Technical University of Denmark


31015 Introduction to Machine Learning and Data Mining

---

# Report 1: Data Visualization

---

Morten Nissov, s163962
Emma Demarecaux, s176437
Johan Dybkjaer-Knudsen, s180049

# Introduction

When doing data analysis, a common framework is used as described in "Introduction to Machine Learning and Data Mining (Herlau, Schmidt, Mørup (2018), p. 15). Data analysis requires five steps to be done: Data understanding, data preparation, data modelling, evaluation and results.

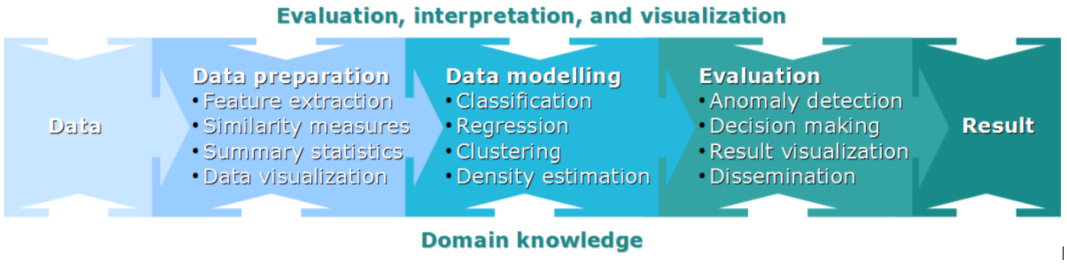


Figure 1: Common framework

This paper will show how step 1 and 2 should be performed so that the data is correctly prepared for data modelling. Data of US monthly wages is used to demonstrate this.
The first step is to get an understanding of the data. It is important that the analyst knows what each feature mean, thus being able to extract only those that are relevant for the analysis. This is done in the first part of this paper.
Secondly, the data is prepared for the data modelling. The data may hold missing values, outliers or noise which can effect the results. These issues should be handled before the modelling is performed. Part two and three focuses on these issues.
The actual data modelling will not be performed in this paper. Later projects will cover this, with the focus on step 3-5 from the common framework.
Lastly, part 4 will discuss how applicable the data is for future modelling.

# 1 Description of dataset

The dataset chosen for this project was found using this link suggested in the course material provided on Campusnet. Among the excel files we decided to work on the dataset "wage2.xls". The source of this dataset is the book "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials" written by M. Blackburn and D. Neumark in 1992. The sample studied in this report gathers 935 observations from the 1980's provided by the professor Neumark. This dataset provides the monthly wages of people depending on the following features:

- Monthly earnings
- Average weekly hours
- IQ score
- Knowledge of world work score
- Years of education
- Years of work experience
- Years with current employer
- Age in years
- =1 if married
- =1 if black
- =1 if live in south
- =1 if live in SMSA
- Number of siblings
- Birth order
- Mother's education
- Father's education
- Natural log of wage

Barry Murphy, of the University of Portsmouth in the UK, has pointed out that for several observations, the values for `exper` (years of work experience) and `tenure` (years with current employer) are in logical conflict. In particular, for some workers the number of years with current employer (`tenure`) is greater than the overall work experience (`exper`). At least some of these conflicts are due to the definition of `exper` as "potential" work experience, but probably not all. Nevertheless, we are using the dataset as it was supplied to us.
We believe that the dataset have potential for an interesting prediction analysis. Both in terms of the social-scientific issues it addresses, but also in terms of how the features relate. There may be unknown correlations between some features that can highlight unseen problems in the debate of racial equality.
The prediction score of later analysis is expected to be fairly good, since there is a sufficient amount of observations. This will only increase the relevance of the analysis.

# 2 Data Attributes

The original dataset contained a lot of attributes that, for one reason or another, weren't particularly interesting to us. So for this data visualization we have chosen to look at the following attributes:

- `wage`
- `hours`
- `IQ`
- `educ`
- `exper`
- `tenure`
- `age`
- `black`

## 2.1 Attribute Types

Our dataset contains the following attributes:

| Attribute Name | Description | Discrete/Continuous | Type |
|---|---|---|---|
| `wage` | monthly earnings | Continuous | Ratio |
| `hours` | average weekly hours | Continuous | Ratio |
| `IQ` | IQ score | Discrete | Interval |
| `educ` | years of education | Discrete | Ratio |
| `exper` | years of work experience | Discrete | Ratio |
| `tenure` | years with current employer | Discrete | Ratio |
| `age` | age in years | Discrete | Ratio |
| `black` | =1 if black | Discrete | Nominal |

Table 1: Data Attributes and Their Attribute types

Some values are simple to identify, `black` is clearly a binary attribute, which also makes it discrete. `wage`, being the monthly earnings is continuous, in that there are infinitely many real number values, by similar logic `hours` is also continuous, because there can be infinitely many values of average weekly hours. `IQ`, `educ`, `exper`, `tenure`, and `age` are discrete because there is a finite set of possible values for these attributes, most of them are also ratio excluding `IQ`. `IQ` is an interval value because an IQ score of 0 does not have any meaning, as well as IQ score not possessing any multiplicative properties.

## 2.2 Issues

We are so fortunate in that the attributes we are analyzing do not have any missing values. The dataset also has attributes for the mother's and father's level of education, wherein there are some missing values. Hypothetically, if our analysis used these attributes we could set the missing values to NaN. But, being that this is an education level, a lack of a value could be interpreted as the mother or father not having an education.

## 2.3 Summary Statistics

We calculate summary statistics with the following equations:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\text{median}(x) = \frac{1}{2}(x_R + x_{R+1})$$

$$\text{var}(x) = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2$$

$$\text{std}(x) = \sqrt{\text{var}(x)}$$

| **Summary Statistic** | `hours` | `IQ` | `educ` | `exper` | `tenure` | `age` | `black` |
|---|---|---|---|---|---|---|---|
| $\bar{x}$ | 43.93 | 101.28 | 13.47 | 11.56 | 7.23 | 33.08 | 0.13 |
| median(x) | 40 | 102 | 12 | 11 | 7 | 33 | 0 |
| range(x) | 20-80 | 50-145 | 9-18 | 1-23 | 0-22 | 28-38 | 0-1 |
| std(x) | 7.22 | 15.05 | 2.20 | 4.37 | 5.08 | 3.11 | 0.33 |
| var(x) | 52.19 | 226.58 | 4.83 | 19.14 | 25.76 | 9.66 | 0.11 |

Table 2: Summary Statistics for Attributes

As we can see from the difference in variances, it will be necessary to later do some form of standardizing to the data, such that some attributes don't count for too much of the data's variance.

| Summary Statistic | `hours` | `IQ` | `educ` | `exper` | `tenure` | `age` | `black` |
|---|---|---|---|---|---|---|---|
| `hours` | 52.19 | 8.03 | 1.44 | -1.96 | -2.04 | 0.56 | -0.26 |
| `IQ` | 8.03 | 226.58 | 17.05 | -14.81 | 3.22 | -2.05 | -1.95 |
| `educ` | 1.44 | 17.05 | 4.83 | -4.38 | -0.40 | -0.08 | -0.13 |
| `exper` | -1.96 | -14.81 | -4.38 | 19.14 | 5.41 | 6.73 | 0.08 |
| `tenure` | -2.04 | 3.22 | -0.40 | 5.41 | 25.76 | 4.27 | -0.13 |
| `age` | 0.56 | -2.05 | -0.08 | 6.73 | 4.27 | 9.66 | -0.04 |
| `black` | -0.26 | -1.95 | -0.13 | 0.08 | -0.13 | -0.04 | 0.11 |

Table 3: Covariance

Covariance was calculated according to the following equation:

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

Correlation can then also be found:

| **Summary Statistic** | `hours` | `IQ` | `educ` | `exper` | `tenure` | `age` | `black` |
|---|---|---|---|---|---|---|---|
| `hours` | 1.0 | 0.074 | 0.091 | -0.062 | 0.056 | 0.025 | -0.11 |
| `IQ` | 0.074 | 1.0 | 0.52 | -0.22 | 0.042 | -0.044 | -0.39 |
| `educ` | 0.091 | 0.52 | 1.0 | -0.46 | -0.036 | -0.012 | -0.18 |
| `exper` | -0.062 | -0.22 | -0.46 | 1.0 | 0.24 | 0.50 | 0.056 |
| `tenure` | -0.056 | 0.043 | -0.036 | 0.24 | 1.0 | 0.27 | -0.078 |
| `age` | 0.025 | -0.044 | -0.012 | 0.50 | 0.27 | 1.0 | -0.036 |
| `black` | -0.11 | -0.39 | -0.18 | 0.056 | -0.078 | -0.036 | 1.0 |

Table 4: Correlation

$$\text{corr}(x, y) = \frac{\text{cov}(x, y)}{\text{std}(x)\text{std}(y)}$$

As expected, one of the two most correlated attributes were age and experience, which is obvious because older people have had a lot more time and opportunity to gain work experience. A lot of attributes also had a relatively high negative correlation with black. Since the data was taken during the 1980s, this isn't surprising given the unequal nature of education between black and white Americans during that time period.

# 3 Principal Component Analysis (PCA)

In this part we are going to focus on the Principal Component Analysis. As we have a high-dimensional dataset, we want to find a lower-dimensional representation of these data. The PCA algorithm will help us determine the main dimensions to project the data and visualize it better. This procedure will determine the directions that maximize the variance of the projected data. The dimension of our matrix X is 935x7. We can define the PCA algorithm on matrix X where the dataset is projected onto the first n components as follows:

- Subtract the mean value in order to set the axis origin at the center of the data : $\bar{x_i} = x_i - \frac{1}{N}\sum_1^N x_i$
- Compute the SVD: $U\Sigma V^T = \bar{X}$ with U and V being orthonormal matrices and $\Sigma$ is a diagonal matrix gathering the sorted singular values of $X$ $\sigma_i$. In fact each $v_i$ vector in $V$ is an eigenvector of $\bar{X}^T\bar{X}$ with associated eigenvalue $\sigma_i^2$ and the eigenvectors are sorted according to their eigenvalues.
- The $n$ first principal components are $v_1, ..., v_n$ and coordinates of observation i when projected onto the subspace spanned by the first n principal components are

  $$b_i^T = \bar{x_i}^T V_n \text{ or alternatively } B = \bar{X}V_n$$

  where $V_n = [v_1, ..., v_n]$.

First, let's take a look at the data. The mean value of each attribute is quite different and we can observe the same problem with the variance.

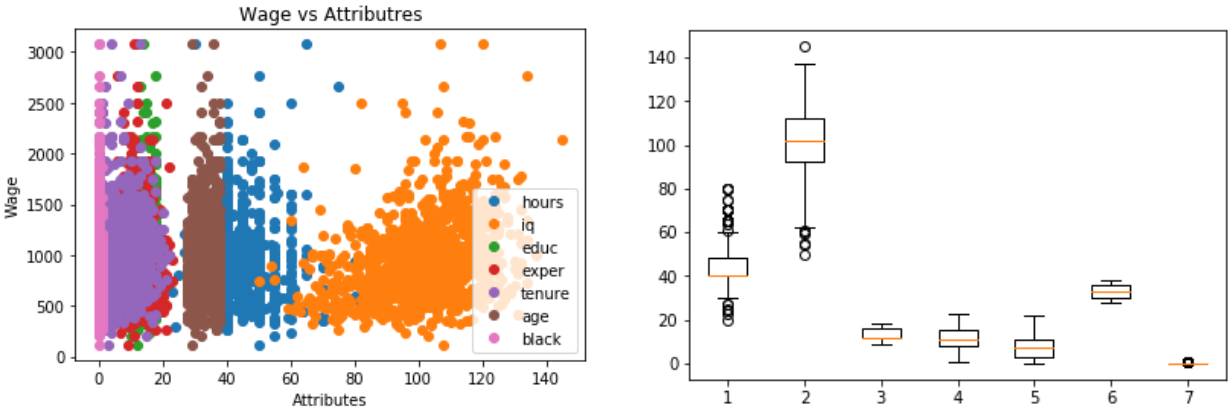


Figure 2: Data

Since we want the projection to be invariant under addition of a constant, we first subtract the mean from each $x_i$.

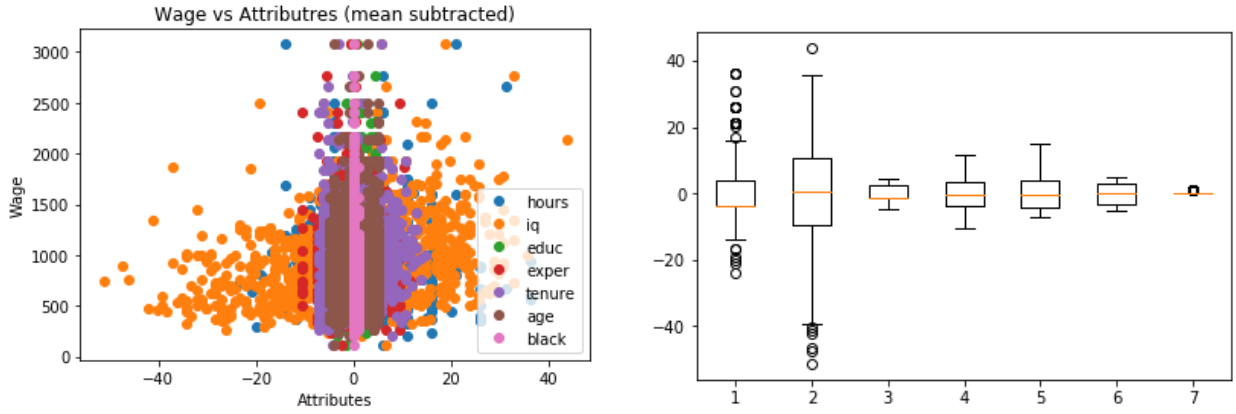


Figure 3: Data when the mean is subtracted

Now that the mean values have been subtracted, the variances are more clear, as seen in Figure 2. However, there still appear to be significantly differences in variance, which can be confirmed by Table 2. To accommodate this, the data can be further standardized, using the formula:

$$x_k^* = \frac{x_k - \bar{x_k}}{\text{std}(x_k)}$$

The formula takes the attributes corrected for their mean values, and divides them by their respective standard deviation. This will scale the variances such that the data can be compared more easily. The result is shown in Figure 4.

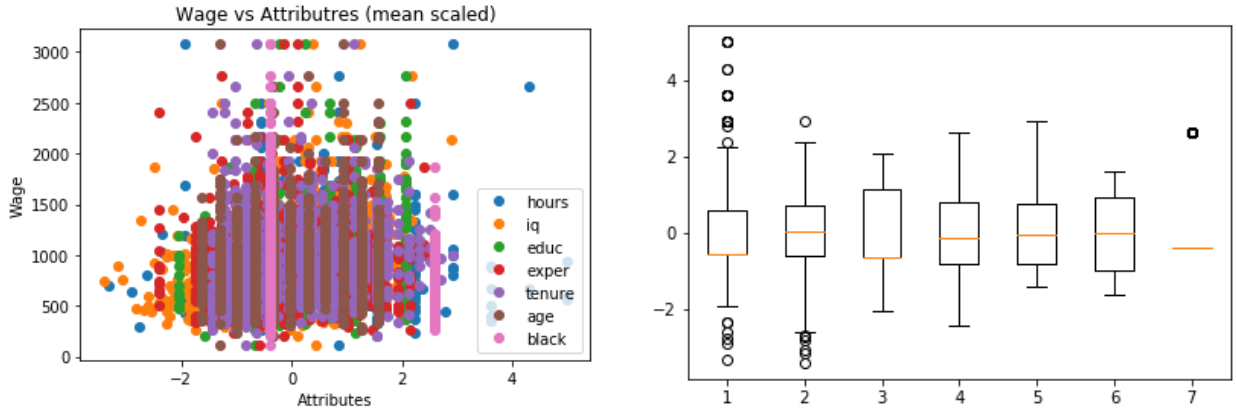


Figure 4: Data standardized

First, we are going to apply the PCA algorithm to the non-standardized data:

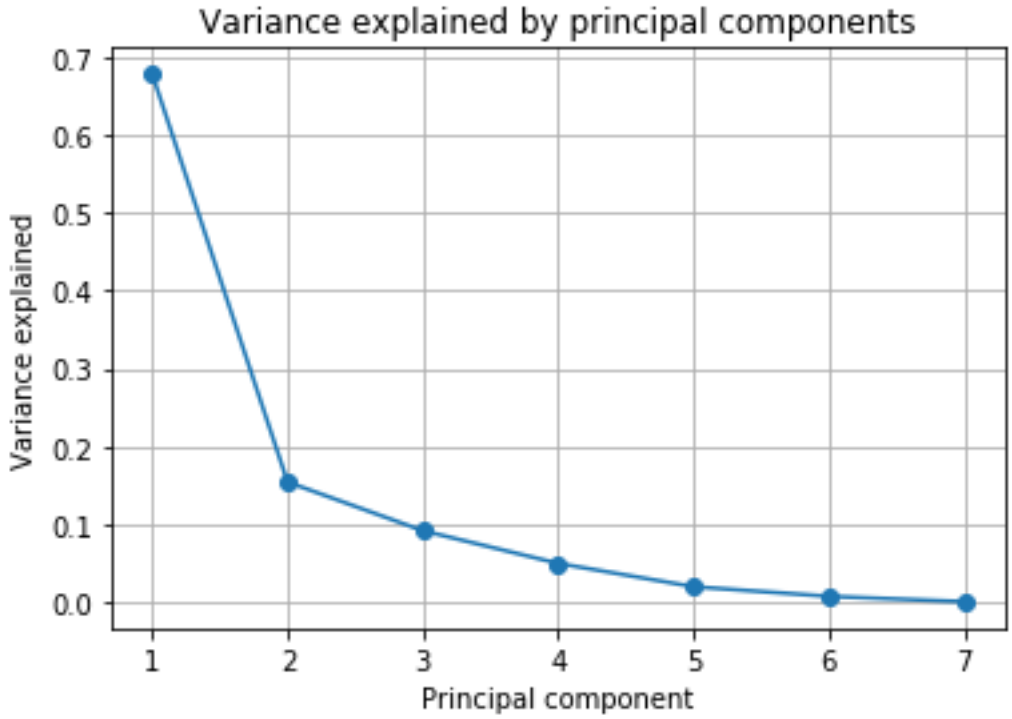


Figure 5: PCA results

From the figure, we can easily see that the two first directions account for most of the variance. The PCA enables a reduction of the data's dimensions from 8 to 2. However, the third component account for exactly 10% of the variance therefore, if we want to make sure that more than 90% is covered, we have to keep PC1, PC2 and PC3.

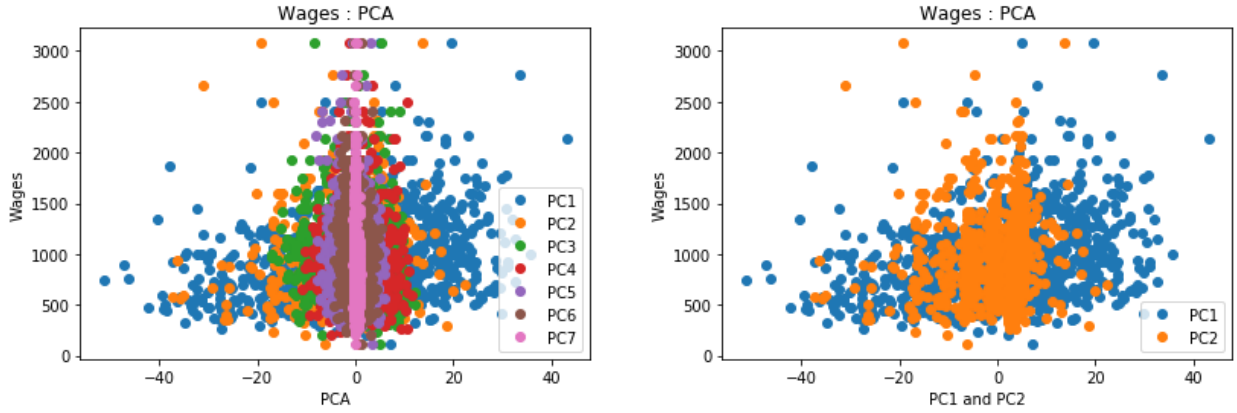


Figure 6: PCA results - projection on the 2 first directions

We can check if the first components have a larger variance than the last.

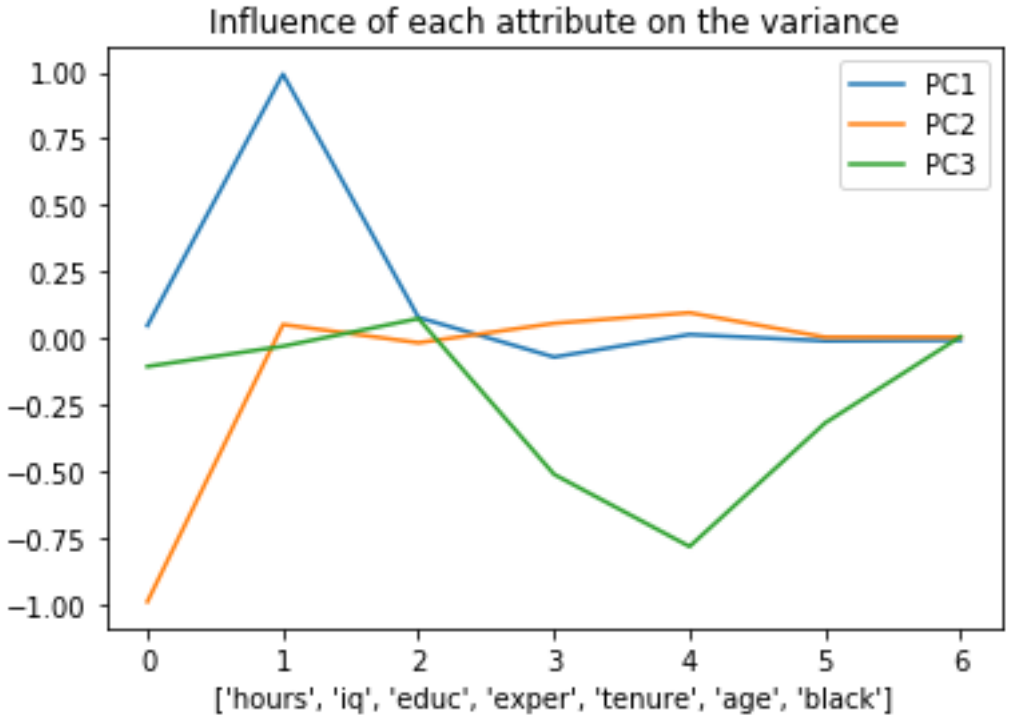


Figure 7: Influence of attributes on principal components

For the projection onto the two main directions PC1 and PC2, attributes 3, 4, 5, 6 and 7 (respectively 'educ', 'exper', 'tenure', 'age' and 'black') have almost no influence on the variance. PC1 is almost completely determined by the attribute 'iq' and PC2 by the attribute 'hours'. To see if we are right, let us project the data onto 'iq' and 'hours' to see if we obtain the same results as before when we projected the data onto the two first principal components.

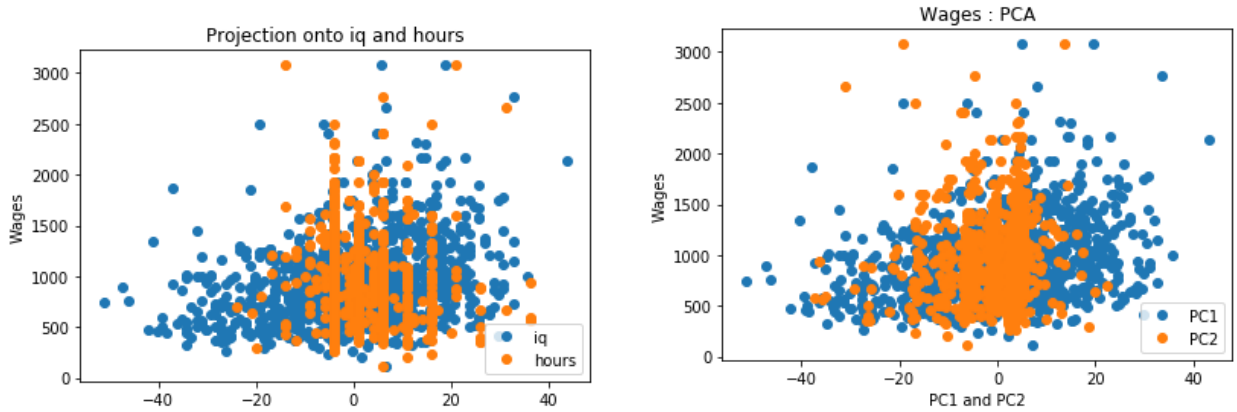


Figure 8: Comparison PC1-'iq' and PC2-'hours'

The influence on the variance is almost completely determined by the attributes 'iq' and 'hours' and also by the attribute 'tenure' as we can see in the figure which represent the influence of each attribute on the variance (through the three first principal components). The data seems to be of a good quality (no bias or illogical data), and therefore hopefully, this set of data will help us predict wages according to these main attributes.

Let's see if our results are the same with the data standardized:

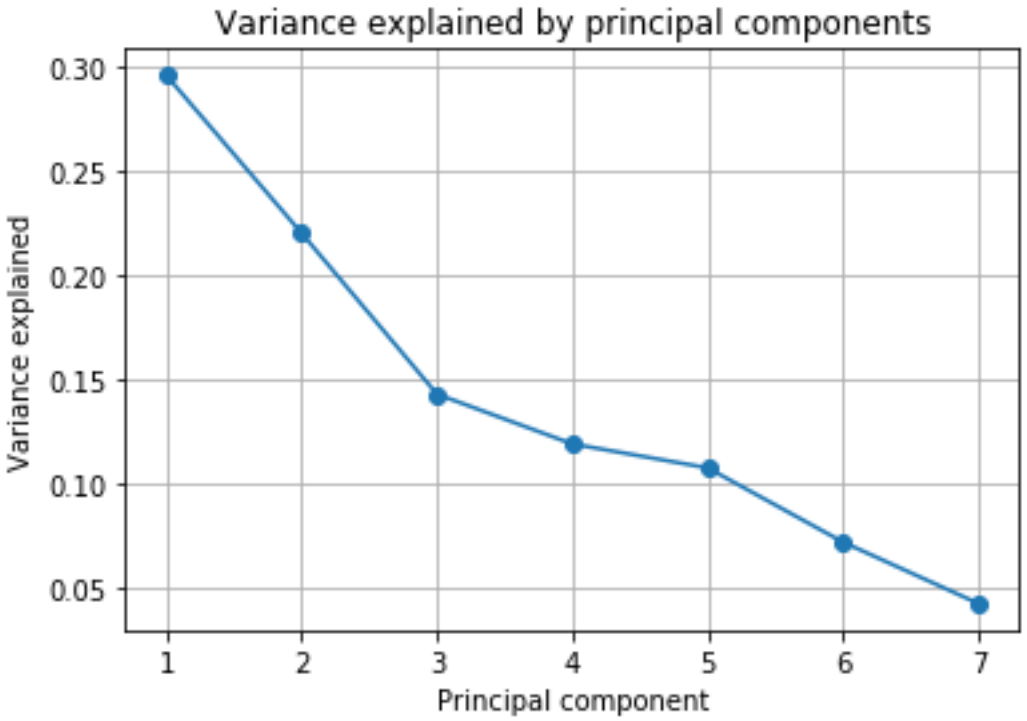


Figure 9: PCA results

The results are quite different, the influence of the variance is not dominated anymore by the two or three first components and each principal components seem to have their significance.

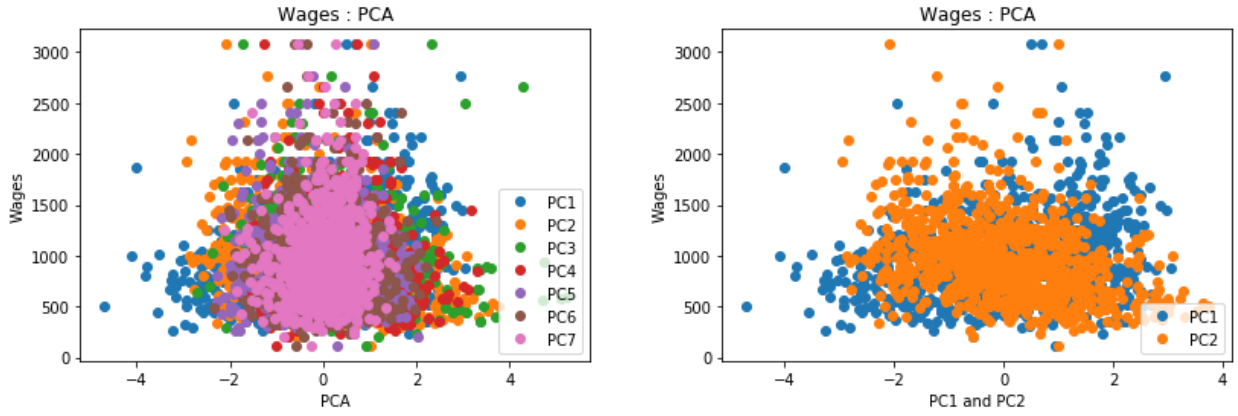


Figure 10: PCA results - projection on the 2 first directions

These figures show that even if the first principal components account for more variance than the last ones, they represent much less variance than before the standardization.

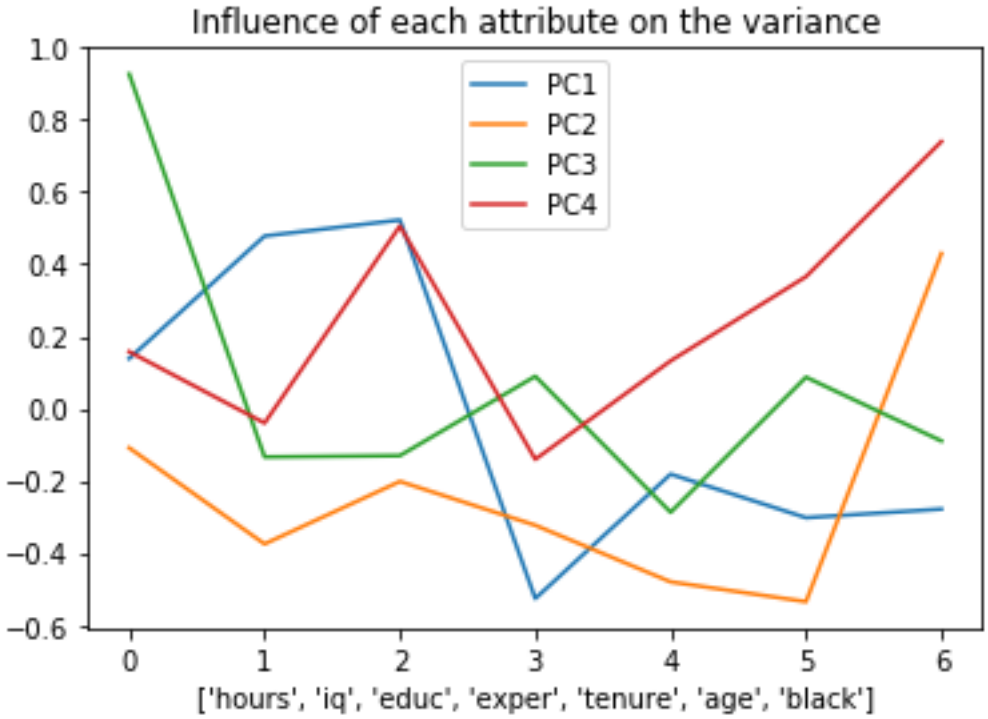


Figure 11: Influence of attributes on principal components, after standardization with mean and std(x)

With the standardized data, we can see that in general, all the attributes have an inuence on the variance and are relevant to carry for the rest of the study.

# 4 Discussion

As mentioned in the introduction, the purpose of this project is to get an understanding of the data, and prepare the data for later modeling. So have these goals been achieved? Will the data be suitable for more advanced data modeling?

Starting at the data understanding, we learned that the data contains both discrete and continuous attributes with ratio, interval and nominal types.

The summary statistics revealed that the data sample is largely representative of the US population, for example the US is about 13% black as is our data. This information will turn out to be of great importance for later projects, since this ultimately defines the usefulness of the prediction functions. More interestingly, the correlation table shows some interesting results. For example, the attribute pairs with the highest correlations were `IQ` and `educ` as well as `age` and `exper`, which is expected. Another interesting observation is that our `black` binary attribute was relatively highly negatively correlated with `hours`, `IQ`, `educ`. Likely as a result of the fact that this data is from the 1980s, a time where education wasn't equal between black and white Americans.

The principal component analysis revealed that nearly 90% of the variance could be explained by two principal components, thus enabling us reduce the dimension of our data significantly. The attributes `hours` and `IQ` turned out be very influencing on the variance. We further did a PCA analysis on a standardized, following the formula: $X = \frac{x-\bar{x}}{std(x)}$, version of the of the data to ensure that there were no scaling issues due to the binary attribute. This resulted in much less dominant principal components, where five components would account for nearly 90% of the variance. Here redimensioning the data as drastically as before wouldn't make any sense.

Since PCAs are very sensitive to scaling of the data, scaled data should only be used when there is a strong need for it. Our primary concern was the binary attribute dominating the variance plot, but it didn't seem to be that way. For this reason the first PCA, based on $X = x - \bar{x}$, will be the one used for later modelling.

The data did not contain any outliers that we would be able to remove, e.g. a negative `age`, so we have not had to do much in terms of data preparation.

In conclusion, the principal component analysis and summary statistics have shown to be two very useful methods when the analysts seeks a better understanding of the data. The data is expected to perform well in future machine learning models since the data is of good quality, the sample size is relatively big, and because of the diversity of the data types.