Lab 1 - Data retrieval and analysis with SQL

- Due Nov 7, 2024 by 11:59pm
- Points 100
- Submitting a file upload

Updates and Clarifications:

I have released a revised version of the part 1 Excel workbook (v1.1) that includes the following clarifications and corrections:

- Question 2 - answer this question considering all -four- cities from question 1, not three cities (typo corrected).
- Question 5 - write a query to find all listings for which a single reviewer has written exactly 14 reviews. That is, the same reviewer has reviewed the same property exactly -fourteen- times. (again, correcting a typo)
- Question 8 - clarified that there is no obviously correct definition for what cities constitute "The New York City Metropolitan Area" and "The San Francisco Bay Area". You need to assess the data available, read the data dictionary, and do some outside research (c.f. Google Maps) to decide which cities should be classified in each category. Likewise, for what variations of spelling, words, etc., constitute the skills requested. Use your judgment after doing some research. This is the challenging part of Q7 and Q8 (and the size of the tables), not the queries themselves, which should be pretty straightforward once you've made those decisions.
- **MDM M2-24 Lab 1 Part 1 - Workbook (v1.1).xlsx (https://canvas.cmu.edu/courses/42925/files/11842829?wrap=1) (https://canvas.cmu.edu/courses/42925/files/11842829/download?download_frd=1)**

Lab Overview:

In this lab, you will develop and improve your ability to use SQL to perform basic queries and analyses on relatively simple datasets. This lab is a team effort that will require a substantial amount of work to complete. You will work in the four-person teams assigned by the MSBA program to complete it. Plan how you will complete the lab and budget your time well - waiting to start the lab until a few hours before it's due will likely end poorly.

You will work with the four datasets below. Data dictionaries and access instructions can be found for each dataset by following the appropriate link: **(https://canvas.cmu.edu/courses/42925/pages/g84bceaf8e0fa96bb1ad613fc5ccf3913)**

- **LinkedIn Jobs (https://canvas.cmu.edu/courses/42925/pages/linkedin-jobs)**
- **BikeShare (https://canvas.cmu.edu/courses/42925/pages/bikeshare)**
- **Airbnb (https://canvas.cmu.edu/courses/42925/pages/airbnb)**
- **US Demographics (https://canvas.cmu.edu/courses/42925/pages/us-demographics)**

There are three parts to the lab, each with a specific deliverable. Read the instructions for each section carefully, especially the directions regarding the format and expectations for deliverables:

- Part 1 - Use SQL to complete specific, directed data retrieval and analysis tasks
- Part 2 - Complete three data analysis tasks and present your insights and conclusions from the analyses in a clear, concise, and compelling written document
- Part 3 - Each team member needs to write a brief (~1/2 of a page) reflective essay on the assignment

Part 1 - Use SQL to complete specific, directed data retrieval and analysis tasks (40 points - 5 points per question)

For the first part of the lab, you need to write SQL to complete the data retrieval and analysis tasks contained in this MS Excel Workbook:

> **MDM M2-24 Lab 1 Part 1 - Workbook (v1.0).xlsx (https://canvas.cmu.edu/courses/42925/files/11821045?wrap=1) (https://canvas.cmu.edu/courses/42925/files/11821045/download?download_frd=1)**

Each worksheet in the Excel document poses a question or lays out an analysis to perform by writing one or more SQL queries against one of the assigned databases and evaluating the results. Follow the directions in the workbook to answer each

question. Most questions require you to write a SQL query (or sequence of queries) to answer the question posed. You should then paste your SQL query(ies) and the result set returned into the appropriate boxes on the worksheet. Some queries require additional analysis and/or formatting of the results. Read the instructions for each question carefully.

(%24CANVAS_COURSE_REFERENCE%24/file_ref/g8ca734397be314e9e1389e103fe8d404/download?wrap=1)

Deliverable: Upload one completed workbook to Canvas per team. This single submitted workbook (.xlsx format) should contain all of the queries you have written, result sets obtained, and, where requested, analysis statements, visualizations and/or presentations of your analysis results.

Grading criteria: The Excel workbook describes each question's grading criteria, rubric, and points. See the "Grading Criteria" worksheet tab for details and expectations that apply to all questions. Some questions include additional grading details specific to the individual question listed on that question's worksheet.

Part 2 - Complete three data analysis tasks and present your insights and conclusions from the analyses in a clear, concise, and compelling written document. (48 points - 16 points per analysis)

For the second part of the lab, download this MS Word template file and complete the analysis tasks described within it.

MDM M2-24 Lab 1 Part 2 - Analysis Questions.docx (https://canvas.cmu.edu/courses/42925/files/11817741?wrap=1) (https://canvas.cmu.edu/courses/42925/files/11817741/download?download_frd=1)

As described in the template, you will need to retrieve and analyze data in the assigned datasets to generate a compelling answer to the question posed. These questions are much more open-ended than the questions in Part 1. This assignment is less about writing individual SQL queries and more about figuring out how to structure your analysis, where the data that you need to use for that analysis lives in the datasets, how to complete the analyses in a statistically appropriate way, and how to present those results in a concise, clear, and compelling way.

The first page of the template file provides details on what is expected. Read the instructions and the grading rubric attached to this Canvas assignment carefully to ensure you approach the assignment appropriately.

Deliverable: Upload a single PDF file containing your three analyses. The one-page limit for presenting your analysis results and conclusions is strict. When grading, we will stop reading after the first page of your response for each question and grade your submission based only on what you have presented in a single page (though we may look at the appendix to see how you derived your answers).

Grading criteria: We will use the criteria in this Canvas assignment's rubric to grade part 2 of the lab.

Part 3 - Each team member needs to write a brief (about 1/2 of a page) reflective essay on the assignment (12 points, 3 points per team member)

After completing the assignment, each team member needs to personally write a brief reflective set of observations about the assignment. Specifically, in up to one-half of a single-spaced written page, you should address the following questions:

- Your name (so I know which observation is made by which team member) (two to three words :-)
- How your team chose to divide the work and/or work together to complete the lab (1-2 sentences)
- A brief summary of your individual contributions to completing the lab (1-2 sentences)
- Lessons that you learned from the assignment that you can use to complete future analytics projects. (The rest of your allotted 1/2 page)
  - What you choose to write about for this last point is up to you. You can include observations about how your team chose to divide the work (and whether that proved to be a good choice), which tools you used, for what purposes, and whether they worked well or poorly, how you went about refining and clarifying the questions you had to answer, what you found surprisingly difficult (or easy), how you went about discovering and understanding the structure and contents of the

assigned datasets, etc. Try to capture a small number of thoughtful observations in clear, straightforward language. In general, simple and clear is better than complicated here.

Part 3 must be completed individually and, unlike the rest of this lab, without the assistance of a GenerativeAI tool (such as ChatGPT). As discussed in class, this part of the assignment is designed to help each student collect and synthesize their thoughts and observations from their work on the lab. Learning Science tells us that writing reflective observations down in a clear and straightforward way is an excellent way to remember the lessons learned so that you are ready to apply them to future projects and assignments.

Grading criteria: We will use the grading criteria in the attached rubric to grade part 3. Please note that your grade for this part mainly depends on the depth and relevance of your observations rather than the brilliance of your writing or command of the English language. Further, it should be clear that you wrote the reflection yourself without the help of GenAI. The purpose here is to take some time to think deeply about how to do this type of work more effectively in the future. So take a few minutes at the end of the lab to gather your thoughts and write them down.

A note on deliverables:

Please upload three separate files per team as your lab submission - the Excel workbook for part 1, one PDF file containing your responses to the three analysis questions in part 2, and a single document that collects the individual reflective writings from each team member, captured in a single PDF document. That should be three documents in total. Canvas allows you to upload multiple files for an assignment. Packaging all your files in a .zip file makes it much more difficult for us to grade your submission and provide feedback. Failure to follow these instructions will result in the loss of some points for the assignment.

MDM Lab 1 Rubric (M2-24)

Criteria	Ratings				Pts
Part 1, Question 1	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 2	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 3	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 4	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 5	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 6	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Question 7	**5 to >4.0 pts** **Exactly correct**	**4 to >2.0 pts** **Almost correct**	**2 to >0.0 pts** **Materially incorrect**	**0 pts**	5 pts

Criteria	Ratings				Pts
	The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	
Part 1, Question 8	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset.	**4 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set but the values calculated and retrieved are materially correct overall.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect.	**0 pts** **Not attempted, or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 2 - Airbnb analysis - problem framing and metric specification	**4 to >3.0 pts** **Sophisticated and nuanced** The problem statement has been refined and focused on a clear and precisely defined question that is amenable to quantitative analysis. The metrics selected are clearly and precisely defined, demonstrating a thorough understanding of the question, database structure, and data available. Selected metrics will provide a clear answer to the refined and restated question.	**3 to >1.0 pts** **Meets baseline expectations** The problem statement has been restated to something more amenable to quantitative analysis. Metrics are defined adequately, with sufficient clarity to guide and support the analysis. Selected metrics will help to answer the refined and restated problem statement.	**1 to >0 pts** **Not yet up to expectations** The problem statement has not been restated to something more clear and quantifiable, or the restatement does not match the intent of the original question. Metrics are poorly defined, unclear, or inappropriate for the question. Metric selection demonstrates a lack of understanding of the question or the data.		4 pts
Part 2 - Airbnb analysis - data retrieval, calculation, and analysis	**6 to >4.0 pts** **Sophisticated and nuanced** The analysis is thorough and accurate, employing appropriate techniques and calculations. Considers potential confounding factors or limitations of available data. Calculations are accurate and provided SQL runs properly, generating results presented. Demonstrates critical thinking and insightful interpretation of the data. Addresses potential limitations of the data available and metrics chosen.	**4 to >2.0 pts** **Meets baseline expectations** The analysis is generally accurate and addresses the question, but may have minor errors or omissions. Calculations are broadly correct, but may not demonstrate a subtle and nuanced understanding of the dataset or limitations of the data it contains. Provided SQL runs correctly and generates results presented. Analysis presents a straightforward analysis of the data and interpretation of metrics.	**2 to >0 pts** **Not yet up to expectations** The analysis is flawed, inaccurate, or incomplete. Calculations and/or SQL code are incorrect or do not produce the results presented. Demonstrates a significant misunderstanding of the data or analytical techniques. Fails to address limitations in data or metrics.		6 pts
Part 2 - Airbnb analysis - presentation, writing, and visualization(s)	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The single-sentence answer is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact of the findings. Excellent use of whitespace and formatting.	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The single-sentence answer is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.		6 pts
Part 2 - BikeShare analysis - problem framing and metric specification	**4 to >3.0 pts** **Sophisticated and nuanced** The problem statement has been refined and focused on a clear and precisely defined question that is amenable to quantitative analysis. The metrics selected are clearly and precisely defined, demonstrating a thorough understanding of	**3 to >1.0 pts** **Meets baseline expectations** The problem statement has been restated to something more amenable to quantitative analysis. Metrics are defined adequately, with sufficient clarity to guide and support the analysis. Selected metrics	**1 to >0 pts** **Not yet up to expectations** The problem statement has not been restated to something more clear and quantifiable, or the restatement does not match the intent of the original question. Metrics are poorly defined, unclear, or inappropriate for the question. Metric		4 pts

Criteria	Ratings			Pts
	the question, database structure, and data available. Selected metrics will provide a clear answer to the refined and restated question.	will help to answer the refined and restated problem statement.	selection demonstrates a lack of understanding of the question or the data.	
Part 2 - BikeShare analysis - data retrieval, calculation, and analysis	**6 to >4.0 pts** **Sophisticated and nuanced** The analysis is thorough and accurate, employing appropriate techniques and calculations. Considers potential confounding factors or limitations of available data. Calculations are accurate and provided SQL runs properly, generating results presented. Demonstrates critical thinking and insightful interpretation of the data. Addresses potential limitations of the data available and metrics chosen.	**4 to >2.0 pts** **Meets baseline expectations** The analysis is generally accurate and addresses the question, but may have minor errors or omissions. Calculations are broadly correct, but may not demonstrate a subtle and nuanced understanding of the dataset or limitations of the data it contains. Provided SQL runs correctly and generates results presented. Analysis presents a straightforward analysis of the data and interpretation of metrics.	**2 to >0 pts** **Not yet up to expectations** The analysis is flawed, inaccurate, or incomplete. Calculations and/or SQL code are incorrect or do not produce the results presented. Demonstrates a significant misunderstanding of the data or analytical techniques. Fails to address limitations in data or metrics.	6 pts
Part 2 - BikeShare analysis - presentation, writing, and visualization(s)	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The single-sentence answer is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact of the findings. Excellent use of whitespace and formatting.	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The single-sentence answer is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.	6 pts
Part 2 - US Demographics analysis - problem framing and metric specification	**4 to >3.0 pts** **Sophisticated and nuanced** The problem statement has been refined and focused on a clear and precisely defined question that is amenable to quantitative analysis. The metrics selected are clearly and precisely defined, demonstrating a thorough understanding of the question, database structure, and data available. Selected metrics will provide a clear answer to the refined and restated question.	**3 to >1.0 pts** **Meets baseline expectations** The problem statement has been restated to something more amenable to quantitative analysis. Metrics are defined adequately, with sufficient clarity to guide and support the analysis. Selected metrics will help to answer the refined and restated problem statement.	**1 to >0 pts** **Not yet up to expectations** The problem statement has not been restated to something more clear and quantifiable, or the restatement does not match the intent of the original question. Metrics are poorly defined, unclear, or inappropriate for the question. Metric selection demonstrates a lack of understanding of the question or the data.	4 pts
Part 2 - US Demographics analysis - data retrieval, calculation, and analysis	**6 to >4.0 pts** **Sophisticated and nuanced** The analysis is thorough and accurate, employing appropriate techniques and calculations. Considers potential confounding factors or limitations of available data. Calculations are accurate and provided SQL runs properly, generating results presented. Demonstrates critical thinking and insightful interpretation of the data. Addresses potential limitations of the data available and metrics chosen.	**4 to >2.0 pts** **Meets baseline expectations** The analysis is generally accurate and addresses the question, but may have minor errors or omissions. Calculations are broadly correct, but may not demonstrate a subtle and nuanced understanding of the dataset or limitations of the data it contains. Provided SQL runs correctly and generates results presented. Analysis presents a straightforward analysis of the data and interpretation of metrics.	**2 to >0 pts** **Not yet up to expectations** The analysis is flawed, inaccurate, or incomplete. Calculations and/or SQL code are incorrect or do not produce the results presented. Demonstrates a significant misunderstanding of the data or analytical techniques. Fails to address limitations in data or metrics.	6 pts
Part 2 - US Demographics analysis - presentation, writing, and	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The single-sentence answer is direct, insightful, and fully supported by the presented evidence.	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The single-sentence answer is missing, weak, or unsupported by the evidence.	6 pts

Criteria	Ratings			Pts
visualization(s)	Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact of the findings. Excellent use of whitespace and formatting.	Visualizations are included and generally interpretable.	Visualizations are missing, unclear, or ineffective.	
Part 3 - Reflective writing - Student #1	**3 pts** **Sophisticated and nuanced** Reflection demonstrates deep personal insight into team dynamics and process learnings, includes specific examples, and articulates clear, actionable lessons for future projects while addressing all required components with thoughtful detail.	**2 pts** **Meets baseline expectations** Addresses all required components adequately with basic description of team process and personal contributions, includes some general lessons learned but may lack specific examples or deeper insights.	**1 pts** **Not yet up to expectations** Missing required components, provides superficial observations without meaningful reflection on process or learning, or fails to articulate useful lessons for future projects.	3 pts
Part 3 - Reflective writing - Student #2	**3 pts** **Sophisticated and nuanced** Reflection demonstrates deep personal insight into team dynamics and process learnings, includes specific examples, and articulates clear, actionable lessons for future projects while addressing all required components with thoughtful detail.	**2 pts** **Meets baseline expectations** Addresses all required components adequately with basic description of team process and personal contributions, includes some general lessons learned but may lack specific examples or deeper insights.	**1 pts** **Not yet up to expectations** Missing required components, provides superficial observations without meaningful reflection on process or learning, or fails to articulate useful lessons for future projects.	3 pts
Part 3 - Reflective writing - Student #3	**3 pts** **Sophisticated and nuanced** Reflection demonstrates deep personal insight into team dynamics and process learnings, includes specific examples, and articulates clear, actionable lessons for future projects while addressing all required components with thoughtful detail.	**2 pts** **Meets baseline expectations** Addresses all required components adequately with basic description of team process and personal contributions, includes some general lessons learned but may lack specific examples or deeper insights.	**1 pts** **Not yet up to expectations** Missing required components, provides superficial observations without meaningful reflection on process or learning, or fails to articulate useful lessons for future projects.	3 pts
Part 3 - Reflective writing - Student #4	**3 pts** **Sophisticated and nuanced** Reflection demonstrates deep personal insight into team dynamics and process learnings, includes specific examples, and articulates clear, actionable lessons for future projects while addressing all required components with thoughtful detail.	**2 pts** **Meets baseline expectations** Addresses all required components adequately with basic description of team process and personal contributions, includes some general lessons learned but may lack specific examples or deeper insights.	**1 pts** **Not yet up to expectations** Missing required components, provides superficial observations without meaningful reflection on process or learning, or fails to articulate useful lessons for future projects.	3 pts
				Total Points: 100