FIT5196-S1-2023 评估

这是一项个人评估，占你 FIT5196 总分的 30%。

截止日期:见 Moodle

对于这个评估，你需要编写 Python 代码*(允许使用任何 Python 包)*来将多个数据集集成到一个模式中。这个评估的输入和输出如下所示:

表 1。任务的输入和输出

输入	**输出文件**	**Jupyter Notebook 和 .py 文件**
< student_no > . xml(文件), < student_no >。json(文件), Vic_suburb_boundary(目录), Vic_GTFS_data(目录), Lga_to_suburb.pdf(文件), house.speakingsame.com	< student_no > _A3_solution.c sv	< student_no > _ass3.ipynb < student_no > _ass3.py
(网站)(你需要弄清楚这个网站怎么勉强过关你自己。*检查第九周的材料了解更多信息。任何 Python 包是允许的*)		

.py 文件应该是从你的 Jupyter Notebook 文件中生成的，它将用于抄袭检查。(见附录)

每个人都有几个不同格式的数据集，最初的数据是关于澳大利亚维多利亚州的住房信息。**你可以在这里找到你自己的数据集(每个学生不同)**，你也可以在这里找到**补充数据(所有学生都一样)**。在这个作业中，你需要完成以下任务。

任务 1:数据集成(55%)

在这个任务中，你需要使用以下模式将来自多个来源的输入数据集集成到一个数据集中。

表二。最终模式描述。有关的更多信息，请参阅示例输出

列的数据类型和值格式。(阅读 Note3 中默认值的含义)

列	描述
property_id	属性的唯一 id
纬度	属性纬度
液化天然气	属性经度
addr_street	属性地址
郊区(15%)	房产郊区。**默认值:“NA”**
达到(10%)	物业**地方政府辖区**(LGA)。**默认值:“NA”**
closest_train_station_id (5%)	使用 Haversine distance 的距离该物业最近的火车站。**默认值:“NA”**
distance_to_closest_train_station (5%)	距离最近的火车站到该物业的 Haversine 距离。**默认值:“NA”**
travel_min_to_MC (15%)	所以我们假设那里有一个很大的疫苗接种中心 墨尔本中央大楼。这一栏是**四舍五入的平均出行** **直达行程的时间**(分钟)(见直达的定义 从最近的火车站到“墨尔本中心”的行程(见**注 2**) 在所有工作日(星期一至星期五)**7 点**至 9 点发车 **点**。例如，如果从最近的火车出发有 3 次直达旅行 站到墨尔本中央车站在工作日早上 7 点到 9 点之间 分别需要 6、7、8 分钟，那么这一列的值 对于属性应该是整数((6+7+8)/3)。如果**没有直接** 最近的车站和墨尔本中央车站之间**的旅程** 值应设置为**“无直达行程”**。**如果最近的车站** **如果离房产最近的是墨尔本中央车站本身，那么价值** **应该设为 0**。**默认值:“NA”**

direct_journey_flag (5%)	布尔属性，指示是否有直达的旅程 7-9 点之间从最近的车站到墨尔本中央车站 平日(即服务由星期一至星期五全日提供)。 如果有直达行程(即列车之间没有换乘)，则此标志为 1 需要从最近的火车站到墨尔本中央车站 Station)， 否则为 0。**如果距离某属性最近的站点为**

	墨尔本中央车站本身，那么这个值应该设为 1。默认值:“NA”
number_of_houses (5%)	房源郊区的房源数量必须从 house.speakingsame.com 网站上报废。**默认值:“NA”**
number_of_units (5%)	房产郊区的户数必须从 house.speakingsame.com 网站上取消。**默认值:“NA”**
直辖市(5%)	房产郊区的市政当局必须从 house.speakingsame.com 网站上删除。**默认值:“NA”**
aus_born_perc (5%)	房产郊区澳大利亚出生人口的比例必须从 house.speakingsame.com 网站上取消。**默认的** **价值:“NA”**
median_income (5%)	房产郊区人口的收入中位数必须从 house.speakingsame.com 网站上取消。**默认值:** **“NA”**
median_house_price (5%)	房产郊区的“房子”价格中位数必须从 house.speakingsame.com 网站上取消。**默认值:“NA”**
人口(5%)	房产郊区的人口必须从 house.speakingsame.com 网站上剔除。**默认值:“NA”**

注 1:输出的 CSV 文件必须具有与模式上指定的完全相同的列。请注意，在集成模式中指定的格式不正确的输出文件将不会被标记。

注 2:直行是指您可以在旅途的任何地点换乘火车即可到达墨尔本中央车站。所以，当你在最近的车站上火车时，你可以直接去墨尔本中央车站，而不用换另一辆车。

注 3:如果你决定不计算任何必需的列，那么你必须在你的最终数据帧中仍然有该列，所有的值都作为“默认值”。请注意，在集成模式中指定的格式不正确的输出文件将不会被标记。

注 4:不允许使用外部数据来计算集成模式的值。例如，要计算郊区，只能使用 Google Drive 中提供的形状文件。唯一的外部信息来源是 http://house.speakingsame.com/suburb.php 网站。

注 5:shapefile 数据和 lga_to_suburban .pdf 数据可能过时且不正确。您不需要修复它们或检查它们的有效性。

注 6:对于哈弗斯距离(链接)，使用 6378 公里作为地球的半径。

注 7:有关 GTFS 文件的更多信息请阅读此处(链接)。

注 8:在表 2 中，(a%)格式的某些列前面的数字是与该列相关联的分配标记。例如，列“郊区”携带任务 1的总输出标记的 15%。另外，请注意，我们知道百分比的总和是 90%。另外 10%用于数据集成任务期间可能出现的问题，您应该找到并解决这些问题。

任务 2:数据重塑(20%)

在这个任务中，你需要研究不同的归一化/变换方法(即标准化、最小-最大归一化、log、power、box-cox 变换)对网站上被抛弃的列(即 **number_of_houses**、**number_of_units**、**population**、**aus_born_perc**、**median_income**、**median_house_price**)的影响，并观察和解释它们的效果，假设**我们想要开发一个线性模型来使用其他属性预测“median_house_price”**。在重塑数据时，我们有两个主要的标准。首先，我们希望我们的特征在相同的尺度上，其次，我们希望我们的特征与目标变量(即 **median_house_price**)有尽可能多的线性关系。你需要首先探索数据，看看是否需要进行任何缩放或转换(如果是，为什么?如果没有，为什么?)然后执行适当的操作并记录您的结果和观察结果。请注意，您不需要实际构建线性模型，只需要为线性回归模型准备数据。

任务 3:文档(25%)

文档的主要重点是你在任务 2 中解释的质量，但与之前的作业类似，你的笔记本文件应该是一个体面的格式，有适当的章节和子章节。

可交付成果

您必须在 Moodle 上提交以下文件才能成功提交。**我们不会对未完成的提交进行标记!**

1.< student_no >
_A3_solution.csv

2. < student_no > _ass3。ipynb(其中包含 task 1和 task 2文档，每个文档都有自己的部分)(强烈建议使用 ToC)**(所有输出必须保留在.ipynb 中)**

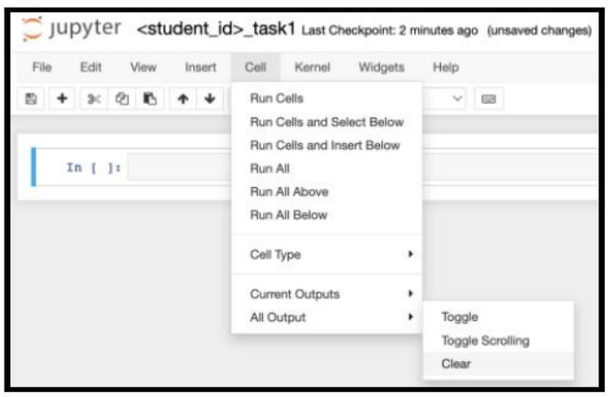
3.<student_no>_ass3.py(此文件用于抄袭检查)

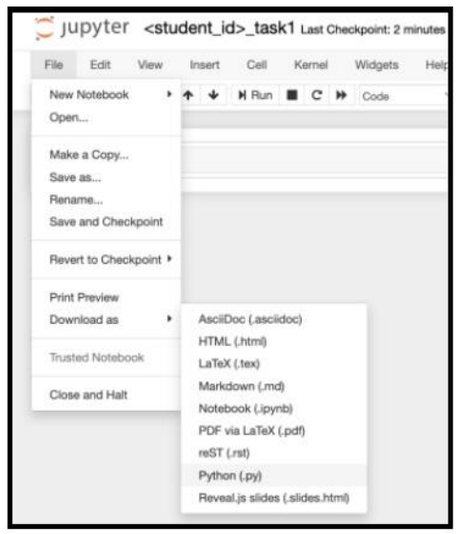
以上文件请以<student_no>_ass3.zip 格式压缩(请勿使用 rar、7z 等格式)。

附录

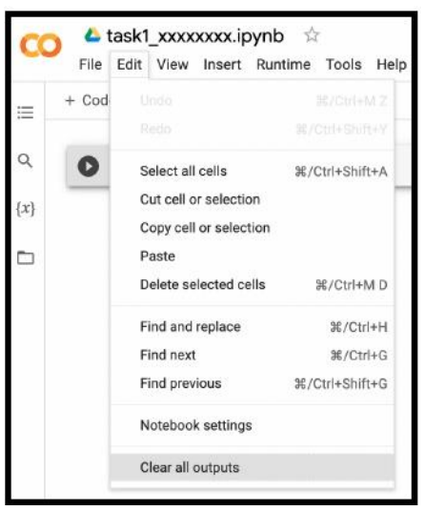
1.要生成.py 文件，需要清除所有单元格输出，然后下载。

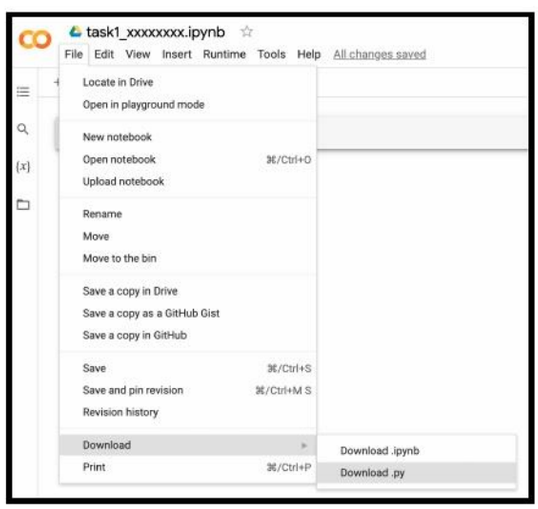
Jupyter 笔记本:





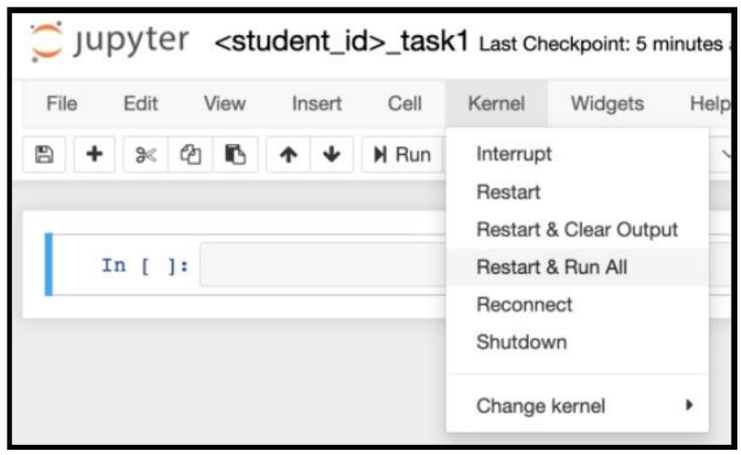
谷歌 colab:





2.为了在提交.ipynb 文件之前生成单元格输出，你需要在保存文件之前运行所有单元格。

Jupyter 笔记本:



谷歌 Colab: