Problem Set 5

Ashwed Patil, Vatsal Jatakia, Saniya Ambavanekar, Akshay Naik (Team Hong Kong)

February 27, 2018

Introduction

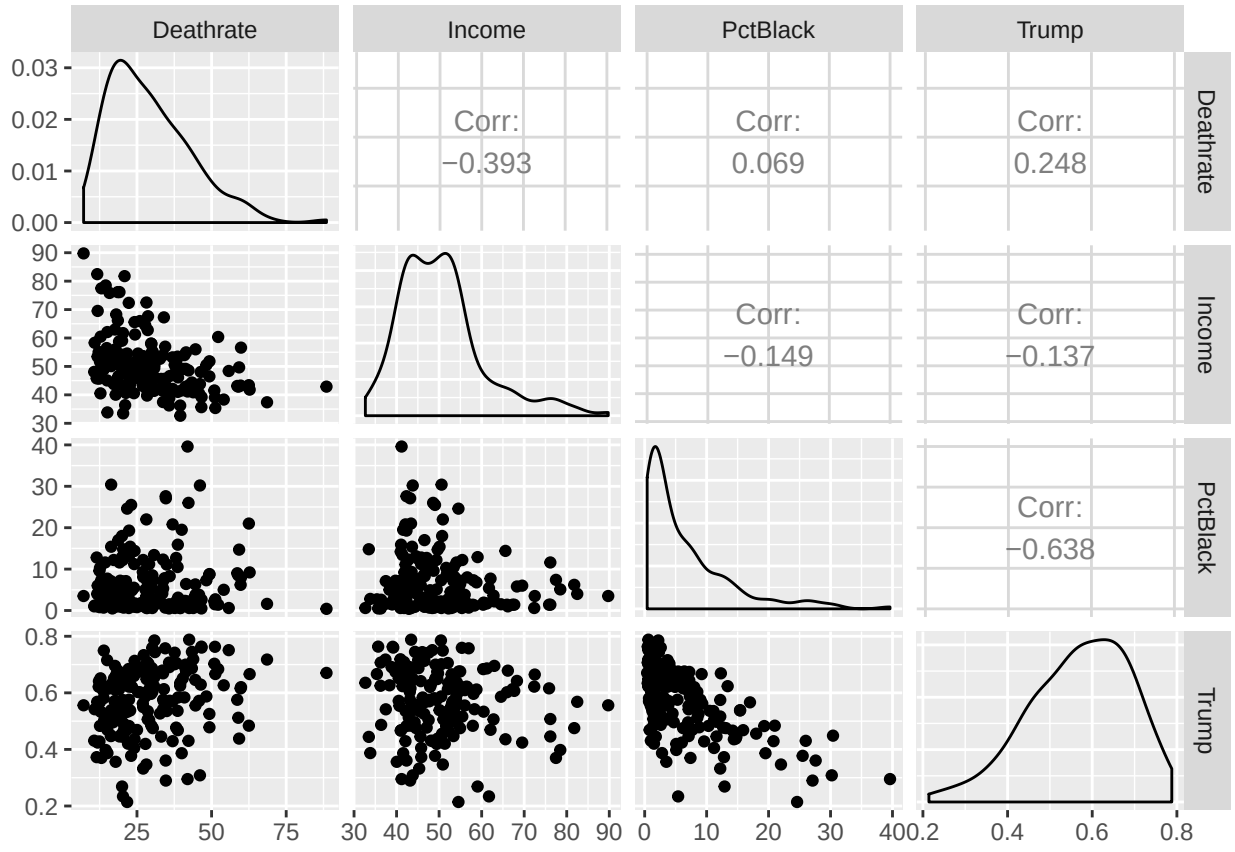
We are investigatin drug related deaths in 2016 in the East North Central Region using county level data from six states (Ohio, Michigan, Indiana, Illinois, and Wisconsin). The goal is to fit a model that explains variation in drug-related deaths in these counties in 2016.

1. Create a variable DeathRate giving the rate of drug-related deaths per 100,000 population. Put this in a data frame with three numerical predictors:

- **Income (might be better to express this in thousands of dollars rather than dollars)**
- **Trump percentage (percentages are more widely understood than proportions)**
- **One other quantitative variable of your choice.**

Draw a pairs plot of your data and comment on potential problems with model-building (outliers, extreme skewness, multicollinearity), or say "it's all good" if there are no such problems.

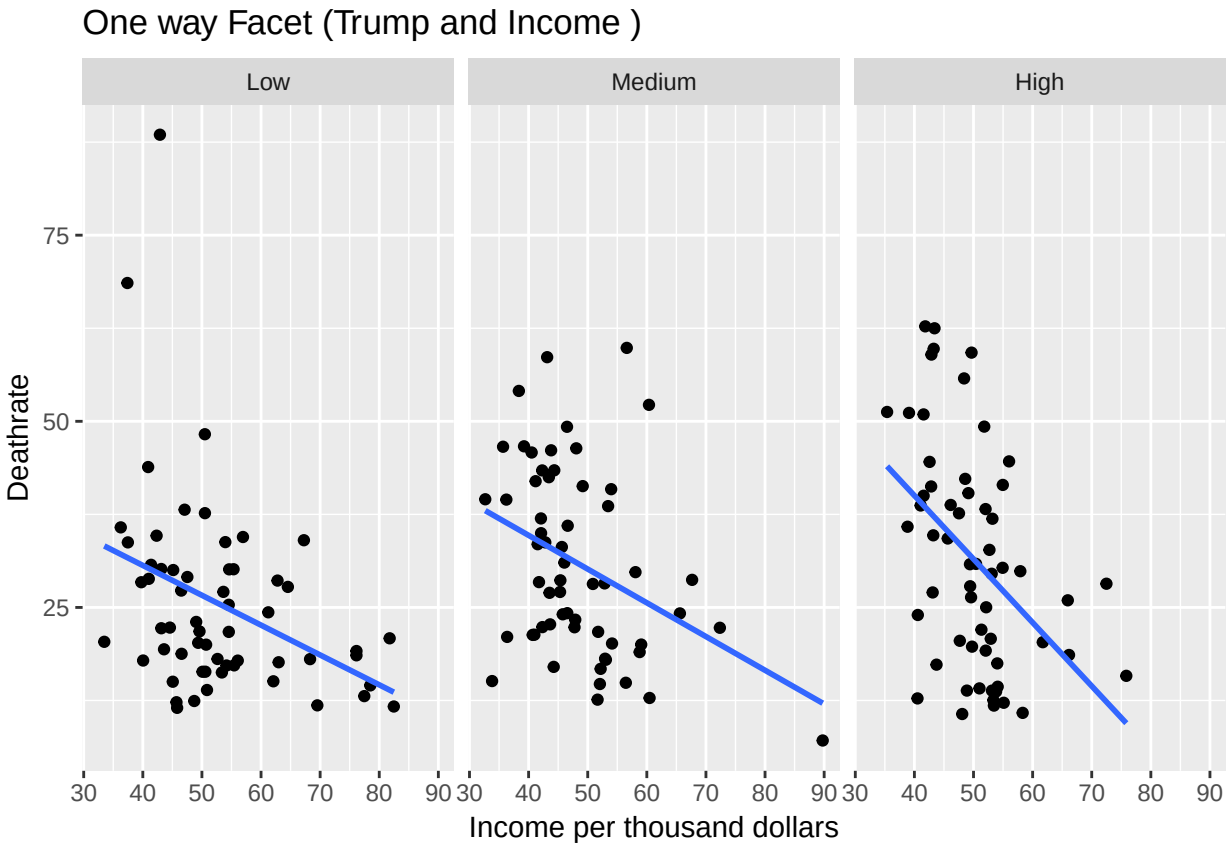
We decided to pick the *PctBlack* variable that gives the percentage of the population who are black. Our new data frame has four variables - DeathRate, Income, Trump and PctBlack.



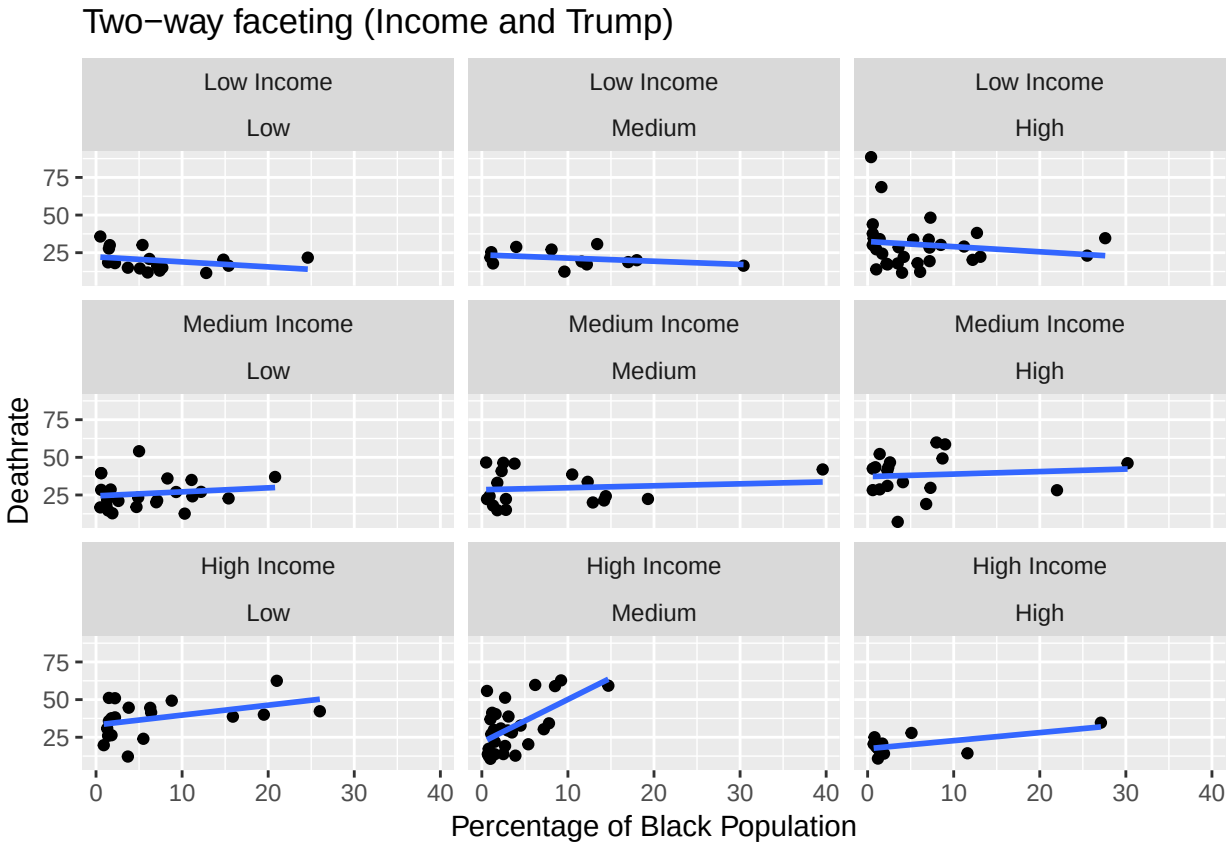
The pair plot shows that Deathrate is related to all the three variables. There isn't any collinearity between these variables, so we will keep all the predictors.

The plot for PctBlack suggests a high skewness to the right. The deathrate and income variables are also skewed a little to the right and the income density plot has two modals.The trump variable is skewed to the left. There is a high negative correlation between Trump and PctBlack.

2. We wish to determine whether our model requires an income:Trump interaction. Draw sets of one-way and two-way faceted plots to graphically examine this potential interaction. Explain what your plots tell you (or don't tell you) about the need for this interaction in your model.



From the above one way faceted plot we can observe that there is change in slope for each interaction of percentage of votes given to trump and Income. Hence it looks like there is an interaction present between the two. As we can see that the slope is decreasing in each facet that is deathrate decreases as income increases, a steep slope is observed within High percentage of votes given to trump facet.



From the above graph we can observe that there is change in slope for every facet hence there is mixture of results/observations we get:

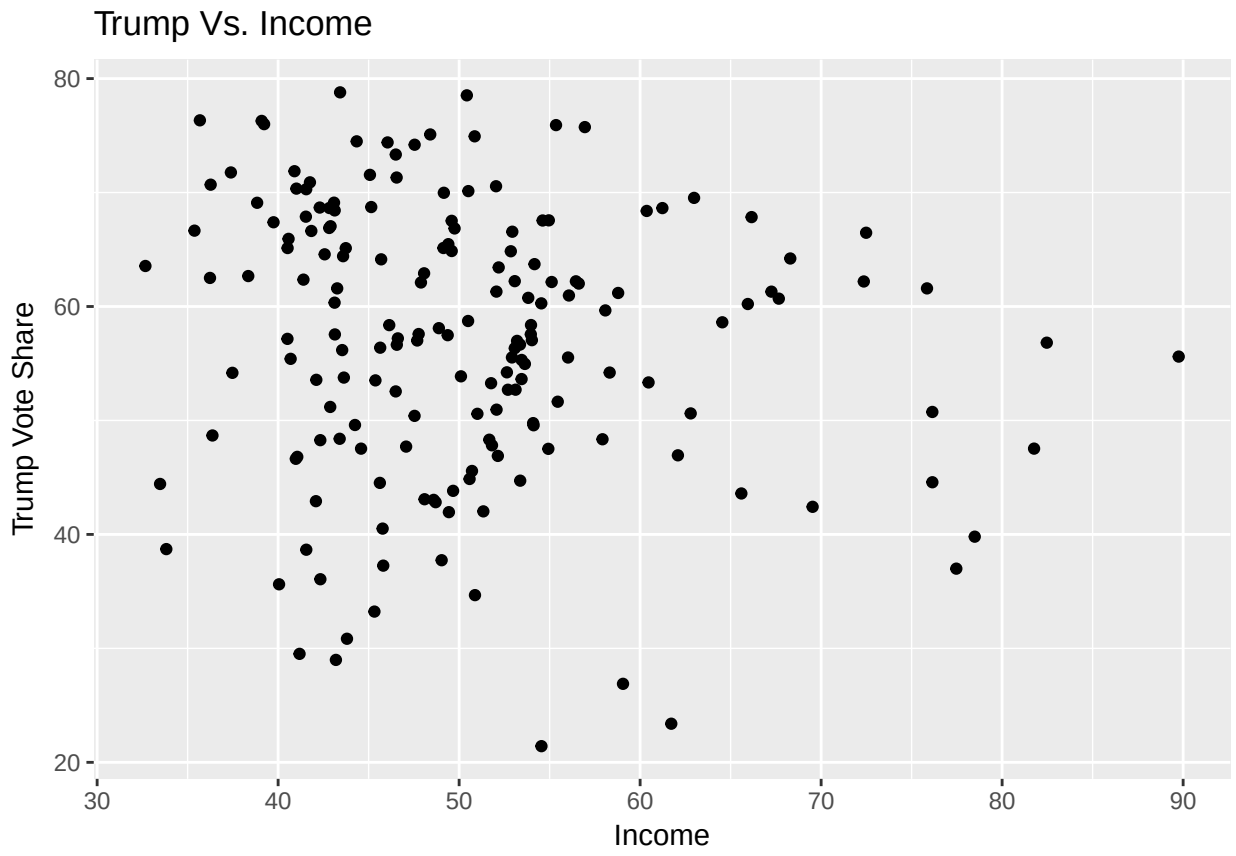
-**1.With Low Income and Variation in percentage of votes given to trump(Low, Medium, High)**: The deathrate of black population tends to decrease.

-**2.With Medium Income and Variation in percentage of votes given to trump(Low, Medium, High)**: The deathrate of black population tends to remain constant

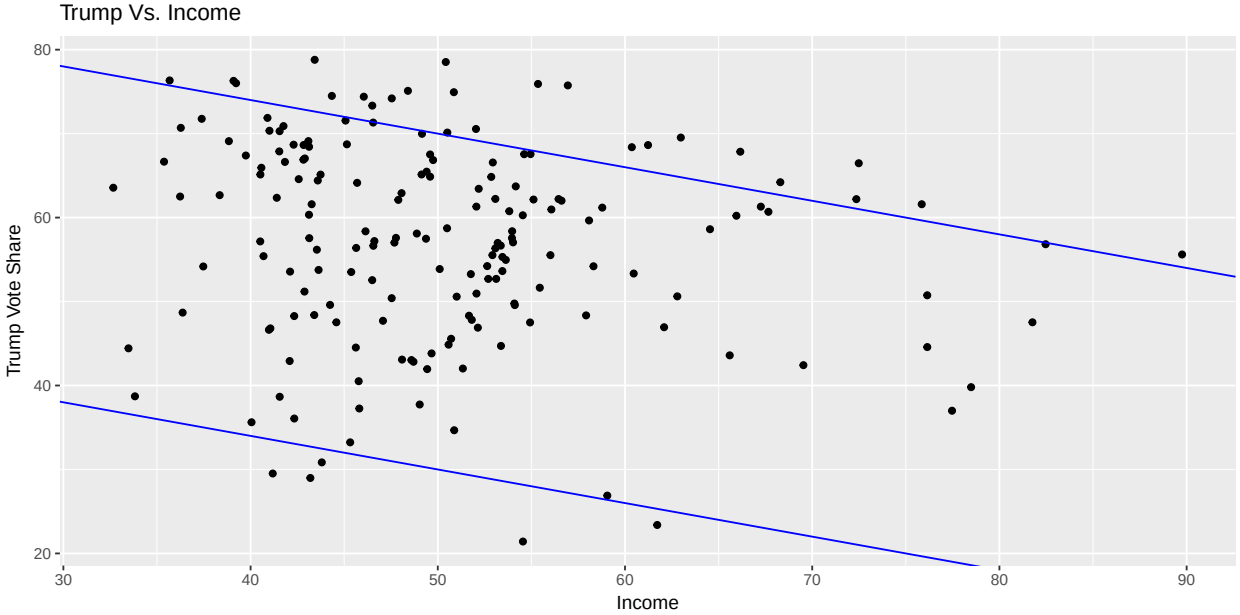
-**3.With High Income and Variation in percentage of votes given to trump(Low, Medium, High)**: The deathrate of black population tends to increase.

3. Fit a model to explain the drug death rate in these counties. Your model must include at least one interaction. You can decide on your model by looking at graphs or by using your favorite numerical criterion (AIC/BIC/etc.); there are many valid choices. Display the model fit graphically in a way that emphasizes the interaction. You must truncate your plots to remove regions that are far away from the observed values of the explanatory variables (e.g. very high incomes.) Explain how the interaction shows up (or doesn't show up) in your plots.

We observe that there is a curse of dimensionality for the income and trump data. Let us investigate this using a scatter plot.



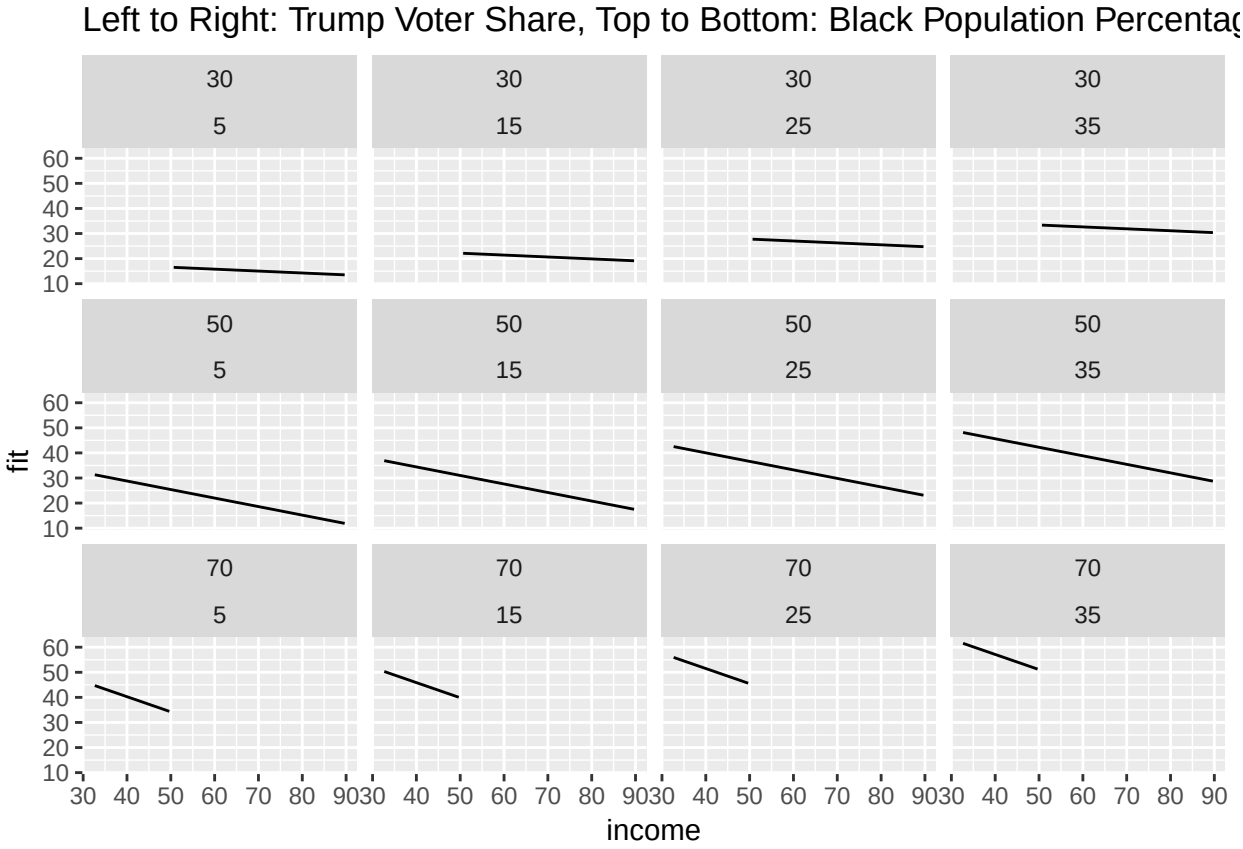
We can see that vast majority of the data falls within a fairly narrow diagonal band of predictor space. For each value of wind, there's a range of about 40% of trump vote share for which the observations are dense. We add two reference lines to the plot:



Model Fitting and Visualization

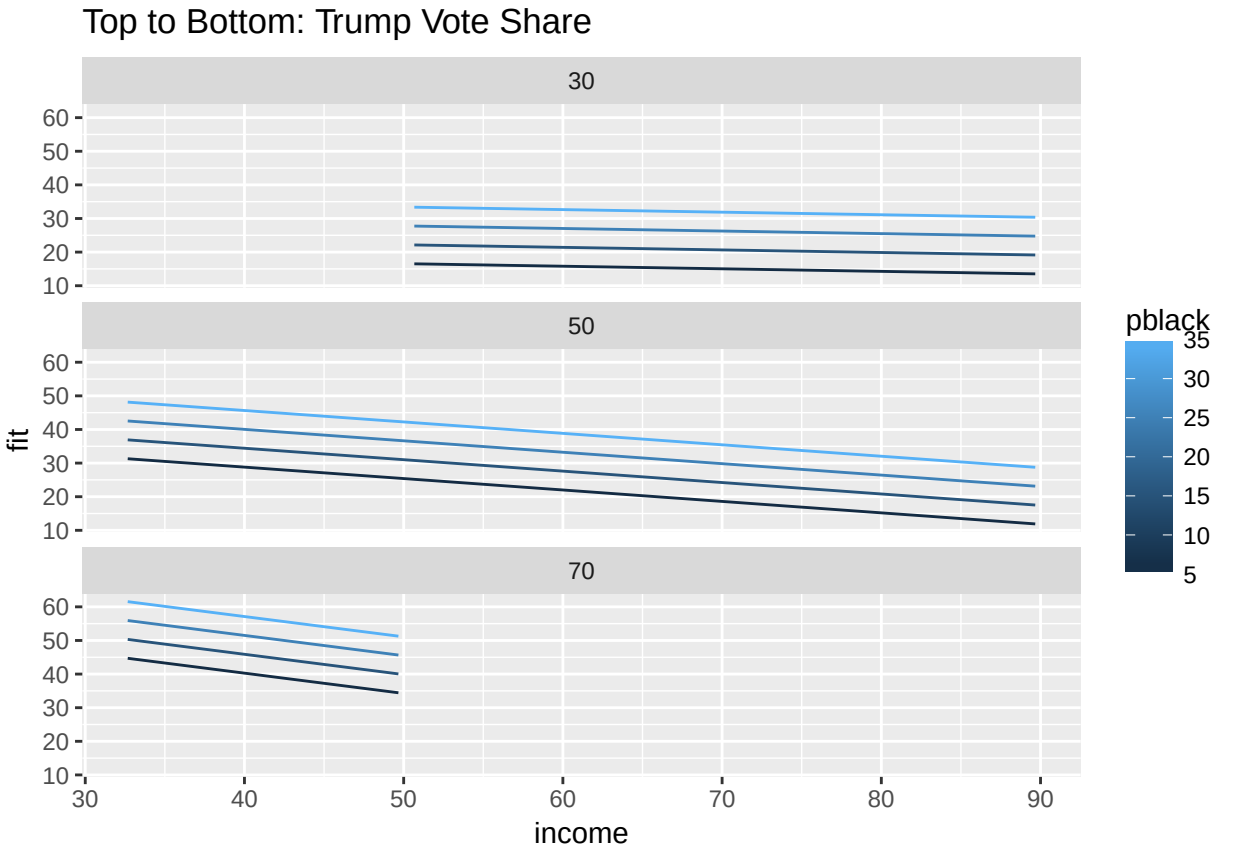
Based on our graphs from Q.1 and Q.2, we decide to fit a linear model for our data. The death rate can be

predicted using our three predictor variables - Income, Trump and Percentage of Black Population and our model also includes an interaction between Income and Trump.



The curve seem to change slope going downwards, justifying our Income:Trump interaction.

Let's draw a quick one way faceted plot to summarize our model and see if there is any interaction between Income and Black Population percentage.



There seems to be some interaction between Income and Black Population Percentage but it doesn't really affect our model and hence our decision of not including this interaction is justifiable.