Question 1.

To approach this problem, I would first understand how my data is formatted, what is my unit of analysis, i.e. what does each row represent in my data?

- First, reshape the data to long format (if necessary) so that each row is a unique instance of a survey respondent and their annual income.
- Second, with the income variable, sort in ascending order so the lowest income is the first observation, and the highest income is the last observation.
- Create a counting feature that goes from 1 to n, where n is total number of rows.
- Create a quintile feature that is set to 5.
- Multiply 0.2*n = p. Replace all rows <= p as quintile 1 (lowest quintile)
- Multiply 0.4*n = w. Replace all rows > p & <= w as quintile 2
- Multiply 0.6*n = x. Replace all rows > w & <= x as quintile 3
- Multiply .8*n = y. Replace all rows > x & <= y as quintile 4
- The quintile feature should have 20% of observations set at 1, 2, 3, 4, and the remaining still set as 5.

Question 2.

The merging of these two data could be considered big data for several reasons. First, there is a variety of data being used. One is a table of 150,000 units and the other is millions of rows. We could be merging data of 150,000 transactions between millions of unique firms. Then, there is also a high volume of data, so much so, the dataset would not be able to fit on a normal computer, but instead the tables were compressed to save space and are pulled in from a remote server. For instance, these 150,000 transactions once merged into the millions of unique firms will create 150,000*2,000,000 more cells, with many being blank (as two firms never traded).

Question 3.

Problem 1: Without context, I know nothing about the baseline day as to compare all other day measures to. While I know relative differences between days, it does not do much to tell me a story of if what Ontario's public health system did is impressive or a failure.

Problem 2: The value of dosages and days listed on and next to the needle are confusing to read and seem misleading. First, the days are not evenly spaced which confuses the reader, as it does not give an adequate picture of how dosages have ramped up in intensity over the past few weeks. Moreover, days are not ordered in time as 9 days come after 8 days. Lastly, the size in dosages is increasing in absolute terms, yet the size on each needle portion is constant in size which paints a picture that dosages are not increasing.

Suggestion: Instead of a needle that counts the total dosages, they could instead make a time series plot, where the X axis is the date, and Y axis is number of dosages. They can add dashed lines at every one million new doses. This would show how dosage administration has grown more so in the last few days.

Question 4.

In this case it's hard to determine if the algorithm is truly fair, as there are multiple terms of fairness, with each being mutually exclusive from each other. The company is arguing that it took out features that are telling of protected status, but then other features might proxy for these measures. For instance, Amazon's job hiring algorithm did not account for sex of applicant to be fair, but the algorithm proxied sex by considering language used on the resume, resulting in less women being hired. Regardless, this "fairness" might then be resulting in the same proportion of protected/unprotected applicants being offered a position, and this is fair in the sense that they are hiring protected people in the same size as unprotected people.

Critics could be focusing on a different metric of fairness, such that protected applicants are more likely to have their applications thrown out, even when equally as effective unprotected applicants who were hired. Therefore, critics are focused on systematic inaccuracies in positives and the firm is focusing on same chance of positive prediction.

In sum, if I were to take a normative approach, I would argue the critics are right, especially after what I learned with Compas and Amazon's algorithms. I think ignoring protected status will only let other features proxy for these variables and punish marginalized groups more often. Taking a positive approach, I would argue that it is not possible to say who is right, as fairness can take on many forms with different metrics for success.

Question 5.

D

Question 6a.

The pseudo is do kfold cross validation, by taking a subset of k observations over n occurrences, making multiple folds. It runs a loop to partition different subsets of data into testing and training, runs the model, then calculates the RSME between the held back testing data and predictions from each model to test accuracy.

Question 6b.

Root Mean Squared Error

Question 6c.

Your K value – how many folds you want to estimate and calculate RMSE for.

Your n value – how many observations your model will need to calculate a predicted outcome for.

Question 7.

Question 4 confused me, just because in lecture I feel like we never took a hard stance on who was right or not right with Compas and I do not know if I should be arguing for one side or another for full credit or instead talk about why both sides could both be right as defined by different definitions of fairness. This question is a necessary one to keep but I think rewriting it as “Take a stance and argue in favor or against the company in this situation, what measure of fairness did the company use? Why do you think this fairness is the best metric in this context? If you are against the firm, which fairness should we focus on and how can we improve this issue?” I felt this way about question 2 as well, I could argue both sides of the coin, but with three sentences I kind of had to throw a hail mary and hope I was right.

Question 8.

Group 6

Person	Did not do their fair share	Did do their fair share	Did more than fair share
Eli Mogel			X
Kelli Maples			X
April Kuang		X	
Yuki Nakagawa		X	