# 项目分析报告

问卷数据分析

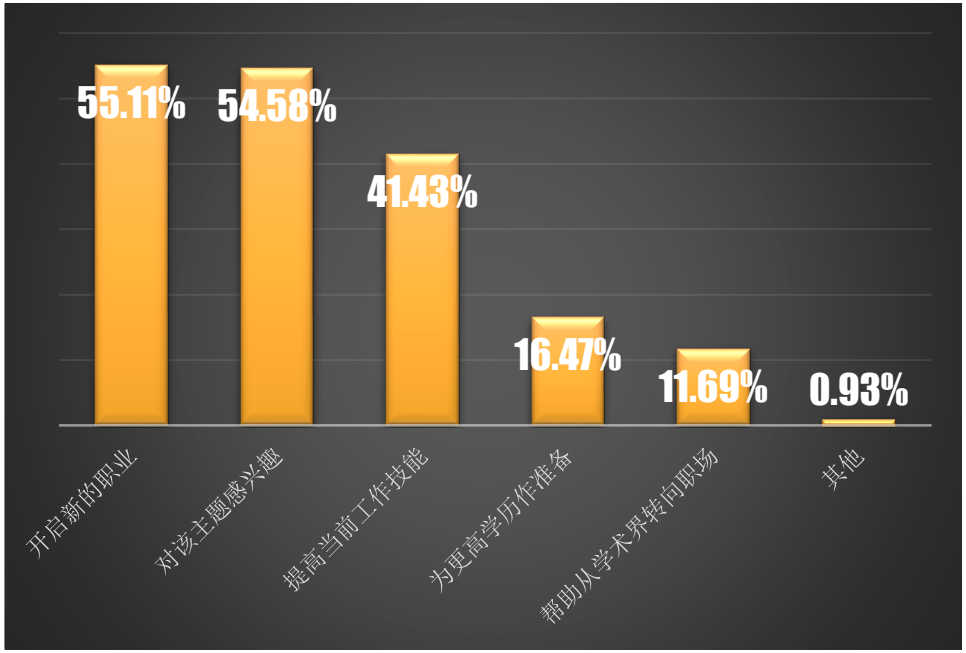
# 优达学城学员们的主要学习目的是?

主要探讨优达学城的学员们因何动机和目的而选择在优达学城进行课程学习和项目培训。

在Excel上的数据清理工作:

- 计算原始调查数据表中的总记录条数
- 计算每个选项占总记录条数的百分比，并汇总为一个数据表
- 用该数据表作出柱状图

# 优达学城学员们的主要学习目的是?




学员们最主要的学习目的是开启新的职业（有55.11%的受访者选了该选项）、对该课题项目感兴趣（54.58%）以及提升当前工作技能（41.43%），而为学历提升做准备和帮助从学术界适应职场则占比较少（16.47%和11.69%）。可见优达学城学员的主要学习目的是转行/转岗、兴趣以及提升技能。

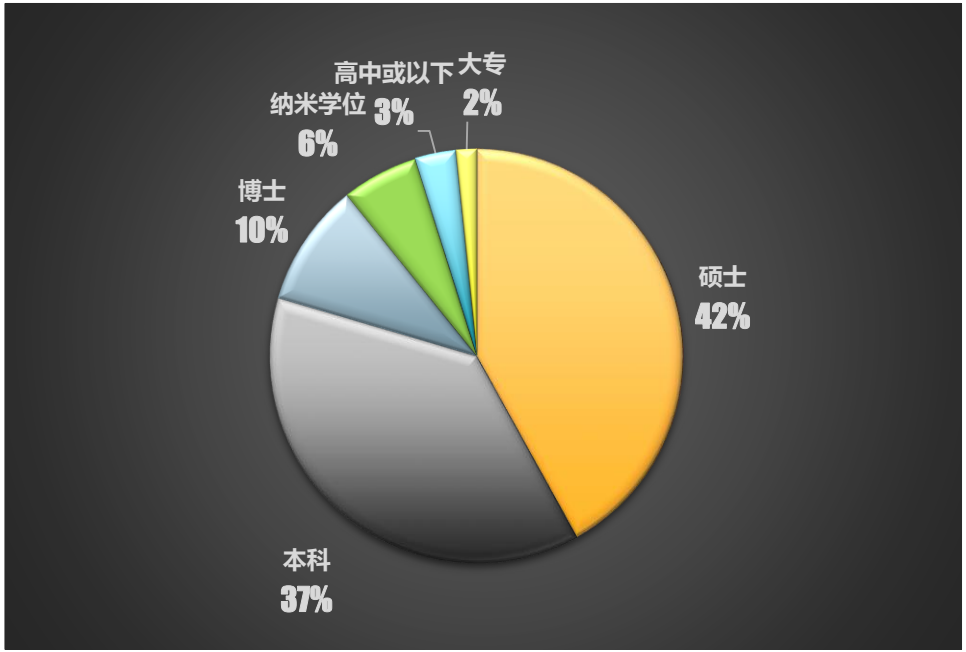
# 优达学城学员的最高学历是?

探讨优达学城学员的学历结构情况。

在Excel上的数据清理工作:

- 在原始数据表中选取最高学历那一列数据，检查没有异常和空缺
- 用数据透视表汇总
- 在数据透视表汇总的基础上作饼图

# 优达学城学员的最高学历是?




大部分优达学城学员都是硕士或本科学历，分别占到42%和35%，其次为博士学历，占到10%。

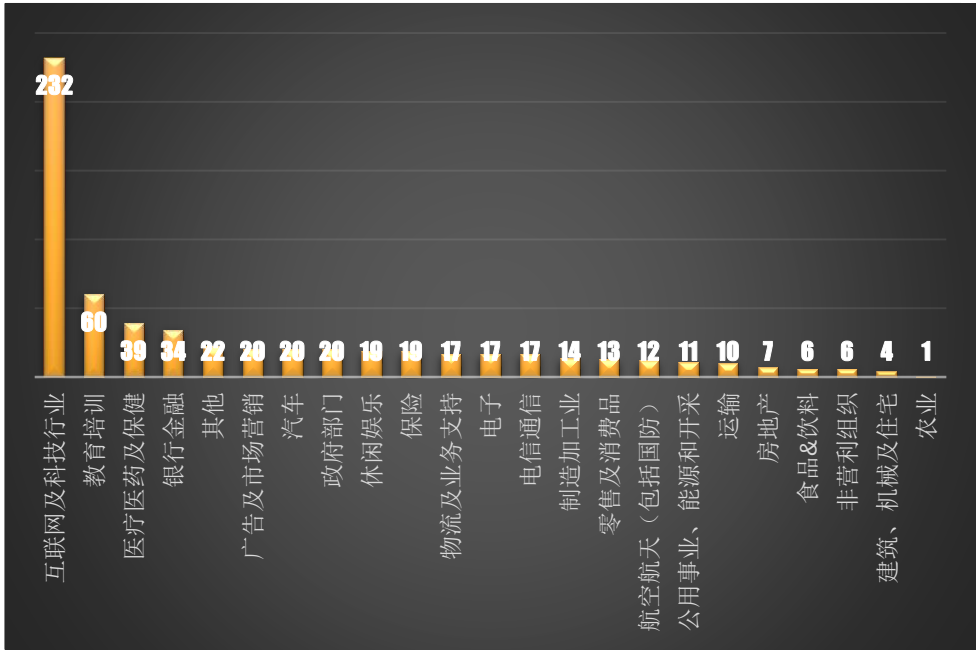
# 学员们主要来自哪些行业?

分析学员们主要来自哪些行业?

在Excel上的数据清理工作:

- 在原始数据表中提取学员行业所处信息那一列
- 有些学员选择了“其他”选项，并自己填写了内容。我根据“其他”中的内容，将其重新归类或与现有的分类合并，其余实在难以归类的仍然放入其他选项。
- 利用数据透视表汇总排列，并生成柱状图。

# 学员们主要来自哪些行业?




来自互联网及科技行业的人数远远多于其他行业的人数（232人），其次是教育行业/医药医疗保健、银行金融等。互联网及科技行业学员人数多，应当与优达学城所提供的课程内容有很大关联。

# 学员的年龄分布

探查目前优达学城学员的年龄分布，查看主要是哪一年龄阶段的学员较多。

在Excel上的数据清理工作：

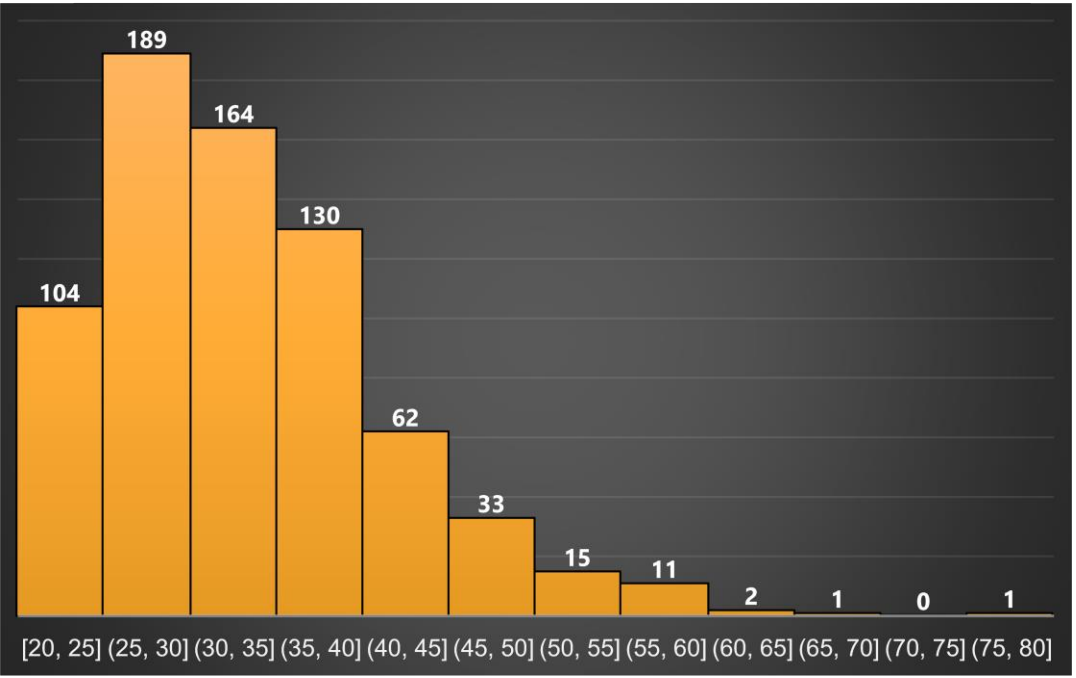
- 提取原始数据表中的出生日期那一列，利用Excel的DATEDIF（）函数计算年龄
- 将一些异常值剔除，例如只有1、2岁的那些记录，同时剔除空缺的记录。
- 利用直方图展示，同时使用表格展示一些描述性统计的指标。

# 学员的年龄分布

| | 学员年龄 |
|---|---|
| 平均值 | 34 |
| 中位数 | 32 |
| 众数 | 28 |
| 最大值 | 79 |
| 最小值 | 20 |
| 值域 | 59 |
| 标准偏差 | 8.34 |

关于学员年龄的一些描述性统计指标，优达学城学员的平均年龄大约是34岁，中位数是32岁，而28岁的学员人数最多。由此反映出学员的年龄应该多数集中在30岁左右这个区间里。

# 学员的年龄分布




以5年为组距绘制的直方图，可以看出25-30岁的学员人数最多，其次30-35岁，35-40岁、20-25岁。可见优达学城的学员主要还是中青年人群占多数。

# 学员们主要来自哪些国家地区？

查看学员主要来自哪些国家和地区？

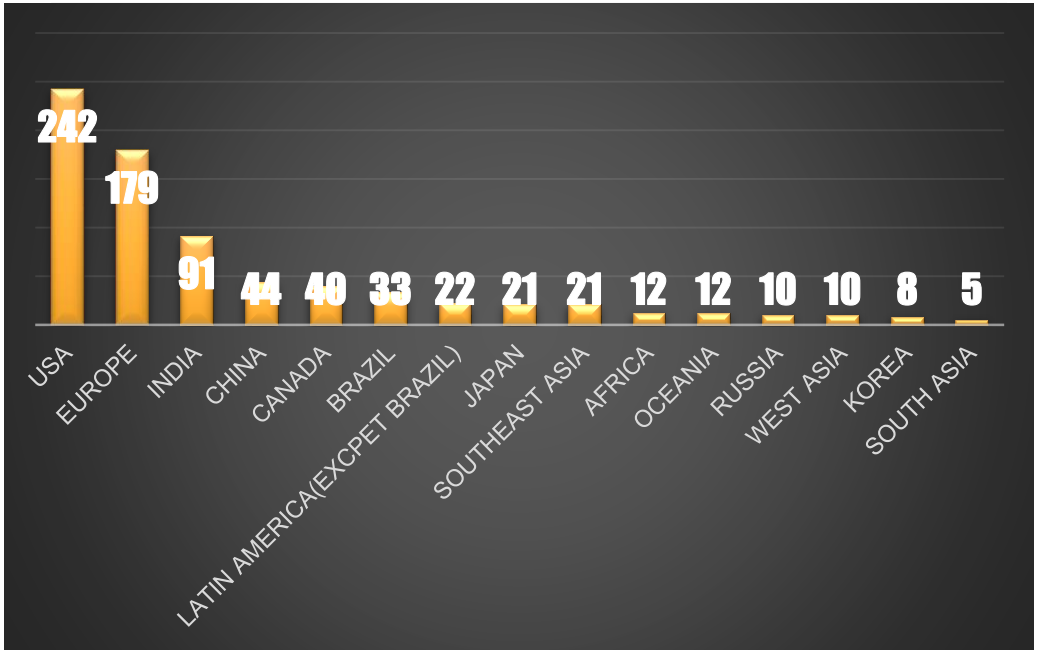
在Excel上的数据清理工作：

- 对原始数据表中的“What city and state / province / country do you live in?”一列进行重新整理，先使用excel的分列功能对其进行拆分，提取当中的国家名。存在相当多记录未写国家名的情况，对其筛选出来后运用数据透视表汇总，再人工判断属于哪个国家，制成一张数据匹配表，然后再用VLOOKUP函数对原数据进行匹配，从而使得每条记录都有国家名。

- 由于国家繁多不便于作图，因此将一些国家进一步汇总成洲/地区，比如将欧洲国家统一汇总为EUROPE

- 选择柱形图来展示数据

# 学员们主要来自哪些国家地区？




由左图可以看出优达的学员主要来自美国和欧洲，其次印度学员也较多，其后是中国、加拿大、巴西等国。

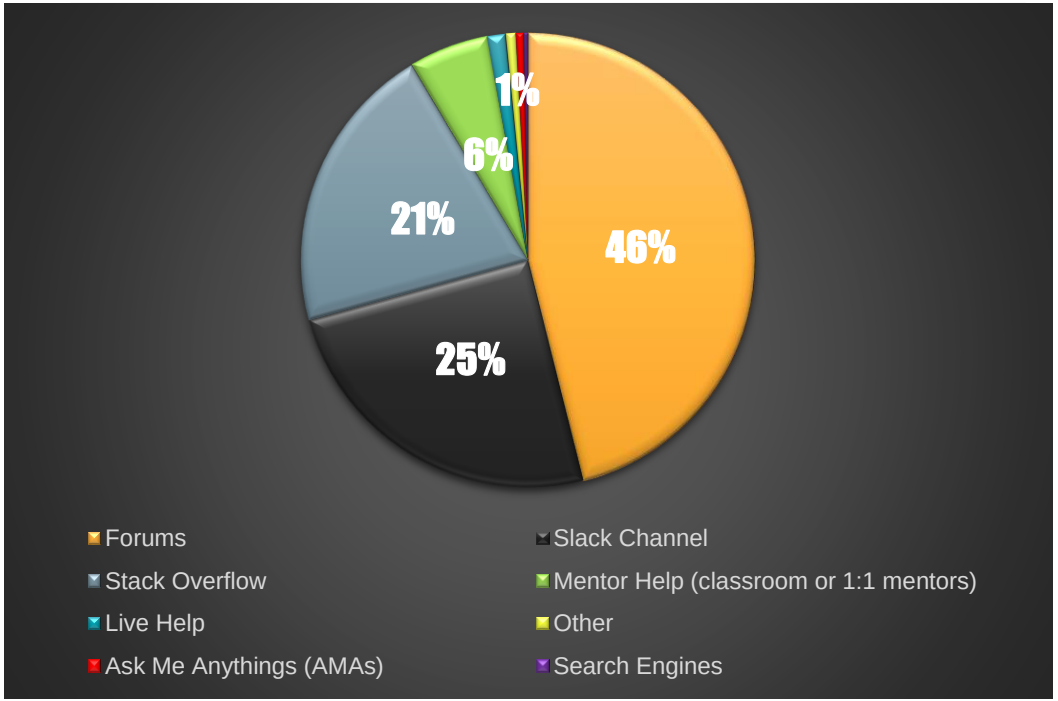
# 当学员们遇到不懂之处时，最常求助于?

当学员们遇到不懂之处时，最常求助于?

在Excel上的数据清理工作：

- 提取问卷调查数据表中关于当学员遇到不懂或疑惑时，会向什么渠道求助的那一列数据，对Other的内容进行判断重新归类和汇总
- 利用饼图展示

# 当学员们遇到不懂之处时，最常求助于?




当遇到不懂之处时，学员们大多还是会求助于论坛（46%）。除此之外，Slack Channel和Stack Overflows等平台工具也是学员们常用的学习求助渠道。相比之下，求助于教师或在教师处得到帮助的占比要小得多（6%）。

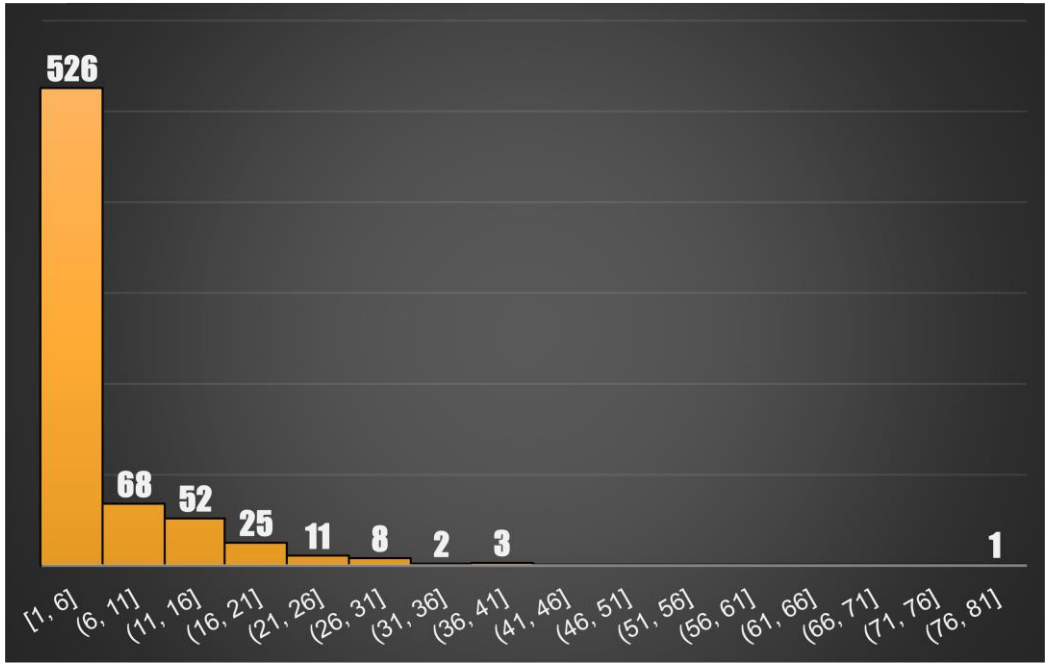
# 学员通常每周花多少时间学习和掌握相关知识?

参看大多数学员每周花多少时间去学习和消化所学知识?

在Excel上的数据清理工作:

- 在数据表上找到相关列
- 对一些非数值型的数据进行修正，比如对于2-4小时，取中值3作为其数值。对少数不确定的文字性描述进行剔除。
- 将所有记录转化为合理的数值型数据后，查看是否有不合理的数字（比如小时数超出一周）
- 生成统计图表

# 学员通常每周花多少时间学习和掌握相关知识?




左图是一个明显的右偏分布。绝大部分学员每周在学习和消化相关知识上要花费1-6小时，其次是6-11小时，每周需要花费较长时间学习的人数较少。

# 在职与否对每周学习和掌握知识的时长有无影响？

工作需要占用相当长时间，通常会对学习时间产生影响。此处查看在职和不在职的学员之间每周花费在学习上的时间是否有明显差别？

在Excel上的数据清理工作：

- 找到相关列进行前述操作
- 将“0”替换为unemployed，“1”替换为employed
- 利用箱线图对比展示

# 在职与否对每周学习和掌握知识的时长有无影响？

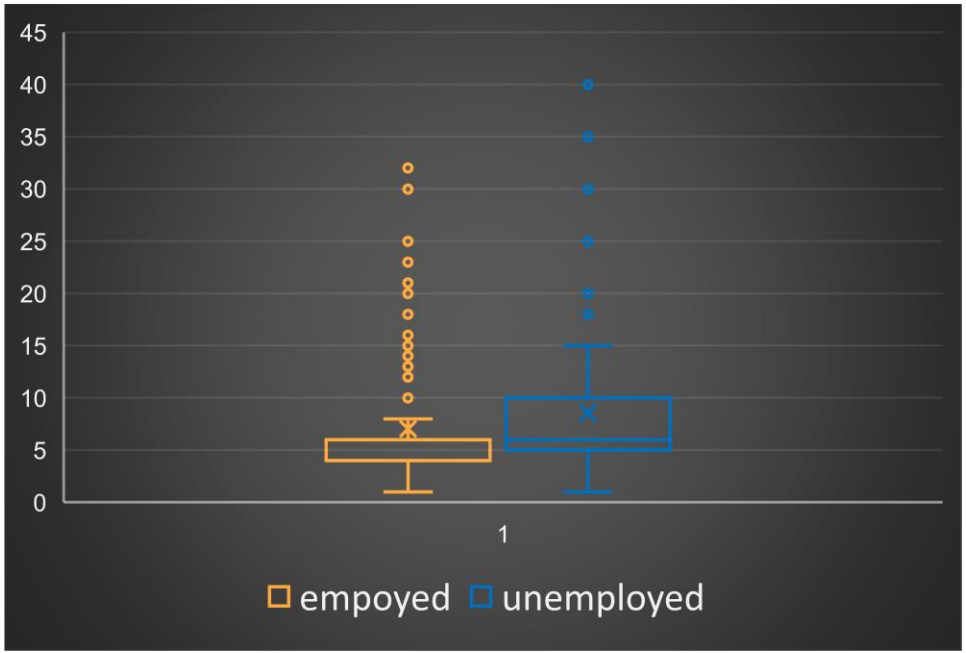
| | employed | unemployed |
|---|---|---|
| 平均值 | 7.0 | 8.6 |
| 中位数 | 6 | 6 |
| 众数 | 6 | 6 |
| 最大值 | 32 | 40 |
| 最小值 | 1 | 1 |
| 值域 | 31 | 39 |
| 标准偏差 | 5.20 | 7.68 |
| 离散系数 | 74.10% | 89.66% |

从平均值来看，不在职的学员平均每周花费8.6小时在学习上，而在职学员花费的时间只有7小时。而无论是在职还是不在职，都是每周花费6小时左右学习的人数最多（两者的中位数和众数都是6）。

相比于在职学员，不在职学员每周学习时间的内部差异要更大一些。比如从值域来看，不在职学员的值域有39，而在职学员只有31。标准偏差也同样反映出这个趋势。

由于双方均值不同，此处引入了离散系数进一步描述双方离散程度的差异。不在职的学员的离散系数89.66%同样大于在职学员（74.10%），反映出不在职学员每周学习时长彼此差异程度更大。

# 在职与否对每周学习和掌握知识的时长有无影响？




从左边的箱线图可以看出，不在职状态（unemployed）的箱体明显要高于在职状态，这显示出不在职状态的学员们确实普遍比在职状态学员有更多的时间可以花在学习上。

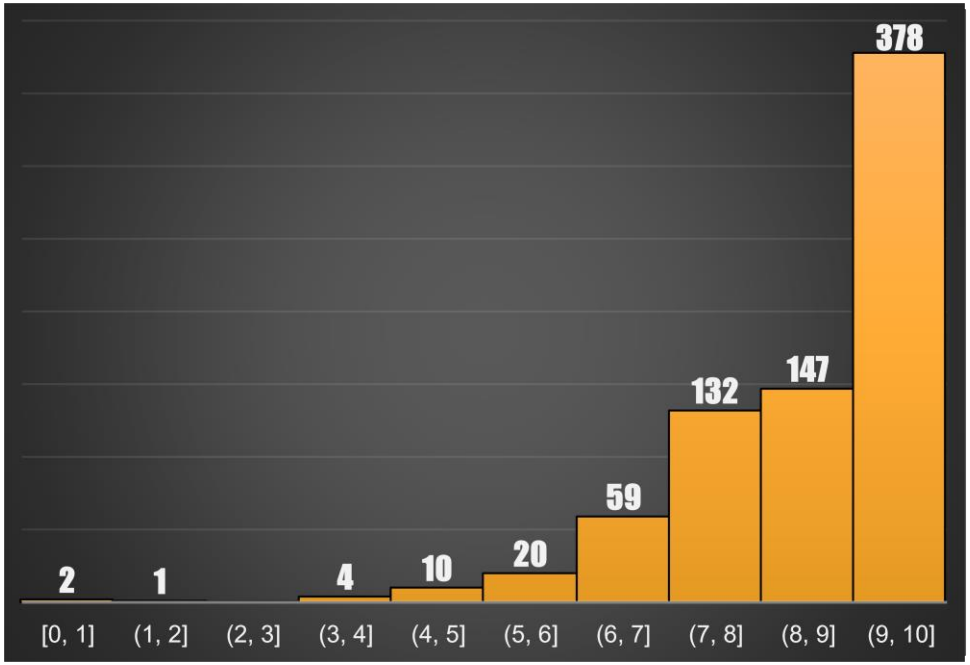
# 学员们有多愿意向同事及朋友推荐优达学城?

通过人们向同事朋友推荐优达学城的意愿，来大致查看人们对于优达学城的看法。

在Excel上的数据清理工作：

- 在问卷数据表中选取“How likely is it that you would recommend Udacity to a friend or colleague?”一列
- 绘制直方图

# 学员们有多愿意向同事及朋友推荐优达学城?




看来大多数人对于优达学城还是比较满意的，满分仅有10份，而推荐程度在9-10之间的人数远多于其他分数区间的人数，整个分布也是呈现出左偏形态的，显示出愿意推荐程度越高，所对应的学员越多。

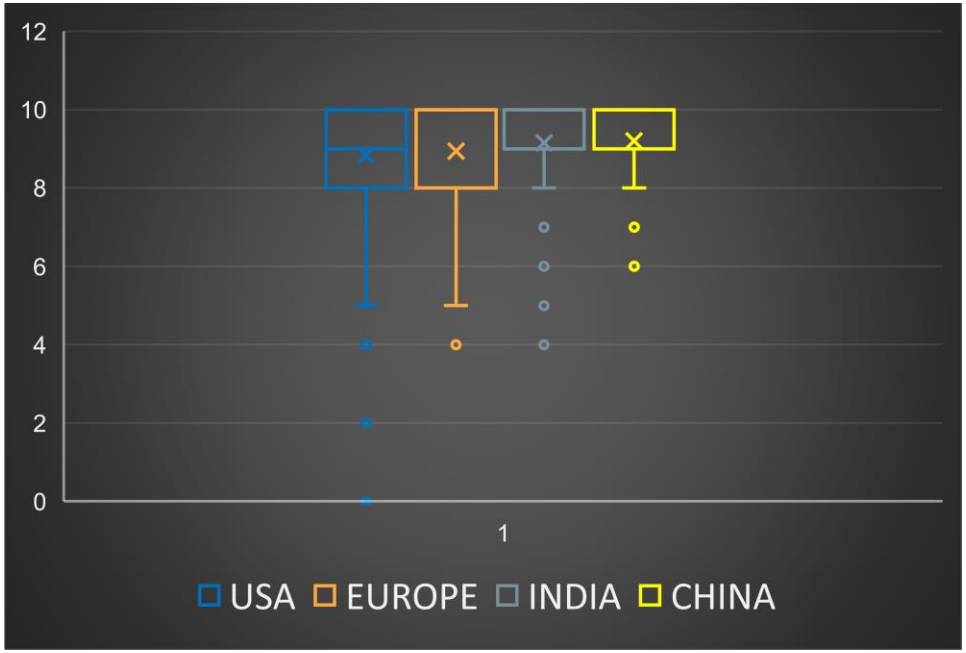
# 不同地区学员对于优达学城的推荐度有何差异?

由于涉及到不同的文化和语言背景，尝试查看不同国家地区的人们对于优达学城的满意度和推荐程度是否会有所不同。

在Excel上的数据清理工作:

- 按照前述所说，对国家和地区的数据进行清理汇总
- 利用箱线图对比展示

# 不同地区学员对于优达学城的推荐度有何差异?




主要选取了学员人数最多的四个国家/地区进行对比，可见四个国家/地区总体上都对优达学城有较高评价（大部分都在8-10之间）。美国学员对优达学城的看法差异相对较大一些，而相比于美国和欧洲，印度和中国学员对优达学城的评价差异度更小。

# 参考资料

[1] 人大经济论坛. 从零进阶！数据分析的统计基础[M]. 北京： 电子工业出版社，2015.2