FALL 2019 , DS - 5110 , HW-4

PART A

Problem 1

Use the following definitions when transforming the dataset: (1) trans women are women who were assigned-male-at-birth; (2) trans men are men who were assigned-female-at-birth; (3) combine the "Genderqueer" and "Androgynous" categories to create a single "Non-binary" category. Filter the dataset to include only participants in these categories.

Create a bar plot showing the number of participants of each of the above genders.

Then create bar plots showing the proportion of participants who have ever been homeless, for each of the above genders. (Do not include missing data in the plot.)

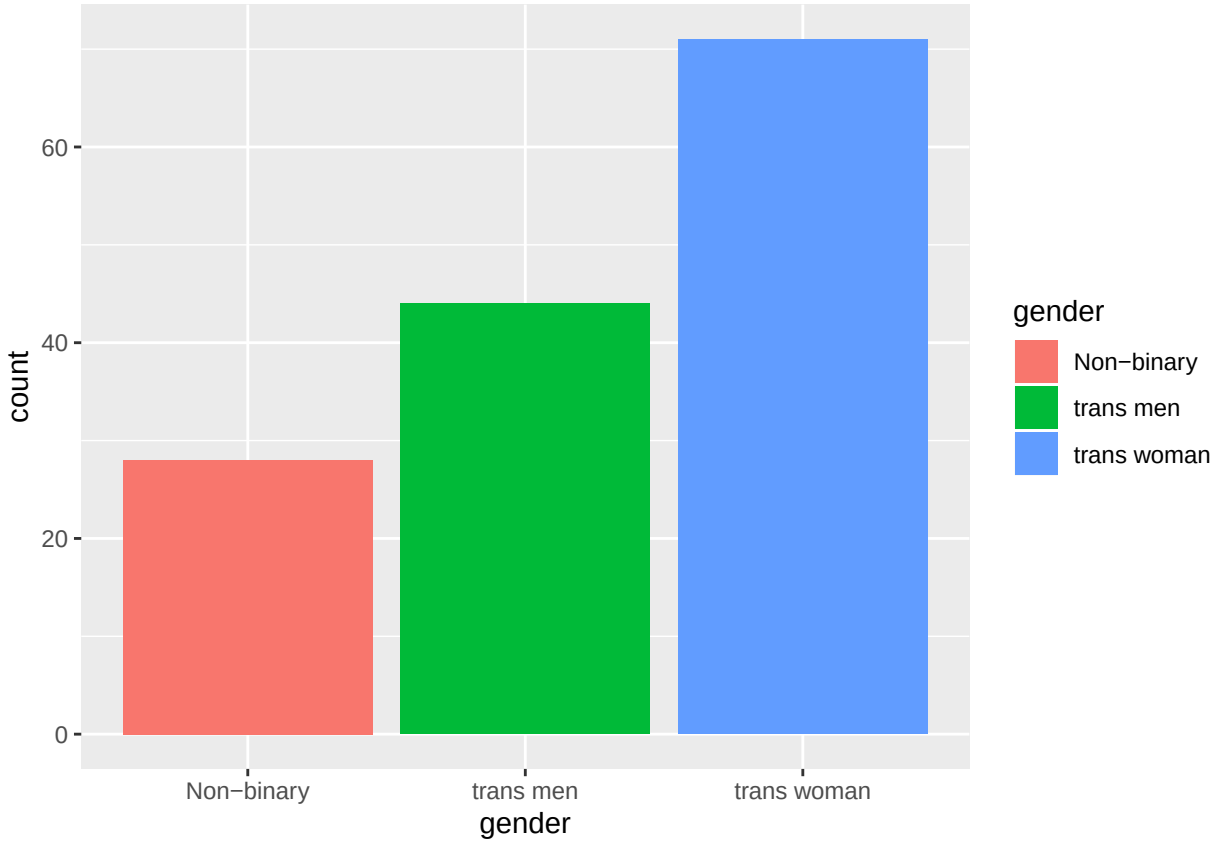
According to statistics from a 2003 study (https://ourworldindata.org/homelessness), roughly 6.2% of the general U.S. population has ever been homeless. How does that compare to the participants of this survey?

Solution :

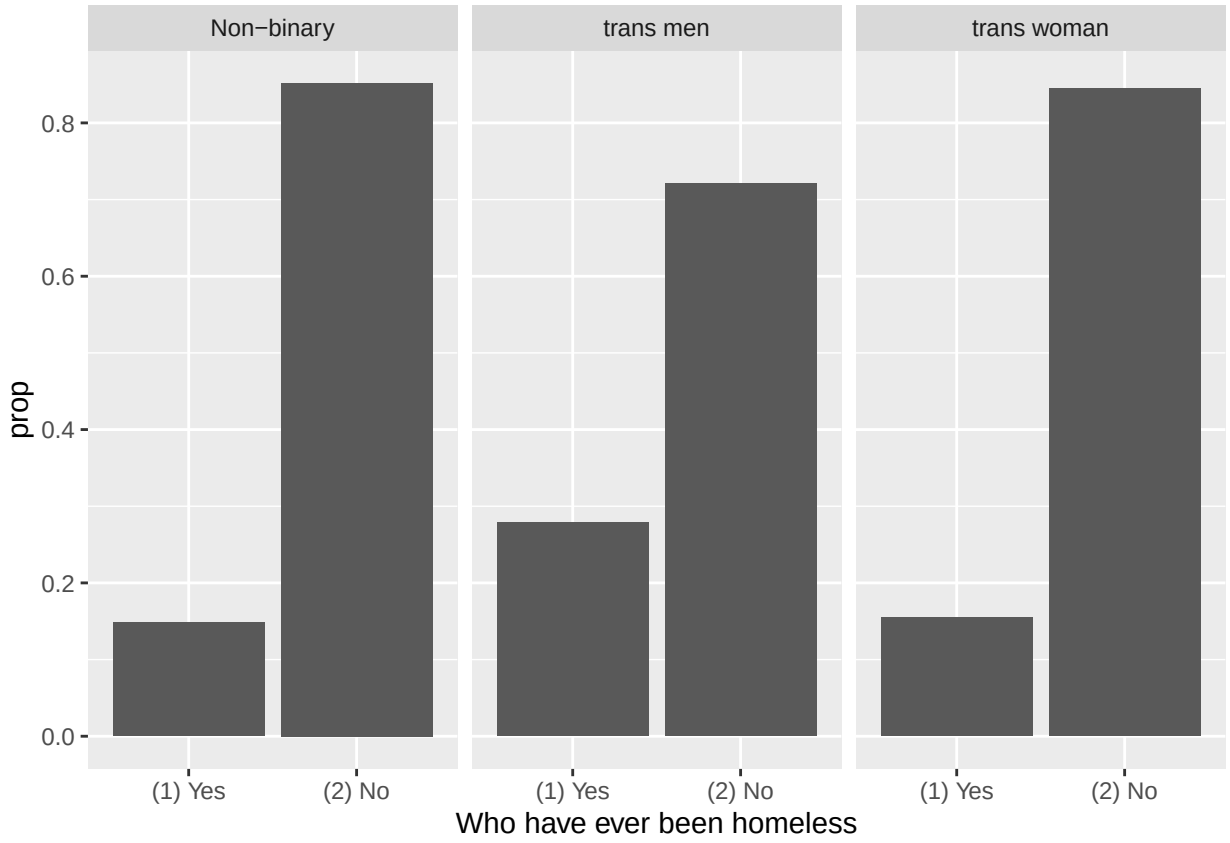
Code

```
load(file = "E:\\FALL_2019\\DS-5110-DM\\HW4-DM\\data\\31721-0001-Data.rda")
df<-as_tibble(da31721.0001)

df %>%
  mutate(gender = case_when(
    Q6 =='(2) Woman' & Q5 == '(1) Male'~'trans woman',
    Q6 =='(1) Man' & Q5 == '(2) Female'~'trans men',
    Q6 =='(4) Androgynous' | Q6 == '(6) Gender Queer'~'Non-binary')) %>%
  filter(gender %in% c('trans woman','trans men','Non-binary')) %>%
  ggplot() +
  geom_bar(aes(x=gender,fill=gender))
```




```
df %>%
  mutate(gender = case_when(
    Q6 =='(2) Woman' & Q5 == '(1) Male'~'trans woman',
    Q6 =='(1) Man' & Q5 == '(2) Female'~'trans men',
    Q6 =='(4) Androgynous' | Q6 == '(6) Gender Queer'~'Non-binary')) %>%
  filter(gender %in% c('trans woman','trans men','Non-binary')) %>%
  filter(Q88!='NA') %>%
  ggplot() +
  geom_bar(aes(x=Q88,y=..prop..,group=gender),position="dodge") +
  labs(x = 'Who have ever been homeless') +
  facet_wrap(~gender)
```



Conclusion

According to statistics we roughly 6.2 % of general U.S population has ever been homeless. From graph we see that proportion for all the genders above are way more that proportion of general U.S population. Around 16% of non binary have ever been homeless which almost 2.5 time of general population. Around 28% of trans men have ever been homeless which is almost more than 4 times of general population. Around 17 % of trans woman have ever been homeless which is almost 2,5 times that of general population. Considering the combine proportion of above 3 categories which is around 20% is 3 times that of general population.

Problem 2

Using the full dataset again, transform the dataset to have a column for race indicating the race of the participant. Include only the racial demographics with publicly available data (i.e., African American, Caucasian, Hispanic/Latinx, and Native American).

(Participants with two or more races may create multiple rows: this is fine for now. Do NOT use the pre-calculated 'RACE' column in the dataset, which does not properly disaggregate multiracial participants.)

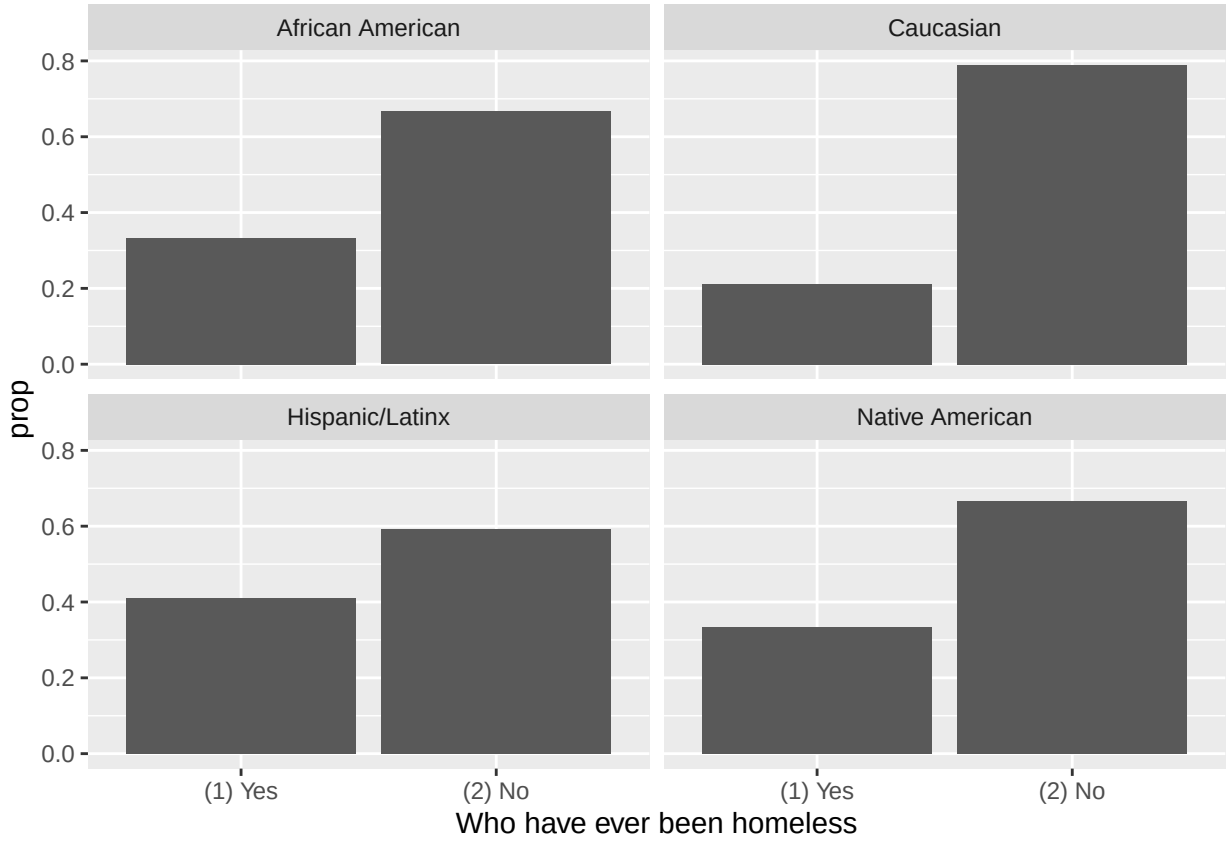
Then create a bar plot showing the proportions of participants who have ever been homeless, for African American, Caucasian, Hispanic/Latinx, and Native American demographics.

(Do not include missing data in the plot.) How do these numbers compared to the statistic of 6.2% of the U.S. general population experiencing homelessness in their lifetime

Solution :

Code

```
df %>%
  gather(D9_1,D9_2,D9_3,D9_4,key = 'race',value = 'answer') %>%
  filter(answer == '(1) Selected' ) %>%
  filter(Q88 != 'NA' ) %>%
  mutate(race = fct_recode(race,
          'African American' = 'D9_1',
          'Caucasian' = 'D9_2',
          'Hispanic/Latinx' = 'D9_3',
          'Native American' = 'D9_4')) %>%
  ggplot()+
  geom_bar(aes(x=Q88,y=..prop..,group='group'),position = 'dodge') +
  labs(x = 'Who have ever been homeless') +
  facet_wrap(~race,nrow=2)
```



Conclusion

From the above graph we see that around 34% of african american have ever been homeless which is almost 6 times that of general population. Around 21% of Caucasian have ever been homeless which is almost 3.5 times that of general population. Around 41% of Hispanic/Latinx have ever been homeless which is almost 7 times that of general population. Around 34% of Native American have ever been homeless which is almost 6 times that of general population. Considering combine proportion of all above category which is around 33% and ehich is 5 times that of general population.

Problem 3

One of the findings reported in the 2015 U.S. Transgender Survey (http://www.ustranssurvey.org) was that a staggering 40% of the respondents reported attempting suicide in their lifetime, nearly nine times the attempted suicide rate of the general U.S. population (4.6%).

Using the full dataset, calculate the total proportion of participants who have attempted suicide in the Virginia THIS survey. Is it higher or lower than the national average for trans people? Is it higher or lower than the national average for the general population?

We would like to know if having a birth family who is supportive of one's gender identity and expression reduces the risk of suicide. Using the full dataset, filter the dataset to remove participants who answered "Not applicable to me" to the question about familial support, and then create bar plots showing the proportions of participants who have thought about killing themselves for each level of familial support. (Do not include missing data in the plot.)

What do you notice?

Solution :

Code

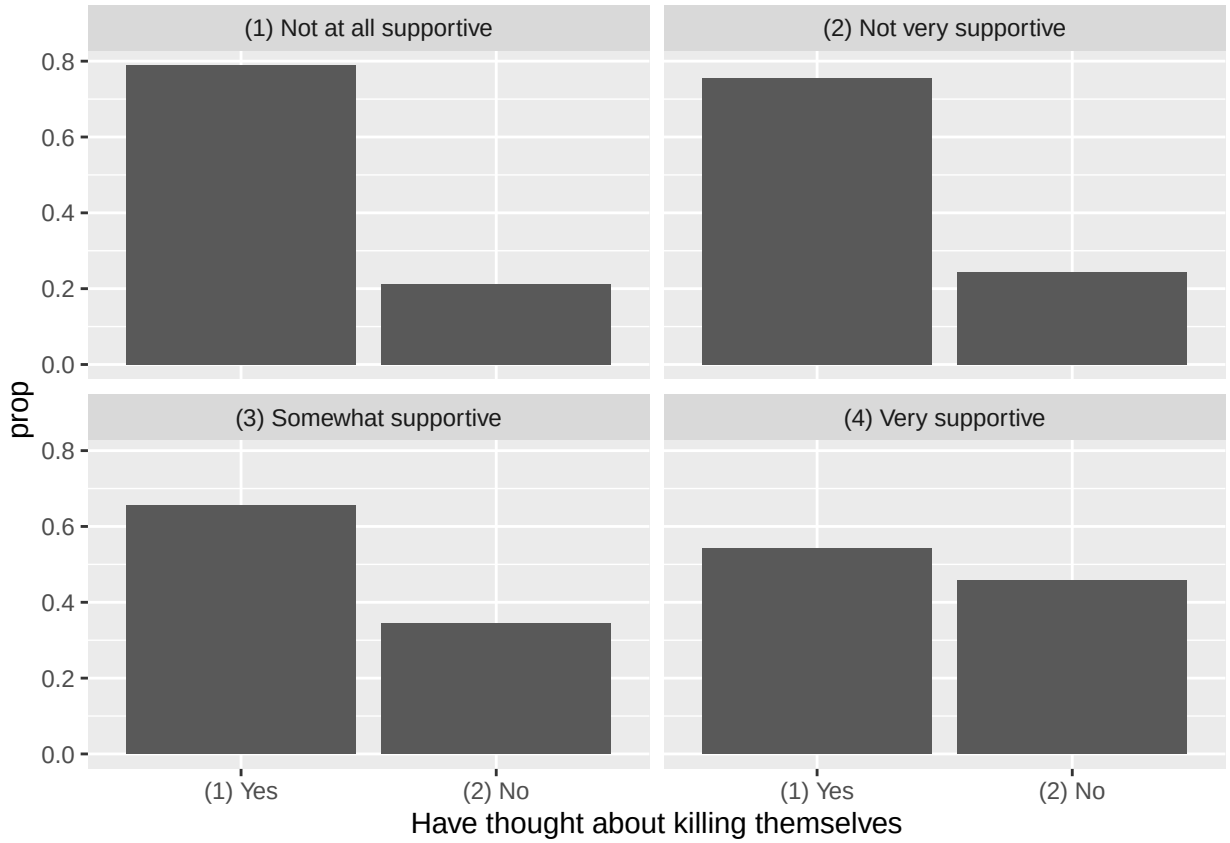
```
df %>%
  summarize(Percantage = sum(Q133=='(1) Yes',na.rm=TRUE)*100/n())
```

```
## # A tibble: 1 x 1
##   Percantage
##        <dbl>
## 1       25.4
```

Conclusion

Total proportion of participants who have attempted suicide in the Virginia THIS survey is 25.4% which less than the national average for trans people which is 40%. But the proportion is higher than the general population which is 4.6% by almost 5.5 times.

```
df %>%
  filter(!(Q119 %in% c('(5) Not applicable to me',NA))) %>%
  filter(!(Q131 == 'NA')) %>%
  ggplot() +
  geom_bar(aes(x=Q131,y=..prop..,group='group',fill=Q131),position="dodge") +
  labs(x = 'Have thought about killing themselves') +
  facet_wrap(~Q119)
```



Conclusion

As we see from above graph that if birth family is not at all supportive the proportion for thought about killing themselves is around 80%. For not very supportive family it drop down slightly to 76%. For somewhat supportive it drop down further to 65%. For very supportive family it is least which is around 55%. So we see from graph that if family is supportive than we have less proportion of participants who thought about killing themselves.

Part B

Problem 4

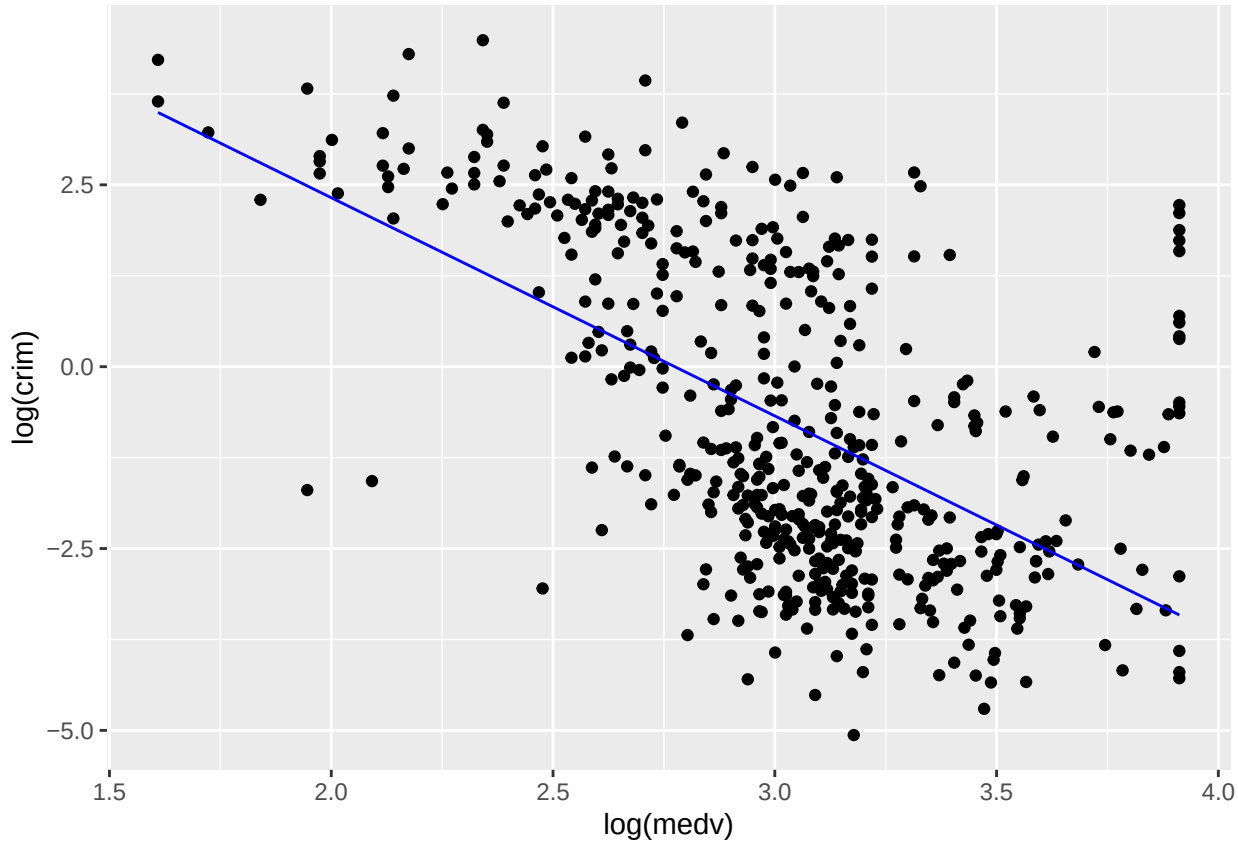
Fit a model that predicts per capita crime rate by town (crim) using only one predictor variable. Use plots to justify your choice of predictor variable and the appropriateness of any transformations you use. Print the values of the fitted model parameters.

Solution :

We will try different combination for predictor and response variable by adding transformation. After trying all combination of predcitor variable with response , none of them show clear linear relationship . Even applying transformation on one axis doesnt help. So we have plotted some plots for both axis transformation(Showing only few of them).

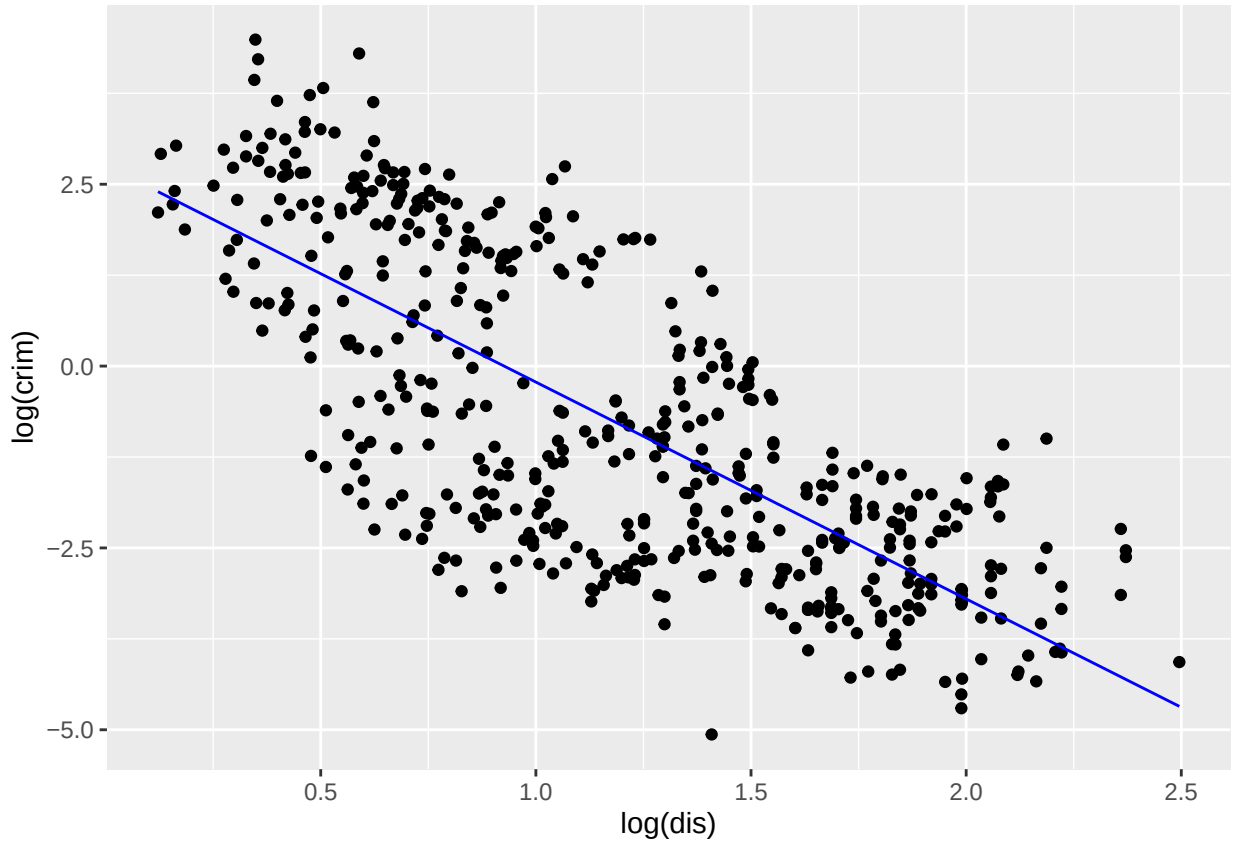
```
data(BostonHousing)

# 1st plot
fit_medv <- lm(log(crim) ~ log(medv), data = BostonHousing)
BostonHousing %>%
  add_predictions(fit_medv) %>%
  ggplot(aes(x=log(medv))) +
  geom_point(aes(y=log(crim))) +
  geom_line(aes(y=pred),color = 'blue')
```

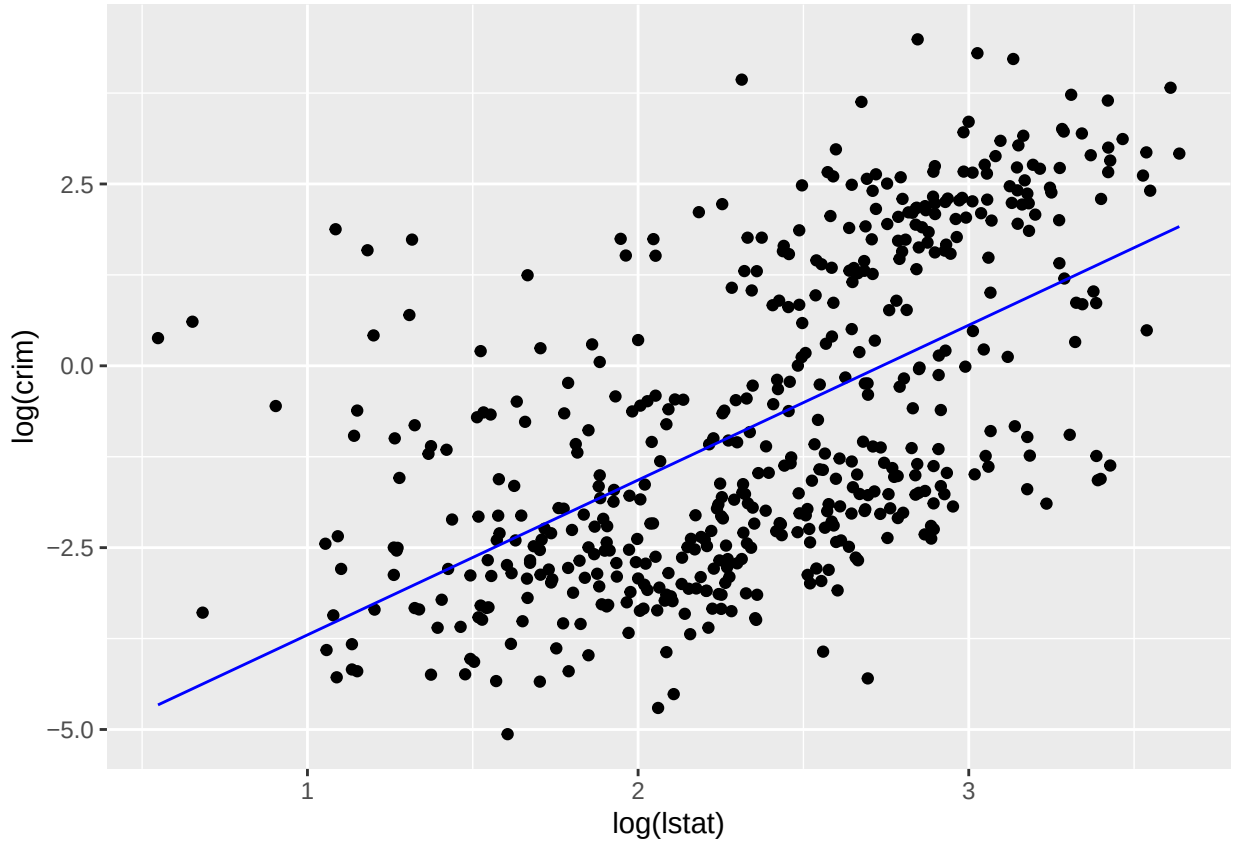



```
# 2nd plot
fit_dis <- lm(log(crim) ~ log(dis), data = BostonHousing)
BostonHousing %>%
  add_predictions(fit_dis) %>%
```

```
  ggplot(aes(x=log(dis))) +
  geom_point(aes(y=log(crim))) +
  geom_line(aes(y=pred), color = "blue")
```




```
# 3rd plot
fit_lstat <- lm(log(crim) ~ log(lstat), data = BostonHousing)
BostonHousing %>%
  add_predictions(fit_lstat) %>%
  ggplot(aes(x=log(lstat))) +
  geom_point(aes(y=log(crim))) +
  geom_line(aes(y=pred),color = 'blue')
```



Conclusion

From above graphs we see that we get some linear trend when we use log(crim) vs log(dis). For other graphs and combinations tried some of them were somewhat linear but not as log(dis) and hence we will choose 'dis' as our first predictor variable using tranformation log(dis) and log(crim).

The model parameters are

```
fit_dis <- lm(log(crim) ~ log(dis), data = BostonHousing)
print(fit_dis)
```

```
##
## Call:
## lm(formula = log(crim) ~ log(dis), data = BostonHousing)
##
## Coefficients:
## (Intercept)     log(dis)
##       2.761       -2.981
```

Problem 5

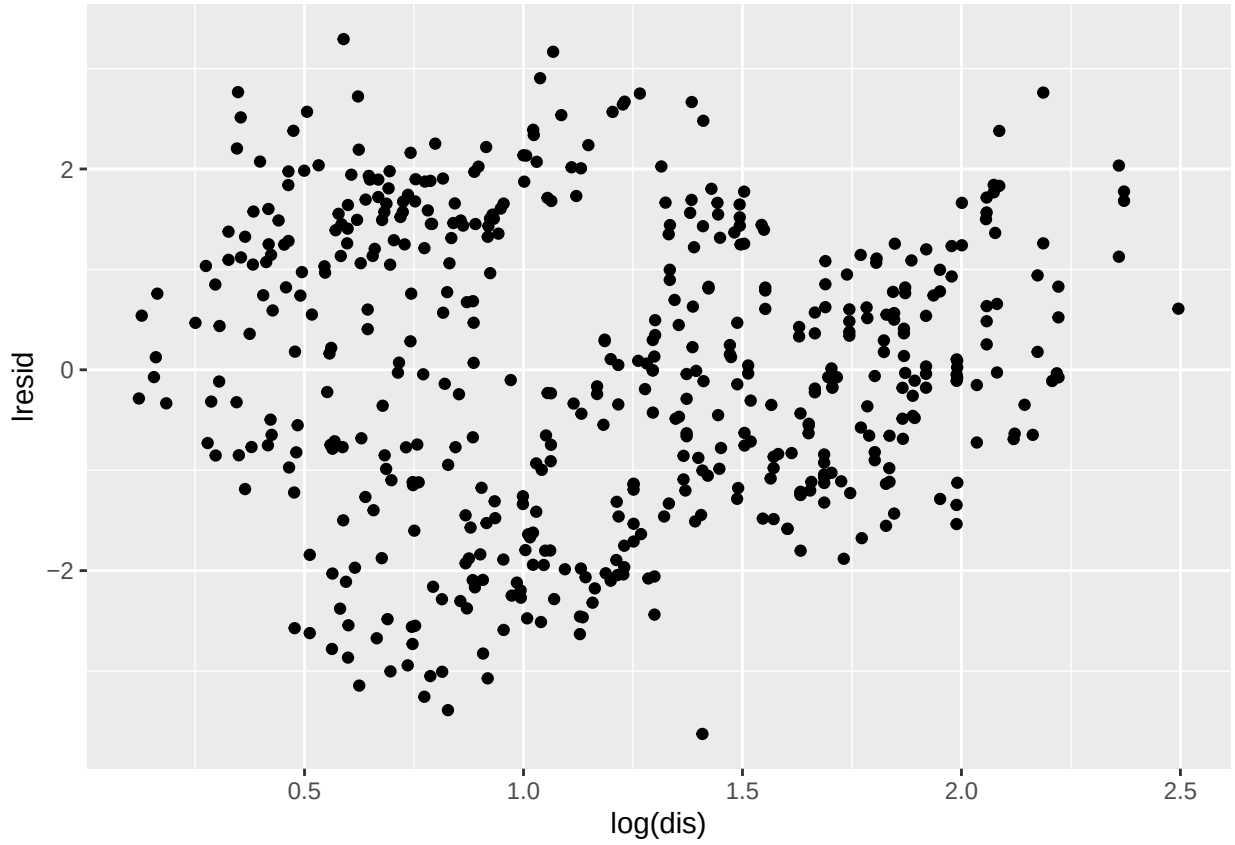
Plot the residuals of the fitted model from Problem 4 against the predictor variable you used to fit the model. Comment on what you observe in the residual plot. Is your model appropriate?

Now choose another variable that you would add to your current model to improve it. Use a residual plot to justify your choice.

Solution :

Residual plot for log(dis)

```
BostonHousing %>%
  add_residuals(fit_dis,'lresid') %>%
  ggplot(aes(x=log(dis))) +
  geom_point(aes(y=lresid))
```



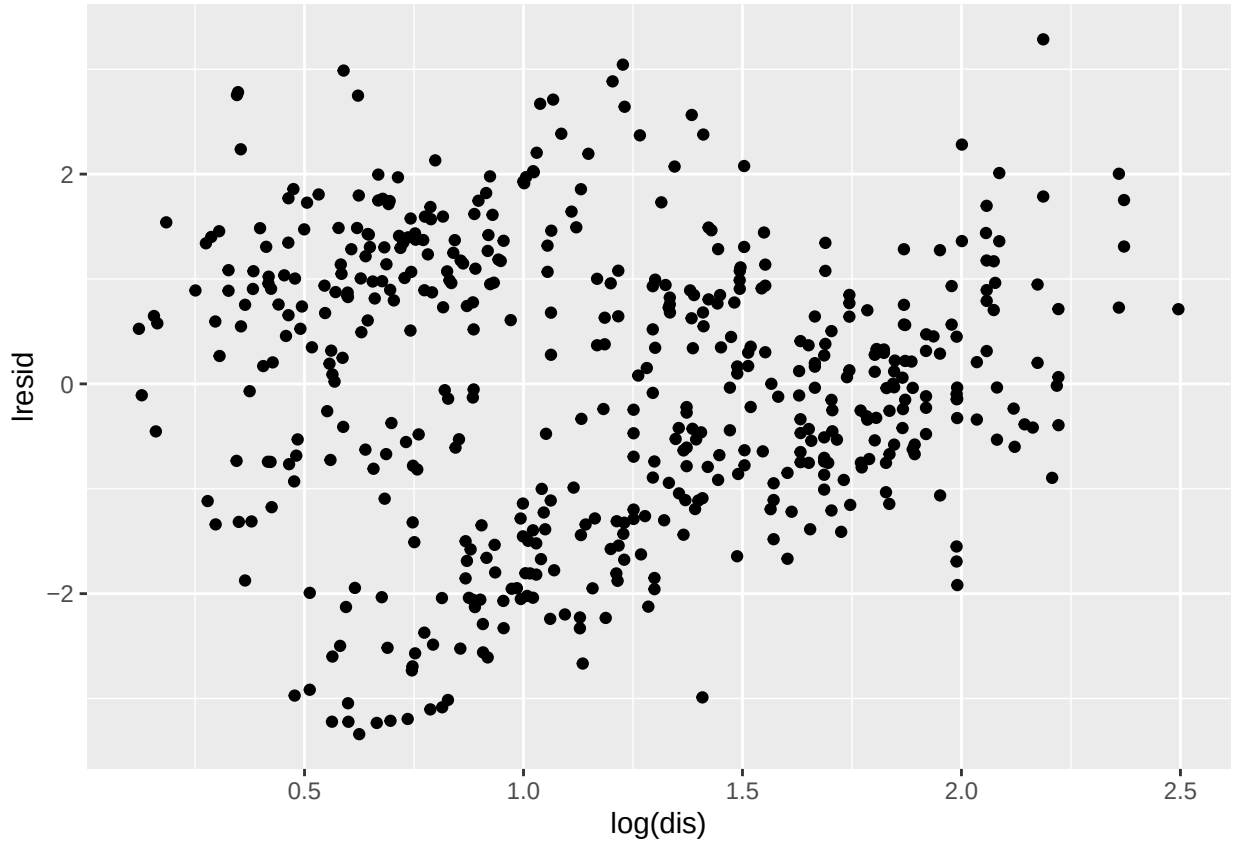
Conclusion

For log(dis), the residual plot shows randomness without any systematic pattern, suggesting no violation of model assumptions. Also we can say that model has captured almost all the relationship between log(crim) and log(dis). And hence our model is appropriate.

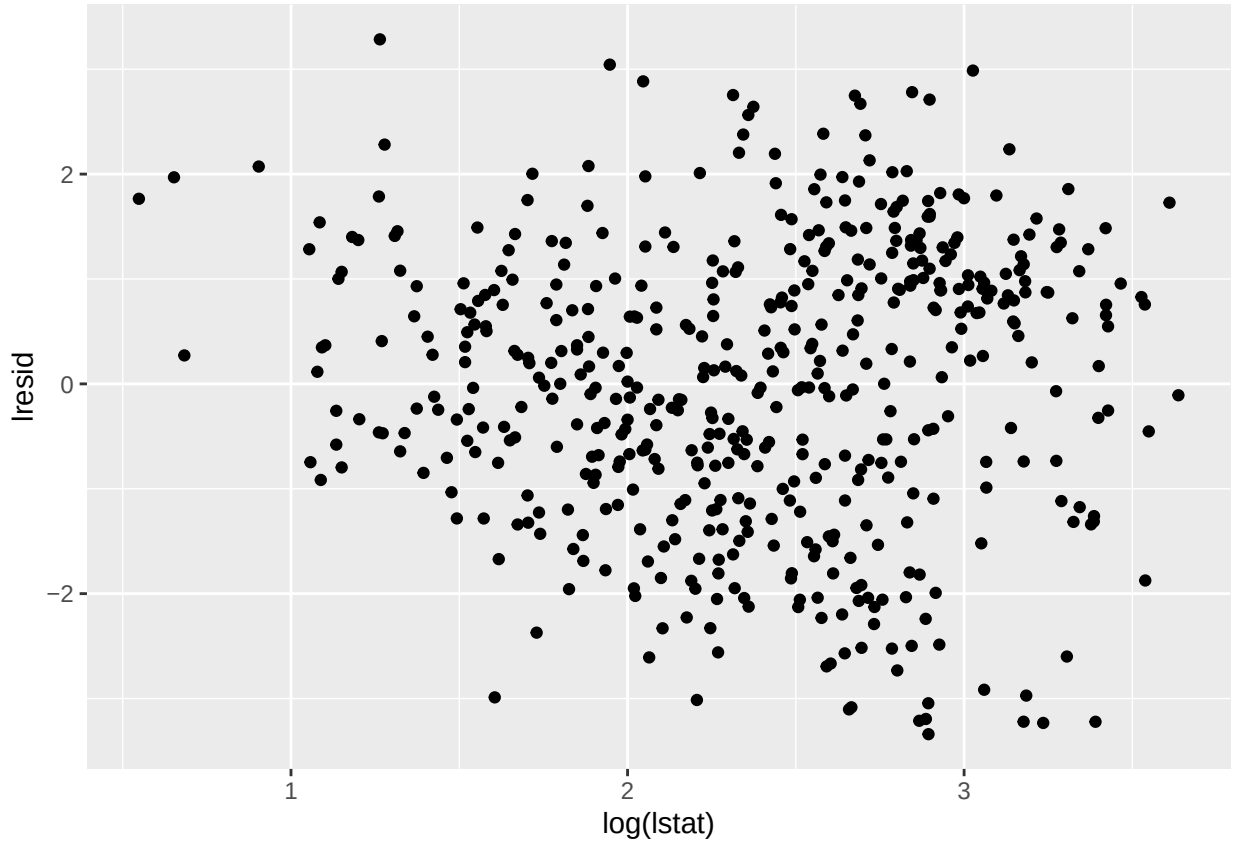
After trying different combination for second predcitor we can select 'lstat' as our second predictor variable and can justify that based on residual plot below:

```
fit_dis_lstat <- lm(log(crim) ~ log(dis)+log(lstat), data = BostonHousing)

BostonHousing %>%
  add_residuals(fit_dis_lstat,'lresid') %>%
  ggplot(aes(x=log(dis))) +
  geom_point(aes(y=lresid))
```




```
BostonHousing %>%
  add_residuals(fit_dis_lstat,'lresid') %>%
  ggplot(aes(x=log(lstat))) +
  geom_point(aes(y=lresid))
```



Conclusion

After plotting the residul plot for both predictor variable , plot shows randomness without any systematic pattern, suggesting no violation of model assumptions. Also we can say that model has captured almost all the relationship between log(crim) and log(dis)+log(lstat). And hence our model is appropriate.