Mini Project 1 - Capital BikeShare Data Analysis

Sarah Adilijiang, Yanqing Gui, Hanyang Peng

1 Background and Purpose

Capital Bikeshare makes its data public recently, which provides our statisticians an opportunity to incorporate data mining tools with real-life data. Utilizing multiple testing techniques, we could delve into the patterns of changes in route in-depth. Their dataset comes from Washington D.C.'s bikeshare program, which records every individual ride taken. Discovering changes in route could have further implications in many disciplines. For example, these changes in riding time could be used to investigate the causal effect of traffic factors on students' performance by researchers focusing on urban education if they could join this data with some education datasets.

Our main task in this bikeshare data mini project is whether we can detect any routes where the average time it takes to travel the route changes over the courses of the time period. And whether there are any possible confounders. We can perform some task-specific permutation tests to remove the influence of these confounders.

2 Data Preprocessing

1.1 Construct Data

The original data has 8 variables, and they are station_start, station_end, duration, starttime, day_of_week, days_since_Jan1_2010, member, and bikenum. Starttime consists of 6 variables, which are year, month, day, hour, minute, and second.

We think the important variables in our tasks are start time, start station, end station, duration, member, day of week, and days since Jan 1 2010. And we can process this data to create more specific variables, like routes, weekdays, and average duration.

For the routes variable, we can decide a specific route by looking at the data's start and end station. For the sake of simplicity, we assume that if two routes have the same start and end station, then we consider them as the same route. Because it is possible one can travel from spot A to spot B through different ways. However, according to the given data, we only know the start and end place of one ride, so we assume that they are the same route.

Therefore, we add the route variable in the data and construct a data frame. We check the dimension of the data frame and summarize it. From the R code, we find that the data frame has 14 variables and 1342364 rows. By summarizing the data, we find that there are many NA's in the bikenum variable. As there are many different bikes and we assume all bikes have similar conditions, so in this project we will not consider the bikenum as a confounder, and remove it for our future discussion.

For the unit of the time period, we think one day is a good unit. We can also make great use of the important variables day_of_week and days_since_Jan1_2010. So we compute the average duration of one day of the same route to remove the influence of the rush hours on this day. And the average duration varies between different days can be a good response to detect whether the routes are changed.

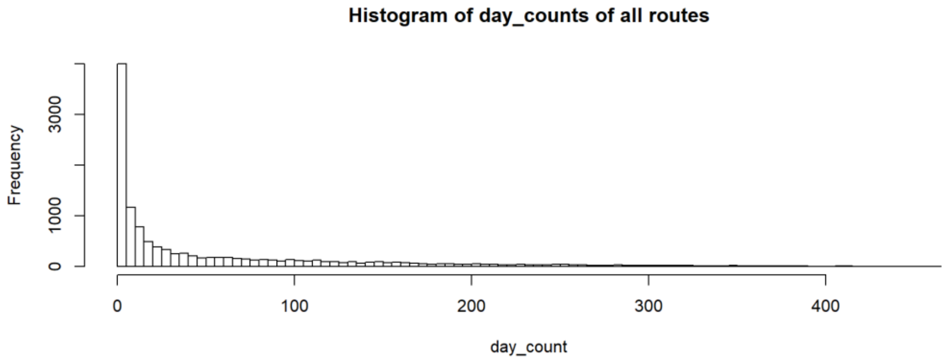
Now as we decide to research the relationship between routes and dates. We construct a new data frame that we aggregate on the station start, station end, year, month, and day. Then we compute the average duration of one day, specific route, day_of_week and days_since_Jan1_2010 on that day, and the length of the duration on that day for a specific route. We can see that after the aggregation and summary, there are 729321 data points left. However, we find that there are many days that only have one route on that day for a specific route. Actually, according to R, we find that there are 432793 data points, i.e. 59.34% of the data points, have only one record for that day.

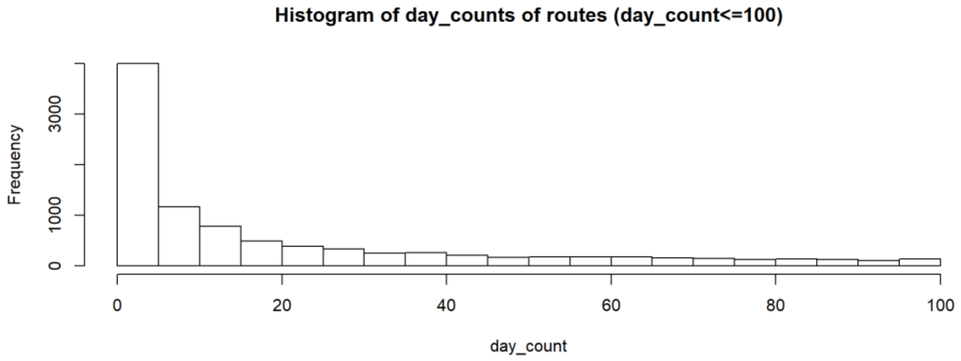
1.2 Select the routes with potential changes

As we discussed before, there are many data points that have little days' records, like one or two, which can be considered as not stable. Therefore, we should select routes first to make our computation stable and reasonable.

First, we should select the routes with day_count cutoff. To do this, we continue to aggregate on the new data frame we have constructed before. Now we aggregate the data frame by grouping by the routes, and count the number of the days shown in one route. We find that there are 12179 distinct routes in total. We plot the histogram of the day counts of the routes, and we find that most routes only have data less than five days.

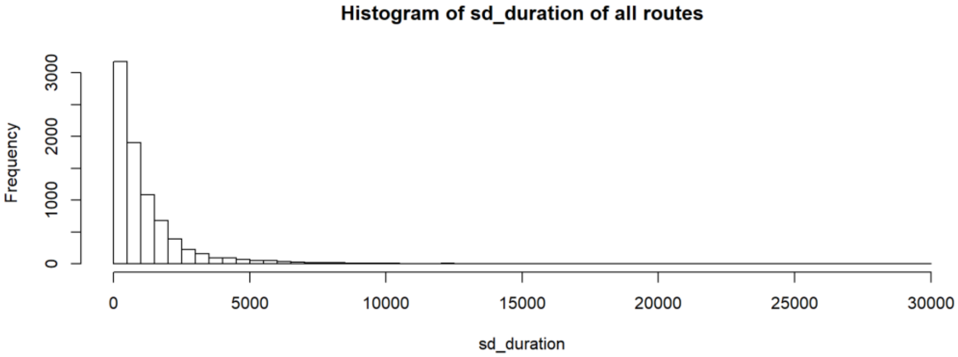
Because later we will permute the data within each group of {member, weekday}, which has four sub-groups for each route. To better compare, we only keep routes with day_counts over 4. And it also makes sense to remove the routes with low day_counts since it may be occasional data which does not well represent the average duration change of one route. So we remove the routes whose day_counts are no larger than 5, and there are 8187 distinct routes left.

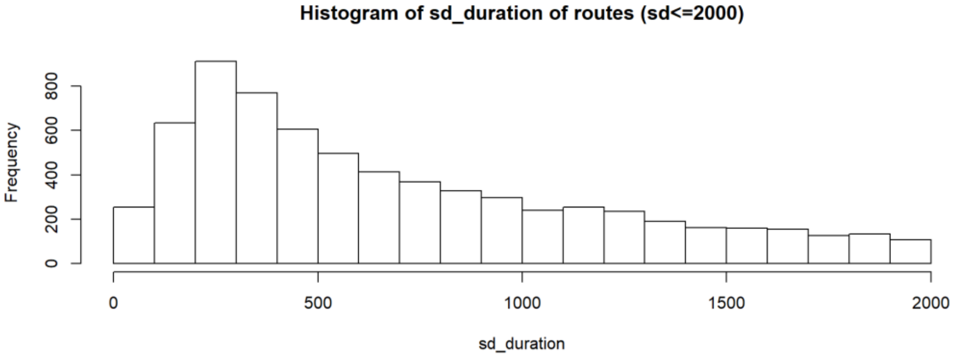




Then we will select the routes according to its duration time. We assume that if routes change over time, there should be a large trend change in its duration. So we may find a larger value in its standard deviation. We also plot the histogram of the standard deviation of the duration for the routes selected before.

From the plots, we find that most of the routes have standard deviation of duration equal to or lower than 300 seconds, i.e. 5 minutes. We think it makes sense to remove these routes since the standard deviations less than 5 minutes indicates no significant time changes according to common sense. But for some short routes, 5 minutes could be a large change in duration. Therefore, we also keep one route if its standard deviation is larger than 1/3 of its mean of duration. After removing these routes, there are finally 6641 distinct routes left in our data, while 5538 routes have been removed from the original data.





As we have already decided which routes we should select to detect the routes that change over courses of the time period. Now we remove these routes from originally constructed data frame before aggregation. Now we have 655290 data points left in total, which is about 90% of the original data. Therefore, we can see that even though we have removed many routes, most of the data points have been kept.

Now we can perform the permutation test and find the routes that may change their average duration over the courses of time period.

3 Permutation Test

We use permutation strategies to find out the routes that change over the courses of the time period. We can assume that if one route changes, for example, a bike lane is added to a major road, then there will be some change in average duration for different dates and this change has some pattern over dates, which will be significantly different from a random duration change. For example, if one route is modified and a bike can travel faster on it, then the average duration after some date may be much less than the duration before that date. In contrast, if one route is not modified, then the average duration will just vary randomly, and we can't find any trend by looking at its average duration and date.

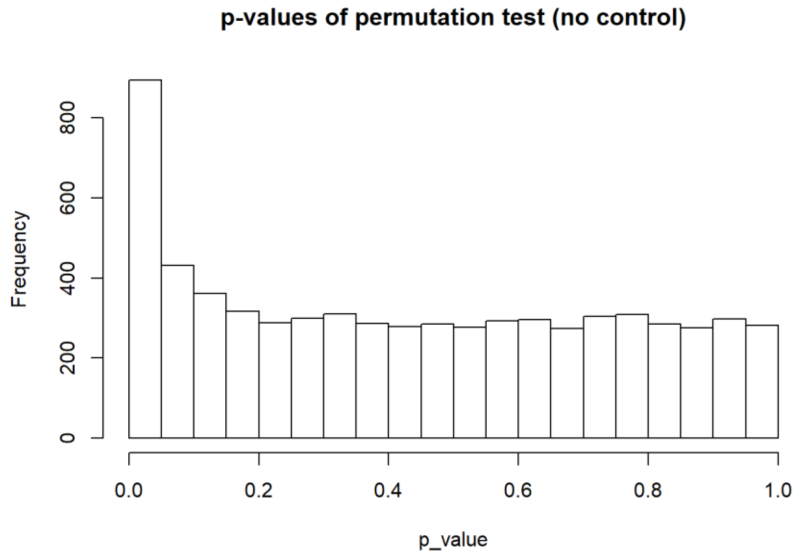
That is to say, the average duration will be correlated with the dates if there is any route change. By using a permutation test, we can find out whether there is a significant correlation or not. The formula of computing the p-value is shown below, while the test statistic T is defined as the absolute value of correlation between the average duration per day and the days since Jan 1 2010 (will be mentioned as date from now on).

$$p = \frac{1 + \sum_{m=1}^{M} \quad 1\{{T_m}^{perm} \quad \geq T \ \}}{1 + M}$$

Recall that after the above data preprocessing, there are 6641 distinct routes left in our data now. We will run permutation 500 times and compute a p-value for each route.

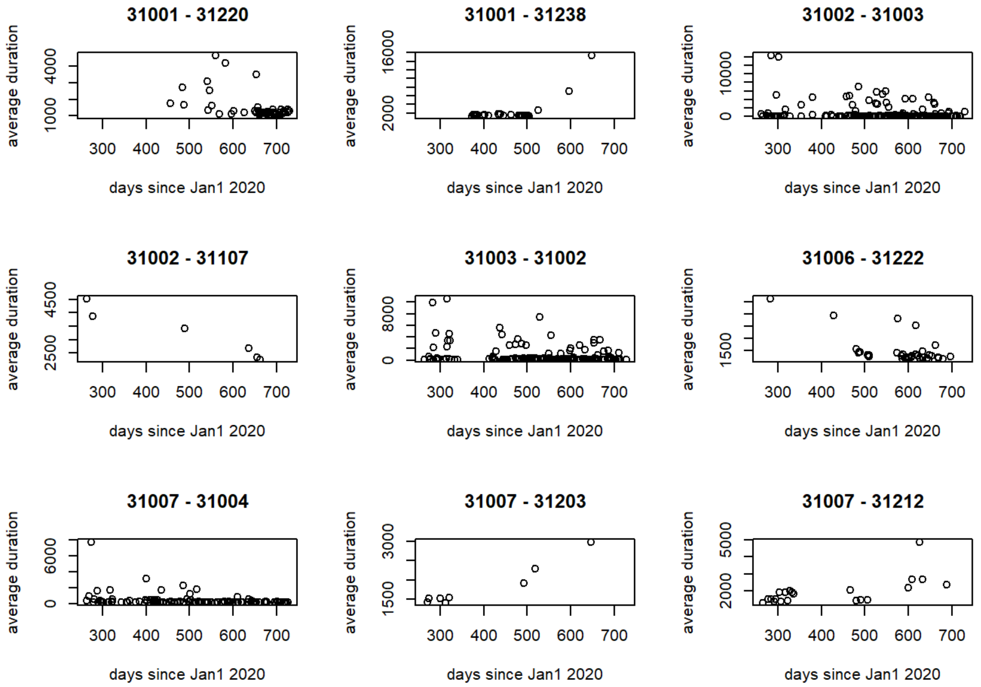
3.1 Permutation test (no control for confounders)

First, we only perform a permutation test for each route without controlling for any confounders. Thus the average duration will be randomly permuted. The histogram of the p-values is shown below.

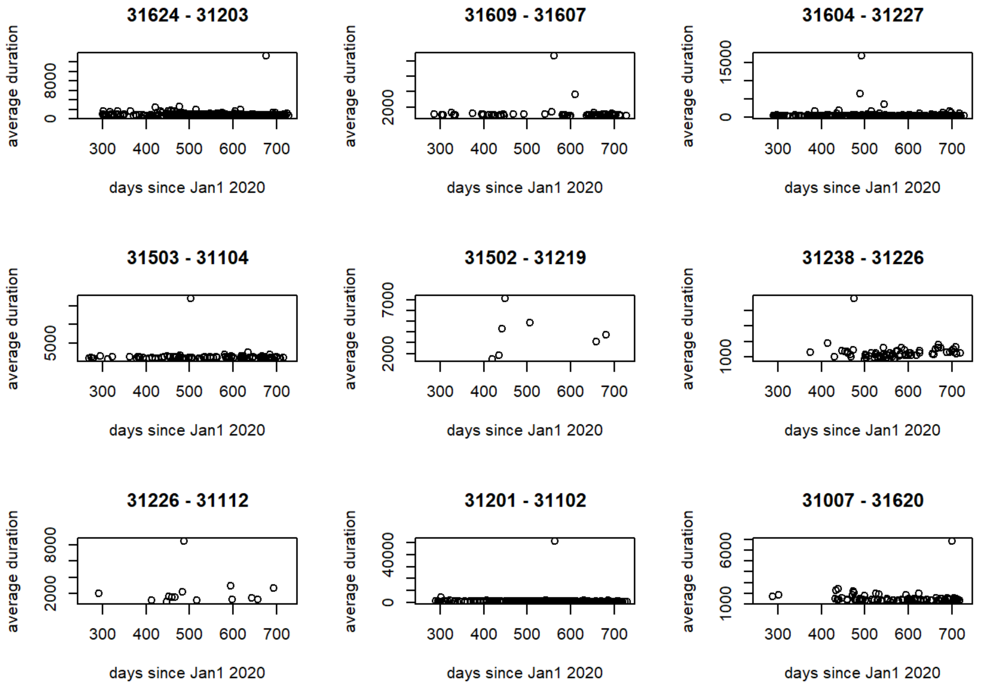


We find that there are 1324 unique routes with p-values equal to or smaller than 0.1, and there are 893 unique routes with p-values equal to or smaller than 0.05. As a result, when not considering any confounders and not yet controlling for FDR, there are 893 routes that change over time while the significant level is 0.05.

Now we order the p-values and plot figures for some most significant and most non-significant routes, showing the changes of average durations per day over the dates.



Routes with most significant p-values (no control)



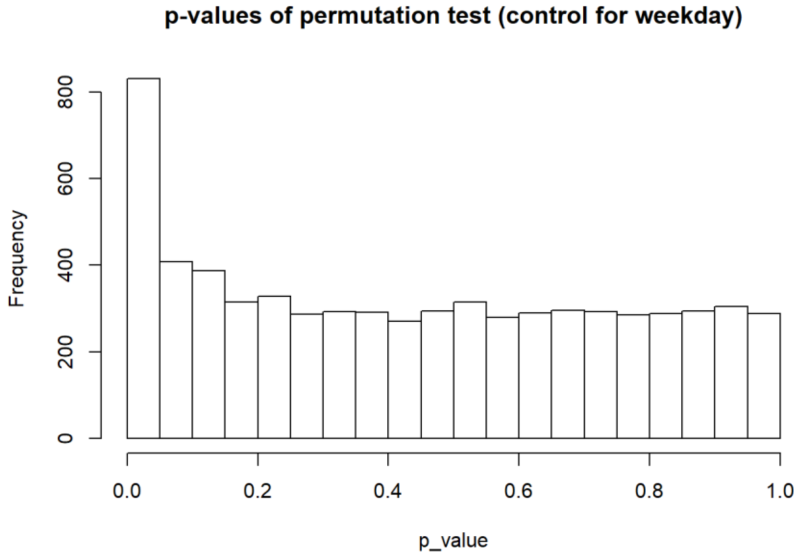
Routes with most non-significant p-values (no control)

We can see that in the plots of routes with most significant p-values, many of them show a relatively obvious trend of increasing or decreasing in average duration per day over the dates.

And in the plots of routes with most non-significant p-values, almost all of them have no such trends, which only show some distribution with no clear patterns, or almost flat lines with some random spikes. From now on, for the other following tests, we will only show the plots for most significant p-values, because all the non-significant p-values have this kind of similar results.

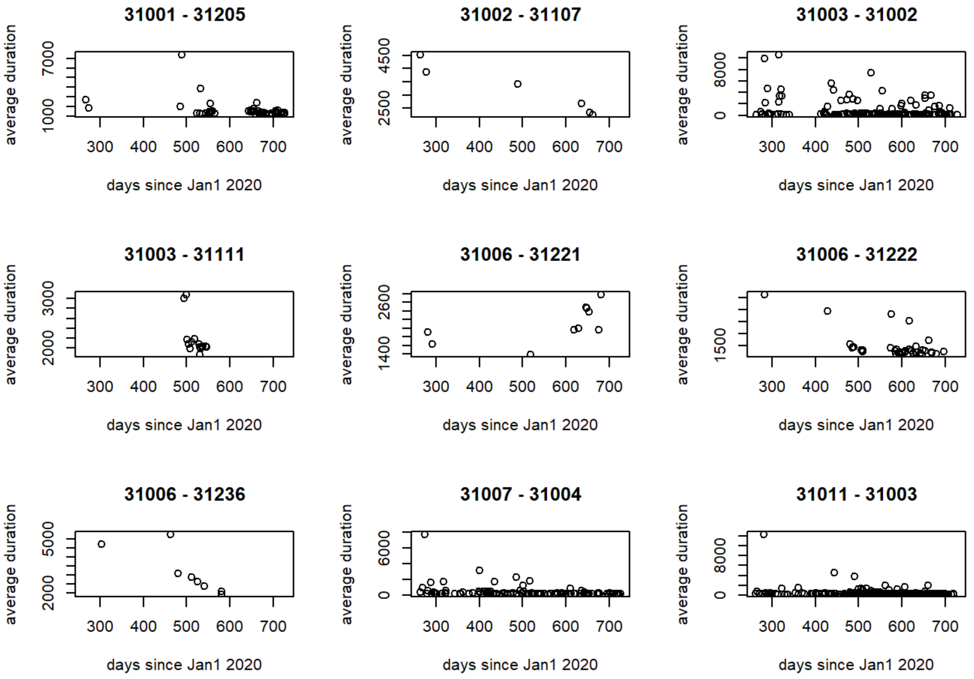
3.2 Permutation test (control for weekday)

Since weekdays' riding durations may be systematically different from that of weekend ridings, we add a dummy variable in our data frame to show whether a day is a weekday or weekend. There are about 74% of the data points are weekdays' data, which is close to 5:2 ratio. Then we carry out a task-specific permutation strategy, where we permute the average duration within the weekday group and the weekend group for each route. Again, for each group, we permute 500 times and calculate the P-values. The histogram of the p-values is shown below.



We find that there are 1238 unique routes with p-values equal to or smaller than 0.1, and there are 830 unique routes with p-values equal to or smaller than 0.05. As a result, when not considering any confounders and not yet controlling for FDR, there are 830 routes that change over time while the significant level is 0.05. We can see that the number of significant routes are reduced compared to the case with no control.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.



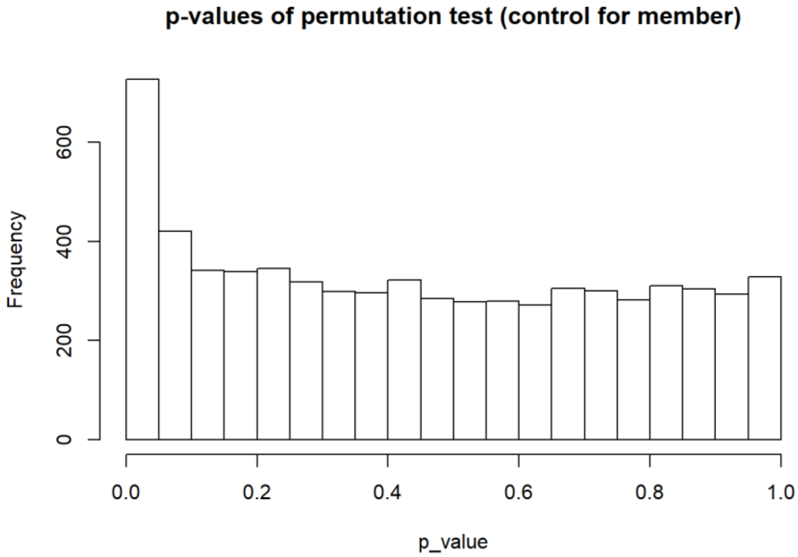
Routes with most significant p-values (control for weekday)

We can see that in the plots of routes with most significant p-values, many of them still show a relatively obvious trend of increasing or decreasing in average duration per day over the dates.

As a result, we can say that, our task-specific permutation test controlling for weekday factor, generally does a good job in selecting routes with ride time changes.

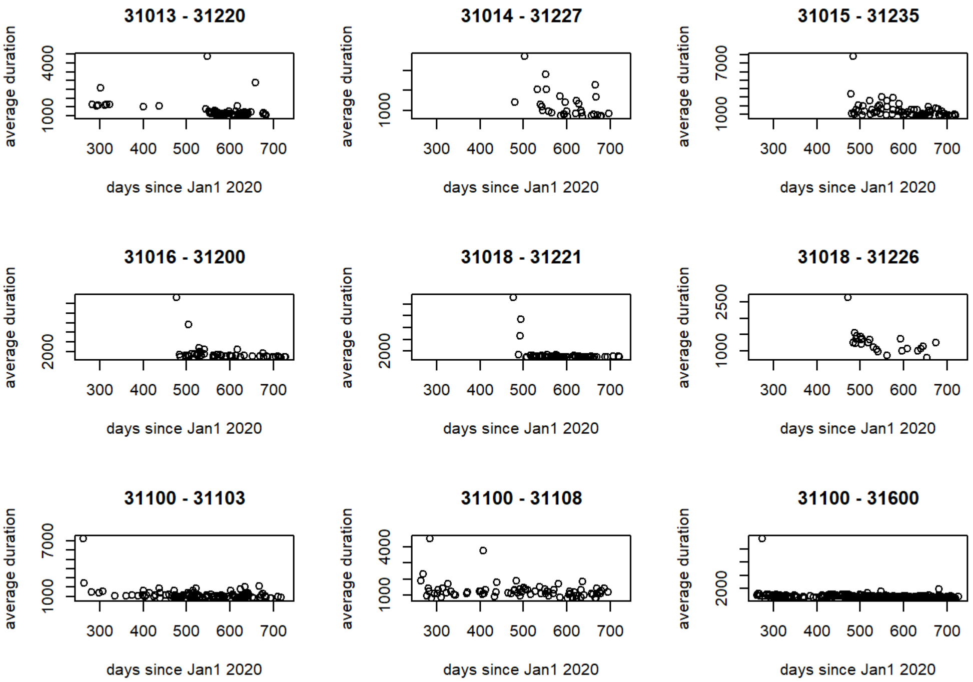
3.3 Permutation test (control for member)

Since members may ride faster than one time users (i.e. not members), so whether the rider of one data point is a member or not may also affect the results. Therefore, we construct a new data frame which is similar to the former one, but this time, we also group the data by its member variable. Now there are 726701 data points, and around 80% of the data points are the members' data. Then we perform a similar task-specific permutation test, where we permute the average duration within the member group and the non-member group for each route. Again, for each group, we permute 500 times and calculate the P-values. The histogram of the p-values is shown below.



We find that there are 1146 unique routes with p-values equal to or smaller than 0.1, and there are 726 unique routes with p-values equal to or smaller than 0.05. As a result, when not considering any confounders and not yet controlling for FDR, there are 726 routes that change over time while the significant level is 0.05. We can see that the number of significant routes are also reduced compared to the case with no control.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.



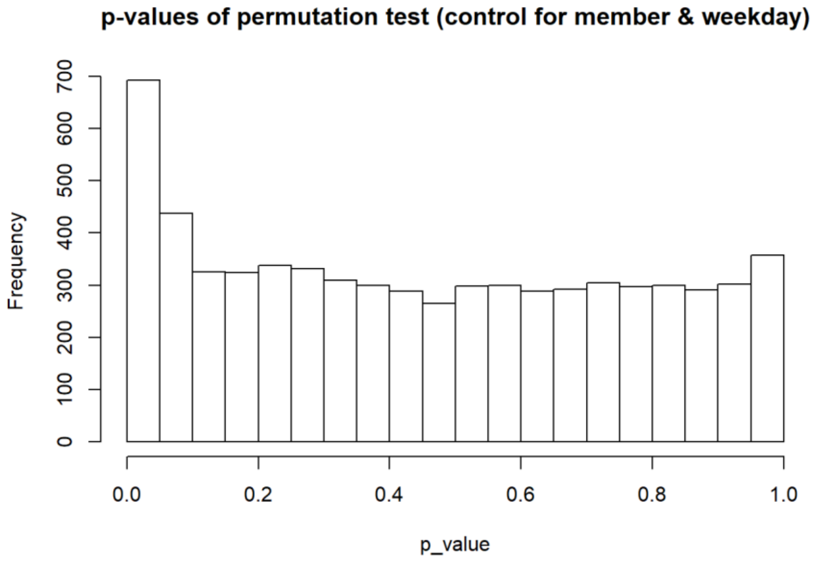
Routes with most significant p-values (control for member)

We can see that in the plots of routes with most significant p-values, many of them still show a relatively obvious trend of increasing or decreasing in average duration per day over the dates, but these trends are not as obvious as in the case controlling for weekday.

As a result, we can say that, our task-specific permutation test controlling for member factor, basically does a good job in selecting routes with ride time changes, but not as good as the case controlling for weekday.

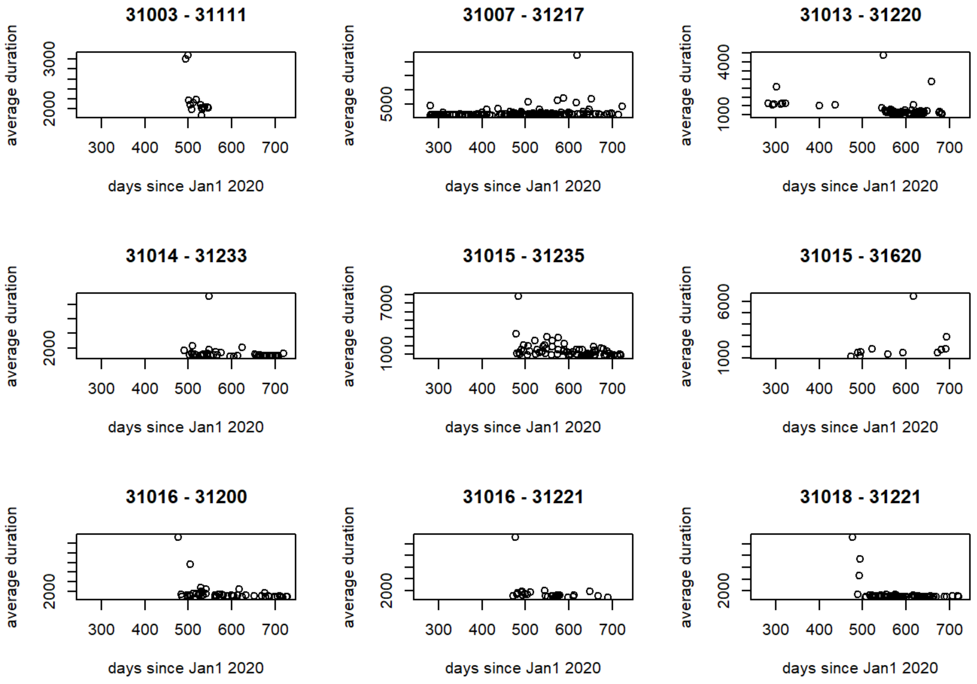
3.4 Permutation test (control for member and weekday)

Now we control for both the weekday factor and the member factor. We perform a task-specific permutation test, where we permute the average duration within the four combinatorial groups for each route. Again, for each group, we permute 500 times and calculate the P-values. The histogram of the p-values is shown below.



We find that there are 1130 unique routes with p-values equal to or smaller than 0.1, and there are 692 unique routes with p-values equal to or smaller than 0.05. As a result, when not considering any confounders and not yet controlling for FDR, there are 692 routes that change over time while the significant level is 0.05. We can see that the number of significant routes are also further reduced compared to the case with only controlling for weekday or member.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.



Routes with most significant p-values (control for member and weekday)

We can see that in the plots of routes with most significant p-values, now there are almost no clear trends of increasing or decreasing in average duration per day over the dates as shown in previous cases.

As a result, we can say that, our task-specific permutation test controlling for both member and weekday, basically does not do a good job in selecting routes with ride time changes. In this case, we might have “over-grouped” our data for the permutation test.

3.5 Modified BH method to control FDR and the results

In the above tests, we are testing the correlations for 6641 unique routes, so there are multiple testing problems. In this case, we need to avoid false positive discoveries by controlling the FDR. Here we use the modified BH method as in homework 2, with alpha=0.1 and gamma=0.5. We perform the modified BH method for all the four cases of p-values and get the following results.

Case	No control	Weekday	Member	Member and Weekday
# significant p-values (alpha = 0.05)	893	830	726	692
# significant p-values after modified BH method	157	133	0	0

We can see that the modified BH method will largely reduce the discovered significant p-values in all the four cases. However, for the cases when we are controlling for member or both member and weekday, the number of discoveries has been reduced to zero. This may because when we are generating the data with member factor added, there might be two data points for a single day showing the average duration time spent by members and non-members, while most of the days only have one data point of member data, since about 80% of the whole data is the member's data. This may cause some problems in the correlation calculation. Though the permutation results show less significant p-values than the no-control case, the BH method does not report any significant p-values in the end.

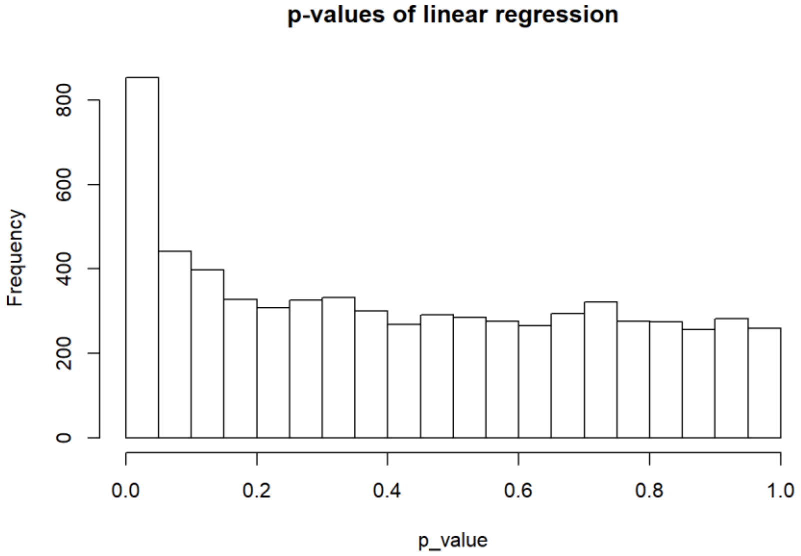
According to this result, we may find that the permutation test where we only control for the weekday is the most reasonable way. And it finds out 133 unique routes with significant ride time change over the courses of time period after using the modified BH method to control for FDR.

4 Linear Regression

We also tried the linear regression way to find the routes with significant ride time change over the time period. Here we use the same 6641 routes as the candidates. If one route has an increasing or decreasing trend over the date, then the estimated coefficient of the variable date would be significant.

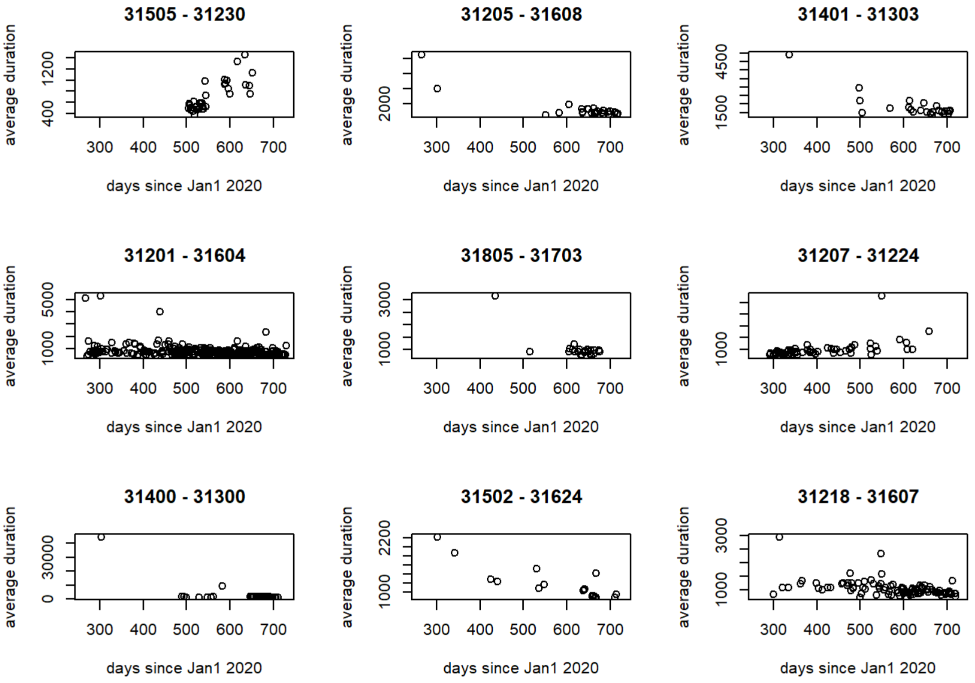
4.1 Linear regression (no control for confounders)

First, we only regress the average duration per day over the date and record the p-value of the coefficient of the date. The histogram of the p-values is shown below.



We find that there are 852 unique routes with p-values equal to or smaller than 0.05. After the modified BH method to control for FDR, there are 180 unique routes with smallest p-values left.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.

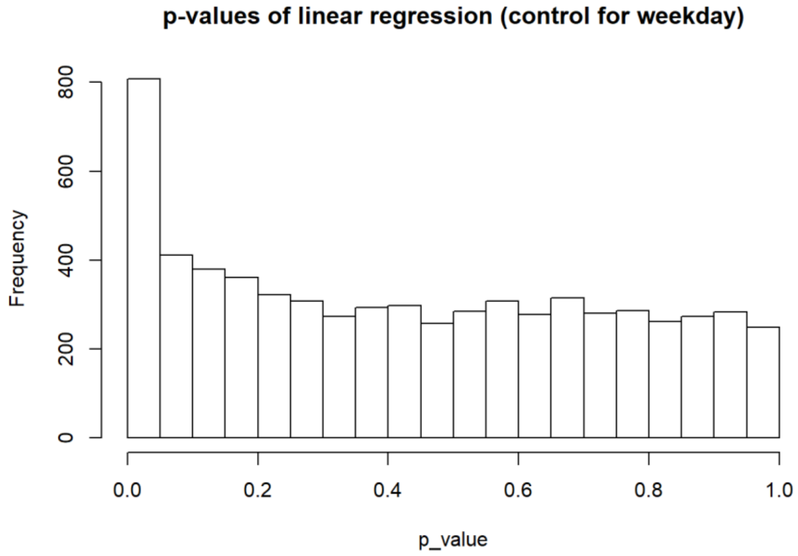


Routes with most significant p-values (no control)

We can see that in the plots of routes with most significant p-values, some of them show a relatively obvious trend of increasing or decreasing in average duration per day over the dates, but some of them just have no clear trends. This is not as good as the results from the permutation test with no control.

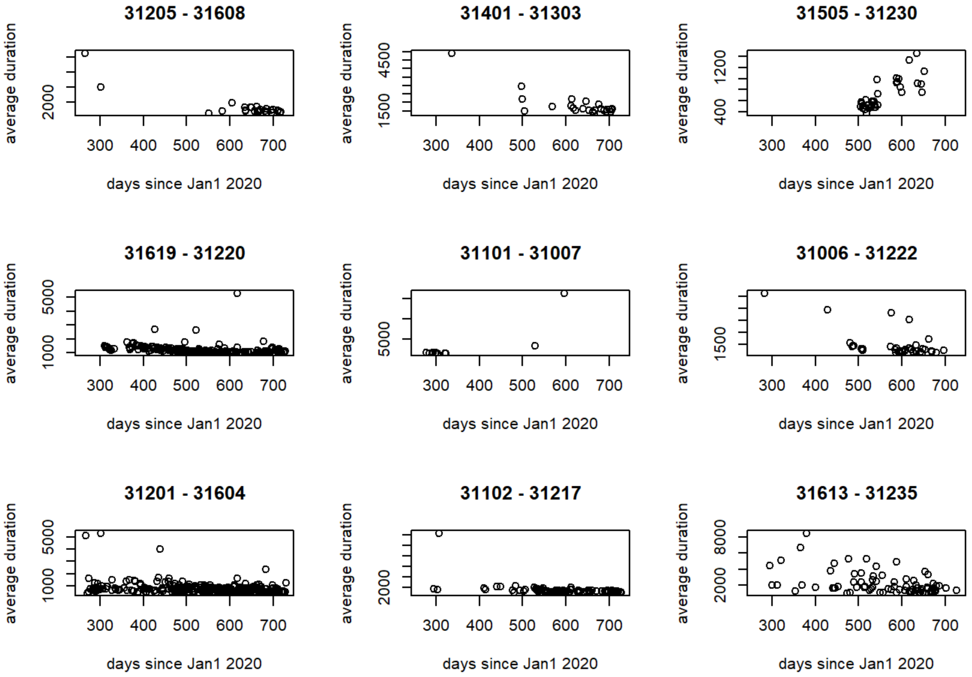
4.2 Linear regression (control for weekday)

Now we regress the average duration per day over the date and weekday factor. The p-values of the coefficients of the dates and weekdays are recorded. The histogram of the p-values is shown below.



We find that there are 807 unique routes with p-values equal to or smaller than 0.05. After the modified BH method to control for FDR, there are 120 unique routes with smallest p-values left. We can see that the number of significant routes are reduced compared to the case with no control.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.

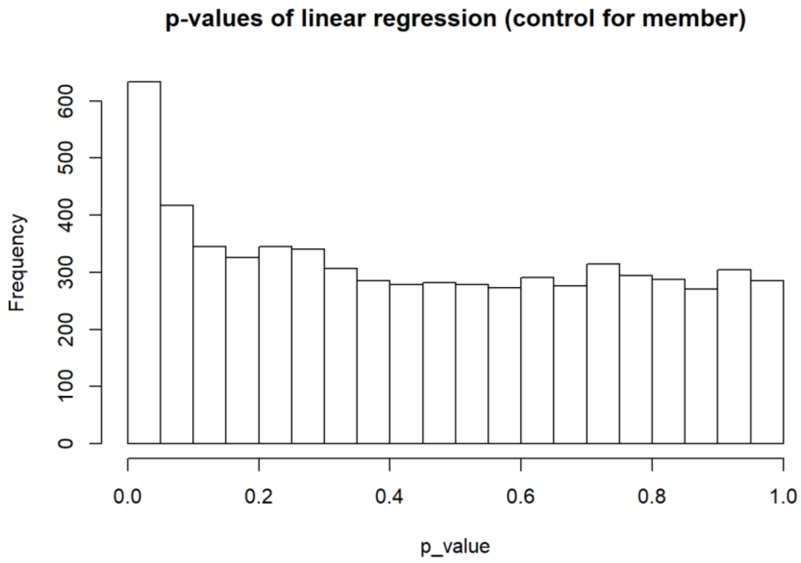


Routes with most significant p-values (control for weekday)

We can see that in the plots of routes with most significant p-values, again, some of them show a relatively obvious trend of increasing or decreasing in average duration per day over the dates, while some of them have no clear trends. This is also not as good as the results from the permutation test when controlling for weekday.

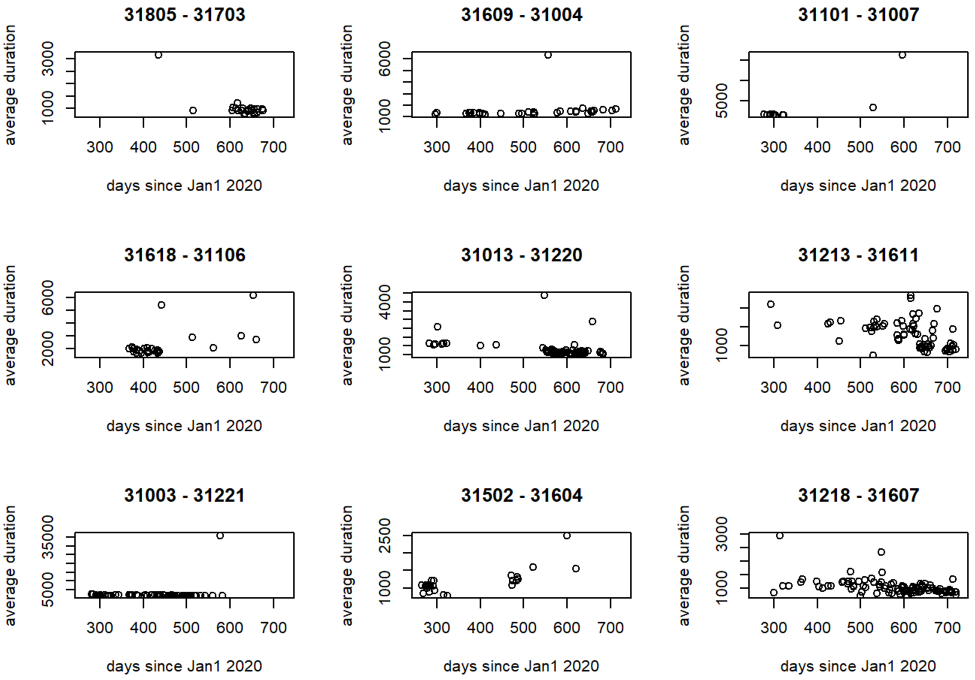
4.3 Linear regression (control for member)

Now we regress the average duration per day over the date and member factor. The p-values of the coefficients of the dates and members are recorded. The histogram of the p-values is shown below.



We find that there are 633 unique routes with p-values equal to or smaller than 0.05. After the modified BH method to control for FDR, there are 77 unique routes with smallest p-values left. We can see that the number of significant routes are also reduced compared to the case with no control.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.

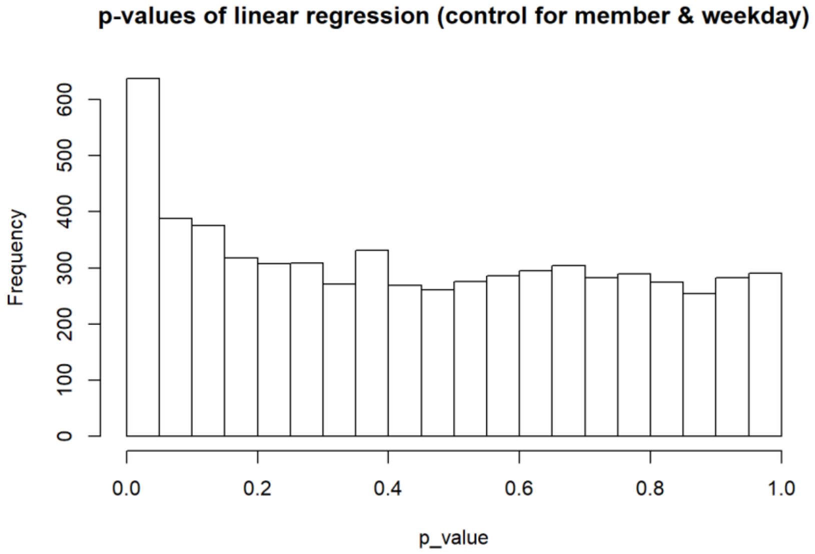


Routes with most significant p-values (control for member)

We can see that in the plots of routes with most significant p-values, again, some of them show a relatively obvious trend of increasing or decreasing in average duration per day over the dates, while some of them have no clear trends. This is also not as good as the results from the permutation test when controlling for member.

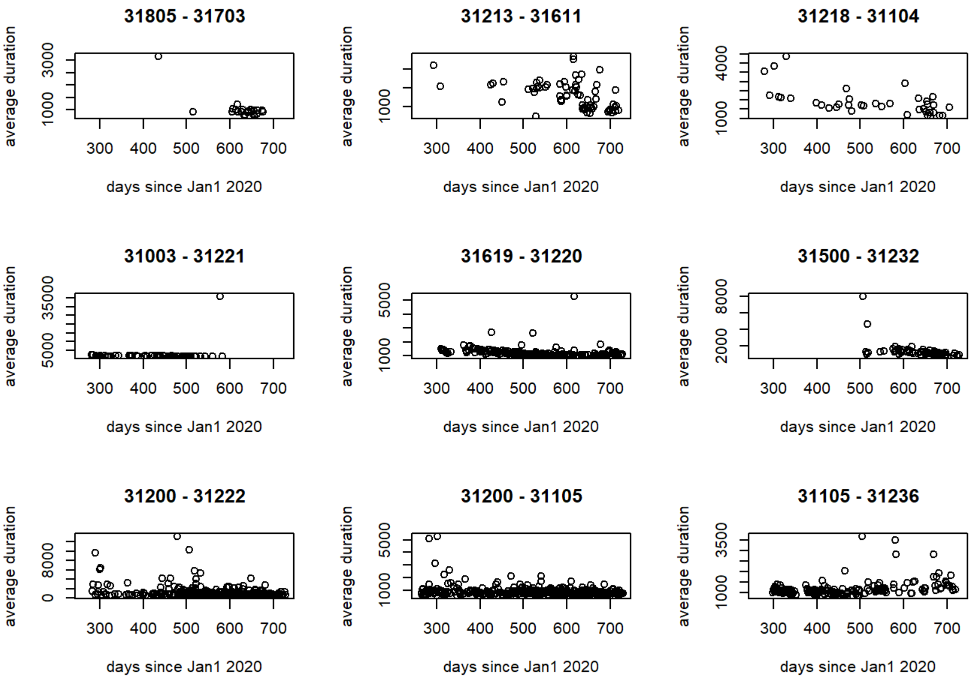
4.4 Linear regression (control for member and weekday)

Now we regress the average duration per day over the date, member factor, and weekday factor. The p-values of the coefficients of the dates, members, and weekdays are recorded. The histogram of the p-values is shown below.



We find that there are 637 unique routes with p-values equal to or smaller than 0.05. After the modified BH method to control for FDR, there are 62 unique routes with smallest p-values left. We can see that the number of significant routes are further less than the cases when only controlling for member or weekday.

Now we order the p-values and plot figures for some routes with the most significant p-values, showing the changes of average durations per day over the dates.



Routes with most significant p-values (control for member and weekday)

We can see that in the plots of routes with most significant p-values, again, some of them show a relatively obvious trend of increasing or decreasing in average duration per day over the dates, while some of them have no clear trends. This is slightly better than the results from the permutation test when controlling for both member and weekday.

4.5 Combine results

In the above tests, we are still testing the correlations between the response ride time with the variable date for 6641 unique routes, so there are multiple testing problems. We have performed a similar modified BH method to control FDR, with alpha=0.1 and gamma=0.5. The results after the BH method have been mentioned above in each case. Here we organized them together to have a better look.

Case	No control	Weekday	Member	Member and Weekday
# significant p-values (alpha = 0.05)	852	807	633	637
# significant p-values after modified BH method	180	120	77	62

We can see that the modified BH method will largely reduce the discovered significant p-values in all the four cases. This time, different from the results from the permutation tests, for the cases when we are controlling for member or both member and weekday, the number of discoveries are not reduced to zero. If only based on the results of linear regression method, we may find that the linear regression where we control for both member and weekday is the most reasonable way, which finds out 62 unique routes with significant ride time change.

However, according to the plots for discovered routes with changes in the above four cases, we find that the routes' ride time changes reported by using linear regression method is not as good as the results from corresponding permutation tests. Therefore, comparing the two methods, we would prefer the permutation test as the tool to find the routes with change in average duration per day over the courses of time period.

5 Conclusion

We used two methods, permutation test and linear regression, to find out the routes that change over the courses of the time period within the 6641 unique routes selected by data preprocessing.

For permutation test, the task-specific permutation test where we only control for the weekday is the most reasonable way. And it finds out 133 unique routes with significant ride time change over the courses of time period after using the modified BH method to control for FDR.

For linear regression, the linear regression where we control for both the member and the weekday is the most reasonable way. And it finds out 62 unique routes with significant ride time change over the courses of time period after using the modified BH method to control for FDR.

However, in general, the results obtained from linear regression are not as good as the ones from permutation tests, because the plots of discovered routes are not showing as many as relatively clear trends in increasing or decreasing of the average duration from the permutation tests. Therefore, based on these findings, we would conclude that the permutation test performs better than linear regression to find significant correlations for this bikeshare dataset. And we would finally report the results from the task-specific permutation tests controlling for the weekday, where 133 unique routes are found for having significant ride time change over the courses of time period. And its FDR is controlled under level 0.1.