Московский государственный технический университет им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

Лабораторная работа №1 по курсу «Методы машинного обучения»

«Разведочный анализ данных. Исследование и визуализация данных»

Писчая бумага

(вид носителя)

Исполнитель:

Студент группы ИУ5-24М

Харлашкин А. И.

«\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2019 г.

Москва

2019 год

**Описание задания**

Выбрать набор данных (датасет).

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

Сформировать отчет и разместить его в своем репозитории на github.

# Текст программы

*import numpy as np*

*import pandas as pd*

*import seaborn as sns*

*import matplotlib.pyplot as plt*

*%matplotlib inline*

*sns.set(style="ticks")*

*# Будем анализировать данные только на обучающей выборке*

*data = pd.read\_csv('economic\_freedom\_index2019\_data.csv', sep=",", encoding='latin-1')*

*# Первые 5 строк датасета*

*data.head()*

*# Размер датасета - 8143 строк, 7 колонок*

*data.shape*

*total\_count = data.shape[0]*

*print('Всего строк: {}'.format(total\_count))*

*# Список колонок*

*data.columns*

*# Список колонок с типами данных*

*data.dtypes*

*# Проверим наличие пустых значений*

*# Цикл по колонкам датасета*

*for col in data.columns:*

*# Количество пустых значений - все значения заполнены*

*temp\_null\_count = data[data[col].isnull()].shape[0]*

*print('{} - {}'.format(col, temp\_null\_count))*

*# Основные статистические характеристки набора данных*

*data.describe()*

*# Определим уникальные значения для целевого признака*

*data['Financial Freedom'].unique()*

*fig, ax = plt.subplots(figsize=(10,10))*

*sns.scatterplot(ax=ax, x='Property Rights', y='Government Integrity', data=data)*

*fig, ax = plt.subplots(figsize=(10,10))*

*sns.scatterplot(ax=ax, x='Property Rights', y='Government Integrity', data=data, hue='Financial Freedom')*

*fig, ax = plt.subplots(figsize=(10,10))*

*sns.distplot(data['Government Integrity'].fillna(0))*

*sns.jointplot(x='Property Rights', y='Government Integrity', data=data)*

*sns.jointplot(x='Property Rights', y='Government Integrity', data=data, kind="hex")*

*sns.jointplot(x='Property Rights', y='Government Integrity', data=data, kind="kde")*

*sns.pairplot(data)*

*subset1 = ['World Rank', 'Region Rank', '2019 Score']*

*data[subset1] = data[subset1].astype(float)*

*#sns.pairplot(data, hue="Financial Freedom")*

*sns.pairplot(data)*

*plt.show()*

*sns.boxplot(x=data['Financial Freedom'])*

*# По вертикали*

*sns.boxplot(y=data['Financial Freedom'])*

*# Распределение параметра Humidity сгруппированные по Occupancy.*

*sns.boxplot(x='Financial Freedom', y='Property Rights', data=data)*

*sns.violinplot(x=data['Financial Freedom'])*

*fig, ax = plt.subplots(2, 1, figsize=(10,10))*

*sns.violinplot(ax=ax[0], x=data['Financial Freedom'])*

*sns.distplot(data['Financial Freedom'].fillna(0), ax=ax[1])*

*# Распределение параметра Humidity сгруппированные по Occupancy.*

*sns.violinplot(x='Financial Freedom', y='Property Rights', data=data)*

*sns.catplot(y='Property Rights', x='Financial Freedom', data=data, kind="violin", split=True)*

*data.corr()*

*data.corr(method='pearson')*

*data.corr(method='kendall')*

*data.corr(method='spearman')*

*sns.heatmap(data.corr())*

*# Вывод значений в ячейках*

*sns.heatmap(data.corr(), annot=True, fmt='.3f')*

*# Изменение цветовой гаммы*

*sns.heatmap(data.corr(), cmap='YlGnBu', annot=True, fmt='.3f')*

*# Треугольный вариант матрицы*

*mask = np.zeros\_like(data.corr(), dtype=np.bool)*

*# чтобы оставить нижнюю часть матрицы*

*# mask[np.triu\_indices\_from(mask)] = True*

*# чтобы оставить верхнюю часть матрицы*

*mask[np.tril\_indices\_from(mask)] = True*

*sns.heatmap(data.corr(), mask=mask, annot=True, fmt='.3f')*

*fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))*

*sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')*

*sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')*

*sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.2f')*

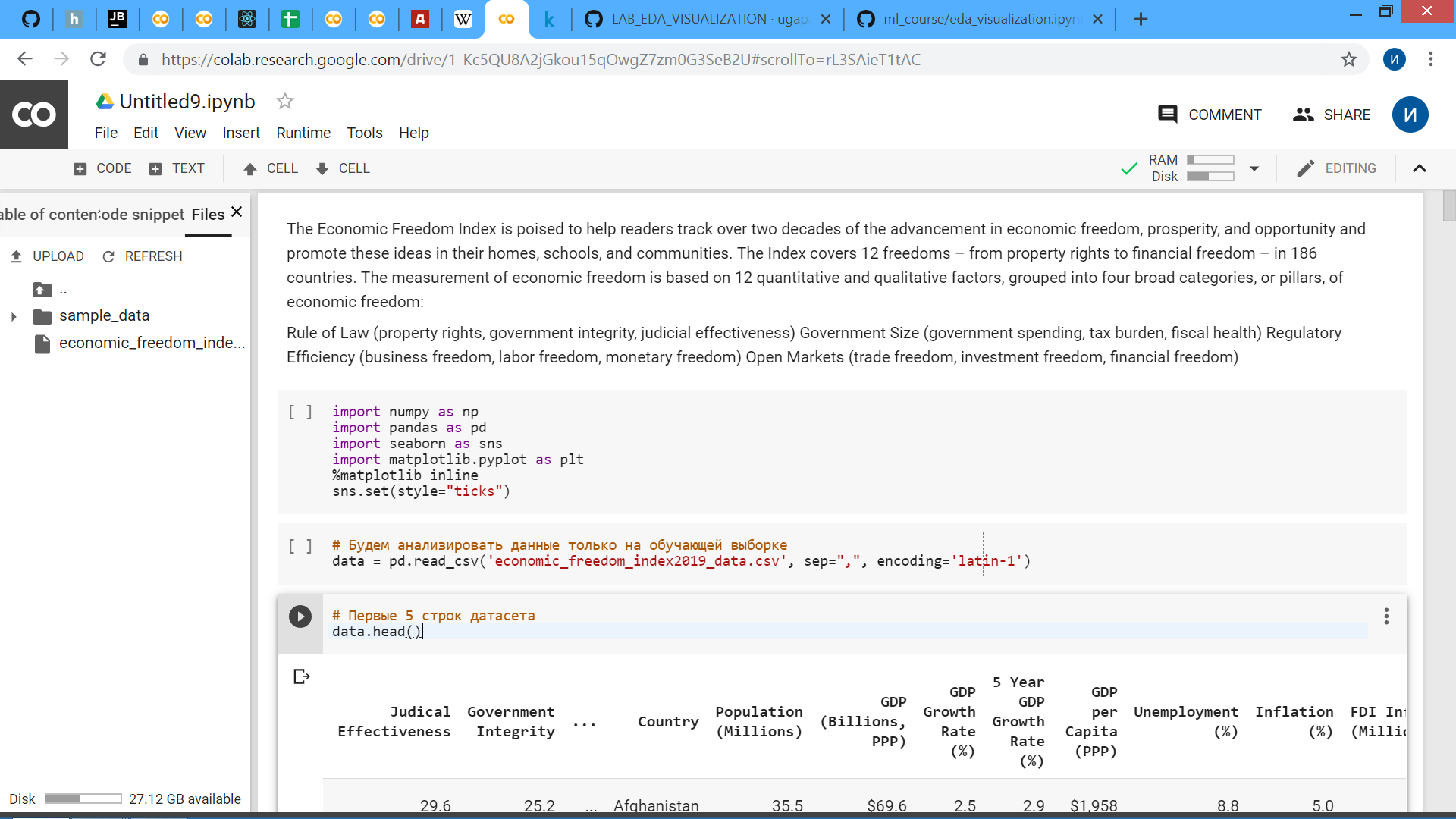
*fig.suptitle('Корреляционные матрицы, построенные различными методами')*

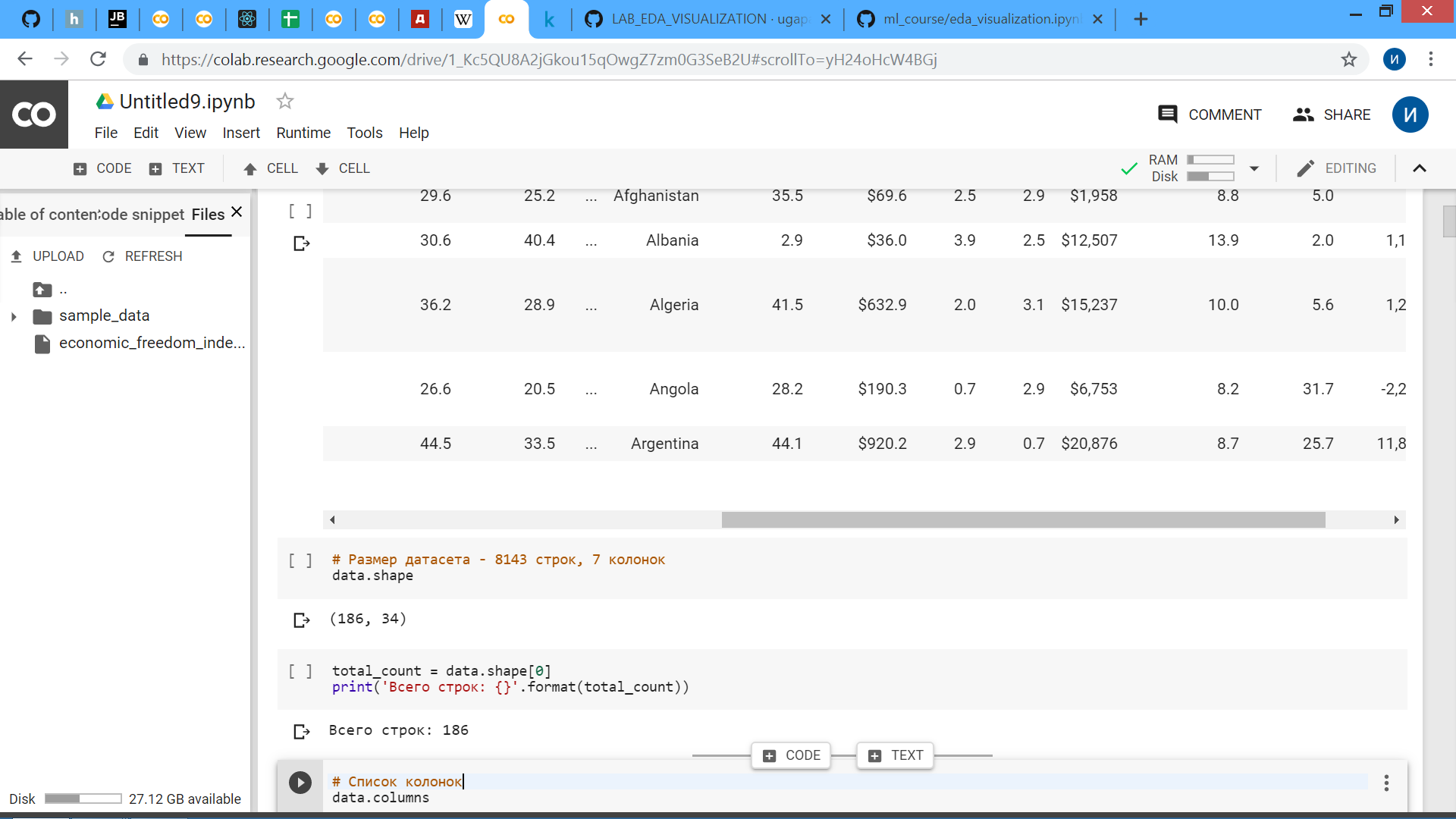
*ax[0].title.set\_text('Pearson')*

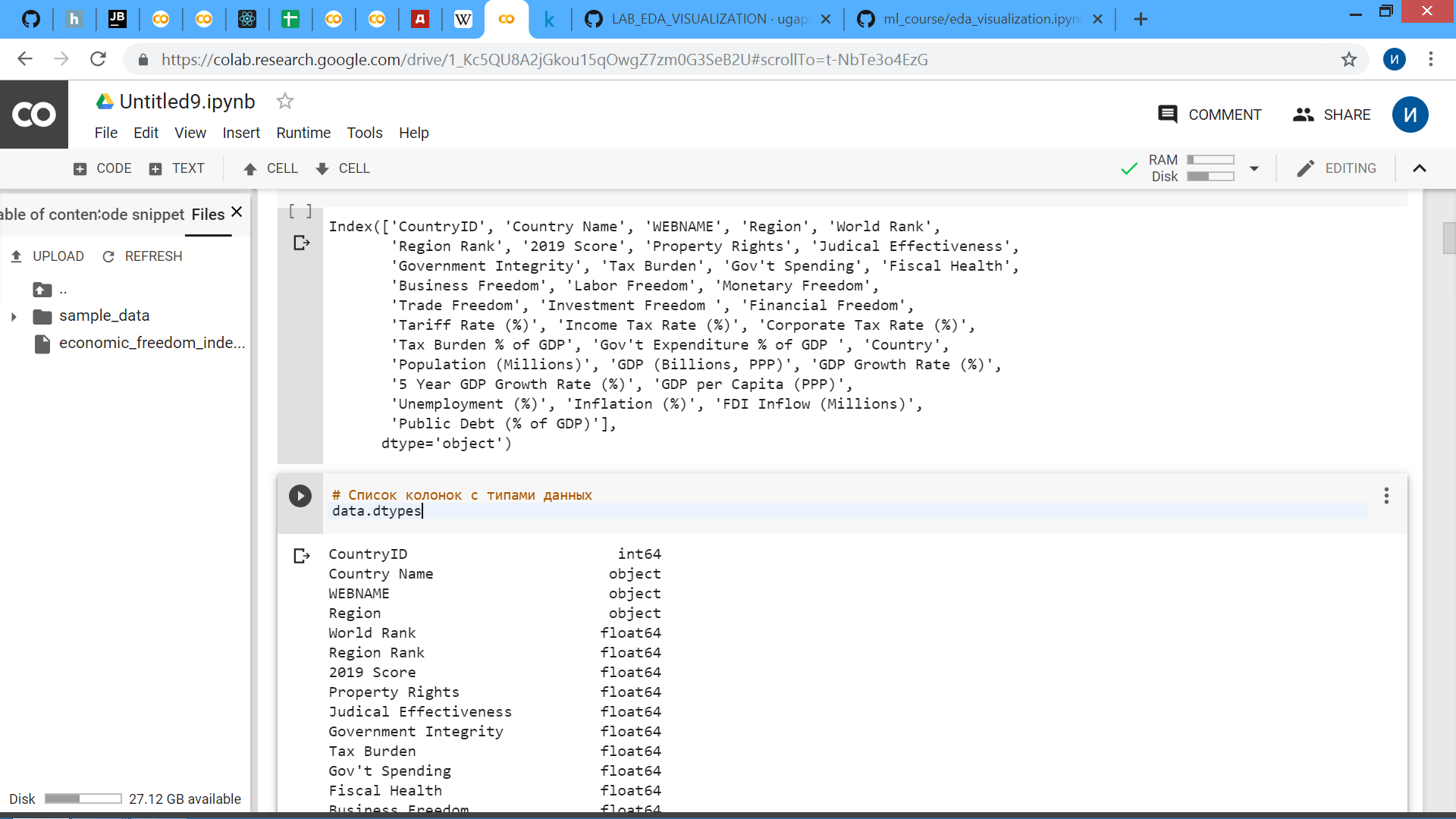
*ax[1].title.set\_text('Kendall')*

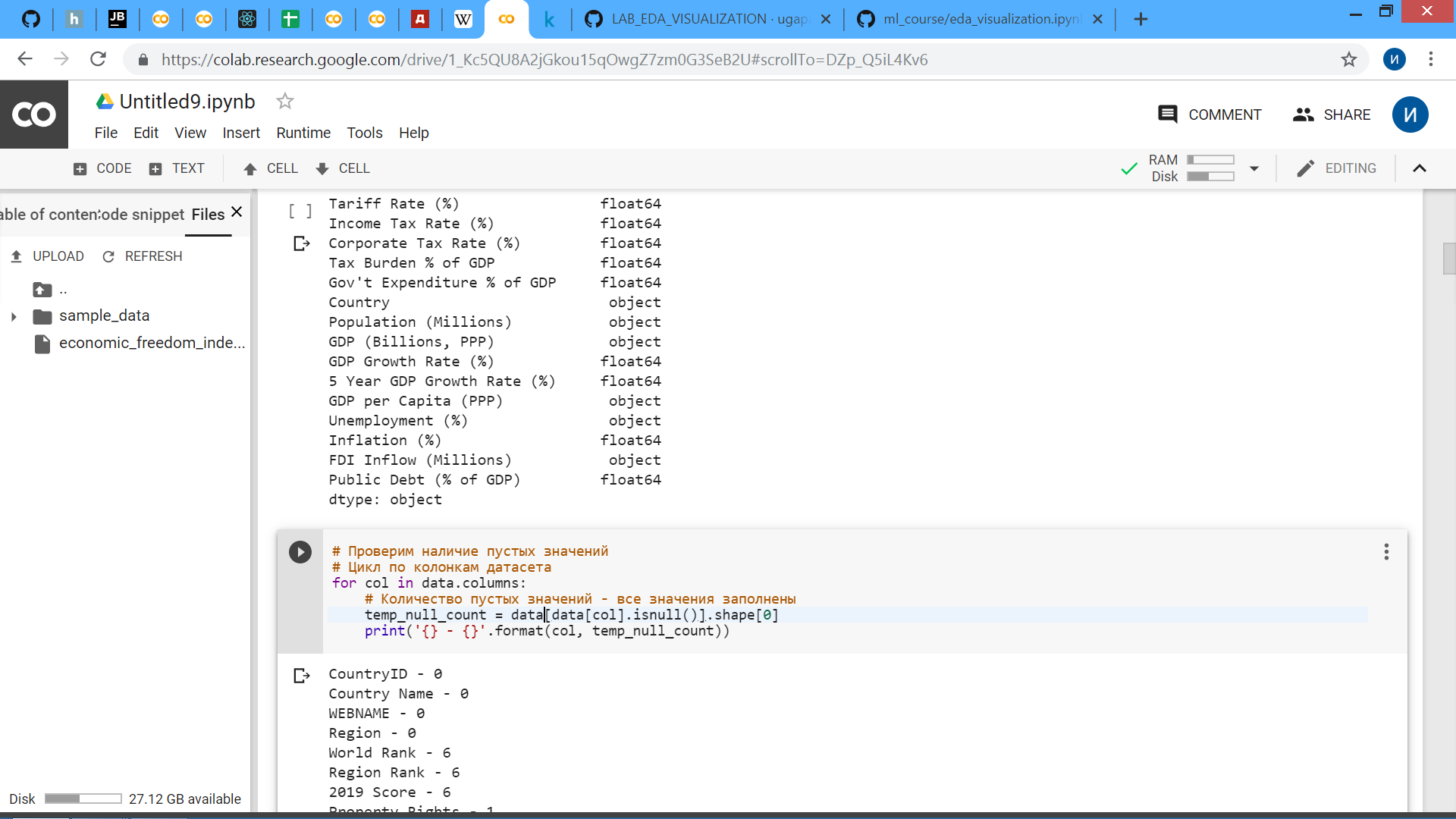
*ax[2].title.set\_text('Spearman')*

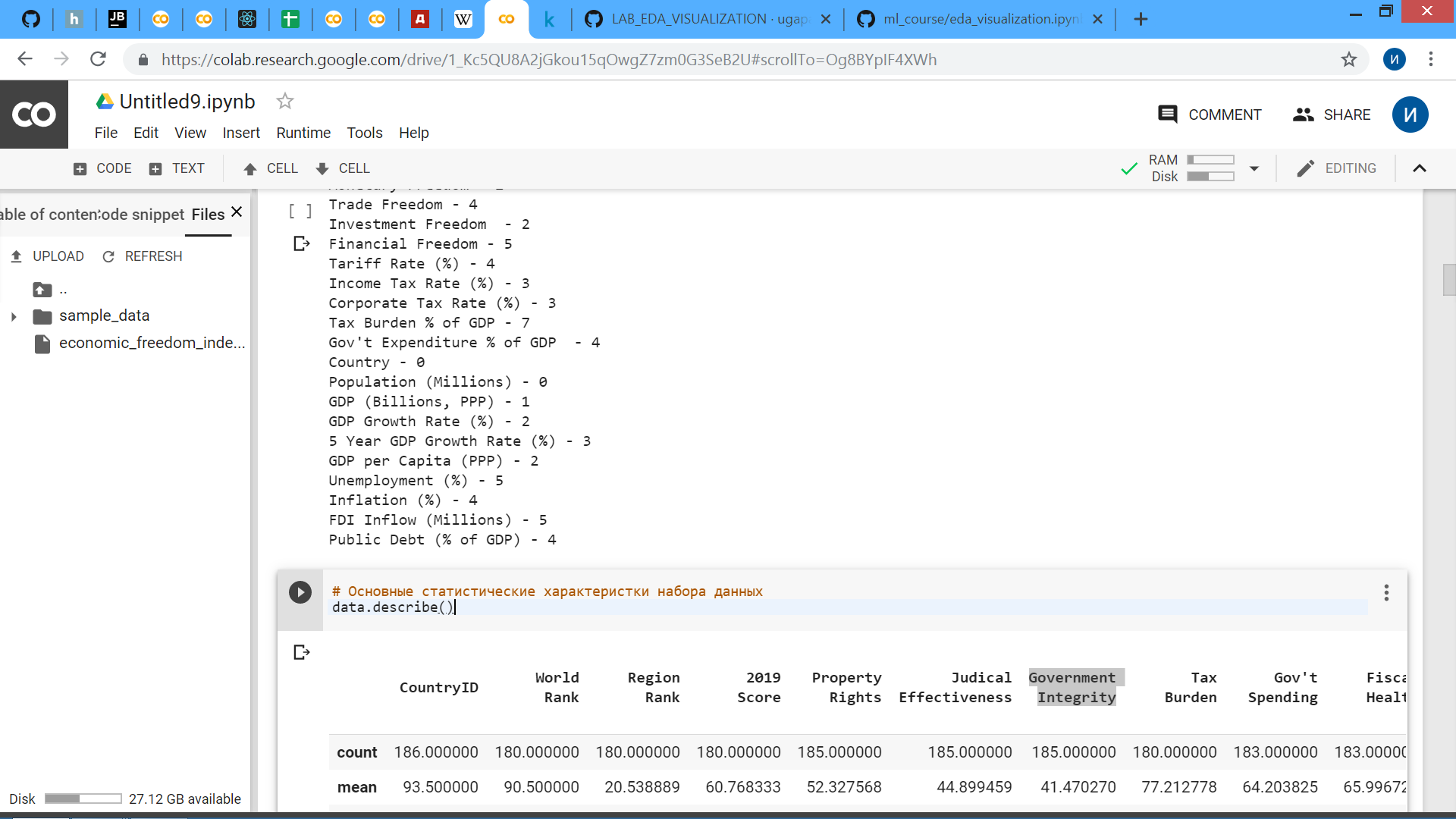
**Экранные формы с примерами выполнения программы**

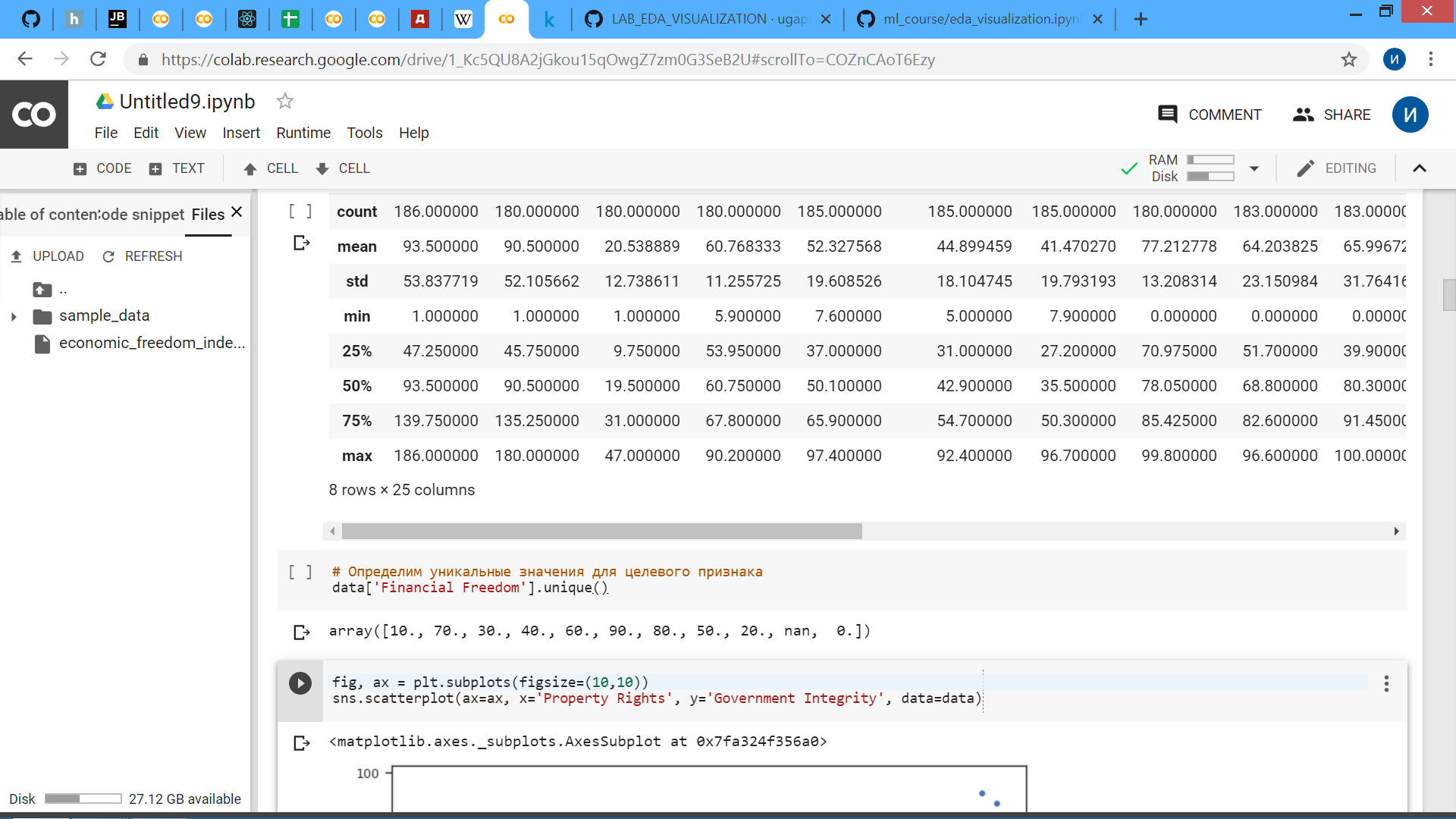


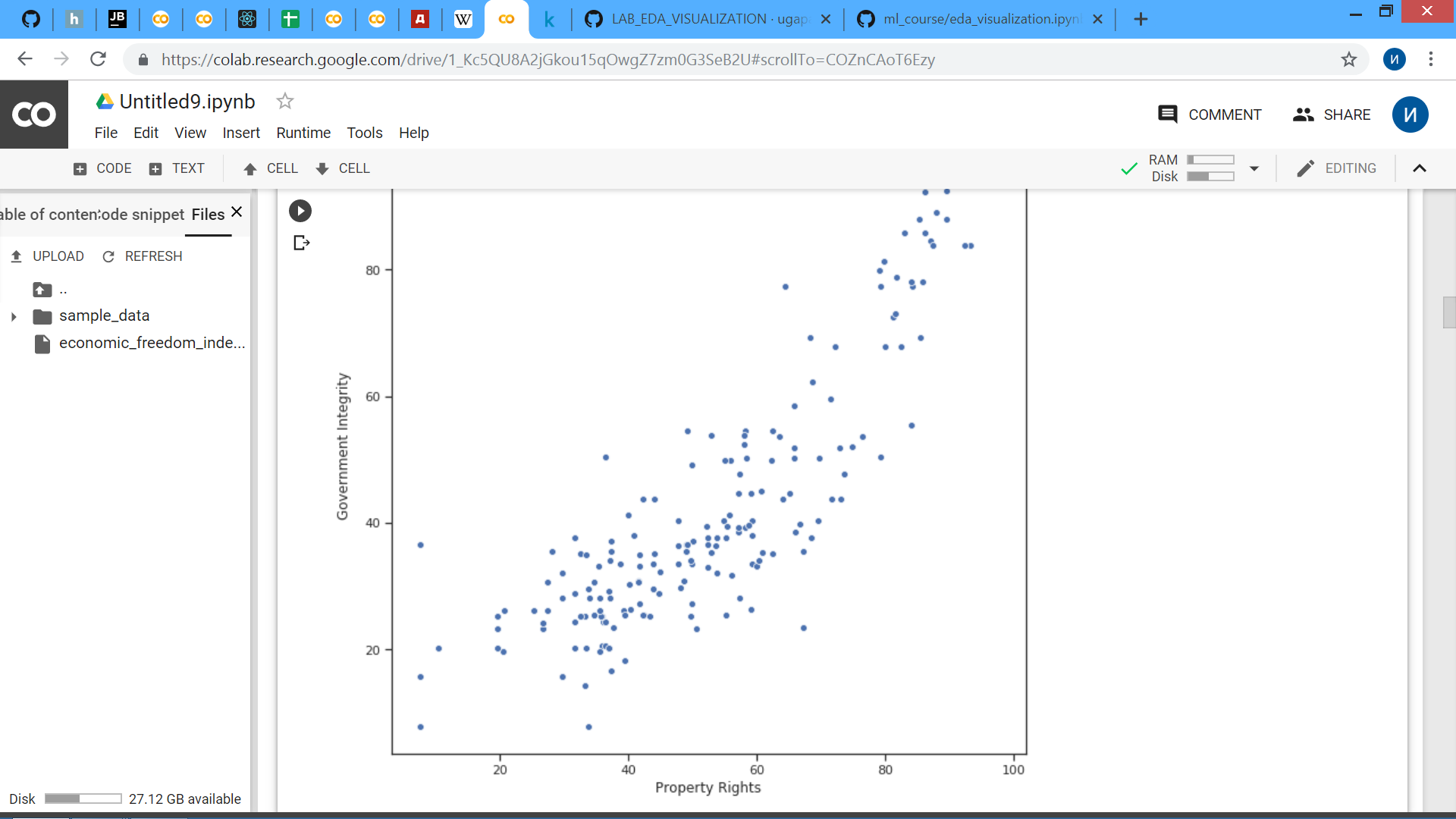


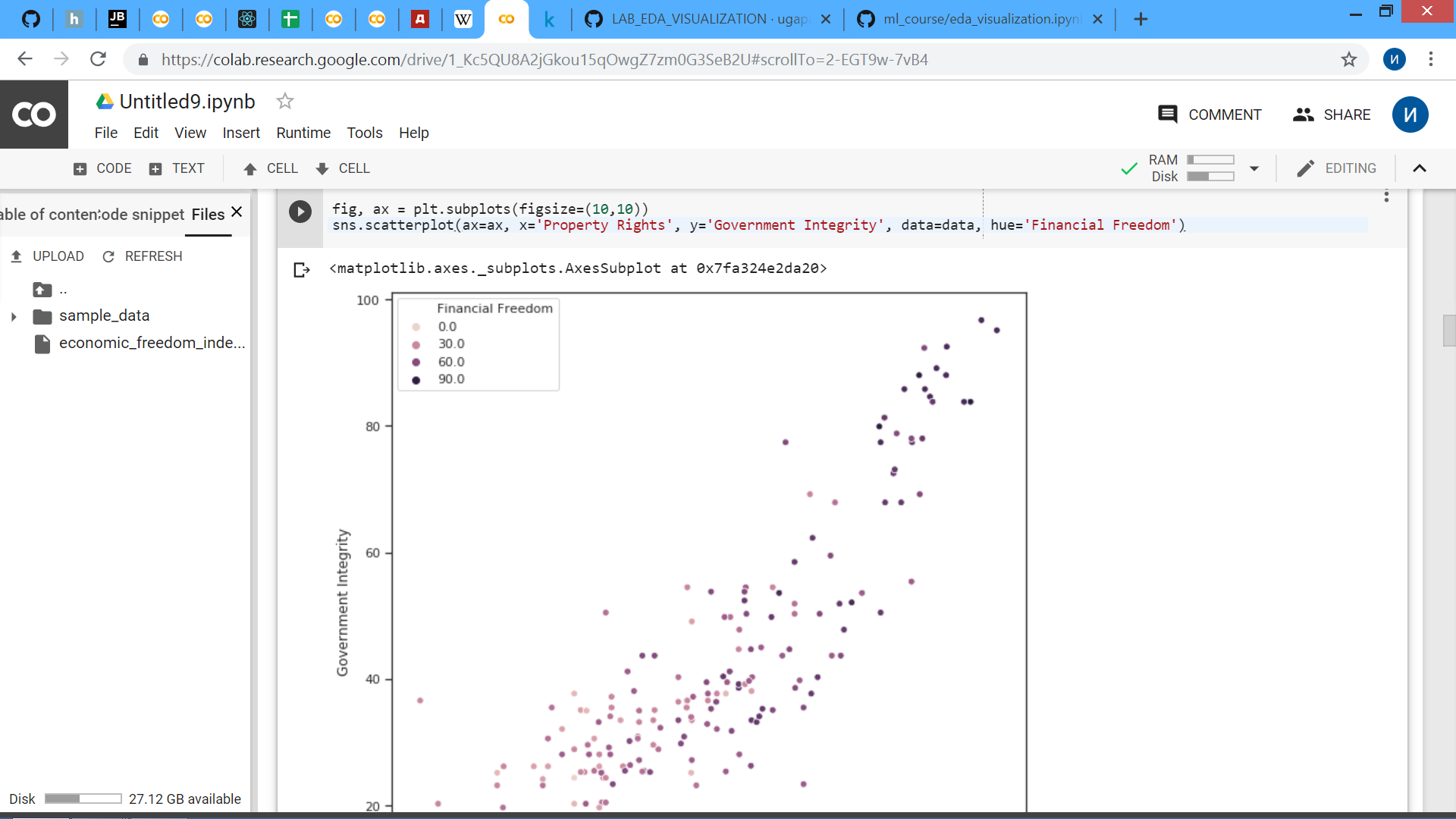


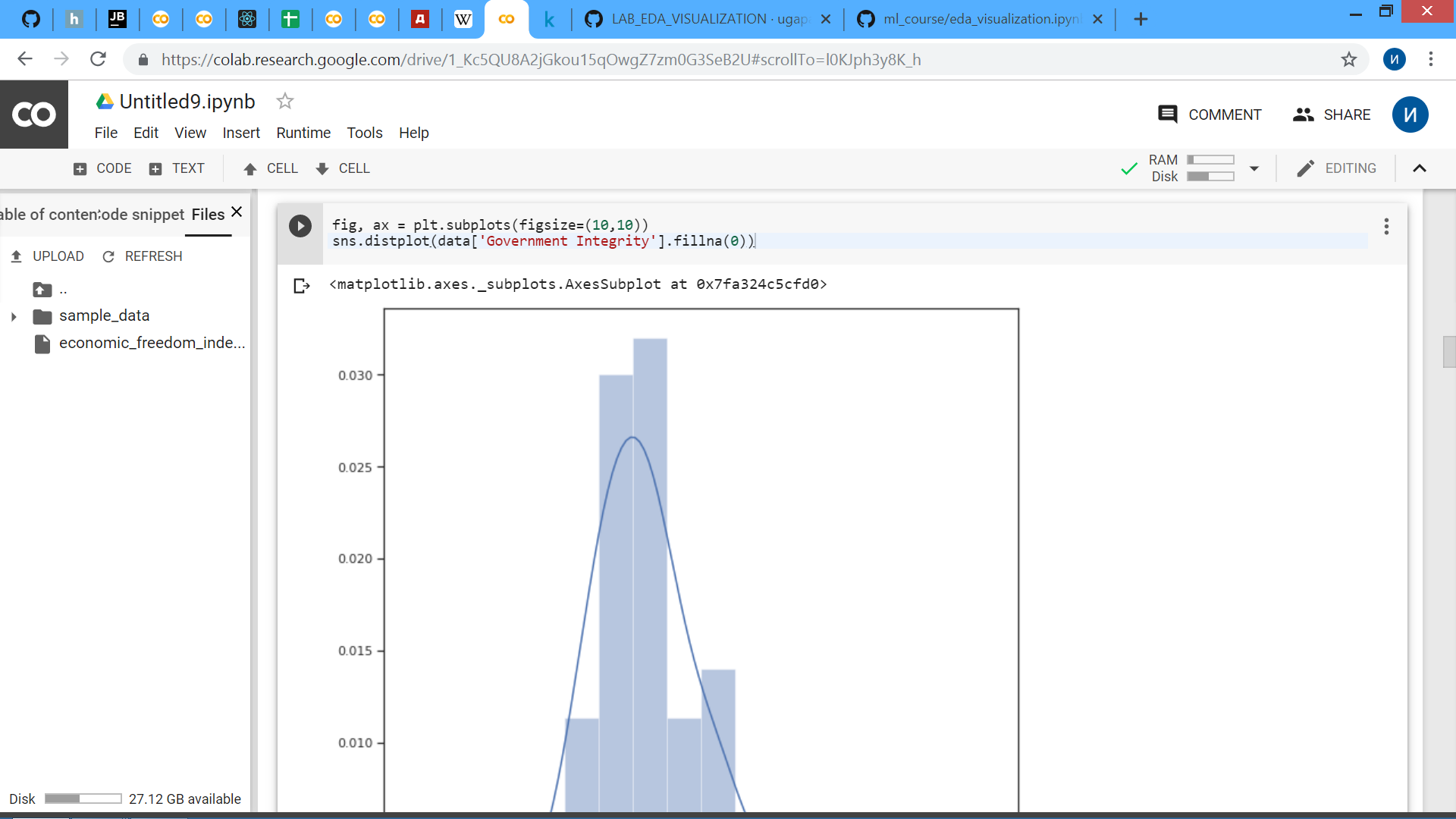


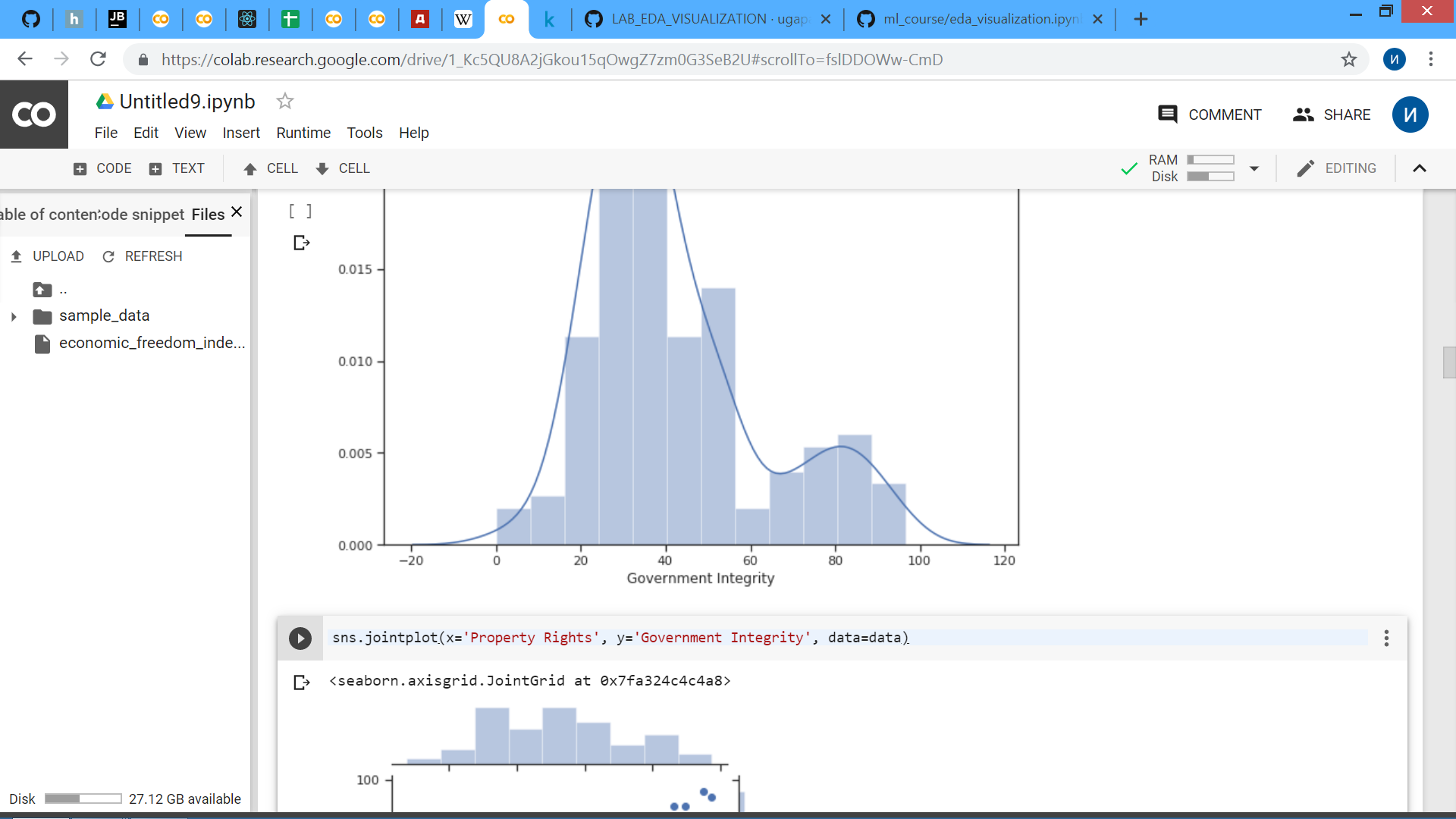


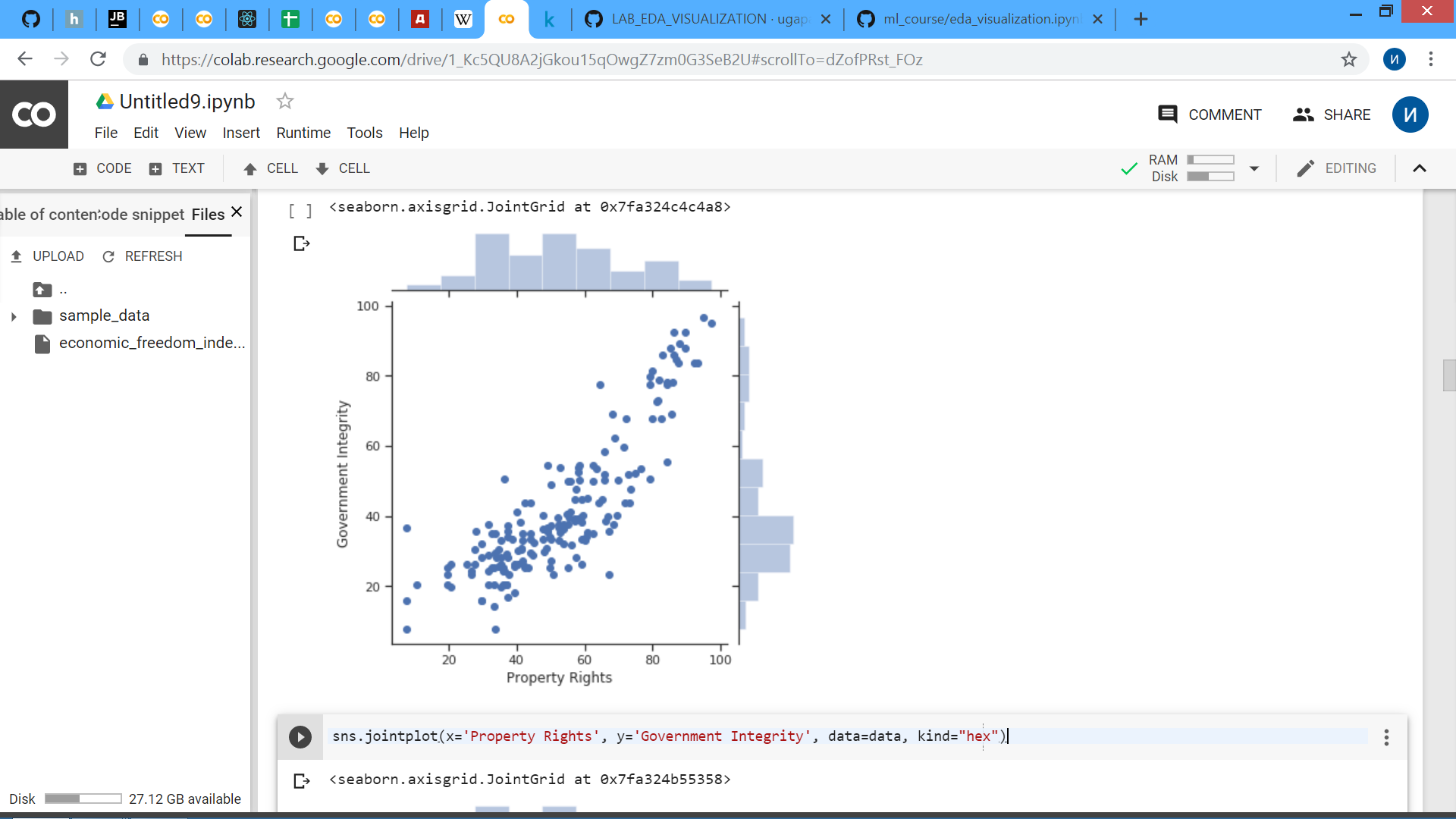


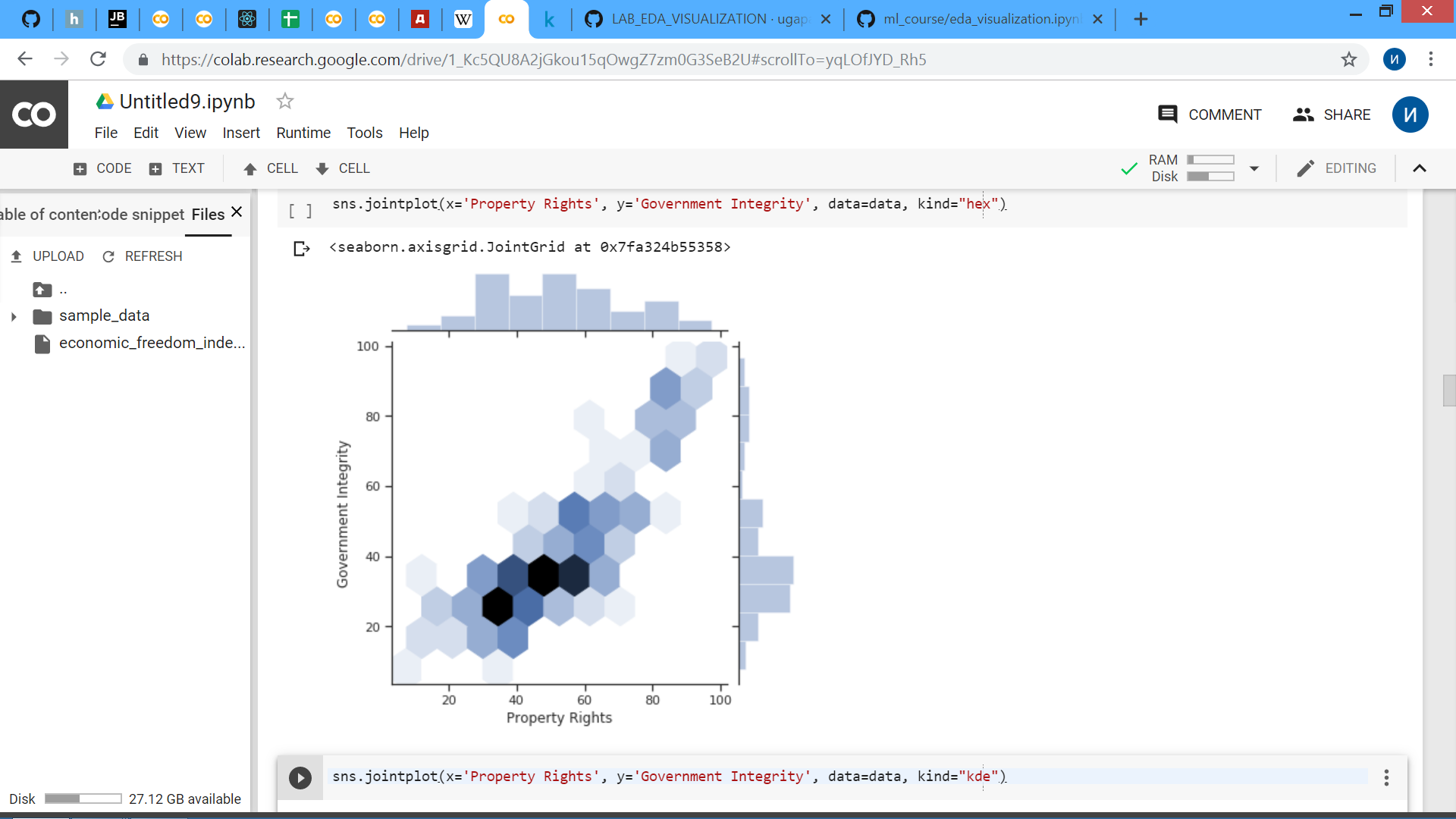


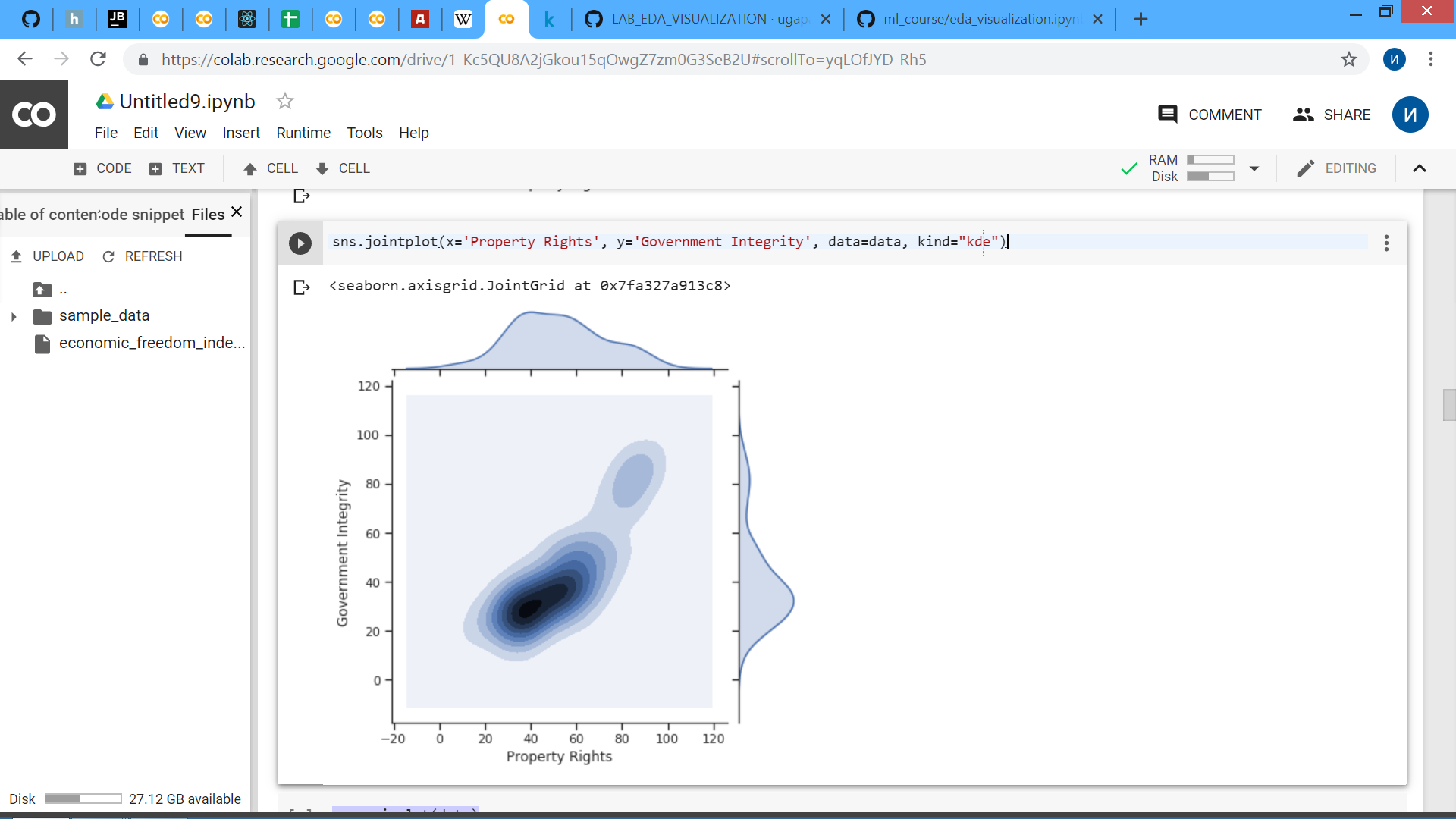


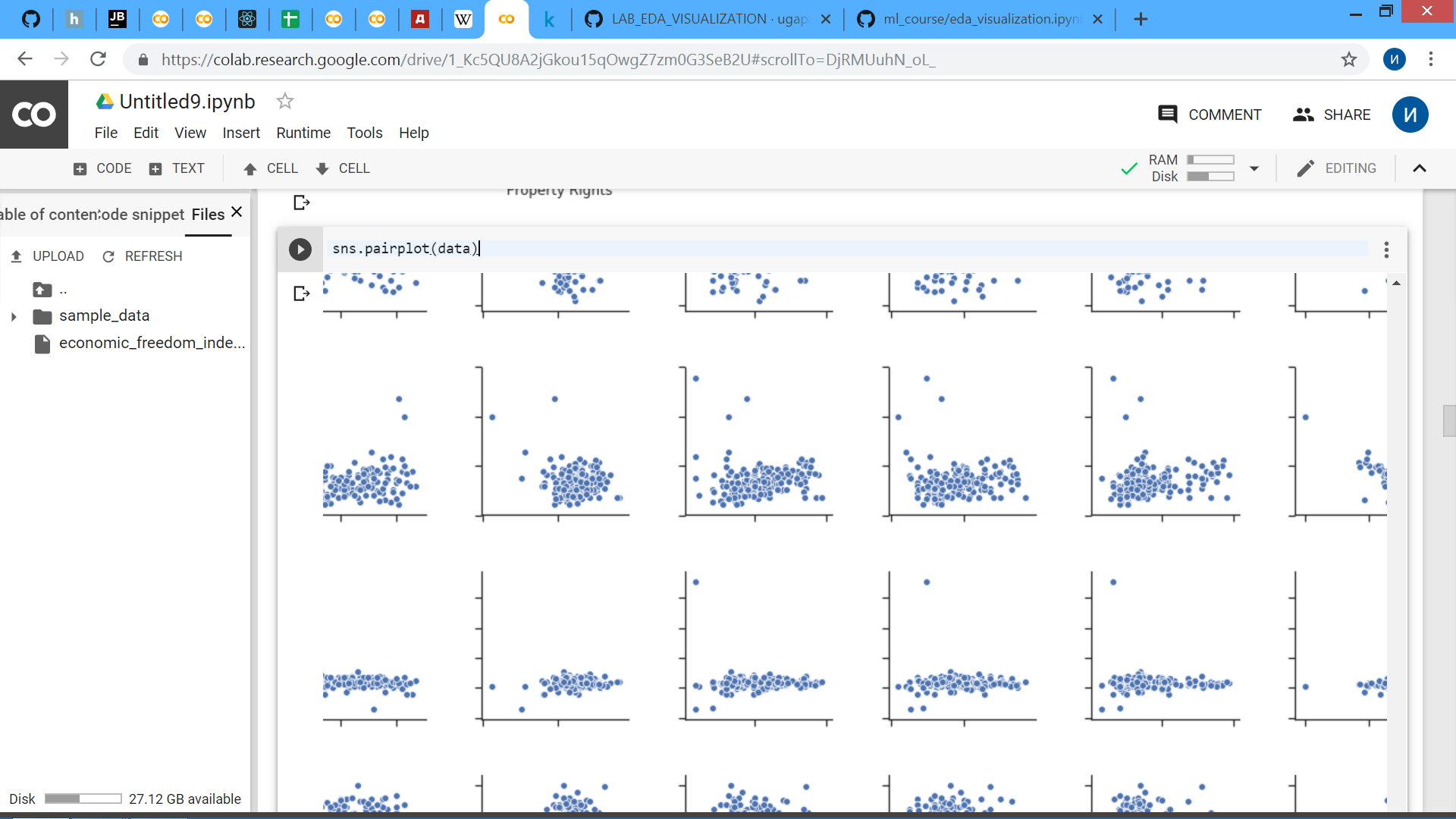




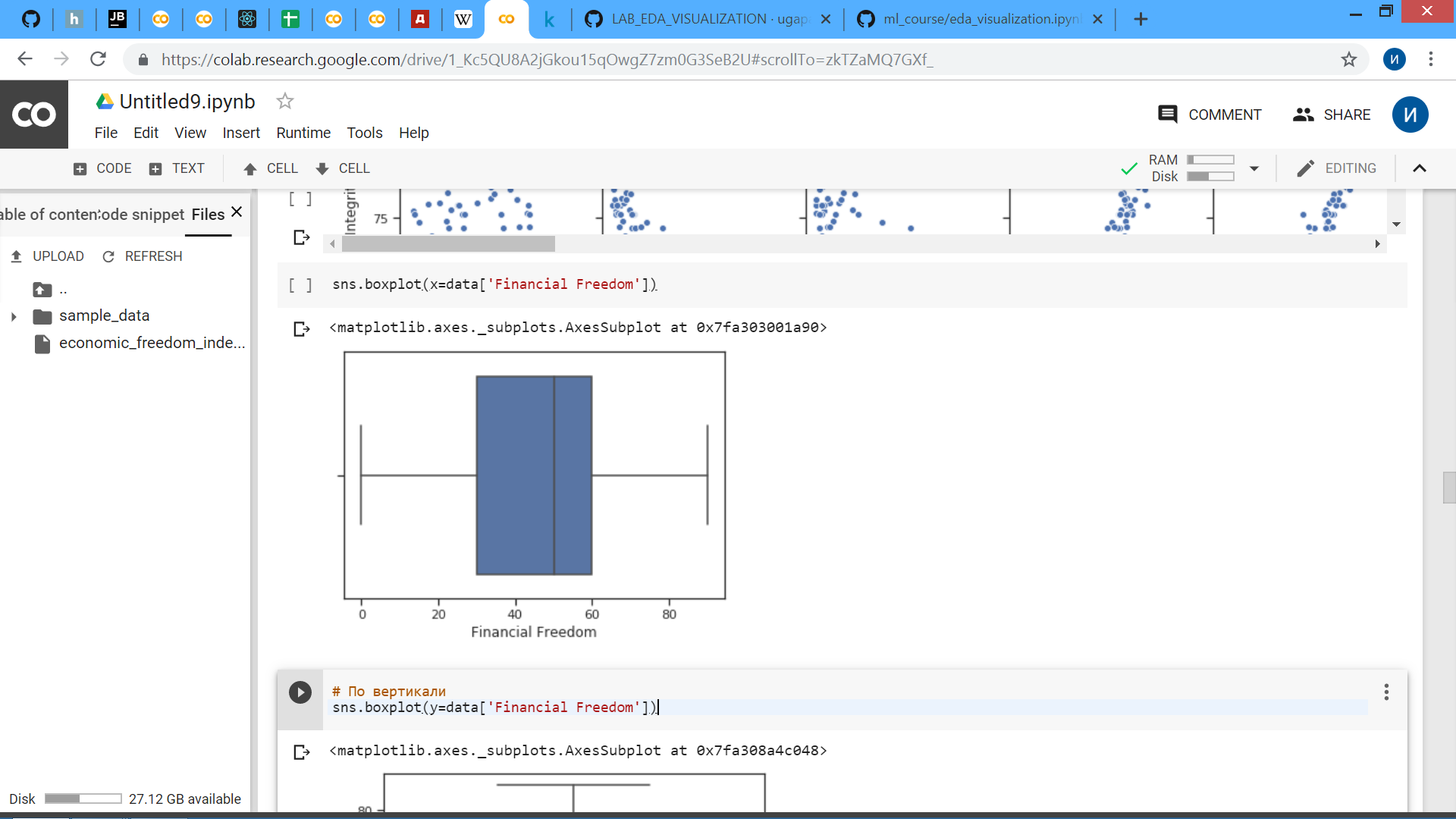


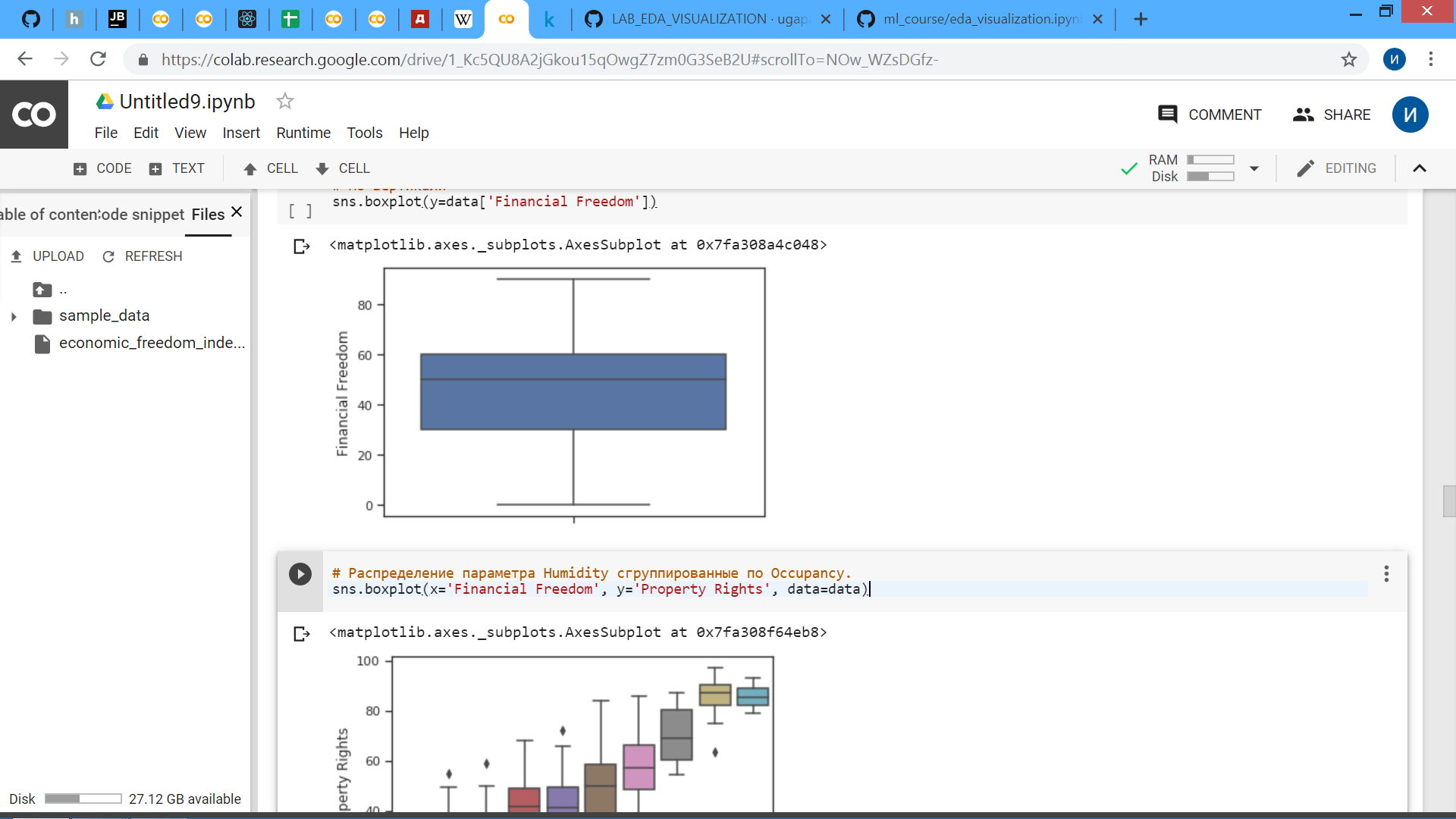


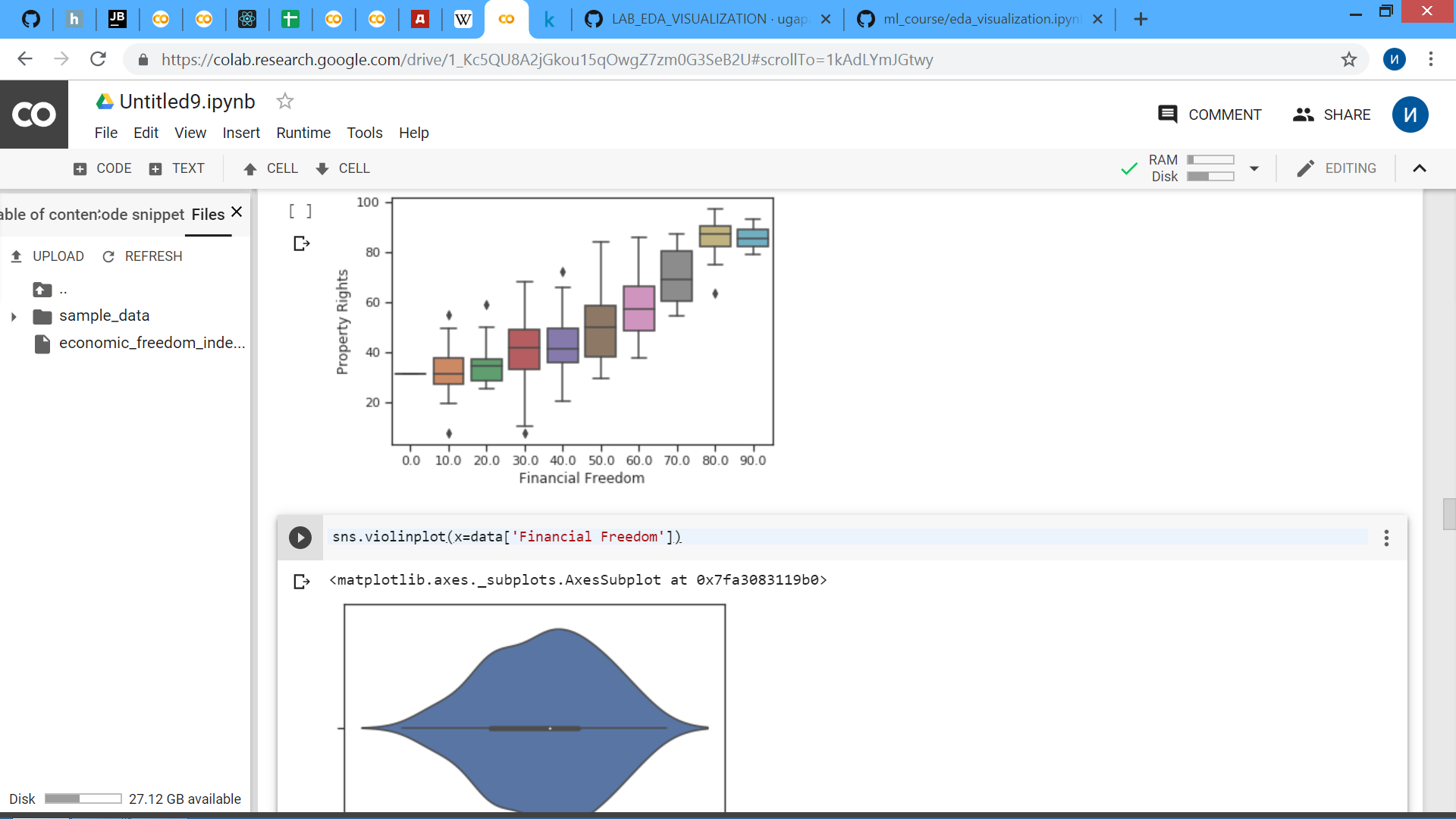


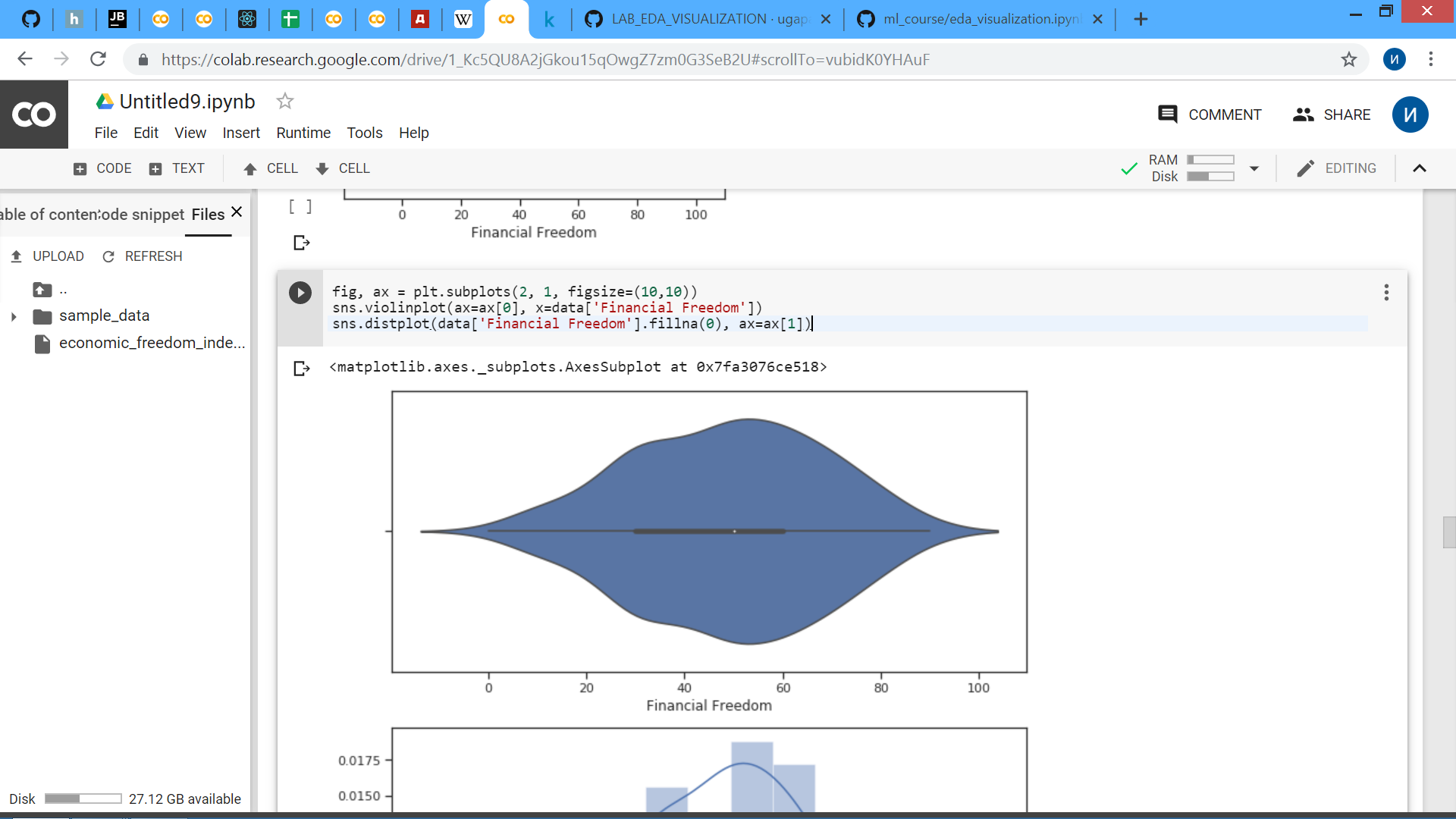


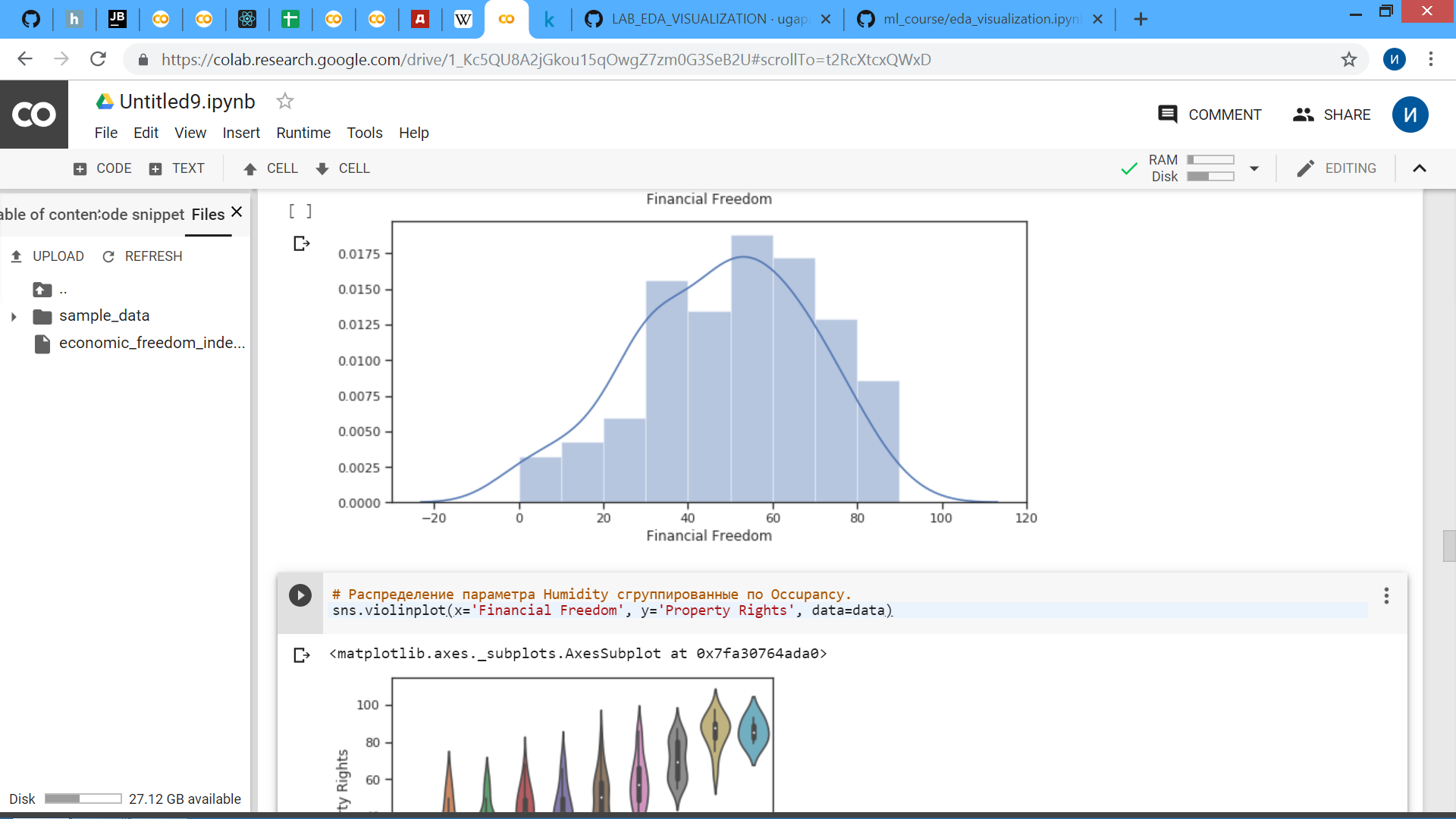


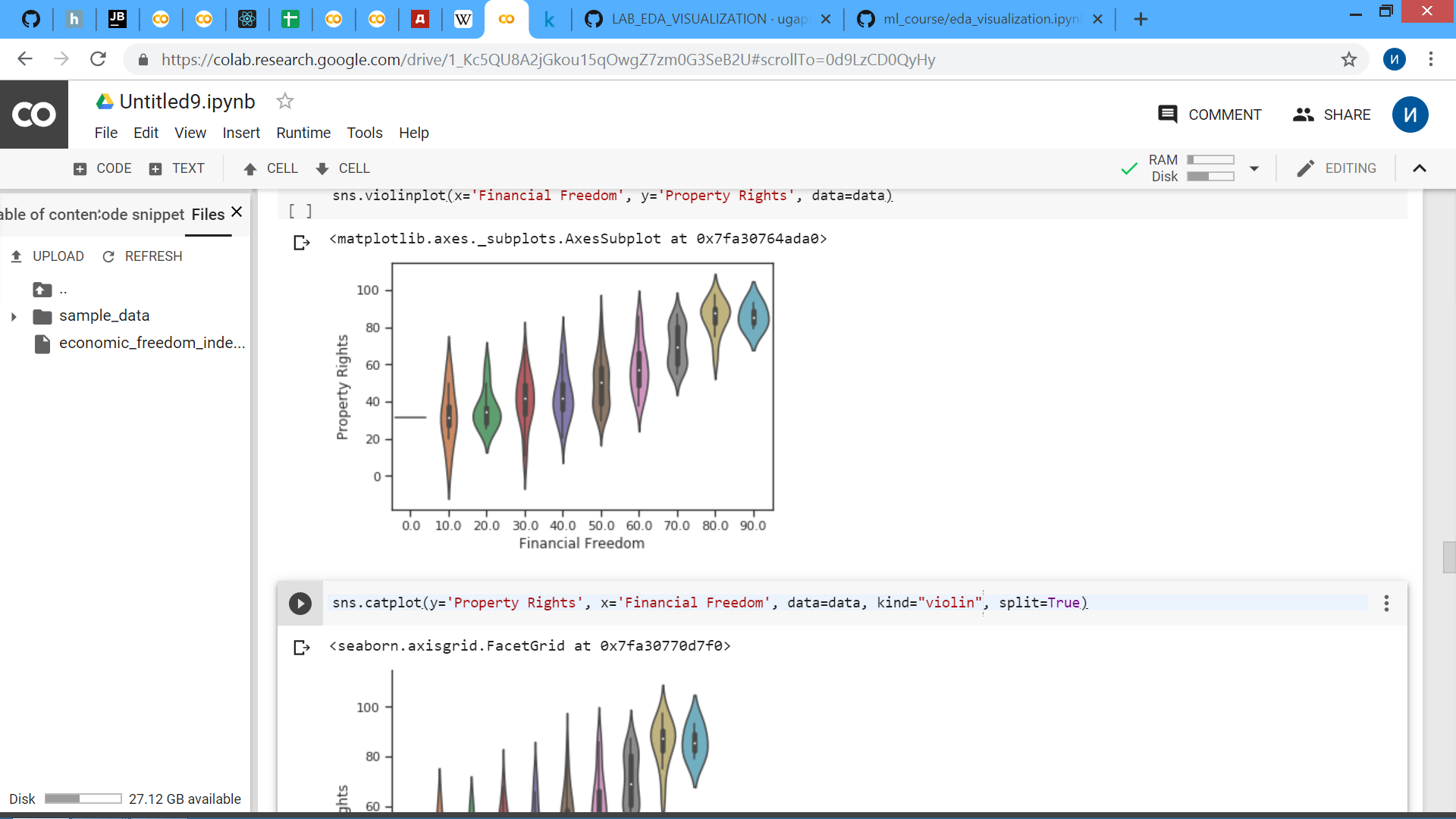


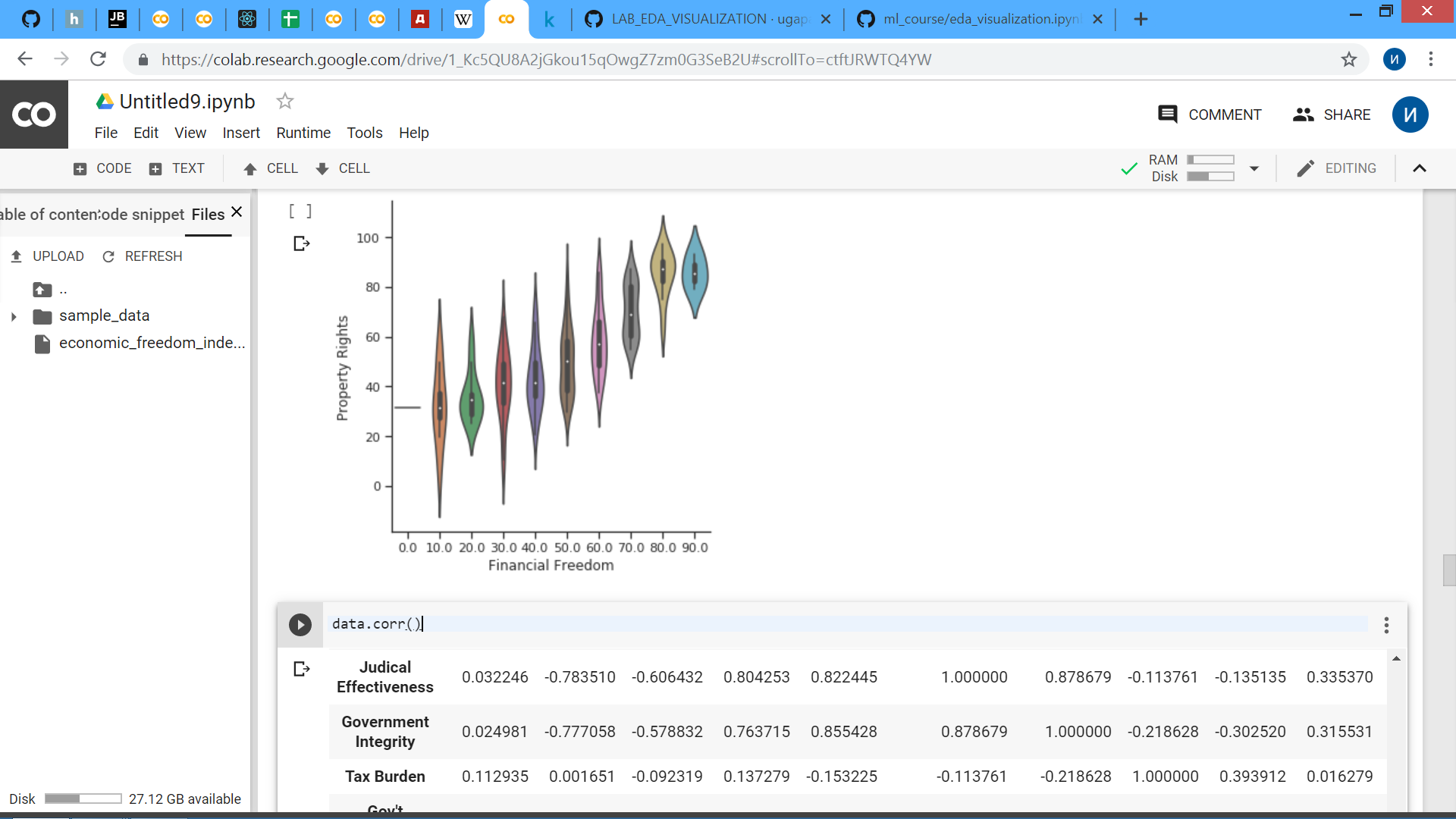


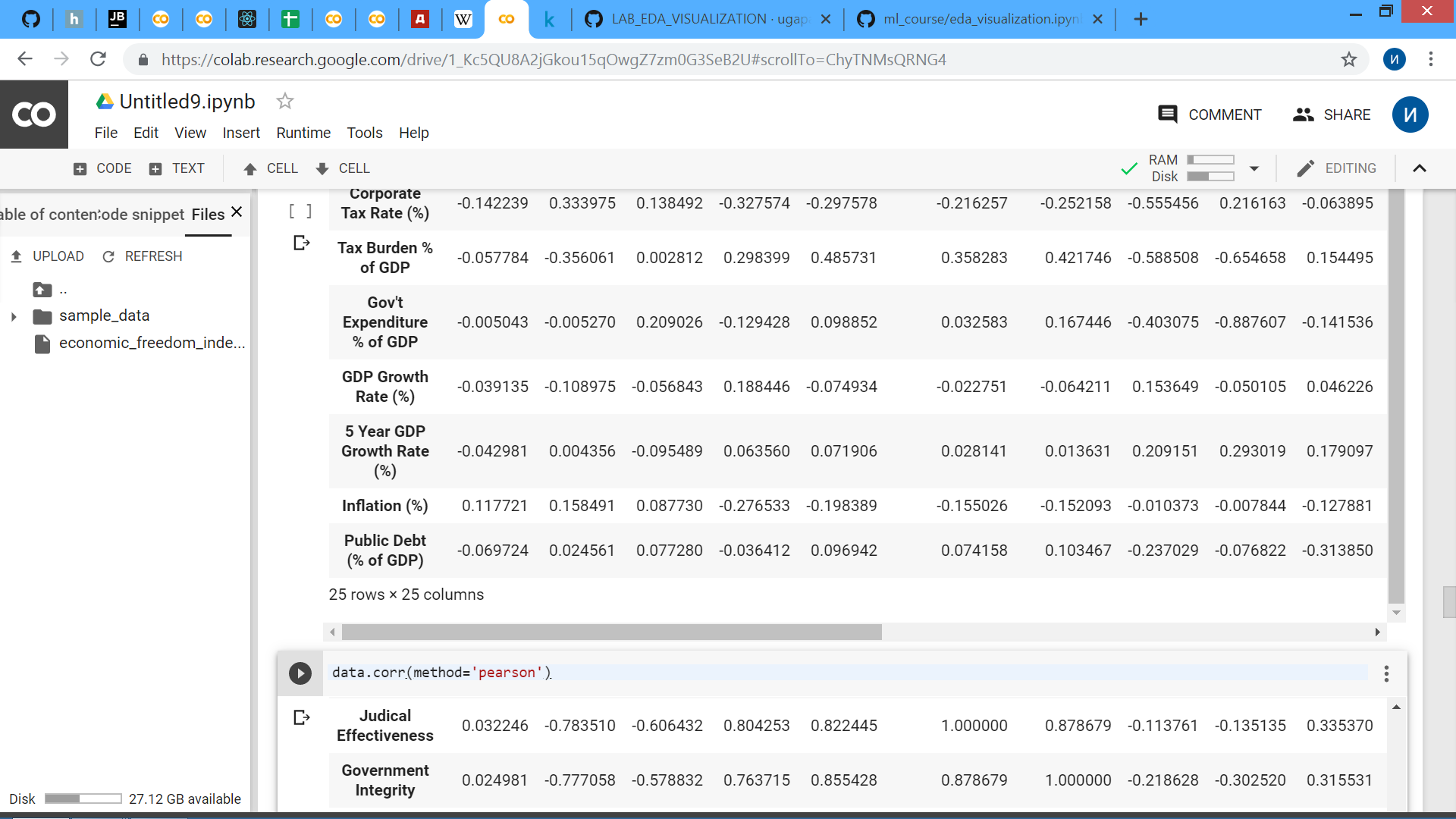


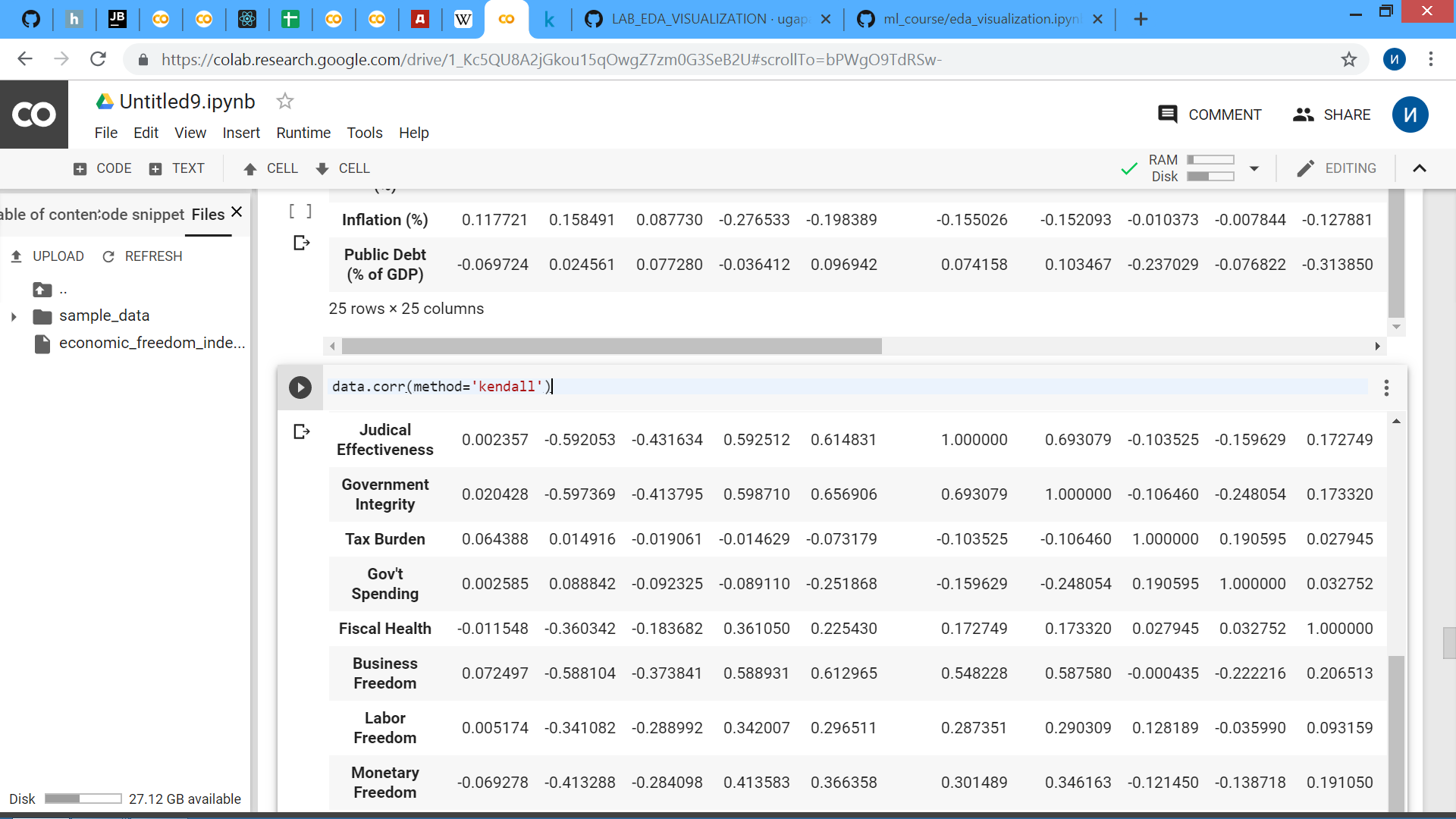




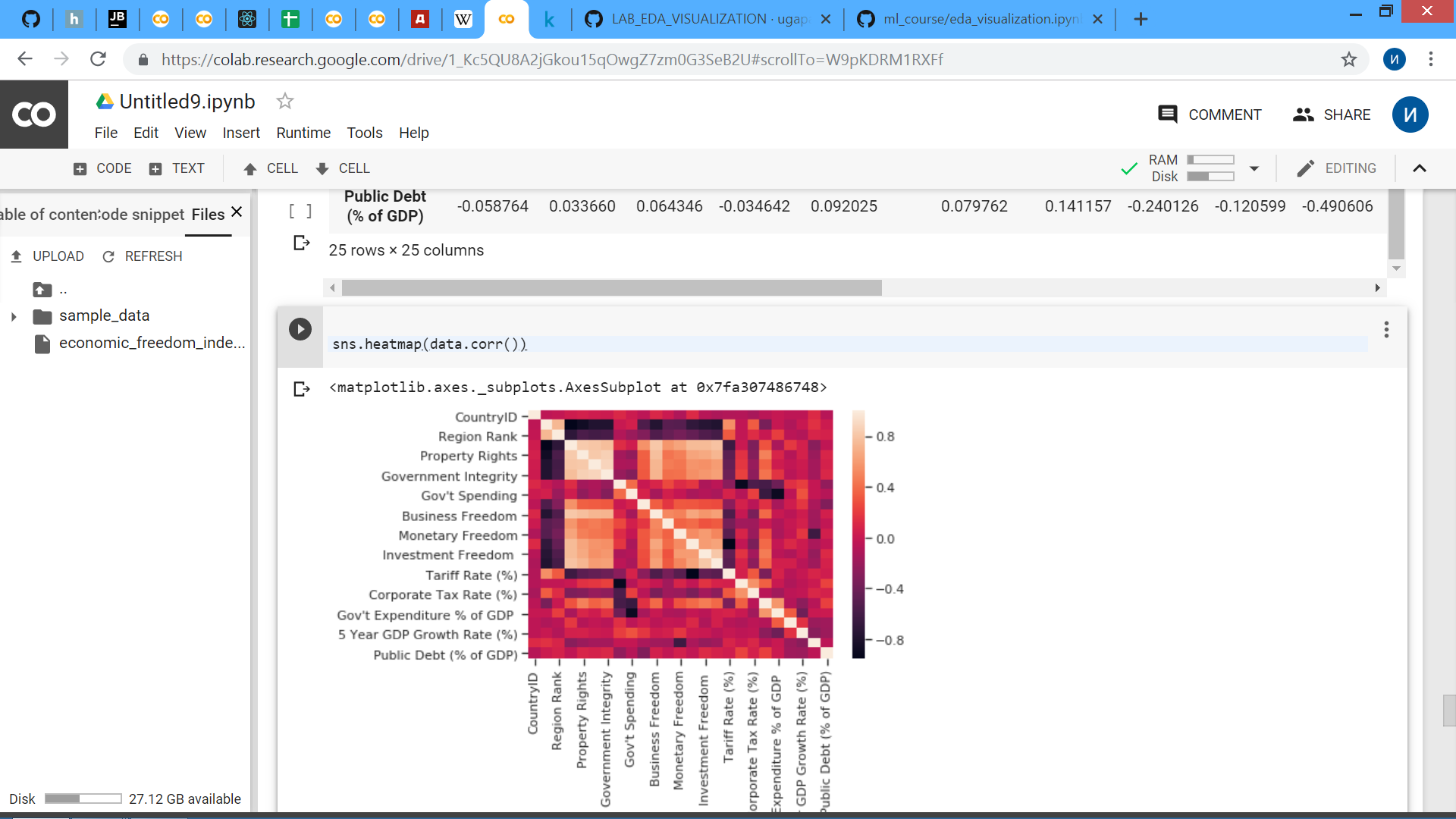


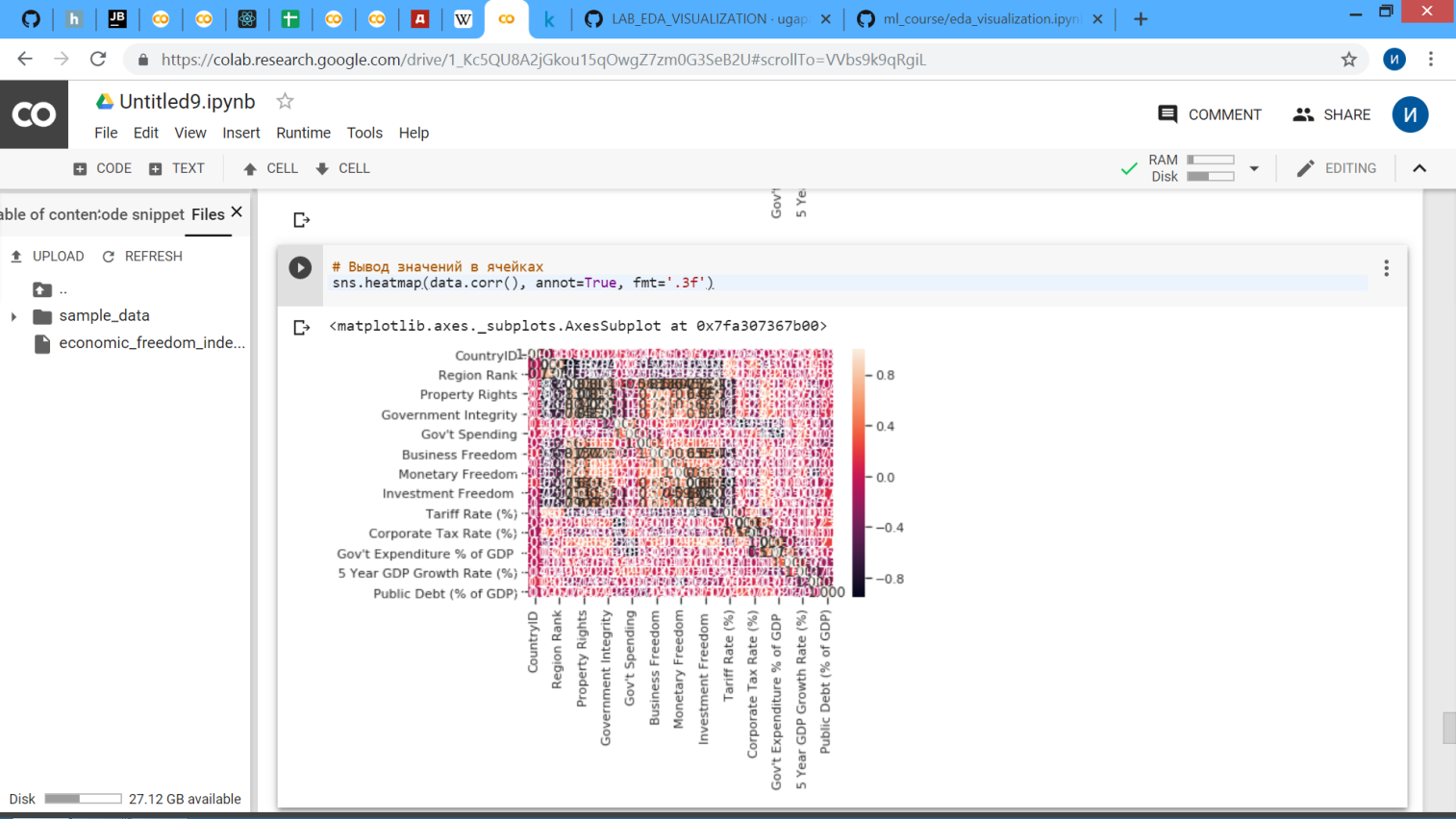


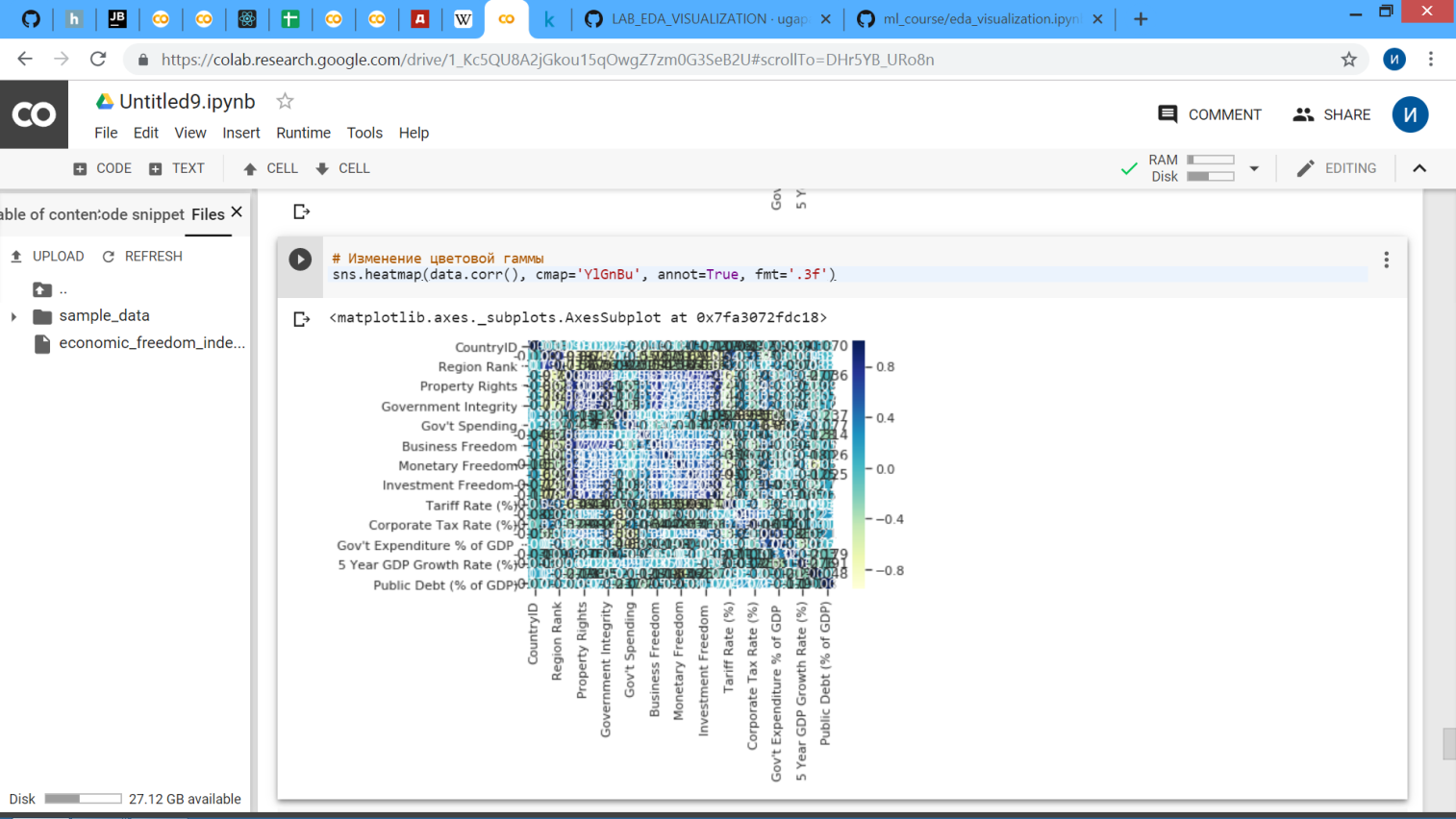


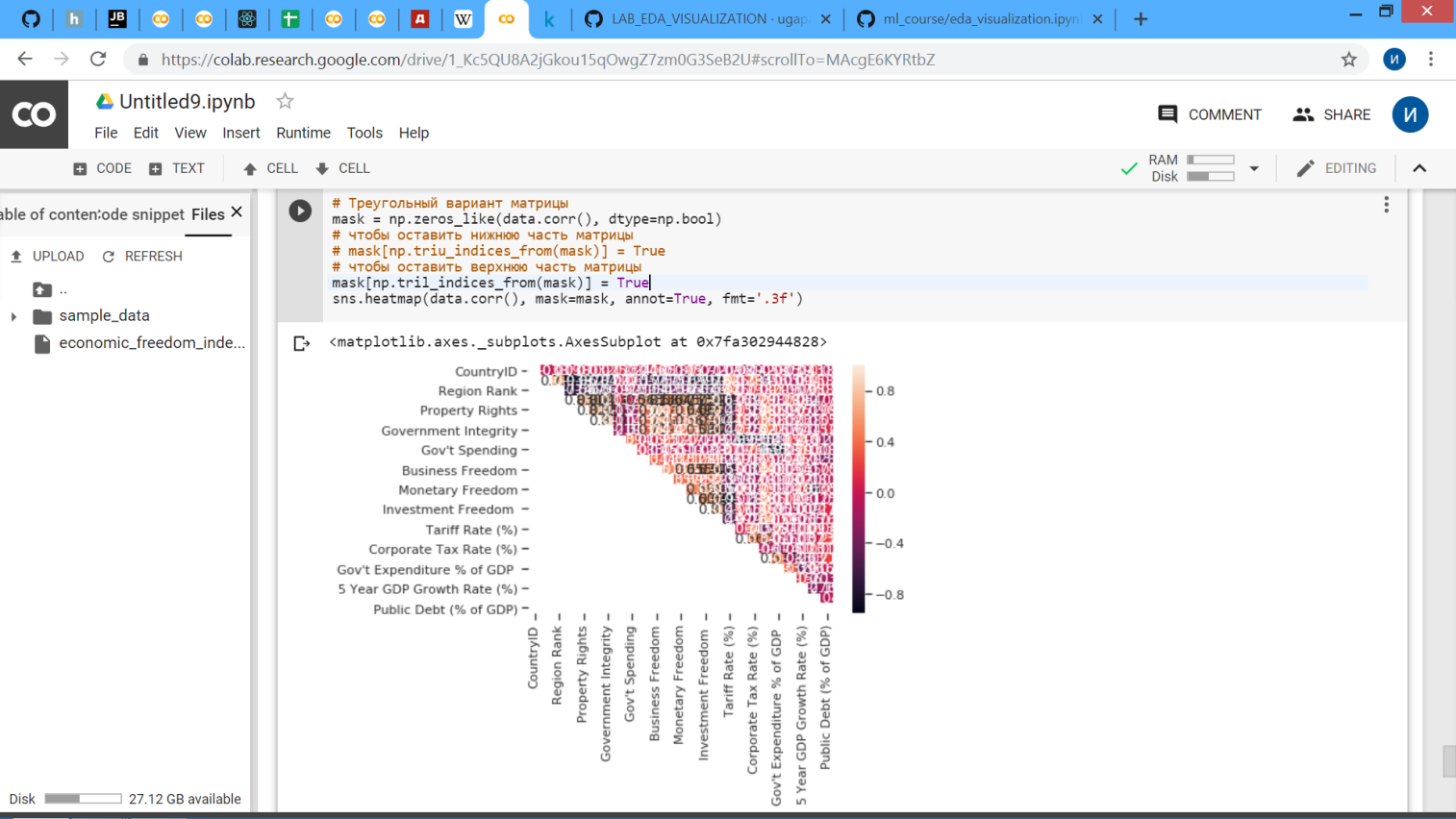


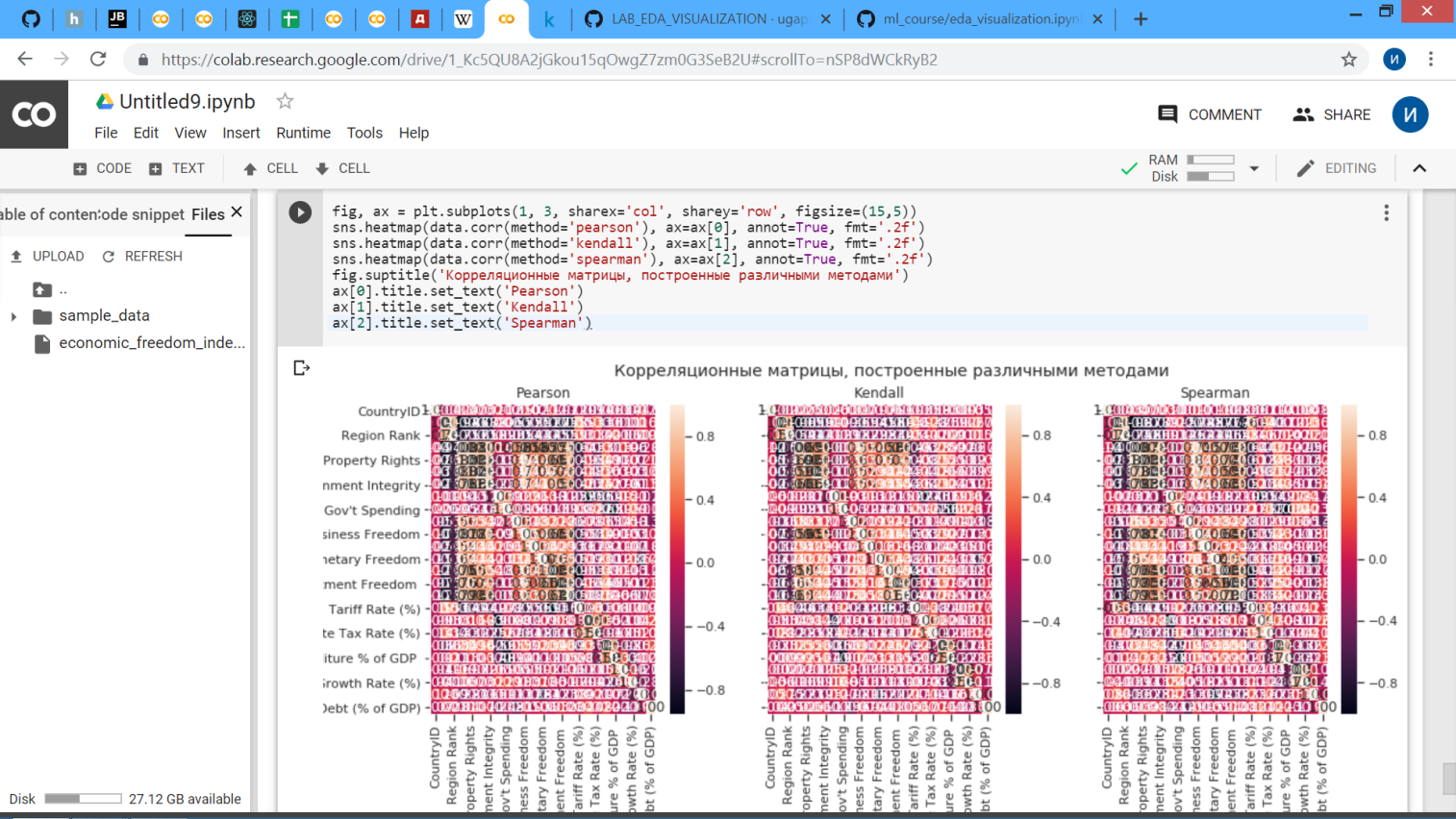












**Список литературы**

[1]. Методические указания. Режим доступа: <https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION> (дата обращения 11.05.2019)

[2]. Dataset «The Economic Freedom Index». Режим доступа: <https://www.kaggle.com/lewisduncan93/the-economic-freedom-index/downloads/the-economic-freedom-index.zip/1> (дата обращения 11.05.2019)

[3].Среда выполнения работы. Режим доступа: [https://colab.research.google.com](https://colab.research.google.com/) (дата обращения 11.05.2019)