EVOLUCIÓN DEL PARO MADRILEÑO EN 2020

VISION GENERAL

Como parte del bootcamp de Data Science, se muestra aquí el proyecto individual EDA (análisis exploratorio de datos).

Durante este proyecto analizo la situación del paro producido en Madrid durante el año 2020.

Objetivos

El objetivo de este proyecto es poner en práctica los conceptos que aprendidos durante el bootcamp de Data Science: obtención de datos, minería de datos, visualización y análisis de los datos para sacar conclusiones y así consecuentemente poder confirmar o rechazar una hipótesis.

Especificaciones

Para que podamos hacer este proyecto y aprovechar al máximo la entrega, se requiere lo siguiente:

Software

Visual Studio v1.48.0

Hardware

Procesador i5 y al menos 4G RAM.

Requisitos

Información requerida:

- Datos sobre Covid-19: https://covid.ourworldindata.org/data/owid-covid-data.csv

Lenguaje de programación requerido:

- Python v3.8.2.64 bits

Bibliotecas necesarias:

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Missingno

Apartados:

I. Definir la hipótesis del proyecto.

Investigando diferentes archivos con información actualizada sobre la evolución del desempleo durante los últimos meses. Mi hipótesis se realiza en torno a la "confirmación de que el covid-19 ha influido negativamente en la empleabilidad de la población madrileña".

II. Obtener datos

Obtención de datos oficiales a través del servidor https://datos.gob.es/, del cual me descargo el archivo agencia_empleo_inscritos_2020 en formato csv para obtener los datos que

confirmarán o rechazarán mi hipótesis. Este dataset contiene información desde enero a junio de 2020.

III. Minería / limpieza de datos

• El dataset es bastante ordenado. Se limpia el conjunto de datos a través de la eliminación filas y valores no deseados. En este caso, decido dejar el resto de las columnas porque me serán útiles para la exploración del dataset.

• Se verifican los tipos de columnas

• Se corrige la columna de fecha y hora y se establece como nuevo índice

• Se comprueban datos para los valores NaN

• Comprobación de los datos en busca de valores duplicados

IV. Análisis y visualización de datos

• Para analizar los datos y facilitar al lector la información, he utilizado la visualización de datos con tablas y gráficos.

En este caso, se reproduce la visualización de cada una de las columnas según distintos criterios, ya que algunas necesitaban verse en una correlación respecto al índice (los inscritos al paro se registran el 1 de cada mes). Así se ha hecho con la evolución del mismo paro, con el género y la nacionalidad.

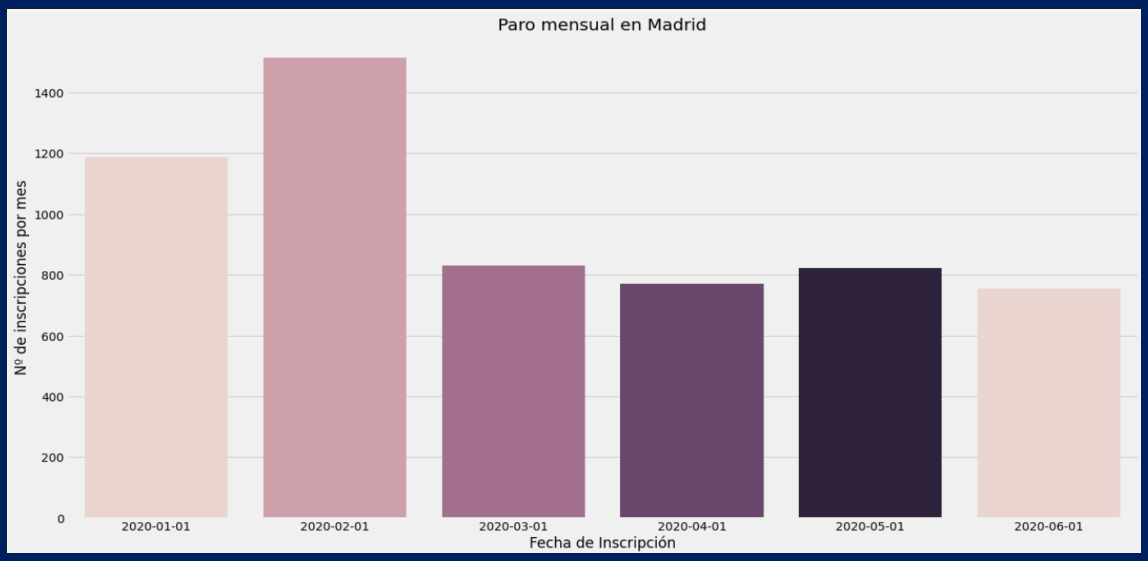
Por otro lado, columnas como la de los distritos marileños, eran más interesantes de ver mostrando el porcentaje que presentaban entre ellos mismos. Así pasó también al observar los distintos tramos de edades y los objetivos profesionales a los que se adscribían los desempleados por orden de prioridad.

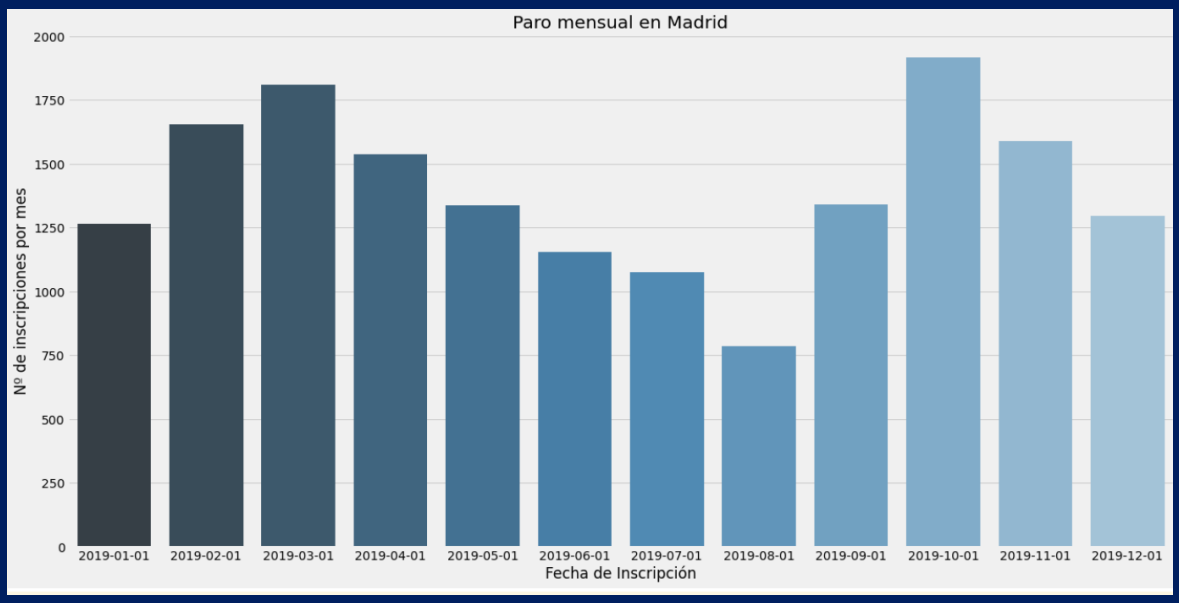
V. Conclusiones

Después de analizar los datos y visualizarlos, se infieren las siguientes conclusiones:

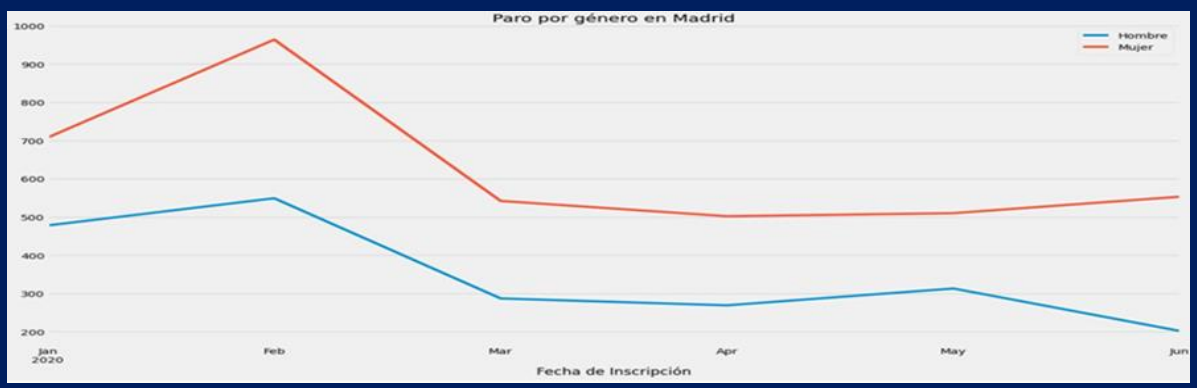
• Al estar hablando de fechas muy recientes, es difícil comprobar con certeza la incidencia del COVID-19 en el paro madrileño. Hay otra variable que se muestra en nuestros gráficos, y es que el paro sufría un fuerte repunte a principios de año debido al paro estacional producido cada año, inmediatamente después de las festividades navideñas. No obstante, es interesante ver que durante los meses de abril y mayo se mantiene en unas cifras menos altas de lo común anualmente, y que en mayo suben ligeramente, esto último muy probablemente debido a la falta de actividad durante la cuarentena.

El primer gráfico de barras muestra la evolución del paro en el 2020 y el segundo en el 2019:

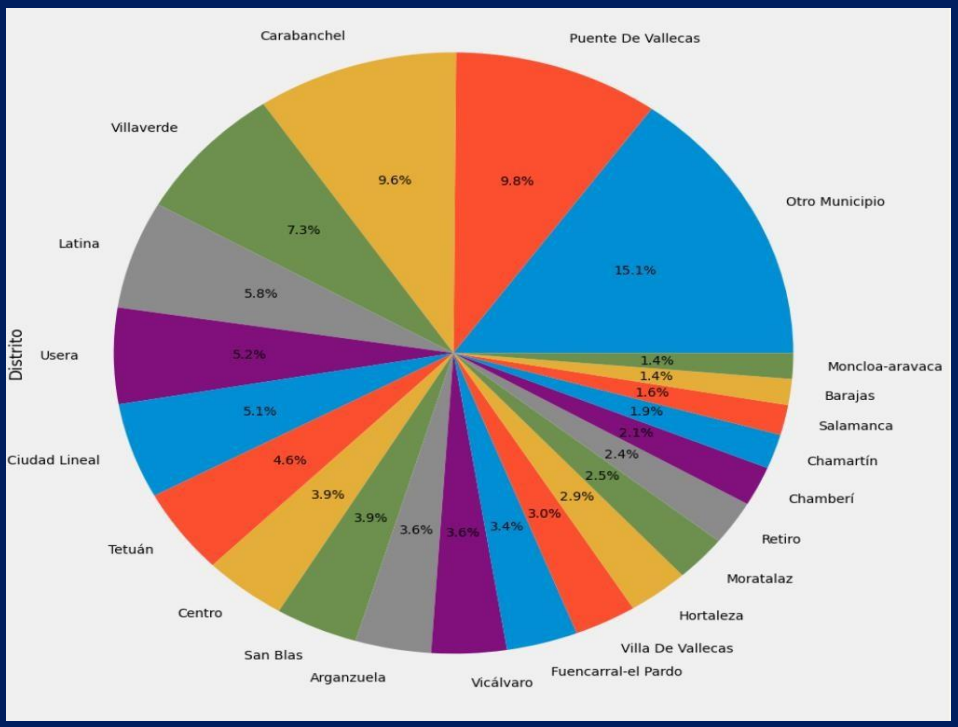




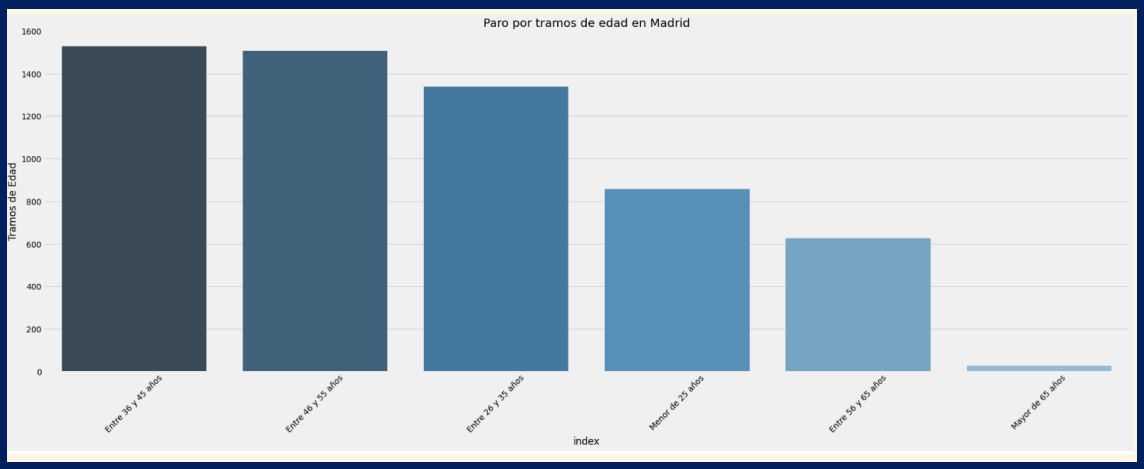
• En cuanto al resto de datos, es interesante observar que el género femenino supera con bastante claridad al paro masculino.



- Por otro lado, los barrios obreros madrileños son también los que cuentan con mayor desempleabilidad (liderando claramente el distrito de “Puente de Vallecas”, y los distritos obreros del sur (Carabanchel, Villaverde, Latina y Usera).



• En cuanto a los tramos de edad, lidera aquella población comprendida entre los 36 y 45 años, seguida con muy poca diferencia de aquella comprendida entre 46 y 55 años.



• Personal de limpieza, vendedores y trabajadores de cuidados a domicilio son los perfiles más demandados, probablemente debido a que son justamente los perfiles que coinciden con la mayor cifra de parados.

•Como conclusión, y además comparando con la gráfica de desempleo de 2019 (donde los datos son aún más altos), debo decir que los datos no son suficientes para apoyar la hipótesis planteada. Aún así el covid19 ha producido un escenario con acontecimientos tan recientes que tendremos que observar cómo siguen evolucionando las cifras de aquí en adelante durante los próximos meses/años.