

0	alcohol	178	non-null	float64
1	malic_acid	178	non-null	float64
2	ash	178	non-null	float64
3	alcalinity_of_ash	178	non-null	float64
4	magnesium	178	non-null	float64
5	total_phenols	178	non-null	float64
6	flavanoids	178	non-null	float64
7	nonflavanoid_phenols	178	non-null	float64
8	proanthocyanins	178	non-null	float64
9	color_intensity	178	non-null	float64
10	hue	178	non-null	float64
11	od280/od315_of_diluted_wines	178	non-null	float64
12	proline	178	non-null	float64
13	target	178	non-null	int64

dtypes: float64(13), int64(1)

memory usage: 19.6 KB

None

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium
total_phenols \					
0	14.23	1.71	2.43	15.6	127.0
2.80					
1	13.20	1.78	2.14	11.2	100.0
2.65					
2	13.16	2.36	2.67	18.6	101.0
2.80					
3	14.37	1.95	2.50	16.8	113.0
3.85					
4	13.24	2.59	2.87	21.0	118.0
2.80					

	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity
hue \				
0	3.06	0.28	2.29	5.64
1.04				
1	2.76	0.26	1.28	4.38
1.05				
2	3.24	0.30	2.81	5.68
1.03				
3	3.49	0.24	2.18	7.80
0.86				
4	2.69	0.39	1.82	4.32
1.04				

	od280/od315_of_diluted_wines	proline	target
0	3.92	1065.0	0
1	3.40	1050.0	0
2	3.17	1185.0	0
3	3.45	1480.0	0
4	2.93	735.0	0

Шаг 2: Проверка на пропущенные значения

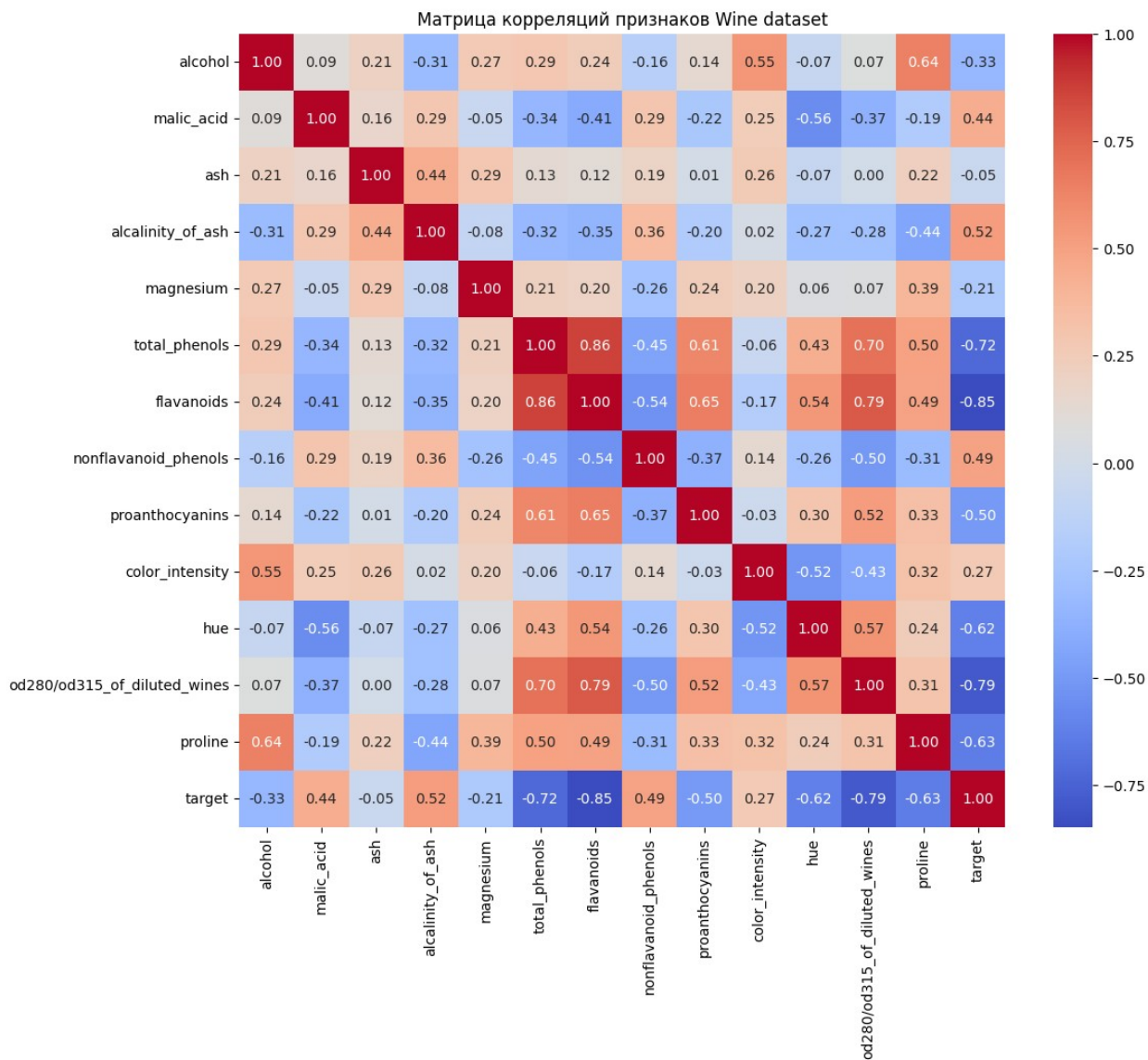
```
# Проверка на пропущенные значения
print("\nКоличество пропущенных значений по колонкам:")
print(wine_df.isnull().sum())
```

```
Количество пропущенных значений по колонкам:
alcohol      0
malic_acid   0
ash          0
alcalinity_of_ash  0
magnesium    0
total_phenols 0
flavanoids   0
nonflavanoid_phenols 0
proanthocyanins 0
color_intensity 0
hue          0
od280/od315_of_diluted_wines 0
proline      0
target       0
dtype: int64
```

1. Корреляционный анализ

```
# Построение матрицы корреляций
plt.figure(figsize=(12, 10))
correlation_matrix = wine_df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
            fmt=".2f")
plt.title("Матрица корреляций признаков Wine dataset")
plt.show()

# Вывод наиболее коррелированных признаков
high_corr =
correlation_matrix.abs().unstack().sort_values(ascending=False)
high_corr = high_corr[high_corr != 1].drop_duplicates()
print("\nНаиболее коррелированные пары признаков:")
print(high_corr.head(10))
```



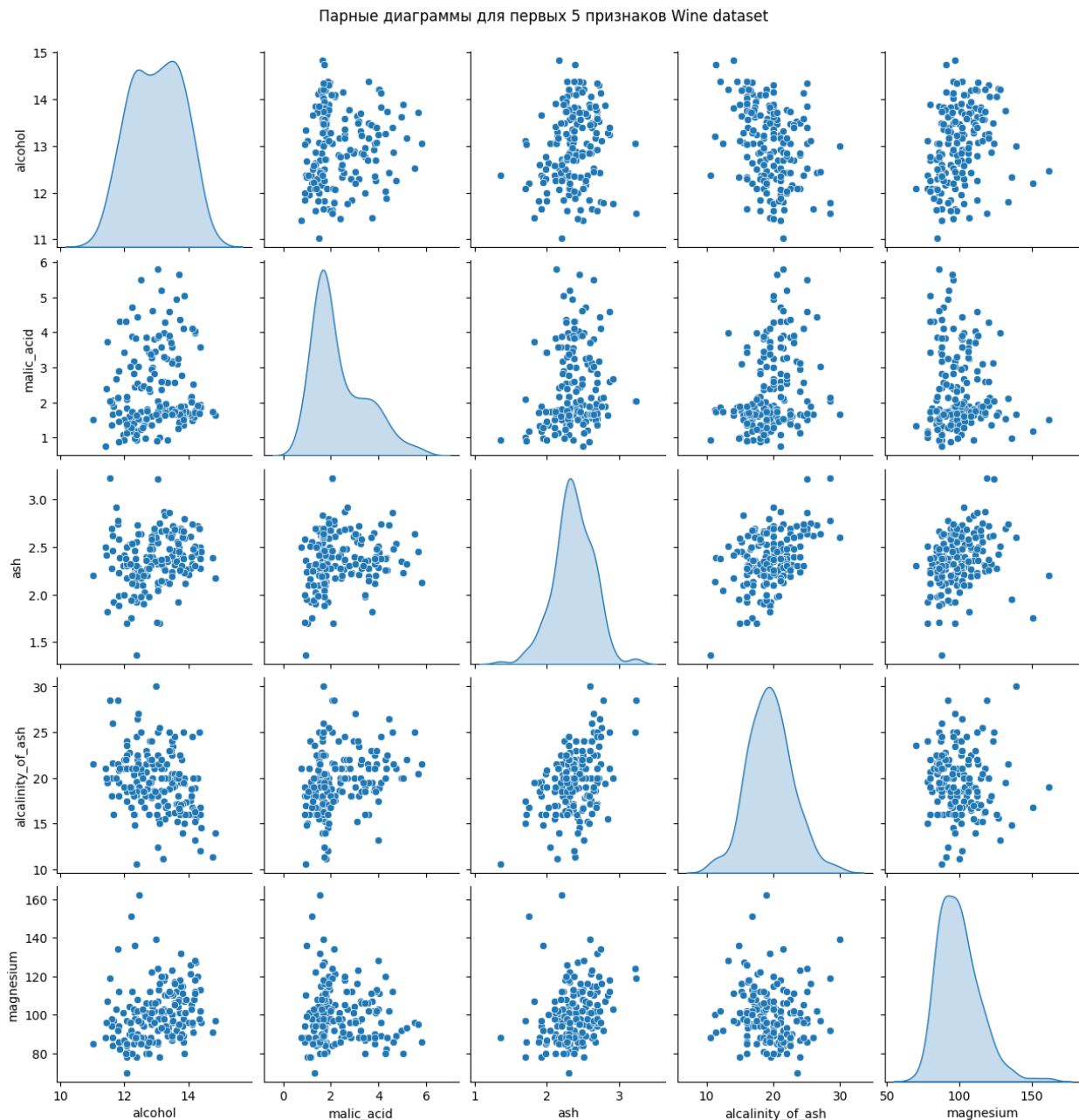
```

Наиболее коррелированные пары признаков:
flavanoids      total_phenols      0.864564
target          flavanoids        0.847498
od280/od315_of_diluted_wines target      0.788230
flavanoids      od280/od315_of_diluted_wines 0.787194
total_phenols   target          0.719163
                od280/od315_of_diluted_wines 0.699949
flavanoids      proanthocyanins 0.652692
proline         alcohol        0.643720
                target          0.633717
hue             target          0.617369
dtype: float64

```

1. Парные диаграммы (pairplot)

```
# Построение парных диаграмм для первых 5 признаков (чтобы не
перегружать визуализацию)
sns.pairplot(wine_df.iloc[:, :5], diag_kind='kde')
plt.suptitle("Парные диаграммы для первых 5 признаков Wine dataset",
y=1.02)
plt.show()
```



Качество данных: В наборе данных отсутствуют пропущенные значения, что позволяет использовать все имеющиеся наблюдения для анализа и построения моделей.

Корреляционный анализ:

Наибольшая положительная корреляция наблюдается между `total_phenols` и `flavanoids` (0.86), что логично, так как флавоноиды являются подклассом фенолов.

Сильная положительная корреляция также между `flavanoids` и `od280/od315_of_diluted_wines` (0.79).

Отрицательные корреляции наиболее выражены между `hue` и `color_intensity` (-0.52).

Парные диаграммы:

Позволяют визуально оценить распределения признаков и их взаимосвязи.

Видно, что некоторые признаки имеют близкое к нормальному распределение (например, `alcohol`), в то время как другие (`malic_acid`, `ash`) имеют более сложные распределения.

На точечных диаграммах можно заметить как линейные, так и нелинейные зависимости между признаками.

Рекомендации для построения моделей:

При построении моделей можно рассмотреть удаление одного из сильно коррелированных признаков (например, оставить только `flavanoids` вместо пары `flavanoids` и `total_phenols`), чтобы избежать мультиколлинеарности.

Набор данных хорошо подходит для задач классификации (предсказание класса вина), так как признаки имеют выраженные различия между классами.

Наибольший вклад в модель, вероятно, внесут такие признаки как `flavanoids`, `total_phenols`, `od280/od315_of_diluted_wines`, `proline` и `color_intensity`, так как они демонстрируют значимые корреляции с другими признаками и, вероятно, с целевой переменной.