**Data** **Analytics** **and** **Mathematical** **Statistics**

**Group** **Project**

Since PM2.5 is a component of PM10, some studies (notably the WHO ambient outdoor air pollution for cities database) have applied a factor to relate the annual mean PM10, based on the annual mean PM2.5, or vice versa, where data for one or the other is not available. Carry out a linear regression modelling study to investigate if an increase in overall particulate matter (PM10) causes fine particulate matter (≤ 2.5 m) to increase for either (choose one country), which would justify the application of such a factor:

a. Cities in Germany

b. Cities in France

(When using the database as a data source, be wary that some of the PM2.5 values have already been converted from PM10, and vice versa, and these should be excluded from your study. Why should this be so?)

You may refer to the WHO air pollution for cities database (filename: “WHO\_DB\_FR\_GER - clean.xlsx”, as uploaded on Canvas, for your data source.

翻译：由于PM2.5是PM10的一个组成部分，一些研究（尤其是世界卫生组织的城市室外空气污染数据库）在缺少其中一种数据时，用一个系数来将年均PM10和年均PM2.5相互转换。进行一个线性回归建模研究，探究总悬浮颗粒物（PM10）的增加是否会导致细颗粒物（≤ 2.5 m）的增加，从而为使用这样一个系数提供依据。选择以下两个国家中的一个进行研究： a. 德国的城市 b. 法国的城市 （在使用数据库作为数据来源时，要注意有些PM2.5的值已经从PM10转换过来，反之亦然，这些值应该从你的研究中排除。为什么要这样做？）

你可以参考世界卫生组织的城市空气污染数据库（文件名：“WHO\_DB\_FR\_GER - clean.xlsx”，已上传到Canvas）作为你的数据来源。

Group Presentation:

Work in your assigned groups. Each presentation should be around 20 minutes. You should include the following in your presentation:

1. Introduce the problem – background of the topic and why it is important to study this problem. You may include summaries of any research your group has done on the topic.

2. Application of the quantitative reasoning framework to the study of the problem:

o Frame a question to describe your problem

o Formulate a hypothesis

o Operationalise any concepts in your hypothesis that are not quantifiable, into measurable variables

o Deduce a prediction involving these variables

3. Is the data that you have for the study reliable and representative? You may discuss this question by commenting on the source of data, period of sampling, geographical spread etc.

4. Clean the dataset by removing those rows highlighted in grey, which indicates converted data (i.e. PM10 converted from PM2.5 using a formula, or vice versa). Why should these values be removed?

5. Carry out initial exploratory analysis of the data using your choice of descriptive statistics and visualisation methods. What are the insights obtained? Are the data normally distributed? Any trends between variables? Any outliers? Should these outliers be removed, or included? Provide justification for your answers.

1. Does your hypothesis appear to be supported from this preliminary analysis? If so, how could your findings be useful? Recommend also a follow-up study, and articulate your motivations for choosing this study.

7. Let’s say your initial exploration reveals that your hypothesis is not supported. What do you think could be some possible reasons? For example, could it be that assumptions you made for the study are not met, and if so, what are these assumptions? Or perhaps there are problems with your data, and if so, what are they? Propose some solutions.

Five group members will give the presentation. Each presentation will be followed by a 5-minute question-and-answer session. Remaining group members who did not partake in the presentation will have to contribute in answering the questions. The session will also be open to audience members.

Please upload a copy of your presentation slides to the submission folder on Canvas just before coming to class for Day 5. Just one group member per team needs to do the upload, on behalf of the other members.

The presentation will be in class, but we will be using Zoom for you to share your slides on the screen. The zoom details are:

Link:<https://nus-sg.zoom.us/j/82785187209?pwd=QzB3Syt1Q2dJSngyRVJQMENrdVdtUT09>

ID: 827 8518 7209

Passcode: 532210

The group presentation will be graded based on the following:

• Content of presentation

• Delivery of presentation

• Good time management and ability to answer questions

小组展示：

按照分配的小组进行工作。每个展示应该持续大约20分钟。你的展示应该包括以下内容：

1. 介绍问题——问题的背景和研究这个问题的重要性。你可以包括你们小组对这个话题做过的任何研究的总结。

2. 应用定量推理框架来研究问题：

o 提出一个描述你的问题的问题

o 形成一个假设

o 将你的假设中不可量化的概念，转化为可测量的变量

o 推导出涉及这些变量的预测

3. 你们用于研究的数据是否可靠和具有代表性？你可以通过评论数据的来源，抽样期间，地理分布等来讨论这个问题。

4. 清理数据集，去除那些用灰色标记的行，这些行表示转换过的数据（即用公式从PM10转换成PM2.5，或反之亦然）。为什么要去除这些值？

5. 使用你选择的描述性统计和可视化方法对数据进行初步的探索性分析。得到了什么见解？数据是否服从正态分布？变量之间有什么趋势？有没有异常值？这些异常值应该去除还是保留？给出你的答案的理由。

6. 从这个初步分析中，你的假设是否得到了支持？如果是，你的发现有什么用处？还可以推荐一个后续研究，并说明你选择这个研究的动机。

7. 假设你的初步探索显示你的假设没有得到支持。你认为可能有什么原因？例如，可能是你对研究做了一些不成立的假设，如果是这样，这些假设是什么？或者可能是你的数据有问题，如果是这样，问题在哪里？提出一些解决方案。

五名小组成员将进行展示。每个展示后将有一个5分钟的问答环节。没有参与展示的其他小组成员必须参与回答问题。环节也对观众开放。

请在来上第五天课之前，将你们的展示幻灯片上传到Canvas上的提交文件夹中。每个团队只需要一个小组成员代表其他成员上传即可。

展示将在课堂上进行，但我们将使用Zoom让你在屏幕上分享你的幻灯片。Zoom的细节如下：

链接：https://nus-sg.zoom.us/j/82785187209?pwd=QzB3Syt1Q2dJSngyRVJQMENrdVdtUT09

ID: 827 8518 7209

密码：532210

小组展示将根据以下标准进行评分：

• 展示内容

• 展示方式

• 良好的时间管理和回答问题能力