

# Laboratorio

# 3

## AlpesGamesInsight

Juan Camilo Castiblanco - 201630025

Maria Alejandra Escalante - 201631008

Juan Sebastian Cabra - 201630689

# Contenido

**01** Introducción

**02** Análisis y perfilamiento de los datos

**03** Entendimiento de las variables

**04** Selección de variables del modelo

**05** Modelos

**06** Análisis de resultados y recomendación

# Introducción

01

## AlpesGamesInsight

Uno de sus factores diferenciadores es el cálculo de un ranking de los diferentes jugadores por video juego. Este ranking sirve como referencia para los equipos en el momento de seleccionar jugadores para competir y para los patrocinadores quienes apoyan económicamente las carreras de los jugadores.

La empresa quiere parar de utilizar expertos para el ranking de sus jugadores y comenzar a utilizar modelos basados en aprendizaje de máquina. Para lo cual, su equipo debe desarrollar un modelo que permita estimar el ranking de un cierto jugador dadas sus estadísticas de juego y evaluar su desempeño.

02

# Análisis y perfilamiento de datos

**Para este laboratorio**

Se trabajo con los datos suministrados por AlpesGamesInsight, estos datos contienen información estadística de jugadores de Starcraft 2

Los datos suministrados por AlpesGamesInsight están organizados en las siguientes variables. Cada variable tiene una serie de restricciones dadas por la empresa.

Variable	Descripción	Valores posibles
LeagueIndex	Ranking del Jugador	enteros entre 1 y 10
Age	Edad del jugador	enteros entre 10 y 100
HoursPerWeek	Horas jugadas por semana	Reales positivos
TotalHours	Total de horas jugadas	Reales positivos
APM	Acciones por minuto	Continua
SelectByHotkeys	Número de selecciones de unidades o edificios realizadas mediante teclas de acceso rápido por marca de tiempo	Continua
AssignToHotkeys	Numero de asignaciones a teclas de acceso rápido de unidades o edificios realizadas por marca de tiempo	Continua
UniqueHotkeys	Numero de teclas de acceso rapido unicas usadas por marca de tiempo	Continua
MinimapAttacks	Numero de acciones de ataque en minimap por marca de tiempo	continua
MinimapRightClicks	Numero de clicks en minimap por marca de timepo	Continua
NumberOfPACs	numero de ciclos de percepcion accion por marca de tiempo	Continua
GapBetweenPACs	Duracion media en milisegundos entre PACs	Continua
ActionLatency	Latencia media desde el inicio de un PACs hasta su primera accion en milisegundos	Continua
ActionsInPAC	Numero promedio de acciones en cada PAC	Continua
TotalMapExplored	El numero de 24x24 coordenadas vistas por el jugador por unidad de tiempo	continua
WorkersMade	Numero de SCVs, drones y sondas entrenadas por unidad de tiempo	Continua
UniqueUnitsMade	Unidades unicas hechas por unidad de tiempo	Continua
ComplexUnitsMade	Numero de fantasmas, infestors, y high templars entrenados por unidad de tiempo	Continua
ComplexAbilitiesUsed	Habilidades que requieren intrucciones de orientacion especifica usadas por unidad de tiempo	Continua

# Este laboratorio

## 2.1. Limpieza y preparación de los datos

Lo primero que se hizo fue eliminar los registros nulos y lo duplicados ya que estos no representan una cantidad significativa de los datos.

### Lo siguiente fue establecer restricciones para algunas de las variables

HoursPerWeek: Si un jugador pasa 16 horas de cada día en el juego puede lograr máximo 112 horas a la semana. Se eliminan los registros con mas de 112 horas a la semana. También, se eliminan los registros con 0 horas a la semana ya que no nos interesan los datos de personas que nunca han jugado el juego (esas personas no tendrán una liga asignada)

TotalHours: el juego tiene 10 años por lo que el máximo de horas posibles es 80640. Se eliminaron los registros con un valor superior

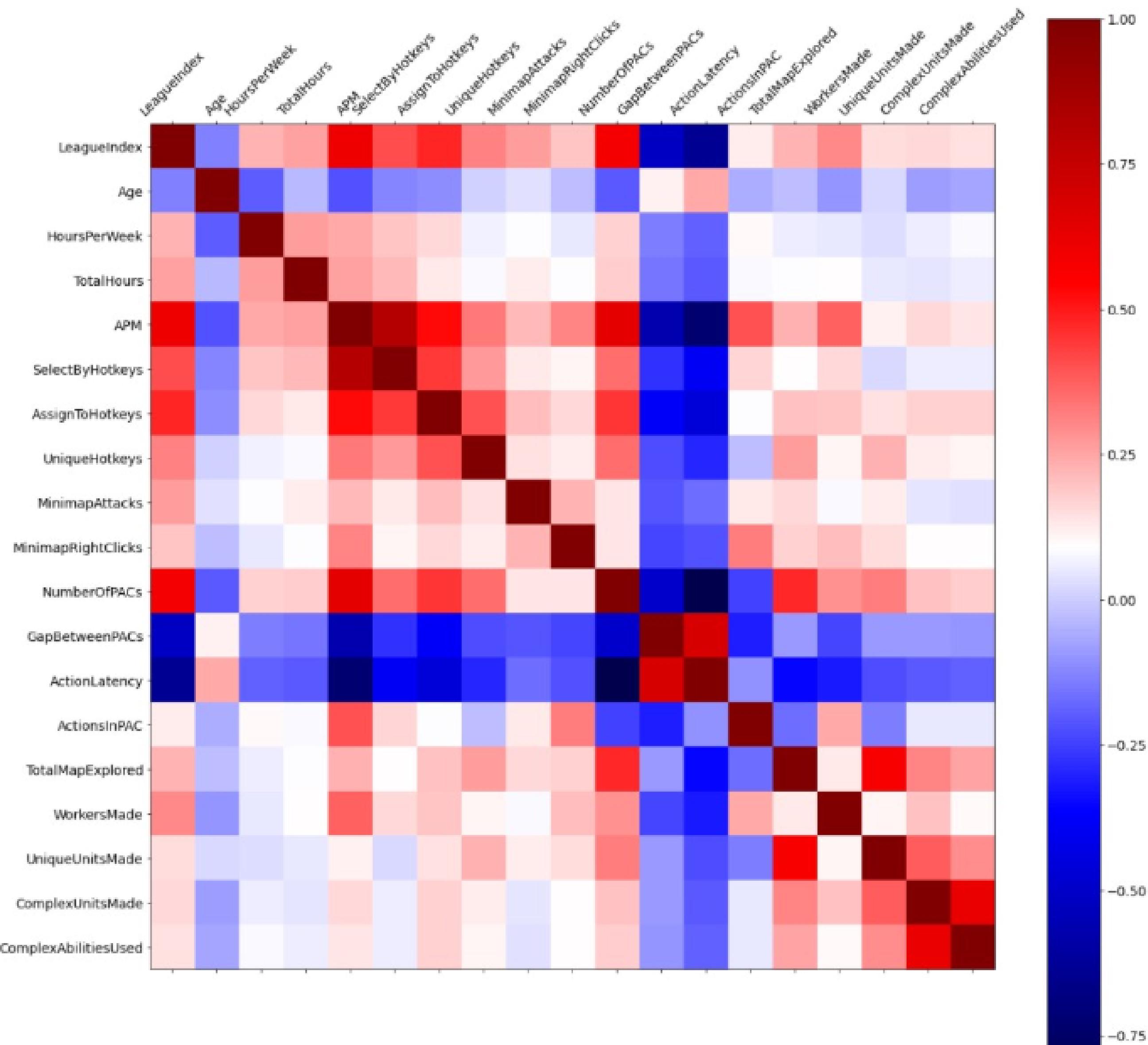
LeagueIndex: Se eliminaron los registros con league index mayor a 10, ya que el numero de ligas va de 0 a 10

03

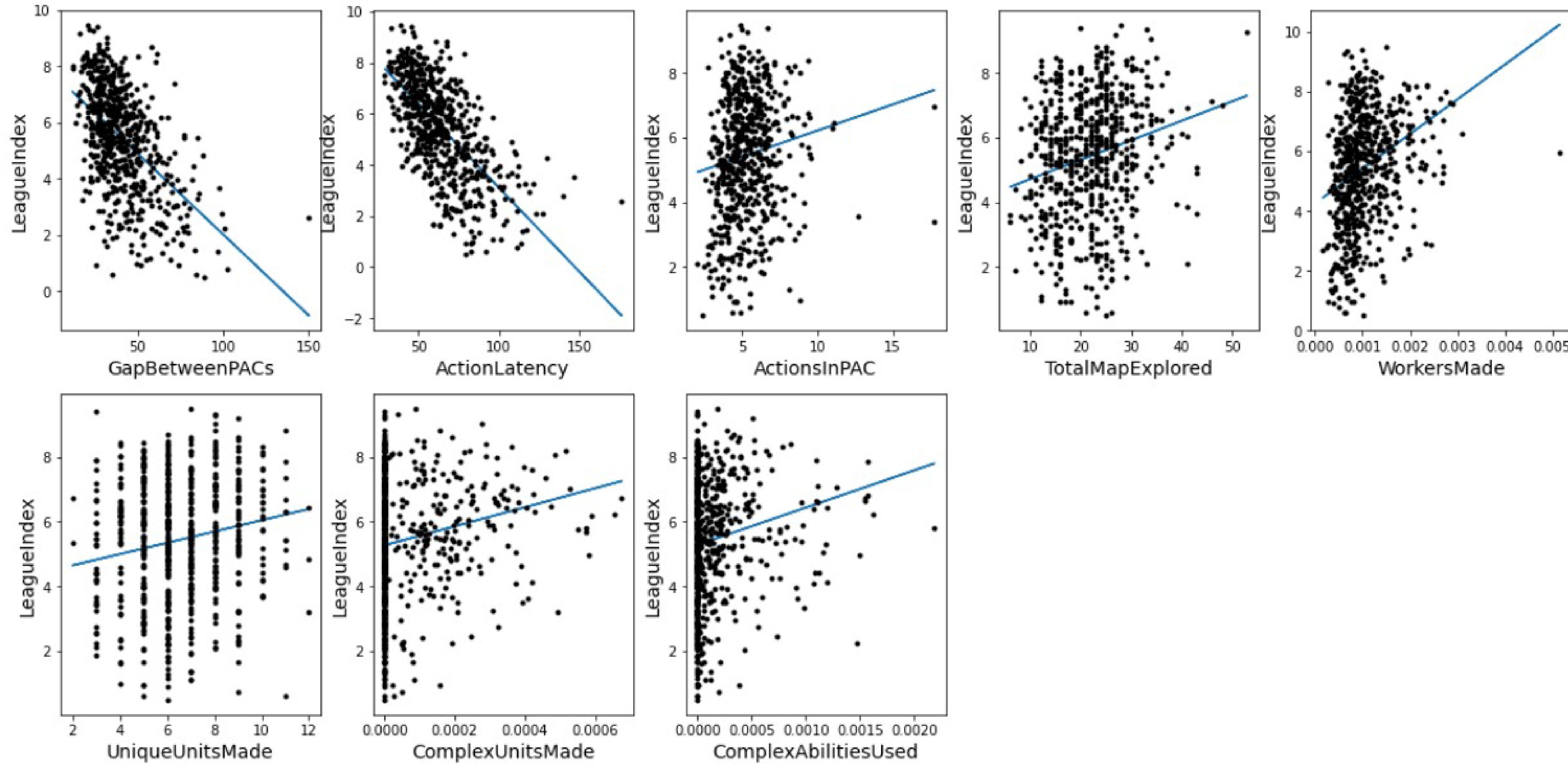
# Entendimiento de las variables

Lo siguiente que se realizo fue una matriz de correlación con el objetivo de verificar que no hubiese colinealidad entre las variables ya que esto lleva a problemas en la regresión

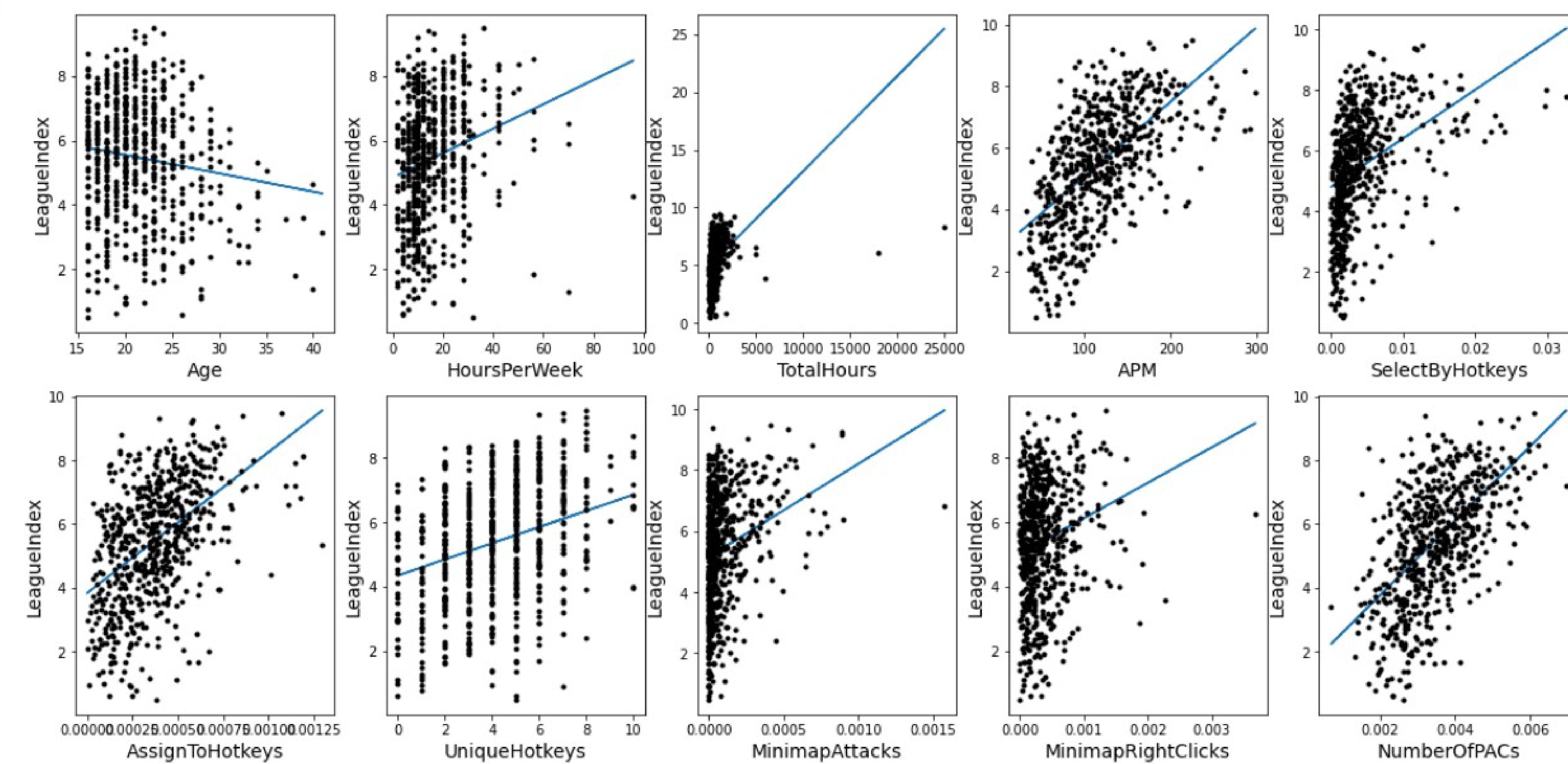
# Matriz de correlación



# Linealidad



# Linealidad

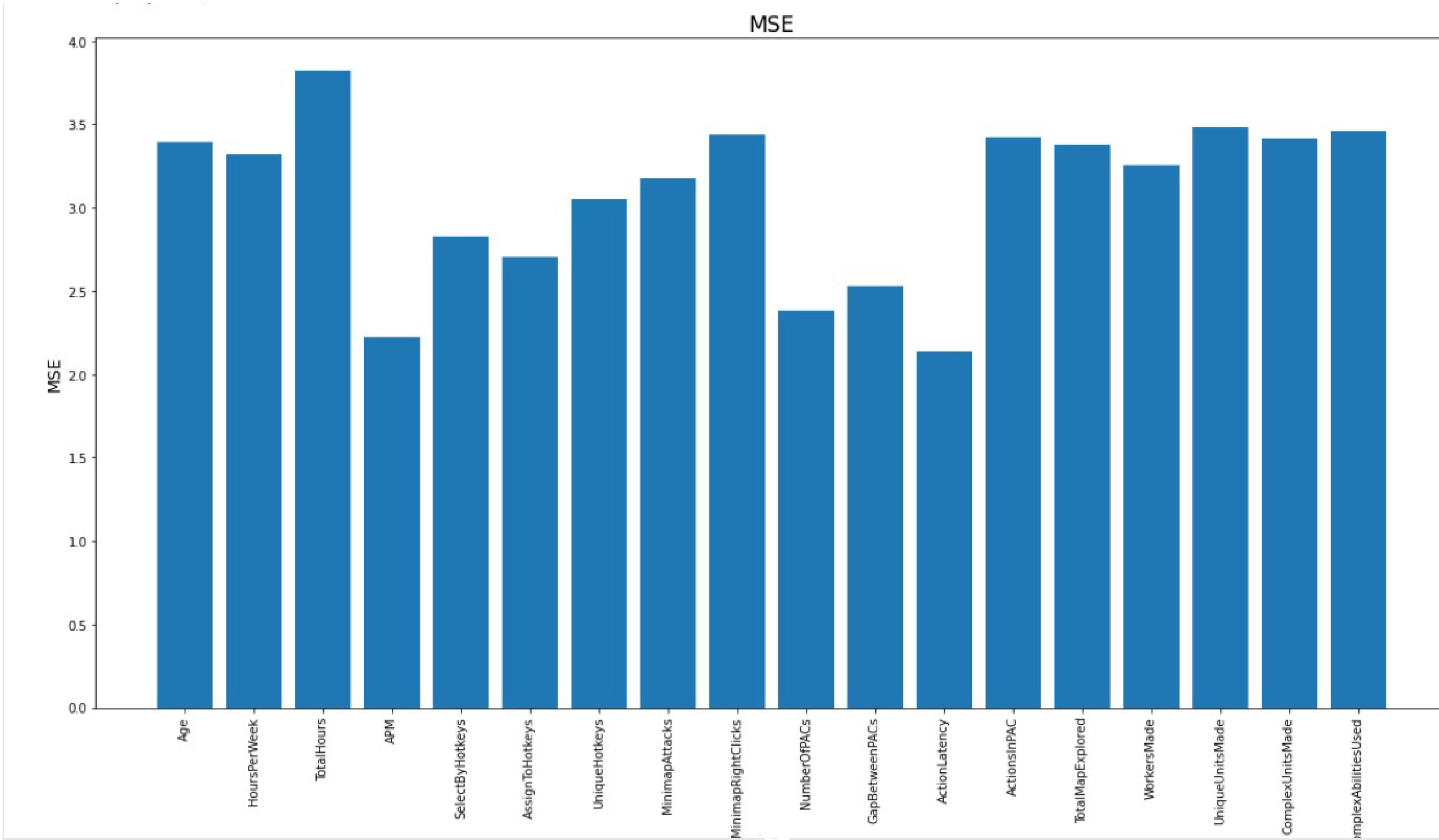


# Selección de variables del modelo

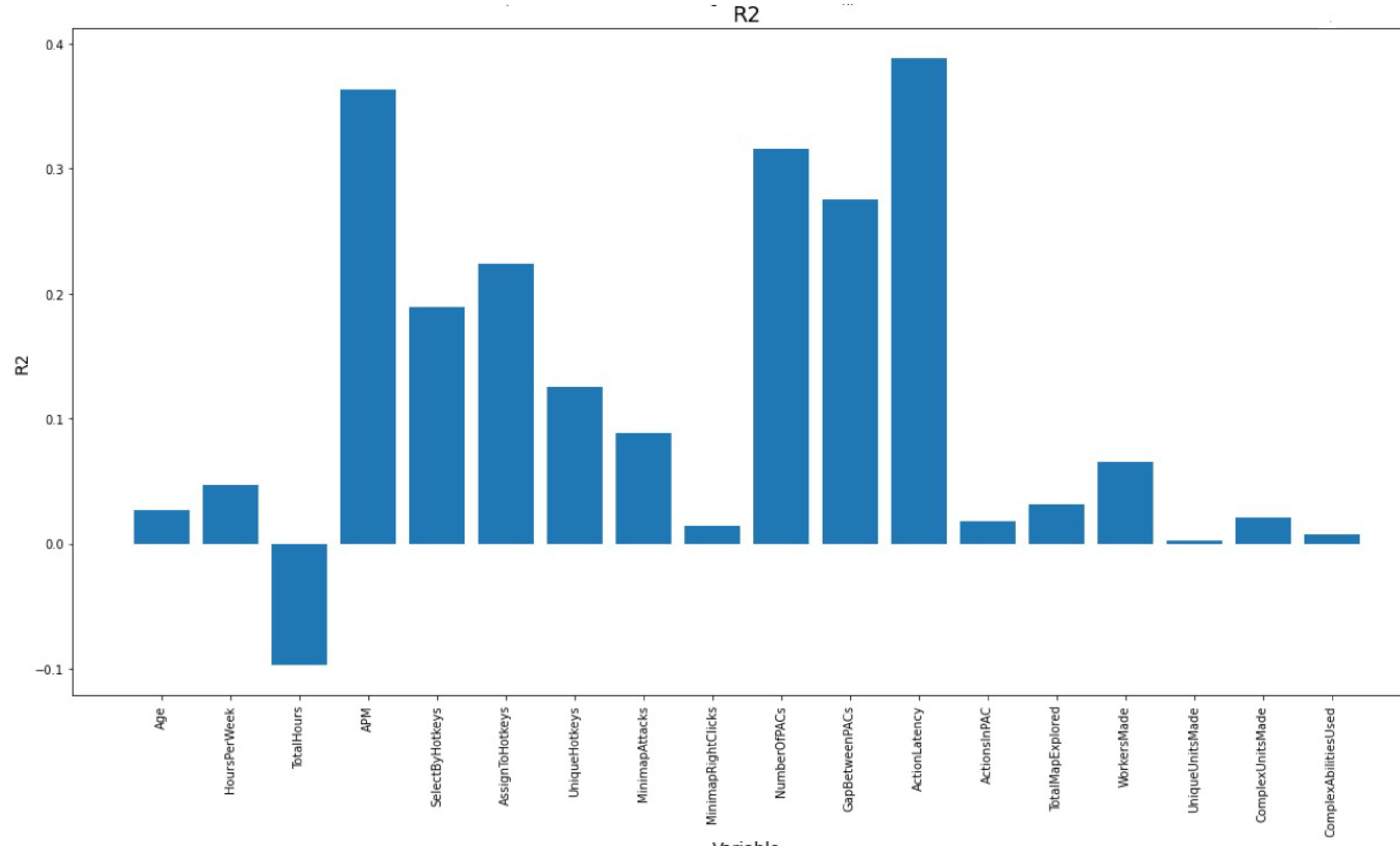
04

- Se separaron los datos en train y test donde el conjunto de entrenamiento quedo con 2493 muestras y el conjunto de pruebas con 624 muestras
- Luego se pasa a probar el desempeño individual de las variables del modelo, para esto se mide el error cuadrático medio (MSE) de cada una de las variables y su coeficiente de determinación ajustado (R<sup>2</sup>).
- El MSE sea bajo, puesto que esto nos indicara que el error o la diferencia que se logra obtener entre los puntos y la regresión planteada es mínima
- También se busca tener el R<sup>2</sup> alto, esto ya que el tener esta variable con un alto valor indica que la variable esta fuertemente relacionada con la salida del modelo
- También se pueden usar métodos visuales que van a permitir observar de una manera mas clara como es la relación entre las variables y la salida del modelo. Para lograr ese objetivo se gráfica el modelo lineal que se obtiene al usar solo una variable del conjunto de datos.

# MSE



# R2



# 05

# Modelos

Para modelar el comportamiento de los datos y poder obtener predicciones se usa un modelo de regresión lineal. La regresión Lineal busca una aproximación con coeficientes  $\text{Beta} = [ \text{Beta}_1, \dots, \text{Beta}_n ]$  que minimizan el error cuadrático medio entre los datos observados y la predicción modelada por una aproximación lineal.

1

## Selección de variables:

Dado que se construye una función para poder probar pipes diferentes las variables que le entran al pipe se dan por parámetro como una lista, en el pipe final quedarán fijas las variables seleccionadas.

2

## Limpieza de valores nulos:

Luego de seleccionar las variables es necesario asegurar de suministrar al modelo una buena calidad de datos por esto se eliminan las filas con valores nulos.

3

## Transformación de las variables:

Al igual que la selección de variables existe un parámetro que modifica el pipe de acuerdo al modelo que se esté probando. En nuestro caso se puede decidir usar o no una transformación sobre los datos, la transformación que se realiza sobre cada variable es sacar la raíz cuadrada para cada dato, esto ya que puede llevar a mejores resultados en el modelo mejorando la linealidad.

4

## Estandarización de los datos:

Al tener los datos debidamente preprocesado lo único restante antes de construir el modelo es estandarizar los datos. La estandarización se usa para escalar todas las variables a un rango similar.

5

## Modelo

Como último paso se construye el modelo de regresión lineal

# Modelo 1

---

En el caso del primero modelo de regresión lineal que se implemento se utilizaron las 6 variables con el mejor MSE, luego se aplico el pipe construido previamente sin ningún tipo de transformación y finalmente se calculo el R2 y MSE del modelo como métricas que nos ayudan a medir que tan bueno es el modelo. Las variables usadas son ActionLatency, APM, NumberOfPACs, GapBetweenPACs, AssignToHotkeys, SelectByHotkeys

## Resultados

R2: 0.46123536744407945

ECM: 1.8786678699559214

# Modelo 2

---

Para el modelo 2 se tomaron las 8 variables con mayor coeficiente de determinación ajustado (R2), y se uso el pipeline sin ninguna transformación. Las variables seleccionadas fueron las siguientes  
ActionLatency, APM, NumberOfPACs,  
GapBetweenPACs, AssignToHotkeys,  
SelectByHotkeys, UniqueHotkeys, MinimapAttacks

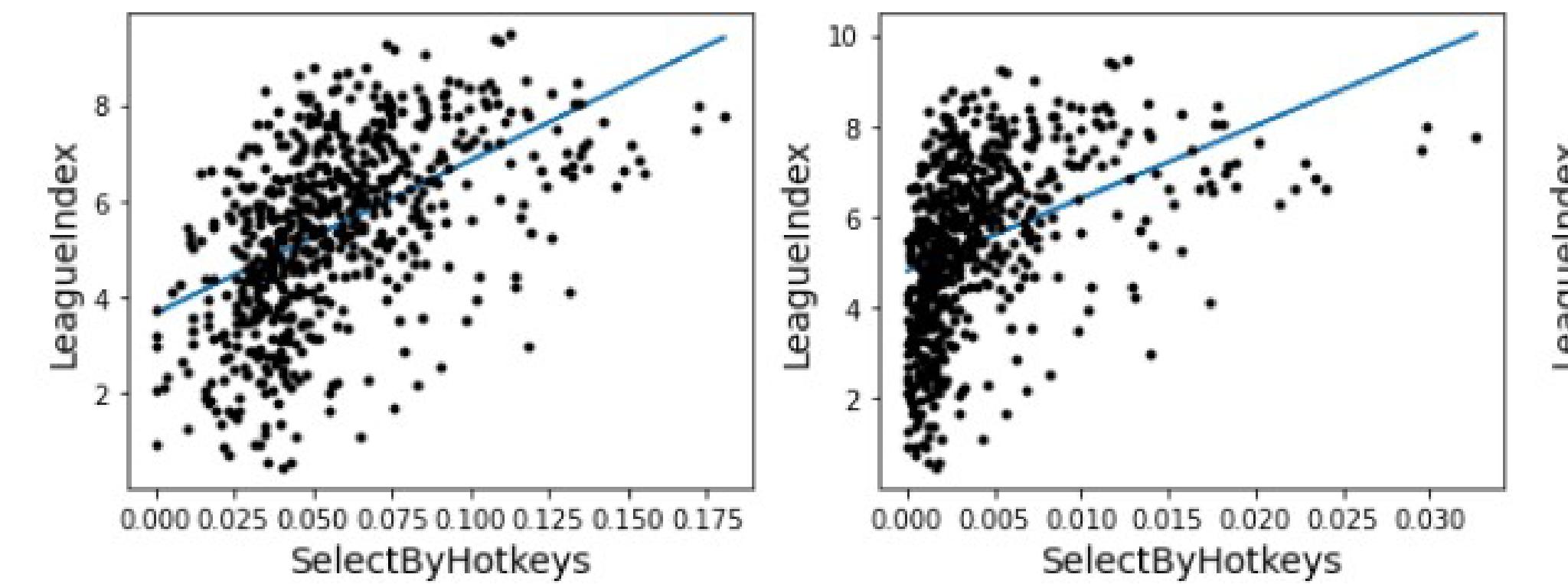
## Resultados

R2: 0.4814700409277356

ECM: 1.8081097289130452

# Modelo 3

Finalmente Para el modelo 3 se realizan transformaciones sobre los datos de entrada para tratar de mejorar la relación lineal entre las variables de entrada y la de salida. La transformación realizada fue la raíz cuadrada sobre las siguientes variables ActionLatency, APM, NumberOfPACs, GapBetweenPACs, AssignToHotKeys, SelectByHotKeys, MiniMapAttacks



El resultado de dicha transformación se ve ejemplificado con la variable AssignToHotkeys en la figura \ref{fig:transformacion} donde la figura de la izquierda es la transformación y la de la derecha la original

## Resultados

R2: 0.5280549676348534

ECM: 1.6456684702622562

# Análisis de resultados y recomendación

06

- El modelo final corresponde al tercer modelo en el que se incluyen las transformaciones por lo tanto se va a guardar el pipe obtenido de este modelo y adicionalmente se revisan los coeficientes de este modelo.
- De los modelos utilizados podemos concluir que las variables mas influyentes son APM, NumberOfPACs, GapBetweenPACs, AssignToHotkeys, SelectByHotkeys. Sin embargo, no se recomienda el uso en el sistema de liga del juego ya que los resultados de regresión no dan muy buenos resultados y las variables no tienen una buena relación lineal con el LeagueIndex. Se recomienda cambiar a una tarea de clasificación para decir si contratar o no a un jugador, y también usar otras variables que pueden ser mas dicientes como average\\_unspent\\_resources, EPM(effective actions per minute), SPM(Screens per minute), Spending Quotient Avg, My Supply Block Time Avg, Unspent Resources Avg. Aunque en muchas situaciones estas variables son útiles es importante tener en

# Coeficientes de las variables

---

Variable	Coeficiente
ActionLatency	-0.569885
APM	0.133822
NumberOfPACs	0.209536
GapBetweenPACs	-0.212991
AssignToHotkeys	0.240208
SelectByHotkeys	0.206023
MinimapAttacks	0.279875

# GRACIAS