Census Data Analytics

Paulo C. Rios, Jr.

Sep 21, 2017

Goals

- Using the “Adult” dataset, predict whether income exceeds $50K/yr based on census data.

- Choose an accuracy statistic to measure model’s accuracy.

- “Adult” dataset available at http://archive.ics.uci.edu/ml/datasets/Census+Income

Actions

- Using a Python Jupyter notebook, I have done the following:
- Training Data Set Importing and Reading
- Data Preprocessing
- Feature Visualization and Transformation
- Model Training and Performance Measurements
- With two models: Random Forest and Stochastic Gradient Descent (SGD)

Data Preprocessing

Steps

- Columns were renamed
- Target feature had its value reset as True/False
- Null values were removed
- Original training data set had 32,560 rows
- Without null values, training data set has 30,161 rows

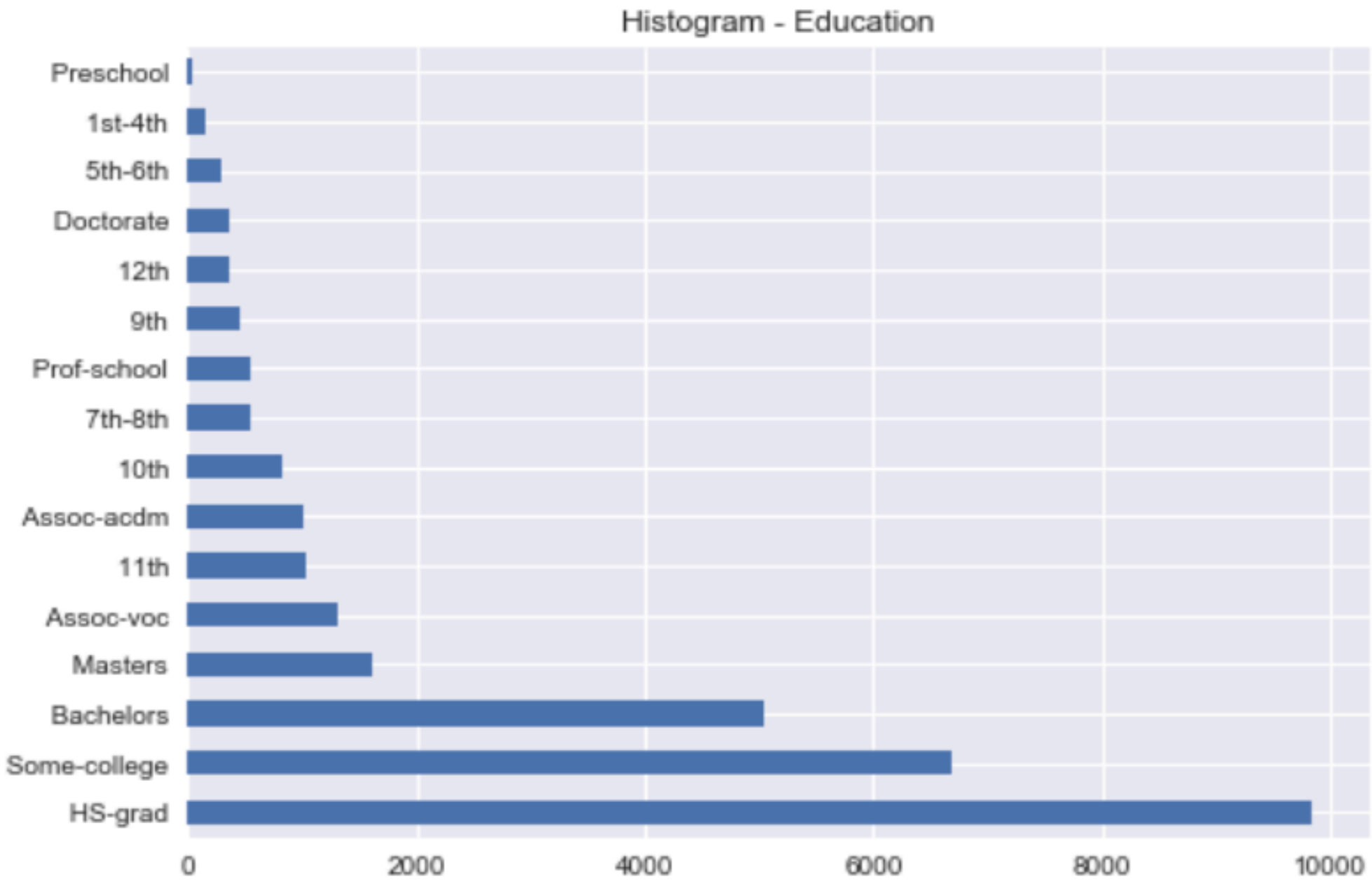
Null Values

```
In [26]: data.isnull().any()
```

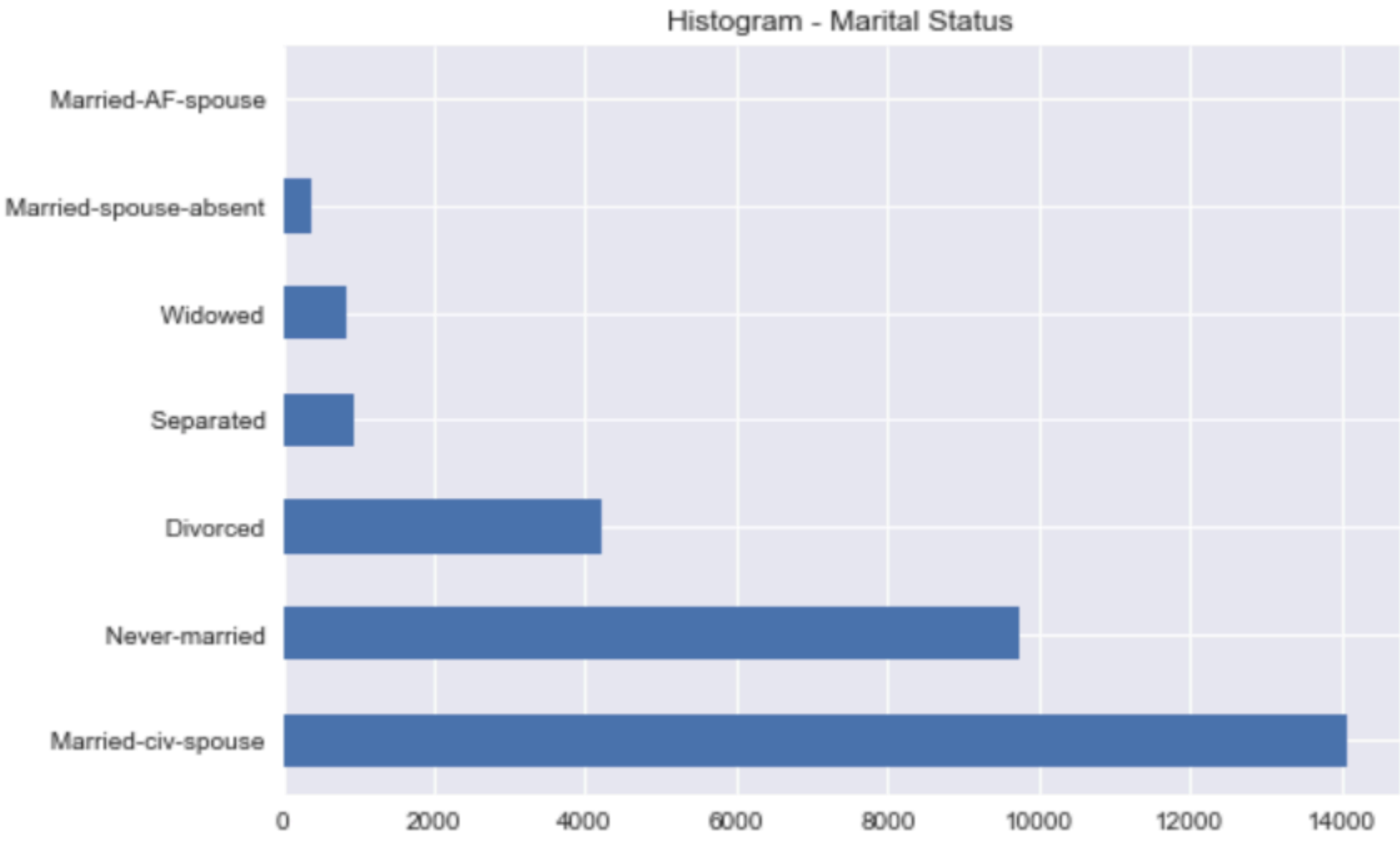
```
Out[26]: Age                 False
         Work Class           True
         Final Weight        False
         Education           False
         Education Num       False
         Marital Status      False
         Occupation           True
         Relationship        False
         Race                False
         Sex                 False
         Capital Gain        False
         Capital Loss        False
         Hours per Week      False
         Native Country       True
         <=50K               False
         dtype: bool
```

Feature Visualization

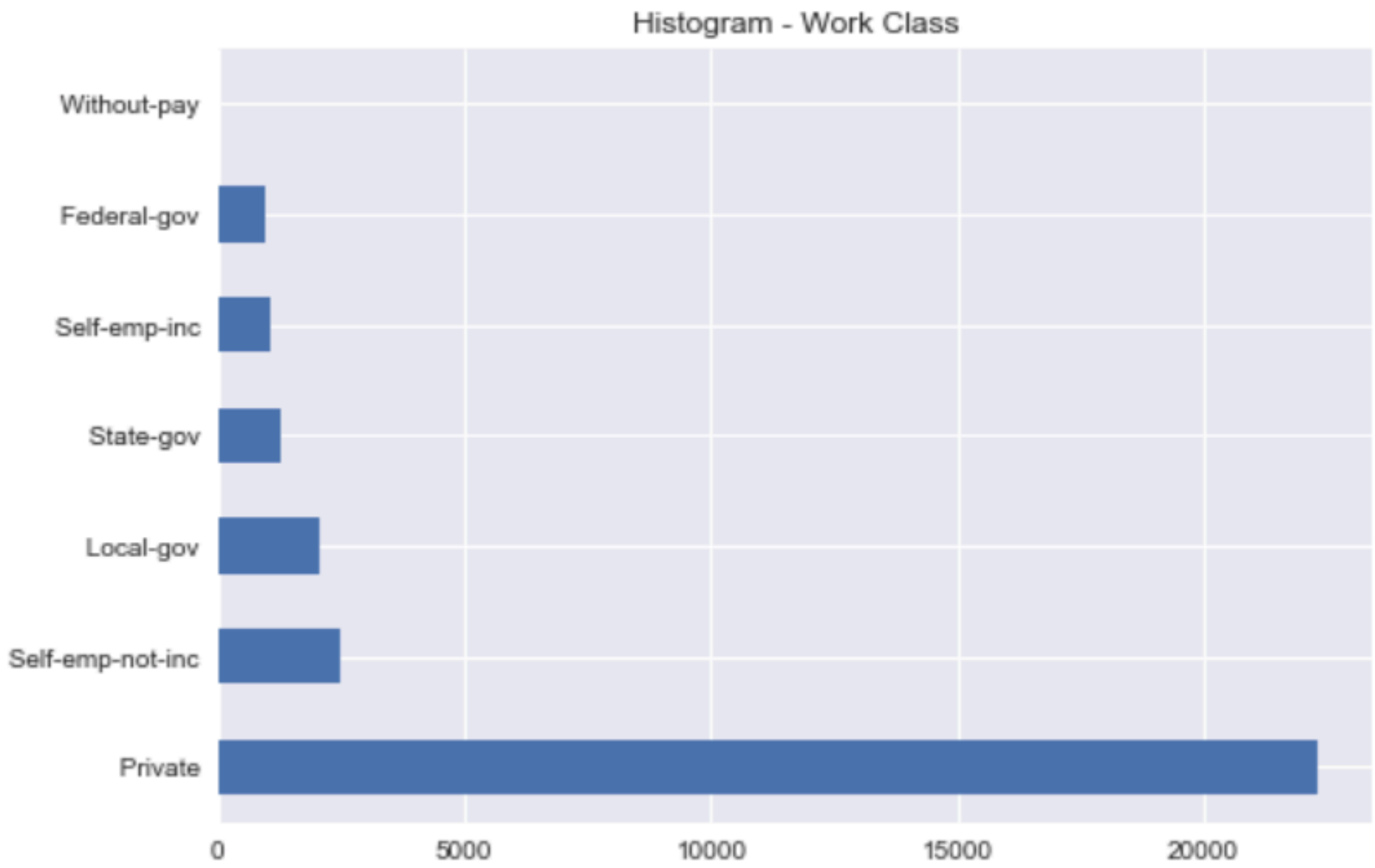
Education


Histogram - Education
Preschool
1st-4th
5th-6th
Doctorate
12th
9th
Prof-school
7th-8th
10th
Assoc-acdm
11th
Assoc-voc
Masters
Bachelors
Some-college
HS-grad
0
2000
4000
6000
8000
10000

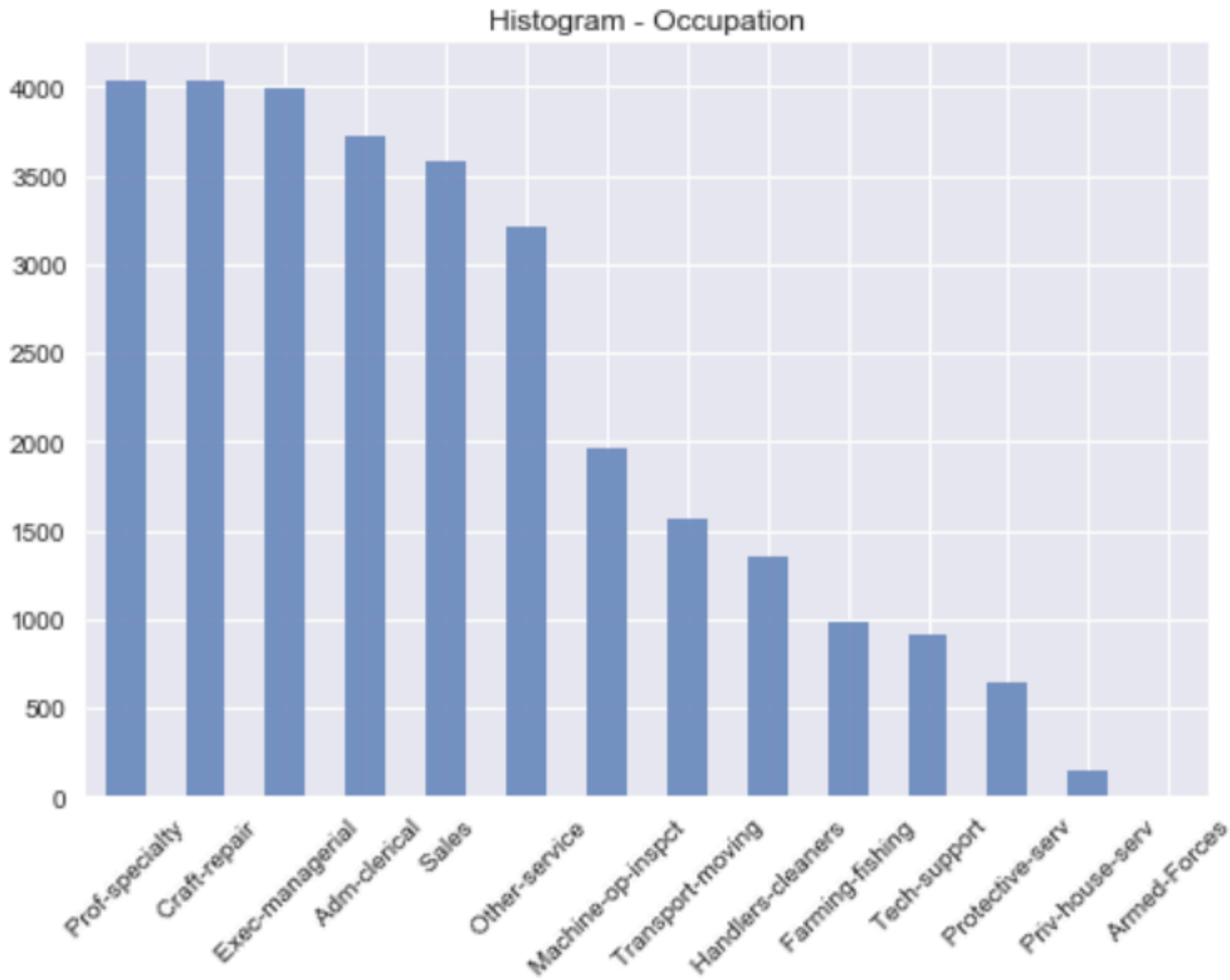
Marital Status



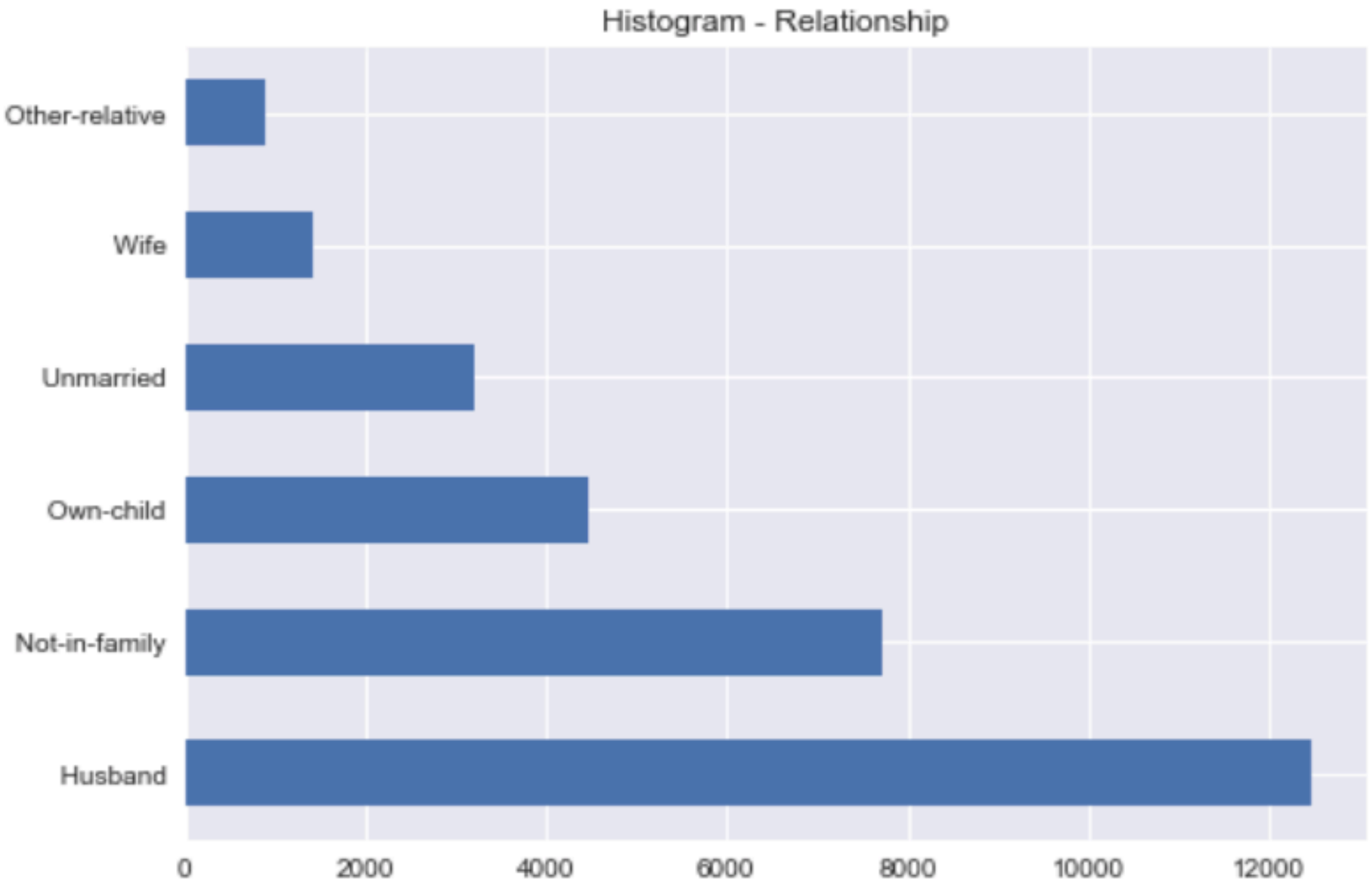
Work Class


Histogram - Work Class
Without-pay
Federal-gov
Self-emp-inc
State-gov
Local-gov
Self-emp-not-inc
Private
0
5000
10000
15000
20000

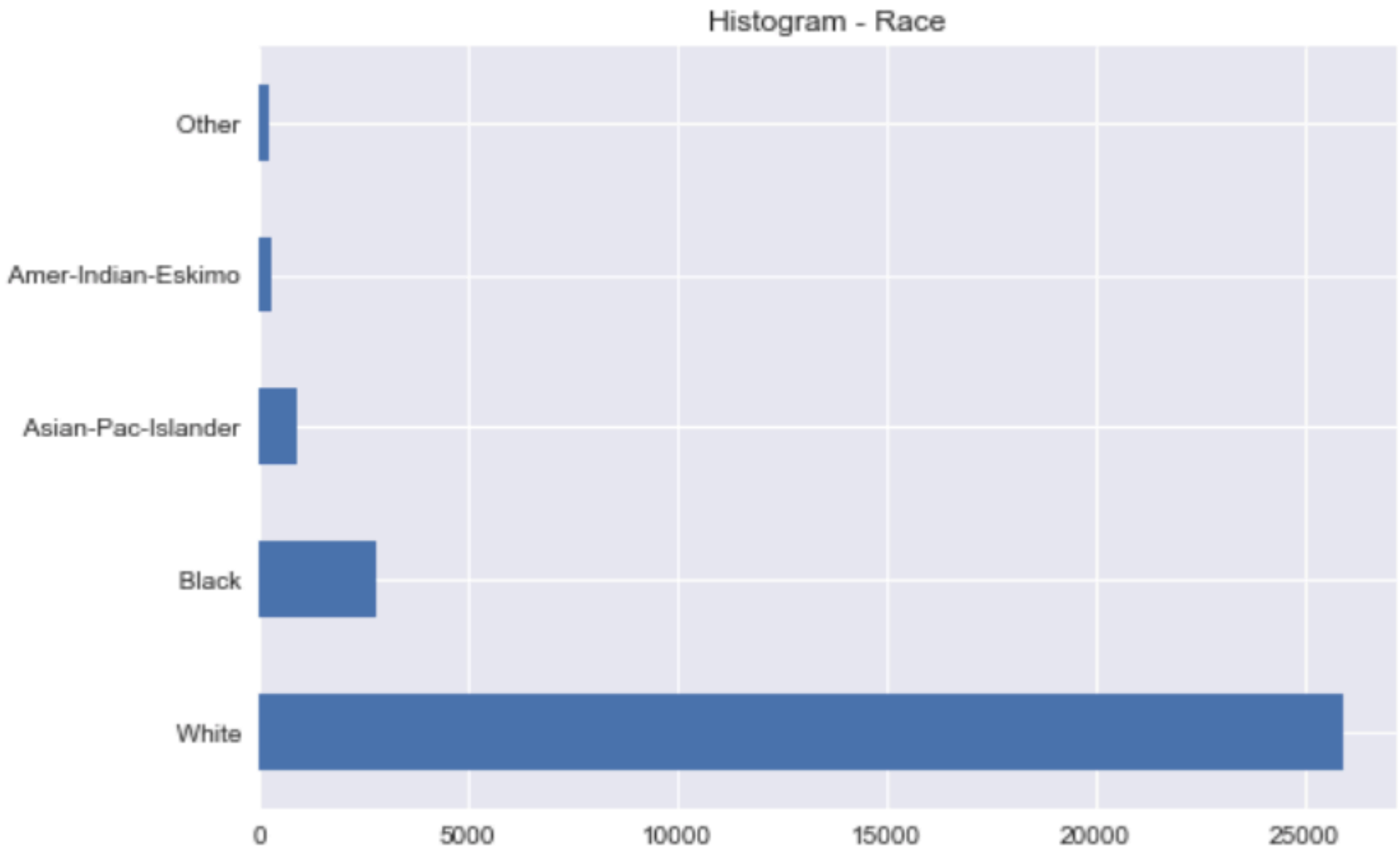
Occupation


Histogram - Occupation
4000
3500
3000
2500
2000
1500
1000
500
0
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces

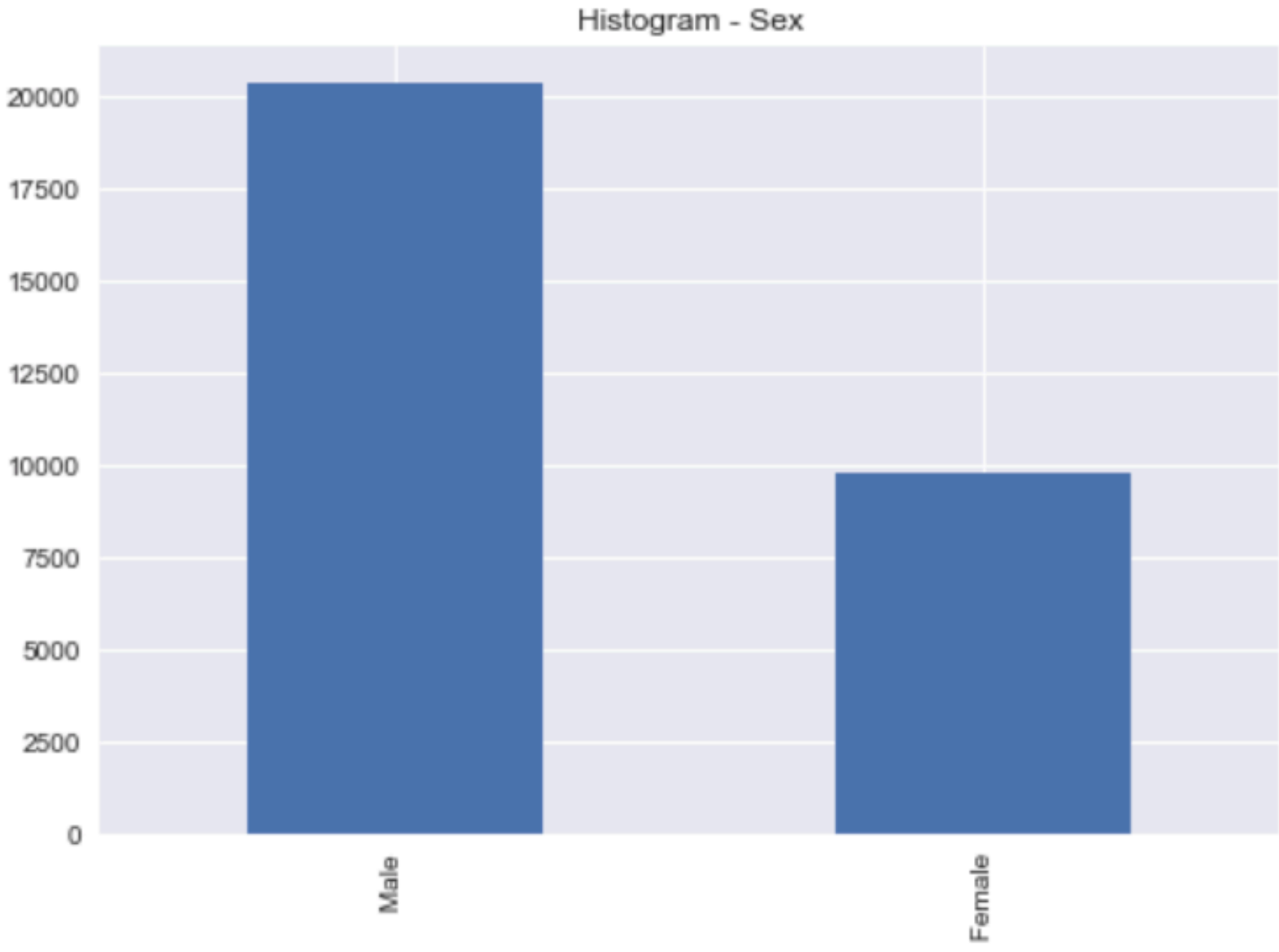
Relationship



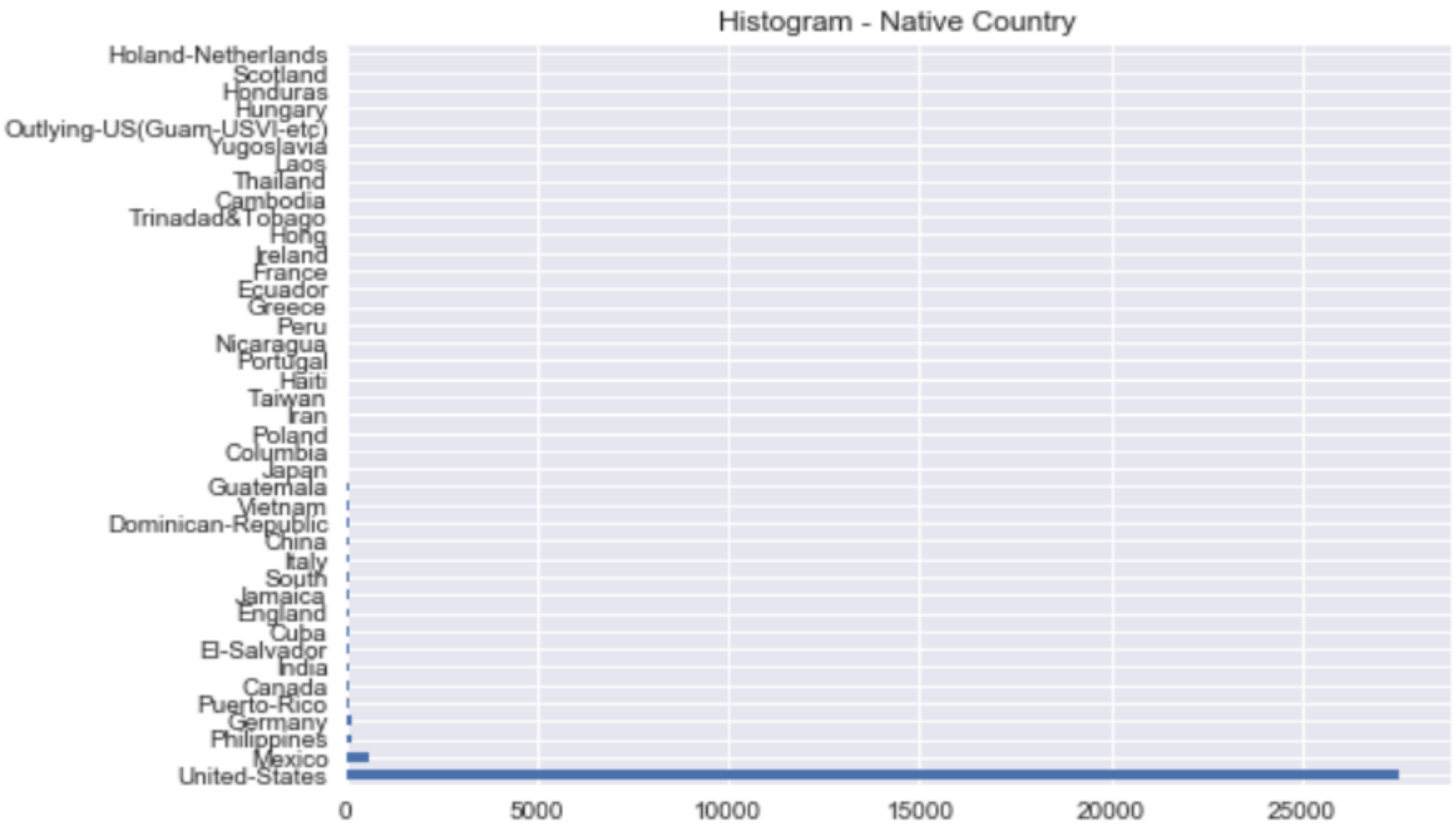
Race


Histogram - Race
Other
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
White
0
5000
10000
15000
20000
25000

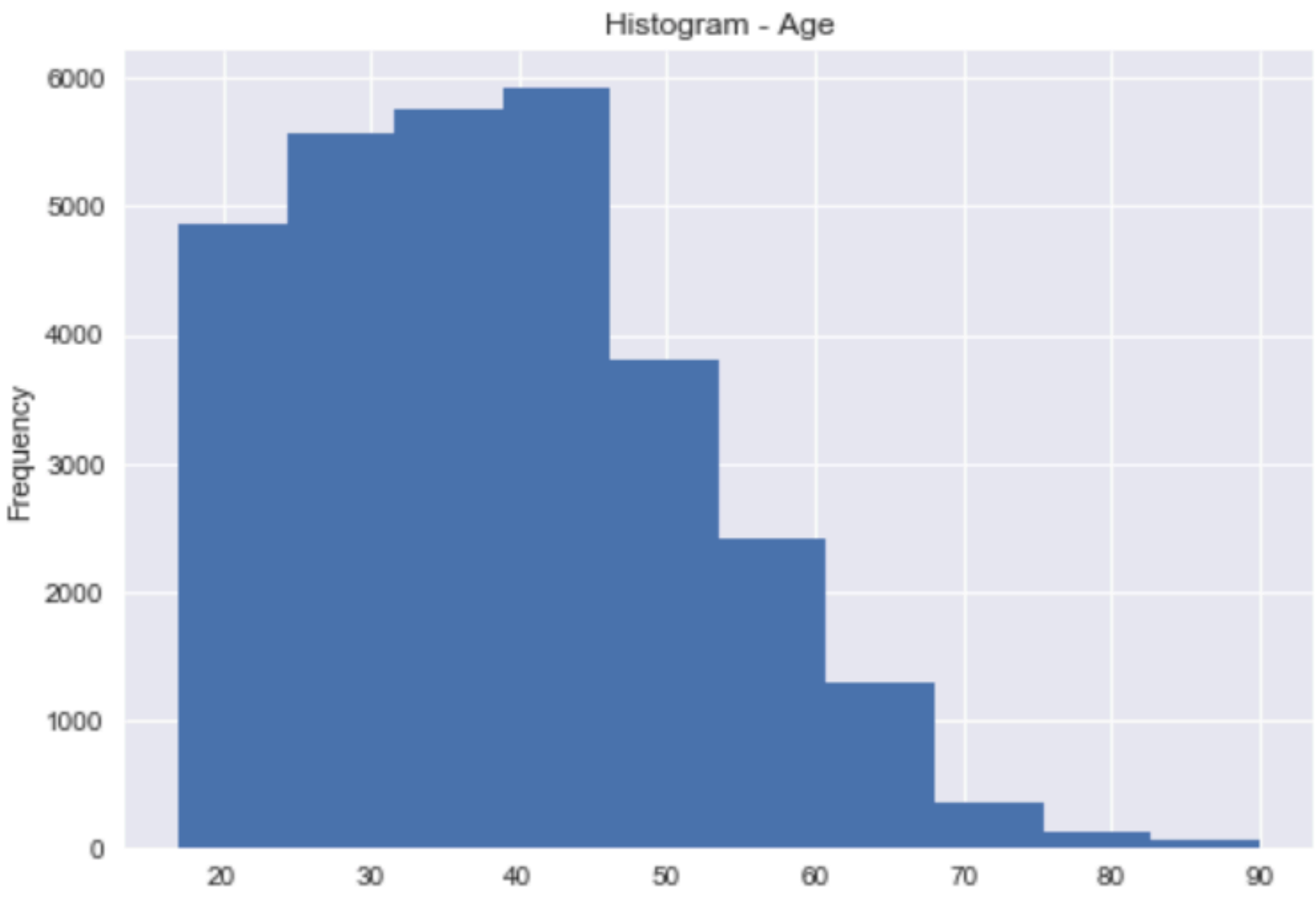
Sex


Histogram - Sex
20000
17500
15000
12500
10000
7500
5000
2500
0
Male
Female

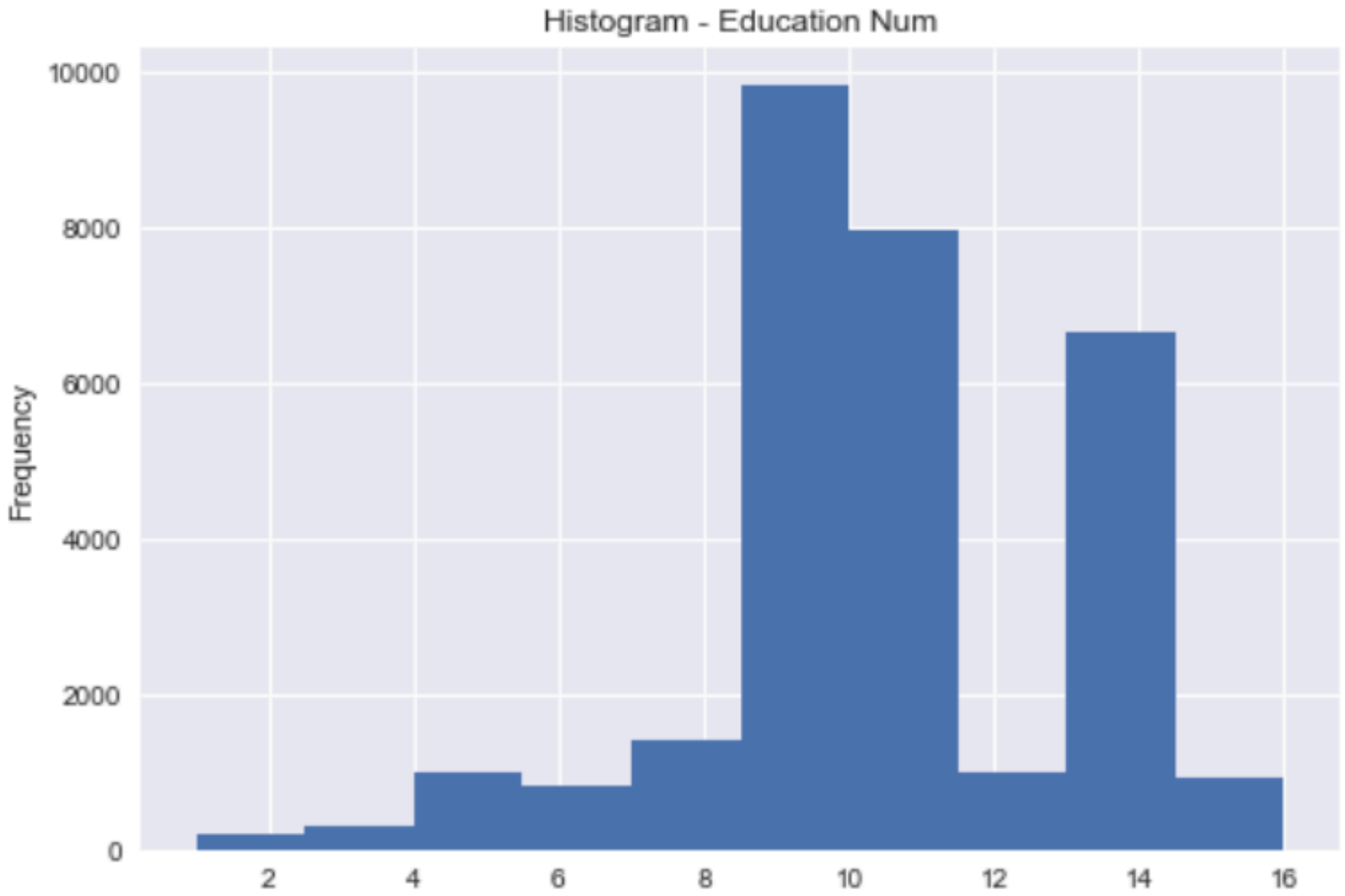
Native Country



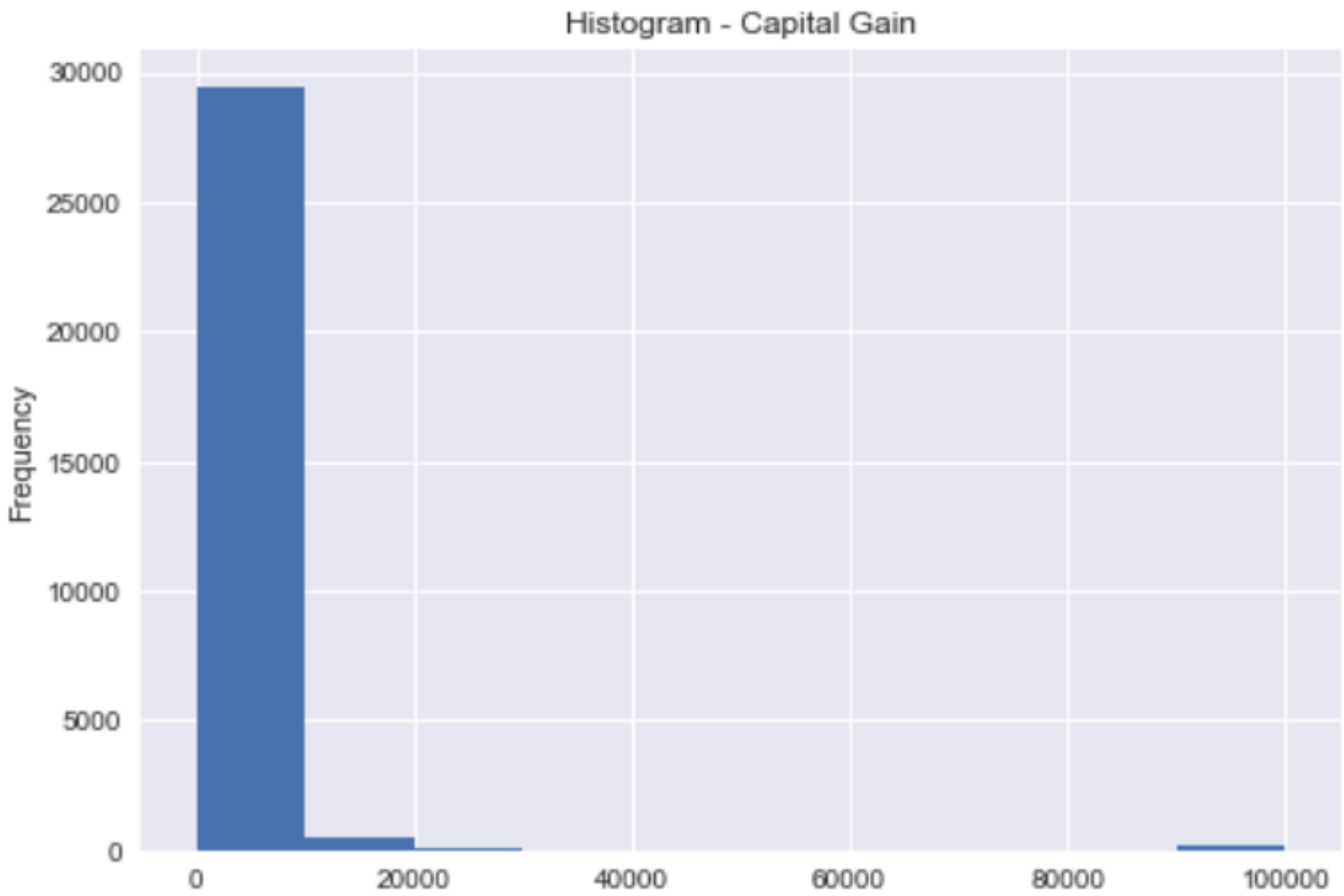
Age


Histogram - Age
Frequency
6000
5000
4000
3000
2000
1000
0
20
30
40
50
60
70
80
90

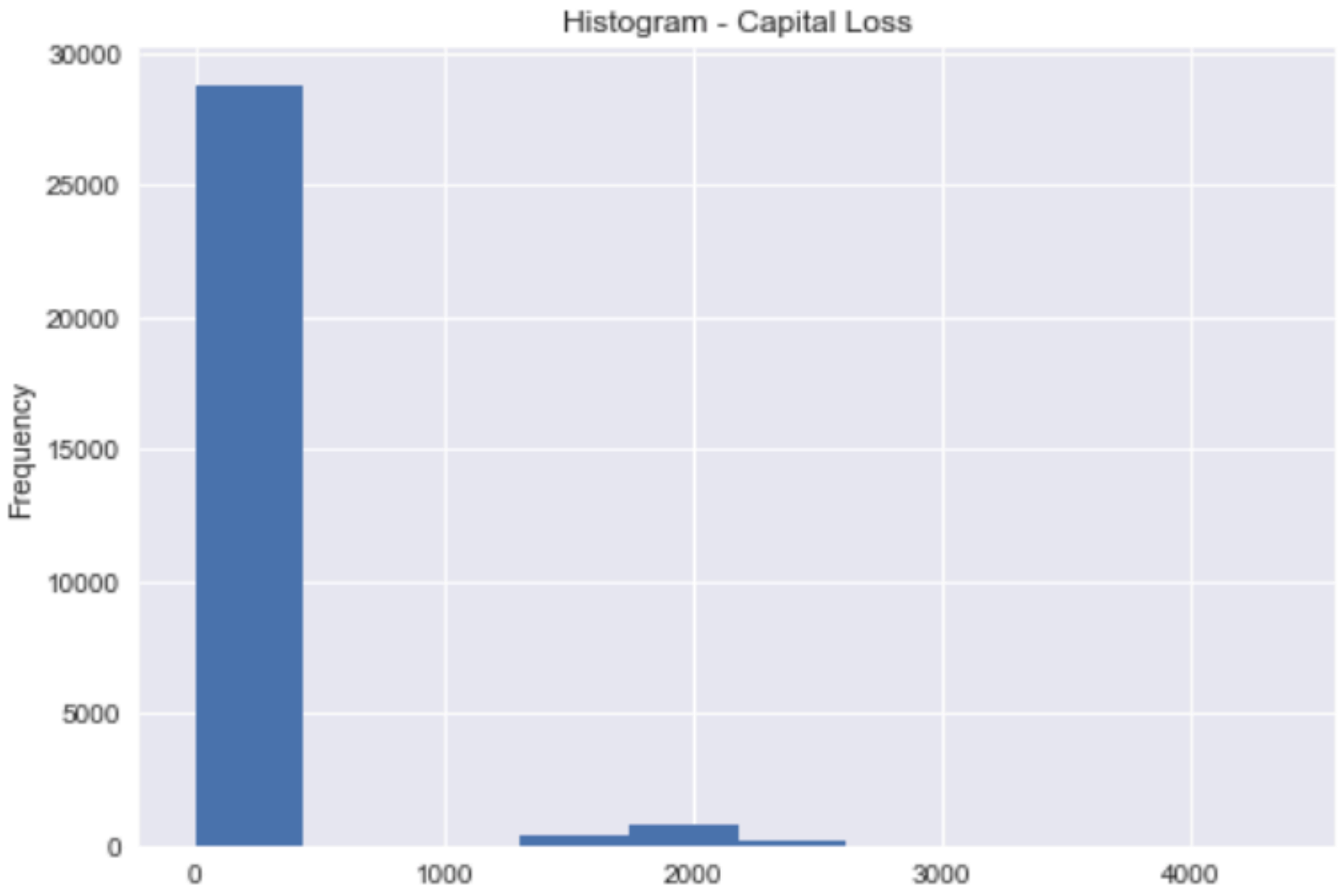
Education - Num



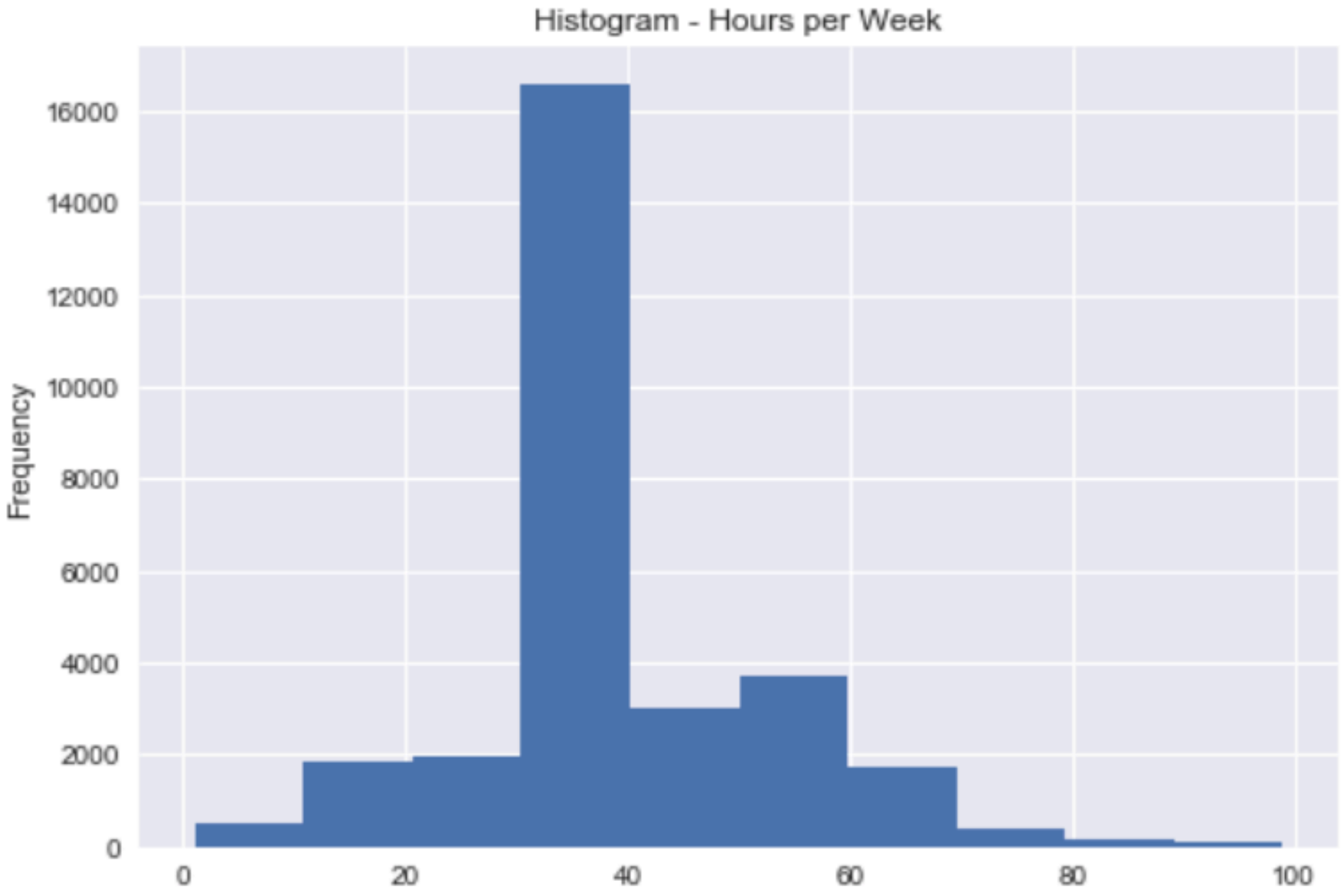
Capital Gain



Capital Loss



Hours Per Week



Model Training and Performance Measurements

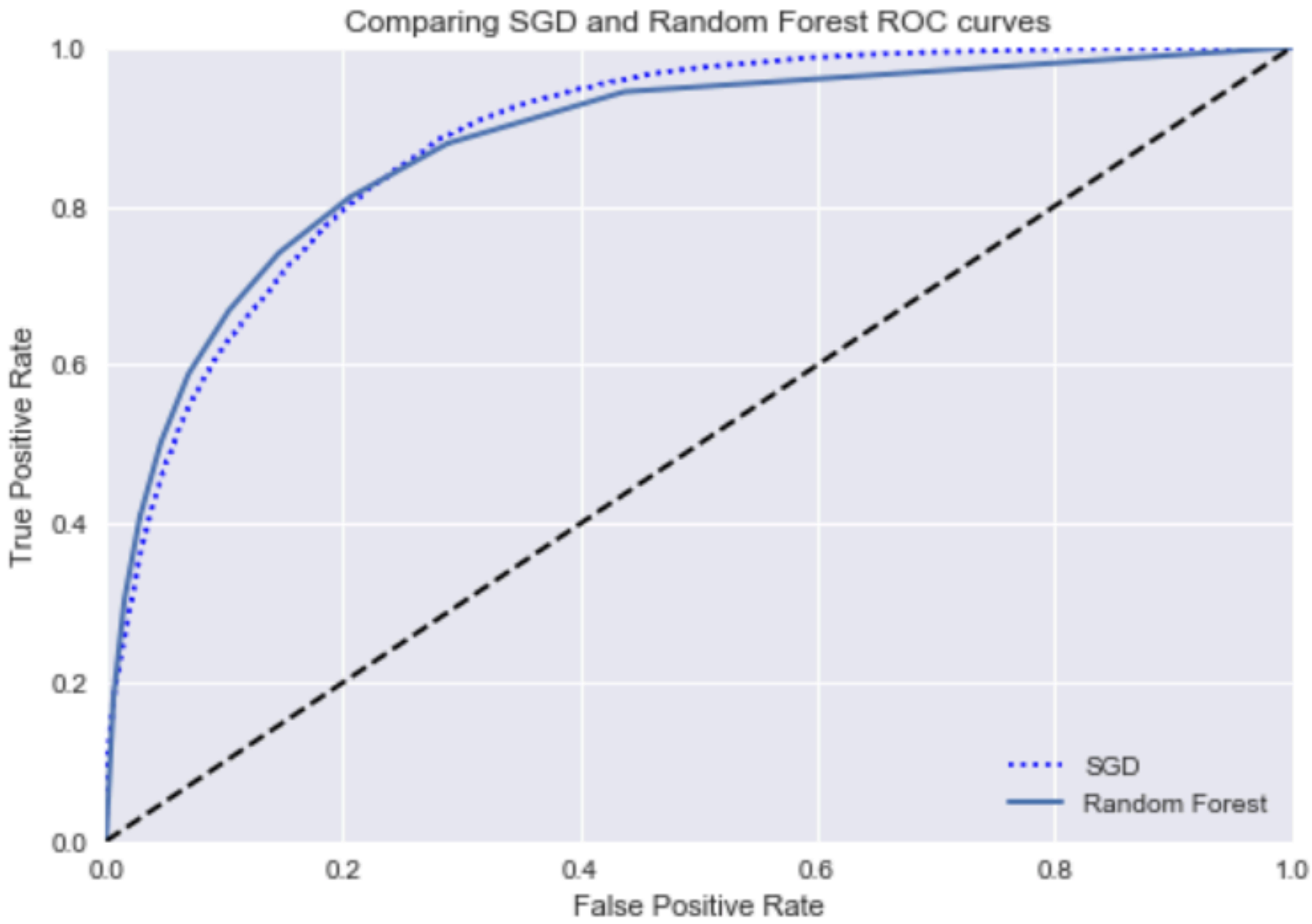
Cross-Validation Accuracy Scores

	Random Forest	SGD
1st CV run	0.845236	0.816590
2nd CV run	0.841058	0.838074
3rd CV run	0.849597	0.844425

Confusion Matrix Scores

	Random Forest	SGD
Precision Score	0.736991	0.684533
Recall Score	0.588572	0.610682
F1 Score	0.654473	0.645502

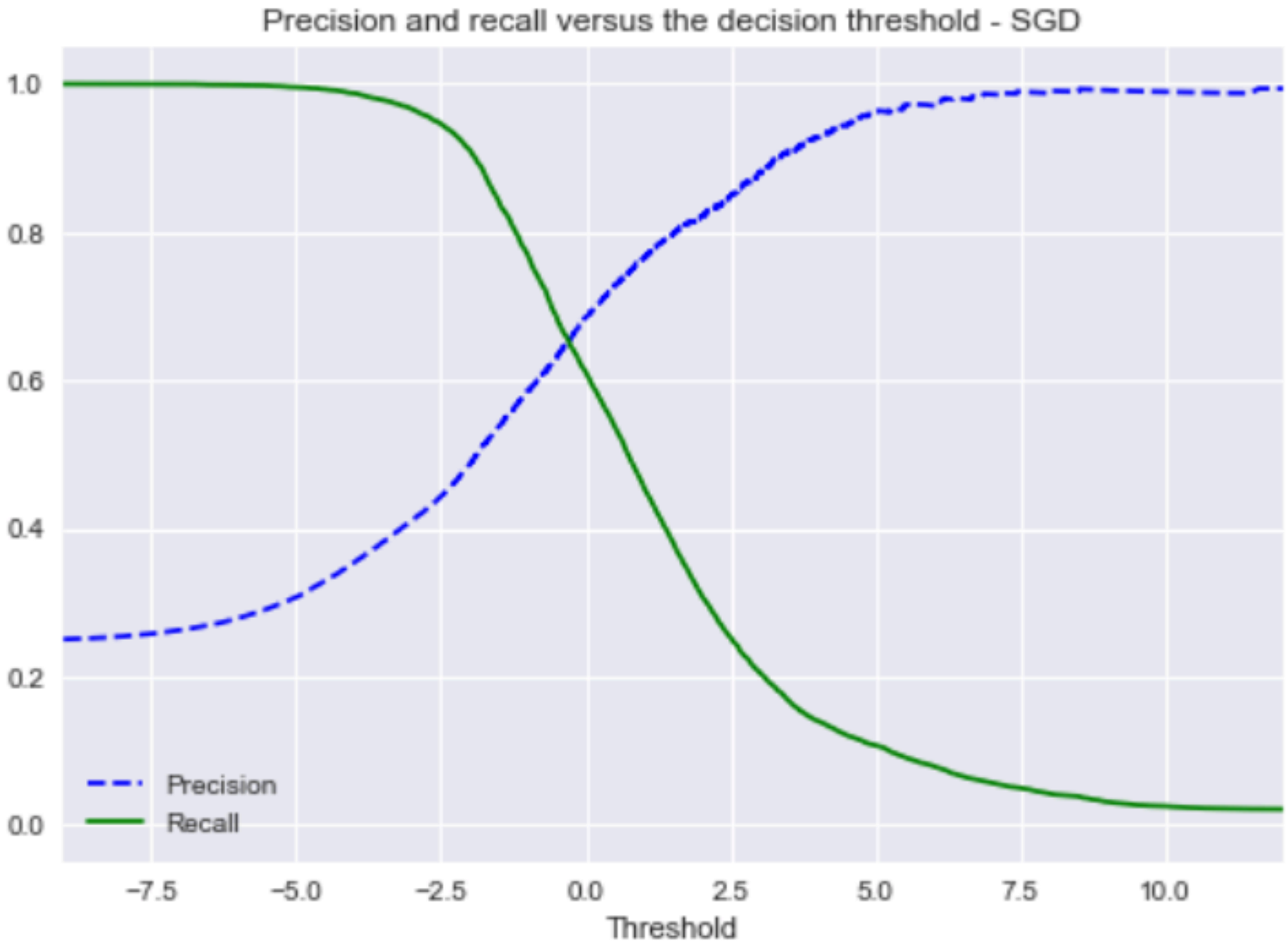
ROC Curves


Comparing SGD and Random Forest ROC curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
SGD
Random Forest

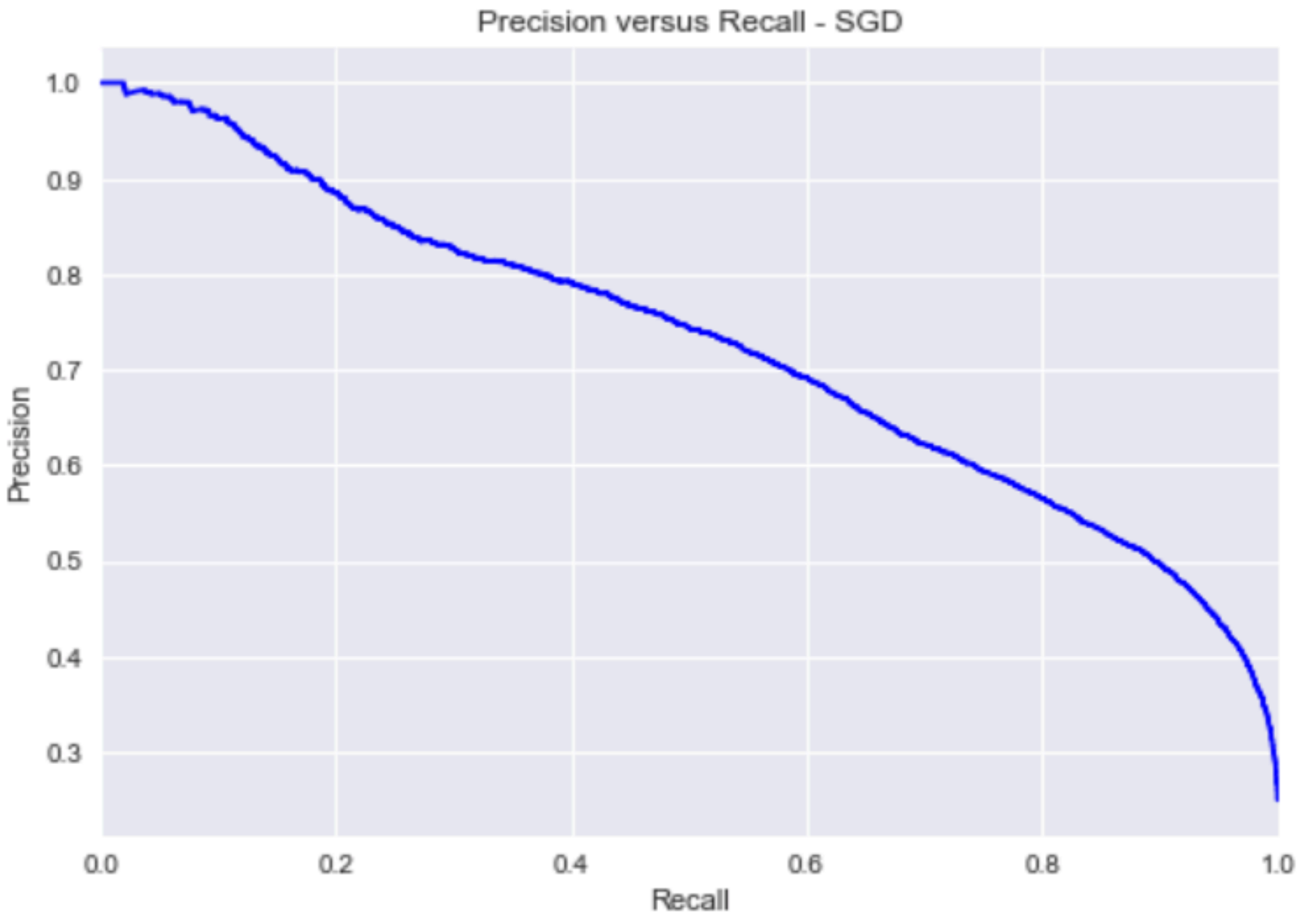
ROC AUC Scores

	Random Forest	SGD
ROC AUC Scores	0.87842	0.885639

Precision and Recall - SGD

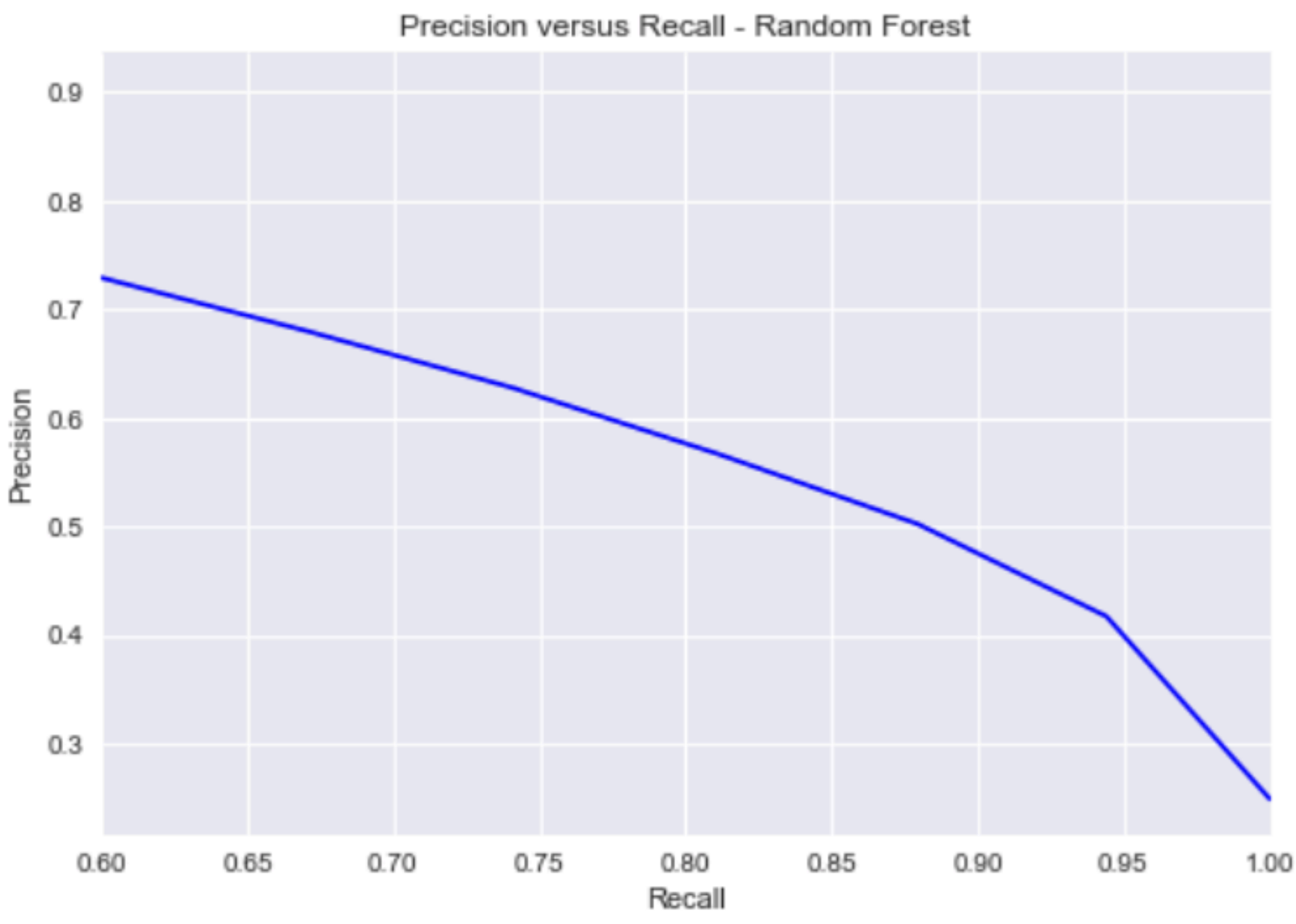


Precision and Recall - SGD



Precision and Recall Random Forest

Precision and Recall Random Forest



Next Steps

- Remove outliers and check effect on performance measurements
- Choose best precision/recall tradeoff for best classifier
- Fine tune best classifier with Grid Search
- Perform data preprocessing and transformation on the test set
- Evaluate best performing classifier on the test set