Arrests for Marijuana Possession Analysis Report

Team 1

```
knitr::opts_chunk$set(comment=" ", error=TRUE, echo=FALSE, message=FALSE, warning=FALSE)
```

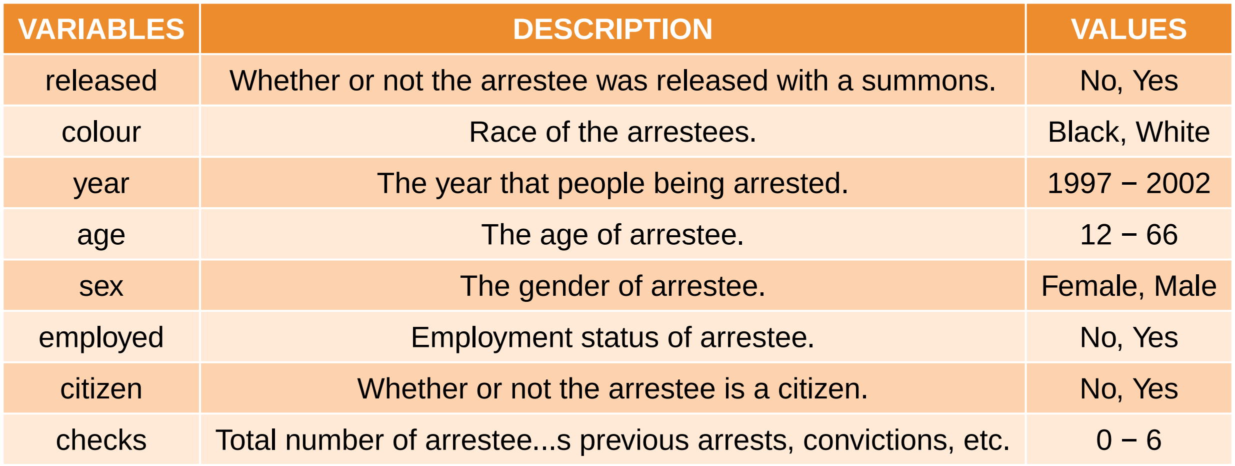
Background —-

A Toronto Star analysis of crime data shows that blacks arrested are treated more harshly than whites. Blacks have higher rate to be taken to police stations and held overnight than whites when they are facing the same charge. This reveals the social phenomenon that race matters in Canadian society and people are unfairly targeted by police. Other than race, there are definitely more factors will lead to discrimination. The dataset Arrests collected by Michael Friendly focus on police treatment of individuals arrested for marijuana possession in Toronto. Our team aim to develop a model that can reveals possible patterns of discrimination by using the dataset Arrests.

Data Visualization —-

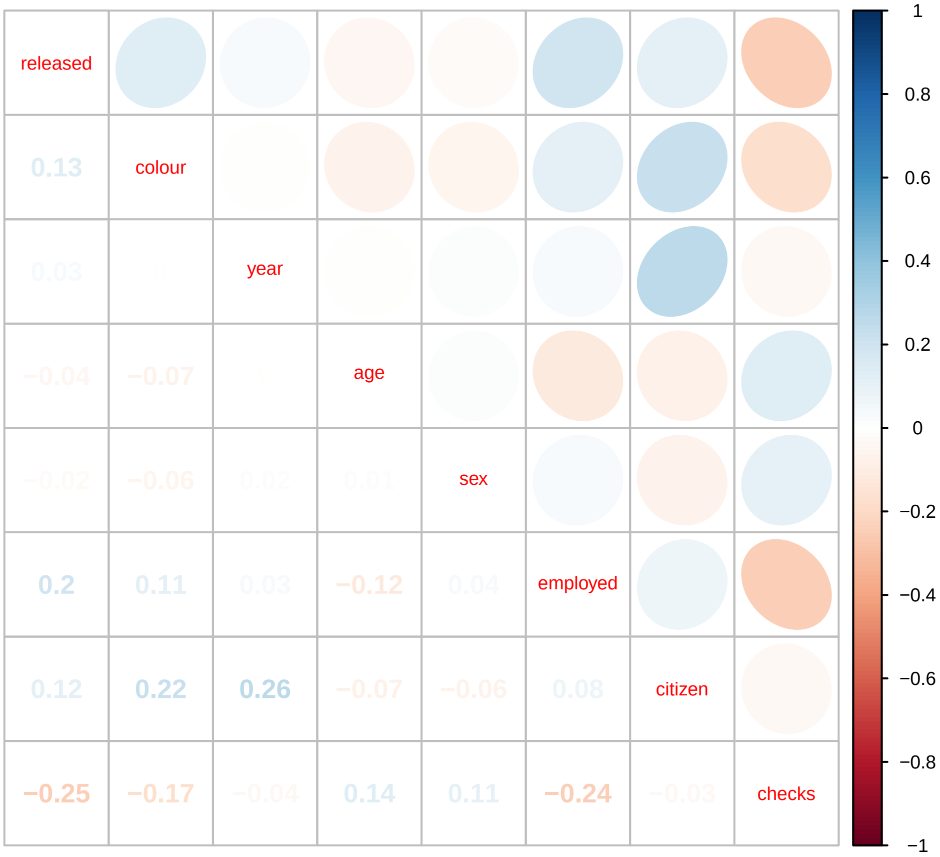
There are 5226 observations with 8 variables, and dataset not include any missing value.

Summary of Arrests data



VARIABLES	DESCRIPTION	VALUES
released	Whether or not the arrestee was released with a summons.	No, Yes
colour	Race of the arrestees.	Black, White
year	The year that people being arrested.	1997 − 2002
age	The age of arrestee.	12 − 66
sex	The gender of arrestee.	Female, Male
employed	Employment status of arrestee.	No, Yes
citizen	Whether or not the arrestee is a citizen.	No, Yes
checks	Total number of arrestee...s previous arrests, convictions, etc.	0 − 6

Correlation plot



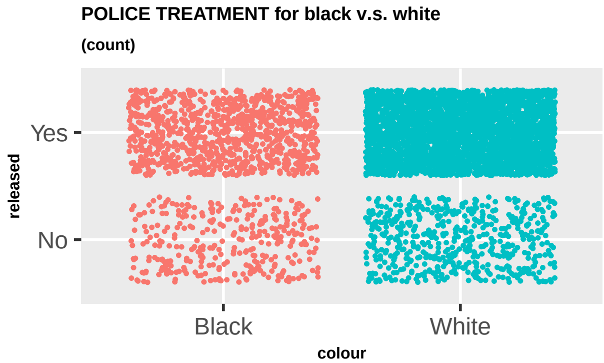
Using correlation plot rather than scatter plot to visualize the relationship between paired variables, since scatterplotMatrix does not work well for categorical data. Noted:

- Need to convert categorical data to numerical data firstly.
- Choose significance level equal to 0.05, correlation between two variables significant if it greater than the preassigned significance level.
- The correlation between released and year is 0.03 doesn't imply there's a positive relationship between them, since year should be a factor instead of a continuous variable.

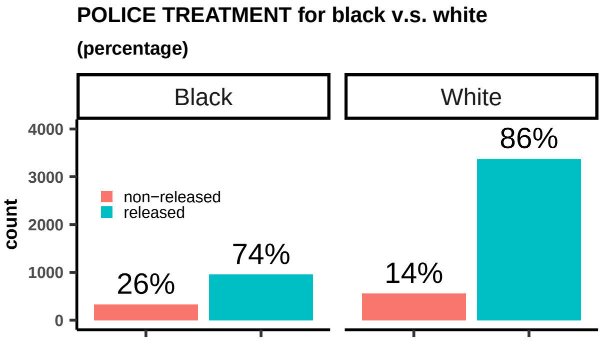
From the correlation plot, released has positive association with colour, employed and citizen, and has negative association with checks. Which means arrestees are more likely to be released if they are white, employed, citizen or less criminal record.

Expecting to build a model able to reflect this relationship, if model include them as predictor variables.

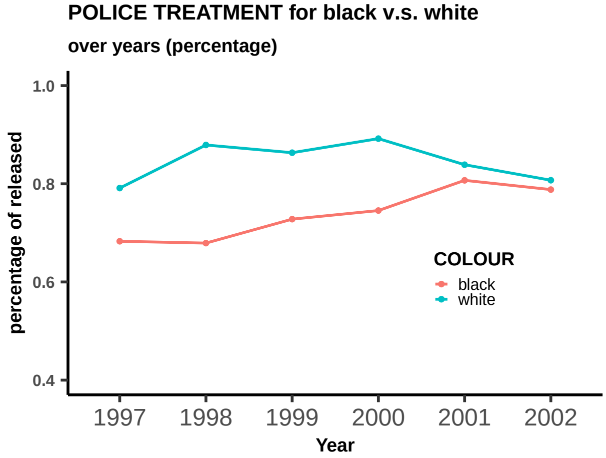
Visualizing data


POLICE TREATMENT for black v.s. white

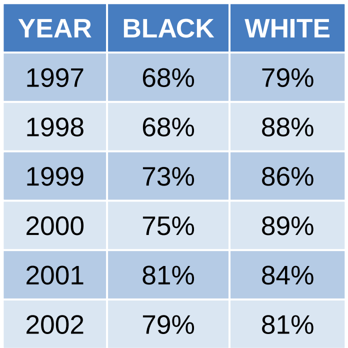
(count)


POLICE TREATMENT for black v.s. white

(percentage)


POLICE TREATMENT for black v.s. white

over years (percentage)



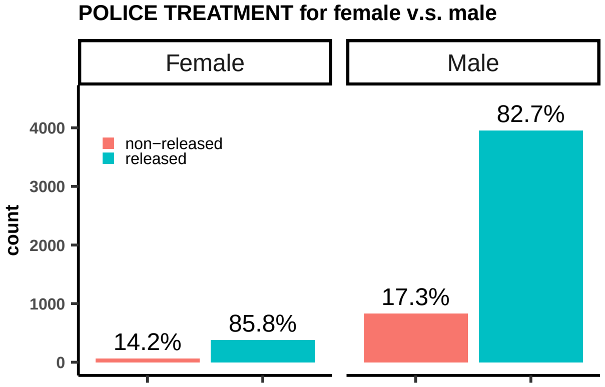
YEAR	BLACK	WHITE
1997	68%	79%
1998	68%	88%
1999	73%	86%
2000	75%	89%
2001	81%	84%
2002	79%	81%

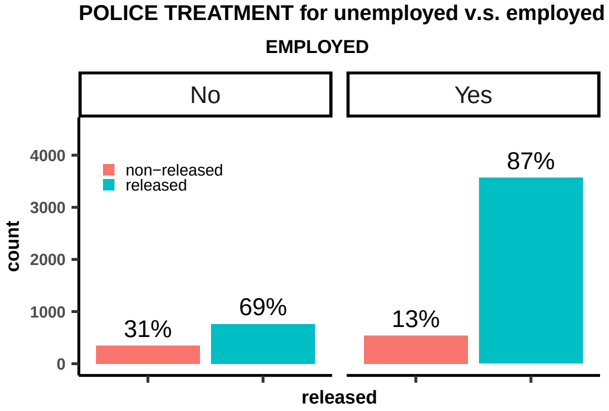
```
[1] 3.057453
```

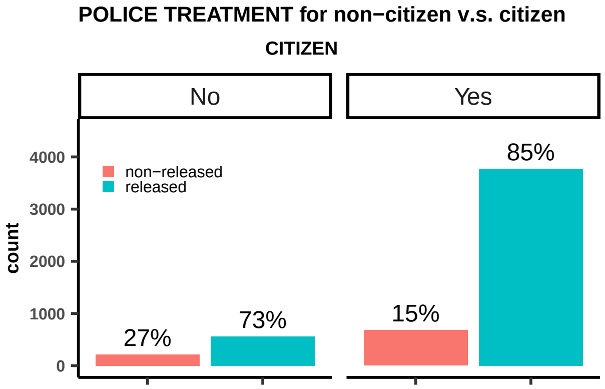
Form the scatterplot, the number for white is greater than black for both released and unreleased groups. The scatterplot may indicate different groups are treated fairly, which is not true. This might relate to the population distribution in Toronto and not indicate different groups are treated fairly. Back to year 2006, the census from Statistical Canada shows the percentage for white and black among total population are 52.5% and 8.4%. The number of white arrestees is three times the number of black arrestees roughly from year 1997-2002.

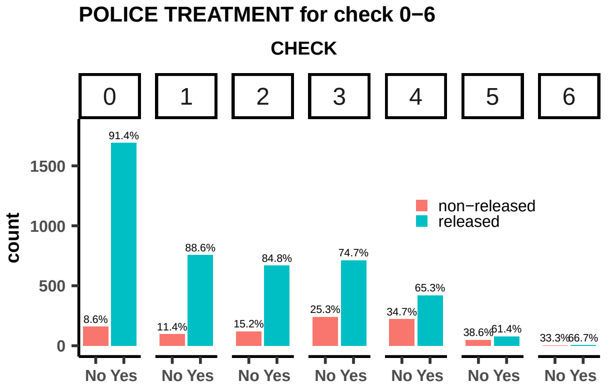
Looking at the bar plot condition on race, 86% white have been released, but only 74% black have been released. Black arrestees are less likely to be released and it reveals a social phenomenon that black racial group is treated unfairly.

For the line plot, the blue and red lines is percentage of released for white and black. The blue line is in the top of red line between year 1997-2002, the black arrestees are less likely to be released over years. The table in the right side record the specific number for the line plot.


POLICE TREATMENT for female v.s. male


POLICE TREATMENT for unemployed v.s. employed


POLICE TREATMENT for non−citizen v.s. citizen


POLICE TREATMENT for check 0−6

Among all graphs, there's a sharp gender imbalance for both released and non-released arrestees. We could consider combining male and female to avoid selection effect. And compare the models before and after combine to see which model fit better.

Build models —-

Taking employed, citizen and checks as mediator factors or confounding factors during the building model process, mainly consider those two situations in different c ondition assumptions.

CASE 1: Consider employed, citizen and checks as mediate factors

From the social and ethnic perspective, it is commonly believed that racism exists in hiring, immigration and judgement process. In other word, the color of skin will have an effect on the possibility a person to be hired, to become a citizen and to be judged guilty. In the meantime, employed, citizen and checks will influence the decision of police whether or not the arrest should be released. For example, black people have a lower chance to enter the job market due to racial discrimination, police might think they have more crime motive since they do not have stable income, so they are less likely to be released.
Under this circumstance, employed, citizen and checks are considering as mediate factors.

1. without combine male and female —-

ID	MODEL	FORMULA	AIC
1	model1_1	released ~ colour* (age+ year+ sex)	4665.5
2	model1_2	released ~ colour* (age+ year)+ sex	4665.4
3	model1_3	released ~ colour* year+ age+ sex	4680.9
4	model1_4	released ~ colour* age+ year+ sex	4671.7

LRT	P_VAL	RESULT	CHOICE
1 v.s. 2	0.1694	interaction between colour and sex is not significant	model1_2
2 v.s. 3	2.782e−05	interaction between colour and age is significant	model1_2
2 v.s. 4	0.00402	interaction between colour and year is significant	model1_2

All coefficients of model1_2 are significant except sex. Consider combining male and female, since the number of female is too little for both released and non-released arrestees.

2. combine male and female —-

ID	MODEL	FORMULA	AIC
1	model2_1	released ~ colour* (year+ age)	4648.9
2	model2_2	released ~ colour* year+ age	4663.9
3	model2_3	released ~ colour* age+ year	4663.7

LRT	P_VAL	RESULT	CHOICE
1 v.s. 2	3.793e−05	interaction between colour and age is significant	model2_1
1 v.s. 3	0.0001516	interaction between colour and year is significant	model2_1

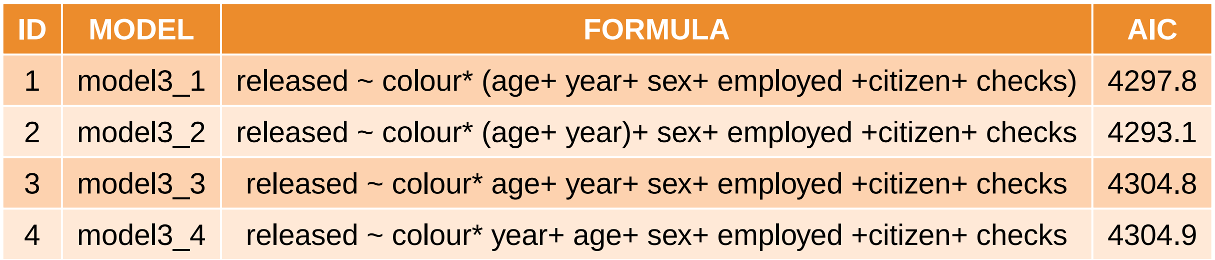
For case 1, choose model2_1 according to smaller AIC criterion.

The effect of colour:
1.495160 - 0.038089 $*$ age + 0.661839 $*$ year1998 + 0.298193 $*$ year1999 + 0.460083 $*$ year2000 -0.348932 $*$ year2001 - 0.479314 $*$ year2002

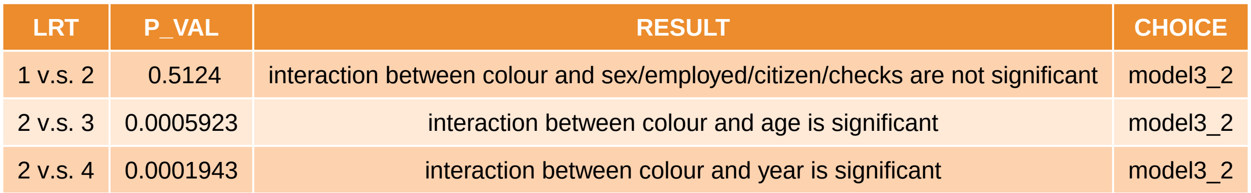
CASE 2: Consider employed, citizen and checks as confounding factors

Firstly, taking variable “employed” as an example to make a process of reasoning If a person is arrested by the police, the confounding variable(employed) will have an impact on his skin color and release rate at the same time. If the person is not working, it will have an adverse impact on his skin color and release rate. It sounds a little weird, because “employed” affects the color of the person’s skin. In this condition, “employed”, “citizen” and “checks” are supposed to be the confounding variables. They have influence both on dependent and independent variables.

1. without combine male and female —-

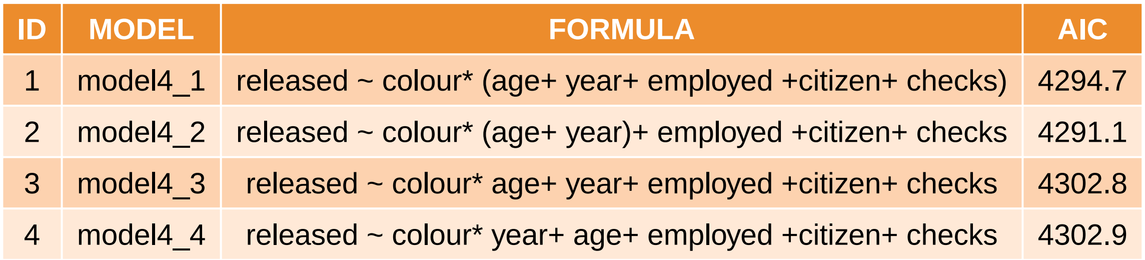


ID	MODEL	FORMULA	AIC
1	model3_1	released ~ colour* (age+ year+ sex+ employed +citizen+ checks)	4297.8
2	model3_2	released ~ colour* (age+ year)+ sex+ employed +citizen+ checks	4293.1
3	model3_3	released ~ colour* age+ year+ sex+ employed +citizen+ checks	4304.8
4	model3_4	released ~ colour* year+ age+ sex+ employed +citizen+ checks	4304.9

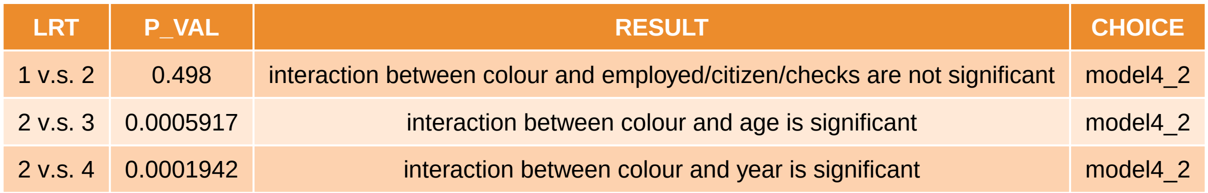


LRT	P_VAL	RESULT	CHOICE
1 v.s. 2	0.5124	interaction between colour and sex/employed/citizen/checks are not significant	model3_2
2 v.s. 3	0.0005923	interaction between colour and age is significant	model3_2
2 v.s. 4	0.0001943	interaction between colour and year is significant	model3_2

2. combine male and female —-



ID	MODEL	FORMULA	AIC
1	model4_1	released ~ colour* (age+ year+ employed +citizen+ checks)	4294.7
2	model4_2	released ~ colour* (age+ year)+ employed +citizen+ checks	4291.1
3	model4_3	released ~ colour* age+ year+ employed +citizen+ checks	4302.8
4	model4_4	released ~ colour* year+ age+ employed +citizen+ checks	4302.9



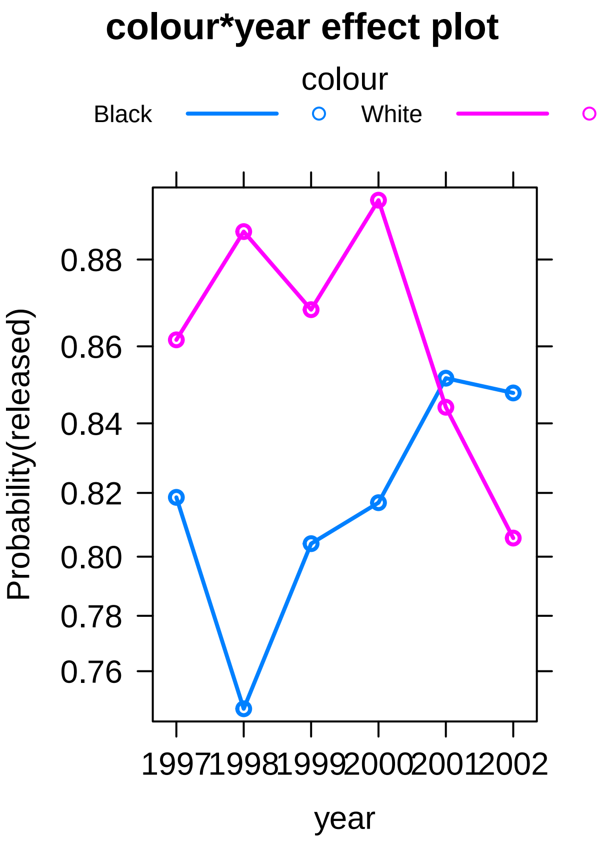
LRT	P_VAL	RESULT	CHOICE
1 v.s. 2	0.498	interaction between colour and employed/citizen/checks are not significant	model4_2
2 v.s. 3	0.0005917	interaction between colour and age is significant	model4_2
2 v.s. 4	0.0001942	interaction between colour and year is significant	model4_2

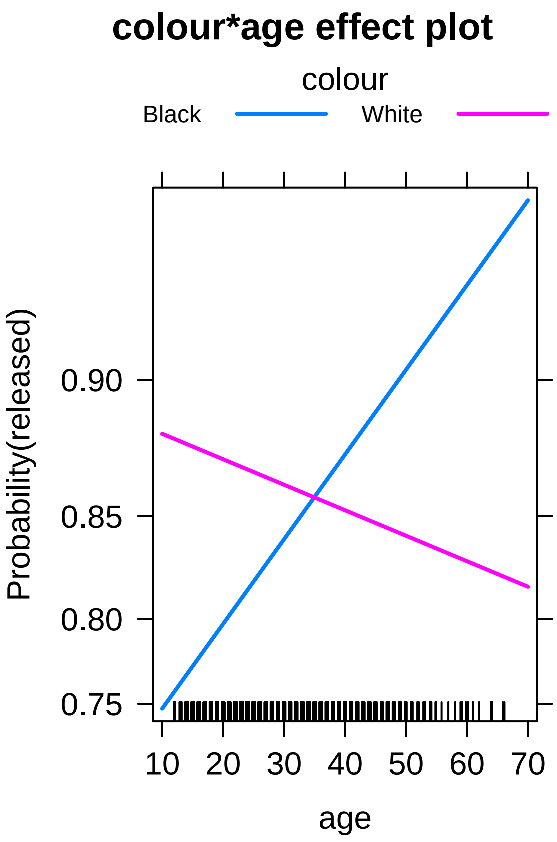
For case 2, choose model4_2 according to smaller AIC criterion.

The effect of colour:
$1.212517 - 0.037373 \ast \text{age} + 0.651956 \ast \text{year1998} + 0.155950 \ast \text{year1999} + 0.295754 \ast \text{year2000} -0.380541 \ast \text{year2001} - 0.617318 \ast \text{year2002}$

Analysis —-

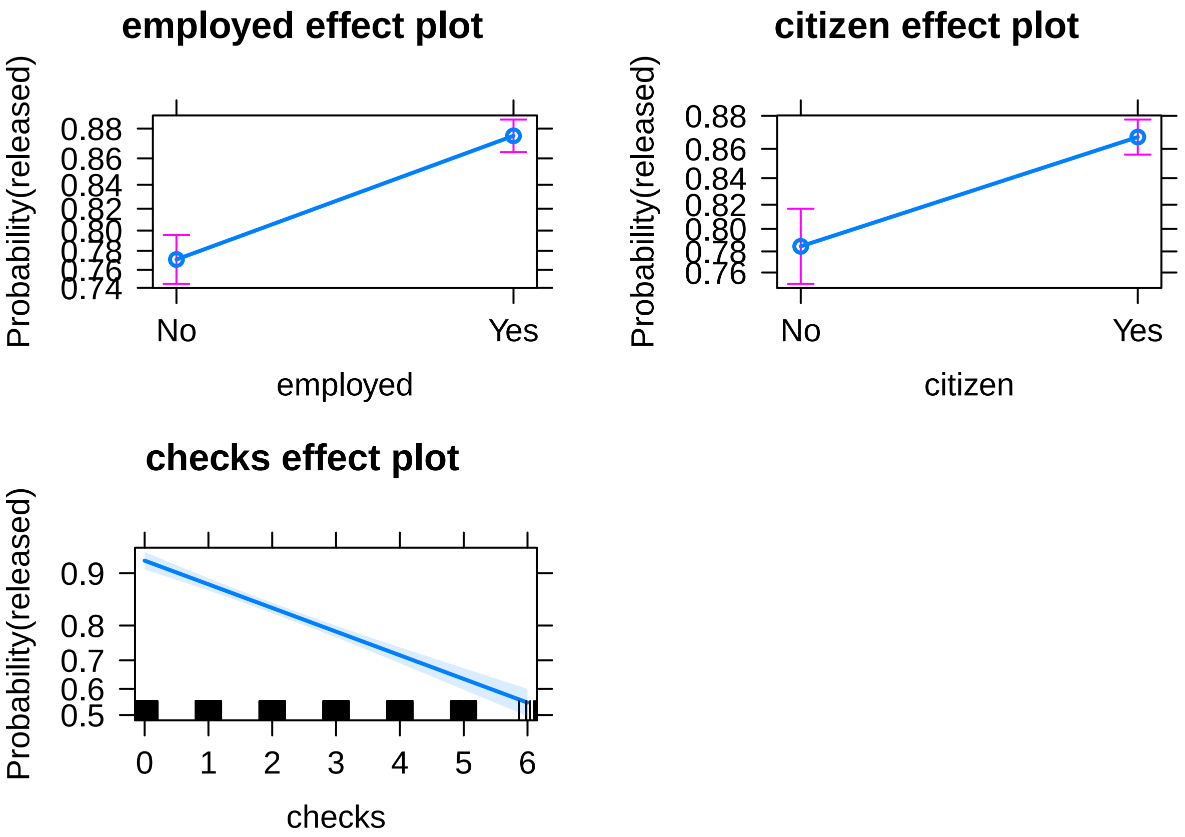
model4_2: released ~ colour* (age+ year)+ employed +citizen+ checks





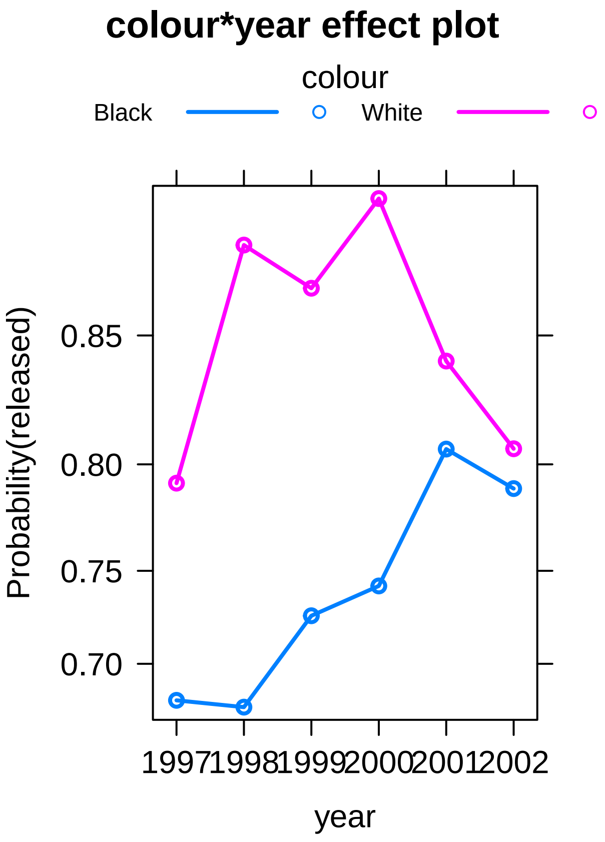
The colour-year plot shows that the probability of being released based on races as time changed. The blue and pink lines represent the probability of being released of black and white people, respectively. As shown from the graph, there is a huge gap between these two races in 1998, and it gradually becomes narrower over the course of 4 years. The two lines finally intersect in the second half of 2001, and suddenly the relationship between black and white races is reversed since then. It implies that starting from 1999, the discrimination towards the black race had gradually reduced to a point where the government viewed the both races more equally, likely due to various Right Acts regarding race discrimination being established at the time.

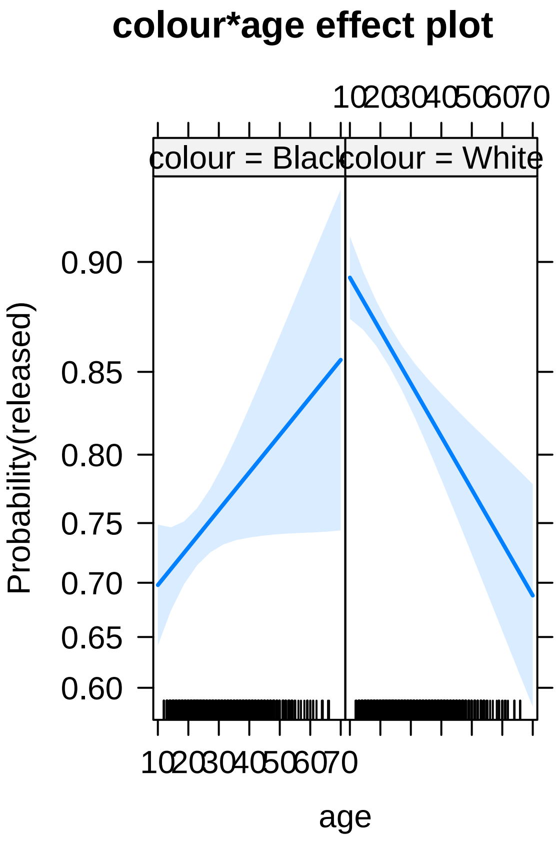
Whereas the colour-age plot reveals that the police had a bias in releasing more white young arrestees than young black arrestees. However, there is an intersection of the two lines at around age 35. We can assume that there was a flip when the criminals are in their middle age.



The three effects plot demonstrate the three confounding factors. The employed effect plot represents how the Toronto police decided whether to release a criminal with and without employment. The result shows that employed criminal had a much higher chance to be released by the police, controlling for other predictors. Having a stable job could somehow imply social responsibilities and could possibily make more contributions to the society than unemployed, and this might be one of the reasons that the policed favored in releasing employed people.The last two plots examine the relationship between the probability of release and the number of arrests one had previously experienced, and the relationship between the probability of release and the number of arrests one had previously experienced, respectively. Obviously, the Toronto police had a more positive attitudes towards Canadian citizens with less checks.

model2_1: released ~ colour* (year+ age)





This second model generates a genuinely different results from the last one. After eliminating the mediating factors, the probability of releasing the whites remained generally stable at a much higher level than the black. However, one common characteristic that both models share is that more blacks are gradually getting higher probability to be released in a relatively rapid pace. Hence, without taking the factors of whether the person arrested is employed, a citizen or received many checks into account, the chance of white arrestee being released remained high. Comparing with the first model, the colour-age plot of model 2 is quite similiar. What is different is the position of the intersection point with nearly the same possibility. Without those mediators, the intersection moved horizontally to the age of 40s. However, this difference does not contribute much extra information to the interpretation of this model. Therefore, a final conclusion of the patterns of discrimination will be given in the next section.

Conclusion —-

According to the logistic regression models and effect displays, racial discrimination manifests itself in the analysis results. For colour*age and colour*year effect plots, blacks and whites have different possibilities of being released and this relationship between released and race along with age changes over time.

As time goes by, black arrestees are more likely to be released than the white since the released probability for white was decreasing from 1997 while the rate of black people's release is increasing. That means the behaviour of the police varies over time.

The data also reveals other types of discriminations: age discrimination, discrimination against employed status and citizenship status as well as total number of arrestee's previous arrests. Specifically, an arrestee who is an employed citizen with no checks has more probabilities of being released.