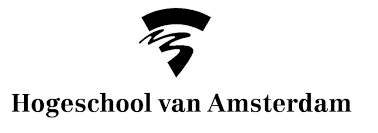
Tennis topsport dashboarding





Big Data C1 Team 1

Youp Schaefers 500738283

Sander van de Kamp 500734351

Stijn Matla 500783587

Arvid de Haas 500772805 11-1-2021

# Voorwoord

Tijdens het semester Big Data hebben wij (Sander van de Kamp, Stijn Matla, Arvid de Haas en Youp Schaefers) voor het project tennis gekozen. Bij het vormen van het team kende Stijn, Arvid en Youp elkaar al maar Sander nog niet. Het klikte gelijk met Sander en hierdoor konden wij ervoor zorgen dat er een mooi eindproduct geleverd kon worden.

Tijdens het project hebben wij kennis gemaakt met Power BI en moesten wij Python leren. Met de kennis binnen het team lukte het erg goed om deze nieuwe dingen makkelijk te leren en te gebruiken. Hierdoor is het gelukt om de eisen van de opdrachtgever te realiseren.

Uiteindelijk is er een dashboard in Power BI uitgekomen waar wij trots op zijn.

Wij willen graag Dennis van Kooij van Sport Data valley en Aldo van de KNLTB bedanken voor de mogelijkheid om voor hun dit project te kunnen uitvoeren. Ook willen wij graag Peter Odenhoven en Kees Rijsenbrij bedanken voor de begeleiding tijdens het project.

# Inhoudsopgave

[**Organisatie beschrijving**](#_mty1yxbuef6b) **1**

[**Ideeën van de opdrachtgever, waarom heeft de opdrachtgever dit project nodig?**](#_57aoqx1mwt7i) **2**

[**Voorafgaande analyse en uitbreiding van de opdracht**](#_rjcq51s4tzgi) **3**

[**Scope van de data oplossing**](#_m9vve9y1jj7f) **4**

[**Aannames en randvoorwaarden**](#_t1sc61ernecf) **5**

[**Voorlopige alternatieven**](#_30r8kbq7ypei) **6**

[**Budget**](#_6dblm0dz9urx) **7**

[**Omzet (financieel/niet financieel)**](#_39kue5gxq5sq) **8**

[**Project-, gebruiker en product-risico’s**](#_oj8cn8e8vf10) **9**

[Project risico’s](#_ioq1uxsoidpr) 9

[Gebruikers risico’s](#_nndh2edu7ejx) 9

[Product risico’s](#_g2zu12uyd6tb) 9

[**Ethische en legale vraagstukken**](#_yf37j97krlrm) **10**

[**Toekomstige Ontwikkeling**](#_dyg9rxfavn24) **10**

[**Planning**](#_1ktj3pw0b20i) **11**

[**User Manual**](#_6vdcrfgqwwrt) **12**

[**Technische Handleiding**](#_8tdzxk52p0po) **17**

# Organisatie beschrijving

De KNLTB is de nationale tennisbond van Nederland die zowel op professioneel als amateurniveau tennis is Nederland ondersteund. Ze zijn een overkoepelende organisatie die zowel administratief als operationeel invloed heeft op tennis in Nederland. Ze zorgen ervoor dat de lokale clubs extra financiering kunnen krijgen en dat er op professioneel niveau begeleiding beschikbaar is.

Tijdens deze opdracht zal het projectteam te maken hebben met de IT afdeling van de KNLTB, deze is is onder andere verantwoordelijk voor het leveren van data waarmee coaches hun eigen en potentiële tegenstanders mee kunnen analyseren.

# Ideeën van de opdrachtgever, waarom heeft de opdrachtgever dit project nodig?

De opdrachtgever heeft de wens om data beter en sneller te kunnen analyseren door het gebruik van power BI. Zo moet het mogelijk worden om de Spelers van de FED cup teams individueel te analyseren en daaruit conclusies te kunnen trekken, een voorbeeld hiervan is het zien van de momenten waarop een speler een Forehand Winners Cross speelt.

Het is bij de opdracht van belang dat het zowel mogelijk wordt om deze data goed te kunnen analyseren maar ook om deze data te kunnen importeren naar een database die makkelijk kan worden aangevuld met nieuwe data. Daarnaast is het ook van belang dat de data die wordt geïmporteerd is gecontroleerd en dat foutieve data is verwijderd.

# Voorafgaande analyse en uitbreiding van de opdracht

In deze opdracht zal er gebruik worden gemaakt van meer klassieke data, het is gestructureerd en relationeel en wordt allemaal vanuit een bron gehaald. Dit zorgt ervoor dat de data zeer goed te analyseren is en dat er met het inrichten van de database en de power BI analyses weinig complicaties oplevert.

Momenteel wordt deze data in eigen excel sheets geanalyseerd, dit geeft al inzicht in enkele zaken m.b.t. simpele berekeningen echter wanneer er complexere vragen worden gesteld in dit niet meer voldoende. Daarnaast biedt het veel kansen t.o.v. de concurrentie om deze data op een goede en geautomatiseerde wijze te kunnen analyseren.

Om de opdrachtgever hierin te helpen zal er een Power BI dashboard worden gebouwd die gekoppeld is aan een database met daarin de gegevens vanuit het programma Dartfish. Deze moet de grafieken en tabellen in de gegeven excel sheet doen vervangen.

# Scope van de data oplossing

Gedurende het project zal er worden gekeken naar de technische oplossingen die gebruikt gaan worden. Dit betekent dat nog niet alle tooling vooraf staat vastgesteld. Wel is er vastgesteld dat er visualisatie zal worden gedaan met Power BI en dat er aan het einde van het project een handleiding zal worden geleverd met hierin omschrijvingen over de gebouwde oplossing. Ook is er vastgesteld dat de data zal worden opgeslagen in MySQL server met een PHPMyAdmin omgeving daar aan vast. Een product backlog zal worden bijgehouden op Trello.

# Aannames en randvoorwaarden

Aangezien het project wordt uitgevoerd in samenwerking met de HvA en er geen aangeleverde tooling buiten Power BI zijn aangeleverd wordt er vanuit gegaan dat er geen verdere vereisten zijn op het gebied van software.

Ook wordt er vanuit gegaan dat de tijd die kan worden besteed aan het project niet meer is dan de door de HvA bepaalde 308 uur per persoon zal worden besteed. Er wordt vanuit gegaan dat alle project gerelateerde taken voor 18 januari zijn voltooid.

# Voorlopige alternatieven

Mogelijke alternatieven voor het project zijn o.a. het blijven werken met de huidige oplossing. Schaduw implementatie is tijdens dit project goed mogelijk. Dat wil zeggen dat beide systemen tegelijkertijd gebruikt kunnen worden tijdens de ontwikkeling van het nieuwe systeem. Daarnaast is het mogelijk om een investering te doen in een van de beschikbare analyse systemen of verbeterd abonnement van Dartfish.

# Budget

Er is geen budget aanwezig voor zover het team weet. De studenten maken gebruik van de tools die door de HvA worden aangeboden zonder externe financiering vanuit de opdrachtgever. In een later stadium van het project zou een externe licentie voor PowerBI moeten worden aangeschaft. Deze kosten zijn $10 per maand per gebruiker voor Power BI Pro, of een complete Power BI Premium voor $ 4.995,- per maand. De inschatting van het projectteam verwacht dat de opdrachtgever voor de Power BI Pro versie zullen opteren.

# Omzet (financieel/niet financieel)

Geen directe omzet, het opgeleverde project zal geen directe kosten/opbrengsten bieden. De baten die worden gemaakt zijn op het gebied van analyse voor tennis coach. Dit zal op hele indirecte wijze naar alle waarschijnlijk wel voordelen bieden maar deze zijn niet meetbaar binnen het tijdsbestek van de opdracht. Op langere termijn wordt verwacht dat de kwaliteit van de coaching van de spelers zal verbeteren en uiteindelijk leiden tot betere resultaten.

# Project-, gebruiker en product-risico’s

## Project risico’s

* Een lid van de projectgroep kan ziek worden of stoppen met het project.
* De opdracht kan niet goed begrepen zijn
* Geen contact met de opdrachtgever kunnen krijgen
* Opdrachtgever verandert constant zijn wensen
* Verkeerd inschatten van hoeveel tijd het kost om het gewenste product te kunnen maken.

## Gebruikers risico’s

* Het gemaakte product wordt niet begrepen door de gebruiker
* Data wordt verkeerd geanalyseerd
* Geen toegang tot het dashboard
* Data wordt op verkeerde wijze geupload

## Product risico’s

* Verbinding met de database verbreekt.
* Power BI update naar nieuwere versies die de huidige versie niet meer ondersteund
* Data wordt niet goed verwerkt

# Ethische en legale vraagstukken

Bij het implementeren van interne software zijn er enkele overwegingen die moeten worden gemaakt. Zo moet er worden nagedacht of er wordt gewerkt met gevoelige informatie en wie deze mag inzien.

In het geval van het huidige project wordt er gewerkt met data die niet gevoelig is maar wel vertrouwelijk is. Om die reden moet ervoor worden gezorgd dat de gegevens niet zomaar te vinden zijn.Bij dit project zijn er geen vraagstukken wat betreft ethiek, de gegevens zijn niet gevoelig en worden met toezegging   
verwerkt.

# Toekomstige Ontwikkeling

Database

De database wordt op het moment gehost op een server van de Hogeschool van Amsterdam, voor toekomstige ontwikkelingen zou de database met alle data en procedures verplaatst moeten worden naar een eigen server.  
  
Glitch

Het uploaden van het CSV bestand gaat nu via een persoonlijke server op <https://glitch.com> , de code die op de server staat zal verplaatst moeten worden naar een eigen server.

Power BI

De power BI omgeving heeft op het moment een groot deel van de bestaande grafieken van het excel bestand, voor verdere ontwikkelingen kunnen er extra grafieken ontworpen worden om beter te passen bij de eisen van de KNLTB / SDV. Daarnaast kunnen er extra filters gecreëerd worden om te zorgen dat de grafieken zo gedetailleerd worden als nodig is.

# Planning

|  |  |
| --- | --- |
|  | **Activiteiten** |
| Sprint 1, week 1 t/m 3 | * Alle nodige programma's installeren. * Oege omgeving klaar gezet * Data van excel importeren in de database |
| Sprint 2, week 4 t/m 6 | * Data controleren en cleanen * Power Bi omgeving creeren * Mogelijke visualisaties testen met de data |
| Sprint 3, week 7 t/m 9 | * Data verder cleanen * Verschillende dashboard maken * Product versie 1.0 opleveren |
| Sprint 4, week 10 t/m 12 | * Contact met de opdrachtgever * Feedback verwerken |
| Sprint 5, week 13 t/m 16 | * Opleveren product versie 2.0 * Eindpresentatie |

# User Manual

Big Data C1 Team 1

Youp Schaefers 500738283

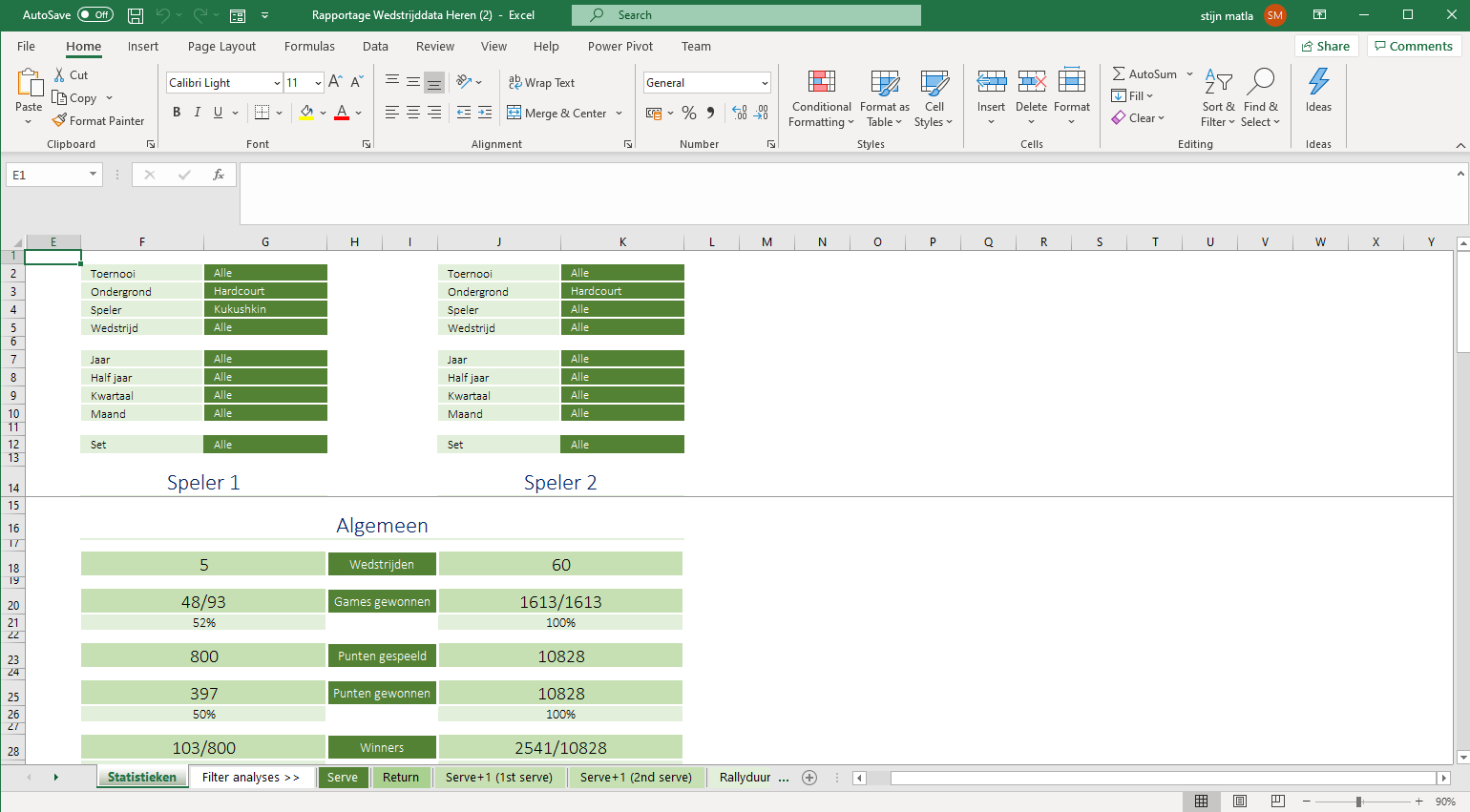
Sander van de Kamp 500734351

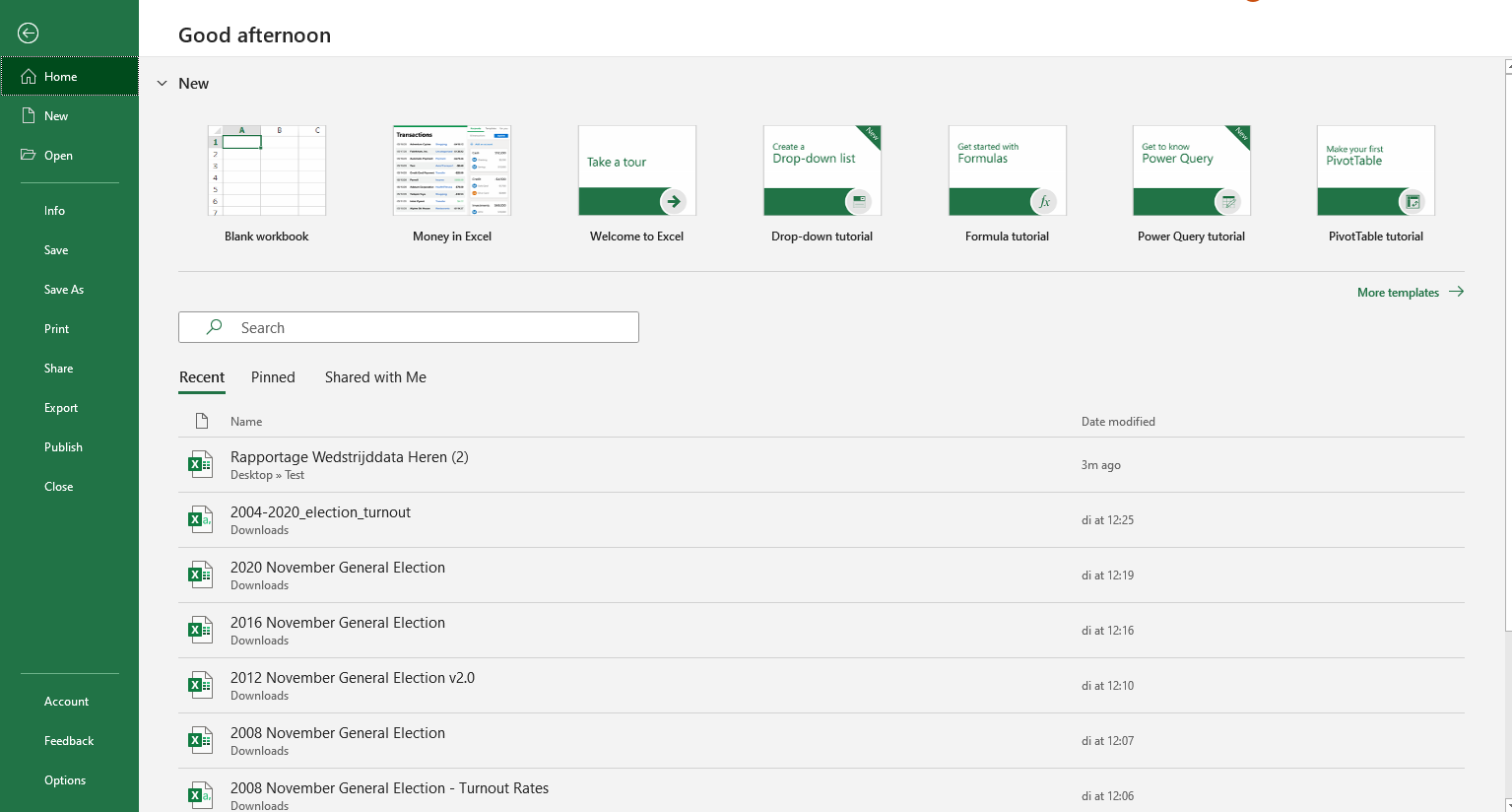
Stijn Matla 500783587  
Arvid de Haas 500772805 11-1-2021

# Data Upload

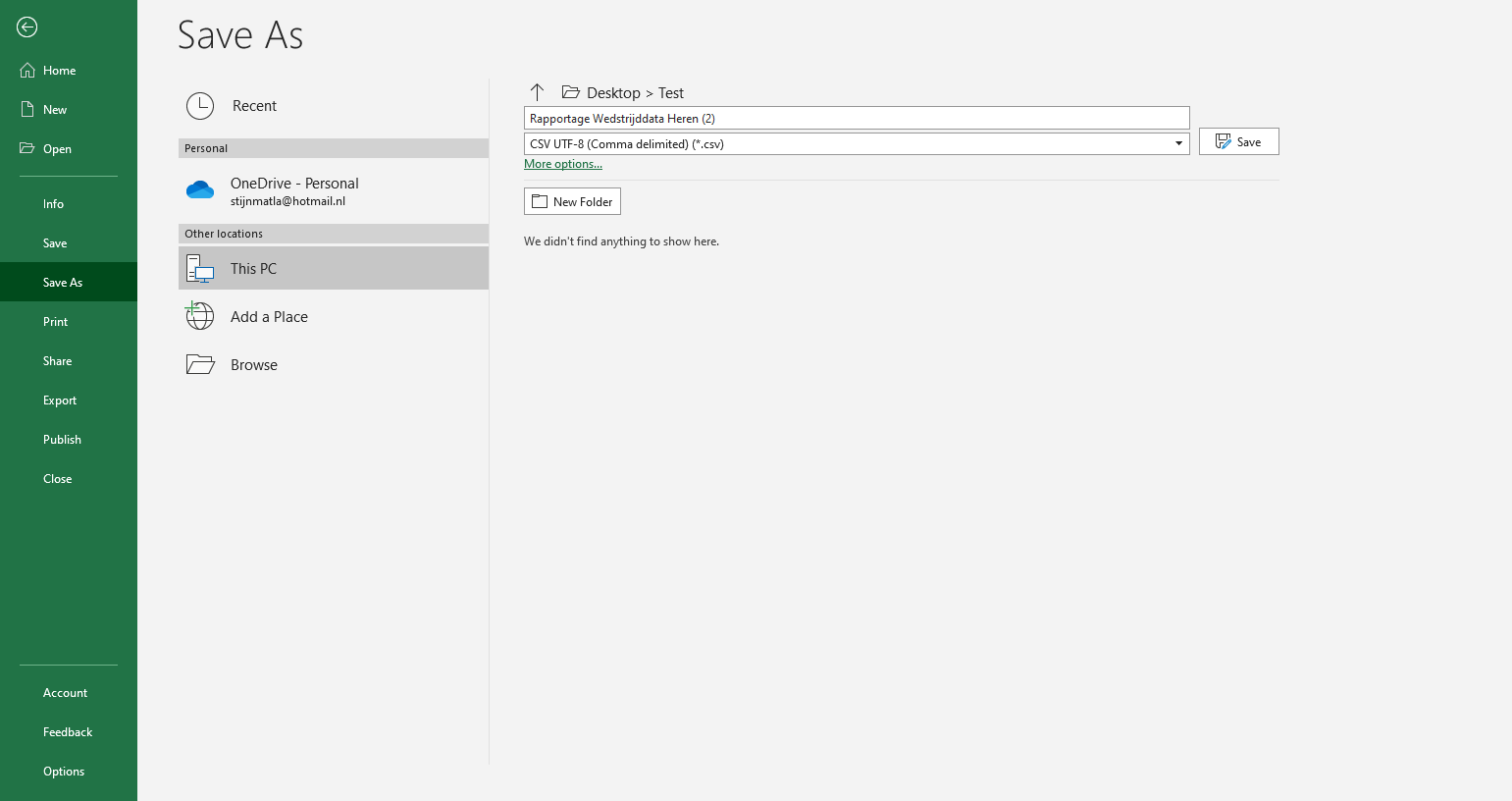
In dit hoofdstuk zal er worden uitgelegd hoe nieuwe data bestanden kunnen worden geüpload in SQL omgeving waarmee het power BI dashboard is verbonden. Dit zal gedaan worden op de volgende website: <https://bd-tennis-1.glitch.me/>

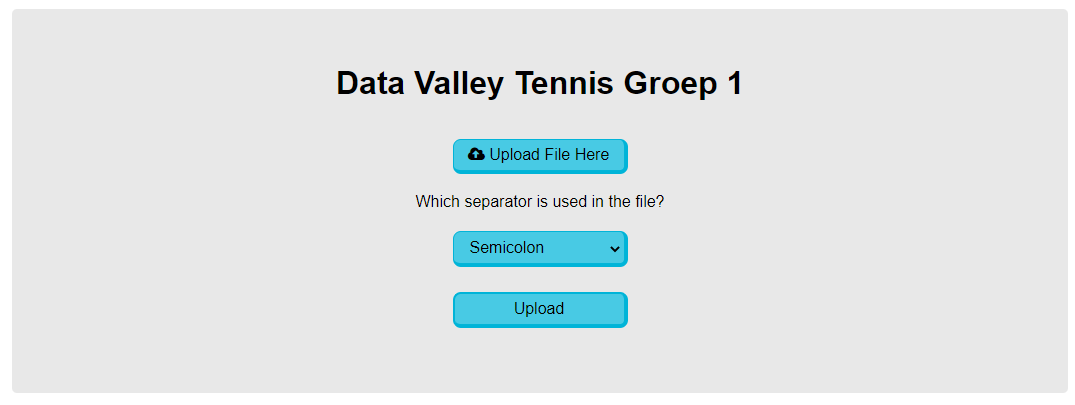
Omdat het voor het proces van belang is dat er een CSV bestand wordt gebruikt zullen we eerst van het bestand een CSV bestand moeten maken. Dit kunnen we doen in Excel. We doet dit door te navigeren naar “file” of “bestand”.

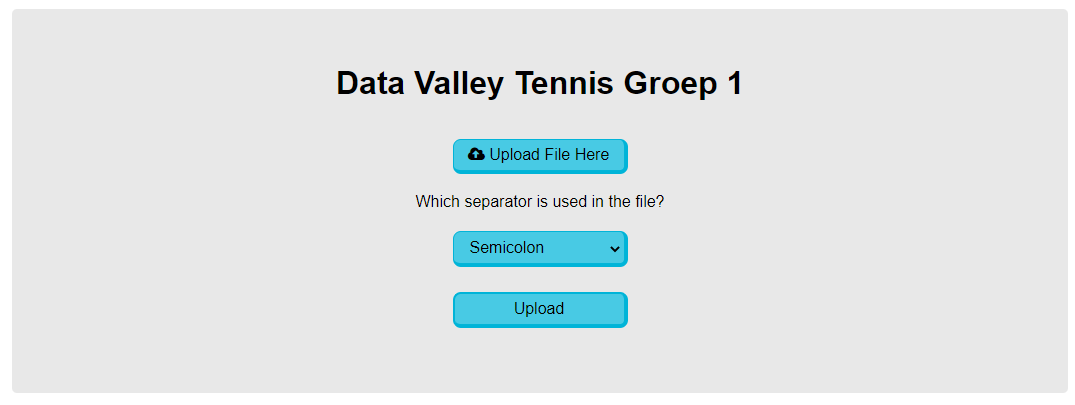


Zodra het volgende scherm is geopend kunnen we “Save As” of “Opslaan Als” selecteren.

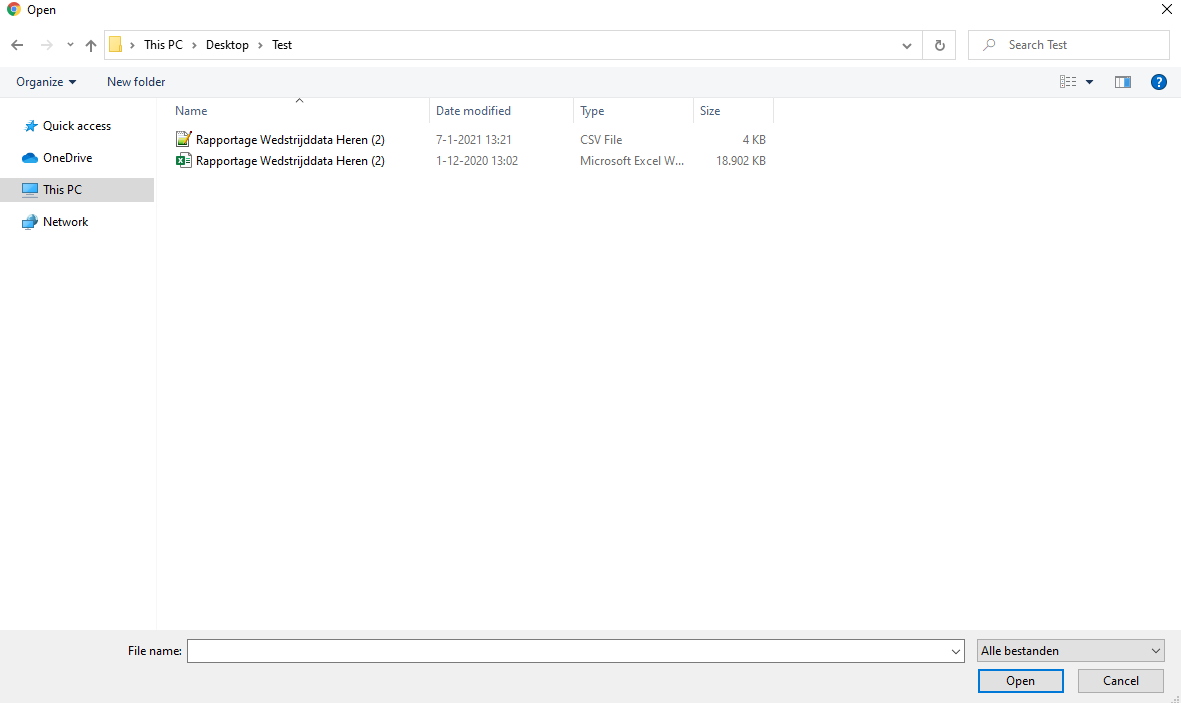
Vervolgens kunnen we in het volgende scherm het bestandstype aanpassen naar de optie CSV UTF-8. En vervolgens dit bestand opslaan.



Wanneer de website(<https://bd-tennis-1.glitch.me/>) wordt geopend zal de gebruiker het volgende scherm worden getoond:

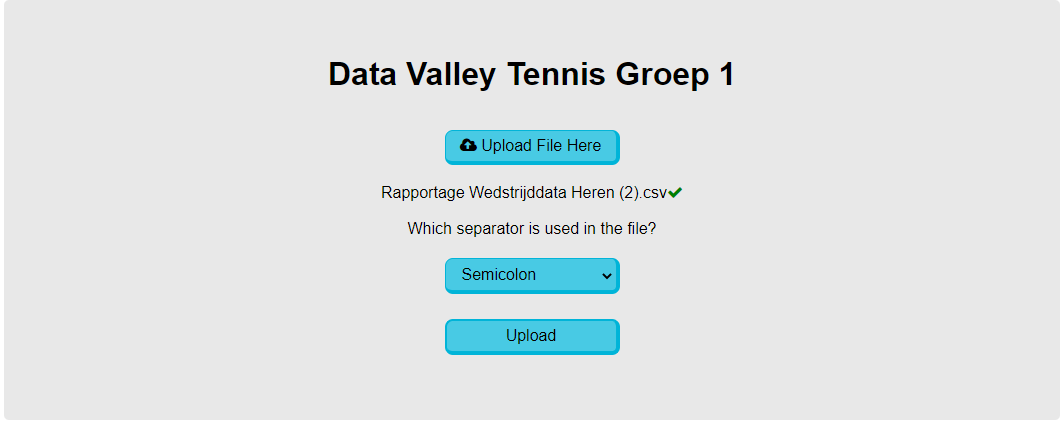
Om een bestand te uploaden op de website zal de knop “Upload File Here” moeten worden ingedrukt:

Nadat deze knop is ingedrukt wordt een venster geopend met uw bestanden, hier kunt u het nieuwe bestand selecteren en vervolgens uploaden op de website door op open te klikken.



Nadat de data is geselecteerd moeten we aangeven welke separator is gebruikt in het bestand, dit is een “semicolon” wanneer er een Nederlandse versie van Excel wordt gebruikt of een “comma” wanneer het over een Engelse versie gaat.

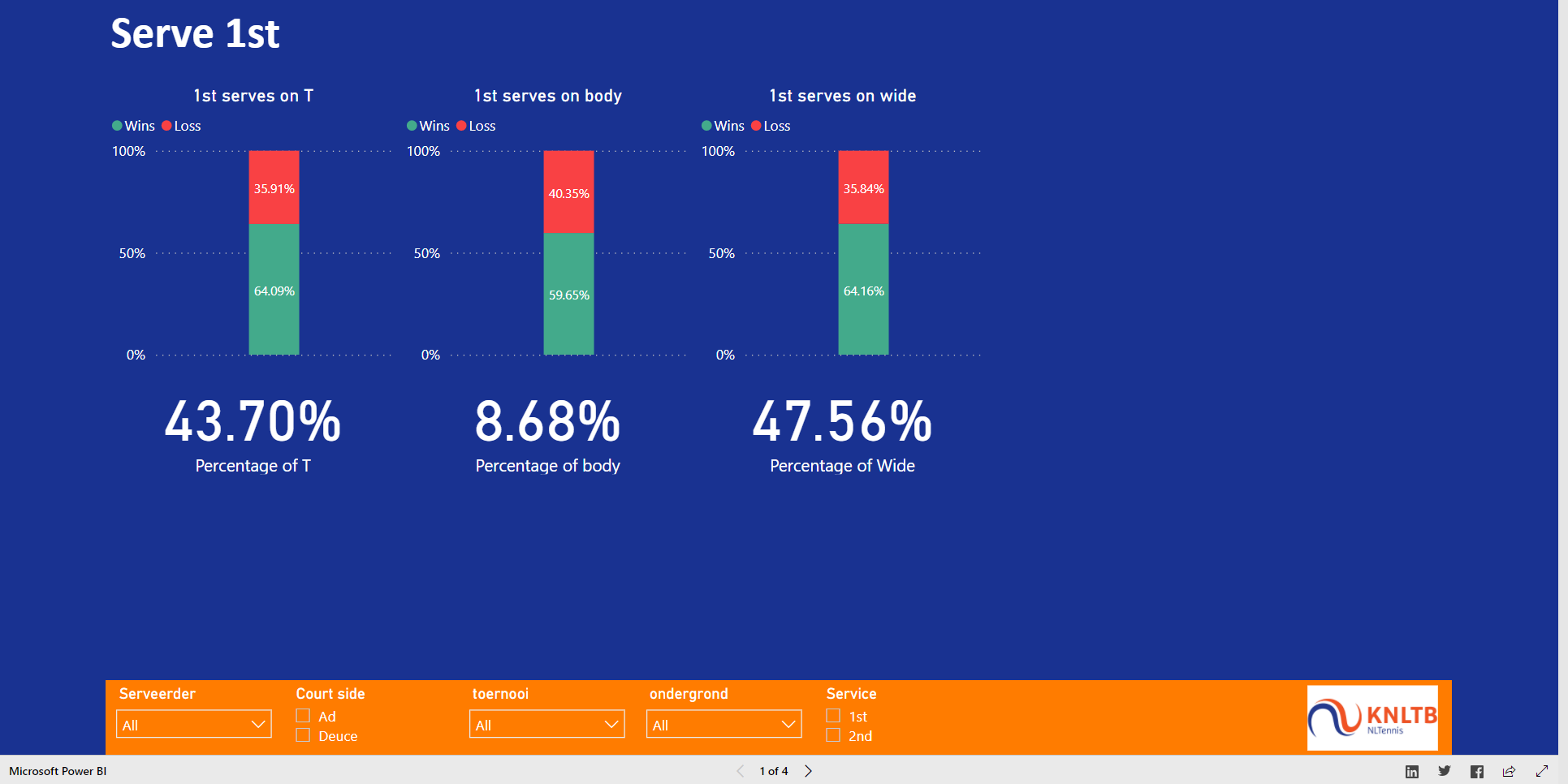
Zodra het bestand is geselecteerd en de juiste separator is gekozen kan het bestand geüpload worden. Dit kunnen we doen door op de “Upload” knop te drukken.



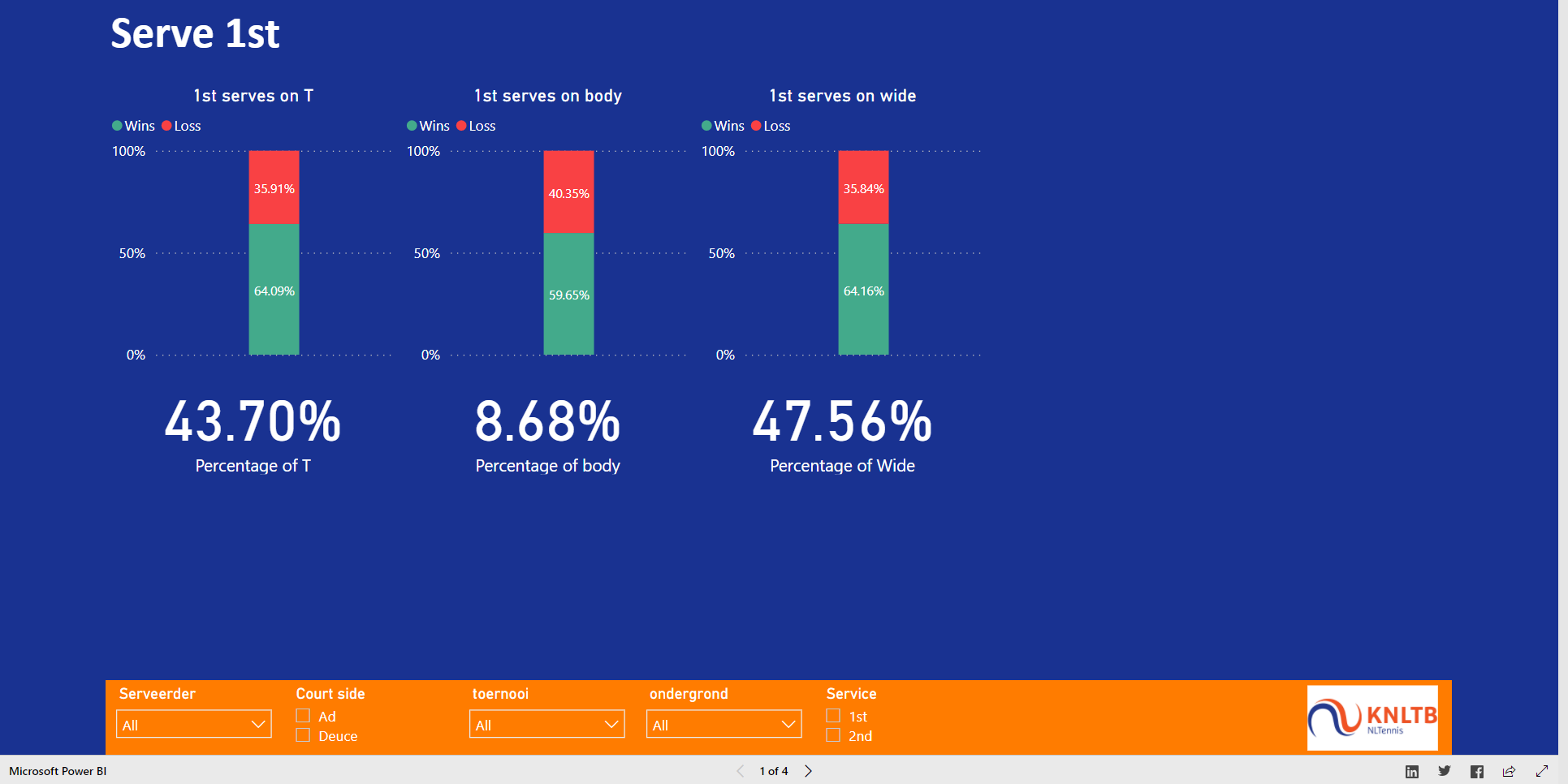
# Power BI

Om inzicht te krijgen in Power BI is er een link gecreëerd waarop er snel toegang worden verkregen tot een viertal grafieken: <https://app.powerbi.com/view?r=eyJrIjoiYmFiNmEzZmEtMTM3My00ZDkwLWJmZmQtYjA1ZmY4NTI1MDE1IiwidCI6IjA5MDdiYjFlLTIxZmMtNDc2Zi04ODQzLTAyZDA5Y2ViNTlhNyIsImMiOjh9>

Er kan gewisseld worden tussen de verschillende grafieken door de pijltjes knoppen onderaan het scherm te gebruiken.



Per pagina worden er relevante filters gebruikt om toegang te krijgen tot de gewenste inzichten. Deze zijn te vinden onderaan de pagina.



# Technische Handleiding

Big Data C1 Team 1

Youp Schaefers 500738283

Sander van de Kamp 500734351

Stijn Matla 500783587  
Arvid de Haas 500772805 11-1-2021

# Introductie

Dit document is opgesteld voor de technische staf van () en mogelijke studenten die dit project verder gaan ontwikkelen. Het document heeft als doel om de werkwijze en methodieken toe te lichten, zodat een begrip van de structuur van het project kan worden gevormd. Het document is opgedeeld in verschillende secties waar in elke sectie een deel van de structuur en de onderliggende connecties worden uitgelicht. De secties staan in chronologische volgorde van beginpunt tot eindpunt.

# Inhoudsopgave

[**Introductie**](#_qi0du4af8eny) **2**

[**Inhoudsopgave**](#_vfoco98qsapz) **3**

[**Data bronnen**](#_hs5xtn4fzg37) **4**

[**Python**](#_ss348ob2wmlf) **5**

[Cleaning](#_h9bs9avu0i0) 6

[Kolomnamen](#_9qia2ai4qbtn) 6

[Spelersnamen](#_77u9au4eehve) 6

[Scores](#_sy82i8ek9hv4) 6

[Tekst naar nummers](#_pmxl6uv0nisk) 6

[Python to database](https://docs.google.com/document/d/1E_Z95Pql0i9Pf5_0Q4E-_le7vxYaWFNdtSN1_r7Wbow/edit#heading=h.owp5qfvjihyc) [7](https://docs.google.com/document/d/1E_Z95Pql0i9Pf5_0Q4E-_le7vxYaWFNdtSN1_r7Wbow/edit#heading=h.owp5qfvjihyc)

[**SQL**](#_i1larm43t552) **8**

[Database](#_80zetibxduzu) 8

[Hoofdtabel](#_755h8o58eiwv) 8

[Log Tabel](#_d0dtlbj1fkrr) 8

[Updatedata tabel](#_6ins69ddwefz) 8

[Procedures](#_movcumltsiv1) 8

[Update](#_7dp50d5l5ymw) 8

[Insert](#_dr00yyqata4c) 9

[Add](#_496b7ih66l90) 9

[Logging](#_ljsiq7k4qdzx) 9

[**Power BI**](#_ifnt9teirjvh) **10**

[Data](#_k1fz6cdg428x) 10

[Berekeningen](#_il4yaksvwkte) 10

[Grafieken](#_2ua5isr6whcm) 10

[Slicers](#_ssj0vmamxjva) 10

# 

# Data bronnen

Vanuit de KNLTB is een Excel (.xlsx) workbook aangeleverd. In dit workbook wordt data verzameld uit onder andere Dartfish. Vervolgens wordt in dit workbook statistische analyses gedaan aan de hand van verscheidene filters.

Tijdens dit project worden deze analyses uitgevoerd door PowerBI, dit zorgt ervoor dat alleen de sheet ‘DATA’ wordt gebruikt voor verdere verwerking.

De ‘DATA’ sheet wordt geëxporteerd naar een .csv bestand en vervolgens via Python ingeladen.

# 

# Python

Python is een programmeertaal die veel wordt gebruikt bij Data Science en Data Engineering. Er zijn vele libraries die het makkelijk maken om data te verzamelen, verwerken en op te slaan naar verschillende databases.

De libraries die zijn gebruikt in dit project zijn:

**Pandas**

*‘Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.’*

Pandas wordt in dit project gebruikt voor het maken van zogenaamde DataFrames vanuit het .csv bestand. Deze DataFrames zorgen ervoor dat de data makkelijk kan worden aangepast voor gebruik.

Meer informatie kunt u vinden in de [Pandas documentatie](https://pandas.pydata.org/docs/)

**Numpy**

*‘The fundamental package for scientific computing with Python.’*

Numpy wordt in deze context vooral gebruikt om missende gegevens in een dataset te vinden.

Meer informatie kunt u vinden in de [Numpy documentatie](https://numpy.org/doc/)

**SQLAlchemy**

*‘SQLAlchemy is the Python SQL toolkit and Object Relational Mapper that gives application developers the full power and flexibility of SQL. It provides a full suite of well known enterprise-level persistence patterns, designed for efficient and high-performing database access, adapted into a simple and Pythonic domain language*.*’*

SQLAlchemy wordt in de context van het project gebruikt om een connectie te leggen met een database en hier onze data naar toe te sturen.

In het project wordt een MySQL Database gehost op HvA domein: oege.ie.hva.nl

Meer informatie kunt u vinden in de [SQLAlchemy documentatie](https://docs.sqlalchemy.org/en/14/)

## Cleaning

Het ‘cleanen’ van de data gebeurd nadat de data vanuit CSV naar een pandas dataframe omgezet is, deze DataFrames zijn ervoor ontwikkeld om snel en efficiënt met grote hoeveelheden te werken.

### **Kolomnamen**

De kolom namen worden genormaliseerd voor de database, waarbij de hoofdletters vervangen worden door kleine letters, de spaties vervangen worden door lage streepjes en de lege ruimtes verwijderd worden.

### **Spelersnamen**

De namen van de spelers zijn soms verkeerd gespeld, missen één letter of hebber er één te veel, om dit op te lossen is er een “sequency matcher” gebruikt. Deze sequency matcher gebruikt een richtlijn waarbij een naam die voor 90% hetzelfde is als een andere naam omgezet wordt naar de andere naam.

### **Scores**

Om de scores volledig bruikbaar te maken in Power BI was het nodig om de scores op te splitsen, hierbij werden de voor en tegen standen in aparte kolommen opgeslagen met de passende naamgeving erbij. Om dit voor elkaar te krijgen moest er eerst consistentie komen in de notering van de scores, hierbij moesten er voor alle 0 waardes een 0:0 komen en voor de waardes waarbij scores opgeslagen werden met een ‘-’, werden ze vervangen met een ‘:’. Als deze waardes opgeschoont waren, werden ze gesplitst in voor en tegen score kolommen.

### **Tekst naar nummers**

Het excel bestand heeft veel nummers opgeslagen als teksten, ook wordt er regelmatig een extra icoon gebruikt om aan te geven dat het getal groter dan een specifiek aantal is. Om te zorgen dat er gerekend kan worden met de getallen worden de extra icoontjes verwijderd en alle teksten met getallen omgezet naar getallen.

## 

# SQL

## Database

De database is onderverdeeld in drie losse tabellen; de hoofdtabel waar alle data in wordt opgeslagen van de tennissers, de log tabel die bijhoudt wanneer de hoofdtabel wordt geüpdate en de hoeveelheid die waardes er in de hoofdtabel zitten en de update tabel waar de nieuwe data naar wordt gestuurd voordat het verwerkt wordt.

### **Hoofdtabel**

De hoofdtabel bestaat uit alle berekeningen en waardes die in het data tabblad van excel zitten, deze waardes worden opgeslagen geïdentificeerd aan de hand van de naam en positie van de rijen. Daarnaast wordt er aan de hand van drie sleutel waardes geïdentificeerd of de rijen actieve waardes zijn, wanneer ze geüpload en wanneer ze gecreëerd zijn, hierdoor worden oude records die geupdate worden niet verwijderd en kan de rest van de applicatie zien welke niet waardes niet gebruikt moeten worden.

### **Log Tabel**

In de log tabel wordt er alleen bijgehouden wanneer de hoofdtabel geüpdate is en hoeveel waardes de hoofdtabel heeft na de update, hierdoor is er bij te houden of er ergens iets fout gegaan is en wat er ongeveer gebeurd is.

### **Updatedata tabel**

De update tabel is een lege tabel die alleen bestaat voor het vasthouden van de nieuwe data die de hoofdtabel in moet, deze tabel wordt voor elke update geleegd en opnieuw gevuld.

## **Procedures**

In de MySQL database staan een aantal gemaakte procedures die nodig zijn om de data volledig in te kunnen voegen in de hoofdtabel van de database, deze procedures kunnen van buitenaf aangeroepen worden als de nieuwe data in update tabel zit.

### **Update**

De update procedure wordt gebruikt om in de update tabel voor elke rij te kijken of het al in de hoofdtabel zit, als het er al in zit worden alle waardes van beide rijen met elkaar vergeleken. Indien één van de waardes anders is wordt de insert procedure hieronder aangeroepen en worden de actief en update kolommen gewijzigd, als het hetzelfde is wordt de rij in de update tabel overgeslagen en wordt het proces opnieuw gedaan totdat de gehele update tabel bekeken is. Als dit proces compleet is worden de add en logging procedures aangeroepen, waarna de procedure compleet is.

### **Insert**

De insert procedure wordt gebruikt om een nieuwe rij toe te voegen in de hoofdtabel, bij deze procedure wordt een nieuwe rij toegevoegd in de hoofdtabel voor een rij die gewijzigd is.

### **Add**

De add procedure voegt alle rijen toe uit de update tabel die nog niet bestaan in de hoofdtabel, hierbij wordt er voor elke rij toegevoegd dat het actief is en wanneer het gecreëerd is.

### **Logging**

Bij de logging procedure wordt in de log tabel een nieuwe rij aangemaakt elke keer dat de update procedure wordt aangeroepen, het voegt in de log tabel een rij toe met de totale hoeveelheid rijen die in de hoofdtabel zitten en de datum en tijd dat het is aangepast.

# 

# Power BI

### **Data**

Binnen Power BI is allereerst een connectie gemaakt met een SQL server met daarop de relevante data. Deze data is geïmporteerd de data die wordt gebruikt kan binnen Power BI gezien worden door in de linkerbalk “data” te selecteren.

### **Berekeningen**

De grafieken die worden getoond zijn opgebouwd aan de hand van enkele berekeningen deze berekeningen zijn terug te vinden in de rechterbalk. Om inzicht te krijgen in hoe een berekeningen is opgebouwd kan er gedubbelklikt worden op de berekeningen. Daar wordt door het gebruik van namen getoond wat de berekening doet.

### **Grafieken**

De grafieken zijn opgebouwd vanuit de eerder genoemde berekeningen in samenwerking met filters. Waar mogelijk zijn berekeningen vermeden om de complexiteit zo laag mogelijk te houden. Door op een grafiek te drukken kan gezien worden welke waardes per grafiek worden gefilterd om tot de gewenste waardes te komen. Deze filters zijn specifiek per grafiek.

### **Slicers**

Aan de onderkant van iedere pagina is ook een set met slicers gemaakt (voor gemak voor de eindgebruiker noemen we deze op de front end filters). Deze slicers kunnen snel en gemakkelijk een specifiek deel van de data filteren. Deze slicers bevatten geen berekende waardes en zijn aanwezig om de gebruiker te helpen om de precieze data waar hij/zij op zoek naar is te vinden.