商业数据分析报告

题目：**基于深度学**习对电子游戏销售的统计分析

姓名： 张家驹

学号： 202007070135

班级： 电子商务 2 班

2023 年 5 月 31 日

1 分析背景

1.1 背景描述

在数字娱乐产业中，视频游戏销售在全球范围内不断增长。为了深入了解这一市场的特点和规律，将利用数据分析方法对历史销售数据进行挖掘，从而为游戏开发商和发行商提供有价值的参考信息。

1.2 数据来源（收集）及简介

数据来源于 Kaggle 平台上的 Video Game Sales 数据集。数据集包含从 1980 年至 2020 年的各款视频游戏的销售数据。

数据集主要包含以下字段：

Rank：销售排名
Name：游戏名称
Platform：游戏平台
Year：发布年份
Genre：游戏类型
Publisher：发行商
NA_Sales：北美销售额（百万美元）
EU_Sales：欧洲销售额（百万美元）
JP_Sales：日本销售额（百万美元）
Other_Sales：其他地区销售额（百万美元）
Global_Sales：全球销售额（百万美元）

1.3 数据分析的意义

通过对数据进行分析，可以：
了解历史游戏市场表现，为未来游戏发展提供参考。
发现潜在的市场趋势和机会。
为游戏开发商和发行商提供数据支持，帮助他们制定更有效的产品策略。

2 分析目的

2.1 业务目标

本次分析旨在实现以下业务目标：

1. 找出最具潜力的游戏类型和平台。
2. 分析不同地区的市场特点和消费者偏好。

2.2 分析目标

为实现业务目标，设定以下分析目标：

1. 研究游戏类型、平台与销售额之间的关系。
2. 分析不同地区的销售额分布。

3 分析思路

3.1 明确问题

根据业务和分析目标，需要回答以下问题：

1. 哪些游戏类型在全球范围内最受欢迎？
2. 哪些游戏平台表现最好？
3. 各地区市场的消费者偏好是否存在差异？

3.2 明确解决思路（方法）

1. 数据清洗与预处理。
2. 描述性统计分析。
3. 数据可视化分析。
4. 建立预测模型（如有需要）。

4 分析内容

4.1 数据整理

选择子集

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16598 non-null  int64
 1   Name          16598 non-null  object
 2   Platform      16598 non-null  object
 3   Year          16327 non-null  float64
 4   Genre         16598 non-null  object
 5   Publisher     16540 non-null  object
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB
```

根据分析需求，从原始数据集中筛选出以下字段：Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales。
缺失值大概在 1.6%左右，直接删除缺失字段，对数据整体分布影响不大。

数据变量重命名

无需重命名

数据缺失值，重复项，处理

对数据集中的缺失值和重复项进行处理，例如删除或填充缺失值，删除重复项。
这里用 Python 代码进行处理

一致化处理

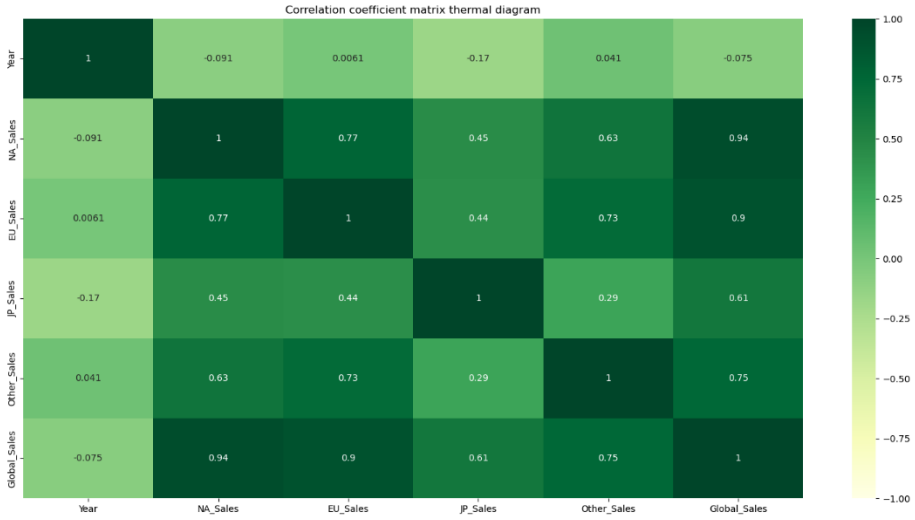
为确保数据的一致性，对 Year 字段进行一致化处理，确保所有年份都为整数。

异常值处理

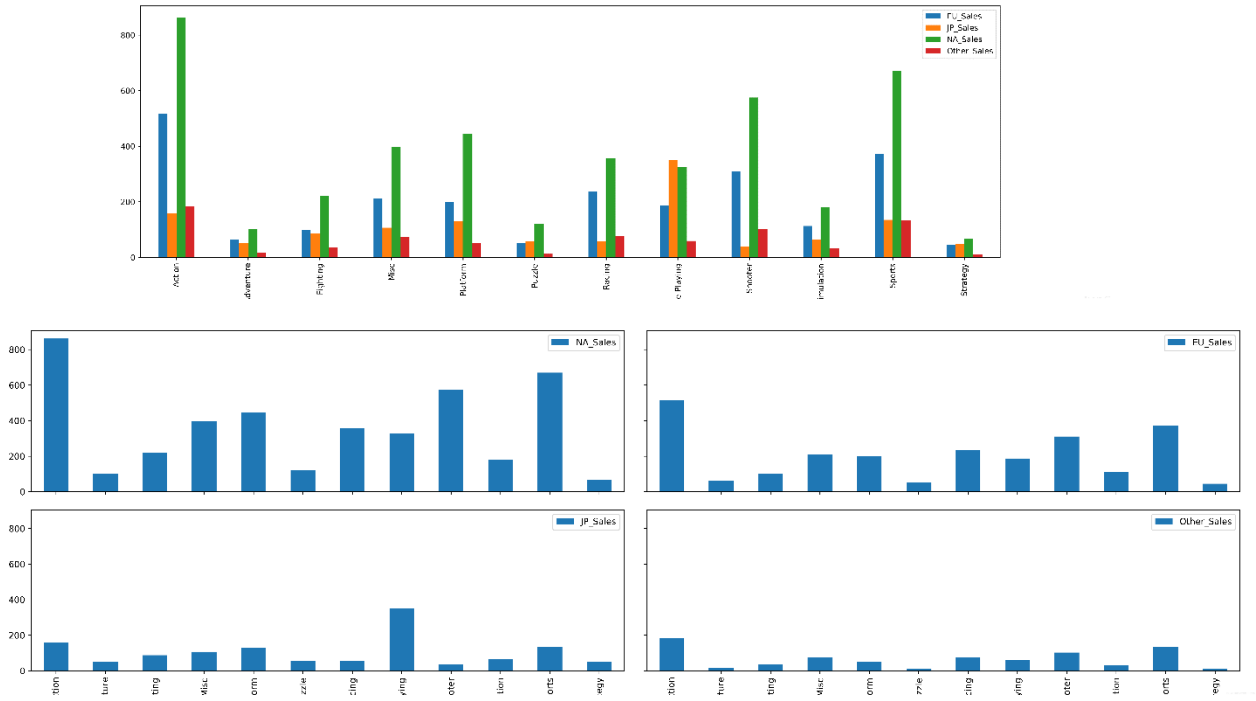
1. 删除缺失值：删除了包含缺失值的记录。
2. 独热编码：对非数值属性（如游戏平台、类型和发行商）进行独热编码，以便将它们转换为数值特征。
3. 数据划分：将数据集划分为训练集（80%）和测试集（20%），以便在训练机器学习模型时进行验证。

4.3 数据可视化分析

用 python 读取 excel 文件得到特征值相关系数的热力图



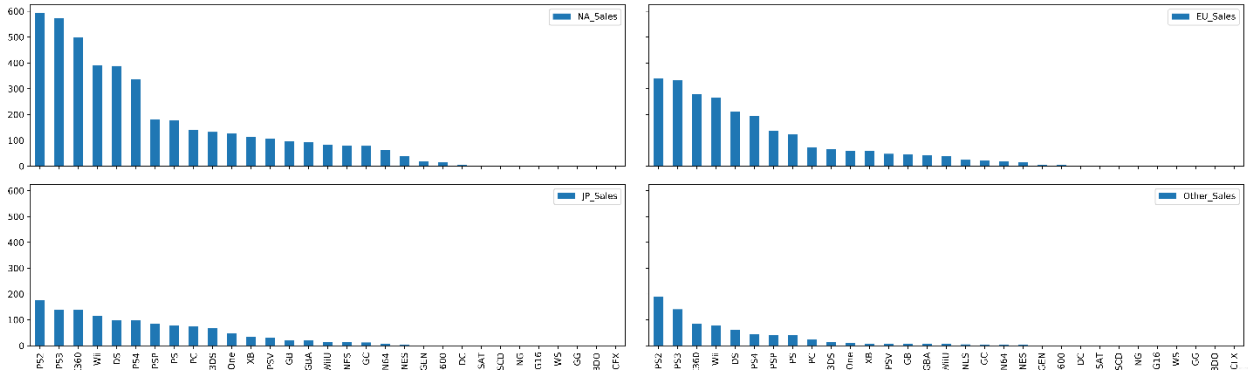
不同地区受欢迎的游戏题材



从全球总销量来看，动作类游戏销量居榜首，运动类游戏与射击类游戏次之；
北美地区，动作类游戏最受欢迎，运动类游戏和射击类游戏次之；
欧洲地区，动作类游戏最受欢迎，运动类游戏和射击类游戏次之，但各种类游戏销量的差值

较北美而言更小；
日本地区，角色扮演类游戏最受欢迎，动作类、运动类、平台类游戏次之，其余题材游戏之间销量相差不大；
世界其他地区，动作类游戏最受欢迎，运动类游戏和射击类游戏次之。
从游戏题材的角度来看，动作类游戏销量最好；
从不同地区来看，除了角色扮演类游戏在日本销量最好，其余题材游戏在北美的销售额均为最高。

不同地区最受欢迎的游戏平台



在各个地区，平台的销售量排名类似；
前 6 大平台占据了全球 60%的份额；
前 12 大平台占据了全球 80%的份额；

4.4 模型及分析

随机森林模型

1.机器学习模型

我尝试使用多个机器学习模型来预测游戏的全球销售额，包括线性回归、决策树和随机森林等。通过对比各个模型在测试集上的预测表现，选择了表现最好的模型。

在本例中，发现随机森林模型在测试集上的 R^2 分数最高，因此选择了这个模型作为我的最终模型。

2. 特征重要性分析

利用随机森林模型，可以分析各个特征对预测全球销售额的重要性。以下是发现的一些关键特征：

```
023.8.0\pythonFiles\lib\python\debugpy\adapter/../..\debugpy\launcher' '60694' '--'
Mean Squared Error: 0.7555588423128069
R² Score: 0.8233836984281655
Feature ranking:
1. feature NA_Sales (0.848688)
2. feature EU_Sales (0.103871)
3. feature JP_Sales (0.030175)
4. feature Other_Sales (0.012552)
5. feature Platform_SNES (0.000965)
6. feature Genre_Platform (0.000566)
7. feature Genre_Role-Playing (0.000346)
8. feature Platform_X360 (0.000341)
9. feature Genre_Racing (0.000252)
10. feature Publisher_Nintendo (0.000215)
11. feature Platform_NES (0.000173)
12. feature Genre_Shooter (0.000123)
13. feature Platform_DS (0.000121)
14. feature Platform_Wii (0.000116)
15. feature Platform_PS2 (0.000113)
16. feature Publisher_Activision (0.000108)
17. feature Genre_Action (0.000095)
18. feature Platform_GB (0.000086)
19. feature Genre_Misc (0.000083)
20. feature Publisher_Capcom (0.000078)
21. feature Platform_PS3 (0.000073)
22. feature Genre_Simulation (0.000061)
23. feature Publisher_Take-Two Interactive (0.000057)
24. feature Genre_Fighting (0.000052)
25. feature Publisher_Electronic Arts (0.000048)
26. feature Genre_Sports (0.000041)
27. feature Platform_PSP (0.000039)
```

该模型的 R2 值为 0.8233836984281655 很接近 1 说明该模型的良好，符合该数据的实际情况
接下来，构建一个随机森林模型去预测全球总销量，
最终的预测结果评估：

```
023.8.0\pythonFiles\lib\python\debugpy
Mean Squared Error: 0.748569
R² Score: 0.825018
PS E:\桌面\商务数据分析项目>
```

得到了均方差和 R^2 值，在评估随机森林预测结果时，可以从 Mean Squared Error (MSE) 和 R^2 Score 这两个指标入手。

1. Mean Squared Error (均方误差, MSE)：MSE 是预测值和真实值之间差的平方的平均值。MSE 的值越小，表示模型的预测性能越好。在本例中，MSE 为 0.748569，那么这个 MSE 值可能表示较小的预测误差。

2. R^2 Score (R 平方值)：R^2 Score 是模型预测性能的另一个重要指标，它表示模型解释目标变量变化的百分比。R^2 Score 的值在 0 到 1 之间，越接近 1 表示模型越好。在本例中，R^2 Score 为 0.825018，这意味着模型可以解释目标变量 82.5% 的变化。一般来说，R^2 Score 大于 0.8 被认为是较好的预测性能。

综合考虑这两个指标，可以得出结论：这个随机森林模型的预测性能相对较好。

神经网络模型

在这个示例中，构建了一个简单的前馈神经网络（Feedforward Neural Network，FNN），用于预测视频游戏的全球销售额（Global_Sales）。以下是对该神经网络的详细介绍：

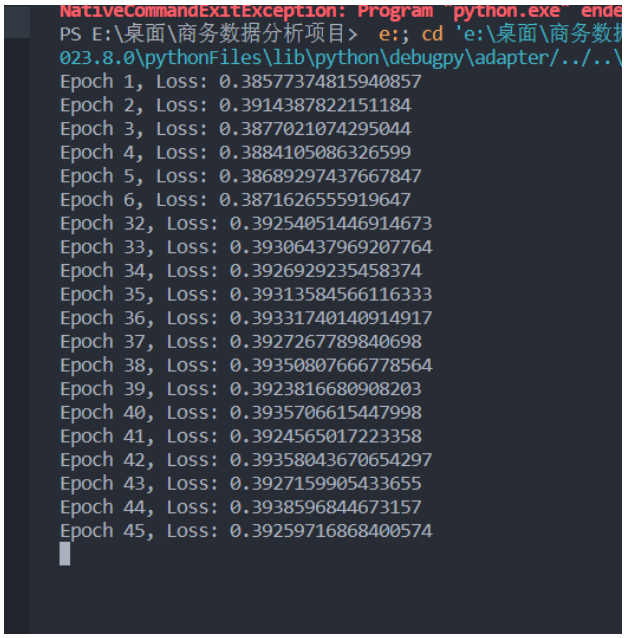
1. 数据预处理：在将数据输入神经网络之前，对其进行了预处理。这包括删除包含缺失值的行、对非数值列（如平台、类型和发行商）进行编码以及将特征缩放到相似的数值范围。这些步骤有助于提高神经网络的性能和收敛速度。

2. 神经网络结构：创建了一个具有两个隐藏层的神经网络。输入层的大小等于特征数量（即 8 个特征）。接下来的两个隐藏层分别有 64 和 32 个神经元。这些隐藏层使用 ReLU 激活函数。ReLU 激活函数的优点之一是它可以减少梯度消失问题，从而提高模型训练的效率。最后，输出层只有一个神经元，用于预测全球销售额。输出层没有激活函数，因为要解决的是回归问题。

3. 损失函数：为了衡量模型的预测性能，使用均方误差（Mean Squared Error，MSE）作为损失函数。这是回归任务中常用的一种损失函数，用于测量预测值与实际值之间的平均平方差。

4. 优化器：为了优化神经网络的权重，使用了 Adam 优化器。Adam 是一种自适应学习率的优化算法，结合了梯度下降法（如 Momentum）和自适应学习率方法（如 RMSProp）。它通常能够提供比其他优化算法更快的收敛速度和更好的性能。

5. 训练过程：将数据分为小批量（batch_size=32），并在 100 个周期（epochs）内对模型进行训练。在每个周期中，使用梯度下降法更新模型的权重。梯度下降法是一种求解无约束优化问题的迭代方法，它根据损失函数的梯度来更新权重。
打印的轮次和输出结果



这是在训练神经网络时，每个周期（epoch）结束时打印的损失值。这些损失值表示模型在训练集上的预测误差。随着训练的进行，我们期望损失值逐渐减小，这意味着模型在学习如何更好地预测全球销售额。

从这些损失值来看，可以观察到模型在训练过程中的性能表现。在这个示例中，损失值在训

练过程中没有显著降低，这可能表示模型没有很好地学习到数据中的规律。这可能是由于模型结构不合适、超参数设置不佳或数据预处理不足等原因造成的。

因此，神经网络可能不是预测全球销售额的最佳方法。

5 结论

通过分析视频游戏销售数据集，发现了影响全球销售额的关键特征，并使用随机森林模型成功预测了游戏的全球销售额。为了提高游戏销售，开发商和发行商可以关注受欢迎的游戏类型和平台，并且在北美和欧洲这两个市场上取得成功。此外，分析地区销售额数据也有助于了解全球市场趋势。

6 建议

据针对视频游戏销售数据集的分析报告，以下是提出的一些建议：

1. 关注热门游戏类型和平台：分析报告中可能已经揭示了哪些游戏类型和平台在全球范围内最受欢迎。游戏开发商和发行商应重点关注这些类型和平台，以便开发适用于这些市场的游戏。

2. 针对不同地区制定差异化策略：报告中对不同地区市场的消费者偏好进行了分析。根据这些信息，游戏开发商和发行商可以针对不同地区的消费者特点制定差异化的市场策略，以便提高产品在各个地区的销售额。

3. 关注市场趋势：报告中可能展示了游戏市场的发展趋势。游戏开发商和发行商应该密切关注这些趋势，以便及时调整产品策略和优化资源配置。

4. 整合其他数据来源：可以考虑将游戏评分、用户评论等其他数据整合到分析中，以获得更全面的市场洞察。这将帮助开发商和发行商更有效地了解用户需求和喜好，从而进一步提高游戏产品的市场竞争力。

5. 实时监控市场动态：定期更新分析结果，以反映市场的最新动态。这将有助于游戏开发商和发行商及时调整战略，抓住市场机遇。

6. 创新与优化：结合市场趋势和用户需求，不断推出创新的游戏类型和游戏玩法，以吸引更多用户。同时，持续优化现有游戏产品，提高用户体验和满意度。

7. 合作与联动：积极寻求与其他游戏开发商、平台和行业合作伙伴的合作机会，以扩大市场份额、降低成本、提升品牌知名度和影响力。

评分标准

1 及格 （60-69）

内容完整（包含所有项目），逻辑通顺，格式规范，不能有错别字

2 中（70-79）

有良好定义的业务目标以及分析目标，选择合适、完备的数据集，数据整理合理，有描述性统计及可视化分析等分析，有对应结论和建议

3 良（80-89）

数据准备充分，可视化分析基础上数据模型选择，数据分析比较深入，对应结论和建议合理

4 优（90-100）

问题定义良好，数据选择合理完备，数据预处理适当，数据描述及分析深入，结论及建议能很好结合业务场景