

ANALISIS EXPLORATORIO DE DATOS



Proyecto de Inversión
Gastronómica y Afines en
Mercado De Estados
Unidos

El análisis exploratorio de datos realizado en el conjunto de datos "df_estado", proporcionado por el dataset limpio "Estado.csv", reveló información valiosa sobre la población y la densidad demográfica en los diferentes estados. En primer lugar, se verificó la integridad de los datos y se observó que no había duplicados ni valores nulos en ninguna de las columnas. Además, se obtuvo una visión general de la estructura del conjunto de datos a través de la función "info()".

Se examinó más detenidamente la columna "Total_Poblacion", utilizando la función "describe()" para obtener estadísticas descriptivas, lo que permitió comprender mejor la distribución de la población en los estados. Además, se identificaron los nombres únicos de los estados a través de la función "unique()".

Para destacar los estados con la población más alta, se generó un gráfico de barras que mostraba los 10 estados con la población más grande. Esta representación visual ayudó a identificar rápidamente los estados más poblados, con etiquetas que indicaban la cantidad exacta de habitantes en cada estado.

Posteriormente, se realizó un análisis para encontrar el estado con una población cercana a la mediana de población total. Se calculó la población intermedia y se creó otro gráfico de barras que mostraba los 5 estados más cercanos a esa población intermedia en millones. Nuevamente, se incluyeron etiquetas con la población exacta en cada barra.

Finalmente, se exploró la población en los estados con menor población total. Se identificaron y presentaron visualmente los 5 estados con la población más baja mediante un gráfico de barras que resaltaba esta característica.

En el análisis exploratorio de datos de densidad demográfica, se calculó la densidad para cada estado y se presentaron visualmente los estados con la densidad más baja en otro gráfico de barras. Cada barra representaba un estado, y su altura reflejaba la densidad demográfica, lo que permitía identificar fácilmente los estados menos densamente poblados.

En lo que respecta al segundo análisis exploratorio de datos (EDA) del conjunto de datos de franquicias obtenido a través del dataset “Franquicias_inversion.csv”, hemos llevado a cabo una exploración exhaustiva para comprender y revelar información esencial sobre las franquicias y sus características clave. El conjunto de datos consta de información variada, que incluye detalles sobre inversiones, unidades, nombres de franquicias y años de fundación.

Primero, evaluamos la forma del conjunto de datos, que comprende el número de filas (instancias) y columnas (atributos), brindando una visión inicial de su tamaño y estructura.

Luego, realizamos una revisión general del conjunto de datos para identificar columnas con valores faltantes y comprender las características de las variables.

Para comprender mejor las inversiones, calculamos estadísticas descriptivas para las columnas 'Max_Inversion' y 'Unidades'. Esto nos proporcionó información sobre la inversión máxima y la cantidad de unidades de franquicias, incluyendo medidas como la media, la desviación estándar y los valores extremos

Exploramos las franquicias únicas en el conjunto de datos mediante la identificación de los valores únicos en la columna 'Nombre_Franquicia', lo que nos permitió conocer qué franquicias están representadas en los datos. Luego, presentamos un resumen completo de estadísticas descriptivas que abarcan todas las columnas numéricas del conjunto de datos, proporcionando una visión completa de las características clave.

Para profundizar en el análisis, identificamos las 10 franquicias con más unidades y las representamos en un gráfico de barras, destacando también su año de fundación. Esto nos permitió visualizar las franquicias líderes en términos de unidades y comprender mejor su historia.

A continuación, exploramos la relación entre la inversión mínima y máxima en las 10 franquicias principales a través de un gráfico de dispersión, lo que nos brindó una visión de cómo estas variables se correlacionan.

Además, examinamos las 20 franquicias con más unidades y presentamos la inversión máxima y mínima de estas franquicias en gráficos de barras horizontales, lo que resaltó las franquicias líderes en términos de inversión.

Para identificar posibles relaciones entre las variables numéricas, calculamos y visualizamos una matriz de correlación en un mapa de calor interactivo.

Finalmente, creamos una nube de palabras que destaca las palabras clave en los nombres de las franquicias y sus inversiones, lo que proporcionó una vista general de los patrones en los nombres de las franquicias y su inversión.

Este EDA exhaustivo nos ha brindado una comprensión completa de las características y relaciones clave en el conjunto de datos de franquicias, lo que puede servir como base sólida para futuros análisis y toma de decisiones en el ámbito de las franquicias y las inversiones.

Por último, para finalizar con los EDA correspondiente a los Datasets limpios, hicimos la exploración de datos del dataset “Reviews_Completos”.

En esta exploración en profundidad de los datos de revisiones de franquicias, hemos llevado a cabo un minucioso examen para extraer información clave sobre estas revisiones y sus aspectos más destacados. El conjunto de datos es bastante diverso y contiene información sobre diversas categorías, nombres de franquicias, ubicaciones geográficas y calificaciones promedio de las revisiones.

Iniciamos el proceso revisando la estructura general del conjunto de datos, lo que nos proporcionó una visión inicial de su tamaño y características. Luego, evaluamos si había duplicados en el conjunto de datos y cuantificamos la cantidad de registros duplicados que encontramos.

Realizamos un análisis de los valores faltantes en el conjunto de datos para comprender la integridad de los

datos y su calidad. Enfocamos nuestra atención en la columna "Categoría" para analizar las diferentes categorías a las que pertenecen las franquicias y proporcionamos estadísticas descriptivas para esta variable.

También examinamos detenidamente la columna "Nombre_Franquicia" para comprender qué franquicias estaban representadas en el conjunto de datos.

Identificamos la categoría con la mayor cantidad de revisiones y filtramos el conjunto de datos para explorar específicamente las revisiones de esa categoría. Además, excluimos las categorías que no se relacionaban con "Restaurant", "Caterer" o "=Catering", y analizamos el subconjunto resultante.

Calculamos la frecuencia de las franquicias en cada categoría filtrada y creamos un gráfico de pastel que destaca las 10 categorías más comunes.

Destacamos las categorías que presentaban cantidades extremas de revisiones y las ordenamos según su magnitud. Investigamos las franquicias más exitosas mediante un análisis de la frecuencia de sus nombres en las revisiones y mostramos las 20 franquicias más destacadas.

Realizamos un análisis espacial para visualizar la distribución geográfica de las sucursales de Subway.

Exploramos la relación entre la cantidad de revisiones y el promedio de calificaciones mediante un gráfico de dispersión. Seleccionamos un grupo específico de franquicias y calculamos, visualizamos y comparamos sus promedios de calificaciones.

Identificamos la categoría "Fast food restaurant" y calculamos el promedio de calificaciones para esta categoría. Presentamos las franquicias mejor calificadas en la categoría "Fast food restaurant" y destacamos la presencia de Subway.

Finalmente, calculamos el promedio de revisiones por estado y mostramos los 20 estados con la mayor cantidad promedio de revisiones.

Este análisis exhaustivo de los datos de revisiones de franquicias proporciona una comprensión completa de las características clave y las relaciones dentro del conjunto de datos. Estos insights pueden ser de gran utilidad para futuros análisis y decisiones relacionadas con franquicias.

Este análisis exploratorio de datos (EDA) realizado en los conjuntos de datos "df_estado", "Franquicias_inversion.csv" y "Reviews_Completos" nos proporcionó una base sólida para desarrollar un modelo de machine learning de sistema de recomendación de inversión por franquicias, estado, categorías y monto a invertir. A través del EDA, hemos examinado la población y densidad demográfica en los estados, explorado las características clave de las franquicias, como inversiones y unidades, y profundizado en las revisiones de franquicias para entender la satisfacción del cliente. Utilizando esta información, hemos definido indicadores clave de rendimiento (KPIs) que incluyen la "Tasa de Satisfacción del Cliente" y la "Cantidad de Sucursales por Conglomerado de Estados". Además, hemos establecido KPIs adicionales como la "Comparación de Satisfacción del Cliente Promedio por Estado con el Nivel de Satisfacción de la Franquicia Elegida" y la

"Comparación del Nivel de Satisfacción del Cliente en Franquicias Seleccionadas frente al Promedio Nacional". Estos KPIs serán fundamentales para evaluar el éxito de la inversión una vez que haya transcurrido un tiempo desde su ejecución, brindándonos información valiosa para tomar decisiones informadas en el ámbito de las franquicias y las inversiones