

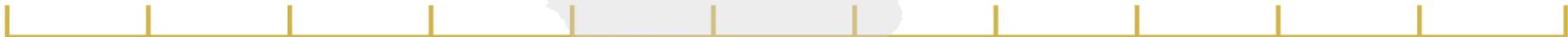
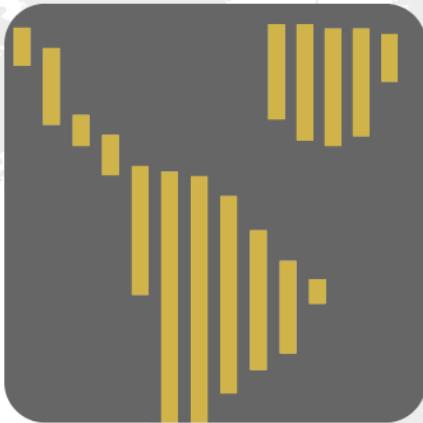


PROMiDAT
IBEROAMERICANO

Programa Iberoamericano de
Formación en Minería de Datos

Tarea №2

Análisis Exploratorio de Datos



Instrucciones:

1. Las tareas tienen fecha de entrega una semana después a la clase, y deben ser entregadas antes del inicio de la clase siguiente (6pm hora de Costa Rica).
2. **Cada día de atraso en implicará una pérdida de 10 puntos.**
3. Las tareas son estrictamente de carácter individual, tareas iguales se les asignará cero puntos.
4. En nombre del archivo debe tener el siguiente formato: **Tarea2_nombre_apellido.pdf**. Por ejemplo, si el nombre del estudiante es Luis Pérez: **Tarea2_luis_perez.pdf**. Para la tarea número 3 sería: **Tarea3_luis_perez.pdf**, y así sucesivamente.
5. El puntaje de cada pregunta se indica en su encabezado.
6. Esta tarea tiene un valor de un 25% respecto a la nota total del curso.

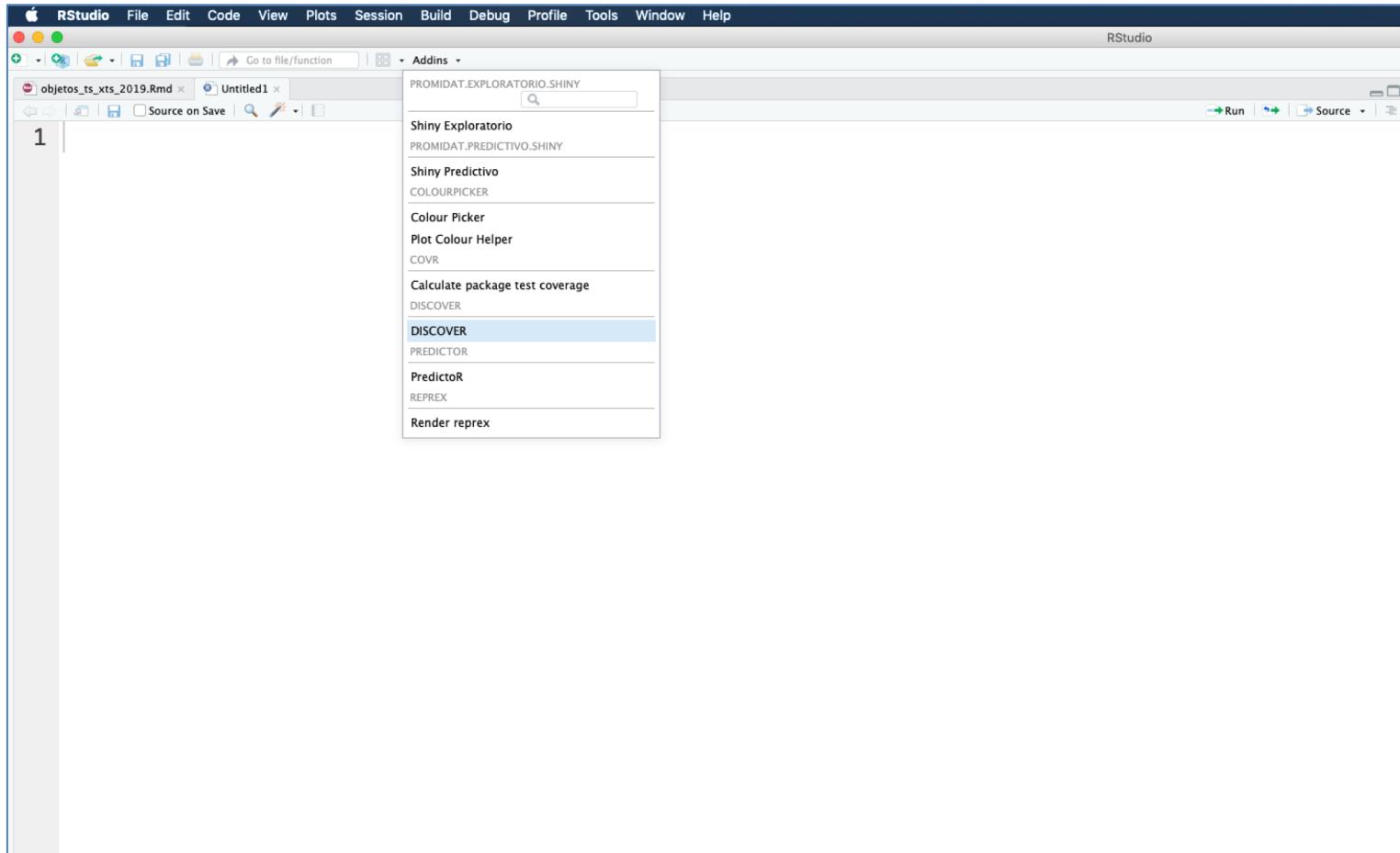
Ejercicio Número 1 (20 puntos)

En este ejercicio vamos a usar los datos “*Incendios_Forestales.csv*”. Estos datos corresponden a información de incendios forestales en regiones de Estados Unidos y Canadá, y de las condiciones imperantes a la hora del incidente. Las observaciones en la tabla son acontecimientos de incendio forestal, de los cuales se tomaron diversas mediciones para su posterior análisis. La idea es usar esta información para realizar un estudio no supervisado sobre los eventos de incendio forestal y detectar relaciones entre ellos. Las variables son:

- FFMC: Índice FFMC, que es una calificación del contenido de humedad de combustibles finos curados, equivalente al nivel de inflamabilidad del combustible.
- DMC: Índice DMC, que mide la humedad promedio de capas orgánicas de profundidad moderada y no tan compactas en los bosques.
- DC: Índice DC, que está relacionado con el nivel de humedad promedio de capas orgánicas profundas y compactas en el bosque.
- ISI: Índice ISI, que describe la velocidad de propagación del fuego a partir de la combinación del viento y el índice FFMC.
- Temperatura: Temperatura en grados Celsius.
- HR: Humedad relativa en el área del incendio.
- Viento: Velocidad del viento en _km/h_.
- Area: Área quemada del bosque en hectáreas.

Nota: Todas son variables numéricas y no tiene NAs.

2. Cargue el paquete *discoveR* en RStudio, como se muestra en la Figura:



3. Cargue el archivo “*Incendios_Forestales.csv*” usando el programa “Bloc de Notas” para ver cual es el separador de decimales y el separador de datos.
4. Desde “*discoveR*” cargue este archivo y verique que fueron bien leídos los tipos de las variables.

5. En *discoveR* calcule el resumen numérico para tres variables, interprete los resultados.
 6. Realice el test de normalidad para tres variables, es decir, determine el tipo de asimetría e interprete los resultados.
 7. Realice dos gráficos de dispersión e interprete 2 similitudes en cada gráfico.
 8. Para tres variables identifique los datos atípicos, si los hay.
 9. Calcule la matriz de correlaciones, incluya alguna de las imágenes que ofrece *discoveR* e interprete 3 de las correlaciones. Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.
 10. Desde el menú ACP gráfique el círculo de correlación (gráfico de variables), incluya la imágenes del círculo e interprete 3 de la correlaciones (las mismas 3 del ejercicio anterior). Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.
- ***ENTREGABLE: Los gráficos e interpretaciones dentro del Reporte PDF.***

Ejercicio Número 2 (20 puntos)

En este ejercicio vamos a usar los datos “**consumo_energetico.csv**”. Estos datos corresponden a mediciones de las características que ayudan con la eficiencia energética (carga calentamiento y carga refrigeracion) de un edificio.

La tabla de estudio está compuesta de 768 observaciones y de 8 variables. (0 categóricas y 8 numéricas). Las variables se explican seguidamente:

- Superficie: Superficie del edificio.
- Superficie_Pared: Superficie de paredes.
- Area_techo: Área del techo.
- Altura_Total: Altura total.
- Area_acristalamiento: Area de acristalamiento.
- Distri_superficie_acristalada: Distribución de la superficie acristalada.
- Carga_calentamiento: Carga de calentamiento. (Entre más pequeño mejor rendimiento)
- Carga_refrigeracion: Carga de refrigeración. (Entre más pequeño mejor rendimiento)

Nota: Todas son variables numéricas y no tienen NAs.

Usando ***discoveR*** repita el ejercicio 1 con esta tabla de datos (Recuerde que los separadores de espacios y decimales en CSV podrían ser diferentes).

Entregable: En el reporte PDF incluya los gráficos e interpretaciones.

Ejercicio Número 3 (20 puntos)

En este ejercicio vamos a usar los datos “*tenis_olimpiadas.csv*”. Es una tabla de datos que posee información de las jugadoras de tenis que han ido a las olimpiadas (1900 - 2016) y si durante su participación han logrado obtener o no una medalla. La tabla de estudio está compuesta de 487 observaciones y de 8 variables. (5 categóricas y 3 numéricas)

La siguiente tabla contiene la descripción, el tipo y el rango numérico de las variables.

- Nombre: El nombre de la atleta (Se puede repetir debido a que hay información de las olimpiadas desde 1900 - 2016)
- Edad: La edad de la atleta
- Altura: La altura de la atleta
- Peso: El peso de la atleta
- País: El país de la atleta
- Temporada: La temporada de la competencia (Verano)
- Deporte: El deporte de la muestra de datos (Tenis)
- Medalla: Si la atleta logró obtener una medalla de oro, plata o bronce.

Nota: Esta tabla contiene dos variables que no varían Temporada y Deporte, por lo que debe excluirse desde *discoveR*. Además debe excluir la variable Nombre que realmente no tiene sentido en el análisis.

1. Usando ***discoveR*** lea esta tabla de datos, recuerde que los separadores de espacios y decimales en CSV podrían ser diferentes, además verifique bien si existen o no nombres de filas y de variables. Verifique que queden bien los tipos de las variables (esta tabla incluye variables categóricas) .

2. En “discoveR” calcule el resumen numérico para tres variables, interprete los resultados. Incluya al menos una variable categórica.
3. Realice el test de normalidad para tres las variables, es decir, determine el tipo de asimetría e interprete los resultados.
4. Realice dos gráficos de dispersión e interprete 2 similitudes en cada gráfico.
5. Grafique la distribución de la(s) variable(s) categórica(s) e interprete.
6. ¿Existe algún dato atípico?
7. Calcule la matriz de correlaciones, incluya alguna de las imágenes que ofrece discoveR e interpréte 2 de las correlaciones. Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.
8. Desde el menú ACP gráfique el círculo de correlación (gráfico de variables), incluya la imagen del círculo e interprete 2 de las correlaciones (las mismas 3 del ejercicio anterior). Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.

- ***ENTREGABLE: Los gráficos e interpretaciones dentro del Reporte PDF.***

Ejercicio Número 4 (40 puntos)

NOTA: Genere un reporte aparte para este ejercicio

Considérese la siguiente tabla datos (que se muestra parcialmente en la imagen), la cual contiene variables que miden la probabilidad de sobrevivir en el hundimiento del Titanic (*titanic.csv*).

ID	Survived	Pclass	Sex	Age	SibSp	F
1	0	3	male	22	1	
2	1	1	female	38	1	
3	1	3	female	26	0	
4	1	1	female	35	1	
5	0	3	male	35	0	
7	0	1	male	54	0	
8	0	3	male	2	3	
9	1	3	female	27	0	
10	1	2	female	14	1	
11	1	3	female	4	1	

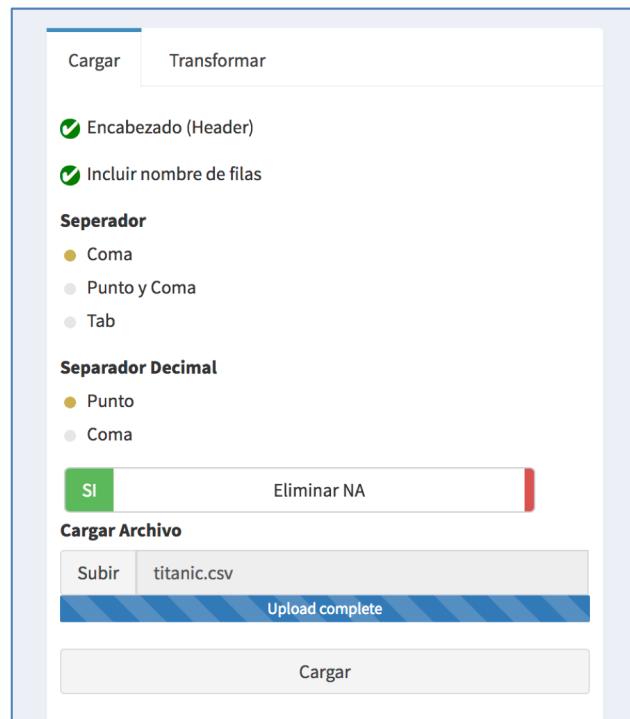
1. Usando **discoveR** lea esta tabla de datos, recuerde que los separadores de espacios y decimales en CSV podrían ser diferentes, además verifique bien si existen o no nombres de filas y de variables. Verifique que queden bien los tipos de las variables, esta tabla incluye variables categóricas en las que debe establecer el tipo adecuadamente.

Nota: *Este ejercicio es de mucho más cuidado porque tiene Id, variables que deben ser desactivadas y valores nulos NA en algunas filas, las columnas de esta tabla de datos se explican en la siguiente filmación:*

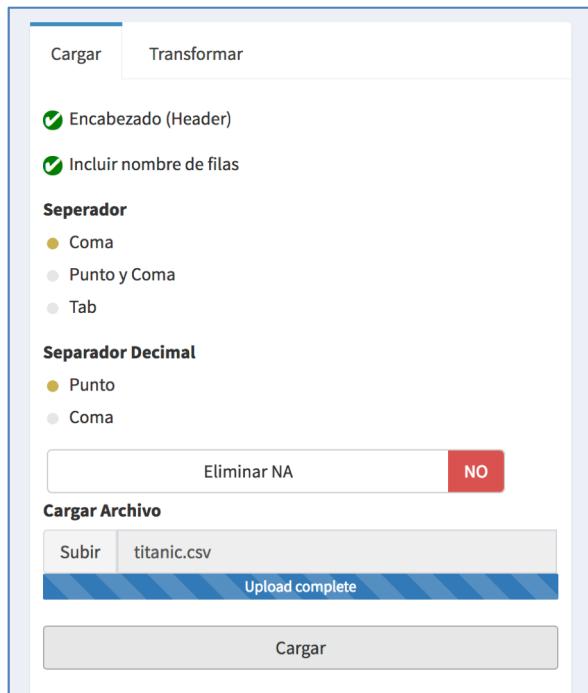
La tabla titanic.csv contiene los datos sobre la conocida historia y tragedia del Titanic, en cada fila está un pasajero y se trata de predecir la supervivencia o no de estos pasajeros. La tabla contiene 12 columnas y 1309 observaciones, las variables son:

- **PassegerId:** El código de identificación del pasajero (valor único).
- **Survived:** Variable a predecir, 1 (el pasajero sobrevivió) 0 (el pasajero no sobrevivió).
- **Pclass:** En que clase viajaba el pasajero (1 = primera, 2 = segunda , 3 = tercera).
- **Name:** Nombre del pasajero (valor único).
- **Sex:** Sexo del pasajero.
- **Age:** Edad del pasajero.
- **SibSp:** Cantidad de hermanos o cónyuges a bordo del Titanic.
- **Parch:** Cantidad de padres o hijos a bordo del Titanic.
- **Ticket:** Número de tiquete (valor único).
- **Fare:** Tarifa del pasajero.
- **Cabin:** Número de cabina (valor único).
- **Embarked:** Puerto donde embarcó el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton).

2. Usando NotePad identifique al menos 5 valores NA en esta tabla de datos.
3. En **discoveR** cargue la tabla de datos titanic.csv dejando la opción Eliminar NA activa como se muestra a continuación.



4. En **discoveR** cargue de nuevo la tabla de datos titanic.csv desactivando la opción Eliminar NA activa como se muestra a continuación (no olvide dar click nuevamente en el botón Cargar).
5. ¿Existe algún cambio en la lectura de Datos? Observe, por ejemplo, la filas 6. Explique que sucedió.



6. En *discoveR* cargue de nuevo la tabla de datos ***titanic.csv*** dejando la opción Eliminar NA activa.
7. Realice la transformación de variables necesaria para que la tabla incluya solamente las columnas que realmente son variables.
8. En *discoveR* calcule el resumen numérico para tres variables, interprete los resultados. Incluya al menos una variable categórica.
9. Realice el test de normalidad para tres de las variables, es decir, determine el tipo de asimetría e interprete los resultados.
10. Realice dos gráficos de dispersión e interprete 2 similitudes en cada gráfico.
11. Grafique la distribución para dos de las variables categóricas e interprete.
12. ¿Existe algún dato atípico?
13. Calcule la matriz de correlaciones, incluya alguna de las imágenes que ofrece *discoveR* e interprete 3 de las correlaciones. Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.
14. Desde el menú ACP gráfique el círculo de correlación (gráfico de variables), incluya la imágenes del círculo e interprete 3 de la correlaciones (las mismas 3 del ejercicio anterior). Debe ser una interpretación dirigida a una persona que no sabe nada de estadística.
 - ***ENTREGABLE: Respuestas a preguntas, los gráficos e interpretaciones dentro del Reporte PDF.***



Programa Iberoamericano de
Formación en Minería de Datos

Gracias ...