**Домашнее задание № 2 .**

**Визуальный анализ данных о публикациях на Хабрахабре**

import pandas as pd

%matplotlib inline

import matplotlib.pyplot as plt

import seaborn as sns

**Загрузка и знакомство с данными**

Для работы вам понадобятся предобработанные данные учебного конкурса на kaggle [«Прогноз популярности статьи на Хабре»](https://www.kaggle.com/c/howpop-habrahabr-favs). Скачайте [данные](https://drive.google.com/file/d/1nV2qV9otN3LnVSDqy95hvpJdb6aWtATk/view?usp=sharing) соревнования (данные были удалены с Kaggle ради организации последующего идентичного соревнования, так что тут ссылка на Google Drive).

# при необходимости поменяйте путь к данным

df = pd.read\_csv("howpop\_train.csv")

df.shape

df.head(3).T

Избавимся сразу от переменных, названия которых заканчиваются на \_lognorm (нужны для соревнования на Kaggle). Выберем их с помощью filter() и удалим drop-ом:

df.drop(

filter(lambda c: c.endswith("\_lognorm"), df.columns),

axis=1, # axis = 1: столбцы

inplace=True,

) # избавляет от необходимости сохранять датасет

df.describe().T

df.describe(include=["object", "bool"]).T # бинарные и категориальные переменные

# настройка внешнего вида графиков в seaborn

sns.set\_style("dark")

sns.set\_palette("RdBu")

sns.set\_context(

"notebook", font\_scale=1.5, rc={"figure.figsize": (15, 5), "axes.titlesize": 18}

)

Столбец **published** (время публикации) содержит строки. Чтобы мы могли работать с этими данными как с датой/временем публикации, приведём их к типу datetime:

print(df.published.dtype)

df["published"] = pd.to\_datetime(df.published, yearfirst=True)

print(df.published.dtype)

Создадим несколько столбцов на основе данных о времени публикации:

df["year"] = [d.year for d in df.published]

df["month"] = [d.month for d in df.published]

df["dayofweek"] = [d.isoweekday() for d in df.published]

df["hour"] = [d.hour for d in df.published]

Теперь Ваша очередь. В каждом пункте предлагается построить картинку и с ее помощью ответить на вопрос в [форме](https://docs.google.com/forms/d/e/1FAIpQLSf3b5OG8zX_nLQBQ-t20c6M5Auz-VUL-yxj8Fm9_o_XWDBTrg/viewform?c=0&w=1). Конечно, можно попытаться ответить на все вопросы только с Pandas, без картинок, но советую Вам потренироваться строить (красивые) визуализации.

**1. В каком месяце (и какого года) было больше всего публикаций?**

* март 2016
* март 2015
* апрель 2015
* апрель 2016

Ваш Код:

**2. Проанализируйте публикации в месяце из предыдущего вопроса**

Выберите один или несколько вариантов:

* Один или несколько дней сильно выделяются из общей картины
* На хабре *всегда* больше статей, чем на гиктаймсе
* По субботам на гиктаймс и на хабрахабр публикуют примерно одинаковое число статей

Подсказки: постройте график зависимости числа публикаций от дня; используйте параметр hue; не заморачивайтесь сильно с ответами и не ищите скрытого смысла :)

Ваш Код:

**3. Когда лучше всего публиковать статью?**

* Больше всего просмотров набирают статьи, опубликованные в 12 часов дня
* У опубликованных в 10 утра постов больше всего комментариев
* Больше всего просмотров набирают статьи, опубликованные в 6 часов утра
* Максимальное число комментариев на гиктаймсе набрала статья, опубликованная в 9 часов вечера
* На хабре дневные статьи комментируют чаще, чем вечерние

Ваш Код:

**4. Кого из топ-20 авторов чаще всего минусуют?**

* @Mordatyj
* @Mithgol
* @alizar
* @ilya42

Ваш Код:

**5. Сравните субботы и понедельники**

Правда ли, что по субботам авторы пишут в основном днём, а по понедельникам — в основном вечером?

Ваш Код: