1/09/2024

Imagen que contiene Patrón de fondo

Descripción generada automáticamente

Proyecto Programación

C++

**Integrantes:**

Juan David Delgado Burbano

Nicolas Martínez

Camila Ariza

Reporte análisis:

Para empezar, el archivo csv el cual el proyecto e c++ lee, es un archivo el cual contiene 414.120 filas de datos separados por comas, de los cuales, algunos de dichos datos no eran útiles para la realización de una regresión lineal en donde se predisca el precio de alguna propiedad en base a alguno de estos datos como el numero de habitaciones, o los metros cuadrado de la vivienda.

Para ello, el equipo realizo el siguiente análisis, para ver datos que tuviesen relación entre ellos para la regresión lineal, en donde tomando 16 datos iniciales del csv, los cuales pertenecieran a la misma ciudad, mismo barrio, y sean el mismo tipo de vivienda (departamentos en este caso), se realizó una tabla con ellos utilizando el número de habitaciones como eje x, y el precio propiedad como eje y, (utilizando las funciones básicas de Excel).

Gráfico

Descripción generada automáticamente con confianza media

De esta forma se comprobó una relación de a mayor numero de habitaciones, mayor valor posee la propiedad por lo cual se utilizaron estos datos como punto de partida.

* Limpieza de Datos:

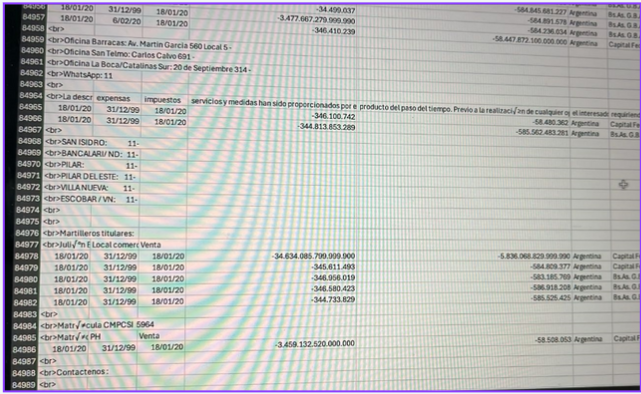
La base de datos del csv, tenia varios problemas, como:



Apartir de la línea 80 mil, el csv contenía datos que interrumpían la lectura del código, como los que se muestran en la siguiente imagen, los cuales no correspondían al patron de lectura del csv siendo (

* Fecha de inicio,
* fecha termiando
* ,creado en ,
* lat,
* lon ,
* pais,
* ciudad,
* barrio ,
* #habitaciones
* #baños,
* #alcobas,
* superficie total,
* superficie cubierta ,
* precio ,
* usd ,
* nombre
* ,descripccion de la propiedad ,
* tipo de propiedad.

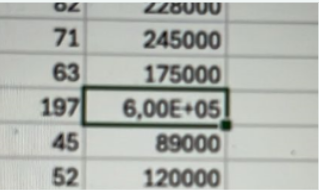
) al romper este ciclo de lectura, el código empezaba a tomar datos de tipo carácter como (tipo de propiedad) como valores enteros ( INT ), lo cual producía errores,



1. caracteres en valores numéricos:

Tabla

Descripción generada automáticamente



Completamente aleatorio en el csv, existían datos de carácter numérico, los cuales contenían caracteres, por lo cual esto producía error en la lectura del código; por otro lado a lo largo del csv existían datos que no se tomaron los cuales aparecían como ( NA ), y en otros casos, aparecían simplemente en blanco, porque no fueron llenados.

Este problema fue muy importante resolverlo puesto que, la forma de lectura del csv, era leer carácter por carácter, y cuando encontrase una coma significaba que hasta ese ultimo carácter antes de la coma era parte de una variable. Y si por alguna razón alguna variable no estaba en la fila, entonces el código emparejaría datos erróneamente con sus respectivas variables de lectura en el código.

1. Eliminación de datos no necesarios:

Por ultimo, solo de dejo datos que serian necesarios para la implementación del código, los cuales fueron: 7



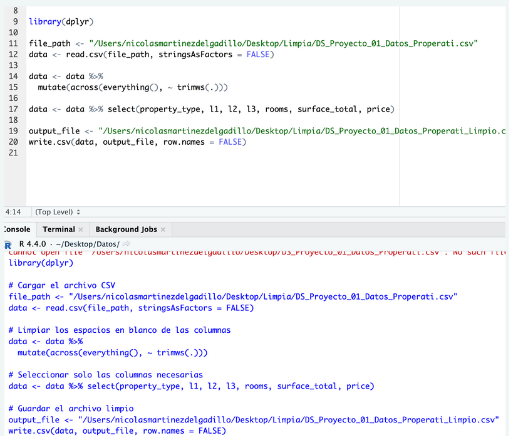
Estos datos eran necesarios para hacer la regresión lineal, puesto que variables como país, ciudad barrio de procedencia de la propiedad, ayudaban a entender la valorización de las propiedades dependiendo su ciudad u barrio. Además el numero de habitaciones y el precio fueron los datos seleccionados para la regresión, y datos adicionales como tipo de propiedad, y superficie total, eran necesarios para poder diferencia las propiedades, ( si eran departamentos o casas, o porque el valor de un apartamento con 2 habitaciones era mas caro que otro de dos habitaciones, esto debido a los metros cuadrados ).

**Solución de problemas:**

Se elaboraron dos código en “ R “ para la limpieza de datos del csv,

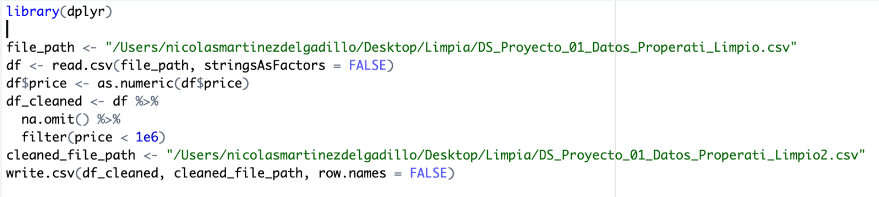


Este código busca eliminar, en la columnas los elementos NA y los números con notación científica.





Este código busca en el csv en la carpeta del escritorio, luego limpia la base de datos eliminando las columnas y filas vacías y luego elimina todas las columnas , menos las columnas de los datos escogidos:



Dejando únicamente estos datos por fila:



**Resultado final csv:**

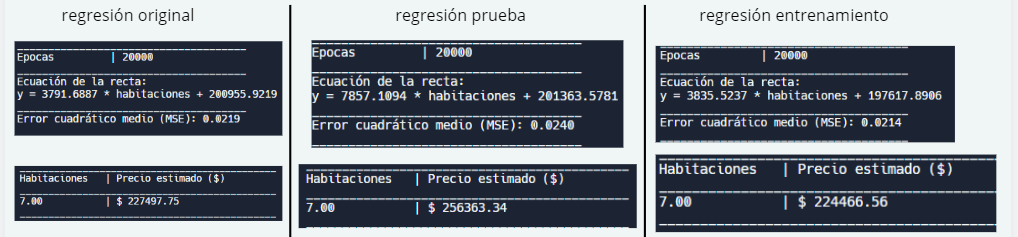
Finalmente, la tabla quedaría con 123517 filas, cada fila contiene 7 datos, separados por comas como se muestra en la siguiente imagen:



**Análisis de Validación de modelo / Modelo:**

( la explicación del código de c++ de validación del modelo, y el modelo de regresión lineal, esta detallada en el github del proyecto, en apartado presentación )

Análisis:



En la imagen se muestra las impresiones finales del modelo original (100% datos csv ) , el modelo de prueba (20% datos csv ), y el modelo de entrenamiento (80% datos csv ),

* Ecuación de la recta:

Como se ve en la imagen las ecuaciones de las rectas del modelo de regresión lineal original, y el modelo de entrenamiento son muy cercanas, sobre todo en la pendiente de la recta, sus valores son muy similares, y diferencian un poco en el corte con el eje y, sin embargo, su cercanía comprueba que el modelo original funciona correctamente.

Por otro lado frente a la pendiente, las ecuaciones del modelo original, y el modelo de prueba, diferencian bastante, sin embargo, en el caso de la pendiente, este valor tiene sentido puesto que dentro del rango del 20% que toma el modelo de prueba, puede tener datos que están más dispersos en proporción al 100% que posee el original. Prueba de ello, es que ambos tienen un corte con el eje y, muy similar.

* MSE:

Todos los modelos poseen un error cuadrático promedio de 0.02, variando únicamente en el decimal después del 2. Lo que significa que la predicción es casi perfecta.

* PREDICCION:

Todos los precios predichos para #7 habitaciones, son muy similares, especialmente entre el modelo original y el modelo de entrenamiento, mientras que el modelo de prueba varia significativamente 2000 dólares mas alto que los demás, pero esto nuevamente se debe a la pendiente de su recta, la cual ya se explico en el punto anterior la razón de ello.

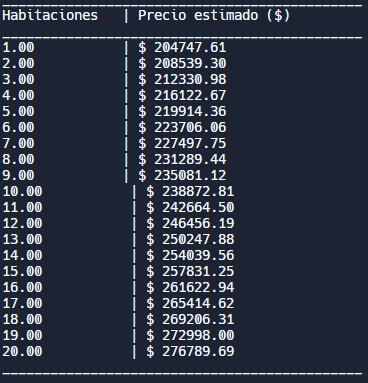
* EPOCAS:

Todos los modelos fueron probados con 20 mil épocas.

**Ejemplo de impresión:**

**Escala de tiempo

Descripción generada automáticamente**

****