

Introduction to Data Mining and Programming in R

Instructions

Download `forestfires.tsv` data and documentation, then use `dplyr` verbs and `ggplot2/ggpairs` visualizations to conduct an exploratory data analysis on the forest fires data set. Start by answering the following questions, and then pursue what interests you on your own. Submit a document with your answers, brief explanations, and any relevant figures.

1. We hypothesize that temp and wind are correlated with burn area. Test that hypothesis with `ggpairs`.
 - A. What is the correlation between temp and burn area?
 - B. What is the correlation between wind and burn area?
 - C. Is there evidence to support the hypothesis that temp and wind are correlated with burn area?
2. Make some hypotheses about three additional variables that could be positively and negatively correlated with `burn_area`. Test your hypotheses using `ggpairs`.
3. Does burn area seems to be related to the season of the year?
 - A. Use `dplyr` verbs to aggregate by month and compute the mean burn area. Do you see evidence of a seasonal relationship?
 - B. Which month has the largest average burn area?
4. Are there limitations in your data that could invalidate your findings?