TC300C_Etapa 4.3-Correlaciones

April 23, 2024

1 4.3 Correlaciones (Avance Evidencia 1)

Pandalytics - Equipo 1 * **A00832444** | Andrea Garza * **A01197991** | Hiram Maximiliano Muñoz Ramírez * **A00517124** | Erick Orlando Hernández Vallejo * **A01197655** | Raúl Isaí Murillo Alemán * **A01235692** | David Gerardo Martínez Hidrogo

---

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
import statsmodels.api as sm
import scipy.stats as stats
from scipy.stats import f_oneway
from scipy.stats import chi2_contingency


def print_chi_summary(chi_result):
    chi2, p, dof, ex = chi_result
    print(
        f'\tChi_square value \033[1m{chi2:.5f}\033[0m\n\tp value␣
 ↪\033[1m{p}\033[0m\n\tdegrees of freedom \033[1m{dof}\033[0m')


#ANOVA considerando las variables numericas
# Seleccionar las columnas numéricas relevantes
def do_anova(df: pd.DataFrame, group: str):
    grouped_values = df.groupby(group, observed=True)

    f_statistics, p_values = f_oneway(
        *map(lambda group_name: grouped_values.get_group(group_name).
 ↪_get_numeric_data().fillna(0),
             grouped_values.groups.keys()),
        nan_policy='omit')

    alpha = 0.05
```

```
    print(f"Values grouped by {group}")
    for idx, col_name in enumerate(df._get_numeric_data().columns):
        p_value = p_values[idx]
        if p_value < alpha:
            print(
                f"p value of {p_value}. Reject the null hypothesis for␣
 ↪{col_name}. There is a significant difference in the means.")
        else:
            print(
                f"p value of {p_value}. Fail to reject the null hypothesis for␣
 ↪{col_name}. There is no significant difference in the means.")

%matplotlib inline
paint_per_date_df = pd.read_feather('data/paint_per_date.feather')
paint_catalog_df = pd.read_feather('data/paint_catalog.feather')
```

[2]:
```
paint_per_date_df
```

[2]:
```
length_m  \
paint_name                    date       production_line user
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
2193.000000
                              2022-01-17 Pintado 2       ALEINSUMOS
5415.000000
                              2022-01-20 Pintado 1       NaN
9803.000000
                              2022-01-21 Pintado 1       ALEINSUMOS
3553.000000
                              2022-01-22 Pintado 1       ALEINSUMOS
4024.000000
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
1308.000000
                                         Pintado 2       ALEINSUMOS
51680.000000
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
33450.924101
                              2022-07-24 Pintado 2       ALEINSUMOS
10482.000000
                              2022-08-30 Pintado 2       ALEINSUMOS
1212.000000

m2  \
paint_name                    date       production_line user
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
2616.249000
```

```
                              2022-01-17 Pintado 2       ALEINSUMOS
4972.966000
                              2022-01-20 Pintado 1       NaN
12378.306000
                              2022-01-21 Pintado 1       ALEINSUMOS
3985.378000
                              2022-01-22 Pintado 1       ALEINSUMOS
3399.081000
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
1278.516000
                                         Pintado 2       ALEINSUMOS
43799.101917
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
32844.694259
                              2022-07-24 Pintado 2       ALEINSUMOS
10166.061000
                              2022-08-30 Pintado 2       ALEINSUMOS
1481.064000

input_weight_kg  \
paint_name                    date       production_line user
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
48915.00000
                              2022-01-17 Pintado 2       ALEINSUMOS
78318.00000
                              2022-01-20 Pintado 1       NaN
120892.00000
                              2022-01-21 Pintado 1       ALEINSUMOS
43759.00000
                              2022-01-22 Pintado 1       ALEINSUMOS
40393.00000
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
14305.00000
                                         Pintado 2       ALEINSUMOS
863001.00000
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
382250.54132
                              2022-07-24 Pintado 2       ALEINSUMOS
150833.00000
                              2022-08-30 Pintado 2       ALEINSUMOS
21800.00000

weight_kg  \
```

```
paint_name                    date       production_line user
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
16060.000000
                              2022-01-17 Pintado 2       ALEINSUMOS
23810.000000
                              2022-01-20 Pintado 1       NaN
72622.000000
                              2022-01-21 Pintado 1       ALEINSUMOS
23425.000000
                              2022-01-22 Pintado 1       ALEINSUMOS
20327.000000
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
6109.000000
                                         Pintado 2       ALEINSUMOS
420438.000000
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
161899.807635
                              2022-07-24 Pintado 2       ALEINSUMOS
57846.000000
                              2022-08-30 Pintado 2       ALEINSUMOS
7170.000000

avg_thickness_mm  \
paint_name                    date       production_line user
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
0.770000
                              2022-01-17 Pintado 2       ALEINSUMOS
0.602967
                              2022-01-20 Pintado 1       NaN
0.732908
                              2022-01-21 Pintado 1       ALEINSUMOS
0.727200
                              2022-01-22 Pintado 1       ALEINSUMOS
0.737667
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
0.600667
                                         Pintado 2       ALEINSUMOS
1.209975
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
0.659627
                              2022-07-24 Pintado 2       ALEINSUMOS
0.709167
                              2022-08-30 Pintado 2       ALEINSUMOS
```

```
0.605000

total_liters_used  \
paint_name                     date       production_line user
0001-PRIMER 4457               2022-01-16 Pintado 2       ALEINSUMOS
70.00
                               2022-01-17 Pintado 2       ALEINSUMOS
70.00
                               2022-01-20 Pintado 1       NaN
NaN
                               2022-01-21 Pintado 1       ALEINSUMOS
400.00
                               2022-01-22 Pintado 1       ALEINSUMOS
300.00
…
…
2453-GRAY BACKER EDGE          2023-08-21 Pintado 1       NaN
NaN
                                          Pintado 2       ALEINSUMOS
800.00
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
218.25
                               2022-07-24 Pintado 2       ALEINSUMOS
160.25
                               2022-08-30 Pintado 2       ALEINSUMOS
59.00

monetary_value_usd  \
paint_name                     date       production_line user
0001-PRIMER 4457               2022-01-16 Pintado 2       ALEINSUMOS
415.80
                               2022-01-17 Pintado 2       ALEINSUMOS
415.80
                               2022-01-20 Pintado 1       NaN
NaN
                               2022-01-21 Pintado 1       ALEINSUMOS
2376.00
                               2022-01-22 Pintado 1       ALEINSUMOS
1782.00
…
…
2453-GRAY BACKER EDGE          2023-08-21 Pintado 1       NaN
NaN
                                          Pintado 2       ALEINSUMOS
7656.00
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
2490.23
```

```
                            2022-07-24 Pintado 2       ALEINSUMOS
1828.45
                            2022-08-30 Pintado 2       ALEINSUMOS
673.19

expected_yield  \
paint_name                  date       production_line user
0001-PRIMER 4457            2022-01-16 Pintado 2       ALEINSUMOS
58.267717
                            2022-01-17 Pintado 2       ALEINSUMOS
58.267717
                            2022-01-20 Pintado 1       NaN
58.267717
                            2022-01-21 Pintado 1       ALEINSUMOS
58.267717
                            2022-01-22 Pintado 1       ALEINSUMOS
58.267717
…
…
2453-GRAY BACKER EDGE       2023-08-21 Pintado 1       NaN
70.866142
                                       Pintado 2       ALEINSUMOS
70.866142
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2     ALEINSUMOS
81.102362
                            2022-07-24 Pintado 2       ALEINSUMOS
81.102362
                            2022-08-30 Pintado 2       ALEINSUMOS
81.102362

                                                                  real_yield
\
paint_name                  date       production_line user
0001-PRIMER 4457            2022-01-16 Pintado 2       ALEINSUMOS   37.374986
                            2022-01-17 Pintado 2       ALEINSUMOS   71.042371
                            2022-01-20 Pintado 1       NaN                NaN
                            2022-01-21 Pintado 1       ALEINSUMOS    9.963445
                            2022-01-22 Pintado 1       ALEINSUMOS   11.330270
…                                                                         …
2453-GRAY BACKER EDGE       2023-08-21 Pintado 1       NaN                NaN
                                       Pintado 2       ALEINSUMOS   54.748877
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2     ALEINSUMOS  150.491154
                            2022-07-24 Pintado 2       ALEINSUMOS   63.438758
                            2022-08-30 Pintado 2       ALEINSUMOS   25.102780

yield_difference
paint_name                  date       production_line user
```

```
0001-PRIMER 4457              2022-01-16 Pintado 2       ALEINSUMOS
20.892731
                              2022-01-17 Pintado 2       ALEINSUMOS
12.774655
                              2022-01-20 Pintado 1       NaN
NaN
                              2022-01-21 Pintado 1       ALEINSUMOS
48.304272
                              2022-01-22 Pintado 1       ALEINSUMOS
46.937447
…
…
2453-GRAY BACKER EDGE         2023-08-21 Pintado 1       NaN
NaN
                                         Pintado 2       ALEINSUMOS
16.117264
2470-HG GRAY POLYESTER BACKER 2022-06-12 Pintado 2       ALEINSUMOS
69.388791
                              2022-07-24 Pintado 2       ALEINSUMOS
17.663604
                              2022-08-30 Pintado 2       ALEINSUMOS
55.999583

[3872 rows x 10 columns]
```

[3]:
```
paint_catalog_df
```

[3]:
```
      group              paint_name    paint_family paint_code  \
0         1        0001-PRIMER 4457  URETANO PRIMER   NI500000
1        60        0001-PRIMER 4457  URETANO PRIMER   NI500000
2      1230        0001-PRIMER 4457  URETANO PRIMER   NI500000
3      1456        0001-PRIMER 4457  URETANO PRIMER   NI500000
4      1952        0001-PRIMER 4457  URETANO PRIMER   NI500000
…     …                        …               …        …
3168   3583          8703-YELLOW 53   POLIESTER STD   NI501752
3169   3584            8900-RED 254   POLIESTER STD   NI501753
3170   3585          8901-VIOLET 19   POLIESTER STD   NI501754
3171   3586     8902-RED IRON OXIDE   POLIESTER STD   NI501755
3172   3587  8903-RED IRON OXIDE BS   POLIESTER STD   NI501756

                              supplier      product_class unified_key  \
0     VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  I1001_VALS
1     VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  I1001_VALS
2     VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  I1001_VALS
3     VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  I1001_VALS
4     VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  I1001_VALS
…                                  …                …         …
```

```
3168  VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  U8703_VALS
3169  VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  U8900_VALS
3170  VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  U8901_VALS
3171  VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  U8902_VALS
3172  VALS - VALSPAR ARIES COATINGS, S  N-I-PINTURA-LIQ-_  U8903_VALS

     clear_desc  density primer  …    solvent_3 paint_catalog_yield  \
0          None     2.30   None  …         None                49.3
1          None     2.30   None  …         None                49.3
2          None     2.30   None  …         None                49.3
3          None     2.26   None  …  AROMINA 150                48.0
4          None     2.26   None  …  AROMINA 150                48.0
…           …       …    …  …  …          …                   …
3168       None     4.28   None  …         None                 NaN
3169       None     2.26   None  …         None                 NaN
3170       None     2.09   None  …         None                 NaN
3171       None     4.34   None  …         None                 NaN
3172       None     3.98   None  …         None                 NaN

     solid_by_weight  solid_by_volume  substratum_1  substratum_2  \
0              102.0             79.0   GALVANIZADO   ACERO NEGRO
1              102.0             79.0   GALVANIZADO          None
2              102.0             79.0   GALVANIZADO    ZINTROALUM
3              100.0             77.0   GALVANIZADO   ACERO NEGRO
4              100.0             77.0   ACERO NEGRO   GALVANIZADO
…              …                …           …           …
3168           164.4              0.0         None          None
3169           130.0              0.0         None          None
3170           108.2              0.0         None          None
3171           158.0              0.0         None          None
3172           157.4              0.0         None          None

     substratum_3 metal_temp viscosity  canning_yield
0            None      481.0      40.0      59.055118
1            None      481.0      40.0      59.055118
2     ACERO NEGRO      481.0      40.0      59.055118
3            None      473.0      35.0      57.480315
4            None      473.0      35.0      57.480315
…           …        …       …            …
3168         None        NaN     160.0            NaN
3169         None        NaN     170.0            NaN
3170         None        NaN     150.0            NaN
3171         None        NaN     170.0            NaN
3172         None        NaN     140.0            NaN

[3173 rows x 22 columns]
```

1.1 Correlación y Mapas de calor

[4]:
```
#. Análisis de la correlación de todas las variables numéricas
numeric_columns_paint = paint_per_date_df[
    ['length_m', 'm2', 'input_weight_kg', 'weight_kg', 'avg_thickness_mm',␣
 ↪'total_liters_used', 'expected_yield',
     'real_yield', 'yield_difference']]
numeric_columns_paint.corr()
```

[4]:
```
                    length_m        m2  input_weight_kg  weight_kg  \
length_m            1.000000  0.981958         0.851163   0.918102
m2                  0.981958  1.000000         0.901572   0.944364
input_weight_kg     0.851163  0.901572         1.000000   0.936647
weight_kg           0.918102  0.944364         0.936647   1.000000
avg_thickness_mm   -0.193886 -0.193762        -0.010013   0.032891
total_liters_used   0.614047  0.662729         0.685833   0.672199
expected_yield     -0.153932 -0.115686        -0.006942  -0.033002
real_yield          0.303295  0.285124         0.208675   0.242395
yield_difference    0.203597  0.179853         0.121940   0.165699

                    avg_thickness_mm  total_liters_used  expected_yield  \
length_m                   -0.193886           0.614047       -0.153932
m2                         -0.193762           0.662729       -0.115686
input_weight_kg            -0.010013           0.685833       -0.006942
weight_kg                   0.032891           0.672199       -0.033002
avg_thickness_mm            1.000000          -0.046754        0.077539
total_liters_used          -0.046754           1.000000        0.005182
expected_yield              0.077539           0.005182        1.000000
real_yield                 -0.161590          -0.227571       -0.030187
yield_difference           -0.035275          -0.112363       -0.012453

                    real_yield  yield_difference
length_m              0.303295          0.203597
m2                    0.285124          0.179853
input_weight_kg       0.208675          0.121940
weight_kg             0.242395          0.165699
avg_thickness_mm     -0.161590         -0.035275
total_liters_used    -0.227571         -0.112363
expected_yield       -0.030187         -0.012453
real_yield            1.000000          0.857967
yield_difference      0.857967          1.000000
```

[5]:
```
#. Análisis de la correlación de todas las variables numéricas
numeric_columns_catalog = paint_catalog_df[
    ['density', 'paint_catalog_yield', 'solid_by_weight', 'solid_by_volume',␣
 ↪'metal_temp', 'viscosity',
     'canning_yield']]
```

```
numeric_columns_catalog.corr()
```

[5]:

	density	paint_catalog_yield	solid_by_weight	\
density	1.000000	-0.168144	0.789069	
paint_catalog_yield	-0.168144	1.000000	-0.162485	
solid_by_weight	0.789069	-0.162485	1.000000	
solid_by_volume	0.249213	-0.092267	0.665292	
metal_temp	0.041121	0.020195	0.094094	
viscosity	0.234810	-0.250440	0.207961	
canning_yield	-0.237723	0.954408	-0.166872	

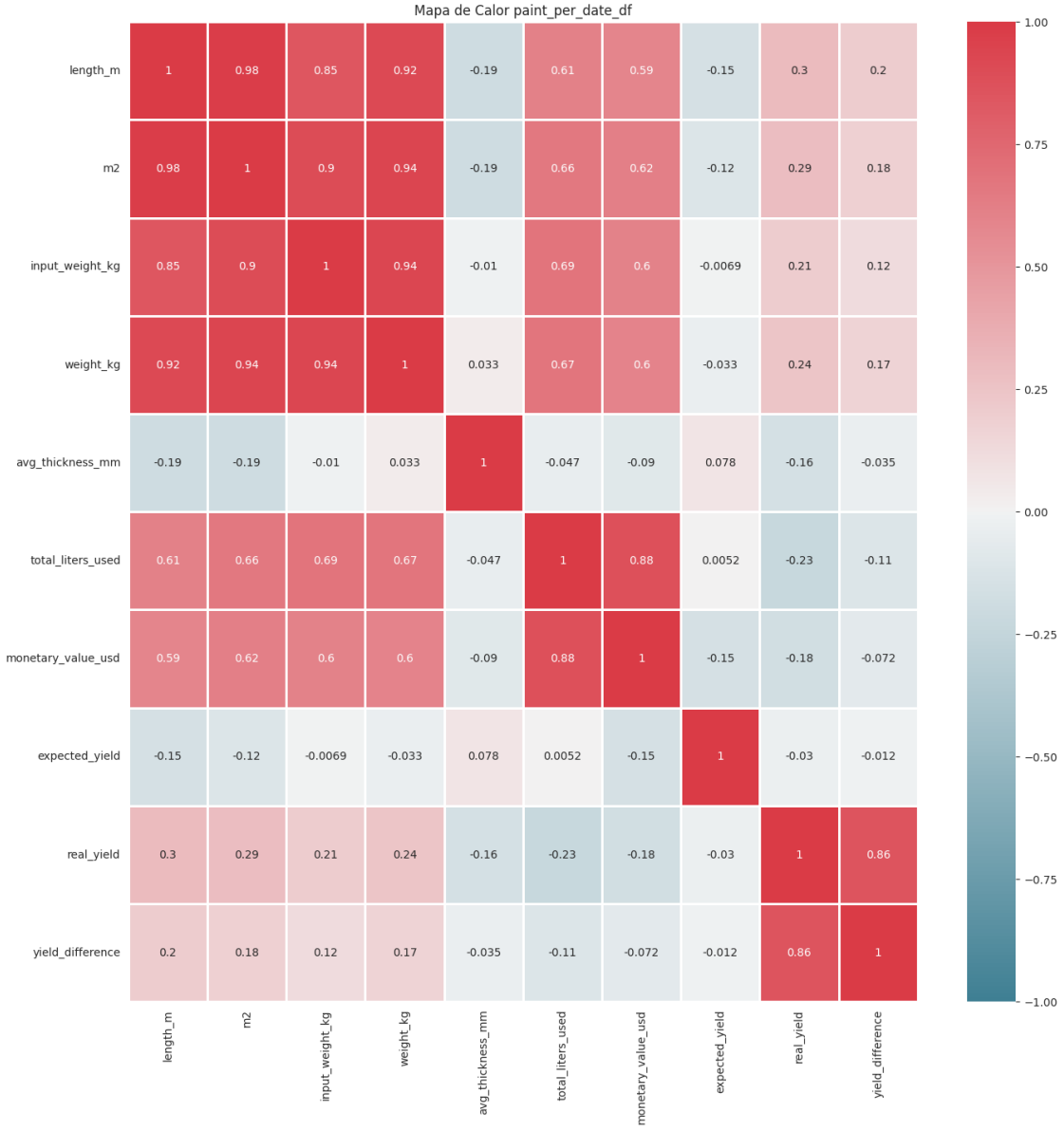
	solid_by_volume	metal_temp	viscosity	canning_yield
density	0.249213	0.041121	0.234810	-0.237723
paint_catalog_yield	-0.092267	0.020195	-0.250440	0.954408
solid_by_weight	0.665292	0.094094	0.207961	-0.166872
solid_by_volume	1.000000	0.102213	-0.213371	-0.145608
metal_temp	0.102213	1.000000	0.010146	-0.015490
viscosity	-0.213371	0.010146	1.000000	-0.269917
canning_yield	-0.145608	-0.015490	-0.269917	1.000000

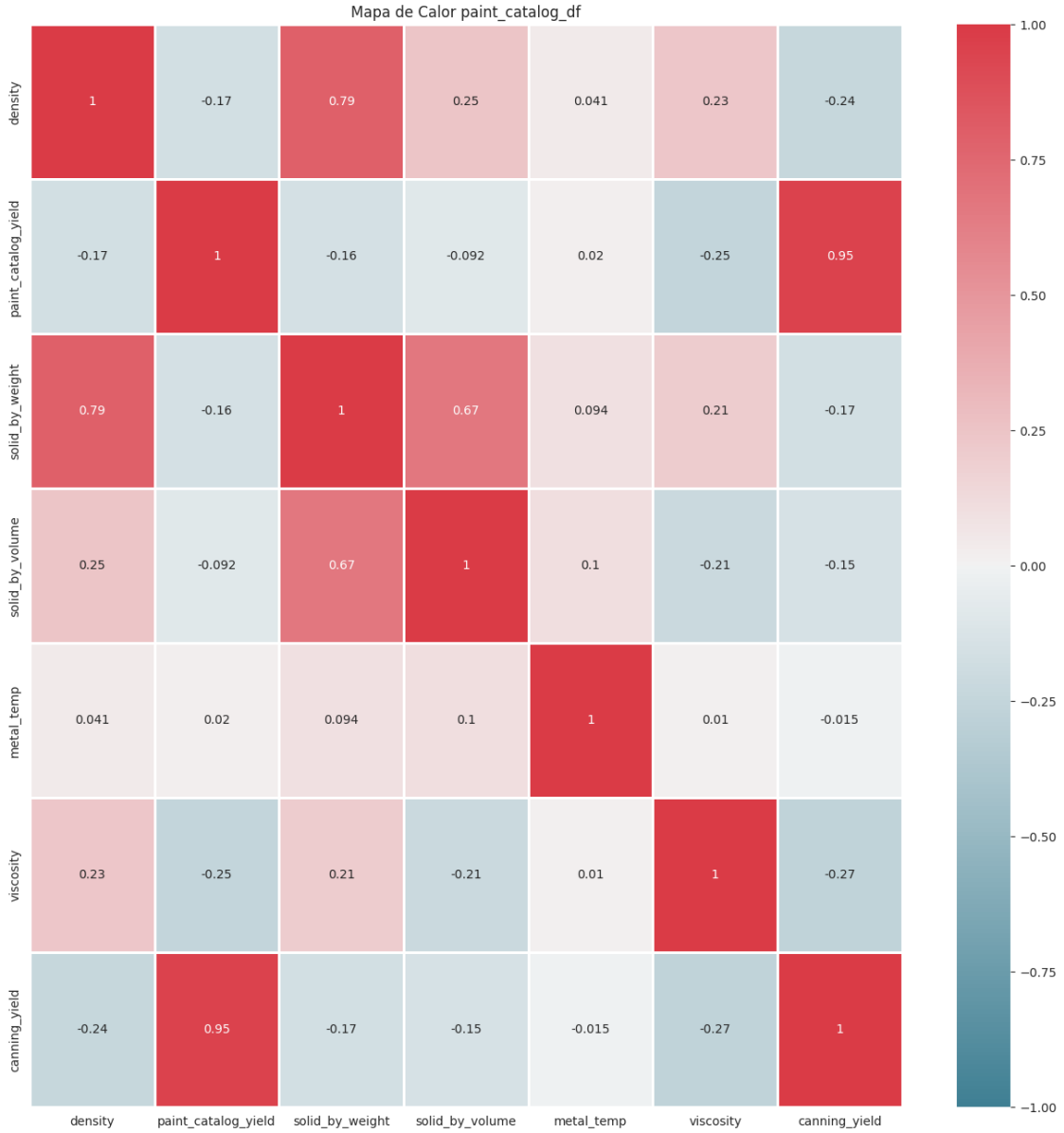
[6]:
```
#Mapa de Calor con  variables numéricas de la base de datos, incluyendo las que␣
 ↪fueron creando (en caso de que aplique)
plt.figure(figsize=(16, 16))
sns.heatmap(paint_per_date_df.corr(), annot=True, linewidths=0.75,␣
 ↪linecolor='white',
            cmap=sns.diverging_palette(220, 10, as_cmap=True), vmin=-1, vmax=1)
plt.title('Mapa de Calor paint_per_date_df')

plt.figure(figsize=(16, 16))
sns.heatmap(numeric_columns_catalog.corr(), annot=True, linewidths=0.75,␣
 ↪linecolor='white',
            cmap=sns.diverging_palette(220, 10, as_cmap=True), vmin=-1, vmax=1)
plt.title('Mapa de Calor paint_catalog_df')

#Explicacion del comportamiento de las variables
```

[6]:
```
Text(0.5, 1.0, 'Mapa de Calor paint_catalog_df')
```


Mapa de Calor paint_per_date_df
length_m 1 0.98 0.85 0.92 -0.19 0.61 0.59 -0.15 0.3 0.2
m2 0.98 1 0.9 0.94 -0.19 0.66 0.62 -0.12 0.29 0.18
input_weight_kg 0.85 0.9 1 0.94 -0.01 0.69 0.6 -0.0069 0.21 0.12
weight_kg 0.92 0.94 0.94 1 0.033 0.67 0.6 -0.033 0.24 0.17
avg_thickness_mm -0.19 -0.19 -0.01 0.033 1 -0.047 -0.09 0.078 -0.16 -0.035
total_liters_used 0.61 0.66 0.69 0.67 -0.047 1 0.88 0.0052 -0.23 -0.11
monetary_value_usd 0.59 0.62 0.6 0.6 -0.09 0.88 1 -0.15 -0.18 -0.072
expected_yield -0.15 -0.12 -0.0069 -0.033 0.078 0.0052 -0.15 1 -0.03 -0.012
real_yield 0.3 0.29 0.21 0.24 -0.16 -0.23 -0.18 -0.03 1 0.86
yield_difference 0.2 0.18 0.12 0.17 -0.035 -0.11 -0.072 -0.012 0.86 1
length_m
m2
input_weight_kg
weight_kg
avg_thickness_mm
total_liters_used
monetary_value_usd
expected_yield
real_yield
yield_difference
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



1.2 ANOVA

[7]:
```
do_anova(paint_per_date_df, 'paint_name')
```

```
Values grouped by paint_name
p value of 2.1231760211666068e-213. Reject the null hypothesis for length_m.
There is a significant difference in the means.
p value of 1.8870328080734653e-155. Reject the null hypothesis for m2. There is
a significant difference in the means.
p value of 2.5908960037089756e-98. Reject the null hypothesis for
input_weight_kg. There is a significant difference in the means.
```

```
p value of 1.0153328316716487e-130. Reject the null hypothesis for weight_kg.
There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for avg_thickness_mm. There is a
significant difference in the means.
p value of 2.972197261453037e-129. Reject the null hypothesis for
total_liters_used. There is a significant difference in the means.
p value of 4.2166326213172925e-166. Reject the null hypothesis for
monetary_value_usd. There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for expected_yield. There is a
significant difference in the means.
p value of 1.4467592434239122e-77. Reject the null hypothesis for real_yield.
There is a significant difference in the means.
p value of 6.606080708960725e-54. Reject the null hypothesis for
yield_difference. There is a significant difference in the means.

/home/hiram/.cache/pypoetry/virtualenvs/pandalytics-
Dnh4JP0f-py3.11/lib/python3.11/site-
packages/scipy/stats/_axis_nan_policy.py:563: ConstantInputWarning: Each of the
input arrays is constant; the F statistic is not defined or infinite
  res = hypotest_fun_out(*samples, axis=axis, **kwds)
```

[8]:
```
do_anova(paint_per_date_df, 'production_line')
```

```
Values grouped by production_line
p value of 3.964373257537889e-27. Reject the null hypothesis for length_m. There
is a significant difference in the means.
p value of 1.1732222398219633e-20. Reject the null hypothesis for m2. There is a
significant difference in the means.
p value of 1.1960528131106033e-15. Reject the null hypothesis for
input_weight_kg. There is a significant difference in the means.
p value of 1.3420178303802961e-33. Reject the null hypothesis for weight_kg.
There is a significant difference in the means.
p value of 0.5311094429773557. Fail to reject the null hypothesis for
avg_thickness_mm. There is no significant difference in the means.
p value of 0.0013964629569917658. Reject the null hypothesis for
total_liters_used. There is a significant difference in the means.
p value of 0.11090475335599351. Fail to reject the null hypothesis for
monetary_value_usd. There is no significant difference in the means.
p value of 0.025666942120083764. Reject the null hypothesis for expected_yield.
There is a significant difference in the means.
p value of 5.387605980758937e-08. Reject the null hypothesis for real_yield.
There is a significant difference in the means.
p value of 0.0004080628635374767. Reject the null hypothesis for
yield_difference. There is a significant difference in the means.
```

[9]:
```
do_anova(paint_per_date_df, 'user')
```

```
Values grouped by user
p value of 0.35685659281245957. Fail to reject the null hypothesis for length_m.
```

```
There is no significant difference in the means.
p value of 0.27179268981969906. Fail to reject the null hypothesis for m2. There
is no significant difference in the means.
p value of 0.24544711772992098. Fail to reject the null hypothesis for
input_weight_kg. There is no significant difference in the means.
p value of 0.3643146885206361. Fail to reject the null hypothesis for weight_kg.
There is no significant difference in the means.
p value of 0.6816529348787248. Fail to reject the null hypothesis for
avg_thickness_mm. There is no significant difference in the means.
p value of 6.688338137918233e-05. Reject the null hypothesis for
total_liters_used. There is a significant difference in the means.
p value of 0.00015378262986798905. Reject the null hypothesis for
monetary_value_usd. There is a significant difference in the means.
p value of 0.5503855664964049. Fail to reject the null hypothesis for
expected_yield. There is no significant difference in the means.
p value of 9.19507469176006e-28. Reject the null hypothesis for real_yield.
There is a significant difference in the means.
p value of 2.1012497170409713e-27. Reject the null hypothesis for
yield_difference. There is a significant difference in the means.
```

[10]:
```
do_anova(paint_per_date_df, 'date')
```

```
Values grouped by date
p value of 1.3437859013201845e-35. Reject the null hypothesis for length_m.
There is a significant difference in the means.
p value of 6.445919167347926e-39. Reject the null hypothesis for m2. There is a
significant difference in the means.
p value of 1.2987244842926618e-49. Reject the null hypothesis for
input_weight_kg. There is a significant difference in the means.
p value of 1.5849689780277744e-38. Reject the null hypothesis for weight_kg.
There is a significant difference in the means.
p value of 1.7542212842067406e-10. Reject the null hypothesis for
avg_thickness_mm. There is a significant difference in the means.
p value of 3.1839157844594584e-35. Reject the null hypothesis for
total_liters_used. There is a significant difference in the means.
p value of 9.103677580808844e-24. Reject the null hypothesis for
monetary_value_usd. There is a significant difference in the means.
p value of 0.9945251522611709. Fail to reject the null hypothesis for
expected_yield. There is no significant difference in the means.
p value of 9.474730131113608e-15. Reject the null hypothesis for real_yield.
There is a significant difference in the means.
p value of 8.555469077280981e-07. Reject the null hypothesis for
yield_difference. There is a significant difference in the means.
```

[11]:
```
do_anova(paint_catalog_df, 'paint_name')
```

```
Values grouped by paint_name
p value of 3.554193854400297e-193. Reject the null hypothesis for group. There
is a significant difference in the means.
```

```
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 0.0. Reject the null hypothesis for paint_catalog_yield. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_volume. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for metal_temp. There is a
significant difference in the means.
p value of 2.394244846763394e-250. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for canning_yield. There is a
significant difference in the means.
```

[12]:
```
do_anova(paint_catalog_df, 'paint_family')
```

```
Values grouped by paint_family
p value of 1.0969074253557855e-159. Reject the null hypothesis for group. There
is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 0.0. Reject the null hypothesis for paint_catalog_yield. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_volume. There is a
significant difference in the means.
p value of 3.058417180954786e-41. Reject the null hypothesis for metal_temp.
There is a significant difference in the means.
p value of 1.2691801321848717e-185. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.00012753415734051643. Reject the null hypothesis for canning_yield.
There is a significant difference in the means.
```

[13]:
```
do_anova(paint_catalog_df, 'paint_code')
```

```
Values grouped by paint_code
p value of 5.211104002649445e-185. Reject the null hypothesis for group. There
is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 0.0. Reject the null hypothesis for paint_catalog_yield. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_volume. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for metal_temp. There is a
```

```
significant difference in the means.
p value of 7.645431239209812e-248. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for canning_yield. There is a
significant difference in the means.
```

[14]:
```
do_anova(paint_catalog_df, 'supplier')
```

```
Values grouped by supplier
p value of 5.794954484687814e-25. Reject the null hypothesis for group. There is
a significant difference in the means.
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 9.055142629446494e-44. Reject the null hypothesis for
paint_catalog_yield. There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 1.3633644263055647e-276. Reject the null hypothesis for
solid_by_volume. There is a significant difference in the means.
p value of 2.4985436323451852e-27. Reject the null hypothesis for metal_temp.
There is a significant difference in the means.
p value of 5.606448657244857e-119. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.3167601516155411. Fail to reject the null hypothesis for
canning_yield. There is no significant difference in the means.
```

[15]:
```
do_anova(paint_catalog_df, 'product_class')
```

```
Values grouped by product_class
p value of 6.235895718078614e-14. Reject the null hypothesis for group. There is
a significant difference in the means.
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 4.782886386151387e-28. Reject the null hypothesis for
paint_catalog_yield. There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 4.473368443085131e-260. Reject the null hypothesis for
solid_by_volume. There is a significant difference in the means.
p value of 8.49253720975088e-14. Reject the null hypothesis for metal_temp.
There is a significant difference in the means.
p value of 9.972271616150101e-115. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.0067667976817338315. Reject the null hypothesis for canning_yield.
There is a significant difference in the means.
```

[16]:
```
do_anova(paint_catalog_df, 'unified_key')
```

```
Values grouped by unified_key
```

```
p value of 1.2525204572445463e-199. Reject the null hypothesis for group. There
is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for density. There is a significant
difference in the means.
p value of 0.0. Reject the null hypothesis for paint_catalog_yield. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_weight. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for solid_by_volume. There is a
significant difference in the means.
p value of 0.0. Reject the null hypothesis for metal_temp. There is a
significant difference in the means.
p value of 1.4735989734164414e-243. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 0.0. Reject the null hypothesis for canning_yield. There is a
significant difference in the means.
```

[17]:
```
do_anova(paint_catalog_df, 'solvent_1')
```

```
Values grouped by solvent_1
p value of 4.146291890974436e-33. Reject the null hypothesis for group. There is
a significant difference in the means.
p value of 4.2117046244627074e-10. Reject the null hypothesis for density. There
is a significant difference in the means.
p value of 1.5493470116837203e-26. Reject the null hypothesis for
paint_catalog_yield. There is a significant difference in the means.
p value of 2.205315540368534e-21. Reject the null hypothesis for
solid_by_weight. There is a significant difference in the means.
p value of 2.496493575289145e-30. Reject the null hypothesis for
solid_by_volume. There is a significant difference in the means.
p value of 1.7813823629405235e-141. Reject the null hypothesis for metal_temp.
There is a significant difference in the means.
p value of 0.0002345054021749633. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 3.484179735341677e-16. Reject the null hypothesis for canning_yield.
There is a significant difference in the means.
```

[18]:
```
do_anova(paint_catalog_df, 'solvent_2')
```

```
Values grouped by solvent_2
p value of 9.079543137275284e-34. Reject the null hypothesis for group. There is
a significant difference in the means.
p value of 0.002429930829461564. Reject the null hypothesis for density. There
is a significant difference in the means.
p value of 3.0125538241695835e-22. Reject the null hypothesis for
paint_catalog_yield. There is a significant difference in the means.
p value of 5.4670663167896286e-05. Reject the null hypothesis for
solid_by_weight. There is a significant difference in the means.
p value of 2.141821943124299e-08. Reject the null hypothesis for
```

```
solid_by_volume. There is a significant difference in the means.
p value of 3.134966857487814e-14. Reject the null hypothesis for metal_temp.
There is a significant difference in the means.
p value of 5.848341216109666e-07. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 4.366689071397412e-15. Reject the null hypothesis for canning_yield.
There is a significant difference in the means.
```

[19]:
```
do_anova(paint_catalog_df, 'solvent_3')
```

```
Values grouped by solvent_3
p value of 3.1328644018574936e-29. Reject the null hypothesis for group. There
is a significant difference in the means.
p value of 2.990422113669359e-08. Reject the null hypothesis for density. There
is a significant difference in the means.
p value of 2.961008662063079e-14. Reject the null hypothesis for
paint_catalog_yield. There is a significant difference in the means.
p value of 9.429161126168314e-14. Reject the null hypothesis for
solid_by_weight. There is a significant difference in the means.
p value of 1.826349714057392e-13. Reject the null hypothesis for
solid_by_volume. There is a significant difference in the means.
p value of 0.1274440568491328. Fail to reject the null hypothesis for
metal_temp. There is no significant difference in the means.
p value of 1.1900527740495174e-06. Reject the null hypothesis for viscosity.
There is a significant difference in the means.
p value of 6.279829825364454e-05. Reject the null hypothesis for canning_yield.
There is a significant difference in the means.
```

1.3 Chi-Cuadrado

[20]:
```
paint_per_date_reindex_df = paint_per_date_df.reset_index(level=['paint_name',␣
↪'date', 'production_line', 'user'])
```

1.3.1 paint__per__date__df

[21]:
```
# Lista de variables categóricas de paint_per_date_df
paint_per_date_chi_square_list = ['paint_name', 'date', 'production_line',␣
↪'user']

# Bucle para comparar cada columna categórica
for column in paint_per_date_chi_square_list:
    for col in paint_per_date_chi_square_list:
        if column != col:
            comparison = pd.crosstab(paint_per_date_reindex_df[column],␣
↪paint_per_date_reindex_df[col])
            print(f'\n\033[1mComparación de {column} con {col}:\033[0m')
            chi_result = stats.chi2_contingency(comparison)
            print_chi_summary(chi_result)
```

```
    paint_per_date_chi_square_list.remove(column)
```

```
Comparación de paint_name con date:
        Chi_square value 35708.08576
        p value 1.0
        degrees of freedom 39732

Comparación de paint_name con production_line:
        Chi_square value 1300.56487
        p value 2.6724946653429326e-217
        degrees of freedom 84

Comparación de paint_name con user:
        Chi_square value 440.28136
        p value 0.011245557076248526
        degrees of freedom 375

Comparación de production_line con date:
        Chi_square value 865.34260
        p value 2.1600869132981054e-25
        degrees of freedom 473

Comparación de production_line con user:
        Chi_square value 13.37314
        p value 0.02012200275250467
        degrees of freedom 5
```

1.3.2 paint_catalog_df

[22]:
```
# Lista de variables categóricas de paint_catalog_df
paint_catalog_chi_square_list = ['paint_name', 'paint_family', 'paint_code',␣
 ↪'supplier', 'product_class', 'unified_key',
                                 'solvent_1', 'solvent_2', 'solvent_3']

# Bucle para comparar cada columna categórica
for column in paint_catalog_chi_square_list:
    for col in paint_catalog_chi_square_list:
        if column != col:
            comparison = pd.crosstab(paint_catalog_df[column],␣
 ↪paint_catalog_df[col])
            print(f'\n\033[1mComparación de {column} con {col}:')
            chi_result = stats.chi2_contingency(comparison)
            print_chi_summary(chi_result)
    paint_catalog_chi_square_list.remove(column)
```

Comparación de paint_name con paint_family:

Chi_square value 116432.25595
p value 0.0
degrees of freedom 59940

Comparación de paint_name con paint_code:

Chi_square value 5133914.00000
p value 0.0
degrees of freedom 2664900

Comparación de paint_name con supplier:

Chi_square value 22520.09371
p value 2.5370201936165087e-117
degrees of freedom 17820

Comparación de paint_name con product_class:

Chi_square value 3173.00000
p value 2.616743270351518e-103
degrees of freedom 1620

Comparación de paint_name con unified_key:

Chi_square value 5120417.99286
p value 0.0
degrees of freedom 2920860

Comparación de paint_name con solvent_1:

Chi_square value 16562.50334
p value 0.0
degrees of freedom 9552

Comparación de paint_name con solvent_2:

Chi_square value 16007.10321
p value 1.4328322561035247e-192
degrees of freedom 11016

Comparación de paint_name con solvent_3:

Chi_square value 7527.42607
p value 9.259628491260295e-159
degrees of freedom 4487

Comparación de paint_code con paint_family:

Chi_square value 117401.00000

p value 0.0
degrees of freedom 60865

Comparación de paint_code con supplier:

Chi_square value 22727.37831
p value 8.397603813895209e-113
degrees of freedom 18095

Comparación de paint_code con product_class:

Chi_square value 3173.00000
p value 1.0860808087253163e-99
degrees of freedom 1645

Comparación de paint_code con unified_key:

Chi_square value 5137863.99286
p value 0.0
degrees of freedom 2964132

Comparación de paint_code con solvent_1:

Chi_square value 16649.53264
p value 0.0
degrees of freedom 9702

Comparación de paint_code con solvent_2:

Chi_square value 16030.69812
p value 1.2496896537951174e-182
degrees of freedom 11160

Comparación de paint_code con solvent_3:

Chi_square value 7535.18505
p value 7.216703214593032e-158
degrees of freedom 4501

Comparación de product_class con paint_family:

Chi_square value 3173.00000
p value 0.0
degrees of freedom 37

Comparación de product_class con supplier:

Chi_square value 3173.00000
p value 0.0
degrees of freedom 11

Comparación de product_class con unified_key:

Chi_square value 3172.00000
p value 1.3761299624562628e-78
degrees of freedom 1803

Comparación de product_class con solvent_1:

Chi_square value 0.00000
p value 1.0
degrees of freedom 0

Comparación de product_class con solvent_2:

Chi_square value 0.00000
p value 1.0
degrees of freedom 0

Comparación de product_class con solvent_3:

Chi_square value 0.00000
p value 1.0
degrees of freedom 0

Comparación de solvent_1 con paint_family:

Chi_square value 6556.70582
p value 0.0
degrees of freedom 198

Comparación de solvent_1 con supplier:

Chi_square value 1950.81920
p value 0.0
degrees of freedom 54

Comparación de solvent_1 con unified_key:

Chi_square value 17951.30818
p value 0.0
degrees of freedom 10620

Comparación de solvent_1 con solvent_2:

Chi_square value 2443.70110
p value 0.0
degrees of freedom 40

Comparación de solvent_1 con solvent_3:

Chi_square value 1230.90316

```
        p value 6.5103496269445325e-236
        degrees of freedom 35

Comparación de solvent_3 con paint_family:

        Chi_square value 514.45922
        p value 2.226805521459802e-48
        degrees of freedom 126

Comparación de solvent_3 con supplier:

        Chi_square value 853.92578
        p value 1.1434923196466271e-156
        degrees of freedom 35

Comparación de solvent_3 con unified_key:

        Chi_square value 7719.77809
        p value 3.073168304334859e-152
        degrees of freedom 4697

Comparación de solvent_3 con solvent_2:

        Chi_square value 1474.50571
        p value 6.212782513261318e-282
        degrees of freedom 42
```

[23]:
```
paint_catalog_df.columns
```

[23]:
```
Index(['group', 'paint_name', 'paint_family', 'paint_code', 'supplier',
       'product_class', 'unified_key', 'clear_desc', 'density', 'primer',
       'solvent_1', 'solvent_2', 'solvent_3', 'paint_catalog_yield',
       'solid_by_weight', 'solid_by_volume', 'substratum_1', 'substratum_2',
       'substratum_3', 'metal_temp', 'viscosity', 'canning_yield'],
      dtype='object')
```