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.

The mathematical definition of independence says

$$P(a \cap b) = P(a)P(b).$$

Take for example when $X = 1$ and $Y = 1$. The joint probability in this case is

$$P((X = 1) \cap (Y = 1)) = \frac{1}{3}.$$

This means, if these variables are independent we constitute that

$$P((X = 1) \cap (Y = 1)) = P(X = 1)P(Y = 1)$$

is true. We know from the previous parts of this problem that $P(X = 1)\frac{5}{12}$ and $P(Y = 1) = \frac{7}{12}$. Putting this together we can see

$$P(X = 1)P(Y = 1) = \frac{5}{12} \cdot \frac{7}{12} = \frac{35}{144}$$

which is **NOT** equal to the value of $P((X = 1) \cap (Y = 1)) = \frac{1}{3}$. Because of this we can then say

$$P((X = 1) \cap (Y = 1)) \neq P(X = 1)P(Y = 1)$$

and therefore we can conclude from this one example that X and Y are **NOT INDEPENDENT**. Showing that one of these cases is not independent is enough to conclude that these variables are not independent.

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

If the covariance of two variables is 0, this implies that there is **no linear relationship** between the two variables. It does not implicate necessarily that the two variables are independent, it just tells us that there is no linear relationship between these two variables.

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 $\text{Cov}(X,Y)$.
- ii). Calculate $\rho(X,Y)$

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

0.0.1 Part i

The mathematical formula for calculating the covariance of two variables is

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y].$$

We first need to fill out a table so that we have the marginal PMF of each variable:

	$Y=0$	$Y=1$	$Y=2$	
$X=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{13}{24}$
$X=1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{11}{24}$
	$\frac{7}{24}$	$\frac{10}{24}$	$\frac{7}{24}$	1

We next calculate $E(X)$ which is

$$E(X) = \sum_{i=0}^{1} X_i P(X_i) = 0 \cdot \frac{13}{24} + 1 \cdot \frac{11}{24}.$$

Now $E(y)$ which is

$$E(Y) = \sum_{i=0}^{1} Y_i P(Y_i) = 0 \cdot \frac{7}{24} + 1 \cdot \frac{10}{24} + 2 \cdot \frac{7}{24} = \frac{10}{24} + \frac{14}{24} = 1.$$

Now $E(XY)$ which is

$$E(XY) = (1 \cdot 0) \cdot \frac{1}{8} + (1 \cdot 1) \cdot \frac{1}{6} + (1 \cdot 2) \cdot \frac{1}{6} = 0 + \frac{3}{6} = \frac{12}{24}.$$

Now, combining these results we then have

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = \frac{12}{24} - \frac{11}{24} = \frac{1}{24}.$$

0.0.2 Part ii

The mathematical for correlation is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$$

All we need to do is calculate the standard deviation of X and Y. The formula for standard deviation is

$$\sigma_\alpha = \sqrt{\text{Var}(\alpha)} = \sqrt{E(\alpha^2) - E(\alpha)^2}.$$

So we first need to calculate the variances of X and Y. The variance of X is

$$\begin{aligned}
\text{Var(X)} &= E(X^2) - E(X)^2 && (1)\\
&= \left(\sum_i X_i^2 P(X_i)\right) - \left(\sum_i X_i P(X_i)\right)^2 && (2)\\
&= \left(0^2 \cdot \frac{13}{24} + 1^2 \cdot \frac{11}{24}\right) - \left(0 \cdot \frac{13}{24} + 1 \cdot \frac{11}{24}\right)^2 && (3)\\
&= \frac{11}{24} - \left(\frac{11}{24}\right)^2 = \frac{11}{24}\left(1 - \frac{11}{24}\right) = \frac{11}{24} \cdot \frac{13}{24} = \frac{143}{576}. && (4)
\end{aligned}$$

The variance of Y is then

$$\begin{aligned}
\text{Var(Y)} &= E(Y^2) - E(Y)^2 && (5)\\
&= \left(\sum_i Y_i^2 P(Y_i)\right) - \left(\sum_i Y_i P(Y_i)\right)^2 && (6)\\
&= \left(0^2 \cdot \frac{7}{24} + 1^2 \cdot \frac{10}{24} + 2^2 \cdot \frac{7}{24}\right) - \left(0 \cdot \frac{7}{24} + 1 \cdot \frac{10}{24} + 2 \cdot \frac{7}{24}\right)^2 && (7)\\
&= \left(\frac{10}{24} + \frac{28}{24}\right) - \left(\frac{10}{24} + \frac{14}{24}\right)^2 && (8)\\
&= \frac{38}{24} - 1^2 = \frac{14}{24}. && (9)
\end{aligned}$$

This then means our standard deviations are then

$$\begin{aligned}
\sigma_Y &= \sqrt{\frac{143}{576}} && (10)\\
\sigma_Y &= \sqrt{\frac{14}{24}} && (11)
\end{aligned}$$

Putting this altogether the correlation of X and Y is then

$$\begin{aligned}
\rho(X,Y) &= \frac{1/24}{\sqrt{\frac{143}{576}}\sqrt{\frac{14}{24}}} && (12)\\
&= \frac{\sqrt{\frac{1}{576}}}{\sqrt{\frac{143}{576}}\sqrt{\frac{14}{24}}} && (13)\\
&= \sqrt{\frac{24}{143 \cdot 14}} = \sqrt{\frac{12}{143 \cdot 7}} && (14)\\
&= \sqrt{\frac{12}{1001}} = 2\sqrt{\frac{3}{1001}}. && (15)
\end{aligned}$$

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). $\text{Cov}(X, Y)$

ii). $\text{Cov}(Y, X)$

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

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

0.0.3 Part i

Using the given information the covariance of X and Y can be calculated with

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[XY] - \mu_X\mu_Y = 27 - (6)(5) = 27 - 30 = -3.$$

0.0.4 Part ii

The covariance of Y and X is the same as the covariance of X and Y in this context, so

$$\text{Cov}(Y, X) = -3.$$

0.0.5 Part iii

The correlation of X and Y is calculated with

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-3}{\sqrt{4}\sqrt{9}} = \frac{-3}{2(3)} = -\frac{1}{2}.$$

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.

0.1.1 Part i

The population of interest are the people in the United States who are eligible to vote. This takes into account the people who are going to vote as well as the people who are not going to vote.

0.1.2 Part ii

The sampling frame is the list of people who are registered to vote. This takes into account people who are most likely going to vote in the election.

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 [30]: def draw_state_sample(N, state):
             trump = 0
             clinton = 0
             if (state == "florida"):
                 trump = 0.4902
                 clinton = 0.4782
             elif (state == "michigan"):
                 trump = 0.4750
                 clinton = 0.4727
             elif (state == "pennsylvania"):
                 trump = 0.4818
                 clinton = 0.4746
             elif (state == "wisconsin"):
                 trump = 0.4722
                 clinton = 0.4645
             other = 1 - (trump + clinton)
             sample = np.random.multinomial(N, [trump, clinton, other])
             return sample.tolist()
```

```
In [31]: grader.check("q5a")
```

```
Out[31]: q5a results: All test cases passed!
```

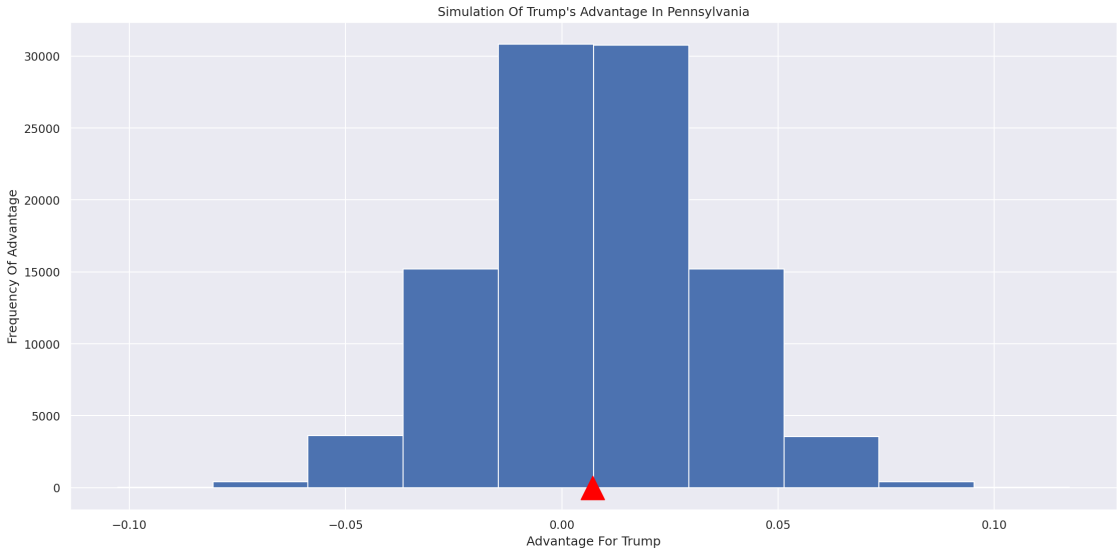
Part D i). Make a **frequency** histogram of `simulations`. This is a histogram of the sampling distribution of Trump's proportion advantage in Pennsylvania.

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 [36]:
```
# Part (i):
plt.hist(simulations)
plt.xlabel("Advantage For Trump")
plt.ylabel("Frequency Of Advantage")
plt.title("Simulation Of Trump's Advantage In Pennsylvania")
# 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);
```



In [37]:
```
# Part (ii):
prob_penn_1500_random_correct = \
sum(simulations[i] > 0 for i in range(len(simulations))) / len(simulations)

prob_penn_1500_random_correct
```

```
Out[37]: 0.60607
```

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 [43]: def draw_biased_state_sample(N, state):
             trump = 0
             clinton = 0
             bias = 0.005
             if (state == "florida"):
                 trump = 0.4902 - bias
                 clinton = 0.4782 + bias
             elif (state == "michigan"):
                 trump = 0.4750 - bias
                 clinton = 0.4727 + bias
             elif (state == "pennsylvania"):
                 trump = 0.4818 - bias
                 clinton = 0.4746 + bias
             elif (state == "wisconsin"):
                 trump = 0.4722 - bias
                 clinton = 0.4645 + bias
             other = 1 - (trump + clinton)
             sample = np.random.multinomial(N, [trump, clinton, other])
             return sample.tolist()

         biased_simulations = \
         np.array([trump_advantage(draw_biased_state_sample(1500, "pennsylvania")) \
         for _ in range(100000)])
```

```
In [44]: grader.check("q6a")
```

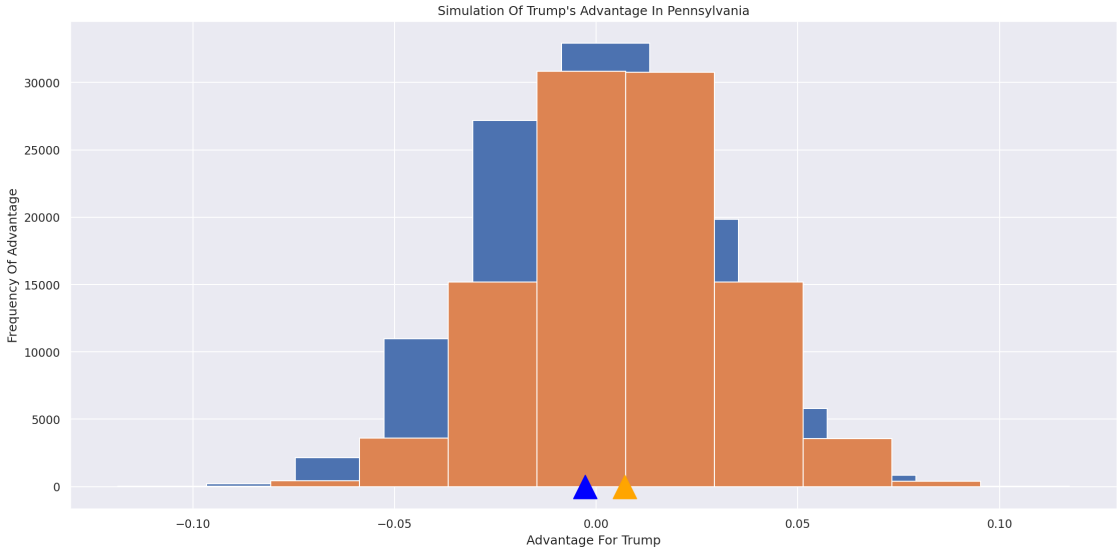
```
Out[44]: q6a results: All test cases passed!
```

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 [45]:
```
plt.hist(biased_simulations)
plt.hist(simulations)
plt.xlabel("Advantage For Trump")
plt.ylabel("Frequency Of Advantage")
plt.title("Simulation Of Trump's Advantage In Pennsylvania")
# your code for the histogram above here.  The code below plots a red marker at the mean:
plt.scatter(biased_simulations.mean(), -1, marker='^', color='blue', s=500);
plt.scatter(simulations.mean(), -1, marker='^', color='orange', s=500);
```



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.

0.3.1 Part i

To calculate the chance of error in correctly predicting that Trump wins using the unbiased samples of 1500 people from each state is simply 1 minus proportion of times Trump wins. Calculating this we find

$$\text{C.O.E} = 1 - \text{proportion\_trump} = 1 - 0.6931 = 0.3069.$$

Non biased sample can be wrong because of some of the aforementioned reasons from before. For instance, we people polled, some hid their support for Trump due to many reasons. This can cause the polls to be incorrect and thus not predict correct results. The chance of error in this case is the proportion of times Trump loses, and this is not close to zero because if it were, it would essentially say that Clinton wins almost 100% of the time and this is not realistic.

0.3.2 Part ii

To calculate this chance of error, we do the same as we did in the prior part for the unbiased scenario. So

$$\text{C.O.E} = 1 - \text{proportion\_trump\_biased} = 1 - 0.46429 = 0.53571$$

is the chance of error in the biased example. This essentially is the percentage of times Trump would lose in the biased scenario. This is drastically different than the unbiased example and thus explains why so many predicted that Clinton would win even though she did not.

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 in the unbiased example (drastically) and increased (minutely). The drastic change in the unbiased sample implicates that having a higher sample size will deduce the chance of error significantly. Conversely, having a higher sample size with a biased sample will only slightly affect the chance of error and still not give an accurate representation of the results that are being sought after.

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 [52]: very_high_trump_proportion_incorrect = \
         1 - sum([trump_wins(30000) for _ in range(100000)]) / 100000
         print(f"Incorrect Prediction: {100 * np.round(very_high_trump_proportion_incorrect,2)}%")
         print(f"Sample Size: 30,000")
         # your code above this line.
         # output the number of samples you used to get to at least 99% accuracy.
```

```
Incorrect Prediction: 1.0%
Sample Size: 30,000
```

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.

If we increase the sample size, the proportion of times Trump wins will actually decrease. This means that as the sample size grows, the error actually increases. This is because we calculate the error by subtracting the number of times Trump wins (the proportion) from 1. Because of this it is **highly unlikely** that we can correctly predict Trump's victory with less than 1% error. The existence of the bias in the sample will significantly skew the results and thus make generating an accurate prediction extremely difficult.