Практическая работа №5
Визуализация данных средствами MatplotLib. Диаграммы

Цель занятия: получить навыки использования библиотеки визуализации данных Matplotlib с использованием языка программирования Python.

Пояснения к работе

matplotlib – это основная библиотека для построения научных графиков в Python. Она включает функции для создания высококачественных визуализаций типа линейных диаграмм, гистограмм, диаграмм разброса и т.д. Визуализация данных и различных аспектов вашего анализа может дать важную информацию.

В данной работе взаимодействие с matplotlib будет проходить в Jupyter Notebook на базе Google Colab (см. https://colab.research.google.com/notebooks/intro.ipynb, https://github.com/deepmipt/dlschl/wiki/Инструкция-по-работе-с-Google-Colab).

В среде Jupyter Notebook возможно вывести рисунок прямо в браузере с помощью встроенных команд %matplotlib notebook и %matplotlib inline. Рекомендуется использовать %matplotlib inline.

Использование Google Colab позволяет не устанавливать на свой компьютер Jupyter Notebook.

1. Подготовительная часть.

1.1. Зарегистрировать электронную почту google (либо использовать существующий аккаунт).

1.2. Перейти по ссылке https://colab.research.google.com/notebooks/intro.ipynb

1.3. В правом верхнем углу нажать кнопку «Войти» и затем ввести свои учетные данные google.

1.4. В верхнем левом углу найдите подменю «Файл», далее «Открыть блокнот» и выберите блокнот, который вы создали в рамках П.р. №4.

1.5. Выполните заново все ячейки этого блокнота.

2. Построение Гистограмм.

Гистограмма - это способ представления частотного распределения числового набора данных. Она работает так: ось x разбивается на ячейки, каждая точка данных в наборе данных назначается ячейке, а затем подсчитывается количество точек данных, назначенных каждой ячейке. Таким образом, ось Y - это частота или количество точек данных в каждой ячейке. Обратите внимание, что мы можем изменять размеры, и обычно их нужно настроить, чтобы распределение отображалось красиво.

2.1. Обзор данных

df_can['2013'].head()

```
df_can['2013'].head()
```

```
Country
India                                                33087
China                                                34129
United Kingdom of Great Britain and Northern Ireland     5827
Philippines                                          29544
Pakistan                                             12603
Name: 2013, dtype: int64
```

2.2. Подготовка данных для гистограммы.

 # np.histogram возвращает два значения
count, bin_edges = np.histogram(df_can['2013'])

print(count) # подсчет частоты появления данных
print(bin_edges) # количество столбцов, по умолчанию – 10

```
# np.histogram returns 2 values
count, bin_edges = np.histogram(df_can['2013'])

print(count) # frequency count
print(bin_edges) # bin ranges, default = 10 bins
```

```
[178  11   1   2   0   0   0   0   1   2]
[    0.   3412.9  6825.8 10238.7 13651.6 17064.5 20477.4 23890.3 27303.2
 30716.1 34129. ]
```

2.3. Построение гистограммы:

df_can['2013'].plot(kind='hist', figsize=(8, 5))

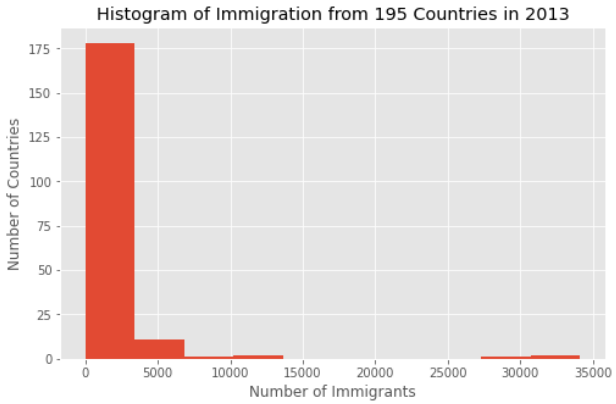
plt.title('Histogram of Immigration from 195 Countries in 2013') # добавление названия
plt.ylabel('Number of Countries') # добавление наименования оси y
plt.xlabel('Number of Immigrants') # наименование оси x

plt.show()

```
df_can['2013'].plot(kind='hist', figsize=(8, 5))

plt.title('Histogram of Immigration from 195 Countries in 2013') # add a title to the histogram
plt.ylabel('Number of Countries') # add y-label
plt.xlabel('Number of Immigrants') # add x-label

plt.show()
```



3. Построение Bar Charts (Dataframe):

Bar Charts – это способ представления данных, где длина полосок представляет величину / размер функции / переменной. Bar Charts обычно представляют числовые и категориальные переменные, сгруппированные по интервалам.

3.1. Извлекаем часть данных из df_can:

 # step 1: get the data
df_iceland = df_can.loc['Iceland', years]
df_iceland.head()

```
# step 1: get the data
df_iceland = df_can.loc['Iceland', years]
df_iceland.head()
```

```
1980    17
1981    33
1982    10
1983     9
1984    13
Name: Iceland, dtype: object
```

3.2. Построение графика (горизонтальный Bar Chart):

 # step 2: plot data
df_iceland.plot(kind='barh', figsize=(10, 6))

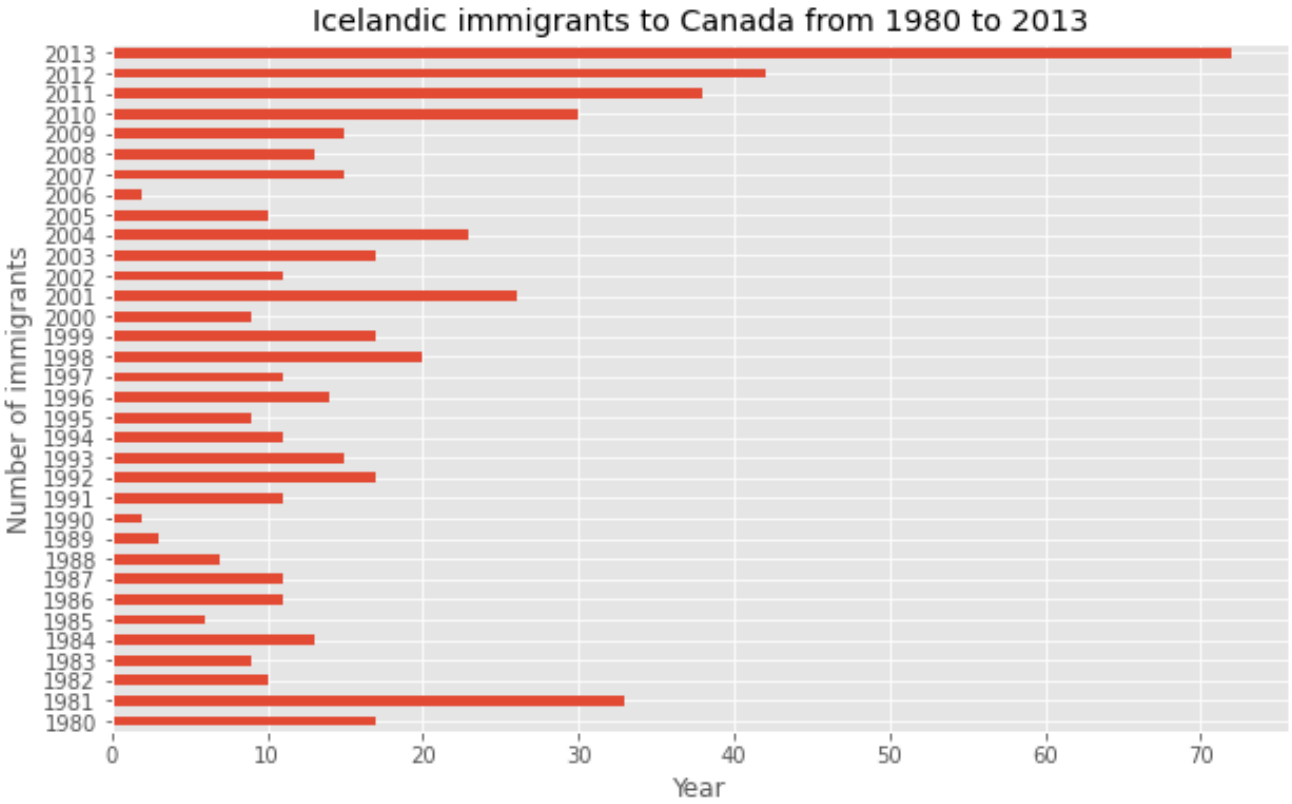
plt.xlabel('Year') # add to x-label to the plot
plt.ylabel('Number of immigrants') # add y-label to the plot
plt.title('Icelandic immigrants to Canada from 1980 to 2013') # add title to the plot

plt.show()

```
# step 2: plot data
df_iceland.plot(kind='barh', figsize=(10, 6))

plt.xlabel('Year') # add to x-label to the plot
plt.ylabel('Number of immigrants') # add y-label to the plot
plt.title('Icelandic immigrants to Canada from 1980 to 2013') # add title to the plot
```

```
plt.show()
```



КОНТРОЛЬНЫЕ ВОПРОСЫ

1. В п. 3.2 попробуйте изменить *kind='barh' на kind='bar'*, что получится?