How’s the Air Out There? Using a National Air Quality Database to Introduce First Year Students to the Fundamentals of Data Analysis

David Hall and Jessica D’eon (Corresponding Author)

19/03/2021

1 Abstract

2 Introduction

Whether we like it or not we’re living in an age of data, and the world of chemistry is no exception. From big-data atmospheric chemistry in climate-change models[1] to machine-learning organic synthesis[2], every domain of chemistry is increasingly relying upon data-driven science. Undergraduate chemistry curricula need to adapt to better equip and prepare the next-generation of chemists with the skills and knowledge needed in response to this trend. As we ourselves work on new undergraduate teaching material we notice that an oft-overlooked aspect of how data analysis is presently tough is how exactly data (measurements, signals, etc.) is transformed into information (trends, correlation) and finally into knowledge. The explicit teaching of these concepts is often neglected in current teaching labs, resulting in increasing student frustration. Motivated by this, and the need to transfer to a virtual laboratory environment as a result of Covid-19 social distancing restrictions, we sough to develop a new, remove learning compatible, experiment.

An obvious assumption of teaching data science is that students will be prepared to analysis *real* data, which is often permeated with outliers and the fingerprints of gross experimental errors. The data soon to be collected by undergraduate chemist is no different. However, acquiring sufficient data for analysis if often stressful for undergraduate students given time- and equipment-constraints in the teaching laboratory. As well, with the Covid-19 restrictions students were unable to attend labs, and hence produce their own data. We saw this as an opportunity to integrate actual measurements from published data repositories such as the air quality data from Environment and Climate Change Canada’s national Air Pollution Surveillance Programs (NAPS). There is no shortage of data to be analyzed as the NAPS program, since 1975, has been collecting hourly measurements across Canada for several major atmospheric pollutants.

Prominent atmospheric pollutants are structurally simple, and undergo reaction schemes comparable to those covered in introductory chemistry lectures. Ozone (O_3) and nitrogen dioxide (NO_2) are two choice candidates for analysis by undergraduate students. They are structurally simple molecules, and undergo reaction schemes comparable to those covered in introductory chemistry lectures (see reactions (**??**), (**??**), and (**??**)). Notable of these compounds is their interdependent diurnal cycles. The relationship between O_3 and NO_2 is so intimate, the term “odd-oxygen” (O_x) is used to express the sum of these two compounds (see reaction (**??**), and ref.),[3] although the relationship between O_3 and NO_2 can vary with environmental and anthropogenic influences including temperature, sunlight, and motor vehicle emissions.

$$NO_2 + photon \rightarrow NO + O(\#eq : O3 - 1)$$
$$O + O_2 \rightarrow O_3(\#eq : O3 - 2)$$
$$O_3 + NO \rightarrow O_2 + NO_2(\#eq : O3 - 3)$$
$$[O_3] + [NO_2] = [O_x](\#eq : odd - ox)$$

Our Air Quality lab explicitly introduces data analysis work-flows to undergraduate students while they investigate seasonal differences in the relationship between atmospheric O_3 and NO_2 from subsets of the NAPS dataset. Students are encouraged to qualitatively explore the difference in this relationship, while generating hypothesis through probing questions, and a new interactive application we developed that allows them to quickly analysis the entire NAPS dataset with a minimal increase in their workload. As the entire lab uses previously acquired data, and the ubiquitous Microsoft Excel software package, students were able to explore real data from home, and complying with Covid-19 restrictions.

3 Experimental Overview And Pedagogical Goals

This 3 hr data-analysis laboratory exercise uses publicly available data and open-source code (described in the [Supplementary information]), and has been run successfully in the one-semester “CHM135: Chemistry: Physical Principles” undergraduate general chemistry at the University of Toronto since Summer 2020. This course is most often conducted in the first-term of the first-year of life-sciences/chemistry students. As our *Air Quality Lab* being the first-lab of five, it is designed as much as an tutorial on data-analysis and Microsoft excel as it is to explore atmospheric chemistry. The lab is divided into three parts: the prelab, data analysis in Excel, and data exploration & hypothesis generation.

The prelab follows a traditional approach, and is written to situate students in the relevant chemistry for the upcoming analysis. Specifically for this lab we create explanatory videos and material introducing gas phase chemistry, and relating the lab content to concurrent lecture material of gas phase chemistry (i.e. ideal gas law).

In the data analysis portion of the lab students are randomly assigned two datasets. Each dataset is a 7-day snapshots of hourly O_3 and NO_2 measurements taken from the NAPS program. The datasets are all from the same NAPS surveillance station for a given year. For our purposes we chose a different downtown Toronto NAPS station for each successive iteration of the lab. The two datasets correspond to 7-days in the winter and 7-days in the summer, and were generated from original NAPS data using R as described below. Alongside their data, students are provided with a written handbook detailing the necessary Excel operations, and an synchronous online session with their TA. Working through the lab exercises students are explicitly taught data analysis workflows, modelled after that recommended by Hadley and Grolemund[4]:

1. *Importing* their assigned comma serrated values (.csv) data sets into Excel.
2. *Tidying* their data and setting up their worksheets. This step consists of formatting cells to properly display values and handling missing data. Specifically for this lab, the NAPS dataset stores missing values as ‘-999,’ which can be erroneously interpreted literally by Excel.
3. *Visualizing* their data by creating a time-series plot of time vs. concentration of each pollutant, see Figure 3.1.
4. *Transforming* their data using mathematical operators in Excel to calculate total oxidant and adding it to their time-series plot as well as calculating 8hr moving averages.
5. *Modelling* a linear relationship between $[O_3]$ and $[NO_2]$ to qualitatively assess the negative relationship between these two contaminants.
6. *Communicating* and exploring their results through a series of accompanying questions.

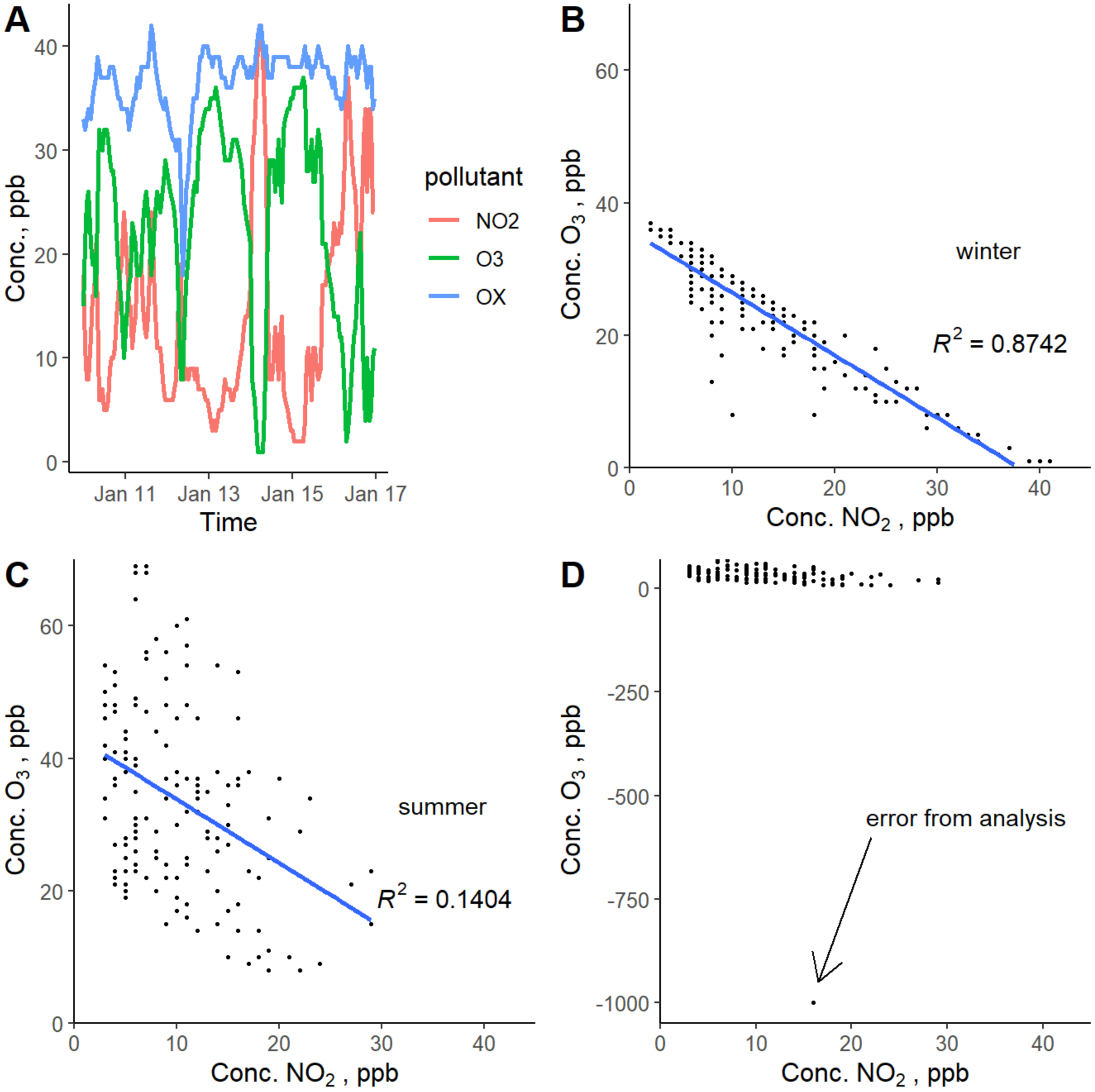

Figure 3.1: Example of plots students are expected to create. (A) time-series of pollutants across 7 winter days. (B) Correlation plot of O3 and NO2 concentrations with linear regression in the winter and (C) summer data sets. (D) Example plot if a ‘-999’ value wasn’t removed.

The last step in this workflow is expanded in the final part of the lab where students compare their results to the complete NAPS dataset from which their assigned datasets originated from. Here they are encouraged to generate hypotheses based on their own data, and their *a priori* chemical knowledge introduced in the prelab. To this end, we created an interactive online application that students visit using *R* and *Shiny*. This application consist of an interactive map showing the location, and local population, of every NAPS surveillance station. Students can then select any station and time-span, and a time-series and correlation plot, similar to the ones they created themselves, are automatically generated. This allows them to rapidly compare their data to any number of stations, simultaneously relieving them of the burden of repetitive and tedious data analysis and facilitating hypothesis generation and data exploration. Accompanying questions prompt students to explore and reason differences in O_3 and NO_2 correlation between urban and rural areas, as well as between winter and summer datasets. See the Supporting Information or https://davidrosshall.shinyapps.io/AirQualityApp/ (https://davidrosshall.shinyapps.io/AirQualityApp/) for details on the application.

3.1 Leveraging R to automate and expand the lab

We made prodigeous use of *R* and various associated frameworks to greatly facilitate many aspects of this lab. Greater details and source code can be found in the Supporting Information, but a brief discussion is warranted, if anythigne lse, to help encourage readers to harness *R* (or similar data science languages) to both simplify their own workload, while expanding course content.

Firstly, the generation of the 7-day datasets. The NAPS hourly information is recorded for each individual station in separate `.csv` files. These expansive datasets include the measurements from every NAPS stations (n > 150), and are organized in a matrix style. For first-year students, accessing, subsetting, and tidying these datasets is immediately intimidating and tedious as they often exceed tens of thousands of rows and contain a number of information not necessary for their work. To counteract this, we used *R* to combine O_3 and NO_2 measurements, remove bilingual headers, ancillary columns, etc., and transformed the data from the ‘wide’ matrix style to the ‘long’ columnar format so data is easier to manipulate in Excel. Then, using *R*, we generate a specific number of student datasets using a 7-day moving window of the year-long data. I.e. data set 1 is Jan. 1st to the 7th, data set 2 is Jan 2nd to 8th, etc, with complimentary summer data sets taken from July 1st onward. The rolling window ensure every student can be assigned a unique dataset, while largely looking at similar data. We also randomly insert a ‘-999’ missing value into each data set ensuring students will encounter it during their analysis. Each data set is then automatically saved as a .csv file.

Secondly, we wrote an *R markdown* script that generates a PDF with the analysis results of every generated data sets. The answer sheet analysis mirrors the one students carry out, providing TAs with an actual analysis of every dataset, allowing them to quickly check each students submission, and relieving them of the burden of verifying each dataset.

Thirdly, using we wrote an interactive application using *R* and *Shiny*. This was created in-house specifically for this laboratory exercise. Thanks to the this, we were able to expand students working data from that directly provided to them, to the entire NAPS dataset. This would be impossible otherwise as students would not have the ability or time to explore the larger NAPS data in any practical manner given the tools we can provide to them. As well, by allowing students to explore the entire dataset, it creates opportunities for them to explore data ‘unknown’ to the instructors. In other words, students are excited to make real connections and discoveries with real data, rather then tediously analysis an increasing number of provided data. See the [Supplementary Information] for more details on the app, and how you can recycle our code to create your own version of the application if you want to run this lab.

Lastly, our code can accept any standard formatted NAPS dataset, and instructors can readily select the NAPS station, the number of datasets, the overlap between datasets, etc. so course material can easily be updated for each iteration of the lab, or between different lab sections. See the [Supplementary Information] for the source-code for instructions on generating datasets.

4 Results and Pedagogical Outcomes

Based on the our surveys and feedback from students, the majority of students responded positively to the new learning experience. From our survey of students in the most recent interation of the Air Quality Lab in the winter 2021 term (n=28), 68% of respondents stated they felt the Excel component of Lab 1 significantly helped them complete the other CHM135 labs more efficiently. (Figure 4.1). This is a welcomed improvement as students frequently complain about the time commitment required for the CHM135 lab component. Students also feel more confident with regards to overall interpretation of data (plots, trendlines) as well as towards their use of Excel in general (57%, and 64%, respectively) (Figure 4.1).

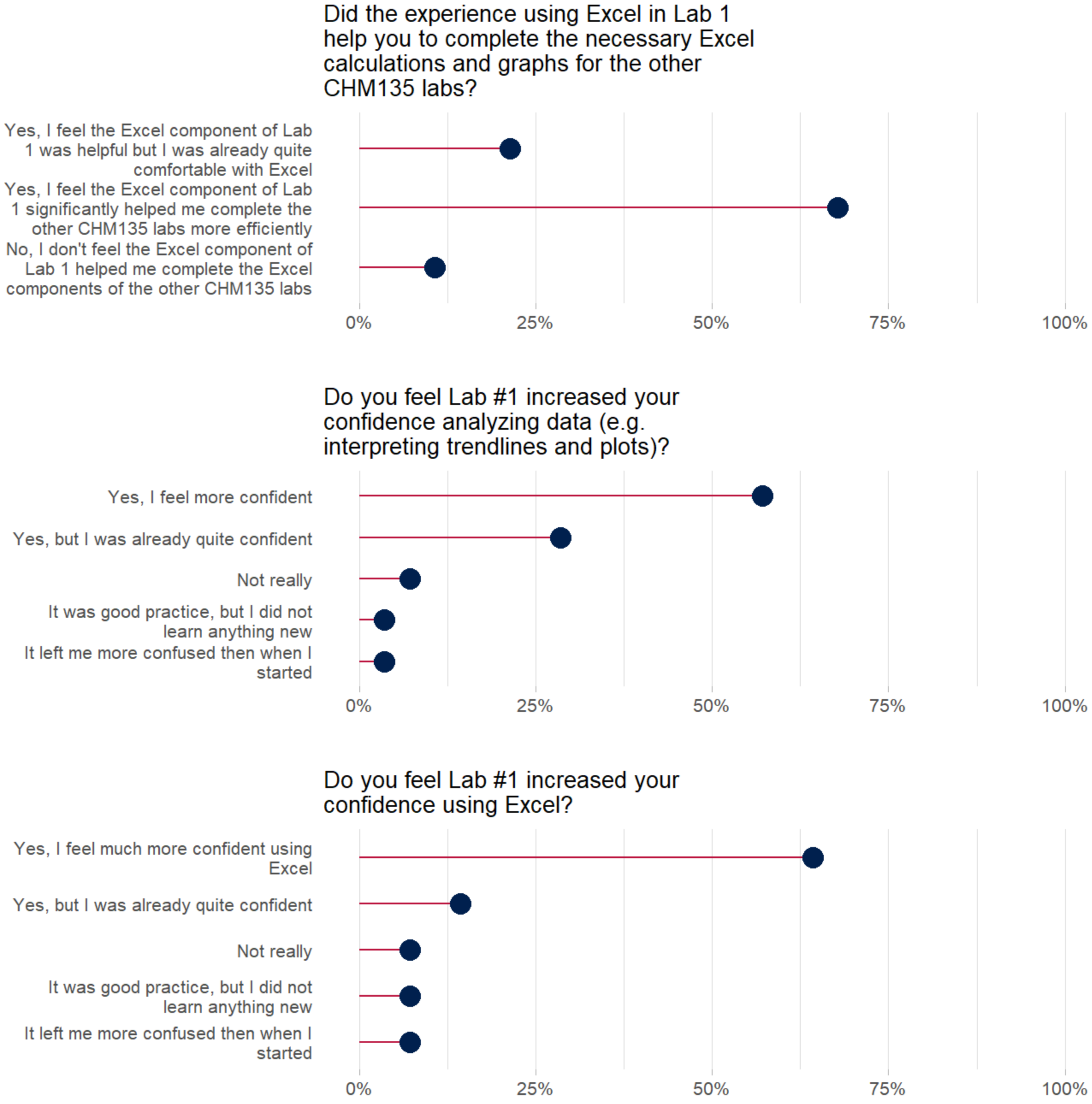

Figure 4.1: End of term student survey results for Lab 1.

Included in the survey was the option for students to provide any additional feedback on the lab. STudents expressed both positive and negative feedback to the Air Quality lab (See Supporting Information Table **??** for complete feedback). Students appreciated the introduction to Excel, the practical usefulness of the incorporated material, and the opportunity to analyze real world data offering a glimpse into environmental chemistry. However, students were also critical of how the material was implemented. Some experienced trouble with inconsistencies between the Excel instructions, and their version of Excel (although all UofT students are provided with free access to the latest version of Excel, with instructions provided in the aforementioned instructions document). Likewise, many students felt the Excel component should have been explicitly tough during the synchronous session, rather than these sessions focusing explicitly on lab material (i.e. data analysis vs. Excel operations). As there are no explicit prerequisites to knowing Excel for this course (or UofT overall), we opted not to directly teach students the basic workings of Excel in the synchronous session. Going forward, we feel that incorporating optional Excel help-sessions specifically to assist students with this component of the lab.

4.1 Notes on the Air Quality Shiny App

A major addition to this lab is the development of the Air Quality App using Shiny. This app is what allows students to readily explore the larger NAPS dataset without burdening them with lengthy data prep/analysis. While the idea of creating an app for a specific lab exercise may seem daunting, the flexibility and support of the Shiny framework greatly relieve one from the minutia of app development, allowing them to focus on how best to present their data to the target audience. The version of our app presented to students in the Winter 2021 term is available for viewing here: https://davidrosshall.shinyapps.io/AirQualityApp/ (https://davidrosshall.shinyapps.io/AirQualityApp/). We have also provided the complete source code and example datasets on Github: https://github.com/DavidRossHall/AirQualityApp (https://github.com/DavidRossHall/AirQualityApp), and as a `.zip` in the supplementary information. Alongside these are instructions modifying, and running, this app for your course.

We chose not to capitalize on the abilities of Google analytics (although that is an option when creating Shiny Apps) to track student usage of the app. Instead, we were inferred app usage by tracking the number of accessions (i.e. times the app was used). As shown in Figure 4.2, despite publishing the course material more than a week prior to the synchronous lab sessions, students did not access the app in any meaningful numbers. Usage did increase after the synchronous session, where students were explicitly instructed to work on their Lab 1 reports, of which two questions specifically instruct students to access the Air Quality App. Predictably, app usage was highest immediately preceding the due date for the Lab 1 report; the implications of this are discussed below. As best we can tell, usage/interaction with the Air Quality app lasted on average 25 minutes, a respectable time given the brevity of the prompting questions, and the richness of the dataset.

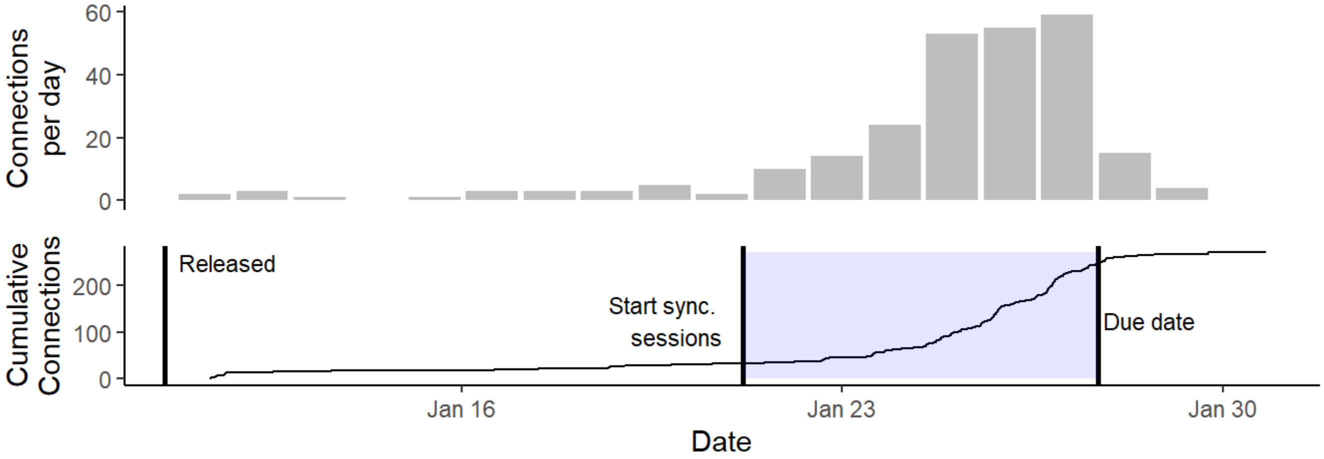

Figure 4.2: Connections to Air Quality app per day (top) and cummulative connections over time (bottom). Vertical lines indicate when the lab material was made available online to students, when the synchronous sessions with TAs started, and the due date. Blue rectangle highlights period when students worked on report sheet.

A brief comment for the unaware: running a Shiny app requires server side computing. In other words, as students select data to plot, a dedicated server must perform the necessary computations. While these requirements are not egregious, they are still to be considered if you plan on hosting your own instance of the app. While Shiny provides instructions and software for running the app on premise, we opted to host our app on the Shiny server cloud.As mentioned above, many students ‘flashed’ the app as the deadline appeared. Without adequate server space, this could increase load times, decrease responsiveness, and possibly crash the app. Shiny offers 15hrs/month of server time (time it takes to run the app), but this is unsuitable for the anticipated server loads once the App was released for all of the CHM135 lab sections. Consequently, for the 2021 Winter Term, there were slightly over ??? students, so we payed the *Standard* hosting package costing $99USD/month for 2000 hrs of server time, in addition to using the reliable Shiny servers. This was admittedly an excess of server time, but we chose to play it safe. Likewise, as the Air Quality App was only used during the first lab, we only needed to pay for one month of server time to account for all the lab sections. As we continue to experiment with creating purpose built apps for individual labs/courses, we will transition to local hosting.

5 Conclusions

We sought to create a new introductory lab experiment to expressly teach incoming students foundational components of data analysis/science as well as practical instructions on how to use Microsoft Excel for future courses. To this end, we leveraged the R computer language for the automated, and scalable, generation of unique data from real world atmospheric measurements from the NAPS that served as the foundation of our introductory data science lab. Alongside written instructions on how to use Excel, we developed an online interactive App, allowing students to readily explore the entire NAPS dataset to compliment their individual analyses. Our efforts were rewarded with students being better equipped to tackle subsequent data analysis challenges in the following lab, in addition to arming them with skills they will assuredly make use of outside of the first-year chemistry laboratory.

6 Supporting Information

Note this will be a seperate document, obviously.

Stuff to include:

- Images of app
-

Table 6.1: Student responses to: Any additional feedback on Lab 1 that you would like the teaching team to know? Both positive feedback and constructive criticism are welcome.

Answer
Fun lab
I feel that lab 1 was a good start for us, it helped us tackle the rest of the labs
I had a really difficult time completing every lab that had an excel component. The reason for this is that most of the instructions, seemingly detailed to students without online tasks and computer skills, was very lacking in detail to someone like myself. I was late doing almost every single lab, and penalized for it, when I should not have been, as the time to complete it along with all other courses was too much to ask. On average, I had to spend between 18-30 hours to finish every lab. My mark suffered as a result
I have been using excel for awhile with programming so i have been pretty comfy with it, but I think it could be really helpful to someone newer with excel

Answer
I really enjoyed it! It was interesting and really did help me with excel, but was also a cool glimpse into environmental chemistry. I feel like it was one of the few real world/career-type applications of class content we got in this class.
I received a good mark on the lab. Unfortunately, I don’t know where I lost the few points that I lost.
I think it would be better if there is an instruction video made by TA. When I took another course, the TA made a short video, which was about 3 minutes, that showed procedures of making a graph and slight tip for the interpretation. This actually made the students have less questions regarding the lab assignment.
I think that it’s not really fair that even if students had successfully uploaded photos and scans on their end and their TA’s were not able to access it, there would be marks off. On top of that, only some students were allowed to have their labs regraded because of that while others did not get the opportunity or were not allowed to. It is completely out of the students control about whether or not something had been uploaded properly after they have checked on their end and I believe that the marking team should be more fair in letting every student have the chance the redeem themselves since they did all the hard work on their end
It was a great practice, but doing all of those learning alone was very painful, and it took such a long time… I wish things were not as condensed.
It was interesting to learn about atmospheric chemistry and I liked that we got to use real data. I found it to be a lot more work than the other labs.
It would have been awesome if how to use excel was guided in practical session. Also, it would have been better if we did it as a group since we can help each other and learn from them.
Some aspects of the excel tipsheet were written as if the reader had prior knowledge of Excel, eg referring to ‘copying’ without specifying that in Excel this isn’t the same as the copy of copy/paste function. The tipsheet also did not account for differences in versions of Excel or for how it presents on different devices, e.g. I had to try 3 different makes/ages of computer to get the proper date format option. Other students may not have the option of switching devices. After the initial learning curve it was a very helpful resource and much appreciated!
The lab was very easy to understand and interesting
This is not my favourite lab, however I learned a lot from it. It was a lot of work and I had to did a bit of extra research how to interpret dataset. At the end I learned, which what matters.

Table 6.2: Student responses to: Is there anything that you would like to share about the CHM135 labs overall?

Answer
(1) I really enjoyed all of the labs. (2) I consistently underestimated the time required to complete the labs - but maybe I am simply slow. (3) I think it would be useful for the professor to spend five to ten minutes introducing each lab. What are the learning objectives?
all labs had very clear instructions except for lab#4, the instructions were not clear and were not easy to follow.
Having more hands on labs will make it easier to learn the material

Answer
I feel like we shouldn’t have to do this much excel alone prior to labs. I believe it would be highly beneficial to have 3 hours labs as normal, with mandatory attendance. Make lab sections that vary in time, aka early in the day, mid day, and evening, so people in all time zones can attend, but still have mandatory attendance. It doesn’t feel like I did one lab the entire semester. I just feel like a professional data entry specialist. Other than that, the actual classes were fine, and all of the professors seemed to care about concerns.
I found that they complemented the course material very well and actually helped prepare me for the tests.
I know this might be unavoidable because of online courses but at times I felt that the majority of material for the labs were more tedious than educational. For instance, most of the time I spent on some of the labs was for filling in Excel spreadsheets rather than practicing calculations and gaining a better understanding of the material. Again though, I know that this is probably difficult to reform because of CHM135 being online but I thought I would mention it nevertheless. Besides that I thought the labs were fine :)
I loved all of them except the first one. I especially enjoyed the last two. Thank you!
I thought they were really effective. I honestly feel like the excel and data analysis skills I developed were more useful to me as a science student than doing the actual wet lab work, like titrations. Obviously that stuff is important too, but it can be learned easily and isn’t applicable in many other classes beyond chem. Using excel, reading trendlines, and working through methodical steps to solve an end problem are applicable skills I can use in any class.
I was very intimidated by the idea of labs, but the straightforward nature of the assignments plus having an excellent lab TA made it one of my favourite parts of the course. (seriously grateful for the calibre of TAs, I can’t say this enough)
I would REALLY have appreciated a full set of the instructions in a PDF in the same way we were able to download a PDF of the report to work on. I spent so much time clicking around in the modules trying to find relevant info, and i really think that could have helped me feel less overwhelmed by the labs
It took soooooooo time to finish each labs. I hope that we can cut down some components… (try to keep important calculation parts, but maybe cut down the simulation records)
Maybe do 4 instead of 5
Personally, I love the labs and I’m grateful for the experience I had. Thank you so much
The lab was fun and I loved it!
The labs are well organized, I think that I clearly understood more of the content as it flows well with the lectures to help me understand the lab calculations and what is expected to be done. Thank you!

7 Author Information

David Hall, Department of Chemistry and School of the Environment, University of Toronto.

Jessica D’eon, Department of Chemistry and School of the Environment, University of Toronto.

8 Acknowledgements

References

(1) *The Future of Atmospheric Chemistry Research: Remembering Yesterday, Understanding Today, Anticipating Tomorrow*; National Academies of Sciences, Engineering, and Medicine (U.S.), Ed.; The National Academies Press: Washington, DC, 2016.

(2) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nature Reviews Chemistry* **2019**, *3* (10), 589–604. https://doi.org/10.1038/s41570-019-0124-0 (https://doi.org/10.1038/s41570-019-0124-0).

(3) Kley, D.; Geiss, H.; Mohnen, V. A. Tropospheric Ozone at Elevated Sites and Precursor Emissions in the United States and Europe. *Atmospheric Environment* **1994**, *28* (1), 149–158. https://doi.org/10.1016/1352-2310(94)90030-2 (https://doi.org/10.1016/1352-2310(94)90030-2).

(4) Wickham, H.; Grolemund, G. *R for Data Science*; Beaugureau, M., Loukides, M., Eds.; O’Reilly Media Inc.: Sebastopol, CA, 2017.