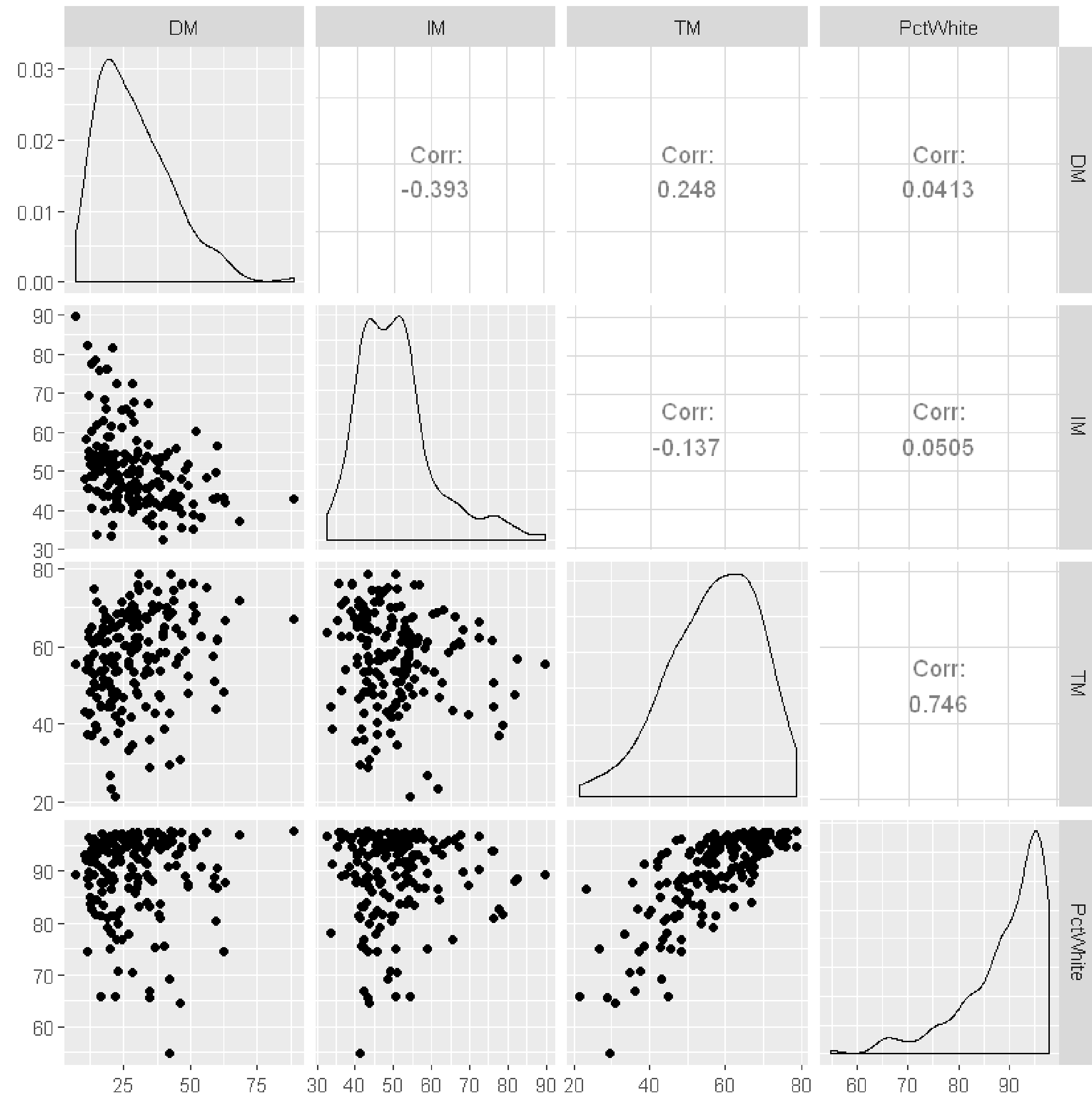
Q 1

In [8]:
```
library(ggplot2)
library(GGally)
library(broom)
library(tidyr)
```

In [10]:
```
df = read.table("rustdrugs2016.txt", header = T)
df$DM = 100000*df$Deaths/df$Population
df$IM = df$Income/1000
df$TM = 100*df$Trump
dfw = df[, c("DM","IM","TM","PctWhite")]
head(dfw)
```

DM	IM	TM	PctWhite
15.01998	45.073	71.55178	93.7
11.51526	45.808	37.26064	74.5
30.02192	45.145	68.72632	96.5
21.71616	54.548	21.41932	65.9
16.26358	53.375	44.71541	87.9
14.52600	78.487	39.80224	81.7

In [11]:
```
ggpairs(dfw)
```



ggpairs plot observations:

Skewness:

• The Explanatory variable PctWhite (Percentage of White population) is extremely left skewed. This implies that most of the counties are majorly white populated.

• The Death rate and Income are fairly right skewed. This can be easily understood since the death rates are going to relative smaller values.

Outliers:

The graph between Income and Trump percentage has a fair number of outliers. There are a few counties which irrespective of the incomes are either extreme supporters of Trump or extreme haters of Trump.

Multicollinearity:

It can be observed from the graph between Percentage of white population and Trump supporters that there is clearly a linear trend. With increasing percentage of white population in a county, the support for trump increases.

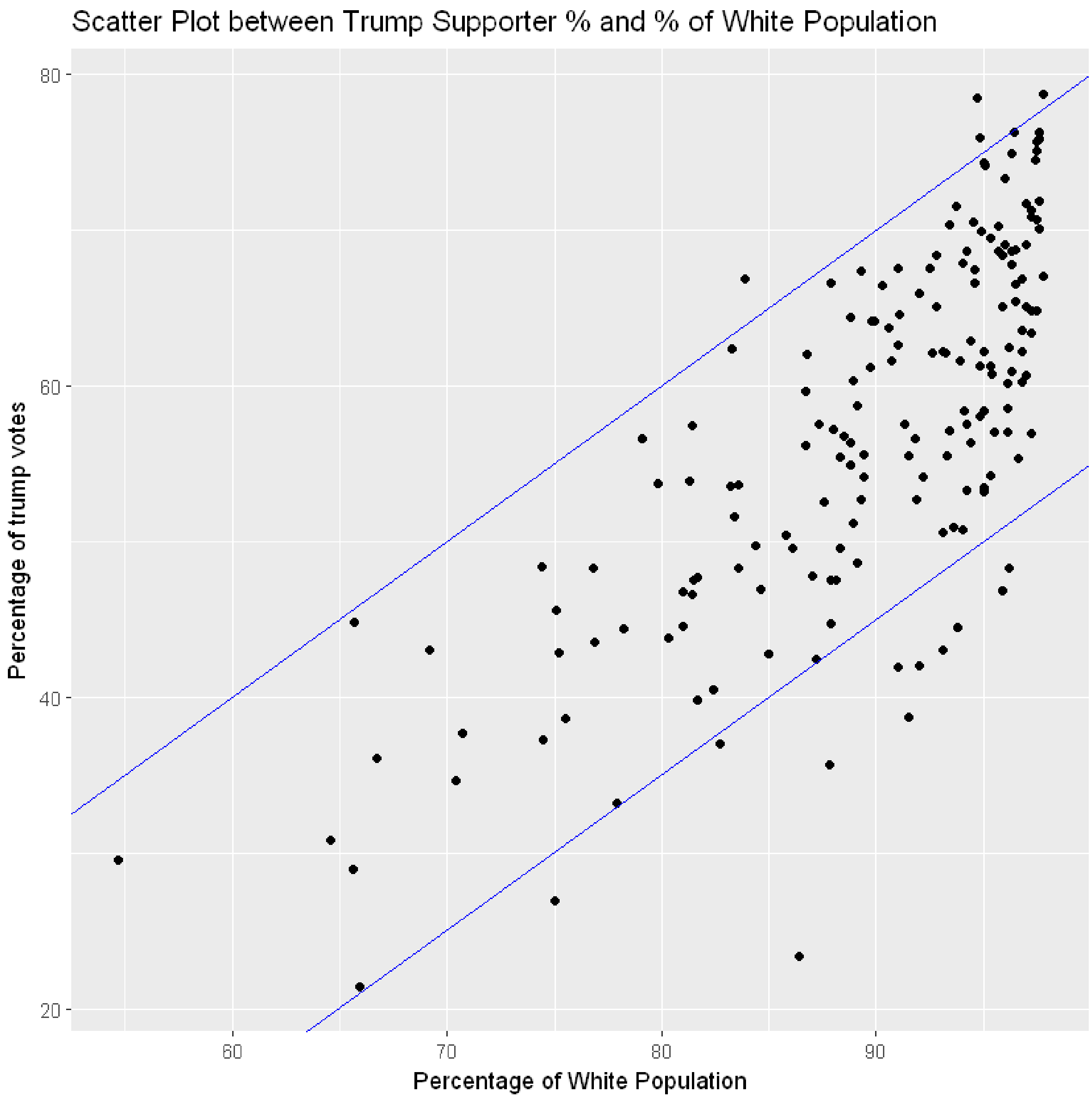
High Correlation:

Among all the variable, there exists a strong correlation only between Trump and PctWhite variable.

Q 2

We are taking Trump as a variable for our model since PCTWhite values are left skewed.

In [16]:
```
ggplot(dfw, aes(y= TM, x=PctWhite)) + geom_point() +
  geom_abline(intercept = -20 , slope = 1, color = "blue") +
  geom_abline(intercept = -45 , slope = 1, color = "blue") +
ggtitle("Scatter Plot between Trump Supporter % and % of White Population") +
 ylab("Percentage of trump votes") +
xlab("Percentage of White Population")
```

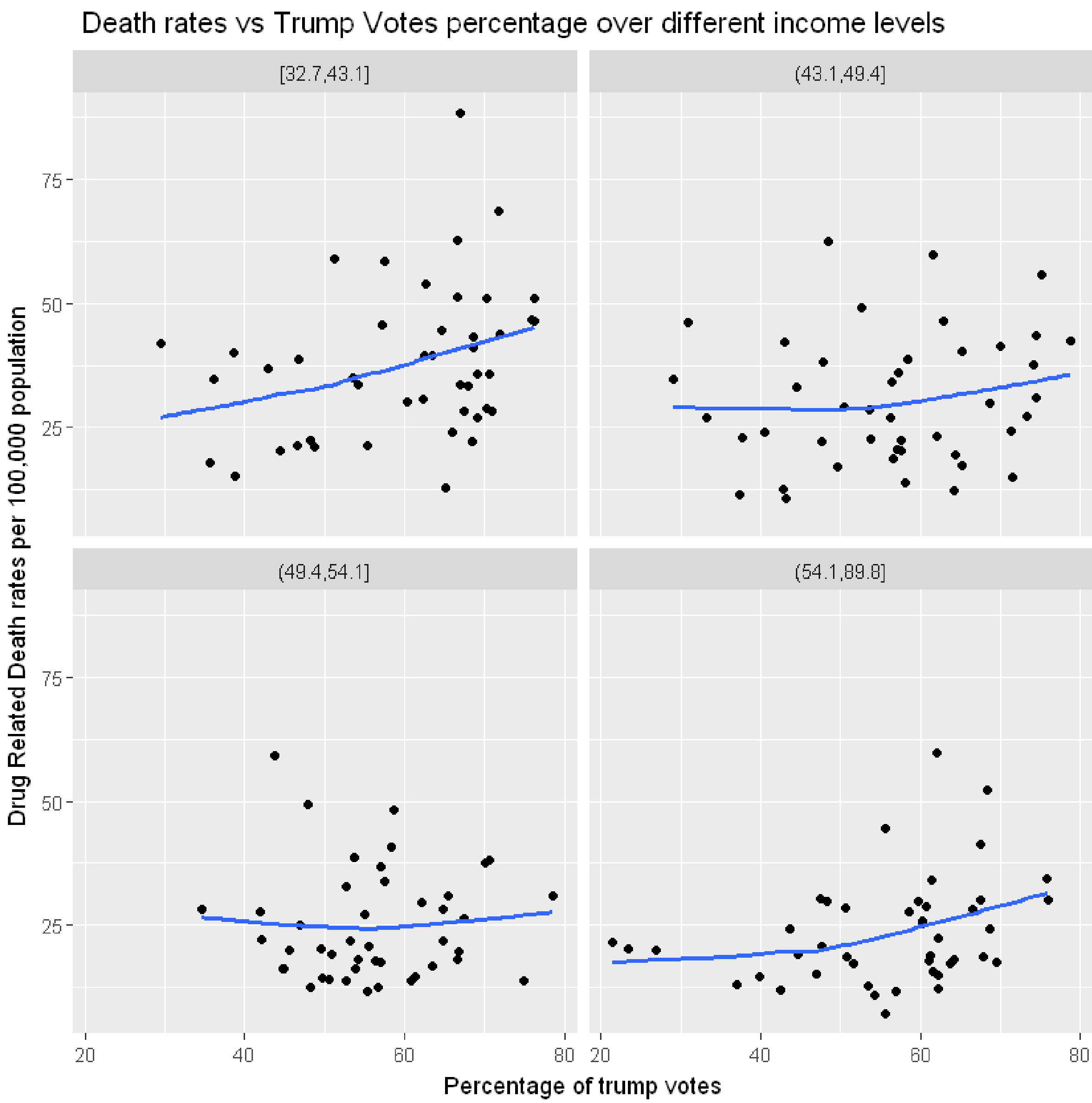


It can be observed from the ggpairs graph that there is a strong correlation of 0.746 between Trump and Percentage of White population. Hence we choose Trump from the two as explanatory variable for the response variable.

We are taking Trump as a variable for our model since PCTWhite values are left skewed.

In [13]:
```
ggplot(dfw, aes(y= DM, x=TM)) + geom_point() +
  geom_smooth(span = 1,formula = y ~ x, method.args = list(degree = 1),se = FALSE) +
  facet_wrap(~cut_number(IM, n =4)) +
  ggtitle(" Death rates vs Trump Votes percentage over different income levels")+
  ylab("Drug Related Death rates per 100,000 population ") +
  xlab("Percentage of trump votes")
```

```
`geom_smooth()` using method = 'loess'
```



In [14]:
```
ggplot(dfw, aes(y= DM, x=TM)) + geom_point() +
  geom_smooth(span = 1,formula = y ~ x, method.args = list(degree = 1),se = FA
LSE) +
  facet_wrap(~cut_number(IM, n =3) + ~cut_number(PctWhite, n =3)) +
  ggtitle("Left to right: Increasing % of White Population; Top to bottom: Inc
reasing Income")+
  ylab("Drug Related Death rates per 100,000 population ") +
  xlab("Percentage of trump votes")
```

```
`geom_smooth()` using method = 'loess'
```

We can observe that the one-way faceted plot follows a particular trend of Death Rate and the two-way faceted plot does not have any particular trend for building a model. So, we consider only one-way faceted plot.

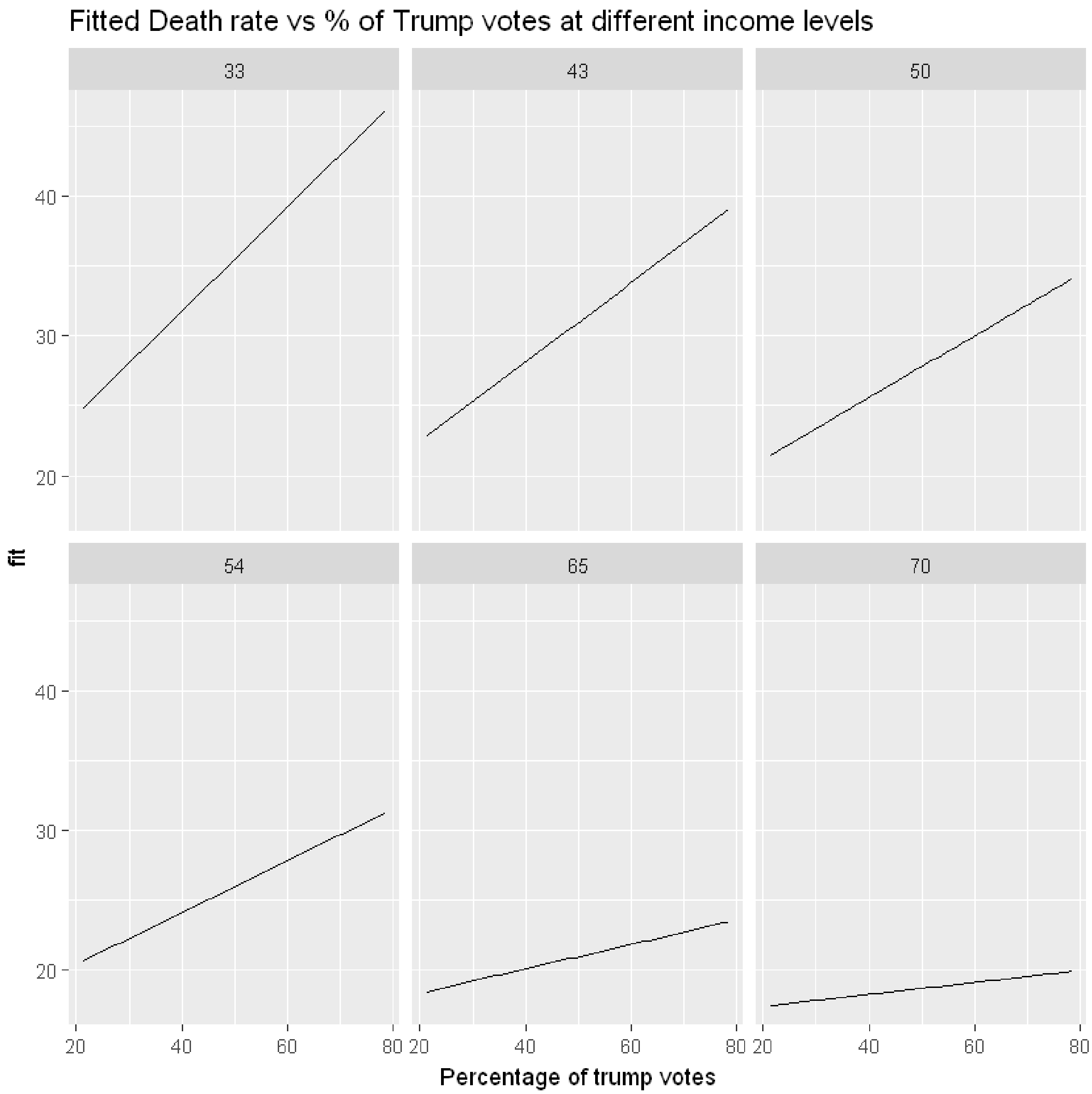
Hence Income and Trump interaction is needed to explain our model

From the one-way faceted plot we may say that death rate is increasing with Percent of votes for Trump. But we need to fit a model to confirm this trend.

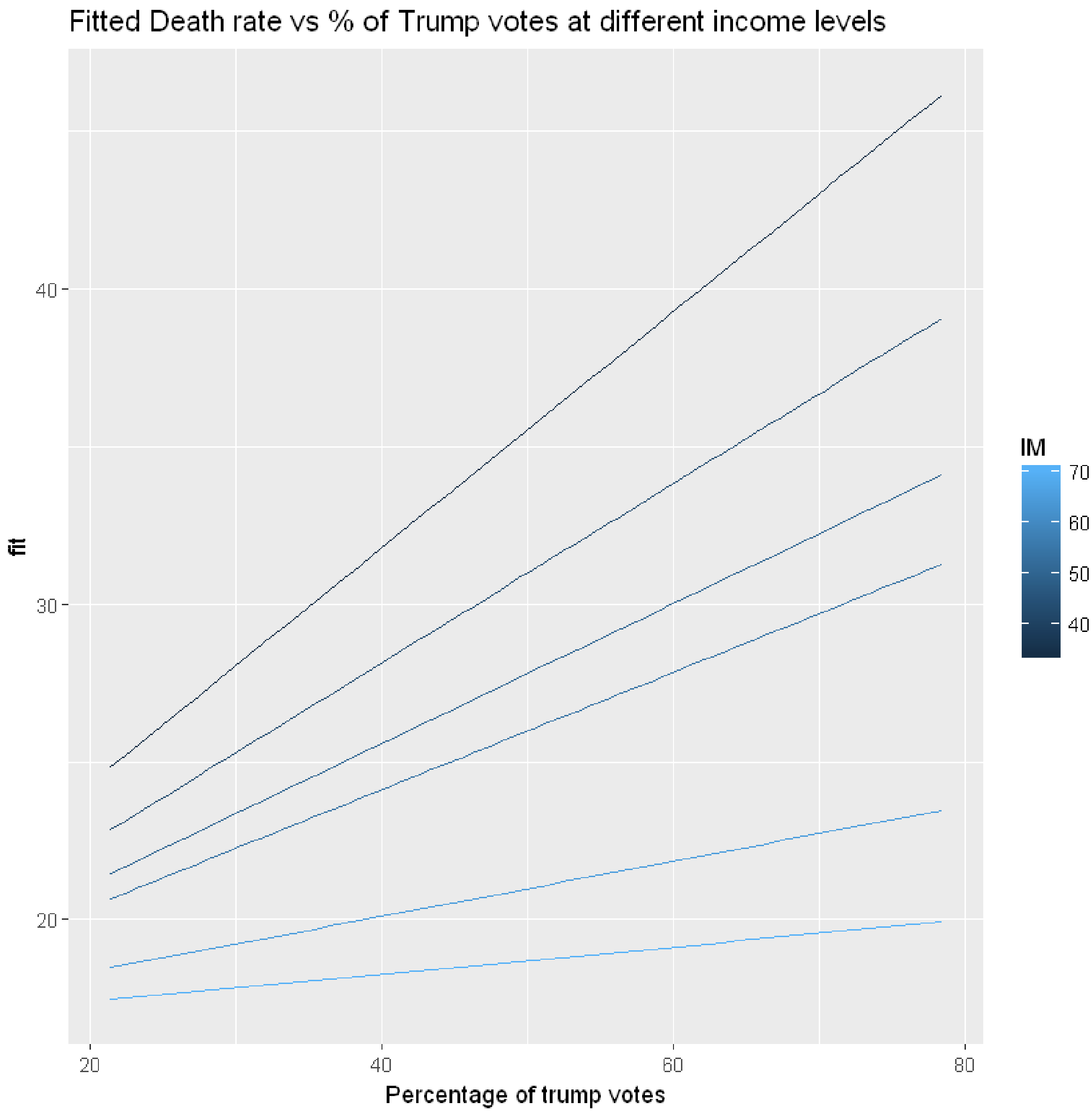
Q 3

In [18]:
```
attach(dfw)
dfw.lm = lm(DM ~ IM*TM)
lm.grid = expand.grid(TM= min(TM):max(TM), IM= c(33,43,50,54,65,70) )
DM.predict = predict(dfw.lm,lm.grid)
df.DM = data.frame(lm.grid, fit = as.vector(DM.predict))
```

In [19]:
```
ggplot(df.DM, aes(y= fit, x= TM)) + geom_line() + facet_wrap(~IM, drop = FALSE
)+
  xlab("Percentage of trump votes")+
  ggtitle("Fitted Death rate vs % of Trump votes at different income levels")
```



In [20]:
```
ggplot(df.DM, aes(y= fit, x= TM, group = IM,color = IM)) +
  geom_line() +xlab("Percentage of trump votes")+
  xlab("Percentage of trump votes")+
  ggtitle("Fitted Death rate vs % of Trump votes at different income levels")
```



For higher income levels the death rate has a little effect from the percentage of votes that went to Trump in 2016 election. But for lower income levels percentage of votes for Trump has a significant effect on the Death rate.

We have analyzed the fitted model for different income levels. For a median income higher than 70, the trend is drastically changing. So, we have filtered out these values as these comprise only 5% of the observations.