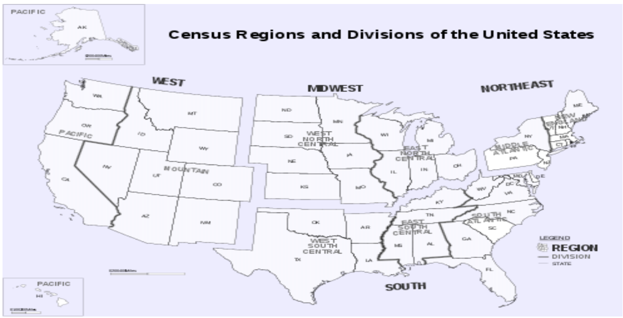
Full Name	:	Peter Sunny Shanthveer Markappa
Student No.	:	R00208303
Subject	:	Practical Machine Learning
Assignment	:	02
Data Set Used	:	Adult Data Set

****

Introduction

United states census data is selected for study, data extraction was done by Barry Becker from the 1994 Census database. The United States Census Bureau (USCB) is the main agency which comes under the U.S federal statistical system (USFSS) is responsible for data collection of American peoples, which is directly under the control of the US president. The population census is considered as a part of the constitution which has to be conducted at least once in ten years. The United State census bureau defined 9 statistical divisions in the country which are distributed in 4 regions as follows.



(Source : https://en.wikipedia.org/wiki/United_States_Census_Bureau)

Why this study only:

The importance of the study,

1. To know the new area where housing and public facilities are needed.
2. To know the demographic distribution of communities in the country.
3. For better planning of cities and transportation systems.
4. Providing domestic facilities such as police stations, electricity, schools ,colleges and hospitals.
5. For creating localised regions for conducting the election for people's choice to choose the leader in democratic country.

--Next Page--

Detailed Descriptions of the data.

The datasets contain 48842 records with six features continuous and eight features are nominal in nature.

A. **Continuous features are,**

1.age: continuous.

2. fnlwgt: continuous.

3. education-num: continuous.

4. capital-gain: continuous.

5.capital-loss: continuous.

6. hours-per-week: continuous.

B. **Nominal Features are,**

1. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
2. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
3. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
4. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
5. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
6. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
7. sex: Female, Male.
8. Native-country.

C.Target Feature

1. income: >50K, <=50K.

Objective:

Building a binary classification model, i.e which can predict/classify an individual based on income that is >50k or <= 50k for the given input data.

--Next Page--

Report - Research

From the detailed statistical descriptions of the data revealed major pre-processing which are need to address before predictive model development and the findings are

a. Handling Of special characters.
b. Appropriate imputation of null value.
c. Handling of outlier values.
d. Handling imbalanced datasets.
e. Optimization of Hyper-parameters.

The current study utilizes 'sklearn'/scikit-learn python library for processing the data.

A. Handling of Special Characters.

Some of the nominal features of the datasets contain special characters such as "?" in the data which will be replaced with null value. This operation can be performed using the pandas library using replace function with regular expression argument to the dataframe.

B. Appropriate imputation of null value.

From the above procedure we will end up creating null/empty cells, these cells/values should be treated. Either we can drop the missing value records from the further analysis, because of which we end up losing the data/information and another way is imputing missing values with appropriate values such mean or mode or some constant values. In the current study missing cells/values are identified in categorical features are education,occupation and native-country. In these features imputation are carried out using SimpleImputer, a function in sklearn library with mode value is imputed.

C. Handling of outlier values.

In the current study outliers are identified using boxplot visualization as shown in the following figure. Outliers identified in the features are educational-num, age, fnlwgt, capital-gain and capital-loss.

Educational-num : As shown in the first row of figure 1 box plot the observations less than the 5 are considered as an outlier; these values are replaced with the mean value.

age: From the below chart's second row box plot it is clear that age greater than the 80 years is considered as an outlier, these values are treated with the mean value.

Hours-per-week : In this variable there are more outliers identified in lower and upper bound with skewness value of 0.2. The distribution of the variable is found to be normal hence no outlier treatment is carried out.

Fnlwgt : From the below chart the fourth row box plot shows a large number of outliers in the data with skewness of 1.4, When keenly looked at the data all observations are found to be unique hence no treatment is carried out.

Capital - loss and Capital gain :In these two features behaviour is found to be more or less same with a large number of outliers in the data from plots, but approximately 97% of observations are zero only 3% observations are non-zero hence no outlier treatment was carried out.

D. **Optimization of Hyper-parameters :** GridSearchCV function is used to tune the hyper parameters for randomforestclassifier.

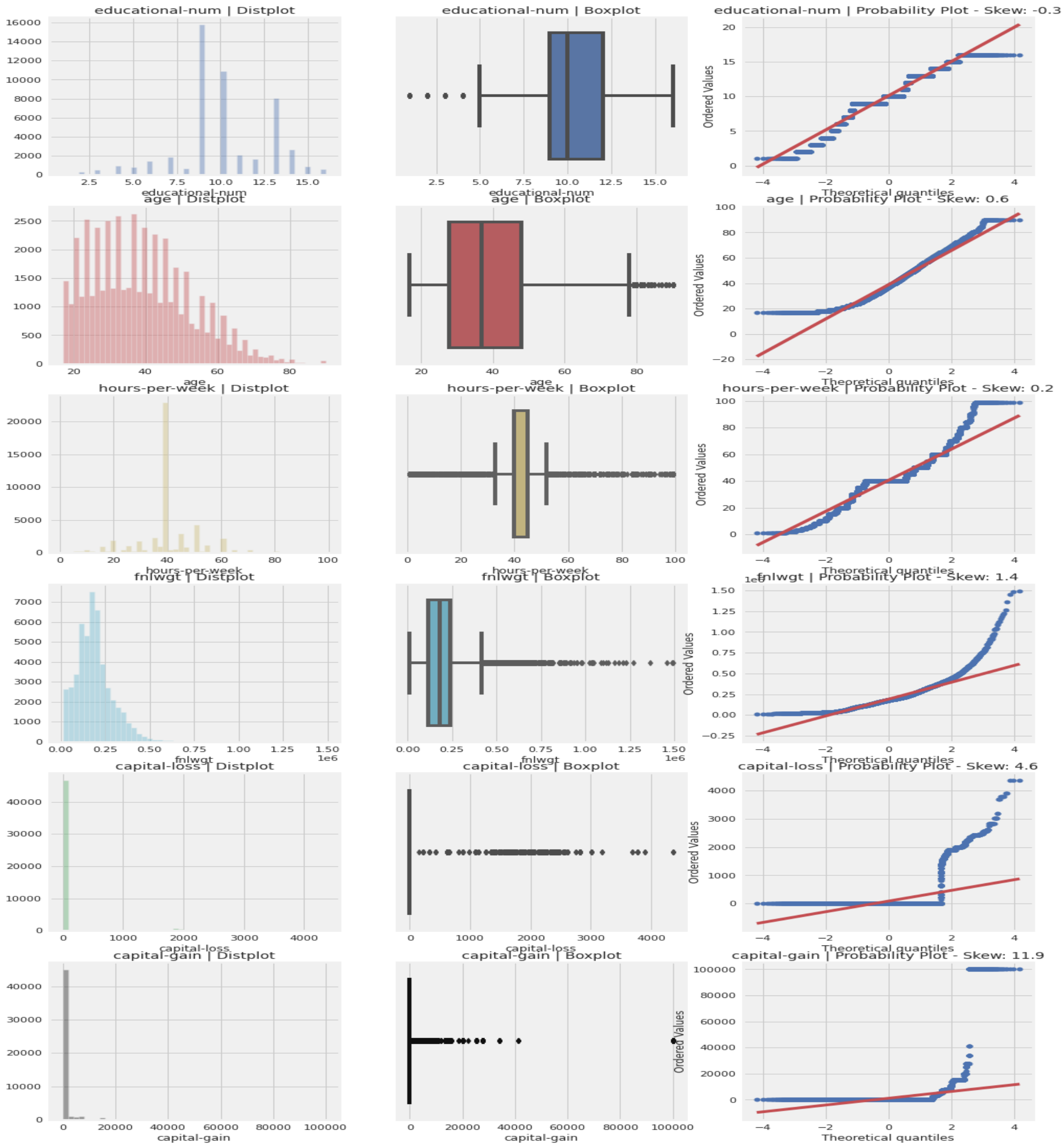
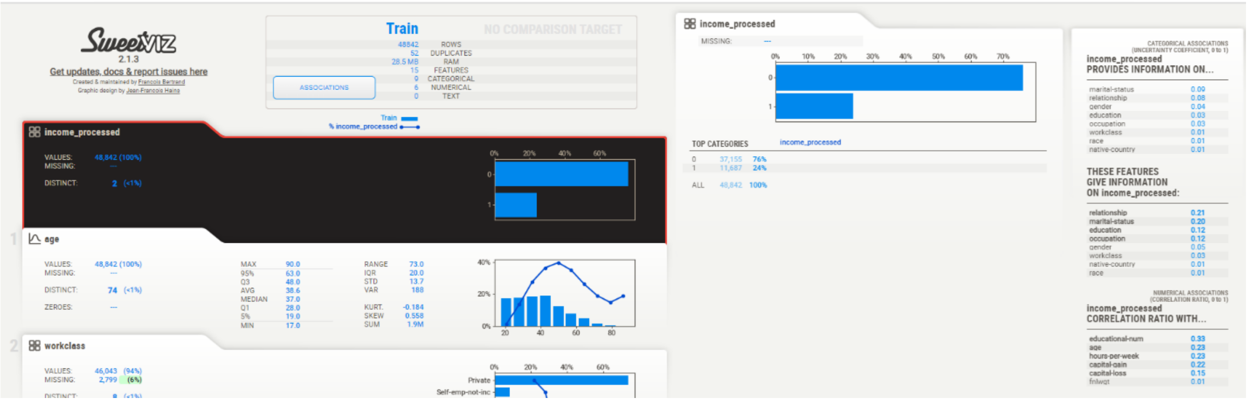


Fig:1

Chart Created using Sweetviz



Train

48842	ROWS
52	DUPLICATES
28.5 MB	RAM
15	FEATURES
9	CATEGORICAL
6	NUMERICAL
0	TEXT

income_processed

VALUES: 48,842 (100%)
MISSING: ---
DISTINCT: 2 (<1%)

1 age

VALUES:	48,842 (100%)	MAX	90.0	RANGE	73.0
MISSING:	---	95%	63.0	IQR	20.0
		Q3	48.0	STD	13.7
DISTINCT:	74 (<1%)	AVG	38.6	VAR	188
		MEDIAN	37.0		
ZEROES:	---	Q1	28.0	KURT.	-0.184
		5%	19.0	SKEW	0.558
		MIN	17.0	SUM	1.9M

2 workclass

VALUES: 46,043 (94%)
MISSING: 2,799 (6%)
DISTINCT: 8 (<1%)

income_processed

MISSING: ---

TOP CATEGORIES		
0	37,155	76%
1	11,687	24%
ALL	48,842	100%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

income_processed PROVIDES INFORMATION ON...

marital-status	0.09
relationship	0.08
gender	0.04
education	0.03
occupation	0.03
workclass	0.01
race	0.01
native-country	0.01

THESE FEATURES GIVE INFORMATION ON income_processed:

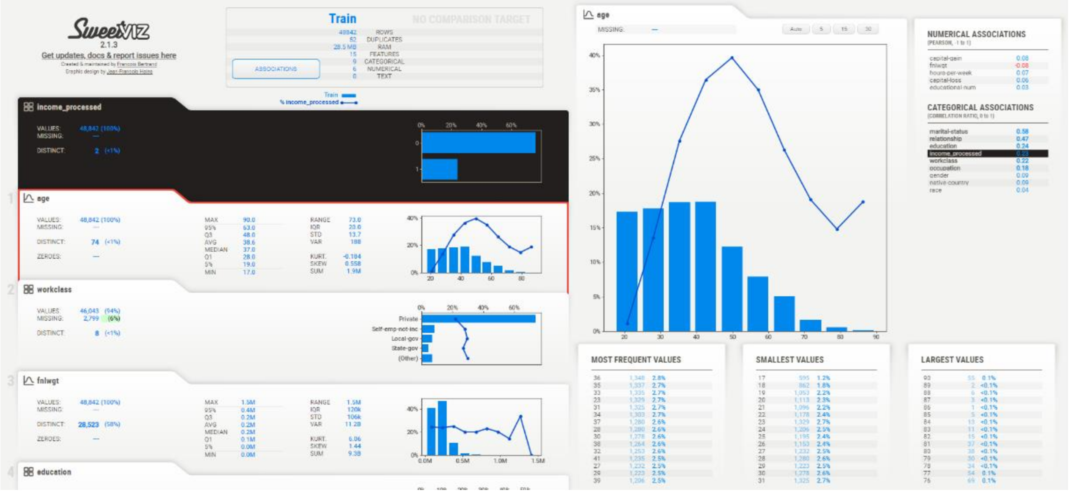
relationship	0.21
marital-status	0.20
education	0.12
occupation	0.12
gender	0.05
workclass	0.03
native-country	0.01
race	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

income_processed CORRELATION RATIO WITH...

educational-num	0.33
age	0.23
hours-per-week	0.23
capital-gain	0.22
capital-loss	0.15
fnlwgt	0.01

Income_Processed



Train

48842	ROWS
52	DUPLICATES
28.5 MB	RAM
15	FEATURES
9	CATEGORICAL
6	NUMERICAL
0	TEXT

income_processed

VALUES: 48,842 (100%)
MISSING: ---
DISTINCT: 2 (<1%)

1 age

VALUES:	48,842 (100%)	MAX	90.0	RANGE	73.0
MISSING:	---	95%	63.0	IQR	20.0
		Q3	48.0	STD	13.7
DISTINCT:	74 (<1%)	AVG	38.6	VAR	188
		MEDIAN	37.0		
ZEROES:	---	Q1	28.0	KURT.	-0.184
		5%	19.0	SKEW	0.558
		MIN	17.0	SUM	1.9M

2 workclass

VALUES: 46,043 (94%)
MISSING: 2,799 (6%)
DISTINCT: 8 (<1%)

3 fnlwgt

VALUES:	48,842 (100%)	MAX	1.5M	RANGE	1.5M
MISSING:	---	95%	0.4M	IQR	120k
		Q3	0.2M	STD	106k
DISTINCT:	28,523 (58%)	AVG	0.2M	VAR	11.2B
		MEDIAN	0.2M		
ZEROES:	---	Q1	0.1M	KURT.	6.06
		5%	0.0M	SKEW	1.44
		MIN	0.0M	SUM	9.3B

4 education

age

MISSING: ---

NUMERICAL ASSOCIATIONS (PEARSON, -1 to 1)

capital-gain	0.08
fnlwgt	-0.08
hours-per-week	0.07
capital-loss	0.06
educational-num	0.03

CATEGORICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

marital-status	0.58
relationship	0.47
education	0.24
income_processed	0.23
workclass	0.22
occupation	0.18
gender	0.09
native-country	0.09
race	0.04

MOST FREQUENT VALUES

36	1,348	2.8%
35	1,337	2.7%
33	1,335	2.7%
23	1,329	2.7%
31	1,325	2.7%
34	1,303	2.7%
37	1,280	2.6%
28	1,280	2.6%
30	1,278	2.6%
38	1,264	2.6%
32	1,253	2.6%
41	1,235	2.5%
27	1,232	2.5%
29	1,223	2.5%
39	1,206	2.5%

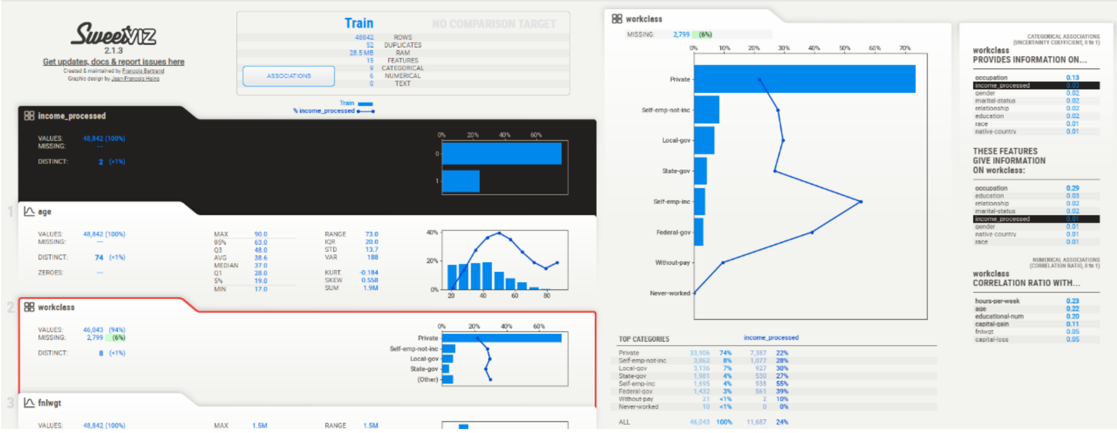
SMALLEST VALUES

17	595	1.2%
18	862	1.8%
19	1,053	2.2%
20	1,113	2.3%
21	1,096	2.2%
22	1,178	2.4%
23	1,329	2.7%
24	1,206	2.5%
25	1,195	2.4%
26	1,153	2.4%
27	1,232	2.5%
28	1,280	2.6%
29	1,223	2.5%
30	1,278	2.6%
31	1,325	2.7%

LARGEST VALUES

90	55	0.1%
89	2	<0.1%
88	6	<0.1%
87	3	<0.1%
86	1	<0.1%
85	5	<0.1%
84	13	<0.1%
83	11	<0.1%
82	15	<0.1%
81	37	<0.1%
80	38	<0.1%
79	30	<0.1%
78	34	<0.1%
77	54	0.1%
76	69	0.1%

Age Details



Train

48842	ROWS
52	DUPLICATES
28.5 MB	RAM
15	FEATURES
9	CATEGORICAL
6	NUMERICAL
0	TEXT

income_processed

VALUES: 48,842 (100%)
MISSING: ---
DISTINCT: 2 (<1%)

1 age

VALUES:	48,842 (100%)	MAX	90.0	RANGE	73.0
MISSING:	---	95%	63.0	IQR	20.0
		Q3	48.0	STD	13.7
DISTINCT:	74 (<1%)	AVG	38.6	VAR	188
		MEDIAN	37.0		
ZEROES:	---	Q1	28.0	KURT.	-0.184
		5%	19.0	SKEW	0.558
		MIN	17.0	SUM	1.9M

2 workclass

VALUES: 46,043 (94%)
MISSING: 2,799 (6%)
DISTINCT: 8 (<1%)

3 fnlwgt

VALUES: 48,842 (100%) MAX 1.5M RANGE 1.5M

workclass

MISSING: 2,799 (6%)

TOP CATEGORIES			income_processed	
Private	33,906	74%	7,387	22%
Self-emp-not-inc	3,862	8%	1,077	28%
Local-gov	3,136	7%	927	30%
State-gov	1,981	4%	530	27%
Self-emp-inc	1,695	4%	938	55%
Federal-gov	1,432	3%	561	39%
Without-pay	21	<1%	2	10%
Never-worked	10	<1%	0	0%
ALL	46,043	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

workclass PROVIDES INFORMATION ON...

occupation	0.13
income_processed	0.03
gender	0.02
marital-status	0.02
relationship	0.02
education	0.02
race	0.01
native-country	0.01

THESE FEATURES GIVE INFORMATION ON workclass:

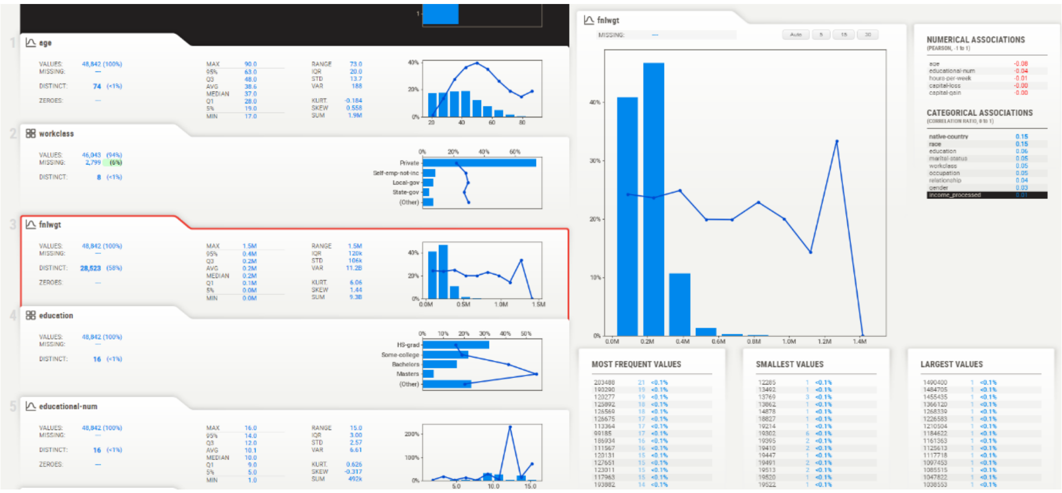
occupation	0.29
education	0.03
relationship	0.02
marital-status	0.02
income_processed	0.01
gender	0.01
native-country	0.01
race	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

workclass CORRELATION RATIO WITH...

hours-per-week	0.23
age	0.22
educational-num	0.20
capital-gain	0.11
fnlwgt	0.05
capital-loss	0.05

WorkClass



```
1 age
VALUES:    48,842 (100%)     MAX     90.0     RANGE  73.0
MISSING:   —                 95%     63.0     IQR    20.0
                             Q3      48.0     STD    13.7
DISTINCT:  74  (<1%)         AVG     38.6     VAR    188
                             MEDIAN  37.0
ZEROES:    —                 Q1      28.0     KURT.  -0.184
                             5%      19.0     SKEW   0.558
                             MIN     17.0     SUM    1.9M

2 workclass
VALUES:    46,043 (94%)
MISSING:   2,799  (6%)
DISTINCT:  8  (<1%)

3 fnlwgt
VALUES:    48,842 (100%)     MAX     1.5M     RANGE  1.5M
MISSING:   —                 95%     0.4M     IQR    120k
                             Q3      0.2M     STD    106k
DISTINCT:  28,523 (58%)      AVG     0.2M     VAR    11.2B
                             MEDIAN  0.2M
ZEROES:    —                 Q1      0.1M     KURT.  6.06
                             5%      0.0M     SKEW   1.44
                             MIN     0.0M     SUM    9.3B

4 education
VALUES:    48,842 (100%)
MISSING:   —
DISTINCT:  16  (<1%)

5 educational-num
VALUES:    48,842 (100%)     MAX     16.0     RANGE  15.0
MISSING:   —                 95%     14.0     IQR    3.00
                             Q3      12.0     STD    2.57
DISTINCT:  16  (<1%)         AVG     10.1     VAR    6.61
                             MEDIAN  10.0
ZEROES:    —                 Q1      9.0      KURT.  0.626
                             5%      5.0      SKEW   -0.317
                             MIN     1.0      SUM    492k
```

fnlwgt

MISSING: —

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

age	-0.08
educational-num	-0.04
hours-per-week	-0.01
capital-loss	-0.00
capital-gain	-0.00

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

native-country	0.15
race	0.15
education	0.06
marital-status	0.05
workclass	0.05
occupation	0.05
relationship	0.04
gender	0.03
income_processed	0.01

MOST FREQUENT VALUES

203488	21	<0.1%
190290	19	<0.1%
120277	19	<0.1%
125892	18	<0.1%
126569	18	<0.1%
126675	17	<0.1%
113364	17	<0.1%
99185	17	<0.1%
186934	16	<0.1%
111567	16	<0.1%
120131	15	<0.1%
127651	15	<0.1%
123011	15	<0.1%
117963	15	<0.1%
193882	14	<0.1%

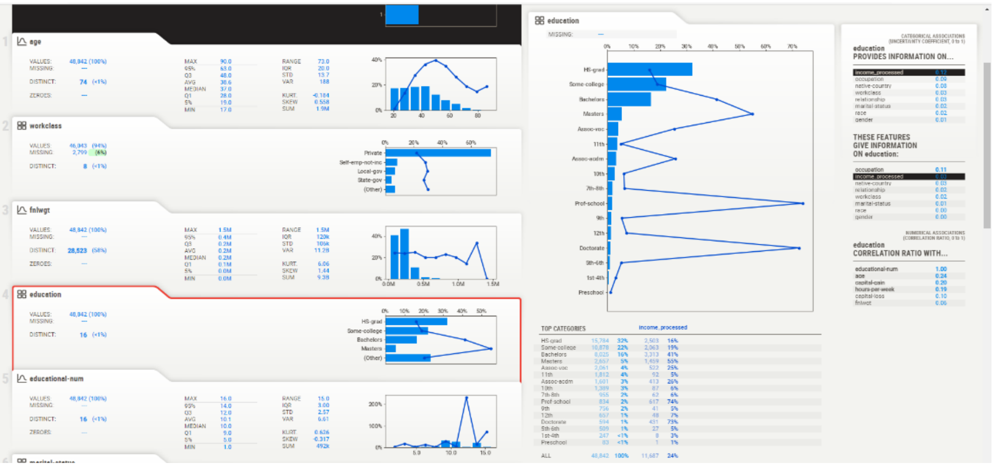
SMALLEST VALUES

12285	1	<0.1%
13492	1	<0.1%
13769	3	<0.1%
13862	1	<0.1%
14878	1	<0.1%
18827	1	<0.1%
19214	1	<0.1%
19302	6	<0.1%
19395	2	<0.1%
19410	2	<0.1%
19447	1	<0.1%
19491	2	<0.1%
19513	2	<0.1%
19520	1	<0.1%
19522	1	<0.1%

LARGEST VALUES

1490400	1	<0.1%
1484705	1	<0.1%
1455435	1	<0.1%
1366120	1	<0.1%
1268339	1	<0.1%
1226583	1	<0.1%
1210504	1	<0.1%
1184622	1	<0.1%
1161363	1	<0.1%
1125613	1	<0.1%
1117718	1	<0.1%
1097453	1	<0.1%
1085515	1	<0.1%
1047822	1	<0.1%
1038553	1	<0.1%

Fnlwgt



```
1 age
VALUES:    48,842 (100%)     MAX     90.0     RANGE  73.0
MISSING:   —                 95%     63.0     IQR    20.0
                             Q3      48.0     STD    13.7
DISTINCT:  74  (<1%)         AVG     38.6     VAR    188
                             MEDIAN  37.0
ZEROES:    —                 Q1      28.0     KURT.  -0.184
                             5%      19.0     SKEW   0.558
                             MIN     17.0     SUM    1.9M

2 workclass
VALUES:    46,043 (94%)
MISSING:   2,799  (6%)
DISTINCT:  8  (<1%)

3 fnlwgt
VALUES:    48,842 (100%)     MAX     1.5M     RANGE  1.5M
MISSING:   —                 95%     0.4M     IQR    120k
                             Q3      0.2M     STD    106k
DISTINCT:  28,523 (58%)      AVG     0.2M     VAR    11.2B
                             MEDIAN  0.2M
ZEROES:    —                 Q1      0.1M     KURT.  6.06
                             5%      0.0M     SKEW   1.44
                             MIN     0.0M     SUM    9.3B

4 education
VALUES:    48,842 (100%)
MISSING:   —
DISTINCT:  16  (<1%)

5 educational-num
VALUES:    48,842 (100%)     MAX     16.0     RANGE  15.0
MISSING:   —                 95%     14.0     IQR    3.00
                             Q3      12.0     STD    2.57
DISTINCT:  16  (<1%)         AVG     10.1     VAR    6.61
                             MEDIAN  10.0
ZEROES:    —                 Q1      9.0      KURT.  0.626
                             5%      5.0      SKEW   -0.317
                             MIN     1.0      SUM    492k
```

education

MISSING: —

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

education
PROVIDES INFORMATION ON...

income_processed	0.12
occupation	0.09
native-country	0.08
workclass	0.03
relationship	0.03
marital-status	0.02
race	0.02
gender	0.01

THESE FEATURES GIVE INFORMATION ON education:

occupation	0.11
income_processed	0.03
native-country	0.03
relationship	0.02
workclass	0.02
marital-status	0.01
race	0.00
gender	0.00

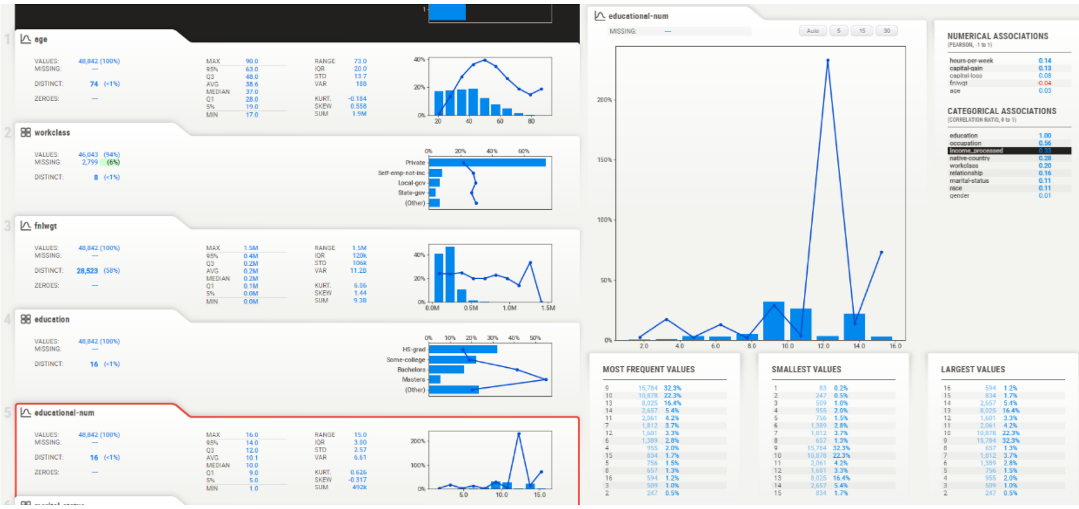
NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

education
CORRELATION RATIO WITH...

educational-num	1.00
age	0.24
capital-gain	0.20
hours-per-week	0.19
capital-loss	0.10
fnlwgt	0.06

TOP CATEGORIES			income_processed	
HS-grad	15,784	32%	2,503	16%
Some-college	10,878	22%	2,063	19%
Bachelors	8,025	16%	3,313	41%
Masters	2,657	5%	1,459	55%
Assoc-voc	2,061	4%	522	25%
11th	1,812	4%	92	5%
Assoc-acdm	1,601	3%	413	26%
10th	1,389	3%	87	6%
7th-8th	955	2%	62	6%
Prof-school	834	2%	617	74%
9th	756	2%	41	5%
12th	657	1%	48	7%
Doctorate	594	1%	431	73%
5th-6th	509	1%	27	5%
1st-4th	247	<1%	8	3%
Preschool	83	<1%	1	1%
ALL	48,842	100%	11,687	24%

Education



```
1 age
VALUES:    48,842 (100%)     MAX     90.0     RANGE  73.0
MISSING:   —                 95%     63.0     IQR    20.0
                             Q3      48.0     STD    13.7
DISTINCT:  74  (<1%)         AVG     38.6     VAR    188
                             MEDIAN  37.0
ZEROES:    —                 Q1      28.0     KURT.  -0.184
                             5%      19.0     SKEW   0.558
                             MIN     17.0     SUM    1.9M

2 workclass
VALUES:    46,043 (94%)
MISSING:   2,799  (6%)
DISTINCT:  8  (<1%)

3 fnlwgt
VALUES:    48,842 (100%)     MAX     1.5M     RANGE  1.5M
MISSING:   —                 95%     0.4M     IQR    120k
                             Q3      0.2M     STD    106k
DISTINCT:  28,523 (58%)      AVG     0.2M     VAR    11.2B
                             MEDIAN  0.2M
ZEROES:    —                 Q1      0.1M     KURT.  6.06
                             5%      0.0M     SKEW   1.44
                             MIN     0.0M     SUM    9.3B

4 education
VALUES:    48,842 (100%)
MISSING:   —
DISTINCT:  16  (<1%)

5 educational-num
VALUES:    48,842 (100%)     MAX     16.0     RANGE  15.0
MISSING:   —                 95%     14.0     IQR    3.00
                             Q3      12.0     STD    2.57
DISTINCT:  16  (<1%)         AVG     10.1     VAR    6.61
                             MEDIAN  10.0
ZEROES:    —                 Q1      9.0      KURT.  0.626
                             5%      5.0      SKEW   -0.317
                             MIN     1.0      SUM    492k
```

educational-num

MISSING: —

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

hours-per-week	0.14
capital-gain	0.13
capital-loss	0.08
fnlwgt	-0.04
age	0.03

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

education	1.00
occupation	0.56
income_processed	0.33
native-country	0.28
workclass	0.20
relationship	0.16
marital-status	0.11
race	0.11
gender	0.01

MOST FREQUENT VALUES

9	15,784	32.3%
10	10,878	22.3%
13	8,025	16.4%
14	2,657	5.4%
11	2,061	4.2%
7	1,812	3.7%
12	1,601	3.3%
6	1,389	2.8%
4	955	2.0%
15	834	1.7%
5	756	1.5%
8	657	1.3%
16	594	1.2%
3	509	1.0%
2	247	0.5%

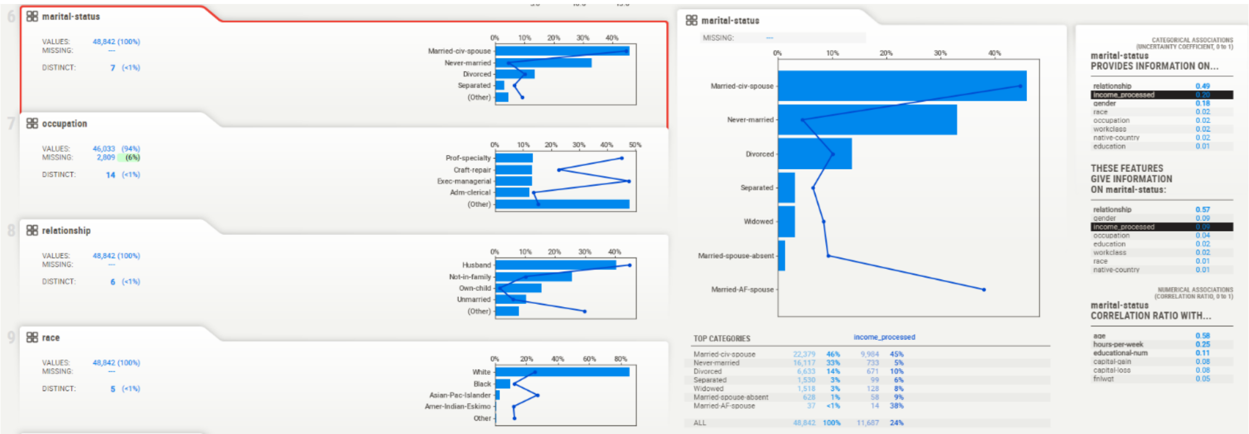
SMALLEST VALUES

1	83	0.2%
2	247	0.5%
3	509	1.0%
4	955	2.0%
5	756	1.5%
6	1,389	2.8%
7	1,812	3.7%
8	657	1.3%
9	15,784	32.3%
10	10,878	22.3%
11	2,061	4.2%
12	1,601	3.3%
13	8,025	16.4%
14	2,657	5.4%
15	834	1.7%

LARGEST VALUES

16	594	1.2%
15	834	1.7%
14	2,657	5.4%
13	8,025	16.4%
12	1,601	3.3%
11	2,061	4.2%
10	10,878	22.3%
9	15,784	32.3%
8	657	1.3%
7	1,812	3.7%
6	1,389	2.8%
5	756	1.5%
4	955	2.0%
3	509	1.0%
2	247	0.5%

Educational-num



6 marital-status

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 7 (<1%)

7 occupation

VALUES: 46,033 (94%)
MISSING: 2,809 (6%)
DISTINCT: 14 (<1%)

8 relationship

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 6 (<1%)

9 race

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 5 (<1%)

marital-status

MISSING: —

TOP CATEGORIES			income_processed	
Married-civ-spouse	22,379	46%	9,984	45%
Never-married	16,117	33%	733	5%
Divorced	6,633	14%	671	10%
Separated	1,530	3%	99	6%
Widowed	1,518	3%	128	8%
Married-spouse-absent	628	1%	58	9%
Married-AF-spouse	37	<1%	14	38%
ALL	48,842	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

marital-status PROVIDES INFORMATION ON...

relationship	0.49
income_processed	0.20
gender	0.18
race	0.02
occupation	0.02
workclass	0.02
native-country	0.02
education	0.01

THESE FEATURES GIVE INFORMATION ON marital-status:

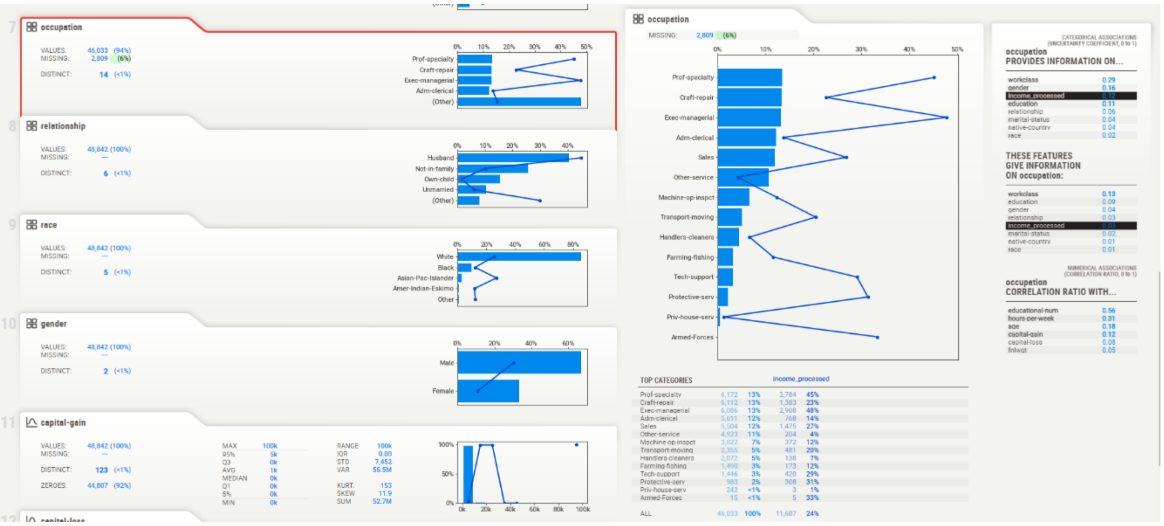
relationship	0.57
gender	0.09
income_processed	0.09
occupation	0.04
education	0.02
workclass	0.02
race	0.01
native-country	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

marital-status CORRELATION RATIO WITH...

age	0.58
hours-per-week	0.25
educational-num	0.11
capital-gain	0.08
capital-loss	0.08
fnlwgt	0.05

Marital-status



7 occupation

VALUES: 46,033 (94%)
MISSING: 2,809 (6%)
DISTINCT: 14 (<1%)

8 relationship

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 6 (<1%)

9 race

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 5 (<1%)

10 gender

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 2 (<1%)

11 capital-gain

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 123 (<1%)
ZEROES: 44,807 (92%)

MAX	100k	RANGE	100k
95%	5k	IQR	0.00
Q3	0k	STD	7,452
AVG	1k	VAR	55.5M
MEDIAN	0k		
Q1	0k	KURT.	153
5%	0k	SKEW	11.9
MIN	0k	SUM	52.7M

12 capital-loss

occupation

MISSING: 2,809 (6%)

TOP CATEGORIES			income_processed	
Prof-specialty	6,172	13%	2,784	45%
Craft-repair	6,112	13%	1,383	23%
Exec-managerial	6,086	13%	2,908	48%
Adm-clerical	5,611	12%	768	14%
Sales	5,504	12%	1,475	27%
Other-service	4,923	11%	204	4%
Machine-op-inspct	3,022	7%	372	12%
Transport-moving	2,355	5%	481	20%
Handlers-cleaners	2,072	5%	138	7%
Farming-fishing	1,490	3%	173	12%
Tech-support	1,446	3%	420	29%
Protective-serv	983	2%	308	31%
Priv-house-serv	242	<1%	3	1%
Armed-Forces	15	<1%	5	33%
ALL	46,033	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

occupation PROVIDES INFORMATION ON...

workclass	0.29
gender	0.16
income_processed	0.12
education	0.11
relationship	0.06
marital-status	0.04
native-country	0.04
race	0.02

THESE FEATURES GIVE INFORMATION ON occupation:

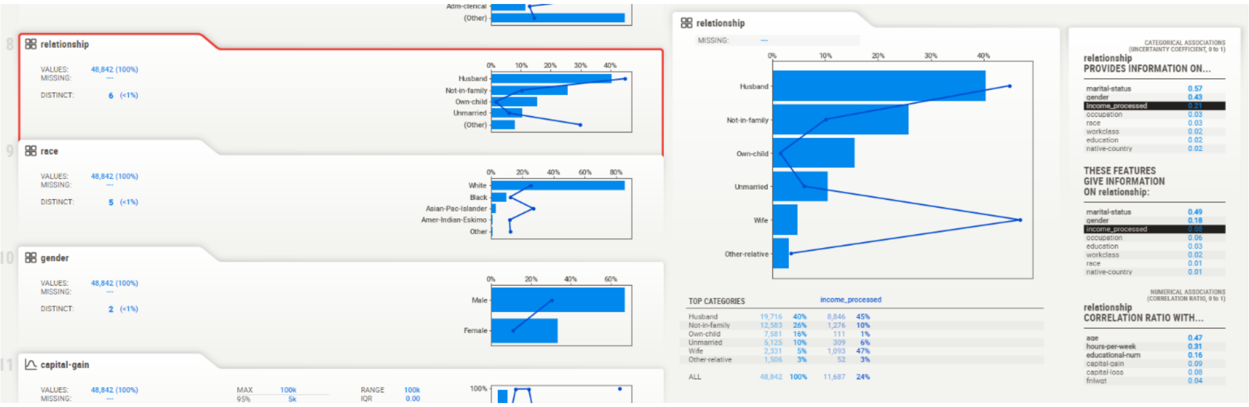
workclass	0.13
education	0.09
gender	0.04
relationship	0.03
income_processed	0.03
marital-status	0.02
native-country	0.01
race	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

occupation CORRELATION RATIO WITH...

educational-num	0.56
hours-per-week	0.31
age	0.18
capital-gain	0.12
capital-loss	0.08
fnlwgt	0.05

Occupation



8 relationship

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 6 (<1%)

9 race

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 5 (<1%)

10 gender

VALUES: 48,842 (100%)
MISSING: —
DISTINCT: 2 (<1%)

11 capital-gain

VALUES: 48,842 (100%)
MISSING: —

MAX	100k	RANGE	100k
95%	5k	IQR	0.00

relationship

MISSING: —

TOP CATEGORIES			income_processed	
Husband	19,716	40%	8,846	45%
Not-in-family	12,583	26%	1,276	10%
Own-child	7,581	16%	111	1%
Unmarried	5,125	10%	309	6%
Wife	2,331	5%	1,093	47%
Other-relative	1,506	3%	52	3%
ALL	48,842	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

relationship PROVIDES INFORMATION ON...

marital-status	0.57
gender	0.43
income_processed	0.21
occupation	0.03
race	0.03
workclass	0.02
education	0.02
native-country	0.02

THESE FEATURES GIVE INFORMATION ON relationship:

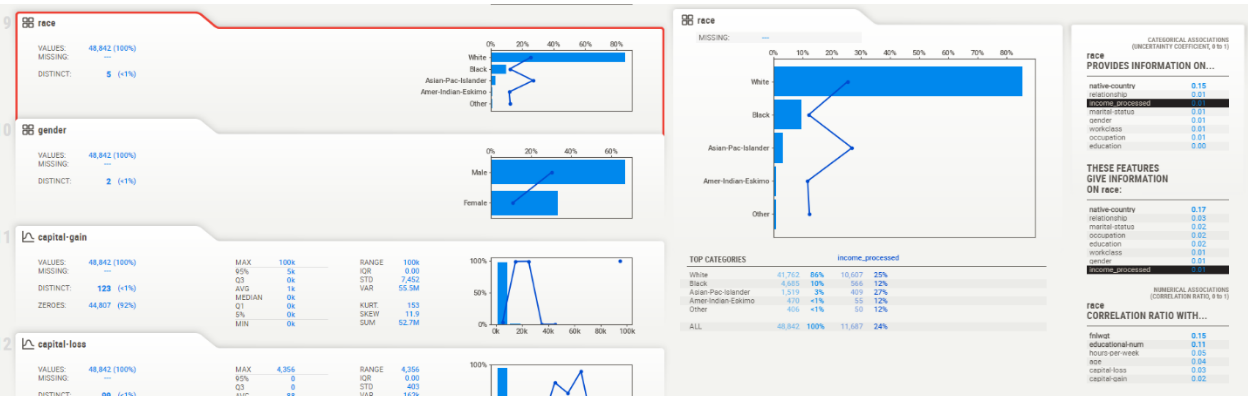
marital-status	0.49
gender	0.18
income_processed	0.08
occupation	0.06
education	0.03
workclass	0.02
race	0.01
native-country	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

relationship CORRELATION RATIO WITH...

age	0.47
hours-per-week	0.31
educational-num	0.16
capital-gain	0.09
capital-loss	0.08
fnlwgt	0.04

Relationship



9 race

VALUES:	48,842 (100%)
MISSING:	---
DISTINCT:	5 (<1%)

0 gender

VALUES:	48,842 (100%)
MISSING:	---
DISTINCT:	2 (<1%)

1 capital-gain

VALUES:	48,842 (100%)	MAX	100k	RANGE	100k
MISSING:	---	95%	5k	IQR	0.00
		Q3	0k	STD	7,452
DISTINCT:	123 (<1%)	AVG	1k	VAR	55.5M
		MEDIAN	0k		
ZEROES:	44,807 (92%)	Q1	0k	KURT.	153
		5%	0k	SKEW	11.9
		MIN	0k	SUM	52.7M

2 capital-loss

VALUES:	48,842 (100%)	MAX	4,356	RANGE	4,356
MISSING:	---	95%	0	IQR	0.00
		Q3	0	STD	403

race

MISSING: ---

TOP CATEGORIES			income_processed	
White	41,762	86%	10,607	25%
Black	4,685	10%	566	12%
Asian-Pac-Islander	1,519	3%	409	27%
Amer-Indian-Eskimo	470	<1%	55	12%
Other	406	<1%	50	12%
ALL	48,842	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

race PROVIDES INFORMATION ON...

native-country	0.15
relationship	0.01
income_processed	0.01
marital-status	0.01
gender	0.01
workclass	0.01
occupation	0.01
education	0.00

THESE FEATURES GIVE INFORMATION ON race:

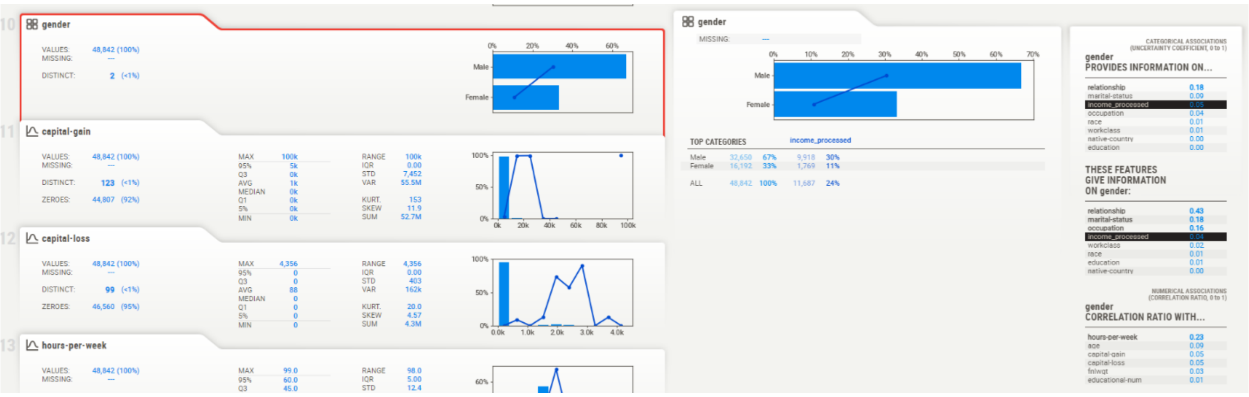
native-country	0.17
relationship	0.03
marital-status	0.02
occupation	0.02
education	0.02
workclass	0.01
gender	0.01
income_processed	0.01

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

race CORRELATION RATIO WITH...

fnlwgt	0.15
educational-num	0.11
hours-per-week	0.05
age	0.04
capital-loss	0.03
capital-gain	0.02

Race



10 gender

VALUES:	48,842 (100%)
MISSING:	---
DISTINCT:	2 (<1%)

11 capital-gain

VALUES:	48,842 (100%)	MAX	100k	RANGE	100k
MISSING:	---	95%	5k	IQR	0.00
		Q3	0k	STD	7,452
DISTINCT:	123 (<1%)	AVG	1k	VAR	55.5M
		MEDIAN	0k		
ZEROES:	44,807 (92%)	Q1	0k	KURT.	153
		5%	0k	SKEW	11.9
		MIN	0k	SUM	52.7M

12 capital-loss

VALUES:	48,842 (100%)	MAX	4,356	RANGE	4,356
MISSING:	---	95%	0	IQR	0.00
		Q3	0	STD	403
DISTINCT:	99 (<1%)	AVG	88	VAR	162k
		MEDIAN	0		
ZEROES:	46,560 (95%)	Q1	0	KURT.	20.0
		5%	0	SKEW	4.57
		MIN	0	SUM	4.3M

13 hours-per-week

VALUES:	48,842 (100%)	MAX	99.0	RANGE	98.0
MISSING:	---	95%	60.0	IQR	5.00
		Q3	45.0	STD	12.4

gender

MISSING: ---

TOP CATEGORIES			income_processed	
Male	32,650	67%	9,918	30%
Female	16,192	33%	1,769	11%
ALL	48,842	100%	11,687	24%

CATEGORICAL ASSOCIATIONS (UNCERTAINTY COEFFICIENT, 0 to 1)

gender PROVIDES INFORMATION ON...

relationship	0.18
marital-status	0.09
income_processed	0.05
occupation	0.04
race	0.01
workclass	0.01
native-country	0.00
education	0.00

THESE FEATURES GIVE INFORMATION ON gender:

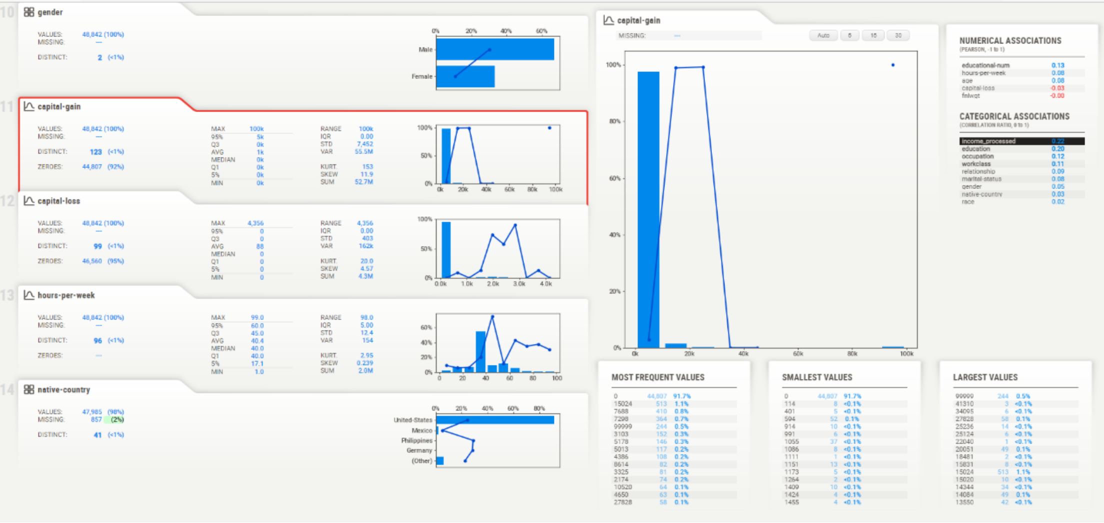
relationship	0.43
marital-status	0.18
occupation	0.16
income_processed	0.04
workclass	0.02
race	0.01
education	0.01
native-country	0.00

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

gender CORRELATION RATIO WITH...

hours-per-week	0.23
age	0.09
capital-gain	0.05
capital-loss	0.05
fnlwgt	0.03
educational-num	0.01

gender



10 gender

VALUES:	48,842 (100%)
MISSING:	---
DISTINCT:	2 (<1%)

11 capital-gain

VALUES:	48,842 (100%)	MAX	100k	RANGE	100k
MISSING:	---	95%	5k	IQR	0.00
		Q3	0k	STD	7,452
DISTINCT:	123 (<1%)	AVG	1k	VAR	55.5M
		MEDIAN	0k		
ZEROES:	44,807 (92%)	Q1	0k	KURT.	153
		5%	0k	SKEW	11.9
		MIN	0k	SUM	52.7M

12 capital-loss

VALUES:	48,842 (100%)	MAX	4,356	RANGE	4,356
MISSING:	---	95%	0	IQR	0.00
		Q3	0	STD	403
DISTINCT:	99 (<1%)	AVG	88	VAR	162k
		MEDIAN	0		
ZEROES:	46,560 (95%)	Q1	0	KURT.	20.0
		5%	0	SKEW	4.57
		MIN	0	SUM	4.3M

13 hours-per-week

VALUES:	48,842 (100%)	MAX	99.0	RANGE	98.0
MISSING:	---	95%	60.0	IQR	5.00
		Q3	45.0	STD	12.4
DISTINCT:	96 (<1%)	AVG	40.4	VAR	154
		MEDIAN	40.0		
ZEROES:	---	Q1	40.0	KURT.	2.95
		5%	17.1	SKEW	0.239
		MIN	1.0	SUM	2.0M

14 native-country

VALUES:	47,985 (98%)
MISSING:	857 (2%)
DISTINCT:	41 (<1%)

capital-gain

MISSING: ---

NUMERICAL ASSOCIATIONS (PEARSON, -1 to 1)

educational-num	0.13
hours-per-week	0.08
age	0.08
capital-loss	-0.03
fnlwgt	-0.00

CATEGORICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

income_processed	0.22
education	0.20
occupation	0.12
workclass	0.11
relationship	0.09
marital-status	0.08
gender	0.05
native-country	0.03
race	0.02

MOST FREQUENT VALUES

0	44,807	91.7%
15024	513	1.1%
7688	410	0.8%
7298	364	0.7%
99999	244	0.5%
3103	152	0.3%
5178	146	0.3%
5013	117	0.2%
4386	108	0.2%
8614	82	0.2%
3325	81	0.2%
2174	74	0.2%
10520	64	0.1%
4650	63	0.1%
27828	58	0.1%

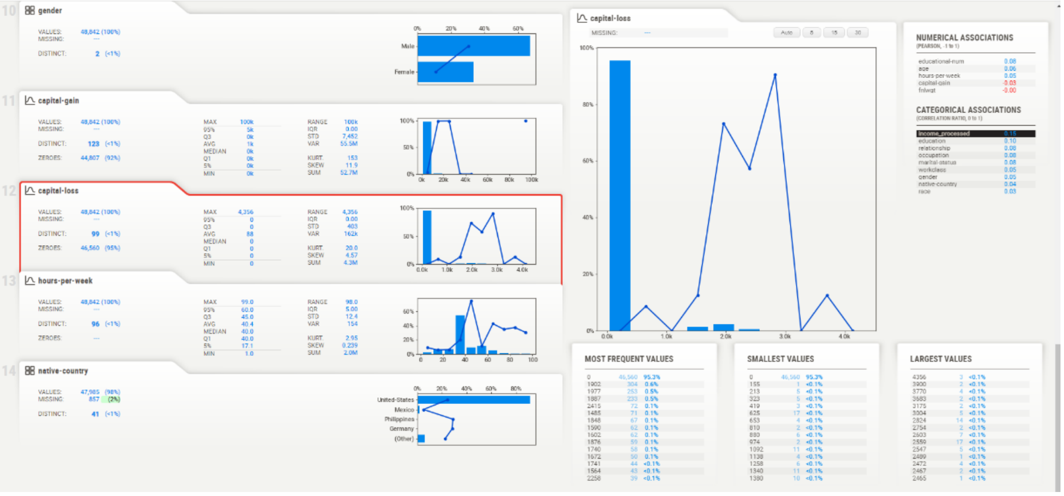
SMALLEST VALUES

0	44,807	91.7%
114	8	<0.1%
401	5	<0.1%
594	52	0.1%
914	10	<0.1%
991	6	<0.1%
1055	37	<0.1%
1086	8	<0.1%
1111	1	<0.1%
1151	13	<0.1%
1173	5	<0.1%
1264	2	<0.1%
1409	10	<0.1%
1424	4	<0.1%
1455	4	<0.1%

LARGEST VALUES

99999	244	0.5%
41310	3	<0.1%
34095	6	<0.1%
27828	58	0.1%
25236	14	<0.1%
25124	6	<0.1%
22040	1	<0.1%
20051	49	0.1%
18481	2	<0.1%
15831	8	<0.1%
15024	513	1.1%
15020	10	<0.1%
14344	34	<0.1%
14084	49	0.1%
13550	42	<0.1%

Capital-gain



```
10 gender
   VALUES:    48,842 (100%)
   MISSING:   ---
   DISTINCT:  2     (<1%)

11 capital-gain
   VALUES:    48,842 (100%)     MAX     100k     RANGE  100k
   MISSING:   ---               95%     5k       IQR    0.00
   DISTINCT:  123   (<1%)       Q3      0k       STD    7,452
   ZEROES:    44,807 (92%)      AVG     1k       VAR    55.5M
                                MEDIAN  0k
                                Q1      0k       KURT.  153
                                5%      0k       SKEW   11.9
                                MIN     0k       SUM    52.7M

12 capital-loss
   VALUES:    48,842 (100%)     MAX     4,356    RANGE  4,356
   MISSING:   ---               95%     0        IQR    0.00
   DISTINCT:  99    (<1%)       Q3      0        STD    403
   ZEROES:    46,560 (95%)      AVG     88       VAR    162k
                                MEDIAN  0
                                Q1      0        KURT.  20.0
                                5%      0        SKEW   4.57
                                MIN     0        SUM    4.3M

13 hours-per-week
   VALUES:    48,842 (100%)     MAX     99.0     RANGE  98.0
   MISSING:   ---               95%     60.0     IQR    5.00
   DISTINCT:  96    (<1%)       Q3      45.0     STD    12.4
   ZEROES:    ---               AVG     40.4     VAR    154
                                MEDIAN  40.0
                                Q1      40.0     KURT.  2.95
                                5%      17.1     SKEW   0.239
                                MIN     1.0      SUM    2.0M

14 native-country
   VALUES:    47,985 (98%)
   MISSING:   857    (2%)
   DISTINCT:  41     (<1%)
```

capital-loss

MISSING: ---

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

educational-num	0.08
age	0.06
hours-per-week	0.05
capital-gain	-0.03
fnlwgt	-0.00

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

income_processed	0.15
education	0.10
relationship	0.08
occupation	0.08
marital-status	0.08
workclass	0.05
gender	0.05
native-country	0.04
race	0.03

MOST FREQUENT VALUES

0	46,560	95.3%
1902	304	0.6%
1977	253	0.5%
1887	233	0.5%
2415	72	0.1%
1485	71	0.1%
1848	67	0.1%
1590	62	0.1%
1602	62	0.1%
1876	59	0.1%
1740	58	0.1%
1672	50	0.1%
1741	44	<0.1%
1564	43	<0.1%
2258	39	<0.1%

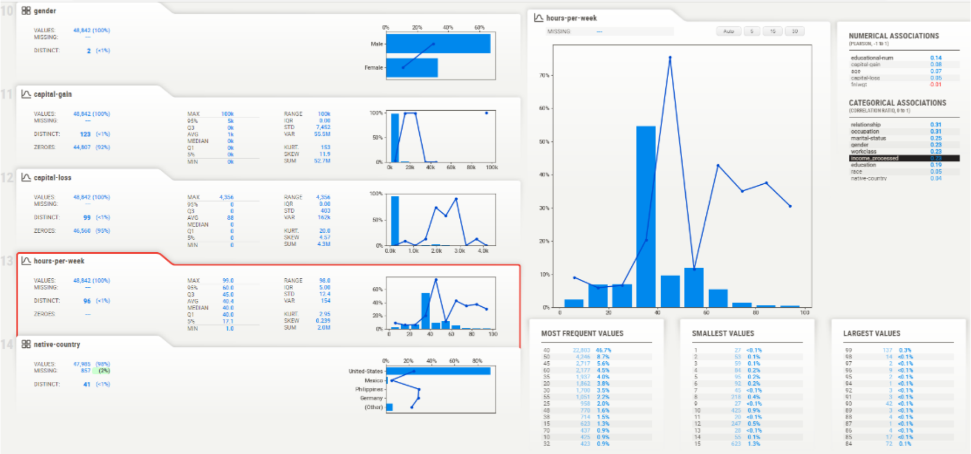
SMALLEST VALUES

0	46,560	95.3%
155	1	<0.1%
213	5	<0.1%
323	5	<0.1%
419	3	<0.1%
625	17	<0.1%
653	4	<0.1%
810	2	<0.1%
880	6	<0.1%
974	2	<0.1%
1092	11	<0.1%
1138	4	<0.1%
1258	6	<0.1%
1340	11	<0.1%
1380	10	<0.1%

LARGEST VALUES

4356	3	<0.1%
3900	2	<0.1%
3770	4	<0.1%
3683	2	<0.1%
3175	2	<0.1%
3004	5	<0.1%
2824	14	<0.1%
2754	2	<0.1%
2603	7	<0.1%
2559	17	<0.1%
2547	5	<0.1%
2489	1	<0.1%
2472	4	<0.1%
2467	2	<0.1%
2465	1	<0.1%

Capital-loss



```
10 gender
   VALUES:    48,842 (100%)
   MISSING:   ---
   DISTINCT:  2     (<1%)

11 capital-gain
   VALUES:    48,842 (100%)     MAX     100k     RANGE  100k
   MISSING:   ---               95%     5k       IQR    0.00
   DISTINCT:  123   (<1%)       Q3      0k       STD    7,452
   ZEROES:    44,807 (92%)      AVG     1k       VAR    55.5M
                                MEDIAN  0k
                                Q1      0k       KURT.  153
                                5%      0k       SKEW   11.9
                                MIN     0k       SUM    52.7M

12 capital-loss
   VALUES:    48,842 (100%)     MAX     4,356    RANGE  4,356
   MISSING:   ---               95%     0        IQR    0.00
   DISTINCT:  99    (<1%)       Q3      0        STD    403
   ZEROES:    46,560 (95%)      AVG     88       VAR    162k
                                MEDIAN  0
                                Q1      0        KURT.  20.0
                                5%      0        SKEW   4.57
                                MIN     0        SUM    4.3M

13 hours-per-week
   VALUES:    48,842 (100%)     MAX     99.0     RANGE  98.0
   MISSING:   ---               95%     60.0     IQR    5.00
   DISTINCT:  96    (<1%)       Q3      45.0     STD    12.4
   ZEROES:    ---               AVG     40.4     VAR    154
                                MEDIAN  40.0
                                Q1      40.0     KURT.  2.95
                                5%      17.1     SKEW   0.239
                                MIN     1.0      SUM    2.0M

14 native-country
   VALUES:    47,985 (98%)
   MISSING:   857    (2%)
   DISTINCT:  41     (<1%)
```

hours-per-week

MISSING: ---

NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

educational-num	0.14
capital-gain	0.08
age	0.07
capital-loss	0.05
fnlwgt	-0.01

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

relationship	0.31
occupation	0.31
marital-status	0.25
gender	0.23
workclass	0.23
income_processed	0.23
education	0.19
race	0.05
native-country	0.04

MOST FREQUENT VALUES

40	22,803	46.7%
50	4,246	8.7%
45	2,717	5.6%
60	2,177	4.5%
35	1,937	4.0%
20	1,862	3.8%
30	1,700	3.5%
55	1,051	2.2%
25	958	2.0%
48	770	1.6%
38	714	1.5%
15	623	1.3%
70	437	0.9%
10	425	0.9%
32	423	0.9%

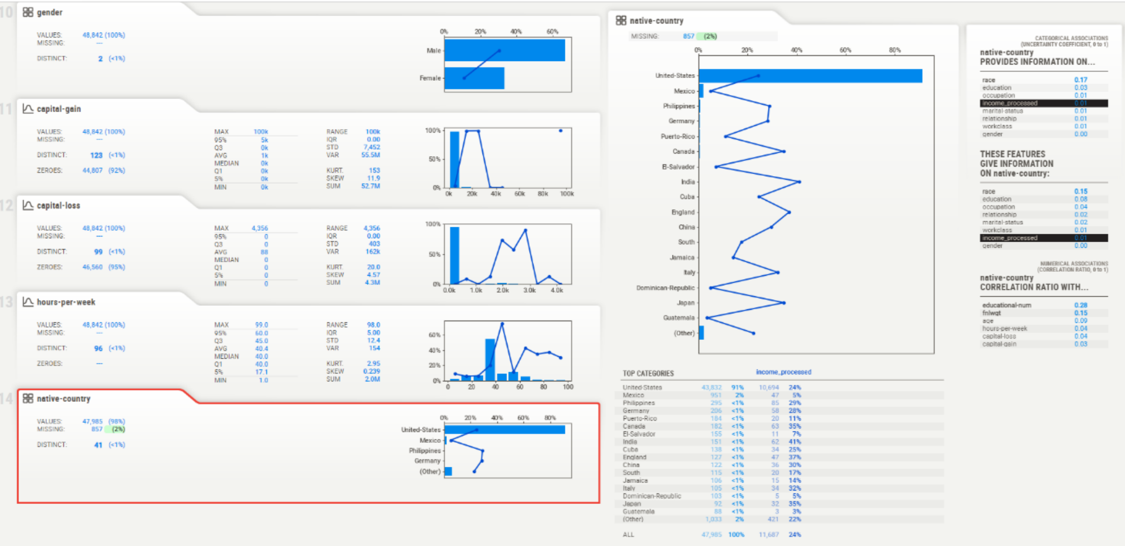
SMALLEST VALUES

1	27	<0.1%
2	53	0.1%
3	59	0.1%
4	84	0.2%
5	95	0.2%
6	92	0.2%
7	45	<0.1%
8	218	0.4%
9	27	<0.1%
10	425	0.9%
11	20	<0.1%
12	247	0.5%
13	28	<0.1%
14	55	0.1%
15	623	1.3%

LARGEST VALUES

99	137	0.3%
98	14	<0.1%
97	2	<0.1%
96	9	<0.1%
95	2	<0.1%
94	1	<0.1%
92	3	<0.1%
91	3	<0.1%
90	42	<0.1%
89	3	<0.1%
88	4	<0.1%
87	1	<0.1%
86	4	<0.1%
85	17	<0.1%
84	72	0.1%

Hours-per-week



```
10 gender
   VALUES:    48,842 (100%)
   MISSING:   ---
   DISTINCT:  2     (<1%)

11 capital-gain
   VALUES:    48,842 (100%)     MAX     100k     RANGE  100k
   MISSING:   ---               95%     5k       IQR    0.00
   DISTINCT:  123   (<1%)       Q3      0k       STD    7,452
   ZEROES:    44,807 (92%)      AVG     1k       VAR    55.5M
                                MEDIAN  0k
                                Q1      0k       KURT.  153
                                5%      0k       SKEW   11.9
                                MIN     0k       SUM    52.7M

12 capital-loss
   VALUES:    48,842 (100%)     MAX     4,356    RANGE  4,356
   MISSING:   ---               95%     0        IQR    0.00
   DISTINCT:  99    (<1%)       Q3      0        STD    403
   ZEROES:    46,560 (95%)      AVG     88       VAR    162k
                                MEDIAN  0
                                Q1      0        KURT.  20.0
                                5%      0        SKEW   4.57
                                MIN     0        SUM    4.3M

13 hours-per-week
   VALUES:    48,842 (100%)     MAX     99.0     RANGE  98.0
   MISSING:   ---               95%     60.0     IQR    5.00
   DISTINCT:  96    (<1%)       Q3      45.0     STD    12.4
   ZEROES:    ---               AVG     40.4     VAR    154
                                MEDIAN  40.0
                                Q1      40.0     KURT.  2.95
                                5%      17.1     SKEW   0.239
                                MIN     1.0      SUM    2.0M

14 native-country
   VALUES:    47,985 (98%)
   MISSING:   857    (2%)
   DISTINCT:  41     (<1%)
```

native-country

MISSING: 857 (2%)

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

native-country
PROVIDES INFORMATION ON...

race	0.17
education	0.03
occupation	0.01
income_processed	0.01
marital-status	0.01
relationship	0.01
workclass	0.01
gender	0.00

THESE FEATURES
GIVE INFORMATION
ON native-country:

race	0.15
education	0.08
occupation	0.04
relationship	0.02
marital-status	0.02
workclass	0.01
income_processed	0.01
gender	0.00

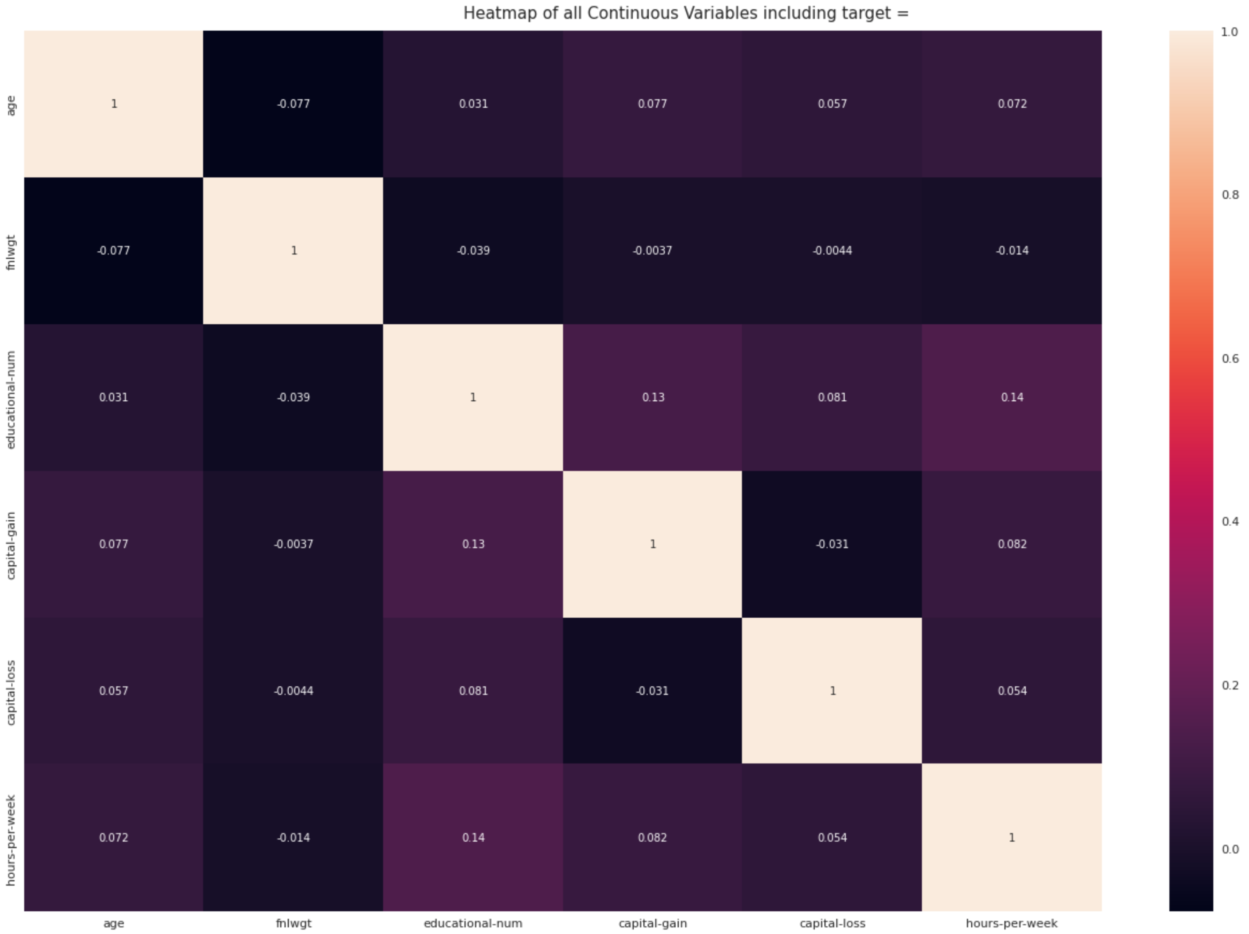
NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

native-country
CORRELATION RATIO WITH...

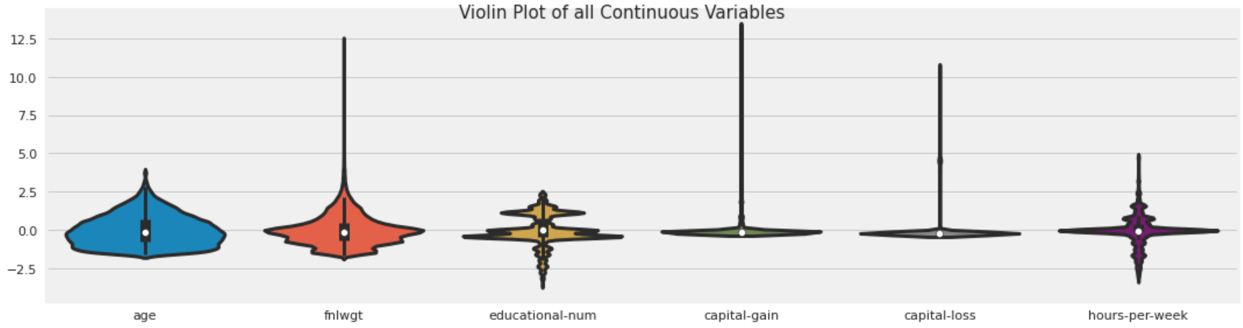
educational-num	0.28
fnlwgt	0.15
age	0.09
hours-per-week	0.04
capital-loss	0.04
capital-gain	0.03

TOP CATEGORIES			income_processed	
United-States	43,832	91%	10,694	24%
Mexico	951	2%	47	5%
Philippines	295	<1%	85	29%
Germany	206	<1%	58	28%
Puerto-Rico	184	<1%	20	11%
Canada	182	<1%	63	35%
El-Salvador	155	<1%	11	7%
India	151	<1%	62	41%
Cuba	138	<1%	34	25%
England	127	<1%	47	37%
China	122	<1%	36	30%
South	115	<1%	20	17%
Jamaica	106	<1%	15	14%
Italy	105	<1%	34	32%
Dominican-Republic	103	<1%	5	5%
Japan	92	<1%	32	35%
Guatemala	88	<1%	3	3%
(Other)	1,033	2%	421	22%
ALL	47,985	100%	11,687	24%

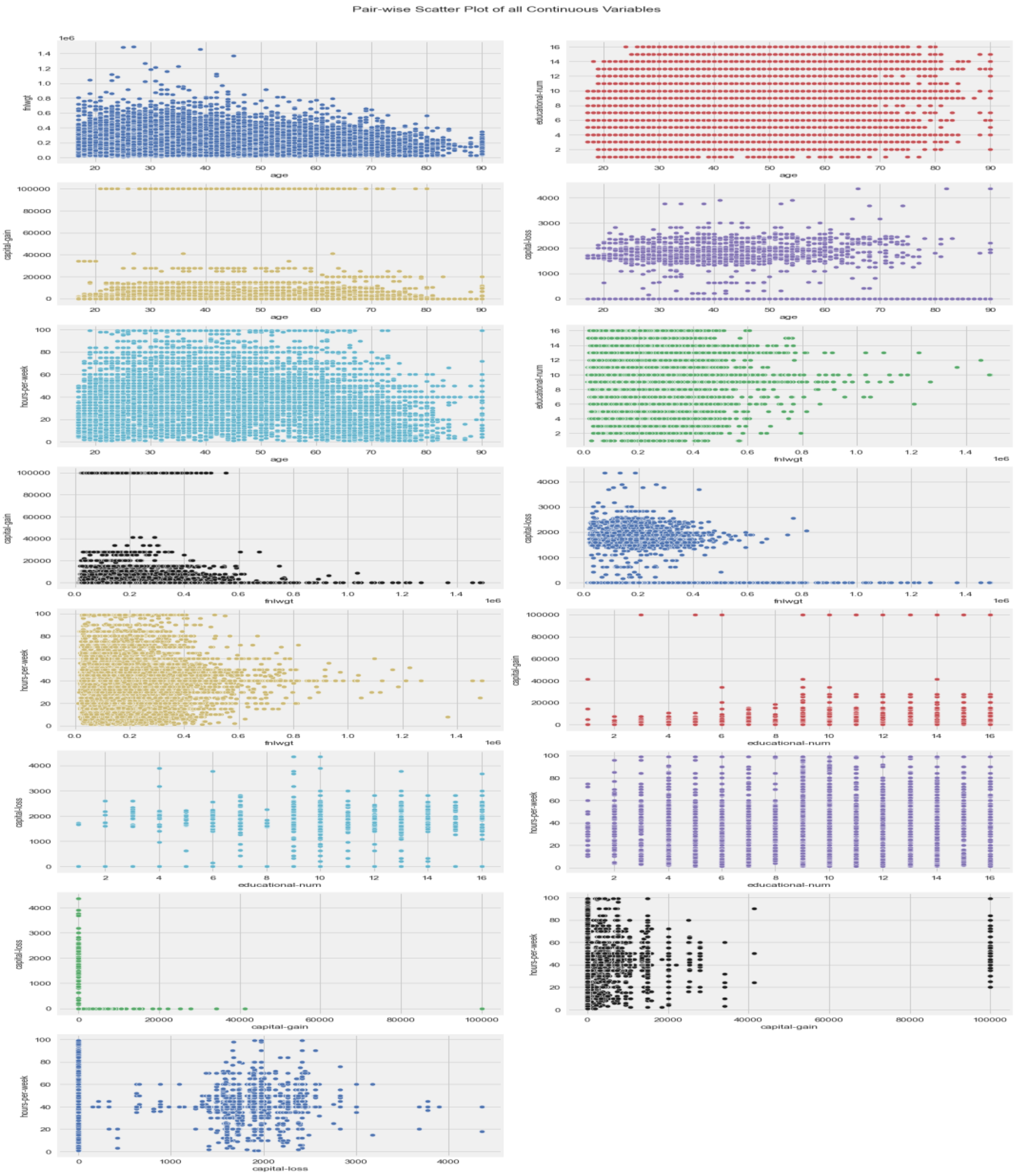
Native-country



Heatmap



Violin Plot



Scatter-plot

--Next Page--

Methodology

In this section detailed analytical methodology will be described such as basic statistical analysis and data pre-processing steps are explained in the following,

1. Establishing a baseline.
2. Basic experimentation.

1. Establishing a baseline: Basic descriptive statistical analysis is carried out to know the distribution of variables using histogram, boxplot for outlier identification and probability chart to know the skewness for the continuous variables.

2. Basic experimentation: Data pre-processing includes data cleaning such as outlier handling, features selection, data transformations and data balancing.

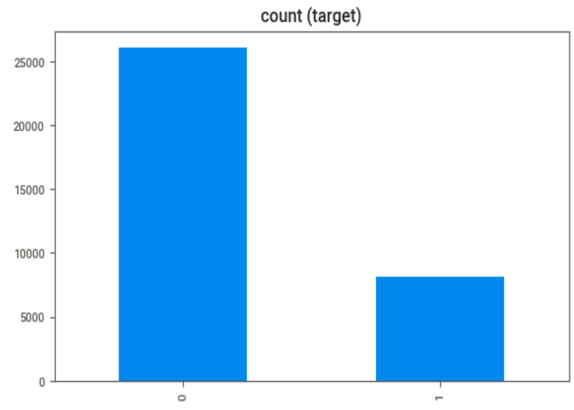
Data Cleaning : Removal of symbols or special characters from the data because mathematical operations can not be carried out in such data.

Missing value: Missing value is imputed with the most frequent value of the variable in the datasets, because missing values are identified in categorical variables imputed with the mode value.

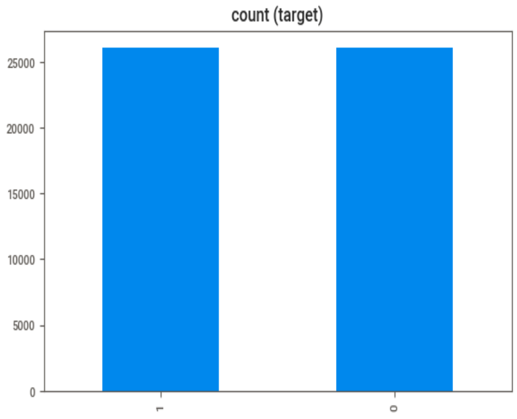
Feature selection : some of features are dropped from the analysis such as educational-num and fnlwgt because of low correlation with the target variables.

Data transformations : One hot encoding transformation is carried out for the categorical variables.

Handling imbalanced datasets: In the current study target variables(income) with class "<=50K" are more number of observations compared to class ">50K". Data balance is carried out using "SMOTE".(Fig-2: Data Imbalance and Fig -3: Data Balance with SMOTE).



(Fig-2)



(Fig-3)

Evaluation and Conclusions

In the study of census data of the US found that the mean age of population was 38.64 years with the standard deviation(SD) of 13.71 years, where as minimum and maximum age of respondents are 17 and 90 years respectively.The average working hours per week is 40 hrs with a SD of 12.39 hrs. 32% of respondents from the datasets have HS-grade followed by some-college is 22%. 46% of respondents are Married-civ-spouses who have marital status, followed by Never-married 33%. 86% of respondents are white race and 10% are black race. 67% of respondents are male and 33% respondents are female. 74% of respondents are belong private workclass and 8% of respondents are self-employed.

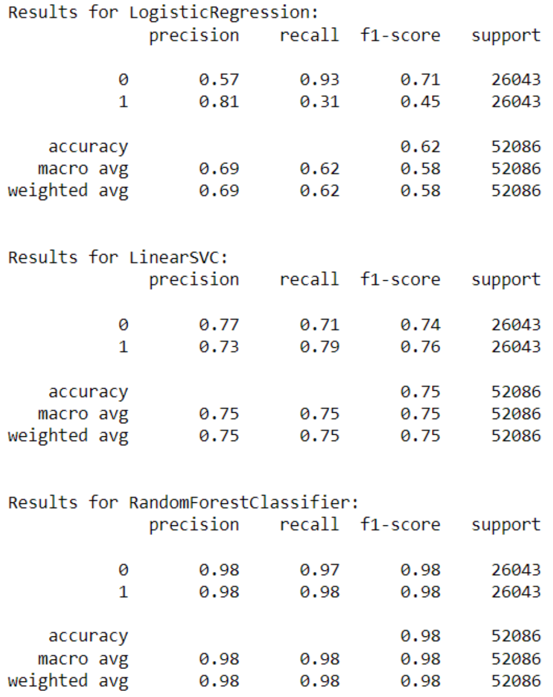
The classification models build for the study are **LogisticRegression, LinearSVC, RandomForestClassifier, GradientBoostingClassifier, BaggingClassifier, SVC, DecisionTreeClassifier and XGBClassifier**

And results of 10 fold cross validation are presented in table-1. The accuracy values for models in test datasets are found to be

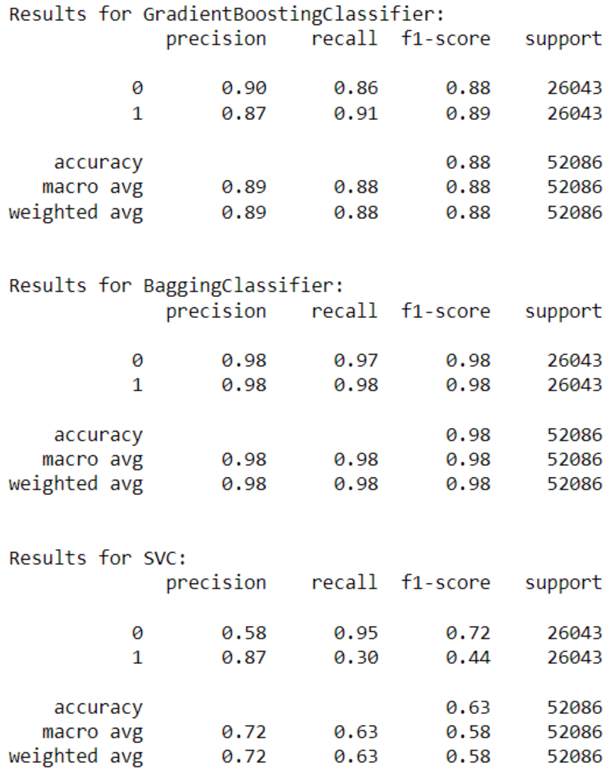
LogisticRegression	: 62%
LinearSVC	: 75%
RandomForestClassifier	: 98%
GradientBoostingClassifier	: 88%
BaggingClassifier	: 98%
SVC	: 63%
DecisionTreeClassifier	: 98%
XGBClassifier	: 92%

Respectively and results are saved in csv file.

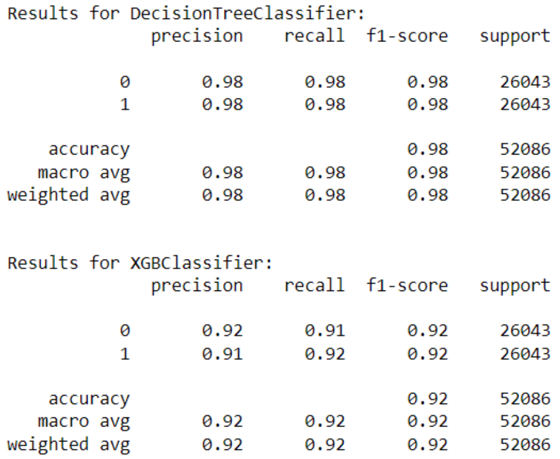
-- Next Page --



```
Results for LogisticRegression:
              precision    recall  f1-score   support

           0       0.57      0.93      0.71     26043
           1       0.81      0.31      0.45     26043

    accuracy                           0.62     52086
   macro avg       0.69      0.62      0.58     52086
weighted avg       0.69      0.62      0.58     52086


Results for LinearSVC:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74     26043
           1       0.73      0.79      0.76     26043

    accuracy                           0.75     52086
   macro avg       0.75      0.75      0.75     52086
weighted avg       0.75      0.75      0.75     52086


Results for RandomForestClassifier:
              precision    recall  f1-score   support

           0       0.98      0.97      0.98     26043
           1       0.98      0.98      0.98     26043

    accuracy                           0.98     52086
   macro avg       0.98      0.98      0.98     52086
weighted avg       0.98      0.98      0.98     52086
```



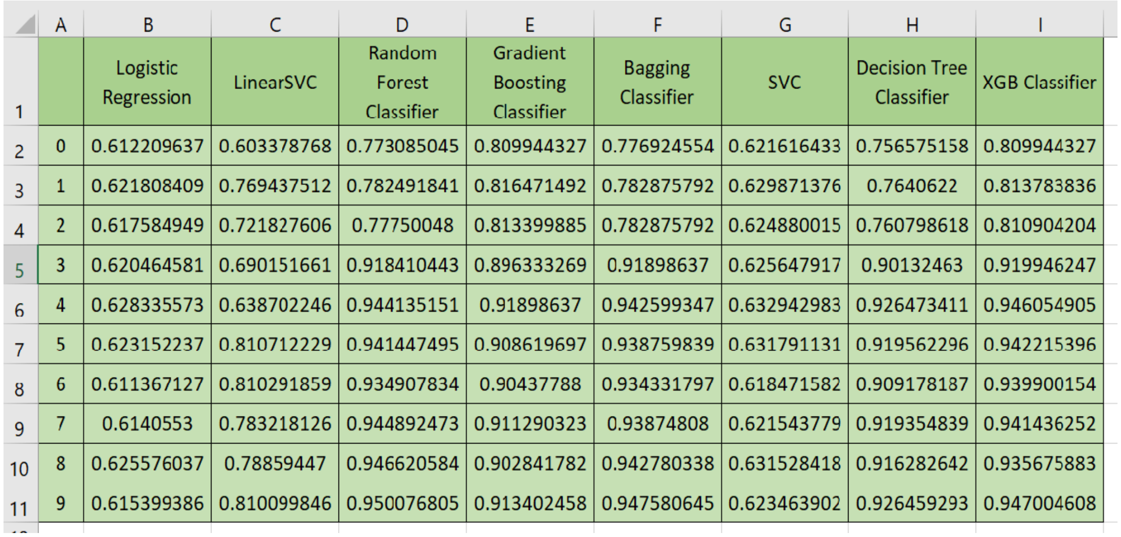
```
Results for GradientBoostingClassifier:
              precision    recall  f1-score   support

           0       0.90      0.86      0.88     26043
           1       0.87      0.91      0.89     26043

    accuracy                           0.88     52086
   macro avg       0.89      0.88      0.88     52086
weighted avg       0.89      0.88      0.88     52086


Results for BaggingClassifier:
              precision    recall  f1-score   support

           0       0.98      0.97      0.98     26043
           1       0.98      0.98      0.98     26043

    accuracy                           0.98     52086
   macro avg       0.98      0.98      0.98     52086
weighted avg       0.98      0.98      0.98     52086


Results for SVC:
              precision    recall  f1-score   support

           0       0.58      0.95      0.72     26043
           1       0.87      0.30      0.44     26043

    accuracy                           0.63     52086
   macro avg       0.72      0.63      0.58     52086
weighted avg       0.72      0.63      0.58     52086
```



```
Results for DecisionTreeClassifier:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     26043
           1       0.98      0.98      0.98     26043

    accuracy                           0.98     52086
   macro avg       0.98      0.98      0.98     52086
weighted avg       0.98      0.98      0.98     52086


Results for XGBClassifier:
              precision    recall  f1-score   support

           0       0.92      0.91      0.92     26043
           1       0.91      0.92      0.92     26043

    accuracy                           0.92     52086
   macro avg       0.92      0.92      0.92     52086
weighted avg       0.92      0.92      0.92     52086
```

-- Next Page --

After the evaluation of model accuracy RandomForestClassifier and BaggingClassifier is selected for model hyper parameter tuning using GridSearchCV with different combinations of parameters and the accuracy remains the same.



	A	B	C	D	E	F	G	H	I
1		Logistic Regression	LinearSVC	Random Forest Classifier	Gradient Boosting Classifier	Bagging Classifier	SVC	Decision Tree Classifier	XGB Classifier
2	0	0.612209637	0.603378768	0.773085045	0.809944327	0.776924554	0.621616433	0.756575158	0.809944327
3	1	0.621808409	0.769437512	0.782491841	0.816471492	0.782875792	0.629871376	0.7640622	0.813783836
4	2	0.617584949	0.721827606	0.77750048	0.813399885	0.782875792	0.624880015	0.760798618	0.810904204
5	3	0.620464581	0.690151661	0.918410443	0.896333269	0.91898637	0.625647917	0.90132463	0.919946247
6	4	0.628335573	0.638702246	0.944135151	0.91898637	0.942599347	0.632942983	0.926473411	0.946054905
7	5	0.623152237	0.810712229	0.941447495	0.908619697	0.938759839	0.631791131	0.919562296	0.942215396
8	6	0.611367127	0.810291859	0.934907834	0.90437788	0.934331797	0.618471582	0.909178187	0.939900154
9	7	0.6140553	0.783218126	0.944892473	0.911290323	0.93874808	0.621543779	0.919354839	0.941436252
10	8	0.625576037	0.78859447	0.946620584	0.902841782	0.942780338	0.631528418	0.916282642	0.935675883
11	9	0.615399386	0.810099846	0.950076805	0.913402458	0.947580645	0.623463902	0.926459293	0.947004608

Table: 10 Fold Cross Validation