TAREA_1_JAVIERA_MONTESINOS

April 30, 2025

Tarea 1 2025

Instrucciones

Su notebook con las respuestas a la tarea se deben entregar a mas tardar el dia 21/04/25 hasta las 21:00, subiendolo al repositorio en la carpeta tareas/2025.

Es importante considerar que el código debe poder ejecutarse en cualquier computadora con la data original del repositorio. Recordar la convencion para el nombre de archivo ademas de incluir en su documento titulos y encabezados por seccion. La data a utilizar es **machine_failure_data.csv**.

Las variables tienen la siguiente descripcion:

- Date: data medida en frecuencia diaria
- Location: ubicacion del medidor
- Min_Temp: temperatura minima observada
- Max_Temp: temperatura maxima observada
- Leakage: Filtracion medida en el area
- Evaporation: Tasa de evaporacion
- Electricity: Consumo electrico KW
- Parameter#: Diferentes sensores de reportando direccion y velocidad de viento en distintos momentos del dia, asi como otras metricas relevantes.
- Failure today: El sensor reporta fallo (o no)

0.1 TAREA

[1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn
import scipy
from scipy.stats import nbinom
import seaborn as sns
from statsmodels.iolib.summary2 import summary_col
import math
import warnings
warnings.filterwarnings("ignore")
```

```
%matplotlib inline
```

0.1.1 1. Cargar la base de datos en el ambiente. Identifique los tipos de datos que se encuentran en la base, realice estadisticas descriptivas sobre las variables importantes (Hint: Revisar la distribuciones, datos faltantes, outliers, etc.) y limpie las variables cuando sea necesario.

R: Primero que nada cargamos la data, trabajamos la data para que el formato sea adecuado para los procesamientos posteriores y las limpiezas estimadas, los procesos aplicados estaran escritos a lo largo del codigo

[2]:
```
#Cargar la data
df = pd.read_csv('DATA/machine_failure_data.csv')
```

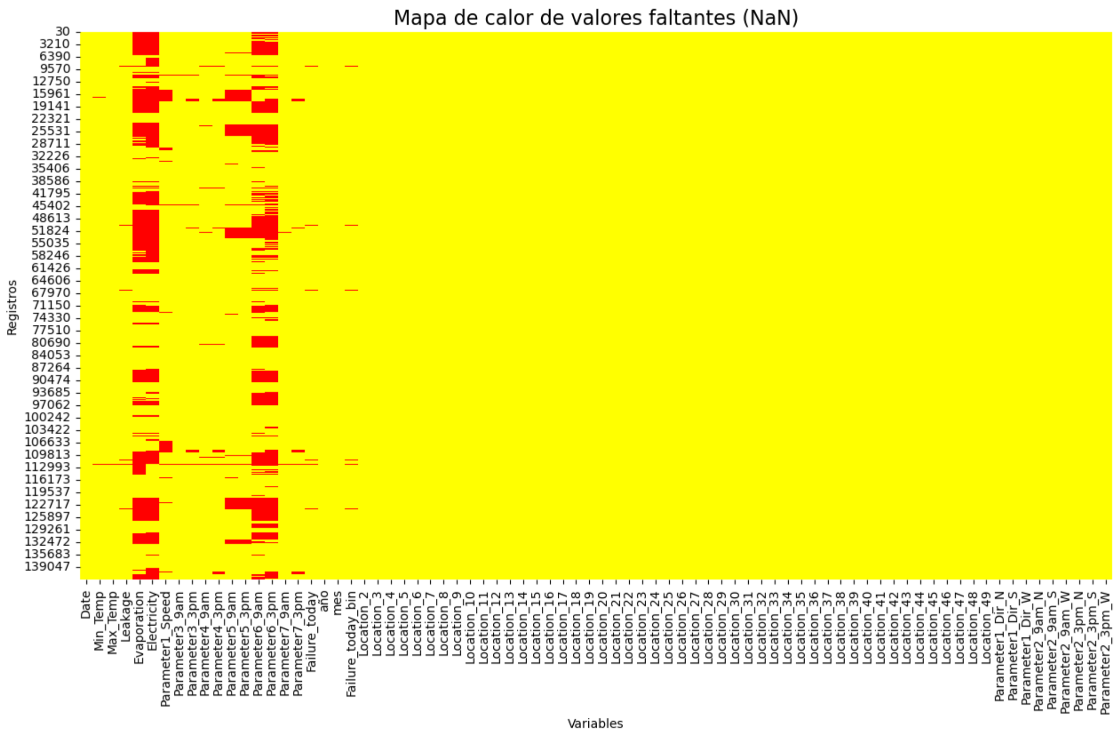
[3]:
```
#filtrar por las fechas de interes(posterior a 2009) y generar columnas de año␣
↪y mes
df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%Y')
df['año'] = df['Date'].dt.year
df['mes'] = df['Date'].dt.month
df_02 = df[df['año'] >= 2009]
```

[4]:
```
#Generar una variable binaria en base a la columna Failure
df_02['Failure_today_bin'] = df_02['Failure_today'].map({'Yes': 1, 'No': 0})
```

[5]:
```
#Generalizar las direcciones en 4 opciones solamente, norte(N), sur(S), este(E)␣
↪y oeste(W), para simplificar analisis
df_02['Parameter1_Dir'] = df['Parameter1_Dir'].str[0]
df_02['Parameter2_9am'] = df['Parameter2_9am'].str[0]
df_02['Parameter2_3pm'] = df['Parameter2_3pm'].str[0]
```

[6]:
```
#Generar variables binarias en base a cada categoria de las variables␣
↪categoricas
df_02 = pd.get_dummies(df_02, prefix=['Location','Parameter1_Dir',␣
↪'Parameter2_9am','Parameter2_3pm'], columns=['Location','Parameter1_Dir',␣
↪'Parameter2_9am','Parameter2_3pm'],dtype = int,drop_first=True)
```

[7]:
```
#Generamos un mapa de calor para identificar visualmente las variables con mas␣
↪NaN
plt.figure(figsize=(15, 8))
sns.heatmap(df_02.isnull(), cbar=False, cmap=sns.color_palette(["yellow",␣
↪"red"]))  # yellow = no NaN, red = NaN
plt.title("Mapa de calor de valores faltantes (NaN)", fontsize=16)
plt.xlabel("Variables")
plt.ylabel("Registros")
plt.show()
```

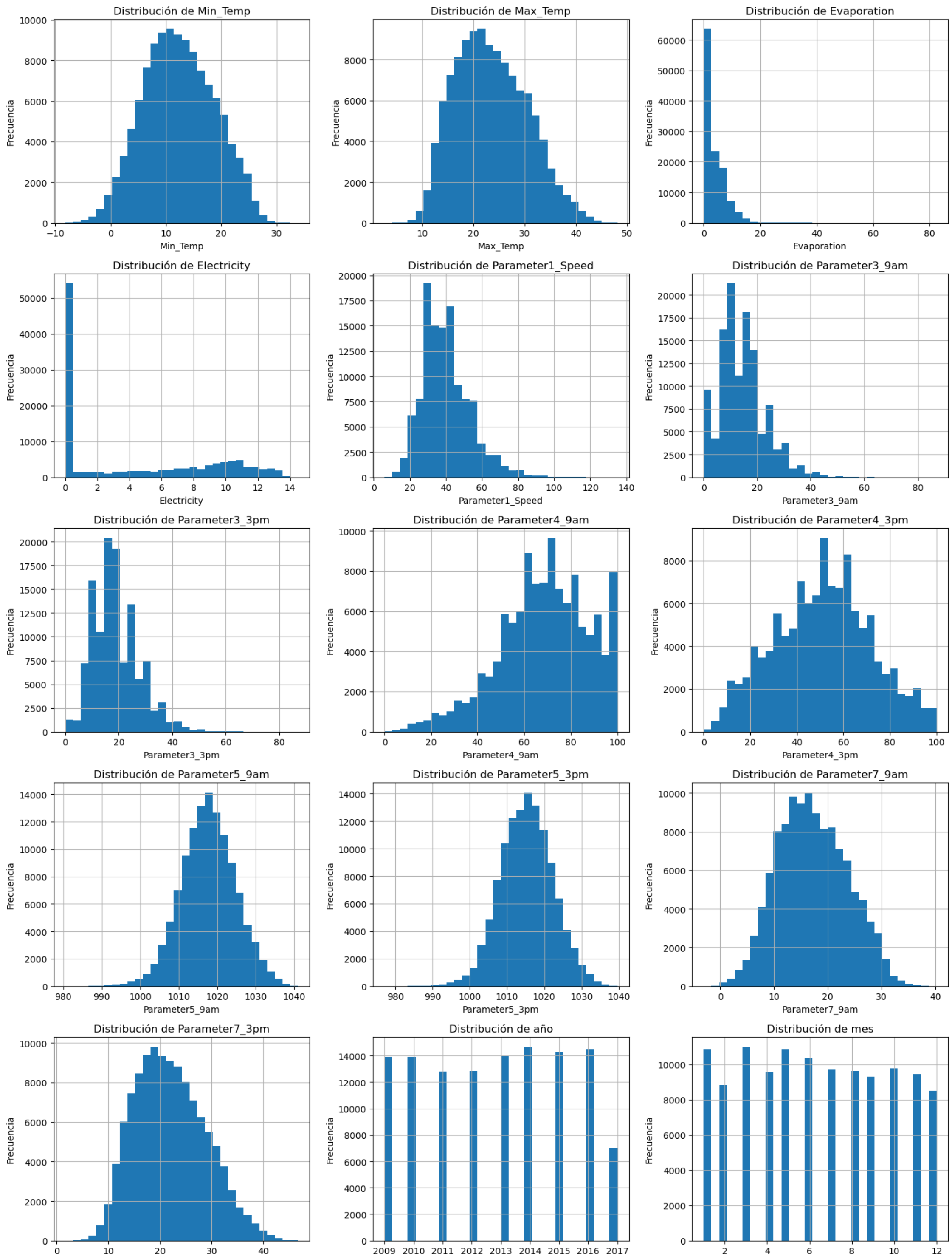


[8]:
```
#Habiendo obtenido que las variables mas destacadas por su cantidad de NaN son Electricity y Evaporation
#Por lo cual, para no eliminar tan gran cantidad de datos generamos una columna indicadora de las veces que las variables no tuvieran valor para poder reconocerlos mas adelante en el analisis
df_03=df_02.copy()
df_03['Electricity_bin'] = df_03['Electricity'].isna().astype(int)
df_03.Electricity=df_03.Electricity.fillna(0)
df_03['Evaporation_bin'] = df_03['Evaporation'].isna().astype(int)
df_03.Evaporation=df_03.Evaporation.fillna(0)
```

[9]:
```
#Posteriormente eliminamos las variablea continuación en los casos que, se hayan generado otras columnas con su informacion y ya no sean necesarias
#si es que no tienen valores o en el caso que su pertenencia en el df pueda afectar negativamente las estimaciones a continuacion como en los casos de 'Parameter6_9am', 'Parameter6_3pm' y 'Leakage'
df_03 = df_03.drop(['Date','Parameter6_9am', 'Parameter6_3pm','Leakage','Failure_today','Location_2','Location_24','Location_25','Locati
axis=1)
df_03=df_03.dropna()
```

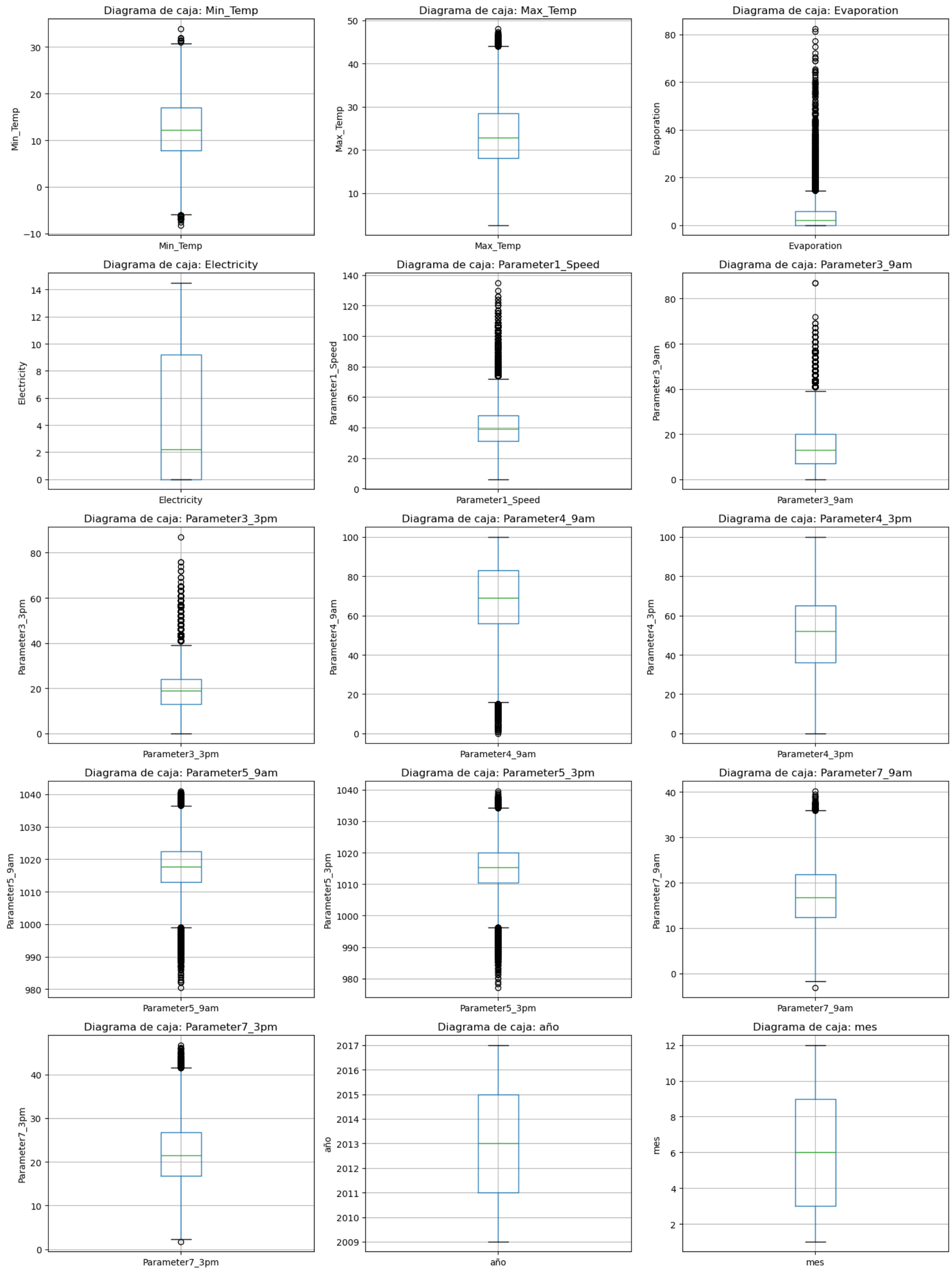
[10]:
```
#Graficamos las distribuciones de las variables para facilitar su analisis
```

```
columnas_numericas = (df.drop(['Location','Leakage','Parameter6_9am',␣
↪'Parameter6_3pm'],axis=1)).select_dtypes(include='number').columns

num_columnas = 3
num_graficos = len(columnas_numericas)
num_filas = math.ceil(num_graficos / num_columnas)

fig, axes = plt.subplots(num_filas, num_columnas, figsize=(num_columnas * 5,␣
↪num_filas * 4))
axes = axes.flatten()

# Generar cada histograma
for i, col in enumerate(columnas_numericas):
    df_03[col].hist(ax=axes[i], bins=30)
    axes[i].set_title(f'Distribución de {col}')
    axes[i].set_xlabel(col)
    axes[i].set_ylabel('Frecuencia')

for j in range(i+1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```



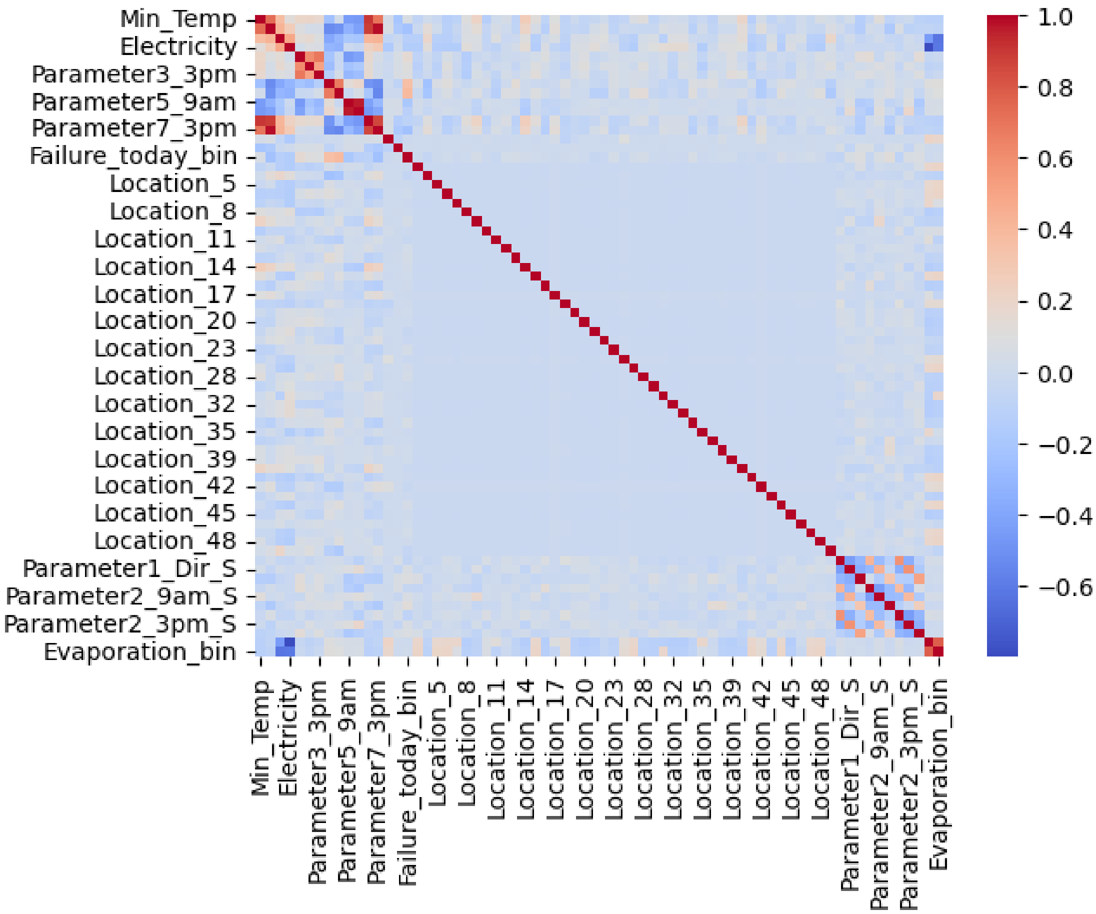
[11]:
```
#Graficar con boxplot las variables para identificar outliers
```

```
#Y aun que se reconoce la existencia de estos mismos de todas formas se␣
↪mantendran en el df para los futuros analisis

columnas_numericas = (df.drop(['Location','Leakage','Parameter6_9am',␣
↪'Parameter6_3pm'],axis=1)).select_dtypes(include='number').columns

num_columnas = 3
num_graficos = len(columnas_numericas)
num_filas = math.ceil(num_graficos / num_columnas)
fig, axes = plt.subplots(num_filas, num_columnas, figsize=(num_columnas * 5,␣
↪num_filas * 4))
axes = axes.flatten()

for i, col in enumerate(columnas_numericas):
    df_03.boxplot(column=col, ax=axes[i])
    axes[i].set_title(f'Diagrama de caja: {col}')
    axes[i].set_ylabel(col)

for j in range(i+1, len(axes)):
    fig.delaxes(axes[j])
plt.tight_layout()
plt.show()
```



[12]:
```python
correlation_matrix = df_03.corr()
sns.heatmap(correlation_matrix, cmap='coolwarm', annot=False)
```

```
plt.show()
```



[13]:
```
#REVISAMOS LAS CORRELACIONES SOBRE 0.8 O BAJO -0.8 PARA PREVEER␣
↪MULTICOLINEALIDAD

upper = correlation_matrix.where(np.triu(np.ones(correlation_matrix.shape),␣
↪k=1).astype(bool))
high_corr = upper.stack().reset_index()
high_corr.columns = ['Variable1', 'Variable2', 'Correlacion']
high_corr_filtrada = high_corr[(high_corr['Correlacion'] > 0.8) |␣
↪(high_corr['Correlacion'] < -0.8)]
print(high_corr_filtrada.sort_values(by='Correlacion', ascending=False))
```

```
          Variable1       Variable2  Correlacion
79         Max_Temp  Parameter7_3pm     0.984704
585  Parameter5_9am  Parameter5_3pm     0.961581
10         Min_Temp  Parameter7_9am     0.902489
78         Max_Temp  Parameter7_9am     0.882529
```

```
704  Parameter7_9am  Parameter7_3pm     0.857249
```

[14]:
```
#Dado que parte de la correlación ocurre por parametros que miden lo mismo en
 ↪distintas horas(por ende tienden a ser parecidos) dejaremos solo 1 horario
 ↪por parametros
#asumiendo la correlación que pueda quedar del parametro 7 restante con el
 ↪maximo y minimo de la temperatura, por que el parametro 7 trabaja con
 ↪temperatura
df_03=df_03.drop(['Parameter5_3pm','Parameter7_3pm','Max_Temp'],axis=1)
```

0.1.2 2. Ejecute un modelo de probabilidad lineal (*MCO*) que permita explicar la probabilidad de que un dia se reporte fallo medido por sensor, a partir de las informacion disponible. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: El modelo explica el 28% de la varianza en la variable dependiente, lo demas del resultado del modelo podemos mencionar que, luego de haber excluido las variables que generaran alta correlación en el recuadro anterior, podemos mencionar que las variables que mas afectan positivamente a la estimacion del fallo son las variables Parameter2_9am y Parameter2_3pm para el oeste(W) y sur(S). Y por otro lado las que mas afectan negativamente son las variables de locaciones destacando entre ellas la locacion 36, 6, 26 y 20, con mayor proporción negativa

[15]:
```
y=df_03['Failure_today_bin']
X=df_03.drop(['Failure_today_bin'], axis=1)
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Failure_today_bin   R-squared:                       0.280
Model:                            OLS   Adj. R-squared:                  0.280
Method:                 Least Squares   F-statistic:                     701.1
Date:                Thu, 24 Apr 2025   Prob (F-statistic):               0.00
Time:                        23:55:47   Log-Likelihood:                -44179.
No. Observations:              117793   AIC:                         8.849e+04
Df Residuals:                  117726   BIC:                         8.914e+04
Df Model:                          66
Covariance Type:                  HC0
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const                7.7120      0.951      8.113      0.000       5.849
9.575
```

Min_Temp	0.0146	0.001	29.089	0.000	0.014	0.016
Evaporation	-0.0064	0.000	-14.388	0.000	-0.007	-0.005
Electricity	-0.0045	0.000	-9.912	0.000	-0.005	-0.004
Parameter1_Speed	0.0048	0.000	34.053	0.000	0.004	0.005
Parameter3_9am	0.0040	0.000	23.371	0.000	0.004	0.004
Parameter3_3pm	-0.0032	0.000	-17.168	0.000	-0.004	-0.003
Parameter4_9am	0.0068	0.000	65.079	0.000	0.007	0.007
Parameter4_3pm	0.0026	9.08e-05	28.906	0.000	0.002	0.003
Parameter5_9am	-0.0090	0.000	-42.639	0.000	-0.009	-0.009
Parameter7_9am	-0.0126	0.001	-22.826	0.000	-0.014	-0.011
año	0.0005	0.000	1.093	0.275	-0.000	0.001
mes	0.0068	0.000	21.394	0.000	0.006	0.007
Location_3	-0.1011	0.009	-10.895	0.000	-0.119	-0.083
Location_4	0.0895	0.008	10.830	0.000	0.073	0.106
Location_5	-0.1303	0.010	-13.324	0.000	-0.149	-0.111
Location_6	-0.2135	0.010	-20.889	0.000	-0.234	-0.193
Location_7	-0.1284	0.009	-14.005	0.000	-0.146	-0.110
Location_8	-0.0138	0.010	-1.423	0.155	-0.033	0.005
Location_9	-0.0896	0.010	-8.661	0.000	-0.110	-0.069
Location_10	-0.1099	0.009	-11.686	0.000	-0.128	-0.091
Location_11	-0.0427	0.009	-4.805	0.000	-0.060	-0.025
Location_12	-0.0426	0.010	-4.148	0.000	-0.063	-0.022
Location_13	-0.1554	0.010	-15.310	0.000	-0.175	-0.136
Location_14	-0.1187	0.010	-12.052	0.000	-0.138	-0.099

Location_15	-0.1027	0.010	-10.149	0.000	-0.123	-0.083
Location_16	-0.1455	0.010	-14.495	0.000	-0.165	-0.126
Location_17	-0.1375	0.014	-9.551	0.000	-0.166	-0.109
Location_18	-0.1301	0.011	-11.654	0.000	-0.152	-0.108
Location_19	-0.1249	0.011	-11.249	0.000	-0.147	-0.103
Location_20	-0.1738	0.010	-17.689	0.000	-0.193	-0.155
Location_21	-0.1224	0.009	-14.247	0.000	-0.139	-0.106
Location_22	-0.0647	0.009	-7.361	0.000	-0.082	-0.047
Location_23	-0.1049	0.010	-10.530	0.000	-0.124	-0.085
Location_26	-0.2028	0.011	-19.020	0.000	-0.224	-0.182
Location_27	-0.1649	0.010	-16.000	0.000	-0.185	-0.145
Location_28	-0.1413	0.010	-13.645	0.000	-0.162	-0.121
Location_29	-0.0851	0.009	-9.298	0.000	-0.103	-0.067
Location_30	-0.0514	0.010	-5.026	0.000	-0.071	-0.031
Location_32	-0.0328	0.009	-3.642	0.000	-0.050	-0.015
Location_33	-0.0429	0.009	-4.722	0.000	-0.061	-0.025
Location_34	-0.1156	0.010	-11.068	0.000	-0.136	-0.095
Location_35	-0.1217	0.010	-12.749	0.000	-0.140	-0.103
Location_36	-0.2171	0.010	-21.957	0.000	-0.236	-0.198
Location_38	-0.1198	0.011	-11.077	0.000	-0.141	-0.099
Location_39	-0.0992	0.010	-9.995	0.000	-0.119	-0.080
Location_40	-0.1170	0.009	-12.479	0.000	-0.135	-0.099
Location_41	-0.0848	0.009	-8.984	0.000	-0.103	-0.066
Location_42	0.0016	0.009	0.172	0.864	-0.017	0.020

Location_43	-0.0749	0.009	-8.246	0.000	-0.093	-0.057
Location_44	-0.1016	0.011	-9.556	0.000	-0.122	-0.081
Location_45	-0.1561	0.010	-15.977	0.000	-0.175	-0.137
Location_46	-0.0790	0.011	-7.304	0.000	-0.100	-0.058
Location_47	-0.0606	0.011	-5.749	0.000	-0.081	-0.040
Location_48	-0.1705	0.010	-16.820	0.000	-0.190	-0.151
Location_49	-0.0974	0.008	-11.528	0.000	-0.114	-0.081
Parameter1_Dir_N	-0.0203	0.004	-5.753	0.000	-0.027	-0.013
Parameter1_Dir_S	0.0057	0.003	1.652	0.098	-0.001	0.012
Parameter1_Dir_W	0.0075	0.004	1.806	0.071	-0.001	0.016
Parameter2_9am_N	0.0054	0.003	1.801	0.072	-0.000	0.011
Parameter2_9am_S	0.0461	0.003	15.153	0.000	0.040	0.052
Parameter2_9am_W	0.0700	0.004	17.370	0.000	0.062	0.078
Parameter2_3pm_N	-0.0130	0.003	-3.763	0.000	-0.020	-0.006
Parameter2_3pm_S	0.0326	0.003	9.565	0.000	0.026	0.039
Parameter2_3pm_W	0.0404	0.004	9.844	0.000	0.032	0.048
Electricity_bin	-0.0343	0.006	-5.330	0.000	-0.047	-0.022
Evaporation_bin	-0.0284	0.006	-5.161	0.000	-0.039	-0.018

==

Omnibus:	9978.337	Durbin-Watson:	1.799
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12730.953
Skew:	0.803	Prob(JB):	0.00
Kurtosis:	2.883	Cond. No.	2.08e+06

==

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 2.08e+06. This might indicate that there are strong multicollinearity or other numerical problems.

0.1.3 3. Ejecute un modelo *probit* para responder a la pregunta 2. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Los resultados sugieren que factores como la Min_Temp (Coef: 0.0223, aumento de 1grado de temperatura aumenta la probabilidad “Failure_today” en un 2.23%), Evaporation y los valores de algunos parámetros a las 9 am y 3 pm son los que mas influyen por si solos en la probabilidad de que ocurra el evento “Failure_today”, ademas de mencionar que son estadisticamente significativos. La ubicación también tiene un impacto relevante, con algunas locaciones siendo más propensas a fallos que otras, sin embargo estas ultimas en algunos casos no son significativas

[16]:
```python
model = sm.Probit(y, X)
probit_model = model.fit(cov_type='HC0')
print(probit_model.summary())

mfxp = probit_model.get_margeff()
print(mfxp.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.361360
         Iterations 7
                          Probit Regression Results
==============================================================================
Dep. Variable:      Failure_today_bin   No. Observations:               117793
Model:                         Probit   Df Residuals:                   117726
Method:                           MLE   Df Model:                           66
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.3158
Time:                        23:55:49   Log-Likelihood:                -42566.
converged:                       True   LL-Null:                       -62216.
Covariance Type:                  HC0   LLR p-value:                     0.000
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const               25.9865      4.476      5.806      0.000      17.214
34.759
Min_Temp             0.1099      0.003     37.108      0.000       0.104
0.116
Evaporation         -0.0457      0.004    -11.822      0.000      -0.053
-0.038
Electricity          0.0021      0.002      0.949      0.343      -0.002
0.006
Parameter1_Speed     0.0176      0.001     29.723      0.000       0.016
0.019
Parameter3_9am       0.0154      0.001     18.572      0.000       0.014
0.017
Parameter3_3pm      -0.0084      0.001     -9.931      0.000      -0.010
-0.007
```

Parameter4_9am	0.0336	0.001	62.075	0.000	0.033	0.035
Parameter4_3pm	0.0105	0.000	27.511	0.000	0.010	0.011
Parameter5_9am	-0.0355	0.001	-38.626	0.000	-0.037	-0.034
Parameter7_9am	-0.1058	0.003	-32.824	0.000	-0.112	-0.099
año	0.0030	0.002	1.362	0.173	-0.001	0.007
mes	0.0326	0.002	20.091	0.000	0.029	0.036
Location_3	-0.3678	0.046	-7.959	0.000	-0.458	-0.277
Location_4	0.0694	0.062	1.125	0.261	-0.051	0.190
Location_5	-0.4222	0.047	-8.956	0.000	-0.515	-0.330
Location_6	-1.0082	0.049	-20.740	0.000	-1.103	-0.913
Location_7	-0.5441	0.047	-11.661	0.000	-0.636	-0.453
Location_8	0.2281	0.046	5.008	0.000	0.139	0.317
Location_9	-0.2103	0.045	-4.689	0.000	-0.298	-0.122
Location_10	-0.3205	0.047	-6.823	0.000	-0.413	-0.228
Location_11	-0.3051	0.054	-5.642	0.000	-0.411	-0.199
Location_12	-0.0216	0.046	-0.474	0.636	-0.111	0.068
Location_13	-0.6704	0.044	-15.143	0.000	-0.757	-0.584
Location_14	-0.2921	0.046	-6.315	0.000	-0.383	-0.201
Location_15	-0.2695	0.048	-5.664	0.000	-0.363	-0.176
Location_16	-0.4068	0.047	-8.693	0.000	-0.498	-0.315
Location_17	-0.4513	0.078	-5.776	0.000	-0.604	-0.298
Location_18	-0.4473	0.051	-8.706	0.000	-0.548	-0.347
Location_19	-0.3618	0.048	-7.510	0.000	-0.456	-0.267
Location_20	-0.6167	0.046	-13.293	0.000	-0.708	-0.526

Location_21	-0.7632	0.051	-15.012	0.000	-0.863	-0.664
Location_22	-0.1318	0.051	-2.564	0.010	-0.233	-0.031
Location_23	-0.4042	0.045	-9.084	0.000	-0.491	-0.317
Location_26	-1.0020	0.058	-17.348	0.000	-1.115	-0.889
Location_27	-0.6291	0.046	-13.559	0.000	-0.720	-0.538
Location_28	-0.4636	0.045	-10.373	0.000	-0.551	-0.376
Location_29	-0.5132	0.050	-10.351	0.000	-0.610	-0.416
Location_30	-0.0622	0.053	-1.171	0.242	-0.166	0.042
Location_32	-0.0325	0.045	-0.718	0.473	-0.121	0.056
Location_33	-0.0471	0.047	-1.004	0.315	-0.139	0.045
Location_34	-0.4811	0.044	-10.951	0.000	-0.567	-0.395
Location_35	-0.4266	0.047	-9.020	0.000	-0.519	-0.334
Location_36	-0.7792	0.046	-16.829	0.000	-0.870	-0.688
Location_38	-0.3397	0.047	-7.174	0.000	-0.433	-0.247
Location_39	-0.2721	0.046	-5.852	0.000	-0.363	-0.181
Location_40	-0.2692	0.048	-5.666	0.000	-0.362	-0.176
Location_41	-0.2085	0.046	-4.509	0.000	-0.299	-0.118
Location_42	-0.2135	0.073	-2.942	0.003	-0.356	-0.071
Location_43	-0.3031	0.049	-6.176	0.000	-0.399	-0.207
Location_44	-0.3465	0.045	-7.627	0.000	-0.436	-0.257
Location_45	-0.5987	0.045	-13.201	0.000	-0.688	-0.510
Location_46	-0.1319	0.048	-2.732	0.006	-0.227	-0.037
Location_47	-0.1221	0.047	-2.607	0.009	-0.214	-0.030
Location_48	-0.6197	0.047	-13.169	0.000	-0.712	-0.527

```
Location_49          -0.8197      0.059    -13.861      0.000      -0.936
-0.704
Parameter1_Dir_N     -0.1356      0.020     -6.927      0.000      -0.174
-0.097
Parameter1_Dir_S      0.0159      0.018      0.896      0.370      -0.019
0.051
Parameter1_Dir_W      0.0298      0.021      1.444      0.149      -0.011
0.070
Parameter2_9am_N      0.0413      0.017      2.440      0.015       0.008
0.075
Parameter2_9am_S      0.2823      0.016     17.685      0.000       0.251
0.314
Parameter2_9am_W      0.3251      0.018     17.751      0.000       0.289
0.361
Parameter2_3pm_N     -0.0886      0.019     -4.682      0.000      -0.126
-0.052
Parameter2_3pm_S      0.1088      0.017      6.318      0.000       0.075
0.143
Parameter2_3pm_W      0.1324      0.020      6.470      0.000       0.092
0.173
Electricity_bin       0.0595      0.029      2.050      0.040       0.003
0.116
Evaporation_bin      -0.2221      0.029     -7.691      0.000      -0.279
-0.165
================================================================================
====
       Probit Marginal Effects
=====================================
Dep. Variable:      Failure_today_bin
Method:                          dydx
At:                           overall
================================================================================
====
                       dy/dx    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
Min_Temp              0.0223      0.001     38.012      0.000       0.021
0.023
Evaporation          -0.0093      0.001    -11.930      0.000      -0.011
-0.008
Electricity           0.0004      0.000      0.949      0.342      -0.000
0.001
Parameter1_Speed      0.0036      0.000     30.170      0.000       0.003
0.004
Parameter3_9am        0.0031      0.000     18.674      0.000       0.003
0.003
Parameter3_3pm       -0.0017      0.000     -9.944      0.000      -0.002
```

```
-0.001
Parameter4_9am        0.0068      0.000     66.133      0.000       0.007
0.007
Parameter4_3pm        0.0021   7.65e-05     27.714      0.000       0.002
0.002
Parameter5_9am       -0.0072      0.000    -39.483      0.000      -0.008
-0.007
Parameter7_9am       -0.0215      0.001    -33.454      0.000      -0.023
-0.020
año                   0.0006      0.000      1.362      0.173      -0.000
0.001
mes                   0.0066      0.000     20.183      0.000       0.006
0.007
Location_3           -0.0746      0.009     -7.971      0.000      -0.093
-0.056
Location_4            0.0141      0.012      1.125      0.260      -0.010
0.039
Location_5           -0.0856      0.010     -8.971      0.000      -0.104
-0.067
Location_6           -0.2044      0.010    -20.943      0.000      -0.224
-0.185
Location_7           -0.1103      0.009    -11.692      0.000      -0.129
-0.092
Location_8            0.0462      0.009      5.010      0.000       0.028
0.064
Location_9           -0.0426      0.009     -4.689      0.000      -0.060
-0.025
Location_10          -0.0650      0.010     -6.830      0.000      -0.084
-0.046
Location_11          -0.0619      0.011     -5.647      0.000      -0.083
-0.040
Location_12          -0.0044      0.009     -0.474      0.636      -0.022
0.014
Location_13          -0.1359      0.009    -15.217      0.000      -0.153
-0.118
Location_14          -0.0592      0.009     -6.316      0.000      -0.078
-0.041
Location_15          -0.0546      0.010     -5.668      0.000      -0.074
-0.036
Location_16          -0.0825      0.009     -8.718      0.000      -0.101
-0.064
Location_17          -0.0915      0.016     -5.776      0.000      -0.123
-0.060
Location_18          -0.0907      0.010     -8.721      0.000      -0.111
-0.070
Location_19          -0.0734      0.010     -7.522      0.000      -0.092
-0.054
Location_20          -0.1250      0.009    -13.341      0.000      -0.143
```

						-0.107
Location_21	-0.1547	0.010	-15.065	0.000	-0.175	-0.135
Location_22	-0.0267	0.010	-2.565	0.010	-0.047	-0.006
Location_23	-0.0820	0.009	-9.104	0.000	-0.100	-0.064
Location_26	-0.2031	0.012	-17.434	0.000	-0.226	-0.180
Location_27	-0.1275	0.009	-13.617	0.000	-0.146	-0.109
Location_28	-0.0940	0.009	-10.397	0.000	-0.112	-0.076
Location_29	-0.1040	0.010	-10.378	0.000	-0.124	-0.084
Location_30	-0.0126	0.011	-1.171	0.242	-0.034	0.009
Location_32	-0.0066	0.009	-0.718	0.473	-0.025	0.011
Location_33	-0.0096	0.010	-1.004	0.315	-0.028	0.009
Location_34	-0.0975	0.009	-10.985	0.000	-0.115	-0.080
Location_35	-0.0865	0.010	-9.033	0.000	-0.105	-0.068
Location_36	-0.1580	0.009	-16.949	0.000	-0.176	-0.140
Location_38	-0.0689	0.010	-7.182	0.000	-0.088	-0.050
Location_39	-0.0552	0.009	-5.857	0.000	-0.074	-0.037
Location_40	-0.0546	0.010	-5.664	0.000	-0.073	-0.036
Location_41	-0.0423	0.009	-4.511	0.000	-0.061	-0.024
Location_42	-0.0433	0.015	-2.943	0.003	-0.072	-0.014
Location_43	-0.0615	0.010	-6.184	0.000	-0.081	-0.042
Location_44	-0.0702	0.009	-7.639	0.000	-0.088	-0.052
Location_45	-0.1214	0.009	-13.251	0.000	-0.139	-0.103
Location_46	-0.0267	0.010	-2.732	0.006	-0.046	-0.008
Location_47	-0.0248	0.009	-2.608	0.009	-0.043	-0.006
Location_48	-0.1256	0.010	-13.222	0.000	-0.144	

```
-0.107
Location_49          -0.1662       0.012    -13.908       0.000      -0.190
-0.143
Parameter1_Dir_N     -0.0275       0.004     -6.935       0.000      -0.035
-0.020
Parameter1_Dir_S      0.0032       0.004      0.896       0.370      -0.004
0.010
Parameter1_Dir_W      0.0060       0.004      1.444       0.149      -0.002
0.014
Parameter2_9am_N      0.0084       0.003      2.440       0.015       0.002
0.015
Parameter2_9am_S      0.0572       0.003     17.763       0.000       0.051
0.064
Parameter2_9am_W      0.0659       0.004     17.831       0.000       0.059
0.073
Parameter2_3pm_N     -0.0180       0.004     -4.684       0.000      -0.025
-0.010
Parameter2_3pm_S      0.0221       0.003      6.321       0.000       0.015
0.029
Parameter2_3pm_W      0.0268       0.004      6.473       0.000       0.019
0.035
Electricity_bin       0.0121       0.006      2.051       0.040       0.001
0.024
Evaporation_bin      -0.0450       0.006     -7.714       0.000      -0.056
-0.034
================================================================================
====
```

0.1.4 4. Ejecute un modelo *logit* para responder a la pregunta 2. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: En este modelo podemos mencionar que dentro de las variables que generan un mayor impacto en el valor resultante de la variable dependiente podemos mencionar que dentro de los que aumentan la probabilidad de failure destacan Min_Temp con un coeficiente estimado de 0.2039 y Electricity con 0.0098, ambos coeficientes significativos. y dentro de los que reducen la probabilidad destaca Evaporation con un coeficiente estimado de -0.1035 igualmente siendo significativo.

[17]:
```
model = sm.Logit(y, X)
logit_model = model.fit(cov_type='HC0')
print(logit_model.summary())

mfxl = logit_model.get_margeff()
print(mfxl.summary())

params = logit_model.params
conf = logit_model.conf_int()
conf['Odds Ratio'] = params
conf.columns = ['Odds Ratio', '5%', '95%']
```

```
print("Odds Ratios")
print(np.exp(conf).iloc[1:17 , ])
```

```
Optimization terminated successfully.
         Current function value: 0.359987
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:      Failure_today_bin   No. Observations:               117793
Model:                          Logit   Df Residuals:                   117726
Method:                           MLE   Df Model:                           66
Date:                Thu, 24 Apr 2025   Pseudo R-squ.:                  0.3184
Time:                        23:57:23   Log-Likelihood:                -42404.
converged:                       True   LL-Null:                       -62216.
Covariance Type:                  HC0   LLR p-value:                     0.000
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const               45.1758      7.932      5.696      0.000      29.630
60.722
Min_Temp             0.2039      0.005     39.053      0.000       0.194
0.214
Evaporation         -0.1035      0.007    -15.582      0.000      -0.117
-0.090
Electricity          0.0098      0.004      2.565      0.010       0.002
0.017
Parameter1_Speed     0.0311      0.001     29.704      0.000       0.029
0.033
Parameter3_9am       0.0268      0.001     18.199      0.000       0.024
0.030
Parameter3_3pm      -0.0137      0.001     -9.209      0.000      -0.017
-0.011
Parameter4_9am       0.0605      0.001     62.779      0.000       0.059
0.062
Parameter4_3pm       0.0182      0.001     27.166      0.000       0.017
0.019
Parameter5_9am      -0.0621      0.002    -38.284      0.000      -0.065
-0.059
Parameter7_9am      -0.1967      0.006    -34.469      0.000      -0.208
-0.185
año                  0.0054      0.004      1.378      0.168      -0.002
0.013
mes                  0.0566      0.003     19.419      0.000       0.051
0.062
Location_3          -0.6673      0.082     -8.128      0.000      -0.828
```

						-0.506
Location_4	0.0776	0.110	0.704	0.482	-0.139	0.294
Location_5	-0.7510	0.084	-8.894	0.000	-0.917	-0.586
Location_6	-1.8410	0.086	-21.499	0.000	-2.009	-1.673
Location_7	-0.9748	0.083	-11.741	0.000	-1.138	-0.812
Location_8	0.4392	0.081	5.391	0.000	0.279	0.599
Location_9	-0.3206	0.079	-4.038	0.000	-0.476	-0.165
Location_10	-0.5741	0.084	-6.829	0.000	-0.739	-0.409
Location_11	-0.6053	0.097	-6.234	0.000	-0.796	-0.415
Location_12	-0.0372	0.081	-0.460	0.646	-0.196	0.121
Location_13	-1.2115	0.078	-15.492	0.000	-1.365	-1.058
Location_14	-0.4609	0.083	-5.576	0.000	-0.623	-0.299
Location_15	-0.4729	0.085	-5.572	0.000	-0.639	-0.307
Location_16	-0.7847	0.084	-9.338	0.000	-0.949	-0.620
Location_17	-0.6920	0.141	-4.915	0.000	-0.968	-0.416
Location_18	-0.8018	0.091	-8.819	0.000	-0.980	-0.624
Location_19	-0.6528	0.086	-7.624	0.000	-0.821	-0.485
Location_20	-1.0984	0.083	-13.254	0.000	-1.261	-0.936
Location_21	-1.3801	0.091	-15.165	0.000	-1.558	-1.202
Location_22	-0.2956	0.093	-3.192	0.001	-0.477	-0.114
Location_23	-0.7440	0.079	-9.409	0.000	-0.899	-0.589
Location_26	-1.7827	0.103	-17.270	0.000	-1.985	-1.580
Location_27	-1.1515	0.083	-13.826	0.000	-1.315	-0.988
Location_28	-0.8242	0.080	-10.324	0.000	-0.981	-0.668
Location_29	-0.9530	0.089	-10.746	0.000	-1.127	

						-0.779
Location_30	-0.1451	0.094	-1.542	0.123	-0.330	0.039
Location_32	-0.0282	0.080	-0.352	0.725	-0.186	0.129
Location_33	-0.0675	0.084	-0.806	0.420	-0.232	0.097
Location_34	-0.8837	0.078	-11.284	0.000	-1.037	-0.730
Location_35	-0.7539	0.085	-8.909	0.000	-0.920	-0.588
Location_36	-1.4282	0.082	-17.321	0.000	-1.590	-1.267
Location_38	-0.5941	0.084	-7.084	0.000	-0.758	-0.430
Location_39	-0.4910	0.084	-5.846	0.000	-0.656	-0.326
Location_40	-0.3694	0.085	-4.349	0.000	-0.536	-0.203
Location_41	-0.3699	0.082	-4.489	0.000	-0.531	-0.208
Location_42	-0.4384	0.130	-3.366	0.001	-0.694	-0.183
Location_43	-0.6048	0.088	-6.895	0.000	-0.777	-0.433
Location_44	-0.6341	0.081	-7.794	0.000	-0.794	-0.475
Location_45	-1.0814	0.080	-13.443	0.000	-1.239	-0.924
Location_46	-0.2418	0.086	-2.804	0.005	-0.411	-0.073
Location_47	-0.2383	0.083	-2.877	0.004	-0.401	-0.076
Location_48	-1.1328	0.085	-13.352	0.000	-1.299	-0.967
Location_49	-1.4955	0.105	-14.248	0.000	-1.701	-1.290
Parameter1_Dir_N	-0.2482	0.035	-7.166	0.000	-0.316	-0.180
Parameter1_Dir_S	0.0163	0.031	0.520	0.603	-0.045	0.078
Parameter1_Dir_W	0.0349	0.036	0.962	0.336	-0.036	0.106
Parameter2_9am_N	0.0751	0.030	2.504	0.012	0.016	0.134
Parameter2_9am_S	0.5099	0.028	18.014	0.000	0.454	0.565
Parameter2_9am_W	0.5843	0.032	18.102	0.000	0.521	

```
0.648
Parameter2_3pm_N    -0.1650      0.033     -4.931      0.000      -0.231
-0.099
Parameter2_3pm_S     0.1854      0.030      6.088      0.000       0.126
0.245
Parameter2_3pm_W     0.2275      0.036      6.300      0.000       0.157
0.298
Electricity_bin      0.1382      0.051      2.710      0.007       0.038
0.238
Evaporation_bin     -0.4631      0.049     -9.410      0.000      -0.559
-0.367
================================================================================
====
        Logit Marginal Effects
=====================================
Dep. Variable:      Failure_today_bin
Method:                          dydx
At:                           overall
================================================================================
====
                      dy/dx    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
Min_Temp             0.0232      0.001     39.936      0.000       0.022
0.024
Evaporation         -0.0118      0.001    -15.753      0.000      -0.013
-0.010
Electricity          0.0011      0.000      2.566      0.010       0.000
0.002
Parameter1_Speed     0.0035      0.000     30.263      0.000       0.003
0.004
Parameter3_9am       0.0030      0.000     18.302      0.000       0.003
0.003
Parameter3_3pm      -0.0016      0.000     -9.225      0.000      -0.002
-0.001
Parameter4_9am       0.0069      0.000     67.380      0.000       0.007
0.007
Parameter4_3pm       0.0021   7.54e-05     27.395      0.000       0.002
0.002
Parameter5_9am      -0.0071      0.000    -39.196      0.000      -0.007
-0.007
Parameter7_9am      -0.0224      0.001    -35.067      0.000      -0.024
-0.021
año                  0.0006      0.000      1.378      0.168      -0.000
0.001
mes                  0.0064      0.000     19.498      0.000       0.006
0.007
```

Location_3	-0.0759	0.009	-8.140	0.000	-0.094	-0.058
Location_4	0.0088	0.013	0.704	0.482	-0.016	0.033
Location_5	-0.0854	0.010	-8.910	0.000	-0.104	-0.067
Location_6	-0.2093	0.010	-21.715	0.000	-0.228	-0.190
Location_7	-0.1108	0.009	-11.770	0.000	-0.129	-0.092
Location_8	0.0499	0.009	5.391	0.000	0.032	0.068
Location_9	-0.0364	0.009	-4.038	0.000	-0.054	-0.019
Location_10	-0.0653	0.010	-6.838	0.000	-0.084	-0.047
Location_11	-0.0688	0.011	-6.240	0.000	-0.090	-0.047
Location_12	-0.0042	0.009	-0.460	0.646	-0.022	0.014
Location_13	-0.1377	0.009	-15.575	0.000	-0.155	-0.120
Location_14	-0.0524	0.009	-5.578	0.000	-0.071	-0.034
Location_15	-0.0538	0.010	-5.577	0.000	-0.073	-0.035
Location_16	-0.0892	0.010	-9.374	0.000	-0.108	-0.071
Location_17	-0.0787	0.016	-4.914	0.000	-0.110	-0.047
Location_18	-0.0911	0.010	-8.837	0.000	-0.111	-0.071
Location_19	-0.0742	0.010	-7.639	0.000	-0.093	-0.055
Location_20	-0.1249	0.009	-13.311	0.000	-0.143	-0.106
Location_21	-0.1569	0.010	-15.225	0.000	-0.177	-0.137
Location_22	-0.0336	0.011	-3.194	0.001	-0.054	-0.013
Location_23	-0.0846	0.009	-9.432	0.000	-0.102	-0.067
Location_26	-0.2027	0.012	-17.355	0.000	-0.226	-0.180
Location_27	-0.1309	0.009	-13.890	0.000	-0.149	-0.112
Location_28	-0.0937	0.009	-10.353	0.000	-0.111	-0.076

Location_29	-0.1083	0.010	-10.771	0.000	-0.128	-0.089
Location_30	-0.0165	0.011	-1.542	0.123	-0.037	0.004
Location_32	-0.0032	0.009	-0.352	0.725	-0.021	0.015
Location_33	-0.0077	0.010	-0.806	0.420	-0.026	0.011
Location_34	-0.1005	0.009	-11.322	0.000	-0.118	-0.083
Location_35	-0.0857	0.010	-8.924	0.000	-0.105	-0.067
Location_36	-0.1624	0.009	-17.464	0.000	-0.181	-0.144
Location_38	-0.0675	0.010	-7.095	0.000	-0.086	-0.049
Location_39	-0.0558	0.010	-5.852	0.000	-0.075	-0.037
Location_40	-0.0420	0.010	-4.348	0.000	-0.061	-0.023
Location_41	-0.0420	0.009	-4.491	0.000	-0.060	-0.024
Location_42	-0.0498	0.015	-3.367	0.001	-0.079	-0.021
Location_43	-0.0687	0.010	-6.906	0.000	-0.088	-0.049
Location_44	-0.0721	0.009	-7.808	0.000	-0.090	-0.054
Location_45	-0.1229	0.009	-13.504	0.000	-0.141	-0.105
Location_46	-0.0275	0.010	-2.805	0.005	-0.047	-0.008
Location_47	-0.0271	0.009	-2.879	0.004	-0.046	-0.009
Location_48	-0.1288	0.010	-13.414	0.000	-0.148	-0.110
Location_49	-0.1700	0.012	-14.291	0.000	-0.193	-0.147
Parameter1_Dir_N	-0.0282	0.004	-7.172	0.000	-0.036	-0.021
Parameter1_Dir_S	0.0019	0.004	0.520	0.603	-0.005	0.009
Parameter1_Dir_W	0.0040	0.004	0.962	0.336	-0.004	0.012
Parameter2_9am_N	0.0085	0.003	2.503	0.012	0.002	0.015
Parameter2_9am_S	0.0580	0.003	18.075	0.000	0.052	0.064

```
Parameter2_9am_W     0.0664      0.004     18.183      0.000       0.059
0.074
Parameter2_3pm_N    -0.0188      0.004     -4.933      0.000      -0.026
-0.011
Parameter2_3pm_S     0.0211      0.003      6.093      0.000       0.014
0.028
Parameter2_3pm_W     0.0259      0.004      6.304      0.000       0.018
0.034
Electricity_bin      0.0157      0.006      2.710      0.007       0.004
0.027
Evaporation_bin     -0.0526      0.006     -9.435      0.000      -0.064
-0.042
==============================================================================
====
Odds Ratios
                  Odds Ratio        5%       95%
Min_Temp            1.213742  1.238844  1.226229
Evaporation         0.890020  0.913497  0.901682
Electricity         1.002322  1.017494  1.009879
Parameter1_Speed    1.029508  1.033747  1.031625
Parameter3_9am      1.024201  1.030131  1.027162
Parameter3_3pm      0.983470  0.989240  0.986351
Parameter4_9am      1.060347  1.064359  1.062351
Parameter4_3pm      1.016996  1.019665  1.018329
Parameter5_9am      0.936823  0.942797  0.939805
Parameter7_9am      0.812337  0.830709  0.821472
año                 0.997726  1.013138  1.005402
mes                 1.052227  1.064323  1.058257
Location_3          0.436831  0.602666  0.513092
Location_4          0.870589  1.341532  1.080705
Location_5          0.399895  0.556805  0.471872
Location_6          0.134147  0.187655  0.158661
```

0.1.5 5. Comente los resultados obtenidos en 2, 3 y 4. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R= Desde la base que el modelo MCO no es adecuado cuando la variable dependiente es binaria, ya que puede generar predicciones fuera del rango 0 y 1, y si bien el modelo probit y logit logran una mejor estimacion del resultado opinaria que en este caso seria mas adecuado el modelo Logit, al es más interpretable (por ejemplo, en términos de odds) y es el más utilizado en análisis de variables binarias. Y las variables robustas podriamos determinar que son Min_Temp, Evaporation y Electricity. dado que las tres cumplen con ser significativas, tener el mismo signo y una magnitud relativamente similar en los 3 modelos.

0.1.6 6. Agregue la data a nivel mensual, usando la data promedio de las variables (ignorando aquellas categoricas, como la direccion del viento). En particular, genere una variable que cuente la cantidad de fallos observados en un mes, utilice un valor de 0 si en ese mes no se reporto fallos en ningun dia. Use un modelo Poisson para explicar el numero de fallas por mes. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Primero que nada se genero una data a nivel mensual y separada por locaciones, que contiene el valor promediado de todas las demas variables exepto por los indicadores NaN de Evaporation y Electricity(1 cuando tienen NaN y 0 cuando no) que fueron generados posteriormente al promediado. Dentro de las variables cuyo coeficiente era de mayor magnitud y a su vez era significativo podemos mencionar la variable Parameter1_Speed dado que ademas de ser significativa aumentaria las fallas en un 4.78%(exp(0.0467)) por unidad y la variable Parameter7_9am que las aumentaria en un 20%(exp(0.1826))

[18]:
```
#filtrar por las fechas de interes(posterior a 2009) y generar columnas de año␣
↪y mes
df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%Y')
df['año'] = df['Date'].dt.year
df['mes'] = df['Date'].dt.month
df_04 = df[df['año'] >= 2009]
#Generar una variable binaria en base a la columna Failure
df_04['Failure_month'] = df_04['Failure_today'].map({'Yes': 1, 'No': 0})

df_05 = df_04.drop(['Date','Parameter6_9am',␣
↪'Parameter6_3pm','Leakage','Failure_today','Parameter1_Dir','Parameter2_9am','Parameter2_3p
↪axis=1)
```

[19]:
```
# Definir cómo queremos agregar
agg_dict = {col: 'mean' for col in df_05.columns if col not in ['año',␣
↪'mes','Location', 'Failure_month']}
agg_dict['Failure_month'] = 'sum'

# Agrupar y aplicar agregaciones
df_mensual = df_05.groupby(['año', 'mes','Location'], as_index=False).
↪agg(agg_dict)

df_mensual['Electricity_bin'] = df_mensual['Electricity'].isna().astype(int)
df_mensual.Electricity=df_mensual.Electricity.fillna(0)
df_mensual['Evaporation_bin'] = df_mensual['Evaporation'].isna().astype(int)
df_mensual.Evaporation=df_mensual.Evaporation.fillna(0)

df_mensual=df_mensual.dropna()

df_mensual
```

[19]:
```
      año  mes  Location   Min_Temp   Max_Temp  Evaporation  Electricity  \
0    2009    1         1  17.932258  32.003226    13.419048    12.180000
```

```
2     2009    1        3  16.312903  34.658065    0.000000    0.000000
3     2009    1        4  22.422581  36.058065   13.561290   10.525806
4     2009    1        5  16.154839  32.780645    0.000000    0.000000
5     2009    1        6  10.467742  28.529032    0.000000    0.000000
…     …  …      …       …          …          …           …
4687  2017    6       45   4.424000  14.744000    1.344000    4.632000
4688  2017    6       46  10.100000  18.356000    0.000000    0.000000
4689  2017    6       47   8.736000  18.616000    0.000000    0.000000
4690  2017    6       48  11.657895  17.700000    0.000000    0.000000
4691  2017    6       49   5.800000  18.754167    2.977273    0.000000

      Parameter1_Speed  Parameter3_9am  Parameter3_3pm  Parameter4_9am  \
0            39.645161       10.161290       17.966667       37.612903
2            42.677419       11.935484       18.548387       41.903226
3            51.258065       18.516129       25.032258       37.096774
4            41.935484        7.419355       17.466667       65.516129
5            48.000000       20.500000       21.806452       50.354839
…                  …               …               …              …
4687         24.040000        4.960000        9.280000       97.840000
4688         34.120000       16.440000       16.440000       87.200000
4689         34.000000        9.520000       16.320000       88.520000
4690         38.894737       15.052632       19.842105       73.315789
4691         27.666667       11.375000       12.833333       66.041667

      Parameter4_3pm  Parameter5_9am  Parameter5_3pm  Parameter7_9am  \
0          23.827586     1014.025806     1012.166667       23.658065
2          17.870968     1013.064516     1009.770968       22.993548
3          24.516129     1008.461290     1004.732258       29.241935
4          35.933333     1015.451613     1012.353333       22.390323
5          24.225806     1012.873333     1011.496667       18.577419
…                …               …               …              …
4687       67.760000     1028.816000     1026.476000        6.736000
4688       70.880000     1025.720000     1023.492000       13.168000
4689       67.280000     1024.156000     1022.168000       12.948000
4690       69.421053     1026.163158     1024.126316       14.726316
4691       35.875000     1029.704167     1027.033333       10.495833

      Parameter7_3pm  Failure_month  Electricity_bin  Evaporation_bin
0          30.750000            0.0                0                0
2          32.964516            1.0                1                1
3          34.487097            3.0                0                0
4          31.156667            3.0                1                1
5          26.593548            0.0                1                1
…                …              …                …                …
4687       13.696000            3.0                0                0
4688       17.304000           13.0                1                1
4689       17.360000            9.0                1                1
```

```
4690      16.757895        4.0               1                1
4691      18.070833        0.0               0                0

[4076 rows x 19 columns]
```

[20]:
```python
y = df_mensual['Failure_month']
X2=df_mensual.drop(['Failure_month','año', 'mes'], axis=1)
X2=sm.add_constant(X2)
poisson=sm.GLM(y,X2,family=sm.families.Poisson()).fit()
print(poisson.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Failure_month   No. Observations:                 4076
Model:                            GLM   Df Residuals:                     4059
Model Family:                 Poisson   Df Model:                           16
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -9367.4
Date:                Thu, 24 Apr 2025   Deviance:                       4925.1
Time:                        23:58:56   Pearson chi2:                 4.55e+03
No. Iterations:                     5   Pseudo R-squ. (CS):             0.8668
Covariance Type:            nonrobust
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const               22.0480      2.467      8.936      0.000      17.212
26.884
Location            -0.0022      0.000     -4.899      0.000      -0.003
-0.001
Min_Temp            -0.0139      0.007     -2.009      0.044      -0.027
-0.000
Max_Temp            -0.0817      0.021     -3.961      0.000      -0.122
-0.041
Evaporation         -0.0069      0.005     -1.538      0.124      -0.016
0.002
Electricity         -0.0501      0.006     -7.861      0.000      -0.063
-0.038
Parameter1_Speed     0.0467      0.002     20.519      0.000       0.042
0.051
Parameter3_9am      -0.0055      0.003     -2.079      0.038      -0.011
-0.000
Parameter3_3pm      -0.0570      0.003    -19.372      0.000      -0.063
-0.051
Parameter4_9am       0.0343      0.002     17.833      0.000       0.030
0.038
```

```
Parameter4_3pm      -0.0031      0.002     -1.317      0.188      -0.008
0.001
Parameter5_9am      -0.0506      0.012     -4.205      0.000      -0.074
-0.027
Parameter5_3pm       0.0290      0.012      2.397      0.017       0.005
0.053
Parameter7_9am       0.1826      0.011     16.142      0.000       0.160
0.205
Parameter7_3pm      -0.0783      0.023     -3.380      0.001      -0.124
-0.033
Electricity_bin     -0.4149      0.051     -8.095      0.000      -0.515
-0.314
Evaporation_bin     -0.0385      0.033     -1.157      0.247      -0.104
0.027
================================================================================
====
```

0.1.7 7. Determine sobre dispersion en la data y posible valor optimo de alpha para un modelo Binomial Negativa.

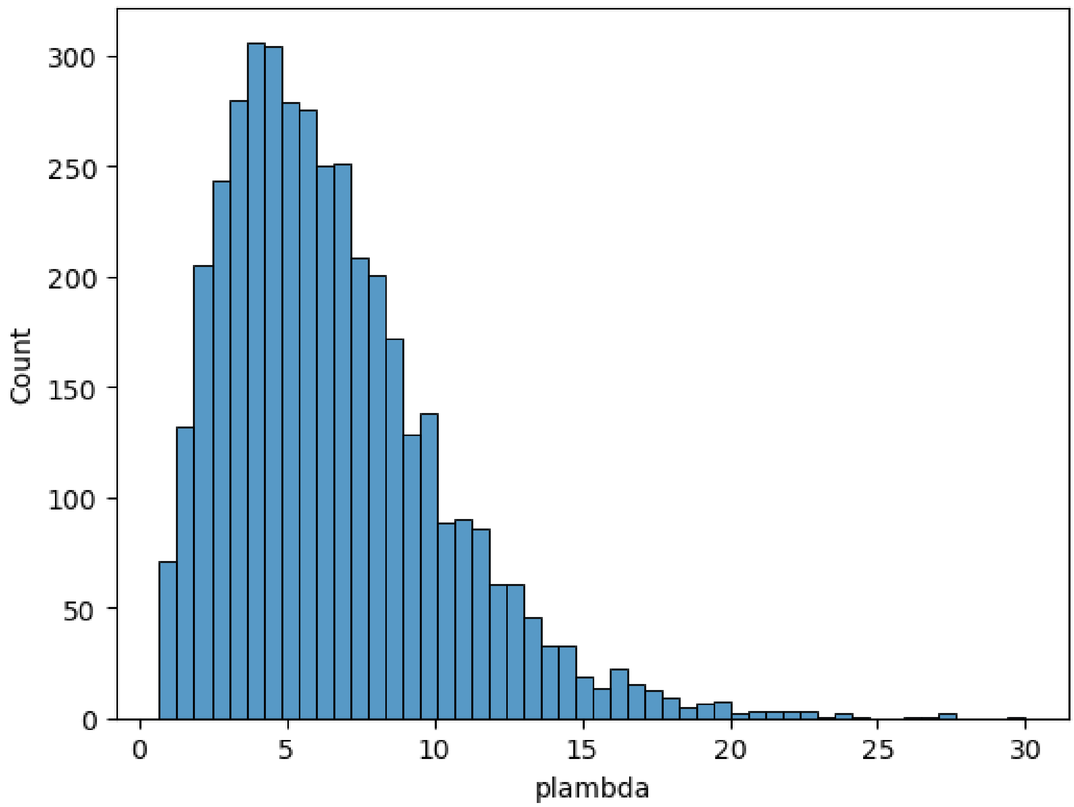
R: Según el análisis, y dado que el valor de alpha es mayor a 1 y es significativo, podemos concluir que el modelo presenta sobre-dispersión. Además, el valor p es igual a 0.000, lo que refuerza la evidencia de que existe una sobre-dispersión importante en los datos.

[21]:
```python
print(df_mensual['Failure_month'].describe())
```

```
count    4076.000000
mean        6.547596
std         4.482926
min         0.000000
25%         3.000000
50%         6.000000
75%         9.000000
max        25.000000
Name: Failure_month, dtype: float64
```

[22]:
```python
df_mensual['plambda'] = poisson.mu
sns.histplot(data=df_mensual, x="plambda")
```

[22]: `<Axes: xlabel='plambda', ylabel='Count'>`



[23]:
```
aux=((y-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:          Failure_month   R-squared (uncentered):
0.000
Model:                            OLS   Adj. R-squared (uncentered):
0.000
Method:                 Least Squares   F-statistic:
1.698
Date:                Thu, 24 Apr 2025   Prob (F-statistic):
0.193
Time:                        23:58:57   Log-Likelihood:
-10575.
No. Observations:                4076   AIC:
2.115e+04
Df Residuals:                    4075   BIC:
2.116e+04
```

```
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0088      0.007      1.303      0.193      -0.004       0.022
==============================================================================
Omnibus:                    12416.645   Durbin-Watson:                   1.961
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       1073304040.529
Skew:                          44.800   Prob(JB):                         0.00
Kurtosis:                    2515.314   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

[24]:
```python
model_nb = smf.glm(formula = "Failure_month ~ Location + Min_Temp + Max_Temp +
 ↪Evaporation + Electricity + Parameter1_Speed + Parameter3_9am +
 ↪Parameter3_3pm + Parameter4_9am + Parameter4_3pm + Parameter5_9am +
 ↪Parameter5_3pm + Parameter7_9am + Parameter7_3pm + Electricity_bin +
 ↪Evaporation_bin", data=df_mensual, family=sm.families.NegativeBinomial()).
 ↪fit()
print(model_nb.summary())

alpha = np.exp(auxr.params[0])
print(f"Alpha (sobre-dispersión): {alpha}")
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Failure_month   No. Observations:                 4076
Model:                            GLM   Df Residuals:                     4059
Model Family:        NegativeBinomial   Df Model:                           16
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -11410.
Date:                Thu, 24 Apr 2025   Deviance:                       1122.2
Time:                        23:58:57   Pearson chi2:                     848.
No. Iterations:                     9   Pseudo R-squ. (CS):             0.2628
Covariance Type:            nonrobust
================================================================================
====
                       coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
Intercept           24.8161      7.263      3.417      0.001      10.580
```

```
39.052
Location              -0.0024      0.001     -1.944      0.052      -0.005
2.02e-05
Min_Temp              -0.0008      0.018     -0.044      0.965      -0.035
0.034
Max_Temp              -0.0359      0.056     -0.639      0.523      -0.146
0.074
Evaporation           -0.0014      0.010     -0.130      0.897      -0.022
0.019
Electricity           -0.0787      0.017     -4.731      0.000      -0.111
-0.046
Parameter1_Speed       0.0533      0.007      8.169      0.000       0.041
0.066
Parameter3_9am        -0.0028      0.007     -0.397      0.691      -0.016
0.011
Parameter3_3pm        -0.0713      0.008     -8.808      0.000      -0.087
-0.055
Parameter4_9am         0.0414      0.005      8.076      0.000       0.031
0.051
Parameter4_3pm        -0.0142      0.007     -2.165      0.030      -0.027
-0.001
Parameter5_9am        -0.0894      0.033     -2.723      0.006      -0.154
-0.025
Parameter5_3pm         0.0656      0.033      1.980      0.048       0.001
0.131
Parameter7_9am         0.2186      0.030      7.175      0.000       0.159
0.278
Parameter7_3pm        -0.1752      0.063     -2.791      0.005      -0.298
-0.052
Electricity_bin       -0.6144      0.143     -4.288      0.000      -0.895
-0.334
Evaporation_bin       -0.0062      0.087     -0.071      0.943      -0.177
0.164
================================================================================
====
Alpha (sobre-dispersión): 1.008796732767034
```

0.1.8 8. Usando la informacion anterior, ejecute un modelo Binomial Negativa para responder a la pregunta 6. Seleccione las variables dependientes a incluir en el modelo final e interprete su significado.

R: Ejecutamos un modelo de binomial Negativa con el aplha estimado anteriormente y podemos mencionar que Electricity, Parameter1_Speed, Parameter3_3pm y Parameter7_9am son clave en la predicción del número de fallas mensuales. dado que todas ellas son significativas y tienen valores de coeficiente de una magnitud razonable.

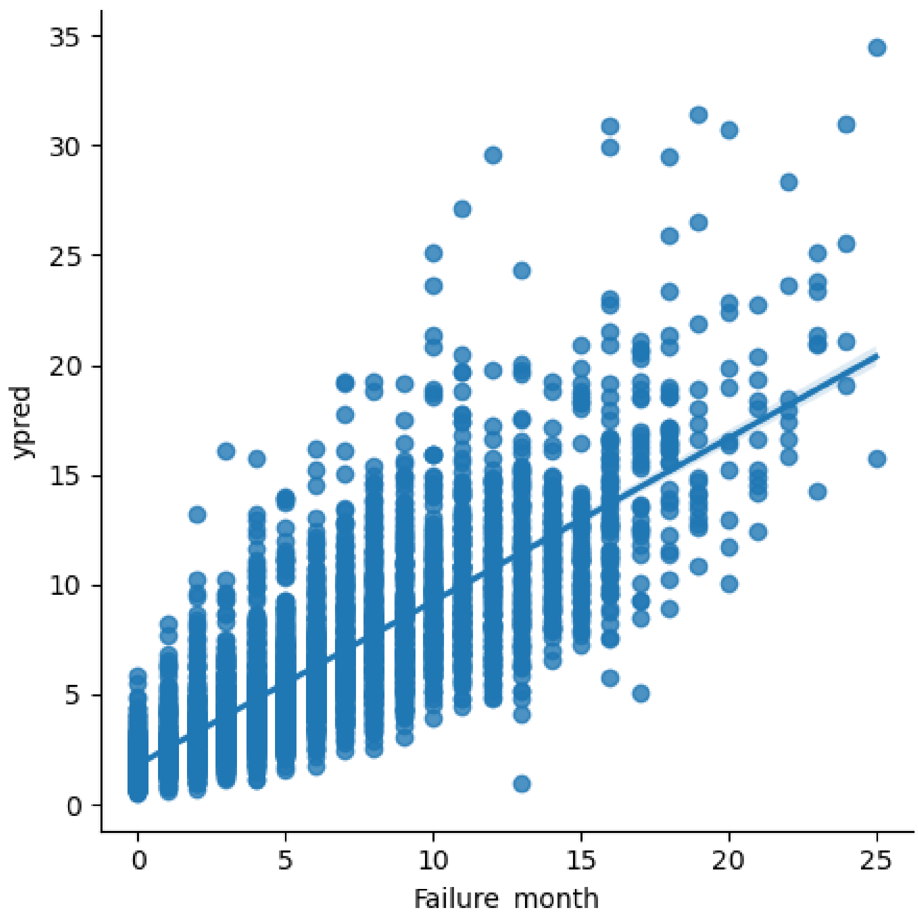
[25]:
```python
negbin=sm.GLM(y,X2,family=sm.families.NegativeBinomial(alpha=1.0088)).fit()
print(negbin.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Failure_month   No. Observations:                 4076
Model:                            GLM   Df Residuals:                     4059
Model Family:        NegativeBinomial   Df Model:                           16
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -11424.
Date:                Thu, 24 Apr 2025   Deviance:                       1116.1
Time:                        23:58:57   Pearson chi2:                     843.
No. Iterations:                     9   Pseudo R-squ. (CS):             0.2612
Covariance Type:            nonrobust
================================================================================
====
                      coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
----
const              24.8281      7.291      3.406      0.001      10.539
39.117
Location           -0.0024      0.001     -1.936      0.053      -0.005
2.98e-05
Min_Temp           -0.0007      0.018     -0.042      0.967      -0.035
0.034
Max_Temp           -0.0358      0.056     -0.633      0.526      -0.146
0.075
Evaporation        -0.0013      0.010     -0.128      0.898      -0.022
0.019
Electricity        -0.0788      0.017     -4.718      0.000      -0.111
-0.046
Parameter1_Speed    0.0533      0.007      8.139      0.000       0.040
0.066
Parameter3_9am     -0.0028      0.007     -0.396      0.692      -0.017
0.011
Parameter3_3pm     -0.0714      0.008     -8.779      0.000      -0.087
-0.055
Parameter4_9am      0.0414      0.005      8.052      0.000       0.031
0.051
Parameter4_3pm     -0.0142      0.007     -2.163      0.031      -0.027
-0.001
Parameter5_9am     -0.0895      0.033     -2.716      0.007      -0.154
-0.025
Parameter5_3pm      0.0657      0.033      1.976      0.048       0.001
0.131
Parameter7_9am      0.2187      0.031      7.154      0.000       0.159
0.279
Parameter7_3pm     -0.1756      0.063     -2.786      0.005      -0.299
-0.052
Electricity_bin    -0.6148      0.144     -4.275      0.000      -0.897
```

```
-0.333
Evaporation_bin     -0.0061      0.087      -0.070      0.944      -0.177
0.165
================================================================================
====
```

[26]:
```
df_mensual['ypred'] = negbin.predict(X2)
sns.lmplot(data=df_mensual, x='Failure_month', y='ypred')
```

[26]: <seaborn.axisgrid.FacetGrid at 0x1a907492050>



0.1.9 9. Comente los resultados obtenidos en 6, 7 y 8. ¿Cuáles y por qué existen las diferencias entre los resultados?. En su opinión, ¿Cuál sería el más adecuado para responder la pregunta de investgación y por qué? ¿Qué variables resultaron ser robustas a la especificación?

R: existen diferencias entre los modelosprincipalmente por que el modelo de Poisson subestima la variabilidad de los datos mientras que en binomial negativa esta puede adaptarse mejor a los datos llegando a estimaciones más precisas y confiables. Considero que seria mejor utilizar binomial negativa dado que Poisson no captura adecuadamente la sobre-dispersión presente en los datos y el modelo Binomial Negativa ajusta la sobre-dispersión y finalmente las variables robustas son Electricity, Parameter1_Speed y Parameter7_9am.