1) Simpson's Paradox

a)

Machine 1	Wins	Losses	Total games
You	$\frac{40}{100} = 0.4$	$\frac{60}{100} = 0.6$	100
Friend	$\frac{30}{100} = 0.3$	$\frac{70}{100} = 0.7$	100

Machine 2	Wins	Losses	Total games
You	$\frac{210}{1040} = 0.202$	$\frac{830}{1040} = 0.798$	1040
Friend	$\frac{14}{84} = 0.167$	$\frac{70}{84} = 0.833$	84

After finding the winning probability, we can see that in both machines I would win more often than my friend.

b)

	Wins	Losses	Total games
You	$\frac{(40+210)}{1140} = 0.219$	$\frac{(60+830)}{1140} = 0.781$	40+60+210+830= 1140
Friend	$\frac{(30+14)}{184} = 0.239$	$\frac{(70+70)}{184} = 0.761$	30+70+14+70= 184

The overall winning probability shows that my friend will have a higher probability of winning than me.
The machines are independent so

c)

When individually comparing each machine, I will have a higher probability of winning. However, if we calculate overall probability then the case is reversed. This manifests in the Simpson's Paradox where groups of data may show a particular trend but when combined the trend is reversed.
Mathematically speaking, the equation used for calculating overall probability throughout the casino was:

$P(X) = \frac{\Sigma Wins(X)}{\Sigma Wins(X)+\Sigma Losses(X)}$ X being the player i.e. me or my friend.

$(\frac{1040}{1140} * 0.202) + (\frac{100}{1140} * 0.4) = 0.219 - Me$

$(\frac{84}{184} * 0.167) + (\frac{100}{184} * 0.3) = 0.239 - Friend$

If the denominator is significantly larger, more losses overall, this will reduce the probability of winning. Since I played significantly more than my friend, my total overall winnings and losses will be much greater. Even if mathematically I have a higher probability of winning in both machines since I played more, the overall winning probability is reduced due to my high number of plays. Since my friend played less, their denominator calculation is lower which will tilt the importance to their winnings.

2) Matrix as Operations

a)

W * (1, 1) = (-0.8, 2.6)
(1, -1) = (1.6, -0.2)

(a, c) * (1, 1) = (-0.8, 2.6)
(b, d) (1, -1) = (1.6, -0.2)

a+c= -0.8
a-c= 2.6
b+d= 1.6
b-d= - 0.2

2a= 1.8 0.9+c= -0.8 2b= 1.4 0.7 +d= 1.6
a= 0.9 c= -1.7 b= 0.7 d= 0.9

W= (0.9, -1.7)
(0.7, 0.9)

b) We use the rotation equation of : (cos x , - sin x)
(sin x , cos x)

$sin^2x + cos^2x = 1$

$sin^2x + \frac{1}{1+tan3} = 1$

$sin^2x + \frac{1}{10} = 1$

$sin^2x = 1 - \frac{1}{10}$ $sinx = \sqrt{\frac{9}{10}}$ $sinx = \frac{3}{\sqrt{10}}$

$cos^2x = 1 - sin^2x$

$cos^2x = 1 - \frac{9}{10} = \frac{1}{\sqrt{10}}$

V= ($\frac{1}{\sqrt{10}}$, -$\frac{3}{\sqrt{10}}$) V= (1, -3)
($\frac{3}{\sqrt{10}}$, $\frac{1}{\sqrt{10}}$) (3, 1)

$\Sigma =$ (1, 0)
(0, 2)

tan= ⅓

sin= $\frac{1}{\sqrt{1^2+3^2}} = \frac{1}{\sqrt{10}}$

cos= $\frac{3}{\sqrt{1^2+3^2}} = \frac{3}{\sqrt{10}}$

U= ($\frac{3}{\sqrt{10}}$, -$\frac{1}{\sqrt{10}}$) U= (3, -1)
($\frac{1}{\sqrt{10}}$, $\frac{3}{\sqrt{10}}$) (1, 3)

UΣV= (3, -1)(1, 0)(1, -3)
(1, 3)(0, 2)(3, 1)
= (3, -2)(1, -3) UΣV= (-3, -11)
(1, 6)(3, 1) (19, 3)
UΣV is a transpose of W, reflecting elements on the diagonal.
c)
Wt= (0.9, 0.7)
(-1.7, 0.9)
WtW= (0.9, -1.7) (0.9, -1.7)
(0.7, 0.9) (0.7, 0.9)
WtW= (1.3-λ, -0.9)
(-0.9, 3.7-λ)

(1.3-λ)(3.7-λ)-0.81
=4.81-1.3λ-3.7λ+$\lambda^2 - 0.81$
4.81-5λ+$\lambda^2 - 0.81$
$\lambda^2 - 5\lambda + 4$
Using the quadratic equation:
$\frac{5\pm\sqrt{25-16}}{2} = \frac{5\pm\sqrt{9}}{2}$
λ2= 4, λ1= 1
For λ2= 4:
(-2.7, -0.9)(v_{21})= 0
(-0.9, -0.3)(v_{22})
-2.7v_{21}- 0.9$v_{22} = 0$
-0.9v_{21}-0.3v_{22}= 0
-3v_{21}= v_{22}
$v_{21} =$ (1)
(-3)

For λ1= 1:
(-2.7, -0.9)(v_{11})= 0
(-0.9, -0.3)(v_{12})
0.3v_{11}- 0.9$v_{12} = 0$
-0.9v_{11}+2.7v_{12}= 0
v_{11}= 3v_{12}
$v_{11} =$ (3)
(1)

(0.9, -1.7)(x) =(0.9x -1.7y)
(0.7, 0.9)(y) (0.7x + 0.9y)

$y^2 = (0.7x + 0.9y)^2$ $x^2 = (0.9x - 1.7y)^2$

We get an ellipse, in comparison to the unit circle we had originally. We can see the transformation breakdown, notably by the y axis scaling where the new shape is twice as big as the original. The eigenvectors are not knocked off their own span and do not change while the rest of the vectors do. They instead are the scaling factor of the vector in the unit circle in their respective directions.

d)

(0.9*0.9)-(-1.7*0.7)= 0.81+1.19= 2= determinant

We first find the area of the original unit circle through πr^2

$\pi * 1^2 = \pi$

Then we scale it by the determinant: $\pi * 2 = 2\pi$

The relationship is very important, we need the determinant to use as a factor to scale the new area. According to the determinant which is greater than 1, the new area is stretched. Since det=2, the area is doubled, which we can prove to be correct from our past manual transformations in part b.

Multiplying two matrices is composing the linear transformations, and the determinant represents the scaling factor by which the volume shape will be affected. The determinant of the product shows how much the combined transformation affects space.

3) Some Practices

a) $(\mathbb{E}[X^3])^2 \leq \mathbb{E}[X^6]$.

From the slides, we know that $E[X] = \Sigma x p_X(x)$ for a discrete random variable X

$(E[X^3])^2 \leq E[X^6]$

$(E[X^3])(E[X^3]) \leq E[X^6]$

$(\Sigma x^3 p_X(x))(\Sigma y^3 pX_X y)) \leq \Sigma x^6 p_X(x)$

For sequence of non-negative real numbers, product of their sums greater than/equal the sum of their products

$\Sigma x^3 \Sigma y^3 p_X(x) p_X(y) \leq \Sigma x \Sigma y\ x^6 p_X(x) p_X(y)$

$\Sigma x \Sigma y\ x^6 p_X(x) p_X(y)= E[X^6]$

Ergo we have

$\Sigma x^3 \Sigma y^3 p_X(x) p_X(y) \leq E[X^6]$

$(E[X^3])^2 \leq E[X^6]$

b)

We can use another approach using the variance equation to prove this equation:

$Var(X) = E[X^2] - E[X]^2$

$Var(X^3) = E[X^6] - E[X^3]^2$

We know that variance is always $Var(X) \geq 0$

$Var(X^3) \geq 0 \Rightarrow E[X^6] - E[X^3]^2 \geq 0$

$E[X^6] \geq E[X^3]^2$

= $E[X^3]^2 \leq E[X^6]$

c)

A M positive semi definite matrix is true if z*Mz is strictly positive for any non-zero column vector

χ, $\chi^T A\chi \geq 0$. We can show λA + (1 − λ)B is also PSD for 0 ≤ λ ≤ 1 by contradiction.

Assuming χ^T (λA + (1 − λ)B)x <0, $\chi^T A\chi \geq 0$, $\chi^T B\chi \geq 0$

Using an example linear combination such as:

χ^T (λA + (1 − λ)B)x= λ χ^T Ax (1 − λ) χ^T Bx

Because $\chi^T A\chi \geq 0$, $\chi^T B\chi \geq 0$, the expression above needs to be greater than zero, the expression is a combination of two not negative values.

Our initial theory is that there exists a vector where χ^T (λA + (1 − λ)B)x <0, where the left hand side of our expression is actually negative. This contradicts our theory since the properties of PSD are not respected. Hence, we stand corrected since we need positive values for a correct expression.

4) Density estimation of multivariate Gaussian

a)

Mean ≈ [0.019, 2,061]

Covariance ≈ [[1.036, 2.0624]
[2.062, 5.088]]

b)

See notebook

c)

We can see a Gaussian distribution depending on the bin interval that's inputted.
We calculated the mean and variance and the data seems to align with the original, although we can see a bit of asymmetry and irregularities in the graphs

d)

See notebook- the difference is mainly caused by the fact that we individually sampled each point randomly from the original data, yet still used the same mean for both coordinates

e)

See notebook

f)

See notebook

5) KNN for Iris flowers classification

a)

Class 0: 50 elements
Class 1: 50 elements
Class 2: 50 elements

b)

100% accuracy, this is not too meaningful since, while the model is perfectly trained for the data, it may not be able to interact with other data. To solve this, we should split data to create a training set and a testing set

c)

According to the code, the best k value is 9

d)

The predicted class was found to be versicolor

6) K means clustering

a)

k=4

b)

Number of observations in each cluster:

0 25

1 25

2 25

3 25

Value of inertia: 4844.925817623822

c)

from the graph, we can see that there is quite some overlap near (2, 3)

i've analyzed different centers, higher or lower than k=4

4 was a good choice of value as the clusters are more compact

there is slight overlap as mentioned although those are mainly with the ones with little concentration

a lower value shows more spread data with more outliers showing that the value does not encapsulate all the data

A higher value shows similar results although the clusters seen are less highlighted (more blue)

this would mean that the clusters don't seem to be too related to each other