MACHINE LEARNING FOR PUBLIC POLICY
HOMEWORK 1 – CARLOS GRANDET

Problem A

1. The first task is to load the file and generate summary statistics for each field as well as probability distributions or histograms. The summary statistics should include mean, median, mode, standard deviation, as well as the number of missing values for each field.

The summary statistics for the variables that are floats are:

	Mean	Median	S.D
Age	17	16.99611	1.458067
GPA	3	2.988447	0.818249
Days_missed	18	18.01114	9.629371

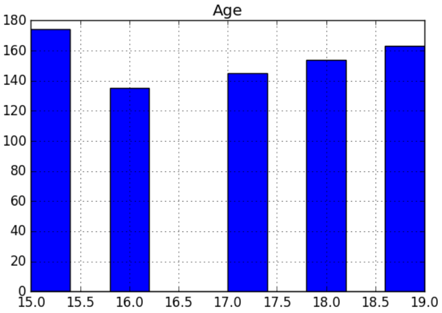
The mode of each variable is:

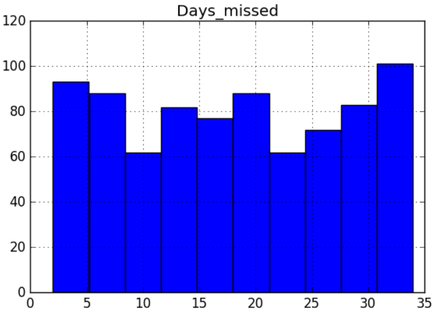
First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
Amy	Ross	Texas	Female	15	2	6	Yes

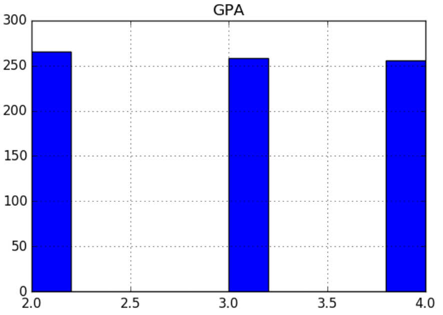
The number of missing values are:

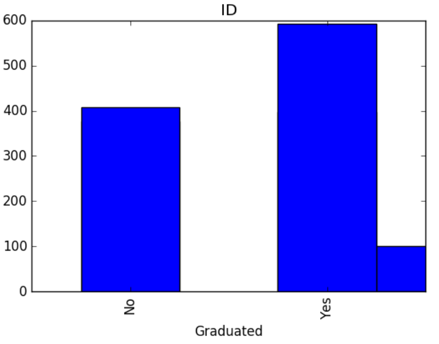
First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
0	0	116	226	229	221	192	0

Histograms for Age, GPA and Days_missed:



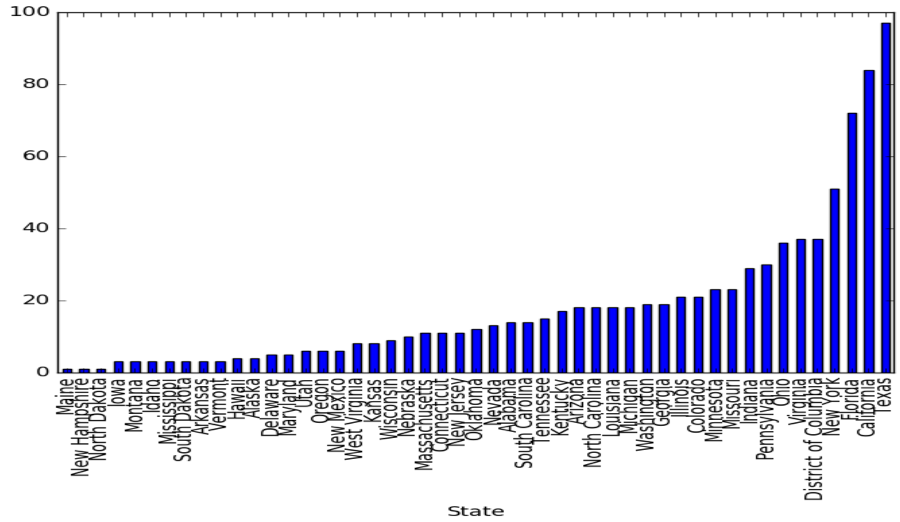


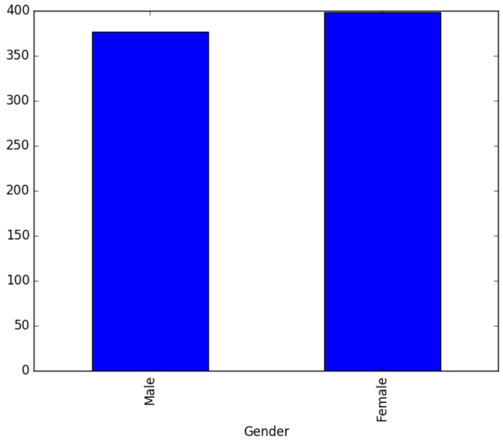


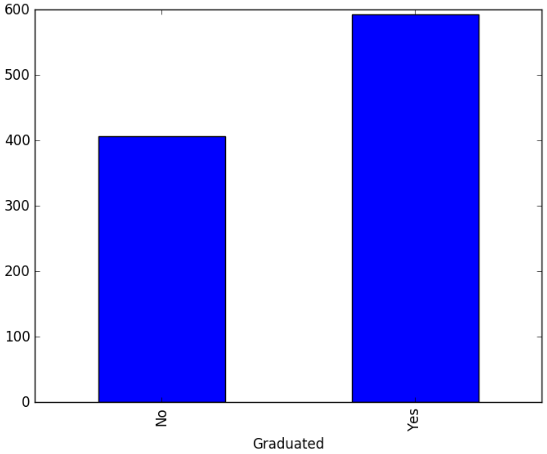


MACHINE LEARNING FOR PUBLIC POLICY
HOMEWORK 1 – CARLOS GRANDET

Histogram for state, gender and graduation







2. You will notice that a lot of students are missing gender values. Your task is to infer the gender of the student based on their name. Please use the API at www.genderize.io to infer the gender of each student and generate a new data file.

The new data file is available at names_predictions.csv

3. You will also notice that some of the other attributes are missing. Your task is to fill in the missing values for Age, GPA, and Days_missed using the following approaches:

A. Fill in missing values with the mean of the values for that attribute
B. Fill in missing values with a class-conditional mean (where the class is whether they graduated or not).
C. Is there a better, more appropriate method for filling in the missing values? If yes, describe and implement it.
I added a condition on gender using the new data file from problem 2. Therefore, I group based on gender and graduation and obtained the mean for these groups to obtain a more granular dataset.

The new data files are available at model A.csv, model B.csv and model C.csv respectively

Problem B

A larger data set than the one in the previous problem was used to build a logistic regression model that predicts the probability an individual student will graduate. Below are coefficients from this model. The definitions of the variables are below.

Variable	coefficient	standard error	z score
Male	1.45	0.09	15.81
Female	-2.11	0.1	-21.95
Ln_Family_Income	-0.109	0.041	-2.63
Age	-0.013	0.0096	-1.4
Age_Sq	0.0001	0.00009	1.52
Census_College	1.77	0.33	5.37
AfAm	2.07	0.399	5.18
AfAm_Male	-0.872	0.437	-2.01
Constant	1.2	0.484	2.48

Male—Coded as 1 if the student is M, 0 if he/she is not Female—Coded as 1 if the student is Female, 0 if he/she is not

Ln_Family_Income—The natural logarithm of the student's Family income (in dollars)

Age—The student's age (in years)

Age_Sq—The student 's age (in years) squared

Census_College—The percentage of residents in the student 's neighborhood who have a college degree (scaled from 0 to 100)

AfAm—Coded as 1 if the student is African American, 0 if he/she is not AfAm_Male—Coded as 1 if the student is both African American and Male, 0 if he/she is not

A. Consider 4 students, Adam, Bob, Chris and David. Adam and Chris share identical characteristics except for their family incomes. Bob and David also share identical characteristics (with each other, not necessarily Adam and Chris), except for their incomes.

Name	Family Income	Modeled Probability of Graduation
Adam	$50,000	50%
Bob	$200,000	50%
Chris	$40,000	?
David	$190,000	?

Based on the coefficients above, who would you think has a higher probability of graduating?

According to the coefficients above, Chris has a higher probability of graduation. This is because larger family income decreases the probability of graduation by a negative logarithmic factor. If I double my income, the probability of graduation will be multiplied by a 2^(-.109)=.92 factor (which means a decrease in the probability), if I halved my income, it will be multiplied by a .5^(-.109)=1.07 factor. The z-statistics shows that this coefficient is significant to 99.5% percent.

MACHINE LEARNING FOR PUBLIC POLICY
HOMEWORK 1 – CARLOS GRANDET

The coefficient for AfAm_Male is negative. How do you interpret this? Does this mean that African-American Males are more likely to not graduate than African-American Females? What about relative to non African American males?

In order to know that you will need to know who is your reference group. If your reference group is non-African American males, then the negative coefficient in AfAm_Male means that they are less likely to graduate in reference to African-American females. Relative to non-African American male you will need to weight the impact of being African American which appears to be positive with the effect of being African American male, depending on the direction of the coefficient, it could be more or less probable to graduate. In this case e^-.872*e^2.07 = 3.3, therefore there is a positive probability.

How do we interpret the difference in graduation probability between students of different ages? How do the variables in the model estimate such probability?

The coefficient is telling us that as age increases, the probability of graduation decreases not in a linear way. Furthermore, since the effect of age squared is positive, it means that it is decreasing at a faster rate and then it will reduce the effect, following a pattern similar to a function 1/x. The variables in the model are estimating this using a squared variable to calculate this non-linear behavior

Are there any variables in this model that you would choose to drop? Why or why not? Would you need more information in order to make this decision?

If you analyze the z-scores, the only variables not significant to 5% are age and age2. I wouldn’t drop them without first doing an f-test with the restricted and unrestricted model to determine the model general f-value. I would also need to consider endogeneity. My intuition is that dropping those variables will create endogeneity since the age of the individual might be correlated with the family income, therefore violating the non-endogeneity assumption for logistic regression.