Minería de datos: PRA1 - Selección y preparación de un juego de datos

Autor: Gabriel Patricio Bonilla Sanchez

Diciembre 2020

Contents

Introducción

Presentación

Esta práctica cubre de forma transversal la asignatura.

Las Prácticas 1 y 2 de la asignatura se plantean de una forma conjunta de modo que la Práctica 2 será continuación de la 1.

El objetivo global de las dos prácticas consiste en seleccionar uno o varios juegos de datos, realizar las tareas de **preparación y análisis exploratorio** con el objetivo de disponer de datos listos para **aplicar algoritmos** de clustering, asociación y clasificación.

Competencias

Las competencias que se trabajan en esta prueba son:

- Uso y aplicación de las TIC en el ámbito académico y profesional.
- Capacidad para innovar y generar nuevas ideas.
- Capacidad para evaluar soluciones tecnológicas y elaborar propuestas de proyectos teniendo en cuenta los recursos, las alternativas disponibles y las condiciones de mercado.
- Conocer las tecnologías de comunicaciones actuales y emergentes así como saberlas aplicar convenientemente para diseñar y desarrollar soluciones basadas en sistemas y tecnologías de la información.
- Aplicación de las técnicas específicas de ingeniería del software en las diferentes etapas del ciclo de vida de un proyecto.
- Capacidad para aplicar las técnicas específicas de tratamiento, almacenamiento y administración de datos.
- Capacidad para proponer y evaluar diferentes alternativas tecnológicas para resolver un problema concreto.

Objetivos

La correcta asimilación de todos los aspectos trabajados durante el semestre.
En esta práctica abordamos un caso real de minería de datos donde tenemos que poner en juego todos los conceptos trabajados. Hay que trabajar todo el ciclo de vida del proyecto. Desde el objetivo del proyecto hasta la implementación del conocimiento encontrado pasando por la preparación, limpieza de los datos, conocimiento de los datos, generación del modelo, interpretación y evaluación.

Descripción de la PRA a realizar

Recursos Básicos

Material docente proporcionado por la UOC.

Criterios de valoración

Ejercicios prácticos

Para todas las PRA es **necesario documentar** en cada apartado del ejercicio práctico que se ha hecho y como se ha hecho.

Formato y fecha de entrega PRA_1

El formato de entrega es: usernameestudiant-PRAn.html/doc/docx/odt/pdf
Fecha de entrega: 02/12/2020
Se debe entregar la PRA_1 en el buzón de entregas del aula

Nota: Propiedad intelectual

> A menudo es inevitable, al producir una obra multimedia, hacer uso de recursos creados por terceras personas. Es por lo tanto comprensible hacerlo en el marco de una práctica de los estudios de Informática, Multimedia y Telecomunicación de la UOC, siempre y cuando esto se documente claramente y no suponga plagio en la práctica.
>
> Por lo tanto, al presentar una práctica que haga uso de recursos ajenos, se debe presentar junto con ella un documento en que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y su estatus legal: si la obra esta protegida por el copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL ...). El estudiante deberá asegurarse de que la licencia no impide específicamente su uso en el marco de la práctica. En caso de no encontrar la información correspondiente tendrá que asumir que la obra esta protegida por copyright.
>
> Deberéis, además, adjuntar los ficheros originales cuando las obras utilizadas sean digitales, y su código fuente si corresponde.

Enunciado

Todo estudio analítico debe nacer de una necesidad por parte del **negocio** o de una voluntad de dotarle de un conocimiento contenido en los datos y que solo podremos obtener a través de una colección de buenas prácticas basadas en la Minería de Datos.

El mundo de la analítica de datos se sustenta en 3 ejes:

1. Uno de ellos es el profundo **conocimiento** que deberíamos tener **del negocio** al que tratamos de dar respuestas mediante los estudios analíticos.
2. El otro gran eje es sin duda las **capacidades analíticas** que seamos capaces de desplegar y en este sentido, las dos prácticas de esta asignatura pretenden que el estudiante realice un recorrido sólido por este segundo eje.
3. El tercer eje son los **Datos**. Las necesidades del Negocio deben concretarse con preguntas analíticas que a su vez sean viables responder a partir de los datos de que disponemos. La tarea de analizar los datos es sin duda importante, pero la tarea de identificarlos y obtenerlos va a ser para un analista un reto permanente.

Como **primera parte** del estudio analítico que nos disponemos a realizar, se pide al estudiante que complete los siguientes pasos:

25% Justificación de la elección del juego de datos donde se detalle el potencial analítico que se intuye

1. Seleccionar un juego de datos y justificar su elección. El juego de datos deberá tener capacidades para que se le puedan aplicar algoritmos supervisados, algoritmos no supervisados y reglas de asociación.

El dataset seleccionado ha sido obtenido desde el siguiente enlace: https://www.kaggle.com/aitzaz/stack-overflow-developer-survey-2020. Este juego de datos contiene los resultados de la Encuesta Anual a Desarrolladores StackOverflow 2020. Se obtuvo alrededor de 65000 participaciones de programadores y desarrolladores de 180 paises. La encuesta aborda varios ámbitos, tanto a nivel de experiencia, formación académica y skills (habilidades técnicas) en diferentes tecnologías que el encuestado ha ido adquiriendo a lo largo del tiempo.

Esta encuesta anual ha recolectado datos sobre 61 variables que se pasan a detallar a continuación:

- *Respondent:* número de identificación del encuestado aleatorizado (no en orden de tiempo de respuesta de la encuesta)
- *MainBranch:* ¿Cuál de las siguientes opciones te describe mejor hoy?
- *Hobbyist:* ¿Desarrollas como pasatiempo?
- *Age:* ¿Cuál es su edad (en años)?
- *Age1stCode:* ¿A qué edad escribiste tu primera línea de código o programa?
- *CompFreq:* ¿Esa compensación es semanal, mensual o anual?
- *CompTotal:* ¿Cuál es su compensación total actual (salario, bonificaciones y beneficios, antes de impuestos y deducciones), en "CurrencySymbol"?. Número entero.
- *ConvertedComp:* Salario anual en USD, utilizando el tipo de cambio del 19 de febrero de 2020, asumiendo 12 meses laborales y 50 semanas laborales".
- *Country:* País dónde vive.
- *CurrencyDesc:* ¿Qué moneda utiliza a diario? Descripción.
- *CurrencySymbol:* ¿Qué moneda usa a diario? Forma abreviada.
- *DatabaseDesireNextYear:* ¿En qué entornos de base de datos desea trabajar durante el próximo año?
- *DatabaseWorkedWith:* ¿En qué entornos de base de datos ha realizado un trabajo de desarrollo extenso durante el año pasado?
- *DevType:* ¿Cuál de los siguientes lo describe?
- *EdLevel:* ¿Cuál de las siguientes opciones describe mejor el nivel más alto de educación formal que ha completado?
- *Employment:* ¿cuál de las siguientes opciones describe mejor su situación laboral actual?
- *Ethnicity:* ¿Cuál de los siguientes grupos étnicos lo describe?
- *Gender:* ¿Cuál de las siguientes opciones de sexo lo describe?
- *JobFactors:* Para el caso de decidiendo entre dos ofertas de trabajo con la misma compensación, beneficios y ubicación. ¿Qué factores son los más importantes para usted?
- *JobSat:* ¿Qué tan satisfecho está con su trabajo actual?
- *JobSeek:* ¿Cuál de las siguientes opciones describe mejor su estado actual de búsqueda de empleo?
- *LanguageDesireNextYear:* "¿En qué lenguajes de programación, scripting y marcado desea trabajar durante el próximo año?.
- *LanguageWorkedWith:* ¿En qué lenguajes de programación, scripting y marcado ha realizado un trabajo de desarrollo extenso durante el año pasado?.
- *MiscTechDesireNextYear:* ¿En qué otros frameworks, bibliotecas y herramientas desea trabajar durante el próximo año?.
- *MiscTechWorkedWith:* ¿En qué otros frameworks, bibliotecas y herramientas ha realizado un trabajo de desarrollo extenso durante el año pasado?.
- *NEWCollabToolsDesireNextYear:* ¿En qué herramientas de colaboración desea trabajar durante el próximo año?
- *NEWCollabToolsWorkedWith:* ¿En qué herramientas de colaboración ha realizado un trabajo de desarrollo extenso durante el año pasado?
- *NEWDevOps:* ¿Su empresa tiene una persona dedicada a DevOps?
- *NEWDevOpsImpt:* ¿Qué importancia tiene la práctica de DevOps para escalar el desarrollo de software?

- *NEWEdImpt:* ¿Qué importancia tiene una educación formal, como un título universitario en ciencias de la computación, para su carrera?
- *NEWJobHunt:* En general, ¿Cuáles son las motivaciones que lo impulsan a buscar un nuevo trabajo?.
- *NEWJobHuntResearch:* Cuando busca trabajo, ¿cómo puede obtener más información sobre una empresa?
- *NEWLearn:* ¿Con qué frecuencia aprende un nuevo lenguaje o marco?
- *NEWOffTopic:* ¿Crees que Stack Overflow debería relajar las restricciones sobre lo que se considera "fuera de tema"?
- *NEWOnboardGood:* ¿Cree que su empresa tiene un buen proceso de incorporación? (Por incorporación, nos referimos al proceso estructurado para que se adapte a su nuevo puesto en una empresa)
- *NEWOtherComms:* ¿Es miembro de alguna otra comunidad de desarrolladores en línea?
- *NEWOvertime:* ¿Con qué frecuencia trabaja horas extraordinarias o más allá de las expectativas formales de su trabajo?
- *NEWPurchaseResearch:* Al comprar una nueva herramienta o software, ¿cómo descubre e investiga las soluciones disponibles?
- *NEWPurpleLink:* Busca una solución de codificación en línea y el primer enlace de resultado es violeta porque ya lo visitó. ¿Cómo se siente?
- *NEWSOSites:* ¿Cuál de los siguientes sitios de Stack Overflow ha visitado?
- *NEWStuck:* ¿Qué hace cuando se queda atascado en un problema?
- *OpSys:* ¿Cuál es el sistema operativo principal en el que trabaja?
- *OrgSize:* Aproximadamente, ¿cuántas personas emplea la empresa u organización para la que trabaja actualmente?
- *PlatformDesireNextYear:* ¿En qué plataformas desea trabajar durante el próximo año?
- *PlatformWorkedWith:* ¿En qué plataformas ha realizado un trabajo de desarrollo extenso durante el año pasado?
- *PurchaseWhat:* ¿Qué nivel de influencia tiene usted, personalmente, sobre las compras de nueva tecnología en su organización?
- *Sexuality:* ¿Cuál de los siguientes lo describe a usted sobre su sexualidad?.
- *SOAccount:* ¿Tiene una cuenta de Stack Overflow?
- *SOComm:* ¿Te consideras miembro de la comunidad de Stack Overflow?
- *SOPartFreq:* ¿Con qué frecuencia diría que participa en preguntas y respuestas en Stack Overflow? Por participar nos referimos a preguntar, responder, votar o comentar preguntas.
- *SOVisitFreq:* ¿Con qué frecuencia visita Stack Overflow?
- *SurveyEase:* ¿Qué tan fácil o difícil fue completar esta encuesta?
- *SurveyLength:* ¿Qué opina de la duración de la encuesta este año?
- *Trans:* ¿Eres transgénero?
- *UndergradMajor:* ¿Cuál fue su campo de estudio principal?
- *WebframeDesireNextYear:* ¿En qué frameworks web desea trabajar durante el próximo año?
- *WebframeWorkedWith:* ¿En qué frameworks web ha realizado un extenso trabajo de desarrollo durante el año pasado?
- *WelcomeChange:* En comparación con el año pasado, ¿qué tan bienvenido se siente en Stack Overflow?
- *WorkWeekHrs:* En promedio, ¿cuántas horas por semana trabaja?
- *YearsCode:* Incluyendo cualquier educación, ¿cuántos años ha estado programando en total?
- *YearsCodePro:* NO incluye educación, ¿cuántos años ha programado profesionalmente (como parte de su trabajo)?

Las capacidades análiticas del dataset, que se tomaron en cuenta para elegirlo son:

- Cuenta con una cantidad suficientes variables, tanto numéricas, categóricas. Las variables categóricas también pueden volverse a convertir a variables numéricas. Esto permitiría aplicar algoritmos supervisados y no supervisados, donde se puede clasificar a los programadores o desarrolladores según la experticia actual.
- También permite agregar nuevas variables númericas que representen el número de tecnologías que domina cada encuestado.

- Al incluir las tecnologías usadas por desarrolladores en: base de datos, lenguages de programación, frameworks y demás herramientas, permite tener una gran cantidad de preferencias de las que se puede extraer reglas de asociación interesantes sobre las tecnologías más usadas entre los distintos tipos de desarrolladores.
- Cuenta con variables que pueden discretizarse y otras donde se puede aplicar tareas de limpieza y preparación previa antes de aplicar los distintos métodos.

Sin embargo, para efectos del análisis, del dataset original, se excluirán las siguientes variables:

1. Respondent
2. MainBranch
3. Hobbyist
4. Age1stCode
5. CompFreq
6. CompTotal (+)
7. CurrencyDesc
8. CurrencySymbol
9. DatabaseDesireNextYear
10. Ethnicity
11. JobFactors
12. JobSat
13. JobSeek
14. LanguageDesireNextYear
15. MiscTechDesireNextYear
16. NEWCollabToolsDesireNextYear
17. NEWDevOps
18. NEWDevOpsImpt
19. NEWEdImpt
20. NEWJobHunt
21. NEWJobHuntResearch
22. NEWLearn
23. NEWOffTopic
24. NEWOnboardGood
25. NEWOtherComms
26. NEWOvertime
27. NEWPurchaseResearch
28. NEWPurpleLink
29. NEWSOSites
30. NEWStuck
31. PlatformDesireNextYear
32. PurchaseWhat
33. Sexuality
34. SOComm
35. SOVisitFreq (+)
36. SurveyEase
37. SurveyLength
38. Trans
39. UndergradMajor (+)
40. WebframeDesireNextYear
41. WelcomeChange
42. YearsCodePro (+)

Muchos de estos campos no son relevantes para el alcance de la Práctica #1 y #2; otros reflejan deseos de

los programadores respecto a tecnologías, para lo cual solo tomaremos los datos que reflejan la experiencia actual del programador.

Los campos marcados con (+) se los ha excluído, ya que se existe otra variable similar, que en caso de mantenerse significaría agregar información redundante al dataset.

Con la finalidad de disminuir el número de observaciones o individuos, vamos a limitar el estudio de este dataset a mi país natal, *Ecuador*. Con esto no incluiremos la variable **Country**.

En conclusión, vamos a trabajar a con 18 variables propias del dataset original, de las cuales 4 son númericas (Age, ConvertedComp, WorkWeekHrs y YearsCode). También tenemos variables no numéricas, las cuales vamos a realizar un análisis más detallado posteriormente, generando variables númericas a partir de ellas, las cuales son:

- DatabaseWorkedWith
- LanguageWorkedWith
- MiscTechWorkedWith
- NEWCollabToolsWorkedWith
- PlatformWorkedWith
- WebframeWorkedWith

Estas nuevas variables numéricas a generarse posteriormente servirán principalmente cuando se intente aplicar algoritmos no supervisados, como K-Means, y también serán usadas para crear un nuevo dataset sobre el que se aplicará el algoritmo SVD o PCA.

Este nuevo dataset de variables numéricas, ayudará a dar respuesta a las siguientes preguntas:

> ¿Hay relación directa entre el número de tecnologías que domina el programador y su sueldo anual, en Ecuador? ¿Hay relación directa entre el número de años de experiencia que tiene el programador y el número de tecnologías que domina, en Ecuador? ¿En Ecuador, influye el número de años de experiencia del programador con el número de tecnologías que domina o conoce? ¿Qué relación hay entre el número de años de experiencia del programador y el sueldo que percibe anualmente en el mercado ecuatoriano? ¿Como afecta el número de horas trabajadas a la semana sobre el sueldo que percibe anualmente el programador ecuatoriano? ¿Hay relación directa entre el número de años de experiencia que tiene el programador y la edad del mismo en Ecuador?

25% Información extraída del análisis exploratorio

2. Realizar un análisis exploratorio del juego de datos seleccionado.

- Cargar el dataset

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
library(car)

options('max.print' = 100000)   # or whatever value you want

# Cargamos el fichero de datos_original
datos_original <- read.csv('survey_results_public.csv', sep=",", encoding = "UTF-8")
filas_original=dim(datos_original)[1]

# Verificamos la estructura del conjunto de datos_original
str(datos_original)
```

```
## 'data.frame':    64461 obs. of  61 variables:
##  $ Respondent                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MainBranch                   : chr  "I am a developer by profession" "I am a developer by professio
##  $ Hobbyist                     : chr  "Yes" "No" "Yes" "Yes" ...
##  $ Age                          : num  NA NA NA 25 31 NA NA 36 30 22 ...
##  $ Age1stCode                   : chr  "13" "19" "15" "18" ...
##  $ CompFreq                     : chr  "Monthly" NA NA NA ...
##  $ CompTotal                    : num  NA NA NA NA NA NA NA 116000 NA 25000 ...
##  $ ConvertedComp                : num  NA NA NA NA NA ...
##  $ Country                      : chr  "Germany" "United Kingdom" "Russian Federation" "Albania" ...
##  $ CurrencyDesc                 : chr  "European Euro" "Pound sterling" NA "Albanian lek" ...
##  $ CurrencySymbol               : chr  "EUR" "GBP" NA "ALL" ...
##  $ DatabaseDesireNextYear       : chr  "Microsoft SQL Server" NA NA NA ...
##  $ DatabaseWorkedWith           : chr  "Elasticsearch;Microsoft SQL Server;Oracle" NA NA NA ...
##  $ DevType                      : chr  "Developer, desktop or enterprise applications;Developer, full-
##  $ EdLevel                      : chr  "Master's degree (M.A., M.S., M.Eng., MBA, etc.)" "Bachelor's d
##  $ Employment                   : chr  "Independent contractor, freelancer, or self-employed" "Employe
##  $ Ethnicity                    : chr  "White or of European descent" NA NA "White or of European desc
##  $ Gender                       : chr  "Man" NA NA "Man" ...
##  $ JobFactors                   : chr  "Languages, frameworks, and other technologies I'd be working w
##  $ JobSat                       : chr  "Slightly satisfied" "Very dissatisfied" NA "Slightly dissatisi
##  $ JobSeek                      : chr  "I am not interested in new job opportunities" "I am not interc
##  $ LanguageDesireNextYear       : chr  "C#;HTML/CSS;JavaScript" "Python;Swift" "Objective-C;Python;Swi
##  $ LanguageWorkedWith           : chr  "C#;HTML/CSS;JavaScript" "JavaScript;Swift" "Objective-C;Python
##  $ MiscTechDesireNextYear       : chr  ".NET Core;Xamarin" "React Native;TensorFlow;Unity 3D" NA NA .
##  $ MiscTechWorkedWith           : chr  ".NET;.NET Core" "React Native" NA NA ...
##  $ NEWCollabToolsDesireNextYear: chr  "Microsoft Teams;Microsoft Azure;Trello" "Github;Slack" NA NA
##  $ NEWCollabToolsWorkedWith     : chr  "Confluence;Jira;Slack;Microsoft Azure;Trello" "Confluence;Jira
##  $ NEWDevOps                    : chr  "No" NA NA "No" ...
##  $ NEWDevOpsImpt                : chr  "Somewhat important" NA NA NA ...
##  $ NEWEdImpt                    : chr  "Fairly important" "Fairly important" NA "Not at all important/
##  $ NEWJobHunt                   : chr  NA NA NA "Curious about other opportunities;Wanting to work wi
##  $ NEWJobHuntResearch           : chr  NA NA NA NA ...
##  $ NEWLearn                     : chr  "Once a year" "Once a year" "Once a decade" "Once a year" ...
##  $ NEWOffTopic                  : chr  "Not sure" "Not sure" NA "Not sure" ...
##  $ NEWOnboardGood               : chr  NA NA NA "Yes" ...
##  $ NEWOtherComms                : chr  "No" "No" "No" "Yes" ...
##  $ NEWOvertime                  : chr  "Often: 1-2 days per week or more" NA NA "Occasionally: 1-2 day
##  $ NEWPurchaseResearch          : chr  "Start a free trial;Ask developers I know/work with" NA NA NA
##  $ NEWPurpleLink                : chr  "Amused" "Amused" NA NA ...
##  $ NEWSOSites                   : chr  "Stack Overflow (public Q&A for anyone who codes)" "Stack Overi
##  $ NEWStuck                     : chr  "Visit Stack Overflow;Go for a walk or other physical activity
##  $ OpSys                        : chr  "Windows" "MacOS" "Linux-based" "Linux-based" ...
##  $ OrgSize                      : chr  "2 to 9 employees" "1,000 to 4,999 employees" NA "20 to 99 empl
##  $ PlatformDesireNextYear       : chr  "Android;iOS;Kubernetes;Microsoft Azure;Windows" "iOS;Kubernete
##  $ PlatformWorkedWith           : chr  "Windows" "iOS" NA NA ...
##  $ PurchaseWhat                 : chr  NA "I have little or no influence" NA "I have a great deal of i
##  $ Sexuality                    : chr  "Straight / Heterosexual" NA NA "Straight / Heterosexual" ...
##  $ SOAccount                    : chr  "No" "Yes" "Yes" "Yes" ...
##  $ SOComm                       : chr  "No, not at all" "Yes, definitely" "Yes, somewhat" "Yes, defini
##  $ SOPartFreq                   : chr  NA "Less than once per month or monthly" "A few times per month
##  $ SOVisitFreq                  : chr  "Multiple times per day" "Multiple times per day" "Daily or alm
##  $ SurveyEase                   : chr  "Neither easy nor difficult" NA "Neither easy nor difficult" NA
##  $ SurveyLength                 : chr  "Appropriate in length" NA "Appropriate in length" NA ...
```

```
##  $ Trans                       : chr  "No" NA NA "No" ...
##  $ UndergradMajor              : chr  "Computer science, computer engineering, or software engineeri
##  $ WebframeDesireNextYear      : chr  "ASP.NET Core" NA NA NA ...
##  $ WebframeWorkedWith          : chr  "ASP.NET;ASP.NET Core" NA NA NA ...
##  $ WelcomeChange               : chr  "Just as welcome now as I felt last year" "Somewhat more welco
##  $ WorkWeekHrs                 : num  50 NA NA 40 NA NA NA 39 50 36 ...
##  $ YearsCode                   : chr  "36" "7" "4" "7" ...
##  $ YearsCodePro                : chr  "27" "4" NA "4" ...
```

```r
#Resumen del dataset original
summary(datos_original)
```

```
##    Respondent     MainBranch          Hobbyist              Age
##  Min.   :    1   Length:64461       Length:64461       Min.   :  1.00
##  1st Qu.:16116   Class :character   Class :character   1st Qu.: 24.00
##  Median :32231   Mode  :character   Mode  :character   Median : 29.00
##  Mean   :32554                                         Mean   : 30.83
##  3rd Qu.:49142                                         3rd Qu.: 35.00
##  Max.   :65639                                         Max.   :279.00
##                                                        NA's   :19015
##   Age1stCode          CompFreq           CompTotal          ConvertedComp
##  Length:64461       Length:64461       Min.   : 0.000e+00   Min.   :      0
##  Class :character   Class :character   1st Qu.: 2.000e+04   1st Qu.:  24648
##  Mode  :character   Mode  :character   Median : 6.300e+04   Median :  54049
##                                        Mean   :3.190e+242   Mean   : 103756
##                                        3rd Qu.: 1.250e+05   3rd Qu.:  95000
##                                        Max.   :1.111e+247   Max.   :2000000
##                                        NA's   :29635        NA's   :29705
##    Country          CurrencyDesc       CurrencySymbol
##  Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  DatabaseDesireNextYear DatabaseWorkedWith   DevType
##  Length:64461           Length:64461       Length:64461
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##    EdLevel           Employment         Ethnicity            Gender
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   JobFactors           JobSat            JobSeek
##  Length:64461       Length:64461       Length:64461
```

```
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  LanguageDesireNextYear LanguageWorkedWith MiscTechDesireNextYear
##  Length:64461           Length:64461       Length:64461
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##  MiscTechWorkedWith NEWCollabToolsDesireNextYear NEWCollabToolsWorkedWith
##  Length:64461       Length:64461                 Length:64461
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##   NEWDevOps          NEWDevOpsImpt       NEWEdImpt          NEWJobHunt
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  NEWJobHuntResearch   NEWLearn         NEWOffTopic        NEWOnboardGood
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  NEWOtherComms      NEWOvertime        NEWPurchaseResearch NEWPurpleLink
##  Length:64461       Length:64461       Length:64461        Length:64461
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##
##
##
##
##   NEWSOSites           NEWStuck            OpSys              OrgSize
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
##  PlatformDesireNextYear PlatformWorkedWith PurchaseWhat
##  Length:64461           Length:64461       Length:64461
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##
##
##
##
##   Sexuality           SOAccount            SOComm           SOPartFreq
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  SOVisitFreq         SurveyEase         SurveyLength          Trans
##  Length:64461       Length:64461       Length:64461       Length:64461
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  UndergradMajor     WebframeDesireNextYear WebframeWorkedWith
##  Length:64461       Length:64461           Length:64461
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##  WelcomeChange       WorkWeekHrs      YearsCode         YearsCodePro
##  Length:64461       Min.   :  1.00   Length:64461       Length:64461
##  Class :character   1st Qu.: 40.00   Class :character   Class :character
##  Mode  :character   Median : 40.00   Mode  :character   Mode  :character
##                     Mean   : 40.78
##                     3rd Qu.: 44.00
##                     Max.   :475.00
##                     NA's   :23310
```

Antes de proceder a hacer el análisis exploratio, Vamos a proceder a quitar las variables detalladas previamente, para ello creamos un nuevo juego de datos resumido únicamente con las columnas detalladas a continuación

```
# Creamos un juego de datos resumido
datos <- datos_original[, c(4, 8:9, 13:16, 18, 23, 25, 27, 42:43, 45, 48, 50, 57, 59:60)]
filas=dim(datos)[1]

# Filtramos solo los registros que corresponden con Ecuador
datosEcuador <- datos[datos$Country %in% c("Ecuador"), ]
filasEcuador=dim(datosEcuador)[1]

# Anulamos la variable Country del dataset datosEcuador
```

```r
datosEcuador$Country = NULL

# Verificamos la estructura del conjunto de datos
str(datosEcuador)
```

```
## 'data.frame':    49 obs. of  18 variables:
##  $ Age                     : num  35 31 24 30 36 28 32 30 47 32 ...
##  $ ConvertedComp           : num  48000 NA 4200 12000 38400 9600 42000 NA 26400 21600 ...
##  $ DatabaseWorkedWith      : chr  "Elasticsearch;MariaDB;MongoDB;MySQL;Redis" "Microsoft SQL Server"
##  $ DevType                 : chr  "Developer, back-end;Developer, desktop or enterprise applications
##  $ EdLevel                 : chr  "Professional degree (JD, MD, etc.)" "Bachelor's degree (B.A., B.S
##  $ Employment              : chr  "Employed full-time" "Not employed, but looking for work" "Employe
##  $ Gender                  : chr  "Man" "Man" "Man" "Man" ...
##  $ LanguageWorkedWith      : chr  "Bash/Shell/PowerShell;HTML/CSS;JavaScript;PHP;Python;SQL;TypeScri
##  $ MiscTechWorkedWith      : chr  "Node.js" ".NET" "Flutter;Unity 3D" NA ...
##  $ NEWCollabToolsWorkedWith: chr  "Confluence;Jira;Github;Gitlab;Slack;Google Suite (Docs, Meet, etc
##  $ OpSys                   : chr  "Linux-based" "Windows" "Windows" "Windows" ...
##  $ OrgSize                 : chr  "2 to 9 employees" NA "2 to 9 employees" "1,000 to 4,999 employees
##  $ PlatformWorkedWith      : chr  "Android;AWS;Docker;Google Cloud Platform;Kubernetes;Linux;Microso
##  $ SOAccount               : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ SOPartFreq              : chr  "A few times per month or weekly" "A few times per week" "A few ti
##  $ WebframeWorkedWith      : chr  "Angular;Symfony" "Angular;Angular.js;ASP.NET" NA NA ...
##  $ WorkWeekHrs             : num  60 NA 20 40 20 60 40 NA 9 40 ...
##  $ YearsCode               : chr  "15" "6" "4" "9" ...
```

```r
# Resumen del conjunto de datos
summary(datosEcuador)
```

```
##       Age        ConvertedComp     DatabaseWorkedWith   DevType
##  Min.   :15.00   Min.   :   4200   Length:49          Length:49
##  1st Qu.:24.75   1st Qu.:  12300   Class :character   Class :character
##  Median :30.00   Median :  18102   Mode  :character   Mode  :character
##  Mean   :31.53   Mean   :  66289
##  3rd Qu.:35.25   3rd Qu.:  27300
##  Max.   :55.00   Max.   :1080000
##  NA's   :13      NA's   :23
##    EdLevel           Employment           Gender          LanguageWorkedWith
##  Length:49          Length:49          Length:49          Length:49
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  MiscTechWorkedWith NEWCollabToolsWorkedWith    OpSys
##  Length:49          Length:49                Length:49
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##
##    OrgSize          PlatformWorkedWith  SOAccount          SOPartFreq
```

```
##  Length:49          Length:49          Length:49          Length:49
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  WebframeWorkedWith  WorkWeekHrs     YearsCode
##  Length:49          Min.   : 8.00   Length:49
##  Class :character   1st Qu.:20.00   Class :character
##  Mode  :character   Median :40.00   Mode  :character
##                     Mean   :34.35
##                     3rd Qu.:40.00
##                     Max.   :60.00
##                     NA's   :18
```

En el resumen del dataset preliminar, vemos que en varios campos numéricos hay una gran cantidad de valores NA, los cuales no permitirán tener aplicar correctamente los algoritmos sin caer en un sesgo. Estos casos deben ser tratados con cuidado, ya que pueden haber muchas causas para no que no haya llenado algún campo, como en el caso de la variable *WorkWeekHrs*. Existen casos donde los programadores respondieron que trabajan a tiempo completo y no han llenado el número de horas, por lo que hay una irregularidad para esos casos, que deberán ser tratados. Se va a proceder a llenar estos campos con las medias de cada variable:

```
# Rellanamos los campos con NA con valores medios de cada variable
datosEcuador$Age[is.na(datosEcuador$Age)] <- mean(datosEcuador$Age,na.rm=T)

datosEcuador$ConvertedComp[is.na(datosEcuador$ConvertedComp)] <- mean(datosEcuador$ConvertedComp,na.rm=

datosEcuador$WorkWeekHrs[is.na(datosEcuador$WorkWeekHrs)] <- mean(datosEcuador$WorkWeekHrs,na.rm=T)
```

Posteriormente se va a realizar el tratamiento de valores faltantes para el campo **YearsCode**, ya que tiene valores categóricos, para lo cual primero se deberá convertir a valor numéricos los campos categóricos y luego llenar las observaciones con valores faltantes o no disponibles.

Ahora que ya tenemos un nuevo objeto con datos limpios de valores no disponibles, podemos ver los tipos de datos de cada columna, para poder determinar como debemos tratarlas y si necesitan o no una conversión de tipos, para lo cual vamos a usar 2 funciones: **glimpse** y **sapply**.

```
#Vemos el tipo de dato de las variables
glimpse(datosEcuador)
```

```
## Rows: 49
## Columns: 18
## $ Age                     <dbl> 35.00000, 31.00000, 24.00000, 30.00000, 36...
## $ ConvertedComp           <dbl> 48000.00, 66289.23, 4200.00, 12000.00, 384...
## $ DatabaseWorkedWith      <chr> "Elasticsearch;MariaDB;MongoDB;MySQL;Redis...
## $ DevType                 <chr> "Developer, back-end;Developer, desktop or...
## $ EdLevel                 <chr> "Professional degree (JD, MD, etc.)", "Bac...
## $ Employment              <chr> "Employed full-time", "Not employed, but l...
## $ Gender                  <chr> "Man", "Man", "Man", "Man", "Man", "Man", ...
## $ LanguageWorkedWith      <chr> "Bash/Shell/PowerShell;HTML/CSS;JavaScript...
## $ MiscTechWorkedWith      <chr> "Node.js", ".NET", "Flutter;Unity 3D", NA,...
## $ NEWCollabToolsWorkedWith <chr> "Confluence;Jira;Github;Gitlab;Slack;Googl...
```

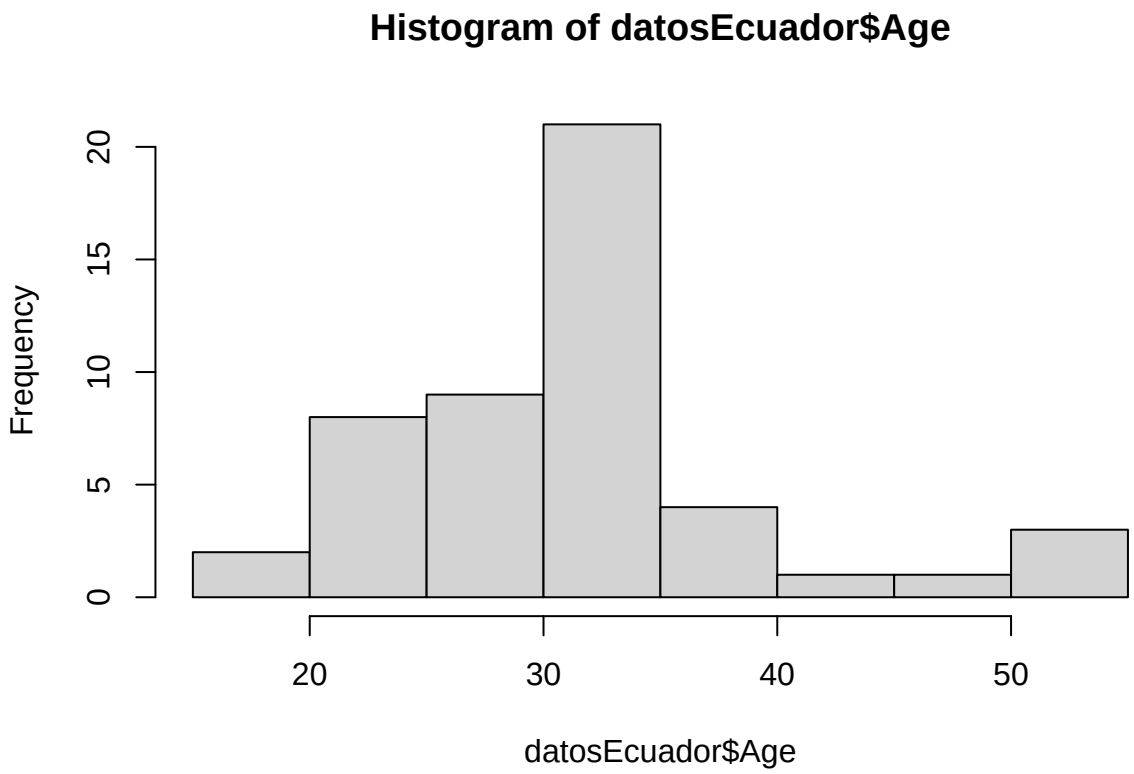
```
## $ OpSys                    <chr> "Linux-based", "Windows", "Windows", "Wind...
## $ OrgSize                  <chr> "2 to 9 employees", NA, "2 to 9 employees"...
## $ PlatformWorkedWith       <chr> "Android;AWS;Docker;Google Cloud Platform;...
## $ SOAccount                <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", ...
## $ SOPartFreq               <chr> "A few times per month or weekly", "A few ...
## $ WebframeWorkedWith       <chr> "Angular;Symfony", "Angular;Angular.js;ASP...
## $ WorkWeekHrs              <dbl> 60.00000, 34.35484, 20.00000, 40.00000, 20...
## $ YearsCode                <chr> "15", "6", "4", "9", "29", "8", "13", "7",...
```

```
# Otra forma de ver el tipo de dato de cada columna
sapply(datosEcuador, class)
```

```
##                       Age             ConvertedComp       DatabaseWorkedWith
##                 "numeric"                 "numeric"               "character"
##                   DevType                   EdLevel                Employment
##                "character"               "character"               "character"
##                    Gender       LanguageWorkedWith       MiscTechWorkedWith
##                "character"               "character"               "character"
## NEWCollabToolsWorkedWith                     OpSys                   OrgSize
##                "character"               "character"               "character"
##       PlatformWorkedWith                 SOAccount                SOPartFreq
##                "character"               "character"               "character"
##       WebframeWorkedWith                WorkWeekHrs                  YearsCode
##                "character"                 "numeric"               "character"
```

ANÁLISIS UNIVARIADO

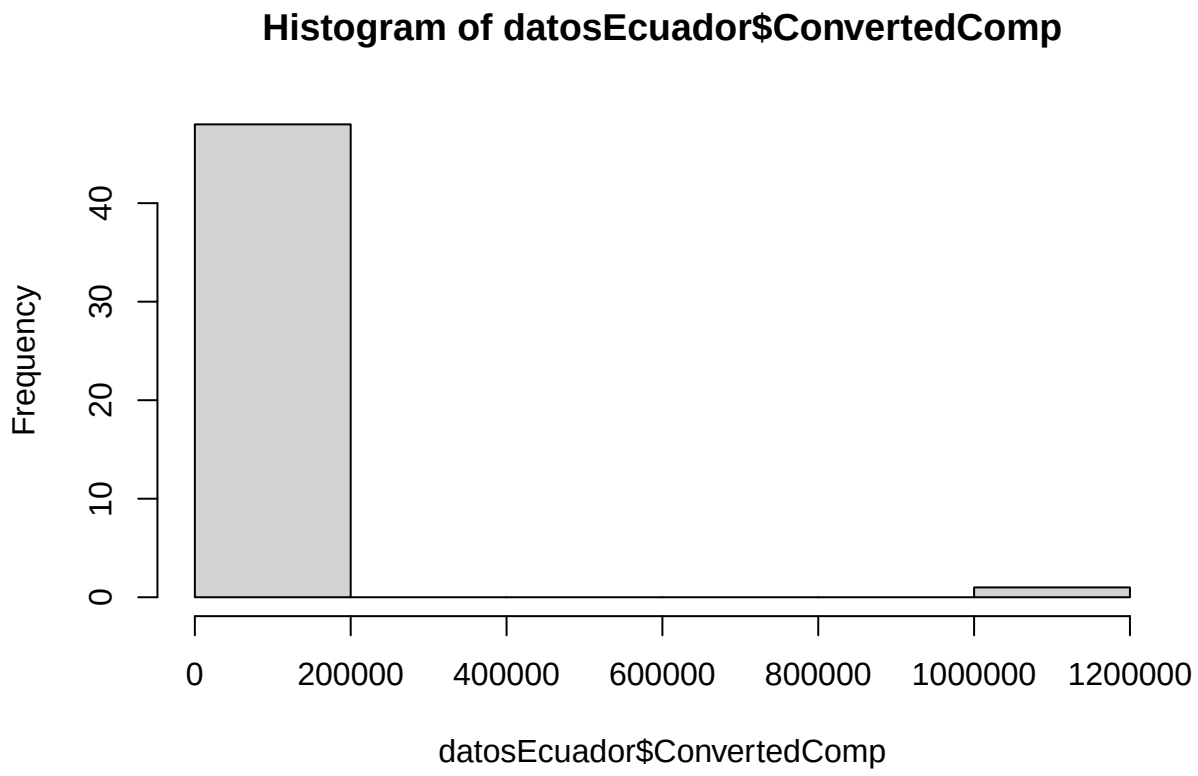
```
#Vemos la variable Age
hist(datosEcuador$Age)
```



Vemos que la edad entre los desarrolladores encuestados oscila entre un rango de edad entre 20 - 40 mayoritariamente.

Ahora vamos a analizar el histograma para la variable Salario Anual

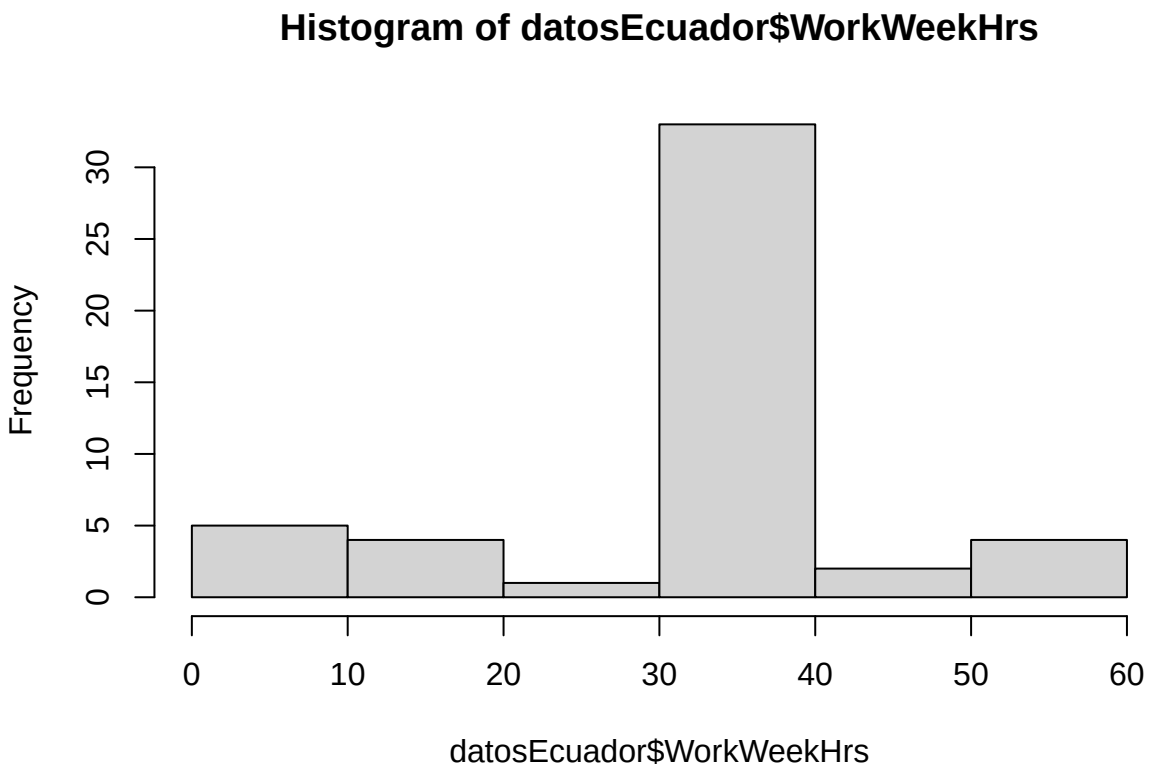
```
#Vemos la variable ConvertedComp
hist(datosEcuador$ConvertedComp)
```



La distribución de los encuestados respecto a su Salario Anual está concentrada en 1 solo grupo, en el rango de $0 a $200000. Luego hay un pequeño grupo, no representativo, donde los encuestados han manifestado ganar cantidades superiores al millón de dolares anuales.

Ahora vamos a analizar el histograma para la variable WorkWeekHrs (Horas semanales de Trabajo)

```
#Vemos la variable WorkWeekHrs
hist(datosEcuador$WorkWeekHrs)
```

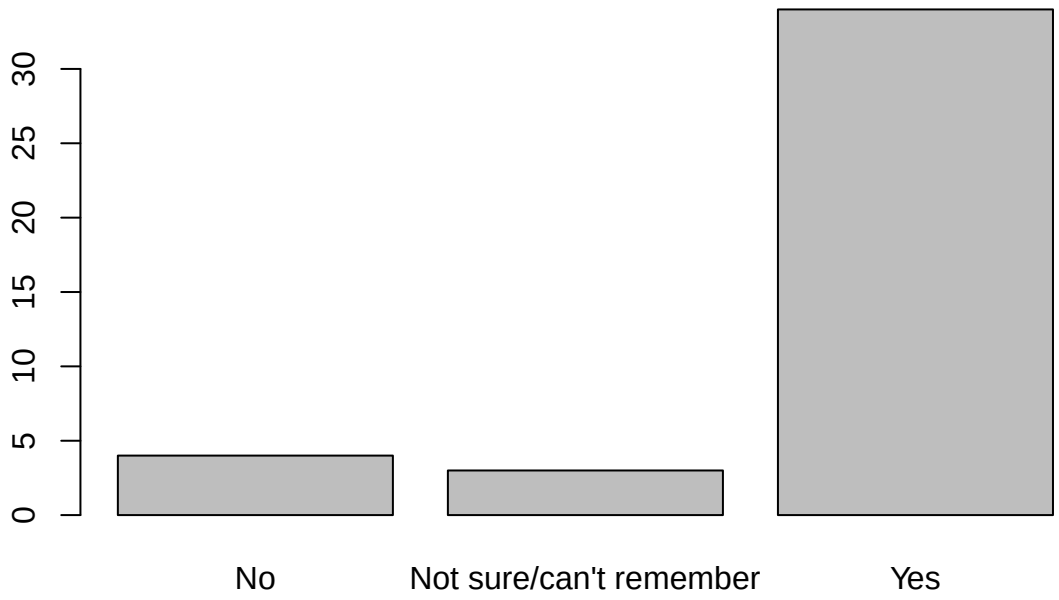


La distribución de los encuestados respecto a su Carga semanal de horas de trabajo está concentrada en 1 grupo, donde más de 30 encuestados manifiestan trabajar entre 30 y 40 horas semanales. Hay un pequeño grupo que manifiesta trabajar más de 40 horas.

> **NOTA**: En el dataset, la variable YearsCode no es númerica del todo, ya que varios desarrolladores seleccionaron tener menos de un año de experiencia (Less than 1 year) y otros tener más de 50 años de experiencia (More than 50 years). Para esto vamos a realizar una adaptación o conversión en estos casos más adelante.

Ahora vamos a analizar un poco las variables categóricas. Primero vamos a comenzar graficando la variable SOAccount

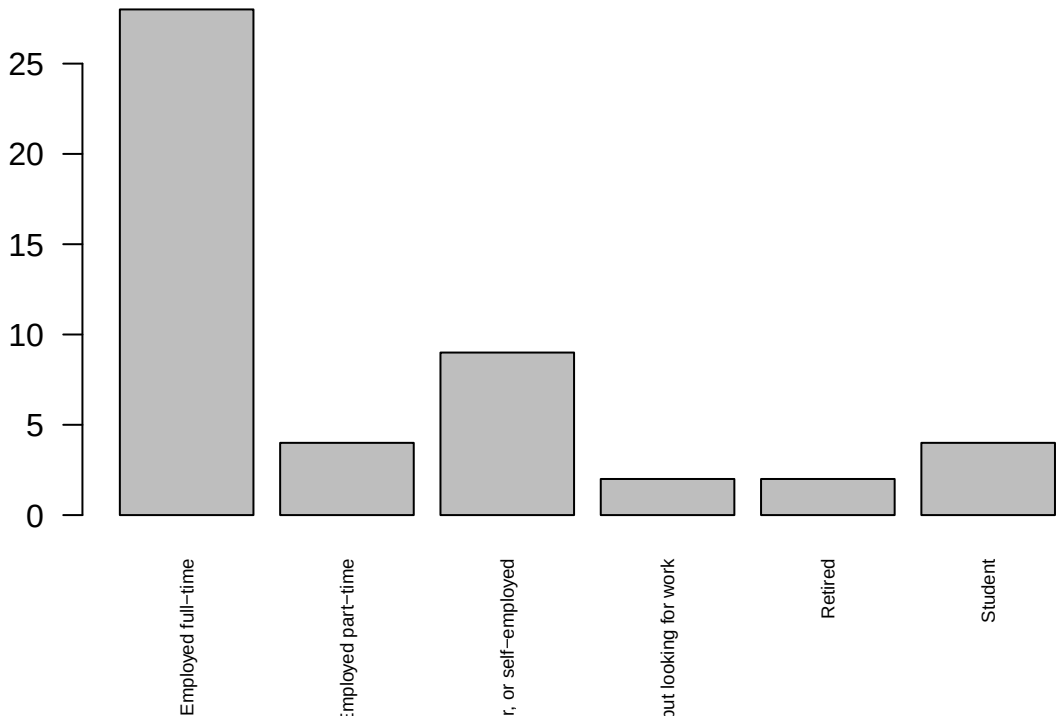
```
barplot(table(datosEcuador$SOAccount))
```



La mayoría tiene una cuenta de StackOverflow, sobrepasando los 30 desarrolladores encuestados. Hay aproximadamente 5 desarrolladores que aseguraron no tener una cuenta o no estar seguros de tenerla.

Ahora vamos a analizar la variable Employment.

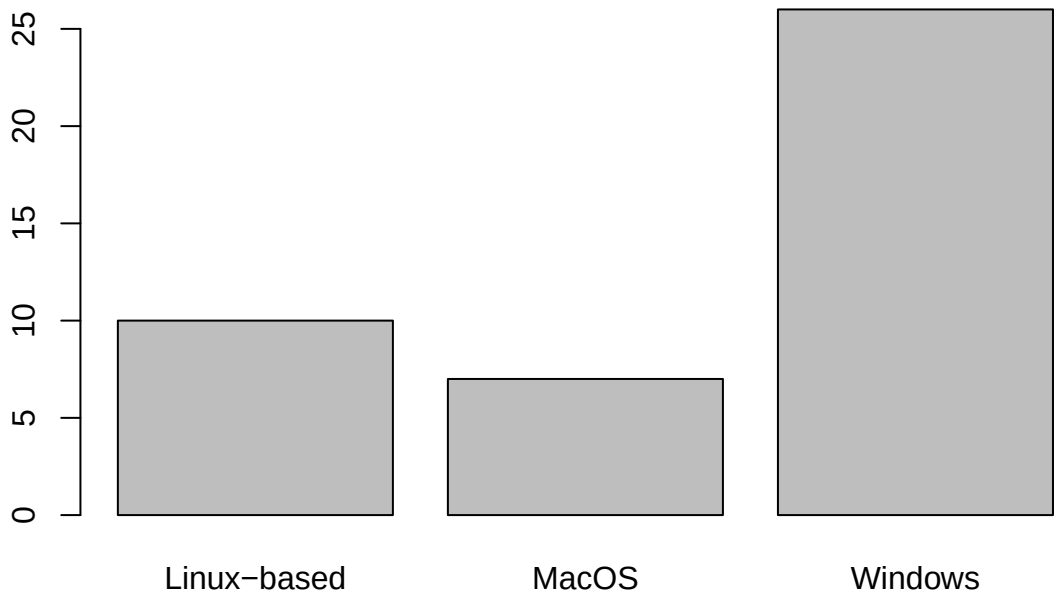
```
barplot(table(datosEcuador$Employment), las=2, cex.names = 0.6)
```



Más del 50% de los encuestados son desarrolladores a tiempo completo. También vemos que una cantidad considerable, aproximadamente un 20% de los participantes son freelance o emprendedores. Hay 2 grupos adicionales que son representativos, los estudiantes y los que trabajan a tiempo parcial, representando un 20% aproximadamente.

Ahora vamos a analizar la variable OpSys.

```
barplot(table(datosEcuador$OpSys))
```



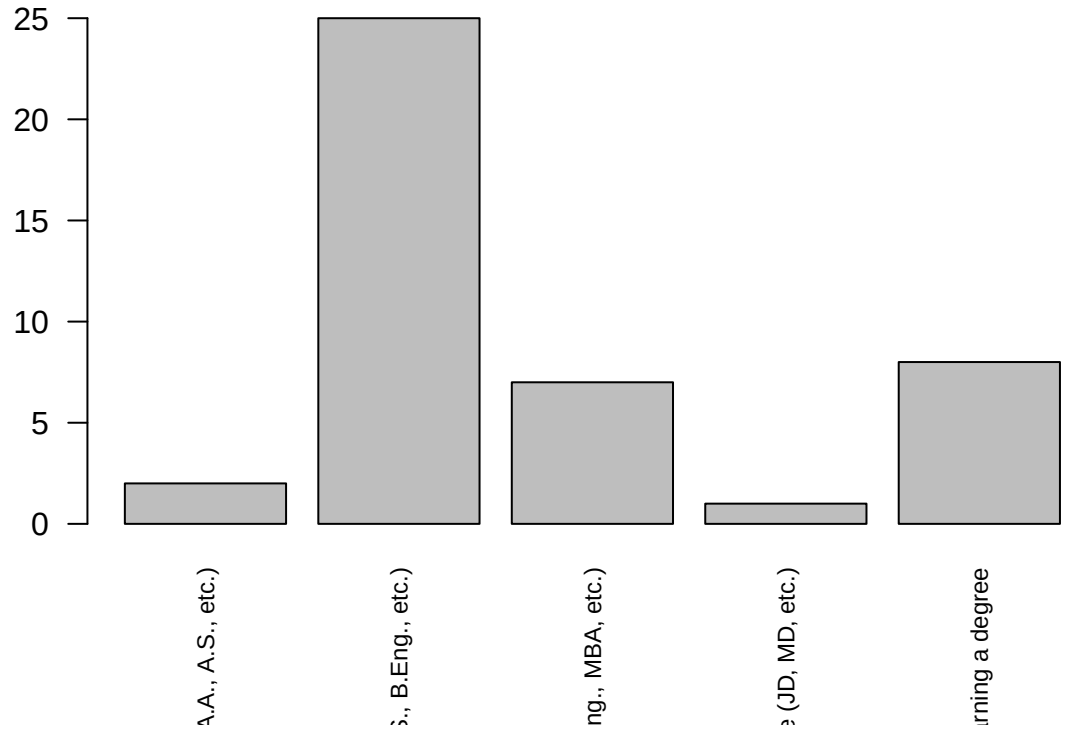
Vemos, como era de esperarse que Windows es el sistema operativo más usado por los encuestados, para sus tareas de desarrollo. También es importante resaltar a las distros basadas en Linux está en 2 lugar entre los *SO* (sistemas operativos) más usados.

Y finalmente para finalizar el análisis univariado, vamos a analizar la formación académica (EdLevel) de los encuestados:

```
table(datosEcuador$EdLevel)
```

```
##
##                     Associate degree (A.A., A.S., etc.)
##                                                       2
##            Bachelor's degree (B.A., B.S., B.Eng., etc.)
##                                                      25
##         Master's degree (M.A., M.S., M.Eng., MBA, etc.)
##                                                       7
##                      Professional degree (JD, MD, etc.)
##                                                       1
## Some college/university study without earning a degree
##                                                       8
```

```
barplot(table(datosEcuador$EdLevel), las=2, cex.names = 0.75)
```

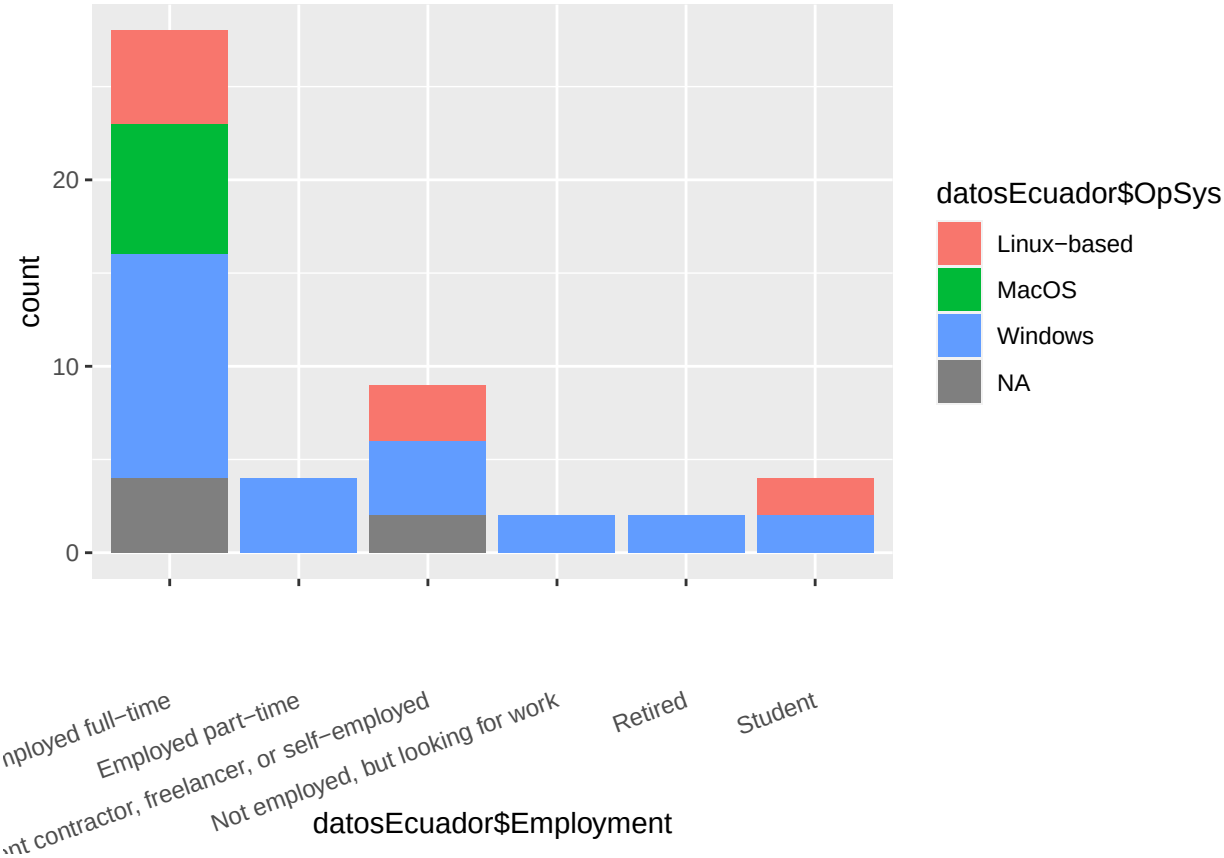


Vemos que se diferencian 2 grupos principales: los que tienen titulos de grado, que son aproximadamente el 50%; y el otro grupo con un menor porcentaje, que no sobrepasa el 20% de los encuestados, han logrado un grado de maestría o títulos relacionados. Hay un tercer grupo, que no es pequeño, agrupando a los desarrolladores que han iniciado sus estudios universitarios y no han concluido, bordeando casi un número de 8 a 10 desarrolladores.

ANÁLISIS MULTIVARIADO

Vamos a realizar un análisis multivariado respecto al sistema operativo usado por los encuestados. Ahora vamos a analizar el sistema operativo usado contra los resultados de la variable Employment

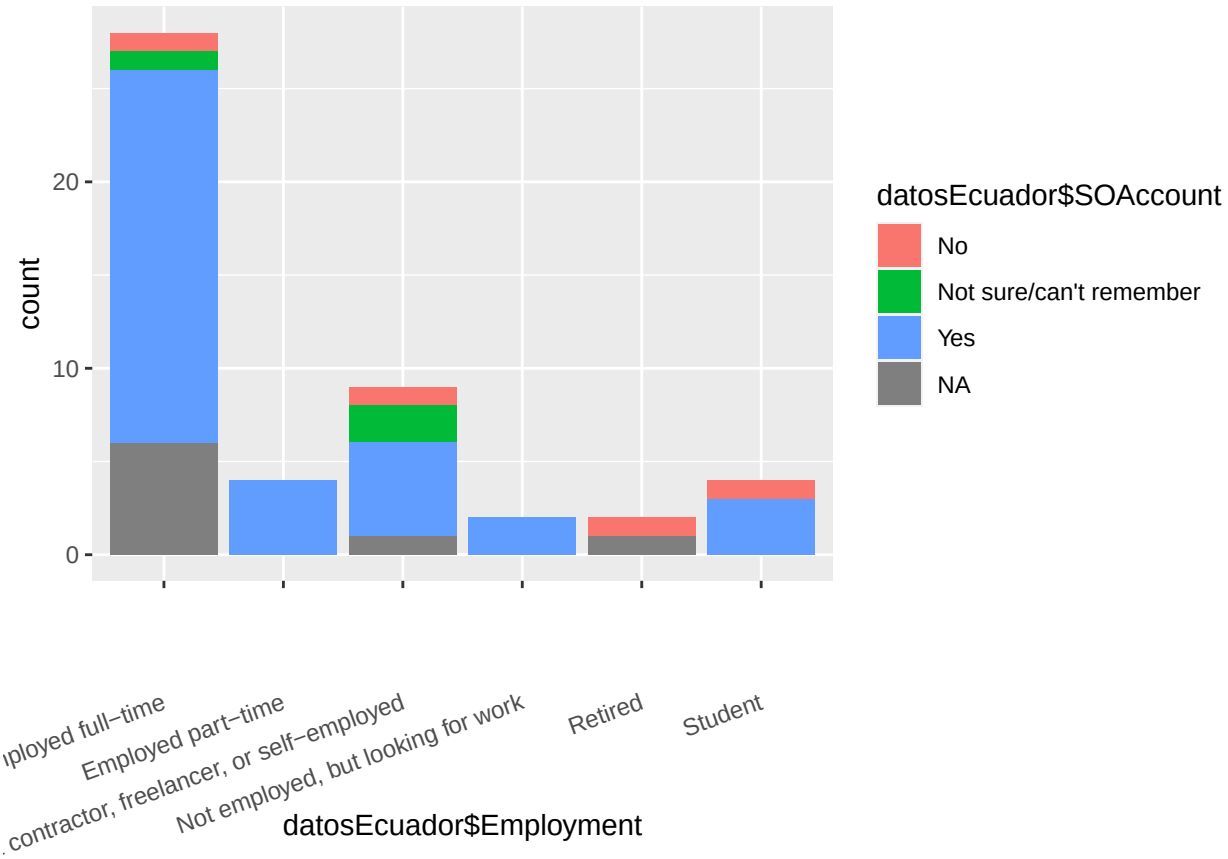
```
ggplot(data=datosEcuador[1:filasEcuador,],aes(x=datosEcuador$Employment,fill=datosEcuador$OpSys))+geom_
```



En este caso vemos que los principales grupos son 3: los que trabajan a tiempo completo, que mayoritariamente usan Windows; los que trabajan de manera independiente, que están más divididos en el sistema operativo que usan, siendo Windows el que mayor porcentaje tiene; y los estudiantes que de igual manera están divides entre Windows y distros basadas en Linux. Existen 3 grupos no representantivos entre los retirados, los no empleados y los que trabajan a tiempo parcial que prefieren Windows por sobre el resto de sistemas operativos.

Ahora vamos a analizar la relación que existe entre las variables Employment vs. SOAccount (¿Tiene cuenta en StackOverflow?)

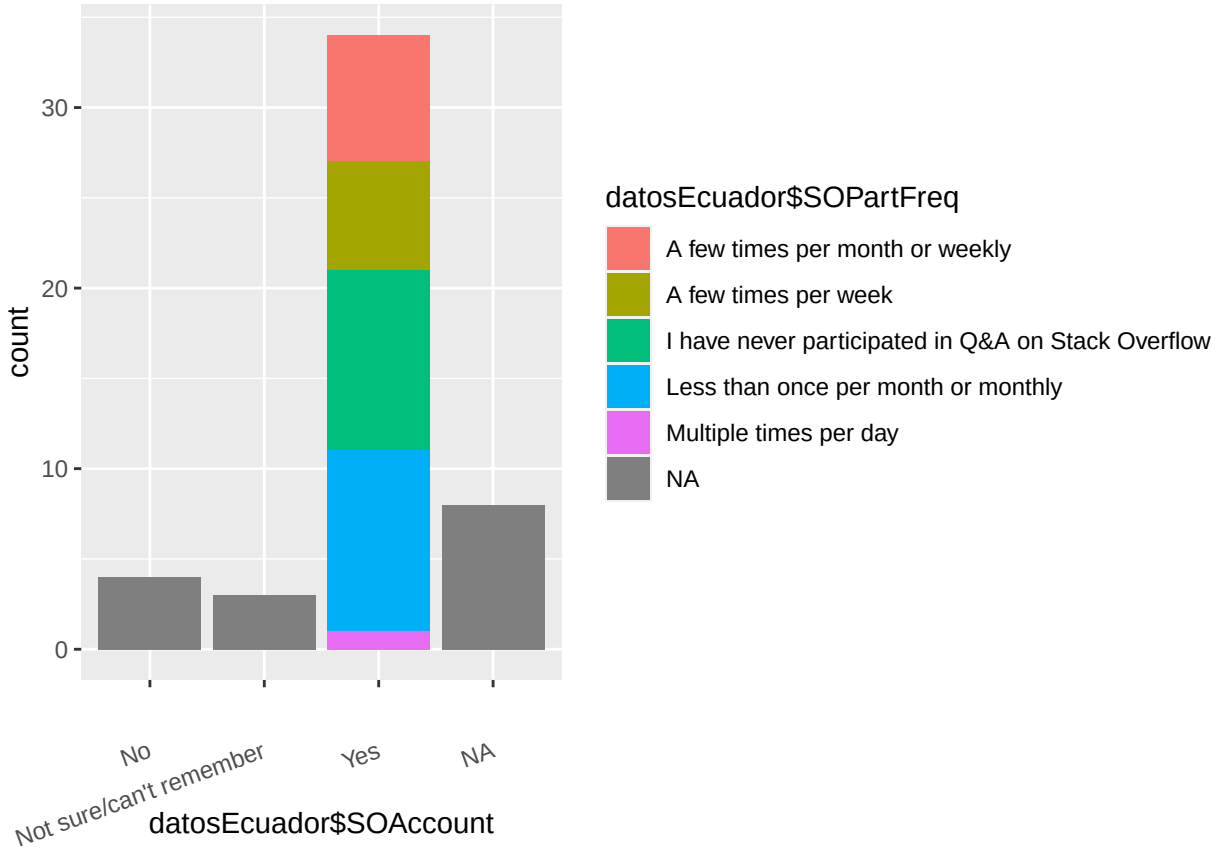
```
ggplot(data=datosEcuador[1:filasEcuador,],aes(x=datosEcuador$Employment,fill=datosEcuador$SOAccount))+g
```



Entre los 5 grupos más representativos gráficamente, se ve que la mayoría de los encuestados tienen una cuenta de StackOverflow. Solo el porcentaje de retirados no están registrados en la plataforma o no están seguros de tenerla, seguramente por ser ya no estar activos.

Ahora vamos a analizar la frecuencia de participación en la comunidad de los desarrolladores que tienen una cuenta en StackOverflow

```
ggplot(data=datosEcuador[1:filasEcuador,],aes(x=datosEcuador$SOAccount,fill=datosEcuador$SOPartFreq))+g
```



Se ve entre los usuarios que tienen una cuenta de StackOverflow, la mayoría participa menos de una vez por mes o nunca ha participado, llegando casi al 40% de los encuestados. Sin embargo, visualmente vemos que los otros grupos de frecuencia (participación diaria, semanal y mensualmente), sumados son un aproximadamente un 20% que participa de manera activa y frecuente en la comunidad.

Se ha realizado un análisis personal y sobre la conducta de los desarrolladores. Ahora para realizar un análisis más técnico debemos aplicar algunas tareas de limpieza o tratamiento de los datos para poder en su momento encontrar también las reglas de asociacion. Para ellos vamos a trabajar sobre los campos:

- DatabaseWorkedWith
- LanguageWorkedWith
- MiscTechWorkedWith
- NEWCollabToolsWorkedWith
- PlatformWorkedWith
- WebframeWorkedWith

Vemos que para la variable **YearsCode**, que corresponde a los años de experiencia de cada desarrollador hay una anomalía en el tipo de dato. Ya que la columna como tal no tiene solo valores numéricos, sino también valos categóricos. Para solucionar esto vamos a convertir los valores categóricos en numéricos y agregaremos otra columna donde posteriormente también se discretice dicha variable y nos sirva para realizar otros análisis posteriores.

25% Explicación clara de cualquier tarea de limpieza o acondicionado que se realiza

3. Realizar tareas de limpieza y acondicionado para poder ser usado en procesos de modelado.

En el análisis previo de los tipos de datos vimos que la variable YearsCode es de tipo caracter y no numérico como se esperaba, por lo que debemos realizar una conversión. **Aunque en el caso de Ecuador no se**

dan casos de respuestas no numéricas (más de 50 años y menos de un año), indicaremos el tratamiento que se debe dar a los casos donde los programadores, hayan optado por estas respuestas categóricas. Para los registros que tienen *"More than 50 years"* definiremos el valor a 50 y para el caso de *"Less than 1 year"* el valor será 1:

```
datosEcuador$YearsCode[datosEcuador$YearsCode=="More than 50 years"] <- 50
datosEcuador$YearsCode[datosEcuador$YearsCode=="Less than 1 year"] <- 1

# Finalmente convertimos dicha columna en númerica
datosEcuador$YearsCode <- as.numeric(datosEcuador$YearsCode)

# Llenamos con la media los valores faltantes
datosEcuador$YearsCode[is.na(datosEcuador$YearsCode)] <- mean(datosEcuador$YearsCode, na.rm=T)
```

Analizamos nuevamente el tipo de dato de la columna antes mencionada

```
#Vemos el tipo de dato de las variables
glimpse(datosEcuador)
```

```
## Rows: 49
## Columns: 18
## $ Age                      <dbl> 35.00000, 31.00000, 24.00000, 30.00000, 36...
## $ ConvertedComp            <dbl> 48000.00, 66289.23, 4200.00, 12000.00, 384...
## $ DatabaseWorkedWith       <chr> "Elasticsearch;MariaDB;MongoDB;MySQL;Redis...
## $ DevType                  <chr> "Developer, back-end;Developer, desktop or...
## $ EdLevel                  <chr> "Professional degree (JD, MD, etc.)", "Bac...
## $ Employment               <chr> "Employed full-time", "Not employed, but l...
## $ Gender                   <chr> "Man", "Man", "Man", "Man", "Man", "Man", ...
## $ LanguageWorkedWith       <chr> "Bash/Shell/PowerShell;HTML/CSS;JavaScript...
## $ MiscTechWorkedWith       <chr> "Node.js", ".NET", "Flutter;Unity 3D", NA,...
## $ NEWCollabToolsWorkedWith <chr> "Confluence;Jira;Github;Gitlab;Slack;Googl...
## $ OpSys                    <chr> "Linux-based", "Windows", "Windows", "Wind...
## $ OrgSize                  <chr> "2 to 9 employees", NA, "2 to 9 employees"...
## $ PlatformWorkedWith       <chr> "Android;AWS;Docker;Google Cloud Platform;...
## $ SOAccount                <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", ...
## $ SOPartFreq               <chr> "A few times per month or weekly", "A few ...
## $ WebframeWorkedWith       <chr> "Angular;Symfony", "Angular;Angular.js;ASP...
## $ WorkWeekHrs              <dbl> 60.00000, 34.35484, 20.00000, 40.00000, 20...
## $ YearsCode                <dbl> 15, 6, 4, 9, 29, 8, 13, 7, 32, 1, 10, 35, ...
```

```
# Otra forma de ver el tipo de dato de cada columna
sapply(datosEcuador, class)
```

```
##                       Age             ConvertedComp       DatabaseWorkedWith
##                 "numeric"                 "numeric"               "character"
##                   DevType                   EdLevel                Employment
##               "character"               "character"               "character"
##                    Gender       LanguageWorkedWith       MiscTechWorkedWith
##               "character"               "character"               "character"
## NEWCollabToolsWorkedWith                     OpSys                   OrgSize
##               "character"               "character"               "character"
##       PlatformWorkedWith                 SOAccount                SOPartFreq
```

```
##              "character"              "character"              "character"
##       WebframeWorkedWith             WorkWeekHrs                YearsCode
##              "character"                "numeric"                "numeric"
```

Vamos también a crear nuevas variables que nos van a servir para aplicar los algoritmos supervisados, no supervisados y reglas de asocioción en su momento.

Primero concantenemos todas las tecnologías usadas en una sola variable, llamada **techs**. Las columnas a concatenar son: *DatabaseWorkedWith*, *LanguageWorkedWith*, *MiscTechWorkedWith*, *NEWCollabToolsWorkedWith*, *PlatformWorkedWith*, *WebframeWorkedWith*.

```
datosEcuador$techs <- paste(datosEcuador$DatabaseWorkedWith, ";", datosEcuador$LanguageWorkedWith, ";",
```

Ahora vamos a agregar nuevas variables que contabilizan el número de tecnologías o herramientas de: bases de datos, lenguajes de programación, de colaboración, entre otros. Primero para la base de datos, vamos a usar la columna *DatabaseWorkedWith*. La variable a crearse será **db_techs**:

```
datosEcuador$db_techs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$DatabaseWorkedWith[i])) {
    datosEcuador$db_techs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$DatabaseWorkedWith[i], ";"), length)
    datosEcuador$db_techs[i] <- longitud
  }
}

summary(datosEcuador$db_techs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   4.000   3.469   5.000  10.000
```

Ahora vamos a agregar una nueva variable para el número de carreras u oficios que tiene el encuestado. Para esto vamos a usar la columna *DevType*, ya que la misma es una concatenación de todas las opciones que seleccionó el participante durante la encuesta. La variable a crearse será **num_types**:

```
datosEcuador$num_types <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$DevType[i])) {
    datosEcuador$num_types[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$DevType[i], ";"), length)
    datosEcuador$num_types[i] <- longitud
  }
}

summary(datosEcuador$num_types)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   4.000   3.857   6.000  10.000
```

Ahora vamos a agregar una nueva variable para el número de lenguajes de programación que usa. Este dato se basa en la experiencia ya adquirida y no en los deseos para usar o aprender el siguiente año. Para esto usaremos la columna *LanguageWorkedWith*. La variable a crearse será **prog_langs**:

```
datosEcuador$prog_langs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$LanguageWorkedWith[i])) {
    datosEcuador$prog_langs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$LanguageWorkedWith[i], ";"), length)
    datosEcuador$prog_langs[i] <- longitud
  }
}

summary(datosEcuador$prog_langs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00    6.00    5.98    8.00   21.00
```

Ahora vamos a agregar una nueva variable para el número de frameworks, librerías y demás herramientras que usa el desarrollador. Este dato se basa en la experiencia ya adquirida y no en los deseos para usar o aprender el siguiente año. Para esto usaremos la columna *MiscTechWorkedWith*. La variable a crearse será **misc_techs**:

```
datosEcuador$misc_techs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$MiscTechWorkedWith[i])) {
    datosEcuador$misc_techs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$MiscTechWorkedWith[i], ";"), length)
    datosEcuador$misc_techs[i] <- longitud
  }
}

summary(datosEcuador$misc_techs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.898   3.000   7.000
```

Haremos lo mismo para el número de herramientras colaborativas que usa el desarrollador, según el contenido de la columna *NEWCollabToolsWorkedWith*. Este dato se basa en la experiencia ya adquirida y no en los deseos para usar o aprender el siguiente año. La variable a crearse será **collab_techs**:

```
datosEcuador$collab_techs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$NEWCollabToolsWorkedWith[i])) {
    datosEcuador$collab_techs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$NEWCollabToolsWorkedWith[i], ";"), length)
```

```
    datosEcuador$collab_techs[i] <- longitud
  }
}

summary(datosEcuador$collab_techs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   3.163   5.000   8.000
```

También vamos a agregar una variable para el número de plataformas que usa el desarrollador. Este dato se basa en la experiencia ya adquirida y no en los deseos para usar o aprender el siguiente año. Usaremos el contenido de la columna *PlatformWorkedWith*. La variable a crearse será **plat_techs**:

```
datosEcuador$plat_techs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$PlatformWorkedWith[i])) {
    datosEcuador$plat_techs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$PlatformWorkedWith[i], ";"), length)
    datosEcuador$plat_techs[i] <- longitud
  }
}

summary(datosEcuador$plat_techs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   3.837   6.000  10.000
```

Finalmente, agregaremos una variable para el número de *frameworks web* que usa el desarrollador. Este dato se basa en la experiencia ya adquirida y no en los deseos para usar o aprender el siguiente año. Para esto usaremos la columna *WebframeWorkedWith*. La variable a crearse será **web_techs**:

```
datosEcuador$web_techs <- 0

for(i in 1:filasEcuador) {
  if (is.na(datosEcuador$WebframeWorkedWith[i])) {
    datosEcuador$web_techs[i] <- 0
  } else {
    longitud  <- sapply(strsplit(datosEcuador$WebframeWorkedWith[i], ";"), length)
    datosEcuador$web_techs[i] <- longitud
  }
}

summary(datosEcuador$web_techs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   3.429   5.000  11.000
```

4. Realizar métodos de discretización

También vamos a aplicar discretización sobre los campos: age, ConvertedComp y WorkWeekHrs

```
# Discretizamos para la variable Age
datosEcuador$segmento_edad <- cut(datosEcuador$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c(

# Discretizamos para la variable ConvertedComp
datosEcuador$segmento_salario <- cut(datosEcuador$ConvertedComp, breaks = c(0,25000,50000,75000,100000,

# Discretizamos para la variable WorkWeekHrs
datosEcuador$segmento_horas_trab <- cut(datosEcuador$WorkWeekHrs, breaks = c(0,20,40,60,80,168), labels

# Discretizamos para la variable YearsCode
datosEcuador$segmento_years_code <- cut(datosEcuador$YearsCode, breaks = c(0,5,10,15,20,30,40,50), label

head(datosEcuador)
```

```
##      Age ConvertedComp                            DatabaseWorkedWith
## 614   35      48000.00    Elasticsearch;MariaDB;MongoDB;MySQL;Redis
## 941   31      66289.23                          Microsoft SQL Server
## 2697  24       4200.00                   MariaDB;PostgreSQL;SQLite
## 2772  30      12000.00 Microsoft SQL Server;MySQL;Oracle;PostgreSQL
## 4669  36      38400.00         Elasticsearch;MariaDB;MongoDB;SQLite
## 5333  28       9600.00                                  Elasticsearch
##
## 614  Developer, back-end;Developer, desktop or enterprise applications;Developer, mobile;DevOps speci
## 941                                                    Developer, back-end;Developer, desktop or
## 2697
## 2772
## 4669                                                                                Developer, back-end;Deve
## 5333                                                                  Data scientist or machine learning
##                                          EdLevel
## 614               Professional degree (JD, MD, etc.)
## 941     Bachelor's degree (B.A., B.S., B.Eng., etc.)
## 2697    Bachelor's degree (B.A., B.S., B.Eng., etc.)
## 2772    Bachelor's degree (B.A., B.S., B.Eng., etc.)
## 4669 Master's degree (M.A., M.S., M.Eng., MBA, etc.)
## 5333 Master's degree (M.A., M.S., M.Eng., MBA, etc.)
##                                          Employment Gender
## 614                                Employed full-time    Man
## 941                 Not employed, but looking for work    Man
## 2697                               Employed part-time    Man
## 2772                               Employed full-time    Man
## 4669 Independent contractor, freelancer, or self-employed    Man
## 5333 Independent contractor, freelancer, or self-employed    Man
##                                                  LanguageWorkedWith
## 614  Bash/Shell/PowerShell;HTML/CSS;JavaScript;PHP;Python;SQL;TypeScript
## 941                                                        C#;JavaScript
## 2697                                  HTML/CSS;Java;JavaScript;Python
## 2772                                          HTML/CSS;JavaScript;SQL
## 4669 Assembly;Bash/Shell/PowerShell;C;C#;C++;Java;JavaScript;Python;Ruby
## 5333                       Go;Java;JavaScript;Python;R;Scala;TypeScript
##                MiscTechWorkedWith
## 614                       Node.js
## 941                          .NET
```

```
## 2697               Flutter;Unity 3D
## 2772                           <NA>
## 4669 Node.js;Unity 3D;Unreal Engine
## 5333                   React Native
##                                                NEWCollabToolsWorkedWith
## 614  Confluence;Jira;Github;Gitlab;Slack;Google Suite (Docs, Meet, etc)
## 941                                                          Jira;Github
## 2697                                                      Github;Gitlab
## 2772              Github;Microsoft Teams;Google Suite (Docs, Meet, etc)
## 4669            Jira;Github;Gitlab;Slack;Google Suite (Docs, Meet, etc)
## 5333                                                             Jira
##            OpSys                                          OrgSize
## 614  Linux-based                                2 to 9 employees
## 941      Windows                                            <NA>
## 2697     Windows                                2 to 9 employees
## 2772     Windows                          1,000 to 4,999 employees
## 4669 Linux-based                                2 to 9 employees
## 5333 Linux-based Just me - I am a freelancer, sole proprietor, etc.
##                                                    PlatformWorkedWith
## 614  Android;AWS;Docker;Google Cloud Platform;Kubernetes;Linux;Microsoft Azure
## 941                                                                <NA>
## 2697                                                 Android;Windows
## 2772                                                 Windows;WordPress
## 4669           Android;Arduino;AWS;Heroku;Linux;Raspberry Pi;Windows;WordPress
## 5333                                Android;AWS;Docker;Kubernetes;WordPress
##      SOAccount                                    SOPartFreq
## 614        Yes                   A few times per month or weekly
## 941        Yes                             A few times per week
## 2697       Yes                             A few times per week
## 2772       Yes              Less than once per month or monthly
## 4669       Yes              Less than once per month or monthly
## 5333       Yes I have never participated in Q&A on Stack Overflow
##              WebframeWorkedWith WorkWeekHrs YearsCode
## 614             Angular;Symfony    60.00000        15
## 941  Angular;Angular.js;ASP.NET    34.35484         6
## 2697                       <NA>    20.00000         4
## 2772                       <NA>    40.00000         9
## 4669       jQuery;Ruby on Rails    20.00000        29
## 5333                   React.js    60.00000         8
##
## 614    Elasticsearch;MariaDB;MongoDB;MySQL;Redis ; Bash/Shell/PowerShell;HTML/CSS;JavaScript;PHP;Pyt
## 941
## 2697
## 2772
## 4669 Elasticsearch;MariaDB;MongoDB;SQLite ; Assembly;Bash/Shell/PowerShell;C;C#;C++;Java;JavaScript;
## 5333
##      db_techs num_types prog_langs misc_techs collab_techs plat_techs web_techs
## 614         5         7          7          1           6          7         2
## 941         1         4          2          1           2          0         3
## 2697        3         1          4          2           2          2         0
## 2772        4         3          3          0           3          2         0
## 4669        4         4          9          3           5          8         2
## 5333        1         4          7          1           1          5         1
##      segmento_edad segmento_salario segmento_horas_trab segmento_years_code
```

```
## 614          30-39      25000-49999              40-59             10-14
## 941          30-39      50000-74999              20-39               5-9
## 2697         20-29          0-24999               0-19               0-4
## 2772         20-29          0-24999              20-39               5-9
## 4669         30-39      25000-49999               0-19             20-29
## 5333         20-29          0-24999              40-59               5-9
```

```
str(datosEcuador)
```

```
## 'data.frame':    49 obs. of  30 variables:
##  $ Age                      : num  35 31 24 30 36 28 32 30 47 32 ...
##  $ ConvertedComp            : num  48000 66289 4200 12000 38400 ...
##  $ DatabaseWorkedWith       : chr  "Elasticsearch;MariaDB;MongoDB;MySQL;Redis" "Microsoft SQL Server"
##  $ DevType                  : chr  "Developer, back-end;Developer, desktop or enterprise applications
##  $ EdLevel                  : chr  "Professional degree (JD, MD, etc.)" "Bachelor's degree (B.A., B.S
##  $ Employment               : chr  "Employed full-time" "Not employed, but looking for work" "Employe
##  $ Gender                   : chr  "Man" "Man" "Man" "Man" ...
##  $ LanguageWorkedWith       : chr  "Bash/Shell/PowerShell;HTML/CSS;JavaScript;PHP;Python;SQL;TypeScri
##  $ MiscTechWorkedWith       : chr  "Node.js" ".NET" "Flutter;Unity 3D" NA ...
##  $ NEWCollabToolsWorkedWith: chr  "Confluence;Jira;Github;Gitlab;Slack;Google Suite (Docs, Meet, etc
##  $ OpSys                    : chr  "Linux-based" "Windows" "Windows" "Windows" ...
##  $ OrgSize                  : chr  "2 to 9 employees" NA "2 to 9 employees" "1,000 to 4,999 employees
##  $ PlatformWorkedWith       : chr  "Android;AWS;Docker;Google Cloud Platform;Kubernetes;Linux;Microso
##  $ SOAccount                : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ SOPartFreq               : chr  "A few times per month or weekly" "A few times per week" "A few ti
##  $ WebframeWorkedWith       : chr  "Angular;Symfony" "Angular;Angular.js;ASP.NET" NA NA ...
##  $ WorkWeekHrs              : num  60 34.4 20 40 20 ...
##  $ YearsCode                : num  15 6 4 9 29 8 13 7 32 1 ...
##  $ techs                    : chr  "Elasticsearch;MariaDB;MongoDB;MySQL;Redis ; Bash/Shell/PowerShell
##  $ db_techs                 : num  5 1 3 4 4 1 4 4 5 9 ...
##  $ num_types                : num  7 4 1 3 4 4 5 0 6 4 ...
##  $ prog_langs               : num  7 2 4 3 9 7 7 6 9 10 ...
##  $ misc_techs               : num  1 1 2 0 3 1 3 3 3 2 ...
##  $ collab_techs             : num  6 2 2 3 5 1 7 2 4 5 ...
##  $ plat_techs               : num  7 0 2 2 8 5 2 3 5 9 ...
##  $ web_techs                : num  2 3 0 0 2 1 6 5 4 5 ...
##  $ segmento_edad            : Factor w/ 8 levels "0-9","10-19",..: 4 4 3 3 4 3 4 3 5 4 ...
##  $ segmento_salario         : Factor w/ 6 levels "0-24999","25000-49999",..: 2 3 1 1 2 1 2 3 2 1 ...
##  $ segmento_horas_trab      : Factor w/ 5 levels "0-19","20-39",..: 3 2 1 2 1 3 2 2 1 2 ...
##  $ segmento_years_code      : Factor w/ 7 levels "0-4","5-9","10-14",..: 3 2 1 2 5 2 3 2 6 1 ...
```

25% Se realiza un proceso de PCA o SVD donde se aprecia mediante explicaciones y comentarios que el estudiante entiende

5. Aplicar un estudio PCA sobre el juego de datos. A pesar de no estar explicado en el material didáctico, se valorará si en lugar de PCA investigáis por vuestra cuenta y aplicáis SVD (Single Value Decomposition).

Tanto PCA y SVD son métodos o técnicas que permiten reducir la dimensionalidad de un dataset. Es decir, en nuestro dataset, que tiene 11 variables, PCA o SVD permite reducir el número de variables, tratando que conservar la mayor representativad posible en los componentes resultantes. Para ello vamos a aplicar de manera simultanea las funciones ya precargadas que provee RStudio, como son: *prcomp*, para aplicar el método PCA (Análisis de componentes principales) y la función *svd* para aplicar descomposición de valores singulares.

La determinación final del número de componentes se tomará en cuenta en base a la sumatoria de la varianza de cada uno. Antes de comenzar a aplicar estos métodos es necesario preparar el dataset:

Vamos a preparar una nueva matriz o dataset con las variables numéricas del objeto dataset **datosEcuador**:

```
# Solo tomamos las variables numéricas del dataset original
A <- datosEcuador[,c(1:2, 17:18, 20:26)]
# Cambiamos los nombres de las filas al índice correspondiente.
rownames(A) <- 1:nrow(A)

head(A)
```

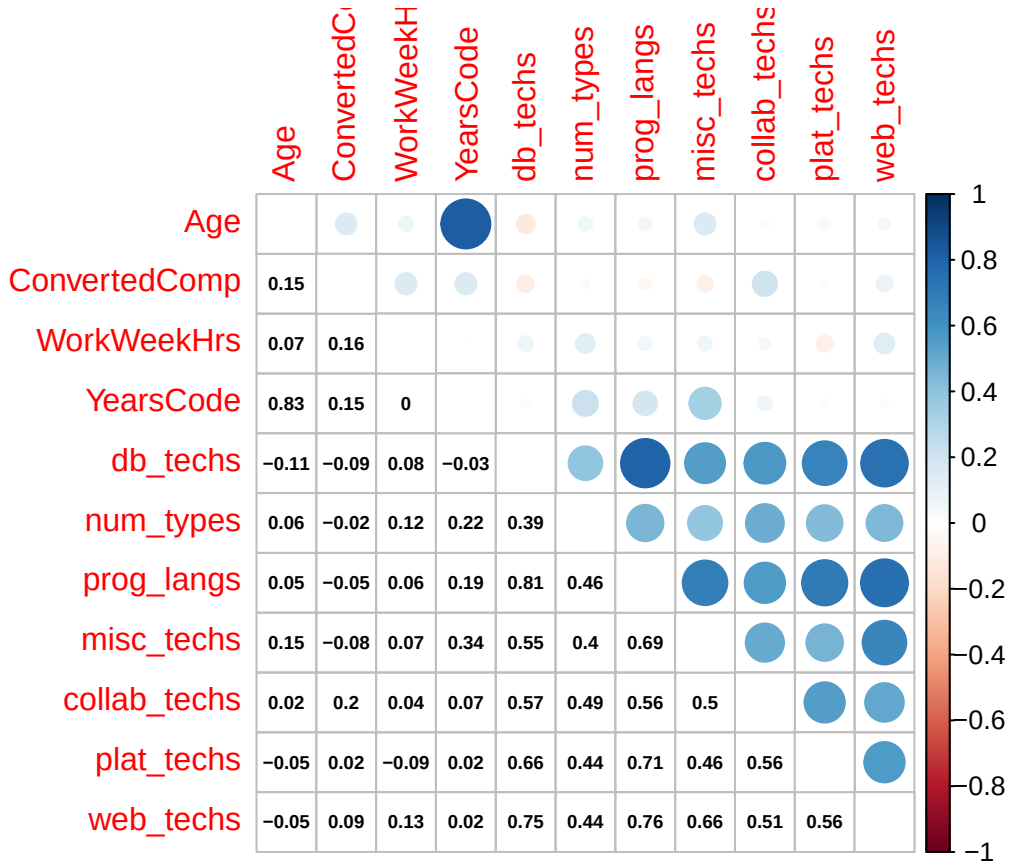
```
##   Age ConvertedComp WorkWeekHrs YearsCode db_techs num_types prog_langs
## 1  35      48000.00    60.00000        15        5         7          7
## 2  31      66289.23    34.35484         6        1         4          2
## 3  24       4200.00    20.00000         4        3         1          4
## 4  30      12000.00    40.00000         9        4         3          3
## 5  36      38400.00    20.00000        29        4         4          9
## 6  28       9600.00    60.00000         8        1         4          7
##   misc_techs collab_techs plat_techs web_techs
## 1          1            6          7         2
## 2          1            2          0         3
## 3          2            2          2         0
## 4          0            3          2         0
## 5          3            5          8         2
## 6          1            1          5         1
```

Ahora vamos a comenzar a introducirnos en los conceptos que involucra aplicar el análisis de componentes principales. Vamos a ver la correlación entre las variables que tenemos. Usaremos la función *corrplot* [1]. Analizaremos visualmente las correlaciones entre ellas:

```
library(corrplot)

correlaciones <- cor(A)

# Graficamos la correlación entre las variables
# corrplot.mixed(correlaciones, method = "circle")
corrplot.mixed(correlaciones, tl.pos = "lt", lower.col = "black", number.cex = .6)
```



En esta gráfica podemos determinar la correlación que existe entre las variables de nuestro dataset, entre más cercano es el color a azul o más grande el círculo, es mejor la correlación entre las variables. Es notorio que las últimas variables calculadas, correspondientes a los diferentes tipos de tecnologías usadas no se relacionan fuertemente con la edad, sueldo anual, horas semanales de trabajo ni con los años de experiencia como programadores.

También podemos concluir que es factible reducir la dimensionalidad del dataset, al menos de 11 a 8 variables, ya que el resto no se relacionan fuertemente con las demás.

Ahora vamos a aplicar el método PCA, con la función integrada *prcomp*. Usaremos 2 parámetros (*center* y *scale*) para estandarizar o normalizar las variables. Cuando el parámetro scale es TRUE, generará el modelo utilizando la matriz de correlación [2]

```
# Aplicamos PCA a la matriz A
pca <- prcomp(A, center = TRUE, scale = TRUE)

summary(pca)
```

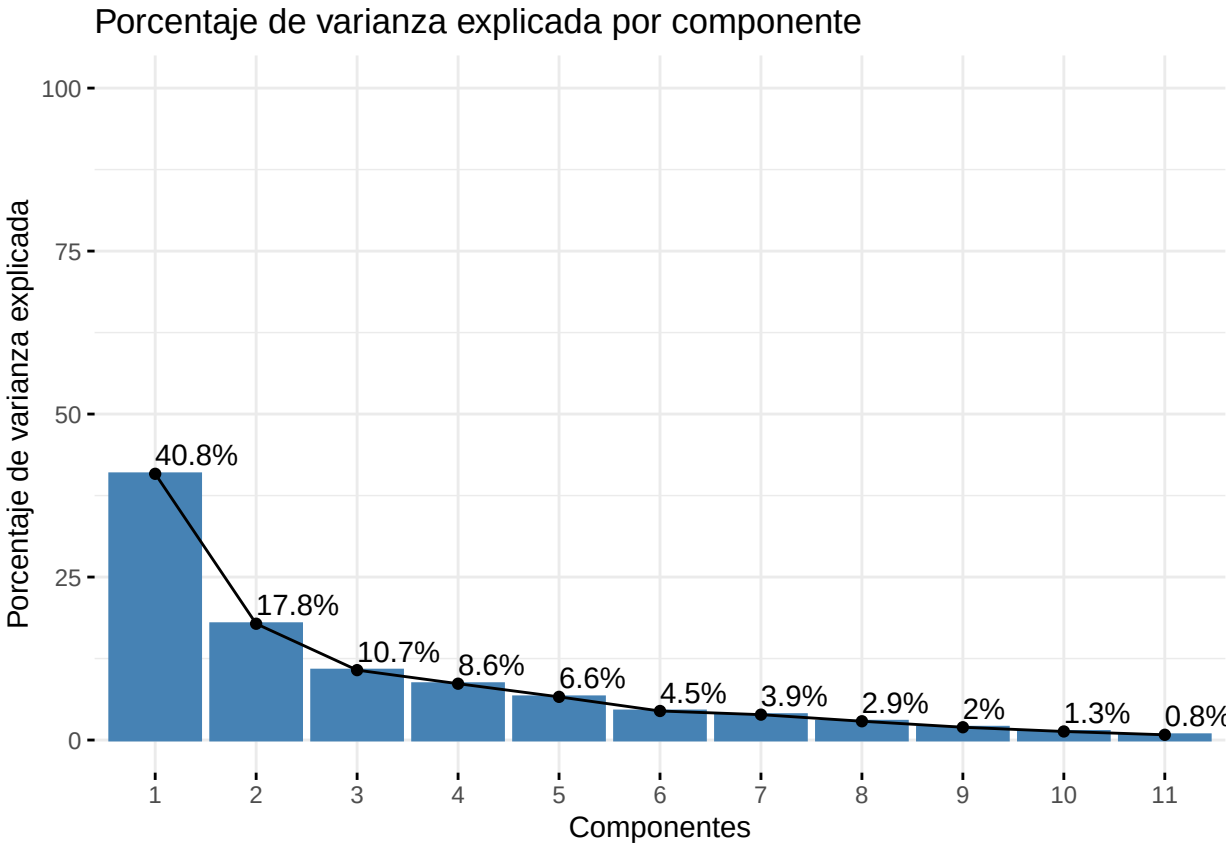
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1195 1.4009 1.0861 0.97511 0.85371 0.70027 0.65471
## Proportion of Variance 0.4084 0.1784 0.1072 0.08644 0.06626 0.04458 0.03897
## Cumulative Proportion  0.4084 0.5868 0.6940 0.78046 0.84672 0.89130 0.93027
##                           PC8     PC9     PC10   PC11
## Standard deviation     0.56324 0.46475 0.38049 0.2984
## Proportion of Variance 0.02884 0.01964 0.01316 0.0081
## Cumulative Proportion  0.95911 0.97874 0.99190 1.0000
```

La reducción de la dimensionalidad a n componentes que cubran una proporción de varianza, superior al 90%, en base al resumen del objeto *pca*, nos deja que entre el componente 6 (PC6) y el componente 7 (PC7) se alcanza este umbral.

También podemos ver esto de manera visual[3]:

```
library(factoextra)

fviz_eig(
    X          = pca,
    choice     = "variance",
    addlabels = TRUE,
    ncp        = 11,
    ylim       = c(0, 100),
    main       = "Porcentaje de varianza explicada por componente",
    xlab       = "Componentes",
    ylab       = "Porcentaje de varianza explicada"
)
```



Ahora vamos a contrastar este resultado preliminar con el resultado de aplicar el algoritmo SVD, para esto vamos aplicar la función *svd* y el parámetro será la matriz escalada de A, la cual también se usó previamente para aplicar PCA.

```
# Aplicamos SVD a la matriz escalada de A
svd <- svd(scale(A))

str(svd)
```

```
## List of 3
##  $ d: num [1:11] 14.68 9.71 7.52 6.76 5.91 ...
##  $ u: num [1:49, 1:11] -0.0853 0.1261 0.1152 0.1055 -0.0962 ...
##  $ v: num [1:11, 1:11] -0.0306 -0.0126 -0.0447 -0.0933 -0.4029 ...
```

Vemos que *svd* retorna 3 matrices, las cuales tienen las siguientes dimensiones:

- d = [1, 11]
- u = [49, 11]
- v = [11, 11]

SVD genera 3 matrices: d, u y v a partir de la matriz o dataset original, en nuestro caso A.

```
# Definimos U
U <- svd$u
# Definimos D
D <- svd$d
# Definimos V
V <- svd$v
```

Según la definición de SVD, para la matriz A es:

$$A = UDV^T$$

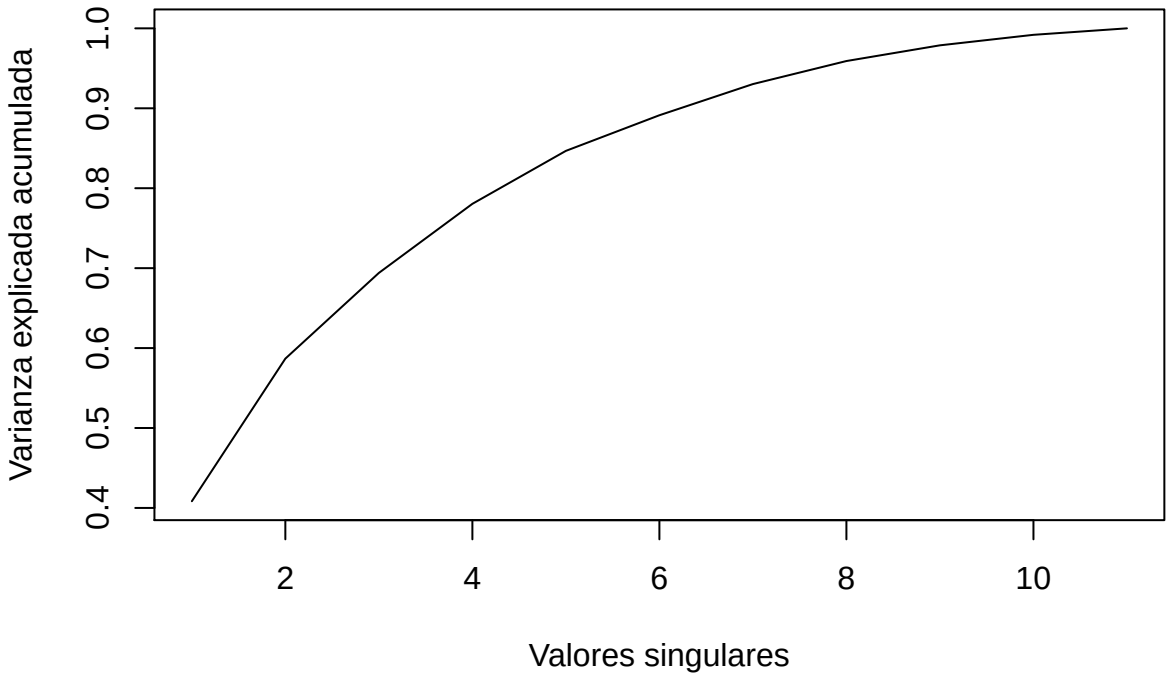
, siendo

$$V^T$$

la matriz transpuesta de V.

Vamos a graficar la varianza explicada acumulada por los valores singulares (D) generados.

```
plot(cumsum(svd$d^2/sum(svd$d^2)), type="l", xlab="Valores singulares", ylab="Varianza explicada acumula
```



Al igual que en PCA, para el valor entre 6 y 8 se alcanza una varianza de 90%

Entre PCA y SVD se genera una matriz que es identica, la cual vamos a comprobar en este momento.

```
# Matriz de rotación generada por el método PCA
pca$rotation
```

```
##                       PC1          PC2         PC3          PC4          PC5
## Age           -0.03057629  0.661771481 -0.10585921  0.026397463 -0.080632950
## ConvertedComp -0.01263164  0.214858227  0.67629612 -0.565476337 -0.164791002
## WorkWeekHrs   -0.04465390  0.083639820  0.64210436  0.686150671 -0.005087036
## YearsCode     -0.09332484  0.660933189 -0.17816160  0.003895437 -0.005054729
## db_techs      -0.40289576 -0.171790438 -0.03188994  0.065954540 -0.211202422
## num_types     -0.29763380  0.075203051  0.03840145  0.098434662  0.834832425
## prog_langs    -0.42825888 -0.017707232 -0.09048257  0.052086869 -0.210110351
## misc_techs    -0.36738964  0.125947138 -0.15250255  0.180796012 -0.201444584
## collab_techs  -0.35150173 -0.009404204  0.18546446 -0.291221208  0.265242837
## plat_techs    -0.36817703 -0.113947317 -0.08279328 -0.262544754  0.045034725
## web_techs     -0.39975311 -0.087464494  0.11040957  0.074448670 -0.260059636
##                       PC6          PC7         PC8          PC9          PC10
## Age            0.261485303 -0.120585131  0.22145954 -0.450614622  3.989995e-02
## ConvertedComp -0.099495523  0.259767620 -0.07503420  0.165459125 -7.831701e-02
## WorkWeekHrs    0.211034099 -0.166755587 -0.16629222  0.012966341  6.070915e-05
## YearsCode      0.004024284  0.068632994 -0.02792596  0.367682147 -1.095287e-01
## db_techs       0.242253418 -0.047742936  0.47270369  0.318871802 -5.886574e-01
## num_types     -0.077882553  0.386211998  0.06643294  0.044069151 -4.483116e-02
## prog_langs     0.160304683  0.096394717  0.01102251  0.402842608  7.238667e-01
```

```
## misc_techs    -0.604008754 -0.108388438 -0.48645701  0.005336468 -2.327109e-01
## collab_techs  -0.185067594 -0.740855312  0.19658723 -0.086436738  1.494556e-01
## plat_techs     0.583042092  0.007513017 -0.57888246 -0.242115021 -1.487633e-01
## web_techs     -0.218425775  0.405043167  0.28020903 -0.549716745  9.563329e-02
##                      PC11
## Age             0.45399749
## ConvertedComp   0.18127714
## WorkWeekHrs    -0.08786198
## YearsCode      -0.60826168
## db_techs        0.15993121
## num_types       0.18455009
## prog_langs      0.19964497
## misc_techs      0.29195912
## collab_techs   -0.18768601
## plat_techs     -0.13341460
## web_techs      -0.38152642
```

Y para la matriz

$$V$$

que corresponde a los vectores singulares izquierdos generados por SVD tenemos:

```
# Matriz V que corresponde a los vectores singulares izquierdos
svd$v
```

```
##              [,1]         [,2]        [,3]         [,4]         [,5]
##  [1,] -0.03057629  0.661771481 -0.10585921  0.026397463 -0.080632950
##  [2,] -0.01263164  0.214858227  0.67629612 -0.565476337 -0.164791002
##  [3,] -0.04465390  0.083639820  0.64210436  0.686150671 -0.005087036
##  [4,] -0.09332484  0.660933189 -0.17816160  0.003895437 -0.005054729
##  [5,] -0.40289576 -0.171790438 -0.03188994  0.065954540 -0.211202422
##  [6,] -0.29763380  0.075203051  0.03840145  0.098434662  0.834832425
##  [7,] -0.42825888 -0.017707232 -0.09048257  0.052086869 -0.210110351
##  [8,] -0.36738964  0.125947138 -0.15250255  0.180796012 -0.201444584
##  [9,] -0.35150173 -0.009404204  0.18546446 -0.291221208  0.265242837
## [10,] -0.36817703 -0.113947317 -0.08279328 -0.262544754  0.045034725
## [11,] -0.39975311 -0.087464494  0.11040957  0.074448670 -0.260059636
##               [,6]         [,7]        [,8]         [,9]         [,10]
##  [1,]  0.261485303 -0.120585131  0.22145954 -0.450614622  3.989995e-02
##  [2,] -0.099495523  0.259767620 -0.07503420  0.165459125 -7.831701e-02
##  [3,]  0.211034099 -0.166755587 -0.16629222  0.012966341  6.070915e-05
##  [4,]  0.004024284  0.068632994 -0.02792596  0.367682147 -1.095287e-01
##  [5,]  0.242253418 -0.047742936  0.47270369  0.318871802 -5.886574e-01
##  [6,] -0.077882553  0.386211998  0.06643294  0.044069151 -4.483116e-02
##  [7,]  0.160304683  0.096394717  0.01102251  0.402842608  7.238667e-01
##  [8,] -0.604008754 -0.108388438 -0.48645701  0.005336468 -2.327109e-01
##  [9,] -0.185067594 -0.740855312  0.19658723 -0.086436738  1.494556e-01
## [10,]  0.583042092  0.007513017 -0.57888246 -0.242115021 -1.487633e-01
## [11,] -0.218425775  0.405043167  0.28020903 -0.549716745  9.563329e-02
##             [,11]
##  [1,]  0.45399749
##  [2,]  0.18127714
##  [3,] -0.08786198
##  [4,] -0.60826168
```

```
##  [5,]  0.15993121
##  [6,]  0.18455009
##  [7,]  0.19964497
##  [8,]  0.29195912
##  [9,] -0.18768601
## [10,] -0.13341460
## [11,] -0.38152642
```

Podemos corroborar que se cumple la condición:

```
all(pca$rotation == svd$v)
```

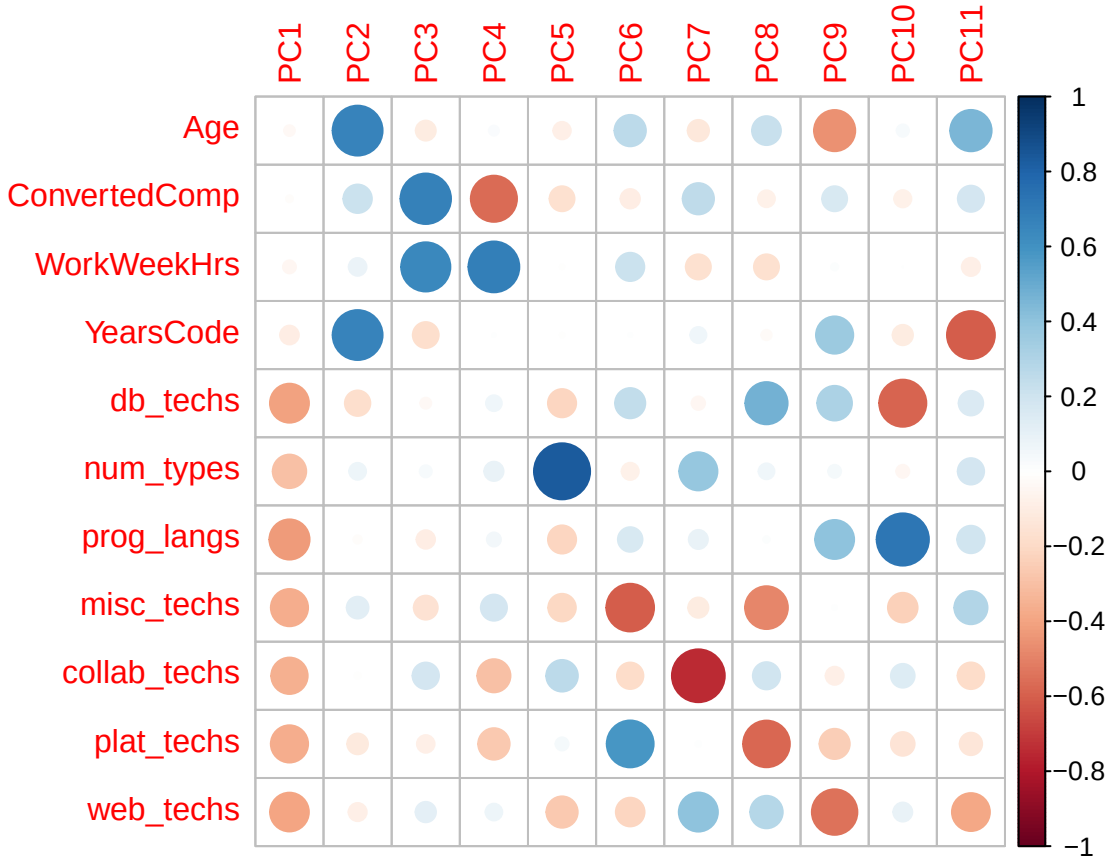
```
## [1] TRUE
```

Conclusión:

Según los resultados obtenidos tanto por PCA y por SVD, la matriz A se puede reducir a una matriz de 7 componentes.

Para lo cual vamos a verificar nuevamente la nueva correlación entre las variables en cada componente.

```
corrplot(pca$rotation, method = "circle")
```



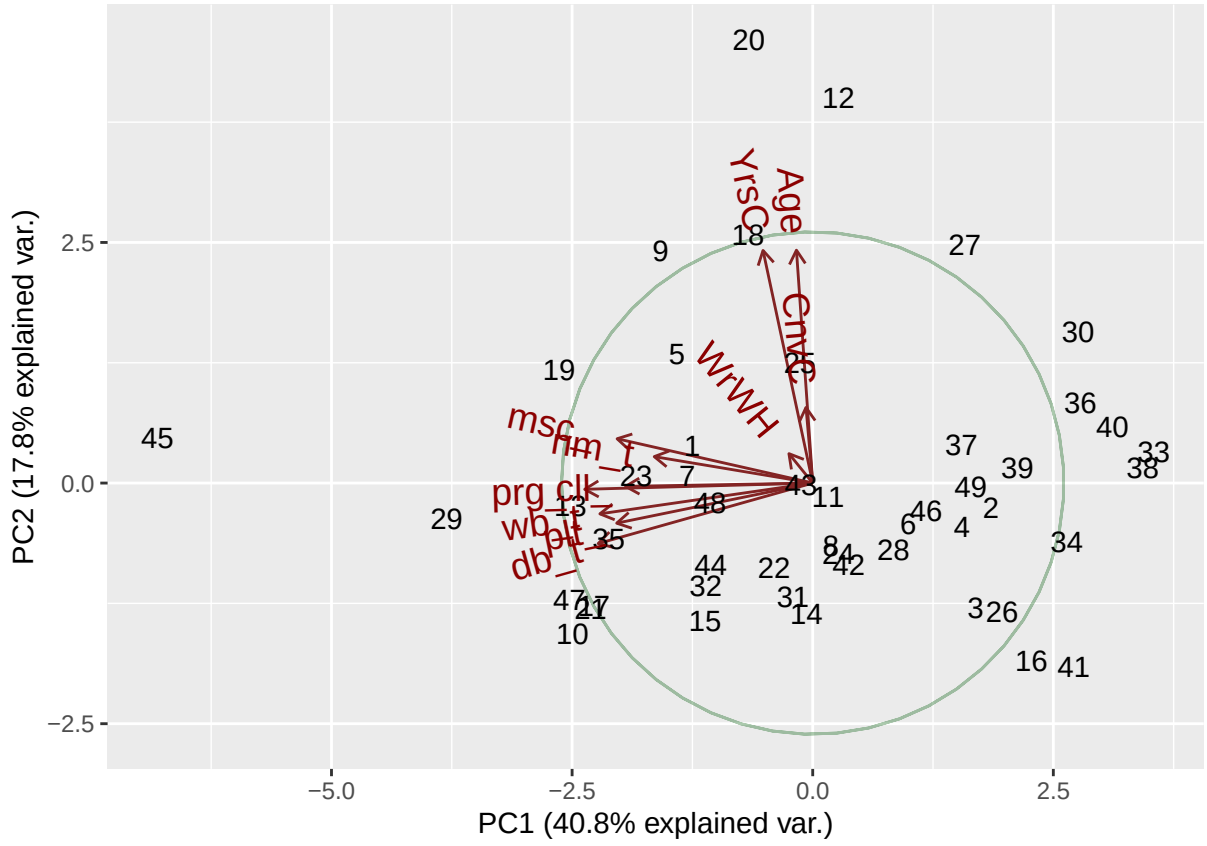
Analizando la gráfica de correlación anterior, ayuda a definir los 7 componentes principales que se van a tomar en cuenta:

- El primer componente (PC1) da mayor ponderación a las variables que corresponden a las tecnologías usadas por cada desarrollador: base de datos, lenguajes de programación, de colaboración, web, entre otras herramientas; así también pondera mejor a la variable que corresponde al número de ocupaciones que tiene un desarrollador. Este componente no pondera mucho a las variables que representan a: la edad, sueldo anual, horas semanales trabajadas y años de experiencia.
- El segundo componente (PC2) da mayor ponderación a las variables que corresponden a la edad y los años de expeiencia del programador. Este componente no pondera mucho a las variables que representan a las tecnologías usadas por el programador.
- El tercer componente (PC3) da una fuerte ponderación a las variables que corresponden al sueldo anual y horas semanales trabajadas por el desarrollador. Este componente da una baja ponderación al resto de variables
- El cuarto componente (PC4) da una ponderación similar a la que da el componente 3 (PC3), a las mismas variables, por lo que este componente no será tomado en cuenta. Por lo que el siguiente componente a tomarse en cuenta será el PC5, el cual pondera mayormente a la variable de los tipos de programadores del desarrollador. PC5 no pondera a las variables de horas semanales trabajadas y años de experiencia. Al resto de variables les da un peso muy bajo.
- El quinto componente (PC6) da una fuerte ponderación a las variables que corresponden a las tecnologías miscelaneas y plataformas usadas por el desarrollador. Este componente da una casi nula ponderación a los años de experiencia y una baja ponderación al resto de variables
- El sexto componente (PC7) da una fuerte ponderación a la variable que corresponden a las herramientas de colaboración usadas por el desarrollador. También pondera significativamente a la variable *num_types* (tipos de desarrolladores) y *web_techs* (tecnologías web).
- El séptimo componente (PC8) da una mejor ponderación a las variables: *db_techs* (tecnologías base de datos), *misc_techs* (tecnologías miscelaneas) y a la variable *plat_techs* (tecnologías de plataformas que domina). Hay una baja correlación para el resto de variables.

Ahora vamos a analizar las relaciones entre el componente 1 (PC1) y el resto de componentes para ver como se comportan las observaciones y relaciones entre componentes. Usaremos la función para visualización **ggbiplot** [4]. Comenzamos por PC1 y PC2:

```
library(devtools)
install_github("vqv/ggbiplot")
require(ggbiplot)


ggbiplot(pcobj = pca,
                  choices = c(1, 2),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,      # Remove variable vectors (TRUE)
                  circle = TRUE,        # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```

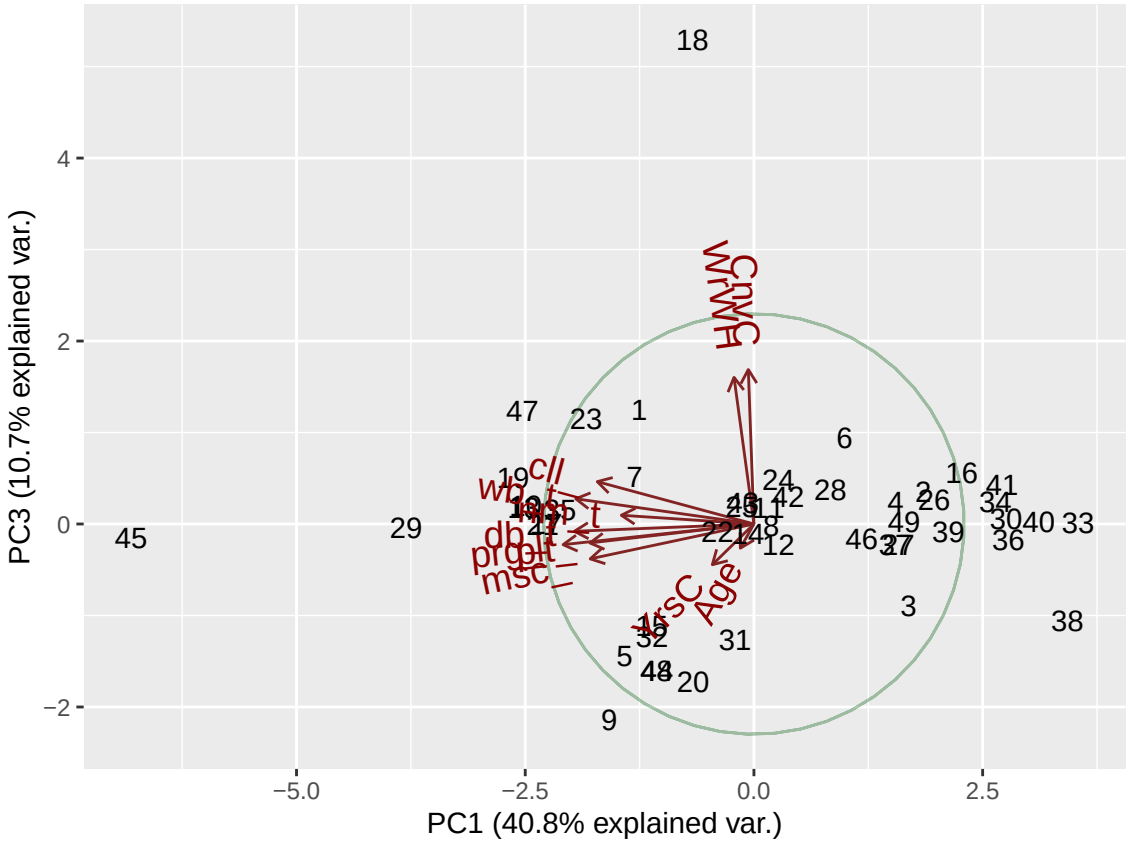


En esta gráfica podemos ver que:

- Las variables: salario anual y horas semanales trabajadas por el desarrollador no está bien representadas, ya que su vector está muy cerca del origen.
- Las variables Edad y años de experiencia programando están muy relacionadas ya que el ángulo que las separa es muy pequeño, además de ser variables que están muy bien representadas por estár muy cerca de la línea del circulo.
- Podemos también corroborar que las variables correspondientes a las tecnologías no están correlacionadas con la edad, experiencia y salario anual.
- Entre los dos componentes (PC1 y PC2) suman una varianza de 58.6% de representatividad respecto al dataset inicial (A).
- La observaciones: *45*, para las variables relacionadas con la tecnología, y *20* para Edad, son casos atípicos, ya que sus valores exceden a la media de cada variable.

Ahora vamos a analizar PC1 y PC3

```
ggbiplot(pcobj = pca,
                  choices = c(1, 3),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,      # Remove variable vectors (TRUE)
                  circle = TRUE,        # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```

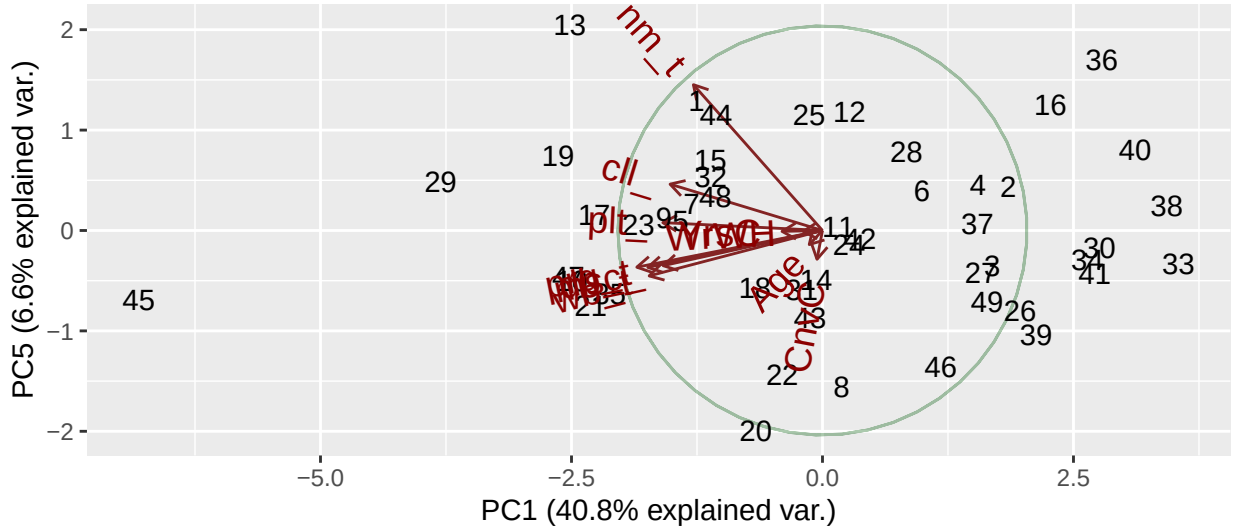


En esta gráfica podemos ver que:

- Las variables: salario anual y horas semanales trabajadas por el desarrollador no está bien representadas en PC1, ya que su vector está muy cerca del origen.
- Las variables Edad y años de experiencia programando están ahora más correlaciondas con las variables de tecnologías. Las variables sueldo anual y horas semanales están fuertemente relacionadas, debido al ángulo muy pequeño entre ambas. Sin embargo, no están correlacionadas con el resto de variables.
- Entre los dos componentes (PC1 y PC3) suman una varianza de 51.5% de representatividad respecto al dataset inicial (A).
- La observaciones: *29*, para las variables que corresponde con lenguajes de programación, y *18* para la variable que corresponde con sueldo anual, son casos atípicos, ya que sus valores exceden a la media de cada variable.

Ahora vamos a analizar PC1 y PC5

```
ggbiplot(pcobj = pca,
                  choices = c(1, 5),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,       # Remove variable vectors (TRUE)
                  circle = TRUE,         # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```

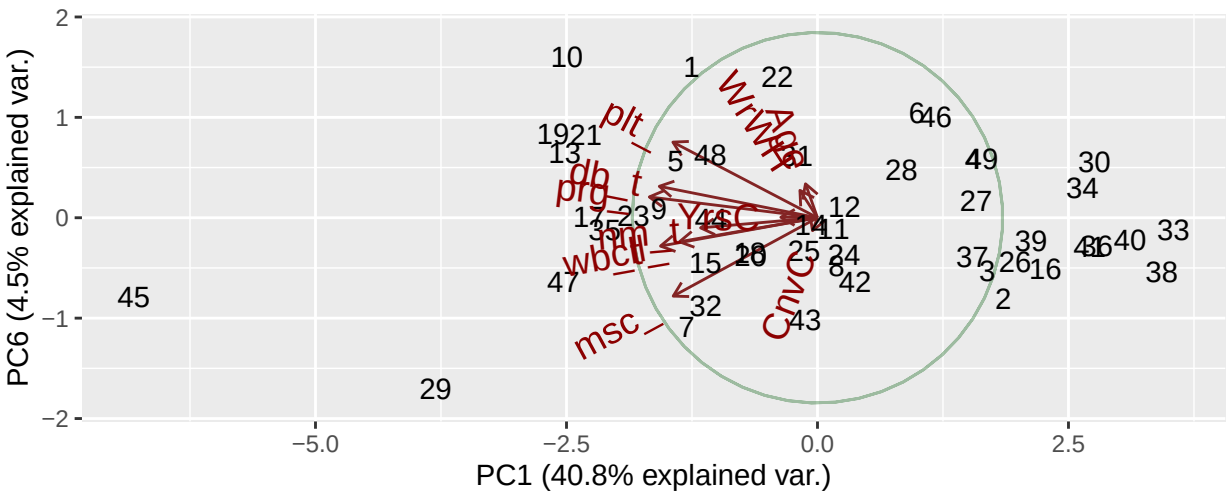


En esta gráfica podemos ver que:

- Las variables: tipos de programadores ahora está mejor representada en el componente 5, y está mejor correlacionada con las variables que corresponden a las tecnologías usadas por el desarrollador.
- Las variables sueldo anual, edad, horas semanales trabajadas y años de experiencia están mal representadas.
- Entre los dos componentes (PC1 y PC5) suman una varianza de 47.4 de representatividad respecto al dataset inicial (A).
- En este análisis vemos que hay muchos valores dispersos que corresponden a las obsrervaciones.

Ahora vamos a analizar PC1 y PC6

```
ggbiplot(pcobj = pca,
                  choices = c(1, 6),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,       # Remove variable vectors (TRUE)
                  circle = TRUE,         # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```

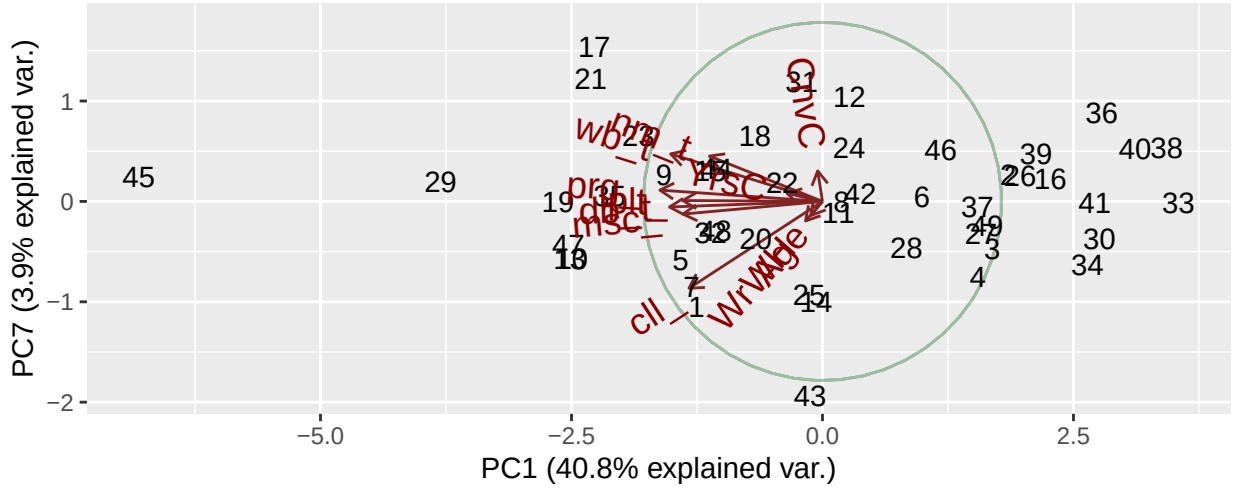


En esta gráfica podemos ver que:

- La mayoría de variables están correlacionadas, a excepción las variables que corresponden a horas semanales trabajadas y tecnologías miscelaneas usadas por el programador, ya que el ángulo que las separa es de 90, es decir perpendicular.
- Las mismas variables ya mencionadas en los análisis anteriores siguen sin estar bien representadas en el componente 6.
- Entre los dos componentes (PC1 y PC6) suman una varianza de 45.3 de representatividad respecto al dataset inicial (A).
- En este análisis vemos que hay muchos valores dispersos que corresponden a las obsrervaciones.

Ahora vamos a analizar PC1 y PC7

```
ggbiplot(pcobj = pca,
                  choices = c(1, 7),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,       # Remove variable vectors (TRUE)
                  circle = TRUE,         # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```

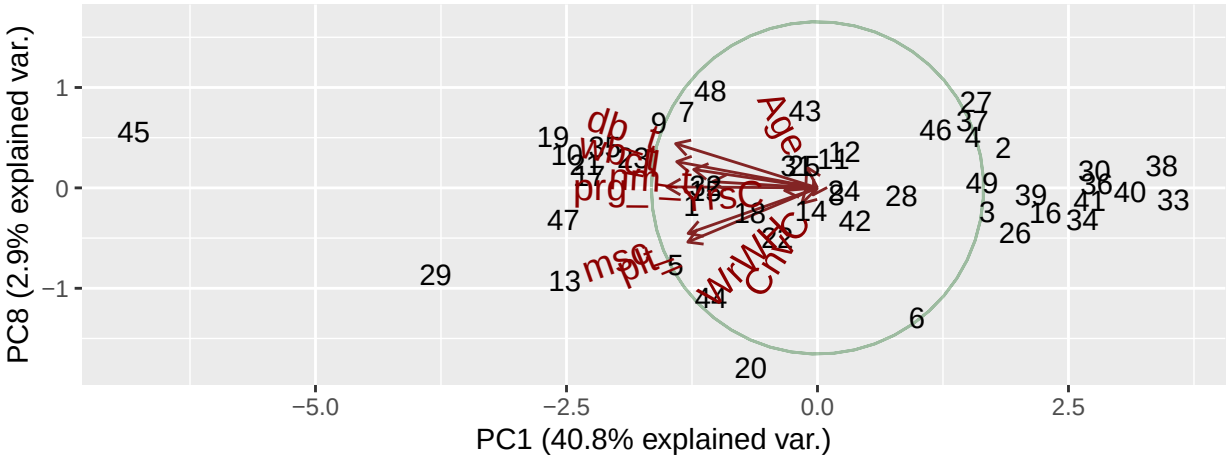


En esta gráfica podemos ver que:

- La mayoría de variables siguen estando muy correlacionadas en PC7, la que tiene mayor representativad es tecnologías colaborativas.
- Entre los dos componentes (PC1 y PC7) suman una varianza de 44.7% de representatividad respecto al dataset inicial (A).
- En este análisis vemos que hay muchos valores dispersos que corresponden a las obsrervaciones, formando pequeñas agrupaciones en todo el espacio bidimensional.

Y Finalmente vamos a analizar PC1 y PC8

```
ggbiplot(pcobj = pca,
                  choices = c(1, 8),
                  obs.scale = 1, var.scale = 1,  # Scaling of axis
                  labels = row.names(A),     # Add labels as rownames
                  labels.size = 4,
                  varname.size = 5,
                  varname.abbrev = TRUE,  # Abbreviate variable names (TRUE)
                  var.axes = TRUE,       # Remove variable vectors (TRUE)
                  circle = TRUE,         # Add unit variance circle (TRUE)
                  ellipse = TRUE, groups = A$type) # Adding ellipses
```



En esta gráfica podemos ver que:

- La mayoría de variables siguen estando muy correlacionadas en PC8. La mayoría de variables que corresponden a tecnologías se correlacionan fuertemente en este componente.
- Entre los dos componentes (PC1 y PC8) suman una varianza de 43.7% de representatividad respecto al dataset inicial (A).

Rúbrica

- 25%. Justificación de la elección del juego de datos donde se detalle el potencial analítico que se intuye. El estudiante deberá visitar los siguientes portales de datos abiertos para seleccionar su juego de datos:
  - Datos.gob.es
  - UCI Machine Learning
  - Datasets Wikipedia
  - Datos abiertos Madrid
  - Datos abiertos Barcelona
  - London Datastore
  - NYC OpenData
- 25%. Información extraída del análisis exploratorio. Distribuciones, correlaciones, anomalías,...

- 25%. Explicación clara de cualquier tarea de limpieza o acondicionado que se realiza. Justificando el motivo y mencionando las ventajas de la acción tomada.
- 25%. Se realiza un proceso de PCA o SVD donde se aprecia mediante explicaciones y comentarios que el estudiante entiende todos los pasos y se comenta extensamente el resultado final obtenido.

Recursos de programación

- Incluimos en este apartado una lista de recursos de programación para minería de datos donde podréis encontrar ejemplos, ideas e inspiración:
  - Material adicional del libro: Minería de datos Modelos y Algoritmos
  - Espacio de recursos UOC para ciencia de datos
  - Buscador de código R
  - Colección de cheatsheets en R

Bibliografía

---

[1] An Introduction to corrplot Package. Alojado en https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

[2] 21 - Análisis de componentes principales en RStudio. Alojado en https://www.youtube.com/watch?v=6BeuHCo1gZQ. Por Juan Gabriel Gomila Salas

[3] Detección de anomalías: Autoencoders y PCA. Alojado en https://www.cienciadedatos.net/documentos/52_deteccion_anomalias_autoencoder_pca.html. Por Joaquín Amat Rodrigo.

[4] Biplot of PCs using ggbiplot function. Alojado en https://agroninfotech.blogspot.com/2020/06/biplot-for-principal-component-analysis.html

---