How To Get Rich?

Aaron Portal, Victoria Routon, Samantha Wolownik

Motivation

Why did we choose this topic?

What did we want to find?

What did our work process look like?

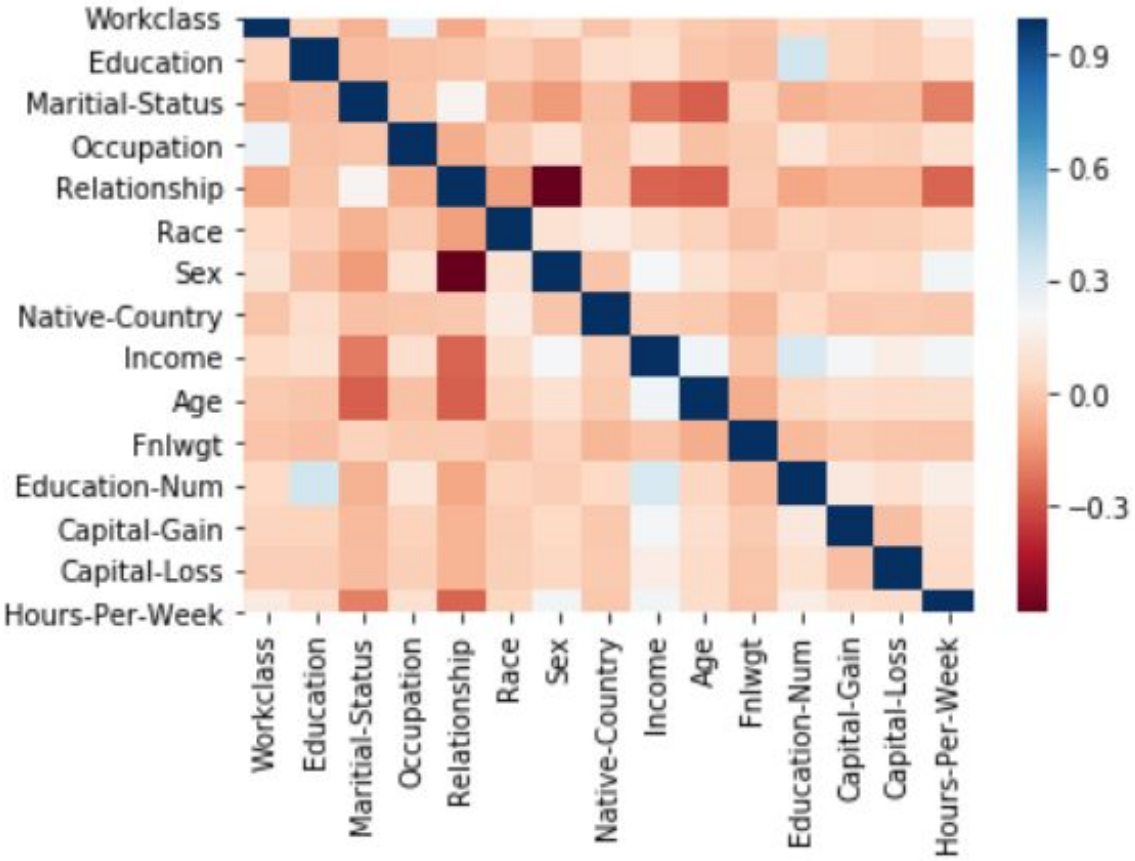
Data Preprocessing

39	State-gov	77516	Bachelors	13	Never-mar	Adm-cleric	Not-in-fam	White	Male	2174	0	40	United-Sta	<=50K
50	Self-emp-r	83311	Bachelors	13	Married-ci	Exec-mana	Husband	White	Male	0	0	13	United-Sta	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-c	Not-in-fam	White	Male	0	0	40	United-Sta	<=50K
53	Private	234721	11th	7	Married-ci	Handlers-c	Husband	Black	Male	0	0	40	United-Sta	<=50K
28	Private	338409	Bachelors	13	Married-ci	Prof-specia	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-ci	Exec-mana	Wife	White	Female	0	0	40	United-Sta	<=50K
49	Private	160187	9th	5	Married-sp	Other-serv	Not-in-fam	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-r	209642	HS-grad	9	Married-ci	Exec-mana	Husband	White	Male	0	0	45	United-Sta	>50K
31	Private	45781	Masters	14	Never-mar	Prof-specia	Not-in-fam	White	Female	14084	0	50	United-Sta	>50K
42	Private	159449	Bachelors	13	Married-ci	Exec-mana	Husband	White	Male	5178	0	40	United-Sta	>50K
37	Private	280464	Some-colle	10	Married-ci	Exec-mana	Husband	Black	Male	0	0	80	United-Sta	>50K
30	State-gov	141297	Bachelors	13	Married-ci	Prof-specia	Husband	Asian-Pac-	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-mar	Adm-cleric	Own-child	White	Female	0	0	30	United-Sta	<=50K
32	Private	205019	Assoc-acdr	12	Never-mar	Sales	Not-in-fam	Black	Male	0	0	50	United-Sta	<=50K
40	Private	121772	Assoc-voc	11	Married-ci	Craft-repai	Husband	Asian-Pac-	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-ci	Transport-	Husband	Amer-India	Male	0	0	45	Mexico	<=50K
25	Self-emp-r	176756	HS-grad	9	Never-mar	Farming-fi	Own-child	White	Male	0	0	35	United-Sta	<=50K
32	Private	186824	HS-grad	9	Never-mar	Machine-o	Unmarried	White	Male	0	0	40	United-Sta	<=50K
38	Private	28887	11th	7	Married-ci	Sales	Husband	White	Male	0	0	50	United-Sta	<=50K
43	Self-emp-r	292175	Masters	14	Divorced	Exec-mana	Unmarried	White	Female	0	0	45	United-Sta	>50K
40	Private	193524	Doctorate	16	Married-ci	Prof-specia	Husband	White	Male	0	0	60	United-Sta	>50K
54	Private	302146	HS-grad	9	Separated	Other-serv	Unmarried	Black	Female	0	0	20	United-Sta	<=50K
35	Federal-go	76845	9th	5	Married-ci	Farming-fi	Husband	Black	Male	0	0	40	United-Sta	<=50K
43	Private	117037	11th	7	Married-ci	Transport-	Husband	White	Male	0	2042	40	United-Sta	<=50K
59	Private	109015	HS-grad	9	Divorced	Tech-suppo	Unmarried	White	Female	0	0	40	United-Sta	<=50K
56	Local-gov	216851	Bachelors	13	Married-ci	Tech-suppo	Husband	White	Male	0	0	40	United-Sta	>50K
19	Private	168294	HS-grad	9	Never-mar	Craft-repai	Own-child	White	Male	0	0	40	United-Sta	<=50K

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
5	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
6	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
7	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
8	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
9	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
10	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
11	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
12	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
13	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
14	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
15	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
16	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
17	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K

- Added column names
- Checked for missing values
- Converted to numerical values
- Found the outliers using interquartile range

Multiple Regression


Workclass
Education
Maritial-Status
Occupation
Relationship
Race
Sex
Native-Country
Income
Age
Fnlwgt
Education-Num
Capital-Gain
Capital-Loss
Hours-Per-Week
0.9
0.6
0.3
0.0
−0.3

Intercept and coefficients

```
The intercept for our model is -0.5802
The Coefficient for Workclass is -0.0018
The Coefficient for Education is -0.0038
The Coefficient for Maritial-Status is -0.023
The Coefficient for Occupation is 0.0016
The Coefficient for Relationship is -0.017
The Coefficient for Race is 0.014
The Coefficient for Sex is 0.1
The Coefficient for Native-Country is 6.4e-05
The Coefficient for Age is 0.0046
The Coefficient for Education-Num is 0.047
The Coefficient for Capital-Gain is 9.3e-06
The Coefficient for Capital-Loss is 0.00011
The Coefficient for Hours-Per-Week is 0.0036
```

OLS Regression Results

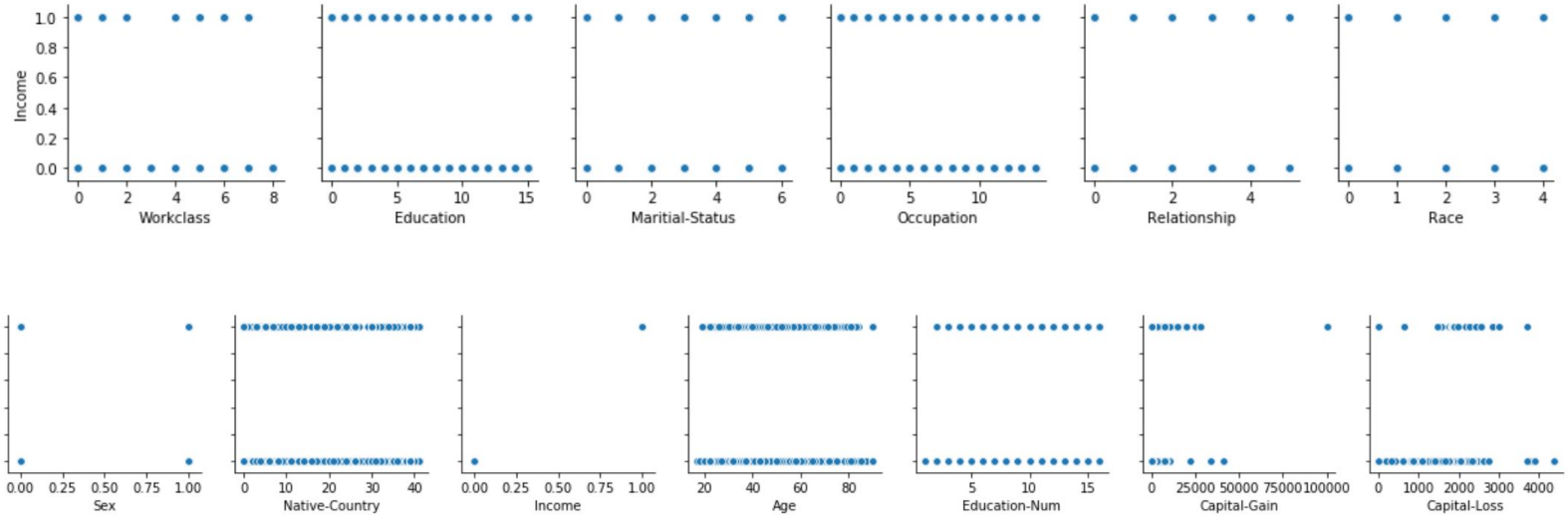
Dep. Variable:	Income	R-squared:	0.262
Model:	OLS	Adj. R-squared:	0.262
Method:	Least Squares	F-statistic:	825.5
Date:	Thu, 12 Nov 2020	Prob (F-statistic):	0.00
Time:	11:48:00	Log-Likelihood:	-13590.
No. Observations:	32560	AIC:	2.721e+04
Df Residuals:	32545	BIC:	2.734e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.5952	0.021	-28.649	0.000	-0.636	-0.554
Workclass	-0.0036	0.001	-2.433	0.015	-0.006	-0.001
Education	-0.0037	0.001	-6.485	0.000	-0.005	-0.003
Maritial-Status	-0.0239	0.001	-16.630	0.000	-0.027	-0.021
Occupation	0.0021	0.001	4.222	0.000	0.001	0.003
Relationship	-0.0153	0.002	-9.319	0.000	-0.019	-0.012
Race	0.0148	0.002	6.043	0.000	0.010	0.020
Sex	0.1035	0.005	19.160	0.000	0.093	0.114
Native-Country	-6.312e-06	0.000	-0.024	0.981	-0.001	0.001
Age	0.0047	0.000	29.403	0.000	0.004	0.005
Fnlwgt	6.706e-08	1.94e-08	3.455	0.001	2.9e-08	1.05e-07
Education-Num	0.0471	0.001	53.951	0.000	0.045	0.049
Capital-Gain	9.272e-06	2.8e-07	33.167	0.000	8.72e-06	9.82e-06
Capital-Loss	0.0001	5.09e-06	22.292	0.000	0.000	0.000
Hours-Per-Week	0.0036	0.000	20.343	0.000	0.003	0.004

Omnibus:	2971.366	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3720.360
Skew:	0.814	Prob(JB):	0.00
Kurtosis:	2.698	Cond. No.	2.22e+06

OLS Regression Results

Non - linear relationship



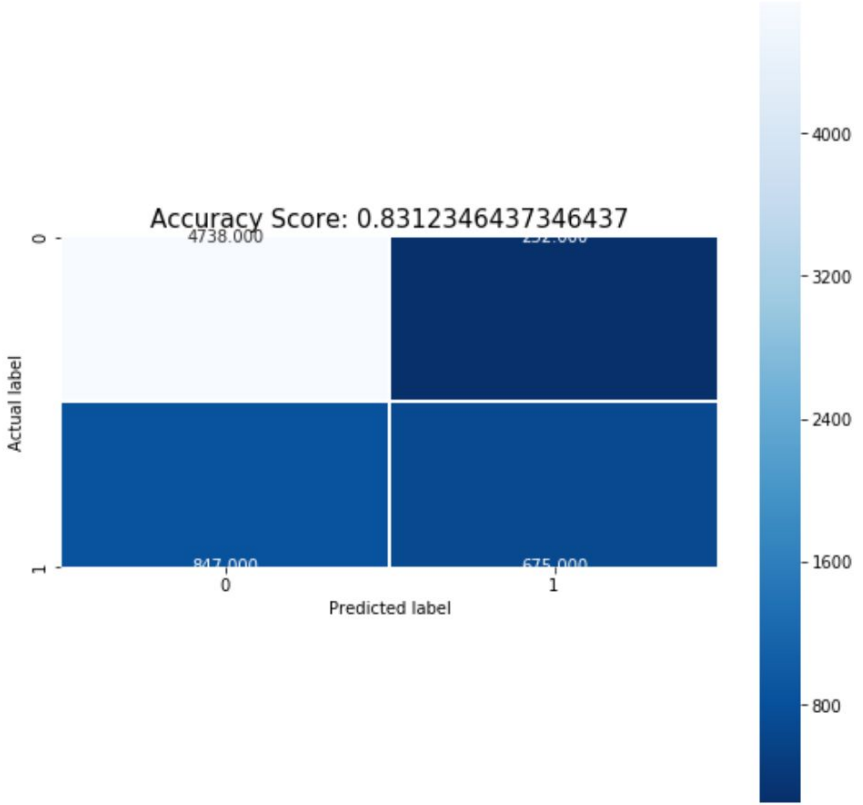
Logistic Regression

- **Confusion matrix**

```
array([[4738,  252],
       [ 847,  675]], dtype=int64)
```

- **Accuracy Score**

```
0.8312346437346437
```



Accuracy

	precision	recall	f1-score	support
0	0.85	0.95	0.90	4990
1	0.73	0.44	0.55	1522
accuracy			0.83	6512
macro avg	0.79	0.70	0.72	6512
weighted avg	0.82	0.83	0.82	6512

K-Means and Hierarchical Clustering

The Data So Far:

	Workclass	Education	Maritial-Status	Occupation	Relationship	Race	Sex	Native-Country	Income	Age	Education-Num	Capital-Gain	Capital-Loss	Hours-Per-Week
0	6	9	2	4	0	4	1	39	0	50	13	0	0	13
1	4	11	0	6	1	4	1	39	0	38	9	0	0	40
2	4	1	2	6	0	2	1	39	0	53	7	0	0	40
3	4	9	2	10	5	2	0	5	0	28	13	0	0	40
4	4	12	2	4	5	4	0	39	0	37	14	0	0	40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	4	7	2	13	5	4	0	39	0	27	12	0	0	38
32556	4	11	2	7	0	4	1	39	1	40	9	0	0	40
32557	4	11	6	1	4	4	0	39	0	58	9	0	0	40
32558	4	11	4	1	3	4	1	39	0	22	9	0	0	20
32559	5	11	2	4	5	4	0	39	1	52	9	15024	0	40

14 attributes

means:

Workclass	3.868796
Education	10.298249
Maritial-Status	2.611794
Occupation	6.572912
Relationship	1.446376
Race	3.665848
Sex	0.669195
Native-Country	36.718796
Income	0.240817
Age	38.581634
Education-Num	10.080590
Capital-Gain	1077.615172
Capital-Loss	87.306511
Hours-Per-Week	40.437469
dtype: float64	

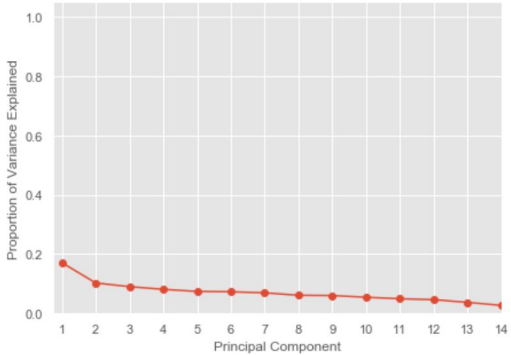
variances:

Workclass	2.119583e+00
Education	1.497935e+01
Maritial-Status	2.268714e+00
Occupation	1.788283e+01
Relationship	2.581786e+00
Race	7.204897e-01
Sex	2.213797e-01
Native-Country	6.121328e+01
Income	1.828298e-01
Age	1.860671e+02
Education-Num	6.618831e+00
Capital-Gain	5.454418e+07
Capital-Loss	1.623817e+05
Hours-Per-Week	1.524637e+02
dtype: float64	

PCA

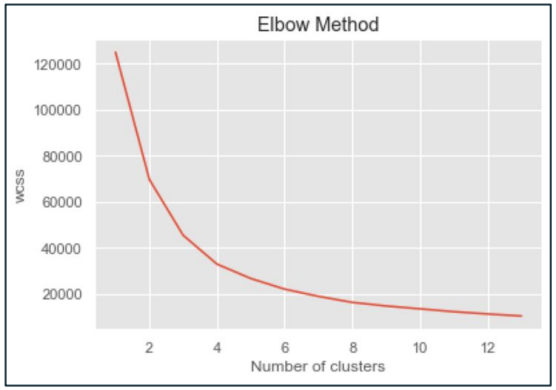
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
0	0.462956	-0.339835	0.077813	-0.679753	0.471427	0.156030	-0.297900	-1.923694	-0.753053	1.337059	0.097912	-1.459951	-1.504416	0.232419
1	0.565247	-0.736502	0.323068	-0.670153	0.424562	0.123653	-1.005439	0.300293	-0.045800	0.237645	1.262745	-0.131694	-0.092552	0.227132
2	-0.063610	-2.785778	0.316906	1.207720	0.472513	0.176105	-0.113421	-0.527691	1.866890	0.133657	-0.272449	-0.329151	-0.766067	-0.029340
3	-1.947225	1.136443	-1.264851	4.125724	-0.247434	0.526482	-1.078404	0.801313	-1.507991	-1.019436	0.529795	-0.331118	-0.711749	0.602179
4	-1.312459	2.190579	0.362760	-0.280111	0.860473	0.550972	-0.633968	0.797110	-0.368523	-0.208467	-0.232667	-0.294319	-0.822201	0.896415
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	-1.623407	1.204123	-1.234905	-0.104398	1.480340	0.922646	0.183357	0.509974	0.009671	-1.456025	0.518314	-0.101822	-0.569992	0.619677
32556	1.501742	-0.390396	0.261677	-0.356967	0.024659	-0.188867	0.189864	-0.298795	-0.030828	-0.300455	-0.026956	-0.812252	1.207602	-0.441513
32557	-2.036807	0.677838	0.836284	-0.243847	0.498208	0.114641	0.385641	-0.064853	-0.349757	0.848114	-2.529436	0.888944	0.240170	0.589741
32558	-1.814360	-0.031788	0.169398	-1.014923	-0.661759	-0.748845	0.550092	-0.255220	-0.461419	1.009650	0.364968	-1.121754	0.266247	0.814218
32559	-0.117196	1.596024	0.893891	0.814259	2.703352	0.178646	0.914891	0.853360	-0.222851	0.703187	-0.648029	-0.429524	1.362349	0.551483

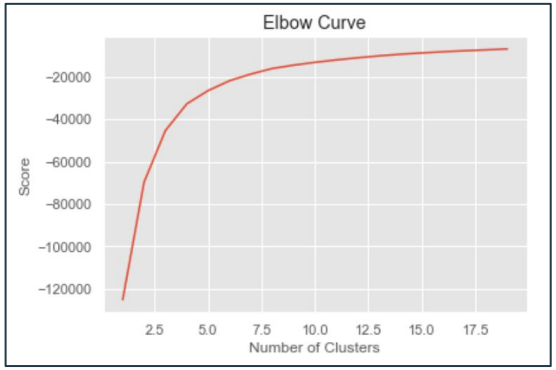
```
Variances = [2.39926508, 1.44148584, 1.26526999, 1.13964582, 1.04076953, 1.02380759, 0.9735611 , 0.85948076, 0.84646794, 0.76488965, 0.69279312, 0.65122761, 0.52203703, 0.37972894]
```



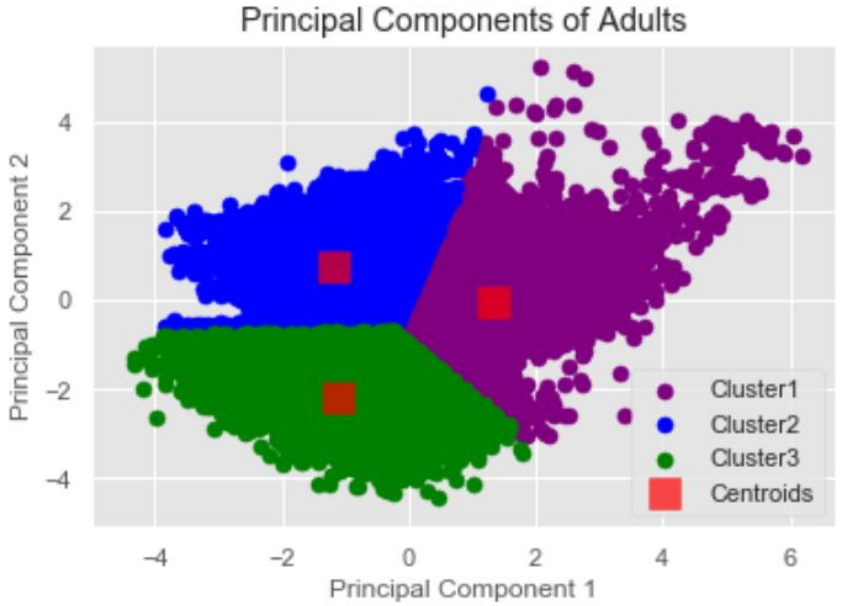
	PC1	PC2
0	0.462956	-0.339835
1	0.565247	-0.736502
2	-0.063610	-2.785778
3	-1.947225	1.136443
4	-1.312459	2.190579
...	...	...
32555	-1.623407	1.204123
32556	1.501742	-0.390396
32557	-2.036807	0.677838
32558	-1.814360	-0.031788
32559	-0.117196	1.596024

K-Means

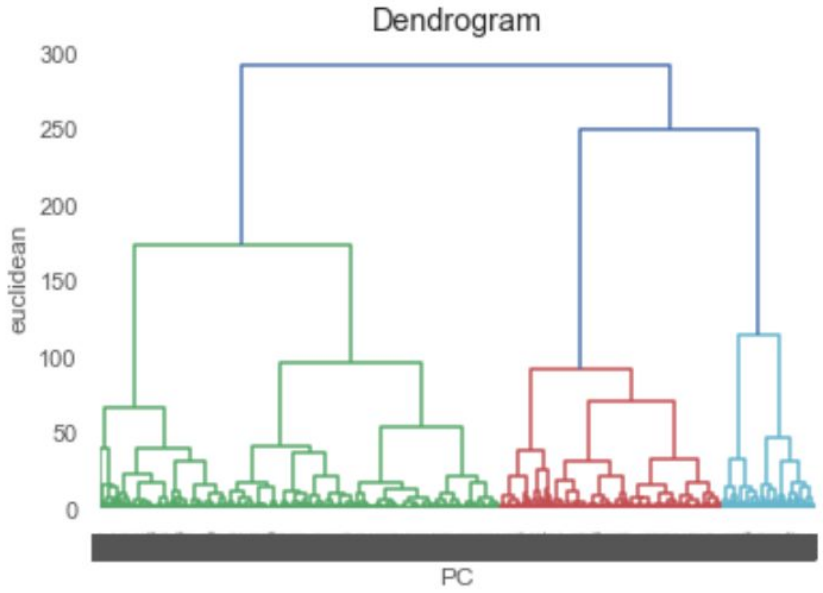


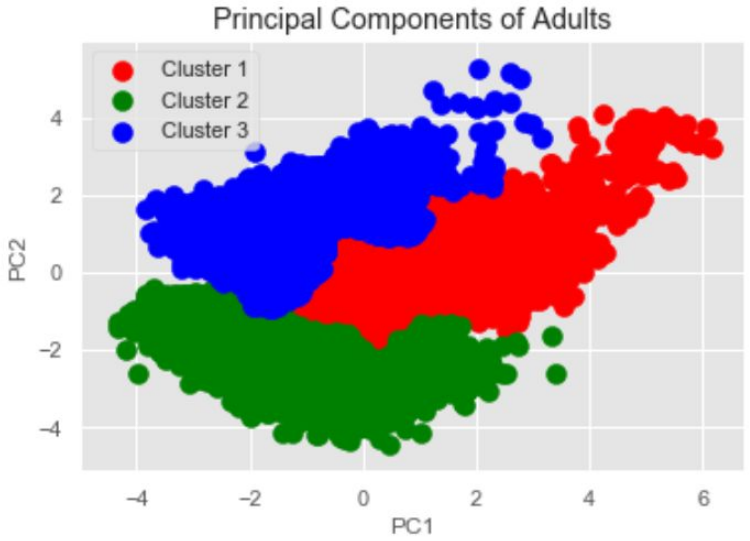



```
For n_clusters=2, The Silhouette Coefficient is 0.42330051043783395
For n_clusters=3, The Silhouette Coefficient is 0.4753867510721339
For n_clusters=4, The Silhouette Coefficient is 0.42841889785735293
For n_clusters=5, The Silhouette Coefficient is 0.4435039368610043
For n_clusters=6, The Silhouette Coefficient is 0.39624900373567445
For n_clusters=7, The Silhouette Coefficient is 0.4030649801247899
For n_clusters=8, The Silhouette Coefficient is 0.38805541881738415
For n_clusters=9, The Silhouette Coefficient is 0.3707888943766242
For n_clusters=10, The Silhouette Coefficient is 0.38470758461805193
```

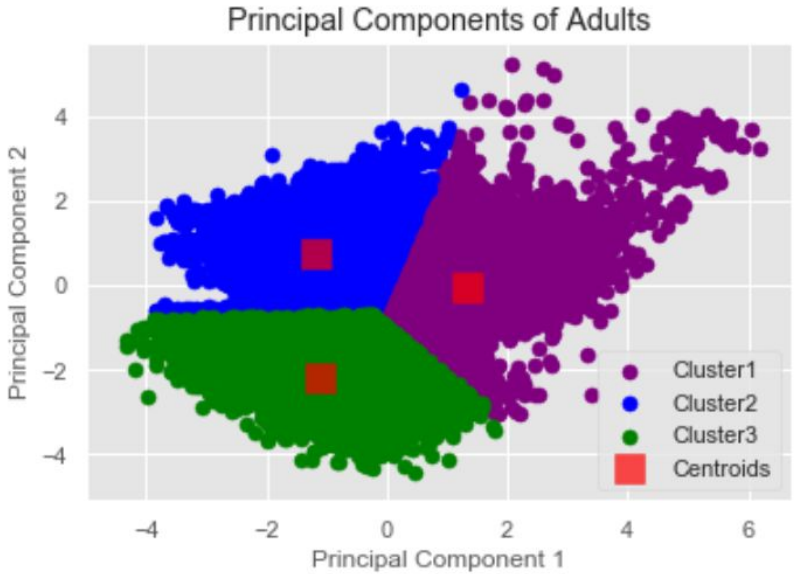


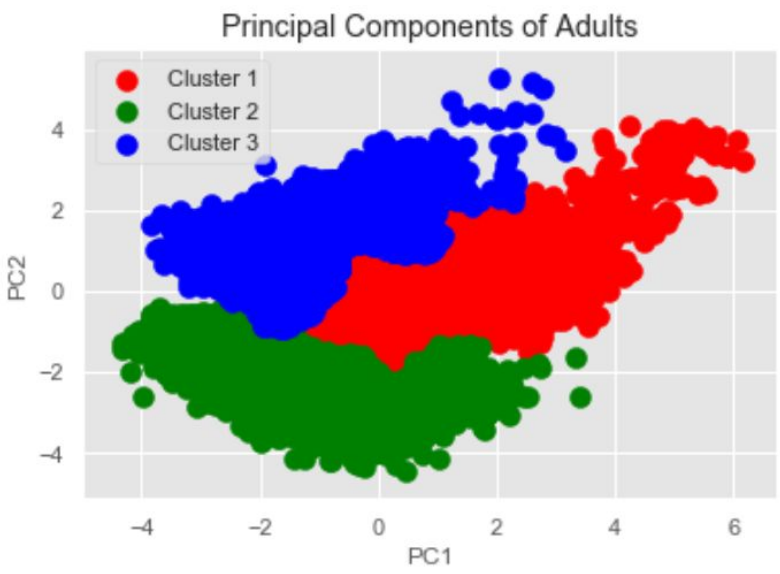
Hierarchical Clustering:





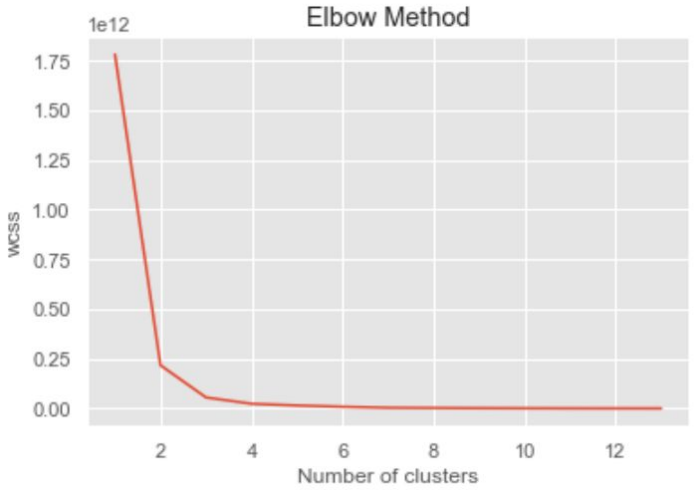
K-Means and Hierarchical:





Interpretation issues due to unnecessary PCA

	Workclass	Education	Maritial-Status	Occupation	Relationship	Race	Sex	Native-Country	Income	Age	Education-Num	Capital-Gain	Capital-Loss	Hours-Per-Week
0	6	9	2	4	0	4	1	39	0	50	13	0	0	13
1	4	11	0	6	1	4	1	39	0	38	9	0	0	40
2	4	1	2	6	0	2	1	39	0	53	7	0	0	40
3	4	9	2	10	5	2	0	5	0	28	13	0	0	40
4	4	12	2	4	5	4	0	39	0	37	14	0	0	40




```
For n_clusters=2, The Silhouette Coefficient is 0.986216325344319
For n_clusters=3, The Silhouette Coefficient is 0.9376088842254818
For n_clusters=4, The Silhouette Coefficient is 0.930204495310775
For n_clusters=5, The Silhouette Coefficient is 0.9319795245204494
For n_clusters=6, The Silhouette Coefficient is 0.9259120706341673
For n_clusters=7, The Silhouette Coefficient is 0.9524809678699202
For n_clusters=8, The Silhouette Coefficient is 0.9533397493211551
For n_clusters=9, The Silhouette Coefficient is 0.9548775273971718
For n_clusters=10, The Silhouette Coefficient is 0.9564792909805154
For n_clusters=11, The Silhouette Coefficient is 0.9567385241351869
For n_clusters=12, The Silhouette Coefficient is 0.9554632944103004
For n_clusters=13, The Silhouette Coefficient is 0.9529239180835452
```

	Workclass	Education	Maritial-Status	Occupation	Relationship	Race	Sex	Native-Country	Income	Age	Education-Num	Capital-Gain	Capital-Loss	Hours-Per-Week	Cluster
0	6	9	2	4	0	4	1	39	0	50	13	0	0	13	0
1	4	11	0	6	1	4	1	39	0	38	9	0	0	40	0
2	4	1	2	6	0	2	1	39	0	53	7	0	0	40	0
3	4	9	2	10	5	2	0	5	0	28	13	0	0	40	0
4	4	12	2	4	5	4	0	39	0	37	14	0	0	40	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	4	7	2	13	5	4	0	39	0	27	12	0	0	38	0
32556	4	11	2	7	0	4	1	39	1	40	9	0	0	40	0
32557	4	11	6	1	4	4	0	39	0	58	9	0	0	40	0
32558	4	11	4	1	3	4	1	39	0	22	9	0	0	20	0
32559	5	11	2	4	5	4	0	39	1	52	9	15024	0	40	0

Cluster1:

	Workclass	Education	Maritial-Status	Occupation	Relationship	Race	Sex	Native-Country	Income	Age	Education-Num	Capital-Gain	Capital-Loss	Hours-Per-Week
6	6	9	2	4	0	4	1	39	0	50	13	0	0	13
4	4	11	0	6	1	4	1	39	0	38	9	0	0	40
4	4	1	2	6	0	2	1	39	0	53	7	0	0	40
4	4	9	2	10	5	2	0	5	0	28	13	0	0	40
4	4	12	2	4	5	4	0	39	0	37	14	0	0	40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	4	7	2	13	5	4	0	39	0	27	12	0	0	38
4	4	11	2	7	0	4	1	39	1	40	9	0	0	40
4	4	11	6	1	4	4	0	39	0	58	9	0	0	40
4	4	11	4	1	3	4	1	39	0	22	9	0	0	20
5	5	11	2	4	5	4	0	39	1	52	9	15024	0	40

Cluster2:

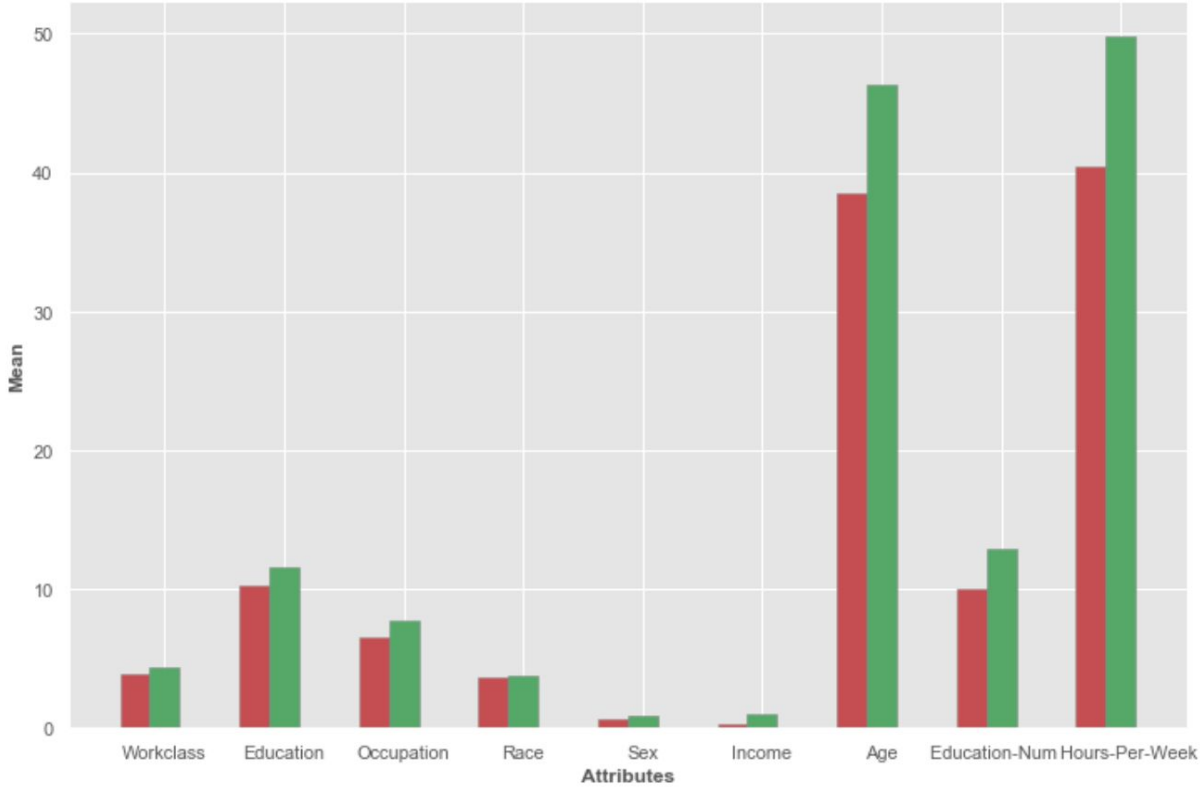
	Workclass	Education	Maritial-Status	Occupation	Relationship	Race	Sex	Native-Country	Income	Age	Education-Num	Capital-Gain	Capital-Loss	Hours-Per-Week
5	5	14	2	10	0	4	1	39	1	54	15	99999	0	60
4	4	11	2	4	0	1	1	24	1	52	9	99999	0	40
5	5	11	2	12	0	4	1	39	1	53	9	99999	0	40
4	4	9	2	4	0	4	1	39	1	52	13	99999	0	50
4	4	14	2	10	0	4	1	39	1	46	15	99999	0	60
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	4	12	2	4	0	4	1	39	1	47	14	99999	0	55
5	5	14	2	4	0	4	1	39	1	43	15	99999	0	40
4	4	9	2	4	0	4	1	0	1	66	13	99999	0	55
4	4	14	2	4	0	4	1	39	1	47	15	99999	0	40
2	2	11	2	3	0	4	1	39	1	57	9	99999	0	40

Cluster1:

Workclass	3.866115
Education	10.292213
Occupation	6.567050
Race	3.665628
Sex	0.668251
Income	0.237091
Age	38.543471
Education-Num	10.066665
Hours-Per-Week	40.391531
dtype: float64	

Cluster2:

Workclass	4.415094
Education	11.528302
Occupation	7.767296
Race	3.710692
Sex	0.861635
Income	1.000000
Age	46.358491
Education-Num	12.918239
Hours-Per-Week	49.798742
dtype: float64	



Naive Bayes and Random Forest

Naive Bayes Accuracy:

Classification Report:

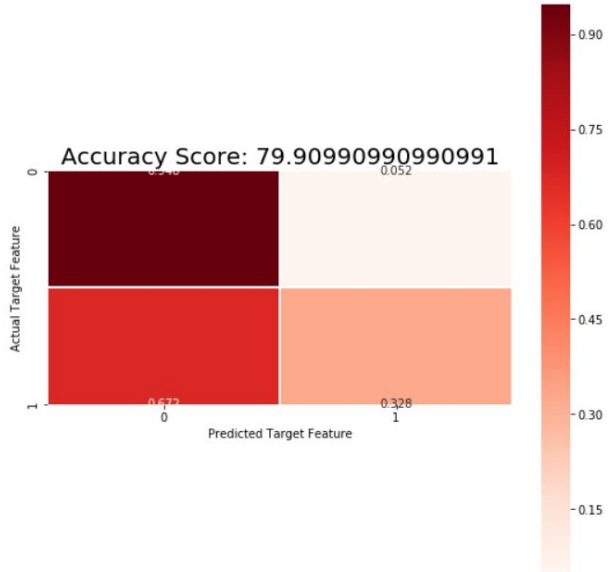
	precision	recall	f1-score	support
0	0.82	0.95	0.88	6194
1	0.67	0.34	0.45	1946
accuracy			0.80	8140
macro avg	0.75	0.64	0.66	8140
weighted avg	0.78	0.80	0.78	8140

Accuracy of Naive Bayes predictions: 80.54054054054053

Accuracy: 0.801

Standard Deviation: 0.0056387775176976 01

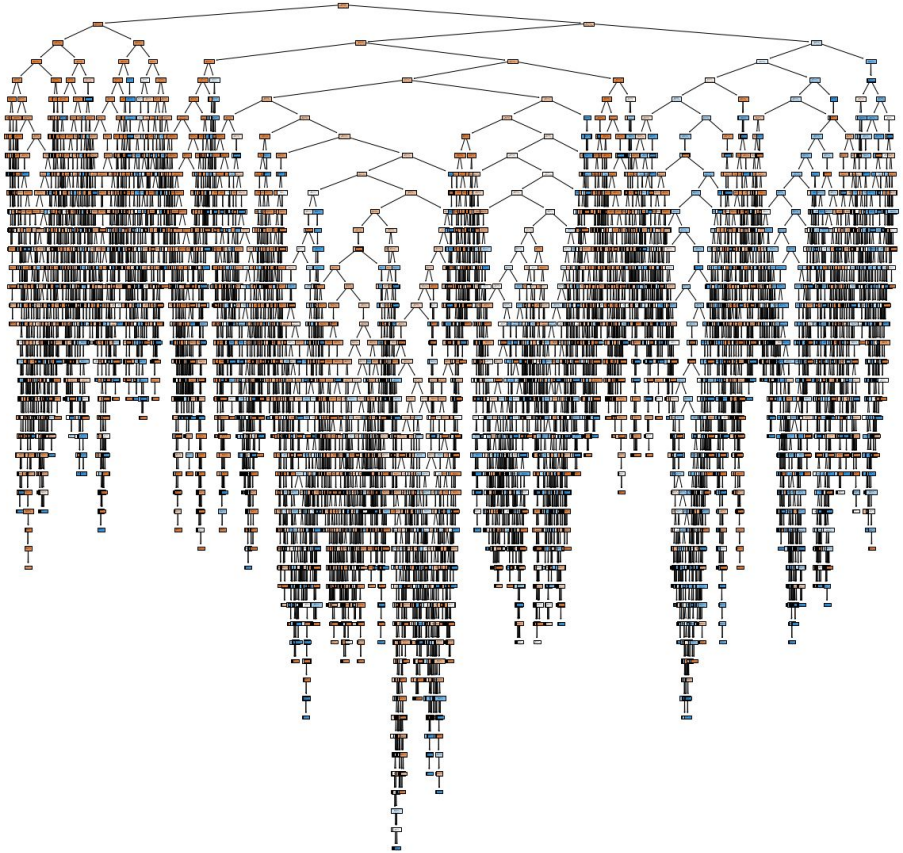
Confusion Matrix:




```
Confusion Matrix Normalized:
[[0.88188296 0.11811704]
 [0.88246269 0.11753731]]

Confusion Matrix Not Normalized:
[[16336  2188]      TrueNeg     FalsePos
 [ 5203   693]]     FalseNeg    TruePos
```

Random Forest Decision Trees:



Random Forest Accuracy

```
Accuracy: 0.801
Standard Deviation:  0.005638777517697601
```

```
Mean Absolute Error:  0.2 degrees
```

Conclusion:

- Improve algorithms by comparing with results from other algorithms
- Remove unnecessary features
  - Fwnlgnt
  - Native_Country
- Best Algorithms
  - Naive Bayes
  - Logistic Regression
- Worst Algorithms
  - Multivariable Linear Regression