

# M2.851 Tipología y ciclo de vida de los datos

## Práctica 1: Web scraping

### Matriculaciones mensuales de turismos en España desde el 2019 hasta la actualidad

Componentes: Carlos Lavado Mahia & Dionisio González Jiménez

Profesor: Diego Perez Trenard

Fecha de elaboración: 12/04/2021

#### 1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

En 2020, el virus SARS-COVID19 afectó a los ingresos de explotación de numerosos sectores, entre ellos el de la **automoción**; el cual tiene una importancia estratégica en España. Concretamente, dicho sector supone aproximadamente un **11 % del PIB Nacional**, considerando todas las ramas de la industria; como pueden ser proveedores de componentes, fabricantes de automóviles y camiones, o concesionarios. Sin embargo, todas ellas se vieron afectadas por las **caídas generales en la demanda de mercado**, así como por las medidas urgentes de paralización de actividades no esenciales tomadas en el año 2020.

Dado que el sector de la automoción dispone de numerosas ramas, hemos decidido optar por ver la evolución de los datos de matriculación de turismos, relativos a la venta de coches al consumidor final. Esto es, la rama de **concesionarios y establecimientos tradicionales de venta de vehículos**. En concreto, dicha rama aporta aproximadamente un 3% del PIB nacional, generando 153.425 empleos directos, así como 459.000 empleos indirectos.

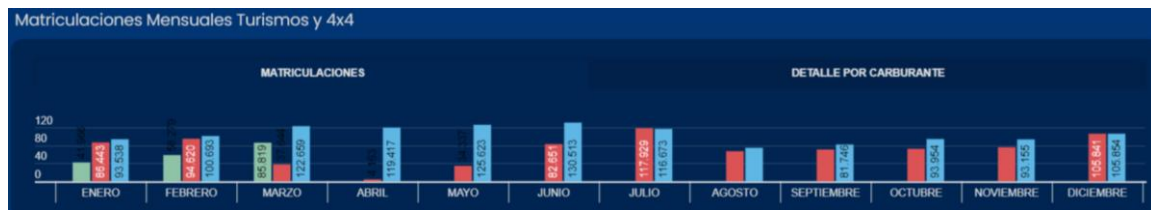
Con el objetivo de estudiar el impacto de la pandemia en las ventas de turismos a través de concesionarios, hemos accedido a la web correspondiente a la patronal de concesionarios de España, conocida como **Faconauto**. En concreto, 2.042 concesionarios españoles se encuentran registrados en esta patronal, acumulando un total de 5.309 puntos de venta.

En términos generales, Faconauto estima una facturación total en el sector de 43.073 millones de euros en 2019, que habría descendido a los 35.895 millones de euros en 2020. Esto representaría un descenso anual de, aproximadamente, un 17%.

En este trabajo, nos planteamos **las siguiente hipótesis**: ¿Cuál ha sido el impacto de cada mes en la evolución general de la facturación? ¿Se han mantenido los patrones de evolución intermensual de años anteriores? ¿Cómo ha afectado la pandemia al número de turismos matriculados?

Con el objetivo de responder estas preguntas -de manera directa o indirecta-, hemos accedido a la sección de "Estadísticas" de la web de la patronal Faconauto, concretamente a los datos relativos a **matriculaciones mensuales de turismos**. En esta URL, podemos encontrar un gráfico con información relativa a los turismos matriculados por los concesionarios registrados en la patronal, a lo largo de los años 2019, 2020 y 2021 (hasta marzo):

Ilustración 1. Matriculaciones mensuales de turismos en 2019, 2020 y 2021 (hasta marzo) - Faconauto



Fuente: Faconauto.com. <https://www.faconauto.com/matriculaciones-mensuales-turismos/>

La web **no dispone de ninguna interfaz API** para extraer estos datos.

Sin embargo, si examinamos el archivo `robots.txt` de la web de la patronal, podemos observar que **es posible realizar web scraping** de datos con cualquier agente, siempre que se deje un delay de 60 segundos.

Ilustración 2. Output del archivo `robots.txt` de faconauto.com

```
User-agent: *  
Crawl-delay: 60  
User-agent: MJ12bot  
Disallow:
```

Fuente: Faconauto.com. <https://www.faconauto.com/robots.txt>

Con la intención de poder generar un fichero para calcular las variaciones interanuales, así como la evolución mensual durante cada año, hemos investigado la forma de extraer los datos mostrados en la *Ilustración 1* mediante el lenguaje de programación Python, así como las librerías disponibles.

De este modo, **a través de una araña o web scraper**, podremos responder a las hipótesis planteadas anteriormente.

Referencias empleadas:

[La Automoción ibérica reivindica su puesto como sector estratégico de la economía \(posventa.info\)](#)

[Sector - Faconauto](#)

[FACONAUTO-Interactivo.pdf](#)

## 2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Nuestro objetivo es evaluar el impacto que ha tenido el SARS-COVID19 en el sector de la automoción y en concreto la matriculación mensual de los turismos en este período de pandemia, considerando a modo comparativo datos desde 2019. Por ello, el título para nuestro dataset y que supone un carácter descriptivo del mismo es:

**Matriculaciones mensuales turismos España 2019-actualidad**

**3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

En nuestro dataset, primeramente hemos incluido los **datos de matriculación mensual de turismos**, extraíbles de la web de Faconauto, correspondientes al **periodo 2019-actualidad**. Esto ha sido logrado a través de un proceso de **web scraping** (principalmente con las librerías BeautifulSoup y Selenium), convirtiendo más tarde los datos en un **dataframe** de la librería pandas. A partir de este primer paso, hemos obtenido tres columnas, correspondientes a las matriculaciones mensuales del año 2019, 2020 y 2021 (primer trimestre).

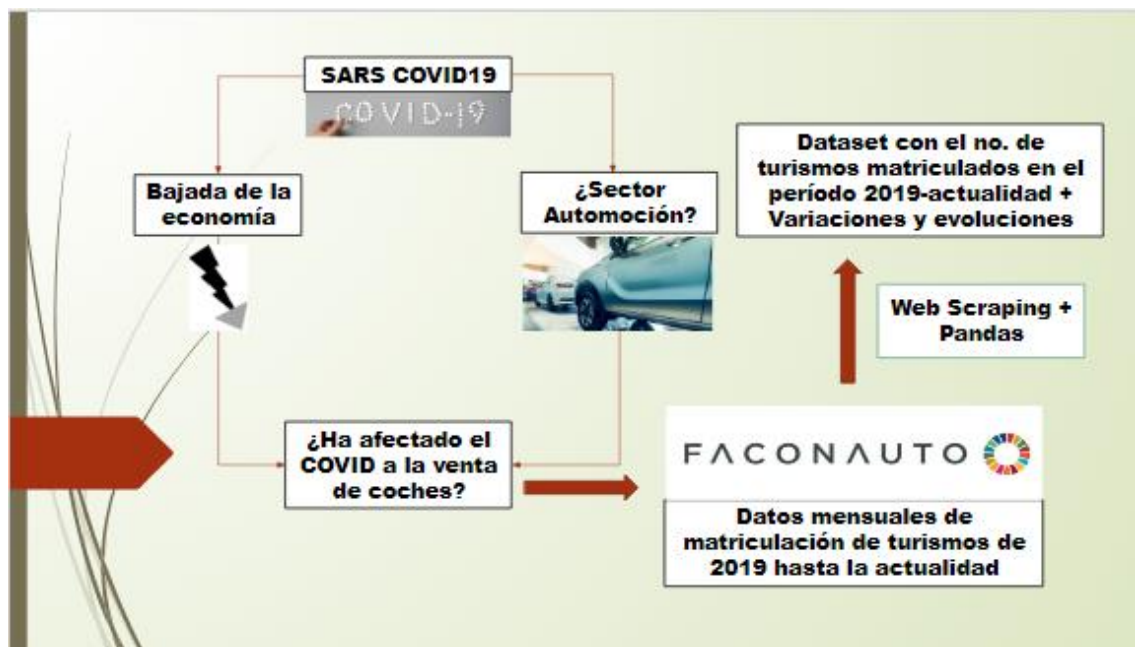
Por otro lado, también hemos aprovechado las funcionalidades de la librería pandas para crear nuevas columnas a través de las ya existentes. En concreto, hemos creado **columnas de variación interanual y evolución intermensual** para los años 2019, 2020 y 2021.

Más adelante, también hemos calculado las **diferencias en evoluciones intermensuales** entre 2019 y 2020 (así como entre 2019 y 2021); con el objetivo de poderse examinar cuál ha sido el impacto de la pandemia en los patrones de crecimiento durante el año natural.

Según nuestro punto de vista, y dada la naturaleza y fuente de estos, los datos extraídos cumplen con los **requisitos** de completitud, unicidad, puntualidad, validez, exactitud y consistencia.

**4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.**

*Ilustración 3. Esquema representativo del proyecto elegido para web scraping.*



*Fuente: elaboración propia, e imágenes extraídas de Faconauto.com*

**5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

Como visión general, el dataset consiste en las matriculaciones de turismos que se han realizado durante los últimos ejercicios (2019, 2020 y 2021) y su comparativa. El título que le hemos dado es *matr\_turismos.csv*.

Para recoger dicho CSV, nos hemos apoyado de la librería **pandas**, una vez realizado el *scraping*. A través de esta librería, **hemos transformado el diccionario en el que teníamos los diferentes meses y valores almacenados, convirtiéndolos en un objeto dataframe**. Una vez generado el *dataframe*, hemos podido crear las columnas accesorias de variaciones interanuales por mes, y de evolución intermensual; haciendo cálculos con las columnas originales, como puede verse al final del código.

Estos datos se han extraído en un archivo csv con 11 columnas:

- **Cabecera**→ nos muestra los meses del año con un período de ENERO a DICIEMBRE. Tipo de dato *string*.
- **2021**→ datos reales del número de turismos matriculados durante el año 2021 dividido por meses; como se trata del año actual, sólo tenemos datos para los meses de ENERO, FEBRERO y MARZO. Por lo que no serán muy relevantes. Utiliza un tipo de dato *float*, a pesar de que los datos originales sean números enteros. Esto se debe, a que la columna registra valores NaN para los meses aún no acontecidos, los cuales son interpretados con el tipo *float* por Python.
- **2020**→ datos reales del número de turismos matriculados durante el año 2020 dividido por meses, en el período ENERO a DICIEMBRE con un tipo de dato *integer*.
- **2019**→ datos reales del número de turismos matriculados durante el año 2019 dividido por meses, en el período ENERO a DICIEMBRE con un tipo de dato *integer*.
- **Variación 2019-2020**→ tasa de variación interanual dividida en meses en el período 2019-2020. Tipo de datos *float*.
- **Variación 2019-2021**→ tasa de variación interanual dividida en meses en el período 2019-2021. Tipo de datos *float*.
- **Evolución 2019**→ evolución intermensual respecto al mes anterior, dividida por meses en el período del año 2019. Tipo de datos *float*.
- **Evolución 2020**→ evolución intermensual respecto al mes anterior, dividida por meses en el período del año 2020. Tipo de datos *float*.
- **Evolución 2021**→ evolución intermensual respecto al mes anterior, dividida por meses en el período del año 2021. Tipo de datos *float*.
- **Difs evolución 19-20**→ diferencias de evolución intermensual respecto al mes anterior, comparativa entre 2019 (base) y 2020. Tipo de datos *float*.
- **Difs evolución 19-21**→ diferencias de evolución intermensual respecto al mes anterior, comparativa entre 2019 (base) y 2021. Tipo de datos *float*.

*Ilustración 4. Representación del dataset matr\_turismos.csv.*

El dataset final es el siguiente:

	2021	2020	2019	Variación 2019-2020	Variación 2019-2021 \
ENERO	41966.0	86443	93538	-0.08	-0.55
FEBRERO	58279.0	94620	100693	-0.06	-0.42
MARZO	85819.0	37644	122659	-0.69	-0.30
ABRIL	NaN	4163	119417	-0.97	NaN
MAYO	NaN	34337	125623	-0.73	NaN
JUNIO	NaN	82651	130513	-0.37	NaN
JULIO	NaN	117929	116673	0.01	NaN
AGOSTO	NaN	66925	74424	-0.10	NaN
SEPTIEMBRE	NaN	70729	81746	-0.13	NaN
OCTUBRE	NaN	72228	93954	-0.23	NaN
NOVIEMBRE	NaN	75708	93155	-0.19	NaN
DICIEMBRE	NaN	105841	105854	-0.00	NaN

	Evolución 2019	Evolución 2020	Evolución 2021 \
ENERO	NaN	NaN	NaN
FEBRERO	0.08	0.09	0.39
MARZO	0.22	-0.60	0.47
ABRIL	-0.03	-0.89	NaN
MAYO	0.05	7.25	NaN
JUNIO	0.04	1.41	NaN
JULIO	-0.11	0.43	NaN
AGOSTO	-0.36	-0.43	NaN
SEPTIEMBRE	0.10	0.06	NaN
OCTUBRE	0.15	0.02	NaN
NOVIEMBRE	-0.01	0.05	NaN
DICIEMBRE	0.14	0.40	NaN

	Difs evolución 19-20	Difs evolución 19-21
ENERO	NaN	NaN
FEBRERO	0.02	0.31
MARZO	-0.82	0.25
ABRIL	-0.86	NaN
MAYO	7.20	NaN
JUNIO	1.37	NaN
JULIO	0.53	NaN
AGOSTO	-0.07	NaN
SEPTIEMBRE	-0.04	NaN
OCTUBRE	-0.13	NaN
NOVIEMBRE	0.06	NaN
DICIEMBRE	0.26	NaN

*Fuente: elaboración propia*

**6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.**

El propietario de los datos es **Faconauto**, una patronal de concesionarios oficiales de turismos y maquinaria agrícola en el mercado español. Como ya se ha explicado anteriormente, Faconauto cuenta con 2.042 concesionarios oficiales asociados.

La misión principal de Faconauto es la de “*representar, promover y fomentar la competitividad de los concesionarios*”. En este sentido, la patronal impulsa varias iniciativas del sector ante las Administraciones Públicas, con tal de propiciar el entorno competitivo de los concesionarios españoles. Además, también acompaña a estos en sus procesos de transformación digital o de transición sostenible.

Queremos agradecer a Faconauto por ofrecer los datos relativos a matriculaciones mensuales de turismos en su página web, y por permitirnos realizar web scraping a través del archivo robots.txt incrustado en esta.

En referencia a los análisis anteriores realizados, queremos destacar que **Faconauto ya elabora notas de prensa para comentar los resultados de los datos que recopila.**

A modo de ejemplo, en la nota de prensa *“El mercado de vehículos en España cierra 2020 con una fuerte caída del 32,3%”*, la patronal ya comenta que **las unidades matriculadas registradas han descendido un 32,3% en comparación al año anterior.** Además, esta nota de prensa también nos informa de que la diferencia interanual de matriculaciones en el **mes de diciembre** es escasa, lo que nos indicaría que el impacto de la crisis ya se estaría revirtiendo.

Fuente: [El mercado de vehículos cierra 2020 con una caída del 32,3% - Faconauto](#)

En otras notas de prensa, Faconauto también nos informa de los datos mensuales de otros meses intermedios. A modo de ejemplo, la siguiente nota de prensa recopila la **evolución de las matriculaciones de turismos en septiembre de 2020**, comparándolas incluso con 2019 como haremos en nuestro dataset: [Las matriculaciones de vehículos en septiembre - Faconauto](#)

Sin embargo, todos estos datos **no se encuentran recopilados en un mismo lugar**, y por ello se hace muy difícil disponer de un gráfico interpretable que ayude a generar conclusiones generales sobre el año en cuestión.

Existen otras asociaciones que aprovechan los datos de Faconauto, como es el caso de **ANFAC - Asociación Española de Fabricantes de Automóviles y Camiones**. Esta asociación toma prestadas muchas de las indagaciones de Faconauto para sus propias notas de prensa, como puede ser observado en la siguiente noticia: [ANFAC | El mercado de vehículos en España cierra 2020 con una fuerte caída del 32,3%](#)

ANFAC se aproxima más a nuestros objetivos para el presente análisis, al mostrar en su página web un **grafo con la evolución mensual de las matriculaciones de turismos y todoterrenos**: [ANFAC | Matriculaciones Turismos y Todoterreno](#)

Sin embargo, este gráfico **no puede satisfacer todas las necesidades que tratamos de solventar** con nuestro análisis ya que:

- 1) Se trata de un gráfico dinámico, que tan solo incluye información de los 12 meses más recientes.
- 2) No incluye las variaciones en la evolución intermensual de año a año (por ejemplo, enero-febrero 2019 vs enero-febrero 2020), siendo este uno de los principales objetivos de nuestro análisis.

Por tanto, consideramos que nuestro conjunto de datos solventa la falta de un **resumen más directo de las relaciones de matriculaciones entre 2019 y 2020**, obteniéndose un dataset que puede reunir todas las tasas de variación interesantes para un análisis del sector.

Además, con el objetivo de potenciar la reutilización del código, nuestro CSV podrá considerar múltiples periodos, llegando incluso a almacenar las **primeras estimaciones de 2021**. En este sentido, consideramos que podría ser interesante investigar si las matriculaciones de turismos se están recuperando en 2021, en relación a las matriculaciones de 2019.

**7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.**

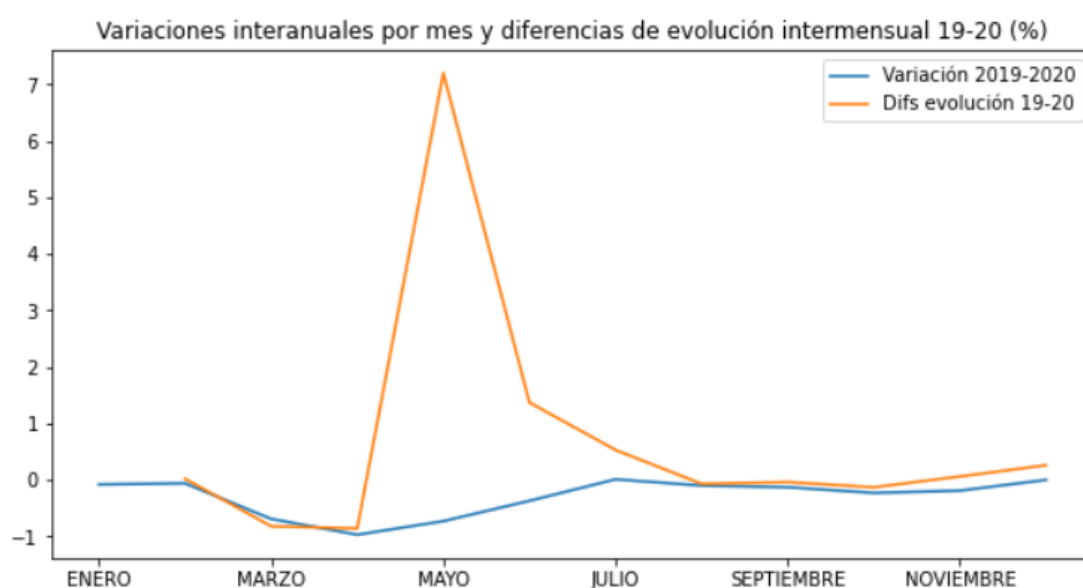
Como ya se ha introducido en el apartado anterior, este conjunto de datos resulta sumamente interesante para **poder recopilar todos los datos de matriculaciones mensuales de turismos; así como las variaciones interanuales e intermensuales de dichos datos en 2019 y 2020, en un solo lugar** (ya que sino, sería necesario hacer cálculos a mano en una *spreadsheet*, o ir consultando las diferentes notas de prensa de Faconauto o de asociaciones como ANFAC). Además, también podemos incluir en nuestro código los **primeros datos de 2021**, para estudiar la recuperación del sector.

Las **preguntas que se pretenden responder** son principalmente las siguientes:

- ¿Ha afectado la pandemia del COVID-19 a las matriculaciones de turismos?
- ¿Cuál ha sido la variación interanual, mes a mes, de las matriculaciones de turismos entre 2019 y 2020?
- Si consideramos la evolución de mes a mes durante el año, ¿han sido también menores estas evoluciones en el ejercicio 2020, en comparación al 2019?
- ¿Se puede entrever una recuperación en el 2021?

Con el objetivo de encontrar una respuesta a las tres primeras preguntas, puede encontrarse en el código una **representación gráfica de las variaciones interanuales 2019-2020, así como las diferencias en evolución intermensual de matriculaciones de turismos entre los dos años**:

*Ilustración 5. Variaciones interanuales por mes y diferencias de evolución intermensual 19-20 (%)*



*Fuente: elaboración propia*

Observamos a través del gráfico que la **tasa de variación** es negativa durante todo su período, ya que obviamente la crisis del COVID provocó un descenso en el año 2020 de las matriculaciones de vehículos. La recta es bastante constante, aunque se puede destacar un punto de descenso para el mes de Abril, momento en el que comenzaron los efectos de esta gran crisis.

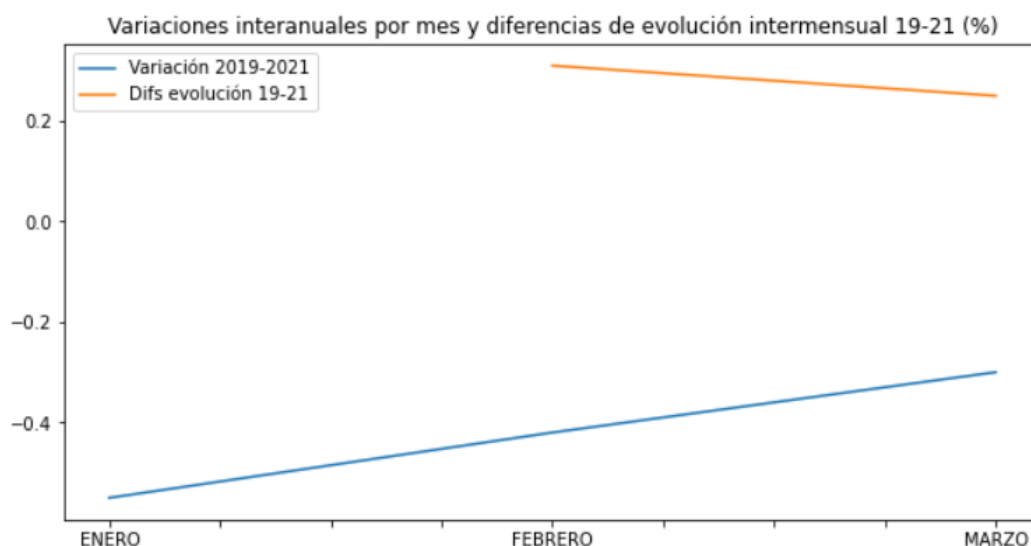
Por otro lado, se muestra la **diferencia entre la evolución intermensual** entre el año 2020 y 2019, con un gran descenso en la curva para el mes marzo-abril y posteriormente una gran subida indicada en el gráfico con un gran pico en el mes de mayo (720%). Una de las causas de esta gran diferencia fue debido a que en 2020 (año COVID), **en los meses de mayo, junio y julio mejoró la economía** gracias a la reducción de restricciones; y si esto se compara con meses anteriores (meses en los que surge el COVID), podemos decir que la diferencia es bastante destacable.

Por último, a pesar de toda la situación económica, nos gustaría destacar que **el mes de diciembre es bastante similar en comparativa del período**; una de las posibles causas, es debido a la liquidación de vehículos en este mes de cierre para sacar los nuevos modelos a principios de año. A pesar de la bajada de la economía, no ha habido una gran diferencia de 2020 a 2019.

En cuanto a la última pregunta, cabe destacar que ha sido posible identificar los datos de **matriculaciones mensuales de Faconauto en el primer trimestre de 2021**. Si bien se trata de unos primeros datos estimatorios, así favorecemos la reutilización del código en meses posteriores; en los cuales habrán más datos almacenados.

Con estos datos del primer trimestre de 2021, podemos observar que, **si bien sigue habiendo menos matriculaciones que en 2019; el ritmo de evolución intermensual también es mucho mayor**, con unos incrementos intermensuales de más del 20% mayores. Esto puede estudiarse en el siguiente gráfico:

*Ilustración 6. Variaciones interanuales por mes y diferencias de evolución intermensual 19 – 21 (%)*



*Fuente: elaboración propia*

**8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:**

- ☐ Released Under CC0: Public Domain License
- ☐ Released Under CC BY-NC-SA 4.0 License
- ☐ Released Under CC BY-SA 4.0 License
- ☐ Database released under Open Database License, individual contents under Database Contents License
- ☐ Other (specified above)
- ☐ Unknown License

Para nuestro proyecto, hemos seleccionado la licencia **CC0: Public Domain License**, cuyas condiciones pueden ser consultadas en el documento *LICENSE* del repositorio, creado a través de la plantilla de GitHub.

En primer lugar, queremos destacar que estos no son datos propios, sino que corresponden a una organización tercera, concretamente a una patronal. Es por eso, que **no consideramos necesario que se nos atribuya ningún reconocimiento** por la utilización de estos datos (que vendría dado por la cláusula BY de las licencias Creative Commons).

En segundo lugar, tampoco consideramos que FACONAUTO restrinja el uso de estos datos a los fines comerciales; puesto que se encuentran expuestos en la web de la organización pública, sin ningún control de acceso - existiendo muchas organizaciones dedicadas a la explotación de este tipo de datos públicos para la consultoría de negocio.



Por tanto, también **hemos desestimado el uso de una cláusula NC (Non-Commercial)**, ya que esta debería ser impuesta en todo caso por la patronal.

Por último, y como consideramos que estos son datos de dominio público sobre los cuales no podemos aplicar restricciones de uso, **tampoco hemos indicado la voluntad de que las modificaciones de este dataset se lancen bajo el mismo tipo de licencia (CC0)**. Si bien consideramos que esto sería lo correcto, en caso de que otra persona o entidad decida aplicar una licencia de *copyright* sobre los datos de la patronal, esto correría bajo su propio riesgo. **Nuestros objetivos con este dataset son puramente académicos.**


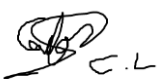


**9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.**

Podemos observar el código utilizado para generar el dataset *matr\_turismos.csv* a través del siguiente [enlace](#).

**10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.**

La publicación de nuestro dataset en Zenodo puede ser consultada en el siguiente [enlace](#).

DOI: `10.5281 / zenodo.4679639`

Contribuciones	Firma
Investigación previa	Dionisio González (D.G.), Carlos Lavado (C.L.)  
Redacción de las respuestas	Dionisio González (D.G.), Carlos Lavado (C.L.)  
Desarrollo de código	Dionisio González (D.G.), Carlos Lavado (C.L.) 