数据清洗代码

|  |
| --- |
| import csv  # 定义输入和输出文件的路径  input\_file\_path = 'clean\_test.csv'  output\_file\_path = 'cleaned\_data.csv'  # 打开输入文件和输出文件  with open(input\_file\_path, mode='r', encoding='utf-8') as csvfile, \  open(output\_file\_path, mode='w', newline='', encoding='utf-8') as output\_file:  # 创建csv阅读器和写入器  reader = csv.reader(csvfile)  writer = csv.writer(output\_file)  # 写入表头（假设第一行为表头）  headers = next(reader)  writer.writerow(headers)  # 遍历csv文件的每一行  for row in reader:  # 检查coupon\_id是否为空  if row[8]: # 这里假设coupon\_id是第9列（索引为8，因为索引是从0开始的）  writer.writerow(row) # 如果不为空，则写入到输出文件  print("处理完成，已移除coupon\_id为空的行。") |

MapReduce功能代码

|  |
| --- |
| import sys  import csv  from collections import defaultdict  from datetime import datetime  def map\_func(row):  # Convert each field to the appropriate data type  user\_id = int(float(row[0])) if '.' in row[0] else int(row[0])  user\_use\_coupon\_times = int(float(row[1])) if row[1] != '0.0' else 0  user\_consume\_times = int(float(row[2])) if row[2] != '' else 0  user\_use\_coupon\_rate = float(row[3])  user\_receive\_coupon\_unused\_times = float(row[4])  user\_mean\_use\_coupon\_interval = float(row[5])  number\_received\_coupon = int(float(row[6]))  merchant\_id = int(float(row[7])) if '.' in row[7] else int(row[7])  coupon\_id = int(float(row[8])) if '.' in row[8] else int(row[8])  discount\_rate = float(row[9])  distance = float(row[10])    # Convert date\_received to datetime object  date\_received\_str = row[11]  date\_received = datetime.strptime(date\_received\_str, "%Y-%m-%d") if date\_received\_str else None  date\_str = row[12]  date = datetime.strptime(date\_str, "%Y-%m-%d") if date\_str else None  label = int(row[13])    # Convert possible float strings to integers by first converting to float then to int  merchant\_launch\_coupon\_used\_count = int(float(row[14]))  merchant\_launch\_coupon\_used\_rate = float(row[15])  merchant\_launch\_coupon\_count = int(float(row[16]))  merchant\_receive\_coupon\_unused\_times = int(float(row[17]))  merchant\_mean\_launch\_coupon\_interval = float(row[18])  coupon\_used\_rate = float(row[19])  user\_merchant\_cus = int(float(row[20]))  user\_merchant\_received\_coupon = int(float(row[21]))  user\_merchant\_used\_coupon = int(float(row[22]))  return (user\_id, (user\_use\_coupon\_times, user\_consume\_times,  user\_use\_coupon\_rate,user\_receive\_coupon\_unused\_times,  user\_mean\_use\_coupon\_interval, number\_received\_coupon, merchant\_id,  coupon\_id, discount\_rate,  distance, date\_received, date, label, merchant\_launch\_coupon\_used\_count,  merchant\_launch\_coupon\_used\_rate, merchant\_launch\_coupon\_count,  merchant\_receive\_coupon\_unused\_times,  merchant\_mean\_launch\_coupon\_interval, coupon\_used\_rate,  user\_merchant\_cus,user\_merchant\_received\_coupon,  user\_merchant\_used\_coupon))  def reduce\_func(key, values):  # 对相同user\_id的数据进行汇总计算  total\_use\_coupon\_times = 0  total\_consume\_times = 0  total\_use\_coupon\_rate = 0  total\_receive\_coupon\_unused\_times = 0  total\_mean\_use\_coupon\_interval = 0  total\_number\_received\_coupon = 0  user\_data\_count = 0    for value in values:  total\_use\_coupon\_times += value[0]  total\_consume\_times += value[1]  total\_use\_coupon\_rate += value[2]  total\_receive\_coupon\_unused\_times += value[3]  total\_mean\_use\_coupon\_interval += value[4]  total\_number\_received\_coupon += value[5]  user\_data\_count += 1    avg\_use\_coupon\_times = total\_use\_coupon\_times / user\_data\_count  avg\_consume\_times = total\_consume\_times / user\_data\_count  avg\_use\_coupon\_rate = total\_use\_coupon\_rate / user\_data\_count  avg\_receive\_coupon\_unused\_times = total\_receive\_coupon\_unused\_times / user\_data\_count  avg\_mean\_use\_coupon\_interval = total\_mean\_use\_coupon\_interval / user\_data\_count  avg\_number\_received\_coupon = total\_number\_received\_coupon / user\_data\_count    return (key, (avg\_use\_coupon\_times, avg\_consume\_times, avg\_use\_coupon\_rate, avg\_receive\_coupon\_unused\_times,  avg\_mean\_use\_coupon\_interval, avg\_number\_received\_coupon)) |

显示平均值代码

|  |
| --- |
| import sys  import csv  import pandas as pd  import matplotlib.pyplot as plt  from collections import defaultdict  from baseclean\_data import map\_func, reduce\_func # 正确的导入方式  def main():  input\_data = sys.stdin    reader = csv.reader(input\_data)  next(reader) # 跳过表头    # 执行Map操作  mapped\_data = map(map\_func, reader)    # 排序以便进行Reduce操作  sorted\_data = sorted(mapped\_data, key=lambda x: x[0])    # 执行Reduce操作  reduced\_data = defaultdict(list)  for key, value in sorted\_data:  reduced\_data[key].append(value)    result = []  for key, values in reduced\_data.items():  result.append(reduce\_func(key, values))    # 输出结果  for r in result:  print(r)  if \_\_name\_\_ == "\_\_main\_\_":  main() |

绘制折线图代码

|  |
| --- |
| import sys  import csv  import pandas as pd  import matplotlib.pyplot as plt  from collections import defaultdict  from baseclean\_data import map\_func, reduce\_func  def main():  input\_data = sys.stdin    reader = csv.reader(input\_data)  next(reader) # 跳过表头    # 执行Map操作  mapped\_data = map(map\_func, reader)    # 排序以便进行Reduce操作  sorted\_data = sorted(mapped\_data, key=lambda x: x[0])    # 执行Reduce操作  reduced\_data = defaultdict(list)  for key, value in sorted\_data:  reduced\_data[key].append(value)    result = []  for key, values in reduced\_data.items():  result.append(reduce\_func(key, values))    ids = []  avg\_use\_coupon\_times = []  avg\_consume\_times = []  avg\_use\_coupon\_rate = []  avg\_receive\_coupon\_unused\_times = []  avg\_mean\_use\_coupon\_interval = []  avg\_number\_received\_coupon = []    for item in result:  ids.append(item[0])  avg\_use\_coupon\_times.append(item[1][0])  avg\_consume\_times.append(item[1][1])  avg\_use\_coupon\_rate.append(item[1][2])  avg\_receive\_coupon\_unused\_times.append(item[1][3])  avg\_mean\_use\_coupon\_interval.append(item[1][4])  avg\_number\_received\_coupon.append(item[1][5])      plt.figure(figsize=(15, 8))  # 可视化图1：平均使用优惠券次数和平均消费次数  plt.subplot(2, 3, 1)  plt.plot(avg\_use\_coupon\_times, label='Average Use Coupon Times', marker='o')  plt.plot(avg\_consume\_times, label='Average Consume Times', marker='x')  plt.xlabel('ID')  plt.ylabel('Times')  plt.legend()  # 可视化图2：平均使用优惠券比例和平均收到但未使用的优惠券次数  plt.subplot(2, 3, 2)  plt.plot(avg\_use\_coupon\_rate, label='Average Use Coupon Rate', marker='o')  plt.plot(avg\_receive\_coupon\_unused\_times, label='Average Receive Coupon Unused Times', marker='x')  plt.xlabel('ID')  plt.ylabel('Rate/Times')  plt.legend()  # 可视化图3：平均使用优惠券的时间间隔和平均收到的优惠券数量  plt.subplot(2, 3, 3)  plt.plot(avg\_mean\_use\_coupon\_interval, label='Average Mean Use Coupon Interval', marker='o')  plt.plot(avg\_number\_received\_coupon, label='Average Number Received Coupon', marker='x')  plt.xlabel('ID')  plt.ylabel('Interval/Number')  plt.legend()  plt.tight\_layout()  plt.show()  if \_\_name\_\_ == "\_\_main\_\_":  main() |

绘制散点图代码

|  |
| --- |
| import sys  import csv  import pandas as pd  import matplotlib.pyplot as plt  from collections import defaultdict  from baseclean\_data import map\_func, reduce\_func  def main():  input\_data = sys.stdin    reader = csv.reader(input\_data)  next(reader) # 跳过表头    # 执行Map操作  mapped\_data = map(map\_func, reader)    # 排序以便进行Reduce操作  sorted\_data = sorted(mapped\_data, key=lambda x: x[0])    # 执行Reduce操作  reduced\_data = defaultdict(list)  for key, value in sorted\_data:  reduced\_data[key].append(value)    result = []  for key, values in reduced\_data.items():  result.append(reduce\_func(key, values))    ids = []  avg\_use\_coupon\_times = []  avg\_consume\_times = []  avg\_use\_coupon\_rate = []  avg\_receive\_coupon\_unused\_times = []  avg\_mean\_use\_coupon\_interval = []  avg\_number\_received\_coupon = []    for item in result:  ids.append(item[0])  avg\_use\_coupon\_times.append(item[1][0])  avg\_consume\_times.append(item[1][1])  avg\_use\_coupon\_rate.append(item[1][2])  avg\_receive\_coupon\_unused\_times.append(item[1][3])  avg\_mean\_use\_coupon\_interval.append(item[1][4])  avg\_number\_received\_coupon.append(item[1][5])    # 可视化图1：平均使用优惠券次数和平均消费次数  plt.subplot(2, 3, 1)  plt.scatter(ids, avg\_use\_coupon\_times, label='Average Use Coupon Times')  plt.scatter(ids, avg\_consume\_times, label='Average Consume Times')  plt.xlabel('ID')  plt.ylabel('Times')  plt.legend()  # 可视化图2：平均使用优惠券比例和平均收到但未使用的优惠券次数  plt.subplot(2, 3, 2)  plt.scatter(ids, avg\_use\_coupon\_rate, label='Average Use Coupon Rate')  plt.scatter(ids, avg\_receive\_coupon\_unused\_times, label='Average Receive Coupon Unused Times')  plt.xlabel('ID')  plt.ylabel('Rate/Times')  plt.legend()  # 可视化图3：平均使用优惠券的时间间隔和平均收到的优惠券数量  plt.subplot(2, 3, 3)  plt.scatter(ids, avg\_mean\_use\_coupon\_interval, label='Average Mean Use Coupon Interval')  plt.scatter(ids, avg\_number\_received\_coupon, label='Average Number Received Coupon')  plt.xlabel('ID')  plt.ylabel('Interval/Number')  plt.legend()  plt.show()  if \_\_name\_\_ == "\_\_main\_\_":  main() |