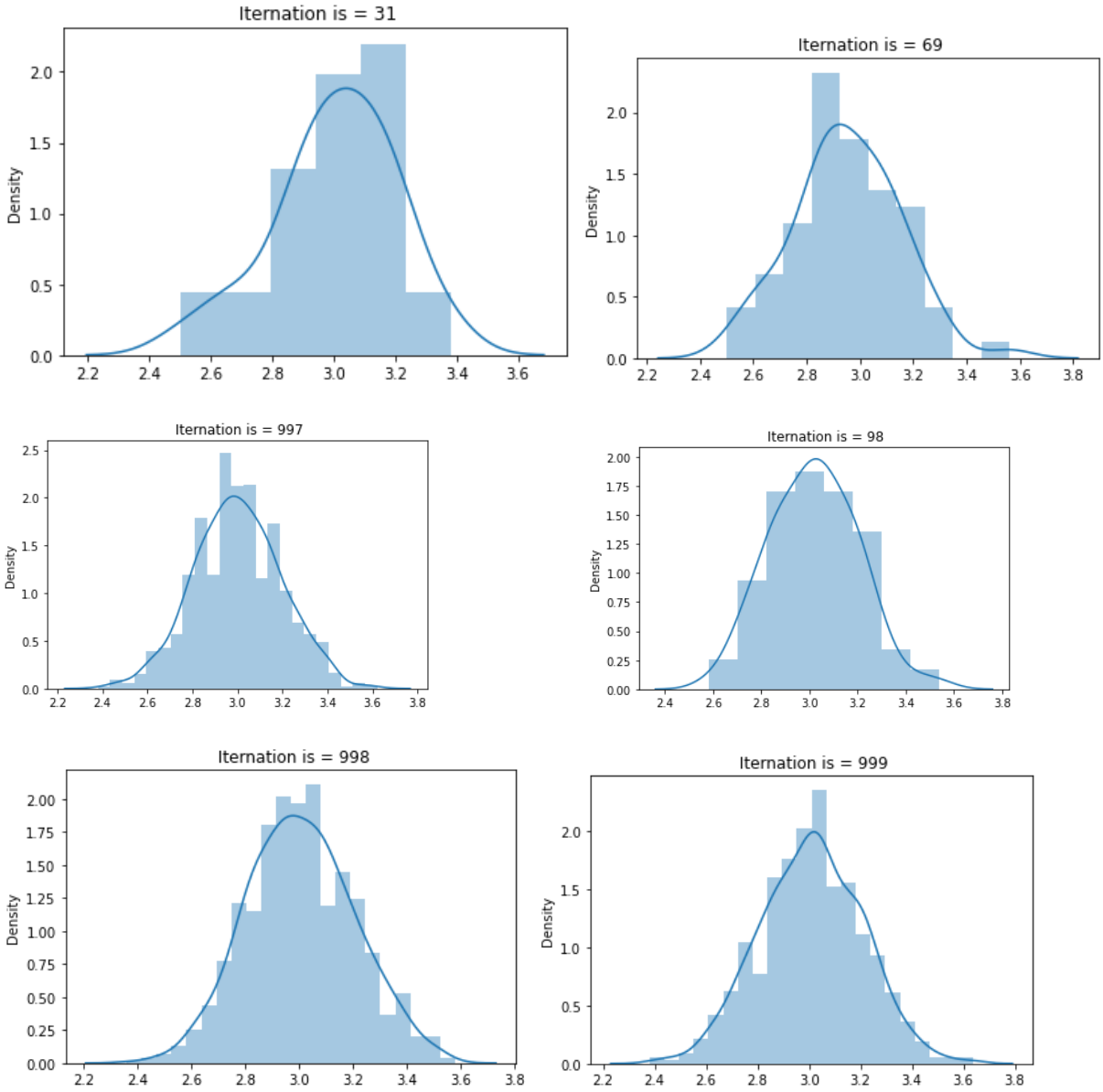
Q1: **Simulate the Central Limit Theorem in any programming language. "The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution".**

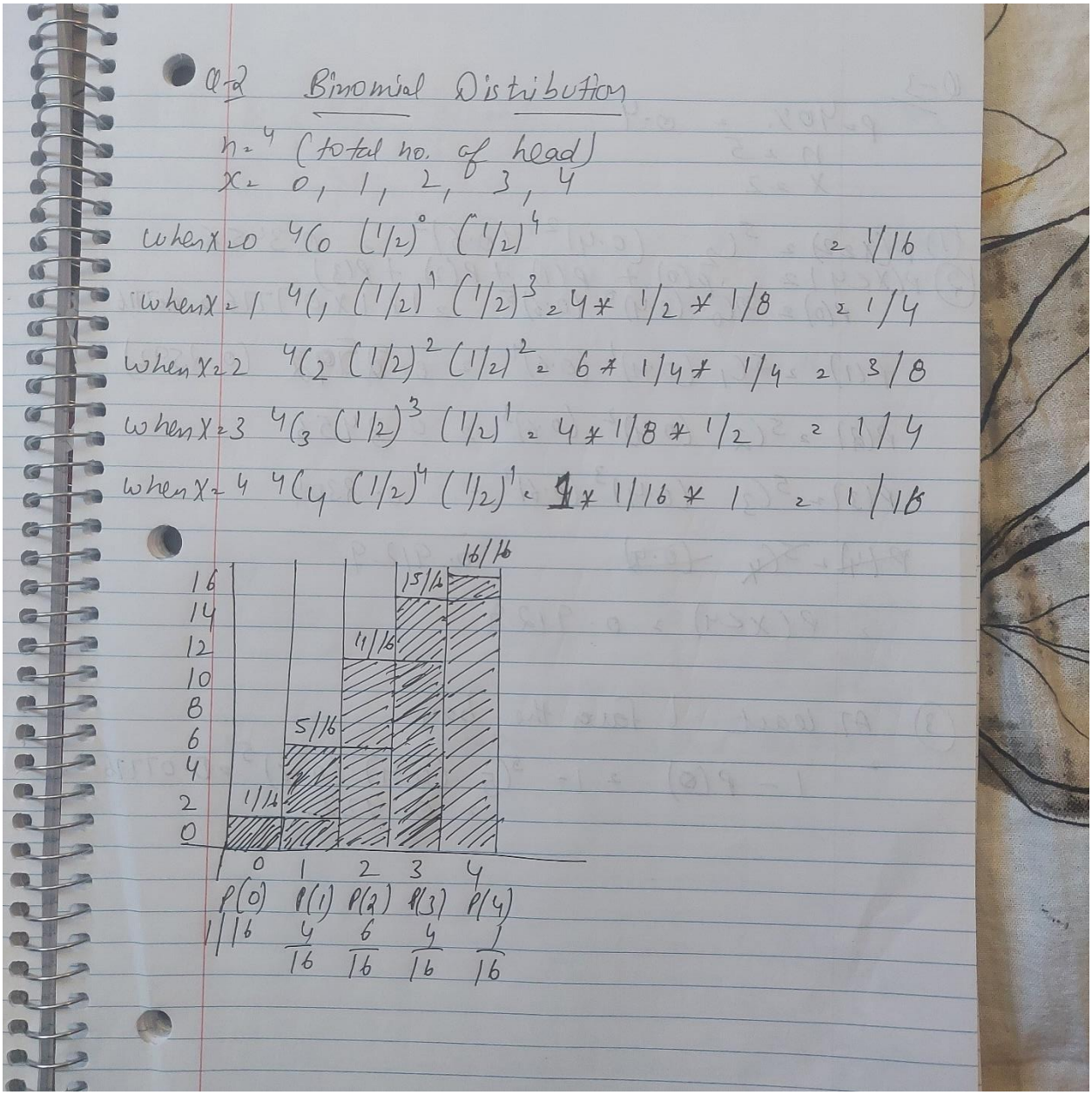
```
1
2 # demonstration of the central limit theorem
3 from numpy.random import seed
4 from numpy.random import randint
5 from numpy import mean
6 from matplotlib import pyplot
7 import seaborn as sbn
8 # seed the random number generator
9 seed(1)
10
11 # calculate the mean of 50 dice rolls 1000 times
12 #no_of_roles = input(int)
13 for i in range(0,1000):
14     no_of_role = i
15     total_no_of_dice = 50
16
17     #Observation: Result from one trial of an experiment.
18     #Sample: no. of dice role as per the rule
19     # population means total_no_of_dice
20     means = [mean(randint(1, 6, total_no_of_dice)) for _ in range(no_of_role)]
21     # plot the distribution of sample means
22     print()
23     means_w = mean(means)
24
25     print(str(means) + "  means  = "  + str(means_w))
26     sbn.distplot(means)
27     pyplot.title(" Iternation is = " + str(i))
28     pyplot.show()
```





In the above Figure, it is seen that the when we increase the number of experiments, the output of the experiment coverages as a normal distribution So, it is clear that as the sampling distribution become normal distribution of the sample as the sample size gets larger — no matter what the shape of the population distribution

Q2: Construct the binomial distribution for the total number of heads in four flips of a balanced coin. Define the PMF(Probability Mass Function) of the following distribution.

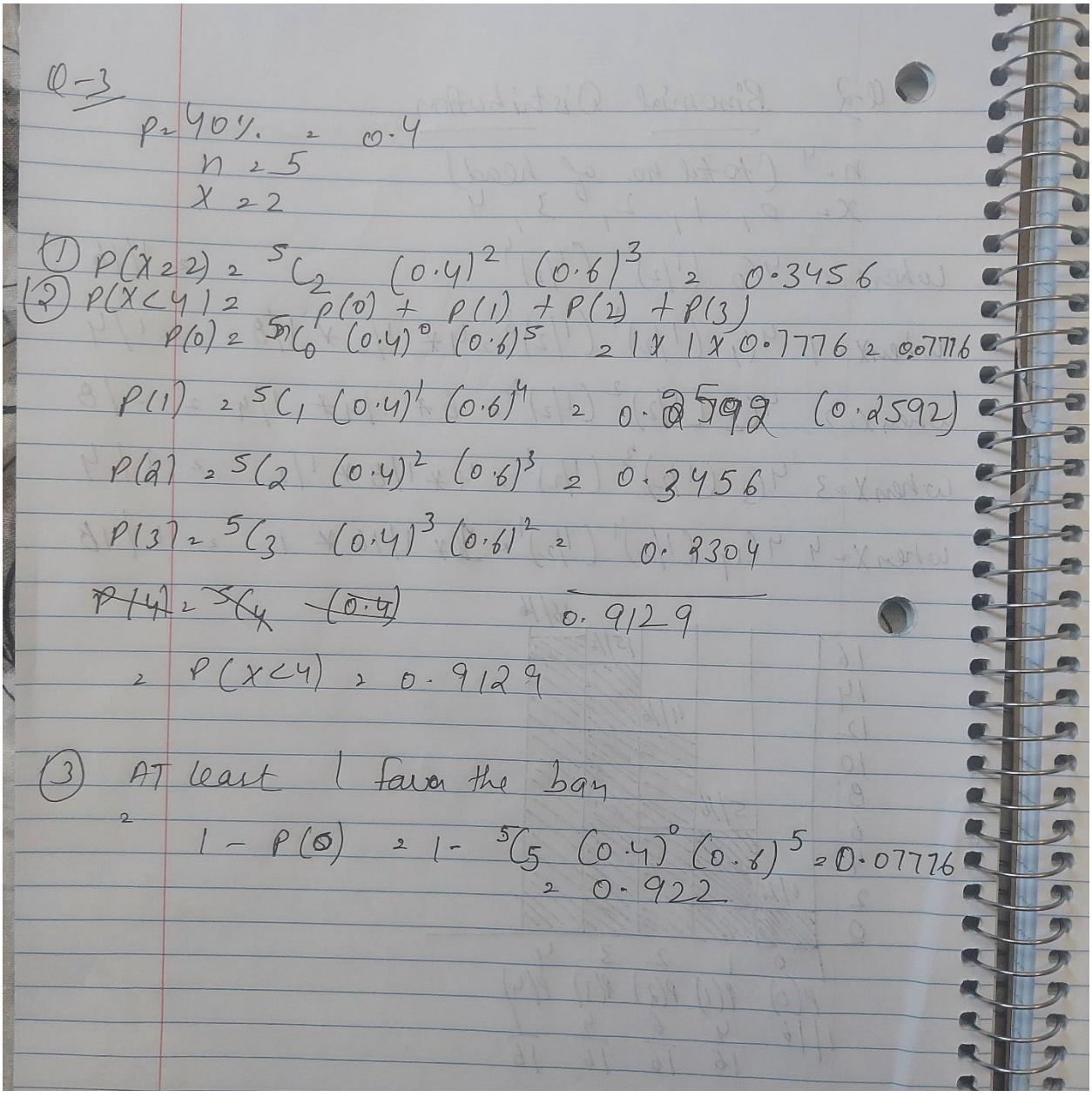


Q 3: - Suppose that 40% of the voters in a city are in favor of a ban of smoking in public buildings. Suppose 5 voters are to be randomly sampled. Find the probability that (10 points):

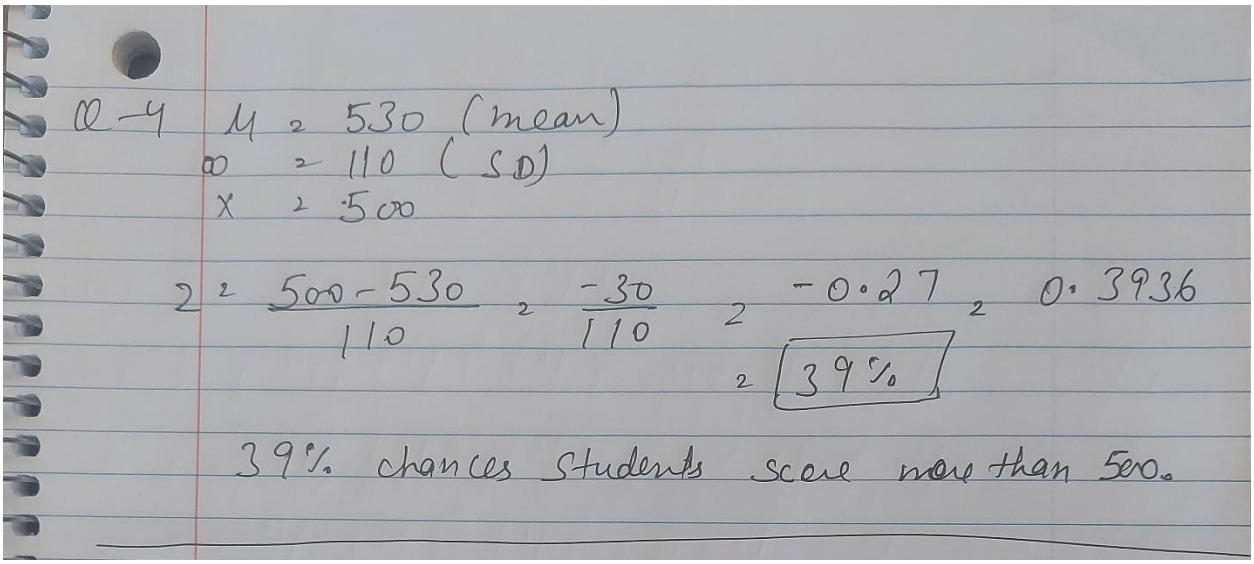
– 2 favor the ban.

– less than 4 favor the ban.

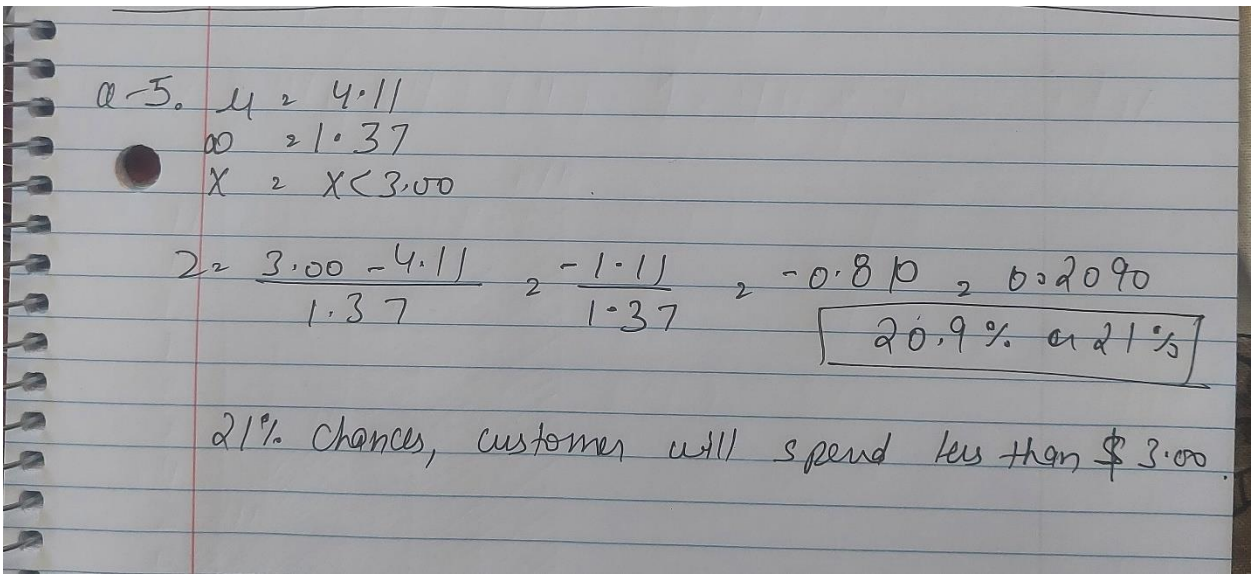
– at least 1 favor the ban.



Q4: Most graduate schools of business require applicants for admission to take the SAT examination. Scores on the SAT are roughly normally distributed with a mean of 530 and a standard deviation of 110. What is the probability of an individual scoring above 500 on the SAT?



Q5: The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of 4.11 dollar and a standard deviation of 1.37 dollar. What percentage of customers will spend less than 3.00 dollar on concessions?



Q6: A data scientist is testing a new model. She choose train and test sets at random from a large population of training data. She randomly choose 8 fold validation to get the accuracy for decision tree model, and choose 5 fold cross validation to get the accuracy for Logistic regression. The data are below: (25 points)

– Decision Trees: 93,94,89,88,78,89,76,98

– Logistic Regression: 78,90,89,76,89

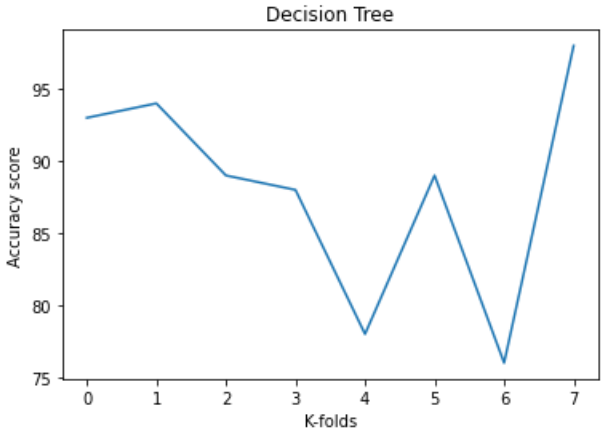
1. Are the two populations paired or independent? Explain your answer.

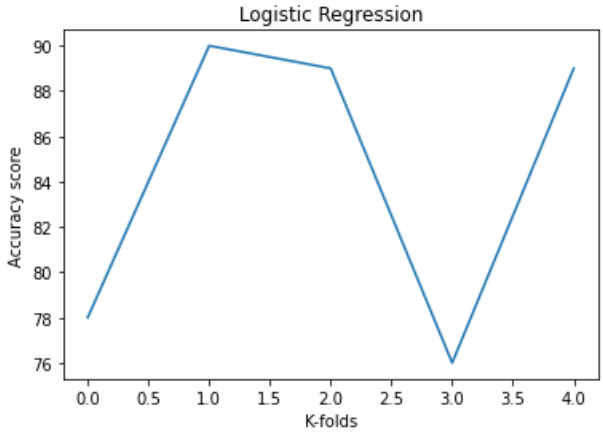
Both of the datasets are paired dataset. In layman terms: in Paired sample, the means of the data is compare within the same group at different times where the independent compared the means and standard deviations of two unrelated dataset of same length.

These datasets are unique in their own... So, we can implement the functions within each data pair instead of collaborative datasets. Thus, Decision Tree has its own pair and the Logistic Regression has its own pair.

2. Graph the data as you see fit. Why did you choose the graph(s) that you did and what does it (do they) tell you?

Both of the datasets are paired dataset.
In layman terms: in Paired sample, the means of the data is compare within the same group at different times where the independent compared the means and standard deviations of two unrelated dataset of same length.
These datasets are unique in their own. So, we can implement the functions within each data pair instead of collaborative datasets. Thus, Decision Tree has its own pair and the Logistic Regression has its own pair.





It is seen that the we have observations of the sequential k-folds cross validations.
These are the output or the accuracy score of a model that are changes with respect to number of K-cross validations.
These are the validated accuracy of a model. So, to observe the change in the accuracy score and patterns, the best fitted graph is the line chart.
It tell me, in case of DT: we have good accuracy in 0th, 1st, and 7th validations
It tell me, in case of LR: we have good accuracy in 1st and 4th Iterations where the 0th and 3rd has inaccurate training/testing score.

3. Choose a test appropriate for the hypothesis above, and justify your choice based on your answers to parts (a) and (b). Then perform the test by computing a p-value, and making a reject or not reject decision. Do use python or any programming language for this, and show your work. Finally, state your conclusion in the context of the problem.

```
print("\n Q3: Choose a test appropriate for the hypothesis above, \n and justify y
from scipy import stats
t_value,p_value=stats.ttest_ind(DT,LR)
print('Test statistic for the DT and LR is %f'%float("{:.6f}".format(t_value)))
print('p-value  of the acquried datasets is %f'%p_value)
alpha = 0.05
if p_value<=alpha:
    print('Conclusion','n','Since p-value(=%f)'%p_value,'<','alpha(=%.2f)'%alpha,'
else:
    print('Conclusion','\n','Since p-value(=%f)'%p_value,'>','alpha(=%.2f)'%alpha,
```

```
 Q3: Choose a test appropriate for the hypothesis above,
 and justify your choice based on your answers to parts (a) and (b).
Test statistic for the DT and LR is 0.890620
p-value  of the acquried datasets is 0.392196
Conclusion
 Since p-value(=0.392196) > alpha(=0.05)
 So, we do not reject the null hypothesis H0.
```

Colab URL: https://colab.research.google.com/drive/1_E6jiRDGIQ-Pgg3Q_e05rJuP3j5Q0ab_#scrollTo=2tXWqLVx3TUL