

# Лабораторная работа 1. Работа с реальным датасетом. Первичная загрузка, очистка, анализ

## 1. Загрузка и подготовка датасета для анализа

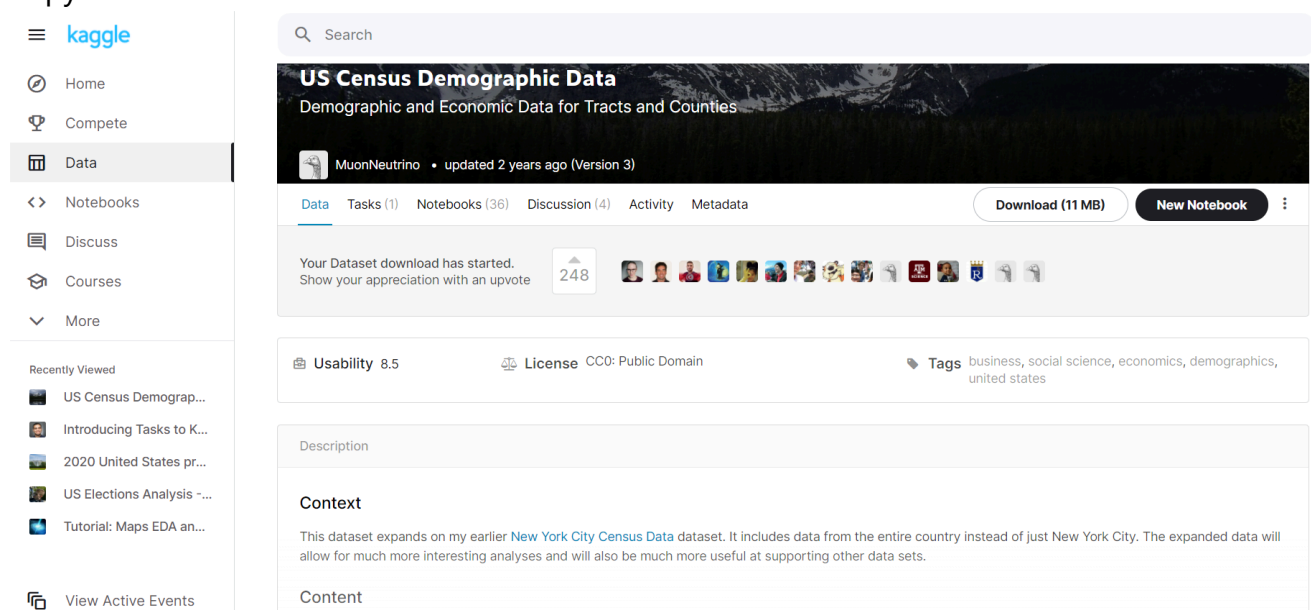
Вам уже дан скачанный датасет. Но если говорить всецело - мы будем использовать открытые источники с датасетами.

Например, как в сегодняшней лабораторной работе - Kaggle.

Кому интересно, ссылка на данный датасет - <https://postimg.cc/BPbXRkF3>

Если рассматривать другие датасеты для других надобностей, конечно мы сначала из текстов извлекаем общее описание, тему и так далее.

Конкретно этот набор данных включает в себя данные переписи по США 2015 года по округам всех штатов.



The screenshot shows the Kaggle interface for the 'US Census Demographic Data' dataset. The left sidebar contains navigation links: Home, Compete, Data (selected), Notebooks, Discuss, Courses, and More. Below these are 'Recently Viewed' items and a 'View Active Events' link. The main content area displays the dataset details: title 'US Census Demographic Data', subtitle 'Demographic and Economic Data for Tracts and Counties', author 'MuonNeutrino' (updated 2 years ago, Version 3), and a 'Download (11 MB)' button. A 'New Notebook' button is also present. Below the download button, a message states 'Your Dataset download has started. Show your appreciation with an upvote' and shows 248 upvotes. The dataset has a 'Usability' score of 8.5, a 'License' of 'CC0: Public Domain', and tags: 'business, social science, economics, demographics, united states'. The 'Description' section includes a 'Context' paragraph: 'This dataset expands on my earlier New York City Census Data dataset. It includes data from the entire country instead of just New York City. The expanded data will allow for much more interesting analyses and will also be much more useful at supporting other data sets.' The 'Content' section is currently empty.

В дальнейшем, по мере работы с датасетом, возможна корректировка - идеи для анализа, предлагаемые автором датасета.

Есть много вопросов, на которые мы могли бы попытаться ответить, используя данные здесь. Можем ли мы предсказать такие вещи, как состояние (классификация) или доход домохозяйства (регрессия)? Какие типы кластеров мы можем найти в данных?

**Context**

This dataset expands on my earlier [New York City Census Data](#) dataset. It includes data from the entire country instead of just New York City. The expanded data will allow for much more interesting analyses and will also be much more useful at supporting other data sets.

**Content**

The data here are taken from the DP03 and DP05 tables of the 2015 American Community Survey 5-year estimates. The full datasets and much more can be found at the [American Factfinder website](#). Currently, I include two data files:

1. acs2015censustract\_data.csv: Data for each census tract in the US, including DC and Puerto Rico.
2. acs2015countydata.csv: Data for each county or county equivalent in the US, including DC and Puerto Rico.

The two files have the same structure, with just a small difference in the name of the id column. Counties are political subdivisions, and the boundaries of some have been set for centuries. Census tracts, however, are defined by the census bureau and will have a much more consistent size. A typical census tract has around 5000 or so residents.

The Census Bureau updates the estimates approximately every year. At least some of the 2016 data is already available, so I will likely update this in the near future.

**Acknowledgements**

The data here were collected by the US Census Bureau. As a product of the US federal government, this is not subject to copyright within the US.

**Inspiration**

There are many questions that we could try to answer with the data here. Can we predict things such as the state (classification) or household income (regression)? What kinds of clusters can we find in the data? What other datasets can be improved by the addition of census data?

## 1.2 Выбор конкретного датасета, если их несколько

Повторюсь, датасет уже у вас есть. Но если бы вы работали по другой цели, то мы видим, что датасетов на странице несколько, ищем описания, по описанию выбираем датасет, с которым будем работать. В зависимости от варианта датасет может быть единственным.

Обращайте внимание на размер файлов.

Для отображения всех столбцов с их расшифровками выберите Select All

**Data Explorer**

29.22 MB

- acs2015\_census\_tract\_data...
- acs2015\_county\_data.csv** (598.86 KB)
- acs2017\_census\_tract\_data...
- acs2017\_county\_data.csv

**acs2015\_county\_data.csv (598.86 KB)**

10 of 37 columns

**About this file**

This data file includes census data for all counties. There are some accents in the data, so we use encoding='latin-1' to avoid errors.

CensusId	State	County
County Census ID	State, DC, or Puerto Rico	County or county equivalent

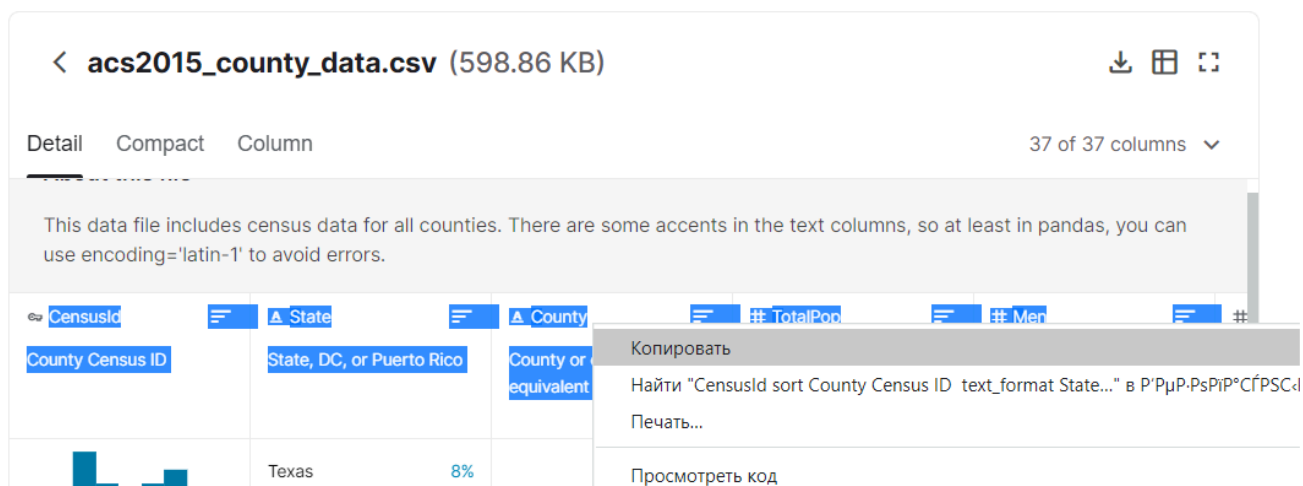
1928 unique values

37 of 37 selected

- ☒ Select all
- ☒ CensusId
- ☒ State
- ☒ County
- ☒ TotalPop

## 1.3 Копирование названий столбцов и их описаний с сайта

Все также :) Если бы вы работали с другим датасетом, по разделу About this file можно перемещаться стрелками. Обычно для анализа копируют названия столбцов и их описания, переводят описания, это пригодится для отчета и для того, чтобы выбрать те столбцы, с которыми дальше работать.



Очищенный текст преобразовывают в таблицу с двумя столбцами (Вставка - Таблица – преобразовать в таблицу), добавляют третий с переводом. (С нашей лабораторной работой переходим к пункту 1.4)

## 1.4 Отбор столбцов (признаков) для дальнейшей работы

Начинаем продумывать тему для анализа данных. На какие вопросы вы бы хотели (сможете) получить ответ, какие картинки нарисовать?

Работаем с таблицей описаний признаков (см. ниже)

Выделяем те признаки, которые в дальнейшем оставим для работы. Оставляйте идентификатор, 2-4 качественных и 3-4 количественных признака. Не нужно замахиваться на масштабные исследования))

Например, относительно данного датасета можно интересоваться распределением рабочих мест по разным формам собственности и уровнем безработицы, оставить данные о подушевом доходе, занятость в разрезах форм собственности рабочих мест и уровень безработицы.

В четвертом столбце укажите тип признака (качественный или количественный).

Определения типов данных выясните самостоятельно.

Название столбца (признака)	Смысл (англ.)	Смысл (на русском языке)	Тип признака
CensusId	County Census ID		идентификатор
State	State, DC, or Puerto Rico		качественный
County	County or county equivalent		качественный
TotalPop	Total population		количественный
Men	Number of men		
Women	Number of women		
Hispanic	% of population that is Hispanic/Latino		

Название столбца (признака)	Смысл (англ.)	Смысл (на русском языке)	Тип признака
White	% of population that is white		
Black	% of population that is black		
Native	% of population that is Native American/Native Alaskan		
Asian	% of population that is Asian		
Pacific	% of population that is Native Hawaiian or Pacific Islander		
Citizen	Number of citizens		
Income	Median household income (\$)		
IncomeErr	Median household income error (\$)		
IncomePerCap	Income per capita (\$)		количественный
IncomePerCapErr	Income per capita error (\$)		
Poverty	% under poverty level		
ChildPoverty	% of children under poverty level		
Professional	% employed in management, business, science, and arts		
Service	% employed in service jobs		
Office	% employed in sales and office jobs		
Construction	% employed in natural resources, construction, and maintenance		
Production	% employed in production, transportation, and material movement		
Drive	% commuting alone in a car, van, or truck		
Carpool	% carpooling in a car, van, or truck		
Transit	% commuting on public transportation		
Walk	% walking to work		
OtherTransp	% commuting via other means		
WorkAtHome	% working at home		

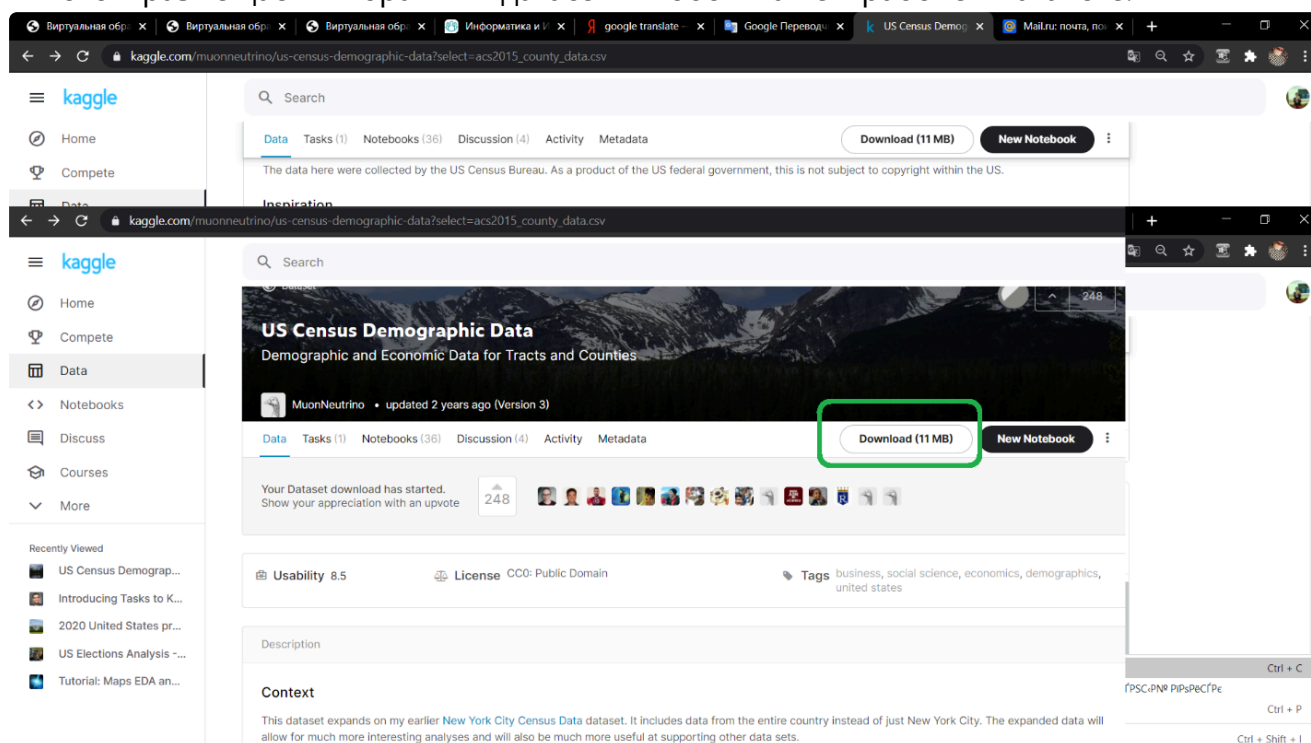
Название столбца (признака)	Смысл (англ.)	Смысл (на русском языке)	Тип признака
MeanCommute	Mean commute time (minutes)		
Employed	Number of employed (16+)		
PrivateWork	% employed in private industry		количественный
PublicWork	% employed in public jobs		количественный
SelfEmployed	% self-employed		количественный
FamilyWork	% in unpaid family work		количественный
Unemployment	Unemployment rate (%)		количественный

Полностью заполненная таблица у вас должна будет быть на отдельном листе в Excel для отчета.

## 1.5 Скачивание датасета (или архива) и сохраняем его

Если бы датасет у нас был не скачан, то пришлось бы это сделать!!! =)

И потом размещаем выбранный датасет в любом вашем рабочем каталоге.



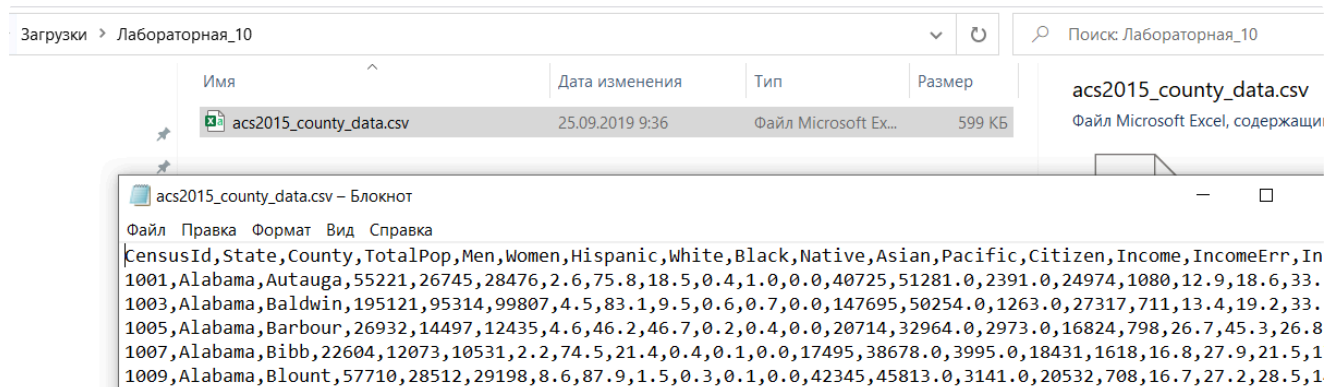
## 1.6 Загрузка датасета в Excel. Только два способа

Ваш файл имеет расширение CSV (от англ. Comma-Separated Values — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных. Строка таблицы соответствует строке текста, которая содержит одно или несколько полей, разделенных запятыми.

Формат CSV стандартизирован не полностью.

Поэтому при открытии в MS Excel данные в некоторых столбцах (даты, десятичные числа, номера версий продуктов) могут отображаться неверно.

Содержимое файла можно увидеть в Блокноте (Открыть с помощью...):



Посмотрите на данные в Блокноте.

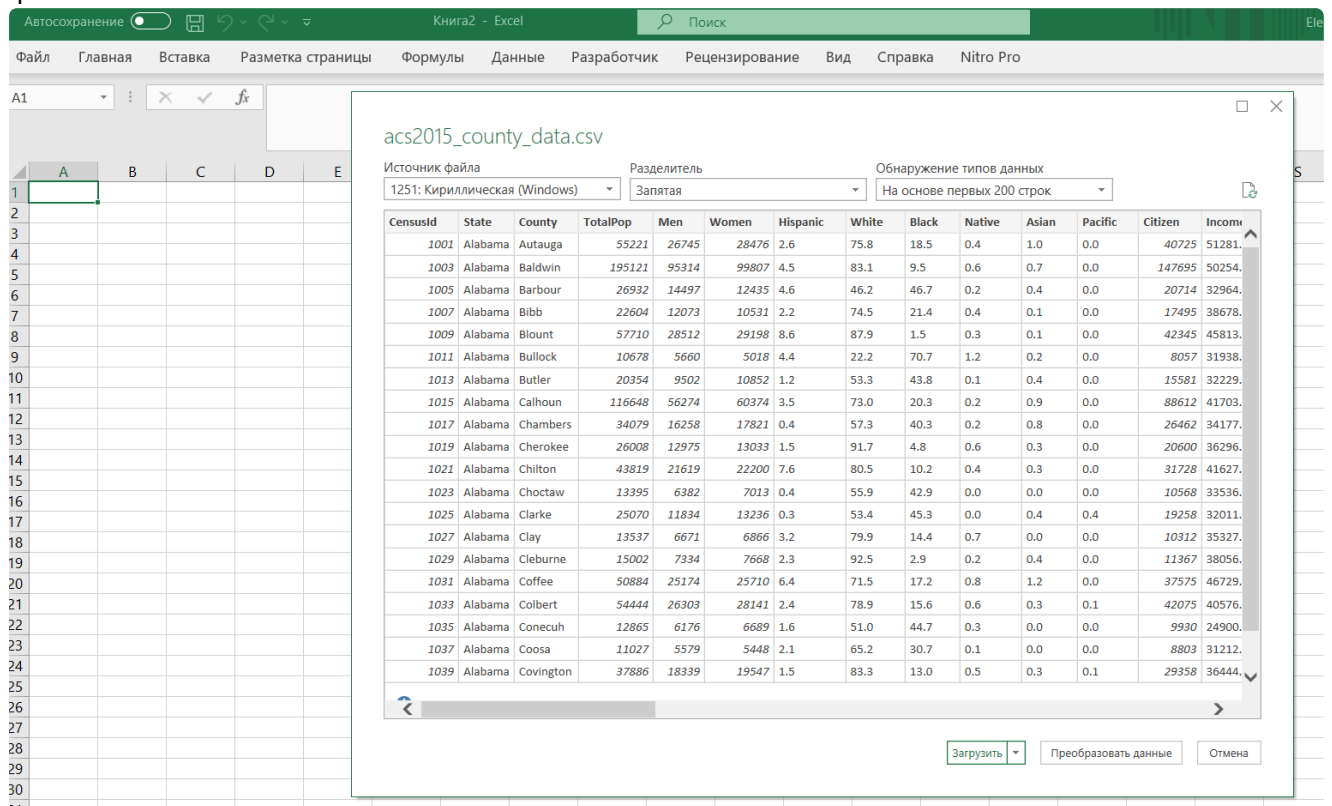
Закройте Блокнот.

Теперь импортируем датасет в MS Excel.

**Способ 1** (неофициальный, но рекомендую): В Блокноте выполнить замену запятой на точку с запятой по всему файлу. Сохранить под новым именем. Из Проводника новый файл открыть в MS Excel.

**Способ 2.**

Запускаем MS Excel. Создаем Новую книгу. Далее меню Данные – из текстового/CSV-файла.



Нажимаем на загрузить.

	H	I	J	K	L	M	N
c	White	Black	Native	Asian	Pacific	Citizen	Income
	75.8	18.5	0.4	1.0	0.0	40725	51281.0
	83.1	9.5	0.6	0.7	0.0	147695	50254.0
	46.2	46.7	0.2	0.4	0.0	20714	32964.0
	74.5	21.4	0.4	0.1	0.0	17495	38678.0
	87.9	1.5	0.3	0.1	0.0	42345	45813.0
	22.2	70.7	1.2	0.2	0.0	8057	31938.0
	53.3	43.8	0.1	0.4	0.0	15581	32229.0
	73.0	20.3	0.2	0.9	0.0	88612	41703.0
	57.3	40.3	0.2	0.8	0.0	26462	34177.0
	91.7	4.8	0.6	0.3	0.0	20600	36296.0
	80.5	10.2	0.4	0.3	0.0	31728	41627.0
	55.9	42.9	0.0	0.0	0.0	10568	33536.0

**Замечание** к обоим способам.

Если какие-то столбцы исказились (например, версии продукта 4.01.03 превратилась в 4 января 2003), то в данной лабораторной работе откажитесь от использования этих столбцов, возьмите для анализа другие. В реальных условиях (на работе) рекомендую открыть файл в Google Table или Libre Office, искажений будет меньше. Дальше исправлять средствами Excel.

Проверьте, что установлен разделитель целой и десятичной части как точка (Файл – Параметры – снять галочку Использовать системные разделители – установить Разделитель точка)

Оставим только выбранные ранее столбцы. Удалите лишние.

!!! Я немножко побуянила и попортила данный датасет =)

Ваша задача его исправить) Подскажу насчет одного - некоторые штаты записаны русским языком, такого быть не должно. Глазками найти и исправить)

Некоторые данные в столбцах также немножко мною такой заразой испорчены.

Найдите как и подумайте каким автоматическим способом это все можно исправить.

Как видите, все остальное - числовые значения, те которые испорчены - по аналогии с правильными вставьте просто свое значение. !!!!

## 2. Основные статистические характеристики

**Ценное Замечание:** Выделить диапазон от позиции курсора до конца вниз Ctrl-Shift-↓  
Создадим новый лист с названием Описательные характеристики, скопируем на него



заголовки столбцов:

3220	72151	Puerto Rico	Yabucoa	36279	7960	8083	65.1	27.6	7.3
3221	72153	Puerto Rico	Yauco	39474	7743	8923	68.0	27.6	4.4
3222									
3223									
3224									

## 2.1 Описательные характеристики

Вспомним немножечко наш всеми любимый тервер)))

Для количественных данных рассчитаем, пользуясь функциями и переходя на нужные листы:

- средние значения (=СРЗНАЧ(...))
- дисперсии (=ДИСП())
- среднеквадратические отклонения (=СТАНДОТКЛОН(...))
- медианы (=МЕДИАНА(...))
- моды (=МОДА(...))

Выяснить самостоятельно смысл этих понятий.

Лабораторная работа 10 Пример.xls									
Поиск									
Файл Главная Вставка Разметка страницы Формулы Данные Разработчик Рецензирование Вид Справка Nitro Pro									
B2 =СРЗНАЧ(acs2015_county_data[TotalPop])									
	A	B	C	D	E	F	G	H	I
1	Количественные данные	TotalPop	IncomePerCap	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment
2	Среднее значение	99409.34596	23981.77174	45593.5183	74.21934783	17.56086957	7.931801242	0.28810559	8.094440994
3	Дисперсия	1.01956E+11	38493834.34	2.241E+10	61.82972496	42.3847063	15.32702192	0.207149563	16.77815175
4	Среднеквадратическое отклонение	319305.4537	6204.33996	149699.504	7.863187964	6.510353777	3.914974063	0.455136862	4.096114226
5	Медиана	26035	23460	10508	75.7	16.2	6.9	0.2	7.6
6	Мода	8697							
7									
8	Качественные данные	State	County						
9	Мода								
10									

Замечание: Обратите внимание на запись диапазона ячеек: если вы используете строку заголовков (щелчок по таблице – Конструктор таблиц – строка заголовков), то диапазон записывается по названию заголовка =СРЗНАЧ(acs2015\_county\_data[TotalPop]), а не =СРЗНАЧ(D2:D3221)

## 2.2 Описательные характеристики для качественных признаков

Для качественных данных мы рассчитаем моды.

Моду можно найти в Excel, если построить частотную таблицу (таблица частоты встречаемости для каждого значения признака) и взять максимальное значение. Сделаем это в разделе Визуализация



## 2.3 Что делать, если числа воспринимаются как текст?

Возможно, при вычислении среднего появится деление на ноль. Причина в том, что, хотя формат ячейки Числовой, данные воспринимаются как текстовые. Исправление: Главная – Заменить – точку на точку (да-да!)

Всё получится.

The screenshot displays an Excel spreadsheet with columns J through N. The 'FamilyWork' and 'Unemployment' columns contain numerical data. A 'Формат ячеек' (Format Cells) dialog box is open, showing the 'Число' (Number) tab. The 'Числовые форматы' (Number formats) list has 'Числовой' (Number) selected. The 'Образец' (Sample) shows '0.4'. The 'Число десятичных знаков' (Number of decimal places) is set to 2. The 'Отрицательные числа' (Negative numbers) section shows four options: '-1234.10' (selected), '1234.10', '-1234.10', and '-1234.10'.

Below the dialog box, the spreadsheet shows columns H through O. The 'PublicWork', 'SelfEmployed', 'FamilyWork', and 'Unemployment' columns contain numerical data. A formula bar shows the formula `=CP3HACH(acs2015[_county_data[@[FamilyWork]:[Unemployment]])`.

The 'Аргументы функции' (Function Arguments) dialog box is open, showing the arguments for the CP3HACH function. The 'Число1' (Number1) argument is `@[FamilyWork]:[Unemployment]` and the 'Число2' (Number2) argument is `число`.

## 3. Визуальный анализ

Формулировать постановку задачи для визуализации нужно самостоятельно. Достаточно одного - двух графиков для каждой комбинации типов данных: два количественных признака, два качественных признака, качественный и количественный признаки.

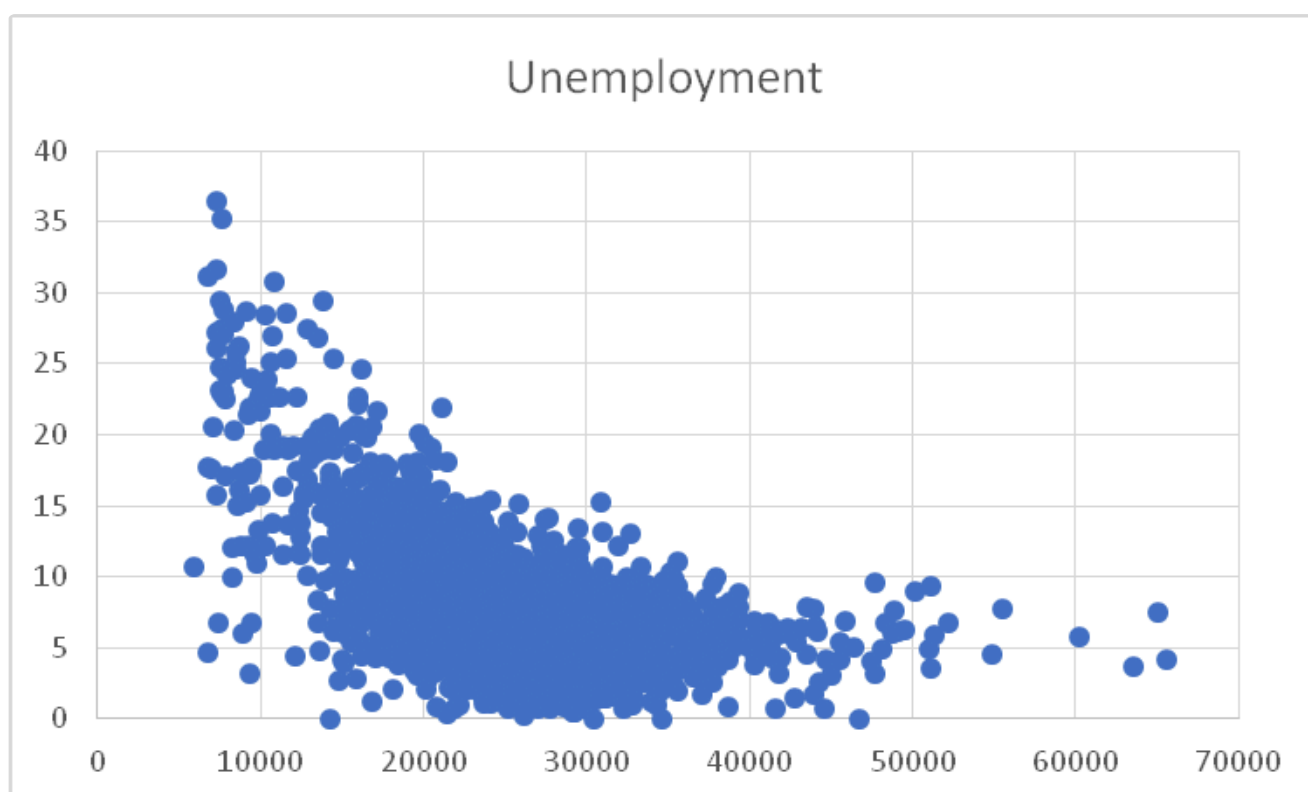
## 3.1 Визуализация: два количественных признаков

Да тут автором датасета уже самостоятельно сформулированы такие задачи

### 3.1.1 Точечная диаграмма

Задача 1. Визуализировать, как распределены значения подушевого дохода и уровня безработицы

Решение. Строим Точечную диаграмму по двум столбцам



Вывод: На основе графика можно выдвинуть гипотезу, что зависимость есть, обратная.

### 3.1.2 Гистограмма распределения

Задача 2. Какие значения численности населения распределены по интервалам с шагом?

Нужно построить гистограмму распределения значений признака численность населения.

Гистограмма распределения отражает частоты попадания значений количественного признака в интервалы. Это НЕ диаграмма Гистограмма.

Построить можно, воспользовавшись надстройкой Пакет анализа.

Но в данной работе сделаем вручную.

## Последовательность действий:

- определить количество интервалов у гистограммы; используем формулу Стёрджеса  $N=1+\log_2(n)=1+\log_2(3221)=13$ . Здесь  $n$  – объём выборки.
- определить ширину интервала (с учетом округления); Найдем минимальное и максимальное значения, их разность разделим на  $N$
- определить границу первого интервала;
- сформировать таблицу интервалов и рассчитать количество значений, попадающих в каждый интервал (частоту); Для вычисления количества значений, попадающих в каждый интервал, использована формула массива на основе функции ЧАСТОТА()
- построить гистограмму. Диаграмма Гистограмма с группировкой

Числовые характеристики TotalPop	
Объём выборки, $n$	3221
Число интервалов, $N$	13
Минимальное значение	85
Максимальное значение	10038388
Ширина интервала	772178

Интервалы	Обозначение интервала	Нижняя граница	Верхняя граница	Частота
1	< 772263	85	772263	1
2	< 1544441	772263	1544441	3144
3	< 2316619	1544441	2316619	0
4	< 3088797	2316619	3088797	52
5	< 3860975	3088797	3860975	0
6	< 4633153	3860975	4633153	14
7	< 5405331	4633153	5405331	0
8	< 6177509	5405331	6177509	3
9	< 6949687	6177509	6949687	0
10	< 7721865	6949687	7721865	2
11	< 8494043	7721865	8494043	0
12	< 9266221	8494043	9266221	2
13	< 10038389	9266221	10038389	0



## 3.2 Визуализация: качественные признаки

### 3.2.1 Частотная таблица

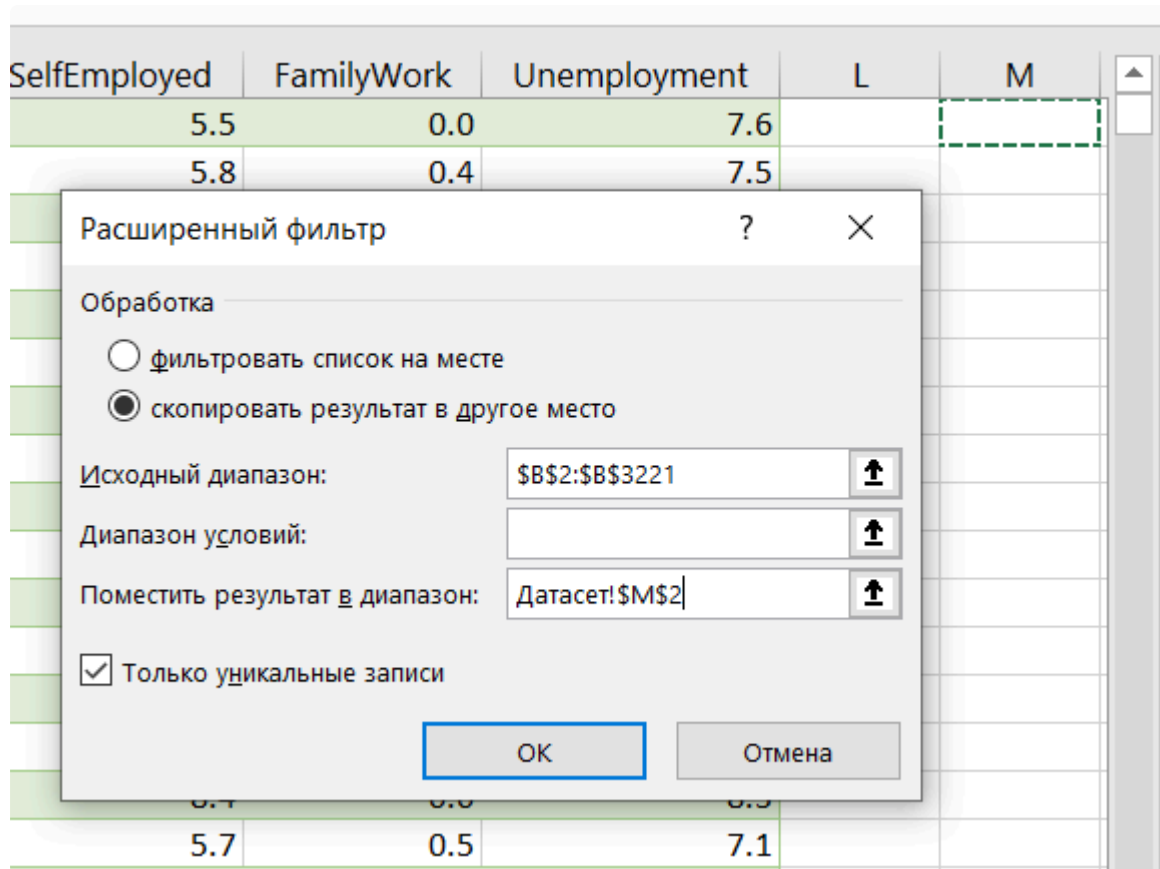
Задача 3. Подсчитать количество округов в каждом штате

Это и будет частотная таблица

Для этого подсчитаем, сколько раз название каждого штата встретилось (частоту появления признака Штат)

Сделать можно с помощью расширенного фильтра

(Данные – блок Сортировка и Фильтр – Дополнительно (Расширенный фильтр))



**Замечание.** Список уникальных значений можно разместить только на тот же лист. Затем подсчитаем количество вхождений каждого названия штата в столбик State с помощью функции =СЧЁТЕСЛИ(B\$2:B\$3221;M3) – аргументы приведены для примера

N3

✕

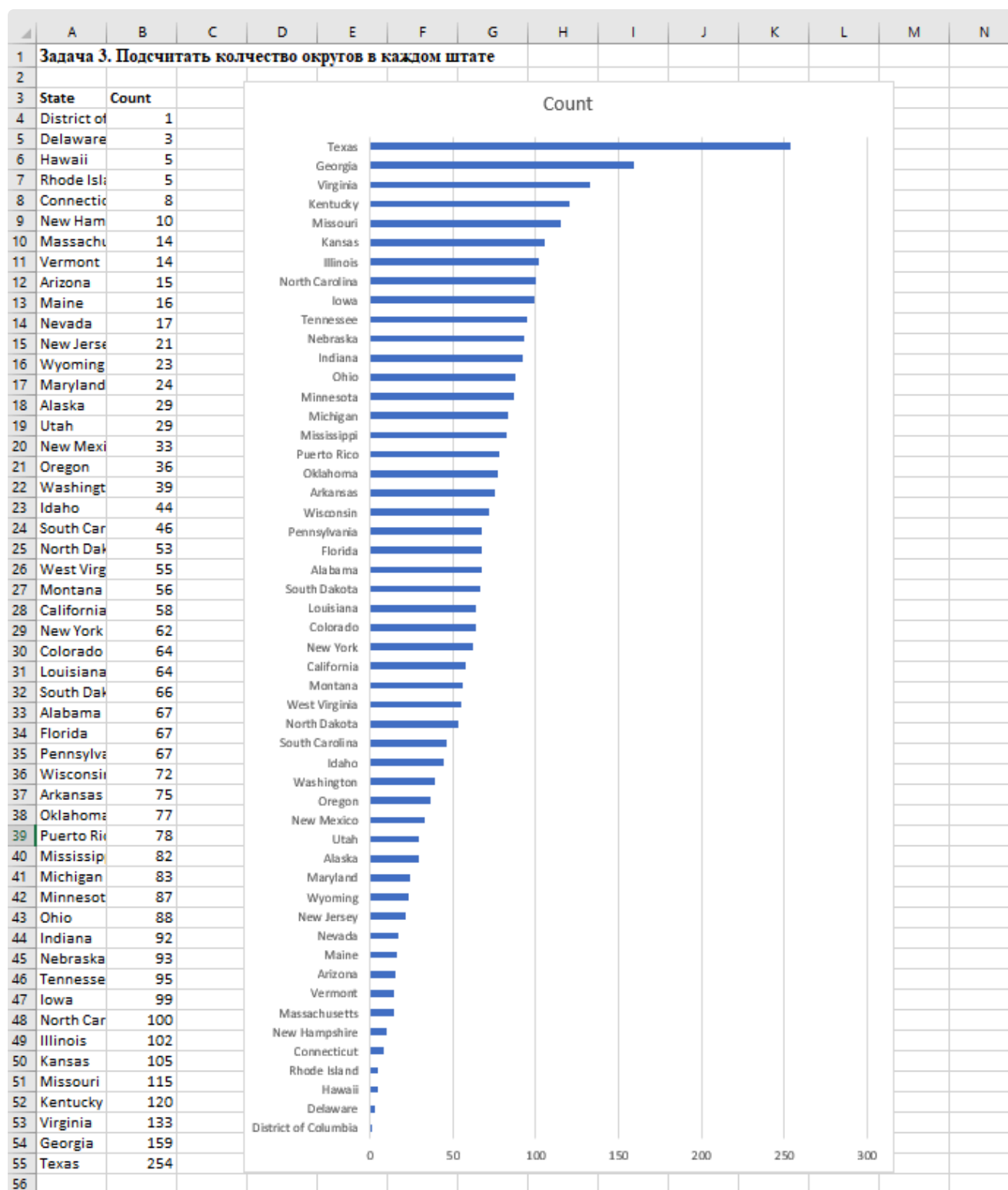
✓

*f<sub>x</sub>*

=СЧЁТЕСЛИ(\$B\$2:\$B\$3221;M3)

	K	L	M	N	O	P
1	Unemployment					
2	7.6		State	Count		
3	7.5		Alabama	67		
4	17.6		Alaska	29		
5	8.3		Arizona	15		
6	7.7		Arkansas	75		

Перенесем полученные данные на лист Качественные признаки, отсортируем и построим диаграмму



### 3.2.2 Таблица сопряженности

Задача 4.

В данном датасете мало качественных признаков, поэтому пришлось изобрести свои. Для второй задачи введем два бинарных признака (да – нет):

PublicWork\_bin – уровень занятых в государственном секторе выше, чем медианное значение по стране

Unemployment\_l\_bin - уровень безработицы выше, чем медианное значение по стране

Выяснить, как распределены округа по этим двум признакам. Составить таблицу сопряженности для этих двух признаков.

Решение.

Для заполнения значений обоих признаков используем функцию =ЕСЛИ(...)

=ЕСЛИ([@PublicWork]>='Описательные характ-ки'!\$F\$5;"да";"нет")				
J	K	L	M	N
FamilyWork	Unemployment	PublicWork_bin	Unemployment	
0.0	7.6	да	да	
0.4	7.5	нет	нет	

Составим таблицу сопряженности для этих двух бинарных признаков. Для этого сформируем сводную таблицу (Вставить – Сводная таблица)

Количество по полю Unemployment	Названия столбцов		
Названия строк	да	нет	Общий итог
да	932	697	1629
нет	700	891	1591
Общий итог	1632	1588	3220

Поиск

☒ PublicWork\_bin  
☒ Unemployment

Другие таблицы...

Перетащите поле в нужную область:

Фильтры

Столбцы

Строки

Значения

PublicWork\_bin

Количество по полю Unemployment

Количество по полю Unemployment

В итоге получаем следующую таблицу сопряженности.

	Unemployment_	_bin %	
PublicWork_bin	да	нет	n
да	57.2	42.8	1629
нет	44	56	1591
Общий итог	1632	1588	3220

Вывод. Округа с различающимся уровнем присутствия гос. сектора сильно различаются по уровню безработицы (гипотеза).



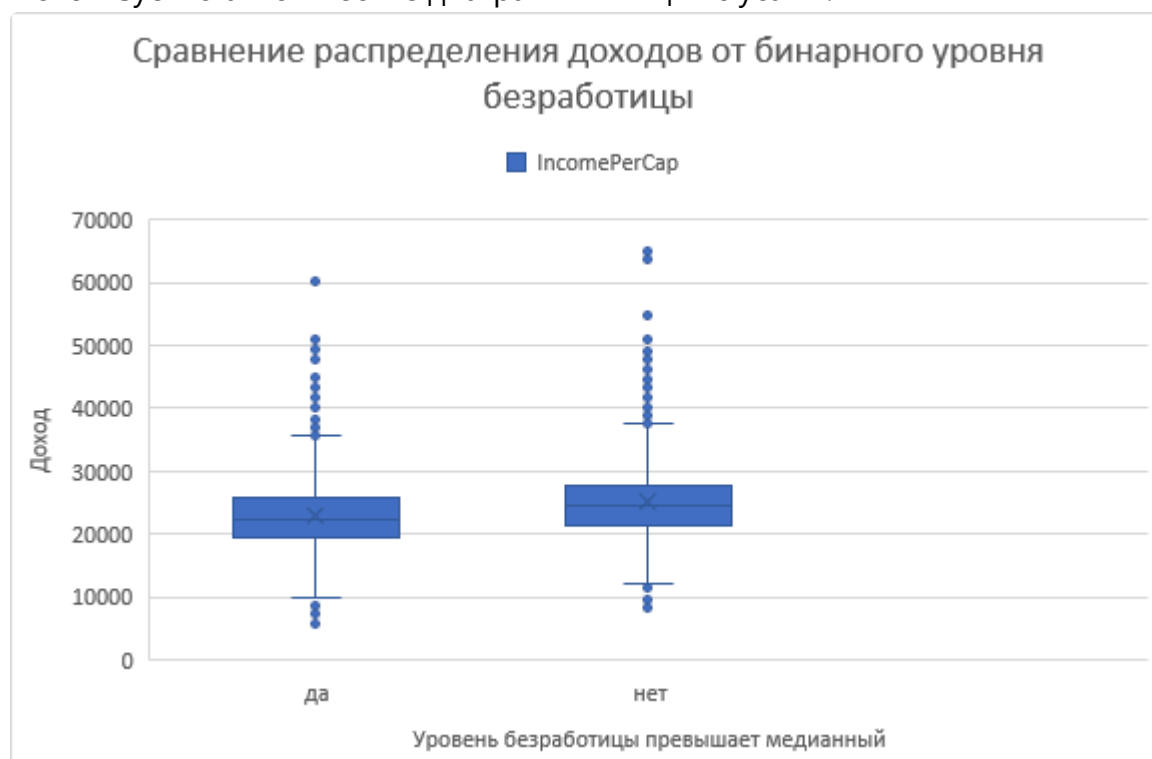
Методики анализа и проверки гипотез таких таблиц широко применяются гуманитариями, но их рассмотрение выходит за рамки данного курса.

### 3.3 Визуализация: Количественный и качественный признаки

При таком сочетании, как правило, интересуются распределением количественного признака при разных значениях качественного.

#### 3.3.1 Распределение количественного признака для разных значений (категорий) качественного

Задача 5. Построим распределение доходов в зависимости от признака Unemployment\_bin: превышает в округе уровень безработицы медианный уровень. Используем статистические диаграммы – ящик с усами.



Вывод. Распределение подушевого дохода зависит от уровня безработицы (гипотеза)

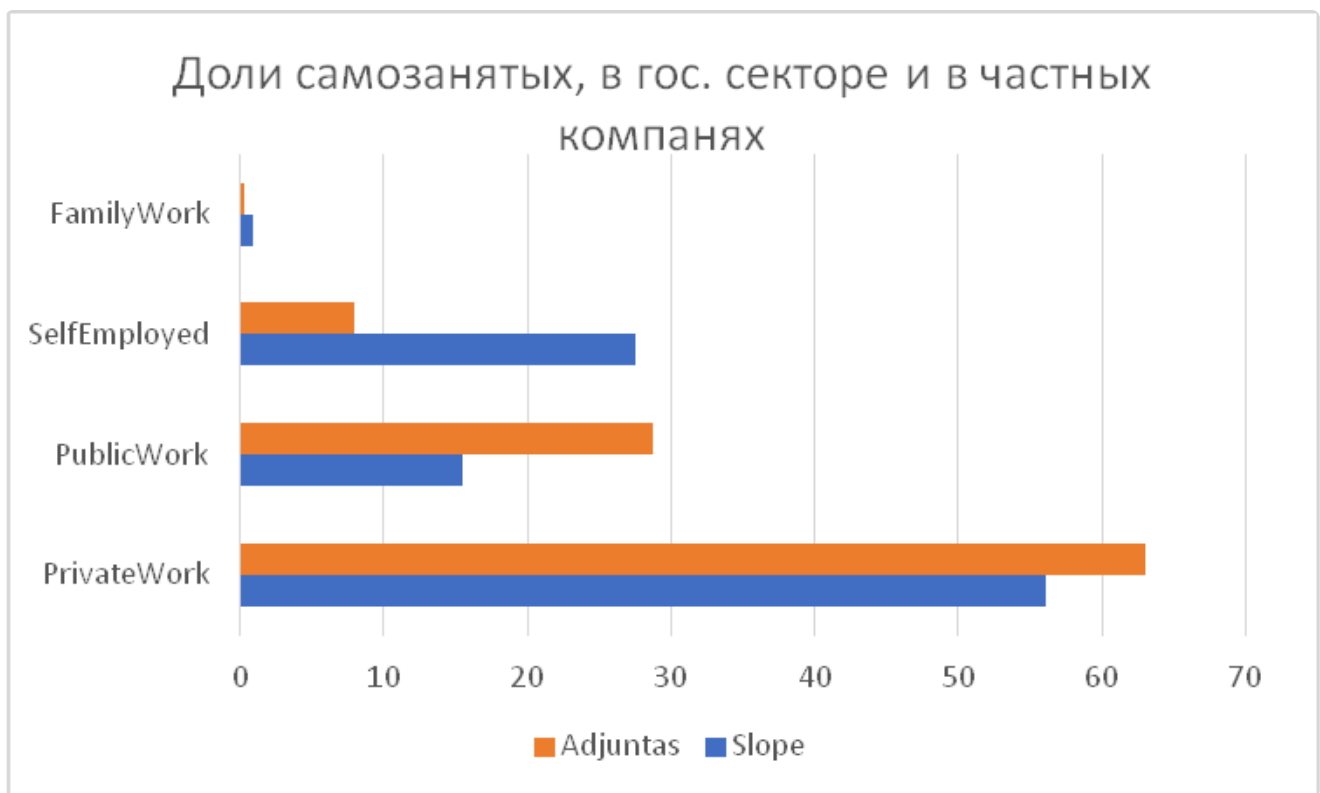
Проверка гипотезы – за рамками курса.

#### 3.3.2 Линейчатая диаграмма с категориями

Задача 6. Визуализировать, как распределены доли рабочих мест по разным видам собственности в округе с минимальным уровнем безработицы в сравнении с округом с максимальным уровнем безработицы.

Имеем пять количественных признаков – уровень безработицы и дол занятых в разных видах собственности. Извлечем из данных нужную информацию и построим диаграмму.

Найдем округ с максимальным уровнем безработицы и с минимальным, скопируем соответствующие строки, построим линейчатую диаграмму



Вывод: В округе с минимальной безработицей доля самозанятых превосходит долю занятых в государственном секторе. В округе с максимальной безработицей - наоборот. Можно выдвинуть гипотезы, что уровень безработицы связан с долей самозанятых и долей рабочих мест, принадлежащих государственным структурам.

Но проверять их нужно на всех выборках с помощью соответствующих методов проверки гипотез, а не по одной паре случаев.