

**Trabajo Teórico Estadística**

**Curso 2022/2023**

**Estudio sobre NFL**

**Grupo**

G11 - 2ºB

**Participantes**

Georgi Angelov Chervenyashki





Índice

Introducción 3

Presentación datos de entrada y sus tipos 3

Variables Cualitativas 4

Variables Cuantitativas 5

Análisis de relaciones entre variables 8

Modelo de regresión lineal 10

Contrastes de hipótesis 11



1. **INTRODUCCION**

Hemos elegido una base de datos sobre la **NFL** (National Football League) o Liga Nacional de Fútbol Americano. Más concretamente datos registrados durante la **temporada 2016** o la 97.ª edición. Estos datos están limitados a los **32 equipos** que participan en la competición registrando así los datos más importantes de cada partido disputado durante las 17 semanas de duración de la competición.

1. **Presentación datos de entrada y sus tipos**

Hemos recogido estos datos en un **dataframe** con las siguientes variables:

**- Team**: Nombre del equipo

**- Wins**: Número de victorias durante la temporada

**- Losses**: Número de derrotas durante la temporada

**- Ties**: Número de empates durante la temporada

**- WinPct**: Porciento de victorias

**- PointsFor**: Puntos a favor

- **PointsAgainst**: Puntos en contra

**- NetPts**: Diferencia de puntos

**- YardsFor**: Yardas a favor

**- YardsAgaint**: Yardas en contra

**- TDs**: Número de Touchdowns

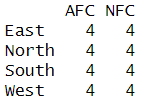
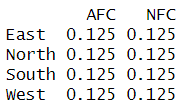
**- Division**: División del equipo (Este, Oeste, Norte, Sur)

**- Conference**: Tipo de federación (AFC o NFC)

* 1. Variables Cualitativas

Uno de los aspectos que podemos analizar con las variables cualitativas de nuestra base de datos es la organización de los 32 equipos y la relación que se da entre el número de equipos de distintas división o federación.

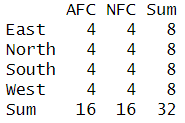


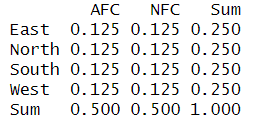


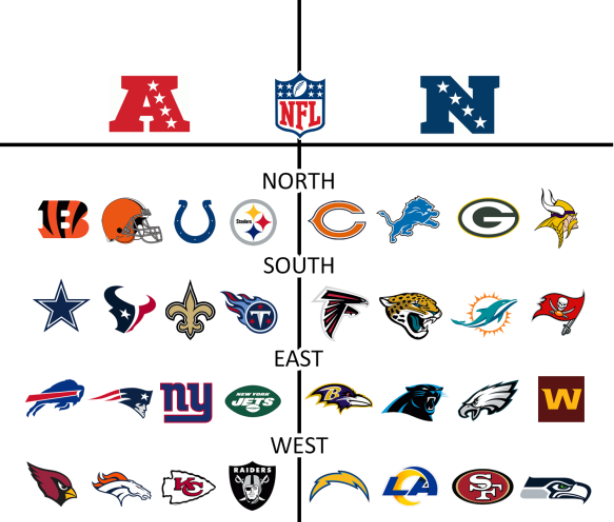
La tabla nos permite ver de un vistazo que existe una paridad o distribución uniforme entre el número de equipos de cada división y federación.

También podemos ver las frecuencias relativas y absolutas en términos de proporciones

El resultado es otro objeto de la clase table al que se le han añadido una o varias filas o columnas, que contienen las frecuencias marginales, tanto absolutas como relativas.

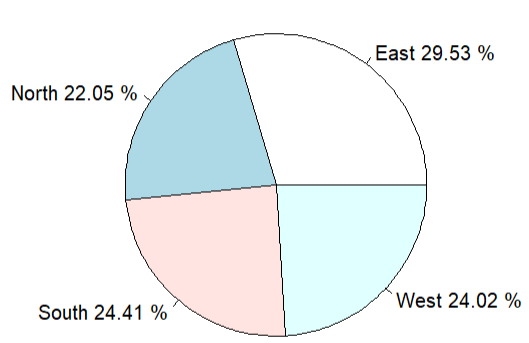




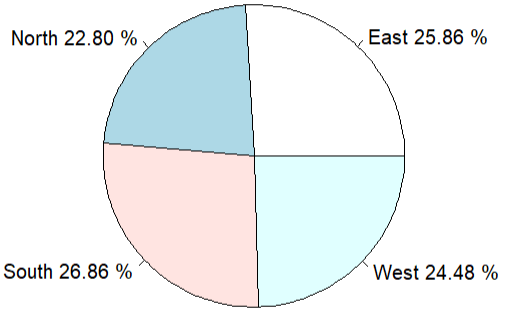
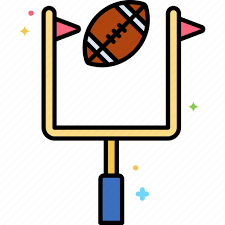


* 1. Variables Cuantitativas

**Porciento de victorias por división**



**Porciento de touchdowns por división**



Vamos a identificar cual es la división más competitiva calculando las medias de Touchdowns, yardas a favor y puntos a favor de cada una de ellas. Después de identificar las variables que vamos a utilizar comprobamos si existen outliers, en el caso de la variable yardas a favor existen 2 outliers uno inferior y otro superior, procedemos a identificarlos, ver a que división pertenecen y eliminar para calcular la media.

|  |  |  |
| --- | --- | --- |
| Boxplot de Touchdowns | Boxplot de yardas a favor | Boxplot de puntos a favor |
|  |  |  |
|  |  |  |

|  |  |
| --- | --- |
| Media de Touchdowns |  |
| Media de puntos a favor |  |
| Media de yardas a favor |  |
| Media yardas a favor sin outliers |  |

Podemos ver que los outliers tenían una influencia significativa en la media, si se tienen en cuenta la división del Sur es la más competitiva en todos los aspectos.



Con las variables cuantitativas de nuestra base de datos hemos medido el rendimiento de cada equipo, los hemos categorizado en 3 grupos según los Touchdowns y la diferencia de puntos obtenidos durante la temporada, donde podemos encontrar los siguientes grupos según el rendimiento de cada equipo:

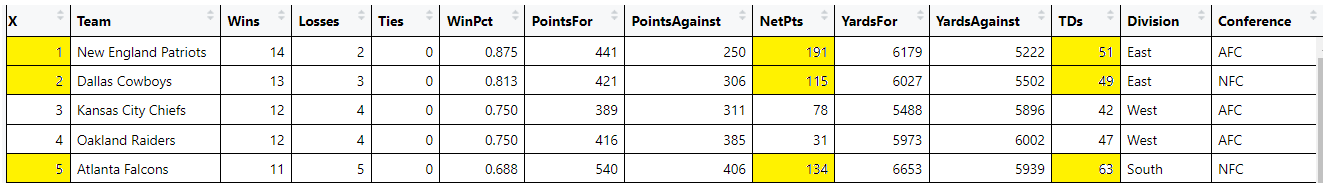
* Good: NetPts <= 10 y TDs <= 30
* Decent: NetPts <= 100 y TDs 30 - 45
* Excelent: NetPts >= 100 y TDs >= 45

|  |  |
| --- | --- |
| Distribución de frecuencias variable NetPts | Distribución de frecuencias variable TDs |
|  |  |

Hemos obtenido los siguientes resultados:

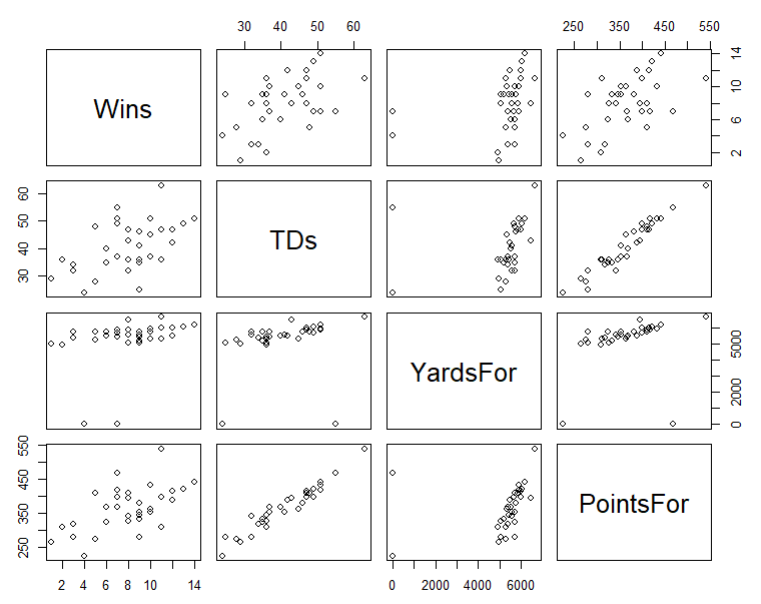
* Good: 14 equipos
* Decent: 15 equipos
* Excelent: 3 equipos

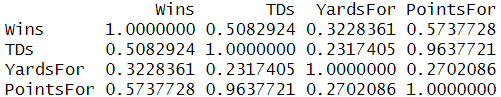
Podemos ver que los 3 equipos que han tenido una temporada excelente según nuestro criterio están muy bien posicionados en la calificación.



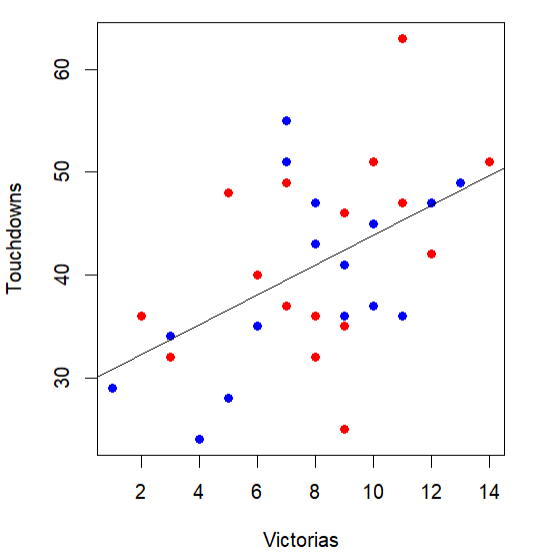
1. **Análisis de relaciones entre variables**

Observamos las relaciones entre las variables de nuestro dataset en la matriz de gráficos y tabla con los coeficientes de correlación.

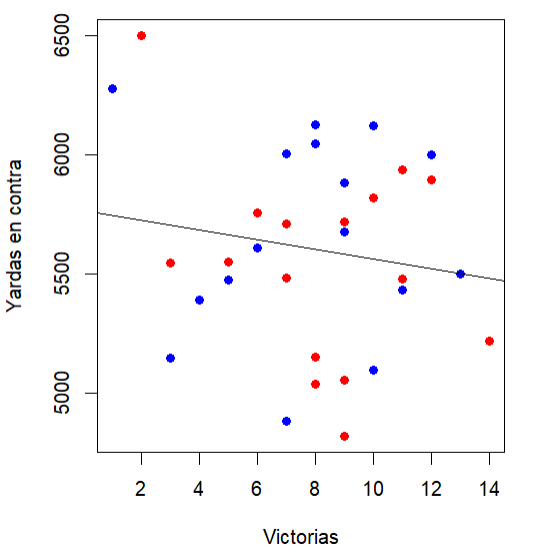




Hemos optado por dos relaciones. La primera es la relación entre victorias y Touchdowns destacar aquí que un Touchdown equivale a 6 puntos a favor. Está relación nos da un coeficiente de Pearson de 0.5082924 lo que quiere decir que existe una relación positiva moderada entre las dos variables, cuantos más touchdowns una mayor posibilidad de ganar. Está relación también se muestra en la siguiente figura:

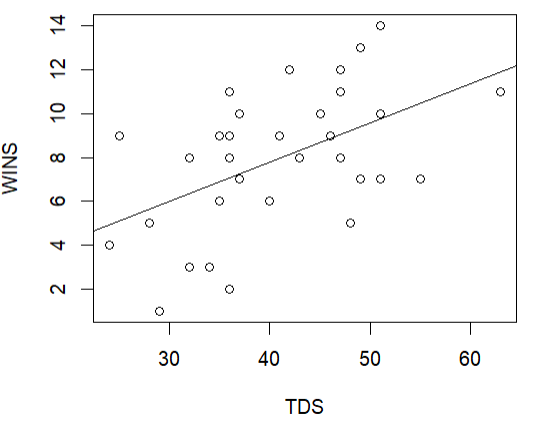
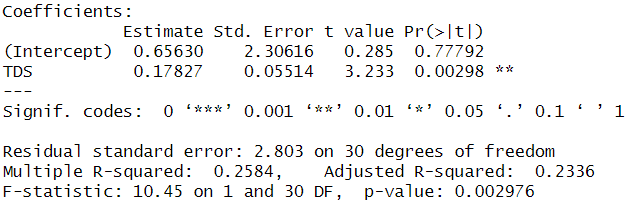


Otra relación podría darse entre las victorias y las yardas en contra en esta relación obtenemos un coeficiente de Pearson de -0.1561484 lo que nos indica que existe cierta relación inversa débil.



1. **Modelo de regresión lineal**

El modelo de regresión muestra las victorias y los touchdowns con un coeficiente de Pearson de 0.5082924 y una bondad de ajuste es de 0,2584.

****

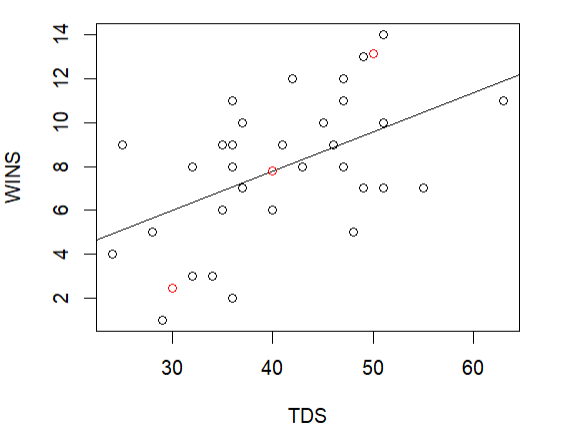
Según el modelo de regresión construido, la relación entre el numero de victorias y el número de touchdowns viene definida por la siguiente función Según el modelo de regresión construido, la relación entre el número de victorias y el número de touchdowns viene definida por la siguiente función:

|  |
| --- |
| y = 0.65630 + 0.17827 x |

En este modelo se toma como variable dependiente (X) el número de victorias y como variable independiente (Y) el número de touchdowns, debido a que el número de touchdowns tiene una mayor variabilidad que el número de victorias.

Según la predicción

* Un equipo con 10 touchdowns tiene entre 2 y 3 victorias (2.438999)
* Un equipo con 30 touchdowns tiene entre 7 y 8 victorias (7.787085)
* Un equipo con 70 touchdowns tiene entre 13 y 14 victorias (13.13517)

****

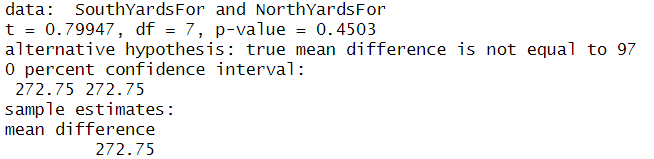
1. **Contrastes de hipótesis**

Se quiere comparar si la media de yardas a favor de la división South (Conferencias NFC y AFC) es igual a la media de yardas a favor de la división North. Para ello se realiza un test de hipótesis con un nivel de confianza del 97%.

* H0: Media yardas a favor South = Media yardas a favor North
* H1: Media yardas a favor South ≠ Media yardas a favor North

Es un contraste bilateral para la media



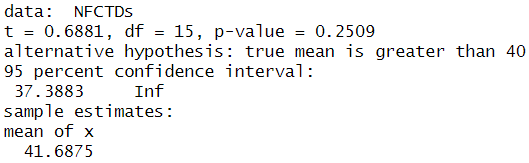


El p-value obtenido es 0,4552, mayor que el nivel de significación α = 0,03. Podemos aceptar la hipótesis H0 y concluir con una confianza del 97% que la media de puntos de la división del norte es igual a la de la división del sur.

Realizaremos un contraste de hipótesis para contrastar con un nivel de confianza del 95% que la media de touchdowns de la federación NFC es mayor a 40.

* H0: Media >= 40
* H1: Media < 40



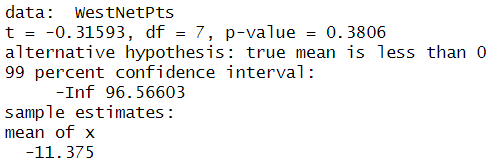


Como podemos observar el p-value es mayor que nuestro nivel de significación 0,05 por tanto, no se descarta H0 y podemos asegurar al 95% que la media de la NFC se encuentra por encima o igual a 40 touchdowns.

Realizaremos un contraste de hipótesis para contrastar con un nivel de confianza del 99% que la media de la NetPts de la división West es menor a 0.

* H0: Media NetPts West <= 0
* H1: Media NetPts West > 0





Como el p-value es mayor que el nivel de significación mayor que el nivel de significación aceptamos H0.