|  |  |
| --- | --- |
|  | **DEPARTMENT OF COMPUTER ENGINEERING** |

**Collaborative Project Report**

|  |  |
| --- | --- |
| Semester | B.E. Semester VII – Computer Engineering |
| Subject | Big Data Analytics |
| Subject Professor In-charge | Prof. Pankaj Vanvari |
| Assisting Teachers | Dr. Umesh Kulkarni |

|  |  |
| --- | --- |
| Roll Number | Name of Students |
| 21102A0003 | Omkar Patil |
| 21102A0005 | Pranav Redij |
| 21102A0006 | Sahil Pokharkar |
| 21102A0014 | Deep Salunkhe |

**Name of the Project:**

Data Visualization in R to get inference (Data set: Forest Fire)

**Project Details:**

**1. Dataset Source**

* The dataset originates from the UCI Machine Learning Repository and focuses on the Montesinho natural park, located in the Trás-os-Montes region of Portugal. It is widely used in machine learning studies for regression and classification problems.

**2. Dataset Description**

* The dataset records forest fire occurrences and related environmental factors. It aims to predict the burned area of forest fires based on meteorological data and seasonal information.

**3. Attributes and Features**

* The dataset consists of 517 instances (rows), each representing a forest fire event.
* There are 13 attributes (columns) in total, which include both categorical and numerical features:

| **Attribute** | **Type** | **Description** |
| --- | --- | --- |
| X | Integer | X-axis spatial coordinate within the Montesinho park map (1 to 9). |
| Y | Integer | Y-axis spatial coordinate within the Montesinho park map (2 to 9). |
| month | Categorical | Month of the year when the fire occurred (jan to dec). |
| day | Categorical | Day of the week when the fire occurred (mon to sun). |
| FFMC | Numeric | Fine Fuel Moisture Code (FFMC) index from the Canadian Forest Fire Weather Index (18.7 to 96.20). |
| DMC | Numeric | Duff Moisture Code (DMC) index from the Canadian Forest Fire Weather Index (1.1 to 291.3). |
| DC | Numeric | Drought Code (DC) index from the Canadian Forest Fire Weather Index (7.9 to 860.6). |
| ISI | Numeric | Initial Spread Index (ISI) from the Canadian Forest Fire Weather Index (0.0 to 56.10). |
| temp | Numeric | Temperature in degrees Celsius (2.2 to 33.30). |
| RH | Numeric | Relative humidity in percentage (15 to 100). |
| wind | Numeric | Wind speed in km/h (0.40 to 9.40). |
| rain | Numeric | Rainfall in mm/m² (0.0 to 6.4). |
| area | Numeric | The burned area of the forest in hectares (ha) (0.00 to 1090.84). Note: Most values are 0.00 (no significant burned area). |

**4. Target Variable**

* The area attribute represents the burned area (in hectares) and serves as the target variable for predictive modeling. Since many fires resulted in no significant burned area, a large portion of the data contains zero values for this attribute, making it a challenging regression problem.

**5. Data Characteristics**

* **Seasonality and Temporal Patterns:** The month and day features capture seasonal and weekly variations in fire occurrences.
* **Meteorological Factors:** Features such as temp (temperature), RH (relative humidity), wind (wind speed), and rain (rainfall) are important for understanding fire behavior and spread.
* **Fire Weather Indices:** The dataset includes Canadian Forest Fire Weather Index components (FFMC, DMC, DC, ISI), which are calculated from meteorological data and help assess fire potential.

**6. Challenges with the Dataset**

* **Imbalanced Target Variable:** Most entries have a burned area of zero, leading to an imbalanced dataset that poses challenges for modeling.
* **Outliers:** The presence of a few very large burned areas can skew the distribution and affect the performance of machine learning models.
* **Correlation Among Features:** Some features may exhibit strong correlations, such as temperature and wind speed, which could impact the predictive power of individual variables.

**7. Potential Uses of the Dataset**

* **Predictive Modeling:** The dataset can be used to develop regression models to predict the burned area based on environmental factors.
* **Clustering and Classification:** It can be employed for classifying the severity of fires or clustering similar fire events.
* **Environmental and Fire Management Studies:** Analysis of this data can provide insights into the factors driving forest fires and help develop strategies for prevention and control.

**Inference:**

**1. Scatter Plot: Area Burned vs. Temperature**

* **Inference:** The scatter plot indicates whether there's a relationship between temperature and the area burned. If there is a positive trend, it suggests that higher temperatures might correlate with larger burned areas. If the points are widely scattered without a clear pattern, it suggests temperature alone isn't a strong predictor of the burned area.

**2. Box Plot: Area Burned per Month**

* **Inference:** The box plot shows the distribution of the area burned across different months. If some months have significantly higher median burned areas or larger interquartile ranges, it suggests that certain months are more prone to severe fires. This could be due to seasonal factors, such as dry weather or wind conditions.

**3. Correlation Heatmap**

* **Inference:** The heatmap displays correlations between numerical variables, with strong correlations (either positive or negative) potentially highlighting key factors influencing the burned area. For example, a high correlation between temperature and wind could indicate that hotter days are often windier, possibly contributing to more severe fires.

**4. Bar Plot: Fire Occurrences by Day**

* **Inference:** The bar plot shows how often fires occur on each day of the week. If there is a noticeable difference in fire occurrences across days, it could be related to human activities (like weekend barbecues or work-related accidents).

**5. Histogram: Distribution of Burned Area**

* **Inference:** If the histogram shows a right-skewed distribution, it indicates that most fires affect small areas, while a few severe fires result in very large burned areas. This kind of distribution is typical for forest fire data, where smaller fires are more frequent than larger ones.

**6. Line Plot: Average Temperature per Month**

* **Inference:** This plot can help identify temperature trends over the year, such as the warmest months. A correlation with the area burned might be observed if months with higher average temperatures also show a greater extent of fires.

**7. Density Plot: Temperature Distribution**

* **Inference:** The density plot provides a smooth estimate of temperature distribution. Peaks in the plot could indicate the most common temperature ranges when fires occur, possibly suggesting conditions under which fires are more likely.

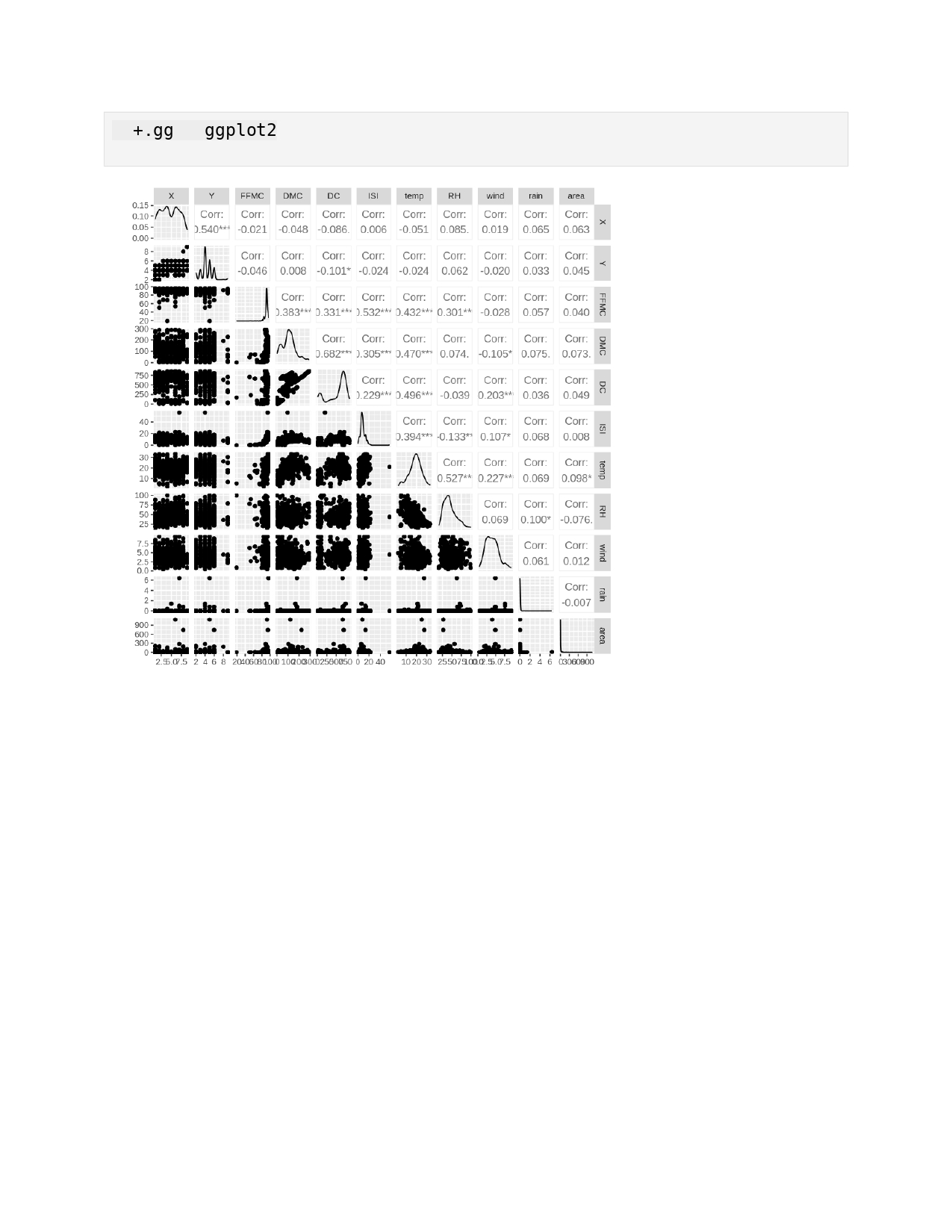
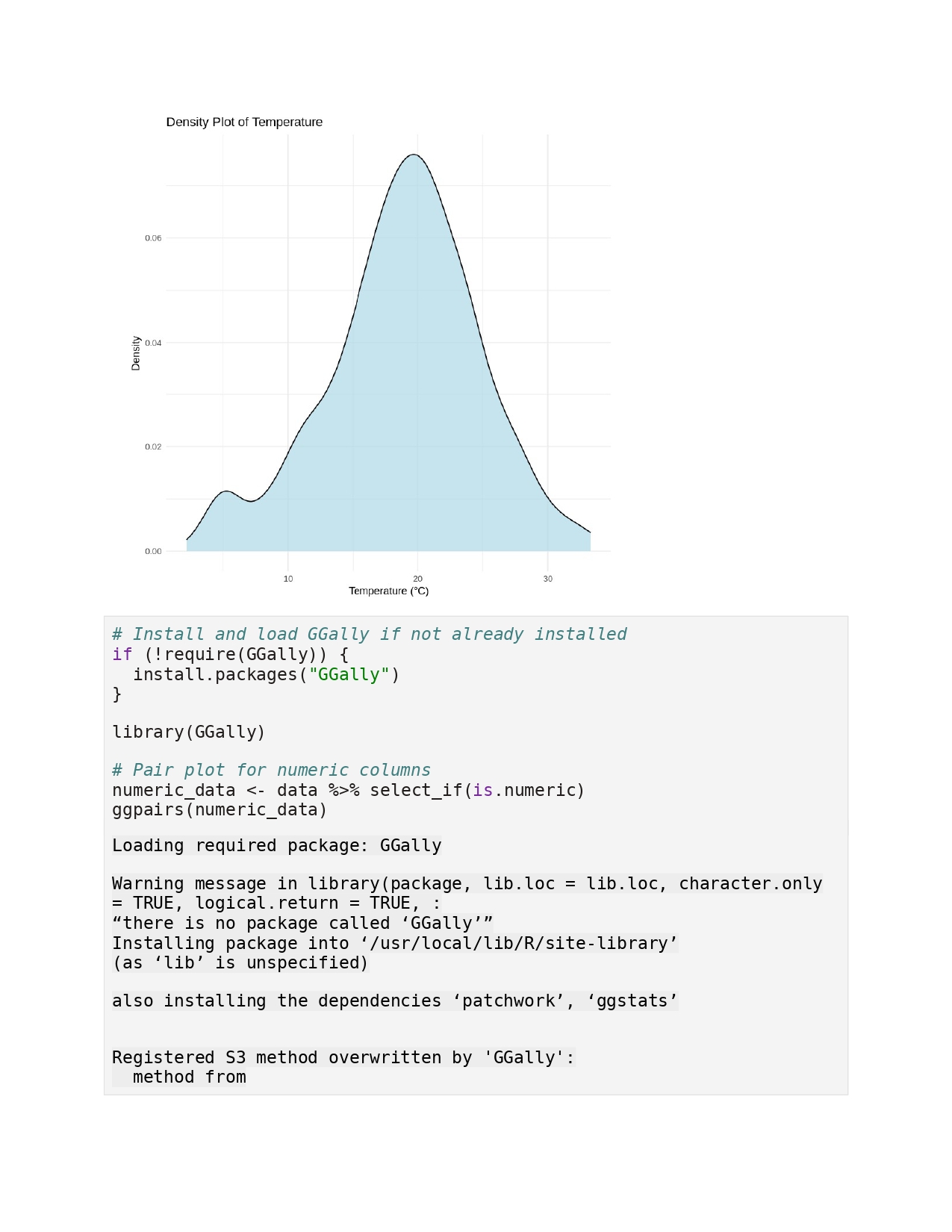
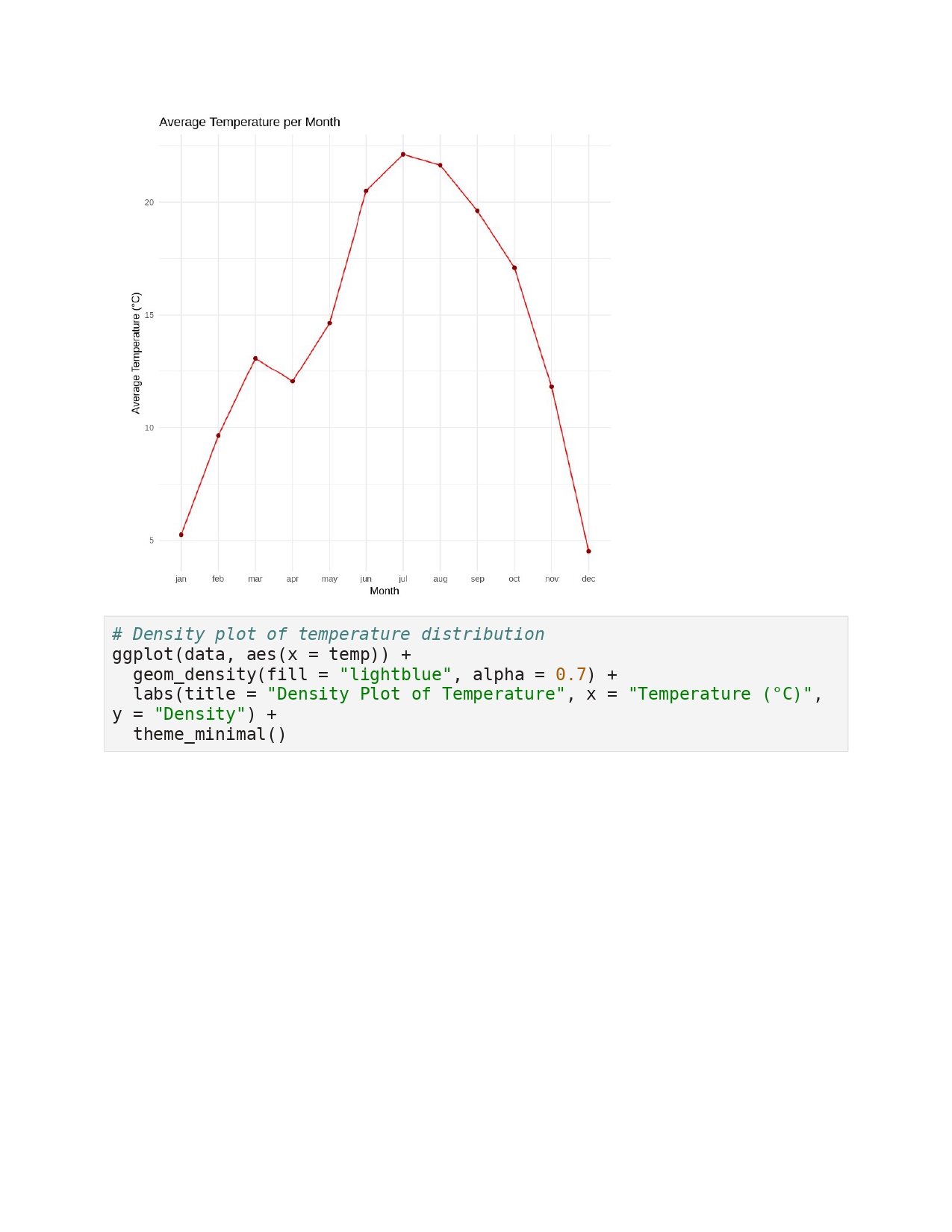
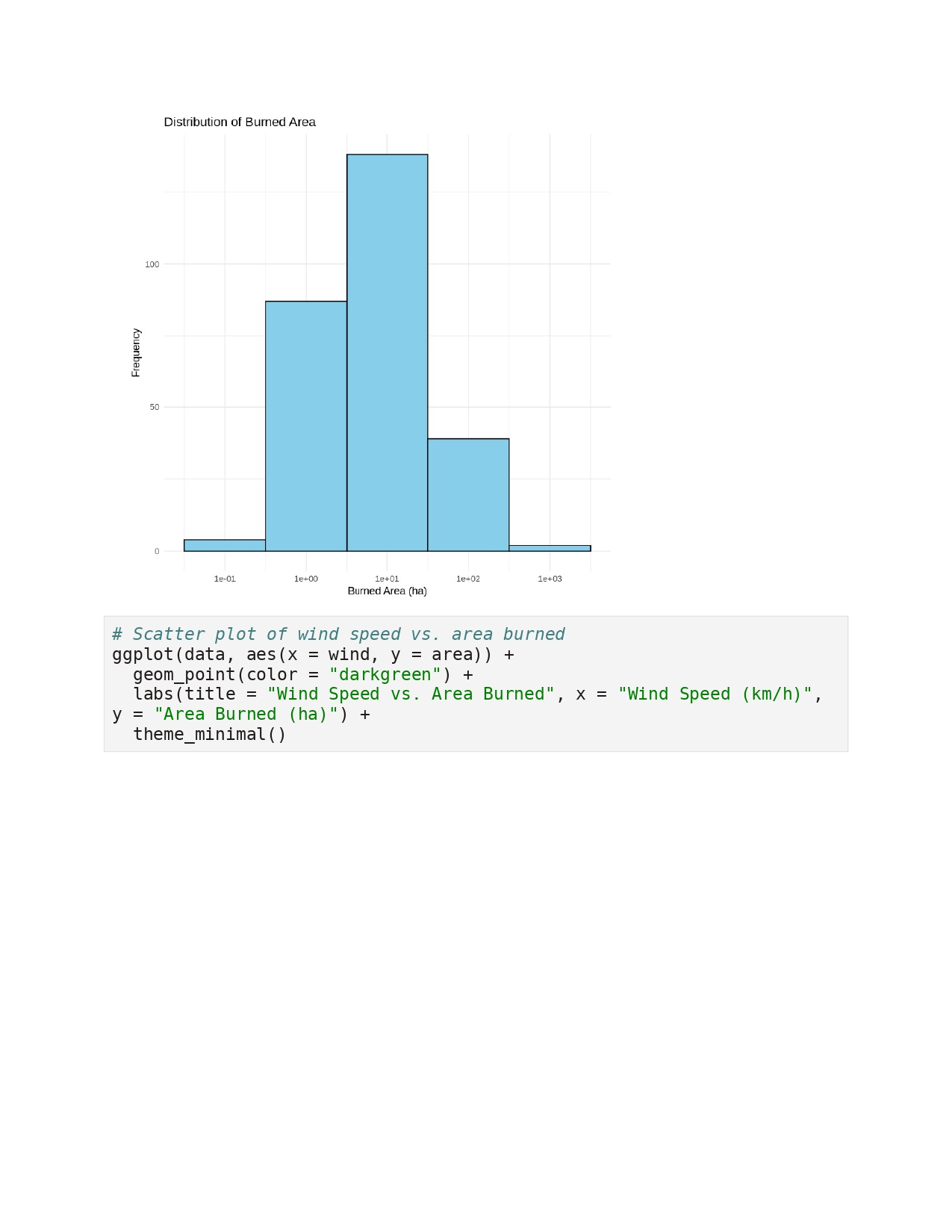
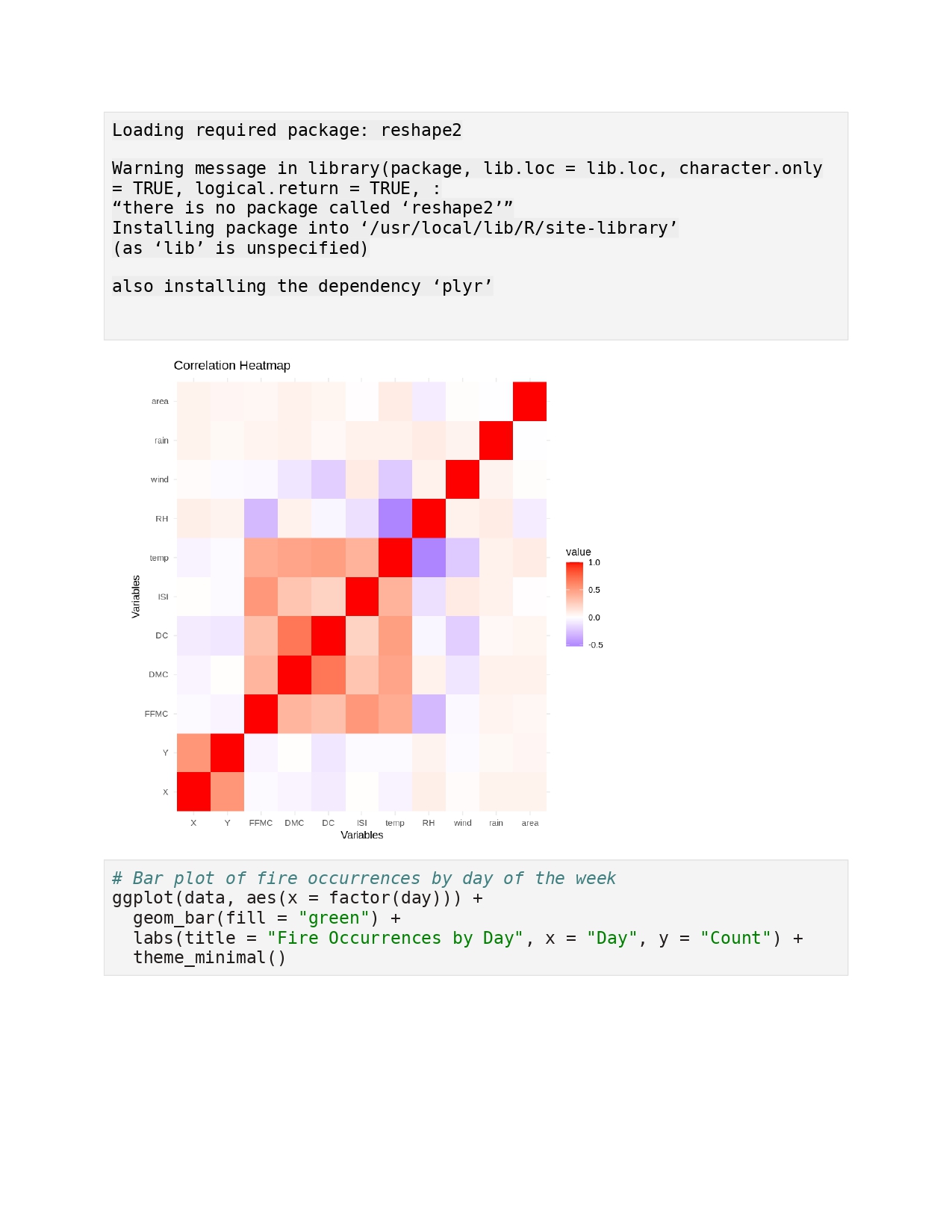
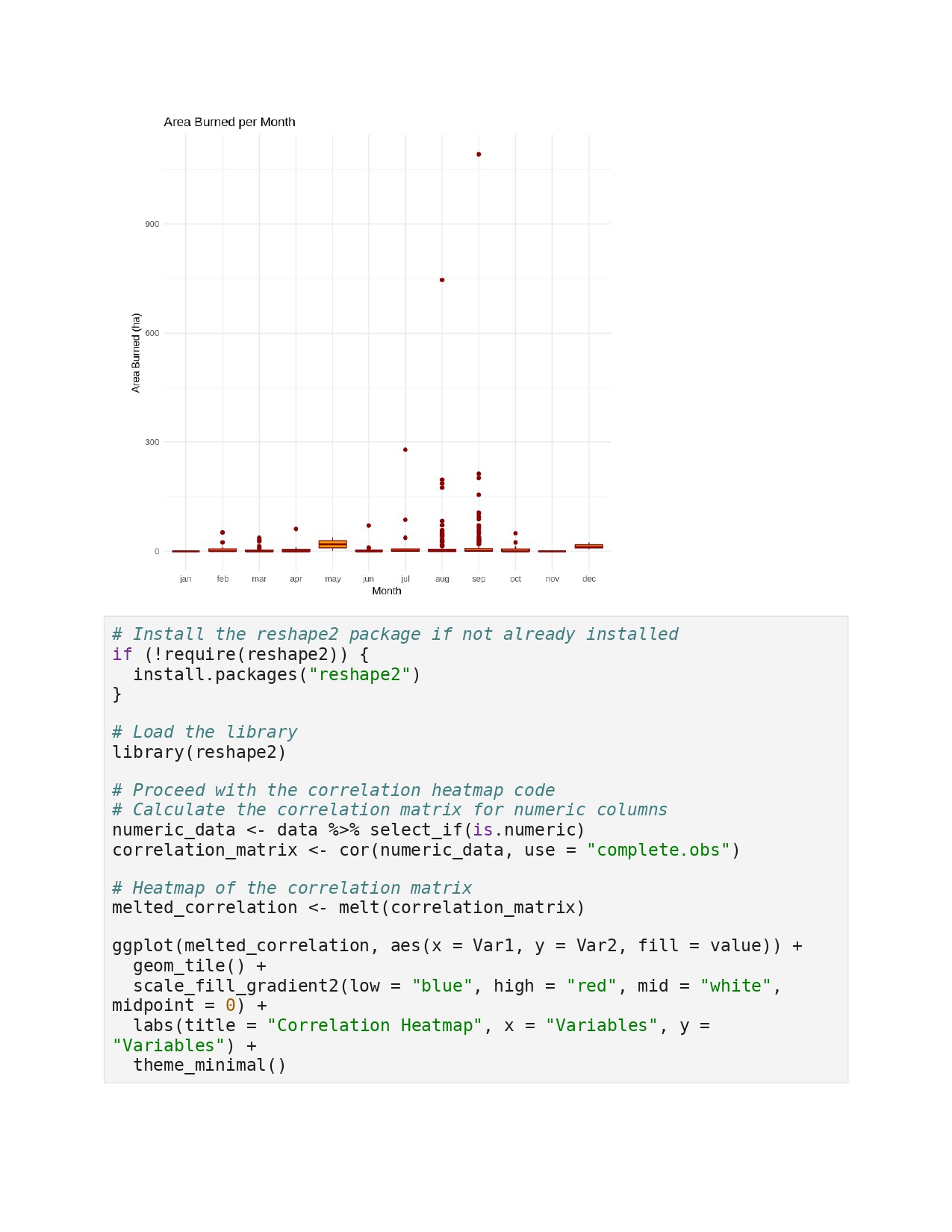
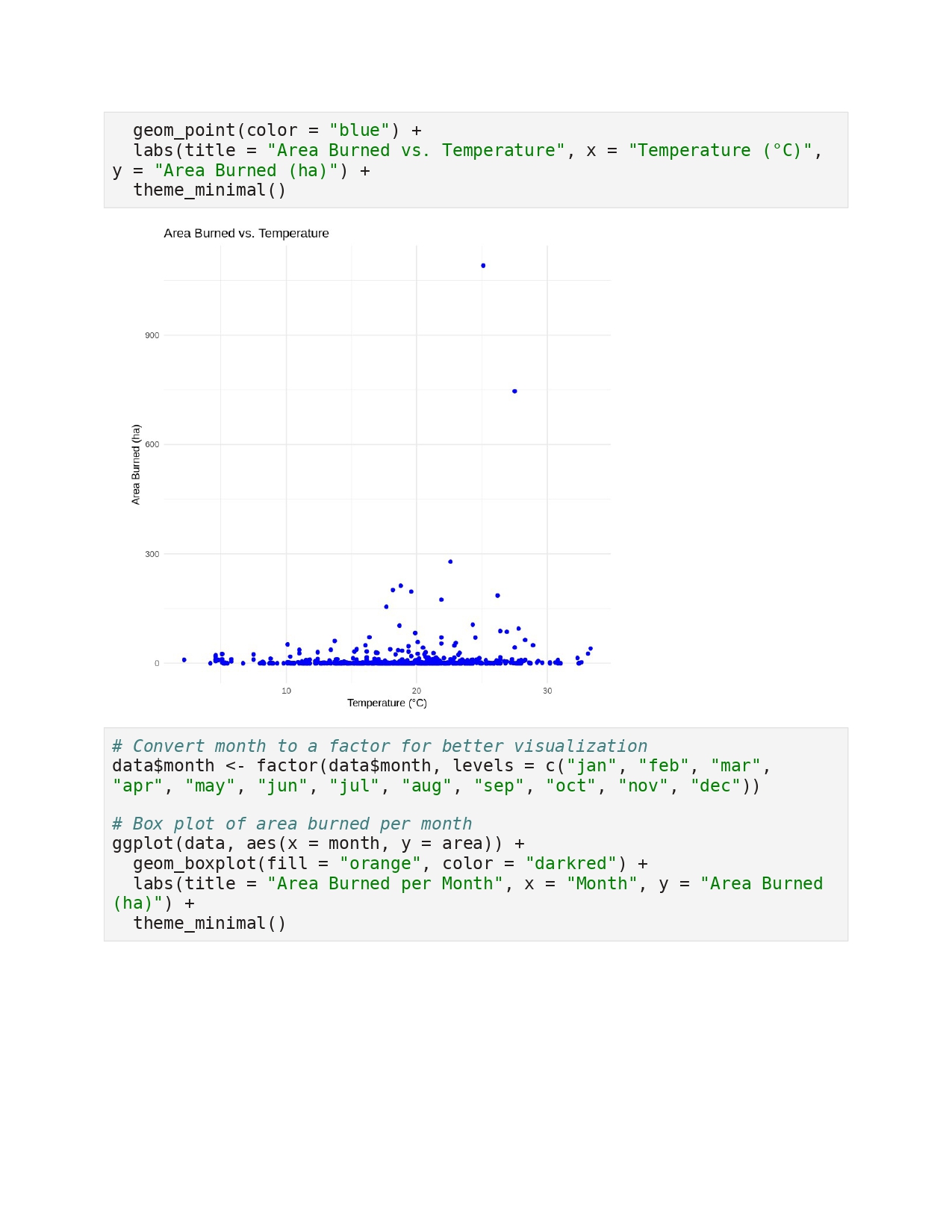
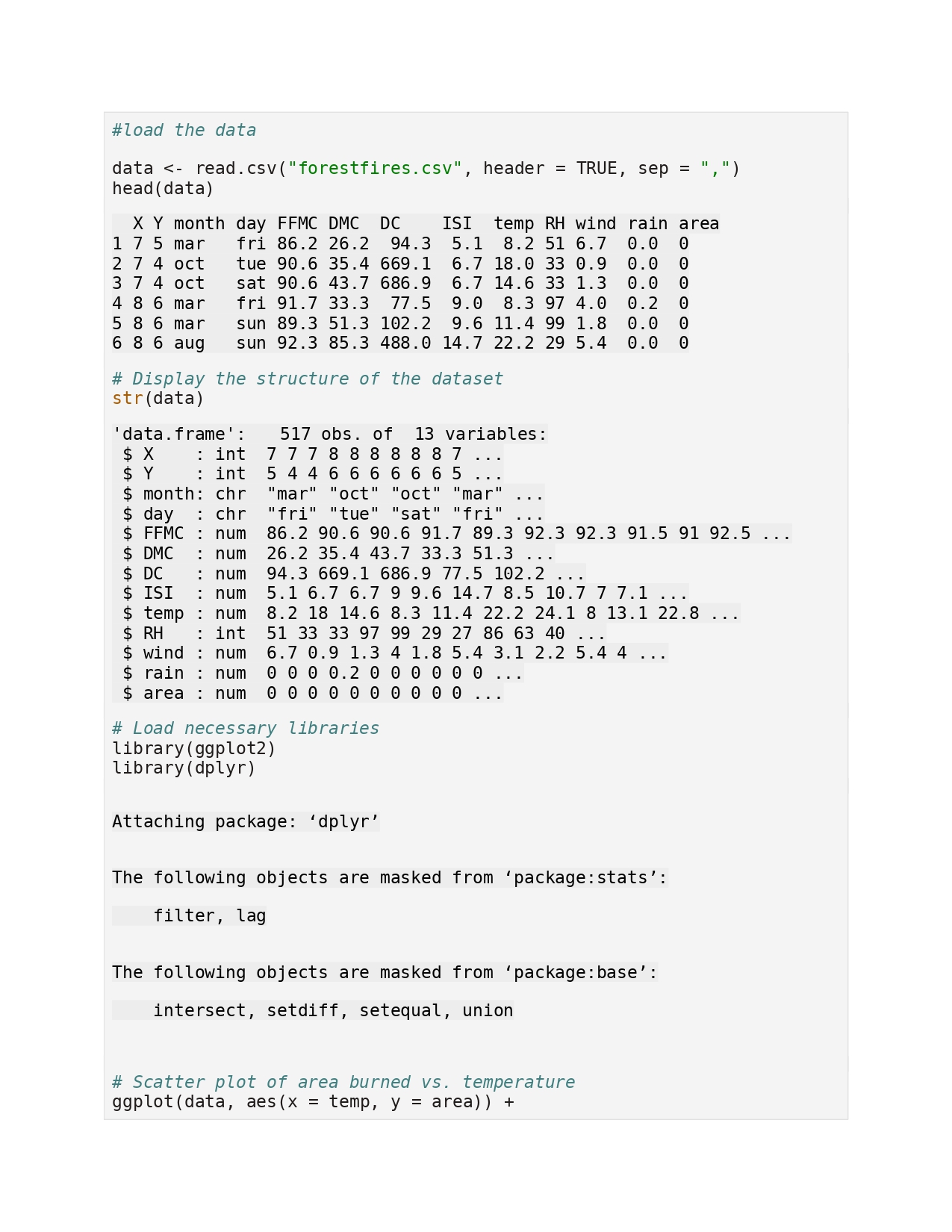
**8. Scatter Plot: Wind Speed vs. Area Burned**

* **Inference:** This plot can reveal if stronger winds are associated with larger fires. If a trend is visible, it suggests that wind might play a significant role in spreading fires.

**GitHub Repository Link (Public):**

[**https://github.com/deepsalunkhee/Learn-and-Practice/tree/master/SEM-7/BDA/R\_Project**](https://github.com/deepsalunkhee/Learn-and-Practice/tree/master/SEM-7/BDA/R_Project)

**Output Screenshots:**

****