Homework 1

AUTHOR
Elly Rokeach and Franklin Paas

PUBLISHED
September 27, 2024

This dataset was initially used for a machine learning regression model, where the aim was to predict the burned area of forest fires in the northeast region of Portugal (Cortez and Jesus Raimundo Morais 2007). Each of the 517 observations in the dataset is a forest fire, and the meteorological data surrounding their occurences including month, day of the week, temperature, humidity, and other weather conditions. In this report we will examine the prevalence of fires in relation to the month and temperature at which they occurred.

With the increasing severity of climate change, forest fires are becoming more and more of a risk. Recently, the 54,000 acre Bridge fire in the East Fork area of the Angeles National Forest had a direct impact on our lives as evacuation warnings nearly reached the Claremont Colleges. It will become increasingly relevant to analyze forest fire and meteorological data in order to be able to make predictions about when forest fires will occur and at what temperatures. We are interested in using this dataset as an example of the kind of analysis that could be conducted on forest fire data in general and in any region.

The following analysis will allow us to see which months and temperature ranges have the highest fire risks.

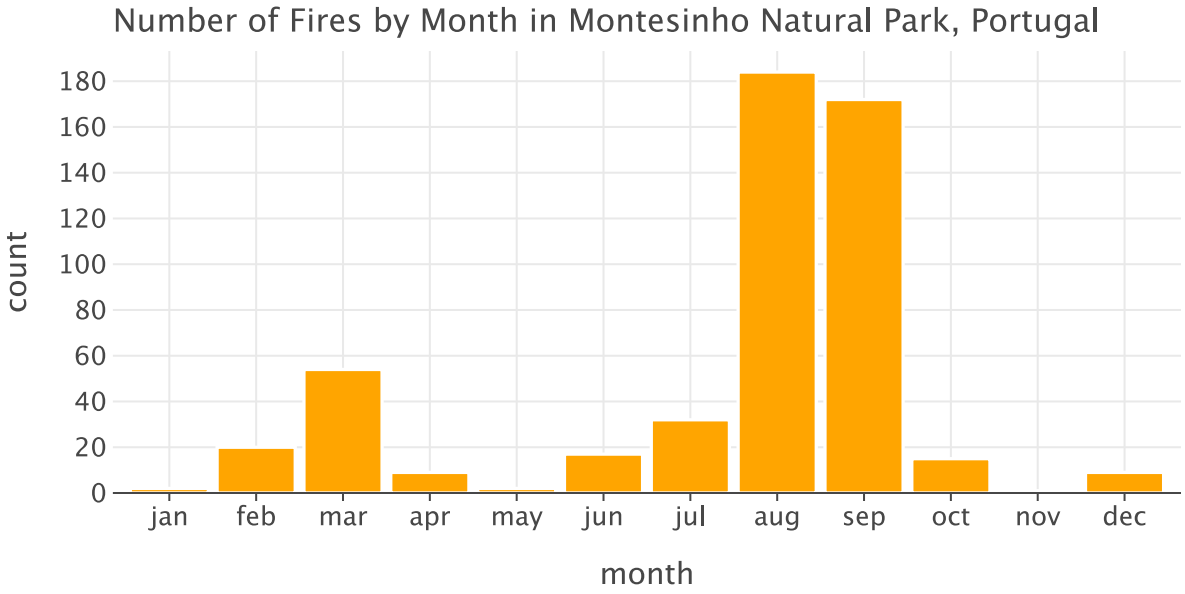
Analysis of Month

We begin by plotting the data in a bar chart, to see what months have the highest concentration of fires.

```
import numpy as np
import pandas as pd
from lets_plot import *
from scipy.stats import multinomial
LetsPlot.setup_html()
```

```
df = pd.read_csv('forestfires.csv')
df.head()
month_type = pd.CategoricalDtype(categories=['jan', 'feb', 'mar', 'apr', 'may', 'jun', 'j
df['month'] = df['month'].astype(month_type)
```

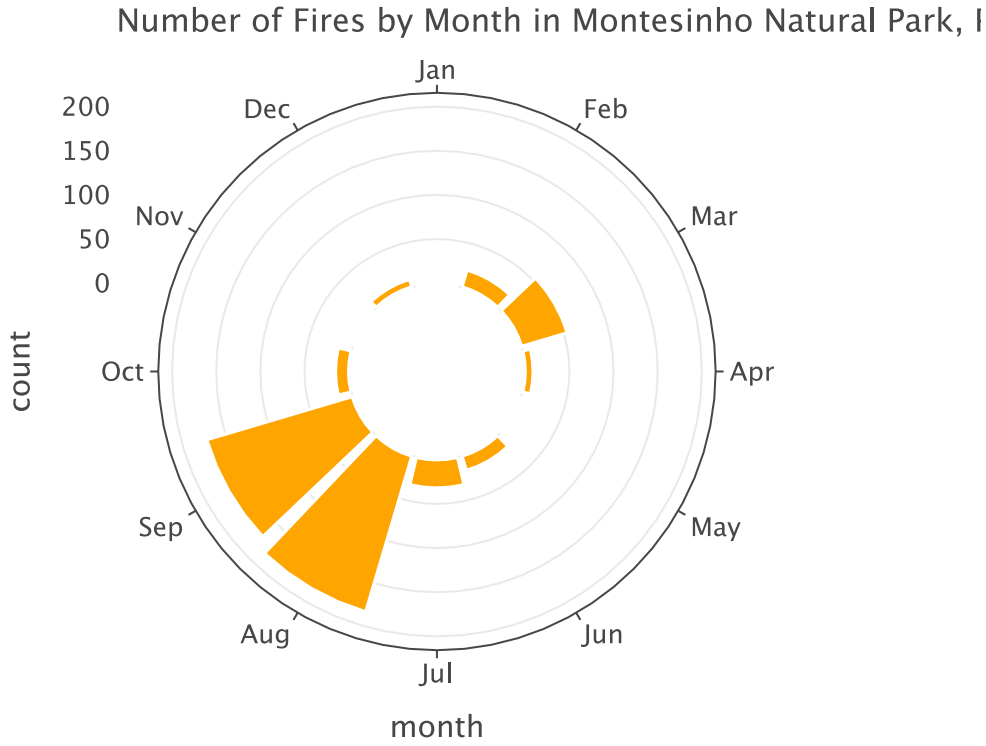
```
ggplot() + geom_bar(
    data=df, stat='count',
    mapping=aes(x='month', y='..count..'), fill='#FFA500') + scale_y_continuous(breaks=li
```



This plot shows that the number of fires reached 184 in August and 172 in September. Since this is largely above the numbers in any of the other months, this indicates that the peak fire season in this region of Portugal is in August and September.

Given the cyclical nature of the calendar, we can also visualize this data with a polar bar chart. This plot also demonstrates that the highest concentration of fires is in August and September. From October into February, it can be seen that there are very few fires, and while there is a slight jump in March, the low fire rate continues from April through July.

```
ggplot() + geom_bar(
    data=df, stat='count',
    mapping=aes(x='month', y='..count..'), fill='#FFA500') + coord_polar() + scale_x_disc
```

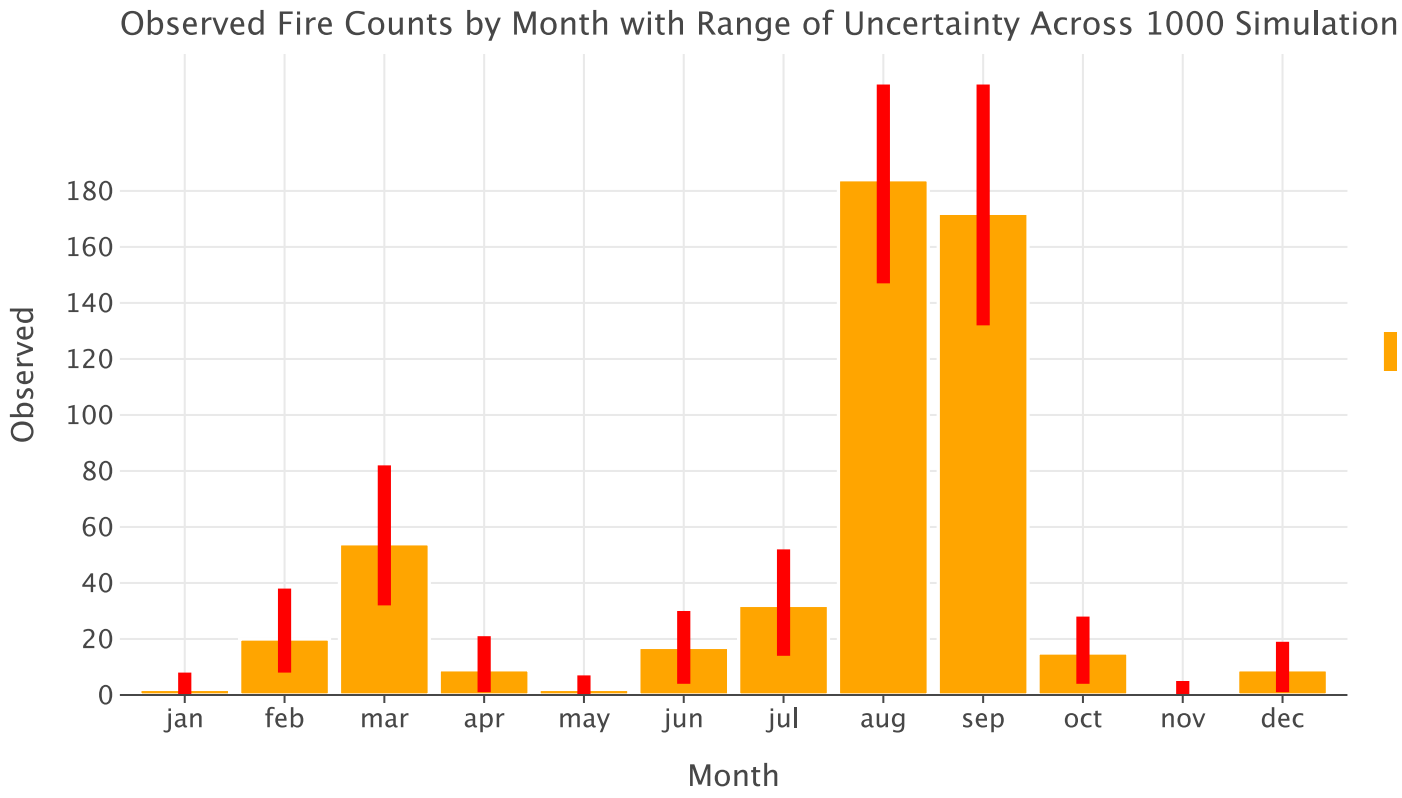


Assuming that our observed fire counts are representitive of the distrbution of actual fire occurrences, we can fit a multinomial distribution to the data and simulate the expected number of fires for any given month.

Below, we run 1000 simulations, where each simulation generates 517 fire counts based on the fitted multinomial distribution. We have plotted the observed fire counts in the original dataset in orange bars, along with the range of uncertainty generated by the simulated counts plotted in red.

```
import pandas as pd
import numpy as np
from scipy.stats import multinomial
from lets_plot import *
LetsPlot.setup_html()

fire_counts = df['month'].value_counts()
total_fires = fire_counts.sum()
probabilities = fire_counts / total_fires

n_simulations = 1000

simulated_counts = np.array([multinomial.rvs(total_fires, probabilities) for _ in range(n


expected_min = simulated_counts.min(axis=0)
expected_max = simulated_counts.max(axis=0)

plot_data = pd.DataFrame({
    'Month': fire_counts.index,
    'Observed': fire_counts.values
```

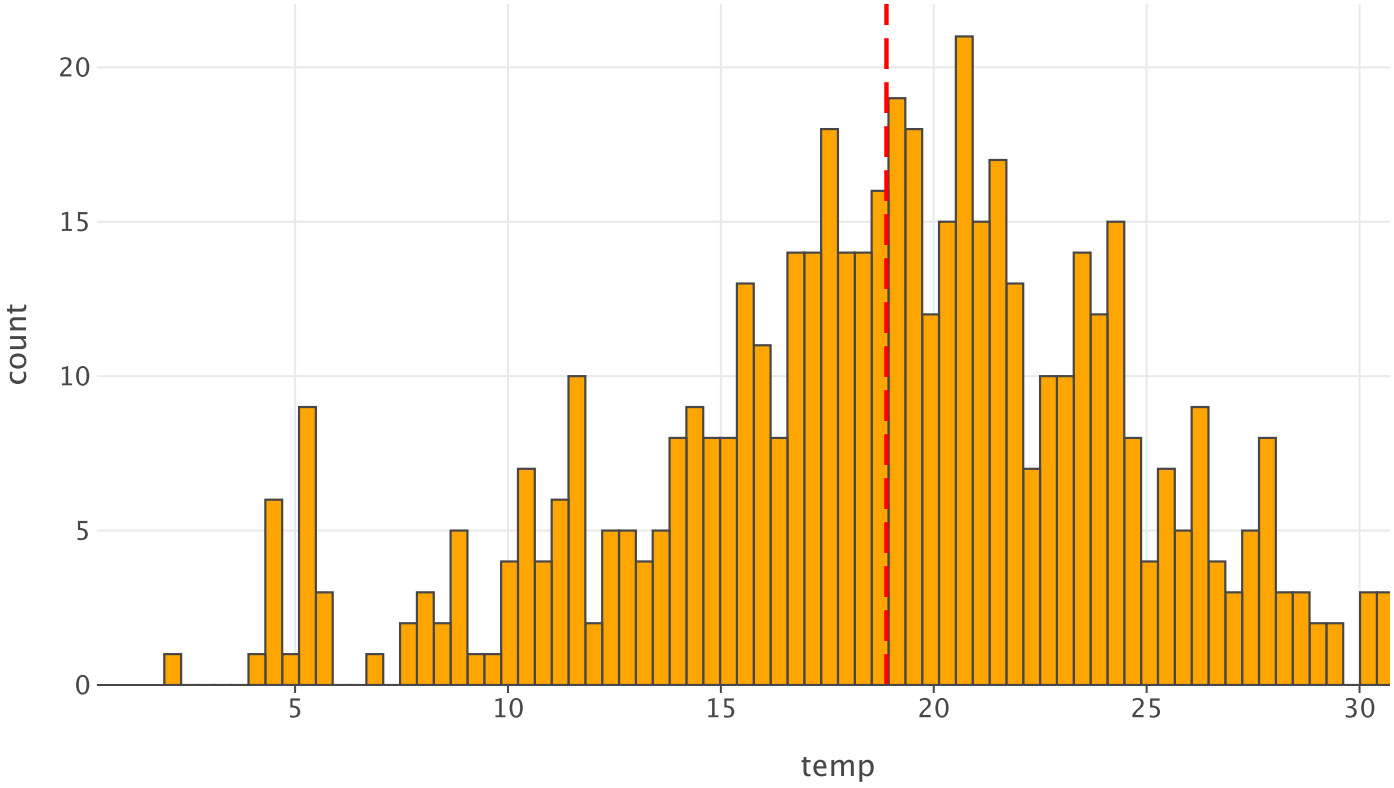
```
})

plot_data['min_expected'] = expected_min
plot_data['max_expected'] = expected_max

plot_data['Legend'] = 'Observed'
plot_data['Range'] = 'Range'

p = (ggplot() +
     geom_bar(aes(x='Month', y='Observed', fill='Legend'), stat='identity', position='dod
     geom_linerange(aes(x='Month', ymin='min_expected', ymax='max_expected', color='Range
     scale_y_continuous(breaks=list(range(0, 200, 20))) +
     ggsize(800, 400) +
     ggtitle("Observed Fire Counts by Month with Range of Uncertainty Across 1000 Simulat
     theme(axis_text_x=element_text(hjust=1)) +
     scale_fill_manual(values=['#FFA500'], name='') +
     scale_color_manual(values=['red'], name='')
    )

p.show()
```



Analysis of Temperature

```
mean_temp = df['temp'].mean()

ggplot() + geom_histogram(
    data=df,
```

```
    mapping=aes(x='temp'), bins=80, fill='#FFA500') + \
    geom_vline(
        xintercept=mean_temp,
        color='red',
        linetype='longdash',
        size=1) + ggsize(800, 400)
```



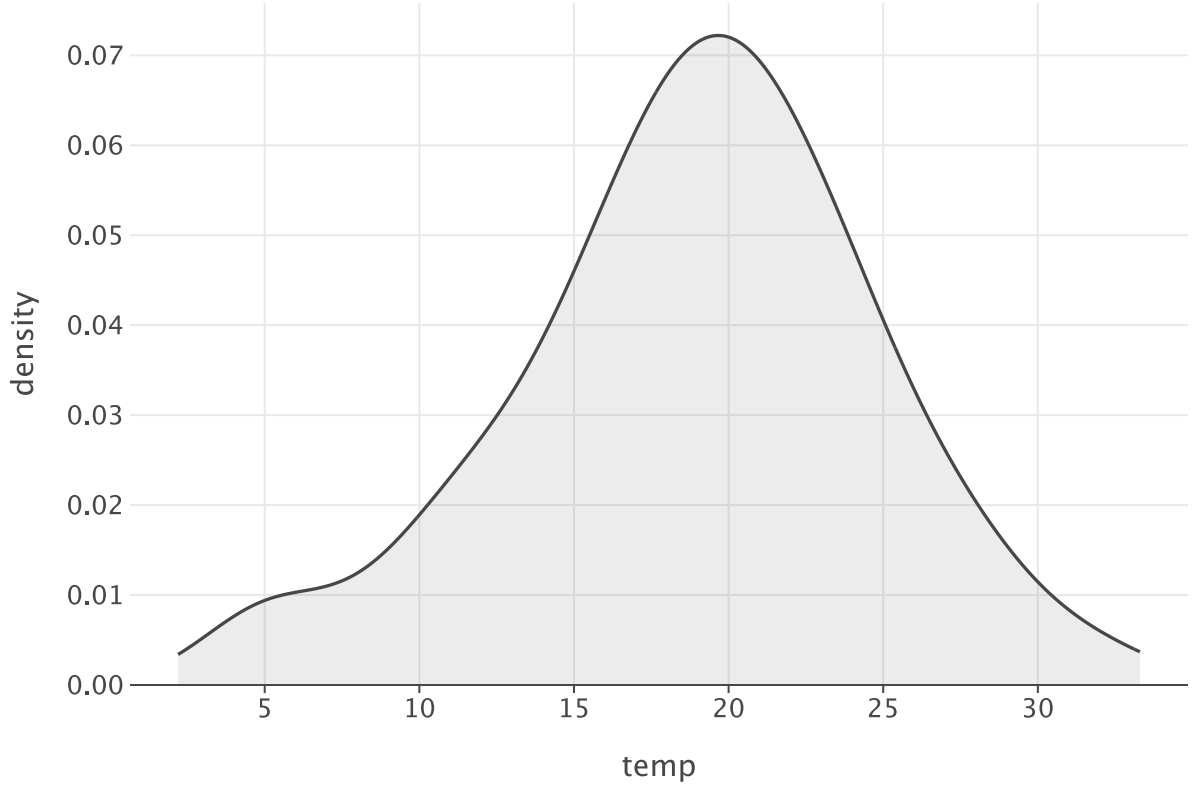

```
min_temp = df['temp'].min()
max_temp = df['temp'].max()

print("Mean temp: " + str(round(mean_temp, 2)))
print("Min temp: " + str(min_temp))
print("Max temp: " + str(max_temp))
```

```
Mean temp: 18.89
Min temp: 2.2
Max temp: 33.3
```

This histogram with 80 bins shows a distribution of the number of fires that occurred at each temperature. The mean is marked with a dashed red line. The mean temperatures at which fires occured in this region was 18.89 degrees Celsius. The minimum temperature is 2.2, and the maximum temperature is 33.3.

The following Kernel Density Estimation plot shows that there is one distinct mode around 20 degrees Celsius. This aligns with the previous histogram plot.

```
ggplot(data=df, mapping=aes(x='temp')) + geom_density(kernel='gaussian', bw=2)
```



One would expect to see higher temperatures correspond with an increased frequency of fires. However, these visualizations do not show a direct relationship between temperature and number of fires. This is an unexpected result, but is explainable by the limitations imposed by the dataset. This dataset only contains observations on days when fires occurred, and does not contain observations of days with no fires. Therefore, these distributions do not account for the proportion of days with temperatures over a specific amount. So, the decrease in fire counts at temperatures exceeding 20 degrees is most likely due to the small percentage of days recorded with temperatures over 20 degrees, rather than decreased likelihood of a fire occuring on any given day over 20 degrees.

Ideally, we would be able to answer the question: given that the temperature is above 25 degrees Celsius, what is the probability of a fire ocurring? However, since we only have observations on days with fires occuring, we can only ask the question: given that a fire is occurring, what is the probability of the temperature being above 25 degrees Celsius? However, relying on the condition that there is a fire occurring in the first place is not a very helpful or realistic prediction for real-world applications. Given that we also only have 517 observations, it is difficult to draw any meaningful conclusions from this dataset. More meaningful analysis could be conducted if we had data points on days without fires, and a larger dataset in general.

References

Cortez, P., and Aníbal de Jesus Raimundo Morais. 2007. “A Data Mining Approach to Predict Forest Fires Using Meteorological Data.” In. https://api.semanticscholar.org/CorpusID:36868619.