05_EDA

August 15, 2024

basic EDA

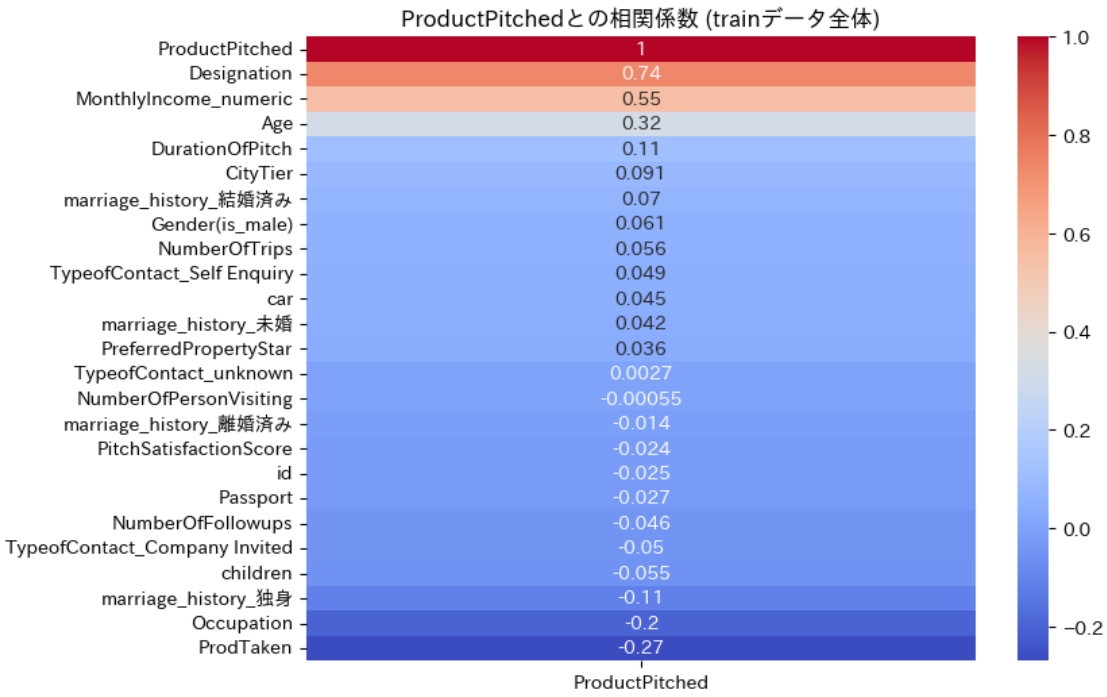
[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import japanize_matplotlib
import seaborn as sns
```

[16]:
```
# https://qiita.com/saka1_p/items/bb4206c6349eb61c073c
palette = sns.color_palette(['#E69F00', '#56B4E9', '#009E73', '#F0E442',␣
↪'#0072B2', '#D55E00', '#CC79A7', '#000000'])
sns.set_palette(palette)
```

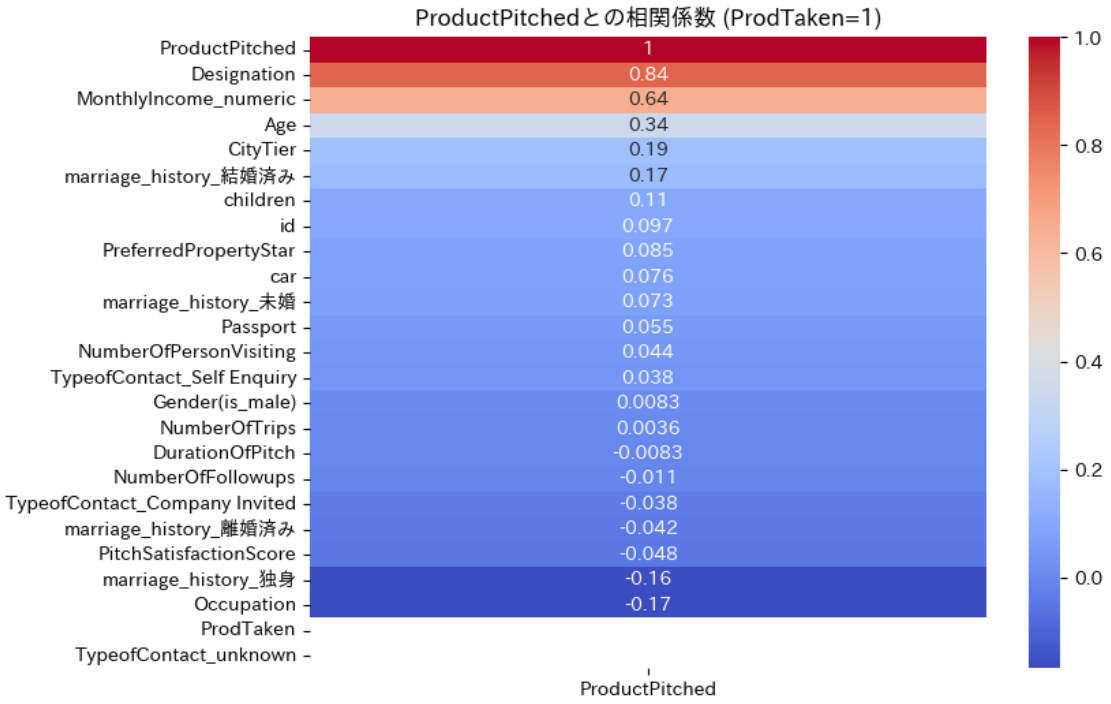
[2]:
```
train_file_path = "../data/20240812/train_preprocessed.csv"
train_df = pd.read_csv(train_file_path)
```

0.1 ProductPitched

[7]:
```
# train_df
correlation = train_df.corr()
correlation = correlation["ProductPitched"].sort_values(ascending=False)
# print(correlation)
# plot correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation.to_frame(), cmap="coolwarm", annot=True)
plt.title("ProductPitched    (train   )")
plt.show()
```

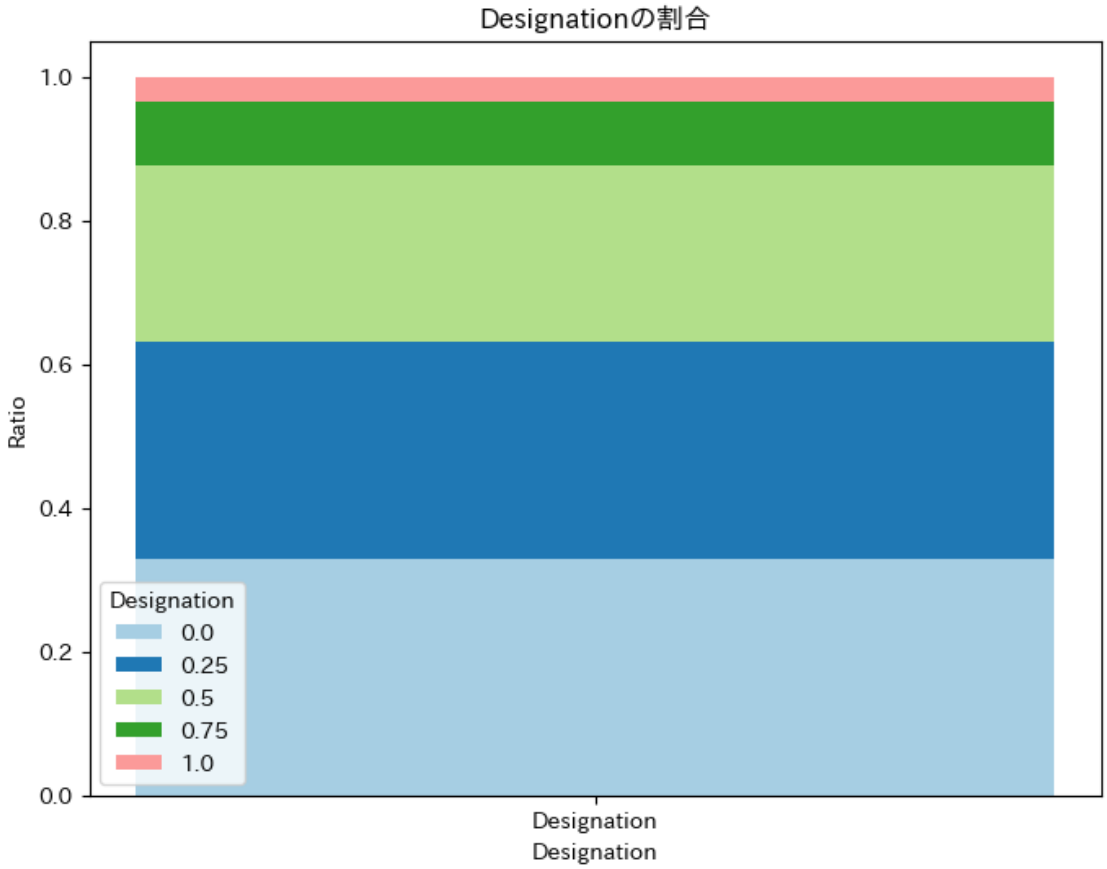


[12]:
```
# ProdTaken=1
prod_taken_df = train_df[train_df["ProdTaken"] == 1]
correlation = prod_taken_df.corr()
correlation_positive = correlation["ProductPitched"].
 ↪sort_values(ascending=False)
# print(correlation)
# plot correlation
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_positive.to_frame(), cmap="coolwarm", annot=True)
plt.title("ProductPitched    (ProdTaken=1)")
plt.show()
```

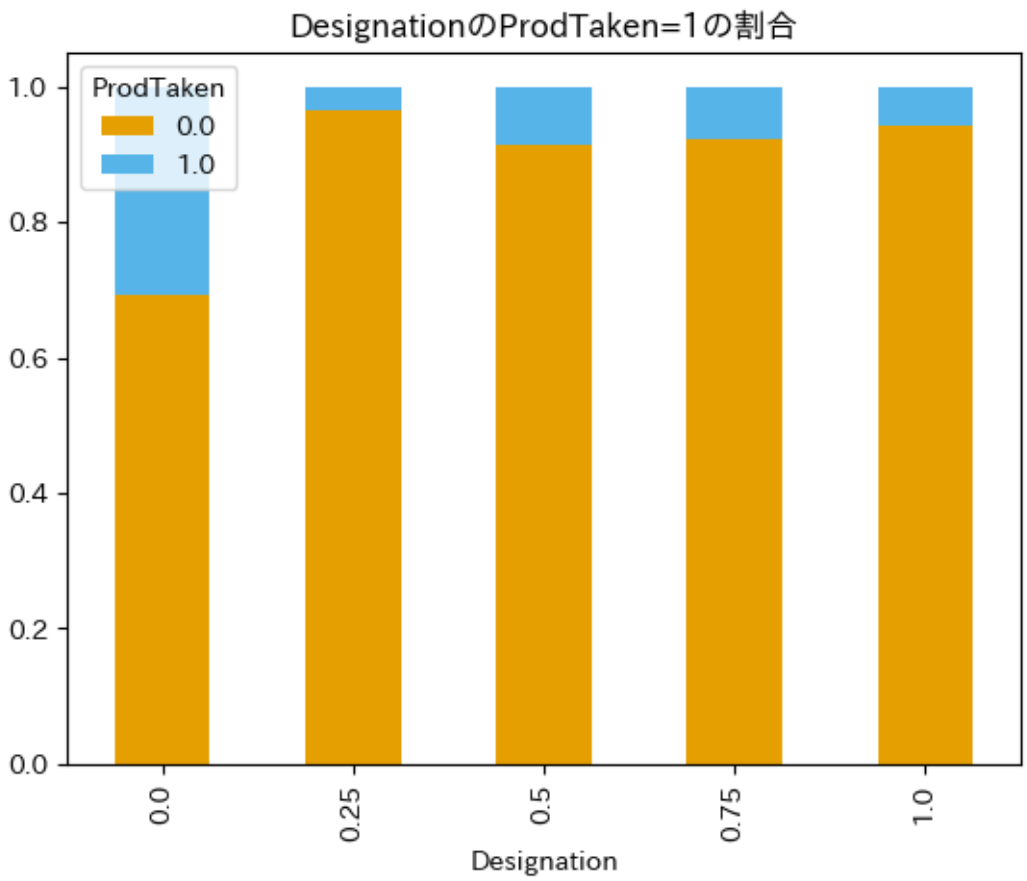


[25]:
```
#
designation_df = train_df["Designation"].value_counts(normalize=True).
 ↪reset_index()
designation_df.columns = ["Designation", "ratio"]
designation_df = designation_df.sort_values("ratio", ascending=False)

#
fig, ax = plt.subplots(figsize=(8, 6))

#
bottom = 0
colors = plt.cm.Paired(range(len(designation_df)))

#
for i, row in designation_df.iterrows():
    ax.bar("Designation",
           row["ratio"],
           bottom=bottom,
           color=colors[i],
           label=row["Designation"])
    bottom += row["ratio"]

#
ax.set_title("Designation  ")
```

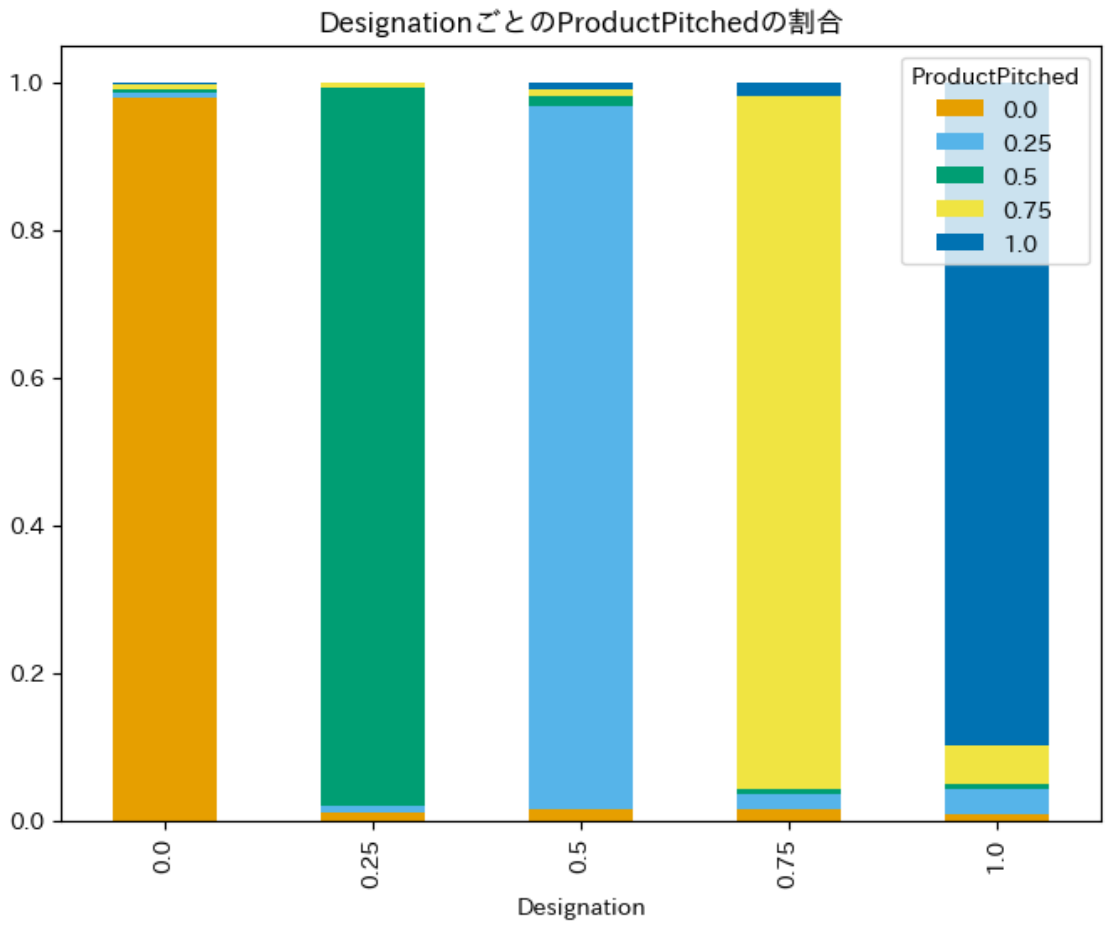
```
ax.set_xlabel("Designation")
ax.set_ylabel("Ratio")
ax.legend(title="Designation")

#
plt.show()
```



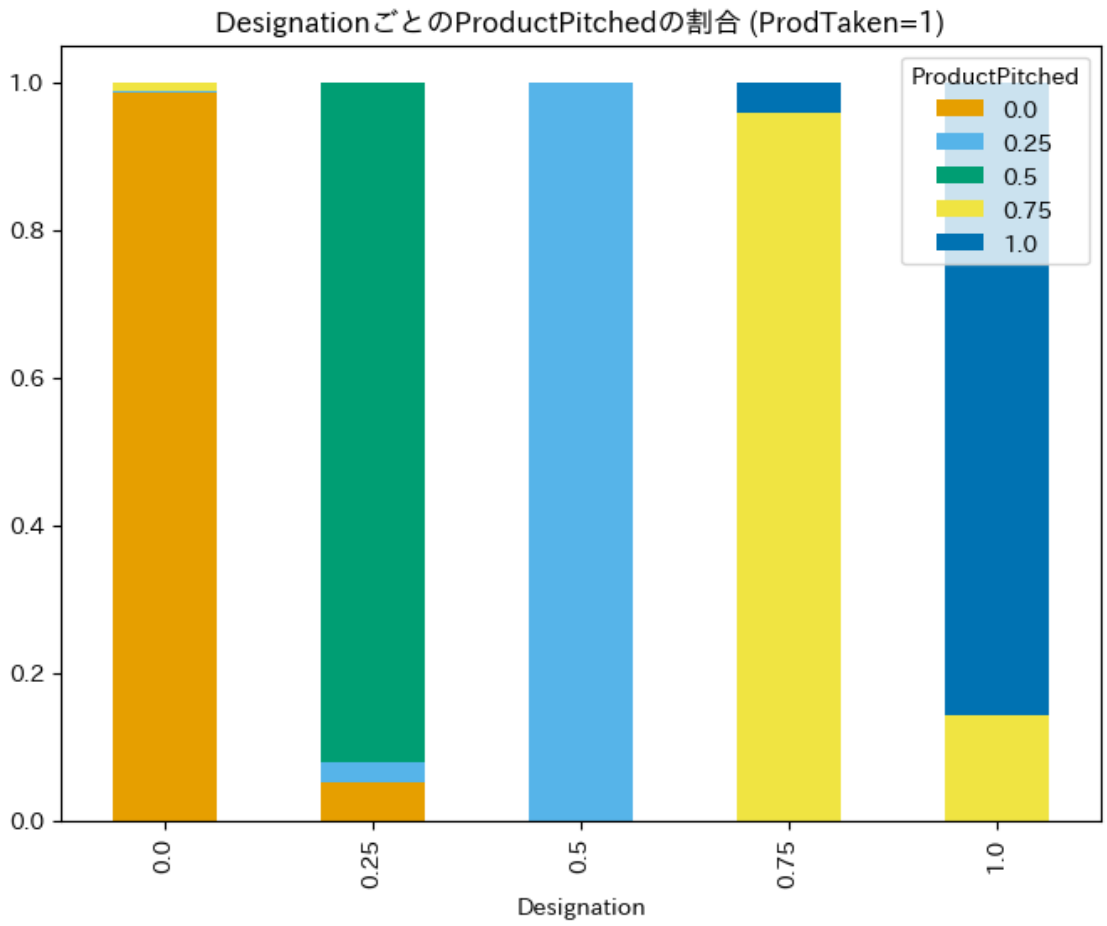
[21]:
```
# Designation      stacked bar plot
designation_df = train_df.groupby(["Designation", "ProdTaken"]).size().unstack()
designation_df = designation_df.div(designation_df.sum(axis=1), axis=0)
designation_df.plot(kind="bar", stacked=True)
plt.title("Designation ProdTaken=1  ")
plt.show()
```



[17]:
```
# Desgnation  ProductPitched stacked bar plot
designation_grouped = train_df.groupby("Designation")["ProductPitched"].
 ↪value_counts(normalize=True).unstack()
designation_grouped.plot(kind="bar", stacked=True, figsize=(8, 6),␣
 ↪title="Designation  ProductPitched  ")
plt.title("Designation  ProductPitched   (train   )")
plt.show()
```



[18]:
```
# ProdTaken=1
prod_taken_df = train_df[train_df["ProdTaken"] == 1]
designation_grouped = prod_taken_df.groupby("Designation")["ProductPitched"].
 ↪value_counts(normalize=True).unstack()
designation_grouped.plot(kind="bar", stacked=True, figsize=(8, 6),␣
 ↪title="Designation  ProductPitched  ")
plt.title("Designation  ProductPitched   (ProdTaken=1)")
plt.show()
```


DesignationごとのProductPitchedの割合 (ProdTaken=1)
ProductPitched
0.0
0.25
0.5
0.75
1.0
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.25
0.5
0.75
1.0
Designation

[ ]: