STAT 133 FINAL

NAME:

SID:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others.

You Signature:

SHORT ANSWER

1. For each of the following relational database terms, provide the letter of the matching term in R.

   ____ relation; ____ attribute; ____ tuple;

   (a) vector; (b) data frame; (c) matrix; (d) row; (e) column; (f) row name.

2. Which of the following are plain text files? Circle all that apply

   (a) an XML file; (b) a JSON file; (c) an XLSX file; (d) an HTML file;

   (e) an RDA file; (f) an Rmd file; (g) a FWF file

3. Write a simple expression to create a vector of 25 0s followed by 75 3s.

4. Consider the following subject lines of 6 emails. Take these 6 subject lines to be the entire corpus of documents.

   Activity Digest since 2:36PM for STAT 133 on Piazza
   Don't miss it! 40% off your order - ends in hours!
   Cookie Time 3:00pm in Dept lounge
   Re: second meeting
   Receive at least 20% off your purchase
   Now open: Women in science fellowship application

   What is the document frequency for the term *in*: _____

5. Write the binary representation of the decimal number 27.

6. The variable `myData` is a sample of 35 values from an unknown distribution. Write one line of R code to carry out a Monte Carlo simulation of the distribution of the median of `myData`. Use 2000 samples in the Monte Carlo.

7. Suppose the following expressions are executed at the command line in R.

```
x = 1:4
y = 12

funcSq = function(y) {
  return(y^2)
}

funcSS = function() {
  func3 = function() {
    return(3*funcSq(x))
  }
  x = 6

  func3()
}
```

Circle the return value from the call `funcSS()`

(a) 3 12 27 48
(b) 432
(c) 108
(d) 9 36 81 144
(e) none of the above

8. Consider the following function signature:

```
myFunc = function(data = NA, nrow, ncol, byrow = FALSE, dimnames = NULL)
```

and the following function call:

```
myFunc(2, byrow = TRUE, 3)
```

Complete the call frame for the function call:

Variable	Value
data	
nrow	
ncol	
byrow	
dimnames	

LONG ANSWER

9. The following XML document is not well-formed. Circle each error and explain what rule is not being met.

```
<catalog>
<plant>
<price>$2.79</price>
<zone>3</zone>
<NAME>Azaela</NAME>
</plant>
<plant>
<price type=discount>$1.49</price>
<zone></zone>
<Name>Geranium</name>
</plant>
</catalog>
<catalog>
<plant>
<price>$1.99 <zone>4 </price>
</zone>
<Name>Camelia</Name>
</plant>
</catalog>
```

10. Consider the following distance matrix for the points $P1$, ..., $P6$.

	P1	P2	P3	P4	P5	P6
P1	0	1	3	10	12	8.5
P2	1	0	6	8	11	13
P3	3	6	0	7	9	12
P4	10	8	7	0	2	5
P5	12	11	9	2	0	4
P6	8.5	13	12	5	4	0

(a) In the process of hierarchical clustering with these 6 observations, suppose that we have two clusters: $\{P1, P2\}$, $\{P4, P5\}$, and the remaining points are in their own clusters, i.e., there are 4 clusters. What is the next step in the agglomerative clustering process, assuming we are using single linkage clustering? That is, which two clusters are joined and what are the points in each of the three clusters?

(b) Using the same distance matrix, suppose that we are carrying out 3-NN classification. Which points are P4's nearest neighbors?

11. Write a function called *yLim* which takes as input the vectors x and y. This function returns a vector of length 2, which contains the values in y that occur in the positions of the minimum and maximum values of x. That is,

```
> x = c(100, 13, 1, 20)
> y = c(5, 7, 9, 0)
> yLim(x, y)
[1]  9  5
```

Since the minimum of x is in the 3rd element of x and the maximum of x occurs in its first position, the return value is a vector containing the 3rd and first values in y (which are 9 and 5, respectively).

Additionally, the input x is required, and y has a default value of x. If x is shorter than y, then a message should be issued but the computation is carried out. If y is shorter than x, then the function is terminated and a message is issued. You may assume that there is a unique minimum and maximum in x and there are no NAs in either x or y.

12. In the table below, a regular expression pattern appears in each row and a string appears in each column. Determine whether or not there is a match for a pattern in each string. If there is a pattern provide the *starting* and *ending positions* of the match. If there is no match, provide a -1. If there is more than one match, provide only the information for the first match.

	`"abcdefg"`	`"abcs!"`	`"ab abc "`	`"abc, 123"`
`"abc*"`				
`"[^[:blank:]]+"`				
`"ab.*c"`				
`"[a-z1,9]+"`				

13. Consider the following relational database for a company. There are four tables (shown below). The tables represent the following entities:
Employee - a record for each current and former employee with: name, address & public ID (called UID).
Security - a record containing each employee's public ID and secure ID.
Payroll - secure ID, salary, position, and department for each current employee.
Vehicle - a record for each vehicle registered with the parking office. An employee may register more than one vehicle with the office or may not have any vehicles registered.

Employee

Name	UID	Address
John	101	100 Elm St
Sara	102	200 Maple Ave
Susan	201	100 Chestnut St
Dave	202	102 Fir Dr

Payroll

SID	Salary	Position	Dept
1	30000	AA2	Stat
2	37000	MS1	Math
3	40000	AA1	Stat

Security

UID	SID
101	1
102	2
201	3
202	4

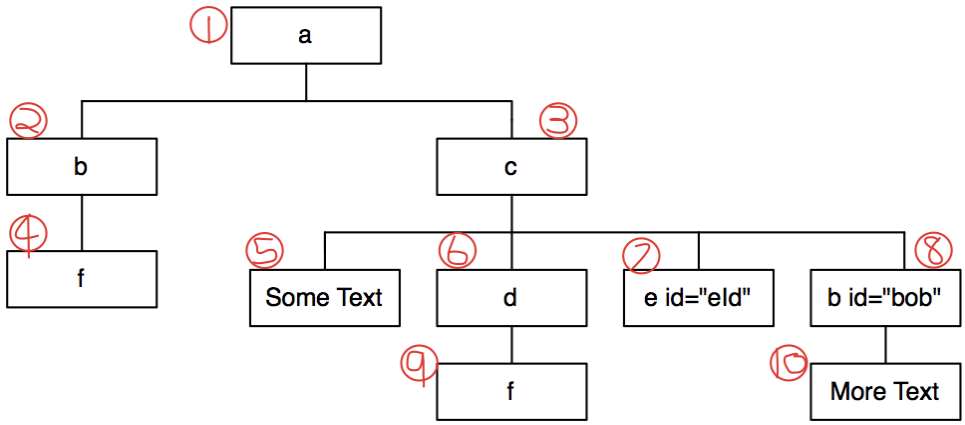
Vehicle

UID	Make	Model	License
101	Toyota	Matrix	2RPM888
101	Honda	Accord	2RPM777
201	Ford	Fiesta	2RPM666
201	Ford	Explorer	2RPM555
201	Subaru	Forester	2RBY447

(a) Write a SELECT statement to retrieve the secure ID and department of all employees earning less than $35,000.

(b) Write a SELECT statement to retrieve the public ID and number of vehicles registered with any employee who has a vehicle registered with the parking office.

14. Below is a tree representation of an XML document. The tree has 10 nodes, which we have been numbered 1 through 10. Two of these are text nodes: one containing "Some Text" (labeled #5), and the other "More Text" (labeled #10). In addition, some of the nodes have attributes, e.g. #7 is the tag <e id='eId'/>.



For each of the following XPath expressions, provide the numbers for the nodes which are located by the expression. If no nodes match, say NULL.

(a) `/a/b`

(b) `//f`

(c) `//b/..`

(d) `//c//f`

15. Suppose we have a function, f(), in R, and we know it is increasing (possibly flat in places, i.e., non-decreasing) on the interval from `A` to `B`. We also know that `f(A)` is negative and `f(B)` is positive. Write a function, called find0() that finds the value `x0` such that `f(x0)` is 0. The function f() may have a flat spot at 0, and in this case, `x0` should be the smallest `x` such that `f(x)` is 0.

    The input arguments to find0() are: two required parameters *`A`* and *`B`*, and one optional argument, *`tol`* which has a default value of 0.001. The return value of your function is `x0` or a value that is within `tol` of it. Be careful that your examination of `f(x)` takes into account the machine tolerance. That is, the function call `f(x)` may not return the exact analytic value.

    Use the bisection method/binary search approach to find `x0`.