CPSC 583 Final Project Report

30002218 | April 14, 2021

1 – Introduction

The Tour de France is an annual 3-week bicycle race that takes place in France primarily as well as surrounding countries. The race dates back to 1903, with a different course route every year. Although there are race maps for the later years, it is not easy to compare courses between years since they are separate. To remedy this problem, this visualization aims to create a way where different years of the Tour can be explored and compared to each other.

This report will cover a description of the dataset, the design process (including initial and refined sketches, as well as the implementation prototypes), the final visualization, and a reflection over the entire process.

2 – Data Description

2.1 – Data Descriptions

https://www.kaggle.com/ralle360/historic-tour-de-france-dataset

The Tour de France is the world's most prestigious and famous bicycle race, consisting of 21 race stages over three weeks. Although it is the most watched sporting event in the world, it can be difficult for a new spectator to understand due to the events focusing on different disciplines and, unlike many other sports, the changing race course from year to year. By visualizing the historical data of the Tour de France, it can provide a greater appreciation for the sport and how it has evolved over time.

I will be using a dataset found on www.kaggle.com, generated by RasmusFiskerBang. It contains race stage information from the first Tour in 1903 up to the 2017 Tour de France. The information was gathered from Wikipedia, and there are 2336 rows and 8 columns.

Each row represents a single stage in each year of the Tour de France; the columns provide details about that stage, including information such as:

Stage - The stage number in the Tour of that year
Date - The date that the stage occurred on
Distance - The number of kilometres covered by the stage
Origin - The name of the city where the stage starts in
Destination - The name of the city where the stage ends in
Type - The main type of the stage (eg. flat, time trial, mountain, etc.)
Winner - The name of the winner of the stage
Winner_Country - The country of origin of the stage winner

2.2 Pros and Cons of Data Sets

The dataset contains a variety of types of data, primarily nominal but some quantitative as well. One notable missing metric is the time of each winner, which could have been used to calculate things like pace. Also, the dataset only includes up to the 2017 tour, but the rows missing can easily be added manually due to the small size. The source of the data is from Wikipedia and therefore may not be as accurate as data from the official Tour de France website, but the convenience of being in .csv format allows it to be more easily compared and cross-referenced with similar datasets for errors, in addition to creating visualizations. The large number of rows can help identify outliers if there are any minor inaccuracies, and is therefore appropriate to be used for viewing the general trends.

2.3 Data Set Decision

I will be using the entire dataset `stages_TDF.csv`, as there are many interesting ways to represent the different types of information. For example, the average timeframe that each Tour occurs within the year displayed over a calendar, or the most common cities to host stage starts/ends and their most recent year that they hosted shown on a map. As an avid cyclist who found the Tour de France confusing to follow at first, I hope that creating interactive visualizations using this data can help demystify and provide insight on the history of the Tour.

3 - Data Description

3.1 – Sketch-able Data Subsets

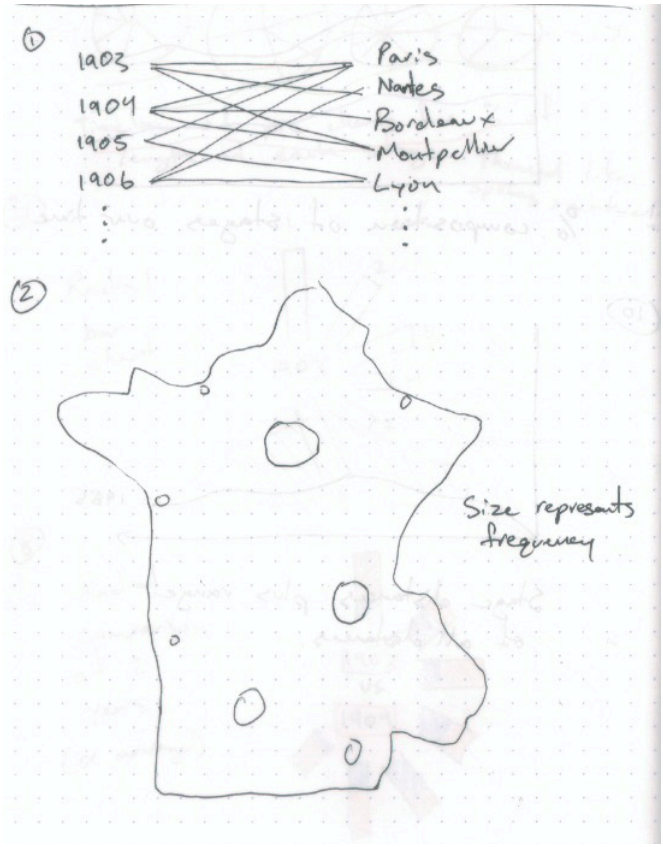
Subset 1: Stage Number, Distance, and Date
The focus for subset 1 was to show how stage distances of the Tour de France changed over time. This subset contains over 2000 rows and 3 columns, and every year of the Tour in this dataset has data in this subset. Because not every year of the Tour has the same number of stages (especially the earlier years), there is "missing data". The varying number of stages, as well as the changing distance covered by each stage, can provide insight to the difficulty of this sporting event and how it has changed over the years. However, this subset is still too large to sketch in its entirety, so sketches were based on a smaller subset. Certain sketches showed data for a few years at a time, or a limited selection of stages. I also explored the idea of visualizing a single year, and comparing it to the rest of the data through averages, minimum/maximum values, etc.

Subset 2: Origin/Destination City, and Date
Subset 2 includes the origin and destination cities, and dates. This subset contains data for every single tour (recorded in this dataset). Although date is included, there is less emphasis on how they change over time and more on the year that they were involved. Being a host city can bring prestige and an economic boom to the area, and visualizations of this subset can be helpful for further analysis.
This subset is of similar size to Subset 1, but due to the categorical nature of the data I found it easier to include the entire subset in my sketches. Some sketches visualize only one year in comparison to the data, and some sketches do not include date at all and focuses on the frequency of popular cities included in the Tour.

3.2 – Design Direction

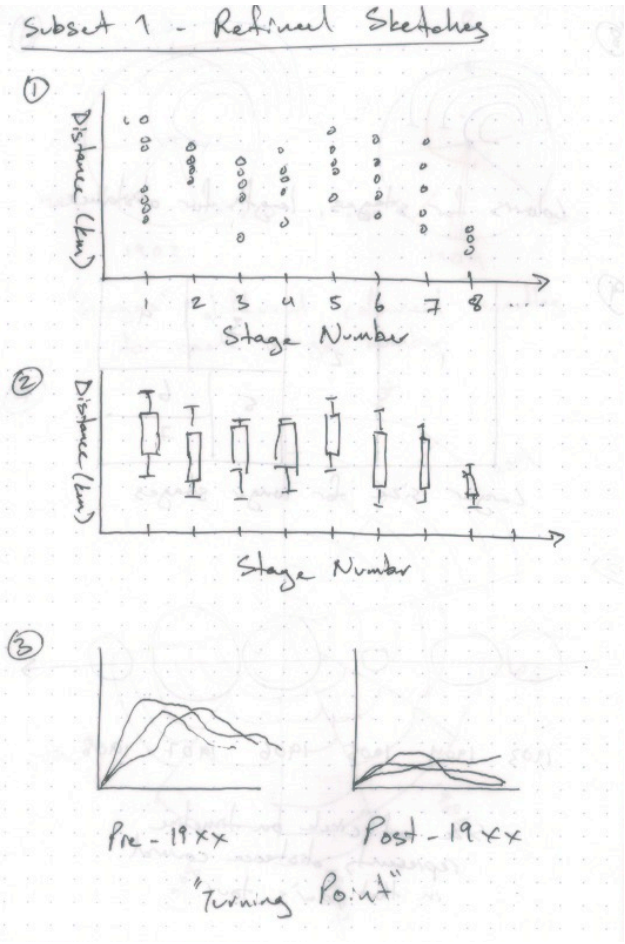
3.2.1 – First Sketches

Not all sketches visualize the same information for the same dataset. For instance, one sketch for Subset 2 represents the overall number of times that a city hosted a stage of the Tour de France, but another sketch for the same subset shows the specific year that the city was an origin or destination city. As expected, the frequency visualizations are generally more compact and more easily represented the entire subset of data compared to the specific year-to-year visualization, which only include a few years in the sketch. Even though some of the visualizations with a limited scope are effective, when they are scaled to include the entire subset it can be overwhelming to comprehend because of the large number of small multiples.



3.2.2 – Variations

The sketches reflect the more common and intuitive ways to represent the data. For example, distances over time are commonly expressed as line or bar charts, and cities are intuitively visualized on a map. There are novel ideas introduced in the initial sketches, but the traditional visualizations proved to be more effective and expressive, and were more likely to be chosen to be developed further.



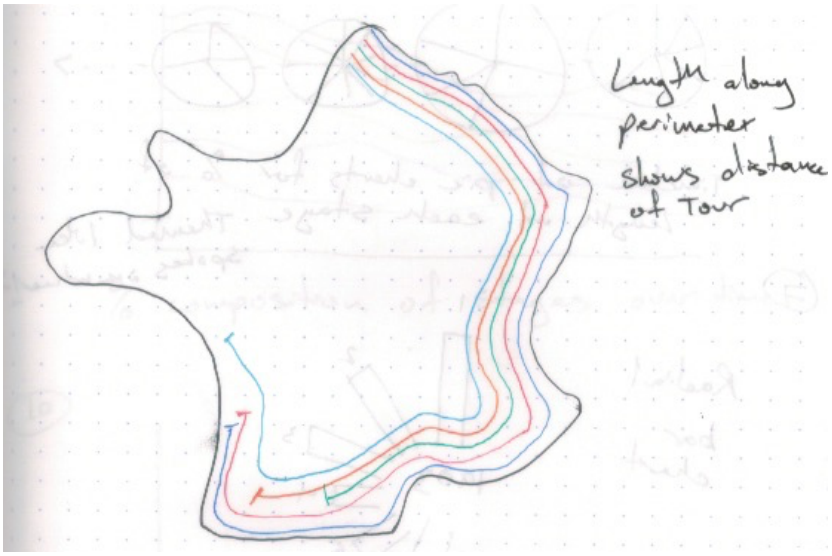
3.3 – Process

The two subsets were created with a clear goal in mind, which drove the results of the sketches. I approached the visualizations first with immediate ideas on how the information could be represented intuitively. Next, I explored ideas that were commonly used to show similar data. Finally, if I did not have ten initial sketches yet, I would try visualizations that were not traditionally used if they could represent the data.

Since some of the sketches are based on a subset of the subset, ideas that could not scale to represent the entire dataset were not explored further. To refine the sketches, I

would modify ideas from the initial sketches by changing the type of encoding, or apply the same visualization to a different subset.

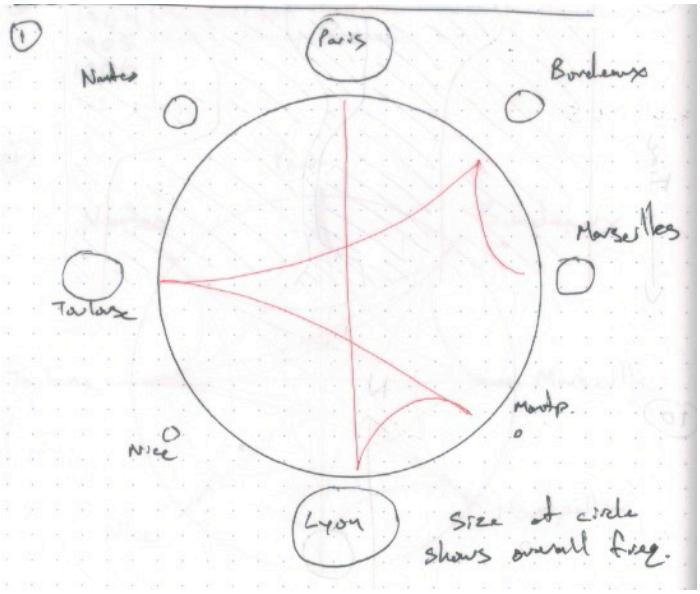
Experimenting with colour made it apparent that there were not enough distinguishable colours to represent the big range of variables. As an example, there are at most 21 stages but there are less than 21 variations of colour that can be easily distinguished. Rather than assigning a colour to every single value, I used colour to highlight certain information from the rest or to represent a range of values.



As well, the visualizations focused only on the subset and may not work as well if the subset was expanded to include other variables. The refined sketches have ideas that are specialized to represent that information, and introducing another column to the subset is difficult.

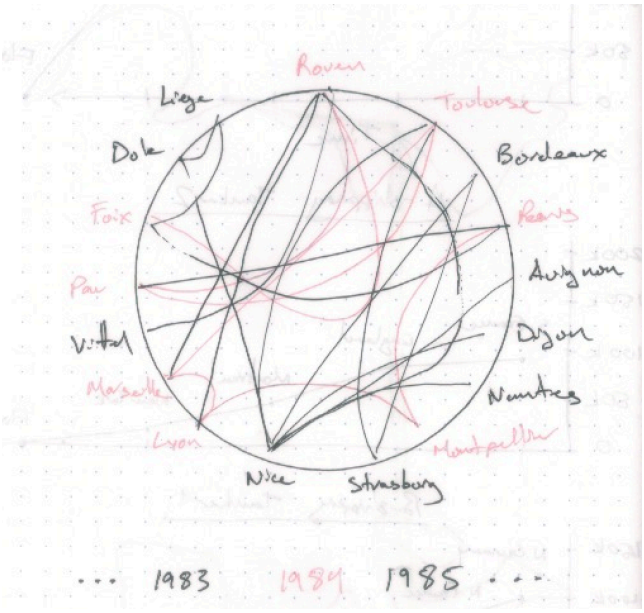
From the sketches done, showing stage numbers and distances over years is likely best with a line or bar chart. Representing the large amount of data all at once in an easily comparable and comprehensible manner is a challenging task, and this traditional visualization is a compact way of accomplishing it. For Subset 2, cities are commonly shown on a map, and by leveraging this familiarity an effective visualization can be created.

On the other hand, non-traditional visualizations are not without their strengths. I found that the circle representation for refined Subset 2 sketches was particularly effective, in that it could show both overall data as well as a specific year's information at the same time. Furthermore, it can be easily adapted to include even more information, such as the order in which cities are visited of a given year's Tour de France.



3.4 – General Design Direction

The circular visualization idea as sketched in P2 was chosen as the main design inspiration for this project component. In the original P2 sketch, year information as well as origin and destination locations were shown, and stage distance has been included for this refined design.



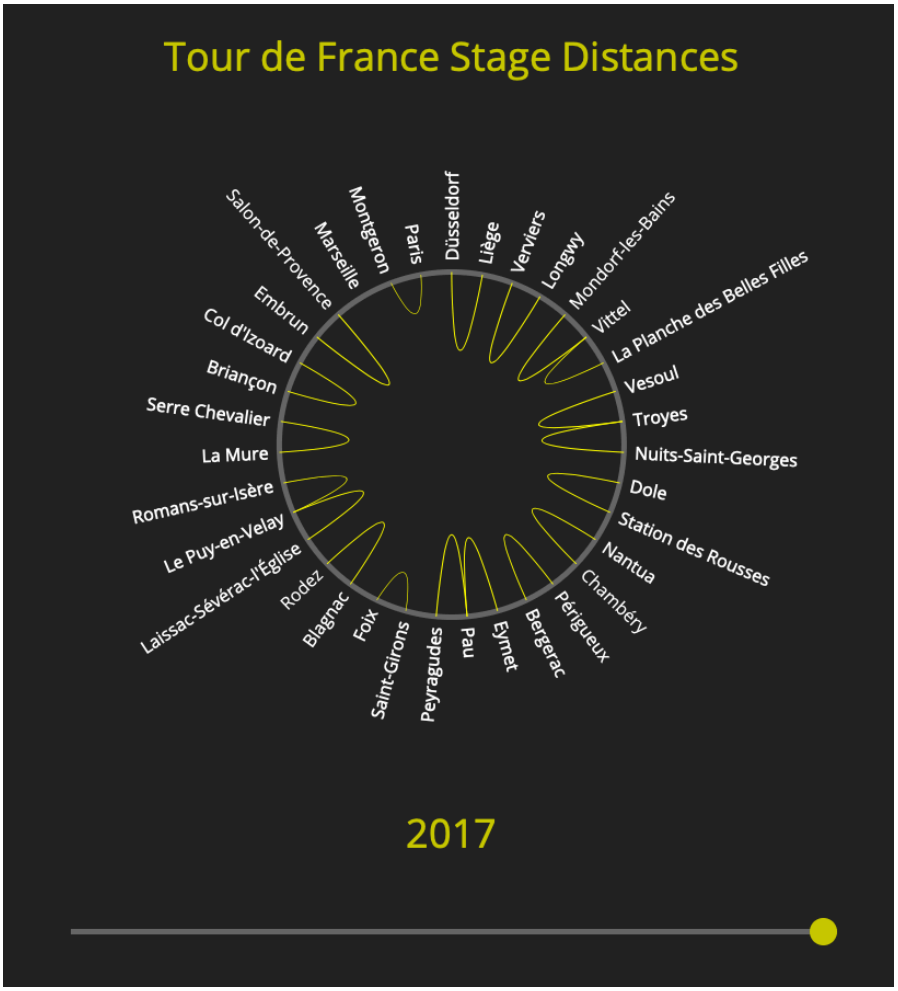
Locations of Tour de France stages are distributed around the circumference of the circle, and for each race stage, a line connects the origin location with the destination. The curvature of the line varies with the distance, with longer stages curving towards the centre of the circle before reaching the destination node on the circumference, and shorter stages represented with a line that is more straight.
A year selector below the visualization allows a specific year's race stage lines to be highlighted in colour among the many other stages, as well as the locations involved in that year's Tour de France. By tracing the path that is created by the highlighted lines, one can follow the order of locations in the Tour.

3.5 – Prototyping Variations

The following design variations were implemented using D3.js, and can be found at https://jordanmklee.github.io/CPSC583/P3_Implementation_Representation_and_Presentation/src/index.html

3.5.1 – Variation A

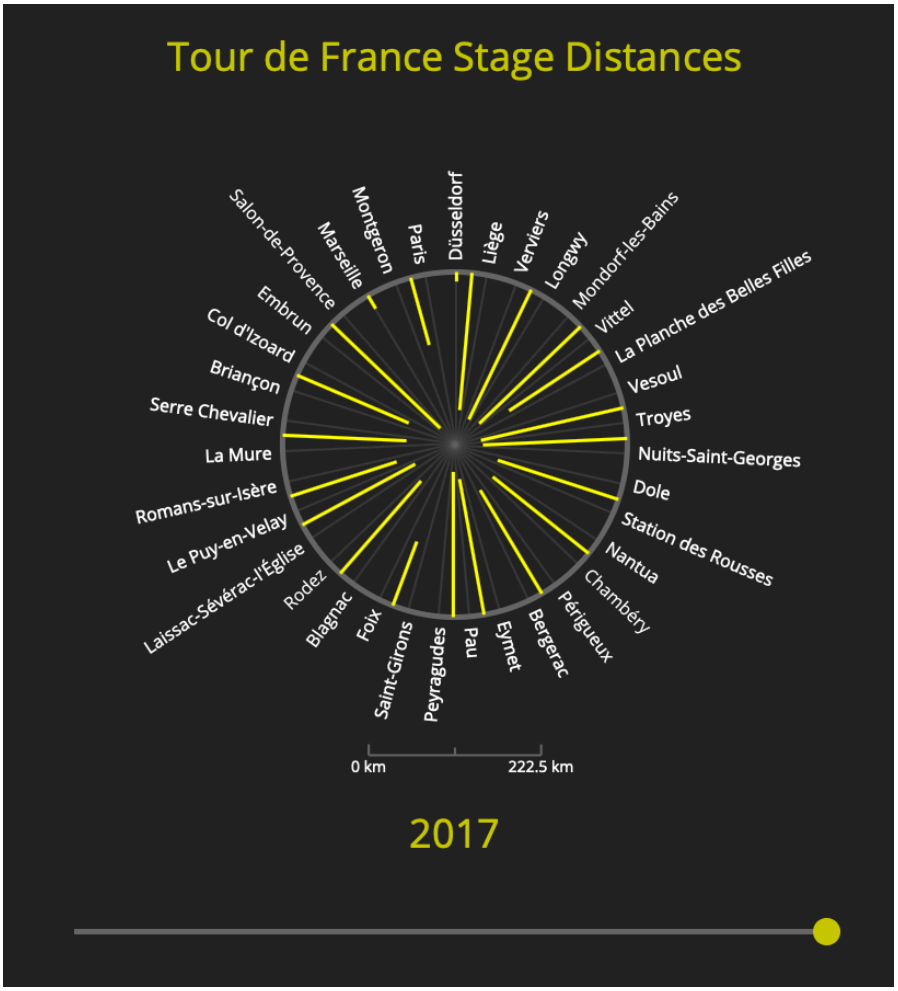
https://jordanmklee.github.io/CPSC583/P3_Implementation_Representation_and_Presentation/src/paths.html



Locations are labelled around the circle and stages are represented as a curved line between two locations. The stage locations are arranged in order of appearance in each year, selected with a slider under the visualization, which produces a visual with minimal overlapping lines since a stage's origin and destination labels are adjacent to one another. Stage distance is represented by the length of the curved line between two locations. Clock-face conventions are followed: the first stage of that year's Tour is located at the 12 o'clock (the intuitive "start" of a circle), with subsequent stages arranged in clockwise.

3.5.2 – Variation B

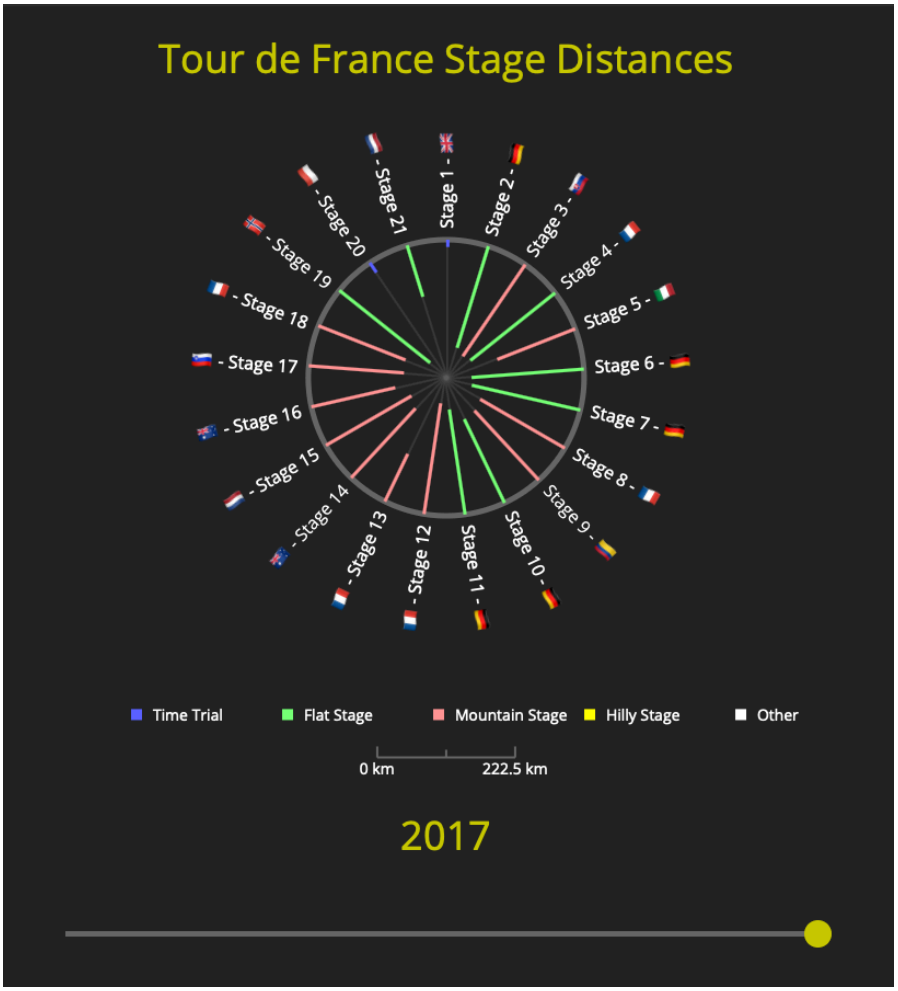
https://jordanmklee.github.io/CPSC583/P3_Implementation_Representation_and_Presentation/src/spokes1.html



A bicycle wheel themed visualization in evolved from Variation 1. Bars (in the style of spokes) are used instead of paths, where the length of each bar represents that stage distance. A distance scale is added to improve readability and context for bar length. Each bar is located in the same place as the curved line in the previous variation; ie. the bar representing a stage and its distance is located between the origin and destination location label.

3.5.3 – Variation C

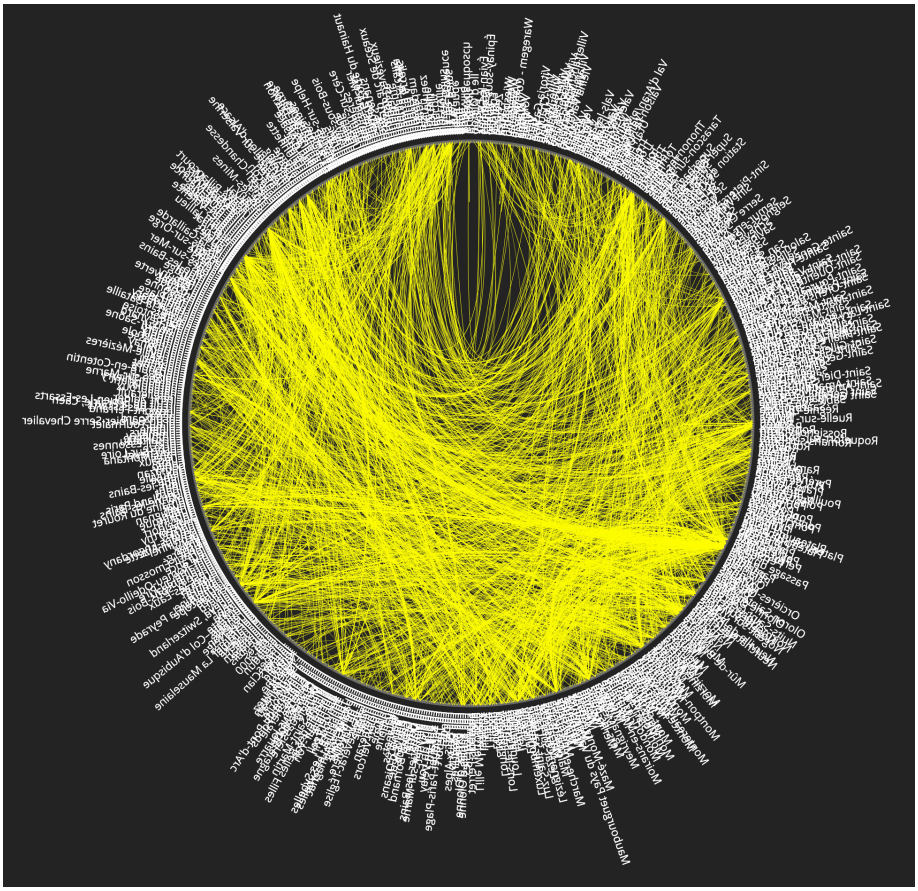
https://jordanmklee.github.io/CPSC583/P3_Implementation_Representation_and_Presentation/src/spokes2.html



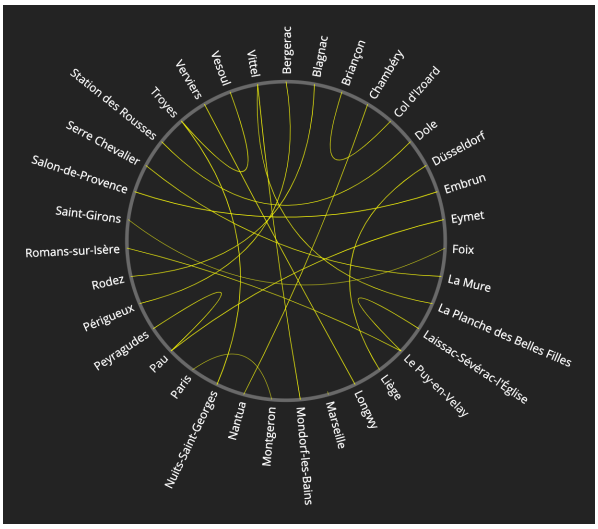
An improved version of Variation 2 with the labels around the circumference representing stage numbers rather than locations, producing a more uniform distribution of "spokes" and improved aesthetics. Origin/destination location information for each stage was omitted from the labels due to readability issues, but was replaced with a flag denoting the country of the stage winner. Also, each spoke is encoded with the stage type using colours loosely inspired by the tour jersey colours (eg. red for mountain stages; red polka dot jersey for leader in the mountains competition), explained with a legend below.

3.6 – Implementation Process

An early implementation showed that there were over 700 unique locations, and, if placed along the circumference of the circle as sketched, the location labels would overlap one another and be near impossible to read. As well, the lines between the locations were much too dense and equally as unreadable. Therefore, following implementations focused on the data for one year of the Tour de France at a time, which consisted of a much smaller number of locations; an average of 21 locations visualized at one time.



Initially, arranging the locations in alphabetical order seemed logical. However, this resulted in a confusing mess of lines connecting stage origin and destinations. The path created by the stage lines was not easily comprehensible, especially with years of the Tour with more locations.



After implementing a visualization with locations arranged in order of occurrence (Variation 1), it produced a wheel-like appearance that was unintended with the sketches, and inspired the bicycle- wheel-themed visualizations of the next two variations.

One weakness of Variation 1 was that the curvature of the lines made it complicated to comprehend and compare distances, so Variation 2 improved upon this by using a straight line along with a scale indicating the largest distance visualized and how long of a line would represent it.

Variation 2 used the same labels and ordering as Variation 1, and due to an assumption that locations were not revisited, the implementation showed that some stages were visualized out of order. To remedy this, Variation 3's labels show stage number in order instead of locations to enforce order and a uniform distribution.

3.7 – Final Static Design

The variations from section 3.5 were developed in an iterative manner from the selected sketch, but with interaction, Variation A can be developed in a different manner. Variation B and C present an interesting design direction, but are more visually complicated, so therefore this final design will be based upon Variation A.

Between prototype variations, more columns of data were represented in the visualization. Variation A only contained origin/destination locations and stage distance, but after two iterations stage number, type of stage, and the country of the stage winner were also encoded. However, the two columns of information left (date, and winner) became a challenge due to the visualization becoming more and more crowded. With interaction, the remaining data can be incorporated into the visualization.

One possible interaction is to have a detailed view when a stage location is clicked on. A separate display would appear, and list pertinent information relating to that stage. The date and winner of the stage could be included in this window, and this can be extended to encode the data in a map, or information not included in the dataset.

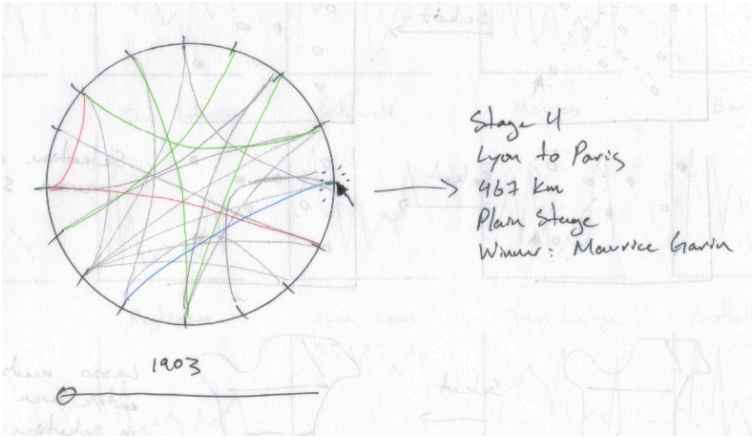
Filtering label information is another way to include the entire dataset without overcrowding the visualization. Checkboxes can be used to select the information that the user is interested, for example choosing to display date and stage number in the spoke labels and omitting the winner as well as country. Filtering stages based on date is already implemented in the form of a slider offering discrete year selections.

Reconfiguring the stage paths into a bar chart could be an interesting way to directly compare stage distances. In the current form, stage distance is not clearly represented and by reconfiguration it could be shown better.

One incompatible interaction would be connection (brushing and linking). Locations are arbitrarily placed along the circumference of the circle, and selecting adjacent lines for comparison is not meaningful.

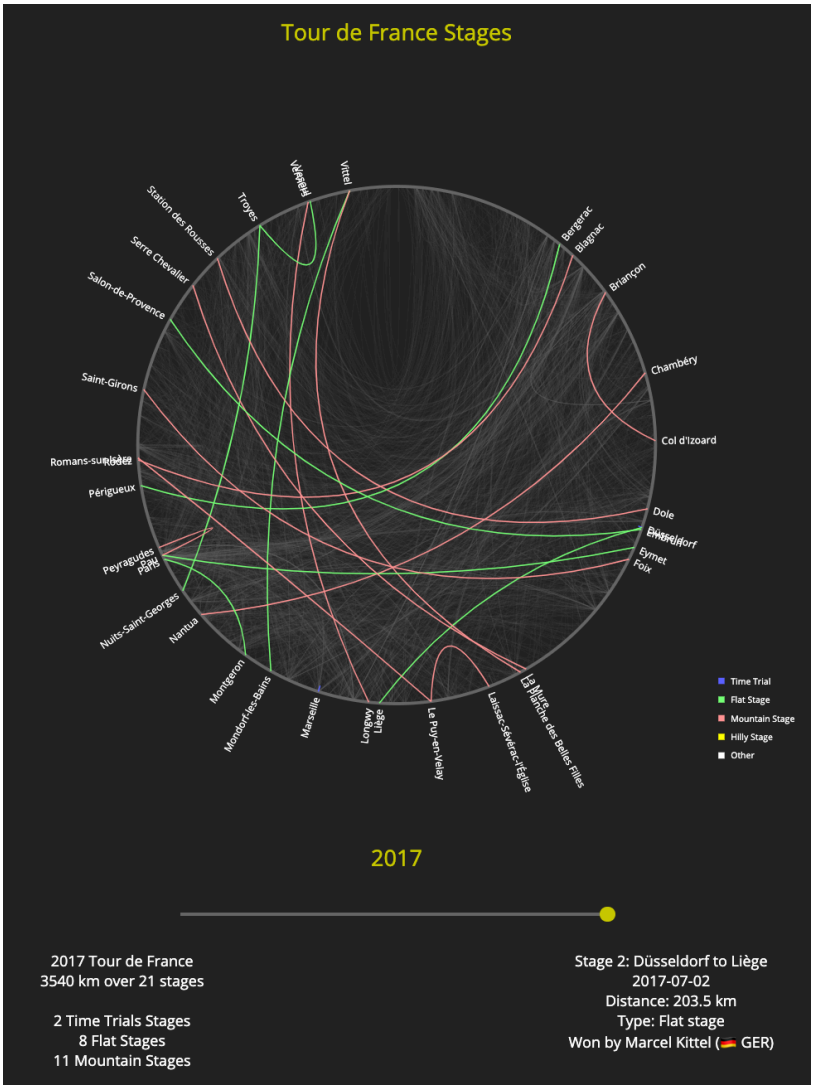
3.8 – Prototyping Interaction(s)

The final implementation will incorporate an overview of every stage in the dataset, shown as the dark grey lines in the background. This gives the user a rough idea of which locations are most popular between years of the Tour. Stages are highlighted if they belong to the selected year, and the location labels only appear when there is a stage going to or from it. The stages are coloured differently to denote the type of stage it is (eg. mountain, plain, etc.). The rest of the stages can be filtered by using the slider to select another year. In addition to filtering, locations can be selected by clicking on them to show secondary information, such as distance, stage type, and winner.



4 – Final Implemented Visualization

Live Demo: https://jordanmklee.github.io/CPSC583/P4_Interaction/src/index.html



The final implemented visualization offers an overview on which locations are most commonly host cities of the Tour de France, in the background as darkened lines. The highlighted lines represent the selected year's route, and are colour coordinated. This allows the user to see the compositions of the types of stages. Also, a secondary display is added to show the selected year's race statistics. By using the colours and selecting different years, one can see that in the early years of the Tour, there were far fewer stages compared to the modern Tour. As well, total distance covered in a year of a Tour peaked in the 1920's at around 5000 km.

4.1 – Process Reflection

Initially, it made sense to continue along the same design direction from the variations in Section 3.5 since they were developed iteratively to solve the problems of the previous variation, but considering interaction it was possible to revisit an older iteration and take it in a different direction.

Instead of presenting all the information at once like in the iterated variations, some of the data could be hidden behind interaction so that the user is not overwhelmed. The previous idea of showing every row of data at once was impossible due to the large number of locations, but filtering allows the visualization to be much more readable and comprehensible.

5 – Discussion

The original idea was to present the data on a map, but due to the large number of unique locations in the dataset, and a lack of geographic location data (only having location names, instead of latitude/longitude), it was not feasible to implement. Given a geojson file with the coordinates, a map visualization would have been possible and could much better represent distances between locations.

Transitions may have been interesting to explore, where the selected year's race could be drawn out in order of locations so that the stage numbers could be interpreted from the animation.

Different encodings for secondary detail displays could present the same information in a much more visually appealing and readable manner. For example, showing the year's race distance in a bar chart, along with the composition of stage type.

As for the current implementation, some of the conclusions that could be drawn were the increase in number of stages from the early years of the Tour compared to modern races. Most years of the Tour are composed primarily of flat and mountain stages, except for 1927 and 1928 which were dominated by time trials (1927 being the first year that time trials were introduced). The visualization also shows that the early races had much more flat stages compared to mountain stages, and have since transitioned to more of a 50/50 split.

A shortcoming of the visualization is that it is still difficult to compare two different years of the Tour. It does a good job at allowing the user to explore the different Tours, but to compare them it is still necessary to flip back and forth. In a further iteration, perhaps multiple Tours could be visualized at the same time.

6 – Conclusion

This implementation accomplishes the goal of creating a visualization where different years of the Tour and details of each stage can be easily explored. It follows the visual information seeking mantra introduced in class: overview first, (zoom and) filter, then details on demand. A user can gain an overview of the Tour's history, and filter out a specific year to see the host locations involved in that year as well as statistics like total distance, and number of each type of stage. It is then possible to gain details on a part of the filtered data, by selecting a single stage to view the origin/destination locations, distance, type, winner, and the winner's country.