# Tract Data Exploration

*Liying Lu*

```
# impoart package
suppressWarnings(suppressPackageStartupMessages(library(dplyr)))
suppressWarnings(suppressPackageStartupMessages(library(tidyverse)))
suppressWarnings(suppressPackageStartupMessages(library(factoextra)))
suppressWarnings(suppressPackageStartupMessages(library(ggplot2)))
suppressWarnings(suppressPackageStartupMessages(library(corrplot)))
suppressWarnings(suppressPackageStartupMessages(library(ggpubr)))
suppressWarnings(suppressPackageStartupMessages(library(FactoMineR)))
suppressWarnings(suppressPackageStartupMessages(library(missMDA)))
suppressWarnings(suppressPackageStartupMessages(library(ggfortify)))
```

**Question: What advice would you give local governments who wish to improve the response to the 2020 census?**

## Data transformation

```
# import data
tractData <-
  read.csv("~/PersonalProject/DataFestPrep/data/2019PlanningDatabaseTractData/pdb2019trv6_us.csv")
```

Get a smaller dataset by subsetting only the data in DMV region. D.C. Maryland and Virginia.

```
tractDMV <- tractData %>%
  filter(State_name=="Virginia" |
           State_name=="Maryland" |
           State_name=="District of Columbia")
```

Summary statistics of the low_response_score on each state.

```
print(summary(tractDMV$Low_Response_Score))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.60   16.30   20.00   20.61   24.60   43.80      59
```

```
maryland <- filter(tractDMV, State_name=="Maryland")
print(summary(maryland$Low_Response_Score))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.60   16.00   20.20   20.69   25.20   43.80      21
```

```
virginia <- filter(tractDMV, State_name=="Virginia")
print(summary(virginia$Low_Response_Score))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.60   16.20   19.40   19.99   23.40   43.30      38
```

```
dc <- filter(tractDMV, State_name=="District of Columbia")
print(summary(dc$Low_Response_Score))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.50   22.20   26.30   26.43   30.70   43.20
```

**Cultural: Race, Language spoken, Citizenship**
- **Total Population**: Columns 14, 15
- **Low Response Rate**: Columns 284
- **Race**: Columns 46-66
- **Language spoken**: Columns 69-96 and 179-188
- **Citizenship**: Columns 145-152

These subsets include both the American Community Survey (ACS) and ACS margin of error (ACSMOE) data, but only the ACS data will be taken into consideration.

```
# get the cultural dataset for DMV region
dmv_cultural <- tractDMV %>% select(c(3,5,284,(14:15),(46:66),(69:96), (179:188), (145:152)))
dmv_race <- tractDMV %>% select(c(3,5,284,(46:66)))
dmv_lang <- tractDMV %>% select(c(3,5,284,(69:96),(179:188)))
dmv_citizen <- tractDMV %>% select(c(3,5,284,(145:152)))
dmv_pop <- tractDMV %>% select(c(3,5,(14:15)))
```

# Race

```
# subset the columns for different races
dmv_race <- dmv_race %>%
  select(-contains("MOE"))
#names(dmv_race)
pop <- dmv_pop$Tot_Population_ACS_13_17
dmv_race_pct <- data.frame(
  state = dmv_race$State_name,
  low_response_rate = dmv_race$Low_Response_Score, # response rate
  white = dmv_race$NH_White_alone_ACS_13_17 / pop, # white
  black = dmv_race$NH_Blk_alone_ACS_13_17 / pop,   # black or African American
  aian = dmv_race$NH_AIAN_alone_ACS_13_17 / pop,   # American Indian and Alaska Native
  asian = dmv_race$NH_Asian_alone_ACS_13_17 / pop, # Asian
  native_hawaiian_PI = dmv_race$NH_NHOPI_alone_ACS_13_17 / pop, # Native Hawaiian and other Pacific Isl
  other = dmv_race$NH_SOR_alone_ACS_13_17 / pop    # Some other race
  )
unique(dmv_race_pct$state)
```

```
## [1] District of Columbia Maryland             Virginia
## 52 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```
# Gather the different columns (race) into a single column with their percentage as a single column.
dmv_race_pct <- dmv_race_pct %>%
  gather(key = "race", value = "pct", -c(state,low_response_rate))
head(dmv_race_pct)
```

```
##                  state low_response_rate  race       pct
## 1 District of Columbia              19.7 white 0.7916828
## 2 District of Columbia              12.5 white 0.6385117
## 3 District of Columbia              21.2 white 0.7690588
## 4 District of Columbia              22.3 white 0.7742391
## 5 District of Columbia              21.6 white 0.6534587
## 6 District of Columbia              23.1 white 0.7607379
```

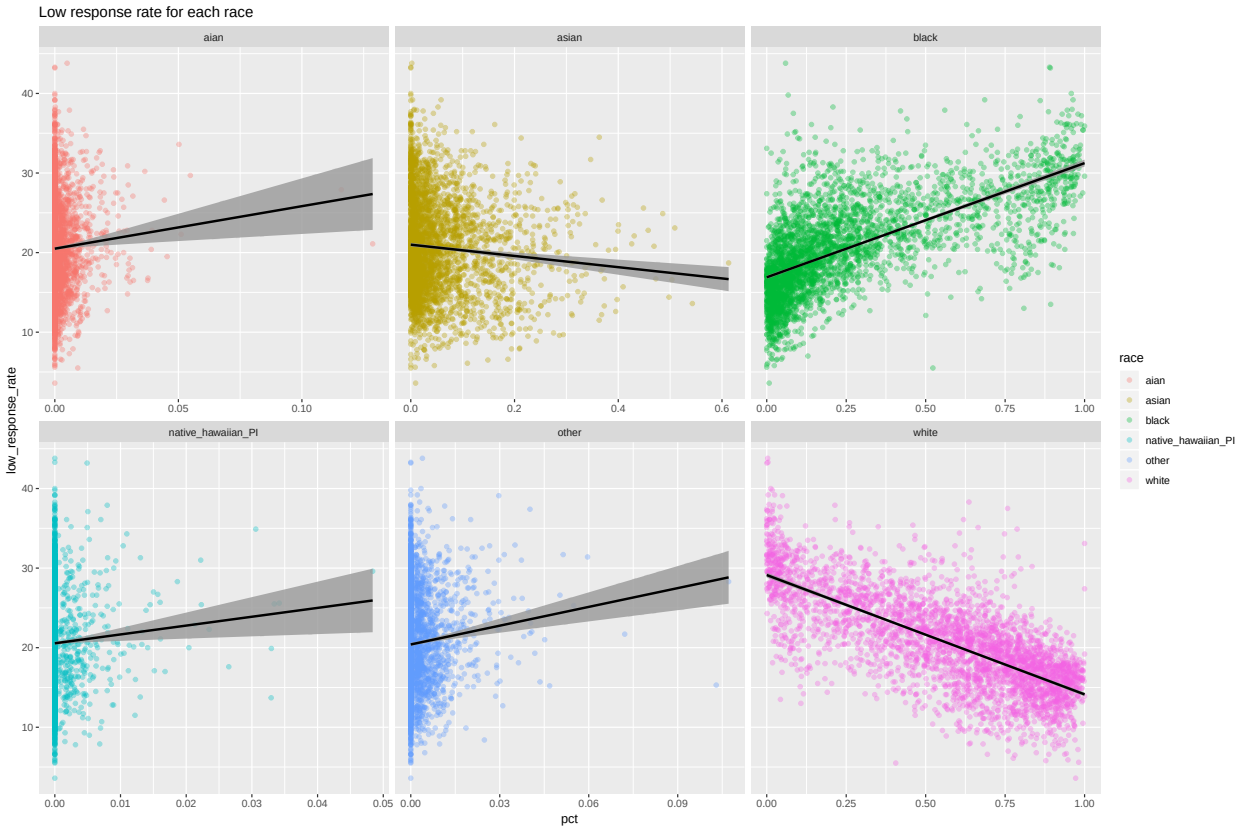
## Race: Plot

The following plots are the plots of low response rate for each race with a linear regression projection. `pct` is the percentage of people who participated in the census belonging to the particular race. `low_response_rate` indicates how likely the particular race would not respond to the census, so the lower the `low_response_rate` is better.
The plots show that the Black, American Indians, Native Hawaiians, and other minority races have an increasing trend of low response rate as the percentage of these races increase in the census participants. Out of which, Black has the most significant increasing trend, indicating that the blacks are least likely to participate in the Census.
Asian and white have a decreasing trend as the percentage of these races increase in the census participants. White has the greatest decreasing trend, indicating that the whites are most likely to participant in the Census.

```
race_stats <- data.frame(summary(dmv_race_pct))
dmv_race_pct %>%
  ggplot(mapping = aes(y = low_response_rate, x = pct, color=race)) +
  geom_point(alpha = 0.35) +
  geom_smooth(color = "black", alpha = 0.8, se=TRUE, method = "lm") +
  facet_wrap(~race, nrow=2, scale = "free_x") +
  ggtitle("Low response rate for each race")
```

```
## Warning: Removed 354 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 354 rows containing missing values (geom_point).
```




The following plots shows the low response rate for each race in each DMV region. I would like to highlight the following races due to the strong and consistent trend shown in their low response rate in all three regions.
- Blacks have an increasing trend in all three regions, indicating that the blacks are less likely to participate

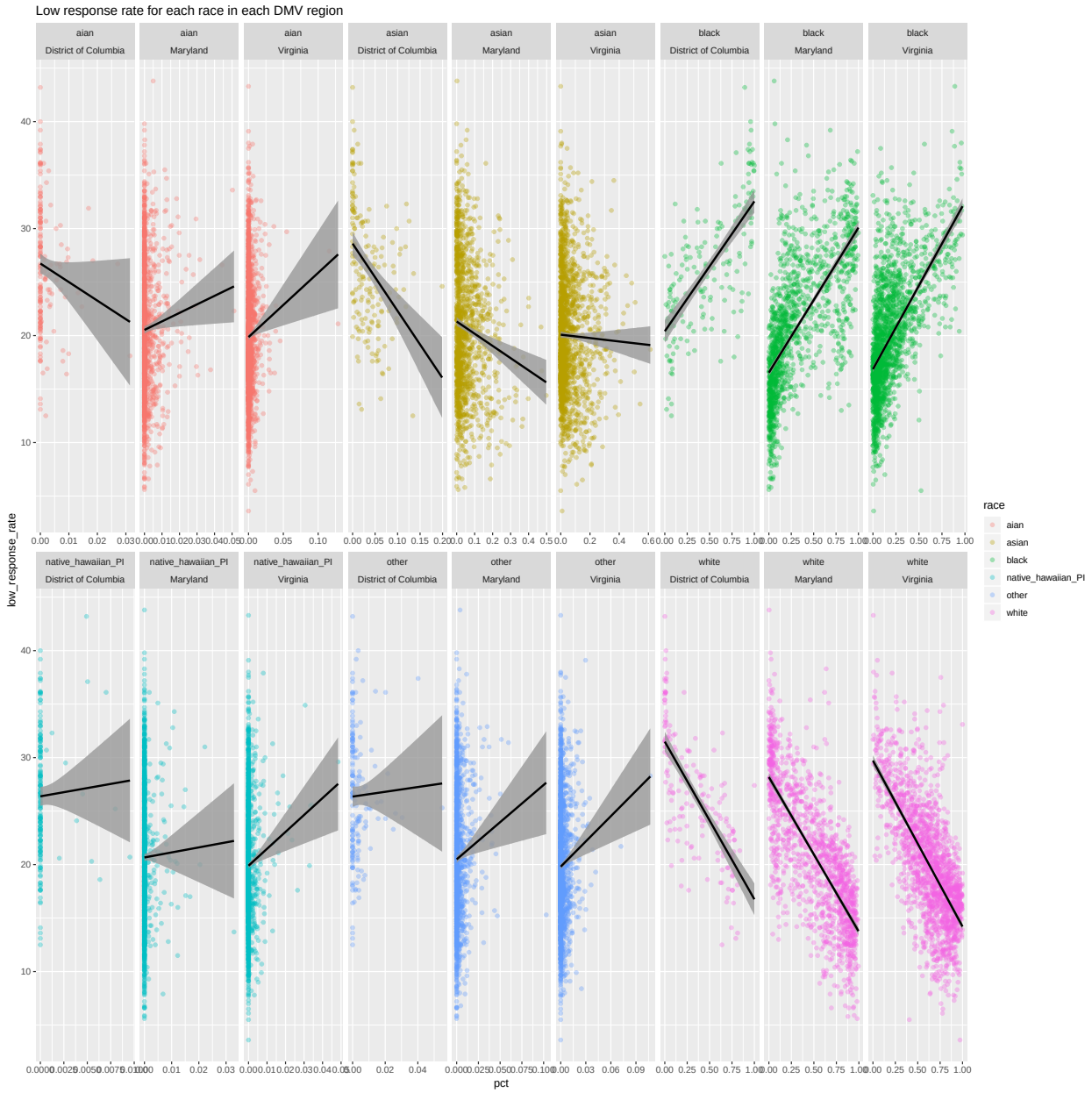
in Census in all three regions.
- Whites have a decreasing trend in all three regions, indicating that the whites are more likely to participate in Census in all three regions.

The rest of the races have less strong trend in their low response rate.
- American Indians show a discrepancy in their low response rate. There is a decreasing trend in D.C. and an increasing trend in Maryland and Virginia.
- Asians, Native Hawaiians, and other minority races maintain a general decreasing trend in all three regions, each with differnt strength.

```
race_stats <- data.frame(summary(dmv_race_pct))
dmv_race_pct %>%
  ggplot(mapping = aes(y = low_response_rate, x = pct, color=race)) +
  geom_point(alpha = 0.35) +
  geom_smooth(color = "black", alpha = 0.8, se=TRUE, method = "lm") +
  facet_wrap(~race*state, nrow=2, scale = "free_x") +
  ggtitle("Low response rate for each race in each DMV region")
```

```
## Warning: Removed 354 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 354 rows containing missing values (geom_point).
```



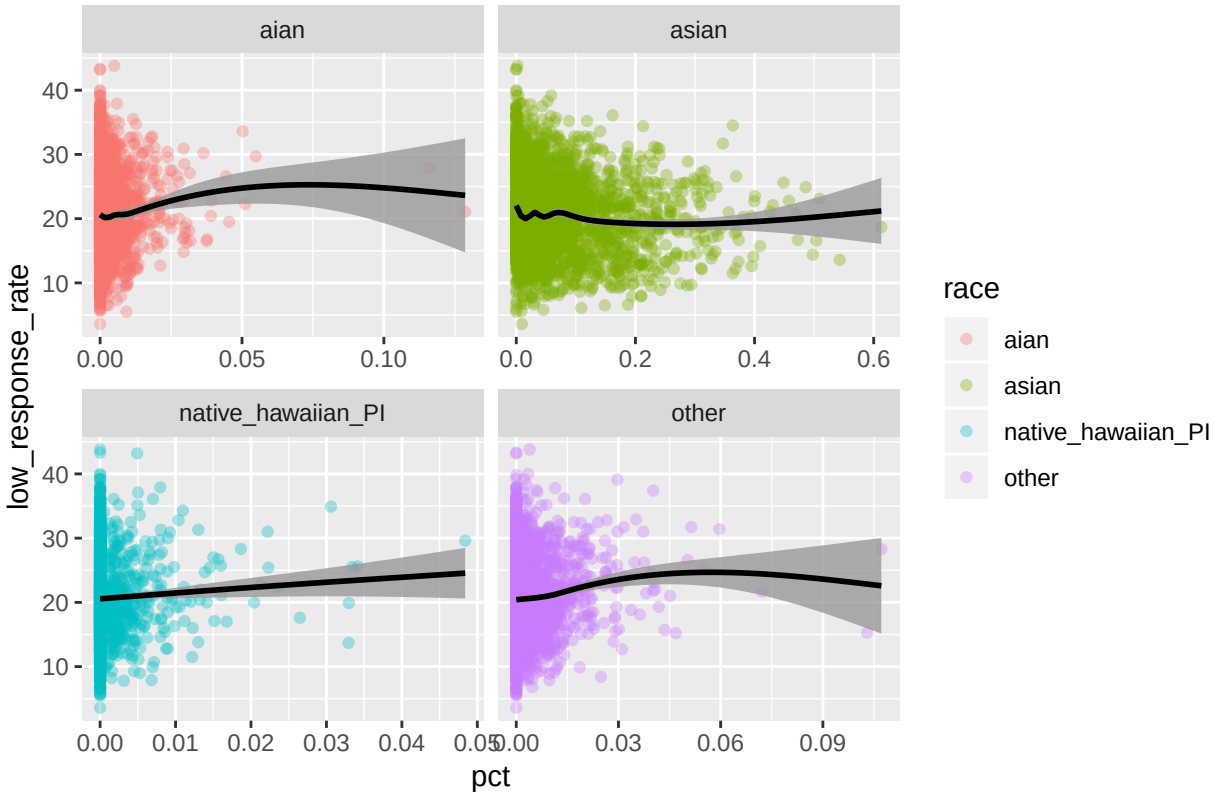

Here, we take a closer look at the other races besides the whites and the blacks. Because these two races have more data points that the other races, the two plots above may not show the trend clearly. The plots show that there is no significant trend in the rest of the races as the trend lines are close to horizontal.

```
dmv_race_pct %>%
  filter(!(race == "white" | race == "black" )) %>%
  ggplot(mapping = aes(y = low_response_rate, x = pct, color=race)) +
  geom_point(alpha = 0.35) +
  geom_smooth(color = "black", alpha = 0.8, se=TRUE) +
  facet_wrap(~race, nrow=2, scale = "free_x") +
  ggtitle("")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 236 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 236 rows containing missing values (geom_point).
```




Next, I will create linear models and correlation tests on the relationship between low response rate and the race white and black.

## Race: Linear model for the black

I used a linear model to examine the relationship between the low response rate and the two races, white and black. The p-value for the slope is less than 0.05, indicating that there is a positive trend between the low response rate and the race black. However, the adjusted r-squared is only 0.4305, indicating that only 43% of variance of the low response rate is explained by the percentage of black in the participants of the Census. Therefore, this is not a very good model for predicting low response rate.

```
# compute a linear model
race_black <- dmv_race_pct %>% filter(race == "black")
lm_black <- lm(low_response_rate ~ pct, data = race_black)
summary(lm_black)
```

```
##
## Call:
## lm(formula = low_response_rate ~ pct, data = race_black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9001  -2.8261  -0.2342   2.4978  26.0522
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  16.8982     0.1059  159.49   <2e-16 ***
## pct          14.3480     0.2816   50.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.511 on 3433 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.4306, Adjusted R-squared:  0.4305
## F-statistic:  2596 on 1 and 3433 DF,  p-value: < 2.2e-16
```

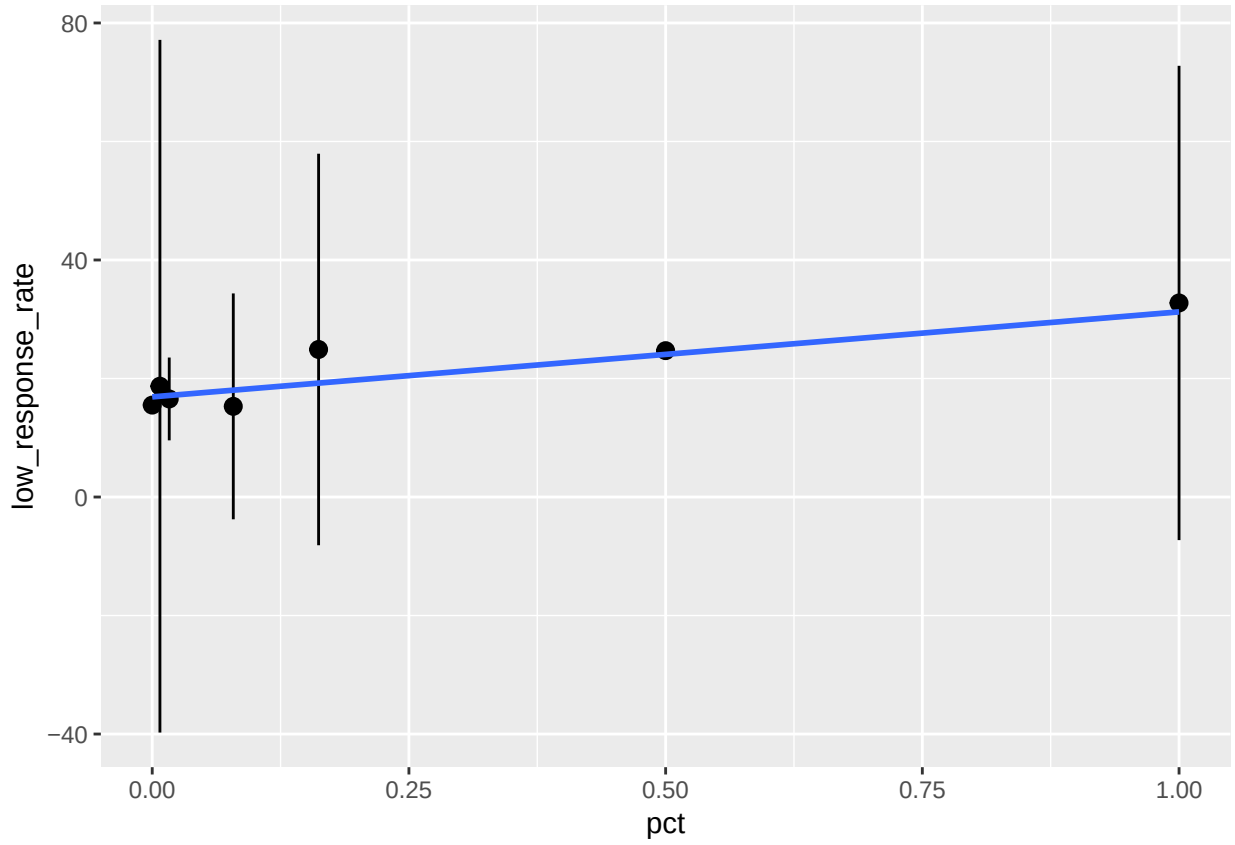
**Assumption checking for LM between percentage of black and low response rate**

```
ggplot(data = race_black, aes(pct,low_response_rate)) +
  stat_summary(fun.data = mean_cl_normal) +
  geom_smooth(method='lm')
```

```
## Warning: Removed 59 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 59 rows containing non-finite values (stat_smooth).
```

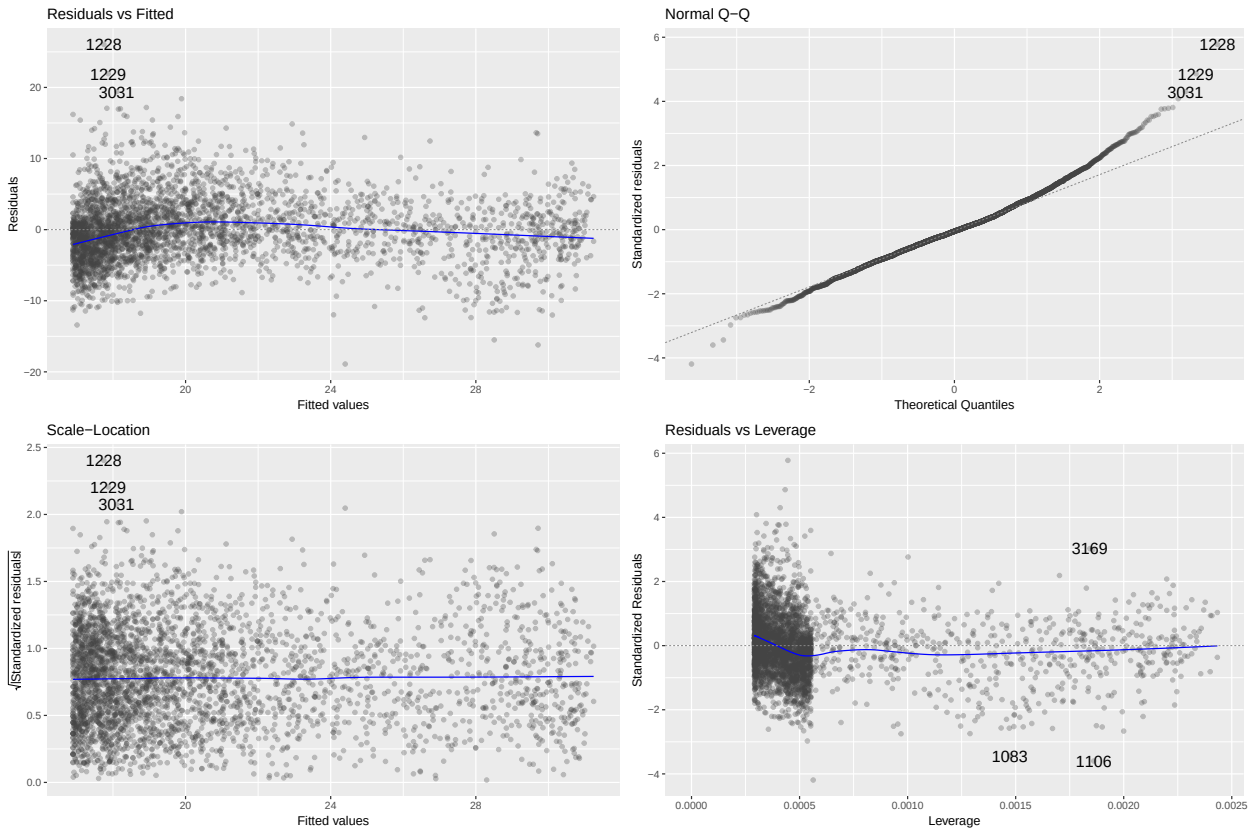
```
## Warning: Removed 3366 rows containing missing values (geom_pointrange).
```




Now I would check the assumptions of the linear model for low response rate and percentage of black. - Homoscedasticity: The Scale-Location shows that the points are concentrated on the left side of the graph, so the points do have equal variances. The conidtion of homoscendasticity is not filfilled.
- Normality: The normal QQ plot curves upwards instead of having a linear line with at least three outliers in the upper right corner. The condition of normality is not fulfilled.
- Indenpendence of observations: Since the data entries are obtained by the American Community Survey for

different area, I assume that the observations are independent.
The Residual vs Leverage plot also shows that there are at least three observations (3169, 1083, and 1106) which could influence the regression result.
Since only one out of three assumptions for a linear relationship is fulfilled, the linear model fitting the low response rate by the percentage of black is not reliable. I would not move on to building a new model with the influential cases removed because the non linearity in the Normal Q-Q plot and the uneven spread of the residuals in the Residuals vs Fitted plot are sufficient to inform that the data is not suitable for fitting a linear model.

```
# Create the plot
autoplot(lm_black, which = c((1:3), 5), ncol=2, label.size = 5, alpha = 0.3)
```




### Correlation between the percentage of blacks and low response rate

The correlation test for the percentage of black and low response rate has a correlation coefficient of 0.66 with a p-value of less than 2.2e-16. The p-value is less than the significance level alpha = 0.05. We can conclude that the percentage of black and the low response rate has a moderate positive correlation.

```
# correlation test for black with low response rate
cor.test(race_black$low_response_rate, race_black$pct, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  race_black$low_response_rate and race_black$pct
## t = 50.955, df = 3433, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6367502 0.6748525
```

```
## sample estimates:
##       cor
## 0.6562195
```

## Race: Linear model for percentage of white and low response rate

Here I created a linear model for the percentage of white and the low response rate. The adjusted r-squared is 0.5216 with a p-value of less than 2.2e16. Since the p-value is less than the significance level alpha 0.05, we can conclude that about 52% of the variance in the low response rate can be explained by the percentage of white.

```
# linear model for white race with low response rate
race_white <- dmv_race_pct %>% filter(race == "white")
lm_white <- lm(low_response_rate ~ pct, data = race_white)
summary(lm_white)
```

```
##
## Call:
## lm(formula = low_response_rate ~ pct, data = race_white)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5403  -2.4516   0.3084   2.5236  19.7367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.127      0.156   186.7   <2e-16 ***
## pct          -14.996      0.245   -61.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.134 on 3433 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.5217, Adjusted R-squared:  0.5216
## F-statistic:  3745 on 1 and 3433 DF,  p-value: < 2.2e-16
```

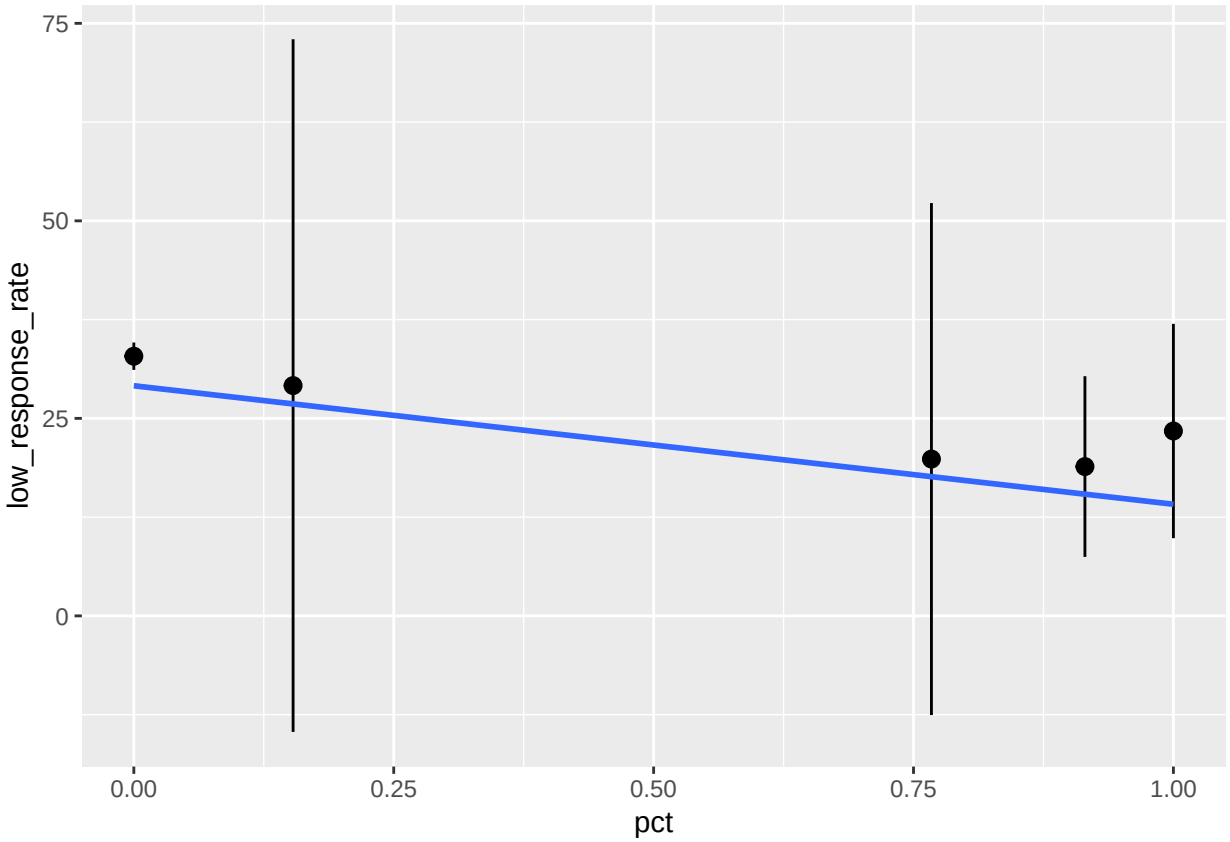
### Assumption checking for LM between percentage of white and low response rate

```
ggplot(data = race_white, aes(pct,low_response_rate)) +
  stat_summary(fun.data = mean_cl_normal) +
  geom_smooth(method='lm')
```

```
## Warning: Removed 59 rows containing non-finite values (stat_summary).
```

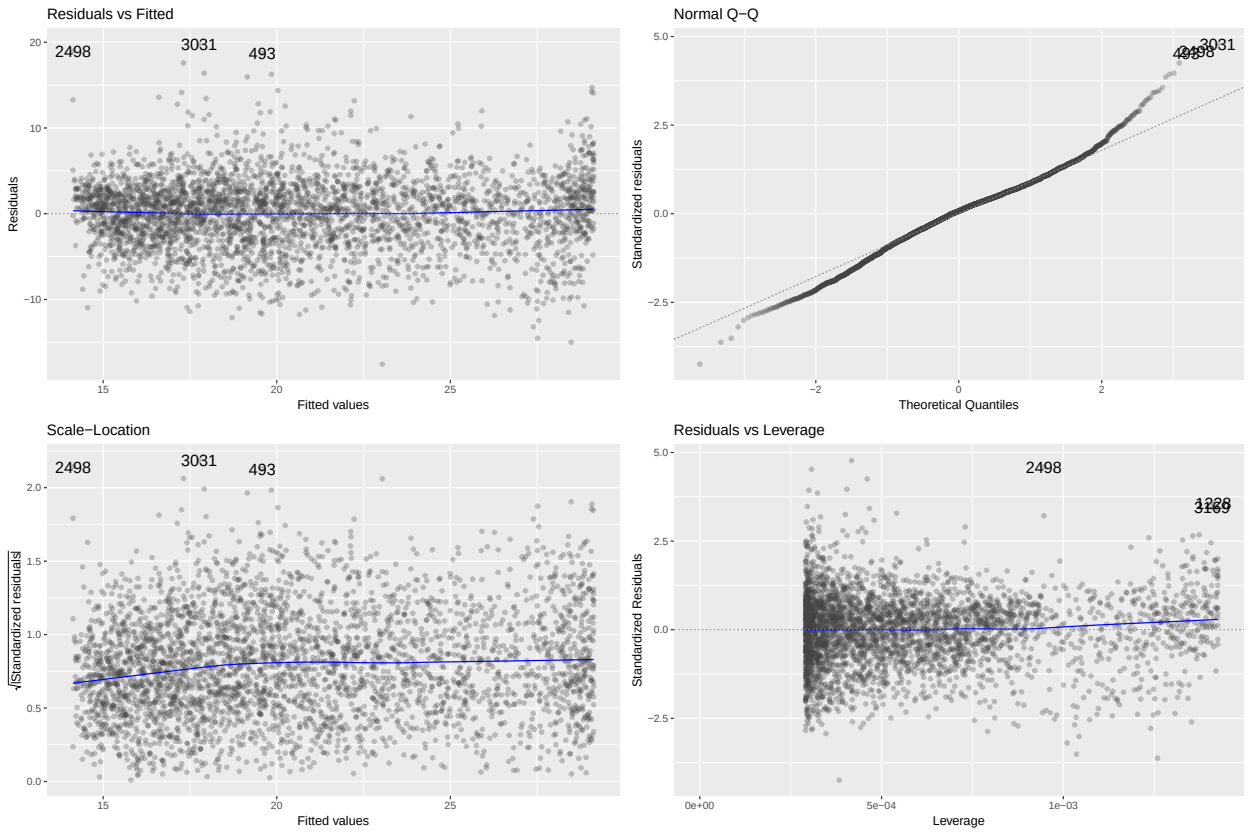
```
## Warning: Removed 59 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3410 rows containing missing values (geom_pointrange).
```




Now I would check the assumptions of the linear model for low response rate and percentage of white.
- Homoscedasticity: The Scale-Location plot shows evenly spread out points with the exception of three outliers on the upper left corner. I would consider that the condition of homoscedasticity is fulfilled.
- Normality: The Normarl Q-Q plot shows a curved line with many outliers in the upper right corner, so the condition of normality is not fulfilled.
- Indenpendence of the observations: Since the data entries are obtained by the American Community Survey for different area, I assume that the observations are independent.
The Residual vs Leverage plot also shows that there are at least three influential points (2498, 3169, and one other points) which could affect the regression result.
Since only two out of three assumptions for a linear relationship is fulfilled, the model is not reliable. I would stop here on the linear model for low response rate and the percentage of white because the Normal Q-Q plot shows that there are too many outliers.

```
# Create the plot
autoplot(lm_white, which = c((1:3), 5), ncol=2, label.size = 5, alpha = 0.3)
```




The correlation test for the percentage of white and low response rate has a correlation coefficient of -0.72 with a p-value of less than 2.2e-16. The p-value is less than the significance level alpha = 0.05. We can conclude that the percentage of black and the low response rate has a moderate positive correlation.

```
# correlation test for percentage of white and low response rate
cor.test(race_white$low_response_rate, race_white$pct, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  race_white$low_response_rate and race_white$pct
## t = -61.196, df = 3433, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7379245 -0.7059155
## sample estimates:
##        cor
## -0.7223066
```

## Race: Conclusion

The scatter plots of low_response_rate by the percentage of race shows that the percentage of black and white have the strongest trend with the low response rate. The low response rate increases as the percentage of black increases and decreases as the percentage of white increases. The higher the low response rate, the corresponding area have less people responding to the Census. Although the linear models built to use either race to estimate the low response rate are both unreliable, there is a high correlation between the two races and low response rate. White has a -0.72 and Black has a 0.66 correlation statistic from the Pearson's correlation test both with p-values less than 2.2e-16. Thus, I conclude that the percentage of white and black

in a community is moderately correlated with the low response rate.

## Language spoken

Here I created a dataset which includes the low response rate and the percentage of people age 5 years or older speaking a certain language.

```
unique(dmv_lang$State_name)
```

```
## [1] District of Columbia Maryland             Virginia
## 52 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```
names(dmv_lang)
```

```
##  [1] "State_name"                      "County_name"
##  [3] "Low_Response_Score"              "Othr_Lang_ACS_13_17"
##  [5] "Othr_Lang_ACSMOE_13_17"          "Age5p_Only_English_ACS_13_17"
##  [7] "Age5p_Only_English_ACSMOE_13_17" "Age5p_Spanish_ACS_13_17"
##  [9] "Age5p_Spanish_ACSMOE_13_17"      "Age5p_French_ACS_13_17"
## [11] "Age5p_French_ACSMOE_13_17"       "Age5p_German_ACS_13_17"
## [13] "Age5p_German_ACSMOE_13_17"       "Age5p_Russian_ACS_13_17"
## [15] "Age5p_Russian_ACSMOE_13_17"      "Age5p_OthEuro_ACS_13_17"
## [17] "Age5p_OthEuro_ACSMOE_13_17"      "Age5p_Korean_ACS_13_17"
## [19] "Age5p_Korean_ACSMOE_13_17"       "Age5p_Chinese_ACS_13_17"
## [21] "Age5p_Chinese_ACSMOE_13_17"      "Age5p_Vietnamese_ACS_13_17"
## [23] "Age5p_Vietnamese_ACSMOE_13_17"   "Age5p_Tagalog_ACS_13_17"
## [25] "Age5p_Tagalog_ACSMOE_13_17"      "Age5p_OthAsian_ACS_13_17"
## [27] "Age5p_OthAsian_ACSMOE_13_17"     "Age5p_Arabic_ACS_13_17"
## [29] "Age5p_Arabic_ACSMOE_13_17"       "Age5p_OthUnSp_ACS_13_17"
## [31] "Age5p_OthUnSp_ACSMOE_13_17"      "ENG_VW_SPAN_ACS_13_17"
## [33] "ENG_VW_SPAN_ACSMOE_13_17"        "ENG_VW_INDO_EURO_ACS_13_17"
## [35] "ENG_VW_INDO_EURO_ACSMOE_13_17"   "ENG_VW_API_ACS_13_17"
## [37] "ENG_VW_API_ACSMOE_13_17"         "ENG_VW_OTHER_ACS_13_17"
## [39] "ENG_VW_OTHER_ACSMOE_13_17"       "ENG_VW_ACS_13_17"
## [41] "ENG_VW_ACSMOE_13_17"
```

```
# person of age 5 years or older who speak english less than very well who speak ----.
dmv_lang_pct <- data.frame(
  state = dmv_lang$State_name,
  low_response_rate = dmv_lang$Low_Response_Score,
  english = dmv_lang$Age5p_Only_English_ACS_13_17 / pop,
  spanish = dmv_lang$Age5p_Spanish_ACS_13_17 / pop,
  french = dmv_lang$Age5p_French_ACS_13_17 / pop,
  german = dmv_lang$Age5p_German_ACS_13_17 / pop,
  chinese = dmv_lang$Age5p_Chinese_ACS_13_17 / pop,
  russian = dmv_lang$Age5p_Russian_ACS_13_17 / pop,
  viet = dmv_lang$Age5p_Vietnamese_ACS_13_17 / pop,
  arabic = dmv_lang$Age5p_Arabic_ACS_13_17 / pop,
  korean = dmv_lang$Age5p_Korean_ACS_13_17 /pop,
  other_euro = dmv_lang$Age5p_OthEuro_ACS_13_17 / pop, # such as Romanian
  tagalog = dmv_lang$Age5p_Tagalog_ACS_13_17 / pop, # phillipine
  others = dmv_lang$Age5p_OthUnSp_ACS_13_17 / pop # unspecified
)
```

```
dmv_lang_pct <- dmv_lang_pct %>%
  gather(key= "lang", value="pct", -c(state, low_response_rate))
```

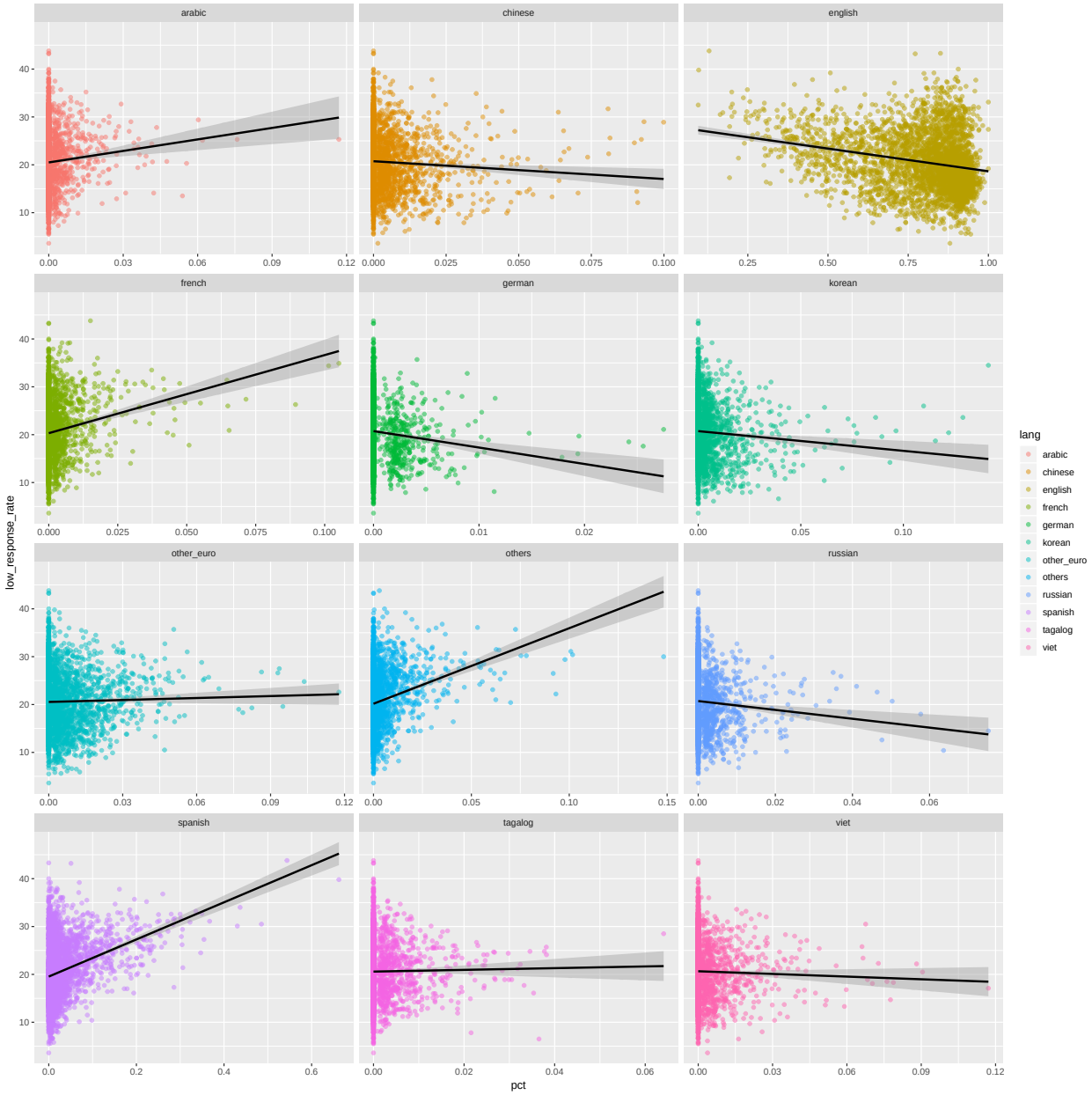
### Language spoken: plots

The plots here shows the low response rate by the percentage of people speaking a certain language. The plot for English shows a decreasing trend line with a narrow confidence interval. Other languages show a large confidence intervals around their trend lines, so I conclude that only English might be useful to estimate the low response rate in a linear model.

```
dmv_lang_pct %>%
  ggplot(mapping = aes(y=low_response_rate, x=pct,  color=lang)) +
  geom_point(alpha = 0.5) +
  geom_smooth(color = "black", method = "lm") +
  facet_wrap(~lang, nrow = 4, scale = "free_x")
```

```
## Warning: Removed 708 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 708 rows containing missing values (geom_point).
```

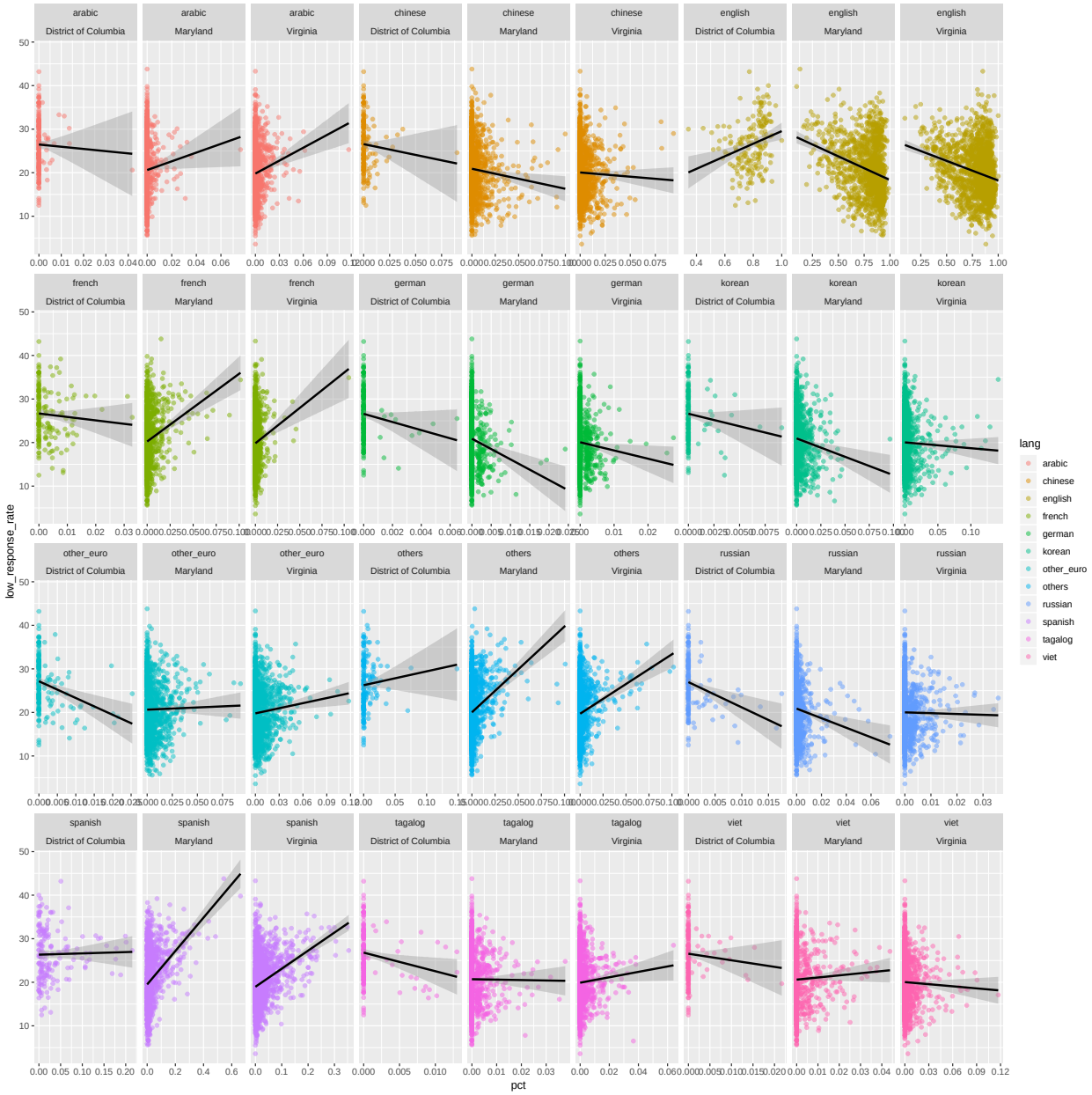



The plots below show the percentage of people speaking certain language by low response rate and states. Plots which have narrow confidence intervals around the linear model line are English in Maryland and Virginia, and Spanish in Virginia.

```
dmv_lang_pct %>%
  ggplot(mapping = aes(y=low_response_rate, x=pct,  color=lang)) +
  geom_point(alpha = 0.5) +
  geom_smooth(color = "black", method = "lm") +
  facet_wrap(~lang*state, nrow = 4, scale = "free_x")
```

```
## Warning: Removed 708 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 708 rows containing missing values (geom_point).
```



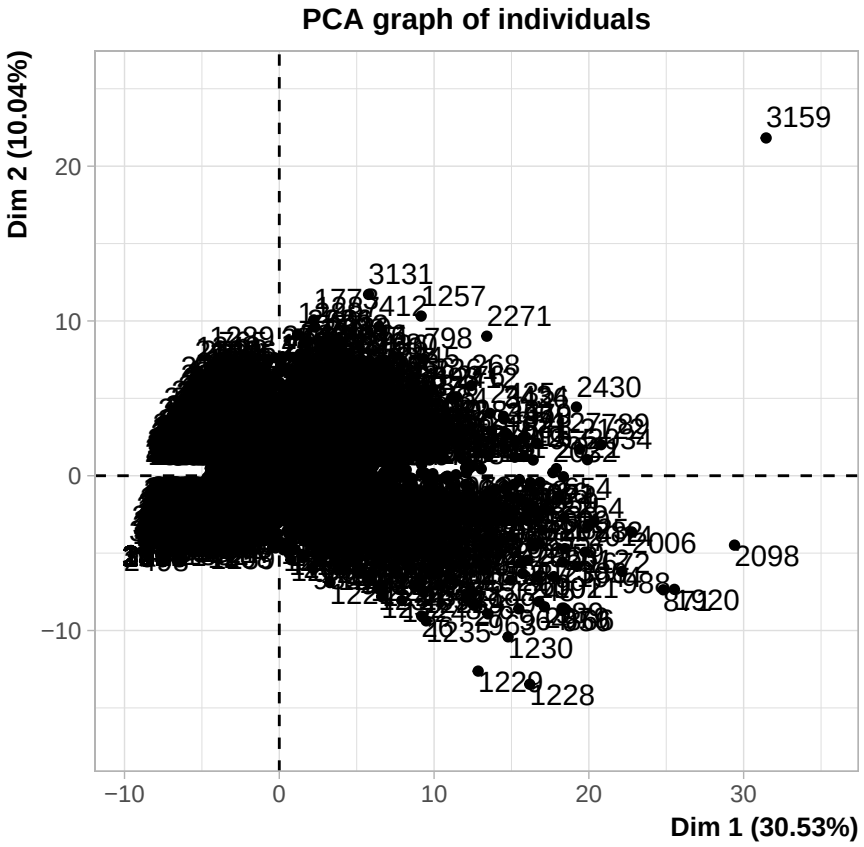

## Language Spoken: Conclusion

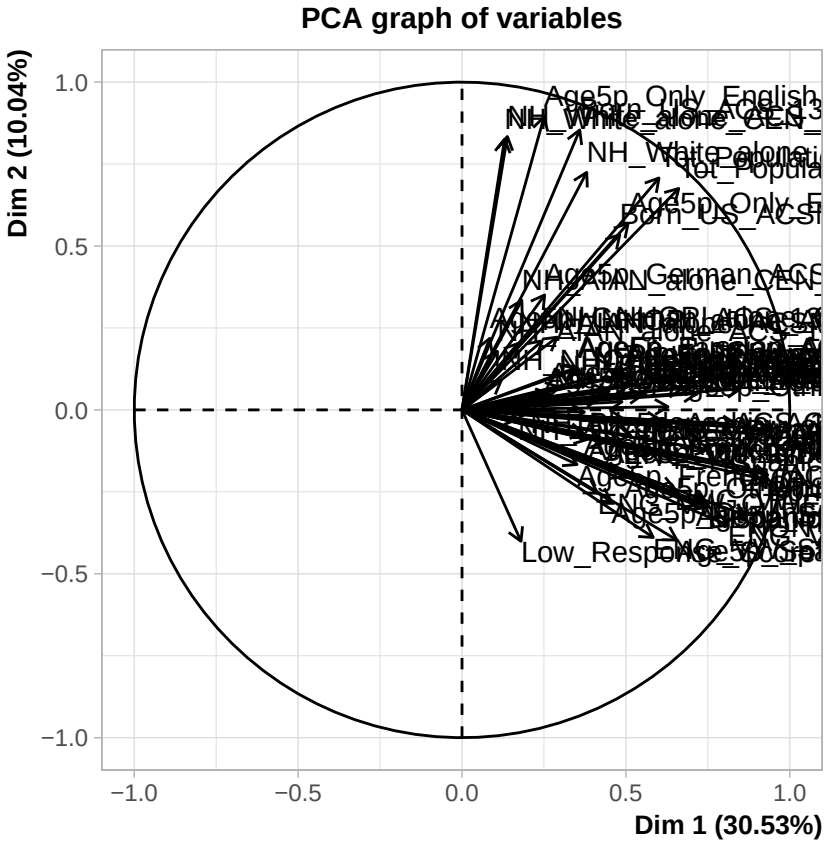
Concluding from the scatter plots and the linear model trend line, I conclude that there may be a negative relation between the percentage of people speaking English and the low response rate in both Maryland and Virginia. There may be a positive relationship between the percentage of people speaking Spanish and the low response rate in Virginia. A positive relationship between the percentage of language spoken and the low response rate indicates that the area would have less people responding to the Census if the percentage of people speaknig a certain language incerases. A negative relationship means more people respond to the Census if the percentage of people speaking the language increases. Further correlation tests and linear model assumption check should be done to confirm this claim.

## Perform Principle Component Analysis on the cultural columns.

This section is a practice for performing PCA for feature grouping. No conclusion will be made here.

```
cultural <- dmv_cultural[, -c(1:2)] # remove the state and county
nb = estim_ncpPCA(cultural, ncp.max = 4)
cultural.comp = imputePCA(cultural, ncp=4)
cultural.pca <- PCA(cultural.comp$completeObs)
```



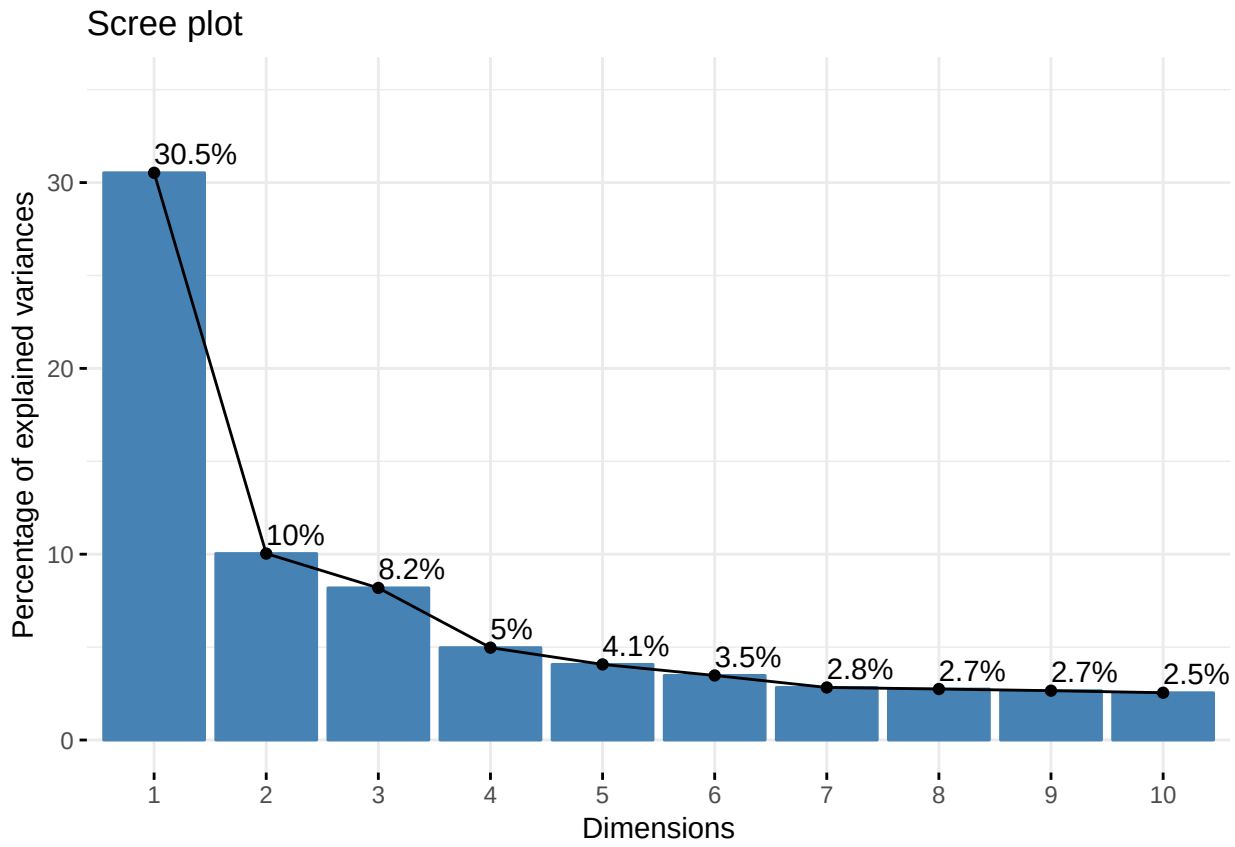



```
get_eig(cultural.pca)
```

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  2.136888e+01     3.052697e+01                     30.52697
## Dim.2  7.024762e+00     1.003537e+01                     40.56235
## Dim.3  5.731008e+00     8.187154e+00                     48.74950
## Dim.4  3.480231e+00     4.971759e+00                     53.72126
## Dim.5  2.849271e+00     4.070387e+00                     57.79165
## Dim.6  2.432549e+00     3.475071e+00                     61.26672
## Dim.7  1.980160e+00     2.828800e+00                     64.09552
## Dim.8  1.921869e+00     2.745527e+00                     66.84104
## Dim.9  1.859221e+00     2.656030e+00                     69.49707
## Dim.10 1.781337e+00     2.544767e+00                     72.04184
## Dim.11 1.729648e+00     2.470926e+00                     74.51277
## Dim.12 1.681790e+00     2.402557e+00                     76.91532
## Dim.13 1.500940e+00     2.144200e+00                     79.05952
## Dim.14 1.426322e+00     2.037603e+00                     81.09713
## Dim.15 1.349810e+00     1.928300e+00                     83.02543
## Dim.16 1.320775e+00     1.886821e+00                     84.91225
## Dim.17 1.205460e+00     1.722085e+00                     86.63433
## Dim.18 1.127345e+00     1.610493e+00                     88.24482
## Dim.19 8.793203e-01     1.256172e+00                     89.50100
## Dim.20 7.363412e-01     1.051916e+00                     90.55291
## Dim.21 6.672962e-01     9.532802e-01                     91.50619
## Dim.22 5.684876e-01     8.121251e-01                     92.31832
## Dim.23 5.252834e-01     7.504048e-01                     93.06872
```

```
## Dim.24 4.729185e-01     6.755978e-01                 93.74432
## Dim.25 4.534356e-01     6.477651e-01                 94.39209
## Dim.26 4.088267e-01     5.840381e-01                 94.97612
## Dim.27 3.653376e-01     5.219109e-01                 95.49803
## Dim.28 3.186082e-01     4.551545e-01                 95.95319
## Dim.29 3.082313e-01     4.403305e-01                 96.39352
## Dim.30 2.478322e-01     3.540459e-01                 96.74757
## Dim.31 1.875369e-01     2.679098e-01                 97.01548
## Dim.32 1.695435e-01     2.422049e-01                 97.25768
## Dim.33 1.621732e-01     2.316760e-01                 97.48936
## Dim.34 1.296643e-01     1.852348e-01                 97.67459
## Dim.35 1.268183e-01     1.811691e-01                 97.85576
## Dim.36 1.153636e-01     1.648051e-01                 98.02057
## Dim.37 1.064885e-01     1.521264e-01                 98.17269
## Dim.38 9.722699e-02     1.388957e-01                 98.31159
## Dim.39 9.073785e-02     1.296255e-01                 98.44121
## Dim.40 8.632251e-02     1.233179e-01                 98.56453
## Dim.41 7.513308e-02     1.073330e-01                 98.67186
## Dim.42 7.412515e-02     1.058931e-01                 98.77776
## Dim.43 6.889571e-02     9.842244e-02                 98.87618
## Dim.44 6.778552e-02     9.683645e-02                 98.97302
## Dim.45 6.444004e-02     9.205721e-02                 99.06507
## Dim.46 6.172001e-02     8.817144e-02                 99.15324
## Dim.47 5.521668e-02     7.888097e-02                 99.23213
## Dim.48 5.349570e-02     7.642243e-02                 99.30855
## Dim.49 5.030362e-02     7.186232e-02                 99.38041
## Dim.50 4.691416e-02     6.702023e-02                 99.44743
## Dim.51 4.476896e-02     6.395566e-02                 99.51139
## Dim.52 4.208234e-02     6.011763e-02                 99.57150
## Dim.53 3.956982e-02     5.652832e-02                 99.62803
## Dim.54 3.723981e-02     5.319973e-02                 99.68123
## Dim.55 3.412724e-02     4.875320e-02                 99.72999
## Dim.56 3.271098e-02     4.672997e-02                 99.77672
## Dim.57 2.752494e-02     3.932135e-02                 99.81604
## Dim.58 2.510619e-02     3.586598e-02                 99.85190
## Dim.59 2.353450e-02     3.362071e-02                 99.88552
## Dim.60 2.009224e-02     2.870320e-02                 99.91423
## Dim.61 1.839870e-02     2.628386e-02                 99.94051
## Dim.62 1.578510e-02     2.255014e-02                 99.96306
## Dim.63 1.259621e-02     1.799458e-02                 99.98105
## Dim.64 9.923974e-03     1.417711e-02                 99.99523
## Dim.65 2.184104e-03     3.120148e-03                 99.99835
## Dim.66 1.050668e-03     1.500954e-03                 99.99985
## Dim.67 1.028758e-04     1.469654e-04                100.00000
## Dim.68 2.998445e-29     4.283493e-29                100.00000
## Dim.69 3.690774e-30     5.272534e-30                100.00000
## Dim.70 2.944161e-30     4.205944e-30                100.00000
```

```r
# visualize eigenvalues/variances
fviz_screeplot(cultural.pca, addlabels=TRUE, ylim=c(0, 35))
```




```
# Extract the result for variables
var <- get_pca(cultural.pca, "var")
var
```

```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

```
head(var$coord)
```

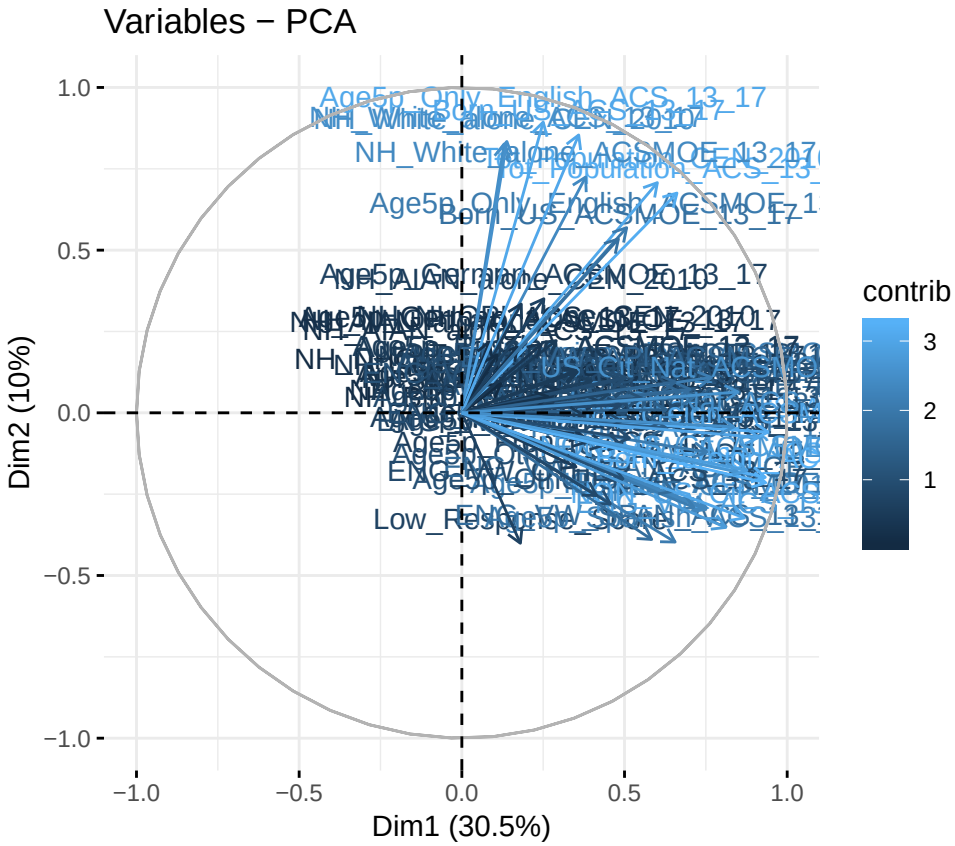
```
##                              Dim.1      Dim.2     Dim.3       Dim.4
## Low_Response_Score       0.1799488 -0.4008368 0.4966171  0.29732528
## Tot_Population_CEN_2010  0.6009922  0.7074561 0.2144726 -0.06755545
## Tot_Population_ACS_13_17 0.6612358  0.6754221 0.2051534 -0.05033372
## Hispanic_CEN_2010        0.7500530 -0.2970018 0.2197723 -0.38030830
## Hispanic_ACS_13_17       0.7521597 -0.2918662 0.2576888 -0.42522372
## Hispanic_ACSMOE_13_17    0.7528241 -0.1206431 0.3027670 -0.26972869
##                                 Dim.5
## Low_Response_Score       -0.232420790
## Tot_Population_CEN_2010  -0.002122991
## Tot_Population_ACS_13_17 -0.007280188
## Hispanic_CEN_2010        -0.066705822
## Hispanic_ACS_13_17       -0.089783565
```

```
## Hispanic_ACSMOE_13_17    -0.083885513
```

```
head(var$contrib)
```

```
##                              Dim.1     Dim.2     Dim.3      Dim.4
## Low_Response_Score       0.1515361 2.2871975 4.3034064 2.54012780
## Tot_Population_CEN_2010  1.6902692 7.1247126 0.8026250 0.13113320
## Tot_Population_ACS_13_17 2.0461191 6.4941005 0.7343893 0.07279641
## Hispanic_CEN_2010        2.6327043 1.2557019 0.8427810 4.15588480
## Hispanic_ACS_13_17       2.6475149 1.2126511 1.1586710 5.19549421
## Hispanic_ACSMOE_13_17    2.6521937 0.2071924 1.5995071 2.09048090
##                                 Dim.5
## Low_Response_Score       1.8959033594
## Tot_Population_CEN_2010  0.0001581839
## Tot_Population_ACS_13_17 0.0018601647
## Hispanic_CEN_2010        0.1561686042
## Hispanic_ACS_13_17       0.2829175865
## Hispanic_ACSMOE_13_17    0.2469677161
```

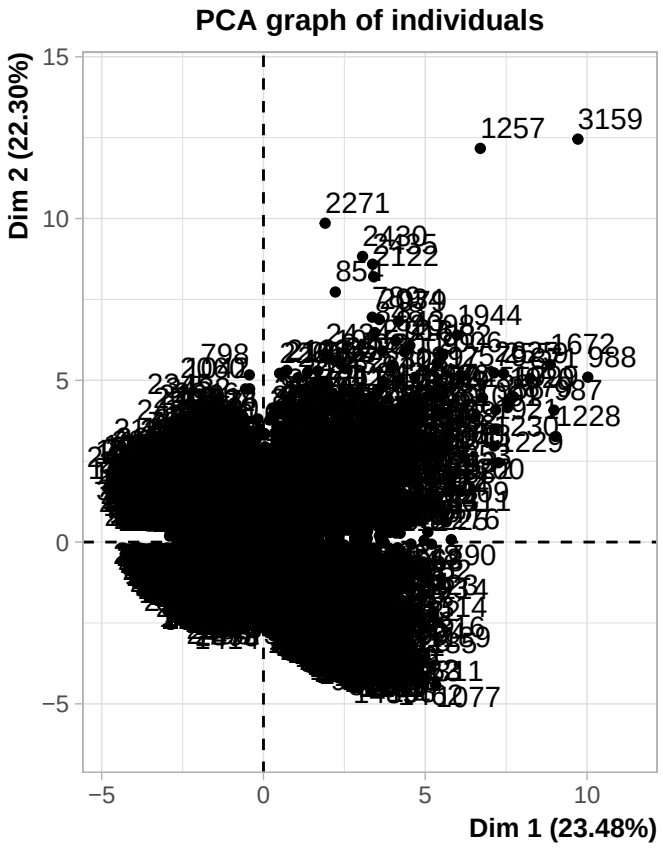
```
fviz_pca_var(cultural.pca, col.var = "contrib",
             gradiant.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = FALSE) # avoid text overlapping
```

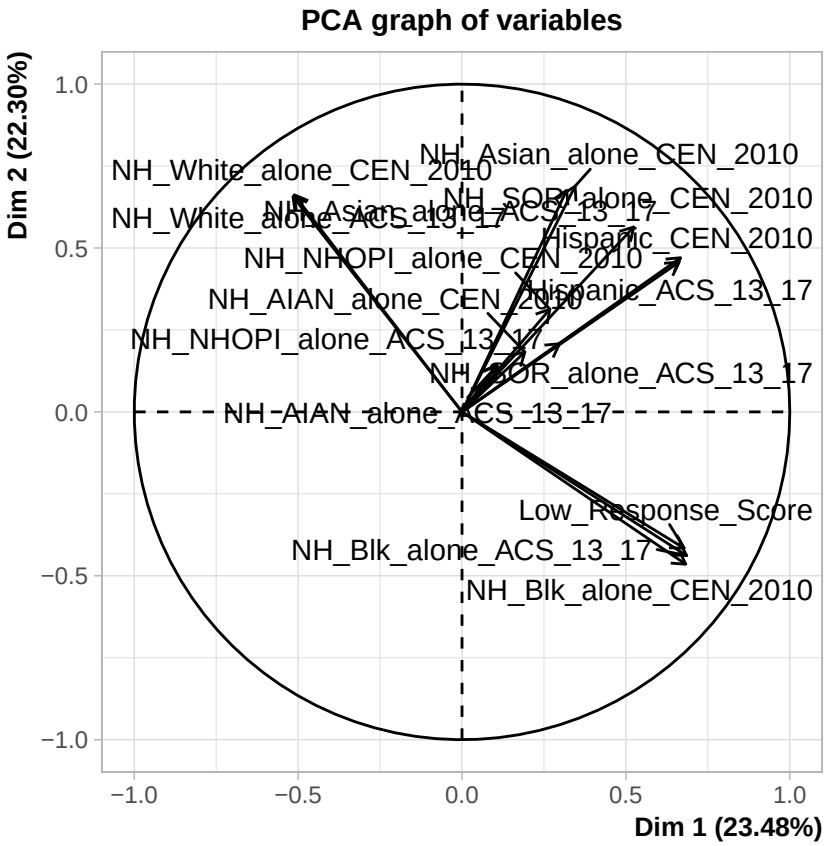



## PCA on race

```
race <- dmv_race[, -c(1:2)] # remove the state and county
nb = estim_ncpPCA(race, ncp.max = 4)
```

```
race.comp = imputePCA(race, ncp=2)
race.pca <- PCA(race.comp$completeObs)
```

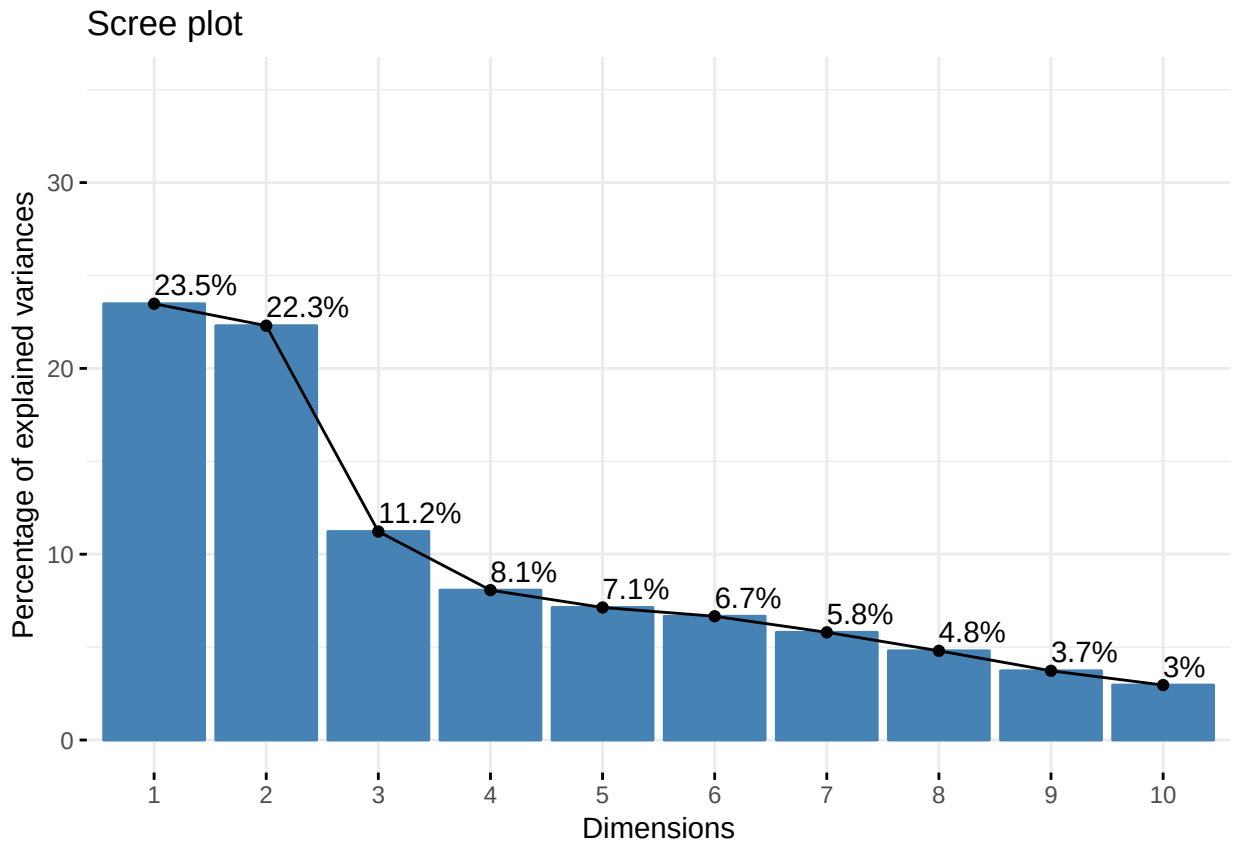





```
get_eig(race.pca)
```

```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1  3.52176220       23.4784146                    23.47841
## Dim.2  3.34481778       22.2987852                    45.77720
## Dim.3  1.68319691       11.2213127                    56.99851
## Dim.4  1.20994360        8.0662907                    65.06480
## Dim.5  1.06985834        7.1323889                    72.19719
## Dim.6  0.99909326        6.6606217                    78.85781
## Dim.7  0.86953300        5.7968866                    84.65470
## Dim.8  0.72013211        4.8008807                    89.45558
## Dim.9  0.55884456        3.7256304                    93.18121
## Dim.10 0.44369677        2.9579785                    96.13919
## Dim.11 0.42656823        2.8437882                    98.98298
## Dim.12 0.05580701        0.3720468                    99.35503
## Dim.13 0.05041282        0.3360855                    99.69111
## Dim.14 0.02634408        0.1756272                    99.86674
## Dim.15 0.01998934        0.1332623                   100.00000
```

```
# visualize eigenvalues/variances
fviz_screeplot(race.pca, addlabels=TRUE, ylim=c(0, 35))
```




```
# Extract the result for variables
var <- get_pca(race.pca, "var")
var
```

```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

```
head(var$coord)
```

```
##                              Dim.1      Dim.2       Dim.3        Dim.4
## Low_Response_Score        0.6774811 -0.4157665  0.01274073 -0.095338995
## Hispanic_CEN_2010         0.6669456  0.4702593 -0.14495416 -0.484293320
## Hispanic_ACS_13_17        0.6625414  0.4569797 -0.12558081 -0.500518201
## NH_White_alone_CEN_2010  -0.5119875  0.6621287  0.39944192 -0.003760494
## NH_White_alone_ACS_13_17 -0.5156038  0.6565848  0.38792966  0.002342630
## NH_Blk_alone_CEN_2010     0.6825499 -0.4646727  0.34199776  0.276838828
##                                 Dim.5
## Low_Response_Score         0.0077507662
## Hispanic_CEN_2010         -0.1130345958
## Hispanic_ACS_13_17        -0.1040131862
## NH_White_alone_CEN_2010    0.0028136643
## NH_White_alone_ACS_13_17  -0.0002462632
```

```
## NH_Blk_alone_CEN_2010     0.0343370346
```

```
head(var$contrib)
```

```
##                              Dim.1     Dim.2       Dim.3        Dim.4
## Low_Response_Score       13.032700  5.168048 0.009643929  0.751235345
## Hispanic_CEN_2010        12.630507  6.611535 1.248321562 19.384376146
## Hispanic_ACS_13_17       12.464247  6.243402 0.936939681 20.704970776
## NH_White_alone_CEN_2010   7.443184 13.107274 9.479214410  0.001168758
## NH_White_alone_ACS_13_17  7.548699 12.888703 8.940690225  0.000453568
## NH_Blk_alone_CEN_2010    13.228443  6.455382 6.948828740  6.334157751
##                                Dim.5
## Low_Response_Score       5.615171e-03
## Hispanic_CEN_2010        1.194253e+00
## Hispanic_ACS_13_17       1.011231e+00
## NH_White_alone_CEN_2010  7.399771e-04
## NH_White_alone_ACS_13_17 5.668562e-06
## NH_Blk_alone_CEN_2010    1.102045e-01
```

```
fviz_pca_var(race.pca, col.var = "contrib",
             gradiant.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE) # avoid text overlapping
```