Feedback on Intermediate Report for Course Project

MATH 448-648 — Spring 2023

General comments

Joshua: I appreciate that you used RMarkdown to create your report. Combined with your mastery of ggplot, the report looks pretty good visually. Also, the research questions you lay out make a lot of sense and the report contains a good variety of statistical analyses. All in all, I see that you have spent a good amount of time and effort preparing this report. Unfortunately, most of these analyses do not make sense. This is in part because you misinterpreted what your two datasets relate to (more below) and in part because you conducted analyses using meaningless variables (row labels) as if they were meaningful. Also, in one case (chi square test), you applied a tool designed for categorical variables to numerical variables (speed limit). So while in most of your report, the statistical analyses are formally correct (e.g., in the hypothesis test and in the regression), the results and their interpretation do not make sense. I give detailed explanations about these mistakes in the next section. In the next paragraph, I explain an important shortcoming in your understanding of the data, which sadly invalidates the conclusions of most analyses in the report.

The labels "January 2023" and "February 2023" of the data files do not refer to the track restrictions in these months. They refer to a reporting date. That is, the first file contains all track restrictions reported up until January 2023 and similarly, the second file contains all restrictions reported up until February 2023. You can in fact check that the data reported in January 2023 are a subset of the data reported in February 2023. Order the data by "Date_Restriction_Reported" and you will see. So in your report, whenever comparing two datasets, you are not comparing the two separate months of January 2023 and February 2023. You are actually comparing all data between December 2020 and January 2023 to all data between December 2020 and February 2023 ☹ Because on this, most of the conclusions and interpretations of your analyses are incorrect. You need to fix that in the final report. One way to do that is to extract the January 2023 and February 2023 data separately and comparing them. But this is very restrictive and the number of data points you will get might be very small. Alternatively, you can repeat your existing analyses on different, non-overlapping data subsets. For example, compare the year 2021 to year 2022.

All in all, you will need to do a major overhaul of the current analyses and report. Please let me know if any of my comments require more unpacking. Also, I encourage you to run your analyses by me before submitting your final report. I will be happy to help.

Specific comments

1. Page 2: Figure 1.1. Nice figure… except that the y-axis scale is completely wrong. For example, in the top left figure, you should not set “scale_y_continuous” to go from 0 to 250. You can check that the actual total restriction distance (in miles) is 0.228 for Green Line B, 0.026 for Green Line C, etc. The largest total restriction distance is 3.826 for the Red Line. You should change your y-scale in all plots of Figure 1.1, making sure that you represent the true scale of the variable displayed.

2. Page 2, Figure 1.1. After doing the above, you will see that the values for the total restriction distance are NOT proportional to the total track lengths. This invalidates the comments you make below the figure, so you should remove them.

3. Page 2, Figure 1.1. I am not convinced that the plots on the right are all that useful. For each branch, you are simply multiplying the branch total length by the number of restrictions in a given month. Is this informative? Wouldn’t it make more sense to report the branches’ lengths and the number of restrictions by branch separately? So maybe report the branches’ total lengths in the text or in a table, and display the counts (monthly numbers of restrictions by branch) in a table or a bar plot.

4. Page 2, bottom. The comment on the normality (or lack thereof) of the graphs makes no sense because the graphs are not histograms. You can’t infer anything about the distribution of the restriction distances for each branch because you only show a single number: the sum of all restriction distances. Alternatively, if your idea was to model the distribution of the total restriction distances and view the MBTA branches are forming a random sample of size 8, well, that’s not wise. These branches are fixed objects, not a random sample, and viewing the measurements on these different branches as coming from the *same* probability distribution is not reasonable.

5. Page 3, Figure 1.2. This is a slightly convoluted way of representing the proportions of different categories: reason for restriction, by branch. Given that the branch name already includes the Line information, it’s maybe not necessary to add Line as a new factor to the plot. Keep it simple when you can! For example, you could drop the Line factor from the graph and keep the jittered scatter plots. Or you could represent this information as a stacked bar plot: stack the counts or percentages of different restriction causes for each branch.

6. Page 3, section 2. Good to do a z-test but the summary table lacks key statistics: sample means, sample standard deviations or variances, z score. You should include them to demonstrate you understand how the z test is performed. In fact, based on the p-value you report (0.6283), I think that you may have conducted a t-test. The z-test yields an almost identical p-value of 0.6287. (In R you would have to calculate it “by hand”. See the examples under Course Materials/Chapter 10 on Blackboard.)

7. Pages 4-5, Figure 2.1. These figures don't make a lot of sense because the report ID is arbitrary, it does not have any order. So the data points could be rearranged in any order and would still represent the same information but show a very different graph. You could more usefully represent the distribution of speed restrictions for each time period with a box plot, for example, or a bar plot (height = average speed limitation) with an error bar (=standard deviation / sqrt(sample size)).

8. Page 6, Figure 2.2. The same comment as for Figure 2.1 applies: this figure does not make much sense. In addition, the legend should say "Histograms detailing the restriction distances […]", not "Histograms detailing the restriction speeds […]".

9. Pages 6-7, Section 3. Alas, the same comment as before applies again. Track ID is just a label. The labels 1, 2, 3… might just as well be called A, B, C, … so it does not make sense to regress the speed restriction distance against this label. That is, the results of this section are meaningless. On the other hand, you could investigate if/how the speed restriction depends on the restriction distance and maybe also the line or the branch. This would make a nice regression analysis with both numerical and categorical predictors. I will not comment on your interpretation of the regression on page 7 because it's based on a meaningless regression model. However you can reproduce this type of comments in the new regression suggested above.

10. Pages 7-8, Section 4. To perform a chi square test in this section, you take all speed restriction data of both datasets, put them in one big vector (length 238 = 70 + 168), and concatenate it with a vector of integers 1, 2, 3, …, 238. Then you run a chi square test on this. Sadly, this makes no sense whatsoever. I am not sure what you had in mind with this approach. If you want to conduct a chi square test, you need one or two categorical variables. For example, you could ask whether the cause for restriction is independent of the branch or not. This is a way to revisit your Figure 1.2 and the associated analysis. Or you could calculate the proportions of restriction by branch for a given time period, say 2020 to mid-2021, take these proportions as the baseline, do the same thing for the period mid-2021 to early 2023, and ask the question: have the proportions changed or are they the same?

Additional suggestions for final report

Only if time permits (you have a lot to fix in the current report): as a follow-up to the regression analysis proposed in my point 9, you could do a residual analysis, see if the residuals have an approximate normal distribution, and if not try to fit a known probability distribution to the residuals using the method of moments and/or maximum likelihood estimation (see Chapter 9).