MA409: Statistical Data Analysis (SAS)

2024 Spring Semester Final Project (Apr 25th – June 7th)

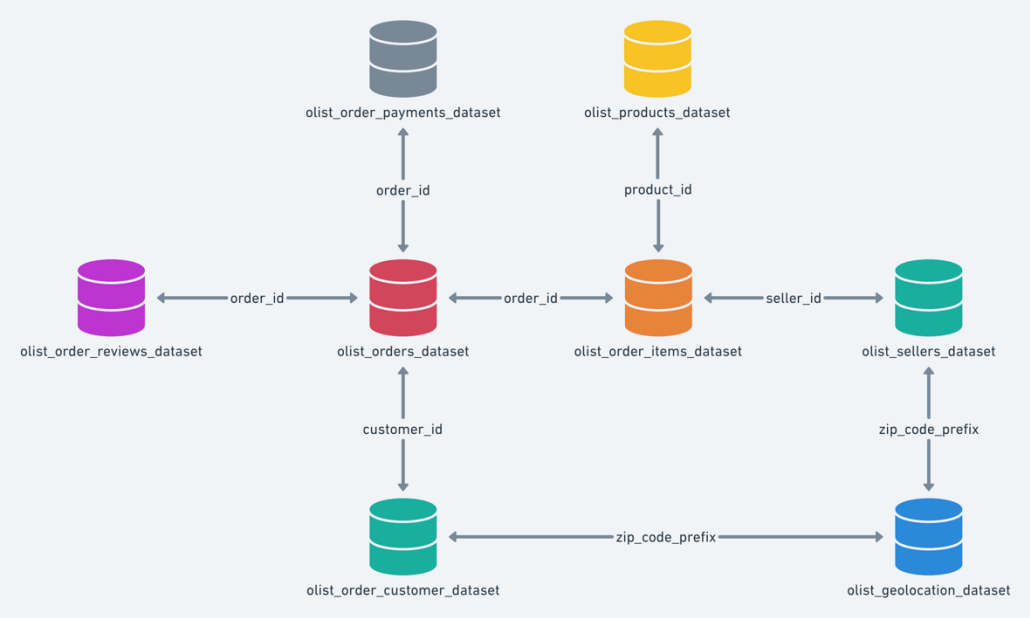
Project Description:

This is a real data analysis project to be accomplished in groups. There are 113 students in total, it would be great to have 29-30 teams of size 4 in total (a few can have size 3). The project schedule is:

Task	Deadline	Remark
Finalize team member	**May 5th, 11:59 pm**	Team leader report to TA (Xiaoling Wu) by QQ
Submit project materials	**May 30th, 11:59 pm**	Submit on Blackboard, team submission.
Submit reviews of other teams' project and team members' performance	**June 7th, 11:59pm**	Submit on Blackboard, individual submission.

In this project, all teams will analyze one of two sets of data, one provided by *Olist* (the largest department store in Brazil), and the other provided by *Airbnb* (a global online marketplace for short- and long-term homestays and experiences, which is no longer available in mainland China). Each set of data consists of several datasets, each containing different information.

The 8 datasets provided by *Olist* include the information of 100k orders (an order might have multiple items) from 2016 to 2018 made in Brazil, enabling viewing an order from multiple perspectives: order status, price, customer location, product attributes, geolocation, and reviews by customers. The data has been anonymized to protect privacy and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses. Some of the text in the data is in Portuguese, hope this won't discourage you from choosing it for this project. The data is divided into multiple datasets for better understanding and organization:



The datasets provided by *Airbnb* include listing information in Sydney, Australia compiled on March 16th, 2024:

Dataset	Description
listings-Sydney.csv.gz	Detailed listings data in Sydney (compressed file)
reviews-Sydney.csv.gz	Detailed reviews on the listings in Sydney (compressed file)
calendar-Sydney.csv.gz	Detailed availability information in the future 365 days on the listings in Sydney (compressed file)
DataDictionary.xlsx	Variable descriptions in the three files above

I do not define the specific problems that should be answered by this project, and you are encouraged to perform comprehensive data analysis on any problem you would like to explore. It is not required to use every piece of information provided. For the data analysis, applying some of the methods learned in this course (using SAS) is required, and other methods or software not covered are also encouraged. Make presentation slides, record a presentation video, and write a project report to present your data analysis and results.

Guidelines on Submissions:

1. Form teams of size 4 (a few teams may be of size 3) and report your team members to our TA (Xiaoling Wu) through QQ by May 5th at 11:59 pm. If you are unable to find a team, please let the TA know as soon as possible. A team number will be assigned to each team.
2. Prepare your presentation slides, presentation video, project report, and SAS code (as well as the code of other software if used) in a zip file; name it "TeamXX_final.zip" (XX is your team number). Submit the zip file on Blackboard. The deadline is May 30th at 11:59 pm. We will be watching the presentation videos of randomly selected teams in the last two lectures.
3. After submission, each team's submitted materials will be uploaded to Blackboard so that all teams can review other teams' projects. Submit your scores and comments on other teams' projects as well as your team members' performance on Blackboard by filling out a form; the deadline is June 7th at 11:59 pm.

For the **presentation video**:

- In MP4 format, ≤150MB (suggest using Tencent Meeting recording).
- Around 15 minutes in total. Each team member has to present.
- Both English and Chinese are accepted.

For the **project report**:

- Include abstract, introduction, exploratory data analysis, statistical modeling, conclusion, discussion, reference, etc. This is a list of recommendations and please exercise your judgment on what should be included in the report.
- Control the report within 15 pages.
- Both English and Chinese are accepted.

Details about the Assessment:

- **Instructor and TA's review based on the presentation (30 points)**
  - Background and problem setup (5 points)
  - Exploratory data analysis and statistical modeling (15 points)
  - Conclusion and discussion (5 points)
  - Presentation skill (5 points)
- **Instructor and TA's review based on the project report (30 points)**
  - Background and problem setup (5 points)
  - Exploratory data analysis and statistical modeling (15 points)
  - Conclusion and discussion (5 points)
  - Writing skill (5 points)
- **Other teams' review based on the project presentation and report (20 points)**
  - n teams in total, each team ranks the others by assigning scores $1, 2, \ldots, n-1$, no tie (higher score indicates better performance).
  - Teams will be stratified to 3 levels (high, middle, low) based on the aggregated score.
  - The high teams will get 20 points, the middle teams will get 15 points, the low teams will get 10 points.
- **Intra-team review (5 Points)**
  - Each team member will give a score ranging from 0 to 10 to other team members (higher score indicates better performance).
- **Review of other teams' projects and team members' performance (10 points)**
  - Finish all other teams' and team members' scoring (5 points)
  - Provide concrete comments on other teams' projects and team members' performance (5 points)
  - You may get bonus points if your comments are very instructive.
- **Attendance of the last two lectures (5 Points)**

Note: Each student will know the aggregated score from other teams' reviews and intra-team reviews. I would like to keep the raw scores confidential.