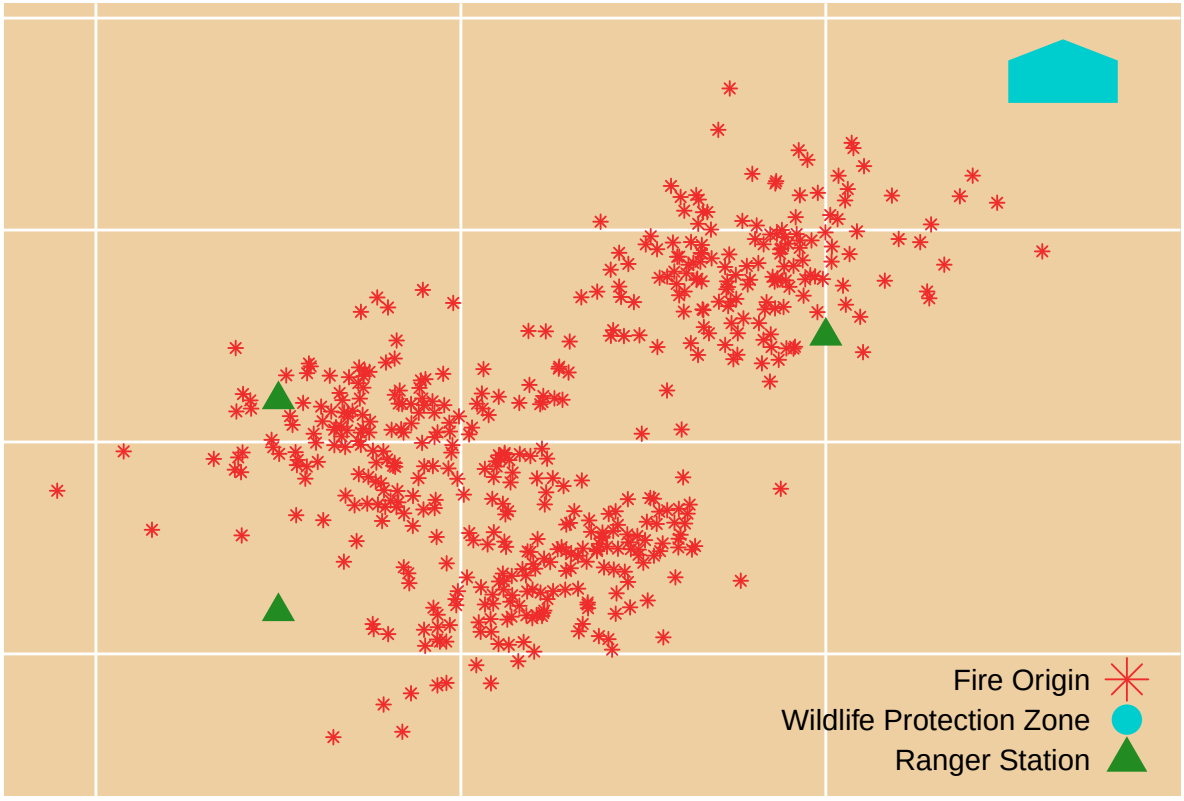
Group Project 1

STAT 301-1

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## -- Column specification --------------------------------------------------------
## cols(
##   x = col_double(),
##   y = col_double(),
##   temp = col_double(),
##   humidity = col_double(),
##   windspd = col_double(),
##   winddir = col_character(),
##   rain = col_double(),
##   days = col_double(),
##   vulnerable = col_double(),
##   other = col_double(),
##   ranger = col_double(),
##   pre1950 = col_double(),
##   heli = col_double(),
##   resources = col_double(),
##   traffic = col_character(),
##   burned = col_double(),
##   wlf = col_double()
## )
```

Assignment

For the group project, you will be using the wildfires dataset (wildfires.csv). This data describes 350 wildfires that started within a large national park. The origin of each fire is shown below in red.



A number of factors may affect how large a fire becomes. For example, if it starts near a ranger station (green triangles on the map) at a time when it is manned, fires may be less likely to spread.

Your job is to make two types of predictions. First, you must predict how large a fire will be. In the context of the data, you will be using all of the variables (except 'wlf') to predict the 'burned' column.

Second, you will need to predict whether a fire will spread to a wildlife protection zone denoted by the light green triangle in the northeastern section of the park. In other words, use all of the columns (except 'burned') to predict the 'wlf' column.

Both prediction tasks are detailed in the next two sections, followed by a description of the dataset.

Task 1: Area Burned

The total area burned by a wildfire is of great concern to government planners. This is captured in the 'burned' column of the wildfires dataset. Given the data (excluding the 'wlf' column), build the best model you can to predict the number of acres a wildfire is likely to burn. You are free to use any model or predictor you choose (e.g., linear regression, ridge regression, lasso, etc.) that minimizes test error.

Task 2: Wildlife Protection

Located in the northeast of the wilderness area is a wildlife protection zone. It is home to several rare and endangered species, and thus conservationists and park rangers are very interested in whether a given fire is likely to reach it. Build a model that predicts whether a fire will reach this zone. As with the first task, you are free to use any model or method you choose that minimizes test error.

Submission & Evaluation

Your team submissions will comprise two products. First, you will need to write a report about your models. Your report should include a description of the data, describe the methods you used, and present your final models. Be sure to justify your methods and clearly explain what you did using the theory we have covered in class. When you present your models, interpret them (e.g., what is a useful predictor?) Moreover, as with the homework, reports should include very little code (preferably none), but rather should communicate why you chose the methods and predictors you did, and how you tweaked and adjusted your approach in light of diagnostics and estimated test error.

Second, your team will submit an R script containing code to fit your model. This code must do the following:

1. Read in the wildfires dataset and store it as 'train'.
2. Fit your model predicting the size of the fire (area burned), and store it as 'model1'.
3. Fit your model predicting whether a fire will reach the wildlife protection zone and store it as 'model2'.

Save your R script with the following naming convention: the word 'gproj' followed by an underscore followed by your team member initials. For example, I would submit an R file named 'gproj_js.R'.

We will evaluate both your report and your models. Reports will be graded for clarity, completeness, and whether you demonstrated sound use of theory and practice. Your models will be tested against a batch of data we have held out.

Data

The dataset contains observations on 350 fires started in a large national park. For each fire, the data contain 15 variables, in addition to two outcomes ('burned' and 'wlf'). These variables are:

- **x**, and **y**: the (x, y) coordinates of where the fire started.
- **temp**: air temperature when the fire started
- **humidity**: air humidity when the fire started
- **windspd**: wind speed when the fire started
- **winddir**: wind direction when the fire started
- **rain**: rainfall in the week preceding the fire
- **days**: number of days since the last fire
- **vulnerable**: amount of vulnerable (unburned) foliage near the fire
- **other**: indicates if another fire is ongoing when the fire starts
- **ranger**: indicates if the fire started near a ranger station, and that ranger station was manned. Note that even if a fire started near a ranger station, that station may not have been manned at the time.
- **pre1950**: indicates if the fire occured before 1950, which it became more common to use planes and helicopters to fight forest fires.
- **heli**: indicates if a helicopter was available to fight the fire
- **resources**: composite score for the resources available to fight the fire, including manpower and available air support
- **traffic**: indicates the level of foot traffic in the national park when the fire started