Assignment 1 – Predictive Analytics

Executive Summary:

The main business problem that Jobs4Ya have identified is creating a model to predict a workers Weekly Earnings; using this model Jobs4Ya would then be able to advise potential job seekers as to the best activities to pursue in order to maximise their Earning potential.

I found that the best 2 predictive indicators of a worker's Weekly Earnings are Weekly Hours worked and Education Level.

*Weekly Earnings = 4.17 + 0.38 * Weekly Hrs Worked + 1.07 * Education level + 8.348 * Standard error*

Correlation Table 6 and Correlation Table 8 help show the interaction between the variables. In Correlation Table 8 we can see that there is little interaction with what activities workers do and their Weekly earnings. Instead if we examine Correlation Table 6, we can see that there is much greater interaction between Weekly earnings and Weekly hours worked and the level of education. We can interpret this to mean that the activities that workers do outside of work have little impact on their earnings; instead it would be more productive to work extra hours or look to increase their Education Level by pursuing higher education.

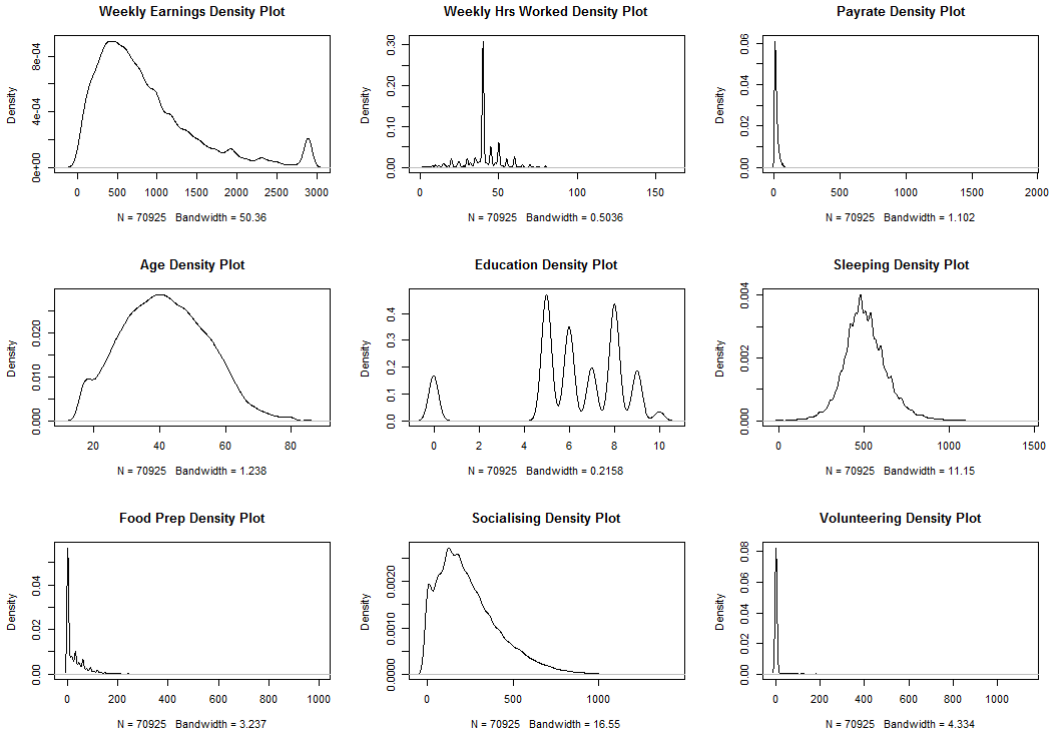
Data exploration:

To estimate Weekly earnings I selected the following variables form the data set:

Weekly Hours Worked, Age, Gender, Education Level, Hours Sleeping, Hours of Food Preparation, Hours, Socialising, Hours Volunteering.

Gender and Education Level are both categorical variables and so had to be converted to numerical variables to be able to be used in the model. For Gender I created a new numeric variable called ‘maleyes’ for converted the values “Female” and “Male” to 1 and 0 respectively. For “Education Level” the various levels were ranked from 1 to 11.

The chosen variables data had the following distributions (Graph 1):



Graph 1 – “Distribution of Chosen Variables”

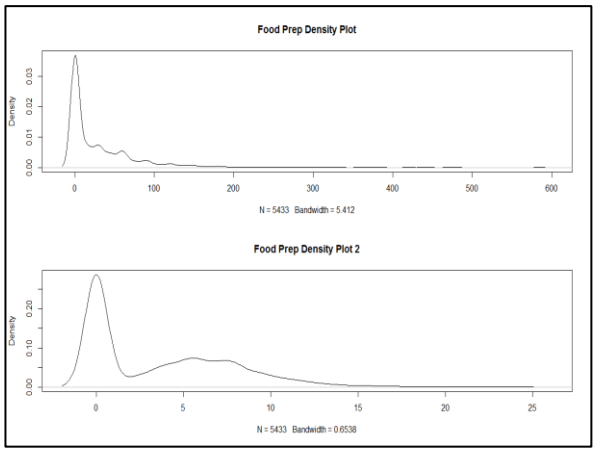
I chose the variable ‘Age’ over ‘Age Range’ as ‘Age’ is a continuous numerical variable instead of a categorical; this means that it has far more intervals and so should be more accurate.

For the ‘Employment.Status’ variable, I removed the records of the unemployed and those searching for employment and in the ‘Weekly.Earnings’ variabke I dropped all instances of workers who earn $0 to remove outliers.

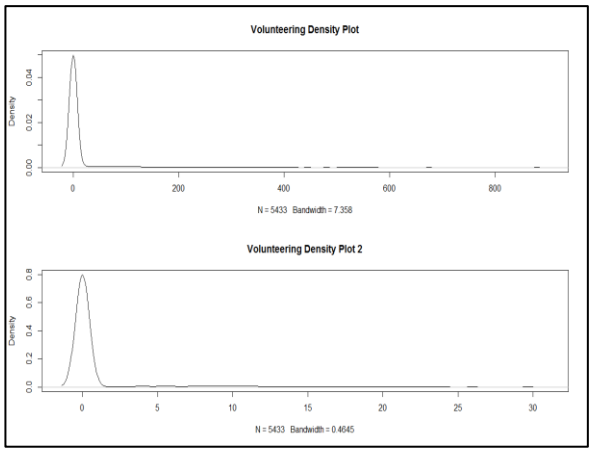
Relationships + Transformations:

Show which indicators have strong correlation here – pick a most appropriate variable.

On examining the distribution of the variables, some of the variables were positive skewed so I used a square root transformation to make these variables more useful. The Transformed variables are displayed in "Transformed Variables Density Graphs 9". Whilst this transformation was useful for the "Food and Drink Prep" variable (Graphs 2 + 3), it had a limited impact for the "Volunteering" variable (Graphs 4 + 5). As the transformed Volunteering variable was still aggressively positively skewed I decided to drop the variable from my model.

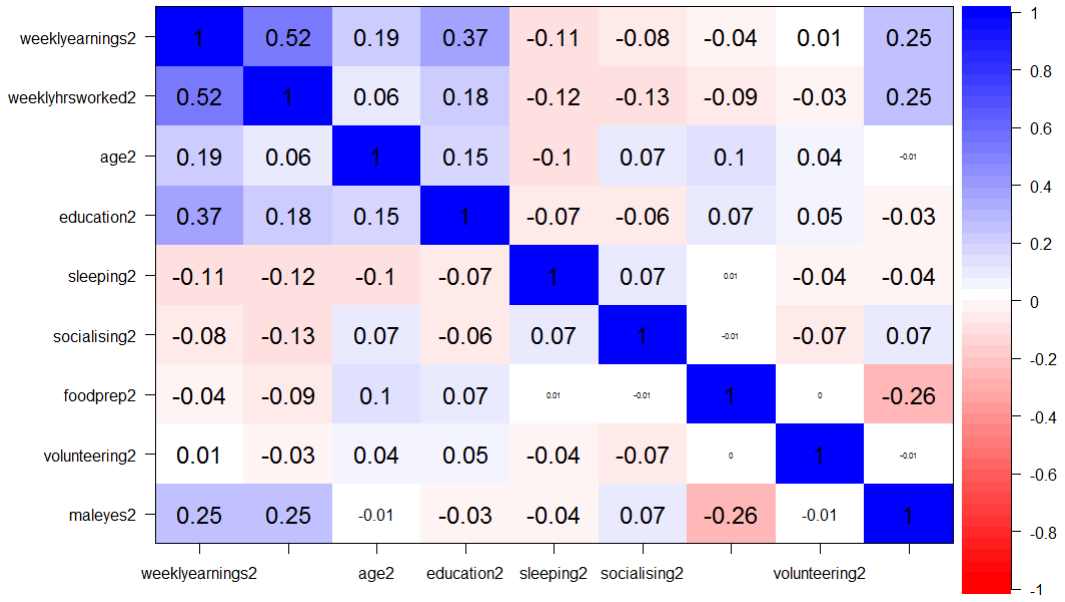


"Graphs 2 + 3"



"Graphs 4 + 5"

Here is a correlation table of the variables that I have chosen after they have been transformed: (weeklyearnings2 is the target variable.)



"Correlation Table 6"

On examining Correlation Table 6 it is clear that 'weeklyhrsworked2' is the most appropriate target variable and the most correlated to 'weeklyearnings2'. I chose to drop all variables that had below .15 correlation – leaving weeklyhrsworked2, age2, education2 and maleyes2.

Multi Regression Model:

After preparing and transforming the variables above, I conducted several tests in which I performed backwards elimination, dropping certain variables to see if I could improve the overall models performance.

The first linear model containing all the chosen variables had the following results:

Adj R^2 = 0.3571 - F-Stat: 302.7 on 8 P Value: < 2.2e-16 = "fit1"

The second linear model dropped 'sleeping2', 'socialising2', 'foodprep2' and 'volunteering2'. This resulted in:

Adj R^2 = 0.3566 - F-Stat: 603.1 on 4 P Value: < 2.2e-16 = "fit2"

Dropping the 'maleyes2' and 'age2' variables helped improve performance more for the final model:

Adj R^2 = 0.3215 - F-Stat: 1031 on 2 P Value: < 2.2e-16 = "fit4"

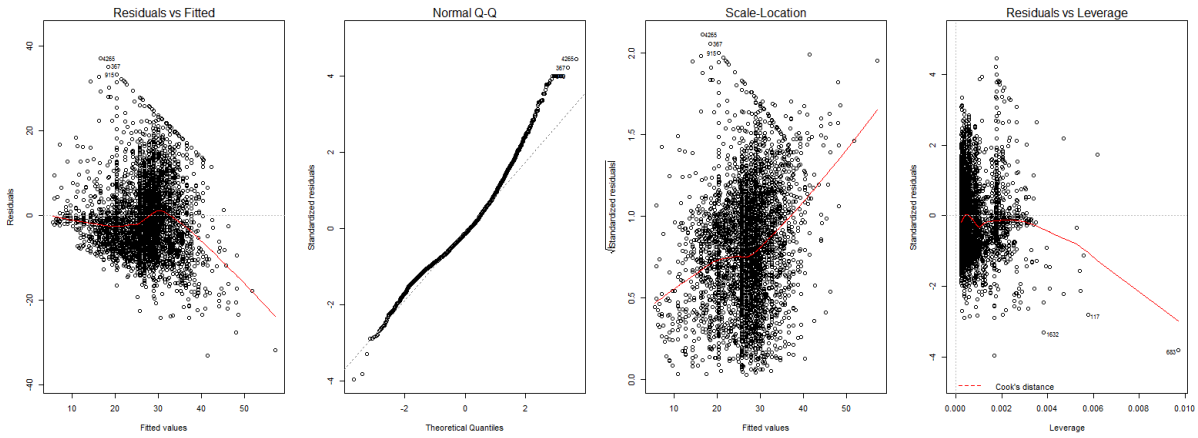
```
> summary(fit4)

Call:
lm(formula = weeklyearnings2 ~ weeklyhrsworked2 + education2,
    data = trainingsample)

Residuals:
    Min      1Q  Median      3Q     Max
-33.024  -5.757  -1.146   4.496  37.091

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.26837    0.49945   10.55   <2e-16 ***
weeklyhrsworked2 0.37842    0.01016   37.26   <2e-16 ***
education2       1.07026    0.05335   20.06   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.342 on 4343 degrees of freedom
Multiple R-squared:  0.3218,    Adjusted R-squared:  0.3215
F-statistic:  1031 on 2 and 4343 DF,  p-value: < 2.2e-16
```

I chose to drop 'maleyes2' and 'age2' as they had relatively low t values compared to the other variables and so had limited predictive power.

The linear model I have chosen ('fit4') is:

Weekly Earnings = 4.17 + 0.38 * Weekly Hrs Worked + 1.07 * Education level + 8.348 * Standard error

The models performance against the training sample was plotted in the following graphs:



"Linear Model Plots 7"

The graphs show that the model is being affected by outliers of those who earn large amounts; this follows the distribution of income in society as the top 1% of earners tend to earn dramatically more than the rest of the population. However with respect to the model this undermines its predictive capabilities.

Evaluate + Improve Model:

I then tested the predictive capabilities of the chosen model above [fit4] by training the model on 80% of the data and then testing the predictive performance on the training data and then the final 20% remaining data for validation. The measures I used to test the model where "Correlation of expected and obtained results" (Cor), "Root Mean Square Error" (RMSE) and "Mean Average Error" (MAE)

On the training data the model had the following results:

Cor: 0.3249 RMSE: 545 MAE: 370

On the validation model the data had the following results:

Cor: 0.36 RMSE: 538 MAE: 354

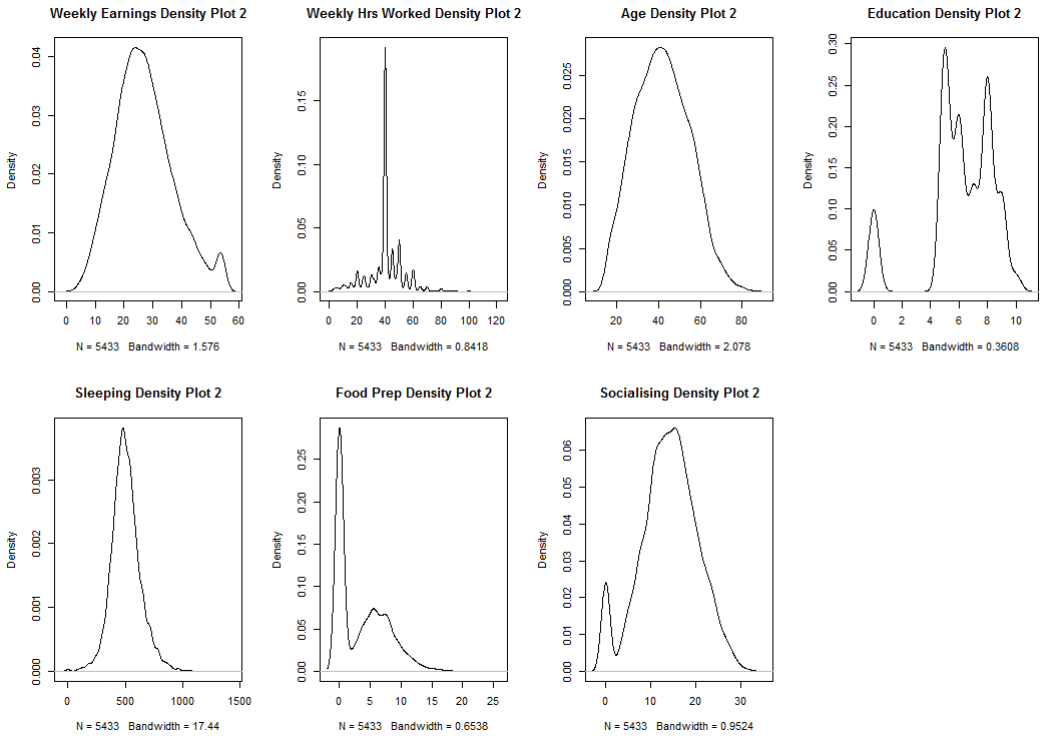
Finally to check for multicollinearity between the predictor variables I performed a VIF test on the final variables:

```
> vif(fit4)
weeklyhrsworked2       education2
        1.026714         1.026714
```

As the VIF values for both indicators are below 5, I am content that there is no multicollinearity in my model.

Appendix

"Correlation Table 8"



"Transformed Variables Density Graphs 9"