Dani Litovsky Alcala - ML ps1

Let's start exploring the student data

Problem A

A.1

Here are summary statistics (mean, median, mode, standard deviation, as well as the number of missing values for each field) and histograms.

Colomn names:

```
array(['ID', 'First_name', 'Last_name', 'State', 'Gender', 'Age', 'GPA',
       'Days_missed', 'Graduated'], dtype=object)
```

Top of the data file:

	ID	First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
0	1	Wayne	Boyd	Florida	Male	19	NaN	9	Yes
1	2	Ann	Matthews	Pennsylvania	Female	18	3	NaN	Yes
2	3	George	Matthews	Texas	NaN	17	NaN	10	Yes
3	4	Jerry	Ramos	California	Male	15	2	28	No
4	5	Andrea	Carroll	North Carolina	Female	NaN	2	29	No

	ID	First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
count	1000.000000	1000	1000	884	774	771.000000	779.000000	808.000000	1000
unique	NaN	200	244	49	2	NaN	NaN	NaN	2
top	NaN	Amy	Ross	Texas	Female	NaN	NaN	NaN	Yes
freq	NaN	12	13	97	398	NaN	NaN	NaN	593
mean	500.500000	NaN	NaN	NaN	NaN	16.996109	2.988447	18.011139	NaN
std	288.819436	NaN	NaN	NaN	NaN	1.458067	0.818249	9.629371	NaN
min	1.000000	NaN	NaN	NaN	NaN	15.000000	2.000000	2.000000	NaN
25%	250.750000	NaN	NaN	NaN	NaN	16.000000	2.000000	9.000000	NaN
50%	500.500000	NaN	NaN	NaN	NaN	17.000000	3.000000	18.000000	NaN
75%	750.250000	NaN	NaN	NaN	NaN	18.000000	4.000000	27.000000	NaN
max	1000.000000	NaN	NaN	NaN	NaN	19.000000	4.000000	34.000000	NaN

There are three modes for missed days. Amy is the most common first name and, thankfully, Yes is the most common outcome to graduation.

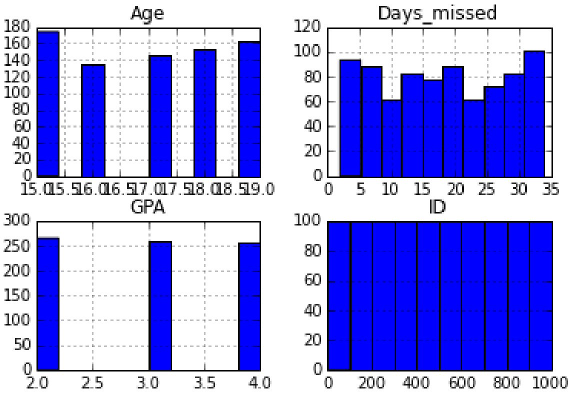
	ID	First_name	Last_name	State	Gender	Age	GPA	Days_missed	Graduated
0	NaN	Amy	Ross	Texas	Female	15	2	6	Yes
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	14	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	31	NaN

There are many missing values.

```
ID                0
First_name        0
Last_name         0
State           116
Gender          226
Age             229
GPA             221
Days_missed     192
Graduated         0
dtype: int64
```

Some histograms to visualize the data

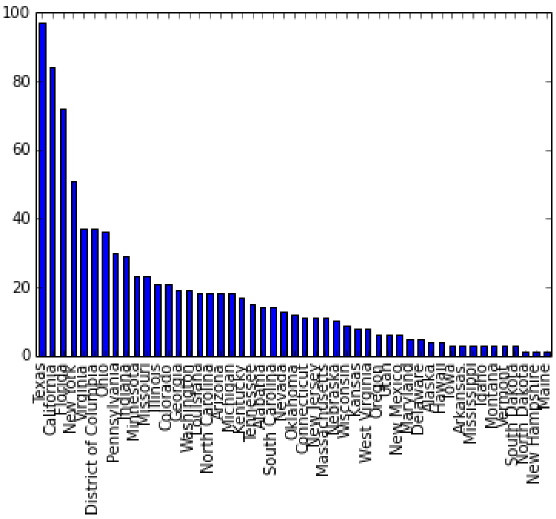
```
array([[<matplotlib.axes.AxesSubplot object at 0x11087fb10>,
        <matplotlib.axes.AxesSubplot object at 0x11097c710>],
       [<matplotlib.axes.AxesSubplot object at 0x1109e1fd0>,
        <matplotlib.axes.AxesSubplot object at 0x110a62e90>]], dtype=object)
```



We can see that GPA is either 2.0, 3.0, or 4.0

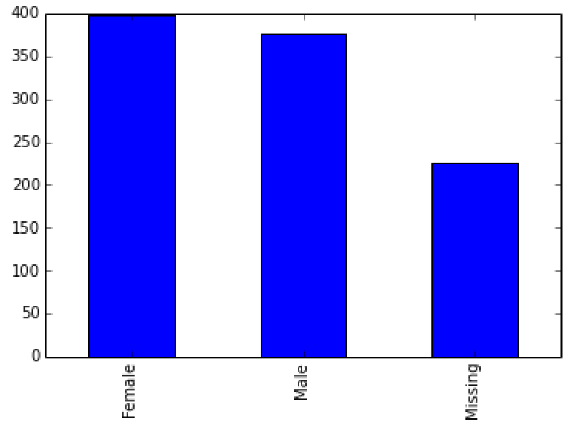
More visualizations for state and graduation

```
<matplotlib.axes.AxesSubplot at 0x110c5e890>
```



```
Female     398
Male       376
Missing    226
dtype: int64
```

```
<matplotlib.axes.AxesSubplot at 0x111319350>
```



We could draw histograms for the remaining categorical data (first and last names) but that would be hard to interpret and not very useful for our purposes.

A.2

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.

A.3

Fill in the missing values for Age, GPA, and Days_missed using the following approaches:

- Fill in missing values with the mean of the values for that attribute
- Fill in missing values with a class-conditional mean (where the class is whether they graduated or not).
- Is there a better, more appropriate method for filling in the missing values? If yes, describe and implement it.

Approach 1: imputing with the mean:

```
ID               0
First_name       0
Last_name        0
State          116
Gender         226
Age              0
GPA              0
Days_missed      0
Graduated        0
dtype: int64
```

Imputing with the mean, we got rid of the missing values for numerical data, but not for state and gender. Note that since we're using the mean, GPA data (which was previously only 2, 3, or 4) will now have the mean of 2.988447.

A better approach would be to fill in missing values with a class-conditional mean (where the class is whether students graduated or not). Note that this will not fill in missing values for State and Gender. I am also using the original dataset, so the previous work of imputing gender does not appear here, to keep tasks separate for now.

Is there a better or more appropriate way to fill missing values?

There are certainly more appropriate, albeit more costly way to impute missing data. I could build statistical / machine learning models that can help predict and classify each individual's age, GPA, and days missed. But for the sake of time, another way to impute missing data would be to take the previous exercise of conditional means and creating means not only based on 'graduation group' (yes or no), but to include more groups to condition on. I will include gender as a group to condition the mean. This is possible since we were able to use the genderize.io API

In general, imputing with means is sensitive to outliers. For GPA data, it may be smarter to simply use the mode of that set because our data only has it

For the missing states data, we could take an approach of imputing it with Texas or California, the most prevalent states in the dataset, but this is also not the most informed choice.

```
Graduated  Gender
No         female    16.967742
           male      17.136364
Yes        female    16.967611
           male      16.948837
Name: Age, dtype: float64
```

Problem B

Let's look at Adam, Bob, Chris and David. Bob and David share identical characteristics (with each other, not necessarily Adam and Chris), except for their incomes, and Adam and Chris share the same characteristics except for income.

Based on the results coefficients, we see that a unit increase in log family income decreases the odds ratio of graduating. Given that Bob and David share identical characteristics, the model indicates that David has a higher (possibly small) probability of graduating since his income is 10,000 dollars less. Likewise, Chris's family income is also 10,000 dollars less, so his probability of graduating is greater than Adam's 50% probability. The question of whether Chris has a higher probability of graduating than David, or vice versa, depends on if a 10,000 dollars change at the 200,000 dollars income is larger or greater than the effect of a 10,000 dollars change at the 50,000 income level. That is, if I am making 200,000 dollars, does decreasing my income increase my odds of graduating more than if I earn 50,000 dollars and decrease my income to 40,000 dollars? This could be calculated mathematically by using the logistic density function, but without it, we cannot certainly determine if Chris' probability of graduating is greater than David's - only that they both most likely have a probability of graduating that is greater than 50%.

We can interpret logistic regression coefficients as the change in the log odds of the outcome for a one unit increase in the covariate. I can interpret the coefficients for the logistic regression model that predicts student graduation as follows:

- The negative coefficient on AfAM_Male means that being an African American male decreases the odds ratio of graduating.
- The difference between AFAm_Male and AfAm is female. That is, If a person is African American but not male, that person is a female.
- African American females are more likely to graduate than African American males and than non-black females.
- If a male is not African American, then his probability of graduating is represented by the coefficient for Male, indicating that being male, unlike being an African American male, makes a person more likely to graduate.
- The coefficient for age indicates that a unit increase in age decreases the probability of graduation by 0.013, while the age squared coefficient indicates that an increase in squared age increases the probability of graduating by 0.0001. Notice that the z scores for both are small, meaning that they are not very statistically significant. Intuitively, a student who is younger, say 12, is of course less likely to graduate than a student who is 17, but the coefficients do not model that very well.
- As a general rule, we look at the z score to decide if to drop a variable because of low statistical significance. I would remove both age variables based on the unintuitive and low-significance results.