Part D) Are X and Y independent or dependent? Fully justify your answer in the cell below using LaTeX and the mathematical definition of independence.

For X and Y to be independent,

$P(X = 1, Y = 2) = P(X = 1)$

However, we calculated $P(X = 1, Y = 2) = \frac{1}{12}$ and $P(X = 1) = \frac{5}{12}$

Since $\frac{1}{12} \neq \frac{5}{12}$ X and Y are dependent.

Part A) If $\mathrm{Cov}(X, Y) = 0$, what does this tell us about the random variables X and Y?

If $\mathrm{Cov}(X, Y) = 0$, then X and Y do not have a linear relationship.

Part B) Given the following joint pmf for discrete random variables X and Y:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
$X=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

- i). Calculate $Cov(X,Y)$.
- ii). Calculate $\rho(X,Y)$

Show all steps for both parts using Markdown and LaTeX in the cell below:

i).

$Cov(X,Y) = E[XY] - E[X]E[Y]$

$E[XY] = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}$

$E[X] = \frac{1}{8} + \frac{1}{6} + \frac{1}{6} = \frac{11}{24}$

$E[Y] = \frac{1}{4} + \frac{1}{6} + \frac{2}{8} + \frac{2}{6} = 1$

$Cov(X,Y) = \frac{12}{24} - \frac{11}{24} = \frac{1}{24}$

ii).

$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

$\sigma_X^2 = E[X^2] - E[X]^2$

$E[X^2] = \frac{1}{8} + \frac{1}{6} + \frac{1}{6} = \frac{11}{24}$

$E[X]^2 = \frac{121}{576}$

$\sigma_X^2 = \frac{143}{576}$

$\sigma_X = \sqrt{\frac{143}{576}}$

$\sigma_Y^2 = E[Y^2] - E[Y]^2$

$E[Y^2] = \frac{1}{4} + \frac{1}{6} + \frac{4}{8} + \frac{4}{6} = \frac{38}{24}$

$E[Y]^2 = 1$

$\sigma_Y^2 = \frac{14}{24}$

$\sigma_Y = \sqrt{\frac{14}{24}}$

$\rho(X,Y) = \frac{1}{24}\sqrt{\frac{576}{143}}\sqrt{\frac{24}{14}} = \frac{1}{24}\sqrt{\frac{6912}{1001}} = 2\sqrt{\frac{3}{1001}}$

Part C) This part is **NOT** related to the parts above.
Suppose you're only given the following information about two joint random variables X and Y:

$$\mu_X = 6,\ \ \mu_Y = 5,\ \ \sigma_X^2 = 4,\ \ \sigma_Y^2 = 9 \text{ and } E[XY] = 27$$

.

For each of the quantities below, calculate if you have enough information, showing all steps. If not, explain what additional info you'd need.

i). $Cov(X, Y)$

ii). $Cov(Y, X)$

iii). $\rho(X, Y)$

Answer all parts in the ONE markdown cell below, fully justifying your answer:

i).

$Cov(X, Y) = E[XY] - E[X]E[Y]$

From above, we know all of those values and can plug them into the equation.

$Cov(X, Y) = 27 - 6 * 5 = -3$

ii).

$Cov(Y, X) = E[YX] - E[Y]E[X]$

Since $E[YX] = E[XY]$ we know that $Cov(Y, X) = Cov(X, Y)$

$Cov(Y, X) = 27 - 5 * 6 = -3$

iii).

$\rho(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Thus

$\sigma_X = \sqrt{4} = 2$ and $\sigma_Y = \sqrt{9} = 3$

So

$\rho(X,Y) = \frac{-3}{2*3} = \frac{-1}{2}$

Back to top

0.1 (2 pts) Problem 4

If we're trying to predict the results of the Clinton vs. Trump 2016 presidential race:

i). What is the population of interest?

ii). What is the sampling frame?

Give both of your answers in the same below in Markdown.

i). The population of interest would be all eligible voters in the United States.

ii). The sampling frame would be the list of 2016 registered voters.

Back to top

0.2 Problem 5 (11 pts)

Part A For your convenience, the actual results of the vote in the four pivotal states is repeated below:

State	% Trump	% Clinton	Total Voters
florida	49.02	47.82	9,419,886
michigan	47.50	47.27	4,799,284
pennsylvania	48.18	47.46	6,165,478
wisconsin	47.22	46.45	2,976,150

Using the table above, write a function `draw_state_sample(N, state)` that returns a sample with replacement of N voters from the given state, using the percentages given in the table above. Your result should be returned as a list, where the first element is the number of Trump votes, the second element is the number of Clinton votes, and the third is the number of Other votes. For example, `draw_state_sample(1500, "florida")` could return `[727, 692, 81]`. You may assume that the state name is given in all lower case.

Hint: You might find `np.random.multinomial` useful.

```
In [29]: def draw_state_sample(N, state):
             sample = 0
             if(state == 'florida'):
                 sample = np.random.multinomial(N, [0.4902, 0.4782, 0.0316])
             if(state == 'michigan'):
                 sample = np.random.multinomial(N, [0.475, 0.4727, 0.0523])
             if(state == 'pennsylvania'):
                 sample = np.random.multinomial(N, [0.4818, 0.4746, 0.0436])
             if(state == 'wisconsin'):
                 sample = np.random.multinomial(N, [0.4722, 0.4645, 0.0633])

             return sample;
```

```
In [30]: grader.check("q5a")
```

```
Out[30]: q5a results: All test cases passed!
```

Part D i). Make a **frequency** histogram of `simulations`. This is a histogram of the sampling distribution of Trump's proportion advantage in Pennsylvania.

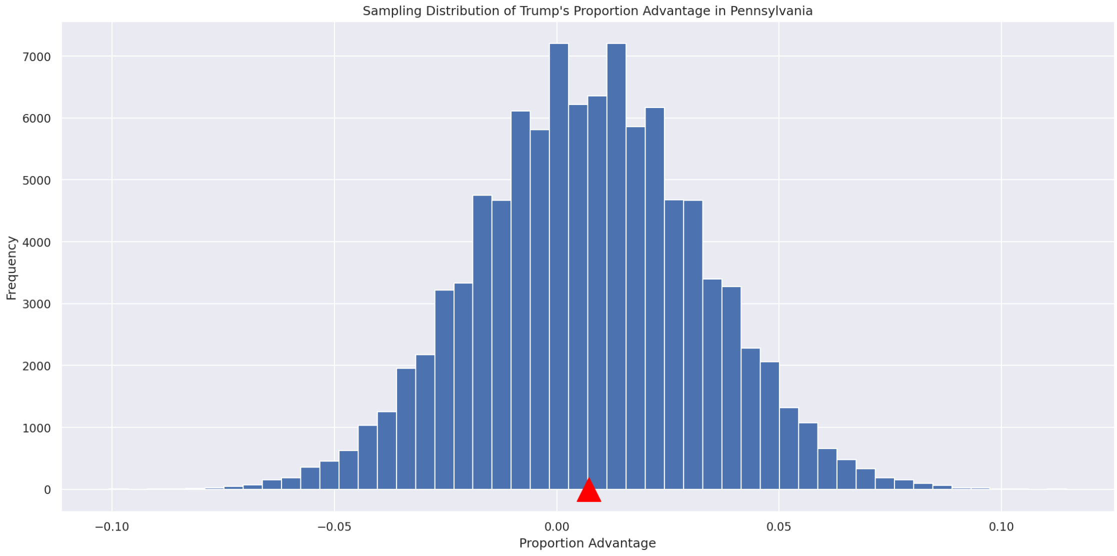
Hint: You should use the `plt.hist` function in your code.

Make sure to include a title as well as axis labels. You can do this using `plt.title`, `plt.xlabel`, and `plt.ylabel`.

ii). Based on your simulation, what is the probability that a random sample of 1500 will correctly predict that Trump wins Pennsylvania? (i.e. what proportion of these simulations predict a Trump victory?) Assign your answer to `prob_penn_1500_random_correct`

```
In [35]: # Part (i):
         plt.hist(simulations, bins=50)
         # your code for the histogram above here.  The code below plots a red marker at the mean:
         plt.scatter(simulations.mean(), -1, marker='^', color='red', s=500)
         plt.title("Sampling Distribution of Trump's Proportion Advantage in Pennsylvania")
         plt.ylabel("Frequency")
         plt.xlabel("Proportion Advantage")
```

```
Out[35]: Text(0.5, 0, 'Proportion Advantage')
```



```
In [36]: # Part (ii):
         prob_penn_1500_random_correct = np.mean(simulations > 0)
```

```
prob_penn_1500_random_correct
```

```
Out[36]: 0.60673
```

Back to top

0.3 Problem 6 (10 pts)

Throughout this problem, adjust the selection of voters so that there is a 0.5% bias in favor of Clinton in each of these states.

For example, in Pennsylvania, Clinton received 47.46% of the votes and Trump 48.18%. Increase the population of Clinton voters to 47.46% + 0.5% and correspondingly decrease the percent of Trump voters.

Part A Simulate Trump's advantage across 100,000 simple random samples of 1500 voters for the **state of Pennsylvania** and store the results of each simulation in an `np.array` called `biased_simulations`.

That is, `biased_simulation[i]` should hold the result of the `i+1`th simulation.

That is, your answer to this problem should be just like your answer from Question 5C, but now using samples that are biased as described above.

```
In [42]: def draw_biased_state_sample(N, state):
             sample = 0
             if(state == 'florida'):
                 sample = np.random.multinomial(N, [0.4852, 0.4832, 0.0316])
             if(state == 'michigan'):
                 sample = np.random.multinomial(N, [0.47, 0.4777, 0.0523])
             if(state == 'pennsylvania'):
                 sample = np.random.multinomial(N, [0.4768, 0.4796, 0.0436])
             if(state == 'wisconsin'):
                 sample = np.random.multinomial(N, [0.4672, 0.4695, 0.0633])

             return sample

         sim2 = np.array([])

         for i in range(100000):
             sample = draw_biased_state_sample(1500, 'pennsylvania')
             adv = trump_advantage(sample)
             sim2 = np.append(sim2, adv)

         biased_simulations = sim2
```

```
In [43]: grader.check("q6a")
```

```
Out[43]: q6a results: All test cases passed!
```

Part B Create a plot of **overlaid DENSITY** histograms of the following: - The new sampling distribution of Trump's proportion advantage in Pennsylvania using these biased samples - The sampling distribution of the unbiased samples from Problem 5D (plotted as a density, not a frequency histogram)

Include 2 markers (of different colors) with the sample means for each distribution (see 5D for code how to do this). The colors of the markers should correspond to the colors of the density histograms.

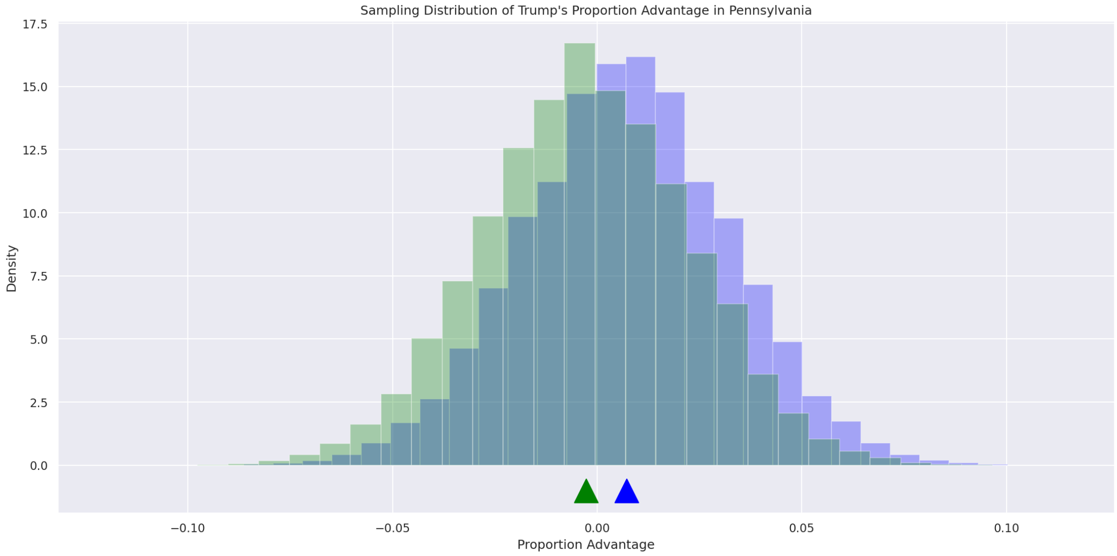
Make sure to give your plot a title, label the x and y axes and include a legend. Use the parameter `alpha` to adjust the transparency of each histogram.

```
In [65]: plt.hist(simulations, bins = 30, density=True, alpha = .3, color='blue')
         plt.hist(biased_simulations, bins = 30, density=True, alpha = .3, color='green')

         plt.scatter(simulations.mean(), -1, marker='^', color='blue', s=500, label='Simulations Mean')
         plt.scatter(biased_simulations.mean(), -1, marker='^', color='green', s=500, \
                     label='Biased Simulations Mean')

         plt.title("Sampling Distribution of Trump's Proportion Advantage in Pennsylvania")
         plt.ylabel("Density")
         plt.xlabel("Proportion Advantage")
```

```
Out[65]: Text(0.5, 0, 'Proportion Advantage')
```



Summarize the findings from these simulations:

i). Based on your simulations, what was the **chance of error** in correctly predicting that Trump wins using the **unbiased** samples of 1500 people from each state? Many people, even well educated ones, assume that this number should be 0%. After all, how could a non-biased sample be wrong? Give a mathematical explanation as to why it isn't 0% (or close to 0%). This is the type of incredibly important intuition we hope to develop in you throughout this class and your future data science coursework.

ii). What was the chance of error in predicting the results using the **biased** samples and how different is it from your answer in part(i)? Recall, we only biased the samples by 0.5%. However, even a bias this small in the percentages can lead to a much larger chance of error in prediction of the final result.

i). The chance of error in correctly predicting that Trump wins using the unbiased samples was 31%. The chance of error is not close to 0% because the real world is complex. For example, voter turnout and preferences can change due entirely to random chance, and in each sample of 1500 voters, there is the potential for variations in voter preferences.

ii). The chance of error in predicting the results using the biased samples was 55% which is 24% greater than the chance using the unbiased samples.

Part B Compare your observations from 7a to your observations in 6d. Did the chance of error increase or decrease in each case and why? What do these changes imply about the impact of sample size on the sampling error and on the bias?

The chance of error decreased for the unbiased sample, but the chance stayed the same for the biased sample. For unbiased samples, larger sample sizes generally lead to more accurate estimates because they provide more information about the population. However, biased samples have the fundamental issue of non-representativeness. Thus, increasing the sample size in a biased sample does not mitigate the bias itself.

Part C Is it possible to correctly predict Trump's victory with less than 1% error using **unbiased sampling?** Rerun the simulation (in each of the 4 states) with increasing sample sizes and 100,000 simulations to determine if you can find an approximate minimum sample size (it doesn't have to be exact) such that the probability of correctly predicting Trump's victory is at least 99% (assuming your sample is unbiased).

```
In [60]: trump20000 = sum(trump_wins(20000) for i in range(100000))
         trump30000 = sum(trump_wins(30000) for i in range(100000))
         trump40000 = sum(trump_wins(40000) for i in range(100000))

         sample_size_20000_incorrect = 1 - trump20000 / 100000
         sample_size_30000_incorrect = 1 - trump30000 / 100000
         sample_size_40000_incorrect = 1 - trump40000 / 100000

         print(sample_size_20000_incorrect, sample_size_30000_incorrect, sample_size_40000_incorrect)
         # your code above this line.
         # output the number of samples you used to get to at least 99% accuracy.
```

```
0.029689999999999994 0.010900000000000002 0.004020000000000236
```

Part D Is it possible to correctly predict Trump's victory with less than 1% error using **biased sampling?** Use the code cell below to rerun the simulation (in each of the 4 states) with increasing sample sizes. What happens to the probability of error? Explain in the markdown cell below.

It is not possible to predict Trump's victory with less than 1% error using biased sampling. When increasing sample size, the error actually increases because the larger sample size magnifies the bias.