|  |  |
| --- | --- |
| Introducción a la ciencia de datos | Trabajo Final Teórico/Práctico  Francisco Pérez Hernández |

Tabla de contenido

[1 Introducción 3](#_Toc470197514)

[2 Descripción del tipo de datos de entrada 3](#_Toc470197515)

[2.1 Baseball 3](#_Toc470197516)

[2.1.1 Cálculo de media, desviación estándar, etc. 4](#_Toc470197517)

[2.1.2 Gráficos que permitan visualizar los datos adecuadamente. 6](#_Toc470197518)

[2.1.3 Descripción del conjunto de datos a partir de los puntos anteriores. 8](#_Toc470197519)

[2.2 Australian 8](#_Toc470197520)

[2.2.1 Cálculo de media, desviación estándar, etc. 9](#_Toc470197521)

[2.2.2 Gráficos que permitan visualizar los datos adecuadamente. 10](#_Toc470197522)

[2.2.3 Descripción del conjunto de datos a partir de los puntos anteriores. 11](#_Toc470197523)

[3 Regresión 12](#_Toc470197524)

[3.1 Regresión simple 12](#_Toc470197525)

[3.2 Regresión múltiple 12](#_Toc470197526)

[3.3 k-NN para regresión 12](#_Toc470197527)

[3.4 Comparación de algoritmos 12](#_Toc470197528)

[4 Clasificación 12](#_Toc470197529)

[4.1 k-NN 12](#_Toc470197530)

[4.2 LDA 13](#_Toc470197531)

[4.3 QDA 13](#_Toc470197532)

[4.4 Comparación de algoritmos 13](#_Toc470197533)

# 1 Introducción

Las bases de datos con las que he trabajado han sido “baseball” para regresión y “australian” para clasificación:

**Dataset Regresión: Baseball**

**Dataset Clasificación: Australian**

Voy a desarrollar lo explicado en clase sobre estos dataset y a analizar los resultados obtenidos. En la primera sección se analizarán las bases de datos asignadas donde se conocerán las estructuras y cuál es el objetivo de cada una de ellas. Esta información se ha extraído de la web Keel, junto a sus archivos asociados, los cuales los he conseguido de los siguientes enlaces:

**Baseball:** [**http://sci2s.ugr.es/keel/dataset.php?cod=76#sub1**](http://sci2s.ugr.es/keel/dataset.php?cod=76#sub1)

**Australian:** [**http://sci2s.ugr.es/keel/dataset.php?cod=53#sub1**](http://sci2s.ugr.es/keel/dataset.php?cod=53#sub1)

De forma que se ha podido extender la información facilitada en los documentos subidos por los profesores.

En la segunda sección se realiza un estudio de regresión sobre el dataset “baseball”, donde buscaré modelos que expliquen el problema. Compararé estos modelos y sus ajustes para encontrar el mejor. Finalmente realizaré el test de Wilcoxon y Friedman para los distintos modelos.

En la tercera sección realizaré un estudio similar a regresión, pero esta vez será sobre el dataset “australian” y es un problema de clasificación.

# 2 Descripción del tipo de datos de entrada

## 2.1 Baseball

El dataset Baseball contiene 17 variables/columnas de las que 16 son entradas y 1 salida. Las entradas son: "Batting\_average", "On-base\_percentage", "Runs", "Hits", "Doubles", "Triples", "HomeRuns", "Runs\_batted\_in", "Walks", "Strike-Outs", "Stolen\_bases", "Errors", "Free\_agency\_eligibility", "Free\_agent", "Arbitration\_eligibility", “Arbitration”. La salida es “Salary”.

Tenemos en total 337 instancias/observaciones y en ninguna hay valores perdidos. Son datos reales, y los tipos de cada variable son 2 reales y 14 enteros. Las dos variables reales son “Batting\_average” y “On-base\_percentage”.

Como podemos ver en la web, tenemos que, de las 14 variables de números enteros, tenemos 4 variables que solo contiene valores de 0 o 1: “Free\_agency\_eligibility”, “Free\_agent”, “Arbitration\_elegibility” y “Arbitration”.

Como información adicional tenemos que: “El dataset contiene los salarios de 1992 de un conjunto de jugadores de la “Major League Baseball” que jugaron al menos un partido en la temporada 1991 y 1992, excluyendo a los pitchers (lanzadores). Para cada jugador, se proporcionan algunas medidas de rendimiento junto con cuatro variables categóricas que indican cuán libre era cada uno para formar parte de otros equipos.”

### 2.1.1 Cálculo de media, desviación estándar, etc.

Para obtener la información la distribución de los datos, su media, valores mínimos, máximos, cuartiles… se ha hecho la función *summary():*

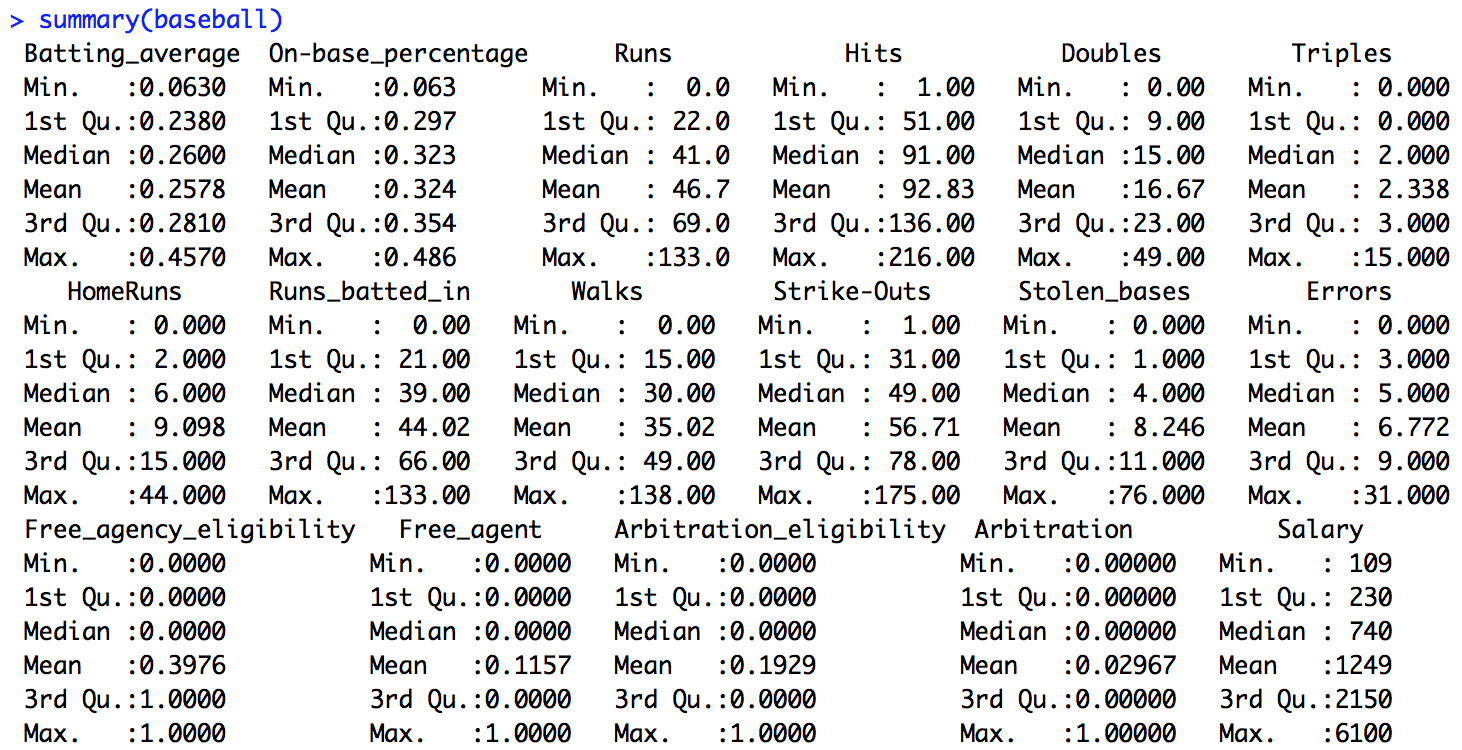


Figura 1: Resumen de distribución de los datos en cada variable para el dataset “baseball”

Si queremos la desviación estándar, con *apply()* podemos obtener la información:

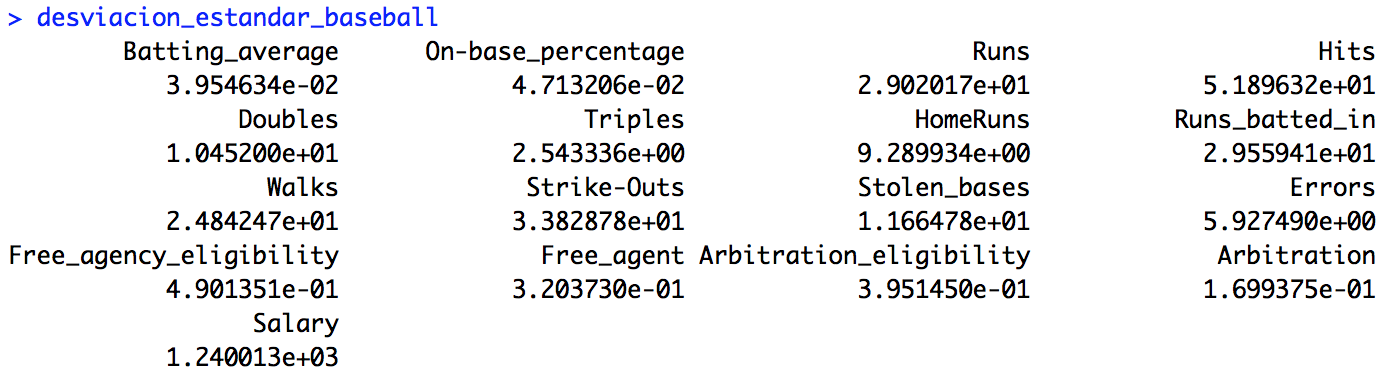


Figura 2: Desviación estándar para el dataset “baseball”

Para la varianza:

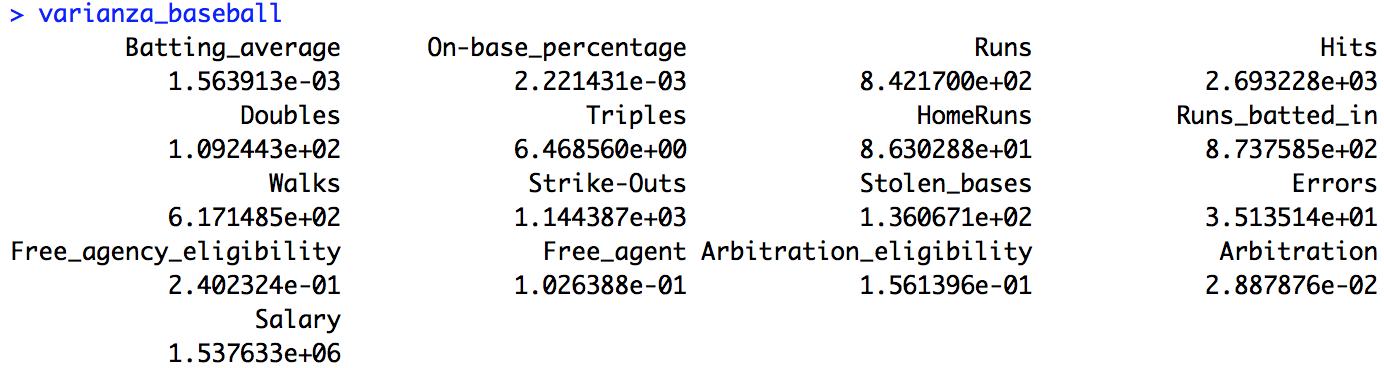


Figura 3: Varianza para el dataset “baseball”

Para la desviación absoluta de la mediana:

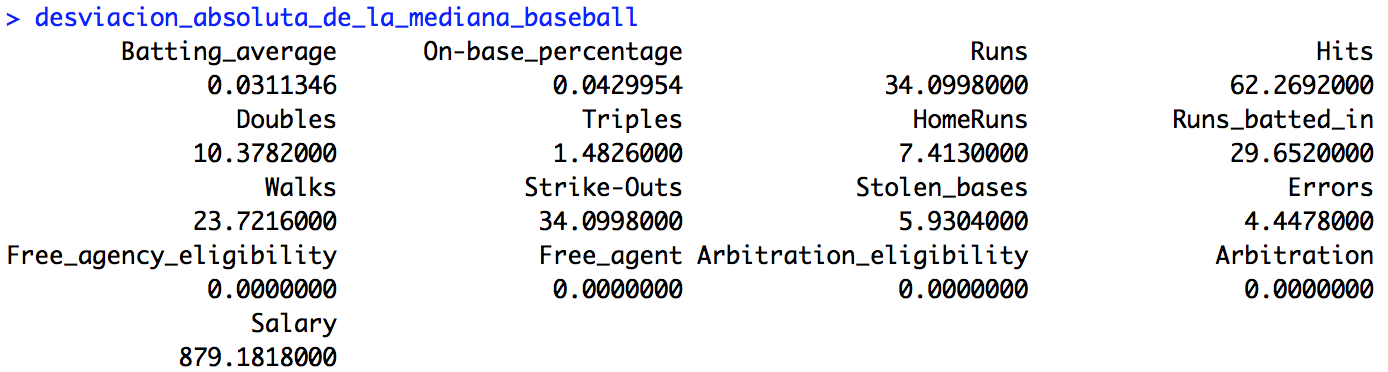


Figura 4: Desviación absoluta de la mediana para el dataset “baseball”

Para el rango intercuartil:

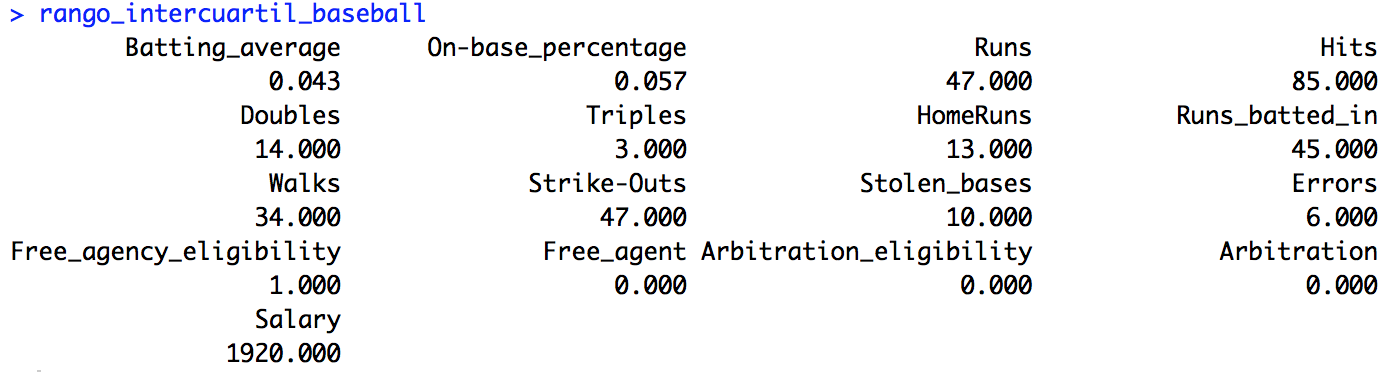


Figura 5: Rango intercuartil para el dataset “baseball”

### 2.1.2 Gráficos que permitan visualizar los datos adecuadamente.

Si queremos ver las gráficas de todas las variables con respecto a todas, podemos hacer uso de la función *plot()*

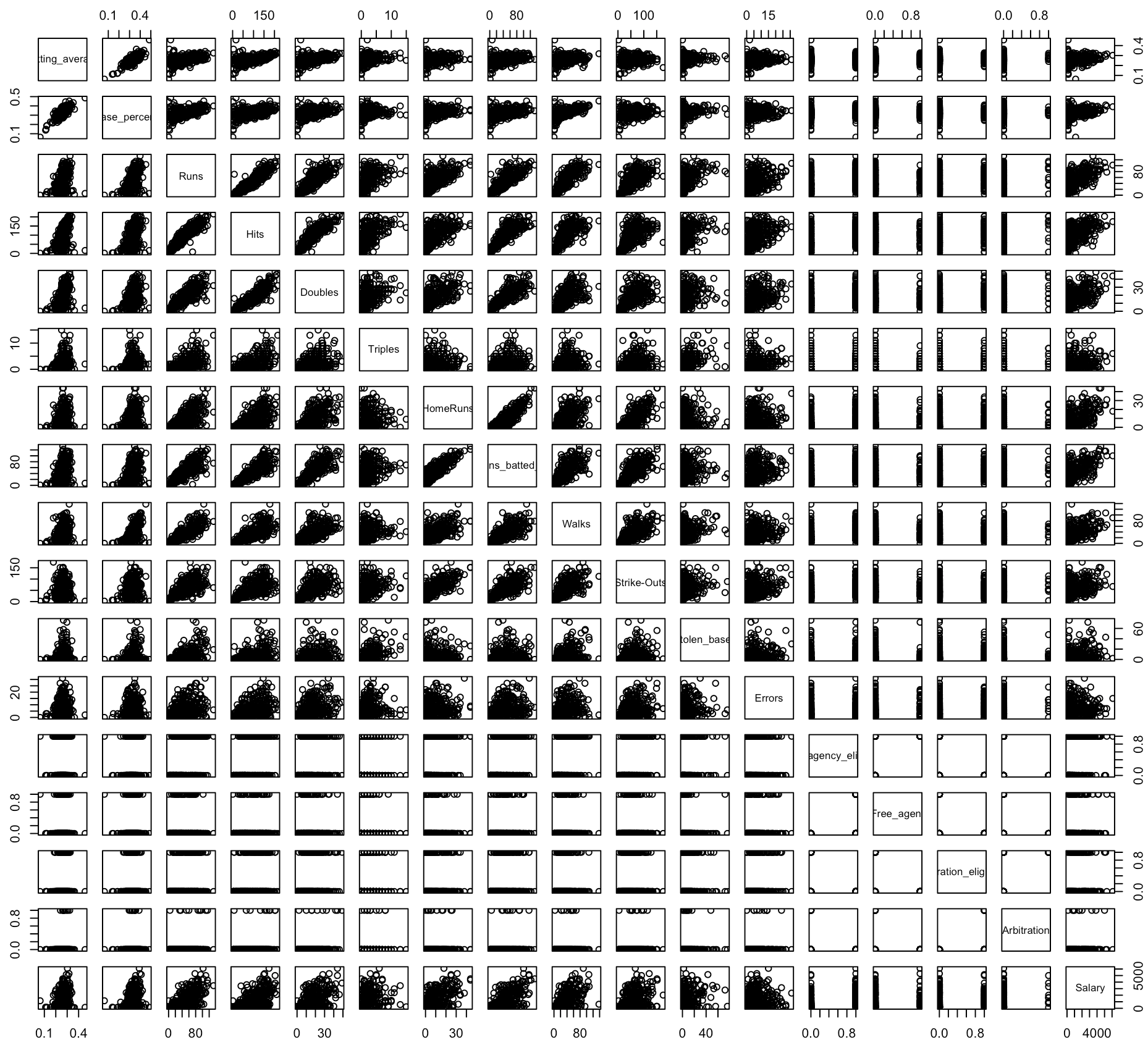


Figura 6: Plot de todas las variables del dataset “baseball”

Pero si lo que queremos es ver todas las variables con respecto de la variable de salida, es decir *Salary,* entonces tendremos:

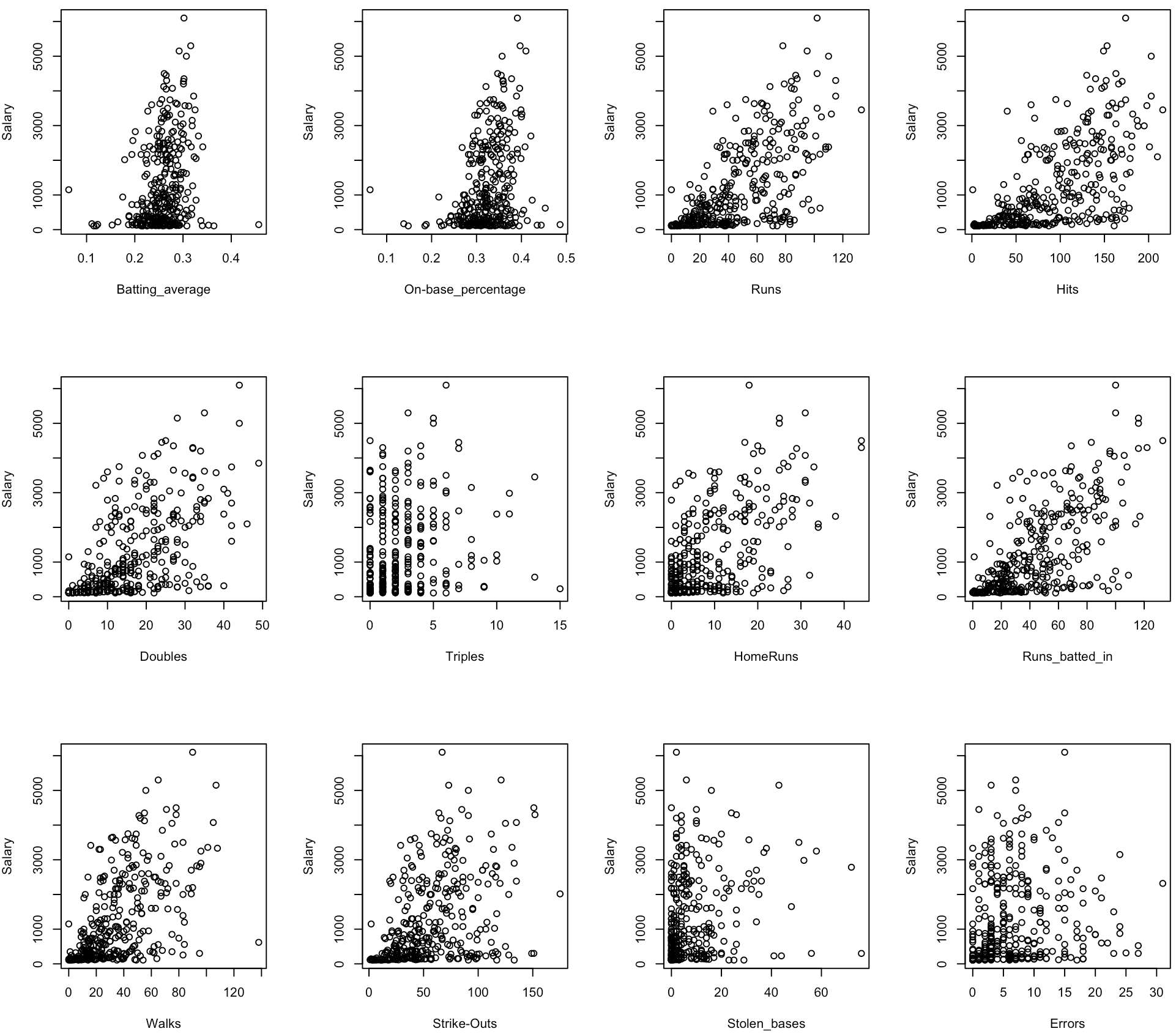


Figura 7: Plot de las primeras 12 variables

Además, tendremos:

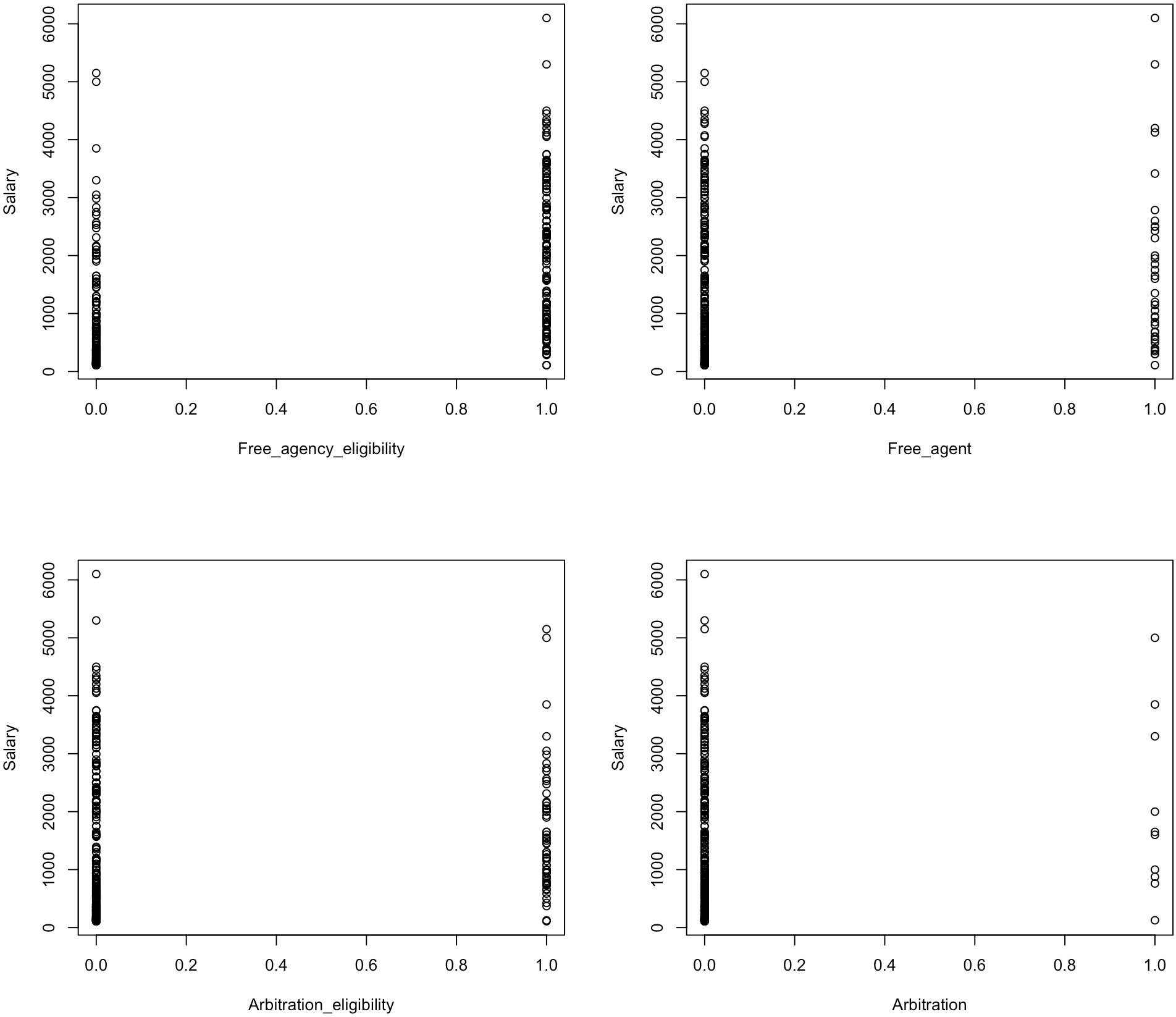


Figura 8: Plot de las variables con valores de 0 o 1.

### 2.1.3 Descripción del conjunto de datos a partir de los puntos anteriores.

En resumen, tenemos un problema de regresión en el que se pretende realizar un modelo para obtener como variable respuesta el salario de los jugadores de un equipo de béisbol. Para ello contamos con 16 variables, de las cuales 4 se mueven entre valores de 0 o 1 y el resto son enteros o valores reales. Hemos podido observar la distribución de estas variables y sus gráficas con respecto a la variable de salida. A simple vista no se podría decir mucho sobre que variables son las más importantes, pero puede decirse que algunas de las más influyentes pueden ser: “Hits”, “Runs”, “HomeRuns” … por lo cual miraremos la matriz de correlación.

Como la correlación simple se da con valores cercanos a 1 o -1, indicando alta correlación entre cada par de variables, y valores cercanos a 0 indican el caso contrario, se tiene que las variables correladas son:

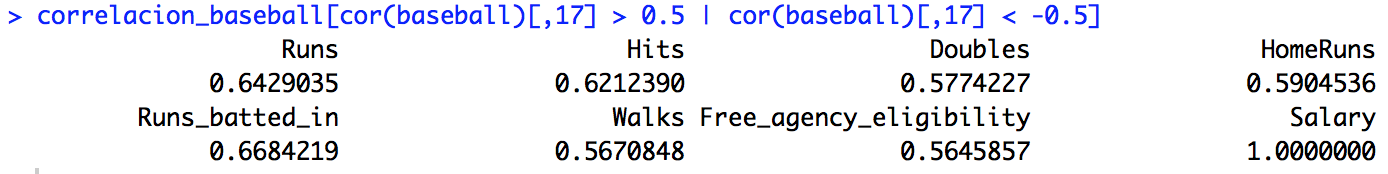


Figura 9: Correlación simple entre cada variable con la variable respuesta

Donde tenemos “Runs”, “Hits”, “Doubles”, “HomeRuns”, “Runs\_batted\_in”, “Walks” y “Free\_agency\_elegibility”. Es decir, estas son las variables que tendremos más en cuenta para nuestros modelos.

## 2.2 Australian

El dataset Australian contiene 15 variables/columnas de las que 14 son entradas y 1 salida. Las variables de este dataset no tienen ningún nombre por lo que no podemos describirlas. Este es un problema de clasificación en el que teneos 2 clases, la 0 y la 1.

Tenemos en total 690 instancias/observaciones y en ninguna hay valores perdidos. Son datos reales, y los tipos de cada variable son 3 reales, 5 enteros y 6 nominales.

Como podemos ver en la web, tenemos que, de las 14 variables de números enteros, tenemos 4 que se mueven entre 0 o 1. Estas son las variables 1, 8, 9 y 11, además de la de salida, la 15.

Como información adicional tenemos que: “El archivo concierne a la aplicación de tarjetas de crédito. Todos los nombres de los atributos y los valores han sido cambiados para proteger la confidencialidad de los datos. El dataset es interesante porque hay una buena mezcla de atributos: continuos, nominales con pequeños valores de números, y nominales con grandes valores numéricos.”

### 2.2.1 Cálculo de media, desviación estándar, etc.

Para obtener la información la distribución de los datos, su media, valores mínimos, máximos, cuartiles… se ha hecho la función *summary():*

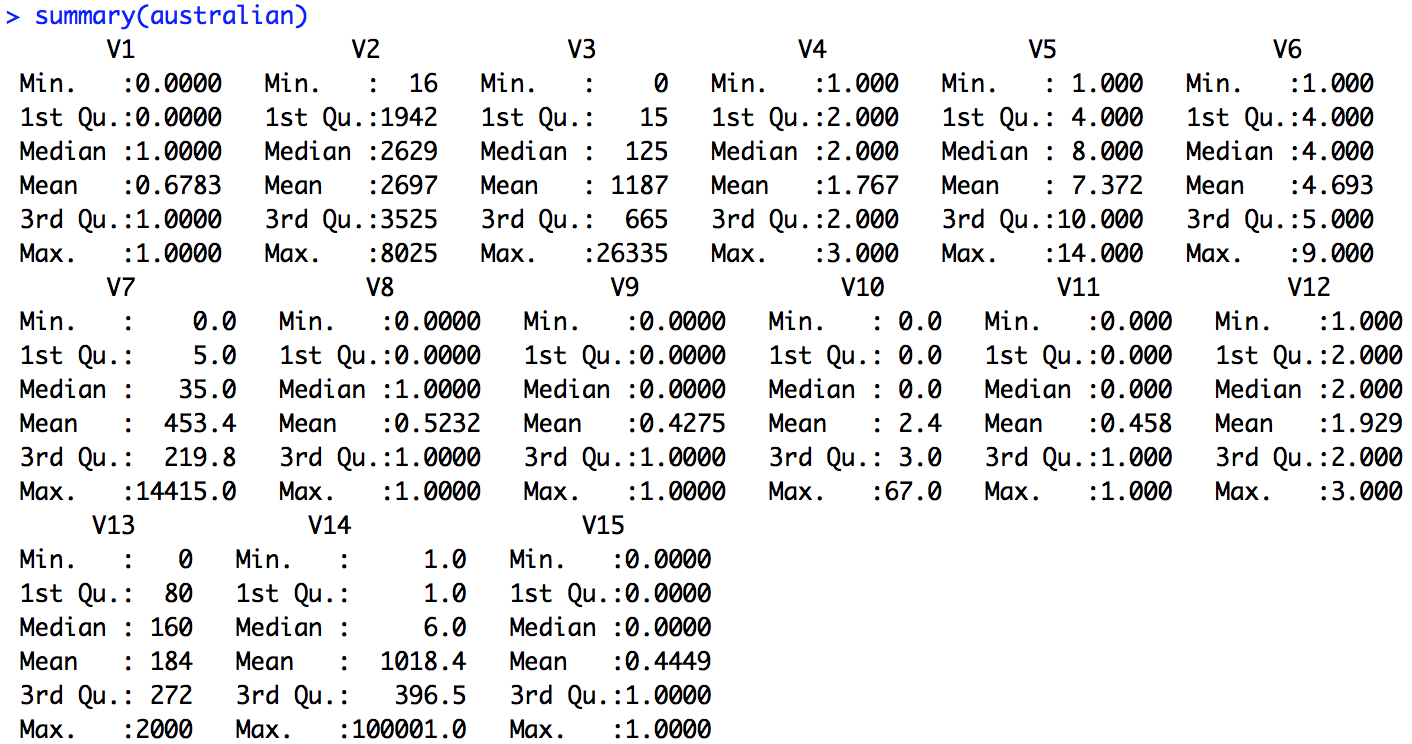


Figura 10: Resumen de distribución de los datos en cada variable para el dataset “australian”

Si queremos la desviación estándar, con *apply()* podemos obtener la información:

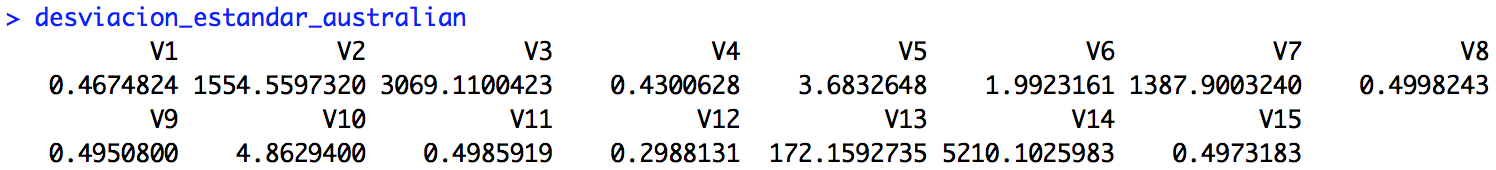


Figura 11: Desviación estándar para el dataset “australian”

Para la varianza:

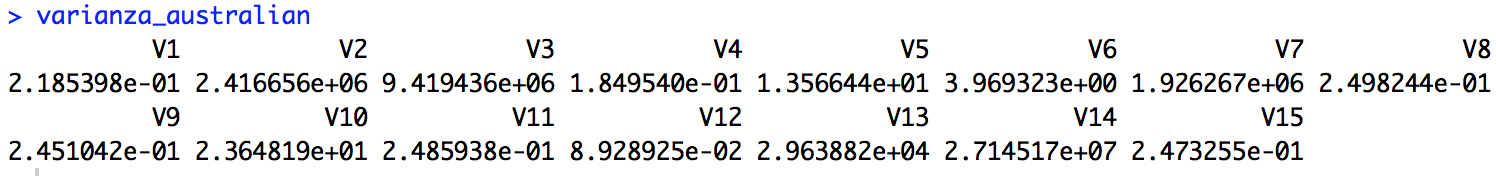


Figura 12: Varianza para el dataset “australian”

Para la desviación absoluta de la mediana:

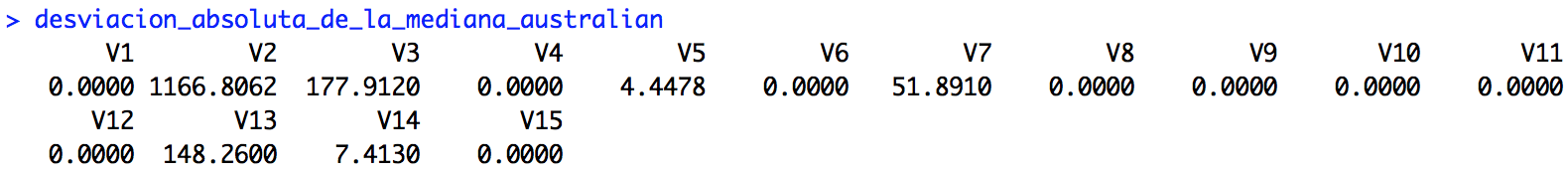


Figura 13: Desviación absoluta de la mediana para el dataset “australian”

Para el rango intercuartil:

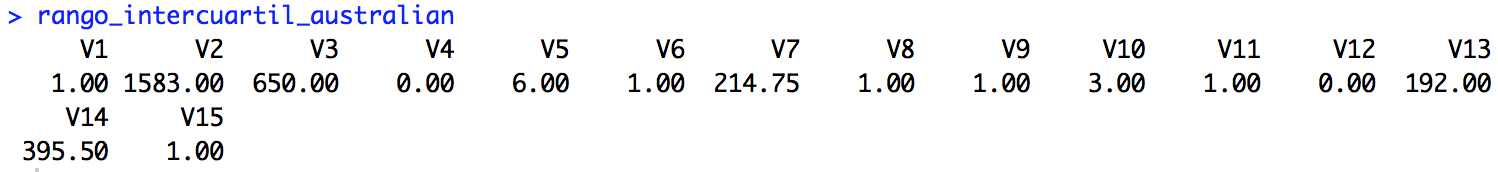


Figura 14: Rango intercuartil para el dataset “australian”

### 2.2.2 Gráficos que permitan visualizar los datos adecuadamente.

Si queremos ver las gráficas de todas las variables con respecto a todas, podemos hacer uso de la función *plot()*

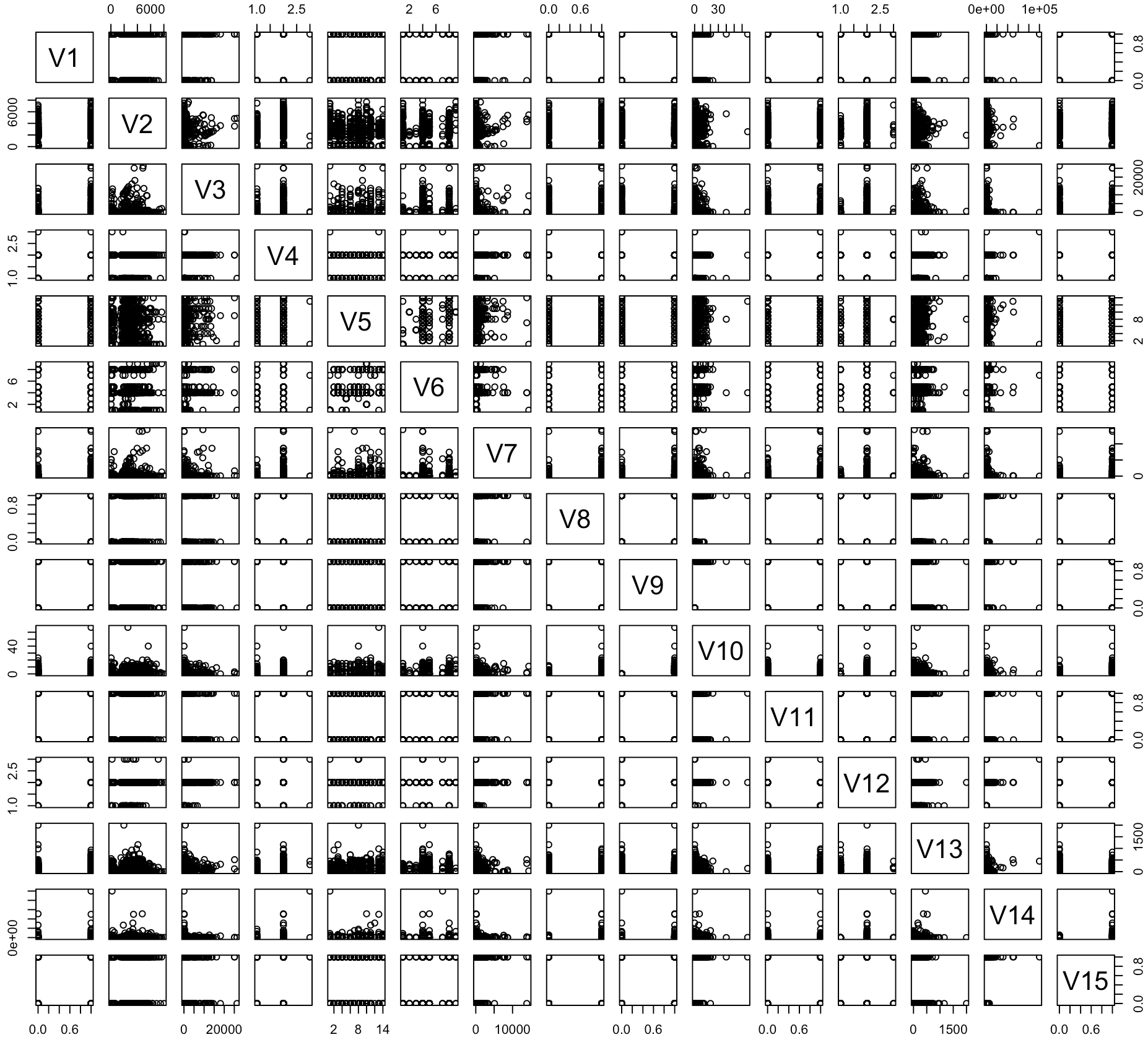


Figura 15: Plot de todas las variables del dataset “australian”

Pero si lo que queremos es ver todas las variables con respecto de la variable de salida, es decir la variable 15, entonces tendremos:

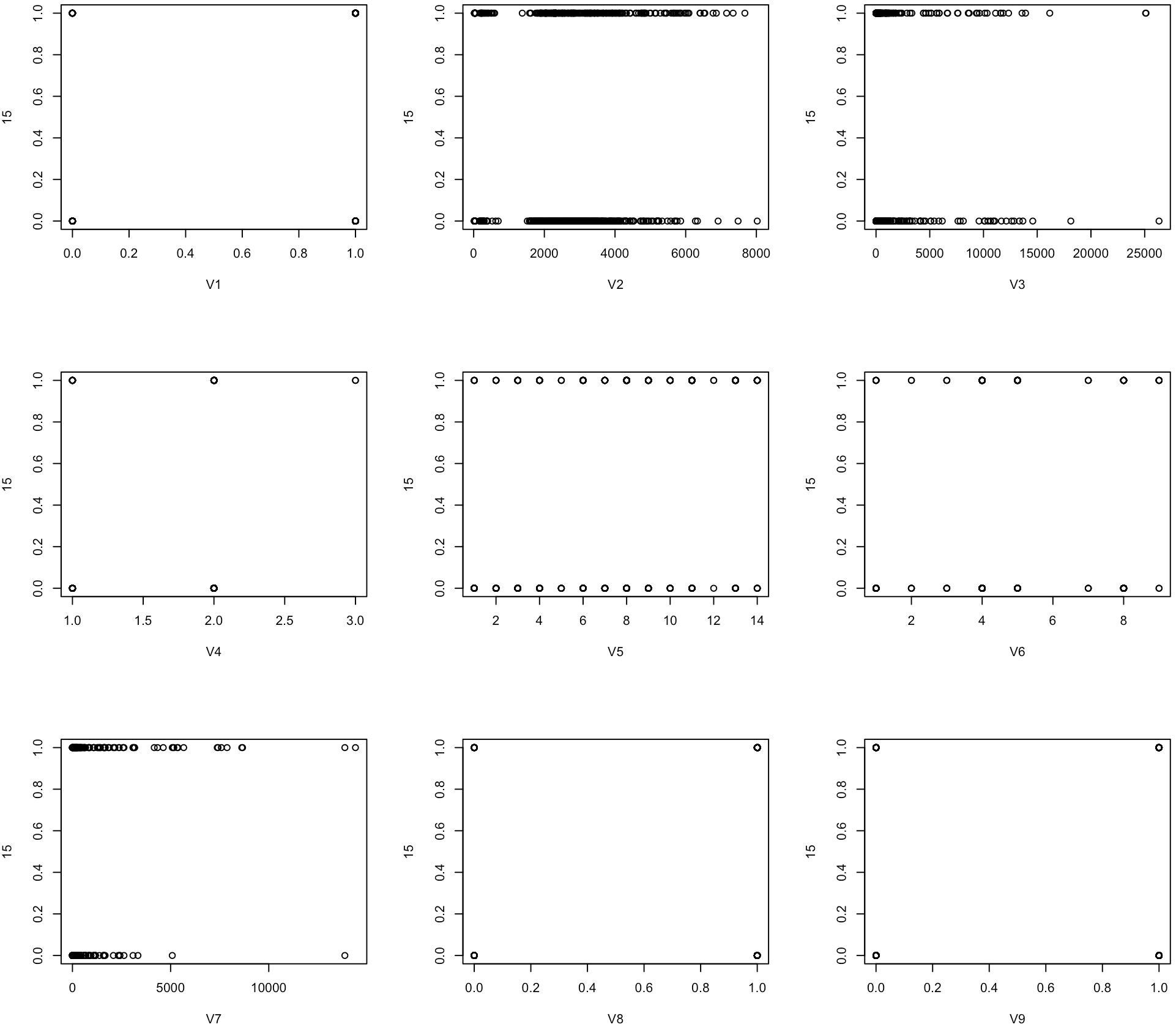


Figura 16: Plot de las primeras 9 variables

Además, tendremos:

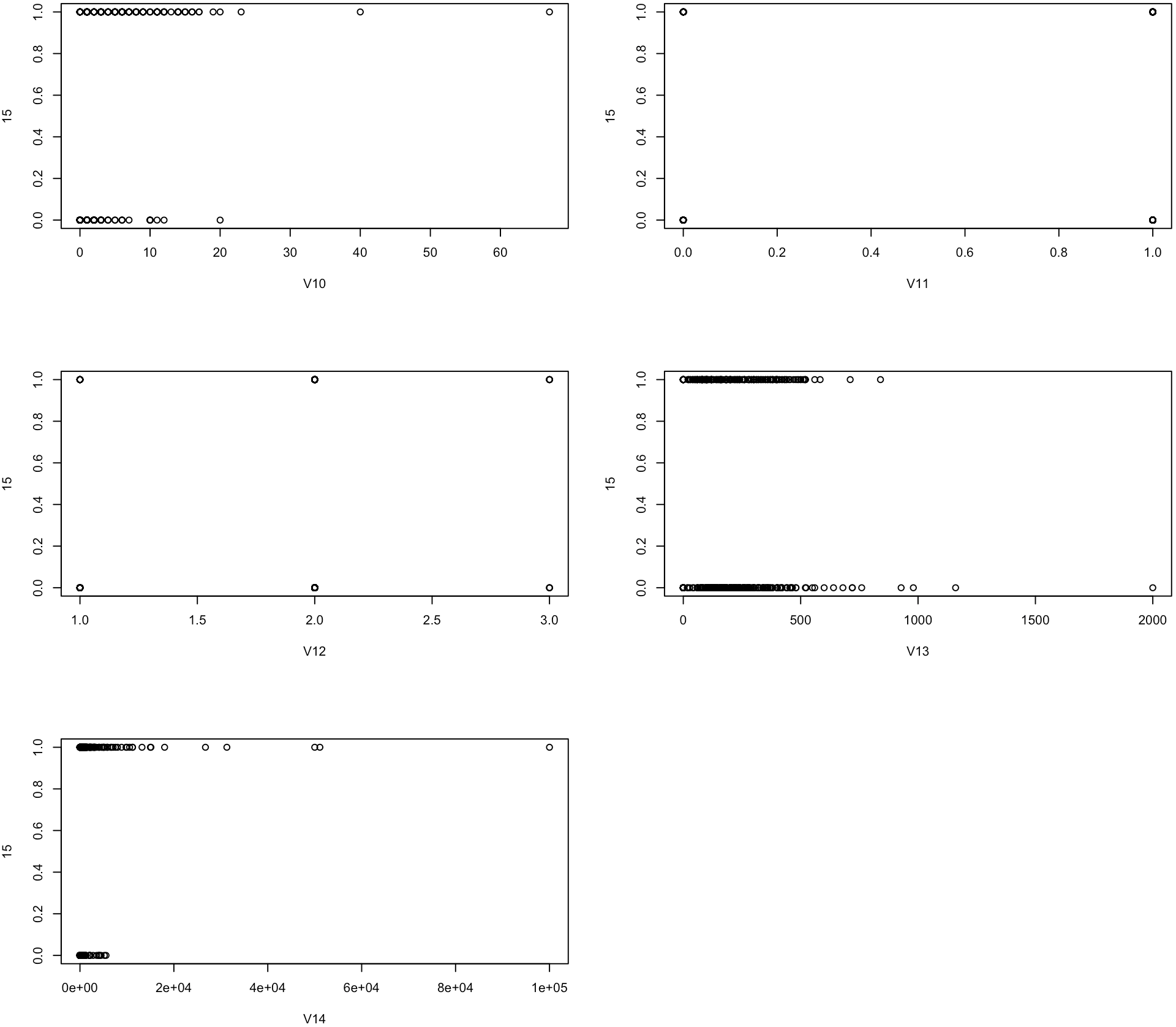


Figura 17: Plot del resto de variables.

### 2.2.3 Descripción del conjunto de datos a partir de los puntos anteriores.

En resumen, tenemos un problema de clasificación en el que se pretende realizar un modelo para obtener como variable respuesta una de las dos clases. Para ello contamos con 14 variables, de las cuales 4 se mueven entre valores de 0 o 1 y el resto son enteros, valores reales o nominales. Hemos podido observar la distribución de estas variables y sus gráficas con respecto a la variable de salida.

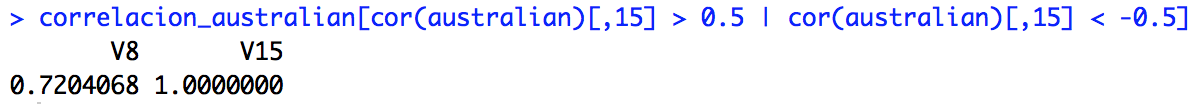


Figura 18: Correlación simple entre cada variable con la variable respuesta.

A simple vista no se podría decir mucho sobre que variables son las más importantes, pero con ayuda de la matriz de correlación, vemos como la Variable 8 está fuertemente correlada con la variable predictora, por lo que la tendremos muy en cuenta cuando realicemos nuestros modelos.

# 3 Regresión

## 3.1 Regresión simple

Para el estudio de la regresión simple, voy a usar las variables que tienen más correlación ya que nuestro dataset tiene más de 5 variables.

Estas variables nos habían salido:

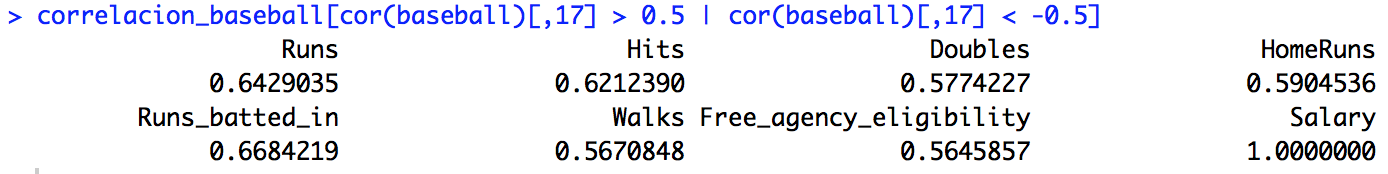


Figura : Correlación simple entre cada variable con la variable respuesta

Donde tenemos “Runs”, “Hits”, “Doubles”, “HomeRuns”, “Runs\_batted\_in”, “Walks” y “Free\_agency\_elegibility”. Por lo tanto, yo voy a elegir como las más importantes a: “Runs”, “Hits”, Runs\_batted\_in”, “HomeRuns” y “Doubles”. Veamos que tal salen estos modelos.

La visualización de estos modelos nos queda:

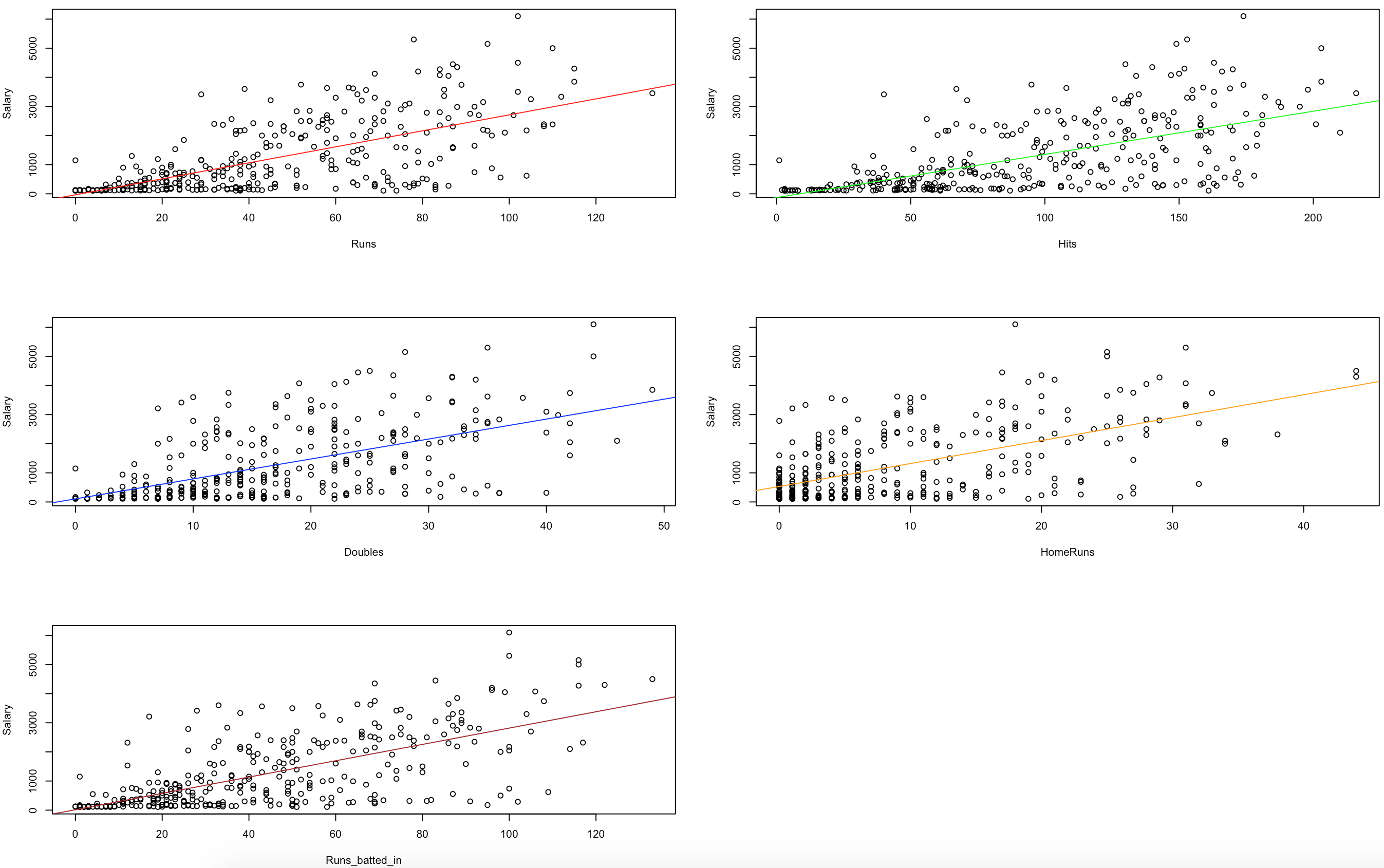


Figura : Visualización de los modelos

Además, si queremos ver los datos más importantes:

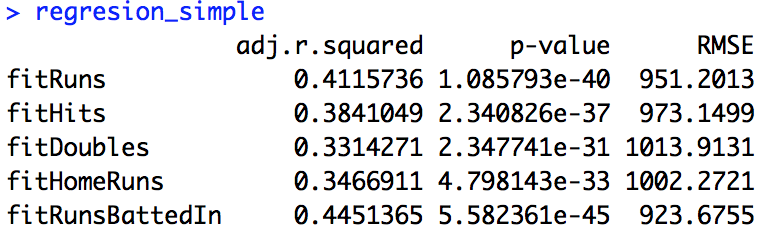


Figura : Datos más importantes de los modelos

De donde vemos como los modelos tiene un bajo valor de R^2 por lo que debemos seguir mejorando los modelos.

## 3.2 Regresión múltiple

## 3.3 k-NN para regresión

## 3.4 Comparación de algoritmos

# 4 Clasificación

## 4.1 k-NN

## 4.2 LDA

## 4.3 QDA

## 4.4 Comparación de algoritmos