# [AAA] Advanced Analytics and Applications

Summer Semester 2021

## Problem Set 1

### a) Multiple Choice Questions

1. Astronomers have been cataloguing distant objects in the sky using long-exposure CCD images. The objects need to be labeled as star, galaxy, nebula etc. The data is highly noisy, and the images are very faint. The cataloguing can take decades to complete. Which Method would you recommend the physicists use to automate the cataloguing process, and improve its effectiveness?
   - ☐ Clustering
   - ☐ Classification
   - ☐ Regression
2. A customer of a consultancy owns a supermarket. Through membership cards, there is basic data about the supermarket customers like Customer ID, age, gender, annual income and spending score. Spending Score was assigned to the customers based on defined parameters like customer behavior and purchasing data. Which method would you recommend using so the consultancy can make a suggestion to the supermarket which of his customers are most likely to react positively to a marketing campaign?
   - ☐ Clustering
   - ☐ Classification
   - ☐ Regression

3. A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company. Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away. Which method would you recommend using to model the probability of attrition where the results can be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay?
   - ☐ Clustering
   - ☐ Classification
   - ☐ Regression

# b) Calculation

| $y$ | $\hat{y}$ | $(y-\bar{y})^2$ | $(\hat{y}-\bar{y})^2$ | $y-\hat{y}$ | $(y-\hat{y})^2$ |
|---|---|---|---|---|---|
| 5 | 5 | | | | |
| 12 | 6 | | | | |
| 16 | 8 | | | | |
| 5 | 6 | | | | |
| 5 | 5 | | | | |

## 1.1 **Revisit MSE**

a) Calculate MSE using the values from the above table (y = true value, $\hat{y}$ = predicted value)

b) When is MSE useful?

## 1.2 **Revisit R-Squared** (R2)

1. Calculate R-Squared using the values from the above table (y = true value, $\hat{y}$ = predicted value)

b) When is R-Squared useful? Compare it to MSE.

c) (BONUS) Derive formula for R-Squared

# c) Programming

1. (Pandas) Create a python notebook that loads the *employee_survey_data.csv* into a Data Frame and find out the following stats:
   (1) Dimensionality: _____
   (2) Sum of rows with missing values: _____
   (3) Number of employees most satisfied with their jobs: _____
   Tip: Find out max value. Filter employees by max value. Sum number of employees with max job satisfaction.

2. (Plotting) In this task, we want to learn how create a plot using the python library matplotlib with multiple axes in one figure. To do so, create a Jupyter Notebook, then download and initialize the California housing data set.

   Then, try to render a plot similar to the one depicted in Figure 1. In the first figure, we can want to render a scatter plot of “age” (Home Age) and “pop” (Population). In the second plot, we want to visualize a histogram of “age” (Home Age). Finally, we want to render a plot a histogram of the “pop” (Population). Follow the template from Figure 2 with 3 axis objects, and by using subplots (for more information, read the docs for matplotlib)

**Data**

```
url =
'http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.tgz'
b = BytesIO(urlopen(url).read())
fpath = 'CaliforniaHousing/cal_housing.data'

with tarfile.open(mode='r', fileobj=b) as archive:
    housing = np.loadtxt(archive.extractfile(fpath),
delimiter=',')
```

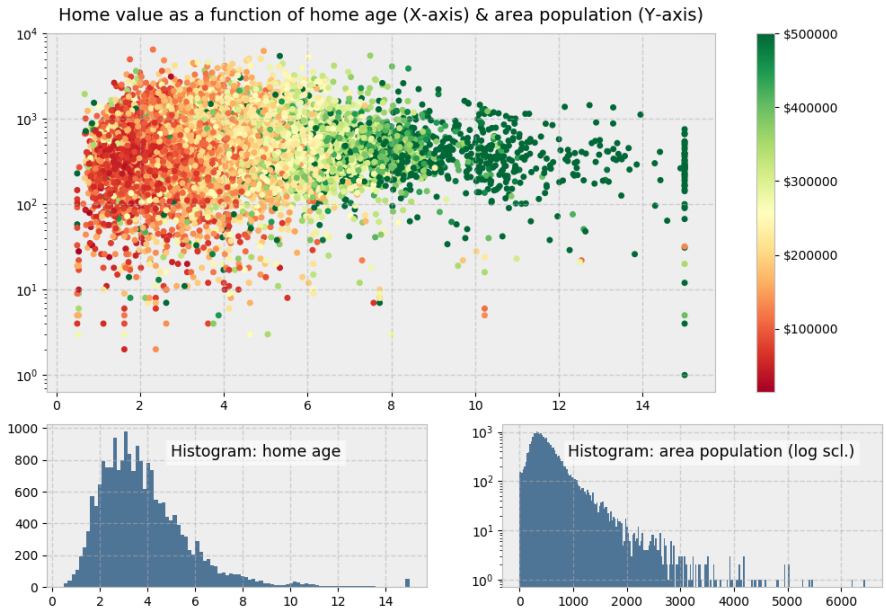


*Figure 1 Home Value Plot*

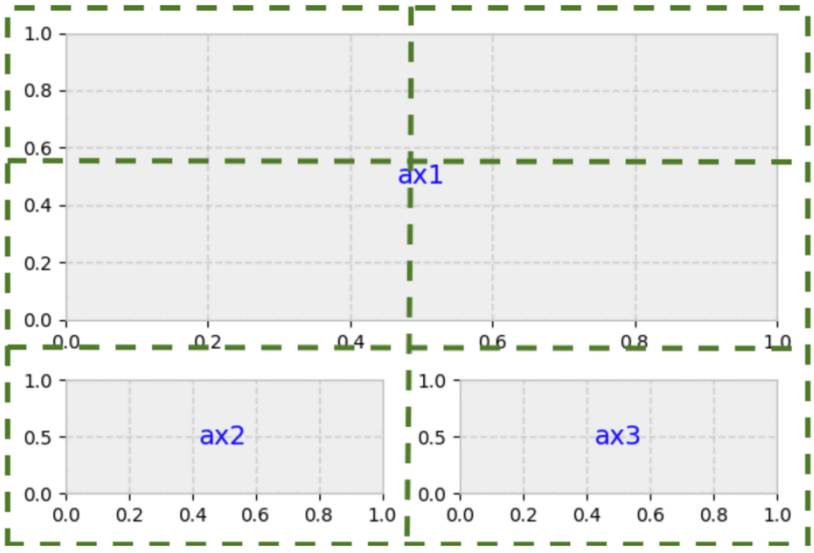


*Figure 2 Plot Structure*