MBTA Speed Restriction Analysis

Joshua Soriano

2023-05-16

Introduction

The MBTA (Massachusetts Bay Transportation Authority) is a transit-oriented organization whose goal is the development and operation of Massachusetts' transit-related systems. It is both the oldest and largest transit system in Massachusetts. Its influence is so large, that it reaches all the way back to the 1800s. Today, the MBTA continues to hold strong, serving almost 200 cities and over a million daily riders.

Data

I found the data for my project on the MBTA data portal, where each data set was readily available to download as a CSV file. The data is available as public domain at https://mbta-massdot.opendata.arcgis.com. The data contains information relating to the speed restrictions on the MBTA railway for all of January through April, in the year 2023.

In my project I will be performing an analysis on these four data sets, all of which contain information regarding speed restriction date of implementation, length in miles, days active, restriction reason, restriction speed, restriction distance, as well as the line, branch, and specific location the speed restriction is assigned to. In particular I will be closely analyzing these variables, and create and analyze the models based on this data to find correlations. There is some data missing in the restriction days to clear and restriction days active columns, which would indicate the value is not applicable for that row in particular.

Research Questions

1. What is the total restriction distance for each line? What is the primary restriction reason for each line? How many speed restrictions were put in place due to track wear and tear?
2. Has the average restriction distance increased between January and April? What about restriction speed? (Hypothesis testing)
3. Can we construct a reliable linear regression model that displays the correlation between restriction speed and restriction distance for the data sets? (Modeling)
4. Is there a difference in the mean restriction speed applied to each specific location? (Chi-square test)

1 Speed Restrictions on the MBTA

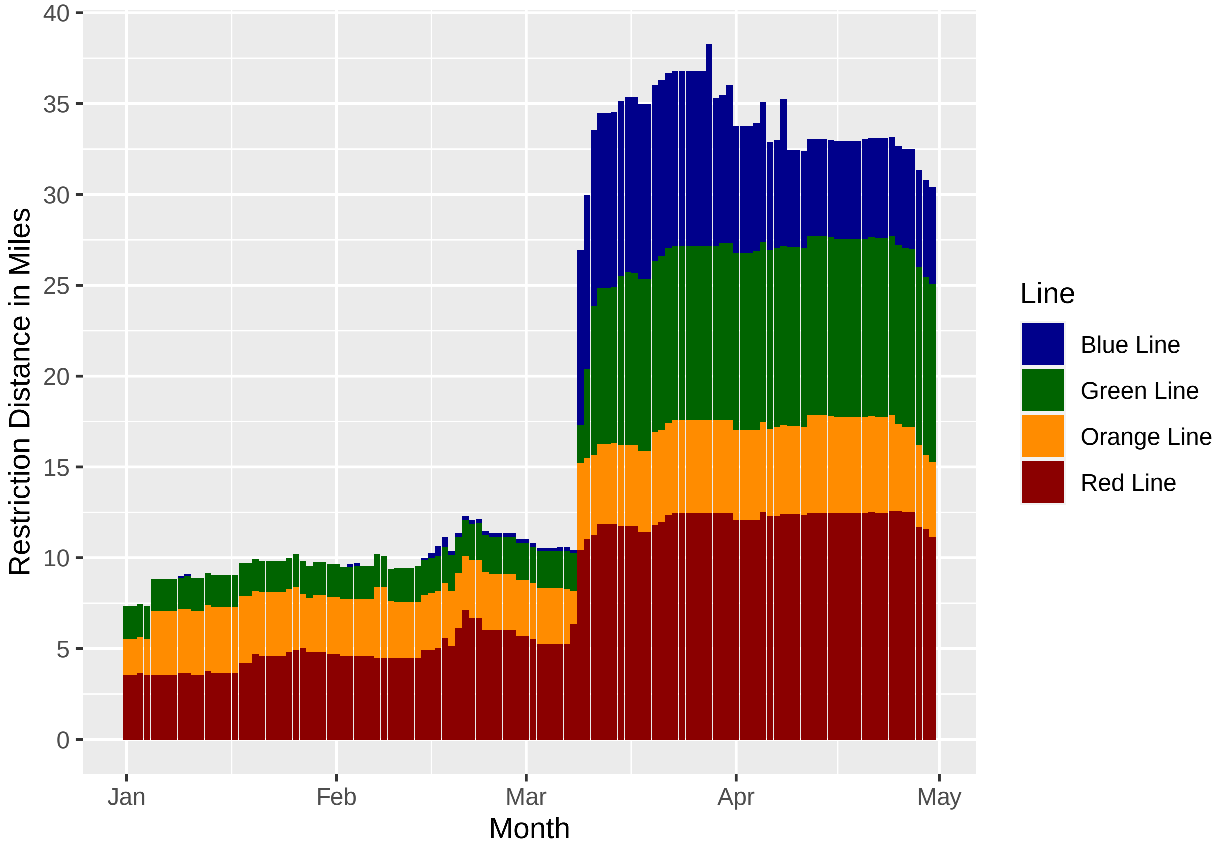

Figure 1.1: Daily restriction distance through months January to the end of April 2023.

The total track length reported each day can be seen in Figure 1.1. It is worth noting that the Red Line Braintree branch easily has the most restriction distance, with the Mattapan line having the least restriction distance. We can observe that the Blue Line and Green Line are relatively free of speed restrictions, up until March and April, where they suddenly suffer an influx of speed restrictions (the Blue Line in particular).

When taking into consideration the sudden influx, we cannot say that the data has a normal distribution, due to the massive negative skew. Although it could be argued that there exists a normal distribution with a higher alpha value, even this is debatable due to the drastic surge in restriction distance, resulting in a curve with little to no slope.

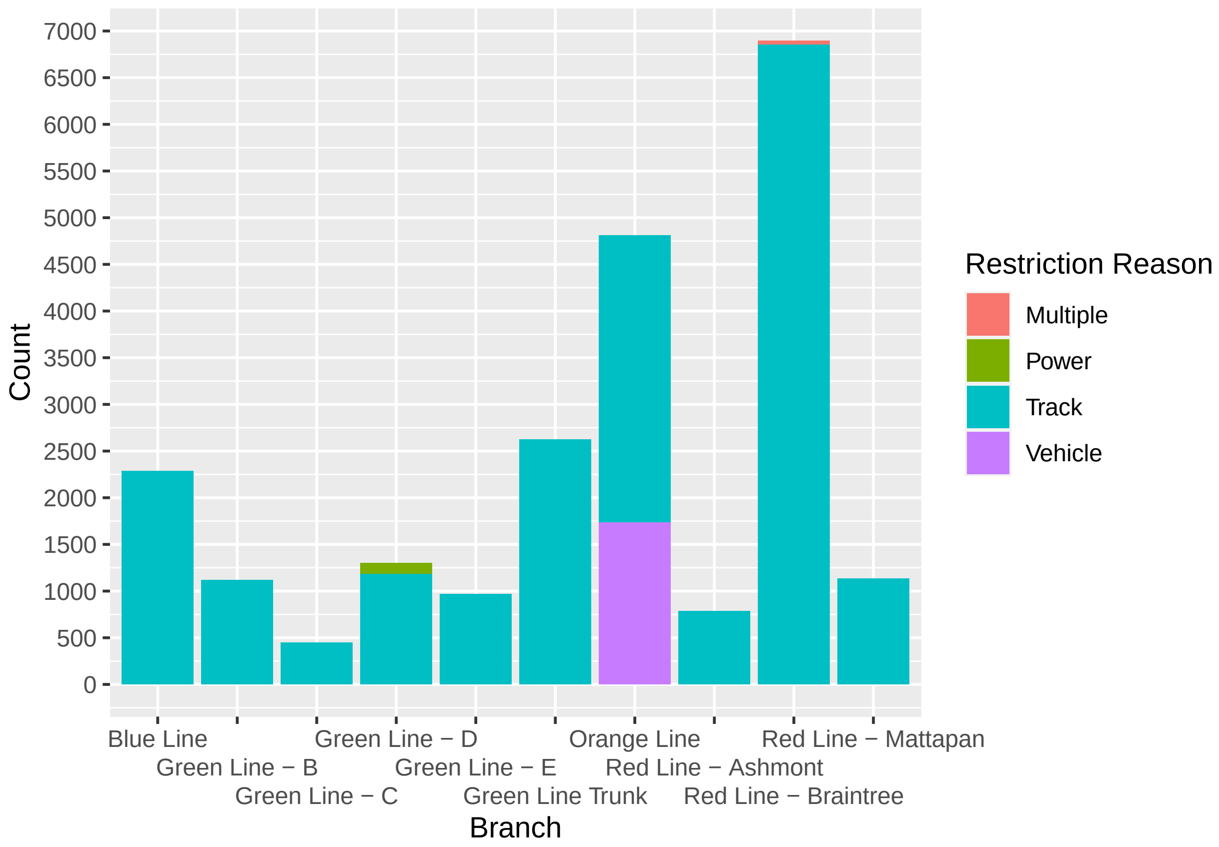

Figure 1.2: Restriction reason per branch for the past four months.

The graph for Figure 1.2 indicates track wear and tear being the primary reason for the speed restriction being applied, with there being only a few instances of vehicle blockage. Other restriction reasons have a minimal appearance rate, with the only exception being vehicle blockage, which, strangely enough, only ever occurred on the Orange Line, which indicates traffic as a prevalent issue there.

Power issues have occurred minimally, to the point that it is somewhat concerning that there exists this many track issues compared to other restriction reasons. We can only assume this is a direct result of the MBTA's somewhat dated infrastructure, as some of the MBTA lines are decades old.

2 Difference Between Mean Restriction Speed and Distance for January and April

Restriction speeds are put in place in order to ensure rider safety, and as discussed in the section above, they are applied mainly due to track wear and tear. Depending on how the situation regarding track maintenance is handled, transit authorities may feel the need to alter said restrictions as they see fit, in order to adapt to any unforeseen circumstances. For further analysis, we will take a look at the mean restriction speeds for January through April, and see if there exists evidence proving they changed significantly.

First consider the following hypothesis test regarding the mean restriction speeds. The data was extracted using the summary function. Since the samples are large enough, we can justify a Z test.

$$=$$

$$=$$

$$=$$

Hypothesis Testing: Comparing Mean Restriction Speeds	
Sample Size (January)	2801
Sample Size (April)	9211
Sample Mean (January)	14.6316
Sample Mean (April)	14.7012
Population Variance (January)	43.6506
Population Variance (April)	54.4697
z-score	-0.4752
Mean Restriction Speed (January)	μ_1
Mean Restriction Speed (April)	μ_2
Null Hypothesis	$H_0 : \mu_1 = \mu_2$
Alternate Hypothesis	$H_a : \mu_1 < \mu_2$

For the above values, we have a p-value of 0.3173. Since this is greater than the α value of 0.05, we fail to reject the null hypothesis, and conclude that the mean restriction speed for January is not less than the one for April. This means the average restriction speed has remained consistent, and that transit authorities felt the need to maintain the restriction speed for further track maintenance.

Given that we previously discussed the primary reason for speed restrictions, it is clear that the majority of these speed restrictions are due to track wear and tear, an issue which seems to have become more prominent within the past few months. Since transit authorities determined it was necessary to maintain the speed restriction, the prospect of track repairs is somewhat dubious at the moment.

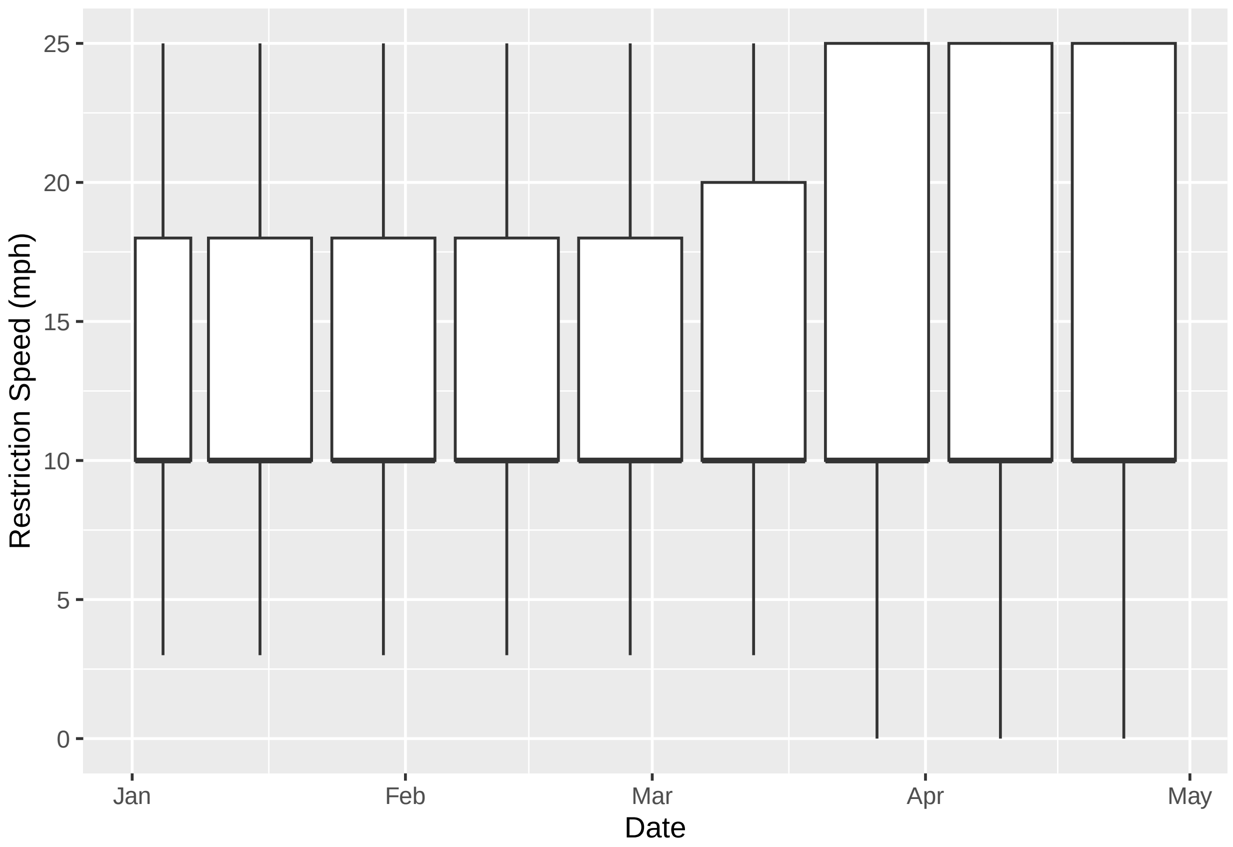

Figure 2.1: Box plot detailing the restriction speeds for each month, with an interval of two weeks.

Observe that for the figure above, the shape of each box is very consistent. This is expected, since the restriction speed is assigned only at whole integers, and given that the restriction speed has an incredibly consistent mean of ten miles per hour, it is likely that the transit authorities are required to keep the restriction speed consistent for the sake of passenger safety.

The maximum and the minimum also stay very consistent in this regard, though we do see an uptick in the third quarter, indicating that there are some higher restriction speeds assigned for the months of March and April. This is in line with our hypothesis testing, since aside from some outliers, the restriction speed overall has not changed.

Next, consider the following hypothesis test regarding mean restriction distance.

Hypothesis Testing: Comparing Mean Restriction Distances	
Sample Size (January)	2801
Sample Size (April)	9211
Sample Mean (January)	0.1011
Sample Mean (April)	0.1073
Population Variance (January)	0.0085
Population Variance (April)	0.0155
z-score	-2.8419
Mean Restriction Distance (January)	μ_1
Mean Restriction Distance (April)	μ_2
Null Hypothesis	$H_0 : \mu_1 = \mu_2$
Alternate Hypothesis	$H_a : \mu_1 < \mu_2$

For the data above, we have a p-value of 0.0022, which is less than $\alpha = 0.05$. Thus we reject the null hypothesis, and must therefore conclude the mean distance in which the speed restriction is applicable has increased within the past four months.

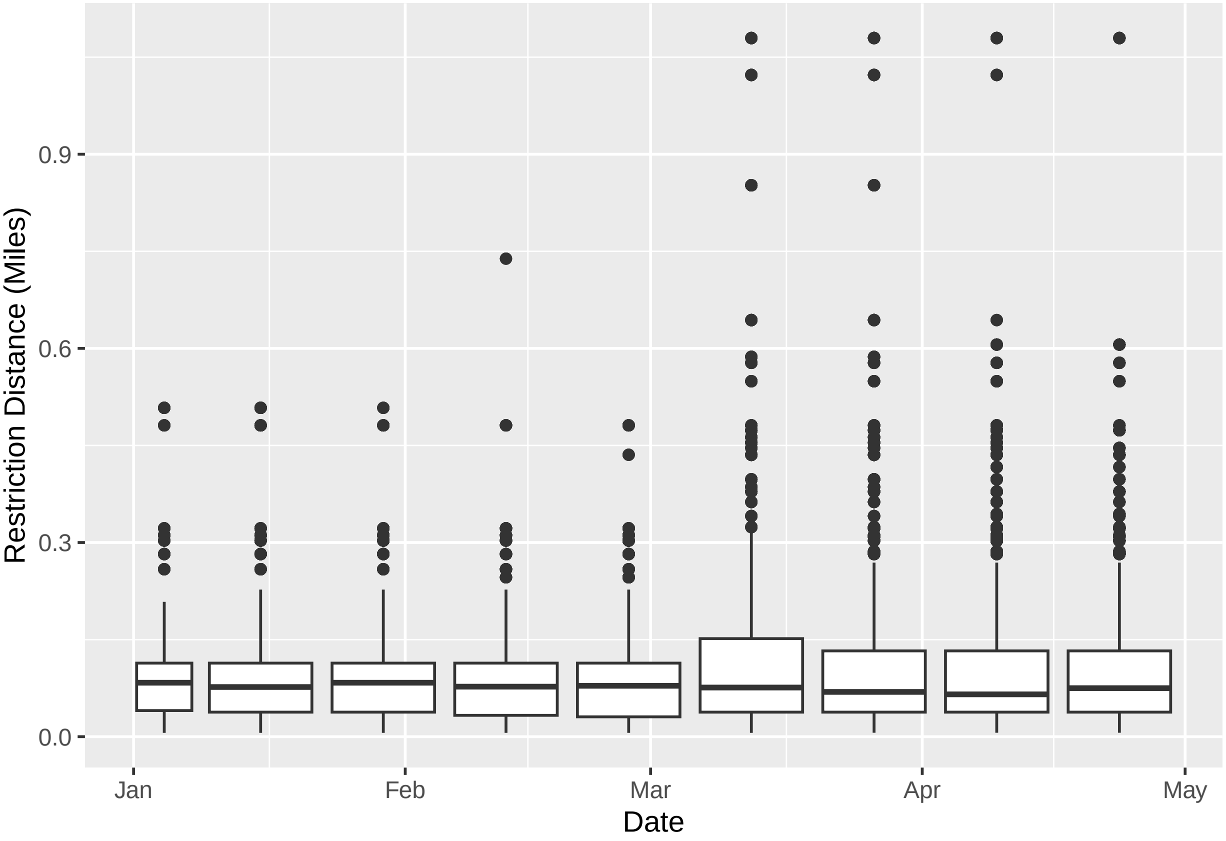

Figure 2.2: Box plot detailing the restriction distance by two-week intervals, for the past four months.

Observe that like our previous box plot, the box plot here is also not as visually indicative of a change in the value. In fact, despite restriction distance having more loose values, and not restricted to whole integers as the restriction speed is, the restriction distance also has a largely consistent median. This is despite the numerous outliers on the graph, which may be the cause of some skew. It is also worth noting that both quarters also remain relatively consistent, although we do see some uptick in the third quarter which does indeed fall in line with the hypothesis testing of an increasing mean restriction distance.

The lack of a visually observable increase in restriction distance is somewhat expected, since for our previous hypothesis test, we also suffered from a clear, visual indication of an increasing mean. The similar results also stems from the fact that we do expect the variables of restriction distance and restriction speed to correlate, since both are aspects derived directly from the speed restriction.

3 Linear Regression Model For Restriction Distance and Speed

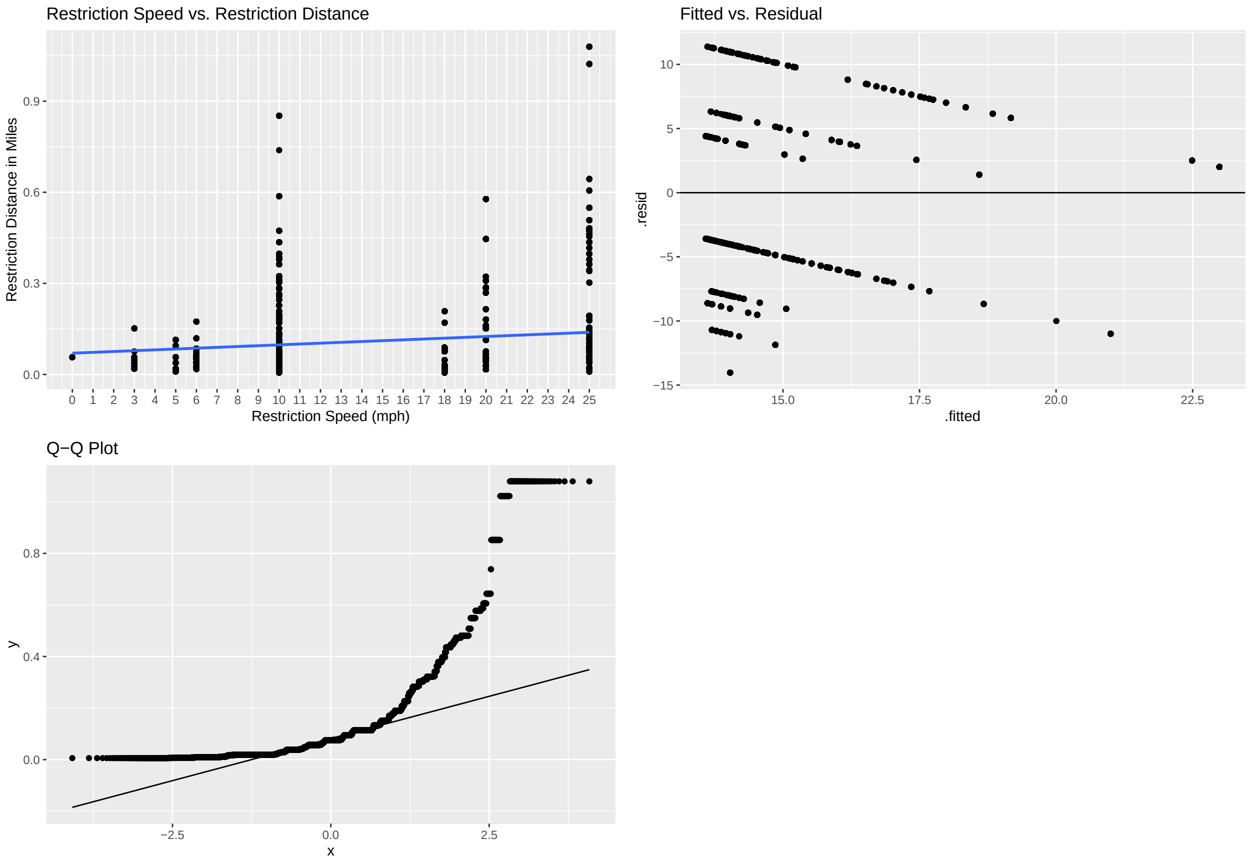

Figure 3.1: Linear regression model of restriction distances for January through April.

The linear regression model was constructed by using the restriction speed as the x-value, and restriction distance as the y-value. As observed in the linear regression model, we have a rather problematic distribution, given the restriction speed, despite having a scale of 25 different speeds, is only ever assigned to seven of those restriction speeds.

Though the predictors are not required to come from a normal distribution, indication of a normal distribution stems from the points being scattered on the fitted vs. residual plot. As observed above, this is not the case for the fitted vs. residual plot, as the points are organized in bizarre, diagonal patterns. We can attribute this unorthodox shape to the rigidness of our original data. Thus, we cannot say the data is not normally distributed.

A lack of normal distribution is also visually indicated by the Q-Q plot, which in the case of a normal distribution, would have the majority of its points tightly knit together on a diagonal line. Instead, the points on the Q-Q plot above form a curved line, not indicative of a normal distribution.

```
##
## Call:
## lm(formula = Restriction_Distance_Miles ~ Restriction_Speed_MPH,
##     data = AllSR)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12900 -0.07305 -0.02955  0.01611  0.94089
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.0701002  0.0018852   37.18   <2e-16 ***
## Restriction_Speed_MPH 0.0027423  0.0001169   23.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1237 on 22367 degrees of freedom
## Multiple R-squared:  0.02402,    Adjusted R-squared:  0.02398
## F-statistic: 550.6 on 1 and 22367 DF,  p-value: < 2.2e-16
```

Figure 3.1: Statistical data for restriction distance.

In practice, we want a high R^2 value, so that the regression model may fit our observations. Unfortunately, as observed from the values above, it is made clear that the R^2 value is in actuality incredibly low, and thus the regression model created above is not suited for our observations. However, this does not mean the regression model has no use. At the very least, it is precisely because the R^2 value is low that the regression model indicates much of the variation in the restriction distance cannot be explained.

Recall that in the previous section we discussed how the majority of speed restrictions were put in place due to track wear and tear. Given that much of the variation in the linear regression model is unexplainable, we can assume that track wear and tear is a mechanical problem which is often difficult to make predictions on. We can also bring human error into the equation, which can be a significant source for variation given its unpredictability. Other observations can be derived from the low R^2 score, but for the time being it would be best to retain the notion that the regression model has a large amount of unexplainable variation, and should not be relied upon for its precision.

4 Potential Correlations Between Location and Restriction Speed

The data set includes speed restriction data on four lines. While this may not seem like many, it is worth noting that each line is extensive, and splits off into multiple branches. Each branch then contains a section of track, in which a speed restriction is reported, the location of which is detailed in a separate column called Location Description.

It is worth keeping in mind that the column used to refer to the exact location in which the speed restriction takes place on is referred to in the data set as "Location Description." These values split off from the Branch column, which splits off of the Line column. For the sake of simplicity, the values in the Location Description will be referred to as individual locations.

With that in mind, we will examine the restriction speed for each location, and see if the exact location of the speed restriction has any effect on the aforementioned speed limit.

The correlation will be determined with a chi-square test.

Chi-square testing independence of restriction speed and branch	
Degrees of Freedom	1638
Test Statistic (X^2)	96612
p-value	2.2e-16

Since the p-value is very close to zero, we can conclude that the restriction speed is not independent of the location it was assigned to. This result is expected, since while restriction distance and restriction speed are separate values, the area in which the speed restriction takes place on would undeniably have some influence to how fast or slow the trains are allowed to travel in that specific area.

Location determines the terrain and environment the track is on, which not only affects how the track may be worn down over time, but also how accessible the track may be in the event the track section requires maintenance. There is even the issue of vehicle blockage, which, as discussed above, seems to be a location-related issue exclusive to the Orange Line. Other contributing factors are connecting tracks, intersecting tracks, track types which would require a specific form of maintenance which may affect the duration and degree of the speed restriction applied.

Given that track maintenance requires direct human intervention, we can recollect from our previous discussion the unexplained variability in our regression model. The human factor involved in the track maintenance may be the catalyst for said unexplained variability, and it would not be unexpected either, as tracks do require constant maintenance to perform optimally.

As it is expected that the restriction speed would vary greatly for each location, we are able to directly observe this by visualizing the data relevant to this observation.

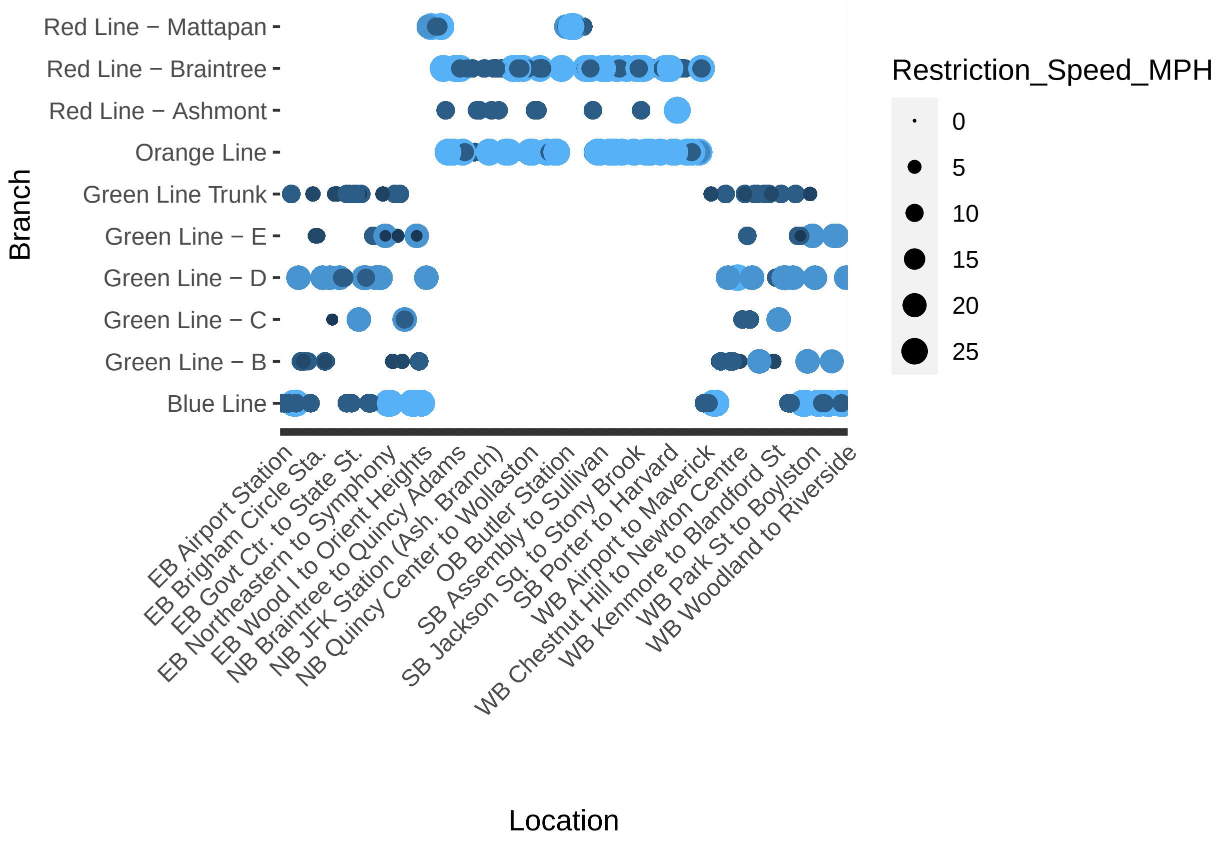

Figure 4.1: Balloon plot of restriction speed per line and branch.

For the figure above, several locations were omitted due to spacing issues. It is worth noting the significant variability of the speed restriction value. Also worth noting are the gaps in the plot, which is mainly due to the location not existing on that branch, and is not in fact due to a lack of speed restrictions. Since the Blue Line and the Orange Line do not have branches, their Line name takes the place of the Branch name in the Branch column, as it is displayed on the plot above.

From this plot we can directly observe the great variability of restriction speeds assigned to each branch and location, despite the fact that only seven restriction speeds are assigned.

5 Conclusion

We summarize our statistical findings below.

- The bar plots listing the daily restriction distance for each line indicates the Red Line as having suffered the most restriction distance. Also notable is the fact that the Green Line and the Blue suffer a sudden influx in restriction distance. This is especially true for the Blue Line, which at one point, had little to no speed restrictions reported.
- A hypothesis test was used to test if the restriction speed and restriction distance increased over the course of January through April 2023. We concluded from the results that while the restriction speed did not increase, the restriction distance did, in fact, increase. We also observed that the box plots displayed were relatively consistent, which was partially expected since we anticipated restriction distance and restriction speed to correlate to an extent.
- A linear regression model as well as a Q-Q plot of the restriction distance continuing between January and February was constructed. Both indicated the data set was not normally distributed, due to the points forming unorthodox patterns.
- We used a chi-square test to test whether the restriction speed was independent of the location it was assigned to. We concluded this was not the case, and that the restriction speed was in fact dependent on the location. We expected this outcome, since the location in which the speed restriction is assigned introduces a myriad of variables which affects how fast the train may travel for that specific location.

6 Potential Extensions For Data Research

- Since the vast majority of speed restrictions are due to track wear and tear, it may be beneficial to specify in a different column any aspects contributing to the wear and tear (constant usage, weather, terrain, etc). This would provide more insight on MBTA operation efficiency, and provide more reliable predictions on when these speed restrictions may be lifted.
- The sudden influx in restriction distance for both the Blue Line and Green Line warrants investigation. Given the Red Line also suffered a similar influx around the same period of time, it may be a prudent course of action to assume the reason for this influx of restriction distance was not exclusive to either line. An investigation may reveal the reason as to why the restriction distance increased so rapidly, and for multiple Lines no less.
- I have personally experienced significant speed restrictions on the Red Line, and this data set indicates that the Red Line indeed has the most taxing speed restrictions. The data sets include speed restriction data for the months of January through April, but it may be beneficial to have speed restriction data for previous years to be uploaded. This way it can be decided whether the MBTA has suffered this degree of severe speed restrictions before, or if the speed restrictions commuters have experienced thus far are exclusive to 2023.