# Segunda entrega – Proyecto Mercado inmobiliario Ruso de Sberbank

**Integrantes:**

Jhon Jader Caro Sanchez

Leider Steven Caro Mejía

José Manuel Ladino Villa

Inteligencia Artificial

Universidad de Antioquia

2023-2

## Objetivo

El objetivo de la primera parte de este proyecto de inteligencia artificial es llevar a cabo la preparación de datos y un análisis exploratorio exhaustivo.
Esto implica la limpieza de los datos, la integración de las fuentes de datos relacionadas con las casas en Rusia y la economía en el momento de la venta, la exploración detallada de las variables disponibles, un análisis de la variable objetivo (el precio de las casas) para identificar posibles sesgos y correcciones necesarias, la evaluación de correlaciones entre variables, la generación de datos nulos sintéticos para alcanzar un 5% de datos faltantes y la conversión de algunas variables numéricas a categóricas con el objetivo de alcanzar un porcentaje específico de variables categóricas en el dataset.
Estas acciones proporcionarán una base sólida para el desarrollo de un modelo predictivo preciso para los precios de las casas en Rusia

## Introducción

En el contexto de este proyecto, se busca desarrollar un modelo de inteligencia artificial capaz de predecir los precios de las casas en Rusia.
Para lograr este objetivo, se inicia con una fase fundamental de preparación de datos y análisis exploratorio. Esta etapa es esencial para garantizar la calidad y la coherencia de los datos, así como para comprender en profundidad las variables disponibles.

Además, se busca corregir posibles sesgos, identificar correlaciones entre las variables y, en algunos casos, transformar variables numéricas en categóricas. La importancia de esta fase radica en su capacidad para establecer una base sólida que permita la construcción de un modelo predictivo preciso y confiable para el mercado de bienes raíces en Rusia.

## Metodológia

**1. Exploración inicial**
Para este primer paso se hará una exploración superficial de la primera base de datos la cual esta nombrada como *Train.csv* la cual tiene variables que son directamente relacionadas con las caracteristicas del inmueble. Lo primero que se realizó como análisis exploratorio fue visualizar las primeras 5 filas, luego visualizamos las columnas, el tipo de dato y la cantidad de valores no nulos, la cantidad de columnas y la cantidad de filas. Se observó que hay 30471 registros y existen 292 variables entre las que se encuentra la variable objetivo la cual se llama *price_doc.*

Luego se realizó el mismo procedimiento con la segunda base de datos que se llama *macro.csv,* esta base de datos contiene la información de los datos economicos de Rusia desde Enero del 2010 hasta Octubre del 2016, contenia variables como el precio del combustible, el crecimiento economico de Rusia, indicadores de PIB, precio del Rublo con respecto al Dolar Estadounidense y entre otras caracteristicas. Se visualizó la cantidad de datos no nulos de cada columna, el tipo de variable, cantidad de columnas y cantidad de filas. Se determinó que la base de datos contiene 100 columnas y 2484 registros.

**2. Datos faltantes iniciales**
En el dataset principal, denominado *Train.csv*, se observaron ciertas columnas con un número significativo de datos faltantes. Específicamente, las columnas que presentaron la mayor cantidad de valores ausentes fueron el *estado del inmueble* con 13,559 datos faltantes, la *cantidad de camas de hospital en el distrito* con 14,441 datos faltantes y el *año de construcción* con 13,605 datos faltantes. En conjunto, estos valores

nulos representaron un total de 261,026, lo que corresponde al 2.93% de los datos en este dataset. Estos hallazgos subrayan la importancia de abordar la imputación de datos en estas columnas para garantizar que el dataset esté completo y sea adecuado para el análisis y la construcción del modelo predictivo.

En relación al dataset secundario, se identificaron columnas con un notable número de datos faltantes. Específicamente, las columnas con la mayor cantidad de valores ausentes fueron *grp_growth* con 1,023 datos faltantes, *salary_growth* con 658 datos faltantes y *pop_migration* con 658 datos faltantes. En conjunto, estos valores nulos sumaron un total de 46,658, lo que representa aproximadamente el 18.78% de los datos contenidos en este dataset. Esta alta proporción de datos faltantes en "macro.csv" resalta la necesidad de abordar la imputación de estos valores de manera efectiva para asegurar la integridad de los datos en el proceso de análisis y modelado subsiguiente.

### 3. Corrección de datos inexactos

En el dataset original, se identificaron valores erróneos relacionados con la ciudad de Tverskoe. Cuando un registro contenía esta designación, los valores de distancia asociados eran inexactos, lo que comprometía la precisión de los datos. Para abordar esta problemática, se ha optado por realizar una actualización utilizando el archivo '*BAD_ADDRESS_FIX.xlsx*', el cual fue obtenido a través del foro de discusión de la competición. Este archivo de corrección proporciona información valiosa para rectificar los datos incoherentes relacionados con la ciudad de Tverskoe y, así, mejorar la calidad y confiabilidad del conjunto de datos en cuestión.
Se aplicó el método *update* para efectuar dicha correción

### 4. Merge del dataset final

Para combinar y enriquecer los datos, se llevó a cabo un proceso de fusión (merge) de los datasets *macro.csv* y *Train.csv* en función de la fecha. Este proceso permitió la creación de un nuevo dataset consolidado con un total de 392 columnas y 30,471 filas. La fusión se basó en la coincidencia de las fechas en ambos datasets, lo que posibilitó la incorporación de información económica y macroeconómica de "macro.csv" a las observaciones correspondientes en "Train.csv". Esta integración de datos es fundamental para enriquecer el análisis y la construcción de un modelo predictivo sólido que tenga en cuenta tanto las características de las propiedades como el contexto económico en el momento de su venta. El nuevo dataset resultante proporciona una base más completa y rica para el análisis y modelado subsiguiente.

### 5. Análisis de la base de datos resultante

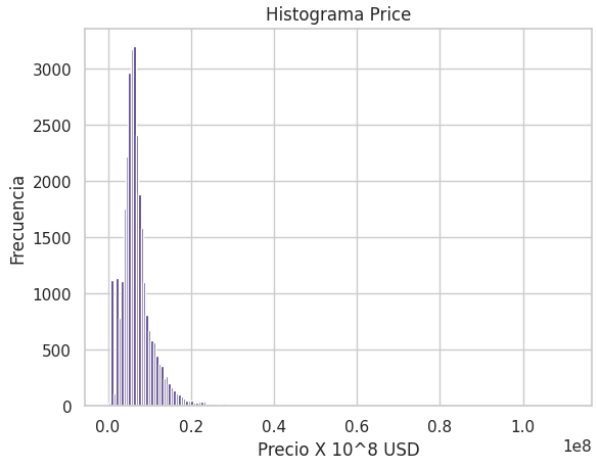
Tras la fusión de los datasets, se llevó a cabo un análisis descriptivo de las variables numéricas en el conjunto de datos consolidado. Este análisis abarcó la obtención del conteo total de observaciones, el cálculo de la media y la desviación estándar, la identificación de los valores mínimo y máximo, así como la determinación de los percentiles 25, 50 (mediana) y 75. Estos aspectos proporcionaron una comprensión profunda de la distribución y la variabilidad de los datos, lo que resultó esencial para orientar las decisiones en etapas posteriores del proyecto.

| | full_sq | life_sq | floor | max_floor | material | build_year | num_room | kitch_sq | state |
|---|---|---|---|---|---|---|---|---|---|
| count | 30471.000000 | 24088.000000 | 30304.000000 | 20899.000000 | 20899.000000 | 1.686600e+04 | 20899.000000 | 20899.000000 | 16912.000000 |
| mean | 54.214269 | 34.403271 | 7.670803 | 12.558974 | 1.827121 | 3.068057e+03 | 1.909804 | 6.399301 | 2.107025 |
| std | 38.031487 | 52.285733 | 5.319989 | 6.756550 | 1.481154 | 1.543878e+05 | 0.851805 | 28.265979 | 0.880148 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000e+00 | 0.000000 | 0.000000 | 1.000000 |
| 25% | 38.000000 | 20.000000 | 3.000000 | 9.000000 | 1.000000 | 1.967000e+03 | 1.000000 | 1.000000 | 1.000000 |
| 50% | 49.000000 | 30.000000 | 6.500000 | 12.000000 | 1.000000 | 1.979000e+03 | 2.000000 | 6.000000 | 2.000000 |
| 75% | 63.000000 | 43.000000 | 11.000000 | 17.000000 | 2.000000 | 2.005000e+03 | 2.000000 | 9.000000 | 3.000000 |
| max | 5326.000000 | 7478.000000 | 77.000000 | 117.000000 | 6.000000 | 2.005201e+07 | 19.000000 | 2014.000000 | 33.000000 |

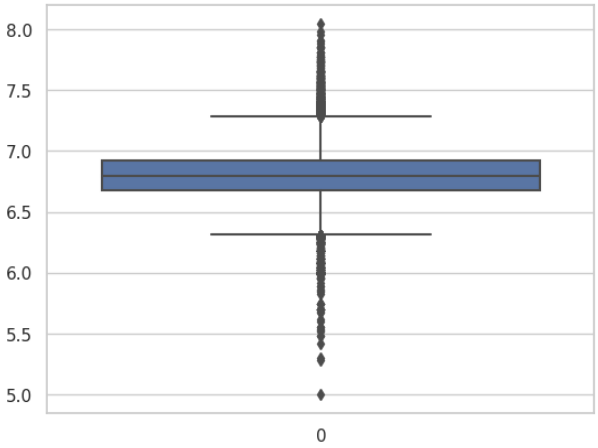
Se identificó inicialmente una cantidad de datos nulos que ascendía a 504,537 observaciones, lo que representaba aproximadamente el 4.2% del total de datos disponibles. No obstante, para cumplir con el objetivo establecido, era necesario incrementar deliberadamente la proporción de datos nulos hasta alcanzar el 5%.

**6.Análisis de la variable objetivo (price_doc)**

Se realizó un un histograma para conocer la distribución de los datos de la variable respuesta se obtuvo el siguiente gráfico




En nuestro caso vemos que la mayoria de los datos se ubican entre $100.000 y $20'000.000 USD aproximadamente.

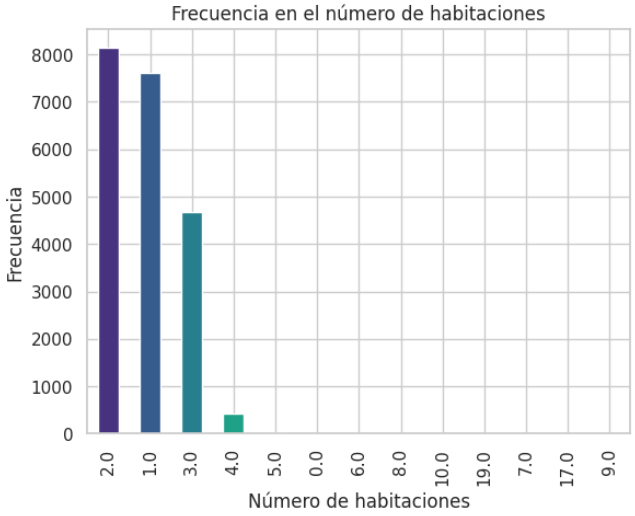


También se elaboró con la misma variable respuesta un gráfico de densidad y un gráfico de caja se determino que existe una cantidad significa de valores atípicos en nuestras muestras

**7.Análisis de variables categóricas**

En nuestro caso se usarón las variables con menos de 20 categorias y aquellas que fueran de tipo string, para este análisis visual se usarón gráficas de barra para representar la frecuencia de cada catégoria.
Se analizó la variable material de inmueble y se encontró que el 50% de los inmuebles estaban fabricadas del material tipo 1.
Del análisis de la variable estado del inmueble se pudó concluir que la tendecia de los inmuebles se encuentra entre mal estado, regular estado y buen estado, siendo excelente estado la catégoria con menos recurrencias.

Luego de la variable número de habitaciones se puede concluir que las casas con 2 habitaciones son las más recurrentes abarcan un poco más del 28% de las muestras, las casas con una habitación representan el 25% de las muestras, las casas con 3 habitaciones representan el 15% de las muestras y por último las casas con 4 habitaciones representan aproximadamente el 2% de las muestras.

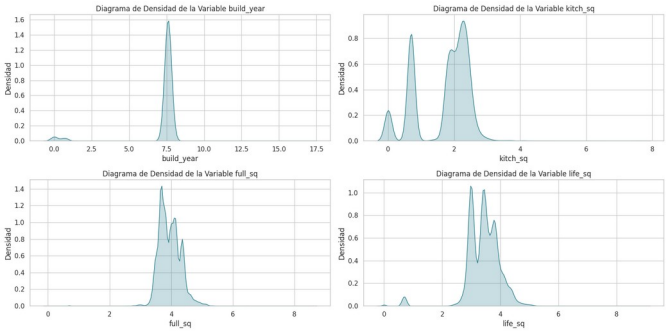



Luego de la variable tipo de producto se puede concluir que casi el 2/3 de las muestras son de tipo inversión mientras el

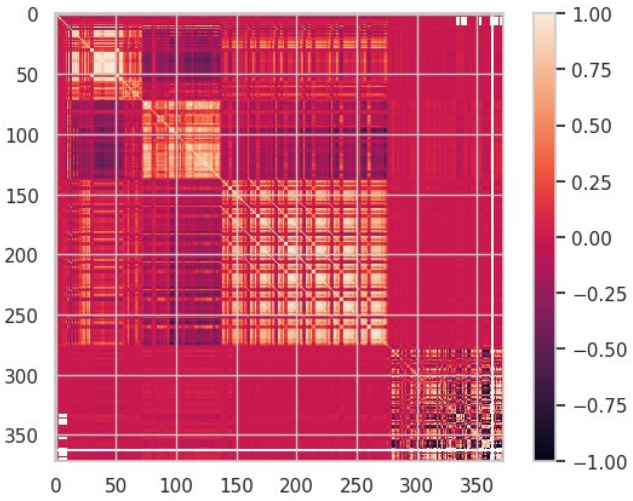
1/3 de las muestras son de tipo dueños propios de la residencia.

**8.Análisis de variables númericas**
El siguiente procedimiento que se ejecutó fue representar algunas variables númericas con diagramas de densidad para conocer un poco su distribución y saber si algunas de estas variables tenían sesgo o una distribución atipica.

Algunas variables como año de construcción, metros cuadrados de la cocina, metros cuadrados totales y metros cuadrados habitables poseían un sesgo hacia la izquierda y por ende se decidio aplicar la función **Log(x+1),** y el resultado fue el siguiente:

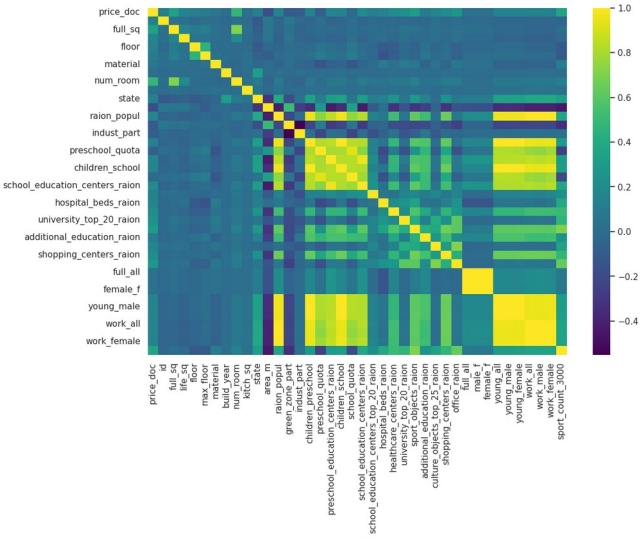



**9.Matriz de correlación general**
Luego se procedió a contruir una matriz de correlación y se represento gráficamente en un mapa de calor y este fue el resultado:




Vemos que el mapa de calor estaba segmentado por zonas, zonas en las que existió una correlación bastante fuerte con otras variables y zonas donde la correlación negativa es bastante pronunciada, y zonas donde no existió correlación entre variables.

**10.Matriz de correlación segmentada**
Luego se realizó un mapa de calor basandonos en la segmentación del anterior mapa de calor general y se tomo el segmento (0,0) para poderlo gráficar y analizarlo más a detalle




En la primera fila vemos a la variable respuesta, vemos que no existe mucha relación de ninguna de estas variables con esta variable respuesta, lo cual significa que no muchas variables de estan pueden explicar el comportamiento del precio de los inmuebles. Tambien se pudó evidenciar que muchas variables independientes tienen correlación muy alta entre ellas.

Luego se tomó el ranking de variable dentro de este mismo segmento con mayor correlación entre variables independientes, se evidenció que existian caracteristicas con más del 95% de correlación

**10. Generación de datos nulos sintéticos**

Para este proceso lo que se hizó fue seleccionar las columnas que ya tuvieran datos nulos, luego de tenerlas identificadas se calculó la cantidad de datos que era necesaria eliminar para lograr el 5% de datos nulos, esta cantidad fue 85.214, se dividió esta cantidad por la cantidad de columnas con datos nulos y aleatoriamente, pero aplicando una semilla para que esta misma aleatoriedad sea reproducible, se eliminaron los datos para completar la cantidad de datos faltantes mínima.

**11. Descargar la base de datos completada**

Se descarga la base de datos fusionada y con la cantidad de valores nulos completada, se descargó la base de datos para poder continuar el proceso en el segundo cuaderno llamado *02-Preprocesamiento*

**12. Generación de variables categóricas sintéticas**

Esta segunda parte esta alojada en el segundo cuaderno llamado *02-Preprocesamiento.*
Otro de los requisitos mínimos es que el 10% de las variables sean categóricas, por ende se analizó que variables son categóricas y cuantas variables categóricas hacían falta para componer el mínimo solicitado. El mínimo solicitado para este caso son 39 columnas categóricas. Para completar el mínimo estipulado se convirtieron 24 columnas a categóricas, formulando la siguiente pregunta ¿el valor es mayor a la media de su columna? En caso de que fuera cierto el valor cambiaría a 1, en caso contrario tomaría el valor de 0.