Criterios utilizados en la normalización, transformación y limpieza de datos

Desempleo

- Archivo: Web.ipynb
  Realizamos Web Scraping de https://datosmacro.expansion.com/ para obtener la tasa de desempleo mensual de EEUU por estado desde 2010 a 2019.
  Una vez obtenida quitamos las columnas innecesarias quedándonos solo con las columnas de Estado, Mes, Año y Tasa de desempleo.
  Convertimos este DataFrame a csv *(Tasa_desempleo_API.csv)*

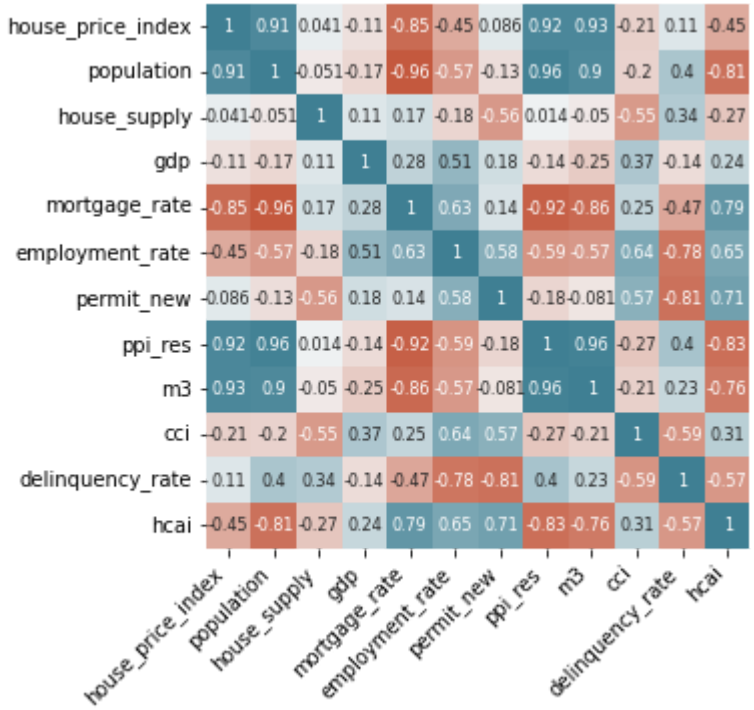
- Archivo: Desempleo.ipynb
  Leemos el archivo csv anteriormente nombrado.
  Creamos una nueva columna Fecha uniendo Año y Mes.
  Cambiamos las columnas a su formato correspondiente.
  Creamos una nueva columna Clasificación para categorizar la tasa de desempleo como:
  - Buena: Por debajo del 4%
  - Normal: Entre el 4 y 6%
  - Crítico: Por arriba del 6%

  Convertimos este DataFrame a csv *(Calificacion_Desempleo.csv)*

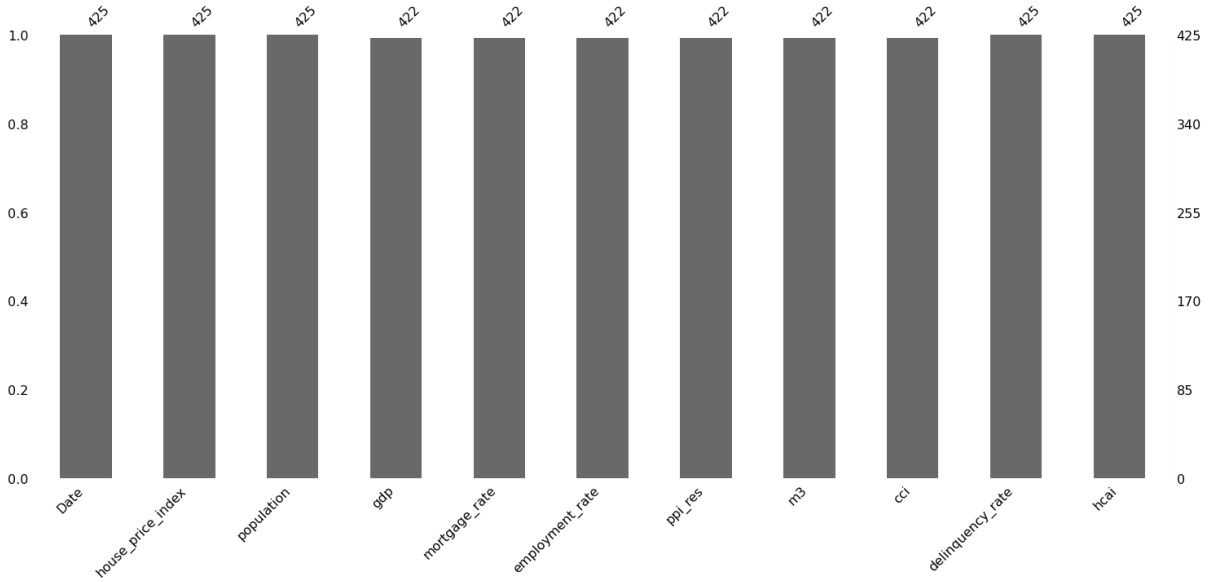
- Archivo: Dataset_Final_Desempleo.ipynb
  Leemos el archivo csv(*Dataset_price_rent_roi_alltypes.csv*)
  Renombramos columnas para poder unir con los demás datasets.
  Leemos el archivo csv(*Calificacion_Desempleo.csv*)
  Leyendo los dos archivos mencionados vemos que la información de las columnas de Estado está uno en español y otro en inglés.
  Modificamos todo a inglés.
  No tomamos en cuenta el Distrito de Columbia ya que no se obtuvo demasiada información, por esta razón lo quitamos.
  Quitamos columnas innecesarias y finalmente unimos los tres DataFrames teniendo las siguientes columnas: ID_Estado, Fecha, Año, Mes, Tipo_Propiedad, Precio, Tasa_Desempleo, Clasificacion.
  (*Dataset_final_desempleo.csv*)

Datos macroeconómicos, Riesgo Pobreza, PIB estados y TD estados

- Archivo: Desarrollo KPIs.ipynb
  comenzamos el ETL de las bases para el estudio del KPI Estabilidad
  Eliminamos duplicados, luego de eso revisamos valores faltantes, también observamos elementos de distribución, valores atípicos y correlaciones para revisar las variables que mejor se desempeñan con lo que vamos a analizar.
  Creamos las matrices de correlación:



revisamos datos faltantes:



reemplazamos caracteres especiales de las bases y luego reordenar columnas, también cambiamos los tipos de datos a float o datetime según corresponda.

Se imputaron valores faltantes con scikit-learn y simpleimputer usando como estrategia la media.
Hasta el momento hemos hecho el ETL y tenemos las bases listas:

- Housing macroeconomic factors
- Annual macroeconomic factors
- PIB estados
- Tasa Riesgo Pobreza por estados ((Web Scraping)
- TD = Tasa desempleo 2000 al 2010 (Web Scraping)
-

les colocamos la abreviación a todas las bases para relacionarlas después exceptuando las que están a nivel general como las macroeconómicas.

PIB

- Usamos una api de la web, donde hice Web Scraping(https://apps.bea.gov/api/signup/) donde encontramos información del PIB desde 2005 hasta el 2022, por trimestre, (variación según el trimestre anterior), luego de obtener esta data nos tocó transformarla primero escogiendo las columnas que solo necesitábamos del rango de años que habíamos decidido estudiar según el grupo de trabajo.
  Esto lo hice en un jupyter notebook, luego por medio de código python acomodamos la tabla y pudimos sacar la información de esta tabla y añadirla al csv *Dataset_price_rent_roi_alltypes,*
  Luego de organizar estos datos hicimos una valoración según investigaciones:
  PIB<0 DECRECIMIENTO
  PIB ENTRE 0 Y 2 ESTABLE
  PIB MAYOR QUE 2 CRECIMIENTO INERCIAL

Precio - Alquiler

- Archivo: unificacion de todas las propiedades.ipynb y las carpetas de datasets Mediana_Alquiler_Estados(Renta) y Property_Precio_Estados(precio de propiedad)

Cabe aclarar que los datasets en las carpetas, están separadas por tipo de propiedad de la misma forma en ambas. Para que al cargarlas los datos coincidan.
Cargamos De forma incremental los csv cambiandolos de Wide format a Long format en el proceso,además de eliminar columnas innecesarias y obteniendo el correspondiente nombre de la propiedad. Uniendo todo de forma ordenada y

secuencial en un solo dataset que incluye los precios y la renta de todos los estados, incluyendo todos los tipos de propiedad En un periodo de 2010-2019.
Rellenamos valores faltantes usando funciones que recogen el promedio de datos de años posteriores filtrando por (estado,año y tipo de propiedad) manteniendo la fiabilidad del dato.
Creamos una nueva columna llamada Roi(retorno de inversion) la cual calculamos dividiendo el precio de la columna Precio por la columna Rent dividido en 12. Lo que nos da el roi en años.
Creamos la columna Year para tener otro parámetro de fecha.
Convertimos el Dataframe completo a csv(Rent_Roi_Estados.csv)

Tiempo - Alquiler

- **EDA Zillow.ipynb:** Se inicia ell proceso de ETL cargando el archivo DaysOnZillow_State.csv, observando que viene como tabla wide format, con un total del 127 columnas, imputando los valores faltantes por la media, luego hacemos una conversión a long format con el método .melt(), obteniendo las columnas RegionName, Date y value en un formato Long Format, finalmente visualizamos cuales son los estados con mayor y menor frecuencia de Días en venta, utilizando un semáforo como forma de medición. el promedio normal es 60
- **El archivo PERMISOS.CSV** Fue obtenida mediante la tecnica de scraping de la página https://www.census.gov/construction/bps/historical_data/, obteniendo el promedio anual por estado
- **El archivo Hospitales.csv**: Contiene el promedio de hospitales por estado, fue obtenida de este enlace: https://hifld-geoplatform.opendata.arcgis.com/datasets/a2817bf9632a43f5ad1c6b0c153b0fab_0