卡迪夫计算机科学与信息学院

课程评估

课程，模块代码：CMT309，

模块标题：计算数据科学，

评估标题：CMT309 Data Science Portfolio评估编号：2

提交日期和时间：**21-05-2021在9:30 am**

返回日期：25-06-2021之前

此作业相当于此项目可用总分数的70％模块。

如果课程作业提交迟了（并且没有任何缓和的情况）：

1. 如果在截止日期之后不迟于24小时提交评估，则评估分数将以最低及格分数为上限;
2. 如果评估是在截止日期后超过24小时提交，则评分为0。

您的**提交必须包括官方的“课程作业提交封面”表**，该表可以在以下位置找到：

https：//docs.cs.cf.ac.uk /downloads/coursework/Coversheet.pdf

提交说明您的课程作业应在上述截止日期之前通过Learning Central提交。

您必须上传以下文件：

表格

描述已自动生成

✅**1.** 对于封面的文件名，将“ Student\_number”替换为您的学生编号，例如“ C1234567890.pdf”。

**2.** 确保在**所有Python文件**中都**包含**您的**学生编号作为注释**！

**3.** 与提交说明（包括提交的文件数量和类型）的任何偏差都可能导致评估或问题部分的分数减少。您可以提交多个on Learning Central。**仅会标记最后一次尝试中包含的文件**，因此请确保您在**最后一次尝试中上载了所有文件**。

Assignment

Start by downloading the following files from the Assessment tab in Learning Central:

•**Part1\_2.ipynb**(Jupyter Notebook file for **Part 1** and **Part 2**)

•**Part3.ipynb** (Jupyter Notebook file for **Part 3**)

•listings.csv

•reviews.csv

Then answer the following questions. You can use any Python expression or package that was used in the lectures and practical sessions. Additional packages are not allowed unless instructed in the question. You answer the questions by filling in the appropriate sections in the Jupyter Notebook. Export your final Jupyter Notebooks as HTML to produce the corresponding HTML files. Before submitting your Jupyter Notebooks and HTML files, make sure to restart the kernel and execute each cell such that all outputs and figures are visible.

首先从Learning Central的“评估”选项卡中下载以下文件：

•Part1\_2.ipynb（对于第1部分和第2部分的Jupyter Notebook文件）

•Part3.ipynb（对于第3部分的Jupyter Notebook文件）

•listings.csv

•reviews.csv，

然后回答以下问题问题。您可以使用在讲座和实践会议中使用的任何Python表达式或程序包。除非问题中另有说明，否则不允许其他程序包。

**4. 您通过在Jupyter Notebook中**的**适当部分填写来回答问题。**

**5.** 将**最终的Jupyter Notebooks**输出为HTML**生成相应的HTML文件**。

**6.** 在**提交Jupyter Notebook和HTML文件之前**，请确保重新启动内核并执行每个单元，以使所有输出和图形都可见。

Scenario

In this assignment you slip into the role of a Data Scientist who has been hired by Airbnb. Airbnb is an online market place for vacation and short-term rentals of rooms or flats which operates in many countries in the world (see <https://en.wikipedia.org/wiki/Airbnb>).

Airbnb collects data on the listing and users interacting with the platform. Let us define these terms first:

**-User: A user is someone using the Airbnb platform (guest or host).**

**-Guest: A guest is a user who uses the Airbnb platform to book a room, flat or house.**

**-Host: A host is a user who offers a room, flat, or house for rent.**

**-Listing: A listing is a room, flat, or house offered for rent. In the dataset, each row corresponds to a listing.**

Since we do not have access to Airbnb’s internal database, we will instead use data published by the website Inside Airbnb. The data has been acquired by web-scraping publicly available data from Airbnb. It is available online (http://insideairbnb.com/about.html) but you do not need to download any data from this website since we have downloaded and renamed all datasets you need and made them available on Learning Central. In our scenario, you are responsible for the Amsterdam branch of Airbnb operations. Your main task is to provide insights on the data collected in Amsterdam as well as write algorithms to improve the experience of Airbnb users. The assignment is split into three parts. In the first two parts, you will focus on the numerical parts of the data. In the last part, you will focus on the text data.

**场景**

在此任务中，您将扮演Airbnb聘请的数据科学家的角色。

Airbnb是在线度假或短期出租房间或公寓的网上市场，在世界许多国家/地区都有营业（请参阅https://en.wikipedia.org/wiki/Airbnb）。 Airbnb会在列表中收集数据并与平台进行交互的用户，让我们首先定义以下术语：

-用户：用户是使用Airbnb平台的人（**来宾或房东**）。

-客人：来宾是使用Airbnb平台进行预订房间，单位或房屋的用户。

-Host：是提供房间，单位或房屋出租的用户。

-Listing列表：列表是指提供出租的房间，单位或房屋。在数据集中，每行对应于一个列表。

由于我们无权访问Airbnb的内部数据库，因此，我们将使用由**Inside Airbnb**网站发布的数据。该数据是通过从Airbnb抓取可公开获取的数据进行网络获取的。它可以在线访问（http://insideairbnb.com/about.html），但是您无需从该网站下载任何数据，因为我们已经下载并重命名了所需的所有数据集，并在Learning Central上提供了这些数据集。

您负责Airbnb的Amsterdam branch业务。

**7. 您的主要任务是：**

**提供有关阿姆斯特丹收集的数据的见解；**

**以及编写算法以改善Airbnb用户的体验。**

任务分为三部分。

* 在前两部分中，您将**重点研究数据的数字部分**。【Part1\_2.ipynb（对于第1部分和第2部分的Jupyter Notebook文件）】
* 在最后一部分中，您将**重点关注文本数据**。【Part3.ipynb（对于第3部分的Jupyter Notebook文件）】

**Part 1: Pre-processing and exploratory analysis**

You start by reading the csv file into a Pandas DataFrame, cleaning the data and removing unwanted columns, performing conversions and calculating new columns. Then, you will perform exploratory analysis to look at the properties and distribution of the data, and answer a couple of questions your manager put forward to you.

**第1部分：预处理和探索性分析**

**8.** 您首先需要将csv文件读取到Pandas DataFrame中，**清理数据并删除不需要的列，执行转换并计算新列**。然后，您将**进行探索性分析**，以**查看数据的属性和分布**，并**回答经理向您提出的几个问题**。

**•Part 2: Statistical analysis and recommender systems**

Starting from the pre-processed DataFrame, you will perform statistical analysis using t-tests and linear regression to identify variables that significantly affect the price of the rent.

Then, you will design a series of recommender systems that have been requested by users: a function that helps in setting the price for someone offering a new property, and a function that helps in selecting a city to visit given a particular budget.

**第2部分：统计分析和推荐系统**

**9.** 从预处理的DataFrame开始，您将使用**t检验**和**线性回归**进行统计分析，**以识别对租金价格有重大影响的变量**。

**10.** 然后，您将**设计一系列用户所要求的推荐系统**：

**10.1** 该功能有助于**为提供新房地产的人设定价格**（a function that helps in setting the price for someone offering a new property）

**10.2** 该功能有助于**在给定特定预算的情况**下**选择**要参观的**城市** （a function that helps in selecting a city to visit **given a particular budget**）

**•Part 3: Text analysis and ethics**

You will mostly work with unstructured text data in a Pandas Dataframe representing user reviews. To get you started with the dataset, you can have a look at a corresponding Jupyter Notebook(note: following and working through this link is not required for this assignment, but maybe you find some inspiration or useful information there). You are allowed to re-use code from the Jupyter Notebook provided that you properly reference it.

**第3部分：文本分析和道德**

您将主要使用**表示用户评论**的Pandas Dataframe中的**非结构化文本数据**。

为了让您开始使用数据集，您可以看一下相应的Jupyter Notebook（注意：

此任务不需要遵循并研究此链接，但是也许您可以在其中找到一些启发或有用的信息）。如果您正确地引用了Jupyter笔记本，则可以重复使用该代码。

Reference: https://github.com/gracecarrillo/Predicting-Airbnb-prices-with-machine-learning-and-location-data/blob/gh-pages/Exploring\_Edinburgh\_Graciela\_Carrillo.ipynb

**Part1–Pre-processing and exploratory analysis [25marks]**

You answer this question by filling in the first part of the Part1\_2.ipynbJupyter Notebook.

**Question 1a–Drop columns (Total 4marks)**

When reading in the dataframe using the load\_csv function, one can see that it contains a lot of textual data which will not be relevant for the numerical analyses in Part 1 and Part 2. Therefore, implement two functions drop\_cols and drop\_cols\_na which remove some of the columns. Detailed instructions:[2marks each]

• drop\_cols(df):

takes the dataframe as an input. It returns the reduced dataframe after dropping the following columns: 'scrape\_id','last\_scraped','description','listing\_url','neighbourhood','calendar\_last\_scraped', 'amenities','neighborhood\_overview', 'picture\_url','host\_url', 'host\_about', 'host\_location','host\_total\_listings\_count','host\_thumbnail\_url','host\_picture\_url', 'host\_verifications','bathrooms\_text','has\_availability','minimum\_minimum\_nights','maximum\_minimum\_nights','minimum\_maximum\_nights','maximum\_maximum\_nights','minimum\_nights\_avg\_ntm','maximum\_nights\_avg\_ntm','number\_of\_reviews\_l30d','calculated\_host\_listings\_count','calculated\_host\_listings\_count\_entire\_homes','calculated\_host\_listings\_count\_private\_rooms','calculated\_host\_listings\_count\_shared\_rooms'

**第1部分–预处理和探索性分析[25分]**

您可以通过填写Part1\_2.ipynbJupyter Notebook的第一部分来回答此问题。

**问题1a –删除列（共4分）**

使用load\_csv函数读取数据框中的内容时，您会看到包含大量文本数据，这些文本数据与第1部分和第2部分中的数值分析无关。因此，实现两个函数drop\_cols和drop\_cols\_na可以删除一些列。详细说明：[每个2分]

• drop\_cols（df）：将数据帧作为输入。

删除以下列后，它返回缩小的数据帧：“ scrape\_id”，“ last\_scraped”，“ description”，“ listing\_url”，“ neighbourhood”，“ calendar\_last\_scraped”，“ amenities”，“ neighborhood\_overview”，“ picture\_url”，“ host\_url” ，'host\_about'，'host\_location'，'host\_total\_listings\_count'，'host\_thumbnail\_url'，'host\_picture\_url'，'host\_verifications'，'bathrooms\_text'，'has\_availability'，'minimum\_minimum\_nights'，'maximum\_minimum\_nights'， minimum\_nights\_avg\_ntm'，'maximum\_nights\_avg\_ntm'，'number\_of\_reviews\_l30d'，'calculated\_host\_listings\_count'，'calculated\_host\_listings\_count\_entire\_homes'，'calculated\_host\_listings\_count\_private\_rooms，

•drop\_cols\_na（df，threshold）：

根据列中包含的NaN值的数量来删除列。 threshold是介于0和1之间的分数。如果一列中NaNs的分数等于或大于阈值，则会删除相应的列。

例如，如果threshold为0.5，则删除NaN大于或等于50％的所有列。阈值的默认值为0.5。要解决此问题，请完成笔记本电脑中问题1a中的功能。

•drop\_cols\_na(df, threshold):

drop columns according to the amount of NaN values they contain. Threshold is a fraction between 0 and 1. If the fraction of NaNs in a column is equal or larger than the threshold, the respective columns is dropped. For instance, if threshold is 0.5, all columns that have at 50%or more NaNs are dropped. The default value for threshold is 0.5. To solve the question, complete the functions in Question 1a in the notebook.

**Question 1b –Recode and add columns (Total 6marks)**

We continue pre-processing by recoding some of the columns and adding columns with additional information. To this end, implement the following functions: binary\_encoding, add\_host\_days, convert\_price. Detailed instructions:[2 marks each]

**问题1b –重新编码并添加列（共6分）**

我们通过对一些列进行重新编码并添加带有其他信息的列来继续进行预处理。为此，请实现以下功能：binary\_encoding，add\_host\_days，convert\_price。详细说明：[每个2分]

**•binary\_encoding(df):** if we have a close look at some of the columns, we can see that some of them are binary variables representing 1(True)and 0(False)values. However, the values are encoded as the strings ‘t’ (for True) and ‘f’ (for False). Recode these columns by turning them into the integer numbers 0 and 1.

**•binary\_encoding（df）：**

如果仔细查看某些列，我们可以看到其中一些是表示1（真）和0（假）值的二进制变量。但是，这些值被编码为字符串“ t”（对于True）和“ f”（对于False）。通过将它们转换为整数0和1来重新编码这些列。

**•add\_host\_days(df):** it would be useful to have a column that represents the number of days (with respect to the current date) that the host has been registered. To this end, create a new column called 'host\_days' (hint: look into Pandas’ to\_datetimemethod).

**•add\_host\_days（df）：**

具有一个代表主机已注册天数（相对于当前日期）的列将很有用。为此，请创建一个名为“ host\_days”的新列（提示：查看熊猫的to\_datetime方法）。

**•convert\_price(df):** the ‘price’ column represents the nightly price in USD. However the prices are encoded as strings. This function should convert the prices into floating point numbers. For instance, the string ‘$40’ should be converted to the floating point number 40.0.To solve the question, complete the functions in Question 1b in the notebook.

convert\_price（df）：

“价格”列表示美元的夜间价格。但是，价格被编码为字符串。此功能应将价格**转换为浮点数**。例如，字符串“ $ 40”应转换为浮点数40.0。要解决该问题，请在笔记本中完成问题1b中的功能。

**Question 1c –Answering questions (Total 6marks)**

Your manager has a couple of questions about the dataset, and provided you with a list of questions they want answered. The questions are as follows:[1 mark each]

**问题1c –回答问题（总计6分）**

您的经理有几个关于数据集，并为您提供了他们想回答的问题列表。问题如下：

[每个标记1分]

•How many hosts offer two or more properties for rent?

•What is the highest price for a listing?

•What is the ID of the listing that has the largest number of bedrooms?

•What is the ID of the listing with the largest advertised price?

•There are different room types. How many listings are there for the most common room type?

•How many hosts are there that have been registered for more than 3000 days?

To solve the questions, provide corresponding Pandas code in Question 1c in the notebook.

•多少房东提供两个或更多房屋出租？

•房源的最高价格是多少？

•卧室数量最多的房源的ID是什么？

•广告价格最高的列表的ID是什么？

•房间类型不同。最常见的房间类型有多少个列表？

•已注册3000天以上的主机有多少个？要解决问题，请在笔记本的问题1c中提供相应的熊猫代码。

**Question 1d –Exploratory analyses (Total 9marks)**

To further explore the dataset, you produce a number of exploratory plots. In particular, you set out to produce the following three plots:[3 marks each]

**问题1d –探索性分析（共9分）**

要进一步探索数据集，您可以生成许多探索性图。特别是，您着手制作以下三个图：[每个3分]

**-Plot 1:**

A barplot with horizontal bars representing average nightly prices. Produce a bar for each neighbourhood (use the neighbourhood\_cleansed column).

**图1：**带有水平条的条形图，代表平均每晚价格。

为每个邻域生成一个栏（使用neighbourhood\_cleansed column）。

图表, 条形图

描述已自动生成

**-Plot 2:**

The review ratings have 5 additional subitems: cleanliness, checkin, communication, location, value. You are interested in the correlations between the subitems. To this end, produce a correlation matrix that depicts all pair-wise Pearson correlations between these 5 variables.

**-图2：**评论评分具有5个其他子项：

清洁度（cleanliness），签到（checkin），沟通（communication），位置（location），价值（value）。

您对子项目之间的相关性感兴趣。为此，生成一个相关矩阵，描述这5个变量之间的所有成对的Pearson相关。

图片包含 正方形

描述已自动生成

**-Plot 3:**

Your manager is interested in the geographical distribution of nightly prices for the more expensive listings. To this end, produce a scatterplot using latitude/longitude as coordinates. The price should be encoded both by color and size of the circles. Make sure to include only those listings with a price larger than $150.

**-图3：**您的经理对较贵商品的夜间价格的地理分布感兴趣。

为此，使用纬度/经度作为坐标生成散点图。价格应使用圆圈的颜色和大小进行编码。请确保仅包括价格大于$ 150的那些商品。

图表, 散点图

描述已自动生成

The following figures are indications of how such plots might look like. Note that title, labels etc are missing. It is just intended for orientation. Your solution can have a different style, colors, etc.

上面图说明了这些图的外观。注意缺少标题，标签等，它仅用于定向。您的解决方案可以具有不同的样式，颜色等。

**Part 2 –Statistical analysis and recommender systems [35marks]**

You answer this question by filling in the second part of the Part1\_2.ipynbJupyter Notebook.

**Question 2a–Linear regression and t-tests(Total 8marks)**

In this question you will perform a linear regression and a number of t-tests. Detailed instructions:[4 marks each]

**第2部分–统计分析和推荐系统[35分]**

您可以通过填写Part1\_2.ipynbJupyter Notebook的第二部分来回答此问题。

**问题2a-线性回归和t检验（总计8分）**

在此问题中，您将执行线性回归和大量的t检验。详细说明：[每个标记4分]

-Linear regression:

Review scores consist of an overall rating as well as additional scores on subitems (accuracy, cleanliness, checkin, communication, location, value). Can the overall rating be predicted from the scores on the subitems? To investigate this, perform a linear regression on rating using statsmodels. After fitting the model, print the summary. Make sure that the variable names shown in the summary are short and readable (accuracy, cleanliness, checkin, communication, location, value).

-线性回归：

评论评分包括总体评分以及子项目的其他评分（准确性，整洁度，入住率，沟通能力，位置，价值）。可以根据子项目的分数预测整体评分吗？要对此进行调查，请使用statsmodels对评分进行线性回归。拟合模型后，打印摘要。确保摘要中显示的变量名称简短易读（准确性，整洁度，值机，沟通，位置，价值）

-T-tests: You want to investigate whether the difference in prices for different room types is significant. To investigate this, perform independent samples t-tests for the prices of each combination of room types using alpha = 0.01.Which room types are significantly different in terms of room type? Do your conclusions change if you perform Bonferroni correction to the alpha level(see https://en.wikipedia.org/wiki/Bonferroni\_correction).Finally, create a Dataframe with 4 rows and 4 columns that holds the p-values for all pairwise combinations. The column/row names should be the different room types, and each entry in the dataframe corresponds to the p-value for the respective combination of room types.

-T-tests：

您想调查不同房型的价格差异是否显着。要对此进行调查，请使用alpha = 0.01对每种房型组合的价格进行独立的样本t检验。最后，创建一个包含4行4列的数据框，其中包含所有成对组合的p值。列/行名称应为不同的房间类型，并且数据框中的每个条目均对应于相应房间类型组合的p值。

To solve this question, provide corresponding code in Question 2ain the notebook and provide short answers in the space designated as YOUR ANSWER.

要解决此问题，请在笔记本中的问题2a中提供相应的代码，并在指定为YOUR的空间中提供简短答案答案

**Question 2b–Linear regression with variable selection**

**(Total 15marks)**

In this question you investigate how well nightly price can be predicted from the other variables in the dataframe. You need to decide yourself which variables to choose, but make sure you have at least 10variables.The only requirement is that you use room\_type as a predictor. Because room\_type is a categorical variable, you first have to use dummy coding to turn it into a number of binary variables(hint:pd.get\_dummies()).In the notebook, provide a short explanation for your choice of variables. Starting from the variables you have chosen, our goal is to derive a sparse model with fewer variables. This process is called variable selection. In variable selection(‘variable’ means the same as ‘predictor’), variables get iteratively added or removed from the regression model. Once finished, the model typically contains only a subset of the original variables. It makes it easier to interpret the model, and in some cases it makes it generalise better to new data. To perform variable selection, implement a function

**2b-带有变量选择的线性回归（总计15分）**

在这个问题中，您将研究如何从数据框中的其他变量，预测夜间价格的水平。

您需要自己决定要选择哪些变量，但要确保至少有10个变量。

唯一的要求是将room\_type用作预测变量。由于room\_type是分类变量，因此首先必须使用伪编码将其转换为许多二进制变量（hint：pd.get\_dummies（））。

从您选择的变量开始，我们的目标是导出具有较少变量的稀疏模型。此过程称为变量选择。在变量选择中（“变量”的含义与“预测变量”相同），变量可以迭代地从回归模型中添加或删除。完成后，模型通常只包含原始变量的一个子集，这使得模型解释更容易，并且在某些情况下，它可以更好地将其推广到新数据中。要执行变量选择，请实现一个函数

**要执行变量选择，请实现一个函数**

**variable\_selection（df， predictors（预测变量），target（目标），alpha），**

其中**df:** 是列表数据框，

**predictors预测变量:** 是列表包含至少10个变量的初始选择 （例如['bedrooms，'beds'，...]）；

**target:** 是回归的目标变量（例如“price(价格)”）；

**alpha:** 是选择重要预测变量的显着性水平（例如0.05）。

该函数返回pred，即原始预测变量的选定子集。

variable\_selection (df, predictors, target, alpha) where df is the listings dataframe, predictors is a list you’re your initial selection of at least 10 variables (e.g. [‘bedrooms’, ‘beds’, ...]), target is the target variable for the regression (e.g. ‘price’), and alpha is the significance level for selecting significant predictors(e.g. 0.05).The function returns pred, the selected subset of the original predictors.

To calculate regression fits and p-values you can use statsmodels. Your approach operates in two stages:

In stage 1, you build a model by adding variables one after the other. You keep adding variables that increase the adjusted R2coefficient. You do not need to calculate it by hand, it is provided by statsmodels package.

In stage 2, starting from these variables, if any of them are not significant, you keep removing variables until all variables in the model are significant. The output of the second stage is your final set of variables.

**要计算回归拟合和p值，您可以使用statsmodels。**

分两个阶段进行：

1. 在阶段1中，通过一个接一个地添加变量来构建模型。您一直在添加变量，以增加调整后的R2系数。（备注：您不需要手动计算它，它由statsmodelspackage提供。）；
2. 在第2阶段中，从这些变量开始（如果它们中的任何一个都不重要），请继续删除变量，直到模型中的所有变量都有效为止。第二阶段的输出是您的最终变量集。

**Let us look at the two stages in detail:**

**Stage 1(add variables)[8 marks]**

•Start with an empty set of variables

•Fit multiple one-variable regression models. In each iteration, use one of the variables provided in predictors. The variable that leads to the largest increase in adjusted R2is added to the model.

•Now proceed by adding a second variable into the model. Starting from the remaining variables, again choose the variable that leads to the largest increase in adjusted R2.

•Continue in thesame way for the third, fourth, ... variable.

•You are finished when there is no variable left that increases adjusted R2.

**让我们详细看两个阶段：**

**阶段1（附加变量）[8分]**

•从一组空变量开始

•拟合多个单变量回归模型。在每次迭代中，使用预测变量中提供的变量之一。导致调整后R2的最大增加的变量被添加到模型中。

•现在，将第二个变量添加到模型中继续进行。从其余变量开始，再次选择导致调整后R2最大增加的变量。

•以相同的方式继续处理第三个，第四个，...变量。

•当没有剩余的变量增加调整后的R2时结束。

**Stage 2(remove non-significant variables)[7 marks]**

It is possible that some of the variables from the previous stage are not significant. We call a variable" significant" if the p-value of its coefficient is smaller or equal to the given threshold alpha.

•Start by fitting a model using the variables that have been added to the model in Stage 1.

•If there is a variable that is not significant, remove the variable with the largest p-value and fit the model again with the reduced set of variables.

•Keep removing variables and re-fitting the model until all remaining variables are significant.

•The remaining significant variables are the output of your function.

To solve this question, provide corresponding code in Question 2b in the notebook and provide a short answer in the space following YOUR ANSWER. To test your function, add a function call with your selection of predictors and alpha level.

**阶段2（删除不重要的变量）[7分]**

可能上阶段部分变量并不重要significant。

如果其系数的p值小于或等于给定的阈值alpha, 我们将其称为“显着”变量(significant variable)：

•首先使用在阶段1中已添加到模型中的变量来拟合模型。

•如果存在不重要的变量，请删除该变量使用最大的p值，并使用减少的变量集再次拟合模型。

•持续移除变量并重新拟合模型，直到所有剩余变量都有效为止。

•剩余的有效变量就是函数的输出。

在笔记本中的问题2b中提供相应的代码，并在您的答案后的空格中提供简短答案。要测试您的功能，请添加带有您选择的预测变量和alpha级别的函数调用。

**Question 2c –Recommender systems (Total 12marks)**

There have been requests from customers to provide automated recommendations. First, guests requested a recommender system that helps them identifying neighbourhoods in a city that fit their budget. Second, hosts who want to offer new listings would like a recommender

system that suggests a nightly price. As a response to this request, you and your team have worked out specifications for two recommender systems.

**问题2c –推荐系统（共12分）**

客户有要求提供自动推荐的要求。

首先，**客人**要求使用推荐系统，以**帮助他们确定适合自己预算的城市社区**。

其次，想要**提供新列表的房东**希望有一个推荐系统**能建议每晚价格的系统**。

为响应此请求，您和您的团队已经制定了两种推荐系统的规范。

**推荐系统1：推荐有预算的邻居[6分]**

预算有限的游客一直在寻求一种工具，使他们可以快速查看城市中哪个社区在预算范围内提供最多的住宿机会。

您想要实现一个提供此功能的Python函数。计划是将Python函数集成到Airbnb网站中。经过仔细考虑，您认为该函数应符合以下规范：

•用户应能够指定其预算范围，即最小和最大预算。

例如，用户可能会寻找价格在$ 10- $ 50范围内的房地产。

另一个正在寻找豪华住宿的用户可能会选择$ 100- $ 500的价格范围。

•您的功能会确定哪个邻里拥有该范围内最多的房产。它返回代表邻居名称的字符串。对邻居名称使用neighbourhood\_cleansed变量。

•某些邻居比其他邻居具有更多的属性，因此，仅考虑绝对数字，算法总是会首选具有更多属性的邻居。另一种选择是考虑相对数量，即预算范围内列表的数量相对于给定社区内列表总数的比例。用户应该能够选择是否要使用绝对数或相对数。

从这些规范中，您将获得以下函数功能：

**recommend\_neighbourhood （df，budget\_min，budget\_max，relative），**

其中

**df:**是您的列表数据框，

**变量budget\_min和budget\_max**： 是表示预算支架的浮点数。这些数字包含在内，即每晚价格也正好等于所考虑的budget\_minor budget\_maxis。

**变量relative：**是一个布尔值，指定在建议中是否应考虑相对数（分数）。如果为False，则改为使用绝对数字。

**推荐系统2：房东的价格推荐者[6分]**

如果新房东想在Airbnb上提供其房间/公寓/房屋，他们需要决定每晚的价格。目前尚无官方指导，但房东一直在要求Airbnb提供一种算法。经过仔细考虑，您认为该功能应符合以下规范：

•用户必须提供其属性的地理位置（纬度和经度）。

•您的算法搜索地理上最接近的属性（用欧几里德距离来衡量，即经度/纬度），已在Airbnb上列出。您的价格建议将是这些邻近物业的每晚价格的平均值。

•用户应该能够设置要考虑的邻近属性（称为邻居）的数量。邻居越多，表示地理区域就越大。

•您可以忽略一个事实，即某些邻居可能位于不同的邻居中

•每晚价格因房间类型而异。用户应该能够设置所需的房间类型。

如果定义了房间类型，则仅考虑各个房间类型的属性。

从这些规范中，您将获得以下函数功能：

recommend\_price（df，latitude（纬度），longitude（经度），n\_neighbours，room\_type）

其中纬度和经度：属性地理位置，

n\_neighbours： 用户要考虑的相邻属性的数量。

room\_type（如果指定）将neighbourssearch限制为给定房间类型的属性；它应默认为None（无），这意味着将考虑任何属性类型。

要测试您的两个推荐系统，请为两个功能中的每个功能提供功能调用。您可以自由选择函数调用的参数。要解决此问题，请在笔记本的问题2c中提供相应的代码

**Recommender system 1:**

**Recommend a neighbourhood given a budget[6 marks]**

Guests who are traveling on a budget have been requesting a tool that allows them to quickly see which neighbourhood in a city offers most accommodation opportunities within their budget bracket. You want to implement a Python function that delivers this functionality. The plan is to integrate the Python function into the Airbnb website. After some deliberation, you work out that the function should meet these specifications:

•The user should be able to specify their budget bracket, that is, a minimum and maximum budget. For instance, a user might look for properties priced in the $10-$50 range. Another user looking for luxury accommodation may opt for a $100-$500 range.

•Your function identifies which neighbourhood has the largest number of properties within this range. It returns a string representing the name of the neighbourhood.Use the neighbourhood\_cleansed variablefor the names of the neighbourhoods.

•Some neighbourhoods have more properties than others, so by considering only absolute numbers neighbourhoods with more properties could always be preferred by your algorithm. An alternative is to consider relative numbers, that is, the proportion of listings within the budget bracket relative to the total number of listings within a given neighbourhood. The user should be able to select whether they want absolute or relative numbers. From these specifications, you arrive at the following function signature: recommend\_neighbourhood(df, budget\_min, budget\_max, relative)with df being your listings dataframe, the variables budget\_min and budget\_max being floating point numbers representing the budget bracket. The numbers are inclusive, i.e., a nightly price exactly equal to budget\_min or budget\_max is considered, too. The variable relative is a Boolean specifying whether relative numbers(fractions)should be considered in the recommendation. If False, absolute numbers are considered instead.

**Recommender system 2:**

**Price recommender for hosts [6 marks]**

If a new host wants to offer their room / flat / house on Airbnb, they need to decide on what the nightly price will be. There is no official guidance but hosts have been requesting for Airbnb to provide an algorithm. After some deliberation, you work out that the function should meet these specifications:

•The user has to provide the geolocation (latitude and longitude) of their property.

•Your algorithm searches for the geographically closest properties(simply measured by Euclidean distance in terms of latitude/longitude)that are already listed on Airbnb. Your price recommendation will be the mean of the nightly prices of these close by properties.

•The user should be able to set the number of close by properties(called neighbours)that are considered. A larger number of neighbours indicates a larger geographical area.

•You can ignore the fact that some neighbours could be in a different neighbourhood.

•Nightly prices are quite different for different room types. The user should be able to set the desired room type. If the room type is defined, only properties of the respective room type are taken into consideration. From these specifications, you arrive at the following function signature: recommend\_price(df, latitude, longitude, n\_neighbours, room\_type)with the variables being latitude and longitude representing geolocation of the property, n\_neighbours the number of neighbouring properties the user wants to take into account. room\_type, if specified, restricts the neighbours search to properties of the given room type; it should default to None which means that any property type is considered. To test your two recommendation system, provide function calls for each of the two functions. You can freely select the parameters of the function call. To solve this question, provide corresponding code in Question 2c in the notebook.

**Part 3 –Text analysis and ethics[40 marks]**

**第3部分–文本分析和道德规范[40分]**

You answer this question by filling in the Part3.ipynb Jupyter Notebook. In this part, you will be working with the reviews.csv file providing reviews for the listings, and more specifically, the ‘comments’ column.

您可以通过填写Part3来回答此问题。 在本部分中，您将使用reviews.csv文件为清单（尤其是'comments'列）提供评论。

**第3部分–文本分析和道德规范[40分]**

您可以通过填写Part3.ipynb Jupyter Notebook来回答此问题。在本部分中，您将使用reviews.csv文件为清单（尤其是'comments'列）提供评论。

**Question 3a –Pointwise Mutual Information**

**(Total 30marks)**

**问题3a – Pointwise Mutual Information（共30分）**

在此问题中，您将实现并应用Pointwise互信息（PMI）度量标准，这是1992年引入Airbnb评论的单词关联度量标准。 PMI的目的是从自由文本中提取词对，而这往往比偶然发生的机会要多。例如，PMI（'new'，'york'）的得分将高于PMI（'new'，'car'），因为在文本中一起找到“ new”和“ york”的机会要高于“ new”和“car”, 尽管“new”比“york”更常见。通过在我们的评论中**提取具有较高PMI分数的单词对**，我们将能够更好地了解人们的感受和谈论某些感兴趣的事物（例如“窗户”或“位置”）。

In this question, you implement and apply the pointwise mutual information (PMI) metric, a word association metric introduced in 1992, to the Airbnb reviews. The purpose of PMI is to extract, from free text, pairs of words than tend to co-occur together more often than expected by chance. For example, PMI(‘new’, ‘york’)would give a higher score than PMI(‘new’, ‘car’) because the chance of finding ‘new’ and ‘york’ together in text is higherthan ‘new’ and ‘car’, despite ‘new’ being a more frequent word than ‘york’. By extracting word pairs with high PMI score in our reviews, we will be able to understand better how people feel and talk about certain items of interest (e.g., ‘windows’ or ‘location’).

**PMI的公式（其中x和y是两个词）是：**

文本

描述已自动生成

观看此视频以了解如何估计这些概率。

您的解决方案将涉及以下步骤：

**1.（4分）处理原始评论，按此特定顺序应用以下步骤：**

A．标记所有评论。使用nltk的word\_tokenizemethod.

B．词性（PoS）标记：能够区分名词与形容词或动词。使用nltk的pos\_tag方法。

C．小写：减少词汇量

**1. (4 marks)** Processing the raw reviews, applying the following steps in this specific order:

a. Tokenize all reviews. Use nltk’s word\_tokenizemethod.

b. Part-of-speech (PoS) tagging: to be able to differentiate nouns from adjectives or verbs. Use nltk’s pos\_tag method.

c. Lower case: to reduce the size of the vocabulary.

**What to implement:** A function process\_reviews(df)that will take as input the original dataframe and will return it with three additional columns: tokenized, tagged and lower\_tagged, which correspond to steps a, band cdescribed above.

实现内容：一个函数process\_reviews（df）将原始数据帧作为输入，返回它的三个附加列：tokenized，tagged和lower\_tagged，它们对应于上述的步骤a,b, c。

**2. (5 marks)** Starting from the output of step 1.c(tokenized, PoS-tagged and lower cased reviews), create a vocabulary of ‘center’ (the x in the PMI equation) and ‘context’ (the yin the PMI equation) words. Your vocabulary of center words will be the 1,000 most frequent NOUNS (words with a PoS tag starting with ‘N’), and the context words will be the 1,000 most frequent words tagged as either VERB or ADJECTIVE (words with any PoS tag starting with either ‘J’ or ‘V’).

**What to implement:** A function get\_vocab(df)which takes as input the DataFrame generated in step 1.c, and returns two lists, one for the 1,000 most frequent center words (nouns) and one for the 1,000 most frequent context words (either verbs or adjectives).

**2.（5分）**从步骤1.c（标记，带PoS标签和小写的评论）的输出开始，创建一个单词“ center”（x在PMI方程中）和“ context”（y在PMI方程中）。

您的中心单词词汇将是1,000个最常见的名词（带有以'N'开头的PoS标签的单词），上下文单词将是被标记为VERB或ADJECTIVE的1,000个最常见的单词（带有PoS标签的单词以开头“ J”或“ V”）。

**实现方法：**一个函数get\_vocab（df），它将在步骤1.c中生成的DataFrame作为输入，并返回两个列表，一个用于1,000个最频繁的中心词（名词），另一个用于1,000个最频繁的上下文词（动词或形容词）。

**3.(8 marks)** With these two 1,000-word vocabularies, create a co-occurrence matrix where, for each center word, you keep track of how many of the context words co-occur with it. Consider this short review with only one sentence as an example, where we want to get co-occurrences for verbs and adjectives for the center word restaurant: a. ‘A big restaurant served delicious food in big dishes’>>> {‘restaurant’: {‘big’: 2, ‘served’:1, ‘delicious’:1}}

**3.（8分）**使用这两个1,000字的词汇表，创建一个共现矩阵，在该矩阵中，对于每个中心词，您将跟踪与之共存的上下文字词的数量。以仅以一个句子为例的简短回顾为例，在此我们希望获得中心单词餐厅的动词和形容词的共现：a。“ bigrestaurant在大碟中提供美味的食物” >>> {'restaurant'：{'big'：2， 'served'：1，'delicious'：1}}

**What to implement:** A function get\_coocs(df, center\_vocab, context\_vocab) which takes as input the DataFrame generated in step 1, and the lists generated in step 2and returns a dictionary of dictionaries, of the form in the example above. It is up to you how you define context (full review? per sentence? a sliding window of fixed size?), and how to deal with exceptional cases (center words occurring more than once, center and context words being part of your vocabulary because they are frequent both as a noun and as a verb, etc). Use comments in your code to justify your approach.

**实现内容：**函数get\_coocs（df，center\_vocab，context\_vocab）将步骤1中生成的DataFrame和步骤2中生成的列表作为输入，以上面示例中的形式返回字典字典。您如何定义上下文（全面审阅？每个句子？固定大小的滑动窗口？），以及如何处理特殊情况（中心词多次出现，中心词和上下文词经常出现在您的词汇表中，因为它们经常出现作为名词和动词等）。在代码中使用注释来证明您的方法合理。

**4.(3marks)** After you have computed co-occurrences from all the reviews, you should convert the co-occurrence dictionary as a pandas DataFrame. The DataFrame should have 1,000 rows and 1,000 columns. **4.（3分）**从所有评论计算出共现后，应将共现字典转换为pandas DataFrame。 DataFrame应该具有1,000行和1,000列。

**What to implement:** A function called cooc\_dict2df(cooc\_dict), which takes as input the dictionary of dictionaries generated in step 3and returns a DataFrame where each row corresponds to one center word, and each column corresponds to one context word, and cells are their corresponding co-occurrence value. Some (x,y) pairs will never co-occur, you should have a 0 value for those cases.

**实现内容：**一个名为cooc\_dict2df（cooc\_dict）的函数，该函数将步骤3中生成的词典字典作为输入，并返回一个DataFrame，其中每一行对应一个中心词，每一列对应一个上下文词，单元格是它们对应的co-出现值。某些（x，y）对将永远不会同时出现，对于这些情况，您应该将其值设置为0。

**5.(5marks)**

Then, convert the co-occurrence values to PMI scores.

**What to implement:** A function cooc2pmi(df)that takes as input the DataFrame generated in step 4, and returns a new DataFrame with the same rows and columns, but with PMI scores instead of raw co-occurrence counts.

**5.（5分）**然后将同现值转换为PMI分数。

**实现内容：**函数cooc2pmi（df）以步骤4中生成的DataFrame作为输入，并返回具有相同行和列，但具有PMI得分而不是原始共现计数的新DataFrame

**6.(5marks)**Finally, implement a method to retrieve context words with highest PMI score for a given center word.

**6.（5分）**最后，实现一种方法来检索给定中心词的PMI得分最高的上下文词。

**What to implement:**

A function to pk(df, center\_word, N=10)that takes as input:

(1) the DataFrame generated in step 5,

(2) a center word(a string like ‘towels’),

and (3) an optional named argument called N with default value of 10; and **returns a list of N strings, in descending order of their PMI score with the center\_word.** You do not need to handle cases for which the word center\_word is not found in df.

**实现的目标：**函数topk（df，center\_word，N = 10）作为输入：

（1）在步骤5，

（2）一个中心词（如“毛巾”之类的字符串），

以及（3）一个名为N的可选命名实参，默认值为10；并以NMI字符串的PMI分数从高到低的顺序返回Nstring的列表。您无需处理在df中找不到单词center\_word的情况。

**Question 3b –Ethical considerations (10marks)**

Local authorities in touristic hotspots like Amsterdam, NYC or Barcelona regulate the price of recreational apartments for rent to, among others, ensure that fair rent prices are kept for year-long residents. Consider your price recommender for hosts in Question 2c. Imagine that Airbnb recommends a new host to put the price of the flat at a price which is above the official regulations established by the local government. Upon inspection, you realize that the inflated price you have been recommended comes from many apartments in the area only being offered during an annual event which brings many tourists, and which causes prices to rise.

In this context, critically reflect on the compliance of this recommender system with one of the five actions outlined in the UK’s Data Ethics Framework. You should prioritize the action that, in your opinion, is the weakest. Then, justify your choice by critically analyzing the three key principles outlined in the Framework, namely transparency, accountability and fairness. Finally, you should propose and critically justify a solution that would improve the recommender system in at least one of these principles. You are strongly encouraged to follow a scholarly approach, e.g., with peer-reviewed references as support.

Your report should be between 500 and 750 words long. Write your answer in Part3.ipynb (under Q3b).

**问题3b –道德考量（10分）**

阿姆斯特丹，纽约市或巴塞罗那等旅游热点的地方当局对休闲公寓的租金进行了规定，以确保为全年居民保持公平的租金价格。考虑问题2c中主机的价格推荐器。想象一下，爱彼迎（Airbnb）建议新的房东以高于当地政府制定的官方规定的价格出售该公寓。通过检查，您会发现推荐的价格虚高来自该地区的许多公寓，这些公寓仅在每年一次的活动期间提供，这会带来许多游客，并导致价格上涨。推荐系统，其中包含英国数据伦理框架概述的五项行动之一。您应该优先考虑最弱的操作。然后，通过批判性地分析框架中概述的三个关键原则（透明性，问责制和公平性）来证明您的选择合理。最后，您应该提出并批判性地提出一种解决方案，以改善至少一个以上原则中的推荐系统。强烈建议您采用学术方法，例如，以同行评审的参考作为支持。您的报告应在500到750个字之间。将您的答案写在Part3.ipynb中（在Q3b下）。

Learning Outcomes Assessed

-Carry out data analysis and statistical testing using code

-Critically analyse and discuss methods of data collection, management and storage

-Extract textual and numeric data from a range of sources, including online

-Reflect upon the legal, ethical and social issues relating to data science and its applications

-使用代码进行数据分析和统计测试

-严格分析和讨论数据收集，管理和存储的方法

-从包括在线在内的多种来源中提取文本和数字数据

-反思与数据科学有关的法律，道德和社会问题及其应用

**Criteria for assessment**

Credit will be awarded against the following criteria. Different criteria are applied to Pandas code (using Pandas outside of a function), function code, and figures obtained with matplotlib or seaborn. Pandas code (e.g. in Question 1c) is exclusively judged by its functionality. Functions are judged by their functionality and additionally their quality will be assessed. Figures are judged by their quality and completeness. The ethics question is judged by its quality. The below table explains the criteria.

评估标准将根据以下标准授予信用。

不同的标准适用于Pandas code（在函数外部使用Pandas），函数代码以及使用matplotlib或seaborn获得的图形。Pandas code（例如问题1c）仅由其功能来判断。通过功能的功能来判断功能，此外还将评估其质量。图片是根据其质量和完整性来判断的。道德问题由其质量来判断。下表说明了标准。

表格

描述已自动生成