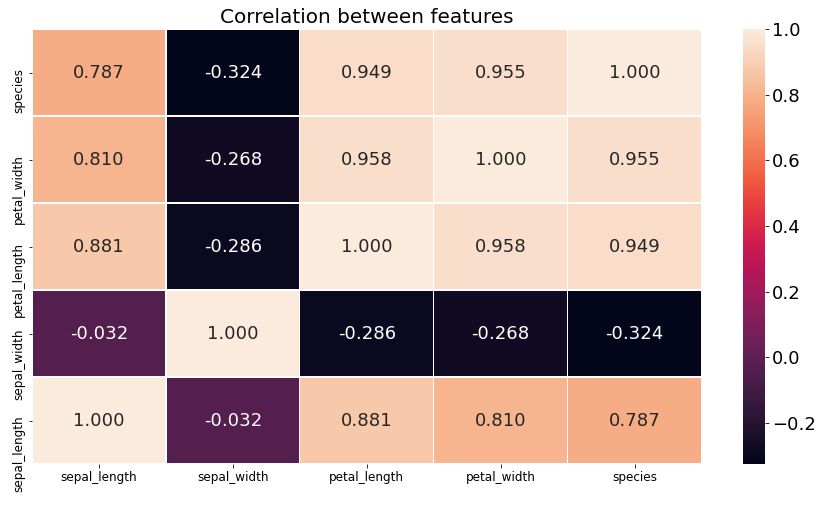
**Q 1.3**

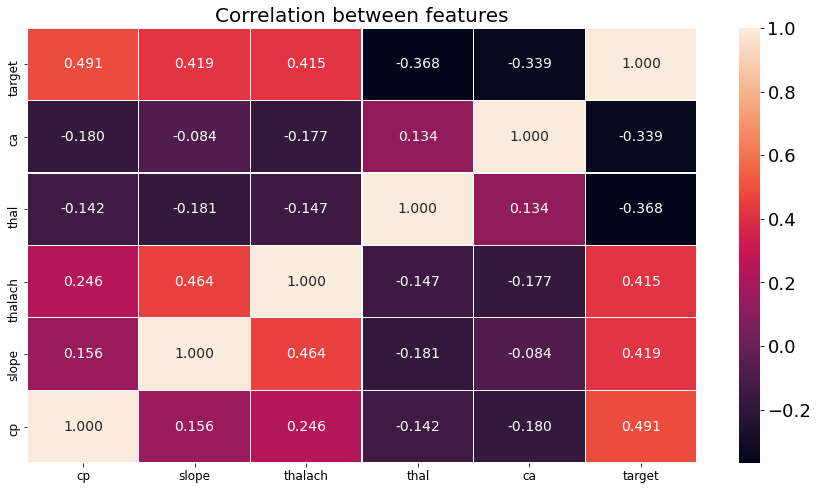
**Iris Dataset Correlation**



**Observations from correlation heatmap**

* Petal width has the highest correlation with the target (0.955)
* Sepal width has the lowest correlation with the target (-0.324)
* Sepal width has a negative correlation with the target
* Petal width, petal length and sepal length have positive correlation with the target
* The highest inter-feature correlation is 0.958 and it exists between petal width and petal length.
* The lowest inter-feature correlation is -0.032 and it exists between sepal width and sepal length.
* Sepal width has a negative correlation with all other features.

**Heart Dataset Correlation**



**Observations from correlation heatmap**

* Cp has the highest correlation with the target (0.491).
* Ca has the lowest correlation with the target (-0.339).
* Ca and thal have a negative correlation with the target.
* Cp, slope, and thalach have positive correlation with the target.
* The highest inter-feature correlation is 0.464 and it exists between slope and thalach.
* The lowest inter-feature correlation is -0.084 and it exists between slope and ca.
* ca has a negative correlation with all other features except thal.

**Q 1.4**

**Iris Dataset Descriptive Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Feature | Mean | Variance | Skew | Kurtosis |
| **Iris Versicolor** | Sepal Length | 5.95 | 0.25 | 0.36 | -0.29 |
| Sepal Width | 2.72 | 0.09 | -0.47 | -0.22 |
| Petal Width | 1.3 | 0.04 | 0.15 | 0.19 |
| Petal Length | 4.26 | 0.23 | -0.51 | 0.03 |
| **Iris Setosa** | Sepal Length | 4.99 | 0.11 | 0.01 | -0.25 |
| Sepal Width | 3.39 | 0.18 | 0.17 | 0.87 |
| Petal Width | 0.24 | 0.01 | 0.47 | -0.8 |
| Petal Length | 1.49 | 0.02 | 0.53 | 2.16 |
| **Iris Virginica** | Sepal Length | 6.68 | 0.44 | 0.23 | -0.97 |
| Sepal Width | 3.0 | 0.12 | 0.32 | 0.52 |
| Petal Width | 2.08 | 0.07 | -0.12 | -0.51 |
| Petal Length | 5.62 | 0.3 | 0.56 | -0.33 |

**Comments about statistical values**

Who has the highest mean?

* Across all the species, Sepal length has the highest mean value while Petal width has the lowest mean value.
* **Iris-Virginica** has the highest mean value of all features. Thus, we can say that Iris-virginica is on average larger than the other species.

What do the variance values say about the features?

* Iris-Virginica has the highest variation in sepal length followed by Iris-versicolor.
* Iris-Setosa has the smallest variation in petal width, petal length and sepal length but has the highest variation in sepal width.

What do positive and negative values of skew and kurtosis say about the dataset?

The skewness measures the degree of symmetry of the distribution of the data.

* For Iris-Setosa, all its features are skewed to the right.
* For Iris-Versicolor, sepal length and petal width are skewed to the right while sepal width and petal length are left skewed.
* For Iris-Virginica, all features are skewed to the right except the petal width.
* The sepal length for the Iris-Setosa specie has the smallest skew (0.01) making it the closest to a mirrored distribution curve.

Kurtosis on the other hand indicates how the tails of the data distribution compares to that of the normal distribution.

* Petal length for Iris-Setosa has the highest positive kurtosis meaning that its curve doesn’t flatten out on the tails. This further indicates that its range of values are small.
* Sepal length for Iris-Virginica has the highest negative kurtosis meaning that the curve flattens out at the tails more than a normal distribution curve (data is light tailed).

**Heart Disease Dataset Descriptive Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Target** | **Feature** | **Mean** | **Mode** | **Variance** | **Skew** | **Kurtosis** |
| 0 (No Heart Disease) | cp | 0.41 | 0 | 0.7 | 1.8 | 1.84 |
|  | slope | 1.14 | 1 | 0.33 | **-0.0** | -0.07 |
|  | thalach | 139.77 | - | 466.98 | -0.17 | -0.28 |
|  | thal | 2.59 | - | 0.46 | -1.33 | -0.61 |
|  | ca | 1.11 | 0 | 1.16 | 0.63 | -0.65 |
| 1 (Heart Disease) | cp | 1.42 | 2 | 0.88 | -0.21 | -0.96 |
|  | slope | 1.66 | 2 | 0.31 | -1.43 | 1.11 |
|  | thalach | 158.12 | - | 352.4 | -0.51 | 0.16 |
|  | thal | 2.15 | - | 0.2 | 0.91 | 1.26 |
|  | ca | 0.41 | 0 | 0.79 | 2.62 | 6.82 |

**Comments about statistical values**

What does the mean and mode?

* for numerical feature thalach we shall compare using the mean. People with heart disease have a higher mean value than those without heart disease.
* For categorical features, comparison between target values will be done using the mode.
  + For ca, people with and without heart disease both have a mode of 0.
  + Most people with heart disease have a flat slope of peak exercise while most people without heart disease have an upsloping slop of peak exercise.
  + Majority of people with heart disease have non-angina chest pain while asymptomatic angina chest pain is most common among those without heart disease.
* Thal is supposed to be a categorical feature but due to noise, the mode is neither definitive nor informative as it now seems like a continuous feature. Looking at the mean, people without heart disease have a slightly higher mean value than those with heart disease.

What do the variance values say about the features?

* For thalach feature; considering the lower mean for those without heart disease, they have a higher variance compared to those with heart disease.
* Generally, the variance for those without heart disease is higher than that of those with heart disease for all the features except cp.

What do positive and negative values of skew and kurtosis say about the dataset?

The skewness measures the degree of symmetry of the distribution of the data.

* The ca values for people with heart disease has the highest skew and highest kurtosis in the entire dataset.
* For those without heart disease, all features except cp have a negative kurtosis (i.e., flatten out more than the normal distribution curve on the tails).
* For those with heart disease, all features except cp have a positive kurtosis (i.e., distribution curve doesn’t flatten out as much as the normal distribution curve on the tails).
* For those without heart disease, all features except cp and ca are skewed to the left.
* For those with heart disease, all features except ca and thal are skewed to the right.