CSE 578：Data Visualizationn

Project Repot

Zhenyu Wang，Yuntao Tan，Yulin He，Shengjian Wu

# Roles and Responsibilities:

Team Members: Yuntao Tan, Shengjian Wu, Yulin He, Shengjian Wu

Stakeholders: UVM College

Product owners: XYZ Corporation

作为XYZ公司的数据分析“X Squad”团队，我们将为UVW大学开发一个用于招生的应用程序，以找到决定个人收入的因素。目前，UVW大学决定以薪资作为关键统计数据，来确定其学位课程的营销标准。我们“X Squad”团队将使用美国人口调查局提供的数据作为依据来开发应用，且将把50,000美元年收入作为薪资的关键数据。

项目主要任务包裹：数据库的理解、分析，探索决定个人收入的关键变量，建立机器学习模型并使用关键变量来预测个人收入。

项目的主要分工如下：

* 每个团队成员需理解数据库。
* 将探讨出来关于数据库变量的问题合理分配至团队成员来分析并得出结果。
  + Shengjian Wu，预测模型的建立和探讨套数据；
  + Yuntao Tan，探索问题： x- x，报告框架起草、前部分；
  + Yulin He, 探索问题：X - X，报告后部分；
  + Zhenyu Wang：探索问题 X - x，建立网站
* 全员参与可视化探索分析。
* 全员参与报告和展示制作。

# Team Goals and Business Objective:

本次项目目标是通过数据可视化分析提供的数据库找到确定个人收入的变量并呈现给UVW大学；其次，我们将基于初始分析，来开发机器学习模型应用，为UVW大学的营销团队提供精准的个人收入预测，以确定其学位课程的营销标准。

# Assumptions:

* **数据准确性和无误性**：我们假定被给到用软件开发的数据库是准确、无误的。数据库中不含有误导性表达及不含有错误的元素。如果不了解数据将如被使用，则确保准确性和精度可能偏离目标或比必要的成本更高。
* **数据合法性和准许性**：我们假定数据库是合法、准许的。例如，在给定的数据集中，性别等项目通常仅限于一组选项，并且不允许提供开放的答案。根据数据集的要求，这些之外的任何答案将不被视为有效或合法。
* **数据及时性和相关性：**我们假定数据库是及时、相关的。收集的数据过早或过晚，可能会歪曲情况，导致不准确的决策。例如，数据集中的工资如果来源于经济衰退期间，它就不能证明数据集的合理性，并且会导致不准确的结果。
* **数据完整性和全面性**：我们假设数据库是完整、全面的。不完整的数据和不准确的数据一样危险，不完整数据可能会导致要显示的整体图像的部分视图缺失；如果没有操作如何运行的完整说明，将会发生未知的操作。必须理解构成数据集的一整套要求，以确定是否满足这些要求。
* **可视化及分析的合理性**：我们假设使用的数据可视化及分析方式均为合理的。对类预测贡献最大的特征可以给我们更多可解释的模式，我们认为大多数的可视化和分析都是对选定的无偏特征进行的。

# User Stories:

Data cleaning and preprocessing:

Individual attribute analysis:

Visualizaiton:

Machine Leaarning:

Executive and systems Report:

# Visualizations:

最初，通过使用数据探索技术来识别模式，对每个特征进行了单独的分析。这是通过利用数据可视化工具，如饼状图、柱状图、直方图和方框图来实现的。此外，还使用其他的统计方法，如平均、中值和标准偏差来理解基本分布。

接下来，我们通过使用马赛克图、平行坐标图和散射图等工具的多元特征分析，加强了我们的初始分析。由此，我们确定了一组重要的特征。这构成了机器学习分析的基础，其中涉及到重要特征的特征工程。接下来，我们训练了一些机器学习模型来确定个人的收入。

由此，我们认识到特征的重要性，并观察到它们与我们对特征的初步分析是一致的。

对个别特征的初步分析：

1. Age：随着年龄的增长，工资会大幅度下降。30-55岁之间的人有更多的机会得到薪水超过5万。
2. Work class ：这不是一个值得考虑的好特性，因为数据在这两个类中都非常稳定。有趣的是，更多的联邦政府成员的收入超过5万。
3. Fnlwgt：此值大于45k的属性的工资为<=50k。
4. Education：从图表中可以明显看出，受过大学教育的人有更高的机会获得>5万。
5. education-num：它与“Education”属性相同。
6. Marital-status：有趣的是，85%的婚姻收入超过5万美元的人都是已婚的公民配偶。它还表明，从未结婚的人的收入更少。
7. Occupation：可以推断，行政管理和专业职业人员使近50%的人的收入超过5万。工艺维修职业的人员同样分布在这两个阶层上。
8. Relationnship：这个特殊的功能对薪水有更大的影响，因为你可以看到，属于关系“丈夫”的人占收入超过5万的人的75%。正如“Marital-status”特征可知，即使这一特征也推断未婚或未结婚人士的收入往往低于5万。
9. Race：“Race”属性并不真正影响人们的工资范围。在这两个阶级中，大多数人都属于“White”种族。
10. Sex：男性统治了这两个阶级，但在工资范围超过5万人之间，男性的统治地位要强得多。
11. Capital-gain：Capital-gian越高，此人的收入越容易超过5万美金。
12. Capital-loss：这两个阶层的资本损失或多或少都是相同的。所以这可能不是一个可以用来对数据进行分类的好因素。
13. Hours-Per-Week：大多数每周工作超过40小时的人属于收入超过5万美金类别。
14. Native-Country：下列国籍的人收入低于5万：

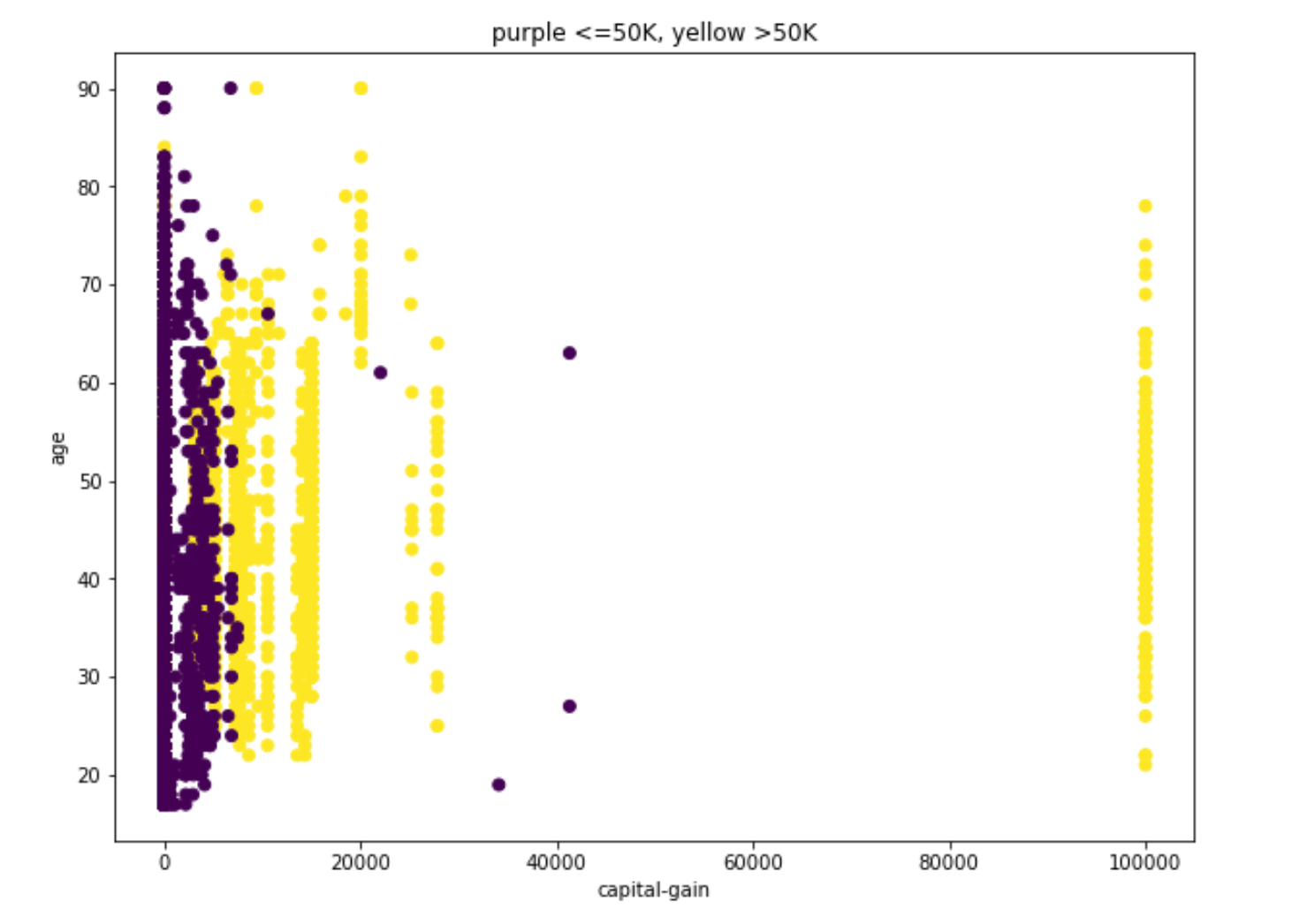
“Jamaica, Mexico, Puerto-Rico, Honduras, Colombia, Ecuador, Laos, Haiti, Portugal, Dominican-Republic, El-Salvador, Guatemala, Peru, Outlying-US, Trinidad & Tobago, Nicaragua, Vietnam.”

1. Income：数据集中超过5万美金的数据为7508条，不超过5万美金的数据为：22654条，因此数据集给出的大多数人都没达到5万美金的年收入。

以下是帮助我们识别数据集中的重要特征的重要图：

**主要分析方法：**

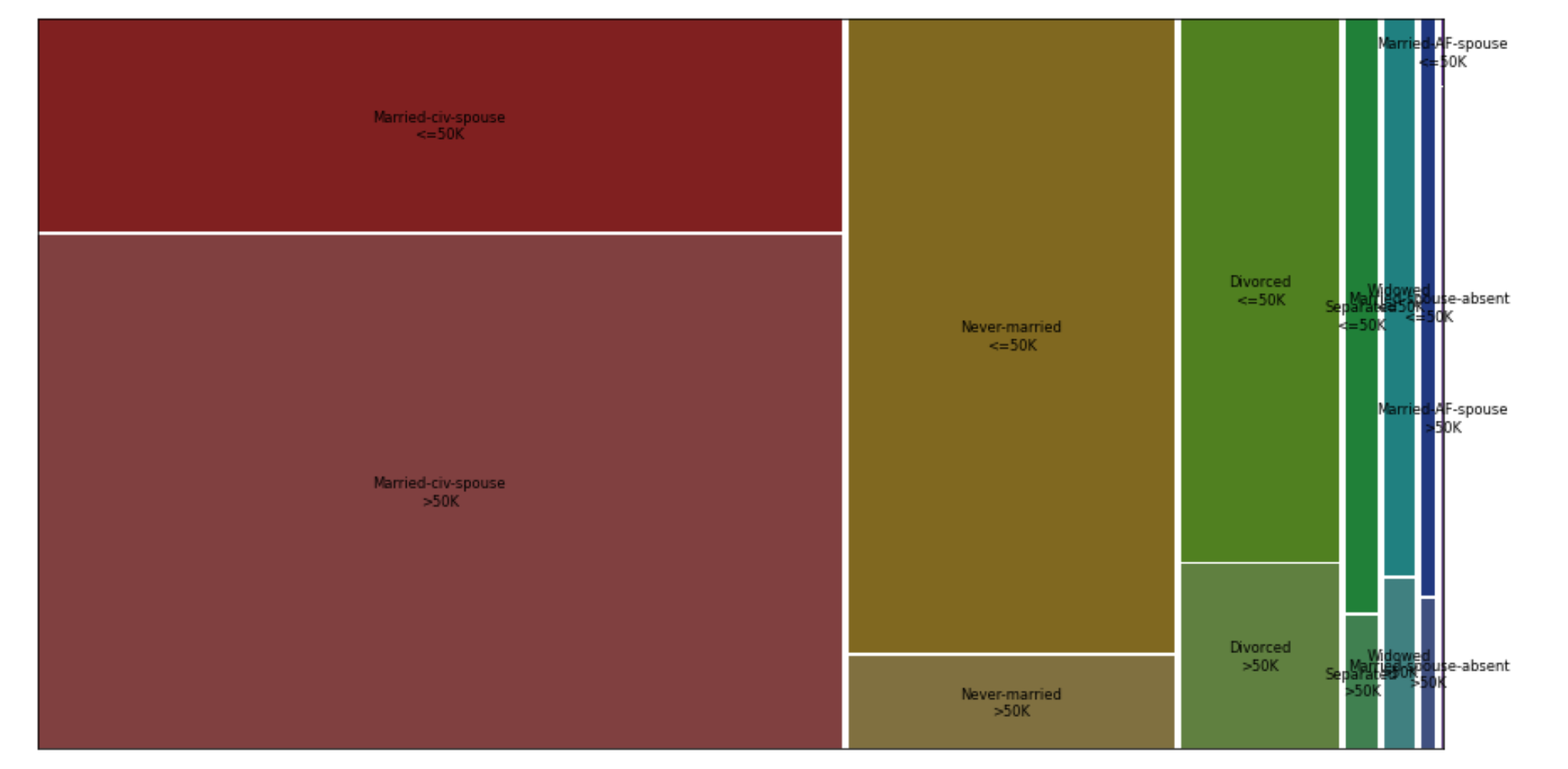
**Scatter plot: Capital-gain and Age**



推论：

1. 除了两个异常值外，这两类数据之间似乎存在着分离性。
2. 资本收益较高的个人更有可能获得的收入超过5万美元。

**Mosaic Plot： Occupation and salary**

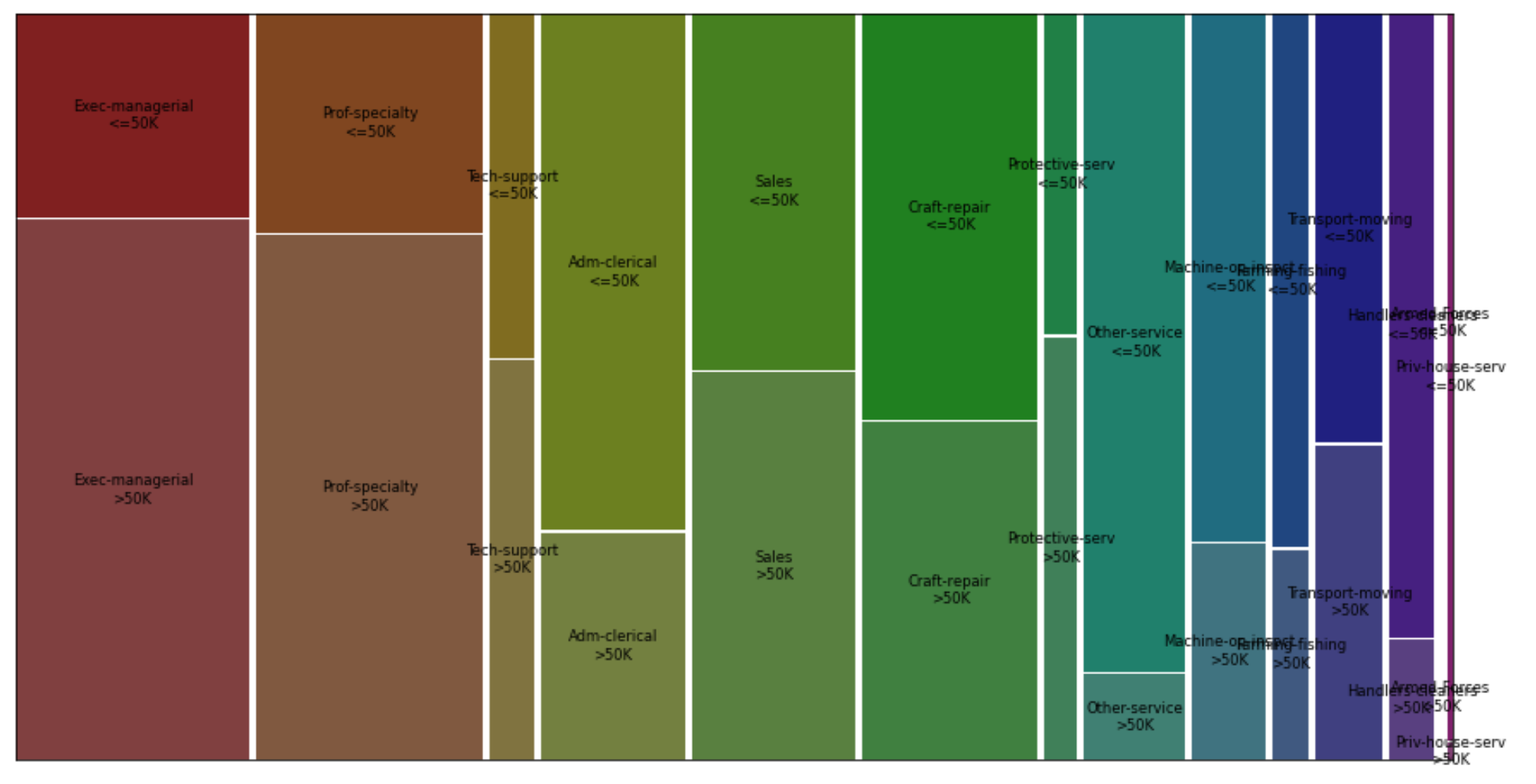


按顺序分类：Never-married, Married-civ-spouse, Divorced, Married-spouse-absent, Separated, Married-AF-spouse, Widowed

推论：

1. 对于大多数分类数据，这两个类的分布是高度倾斜的，这表明这个特性可以用来区分这两个类。
2. 婚姻状况为“Married-civ-spouse”的个人更有可能收入超过5万。
3. 婚姻状况“Never-married”的个人收入低于5万的几率比较大。

**Mosaic Plot：Occupation and Salary**

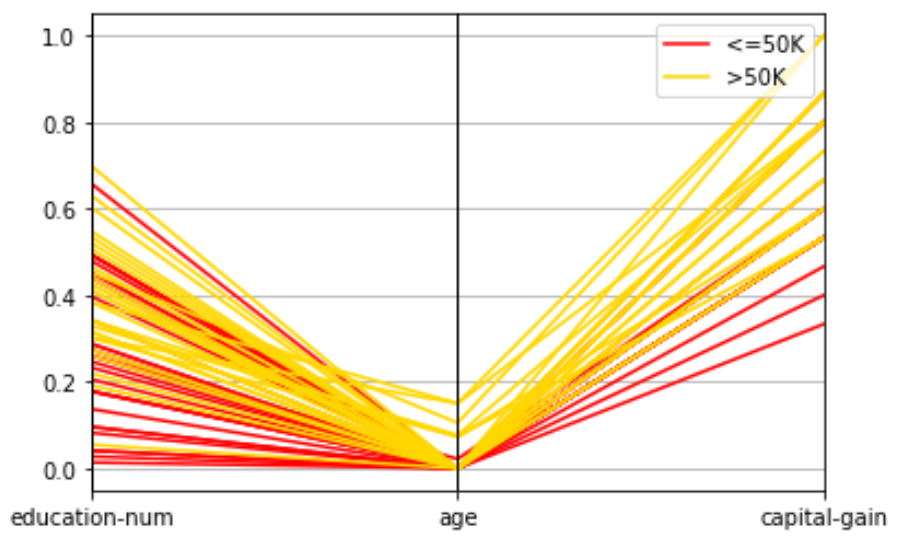


**按顺序分类: Adm-clerical, Exec-managerial, Handlers-cleaners, Prof-specialty, Other-service, Sales, Transport-moving, Farming-fishing, Machine-op-inspct, Tech-support, Craft-repair, Protective-serv, Armed-Forces, Priv-house-serv**

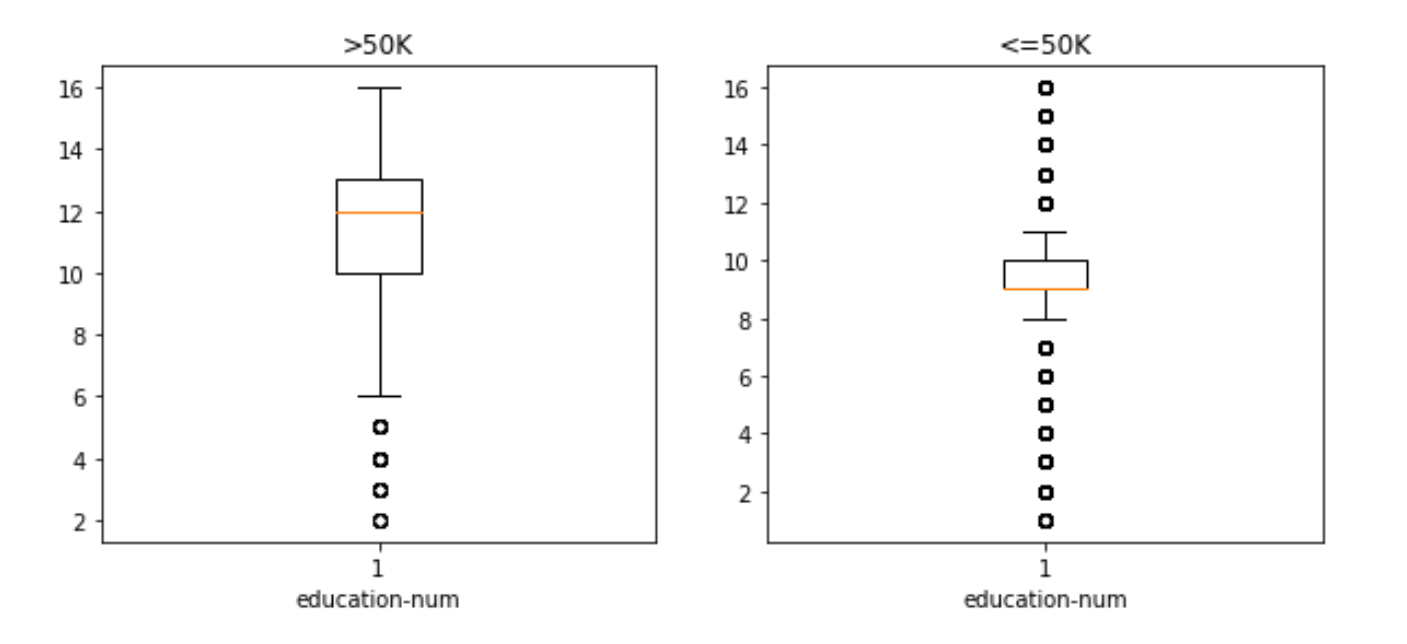
推论：

1. 对于大多数分类数据，这两个类的分布是高度倾斜的，这表明这个特性可以用来区分这两个类。
2. 拥有“Exec-managerial”、“Protective-serv”等职业的个人更有可能获得5万的收入。

**Parallel Coordinate Plot: Capital-gain, Education-num and Age**



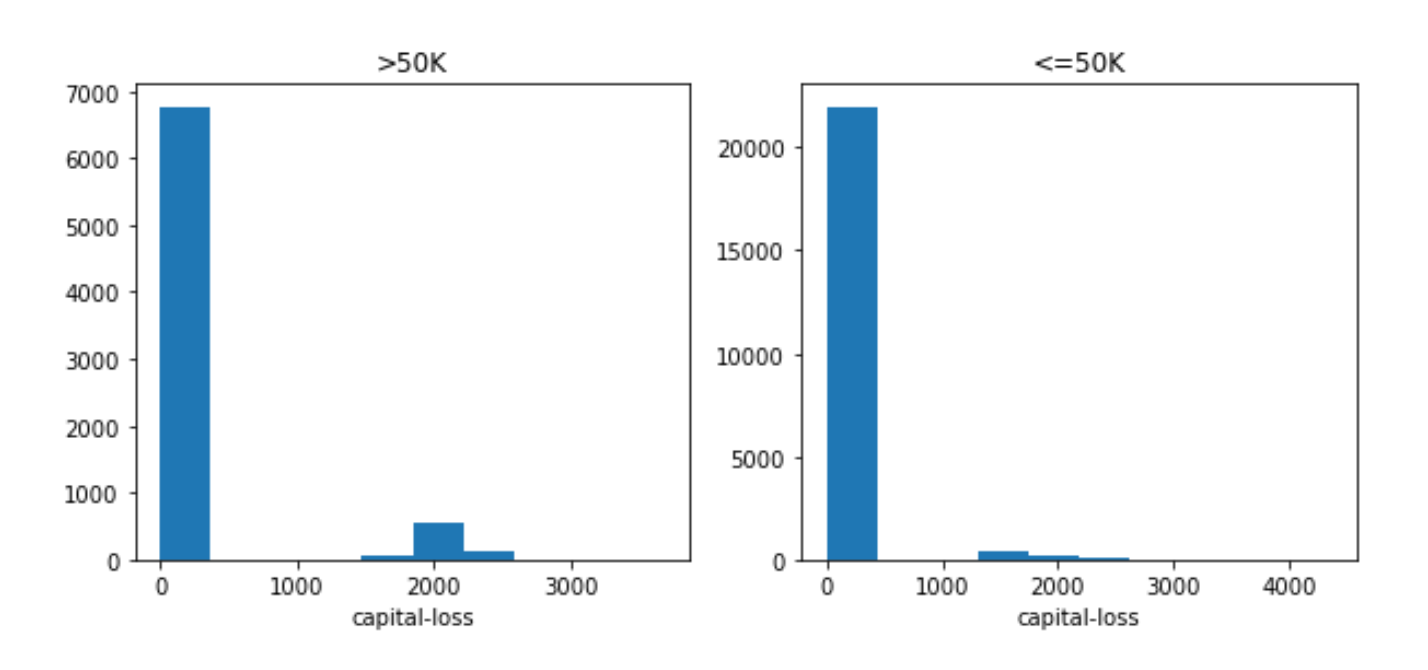
**Box plot: Education-num**



推论：

1. 从平行坐标图中，我们可以通过这三个特征的结合来区分黄线和红线。
2. 从箱型图中，我们可以看到，教育在这两类数据中的分布差异很大。
3. 受过教育程度高的人更有可能收入超过5万。
4. 老年人的收入很可能比年轻人的收入更多。

**Histogram: capital-loss**



推论：

1、如图所示，对于这两类数据，它们都显示出相似的分布情况，这表明这一特性可能不会有助于区分这两类数据。

# Machine Learning Analysis:

# Questions：

# Conclusion and Future Expetations

# （结论及展望）