STAC33 Data Analysis Project

Code

Dwight Sampson

Introduction

In the discipline of sociology and other social sciences, there has always been questions as to whether crowding in a city affects that cities crime rates.

What is crowding?
Crowding occurs when there is a large density of people in a particular area. We know the formula for density is $pop. density = quantity of people /area $.
Our goal is to determine whether crime rate depends on population density or any other factor.

Please note, that all the cities in our dataset are American cities. Our dataset is structured as below.
Variables:
* city: the name of the city
* population: population of the city in thousands
* nonwhite: the percentage of nonwhite citizens in a particular city
* density: the population density for a city measured in people per sq. mile
* crime rate: (number of reported crimes / total population)

It is my hypothesis that as the population of a city increases so too does the crime rate.

Analysis

The first step of our analysis is to read the data into our environment and take a look at the data.

Hide

```
library(tidyverse)
```

```
Registered S3 method overwritten by 'dplyr':
  method           from
  print.rowwise_df
Registered S3 methods overwritten by 'dbplyr':
  method         from
  print.tbl_lazy
  print.tbl_sql
□[30m-- □[1mAttaching packages□[22m --------------------------------------- tidyverse 1.3.0 -
-□[39m
□[30m□[32mv□[30m □[34mggplot2□[30m 3.2.1     □[32mv□[30m □[34mpurrr  □[30m 0.3.3
□[32mv□[30m □[34mtibble □[30m 2.1.3     □[32mv□[30m □[34mdplyr  □[30m 0.8.4
□[32mv□[30m □[34mtidyr  □[30m 1.0.2     □[32mv□[30m □[34mstringr□[30m 1.4.0
□[32mv□[30m □[34mreadr  □[30m 1.3.1     □[32mv□[30m □[34mforcats□[30m 0.5.0□[39m
□[30m-- □[1mConflicts□[22m ------------------------------------------ tidyverse_conflicts() --
□[31mx□[30m □[34mdplyr□[30m::□[32mfilter()□[30m masks □[34mstats□[30m::filter()
□[31mx□[30m □[34mdplyr□[30m::□[32mlag()□[30m    masks □[34mstats□[30m::lag()□[39m
```

Hide

```
url <- "http://www.utsc.utoronto.ca/%7ebutler/c32/crowding.txt"
crime <- read_delim(url, " ")
```

```
Parsed with column specification:
cols(
  city = □[31mcol_character()□[39m,
  population = □[31mcol_character()□[39m,
  nonwhite = □[32mcol_double()□[39m,
  density = □[31mcol_character()□[39m,
  crime = □[32mcol_double()□[39m
)
```

Hide

```
sample_n(crime, 10)
```

city <chr>	population <chr>	nonwhite <dbl>	density <chr>	crime <dbl>
Birmingham	739	32.1	272	2285
Los.Angeles	6860	9.7	1686	4852
Kansas.City	1231	10.9	445	3748
Wichita	406	5.8	166	2532
Duluth	271	0.8	37	1736
Toledo	678	7.3	446	2050
Tampa	924	11.5	709	3247
phoenix	872	5.5	94	3962

city	population	nonwhite	density	crime
<chr>	<chr>	<dbl>	<chr>	<dbl>
Newark	1881	13.4	2683	3261
Cincinnati	1376	10.4	640	1780

1-10 of 10 rows

There are `110` rows in the crime table. Lets remove rows from the table that have missing values. This is due to the fact that all missing values are in the columns population and population density and we want to know if crime rate depends on these columns.

Hide

```
#missing values are marked by a question mark (?)
#remove any missing values
crime_propr <- crime %>% filter(population!='?', density!='?')
# transform population and density columns from character type columns to numeric type columns
crime_propr$population <- as.numeric(strsplit(crime_propr$population, " "))
crime_propr$density <- as.numeric(strsplit(crime_propr$density, " "))
#should we classify people by small city, medium city and large city? Then make a box-plot of aver
age crime rate in each?
```

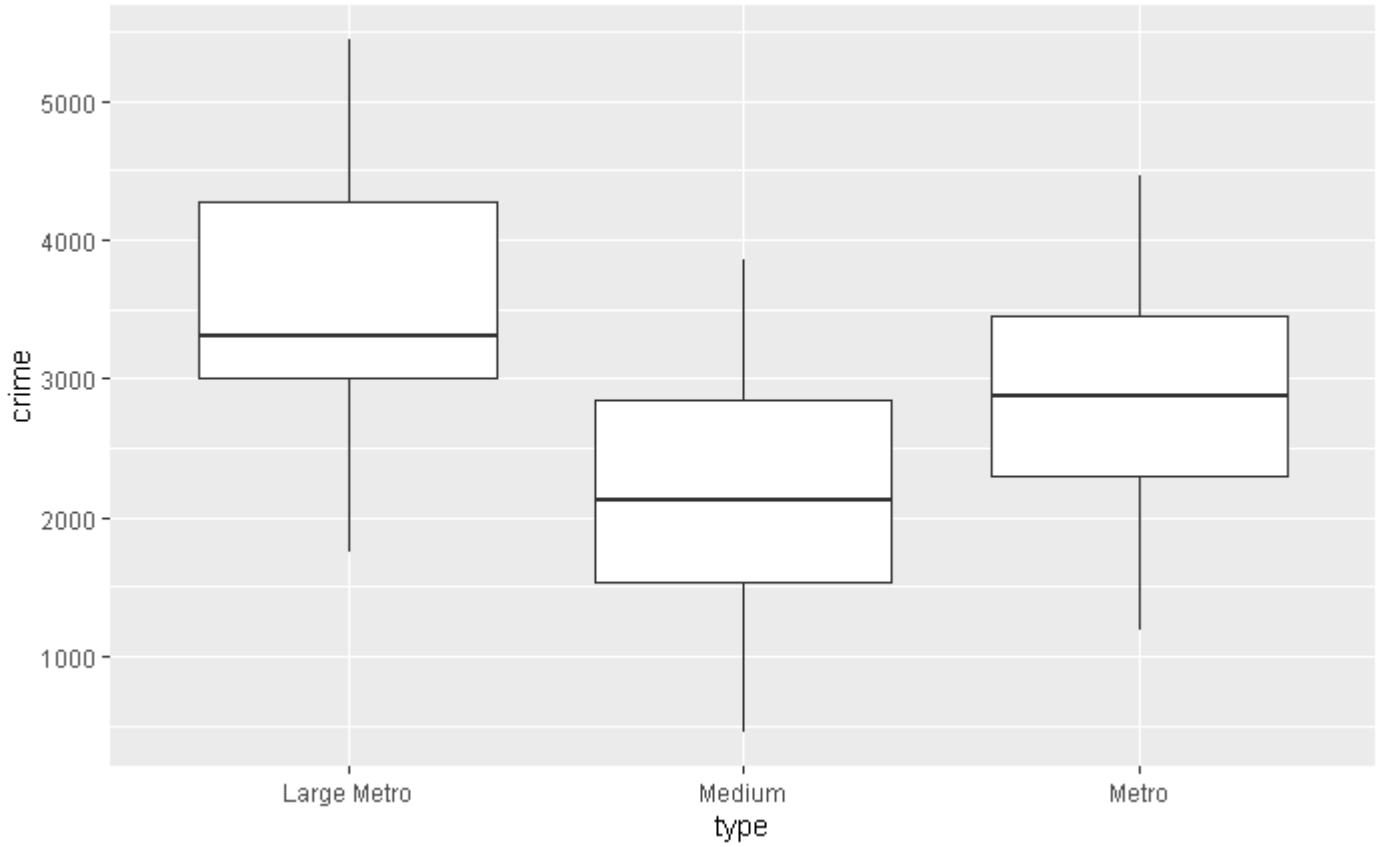
By removing all missing values from the crime table, we are left with `100` oberservations. Meaning there were `10` rows with missing values.

There are different ways to classify cities but we choose to classify them as follows: large metropolitans have a population over 1.5 million, metropolitans have populations between 500,000 and 1.5 million, medium cities have populations between 200,000 and 500,000 and anything else is a small city [1].

Lets investigate if there is a noticable difference in crime rate for each city size. The best way to visualize this is by boxplot.

Hide

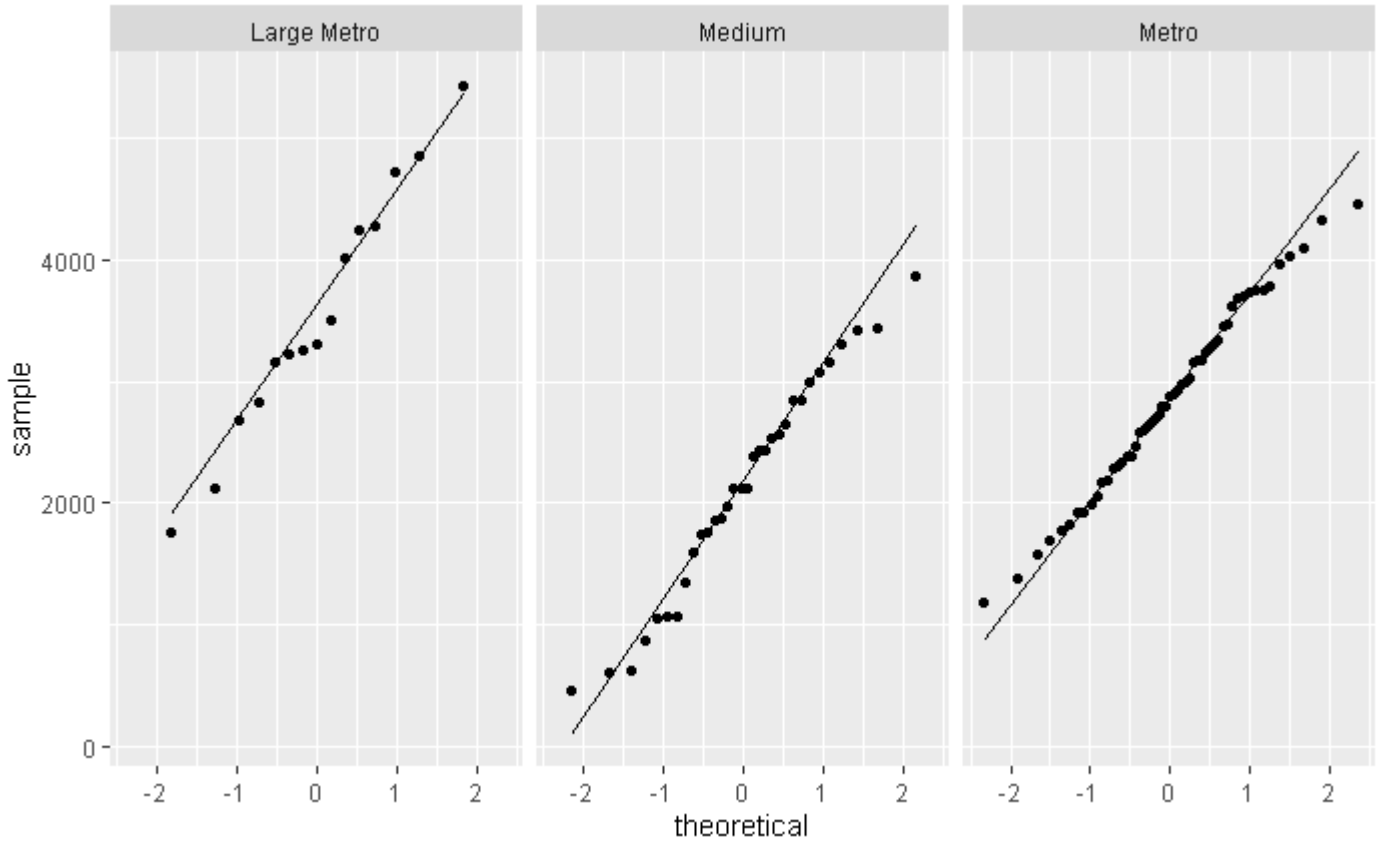
```
# a fcn that classifies every city by population x
class_city <- function(x) {
  ans <- ""
  if (x > 1500){ans <- "Large Metro"}
  else if (x > 500 && x <= 1500){ans <- "Metro"}
  else if (x > 200 && x <= 500){ans <- "Medium"}
  else{ans <- "Small"}
  return(ans)
}
#look at the population of every city and classify it
city_type <- numeric(nrow(crime_propr))
for (i in 1:nrow(crime_propr)){
  row <- crime_propr$population[i]
  city_type[i] <- class_city(row)
}
crime_propr %>% mutate(type = city_type) -> crime_city
ggplot(crime_city, aes(x= type, y=crime)) +geom_boxplot()
```



The box-plot indicates to use that there are no outliers in each of the various types of cities. Notice, each city type also has equal variance. Additionally, there is some overlap between Medium Cities, Metropolitans and Large Metropolitans. Though it is noteworth that crime does seem to increase as the size of the city increases.

Hide

```
ggplot(crime_city, aes(sample=crime)) + stat_qq() + stat_qq_line() + facet_wrap(~type)
```



As we can see above, our data is normally distributed (all points are about the quantile line) and has equal variance (paralel slope). Is there a difference between our groups?

Hide

```
crime_aov <- aov(crime ~ type, data = crime_city)
summary(crime_aov)
```

```
            Df   Sum Sq  Mean Sq F value   Pr(>F)
type         2 22787475 11393738   15.14 1.89e-06 ***
Residuals   97 72987605   752450
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes a difference does exist between the 3 variuos types of cities in our dataset. Where does this difference exist?

Hide

```
TukeyHSD(crime_aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = crime ~ type, data = crime_city)

$type
                          diff        lwr        upr     p adj
Medium-Large Metro -1431.5000 -2077.5770 -785.42296 0.0000024
Metro-Large Metro   -701.4151 -1305.2621  -97.56813 0.0185380
Metro-Medium         730.0849   267.8609 1192.30895 0.0008433
```

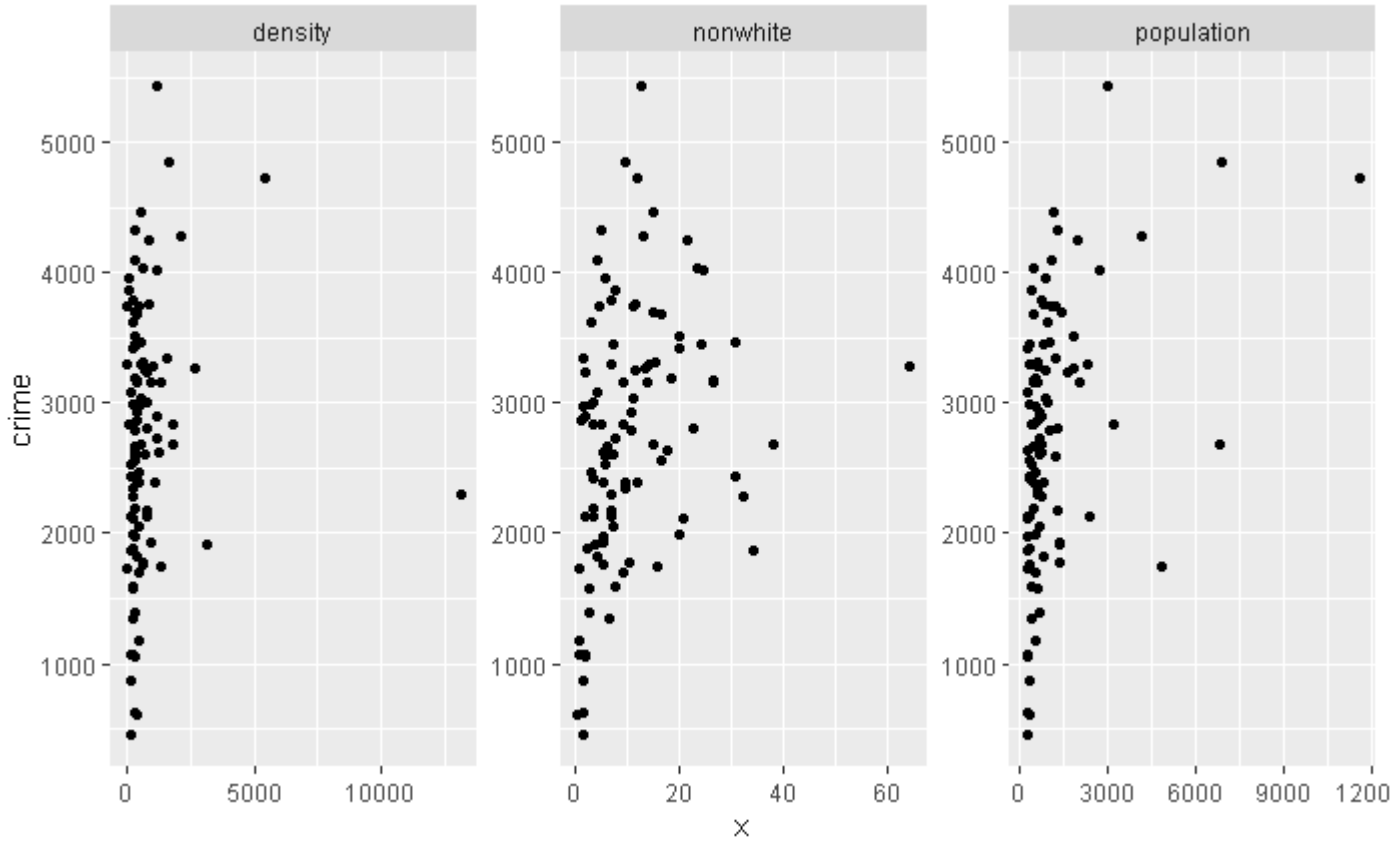
The p-values from the Tukey Test indicate to us that all the sizes of cities in our dataset are different from one another statistically.
Now lets investigate whether crime depends on any of the variables mentioned in the introduction:

Linear Regression

Let's try to observe any obvious relationship in our data.

Hide

```
crime_propr %>% pivot_longer( c(population:density), names_to = "column", values_to = "x") %>%
  ggplot(aes(x=x, y= crime)) + geom_point() + facet_wrap(~ column, scales = "free")
```



Our numeric variables indicate to us that there are strong linear relationships between crime and each variable. Though we seem to have a few outlier values in each plot.

We will choose not to do any analysis on the individual cities. Why? This is due to the fact that our goal is to know if crime rate depends on any variable.

Hide

```
mod_1 <- lm(crime ~ density + nonwhite + population, data= crime_propr)   # the full model
summary(mod_1)
```

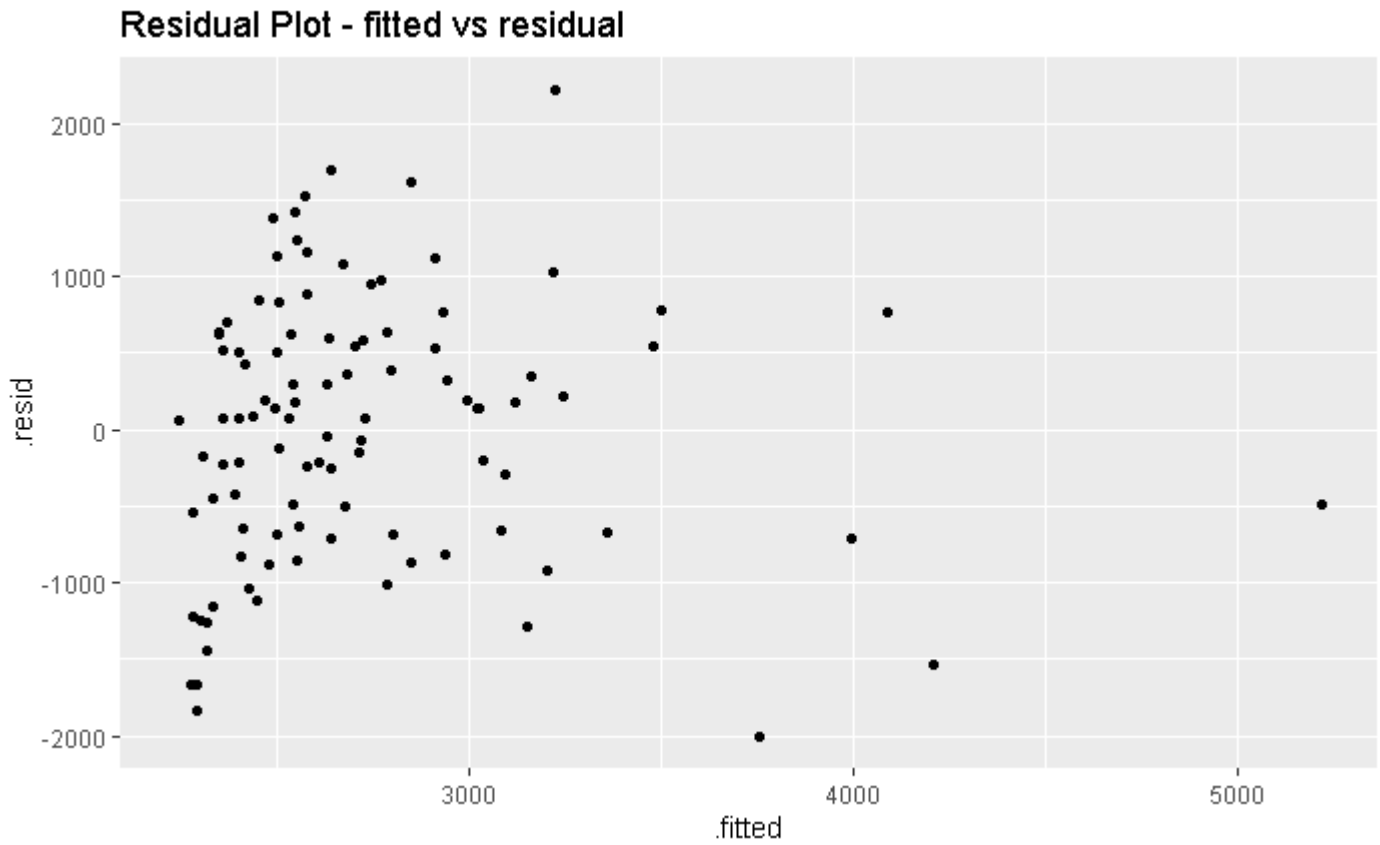
```

Call:
lm(formula = crime ~ density + nonwhite + population, data = crime_propr)

Residuals:
     Min       1Q   Median       3Q      Max 
-2003.18  -657.16    70.94   603.73  2213.20 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2193.70088  143.04566  15.336  < 2e-16 ***
density       -0.02145    0.06578  -0.326 0.745045    
nonwhite      26.03770    8.76746   2.970 0.003764 ** 
population     0.24495    0.06095   4.019 0.000116 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 877.2 on 96 degrees of freedom
Multiple R-squared:  0.2288,    Adjusted R-squared:  0.2047 
F-statistic: 9.493 on 3 and 96 DF,  p-value: 1.496e-05
```

Notice, there is a low R^2 but we have 3 significant variables. Additionally, our p-value for the model is telling us that the model we have fit is much better than the empty model.

Lets check our residuals:

Hide

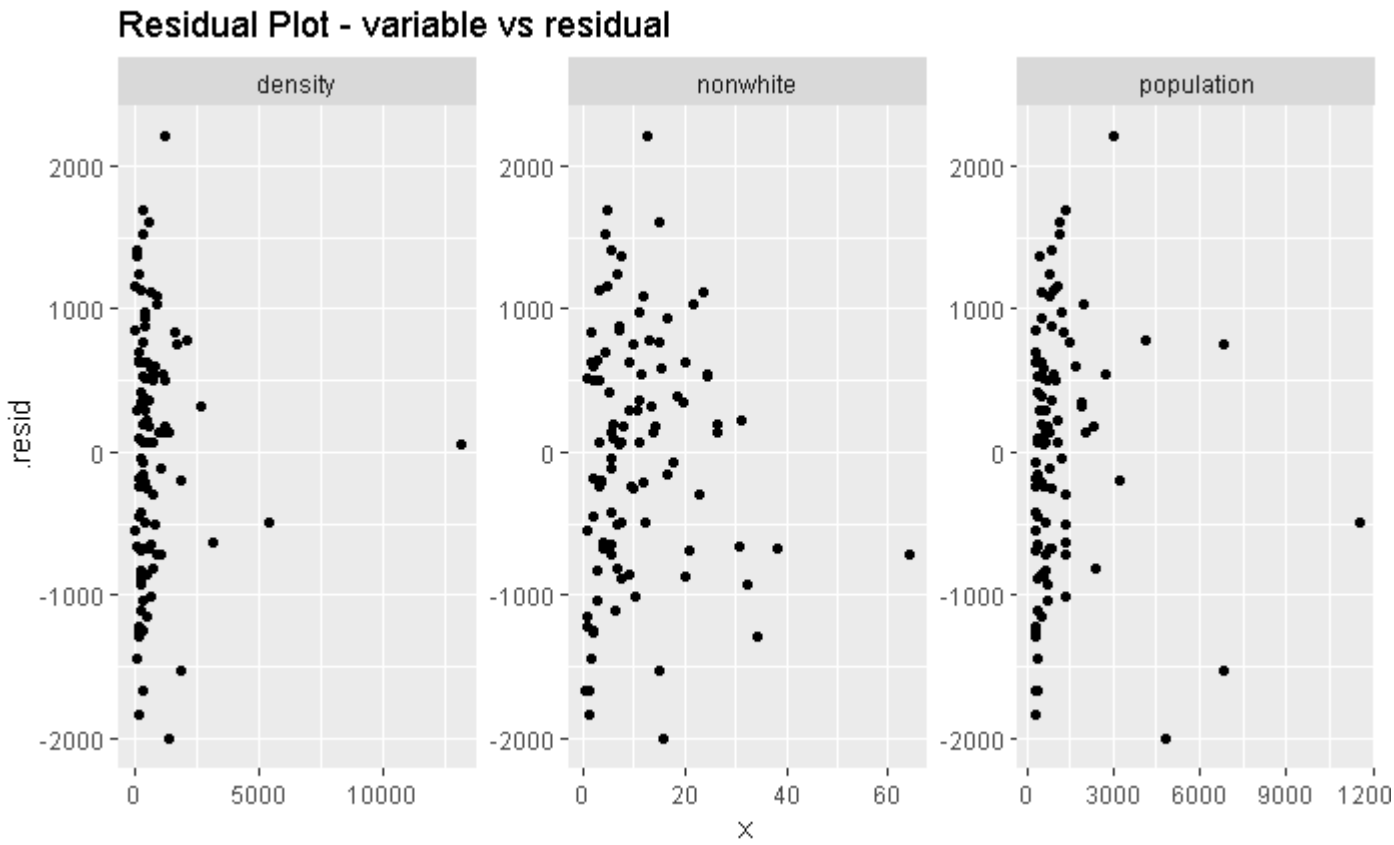
```
ggplot(mod_1, aes(x=.fitted, y= .resid)) + geom_point() + ggtitle("Residual Plot - fitted vs resid
ual")
```



There is no discernable pattern of the residual plot (fitted vs residual). All points are randomly about the line y=0. This is good and indicates that our model assumptions hold.

Hide

```
library(broom) #library need for augment fcn
mod_1 %>% augment(crime_propr) %>%
  pivot_longer( c(population:density), names_to = "column", values_to = "x") %>%
  ggplot(aes(x=x, y= .resid)) + geom_point() + facet_wrap(~ column, scales = "free") +
  ggtitle("Residual Plot - variable vs residual")
```



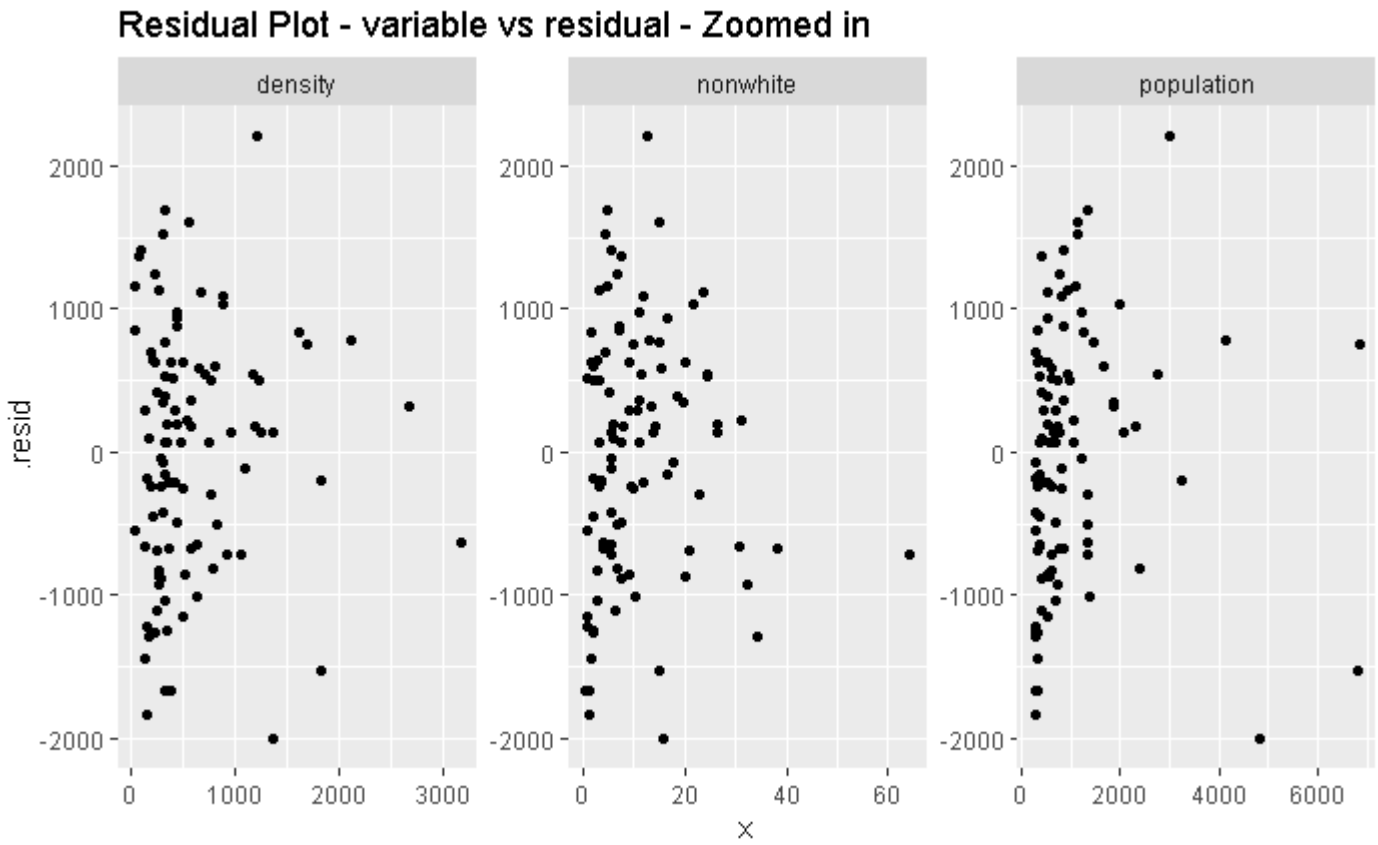
We know that there are more cities with small population density than cities with large population density. This is the nature of the data.

From the graph above, we cannot tell if the points in the residual plot are grouped together because of the nature of the data or because there is a pattern we are not accounting for.

Since we cannot fit all points in the plot if we zoom in, we will look at a smaller subset of the data zoomed in to determine any patterns.

Hide

```
mod_1 %>% augment(crime_propr) %>%
  filter(crime_propr$density < 5000) %>%
  pivot_longer( c(population:density), names_to = "column", values_to = "x") %>%
  ggplot(aes(x=x, y= .resid)) + geom_point() + facet_wrap(~ column, scales = "free") +
  ggtitle("Residual Plot - variable vs residual - Zoomed in")
```



The residual plots are nicely spaced about the y=0 for each variable. There is grouping on the right hand side but this could be due to the nature of the data we expressed above.
Lets double check that we dont need to do any transformations on our data (specifically the response variable):

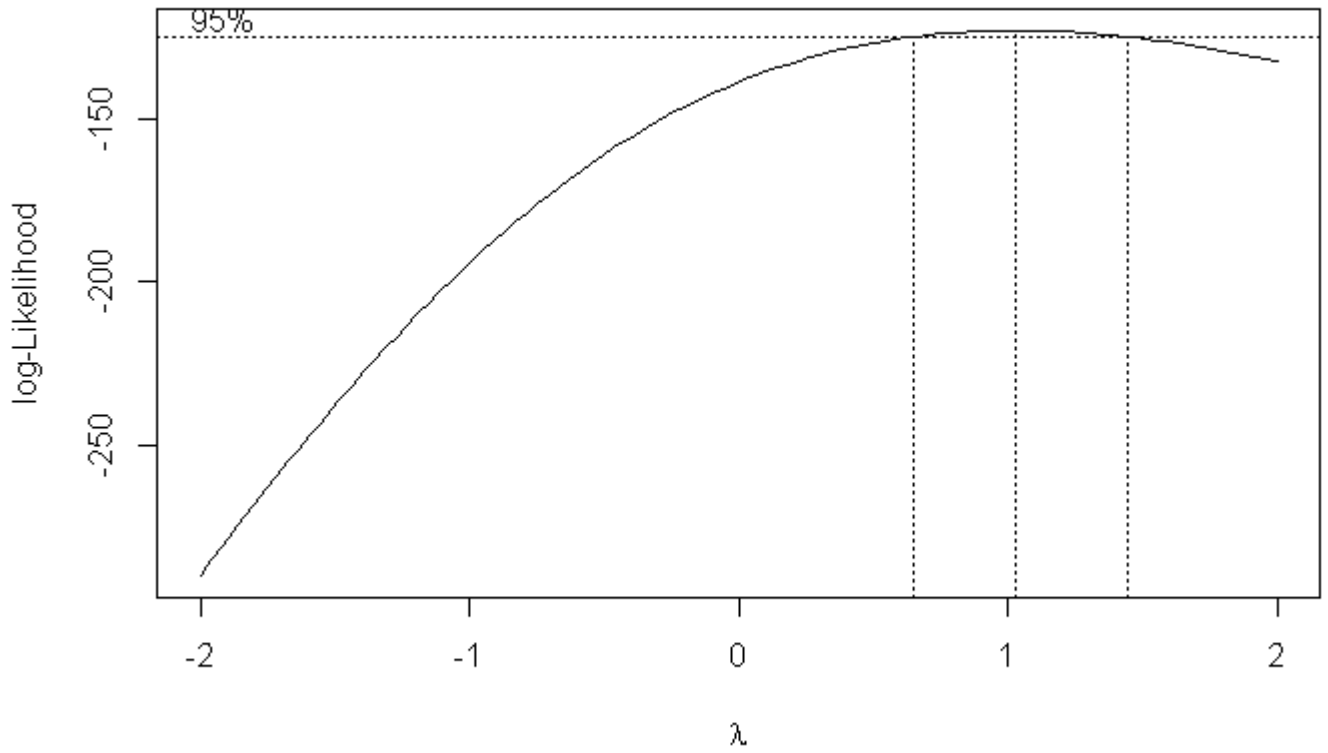
Hide

```
library(MASS) #library need for Box-Cox
```

```
Attaching package: 愀牶MASS愀牷

The following object is masked from 愀牶package:dplyr愀牷:

    select
```

Hide

```
boxcox(crime ~ density + nonwhite + population, data= crime_propr)
```



The Box-Cox test indicates to us that the response is fine as it is and needs no transformation. This is due to the fact that we recieved a lambda value of 1.

Although there isnt any numerical transformation we can do to improve our model is it possible to improve the model by using categorical regression? In particular, by accounting for population using the city type variable instead of the regular population variable can we get a better model fit?

Hide

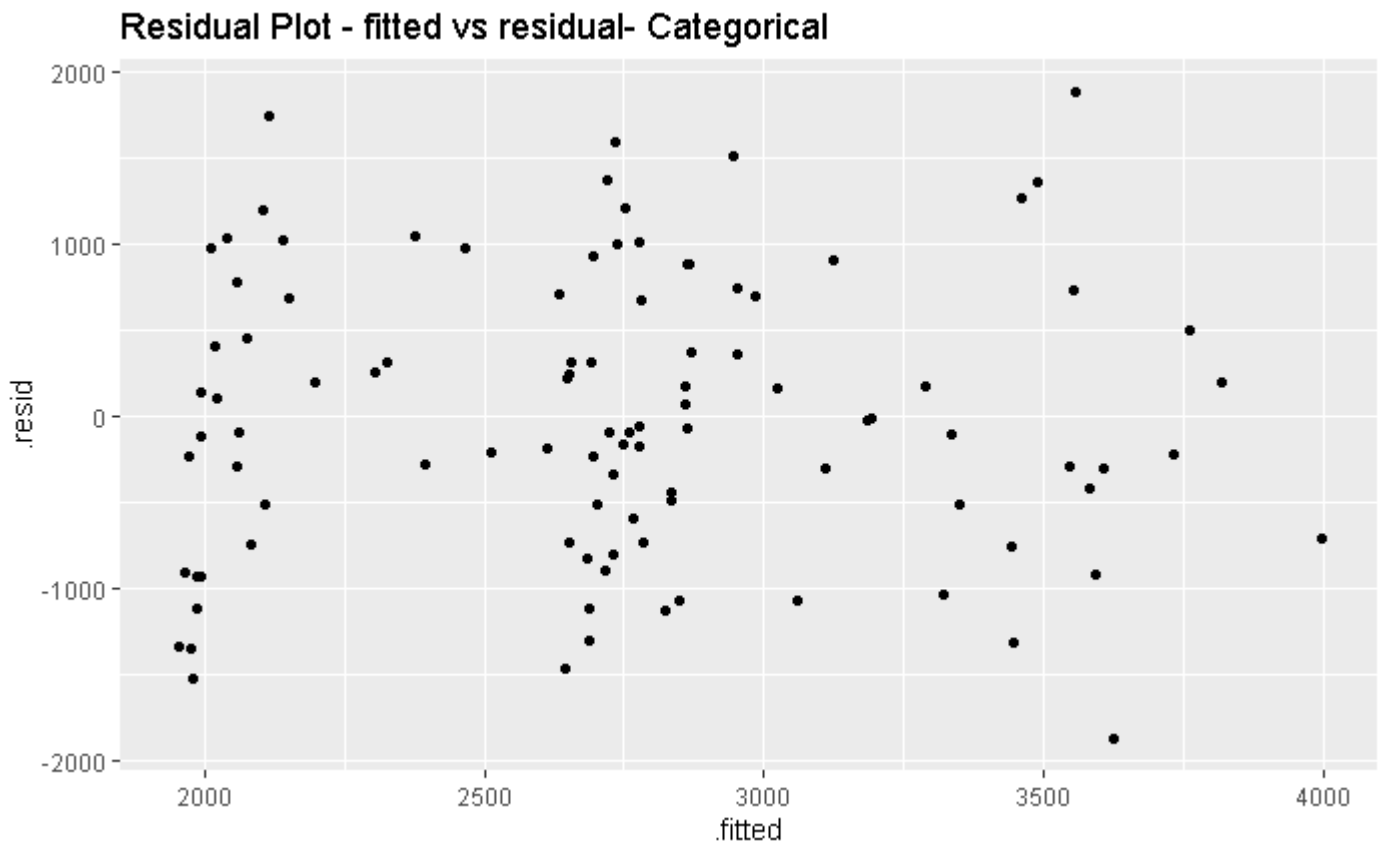
```
city_lm <- lm(crime ~ density + nonwhite + type, data= crime_city)
summary(city_lm)
```

```
Call:
lm(formula = crime ~ density + nonwhite + type, data = crime_city)

Residuals:
     Min       1Q   Median       3Q      Max
-1870.06  -624.00   -76.35   693.78  1883.53

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.314e+03  2.677e+02  12.382  < 2e-16 ***
density     -2.078e-02  6.223e-02  -0.334  0.73916
nonwhite     2.146e+01  8.578e+00   2.502  0.01406 *
typeMedium  -1.360e+03  2.820e+02  -4.822  5.4e-06 ***
typeMetro   -6.774e+02  2.530e+02  -2.677  0.00875 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 848.1 on 95 degrees of freedom
Multiple R-squared:  0.2865,    Adjusted R-squared:  0.2565
F-statistic: 9.538 on 4 and 95 DF,  p-value: 1.585e-06
```

There is a significant improvement in R^2 between this model and our multiple regression model (mod_1). Thus the model does have a better fit.
lets check our residual plots:

Hide

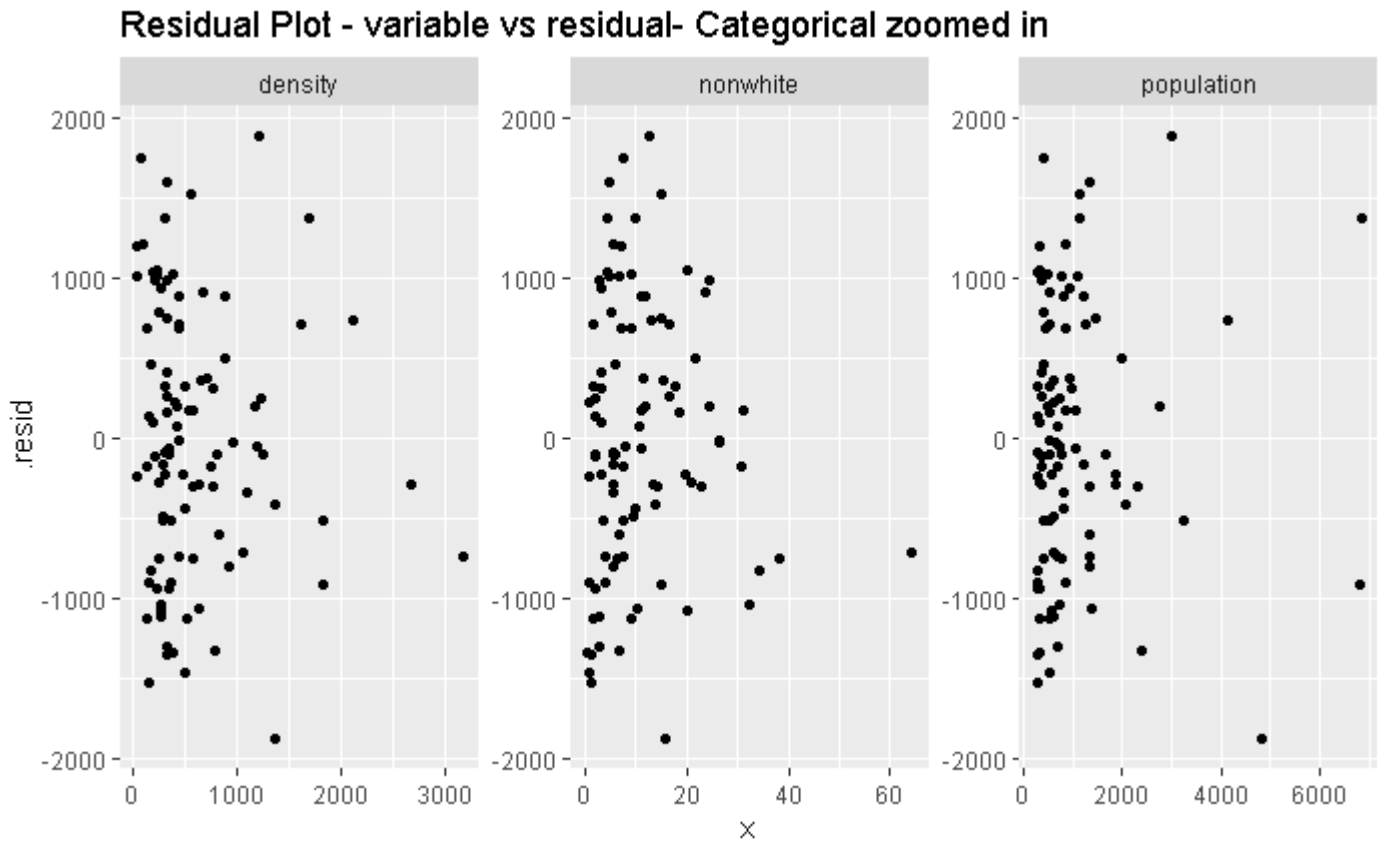
```
ggplot(city_lm, aes(x=.fitted, y= .resid)) + geom_point() + ggtitle("Residual Plot - fitted vs res
idual- Categorical")
```



This fitted vs residual plot looks even better than the one we acheived using population as a predictor of crime. The residual plots are nicely spaced on the x-axis, and randomly placed about the y=0 for each variable.

Hide

```
city_lm %>% augment(crime_city) %>%
  filter(crime_propr$density < 5000) %>%
  pivot_longer( c(population:density), names_to = "column", values_to = "x") %>%
  ggplot(aes(x=x, y= .resid)) + geom_point() + facet_wrap(~ column, scales = "free") +
  ggtitle("Residual Plot - variable vs residual- Categorical zoomed in")
```



The residual plots are nicely spaced, and randomly placed about the y=0 for each variable. Thus, we can move forward with model building.

Model building:

Lets start elliminating variables that arent significant.

Hide

```
#this is the full model
tidy(city_lm) %>% arrange(p.value)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	3.314340e+03	267.68357617	12.3815578	1.563682e-21
typeMedium	-1.359711e+03	282.00540302	-4.8215763	5.403404e-06
typeMetro	-6.773654e+02	253.03365194	-2.6769775	8.752022e-03
nonwhite	2.146240e+01	8.57778361	2.5020917	1.405547e-02
density	-2.078209e-02	0.06223221	-0.3339442	7.391577e-01

5 rows

Density has a very large p-value, meaning it isnt statistically significant to our model. Thus, we will take density out of the model.

Hide

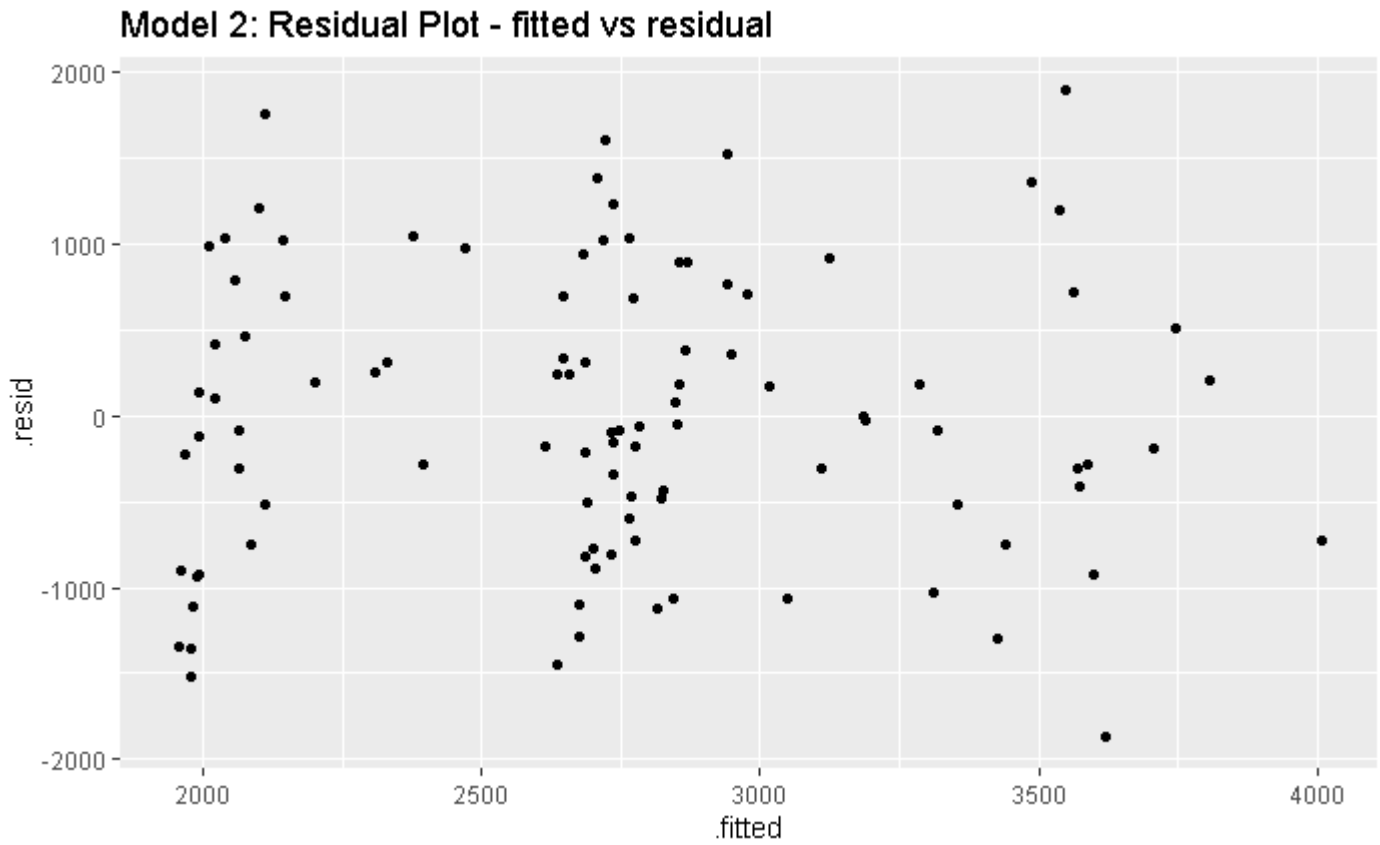
```
#the model that uses categorical regression replacing population with city type
mod_2 <- update(city_lm, .~ .-density)
summary(mod_2)
```

```

Call:
lm(formula = crime ~ nonwhite + type, data = crime_city)

Residuals:
     Min       1Q   Median       3Q      Max 
-1865.45  -626.22   -71.23   692.83  1891.68 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3279.248    245.054  13.382  < 2e-16 ***
nonwhite       21.605      8.527   2.534  0.01291 *  
typeMedium  -1330.787    267.131  -4.982 2.78e-06 ***
typeMetro    -661.520    247.392  -2.674  0.00881 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 844.2 on 96 degrees of freedom
Multiple R-squared:  0.2857,    Adjusted R-squared:  0.2634 
F-statistic:  12.8 on 3 and 96 DF,  p-value: 4.184e-07
```

There are no more non-significant terms to remove, so the model where crime is predicted from city type and percentage of non-white citizens (mod_2) is our final model. It is important that we see significant reductions in the average crime rate as the city type becomes smaller.
Lets look at the residuals for the new model:

Hide

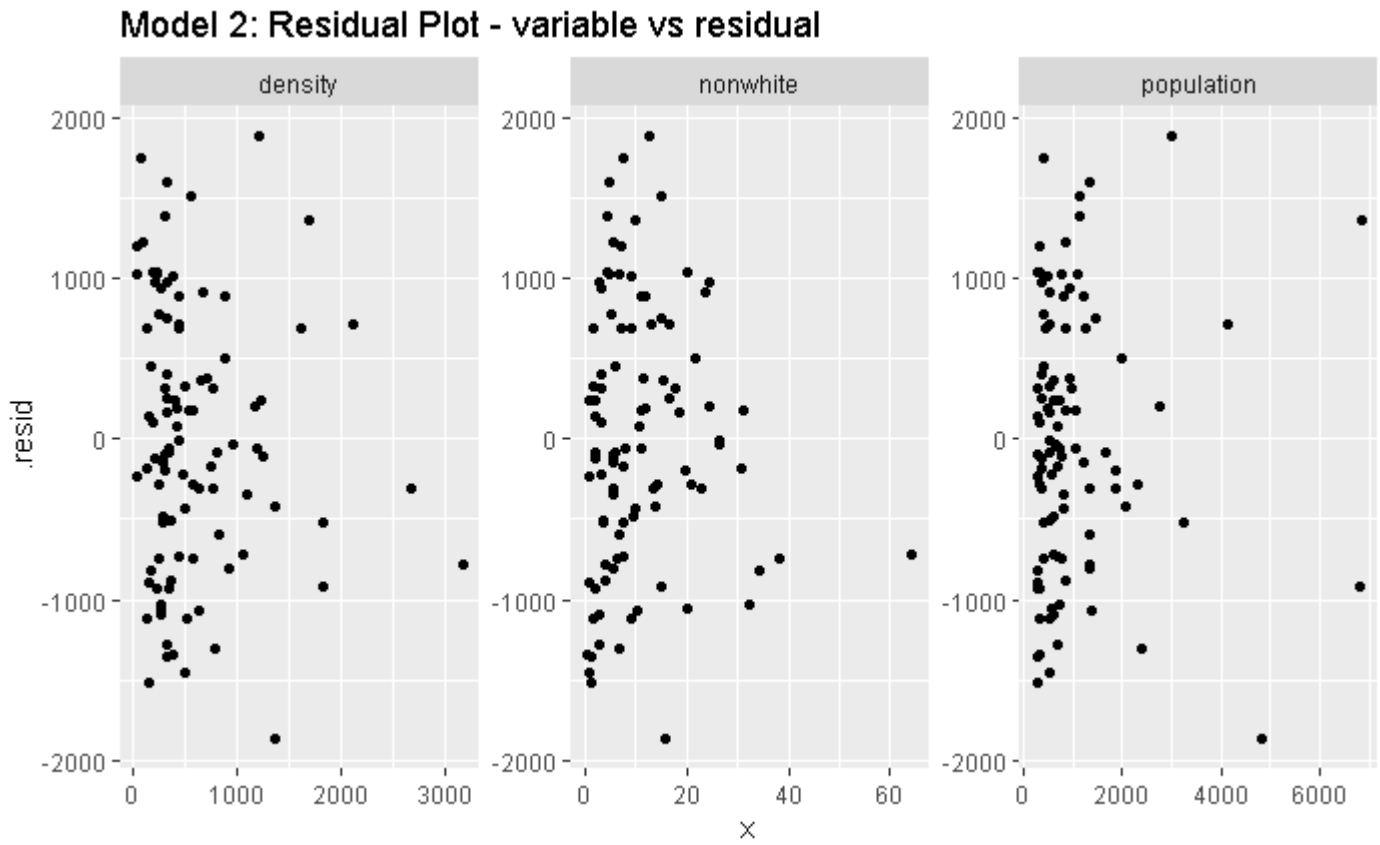
```
ggplot(mod_2, aes(x=.fitted, y= .resid)) + geom_point() + ggtitle("Model 2: Residual Plot - fitted vs residual")
```



There is no discernable pattern of the residual plot (fitted vs residual). All points are randomly about the line y=0. This is good and indicates that our model assumptions hold.

Hide

```
mod_2 %>% augment(crime_city) %>%
  filter(crime_propr$density < 5000) %>%
  pivot_longer( c(population:density), names_to = "column", values_to = "x") %>%
  ggplot(aes(x=x, y= .resid)) + geom_point() + facet_wrap(~ column, scales = "free") +
  ggtitle("Model 2: Residual Plot - variable vs residual")
```



The residual plots are nicely spaced about the y=0 for each variable. There is grouping on the right hand side but this could be due to the nature of the data we expressed above.

Conclusion

Through our analysis we have determined that crime rate does not depend on population density but rather it depends on population size in the form of city type and the percentage of nonwhite citizens in a particular city. Thus, as population increases we also see an increase in crime rate this is evident in both our boxplot and linear regression. Addition there is a significant increase in crime rate of about 22 units for every 1% increase in the percentage of non-white citizens in a particular city.

It should be noted that this model () is only a weak predictor of crime rate. For a stronger model, we need to explore concepts like why crime increases when the percentage of non-white citizens increase in a population. Additionaaly, we need to explore more cities with high population what we classified as large metropolitans.

---

1. https://data.oecd.org/popregion/urban-population-by-city-size.htm (https://data.oecd.org/popregion/urban-population-by-city-size.htm)↩