Opdracht data case - Vlammrs assessment

Tour de France

In het vakgebied van data science wordt steeds vaker gekeken naar het kunnen voorspellen van bepaalde gebeurtenissen gebaseerd op data uit het verleden, en ook naar (commerciële) mogelijkheden die deze kennis biedt.

In deze opdracht krijg je een dataset waarin je verschillende eigenschappen over de historie van de Tour de France ziet. De Tour de France is een wielrenwedstrijd die elk jaar gereden wordt en bestaat uit een aantal etappes per Tour.

De data bestaan uit twee verschillende .csv bestanden met elk hun eigen informatie. 'Stage_results' bevat informatie over de individuele etappes die afgelegd zijn tijdens een editie (de uitslag per renner, de tijd die een renner over de etappe gedaan heeft, etc). 'Stage_info' bevat informatie over de afgelegde etappes (afgelegde afstand, begin- en eindpunt). In de tabellen onderaan dit document vind je meer informatie over de datasets en variabelen.

Recentelijk heeft Marit van Egmond, de CEO van de Albert Heijn het wielrennen ontdekt. Nu wil ze graag zelf investeren in een wielerteam om haar supermarkt te promoten en daarnaast ook op de weg de concurrentie aan te gaan met de Jumbo. Om de grootste kans van slagen te garanderen ben jij gevraagd om op basis van deze data een advies op te stellen. Om tot een goed advies te komen vragen we je om de onderstaande vragen voor te bereiden, te beantwoorden en aan ons te presenteren. De vorm waarin je de presentatie giet mag je zelf bepalen.

Tijdens de assessmentdag hoort het MT van Albert Heijn graag hoe je het hebt aangepakt en wat je bevindingen zijn. Zij zullen ook vragen stellen over je presentatie en de mogelijkheden die de data biedt.

1) Wat vind je van de kwaliteit van de datasets?
    - Kijk bijvoorbeeld naar de outliers en missing values in de dataset.

2) Wat kun je zeggen over de afstand en het type van de etappes (stage type) over de edities heen?

3) Kijkend naar de top 10 renners met de meeste etappewinsten in de historie van de Tour de France, wat valt je op?

4) Welke teams zijn de laatste 5 edities het meeste te vinden op het podium (top 3) van de etappes?

5) Zoals in de introductie al aangegeven, wil de CEO van de Albert Heijn investeren in een wielerteam om de komende jaren successen mee te behalen en haar merk te promoten. Vanuit marketing oogpunt is het interessant wanneer een renner in de top 10 eindigt. Formuleer een advies en baseer dit op de top 10 renners per etappe. Kijk ook naar de variabelen die je hebt gebruikt in de vragen 2 t/m 4 (en voeg er meer toe als je dat nodig vindt).

Databronnen:

Stage_results.csv

Stage_ID	ID of the stage
Edition	Edition of the Tour de France
Year	Year of the edition
Stage_number	Number of the stage
Rank	The rank of the rider on the stage
Rider	Name of the rider
Age	Age of rider
Team	Team of rider
Time_elapsed	The time it took the rider to complete the stage in seconds

Stage_info.csv

Stage_ID	Unique identifier per stage
Year	Year of edition
Stage_number	Number of the stage
Distance	Distance in KM
Origin	Starting city
Destination	Finishing city
Type	Stage type

Extra informatie Tour de France:

De Tour de France is een jaarlijkse wielerwedstrijd waarin een groot aantal wielrenners (184 in 2021, verdeeld over 23 ploegen) van start gaan. Elke dag wordt er een etappe afgelegd. De renner die alle etappes in de minste tijd heeft afgelegd, mag zich de winnaar van de Tour de France noemen. Deze etappes zijn terug te vinden in de 'stage_info' tabel. Hierin wordt informatie *per etappe* gegeven zoals het nummer van de etappe, de afstand, de start en finish plaats, het type etappe en een uniek ID (Stage_ID).

Per etappe is er ook data beschikbaar, dit is te vinden in 'stage_results'. Hierin staat een stage_ID die correspondeert met één van de etappes in 'stage_info'. Verder staan hierin de uitslagen *per etappe* met de rank (plaats waarop de rijder gefinisht is desbetreffende etappe), de naam van de renner, zijn leeftijd en team. Ook wordt de tijd in seconden gegeven in 'time_elapsed', dit is hoe lang desbetreffende renner over de etappe heeft gedaan. De top drie renners van een etappe zijn dus de renners met rank 1, 2 en 3.

Note: het betreft een fictieve dataset.