UNIVERSIDAD EUROPEA MIGUEL DE CERVANTES

ESCUELA POLITECNICA SUPERIOR

TITULACIÓN:

MÁSTER EN GESTIÓN Y ANÁLISIS DE GRANDES VOLÚMENES DE DATOS: BIG DATA



TRABAJO FIN DE MASTER

DESARROLLO DE ALGORITMOS METAHEURISTICOS PARA OPTIMIZACION DE REDES NEURONALES ARTIFICIALES APLICADOS A LA EFICIENCIA ENERGÉTICA BAJO SPARK

AUTOR

RUBÉN CORREOSO CAMPILLO

TUTOR

Mª DEL CARMEN PEGALAJAR JIMÉNEZ

VALLADOLID, JULIO DE 2020

1. Objetivos del trabajo.
2. Análisis de la situación.
3. Obtención, procesado y almacenamiento de los datos.
4. Análisis exploratorio.
5. Diseño e implementación de los modelos o técnicas necesarios.
6. Análisis de los resultados obtenidos.
7. Conclusiones y planes de mejora.
8. Bibliografía.
9. Anexo con el código fuente desarrollado.
10. **Objetivos del trabajo**
11. **Análisis de la situación.**
12. **Obtención, procesado y almacenamiento de los datos.**
    1. **Introducción**

El dataset **“Hourly Energy Consumption”[[1]](#footnote-1)** es un conjunto de datos alojado en Kaggle por el investigador científico Rob Mulla, Kernell Máster de Kaggle por sus aportaciones a la plataforma.

El conjunto de datos utilizado en el estudio contiene información referente a casi 20 años de consumo de energía de **PJM Interconnection LLC[[2]](#footnote-2)**, una organización de transmisión regional (RTO) en los Estados Unidos.

PJM es parte de la red de interconexión oriental que opera un sistema de transmisión eléctrica que opera en Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, Nueva Jersey, Carolina del Norte, Ohio, Pensilvania, Tennessee, Virginia, Virginia Occidental y el Distrito de Columbia.

El dataset se compone de trece archivos en formato “csv” y un archivo en formato “parquet” con información sobre el consumo de energía estimada en megavatios en formato float64, registrada cada hora por diferentes compañías.

Doce de los csv del conjunto de datos contienen una estructura similar de dos columnas, una que indica la potencia en megavatios estimada y la otra el instante de tiempo al que se asocia en formato AAAA-MM-DD HH:mm:ss:

* “AEP\_hourly.csv" contiene información de la compañía “American Electric Power” [[3]](#footnote-3).
* “COMED\_hourly.csv" contiene información de la compañía “Commonwealth Edison” [[4]](#footnote-4).
* “DAYTON\_hourly.csv” contiene información de la compañía “The Dayton Power and Light Company” [[5]](#footnote-5).
* “DEOK\_hourly.csv” contiene información de la compañía “Duke Energy Ohio/Kentucky” [[6]](#footnote-6).
* “DOM\_hourly.csv” contiene información de la compañía “Dominion Virginia Power” [[7]](#footnote-7).
* DUQ\_hourly.csv contiene información de la compañía “Duquesne Light Co.” [[8]](#footnote-8).
* “EKPC\_hourly.csv” contiene información de la compañía “East Kentucky Power Cooperative” [[9]](#footnote-9).
* “FE\_hourly.csv” contiene información de la compañía “FirstEnergy” [[10]](#footnote-10).
* “NI\_hourly.csv” contiene información de la compañía Northern Illinois Hub.
* “PJM\_Load\_hourly.csv” contiene información de la compañía PJM desde el año 1998 hasta el 2001.
* “PJME\_hourly.csv” contiene información de la región Este de la compañía PJM desde el año 2001 hasta el 2018.
* “PJMW\_hourly.csv” contiene información de la región Oeste de la compañía PJM desde el año 2001 hasta el 2018.

El csv restante “pjm\_hourly\_est.csv” y el archivo parquet “est\_hourly.parquet”, tienen la misma información.

Se componen de doce columnas que contienen la información estimada de los megavatios de los archivos anteriores, y una columna con información temporal en formato AAAA-MM-DD HH:mm:ss de cada hora registrada.

He decidido trabajar directamente con el archivo “pjm\_hourly\_est.csv” dado que contiene la información combinada de los demás archivos distribuida entre sus columnas.

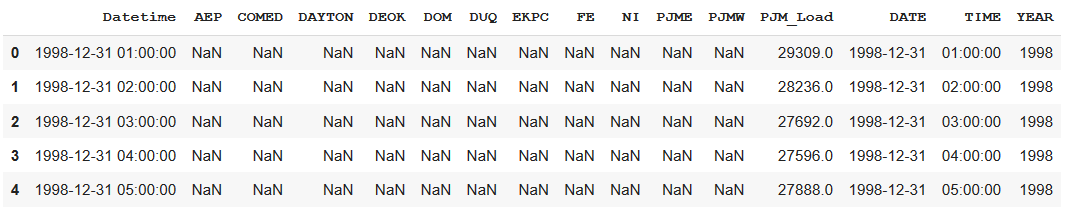


Ilustración 1: Primeras filas del dataset pjm\_hourly\_est.csv

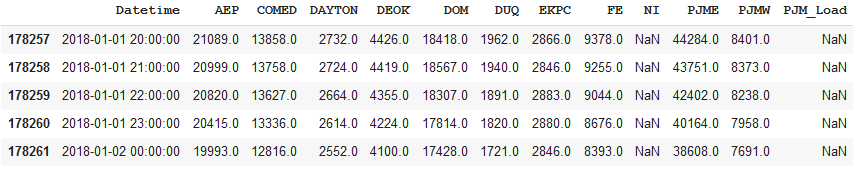


Ilustración 2: Últimas filas de dataset pjm\_hourly\_est.csv

Como se puede apreciar en la ilustración 1 y 2 la columna PJM\_Load tiene información al inicio del dataset mientras que las demás regiones se encuentran sin información. Esto es debido a que PJM\_Load tiene registrado desde 1998 hasta 2002 mientras las demás columnas permanecen vacías hasta 2002. No hay una fecha de inicio común para todas las columnas en la que se empiecen a recibir datos por lo que más adelante profundizaremos en sus periodos de actividad y cómo afrontar los valores nulos.

Las diferentes regiones de PJM se han unificado y separado a lo largo de los años por lo que es normal que falten datos en alguna región en alguna fecha.

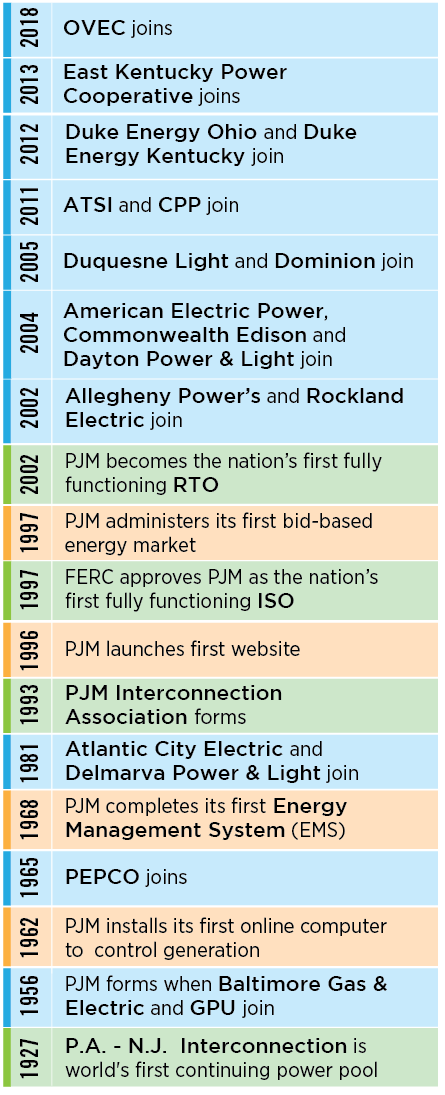
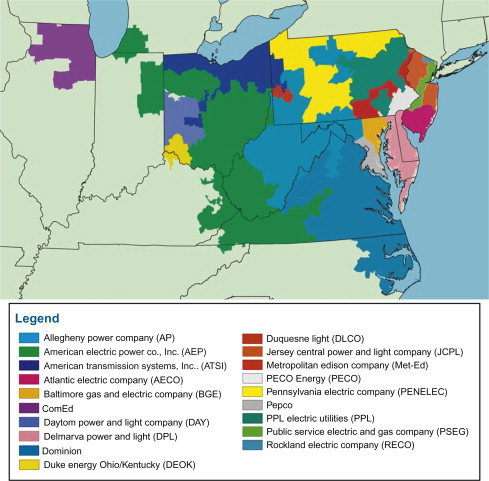


Ilustración 3: Evolución de las regiones de PJM y estado actual

* 1. **Preprocesado de datos**

Antes de comenzar a analizar los datos, ver las posibles tendencias y correlaciones, es necesario revisarlos y preprocesarlos. Es muy común que haya valores perdidos, outliers o incluso que encontremos combinaciones de nuevas variables que puedan ser interesante para el estudio.

* + 1. **Comprobar orden de la serie temporal**

En primer lugar, reviso el orden de los datos en función de la serie temporal en función de la columna “Datetime”, obteniendo como resultado que el dataset no estaba ordenado en función de la fecha.

Tras ordenarlo se tiene que la fecha inicial desde la que empieza a haber registros ahora en el head es el 1 de abril de 1998 a las 1:00 y la fecha fin es el 3 de agosto de 2018 a las 00:00



Ilustración 4:Primeras filas del dataset ordenado en función del tiempo

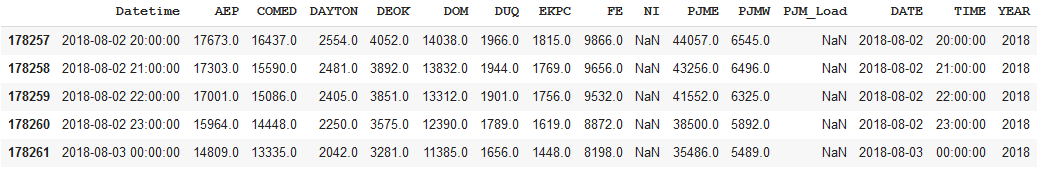


Ilustración 4: Últimas filas del dataset ordenado en función del tiempo

* + 1. **Comprobar registros perdidos**

Desde la fecha del primer registro a la fecha del último debería haber 178296 registros horarios.

En el dataset faltan los registros correspondientes a las horas '1998-04-05 03:00:00', '1998-10-25 02:00:00', '1999-04-04 03:00:00', '1999-10-31 02:00:00', '2000-04-02 03:00:00', '2000-10-29 02:00:00', '2001-04-01 03:00:00', '2001-10-28 02:00:00', '2002-04-07 03:00:00', '2002-10-27 02:00:00', '2003-04-06 03:00:00', '2003-10-26 02:00:00', '2004-04-04 03:00:00', '2004-10-31 02:00:00', '2005-04-03 03:00:00', '2005-10-30 02:00:00', '2006-04-02 03:00:00', '2006-10-29 02:00:00', '2007-03-11 03:00:00', '2007-11-04 02:00:00', '2008-03-09 03:00:00', '2008-11-02 02:00:00', '2009-03-08 03:00:00', '2009-11-01 02:00:00', '2010-03-14 03:00:00', '2010-11-07 02:00:00', '2010-12-10 00:00:00', '2011-03-13 03:00:00', '2011-11-06 02:00:00', '2012-03-11 03:00:00', '2012-11-04 02:00:00', '2013-03-10 03:00:00', '2013-11-03 02:00:00', '2014-03-09 03:00:00', '2015-03-08 03:00:00', '2016-03-13 03:00:00', '2017-03-12 03:00:00', '2018-03-11 03:00:00' las cuales he introducido en el dataframe inicializados con valor NaN en las columnas de Megavatios de las diferentes regiones para decidir en el siguiente apartado como imputar los diversos valores perdidos en función de diferentes criterios de fecha y tiempo.

* + 1. **Valores perdidos**

Una vez registradas todas las fechas que faltaban en el dataset reviso los valores perdidos desde la fecha de inicio en la que se empiezan a registrar valores hasta la fecha en la que se dejan de registrar en cada columna.

Tabla 1: Valores perdidos de cada columna

|  |  |  |  |
| --- | --- | --- | --- |
| **Columna** | **Primer registro** | **Último registro** | **Valores perdidos** |
| Datetime | 1998-04-01 01:00:00 | 2018-08-03 00:00:00 | 0 |
| AEP | 2004-10-01 01:00:00 | 2018-08-03 00:00:00 | 27 |
| COMED | 2011-01-01 01:00:00 | 2018-08-03 00:00:00 | 11 |
| DAYTON | 2004-10-01 01:00:00 | 2018-08-03 00:00:00 | 25 |
| DEOK | 2012-01-01 01:00:00 | 2018-08-03 00:00:00 | 9 |
| DOM | 2005-05-01 01:00:00 | 2018-08-03 00:00:00 | 23 |
| DUK | 2005-01-01 01:00:00 | 2018-08-03 00:00:00 | 24 |
| EKPC | 2013-06-01 01:00:00 | 2018-08-03 00:00:00 | 6 |
| FE | 2011-06-01 01:00:00 | 2018-08-03 00:00:00 | 10 |
| NI | 2004-05-01 01:00:00 | 2011-01-01 00:00:00 | 14 |
| PJME | 2002-01-01 01:00:00 | 2018-08-03 00:00:00 | 30 |
| PJMW | 2002-04-01 01:00:00 | 2018-08-03 00:00:00 | 30 |
| PJM\_Load | 1998-04-01 01:00:00 | 2002-01-01 00:00:00 | 8 |

Observando las fechas de inicio y fin de registro de datos de cada región una solución viable a los posibles valores perdidos es cambiar cualquier valor NaN que se encuentre en medio del intervalo de actividad de la región, realizando una interpolación al valor no nulo más cercano “nearest”.

Aplicando la interpolación se han rellenado todos los valores nulos para los periodos de actividad de cada columna.

Por último, para los valores NaN correspondientes a los tiempos de inactividad de cada región, se ha realizado una sustitución por 0.

* + 1. **Nuevos parámetros añadidos**

Se han creado las columnas DATE, TIME y YEAR a raíz de la columna “Datetime” para realizar análisis sobre períodos más específicos.

DATE tiene la estructura Año-mes-día y hace referencia a un día específico.

TIME tiene el formato Hora:Minutos:Segundos

YEAR hace referencia al año.

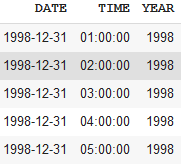


Ilustración 5: Nuevos campos de tiempo

Debido a que el consumo podría ser distinto en un día laborable que en un día de entre semana, se han creado 7 columnas categóricas con el día de la semana Monday, Tuesday, Wednesday, Thursday, Friday, Saturday y Sunday para realizar análisis posteriores. También se han creado las columnas Weekday Y Weekend que indican si es un día entre semana o fín de semana.



Ilustración 6: Nuevas columnas días de la semana

Las nuevas columnas Quarter\_year, Month\_year, Week\_year y Day\_year indican sucesivamente el trimestre, mes, semana y día del año en cada una de las filas del dataset. Están altamente relacionadas con la estación del año la cual influye en el consumo de energía.

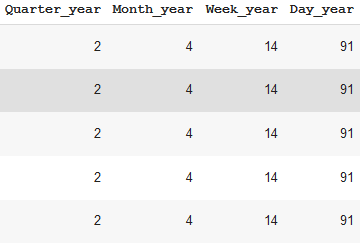


Ilustración 7: Nuevas columnas relacionadas con la época del año

* + 1. **Normalización**

Para la normalización he dividido cada una de las columnas que contienen información de Megavatios por el valor máximo en esa columna, con lo que se consigue que los valores de la columna oscilen entre 0 y 1.

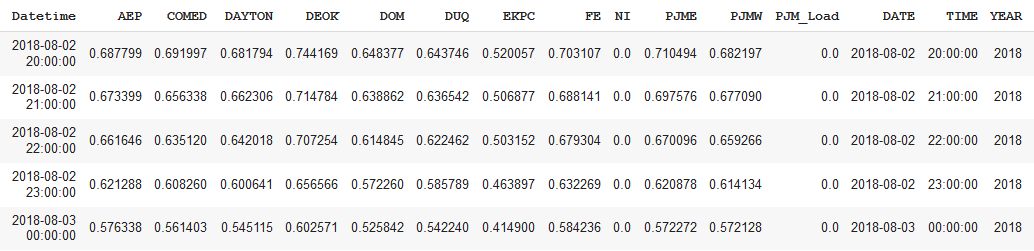


Ilustración 8: Normalización

1. **Análisis exploratorio.**

Cada región dispone de intervalo de tiempo donde registra datos de consumo (Véase la siguiente tabla 1)

En la siguiente gráfica se puede apreciar la evolución del consumo en cada región a través de los años. A priori no se aprecian outliers, pero se estudiará en profundidad más adelante.

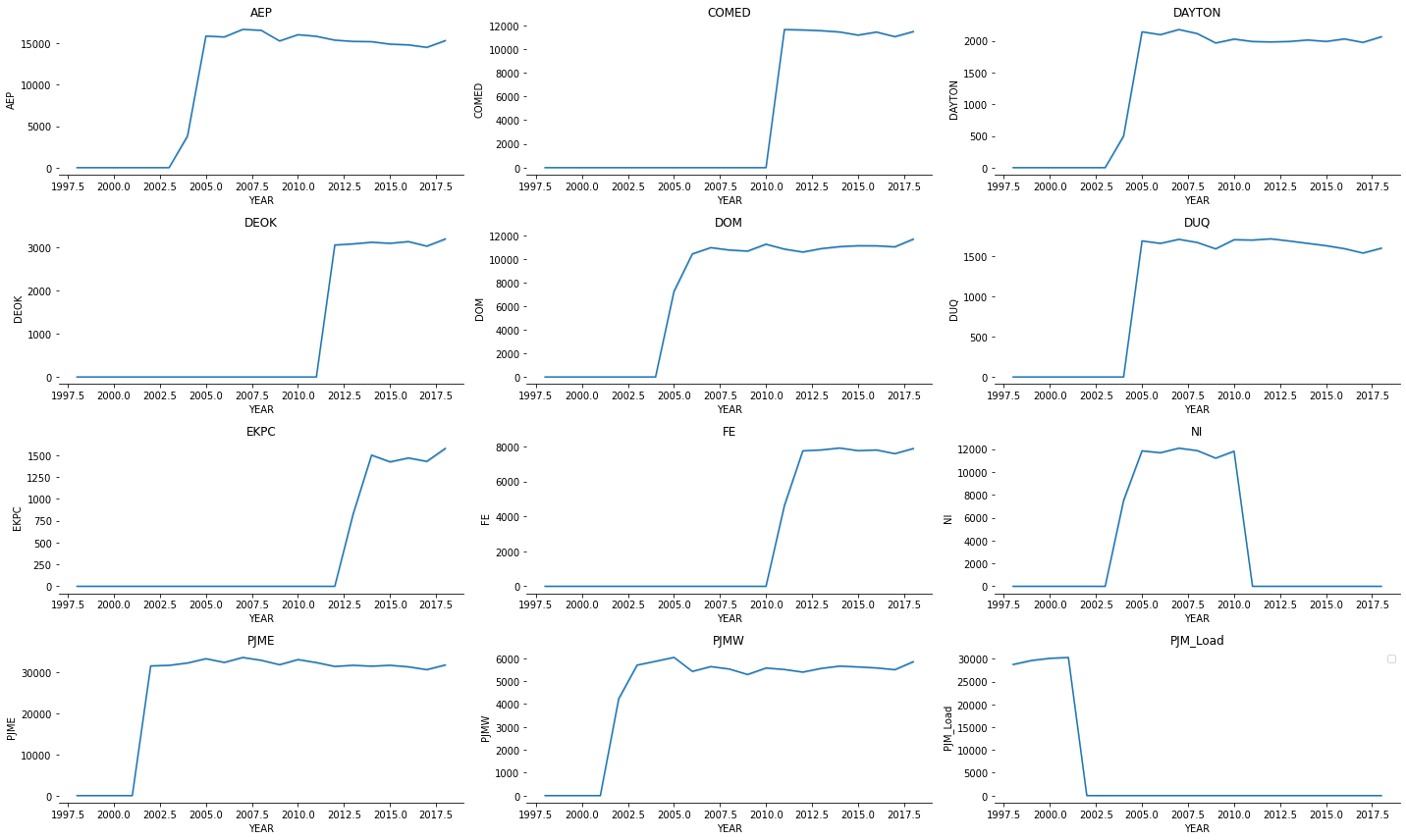


Ilustración 9: Estimación de consumo en las diferentes regiones

El consumo por trimestres devela un comportamiento dispar en las distintas regiones del del conjunto de datos.

El comportamiento en casi todas las regiones menos en “PJM\_Load” Y “NI” comienza el primer trimestre con un consumo elevado, que baja hasta el segundo trimestre, vuelve a subir en el tercero y baja en el cuarto.

Hay ciertos matices en este comportamiento como valor máximo alcanzado en el segundo y cuarto trimestre, siendo a veces superior en el segundo y otras en el cuarto.

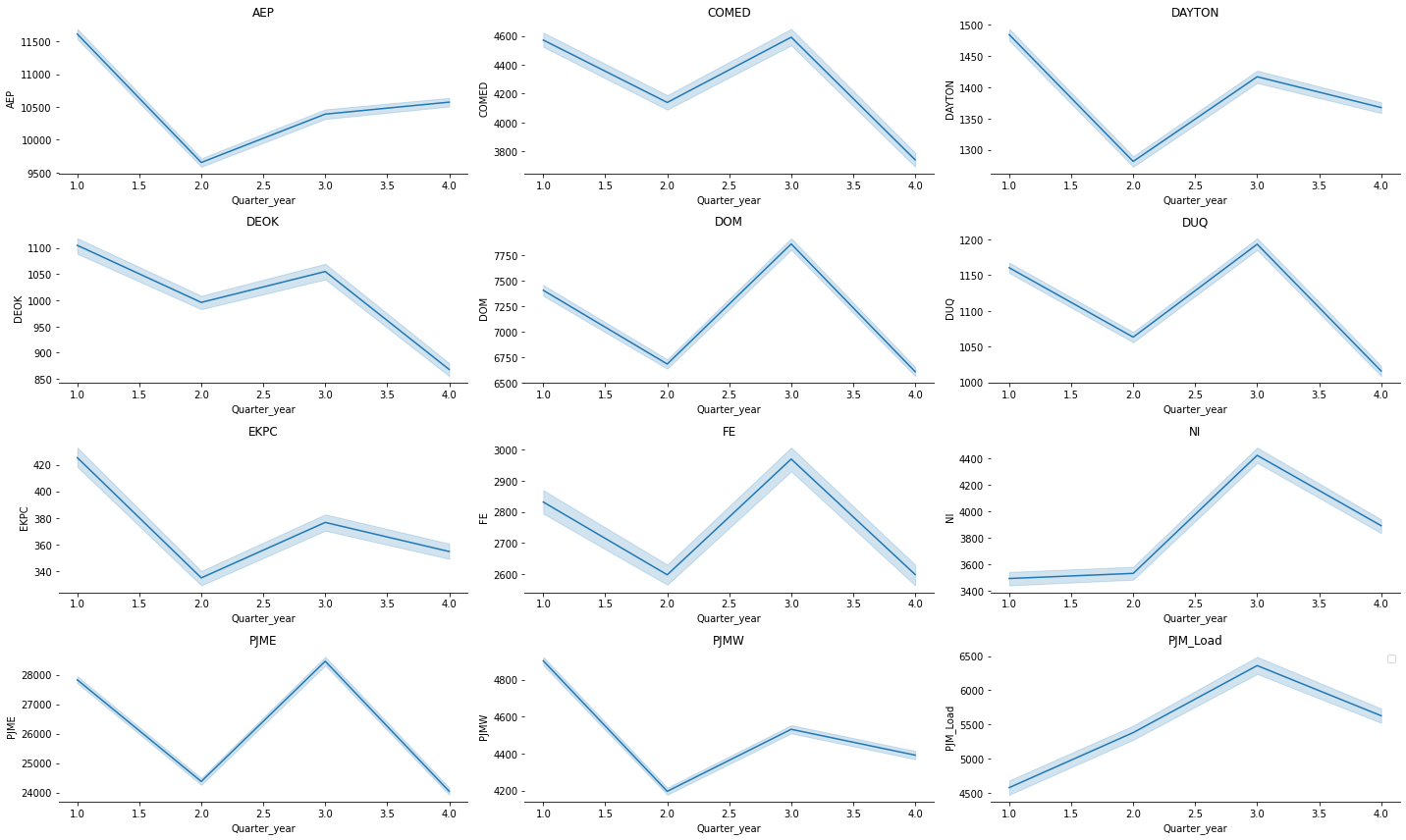


Ilustración 10: Consumo por trimestres

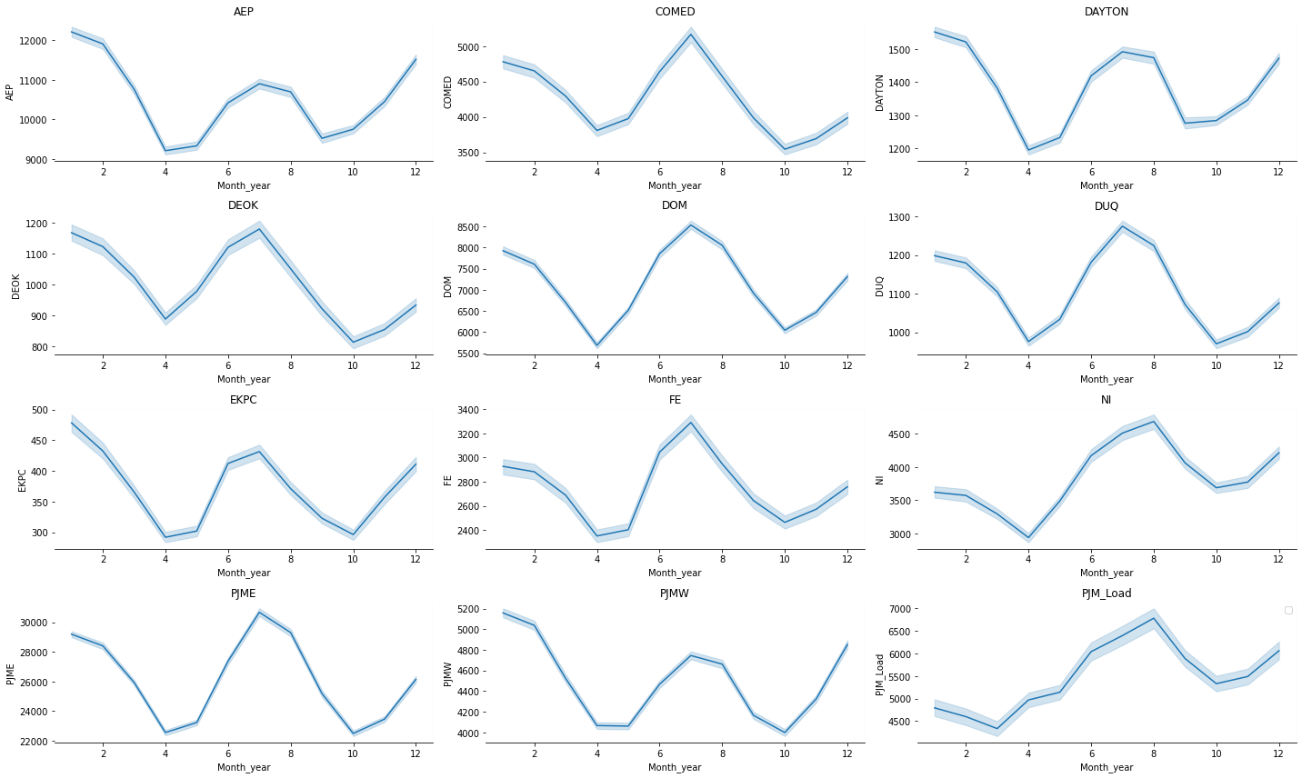
El comportamiento en de “NI” y “PJM\_Load” en los trimestres es de empezar en enero en el mínimo, subiendo hasta el tercer trimestre y bajando en el cuarto. El consumo mensual sigue la misma distribución que el trimestral, pero se puede apreciar con más detalle la tendencia.

Ilustración 11: Consumo por meses

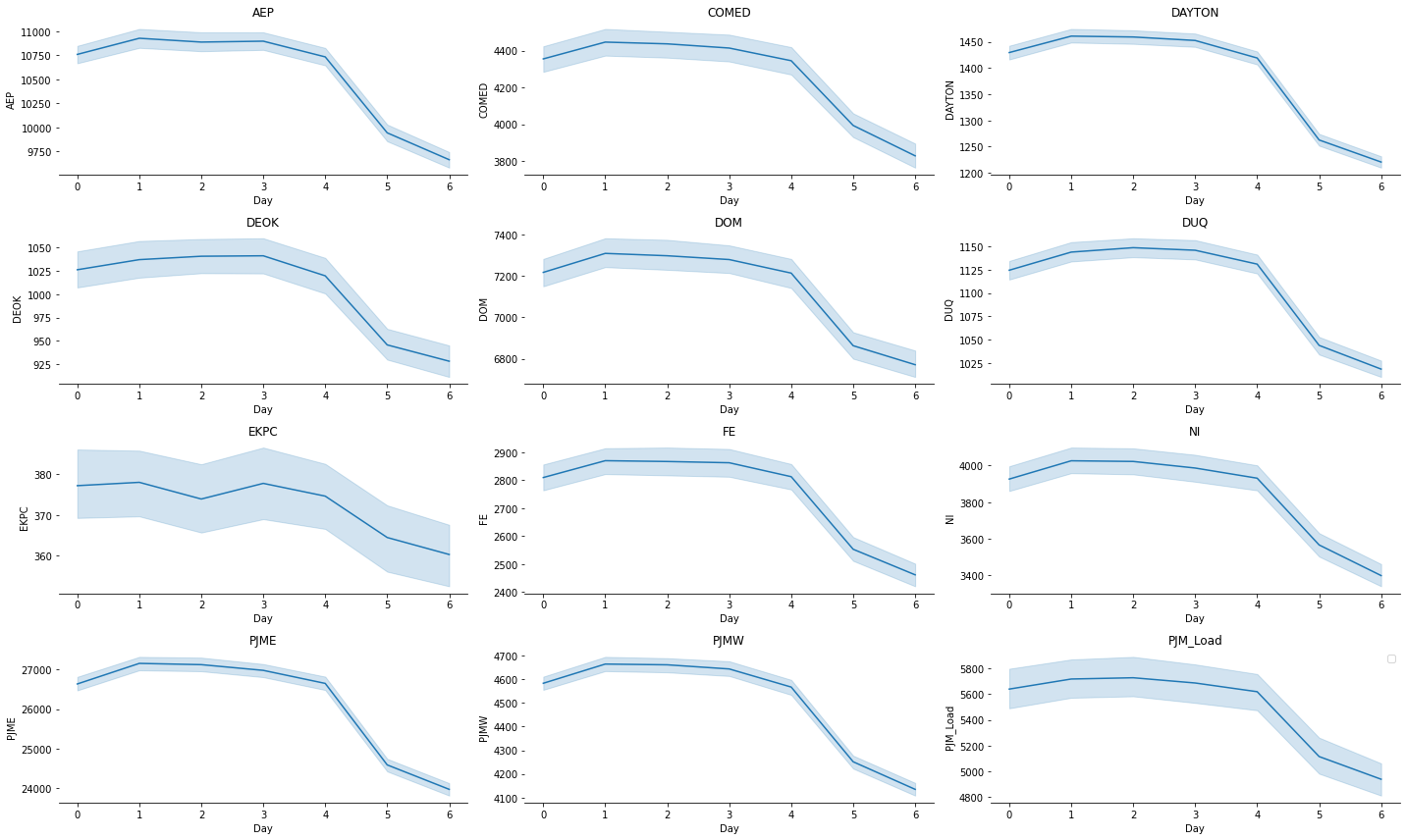
Al visualizar el consumo diario se aprecia que el consumo baja en fín de semana en todas las regiones.

Ilustración 12: Consumo diario 0=Lunes, …, 6=Domingo

Si observamos el consumo entre semana y en fin de semana observamos que en la mayoría de los años hay un consumo más alto entre semana. En 2017 parece que se iguala un poco el consumo en fin de semana.

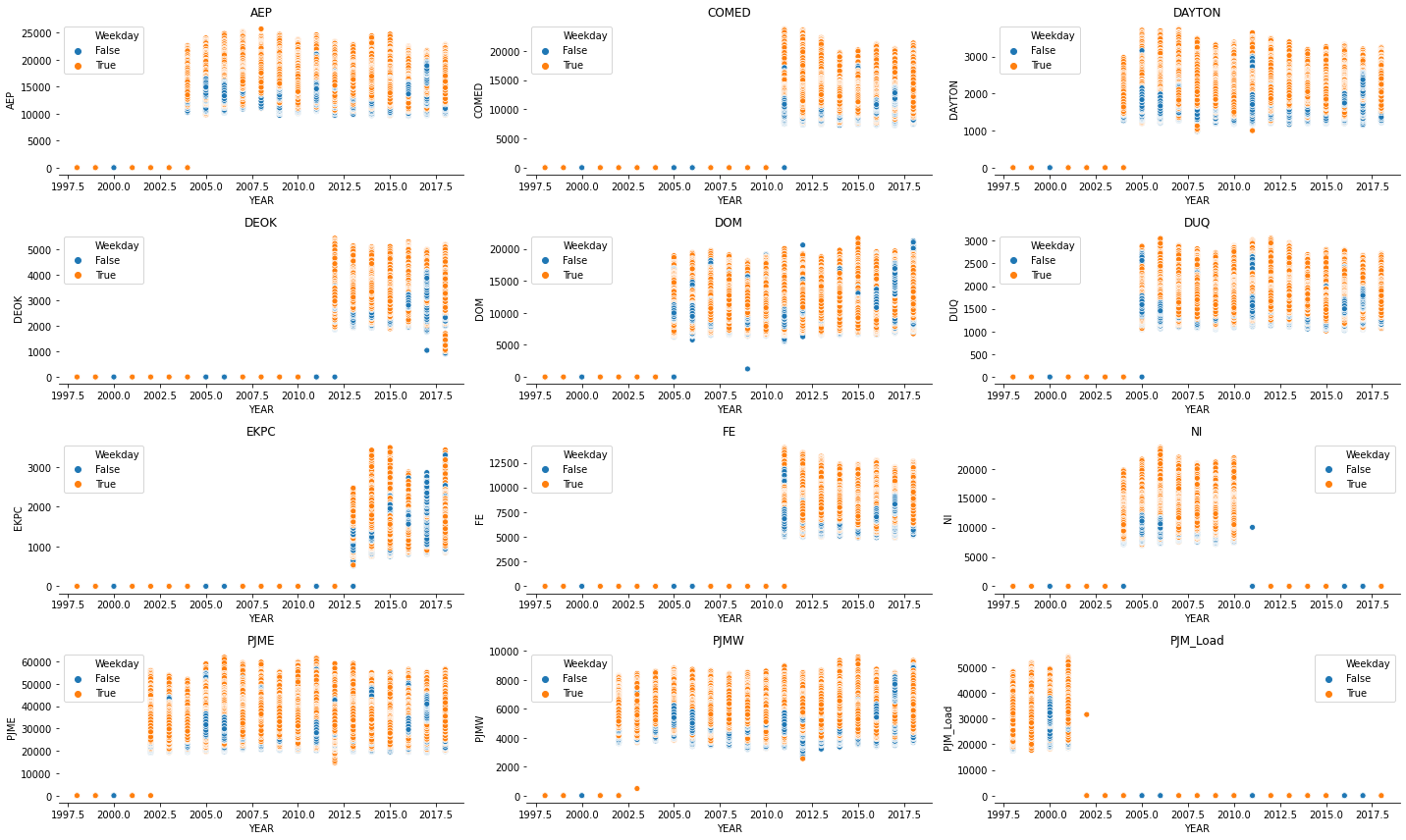
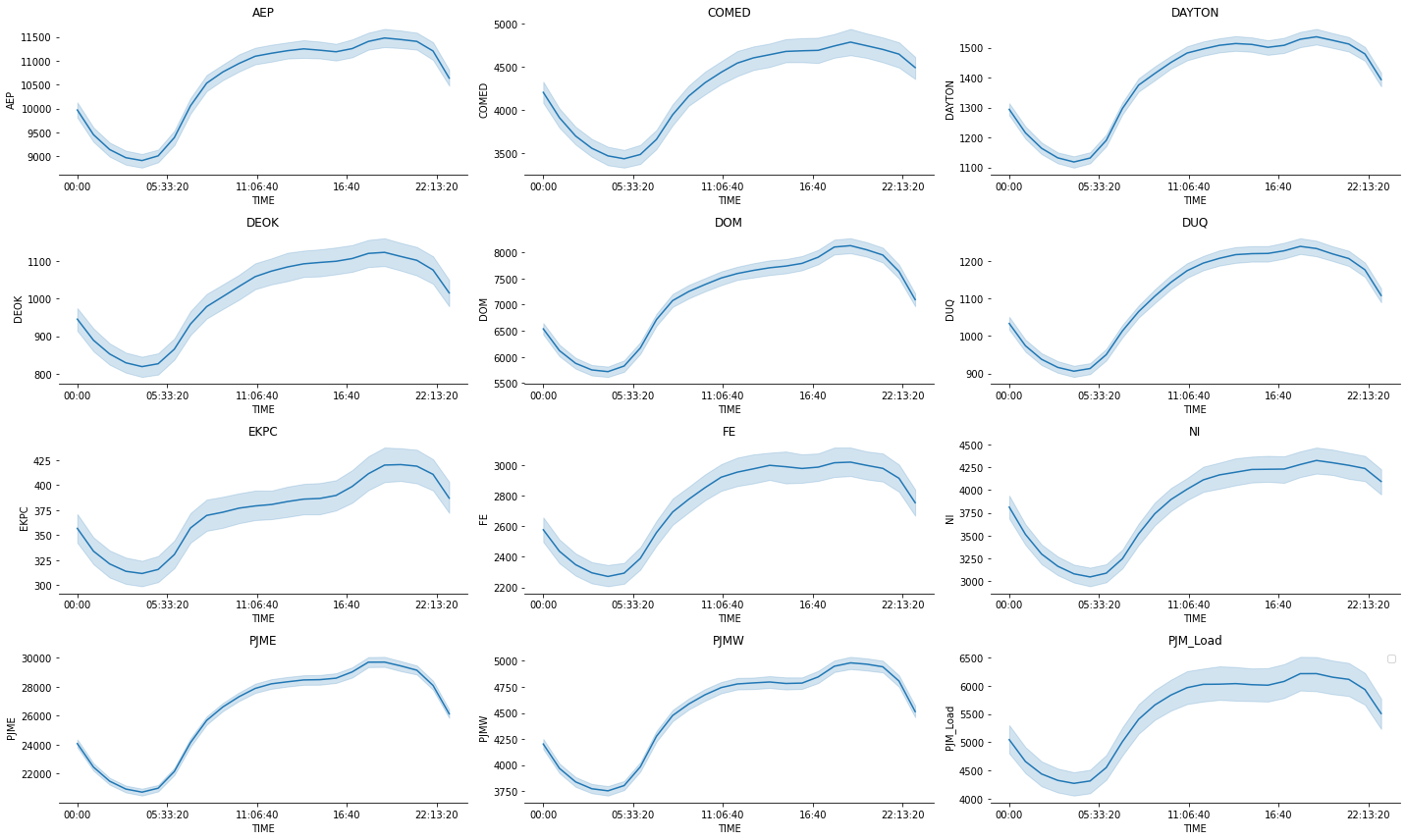


Ilustración 13: Consumo en Fin de semana anual

El consumo parece seguir un patrón horario similar en las diferentes regiones. Desde las 22:13:20 hasta las 5:33:20 se aprecia el pico con menor consumo, lo que parece lógico debido a que es de noche. Hay una subida hasta las 11:06:40 que se estabiliza a las16:40. El consumo se mantiene un poco hasta la bajada de las 22:13:20.

Ilustración 14: Estimación horaria del consumo en las diferentes regiones

Se procede a realizar un análisis de los principales estadísticos en las columnas en sus periodos de actividad y obtenemos la siguiente tabla.

Tabla 2: Principales estadísticos en el intervalo de fecha de actividad

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Región** | **Registros** | **Media** | **STD** | **Mínimo** | **25%** | **50%** | **75%** | **Máximo** |
| **AEP** | 121.300 | 15.499,04 | 2.591,44 | 9.581 | 13.629 | 15.309 | 17.200 | 25.695 |
| **COMED** | 66.508 | 11.419,81 | 2.304,11 | 7.237 | 9.779 | 11.151 | 12.510 | 23.753 |
| **DAYTON** | 121.300 | 2.037,76 | 393,42 | 982 | 1.749 | 2.009 | 2.279 | 3.746 |
| **DEOK** | 57.748 | 3.105 | 599,86 | 907 | 2.687 | 3.013 | 3.449 | 5.445 |
| **DOM** | 116.212 | 10.948,78 | 2.414 | 1.253 | 9.322 | 10.501 | 12.378 | 21.651 |
| **DUQ** | 119.092 | 1.658,76 | 301,76 | 1.014 | 1.444 | 1.630 | 1.819 | 3.054 |
| **EKPC** | 45.340 | 1.464,2 | 378,86 | 514 | 1.185 | 1.386 | 1.699 | 3.490 |
| **FE** | 62.884 | 7.791,93 | 1.331,3 | 0 | 6.806 | 7.700 | 8.556 | 14.032 |
| **NI** | 58.464 | 11.701,17 | 2.371,52 | 7.003 | 9.953 | 11.520 | 12.896 | 23.631 |
| **PJME** | 145.396 | 32.078,76 | 6.464,3 | 14.544 | 27.571 | 31.420 | 35.647 | 62.009 |
| **PJMW** | 143.236 | 5.602,19 | 979,18 | 487 | 4.906 | 5.530 | 6.252 | 9.594 |
| **PJM\_Load** | 32.904 | 29.764,36 | 5.850,61 | 17.461 | 25.471 | 29.653 | 33.071,25 | 54.030 |

Como conclusión sobre los estadísticos de conjunto de datos total podemos destacar:

* En el número de registros tenemos que las regiones AEP, DAYTON, PJME, PJMW son las que más tiempo de actividad han aportado al conjunto de datos, PJM\_Load y EKPC las que menos.
* La menos media de consumo se encuentra en la región EKPC, y la mayor PJME. PJM\_Load aún teniendo pocos registros en el conjunto de datos tiene una media elevada de consumo.
* Hay mucha diferencia de consumo entre las zonas Este y Oeste.
* Las columnas con más desviación son PJM\_Load, junto con AEP, COMED y NI. Hay más desviación en la zona Este que en la Oeste.
* La información de FirstEnergy proporciona un mínimo 0.0, quizás sea un posible outlier que investigaremos más adelante en los boxplot.
* Los mínimos más elevados se encuentran en AEP, PJME y PJM\_Load. EKPC Y PJMW son las columnas que contienen los mínimos más bajos.
* Las regiones con valores más altos son PJME, PJM\_Load, y AEP.

Los máximos y mínimos se encuentran en las siguientes fechas:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Región** | **Mínimo** | **Fecha mínimo** | **Máximo** | **Fecha máximo** |
| **AEP** | 9.581 | 2016-10-02 05:00:00 | 25.695 | 2008-10-20 14:00:00 |
| **COMED** | 7.237 | 2014-05-25 07:00:00 | 23.753 | 2011-07-20 18:00:00 |
| **DAYTON** | 982 | 2008-09-14 18:00:00 | 3.746 | 2007-08-08 15:00:00 |
| **DEOK** | 907 | 2018-04-22 05:00:00 | 5.445 | 2012-06-29 16:00:00 |
| **DOM** | 1.253 | 2009-12-12 00:00:00 | 21.651 | 2015-02-20 08:00:00 |
| **DUQ** | 1.014 | 2015-04-27 06:00:00 | 3.054 | 2012-06-29 14:00:00 |
| **EKPC** | 514 | 2013-09-29 04:00:00 | 3.490 | 2015-02-20 08:00:00 |
| **FE** | 0 | 2011-06-01 01:00:00 | 14.032 | 2011-07-21 16:00:00 |
| **NI** | 7.003 | 2005-05-29 08:00:00 | 23.631 | 2006-08-01 17:00:00 |
| **PJME** | 14.544 | 2012-10-30 04:00:00 | 62.009 | 2006-08-02 17:00:00 |
| **PJMW** | 487 | 2003-05-29 00:00:00 | 9.594 | 2015-02-20 08:00:00 |
| **PJM\_Load** | 17.461 | 1998-05-24 05:00:00 | 54.030 | 2001-08-09 15:00:00 |

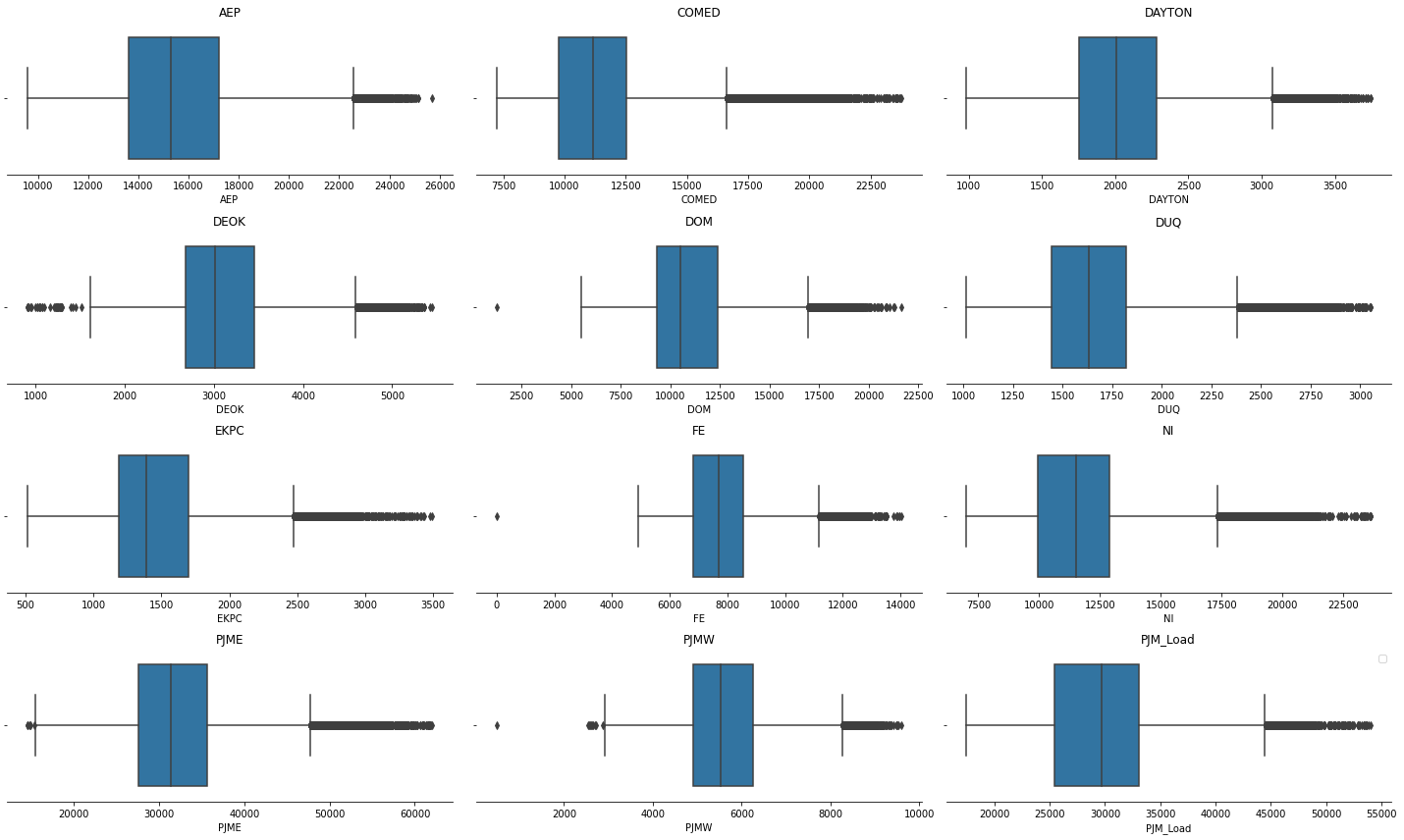
Procedo a revisar los diagramas de caja y bigotes para encontrar posibles outliers.

Ilustración 15: Diagramas boxplot

Se aprecian en las columnas DOM, FE Y PJMW posibles outliers con valor 0 que tras analizar en profundidad obtenemos que solo hay un outlier para la columna FE en la fecha 2011-06-01 01:00:00 que coincide con el inicio de los datos de FE, por lo que la fecha real inicial de los datos es “2011-06-01 02:00:00”

Una vez preparados los datos procedo a analizar la tendencia a lo largo de los años de cada una de las columnas de estimación de megavatios.

Se pueden observar regiones con tendencia decreciente como es el caso de AEP, DATYON o DUQ. Hay regiones con tendencia estable como DEOK o EKPC. DOM al igual que PJM\_Load tienen una tendencia creciente.

Ilustración 16: Regresión a través de los años 1

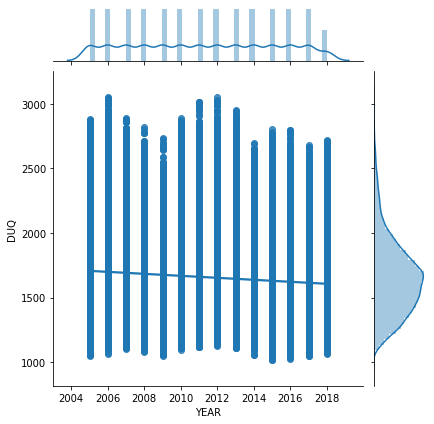
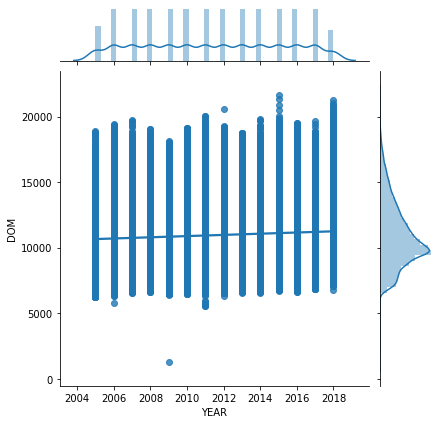
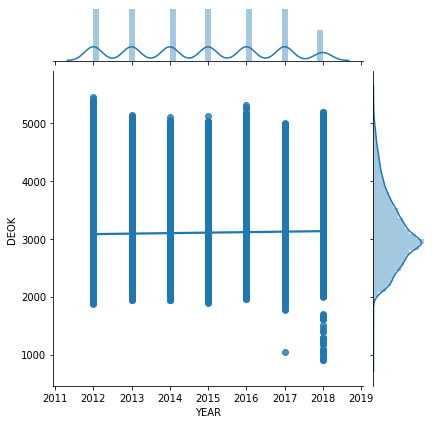
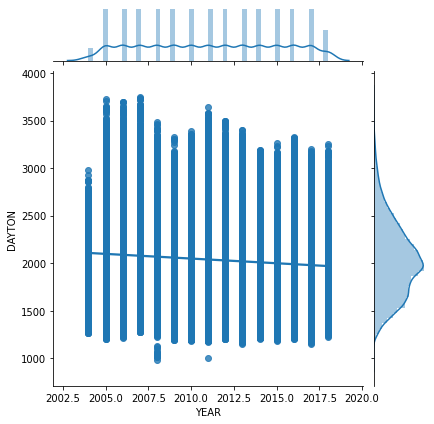
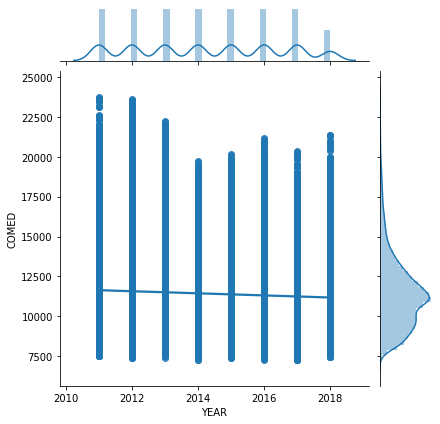
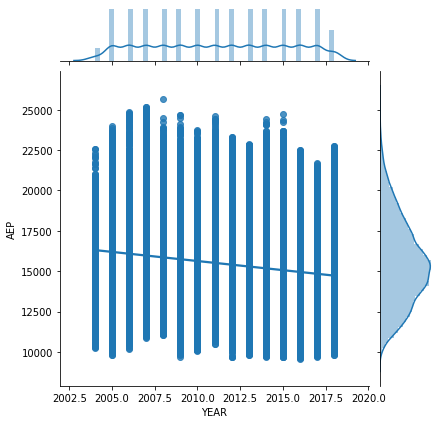
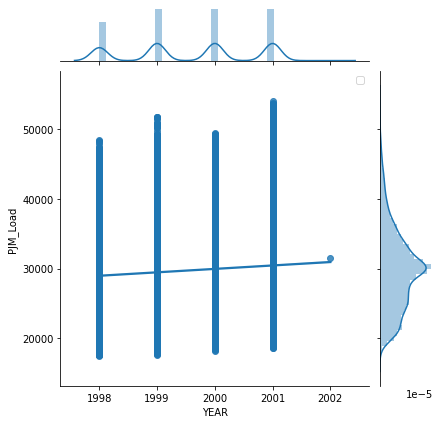
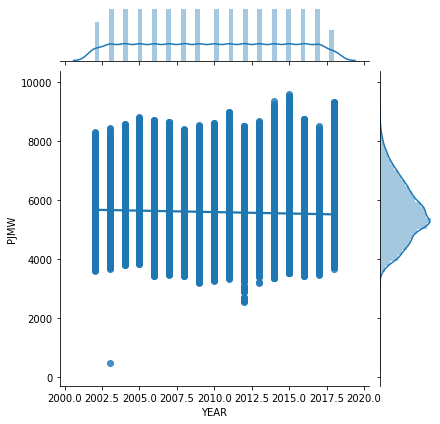
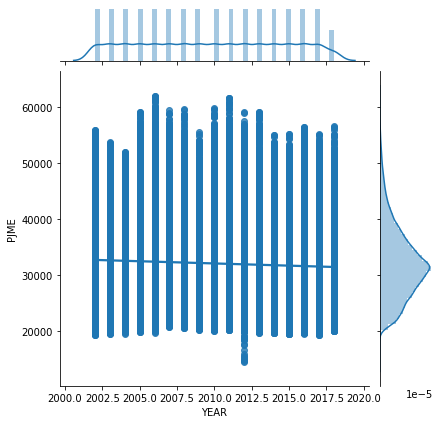
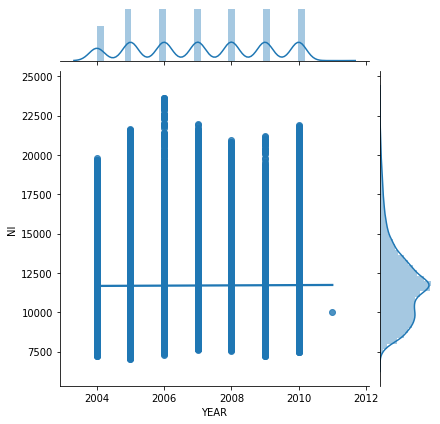
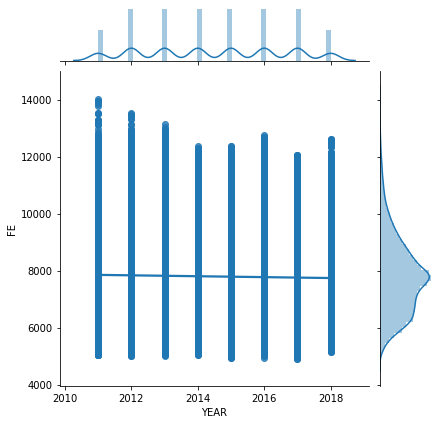
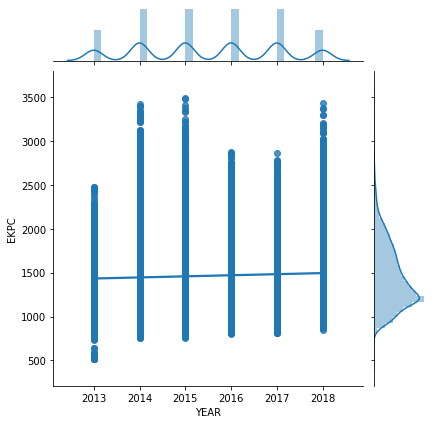


Ilustración 17: Regresión a través de los años 2



Parece que el consumo tiene una tendencia generalizada a decrecer o estabilizarse lo cual es bastante interesante dado que debido al crecimiento de la población y el crecimiento de la industria parece que esta tendencia debería ser positiva.

1. **Diseño e implementación de los modelos o técnicas necesarios.**
   1. Series temporales
   2. Machine Learning
   3. Redes neuronales
      1. Google Colab
      2. Tensorflow & Keras
      3. Redes recurrentes
      4. LSTM
      5. GRU
   4. Algoritmos metaheurísticos
      1. Algoritmo de colonia artificial de abejas
   5. Hadoop
      1. HDFS
      2. YARN
      3. Map Reduce
   6. Spark
      1. PySpark
      2. Spark SQL
      3. SparkML
   7. Esquema general
2. **Análisis de los resultados obtenidos.**
3. **Conclusiones y planes de mejora.**
4. **Bibliografía.**
5. **Anexo con el código fuente desarrollado.**

1. [Dataset](https://www.kaggle.com/robikscube/hourly-energy-consumption) [↑](#footnote-ref-1)
2. [PJM](https://www.pjm.com/) [↑](#footnote-ref-2)
3. [AEP](https://en.wikipedia.org/wiki/American_Electric_Power) [↑](#footnote-ref-3)
4. [COMED](https://en.wikipedia.org/wiki/Commonwealth_Edison) [↑](#footnote-ref-4)
5. [DAYTON](https://en.wikipedia.org/wiki/DPL_Inc.) [↑](#footnote-ref-5)
6. [DUKE](https://en.wikipedia.org/wiki/Duke_Energy) [↑](#footnote-ref-6)
7. [DOM](https://en.wikipedia.org/wiki/Dominion_Energy) [↑](#footnote-ref-7)
8. [DUQ](https://en.wikipedia.org/wiki/DQE) [↑](#footnote-ref-8)
9. [EKPC](http://www.ekpc.coop/) [↑](#footnote-ref-9)
10. [FE](https://en.wikipedia.org/wiki/FirstEnergy) [↑](#footnote-ref-10)