**Proyecto final de Big Data**

**Grupo TELLA**

**Diagrama

Descripción generada automáticamente con confianza media**

**Fecha de entrega: 15 de febrero de 2023**

**Alumnas autoras del informe:**

|  |
| --- |
| **Apellidos, Nombre** |
| Baluja García, Tania |
| Jiménez-Alfaro Ayala, Ana |
| Jiménez González, Laura |
| Manrique Savasta, Emely |
| Montero Martínez, Liliana |

**ÍNDICE**

**Pp.**

**PRESENTACIÓN ……………………….………………………………….………..…….. 1**

**IDEA DEL PROYECTO ………………………………………..…….…….……………… 1**

**DEFINIR DATA SET ………………………………………………………..……….…….. 1**

**ARQUITECTURA, VALIDACIÓN DE LOS DATOS Y ANÁLISIS EXPLORATORIO .. 1**

**VISUALIZACIÓN DE LAS MÉTRICAS …………...…………………………….………. 1**

**PREPROCESAMIENTO Y MODELADO ……………...…………………….………….. 1**

**CONCLUSIONES ……………………………………...……………………….………….. 1**

**BIBLIOGRAFÍA ……………………………………………………………….….………… 1**

**PRESENTACIÓN**

* **Tania Baluja García**

Soy estudiante de último curso del grado de ingeniería aeroespacial en la Universidad Politécnica de Madrid. Decidí realizar el curso de introducción a la programación, puesto que considero que una herramienta de trabajo que se emplea en numerosos ámbitos de la actualidad, ya que constantemente estamos rodeados de gran cantidad de datos que debemos gestionar para poder trabajar con ellos. Con este curso también conocí gente de diferentes sitios que estaban realizando esta formación a la par que trabajan o estudian, lo que me hizo consciente de lo capaces que somos de esforzarnos para conseguir nuestros objetivos. Finalmente, me gustaría destacar que me sentí muy bien acompañada en todo momento por unas compañeras que no sólo luchaban por conseguir sus metas, sino que ayudaban a las demás. Esto me recuerda a las primeras clases de Katas, en las que el profesor nos insistía en ayudarnos las unas a las otras para poder progresar.

* **Ana Jiménez-Alfaro Ayala**
* **Laura Jiménez González**
* **Emely Manrique Savasta**
* **Liliana Montero Martínez**

**IDEA DEL PROYECTO**

En este proyecto crearemos una empresa que gestione el alquiler de una serie de viviendas vacacionales tratando de optimizar los recursos para obtener la máxima rentabilidad.

La idea es mostrar el cliente potencial y la rentabilidad según las características de una vivienda gestionada por una empresa especializada.

**DEFINIR DATASET**

Al inicio de este proyecto se parte de una tabla con distintos datos. Cada columna es una variable (país, ciudad, nombre, descripción…), mientras que cada fila es una vivienda. En este caso hay 14780 listing y 48 variables.

De toda esta información hay columnas que no queremos porque nos faltan datos o porque no es relevante la información que aporta.

Primero, se divide el Dataset en dos grupos. Por un lado, están las viviendas gestionadas por profesionales, considerando que estos tienen un mínimo de unas 4 o 5 viviendas. Por otro lado, están las viviendas gestionadas por particulares, que tendrán menos propiedades que los profesionales.

**ARQUITECTURA, VALIDACIÓN DE DATOS Y ANÁLISIS EXPLORATORIO**

Se comienza por comprobar los tipos de datos que proporciona cada columna (int, object, float). Aquí se observó que ‘Host Since’ aparece como object, por lo que se cambia a formato fecha.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Finalmente, se comprueba si ID es una PK.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Luego, se obtiene el número de filas y de columnas de la tabla de datos.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para normalizar los valores, se empieza por realizar una limpieza de los nombres de las columnas eliminando los espacios por ‘\_’ y las filas vacías.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Posteriormente, se transforman los nombres de las columnas a mayúsculas para una mayor legibilidad.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Finalmente, se realiza una limpieza de los valores nulos. Para ello, antes de normalizar, limpiar y transformar los datos a ASCII y mayúsculas; se buscan las columnas con mayor cantidad de valores nulos y se toma una decisión de cómo tratarlos, puesto que es necesario tener un DF sin NANs.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

En este caso, se obtienen la siguiente cantidad de NANs para los diferentes campos:

* BEDROOMS: 25
* BEDS: 49
* BATHROOMS: 55
* SECURITY DEPOSIT: 8524
* CLEANING\_FEE: 6093
* AMENITIES: 170
* FEATURES: 1

Estos valores de NAN fueron transformados a un valor 0, puesto que es el valor numérico equivalente.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Al quitar los NAN hay que volver a pasar los valores numéricos a datos tipo integer, puesto al realizar la operación anterior se transformaron en datos tipo object. Finalmente, se comprueba que los datos son correctos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Se realiza una modificación en los datos anteriores devolviendo los NAN a las columnas de reviews y se comprueba que está correcto. De esta manera se obtienen 11190 NANs para WEEKLY PRICE y 11219 NANs para MONTHLY PRICE.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

En este punto se asume que estos campos nulos se deben a que el propietario no ofrece descuentos en los alquileres semanales o mensuales, por lo que el coste total será el precio diario multiplicado por los días de alquiler.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Y se comprueba que las columnas WEEKLY\_PRICE\_CLEANSED y MONTHLY\_PRICE\_CLEANSED no contienen valores NAN.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Finalmente, se comprueba si PRICE es NAN y WEEKLY\_PRICE o MONTHLY\_PRICE tienen datos. Luego se calcula la inversa, lo que permite conseguir 4 campos más de 17 NAN a 13. Tras calcular los que faltan, se quitan del modelo los que siguen teniendo NAN, puesto que este campo es clave para el análisis y no aporta la información suficiente.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Al campo anterior se añade la columna TOTAL\_PRICE que es la suma de PRICE\_CLEANSED y CLEANING\_FEE. No interesa sumar al precio total el depósito, ya que ese dinero es recuperable.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

También se debe comprobar que la suma es correcta haciendo la media de PRICE\_CLEANSED y TOTAL\_PRICE y viendo que TOTAL\_PRICE es mayor.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

En este punto del análisis de los datos se procede a añadir columnas al campo anterior.

* **Columna que indica si la vivienda tiene descuento por media/larga temporada**
* Columna que indica el precio por persona
* Columna que indica la ratio de cama por persona
* Columna que indica la tasa de ocupación y ocupación máxima
* Columna que indica la ratio de cama por capacidad de huéspedes (accommodates)
* Limpieza de localización
* Separación en columnas amenities
* Columna booleana de security deposit
* Codificando políticas de cancelación
* Columna que indica la ratio del número de veces alquilado
* Exportar el DF final
* Análisis de la ratio de ocupación > 1
* Exportar el nuevo CSV

**VISUALIZACIÓN DE LAS MÉTRICAS**

**Contestar a una pregunta relevante que hagamos sobre los datos**

* **Diseño del dashboard**
* **Uso de buenas prácticas**
* **Cálculo de KPIs adecuados y el uso de campos calculados avanzados**
* **Uso de vistas interactivas**

**PROCESAMIENTO Y MODELADO**

**Algoritmo de regresión lineal que prediga el precio de un imbueble en función de sus características. Elegir características: tipo de alquilador, barrio, tipo de vivienda, tipo de habitación, número de personas, habitaciones, baños, camas, tipos de camas, si tiene fianza, si tiene precio de limpieza, política de cancelación…**

**CONCLUSIONES**

* **Suposiciones iniciales: ¿Cuáles han demostrado ser válidas y cuáles no? ¿Por qué?**
* **Métricas seleccionadas: ¿Han sido las correctas o no? ¿Por qué?**
* **Teniendo en cuenta lo aprendido: ¿Qué cosas se harían igual y cuáles se harían de otra forma? ¿Por qué?**
* **Conclusiones y lecciones aprendidas.**

**BIBLIOGRAFÍA**