UDACITY

项目

Wrangle and Analyze Data

此部分属于 Data Analyst Nanodegree Program

项目审阅

注释

与大家分享你取得的成绩！

Meets Specifications

针对上一次审阅中提出的问题和建议进行了修改，非常棒！恭喜你完成项目！🎉
在下方的审阅中又提了一些小建议，希望能对以后的学习和工作有所帮助~
数据整理是现实世界中数据分析师工作的一大部分，完成这个项目的你已经具备相当强的数据分析实战经验！
当然，我们完成数据整理的目的是进行数据分析，接下来，我们将学习探索性数据分析，在整理后的数据集中满足好奇心！
继续加油吧！期待你的下一个项目！
Learning by doing! 💪

代码功能与可读性

Jupyter Notebook 具有直观清晰，易于遵循的逻辑结构。 该代码包含清晰有效的注释，并应用在 Jupyter Notebook 的 Markdown 单元格中。 数据清洗过程（即收集、评估和清理）的步骤也通过注释或 Markdown 单元格被清晰地标识出来。

项目中的所有代码都包含在名为 wrangle_act.ipynb 的 Jupyter Notebook 中，并且代码在运行时没有出现错误。

收集数据

学员成功收集数据：

- 在“项目详细信息”页面上使用至少三（3）个不同来源。
- 在“项目详细信息”页面上，使用至少三（3）种不同的文件格式。
  首先将每一条数据导入到一个单独的 pandas 数据框中。

评估数据

学员评估使用了两种方式：

- 可视化评估：每张收集的数据都显示在 Jupyter Notebook 中，以便进行可视化评估。 一旦显示出来，数据可以在外部应用程序（如 Excel，文本编辑器）中进行评估。
- 编程评估：使用 pandas 的功能和/或方法来评估数据。

学员能够检测到至少 八（8）个数据质量问题和两（2）个整洁度问题，包括待清理问题以满足项目要求。每一个问题用一到几句话记录下来。

清理数据

清理过程中的定义，编码和测试步骤都有明确的记录。

在清理之前，保存原始数据的副本。

（如果可能的话）评估阶段确定的所有问题都可以通过 Python 和 pandas 成功清理，并包括满足项目要求所需的清理任务。

学员需要创建一个整洁的主数据集（或者多个数据集，如果有必要的话）与所有收集的数据片段。

按照上一次的审阅建议进行了修改，非常棒！

1. 将清理后的数据集保存到了 csv 文件中；
2. 使用了正确的 np.nan 来替换无效值；
3. 保留了一只狗狗多个地位的情况（使用了建议中的第二种方式）

建议

- 上一次建议的第一种提取 stage 方式参考：

```
# 加了个 .str.lower()，避免提取不到大小写不同的 stage
# 这里将 floofer 地位改为使用 floof 提取，因为有很多其他的写法，比如 floofs floofy floofie 等
twitter_archive_master['stage']=twitter_archive_master.text.str.lower().str.findall('(doggo|pupper|puppo|floof)')
# 对 stage 列调用匿名函数，使其 join 为字符串
twitter_archive_master['stage'] = twitter_archive_master['stage'].apply(lambda x: ','.join(set(x)))
# 替换其中的空值为 np.nan
twitter_archive_master['stage'].replace('', np.nan, inplace = True)
twitter_archive_master['stage'].value_counts()
```

```
pupper          220
doggo            66
floof            33
puppo            27
doggo,pupper      9
doggo,floof       2
doggo,puppo       2
floof,pupper      1
Name: stage, dtype: int64
```

存储并处理清洁过的数据

学员将他们收集、评估和清理过的主数据集保存到 CSV 文件或 SQLite 数据库中。

本次提交在保存清理后的主数据集时，又忘记使用了 index 的参数设定咯~以后还是需要注意这一点！

使用 Jupyter Notebook 中的 pandas 或 SQL 分析主数据集，并生成至少三（3）个独立的结论。

在 Jupyter Notebook 中，使用 Python 绘图库或在 Tableau 中至少生成一（1）个标记的可视化对象。

学员必须在他们的清洗数据中明确他们之后分析和可视化所依据的数据是建立在评估和清理的基础上。

报告

学员需要言简意赅地介绍他们的数据清理。 这一文件（wrangle_report.pdf）大约只需要300-600字。

学员发现至少三（3）个结论，其中至少包含一个（1）可视化。这一文件（act_report.pdf）至少需要 250 个字。

项目文件

提交的文件包括如下：

- wrangle_act.ipynb
- wrangle_report.pdf
- act_report.pdf

并包括所有的数据集文件，如存储的主数据集，并使用在项目提交页面中指定的文件名和扩展名。

下载项目

返回 PATH

给这次审阅打分

学员 FAQ