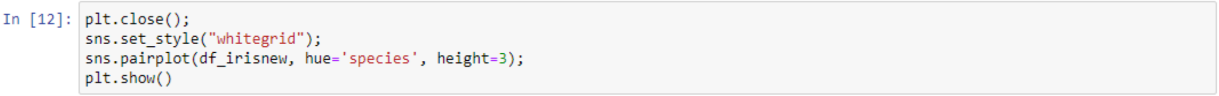
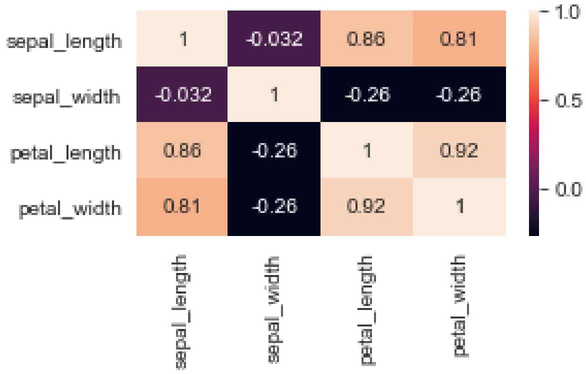
[CM3] Calculate and report the correlation coefficient for all pairs of features. To what extent are the features correlated? Calculate and report the mean, variance, skew, and kurtosis for all the variables in datasets for which is makes sense. Do you find any interesting or significant relationships? Briefly explain your observations about the nature of data and the relationships between the features of the dataset. Do you chosen subset of features from Heart Disease have any particular pattern on any of these statistics?

IRIS DATASET



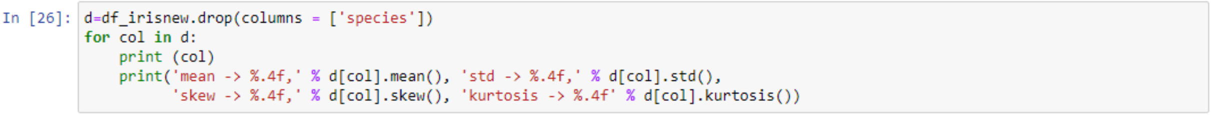
```
In [12]: plt.close();
         sns.set_style("whitegrid");
         sns.pairplot(df_irisnew, hue='species', height=3);
         plt.show()
```



	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1	-0.032	0.86	0.81
sepal_width	-0.032	1	-0.26	-0.26
petal_length	0.86	-0.26	1	0.92
petal_width	0.81	-0.26	0.92	1

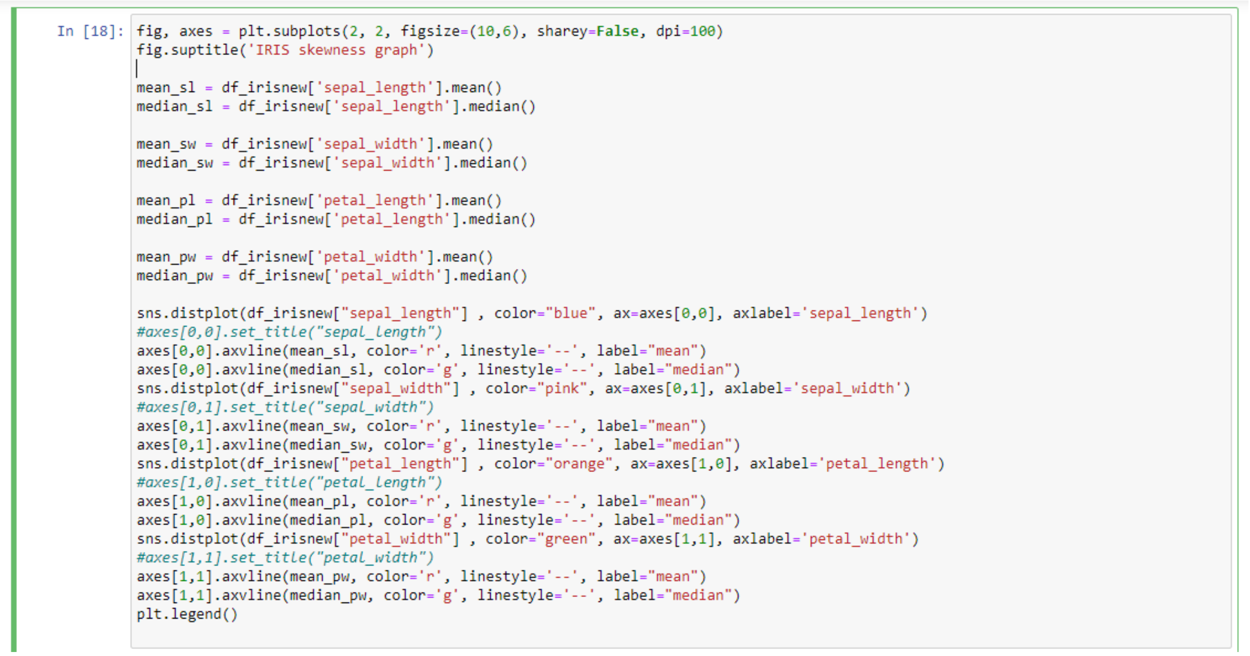
- There is high correlation i.e. (0.92) between petal_width & petal_length.
- Petal_width & sepal_width have low correlation i.e. (-0.26).
- petal_length & sepal_length are the next features having better correlation i.e. (0.86).
- Sepal_width have low correlation with all the other features.

Mean, variance, skew and kurtosis

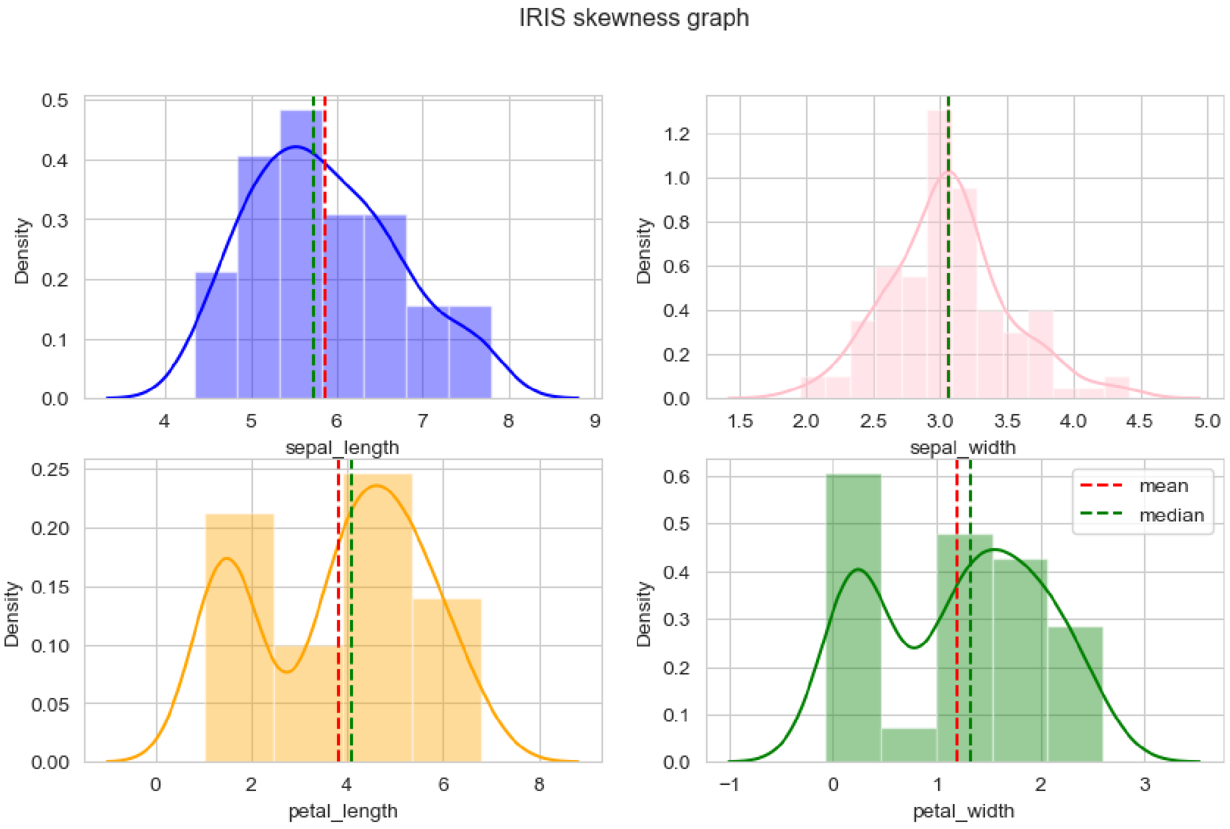


```
In [26]: d=df_irisnew.drop(columns = ['species'])
         for col in d:
             print (col)
             print('mean -> %.4f,' % d[col].mean(), 'std -> %.4f,' % d[col].std(),
                   'skew -> %.4f,' % d[col].skew(), 'kurtosis -> %.4f' % d[col].kurtosis())
```

sepal_length
mean -> 5.8589, std -> 0.8616, skew -> 0.4015, kurtosis -> -0.5448
sepal_width
mean -> 3.0591, std -> 0.4463, skew -> 0.3747, kurtosis -> 0.6496
petal_length
mean -> 3.8124, std -> 1.7231, skew -> -0.2658, kurtosis -> -1.2489
petal_width
mean -> 1.1997, std -> 0.7872, skew -> -0.0748, kurtosis -> -1.3155
We are replacing missing value for Iris dataset with mean as data is symmetric (i.e. skewness <0.5) for all features

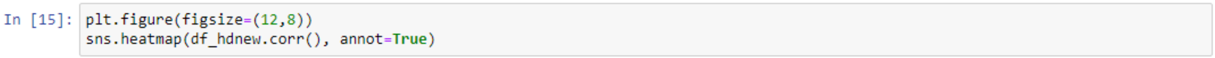


In [18]:
```
fig, axes = plt.subplots(2, 2, figsize=(10,6), sharey=False, dpi=100)
fig.suptitle('IRIS skewness graph')

mean_sl = df_irisnew['sepal_length'].mean()
median_sl = df_irisnew['sepal_length'].median()

mean_sw = df_irisnew['sepal_width'].mean()
median_sw = df_irisnew['sepal_width'].median()

mean_pl = df_irisnew['petal_length'].mean()
median_pl = df_irisnew['petal_length'].median()

mean_pw = df_irisnew['petal_width'].mean()
median_pw = df_irisnew['petal_width'].median()

sns.distplot(df_irisnew["sepal_length"] , color="blue", ax=axes[0,0], axlabel='sepal_length')
#axes[0,0].set_title("sepal_length")
axes[0,0].axvline(mean_sl, color='r', linestyle='--', label="mean")
axes[0,0].axvline(median_sl, color='g', linestyle='--', label="median")
sns.distplot(df_irisnew["sepal_width"] , color="pink", ax=axes[0,1], axlabel='sepal_width')
#axes[0,1].set_title("sepal_width")
axes[0,1].axvline(mean_sw, color='r', linestyle='--', label="mean")
axes[0,1].axvline(median_sw, color='g', linestyle='--', label="median")
sns.distplot(df_irisnew["petal_length"] , color="orange", ax=axes[1,0], axlabel='petal_length')
#axes[1,0].set_title("petal_length")
axes[1,0].axvline(mean_pl, color='r', linestyle='--', label="mean")
axes[1,0].axvline(median_pl, color='g', linestyle='--', label="median")
sns.distplot(df_irisnew["petal_width"] , color="green", ax=axes[1,1], axlabel='petal_width')
#axes[1,1].set_title("petal_width")
axes[1,1].axvline(mean_pw, color='r', linestyle='--', label="mean")
axes[1,1].axvline(median_pw, color='g', linestyle='--', label="median")
plt.legend()
```

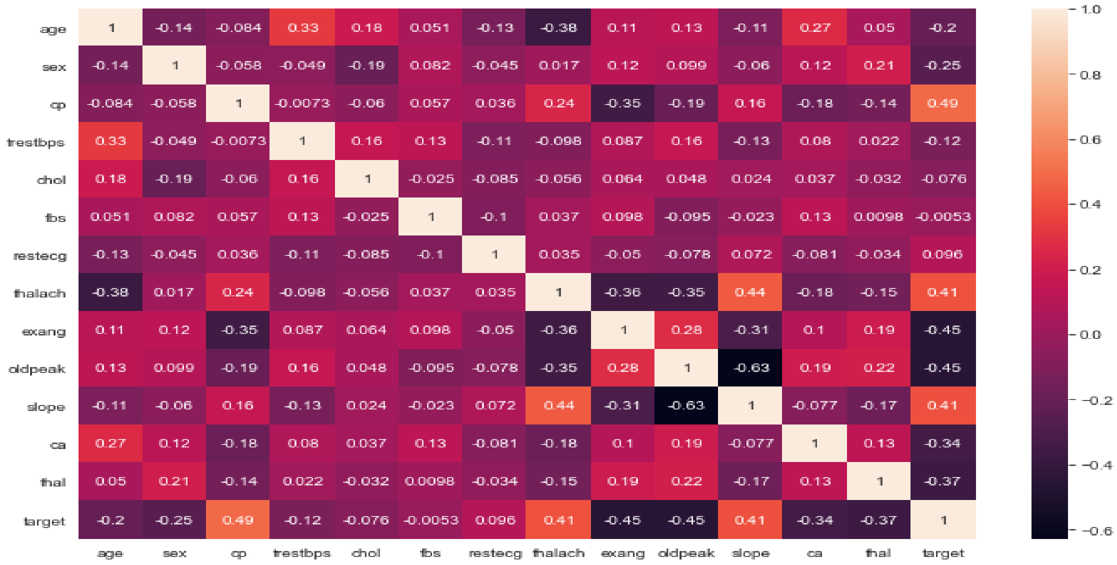


- Since the skewness for all the feature is quite less, we can use mean for replacing the missing values.
- sepal_width follows bell curve and also the mean and median much closer to each other.
- Also as seen in above graph, petal_length and petal_width denotes multimodal distribution.

HEART DISEASE DATASET



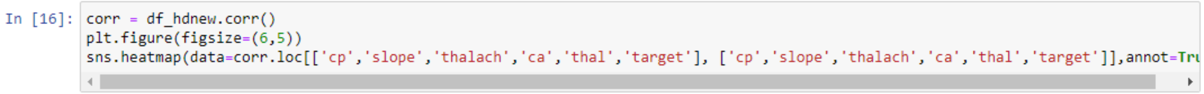
```
In [15]: plt.figure(figsize=(12,8))
         sns.heatmap(df_hdnew.corr(), annot=True)
```



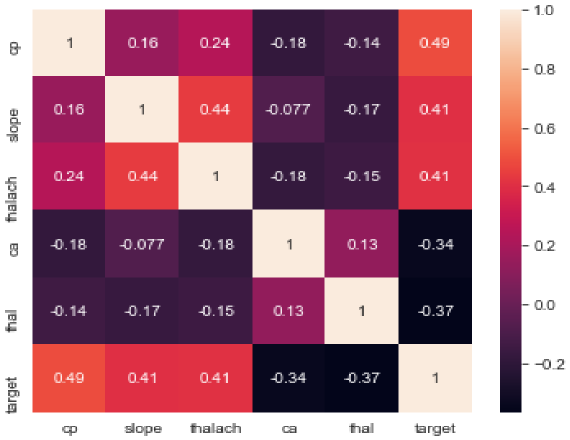
By exploring the above correlation coefficient we select some features i.e. cp, slope, thalach, thal and ca

Reasons for selecting these features are as follows:

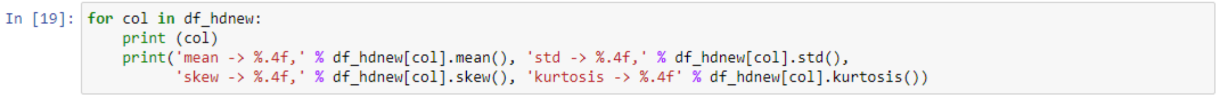
- cp has the maximum correlation with the target i.e. (0.49), so it is selected.
- thalach is selected as it the next numerical variable which has high correlation with target
- slope and oldpeak has high inter-feature correlation i.e. (-0.61), so one of this is dropped.
- oldpeak had higher number of missing values so had an another reason to drop it.
- ca and thal has least relative correlation with other features so both are selected.



```
In [16]: corr = df_hdnew.corr()
         plt.figure(figsize=(6,5))
         sns.heatmap(data=corr.loc[['cp','slope','thalach','ca','thal','target'], ['cp','slope','thalach','ca','thal','target']],annot=Tru
```

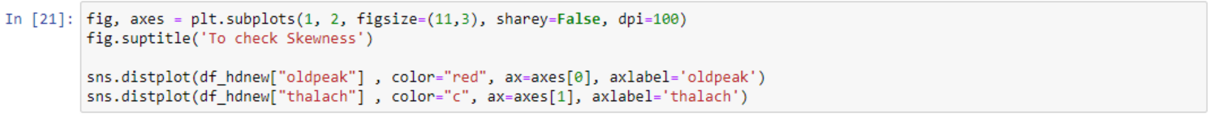


Mean, variance, skew and kurtosis

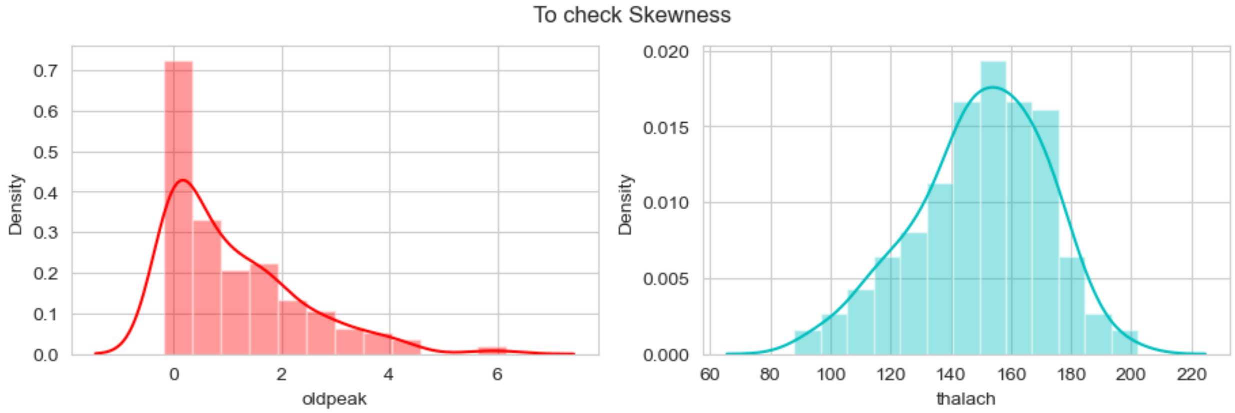


```
In [19]: for col in df_hdnew:
             print (col)
             print('mean -> %.4f,' % df_hdnew[col].mean(), 'std -> %.4f,' % df_hdnew[col].std(),
                   'skew -> %.4f,' % df_hdnew[col].skew(), 'kurtosis -> %.4f' % df_hdnew[col].kurtosis())
```

```
age
mean -> 54.3113, std -> 9.1453, skew -> -0.1060, kurtosis -> -0.5616
sex
mean -> 0.6887, std -> 0.4641, skew -> -0.8208, kurtosis -> -1.3390
cp
mean -> 0.9575, std -> 1.0225, skew -> 0.4614, kurtosis -> -1.2407
trestbps
mean -> 131.7846, std -> 17.7552, skew -> 0.6839, kurtosis -> 0.7265
chol
mean -> 244.1333, std -> 45.3303, skew -> 0.3417, kurtosis -> 0.4159
fbs
mean -> 0.1321, std -> 0.3394, skew -> 2.1889, kurtosis -> 2.8178
restecg
mean -> 0.5708, std -> 0.5330, skew -> 0.0933, kurtosis -> -1.1936
thalach
mean -> 149.6480, std -> 21.8660, skew -> -0.3978, kurtosis -> -0.1602
exang
mean -> 0.3443, std -> 0.4763, skew -> 0.6599, kurtosis -> -1.5795
oldpeak
mean -> 1.0912, std -> 1.2230, skew -> 1.3020, kurtosis -> 1.6673
slope
mean -> 1.4292, std -> 0.6232, skew -> -0.6185, kurtosis -> -0.5557
ca
mean -> 0.7311, std -> 1.0388, skew -> 1.3778, kurtosis -> 1.0203
thal
mean -> 2.3491, std -> 0.6007, skew -> -0.2507, kurtosis -> -0.6354
target
mean -> 0.5425, std -> 0.4994, skew -> -0.1716, kurtosis -> -1.9894
```



```
In [21]: fig, axes = plt.subplots(1, 2, figsize=(11,3), sharey=False, dpi=100)
         fig.suptitle('To check Skewness')

         sns.distplot(df_hdnew["oldpeak"] , color="red", ax=axes[0], axlabel='oldpeak')
         sns.distplot(df_hdnew["thalach"] , color="c", ax=axes[1], axlabel='thalach')
```



Numeric variable oldpeak having skewness as 1.2241 which suggests that the distribution is right skewed and kurtosis value of 1.3632 which is greater than zero suggesting that the distribution is leptokurtic i.e. heavier tails. Since it right skewed, we can replace the missing values with mode rather than mean.

Numeric variable thalach having skewness as -0.3978 which suggests that the distribution is slightly left skewed and kurtosis value of -0.1602 which is less than zero suggesting that the distribution is platykurtic i.e. light tails. Since the skewness is not much, we can replace missing values with mean for this feature.