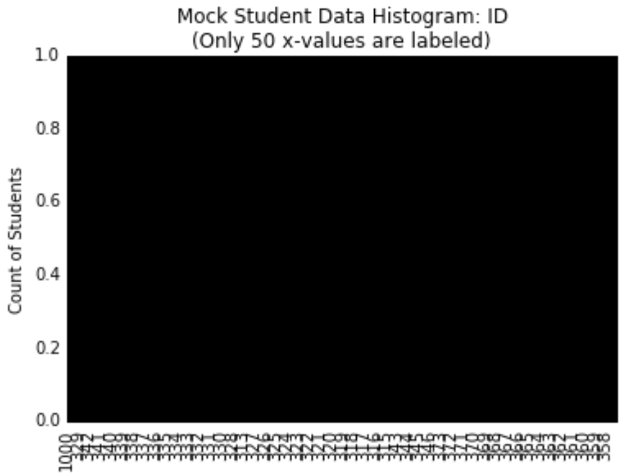
In [1]:
```
# Machine Learning for Public Policy
# Assignment 1: Prepping Student Data
# Name: Vi Nguyen
```

In [1]:
```
%matplotlib inline
```

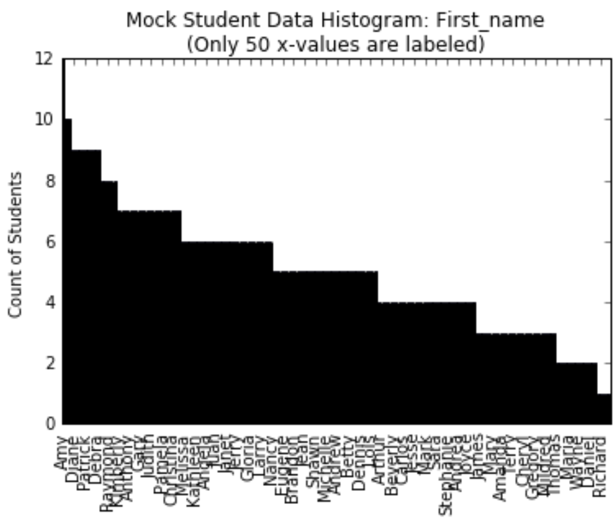
In [ ]:
```
# Problem A #
```

In [2]:
```
import prep_data
```

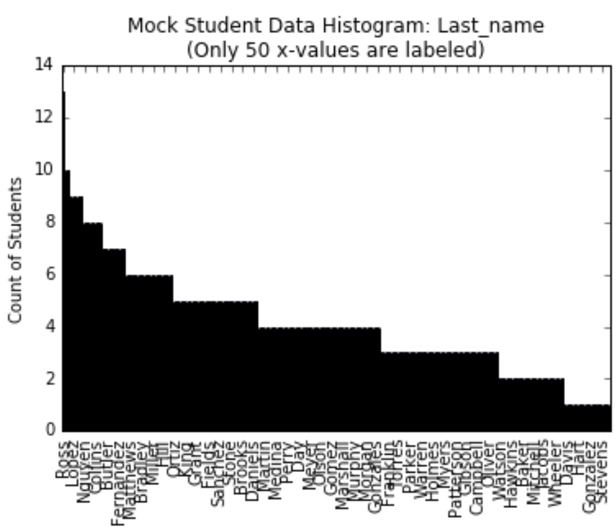
```
hist_ID.png created
```



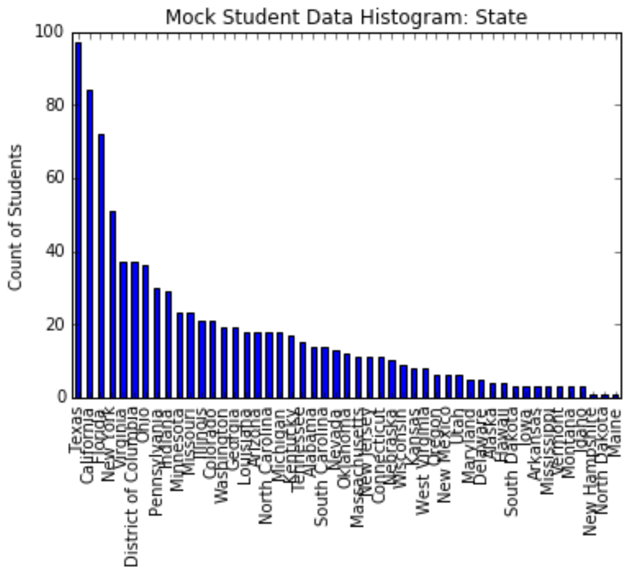

```
hist_First_name.png created
```



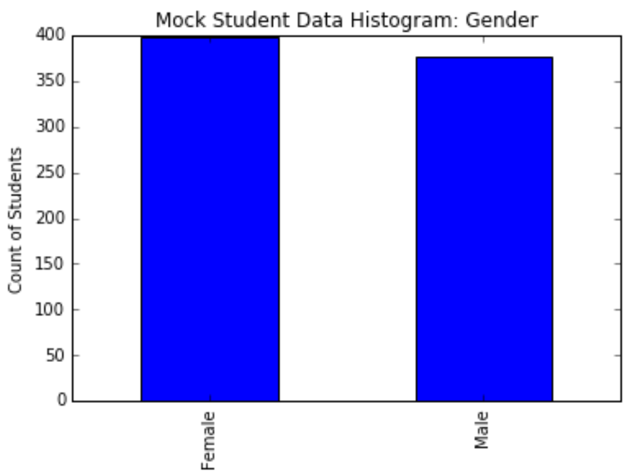

```
hist_Last_name.png created
```



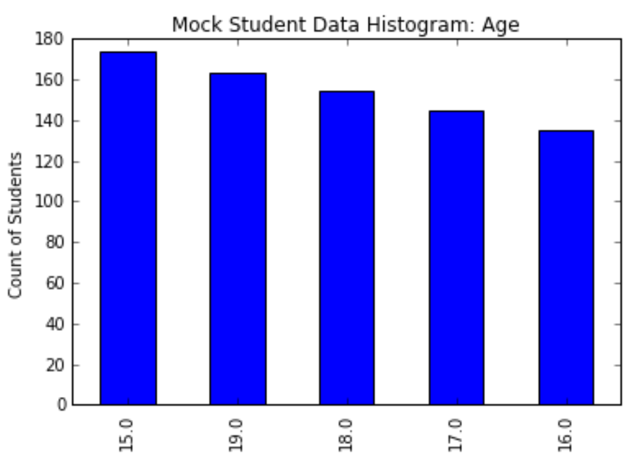
```
hist_State.png created
```



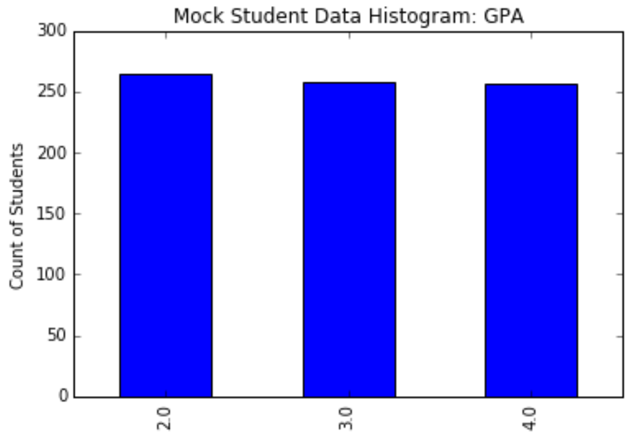

```
hist_Gender.png created
```



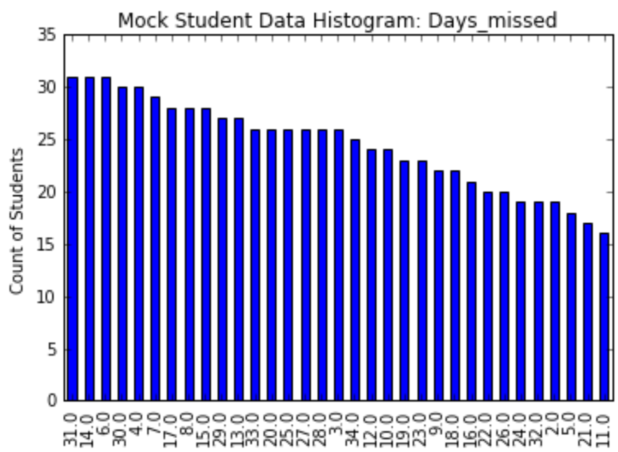

```
hist_Age.png created
```



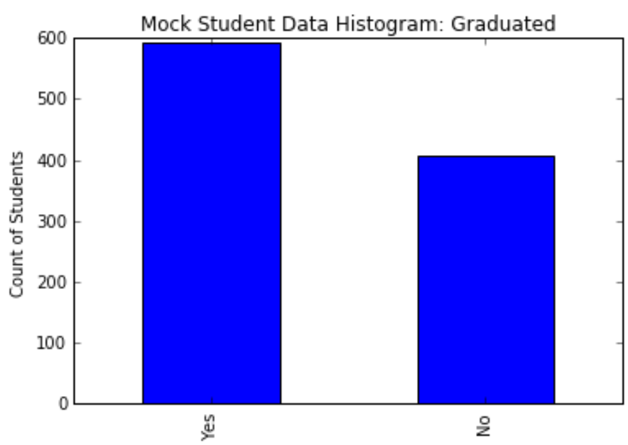

```
hist_GPA.png created
```




```
hist_Days_missed.png created
```




```
hist_Graduated.png created
```




```
summary_stats.csv created
mock_student_data_gender_inferred.csv created
Number of records with missing gender after inference: 0
mock_student_data_gend_inf_fillna_mean.csv created
mock_student_data_gend_inf_fillna_cond_mean.csv created
mock_student_data_gend_inf_fillna_cond_mean2.csv created
```

In [6]:
```
# Problem A (continued) #
# 3c. I'm using conditional mean on Graduation and Gender to infer the
# missing values for Age, GPA, and Days_missed. We could have used padding
# methods in pandas if we thought there was some mathematical relationship
# between one missing value and the next. We could also have used linear
# regression to infer the missing values but there were a lot of missing
# values.
```

In [7]:
```
# Problem B #
```

In [9]:
```
# Problem B, Initial Question A #
# David and Chris would be expected to have the same probability of
# of graduation. Since Chris and Adam differ only on income, Chris's
# probability will be -0.109 * ln(10,000) relative to Adam's (where the
# 10,000 is the difference in income between the two). Similarly, David's
# probability will also be -0.109 * ln(10,000), relative to Bob's. Because
# Bob and Adam have the same probability, and Chris and David differ by
# the same amount relative to Bob's and Adam's probability--Chris and David
# have the same probability of graduation, as predicted by the model.
```

In [10]:
```
# Problem B, Second Question A
# The negative coefficient on AfAm_Male indicates that relative to
# non-African American Males, African American Males are less likely to
# graduate. Since the coefficients on Female, and AfAm are -2.11, and
# 2.07 respectively--the negative coefficient on AfAm_Male indicates that
# relative to African American Females, African American Males are also
# less likely to graduate.
```

In [13]:
```
# Problem B, Question B
# Although the z-scores on Age and Age_Sq are not high enough in absolute
# value to be significant, the coefficients indicate that different ages
# may be correlated differently with the probability of graduation. The
# model variables estimate that the incremental probability of
# graduation changes to positive beyond age 65.
# 0 = -0.013 + .0001 * 2 * Age
# Age = 65
```

In [12]:
```
# Problem B, Question C
# I would drop either the Male or Female variable since having both of them
# is redundant. Having a coefficient on Male also gives us a sense of the
# correlation between being a Female (when Male = 0) and the probability of
# graduating. I would also potentially look into removing the Age variables
# since the z-scores are not significant. However, I would need to know how
# that would affect the model (Would run F-tests), and understand
# what is the reason for our model. If we are trying to understand how Age
# is correlated with probability of graduation, then removing Age and Age_sq
# would not make sense even if it did not affect our model.
```