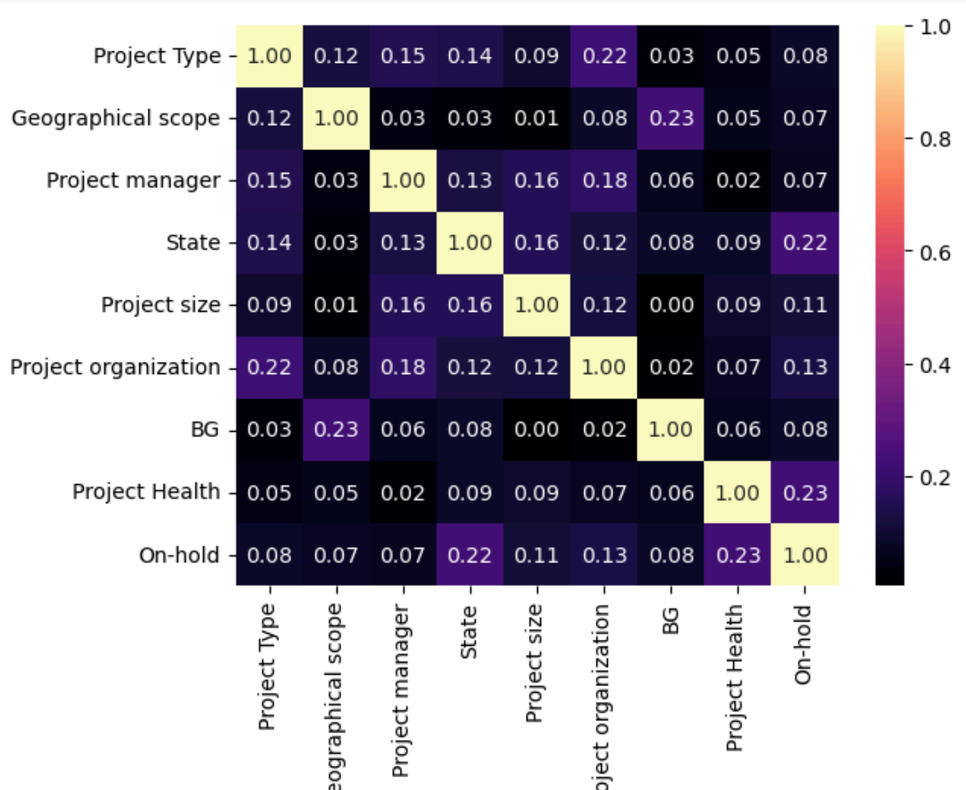
MODELO DE REGRESIÓN LINEAL

PAULA SIMONETTA MADRID PEREZ
ANIA DIAZ PEYROT
IVANNA MALDONADO MIRANDA

FORVIA

Absolutos de la Matriz de correlación:

	Project Type	Geographical scope	Project manager	State	Project size	Project organization	BG	Project Health	On-hold
Project Type	1.000000	0.116070	0.150064	0.141691	0.092805	0.217320	0.025657	0.045141	0.080656
Geographical scope	0.116070	1.000000	0.034761	0.027901	0.009645	0.082748	0.227722	0.054766	0.071808
Project manager	0.150064	0.034761	1.000000	0.128892	0.156898	0.181224	0.061832	0.021136	0.070125
State	0.141691	0.027901	0.128892	1.000000	0.157928	0.120862	0.084619	0.085652	0.221179
Project size	0.092805	0.009645	0.156898	0.157928	1.000000	0.115164	0.004848	0.086617	0.111695
Project organization	0.217320	0.082748	0.181224	0.120862	0.115164	1.000000	0.015659	0.072307	0.129562
BG	0.025657	0.227722	0.061832	0.084619	0.004848	0.015659	1.000000	0.060117	0.082848
Project Health	0.045141	0.054766	0.021136	0.085652	0.086617	0.072307	0.060117	1.000000	0.227482
On-hold	0.080656	0.071808	0.070125	0.221179	0.111695	0.129562	0.082848	0.227482	1.000000



	Par de Variables	Correlación
0	Geographical scope y BG	0.23
1	On-hold y Project Health	0.23
2	Project Type y Project organization	0.22
3	State y On-hold	0.22
4	Project manager y Project size	0.16

pares

0-'Presentan una correlación positiva débil, lo que indica que a mayor alcance geográfico del proyecto podría haber una ligera tendencia a cambios o variaciones en el indicador BG.',

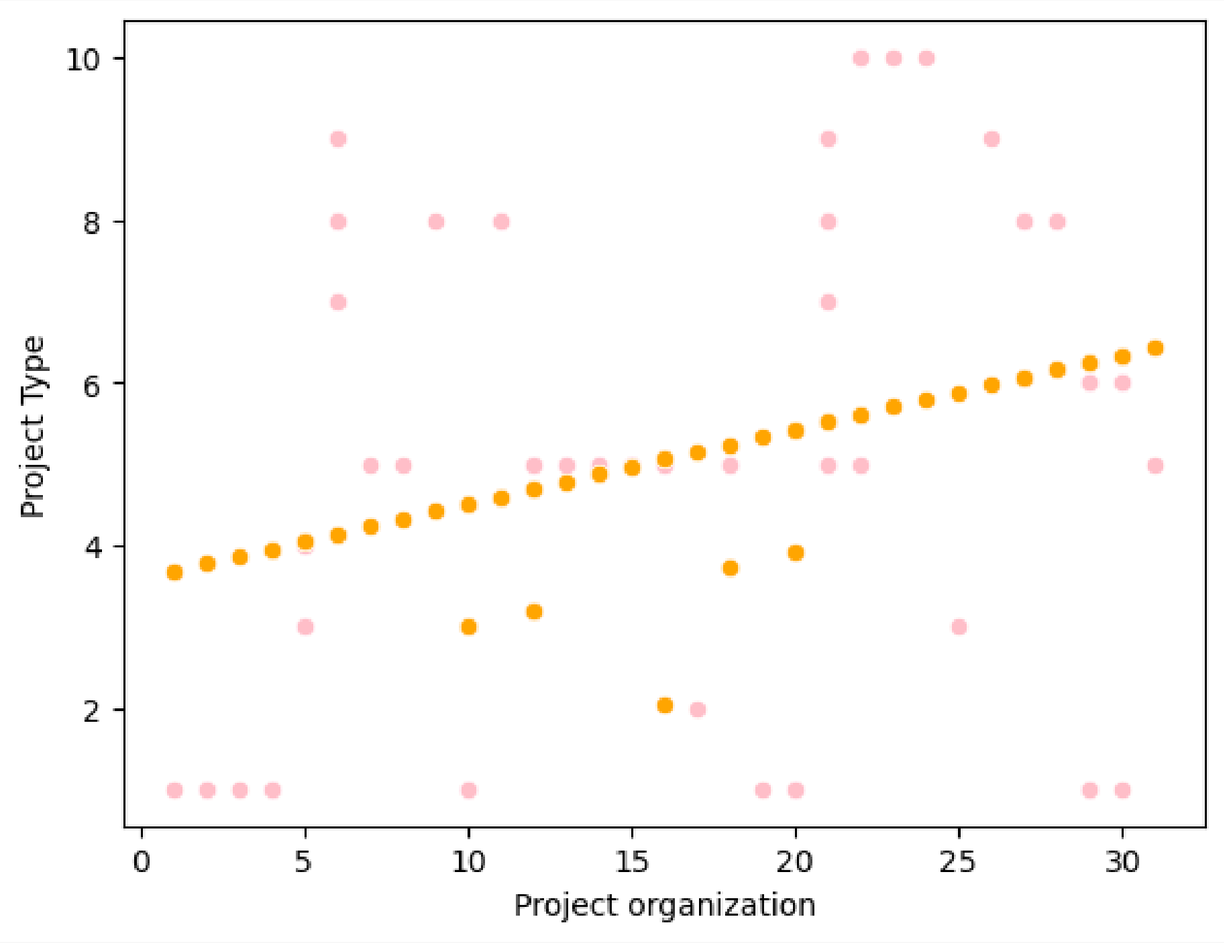
1- 'Presentan una correlación positiva débil, lo que sugiere que los proyectos en espera tienden a tener una salud del proyecto similar.',

2- 'Presentan una correlación positiva débil, lo que indica que el tipo de proyecto podría estar relacionado con la organización del proyecto.',

3-'Presentan una correlación positiva débil, lo que sugiere que el estado del proyecto podría estar relacionado con su estado de espera.',

4- Presentan una correlación positiva débil, lo que indica que el gerente del proyecto podría estar relacionado con el tamaño del proyecto.'

PROYECT TYPE

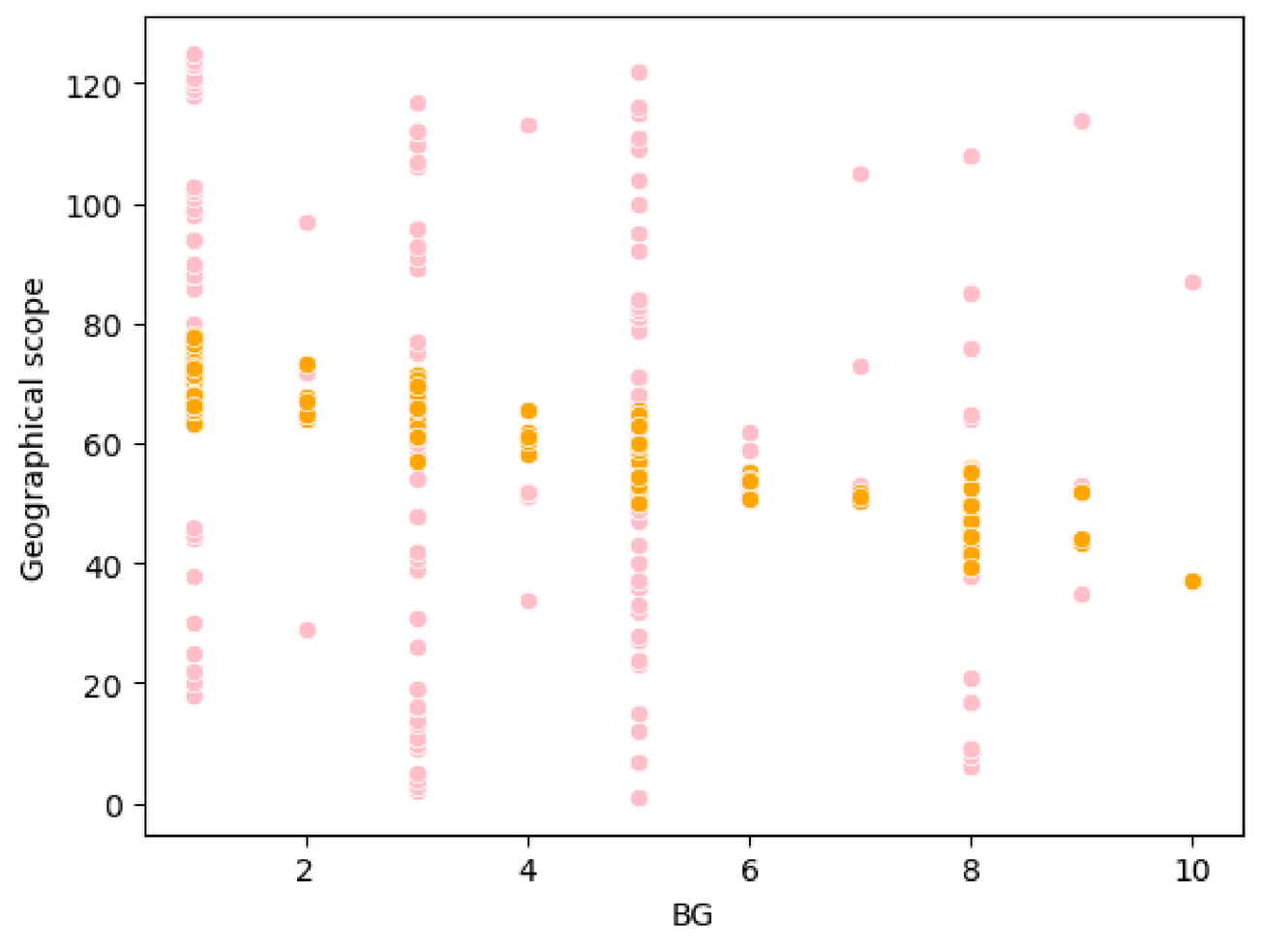


CORRELACIÓN: 0.22

```
sns.scatterplot(x='Project organization', y='Project Type', color="pink", data=df_num)
sns.scatterplot(x='Project organization', y='PrediccionesTname0', color="orange", data=df_num)
```

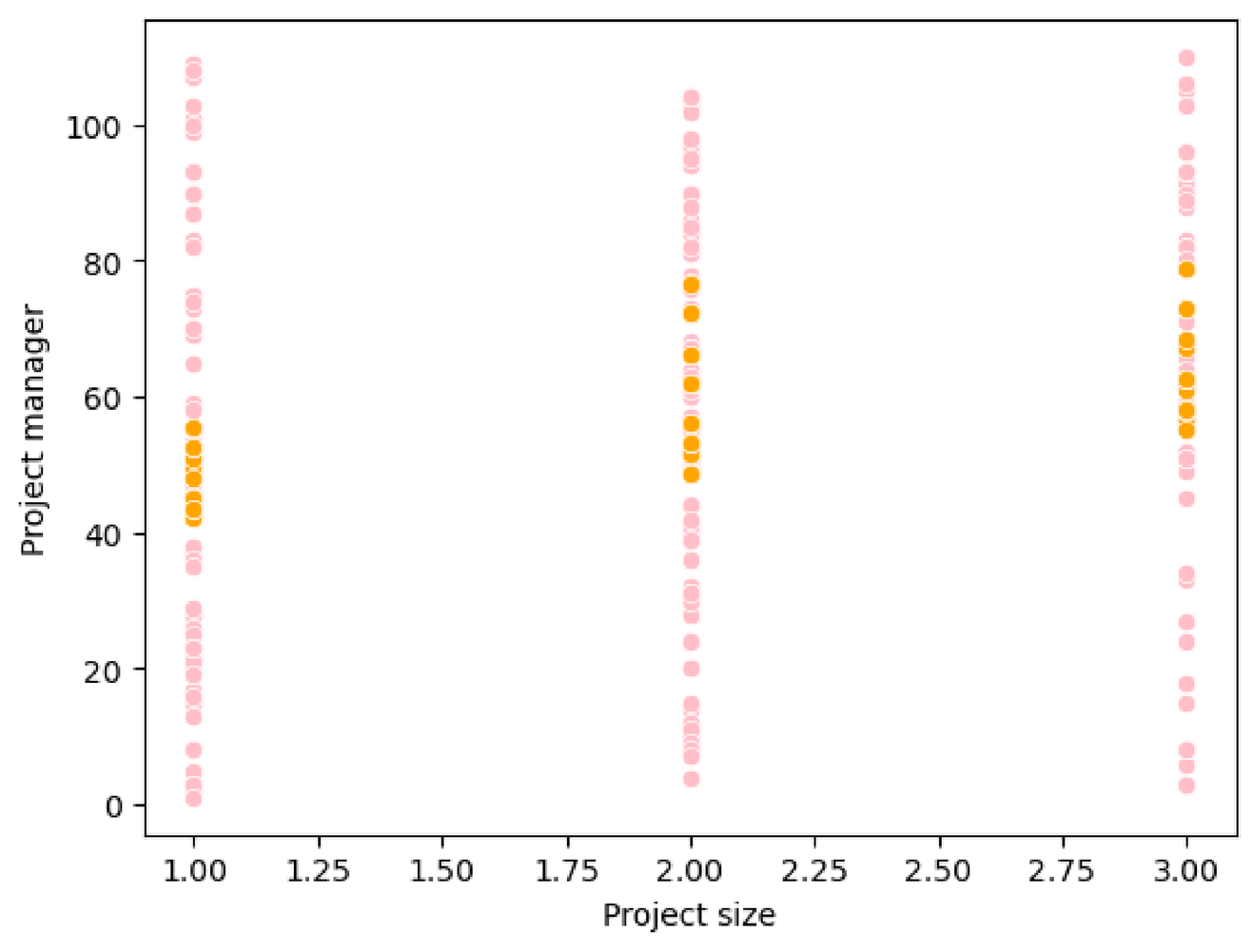
GEOGRAPHICAL SCOPE

```
sns.scatterplot(x='BG', y='Geographical scope', color="pink", data=df_num)
sns.scatterplot(x='BG', y='PrediccionesGeographical scope', color="orange", data=df_num)
```



CORRELACIÓN: 0.23

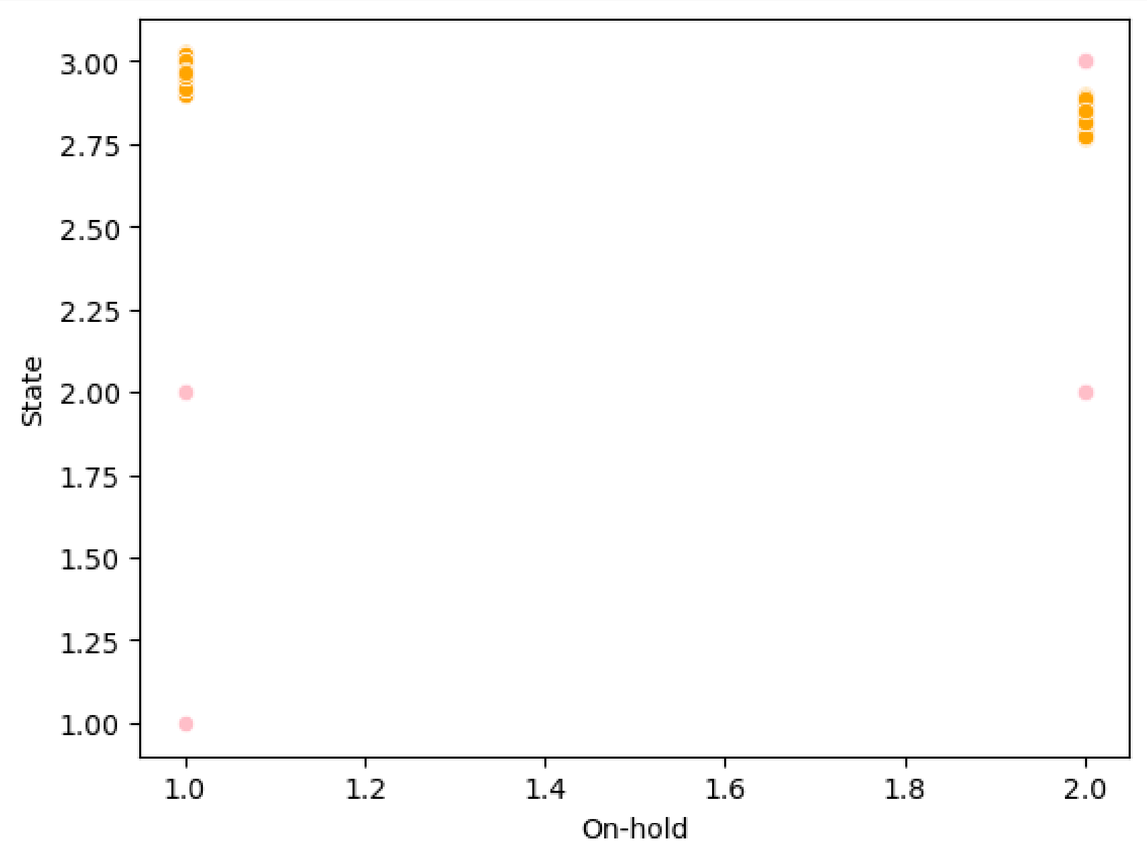
PROJECT MANAGER



CORRELACIÓN: 0.16

```
sns.scatterplot(x='Project size', y='Project manager', color="pink", data=df_num)
sns.scatterplot(x='Project size', y='PrediccionesProject manager', color="orange", data=df_num)
```

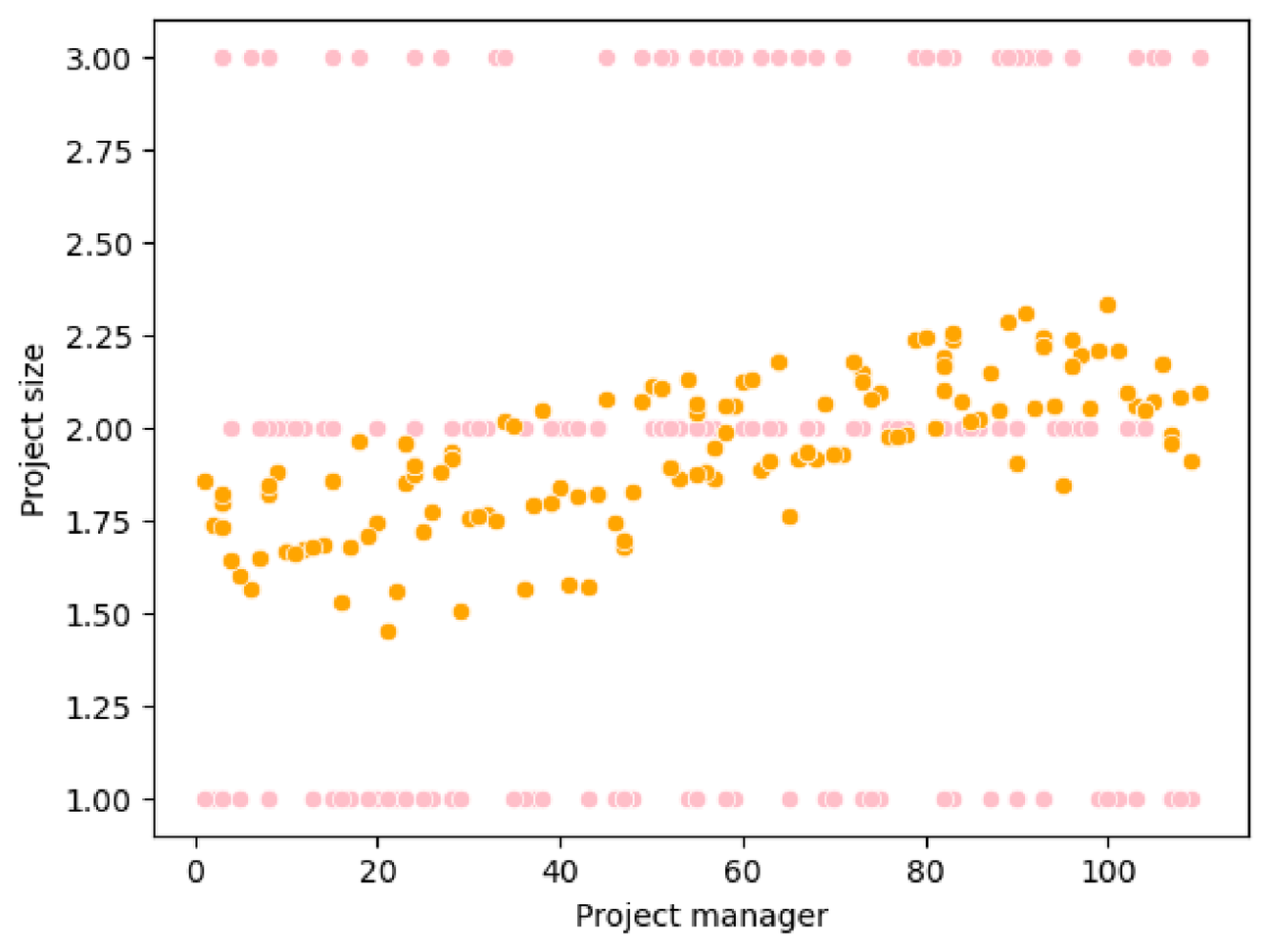
STATE



CORRELACIÓN: 0.22

```
sns.scatterplot(x='On-hold', y='State', color="pink", data=df_num)
sns.scatterplot(x='On-hold', y='PrediccionesState', color="orange", data=df_num)
```

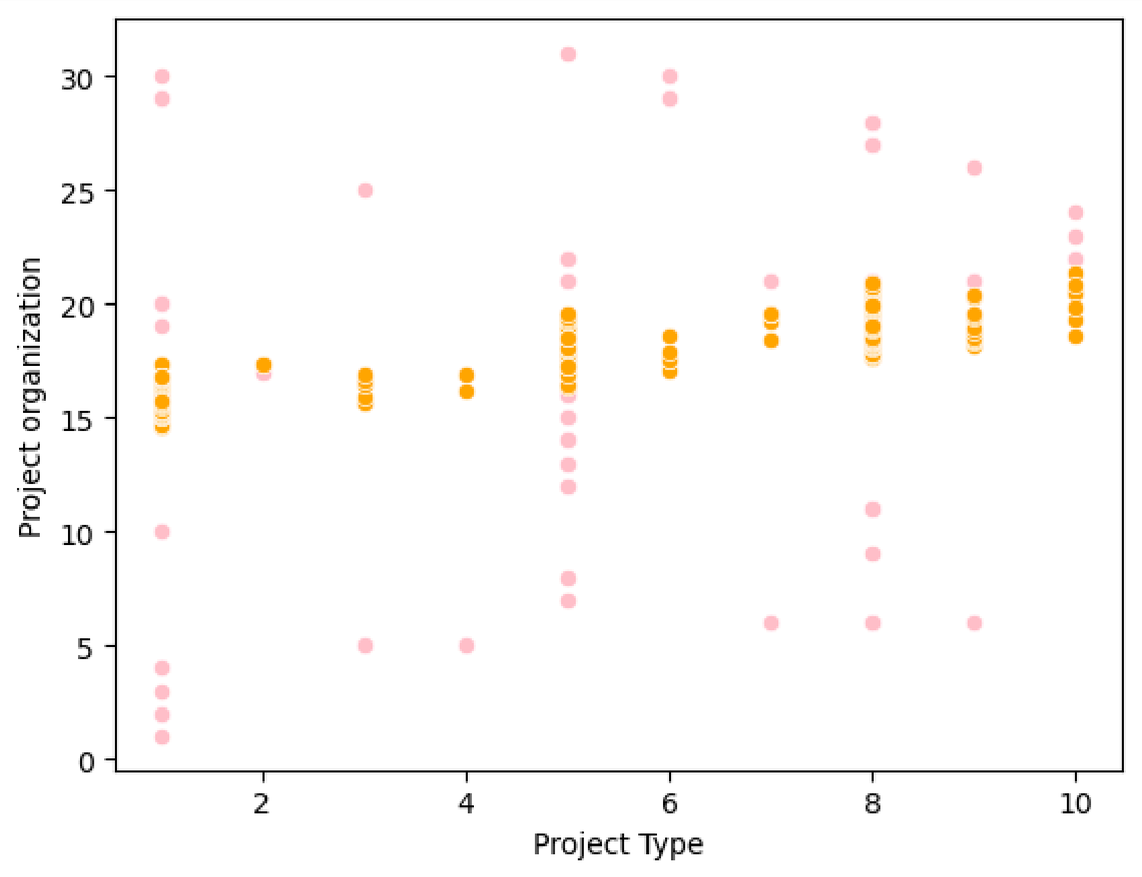
PROJECT SIZE



CORRELACIÓN: 0.16

```
sns.scatterplot(x='Project manager', y='Project size', color="pink", data=df_num)
sns.scatterplot(x='Project manager', y='PrediccionesProjectSize', color="orange", data=df_num)
```

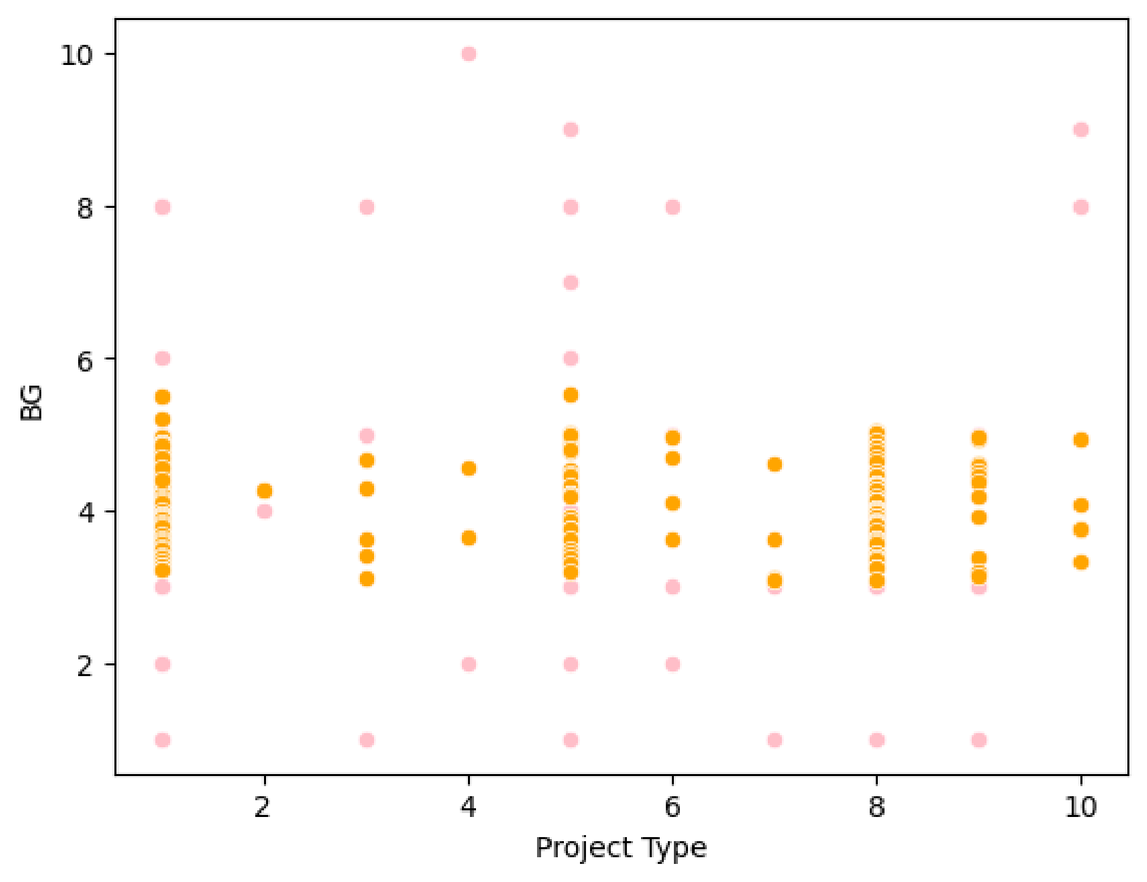
PROJECT ORGANIZATION



CORRELACIÓN: 0.22

```
sns.scatterplot(x='Project Type', y='Project organization', color="pink", data=df_num)
sns.scatterplot(x='Project Type', y='PrediccionesProjectorg', color="orange", data=df_num)
```

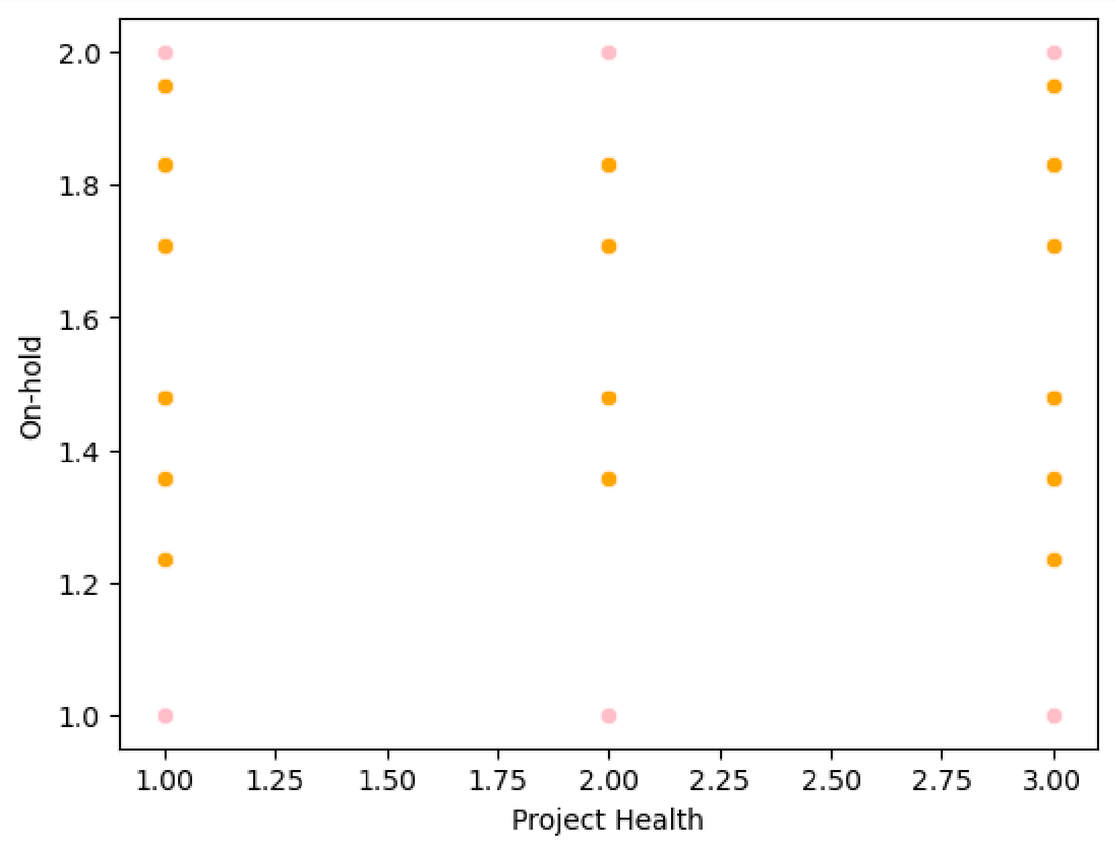
BG



CORRELACIÓN: 0.03

```
sns.scatterplot(x='Project Type', y='BG', color="pink", data=df_num)
sns.scatterplot(x='Project Type', y='PrediccionesProjectBG', color="orange", data=df_num)
```

ON-HOLD



CORRELACIÓN: 0.23

```
sns.scatterplot(x='Project Health', y='On-hold', color="pink", data=df_num)
sns.scatterplot(x='Project Health', y='PrediccionesProjectH', color="orange", data=df_num)
```

REGRESIÓN LINEAL

EXAMINAR LAS FRECUENCIAS DE LAS VARIABLES CATEGÓRICAS "TAXONNAME", "TAXONCODE", "SAMPLINGOPERATIONS_CODE", "CODESITE_SAMPLINGOPERATIONS" Y "DATE_SAMPLINGOPERATION"

```python
#Frecuencias de mayor a menor para cada una
print("\nFrecuencias de TaxonName:")
print(df['TaxonName'].value_counts())
print()

print("\nFrecuencias de TaxonCode:")
print(df['TaxonCode'].value_counts())
print()

print("\nFrecuencias de SamplingOperations_code:")
print(df['SamplingOperations_code'].value_counts())
print()

print("\nFrecuencias de CodeSite_SamplingOperations:")
print(df['CodeSite_SamplingOperations'].value_counts())
print()

print("\nFrecuencias de Date_SamplingOperation:")
print(df['Date_SamplingOperation'].value_counts())
print()
```

✓ 0.2s Python

CONVERTIRLAS A NÚMERICAS SEGÚN SU FRECUENCIA

```python
df_numeric = df.copy()

variables_categoricas = ['TaxonName', 'TaxonCode', 'SamplingOperations_code',
                         'CodeSite_SamplingOperations', 'Date_SamplingOperation']

for variable in variables_categoricas:
    frecuencias = df[variable].value_counts() #Contamos las frecuencias

    mapeo = {}
    for i, valor in enumerate(frecuencias.index): #Mapeamos el valor de la más frecuente a 1, y asi sucesivamente
        mapeo[valor] = i + 1
    df_numeric[variable] = df[variable].map(mapeo) #Aplicamos el mapeo a la columna
```

✓ 0.4s Python

CONVERTIRLAS A NÚMERICAS SEGÚN SU FRECUENCIA

```
print("\nTipos de datos en el dataframe convertido:")
print(df_numeric.dtypes)
```

✓ 0.0s Python

```
Tipos de datos en el dataframe convertido:
TaxonName                           int64
TaxonCode                           int64
SamplingOperations_code             int64
CodeSite_SamplingOperations         int64
Date_SamplingOperation              int64
Abundance_nbcell                    int64
TotalAbundance_SamplingOperation    int64
Abundance_pm                      float64
dtype: object
```

VERIFICAMOS LOS TIPOS DEL DATOS DEL DF CONVERTIDO

```
Tipos de datos en el dataframe convertido:
TaxonName                              int64
TaxonCode                              int64
SamplingOperations_code                int64
CodeSite_SamplingOperations            int64
Date_SamplingOperation                 int64
Abundance_nbcell                       int64
TotalAbundance_SamplingOperation       int64
Abundance_pm                         float64
dtype: object
```

```
print("\nTipos de datos en el dataframe convertido:")
print(df_numeric.dtypes)
```

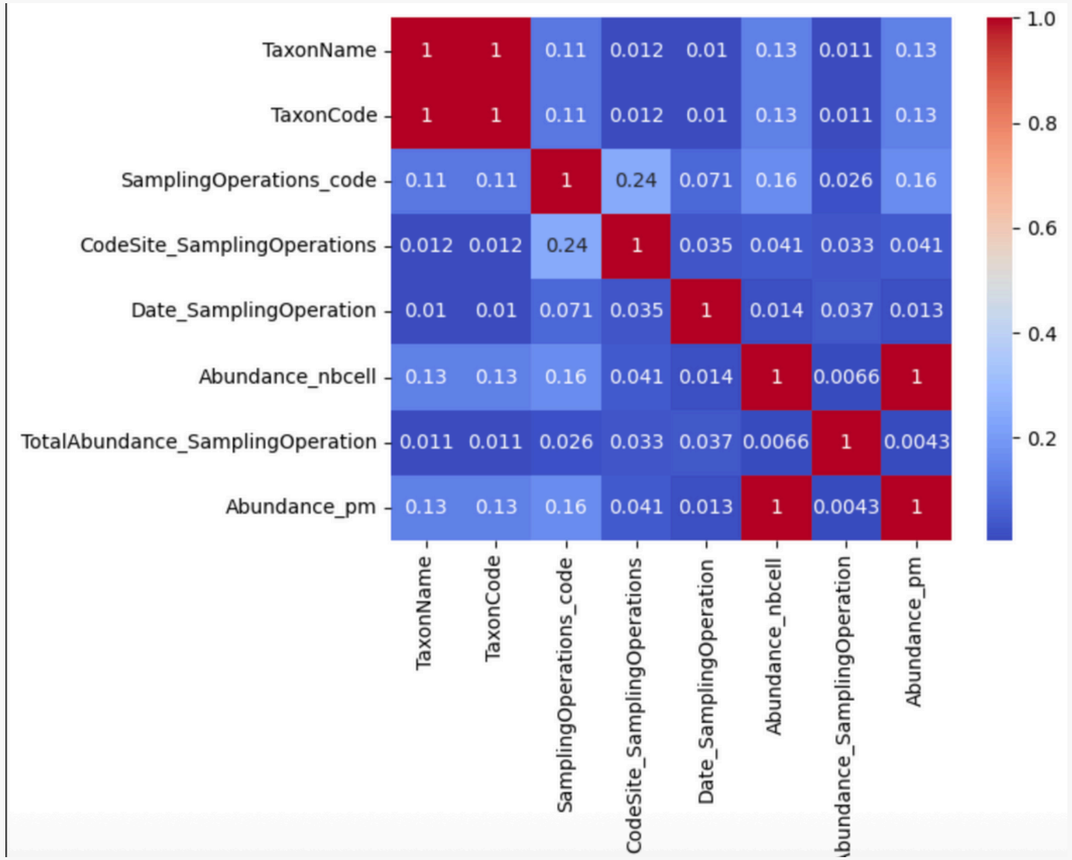
✓ 0.0s Python

CALCULAMOS LAS CORRELACIONES ENTRE TODAS LAS VARIABLES PARA IDENTIFICAR LOS 5 PARES CON MAYOR CORRELACIÓN

```
print("\nMatriz de correlación:")
Corr_Factors = df_numeric.corr()
Corr_Factors
```

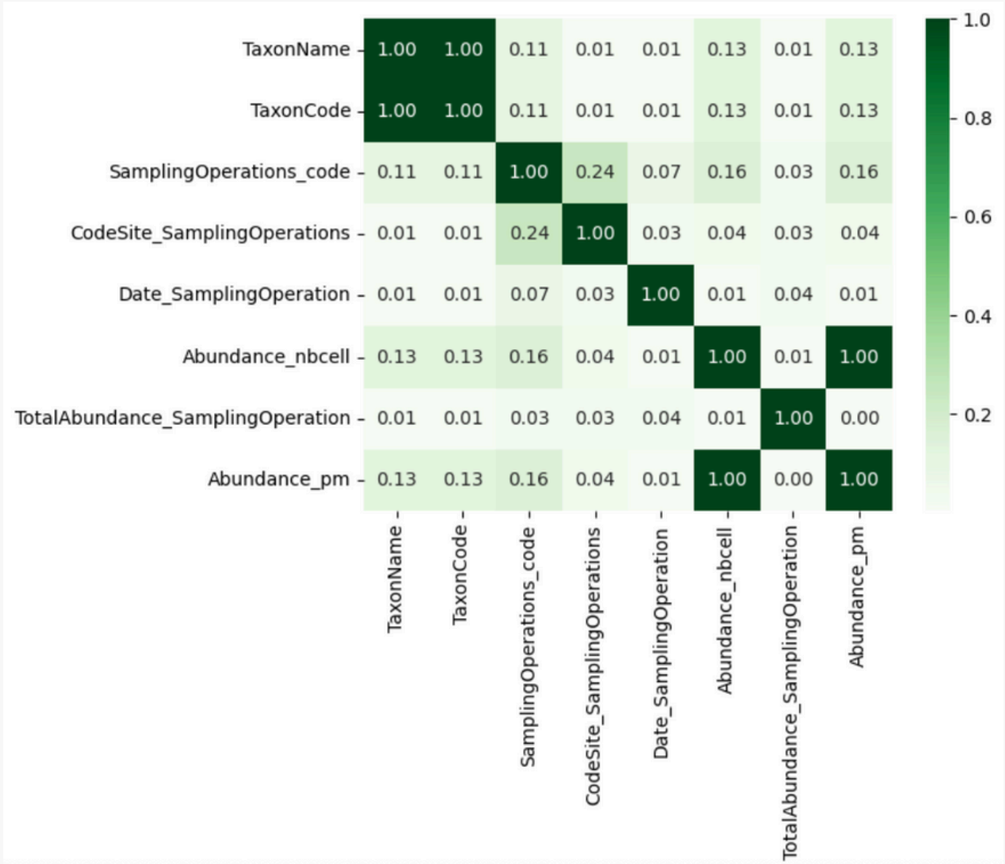
✓ 0.1s Python

GRAFICAMOS EL MAPA DE CALOR DE LOS COEFICIENTES DE CORRELACIÓN




```
Heat_Map= sns.heatmap(Corr_Factors1, cmap = 'coolwarm', annot=True)
Heat_Map
```

AJUSTAMOS EL MAPA DE CALOR DE LOS COEFICIENTES DE CORRELACIÓN




```
Heat_Map= sns.heatmap(Corr_Factors1, cmap = 'Greens', annot=True, fmt=".2f")
Heat_Map
```

TABLA DE LOS 5 PARES

```python
top5 = pd.DataFrame({
    'Par de Variables': [
        'TaxonName y TaxonCode',
        'Abundance_nbcell y Abundance_pm',
        'TotalAbundance_SamplingOperation y Abundance_pm',
        'TotalAbundance_SamplingOperation y Abundance_nbcell',
        'SamplingOperations_code y CodeSite_SamplingOperations'
    ],
    'Correlación': [1.00, 1.00, 1.00, 1.00, 0.24],
    'Interpretación': [
        'Correlación perfecta positiva, Ambas variables representan la misma información (especies) en diferente formato.',
        'Correlación perfecta positiva, Ambas miden abundancia de las mismas especies, solo en diferentes unidades.',
        'Correlación perfecta positiva, La abundancia total está directamente relacionada con la abundancia por muestra.',
        'Correlación perfecta positiva, Misma relación que el anterior, pero con la otra medida de abundancia.',
        'Correlación baja, Muestra una relación débil entre código de operación y sitio de muestreo.'
    ]
})
top5
```

✓ 0.0s

	Par de Variables	Correlación	Interpretación
0	TaxonName y TaxonCode	1.00	Correlación perfecta positiva, Ambas variables...
1	Abundance_nbcell y Abundance_pm	1.00	Correlación perfecta positiva, Ambas miden abu...
2	TotalAbundance_SamplingOperation y Abundance_pm	1.00	Correlación perfecta positiva, La abundancia t...
3	TotalAbundance_SamplingOperation y Abundance_n...	1.00	Correlación perfecta positiva, Misma relación ...
4	SamplingOperations_code y CodeSite_SamplingOpe...	0.24	Correlación baja, Muestra una relación débil e...

REGRESIÓN LINEAL MÚLTIPLE

```
df_numeric.insert(0, 'PrediccionesTotalAb0', y_pred)
df_numeric
```

✓ 0.0s

	PrediccionesTotalAb0	PrediccionesAbnbcellO	PrediccionesDateSamplingO	PrediccionesCodeSiteO	PrediccionesSamplingO	PrediccionesTcode0	PrediccionesTnam
0	405.592225	-19.493875	505.088381	2641.856823	19501.088929	79.550099	79.5500
1	405.322372	-20.717352	489.297200	2494.596972	16268.736808	87.732102	87.7321
2	405.629760	3.275841	477.033281	2266.278194	17118.414345	94.423476	94.4234
3	407.384970	-2.294928	460.721199	1649.721219	26237.734154	111.896494	111.8964
4	405.832620	1.039577	461.347628	2018.917831	17082.195005	101.691599	101.6915
...	...	...	...	...	...	...	
1643867	406.344626	-21.174638	434.586434	1559.520791	17987.962689	115.240699	115.2406
1643868	406.197498	-16.777252	470.007813	2044.797937	19907.303535	100.931171	100.9311
1643869	405.191909	-11.782861	494.995009	2597.532997	15948.640380	84.703279	84.7032
1643870	406.068389	-21.214419	429.465158	1551.851351	16410.119135	111.556630	111.5566
1643871	406.402373	-20.587063	440.726613	1623.382449	18817.273870	113.313509	113.3135

1643872 rows × 15 columns

PrediccionesAbpm0	Pre
-48.011670	
-51.031541	
8.075914	
-5.674245	
2.556212	
...	
-52.195329	
-41.341395	
-29.013892	
-52.293518	
-50.745034	

ws × 16 columns