Maths and Statistics Foundations Project

Group N

06/10/2017

1. Introduction

Discrimination has long been a controversial topic. Forms of discrimination can range greatly but over the last few decades, one notable form of discrimination has been the phenomenon of an income gap that has existed in society (Blinder, 1973). The European Commission defines gender pay gap as "the difference between men's and women's pay, based on the average difference in gross hourly earnings of all employees" (European Commission, 2017). In recent years especially, addressing the gender related pay gap has been a policy matter of many countries namely those belonging to the EU (European Union, 2014).

The United Nations (2017), aims to remove all gender discrimination to ensure gender equality, and according to the European Commission (2017), gender discrimination is prohibited under European law. However, the Office for National Statistics (2017) states that female full-time employees in the UK earn on average 9.4 percent less than male full-time employees in 2016. This raises the question if gender pay discrimination currently exists in the UK, or if there are other drivers affecting the differences in salary levels.

Consequently, this report aims to identify if there currently exists gender pay discrimination in the UK whilst considering other factors which may affect the pay gap such as differences in where and how people in the UK tend to work.

These results would help to make suggestions regarding salary policies leading to a minimisation of the pay gap.

2. Theory

3. Methods

3.1 Data Description

To examine the above mentioned pay gap, data is applied from the Office of National Statistics *Quarterly Labour Force Survey* for the period of January to March 2017 (ONS, 2017). The dataset initially consisted of over 700 variables, which we have condensed to the 7 variables we found the most relevant.

In order to examine the relationship between salary and gender and explore the possibility of gender discrimination in the UK, we have included other variables which also may impact the pay gap. Hence, we define the dependent variable as *Salary* and the independent variables are defined as *sex (gender), age, industry sector, education qualifications, number of employees in the workplace and major occupation group,* respectively.

We have considered variables in our analysis such as ethnicity. It would be logical to include such variables as for instance, discrimination by ethnicity is a common occurrence and is still prevalent in present time and not completely eradicated. Ethnic pay gaps are resultant from either the the majority of the ethnic minority group going into more less well-paid careers or that they are merely generally paid less for doing the same kind of work (Brynin and Güveli, 2012).

However, we eventually dropped ethnicity as a variable as a large proportion of our dataset reflects white people which would greatly distort our analyses and study from the arising biasedness. For similar reasons,

we have also dropped variables that are less significant in our regression models or display multicollinearity such that it influences the other variables in our model to a large extent - namely region of work, marital status and religion. This shortens our dataset further to just 7 variables in our regression model.

We implemented a linear regression model for this dataset. We treated salary as a dependent variable, and xxxxxxxxxxx as predictors. As the participants of the survey obtain weekly, monthly and yearly salaries, we converted the salaries to be on an annual basis. This is done where weekly salaries are multiplied by 52 and monthly salaries are multiplied by 12. The salary data is thus in a range of amounts of earnings. In order to implement a linear regression model, we therefore calculated the mean of each range.

The majority of the variables are categorical in nature - except for the variable age. Therefore, we had to create many dummy variables depending on the level of each variable.

Several statistical tools in R were used in the analysis of the dataset. X and X libraries were used to help visualise the data and X and X used for y.

Our final, clean, dataset contains 570 observations and 7 variables. These are outlined below.

Variable	Description
`Salary`	Salary of respondent (249.5 - 48000.0); N.b. adjusted for inflation
`Age`	Age of respondent (0-99)
`Sex`	Gender of respondent
`Occupation`	Major occupation group of respondent
`Industry`	Industry sector in main job
`Education`	Highest qualification level
`NumEmployee`	Number of employees at workplace
`Religion`	Religion GB level
`MaritalStatus`	Marital status
`Ethnicity`	Ethnicity in GB
`Region of workplace`	Region of place of work

4. Analysis

4.1 Descriptive statistics

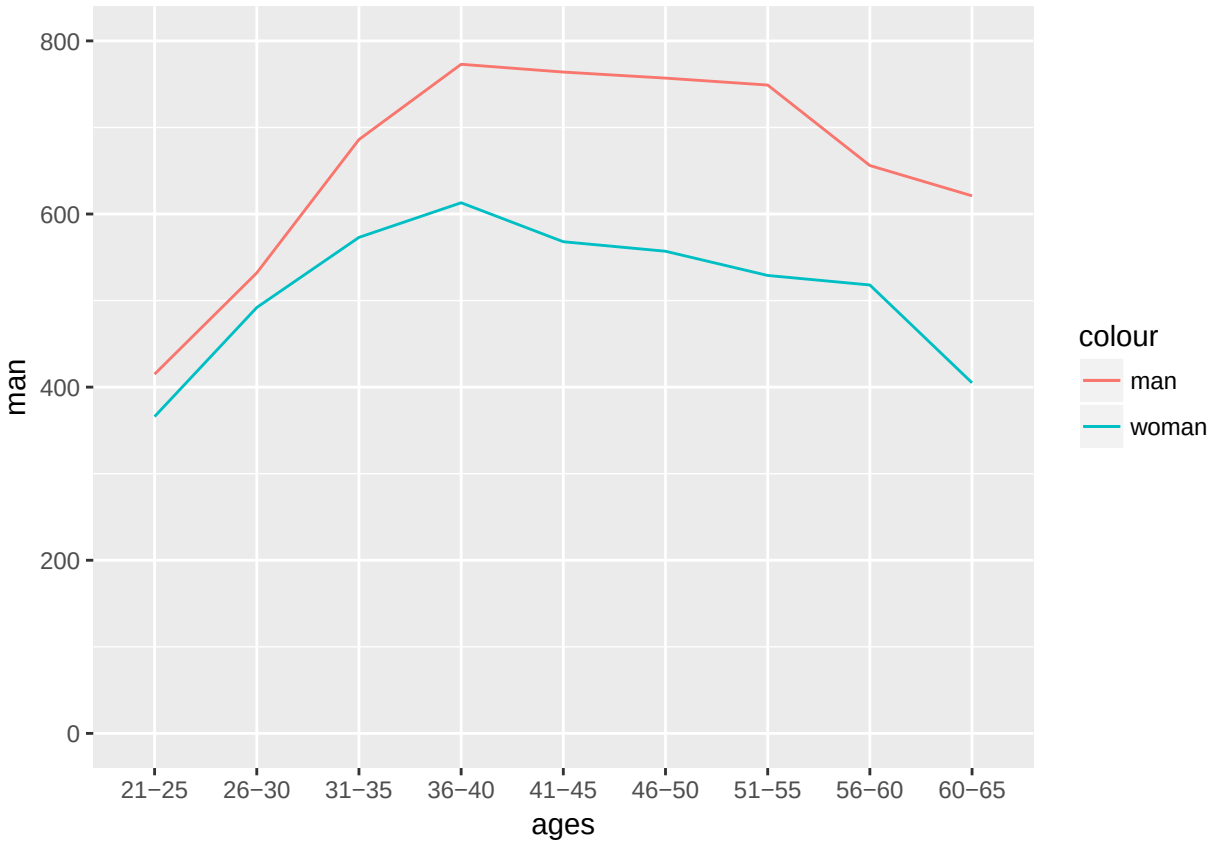
4.1.1 Univariate descriptive of variables

- central tendency (mean, mode, median)
- dispersion (range, variance, maximum, minimum, quartiles & IQR, std.)
- frequency distribution tables
- bar charts
- histograms

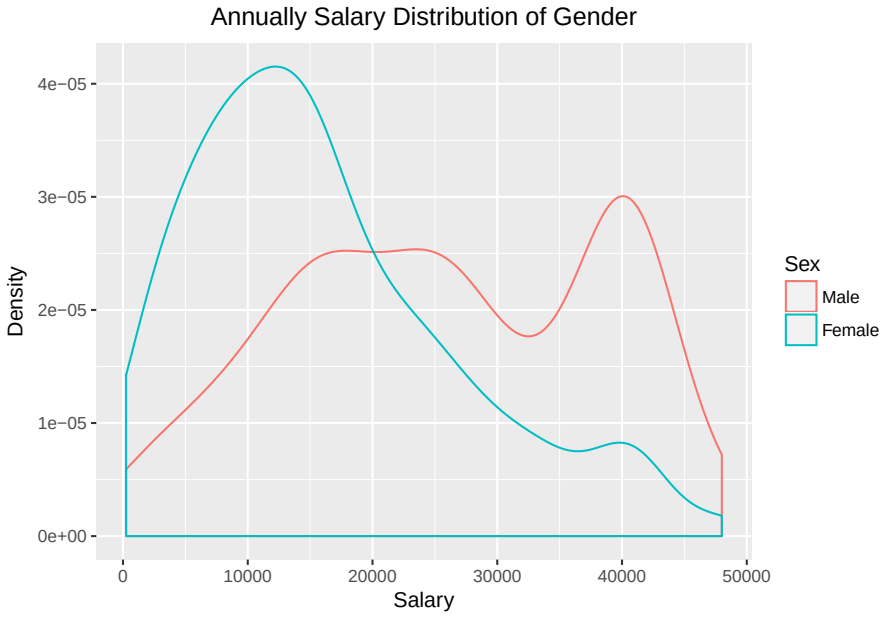
Sex

```
man <- c(415, 532, 686, 773, 764, 757, 749, 656, 621)
woman <- c(366, 492, 573, 613, 568, 557, 529, 518, 405)
ages <- c("21-25", "26-30", "31-35", "36-40", "41-45", "46-50", "51-55", "56-60", "60-65")
dtmw <- data.frame(ages, man, woman)

ggplot(dtmw) +geom_line(aes(x= ages, y=man,col='man'),group=1)+ geom_line(aes(x= ages, y=woman,col='woma
```

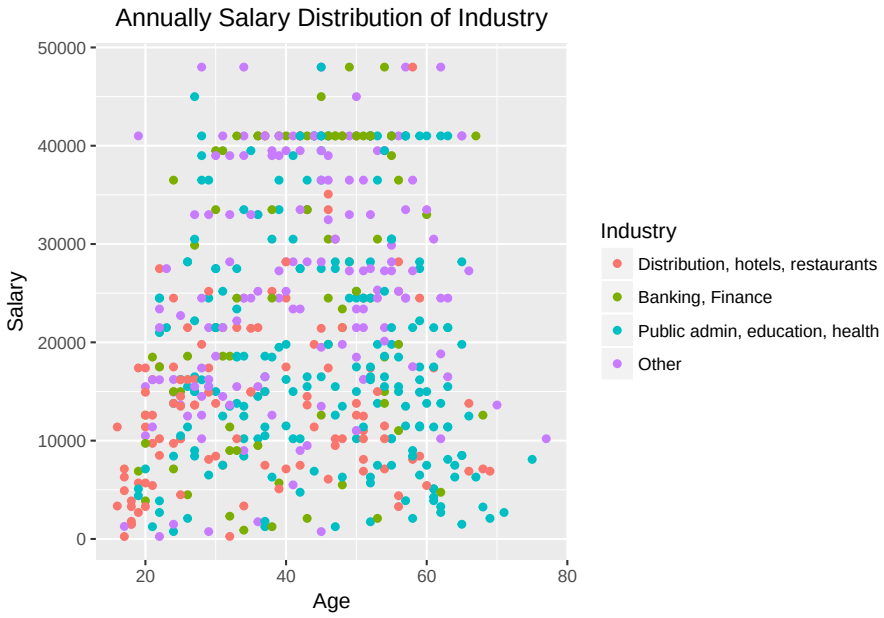


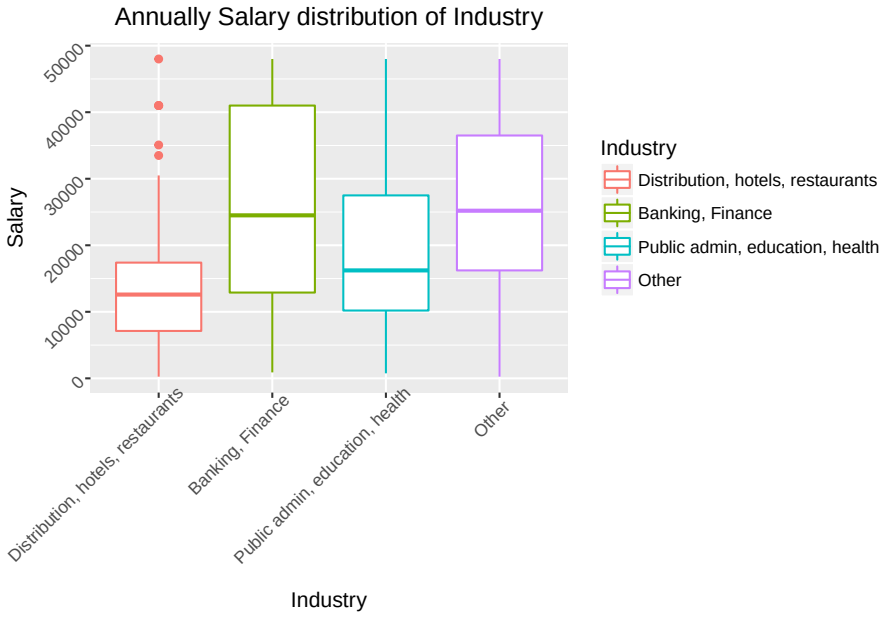

```
## Don't know how to automatically pick scale for object of type labelled. Defaulting to continuous.
```

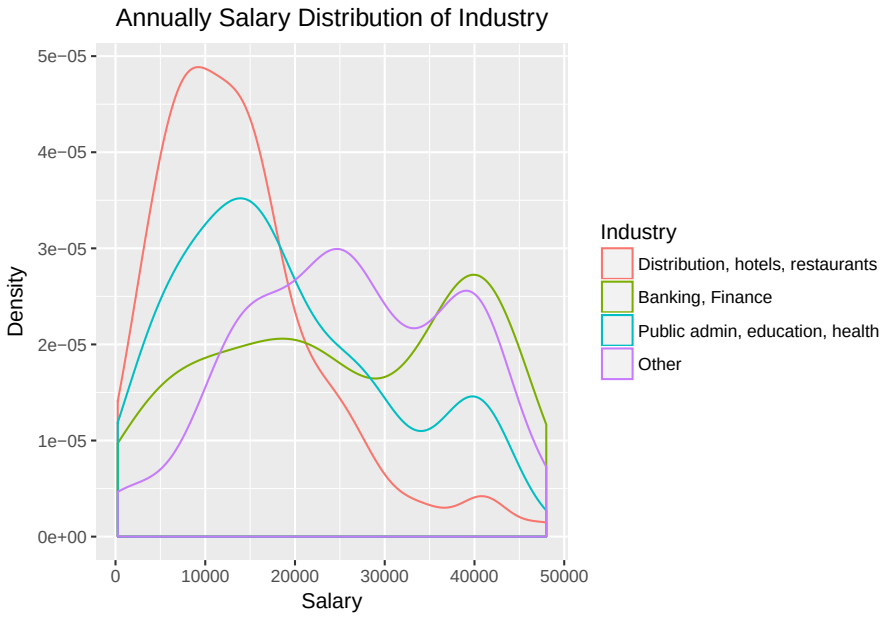


Industry

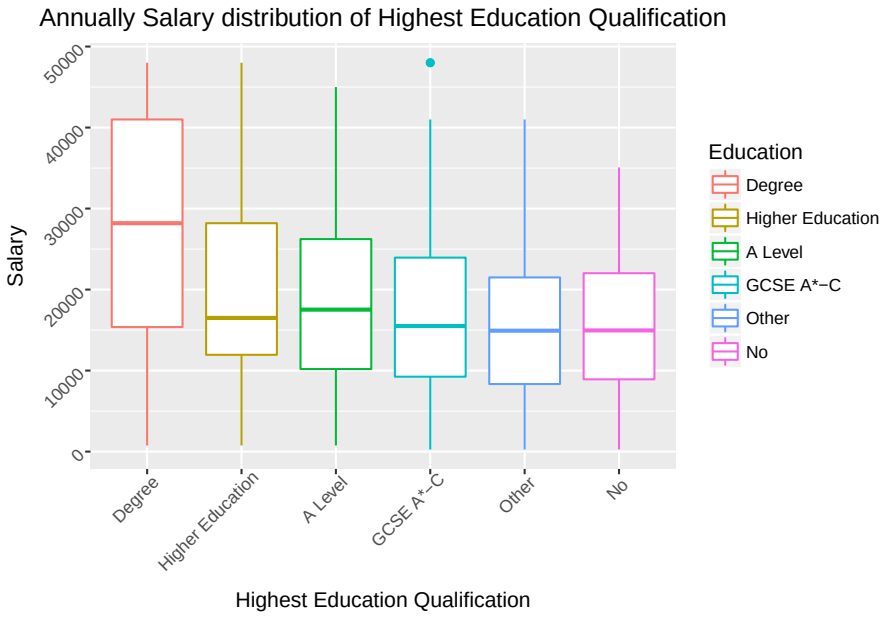
```
## Don't know how to automatically pick scale for object of type labelled. Defaulting to continuous.
```

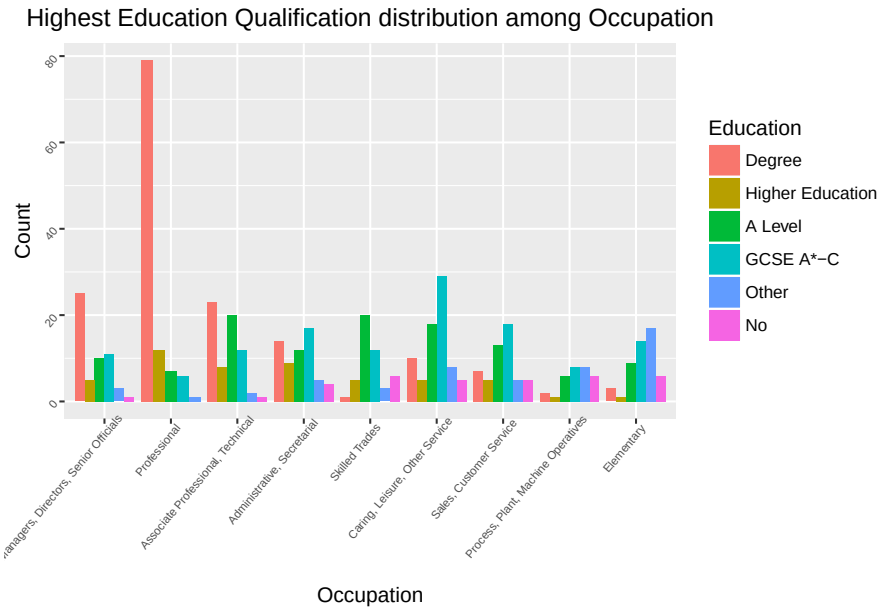

Annually Salary Distribution of Industry
Salary
50000
40000
30000
20000
10000
0
20
40
60
80
Age
Industry
Distribution, hotels, restaurants
Banking, Finance
Public admin, education, health
Other

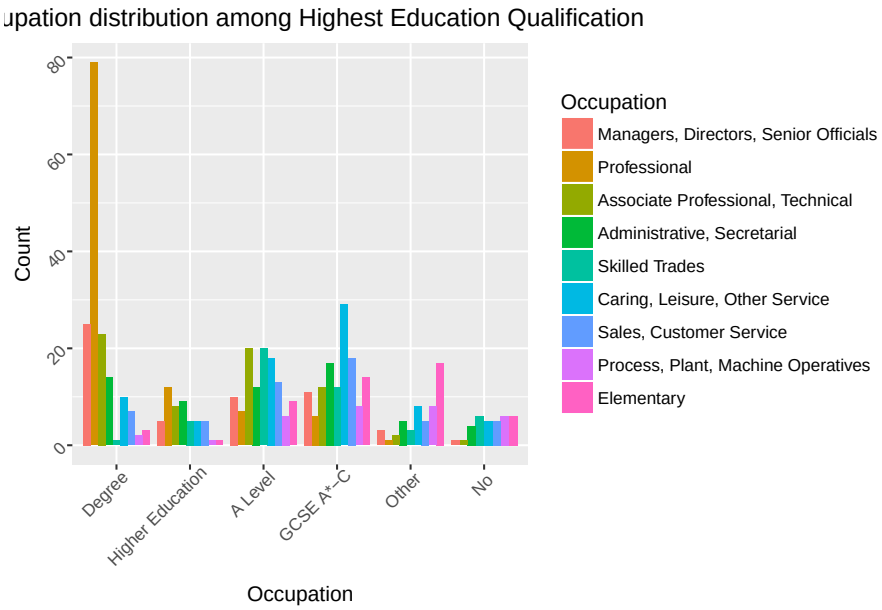

Annually Salary distribution of Industry
Salary
50000
40000
30000
20000
10000
0
Distribution, hotels, restaurants
Banking, Finance
Public admin, education, health
Other
Industry
Industry
Distribution, hotels, restaurants
Banking, Finance
Public admin, education, health
Other



Highest Education Level







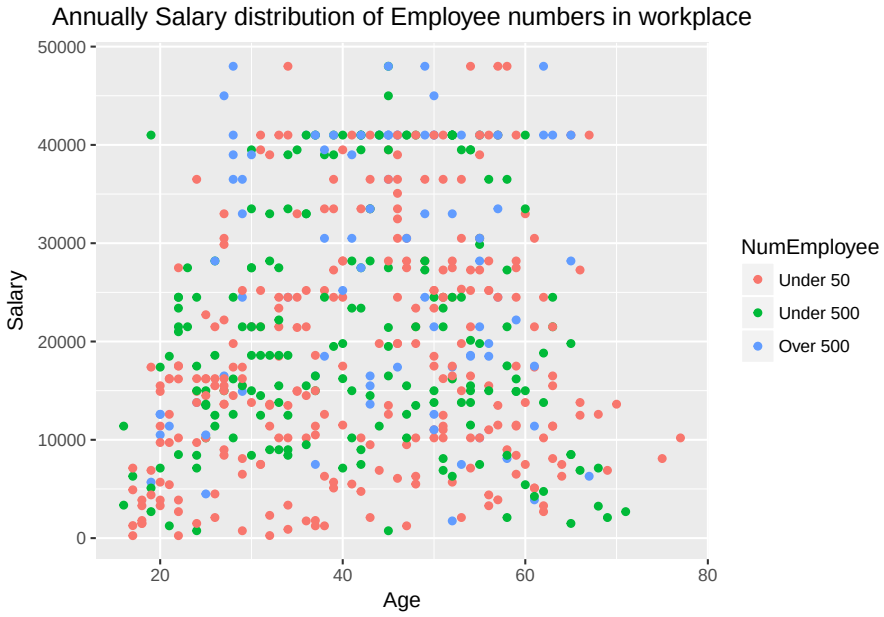

```
## # A tibble: 6 x 2
##          Education `mean(Salary)`
##              <fctr>          <dbl>
## 1           Degree       27478.49
## 2 Higher Education       20436.90
## 3          A Level       19052.11
## 4        GCSE A*-C       17511.61
## 5            Other       15926.26
## 6               No       15737.79
```

```
## # A tibble: 9 x 2
##                              Occupation    percent
##                                  <fctr>      <dbl>
## 1 Managers, Directors, Senior Officials 14.4508671
## 2                        Professional 45.6647399
```

```
## 3     Associate Professional, Technical 13.2947977
## 4           Administrative, Secretarial  8.0924855
## 5                         Skilled Trades  0.5780347
## 6        Caring, Leisure, Other Service  5.7803468
## 7                Sales, Customer Service  4.0462428
## 8    Process, Plant, Machine Operatives  1.1560694
## 9                             Elementary  1.7341040
```

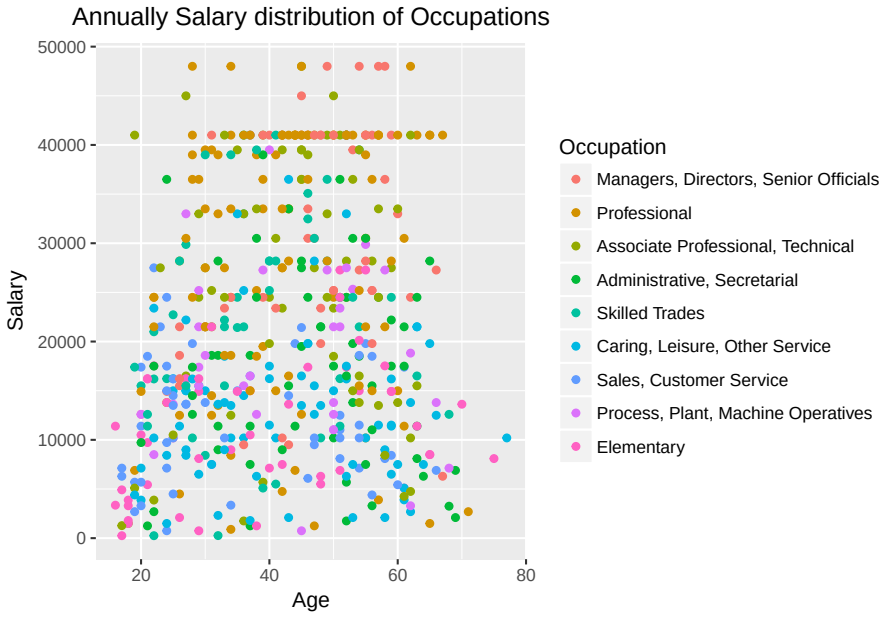
Employee Numbers in workplace

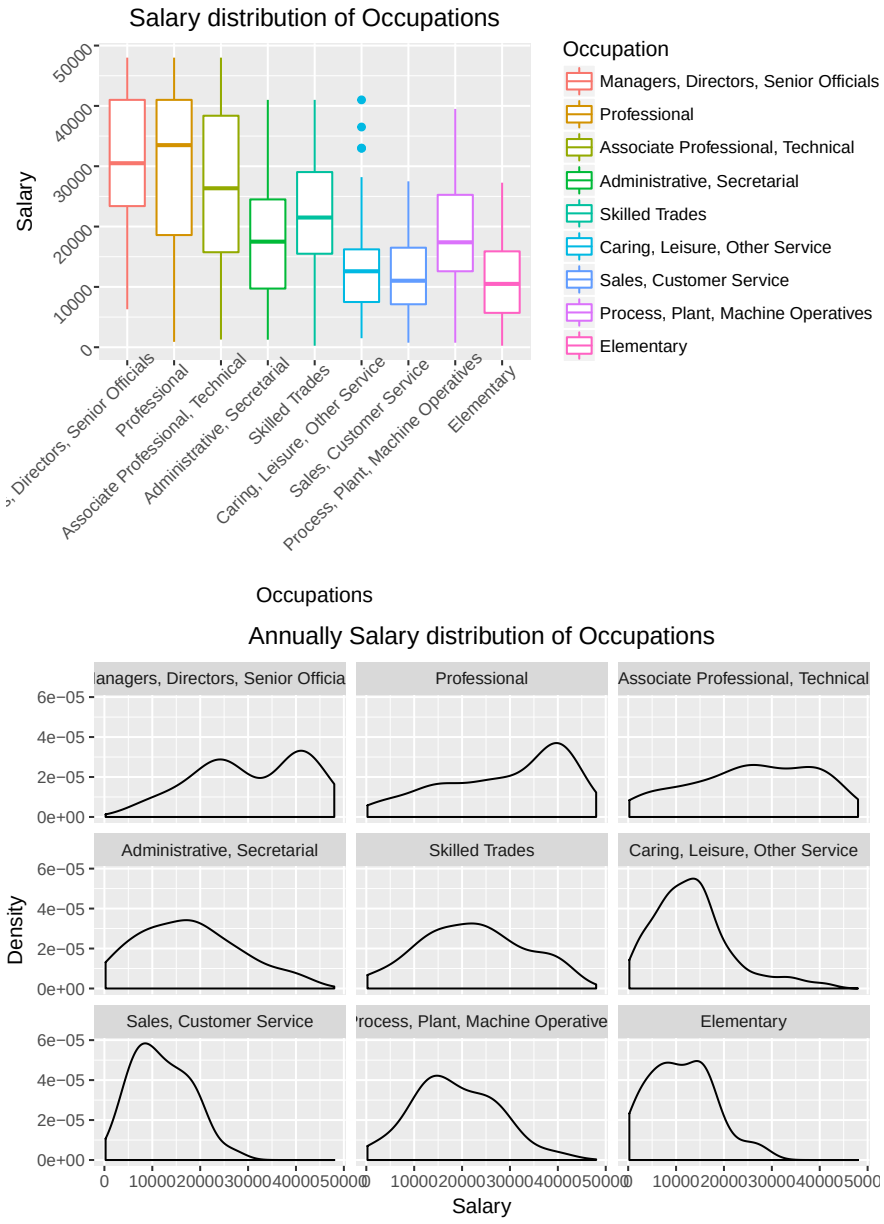
```
## Don't know how to automatically pick scale for object of type labelled. Defaulting to continuous.
```



Occupations

```
## Don't know how to automatically pick scale for object of type labelled. Defaulting to continuous.
```





4.2 Bivariate descriptive

occupation industry occupation education

4.3 Analysis

4.3.1 Correlation analysis

occupation industry occupation education

```
##
##  Pearson's product-moment correlation
##
## data:  qlfs$Education and qlfs$Occupation
```

```
## t = 14.009, df = 541, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4513880 0.5751126
## sample estimates:
##       cor
## 0.5159359
```

```
##
##  Pearson's product-moment correlation
##
## data:  qlfs$Industry and qlfs$Occupation
## t = -5.2834, df = 541, p-value = 1.841e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3000600 -0.1399736
## sample estimates:
##        cor
## -0.2215087
```

4.3.2 Hypothesis test

3.2.1 T-test

```
##
##  Welch Two Sample t-test
##
## data:  caring$Salary and sales$Salary
## t = 1.1784, df = 125.74, p-value = 0.2409
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1010.806  3986.058
## sample estimates:
## mean of x mean of y
##  13441.95  11954.32
```

```
##
##  Welch Two Sample t-test
##
## data:  caring$Salary and elem$Salary
## t = 1.7133, df = 117.36, p-value = 0.08929
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -360.4099 4984.9032
## sample estimates:
## mean of x mean of y
##  13441.95  11129.70
```

```
##
##  Welch Two Sample t-test
##
## data:  sales$Salary and elem$Salary
## t = 0.64894, df = 98.141, p-value = 0.5179
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -1697.030  3346.272
## sample estimates:
## mean of x mean of y
##  11954.32  11129.70
```

3.2.2 Chi-square test

```
## Warning in chisq.test(table(facQlfs$Industry, facQlfs$Occupation)): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(facQlfs$Industry, facQlfs$Occupation)
## X-squared = 352.87, df = 24, p-value < 2.2e-16
```

```
## Warning in chisq.test(table(facQlfs$Education, facQlfs$Occupation)): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(facQlfs$Education, facQlfs$Occupation)
## X-squared = 252.74, df = 40, p-value < 2.2e-16
```

4.3.3 Linear regression

```
##
## Call:
## lm(formula = Salary ~ Education + Sex + Industry + NumEmployee +
##     Occupation, data = facQlfs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28168.4  -5330.9    716.9   5810.2  28070.7
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  30965.20    1838.43  16.843
## EducationHigher Education                    -3962.11    1553.36  -2.551
## EducationA Level                             -4994.07    1306.26  -3.823
## EducationGCSE A*-C                           -4324.68    1288.09  -3.357
## EducationOther                               -3379.98    1701.91  -1.986
## EducationNo                                  -4100.46    1953.10  -2.099
## SexFemale                                    -7073.80     932.32  -7.587
## IndustryBanking, Finance                      3754.20    1589.81   2.361
## IndustryPublic admin, education, health        -36.14    1442.07  -0.025
## IndustryOther                                 5421.75    1406.48   3.855
## NumEmployeeUnder 500                           897.30     924.66   0.970
## NumEmployeeOver 500                           6050.89    1187.47   5.096
## OccupationProfessional                       -1197.35    1647.72  -0.727
## OccupationAssociate Professional, Technical  -2698.83    1739.08  -1.552
## OccupationAdministrative, Secretarial        -8411.84    1812.10  -4.642
## OccupationSkilled Trades                     -8381.08    1950.88  -4.296
## OccupationCaring, Leisure, Other Service     -9424.12    1849.28  -5.096
```

```
## OccupationSales, Customer Service            -11781.38    2042.65  -5.768
## OccupationProcess, Plant, Machine Operatives -12550.62    2227.49  -5.634
## OccupationElementary                         -14916.41    1974.74  -7.554
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## EducationHigher Education                    0.011035 *
## EducationA Level                             0.000148 ***
## EducationGCSE A*-C                           0.000844 ***
## EducationOther                               0.047555 *
## EducationNo                                  0.036255 *
## SexFemale                                    1.51e-13 ***
## IndustryBanking, Finance                     0.018571 *
## IndustryPublic admin, education, health      0.980016
## IndustryOther                                0.000130 ***
## NumEmployeeUnder 500                         0.332295
## NumEmployeeOver 500                          4.86e-07 ***
## OccupationProfessional                       0.467755
## OccupationAssociate Professional, Technical  0.121297
## OccupationAdministrative, Secretarial        4.37e-06 ***
## OccupationSkilled Trades                     2.07e-05 ***
## OccupationCaring, Leisure, Other Service     4.85e-07 ***
## OccupationSales, Customer Service            1.38e-08 ***
## OccupationProcess, Plant, Machine Operatives 2.87e-08 ***
## OccupationElementary                         1.90e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9295 on 523 degrees of freedom
## Multiple R-squared:  0.4795, Adjusted R-squared:  0.4606
## F-statistic: 25.35 on 19 and 523 DF,  p-value: < 2.2e-16
```

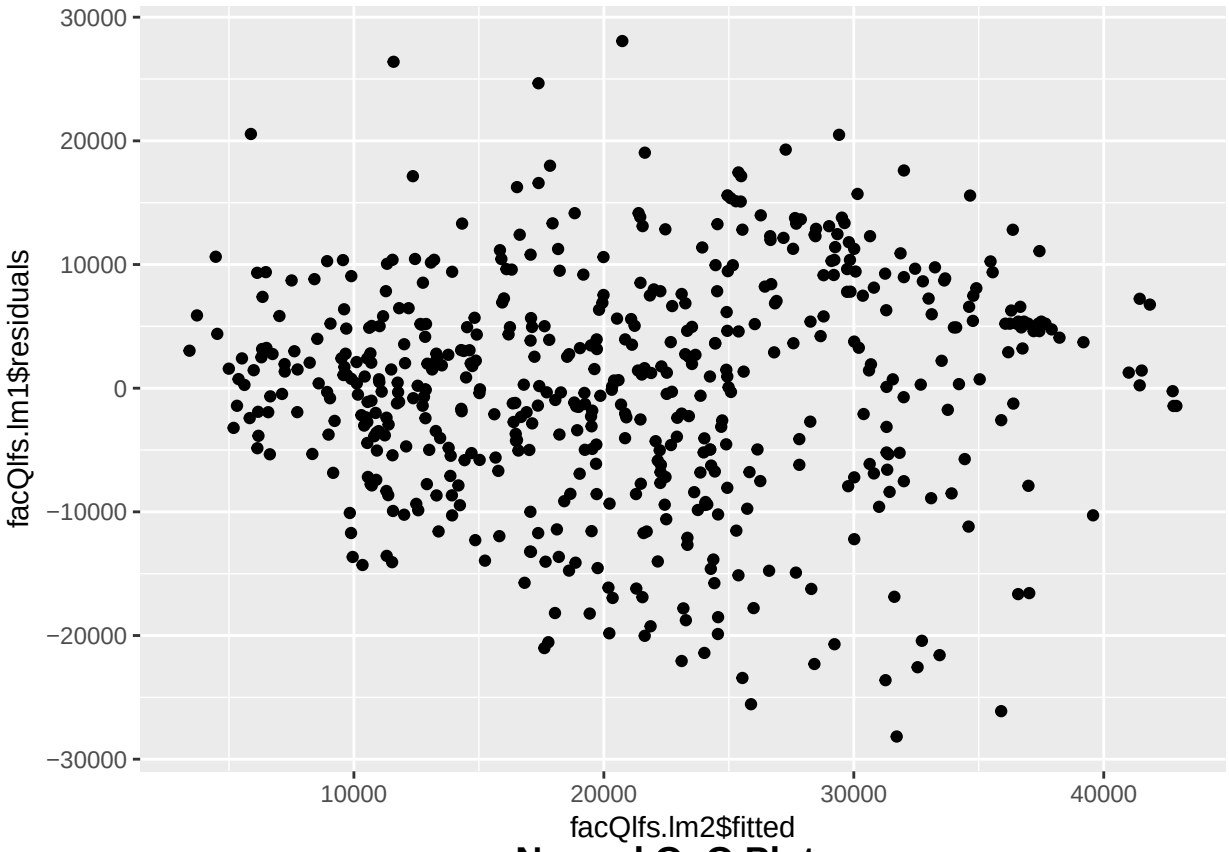
```
##
## Call:
## lm(formula = Salary ~ Education + Sex + Industry + NumEmployee +
##     Occupation + Marital + Religion + Ethnicity + workRegion,
##     data = facQlfs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26967.0  -5948.4    505.5   5967.2  27261.8
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   29672.0     2358.8  12.579
## EducationHigher Education                     -4026.8     1545.4  -2.606
## EducationA Level                              -4842.3     1326.3  -3.651
## EducationGCSE A*-C                            -4643.4     1304.3  -3.560
## EducationOther                                -4196.8     1732.4  -2.422
## EducationNo                                   -4822.0     1955.0  -2.467
## SexFemale                                     -6733.5      950.5  -7.084
## IndustryBanking, Finance                       3323.9     1599.3   2.078
## IndustryPublic admin, education, health        -304.0     1443.3  -0.211
## IndustryOther                                  5704.1     1408.0   4.051
## NumEmployeeUnder 500                            852.8      930.4   0.917
```

```
## NumEmployeeOver 500                              5710.0     1213.1   4.707
## OccupationProfessional                          -1061.2     1657.1  -0.640
## OccupationAssociate Professional, Technical     -2819.2     1756.7  -1.605
## OccupationAdministrative, Secretarial           -8265.5     1820.3  -4.541
## OccupationSkilled Trades                        -8147.8     1961.9  -4.153
## OccupationCaring, Leisure, Other Service        -8917.5     1893.5  -4.709
## OccupationSales, Customer Service              -11633.3     2068.8  -5.623
## OccupationProcess, Plant, Machine Operatives   -12386.7     2232.6  -5.548
## OccupationElementary                           -14151.0     2031.1  -6.967
## MaritalMarried                                   1826.5      929.8   1.964
## MaritalOther                                     2238.0     1385.1   1.616
## ReligionYes                                     -1139.6      860.2  -1.325
## EthnicityAsian                                  -3553.5     2111.1  -1.683
## EthnicityBlack                                  -1550.0     2522.3  -0.614
## EthnicityOther                                   -285.9     2141.3  -0.134
## workRegionYorkshire and the Humber               1826.7     1614.9   1.131
## workRegionEast Midlands                          1476.9     1930.2   0.765
## workRegionWest Midlands                         -3400.1     1879.2  -1.809
## workRegionEast of England                        2152.1     2354.0   0.914
## workRegionLondon                                 2047.1     1635.5   1.252
## workRegionSouth West                             1021.6     1713.9   0.596
## workRegionSouth East                             1133.0     1544.5   0.734
## workRegionOutside UK                             9102.3     9549.4   0.953
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## EducationHigher Education                      0.009435 **
## EducationA Level                               0.000288 ***
## EducationGCSE A*-C                             0.000406 ***
## EducationOther                                 0.015762 *
## EducationNo                                    0.013970 *
## SexFemale                                      4.69e-12 ***
## IndustryBanking, Finance                       0.038183 *
## IndustryPublic admin, education, health        0.833252
## IndustryOther                                  5.89e-05 ***
## NumEmployeeUnder 500                           0.359789
## NumEmployeeOver 500                            3.24e-06 ***
## OccupationProfessional                         0.522218
## OccupationAssociate Professional, Technical    0.109162
## OccupationAdministrative, Secretarial          7.01e-06 ***
## OccupationSkilled Trades                       3.85e-05 ***
## OccupationCaring, Leisure, Other Service       3.21e-06 ***
## OccupationSales, Customer Service              3.10e-08 ***
## OccupationProcess, Plant, Machine Operatives   4.65e-08 ***
## OccupationElementary                           1.00e-11 ***
## MaritalMarried                                 0.050024 .
## MaritalOther                                   0.106753
## ReligionYes                                    0.185834
## EthnicityAsian                                 0.092949 .
## EthnicityBlack                                 0.539165
## EthnicityOther                                 0.893841
## workRegionYorkshire and the Humber             0.258528
## workRegionEast Midlands                        0.444550
## workRegionWest Midlands                        0.070979 .
## workRegionEast of England                      0.361025
```

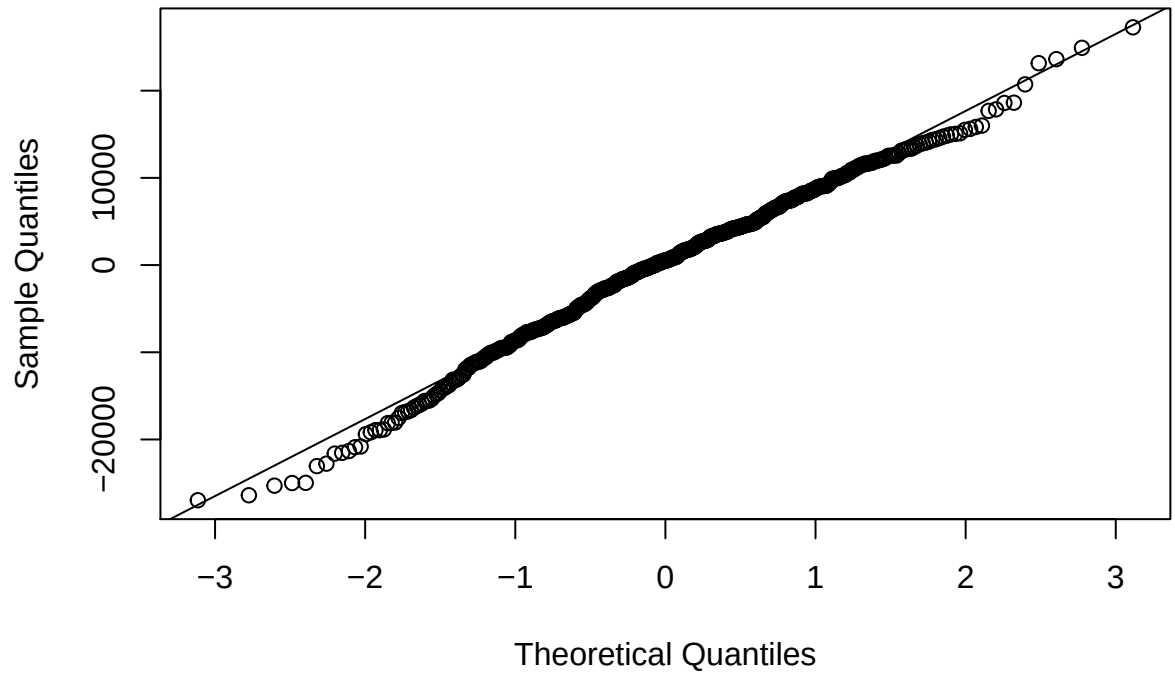
```
## workRegionLondon                              0.211267
## workRegionSouth West                          0.551404
## workRegionSouth East                          0.463538
## workRegionOutside UK                          0.340953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9206 on 509 degrees of freedom
## Multiple R-squared:  0.5031, Adjusted R-squared:  0.4709
## F-statistic: 15.62 on 33 and 509 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: Salary
##              Df     Sum Sq    Mean Sq  F value    Pr(>F)
## Education     5 1.1151e+10 2.2301e+09  26.3161 < 2.2e-16 ***
## Sex           1 1.4246e+10 1.4246e+10 168.1018 < 2.2e-16 ***
## Industry      3 5.1097e+09 1.7032e+09  20.0986 2.521e-12 ***
## NumEmployee   2 2.7248e+09 1.3624e+09  16.0765 1.697e-07 ***
## Occupation    8 8.3922e+09 1.0490e+09  12.3788 2.896e-16 ***
## Marital       2 3.4101e+08 1.7050e+08   2.0120   0.13478
## Religion      1 2.3920e+08 2.3920e+08   2.8226   0.09356 .
## Ethnicity     3 2.9990e+08 9.9965e+07   1.1796   0.31688
## workRegion    8 1.1733e+09 1.4667e+08   1.7307   0.08878 .
## Residuals   509 4.3135e+10 8.4744e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis for anova is that the mean(average value of the dependent variable) is the same for all groups



Normal Q−Q Plot




```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  Salary ~ Education + Sex + Industry + NumEmployee + Occupation
## BP = 37.53, df = 19, p-value = 0.006807
```

5. Conclusion

6. References

Notes:

parallel plots message table for results anova analysis