Stats 506 F23 Midterm

Name: ____

UMID: ____

(Please write your name on the **back** of the test as well.)

Question	Out of	Score
1	18	____
2	14	____
3	20	____
4	20	____
5	20	____
6	20	____
Total	112	____

Instructions:

- Complete all problems.
- Your code does not have to be completely syntactically correct.
- If you make any assumptions, please state them.
- If you do not know the name of a function/command/statement, describe what it does instead.
- If you use additional pages, staple them to your test and write you name on the back of the last sheet.

You have 1 hour and 20 minutes to complete the exam.

Problem 1 - 18 points

For each of the following snippets of code, determine the value of `q` after execution.

a.

```
x <- 1:10
y <- x[which(x %% 2 == 1)]
q <- sum(y)
```

`q =`

b.

```
x <- c(1, 5, 10, 15, 20)
y <- x*c(2, 1)
q <- y[1] + y[5]
```

`q =`

c.

```
x <- matrix(c(1, 3, 2, 2, 2, 1), nrow = 3, byrow = FALSE)
q <- apply(x, 1, sum) - apply(x, 2, sum)
```

`q =`

d.

```
f <- function(x) {
  ifelse(x > 2, f(x/2), x)
}
q <- f(10)
```

`q =`

e.

```
tibble(a = 1:4, b = c(1, 2, 1, 2)) %>%
  mutate(c = a - b,
         d = c + b) %>%
  group_by(b) %>%
  summarize(e = mean(d)) %>%
  ungroup %>%
  arrange(-e) %>%
  filter(row_number() == 1) %>%
  select(e) %>%
  as.numeric -> q
```

q =

f.

```
end = 1
goal = 30
total = 0
while ( total <= goal ) {
  for ( i in 1:end ) {
    total <- total + i
  }
  end <- end + 1
}
q <- total
```

q =

Problem 2 - 14 points

Assume I have the following data.

first	last	year	age	owns_home
Afonso	Pelletier	2012	36	0
Putu	Contreras	2013	24	1
Lisbet	Gruber	2007	62	1
Catrine	Smit	2020	18	0
Vilté	Proudfoot	2017	42	1

The data is stored in “work.housing”. Write SAS code to

a. Sort the file by “age”
b. Generate a variable “birthyear”
c. Compute the average birthyear amongst home owners and non-homeowners.

(Hint: One or more of these may require multiple `data` and/or `proc` steps.)

Problem 3 - 20 points

For each of the following `regress` calls from Stata, produce R code to replicate the model. You may assume the data is stored in equivalent fashion in both software, with the R data being stored in a `data.frame` named “dat”, with all columns being numeric. **You should write only one line of R code.**

a.

```
regress y x1 i.x2 c.x3##c.x4
```

b.

```
regress y i.x1##i.x2##x3, nocons
```

c.

```
regress y x1 c.x1#c.x2
```

d.

```
generate z = log(x3)
regress y c.x1#c.x1 x2 z
```

Complete each of the following pieces of Stata code according to the comments.

e.

```
* Loop over each of the following binary variable currently stored as
* 1/2 and convert to 0/1
foreach var of varlist bin1-bin20 {



}
```

f.

```
regress depress i.race##c.age
* Draw an interaction plot with `age` on the x-axis at values 20
* through 80 by 10s, plotting line per `race`


marginsplot
```

g.

```
mean x1 x2, by(group)
* calculate squared difference between `x1` and `x2`


* replace the data with a dataset that contains the mean squared
* difference between `x1` and `x2` per `group`
```

h.

```
mata:
X = st_matrix("x")
// find the largest value on the diagonal of `X`



st_matrix("q", Q)
end
matrix list q
```

Problem 4 - 20 points

Fill out the following table. Place an X in each cell which the regular expression will match the string. Leave the remaining cells blank.

	`[Aa]*[a-zA-Z]+.{2}o$`	`^..?...$`	`^[^AS]..?[aeiou]`	`(.).\\1`
Adedayo				
Dinah				
Frida				
Liam				
Marlyn				
Milko				
Mirka				
Rupert				
Rustik				
Sebastian				

This second table is for *scratch work only.* **ONLY THE ABOVE TABLE WILL BE GRADED.**

	`[Aa]*[a-zA-Z]+.{2}o$`	`^..?...$`	`^[^AS]..?[aeiou]`	`(.).\\1`
Adedayo				
Dinah				
Frida				
Liam				
Marlyn				
Milko				
Mirka				
Rupert				
Rustik				
Sebastian				

Problem 5 - 20 points

Let data be the following table:

```
  x y q z
1 2 7 1 1
2 9 8 1 2
3 7 5 2 3
4 3 5 2 4
5 1 8 2 5
6 6 4 1 6
```

and data2 be

```
  z p
1 1 4
2 7 8
```

For each of the following SQL queries, what will the output table be? Be sure to provide column names as appropriate.

a.

```
SELECT *
  FROM data
 WHERE y > 7
```

b.

```
SELECT x, x - 1 AS x1
  FROM data
 LIMIT 2
```

c.

```
SELECT q, sum(x) AS xx
  FROM data
 GROUP BY q
HAVING xx < 15
```

d.

```
SELECT x, p
  FROM data AS d1
  LEFT JOIN data2 AS d2 ON d1.z = d2.z
```

e.

```
SELECT x, p
  FROM data AS d1
 RIGHT JOIN data2 AS d2 ON d1.z = d2.z
```

Problem 6 - 20 points

Suppose we have an unfair coin. We know that $P(\text{heads}) = p$, $P(\text{tails}) = q$ and $p \neq q$ but we do not know p or q. One way to obtain a fair result from such a coin is to carry out the following procedure. Flip the coin twice, recording the first and second flip in order as a pair. If the pair has different results, return the result of the first flip. If the pair has the same results, repeat the procedure. (This is called the "Von Neumann Extractor".)

a. Write an R function that takes in a proportion p for a biased coin, and returns a single heads or tails result using the procedure above. Name your function "vonneumann". You do not need to check your input; you can assume the proportion is a valid number strictly between 0 and 1. You can use the `rbinom` function to generate the random coin flips, it takes in two arguments: `n` and `p`. The output should be a binary where 0 represents tails, and 1 represents heads. Do not worry about whether your function finishes in finite time.

b. Using your `vonneumann` function, write a Monte Carlo simulation to demonstrate that it does in fact produce fair results. Start with the following parameters:

   ```
   p <- .7
   reps <- 10000
   ```

   Your code should produce an estimate of the proportion of heads in `reps` loops of the procedure.

Name: ____

Please do not write anything else on this side of this page.