# Matriz de aptitudes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Aptitud / Miembro** | **Alejandro** | **Daniel** | **Eugenia** | **Gastón** | **Matías** |
| **Programación** |  | Sí | Sí | Sí | La peleo |
| **Pandas** |  |  | Me defiendo | Sí | La peleo |
| **Visualización** |  |  | Me defiendo | Me defiendo | No |
| **Estadística** |  |  | No | Sí | Me defiendo |
| **Presentación** | Sí |  | Sí | No | Sí |

# Tareas.

## Distribución.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tarea** | **Alejandro** | **Daniel** | **Eugenia** | **Gastón** | **Matías** |
| Geolocalización |  |  | x |  |  |
| **Precio** |  |  |  | x |  |
| **Superficie** |  |  |  | x |  |
| Piso |  |  |  |  | X |
| **Ambientes** |  |  |  | x |  |
| Visualización |  |  | x | x |  |
| Notebook entrega | x | x | x | x | x |
| Presentación | x |  |  |  |  |

## 

## Descripción.

### Geolocalización.

Hay varios campos que intervienen en la geolocalización. Los más útiles son **lan** y **lon** porque nos dan la ubicación exacta de la propiedad. El problema es que tienen un porcentaje alto de nulos, por lo que el objetivo va a ser tratar de inducir estos valores.

Una estrategia de imputación podría ser utilizando el campo **geonames\_id**, y luego, a su vez, evaluar si podemos imputar **geonames\_id** con el resto de los campos involucrados en la ubicación de la propiedad: **place\_name**, **place\_with\_parent\_names**, **country\_name** y **state\_name**.

Por orden:

1. Imputar **geonames\_id** con **place\_name.**
   1. Agrupar los datos por **place\_name** o **state\_name**.
   2. Dentro de esa agrupación, buscar nulos del campo **geonames\_id**.
   3. Si dentro de la agrupación todos los registros completos tienen el mismo **geonames\_id**, impactar los registros vacíos con este valor. Si los números son distintos, imputar la moda.
2. Imputar **lat** y **lon** con **geonames\_id.**

Por último, se podría hacer un mapa para mostrar la distribución de las propiedades.

### Precio. Hecho.

No vamos a tratar de predecir el precio en esta etapa porque es tema del próximo desafío.

Lo que vamos a hacer es evaluar cuáles son los campos que tienen la información que nos sirve y descartar el resto.

Además, vamos a tratar de completar valores con lo que haya en **description**.

### Superficie. Hecho.

Los campos que juegan en la superficie de la propiedad son **surface\_total\_in\_m2** y **surface\_covered\_in\_m2**. Ambos son muy importantes para tasar la propiedad, por lo cual es importante que estén lo más completos posible.

En el caso de **surface\_total\_in\_m2**, el porcentaje de nulos es bastante alto. No podemos descartar una cantidad tan grande de datos, por lo que vamos a tratar de completar algunos valores con lo que haya en **description**.

El problema principal de esto es que la **regex** que necesitamos para obtener los campos de **description** es complicada.

Franco sugirió buscar:

* dormitorios / dorm
* m2 / m^2 / mts
* baños
* living
* si tiene pileta o no

### Piso.

El campo **floor** tiene un porcentaje muy alto de nulos.

Si la propiedad no es de tipo *apartment*, puede ser que haya sido omitido intencionalmente.

Si es un *apartment*, vamos a tratar de completar algunos valores con lo que haya en **description**.

### Ambientes.

El campo **rooms** también tiene un porcentaje alto de nulos.

La única fuente de la cual podríamos completar los datos es con lo que haya en **description**.

### Visualización.

Mostrar los datos en visuales que cuenten algo de la distribución de los datos.

* Gráfico de barras con las variables más relevantes y porcentajes de nulos.
* Antes y después, con la imputación/deducción de los valores.
* Geolocalización con Seaborn/Geopandas.
* Dispersión de datos (estadístico).
* Mapa de calor: barrio + precio + m2.
* Tamaño de propiedad / tipo de propiedad por barrio.

### Notebook presentación.

Recopilar y unificar todo el material en una Notebook para entregar.

Podría seccionarse de la siguiente forma:

1. Presentación del dataset y conclusiones iniciales.
2. Tratamiento de los campos: limpieza, normalización, imputación, descarte, etc.
3. Visualización: más que nada de la distribución de los datos.
4. Conclusiones finales.

### Presentación.

Hacer la presentación en clase.

Implica la elaboración de la PPT, ya que tratar de explicar la presentación de otro suele ser avergonzante.