Final Project – Visualization For Cycling Performance

Kevin Chu
INFO 5602

ABSTRACT

Visualization helps us to find a certain pattern and predict the behavior according to the data. As a cycling enthusiast, I have collected some data and created several visualizations that describe my performance from the previous 8 months as well as analyze the segment's performance of tons of people. Through these two parts of visualization, we can first check if I am making any progress in my riding and then predict my possible performance on each segment.

1 INTRODUCTION

I have been passionate about cycling for a couple of years. One of the reasons I decided to study for my MS degree at CU Boulder is that it's a perfect place for cycling. The gorgeous view and the altitude here just make the training efficient and enjoyable at the same time. Thus, I believe it's a good chance to visualize the data I got from the training to see if there's a difference over time. What's more, I would also collect the data for some famous routes and check if I can use them to find the relation there. Those visualizations should help me know which level of rider I am now, and I would keep using them to track my performance in the future.

I have collected lots of data about my riding and also obtained the routes data from the Strava API (which is somehow more difficult than I thought, because there isn't much information on the internet). With those two datasets, I would try to use ggplot2 to visualize them. It's a powerful package, and I think that finishing this project with it may let me polish my coding skill as well. I would also use tableau to see if I can make something interesting too.

2 RELATED WORK

2.1 WICKHAM, Hadley. Package 'ggplot2'. Create elegant data visualizations using the grammar of graphics.

ggplot2 is an open-source data visualization package for the statistical programming language R. Created by Hadley Wickham in 2005, ggplot2 is an implementation of Leland Wilkinson's Grammar of Graphics—a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers. ggplot2 can serve as a replacement for the base graphics in R and contains a number of defaults for web and print display of common scales. Since 2005, ggplot2 has grown in use to become one of the most popular R packages. It's also the foundation of many graphic applications such as Vega-Lite and Tableau.

To apply ggplot2 in the project, I first studied the basic concept and coding grammar, and I also join the virtual workshop to learn some advanced applications like how to create an interactive chart with it. It's a powerful tool along with several convenient packages such as patchwork, ggforce, and ggraph that let you make the chart clearer and more appealing. A takeaway for me in this class is polishing my R coding skill and the ability to create vis with ggplot2.

2.2 Jeukendrup, Asker E. "Improving cycling performance."

The paper describes the factors that influence our riding and provides the methods for riders to get stronger. We will observe the relation between Power and Heart Rate and analyze the performance accordingly. For now, I am training myself with the courses that were created with the concept mentioned in the paper, while there is not a scientific way to verify if you are getting better and better through the training. To prove that the training course works, I would like to visualize the chart of the mean and even use a statistical method to define if there is any difference between each month.

2.3 Ko, Inseok. "Interactive visualization of healthcare data using tableau."

This paper provides an example to visualize the health data using Tableau which is similar to what I am going to do. Tableau is a tool that can let users create visualization easily by clicking the mouse. You can not only map the data into the x and y-axis but also represent them with color, shape, size, or text. It's convenient but you would have fewer chances to customize your chart. Instead of ggplot2, I would also use Tableau to create some charts as well and compare the pros and cons between those two.

2.4 Sun, Yeran. "Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data."

I would say the most difficult part of this project is how to obtain the data I need, and the paper briefly describes how they get data. The data of my training is easy, the only thing I need to do is to download the csv file from my Garmin webpage. However, extracting data from Strava for segments analysis is at another level. I have to first create an account to require access and then connect to the Strava API to download the data. The difficult part is that there are few sources online showing you how to do it correctly, which makes me spend lots of time researching how to get the dataset. What's more, there is a limitation that you can only require 150 data per day, which makes me only have the data of 1700 riders. I was hoping to obtain over 5000 riders' data so that we may find a clearer pattern, but I will use this dataset to create the vis for the project.

3 DETAILED DESCRIPTION

3.1 Dataset manipulation

Month	Distance	Calories	Avg HR	Max HR	Normalize	Max Avg	Avg Powe	Max Powe
Jul-21	15.09	310	132	168	185	178	128	360
Jul-21	27.85	551	136	150	152	166	144	232
Jul-21	30.1	620	137	164	156	184	146	255
Jul-21	9.11	198	138	181	182	148	128	313
Jul-21	37.84	567	141	160	162	163	158	276
Jul-21	32.06	499	131	144	135	188	133	198
Jul-21	41.59	644	140	154	151	155	149	202
Jul-21	32.03	646	136	159	164	175	144	229
Jul-21	32.29	542	131	143	135	167	132	189
Jul-21	37.28	548	139	152	158	161	155	265
Jul-21	37.08	617	140	164	163	171	149	257
Jul-21	20.13	362	143	163	179	198	159	265

Fig 1. My training performance – including Distance, avg Power, Heart rate, and other results.

rider	SunshineE	Fourmilet	SuperFlag	SunshineE	Fourmilet	SuperFlag	w_kg_rati	w_kg_ratio
TamiEdwa	4222.268	4661.473	1681.5	70	78	28	3.1	3
GlenCain	4527.811	4267.246	1335.7	75	71	22	3.3	3
ElmerCox	4660.462	4089.863	1460.6	78	68	24	3.3	3
RandyWo	4352.935	4530.608	1492.9	73	76	25	3.9	4
BobbiePer	4937.548	4353.255	1597.5	82	73	27	6.2	6
NadineMu	4922.025	4014.817	1597.5	82	67	27	4.5	5
JaimeBroc	4237.782	4411.325	1597.5	71	74	27	3.6	4
MadelineF	4009.834	3675.275	1597.5	67	61	27	2.3	2
AgnesBlal	4497.916	3849.569	1598.7	75	64	27	3.1	3

Fig 2. Dataset for segments analysis – including segments performance and riders' power.

The period of the data is from 2020-Dec to 2021-July, which is roughly half a year. I usually review myself each quarter so the period includes 2 quarters which should let us see if there's variation. After having the data, I still need to manipulate them to get the data we need for visualization. In this step, I exclude some data that are not used for our visualization here such as the weather, the beginning time of the riding, the wind velocity…etc. We could leave them if we are going to analyze the influence of weather and wind, but they are useless in my case. Since I am going to observe the relation between the Heart rate and power, I only left the columns that contain data about them.

For the segments analysis, I categorized the results from seconds into minutes so that I can create a bar chart that describes the distribution for each minute. I also transfer the continuous data, w/kg into ordinal data that set the riders into 6 groups which would be easier to add interaction and filter with the factors.

3.2 Personal analysis & vis

Currently, I use several indexes to trace my training effect. Monthly distance and ascent is the one I use to monitor how much effort I put in each month. Then I would observe my normalized power and heart rate to see if the training works. If I did improve my riding, my power should be higher while the heart rate remains the same (or lower). Last but not least, I will check my 20 min max avg power which should have a similar trend to the normalized power curve.

With that being said, here are the three visualizations I create. The first one (fig.3) shows the aggregate distance and ascent for each month. I set up a goal that I should do 200 km plus 2000m ascent per month, and I use some pop-ups here to highlight the month that didn't reach the goal (i.e. Dec, Feb, and Apr). With that visualization, we can review why I didn't make it. What's more, we can use some interaction to show the reason when people hover on the bar (Studied for a while but don't know how to do with ggplot2 now).

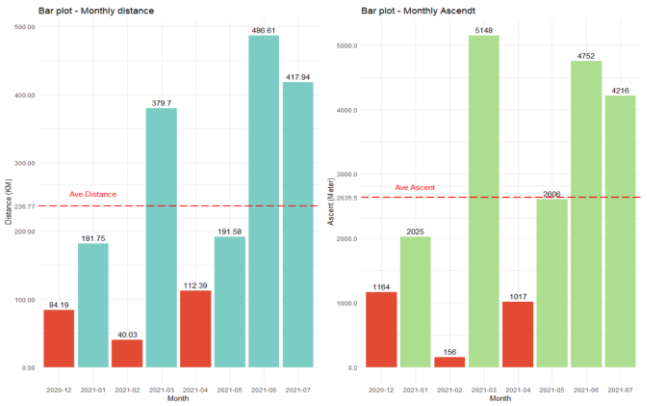

Fig 3. Bar chart with total distance & ascent each month

In my second visualization (fig. 4), I first use boxplots of HR and Power to analyze the distribution for each month, and then I create an area chart that draws the difference variation between HR & Power. In the visualization, we can interpret a few things. First of all, we can see that I have higher HR with lower Power in Dec. However, the Power increased in March and even went up further in June. At the same time, My Heart rate data is almost damping in the same range. As the result, the bottom chart shows the difference between these two and I believe it shows that I did make progress during the first half-year in 2021 and maintained at a certain level in July. To statistically prove it, we could use the ANOVA test to see the difference of the mean, but I am not going to do it in the project.

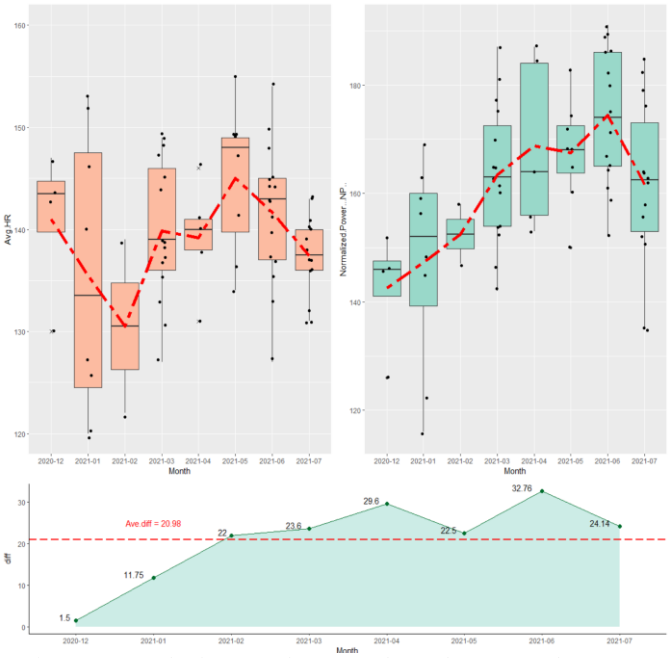

Fig 4. Up: Box plot for HR and Power each month; Bottom: The area chart demonstrates the difference (avg power – avg HR)

In the last visualization (fig. 5), I am trying to use some animation with ggplot2 and extension packages just for fun. It shows the max avg 20 min Power with a line chart and it moves when the time passes. We can see that the maximum data is made in late June and the whole power distribution increases around Feb and Mar. Scatter plot and boxplot may explain the same concept, but I

think the animation of the line chart makes people feel the time better.

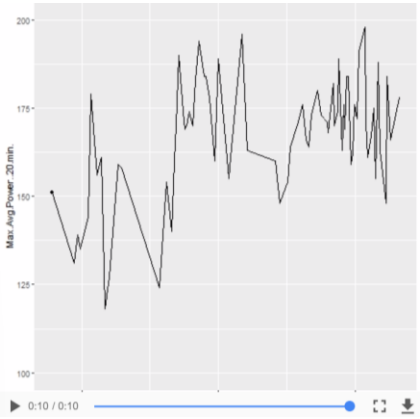

Fig 5. 20 minutes max avg power with animation

3.3 Segments analysis & vis

In segment analysis, I first use a scatter plot (fig. 6) to represent the data from 2 different routes (i.e., "Super Flag" and "Four miles to Gold Hill"), as well as mapping the color with the ordinal power groups to show the riders' level. As you can see, riders tend to have better performance as long as their Power-to-Weight ratio is higher. The whole distribution also shows a relation between 2 segments. Riders usually have better results on the route if they're faster on another one. Thus, I use a box plot (fig. 7) to show the distribution between each minute and predict the possible results according to the chart. From the second chart, I think we can conclude that if you can make it around 30 min on "Super flag", you should be able to finish "Four miles to Gold hill" around 70 minutes, so on and so forth for other groups.

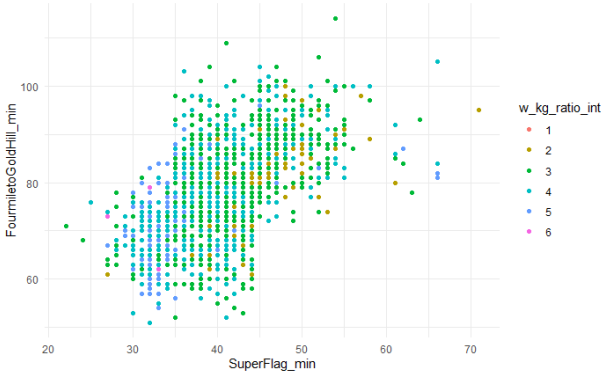

Fig 6. Scatter plot for 2 segments results. Colour mapping with power/kg ratio.

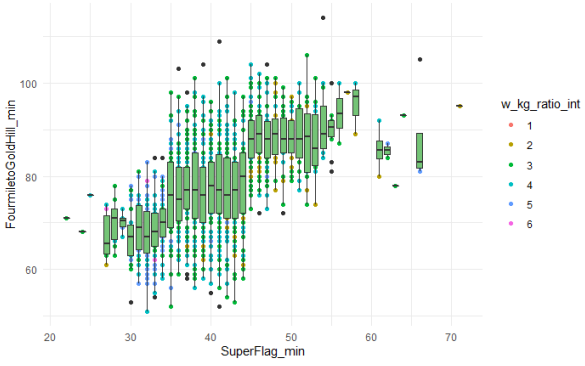

Fig 7. Boxplot overlaps on scatter plot to mark the majority of distribution within each minute.

In fig. 8 and fig. 9, I add some interaction in the visualization. You can filter the data in fig. 8 to show the certain group of Power-to-Weight ratio that you are interested in. In fig. 9, you can hover your mouse on the point, and it will show you the tooltip with the rider's name which is pretty convenient when you want to find out who makes such unbelievable performance. It will also highlight their group of power ratio so you can know which level they are at.

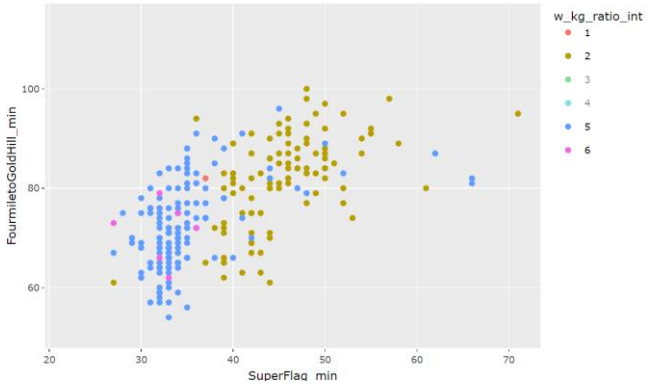

Fig 8. Interaction applied – Filtering

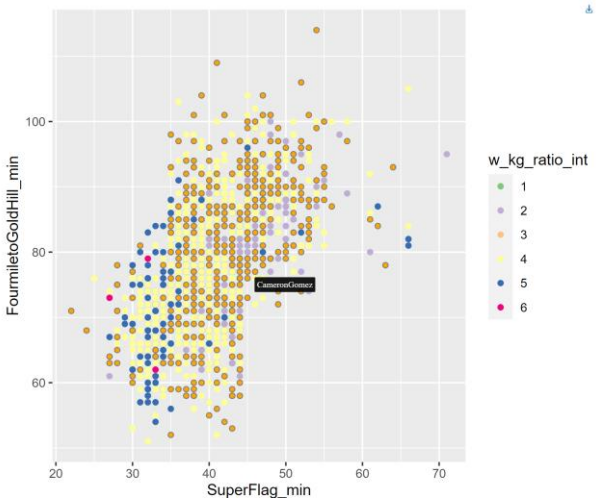

Fig 9. Interaction applied – Showing information when hovering.

3.4 Play around with Tableau for the same dataset

I also try to use Tableau to create some visualizations with the concept I mentioned before. I would show some visualization here and make some simple notes under them. To summarize, I think Tableau is not only easier for people who are not familiar with coding but also a tool that let all of us create visualization more efficiently.

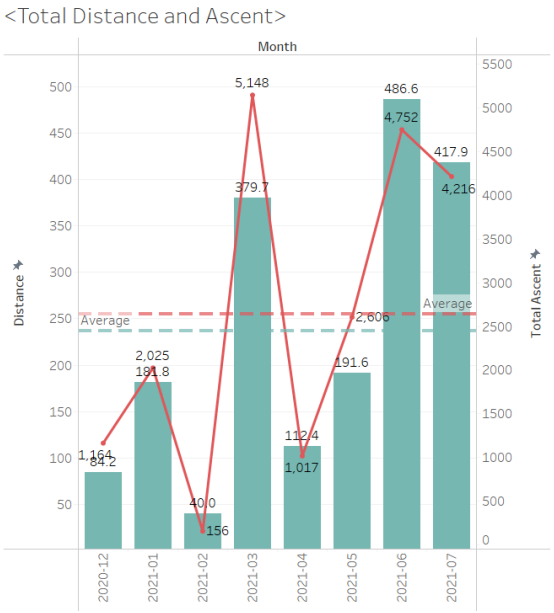

Fig 10. Combine total distance and ascent into one graph with dual-axis.

Fig 11. Use facet to represent 3 segments. Mapping the Power-to-Weight ratio to pie shape so it solved the overlap problem I have when I use scatter plots. Adding mean lines for each group so people know roughly where they would be with the chart. The filter function on the side helps people close the outliers or any category there are interested in.

4 Conclusion

The whole project let me learn a lot of skills. Starting from how to get data and clean them properly for visualization to hands-on practice with ggplot2 and Tableau. I believe the visualization of the training is necessary to monitor your progress, and the segments analysis gives you an idea of how you might perform in those routes. However, I thought I would get more than 4000 data for segments analysis, the limitation from Strava API just makes it impossible to get that number before the due. We should be able to have more precise analysis results if we can get more data and I think I would keep tracking them in the future.

I would also like to discuss the pros and cons between Tableau and ggplot2. In my opinion, ggplot2 is quite flexible that you can almost customize anything you want as long as you know the code or the packages that provide these functions. On the other hand, Tableau provides an efficient way to create the chart as well. The only thing you need to do is dragging variables and map them into corresponding factors such as axis, size, shape, and color. While it sacrifices a little bit of customization, I think it's still powerful enough for most people to make a chart that is clear enough to deliver the meaning behind it. I spent a lot of time finding specific functions with ggplot2, but it turned out I can have similar results with Tableau in just 10 minutes. I like the concept of ggplot2, but I think I will go for Tableau first next time. Both of them are good, and I am glad I spent tons of time studying them for this project.

Other References

[1] ColorBrewer: Color Advice for Maps (colorbrewer2.org)
[2] Welcome | ggplot2 (ggplot2-book.org)
[3] Business Intelligence and Analytics Software (tableau.com)
[4] Make ggplot2 Graphics Interactive • ggiraph package (davidgohel.github.io)
[5] ggplot2 workshop part 1
[6] ggplot2 workshop part 2