Orienteering Challenge Data Modeling

Shivangi Saxena, Patrick Byrne, Chris Stoafer

Introduction

Our team has chosen to analyze the available data from a rogaine held this March outside of Pittsburgh, PA. In a Rogaine, teams of hikers are assembled in one location and given a topograhic map. This map is annotated with a series of checkpoints, a legend giving the point value of each checkpoint, and a very rough description of the points' location, allowing teams who are nearby to locate the point regardless of surrounding terrain. Teams score points by locating the checkpoint, and connecting an RFID tag carried by a team member to one attached to the checkpoint. The points are so widely spread across an area of ~30 square kilometers that it is highly unlikely any team would be able to score on all of them. In fact, in this race, the most points any team was able to reach was 38 of a possible 50, in 6 hours. The difficulty is highlighted by the fact this particular team reached nearly 150% as many checkpoints as the second place team.

Teams are scored based on the points they score in the time limit, not on the number of checkpoints reached. There is a heavy penalty for not returning to the start before the time limit is up. The point value of the checkpoints tends to be correlated to the distance from the start, and the difficulty of the surrounding terrain. Given that teams are tasked with maximizing their scores by selecting a subset of different value checkpoints, each with different costs measured in terms of required time, along a route that returns to the start within the time limit, this activity represents a real-world application of the Traveling Salesman and Knapsack problems. The field of competitors was split into those allowed 3 hours, and those allowed 6 hours to complete the course, which we would naively suspect would lead to vastly different strategies.

Given that teams progress is electronically recorded, there is a large, but still incomplete dataset available for analysis. Hardcore enthusiasts track their progress with GPS, which allows them to retroactively determine which areas of their route gave them the most trouble, and which ones they progressed though fastest, allowing them to adjust their gameplan in later events. However, an analysis of the entire field of competitors is an unexamined problem, which may provide insights into optimal strategies. At the very least, this project should be of interest to the orienteering community, but the fundamental problem - using an incomplete and noisy dataset, maintained and uploaded by people who are not data-scientists, to optimize strategies relevant to many fields of endeavor - should be of much broader interest.

The available data was taken from the results page of event's public website. These results were in poorly formatted html tables, and required significant cleaning and collation in both R and Python to make the dataset useable. The data provided are as follows:
The distance, in meters from any checkpoint to the start/finish, as well as each points elevation above sea level.
The number of teams to visit each checkpoint.
The teams competing, their size, time limit, affiliation with the orienteering club presenting the rogaine, their age level - defined as that of the youngest member of the team, and whether they rented or owned their RFID equipment - taken here to be a proxy for skill and experience with orienteering challenges.
The points scored by each team, their place in the event, and the order in which they progressed through the checkpoints, and the time splits for each team to reach each checkpoint.

Noteable gaps in the data which must be accounted for:
The topology of the terrain. This data *is* publicly available, from the same dataset as that used to build the maps given to the competitors. However, interpreting the data requires specialized cartographic software, and given the time constraints of the project, we will be considering the absolute elevation difference between the points only.
The terrain itself. The green areas on the map represent thick undergrowth. Thornbushes, felled trees, and rain-slicked muddy hillsides all delayed teams passing through them, or caused teams to route their course around them. The subjective and highly variable nature of this variable makes taking proper account of it almost impossible. The time required to pass between two checkpoints, normalized to their straight-line distance and elevation change will be used as a proxy for the difficulty of the intervening terrain.
The real-time paths taken by the teams. While this data was published on a voluntary basis by a few teams, they represent a very small fraction of the number of teams competing. Therefore, the actual paths taken by each team between points is not available, though it *is* seen from the available data that in almost no cases were those paths straight lines, for the reasons just stated. Further, deviations from expected behavior cannot be discerned with high resolution. For instance, the team to which one of the authors belonged took a brief 10 minute

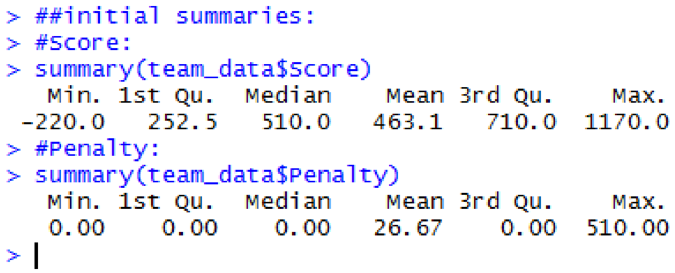
rest/lunch break slightly after noon. This would be represented in the team taking a correspondingly longer time between the last point reached and the next. By comparing the time taken on that leg by that team with the time required by other teams, we might be able to surmise that *something* slowed this team more than simply the terrain, but we will be able to say nothing more than that.

EXPLORATORY DATA ANALYSIS

Once we had the data in CSV files, we used R and iPython to check for patterns and get more insight on the relations between different parameters of the data. There were two kinds of datasets – i) data about the checkpoints, ii) data about the teams.

1. team_info_score.csv
    a. Team ID
    b. Team Name
    c. Time Limit – two kinds of teams, with one kind being given 3 hours and the other 6 hours
    d. Category
    e. Size – number of members in team
    f. Borrow – If the members had to borrow the RSID or had their own (could be an indicator of how frequently they participate in such races)
    g. Test
    h. Rank – how the team fared
    i. Class – basically a combination of Time Limit & Category.
    j. Club – whether the participants are Club members
    k. Gross Score – Score based on number of points visited
    l. Penalty – penalty based on how late after race duration they reached Finish point.
    m. Score – cumulative score (Gross score - penalty)
2. team_routes.csv
    a. Team ID
    b. Team Name
    c. Checkpoint #
    d. Total Time - time taken by team till that checkpoint
    e. Split – time taken to reach that checkpoint by that team from the previous point

We started with some elementary analysis to see how the data is spread out:



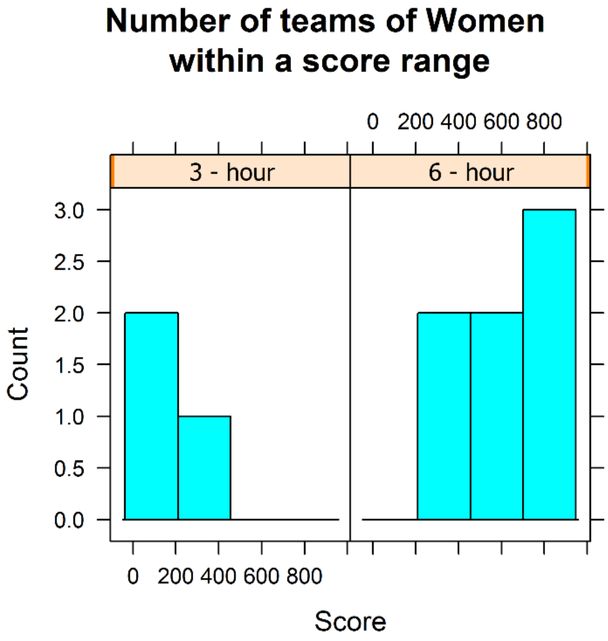
```
> ##initial summaries:
> #Score:
> summary(team_data$Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -220.0   252.5   510.0   463.1   710.0  1170.0 
> #Penalty:
> summary(team_data$Penalty)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00    0.00   26.67    0.00  510.00 
> |
```

As seen above, the mean score for all teams was about 463, with the maximum score being 1170 and lowest score being -220. Negative scores indicate that the no. of points gained by crossing checkpoints was lesser than the penalty received.

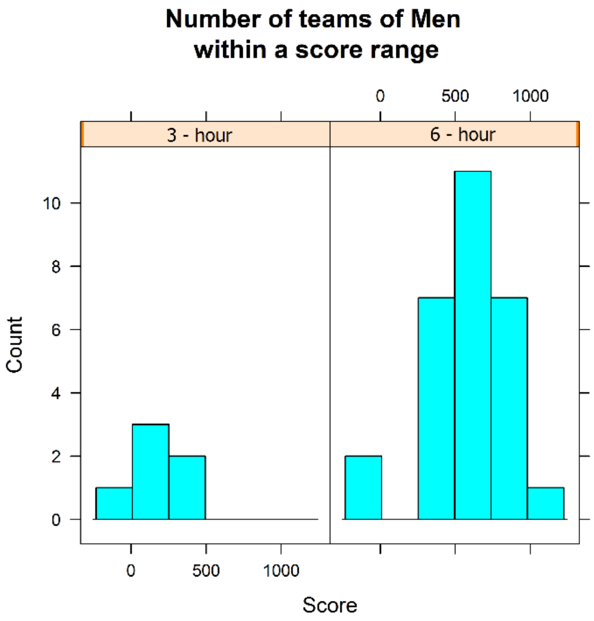
Next, a comparison between the performances of the different kinds of teams, based on gender was made. There were three kinds of teams – all men (M), all women(W), and 'mixed' (B), with a maximum of 5 members and a minimum of 1 members, with most teams having 2 members. We measured the performance separately for teams running for 3 hours and 6 hours.

- **Analyzing performance of different gender types:**

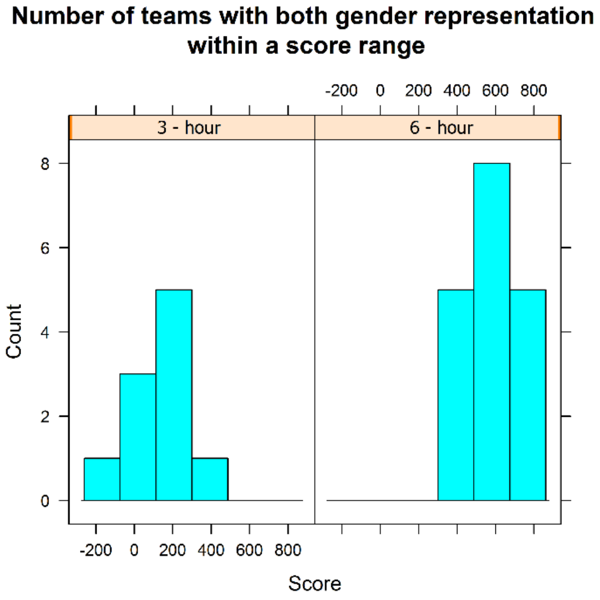
Distribution of scores for W was as follows:



Distribution of scores for M was as follows:



Distribution of scores for B was as follows:

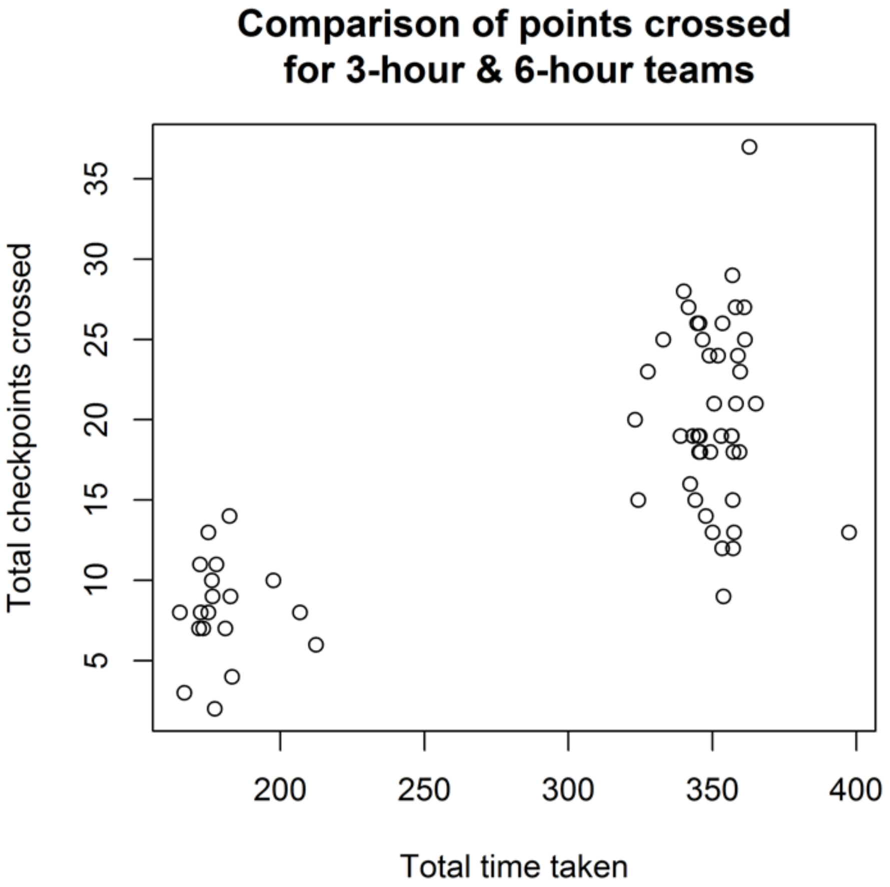


As seen in the above histograms, for 3 hour teams, women performed the worst, while men performed the best. However, it should be noted that the number of 3-hour women teams is very less which makes it difficult to make conclusions from this data.

Among 6 hour teams, the observations were very different. The W teams performed the best, with M and B teams both coming second. Again, women representation seems to be much lower than that of the other teams, which could explain the anomaly.

- **Comparing time taken and checkpoints crossed:**

Looking at the data with reference to the checkpoints, we compared how many were crossed by different teams:



Comparison of points crossed
for 3-hour & 6-hour teams

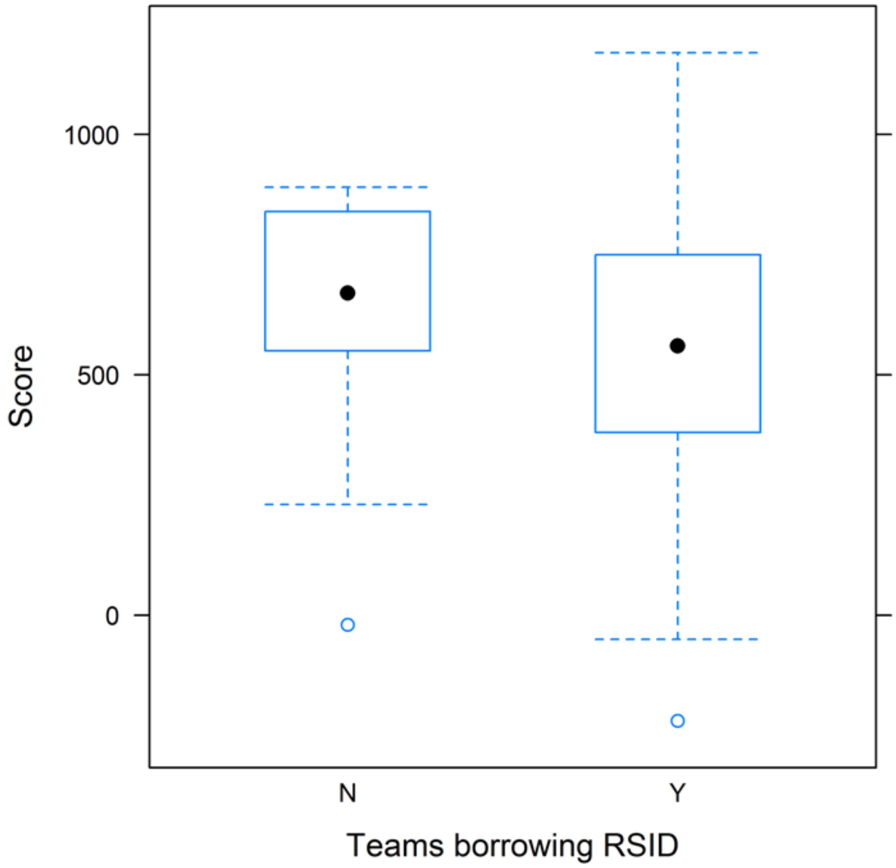
From the many clusters, we can see how much the number of points crossed can vary across teams, despite using the same amount of time to cover them.

For the 6-hour teams, some crossed as many as 30, while one covered less than 10 points. The distribution is not as spread out for the 3-hour teams, with a good majority getting about 6-11 checkpoints. With regard to penalty too we can see at least 3 teams going above the 180 minute limit, while in the 6-hour category, one team was really late, coming in at about 400 minutes.

- **Checking for correlations:**
  - Between "Borrow" and "Score"

  Our hypothesis was that only the team members that regularly took part in such races would buy their own RSIDs, thereby not requiring to "borrow" the RSID from the organizers. So we considered the 'Borrow' attribute to be an important indicator of the team's performance, based on their expertise.

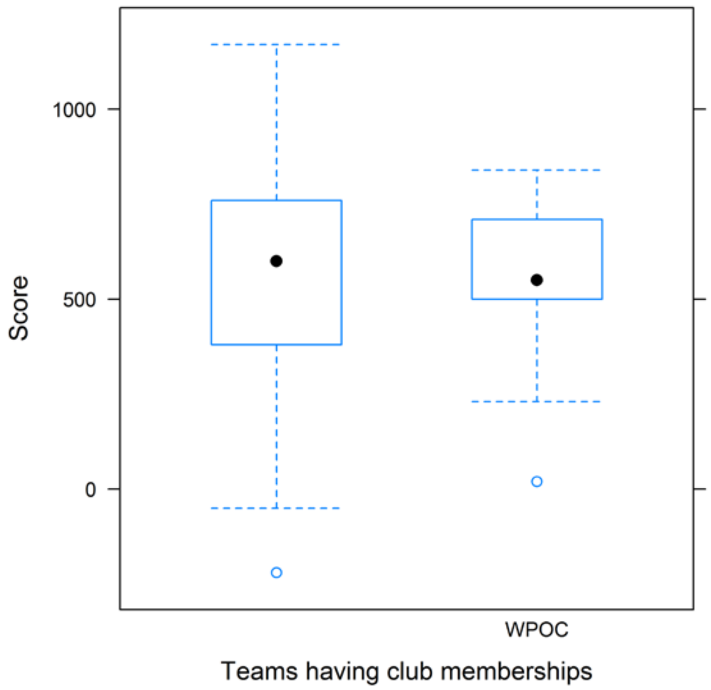
So, we plotted a graph to compare the scores of teams w.r.t. whether or not they borrowed the RSID.



As expected, the median score of the teams that didn't borrow tended to be higher than the ones who did.

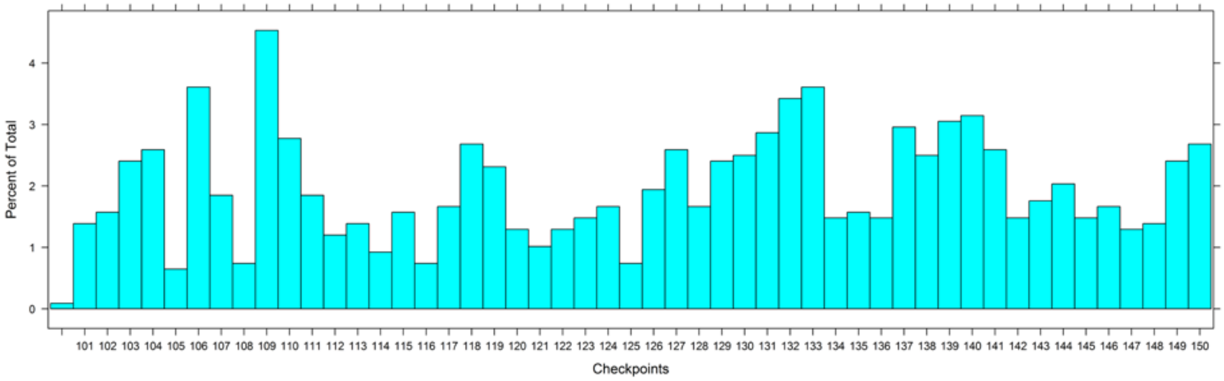
- Between "Club Membership" and "Score"

Similar to the previous attribute, we checked the Club membership attribute to check how well they did in the race. Team members that were members of the "Western Pennsylvania Orienteering Club" could be expected to be regulars in such events, and so could possibly be doing better than the rest.



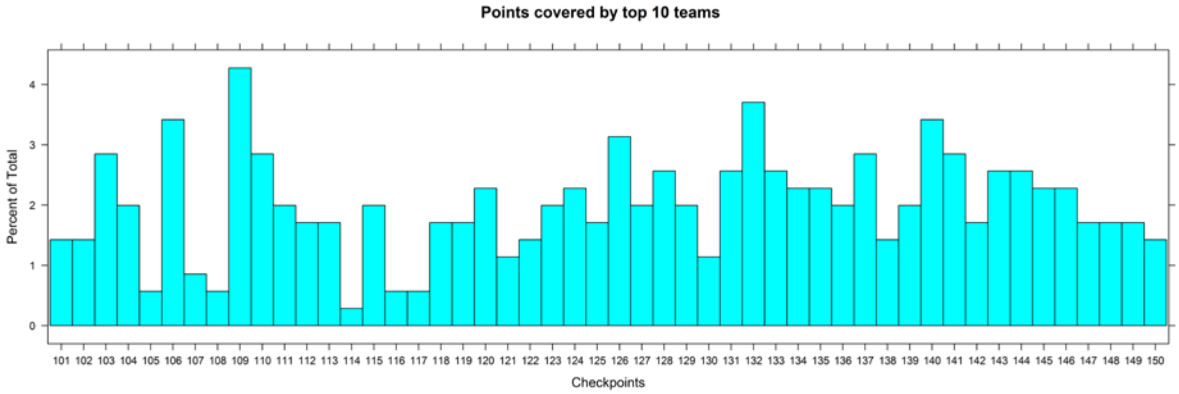
However, as seen from the above figure, membership to the WPOC club does not have much effect the performance of the team. The median for both classes is around the same. However, the scores vary on a much larger range for non-Club members, with the minimum even going negative.

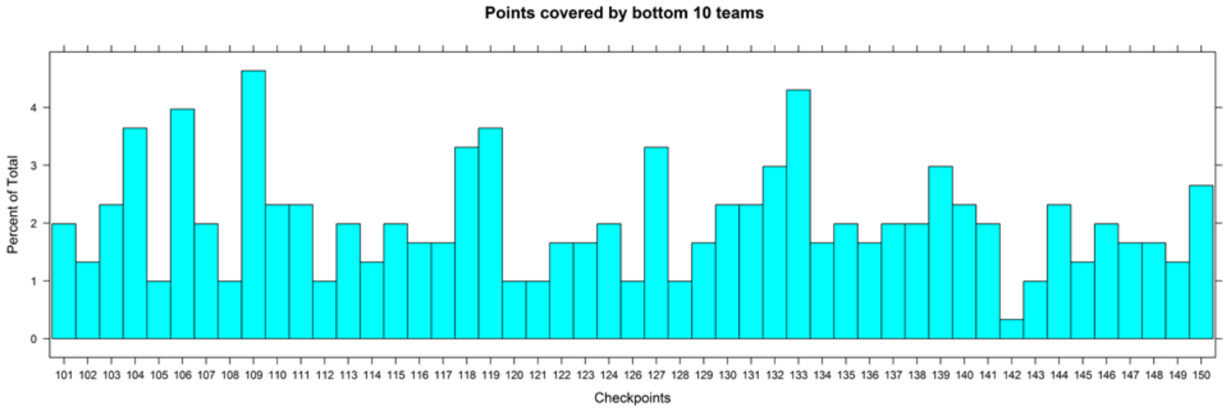
- **Popularity of checkpoints:**
  To see which points were taken most by the teams, we plotted a histogram with each checkpoint's frequency. The final point was removed from this analysis, since all teams crossed it.



Certain points, like 109 & 106, we can compare with the map to see are very close to the start point. So we see a lot of teams going to these points. However, there were also a few surprising points, like 133, that was not very connected nor very close to the starting point, but was still crossed by a lot of teams.

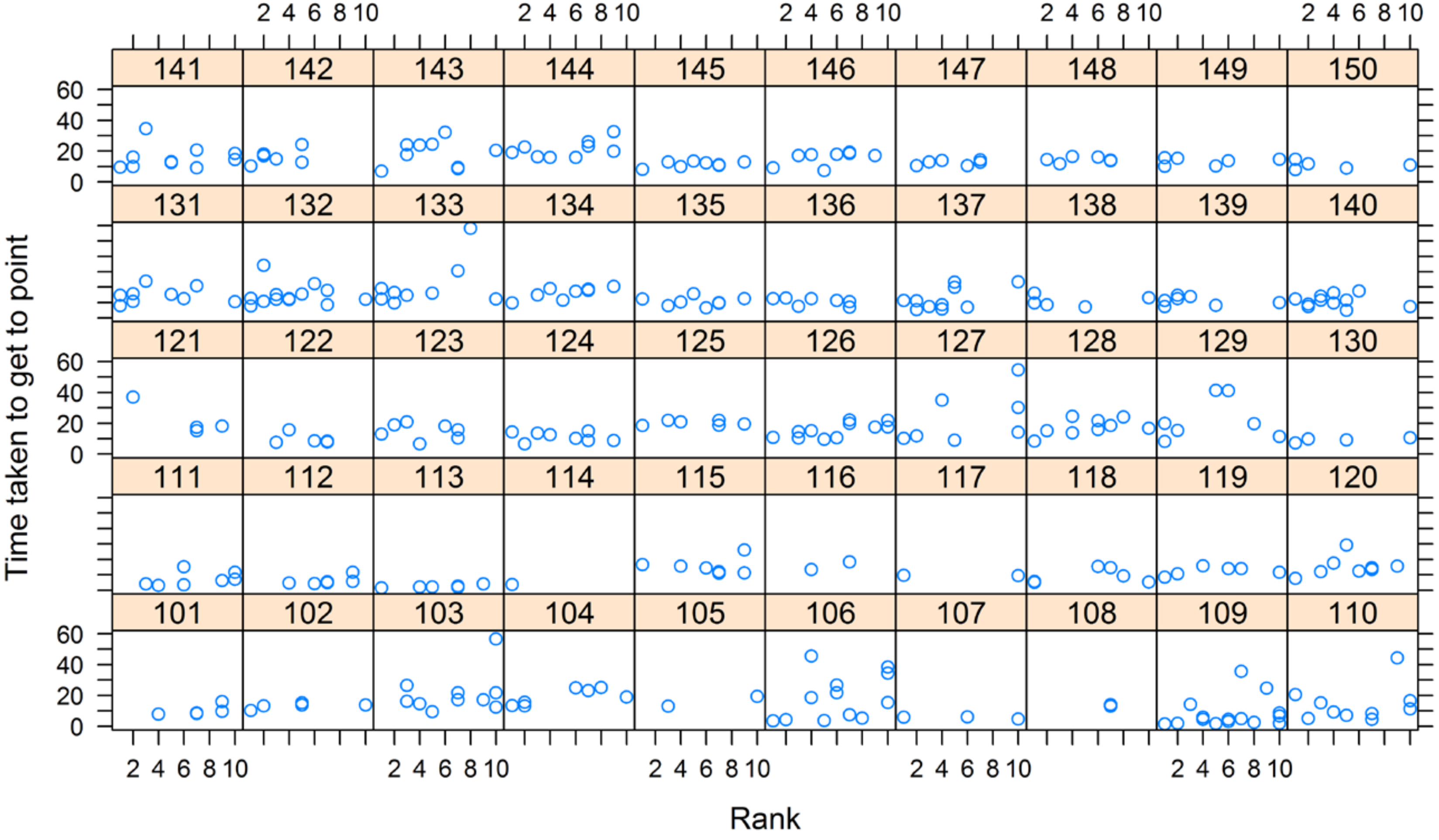
We also wanted to see which points were most frequented by the top-ranked and bottom-ranked teams:





As seen above, some points still overlap, like 109 and 106. Interestingly, a lot more of the top teams ended up going to *point 132*. The topography as seen from the map shows that although on the far-end, this point is well-connected to a proper trail, making it a quicker point to get to.

- **Performance for each checkpoints:**
  While looking into the top 10 teams, we plotted, for all points, the time taken by each team to get to it, w.r.t the team's rank, to see if there was a correlation between how good the team is (their rank) to how fast they were.

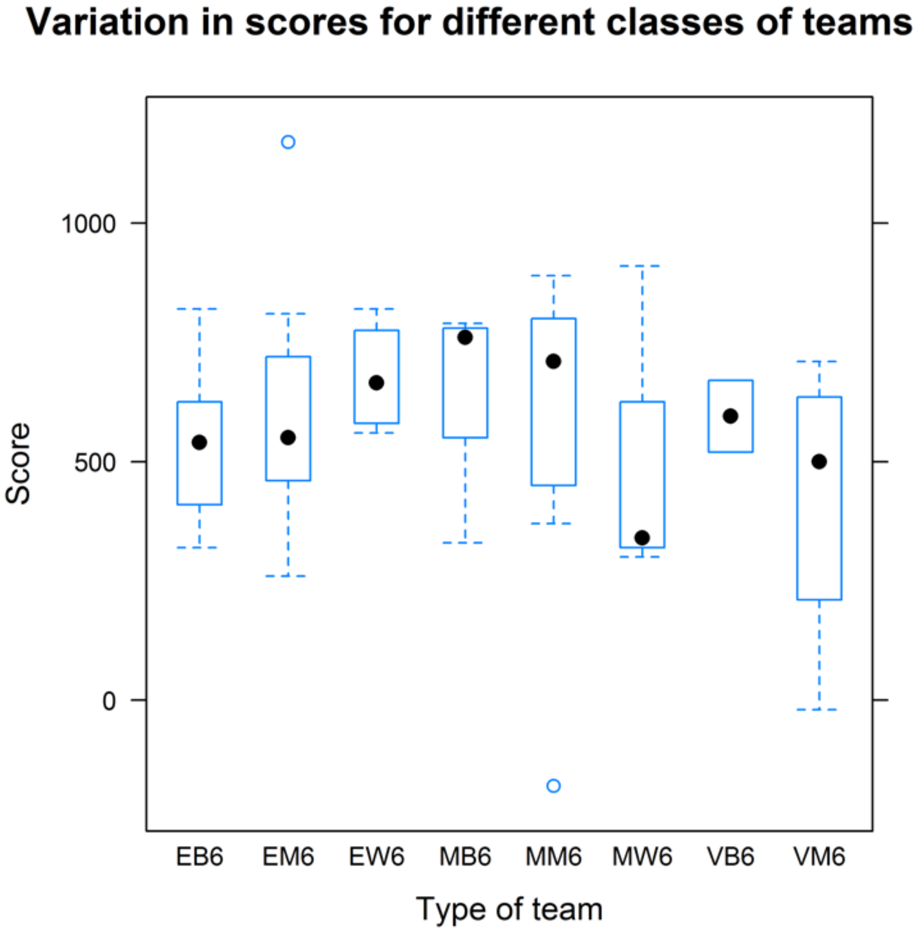

141
142
143
144
145
146
147
148
149
150
131
132
133
134
135
136
137
138
139
140
121
122
123
124
125
126
127
128
129
130
111
112
113
114
115
116
117
118
119
120
101
102
103
104
105
106
107
108
109
110
0
20
40
60
2 4 6 8 10
Time taken to get to point
Rank

For most points, we see that teams take about the same time to get to it. The only point showing a major change in time is **point 133**. We can see positive correlation here; with teams with higher ranks (lower value ranks) taking less time to reach it.
We can also see how certain points like **111, 112** and **113** took very less time to reach. These points were some of the "mystery points", and generally placed very close to the point they were given access from.

- **Performance in different age groups:**
  The teams have been classified not just on gender and time-limits, but also on the age-group of the team members. There are three classes – Elite, Masters & Veterans,

  Elite – age >20
  Masters – age >40
  Veterans – age >55

  Also, due to data sparsity issues, we only looked at 6-hour teams.



As seen above, Masters Women tend to be performing the worst among all 8 categories (there were no Veterans Women teams). Among the Masters, Men & Both performed equally well. We can also see the Elite Women performing just as well.

Model

We used linear regression to predict a team's score based on the given parameters for each team. We also made a separate model to include each of the checkpoints the teams reached to find if certain checkpoints correlated with higher scores. We used LassoCV in scikit-learn to do the model because we had a relatively small dataset and wanted to avoid overfitting with too many parameters. We found that we did not have enough data to make a good model unless we included the checkpoints. We found a RMSE of the test set to be around 125 for the case with checkpoints, compared to an RMSE of 200 when naively just predicting the average score for each team. This is an improvement based on the basic approach, but not great considering we provide complete information for which checkpoints the teams reach.

We found that checkpoints 124, 125, and 149 correlate with higher scores for the 6-hour race. We also found that the coefficient for team size anti-correlates with score, meaning bigger teams scored fewer points.

Please see the Ipython notebook for more information and description of the model. It is provided as a .ipynb, .py, .html, and .pdf file, called 'Clean and Model Data.*'