1(a) Carry out exploratory data analysis for the response variable with the predictors provided.

wage_status ~ maritl:

```
  1. Never Married 2. Married 3. Widowed 4. Divorced 5. Separated
0              446        748          9         104           36
1              202       1326         10         100           19
```

```
        Pearson's Chi-squared test

data:  earnings$wage_status and earnings$maritl
X-squared = 227.98, df = 4, p-value < 2.2e-16
```

The Chi-squared test statistic has p < 0.0001, which indicates an association between wage_status ~ maritl.

wage_status ~ race:

```
  1. White 2. Black 3. Asian 4. Other
0     1072      172       73       26
1     1408      121      117       11
```

```
        Pearson's Chi-squared test

data:  earnings$wage_status and earnings$race
X-squared = 38.224, df = 3, p-value = 2.534e-08
```

The Chi-squared test statistic has p < 0.0001, which indicates an association between wage_status ~ race.

wage_status ~ education:

```
  1. < HS Grad 2. HS Grad 3. Some College 4. College Grad 5. Advanced Degree
0          208        609             286             185                55
1           60        362             364             500               371
```

```
        Pearson's Chi-squared test

data:  earnings$wage_status and earnings$education
X-squared = 505.86, df = 4, p-value < 2.2e-16
```

The Chi-squared test statistic has p < 0.0001, which indicates an association between wage_status ~ education.

wage_status ~ health:

```
  1. <=Good 2. >=Very Good
0       468            875
1       390           1267
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  earnings$wage_status and earnings$health
X-squared = 45.921, df = 1, p-value = 1.231e-11
```

The Chi-squared test statistic has p < 0.0001, which indicates an association between wage_status ~ health.

wage_status ~ year:

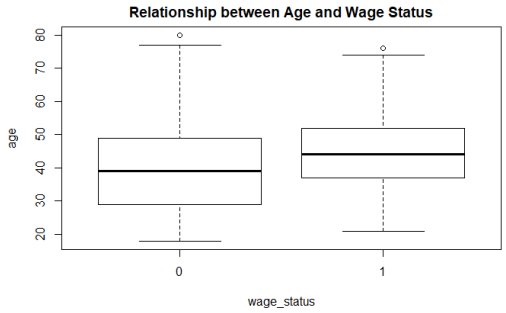
```
  2003 2004 2005 2006 2007 2008 2009
0  262  224  202  166  173  167  149
1  251  261  245  226  213  221  240
```

```
        Pearson's Chi-squared test

data:  earnings$wage_status and earnings$year
X-squared = 16.644, df = 6, p-value = 0.01068
```

The Chi-squared test statistic has p > 0.0001, which indicates **NO** association between wage_status ~ year.

wage_status ~ age:



The boxplots of both categories in the wage_status seems to have similar variances, though the boxplot of wage_status(0) is slightly skewed to its lower whisker. There is just one outlier for both categories in wage_status which is negligible.

1(b) Develop a GLM for *wage status* using the predictors provided.

Single Covariate Regression Models:

Model	Pr(>Chisq)	AIC
model_maritl	**< 2.2e-16**	3905.995
model_race	2.561e-08	4095.755
model_education	**< 2.2e-16**	**3596.223**
model_health	1.005e-11	4083.638
model_year	0.01055	4123.280
model_age	**< 2.2e-16**	4007.556

The p value for all the covariates are less than 0.2, with maritl, education and age being the most significant while education also has the least AIC score.

Multivariate Regression Models:

Model	Covariates	AIC
model1	maritl, race, education, health, year, age	**3334.571**
model2	maritl, education, health, year, age	3342.419
model3	maritl, education, health, age	3351.729
model4	maritl, education, age	3366.290
model5	maritl, education	3390.285

The variables are checked for the significance level of 0.05 and different models are fitted. The lowest AIC score is achieved by model1.

1(c) Write down your final model equation.

Log $\frac{\pi_i}{1-\pi_i}$ = - 3.671683 (Intercept) + 1.210894 (maritl2. Married) + 0.853295 (maritl3. Widowed) + 0.580225 (maritl4. Divorced) + 0.388077 (maritl5. Separated) - 0.460736 (race2. Black) - 0.325503 (race3. Asian) - 0.502806 (race4. Other) + 0.710168 (education2. HS Grad) + 1.619020 (education3. Some College) + 2.298950 (education4. College Grad) + 3.078956 (education5. Advanced Degree) + 0.381752 (health2. >=Very Good) + 0.146517 (year2004) + 0.315777 (year2005) + 0.486721 (year2006) + 0.237894 (year2007) + 0.375433 (year2008) + 0.652628 (year2009) + 0.023955 (age)

1(d) Interpret the model parameters.

An individual's wage *does not* exceed $100k per year if the individual's race is Black, Asian or Other.

The chances of an individual's wage exceeding $100k per year is highest when that individual has an Advanced Degree. The College Graduates follow closely behind with more chances of earning more than $100k per year.

The marital statuses except Never Married also have a nominal positive impact on an individual's wage exceeding $100k per year. It is surprising to see that an individual being Married has a considerable positive impact on an individual's wage exceeding $100k per year rather than an individual who is Never Married.

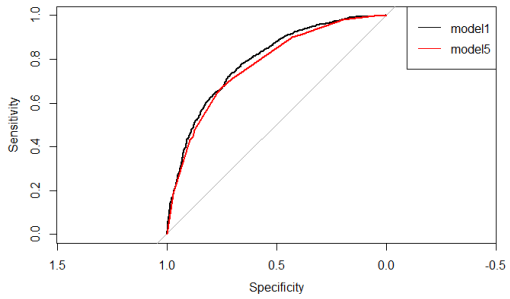
Except year 2003, the rest of the years does seem to be a token positive factor in an individual's wage exceeding $100k per year, including the recession year 2008.

Again, an individual being in Very Good health only has little positive effect on the wage exceeding $100k per year.

The age has a very negligible positive bearing on an individual's wage exceeding $100k per year.

1(e) Obtain the ROC curve for your final model and the corresponding AUC value. Are the predictions from your final model better than tossing a coin?

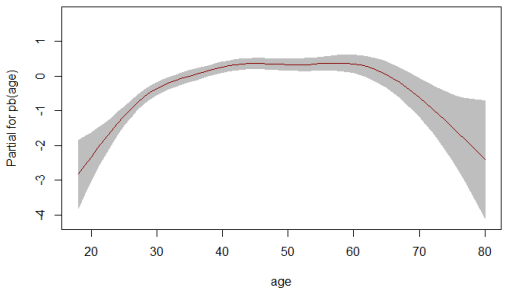
Model	Area Under Curve (AUC)
model1	**0.7906**
model2	0.7881
model3	0.7841
model4	0.7811
model5	**0.7735**



The ROC curve of model1 is almost near to sensitivity = specificity = 1 and the AUC is well above the minimum required 0.7 and is near 0.8 at 0.7906, indicating better predictive ability.

With model5 having the least AUC at 0.7735, model1 indeed produces predictions better than tossing a coin.

1(f) We are also interested in whether your final model can be improved by replacing the linear *age* term by including an additional smooth term. Using a GAM and its partial plot, investigate whether the smooth term of *age* is significant or not.



The linear part and smooth term of age looks significant.

2(a) Ignoring within-subject correlation, and using only *semantic class*, *length of recipient*, *access of rec* and *length of theme* as covariates, find a reasonable model for realization of recipient.

Single Covariate Regression Models:

Model	Pr(>Chisq)	AIC
model_semantic	0.1576	128.4834
model_recipient	2.134e-07	102.1925
model_access	2.648e-07	**100.8120**
model_theme	**8.571e-07**	104.8756

The p value for all the covariates are less than 0.2, with ‘length of theme’ being the most significant while ‘access of recipient’ has the least AIC score.

Multivariate Regression Models:

Model	Covariates	AIC
model_a	semantic class, length of recipient, access of rec, length of theme	92.69076
model_b	length of recipient, access of rec, length of theme	**86.83496**
model_c	length of recipient, length of theme	87.29820

The variables are checked for the significance level of 0.05 and different models are fitted. The lowest AIC score is achieved by model_b.

2(b) To model the correlation, add a random intercept for *speaker* i.e. the subject to your best model in (a).

```
library(lme4)

model_b_cor <- glmer(dative$realization_of_recipient ~ dative$length_of_recipient + dative$access_of_rec + dative$length_of_theme + (1|dative$speaker), family = binomial(), data = dative)
```

2(c) Compare your models in (a) and (b) using a model selection criterion.

Model	AIC
model_b	86.83496
model_b_cor	88.83496

The correlation model produces higher AIC score, indicating that the normal model is indeed the best model.