学 号：

华北理工大学

NORTH CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

数 据 科 学 之 美

期 末 报 告

报告题目：当代中国青年阅读文学类书籍情况分析

学生姓名：

专业班级：

学　　院：

指导教师：

2023 年 4 月 20 日

第一部分：青年阅读情况数据收集的展示

为了较为全面、透彻地分析青年阅读文学类书籍的情况，通过百度、知网等搜索引擎主要对两大方面进行了数据收集：各类型书籍阅读需求及阅读量（阅读时长及阅读数量）。

其中，各类型书籍阅读需求又分为文学类书籍相对于全部书籍类型阅读需求、文学类书籍分支下各种具体书籍类型的阅读比例（如小说、名著等），前者即图-1、图-2，后者即图-3，图-3 存在文学类型以外书籍，会在第二部分进行清洗并重新整合。

一、各类型书籍阅读需求

图-1　中国青年阅读需求指数

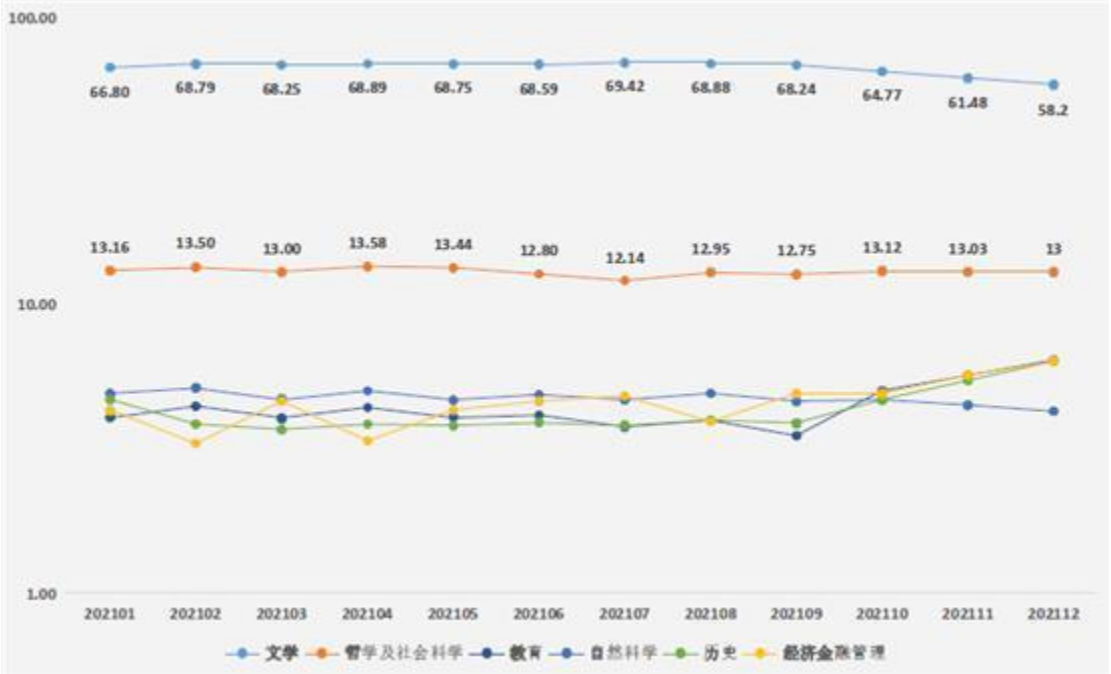

图-2　中国青年阅读域市场映射指数

“本期中国青年阅读指数，文学阅读域需求指数依然占据 2021 年 12 月需求指数排名的第一位，环比上升 1.06，也是当月需求指数上升最多的阅读域。从 2021 年 12 月文学阅读域阅读热度词云图中，我们可以看到，经典、小说、文学史等阅读域的阅读指数表现比较突出。”

——文本 1

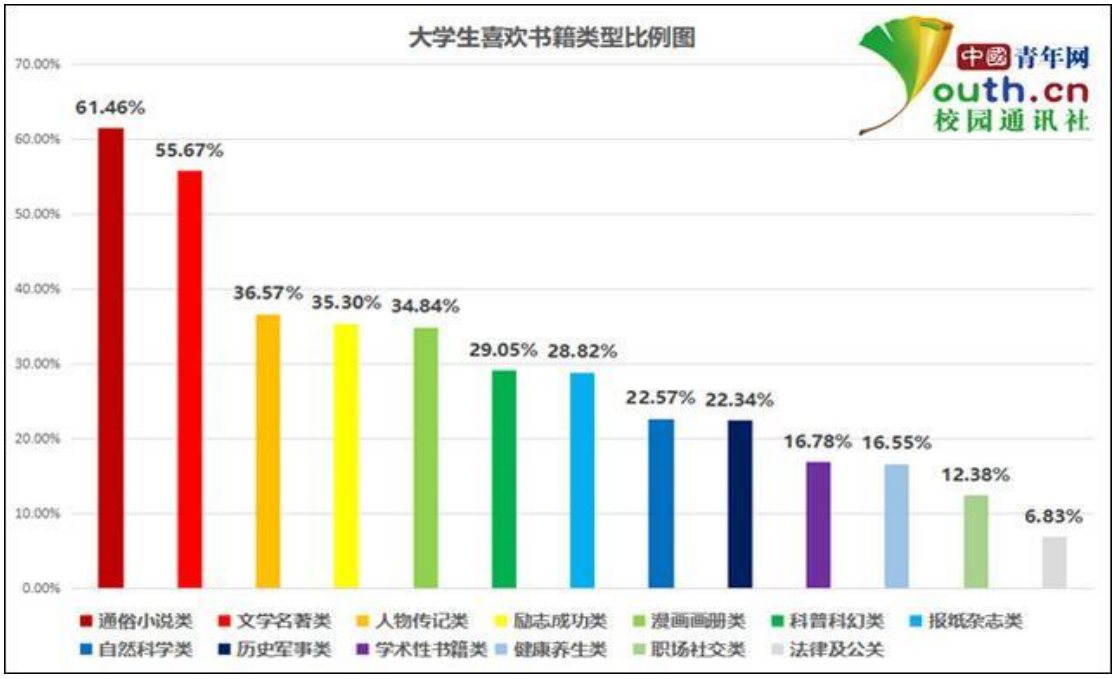

图-3　大学生喜欢书籍类型比例 中国青年网记者 李华锡 制图

二、阅读量（阅读时长及阅读数量）

最后便是阅读量，包括每日平均阅读量（图-4）和每月平均阅读书籍数量（图-5），全面反映实际阅读量。

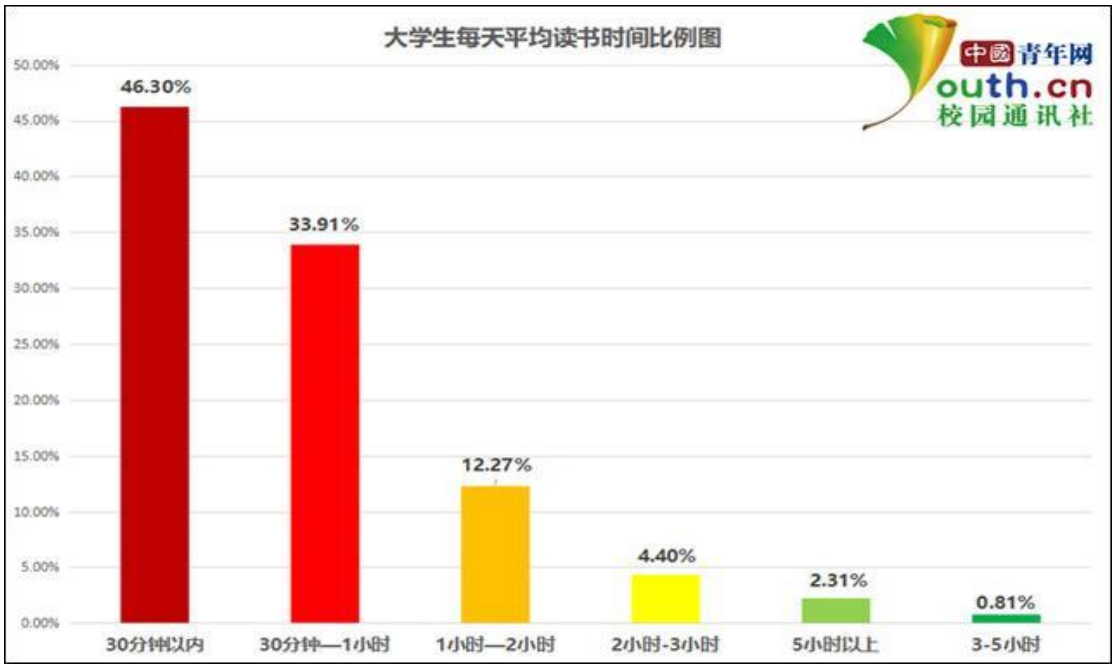

图-4　大学生每日平均读书时间

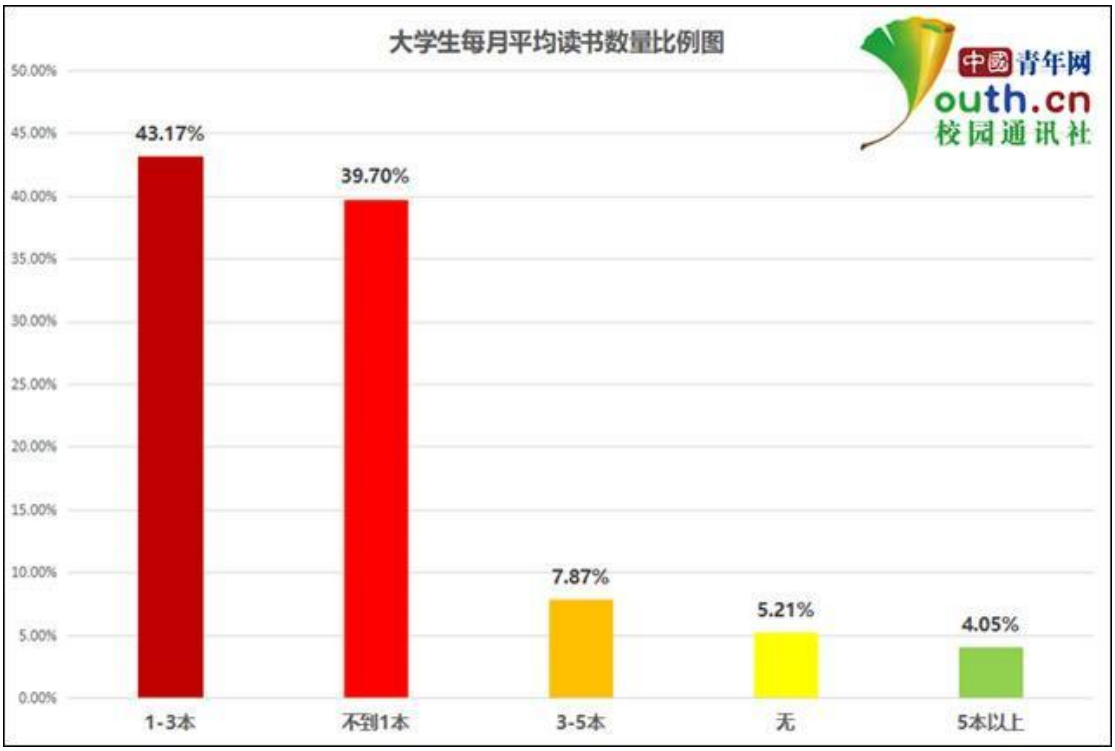

图-5　大学生每月平均读书数量比例

“根据中国青年网记者调查显示，46.3%的受访学生平均每天读书时间不足 30 分钟，每天花费 30 分钟到 1 个小时、1 个小时以上时间读书的学生分别占 33.91%、19.79%。在读书数量上，44.91%的受访学生平均每月读书不到一本，另有 43.17%和 11.92%的学生分别平均每月读书 1-3 本和 3 本以上。”

——文本 2

该部分多为中国青年网、知网等权威网站发布数据，内容相对真实可靠，数据类型多为图片，少数为文本信息。

第二部分：各数据类型的清洗预处理过程展示

我们要根据阅读需求及阅读量两方面来分析青年阅读文学情况，所以要选取图片及文字两种类型数据进行数据清洗、数据标准化、数据集成等预处理。

预处理过程分为三部分：

1.首先清洗数据。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。通过图-1、文本 1、图-2、图-3 都可以看出较其它类型书籍，文学类书籍需求指数最大，都说明了文学类书籍的受欢迎，且处于增长态势。故四者属于重复数据。而图-3 更能清晰具体地表达青年阅读文学类书籍的情况，故选取图-3 中数据进行进一步讨论分析。对于不完整的数据、错误的数据，通过审查并未发现。

2.而后进行数据标准化。数据标准化也就是统计数据的指数化。数据标准化处理主要包括数据同趋化处理和无量纲化处理两个方面。对于图-3，取通俗小说类、文学名著类、人物传记类均为文学类，对三者受喜爱率之和求平均值作为文学类受喜爱比例。即$A = (a[1] + a[2] + a[3])/3$。同理，对剩余 10 项求平均值代表非文学类书籍受喜爱比例。即$B = \sum_{i=1}^{n} b_i /n$。然后将二者用百分制转化并用扇形图形象表示出来。

3.最后着眼阅读量，将其进一步分为阅读时长在 30 分钟以上（达标）和不足 30 分钟的、

每月阅读量在 1 本以上（达标）和不足一本的，进一步呈现数据。

结果展示：1.对于第一部分即提取出“文学类书籍需求最大且处于增长态势”信息。

2.对于第二部分即扇形图（图-6）展示：

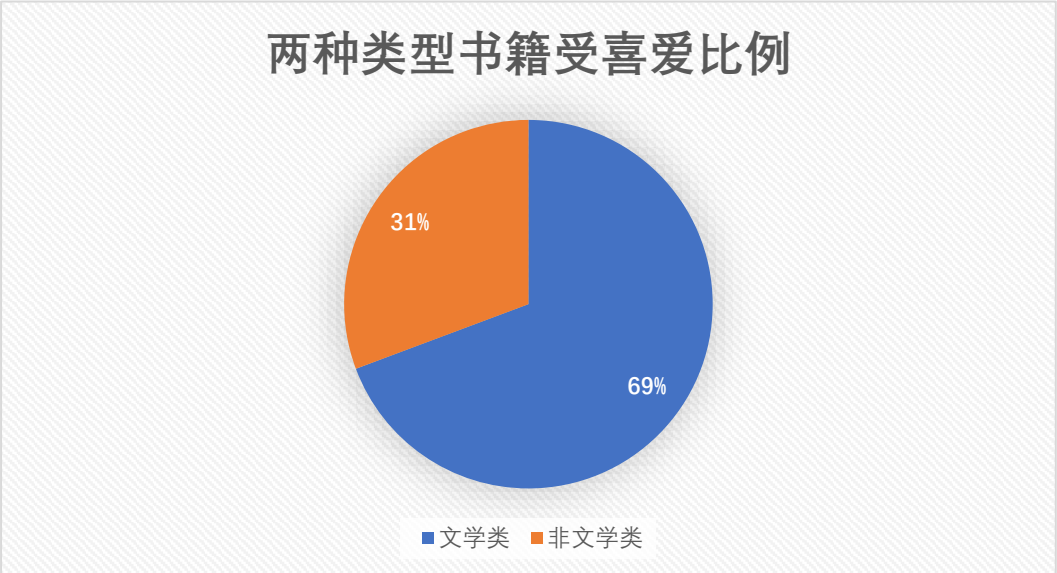

图-6　两种类型书籍受喜爱比例

说明文学类书籍受喜爱程度远高于其他类型书籍，相较于其它书籍，文学类型书籍更受大众偏爱。

3.对于第三部分同样通过图表展示：

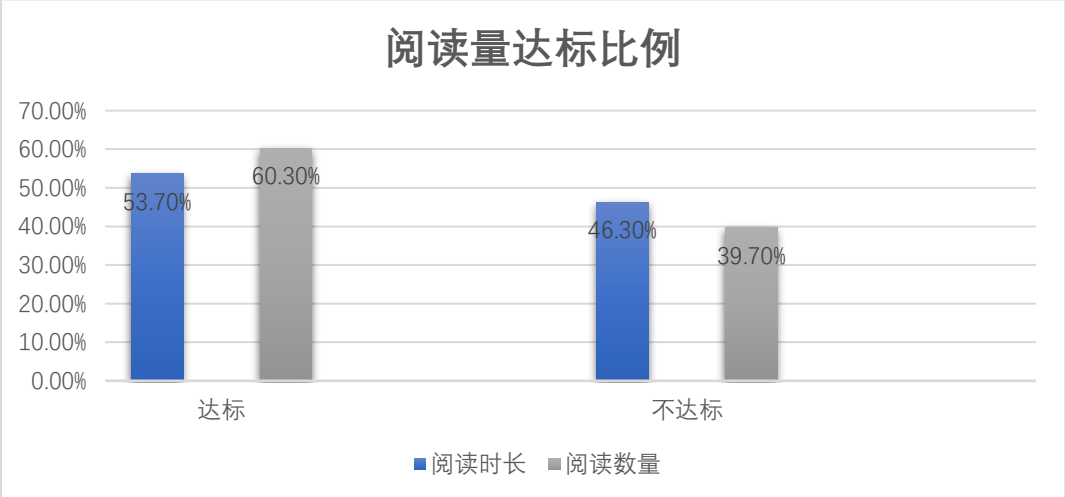

图-7　阅读量达标比例

说明虽然大多数青年会进行阅读，但真正能达到每日阅读 30 分钟以上或每月阅读 1 本以上书籍的青年比例并不突出。尤其对于每日坚持阅读，少于 30 分钟（不达标）者将近占比一半，将阅读纳入日常生活还需广大青年自省自强。

第三部分：各数据类型的挖掘分析过程展示

目的要对青年阅读文学类书籍情况做一个形象化且相对全面的展示，要对第二部分预处理后的图片、数字、文本类型数据进行回归预测估计、关联规则挖掘、聚类等数据挖掘。

挖掘分析过程：

1.进行回归分析，建立数学模型。通过初步分析得出大学生每日阅读时长与每月阅读数

量有强相关性，结合预处理结果建立数学模型：有效阅读文学类书籍青年比例（S）=喜爱文学类书籍（A）*有效阅读比例（N），其中有效阅读比例取大学生每日阅读时长达标比例与每月阅读数量达标比例的平均值，同理可得出不达标比例。

2.关联规则挖掘。而为了深入探索青年阅读文学类书籍的情况，将图-3 中文学类书籍分支下三种书籍按照受喜爱比例做比值并百分比化，直观呈现青年更倾向于哪类文学类书籍。

挖掘结果展示：S 不足 40%，但说明真正喜爱并坚持一定量阅读量的青年总体上仍居首位。如下图 S 数据表格：

	喜爱文学类书籍（0.69）	喜爱其它类书籍（0.31）
阅读达标比例（0.57）	39.3%	17.7%
阅读不达标比例（0.43）	29.7%	13.3%

表-1　喜爱某类书籍且阅读达标与否占比表

通俗小说类	文学名著类	人物传记类
40%	36%	24%

表-2　喜欢不同文学类书籍的比例

第四部分：各数据类型的数据可视化过程展示

选取第三部分中表-1 喜爱某类书籍且阅读达标与否占比表进行图像化展示，采用更加直观易懂的扇形图进行可视化。

可视化过程：

1.将表-1 中各类比例至于表中且划分不同颜色分配比例，使结果一目了然，不仅提高用户的理解和分析数据的效率，而且也能够让用户更好地发现青年阅读文学类书籍数据背后的规律和趋势，帮助用户做出更准确和有意义的决策判断。

2.同理将表-2 中各类比例用柱状图呈现，突出对比。

可视化展示：

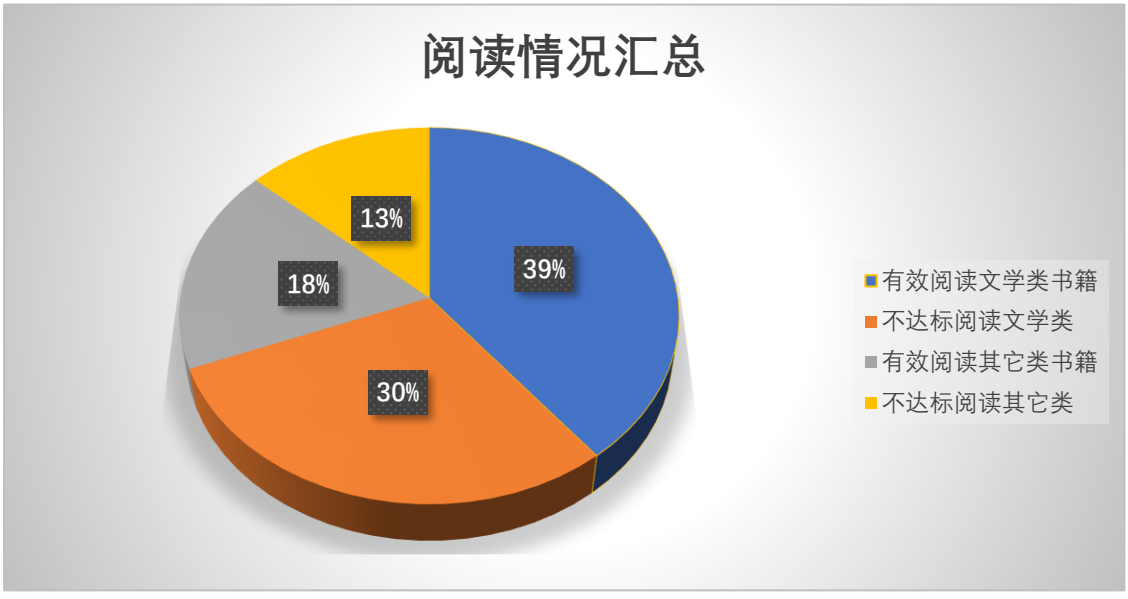

图-8　阅读情况汇总

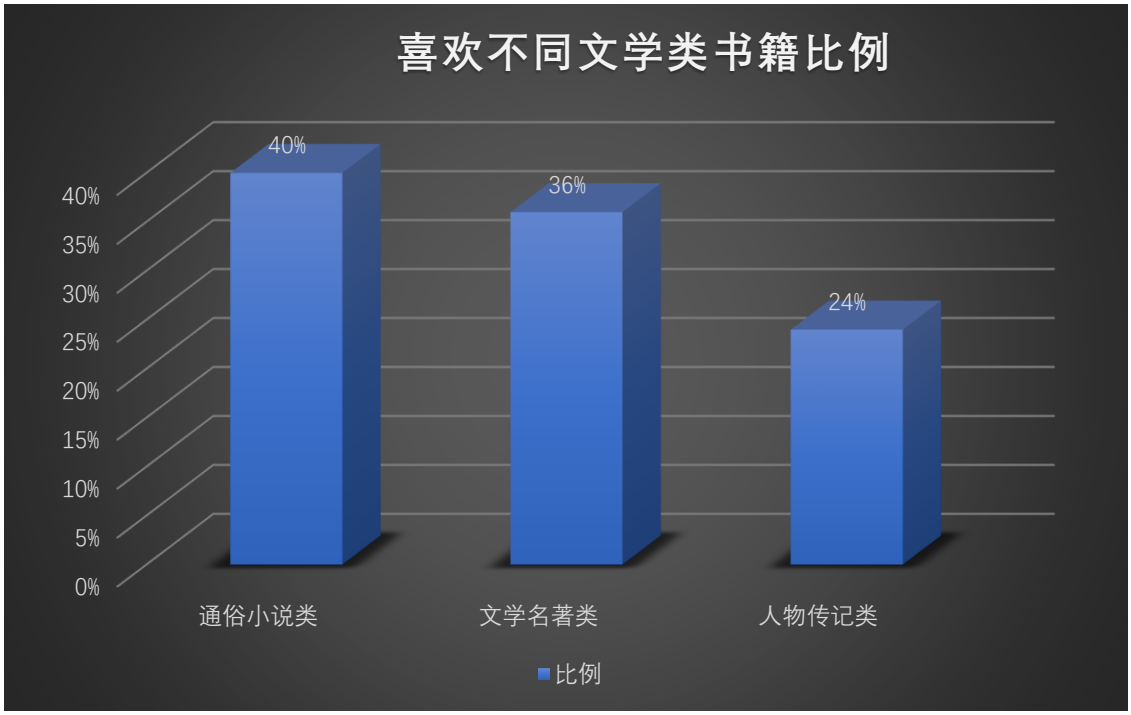

图-9　喜欢不同文学类书籍的比例

据图-8 可以看出有效阅读文学类书籍的比重仍最大，青年阅读文学类书籍的情况比较乐观，而据图-9 观察青年多喜爱通俗小说类，对于文学名著人物传记类等人文价值更大的书籍兴趣相对较少。青年应养成坚持阅读的习惯，从中汲取有价值的知识，努力提升自我，在碎片化信息时代坚守自我，展现青春光彩！