Informe_Ecoenergy_Solutions_Dataset

Informe sobre limpieza, normalización y visualización de datos-Ecoenergy_ Solutions

1. Introducción

El objetivo de este informe se centra en analizar de manera exploratoria un dataset sobre el consumo energético de clientes. Se busca identificar patrones relevantes, para su representación se utilizará Looker Studio, facilitando la visualización de los ppales. indicadores. Se ha utilizado Google Colab para la aplicación de técnicas de limpieza, normalización de datos y análisis estadístico.

El proceso incluyó: revisión del estado inicial del dataset, limpieza y estandarización de variables, visualización y análisis de patrones, detección de valores atípicos, preparación del archivo final para Looker Studio

2. Preparación y limpieza del dataset

```
df = pd.read_csv("ecoenergy_consumption_data.csv")
df.head()
df.info()
df.describe()
df.isnull().sum()
```

El objetivo es conocer la estructura que tenemos en nuestro dataset. Identificamos el número de filas y columnas, los tipos de datos y además, los valores faltantes. Ésto nos permite revisar estadísticas básicas de las columnas numéricas.

Es importante saber qué tipo de datos tenemos y dónde hay problemas o posibles problemas.

Normalización de la fecha

```
df["billing_date"] = pd.to_datetime(df["billing_date"])
df["year"] = df["billing_date"].dt.year
df["month"] = df["billing_date"].dt.month
```

La columna billing_date se encontraba en formato string. Para facilitar el análisis temporal, se transformó a formato datetime y se generaron las columnas(year, month). Estas variables permitieron agrupar y analizar tendencias mensuales y anuales.

Búsqueda de patrones y correlaciones

```
##Distribución del consumo
sns.histplot(df["consumption_kwh"])
      (Búsqueda de consumo homogénea o heterogénea)
```

```
##Coste total
sns.histplot(df["total_cost"])
      (Posibles patrones de consumo, tarifas anómalas)
##Consumo por región
sns.boxplot(x="region", y="consumption_kwh", data=df)
      (IDentificar regiones con mayor consumo,variabilidad)
##Contrato vs Consumo
sns.boxplot(x="contract_type", y="consumption_kwh", data=df)
      (Comprobar si ciertos contratos generan mayor consumo)
#Búsqueda de correlaciones
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap="coolwarm")
      (Posible RElación significativa: consumo, coste,
      emisiones, %de energía renovable)
```

Detección de valores atípicos

```
sns.boxplot(df["consumption_kwh"])
sns.boxplot(df["total_cost"])
sns.boxplot(df["co2_emissions"])
      (Se analizaron outliers en: consumo, coste y emisiones)
```

Creación del Dashboard (Looker Studio)

1. **Ranking de clientes ordenados por consumo**
   Muestra los clientes con mayor consumo. Puede observarse que existen filiales o contratos bajo el mismo grupo. Los consumos son muy elevados por lo que generan mayores ingresos y mayores emisiones.
2. **Consumo por región**
   Muestran diferencias claras de consumo. LAs regiones de centro y norte tienen mayor consumo. Evidencia la demanda dependiendo de la localización geográfica.
3. **Consumo por tipo de contrato**
   Los clientes comerciales consumen mucho más, se evidencia frente a los consumos más homogéneos de los residenciales. Por lo que puede afirmarse que el tipo de contrato es un factor clave.
4. **Dispersión Consumo vs Coste total**
   La relación que presenta es lineal: más consumo, mayor coste. Confirma un sistema de facturación consistente
5. **Renovables por región o tipo de contrato**
   Sistema de facturación consistente, a mayor consumo mayor coste. Se aprecia qué regiones o contratos usan más renovables, no todas adoptan renovables por igual.

Normalización de variables numéricas y reducción de dimensionalidad

Al intentar visualizar las variables region y contract_type, la app no permitia generar los gráficos: se debía al elevado número de categorías dentro de las mismas. Esto hacía que mostrará errores. Para solventarlo se optó por seleccionar las 5 regiones y los 5 tipos de contrato con mayor consumo. Esto evitó los errores y permitió generar los gráficos.

```
# Top 5 regiones por consumo total
top_regions =
df.groupby("region")["consumption_kwh"].sum().sort_values(ascend
ing=False).head(5).index
df_top = df[df["region"].isin(top_regions)]
# Top 5 tipos de contrato por consumo total
top_contracts =
df.groupby("contract_type")["consumption_kwh"].sum().sort_values
(ascending=False).head(5).index
df_top = df[df["contract_type"].isin(top_contracts)]
```

Fue necesario escalar las variables numéricas(*consumption_kwh*, *total_cost*, *co2_emissions* y *cost_per_kwh)* presentaban rangos muy diferentes, dificulta la interpretación de patrones. Para solucionarlo se aplicó StandarScaler.

```
num_cols = ["consumption_kwh", "total_cost", "co2_emissions",
"cost_per_kwh"]
scaler = StandardScaler()
df_top[num_cols] = scaler.fit_transform(df_top[num_cols])
```

Conclusión Final

Tras lo realizado se obtuvieron varias conclusiones:

- El consumo es el ppal. factor que determina el coste, con una correlación lineal.
- Se identificaron diferencias significativas entre regiones y tipos de contrato.
- Aunque no había valores nulos, los valores atípicos explicaban la variabilidad del consumo de los clientes.
- La reducción de la dimensionalidad de ciertas variables, permitió extraer insights de forma más clara