

# STA312H5S Tutorial 5

Matthew C. Scicluna

University of Toronto Mississauga

February 24, 2015

# What We Are Going to Cover

A bit more information about this

Today we are going to discuss

- ▶ Some common mistakes in assignment 2
- ▶ How to write better reports in the future
- ▶ Project 2 details

## Assignment 2 Common Mistakes

```
> cbind(table(racecluster1),table(racecluster2))  
      [,1] [,2]  
Black    157   31  
Hispanic   74   14  
Other      2    0  
White    200   18  
White      1    0  
#Didn't merge the two white categories together
```

## Assignment 2 Common Mistakes

	x	y
Anderson	2	1
Anderson	1	0
Aransas	1	0
Atascosa	1	0
Bailey	0	1
Bastrop	0	1
Bee	1	0
Bell	2	0
Bexar	35	4
Bowie	4	0
Bowie	1	0
Brazoria	3	1
Brazos	10	1
Brazos	1	0
Brown	0	1

Is this really the most effective way of presenting the county data??

## Assignment 2 Common Mistakes

Docs																		
Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
and	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ask	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
can	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
caus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
come	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
death	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
declin	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0
done	0	0	0	2	0	0	1	0	0	1	0	0	0	0	0	0	0	0
dont	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0
famili	1	0	2	1	0	0	5	0	0	2	0	0	1	1	0	1	1	0

## Assignment 2 Common Mistakes

forgiv	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
friend	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	1	0
get	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
give	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
god	0	0	0	2	0	0	5	0	0	1	0	0	0	1	0	0	1	0
heart	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
hope	0	0	3	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
jesus	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
just	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
keep	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
know	0	0	1	0	0	1	1	0	0	0	0	0	0	3	0	1	1	0
last	0	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0

## Assignment 2 Common Mistakes

let	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
life	0	0	0	0	0	1	2	0	0	0	0	0	1	0	0	1	0
like	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0
lord	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
love	3	1	10	8	3	6	9	0	0	3	0	0	1	2	3	0	4
make	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
offend	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0
one	0	0	2	0	0	0	0	0	0	1	0	0	1	0	0	0	0
pain	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0
peac	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0
peopl	0	0	0	0	0	0	3	0	0	1	0	0	0	3	0	0	0
readi	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0
say	0	0	0	0	0	1	6	0	0	0	0	0	0	1	0	0	0
see	0	1	1	0	0	0	2	0	0	0	0	0	0	0	0	1	0

## Assignment 2 Common Mistakes

sorri	0	0	1	0	1	1	2	0	0	0	0	2	0	0	0	0	0
statement	0	0	0	0	0	1	0	1	1	2	1	1	0	0	0	0	0
support	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0
take	0	0	3	0	0	0	0	0	0	0	0	0	2	1	0	0	1
tell	2	0	0	3	0	0	1	0	0	0	0	0	1	0	0	0	0
thank	0	0	2	1	0	0	4	0	0	2	0	0	0	0	0	0	2
that	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
the	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
thing	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
this	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0
want	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0	1	1
warden	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
will	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
yall	0	1	0	3	0	0	3	0	0	0	0	0	0	0	0	0	3
year	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
yes	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
you	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0



## Assignment 2 Common Mistakes

#This goes on for about a hundred or so slides,  
so why print it when it doesn't add to the report?

Docs

Terms	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
and	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ask	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0
can	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1	1
caus	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
come	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0
death	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
declin	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
done	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
dont	0	0	5	0	1	0	0	1	0	1	1	2	0	1	1	1
famili	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0
forgiv	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0
friend	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
get	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0

## Assignment 2 Common Mistakes

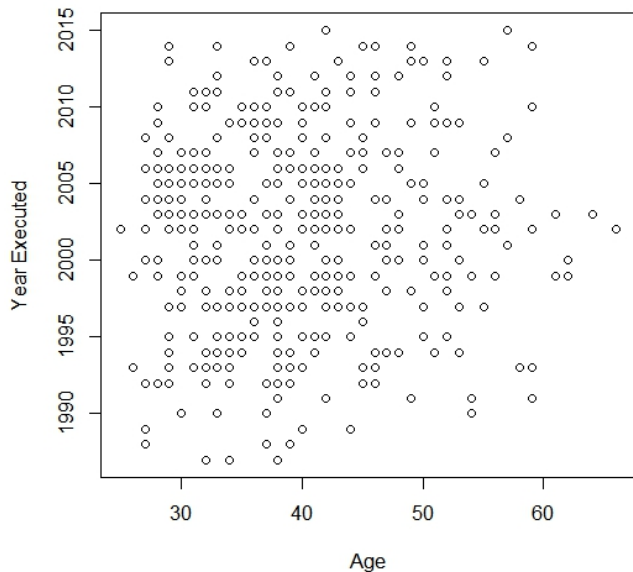
```
> summary(z)
```

year	age	cluster
Min. : 8	Min. :24.00	Min. :0.0000
1st Qu.:1998	1st Qu.:33.00	1st Qu.:0.0000
Median :2002	Median :38.00	Median :0.0000
Mean :1998	Mean :39.38	Mean :0.1268
3rd Qu.:2007	3rd Qu.:44.00	3rd Qu.:0.0000
Max. :2015	Max. :66.00	Max. :1.0000

#Just printing a summary like this wasn't enough  
to be able to conclude anything...  
You had to communicate what you found interesting...

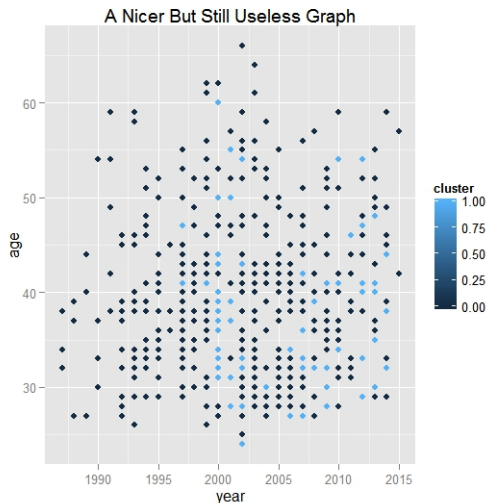
## Assignment 2 Common Mistakes

**A Totally Useless Graph**

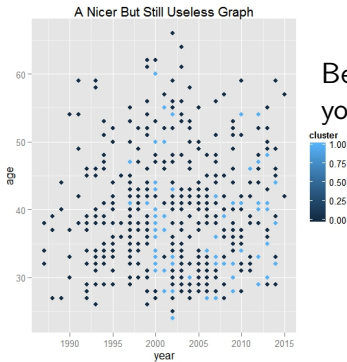


## Assignment 2 Common Mistakes

```
> p=qplot(year,age,data=data.frame(zz),color=cluster)
> p+labs(title="A Nicer But Still Useless Graph")
```



# Assignment 2 Common Mistakes

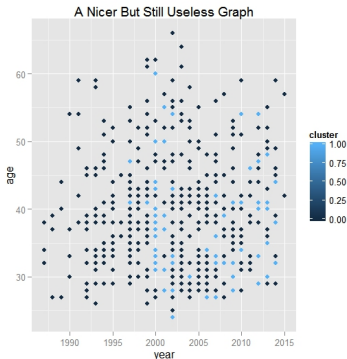


Before adding a graph to your report you should consider a few things:

- ▶ Were the results significant?
- ▶ Were the results interpretable?
- ▶ Were the results interesting?
- ▶ In most cases the answer was **NO**

## Assignment 2 Common Mistakes

Lets think about how we could interpret this graph



- ▶ We see this is relating age with year of execution
- ▶ This could be answering something like "Were older inmates executed in the past?"
- ▶ However it is extremely unlikely that the clustering algorithm would make clusters based on the relationship between age and year of execution
- ▶ Hence this is a useless graph and you should not have included it in assignment 2.

# How to Write a Good Report

- ▶ Merge redundant categories and eliminate sparse ones to tidy up the data
- ▶ Perform statistical tests based on what you think might be true. Have a reason for thinking something may be so before testing for it.
- ▶ Only perform statistical tests that are interpretable
- ▶ This means use statistical tests in such a way that if you got results that were significant, that you could derive a conclusion from them.
- ▶ For example, testing equality of means from each cluster makes sense, but testing to see if the interactions between variables changes between clusters has no interpretation.

# How to Write a Good Report

- ▶ If you find something interesting and interpretable, only then should you bother to display it graphically.
- ▶ You shouldn't include results that were not significant unless you had good reason to suspect otherwise.
- ▶ If you do include a result you must communicate why the result is significant or insignificant (via a statistical test) and why the result is interesting.
- ▶ This basically boils down to including in your report results which...

**answer the research question and exactly nothing else**



# How to Write a Good Report

- ▶ Remember, nobody cares if eating lots of chocolate is uncorrelated to having red hair (it is obvious)
- ▶ Also, nobody cares if if eating lots of chocolate is correlated to having red hair, since clearly there must be a confounding variable, or type 1 error being committed (there is no reason for thinking it is so in the first place!)
- ▶ Finally, nobody cares if if eating lots of chocolate is correlated to having red hair, only when the subject is male. This result is completely uninterpretable (it is too convoluted)
- ▶ The more complex the statistical analysis, the better the reason you should have for expecting to see it in the first place!

## Project 2 Details

Now you will go into groups and discuss project 2, I would like to see you answer these specific questions:

- ▶ How are you going to implement a Kevin Bacon counter for part 2 of the analysis?
- ▶ How you would address the sparsity of the budget data in your logistic regression models in part 3 of the analysis.
- ▶ What ideas do you have for part 4 of the analysis.

I will come around to discuss these with you, along with any other questions you may have about the code provided.