Statistical Analysis of British Workforce Using Machine Learning - Part 2

Filippo Lisanti

The Data

Data for both parts comes from Wave 11 of the British Household Panel Survey. This is a multi-year panel of British households, and the data from Wave 11 were collected in the Autumn of 2001. The dataset here is a subset of variables relevant to income and job satisfaction. It is in Parquet format.

The data consists of households (identified by hid) and individuals within that household (identified by pid).

The relevant variables are:

- real_hourly_wage: The inflation-adjusted wage (in 1991 British pounds) of the respondent at his or her main job.
- real_hh_non_labour_income inflation-adjusted income from non-wage sources (such as financial assets).
- age: The respondent's age
- job_tenure: The number of years that the respondent has been working at his or her main job.
- higher_degree, first_degree, hnd_hnc_teaching, a_level, o_level, cse: These are all dummies for the highest level of education completed, listed here in reverse order of the qualification level.
- separate_before_next is a dummy that is 1 if the employee separated from their employer before the next BHPS wave (one year later), or zero otherwise.
- pay_satisfaction is the employee's self-reported satisfaction with their level of pay (1-7, with 7 being best).
- job_security_satisfaction is the employee's self-reported satisfaction with their job security (1-7, with 7 being best).
- work_satisfaction is the employee's self-reported satisfaction with their daily work (1-7, with 7 being best).
- here is an index variable that was created by the importation process and is of no consequence

Loading the data into R and dropping the same observations as Part 1.

```
file_path <- "/Users/filippolisanti/Desktop/bhps-wave-11.pqt"
data <- arrow::read_parquet(file_path)
data <- subset(data, age <= 65)
```

Structuring the education level variable (e.g., 1=CSE, 2=O level, etc.)

```
library(dplyr)
data <- data %>%
  mutate(
    education_level = case_when(
      higher_degree == 1 ~ 6,
      first_degree == 1 ~ 5,
      hnd_hnc_teaching == 1 ~ 4,
      a_level == 1 ~ 3,
      o_level == 1 ~ 2,
      cse == 1 ~ 1,
      TRUE ~ 0
    )
  )
print(data)
```

```
## # A tibble: 18,793 × 18
##         hid   pno real_hourly_wage real_hh_non_labour_income   age
job_tenure
##       <int> <int>            <dbl>                     <dbl> <dbl>
<dbl>
##  1 10000054     1           NA                       31452.    63      NA
##  2 10000119     1           NA                          NA     45      NA
##  3 10246339     1           NA                       12368.    41      NA
##  4 10000119     2           NA                          NA     19      NA
##  5 10000151     1           NA                       17890.    58      NA
##  6 10000151     2           NA                       18026.    21      NA
##  7 10231617     1           NA                          NA     37      NA
##  8 10273948     1            6.55                        0     39
0.5
##  9 10146865     1            9.90                     1659.    38
3.42
## 10 10146865     2           NA                          NA     12      NA
## # ℹ 18,783 more rows
## # ℹ 12 more variables: higher_degree <dbl>, first_degree <dbl>,
## #   hnd_hnc_teaching <dbl>, a_level <dbl>, o_level <dbl>, cse <dbl>,
## #   separate_before_next <dbl>, pay_satisfaction <dbl>,
## #   job_security_satisfaction <dbl>, work_satisfaction <dbl>,
## #   `__index_level_0__` <int>, education_level <dbl>
```

Scaling data in preparation for cluster analysis and factor analysis

```
data_clean <- na.omit(data)

data_for_analysis <- data_clean %>%
  select(real_hourly_wage, real_hh_non_labour_income, age, job_tenure,
         pay_satisfaction, job_security_satisfaction, work_satisfaction,
         education_level)

scaled_data <- scale(data_for_analysis)
```

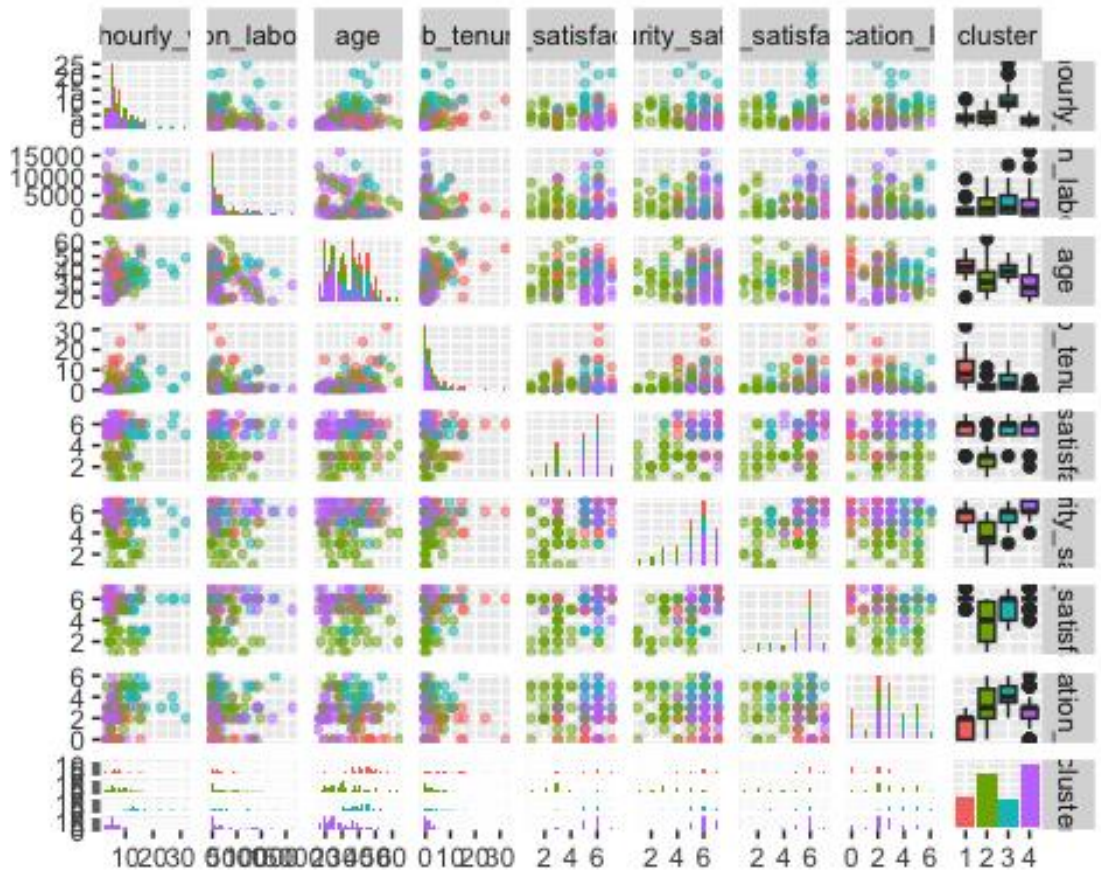
Using a silhouette score to figure out the appropriate number of clusters.

```
max_clusters <- 10
silhouette_scores <- numeric(max_clusters - 1)

for (k in 2:max_clusters) {
  km_res <- kmeans(scaled_data, centers = k, nstart = 25)
  silhouette_avg <- mean(silhouette(km_res$cluster, dist(scaled_data))[,
"sil_width"])
  silhouette_scores[k - 1] <- silhouette_avg
}

optimal_clusters <- which.max(silhouette_scores)
optimal_clusters

## [1] 4
```

Scatter plot matrix displaying relationships among all variables

```
optimal_clusters <- 4
km_res <- kmeans(scaled_data, centers = optimal_clusters, nstart = 25)
data_clean$cluster <- km_res$cluster
set.seed(42)
data_sample <- sample_n(data_clean, size = floor(0.05 * nrow(data_clean)))

data_sample_selected_vars <- data_sample %>%
  select(real_hourly_wage, real_hh_non_labour_income, age, job_tenure,
         pay_satisfaction, job_security_satisfaction, work_satisfaction,
         education_level, cluster) %>%
  mutate(cluster = as.factor(cluster))

ggpairs_plot <- ggpairs(data_sample_selected_vars, aes(color = cluster),
                        upper = list(continuous = wrap("points", size = 1,
alpha = 0.5)),
                        lower = list(continuous = wrap("points", size = 1,
alpha = 0.5)),
                        diag = list(continuous = wrap("barDiag")))
```

```
print(ggpairs_plot)
```




```
ggsave("scatterplot_matrix.png", plot = ggpairs_plot, width = 20, height =
20)
```

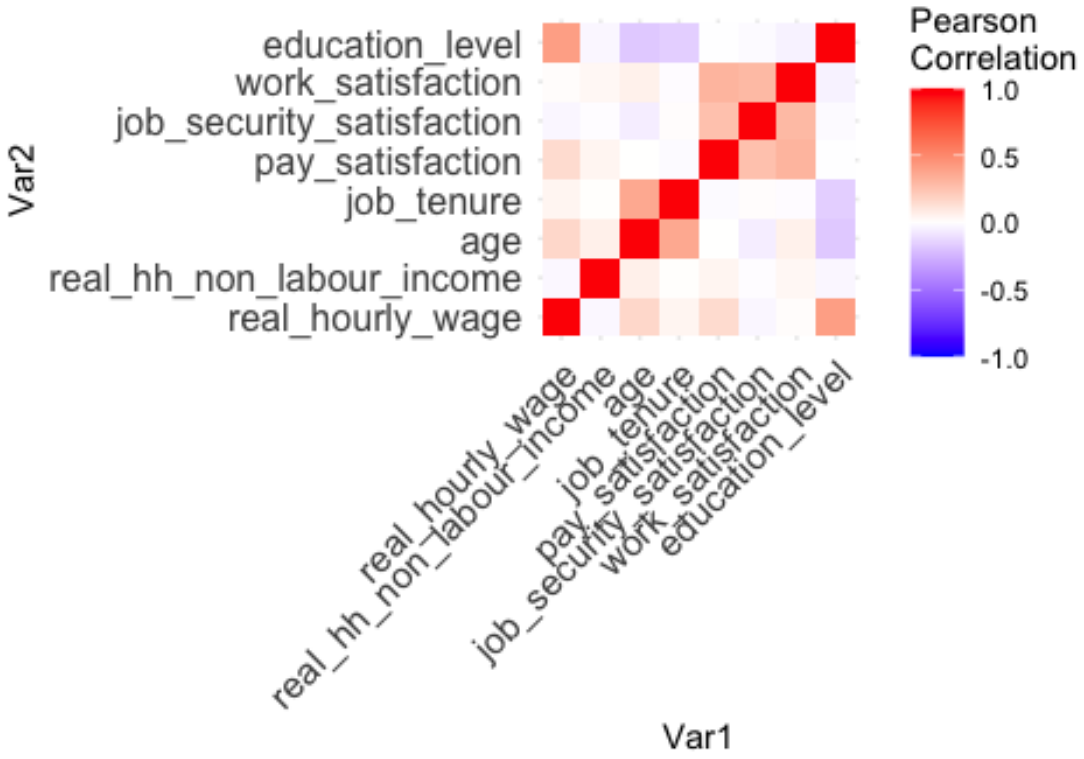
Correlation matrix of the variables

```
scaled_data <- data.frame(scaled_data)
correlation_matrix <- cor(scaled_data, use = "complete.obs")
print(correlation_matrix)
```

```
##                             real_hourly_wage real_hh_non_labour_income
## real_hourly_wage                  1.00000000               -0.026525270
## real_hh_non_labour_income        -0.02652527                1.000000000
## age                               0.16789500                0.064669041
## job_tenure                        0.04066311                0.005720158
## pay_satisfaction                  0.15175583                0.043843555
## job_security_satisfaction        -0.02934420               -0.008971205
## work_satisfaction                 0.01284147                0.033066866
## education_level                   0.41550981               -0.030413803
##                                        age   job_tenure pay_satisfaction
## real_hourly_wage             0.16789500  0.040663115      0.151755830
```

```
## real_hh_non_labour_income  0.06466904  0.005720158      0.043843555
## age                        1.00000000  0.370852789      0.002106060
## job_tenure                 0.37085279  1.000000000     -0.018143023
## pay_satisfaction           0.00210606 -0.018143023      1.000000000
## job_security_satisfaction -0.05716743  0.009086143      0.265855383
## work_satisfaction          0.05943843 -0.011346538      0.316327522
## education_level           -0.18310261 -0.159764918     -0.005937772
##                           job_security_satisfaction work_satisfaction
## real_hourly_wage                       -0.029344196        0.01284147
## real_hh_non_labour_income              -0.008971205        0.03306687
## age                                    -0.057167428        0.05943843
## job_tenure                              0.009086143       -0.01134654
## pay_satisfaction                        0.265855383        0.31632752
## job_security_satisfaction               1.000000000        0.29732196
## work_satisfaction                       0.297321957        1.00000000
## education_level                        -0.017671175       -0.04198162
##                           education_level
## real_hourly_wage              0.415509810
## real_hh_non_labour_income    -0.030413803
## age                          -0.183102606
## job_tenure                   -0.159764918
## pay_satisfaction             -0.005937772
## job_security_satisfaction    -0.017671175
## work_satisfaction            -0.041981616
## education_level               1.000000000
```

```
correlation_melted <- melt(correlation_matrix)
ggplot(data = correlation_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust =
1),
        axis.text.y = element_text(size = 12)) +
  coord_fixed()
```



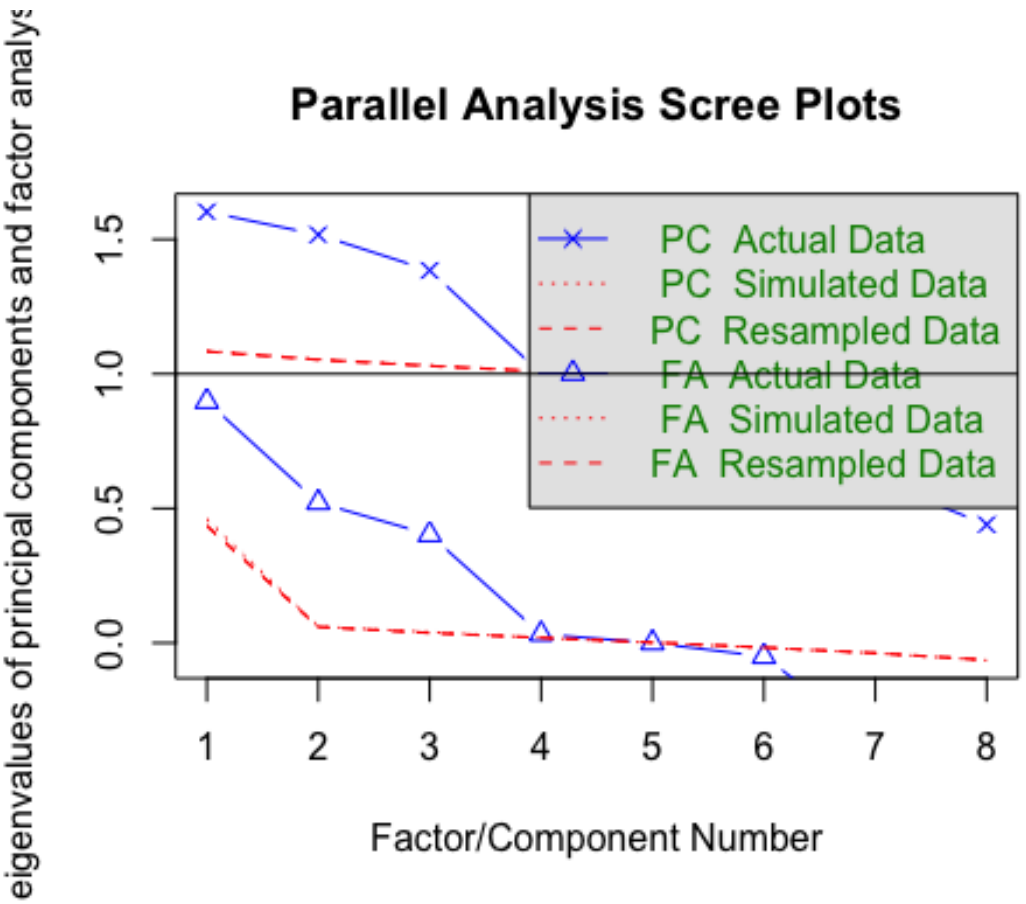
Testing the statistical significance of the correlations

```
correlation_tests <- combn(names(scaled_data), 2, FUN = function(vars) {
  cor.test(scaled_data[[vars[1]]], scaled_data[[vars[2]]], method =
"pearson")
}, simplify = FALSE)

p_values <- sapply(correlation_tests, function(test) test$p.value)
favorite_significance_level <- 0.05
significant_correlations <- p_values < favorite_significance_level
fraction_significant <- sum(significant_correlations) / length(p_values)

answer <- sprintf("The fraction of the correlations that are
                  statistically significant is %f", fraction_significant)
```

The fraction of the correlations that are statistically significant is 0.5357

Using a parallel analysis to determine the number of principal components that should be used to find relationships among your variables

```
pa <- fa.parallel(scaled_data, fa = "both", n.iter = 100)
```




```
## Parallel analysis suggests that the number of factors =  0  and the number
of components =  3

n_factors <- pa$nfact
print(paste("The number of components suggested by parallel analysis is 3
(elbow method)"))

## [1] "The number of components suggested by parallel analysis is 3 (elbow
method)"
```

Varimax factor rotation

```
factor_analysis_result <- fa(scaled_data, nfactors = 3, rotate = "varimax")
loadings <- factor_analysis_result$loadings
dimnames(loadings)[[2]] <- c("Experience", "Career_Stab", "Satisfaction")
print(loadings)
```

```
##
## Loadings:
##                            Experience Career_Stab Satisfaction
## real_hourly_wage            0.986      0.147
## real_hh_non_labour_income
## age                                    0.790
## job_tenure                             0.471
## pay_satisfaction            0.101                  0.549
## job_security_satisfaction                          0.497
## work_satisfaction                                  0.588
## education_level             0.462     -0.276
##
##                 Experience Career_Stab Satisfaction
## SS loadings          1.205       0.955        0.903
## Proportion Var       0.151       0.119        0.113
## Cumulative Var       0.151       0.270        0.383
```

I initially considered four factors, as suggested by the scree plot, but opted for three after observing that the loadings for the fourth factor were consistently below 0.3. This decision ensures that only the most statistically significant results are included in my analysis.

Experience = Factor 1 (Experience): This factor is characterized mainly by a person's real hourly wage (0.986). Education level also has a notable loading of 0.462. Pay satisfaction also plays a small role at 0.101.

Career_Stability = Factor 2 (Career_Stability): This factor has high loadings on age (0.790) and job tenure (0.471). Real hourly wage plays a small role with a loading of 0.147. Education level interestingly has a negative loading of -0.276 according to my varimax factor rotation.

Satisfaction = Factor 3 (Satisfaction): This factor has balanced high loadings on pay satisfaction (0.549), job security satisfaction (0.497) and work satisfaction (0.588) It is evident that work satisfaction plays the largest role on one's overall satisfaction.

Oblique factor rotation

```
factor_analysis_oblique <- fa(scaled_data, nfactors = 3, rotate = "oblimin")

## Loading required namespace: GPArotation

colnames(factor_analysis_oblique$loadings) <- c("Experience", "Career_Stab",
"Satisfaction")
print(factor_analysis_oblique$loadings)

##
## Loadings:
##                            Experience Career_Stab Satisfaction
## real_hourly_wage            0.993
## real_hh_non_labour_income
## age                                    0.786
```

```
## job_tenure                              0.477
## pay_satisfaction           0.103                  0.544
## job_security_satisfaction                         0.500
## work_satisfaction                                 0.591
## education_level            0.468     -0.323
##
##                Experience Career_Stab Satisfaction
## SS loadings         1.224       0.960        0.902
## Proportion Var      0.153       0.120        0.113
## Cumulative Var      0.153       0.273        0.386
```

Experience = Factor 1 (Experience): This factor is still characterized mainly by a person's real hourly wage (0.993). Education also has a notable loading of 0.468 (slightly higher). Pay satisfaction also plays a small role at 0.103.

Career_Stability = Factor 2 (Career_Stability): This factor still has high loadings on age (0.790) and job tenure (0.477). Real hourly wage plays no role. Education level still has a negative loading of -0.323.

Satisfaction = Factor 3 (Satisfaction): This factor still has balanced high loadings on pay satisfaction (0.544), job security satisfaction (0.500) and work satisfaction (0.591).

StatsInfo = "StatsInfo: SS loadings, Proportion Var and Cumulative Var remain almost identical to the results of the varimax factor rotation.