1. Report the mean, standard deviation, mode, median, and the five-number summary of the attributes RH and wind.

Text

Description automatically generated

1. For the attribute RH, use seaborn or matplotlib to show boxplots for

(a) RH;  
(b) RH for various values of month (show month in x-axis);

(c) RH for various values of day (show day in x-axis).

Chart, box and whisker chart

Description automatically generated

1. For attributes wind and RH, use seaborn or matplotlib to show their histograms (use 8 equal-sized bins). Chart, histogram

   Description automatically generated
2. Use seaborn or matplotlib to show the scatter plot for attributes wind and area (show area in x-axis). Report the correlation coefficient between wind and area.

Chart, scatter chart

Description automatically generated

1. Consider the 6 attributes {FFMC, DMC, DC, ISI, RH, wind}, report the top 3 attributes that are most correlated (either positively or negatively) with area.

Text

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

*Remark: I have sorted the value through the notebook, so the values in the above figure have been expressed as absolute values.*

1. χ2-test (using the χ2 table shown on the last page). Show your steps clearly (including contingency table, χ2-value, and p-value) in the report.

(a) By performing the χ2-test at a significance level of 0.01, are the attributes temp and wind independent of each other? Graphical user interface, text

Description automatically generated

(b) By performing the χ2-test at a significance level of 0.01, are the attributes X and Y independent of each other? Graphical user interface, text, application, email

Description automatically generated

8. PCA: In this question, use only the attributes {DC, ISI, temp, RH, wind} after the normalization in Task 7b.

1. (a)  Plot the cumulative explained variance ratio with the number of principal components. Chart, line chart

   Description automatically generated
2. (b)  Transform the data by PCA, by using the smallest number of PCA components such that the proportion of explained variance is at least 0.9. You can use the built-in PCA function from scikit-learn.
3. (c)  For the transformed data, name the new features PC1, PC2, and so on, and output the transformed dimensions to the csv file data reduced.csv without the header (use “,” as field delimiter). Report the five-number summary for each obtained dimension.

Table

Description automatically generated

9. Consider the original data. At first glance, this dataset has no missing value. However, for attribute area , some values got zero. We assume that these zeros can be viewed as missing values. Please fill in the missing values with the attribute mean of transactions whose corresponding attribute has non-zero value (filter non-zero value transactions first, then calculate attribute mean of them). and show the box plot for attribute area again. You can use the “DataFrame.replace” function from pandas. After filling in the missing values, report the correlation coefficient between wind and area again, and compare it with the value obtained in Task 4.

Chart, box and whisker chart

Description automatically generated

Text

Description automatically generated with medium confidence

Compared with the original correlation coefficient done in Task 4, the correlation got weaker as I filled the missing value with the mean. It shows that filling missing values with mean is not a good option in this situation