CS422-HW1-REPORT

SHIQI LIU

A20430674

All the comments in this report are recorded on ipynb markdown.

2.1

From the graphic, we know that red color represents 'died' status in 'survived' column. And blue color represents 'survived' status in 'survived' column.

According to these two graphics, we can clearly see that the number of females survived are much more lager than the number of male survived.

In the first graph, the only two parts show blue higher than red. That implies the children which between 0-18 and the old which older than 70 survived numbers larger than died numbers.

#therefore, I print the table to show data of died/survived in different gender.

As a result, we observe that the most of female were survived, but most of adult male were died. and the most of children and older were survived. It shows male made female and children first to go and let their have a chance to live.

Therefore, we got a relationship: most of adult male were died, most of female were lived, many of boys and the old were lived

The result make sense.

Because in that time, there was a rule called 'women and children first' in people mind.

Honor and shame ruled society in 1912. Gentlemen were expected to be honorable and feared being shamed. Men aboard the Titanic, especially those among the upper class, had to weigh their own survival with doing what society would deem the right thing. Famously several of the wealthiest men, John Jacob Astor, George Widener, and Isidor Strauss, refused to board a lifeboat on grounds of honor, even after it was clear that the Titanic would sink.

Therefore, the data make sense.

2.2

According to the data, we observe that the mean's variance is the lowest.

Because mean calculate the average of the data, and one of mean's important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set. Variance measures how far the set of values are spread out from their average value. When we replace the missing values with the mean/ average value itself the whole set of data is centralized and moved towards the mean. So the variance is lesser compared to replacing with median or mode.

No. Using the mean is the optimum way to minimize variance while handling missing values using imputation

2.3

From the plot, it can be observed that petal width and length are highly correlated and their variability can be primarily attributed to the first component. Likewise, the first component also explains a large part of the Sepal length. The variability in Sepal width is more attributed to the second component.

pca.explained\_variance\_ratio\_: array([0.72770452, 0.23030523, 0.03683832, 0.00515193])

pca.explained\_variance\_: array([2.93035378, 0.92740362, 0.14834223, 0.02074601])

14.994542918291165 4.154401626781096 68.14579712864128 12.705258326286472

We observe that most of the variance (72.77% of the variance to be precise) can be explained by the first principal component alone. The second principal component still bears some information (23.03%) while the third and fourth principal components can safely be dropped without losing to much information. Together, the first two principal components contain 95.8% of the information

PC1 and PC2 are covering maximum detail. The percentage of variance covered by PC1 and PC2 is around 95% with PC1 covering 72.77% and PC2 covering 23.03%When calculating the percentage of variance of each of the 4 features of dataset Percentage of Variance of sepal length = 15%Percentage of Variance of sepal width = 4.11%Percentage of Variance of petal length= 68.13%Percentage of Variance of petal width = 12.74%We find the number of dimensions where variance is spread is 3 in the case of feature variance. And using PCA we can clearly conclude that the number of dimensions have been reduced to 2 covering 95% of variance.

2.4

The scatter plots shown above show that the third feature (petal length in cm) has the closest relationship with PC1. Looking back at Problem 3, the percentage of variance for PC1 was 92% while the highest percentage of variance among the original features was 68%, which resulted in being the third feature.

Since the correlation coefficient between PC1 and third feature is almost close to 1, then I can agree with the visual inspection above.

2.5

total\_var= 4.572957

total\_pca\_var= 4.0268456399999994

The above demonstrates that the first two components are capable of capturing at least 95% variance of the original data. it shows that PCA was capable of capturing the variance of the original features by reducing the original features to two.

The above demonstrates that the first two components are capable of capturing at least 95% variance of the original data. it shows that PCA was capable of capturing the variance of the original features by reducing the original features to two.