实验五　Pandas 数据预处理

1.实验类型

设计型实验

2.实验目的和要求

（1）掌握数据查看的方法；

（2）掌握缺失值的识别与处理方法；

（3）掌握重复值的识别与处理方法；

（4）掌握异常值的识别与处理方法；

（5）掌握探索性分析的方法与应用；

3.实验内容

题目：AppleStore 应用商城数据分析

3.1 项目介绍：

* 随着网络的普及，各行各业都呈现蓬勃向上的发展趋势，而智能手机是改变我们生活的重要方式之一，因为智能手机可以允许安装各种 app 应用，这些应用极大的便利了我们的生活，因此，对 app 应用程序进行分析变得极为重要。

* 现在我们要对具有代表性的苹果应用商城的 app 应用进行分析，根据每个 app 应用的属性进行探索性分析。

注：本次实验需参考给定模板来完成，在要求部分，填入代码，并在实验报告中回答模板中的问题。

3.2 数据集分析

项目 AppleStore 数据集包含应用程序 ID、名称、大小、价格、评分、内容评级、主要类型、支持设备类型数量等信息，共包括 7207 行。

字段的描述：

- size_bytes：　大小（以字节为单位）
- price：　价格金额
- rating_count_tot：　用户评分计数（适用于所有版本）
- rating_count_ver：用户评分计数（当前版本）
- user_rating：平均用户评分值（适用于所有版本），取值范围为(0,5]，左开右闭区间
- user_rating_ver：平均用户评分值（对于当前版本）取值范围(0,5]，左开

右闭区间

- prime_genre：主要类型
- sup_devices.num：支持设备的数量
- lang.num：支持的语言数量

3.3 分析流程：

- 查看数据：导入数据，了解数据集的维度，每列数据的数据类型，以及打印部分数据，进行观察
- 数据清洗：删除不需要的列，删除重复行数据，删除包含有空值的行数据，以及对异常值进行处理
- 分析数据：

  (1)对用户评分进行分析

  (2)对 prime_genre 数据分析

  (3)发散题

4.实验背景知识

（1）数据查看

DataFrame.info（verbose = None，buf = None，max_cols = None，memory_usage = None，null_counts = None ）

打印 DataFrame 的简明信息。此方法打印有关 DataFrame 的信息，包括索引 dtype 和列 dtypes，非 null 值和内存使用情况。

（2）重复值识别与处理

pandas 提供了一个名为 duplicated(subset)的判断重复方法和一个名为 drop_duplicates 的去重方法。方法只对 DataFrame 或者 Series 类型有效。这种方法不会改变数据原始排列，并且兼具代码简洁和运行稳定的特点。该方法不仅支持单一特征的数据去重，还能够依据 DataFrame 的其中一个或者几个特征进行去重操作。

pandas.DataFrame(Series).drop_duplicates(self, subset=None, keep='first', inplace=False)

参数名称	说明
subset	接收string或sequence。表示进行去重的列。默认为None，表示全部列。
keep	接收特定string。表示重复时保留第几个数据。First：保留第一个。Last：保留最后一个。False：只要有重复都不保留。默认为first。
inplace	接收boolean。表示是否在原表上进行操作。默认为False。

（3）缺失值识别与处理

● 利用 isnull 或 notnull 找到缺失值。pandas 提供了识别缺失值的方法 isnull 以及识别非缺失值的方法 notnull，这两种方法在使用时返回的都是布尔值 True 和 False。

结合 sum 函数和 isnull、notnull 函数，可以检测数据中缺失值的分布以及数据中一共含有多少缺失值。

● pandas 中提供了简便的删除缺失值的方法 dropna，该方法既可以删除观测记录，亦可以删除特征。

pandas.DataFrame.dropna(self, axis=0, how='any', thresh=None, subset=None, inplace=False)

参数名称	说明
axis	接收0或1。表示轴向，0为删除观测记录（行），1为删除特征（列）。默认为0。
how	接收特定string。表示删除的形式。any表示只要有缺失值存在就执行删除操作。all表示当且仅当全部为缺失值时执行删除操作。默认为any。
subset	接收类array数据。表示进行去重的列/行。默认为None，表示所有列/行。
inplace	接收boolean。表示是否在原表上进行操作。默认为False。

● pandas 中提供了缺失值替换法，替换法是指用一个特定的值替换缺失值。

pandas 库中提供了缺失值替换的方法名为 fillna，其基本语法如下。

pandas.DataFrame.fillna(value=None, method=None, axis=None, inplace=False, limit=None)

参数名称	说明
value	接收scalar，dict，Series或者DataFrame。表示用来替换缺失值的值。无默认。
method	接收特定string。backfill或bfill表示使用下一个非缺失值填补缺失值。pad或ffill表示使用上一个非缺失值填补缺失值。默认为None。
axis	接收0或1。表示轴向。默认为1。
inplace	接收boolean。表示是否在原表上进行操作。默认为False。
limit	接收int。表示填补缺失值个数上限，超过则不进行填补。默认为None。

（4）异常值识别与处理

异常值是指数据中个别值的数值明显偏离其余的数值，有时也称为离群点，检测异常值就是检验数据中是否有录入错误以及是否含有不合理的数据。异常值的识别与处理需结合业务知识和实际数据开展。

常用的异常值识别方法有：

● 3σ 原则又称为拉依达法则。该法则就是先假设一组检测数据只含有随机误差，对原始数据进行计算处理得到标准差，然后按一定的概率确定一个区间，认

为误差超过这个区间的就属于异常值。

这种判别处理方法仅适用于对正态或近似正态分布的样本数据进行处理，如下表所示，其中 σ 代表标准差， μ 代表均值，x= μ 为图形的对称轴。

数据的数值分布几乎全部集中在区间(μ -3 σ , μ +3 σ)内，超出这个范围的数据仅占不到 0.3%。故根据小概率原理，可以认为超出 3 σ 的部分数据为异常数据。

数值分布	在数据中的占比
$(\mu-\sigma,\mu+\sigma)$	0.6827
$(\mu-2\sigma,\mu+2\sigma)$	0.9545
$(\mu-3\sigma,\mu+3\sigma)$	0.9973

● 箱型图提供了识别异常值的一个标准，即异常值通常被定义为小于 QL-1.5IQR 或大于 QU+1.5IQR 的值。

QL 称为下四分位数，表示全部观察值中有四分之一的数据取值比它小。

QU 称为上四分位数，表示全部观察值中有四分之一的数据取值比它大。

IQR 称为四分位数间距，是上四分位数 QU 与下四分位数 QL 之差，其间包含了全部观察值的一半。

箱线图依据实际数据绘制，真实、直观地表现出了数据分布的本来面貌，且没有对数据做任何限制性要求，其判断异常值的标准以四分位数和四分位数间距为基础。

四分位数给出了数据分布的中心、散布和形状的某种指示，具有一定的鲁棒性，即 25%的数据可以变得任意远而不会很大地扰动四分位数，所以异常值通常不能对这个标准施加影响。鉴于此，箱线图识别异常值的结果比较客观，因此在识别异常值方面具有一定的优越性。

（5）数据分析

探索性数据分析(Exploratory Data Analysis, EDA)注重对数据进行概括性的描述，不受数据模型和科研假设的限制。EDA 是指对已有的数据通过作图、制表、计算特征等手段探索数据的结构和规律的一种数据分析方法。适用于当对数据中的信息没有足够的经验，不知道该用何种传统统计方法进行分析时。另一种是验证性数据分析（Confirmatory Data Analysis, CDA），其注重对数据模型和假设的验证。

5.实验思考

（1） 数据质量不高的情况下如何提高数据质量？

（2） 数据预处理包含哪些内容？

（3） 数据预处理各个步骤是否有先后？

（4） 本实验中，你使用了哪些数据分析的方法？