NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

MW 4:30 – 5:50, NFN 203

Slide Set #18
K-Nearest Neighbors

Lecture Outline

- Data Cleaning
- K Nearest Neighbors Classifier
- Jupyter notebook example

Data Cleaning

- Issues:
  - Irrelevant Columns
    - These can be fully dropped or filtered out
  - Missing or Invalid Values
    - Strategies for missing/invalid values:
      - **<u>Delete rows with missing/invalid values</u>**
      - Impute the missing data
        - E.g., replace with a mean value for the column or for a subgroup of the column associated with a particular row
      - Use a classification or regression model to predict missing values

Simple Example for Cleaning

```
import pandas as pd
import numpy as np
```

```
data = {'sex':['F','M','M','M','F','M','M','M','M','M'],
        'heart_rate':[0,82,75,0,66,69,0,58,73,90],
        'age':[62,51,28,43,17,30,70,21,21,60],
        'bmi':[26, 27, 32, 31, 25, 0, 33, 28, 27, 0]}
df = pd.DataFrame(data)
df
```

	sex	heart_rate	age	bmi
0	F	0	62	26
1	M	82	51	27
2	M	75	28	32
3	M	0	43	31
4	F	66	17	25
5	M	69	30	0
6	M	0	70	33
7	M	58	21	28
8	M	73	21	27
9	M	90	60	0

Occurrences of Invalid Values

We would retain 70% and 80% of our data respectively if we filtered on only one of the features in question.

If we filter on the two combined, we lose 50% of our entries.

If we are training a model using only one or the other of these features as a predictor, we do not need to remove entries

```
df['heart_rate'].value_counts()[0]
```

```
3
```

```
df['bmi'].value_counts()[0]
```

```
2
```

```
mask = ((df['heart_rate'] > 0) & (df['bmi']>0))
mask
```

```
0    False
1     True
2     True
3    False
4     True
5    False
6    False
7     True
8     True
9    False
dtype: bool
```

```
mask.value_counts()
```

```
False    5
True     5
Name: count, dtype: int64
```

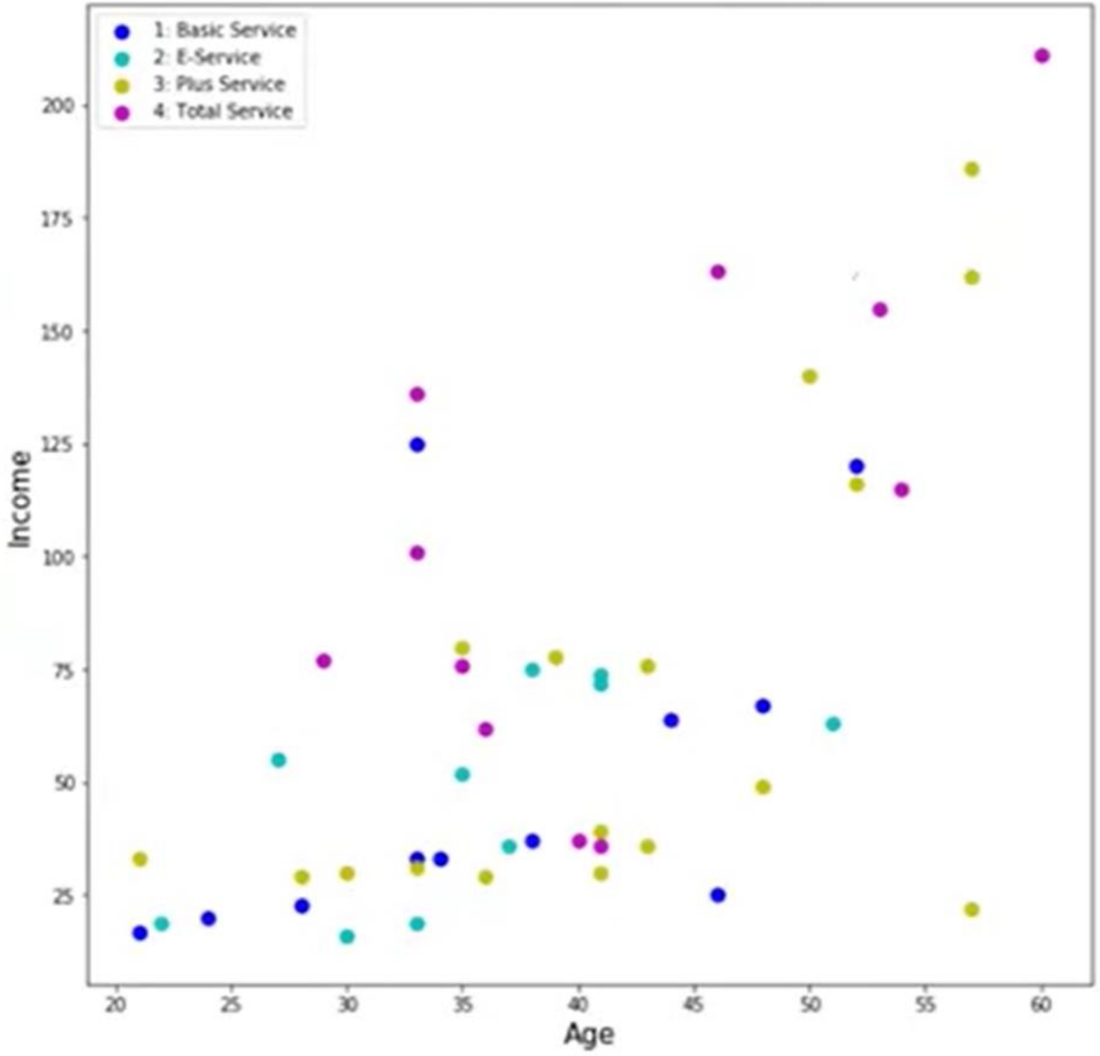
Fully Filtered Dataframe

```
df2 = df[mask]
df2
```

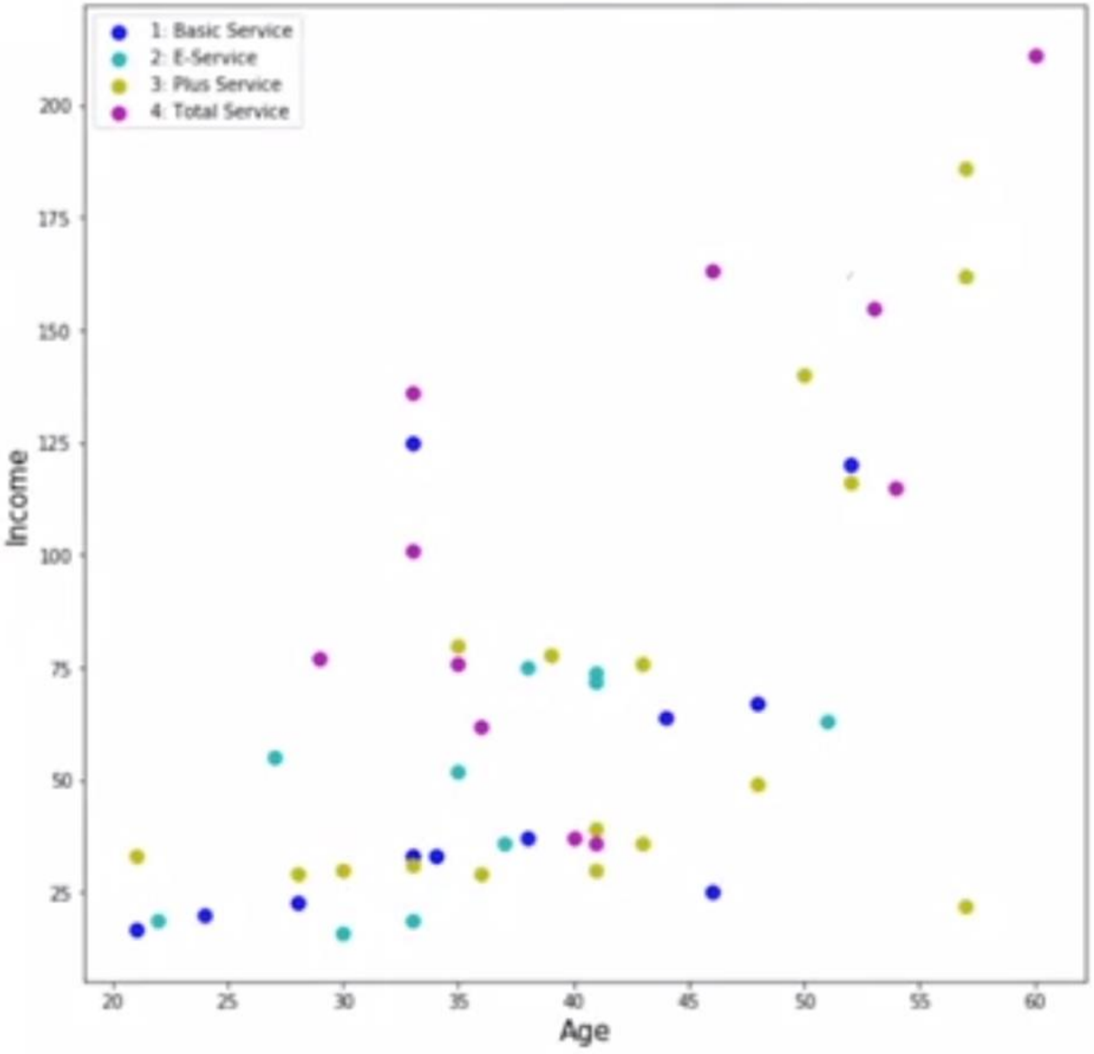
	sex	heart_rate	age	bmi
1	M	82	51	27
2	M	75	28	32
4	F	66	17	25
7	M	58	21	28
8	M	73	21	27

K - Nearest Neighbors

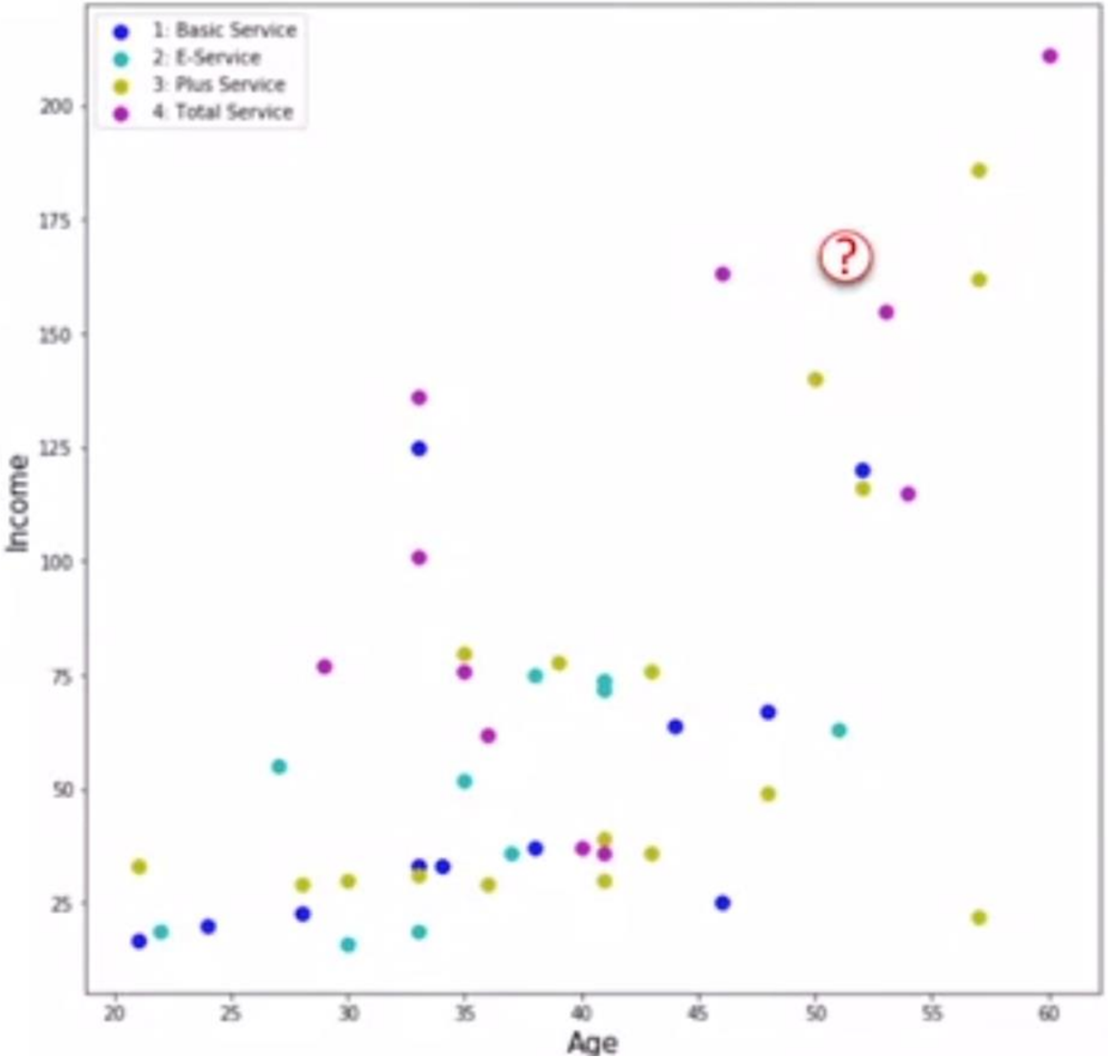
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



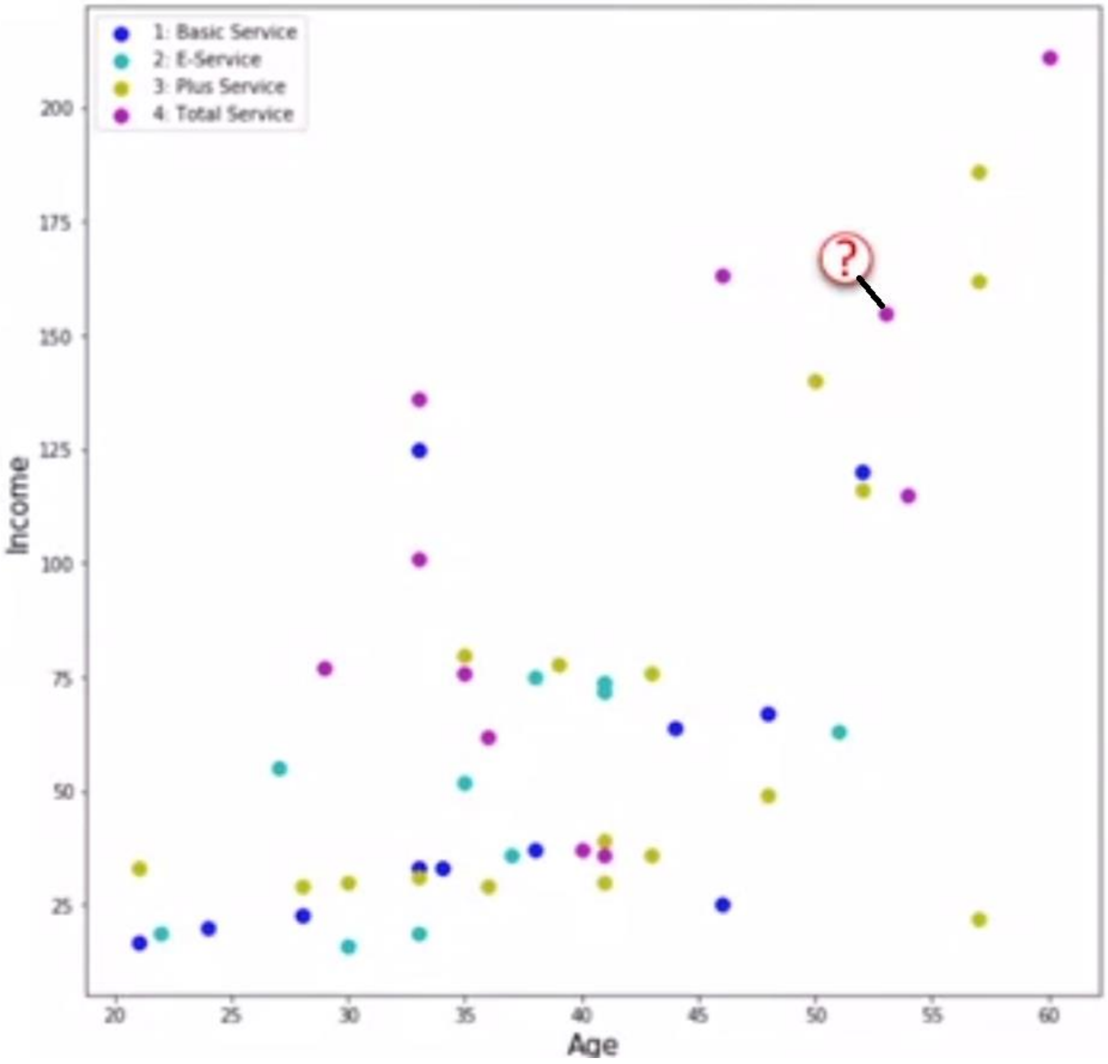
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



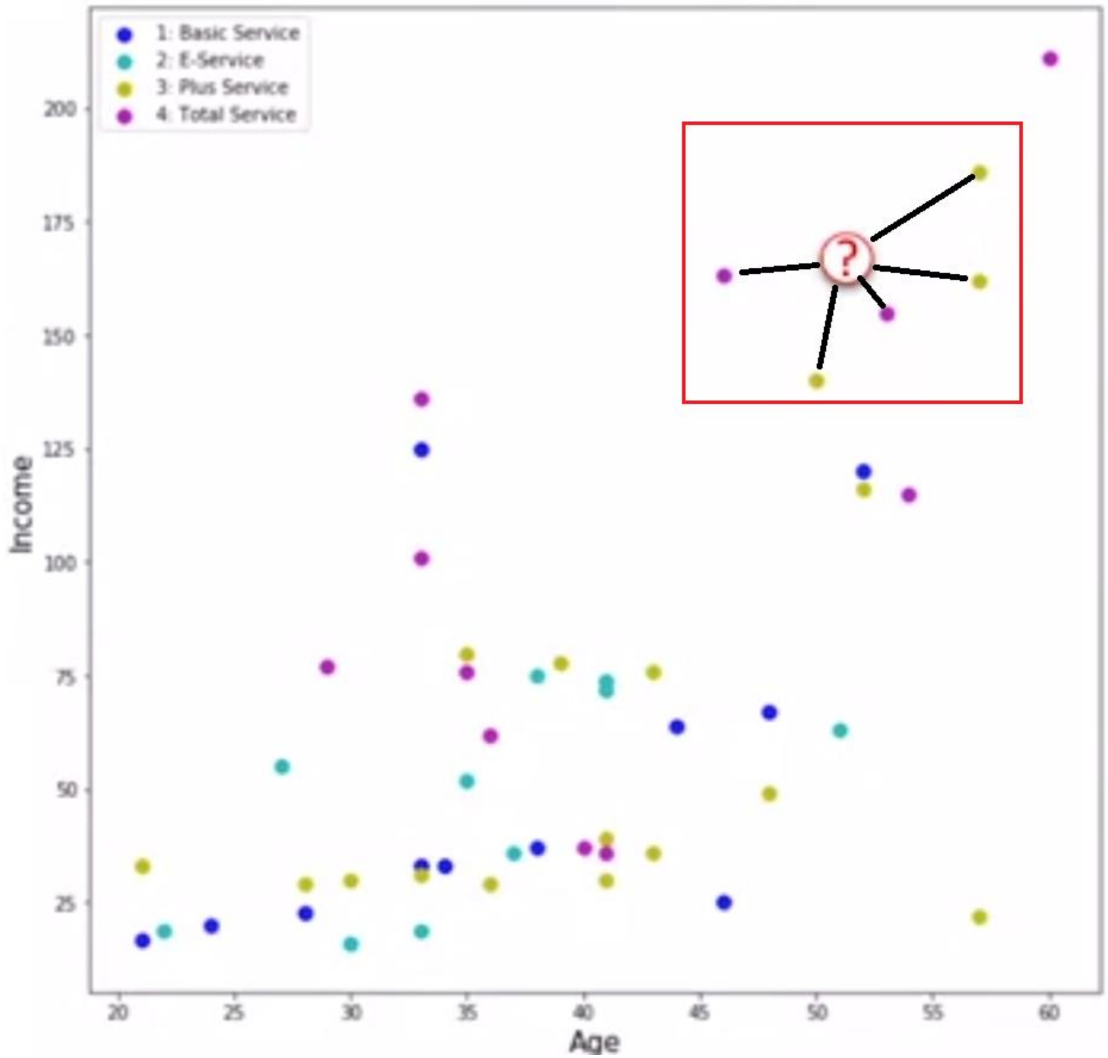
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



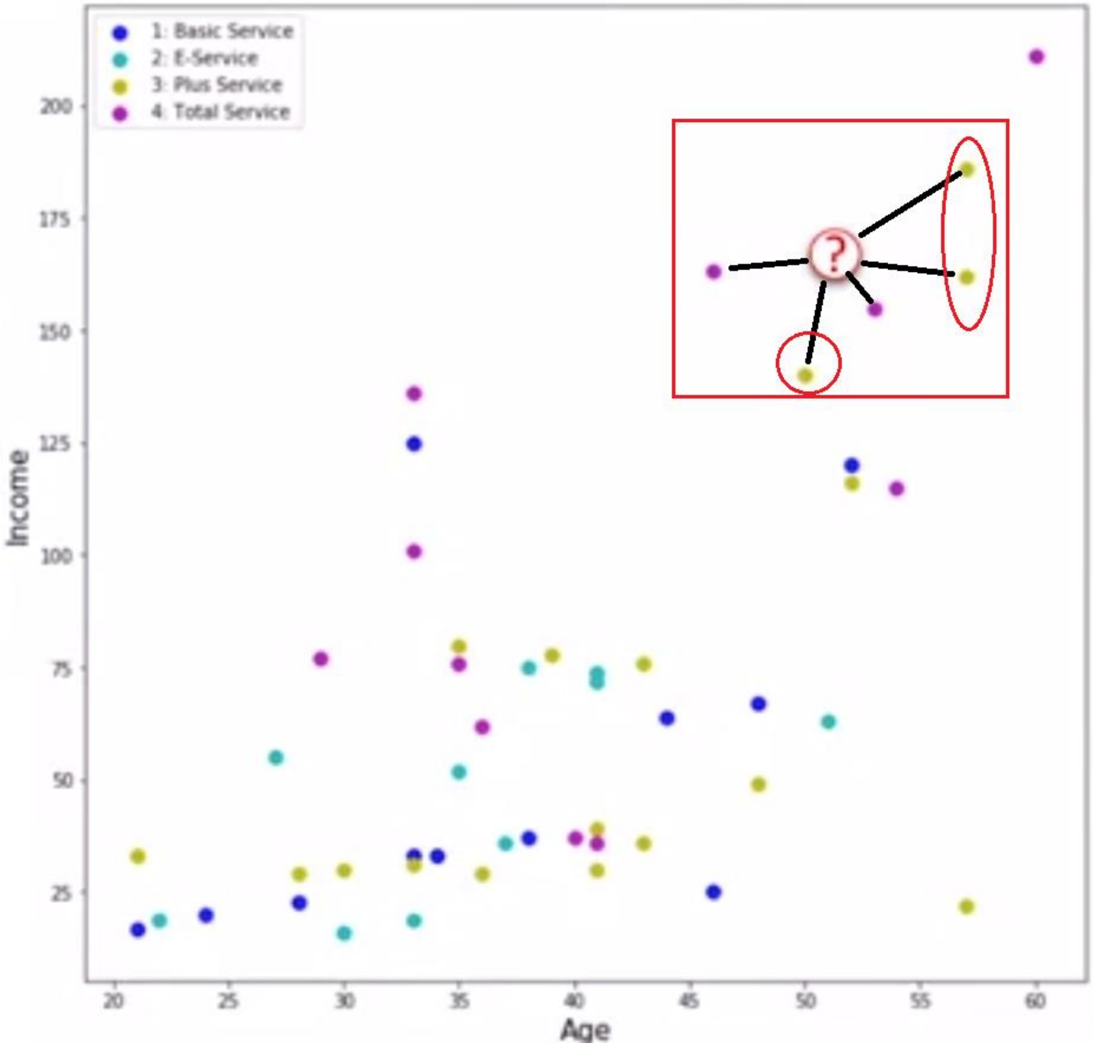
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



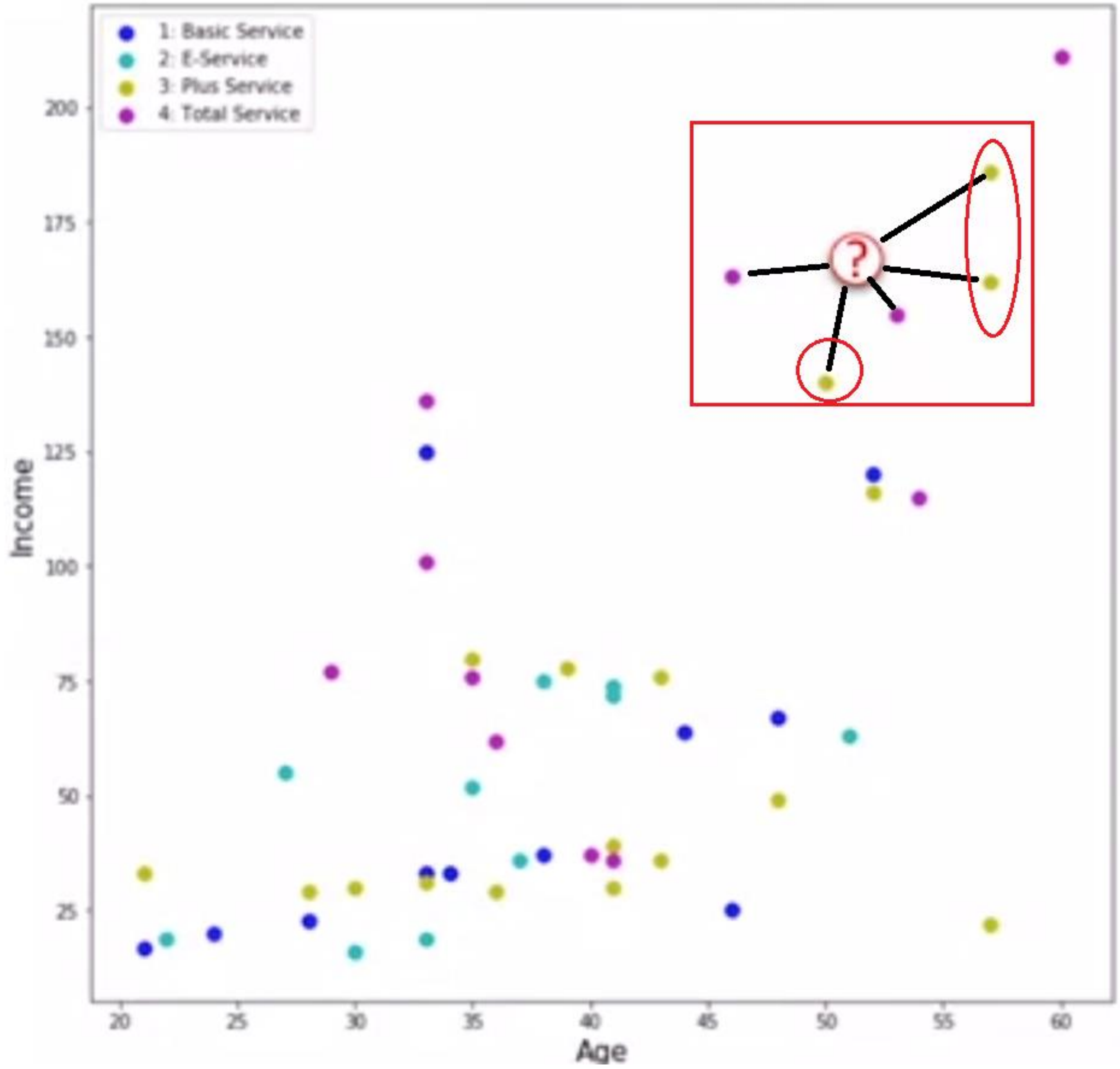
5-NN ⟶ Plus Service (3)

↑

K

What is KNN?

- A method for classifying cases based there on similarity to other cases
- Cases that are near each other are said to be "neighbors"
- Based on similar cases with same class labels are near each other



The K-Nearest Neighbors Algorithm

1. Pick a value for K
2. Calculate the distance of unknown case from all cases
3. Select the K-observations in the training data that are “nearest” to the unknown data point
4. Predict the response (“category”/”class”) of the unknown data point using the most common response value from the K-nearest neighbors

Calculating Distance in 1-Dimensional Space

Customer 1
Age
34

Customer 2
Age
30

Euclidean Distance

Customer 1
Age
34

Customer 2
Age
30

$$\text{Dis}\ (x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

$$\text{Dis}\ (x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

Same for Multidimensional Space

Customer 1		
Age	Income	Education
34	190	3

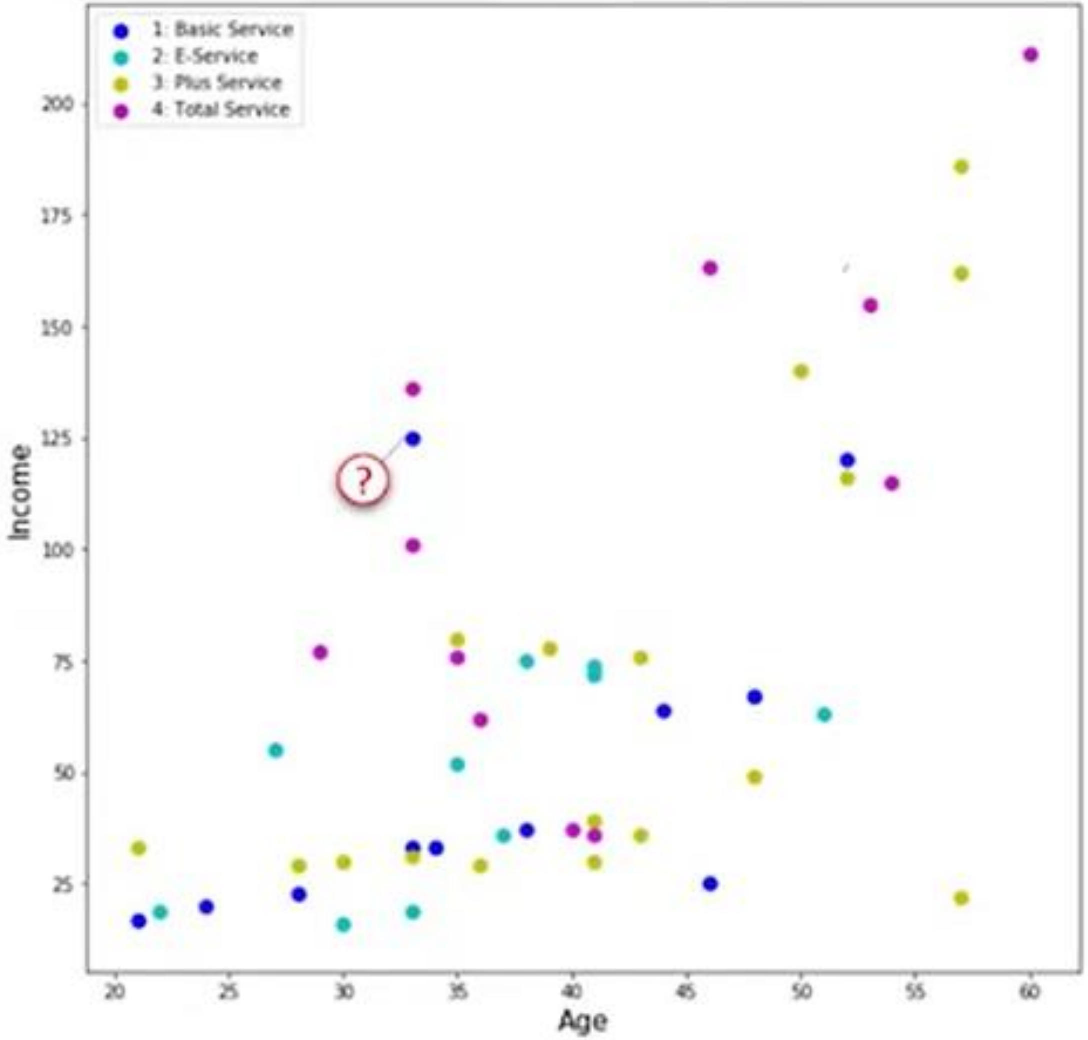
Customer 2		
Age	Income	Education
30	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\textstyle\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(34-30)^2+(190-200)^2+(3-8)^2} = 11.87$$

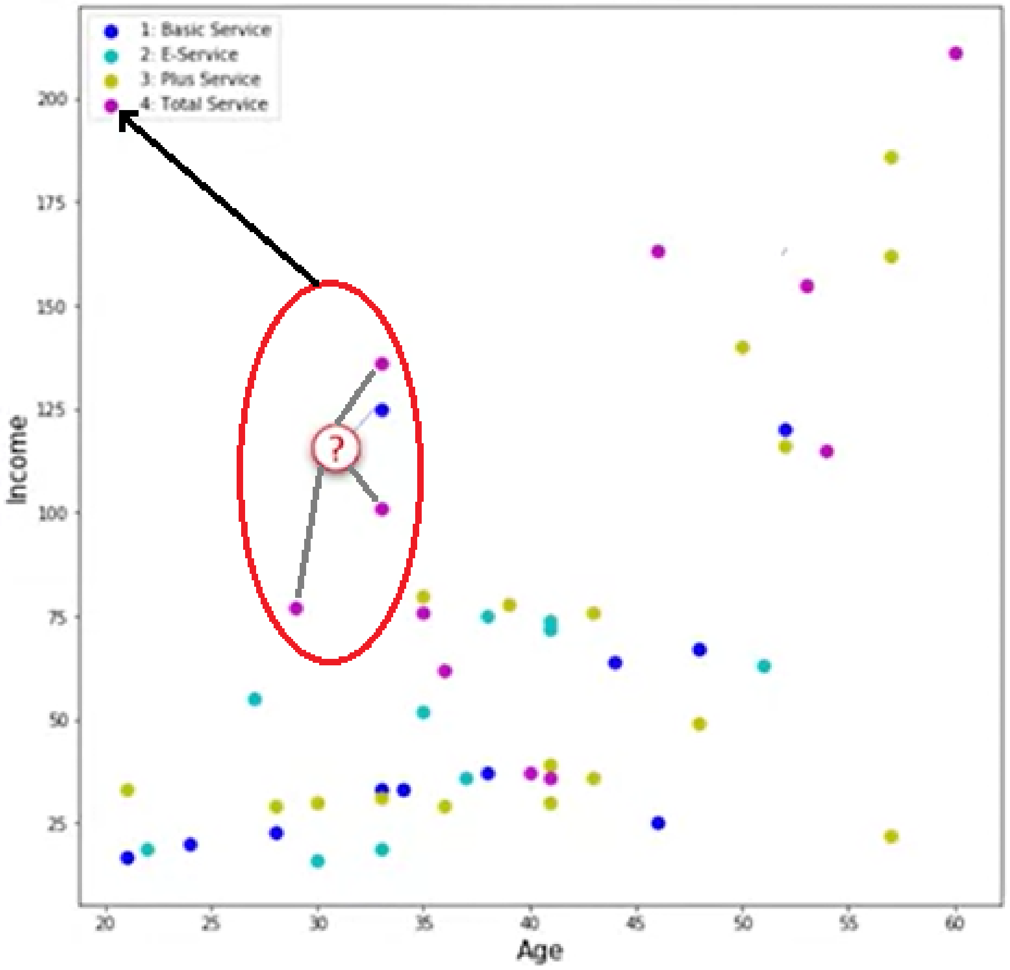
Picking Value of K

- K =1 class 1



Overfit

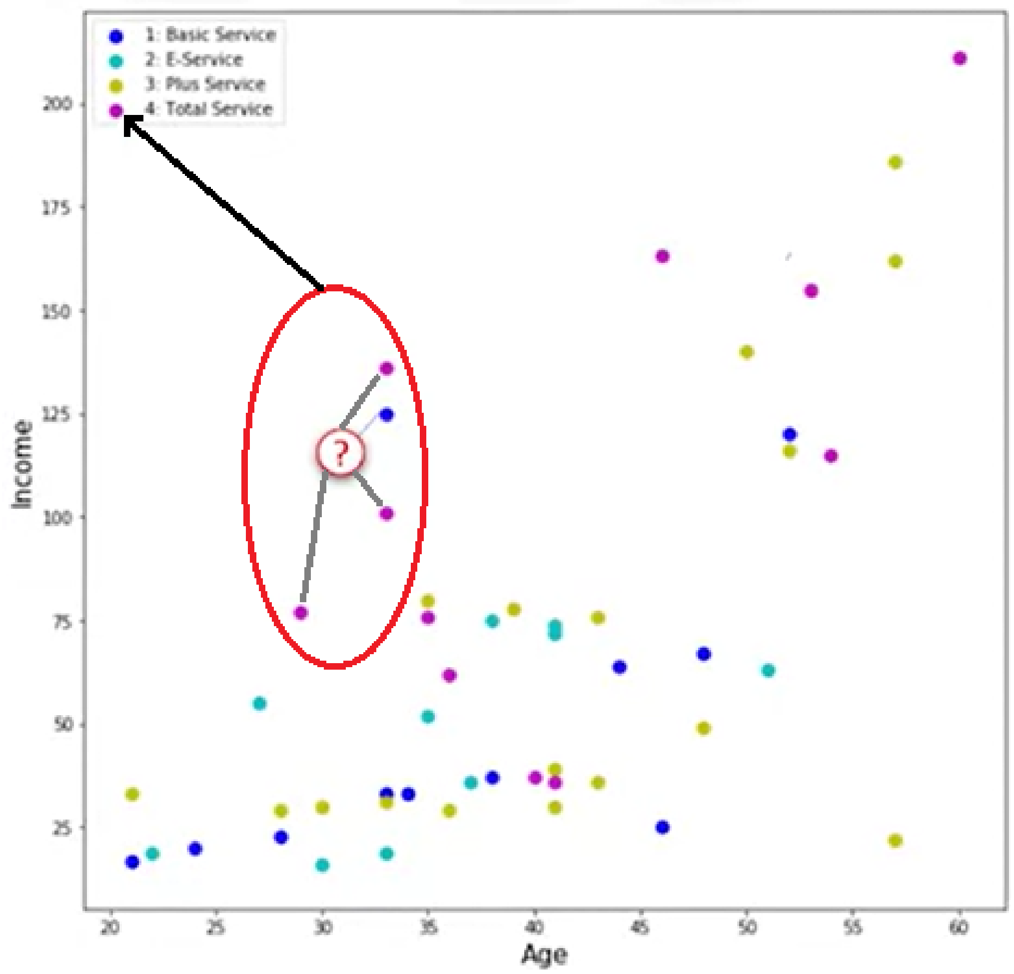
- K =1 class 1



Overfit vs Over Generalized

- K =1 class 1

A very high value such K = 20 would also be bad as it would over generalize the model.



What is the Best Value of K?

K = ?