基于 Tableau 和 Python 的股票分析报告

杨思颖 未央书院 未央-能动 01 2020012981

一、摘要

本文主要介绍了笔者小组在北卡州立大学数据科学项目的大作业中所做的研究工作。我们使用 Tableau 和 Python 作为工具，挑选两只游戏公司的股票 ATVI 和 EA 进行比对，对其进行了多种数据分析和可视化，并使用多种模型对股价进行拟合和预测。最终研究出两只股票的特性，并对比推荐出更适合投资的股票，并作汇报展示。

关键词：Tableau Python 数据分析 数据可视化 股票 模型拟合和预测

二、引言

在如今的信息时代，我们身边充斥着海量的数据，而这些大数据中有隐藏着很多有价值的信息，这些信息仅仅依靠观察这些冰冷而杂乱的数据是无法被有效捕获的。因此，数据处理和可视化就显得格外重要，在清晰易读的图表中，我们可以看到数据的分布、趋势等有效信息。而随着机器学习、深度学习等算法的发展，人们已经可以通过编程实现模型的搭建，甚至用模型去预测未来，对于当下的决定有指引作用。Tableau 是近年来兴起的数据可视化软件，能够快速生成多样的图表并实现交互式展示，Python 作为时下最普及的编程语言之一，拥有着很强大的功能，非常多开源的库支持着数据可视化和机器学习的实现。

在我参与的北卡州立大学数据科学训练营的项目中，教授向我们讲授了 Tableau 和 Python 的使用方法，并且布置了相关的作业要求我们个人完成，同时我们还需要以小组的形式完成一个大作业的最终汇报和展示。在大作业中，我们使用 Tableau 和 Python 作为工具，挑选了两只股票，并作数据分析、预测和可视化，将两只股票进行比较，推荐出更适合投资的股票，并做了汇报展示。本文主要介绍大作业的成果，小作业的成果将在文末简单说明。

三、研究内容及分析

（一） 两支股票的选取及数据介绍

考虑到游戏在年轻人中的流行，我们选取了两家游戏公司的股票（美股）进行分析。第一家动视暴雪（Activision Blizzard，简称 ATVI），是全世界最大的游戏开发商和发行商，旗下推出过多款经典系列作品，例如魔兽世界、星际争霸等等。第二支股票是美国艺电公司

（Electronic Arts，简称 EA），是全球著名的互动娱乐软件公司，旗下游戏包括极品飞车、战地、星球大战等。

我们的数据来源于 https://finance.yahoo.com/quote，选取了自 2000 年到 2019 年疫情前的数据。数据包括日期、开盘价、最高价、最低价、收盘价、调整后的收盘价、成交量。下图是数据示例。

Date	Open	High	Low	Close	Adj Close	Volume
6/29/2010	19.00	25.00	17.54	23.89	23.89	18766300
6/30/2010	25.79	30.42	23.30	23.83	23.83	17187100
7/1/2010	25.00	25.92	20.27	21.96	21.96	8218800
7/2/2010	23.00	23.10	18.71	19.20	19.20	5139800
7/6/2010	20.00	20.00	15.83	16.11	16.11	6866900
7/7/2010	16.40	16.63	14.98	15.80	15.80	6921700
7/8/2010	16.14	17.52	15.57	17.46	17.46	7711400
7/9/2010	17.58	17.90	16.55	17.40	17.40	4050600
7/12/2010	17.95	18.07	17.00	17.05	17.05	2202500

图 1. 股票数据示例

（二） 使用 Tableau 进行数据可视化

2.2.1 基本数据可视化

首先，我们对两支股票的基本数据做了可视化，并进行了分析。

2.2.1.1 收盘价

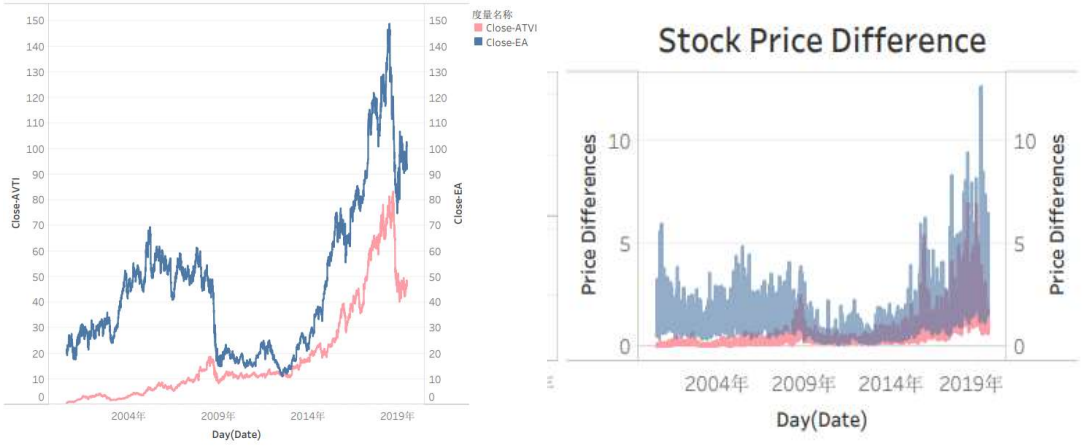


图 2. 收盘价及收盘价差异图

可以看出，二者整体呈现上涨趋势，EA 的股价比 ATVI 高。ATVI 股价持续走高，2018 开始有所下滑，19 年又呈现上涨趋势，而 EA 的股价在 2009 年至 2014 年一直处于低谷期，后飙升，在 18 年同样经历下跌，在 19 年又有上涨趋势。因此，以年为单位来看，EA 股价可能会经历几年的低谷期，但整体倾向于上涨，波动性较大，风险和机遇并存。而 ATVI 基本上持续涨高，适合长期投资，但套利空间不大。

2.2.1.2 成交量

从成交量可以看出 ATVI 的成交量明显更高，成交量高体现了 ATVI 股票更加活跃。

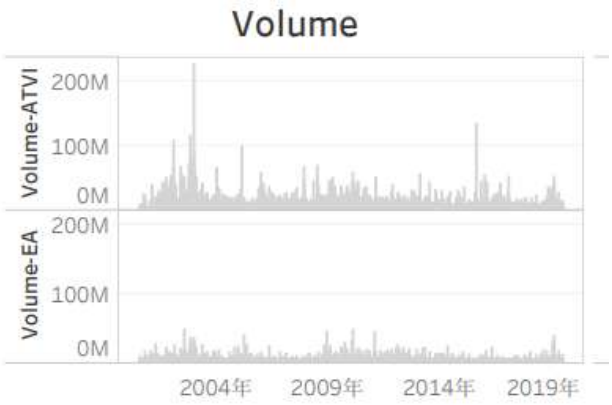


图 3. 成交量图

2.2.1.3 收益率

在基本数据上稍做处理，得到收益率=（收盘价-开盘价）/开盘价，并且分别绘制了年收益率（平均及中位数）和月收益率（平均及中位数）。EA 盈利的年份更多，且波动性更大，如果在合适的买卖点交易可能会赚取很多利润。

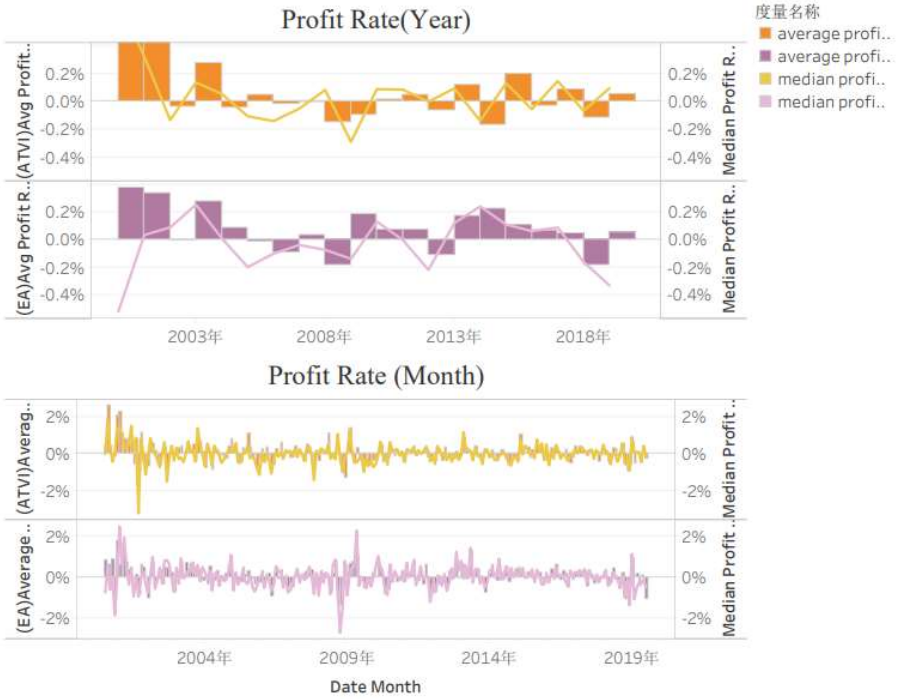


图 4. 利润率图

细看月度利润率图，把二者放在一张图中，可以发现两只股票有非常相似的趋势，因为他们都是游戏公司，受到同样的大环境影响。而紫色所代表的 EA 的确有更高的波动率，适合低买高卖，且近年来 EA 也有更多的盈利日和更高的收益率。

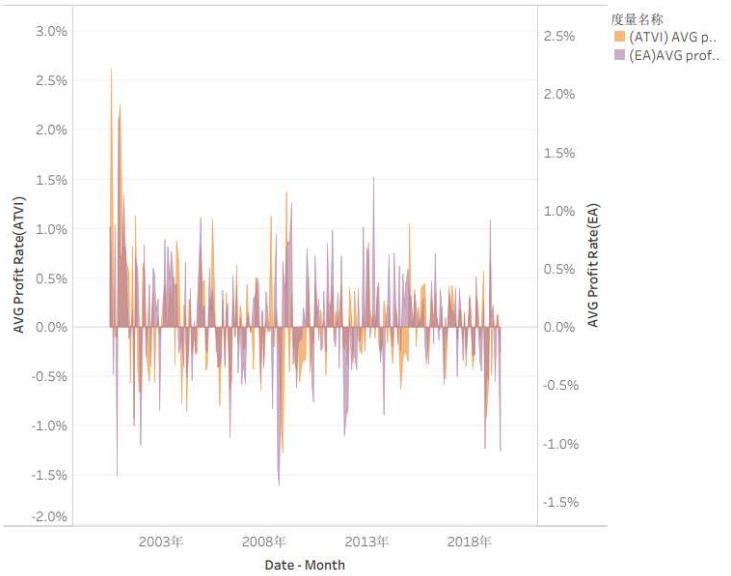


图 5. 月度利润率对比图

对于这 20 年间的每日利润率，我们也做了饼图统计，得到的结果是两只股票的涨和跌都在总数中各占一半。因此此指标没有很大的参考意义。

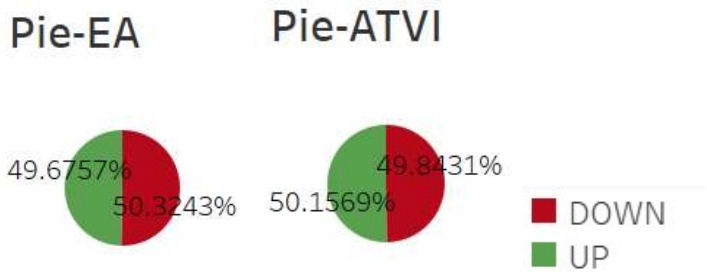


图 6. 日利润率饼图

2.2.2 进一步的指标分析

2.2.2.1 布林带（Bollinger Bands）

布林带是根据统计学中的标准差原理设计出的。由三条轨道组成，上下两条线可以看成价格的压力线和支撑线，算法是首先计出过去 20 天平收市价的标准差 SD，通常再乘 2 得出 2 倍标准差， Up 线为 20 天平均线加 2 倍标准差， Down 线则为 20 天平均线减 2 倍标准差。中间的灰线是 20 天平均价（就是后面要介绍的 MA），蓝线是收盘价，蓝线中的某些部分（当股价平均价不在常规的上下轨之间）用红色标记了，是需要注意的点。

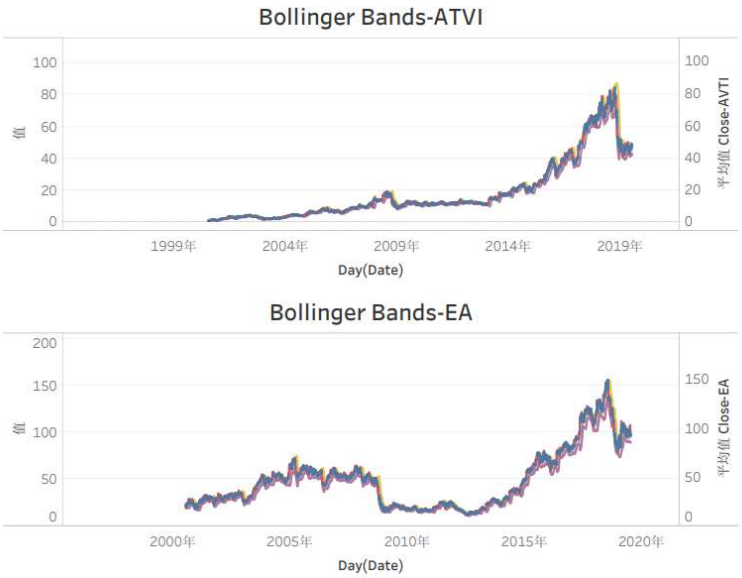


图 7. 布林带分析图

使用数据选取功能，筛选放大 2019 年数据，观察 2019 年末的布林带，ATVI 的上下轨和均线都呈现上涨态势，意味着股价短期内是上涨的，很适合做短期投资；EA 的收盘价比 MA 高，但是呈现下降趋势，因此短期投资中内应当卖出。

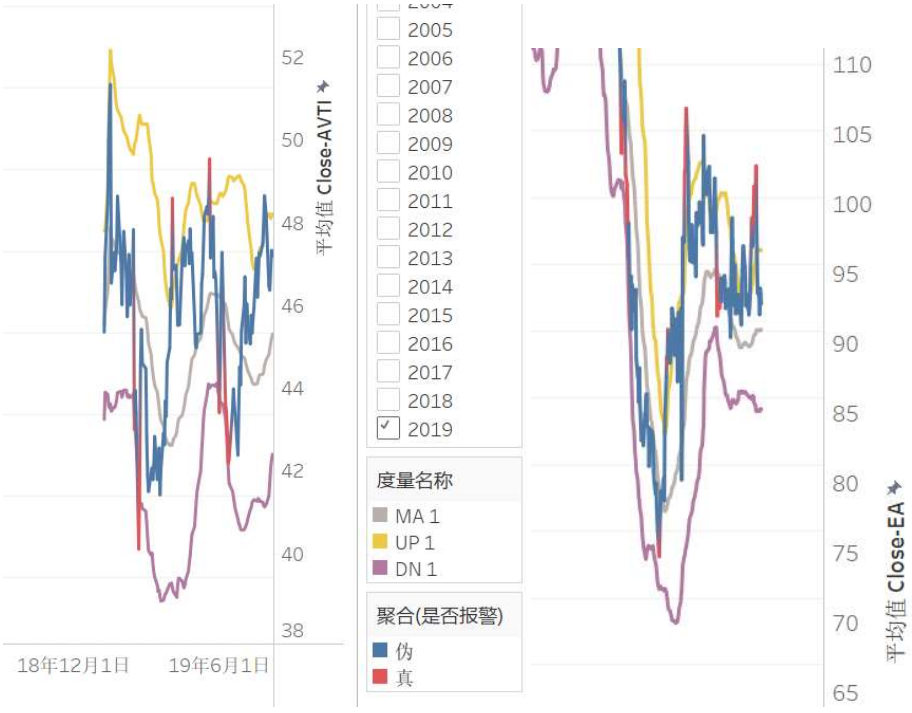


图 8. 2019 年末布林带

2.2.2.2 K 线图（Candlestick chart）

K 线图因为其形状类似蜡烛而又别名蜡烛图，是非常典型的股票指标。我们利用 Tableau 对数据进行叠加绘制，复刻了 K 线图。长方形的上下底边代表着开盘和收盘价，在美国绿色

代表着当天收盘价比开盘价高，红色反之；穿过长方形的上下影线的端点代表着最高价和最低价。

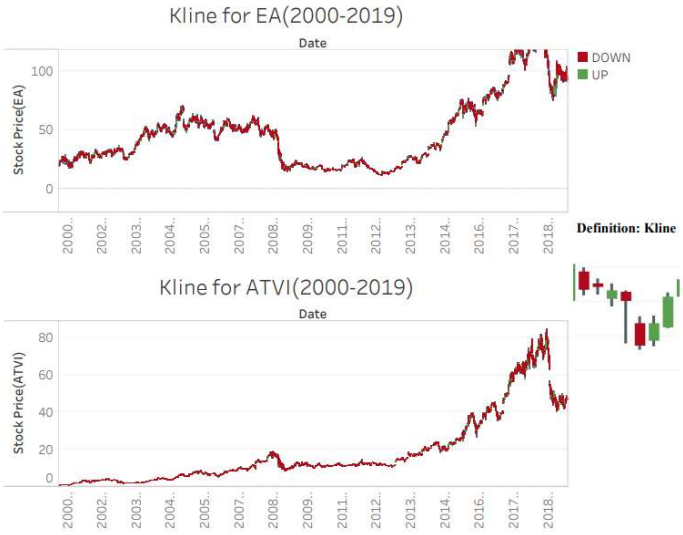


图 9.20 年间 K 线图

仅仅看最近的数据，2019 一年的 K 线图会更加清晰，二者都波动的很厉害，但是如果看一小段时间，比如 3 月 13 号到 3 月 26 号，我们可以看出 ATVI 有比较明显的涨跌趋势，而 EA 的上上下下太多了，以至于没有清晰的趋势。因此，ATVI 比较适合做短期投资，因为短期内的波动没有那么大，趋势比较明显，利于做出判断。但是 EA 的波动太大了，很难做判断，所以短期投资的收益可能不会很好，除非有技术找到正确的买卖点，那么就有很大的套利空间。

图 10.2019 年 K 线图

当然，在 K 线图中还有非常多的学问，是股民们津津乐道的，可以看阴阳、影线长短、十字星等等，不同 K 线的形态都对应着不同的趋势，在此不再赘述。

2.2.2.3 移动平均（MA）

移动平均(Moving Average)是显示价格走向的重要指数，且这个趋势一旦形成，将在一段时间内继续保持，因此我们可以看出买卖机会。MA 的核心是做过去 N 天的股价计算，想做短期观测可以考虑以 3，6 或者 24 天为一组。我们想做长期观测，因此考虑 120 天为一组。由于我们不考虑疫情，故假设我们处于 2019 年年底，根据 MA 的判断规则：移动平均线从下跌趋势中变平，并略微上升，而股价突破移动平均线以上，这就是买入信号。在 EA 的 2019 年年底我们看到了这个趋势，因此适合购买，而 ATVI 的移动平均线一直在下跌，这并不是一个好时机。

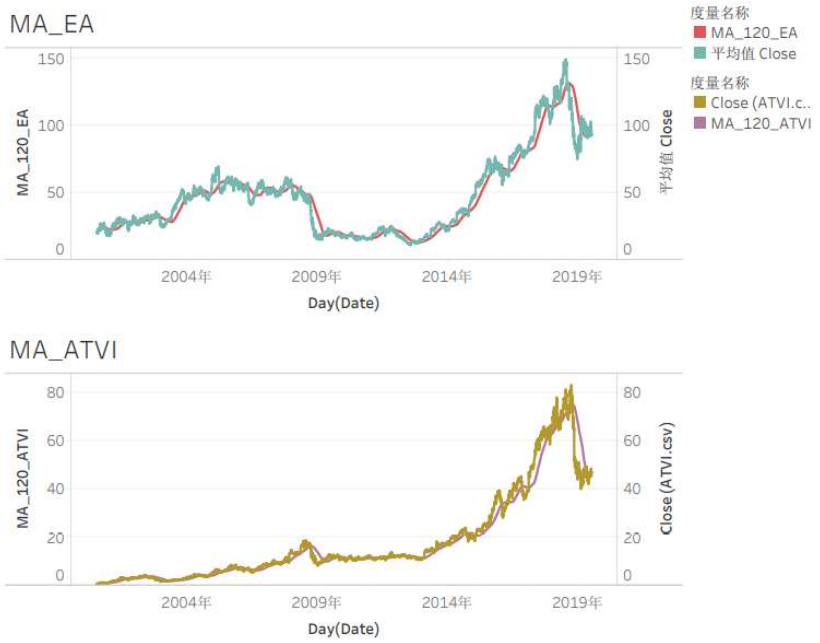


图 11.移动平均线 MA

2.2.2.4 多空指标（BBI）

BBI 是 MA 的升级版本，它更适合做短期预测。BBI 指标线是由多条移动平均线的加权平均值，一般情况下，选用的是 3 日平均价、6 日平均价、12 日平均价与 24 日平均价的和的平均值。判断标准和 MA 是很类似的。2019 年末的 BBI 指标可以看出，EA 有下降趋势，而交叉点在 ATVI 的 BBI 上出现了，因此 ATVI 此时比较利于进行短期投资。

图 12. BBI

2.2.3 预测

接着，使用 Tableau 自带的预测功能，只需要点击一下就可以生成预测线条。同时我们还对之前的股价做了拟合。二者在未来几年都是增长的趋势，但是 ATVI 的斜率更抖，可能增长更快。Tableau 并不适合做精确预测，只能预测大致趋势。

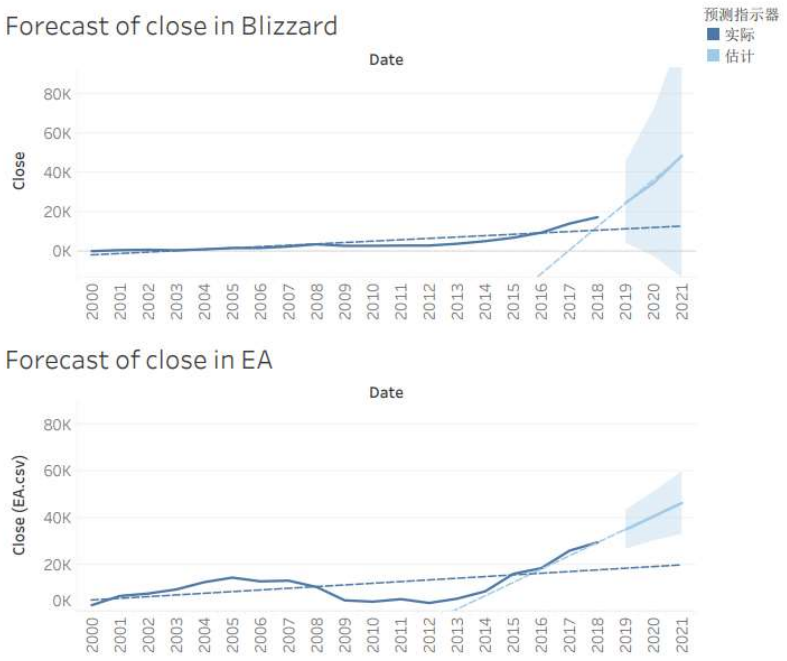


图 13. 使用 Tableau 预测未来股价趋势

2.2.4 总结

总得来看，二者在长期内都呈现出增长的趋势。ATVI 波动性较小，比较稳健，但收益也不会那么大，适合偏保守的投资者。EA 波动性较大，有很大的套利空间，但风险与机遇并存。

2.3 使用 Python 进一步分析

2.3.1 基本回归

首先做最基本的回归，观察股票基本的趋势。

2.3.1.1 对成交量线性回归

选取 2018.12.20-2019.7.12 的成交量，对不同阶段分段进行线性回归，拟合结果如下。ATVI 的成交量在最后呈下降趋势，EA 呈现上涨趋势。

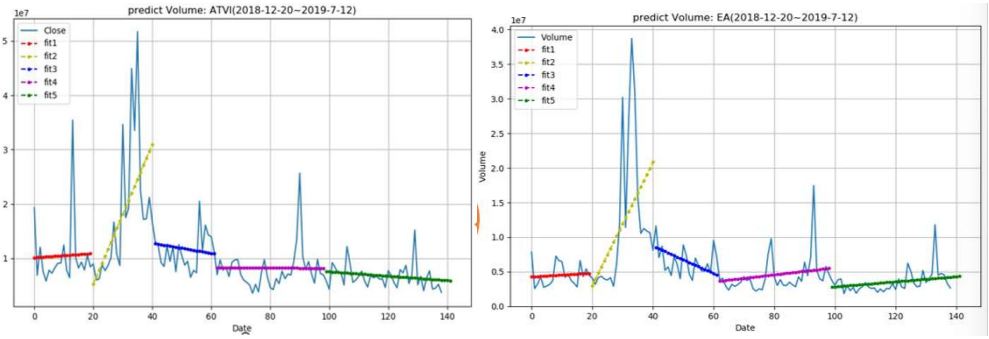


图 14. 对成交量线性回归

2.3.1.2 对收盘价回归

首先我们对这 20 年间的收盘价进行线性拟合，能看出一个大致趋势，但很不准确。

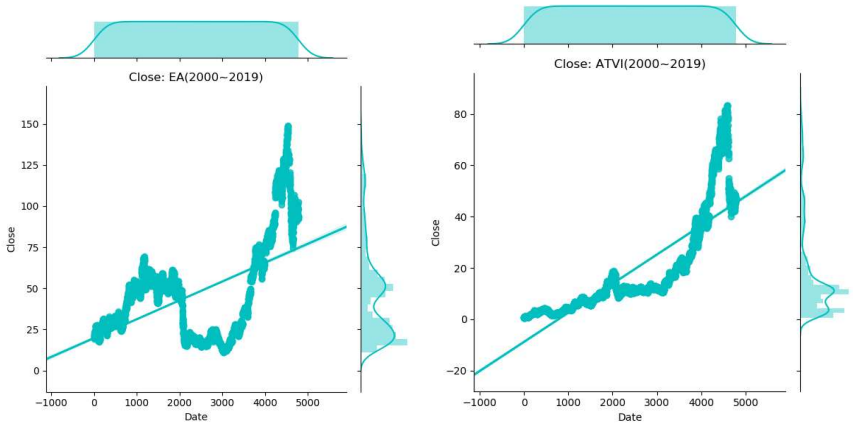


图 15. 20 年收盘价线性回归

选取 2015 以前的数据用来训练线性 LinearRegression 模型，用 2015-2019 的做验证。如下图所示，绿色的预测值和橙色的实际值相差很大，线性回归对于这么长的一段时间还是太过简单了。

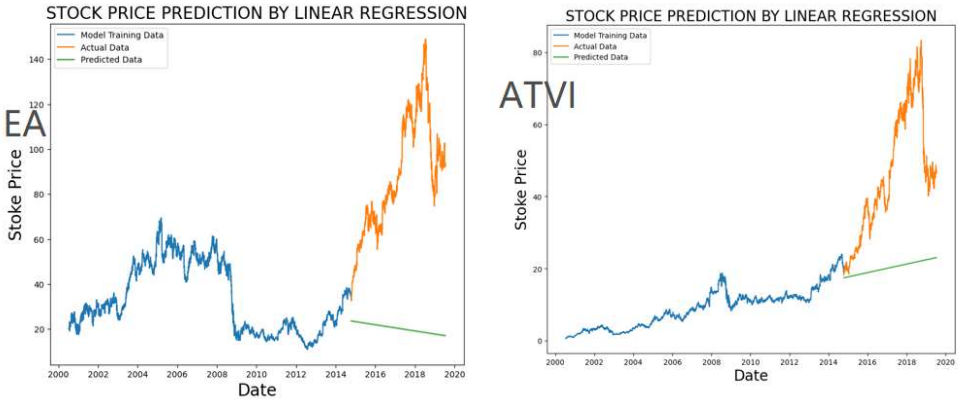


图 16. 简单线性回归预测

于是尝试四阶回归，发现拟合程度明显变好。

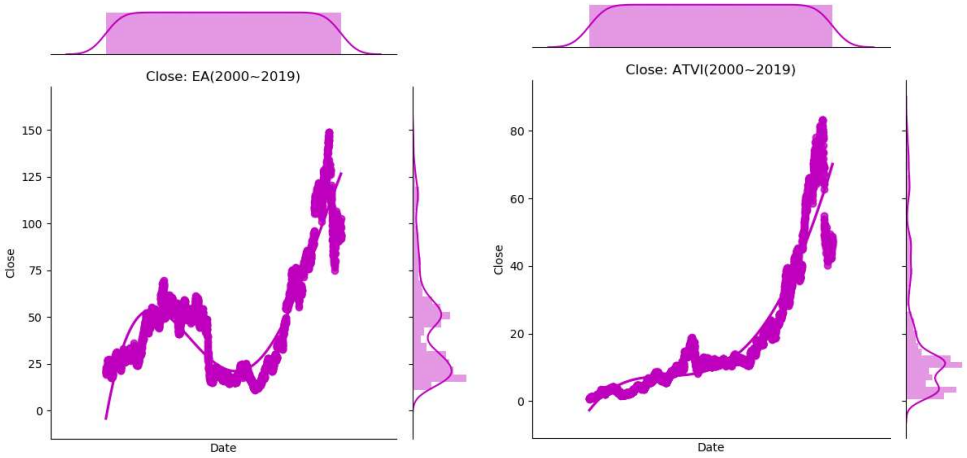


图 17. 四阶回归

接着选取 2018.12-2019.7 半年的收盘价，分段进行线性拟合，可以得到如下结果。

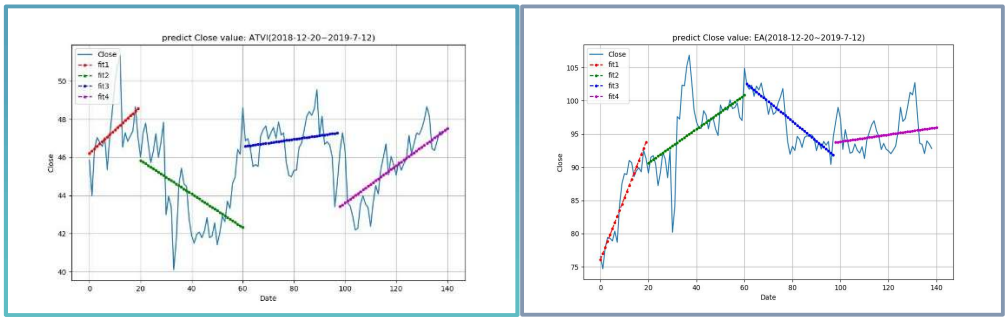


图 18. 对半年的收盘价分段线性拟合

再选取我们数据中最新的一个月，2019.6.10-2019.7.12，做线性拟合，结果如图。

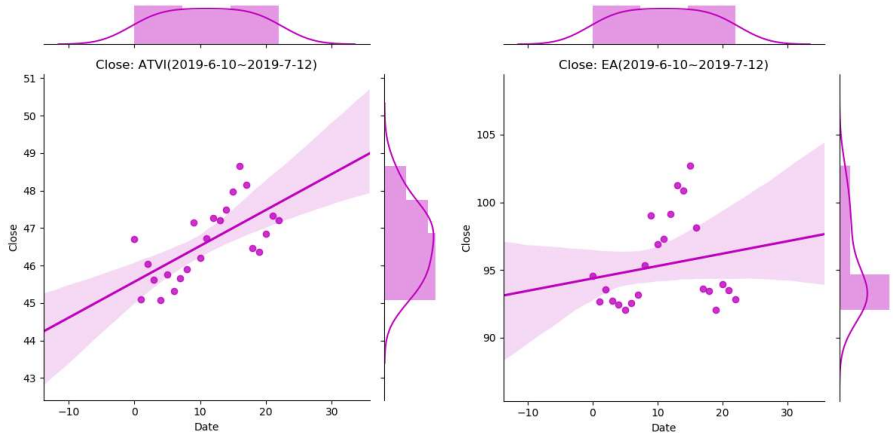


图 19. 对近一个月做线性回归

从以上简单回归的结果，可以发现简单回归只能体现一个大致的趋势，且回归的时间越长，差距越大，因为收盘价的波动已经不足以只用简单回归来表示了，因此以上回归除了告知趋势之外并没有很大的参考意义。

2.3.2 更进一步的模型拟合

2.3.2.1 多元线性回归(Multiple Regression)

与简单线性回归不同，多元回归是由多个自变量拟合出的方程，因此效果应该比简单一元线性回归要好，看一下结果，发现预测结果和实际值还存在不小的差异，因此这种方法或许不是一个好选择。

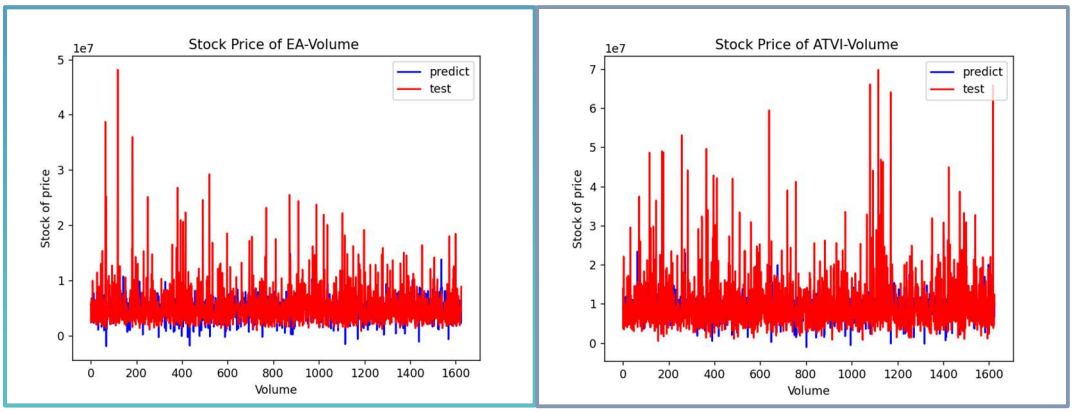


图 20. 多元回归

2.3.2.2 多项式回归（Polynomial Regression）

使用多项式回归模型拟合，在训练集部分的回归如橙线所示，之后的绿线为用于回测的部分，蓝色为真实数据，相差较大，此模型还是太过简单，只能体现基本趋势。

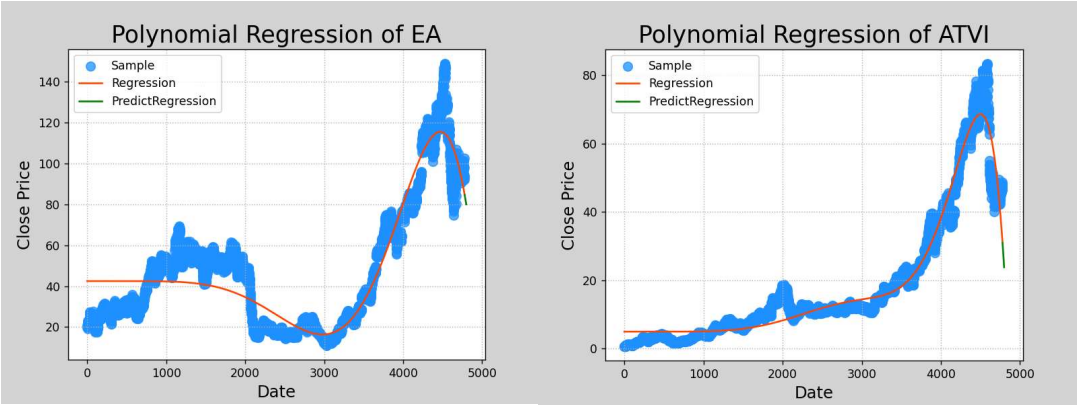


图 21. 多项式回归

2.3.2.3 AutoTS

AutoTS 是专门用来预测时序数据的，使用 AutoTS 预测数据结束后 10 天的数据，如下图所示。

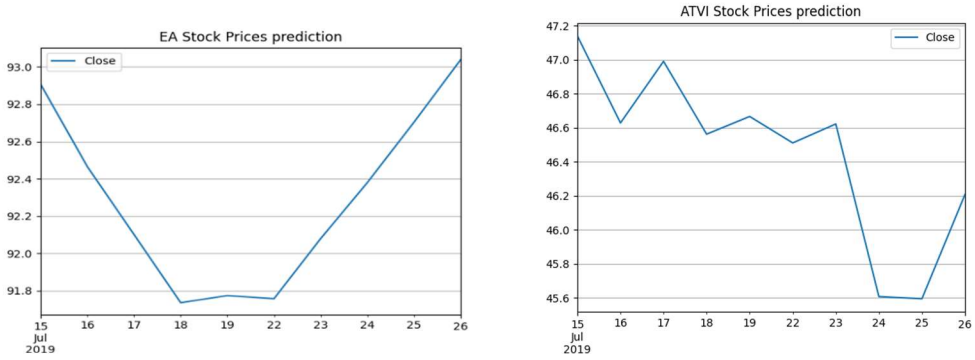


图 22. AutoTS 预测

2.3.2.4 LightGBM

经过多种模型的尝试和比对，我们最终认为 LightGBM 是准确度最高的预测模型。在此进行较为详细的介绍。LGBM 是基于决策树算法的分布式梯度提升框架，使用直方图算法将特征离散化，不像往常的 GBDT 模型的按层生长（level-wise）的决策树生长策略，LGBM 使用带有深度限制的按叶子生长策略（leaf-wise），找到当前分裂增益最大的叶子进行分裂，可以更快且精度更高。

首先我们以 2000-2017 年的数据为训练集，2018-2019 的数据为验证集，设置参数、训练模型，并在验证机上测试模型的精度。

```
model=lgb.LGBMRegressor(max_depth=8,
                        num_leaves=20,
                        n_estimators=500,
                        learning_rate=0.1)
model.fit(x_train,y_train)
```

图 23. LGBM 模型参数设置

关注验证集的表现，绿线是实际收盘价，橙线是预测收盘价，可以看到预测和实际相差非常非常小，尤其是最后的部分，拟合得非常好，MSE(mean squared error)也只有 0.21，因此 LGBM 模型的效果很优秀。

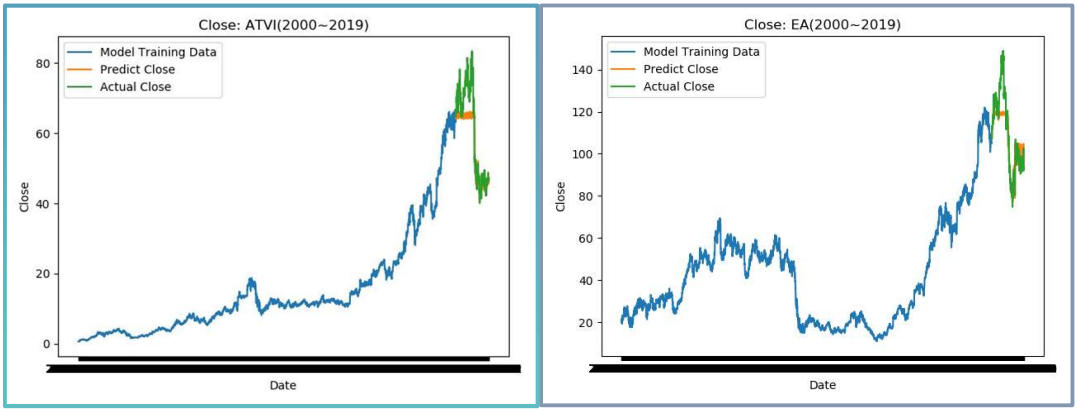


图 24. LGBM 模型效果

因此，我们可以使用 LGBM 模型来往后预测数据结束后的 20 天，对于 ATVI，预测的结果是股价上涨，因此如果我们是在 2019 年末，我们现在可以买入了，但 EA 的预测结果显示要跌，此刻买 EA 并不是一个好主意。

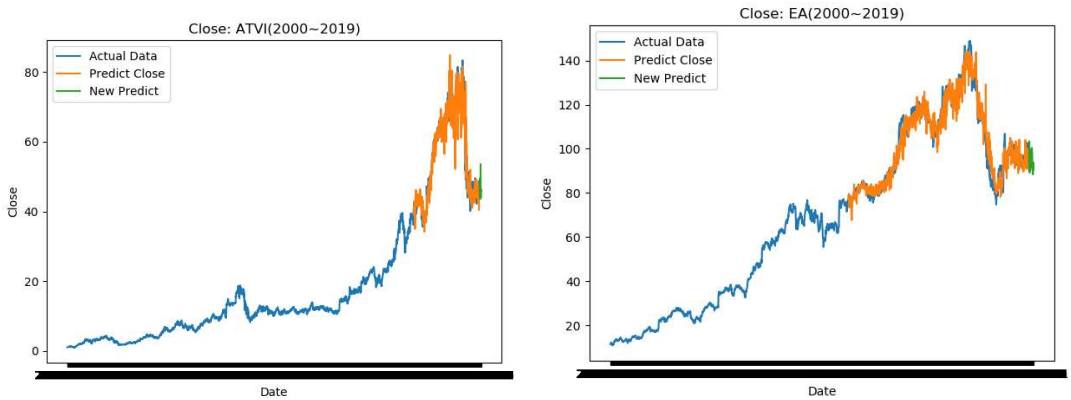


图 25. LGBM 预测未来 20 天

2.3.3 结论

我们使用了基本的回归和更进一步的模型来对股价做预测，基于基本的回归告诉我们，在以年为单位的长跑里，EA 和 ATVI 都会涨，且 ATVI 涨势更好。在未来的短期内，根据最为准确的 LGBM 模型，ATVI 将要涨，而 EA 将要跌，所以此时买 ATVI 是一个好选择。

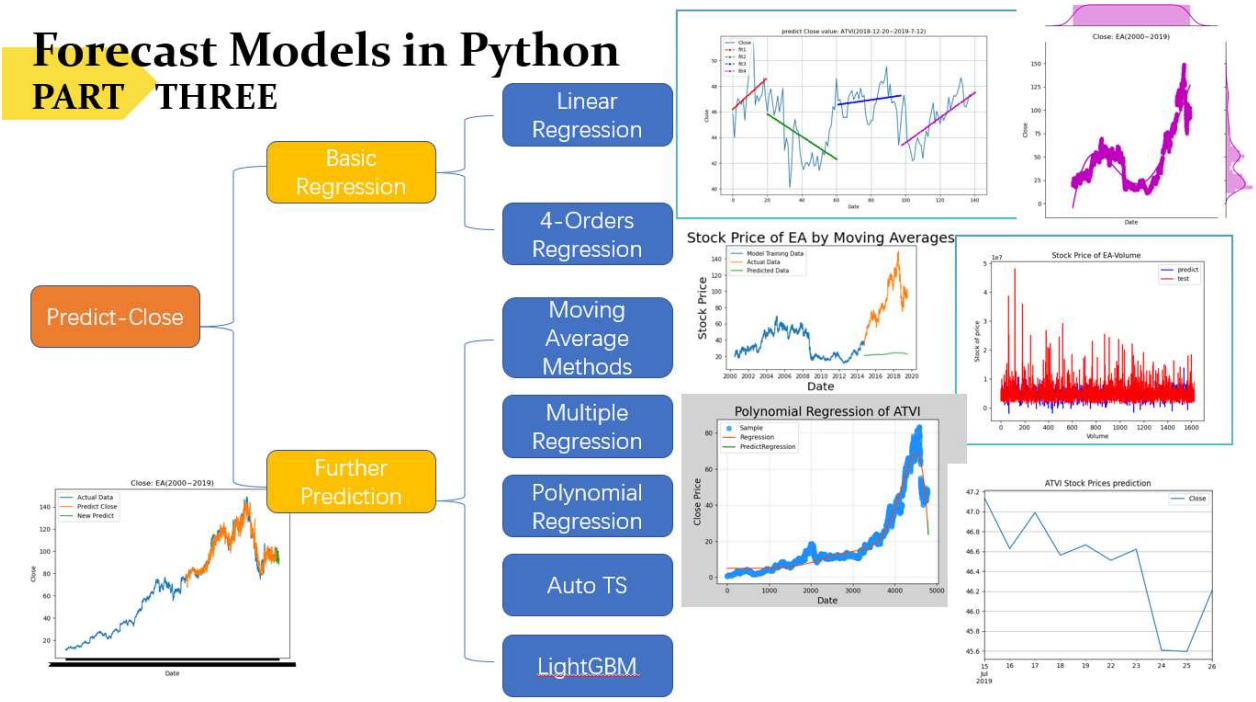


图 26. Python 数据分析所用方法总结

四、讨论及分析结果

4.1 对 Tableau 和 Python 的比较

Tableau 是专门用于数据可视化的软件，在做描述性分析时很有优势，操作简单，动动手指就可以生成非常精美的图表，且在展示时更是有交互式的互动效果，让数据可视化更加清晰易读。Python 在数据可视化方面，有很多的库可以做到，例如 seaborn 等，但是需要编辑代码，比 Tableau 麻烦一些，但是绘制出的图片也很精美，也可以有更多的自定义空间，让图表更好的满足用户需求，但是缺少交互的功能，不能在展示时随意放大、显示点、筛选数据等等，因此交互性不够好。

在数据预测方面，Tableau 提供预测功能，非常简单，只要点击选项即可，但非常不准确，只能是一个大致趋势。但是 Python 可以使用多种开源的库，使用机器学习、深度学习等方法，来实现模型的拟合和预测，效果更佳的精准，能够真实使用到现实生活中用于选股等等，对于复杂的数据也能达到很高的精确度。因此 Python 在更复杂的数据处理上是做得非常非常好的。

4.2 股票分析总结

对于股票本身的特性：成交量 ATVI 更大；ATVI 波动性较小，EA 波动性较大。因此 ATVI 更加稳健，但是套利空间较小；EA 风险与机遇并存，高卖低买策略若是成功，则可以赚取很多的利益。

假设我们是处于 2019 年末的投资者（由于疫情不考虑近年，教授的要求），考虑未来短期 20 天左右的涨跌：ATVI 将要上涨，适合买入；EA 将要下跌，适合卖出。在未来的长期里，二者都会上涨，但是 ATVI 的上涨幅度会更大。

因此，综上所述，我们建议现在买入 ATVI。

五、补充说明

以上介绍的是在项目大作业中所做的工作，除了项目大作业之外，还有 Python 和 Tableau 的小作业。

在 Tableau 小作业中，根据老师提供的销售数据，我需要提出四个问题，并且相应地绘制出能解释这四个问题的四张不同类型的可视化图表。以下是我的四张可视化图表，在图表上已标注问题。

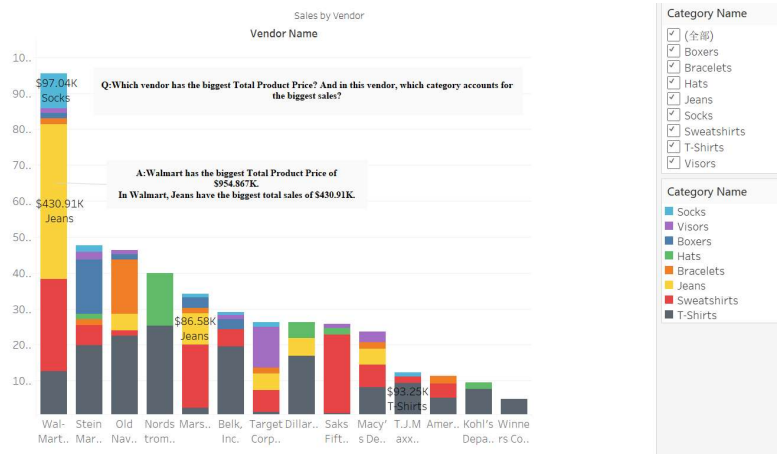


图 27. Tableau 问题一

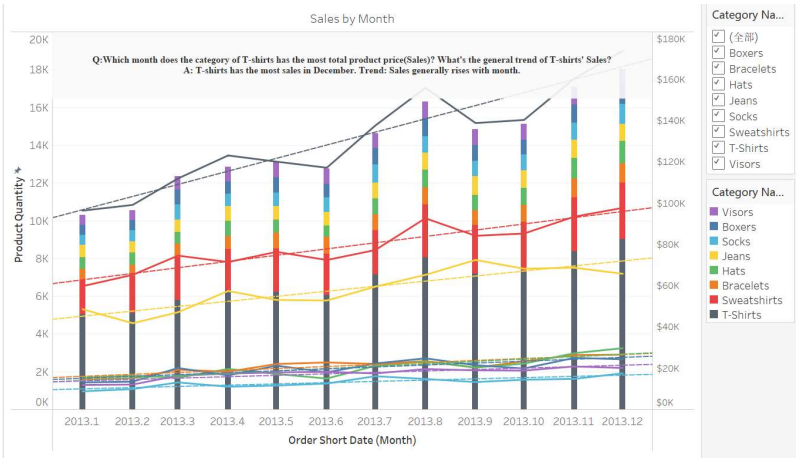


图 28. Tableau 问题二

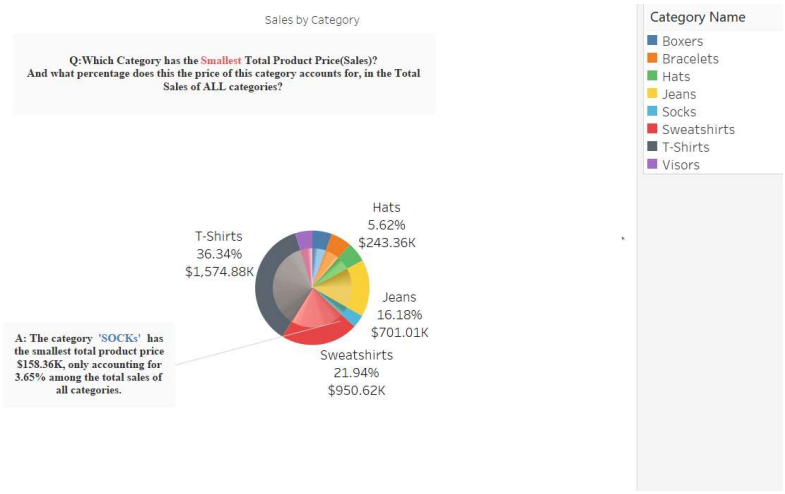


图 29. Tableau 问题三

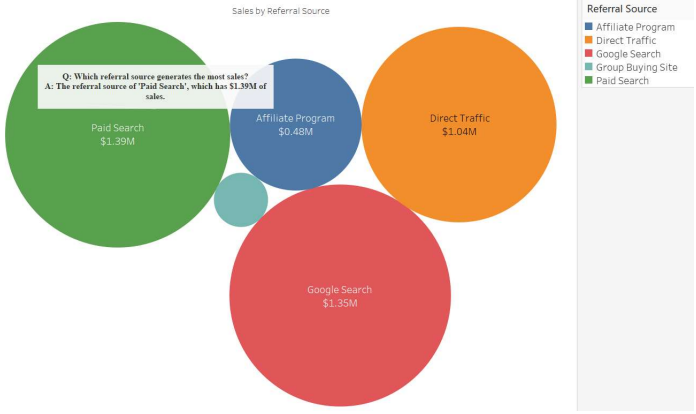


图 30. Tableau 问题四

在 Python 小作业中，需要根据老师的要求，实现文件的读取、一些统计量的呈现、相关性的可视化、价格分布的可视化等。较为简单，在此不做赘述，仅将结果做一个简单呈现。

a. Print head for top 10 records

	order_id	customer_id	...	category_name	product_color
0	59929	11914	...	T-Shirts	Blue
1	59929	11914	...	Sweatshirts	Purple
2	59966	14644	...	Sweatshirts	Yellow
3	59973	23186	...	Jeans	Green
4	59973	23186	...	Visors	Purple
5	59935	7437	...	T-Shirts	Orange
6	59935	7437	...	T-Shirts	Blue
7	59980	23193	...	T-Shirts	Blue
8	59916	360	...	T-Shirts	Red
9	59916	360	...	Boxers	Orange

[10 rows x 13 columns]

b. Sum,mean,max,min for total product price

	order_id	customer_id	...	item_product_price	total_product_price
count	47301.000000	47301.000000	...	47301.000000	47301.000000
mean	45395.930636	10806.926492	...	25.748725	91.612787
std	8354.478877	7477.191597	...	12.831781	100.521693
min	31022.000000	1.000000	...	13.530000	13.530000
25%	38134.000000	3097.000000	...	19.110000	25.720000
50%	45313.000000	11844.000000	...	21.680000	57.390000
75%	52622.000000	17345.000000	...	27.220000	124.320000
max	60011.000000	23224.000000	...	74.440000	893.280000

[8 rows x 5 columns]

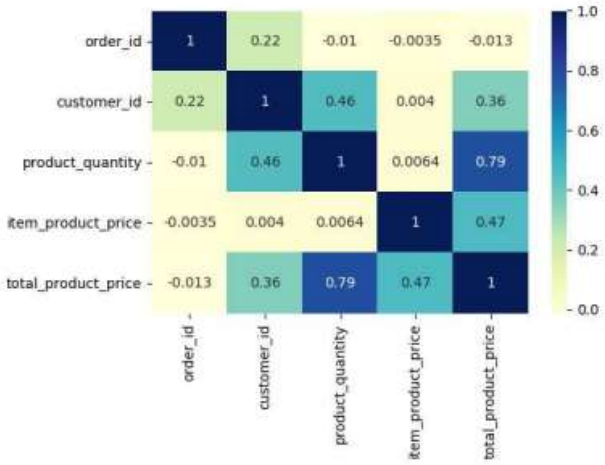
Sum for total product price is: $4,333,376.44
Mean for total product price is: $91.61
Max for total product price is: $893.28
Min for total product price is: $13.53

c. Show correlation

Correlation is:

	order_id	...	total_product_price
order_id	1.000000	...	-0.013055
customer_id	0.218983	...	0.360153
product_quantity	-0.010340	...	0.792602
item_product_price	-0.003488	...	0.465881
total_product_price	-0.013055	...	1.000000

[5 rows x 5 columns]



d. Histogram for item product price

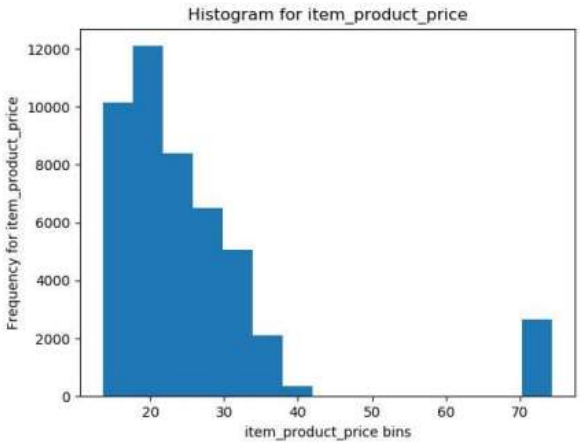


图 31.Python 小作业

六、资金使用情况

申请 5000 元的资金支持，全部用于学费和注册费。（总的学费是 1280 美元，当时实付 8770 元人民币）