REV. INVESTIG. ALTOANDIN

REPORTE DE CASO

Rev. Investig. Altoandin. 2017; Vol 19 Nº 2: 201 - 210

http:huajsapata.unap.edu.pe/ria - http://dx.doi.org/10.18271/ria.2017.279
Abril - Junio - ISSN: 2306-8582 (Versión impresa)
ISSN: 2313-2957 (Versión digital)

Journal of High Andean Research
VOL 19 Nº 2
Revista de Investigaciones Altoandinas

RMySQL para el análisis de datos de postulantes e ingresantes del área biomédicas a la Universidad Nacional del Altiplano – Puno Perú

RMySQL for the analysis of data of postulants and entrants of the biomedical area to the National University of the Altiplano – Puno Perú

Adolfo Carlos Jiménez Chura

Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano Puno – Perú. Correspondencia email: adolfocarlos300@hotmail.com.

INFORMACIÓN DEL ARTÍCULO

Artículo recibido 31-03-2017
Artículo aceptado 18-06-2017
On line: 26-06-2017

PALABRAS CLAVES:

estadística,
funciones,
librerías,
procesos de admisión,
tecnología.

RESUMEN

El objetivo fue aplicar *RMySQL*, *dplyr* y funciones de graficación de la base de datos de los procesos de admisión de la Comisión Central de Admisión de la Universidad Nacional del Altiplano. Se implementó el código fuente de las funciones que retornan *data.frames* de las observaciones de la base de datos y funciones para generar los gráficos de forma dinámica tales como ggplot y barplot; se usó *summarise, select, group_by, merge, arrange, unique* y otras del proyecto R. El resultado indica que el porcentaje de ingresantes, de los procesos de admisión ordinario y extraordinario, en función a la cantidad de postulantes es: el 22.91% ingresan a Medicina, Veterinaria y Zootecnia; el 10.87% a Enfermería; 26.83% a Biología; 2.44% a Medicina Humana; 17.64% a Nutrición Humana y 7.25% a Odontología, comprendidos entre el 17 de marzo del 2013 al 22 de enero del 2017; igualmente, el porcentaje de ingresantes por procedencia de colegio del área urbana y rural se obtuvo que la Escuela Profesional de Biología tiene el mayor porcentaje, 27.52% y 19.48% del área urbana y rural; Medicina, Veterinaria y Zootecnia un 22.87% y 15.53% del área urbana y rural; Nutrición Humana un 18.48% y 8.09% del área urbana y rural; Enfermería un 11.45% y 5.62% del área urbana y rural; Odontología un 7.51% y 3.24% del área urbana y rural; y Medicina Humana un 2.49% y 2.09% área urbana y rural.

ARTICLE INFO

Article received 31-03-2017
Article accepted 18-06-2017
Online: 26-06-2017

KEY WORDS:

statistics,
functions,
libraries,
admission processes,
technology

ABSTRACT

The objective was to apply RMySQL, dplyr and graphing functions of the database to the processes of admission at the Central Admission Commission at the National University of the Altiplano. The source code of the functions that return data.frames of the observations of the database and functions was implemented to generate the graphs of dynamic form such as ggplot and barplot; summarise, select, group_by, merge, arrange, unique and others of project R were used. The result indicates that the percentage of entrants, of the ordinary and extraordinary admission processes, according to the number of applicants is: 22.91% were admitted to Medicine, Veterinary and Animal Science; 10.87% to Nursing; 26.83% a Biology; 2.44% in Human Medicine; 17.64% to Human Nutrition and 7.25% to Dentistry, included between March 17, 2013 and January 22, 2017; also, the percentage of admissions by origin of school in the urban and rural area obtained shows that the Professional School of Biology has the highest percentage, 27.52% and 19.48% of the urban and rural area; Medicine, Veterinary and Animal Husbandry 22.87% and 15.53% of the urban and rural area; Human Nutrition 18.48% and 8.09% of the urban and rural area; Nursing 11.45% and 5.62% of the urban and rural area; Dentistry 7.51% and 3.24% of the urban and rural area; and Human Medicine 2.49% and 2.09% urban and rural area.

1) Tesista Facultad Ciencias Agrarias/Ingeniería Agronómica UNA-PUNO

INTRODUCCIÓN

La Universidad Nacional del Altiplano a través de la oficina de la Comisión Central de Admisión (CCA) organiza procesos de admisión para alcanzar una vacante de ingreso en sus dos modalidades: examen ordinario (general y cepreuna) y extraordinario. En cada proceso de admisión se recopila información de los postulantes tales como Escuela Profesional a la que postula, género, fecha de nacimiento, lugar de nacimiento, tipo de colegio, área al cual pertenece (rural, urbano), nombre del colegio, etc. Se tiene gran cantidad de información que puede ser analizada con un software estadístico tal como STATISTICA, EVIEWS, STATA, SAS, S-PLUS, SPSS, MATLAB, R entre otros (Mirabal Sosa, 2010). Los programas mencionados son comerciales excepto R, que es un software libre especialmente desarrollado para el análisis estadístico y la presentación gráfica de los datos.

R es un proyecto que surgió en 1990 y se deben a Ross Ihaka y Robert Gentleman del Departamento de Estadística de la Universidad de Auckland. R y un software libre y compila en las tres plataformas más importantes: Linux, Windows y MacOS (Flores Sánchez, 2013) y existe un grupo de personas que desde 1997 se ocupan del mantenimiento del sistema denominada *The R Core-Development Team*. (Mirabal Sosa, 2010).

Existen diversos investigadores de diferentes áreas que usan las características de R para crear paquetes avalados por CRAN (The Comprehensive R Archive Network) (R, 2017) de aplicación general o particular, paquetes como ggplot2 para graficación; *dplyr* y sus cinco (5) funciones básicas filter, select, arrange, mutate y summarise; para integración con gestores de base de datos se tiene RPgSQL (PostgreSQL), ROracle (Oracle), RMySQL (MySQL), RODBC (origen de datos ODBC) (Tussell F., 2005) y RSQlite (SQLite). Estos y muchos otros paquetes pueden descargarse desde la página oficial en https://www.r-project.org.

Grandes empresas están incluyendo R en su estrategia de análisis predictivo, el 70% de encuestados por "Rexer Analytics" indica que usan R para minería de datos y análisis científico (Tecnológica, 2014).

R es un lenguaje de programación simple que admite condicionales, iteraciones, creación de funciones, funciones definidas por el usuario y tiene una sintaxis muy similar a C/C++ (Mirabal Sosa, 2010) y gracias a estas instrucciones se pueden realizar cálculos fundamentalmente estádisticos (Team, 2008) además, contiene dos instrucciones muy importantes y son *install.packages(nombPaquete)* que permite descargar e instalar el paquete y *library (nombPaquete)* que permite usar las variables, funciones y objetos implementados del paquete.

Para presentar información en gráficos estadísticos y tomar ciertas decisiones, lo primero es realizar consultas específicas sobre la base de datos, luego exportarlo a un software tal como Excel, darle el formato adecuado para finalmente obtener el gráfico. RMySQL es un paquete que se integra con la base de datos MySQL y tiene la ventaja de crear consultas en una base de datos relacional con información en tablas, filas y columnas, de forma muy similar a un *data.frame* (tabla) con observaciones (filas) y variables (columnas) y, con paquetes adicionales se puede obtener un gráfico estadístico con sólo invocar a las funciones con los parámetros adecuados.

Se tiene antecedentes sobre la comunidad científica que viene produciendo diferentes publicaciones en investigación resaltando las ventajas que ofrece R, por ejemplo en el campo de la investigación psicológica (Anchía, 2010) donde se aplica cálculos de correlaciones policóricas con el paquete *polycor*; en el campo ecológico se tiene el paquete WaterML para gestionar los datos experimentales de diferentes instituciones utilizando tecnología de código abierto para establecer una comunicación con una base de

datos relacional a través del paquete RObsDat (Kadlec, Bryn, Daniel P., y Richard A., 2015); también el paquete ARNN para la predicción de series de tiempo no lineales usando en redes neuronales autoregresivas (Velasquez, Laura, y Cristian, 2011).

Por los puntos mencionados este estudio tiene por objetivo cargar información de postulantes e ingresantes en un *data.frame* realizadas con consultas SQL (Standard Query Language) (Piattini Velthuis, Martines, Muñoz, y Sánchez, 2006) simples o complejas con el paquete RMySQL como medio de interface con la base de datos MySQL, seguidamente realizar el procesamiento de los datos con las ventajas que ofrece el paquete *dplyr* (Wickham, A Grammar of Data Manipulation, 2016) y así evitar realizar consultas SQL de agrupación, filtrado, selección, etc. sobre la base de datos y con el uso de las funciones de graficación como ggplot, barplot obtener e interpretar información mencionada del área biomédicas de la Universidad Nacional del Altiplano.

MATERIALES Y MÉTODOS

La presente investigación tuvo lugar en la oficina de la Comisión Central de Admisión de la Universidad Nacional del Altiplano – Puno, durante los meses de enero y marzo del 2017. La población de estudio estuvo conformada por los postulantes e ingresantes del área biomédicas: Medicina Veterinaria y Zootecnia, Enfermería, Biología, Medicina Humana, Nutrición Humana y Odontología. Los procesos de admisión en estudio están comprendidos desde el 17 de marzo del 2013 al 22 de enero del 2017.

Para el entorno de trabajo se usó el software RStudio (RSudio, 2017), el paquete RMySQL (Jeroen Ooms, 2016) como interface entre la base de datos y R y del paquete dplyr las funciones implementadas (Wickham y Francois, 2017) para realizar las agrupaciones de datos, selección, filtrado de una o más variables almacenados en un objeto de tipo *data.frame* de postulantes e ingresantes.

Para establecer la conexión a la base de datos se usó la función *dbConnect* de la librería RMySQL:

```
Library(RMySQL)

con dbConnect(RMySQL::MySQL(),
        dbname = "database",
        user="root",
        password = "123",
        port = 3306)
```

R al igual que otros lenguajes de programación permite crear funciones con los comandos e instrucciones requeridos. Se creó un script donde se encuentra los paquetes a cargar, funciones y variables a ser usadas para trabajar con la información. Para la carga del script en memoria y trabajar con las funciones se empleó la siguiente instrucción:

```
source("c: R/articulo.R")
```

La información de los procesos de admisión se encuentra almacenada en variables y se clasificó según el tipo de examen: general, cepreuna y extraordinario en las siguientes variables de tipo vector:

Figura 1. Variables de los procesos de admisión

```
proc_general<-
c("c_postulante_01g_22_01_2017",
  "c_postulante_01g_21_08_2016",
  …)
proc_cepreuna<-
c("c_postulante_02c_18_12_2016",
  "c_postulante_02c_18_09_2016",
  …)
proc_extraordinario<-
c("c_postulante_03e_12_03_2016",
  "c_postulante_03e_14_03_2015",
  …)
```

Fuente. Elaboración propia

Se implementó la función "carreras" que retorna una lista con las variables id_carrera, id_grupo y nombre de la Escuela Profesional. El código de la función es el siguiente:

Figura 2. Función que retorna la lista de Escuelas Profesionales y resultado desde interfaz RStudio.

```
function()
{
  SQL <- paste("SELECT id_carrera, id_grupo, nombre",
               "FROM c_carrera ORDER BY orden", sep = " ")
  return (data.frame(dbGetQuery(con, SQL)))

}
source("d:/www/R/articulo.R")
carreras()
 id_carrera id_grupo                          nombre
          1        1 Medicina Veterinaria y Zootecnia
          2        1                      Enfermería
          3        1                        Biología
          4        1                 Medicina Humana
          5        1                Nutrición Humana
          6        1                     Odontología
```

Fuente. Elaboración propia

Para procesar la información de postulantes e ingresantes se implementó una función que retorna en una lista de datos de las variables proceso, departamento, provincia y si ingreso o no a la Escuela Profesional. Ésta recibe dos parámetros que indican el proceso a trabajar y el área biomédica.

Figura 3. Función que retorna relación de postulantes e ingresantes por departamento y provincia.

```
function(proceso, grupo, idcarrera = 0)
{
  switch(proceso, proc<-proc_general, proc<-proc_cepreuna,
         proc<-proc_extraordinario, proc<-proc_todos)
  if(idcarrera==0){
    carrera <- ""
  }else{
    carrera<-paste(" AND p.id_carrera1=",idcarrera, sep = "")
  }
  salida<-list()
  for(i in 1:length(proc)) {
    SQL <- paste("SELECT SUBSTRING('", proc[i],"', 16) AS ",
         "'proceso',c.id_carrera, c.nombre as 'escuela',",
         "d.nombre as 'depart', pro.nombre as 'prov',",
         "if(p.paso=1, 'Si','No') as 'ingreso' FROM ",proc[i],
         " as p RIGHT join c_departamento as d ON ",
         "SUBSTRING(p.ubi_nacimiento,1,2)=d.IdDepartamento ",
         "RIGHT join c_provincia as pro ON ",
         "SUBSTRING(p.ubi_nacimiento,1,4)=pro.IdProvincia ",
         "INNER JOIN c_carrera AS c ON p.id_carrera1=",
         "c.id_carrera WHERE p.habilitado=1 AND p.formulario",
         "=3 AND c.id_grupo=",grupo, carrera," ORDER BY ",
         "c.id_carrera, pro.nombre;", sep = "")
    salida[[i]] <- data.frame(dbGetQuery(con, SQL))
  }
  result <- do.call(rbind,
              lapply(salida,
                function(x)
                  x[match(names(salida[[1]]), names(x))]))
  return(result)
}
```

Fuente. Elaboración propia

La función de la figura 3 tiene por nombre *Data.post.ing.dep.prov* y retorna un *data.frame.*

El primer parámetro toma el valor de uno (1) que indica que se procesó las tablas de la base de datos cuyos nombres se encuentran en la variable tipo vector *proc_general*, si es dos (2) indica que procesó las tablas de la variable *proc_cepreuna,* tres (3) indica que se procesó las tablas de la variable *proc_extraordinario,* el segundo parámetro con el valor de uno (1) indica que es el área biomédica y el tercer parámetro, opcional, es el código de la Escuela Profesional.

Se usó el paquete dplyr para realizar las agrupaciones por *departamento, provincia* y el campo *ingreso* de todos los procesos de admisión, según las variables ya mencionadas. La instrucción usada es:

```
datosresult %>%
group_by(depart,prov) %>%
summarise(total=n())
```

se implementó la siguiente función para generar el gráfico de cantidad de postulantes o ingresantes por proceso.

Figura 4. Uso de barplot para representar la información de postulantes

```
function (proceso, grupo, ingreso, depart="PUNO",
          idcarrera = 0, visible=50){
  result<-Data.post.ing.dep.prov(proceso, grupo, idcarrera)
  filas<-nrow(result)
  if(idcarrera == 0){
    nombrecar<-"◆rea biom◆dicas"
  }else{
    nombrecar<-grupocarrera[grupocarrera$id_carrera==idcarrera,1]
  }
  yTitulo<-"postulantes"
  if(ingreso=="Si"){
    result<-result[result$ingreso=="Si",]
    yTitulo <- "ingresantes"
  }
  datos<-result %>% group_by(depart,prov) %>% summarise(total=n())
  dep<-datos[datos$depart==depart,]

  par(mar = c(8.0, 3.7, 2.8, 1) + 0.2, col.main="blue",
      col.lab="red", font.main=1, font.lab=1, font.sub=1,
      cex.main=0.9, cex.lab=0.8, cex.sub=0.8)
  graf<-barplot(dep$total, names=dep$prov,
      main = paste("Cant. de ", yTitulo,
      " por provincia\nDepartamento de ",depart,"\n",
      nombrecar, sep = ""),
      xlab = "", ylab = yTitulo, xaxs="i", yaxs="i",
      ylim = c(0,max(dep$total)+visible), cex.axis = 0.65,
      cex=0.65, las=2, border = 1, col = rgb(166,182,244,
      maxColorValue = 255))
  text(graf, dep$total, labels=lab, cex = 0.6, pos=3)
  grid(10, NA, lwd = 0)
  box()
}
```

Fuente. Elaboración propia

La siguiente función *Data.grupo.colegio.area* permite determinar la cantidad de postulantes e ingresantes de los colegios de la zona URBANA y zona RURAL de los diferentes procesos de admisión por Escuela Profesional del departamento de Puno. Dependiendo del parámetro proceso, éste tomará valores de las tablas de la base de datos almacenados en las variables *proc_general, proc_cepreuna y proc_extraordinario*, si se asigna el valor de cero (0) tomará todos los procesos de las tres (3) variables mencionadas. Para obtener datos de la Escuela Profesional se tien el parámetro *idcarrera.*

Figura 5. Retorna relación de postulantes por zona urbana/rural por Escuela Profesional

```
function(depart="PUNO",proceso = "0", grupo=0, idcarrera=0){
  observ <- data.frame()
  proc <- switch (proceso,"0" = proc_todo,
    "1" = proc_general,   "2" = proc_cepreuna,
    "3" = proc_extraordinario
  )
  grupo<-ifelse(grupo==0, "", paste(" AND c.id_grupo = ",
                grupo, sep = ""))
  idcar<-ifelse(idcarrera==0, "",
        paste(" AND c.id_carrera = ",idcarrera, sep = ""))
  depart <- ifelse(depart=="", "",
        paste(" AND co.departamento='",depart,"'",sep = ""))
  for (i in 1:length(proc)){
    SQL <- paste("SELECT ", i," AS orden,SUBSTRING('",
            proc[i],"', 16) AS 'proceso',c.id_grupo,",
            "c.id_carrera,c.nombre AS 'carrera',co.codigo,",
            "co.nombre AS 'colegio', co.departamento,",
            "co.provincia, co.distrito, co.tipo_colegio,",
            "co.area,IF(p.paso=1, 'Si','No') as 'ingreso'",
            " FROM ", proc[i], " AS p INNER JOIN c_carrera ",
            "AS c ON p.id_carrera1=c.id_carrera INNER JOIN ",
            "c_colegioperu AS co ON p.id_colegio=co.codigo ",
            "WHERE p.habilitado=1 and p.formulario=3 ",
            grupo, idcar, depart, sep = "")
    temp <- dbGetQuery(con, SQL)
    observ <- rbind(observ, temp)
  }
  return (observ)
}
```

Fuente. Elaboración propia

RESULTADOS Y DISCUSIÓN

Análisis de postulantes: La implementación de la función *Data.post.ing.dep.prov* de la figura 3 realiza una iteración sobre todas las tablas de la base de datos MySql, en esta línea (Velasquez, Montoya y Cataño, 2010) dan un aporte interesante acerca de la abstracción de datos en base a la programación funcional en R, por lo que en la figura 6 se puede apreciar la utilidad de la función mencionada obteniendo un resultado de 14583, 7624 y 287 postulantes (observaciones) de los diferentes departamentos del Perú. La cantidad de postulantes de la región de Puno alcanzan una cifra de 13433, 7196 y 264 postulantes, lo que indica que el 92% de postulantes provienen de la región de Puno para el examen general, el 94% para el examen cepreuna y el 92% para el examen extraordinario, esto indica que cerca al 7.3% provienen de otros departamentos. Estos resultados se obtuvieron con el código de la figura 6.

Figura 6. Código para la obtención de la cantidad de postulantes por proceso

```
result1<-data.post.ing.dep.prov(1,1)
result2<-data.post.ing.dep.prov(2,1)
result3<-data.post.ing.dep.prov(3,1)
general<-result1 %>% group_by(id_carrera, escuela) %>%
  summarise(tot_gen=n())
cepreuna<-result2 %>% group_by(id_carrera, escuela) %>%
  summarise(tot_cepre=n())
extra<-result3 %>% group_by(id_carrera, escuela) %>%
  summarise(tot_extra=n())
res<-merge(general, cepreuna, by="escuela")
resumen<-merge(res, extra, by="escuela")
resumen<-arrange(unique(resumen[,c(6,1,3,5,7)] ),id_carrera)
summarise(resumen, total=sum(tot_gen))
summarise(resumen, total=sum(tot_cepre))
summarise(resumen, total=sum(tot_extra))
```

Fuente. Elaboración propia

Figura 7. Cantidad de postulantes por proceso, área biomédica

```
> resumen
  id_carrera           escuela tot_gen tot_cepre tot_extra total
1          1 Med. Vet. y Zoot.    1372       417         9  1798
2          2        Enfermería    3683      1425        25  5113
3          3          Biología    1065       474         4  1543
4          4   Medicina Humana    5317      3614       201  9132
5          5 Nutrición Humana    1474       621        14  2109
6          6       Odontología    1692      1073        34  2799
7         NA              <NA>   14583      7624       287    NA
```

Fuente. Elaboración propia

Análisis de ingresantes: La elección de la escuela profesional es importante, tal como se indica en (Duque, *et al*,, 2012), ya que es de vital importancia para el proyecto de vida del estudiante, esto se puede observar en la Escuela Profesional de Biología tiene la mayor cantidad de ingresantes en un 26.83%, por lo que existe una mayor preferencia por esta profesión frente a Medicina Veterinaria y Zootecnia que cuenta con un 22.91%, Nutrición Humana con 17.64%, Enfermería con 10.87%, Odontología con 7.25% y Medicina Humana con 2.44%.

Figura 8. Cantidad de ingresantes por proceso, área biomédica

```
> ing
  id_carrera           escuela tot_gen tot_cepre tot_extra total
1          1 Med. Vet. y Zoot.     223       185         4   412
2          2        Enfermería     310       237         9   556
3          3          Biología     227       184         3   414
4          4   Medicina Humana     118        84        21   223
5          5 Nutrición Humana     201       164         7   372
6          6       Odontología     103        87        13   203
7         NA              <NA>    1182       941        57    NA
```

Fuente. Elaboración propia

A la variable *ing* de la figura 8, se agregó una columna *porcent* de ingresantes y se relacionó con la variable post con la función *inner_join* del paquete *dplyr*. El código implementado es el siguiente:

```
ing$porcentpaste(round(ing$total*100/
post$total,2),"%", sep = "")
porcinner_join(ing,post,by="escuela")
porcselect(porc,1,10,5,6)
```

figura 9. Porcentaje de ingresantes por Escuela Prof.

```
> porc
            escuela post ing porcent_ing
1 Med. Vet. y Zoot. 1798 412      22.91%
2       Enfermeria 5113 556      10.87%
3         Biologia 1543 414      26.83%
4   Medicina Humana 9132 223       2.44%
5  Nutrición Humana 2109 372      17.64%
6       Odontologia 2799 203       7.25%
```

Fuente. Elaboración propia

Análisis de postulantes por provincia del departamento de Puno: Este análisis muestra que los postulantes de la provincia de Moho no cuentan con una vocación, tal como indica (Vilaseca, 2009) tiene poco interés por aportar conocimiento en el área biomédica siendo ésta zona rural, o puede darse el caso de que no postularon a la Universidad. El resultado indicaque: 4354 provienen de la provincia de Puno, 3118 de la provincia de San Román, 1240 de Azángaro, 949 de Melgar, 704 de El Collao, 669 de Huancané, 517 de Chucuito, 456 de Sandia, 398 de Carabaya, 314 de San Antonio de Putina, 287 de Yunguyo y 75 de Moho. Para los procesos de admisión cepreuna y extraordinario se muestra en la figura 11. Se puede observar que la proporción de postulantes se mantiene en las diferentes modalidades.

Figura 10. Código para generar cantidad de postulantes por departamento y provincia

```
post1<-result1 %>%
  select(depart, prov) %>%
  filter(depart=="PUNO") %>%
  group_by(depart, prov) %>%
  summarise(total = n())
post2<-result2 %>%
  select(depart, prov) %>%
  filter(depart=="PUNO") %>%
  group_by(depart, prov) %>%
  summarise(total = n())
post3<-result3 %>%
  select(depart, prov) %>%
  filter(depart=="PUNO") %>%
  group_by(depart, prov) %>%
  summarise(total = n())
total.post<-inner_join(post1, post2, by=c("depart", "prov"))
total.post<-inner_join(total.post, post3,by=c("depart", "prov"))
total.post<-rename(total.post, general = total.x,
                   cepreuna=total.y, extraord=total )
```

Fuente. Elaboración propia

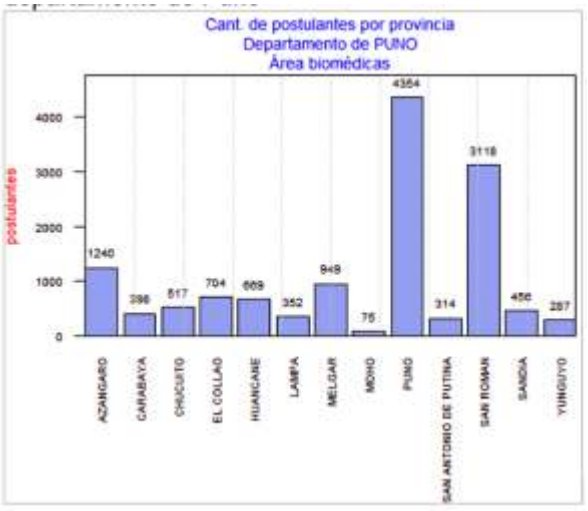
Figura 11. Cantidad de postulantes del departamento de Puno y provincias

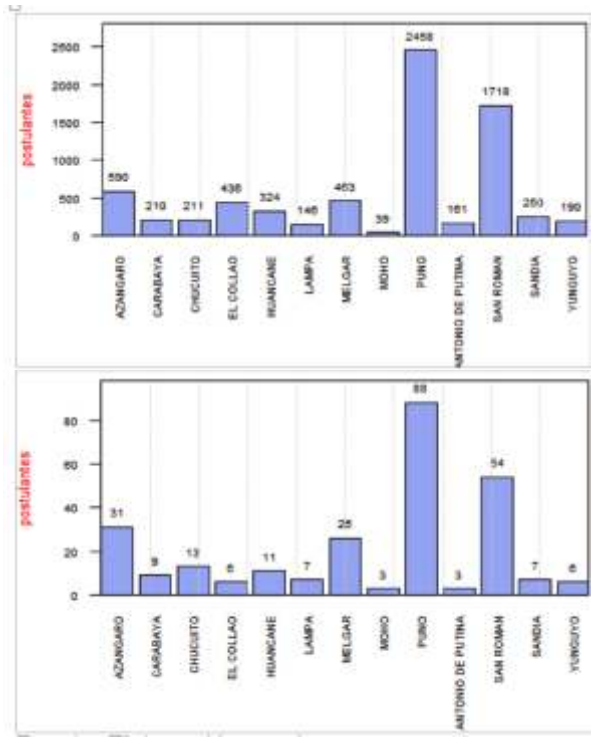
```
> total.post
Source: local data frame [13 x 5]
Groups: depart [1]

   depart                 prov general cepreuna extraord
    <chr>                <chr>   <int>    <int>    <int>
1   PUNO             AZANGARO    1240      590       31
2   PUNO            CARABAYA     398      210        9
3   PUNO            CHUCUITO     517      211       13
4   PUNO           EL COLLAO     704      436        6
5   PUNO           HUANCANE      669      324       11
6   PUNO              LAMPA      352      146        7
7   PUNO             MELGAR      949      463       26
8   PUNO               MOHO       75       39        3
9   PUNO              PUNO      4354     2458       88
10  PUNO SAN ANTONIO DE PUTINA    314      161        3
11  PUNO          SAN ROMAN     3118     1718       54
12  PUNO             SANDIA      456      250        7
13  PUNO           YUNGUYO       287      190        6
```

El resultado de la graficación de la información de los procesos de la variable *proc_general, proc_cepreuna y proc_extraordinario* se usó la función graf.post.dep.prov de la figura 4, con las siguientes invocaciones:

```
graf.post.dep.prov(1,1,"No","PUNO", 0, 250)
graf.post.dep.prov(2,1,"No","PUNO", 0, 150)
graf.post.dep.prov(3,1,"No","PUNO", 0, 10)
```

Figura 12. Procesos general, cepreuna y extraordinario. Cantidad de postulantes de las provincias del departamento de Puno

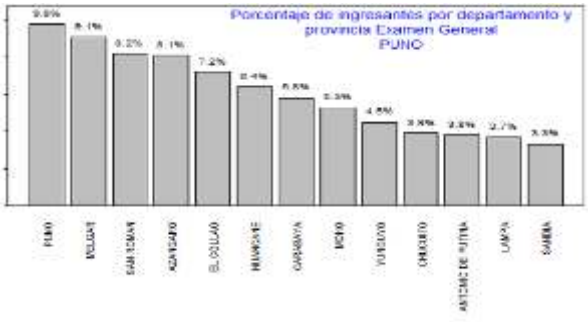


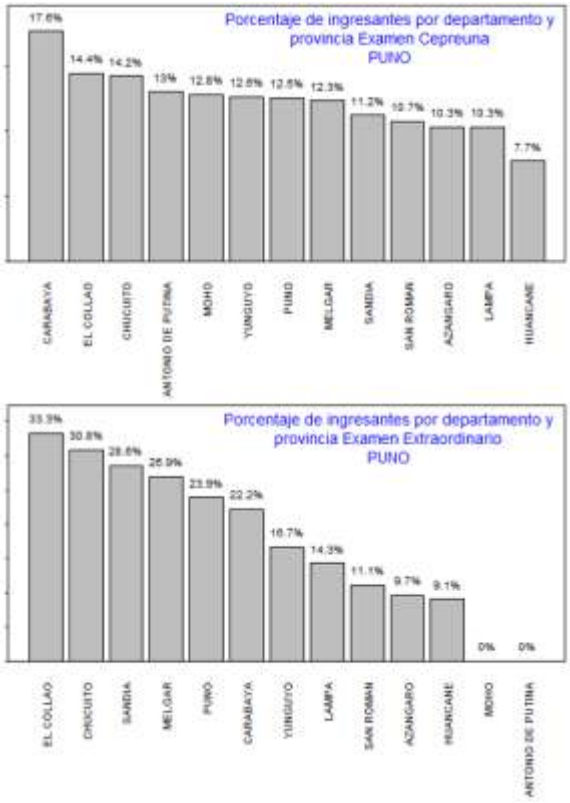


Fuente. Elaboración propia

Análisis de ingresantes por provincia del departamento de Puno: El resultado muestra que para los exámenes generales se tiene que el 9.8% proceden de Puno, el 9.1% proceden de Melgar, el 8.2% proceden de San Román, etc. Para los exámenes cepreuna se tiene que el 17.7% proceden de Carabaya, el 14.5% proceden del Collao, el 14.2% proceden de Chucuito, etc. y para los exámenes extraordinario se tiene que el 33.3% proceden de El Collao, el 30.8% proceden de Chucuito, el 28.6% proceden de Sandia, etc.; tal como se muestra en la siguiente figura.

Figura 13. Porcentaje de ingresantes por provincias. Procesos ordinario y extraordinario





Fuente. Elaboración propia

Para obtener esta información se usó el código de la figura 10, adicionando el siguiente filtro: filter(depart==”PUNO” & ingreso==”Si”).

Figura 14. Cantidad de ingresantes del departamento de Puno y provincias

```
> total.ing
Source: local data frame [13 x 5]
Groups: depart [1]

   depart                 prov general cepreuna extraord
    <chr>                <chr>   <int>    <int>    <dbl>
1   PUNO             AZANGARO     100       61        3
2   PUNO             CARABAYA      23       37        2
3   PUNO             CHUCUITO      20       30        4
4   PUNO           EL COLLAO       51       63        2
5   PUNO             HUANCANE      43       25        1
6   PUNO                LAMPA      13       15        1
7   PUNO              MELGAR       86       57        7
8   PUNO                MOHO        4        5        0
9   PUNO                PUNO      425      307       21
10  PUNO SAN ANTONIO DE PUTINA      12       21        0
11  PUNO           SAN ROMAN       255      184        6
12  PUNO              SANDIA        15       28        2
13  PUNO             YUNGUYO       13       24        1
```

Fuente. Elaboración propia

Análisis de postulantes e ingresantes por procedencia de colegio urbana y rural: Según (Ramos, Duque, y Nieto, 2012), el rendimiento educativo en zonas urbanas es más altos que el de los rurales, esto puede deberse a las características de la familia o de la escuela, es se corrobora con los siguientes resultados, el mayor porcentaje de ingresantes, en función de la

cantidad de postulantes la Escuela Profesional de Biología con un 27.52% y 19.48% en el área urbana y rural; Medicina, Veterinaria y Zootecnia un 22.87% y 15.53% en el área urbana y rural; Nutrición Humana un 18.48% y 8.09% en el área urbana y rural; Enfermería un 11.45% y 5.62% en el área urbana y rural; Odontología un 7.51% y 3.24% en el área urbana y rural; y Medicina Humana un 2.49% y 2.09% área urbana y rural.

Para determinar la cantidad de postulantes e ingresantes de los diferentes colegios del departamento de Puno por zona urbana – rural y por Escuela Profesional, se usó la función de la figura 5, *Data.grupo.colegio.area.* Se implementó el siguiente código cuyo parámetro *grupo* toma el valor de cero (0) lo que indica que se procesó todos los procesos de admisión, el último parámetro toma valores entre 1 y 6, que son los códigos de las Escuelas Profesionales.

Figura 15. Código que muestra la cantidad de postulantes e ingresantes por proceso y área

```
temp<-Data.grupo.colegio.area("PUNO","0",1,1)
temp1 <- temp %>% filter(ingreso %in% c("SI","No")) %>%
  group_by(orden, proceso, id_grupo, carrera, area) %>%
  summarise(total=n())
temp2 <- temp %>% filter(ingreso %in% c("SI")) %>%
  group_by(orden, proceso, id_grupo, carrera, area) %>%
  summarise(total=n())
temp3<-full_join(temp1,temp2, by=c("proceso","area"))
temp3<-tbl_df(temp3)
temp3<-dplyr::select(temp3, -id_grupo.x, -id_grupo.y,
                     -carrera.y, -orden.x, -orden.y)
temp3<-dplyr::rename(temp3, post=total.x, ing=total.y,
                     carrera = carrera.x)
temp3[is.na(temp3)] <- 0
```

Fuente. Elaboración propia

El resultado muestra las 10 primeras observaciones de la Escuela Profesional de Medicina, Veterinaria y Zootecnia. La primera columna contiene las fechas de los procesos de admisión cuyo primer carácter indica si es general (g), cepreuna (c) o extraordinario (e).

Figura 17. Cantidad de postulantes e ingresantes por proceso y área

```
> data.table(head(temp,10))
          proceso          carrera   area post ing
 1: e_17_03_2013 Med. Vet. y Zoot. URBANO    3   2
 2: c_31_03_2013 Med. Vet. y Zoot.  RURAL    2   1
 3: c_31_03_2013 Med. Vet. y Zoot. URBANO   29   9
 4: g_07_04_2013 Med. Vet. y Zoot.  RURAL    8   2
 5: g_07_04_2013 Med. Vet. y Zoot. URBANO  105  20
 6: c_14_07_2013 Med. Vet. y Zoot.  RURAL    3   1
 7: c_14_07_2013 Med. Vet. y Zoot. URBANO   21  12
 8: g_18_08_2013 Med. Vet. y Zoot.  RURAL   14   0
 9: g_18_08_2013 Med. Vet. y Zoot. URBANO  120   8
10: c_08_09_2013 Med. Vet. y Zoot.  RURAL    3   3
```

```
areas<-temp
head(data.table(areas), 10)
resumen1<-data.table(
  areas %>% dplyr::select(carrera, area, post) %>%
    group_by(carrera, area) %>%
    summarise(totalPost=sum(post))
)
head(resumen1,30)
resumen2<-data.table(
  areas %>% dplyr::select(carrera, area, ing) %>%
    group_by(carrera, area) %>%
    summarise(totalIng=sum(ing))
)
head(resumen2,30)
res<-cbind(resumen1, resumen2)
res[,c(1:2,3,6)]
```

Fuente. Elaboración propia

Figura 19. Resumen de postulantes e ingresantes por colegios rurales y urbanos – otras escuelas.

```
> resultado
           carrera   area totalPost totalIng porcIng
1         Biologia URBANO      1348      371  27.52%
2  Med. Vet. y Zoot. URBANO      1399      320  22.87%
3         Biologia  RURAL        77       15  19.48%
4   Nutrición Humana URBANO      1837      339  18.45%
5  Med. Vet. y Zoot.  RURAL       161       25  15.53%
6        Enfermeria URBANO      4400      504  11.45%
7   Nutrición Humana  RURAL       173       14   8.09%
8       Odontologia URBANO      2584      194   7.51%
9        Enfermeria  RURAL       534       30   5.62%
10      Odontologia  RURAL       122        4   3.28%
11   Medicina Humana URBANO      8449      210   2.49%
12   Medicina Humana  RURAL       287        6   2.09%
```

Fuente. Elaboración propia

En el siguiente gráfico se usa el paquete ggplot para representar la información de los postulantes e ingresantes por colegio de procedencia del área rural y urbano.

Figura 20. Código que genera el gráfico de post. e ing. de colegios urbanos y rurales

```
ggplot(resultado, aes(x=carrera, y=totalPost)) + # dentro d
  geom_point(aes(colour = factor(area)), size=2) +
  geom_text(aes(label=totalPost),hjust=1.3, vjust=0.4,
            size=2.5) +
  labs(x = "", y = "Postulantes", colour = "AREA")+
  ggtitle("Cantidad de postulantes del área biomédicas por
           procedencia de colegio\nRural - Urbano")+
  theme(plot.title = element_text(hjust = 0.5, size=8,
                                  face = "bold"),
        legend.title=element_text(size=7),
        legend.text=element_text(size=7),
        axis.text.x = element_text(angle=0,
                                   hjust=0.5, size = 8),
        axis.text.y = element_text(size = 7)) +
  geom_line(data=resultado)
```

El siguiente gráfico muestra el poco interés de los postulantes de colegios rurales de estudiar la Escuela Profesional de Medicina, Veterinaria y Zootecnía, cuando debería ser todo lo contrario, esto se debe a que las escuelas rurales no brindan los conocimientos necesarios a los escolares en su proyecto de vida y su sociedad rural (Salazar, 2017); las cifras indican que sólo se obtuvo 25 ingresantes de colegios del área rural en 32 procesos de admisión.

Figura 21. Cant. de postulantes por procedencia de colegio – rural y urbano

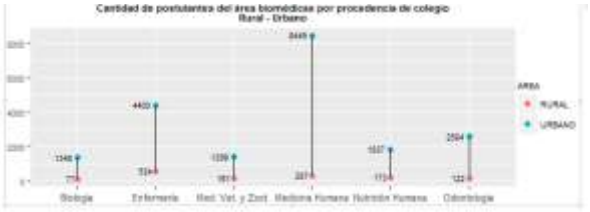


Figura 22. Cant. de ingresantes por procedencia de colegio – rural y urbano

CONCLUSIONES

La efectividad del uso de los paquetes RMySQL, dplyr, ggplot y el propio R ofrece un gran aporte al análisis de datos llevando una clara ventaja frente a los programas de uso comercial, como SAS, SPSS y Excel; sin embargo, existe una limitante para aquellos profesionales que no tienen conocimientos sobre lenguajes de programación, esto dificultaría la preferencia por el uso de R y buscando alguna otra herramienta de análisis.

El mayor porcentaje de ingresantes en función a la cantidad de postulantes de los procesos general, cepreuna y extraordinario lo tienen las provincias de: Puno, Carabaya y El Collao, respectivamente; y un bajo índice de ingresantes las provincias de: Sandia, Huáncane y San Antonio de Putina y Moho, respectivamente.

Para el cálculo del porcentaje de ingresantes provenientes de colegios urbanos se ha considerado los diferentes tipos de colegios ya sea estatal o particular, donde la Escuela Profesional de Biología tiene un mayor índice de ingresantes, y la Escuela Profesional de Medicina tiene casi el mismo porcentaje de ingresantes del área urbana y rural.

REFERENCIAS BIBLIOGRÁFICAS

Anchía, R. J. (2010). Aportaciones del software libre R al proceso de investigación psicológica. MISCELANEA COMILLAS, 165-175.

Duque, D. C., Salazar, J. A., Giraldo, L. A., Castro, V. G., y Olivera, A. P. (2012). Prevalecencia de intereses y preferencias profesionales en estudiantes de grado 11 de instituciones educativas públicas de la ciudad de Ibague. Dialnet, 11.

Flores Sánchez, M. (2013). Desarrollo de una aplicación para gráficos de control de procesos industriales. Universidad de la Coruña, España.

Jeroen Ooms, D. J. (26 de agosto de 2016). Database Interface and 'MySQL' Driver for R. Obtenido de https://cran.r-project.org/web/packages/RMySQL/index.html

Kadlec, J., B. S., D. A., y R. G. (2015). Ecological Informatics. ScienceDirect, 19-28.

Mayelín Mirabal Sosa, M. R. (2010). R: una herramienta poco difundida y muy útil para la investigación clínica. Scielo - Revista Cubana de Investigaciones Biomédicas, 4. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002010000200012

Mirabal Sosa, M. (30 de abril de 2010). R: una herramienta poco difundida y muy útil para la investigación clínica. Obtenido de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002010000200012

Piattini Velthuis, M. G., Martines, E. M., Muñoz, C. C., y Sánchez, B. V. (2006). Tecnología t diseño de base de datos. España: RA-MA EDITORIAL.

Ramos, R., Duque, J., y Nieto, S. (2012). Un análisis de las diferencias rurales y urbanas en el rendimiento educativo de los estudiantes colombianos a partir de los microdatos de Pisa. International Conference on Regional Science, 26.

Rexel, K. P., Gearan, P., y Heather, A. (12 de marzo de 2015). Data Science Survey. Obtenido de http://www.rexeranalytics.com/assets/rexer_analytics_2015_data_miner_survey_summary_report.pdfRSudio. (20 de enero de 2017).

RStudio. Obtenido de https://www.rstudio.com

Salazar, R. A. (2 de mayo de 2017). La educación rural un reto educativo. Obtenido de http://www.docentes.unal.edu.co/lgonzalezg/docs/LaEducacionRuralunRetoEducativo.pdf

Team, R. C. (8 de febrero de 2008). R: A Language and Environment for Statistical Computing.

Tecnológica, G. (08 de julio de 2014). Strata+Hadoop. Obtenido de http://www.gacetatecnologica.com/teradata-amplia-la-capacidad-analitica-del-software-libre-r/

Velasquez, J. D., L. V., y C. Z. (2011). ARNN: Un paquete para la predicción de series de tiempo usando redes neuronales autoregresivas. ResearchGate, 177-181.

Velasquez, J., Montoya, O., y Cataño, N. (2010). ¿Es el proyecto R para la computación estadística apropiado para la inteligencia computacional? Ingeniería y Competitividad, 81-94.

Vilaseca, B. (21 de junio de 2009). La conquista de la vocación profesional. Obtenido de http://borjavilaseca.com/la-conquista-de-la-vocacion-profesional/

Wickham, H. (24 de junio de 2016). A Grammar of Data Manipulation. Obtenido de https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

Wickham, H., y Francois, T. (13 de febrero de 2017). Package dplyr. Obtenido de https://cran.r-project.org/web/packages/dplyr/dplyr.pdf