

ИАД с Pandas

Исследовательский Анализ Данных

Обзор занятия



1. Краткая общая информация и история
2. Введение в методы pandas DataFrame для ИАД
 - а. Откройте чистый notebook и приготовьтесь печатать вместе со мной!
3. Упражнения в iPython notebook

Что такое ИАД?

- Важная часть науки о данных — это знать свои данные
- Основоположником считается Тьюки в 60-х гг.
- Его цитата:
 - “Методы анализа данных, техника интерпретации результатов таких методов, способы планирования сбора данных для их более простого и точного анализования а также все методики и результаты (математической) статистики, которые применимы для анализа данных.”

Джон Вайлдер Тьюки 1915-2000



Что такое ИАД?

Джон Вайлдер Тьюки 1915-2000

- Поддержка Тьюки вдохновила разработчиков на создание языков программирования для ИАД, например, S (который впоследствии стал R)
- Тьюки поддерживал это стремление, поскольку хотел, чтобы статистики больше работали со своими данными: искали гипотезы, а не тестировали их.
- Наше мнение:
 - Вы должны знать Ваши данные, чтобы избежать ошибок при их моделировании.
 - По мере знакомства с данными, пытайтесь удовлетворить своё естественное любопытство
 - Задавайте вопросы и используйте инструменты, чтобы **быстро** на них ответить



Pandas и Графики

Инструменты для ИАД

1. Откройте консоль
2. Перейдите к папке `eda_with_pandas`
(внутри извлечённого архива)
3. Активизируйте Ваше conda окружение
4. Запустите сервер jupyter notebook
5. **Создайте пустой notebook для записей**

Pandas - импортирование и анализ

```
$ df = pd.read_csv("data/flights08.csv")  
$ df.head()  
$ df.shape  
$ df.dtypes  
$ df.describe() # почему не include='all'
```

Pandas - недостающие значения

```
$ df.isnull()  
$ df.isnull().any() # axis=0 по умолчанию  
$ df.isnull().any(axis=1)  
$ df[df.isnull().any(axis=1)].head()  
$ boring_cols = [col for col in df.columns  
                  if 'Delay' in col][2:]  
$ boring_cols  
$ df[df.drop(boring_cols, axis=1).isnull().any(axis=1)].head()  
$ df.isnull().mean(axis=0) # пропорция недостающих
```

Pandas – описательная статистика и вычисления

Документация по всем методам включая mean, var, и т.д...

```
$ df['DayOfWeek'].nunique()  
$ df['DayOfWeek'].value_counts()  
$ df['DepDelay'].nlargest(12)  
$ idx = df['DepDelay'].nlargest(12).index  
  df.loc[idx]  
$ (df['Month'] == 1).all()  
$ (df['DayOfWeek'] == 8).any()
```


Seaborn – график распределения

```
$ import matplotlib.pyplot as plt
$ import seaborn as sns
$ sns.countplot(df['DayofMonth'])
plt.xticks(rotation=90)
$ sns.distplot(df['CRSElapsedTime'])
$ sns.distplot(df['DepTime'].dropna()) # эмм...проблема?
$ # Да! Мы импортировали это время как float => пробелы!
```

Pandas - apply

\$

```
def is_weekend(x):  
    if x in [6, 7]:  
        return True  
    else:  
        return False
```

\$

```
df['DayOfWeek'].apply(is_weekend)
```

\$

```
df['is_weekend'] = df['DayOfWeek'].apply(is_weekend)
```

\$

```
df['delay_gt_60'] = df['ArrDelay'].apply(lambda x: x > 60)
```

Pandas - **DataFrame** groupby

```
$ df.groupby('is_weekend')
$ for group_name, group_df in df.groupby('is_weekend'):
    print("выходной: {}".format(group_name))
    print(group_df['ArrDelay'].mean())
$ def delay_per_dist(df):
    return df['ArrDelay'].sum() / df['Distance'].sum()
$ delay_per_dist(df)
$ df.groupby('is_weekend').apply(delay_per_dist)
```

Pandas - **Series** groupby

```
$ df['delay_gt_60'].groupby(df['DayOfWeek']) # форма должна
$ delay = (                                     # совпадать
    df['delay_gt_60']
      .groupby(df['DayOfWeek'])
      .sum()
)
$ delay
$ delay.rename('Nr flights delayed gt 60 mins', inplace=True)
$ delay.reset_index()
$ sns.barplot(x='DayOfWeek', y='Nr flights delayed gt 60 mins',
              data=delay.reset_index())
```



Практическая часть

eda-with-pandas.ipynb