hw02

По адресу https://raw.githubusercontent.com/SergeyMirvoda/da2016/master/data/calif_penn_2011.csv (https://raw.githubusercontent.com/SergeyMirvoda/da2016/master/data/calif_penn_2011.csv) можно получить набор данных, содержащий информацию о домовладениях в Калифорнии и Пенсильвании за 2011г. Информация сгруппированна по зонам переписи (Census tracts (https://en.wikipedia.org/wiki/Census_tract)).

В построении диаграмм может помочь книга The R Cookbook (http://shop.oreilly.com/product/9780596809164.do). Рецепты 10.1 и 10.2. ## Загрузка и очистка данных - Загрузите данные в датафрейм, который назовите data.

```
data = read.csv("https://raw.githubusercontent.com/SergeyMirvoda/da2016/master/data/calif_penn_2011.csv")
```

- Сколько строк и столбцов в data?

```
nrow(data)
```

```
## [1] 11275
```

```
ncol(data)
```

```
## [1] 34
```

- Выполните следующую команду и объясните, что она делает. colSums(apply(data,c(1,2), is.na))

```
colSums(apply(data,c(1,2), is.na))#количество NA по каждому столбцу построчно
```

```
##                          X                     GEO.id2
##                          0                           0
##                    STATEFP                    COUNTYFP
##                          0                           0
##                    TRACTCE                  POPULATION
##                          0                           0
##                   LATITUDE                   LONGITUDE
##                          0                           0
##           GEO.display.label          Median_house_value
##                          0                         599
##                Total_units                Vacant_units
##                          0                           0
##               Median_rooms  Mean_household_size_owners
##                        157                         215
## Mean_household_size_renters         Built_2005_or_later
##                        152                          98
##          Built_2000_to_2004                 Built_1990s
##                          98                          98
##                 Built_1980s                 Built_1970s
##                          98                          98
##                 Built_1960s                 Built_1950s
##                          98                          98
##                 Built_1940s       Built_1939_or_earlier
##                          98                          98
##                  Bedrooms_0                  Bedrooms_1
##                          98                          98
##                  Bedrooms_2                  Bedrooms_3
##                          98                          98
##                  Bedrooms_4          Bedrooms_5_or_more
##                          98                          98
##                      Owners                     Renters
##                         100                         100
##     Median_household_income       Mean_household_income
##                         115                         126
```

- Функция na.omit() принимает датафрейм и возвращает новый датафрейм, игнорируя строки, содержащие значение NA. Используйте эту функцию для удаления строк с неполными данными.

```
newdata <- na.omit(data)
```

- Сколько строк было удалено?

```
nrow(data) - nrow(newdata)
```

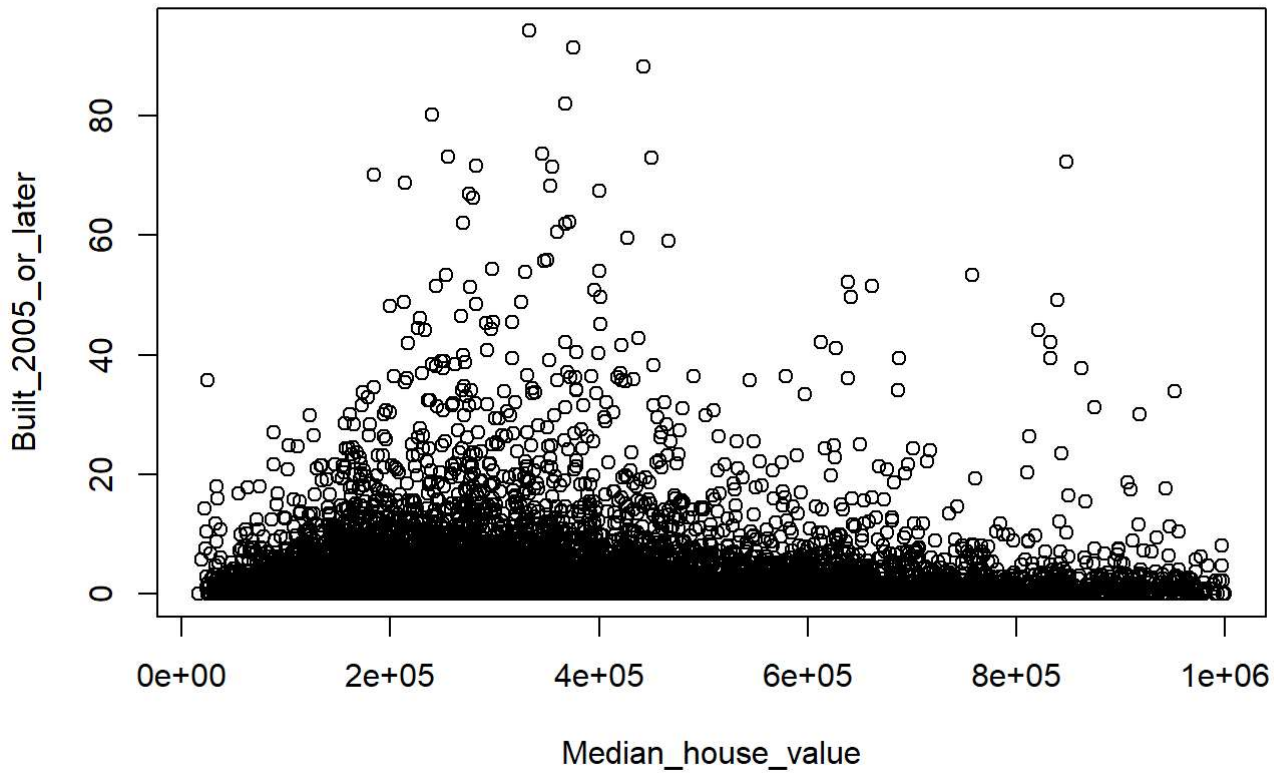
```
## [1] 670
```

- Соответствует ли результат выполнения, значениям из пункта 3? Нет, так как в одной строке мб по несколько NA

Новые дома

- Переменная(колонка) *Built_2005_or_later* содержит данные о проценте домов, построенных с 2005 года. Постройте диаграмму рассеяния (scatterplot) медианы стоимости домов (переменная *Median_house_value*) относительно процента новых домов.

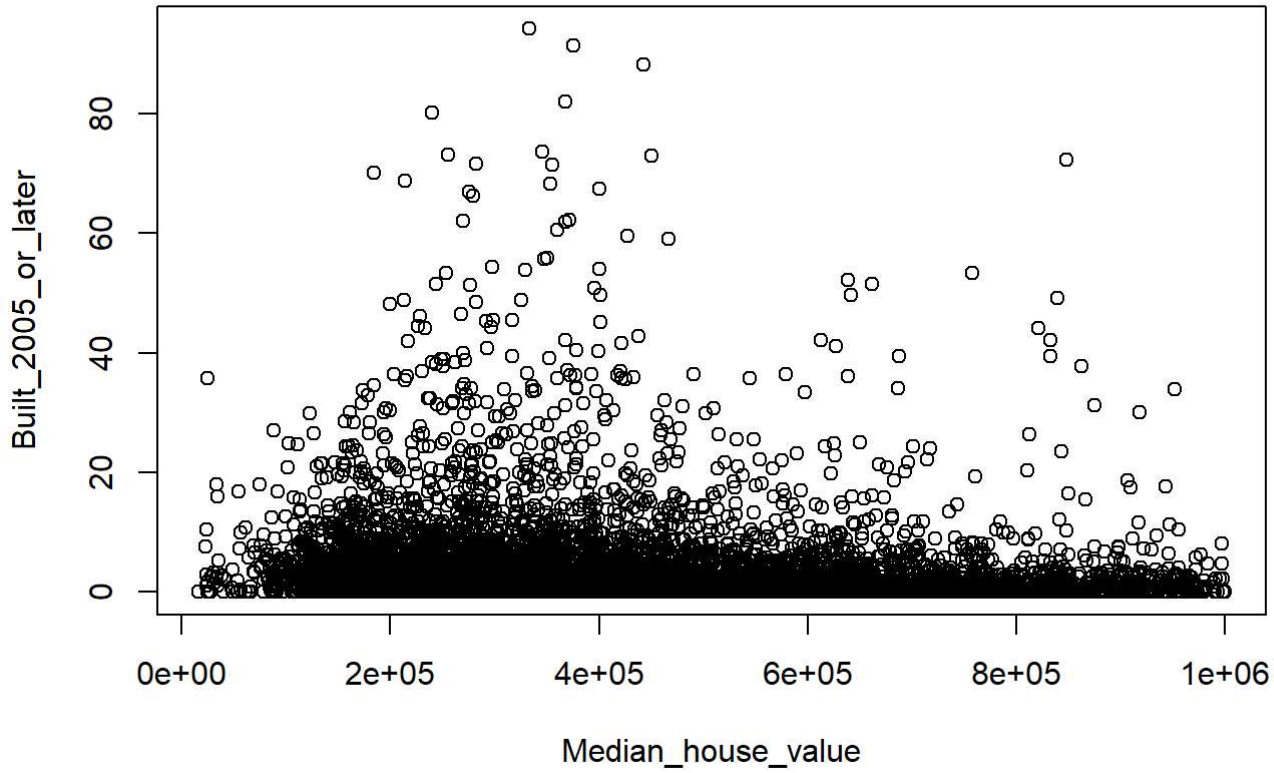
```
plot(newdata$Median_house_value,
     newdata$Built_2005_or_later,
     xlab = "Median_house_value",
     ylab = "Built_2005_or_later")
```



- Постройте ещё два графика для каждого из штатов отдельно. Номер штата содержится в переменной (*STATEFP*), где Калифорния 6-й штат, а Пенсильвания 42. Калифорния

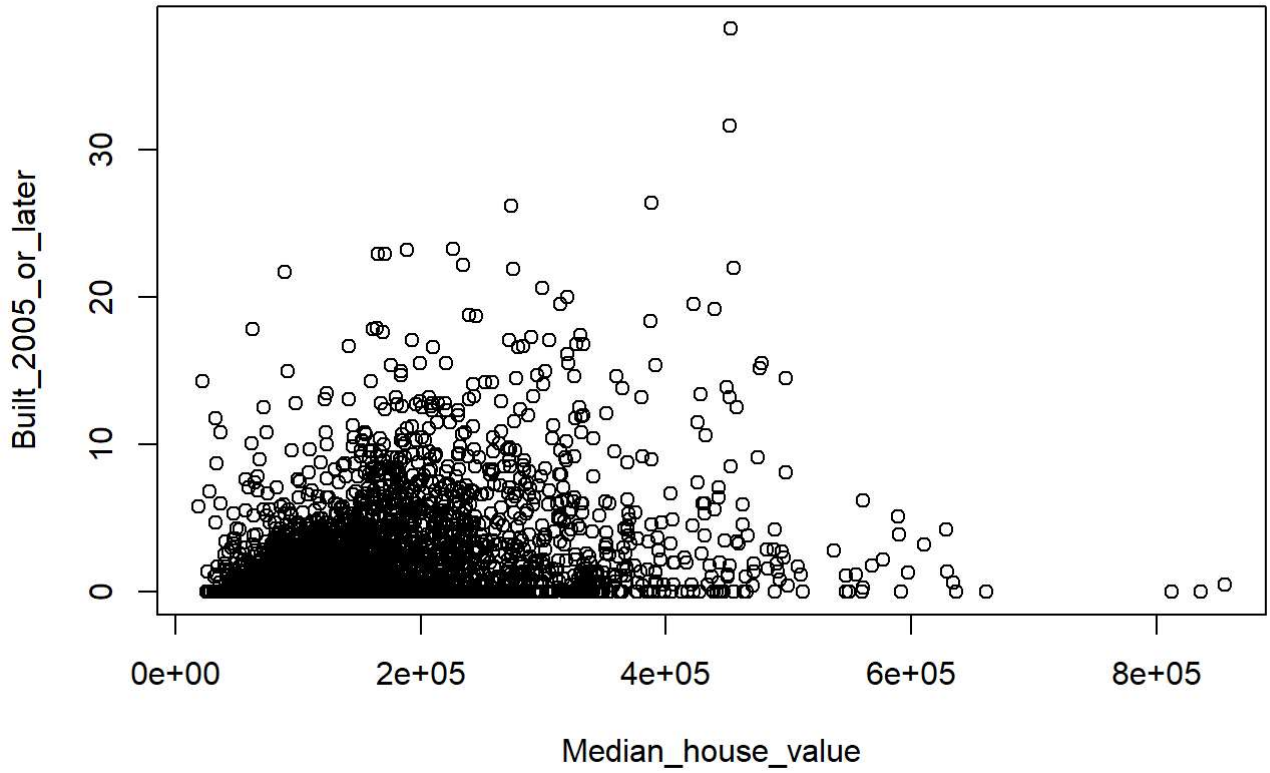
```
plot(newdata$Median_house_value[newdata$STATEFP == 6],
     newdata$Built_2005_or_later[newdata$STATEFP == 6],
     xlab = "Median_house_value",
     ylab = "Built_2005_or_later",
     main = "Калифорния")
```

<U+0430><U+043B><U+0438><U+0444><U+043E><U+0440><U+043D><U+0




```
plot(newdata$Median_house_value[newdata$STATEFP == 42],
     newdata$Built_2005_or_later[newdata$STATEFP == 42],
     xlab = "Median_house_value",
     ylab = "Built_2005_or_later",
     main = "Пенсильвания")
```

<U+043D><U+0441><U+0438><U+043B><U+044C><U+0432><U+0430><U+0



Незанятые дома

Уровень найма (vacancy rate) — доля домов, которые не были заняты. В данных содержатся колонки, содержащие общее количество домовладений и количество не занятых домовладений.

- В датафрейм *data* добавьте новую колонку *vacancy_rate*, которая должна содержать вышеописанный показатель.

```
newdata$vacancy_rate = newdata$Vacant_units / newdata$Total_units
```

- Найдите минимум, максимум, среднее и медиану полученных значений показателя.

```
min(newdata$vacancy_rate)
```

```
## [1] 0
```

```
max(newdata$vacancy_rate)
```

```
## [1] 0.965311
```

```
mean(newdata$vacancy_rate)
```

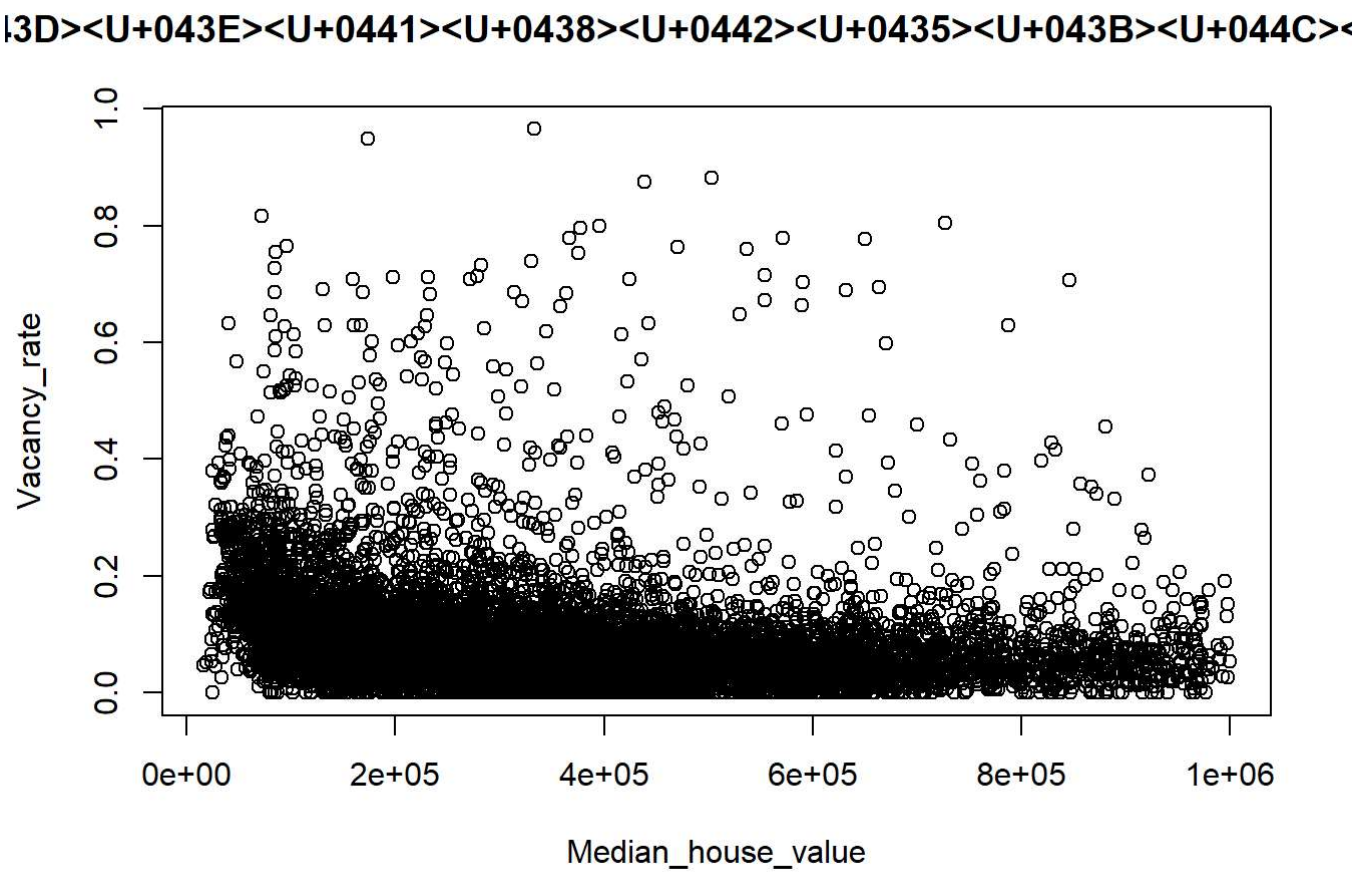
```
## [1] 0.08888789
```

```
median(newdata$vacancy_rate)
```

```
## [1] 0.06767283
```

- Постройте диаграмму уровня найма относительно медианы стоимости домов. Что можно заметить?

```
plot(newdata$Median_house_value,
     newdata$vacancy_rate,
     xlab = "Median_house_value",
     ylab = "Vacancy_rate",
     main = "Диаграмму уровня найма относительно медианы стоимости домов")
```



Чем больше средняя стоимость дома, тем меньше незанятых домов

Корреляция

Колонка *COUNTYFP* содержит числовой код округа внутри штата. Нас интересуют Butte County (округ 7 в Калифорнии), Santa Clara (округ 85 в Калифорнии) и York County (округ 133 в Пенсильвании).

- Объясните, что делает приведённый в конце задания код и как именно он это делает.

```
acc <- c()
for (tract in 1:nrow(newdata)) {
  if (newdata$STATEFP[tract] == 6) {
    if (newdata$COUNTYFP[tract] == 1) {
      acc <- c(acc, tract)
    }
  }
}
accmv <- c()
for (tract in acc) {
  accmv <- c(accmv, newdata[tract,10])
}
fw = median(accmv)
```

acc - индексы для строк штата 6 округа 1 accmv - Median house value для выбранных строк fw - медиана

- Напишите другим способом в одну строку, то же самое, что делает нижеуказанный код. Способов получить тот же ответ множество, достаточно одного.
- Найдите средний процент построенных домовладений в округах (Butte County, Santa Clara, York County)

1. Butte County
2. Santa Clara
3. York County

- Функция `cor` рассчитывает коэффициент корреляции между двумя переменными. Рассчитайте корреляцию между медианы стоимости домовладений (*Median_house_value*) и процентом построенных домов (*Built_2005_or_later*):

1. для всего набора данных
2. для Калифорнии
3. для Пенсильвании
4. для округа Butte County
5. для округа Santa Clara
6. для округа York County

- Постройте три диаграммы медианы стоимости домовладений (*Median_house_value*) относительно медианы дохода (*Median_household_income*) для трёх округов. Допустимо указать все три на одном графике.