Kate Larsen
January 25, 2017

Online Assessment

Data Analysis

1) Generate per-column statistical information that would be useful for understanding the data set and for choosing an appropriate machine learning method, i.e., Nearest Neighbors, SVM, logistic regression, Random Forests, etc. Make explicit any assumptions that influence your choices. These may be in the form of comments to your code.

Numerical Analysis:

- **Key measures:** mean, median, standard deviation, skewness, correlation coefficients
- The capital gain and capital loss columns have heavy skew (use of e1071 library). This marks this feature to be further investigated to check that even with skewed data, it still shows differentiation towards the resulting income level. The histograms in the next section detail this.

	Column	Mean	Median	Standard Deviation	Skewness
1	age	38.582	37	13.64	0.559
2	fnlwgt	189778.367	178356	105549.978	1.447
3	education.num	10.081	10	2.573	-0.312
4	capital.gain	1077.649	0	7385.292	11.953
5	capital.loss	87.304	0	402.96	4.594
6	hours.per.week	40.437	40	12.347	0.228

- Overall, these numeric columns vary widely in the unit of measure; it will be important to standardize them for various machine learning methods, especially nearest neighbors.
- The table below displays correlation analysis on the numeric columns to check for multicollinearity. The highest correlation coefficient was 0.148 between education and hours per work. The others were rather low across the board.

	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
age	1	-0.077	0.037	0.078	0.058	0.069
fnlwgt	-0.077	1	-0.043	0	-0.01	-0.019
education.num	0.037	-0.043	1	0.123	0.08	0.148
capital.gain	0.078	0	0.123	1	-0.032	0.078
capital.loss	0.058	-0.01	0.08	-0.032	1	0.054
hours.per.week	0.069	-0.019	0.148	0.078	0.054	1

Categorical analysis:

- **Key measures**: missing data, number of levels, distribution within levels
- Some categorical columns had '?' values representing the missing data. The following table shows the number of occurrences within each of the categorical columns, as well as the total number of occurrences in the whole data set. Overall, 7.4% of the records had at least one '?' point.

	Column	Number of Unknown	Percent Unknown
1	workclass	1836	0.056
2	education	0	0
3	marital.status	0	0
4	occupation	1843	0.057
5	relationship	0	0
6	race	0	0
7	sex	0	0
8	native.country	583	0.018
9	yearly.income	0	0
10	Total with Overlap	2399	0.074

- Due to the issue of dimensionality, I calculated the number of levels within each column. As seen below, education, occupation, and native country had the highest number of levels. This should be taken into account with bucketing if smaller dimensionality is desired.
- For the highest dimension of native country, I calculated the frequency for each level. The United States dominated the count, suggesting the use of bucketing for United States versus non United States.

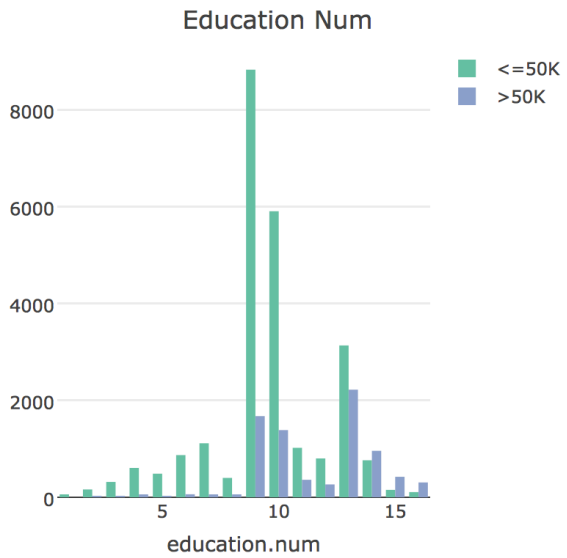
	Column	Number of Levels
1	workclass	9
2	education	16
3	marital.status	7
4	occupation	15
5	relationship	6
6	race	5
7	sex	2
8	native.country	42
9	yearly.income	2

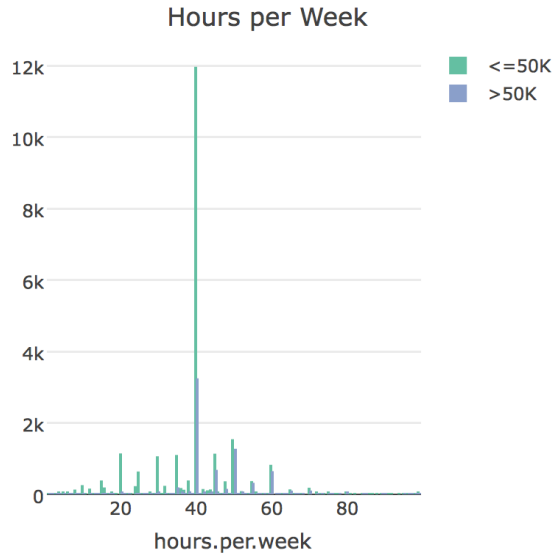
	Country	Freq
40	United-States	29170
27	Mexico	643
1	?	583
31	Philippines	198
12	Germany	137
3	Canada	121
34	Puerto-Rico	114
9	El-Salvador	106
20	India	100
6	Cuba	95

2) Create visualizations (plots, histograms, etc.) that would help to: a) gain insight into the data; b) help to choose the appropriate machine learning method(s); c) suggest issues in the data that may need to be addressed. Include any relevant information that guided your choices.

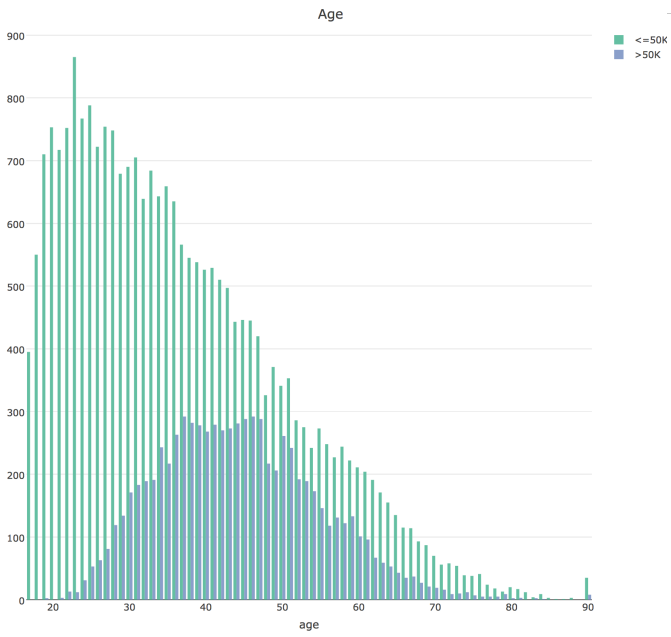
Numeric Columns:

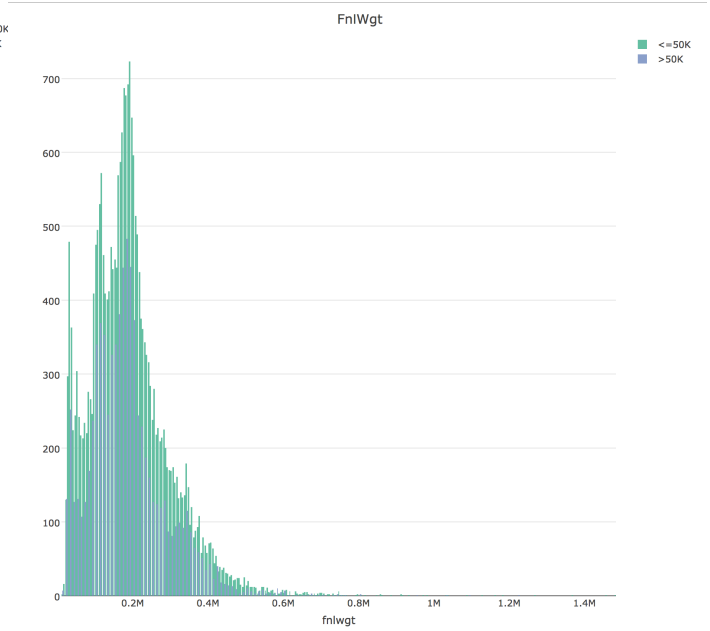
- **Education Num**: Initial education numbers show a strong association with <=50K, however as the number grows past 10, there is an evident split between each interval for income level.
- **Hours per Week**: An interesting piece in the hours per week histogram shows the high number of 40 hour per week data points, as expected from national averages. Within these 40 hour week data points, about 75% of those have an income of <=50 K, which is a rather significant split between the income levels.



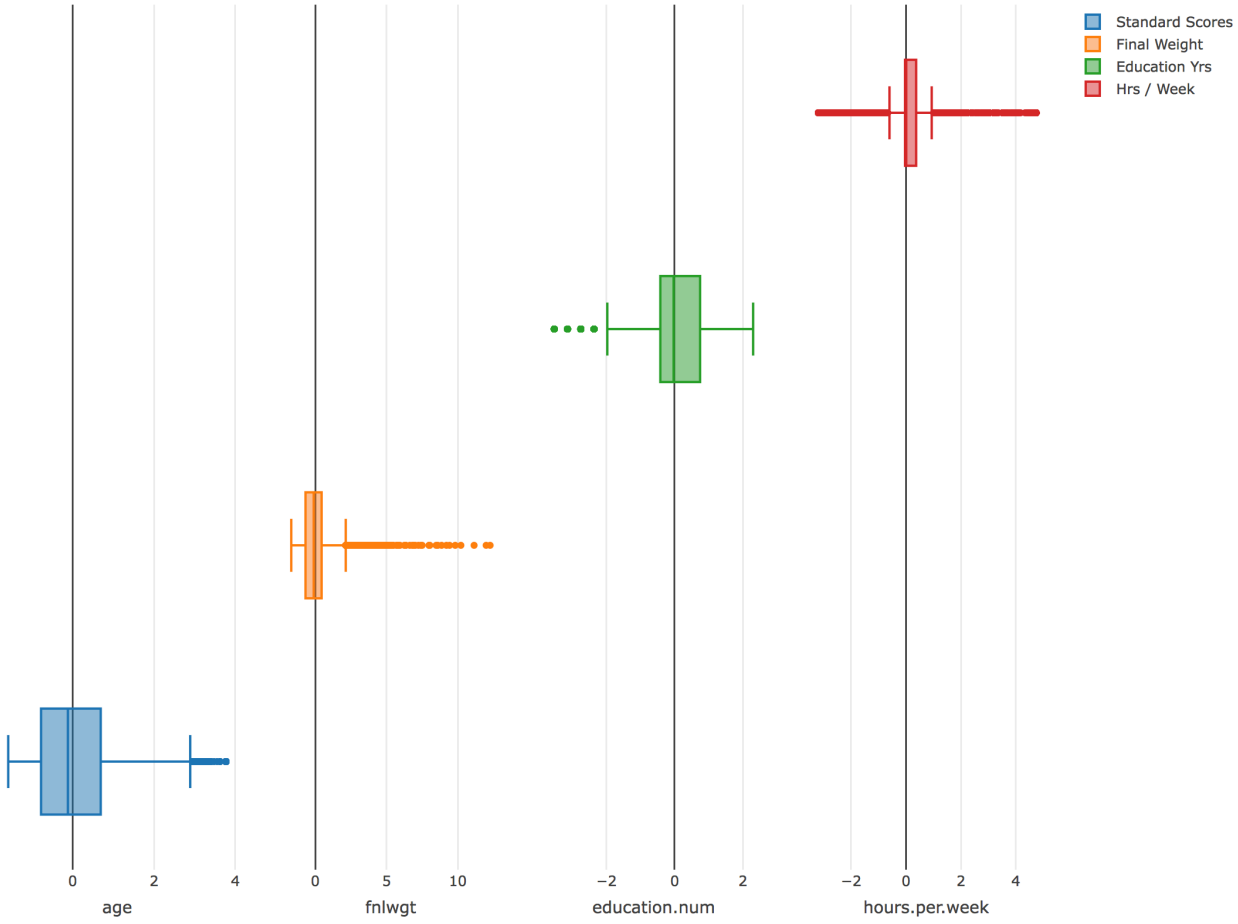


- **Age:** In the Age histogram, there is a sharp increase of income > 50K around the 30 year mark. This is important as those records with age less than 30, which viewing the frequency of the histogram, is quite high, are almost all marked as <= 50K income.
- **Fnl Wgt:** In the FnlWgt histogram, there is a rather even split between the two income levels, perhaps indicating that this may not be as strong of a predictor.

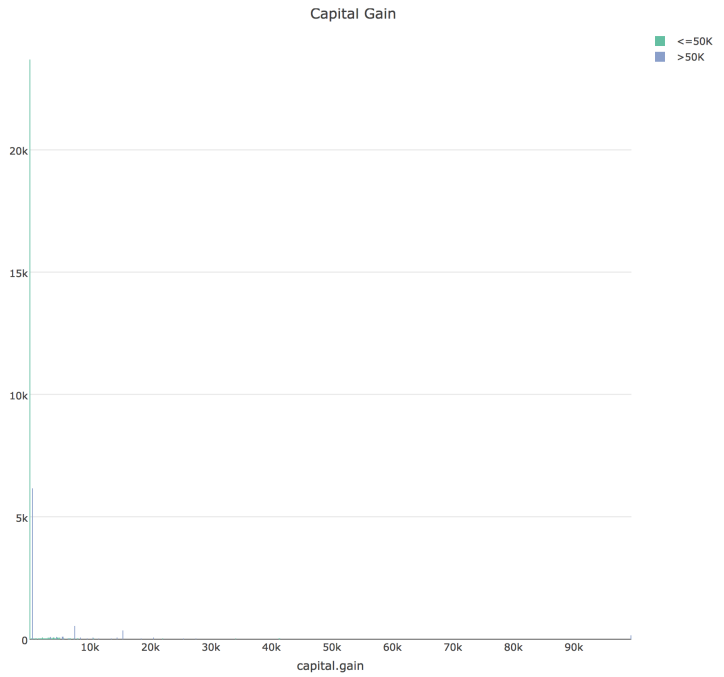


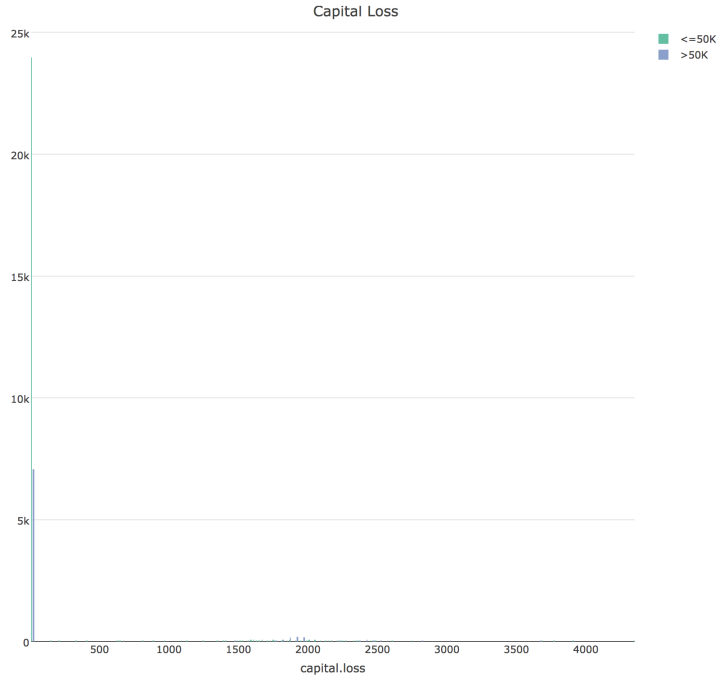


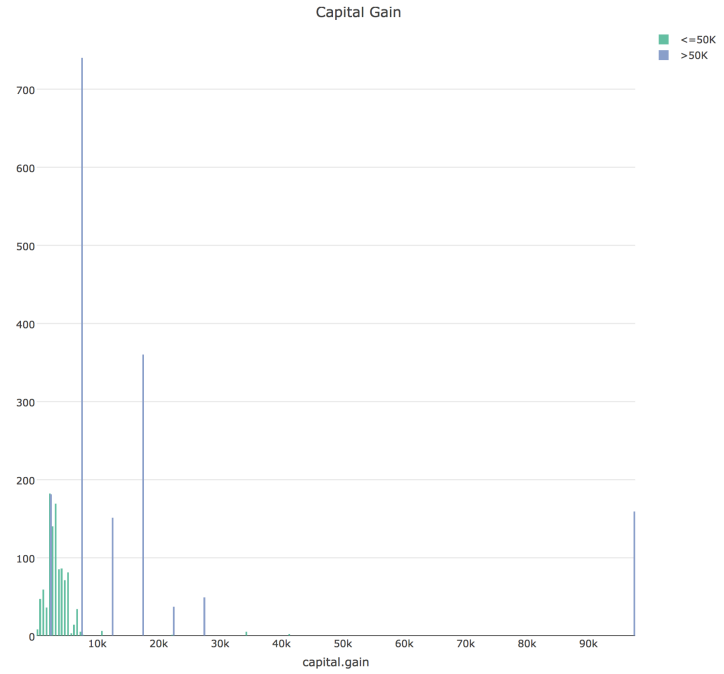
- The following box plots, built after standardizing the numeric columns, show a diverse set of ranges where age has the largest interquartile range, while hour per week is very small with many extreme points. The fnlwgt column also has many extremes in the upper portion, while not as many in lower standard deviation values. Such columns with extreme outliers can dampen the overall performance of a model if only rare instances of those cases exist. This is an important issue to note when choosing and implementing a machine learning algorithm. The capital gains and losses are not viewed here, but are analyzed on the following page.

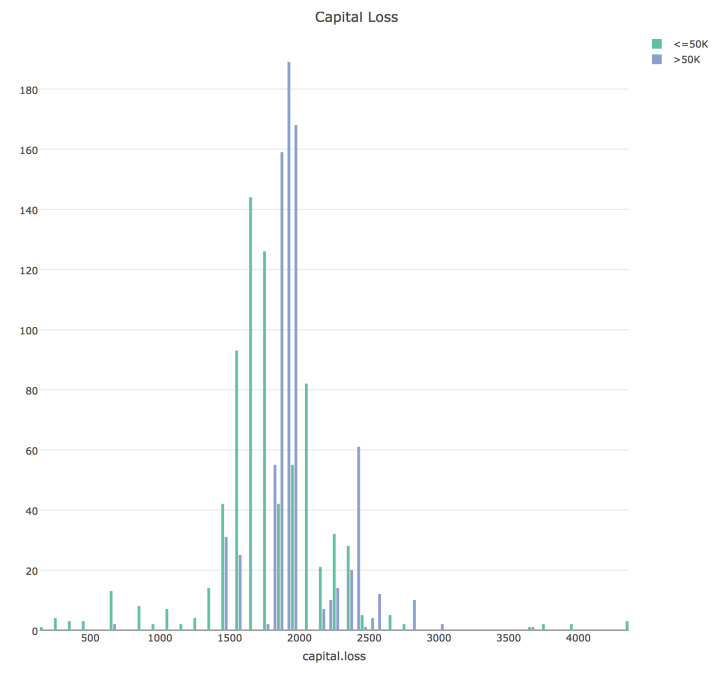


- **Capital Gains and Losses:** As seen in the histograms below, there were many records with 0 capital gains and losses. The first row of histograms shows all the data from these features. The second row of plots shows the subsets of this data to examine those records with actual gains or losses. In this analysis, almost all capital gains after ~$8,000 are marked as income of greater than $50K. In the capital losses plot, the income is much more mixed throughout.



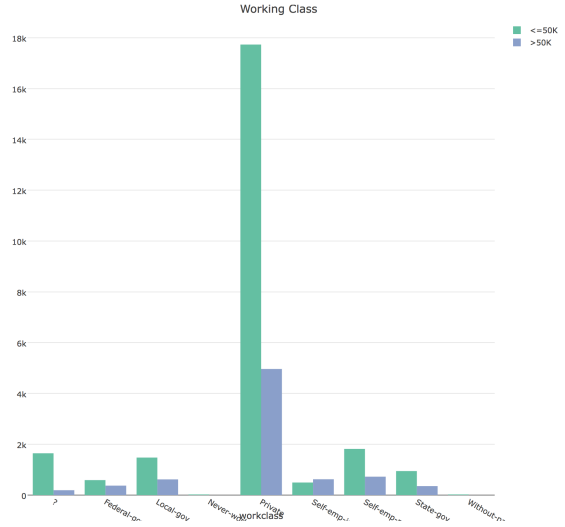


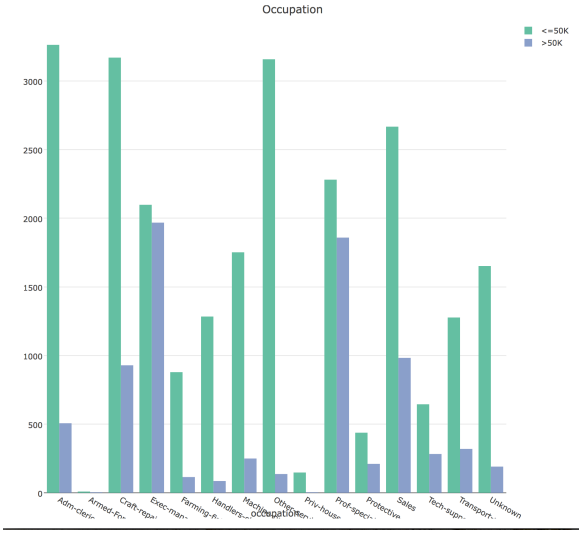


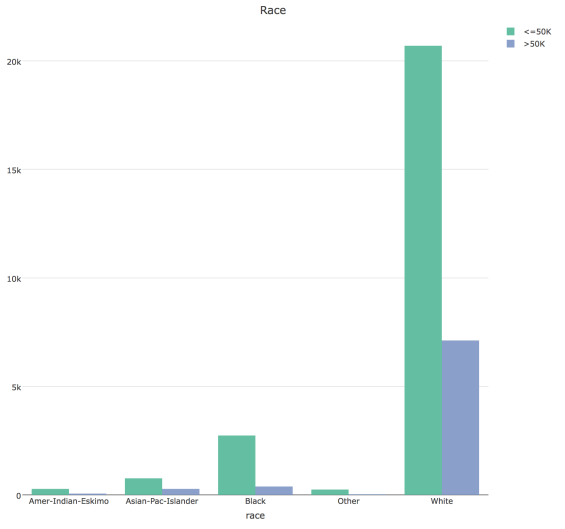


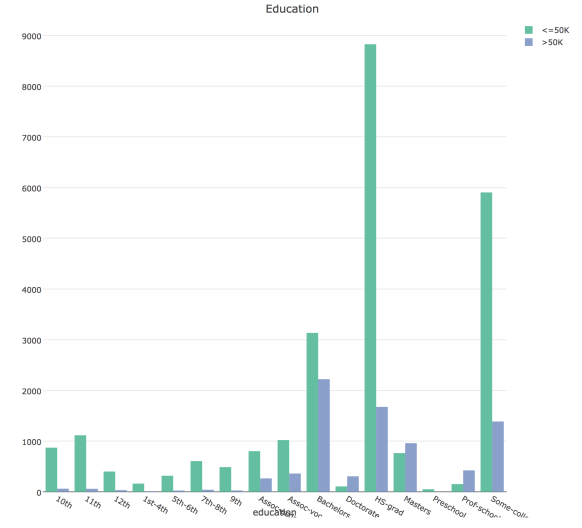
Categorical Columns:

- The full set of histograms can be viewed directly in the R code. Below are the subsets of interesting findings.
- **Working Class:** There is a large distinction between the income levels for Private working class level, while the other are split more or less. This distinction suggests that bucketing private versus not private may help in models.
- **Education, Occupation**: While there are many levels in this category, each have a rather unique split between the income levels, suggesting the need to keep this predictor intact in models.
- **Race**: The race plot shows that a vast majority of the data point are white, suggesting the possibility of bucketing this factor as well.



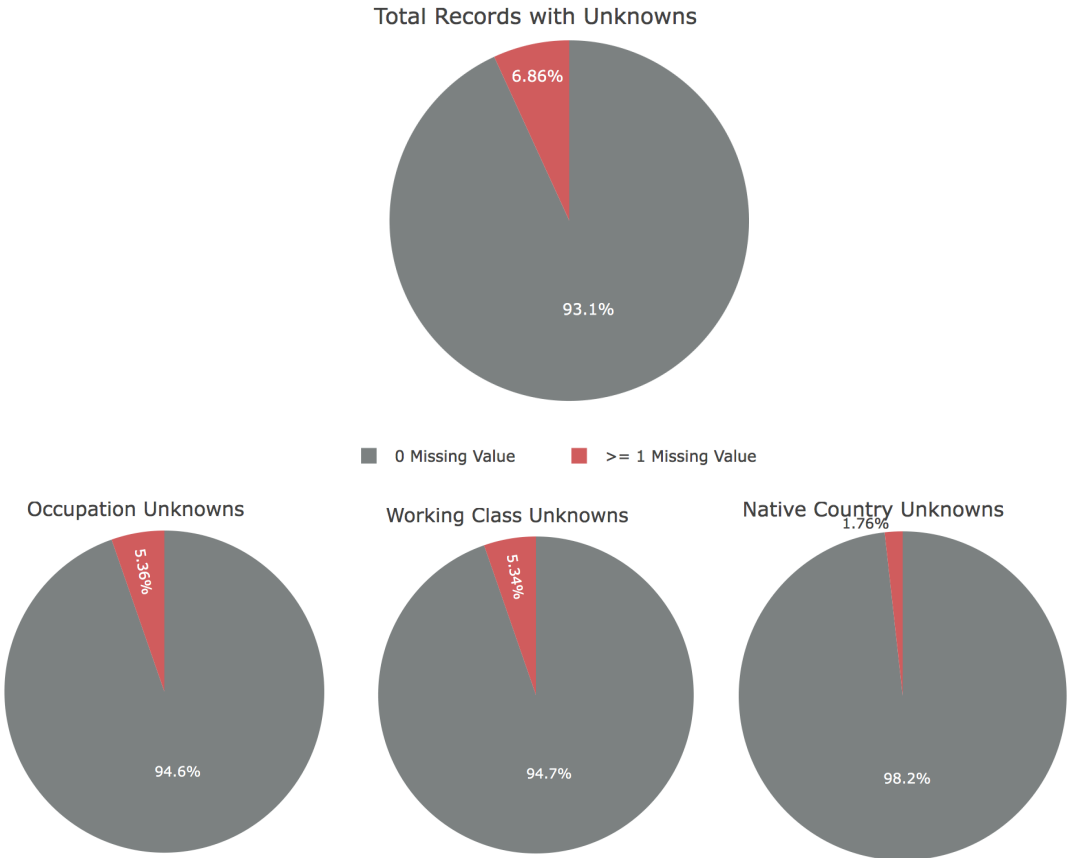






Missing Values

- Overall, just under 7% of the records have at least one '?' in either the occupation, working class, or native country columns. Dealing with these missing values will be critical in machine learning models, especially nearest neighbors. To remedy these values, the techniques of removing the data point or replacing the data value can be applied.
- My approach to replacing the missing values are to examine the common values for that column. If there is an overwhelming majority, such as the native country case and working class case, where most are marked as the United States and Private class, then I looked into replacing such values with those high frequent values. For the missing values in the occupation unknowns, I examined the respective education column value and then looked at the histograms from the education columns to determine the most frequent occupation given the education level presented. This was an assumption that the level of education relates to future occupation levels.



	Occupation for HS-grad	Freq
4	Craft-repair	1922
2	Adm-clerical	1365
9	Other-service	1281
13	Sales	1069
8	Machine-op-inspct	1023
15	Transport-moving	825
5	Exec-managerial	807
7	Handlers-cleaners	611
1	?	533
6	Farming-fishing	404
11	Prof-specialty	233
12	Protective-serv	215
14	Tech-support	159
10	Priv-house-serv	50
3	Armed-Forces	4

	Occupation for Some-college	Freq
2	Adm-clerical	1281
13	Sales	1009
5	Exec-managerial	879
4	Craft-repair	868
9	Other-service	781
1	?	516
11	Prof-specialty	430
8	Machine-op-inspct	310
15	Transport-moving	283
14	Tech-support	273
7	Handlers-cleaners	267
12	Protective-serv	202
6	Farming-fishing	174
10	Priv-house-serv	16
3	Armed-Forces	2

Risk Assessment for Machine Learning Applications

Issue	Explanation	Present Risk Level
Extreme Observations	It is important to determine how unbalanced a data set may be. Unbalanced data sets can be concern for SVM, while not as much for KNN. For the given data set, extreme observations are present and at medium risk.	Medium
Missing Data	Dealing with missing data prior to using SVM and KNN is especially important, while it may not as significant of an issue for Random Forest. For the given data set, missing data must be addressed and is at medium risk.	Medium
Size of Data	The given data set of 30,000+ records is not of too concern for the machine learning algorithms, since often these are run on much larger volumes of data. For the given data set, the overall size is low risk.	Low
Dimensionality of Data	Understanding the number of significant features within a data sets enables a smarter approach to choosing the right model. Those that can withstand high dimensionality like SVM will often perform better than those such as KNN. For the given data set, since there are multiple columns with many levels, the dimensionality is medium risk.	Medium

3) Generate the following tables:

a) calculate 1) the mean hours-per-week and 2) the median education-num for different age ranges; use bins of width 10, i.e, ages 20-30, 30-40, etc.

b) calculate the median age for males and females, broken down by occupation.

Part A:
The median education stays rather constant across the age bucket, while the mean hours per week scales up in the mid 30s to mid 40s range then gradually decreases with age.

	Age Bucket	Mean Hours / Week	Median Education
1	[10,20)	26.87689	9
2	[20,30)	38.61175	10
3	[30,40)	43.23894	10
4	[40,50)	43.23038	10
5	[50,60)	42.70914	10
6	[60,70)	36.14739	9
7	[70,80)	26.77756	9
8	[80,90]	28.47934	9

Part B:
It is interesting to note that there are no records for females in the Armed-Forces occupation.

	Occupation	Median Female Age	Median Male Age
1	Adm-clerical	35	36.0
2	Armed-Forces	NA	29.0
3	Craft-repair	39	38.0
4	Exec-managerial	39	42.0
5	Farming-fishing	39	39.0
6	Handlers-cleaners	32	28.0
7	Machine-op-inspct	37	36.0
8	Other-service	33	31.0
9	Priv-house-serv	41	23.0
10	Prof-specialty	38	41.0
11	Protective-serv	33	37.0
12	Sales	29	38.0
13	Tech-support	33	37.0
14	Transport-moving	37	39.0
15	Unknown	29	46.5

Data Preparation:

4) For this question, you will be preparing the data contained in "income_prediction.data" for training and testing a K-nearest neighbors model. Assume that you have a Linux command-line executable for running KNN. The command 'knn --help' generates the following output:

usage: knn --training_in <trainfile> --training_labels_in <trainlabels> --testing_in <testfile>

class labels of the rows of <testfile> will be printed to standard output.

Prepare the files <trainfile>, <trainlabels> and <testfile>. How does the choice of K-nearest neighbors as a machine learning method influence your representation of the data?

The choice of KNN as a machine learning method influences my representation in a variety of ways. First, I want to be aware of the curse of dimensionality. I do not want to inundate the KNN algorithm with too many predictors that may carry different significance. To remedy this issue, I can weight the predictors, choose significant predictors to include, and bucket categorical features that may have many levels. In my approach, I decided to remove the fnlwgt feature to keep more likely significant predictors and bucket the high dimensionality categories, such as the native country and work class attributes.

Furthermore, I normalized the numerical features to ensure that all the measurements will be assessed as on the same scale. Third, I examined the missing data analysis from above. For the native country and work class unknowns, the same columns that I bucketed, I made those the majority class of United States and Private. These values were overwhelmingly dominant in the data set, and thus have a high likelihood of being such values. To determine how to assign the missing occupation values, I examined what work class they had; however, most of those values overlapped with an unknown workclass. My approach would have been to assign the most common occupation for the most common work class, but there was overlap in both of these as unknown values. Next, I examined the education level associated with the unknown occupations; in this case, the two highest most prevalent values were high school and some college. I assigned the missing values to the most common occupation level given the education level; for those not with high school or college education levels, I marked them as the majority occupation of Prof-specialty.

Once the data was set, I divided the training and the test data into a 2/3 training to 1/3 testing allocation using every third row as a test data point. Note, that the headers are included as the first row in the training and test data file. The output of the test labels are included as a data file as well if needed.