Hrytsyuk Kostyantyn. Course Project

```r
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# require(foreach)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
# require(gridExtra)
require(survey)
```

```
## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart
```

```r
# Set path to data
setwd('./')
getwd()
```

```
## [1] "C:/D/Code/youth-habits-research"
```

```r
fix_weights <- function(df){
  # Weight column is stored as factor in data
  # Firstly, we have to unfactor weight column
  # w <- unfactor(df$weight)
  # Now, w is a vector of characters with ',' as a separator between integer and fractional parts
  # Secondly, we have to substitute ',' with '.'
  # w <- gsub(',', '.', w)
  # Thirdly, we have to convert string vector to numeric vector
  # w <- as.numeric(w)

  # Here is implementation all code above in one line
  df$weight <- as.numeric(gsub(',','.', df$weight))
  return(df)
}
```

```r
get_parts <- function(design, col_name) {
    # Convert col_name to formula
    col_name <- as.formula(paste('~',col_name))
    # Group by category (col_name) in a 1d table
    col_parts <- svytable(col_name, design) %>%
                 as.data.frame() %>%
    # Calculate percentage of each part in data frame
                 mutate(Part = Freq/sum(Freq))

    return(col_parts)
}

get_relation_table <- function(design, var1, var2) {
    # Formula for table creation
    rel_formula <- as.formula(paste('~', var1, '+', var2))

    # rel_table <- svytable(~age + sleep.amount, design) %>%
    rel_table <- svytable(rel_formula, design) %>%
                 as.data.frame() %>%
                 group_by(!!var2) %>%
                 mutate(Part = Freq/sum(Freq))

    return(rel_table)
}

add_dumm_var <- function(df, var_ranges, source_col_name) {
    for (i in 1:length(var_ranges)) {
      source_col_name <- enquo(source_col_name)
      source_col <- df %>% select(!! source_col_name)
      dum_col <- if_else(source_col == i, 1, 0)
      df[, paste('is_', var_ranges[i], sep = '')] <- dum_col
    }

    return(df)
}

convert_to_dumm_var <- function(df, col_name, condition, add_col = FALSE) {
    a <- col_name
```

```
    source_col_name <- enquo(a)
    source_col <- df %>% select(!! source_col_name)
    if (add_col) {
      df[, paste('is_', col_name, sep = '')] <- if_else(condition(source_col), 1, 0)
    } else {
      df[, col_name] <- if_else(condition(source_col), 1, 0)
    }

    return(df)
}

conduct_survey_test <- function(statement, form, des, crit_alpha = 0.001,
                                t_test = FALSE, chi_sq_test = FALSE) {
  if(!xor(t_test, chi_sq_test)) {
    return()
  }
  cat('\nNull hypothesis:', statement, '\n')
  cat('Alpha:', crit_alpha*100,'%')

  if(t_test) {
    res <- svyttest(form, design = des)
  }
  else {
    res <- svychisq(form, design = des, statistic = "Chisq")
  }
  print(res)
  if (res$p.value < crit_alpha) {
    cat('Null hypothesis is wrong!\n')
  } else {
    cat('Null hypothesis is correct!\n')
  }
  cat('-------------------------------------------------------')
}
```

```
# Setting up data descriptions
age_ranges <- c('12 years old \n or younger', '13 years', '14 years'
                , '15 years', '16 years', '17 years', '18 years old \n or older')
race_7_types <- c('American Indian/\nAlaska Native', 'Asian', 'Black or \n African American' ,'Hispanic
sex_types <- c('Female', 'Male')
grade_types <- c('9th_grade', '10th_grade', '11th_grade', '12th_grade', 'ungraded')
sleep_ranges <- c('4 or less hours', '5 hours', '6 hours', '7 hours', '8 hours', '9 hours', '10 or more
marks_ranges <- c('Mostly A\'s', 'Mostly B\'s', 'Mostly C\'s', 'Mostly D\'s', 'Mostly F\'s', 'None of th

bad_habits <- c('is_smoking', 'is_drinking_alcohol', 'using_marijuana', 'using_cocaine',
                'using_inhalant', 'using_heroin', 'using_methamphetamines', 'using_ecstasy',
                'using_synthetic', 'using_needle')

good_habits <- c('is_drinking_juice', 'is_eating_fruits', 'is_eating_salad',
                    'is_eating_potatoes', 'is_eating_carrots', 'is_eating_other_vegetables')

days <- c('0 days', '1 day', '2 days', '3 days', '4 days', '5 days', '6 days', '7 days')

categories <- c('age', 'race7', 'sex', 'grade', 'marks')
```

```
subscripts_categories <- list(age_ranges, race_7_types, sex_types, grade_types, marks_ranges)
```

```
# Data loading
# Each file loaded below contains US states
# which title begins with letters mentioned in the filename
am <- read.csv("A-M.csv", row.names = NULL, sep = ';')
nz <- read.csv("N-Z.csv", row.names = NULL, sep = ';')

# Data cleaning
# We need only response with answer on the question 89:
# During the past 12 months, how would you describe your grades in school?

# Calculating first year when response was received
base_year_am <- am %>% filter(!is.na(q89)) %>% select(year) %>% slice(1)
base_year_nz <- nz %>% filter(!is.na(q89)) %>% select(year) %>% slice(1)

# Defining the latest year one with the purpose to have a common base year
base_year <- max(base_year_am, base_year_nz)

# Setting base year
# Also, we need some important data as age, race , grade, sex and next questions:
# Unhealth questions
# q32 - do you smoke?
# q42 - do you drink alcohol?
# q46 - used marijuana
# q49 - used any form of cocaine
# q50 - used inhalant to get high
# q51 - used heroin
# q52 - used methamphetamines
# q53 - used ecstasy
# q54 - used synthetic marijuana
# q57 - did you use a needle to inject any illegal drug into your body?
# Health questions
# q70 - 100% fruit juices
# q71 - do you eat fruits?
# q72 - do you eat green salad?
# q73 - do you eat potatoes?
# q74 - do you eat carrots?
# q75 - do you eat other vegetables?
# q78 - do you eat breakfast?
# q79 - were you physically active?
# q88 - how many hours of sleep do you get?
# q89 - your grades in school?
am <- am %>% filter(year >= base_year) %>%
             filter_at(vars(age, sex, grade, race7, q89), all_vars(!is.na(.))) %>%
             select(sitename, year, weight, stratum, PSU, age, sex, grade, race7,
                    q32, q42, q46, q49, q50, q51, q52, q53, q54, q57, q70,
                    q71, q72, q73, q74, q75, q75, q78, q79, q88, q89)

nz <- nz %>% filter(year >= base_year) %>%
             filter_at(vars(age, sex, grade, race7, q89), all_vars(!is.na(.))) %>%
             select(sitename, year, weight, stratum, PSU, age, sex, grade, race7,
                    q32, q42, q46, q49, q50, q51, q52, q53, q54, q57, q70,
```

```
                    q71, q72, q73, q74, q75, q75, q78, q79, q88, q89)
```

```
# Creating final data frames
usa <- rbind(am, nz)

# Environment storage optimization
rm(am, nz, base_year_am, base_year_nz, base_year)

# Changing names of columns for informative ones
usa <- usa %>% rename(is_smoking = q32,
                    is_drinking_alcohol = q42,
                    using_marijuana = q46,
                    using_cocaine = q49,
                    using_inhalant = q50,
                    using_heroin = q51,
                    using_methamphetamines = q52,
                    using_ecstasy = q53,
                    using_synthetic = q54,
                    using_needle = q57,
                    is_drinking_juice = q70,
                    is_eating_fruits = q71,
                    is_eating_salad = q72,
                    is_eating_potatoes = q73,
                    is_eating_carrots = q74,
                    is_eating_other_vegetables = q75,
                    having_breakfast = q78,
                    doing_sport = q79,
                    sleep.amount = q88,
                    marks = q89)
```

```
# Data cleaning
usa <- fix_weights(usa)
```

```
# Answers from 2017
usa_2017 <- usa %>% filter(year == 2017)
```

```
# Adding dummy variables

# Dummy variable for race 7
race_7_col_names <- c('alaska_native', 'asian', 'black' ,'latino', 'hawaiian', 'white', 'other_race')

usa <- add_dumm_var(usa, race_7_col_names, 'race7')
usa <- add_dumm_var(usa, grade_types, 'grade')
usa <- add_dumm_var(usa, sex_types, 'sex')

# Converting existing varibles to the dummy ones
# Bad habits
for (i in 1:length(bad_habits)) {
    usa <- convert_to_dumm_var(usa, bad_habits[i], function(x) !is.na(x) & x > 1)
}

# Drugs
usa <- usa %>% mutate(used_drugs := if_else(using_marijuana == 1 |
```

```r
                                                    using_cocaine == 1 |
                                                    using_inhalant == 1 |
                                                    using_heroin == 1 |
                                                    using_methamphetamines == 1 |
                                                    using_ecstasy == 1 |
                                                    using_synthetic == 1 |
                                                    using_needle == 1, 1, 0))

# Good habits
for (i in 1:length(good_habits)) {
    usa <- convert_to_dumm_var(usa, good_habits[i], function(x) !is.na(x) & x >= 2)
}

# Vegetables
usa <- usa %>% mutate(is_eating_vegetables := if_else(is_eating_carrots == 1 |
                                                        is_eating_salad == 1 |
                                                        is_eating_potatoes == 1 |
                                                        is_eating_other_vegetables
, 1, 0))

# Breakfast
usa <- convert_to_dumm_var(usa, 'having_breakfast', function(x) !is.na(x) & x > 3, add_col = TRUE)

#Sport
usa <- convert_to_dumm_var(usa, 'doing_sport', function(x) !is.na(x) & x > 3, add_col = TRUE)

rm(race_7_col_names, good_habits, bad_habits)
```

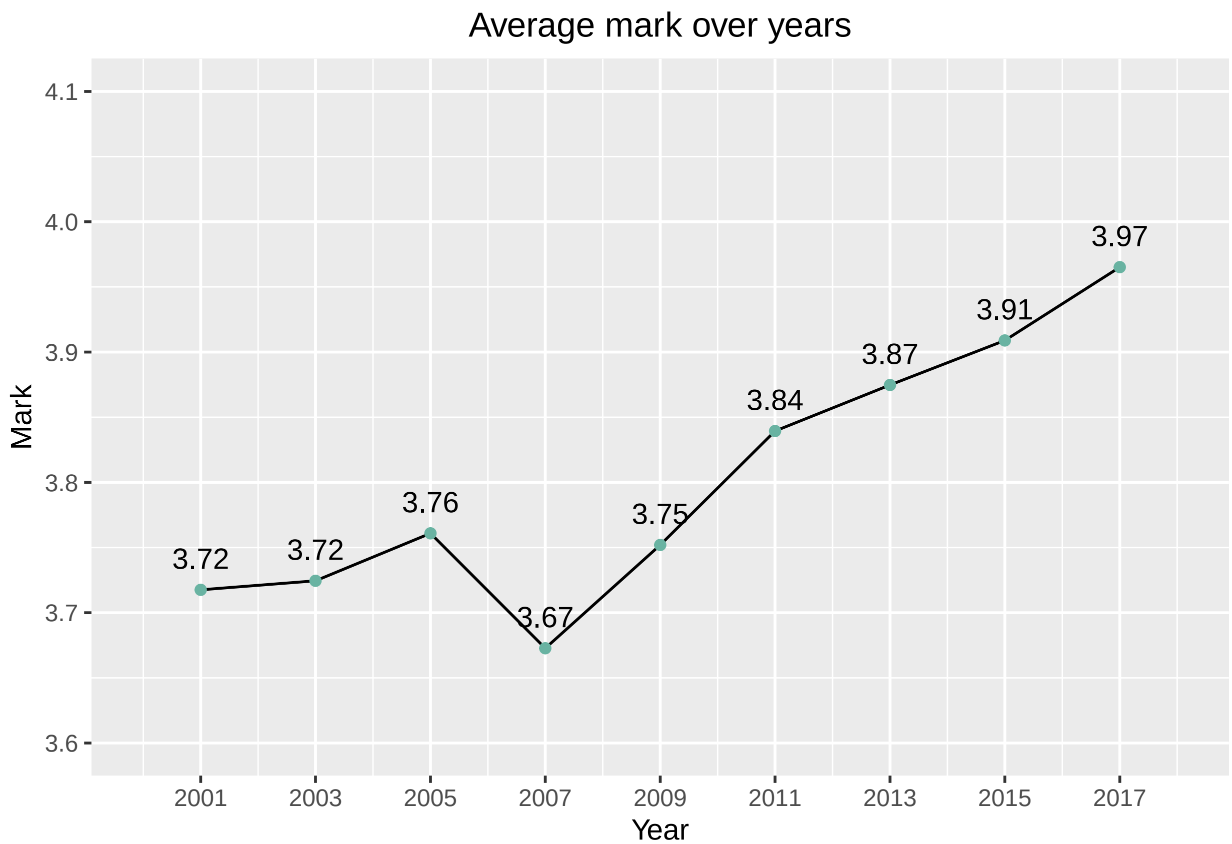
```r
# Changing values of columns for informative ones

#age
usa <- usa %>% mutate(age = case_when(
  age == 1 ~ 12,
  age == 2 ~ 13,
  age == 3 ~ 14,
  age == 4 ~ 15,
  age == 5 ~ 16,
  age == 6 ~ 17,
  age == 7 ~ 18
))

# sleep.amount
usa <- usa %>% mutate(sleep.amount = case_when(
  sleep.amount == 1 ~ 4,
  sleep.amount == 2 ~ 5,
  sleep.amount == 3 ~ 6,
  sleep.amount == 4 ~ 7,
  sleep.amount == 5 ~ 8,
  sleep.amount == 6 ~ 9,
  sleep.amount == 7 ~ 10
))

#Additional dummy variable for sleep.amount
```

```r
usa <- usa %>% mutate( sleep.more.8.hours := case_when(
    sleep.amount >= 8 ~ 1,
    sleep.amount < 8 ~ 0))

# marks
usa <- usa %>% mutate(marks = case_when(
  marks == 1 ~ 5,
  marks == 2 ~ 4,
  marks == 3 ~ 3,
  marks == 4 ~ 2,
  marks == 5 ~ 1,
  marks == 6 ~ 0,
  marks == 7 ~ 0
))

# grades
usa <- usa %>% mutate(grade = case_when(
  grade == 1 ~ 9,
  grade == 2 ~ 10,
  grade == 3 ~ 11,
  grade == 4 ~ 12,
  grade == 5 ~ 0
))

# Converting sex variable in (0,1) form
usa <- usa %>% mutate(sex = case_when(
  sex == 1 ~ 0,
  sex == 2 ~ 1
))
```

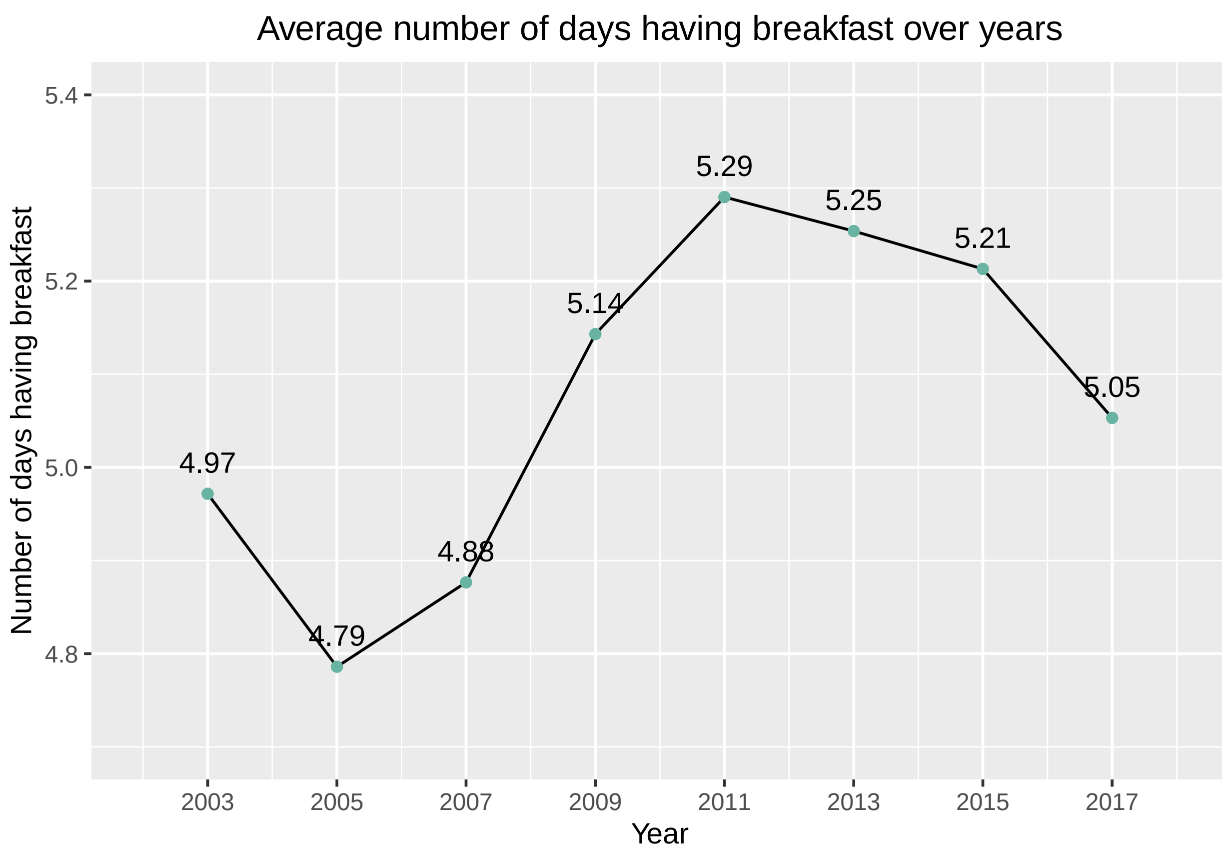
```r
# Answers before 2017
usa_prev <- usa %>% filter(year < 2017)

# Answers from 2017
usa_2017 <- usa %>% filter(year == 2017)
```

```r
# To handle weights properly we have to create survey design object using package "survey"
# We are passing such arguments to svydesign() function
# weights - number of individuals that this record presents/ alternative to sampling probabilities
# strata - id of the part of the entire population
# id - id of the record inside of strata
# data - source data frame
# nest - indicator that shows us that id is nested inside of strata
# Symbol '~' is used to point that ~name is a name of a column of data object
yrbs <- svydesign(id=~PSU, weights = ~weight, strata = ~stratum, data = usa, nest = TRUE)
yrbs_2017 <- svydesign(id=~PSU, weights = ~weight, strata = ~stratum, data = usa_2017, nest = TRUE)
yrbs_prev <- svydesign(id=~PSU, weights = ~weight, strata = ~stratum, data = usa_prev, nest = TRUE)

designs <- list(yrbs, yrbs_2017, yrbs_prev)
names(designs) <- c('respondents since 2009', 'respondents since 2017', 'respondents before 2017')
```

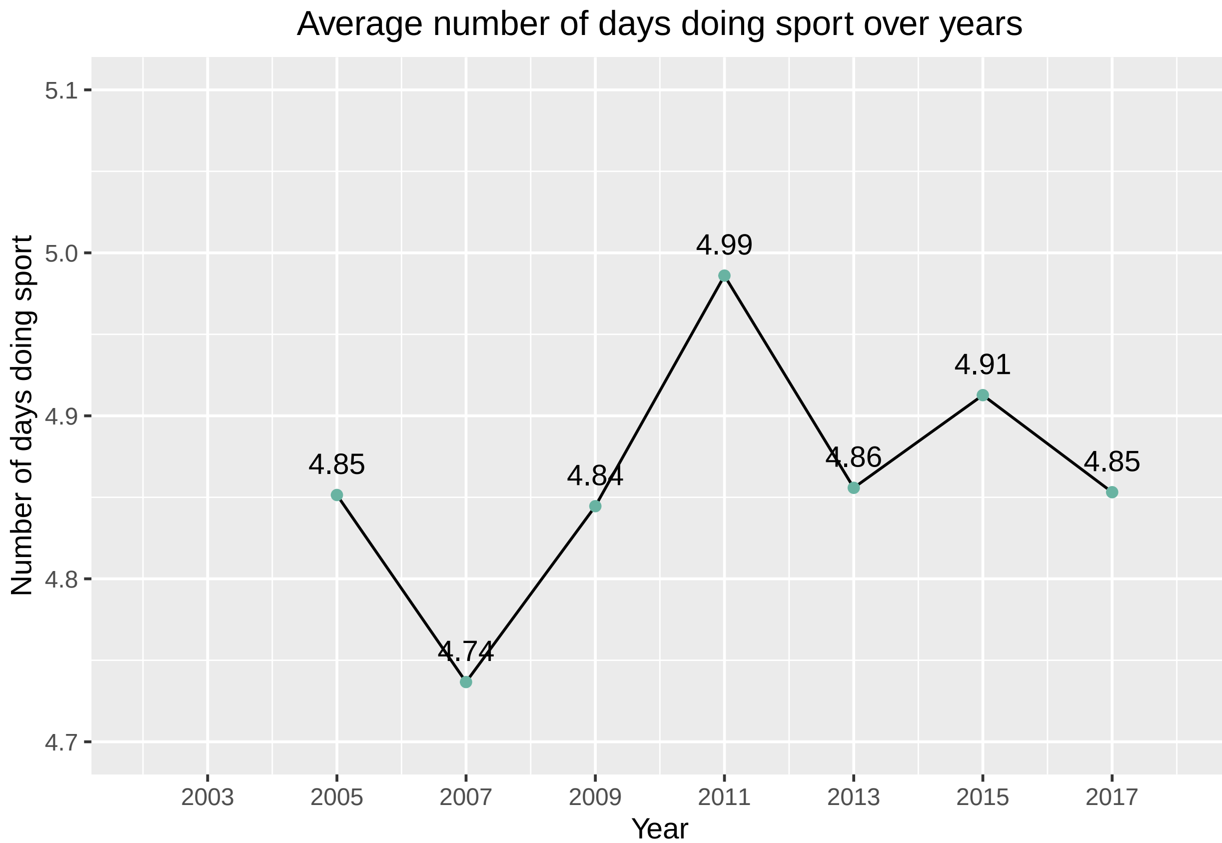
```
# Visualizing average mark over years
years_mean_marks <- svyby(~marks, ~year, yrbs, FUN = svymean, na.rm = TRUE, keep.names = FALSE)

years_mean_marks <- years_mean_marks %>% mutate(year = as.Date(paste(year, 1, 1, sep='-')))

ggplot(years_mean_marks, aes(x = year, y = marks)) +
  geom_line() + geom_point(color="#69b3a2") +
   labs(x = 'Year', y = 'Mark', title = 'Average mark over years') + scale_x_date(date_breaks = '2 year
  expand_limits(x = c(as.Date('2000-1-1'), as.Date('2018-1-1')), y=c(3.6,4.1)) +
  geom_text(aes(label = round(marks,2)), vjust = -1)
```



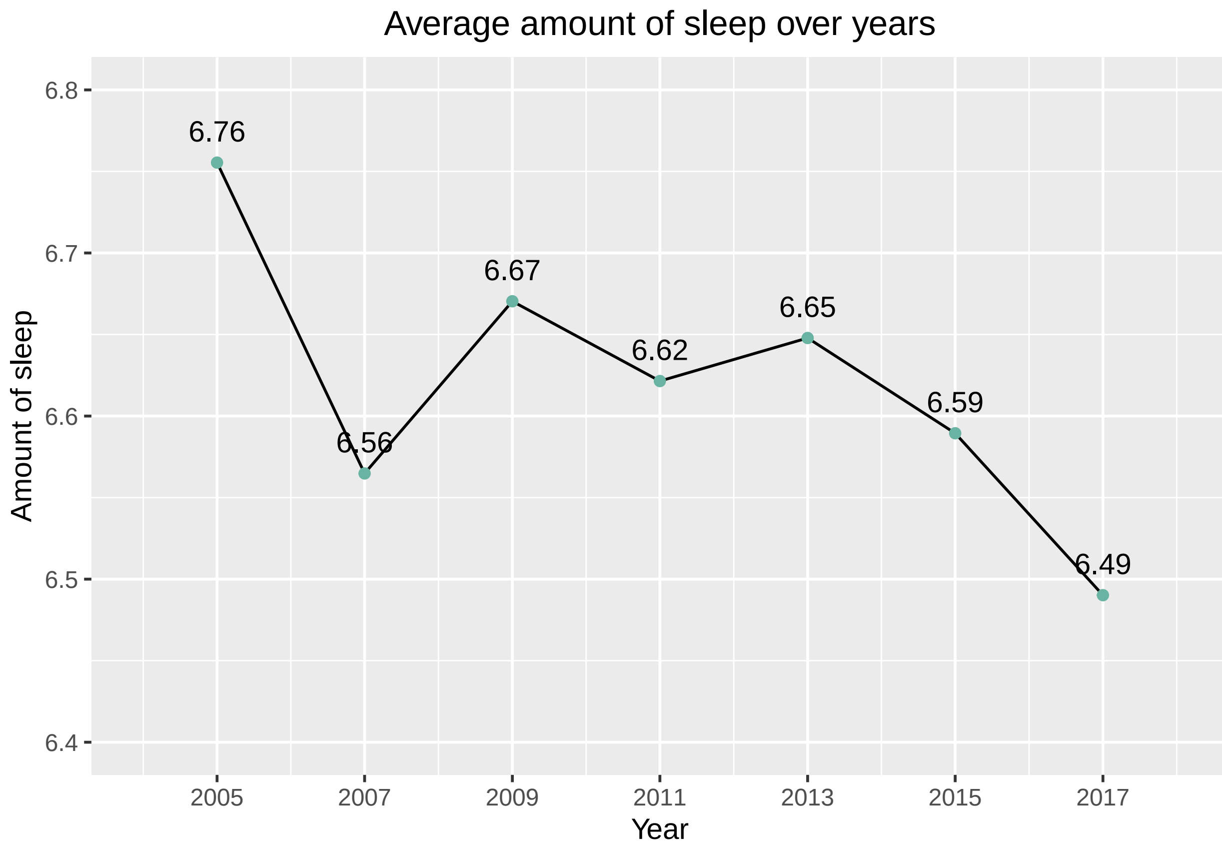

```
# Visualizing average number of days having breakfast over years
years_mean_breakfast <- svyby(~having_breakfast, ~year, yrbs, FUN = svymean, na.rm = TRUE, keep.names =

years_mean_breakfast <- years_mean_breakfast %>% filter(having_breakfast > 0) %>%
                        mutate(year = as.Date(paste(year, 1, 1, sep='-')))

ggplot(years_mean_breakfast, aes(x = year, y = having_breakfast)) +
  geom_line() + geom_point(color="#69b3a2") +
   labs(x = 'Year', y = 'Number of days having breakfast',
        title = 'Average number of days having breakfast over years') + scale_x_date(date_breaks = '2 ye
  expand_limits(x = c(as.Date('2002-1-1'), as.Date('2018-1-1')), y=c(4.7,5.4)) +
  geom_text(aes(label = round(having_breakfast,2)), vjust = -1)
```




```r
# Visualizing average number of days doing sport over years
years_mean_sport <- svyby(~doing_sport, ~year, yrbs, FUN = svymean, na.rm = TRUE, keep.names = FALSE)

years_mean_sport <- years_mean_sport %>% filter(doing_sport > 0) %>%
                    mutate(year = as.Date(paste(year, 1, 1, sep='-')))

ggplot(years_mean_sport, aes(x = year, y = doing_sport)) +
  geom_line() + geom_point(color="#69b3a2") +
   labs(x = 'Year', y = 'Number of days doing sport',
       title = 'Average number of days doing sport over years') + scale_x_date(date_breaks = '2 year',
  expand_limits(x = c(as.Date('2002-1-1'), as.Date('2018-1-1')), y=c(4.7,5.1)) +
  geom_text(aes(label = round(doing_sport,2)), vjust = -1)
```

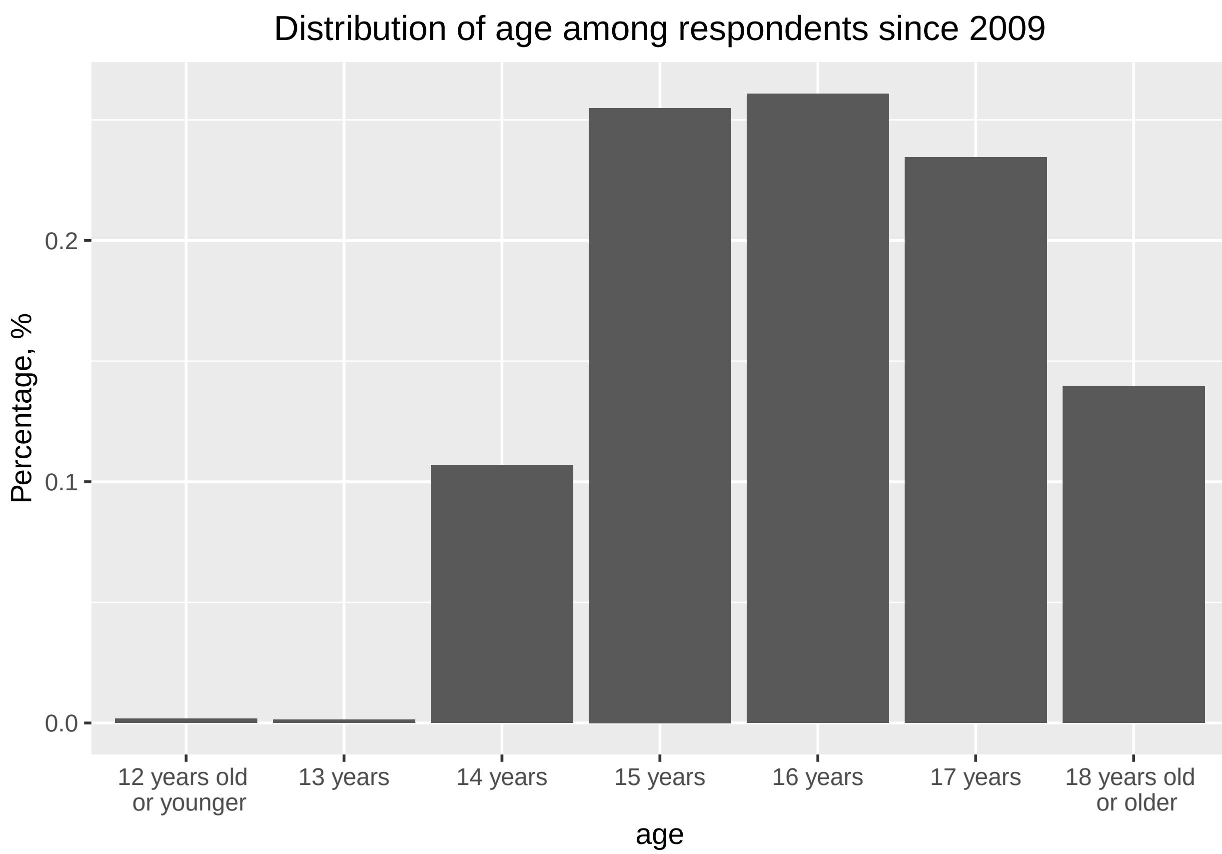


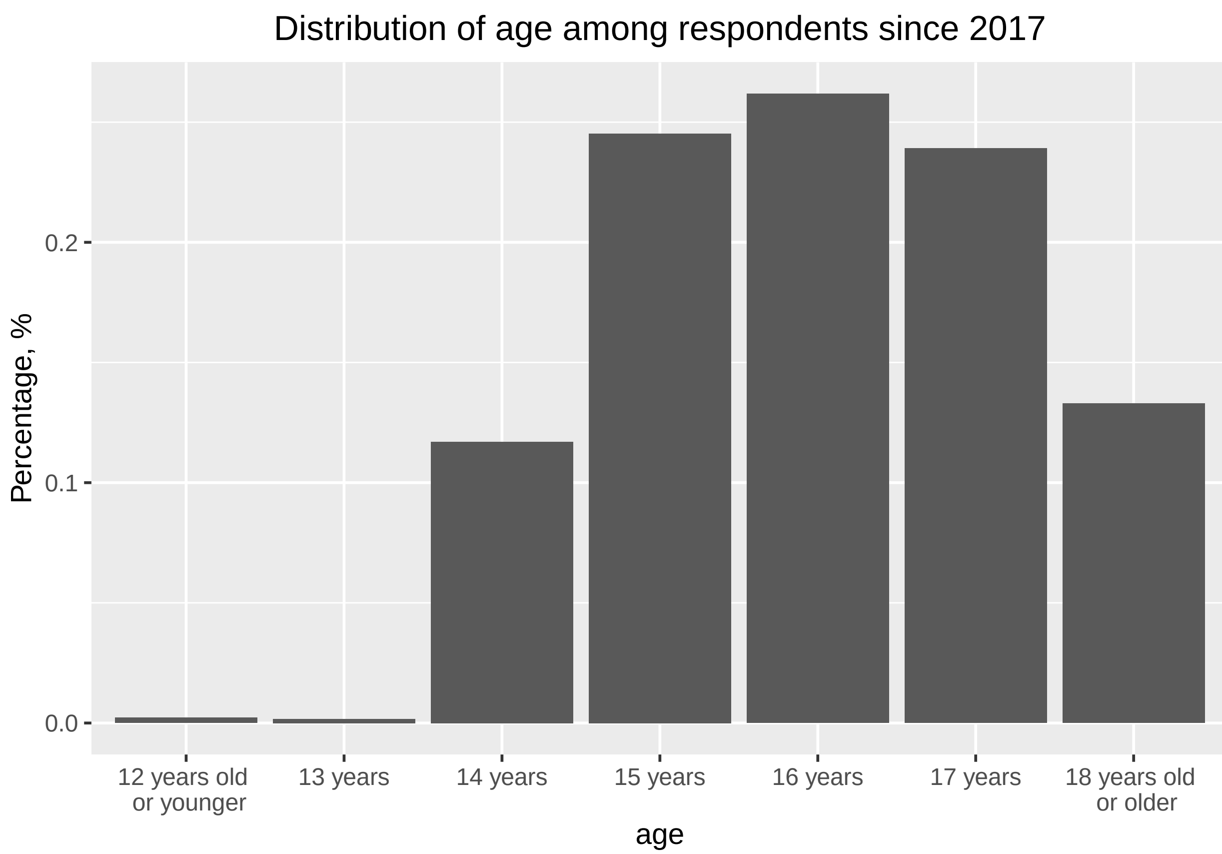

```r
# Visualizing average sleep amount over years
years_mean_sleep <- svyby(~sleep.amount, ~year, yrbs, FUN = svymean, na.rm = TRUE, keep.names = FALSE)

years_mean_sleep <- years_mean_sleep %>% filter(sleep.amount > 0) %>%
                    mutate(year = as.Date(paste(year, 1, 1, sep='-')))

ggplot(years_mean_sleep, aes(x = year, y = sleep.amount)) +
  geom_line() + geom_point(color="#69b3a2") +
  expand_limits(x = c(as.Date('2004-1-1'), as.Date('2018-1-1')), y=c(6.4,6.8)) + labs(x = 'Year', y = '
  geom_text(aes(label = round(sleep.amount,2)), vjust = -1)
```

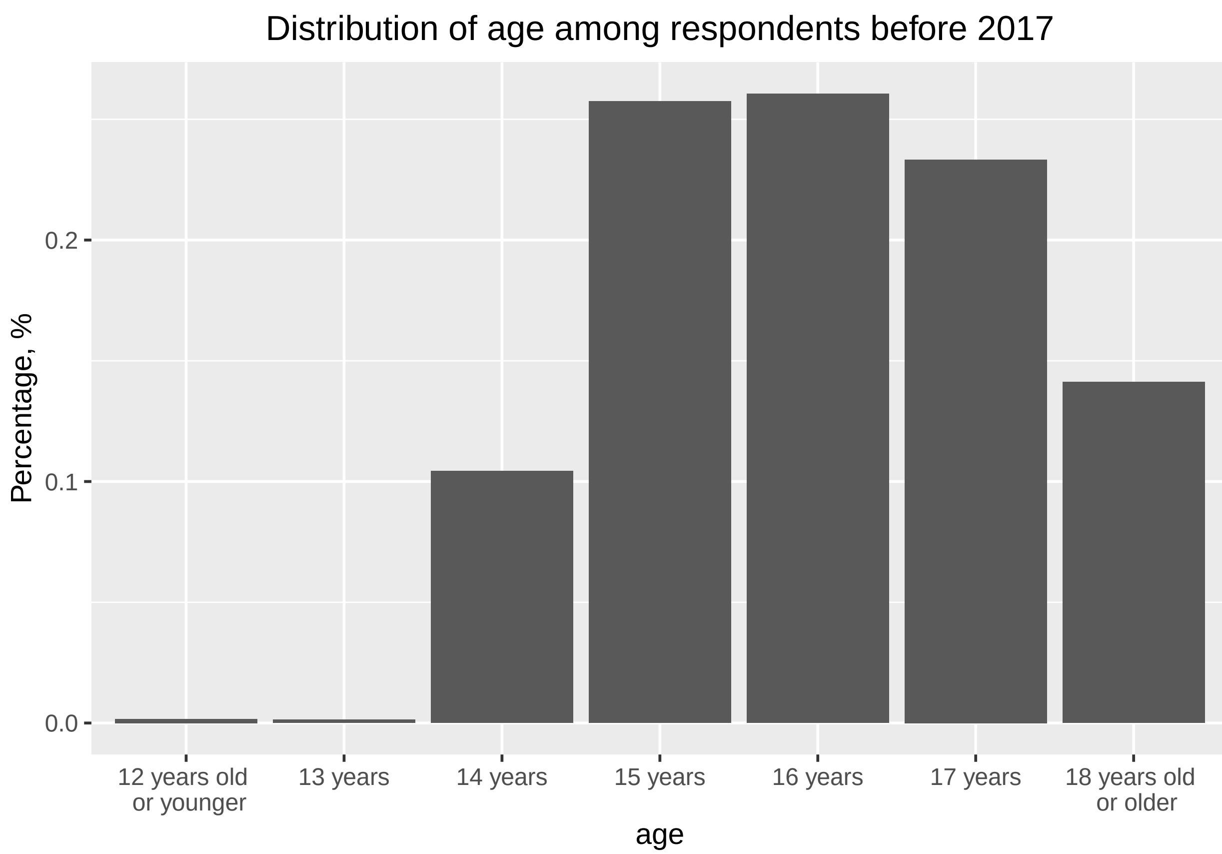


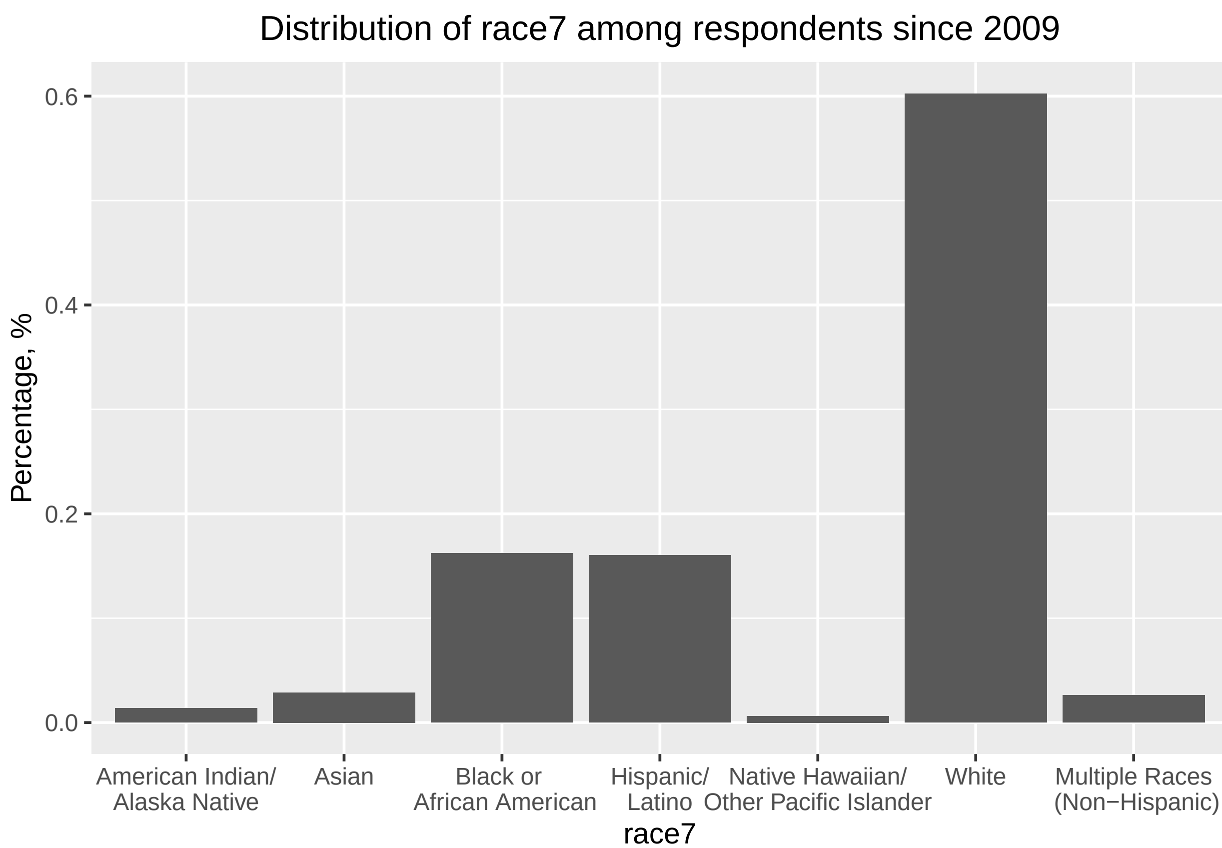

```
rm(years_mean_breakfast, years_mean_marks, years_mean_sport, years_mean_sleep)
```

```
# Visualizing difference in age distribution in different data frames
for(i in 1:length(categories)) {
  for(d in 1:length(designs)) {
      col_name <- as.name(categories[i])
      parts <- get_parts(as.svydesign2(designs[[d]]), categories[i])
      category_names <- subscripts_categories[[i]]
      levels(parts[,categories[i]]) <- category_names

      plot <- ggplot(parts, aes(x = !!col_name, y = Part)) +
        geom_col() +
        labs(title = paste('Distribution of', categories[i], 'among', names(designs)[d]), y = 'Percenta
      print(plot)
  }
}
```

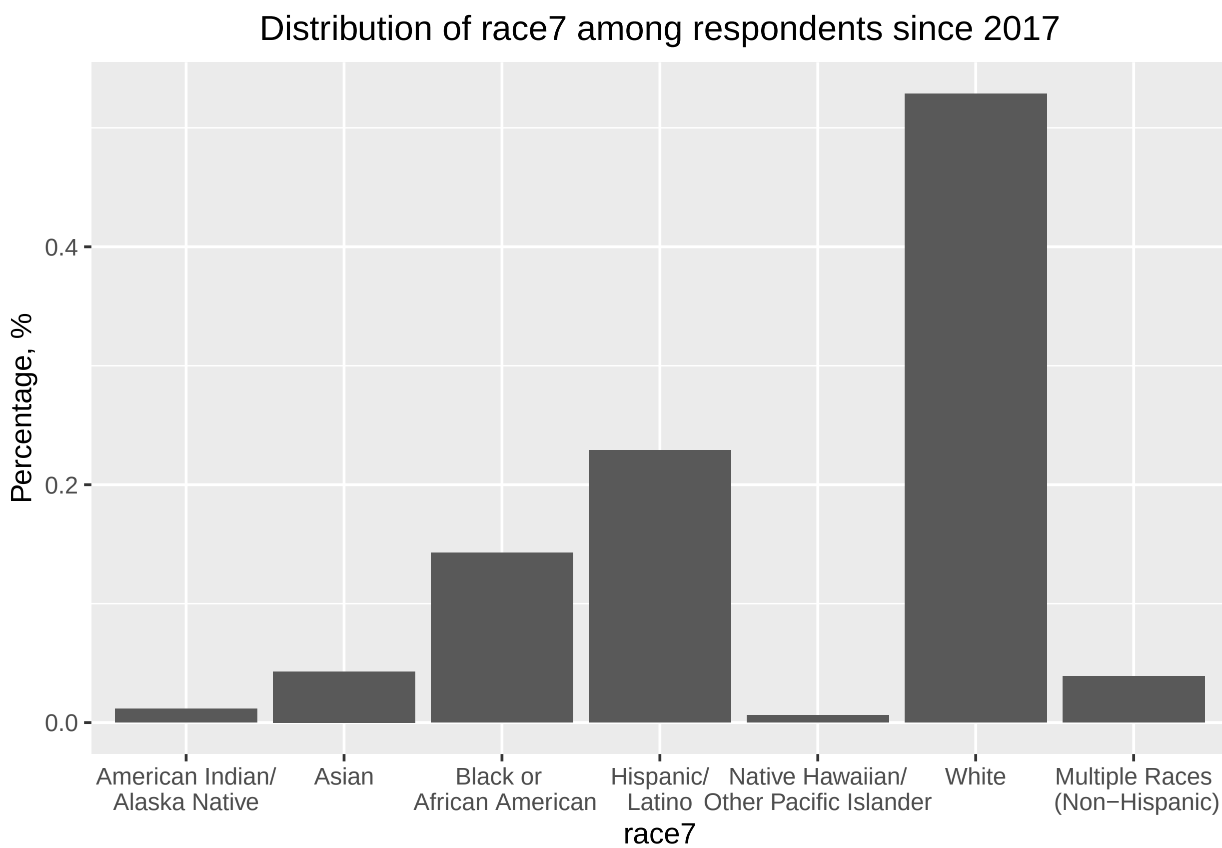

Distribution of age among respondents since 2009
Percentage, %
0.0
0.1
0.2
12 years old
or younger
13 years
14 years
15 years
16 years
17 years
18 years old
or older
age

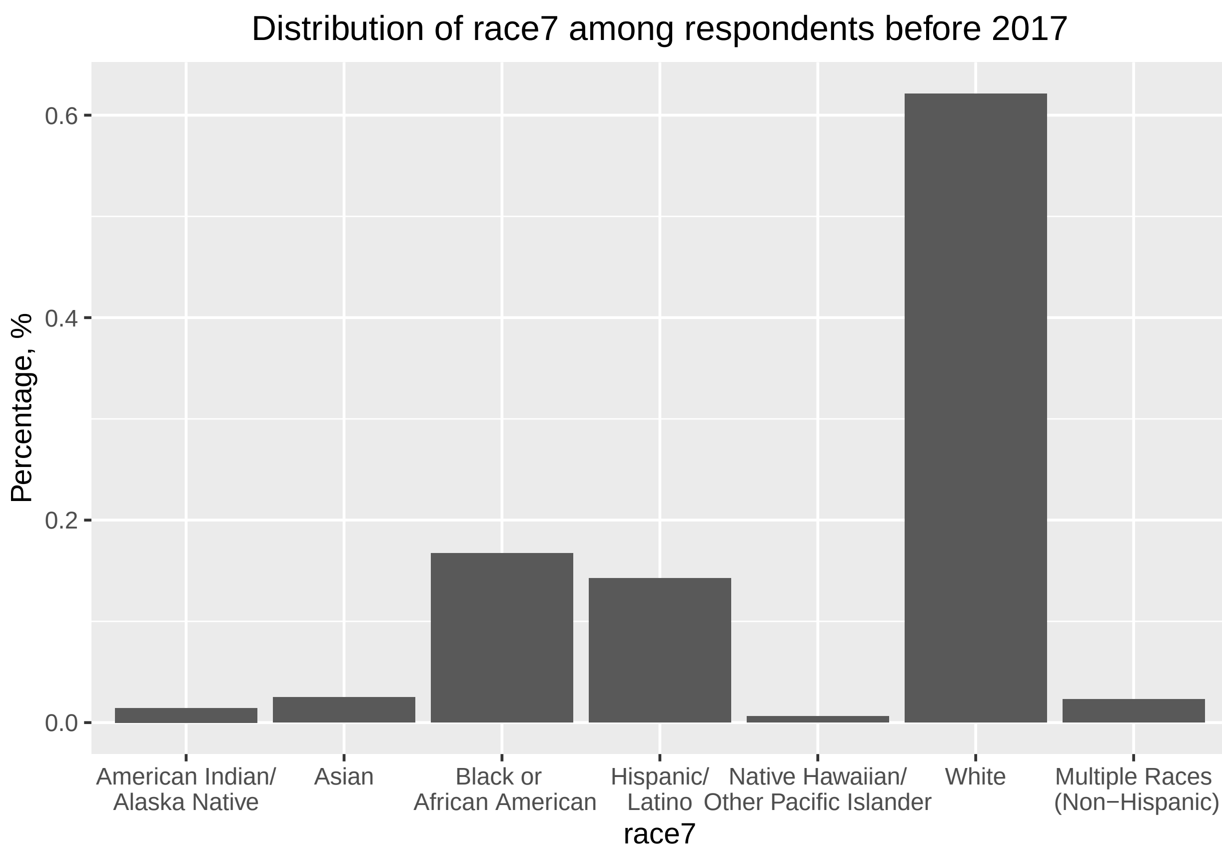

Distribution of age among respondents since 2017
Percentage, %
0.2
0.1
0.0
12 years old or younger
13 years
14 years
15 years
16 years
17 years
18 years old or older
age

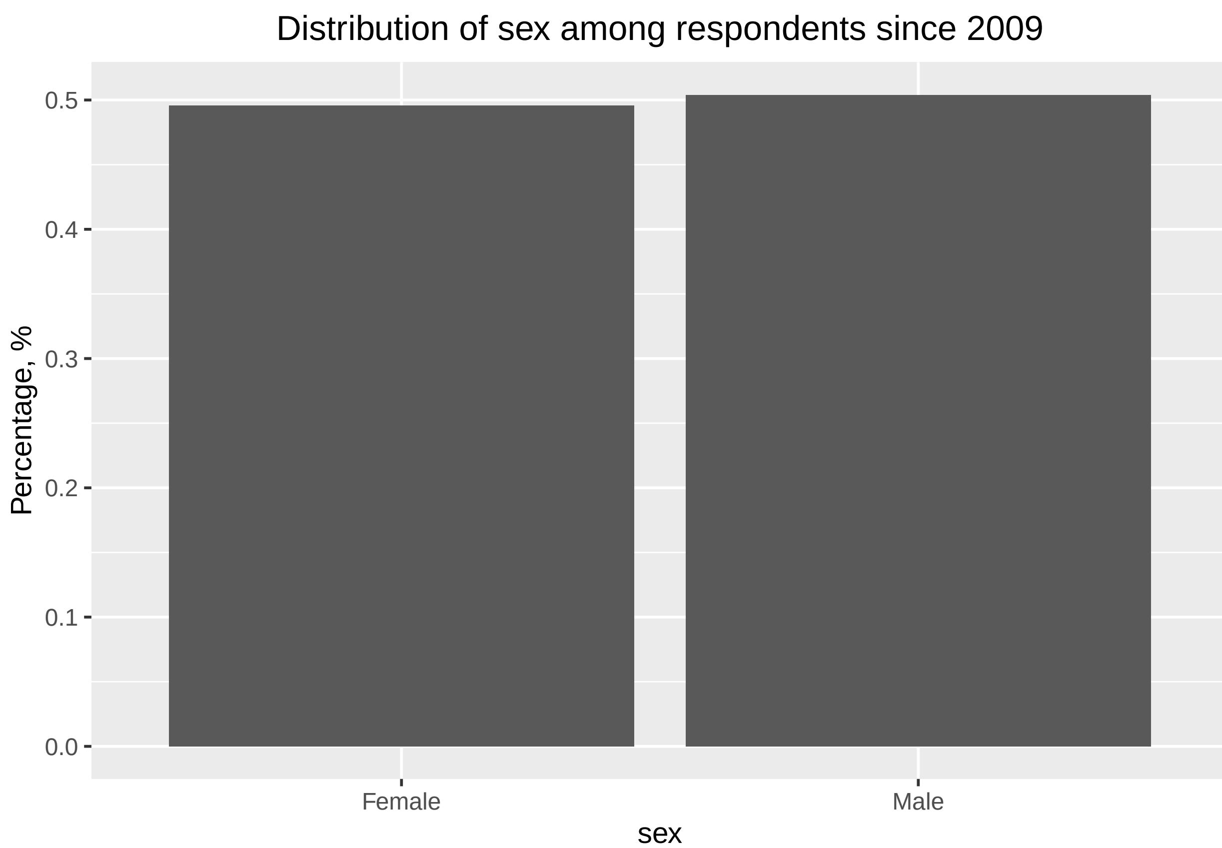

Distribution of age among respondents before 2017
Percentage, %
0.0
0.1
0.2
12 years old or younger
13 years
14 years
15 years
16 years
17 years
18 years old or older
age

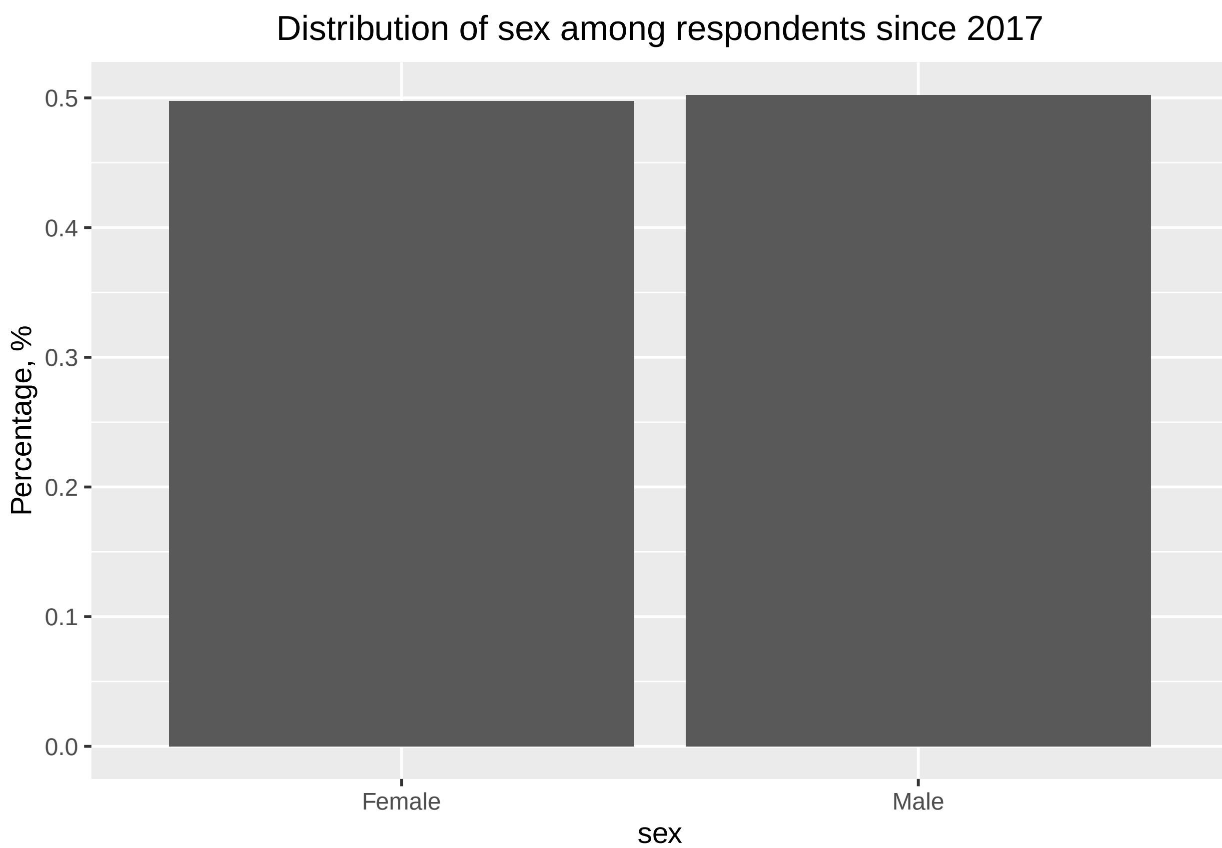


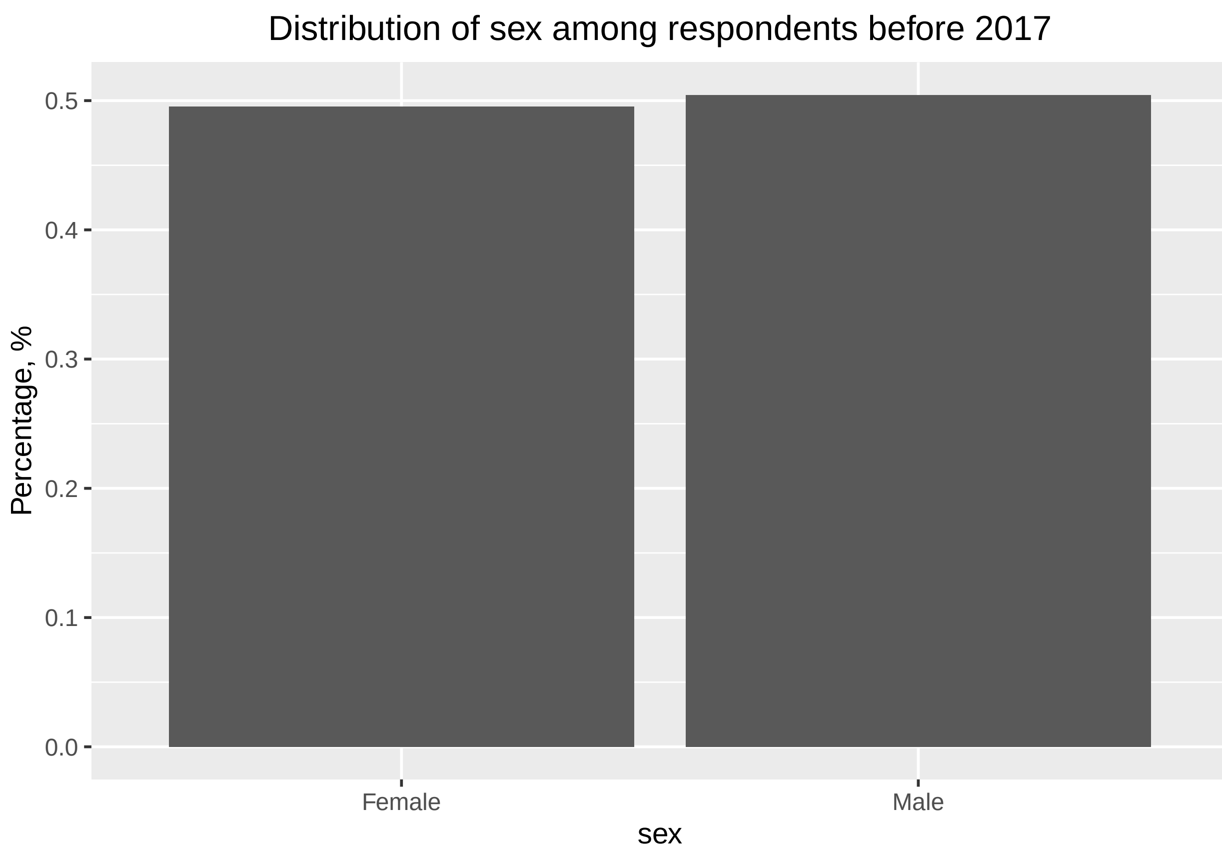
Distribution of race7 among respondents since 2009
Percentage, %
0.6
0.4
0.2
0.0
American Indian/ Alaska Native
Asian
Black or African American
Hispanic/ Latino
Native Hawaiian/ Other Pacific Islander
White
Multiple Races (Non−Hispanic)
race7

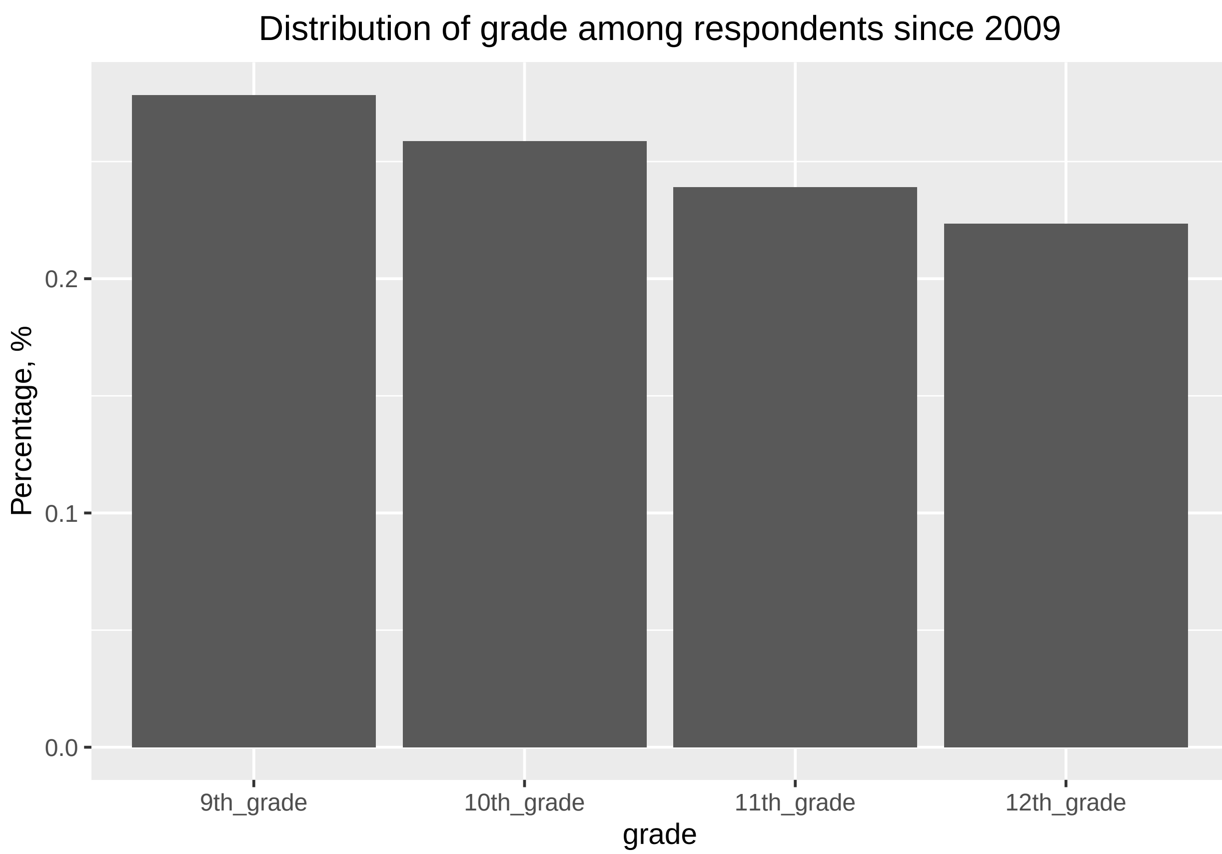

Distribution of race7 among respondents since 2017
Percentage, %
0.0
0.2
0.4
American Indian/ Alaska Native
Asian
Black or African American
Hispanic/ Latino
Native Hawaiian/ Other Pacific Islander
White
Multiple Races (Non−Hispanic)
race7

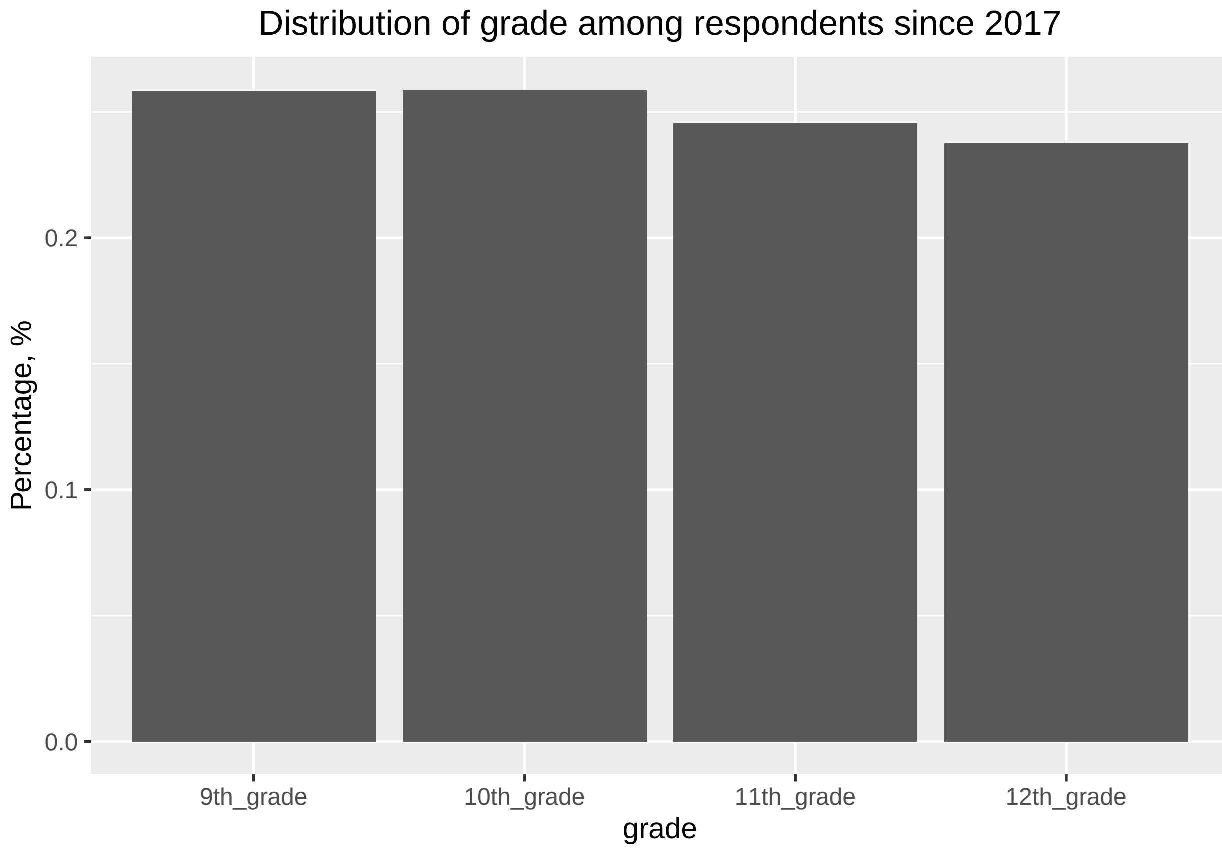

Distribution of race7 among respondents before 2017
Percentage, %
0.0
0.2
0.4
0.6
American Indian/
Alaska Native
Asian
Black or
African American
Hispanic/
Latino
Native Hawaiian/
Other Pacific Islander
White
Multiple Races
(Non−Hispanic)
race7

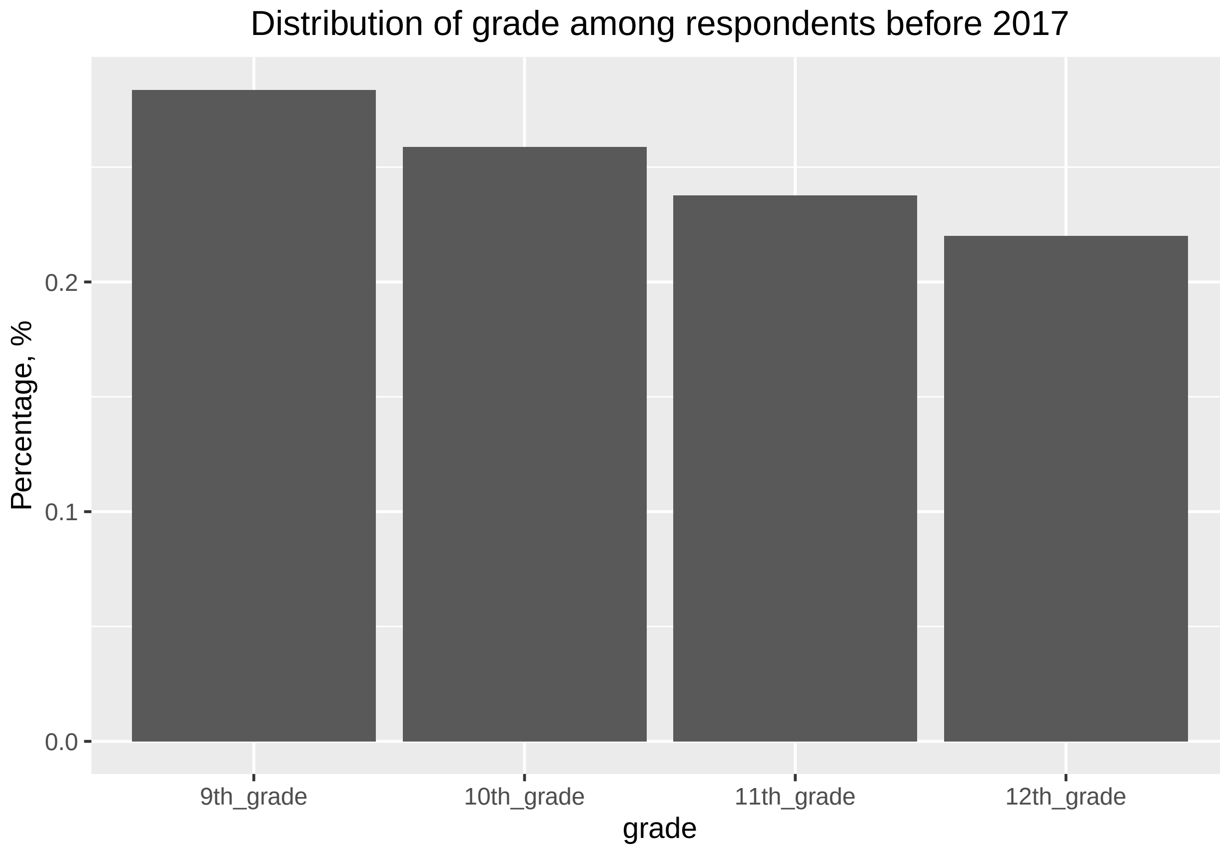

Distribution of sex among respondents since 2009
Percentage, %
0.5
0.4
0.3
0.2
0.1
0.0
Female
Male
sex

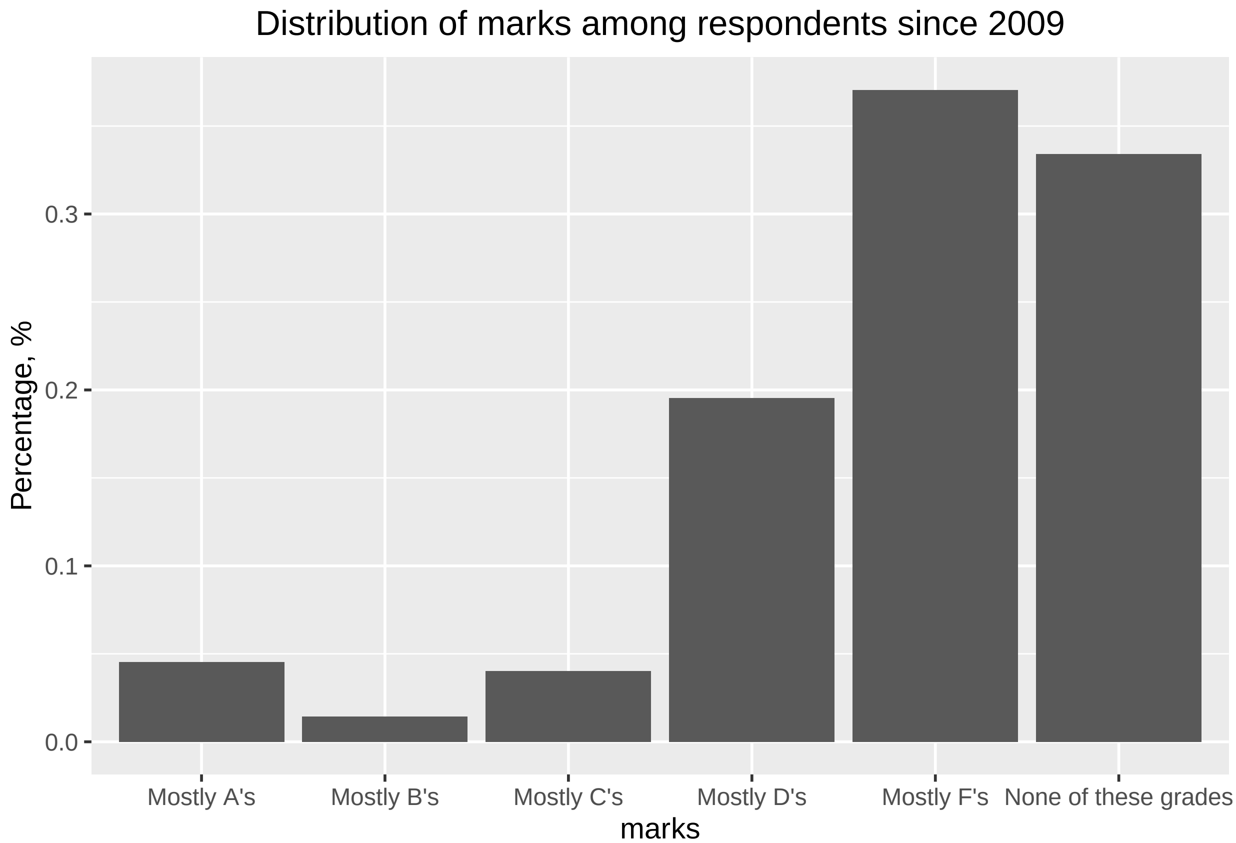

Distribution of sex among respondents since 2017
Percentage, %
0.0
0.1
0.2
0.3
0.4
0.5
Female
Male
sex

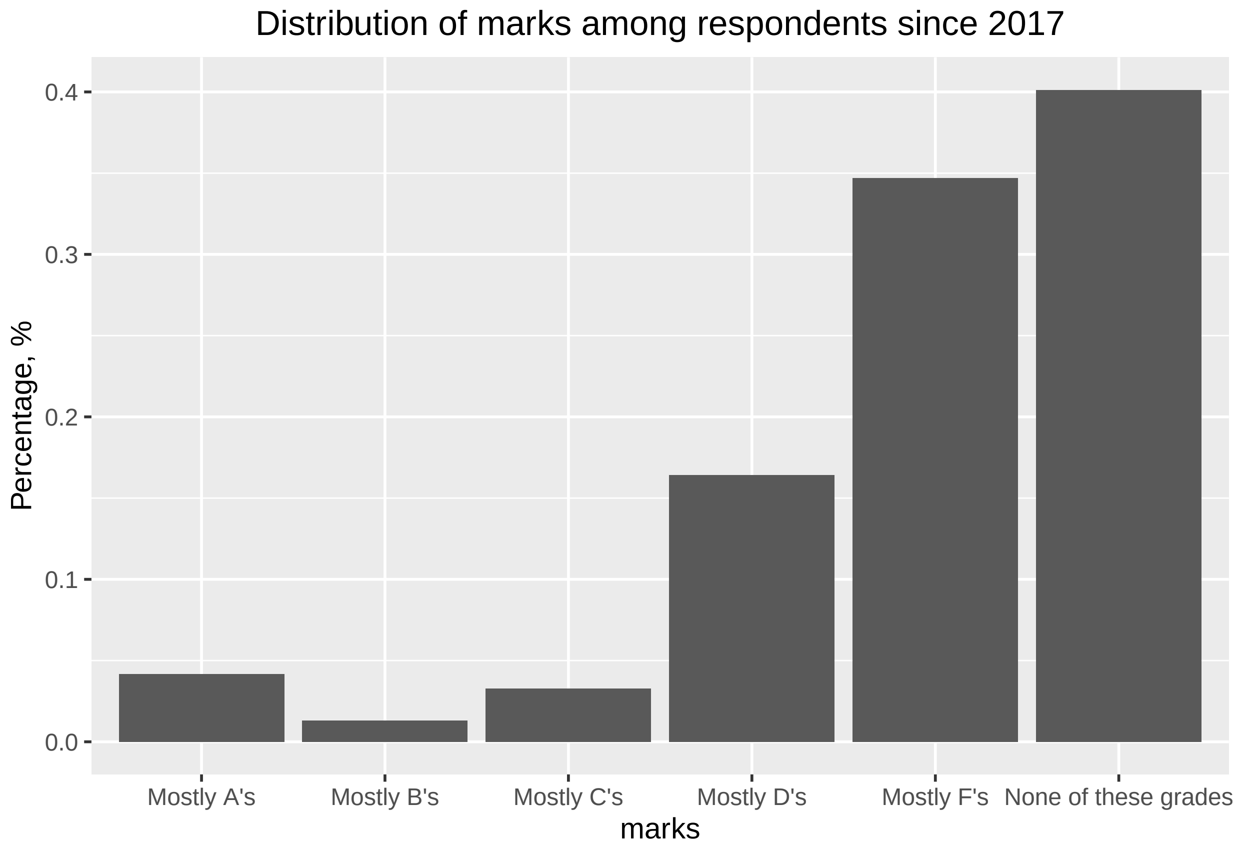

Distribution of sex among respondents before 2017
Percentage, %
0.0
0.1
0.2
0.3
0.4
0.5
Female
Male
sex

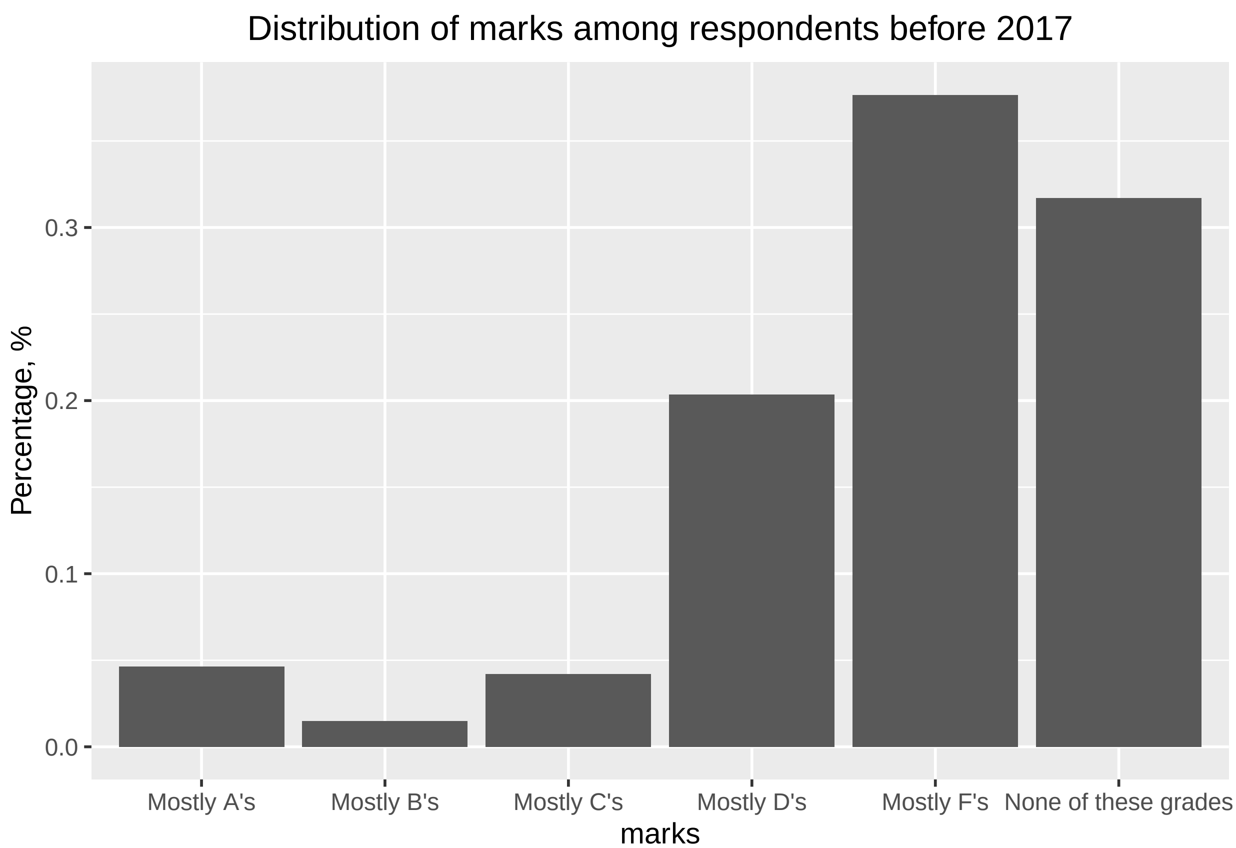

Distribution of grade among respondents since 2009
Percentage, %
0.0
0.1
0.2
9th_grade
10th_grade
11th_grade
12th_grade
grade


Distribution of grade among respondents since 2017
Percentage, %
0.2
0.1
0.0
9th_grade
10th_grade
11th_grade
12th_grade
grade


Distribution of grade among respondents before 2017
Percentage, %
0.0
0.1
0.2
9th_grade
10th_grade
11th_grade
12th_grade
grade

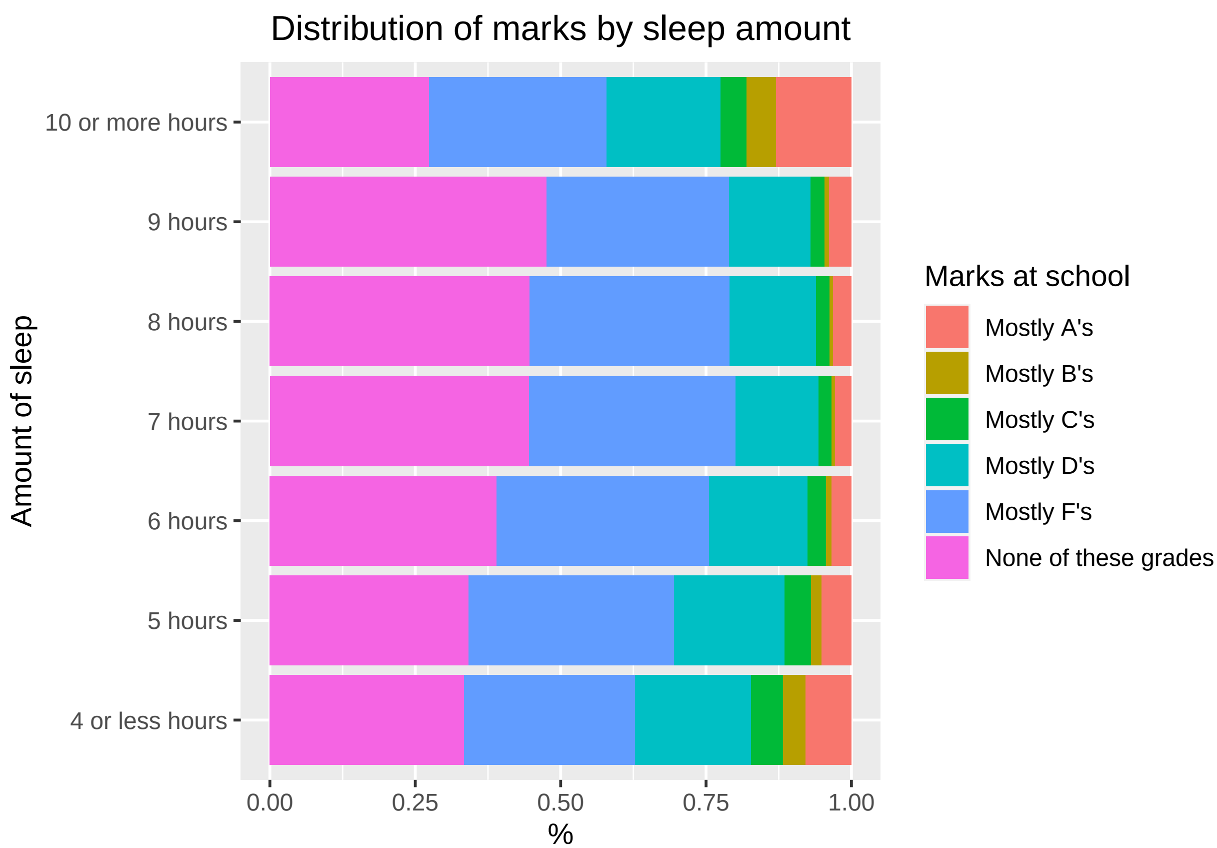

Distribution of marks among respondents since 2009
Percentage, %
0.0
0.1
0.2
0.3
Mostly A's
Mostly B's
Mostly C's
Mostly D's
Mostly F's
None of these grades
marks


Distribution of marks among respondents since 2017
0.4
0.3
0.2
0.1
0.0
Percentage, %
Mostly A's
Mostly B's
Mostly C's
Mostly D's
Mostly F's
None of these grades
marks




```r
# We can see that distributions are similar for both designs
# So, without loss of generality, we can omit usa data frame and yrbs design
rm(usa, yrbs, plot, parts, designs, category_names, subscripts_categories,
   categories, col_name)
```

```r
# Visualizing distribution of marks by sleep.amount
rel_table <- get_relation_table(yrbs_2017, 'sleep.amount', 'marks')
levels(rel_table$marks) <- marks_ranges
levels(rel_table$sleep.amount) <- sleep_ranges


ggplot(rel_table, aes(x = sleep.amount, y = Part, fill = marks)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Amount of sleep', y = '%', title = 'Distribution of marks by sleep amount') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Marks at school')
```

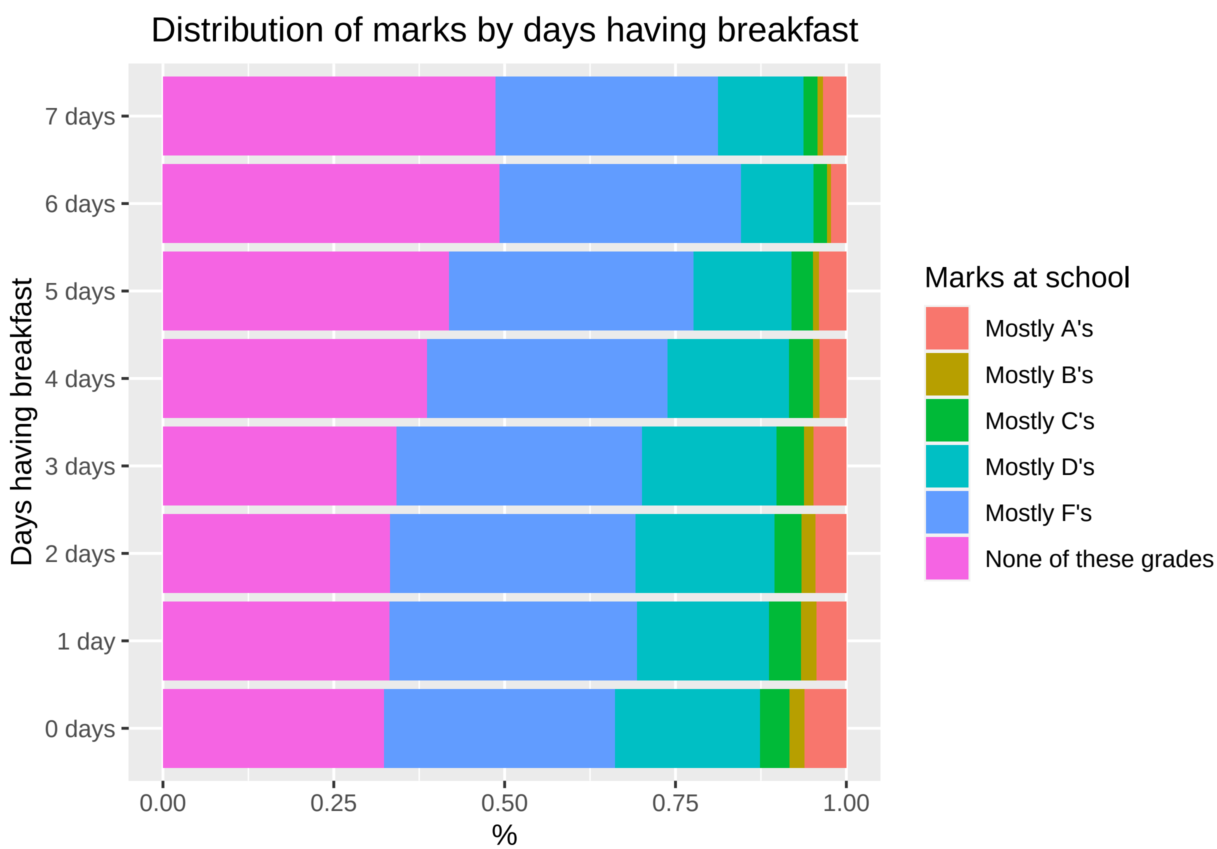



```r
# Hypothesis testing

#t-test
conduct_survey_test('There is no difference in marks by the enough amount of sleep',
                    form = as.formula('marks ~ sleep.more.8.hours'), des = yrbs_2017, t_test = TRUE)
```

```
##
## Null hypothesis: There is no difference in marks by the enough amount of sleep
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ sleep.more.8.hours
## t = 4.8334, df = 981, p-value = 1.557e-06
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.06270994 0.14825893
## sample estimates:
## difference in mean
##          0.1054844
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```r
# Visualizing distribution of marks by number of days having breakfast
rel_table <- get_relation_table(yrbs_2017, 'having_breakfast', 'marks')
```

```r
levels(rel_table$having_breakfast) <- days
levels(rel_table$marks) <- marks_ranges


ggplot(rel_table, aes(x = having_breakfast, y = Part, fill = marks)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Days having breakfast', y = '%',
             title = 'Distribution of marks by days having breakfast') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Marks at school')
```

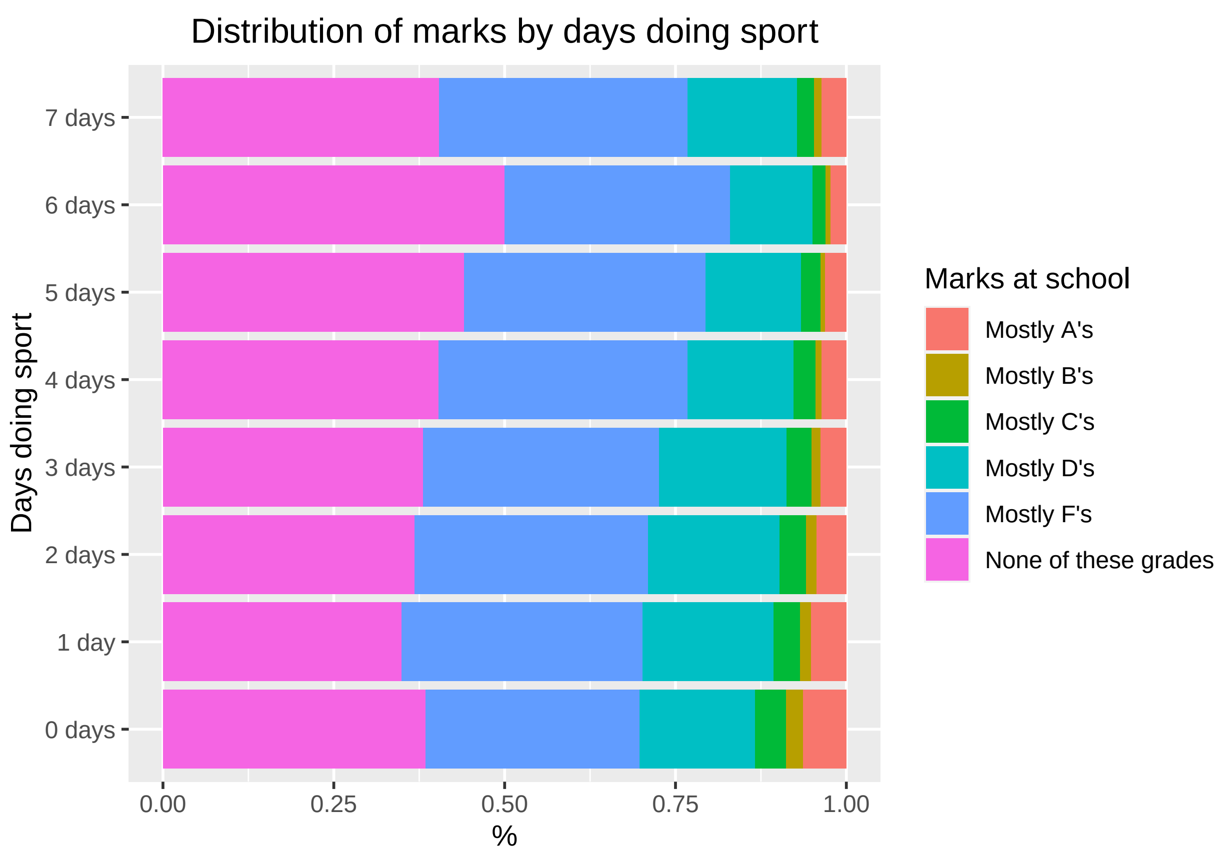



```r
# Hypothesis testing
# t-test
conduct_survey_test('There is no difference in marks by having breakfast',
                    form = as.formula('marks ~ is_having_breakfast'), des = yrbs_2017, t_test = TRUE)
```

```
##
## Null hypothesis: There is no difference in marks by having breakfast
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_having_breakfast
## t = 9.9977, df = 981, p-value < 2.2e-16
## alternative hypothesis: true difference in mean is not equal to 0
```

```
## 95 percent confidence interval:
##  0.2211528 0.3290070
## sample estimates:
## difference in mean
##          0.2750799
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```
# Visualizing distribution of marks by number of days doing sport
rel_table <- get_relation_table(yrbs_2017, 'doing_sport', 'marks')
levels(rel_table$doing_sport) <- days
levels(rel_table$marks) <- marks_ranges


ggplot(rel_table, aes(x = doing_sport, y = Part, fill = marks)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Days doing sport', y = '%',
             title = 'Distribution of marks by days doing sport') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Marks at school')
```

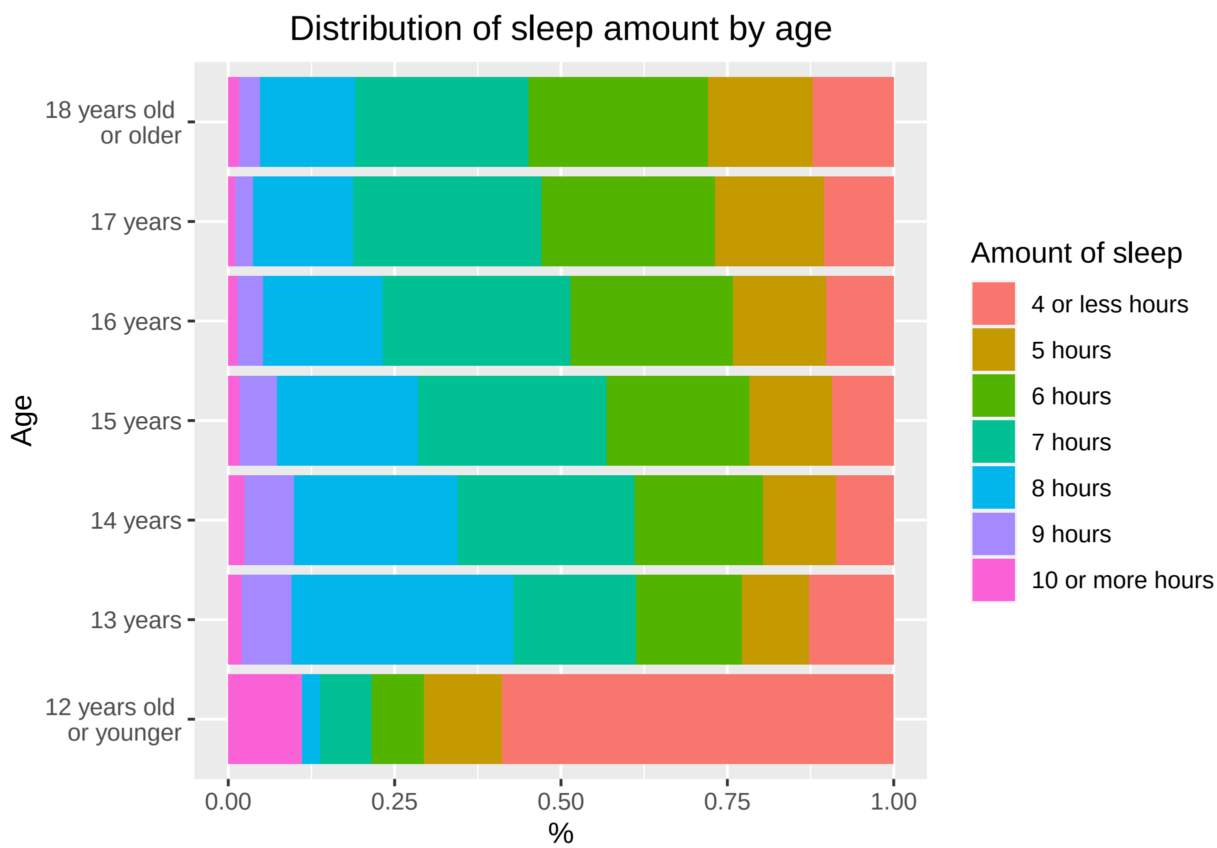



```
# Hypothesis testing
# t-test
```

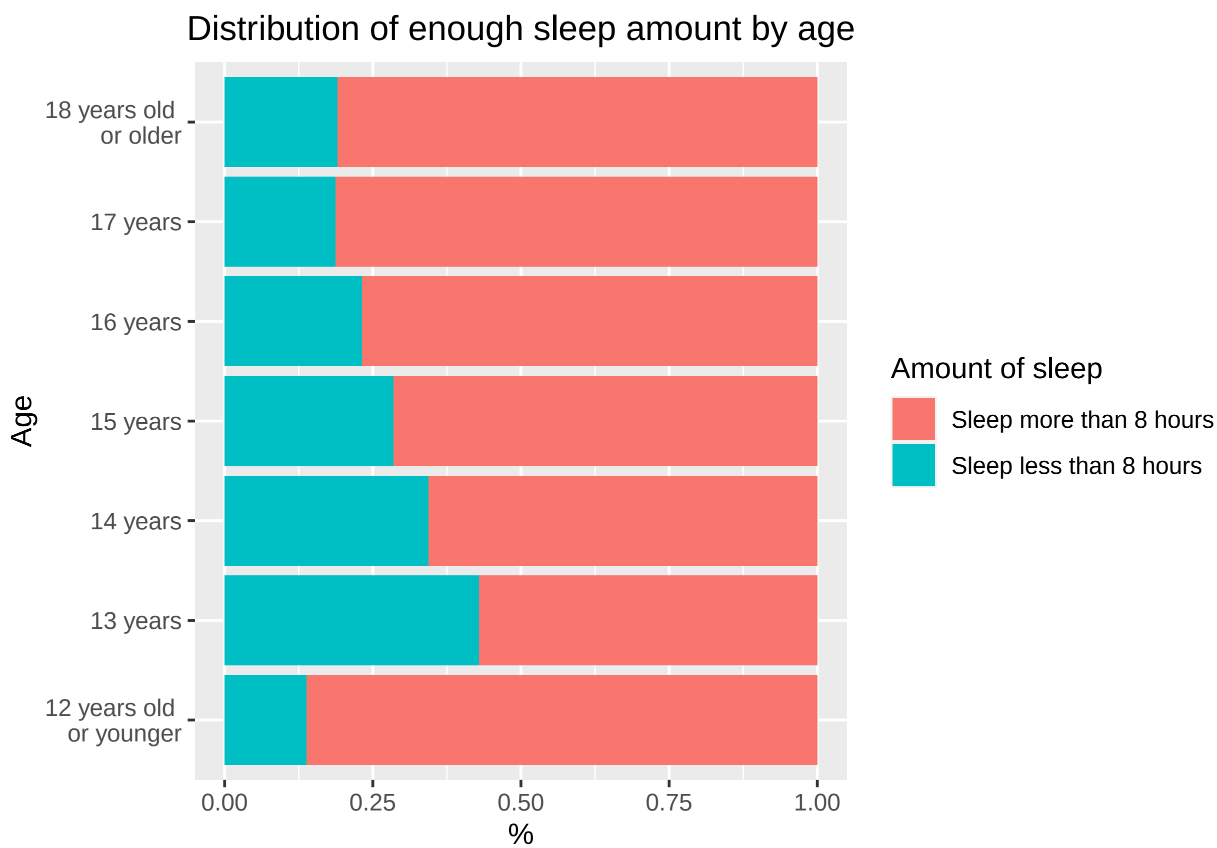
```r
conduct_survey_test('There is no difference in marks by doing sport',
                    form = as.formula('marks ~ is_doing_sport'), des = yrbs_2017, t_test = TRUE)
```

```
##
## Null hypothesis: There is no difference in marks by doing sport
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_doing_sport
## t = 9.0002, df = 981, p-value < 2.2e-16
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.1746942 0.2719623
## sample estimates:
## difference in mean
##          0.2233283
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```r
# Visualizing distribution of sleep amount by age
rel_table <- get_relation_table(yrbs_2017, 'age', 'sleep.amount')
levels(rel_table$age) <- age_ranges
levels(rel_table$sleep.amount) <- sleep_ranges


ggplot(rel_table, aes(x = age, y = Part, fill = sleep.amount)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Age', y = '%', title = 'Distribution of sleep amount by age') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```




```
# Visualizing distribution of enough sleep amount by age
rel_table <- get_relation_table(yrbs_2017, 'age', 'sleep.more.8.hours')
levels(rel_table$age) <- age_ranges
levels(rel_table$sleep.more.8.hours) <- c('Sleep more than 8 hours', 'Sleep less than 8 hours')


ggplot(rel_table, aes(x = age, y = Part, fill = sleep.more.8.hours)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Age', y = '%', title = 'Distribution of enough sleep amount by age') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```

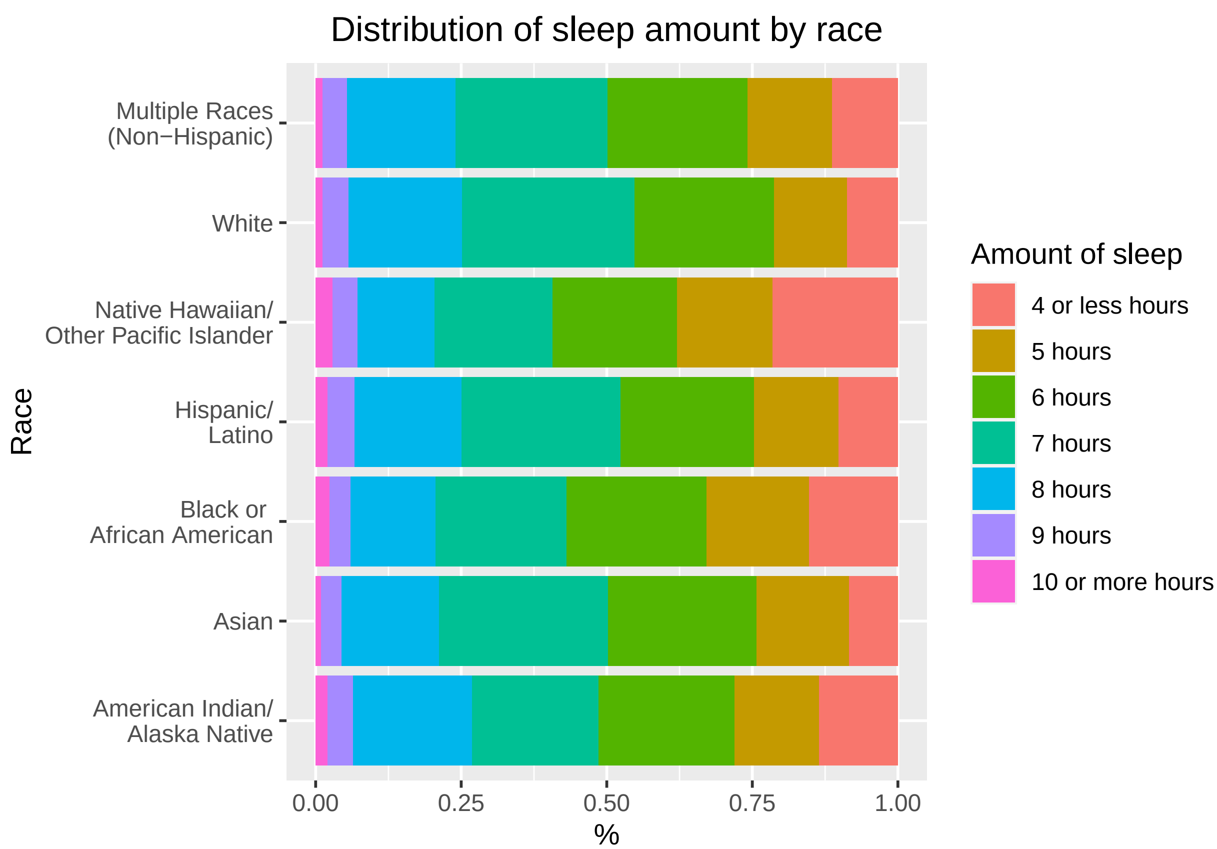



```
# Hypothesis testing
# Chi-squared test
conduct_survey_test(statement = 'Enough amount of the sleep is not depend on age', des = yrbs_2017,
                    form = as.formula('~age + sleep.more.8.hours'), chi_sq_test = TRUE)
```

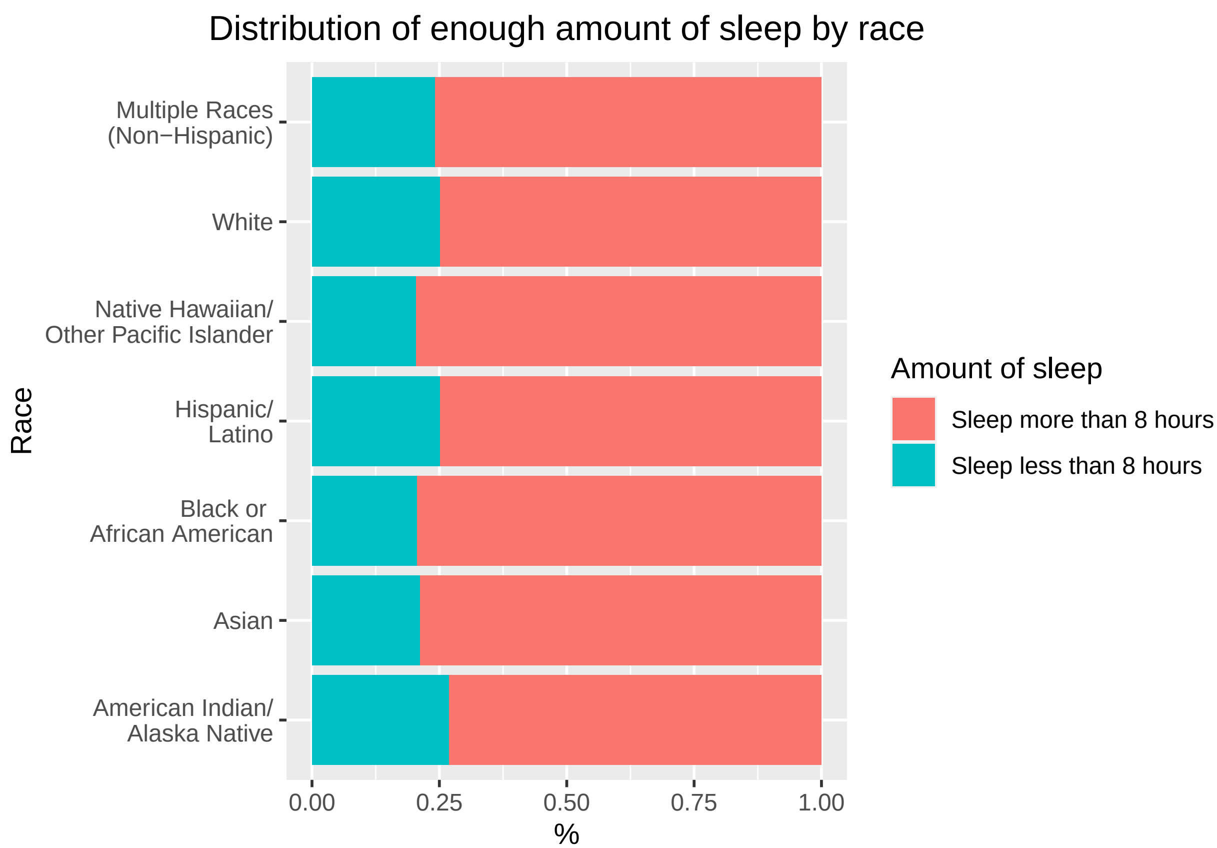
```
##
## Null hypothesis: Enough amount of the sleep is not depend on age
## Alpha: 0.1 %
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(form, design = des, statistic = "Chisq")
## X-squared = 1340.1, df = 6, p-value < 2.2e-16
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```
# Visualizing distribution of sleep amount by race
rel_table <- get_relation_table(yrbs_2017, 'race7', 'sleep.amount')
levels(rel_table$race7) <- race_7_types
levels(rel_table$sleep.amount) <- sleep_ranges


ggplot(rel_table, aes(x = race7, y = Part, fill = sleep.amount)) +
        coord_flip() +
        geom_col(position = 'fill') +
```

```r
        labs(x = 'Race', y = '%', title = 'Distribution of sleep amount by race') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```




```r
# Visualizing distribution of enough sleep amount by race
rel_table <- get_relation_table(yrbs_2017, 'race7', 'sleep.more.8.hours')
levels(rel_table$race7) <- race_7_types
levels(rel_table$sleep.more.8.hours) <- c('Sleep more than 8 hours', 'Sleep less than 8 hours')


ggplot(rel_table, aes(x = race7, y = Part, fill = sleep.more.8.hours)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Race', y = '%', title = 'Distribution of enough amount of sleep by race') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```

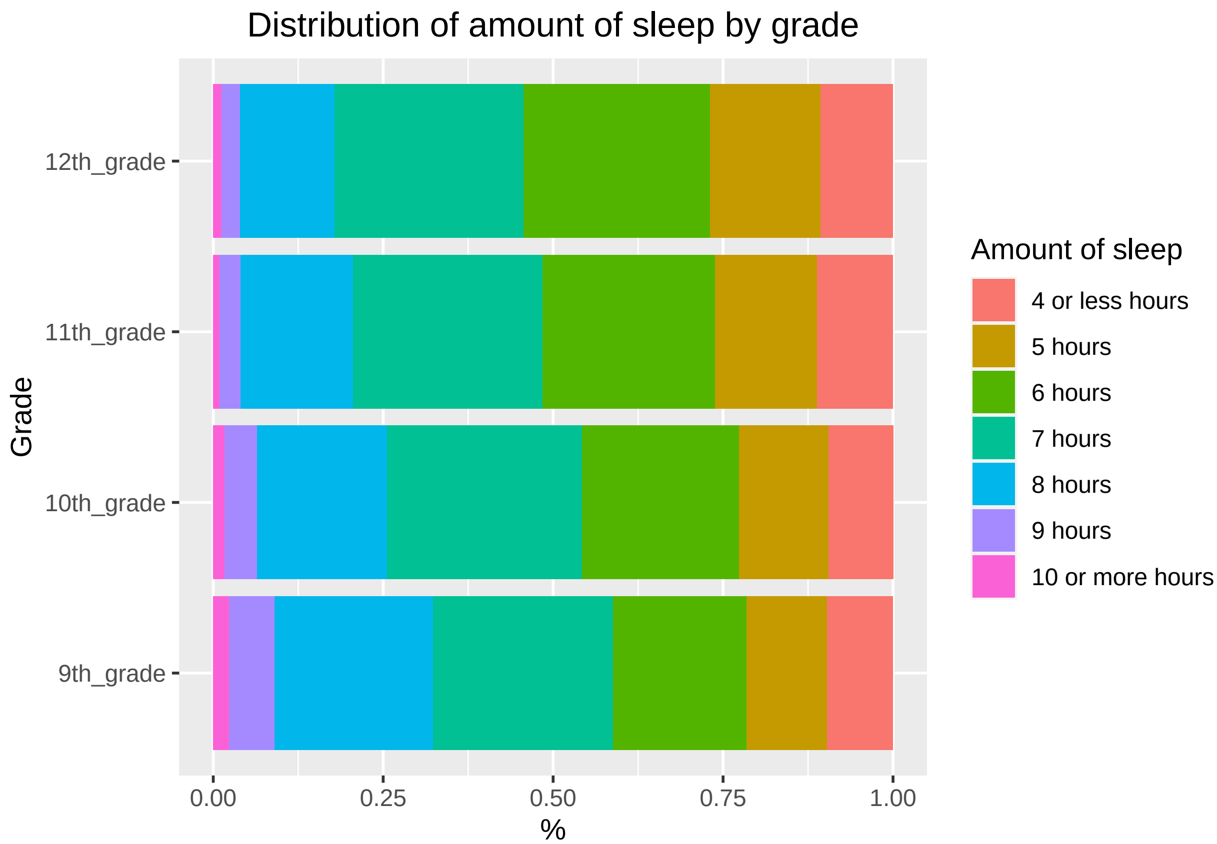



```r
# Hypothesis testing
# Chi-squared test
conduct_survey_test(statement = 'Enough amount of the sleep is not depend on race', des = yrbs_2017,
                    form = as.formula('~race7 + sleep.more.8.hours'), chi_sq_test = TRUE)
```

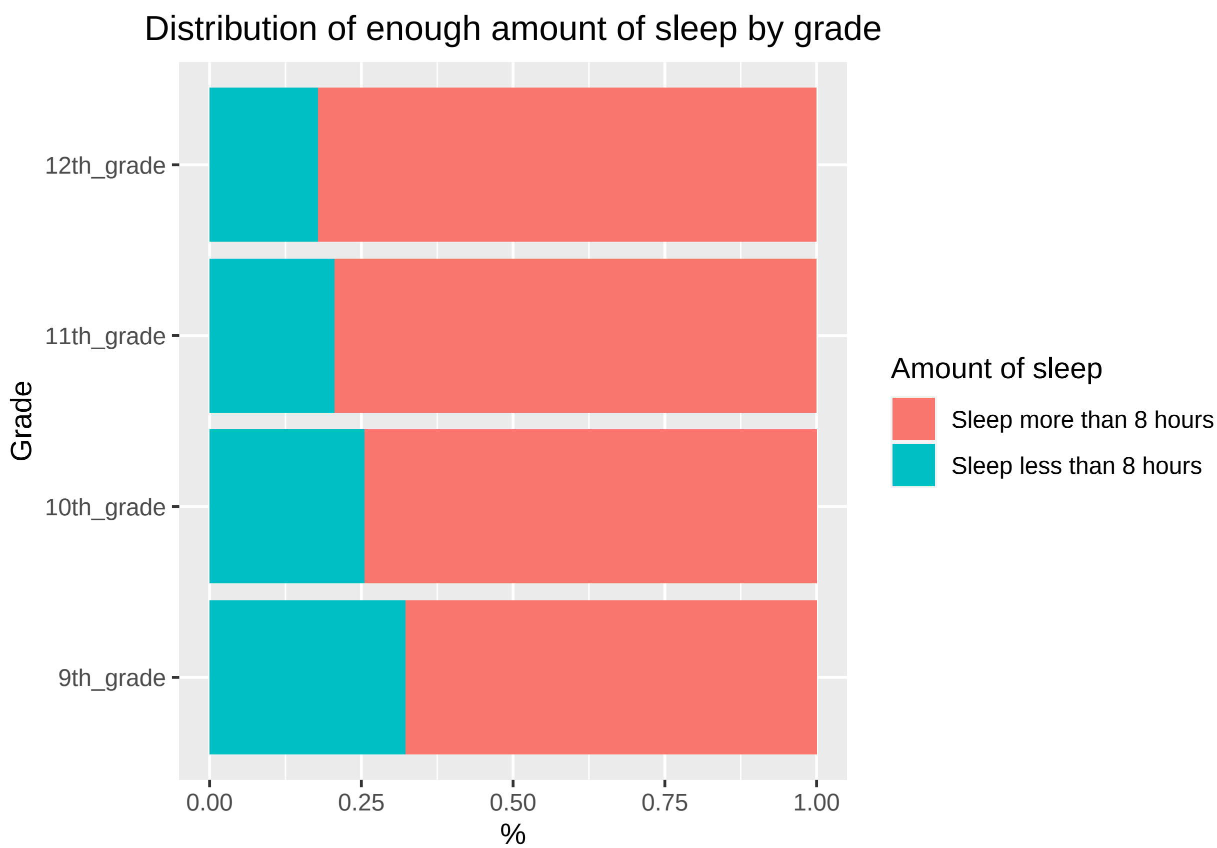
```
##
## Null hypothesis: Enough amount of the sleep is not depend on race
## Alpha: 0.1 %
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(form, design = des, statistic = "Chisq")
## X-squared = 146.38, df = 6, p-value = 0.0001638
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```r
# Visualizing distribution of sleep amount by grade
rel_table <- get_relation_table(yrbs_2017, 'grade', 'sleep.amount')
levels(rel_table$grade) <- grade_types
levels(rel_table$sleep.amount) <- sleep_ranges


ggplot(rel_table, aes(x = grade, y = Part, fill = sleep.amount)) +
        coord_flip() +
        geom_col(position = 'fill') +
```

```
        labs(x = 'Grade', y = '%', title = 'Distribution of amount of sleep by grade') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```




```
# Visualizing distribution of enough sleep amount by grade
rel_table <- get_relation_table(yrbs_2017, 'grade', 'sleep.more.8.hours')
levels(rel_table$grade) <- grade_types
levels(rel_table$sleep.more.8.hours) <- c('Sleep more than 8 hours', 'Sleep less than 8 hours')


ggplot(rel_table, aes(x = grade, y = Part, fill = sleep.more.8.hours)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Grade', y = '%', title = 'Distribution of enough amount of sleep by grade') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```

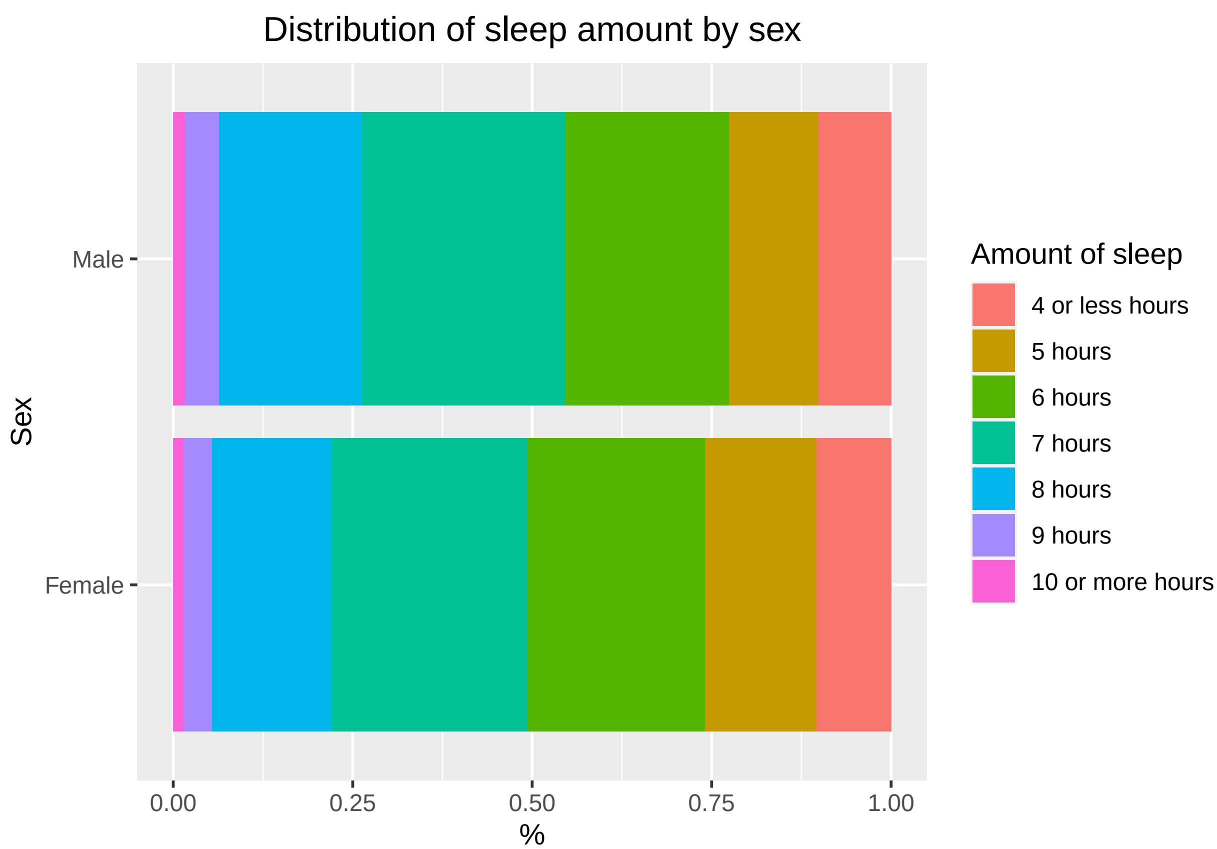



```
# Hypothesis testing
# Chi-squared test
conduct_survey_test(statement = 'Enough amount of the sleep is not depend on grade', des = yrbs_2017,
                    form = as.formula('~grade + sleep.more.8.hours'), chi_sq_test = TRUE)
```

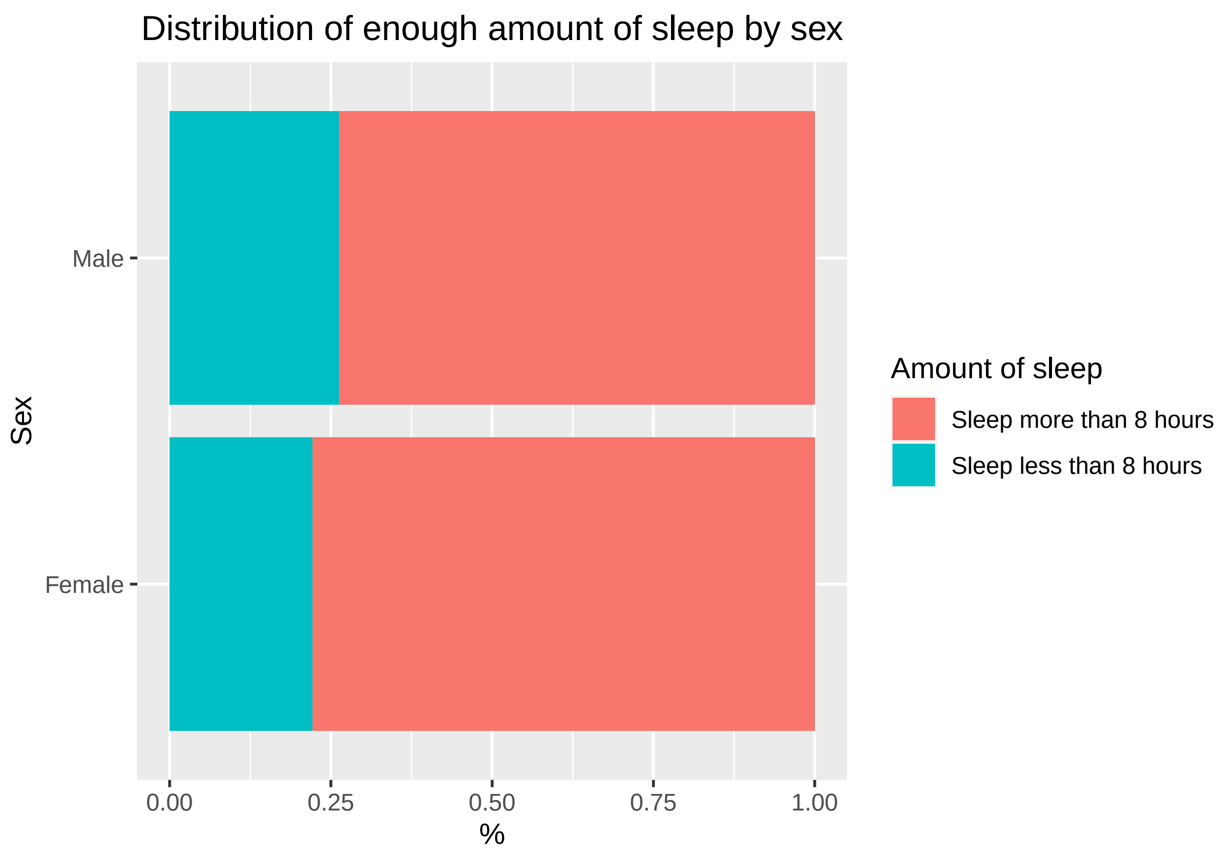
```
##
## Null hypothesis: Enough amount of the sleep is not depend on grade
## Alpha: 0.1 %
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(form, design = des, statistic = "Chisq")
## X-squared = 1410, df = 3, p-value < 2.2e-16
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

```
# Visualizing distribution of sleep amount by sex
rel_table <- get_relation_table(yrbs_2017, 'sex', 'sleep.amount')
levels(rel_table$sex) <- sex_types
levels(rel_table$sleep.amount) <- sleep_ranges


ggplot(rel_table, aes(x = sex, y = Part, fill = sleep.amount)) +
        coord_flip() +
        geom_col(position = 'fill') +
```

```
        labs(x = 'Sex', y = '%', title = 'Distribution of sleep amount by sex') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```




```
# Visualizing distribution of enough sleep amount by sex
rel_table <- get_relation_table(yrbs_2017, 'sex', 'sleep.more.8.hours')
levels(rel_table$sex) <- sex_types
levels(rel_table$sleep.more.8.hours) <- c('Sleep more than 8 hours', 'Sleep less than 8 hours')


ggplot(rel_table, aes(x = sex, y = Part, fill = sleep.more.8.hours)) +
        coord_flip() +
        geom_col(position = 'fill') +
        labs(x = 'Sex', y = '%', title = 'Distribution of enough amount of sleep by sex') +
        theme(plot.title = element_text(hjust = 0.5)) +
        scale_fill_discrete(name = 'Amount of sleep')
```




```
# Hypothesis testing
# Chi-squared test
conduct_survey_test(statement = 'Enough amount of the sleep is not depend on sex', des = yrbs_2017,
                    form = as.formula('~sex + sleep.more.8.hours'), chi_sq_test = TRUE)
```

```
##
## Null hypothesis: Enough amount of the sleep is not depend on sex
## Alpha: 0.1 %
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(form, design = des, statistic = "Chisq")
## X-squared = 198.07, df = 1, p-value = 5.302e-10
##
## Null hypothesis is wrong!
## ------------------------------------------------------
```

```
# Hypothesis testing

#t-tests
null_hypothesis <- list('There is no difference in marks by eating fruits',
                        'There is no difference in marks by eating vegetables',
                        'There is no difference in marks by smoking',
                        'There is no difference in marks by drinking alcohol',
                        'There is no difference in marks by using drugs',
                        'There is no difference in the amount of sleep by sex',
```

```
                        'White people sleep the same as much as other races')

formulas <- list(as.formula('marks ~ is_eating_fruits'),
                 as.formula('marks ~ is_eating_vegetables'),
                 as.formula('marks ~ is_smoking'),
                 as.formula('marks ~ is_drinking_alcohol'),
                 as.formula('marks ~ used_drugs'),
                 as.formula('sleep.amount ~ sex'),
                 as.formula('sleep.amount ~ is_white'))

# Alpha: 0.1%
for(i in 1:length(formulas)) {
    conduct_survey_test(statement = null_hypothesis[[i]], form = formulas[[i]],
                        des = yrbs_2017, t_test = TRUE)
  }
```

```
##
## Null hypothesis: There is no difference in marks by eating fruits
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_eating_fruits
## t = 11.391, df = 981, p-value < 2.2e-16
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.2352370 0.3330125
## sample estimates:
## difference in mean
##          0.2841247
##
## Null hypothesis is wrong!
## -------------------------------------------------------
## Null hypothesis: There is no difference in marks by eating vegetables
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_eating_vegetables
## t = 6.7629, df = 981, p-value = 2.321e-11
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.1559807 0.2832855
## sample estimates:
## difference in mean
##          0.2196331
##
## Null hypothesis is wrong!
## -------------------------------------------------------
## Null hypothesis: There is no difference in marks by smoking
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_smoking
## t = -16.33, df = 981, p-value < 2.2e-16
```

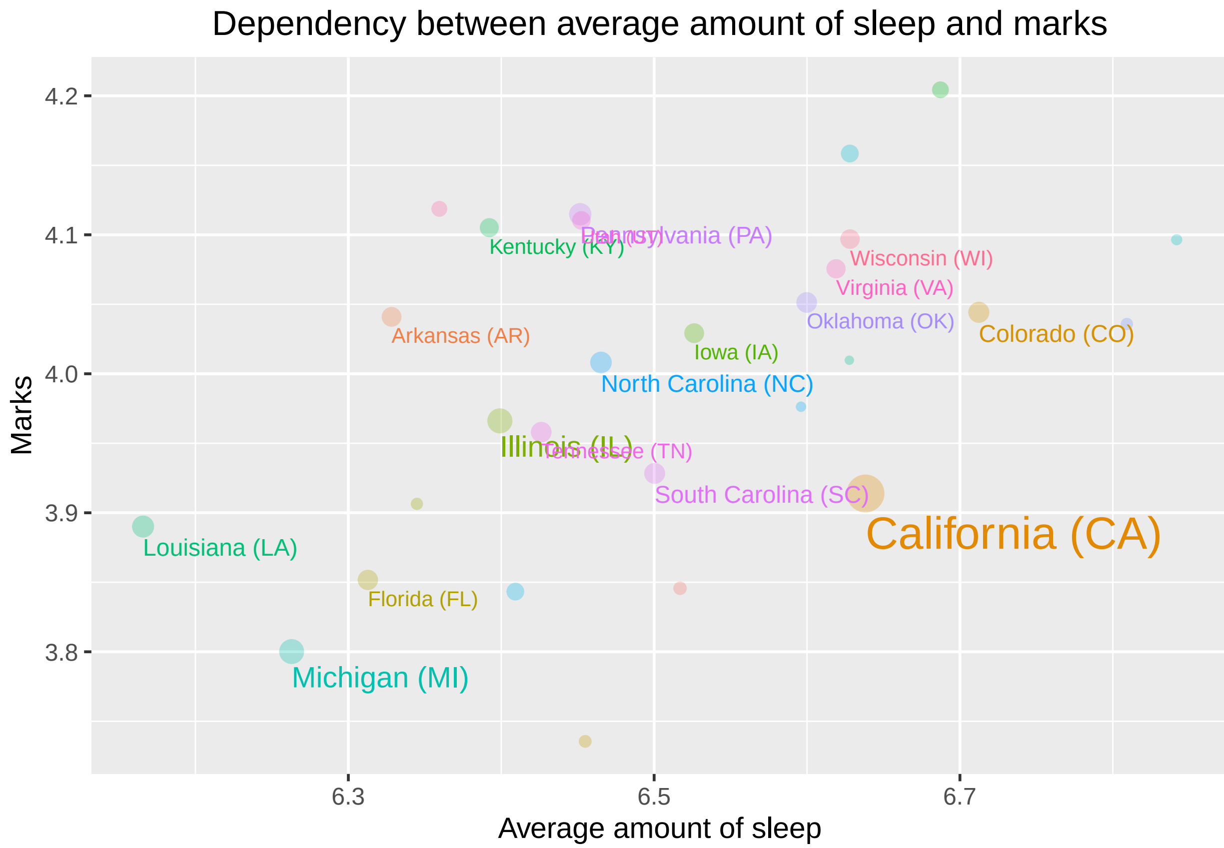
```
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -0.5892652 -0.4629712
## sample estimates:
## difference in mean
##         -0.5261182
##
## Null hypothesis is wrong!
## -----------------------------------------------------------
## Null hypothesis: There is no difference in marks by drinking alcohol
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ is_drinking_alcohol
## t = -6.785, df = 981, p-value = 2.006e-11
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -0.1892688 -0.1044287
## sample estimates:
## difference in mean
##         -0.1468487
##
## Null hypothesis is wrong!
## -----------------------------------------------------------
## Null hypothesis: There is no difference in marks by using drugs
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  marks ~ used_drugs
## t = -17.463, df = 981, p-value < 2.2e-16
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -0.4636626 -0.3700869
## sample estimates:
## difference in mean
##         -0.4168747
##
## Null hypothesis is wrong!
## -----------------------------------------------------------
## Null hypothesis: There is no difference in the amount of sleep by sex
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  sleep.amount ~ sex
## t = 7.0779, df = 981, p-value = 2.785e-12
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.1022298 0.1805294
## sample estimates:
## difference in mean
##          0.1413796
##
## Null hypothesis is wrong!
## -----------------------------------------------------------
```

```
## Null hypothesis: White people sleep the same as much as other races
## Alpha: 0.1 %
##  Design-based t-test
##
## data:  sleep.amount ~ is_white
## t = 6.0457, df = 981, p-value = 2.112e-09
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  0.1091741 0.2139190
## sample estimates:
## difference in mean
##          0.1615466
##
## Null hypothesis is wrong!
## -------------------------------------------------------
```

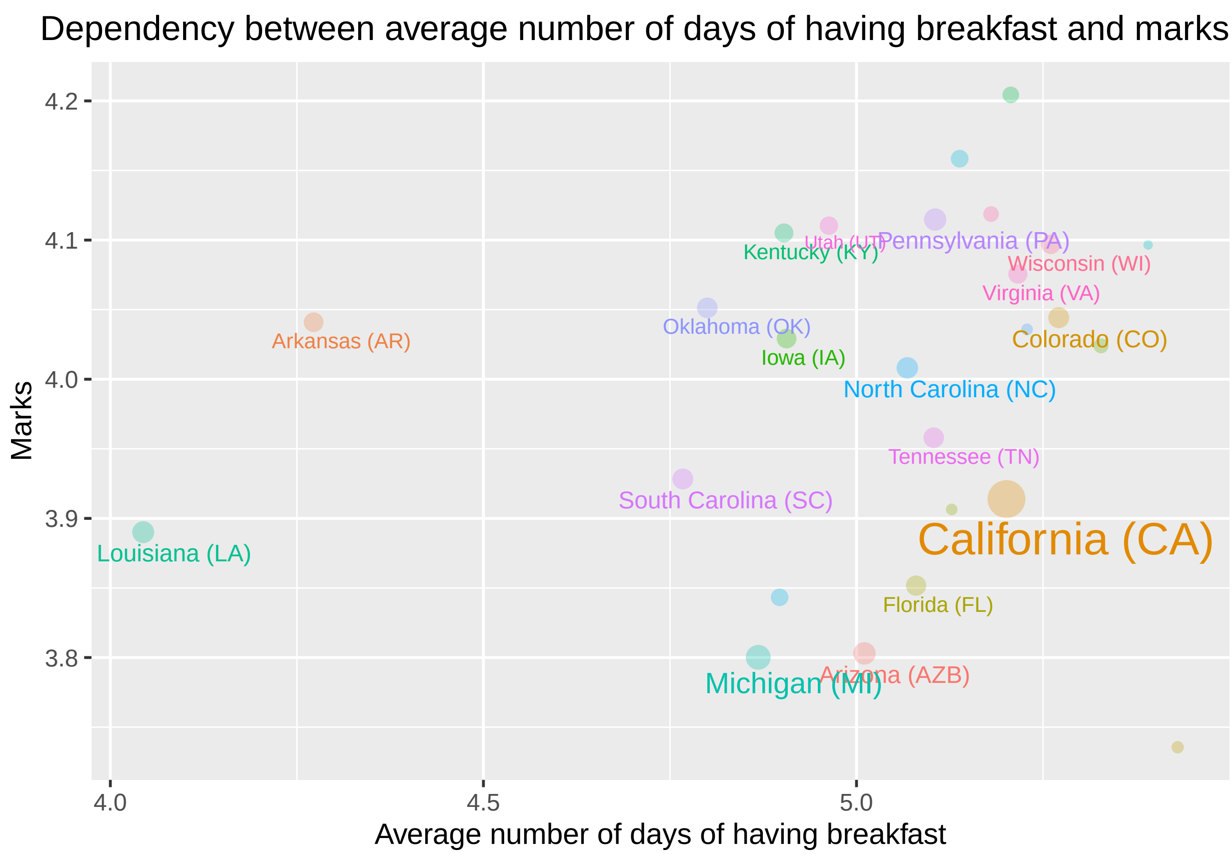
```r
rm(null_hypothesis, formulas, rel_table)
```

```r
# Set up for scatter plots
states_mean_marks <- svyby(~marks, ~sitename, yrbs_2017, FUN = svymean, na.rm = TRUE, keep.names = FALSE

states_mean_weight <- svyby(~weight, ~sitename, yrbs_2017, FUN = svymean, na.rm = TRUE, keep.names = FA

states <- inner_join(states_mean_marks, states_mean_weight, by = 'sitename')
```

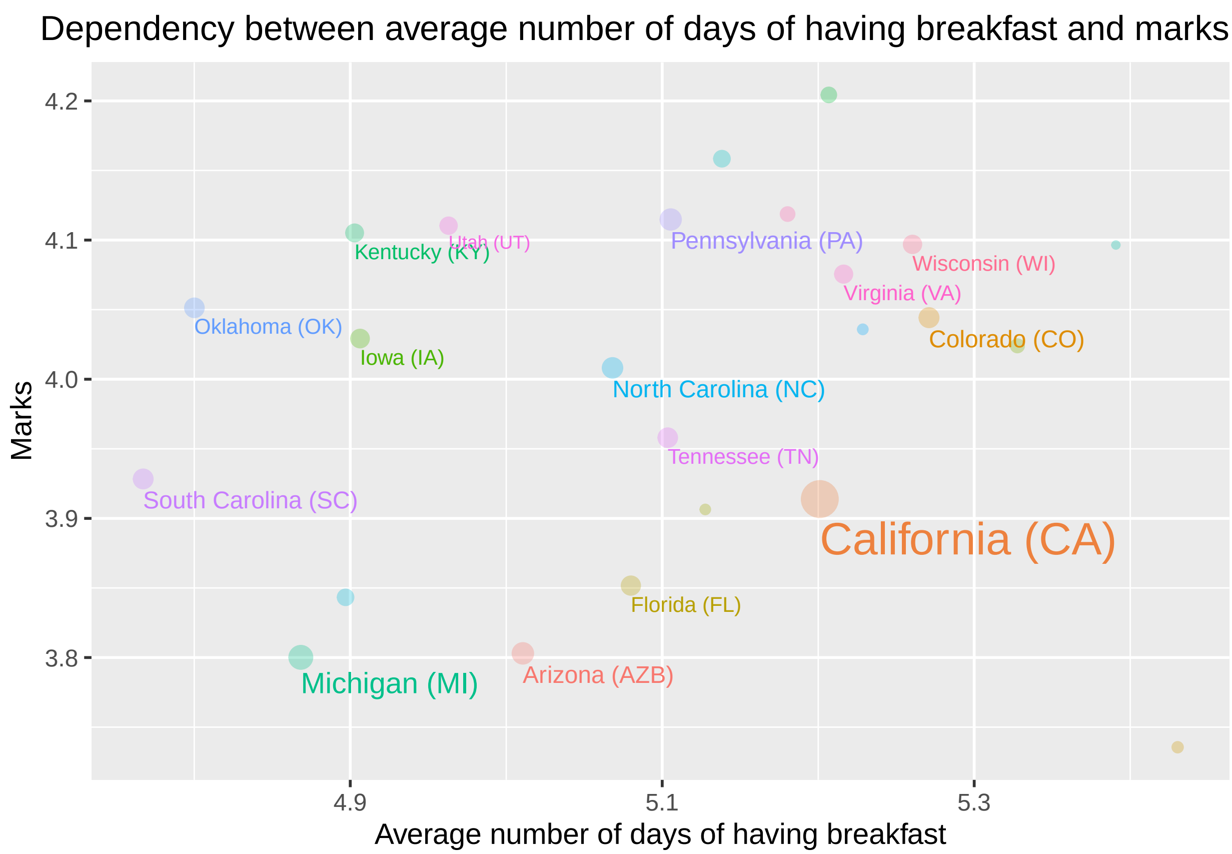
```r
# Scatter plot based on states average number of sleep hours and marks
states_mean_sleep <- svyby(~sleep.amount, ~sitename, yrbs_2017, FUN = svymean, na.rm = TRUE, keep.names

states_mean_sleep <- inner_join(states_mean_sleep, states, by = 'sitename')
states_mean_sleep <- states_mean_sleep %>% filter(sleep.amount > 0)

ggplot(states_mean_sleep, aes(x = sleep.amount, y = marks, size = weight, color = sitename)) +
  geom_point(alpha = 0.3) + guides(color = FALSE, size = FALSE) +
  labs(x = 'Average amount of sleep', y = 'Marks',
       title = 'Dependency between average amount of sleep and marks') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label = ifelse(weight > 100, as.character(sitename), '')), hjust = 0, vjust = 1.7)
```



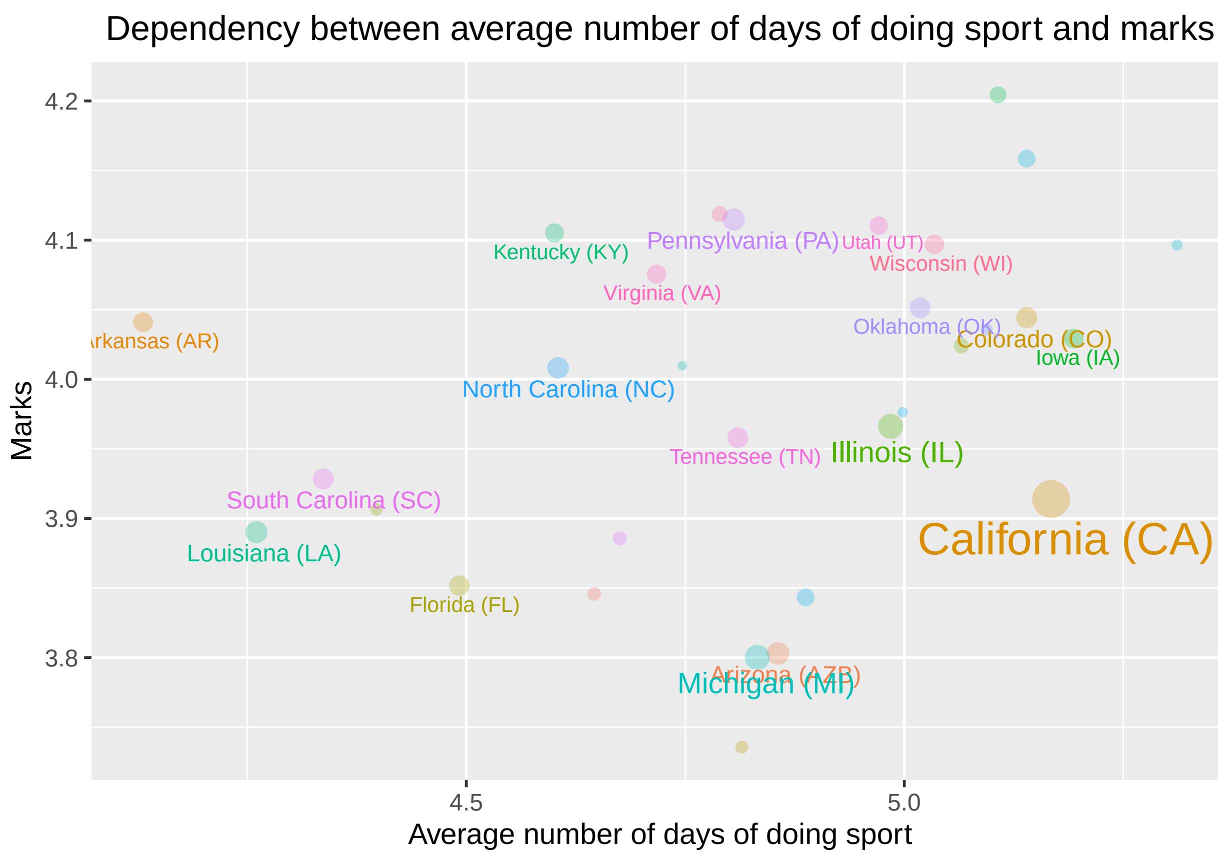

```
# Scatter plot based on states average number of days of having breakfast and marks
states_mean_breakfast <- svyby(~having_breakfast, ~sitename, yrbs_2017, FUN = svymean, na.rm = TRUE, ke

states_mean_breakfast <- inner_join(states, states_mean_breakfast, by = 'sitename')
states_mean_breakfast <- states_mean_breakfast %>% filter(having_breakfast > 0)

ggplot(states_mean_breakfast, aes(x = having_breakfast, y = marks, size = weight, color = sitename)) +
  geom_point(alpha = 0.3) + guides(color = FALSE, size = FALSE) +
  labs(x = 'Average number of days of having breakfast', y = 'Marks',
       title = 'Dependency between average number of days of having breakfast and marks') +
  theme(plot.title = element_text(hjust = 1)) +
  geom_text(aes(label = ifelse(weight > 100, as.character(sitename), '')), hjust = 0.3, vjust = 1.7)
```




```r
# Filtering outliers
states_mean_breakfast <- states_mean_breakfast %>% filter(having_breakfast > 4.5)

ggplot(states_mean_breakfast, aes(x = having_breakfast, y = marks, size = weight, color = sitename)) +
  geom_point(alpha = 0.3) + guides(color = FALSE, size = FALSE) +
  labs(x = 'Average number of days of having breakfast', y = 'Marks',
       title = 'Dependency between average number of days of having breakfast and marks') +
  theme(plot.title = element_text(hjust = 1)) +
  geom_text(aes(label = ifelse(weight > 100, as.character(sitename), '')), hjust = 0, vjust = 1.7)
```




```r
# Scatter plot based on states average number of days of doing sport and marks
states_mean_sport <- svyby(~doing_sport, ~sitename, yrbs_2017, FUN = svymean, na.rm = TRUE, keep.names =

states_mean_sport <- inner_join(states, states_mean_sport, by = 'sitename')
states_mean_sport <- states_mean_sport %>% filter(doing_sport > 0)

ggplot(states_mean_sport, aes(x = doing_sport, y = marks, size = weight, color = sitename)) +
  geom_point(alpha = 0.3) + guides(color = FALSE, size = FALSE) +
  labs(x = 'Average number of days of doing sport', y = 'Marks',
       title = 'Dependency between average number of days of doing sport and marks') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label = ifelse(weight > 100, as.character(sitename), '')), hjust = 0.45, vjust = 1.7)
```




```r
rm(states, states_mean_marks, states_mean_sleep, states_mean_weight,
        states_mean_breakfast, states_mean_sport)
```

```r
# Static data for regressions
headers <- c('Impact of different factors on the marks at school',
             'Impact of fruits and vegetables on the marks at school',
             'Impact of different vegetables on the marks at school',
             'Impact of different drugs on the marks at school',
             'Impact of the amount of sleep on the marks at school',
             'Impact of race on the marks at school',
             'Impact of school grade on the marks at school',
             'Impact of sex on the marks at school',
             'Impact of sex on the amount of sleep',
             'Impact of race on the amount of sleep',
             'Impact of sex on the amount of sleep',
             'Impact of school grade on the amount of sleep')

              # All together
formulas <- c('marks ~ sleep.more.8.hours + is_eating_vegetables + is_eating_fruits +
                is_doing_sport + is_having_breakfast + is_smoking + is_drinking_alcohol +
                used_drugs',
              # Fruits or vegetables?
              'marks ~ is_eating_vegetables + is_eating_fruits',
              # Which vegetables?
              'marks ~ is_eating_salad + is_eating_potatoes + is_eating_carrots +
                is_eating_other_vegetables',
```

```r
              # Different drugs
              'marks ~ using_marijuana + using_cocaine + using_inhalant + using_heroin +
                using_methamphetamines + using_ecstasy + using_synthetic + using_needle',
              # Sleep
              'marks ~ sleep.amount',
              # Impact of race on marks
              'marks ~ is_alaska_native + is_asian + is_black + is_latino + is_white + is_other_race',
              # Impact of grade on marks
              'marks ~ is_9th_grade + is_10th_grade + is_11th_grade + is_12th_grade + is_ungraded',
              # Impact of sex on sleep
              'sleep.amount ~ sex',
              # Impact of race on sleep
              'sleep.amount ~ is_alaska_native + is_asian + is_black +
              is_latino + is_white + is_other_race',
              # Impact of grade on sleep
              'sleep.amount ~ is_9th_grade + is_10th_grade + is_11th_grade +
              is_12th_grade + is_ungraded')
```

```r
for (f in 1:length(formulas)) {
    ols <- svyglm(as.formula(formulas[f]), yrbs_2017)
    cat(headers[f], '\n')
    print(summary(ols))
    cat('-------------------------------------------------------------\n')
}
```

```
## Impact of different factors on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.66385    0.03635 100.806  < 2e-16 ***
## sleep.more.8.hours    0.01259    0.02056   0.612 0.540556
## is_eating_vegetables  0.12521    0.03381   3.703 0.000225 ***
## is_eating_fruits      0.16833    0.02377   7.082 2.72e-12 ***
## is_doing_sport        0.13682    0.02386   5.735 1.30e-08 ***
## is_having_breakfast   0.17804    0.02609   6.823 1.56e-11 ***
## is_smoking           -0.34264    0.03891  -8.807  < 2e-16 ***
## is_drinking_alcohol   0.05783    0.02408   2.402 0.016503 *
## used_drugs           -0.36739    0.02610 -14.075  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.426185)
##
## Number of Fisher Scoring iterations: 2
##
## -------------------------------------------------------------
```

```
## Impact of fruits and vegetables on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.62255    0.02767 130.900  < 2e-16 ***
## is_eating_vegetables  0.15518    0.03235   4.796 1.87e-06 ***
## is_eating_fruits      0.24211    0.02428   9.971  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.469359)
##
## Number of Fisher Scoring iterations: 2
##
## -------------------------------------------------------------
## Impact of different vegetables on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.724040   0.016815 221.471  < 2e-16 ***
## is_eating_salad            0.121077   0.020948   5.780 1.00e-08 ***
## is_eating_potatoes         0.004728   0.017236   0.274   0.7839
## is_eating_carrots          0.059321   0.028524   2.080   0.0378 *
## is_eating_other_vegetables 0.199312   0.024625   8.094 1.71e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.462658)
##
## Number of Fisher Scoring iterations: 2
##
## -------------------------------------------------------------
## Impact of different drugs on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.0996928  0.0188955 216.966  < 2e-16 ***
## using_marijuana        -0.3367023  0.0283929 -11.859  < 2e-16 ***
## using_cocaine          -0.3657897  0.0551609  -6.631 5.51e-11 ***
## using_inhalant         -0.1669351  0.0506700  -3.295  0.00102 **
## using_heroin           -0.1592089  0.1113786  -1.429  0.15320
## using_methamphetamines  0.0769603  0.1095989   0.702  0.48272
## using_ecstasy          -0.1776191  0.0831840  -2.135  0.03299 *
## using_synthetic        -0.1961109  0.0701866  -2.794  0.00531 **
## using_needle            0.0007691  0.1178776   0.007  0.99480
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.43502)
##
## Number of Fisher Scoring iterations: 2
##
## ------------------------------------------------------------
## Impact of the amount of sleep on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.414426   0.048074   71.02   <2e-16 ***
## sleep.amount 0.086497   0.007634   11.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.487468)
##
## Number of Fisher Scoring iterations: 2
##
## ------------------------------------------------------------
## Impact of race on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.677836   0.104197  35.297  < 2e-16 ***
## is_alaska_native  0.011368   0.121529   0.094  0.92549
```

```
## is_asian          0.693231   0.116122   5.970 3.32e-09 ***
## is_black         -0.001792   0.105127  -0.017  0.98640
## is_latino        -0.003876   0.100281  -0.039  0.96918
## is_white          0.463965   0.101477   4.572 5.45e-06 ***
## is_other_race     0.340925   0.107406   3.174  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.425145)
##
## Number of Fisher Scoring iterations: 2
##
## -------------------------------------------------------------
## Impact of school grade on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.05748    0.03227 125.736  < 2e-16 ***
## is_9th_grade  -0.13868    0.04431  -3.129  0.00180 **
## is_10th_grade -0.11749    0.04506  -2.607  0.00927 **
## is_11th_grade -0.10635    0.03225  -3.298  0.00101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.478761)
##
## Number of Fisher Scoring iterations: 2
##
## -------------------------------------------------------------
## Impact of sex on the marks at school
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.41926    0.01772 362.353  < 2e-16 ***
## sex          0.14138    0.01997   7.078 2.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.002079)
##
```

```
## Number of Fisher Scoring iterations: 2
##
## ----------------------------------------------------------------
## Impact of sex on the amount of sleep
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.1152     0.1534  39.875  < 2e-16 ***
## is_alaska_native   0.3058     0.1772   1.726  0.08474 .
## is_asian           0.3244     0.1591   2.039  0.04171 *
## is_black           0.1223     0.1496   0.818  0.41377
## is_latino          0.3954     0.1604   2.465  0.01388 *
## is_white           0.4509     0.1517   2.973  0.00303 **
## is_other_race      0.3198     0.1512   2.115  0.03470 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.993128)
##
## Number of Fisher Scoring iterations: 2
##
## ----------------------------------------------------------------
## Impact of race on the amount of sleep
##
## Call:
## svyglm(formula = as.formula(formulas[f]), design = yrbs_2017)
##
## Survey design:
## svydesign(id = ~PSU, weights = ~weight, strata = ~stratum, data = usa_2017,
##     nest = TRUE)
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.30854    0.02759 228.666  < 2e-16 ***
## is_9th_grade   0.40264    0.03469  11.605  < 2e-16 ***
## is_10th_grade  0.24788    0.03264   7.594 7.21e-14 ***
## is_11th_grade  0.05683    0.03561   1.596    0.111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.981187)
##
## Number of Fisher Scoring iterations: 2
##
## ----------------------------------------------------------------
```