Análisis descriptivo de los equipos participaron en el Torneo Argentino de Programación en 2015.

Probabilidad y Estadística

Licenciatura en Ciencias de la Computación

Universidad Nacional de Rosario
Facultad de Ciencias Exactas, Ingeniería y Agrimensura

El objetivo de este material es presentar a los alumnos una aplicación de técnicas descriptivas a datos reales. En este ejemplo, a través de la utilización de una serie de medidas resúmenes, gráficos y tablas, se caracteriza a un conjunto de unidades, intentando descubrir regularidades y singularidades de los mismos. Para obtener los resultados se utiliza el paquete estadístico R.

1. Comentarios sobre R

Características principales:

- Libre y gratuito. Se puede descargar en www.r-project.org.
- Es un conjunto integrado por paquetes, cada paquete se refiere a uno o más temas de la estadística, a su vez cada uno de esos paquetes está formado por sentencias que permiten implementar las distintas técnicas estadísticas. Los paquetes básicos están incluidos en R, otros están disponibles a través de Internet en CRAN (buscar en la página del ítem anterior)
- Es un lenguaje orientado a "objetos".

Entre otras propiedades dispone de:

• almacenamiento y manipulación efectiva de datos,

• operadores para cálculo sobre variables indexadas (Arrays), en particular matrices,

• una amplia, coherente e integrada colección de herramientas para análisis de datos,

• posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y

• un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas.

• distingue entre mayúscula y minúsculas. Es decir, el objeto "DATOS" es distinto al objeto "Datos".

2. Acerca de los datos

El conjunto de datos que se presenta en la Tabla 5 del Anexo se refiere a los resultados obtenidos por los equipos de Universidades Argentinas que participaron del Torneo Argentino de Programación (TAP). Sólo se consideran los 50 equipos que obtuvieron mayor puntaje. La puntuación se mide en cantidad de problemas resueltos, y en caso de coincidir, se desempata por menor cantidad de puntos de penalidad. Para cada equipo se registra la sede (ciudad donde está su universidad), el score (cantidad de ejercicios resueltos

correctamente de 11) y la penalidad (suma de los segundos que se exceden en cada ejercicio del tiempo reglamentado para cada uno de ellos).

En resumen,

Unidad de análisis: cada equipo

Muestra: los 50 equipos que obtuvieron mayor puntaje en el TAP 2015. (tamaño de la muestra n=50)

Variables:

Sede: ciudad donde está su universidad (cualitativa)

Score: cantidad de ejercicios resueltos correctamente (cuantitativa discreta)

Penalidad: suma de los segundos que se exceden en cada ejercicio del tiempo reglamentado para cada uno de ellos (cuantitativa continua)

Objetivo: resumir la información para conocer el lugar de procedencia de los participantes, el número de ejercicios resueltos correctamente y el tiempo que les lleva resolverlos a los mejores equipos que se presentaron en la competencia.

3. Análisis descriptivo

El análisis descriptivo se realiza utilizando tablas, gráficos y medidas resumen oportunas para cada variable considerada teniendo en cuenta si es cualitativa (nominal o ordinal) o cuantitativa (discreta o continua). Primero, se estudia cada variable por separado. En una segunda parte, se analizan las relaciones que se crean convenientes. Para facilitar el trabajo se usa el paquete estadístico R. Para cada resultado presentado se muestra en rojo la sentencia que se escribe en R para obtenerlo. Tener en cuenta que R no tiene una única forma de pedir los resultados, aquí sólo se muestra una opción.

Cuando se cuenta con datos se aconseja que su disposición sea la implementada en la Tabla 1 del Anexo, las filas se refieren a cada unidad que se estudia y las columnas a cada variable que se registra.

La siguiente sentencia permite importar a R una base de datos que se encuentra en un archivo .txt.

```
datos<-read.table("C:\\Users\\Mara\\Desktop\\LCC\\TAP2015.txt",header=T,sep="")
```

importa un archivo txt en un “objeto” llamado datos

```
datos
```

muestra lo que hay en el “objeto” datos

```
attach(datos)
```

toma cada columna (cada variable) como un objeto vector

R permite importar y exportar archivos con distintos formatos.

3.1. Análisis univariado

Se comienza resumiendo la información de la ciudad donde está la universidad a la que pertenecen los equipos.

Tabla de distribución de frecuencias

```
Sede.ord<-ordered(Sede,levels=c("Bs_As","Cordoba","Rosario","La_Plata","Chilecito",
"Neuquen"))
frecabs<-table(Sede.ord)
frecrel<-prop.table(table(Sede))
porc<- frecrel*100
tabla<-cbind(frecabs,frecrel,porc)
```

Tabla 1: Equipos según la ciudad de origen

Sede	Frecuencia absoluta	Frecuencia relativa	Porcentaje
Buenos Aires	23	0.46	46
Córdoba	9	0.18	18
Rosario	6	0.12	12
La Plata	5	0.10	10
Resistencia	5	0.10	10
Chilecito	1	0.02	2
Neuquén	1	0.02	2
Total	50	1	100

Interpretación de una fila de la tabla:

De los 50 equipos con mayor Score, sólo 1 es de Chilecito.

Chilecito representa sólo el 2 % del total de equipos.

Los mejores equipos provienen de 7 ciudades distintas.

Gráfico de barras

barplot(table(Sede.ord),ylim=c(0,25),xlab="Sede",ylab="Frecuencia")

Gráfico 1: Equipos según la ciudad de origen

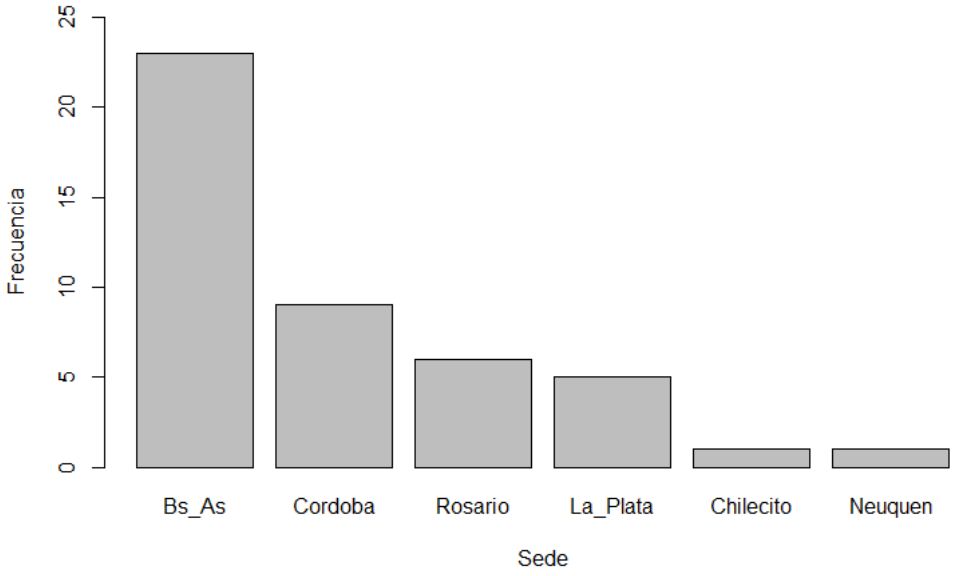
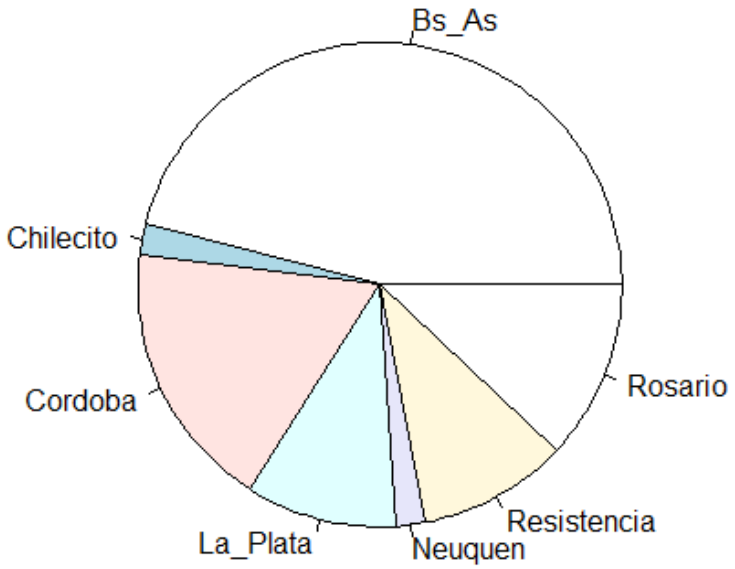


Gráfico de sectores

pie(table(Sede))

Gráfico 2: Equipos según la ciudad de origen



El Gráfico 1 muestra que la ciudad que más equipos aporta es Buenos Aires. En el Gráfico 2 se ve claramente que casi la mitad de los equipos son de Buenos Aires, también hay muchos equipos de Córdoba que se ubicaron entre los 50 mejores.

Medida de interés:

Moda=Buenos Aires

Buenos Aires es la ciudad que presenta el mayor número de equipos.

A continuación se resume la información sobre el número de ejercicios resueltos por cada equipo.

Tabla de distribución de frecuencias

```
frecabs<-table(Score)
frecrel<-prop.table(table(Score))
porc<- frecrel*100
frecabsacum<-cumsum(frecabs)
frecrelacum<-cumsum(frecrel)
porcacum<-cumsum(porc)
tabla<-cbind(frecabs,frecrel,porc,frecabsacum,frecrelacum,porcacum)
```

Tabla 2: Ejercicios resueltos por equipos

Score	Frecuencia absoluta	Frecuencia relativa	Porcentaje	Frec abs acumulada	Frec rel acumulada	Porc acumulado
2	14	0.28	28	14	0.28	28
3	19	0.38	38	33	0.66	66
4	7	0.14	14	40	0.80	80
5	1	0.02	2	41	0.82	82
6	4	0.08	8	45	0.90	90
7	3	0.06	6	48	0.96	96
8	1	0.02	2	49	0.98	98
9	1	0.02	2	50	1	100
Total	50	1	100			

Interpretación de la segunda fila de la tabla:

De los 50 equipos con mayor score, 19 pudieron resolver correctamente 3 ejercicios de 11.

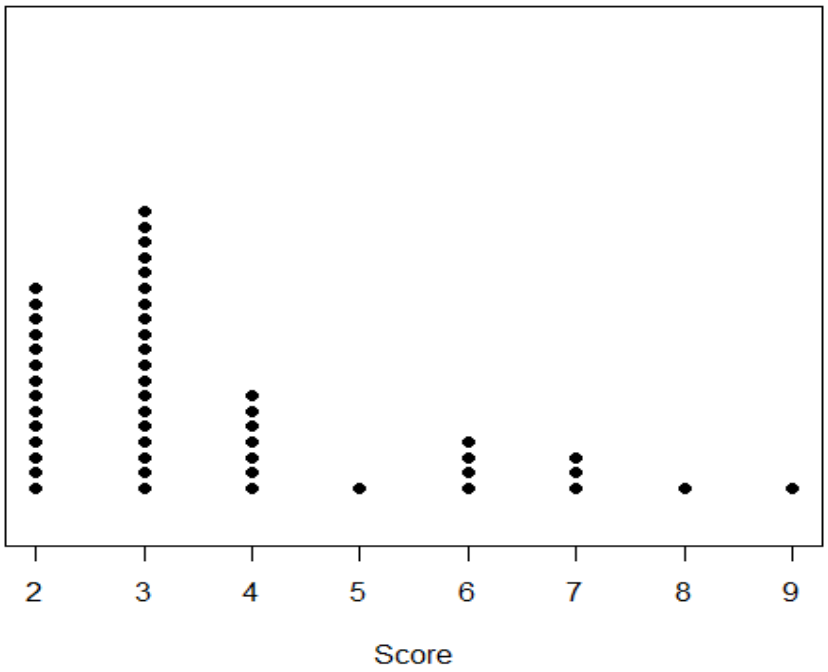
De los 50 equipos con mayor score, el 38% pudo resolver correctamente 3 ejercicios de 11.

De los 50 equipos con mayor score, el 66% pudo resolver correctamente a lo sumo (como máximo) 3 ejercicios.

Gráfico de puntos

```
stripchart(Score,method="stack",xlab="Score",offset=0.5,at=0.15,pch=19)
```

Gráfico 3: Ejercicios resueltos correctamente por equipos

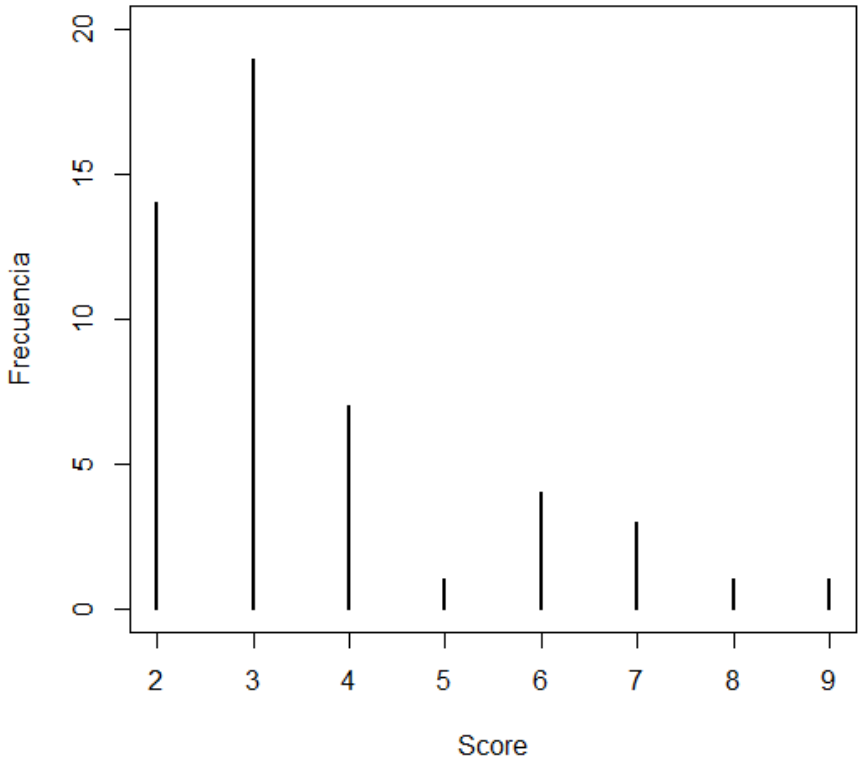


En el gráfico de puntos no hay eje de ordenadas debido a que cada punto representa un equipo.

Gráfico de bastones

```
plot(table(Score),ylim=c(0,20),space=20,xlab="Score",ylab="Frecuencia")
```

Gráfico 4: Ejercicios resueltos por equipos



Los Gráficos 3 y 4 muestran la cantidad de equipos que resolvió correctamente 1 problema, 2 problemas, etc. La mayoría de los equipos pudo resolver 2 o 3 ejercicios.

Medidas resúmenes de interés

summary(Score)

sd(Score)

IQR(Score)

numSummary((Score), statistics=c("cv"))

Medidas de tendencia central

Media=3,6

Mediana=3

Moda=3

Interpretación:

El promedio de ejercicios resueltos por equipo es 3,6 ejercicios correctos.

El 50% de los equipos resolvió 3 problemas o menos.

La cantidad de ejercicios resueltos con mayor frecuencia es 3.

Medidas de posición

Mínimo=2

Cuartil 1=2

Cuartil 3=4

Máximo=9

Interpretación:

Los 50 mejores equipos como mínimo resolvieron 2 ejercicios correctamente.

El 25% de los equipos resolvió 2 problemas o menos.

El 75% de los equipos resolvió 4 problemas o menos.

El número máximo de ejercicios que pudieron resolver fue 9.

Medidas de dispersión

Desvío estándar=1,76

Rango=7

Rango intercuatílico=2

Para tener una medida de variabilidad en relación con la media, que permita en un futuro comparar distintos años, se calcula el coeficiente de variación como

cv=1,76/3,6=0,49

Interpretación:

Considerando que la cantidad de ejercicios propuestos era 11 y la media 3,6 ejercicios, se puede pensar que el desvío estándar resultó grande sobre todo si se tiene en cuenta que se tomaron los 50 mejores equipos. El rango intercuatílico hace referencia a la variabilidad del 50% de las observaciones del centro (si los valores de Score están ordenados de mayor a menor). Si la mediana es 2 ejercicios, que el rango intercuatílico sea 2 implica una gran variabilidad de las observaciones del centro.

Por último se resume la información sobre el tiempo por penalidad que recibe cada equipo por haber excedido el tiempo para resolver cada ejercicio. Al ser ésta una variable continua con muchos valores distintos se decide crear intervalos. Se utiliza el resultado de la raíz cuadrada del número de equipos como guía para saber cuántos intervalos conviene. Además se consideran los valores mínimo y máximo de las observaciones.

Tabla de distribución de frecuencias

Cantidad de intervalos= $\sqrt{50} \approx 7$

Valor mínimo de penalidad=72

Valor máximo de penalidad=1135

Ayuda para crear los intervalos

```
Int1<-sum(Penalidad<=200)
Int2<-sum(Penalidad<=350&Penalidad>200)
Int3<-sum(Penalidad<=500&Penalidad>350)
Int4<-sum(Penalidad<=650&Penalidad>500)
Int5<-sum(Penalidad<=800&Penalidad>650)
Int6<-sum(Penalidad<=950&Penalidad>800)
Int7<-sum(Penalidad<=1100&Penalidad>950)
Int8<-sum(Penalidad>1100)
frecabs<-c(Int1,Int2,Int3,Int4,Int5,Int6,Int7,Int8)
```

Este procedimiento se puede simplificar

```
Intervalos<-cut(Penalidad,breaks<-c(50,200,350,500,650,800,950,1100,1250))
frecabs<-table(Intervalos)

frecrel<-prop.table(frecabs)
porc<- frecrel*100
```

```
frecabsacum<-cumsum(frecabs)
frecrelacum<-cumsum(frecrel)
porcacum<-cumsum(porc)
tabla<-cbind(frecabs,frecrel,porc,frecabsacum,frecrelacum,porcacum)
```

Tabla 3: Penalidad que reciben los equipos

Penalidad	Frecuencia absoluta	Frecuencia relativa	Porcentaje	Frec abs acumulada	Frec rel acumulada	Porc acumulado
(50,200]	15	0.30	30	15	0.30	30
(200,350]	13	0.26	26	28	0.56	56
(350,500]	8	0.16	16	36	0.72	72
(500,650]	7	0.14	14	43	0.86	86
(650,800]	3	0.06	6	46	0.92	92
(800,950]	2	0.04	4	48	0.96	96
(950,1100]	1	0.02	2	49	0.98	98
(1100,1250]	1	0.02	2	50	1	100
Total	50	1	100			

Interpretación de la segunda fila de la tabla:

De los 50 equipos con mayor score, 13 tuvieron más de 200 y 350 o menos segundos de penalidad.

De los 50 equipos con mayor score, 26% tuvo más de 200 y 350 o menos segundos de penalidad.

De los 50 equipos con mayor score, el 56% tuvo a lo sumo 350 segundos de penalidad.

Diagrama de tallo y hoja

```
stem(Penalidad)
```

Gráfico 5: Penalidad que reciben los equipos

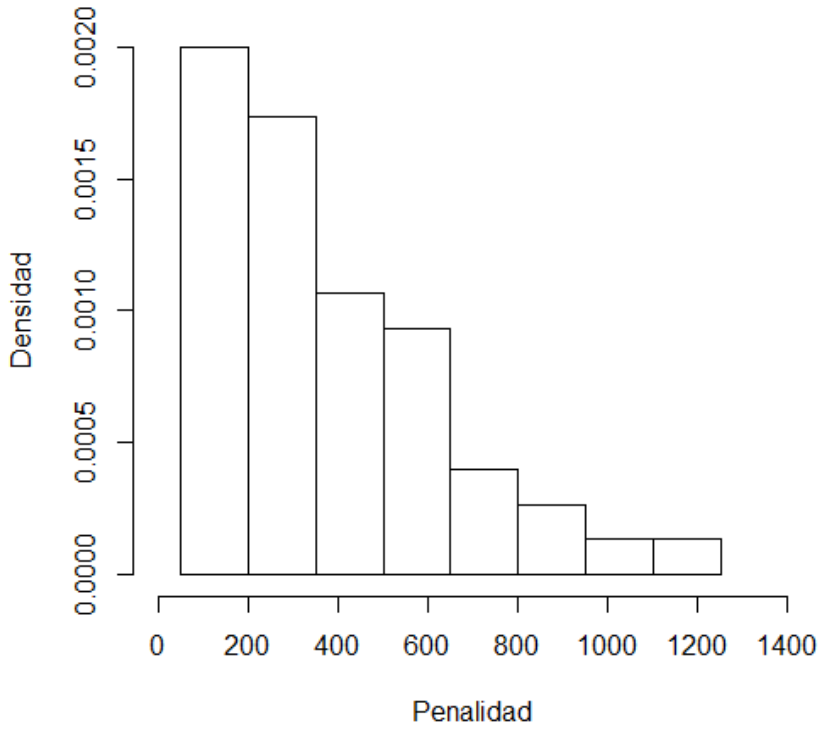
```
 0 | 789123456688999
 2 | 111335566801367
 4 | 336667056677
 6 | 501
 8 | 070
10 | 24
```

Nota: 0|7 significa aproximadamente 70 segundos de Penalidad

Histograma

```
hist(Penalidad,seq(50,1250,by=150),freq=FALSE,xlim= c(0,1400),xlab="Penalidad",
ylab="Densidad",main="")
```

Gráfico 6: Penalidad que reciben los equipos



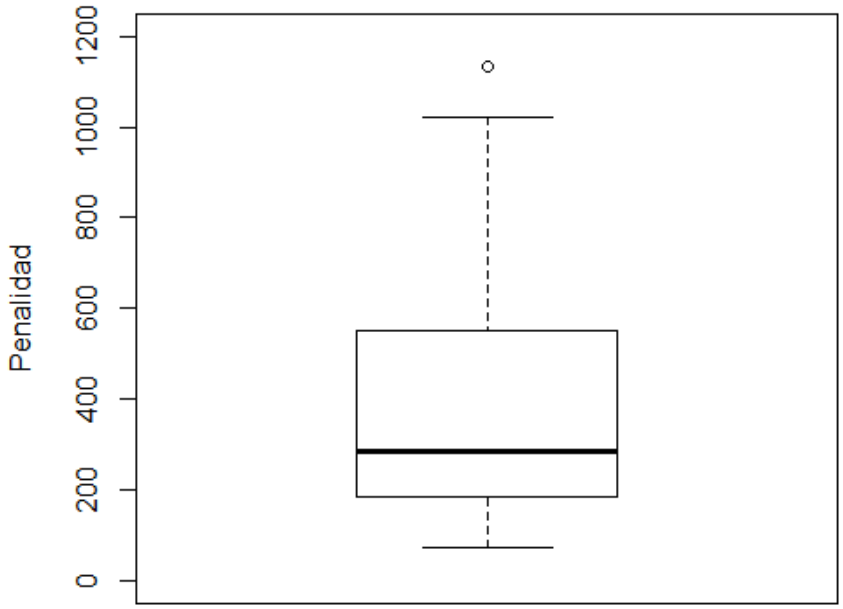
El histograma (Gráfico 6) claramente muestra una distribución asimétrica hacia la derecha. Es decir, hay muchos equipos que recibieron poca penalidad debido a que cumplieron con el tiempo asignado o producto de que no resolvieron muchos ejercicios. Hay dos equipos que tienen penalidad alta, son los que salieron en segundo y cuarto puesto.

Dado que la distribución es asimétrica lo apropiado es informar, como medida descriptiva, la mediana junto con el rango intercuatílico debido a que estas medidas no se ven demasiado influenciada por los valores más extremos.

Diagrama de caja

```
boxplot(Penalidad,ylim=c(0,1200),ylab="Penalidad")
```

Gráfico 7: Penalidad que reciben los equipos



El Boxplot (Gráfico 7) distingue un valor atípico, que pertenece al mayor valor de penalidad.

Este gráfico también muestra la asimetría de la distribución.

Análisis bivariado

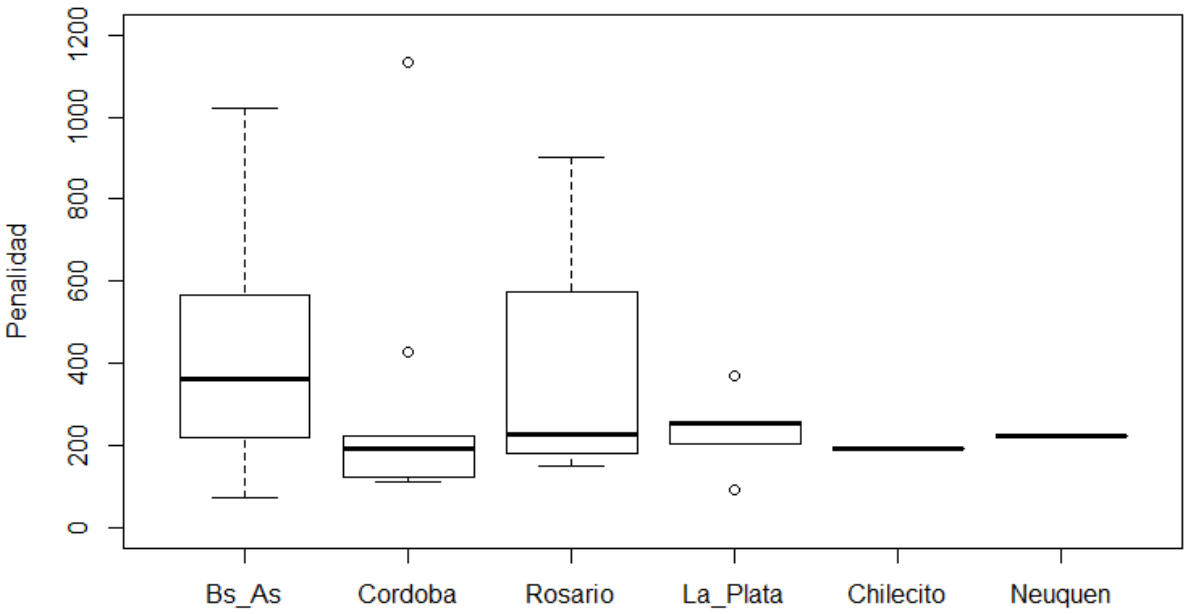
En esta parte del trabajo se pretende estudiar la penalidad obtenida en la competencia según la ciudad que representa el equipo.

Para comparar las ciudades se realiza un gráfico que muestre un boxplot para cada ciudad.

Diagrama de caja

```
boxplot(Penalidad~Sede.ord,ylim=c(0,1200),ylab="Penalidad")
```

Gráfico 8: Penalidad que reciben los equipos según la ciudad que representan



La mediana de la penalidad es mayor para la ciudad de Bs As. Esta ciudad presenta un rango intercuatílico grande, lo cual indica que los equipos varían bastante con respecto a la penalidad. Los equipos de Córdoba no recibieron demasiada penalidad pero hay uno que se destaca siendo el que mayor penalidad recibió. Los equipos de Rosario en comparación con el resto hay varios que recibieron mucha penalidad. De Chilecito y Neuquén sólo un equipo quedó entre los 50 mejores, ninguno de los dos se excedieron mucho del tiempo estipulado.

En algunas ciudades hay sólo un equipo, pero si el tamaño de muestra fuera mayor y cada ciudad estaría representada por varios equipos resulta de interés completar la Tabla 4.

Tabla 4: Penalidad que reciben los equipos según la ciudad que representan

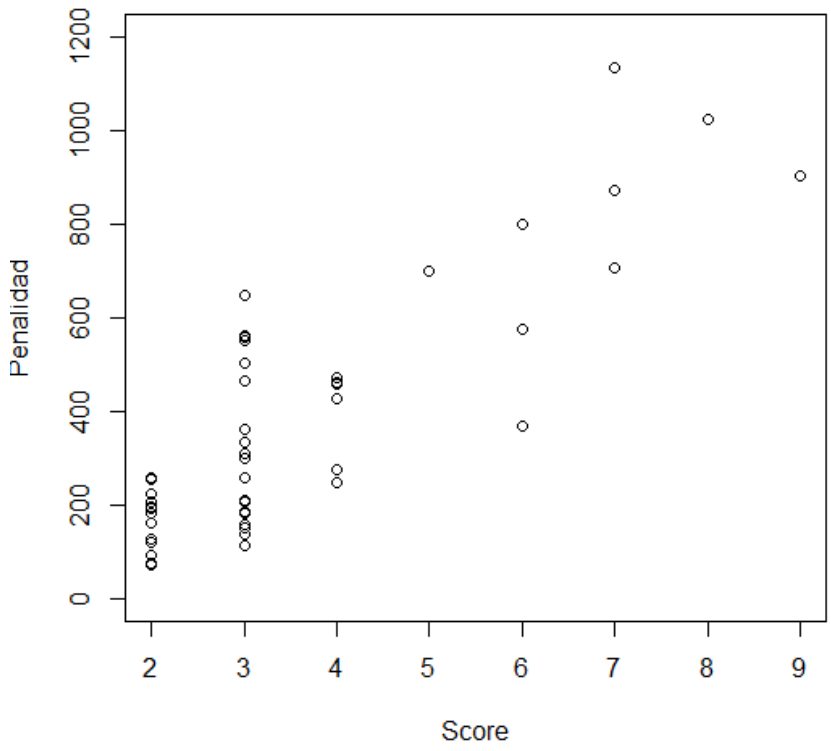
Medidas	**Bs As**	**Córdoba**	**Rosario**	**La Plata**	**Chilecito**	**Neuquén**
Media						
Mediana						
Moda						
Desvío Estándar						
Rango						
Rango intercuatílico						
Mínimo						
Máximo						

También interesa analizar la penalidad obtenida en relación al número de ejercicios resueltos en la competencia.
A continuación se presenta un gráfico de dispersión entre Penalidad y el Score.

Gráfico de dispersión

```
plot(Penalidad~Score,ylim=c(0,1200),ylab="Penalidad",xlab="Score")
```

Gráfico 9: Penalidad vs Score de los equipo



El Gráfico 9 indica que en general a medida que la cantidad de ejercicios resueltos correctamente es mayor, mayor es la penalidad recibida.

Anexo

La base de datos consiste en los resultados del Torneo Argentino de Programación (TAP), año 2015. Se estudiaron los equipos que obtuvieron dentro de los primeros 50 puestos. Toda la información sobre esta competencia se puede encontrar en la página: http://torneoprogramacion.com.ar/2015/09/27/resultados-tap-2015/

Tabla 5: Ranking de equipos del TAP. Año 2015.

id	Sede	Equipo	Universidad	Score	Penalidad
1	Rosario	Caloventor en Dos	U.N.R.	9	903
2	Bs As	Proyecto Mandioca	U.B.A. - FCEN	8	1023
3	Bs As	Pummas	U.B.A. - FCEN	7	708
4	Bs As	Pseudorandom	U.B.A. - FCEN	7	873
5	Córdoba	El nombre es lo de menos	U.N. de Córdoba - FaMAF	7	1135
6	La Plata	The Knights of the Hash Table	U.N. de La Plata	6	369
7	Rosario	I cannot stand this kappa	U.N.R.	6	574
8	Bs As	HelloWorld	Instituto Tecnológico de Bs As	6	574
9	Bs As	BBB	U.B.A. - FCEN	6	800
10	Bs As	Brulston	U.B.A. - FCEN	5	701
11	Rosario	Listas, Turings y Lambdas	U.N.R.	4	246
12	Bs As	Lukas compras cocas	U.B.A. - Facultad de Medicina	4	276
13	Córdoba	Cualquier Co$a	U.N. de Córdoba - FaMAF	4	428
14	Bs As	Salchebomba	U.B.A. - FCEN	4	428
15	Resistencia	time limit exceed	U.T.N.–Regional Resistencia	4	458
16	Resistencia	EstupidoySensualJAVA	U.T.N.– Regional Resistencia	4	463
17	Bs As	Catastrophic Cancellation	U.B.A. - FCEN	4	471
18	Córdoba	b1d0a0d7420d6d54	U.N. de Córdoba - FaMAF	3	112
19	Bs As	El Imperativo Contraataca	U.B.A. - FCEN	3	136
20	Rosario	Flower Power	U.N.R.	3	150
21	Bs As	Buen Guiso de Lentejas	U.B.A. - FCEN	3	159
22	Rosario	Uroboro	U. N. Litoral - FICH	3	182
23	Córdoba	P sii !P	U.N. de Córdoba - FaMAF	3	185

24	Córdoba	De Cara	U.N. de Córdoba - FaMAF	3	206
25	Rosario	Dormitrete	U.N.R.	3	208
26	Bs As	Team Deadlock	U.B.A. - FCEN	3	258
27	Bs As	Kuratowskis	U.B.A. - FCEN	3	299
28	Bs As	Trinidad y Estos Vagos	U.B.A. - FCEN	3	308
29	Bs As	Guerreros de Zorn	U.B.A. - FCEN	3	334
30	Bs As	/tmp	U.B.A. - FCEN	3	363
31	Bs As	Bit Force	U.B.A. - FCEN	3	464
32	Resistencia	mmmmh JAVA	U.T.N. – Regional Resistencia	3	504
33	Bs As	Incrustancia	Instituto Tecnológico de Bs As	3	553
34	Bs As	La esfera en la ingle	U.B.A. - FCEN	3	559
35	Resistencia	D.D	U.T.N. –Regional Resistencia	3	563
36	Resistencia	Mujeres cn Picaporte	U.T.N. –Regional Resistencia	3	648
37	Bs As	merengue	U.B.A. - FCEN	2	72
38	Bs As	30º and the Poppers	U.B.A. - FCEN	2	76
39	La Plata	WASD	U.N. de La Plata	2	92
40	Córdoba	Team While True	U.T.N.- Regional Córdoba	2	121
41	Córdoba	Z2	U.N. de Córdoba - FaMAF	2	125
42	Bs As	Wrong Answer	U.N. de La Matanza	2	162
43	Bs As	CODERS-UP 1	Universidad de Palermo	2	182
44	Córdoba	La Base Ordenada	U.N. de Córdoba - FaMAF	2	193
45	Chilecito	[UNdeC] ManPerQuin	U.N. de Chilecito	2	194
46	La Plata	MMA	U.N. de La Plata - Facultad de Informática	2	205
47	Córdoba	FamaFeroz	U.N. de Córdoba - FaMAF	2	225
48	Neuquén	Scatman	U.N. del Comahue - Facultad de Informática	2	225
49	La Plata	How do you turn this on?	U.T.N. - La Plata	2	253
50	La Plata	UnChifle	U.N de La Plata - Facultad de Informática	2	257