管理学院专业学位大作业

课程名称《商务大数据分析》

学生行为数据分析报告

专业名称：_____

班　　级：_____

学　　号：_____

姓　　名：_____

任课教师：　刘跃文 副教授

提交时间：_____

摘要

本报告的目的是给西安交通大学管理学院《商务大数据分析》课程的学员提供一个课后大作业模板。本报告对大学生的基本情况、就餐情况进行了探索性分析；构建回归模型，对就餐数据与考试成绩之间的关系进行了简单研究。**参加2天课程的学员只需要参考一、二、三、五节，参加4天课程的学员可参考全篇报告。**

本报告发现，（1）女生平均成绩高于男生，但是男生的成绩范围更宽；（2）早餐次数、午餐次数等变量和成绩正相关，也就是说，选择在学校就餐的学生成绩更好；（3）早餐时间和成绩负相关，起早的学生成绩更好。

一、数据描述

本报告中所使用的数据，是一份根据真实消费记录虚构的学生基本情况及食堂就餐数据。分析本数据的主要目的，是教会西安交通大学管理学院《商务大数据分析》课程的学员开展探索性数据分析及简单的建模工作。

数据主要包括两个部分：学生信息（stu.csv）和就餐记录（trans.csv）。学生信息表样例如表1。该数据一共9列，107条。Gender字段0代表女性，1代表男性，本表中所有字段均是虚构的。

表1：学生信息样表

ID	Gender	T1	T2	Class	Group	Prov	City	BirthDay
94	0	76	74	1	8	河北	石家庄市	1996-6-10
95	1	86	87	1	8	湖南	郴州市	1996-7-10
96	1	81	83	1	8	北京	北京	1994-10-10
97	1	73	79	1	8	广西	南宁市	1996-3-11
98	1	77	73	1	9	辽宁	鞍山市	1996-6-11
99	1	84	82	1	9	湖南	长沙市	1994-8-11
100	1	87	82	1	9	河北	迁安	1995-10-11
101	1	58	55	1	9	湖北	武汉市	1995-4-12
102	0	89	88	1	9	山西	晋中市	1995-8-12
103	0	89	82	1	9	江苏	盐城市	1995-10-12

就餐记录样例如表 2 所示。该数据一共 6 列，93279 条。该数据是根据真实的就餐记录虚构的。其中，就餐时间（transtime）与金额（transvalue）是真实的，其它字段均为虚构字段。

表 2：就餐记录样表

stuid	campus	canteen	pos	transtime	transvalue
32	北	2	112	2014-10-31 7:34	2
32	北	2	100	2014-10-31 11:49	1
35	北	2	112	2014-10-31 16:21	2.1
46	北	2	65	2014-10-31 16:48	8
52	北	2	65	2014-10-31 16:48	8
54	北	2	65	2014-10-31 16:48	8
55	北	2	100	2014-10-31 18:36	1
56	北	2	112	2014-10-31 16:20	2.8
61	北	2	112	2014-10-31 7:44	2
72	北	2	89	2014-10-30 11:40	6
73	北	2	112	2014-10-30 7:38	1.5
74	北	2	112	2014-10-30 18:34	2.5
85	北	5	30	2014-10-30 12:08	2.5
87	北	2	112	2014-10-30 7:35	2
90	北	2	89	2014-10-30 11:43	6
90	北	2	65	2014-10-30 18:38	8

二、学生情况探索性分析

1、学生基本情况

如图 1 所示，本次数据分析一共涉及到 107 名学生，其中 92 名男生，占 85.98%。出生年份以 1996 年为主（60 名），其次是 1995 年（33 名）。出生月份 9 月份、10 月份、11 月份的相对较多，可能是因为小学入学年龄的分区时间为 8 月 31 日，9-11 月份的学生需要延迟一年入学，比别的学生年龄大，在入学之初表现较好，从而逐步累积了优势[1]。

[1] 由于出生年月是特别虚构的，这一分析结果并不是真实的，没有现实意义。关于年龄优势的讨论，可参考格拉德威尔《异类：不一样的成功启示录》一书。

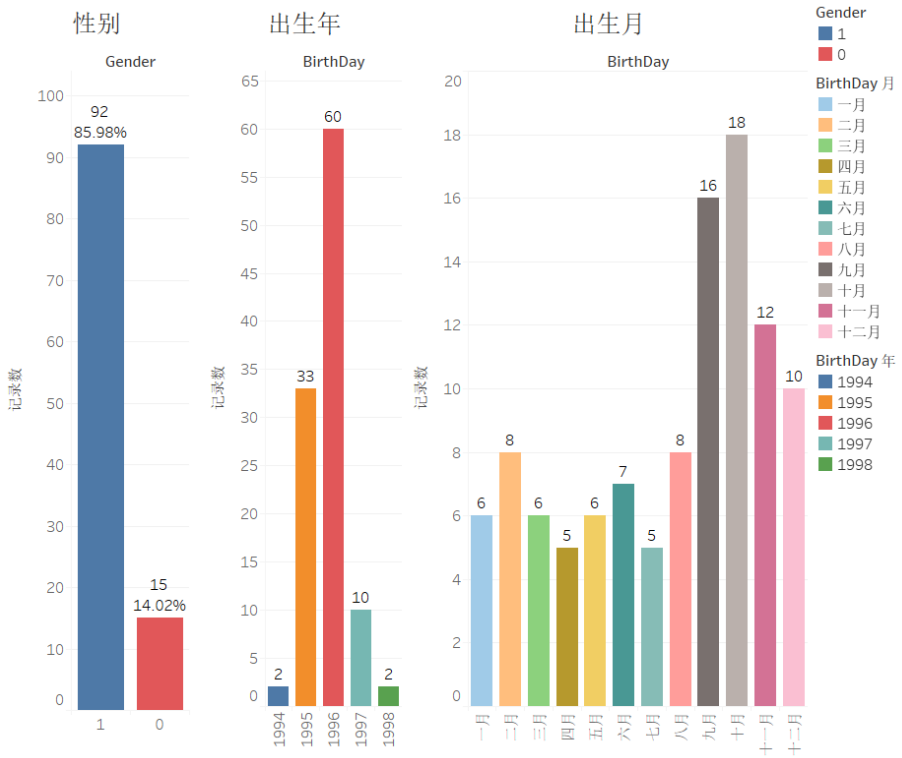


图 1：性别及生日

如图 2 所示，学生主要来自中国的东部地区，其中，来源较多的省份有北京、湖南、河北、江苏等地；来源较多的城市有北京、江苏南通、河北石家庄、湖北武汉等。图 2 的 4 个图分别是填色地图、打点地图、树形图、词云。可以看出，4 个图各有特色，可以用于不同的场合。

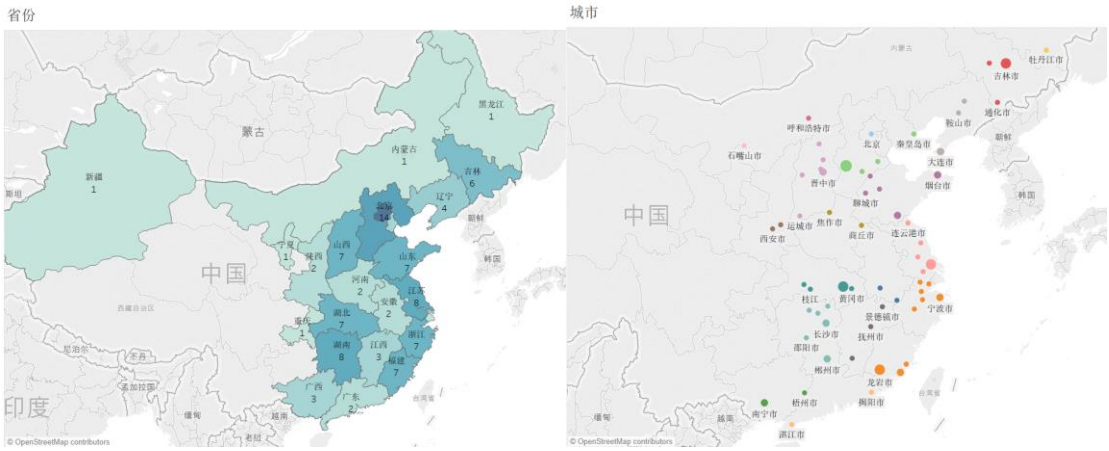


图 2：学生来源分布图

2、学生成绩情况

所研究的学生第一学期平均成绩为 78.32，第二学期平均成绩为 77.47。从图 3 的成绩分布来看，两学期成绩的分布相似，但是第二学期的最低成绩下降了不少，这是导致第二学期平均成绩下降的主要原因。

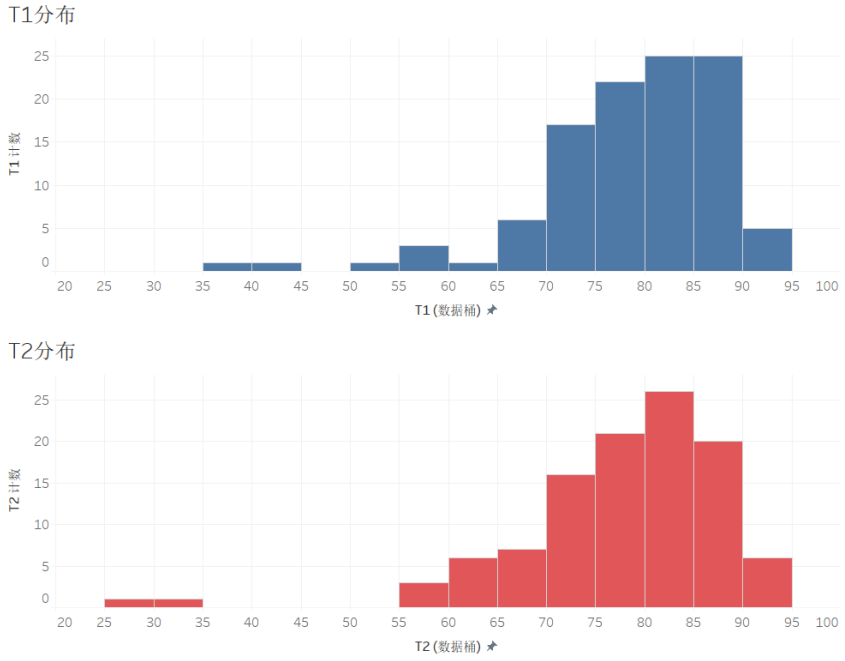


图 3：学生成绩分布图

那么是哪些学生造成了第二学期最低成绩下降呢？图 4 用散点图的形式展示了每位学生两学期的成绩，横坐标为第一学期成绩，纵坐标为第二学期成绩。

可以看出，91 号和 53 号学生第二学期成绩较低。其中，91 号学生成绩从 71 分直接降到 26 分，值得特别关注。

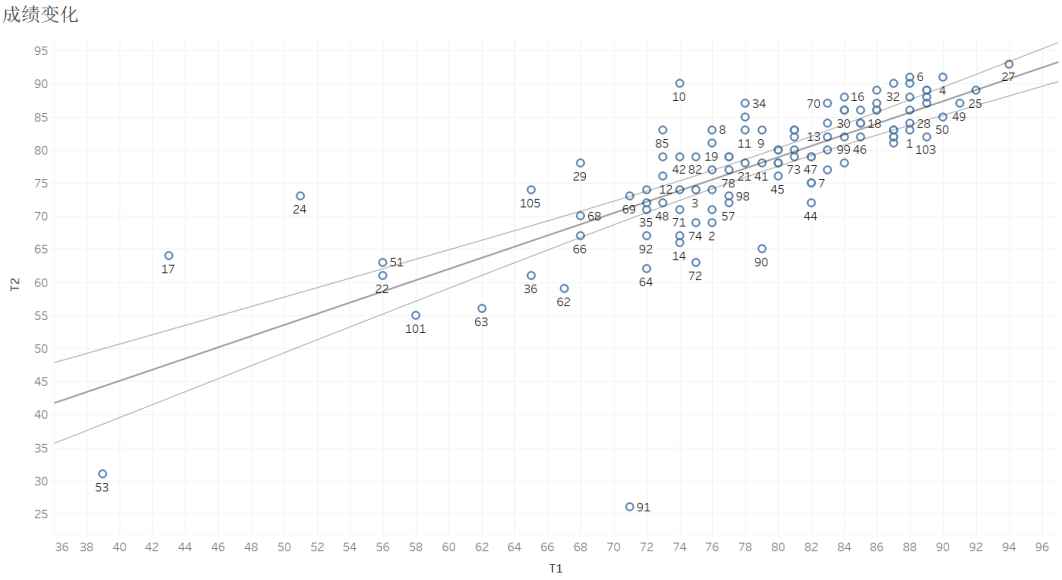


图 4：学生成绩散点图

图 5 对学生的成绩进行了分组比较。可以发现，女生的平均成绩高于男生的平均成绩。然而，箱线图显示，女生的成绩分布较为集中，男生的成绩范围更宽，也就是说，高分更高，而低分更低。

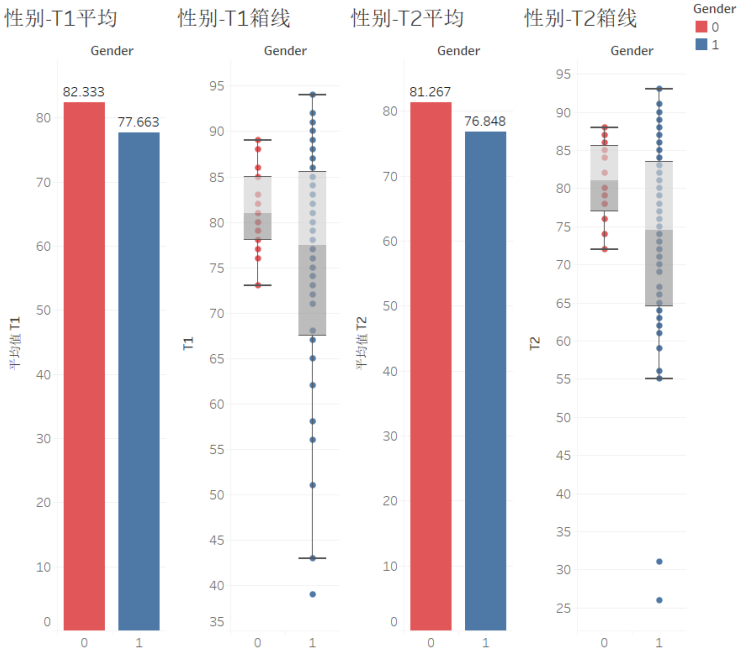


图 5：男女生成绩对比图

同样对 3 个班级的成绩进行比较，各班成绩变动比较大，没有固定的模式。

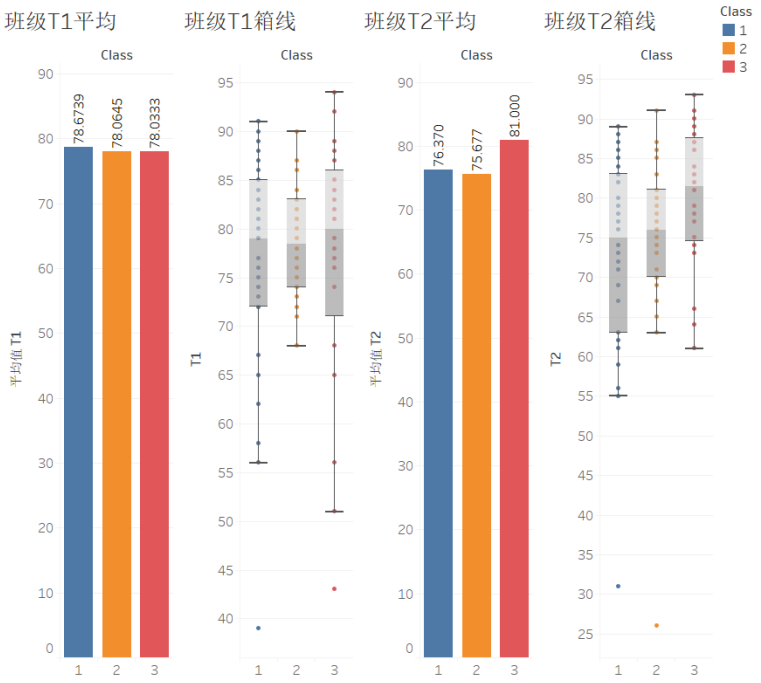


图 6：各班成绩对比图

三、就餐数据探索性分析

1、就餐行为规律性

对每日就餐人次进行分析，可以发现，周一到周五的就餐人次显著地高于周六和周日的就餐人次。节假日（如国庆节）会对就餐人次有显著影响。

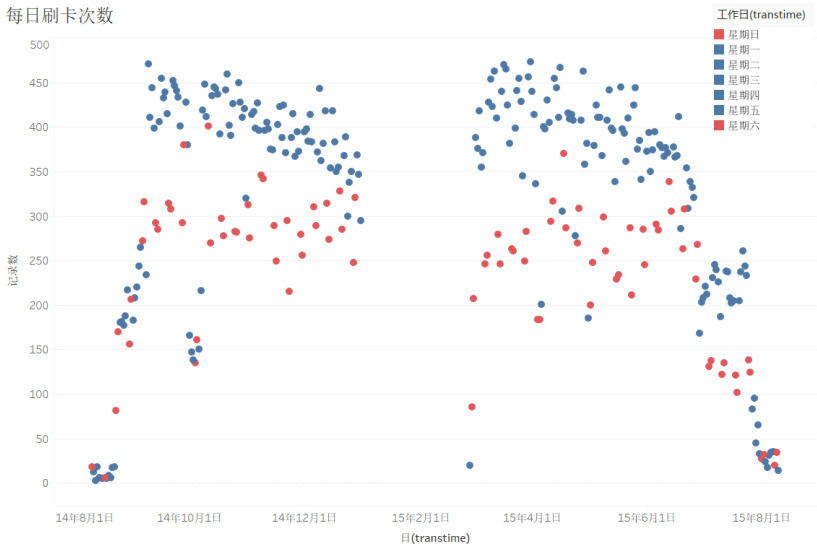


图 7：就餐人次

图 8 展示了本科生的就餐习惯。可以看出，早餐有 3 个高峰，分别是 7:05，7:40，9:55；午餐有 2 个高峰，分别是 12:05 以及 11:40。晚餐是 17:30 和 18:05。由于冬季和夏季作息时间的存在，导致 18:35 左右还有个小高峰。

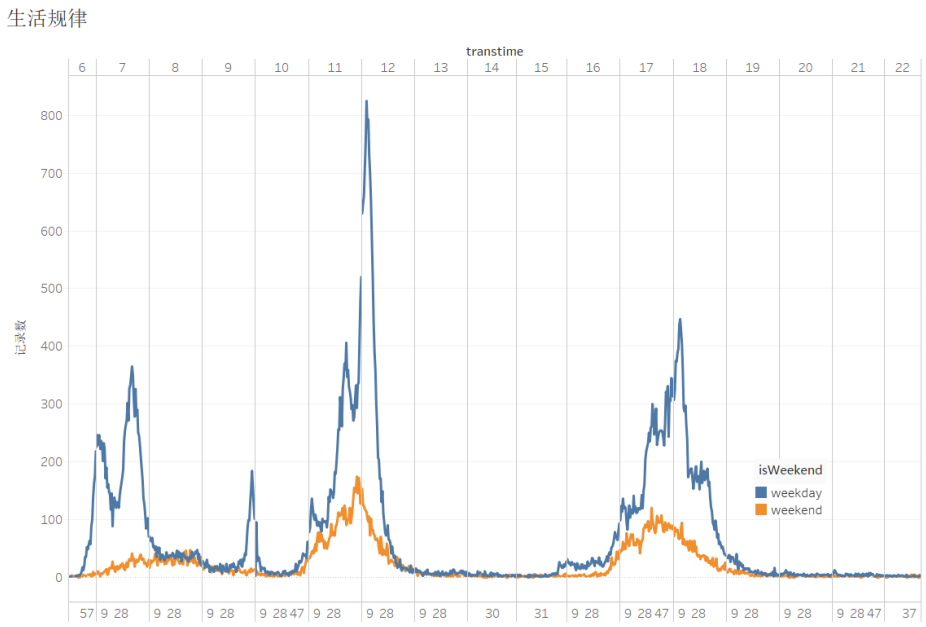


图 8：每日就餐规律

2、就餐偏好

对每个 POS 机的刷卡次数进行统计，可以看出，有一些 POS 机刷卡次数很多，同时也有一些 POS 机几乎没人光顾。进一步用二维表技术来分析每个 POS 机的工作时间（如图 10），横坐标表示时间，纵坐标表示 POS 机标号。可以发现，部分 POS 机没有稳定的销售行为，这是因为 POS 编号是虚构的。

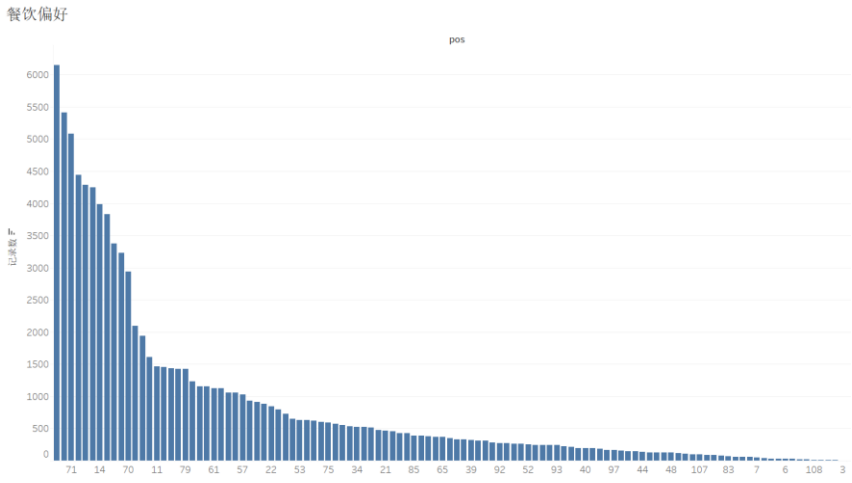


图 9：POS 机的就餐人次

图 10：各 POS 机每日刷卡数据统计

3、学生就餐分析

分析每个学生每日的就餐情况，可以看出，学生就餐行为有一定的规律性。例如，寒暑假就餐学生较少；国庆节就餐情况较少。同时，也发现了若干学生，在学校就餐次数较少。关联其成绩，可以看到其成绩较差。

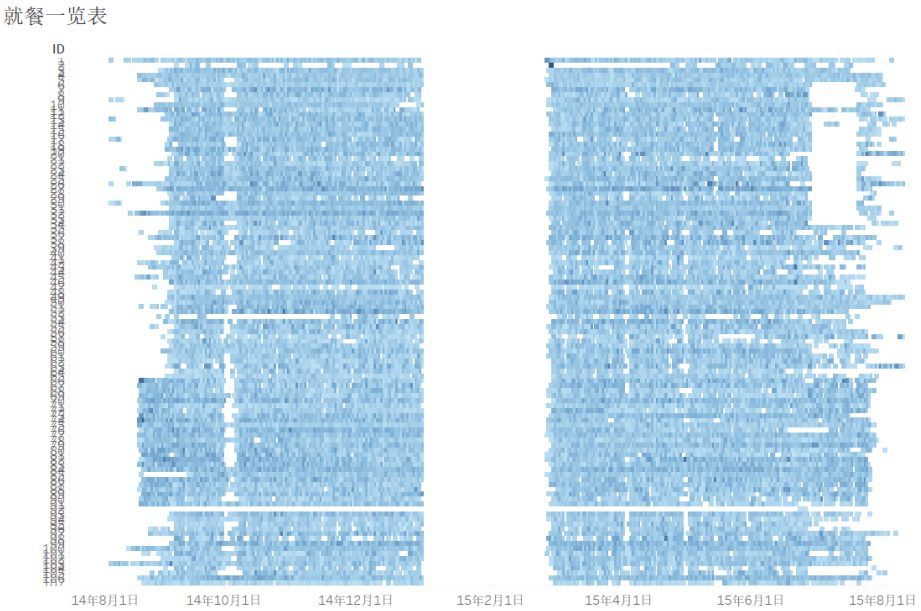


图 11：学生每日就餐统计

分析每个学生与 POS 机的关联，可以看出部分学生有固定的就餐偏好。如 99 号学员在 12 号 POS 机上刷卡 482 次，平均每天 2 次以上。

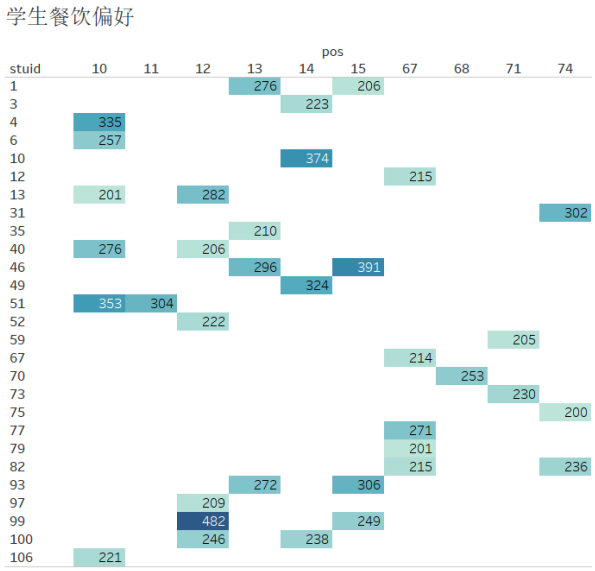
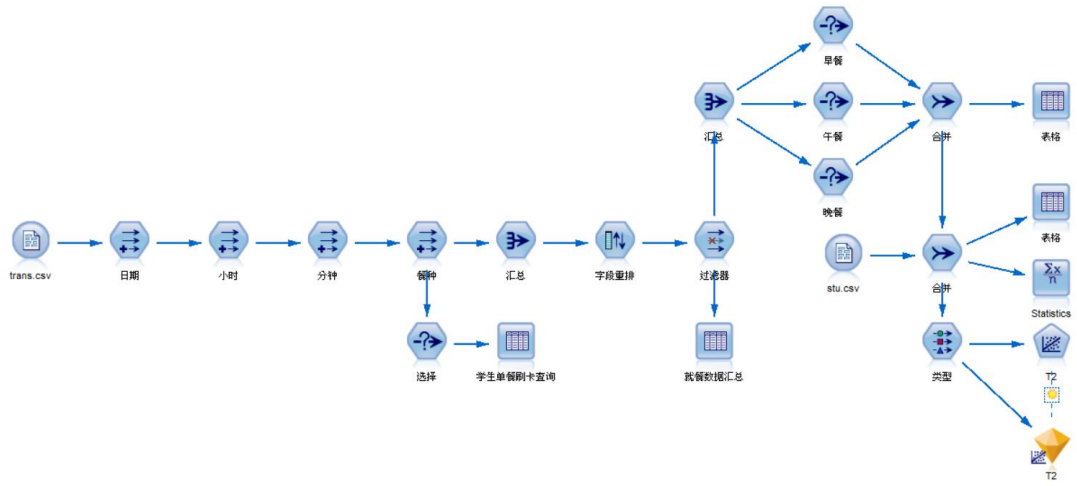


图 12：学生就餐偏好统计

四、建模分析

利用 modeler，对数据进行分析，提取食堂交易数据中的若干特征，形成一张二维表（如图 13 所示）。这些特征包括早餐金额、早餐时间、早餐次数、午餐金额、午餐时间、午餐次数等。



ID	早餐金额	早餐时间	早餐次数	午餐金额	午餐时间	午餐次数	晚餐金额	晚餐时间	晚餐次数	Gender	T1	T2	Class	Group	Prov	City	BirthDay
1	3.745	462.067	151	10.660	720.171	195	9.267	1069.800	162	1	88	83	1	1	福建	龙岩市	1995-02-01
2	3.400	435.750	2	13.756	726.723	43	10.175	1083.815	32	1	76	69	1	1	福建	龙岩市	1996-05-01
3	2.174	475.083	162	5.288	711.319	239	5.847	1078.600	231	1	75	74	1	1	福建	龙岩市	1996-10-01
4	2.618	470.613	238	6.569	722.496	273	6.265	1071.753	246	1	90	91	2	1	北京	朝阳区	1995-11-12
5	2.526	457.124	126	7.468	720.721	192	8.204	1077.223	197	1	86	86	2	1	江西	景德镇市	1996-01-13
6	2.335	479.360	138	6.247	716.202	225	5.834	1070.965	218	1	88	91	3	1	内...	呼和浩特市	1996-02-21
7	4.015	512.095	117	8.087	723.473	180	8.579	1071.355	178	1	82	75	3	1	广东	湛江市	1995-09-21
8	6.431	531.682	63	11.317	771.664	128	10.148	1094.857	97	1	76	83	3	1	辽宁	大连市	1997-10-21
9	2.668	494.025	114	7.442	717.031	142	7.289	1082.273	137	1	79	83	3	1	辽宁	大连市	1996-02-22
10	1.425	471.718	193	4.456	718.788	231	3.840	1076.345	232	1	74	90	3	1	浙江	杭州市	1995-09-22
11	4.869	466.103	77	11.774	743.235	174	10.768	1097.019	163	1	78	83	2	1	天津	河北区	1996-02-13
12	6.059	529.048	108	9.299	765.414	174	8.670	1076.434	127	1	74	74	3	2	湖南	常德市	1998-10-22
13	2.955	456.059	191	8.426	707.940	197	7.796	1084.848	159	1	83	82	3	2	河北	石家庄市	1996-02-23
14	3.838	506.906	131	9.754	727.109	225	8.979	1105.049	192	1	74	66	3	2	江苏	南通市	1995-09-23

图 13：数据处理流程与结果

分析成绩与各指标间的关联关系，发现各类就餐指标和成绩具有强相关关系。例如，早餐时间和成绩强负相关，早餐次数、午餐次数和成绩强正相关。

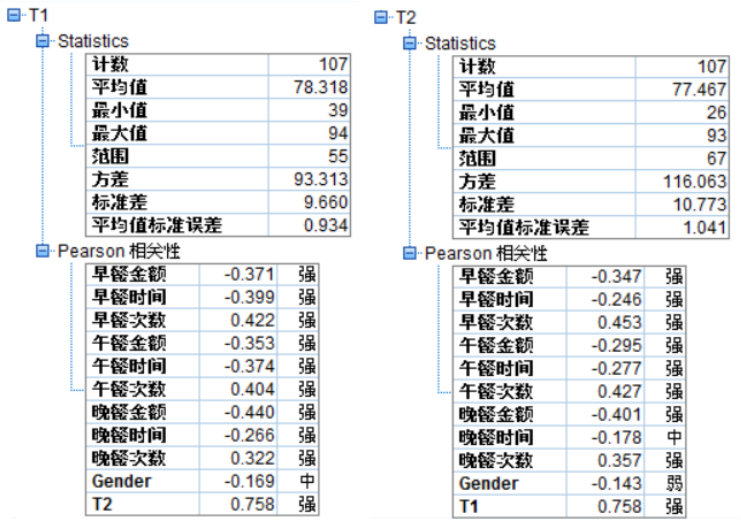


图 14：就餐指标和成绩的关联分析

构建回归模型，可以得到期末成绩的预测模型。期末成绩=72.85-0.012*早餐时间+0.33*早餐次数+0.46*午餐次数+0.19*晚餐次数-5.346*（性别男）。由于样本量较小，回归系数不显著，但是该模型仍具有一定的可解释性。

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	72.851	18.222		3.998	.000
	早餐时间	-.012	.034	-.036	-.358	.721
	早餐次数	.033	.027	.183	1.213	.228
	午餐次数	.046	.045	.212	1.020	.310
	晚餐次数	.019	.038	.098	.507	.613
	Gender	-5.436	3.174	-.177	-1.713	.090

图 15：回归模型的结果

六、总结和结论

为了帮助学员完成《商务大数据分析》的课程作业，撰写了本报告。**本报告中的主要数据为虚构数据。为了报告的趣味性，在虚构数据的过程中进行了一些设计。本数据分析的结果不具有现实意义。在具体撰写数据分析报告时，可以适当讨论数据分析报告的实践意义和价值。**