Data Science: EN.553.436/636
Midterm 1

(attempt 2)

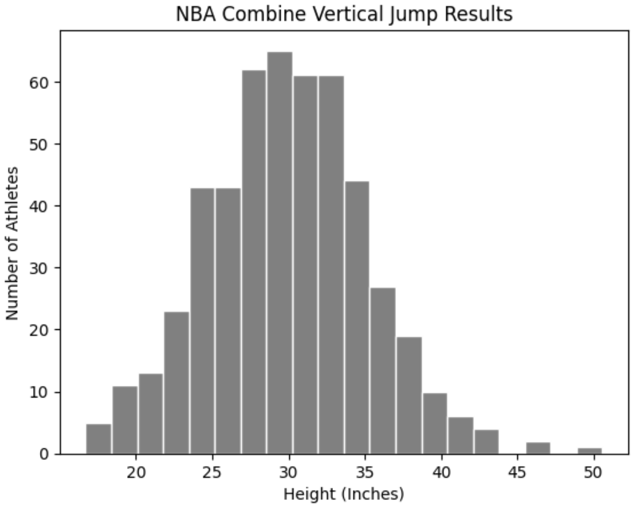
Name:
JHED:

Instructions

- You will have 70 minutes to complete this exam.
- The exam has 3 problems, each having multiple parts.
- You are allowed a single (front and back) note sheet.
    - You must turn in the note sheet along with your exam.
- No notes (other than your note sheet), books, calculators, phones, laptops, or any sort of electronic devices are allowed during the exam.
- Do not look turn the page before the exam begins. Doing so will result in a penalty.
- Good Luck!

Problem 1 (14 pts)

The vertical jump heights were collected from 500 individuals trying out for a basketball team. The samples were plotted in the histogram below.



For the rest of problem 1 the following packages have been imported.

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import norm as gaussian
```

Now consider the following code block. It should look very familiar.

```python
domain = np.linspace(X.min(), X.max(), 500)

g = gaussian(0,2)

variable1 = g.pdf((domain - X[:,np.newaxis])).mean(axis=0)

print(domain.shape, variable1.shape)
pd.DataFrame(variable1).describe()
```

```
(500,) (500,)
                0
count  500.000000
mean     0.029306
std      0.025718
min      0.000471
25%      0.005073
50%      0.021490
75%      0.054005
max      0.072641
```

a) (3 pts)

Describe what the code block above is doing. Be specific, explicitly comment on *variable1* and any other important values that have been coded in.

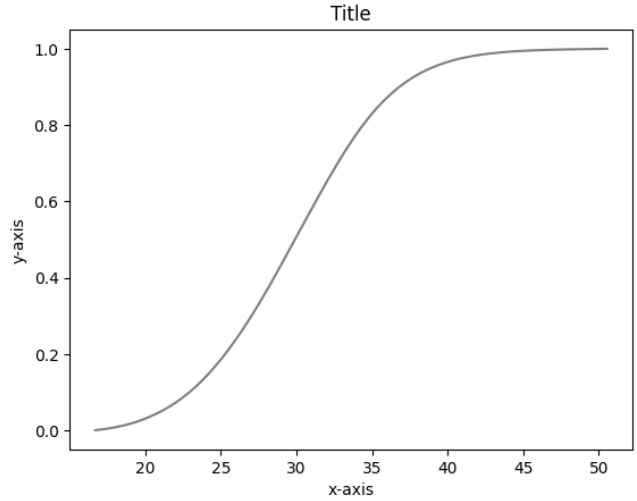
1.a) answer:

b) (6 pts)

The following additional analysis was done. In the appropriate answer box describe what *variable2* represents and why one might want to do these extra calculations.

```
variable2 = np.cumsum(variable1)/np.sum(variable1)

plt.plot(domain,variable2, color = "grey")
```



1.b) answer:

c) (3 pts)

Label the graph created in (b). That is provide an appropriate label for the x-axis, y-axis, and a descriptive title.

1.c) answer:

Title:

x-axis:

y-axis:

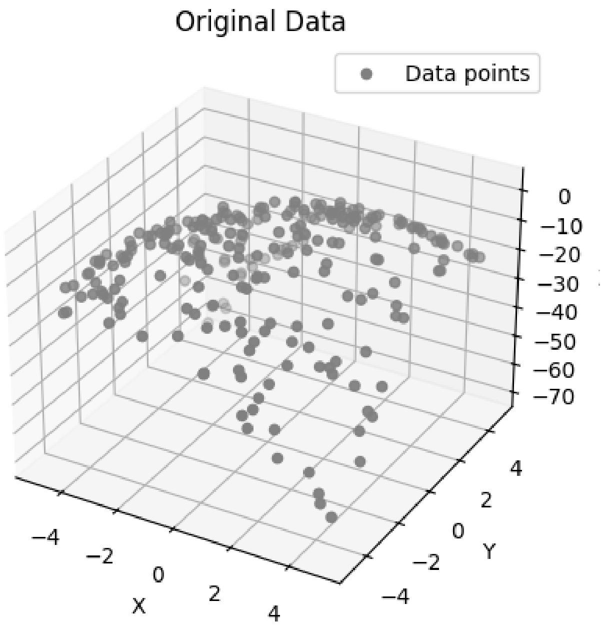
d) (2 pts)

Given the data, what is the probability that somebody at the tryout jumps **at least** 35 inches?

1.d) answer:

Problem 2 (18 pts)

For this next problem imagine now you are in charge of understanding some 2-dimensional data. You have 250 samples distributed in the following way:



a) (4 pts)

Now consider the following code to begin our analysis of the data.

```
X = np.ones((y.shape[0],6))
for count,(i,j) in enumerate([(i,j) for i in range(3) for j in range(3) if i+j < 3]):
    print(f"({count}, i: {i}, j: {j})", end = ", ")
    X[:, count] = x**j * y**i
```

```
(0, i: 0, j: 0), (1, i: 0, j: 1), (2, i: 0, j: 2), (3, i: 1, j: 0), (4, i: 1, j: 1), (5, i: 2, j: 0)
```

What is the purpose of this code? Specifically, what does X represent? Be sure to be specific.

2.a) answer:

b) (3 pts)

What is the underlying equation that we believe to be generating the data? You should explicitly state it, using β's as placeholders for coefficients.

2.b) answer:

c) (4 pts)

Below is the next step in our analysis.

```
variable1 = np.linalg.inv(X.T @ X) @ X.T @ z
print(variable1.shape)
print(variable1)
```

```
(6,)
[ 1.005133    0.00745594 -1.00069387 -0.00307571  1.00140294 -1.01249798]
```

Comment on the purpose of this code. Specifically, mention what *variable1* means in the greater scheme of our problem.

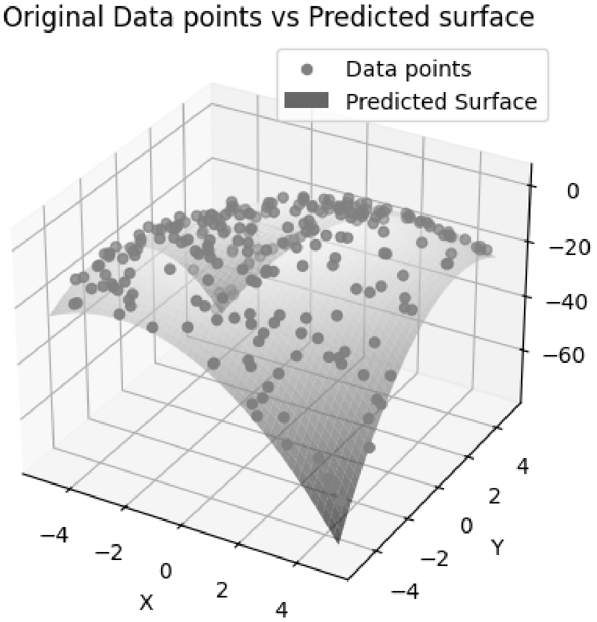
2.c) answer:

d) (3 pts)

Given the previous few parts, what do we believe the generating equation to be?

2.d) answer:

e) (4 pts)



The above image was generated by plotting the original data points against the surface created by the equation from (d). Suppose you guessed the true generating equation exactly (except some homeoscadistic noise term). If you wanted to analyze the noise of the sampled data, how would you do so?. Provide a step by step instruction on how you would recommend to figure out the underlying distribution of the noise.

2.e) answer:

Problem 3 (18 pts)

For the rest of this problem, consider the following context. You were hired by a restaurant to help them improve their customer satisfaction. To do so they have provided a dataset X of n samples, with m features related to different aspects of their service.

a) (7 pts)

Describe the mathematical steps involved in performing PCA, starting from the raw dataset described above up to obtaining the principal components. That is, walk us through step by step the PCA algorithm. (Note: we are not asking for you to write any code just describe the mathematics behind PCA)

3.a) answer:

b) (3 pts)

What do eigenvalues and eigenvectors represent in the context of PCA? How do they help in reducing dimensionality?

3.b) answer:

Suppose you perform PCA on the dataset and obtains 5 principal components. The eigenvalues for each component are as follows:

```
PC1: 7.3
PC2: 1.5
PC3: 0.5
PC4: 0.4
PC5: 0.3
```

c) (2 pts)

Which component captures the most variance?

3.c) answer:

d) (2 pts)

Give an example of a method that can be used to decide on how many principal components to include in the analysis? Hint: maybe include a sketch?

3.d) answer:

e) (4 pts)

Suppose the restaurant does decide to visualize the data using only the first two principal components. What percentage of the total variance will this visualization represent? Discuss potential trade-offs of this decision.

3.e) answer:

MAKE SURE YOU HAVE WRITTEN YOUR NAME ON THE FRONT PAGE BEFORE SUBMITTING!