1. Nenana Ice Classic (8 points)

In 1917, railroad workers were building a bridge across the Tanana River in Alaska. As a diversion, they placed bets on when the ice in the river would start to break up, and a betting pool was started with a pot of $800. Over the years, the contest has grown, and the Nenana Ice Classic (as it has come to be known) now sports a pot of over $300,000 and is regulated by the state of Alaska as a legalized game of chance. Because of the large amount of money at stake, the exact moment of ice breakup has been recorded carefully each year. This dataset is now finding use as a consistently measured and high-quality source of data on local climate change.

First, read the following articles:

older news article
http://news-service.stanford.edu/news/2001/october31/alaskabet-1031.html

more recent news article
http://www.newsminer.com/news/local_news/nenana-ice-classic-tripod-falls-earliest-date-on-record/article_fb399560-5ed6-11e9-9aa0-e300d14b3a97.html

research article:
https://science.sciencemag.org/content/294/5543/811

Also, quickly browse the official site:
http://www.nenanaakiceclassic.com

The dataset `nenana` (available on Moodle) gives the ice breakup date for each year from 1917 to 2019. The breakup date is given in Julian date format, which represents the number of days since the beginning of the year. (A value of 120, for instance, means the ice broke up on the 120th day of the year.)

Read in the dataset using the following commands:

```
# read in dataset
setwd("~/fiil in your pathname")
data <- read.csv("nenana.csv")

# define variables
year <- data[,"year"]
date <- data[,"date"]
```

Investigate this dataset using loess. The command for loess is similar to `lm` for linear regression, except that there is a `span` parameter that governs the degree of smoothing:

```
fit <- loess(date ~ year, span=.75)
plot(year, date)
lines(year, predict(fit), col="red")
```

Is there any evidence for a warming trend? If so, over what time period has the trend occurred? Is the trend linear or nonlinear? How does the curve change when you change the value of `span`? **Hand in your graph(s), indicate the value of** `span` **you used, and summarize your findings in a few sentences.**

(To be clear, this is just one dataset for one location; by itself this cannot prove or disprove global warming. However, it may provide one piece of evidence that can be combined with other evidence to increase our understanding of climate change.)

2. Spam (21 points)

Ever wonder how your email program is able to distinguish spam from real messages? One way to do this is by using logistic regression. In this question we'll create a highly simplified spam filter.

The dataset `spam` (available on Moodle) has data from 1000 email messages sent to George Forman, a Hewlett-Packard computer scientist. For each message, Forman recorded how often (as a percentage of the total number of words in the message) the words `meeting` and `credit` appeared. The units are percentage points; that is, a value of 0.22 for `credit` means 0.22% of the words in the message were "credit" (not 22%). (The dataset is excerpted from a much larger dataset, with 4601 messages and 57 predictors. If you're curious, you can download the entire dataset from `http://archive.ics.uci.edu/ml/datasets/Spambase`.)

1) First, let's make a picture. Make boxplots of `meeting` and `credit` as Y variables against `spam` as the X variable. You can use the `boxplot` command, and you can specify the Y and X variables as in regression. **How does the distribution of each word differ in spam vs real messages? Hand in the boxplots.**

2) Now we'll fit a logistic regression model with `spam` as Y and `credit` and `meeting` as X's and include an interaction term. **Is the interaction term significant? Why or why not? If not, delete the interaction term and re-fit the model. Hand in the glm output.**

3) **What is the coefficient giving the effect of `credit`? Is `credit` a statistically significant predictor? Calculate the odds ratio for `credit` and explain what this quantity means.**

4) What is the coefficient giving the effect of `meeting`? Is `meeting` a statistically significant predictor? Calculate the odds ratio for `meeting` and explain what this quantity means.

5) Using your model, predict the probability that a new message is spam if `credit` makes up 0.2% of the total words in the message and `meeting` does not appear at all. **First, calculate this probability by hand, showing your work.** Next, we'll use R to do the calculation. To do this, you need to create an object that contains the data for the new message, which you can do as follows:

```
new <- as.data.frame(t(c(.2, 0)))
colnames(new) <- c("credit", "meeting")
```

Type `new` to verify that the new object has two columns, appropriately labeled, and that the values are as intended. If that works, then use the following commands to calculate the prediction (assuming `fit2` holds the results of your model from part (2) above):

```
predlogodds <- predict(fit2, newdata=new)
1 / (1 + exp(-predlogodds))        # be sure to include the negative sign!
```

Note that `predlogodds` is just $b_0 + b_1x_1 + b_2x_2$, so we need the second command to convert it to a predicted probability. **Does this result match what you got by hand?**

(more)

6) How successful is your model at distinguishing spam messages? To determine this, use the following commands:

```
predlogodds <- predict(fit2)
predprobs <- 1 / (1 + exp(-predlogodds))     # be sure to include the negative sign!
predspam <- predprobs >= .5
```

The first command calculates the predicted logodds $b_0 + b_1x_1 + b_2x_2$ (if a `newdata` option is not specified, the default is to calculate the predicted logodds for the dataset used to fit the model). The second command converts these logodds to predicted probabilities of being spam. The third command says we predict a message is spam if its predicted probability is at least 0.5.

Now make a table of predicted and actual spam. **(Hand in this table.) Of the messages your model predicted to be spam, what percent actually are spam? Of the messages your model predicted to be non-spam, what percent actually are not spam?**

7) **Is your spam filter conservative (incorrectly misses many spam messages) or aggressive (incorrectly classifies real messages as spam)? Without collecting any new data or running any new analyses, how can you easily change the conservativeness/aggressiveness of your spam filter?**