Stat 341/641 Final Exam

Name:

Student Number:

SFU email:

Instructions:

Fill in your personal information in the spaces above. On the last page of the exam is an Appendix of useful functions. In addition to this exam paper you should recieve the four "cheatsheets" described in an email to the class. You are not allowed any other aids. Write your answers directly on the exam. Throughout, you may assume that all functions you need have already been loaded from the relevant packages. Please try to use the forward pipe, `%>%`, to chain together multiple operations on data frames when you don't need the intermediate results of the computations. When you **do** need results from computations for future work, you must save these results or you will lose marks.

Question 1 (9 marks)

You have a vector of file names of the form `chrNgeneC.txt`, where `chr` is short for chromosome, `N` can be a number between 1 and 22, or the letters `X` or `Y` and `C` is a character string of arbitrary length. Here is a sample of what these filenames might look like.

```
fl <- c("chr1geneC1orf49.txt","chr1geneAMPD2.txt",
        "chr2geneCOL4A4.txt","chr2geneHADHB.txt",
        "chrXgeneCXorf49.txt")
```

a. (3 marks) Write a function called `findChrFiles()` that takes the following inputs: (i) a character vector (such as the example `fl` above) of filenames, and (ii) a character string in the set `cnum <- c(as.character(1:22),"X","Y")` and returns all filenames that include that chromosome. Have your function check that the chromosome identifier is in `cnum` and return the error message "Invalid chromosome" if not.

b. (3 marks) Write a function `findGeneNames()` that takes a vector of filenames and returns the gene names found in the filenames.

c. (3 marks) Write a function `readORF()` that takes a chromosome identifier in the set `cnum` defined in part (a) and a vector of gene names, and reads in all the files from that chromosome with a gene name that contains the string `orf`. Assume that the files all contain tabular data that is read in correctly by `read.table()` with its default settings. Save the data from these files in a list that has one list element per gene, and with list elements named by gene. For example, "chrXgeneCXorf49.txt" should get read into a list element with name "CXorf49". Don't bother with error checking the inputs.

Question 2 (6 marks)

Data frames STATION and STATS are created as follows:

```
STATION <- data.frame(ID=c(13,44,66),
    City = c("Phoenix","Denver","Caribou"),
    State = c("AZ","CO","ME"),
    Lat_N = c(33,40,47),
    Long_W = c(112,105,68),stringsAsFactors=FALSE)
STATS <- data.frame(row = 1:6,
    ID = c(13,13,44,44,66,66),
    Month = c(1,7,1,7,1,7),
    Temp_F = c(57.4,91.7,27.3,74.8,6.7,65.8),
    Rain_I = c(0.31,5.15,0.18,2.11,2.1,4.52))
```

a. (1 mark) Extract the stations from the state of Colorado

b. (1 mark) Select the `City` and `State` from the `STATIONS`.

c. (1 mark) Add Vancouver as a new station in the `STATION` data frame. Use BC as the `State` and the latitude 49 and longitude 123.

d. (3 marks) Do an inner join that returns a table with city, state, and temperatures for July from cities at north latitude 40 or more.

Question 3 (6 marks)

The following data are from a summary of Stat and Act Sci students.

```
stat <- data.frame(
  year = 12:16,
  FTEs = c(446.47,484.8,483.53,443.97,466),
  majors = c(213,245,260,228,233),
  minors = c(17,27,62,77,111),
  p.intl = c(.4,.42,.51,.53,.54))
```

a. (1 mark) Reshape the `FTEs`, `majors` and `minors` columns into a key-value pair of columns with the key variable labeled as `type` and value variable labeled as `students`. Save your results in a data frame called `statlong`.

b. (1 mark) Plot `students` versus `year` by group `type`, with different colors for each type of student. Add a scatterplot smoother **without** standard error bands.

c. (2 marks) Add the following variables to the `stat` data frame: (i) `total = majors+minors`, (ii) `international = total*p.intl` and (iii) `domestic = total*(1-p.intl)`

d. (1 mark) Reshape the `total`, `international` and `domestic` columns into key-value pairs with key labeled `type` and value labeled `students`. Save your results in a data frame `majmin`.

e. (1 mark) Plot `students` versus `year` by group `type`, with different colors for each type of student. Add a scatterplot smoother.

Question 4 (9 marks)

`youthUI` is a data frame of youth unemployment data. The first 4 rows are as follows.

```
##   Country.Name Country.Code    X2010    X2011    X2012    X2013    X2014
## 1  Afghanistan          AFG 20.60000 20.90000 19.70000 21.10000 20.80000
## 2       Angola          AGO 10.80000 10.70000 10.70000 10.60000 10.50000
## 3      Albania          ALB 25.80000 27.00000 28.30000 28.70000 29.20000
## 4   Arab World          ARB 25.02221 28.11752 29.11321 29.33531 29.70457
```

a. (2 marks) Reshape the data from 2010 to 2014 into key-value pairs with key `year` and value `unemplRate`. Remove the `X`'s from the year names and coerce to numeric. Over-write `youthUI` with the reshaped data frame.

b. (1 mark) Plot unemployment rates by year for each "country" in `youthUI`. Represent each time series by a line. Use an alpha level of 0.2 to manage overplotting.

c. (3 marks) Extract the subset of "Countries" whose `Country.Name` contains the string "(IDA & IBRD countries)", and save in a data frame `youthDevel`. Remove the "(IDA * IBRD countries)" from the country names. Remove `Country.Code` from the data frame.

d. (3 marks) Initialize a plot of unemployment rates by year for each region in `youthDevel` with different colors and symbols for each region. Then add the following layers:
   - Add lines.
   - Add points. Do this with `geom_point()`, but add a comment to your R code that would generate the same layer with the generic `layer()` function.
   - In the legend of your plot, modify the legend title from its default to `Region`.
   - Add the world-wide unemployment data from `youthUI` (`Country.Name==World`).
   - Finally, display your plot.

Question 5 (8 marks)

A dataset `FruitFlies` has variables `Longevity` and `Treatment`.

a. (1 mark) Plot histograms of `Longevity` in separate facets for each level of the factor `Treatment`. Use a binwidth of 10 for your histograms.

b. (1 mark) Do boxplots of `Longevity` by `Treatment`

c. (2 marks) Calculate sample size, mean and SD of `Longevity` within each level of `Treatment`.

d. (1 mark) Plot 95% confidence intervals for the mean `Longevity` in each treatment group.

e. (1 mark) The following function `fstatPerm()` is meant to calculate the F statistic on a permutation of the `FruitFlies` data. What is the missing R code marked by `___`?

```r
fstatPerm <- function() {
  permdat <- data.frame(Treatment = FruitFlies$Treatment,
                        Longevity = ___)
  mm <- lm(Longevity ~ Treatment, data=permdat)
  tidy(anova(mm))[1,5]
}
```

f. (2 marks) Write code to (i) call `fstatPerm()` 1000 times to get the permutation distribution of the F test of treatment effect using 1000 permutations and (ii) compute the permutation p-value.

Question 6 (9 marks)

A business collects data on weekly advertising expenditures for TV, radio and newspaper ads in thousands of dollars, and sales in thousands of units sold, over 200 weeks. The data are in a CSV-format spreadsheet called `Advertising.csv` with column labels `Week`, `TV`, `Radio`, `Newspaper` and `Sales`.

a. (1 mark) Read the data into a data frame called `ad`, assuming that the `Advertising.csv` spreadsheet is in your R working directory.

b. (3 marks) Next (i) add the variable `RadNews = Radio + News` to the data frame, (ii) drop the `Radio` and `News` columns and (iii) reshape the `TV`, `RadNews` columns into a key-value pair of columns with the key variable labeled as `Type` and value variable labeled as `Expenditures`. Save your reshaped data in a data frame called `adlong`.

c. (2 marks) Plot advertising expenditures by week as points on a single graph with different colours for `TV`, `RadNews`. Add a scatterplot smoother.

d. (3 marks) Starting with the `ad` data frame do the following: Create new columns (i) `Total`, the total of `TV`, `Radio` and `Newspaper`, (ii) `pTV`, the proportion of ads that are TV adds, `TV` ads, and (iii) a categorical variable `cTV` that is `pTV` cut into the intervals `(0,.25]`, `(.25,.5]`, `(.5,.75]` and `(.75,1]`. Plot `Sales` *versus* `Total` and color points by `cTV`. Use the appropriate scale function to set the title of the legend to "Proportion TV ads".

Question 7 (7 marks)

A data frame called `Credit` contains information the average credit card balance (`Balance`) of 400 people. The other variables in the data frame are `Age`, `Gender` and `Married`.

a. (1 mark) The variable `Married` is a factor with levels "No" and "Yes". Change these levels to "Unmarried" and "Married".

b. (2 marks) Calculate the sample size, median of `Balance` and SD of `Balance` for each combination of the factors `Gender` and `Married`.

c. (2 marks) Sort `Credit` on (i) `Gender`, (ii) `Married` within `Gender` and (iii) `Age` within `Gender` and `Married` and rearrange the variables so that they are in the order `Gender`, `Married`, `Age` and `Balance`.

d. (2 marks) Plot `Balance` *versus* `Age` with a grid of panels on your graph for the combinations of `Gender` and `Married`.

Appendix: Useful functions not on the cheatsheets

- `cut(x,breaks)` cuts a vector `x` according to `breaks`, where `breaks` is either a numeric vector of cut points or a single number of intervals to cut `x` into.
- `sample(x,size,replace=FALSE)` draws a sample of size `size` from `x` either with or without `replace`ment. If `size` is not specified it is taken to be the length of `x`.
- `replicate(n,expr)` executes `expr` `n` times