Data Scientist
Technical Test

2019-04-10

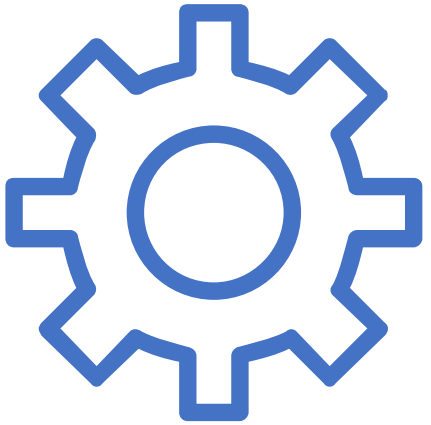
ADRIÁN OTERO RODRÍGUEZ

AGENDA

1. **Definición del problema**
2. **Solución propuesta**
3. **Metodología**
4. **Resultados y conclusiones**

Definición del problema

Objetivos



¿Qué idiomas son los más adecuados para traducir el FIFA? (Localizar el juego)

ANÁLISIS Y TRANSFORMACIÓN DE DATOS

Carga de datos en una BBDD

Análisis de calidad del dato

Análisis exploratorio

VISUALIZACIÓN

Construcción de un dashboard focalizado en los países e idiomas

Perfilar equipos y jugadores

ANÁLISIS DE SENTIMIENTOS

Extracción información desde las redes sociales

Análisis de sentimientos en las distintas localizaciones

Definición del problema

Fuentes de datos

INFORMACIÓN SOBRE LOS JUGADORES DEL FIFA 18

Nacionalidad

Edad

Club

Valor y salario

Atributos

INFORMACIÓN SOBRE LAS LENGUAS DE CADA PAÍS

Información sobre países

Idiomas hablados en cada país

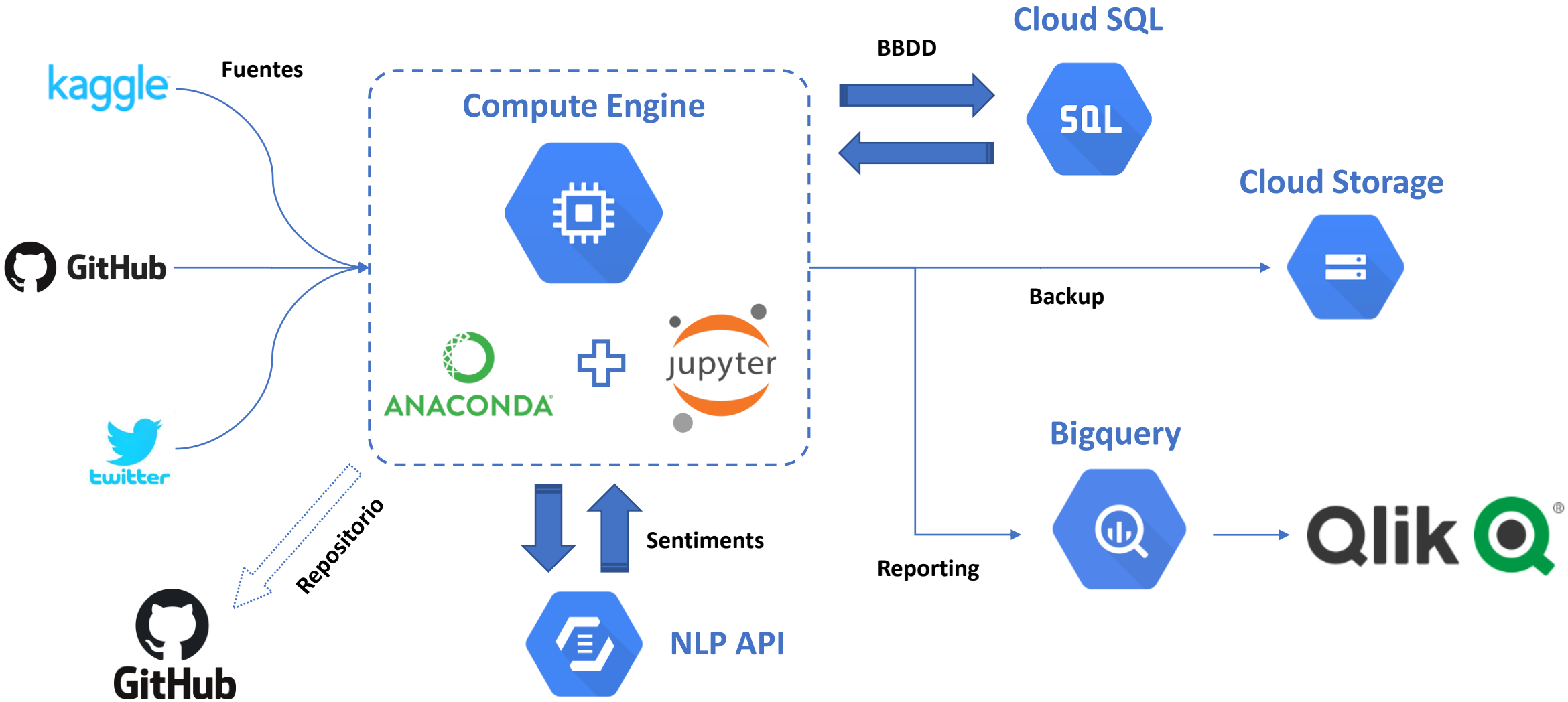
COMENTARIOS EN REDES SOCIALES SOBRE EL FIFA 20

Comentarios con el hashtag #FIFA20

Preferiblemente geolocalizados

Solución propuesta

Arquitectura y componentes

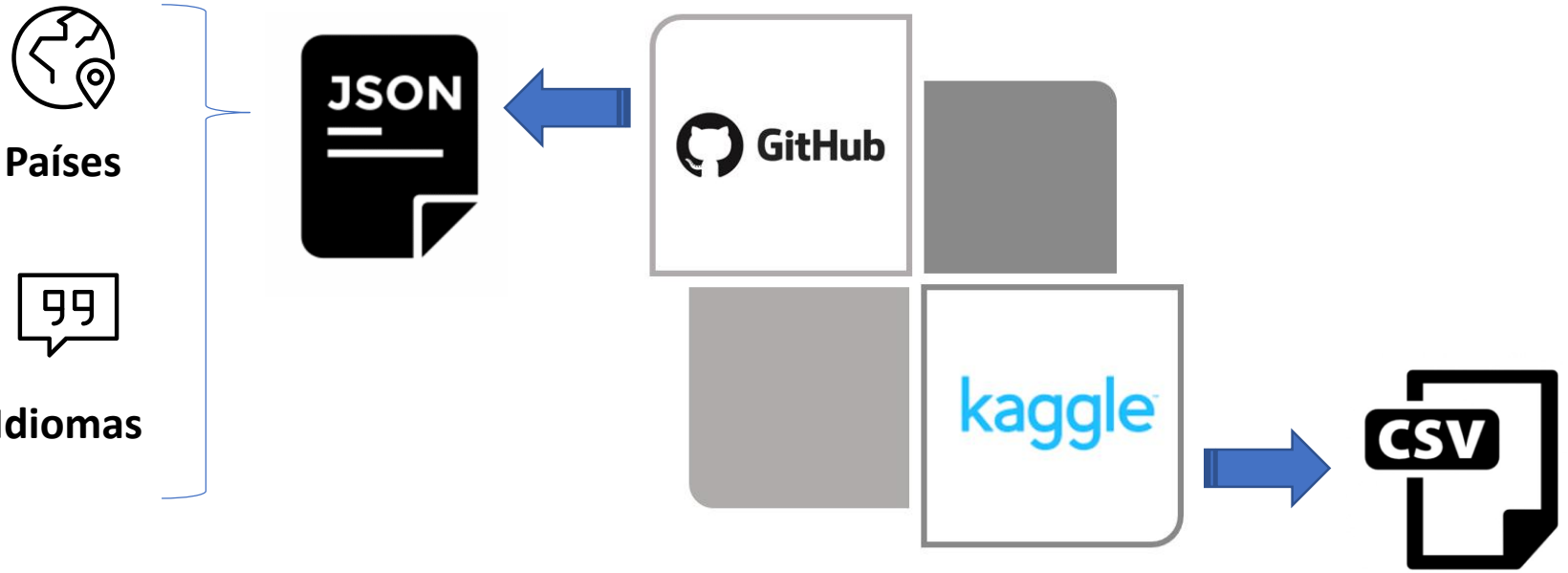


Metodología

Carga de datos

Partiendo de la premisa de que los jugadores prefieren jugar en su primera lengua

Para cada jugador se identifica el idioma principal



- **CompleteDataset.csv**
- PlayerAttributeData.csv
- PlayerAttributeData.csv
- PlayerPlayingPositionData.csv

Cada país puede tener entre 1 – 10 idiomas diferentes

EA SPORTS FIFA 18

Metodología

Calidad del dato - Findings

- Registros **duplicados** (52)
- **Players sin** información de **Club**, son agentes libres con un **Valor de 0€**
- Variables monetarias (**Salary** & **Wage**) con **tipo de datos** erróneo **String** debido a su **formato**, por ejemplo, **€**95.5**M**
- Variables con un alto porcentaje de valores no informados, (> 10%) **missings**. Afecta a las variables:
  - CAM, CB, CDM, CF, CM, LAM, LB, LCB, LCM, LDM, LF, LM, LS, LW, LWB, RAM, RB, RCB, RCM, RDM, RF, RM, RS, RW, RWB, ST
  - Se debe a que cuando **Preferred Positions** = 'GK', estas variables son nulas
- Variables con información irrelevante para el problema: **Photo**, **Flag**, **Club Logo**
- Nacionalidades de jugadores que **no cruzan** con la información de la tabla de países-idiomas, por ejemplo:
  - 'Wales', 'England', 'Bosnia Herzegovina', 'Korea Republic', 'DR Congo', 'Republic of Ireland', 'Northern Ireland', etc.

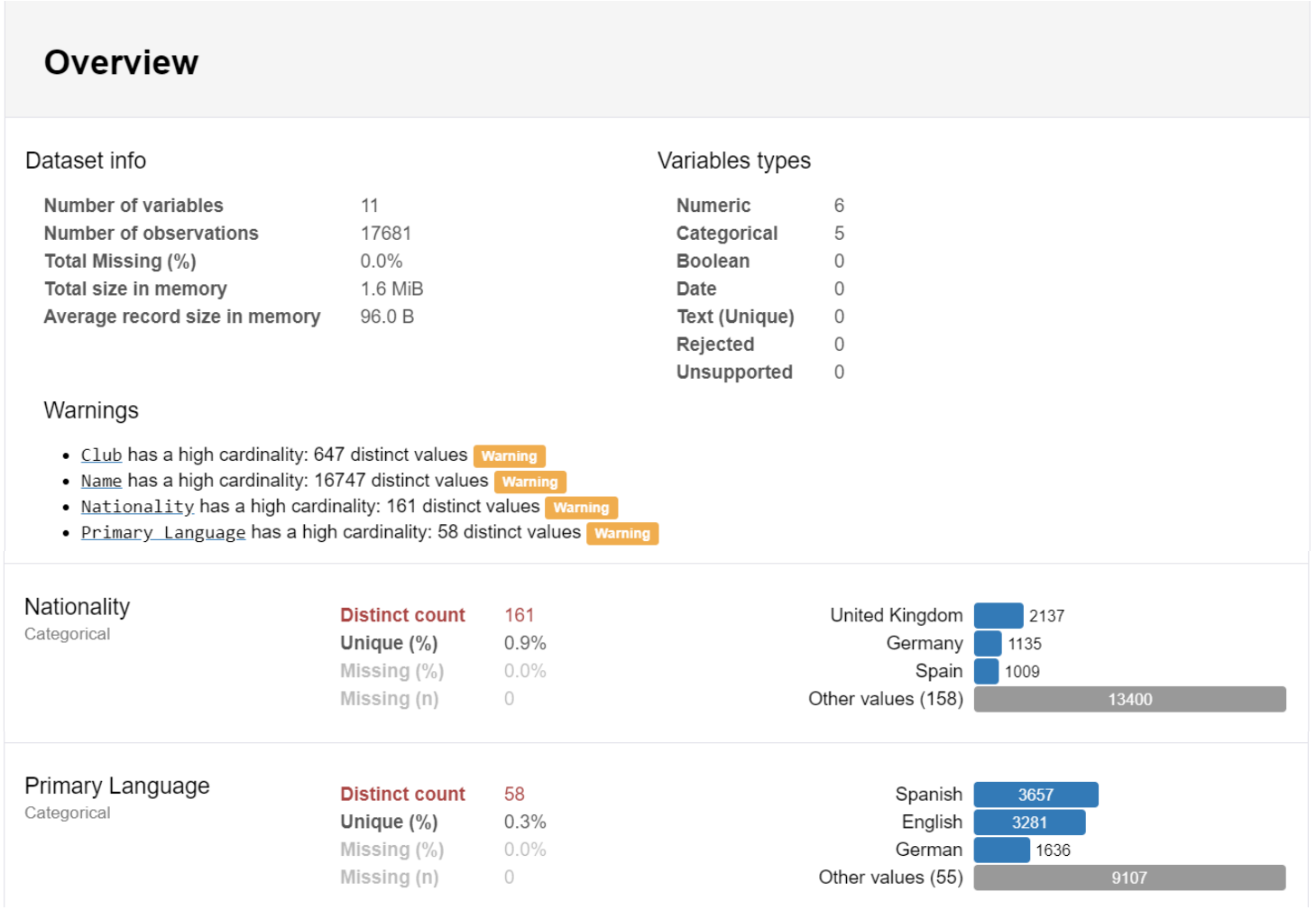
Metodología

Calidad del dato - Report

Finalmente, se ha generado un informe final de calidad del dato, disponible en “Task_1b - Data Quality”

Se han seleccionado las **11 variables más relevantes** en los datos

Se han **eliminado 300** de 17.981 **registros (2%)** por problemas de **calidad**



Overview

Dataset info

Number of variables	11
Number of observations	17681
Total Missing (%)	0.0%
Total size in memory	1.6 MiB
Average record size in memory	96.0 B

Variables types

Numeric	6
Categorical	5
Boolean	0
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

- Club has a high cardinality: 647 distinct values Warning
- Name has a high cardinality: 16747 distinct values Warning
- Nationality has a high cardinality: 161 distinct values Warning
- Primary Language has a high cardinality: 58 distinct values Warning

Nationality
Categorical

Distinct count	161
Unique (%)	0.9%
Missing (%)	0.0%
Missing (n)	0

United Kingdom 2137
Germany 1135
Spain 1009
Other values (158) 13400

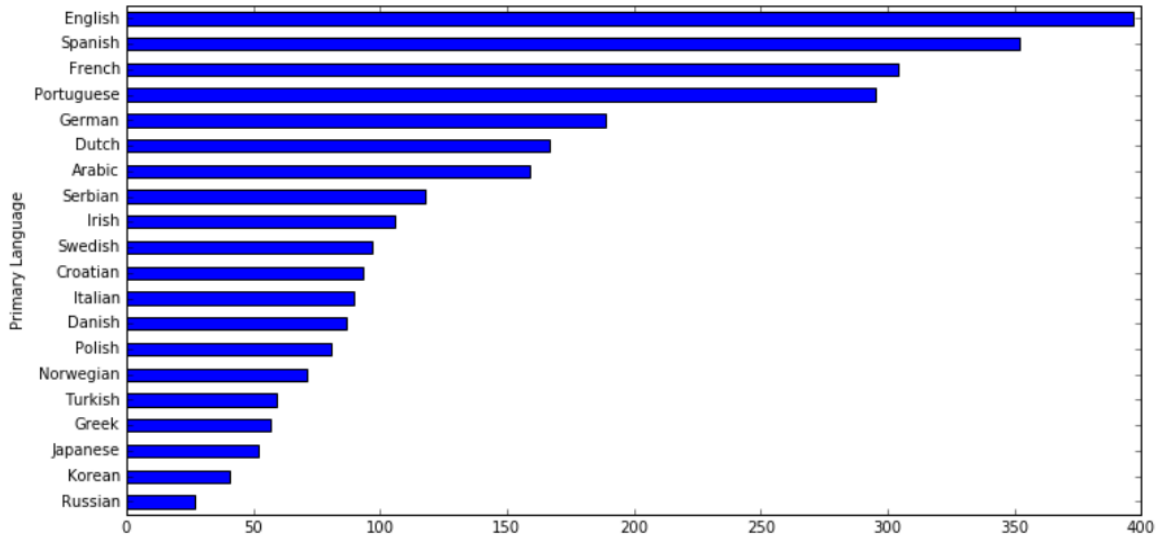
Primary Language
Categorical

Distinct count	58
Unique (%)	0.3%
Missing (%)	0.0%
Missing (n)	0

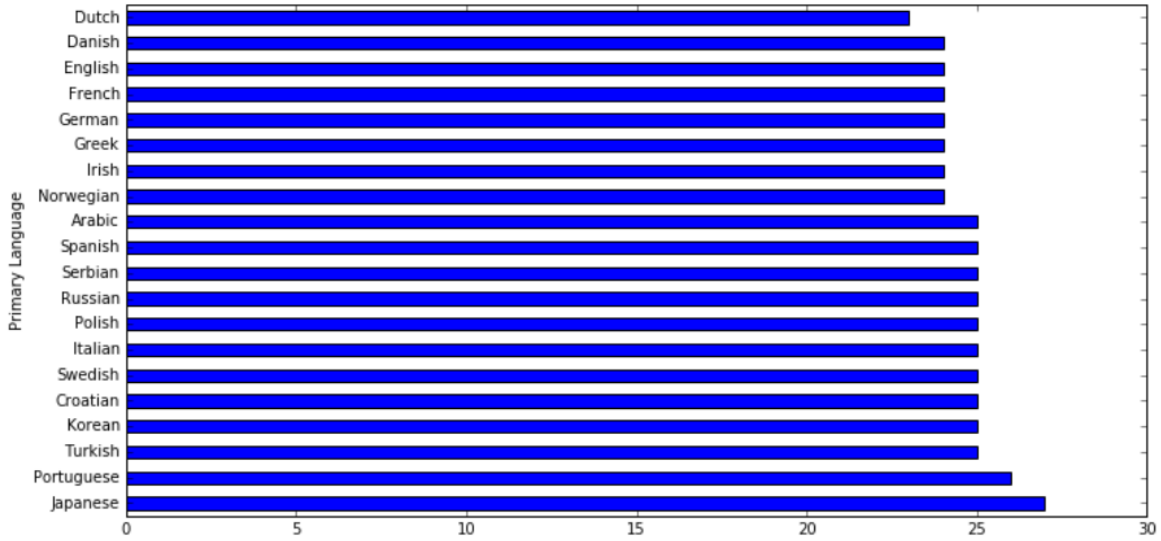
Spanish 3657
English 3281
German 1636
Other values (55) 9107

Metodología
Análisis exploratorio

El inglés, español, francés y portugués son los idiomas que están presentes en una mayor variedad de clubs

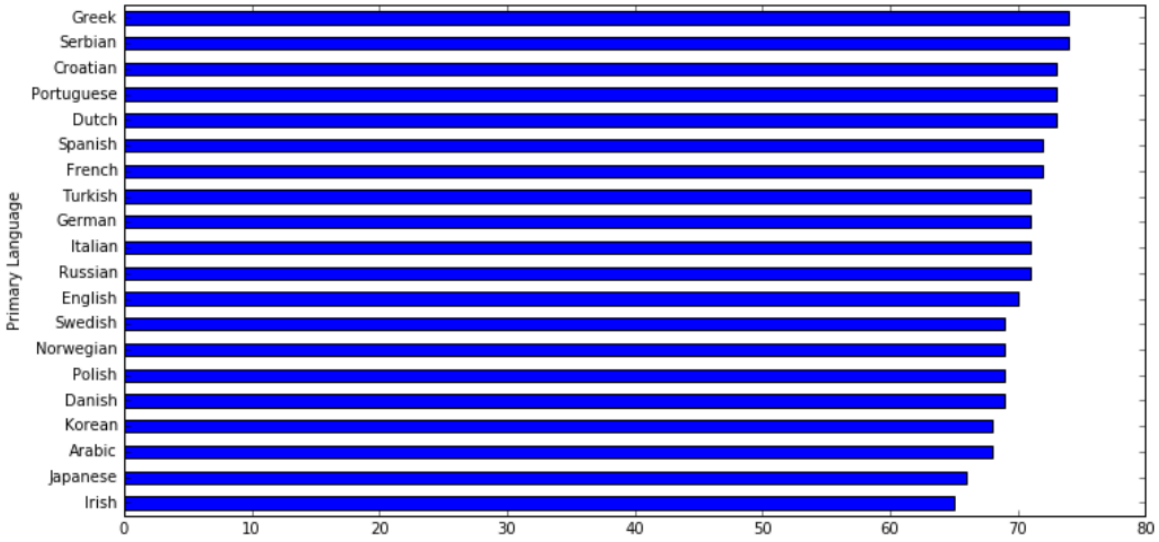


Los jugadores más jóvenes son de habla holandesa, danesa, inglesa, francesa y alemana

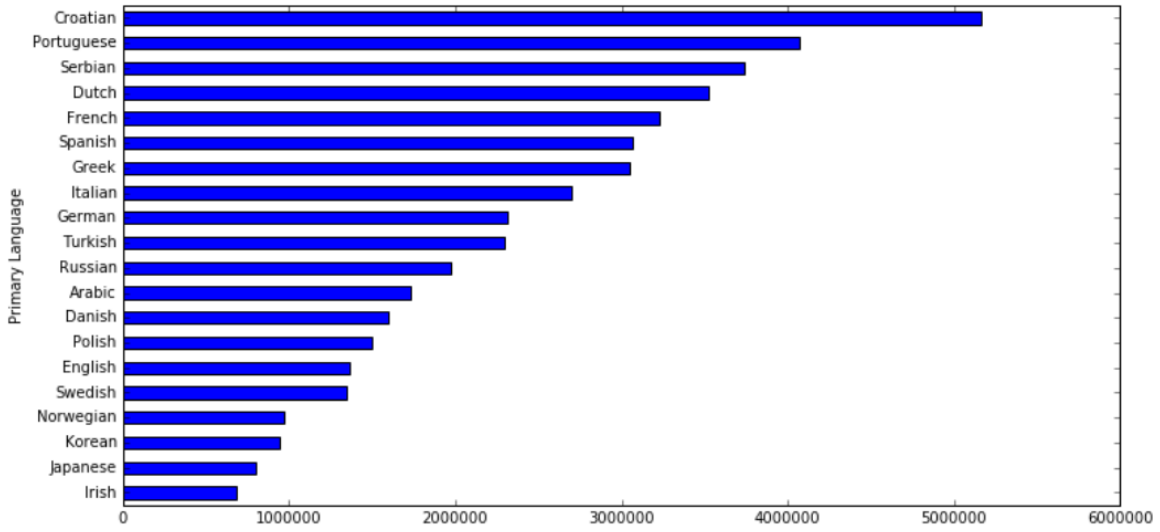


Metodología
Análisis exploratorio

Los jugadores que hablan griego, serbio, croata, portugués y holandés son los que muestran mayor potencial



Los salarios asociados a los idiomas croata, portugués, serbio, holandés y francés tienen mayor valoración

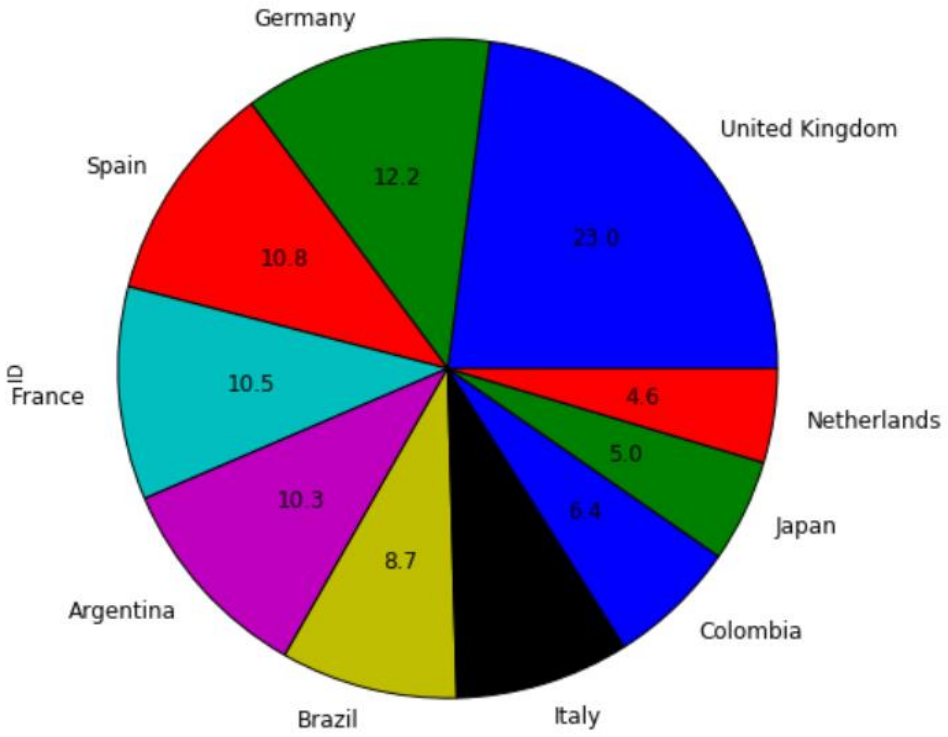


Metodología

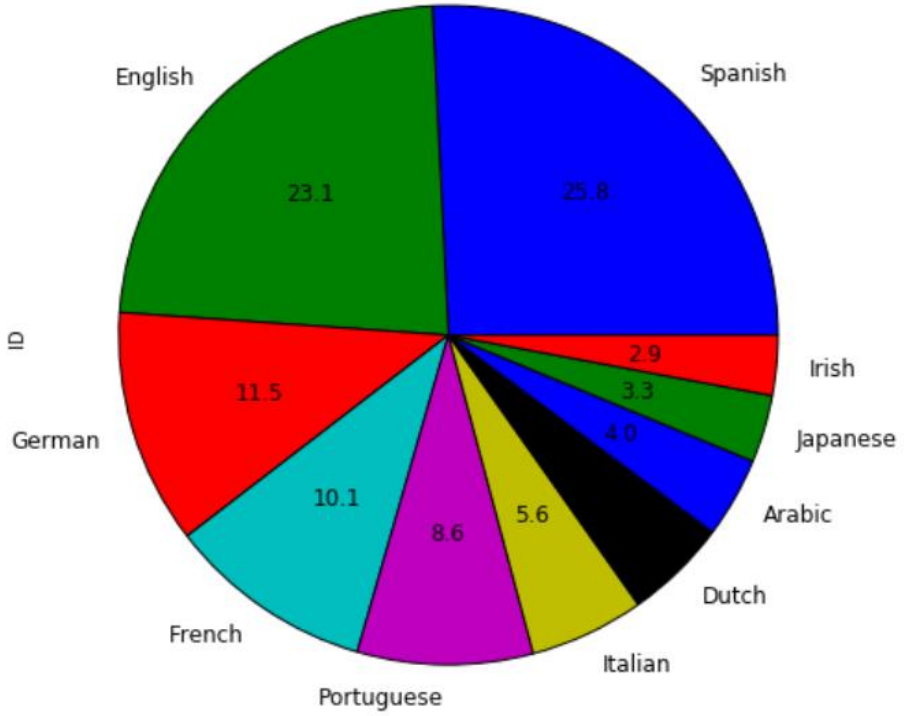
Análisis exploratorio

Reino Unido es la nación con mayor presencia, sin embargo, el Español destaca como primera lengua más hablada

Top Nacionalities



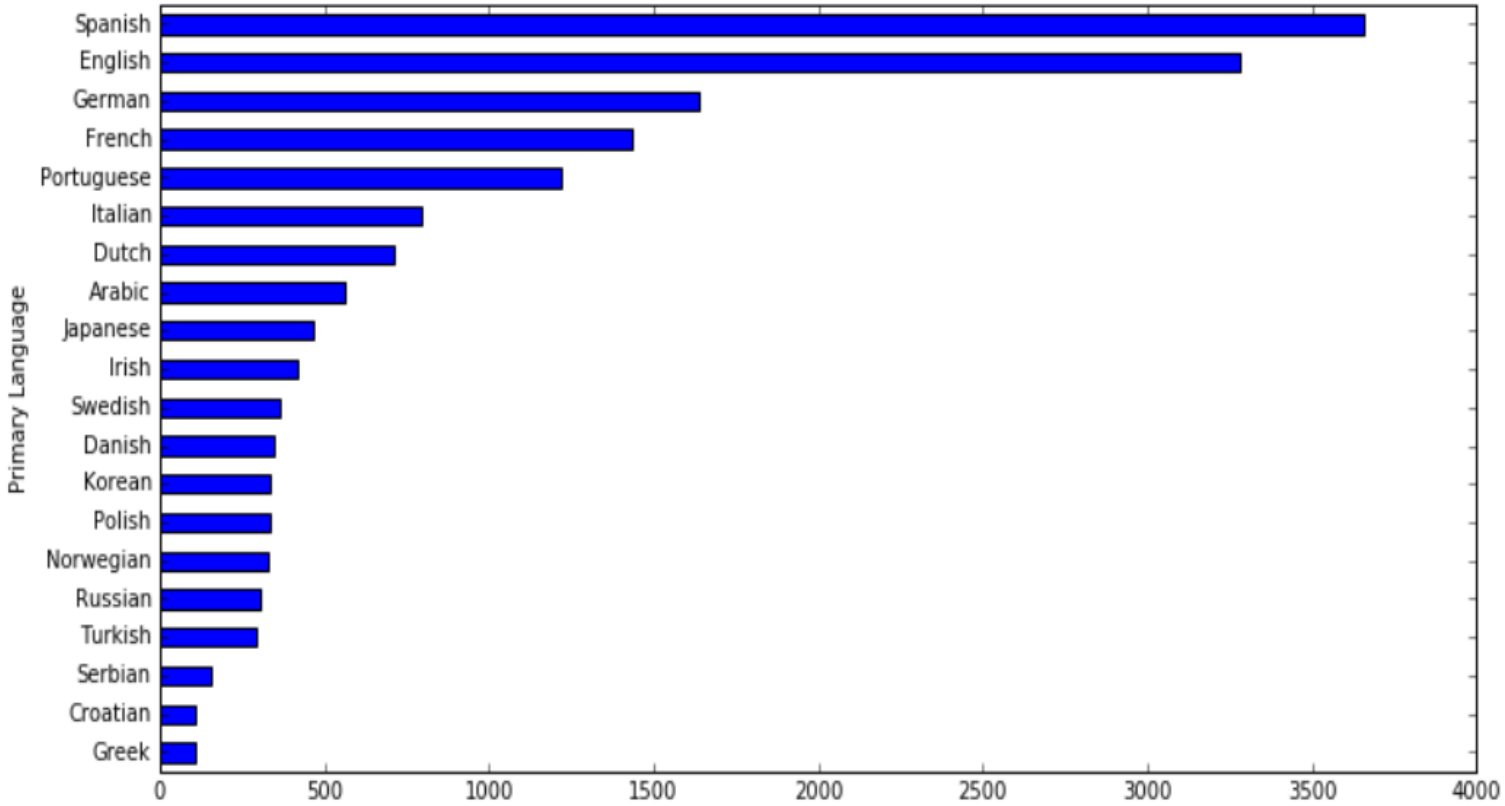
Top Primary Languages



Metodología
Análisis exploratorio

- *Existen numerosos idiomas con un porcentaje muy reducido de observaciones (< 100)*

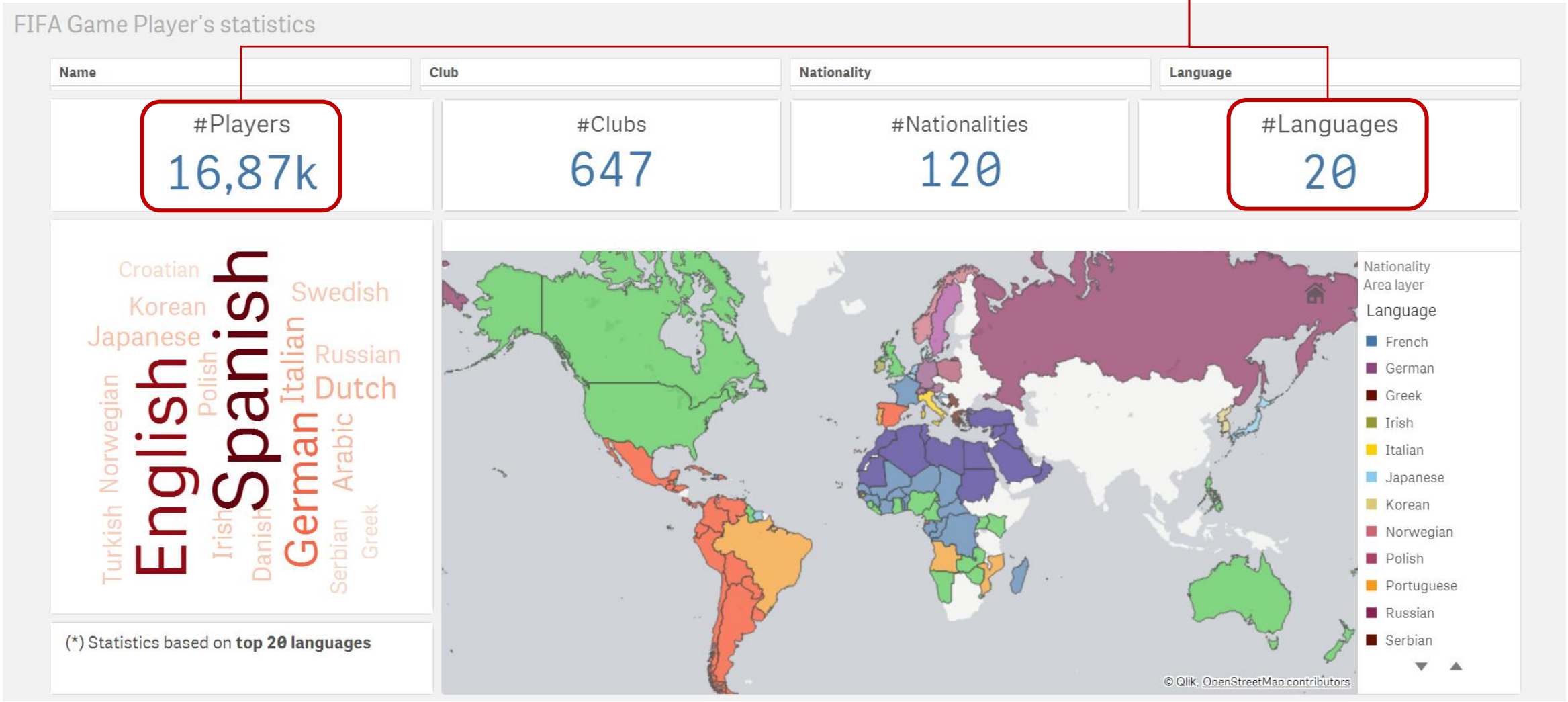
- *Únicamente 20 de los idiomas alcanzan el número mínimo de observaciones (jugadores)*



Resultados y conclusiones

Data Visualization

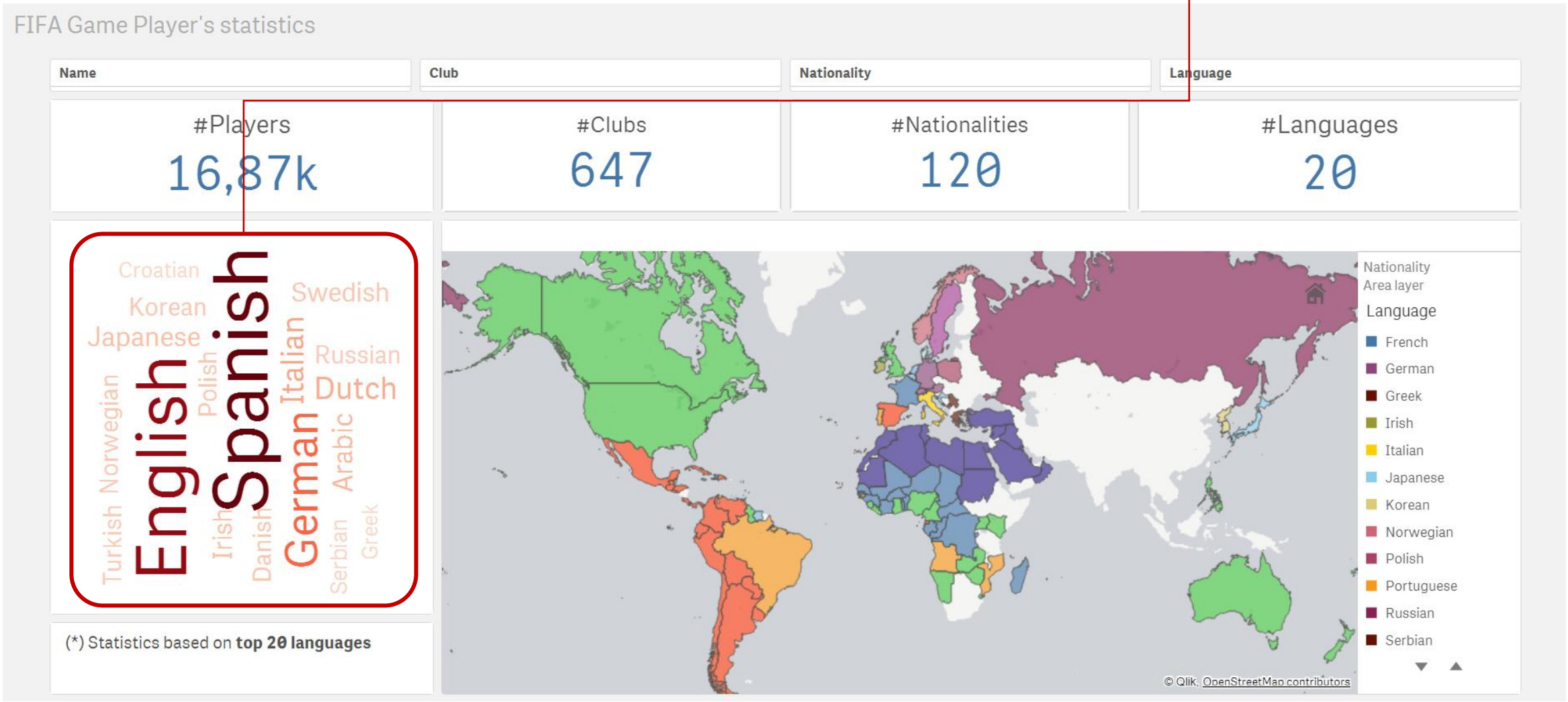
EL 95% de los jugadores hablan alguno de los 20 idiomas más extendidos como lengua principal



Resultados y conclusiones

Data Visualization

Entre estos 20 idiomas, los más hablados son el español, inglés, alemán, francés y portugués

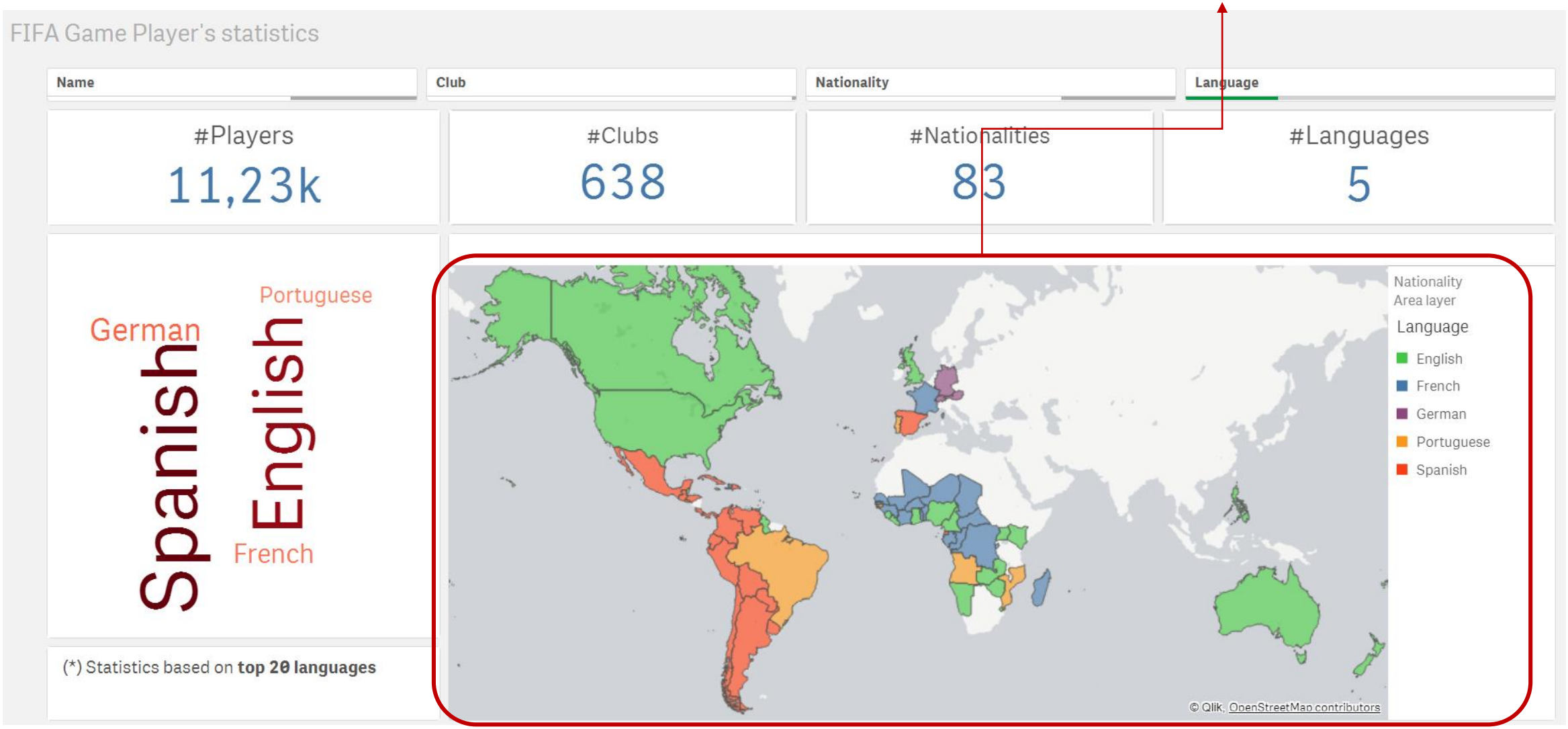


Resultados y conclusiones

Data Visualization

Estos idiomas son la lengua principal en las regiones de América (norte y sur), Europa y sur de África.

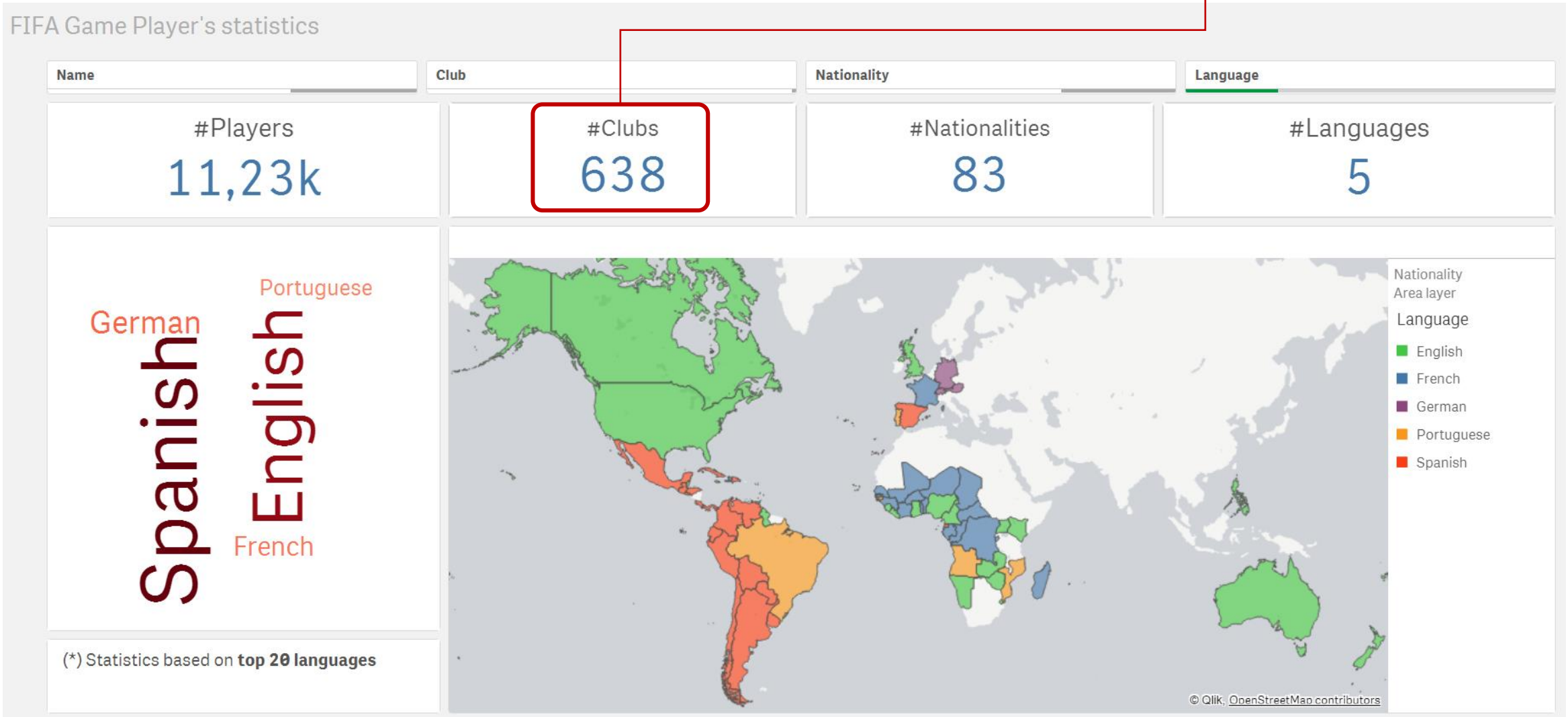
Destaca la escasa presencia de regiones asiáticas



Resultados y conclusiones

Data Visualization

638 de los 647 Club existentes (99%) tienen algún jugadores cuyo idioma principal está entre: español, inglés, alemán, francés y portugués



Resultados y conclusiones

Data Visualization

El español y el inglés son los idiomas más representados, con una gran diferencia frente al resto de idiomas

Details by language

Language | Nationality | Values

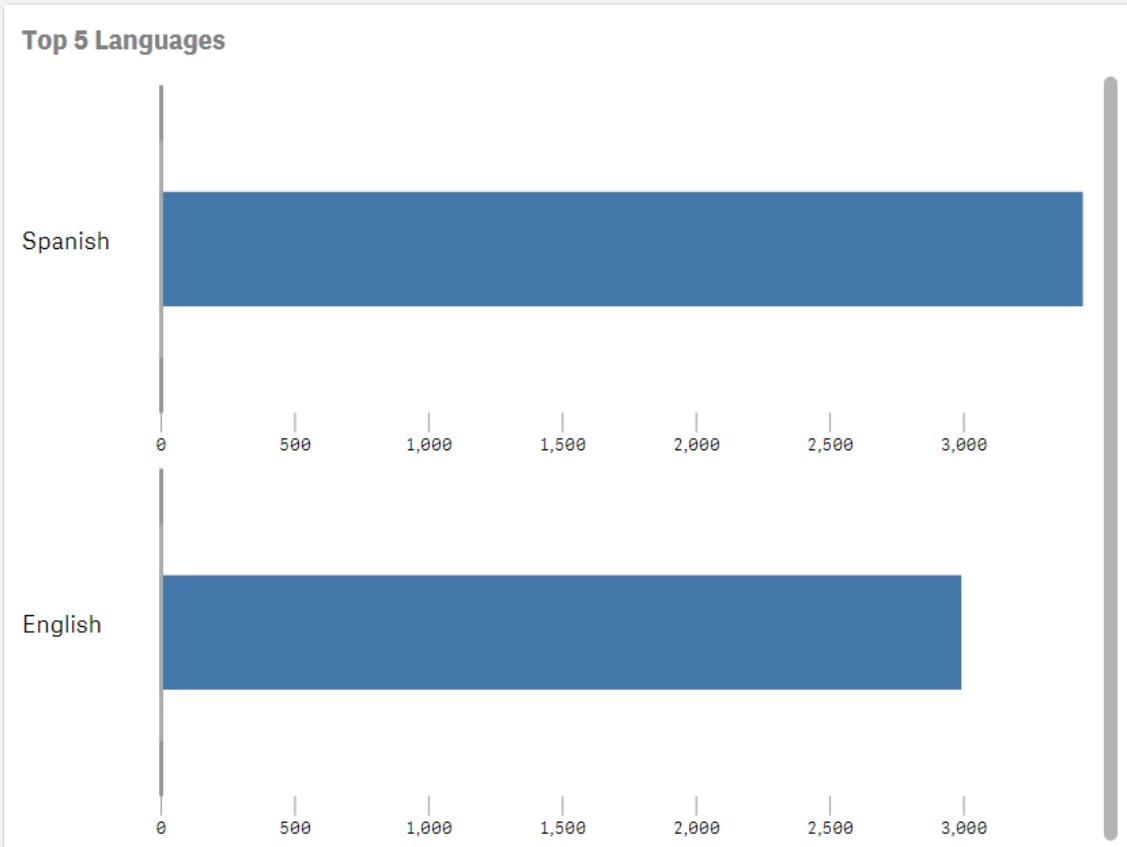
	#Players	Avg Value	Avg Overall	Avg Potential
English	3281	1.367.656 €	64	70
French	1434	3.230.223 €	67	73
German	1636	2.317.796 €	65	71
Portuguese	1222	4.073.830 €	71	73
Spanish	3657	3.067.250 €	68	73

Resultados y conclusiones

Data Visualization

La alta representación de estos idiomas está ligada al numeroso grupo de jugadores procedentes de:

- ***España*** *y* ***Sudamérica*** *(español)*
- ***Reino Unido****, Estados Unidos y Australia (inglés)*



Details by language

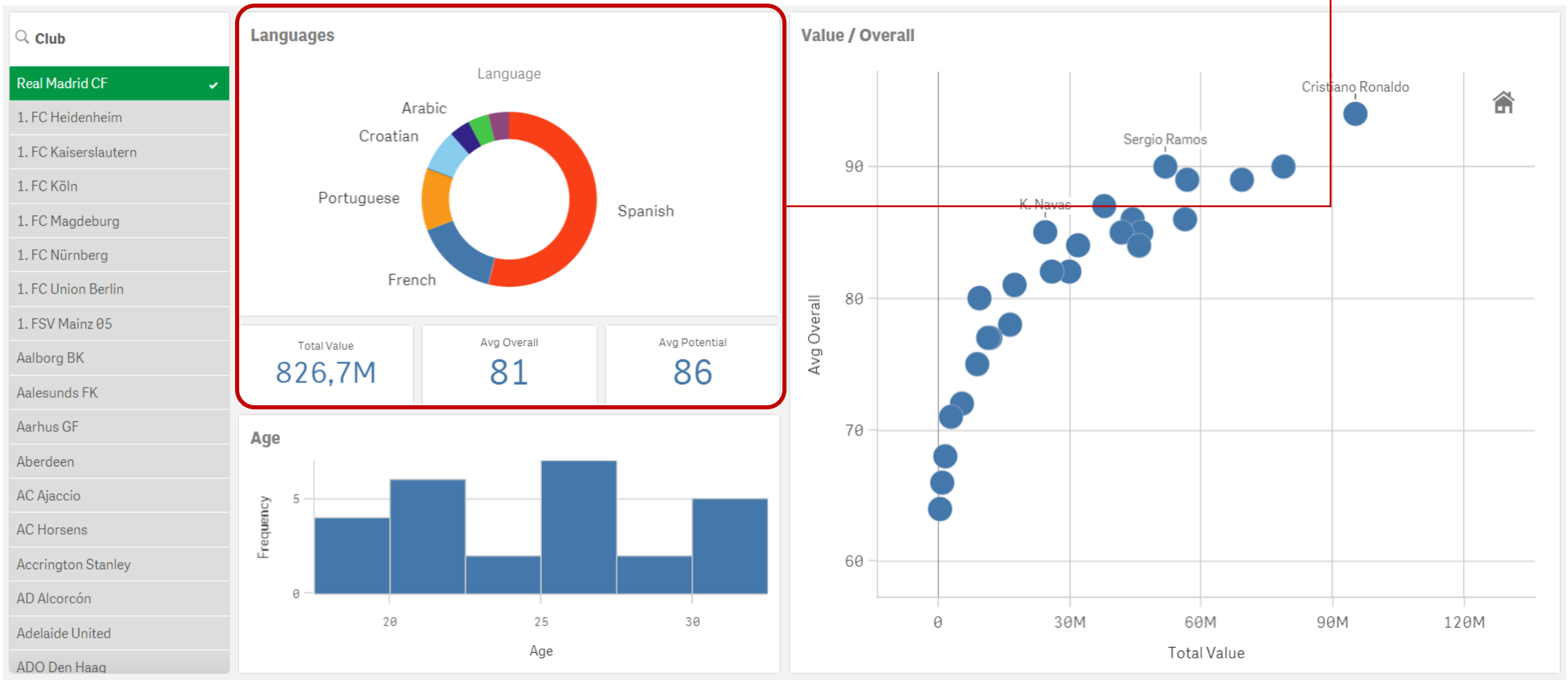
Language | Nationality | Values

	#Players	Avg Value	Avg Overall	Avg Potential
English	**2984**	**1.351.004 €**	**64**	**70**
Australia	224	743.817 €	63	68
Ghana	117	2.336.453 €	67	73
Nigeria	126	2.316.706 €	67	73
United Kingdom	2137	1.345.510 €	63	70
United States	380	1.116.197 €	65	70
Spanish	**3438**	**3.063.871 €**	**68**	**73**
Argentina	961	2.875.021 €	68	72
Chile	367	2.831.839 €	71	71
Colombia	591	1.711.751 €	66	71
Mexico	360	1.704.569 €	66	71
Spain	1009	4.445.268 €	70	75
Uruguay	150	4.138.967 €	69	73

Resultados y conclusiones

Data Visualization – Club Profiling

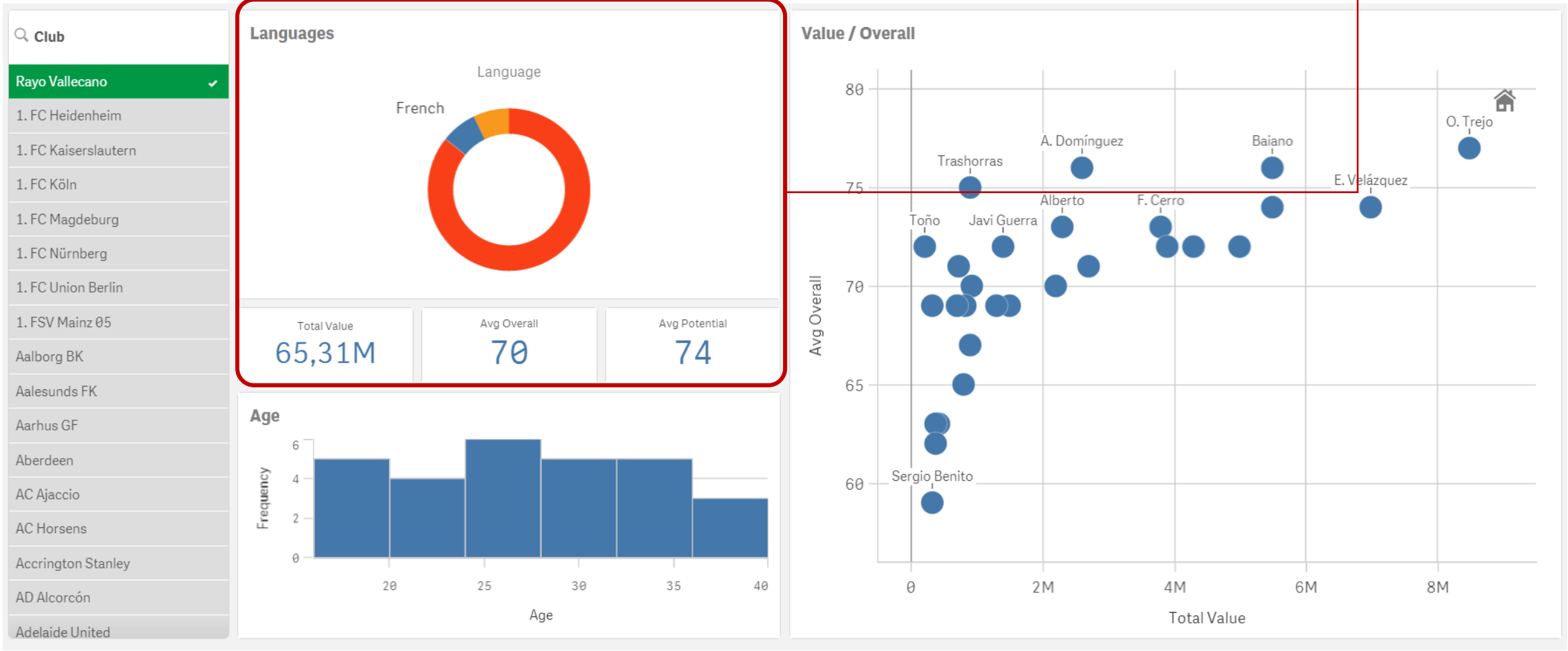
Los Clubes cuyos jugadores tienen un mayor valor, en general, tienen mayor variedad lingüística



Resultados y conclusiones

Data Visualization – Club Profiling

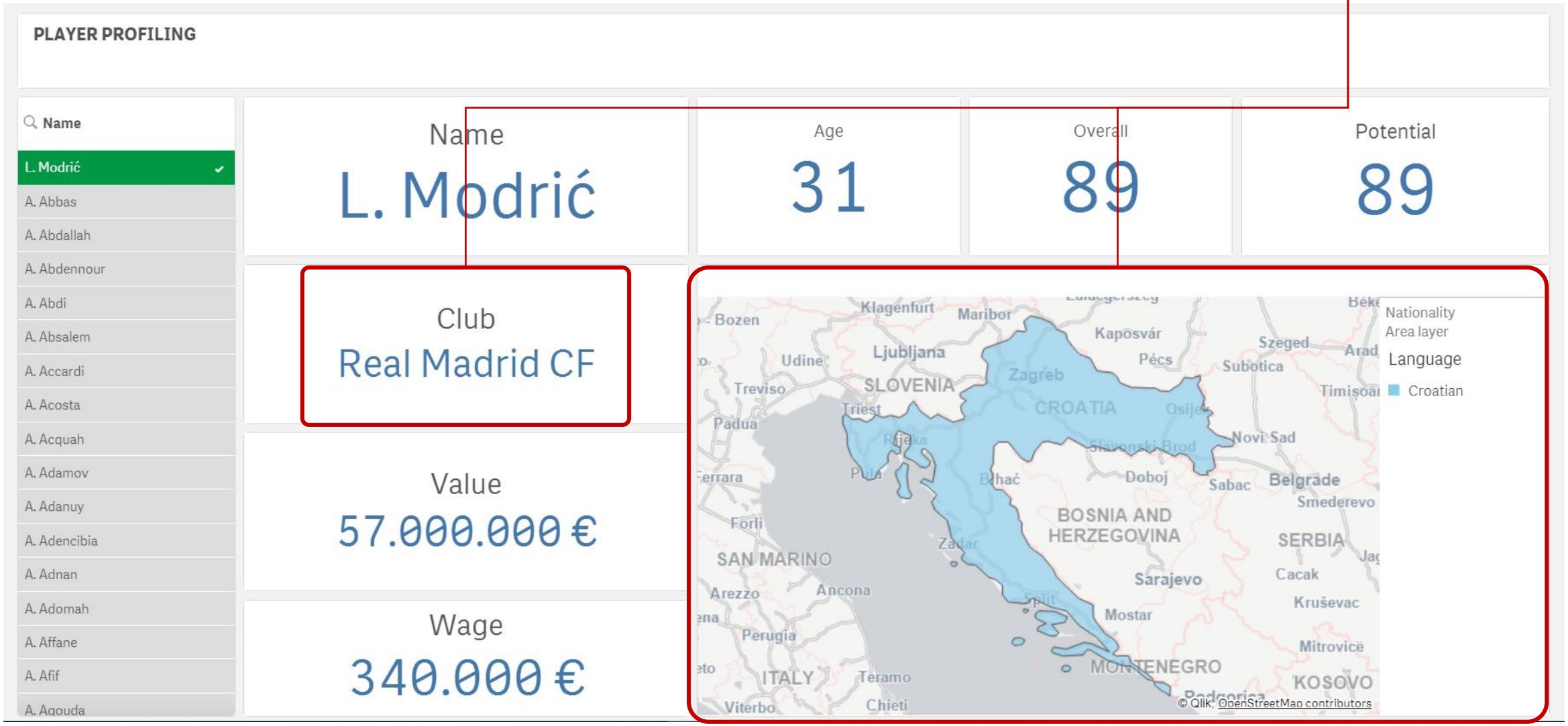
Sin embargo, en aquellos con un valor menor, la variedad de idiomas se reduce significativamente



Resultados y conclusiones

Data Visualization – Player Profiling

Esto se debe a que los Clubes con mayor valor cuentan con numerosos internacionales



Resultados y conclusiones

Data Visualization – Player Profiling

Mientras que los Clubes con menor valor, cuentan mayoritariamente con jugadores locales

Resultados y conclusiones

Sentiment Analysis

Se han recolectado más de 40K de tweets de los **últimos 5 días** utilizando la **Search API** de Twitter, mediante la librería **tweepy**

El **99% de los tweets** recolectados **no tienen informada su geolocalización** por parte del usuario

Los tweets serán analizados mediante el método *analyzeSentiment* la **API de Natural Language** de Google Cloud

Se han considerado únicamente los tweets escritos en alguno de los siguientes **idiomas**:

Chino, coreano, español, francés, inglés, italiano, japonés, portugués y ruso

Métodos

Método	Solicitudes	Errores	Latencia media	Latencia del percentil 99
google.cloud.language.v1.LanguageService.AnalyzeSentiment	45.713	0,01 %	0,048 segundos	0,123 segundos

Resultados y conclusiones

Sentiment Analysis

Los tweets analizados obtienen una puntuación (**score**) entre -1 y 1.

Siendo **(1)** las opiniones más **positivas** y **(-1)** las más **negativas**

Los tweet analizados cuentan con un valor de magnitud (**magnitude**).

Cuanto mayor es la magnitud, mas **emocionalmente significativa** es una opinión.

Los tweets se han clasificado en **5 niveles** en función de su score:

1. Very Negative
2. Negative
3. Neutral
4. Positive
5. Very Positive

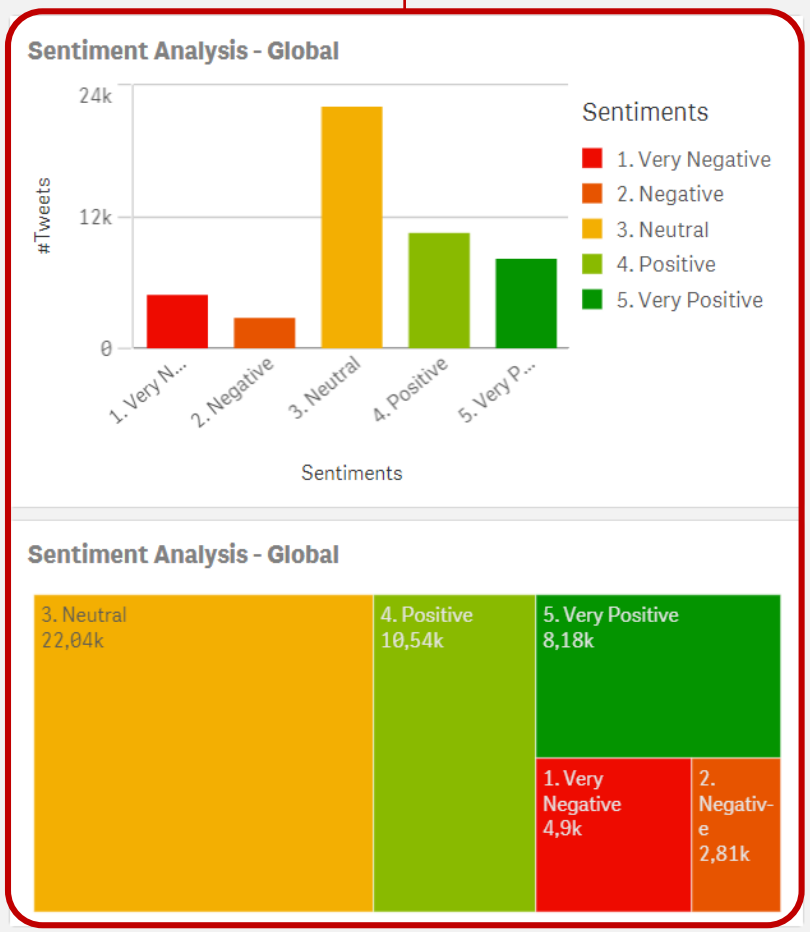
Country	Language	Sentiments	Tweet
Spain	es	5. Very Positive	Buen mensaje de #FIFA20 @EASPORTSEsp https://t.co/XznyjOYmbw
Spain	es	4. Positive	#deathstanding #CallofDutyModernWarfare #fifa20 #TheLastOfUsPartII y creo que ya me paso de los 250 porque cogería... https://t.co/MLbrLhhihO
Spain	es	4. Positive	4-2 en el global, 3-1 ayer Viernes y 1-1 hoy. A la tarde mas... #FutChampions #Fifa20 #GuezarjeCF https://t.co/VjVCqlElAj
Spain	es	4. Positive	Primer partido en #FutChampions de #FIFA20 https://t.co/Ch3rtwLvRd
Spain	es	1. Very Negative	@F_deFIFA No entiendo una cosa. Subiste un video hace tiempo diciendo que dejabas fifa por la conexión, servidores,... https://t.co/giwU0EguxD
Spain	en	1. Very Negative	Is your FUTCHAMPS @EASPORTSFIFA @EASPORTSEsp @EASPORTS not my. YOU decide Who wins with this fake goals. Bastard TH... https://t.co/d312Mj9tT7

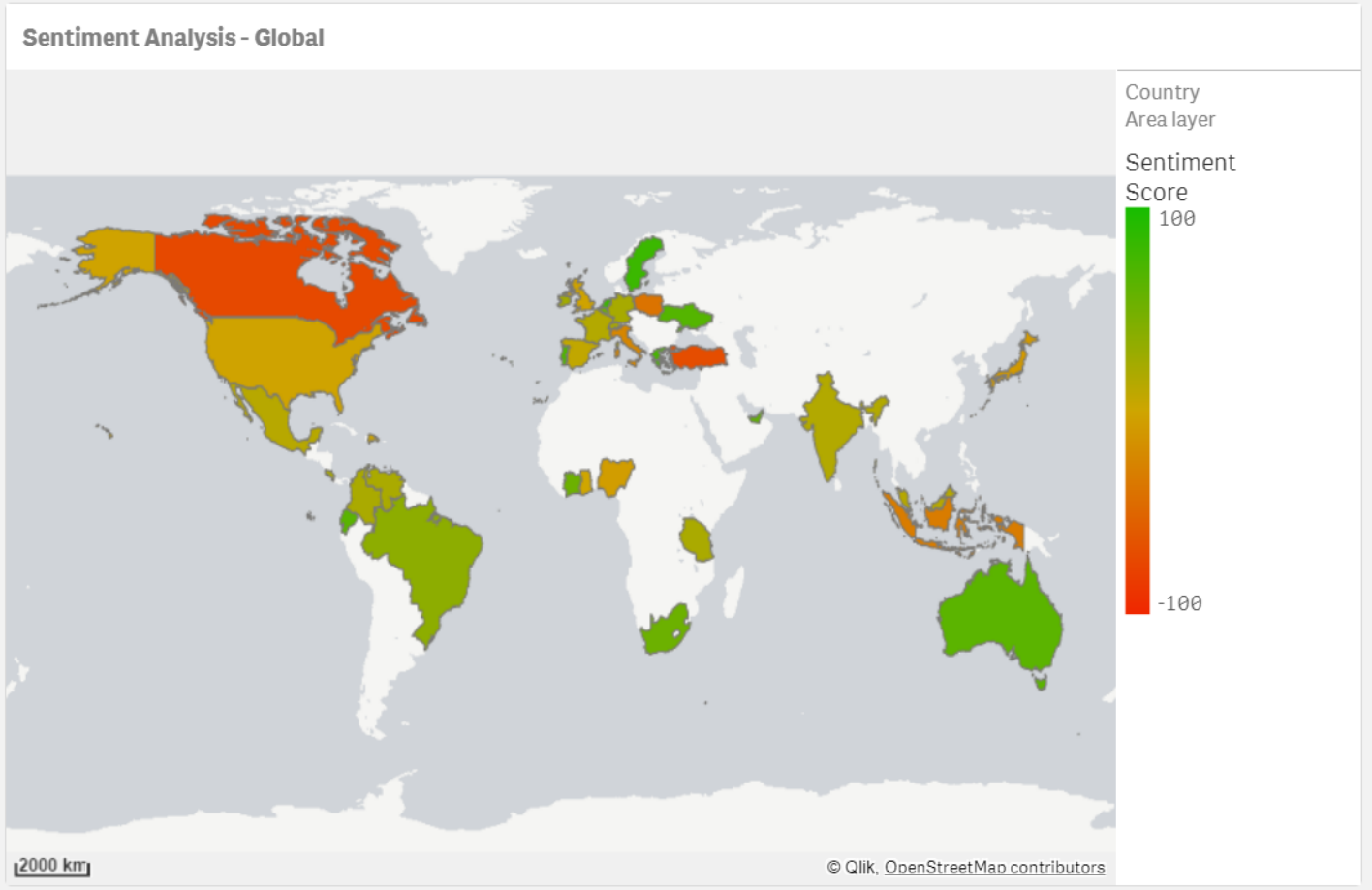
Resultados y conclusiones

Sentiment Analysis

En los datos analizados sobre el FIFA 20, los comentarios positivos o muy positivos, superan ampliamente a los comentarios negativos o muy negativos

FIFA 20 Sentiment Analysis





Resultados y conclusiones

Sentiment Analysis

En general, los países de habla no inglesa tienen mejores opiniones del FIFA 20

FIFA 20 Sentiment Analysis