

FIFA WORLD CUP 2010

ANÁLISIS DE CLUSTERING

¿Los que jugaron bien fueron los mejores?

Lic. Alejandro Saavedra

29 de Septiembre 2010

Comunidad analyticsconosur.com

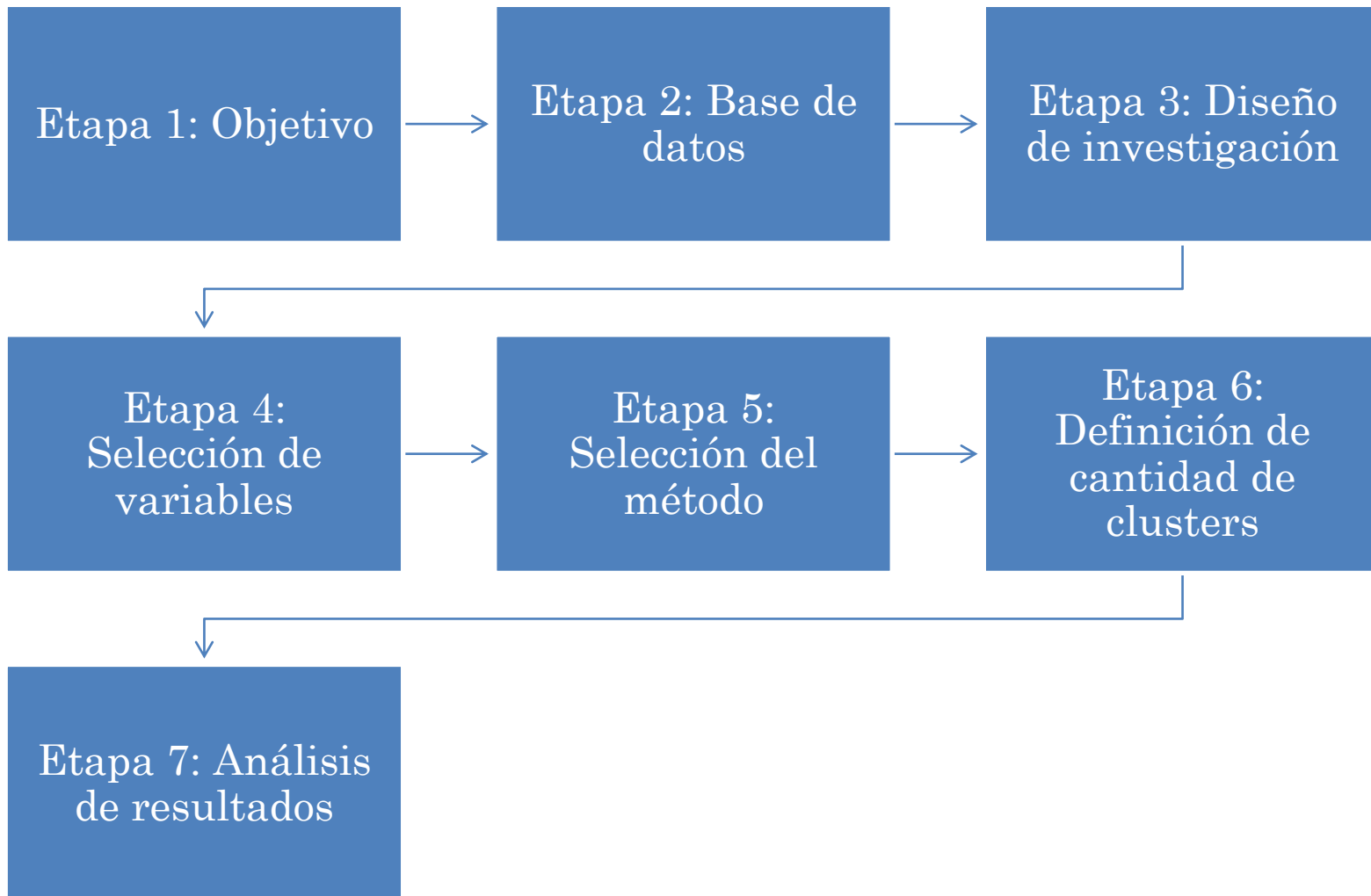


INTRODUCCIÓN

- Para **crear una segmentación o clustering** que genere diferentes grupos para los **32 equipos clasificados** al Mundial de Sudáfrica 2010, se utilizarán diferentes variables relacionadas con su juego.
- Es importante aclarar que se prescindieron de determinadas técnicas necesarias para obtener resultados “estadísticamente correctos”, porque se priorizó la **comprensión de las relaciones causa y efecto del proceso de clustering de manera directa.**
- Esta simplificación se realiza para salir un poco de los análisis de tipo “caja negra”, en los que:
 1. Se ingresan una serie de variables.
 2. Se utiliza el software de análisis estadístico para obtener los resultados esperados.
 3. Se analizan los resultados sin tener una comprensión completa del proceso que los genera.



DIAGRAMA DE TRABAJO



The slide features a dark blue background with a vertical decorative element on the left consisting of several thin, light blue stripes of varying widths. To the right of these stripes are several blue circles of different sizes, arranged in a vertical line that tapers towards the bottom.

ETAPA 1: OBJETIVO

FIFA World cup 2010
Análisis de Clustering

ETAPA 1: OBJETIVO

Segmentar los países clasificados al Mundial de Sudáfrica 2010 en grupos con características de juego similares entre sí y diferentes entre los grupos.





ETAPA 2: BASE DE DATOS

FIFA World cup 2010
Análisis de Clustering

ETAPA 2: BASE DE DATOS

- Para poder realizar la valoración de cada país, se **importaron los datos publicados en el sitio de FIFA** con la función “Importar datos desde web” de Excel.
 1. Se importaron las **62 variables** publicadas en fifa.com.
 2. Los datos se refieren a los **599 jugadores** que jugaron al menos un minuto.
 3. Para generar la base de datos sobre la que se realizará el análisis, **se calcularon los totales por país de todas las variables**.
 4. Para eliminar el efecto de la cantidad de partidos jugados, que afecta a los totales de las variables, **se calcularon los valores “por partido jugado” de cada variable**.
 5. Teniendo las 62 variables calculadas “por partido jugado”, se comenzó con el **proceso de discriminación de las variables de importancia para realizar el análisis de clustering**.





ETAPA 3: DISEÑO DE INVESTIGACIÓN

FIFA World cup 2010
Análisis de Clustering

ETAPA 3: DISEÑO DE INVESTIGACIÓN

- **Criterio de selección de variables:**
 - Variables utilizadas en clustering: 7 de 62.
 - Análisis de frecuencias, correlaciones y pruebas.
- **Definición del método:**
 - Análisis de clusters de K medias
 - Software utilizado: IBM SPSS Statistics 18
- **Definición de cantidad de clusters: 4**





ETAPA 4: SELECCIÓN DE VARIABLES

FIFA World cup 2010
Análisis de Clustering

CRITERIO DE SELECCIÓN DE VARIABLES

- El primer paso fue **examinar que no haya outliers** antes de comenzar con la partición, y no se encontraron casos para eliminar.
- Para seleccionar las variables se utilizaron las siguientes técnicas:
 - **Análisis de frecuencias:**
 - Para conseguir una primera impresión de los datos y conocer la distribución de las variables
 - **Análisis de correlaciones de Pearson:**
 - Buscando correlaciones bajas y significantes (al nivel 0,05 bilateral).
 - En los casos en que se encontraron correlaciones altas, y como no se realiza análisis factorial, se seleccionaron las variables más representativas del juego de un equipo.
 - **Prueba y error:**
 - Aplicando el clustering en diversas oportunidades y analizando los resultados y el papel que tomaban las variables.



ETAPA 4: SELECCIÓN DE VARIABLES

- **No se utiliza ninguna estandarización** porque, si bien las variables estaban en diferentes escalas, la magnitud de la percepción directa es un elemento importante para los objetivos de segmentación.
- **Proceso de transformación de variables:**
 1. Las variables originales están expresadas en **valores por jugador**.
 2. Se calcularon los **totales por país** de todas las variables.
 3. Se realizaron las transformaciones necesarias para que las mismas estén expresadas **por partido jugado**.
 4. **Ejemplo:** Para la variable “Distancia cubierta en posesión“, en el caso de España se sumaron las distancias recorridas de sus 21 jugadores (359,29 km.) y se dividió este valor por los 7 partidos jugados: 51,33 km.
- **Proceso de selección:**
 - Para discriminar cuáles variables incluir en el análisis, se **priorizaron las consideraciones prácticas sobre las teóricas**.
 - Se incluyeron **solo las variables que exhibieron diferencia** a través de los casos a ser clusterizados.



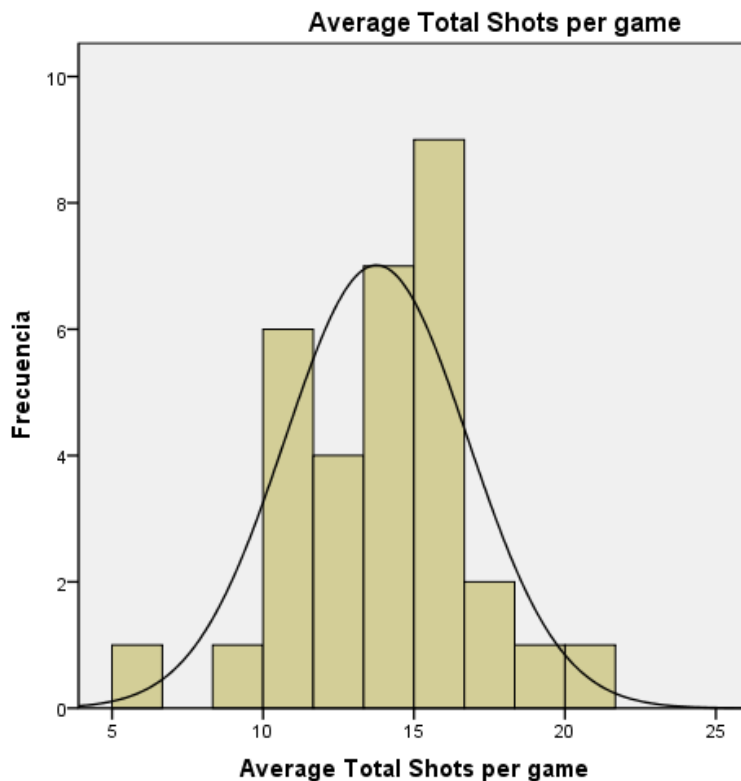


LAS VARIABLES SELECCIONADAS

FIFA World cup 2010
Análisis de Clustering

1. AVERAGE TOTAL SHOTS PER GAME

Media	Desviación típica
13,76	3,033



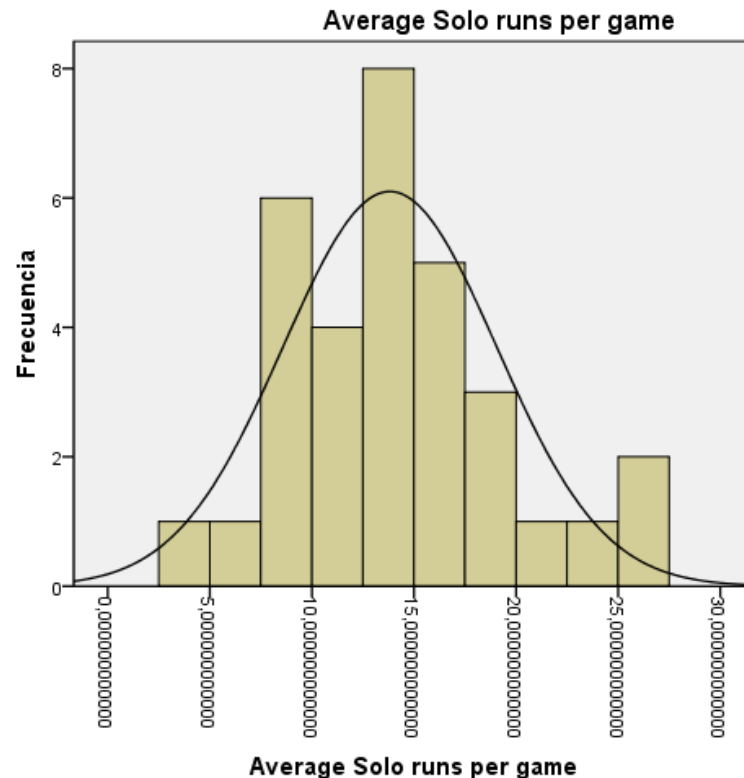
- Es el promedio de tiros totales por partido jugado.
- Ghana y Argentina son los que más tiros realizaron, con cerca de 20 por partido.



2. AVERAGE SOLO RUNS PER GAME

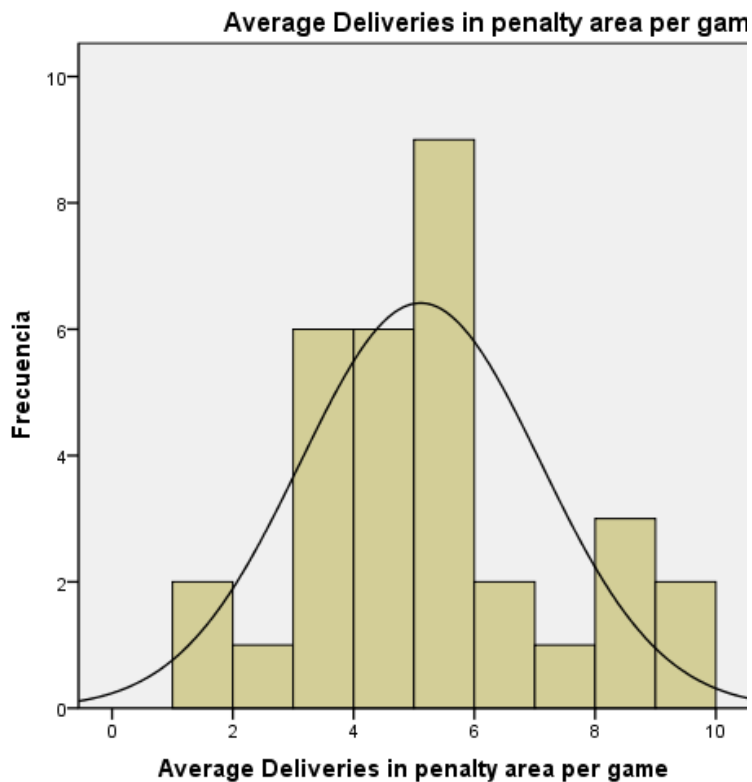
Media	Desviación típica
13,81577380952381	5,231646419568238

- Esta variable es el promedio de corridas solas realizadas por cada jugador, por partido jugado.
- Costa de Marfil y España son los que más corridas de este tipo tienen, con 27 y 25 respectivamente.



3. AVERAGE DELIVERIES IN PENALTY AREA PER GAME

Media	Desviación típica
5,11	1,990



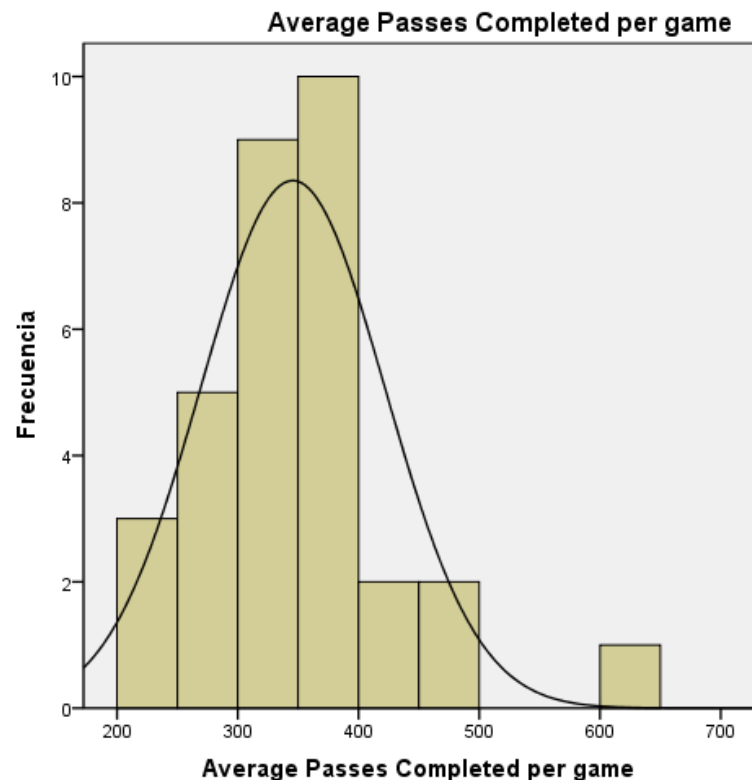
- Esta variable refleja los pases realizados en el área chica.
- Los países con mayores valores son Chile y España con cerca de 9 pases en el área chica por partido.



4. AVERAGE PASSES COMPLETED PER GAME

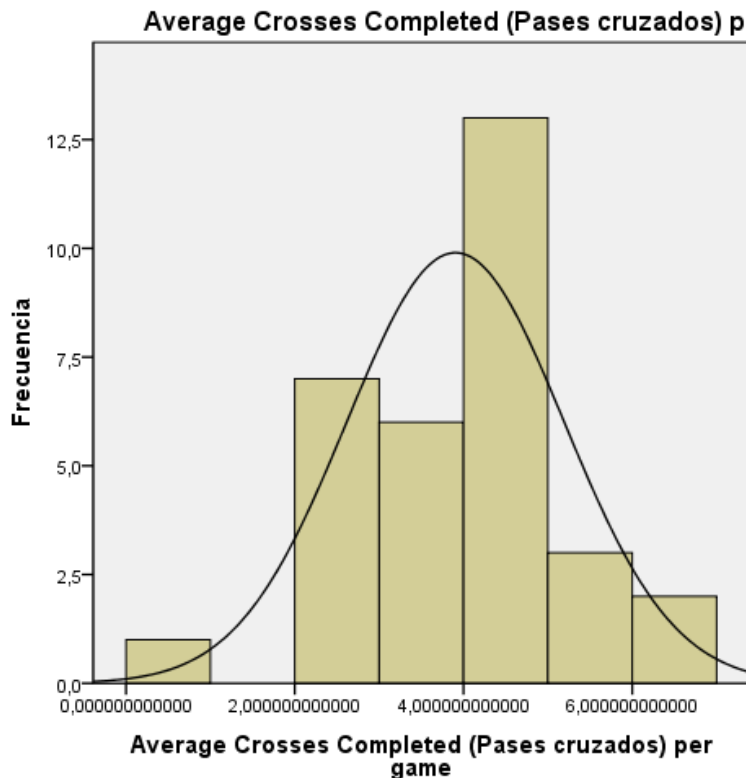
Media	Desviación típica
345,52	76,409

- El total de pases realizados por partido es un indicador del dominio de la pelota en el juego.
- España y Costa de Marfil tienen los mayores valores, con 606 y 466 promedio de pases por partido.



5. AVERAGE CROSSES COMPLETED PER GAME

Media	Desviación típica
3,90424107142857	1,289477018708406



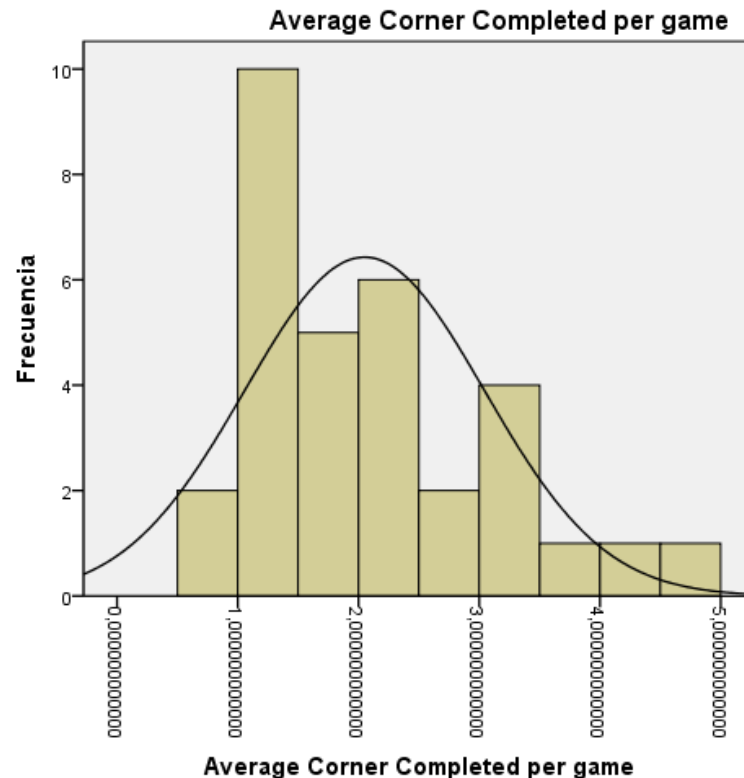
- Los pases cruzados realizados por partido es una variable que refleja un estilo de juego indirecto.
- España y Costa de Marfil son los equipos que tuvieron el mayor valor de esta variable, con un promedio de casi 7 pases cruzados por partido.



6. AVERAGE CORNER COMPLETED PER GAME

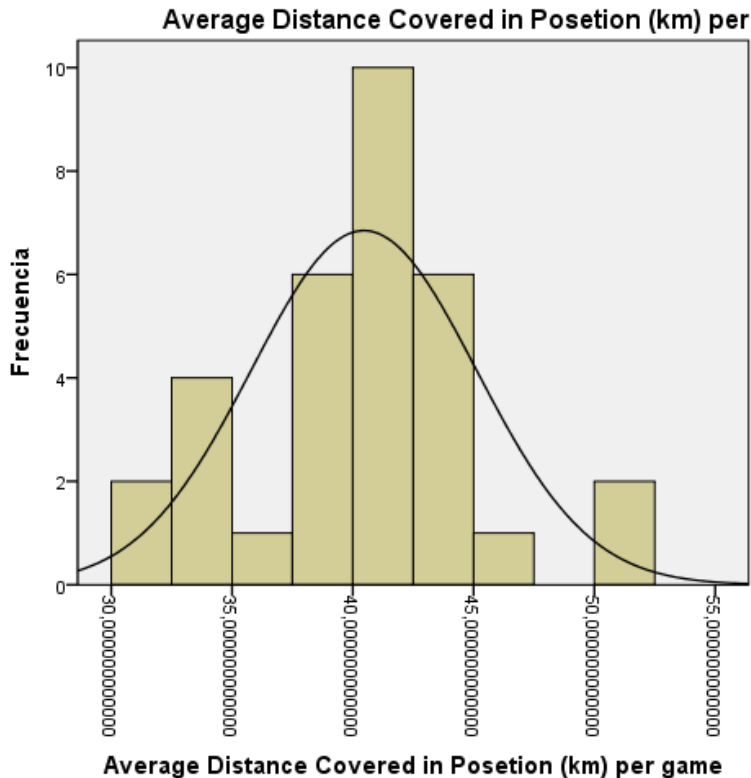
Media	Desviación típica
2,04985119047619	,992715759713989

- El promedio de corners por partido se relaciona con las llegadas al arco, que generalmente son desviadas.
- España y Costa de Marfil son los que más corners tuvieron, con un promedio de 4,5 por partido.



7. AVERAGE DISTANCE COVERED IN POSETION (KM) PER GAME

Media	Desviación típica
40,45700892857143	4,659910355187906



- El promedio de distancia cubierta en posesión de la pelota es una variable que refleja la dominación de un equipo.
- España y Costa de Marfil son los que más kilómetros hicieron en posesión de la pelota, con un promedio de cercano a los 50 km por partido.



CORRELACIONES DE PEARSON

Variables	Average Total Shots per game	Average Solo runs per game	Average Deliveries in penalty area per game	Average Passes Completed per game	Average Crosses Completed per game	Average Corner Completed per game	Average Distance Covered in Posetion (km) per game
Average Total Shots per game	1	,580**	,599**	,602**	,663**	,433*	,629**
Average Solo runs per game	,580**	1	,577**	,818**	,676**	,773**	,619**
Average Deliveries in penalty area per game	,599**	,577**	1	,642**	,658**	,561**	,574**
Average Passes Completed per game	,602**	,818**	,642**	1	,737**	,730**	,833**
Average Crosses Completed per game	,663**	,676**	,658**	,737**	1	,579**	,735**
Average Corner Completed per game	,433*	,773**	,561**	,730**	,579**	1	,515**
Average Distance Covered in Posetion (km) per game	,629**	,619**	,574**	,833**	,735**	,515**	1



ETAPA 5: DEFINICIÓN DEL MÉTODO

FIFA World cup 2010
Análisis de Clustering

ETAPA 5: DEFINICIÓN DEL MÉTODO

- **Análisis de clusters de K medias:**
 - Utiliza la **distancia euclidiana** entre casos, para asignarlos a los grupos.
 - Se seleccionó este método porque es **conceptualmente fácil de entender**, y sirve para variables cuantitativas continuas y ordinales.
 - Si bien este método es **sensible a la métrica de las variables, el rango de variabilidad de cada variable no es muy grande**. Para reducir este efecto se calcularon los valores de cada variable por partido.
- **Método solo iterar y clasificar:**
 - Se clasifican los casos según sus centros iniciales, actualizando los valores iterativamente.
- **Nº máximo de iteraciones máximas: 10.**
- **Criterio de convergencia:** El proceso de iteración se detendrá cuando entre una iteración y la siguiente no se consiga desplazar ninguno de los centros una distancia superior al 0% de la menor de las distancias existentes entre cualquiera de los centros iniciales.
- **Uso de medias actualizadas:** Esto se utiliza para solicitar la actualización de los centros de clusters, y que el orden de los casos de la base de datos no afecte la solución obtenida.



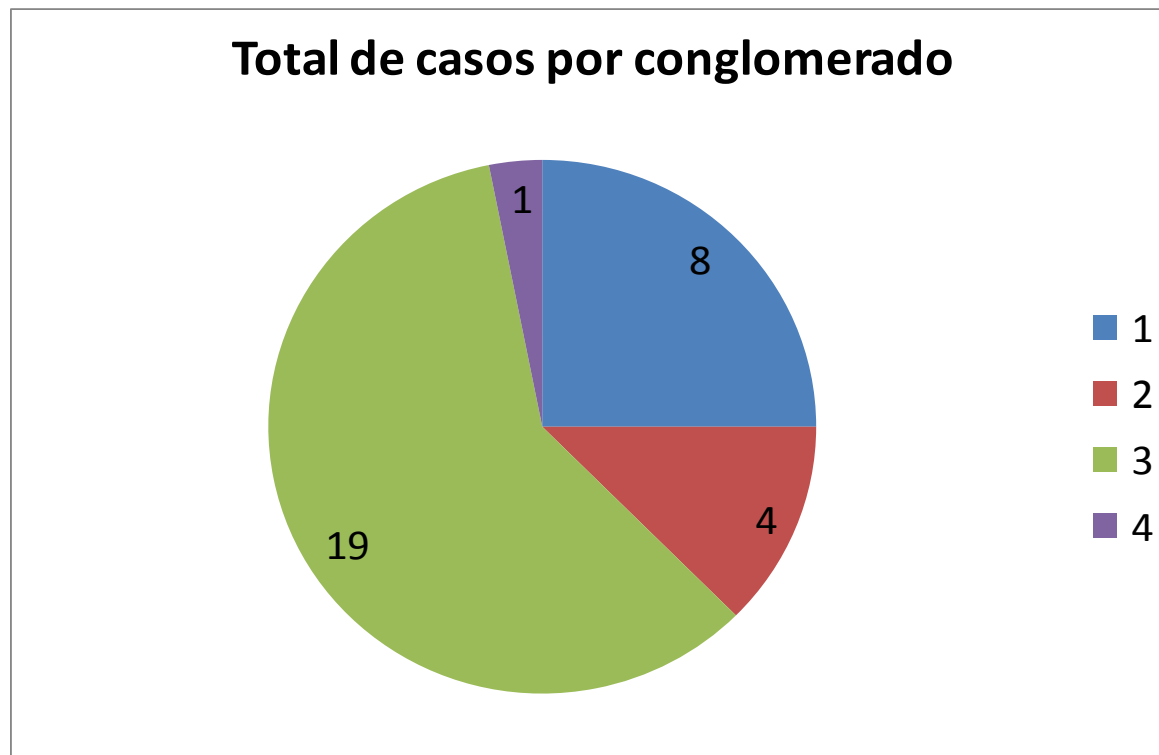


ETAPA 6: DEFINICIÓN DE CANTIDAD DE CLUSTERS

FIFA World cup 2010
Análisis de Clustering

ETAPA 6: DEFINICIÓN DE CANTIDAD DE CLUSTERS

- El número de clusters fue de 4, esta cantidad se definió después de ver los resultados con diferentes cantidades de clusters.
- A continuación se pueden observar los resultados después de aplicar el análisis de clusters de K medias:



CASOS UTILIZADOS COMO CENTROS INICIALES DE LOS CLUSTERS

Variables	New Zealand	Ivory Coast	USA	Spain
Average Total Shots per game	5	15	18	16
Average Solo runs per game	4	27	10	25
Average Deliveries in penalty area per game	3	8	5	9
Average Passes Completed per game	221	466	310	606
Average Crosses Completed per game	1	7	5	7
Average Corner Completed per game	1	4	2	5
Average Distance Covered in Posetion (km) per game	35	51	45	51

- Si bien a continuación se pueden observar los centros finales de los clusters después de la iteración, es importante observar cuáles **equipos fueron detectados inicialmente como centros**.



CENTROS FINALES DE LOS CLUSTERS

Variables	Greece Honduras Japan New Zealand Nigeria North Korea Switzerland Uruguay	Argentina Brazil Germany Ivory Coast	Algeria – Australia - Cameroon Chile – Denmark - England France – Ghana - Italy Mexico – Netherlands - Paraguay Portugal – Serbia - Slovakia Slovenia - South Africa South Korea - USA	Spain
Average Total Shots per game	11	16	14	16
Average Solo runs per game	10	22	13	25
Average Deliveries in penalty area per game	3	6	5	9
Average Passes Completed per game	260	442	348	606
Average Crosses Completed per game	3	5	4	7
Average Corner Completed per game	1	3	2	5
Average Distance Covered in Posetion (km) per game	35	45	41	51

Comparando los centros iniciales, con los finales después de la iteración, se pueden apreciar el **desplazamiento de los centros de los clusters**. Esto no ocurre en Spain porque hay un único caso.



VARIABLES ADICIONALES

Variables	Spain	Argentina Brazil Germany Ivory Coast	Algeria – Australia - Cameroon Chile – Denmark - England France – Ghana - Italy Mexico – Netherlands - Paraguay Portugal – Serbia - Slovakia Slovenia - South Africa South Korea - USA	Greece Honduras Japan New Zealand Nigeria North Korea Switzerland Uruguay
Total games played	7	5	3,79	3,63
Promedio de goal difference per game	0,86	0,93	-0,29	-0,68

- Para enriquecer los resultados obtenidos, se incluyen las siguientes variables para utilizar en el análisis posterior al proceso de clustering:
 - **Total games played:** Es el total de partidos jugados por el equipo.
 - **Average goal difference per game:** Este valor de diferencia de gol se obtiene restando los goles recibidos a los realizados (Diferencia de gol = Goles realizados – Goles recibidos). Para el análisis se lo divide por los partidos jugados.
- **Estas variables no se incluyeron en el proceso de clustering** porque se refieren a los resultados y no a la forma de juego.



CLUSTERS FINALES

El Campeón

- Spain

Los mejores

- Argentina
- Brazil
- Germany
- Ivory Coast

Ni mejores ni peores

- Algeria
- Australia
- Cameroon
- Chile
- Denmark
- England
- France
- Ghana
- Italy
- Mexico
- **Netherlands**
- Paraguay
- Portugal
- Serbia
- Slovakia
- Slovenia
- South Africa
- South Korea
- USA

Los peores

- Greece
- Honduras
- Japan
- New Zealand
- Nigeria
- North Korea
- Switzerland
- **Uruguay**



ETAPA 7: ANÁLISIS DE LOS RESULTADOS

FIFA World cup 2010
Análisis de Clustering

ETAPA 7: ANÁLISIS DE RESULTADOS

El Campeón (Cluster 4)

- Es de esperar que este grupo esté constituido por un solo caso: **España**.
- No solo fue el campeón, sino que fue **el mejor en las 7 variables** utilizadas para determinar los clusters.

Los mejores (Cluster 2)

- Los mejores tuvieron un **desempeño similar al Campeón**, es más, lo superó en la diferencia de gol promedio (0,93 contra 0,86 de España).

Ni mejores ni peores (Cluster 3)

- Este grupo tuvo una participación con valores que se ubican **entre los mejores y los peores**.

Los peores (Cluster 1)

- Como es de esperar, al cluster con el **rendimiento más bajo en las variables** utilizadas, que se relacionan con la calificación del juego, se lo denomina “Los peores”.

Analizando los países incluidos en cada cluster se puede observar que **el campeón fue definitivamente un caso atípico**, porque es el único integrante de su grupo.



ANÁLISIS DE LOS CLUSTERS

Variables	El Campeón (Cluster 4)	Los mejores (Cluster 2)	Ni mejores ni peores (Cluster 3)	Los peores (Cluster 1)
Total games played	7	5	3,79	3,63
Promedio de goal difference per game	0,86	0,93	-0,29	-0,68
Average Total Shots per game	16	16	14	11
Average Solo runs per game	25	22	13	10
Average Deliveries in penalty area per game	9	6	5	3
Average Passes Completed per game	606	442	348	260
Average Crosses Completed per game	7	5	4	3
Average Corner Completed per game	5	3	2	1
Average Distance Covered in Posetion	51	45	41	35

Las variables en rojo no se incluyen en el proceso de clustering, pero se agregan a posteriori para enriquecer los resultados obtenidos.



DISTANCIAS ENTRE LOS CENTROS DE LOS CLUSTERS FINALES

Cluster	El campeón	Los mejores	Ni mejores ni peores	Los peores
Los peores	347,196	183,439	88,333	
Los mejores	163,931		95,25	183,439
Ni mejores ni peores	258,952	95,25		88,333
El campeón		163,931	258,952	347,196

La mayor distancia entre los centros de los clusters se da entre el campeón y los peores, esta distancia se achica a medida que mejora la calificación de cada grupo.



DISTANCIA RESPECTO A LOS CENTROS DE LOS CLUSTERS

Cluster 1

- New Zealand 39,599
- Uruguay 38,362
- Honduras 27,006
- Japan 26,537
- Greece 22,461
- Switzerland 17,835
- North Korea 8,614
- Nigeria 7,752

Cluster 2

- Germany 34,932
- Ivory Coast 24,959
- Argentina 21,764
- Brasil 11,392

Cluster 3

- Netherlands 42,138
- Slovakia 39,167
- USA 38,248
- Paraguay 34,714
- Italy 30,278
- Ghana 24,256
- South Korea 22,735
- Slovenia 18,434
- South Africa 15,145
- Chile 14,662
- France 13,277
- England 12,707
- Serbia 11,812
- Cameroon 9,577
- Mexico 9,064
- Algeria 6,927
- Portugal 5,561
- Australia 5,468
- Denmark 3,173

Cluster 4

- Spain: 0

TABLA ANOVA

Variables	Cluster		F	Sig.
	Media cuadrática	gl		
Average Total Shots per game	34,164	3	5,235	0,005
Average Solo runs per game	184,247	3	17,444	0
Average Deliveries in penalty area per game	15,497	3	5,687	0,004
Average Passes Completed per game	54777,158	3	92,077	0
Average Crosses Completed per game	8,858	3	9,933	0
Average Corner Completed per game	5,621	3	11,5	0
Average Distance Covered in Posetion (km) per game	164,424	3	25,593	0

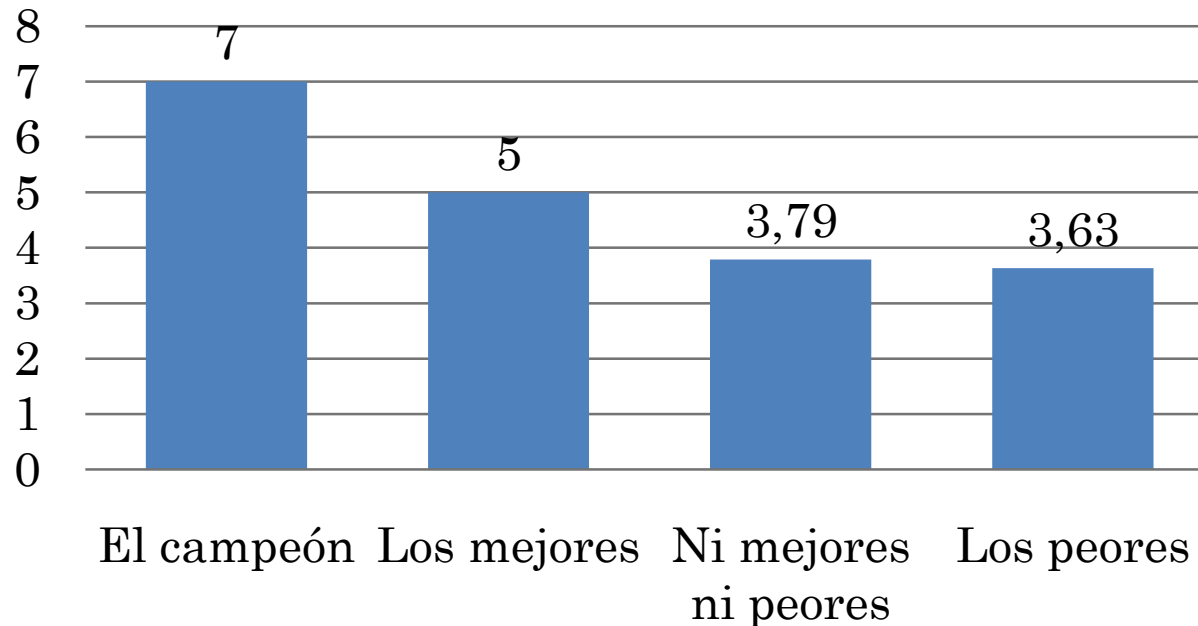
Este análisis muestra las **varianzas del estadístico F univariante**, y se obtiene tomando a los clusters como factor y a cada una de las variables como variable dependiente.

La tabla de ANOVA (análisis de varianza) demuestra que **las medias de las variables tienen diferencias entre los clusters determinados.**

Como en todos los casos el nivel de significancia de F es menor a 0.5, se **rechaza la hipótesis nula de que las medias de cada variable son iguales. Esto significa que todas las variables funcionan para discriminar los diferentes grupos.**



ANÁLISIS DE TOTAL GAMES PLAYED

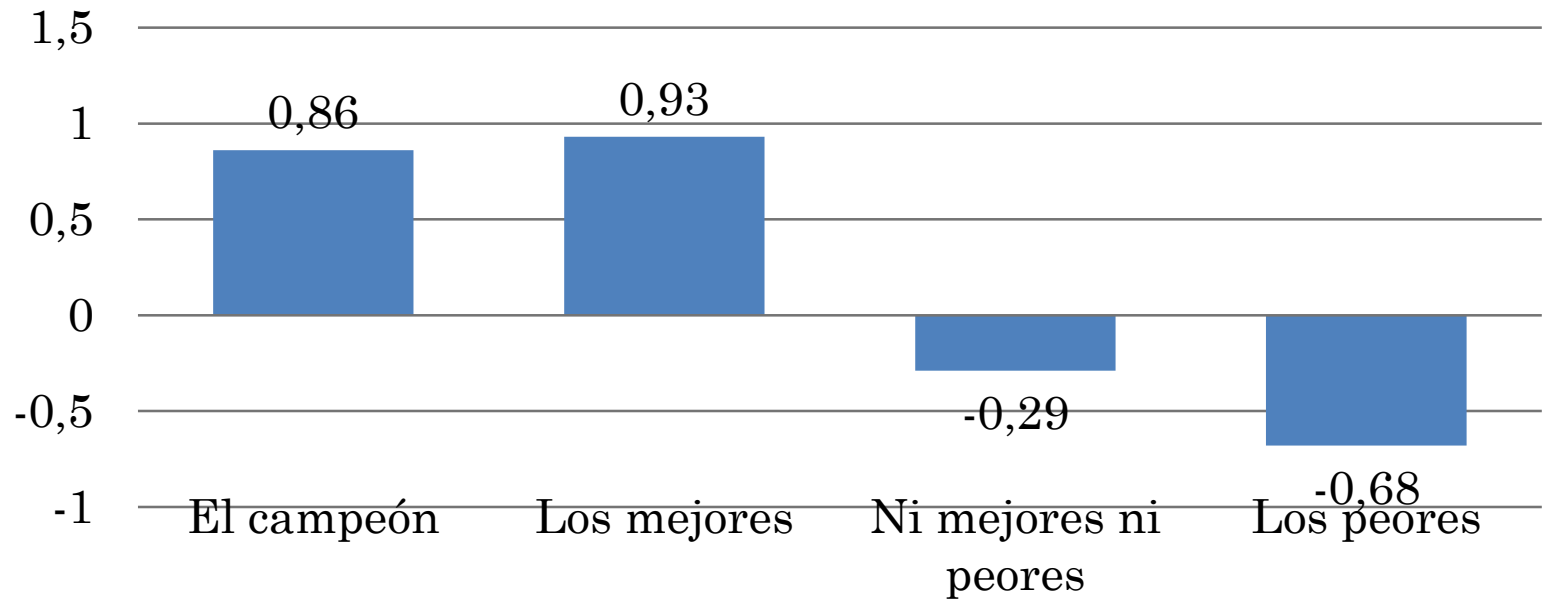


■ Promedio de Total games played (Pld)

- Si bien esta variable no se incluyó en el clustering, es buena para observar la relación entre la calificación determinada para el o los equipos integrantes de cada grupo, y los partidos que en promedio jugaron sus integrantes.



ANÁLISIS DE PROMEDIO DE GOAL DIFFERENCE PER GAME

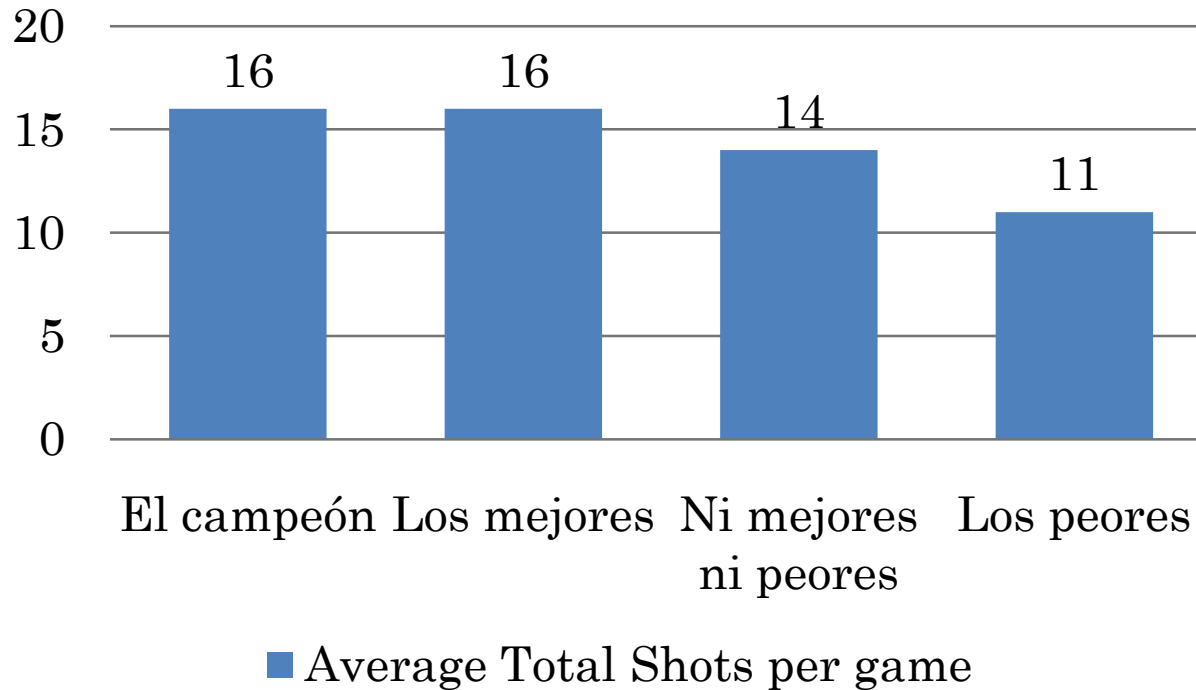


■ Promedio de Average goal difference ($GD = GF - GA$) per game

- El promedio de diferencia de gol es una variable en la que los equipos pertenecientes al grupo de los mejores obtuvieron un mejor resultado que el campeón.
- Se puede observar los últimos dos grupos tuvieron una diferencia de gol negativa.



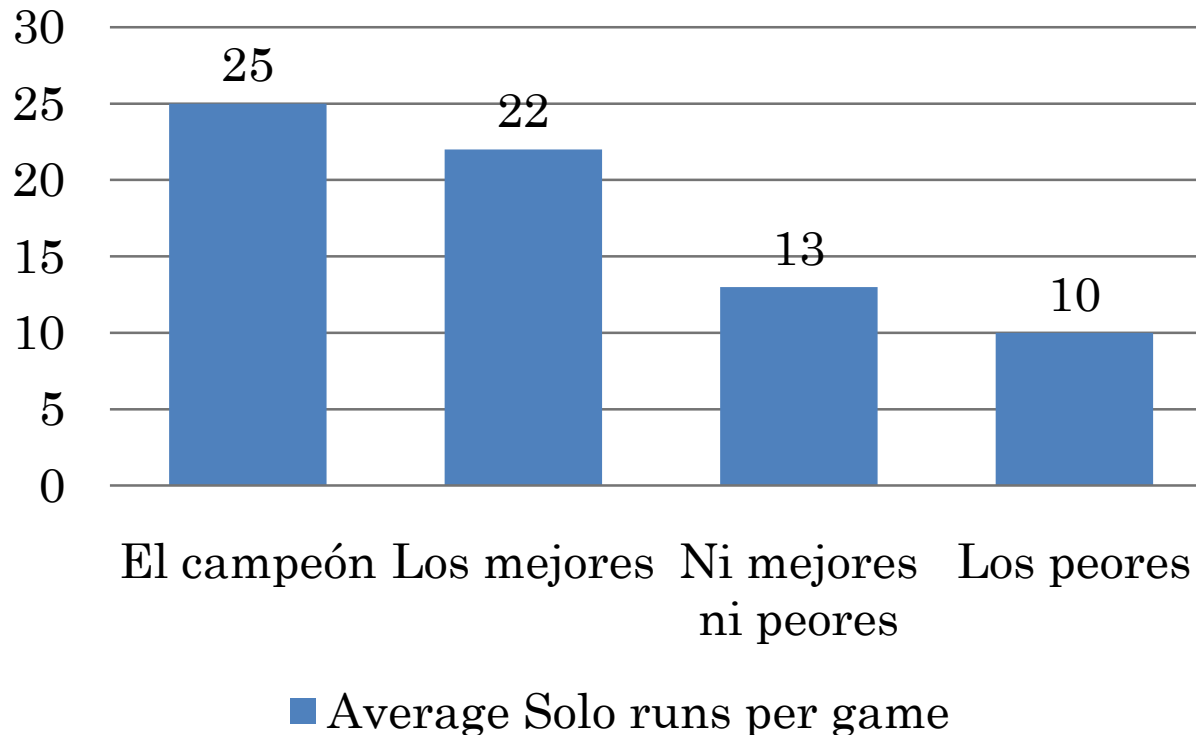
ANÁLISIS DE AVERAGE TOTAL SHOTS PER GAME



- El promedio de tiros totales es similar entre el campeón y los mejores, y baja para los otros dos grupos.



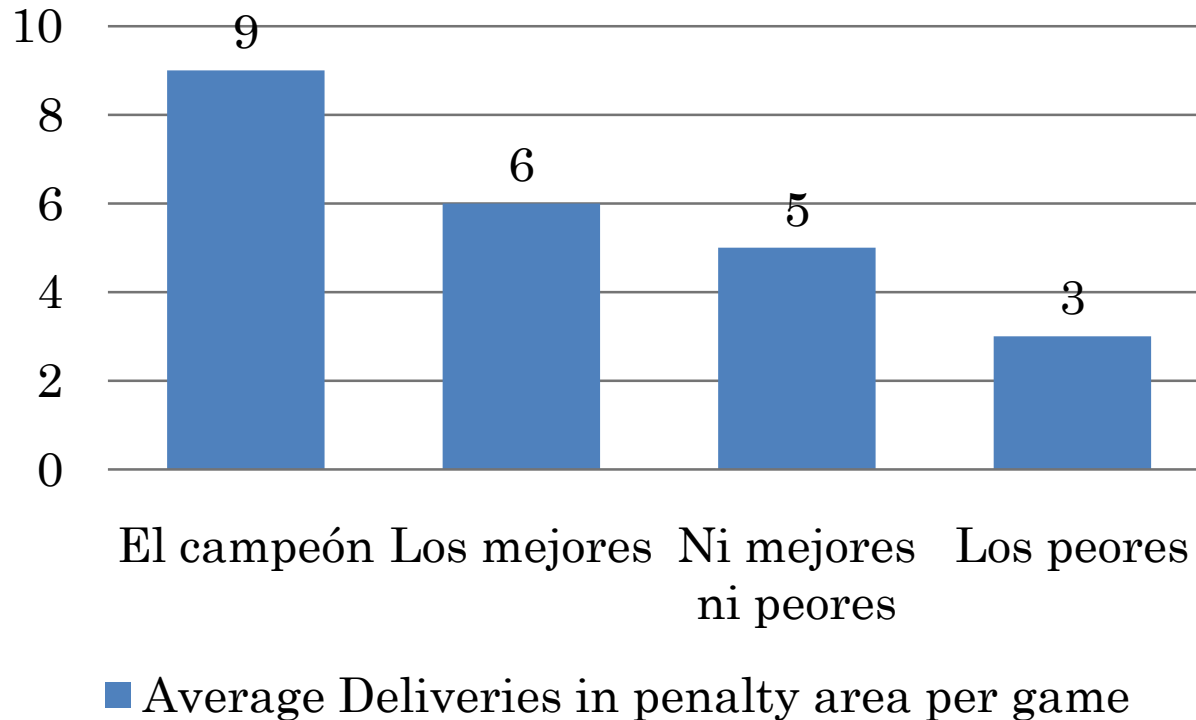
ANÁLISIS DE AVERAGE SOLO RUNS PER GAME



- El promedio de corridas solas realizadas por cada jugador, por partido jugado, es una variable que permite diferenciar a los 4 grupos.



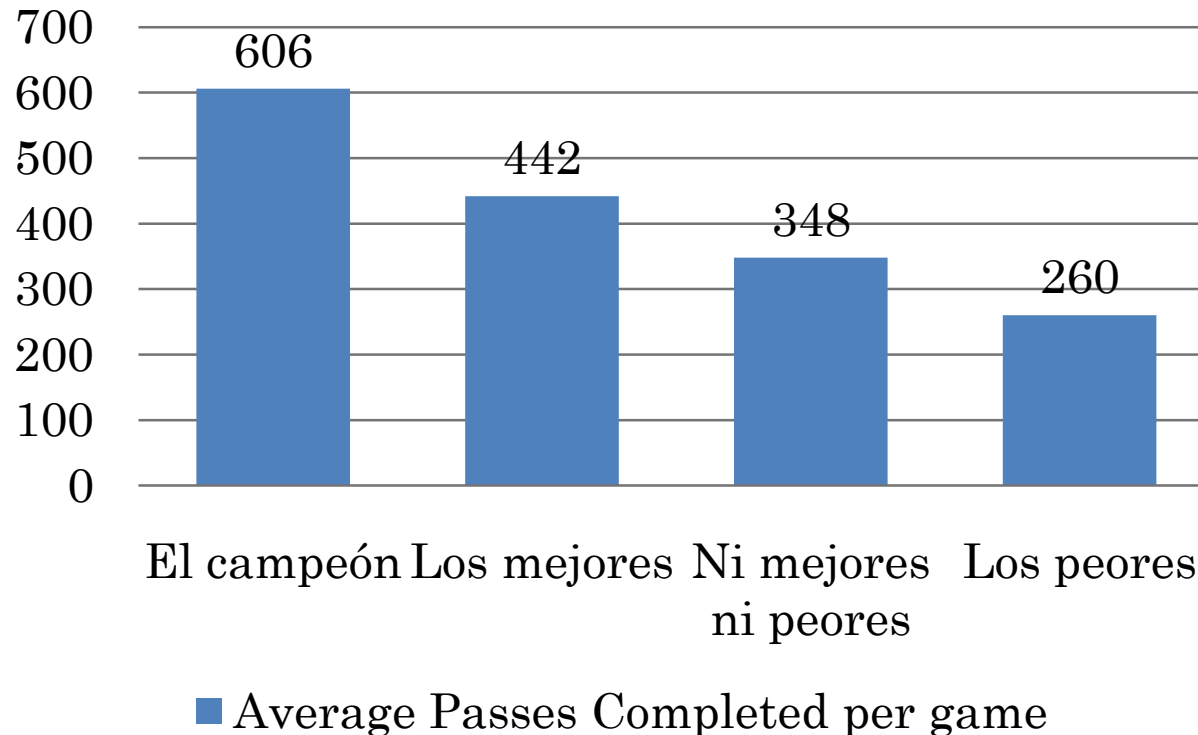
ANÁLISIS DE AVERAGE DELIVERIES IN PENALTY AREA PER GAME



- Esta variable refleja los pases realizados en el área chica, y permite observar que el campeón tuvo más llegadas que los equipos integrantes de los otros clusters.



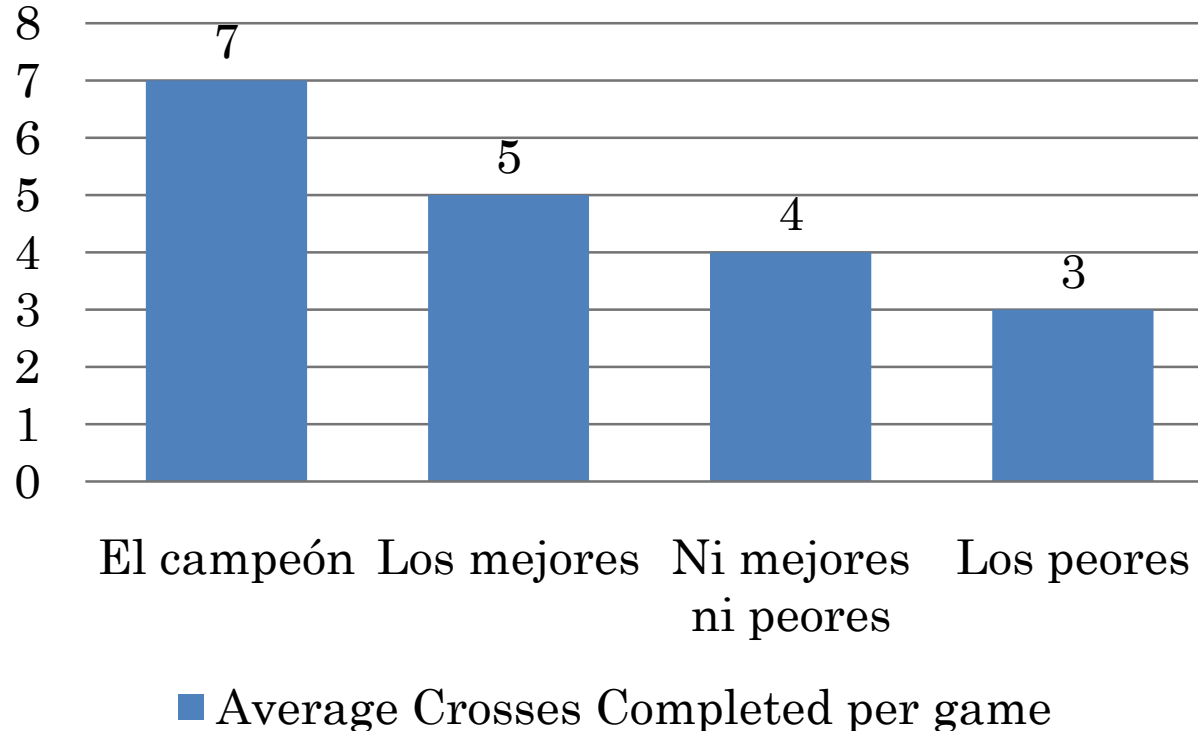
ANÁLISIS DE AVERAGE PASSES COMPLETED PER GAME



- En cuanto al promedio de pases realizados por partido jugado, El campeón vuelve a obtener una puntuación superior al resto.



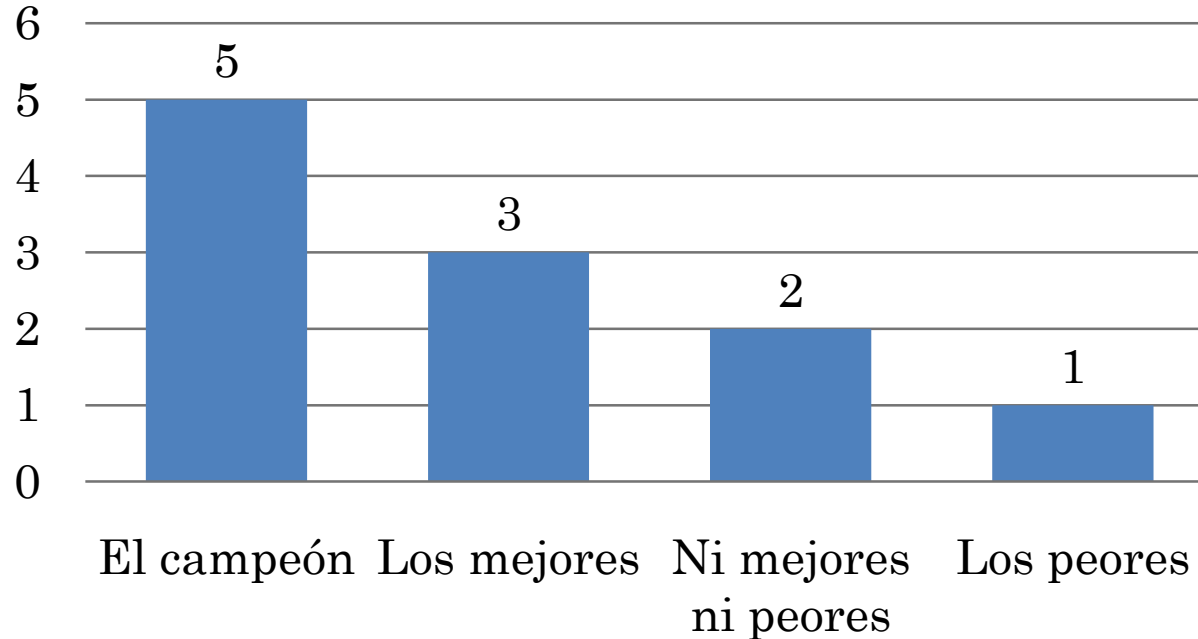
ANÁLISIS DE AVERAGE CROSSES COMPLETED PER GAME



- El promedio de pases cruzados realizados por partido es otra de las variables en la que se observan las diferencias entre grupos.



ANÁLISIS DE AVERAGE CORNER COMPLETED PER GAME

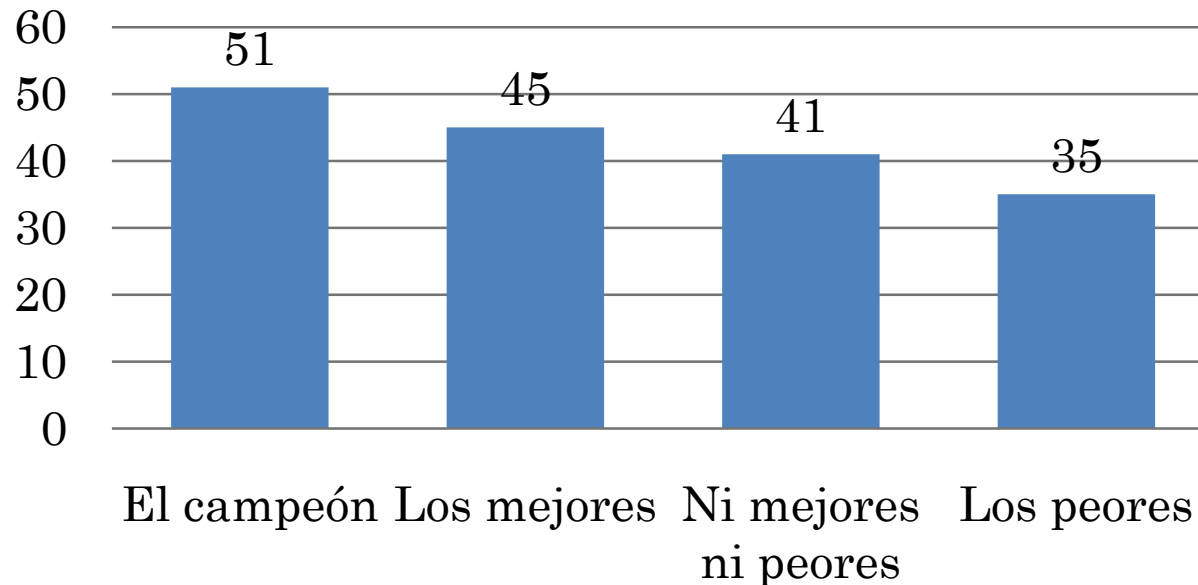


■ Average Corner Completed per game

- El promedio de córner realizados por partido vuelve a demostrar que España ha sido el campeón.



ANÁLISIS DE AVERAGE DISTANCE COVERED IN POSETION (KM) PER GAME



■ Average Distance Covered in Posetion (km) per game

- En este gráfico se puede observar la relación directa entre la posesión de la pelota y la calificación del equipo.



CONCLUSIÓN FINAL

- Los resultados de este análisis **no pueden ser tomados para caracterizar** a un buen equipo, porque hay muchos **factores externos que inciden en los resultados**: como la táctica del entrenador, las características de los jugadores y diferentes circunstancias que marcan un partido.
- A primera vista, la conclusión puede parecer trivial o naïve: **“Mientras más tiros, corridas, pases, llegadas, corners y tenencia de la pelota; mayor probabilidad de salir campeón.”** Pero analizando en profundidad los resultados, se puede observar:
 - A pesar de que **Costa de Marfil** (Ivory Coast) no pasó a octavos de final, se encuentra dentro del cluster de los mejores. **¿Qué hubiera ocurrido si no hubiera estado en el grupo de Brasil y Portugal?**
 - **Uruguay y Holanda** (Netherlands) están en los grupos de “Los peores”, y “Ni mejores ni peores”, pero llegaron a semifinales. **¿Se puede estar cerca de ser campeón jugando mal?**
- Para finalizar, es importante destacar que este mundial ocurrió algo que no es muy frecuente: **El que llegó como uno de los principales candidatos, por ser campeona de la Eurocopa 2008, fue el que mejor jugó y el que se llevó la copa de Sudáfrica 2010.**



BASES DE DATOS UTILIZADAS

- **Base de datos general por jugadores (inicial):**

- <https://docs.google.com/a/alesaavedra.com/leaf?id=0BzyrTciTQFfVNGU0ZmJhYTEtZGU3My00ODRhLTk0YjAtOGZiNzJkODMxZjdk&hl=en>

- **Base de datos por equipos (utilizada):**

- <https://docs.google.com/a/alesaavedra.com/leaf?id=0BzyrTciTQFfVZGMxOWM5YzctMGVjNi00M2M1LTg4ZTEtYTRhZWYyNTI0ZTM3&hl=en>

- **Resultados IBM SPSS Statistics 18:**

- <https://docs.google.com/a/alesaavedra.com/fileview?id=0BzyrTciTQFfVMGFjMThjZjctOWZjYy00ZWYwLTg2N2MtM2ZlMTE1OTgwN2Ux&hl=en>



FUENTES DE DATOS

- <http://www.fifa.com/worldcup/statistics/players/goals.html>
- <http://www.fifa.com/worldcup/statistics/players/shots.html>
- <http://www.fifa.com/worldcup/statistics/players/shotsposition.html>
- <http://www.fifa.com/worldcup/statistics/players/defending.html>
- <http://www.fifa.com/worldcup/statistics/players/attacking.html>
- <http://www.fifa.com/worldcup/statistics/players/passes.html>
- <http://www.fifa.com/worldcup/statistics/players/distanceandspeed.html>



COMENTARIOS

- Alejandro Saavedra
- mail@alesaavedra.com
- [@alesaavedra](#)

