**Introducción**

**Descripción del proyecto**

**Descripción de los datos**

Los datos fueron conseguidos de la página web kaggle.

* <https://www.kaggle.com/juliansimon/weather_madrid_lemd_1997_2015.csv>
* https://www.kaggle.com/grubenm/austin-weather

La unión de ambos conjuntos de datos después de hacer una conversión de medidas, da una cantidad de 8131 registros y 10 variables de información las cuales son:

1. TemperaturaMaxC
2. TemperaturaPromC
3. TemperaturaMinC
4. HumedadMax
5. HumedadProm
6. HumedadMin
7. VientoMaxKMXH
8. VientoPromKMXH
9. PrecipitacionMM
10. Eventos

**TemperaturaMaxC:** Describe la temperatura máxima en grados centígrados.

**TemperaturaPromC:** Describe la temperatura promedio en grados centígrados.

**TemperaturaMinC:** Describe la temperatura mínima en grados centígrados.

**HumedadMax:** Describe el porcentaje de humedad máxima.

**HumedadProm:** Describe el porcentaje de humedad promedio.

**HumedadMin:** Describe el porcentaje de humedad mínima.

**VientoMaxKMXH:** Describe la velocidad del viento máximo dado por kilómetros por hora.

**VientoPromKMXH:** Describe la velocidad del viento promedio dado por kilómetros por hora.

**PrecipitacionMM:** Describe la precipitación está dado por milímetros.

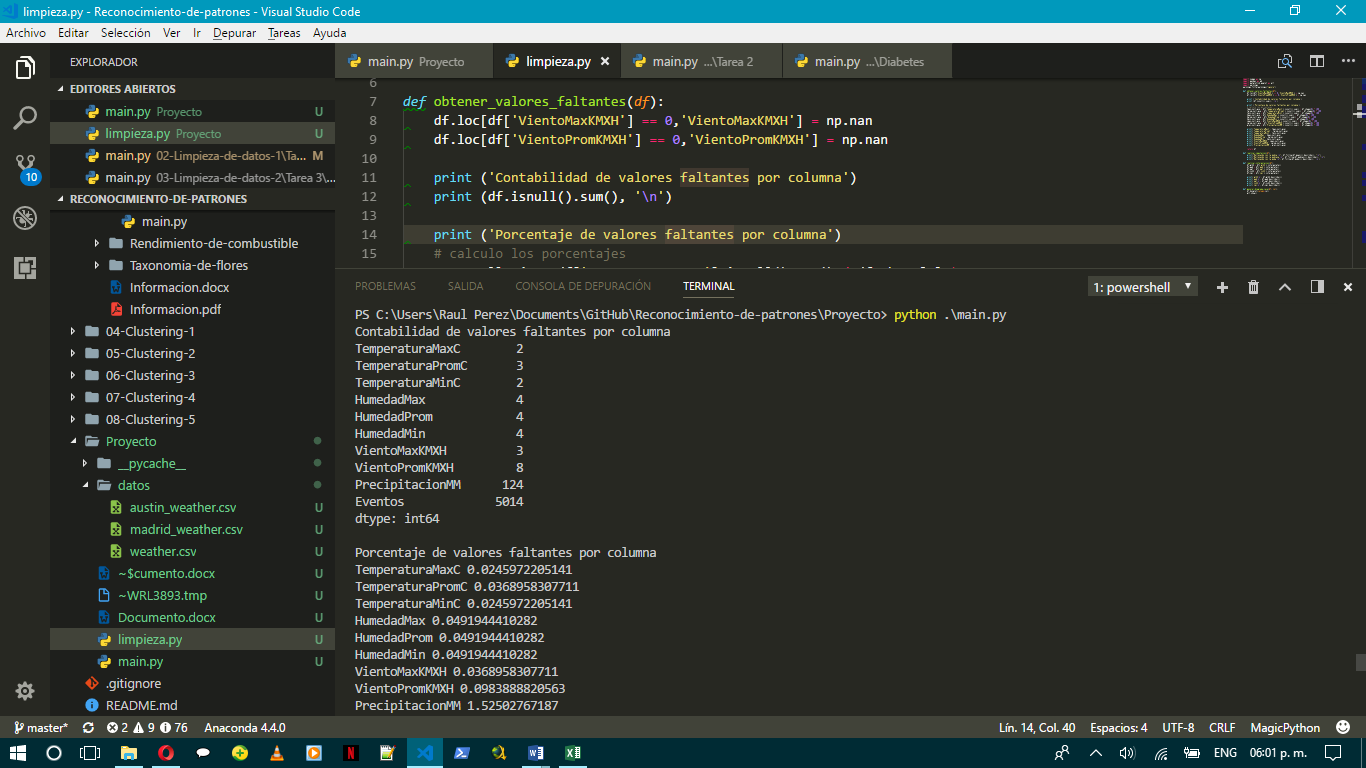
**Eventos:** Describe el evento o eventos sucedidos en el día, se representa con un valor entero. (Nada = 0, Neblina = 1, Nieve = 2, Lluvia = 3, Neblina-Lluvia = 4, Tornado = 5, Lluvia-Nieve = 6, Lluvia-Tormenta = 7, Neblina-Lluvia-Tormenta = 8, Lluvia-Granizo-Tormenta).

**Limpieza de los datos**

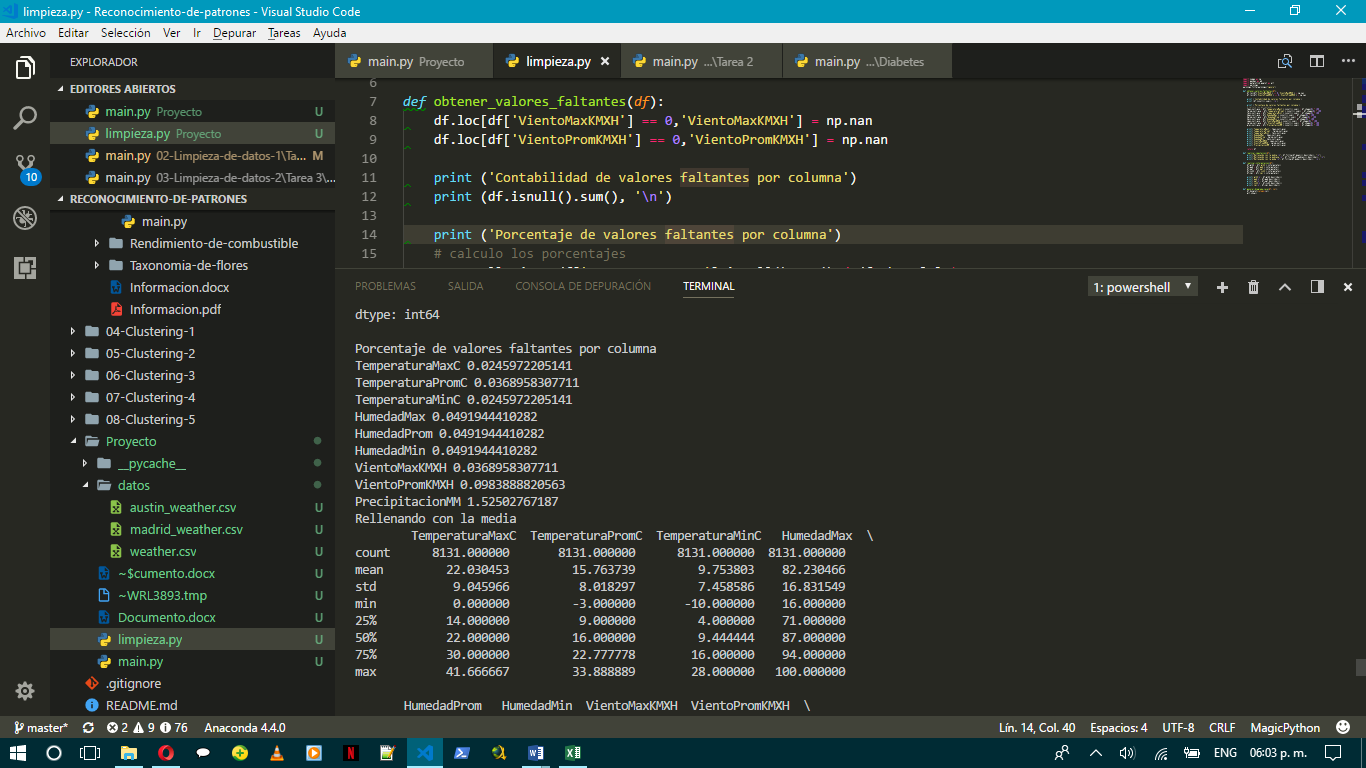
*El código relacionado a esta parte está contenido en el archivo* ***limpieza.py*** *dentro de la carpeta del proyecto.*

**Valores faltantes**

Después de analizar los datos se determinó la cantidad de valores faltantes para cada variable en el conjunto de datos, los cuáles son siguiente:



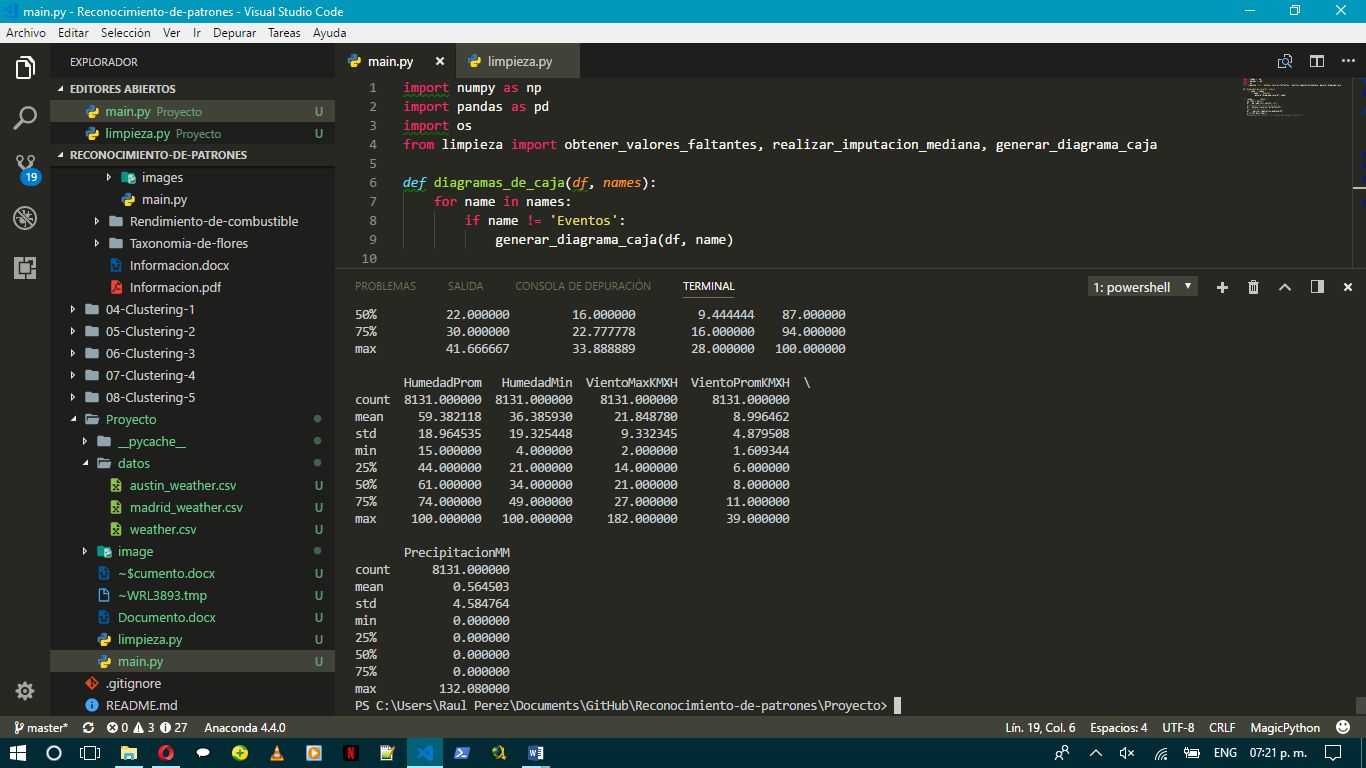
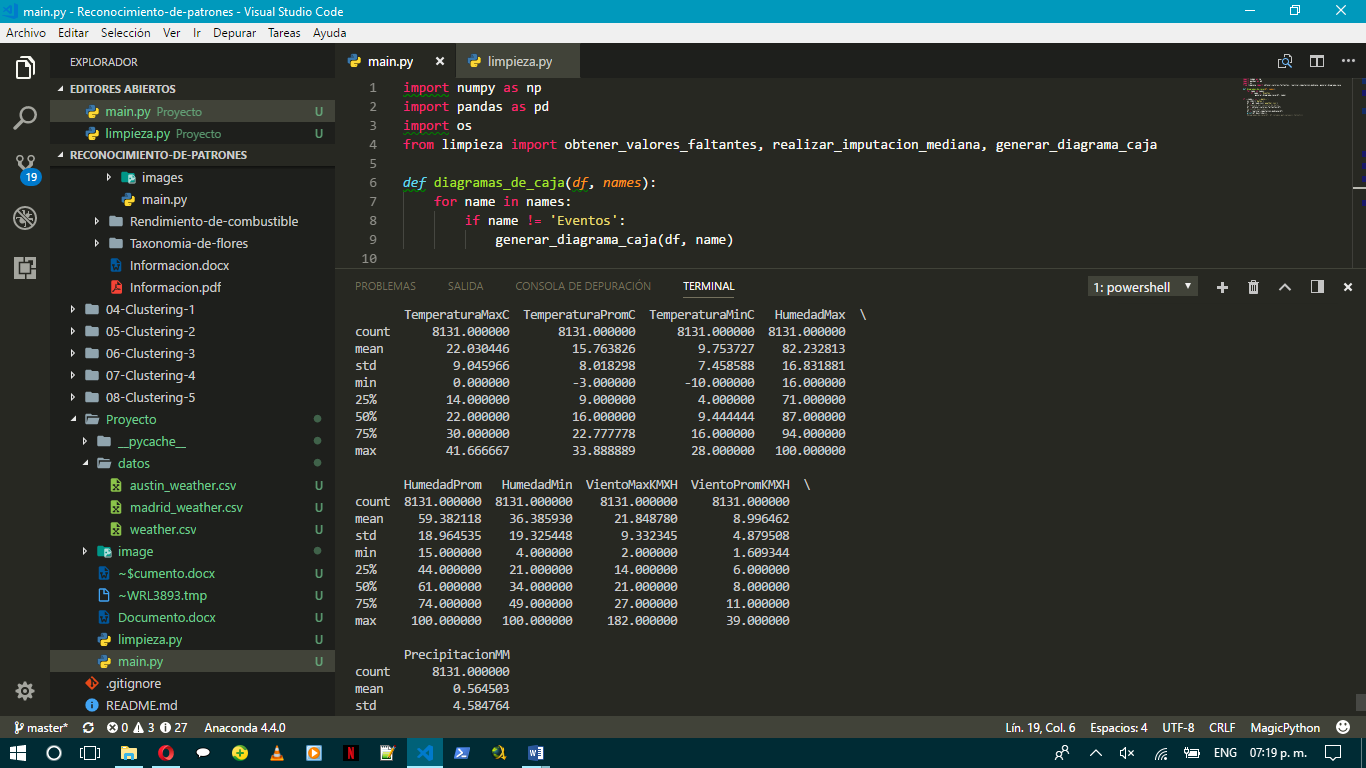
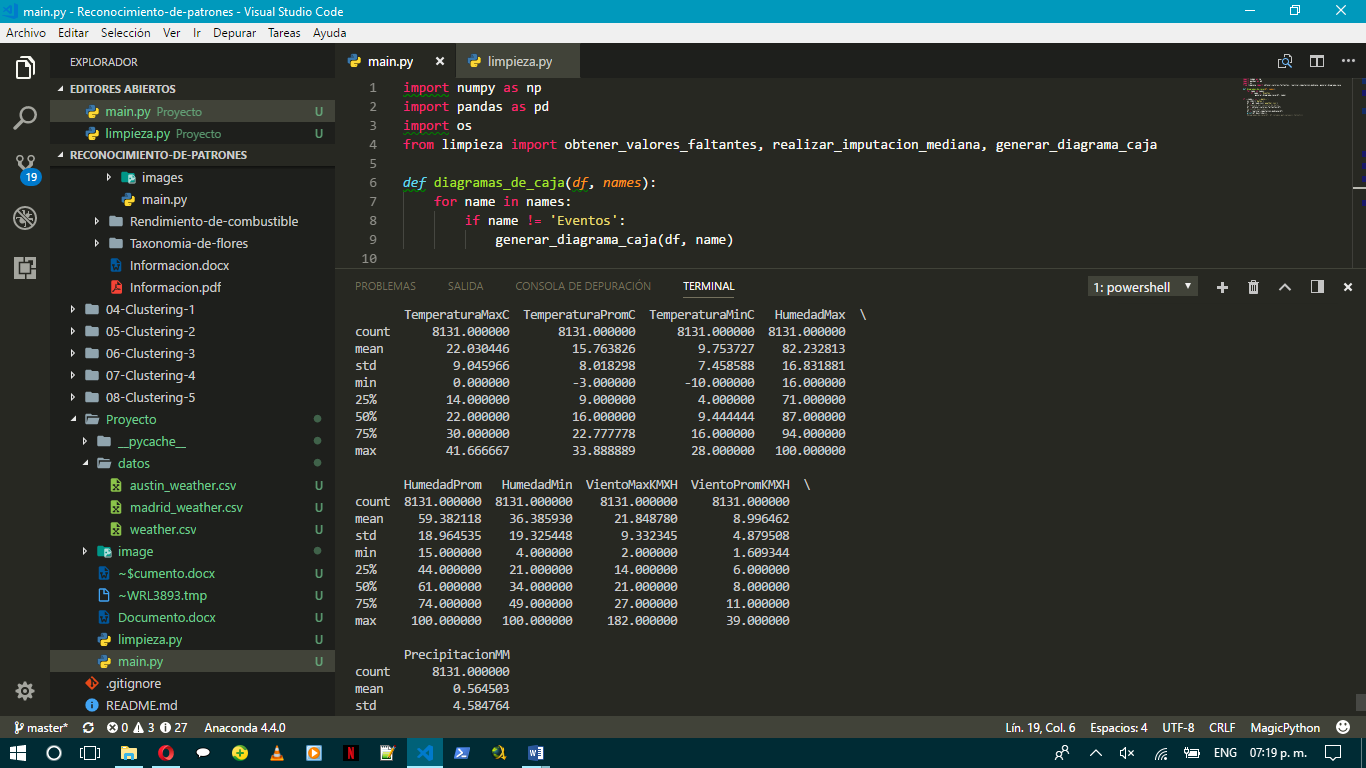
Una vez teniendo el número de valores faltantes se determinó el porcentaje del mismo.



Únicamente la variable de **PrecipitacionMM** tiene un grado de impacto manejable, las otras variables tienen un grado de impacto trivial.

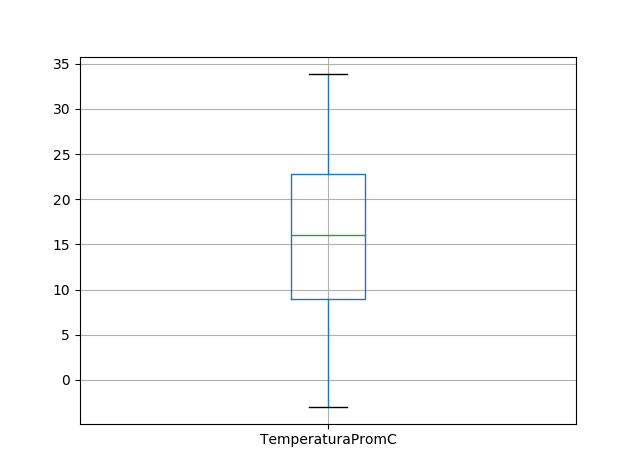
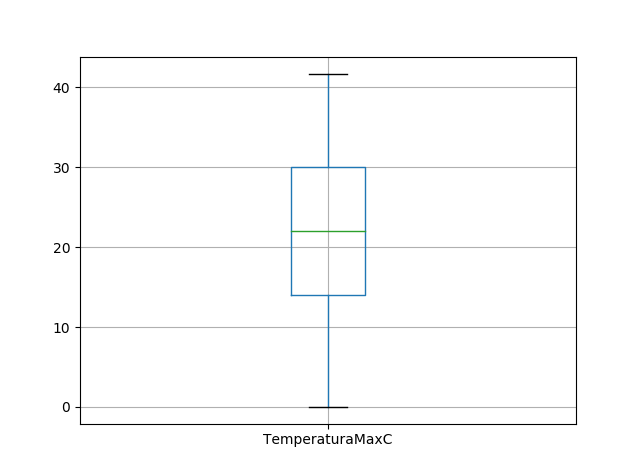
**Imputación**

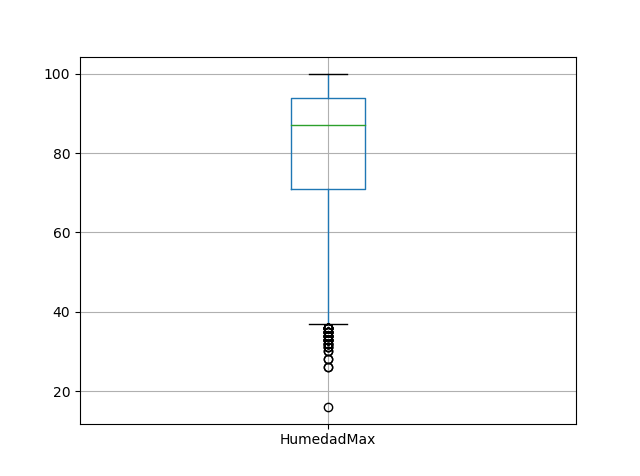
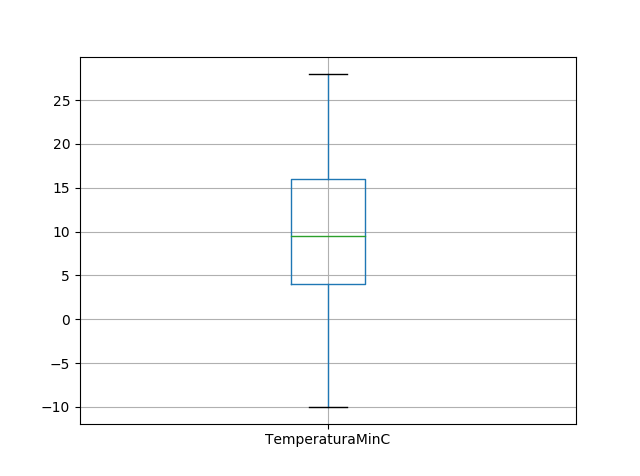
Después se determinó usar imputación con la mediana para el rellenado de los datos faltantes ya que interpolación no tendría sentido ya que los datos ni siquiera tiene una variable de fecha.

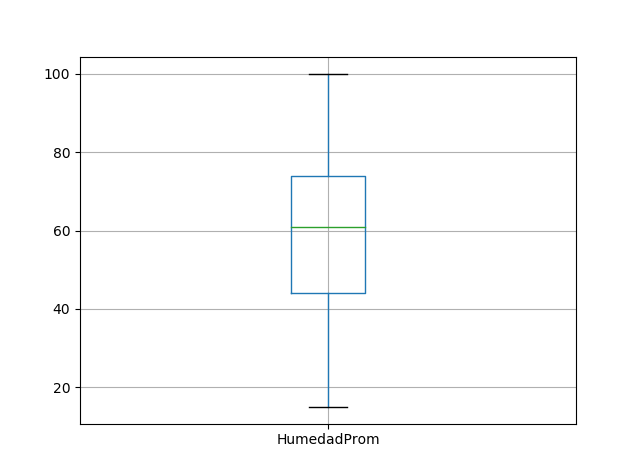
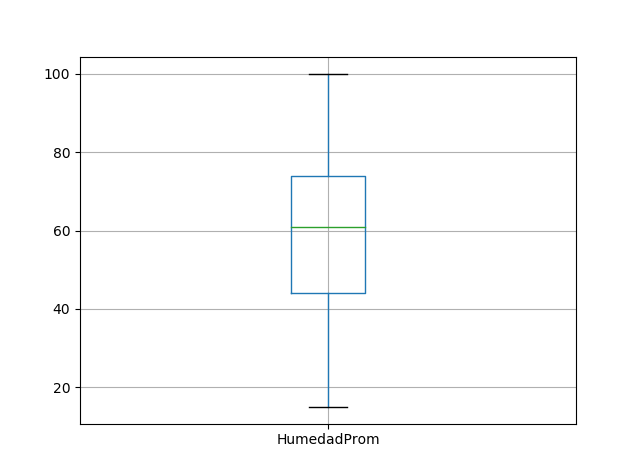


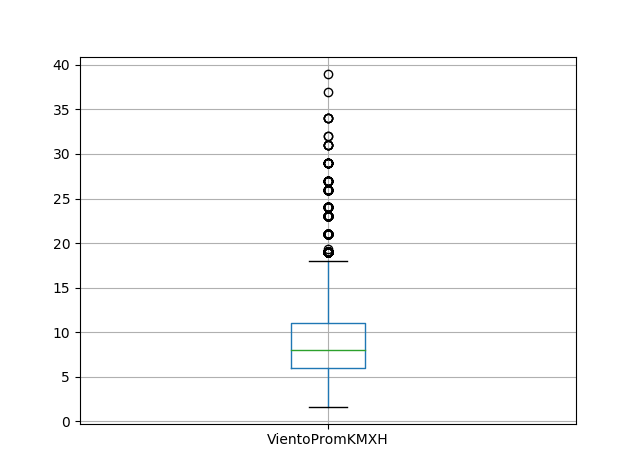
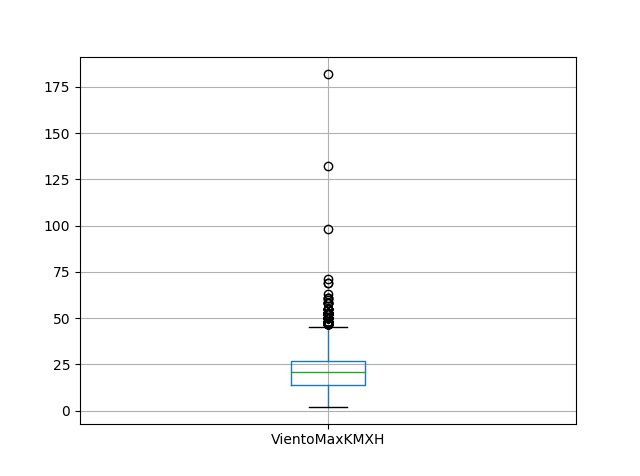
**Valores atípicos**

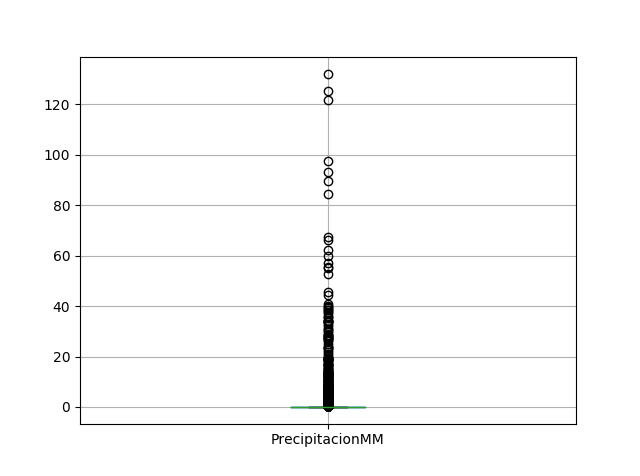
Para el análisis de los valores atípicos se usó diagramas de cajas para cada variable numérica en el conjunto de datos.



****

****

****

****

Los valores atípicos dados por las variables se consideran normales, ya que el conjunto de datos es la unión de mediciones del clima de dos ciudades diferentes y por lo tanto **HumedadMin** contiene valores atípicos. En el caso de las otras variables las de **VientoMaxKMXH**, **VientoPromKMXH** y **PrecipitacionMM**.También son valores atípicos previstos, ya que son las situaciones de las lluvias posiblemente provocadas por tormentas, huracanes, tornados, etc.

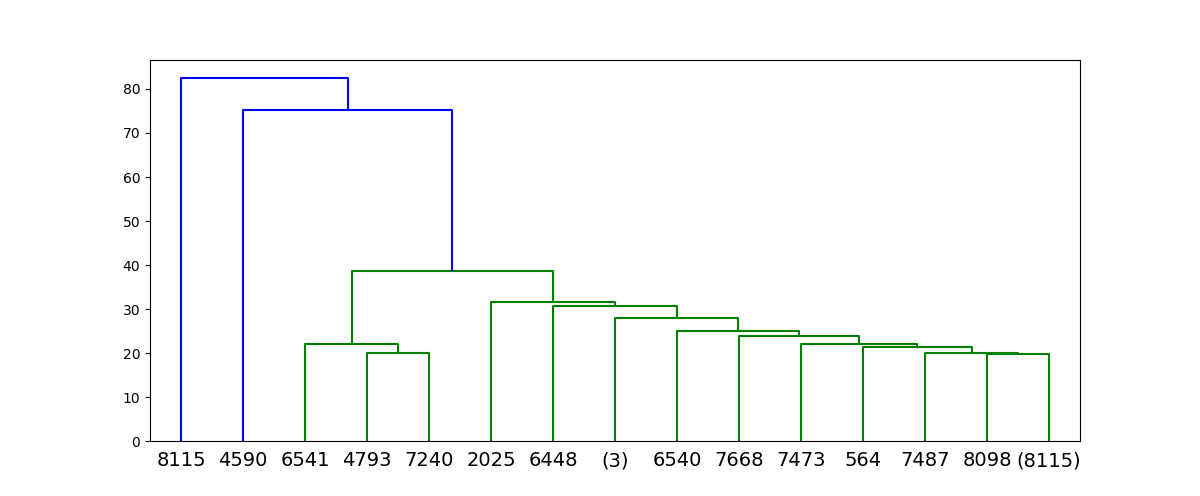
**Clustering**

**Dendrogramas**

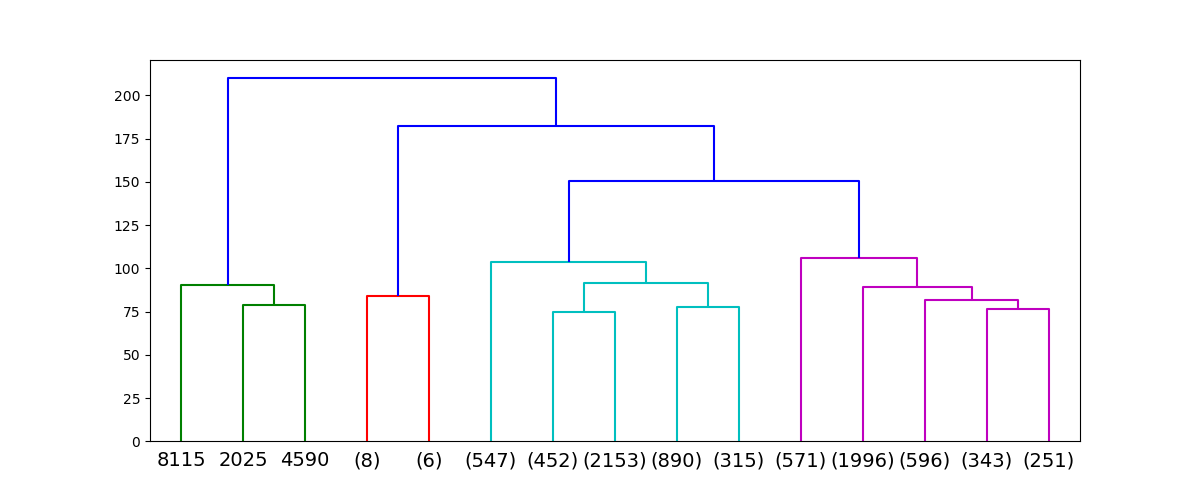
*El código relacionado a esta parte está contenido en el archivo* ***dendrograma.py*** *dentro de la carpeta del proyecto.*

Los dendrogramas son útiles para una primera impresión para saber el número de grupos que puede haber en el conjunto de datos. Se usaran los métodos de singler, complete y centroid para visualizar los últimos 15 clusters formados del conjunto de datos.

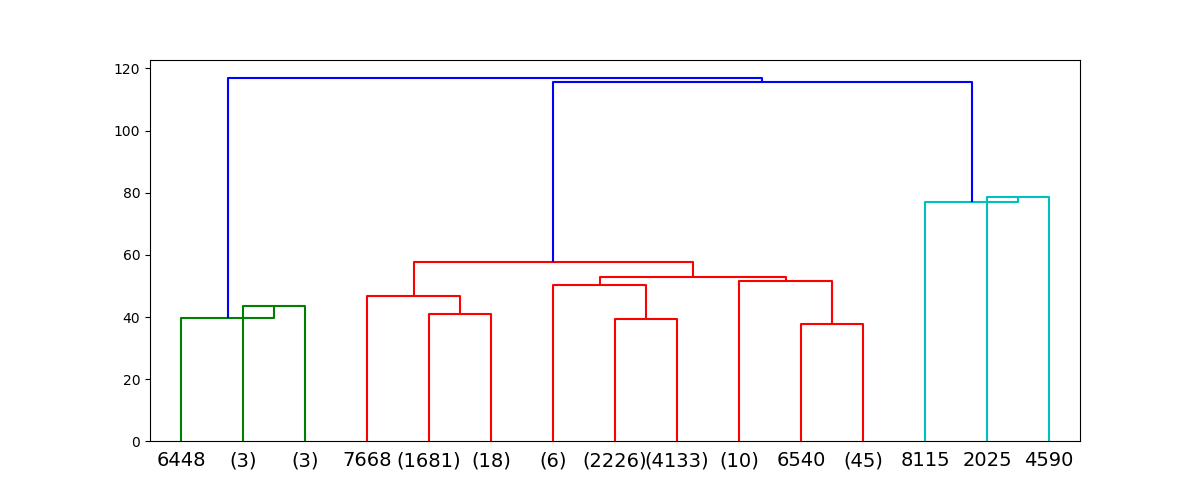
**Dendrograma método singler**



**Dendrograma método complete**

****

**Dendrograma método centroid**

****

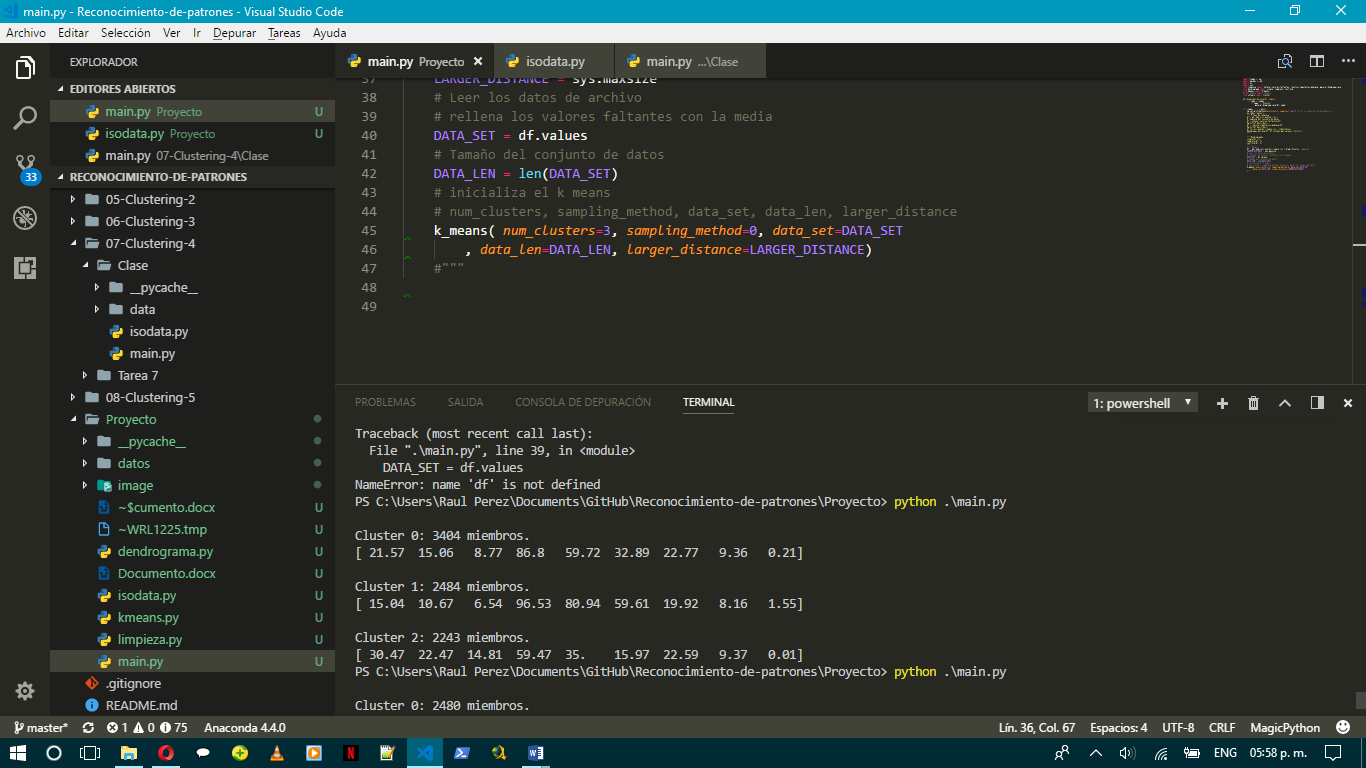
Usando dendrogramas y analizando los resultados con los tres métodos se pueden observar la formación de tres o cuatro grupos en el conjunto de datos. Como los datos son de un conjunto de mediciones del clima, se puede llegar a concluir que existen uno o dos grupos contiene a gran parte de las mediciones que son posiblemente de los tiempos normales del día, es decir, cuando no ocurre un fenómeno natural de gran intensidad ya se por ejemplo un huracán. Los otros grupos que contienen pocas mediciones pueden ser de mediciones de cuando está ocurriendo un fenómeno natural de gran intensidad.

**K-means**

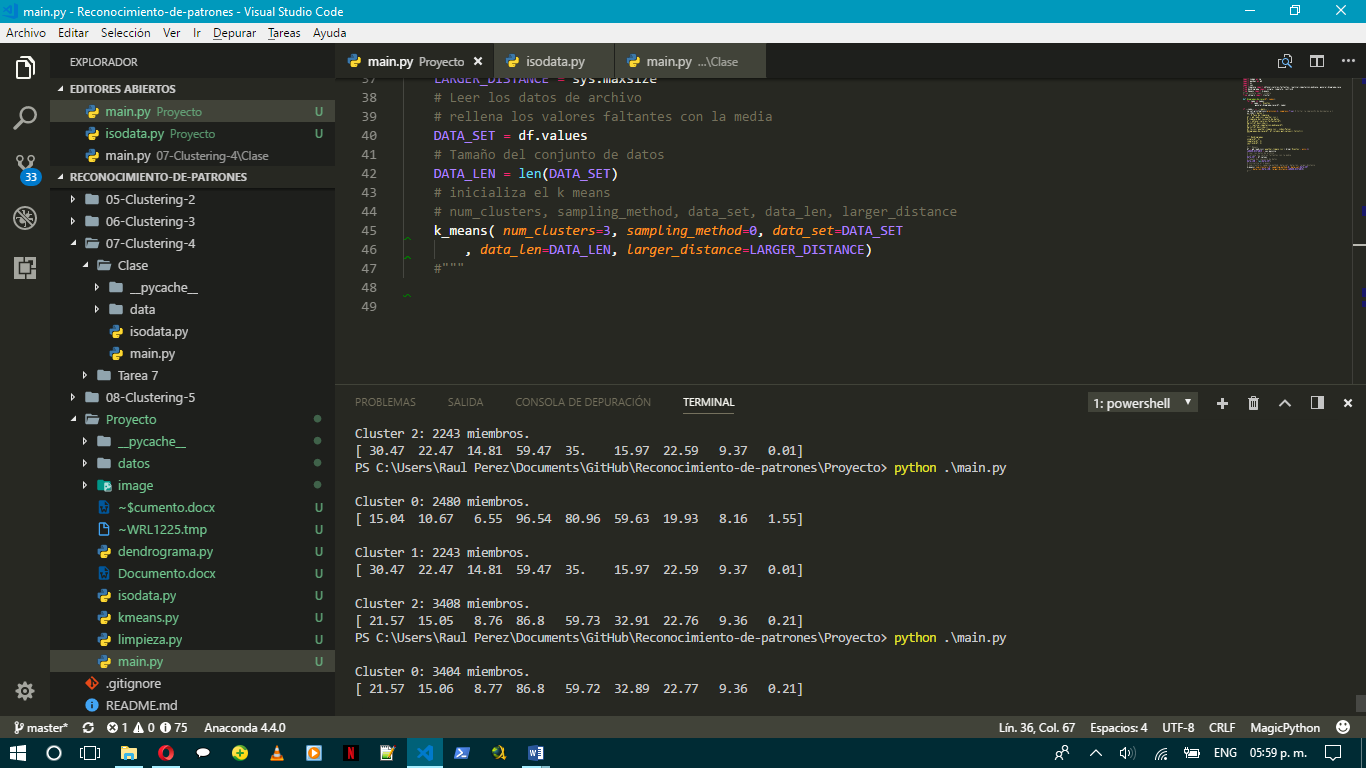
*El código relacionado a esta parte está contenido en el archivo* ***kmeans.py*** *dentro de la carpeta del proyecto.*

K-means es un método de agrupamiento que tiene como objetivo la división de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Como utilizando dendrogramas en el conjunto de datos se concluyó que posiblemente existan cuatro diferentes agrupaciones de datos, se usara k-means con k igual a 3 y 4 con una forma de muestreo aleatoria.

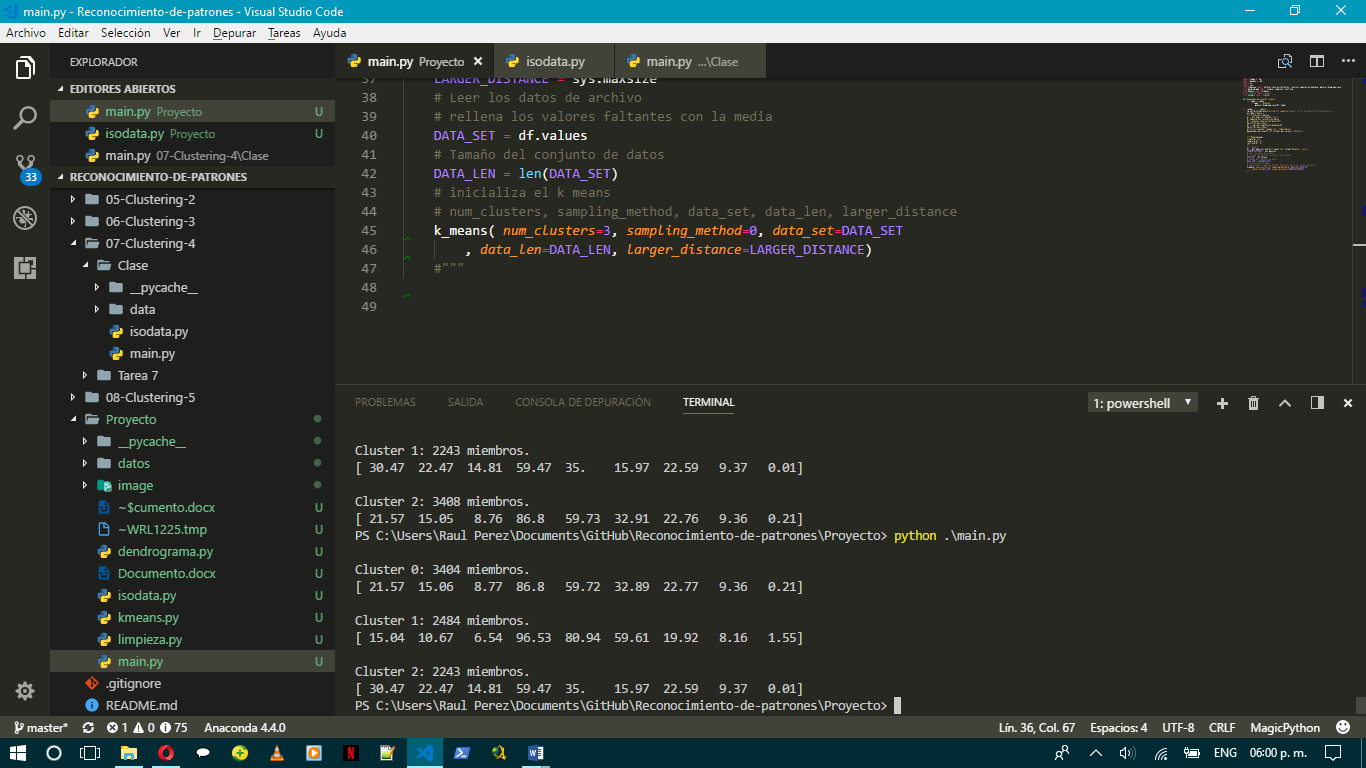
**Primer resultado k = 3**



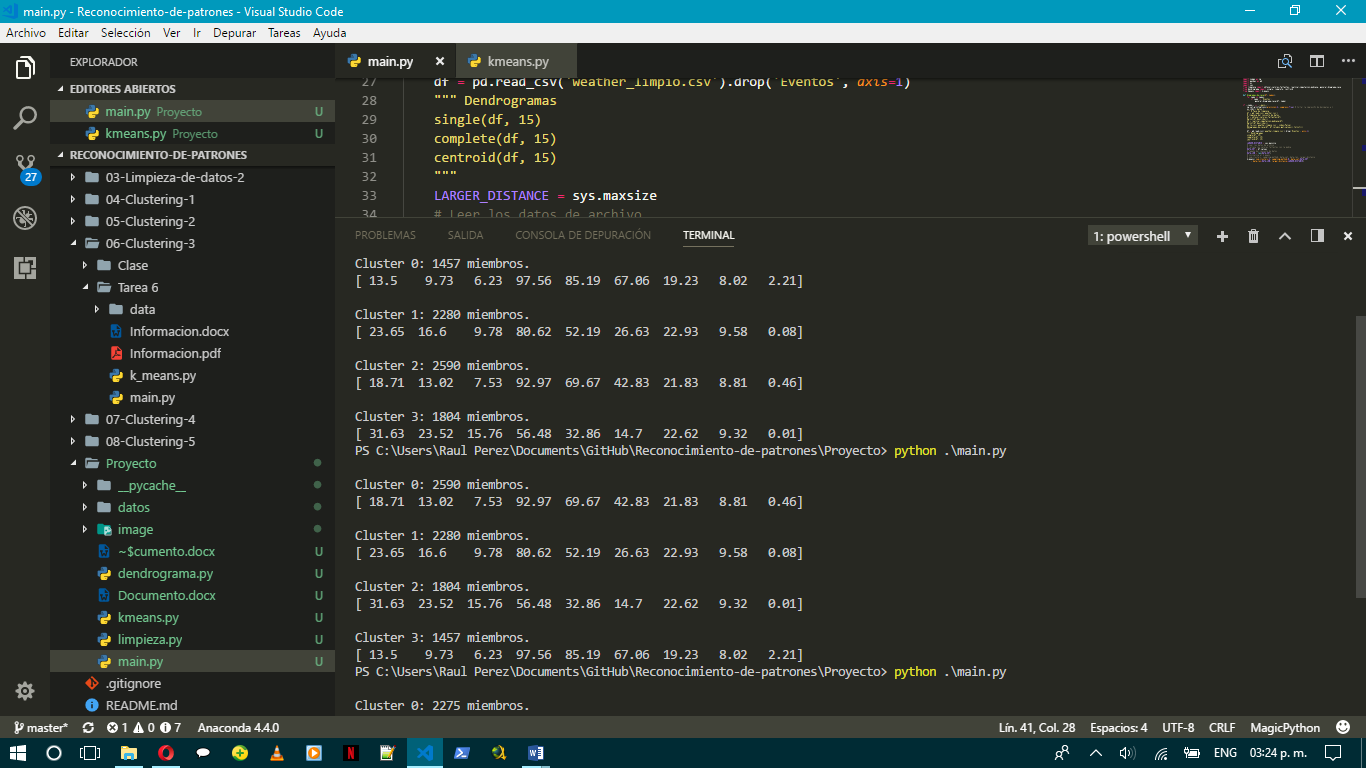
**Segundo resultado k = 3**



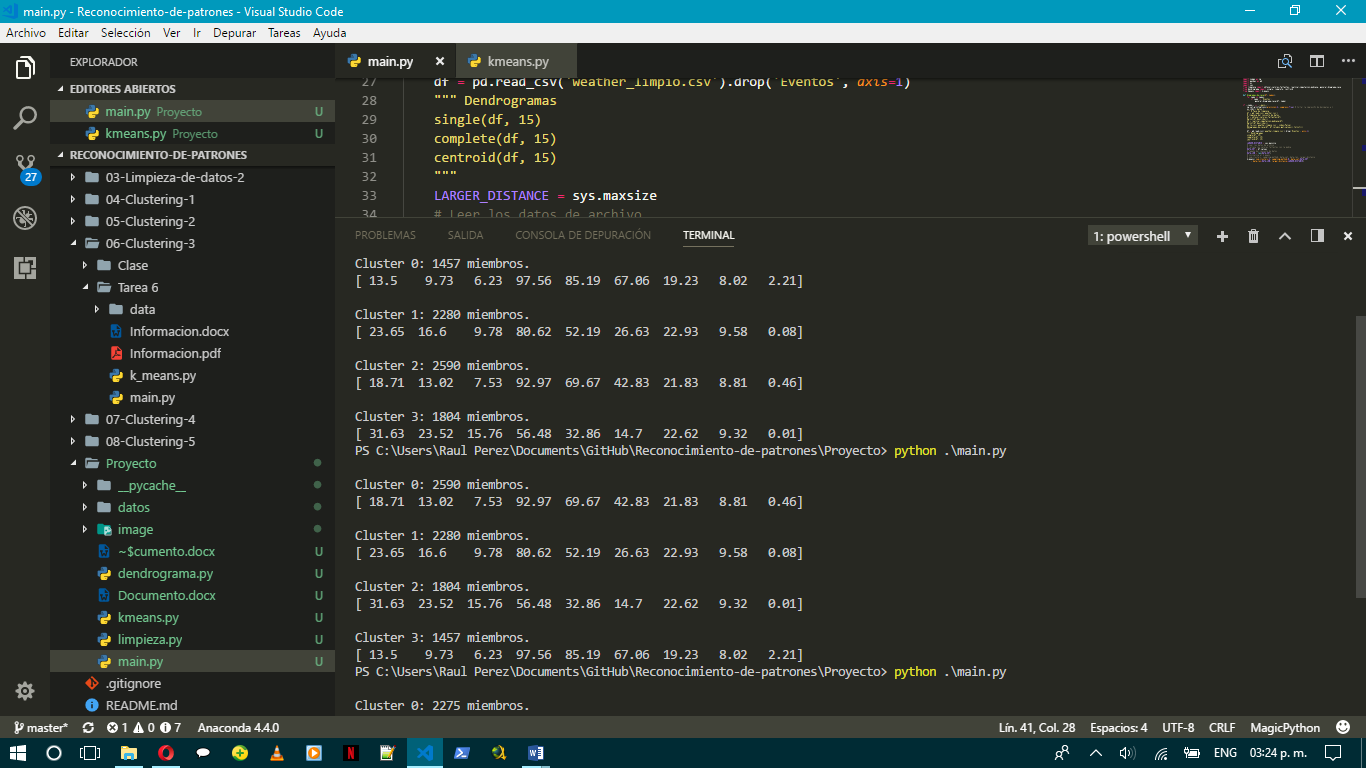
**Tercer resultado k = 3**



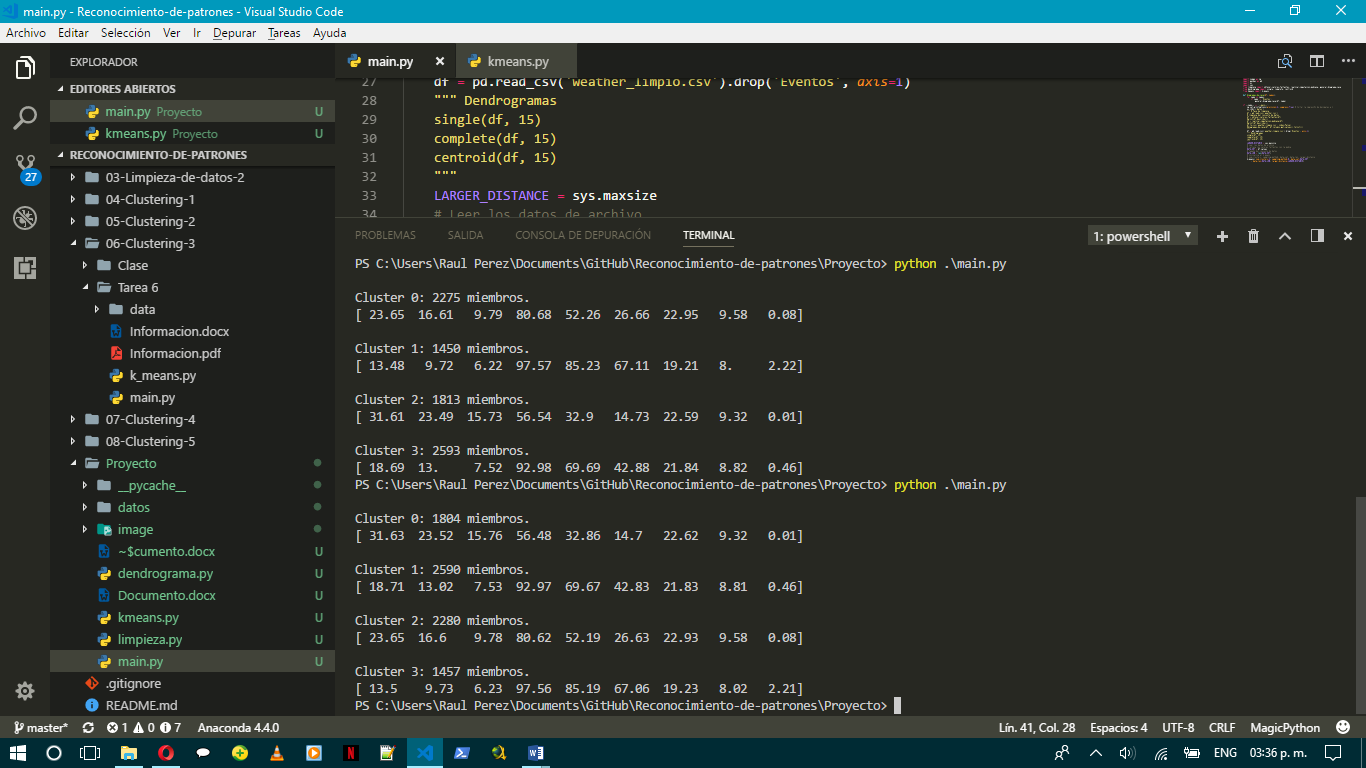
**Primer resultado k = 4**



**Segundo resultado k = 4**



**Tercer resultado k = 4**



Se puede observar tanto con k igual a 3 como k igual a 4, que los grupos que se forman son iguales (k = 4) o casi iguales (k = 3) en los 3 resultados que se hicieron de cada uno. Por lo tanto, se puede pensar que es posible que existan tres o cuatro grupos diferentes como se había concluido en los dendrogramas.

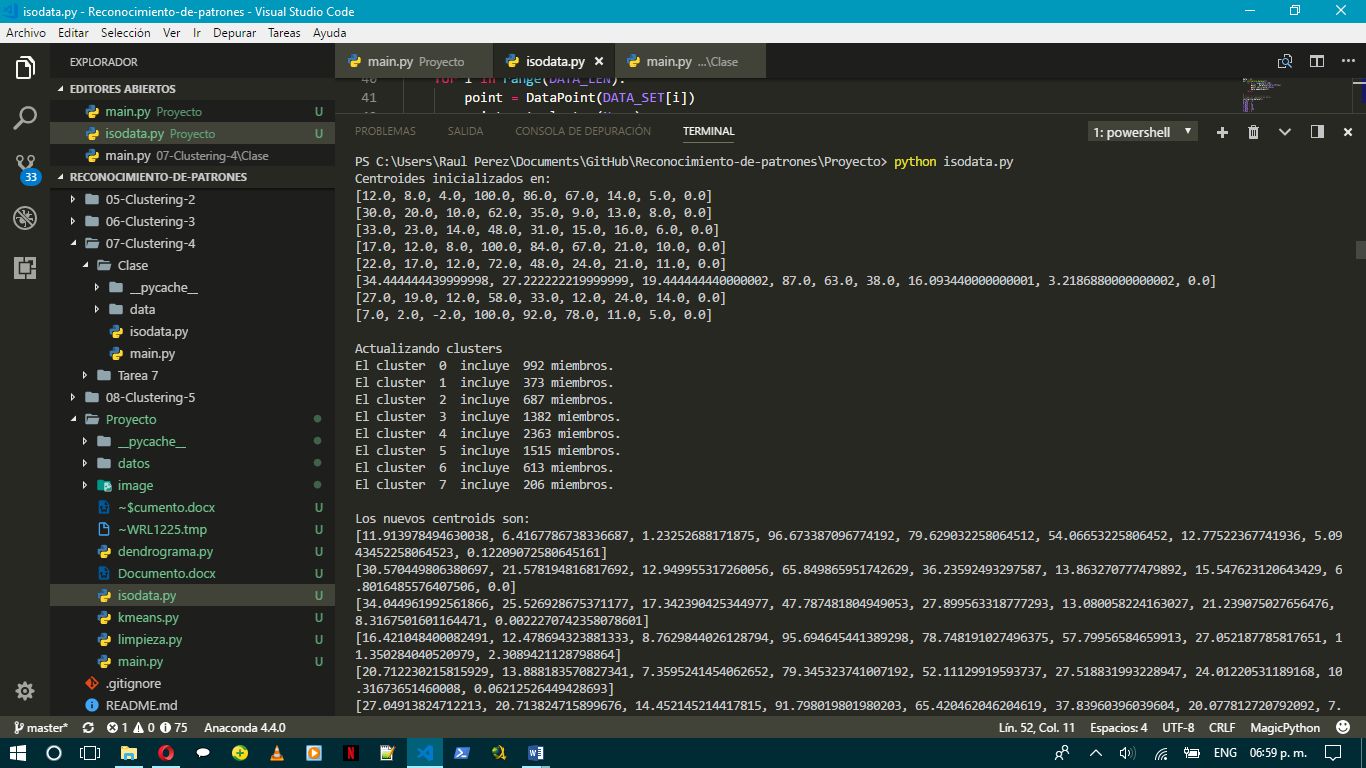
**ISODATA**

*El código relacionado a esta parte está contenido en el archivo* ***isodata.py*** *dentro de la carpeta del proyecto.*

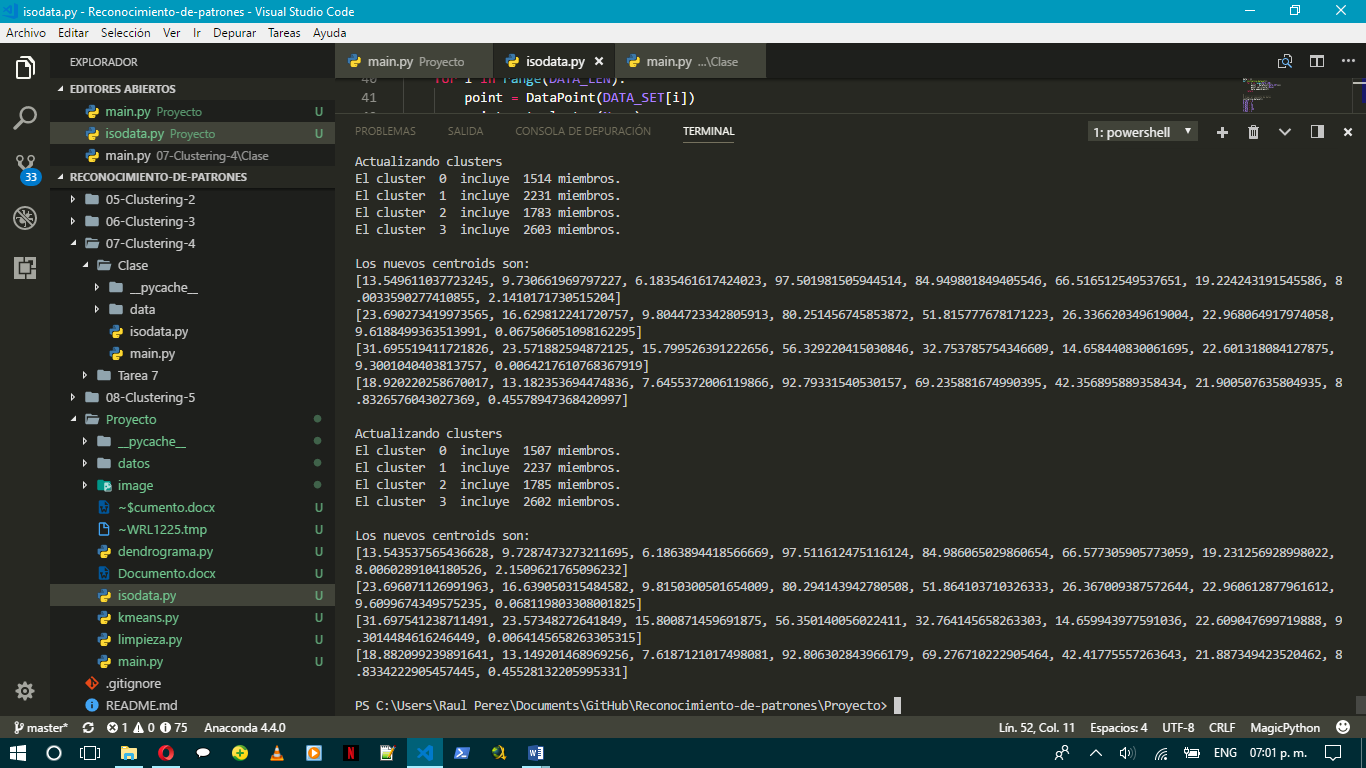
El método ISODATA es un algoritmo similar al k-means. Su objetivo es particionar un conjunto de datos en subconjuntos. Sin embargo, a diferencia de k-means, el método ISODATA  maneja una serie de heurísticas con tres objetivos que le permiten optimizar el número de clusters:

* Eliminar clusters con poco ejemplares.
* Unir clusters muy cercanos.
* Dividir clusters dispersos

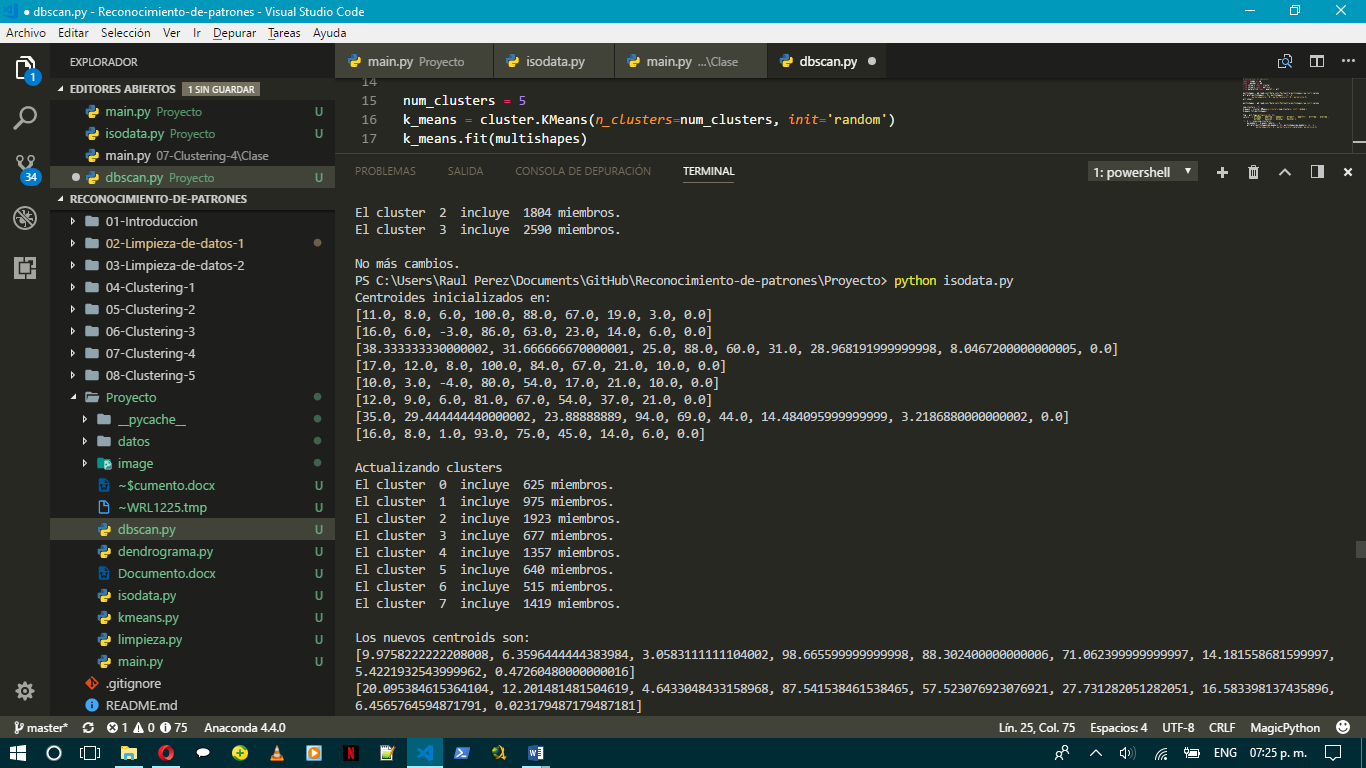
**Centroides iniciales resultado uno**



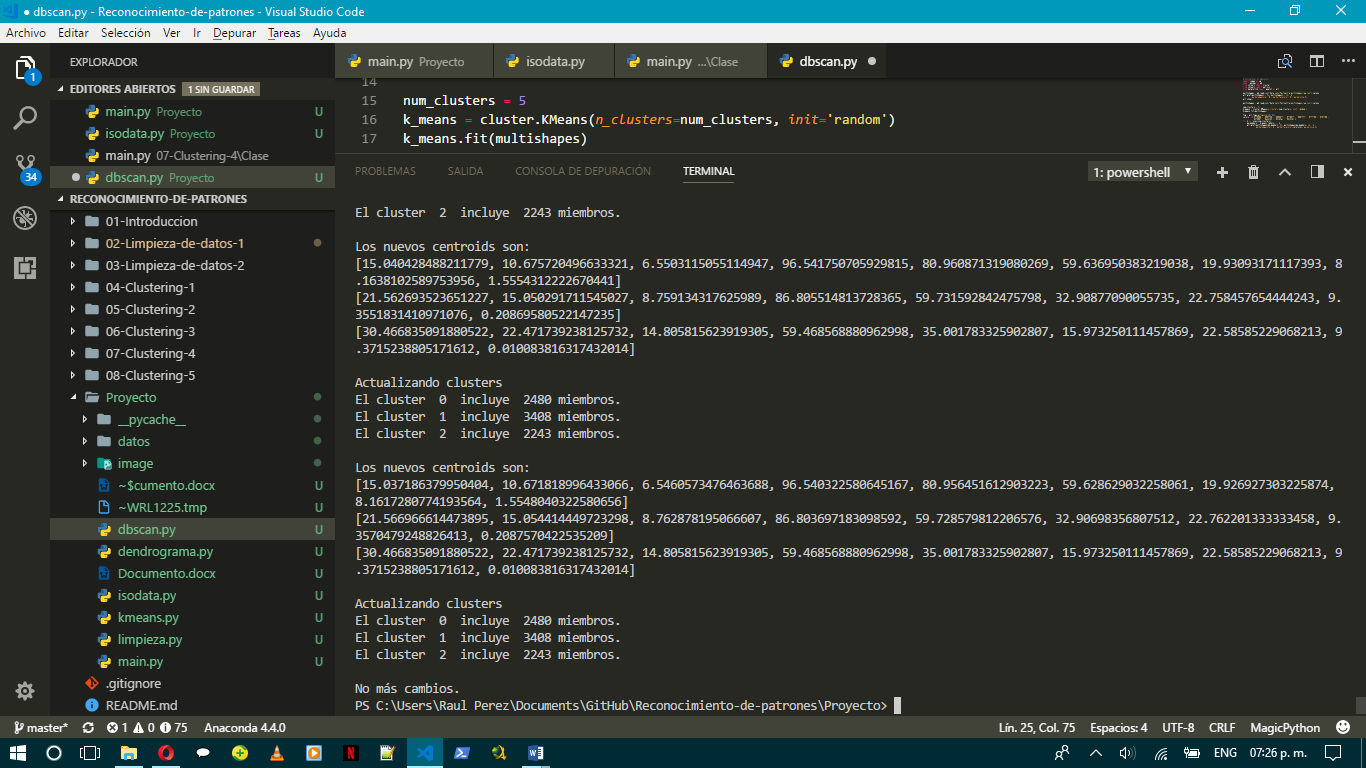
**Centroides finales resultado uno**



**Centroides iniciales resultado dos**



**Centroides finales resultado dos**



Realizando varias ejecuciones del ISODATA con el conjunto de datos, se observó que con una k inicial de 8 y con una forma de muestreo aleatorio, al final de las iteraciones se formaban grupos de 3 y 4 clusters. Mismos que se observaban en los demás técnicas de clustering.

**Clasificación**

**K-Vecinos Próximos**

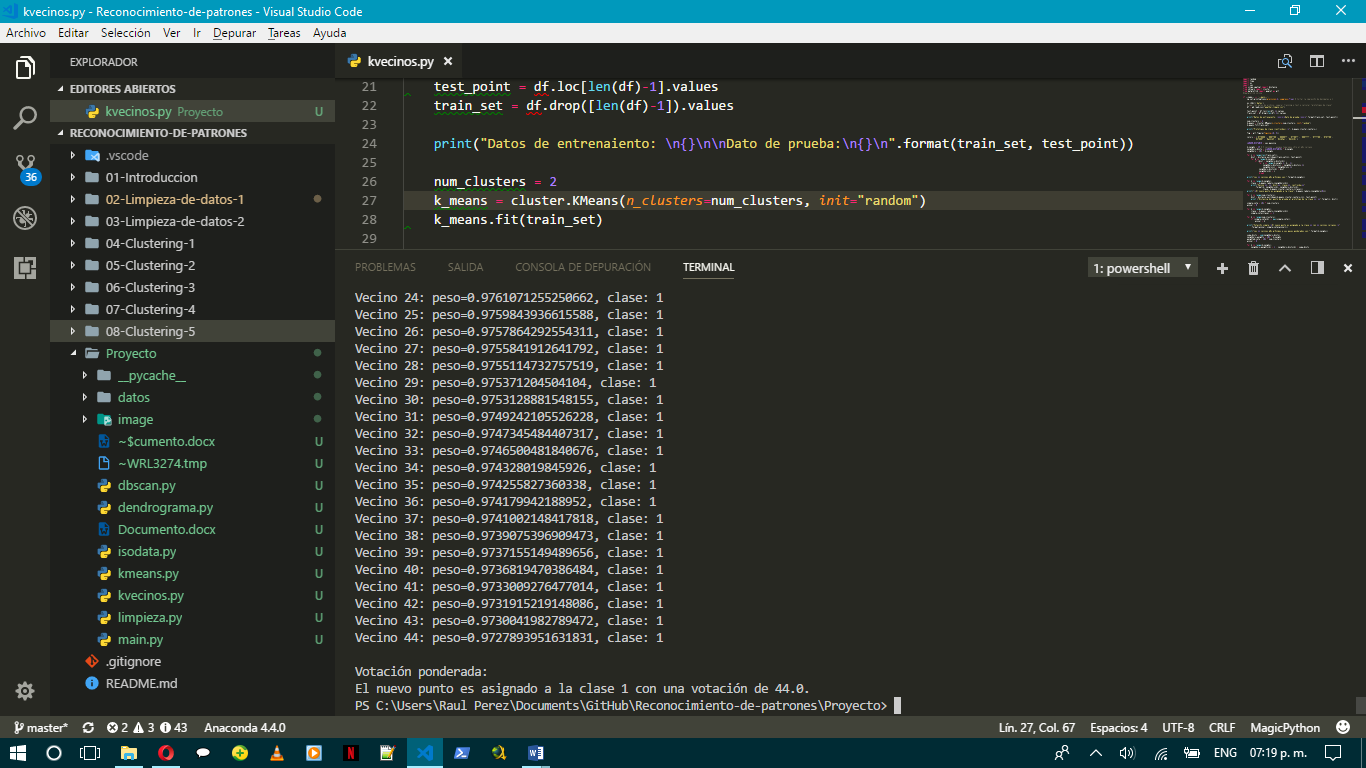
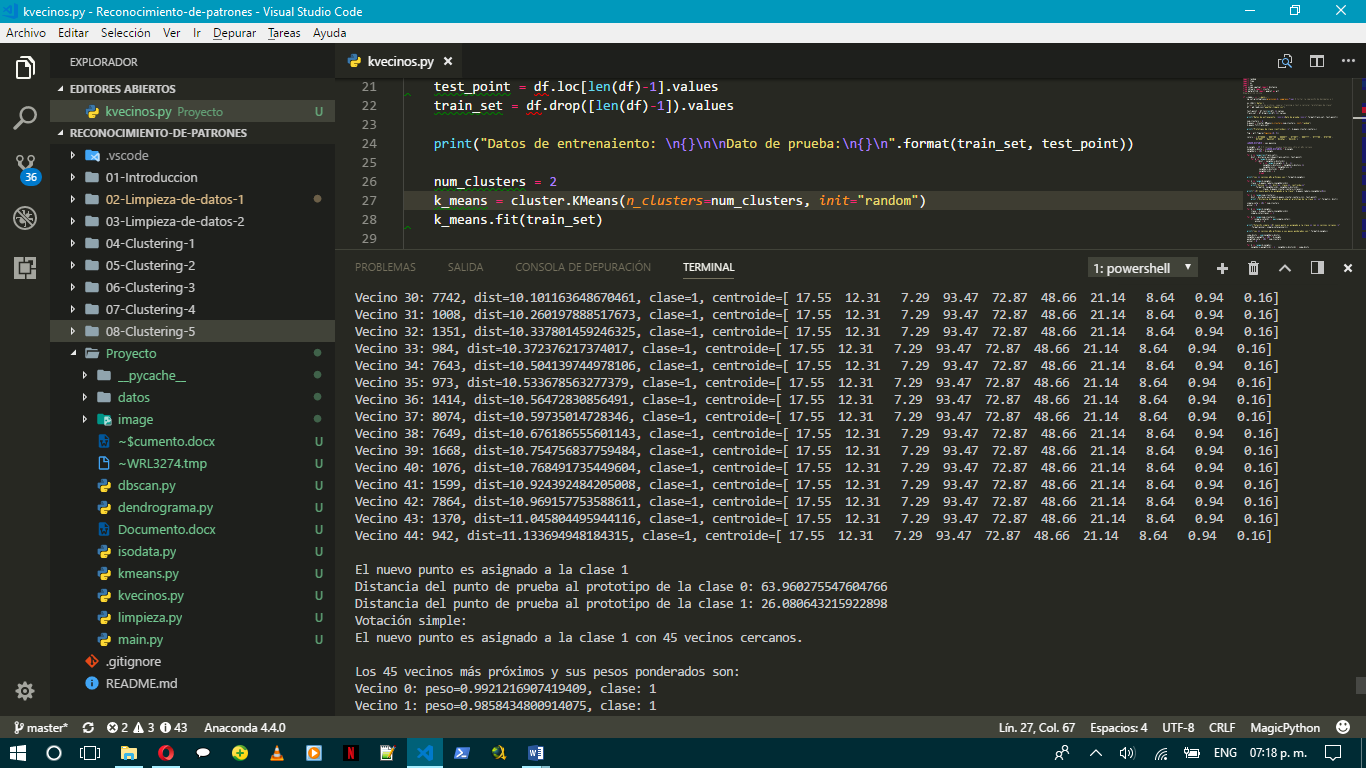
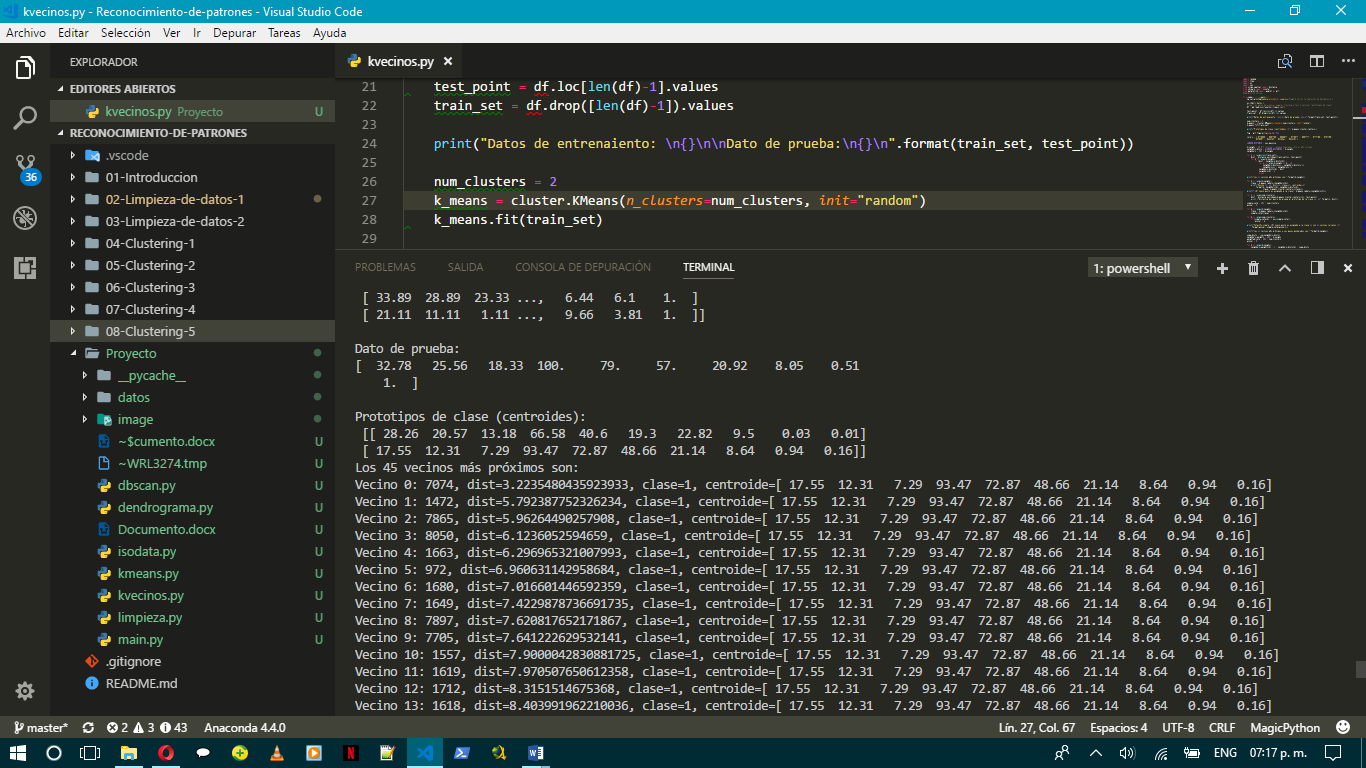
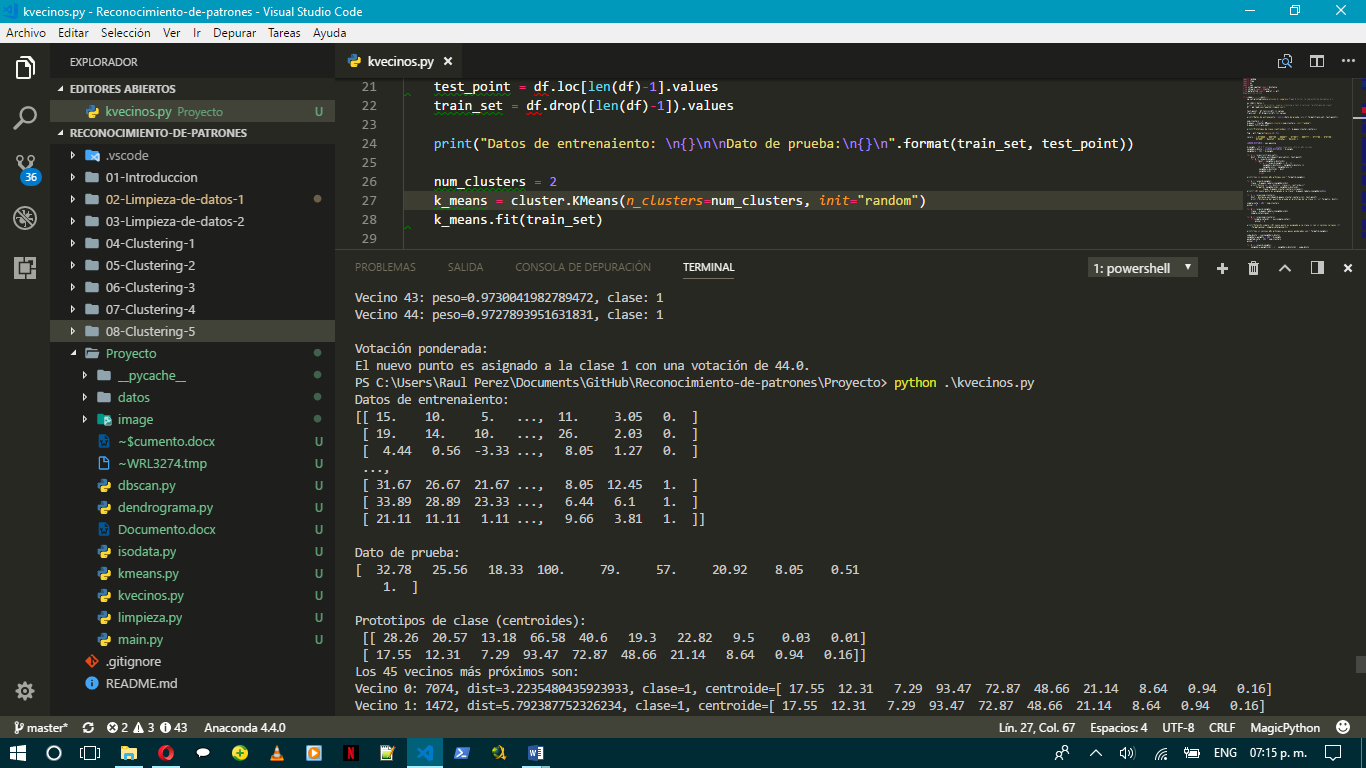
*El código relacionado a esta parte está contenido en el archivo* ***kvecinos.py*** *dentro de la carpeta del proyecto.*

*Debido a que el conjunto de datos contiene varios eventos dependiendo de lo sucedido en el día, se cambiara sus valores a sucedió o no sucedió un evento en el día, para que de esa manera haya dos clases.*

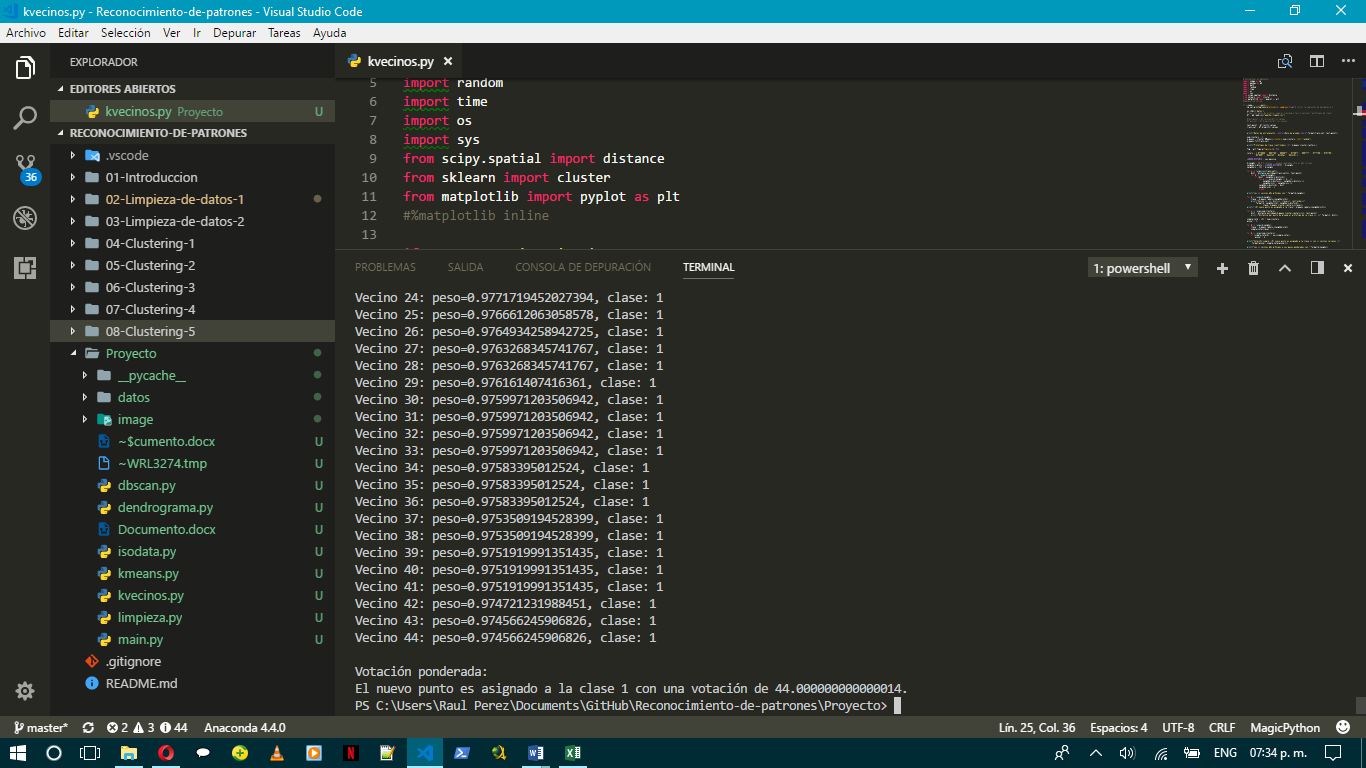
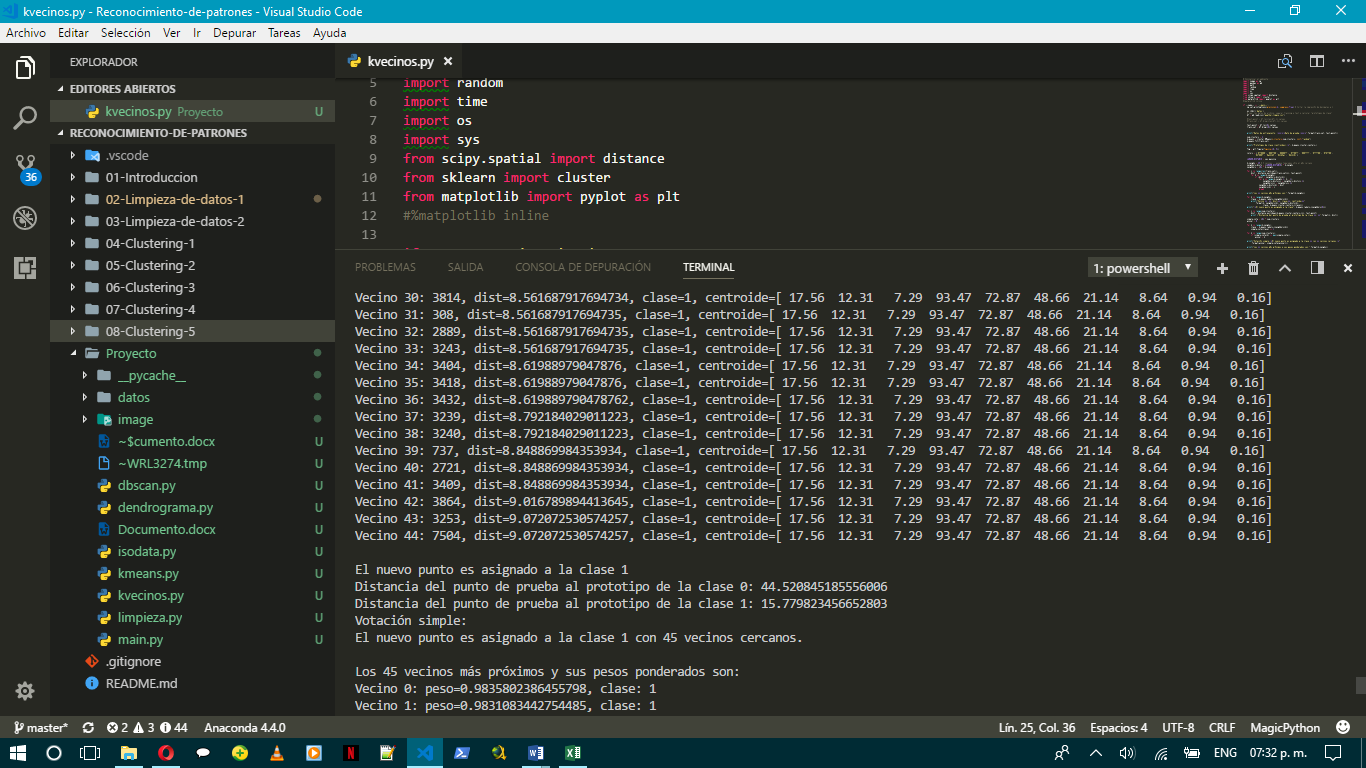
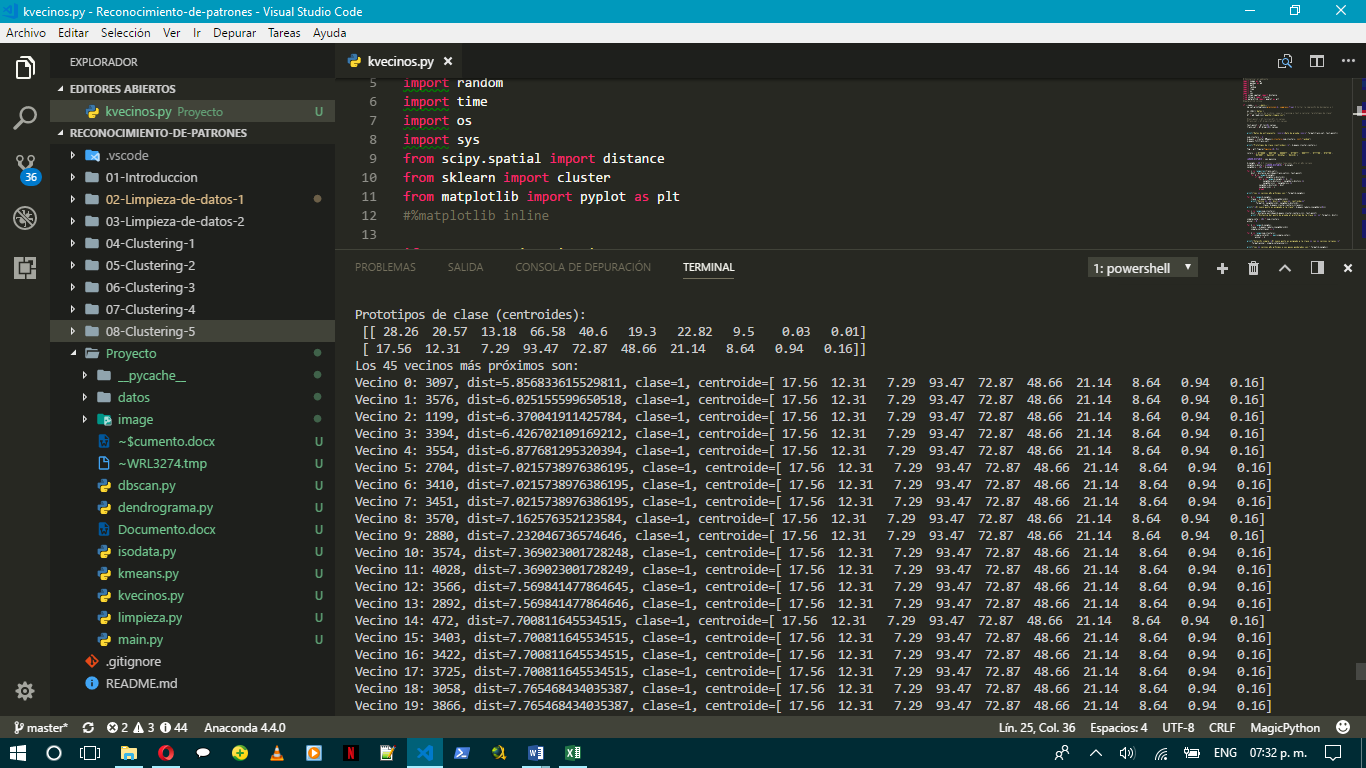
El método de k**-vecinos próximos** es un método de aprendizaje vago (lazy learning). Esto significa que el aprendizaje no conduce a una generalización: no existe una fase de entrenamiento (o es muy breve) y en su lugar el método mantiene todos los datos disponibles y los emplea para realizar la clasificación.

Para la ejecución del método se tomaron 45 debido a que correspondía con la cantidad sugerida: donde es la cantidad de datos. Se tomó la última medición del conjunto de datos, correspondiente a un dato del tipo de evento con valor 1 (sucedió) y la primera medición que corresponde a un evento de no sucedió*.*

**Última medición**



**Primera medición**



Se puede observar que los dos grupos formados por el método, se encuentran mesclados mediciones de eventos de sucedió o no sucedió. Ya que tanto la primera como la segunda medición caen en el mismo grupo.

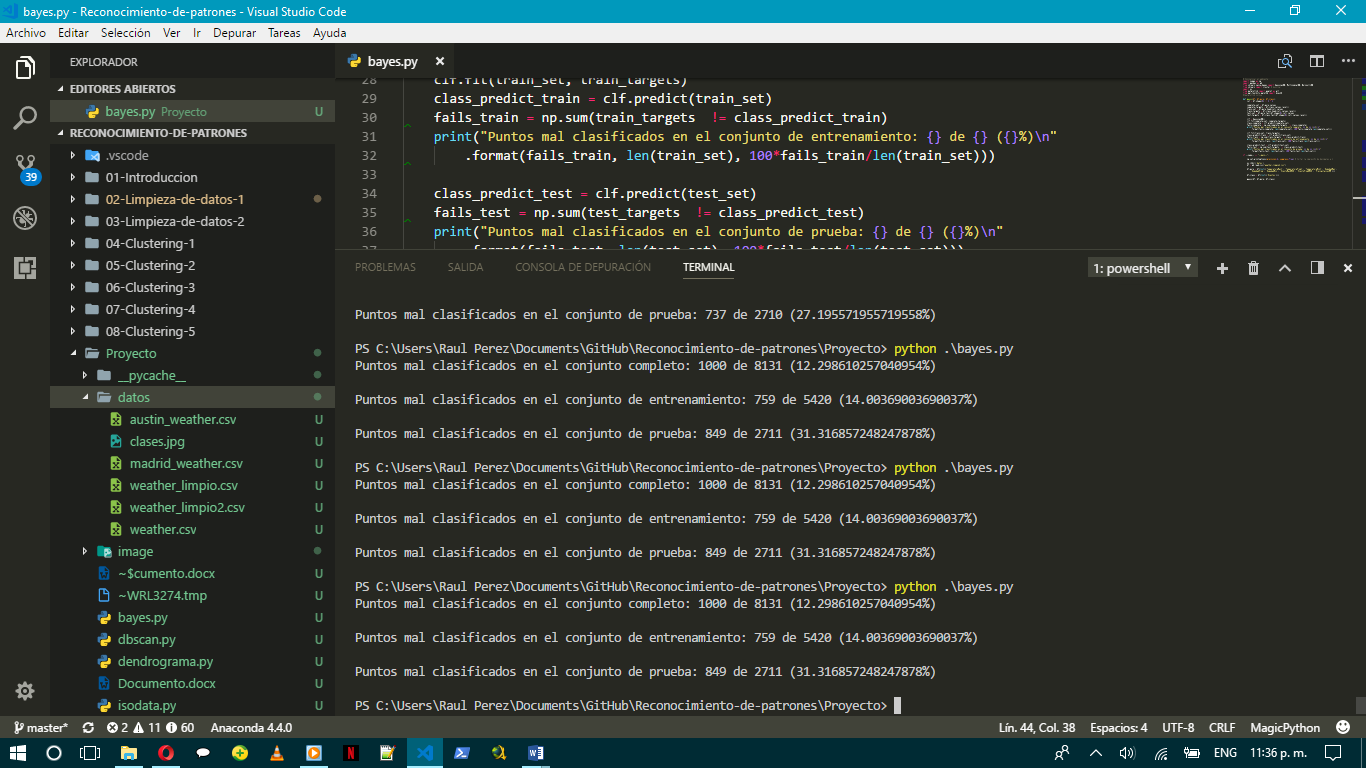
**Teorema de Bayes**

*El código relacionado a esta parte está contenido en el archivo* ***bayes.py*** *dentro de la carpeta del proyecto.*

En teoría de la probabilidad y minería de datos, un clasificador Bayesiano ingenuo es un clasificador probabilístico fundamentado en el teorema de y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de ingenuo.

Para el clasificador de Bayes se dividieron los datos en datos de entrenamiento y datos de prueba para poder hacer las pruebas con la distribución Gaussiana.

**Modelo gaussiano**



Con la distribución Gaussiana se obtuvo un error de 12.29% para el conjunto de datos completos, es un error razonable teniendo en cuanta que hay 10 tipos de clases en el conjunto de datos.

**Árbol de decisión**

*El código relacionado a esta parte está contenido en el archivo* ***arbol.py*** *dentro de la carpeta del proyecto.*

Los [árboles de decisión](https://en.wikipedia.org/wiki/Decision_tree) son uno de los métodos de clasificación más antiguos y robustos y, por lo tanto, de las más utilizados. Los árboles de decisión son un método no paramétrico de aprendizaje supervisado utilizado para clasificación y regresión. El objetivo es generar un conjunto de reglas de decisión simples a partir de las características de los datos de entrenamiento. En un árbol de decisión, los datos de entrenamiento son divididos en cada nodo utilizando reglas de decisión para obtener grupos de elementos similares. Este agrupamiento equivale a una operación de decisión tipo *switch*(caso). Cada nodo interno en el árbol representa una evaluación sobre un atributo, los arcos representan los posibles resultados de la evaluación y las hojas representan una decisión (una asignación de elementos a una clase). Las trayectorias, desde el nodo raíz a las hojas son las reglas de decisión.