A BRIEF LOOK AT THE HISTORIC

Tour de France

Changes to the race, the sport, and it's major players

Why this data set?

"Imagine you are working for a business and your boss doesn't understand data." – Capstone I Prompt

When I read this line in our prompt for this assignment I knew what data I wanted to work with immediately. As I have been completing this bootcamp I have also been working full time in the bicycle industry. A daily occurance in my current position is running SQL queries for my boss and coworkers who do not understand how they work and what we can do with them. I have often been completing lectures and assignments in the breakroom with my coworkers looking over my shoulder and wondering what I was up to. I decided to pick a data set my coworkers would be familiar with to run a full analysis on in part to take this back to my job and show off just what I have learned. My boss might not be able to understand all the steps I took but maybe after this, he will be able to better understand what I have been hunched over my laptop working on during lunch.

Some Background History

The Tour, 1903 to 2017

1903

The first Tour de France is organized to increase newspaper sales. It is the longest of similar races of it's time., only 24 riders completed the course.

1906-1912

The stage and point system is developed as the Tour continues to expand.

1940-1946

World War 2 and the subsequent closing of the organizing newspaper threaten to end the Tour de France for good.

1967

Rider strike leads to limits on stage distances, number of stages, adding rest days, and requiring drug tests.

2017

The last race covered in the data. In the 104th year of the Tour, Chris Froome takes the yellow jersey.

The Data

Untouched CSV file

These are the first 10 entries in our file. We can see it's sorted most recent to oldest race. We can also see the keys and generally how the row data is formatted.

Keys:

- Stage
- Date
- Distance
- Origin
- Destination
- Type
- Winner
- Winner Country

	Stage	Date	Distance	Origin	Destination	Type	Winner	Winner_Country
0	1	2017-07-01	14.0	Düsseldorf	Düsseldorf	Individual time trial	Geraint Thomas	GBR
1	2	2017-07-02	203.5	Düsseldorf	Liège	Flat stage	Marcel Kittel	GER
2	3	2017-07-03	212.5	Verviers	Longwy	Medium mountain stage	Peter Sagan	SVK
3	4	2017-07-04	207.5	Mondorf-les-Bains	Vittel	Flat stage	Arnaud Démare	FRA
4	5	2017-07-05	160.5	Vittel	La Planche des Belles Filles	Medium mountain stage	Fabio Aru	ITA
5	6	2017-07-06	216.0	Vesoul	Troyes	Flat stage	Marcel Kittel	GER
6	7	2017-07-07	213.5	Troyes	Nuits-Saint-Georges	Flat stage	Marcel Kittel	GER
7	8	2017-07-08	187.5	Dole	Station des Rousses	Medium mountain stage	Lilian Calmejane	FRA
8	9	2017-07-09	181.5	Nantua	Chambéry	High mountain stage	Rigoberto Urán	COL
9	10	2017-07-11	178.0	Périgueux	Bergerac	Flat stage	Marcel Kittel	GER
10	11	2017-07-12	203.5	Eymet	Pau	Flat stage	Marcel Kittel	GER

Where to start

Checking for Null/NA data values

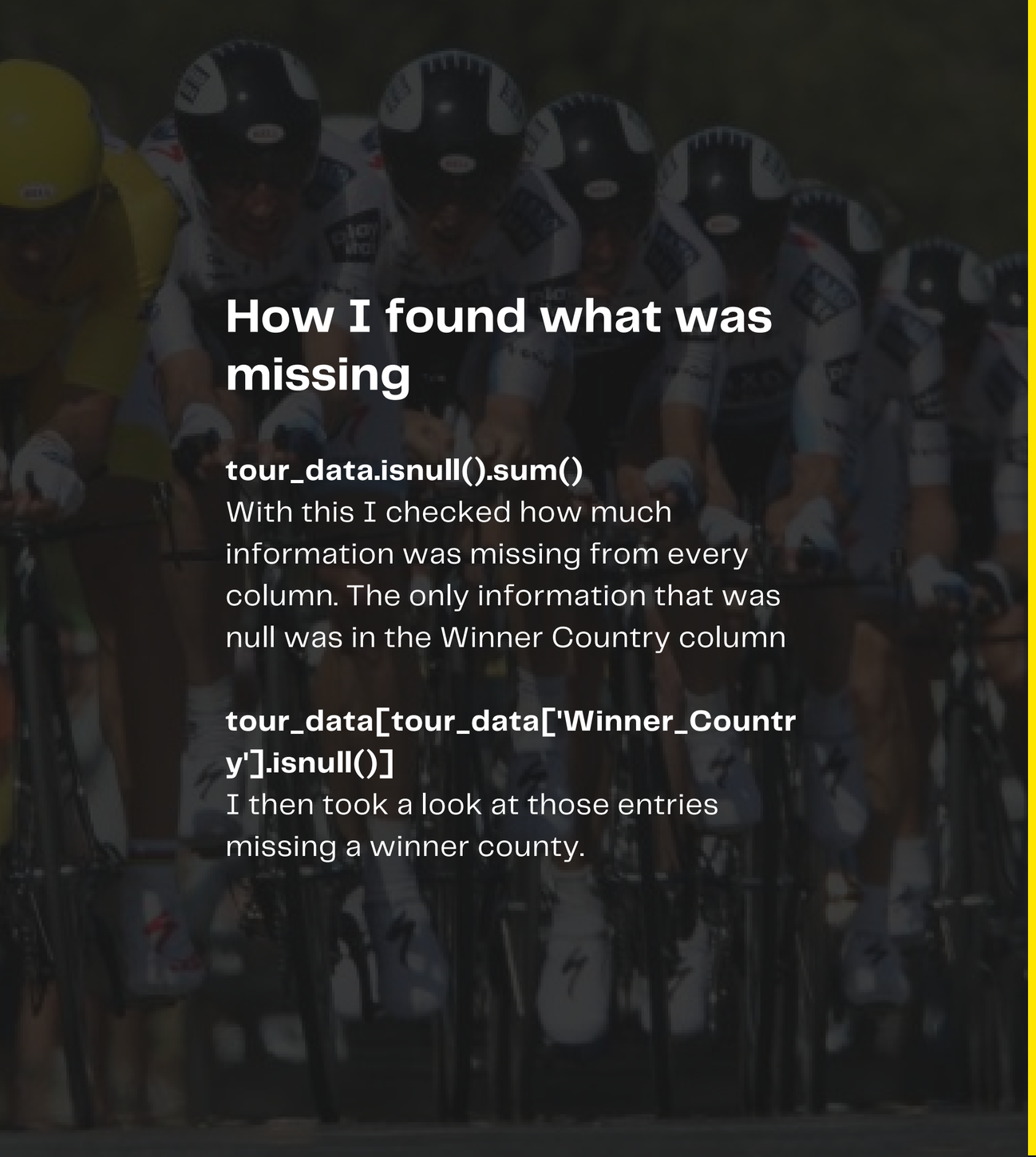
What is missing and why?

Formatting our existing data

Making our information more usable and standardized

Reassess and ask questions

Now that we have a better look where would we like to go



How I found what was missing

tour_data.isnull().sum()
With this I checked how much information was missing from every column. The only information that was null was in the Winner Country column

tour_data[tour_data['Winner_Country'].isnull()]
I then took a look at those entries missing a winner county.

What was missing?

Nearly all of the data missing was the winner country in team time trial stages. This is understandable, teams are international and we can't average out the country between them. I decided to leave these be and make note for when I ran analyses.
The other missing countries were for stages that were listed but were cancelled, therefore no winner exists. I also decided to leave these be and note it instead.

Formatting Methods

Quick early adjustments to make the data more usable.

Date to Year: Added a column for year, data type integer so I can more easily sort by year

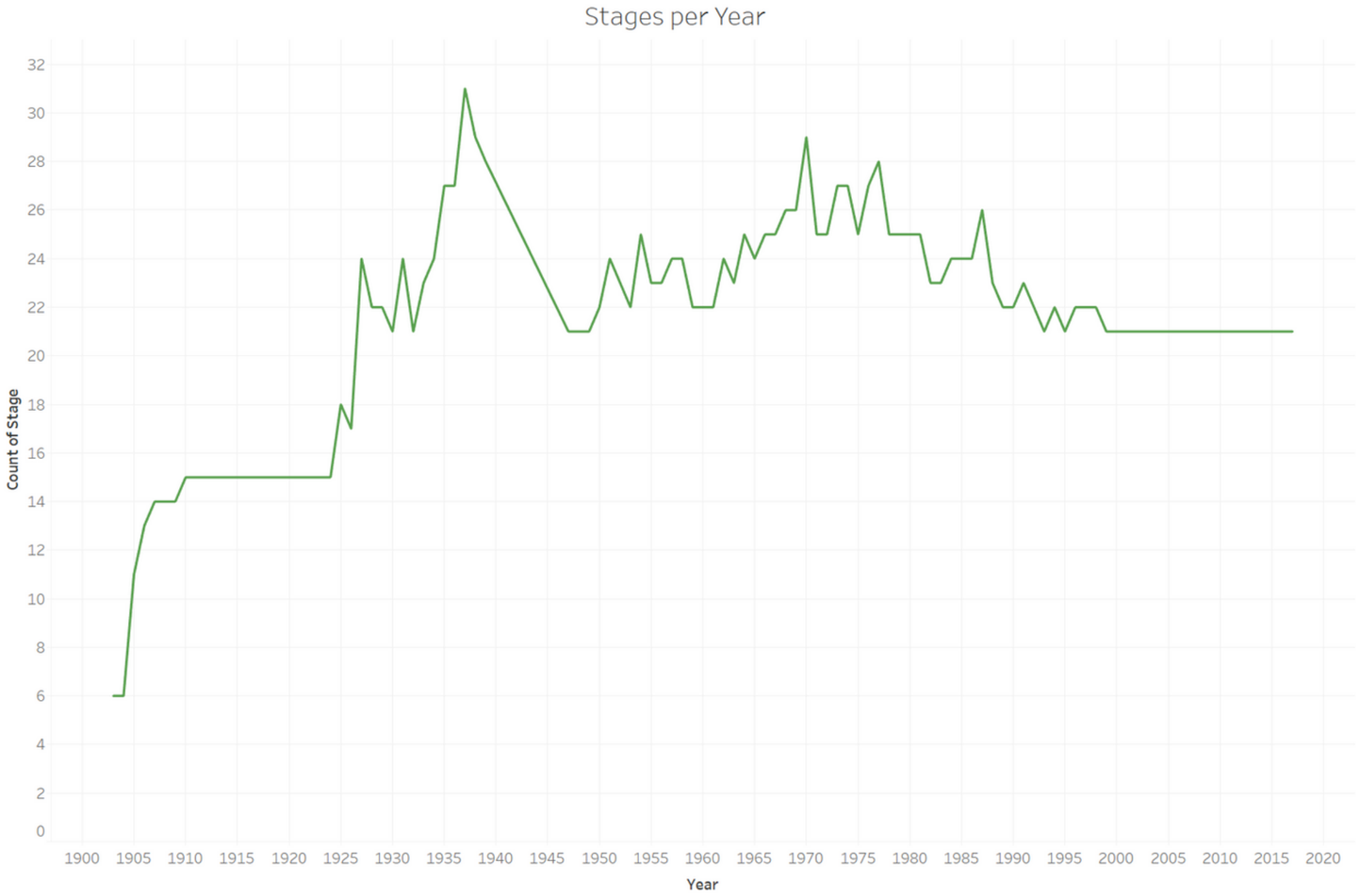
```
def get_year(race_date):
    l_date = race_date.split("-")
    return int(l_date[0])
```

Stage Type: In order to sort the stages by type more effectively I need to standardize the stage types and text formatting. Grouping all variations of mountain stages, flat stages, etc regardless of captialization as well

```
def group_stage(stage):
    if stage in ['Flat Stage', 'Flat cobblestone stage', 'Flat stage', 'P
        return "Flat Stage"
    elif stage in ['High mountain stage', 'Medium mountain stage', 'Mounta
        return "Mountain Stage"
    elif stage in ['Hilly stage']:
        return "Hilly Stage"
    elif stage in ['Individual time trial', 'Mountain time trial']:
        return "Individual Time Trial"
    elif stage in ['Team time trial']:
        return "Team Time Trial"
    else:
        return "Other Type Stage"
```

	Stage	Date	Distance	Origin	Destination	Type	Winner	Winner_Country	Year
0	1	2017-07-01	14.0	Düsseldorf	Düsseldorf	Individual Time Trial	Geraint Thomas	GBR	2017
1	2	2017-07-02	203.5	Düsseldorf	Liège	Flat Stage	Marcel Kittel	GER	2017
2	3	2017-07-03	212.5	Verviers	Longwy	Mountain Stage	Peter Sagan	SVK	2017
3	4	2017-07-04	207.5	Mondorf-les-Bains	Vittel	Flat Stage	Arnaud Démare	FRA	2017
4	5	2017-07-05	160.5	Vittel	La Planche des Belles Filles	Mountain Stage	Fabio Aru	ITA	2017
5	6	2017-07-06	216.0	Vesoul	Troyes	Flat Stage	Marcel Kittel	GER	2017
6	7	2017-07-07	213.5	Troyes	Nuits-Saint-Georges	Flat Stage	Marcel Kittel	GER	2017
7	8	2017-07-08	187.5	Dole	Station des Rousses	Mountain Stage	Lilian Calmejane	FRA	2017
8	9	2017-07-09	181.5	Nantua	Chambéry	Mountain Stage	Rigoberto Urán	COL	2017
9	10	2017-07-11	178.0	Périgueux	Bergerac	Flat Stage	Marcel Kittel	GER	2017
10	11	2017-07-12	203.5	Eymet	Pau	Flat Stage	Marcel Kittel	GER	2017
11	12	2017-07-13	214.5	Pau	Peyragudes	Mountain Stage	Romain Bardet	FRA	2017
12	13	2017-07-14	101.0	Saint-Girons	Foix	Mountain Stage	Warren Barguil	FRA	2017
13	14	2017-07-15	181.5	Blagnac	Rodez	Mountain Stage	Michael Matthews	AUS	2017
14	15	2017-07-16	189.5	Laissac-Sévérac-l'Église	Le Puy-en-Velay	Mountain Stage	Bau		
15	16	2017-07-18	165.0	Le Puy-en-Velay	Romans-sur-Isère	Mountain Stage	Micha		
16	17	2017-07-19	183.0	La Mure	Serre Chevalier	Mountain Stage	P		
17	18	2017-07-20	179.5	Briançon	Col d'Izoard	Mountain Stage	W		
18	19	2017-07-21	222.5	Embrun	Salon-de-Provence	Flat Stage	Edvald Boa		

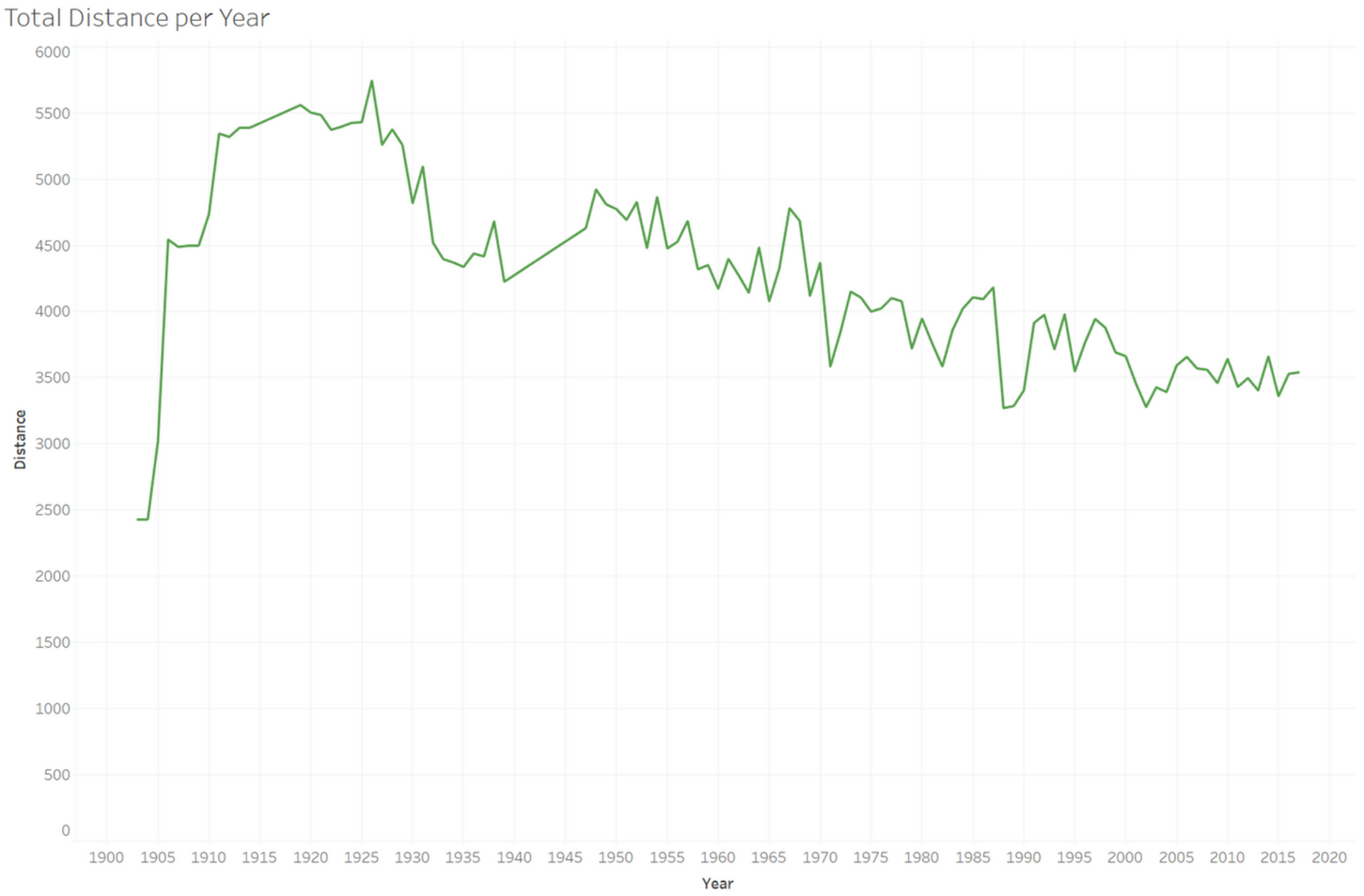
Now that it's been wrangled, let's take a clean look at the dataframe and get into some questions.



Question 1

How did the number of stages change over time?

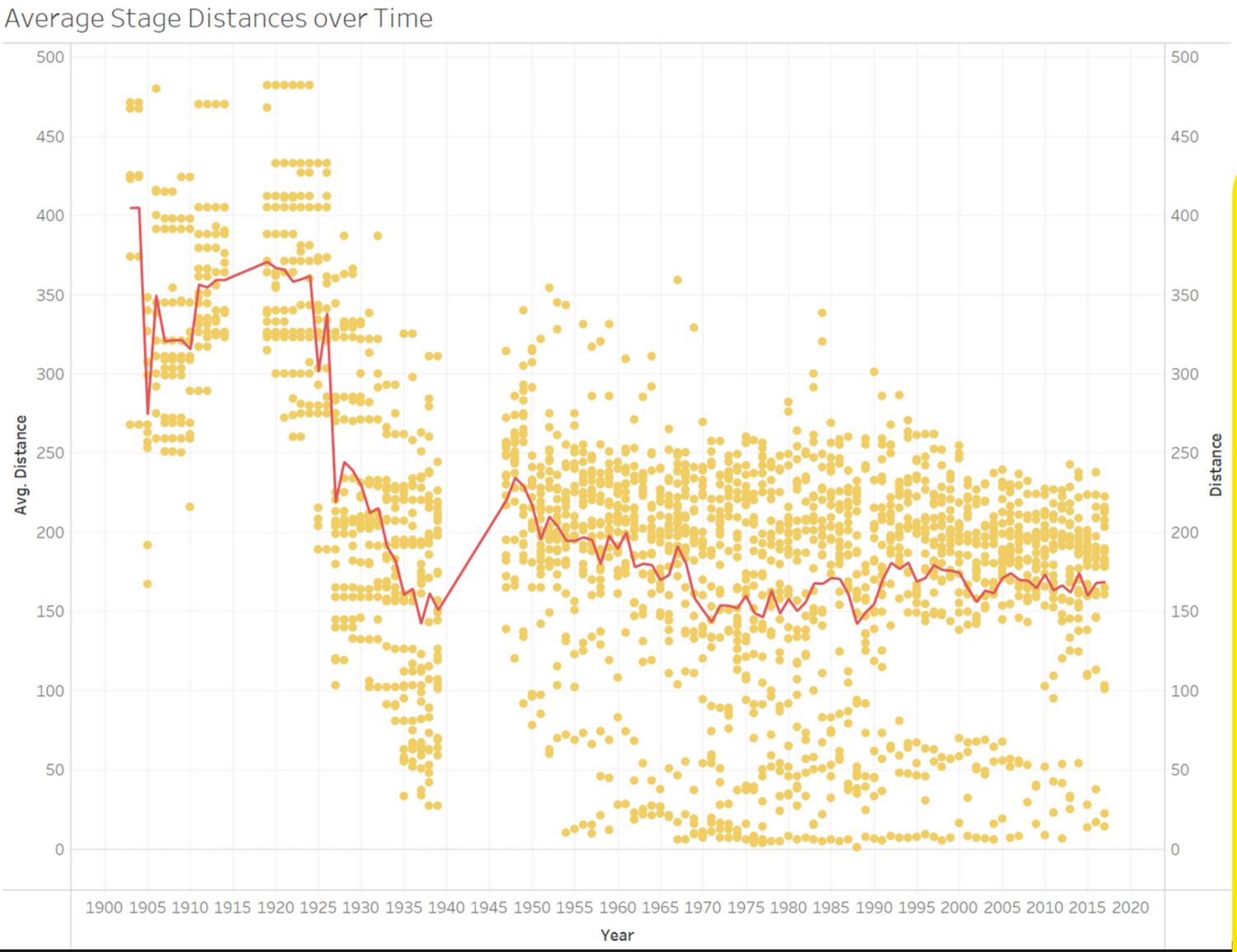
While I had assumed that the number of stages would have only increased over time with a leveling out in the early 2000's, it actually spiked early in the race's development. The move to 21 stages began in the late 1980's and is now considered the standard. Other notes are the near immediate doubling of number of stages and then the variability of stage numbers through the 1960s and 80s.



Question 2

How did the total race distance change over time?

My initial hypothesis was an overall increase in distance over time as technology and cycling as a sport has evolved. Instead again we have an early hike with overall distance peaking in the 1920s, and then a consistent decrease in overall distance until the 2000's where we have settled at roughly 3500km.



Question 3

How did the average stage length change over time?

After seeing the decrease in total distance and an increase in stage numbers it makes sense that the average stage distance took follows those trends. The average stage distance has sat between 150 to 200km since roughly 1960. The other noted features are the spread of distances for stages each year, which now are separated between full group stages (which tend to be longer) and time trials (which tend to be shorter). This divide became most distinct in the 1990s.

Because the Tour is as long as it is with as much variability across the country of France, stages are noted with stage types to denote the terrain and race style. Special jerseys are given to those who prove their merit in the mountains and sprints. The notation for stage types has changed over time which is why we wrangled the formatting at the start of our work.

Question 4

What stage type is the most common and when did the come into use?

Flat stages are the most common stage type by far, nearly doubling the number of mountain stages. The classification for hilly stages was introduced later in the Tour's history and has not seen as consistent of use (note: it has been used since 2006 in Tours after our 2017 data). I assumed that individual Time Trials were introduced before team so that was an interesting discovery, however Individual Time Trials have been used much more often and consistently than Team Time Trials

```
#Years of first (maybe last) introduction:
#flat
flat_stages = tour_data[tour_data['Type'] == 'Flat Stage']
flat_stages.sort_values(by = "Year")
#First introduced in 1903, last used in 2017
```

Type	Stage
Flat Stage	1175
Hilly Stage	76
Individual Time Trial	218
Mountain Stage	668
Other Type Stage	12
Team Time Trial	87

Stage Type	First Year	Most Recent Year
Flat Stage	1903	1927
Hilly Stage	1972	2006
Mountain Stage	1903	2017
Individual Time Trial	1934	2017
Team Time Trial	1927	2015

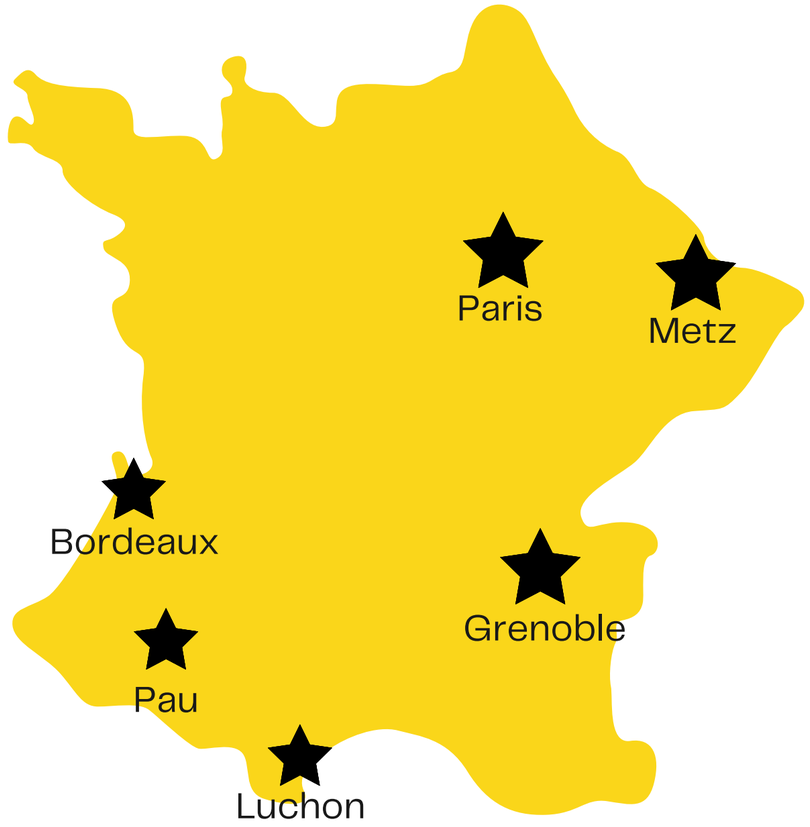
```
# stage type v average distance
ty_st_count = tour_data.groupby(['Type']).mean()
ty_st_count[['Distance']]
```

Type	Distance
Flat Stage	222.665787
Hilly Stage	196.328947
Individual Time Trial	39.482454
Mountain Stage	215.687275
Other Type Stage	148.500000
Team Time Trial	103.278161

Question 5

What is the average stage length for each type?

Initially I believed that flat stages would be the longest by far, as mountain climbs would theoretically shorten the overall stage length. This was an incorrect assumption as flat and mountain stages are within 10 km of each other, and hilly stages not too far behind. It makes sense that the sprint style time trials are significantly shorter than peloton stages, with individual time trials still roughly a quarter of the length of team time trials.




```
# Popular cities? I know Paris but others?

# count of origin
ocity_counts = tour_data.groupby(["Origin"]).count()
ocity_counts.sort_values(by = ['Stage'], ascending = False).head(10)
#I thought Paris was #1 but its not!!
first_ocity = tour_data[tour_data['Stage'] == '1'].groupby(['Origin']).count()
first_ocity.sort_values(by = 'Stage', ascending = False)

# count of destinations
dcity_counts = tour_data.groupby(["Destination"]).count()
dcity_counts.sort_values(by = ['Stage'], ascending = False).head(10)
#Paris is the #1 destination but not origin


cityab_counts = tour_data.groupby(["Origin", "Destination"]).count()
cityab_counts.sort_values(by = ['Stage'], ascending = False).head(10)
#Pau to Bordeaux has been a stage 18 times!
```

Question 6(s)

What are the most popular Tour hotspots?

What is the most common stage origin city?

Top 5 in order of most origin appearances: Pau (62), Bordeaux (56), Luchon (51), Paris (44), Grenoble (40)

What is the most common city to begin the entire Tour in?

Top 2 in order of number of times the first stage of the race began there: Paris (36), Brest (3)

What is the most common stage destinations?

Top 5 in order of most origin appearances: Paris (108), Bordeaux (79), Pau (60), Luchon (43), Metz (38)

What are the common legs of the Tour?

Top 3 in order of appearances: Pau to Bordeaux (18), Luchon to Perpignan (17), Starbourg to Metz (13)

Question 7

What country has the most wins? By stage type?

I pressumed correctly that in the Tour de France those de France had probably won the race the most times. The next several countries on the rank for most wins are neighboring European countries a few of whom host their own Grand Tours. The first non-European country to break into the ranks is the US at number 10 for most stage wins. The first non-Frenchman stage win was actually in the first year of the tour 1903 by a Swiss rider. The tour is only becomming more diverse and international it will be interesting to see what other countries climb the ranks.

Winner_Country	Stage
FRA	691
BEL	460
ITA	262
NED	157
ESP	125
GER	71
LUX	70
GBR	67
SUI	57
USA	38

Stage Type	Most Wins	Second Most Wins	Third Most Wins
Flat Stage	France (318)	Belgium (272)	Italy (152)
Hilly Stage	France (20)	Belgium (12)	Netherlands (11)
Mountain Stage	France (216)	Belgium (117)	Italy (93)
Individual Time Trial	France (59)	Belgium (38)	Spain (22)

```
tour_data.groupby(['Winner_Country']).count()
# The most Tour de France winners are from France, which makes sense
non_french = tour_data[tour_data['Winner_Country'] != 'FRA']
non_french.tail()
# The first non-French stage winner was from Switzerland in 1903

#lets look at stage types
#flat - French, then Belgian, then Italian
wc_flat = flat_stages.groupby(['Winner_Country']).count()
wc_flat.sort_values(by = 'Stage', ascending = False).head()
```

Question 8

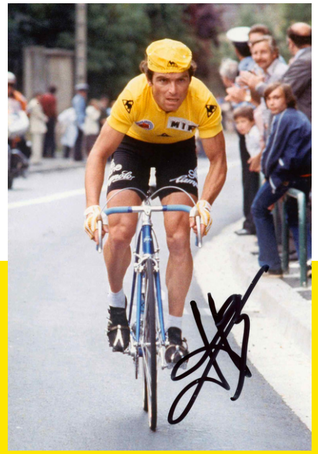
Who has the most stage wins?

The name most Americans know in cycling is Lance Armstrong but he is not the top name in the history of the Tour de France. Several key players have made their names setting records with their performance in the Tour. Here is a look at the riders with the most stage wins.

```
# Most stages won

win_count = tour_data.groupby(['Winner']).count()
win_count.sort_values(by = 'Stage', ascending = False).head(10)
```

Winner	Stage
Eddy Merckx	34
Mark Cavendish	30
Bernard Hinault	28
André Leducq	24
André Darrigade	22
Nicolas Frantz	20
Lance Armstrong[n 1]	20
François Faber	18
Jean Alavoine	17
Charles Pélissier	16



Bernard Hinault (28)

Eddy Merckx (34)

Mark Cavendish (30)

```python
# Winner v stage distance?

# Group by winner
# Sum of distance
winner_dist = tour_data.groupby(["Winner"]).sum()
winner_dist.sort_values(by = 'Distance', ascending = False).head(10)
```

Question 9

Who has won the most kilometers?

This was more a curiosity than a deep goal for my investigation. Especially with changes to stage distance over time, and the steep difference between time trial and peloton stages, I decided to take a look at who had won the most kilometers by summing the distance of their stage wins. Our podium leader fro this question, Jean Alavione, is a racer from the 1910s/20s when stages were the longest they ever were.

Winner	Distance
Jean Alavoine	5965.0
François Faber	5855.0
Nicolas Frantz	5648.0
Mark Cavendish	5557.0
André Leducq	5436.0
André Darrigade	4501.0
Philippe Thys	4089.0
Louis Trousselier	3974.0
Eddy Merckx	3441.9
Charles Pélissier	3425.0

Question 10

Who has the most wins per stage type?

Take advantage
of the tools available.

```
# Winner v stage types

# Group by winner
# Count of type
win_type = tour_data.groupby(["Winner", "Type"]).count()
win_type.sort_values(by = "Stage", ascending = False)

# compare against reverse, group by type, count of winner

type_leaderboards = tour_data.groupby(['Type', 'Winner']).count()
type_leaderboards.sort_values(by = "Stage", ascending = False).head(30)

#individual breakdown
#flat - Cavendish, then Darrigade, then Kittel
w_flat = flat_stages.groupby(['Winner']).count()
w_flat.sort_values(by = 'Stage', ascending = False).head()
```

Flat Stages:

1. Mark Cavendish – 30
2. André Darrigade – 18
3. Marcel Kittel & René Le Grevès – 14

Hilly Stages:

Six way tie for first at 2 hilly stage wins between Henk Lubberding, Laurent Jalabert, Jean-Paul van Poppel, David Moncoutié, Sean Kelly, and Erik Dekker

Mountain Stages:

1. Eddy Merckx – 13
2. François Faber – 11
3. Gino Bartali & Lance Armstrong – 9

Individual Time Trial:

1. Bernard Hinault – 20
2. Eddy Merckx – 16
3. Jacques Anquetil & Lance Armstrong – 11

Conclusion

Between the invention of helmets, carbon fiber, and modern steroids, the Tour de France has seen a century of evolution. Race leaders walk a fine line of maintaining traditions while keeping up with cycling innovation. After another 100 years of riding who knows where we will go