In [40]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Red Wine

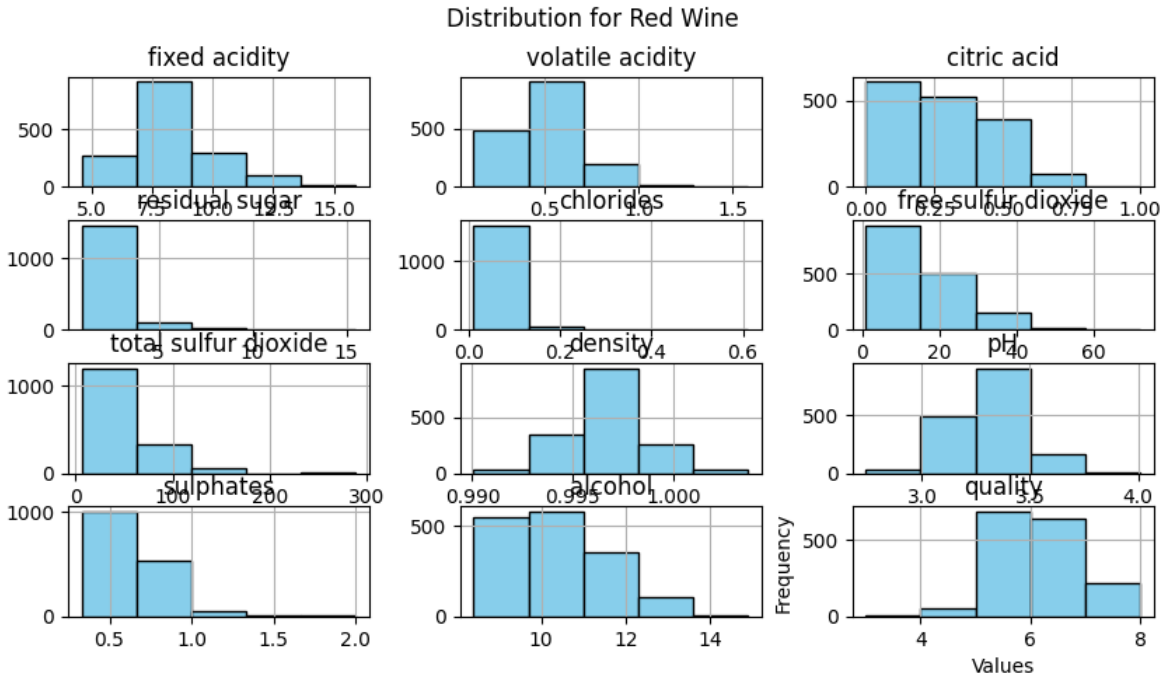
In [31]:
```
df_red = pd.read_csv('winequality-red.csv', sep = ';')
df_red
```

Out[31]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

1599 rows × 12 columns

In [50]:
```
#Histogram plots
df_red.hist(bins=5, color = 'skyblue', edgecolor = 'black', figsize =(10,5))
plt.suptitle('Distribution for Red Wine')
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.show()
```



In [52]:
```
df_red.describe()
```

Out[52]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	159
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	1
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	1
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	1
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	1

White Wine

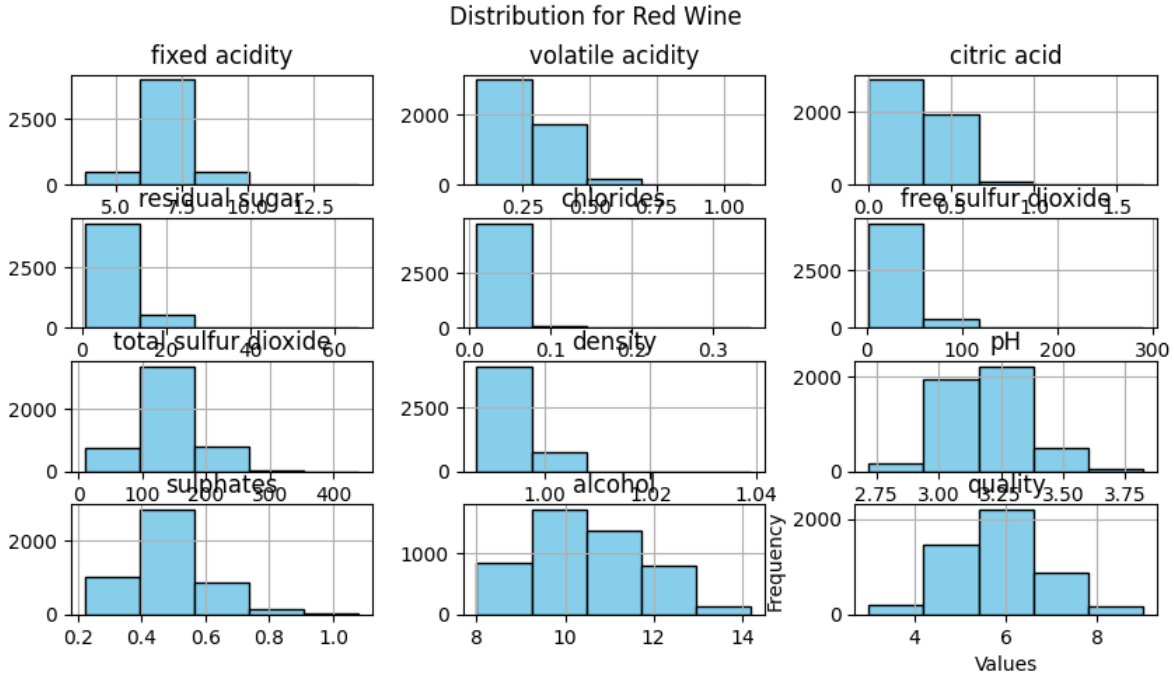
In [30]:
```
df_white = pd.read_csv('winequality-white.csv', sep = ';')
df_white
```

Out[30]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6

4898 rows × 12 columns

In [63]:
```
df_white.hist(bins=5, color = 'skyblue', edgecolor = 'black', figsize =(10,5))
plt.suptitle('Distribution for Red Wine')
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.show()
```



In [64]:
```
df_white.describe()
```

Out[64]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	
count	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	4898.000000	489
mean	6.854788	0.278241	0.334192	6.391415	0.045772	35.308085	138.360657	0.994027	3.188267	0.489847	1
std	0.843868	0.100795	0.121020	5.072058	0.021848	17.007137	42.498065	0.002991	0.151001	0.114126	
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991723	3.090000	0.410000	
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993740	3.180000	0.470000	1
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996100	3.280000	0.550000	1
max	14.200000	1.100000	1.660000	65.800000	0.346000	289.000000	440.000000	1.038980	3.820000	1.080000	1

Hypotheses

1. The wine quality of White Wine is higher than Red Wine
2. There is a correlation with residual sugar and wine quality

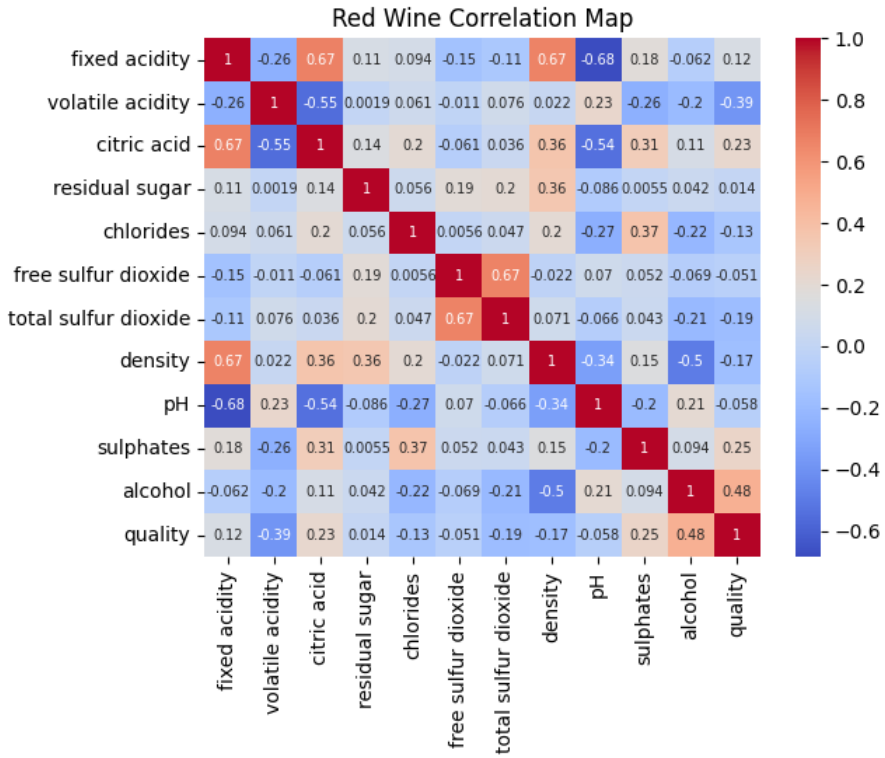
In [72]:
```
#Hypothesis One
white_q = df_white['quality'].mean()
red_q = df_red['quality'].mean()

print(f"We fail to reject the null hypothesis that the wine quality of White Wine is higher\nthan Red Wine because t
```

```
We fail to reject the null hypothesis that the wine quality of White Wine is higher
than Red Wine because the White Wine quality average is 5.87790935075541
while the Red Wine quality average is 5.6360225140712945
```

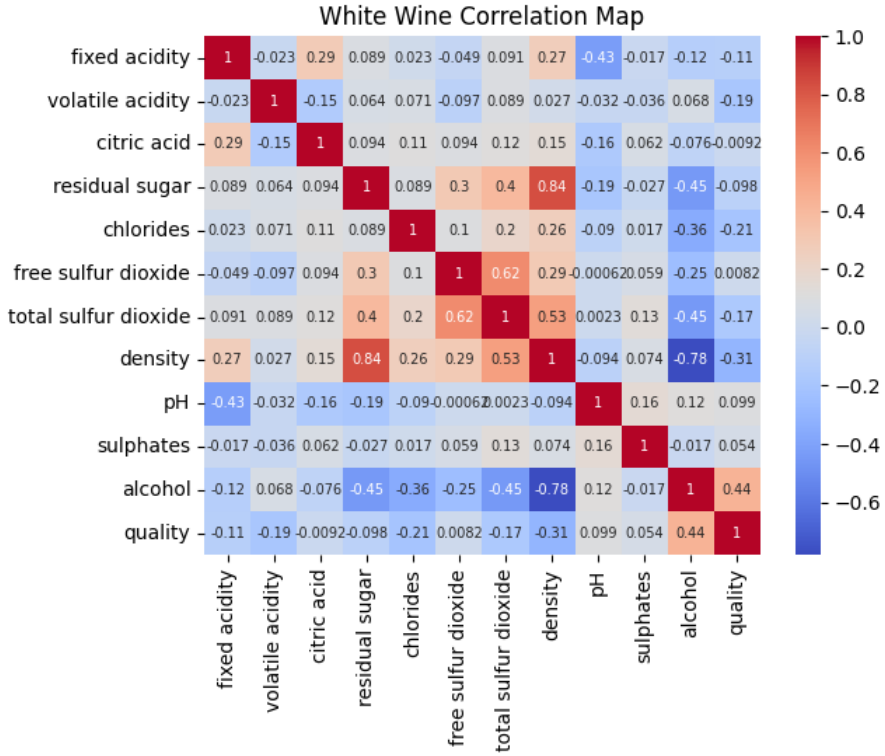
In [102]:

```
#Hypothesis Two
sns.heatmap(df_red.corr(), annot=True, cmap = 'coolwarm', annot_kws={"size": 7})
plt.title("Red Wine Correlation Map")
plt.show()
```



In [103]:

```
sns.heatmap(df_white.corr(), annot=True, cmap = 'coolwarm', annot_kws={"size": 7})
plt.title("White Wine Correlation Map")
plt.show()
```



In [105]:
```
import scipy.stats as stats

# Calculate the Pearson correlation coefficient and p-value
r, p_value = stats.pearsonr(df_red['quality'], df_red['alcohol'])

# Print the results
print(f"Pearson correlation coefficient: {r}")
print(f"P-value: {p_value}")

# Check for significance (using a common significance level of 0.05)
if p_value < 0.05:
    print("The correlation is statistically significant.")
else:
    print("The correlation is not statistically significant.")
```

```
Pearson correlation coefficient: 0.47616632400113584
P-value: 2.8314769747789036e-91
The correlation is statistically significant.
```

In [106]:
```
# Calculate the Pearson correlation coefficient and p-value
r, p_value = stats.pearsonr(df_white['quality'], df_white['alcohol'])

# Print the results
print(f"Pearson correlation coefficient: {r}")
print(f"P-value: {p_value}")

# Check for significance (using a common significance level of 0.05)
if p_value < 0.05:
    print("The correlation is statistically significant.")
else:
    print("The correlation is not statistically significant.")
```

```
Pearson correlation coefficient: 0.43557471546137627
P-value: 5.614770253715724e-226
The correlation is statistically significant.
```

Report

According to the averages for wine quality, We fail to reject the first null hypothesis that the wine quality of White Wine is higher than Red Wine because the White Wine quality average is 5.87790935075541 while the Red Wine quality average is 5.6360225140712945.

Abiding the Heatmaps, there is a positive correlation between the alcohol content and the quality of the wine. After calculating the correlation coefficient and the p-value to check if this statistic was significant at a level of 0.05, the statistics show that the correlation is statistically significant between the alcohol content and quality of wine in Red Wine and White Wine.