Projeto SCC5836 - Visualização Computacional
Visualização e observações sobre a velocidade média no ciclismo de estrada

Nome: Helder Giro Lopes

Nº USP: 7547055

Docente(s): Mara Cristina Ferreira de Oliveira

Eric Macedo Cabral

Parte 1

Contextualização e motivação

No ciclismo de estrada, existem vários fatores que definem o quão bom um atleta é, mas poucos deles conseguem bater a velocidade média, afinal, numa competição ganha o primeiro a cruzar a linha de chegada.

Provas como contra-relógio individual ou a seção de ciclismo do triatlo podem ser decisivas e, diferente das provas de ciclismo tradicionais onde há um pelotão correndo, são mais controladas e podem ter o seu resultado predito de acordo com características do atleta.

Este trabalho, portanto, visa analisar características pessoais do atleta como a a cadência média da pedalada (número de voltas do pedivela por minuto), a potência empregada (fisicamente medida pelo torque aplicado no pedivela multiplicado pela cadência), o seu peso, a elevação do percurso, dentre outros fatores e observar como eles influenciam na velocidade média da atividade.

Obtenção dos dados

Os dados foram exportados do aplicativo Strava[1]. Este é um aplicativo onde o atleta pode subir, armazenar e compartilhar atividades de diversos esportes com seus seguidores.

O Strava permite que o usuário exporte todas as suas atividades de uma única vez, realizando um simples procedimento. A pasta disponibilizada contém vários arquivos, entre eles, um chamado *activities.csv*, que foi o arquivo utilizado neste projeto.

[1] strava.com

O arquivo é bem completo e contém um total de 80 *features.*

Manipulação dos dados

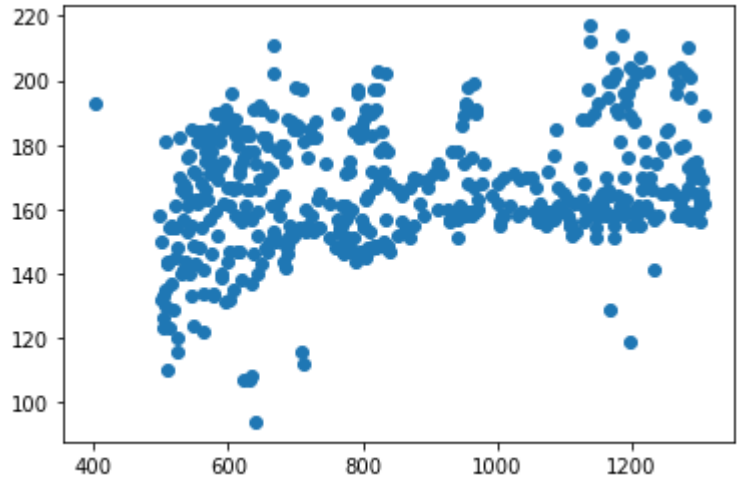
Primeiramente o csv foi transformado num dataframe. Após isso, foram filtradas apenas as atividades de ciclismo, já que o atleta pode possuir atividades de vários outros esportes (corrida, natação, etc).

Das 80 *features* iniciais, a maioria poderia ser diretamente descartada, restando 16 *features*, que podem ser vistas na tabela abaixo

Activity Date	Distance	Athlete Weight	Bike Weight
Moving Time	Max Speed	Average Speed	Elevation Gain
Elevation Loss	Elevation Low	Elevation High	Average Cadence
Average Heart Rate	Average Watts	Weighted Average Power	Wind Speed

Inicialmente, 911 atividades foram contabilizadas. A primeira manipulação foi retirar todas as atividades em que não havia medidor de potência, já que esta é uma *feature* indispensável e fortemente correlacionada à velocidade média da atividade.

A Figura abaixo mostra que as primeiras atividades não continham o valor de potência, já que a contagem se inicia a partir da atividade número 400.




```
df = df.dropna(subset=['Weighted Average Power'])
```

Após isso, notou-se que existiam várias atividades sem o peso do atleta. Elas foram preenchidas com a moda desta *feature*.

```
df['Athlete Weight'] = df['Athlete Weight'].fillna(df['Athlete Weight'].mode()[0])
```

O mesmo foi feito com o peso da bicicleta.

Além disso, foram removidas todas as atividades de ciclismo *indoor*, já que a velocidade neste caso é irrelevante.

```
df = df.dropna(subset=['Elevation Gain'])
df = df.dropna(subset=['Elevation Loss'])
df= df[df['Elevation Gain'] != 0.0]
df= df[df['Elevation Loss'] != 0.0]
```

Os valores relacionados a velocidade foram todos multiplicados por 3.6, pois eles vieram em m/s e a métrica mais utilizada é km/h

```
df['Max Speed'] = df['Max Speed'] * 3.6
df['Wind Speed'] = df['Wind Speed'] * 3.6
```

As *features* de peso do atleta e de peso da bicicleta foram somadas e chamadas de *Total Weight* para reduzir uma dimensionalidade

Uma *feature* importante, mas que não havia inicialmente no dataframe, é a inclinação média da atividade. Ela indica a proporção entre o ganho de elevação e a distância percorrida

```
df['Average Grade'] = df['Elevation Gain'] /  (df['Distance'] * 10)
```

Curiosamente, algumas atividades estavam sem a velocidade média. Mas é possível calcular esta variável, bastando dividir a distância percorrida pelo tempo de movimentação.

```
df['Average Speed'] = (df['Distance'] * 60 * 60) / df['Moving Time']
```

Por fim, restaram 330 atividades que podem ser aproveitadas

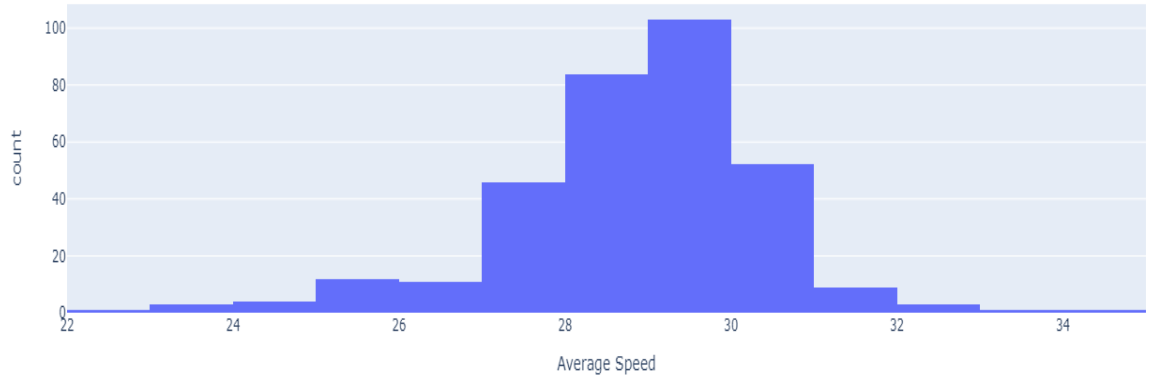
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 330 entries, 499 to 1307
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Activity Date           330 non-null    datetime64[ns]
 1   Distance                330 non-null    float64
 2   Moving Time             330 non-null    float64
 3   Max Speed               330 non-null    float64
 4   Elevation Gain          330 non-null    float64
 5   Elevation Loss          330 non-null    float64
 6   Elevation Low           330 non-null    float64
 7   Elevation High          330 non-null    float64
 8   Average Cadence         330 non-null    float64
 9   Average Heart Rate      330 non-null    float64
 10  Average Watts           330 non-null    float64
 11  Weighted Average Power  330 non-null    float64
 12  Wind Speed              193 non-null    float64
 13  Total Weight            330 non-null    float64
 14  Average Grade           330 non-null    float64
 15  Average Speed           330 non-null    float64
```

Parte 2

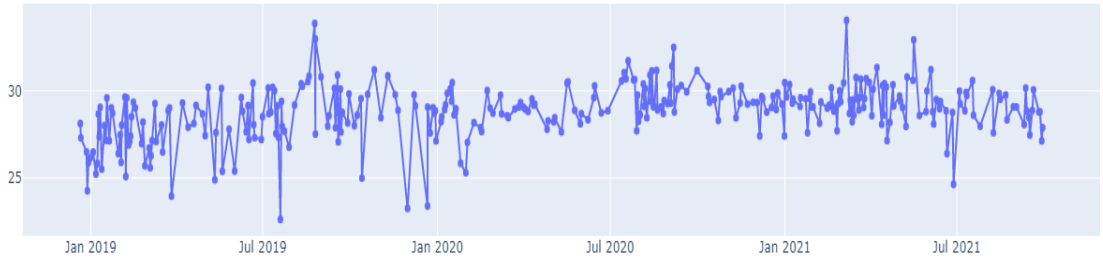
A parte 2 consistiu na visualização dos dados usando as bibliotecas *plotly* e *seaborn*.

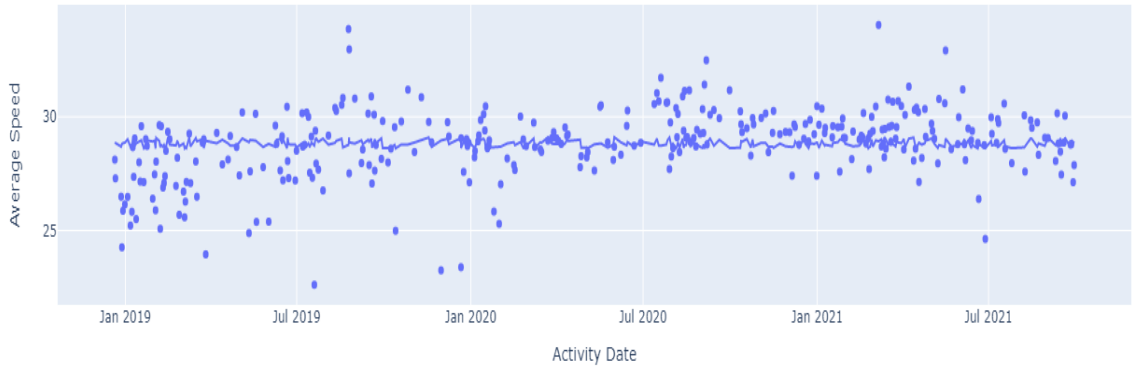
O gráfico abaixo é um *Density Plot,* que mostra a contagem de velocidade média das atividades. Nota-se uma grande queda entre 29 e 30 km/h.



Os próximos dois *Scatter Plots* mostram como a velocidade média mudou de acordo com o tempo.

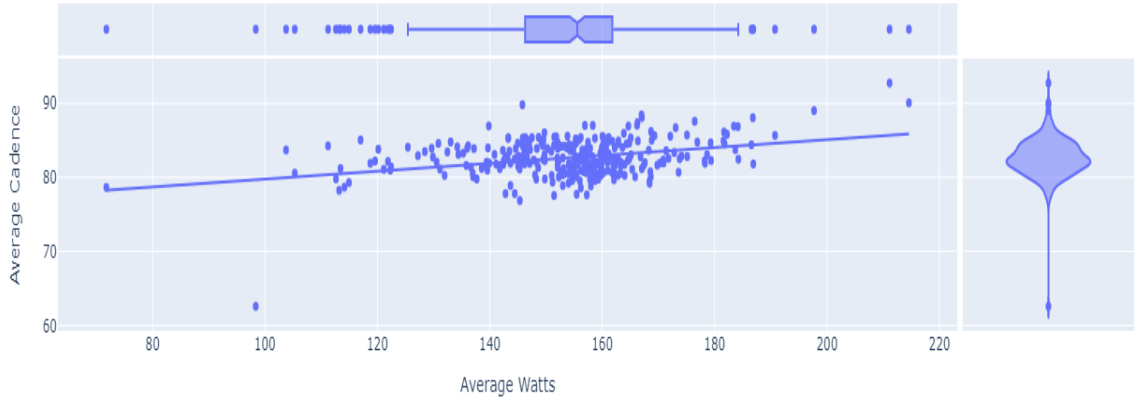
O primeiro é um *Connected Scatter plot,* enquanto o segundo é um *Scatter Plot* tradicional com uma *trendline* para ficar mais clara a tendência de aumento da velocidade com o passar do tempo



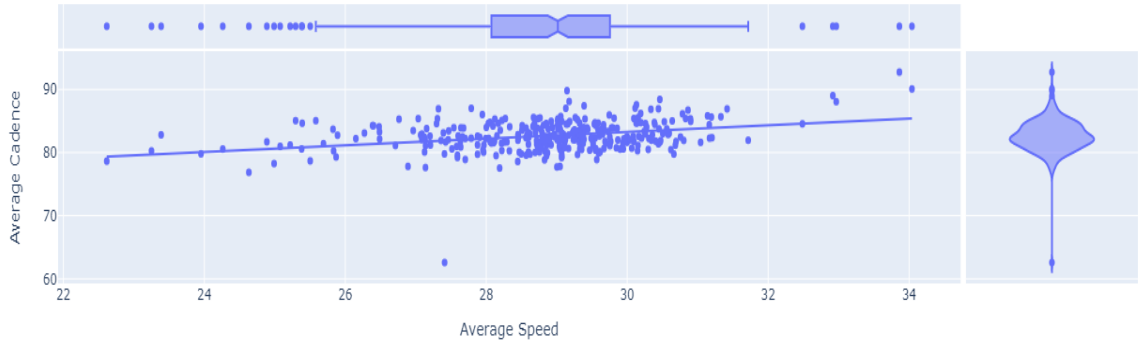


O *Scatter Plot* abaixo mostra a relação entre a potência média (*average watts*) e a cadência média das atividades. Nota-se que existe uma relação linear entre a potência e a cadência.

Foi escolhido como opções adicionais de visualização um *Box Plot* para a potência média e um *Violin Plot* para a cadência.

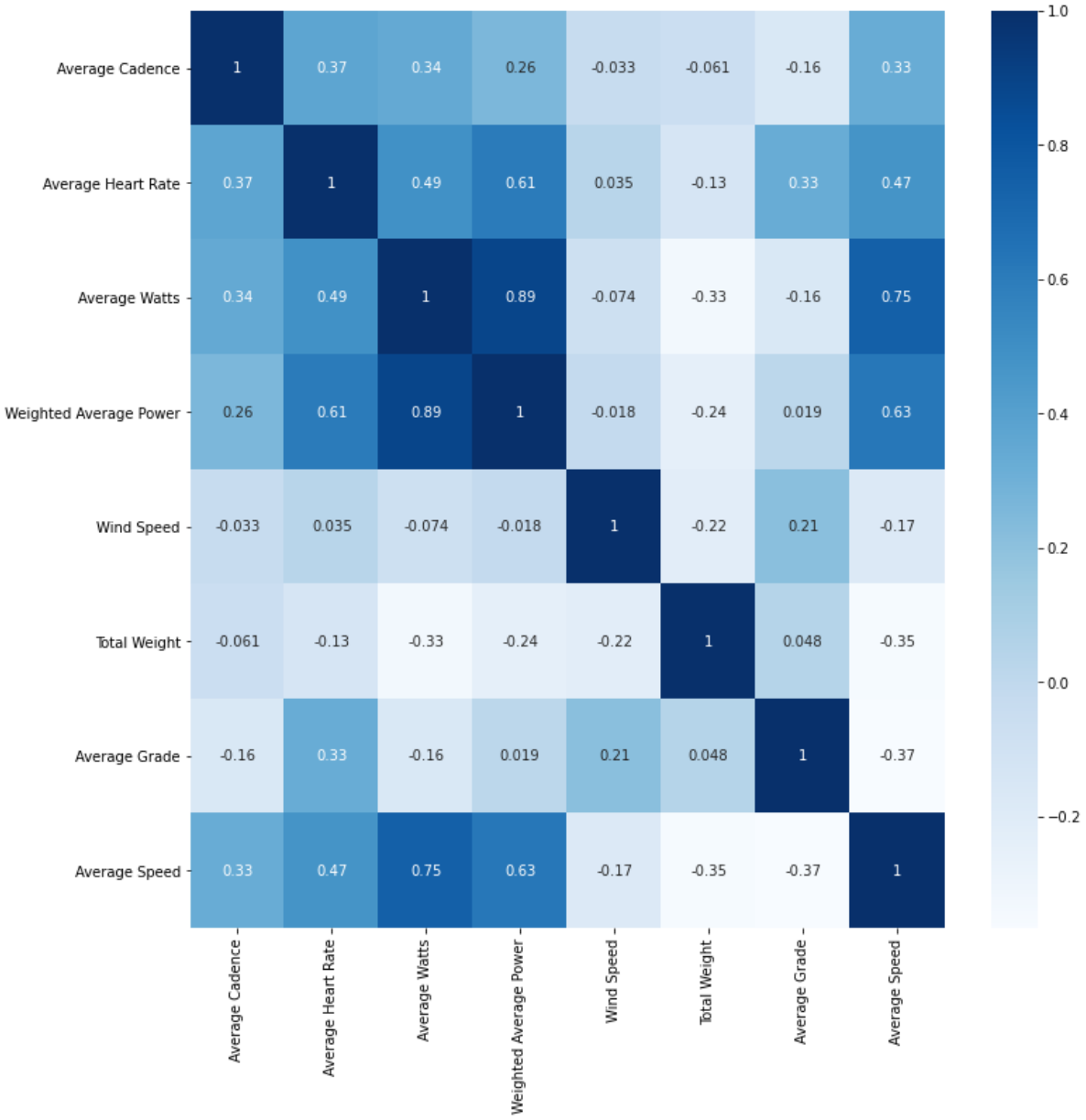


Seguindo a mesma tendência do gráfico anterior, abaixo é apresentado o gráfico de velocidade por cadência

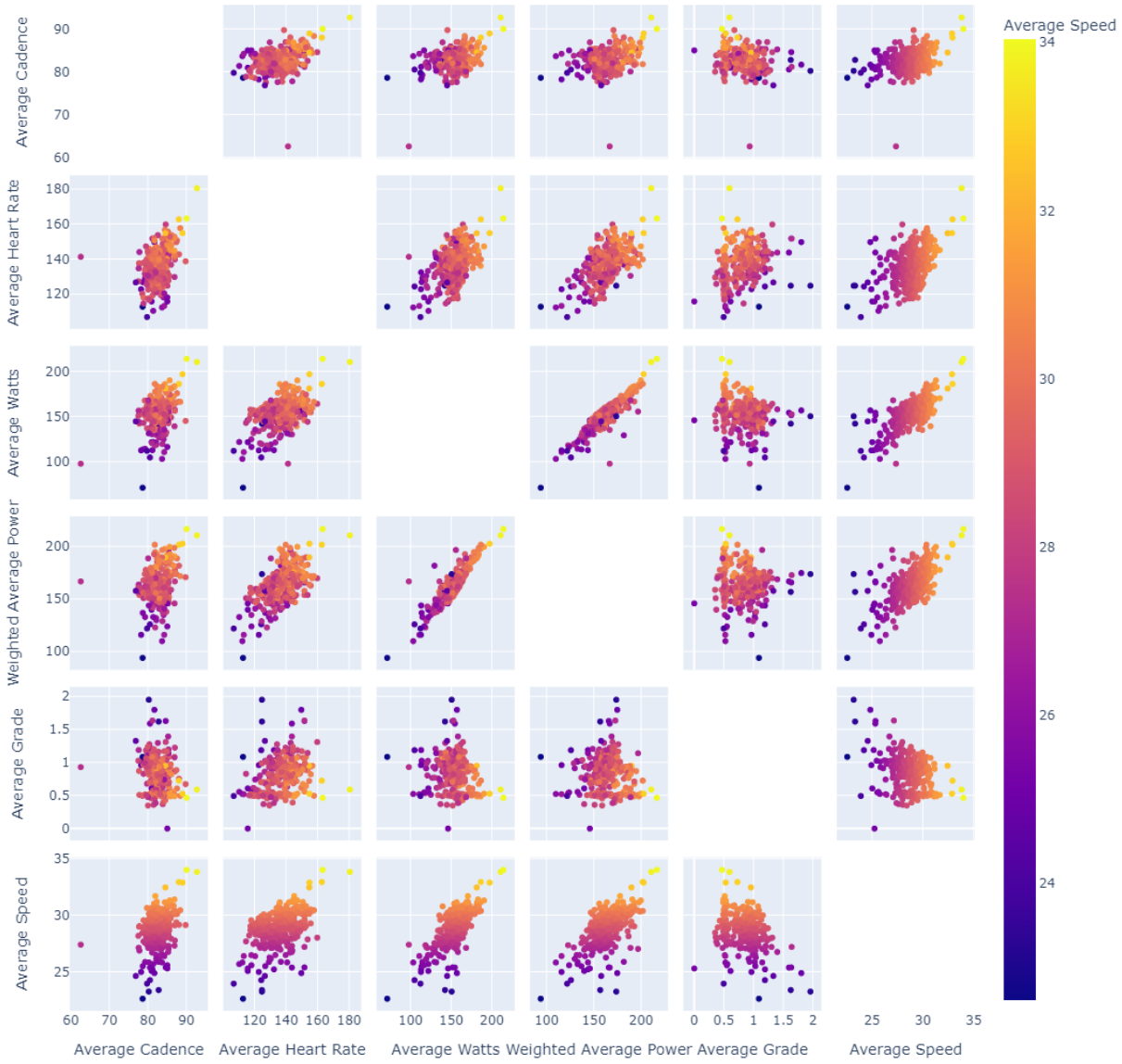


Como todas as variáveis são contínuas, pode-se fazer a correlação entre elas pelo coeficiente de *Spearman*. A correlação é apresentada abaixo.

Como esperado, a correlação mais forte com a velocidade média é a potência média.



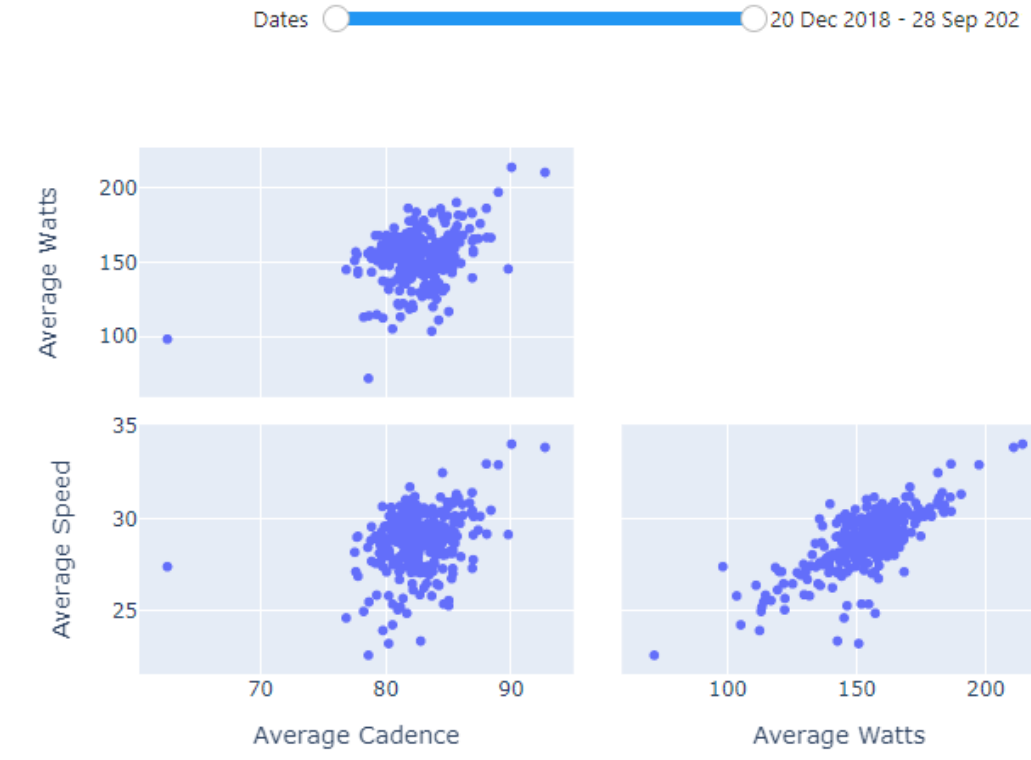
Por último, um correlograma mostra as relações entre algumas das principais *features* e o *target*. É possível notar algumas tendências neste correlograma, por exemplo, o gradiente médio alto reflete numa menor velocidade média (num percurso com mais subidas, o ciclista vai mais devagar).



Parte 3

A última parte do projeto consiste em gerar um *dashboard* com os gráficos obtidos na Parte 2. Além disso, foi adicionado uma seção interativa, onde o usuário pode escolher as datas de início e fim do *dataframe* e assim *scatter plots* de cadência, potência e velocidade são atualizados.

Foi escolhido a variável de potência pois ela tem correlação forte e direta com a velocidade média. A cadência também é interessante para ser observada, pois ela pode mostrar características próprias do ciclista com relação à sua maneira de pedalar.


Dates
20 Dec 2018 - 28 Sep 202
Average Watts
200
150
100
Average Speed
35
30
25
70
80
90
Average Cadence
100
150
200
Average Watts


Average Cadence
90
80
70
Average Watts
200
150
100
Average Speed
30
25
Jan 2019
Jul 2019
Jan 2020
Jul 2020
Jan 2021
Jul 2021