# Analisis exploratorio de datos

# Condiciones del laboratorio

El laboratorio se realizará por parejas, generando los siguientes entregables:

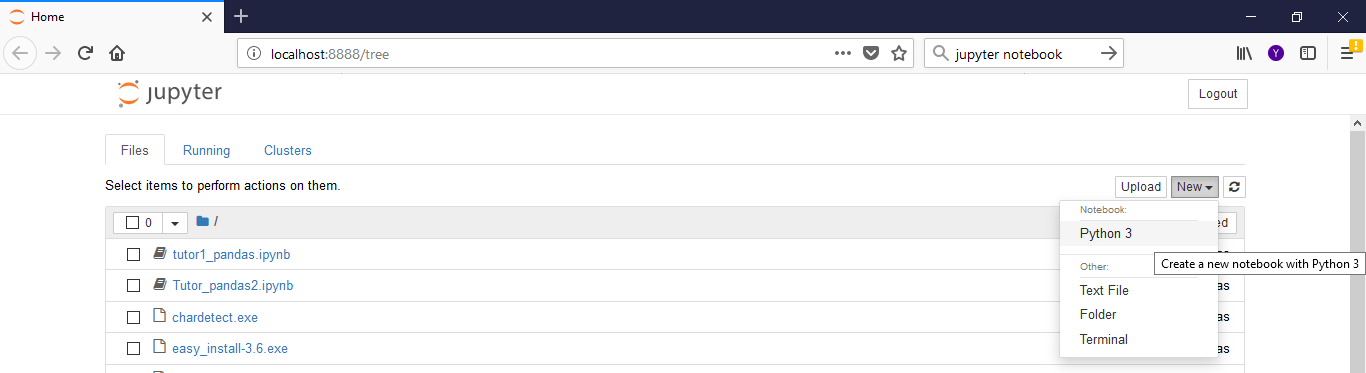
1. Documentación de los resultados obtenidos tras las ejecuciones , asi como su observación de análisis sobre los mismos y los conceptos asociados en la clase teórica.
2. Archivo “.ipynb” del notebook de Jupyter o archivo “.py” de Python preparado para la ejecución.
3. La entrega se realiza a través del Avata, con la fecha límite de entrega programada.

# Desarrollo

La metodología CRISP-DM tiene entre sus etapas el “Entendimiento de los datos”, una de sus principales técnicas es el Análisis exploratorio de datos (AED), este análisis usa métodos estadísticos para conocer las principales características, comportamiento y nivel de calidad de los datos.

Para esta guía se usará como fuente de análisis el archivo Clasificacion.txt, es un conjunto de datos de vinos italianos de tres diferentes clases, a cada uno de los cuales se les han medido trece características físico-químicas. Se busca construir un clasificador que con base en dichas características clasifique los vinos en una de las tres categorías. Para realizar una correcta clasificación se deben analizar y realizar el preprocesamiento necesario a las variables.

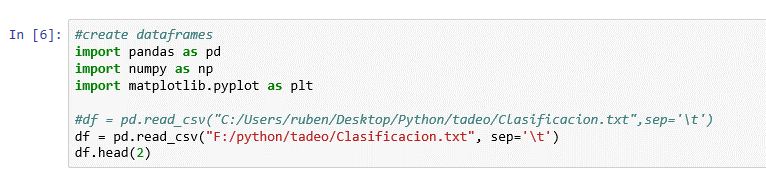
Jupyter usa entornos de trabajo como blocs de notas donde se puede agregar las secuencias de comando de Python y se ejecutan por bloques para ello se crea un cuaderno nuevo de Python. Ingrese a Jupyter notebook y seleccione la opción “New… Python3” en la parte superior derecha.



## Importar un archivo de datos

La primer tarea será leer el conjunto de datos de trabajo.

En el cuadro IN, coloque las instrucciones Python para importar los datos a un nuevo dataframe “df”. Tenga en cuenta la ruta de la carpeta en la cual se encuentra el archivo a importar.

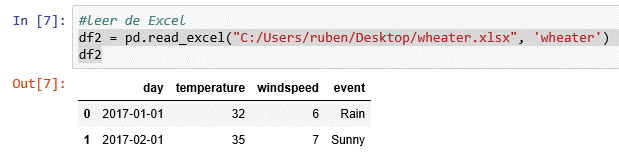


De igual manera, el archivo de datos se encuentra publicado en

Clasificacion.txt

* Tenga en cuenta que el parámetro **sep** ha sido establecido a ***'\t'*** para un archivo separado por tabuladores otras opciones pueden ser ***sep=";"*** o ***sep=","*** y la instrucción head(#) es para mostrar los n primeros registros del dataframe.

Para un archivo de Excel se puede leer con:



Consulte más sobre los comandos read\_csv y read\_excel en

<https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html>

<https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_excel.html>

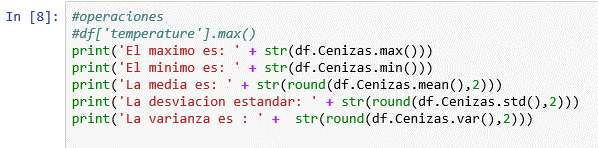
df.dtypes

df.shape

## Calculo de estadísticas descriptivas

Previo a conocer el contexto de los datos, de que variables se componen la data y que representan, el siguiente paso sugerido es calcular las estadísticas descriptivas del conjunto de las variables que componen los datos para revisar su frecuencia, distribución completitud y hacerse una idea de ellos.

Ejecute los siguientes comandos para calcular estadísticas básicas para el conjunto de datos



Luego pruebe con

df.mean()





|  |  |
| --- | --- |
| df.head(10) |  |
| len(df) |  |
| df.tail(5) |  |
| df.columns |  |
| df[“Alcohol”] |  |
| df[“Alcohol”][:5] |  |
| df[“Alcohol”][5] |  |
| df[[“Prolina”, “Alcohol”]][:5] |  |
| df[“Clase Vino”].value\_counts() |  |
|  |  |
|  |  |
|  |  |

Para la matriz de correlaciones:



O para la matriz de correlaciones



Mas de dataframes en:

<https://pandas.pydata.org/pandas-docs/stable/api.html#api-dataframe-stats>

## Medidas agrupadas

Para esta base, la clase de vino es una variable categórica que define las propiedades de 3 tipos de vino y está codificada con 1, 2 y 3.

Es buena idea también calcular los descriptivos por tipo de vino con el método gruopby

Puede hacerlo por cada variable numérica agrupando sus estadísticas



O una estadística para todas las variables



Y la matriz de correlaciones por tipo de vino



## Gráficos

Para una mejor compresión de los datos es esencial el análisis gráfico en el caso de variables continuas con histogramas, gráficos de dispersión y gráficos de cajas y bigotes (boxplot) y en el caso de variables discretas, diagramas de barras o gráficos circulares (pie).

### Histograma

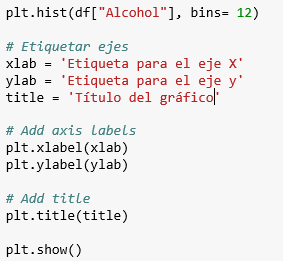
La siguiente instrucción genera el histograma para la variable Alcohol



Ahora verifique otras opciones cambiando el número de caterogias (bins) que se usan para agrupar los datos con



Y finalmente adiciónele algo de personalización agregando títulos a los ejes y al gráfico



### Boxplot

Realice para cada variable continua un boxplot para analizar medidas de tendencia y dispersión



Personalícelo con título, y nombre a los ejes

Vaya un paso más allá graficándolo para varias variables simultáneamente



### Diagrama de dispersión

Permite ver la relación simultánea de variables continuas



### Gráfico de barras

Se una para ver la dimensión de las variables nominales y las discretas o para comparar un parámetro como el máximo, mínimo, promedio u otra medida en variables continuas

* Compare el promedio para todas las variables continuas



* Compare el valor máximo para una variable entre las distintas clases de vino



* Compare la distribución de la clase de vino en el dataset



### Diagramas de pie

* Visualice la distribución del tipo de vino en forma de pie



Tablas de confusión

Resume el numero de ocurrencias de cada valor para una variable dada.

pd.crosstab(index=df["Clase Vino"],

columns="count") # Name the count column

Ahora para dos variables:

tcruzada = pd.crosstab(index=df["Clase Vino"],

columns=seguros["???"])

tcruzada.index= ["Clase 1","Clase 2"]

tcruzada

# Análisis, Comprensión y preprocesamiento de los Datos

## Parte 1

1. Haga una descripción de la metadata que contiene el archivo con base en la información anterior

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Tipo de dato** |
|  |  |  |

Basado en las salidas anteriores haga un análisis de los datos y genere un diagnóstico inicial de los mismos. Luego responda los siguientes interrogantes:

* Selecciones 5 variables y de acuerdo a los histogramas y las medidas vistas en clase (media, varianza, skew, kurtosis) indique si se trata de frecuencias unimodales, bimodales simétricas o asimétricas(desequilibradas), y su apreciación de la distribución de los datos de esa variable en el contexto vinos.
* Indique las variables categóricas ordinales y/o variables categóricas nominales que tenga el conjunto de datos

## Parte 2

Apartir de matrices de correlación y scatterplot indique su análisis (correlaciones fuertes principalmente)

De acuerdo con lo visto en clase anterior, ejecute las respectivas sentencias de Python para identificar ruido existente en el conjunto de datos

Indique que variables tienen valores faltantes

## Operaciones básicas de datos

### Discretizar variables (bining)

* Discretice 3 variables usando le método cut. Explique porqué discretizó estas variables.

Binning de igual tamaño



Bining por cuartiles

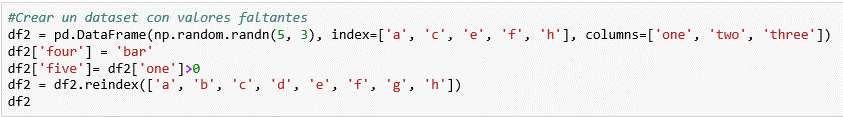


### Contar los missing

Puede con las funciones **isna** y **notna** identificar y contar los registros nulos en una fila. Si no aplica la función suma es decir solo con ***pd.isna(df2['one'])*** enlistara los vacios con del dataset para la columna indicada



Puede reemplazar estos datos con valores fijos, o calculados de las funciones propias del paquete panda como se ilustra a continuación:



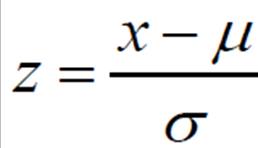
Valide el resultado de los siguientes comandos

* df2.fillna(0)
* df2['four'].fillna('missing')
* df2['four'].fillna('missing')
* dff.fillna(dff.mean())
* dff.fillna(dff.mean()['B':'C'])
* df.interpolate()

Consulta la guía en <http://pandas.pydata.org/pandas-docs/stable/missing_data.html#missing-data>

### Normalizar una variable

Normalizar una variable continua es redimensionarla en medida de desviaciones estándar conocidas como valores Z por la tabla de puntajes Z de la distribución normal, proceso que se hace con la siguiente formula



Que es esencia tomar el dato, restarle el promedio de la variable y dividirlo sobre su desviación estándar, e

Intente realizarlo con la siguiente serie de comandos

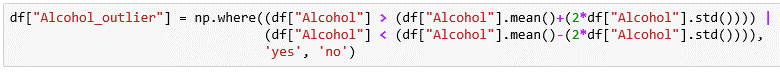


### Identificar valores atípicos

¿Existen variables con valores atípicos? Una regla para determinar valores atípicos es que sean menores o mayores a más o menos 2 desviaciones estándar, es decir

Estos deben ser identificados, analizados y discutir una técnica para reemplazarlos o descartarlos del dataset de análisis

Cree una variable que identifica los valores atípicos



Filtre los valores atípicos (outliers)



Ahora que ya se han valores atípicos, escriba una rutina para suavizar/corregir dichos valores, en una nueva columna, llamada Alcohol\_suav

## Parte final

Teniendo en cuenta los graficos generados y las medidas tomads al dataset Clasificación.csv , podría lanzar alguna hipótesis inicial que discrimine las distintas clase de vino?