中心主题

- Pandas
  - 导入导出数据
    - Sql（sqlalchemy）、CSV、Excel、JSON等等
  - 查看数据
    - loc、iloc、df.col_name、df[col_name]
  - 筛选数据
    - df[df.col_name > x & (~df.col_name.isna())]
    - df.loc[:,df.col_name>10]
  - 索引
    - df.index、df.set_index('col_name')
    - df.col_name.is_unique#是否不重复
    - df.index.has_duplicates #索引是否有重复值
  - 数据概览
    - df.info,df.describe,df.nunique(),df.col_name.unique(),df.value_count()
  - 位置计算
    - diff、shift、rank
  - 常用操作
    - agg、apply、map、applymap、groupby、combine、concat
- **数据预处理**
  - 缺失值
    - 移除行列
    - 统计学，用均值、中位数、众数等填补，可groupby特定列后进行
      - df.groupby(['group_col1', 'group_col2'])['column_to_fill'].transform(lambda x: x.fillna(x.median()))
      - df.groupby(['group_col1', 'group_col2'])['column_to_fill'].apply(lambda x: x.interpolate(method='linear'))
    - 使用插值或者预测模型
  - 重复值
    - 移除
  - 异常值
    - 箱线图、Z-Score、IQR、LOF、孤立森林、时间序列、散点图、自动编码、综合方法
  - 日期处理
    - pd.to_datetime、astype('datetime64[ns]')
  - 数据重采样
    - resample