ADULT INCOME

Group 2 – Cyber Team: Vincenzo Aiello, Gianluca Guidi, Marzia Longo, Elisa Mercanti

A. DATA UNDERSTANDING

Data semantics

This dataset was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

The dataset contains **48.842 records and 15 attributes**, 9 of which are categorical, 5 are continuous plus the categorical target variable.

Categorical Attributes:

- **workclass**: individual work category. There are 8 unique values (plus missing values). The categories are: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked;
- **education**: individual's highest education degree. This variable has 16 categories: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. There are no missing values;
- **educational-num**: an ordinal feature corresponding to the education nominal attribute, ranging from 1 (preschool) to 16 (doctorate). As for education, there are no missing values;
- **marital-status**: individual marital status. There are 7 categories: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. This attribute doesn't have any missing values;
- **occupation**: individual's occupation. This attribute has 14 unique values: 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. There are missing values;
- **relationship**: individual's relation in a family. There are 6 values: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. There are no missing values;
- **race**: race of individual: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. There are no missing values;
- **gender:** individual's gender (Female, Male);
- **native-country**: individual's native country. There are 41 categories: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US (Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands. This attribute presents missing values.

Continuous Attributes:

- **age**: age of an individual. No missing values;
- **fnlwgt**: final weight. It represents the number of units represented by each record in the target population;
- **capital-gain**: profits besides the job salary (financial investments etc). No missing values;
- **capital-loss**: losses besides the job salary (financial investments etc). No missing values;

- **hours-per-week**: individual's working hours per week. No missing values.

Target attribute:

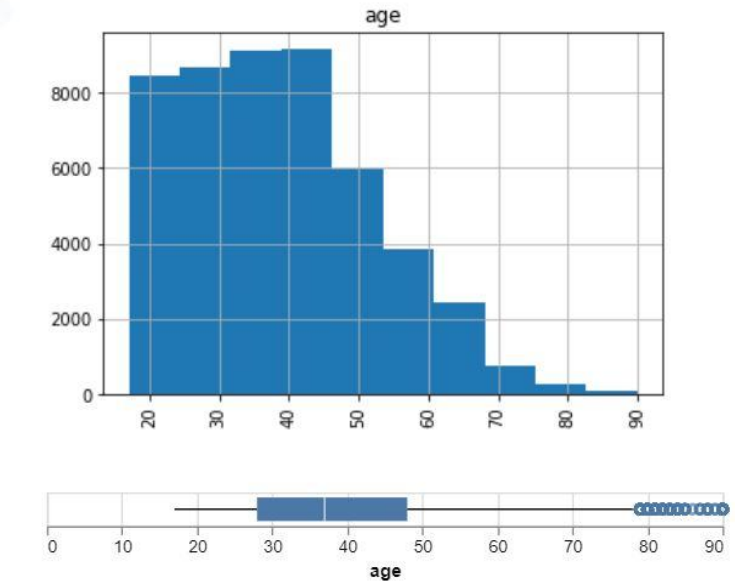
- **Income**: individual's annual income. Dichotomized variable with values: <=50K and >50K.

Distribution of the variables and statistics

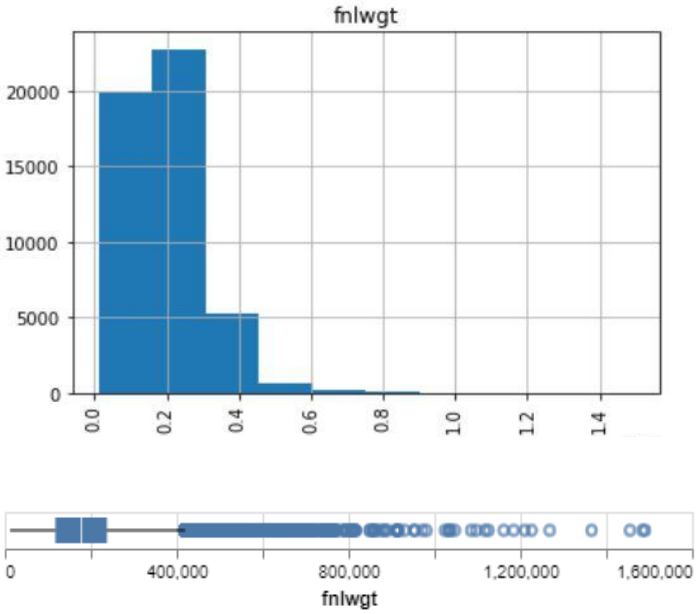
Univariate analysis - numerical attributes

Here the main statistics for the numerical variables:

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
count	48842.000000	4.884200e+04	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1.896641e+05	10.078089	1079.067626	87.502314	40.422382
std	13.710510	1.056040e+05	2.570973	7452.019058	403.004552	12.391444
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175505e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781445e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376420e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000



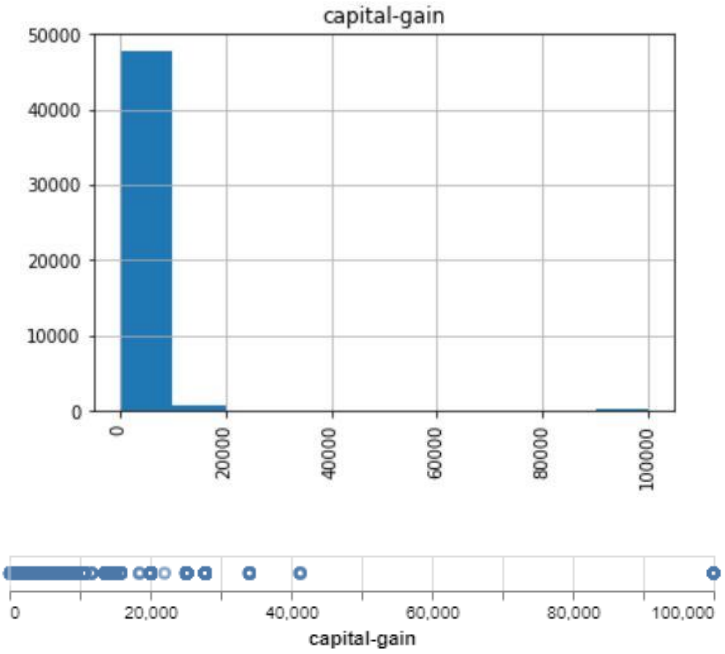
For **Age**:
The value of Age attribute varies from 17 to 90, with an average of 38 and standard deviation of 13.71. 3rd quartile is 48 which indicates that in 75% of the observations the value of age is less than 48. The distribution is therefore right skewed.



For **fnlwgt**:
This is the sampling weight corresponding to the observations. The variable is rightly skewed since there is a very large distance between median and maximum value as compared to minimum and median value.

The histogram shows the fnlwgt's frequency distribution. The distribution is right skewed.

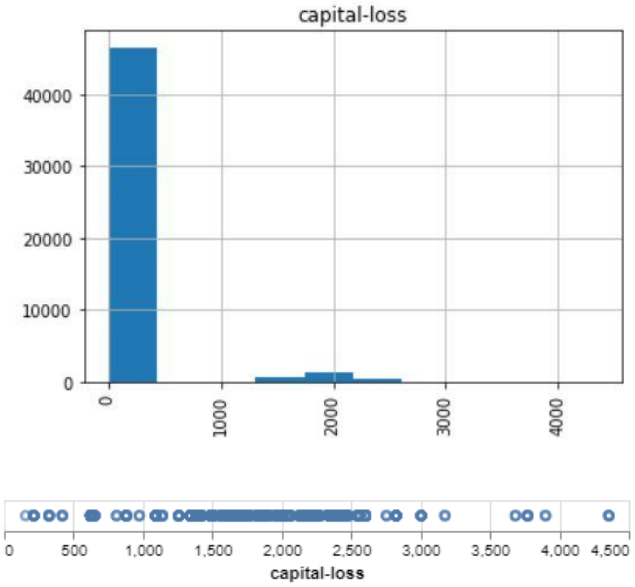
The boxplot shows the fnlwgt's distribution. The values are concentrated in range (50-300).



For **capital-gain**:
It is clearly visible that 75% of observations are having capital gain zero. The distribution is highly right skewed with mean 1079.06. The variable has a large standard deviation (7452.01) which means that either a person has no gain or has gain of very large amounts (10k or 99k).

The histogram shows the capital-gain's frequency distribution. The distribution is right skewed.

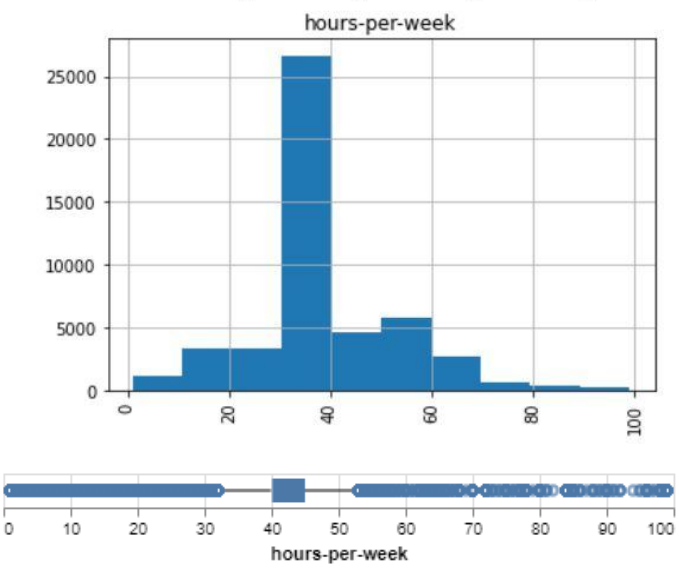
The boxplot shows the capital-gain's distribution. The values are concentrated in range (0-10.000).



For **capital-loss**:
This attribute is similar to the capital-gain, i.e. most of the values are centered on 0. The distribution is rightly skewed with mean 87.

The histogram shows the capital-loss's frequency distribution. The distribution is right skewed.

The boxplot shows the capital-loss's distribution. The values are concentrated in the range (0-500), but a large number of values is present in central area.



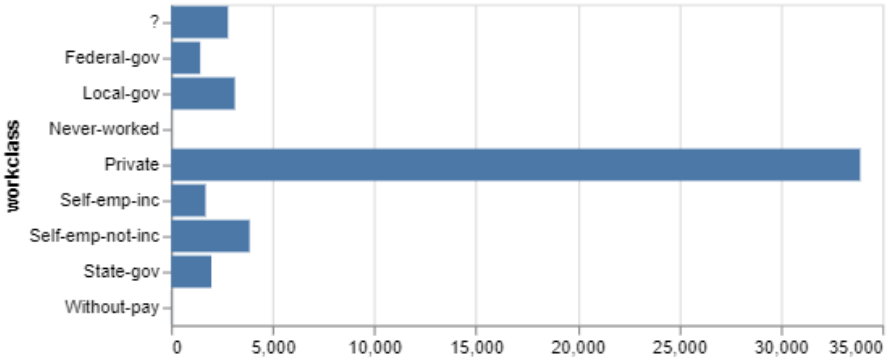
For **hours-per-week**:
The variable ranges between 1 and 99. 75% of the examined people spend 45 or less working hours per week. The IQR is very small (i.e. [40-45]) and it's because the majority of people work 40 hours a week.

The histogram shows the hours-per-week frequency distribution. The distribution is approximately symmetric.

The boxplot shows the hours-per-week's distribution. The 50% of the values are concentrated in the range (40-45).

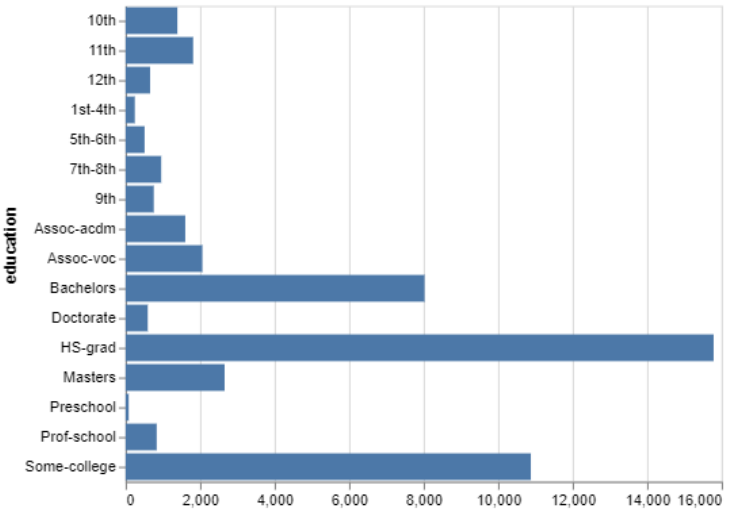
Univariate analysis - categorical attributes

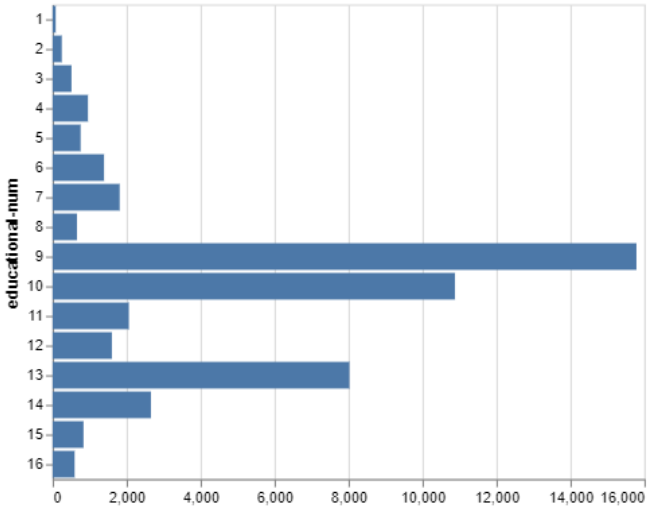
For **workclass**:
We notice that 'Private' is the most frequent value. We also observe the 6% of the values are missing here.



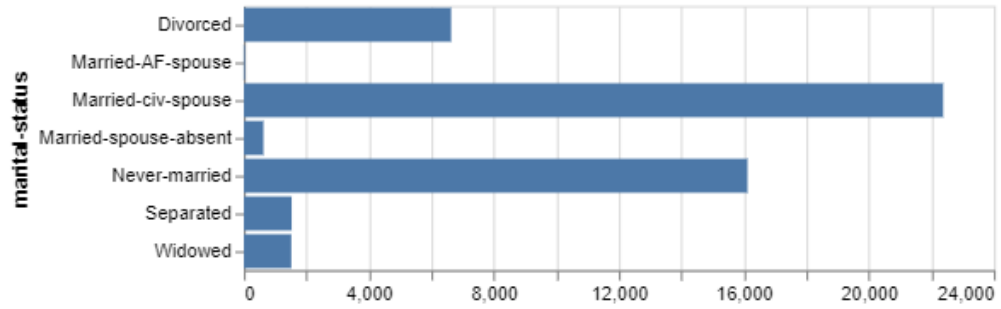
For **education and education-num**:
We report the table showing the correspondences of values between education and educational-num. The histogram at the bottom, is more intuitive to read since the numbers are ordinal, and it emerges that the majority of the people in this dataset have a high school diploma or some college degree and a consistent number of people have a bachelor degree.

education	educational-num
Doctorate	16
Prof-school	15
Masters	14
Bachelors	13
Assoc-acdm	12
Assoc-voc	11
Some-college	10
HS-grad	9
12th	8
11th	7
10th	6
9th	5
7th-8th	4
5th-6th	3
1st-4th	2
Preschool	1

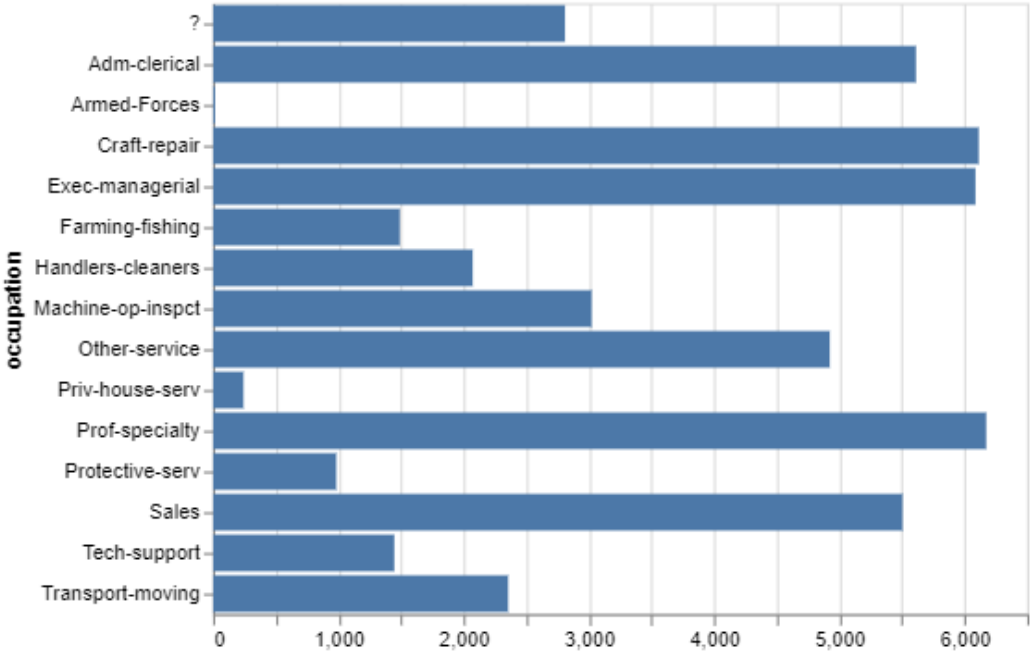




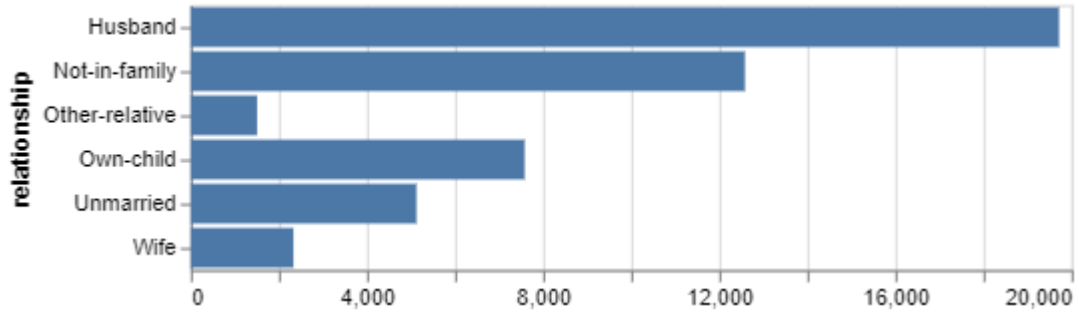
For **marital status**:
We find that on average the population is equally splitted between married and unmarried/divorced people.



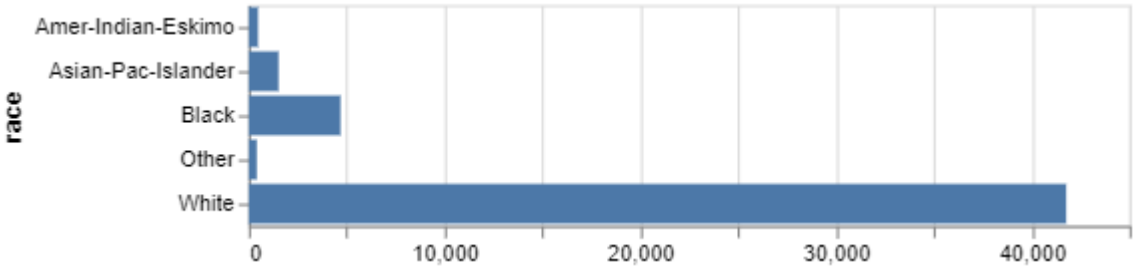
For **occupation**:
Here, the occupation class with highest frequency are Prof-specialty, Craft-repair and Exec-managerial.



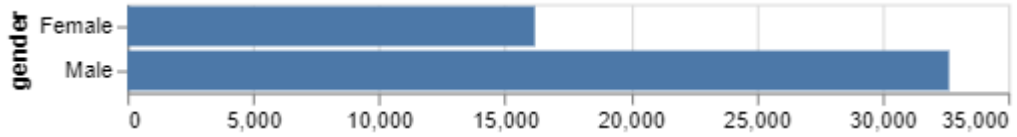
For **relationship**:
The most frequent value is ‘Husband’, this may be due to the fact that the number of males in the dataset doubles the number of females.



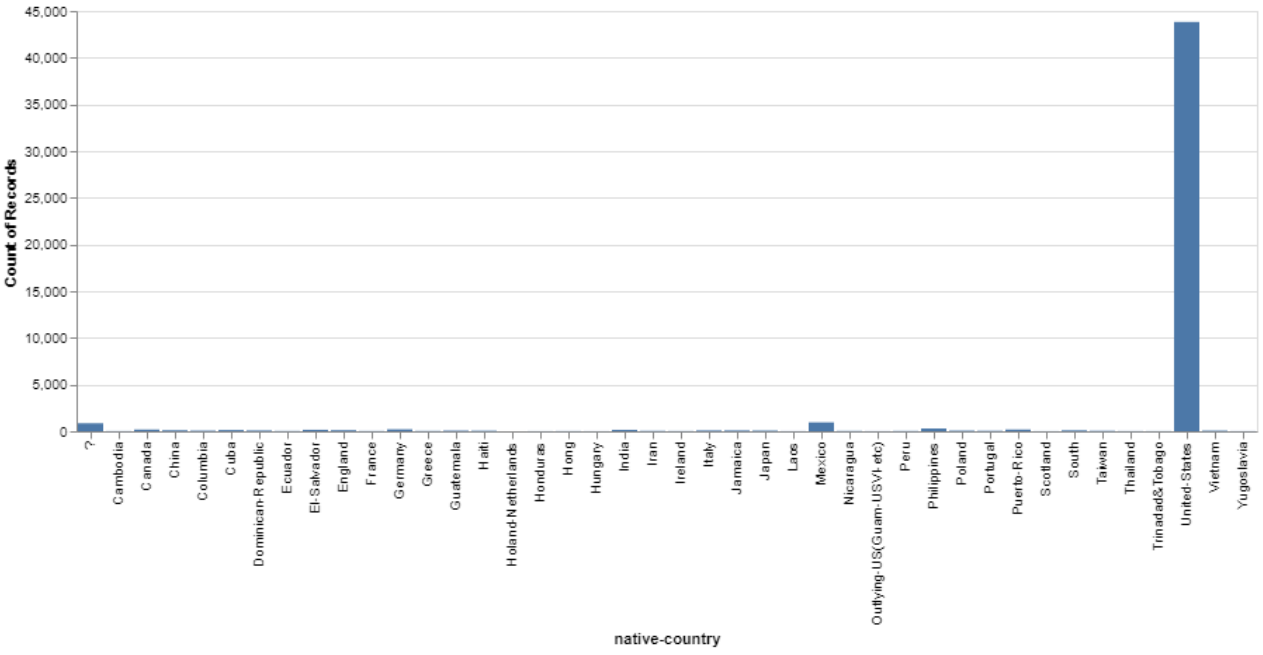
For **race**:
Most frequent race is ‘White’, with 41k records over 48k. Second highest is ‘Black’, with only 4k values.



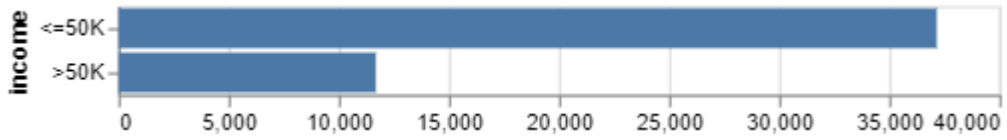
For **gender**:
As previously observed, men are 32k , and women only 16k, this fact influences the relationship attribute.



For **native-country**:
91% of the records are represented by the United States, the rest of the 41 countries share the remaining 9% with missing values.

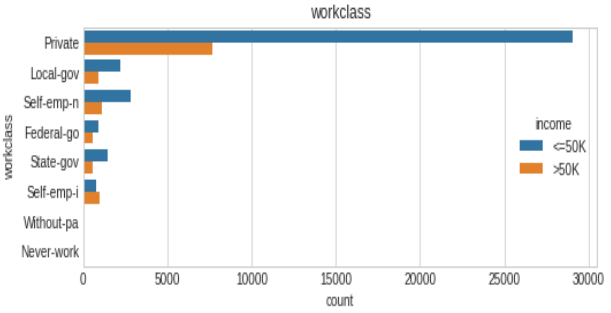


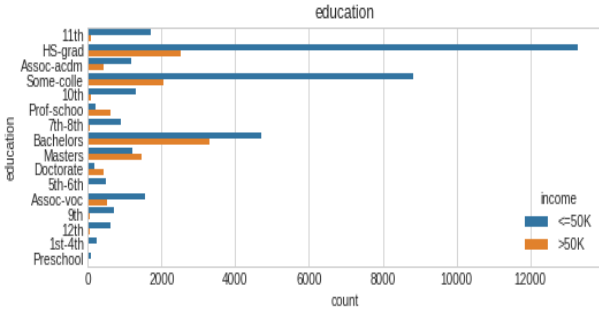
For **income**:
Almost 75% of the records have an income lower than 50K a year. It will be important to keep that in mind for the classification tasks where we will have to predict two labels with very unbalanced frequencies.

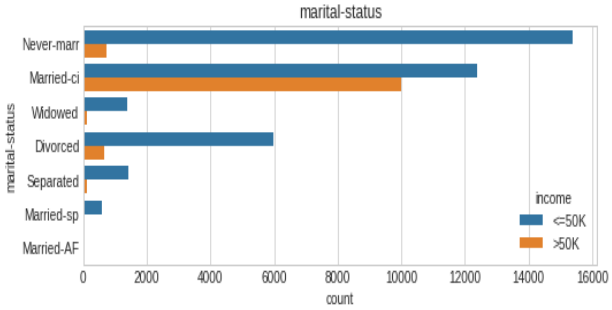


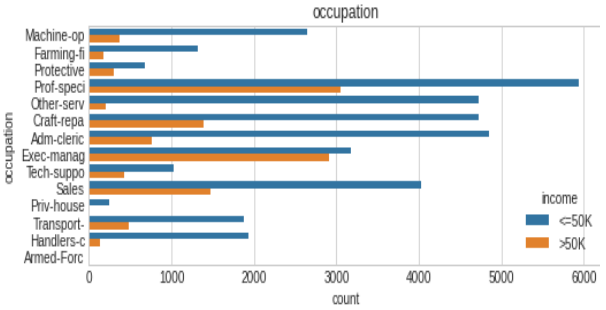
Bivariate analysis

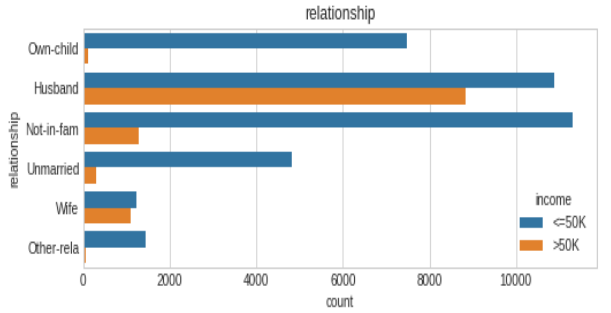
With the following plots we checked how each attribute behaves with respect to the target attribute "income".

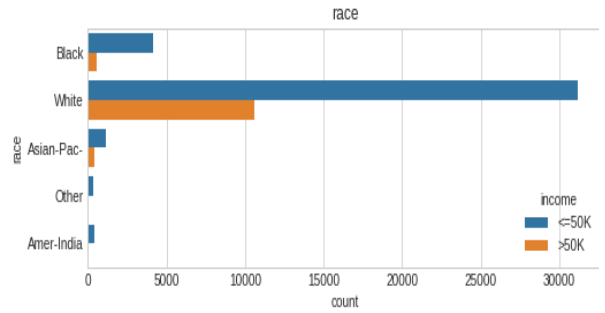


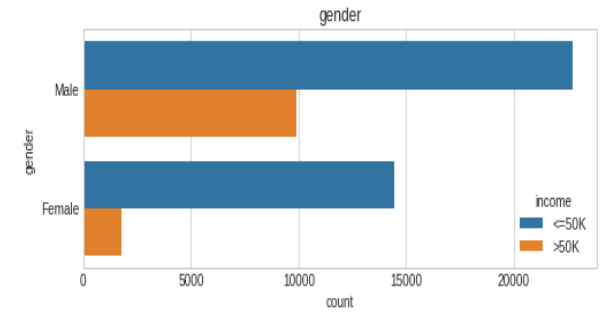


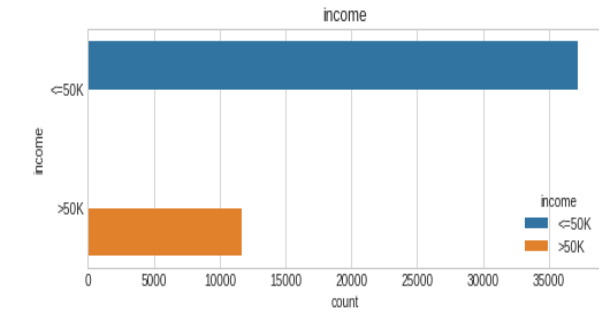












Some interesting facts from these plots are:

- with respect to education, people holding a master, a doctorate or a professional school are the only ones where the frequency of income >50K is higher than >=50K,
- marital status: married people have a much higher probability of gaining more than 50K than the people who are not. We can deduce the same information from the relationship graph,
- gender: men earn more than women, on average,
- with respect to the workclass, the only category where people earning more than 50K overcome the ones earning less than 50K is the self-employed. Executive/managers are the only ones who have almost the same probability of earning more than 50K or less than that. For all the other job categories the probability of earning more than 50K is much smaller.

Assessing data quality (missing values, outliers)

age	0.000000
workclass	5.730724
fnlwgt	0.000000
education	0.000000
educational-num	0.000000
marital-status	0.000000
occupation	5.751198
relationship	0.000000
race	0.000000
gender	0.000000
capital-gain	0.000000
capital-loss	0.000000
hours-per-week	0.000000
native-country	1.754637
income	0.000000

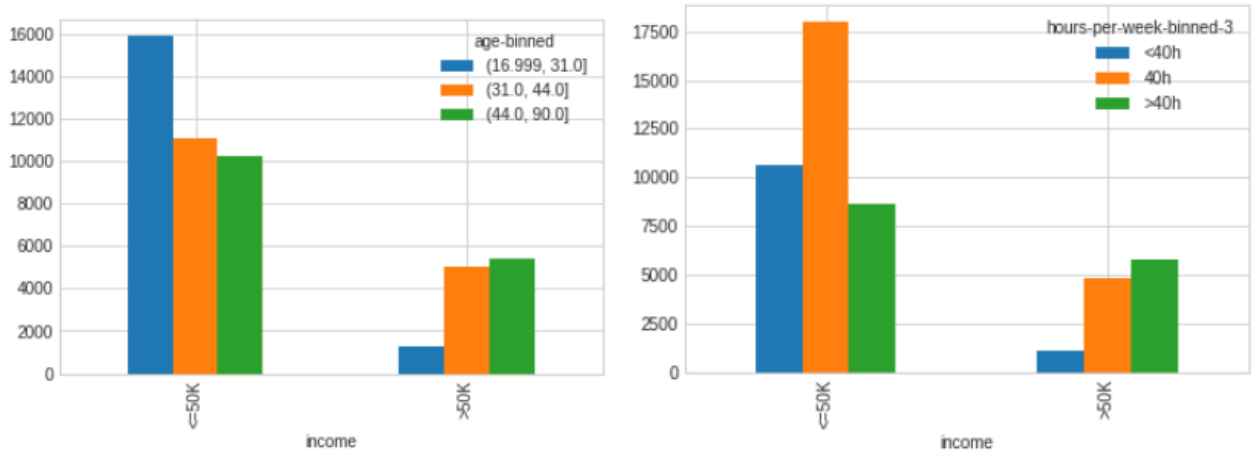
The variables having missing values are “workclass”, “occupation” and “native-country”, with the percentages shown to the left.

Since all the variables with missing values are categorical we decided to replace them with the mode of the single variables.

As for the outliers, the only variable with too many outliers was final-weight and since we didn’t consider it relevant for any of our analysis, we decided to remove it for all of the forthcoming phases.

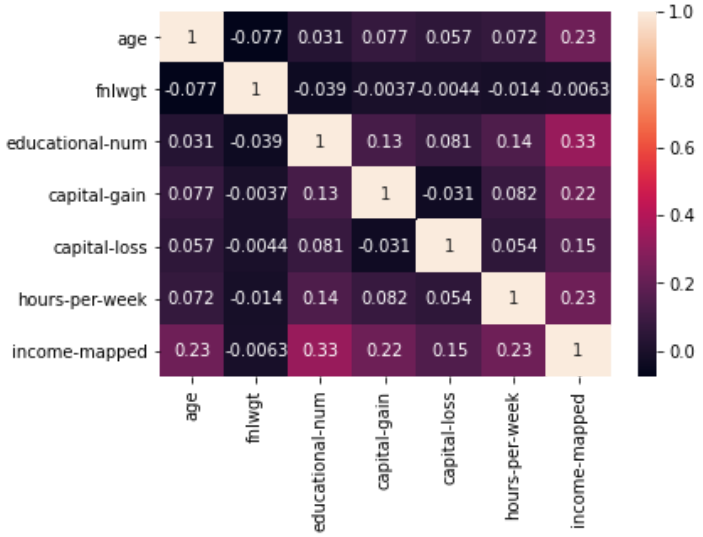
Variables transformations

We discretized some numerical variables in order to better visualize their distribution with histograms. Other transformations will follow in the preprocessing of each analysis.



From these graphs we can see that the people who earn more than 50K a year have usually more than 44 years of age and 40 hours a week or more.

Pairwise correlations and eventual elimination of redundant variables



No interesting correlation between the numerical variables in our dataset emerged, as we can see from the heatmap to the left. For this reason there was no need to eliminate variables which could be redundant.

B. CLUSTERING

Clustering pre-processing

Since both the algorithms we used for clustering (K-means and DB-Scan) compute distances between the points, it was important to exclude all the categorical variables. We retained only the numerical variables, namely:

- Age
- Educational-num
- Hours-per-week
- Capital-gain
- Capital-loss

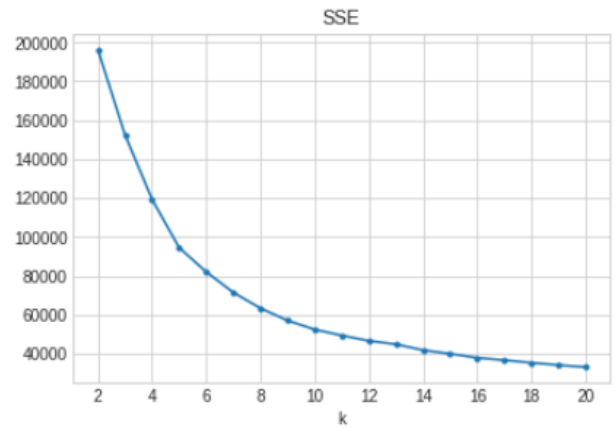
They are all numeric variables besides educational-num which is ordinal (but has many categories so it can be used as a numerical variable for our purposes in this context).

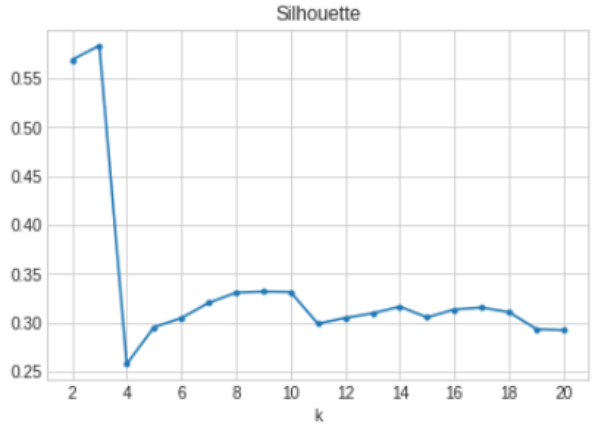
The income attribute was dropped since it is a binary variable of 0s and 1s.

All the variables were normalized with the standard normalization in order to avoid a defective clustering due to the different units of measure and the different scales of the variables in exam.

Analysis by center-based clustering

For the center-based clustering, we used the k-means algorithm.
We decided to use the euclidean distance, being the standard metric for numeric variables like the ones we were dealing with.
For the selection of the best value of k we performed various k-means with a number of clusters ranging from 2 to 20. Here the results:



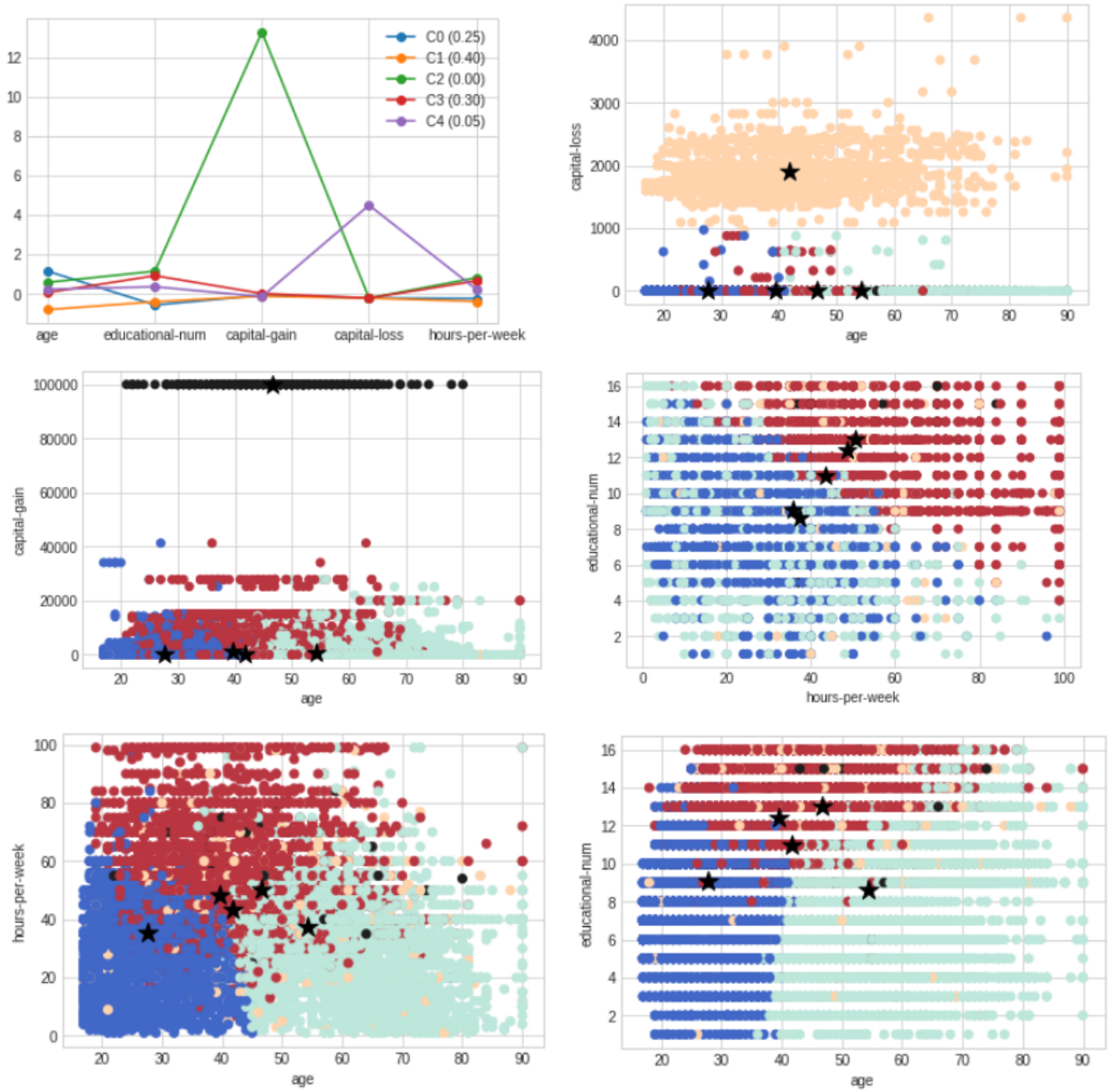


The SSE suggests that a number of clusters between 4 and 8 could be acceptable. The Silhouette coefficient is surely too low for the model with 4 clusters but relatively similar for the models with 5 to 8 parameters. In this case we chose the model with the lowest complexity, which is the one with 5 clusters.

The clusters we obtained have very different dimensions, with cluster 2 and 4 with a really scarce number of records:

```
Cluster 0: 12146 (0.25%)
Cluster 1: 19357 (0.40%)
Cluster 2: 244   (0.00%)
Cluster 3: 14857 (0.30%)
Cluster 4: 2238  (0.05%)
```

Below are the graphs with the characterizations of the clusters with respect to the variables used for k-means:



As we can see from the graphs shown above, the two clusters with very few records are characterized mainly by the features “capital-gain” and “capital-loss”. One is composed of people who have a high capital-gain, have a higher level of education and work more than 40 hours a week. The other one is composed by people who have some capital-loss.

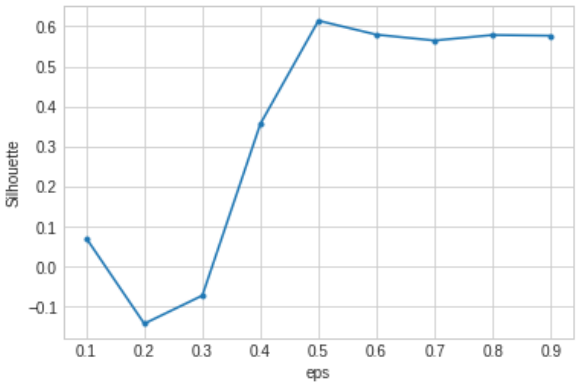
Of the other three, one seems to be characterized by people over 40 years of age, who work mainly 40 hours a week or less and don’t reach the highest levels of education; another one is composed by people under 40, who work 40 hours a week or less and don’t reach the highest levels of education; the last one is characterized by people who are generally not too old, work more than 40 hours a week and have a high level of education.

Analysis by density-based clustering

For what concerns the density based clustering, the algorithm employed was the DB-Scan.

In the graph below, we can see how the Silhouette changes by varying the value of epsilon from 0.1 to 0.9. The Silhouette was calculated excluding the noise points.

The value of the epsilon that maximizes the Silhouette score is 0.5, corresponding to a 0.6 Silhouette score.



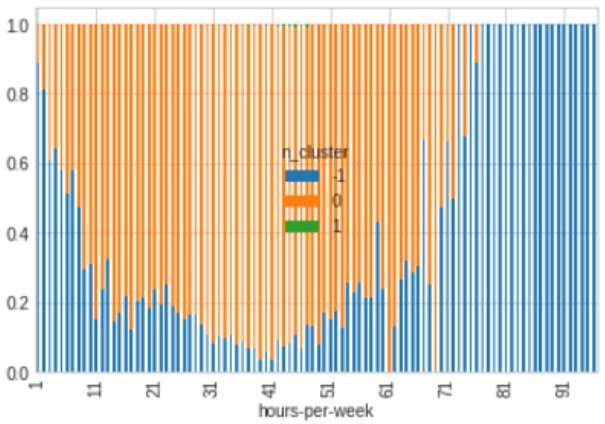
However, if we check the distribution of the points in our clusters, we observe that we only have 2 clusters, and that cluster 0 has 87% of the total points, while the rest of the points are noise points and only 162 records out of 48k are part of cluster 1. The result is an unuseful clustering.

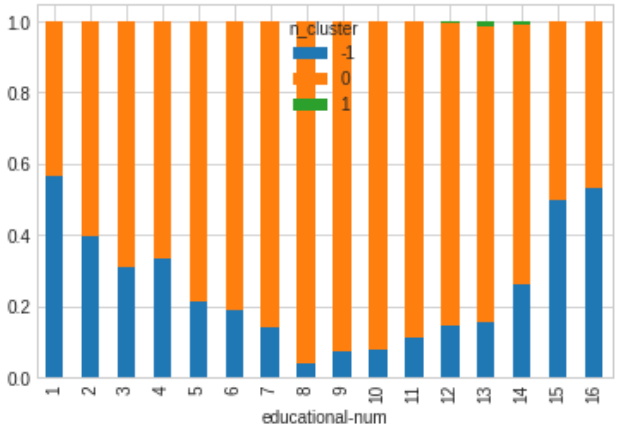
```
Cluster -1: 6428 (0.13)
Cluster 0: 42252 (0.87)
Cluster 1: 162 (0.00)
```

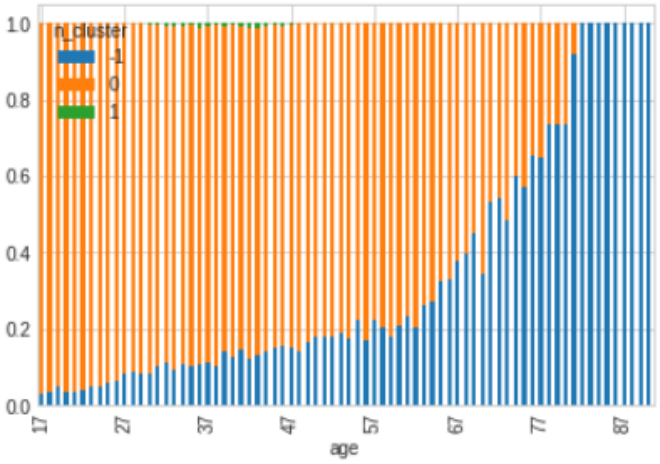
It may be interesting to execute the same analysis by employing the jaccard distance instead of the euclidean.

Unfortunately, due to the high number of records contained in this dataset, colab's RAM proved to be insufficient to execute this analysis (by crashing "because all the memory was in use").

Below, we report the distribution of the clusters with respect to age, hours worked per week and educational-num.







As expected, cluster 1 (in green) never shows up.

Analysis by hierarchical clustering

For the same reason we did not check jaccard distance in the section above, we had to give up hierarchical analysis too. Hierarchical clustering in fact, by firstly employing all records as single clusters and then computing distance matrix for all of them, uses all colab RAM, making it crash.

Final evaluation of the best clustering approach and comparison of the clustering obtained

The only algorithm which gave interesting results was k-means, that proved to be the best one for our kind of analyses and data.

C. CLASSIFICATION

The prediction task is to determine whether a person makes over $50K a year (target variable: “income”).

Preprocessing for Decision Tree Classifier

Classification needs the dataset to be preprocessed. The Decision tree classifier requires categorical or discretized numerical features in order to build the model.
First of all, the 2 values of the target attribute *income*, that are in “string” form (> 50K and <= 50K), are replaced by numeric values 1 and 0 respectively.
Gender is mapped in 1 for females and 0 for males.
We then reduced the number of categories of the categorical features which had too many values:

- For what concerns *Native-Country*, since its distribution sees the United-States covering more than 90% of the total, we created a mapped variable with 1 for the US and 0 for all the other countries.
- The two variables related to *education* had too many categories (16) so we mapped them in a new variable with 6 categories, based on the american educational system. They are: dropout, highschool grad, community college, bachelors, masters and doctorate.
- *Race*, that contains 5 unique values, of which White covers the majority, is again mapped into 1 for White and 0 for the rest.

We binned the numerical variables:

- *Age* class, that ranges from 17 to 90 years old (74 unique values), was mapped into 3 categories, using its terziles. This binning therefore divides age into 17-30 years old people, 31 to 44 and 44 to 90.
- *Hours per week*, that ranges from 1 to 99, are binned into two groups using the 50th percentile (1 to 40 hours-per week and 41 to 99).

We dropped some variables:

- *Occupation* and *work class* attributes, with their 15 and 9 categories respectively, are both hard to be binned, and therefore are dropped.
- *Relationship* and *marital status* are qualitatively not much different, and therefore only marital status is kept.
- *Capital gain* and *capital loss* are also removed, since they’re not believed to bring much added value to the model.

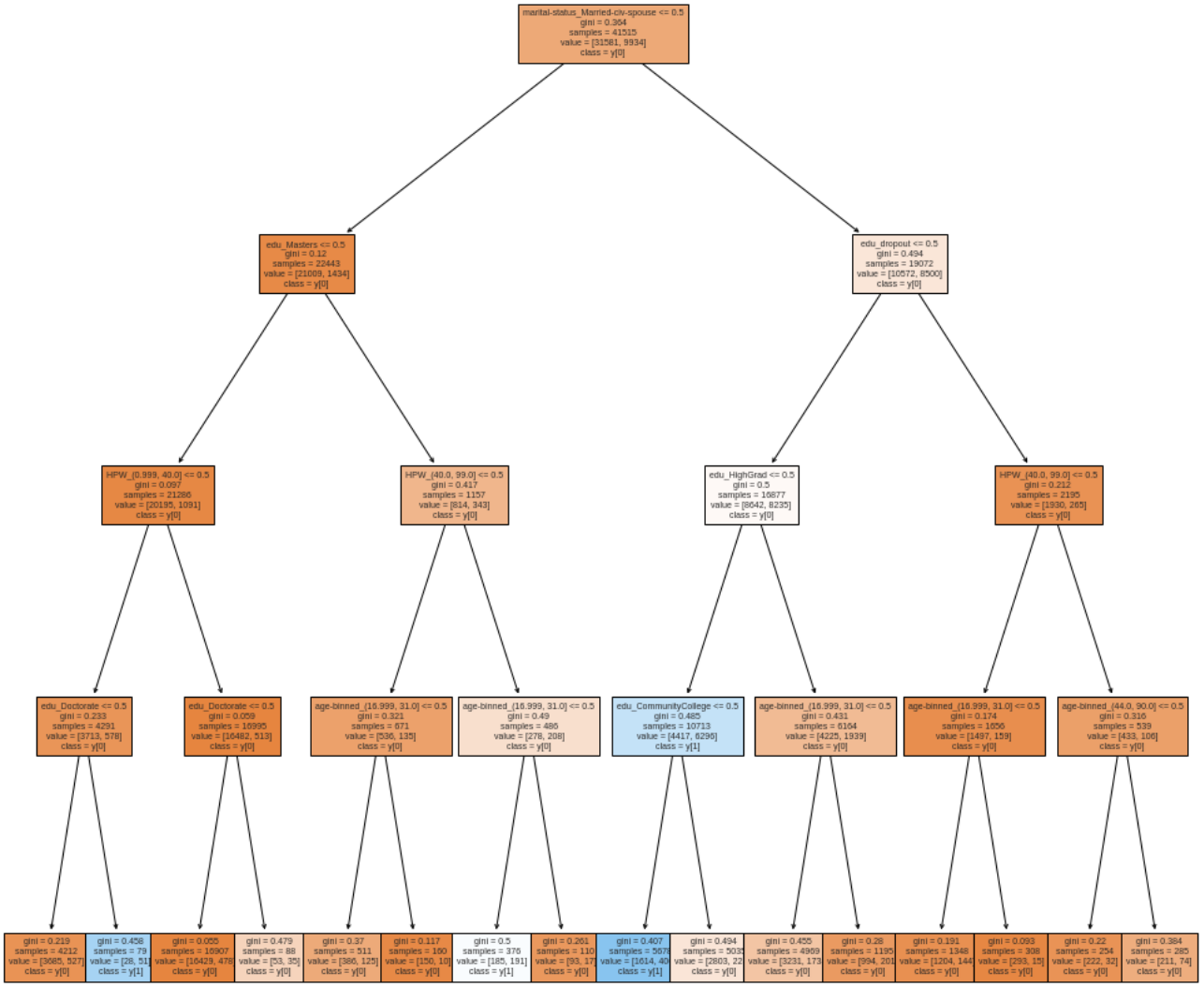
Finally, one hot encoder is applied to the dataset in order to get dummy variables for all the binned attributes.
The resulting dataset is therefore 48.842 rows × 22 columns.

Decision Tree Classifier

The first classification algorithm we applied was the decision tree. We tested different parameters in order to find the best tuning that optimizes the accuracy. The parameters that were tested are the following:

```
'max_depth': [2, 3, 4, 5, None]
'min_samples_leaf': [1, 2, 3]
'criterion': ['gini', 'entropy']
```

The best model resulting from the tuning of these parameters is:
{'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 1} , yielding the tree below:



This model has an accuracy of **0.82**. The tree shows that only three combinations of attributes lead to an income of more than 50K:

- those who are **unmarried**, that **work more than 40 hours per week** and have a **Doctorate**,
- those who are **unmarried, have a master and work more than 40 hours per week, aged 31 years or more**,
- and **those who are married,** and **either have a doctorate or a master degree** (not adding much information after the previous two, if not for the fact that these people are married).

Nevertheless, Gini coefficients of the three combinations range from 0.4 to 0.5, and thus these combinations do not lead to significant conclusions.

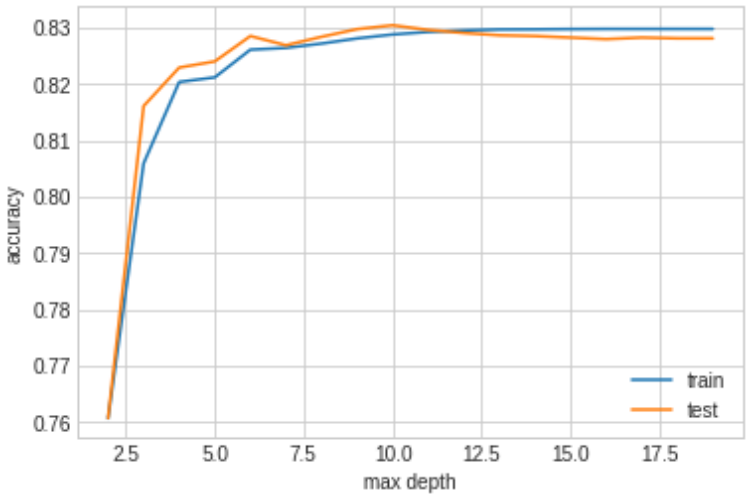
Amongst all, the leaf showing the lowest Gini coefficient (**0.055**), and therefore the **highest** grade of **purity** is the one indicating that: people who are **not married**, who **have neither a Master degree** nor a **Doctorate** and **work less than 40 hours per week**, are going to earn almost surely **less than 50K**.

It is interesting to note that, as highlighted in the table below, the decision tree employed and shown above, lead to a **low value of recall (0.44)** for the **class 1** (income above 50K). This is certainly due to the distribution of our target attribute, where only 26% of the population observed earns more than 50K, leading the classifier to assign an income lower than 50K to most of the records in the model, and therefore

yielding to a **large number of false negatives for income 1** (mostly people with income higher than 50K labelled wrongly). This is confirmed by the performance of the classifier with income 0 (income below 50K), a precision of 0.84 and a recall of 0.94 (very few false negatives in this case).

	precision	recall	f1-score	support
0	0.84	0.94	0.89	5574
1	0.71	0.44	0.54	1753
accuracy			0.82	7327
macro avg	0.78	0.69	0.72	7327
weighted avg	0.81	0.82	0.81	7327

Finally, in the following graph, we studied how accuracy behaves as the max depth is increased. As expected, as max depth increases, accuracy increases, following approximately a log-shape: at the beginning the increase in depth results in a large increase of the accuracy, and at around depth 10 the accuracy stabilizes at around 0.83. The choice of the parameters above stopped at max depth 5, since a further depth of the tree may result in overfitting and it would be more difficult to visualize and read. Moreover, this choice does not result in a huge loss of the accuracy (let's rimind that accuracy with depth 4 is 0.82).



Preprocessing for K- Nearest Neighbours Classifier

The second classifier that is trained and tested is the K-Nearest Neighbours. For this task, since the algorithm is based on distances, a number of attributes that were dropped before are brought back, and similarly the binned attributes are no longer suitable in this case.
Therefore, we need to recover initial values of attributes:

- Age
- Educational-num (an ordinal number ranging from 1 (lowest level of education) to 16 (highest level of education)
- Capital gain and capital loss
- Hours per week

And encode categorical attributes with one hot encoder:

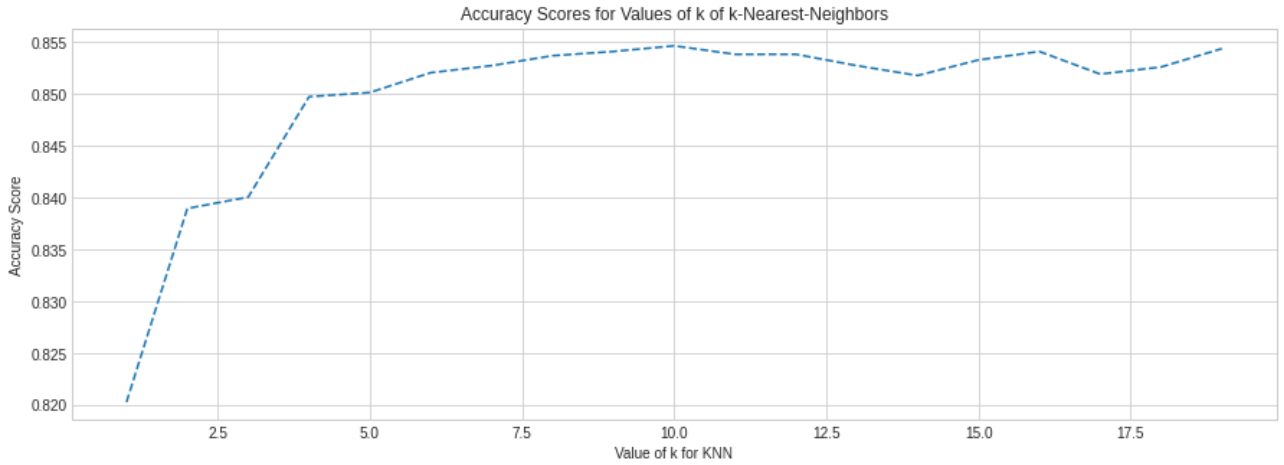
- Marital status
- Race

- Gender

In this case the dataset we work on has the shape 48.842 rows x 20 columns.

K- Nearest Neighbours Classifier

Again, multiple KNN algorithms are tested in order to find the optimal number of k nearest neighbours in terms of accuracy.
From the following graph, we observe that the optimal number of neighbours is 10, with an accuracy score of 0.85.



Therefore, performing a KNN with k set to 10, the algorithm will yield the following classification report:

	precision	recall	f1-score	support
0	0.85	0.95	0.90	5574
1	0.76	0.48	0.59	1753
accuracy			0.84	7327
macro avg	0.81	0.72	0.75	7327
weighted avg	0.83	0.84	0.83	7327

From the table, we once again confirm how the model succeeds at predicting the **income 0** (same as decision tree), while in the case of **income 1**, recall value is still low (**0.48**), with a reasonable precision value (**0.76**).

Comparison of the two classifiers

As seen in the previous two analysis, K-Nearest neighbours proved to be more accurate, with higher precision and recall on average for both target class values 0 and 1 (<=50K and >50K).

D. ASSOCIATION RULES MINING

Preprocessing for pattern mining

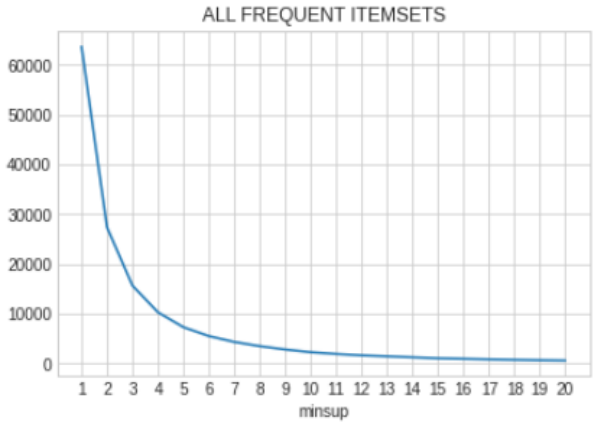
Other manipulations of the variables were needed in order to apply pattern mining techniques.
We kept all the variables we had in the initial dataset (besides the final-weight) and we binned the numerical ones and reduced the number of categories of some categorical variables who had too many.
We started our analysis with the following variables (the number of unique values for each variable is also reported):

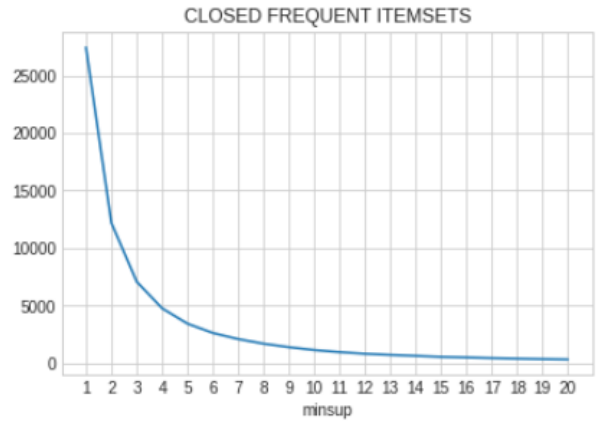
workclass	8
occupation	14
race	5
gender	2
capital-gain	2
capital-loss	2
income	2
age-binned	3
hours-per-week-binned	2
education-grouped	6

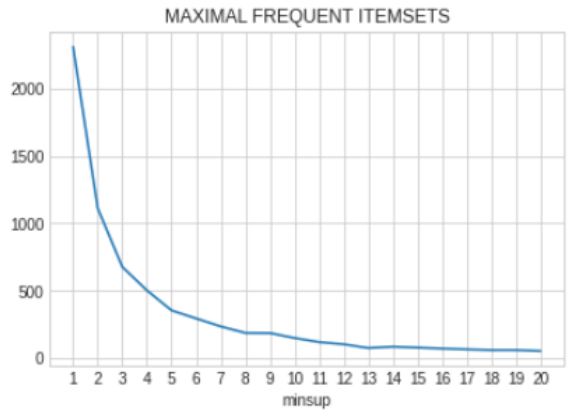
We transformed our data from matricial to transactional data in order to apply the pattern mining techniques.

Frequent patterns extraction

We used the apriori algorithm to extract the frequent itemsets, closed frequent itemsets and maximal frequent itemsets. Here are the results for the three cases with minimum support ranging from 0 to 20:







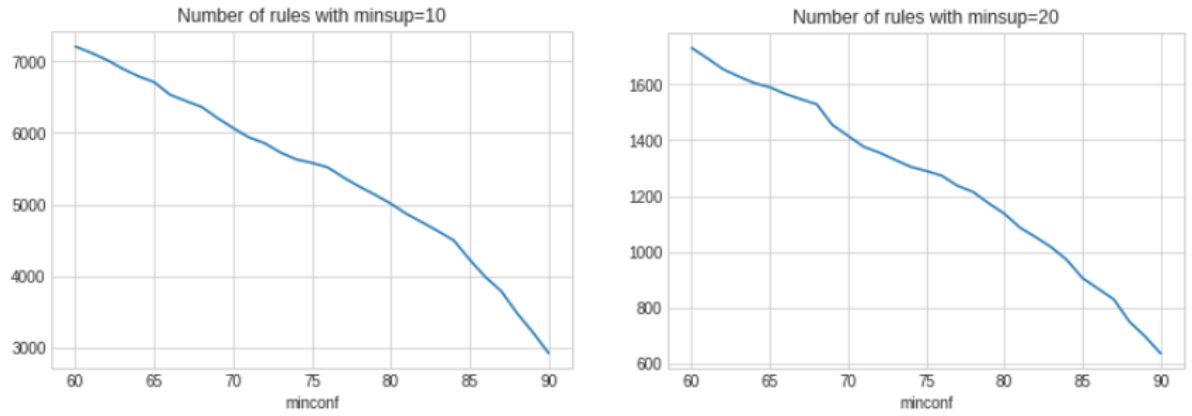
We know that closed frequent itemsets are a subset of all frequent itemsets and maximal frequent itemsets are a subset of the closed frequent itemsets, so it is coherent to observe that the frequency of itemsets is decreasing in the three graphs. With a minimum support of 1 all the frequent itemsets are more than 60.000, the closed frequent itemsets are more than 50.000 and maximal frequent itemsets are more than 3.000. With a minimum support of 10 we have

respectively 1.922, 1.837 and 181 itemsets. With a minimum support of 20 we have respectively 449, 449 and 58 itemsets.

The number of itemset is of course decreasing as the values of the minsup raises (as seen in the plots above).
With both a minimum support of 10 and 20, the most frequent itemset is [0-gain, 0-loss] which is not particularly interesting. Since we dichotomized "capital gain" and "capital loss", we checked the percentages of the two categories and 1-gain/1-loss didn't exceed the minimum support for both 10 and 20 minsup. This means that finding 0-gain or 0-loss in an itemset is not very informative.
The most frequent of the maximal frequent itemsets with minsup=20 could be interesting: *white male* who has a *private* job, *doesn't gain* or *lose* any money outside her salary (this one not particularly interesting, as just said above) and gains *less than 50K* a year. The same itemset but with "female" turns up with minsup=10.

Association rules extraction

We used the Apriori algorithm to extract the association rules. We tried various levels of confidence, ranging from 60% to 90% for both a minimum support of 10% and 20%. Here the results:



As expected the number of rules is decreasing for increasing levels of confidence. The number of rules is much lower for the second grafh, i.e. for a higher level of support.

Interesting rules

To investigate the interesting rules we started from a model with minsup=20 and minconf=80. The number of rules emerging from this model was 1.137 and it seemed a reasonable number of rules to study.
From this set we selected only the 21 rules with interest higher than 2.4. Only 5 of them contained the target variable and here we report them:

```
('Married-civ-spouse', 'Male', '<=50K', 'White')                    -->
Husband
with lift: 2.4463961065617634

('Married-civ-spouse', 'Male', '<=50K', '0_gain', '0_loss')         -->
Husband
with lift: 2.4400491457195357

('Married-civ-spouse', 'Male', '<=50K', '0_gain')                   -->
Husband
with lift: 2.440420169127664
```

```
('Married-civ-spouse', 'Male', '<=50K', '0_loss')                    -->
Husband
with lift: 2.439848504545324

('Married-civ-spouse', 'Male', '<=50K')                              -->
Husband
with lift: 2.440205801761027
```

We can see that all these rules are similar: being a married man who gains less than 50K a year (and some other eventual feature like being white or having no gain or losses outside the salary) implies being a husband.
This doesn't seem very relevant and we notice that marital-status could be related to the relationship attribute. So we tried to delete the "relationship" feature to see if we could have more interesting results.
Following the same steps as before we couldn't find any rule with a lift higher than 2 so we had to lower this standard. With a minimum lift of 1.3 we found 34 interesting rules, 14 of which contained the target variable. The only one where the target variable was implied by other features was the following:

```
('Never-married', '(16.999, 31.0]_age-class', 'Private', '0_gain')   -->
<=50K
with lift: 1.3008126981747583
```

This means that being young, not married, working as a private and making no investments implies gaining less than 50K a year. The interest of this rule was not too high, though.
Other rules which contained the target variable and had a higher lift were the following:

```
('>50K', 'Male')                  --> Married-civ-spouse
with lift: 1.9525364073701417

('>50K', 'White')                 --> Married-civ-spouse
with lift: 1.8728242551429999

('>50K', '0_loss')                --> Married-civ-spouse
with lift: 1.8593519655188522

('>50K',)                         --> Married-civ-spouse
with lift: 1.8644652411055356
```

We can see that gaining more than 50K a year and being either male, white or with no money/investment losses implies being married.