Banking on Sex Discrimination

Data Analysis by Alexander Oakley

This report discusses the evidence for sex discrimination at an anonymous bank (The Bank). In particular we are interested in whether the bank consistently discriminated against women in terms of their base salaries when compared to the Base Salaries of males. The evidence discussed comes from analysis of data about bank employees hired between 1965-1975.

The data contains 93 observations (32 male, 61 female) and six variables;

- Bsal - Base salary at time of hire;
- Sex - M (males), F (females);
- Senior - Months since first hired;
- Age - Age in months;
- Educ - Years of education;
- Exper - work experience prior to job with the bank (months).

At a Glance

Before discussing more rigorous analysis, we will present some simple visualization (figure 1). If we plot Base Salary against each of the other scalar variables and colour code each data point, a trend can be seen. The data points representing males tend to be above most females in terms of Base Salary. If we overlay a line of best fit, as determined by formal statistical methods, the trend becomes clearer still; that is, males tend to be paid more than women of the same age, education, seniority or prior work experience.

The trend suggested by figure 1 is backed up by further analysis (discussed in the technical section of this report). After taking into account all other factors, it was found that being male was linked to a Base Salary increase of approximately $594. While this number is approxiamte, our analysis suggests that we can be 95% confident that the true figure lies between $430.52 and $832.33 (table 1).

Statistic	Benefit of Being Male ($)	Lower Bound of Estimate	Upper Bound of Estimate
Value	$594	$430.52	$832.33

Table 1 – Summary of findings. Males have a base salary, on average, $594 higher than females.

A Word of Caution

It is important to interpret these results with caution. Although there is a clear correlation between Sex and Base Salary, it is not clear that this link is the result of discrimintation. There may be other, unseen variables at play. For example, it is not clear which jobs each employee in this data set had. Perhaps women tend to occupy positions that were lower paid for reasons other than Sex. Of course, if this were true then the next question you might ask is why are women inhabiting the lower paid roles?

This caveat is in no way intended to suggest that discrimination did not occur. On the contrary, the findings of this report are evidence that it did occur. The author of this report only wishes to remind the reader that these findings are somewhat necessary, but not sufficient to determine guilt.

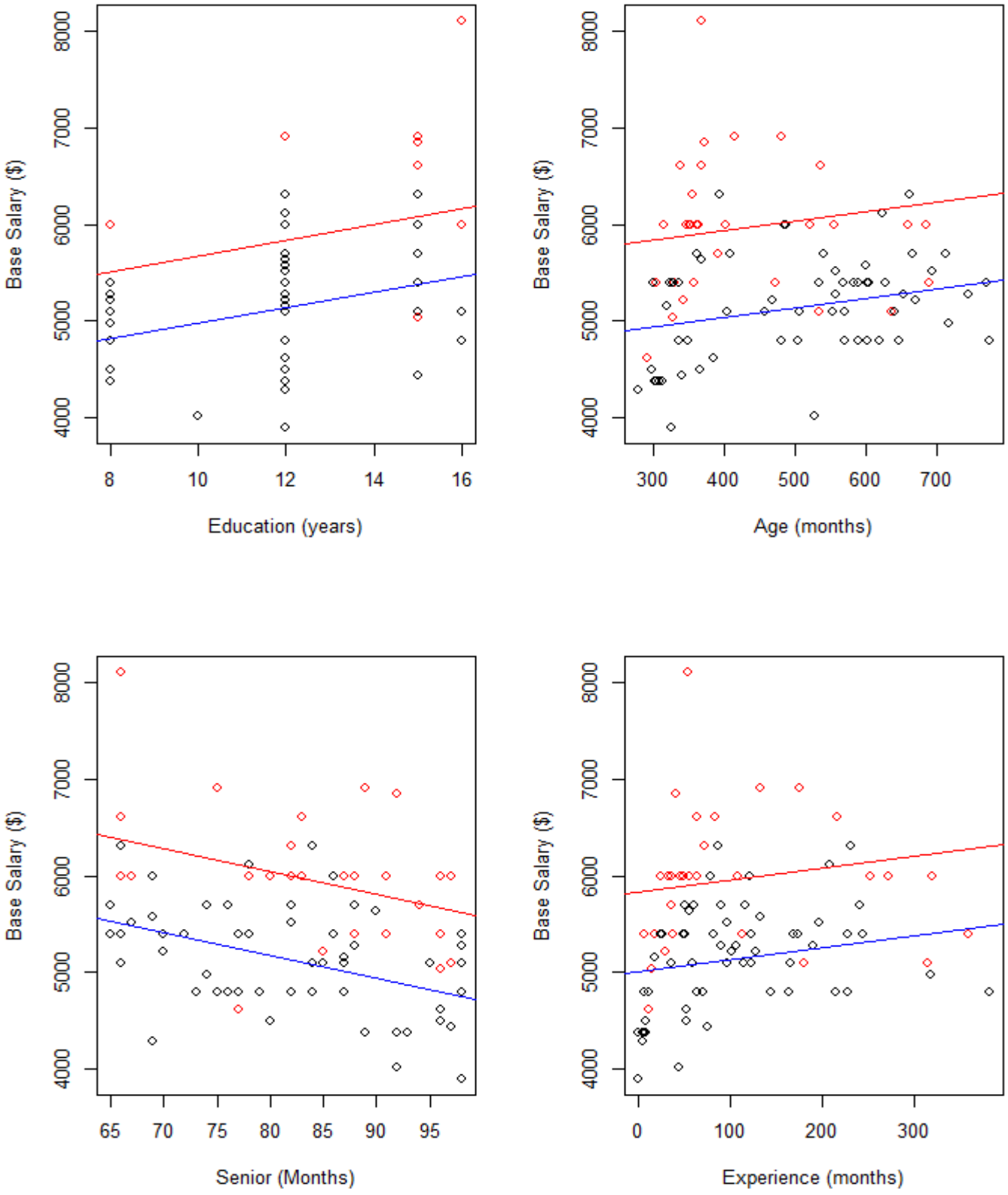

Figure 1 Base Salary (Bsal) plotted against each of the other scalar variables. In each plot, red dots indicate male data points. Each plot has two trendlines. The blue trendline represents females and the red one represents males. There is a consistent tendency for males to have higher base salary.

Statistical Methods and Results

Preliminary Visualisations

The analysis began with visualizations of the data. All variables were looked at *except Sex.* Visualisations included scatter plots of pairs of variable, histograms, and boxplots. The scatter plots did not reveal anything particularly striking besides positive relationships between Base Salary and Experience, and Experience and Age (figure 2).

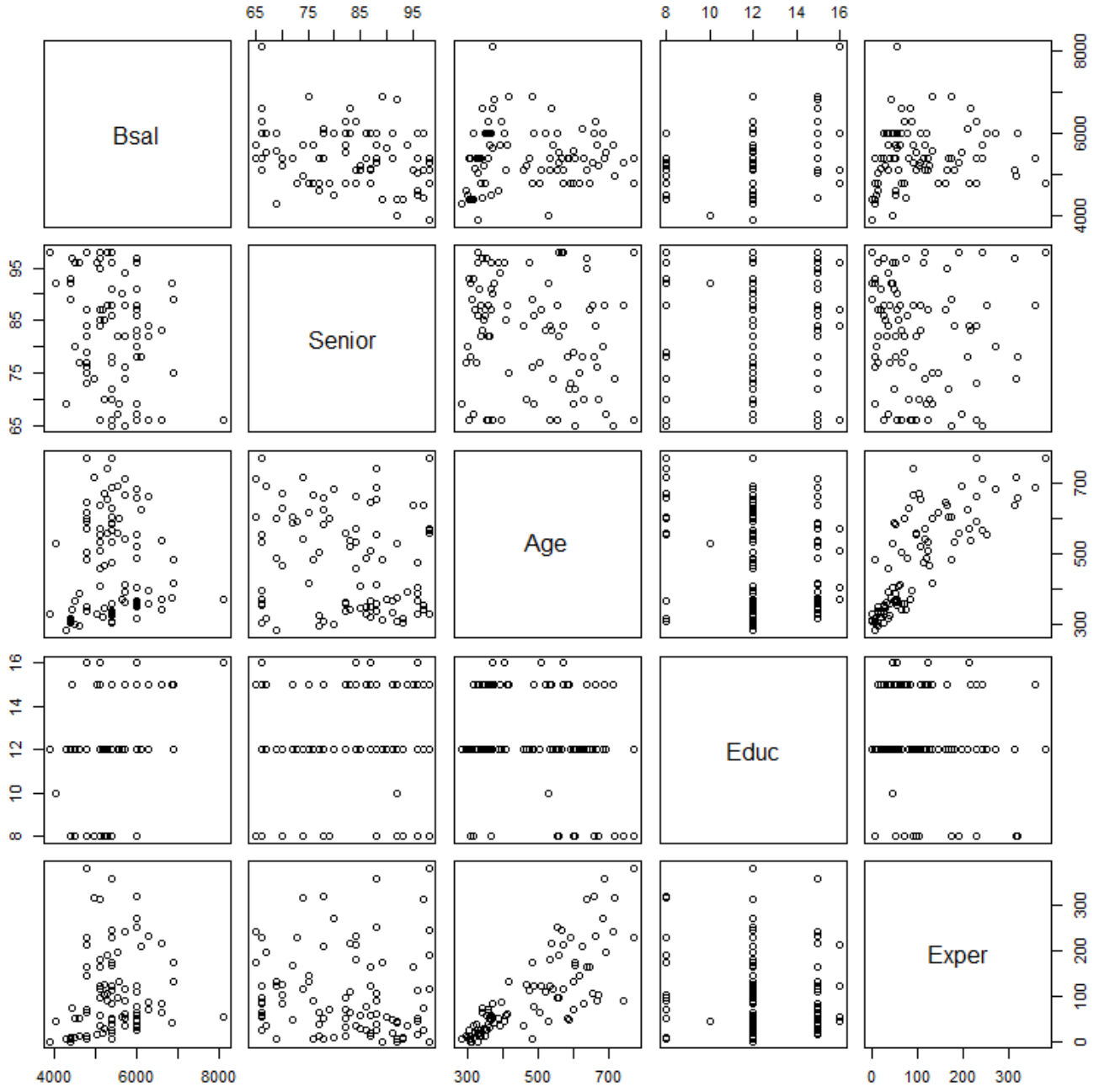

Figure 2 - Scatter plots of all pairs of scalar variables. The most obvious relationships are Bsal ~ Exper and Age ~ Exper

Histograms (figure 3) suggest several noteworthy characteristics of the data;

- Experience had some positive skew
- Base Salary has a bit of a positive skew
- Age had what looked like a bi-modal distribution
- Senior had a peak down at the lower end

These characteristics suggested some transformations of the explanatory variables may be a good idea.

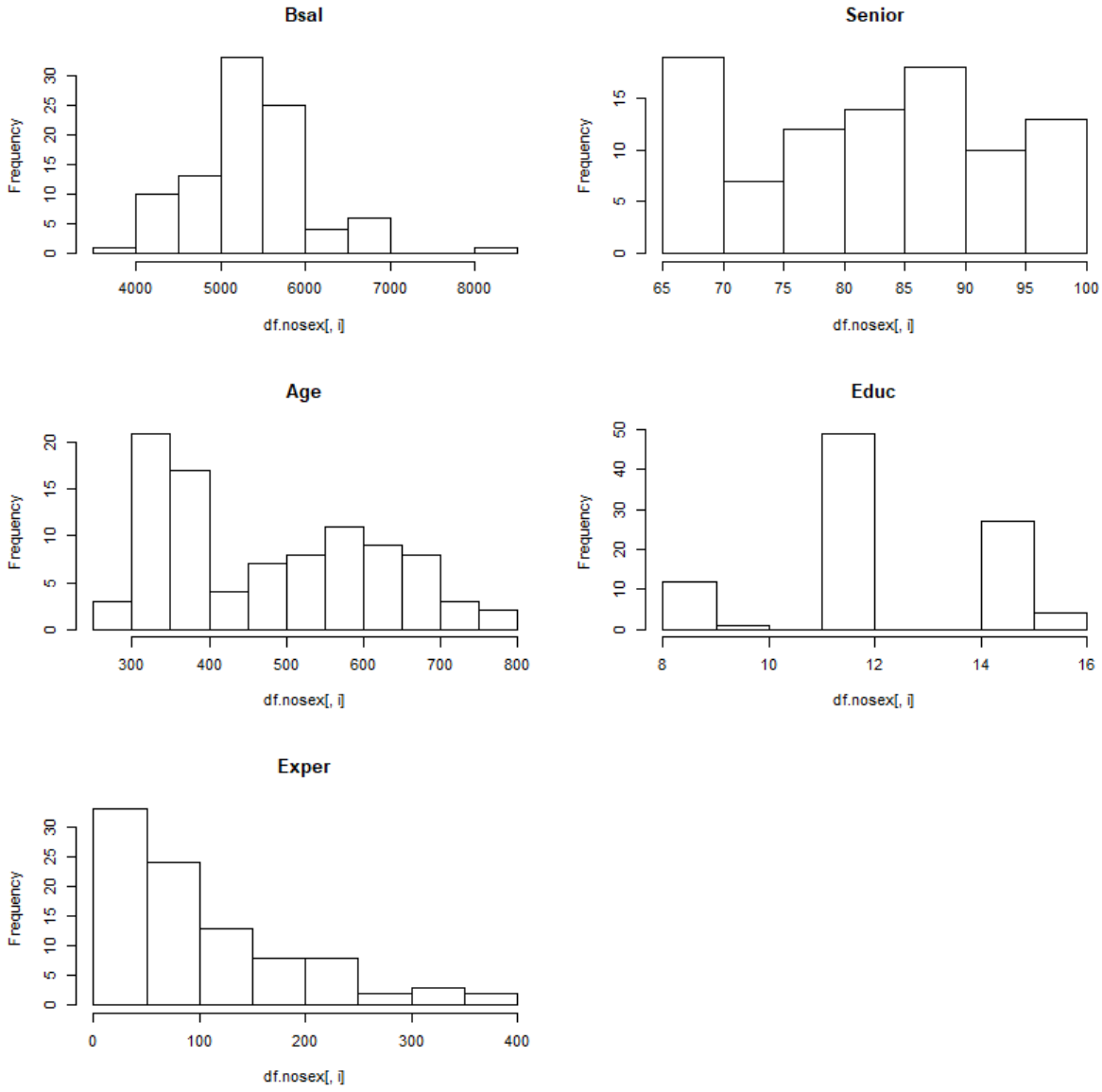

Figure 3 Histograms of all variable besides sex.

The boxplots showed some unusual points in terms of Base Salary and Experience. These points were unusual because they lay more 1.5 times the length of the IQR from the upper end of the IQR. Specificially, there was one employee with an extremely high base salary, and five employees with unusually high levels of experience. It was decided that further investigation was necessary before omitting to these data points from the analysis. This issue is revisted later in the report during discussion of linear model diagnostics (e.g. hatvalues, Cook's d and studentized residuals). We omit a graphic of the boxplots and instead refer to the relevant histograms in figure 3 which illiustrate the extremities of the data.

Preliminary Models – Excluding Sex

A linear model that only considered main effects of the variables Senior, Age, Education and Experience indicated that Age was not a significant predictor of Base Salary. Experimenting with other models that involved different combinations of predictors in concert with Age continued to show that age was not a useful predictor. This is perhaps not surprising given the apparent correlation between Age and Experience that was indicated by the scatter plots early in the analysis (figure 2). It seems that Experience predicts Base Salary better than Age does.

Variable Transforms

Recall that there was some positive skew in the data for some variables. Taking this into consideration, the next sensible step was to try some transforms of that data. Log and squared transforms of all predictor variables were explored using the R's step function. Unfortunately, a log transform of Experience triggered an error due to there being some zero-values (log of zero is undefined). A square-root transform was applied instead.

The best model found by the step function can be seen in table 1. This model conforms nicely with the transformations suggested by the histograms. Specifically, Experience, which had the largest positive skew, has a transform that is known to remedy positive skew. Also, surprisingly, Age is a significant predictor in this model. It seems that all of the explanatory variables can now be made use of.

A Boxcox was run on this new model. It suggested the response variable be transformed by an exponent of negative one (figure 4). This transformation yielded a higher R-squared and generally lower p-values (table 2).

Best No-Sex-Model ***With No*** Response Variable Transform

```
Call:
lm(formula = Bsal ~ Senior + Age + Educ + Exper + sqrt(Exper),
    data = df.nosex)

Residuals:
     Min       1Q   Median       3Q      Max
-1431.68  -212.88   -32.58   248.20  1865.26

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5259.4064   695.5194   7.562 3.81e-11 ***
Senior       -15.9174     5.7403  -2.773 0.006796 **
Age           -2.4063     0.7452  -3.229 0.001752 **
Educ          94.7882    26.3613   3.596 0.000536 ***
Exper         -7.5951     2.3878  -3.181 0.002036 **
sqrt(Exper)  245.3293    50.7967   4.830 5.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 535.3 on 87 degrees of freedom
Multiple R-squared:  0.4619,    Adjusted R-squared:  0.4309
F-statistic: 14.93 on 5 and 87 DF,  p-value: 1.446e-10
```

Table 1- Best model that did not use Sex as a predictor variable

Best No-Sex-Model ***With*** Reponse Variable Transform

```
Call:
lm(formula = 1/Bsal ~ Senior + Age + Educ + Exper + sqrt(Exper),
    data = df.nosex)

Residuals:
       Min         1Q     Median         3Q        Max
-3.727e-05 -9.319e-06 -6.220e-07  6.535e-06  5.267e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.007e-04  2.290e-05   8.765 1.36e-13 ***
Senior       5.275e-07  1.890e-07   2.791 0.006452 **
Age          7.308e-08  2.453e-08   2.979 0.003751 **
Educ        -3.135e-06  8.679e-07  -3.612 0.000508 ***
Exper        3.001e-07  7.862e-08   3.818 0.000252 ***
sqrt(Exper) -9.206e-06  1.672e-06  -5.504 3.69e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.762e-05 on 87 degrees of freedom
Multiple R-squared:  0.4982,    Adjusted R-squared:  0.4694
F-statistic: 17.28 on 5 and 87 DF,  p-value: 7.696e-12
```

Table 2- Summary of the same linear mode shown in figure 4, but with a transform of the response variable. Notice that the R-squared is higher and p-value of the F-statistic is lower.

Diagnostics

The residual diagnostics prior to a transform of the response variable were acceptable (figure 5). The most concerning issue was that two points, 7 and 79, stood out on the Normal Q-Q plot. After the transform to the response variable the diagnostics, particularly the Scale-Location plot, looked slightly better (figure 6). Also, point 7 was no-longer an obvious issue on the Normal Q-Q plot. Point 79, however, was still there.

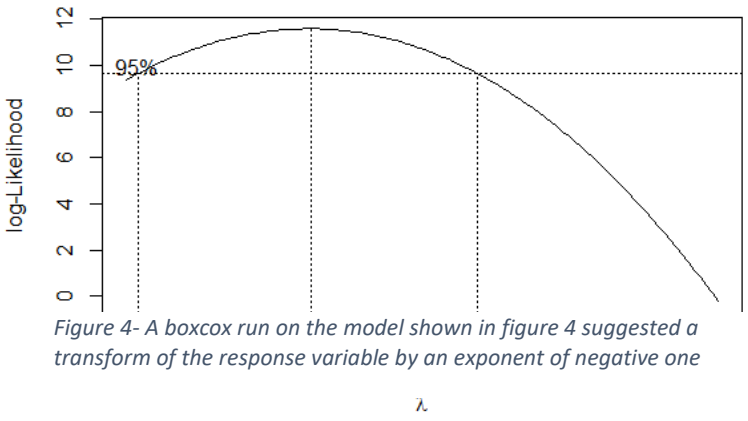

Figure 4- A boxcox run on the model shown in figure 4 suggested a transform of the response variable by an exponent of negative one

Hat-values and studentized residuals also indicate some data points worthy of investigation (figure 7)

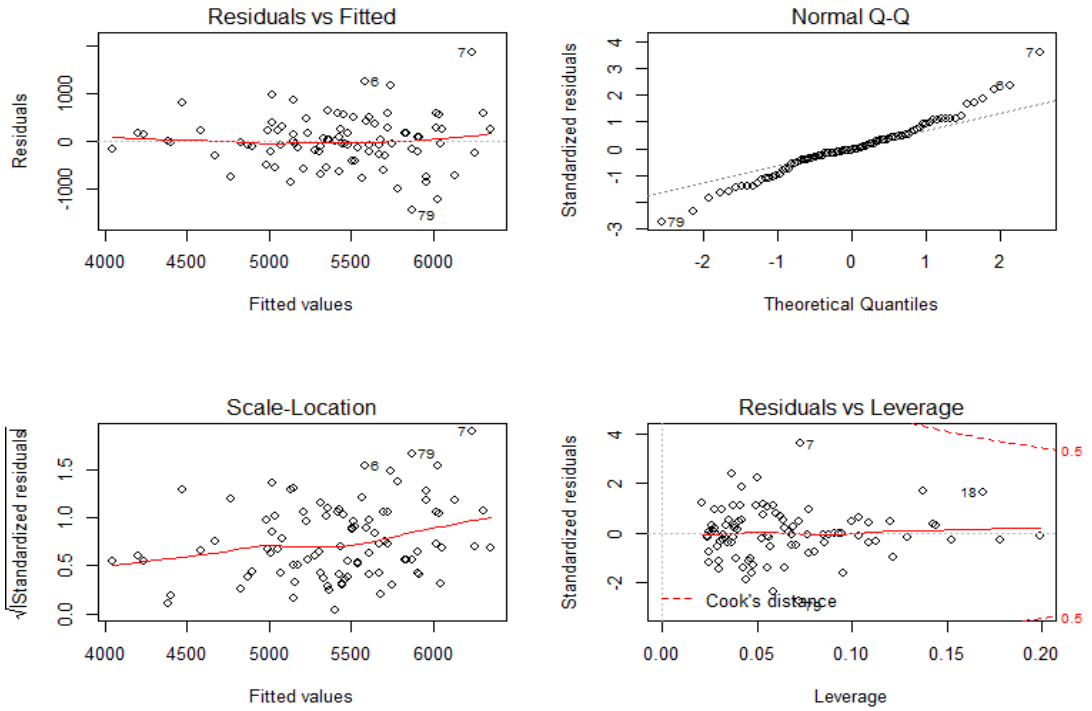

Figure 5- Residual Diagnostics for model: Bsal ~ Senior + Age + Educ + Exper + sqrt(Exper). *Note that points 7 and 79 cause some problems for the assumption of normal distribution of residuals (shown in Normal Q-Q plot)*

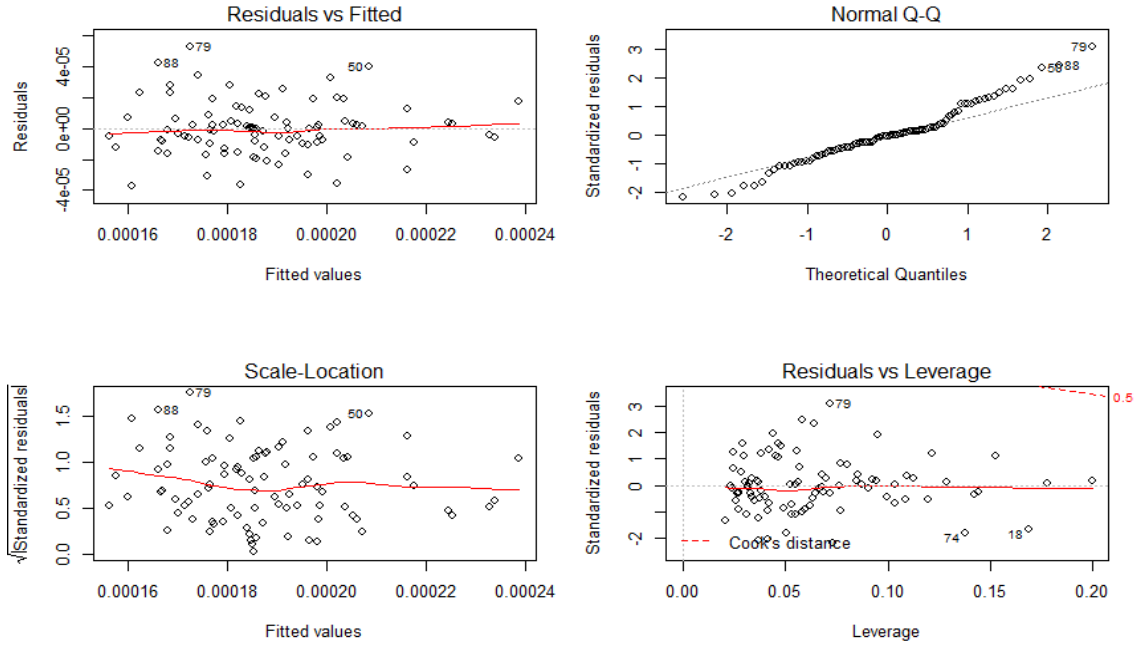

Figure 6- Residual Diagnostic Plots for model: 1/Bsal ~ Senior + Age + Educ + Exper + sqrt(Exper). *Note the (slight) improvement on the diagnostics shown in figure 5, in particular the scale location plot is a bit more horizontal.*

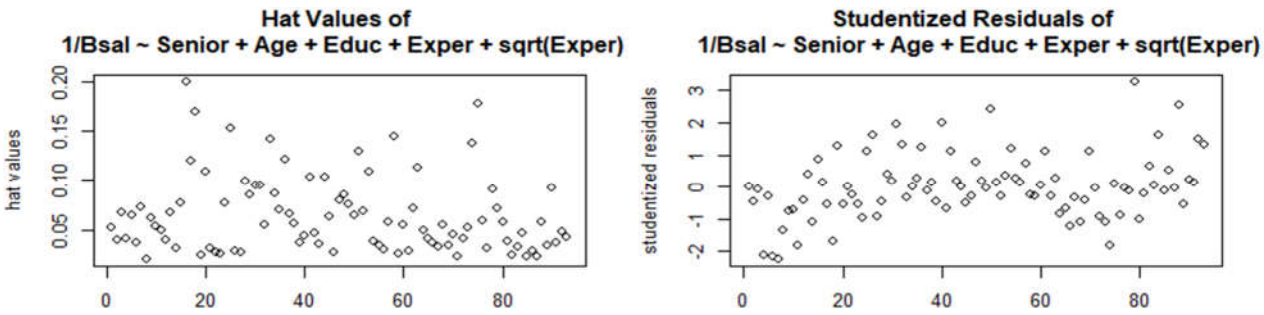

Figure 7 - Point 16 has the highest hatvalue of slightly over three times the average.Point 79 had the greatest absolute studentized residual of slightly about three. These points were both found to be femailes.

Investigating Unusual points – 16 and 79

The figures below show where points 16 and 79 sit in distributions of each variable. Point 16 is show in figure 8 and point 79 is shown in figure 9. Both these points are female. It was noted that both points had lower than average Base Salaries, but not correspondingly low levels of education or experience. Recall that point 16 was identified in the boxplots as having unusually high experience. Although this does raise concerns, it was noted that the experience of point 16 is within 10% of its nearest neighbour, and within 20% of it 4 nearest neighbours (see appendix, table 8).

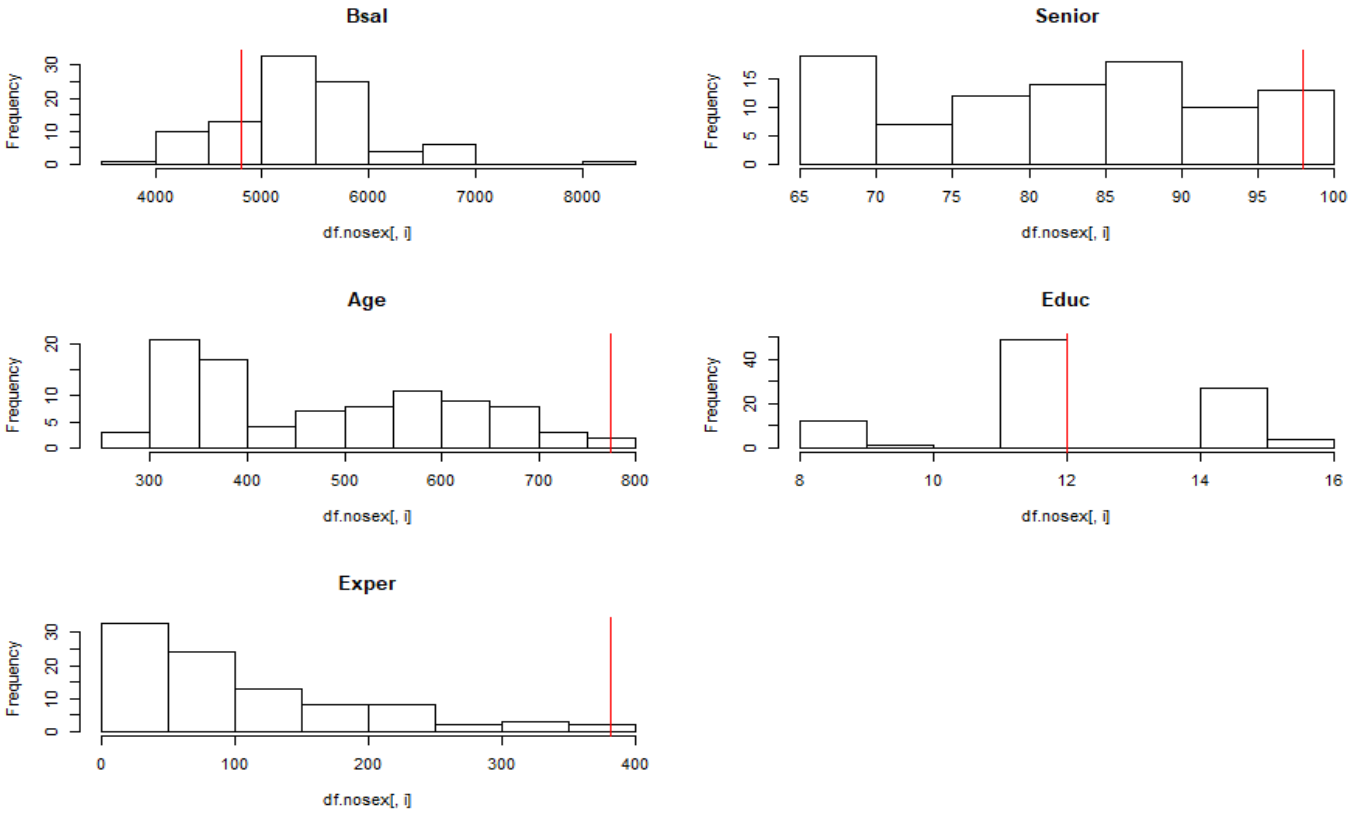

Figure 8- The location of point 16 (shown as a redline) in the distribution of each scalar variable. It seems clear that this point is paid less than average, while being above average in most other measures available. Point 16 is female.

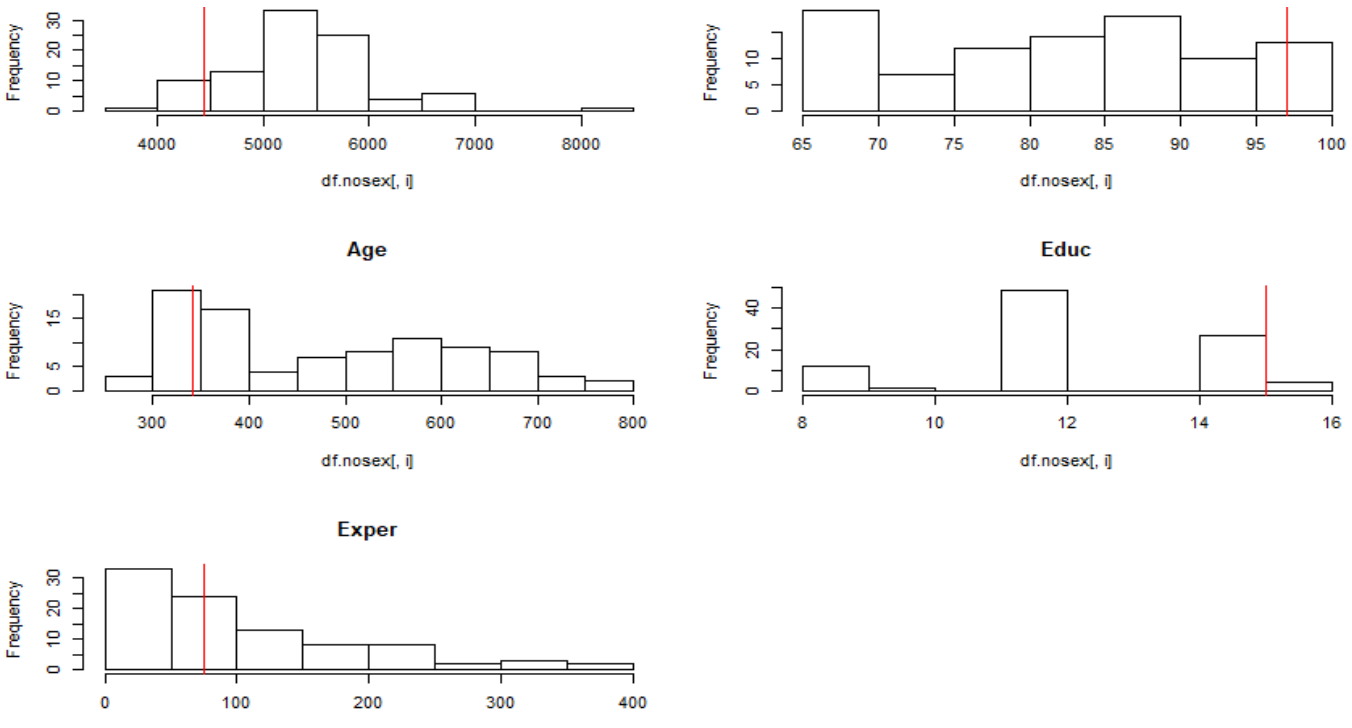

Figure 9 - The location of point 79 (shown as a redline) in the distribution of each scalar variable. It seems clear that this point is paid less thanaverage, while being above average in most other measures available. Point 79 is female.

Furthermore, both points are particularly relevant to the question at hand. That is, are women discriminated against in terms of Base Salary? Because both points were women who were paid less than average despite having either high Education or Experience, it would be a mistake to not include them in further analysis. However, it would also be sensible to check if the inclusion of these points affects the predicted coefficient for sex. This is discussed later in the report.

Is Sex a Useful Predictor for Base Salary?

Now we get down to the bottom line; did The Bank discriminate? In short, it looks like the answer to that question is "yes, they do pay women less". If we include the main effect of Sex in the linear model show in table 2, then we can reject the null hypothesis that Sex is not a predictor of Base Salary. Table 3 shows the summary of the linear model that includes Sex. Notice that Age is no longer a significant predictor. Table 4 shows a summary of the linear model without Age.

```
Call:
lm(formula = 1/Bsal ~ Sex + Senior + Age + Educ + Exper + sqrt(Exper),
    data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-2.839e-05 -9.265e-06 -1.219e-06  8.234e-06  4.117e-05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.086e-04  1.935e-05  10.782  < 2e-16 ***
SexMale     -2.301e-05  3.812e-06  -6.036 3.89e-08 ***
Senior       5.651e-07  1.594e-07   3.544 0.000639 ***
Age          1.449e-08  2.285e-08   0.634 0.527561
Educ        -2.182e-06  7.485e-07  -2.915 0.004533 **
Exper        3.159e-07  6.633e-08   4.763 7.67e-06 ***
sqrt(Exper) -7.952e-06  1.425e-06  -5.580 2.74e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.486e-05 on 86 degrees of freedom
Multiple R-squared:  0.6475,    Adjusted R-squared:  0.6229
F-statistic: 26.33 on 6 and 86 DF,  p-value: < 2.2e-16
```

Table 3 - Adding Sex to the best model found without Sex as a predictor

Because the response variable is transformed, it is not a straight forward task to interpret the size of the effect of Sex. However, there are *some* obvious conclusions that can be drawn.

Firstly, because the transform of the response is to the power of negative one, the ordering of the response is reversed. This means that the negative coefficient shown in table 3 corresponds to a positive relationship between Sex and Base Salary where being female is the baseline. Thus, being male correlates with being paid more.

```
Call:
lm(formula = 1/Bsal ~ Senior + Educ + Exper + sqrt(Exper) + Sex,
    data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-2.848e-05 -9.464e-06 -1.555e-06  9.454e-06  4.229e-05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.144e-04  1.704e-05  12.578  < 2e-16 ***
Senior       5.579e-07  1.585e-07   3.520 0.000689 ***
Educ        -2.270e-06  7.329e-07  -3.098 0.002627 **
Exper        3.165e-07  6.609e-08   4.789 6.83e-06 ***
sqrt(Exper) -7.603e-06  1.310e-06  -5.805 1.03e-07 ***
SexMale     -2.403e-05  3.439e-06  -6.989 5.33e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.481e-05 on 87 degrees of freedom
Multiple R-squared:  0.6459,    Adjusted R-squared:  0.6255
F-statistic: 31.73 on 5 and 87 DF,  p-value: < 2.2e-16
```

Table 4- Including Sex but excluding Age

Next, if we enquire into the size of the effect of Sex, we are met with the challenge of teasing this out from the effects of the variables in the equation. Considering the model in table 4, if we set all other variables to zero, then we are left with only the effect of Sex and the intercept. Rearranging the equation shows that being male would earn you an extra $594 (assuming that the units of Base Salary are in dollars).

Interactions of Sex with other variables were checked, but nothing that would improve the model was found.

Concerns about unusual data points and the shape of the data.

As was mentioned earlier in the report, there were some concerning aspects of the data. Firstly, age is bimodal. Secondly, there were some suspicious data points. Both issues are discussed here.

Bimodal Distribution of Age

It was considered that the bimodal distribution of Age might indicate that two different subpopulations existed. Further investigation revealed that most older employees were female (table 5). If the data is separated into young (Age<450) and old (Age>=450), and we use the step function to find significant predictors, we can see that p-values tend to be higher among the Young partition (table 6 and 7). It is not clear what the meaning of this difference is. We leave that analysis for a different report.

	Age	
	<450	>450
Male (%)	24	11
Female (%)	25	41

Table 5 - Proportions according to Age and Sex

Best linear Model Predicting Base Salary Amoung the Youngest Employees (Age < 450)

```
Call:
lm(formula = 1/Bsal ~ Age + Senior + Educ + Exper + sqrt(Exper) +
    Sex, data = df.young)

Residuals:
       Min         1Q     Median         3Q        Max
-2.122e-05 -9.436e-06 -3.446e-07  8.978e-06  2.474e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.773e-04  3.218e-05   8.616 1.81e-10 ***
Age         -2.791e-07  1.117e-07  -2.499 0.016894 *
Senior       9.052e-07  2.526e-07   3.583 0.000952 ***
Educ        -2.825e-06  1.254e-06  -2.253 0.030117 *
Exper        5.663e-07  2.521e-07   2.246 0.030604 *
sqrt(Exper) -7.951e-06  2.755e-06  -2.886 0.006394 **
SexMale     -1.911e-05  4.859e-06  -3.934 0.000343 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.367e-05 on 38 degrees of freedom
Multiple R-squared:  0.8137,    Adjusted R-squared:  0.7842
F-statistic: 27.65 on 6 and 38 DF,  p-value: 1.946e-12
```

Table 6- Linear model that only uses employees with age less than 450 months

Best linear Model Predicting Base Salary Amoung the Oldest Employees (Age < 450)

```
Call:
lm(formula = 1/Bsal ~ Senior + Educ + Exper + sqrt(Exper) + Sex
    data = df.old)

Residuals:
       Min         1Q     Median         3Q        Max
-2.704e-05 -1.024e-05 -4.020e-07  8.671e-06  3.980e-05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.113e-04  2.779e-05   7.602 2.03e-09 ***
Senior       5.435e-07  2.362e-07   2.301  0.02644 *
Educ        -1.609e-06  9.332e-07  -1.725  0.09197 .
Exper        2.762e-07  1.278e-07   2.162  0.03636 *
sqrt(Exper) -7.291e-06  3.082e-06  -2.366  0.02270 *
SexMale     -2.058e-05  6.150e-06  -3.347  0.00173 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.52e-05 on 42 degrees of freedom
Multiple R-squared:  0.3824,    Adjusted R-squared:  0.3089
F-statistic: 5.201 on 5 and 42 DF,  p-value: 0.0008367
```

Table 7- Linear model that only uses employees with age greater than or equal to 450 months

Unusual Data Points

It has been discussed that points 16 and 79 were unusual. If we exclude these points and try to fit the same linear model, we do not find much difference between models. What is especially relevant is that Sex remains a significant predictor of Base Salary. In this new model Sex has a similar coefficient and p-value. Because points 16 and 79 seemed to be examples of underpaid females, we take this as further evidence of sexual discrimination by the bank.

Appendix – Software Used

All analysis was done using R version 3.5.2

The following table contains information on the functions and packages used.

Aspect of Report	Function and Package used to Produce aspect of Report
All figures	Plot() function – R core
Results	lm() and summary() – built in R functions
Boxcox – to determine transformations	Boxcox – built in R function
All tables	A combination of R and Word

Additional Table

	X	Bsal	Sex	Senior	Age	Educ	Exper
16	16	4800	Female	98	774	12	381
51	51	4980	Female	74	718	8	318
63	63	5100	Male	97	637	12	315
74	74	6000	Male	78	659	8	320
75	75	5400	Male	88	690	15	359

Table 8 - Five highest Experience data points