**数据可视化项目报告**

1. 目标与业务目标

你的团队刚刚获得一个新的项目，要与当地希望扩招学员的 UVW 大学合作。UVW 已经选择将薪资作为关键统计数据，以确定其学位课程的营销标准。你必须利用美国人口调查局提供的数据开发营销档案，并将 50,000 美元作为薪资的关键数据。对于薪资低于和超过 50,000 美元的人员，要评估许多关键变量，包括年龄、性别、教育背景、婚姻状态、职业等等。

项目要求：使用不同的数据可视化来发掘可能与收入相关变量，以回答客户的问题。

2、项目介绍

在本项目中，我们的目标是通过各种数据可视化技术，来发掘与人员收入相关的各项变量，整个项目流程包含如下几个步骤：

1. 数据准备以及加载
2. 数据清洗
3. 数据可视化

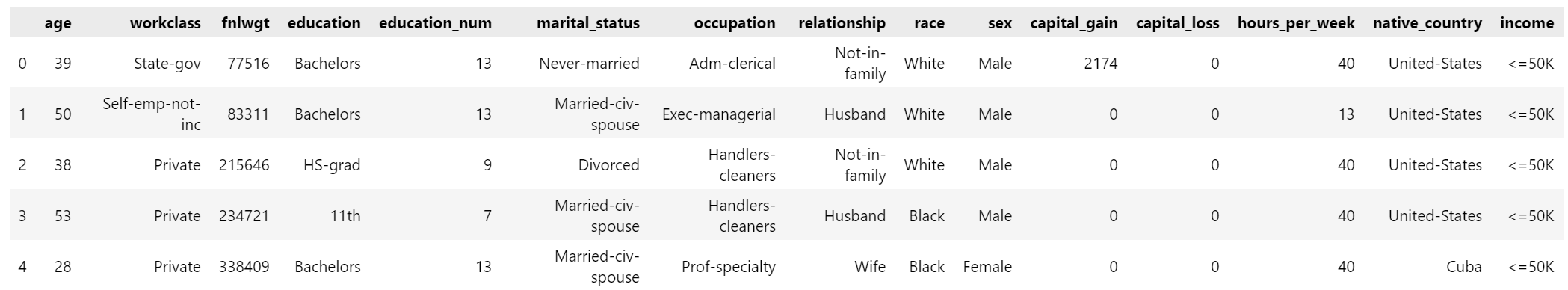
下面我将分别介绍每个步骤的具体做法：

1. **数据准备以及加载**

在本项目中，我们使用了[Adult - UCI Machine Learning RepositoryLinks to an external site.](https://archive.ics.uci.edu/dataset/2/adult" \t "_blank) 数据集，通过对数据文件和字段说明文件的分析，实现了数据的加载以及数据字段的设置，数据加载的示例代码如下：

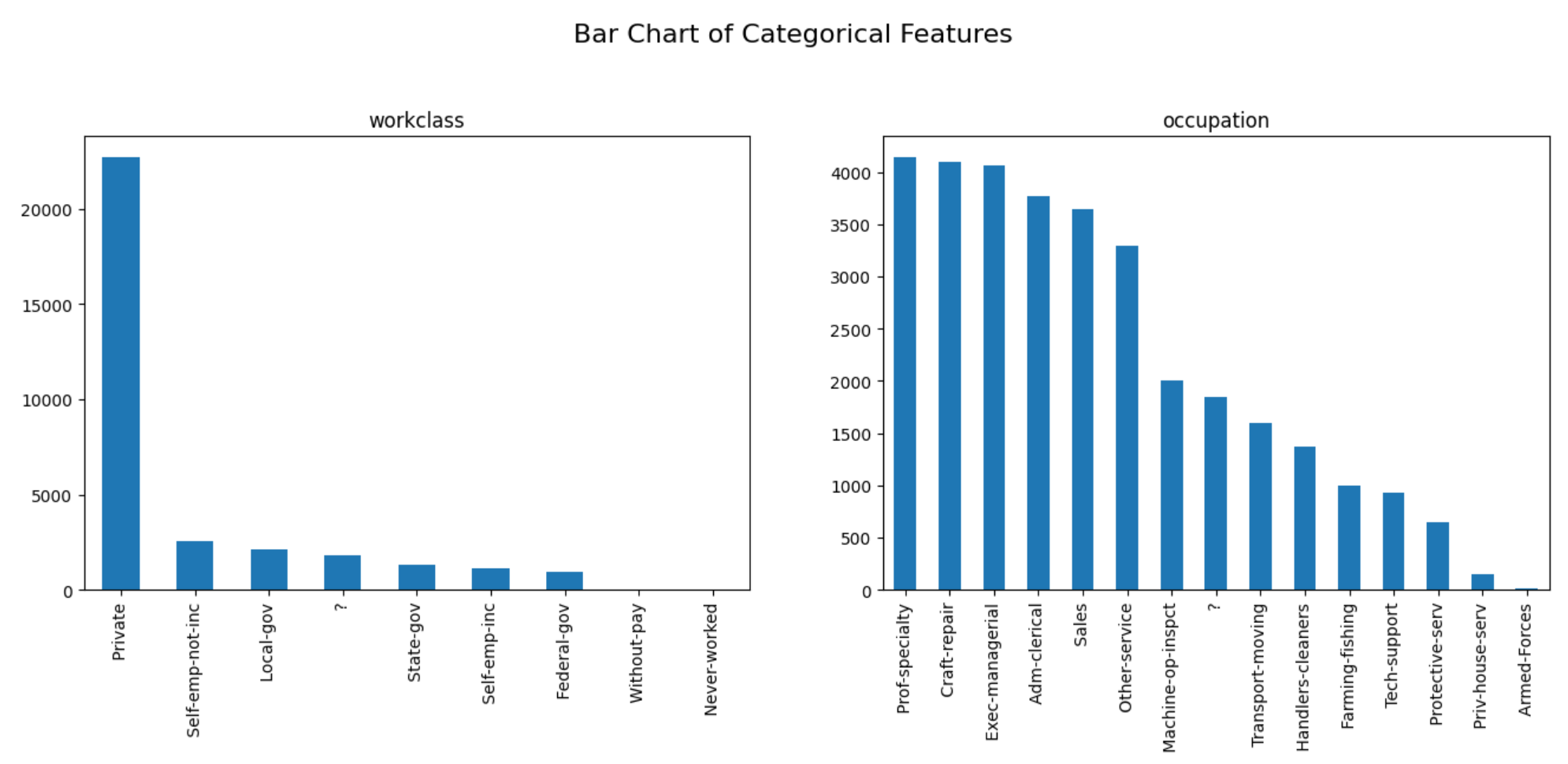
|  |
| --- |
| adult\_data\_path = "./data/adult.data"  adult\_test\_path = "./data/adult.test"  columns = ["age", "workclass", "fnlwgt", "education", "education\_num", "marital\_status",             "occupation", "relationship", "race", "sex", "capital\_gain", "capital\_loss",             "hours\_per\_week", "native\_country", "income"]  adult\_df = pd.read\_csv(adult\_data\_path, names=columns)  adult\_test\_df = pd.read\_csv(adult\_test\_path, names=columns)  adult\_df.head() |

加载完成后的示例数据如下：



1. **数据清洗以及数据处理**

数据清洗步骤主要包括对异常值的过滤以及对特定字段的数据处理，通过分析，我们发现workclass 以及 occupation 两个字段分别存在 5.6% 和 5.7% 的数据缺失（对应柱状图中的?字段）。



过滤掉异常值后，数据量从 32561 减少至 30162，异常数据占比为 7.36%。

完成异常值过滤后，我们对三个特殊字段的数值进行了分桶处理，以方便后续计算与收入的相关性，数据分桶是比较有效的数据转换方法，可以更好的挖掘数值数据的特性，比如针对年龄数据，针对具体的年龄数值很难挖掘出比较明显的特性，但不同年龄段的人群却比较容易发掘出关联，本项目中，我们将年龄字段安装不同的年龄段进行了分桶处理，0-18岁为少年，18-30岁为青年，30-50岁为中年，50-70岁为老年，70岁以上为老老年。受教育年限以及每周工作时长也按照不同的区间划分为了高、中、低三段。

1. **数据可视化**

在完成了数据加载以及清洗操作之后，我们尝试了一系列的数据可视化策略，来挖掘不同变量与收入之间的关系，下面我们将分别介绍每一种可视化策略的实现以及对应的分析结论。

1. **可视化策略1**

|  |
| --- |
| 该策略基于 age, hours\_per\_week, education\_num 三个字段, 创建一个3D散点图，其中 x 轴为 age, y 轴为 hours\_per\_week, z 轴为 education\_num，income 的两种取值用不同颜色表示，通过散点图可以直观的观察高收入和低收入人群在不同象限上的分布情况。  3d_scatter  通过分析，该可视化策略的结果符合预期，低收入的人群主要集中在年龄较小或较大，工作时长短、且受教育年限短的区间；高收入的人群则主要集中在青壮年、每周工作时长较长且受教育年限较多的区间。 |

1. **可视化策略2**

|  |
| --- |
| 该策略基于离散类型的字段 "relationship", "race", 采用热力图来进行可视化展示，在不同的变量取值组合下，高收入人群的比例越高，图像的颜色越深，反之则越浅。通过观察图像的深度变化，可以分别发掘出与高收入和低收入人群最相关的变量组合。  heatmap  通过分析，从该热力图中可以看出，White+Wife 以及 Asian-Pac-Islander+Husband 的变量组合高收入占比最高，black+Own-child 以及 black+Other-relative 的变量组合低收入占比最高。 |

1. **可视化策略3**

|  |
| --- |
| 该可视化策略基于使用（relationship, age\_catrgory, education, race, occupation）等相对区分度较高的变量，使用了平行坐标图可视化发掘各变量之间的关系  parallel_plot  通过平行坐标图，可以看出，高收入人群相关性更强的变量组合包括：relationship-husband, relationship-wife, age-middle, age-old 以及 race-white。  低收入人群相关性最强的变量组合包括： relationship-ownchild, age-veryold, age-child, education-1st-4th, 以及 race-other. |

1. **可视化总结**

通过以上的一系列数据可视化策略的构建，我们挖掘出了以下与收入密切相关的变量以及变量组合：

|  |  |
| --- | --- |
| **变量或变量组合** | **预估收入** |
| **age(<20 or >60) + education\_num(<6) + hours\_per\_week(<40)** | **<50K** |
| **age(30 ~ 50) + education\_num(>8) + hours\_per\_week（>60）** | **>50K** |
| **Race（White）+ Relationship(Wife)** | **>50K** |
| **Race(Asian-Pac-Islander) + Relationship(Husband)** | **>50K** |
| **Race（Black）+ Relationship(Own-child, Other-relative)** | **<50K** |
| **Relationship(husband,wife) + age(middle,old) + race(white)** | **>50K** |
| **Relationship(own-child) + age(child, very-old) + education(1st-4th) + race(other)** | **<50K** |

1. **问题与挑战**

在本项目中，利用数据可视化来进行重要变量的挖掘以及尝试对收入情况进行预估时,我们发现了如下几个问题与挑战：

1. 数据质量的判定：目前的数据清洗技术只能过滤掉存在明确异常值的数据，但对于数值正常但可能存在信息错误的数据没有特别明确的判定手段，比如在本项目的数据集中，capital\_gain 和 capital\_loss 两个字段的数据过于稀疏，99%以上数据为0，但我们也没有确定的手段来判定这两个字段的数据是否完全不可用。
2. 变量挖掘的覆盖面问题：目前基于可视化分析得到的关键变量以及变量组合都是基于分析人员对于可视化结果中存在的特别明显的分布差异总结得来的，但在真实的收入预估场景下，能够满足这些特定条件的用户有限，只能覆盖很小一部分用户。
3. 预估结果的准确性提升问题：目前基于可视化方法总结得到的预估策略，在预估的准确性上存在一定瓶颈，很难再进一步提升。
4. **后续工作计划**

针对用户收入预估的任务，在真实的应用场景下一般会基于前期数据分析以及可视化的基础，进行特征的评估和筛选，然后利用机器学习模型来进行建模。在本次任务场景下，后续我们可以基于各个特征变量，利用机器学习模型来构建一个二分类器，用于预估用户的收入是 “>50k” 还是 “<50K”。

相比于基于可视化分析总结得到的关键特征变量，采用机器学习模型的方式有如下优点：

1. **自动化特征提取：**机器学习算法能够自动从原始数据中提取关键特征，无需人工干预。相比较而言，基于数据可视化的方法需要人工进行特征的选择和提取。
2. **能够发现非显性特征：**机器学习算法可以通过学习数据中的模式和关联性，发现非显性的特征。这些特征可能对于分类任务非常重要，而且很难通过简单的可视化方法发现。
3. **能够处理大规模数据集：**机器学习方法可以有效地处理大规模数据集，即使是具有高维度的数据。相比之下，基于数据可视化的方法有时候可能会受到数据规模的限制。
4. **可以适应复杂模式和非线性关系：**机器学习算法可以建模和捕捉复杂的模式和非线性关系，从而更好地进行分类。而基于数据可视化的方法往往只能处理简单的模式或线性关系。
5. **鲁棒性和泛化能力：**通过机器学习训练的分类器通常具有较好的泛化能力，即可以处理未曾见过的数据样本并做出准确的分类。这使得分类器具有一定的鲁棒性，相对不容易受到异常数据的干扰。