0.0 **Remarks**

1. Since we use different preprocessing method (even the numbers of inputs are different) for different classifier, so we break the original structure of the program.
2. How to run:
   - All of the python code for plotting (Section 3) are independent programs. They can be run by themselves, i.e. not a subprogram of prj_main.py.
   - To get the prediction, you can simply run prj_main.py.
3. Prediction: ./predictions/vote_predictions.csv are the one we predicted by voting in four classifier. It seems to have the best result.

2.2 **Statistic Analysis of Data**

We have done a statistical analysis on the data. We used ANOVA test to analyse the variance of the effect of different factor to y (Happy).

And we find that the below factors are most significant to the result, with setting the threshold to be 0.05.
['Q118237','Q101162','Q107869','Q102289','Q98869','Q102906','Q106997','HouseholdStatus','Q108855','Q119334','Q115610','Q108856','Q120014','Q108343','Q116197','Q98197','Q116448','Q102687','Q114961','Q108342','Q113181','Income','Q117186','Party','Q115390','Q112512','Q102089','Q116953','Q115611','Q121011','Q111580','Q99716','Q106993','Q109367','Q114152','Q106389','Q116441','Q123621','Q113584','Q124742','Q108617','Q116881','Q117193','Q100689','Q115602','Q98578', 'Q120012','Q100680','Q112478','Q106272','Q99982','Q109244','Q98059','EducationLevel','Q96024','Q111848','Q119851','Q118233','Q119650','Q108950','Q102674']

We later use this to train two of our classifier. (* We use different preprocessing method for different classifier.

2.3 **Missing Data Filling Method**

To begin with, we use {False: -1, Missing: 0, True: 1} as our baseline missing data filling method. Any imputing method should be better than it, or else we don't adopt it. We use {False: -1, Missing: 0, True: 1}, because in number analysis, missing is actually uncertain rather than something beyond true or before false, so we set the missing value in between false and true.

We have tried simply using mean, mode or median to fill the data, also tried using different ways to group the users, and by decision tree to get the mean, mode or median. But it turned out none of the method is more successful than the baseline.

Furthermore, as we want to try something we haven't learnt before and other people usually not tried. We have even tried MICE algorithm, a new sklearn imputing method on a subbranch in sklearn GitHub which haven't on the stable release yet, as attached in ./MICE.py. We port it and link to relevant dependencies by ourselves, but turn out still, it is no better than the baseline.

To make sense of it, we try to explore the data again. We find that on average, there are nearly 30% of the data are missing, and in some fields, there are even more than 50% of the data are missing. In such a ratio, to impute any value to the missing may cause serious bias to the data.

Therefore, we follow our baseline method, which we consider missing value as a signal, but a slightly weaker signal between True and False.

3.1.1 Logistic Regression

```
Train the logistic regression classifier
Accuracy: 0.727027027027
Runtime: 0.10343503952026367
Train the logistic regression classifier-sklearn
Accuracy: 0.718918918919
Runtime: 0.027269840240478516
```

The running time of our classifier is longer than that of sklearn, but the accuracy of our classifier is better than that of sklearn in default setting, it probably due to the optimisation algorithm we use in finding the minimum. We adopted the truncated Newton algorithm.

3.1.2 Naïve Bayes

We choose to use MultinomialNB classifier in our classification. Because in this problem, most of the fields are categorised input, which is not as continuous as gaussian distribution and it is not just binary.

```
Train the naive bayes classifier-MultinomialNB
Accuracy: 0.718918918919
Runtime: 0.009047985076904297
Train the naive bayes classifier-sklearn-MultinomialNB
Accuracy: 0.718918918919
Runtime: 0.0053060054779052734
```

The running time of our classifier is slightly longer than that of sklearn, but the accuracy of our classifier is the same as that of sklearn. It is probably due to there is no random factor, such as optimisation or others in this model, just some matrix operation. Therefore, the results are the same.

3.1.3 SVM

Since we don't have too much prior knowledge of the invariance of the problem, therefore we blindly find the best one by searching. We therefore used the GridSearchCV function in sklearn, to use 6-folds of data to cross validate and to search for the best kernel.

We finally choose to use 'rbf' kernel function in our SVM classification, which gives us the best result.

3.3 **Report Writing (Answering the questions)**

Q: What are the characteristics of each of the four classifiers?

Logistic Regression:
It is a relatively simple model which outliners can be really a huge problem. We find that more factors feeding into this model can be a bad thing. By stemming factors, the result can be better in this model, which is not always true for other models.

However, it is a model easily affected by how the optimisation process, because it involves a procedure finding the minimum, which may be trapped into the local minimum.

Naive Bayes:
Same as LR, it is a relatively simple model which outliners can be really a huge problem. We find that more factors feeding into this model can be a bad thing. By stemming factors, the result can be better in this model, which is not always true for other models. H

However, in the model we use, MultinomialNB classifier, the parameter is somehow not relevant, the results seems not changing much when we change the alpha value.

SVM:
SVM is a relatively complicated model compared with the previous two. We can input a lot of dimension into it and sometimes it even gives us better result.

The parameter tuning of it is important, which may affect the result a lot.

Random Forest:
Random Forest is a relatively complicated model too, like SVM. We can input a lot of dimension into it and sometimes it even gives us better result.

The parameter tuning of it is important, which may affect the result a lot. We also find that the accuracy generally increases with the number of estimators.

Q: Different classification models can be used in different scenarios. How do you choose classification models for different classification problems? Please provide some examples.

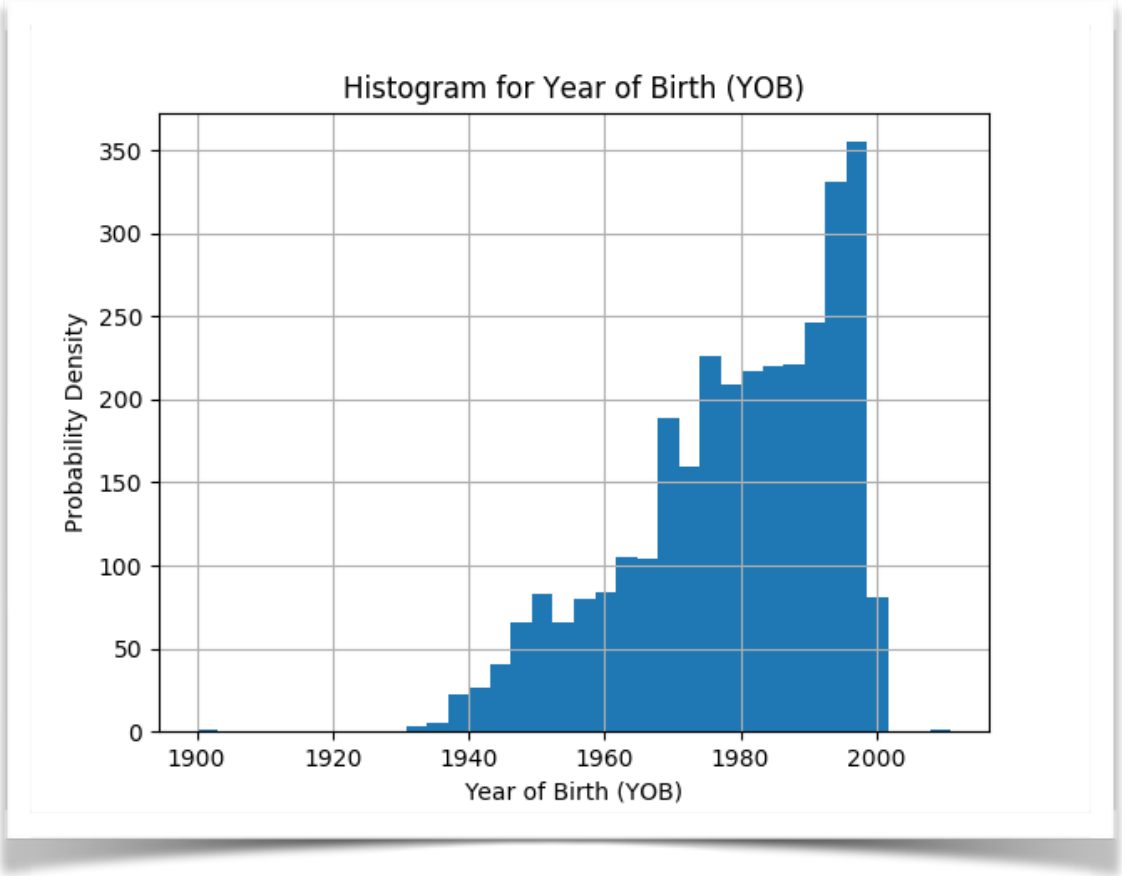
It depends on the distribution/the type of the data. For example, if the problem are with continuous data, feeding it into a model which is boolean based/categories based, which will get a bad result, vice versa. This problem can be solved by data preprocessing.

However, if it is the case of distribution, it will be not that easy to solve. For example, if it is a distribution combined of several Gaussian distributions, and we feed it into a model which assume a single Gaussian distribution, the prediction result will be very bad.

Q: How do the cross validation techniques help in avoiding overfitting?

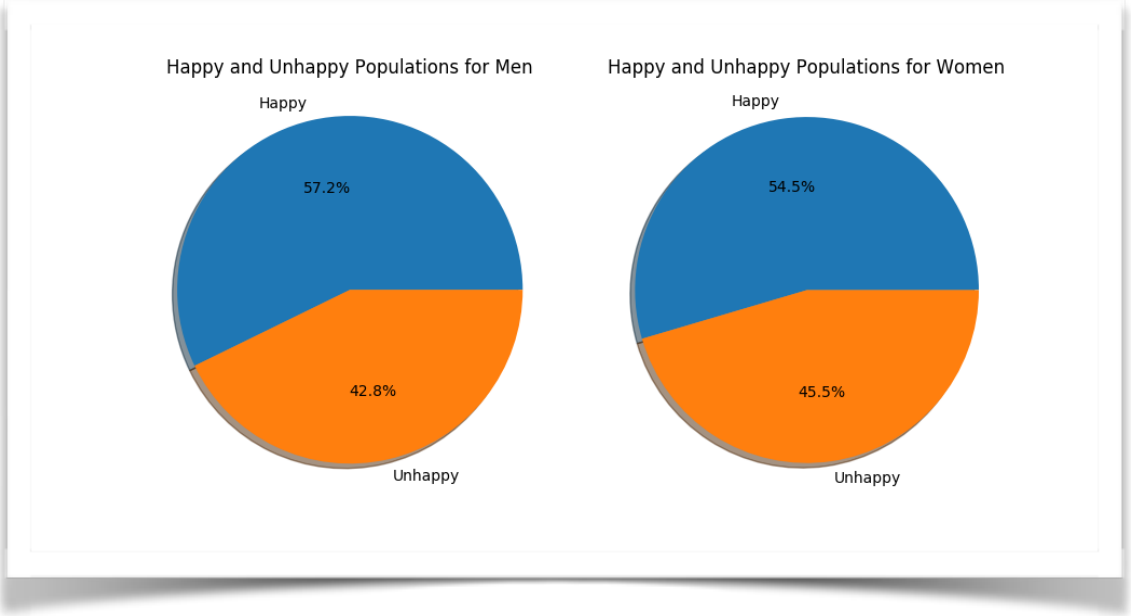
By cross validation technique, we can avoid overfitting by tuning the parameters which fit to the the test data only. In cross validation, we train the model by “rotating”. We divide the data into several folds (n-folds) of data and then, (n-1) and 1 fold(s) of data are used for training and testing rotatively, to calculate its accuracy. Then, we can prevent the accuracy is only valid for certain data.

4.1.1 **Histogram of YOB**



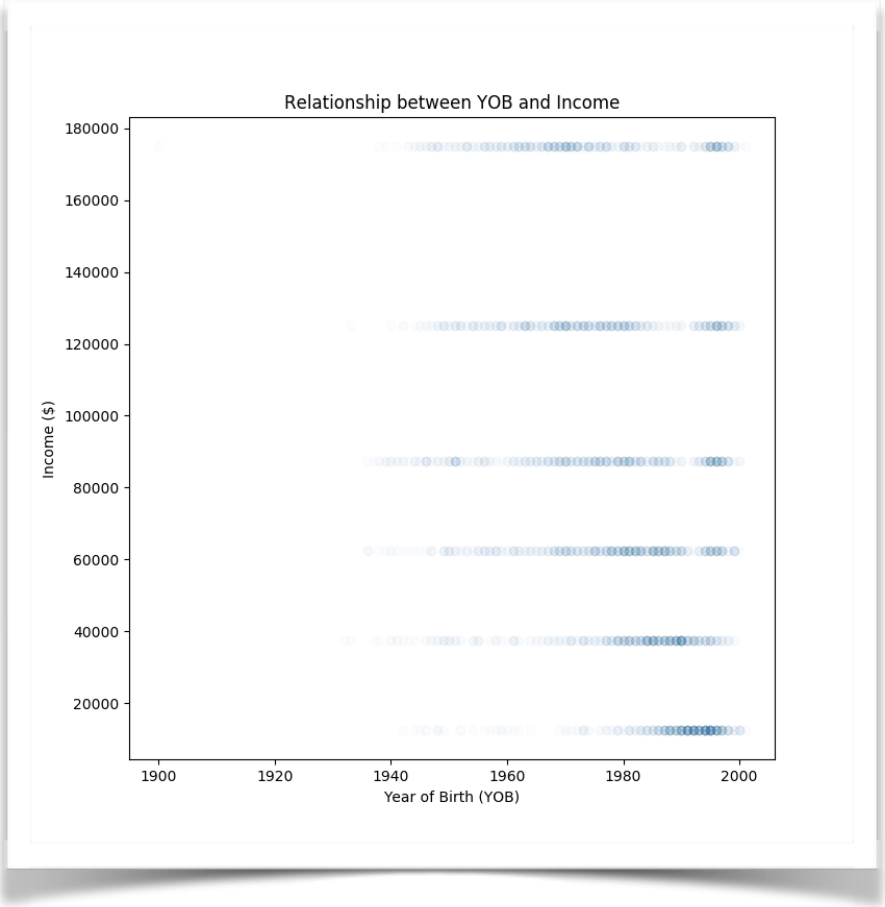
In the histogram, we can find that “Year of Birth” is not normally distributed, which is not a bell shape, and the majority of the people are born relatively near to 1990.

4.1.2 **Pie chart for the fraction of happy men/women**



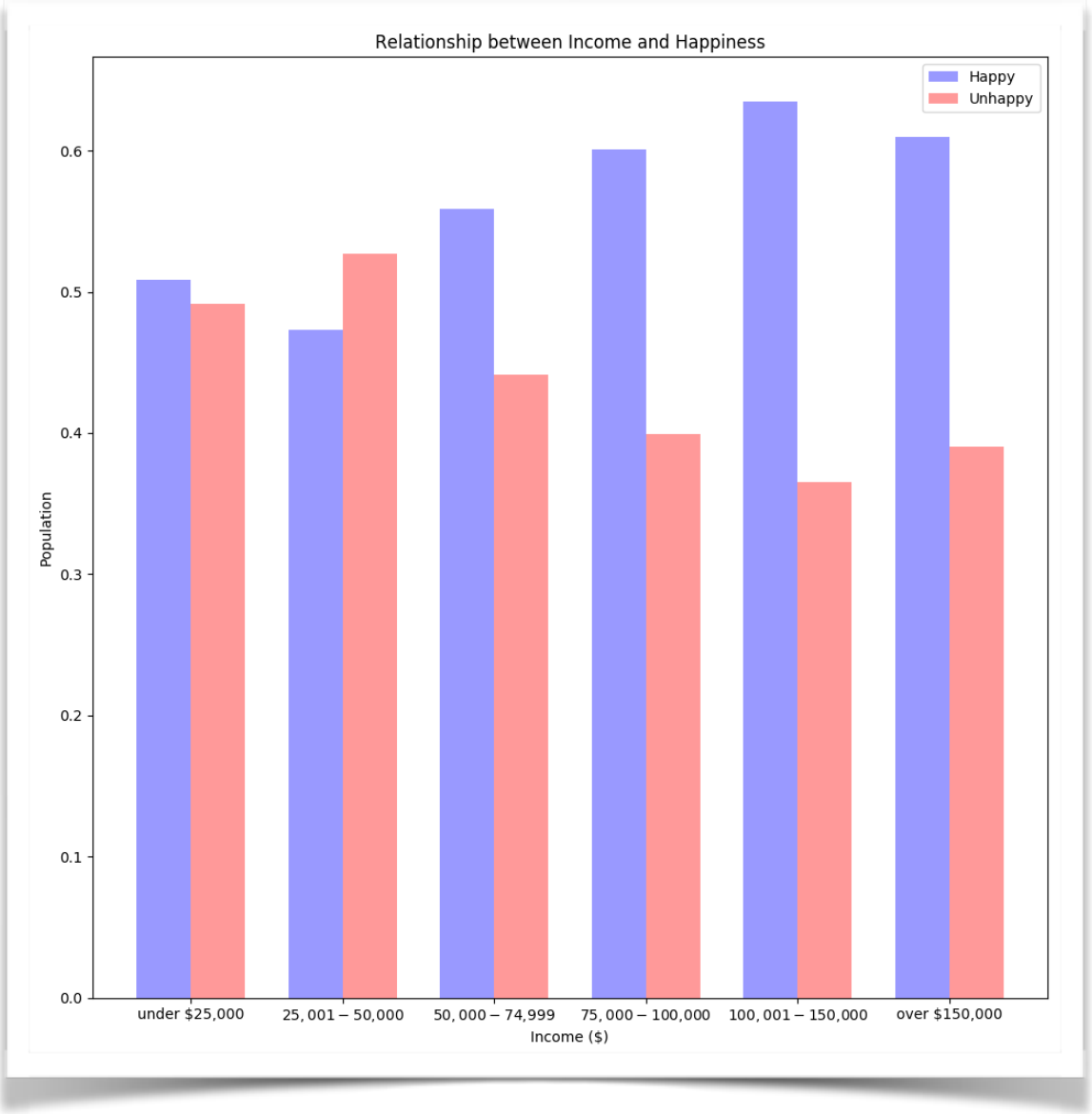
In the pie chart, we can find that more people are happy, and the ratio of men being happy is slightly larger than the women.

4.1.3 **Scatter plot of YOB and income**



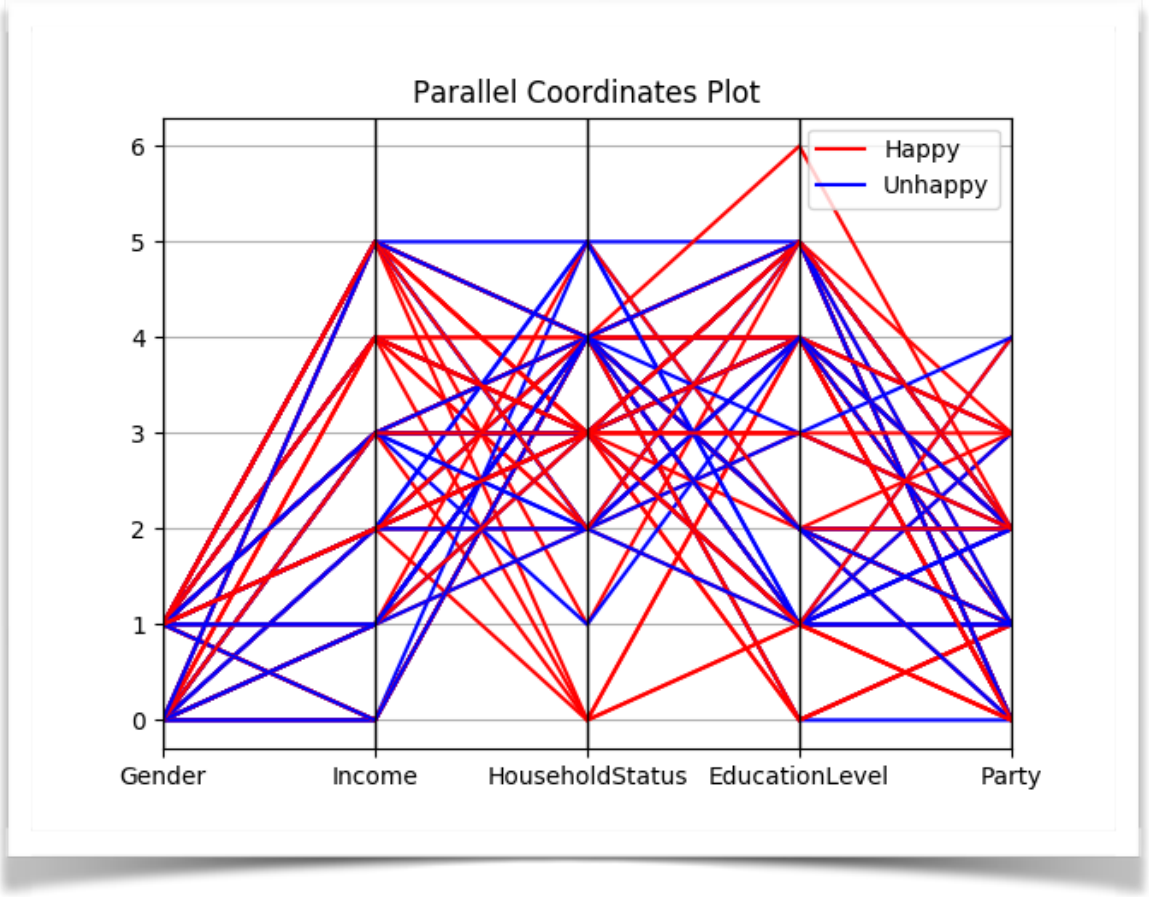
In the scatter plot, we can find that usually older people are with higher income.

4.1.4 **Bar chart of income and happiness**

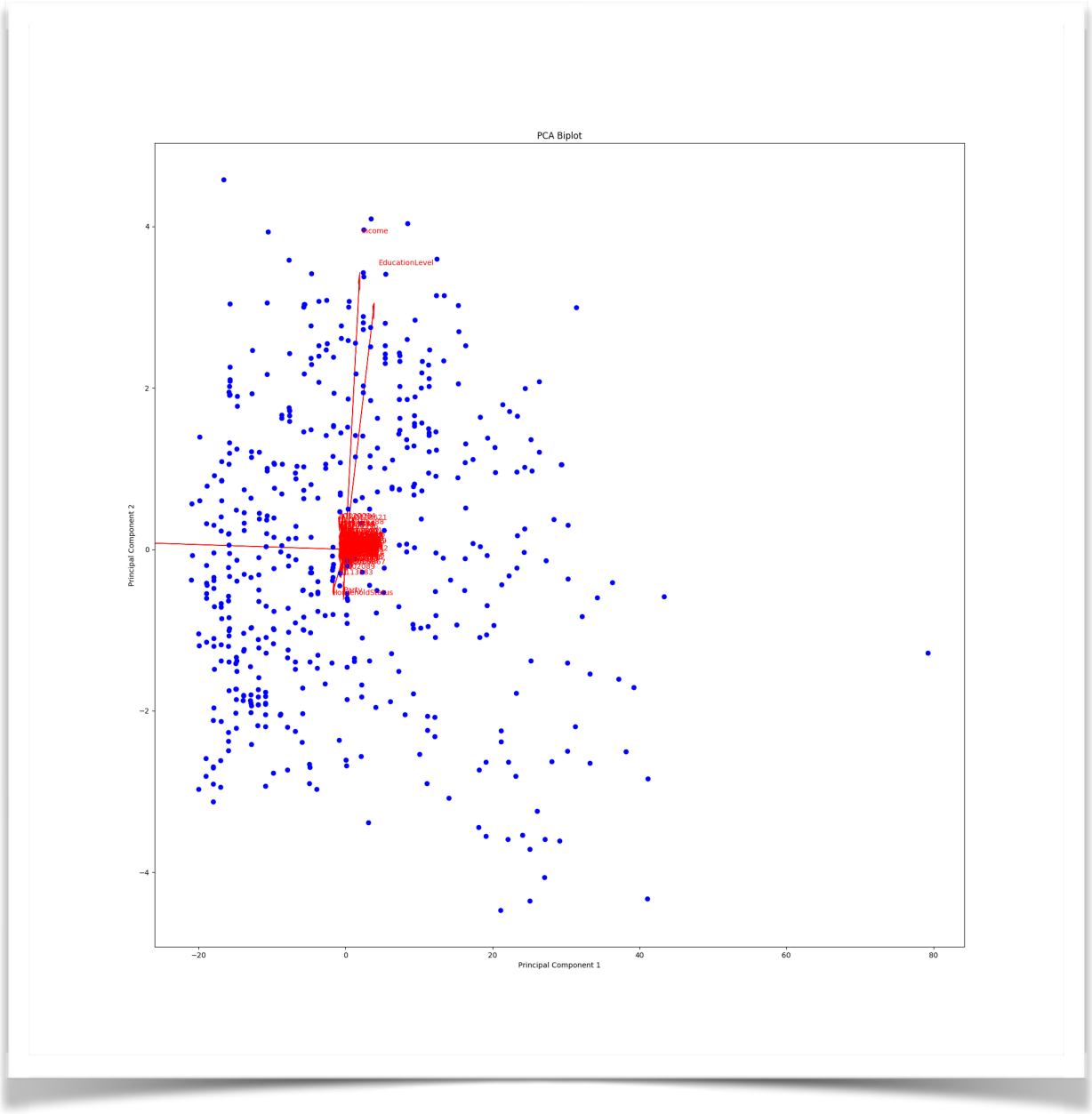


In the bar chart, we can find that the ratio of happy people increases with income, except in the lowest income group. *(It may due to the subsidy of the government? Need further research on the particular background of that society to have a certain answer.)*

4.2.1 **Parallel Coordinates Plot**



4.2.2 **PCA and biplot**



Q1: What's the physical meaning the vector corresponded to each variable? Explain it in one sentence.

The vector corresponded to each variable is the significance of each variable on the first two principal components.

Q2 What are the factors closely related to happiness according to this biplot? Write down your answer and use one more sentence to explain why.

Income and Educationlevel.
Because they are the ones with longest vectors in this biplot.

4.3.1 **Visualise SVM**