

**课程报告**

**课程名称：大数据导论**

**专业班级： 计算机1908**

**学 号： 见任务分工**

**姓 名： 杨天元/周奥洋/邓至廷**

**指导教师： 王蔚**

**报告日期： 5/15/2020**

**计算机科学与技术学院**

目录

[1 报告题目：CSGO饰品数据分析数据级报告](#_Toc28484)

[1.1 测试环境说明](#_Toc14372)

[1.2 数据集特征说明](#_Toc6636)

[1.3 测试应用说明](#_Toc21769)

[1.4 研究目的及意义](#_Toc328)

[1.5 问题挑战](#_Toc9975)

[1.6 测试结果](#_Toc11544)

[1.7 角色分工](#_Toc4566)

[1.8 心得体会与总结](#_Toc19500)

1 报告题目：CSGO饰品数据分析数据级报告

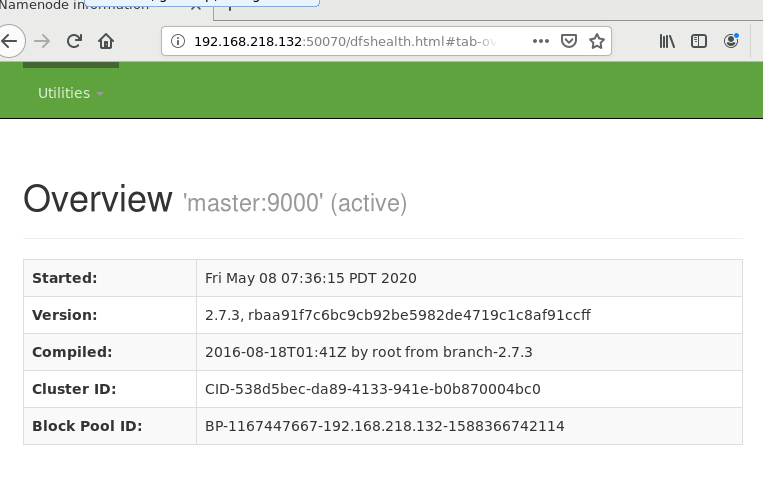
* 1. 测试环境说明

采取在VMware workstation的环境下搭建以HDFS进行存储，MapReduce提供计算的Hadoop集群。搭载集群的虚拟机采用的是操作系统为centos6.5的虚拟机，两台虚拟机配置如下：

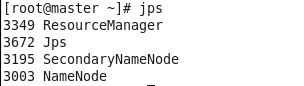


虚拟机名称分别为master和slave1，平台所需的JDK版本为1.8.0，hadoop软件版本为2.7.3

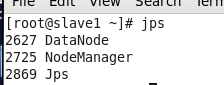
集群搭建web信息如图：



输入jps命令后如图：



master



2slave1

* 1. 数据集特征说明

本项目的数据来源于<https://buff.163.com/>

数据集由饰品、售价信息和饰品交易记录三部分组成，其结构如下：

饰品：{饰品名，ID，类别，类型，}

售价信息：{饰品名，上架时间，上架价格}

交易记录：{饰品ID，交易时间，交易价格}

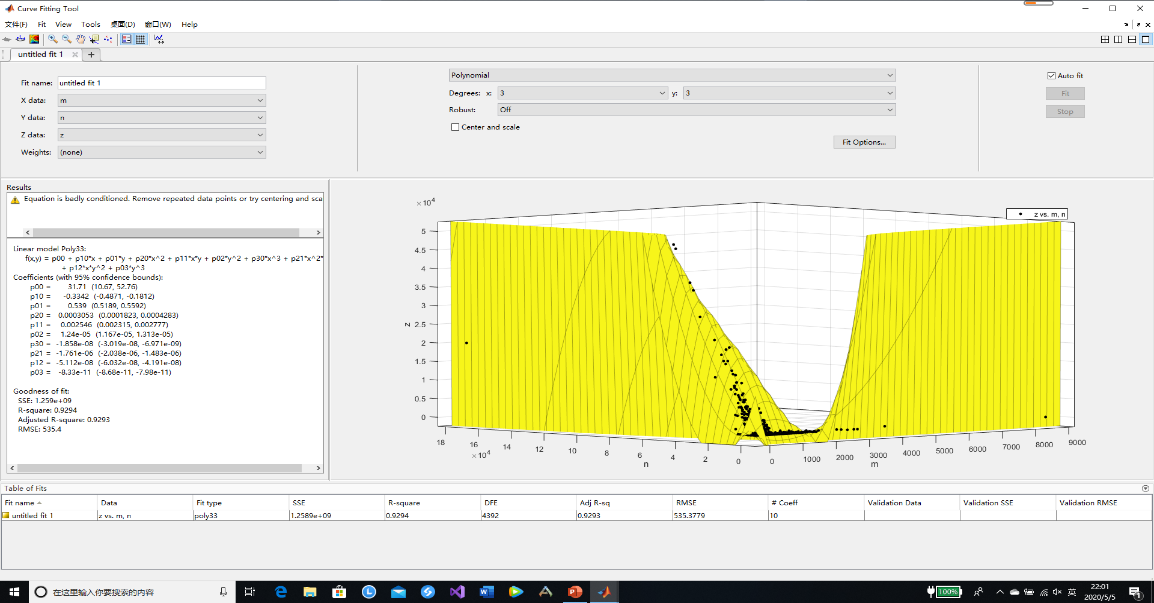
数据量约为2 GB，采用scrapy框架进行采集

* 1. 测试应用说明

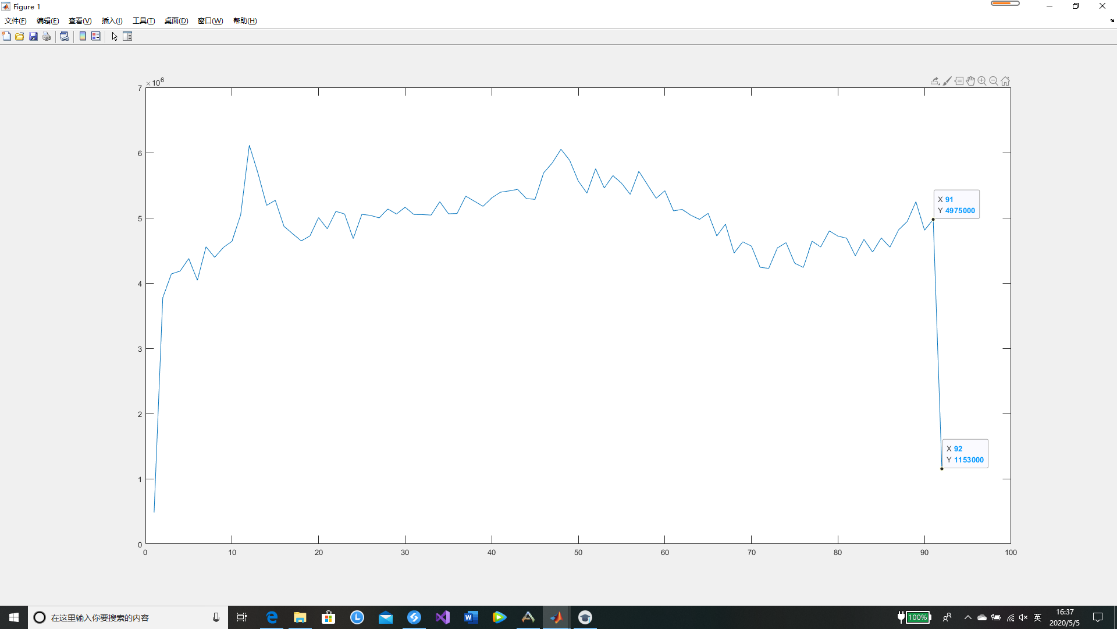
分析数据主要应用了MATLAB，Python，以及Tableau

1.MATLAB（朱昱洋）

我本次使用的是MATLAB进行数据分析，使用它的原因主要在于其提供了大量方便进行数据分析以及绘制成图的函数与工具箱。本次主要针对buff平台上的商品交易数据进行处理，由于爬取数据属性较多，使用MATLAB可以有效针对单一属性进行排序，统计以及删除异样数据。此外使用其自带的cftool工具箱，可以将处理好的数据进行拟合分析。由于MATLAB是基于矩阵进行处理，所以我可以将需要分析的数据放入同一个矩阵之中进行拟合。cftool工具箱可以手动选择拟合方式以及拟合程度，我选择将数据点拟合成三次多项式，并得到一张拟合曲线图。



另外我还利用自带函数将同一天的交易数据进行累计，将每天的成交额以及上架额度与时间形成关联，最终得出多幅折线图。



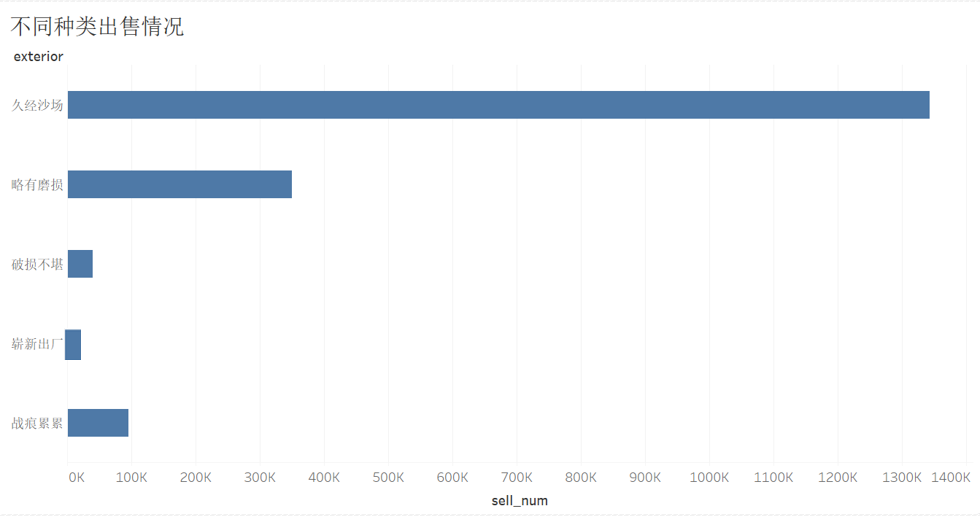
正是由于MATLAB数据分析能力强，且导入数据可以直接变化成矩阵进行分析，所以我们选择了MATLAB作为测试应用之一。

2.Python+Tableau（廖翔）

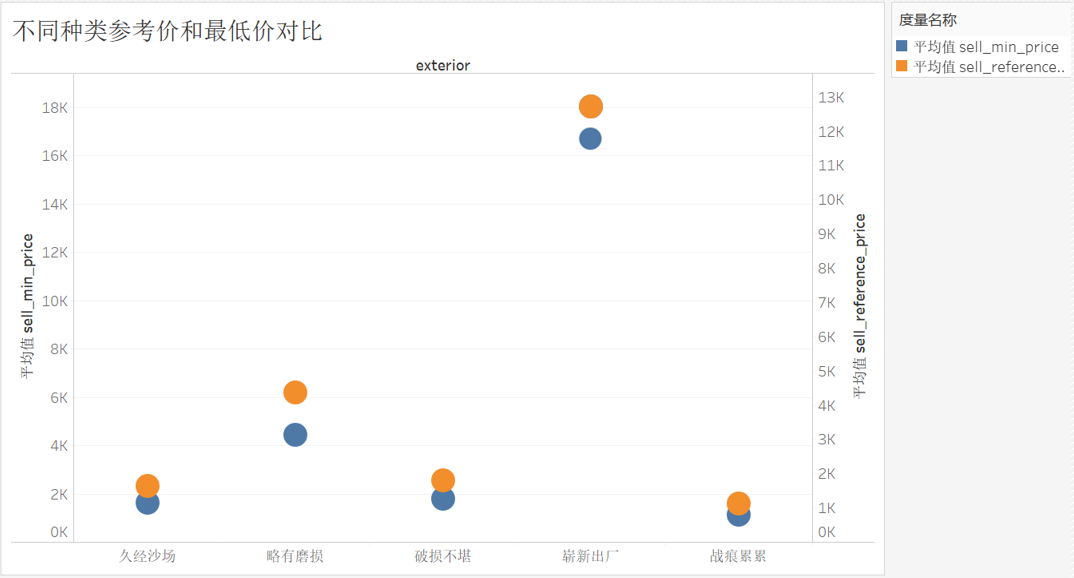
Python由于其语言简单易懂，通用，存在科学计算活跃区域，因此对于大数据分析有很明显的帮助。另外，python拥有Numpy,matplotlib,scikit-learn,pandas,ipython等工具在科学计算方面优势明显，尤其是pandas，在处理中型数据方面更为方便。

得益于Python丰富的第三方库，Python在数据分析方面，能够处理的问题非常之广，从Excel的公式计算到R语言中一些零散的机器学习库所能做的事情，Python都能处理。而在这些工具不擅长的爬虫，Python更是能出色的完成。

但是，仅仅凭借Python官方的作图工具做出来的图并不十分美观，并且由于大小的限制，并不能让我们直观的看到数据变化的规律，因此，在Python进行了数据处理的基础上使用Tableau这一工具来进行数据可视化就显得更为优秀而美观。



如图，通过使用Tableau，让我们直观的看到了五种不同磨损度的饰品的出售数目以及其之间的差距大小，通过分析我们发现并不是磨损程度越高（质量越差）的就一定卖的最多，玩家还是偏向于做出自己的选择。



每一款饰品都有一个官方建议售价，我们同样爬取了这一属性，并与最低价进行了比较，发现官方建议售价总是略高于最低价，这也为我们出售这一类饰品提供了参考价值。

Python在数据分析这方面功能十分强大，而加上Tableau能创建出交互，精美，智能的视图，也就更加方便我们的分析。

* 1. 研究目的及意义

CSGO是一款当下流行的电子游戏，游戏中的饰品因其高流通性、高保值性以及合乎市场规律的波动性而成为很多人的投资选择，堪称“游戏界的股票”，CSGO也因此被称为“理财产品”。

游戏中的饰品种类繁多，我们希望利用大数据来分析近三个月内饰品的价格走势和交易信息，并以此为饰品投资提供帮助。

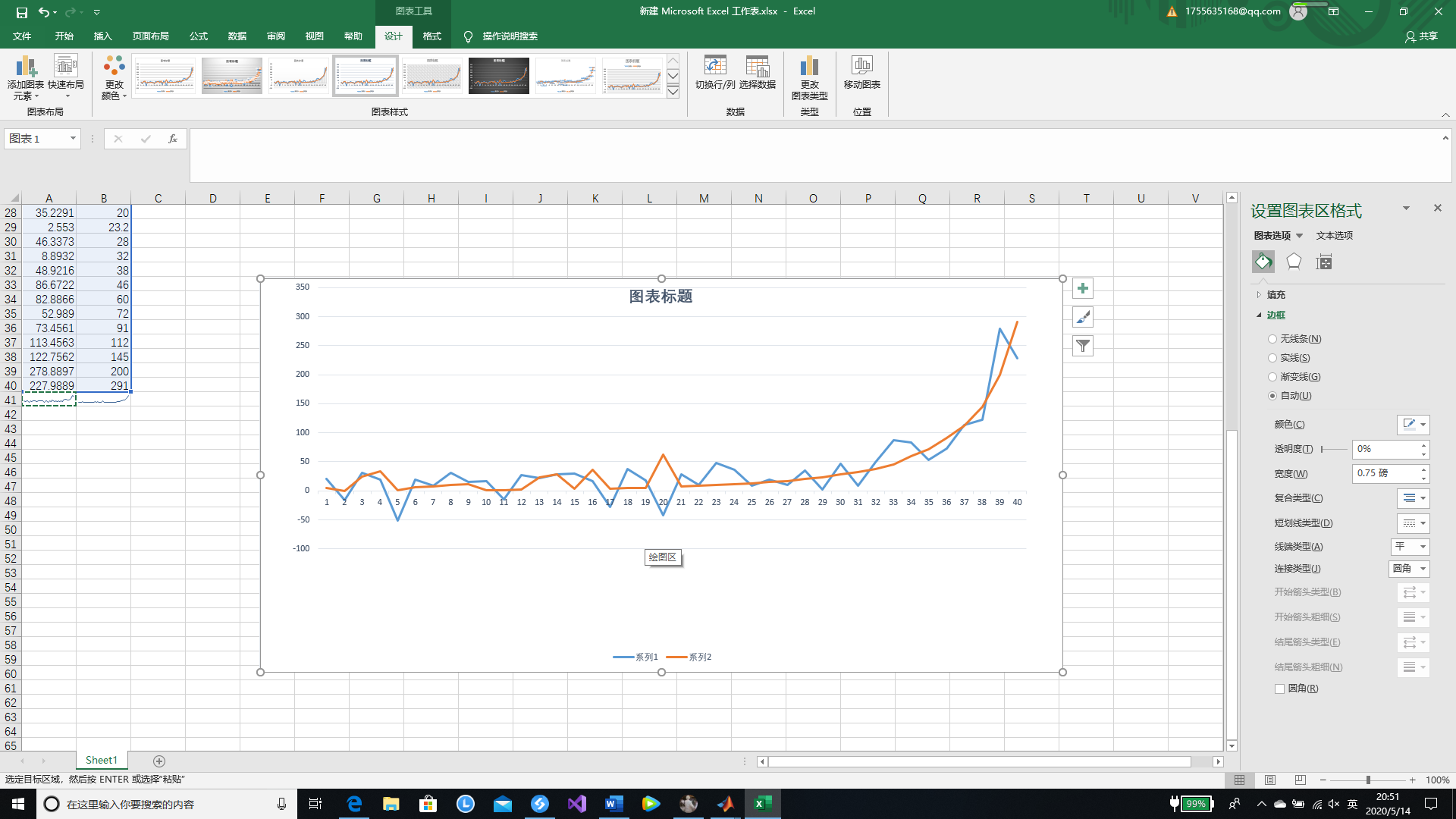
* 1. 问题挑战

网站上的信息会实时更新，由于采集不同页面的时间差，很可能会在两个页面采集到同一件饰品的信息，这为之后的数据清洗带来了不少麻烦。

而且为了减少目标服务器的压力以及防止被封禁，需要在采集两个页面之间设置1到2秒的延时，但是因为总共要采集数十万个页面，采集所需的总时间长达数天，这就使得数据的时效性难以得到保证。

* 1. 测试结果

我们经过MATLAB的拟合得到了一个较为复杂的函数，为了检测该函数是否能够较为正确的反映成交价格与官方定价以及售卖数量这三者的联系，我们又选出了一组测试数据，希望将理论结果与实际结果进行对比。经过对比得到下图，其中蓝色折线为理论的成交价格，而橙色曲线为实际的成交价格。不难看出两者基本上是拟合的，由此推断我们得出的函数具有较高的拟合程度。



经过测试，我们就可以很轻松结合之前分析出的诸多结论对于价格有一个清晰的预测，倘若利用这些进入交易平台进行购买，我们便可以大大减少被坑的可能性。

* 1. 角色分工

饰品信息及历史记录采集：杨天元(U201915108)

售价信息采集：周奥洋(U201915112)

数据清洗：邓至廷(U201915104)

* 1. 心得体会与总结

在对网站进行采集之前，需要对网站的整体框架结构进行分析，根据分析的结果来选取采集方法。当遇到用JavaScript渲染的页面，无法利用常规方法采集时，可以通过selenium来模拟浏览器的行为，或者分析后台数据来源，直接对后台文件进行采集。