Homework 5

Due October 22 at 11 pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Babysitter) A babysitter is taking care of a baby. She gives him some food and then puts him to sleep. We make the following assumptions:

   - The probability that the food is bad is 0.1.
   - If a baby eats food that is bad, they will wake up in the middle of the night. If the food is not bad, they may still wake up (with a probability that depends on whether they are good or bad sleepers).
   - All babies can be classified into *good sleepers* or *bad sleepers*. The probability that a baby that is a *good sleeper* wakes up is 0.1 (assuming the food is fine). The probability for a baby that is a *bad sleeper* is 0.8.
   - A baby is a *good sleeper* with probability 0.6. This is independent from the food.

   We model the problem by defining Bernoulli random variables $\tilde{b}$ indicating whether the baby is a good ($\tilde{b} = 1$) or bad sleeper ($\tilde{b} = 0$), $\tilde{w}$ indicating whether the baby wakes up in the middle of the night ($\tilde{w} = 1$) or not ($\tilde{w} = 0$), and $\tilde{x}$ indicating whether the food is bad ($\tilde{x} = 1$) or not ($\tilde{x} = 0$).

   (a) What is the probability that the baby wakes up in the middle of the night?

   Using the total law of probability, we know:

$$
\begin{aligned}
P(\tilde{w} = 1) &= P(\tilde{w} = 1|\tilde{x} = 1)P(\tilde{x} = 1) + P(\tilde{w} = 1|\tilde{x} = 0)P(\tilde{x} = 0) \\
&\qquad \text{(Total law of Probability)} \\
&= P(\tilde{w} = 1|\tilde{x} = 1)P(\tilde{x} = 1) \\
&+ [P(\tilde{w} = 1|\tilde{b} = 1, \tilde{x} = 0)P(\tilde{b} = 1|\tilde{x} = 0) \\
&+ (\tilde{w} = 1|\tilde{b} = 0, \tilde{x} = 0)P(\tilde{b} = 0|\tilde{x} = 0)] \times P(\tilde{x} = 0) \\
&\qquad \text{(Total law of probability/Chain Rule)} \\
&= 1 \times 0.1 + (0.1 \times 0.6 + 0.8 \times 0.4) \times 0.9 \qquad \text{(Fit in the probability)} \\
&= 0.442
\end{aligned}
$$

(b) If the baby wakes up in the middle of the night, what is the probability that the food was bad?

We have the following derivations:

$$\begin{aligned} P(\tilde{x}=1|\tilde{w}=1) &= \frac{P(\tilde{w}=1|\tilde{x}=1)P(\tilde{x}=1)}{P(\tilde{w}=1)} && \text{(Conditional Probability Formula)} \\ &= \frac{1\times 0.1}{0.442} && \text{(Fit in the value)} \\ &= 0.226 \end{aligned}$$

(c) Compute the probability that the food is bad conditioned on the baby waking up and being a good sleeper. Are $\tilde{b}$ and $\tilde{x}$ conditionally independent given $\tilde{w}$? Justify your answer mathematically and explain it intuitively.

We have the following derivations:

$$\begin{aligned} P(\tilde{x}=1|\tilde{w}=1,\tilde{b}=1) &= \frac{P(\tilde{w}=1,\tilde{b}=1|\tilde{x}=1)P(\tilde{x}=1)}{P(\tilde{w}=1,\tilde{b}=1)} \\ & && \text{(Conditional probability)} \\ &= \frac{P(\tilde{w}=1|\tilde{b}=1,\tilde{x}=1)P(\tilde{b}=1|\tilde{x}=1)P(\tilde{x}=1)}{\sum_{i=0,1}P(\tilde{w}=1|\tilde{b}=1,\tilde{x}=i)P(\tilde{b}=1|\tilde{x}=i)P(\tilde{x}=i)} \\ & && \text{(Chain Rule)} \\ &= \frac{1\times 0.6\times 0.1}{1\times 0.6\times 0.1+0.1\times 0.6\times 0.9} && \text{(Fir in the value)} \\ &= 0.526 \end{aligned}$$

If $\tilde{b}$ and $\tilde{x}$ are conditionally independent given $\tilde{w}$, we should have $P(\tilde{x}=1|\tilde{w}=1,\tilde{b}=1)=P(\tilde{x}=1|\tilde{w}=1)$. But from part (b) these two quantities doesn't match. So $\tilde{b}$ and $\tilde{x}$ are not conditionally independent given $\tilde{w}$.

The intuitive explanations would be that if we know that the baby wakes up, then if the food is good, then we have more confidence to say that the baby is a bad sleeper. Alternatively, given that the baby wakes up, if the baby is a good sleeper, it is very likely that the food is bad.

2. (Earthquake) During a period of high seismic activity, a group of scientists is trying to predict the occurrence of earthquakes by measuring vibrations in the ground. They model the occurrence of an earthquake as a random variable $\tilde{e}$ ($\tilde{e} = 1$ if there is an earthquake, and $\tilde{e} = 0$ if there isn't), and the vibrations as a random variable $\tilde{v}$ ($\tilde{v} = 0$ if there are no vibrations, $\tilde{v} = 1$ if there are small vibrations, and $\tilde{v} = 2$ if there are large vibrations). The joint pmf of $\tilde{e}$ and $\tilde{v}$ is:

		Vibrations		
	$p_{\tilde{e},\tilde{v}}$	0	1	2
Earthquake	0	0.8	0.05	0
	1	0	0.05	0.1

The sensor reading is modeled as a Bernoulli random variable $\tilde{s}$ that is conditionally independent of the earthquake given the vibrations. If there are no vibrations, the reading is always 0, if there are small vibrations the reading is 1 with probability 0.5, and if there are large vibrations the reading is always 1.

(a) Derive the marginal pmf of $\tilde{s}$.

Based on the statement, we have $\begin{cases} p_{\tilde{s}|\tilde{v}}(0|0) = 1 \\ p_{\tilde{s}|\tilde{v}}(1|1) = 0.5, p_{\tilde{s}|\tilde{v}}(0|1) = 0.5 \\ p_{\tilde{s}|\tilde{v}}(1|2) = 1 \end{cases}$ and

$\begin{cases} p_{\tilde{v}}(0) = 0.8 \\ p_{\tilde{v}}(1) = 0.05 + 0.05 = 0.1 \\ p_{\tilde{v}}(2) = 0 + 0.1 = 0.1 \end{cases}$ Then we calculate the following marginal pmf for $\tilde{s}$:

$$\begin{cases} p_{\tilde{s}}(0) = p_{\tilde{s}|\tilde{v}}(0|0)p_{\tilde{v}}(0) + p_{\tilde{s}|\tilde{v}}(0|1)p_{\tilde{v}}(1) + p_{\tilde{s}|\tilde{v}}(0|2)p_{\tilde{v}}(2) = 1 \times 0.8 + 0.5 \times 0.1 + 0 \times 0.1 = 0.85 \\ p_{\tilde{s}}(1) = p_{\tilde{s}|\tilde{v}}(1|0)p_{\tilde{v}}(0) + p_{\tilde{s}|\tilde{v}}(1|1)p_{\tilde{v}}(1) + p_{\tilde{s}|\tilde{v}}(1|2)p_{\tilde{v}}(2) = 0 \times 0.8 + 0.5 \times 0.1 + 1 \times 0.1 = 0.15 \end{cases}$$

(b) What is the probability that there is an earthquake if the sensor reading equals 1?

We have the following derivations:

$$\begin{aligned} p_{\tilde{e}|\tilde{s}}(1|1) &= \frac{p_{\tilde{s},\tilde{e},\tilde{v}}(1,1,1) + p_{\tilde{s},\tilde{e},\tilde{v}}(1,1,0)}{p_{\tilde{s}}(1)} && \text{(Conditional probability)} \\ &= \frac{p_{\tilde{s}|\tilde{e},\tilde{v}}(1|1,1)p_{\tilde{e},\tilde{v}}(1,1) + p_{\tilde{s}|\tilde{e},\tilde{v}}(1|1,0)p_{\tilde{e},\tilde{v}}(1,0)}{0.15} && \text{(Conditional Probability)} \\ &= \frac{p_{\tilde{s}|\tilde{v}}(1|1)p_{\tilde{e},\tilde{v}}(1,1) + p_{\tilde{s}|\tilde{v}}(1|0)p_{\tilde{e},\tilde{v}}(1,0)}{0.15} && \text{(Conditional Independence)} \\ &= \frac{0.5 \times 0.05 + 0 \times 0}{0.15} && \text{(Fit in the value)} \\ &= \frac{1}{6} \end{aligned}$$

(c) Are the random variables $\tilde{s}$ and $\tilde{e}$ independent? Justify your answer mathematically, but also explain it intuitively.

They are not independent. Since if they were we would have at least $p_{\tilde{e}|\tilde{s}}(1|1) = p_{\tilde{e}}(1)$. But the LHS is equal to $\frac{1}{6}$ while the RHS is equal to $0.05 + 0.1 = 0.15$.

The intuitive explanation would be that conditional independence doesn't imply dependence. If we know that the earthquake happens, there will definitely be vibrations in the ground, which is given in the table. If there is vibrations, then the sensors will more likely to deliver 1 then 0. Which means the earthquake provides some information about our sensor's measurement. So $\tilde{s}$ and $\tilde{e}$ should not be independent.

3. (Surgery) A hospital wants to evaluate two surgery procedures: A and B. There are two types of patients that receive the procedure, *mild* and *serious* cases. The truth is that procedure A is better. Mild cases recover with probability 0.9 if they receive A, and 0.8 if they receive B. Serious cases recover with probability 0.5 if they receive A, and 0.2 if they receive B.

   (a) The data shows that patients recover with probability 0.58 if they receive procedure A, and 0.68 if they receive B. How is this possible? Justify your answer mathematically. (Hint: Start by computing what fraction of patients receiving each procedure are mild or serious cases.)

   This is due to simpson's paradox. Let's denote the following:

$$\begin{cases} \tilde{x} = 1 : Serious, \tilde{x} = 0 : Mild \\ \tilde{t} = 1 : Proc\ \ A, \tilde{t} = 0 : Proc\ \ B \\ \tilde{y} = 1 : Recover, \tilde{y} = 0 : Not\ \ Recover \end{cases}$$

   Then we have the following:

   i. $p_{\tilde{y}|\tilde{x},\tilde{t}}(1|0,1) = 0.9$, $p_{\tilde{y}|\tilde{x},\tilde{t}}(1|0,0) = 0.8$
   ii. $p_{\tilde{y}|\tilde{x},\tilde{t}}(1|1,1) = 0.5$, $p_{\tilde{y}|\tilde{x},\tilde{t}}(1|1,0) = 0.2$
   iii. $p_{\tilde{y}|\tilde{t}}(1|1) = 0.58$, $p_{\tilde{y}|\tilde{t}}(1|0) = 0.68$

   , then we can solve the following equations:

   i. $0.58 = p_{\tilde{y}|\tilde{t}}(1|1) = p_{\tilde{y}|\tilde{x},\tilde{t}}(1|0,1)p_{\tilde{x}|\tilde{t}}(0|1) + p_{\tilde{y}|\tilde{x},\tilde{t}}(1|1,1)p_{\tilde{x}|\tilde{t}}(1|1) = 0.9 \times p_{\tilde{x}|\tilde{t}}(0|1) + 0.5 \times p_{\tilde{x}|\tilde{t}}(1|1)$
   ii. $0.42 = p_{\tilde{y}|\tilde{t}}(1|1) = p_{\tilde{y}|\tilde{x},\tilde{t}}(0|0,1)p_{\tilde{x}|\tilde{t}}(0|1) + p_{\tilde{y}|\tilde{x},\tilde{t}}(0|1,1)p_{\tilde{x}|\tilde{t}}(1|1) = 0.1 \times p_{\tilde{x}|\tilde{t}}(0|1) + 0.5 \times p_{\tilde{x}|\tilde{t}}(1|1)$
   iii. $0.68 = p_{\tilde{y}|\tilde{t}}(1|0) = p_{\tilde{y}|\tilde{x},\tilde{t}}(1|1,0)p_{\tilde{x}|\tilde{t}}(1|0) + p_{\tilde{y}|\tilde{x},\tilde{t}}(1|0,0)p_{\tilde{x}|\tilde{t}}(0|0) = 0.2 \times p_{\tilde{x}|\tilde{t}}(0|1) + 0.8 \times p_{\tilde{x}|\tilde{t}}(1|1)$
   iv. $0.32 = p_{\tilde{y}|\tilde{t}}(0|0) = p_{\tilde{y}|\tilde{x},\tilde{t}}(0|1,0)p_{\tilde{x}|\tilde{t}}(1|0) + p_{\tilde{y}|\tilde{x},\tilde{t}}(0|0,0)p_{\tilde{x}|\tilde{t}}(0|0) = 0.8 \times p_{\tilde{x}|\tilde{t}}(0|1) + 0.2 \times p_{\tilde{x}|\tilde{t}}(1|1)$

   , we solve to get the following:

   i. $p_{\tilde{x}|\tilde{t}}(1|1) = 0.8$
   ii. $p_{\tilde{x}|\tilde{t}}(0|1) = 0.2$
   iii. $p_{\tilde{x}|\tilde{t}}(1|0) = 0.2$
   iv. $p_{\tilde{x}|\tilde{t}}(0|0) = 0.8$

   So we see that the majority patients who receive A are of serious cases, which causes low likelihood of recovery, while and the majority of patients who receive B are of mild cases, which ensures high likelihood of recovery. Thus, even if procedure A is more effective, the fact that most patients are under serious conditions and are hard to recover, the overall effect of procedure A is undermined by the imbalanced assignment of the procedure to the individuals.

(b) Explain how to analyze the data in order to obtain an accurate conclusion about the surgery procedures. Under what assumption does this work?

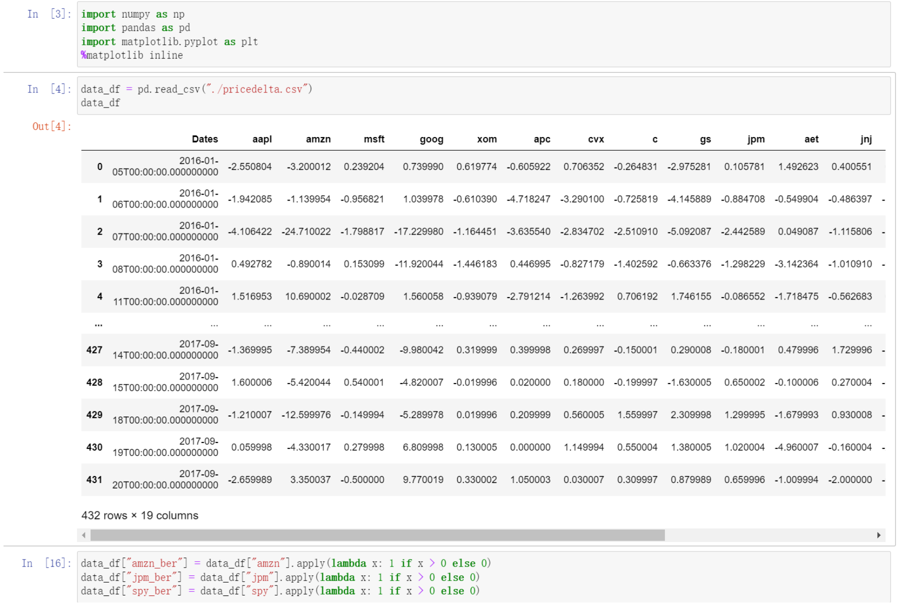
The assumption should be that $\tilde{x}$ and $\tilde{t}$ are independent. Under this assumption, the distribution of the severity type of patients will be the same within the group that receives procedure A and the group with procedure B, which serves to cancel out the confounding factors out.

(c) Suggest how to design a follow-up study that would not require adjusting for confounding factors.

A follow-up study that does not require correcting for confounding factors could be designed by randomly assigning patients to either procedure A or procedure B. This ensures that the distribution of patient severity is the same for both operations, eliminating the need to account for confounding circumstances.

4. (Stock) The table in *pricedelta.csv* records daily share prices changes. For each stock, model whether the price goes up or down with a Bernoulli random variable. Estimate the following probabilities from the data.

   Codes for data preprocessing are attached below:



```
In [3]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline

In [4]: data_df = pd.read_csv("./pricedelta.csv")
        data_df

Out[4]:
```

	Dates	aapl	amzn	msft	goog	xom	apc	cvx	c	gs	jpm	aet	jnj
0	2016-01-05T00:00:00.000000000	-2.550804	-3.200012	0.239204	0.739990	0.619774	-0.605922	0.706352	-0.264831	-2.975281	0.105781	1.492623	0.400551
1	2016-01-06T00:00:00.000000000	-1.942085	-1.139954	-0.956821	1.039978	-0.610390	-4.718247	-3.290100	-0.725819	-4.145889	-0.884708	-0.549904	-0.486397
2	2016-01-07T00:00:00.000000000	-4.106422	-24.710022	-1.798817	-17.229980	-1.164451	-3.635540	-2.834702	-2.510910	-5.092087	-2.442589	0.049087	-1.115806
3	2016-01-08T00:00:00.000000000	0.492782	-0.890014	0.153099	-11.920044	-1.446183	0.446995	-0.827179	-1.402592	-0.663376	-1.298229	-3.142364	-1.010910
4	2016-01-11T00:00:00.000000000	1.516953	10.690002	-0.028709	1.560058	-0.939079	-2.791214	-1.263992	0.706192	1.746155	-0.086552	-1.718475	-0.562683
...	...	...	...	...	...	...	...	...	...	...	...	...	...
427	2017-09-14T00:00:00.000000000	-1.369995	-7.389954	-0.440002	-9.980042	0.319999	0.399998	0.269997	-0.150001	0.290008	-0.180001	0.479996	1.729996
428	2017-09-15T00:00:00.000000000	1.600006	-5.420044	0.540001	-4.820007	-0.019996	0.020000	0.180000	-0.199997	-1.630005	0.650002	-0.100006	0.270004
429	2017-09-18T00:00:00.000000000	-1.210007	-12.599976	-0.149994	-5.289978	0.019996	0.209999	0.560005	1.559997	2.309998	1.299995	-1.679993	0.930008
430	2017-09-19T00:00:00.000000000	0.059998	-4.330017	0.279998	6.809998	0.130005	0.000000	1.149994	0.550004	1.380005	1.020004	-4.960007	-0.160004
431	2017-09-20T00:00:00.000000000	-2.659989	3.350037	-0.500000	9.770019	0.330002	1.050003	0.030007	0.309997	0.879989	0.659996	-1.009994	-2.000000

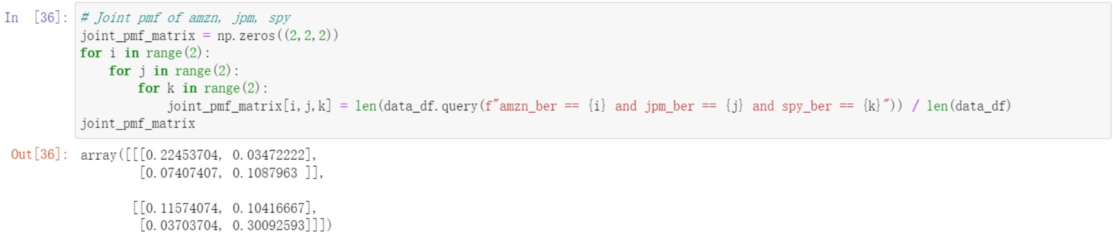
432 rows × 19 columns

```
In [16]: data_df["amzn_ber"] = data_df["amzn"].apply(lambda x: 1 if x > 0 else 0)
         data_df["jpm_ber"] = data_df["jpm"].apply(lambda x: 1 if x > 0 else 0)
         data_df["spy_ber"] = data_df["spy"].apply(lambda x: 1 if x > 0 else 0)
```

Figure 1: Data Preprocessing

(a) What is the joint pmf of the three random variables representing *amzn*, *jpm*, *spy*? (One $2 \times 2 \times 2$ matrix)

   Codes and matrices are attached as below:

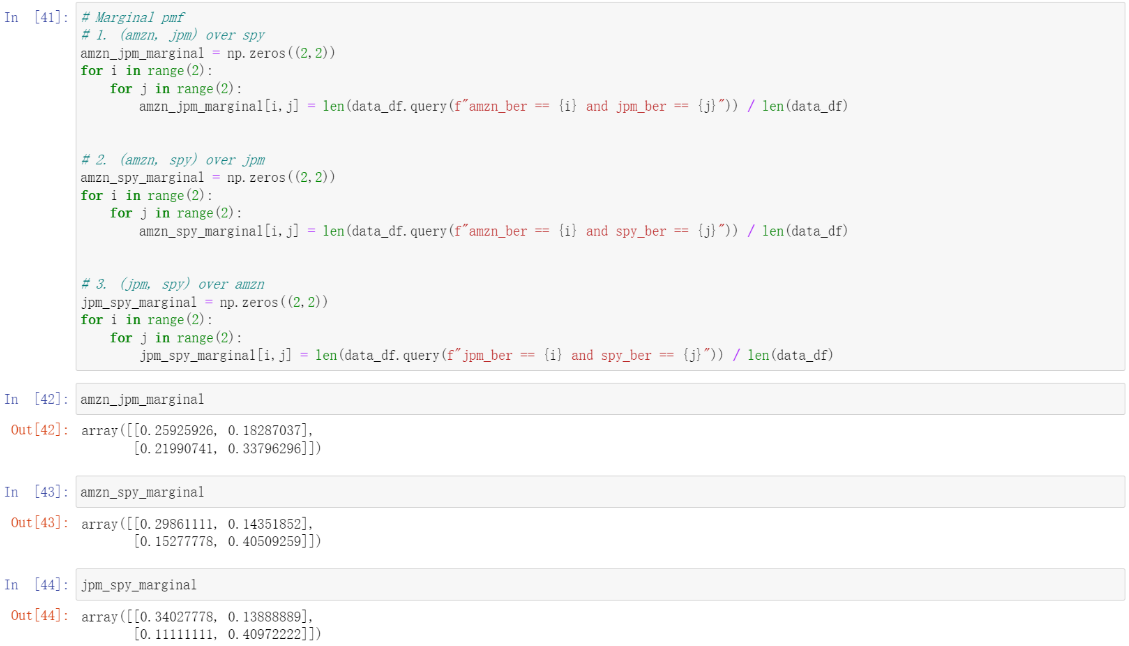


```
In [36]: # Joint pmf of amzn, jpm, spy
         joint_pmf_matrix = np.zeros((2,2,2))
         for i in range(2):
             for j in range(2):
                 for k in range(2):
                     joint_pmf_matrix[i,j,k] = len(data_df.query(f"amzn_ber == {i} and jpm_ber == {j} and spy_ber == {k}")) / len(data_df)
         joint_pmf_matrix

Out[36]: array([[[0.22453704, 0.03472222],
                 [0.07407407, 0.1087963 ]],

                [[0.11574074, 0.10416667],
                 [0.03703704, 0.30092593]]])
```

Figure 2: Joint PMF Matrix

(b) What is the marginal pmf of each possible pair of random variables? (Three 2 × 2 matrices)

Codes and matrices are attached as below:



```
In [41]: # Marginal pmf
         # 1. (amzn, jpm) over spy
         amzn_jpm_marginal = np.zeros((2,2))
         for i in range(2):
             for j in range(2):
                 amzn_jpm_marginal[i,j] = len(data_df.query(f"amzn_ber == {i} and jpm_ber == {j}")) / len(data_df)


         # 2. (amzn, spy) over jpm
         amzn_spy_marginal = np.zeros((2,2))
         for i in range(2):
             for j in range(2):
                 amzn_spy_marginal[i,j] = len(data_df.query(f"amzn_ber == {i} and spy_ber == {j}")) / len(data_df)


         # 3. (jpm, spy) over amzn
         jpm_spy_marginal = np.zeros((2,2))
         for i in range(2):
             for j in range(2):
                 jpm_spy_marginal[i,j] = len(data_df.query(f"jpm_ber == {i} and spy_ber == {j}")) / len(data_df)
```

```
In [42]: amzn_jpm_marginal

Out[42]: array([[0.25925926, 0.18287037],
                [0.21990741, 0.33796296]])
```

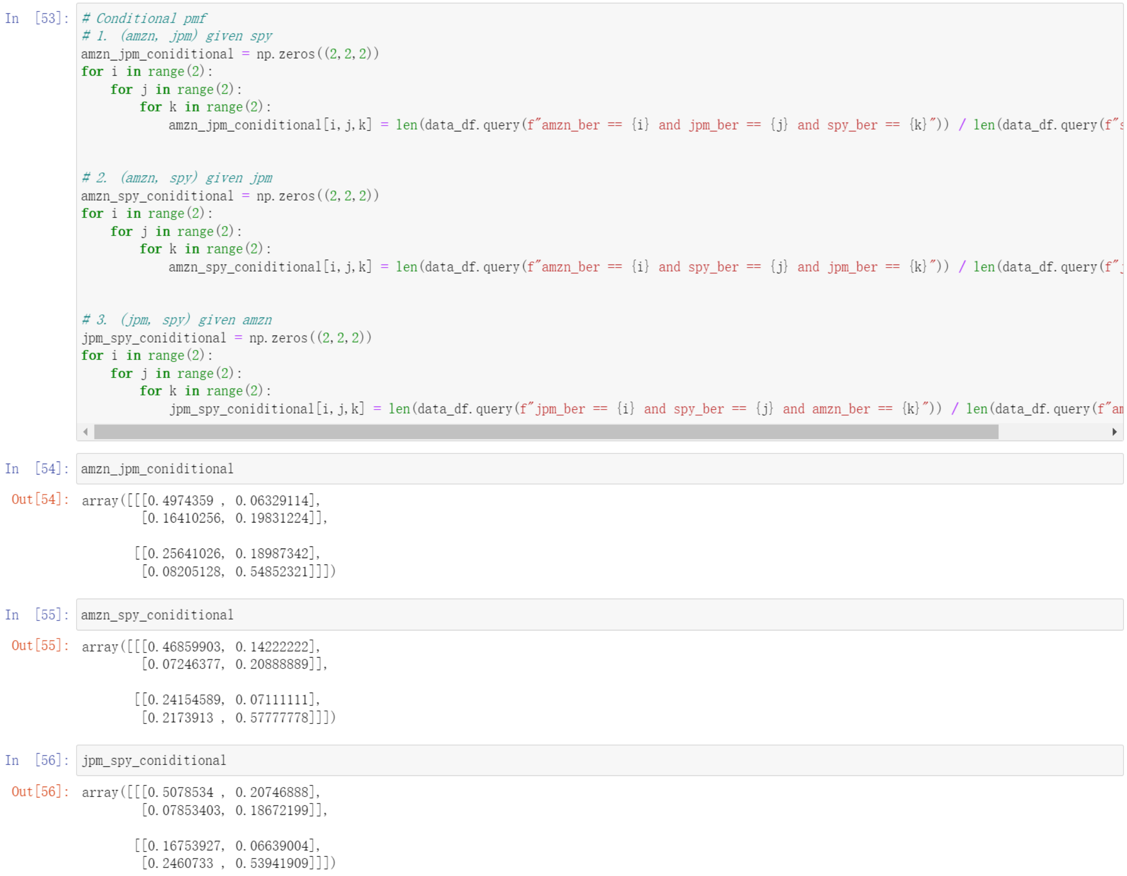
```
In [43]: amzn_spy_marginal

Out[43]: array([[0.29861111, 0.14351852],
                [0.15277778, 0.40509259]])
```

```
In [44]: jpm_spy_marginal

Out[44]: array([[0.34027778, 0.13888889],
                [0.11111111, 0.40972222]])
```

Figure 3: Marginal PMF Matrix

(c) What is the conditional pmf of each possible pair given the remaining random variable? (Three 2 × 2 × 2 matrices)

Codes and matrices are attached as below:



```
In [53]: # Conditional pmf
         # 1. (amzn, jpm) given spy
         amzn_jpm_coniditional = np.zeros((2,2,2))
         for i in range(2):
             for j in range(2):
                 for k in range(2):
                     amzn_jpm_coniditional[i,j,k] = len(data_df.query(f"amzn_ber == {i} and jpm_ber == {j} and spy_ber == {k}")) / len(data_df.query(f"s


         # 2. (amzn, spy) given jpm
         amzn_spy_coniditional = np.zeros((2,2,2))
         for i in range(2):
             for j in range(2):
                 for k in range(2):
                     amzn_spy_coniditional[i,j,k] = len(data_df.query(f"amzn_ber == {i} and spy_ber == {j} and jpm_ber == {k}")) / len(data_df.query(f"j


         # 3. (jpm, spy) given amzn
         jpm_spy_coniditional = np.zeros((2,2,2))
         for i in range(2):
             for j in range(2):
                 for k in range(2):
                     jpm_spy_coniditional[i,j,k] = len(data_df.query(f"jpm_ber == {i} and spy_ber == {j} and amzn_ber == {k}")) / len(data_df.query(f"am

In [54]: amzn_jpm_coniditional

Out[54]: array([[[0.4974359 , 0.06329114],
                 [0.16410256, 0.19831224]],

                [[0.25641026, 0.18987342],
                 [0.08205128, 0.54852321]]])

In [55]: amzn_spy_coniditional

Out[55]: array([[[0.46859903, 0.14222222],
                 [0.07246377, 0.20888889]],

                [[0.24154589, 0.07111111],
                 [0.2173913 , 0.57777778]]])

In [56]: jpm_spy_coniditional

Out[56]: array([[[0.5078534 , 0.20746888],
                 [0.07853403, 0.18672199]],

                [[0.16753927, 0.06639004],
                 [0.2460733 , 0.53941909]]])
```

Figure 4: Conditional PMF Matrix