Course 1 Section 3.4 - Aesthetics

Jiaying Wu

20/09/2020

```
library(tidyverse)
```

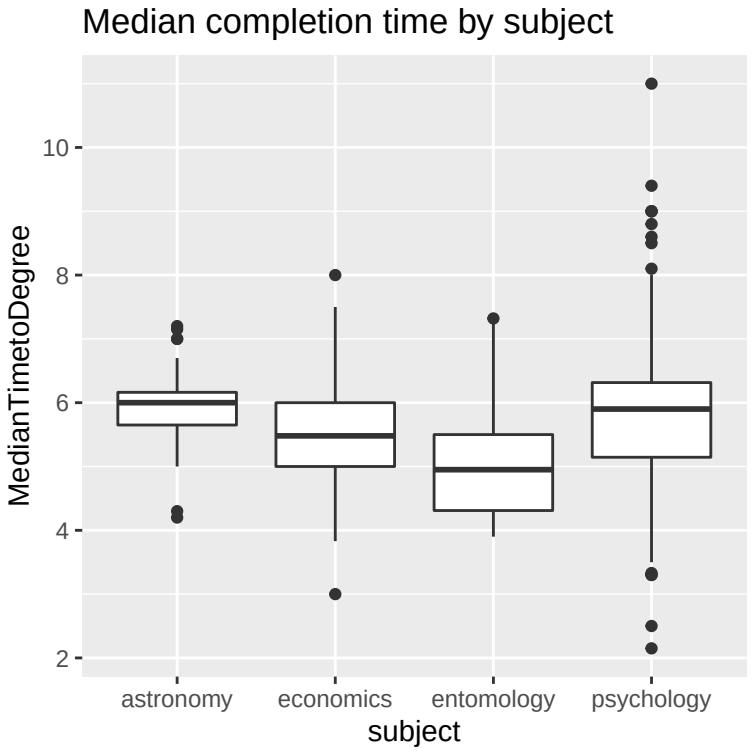
```
grad_programs <- read_csv("https://raw.githubusercontent.com/datascienceprogram/ids_course_data/master/g
grad_programs
```

```
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGrants PctCompletion MedianTimetoDeg~
##    <chr>   <chr>     <dbl>     <dbl>        <dbl>         <dbl>            <dbl>
##  1 econom~ ARIZ~      0.9       1.57         31.3          31.7              5.6
##  2 econom~ AUBU~      0.79      0.64         77.6          44.4              3.84
##  3 econom~ BOST~      0.51      1.03         43.5          46.8              5
##  4 econom~ BOST~      0.49      2.66         36.9          34.2              5.5
##  5 econom~ BRAN~      0.3       3.03         36.8          48.7              5.29
##  6 econom~ BROW~      0.84      2.31         27.1          54.6              6
##  7 econom~ CALI~      0.99      2.31         56.4          83.3              4
##  8 econom~ CARN~      0.43      1.67         35.2          45.6              5.05
##  9 econom~ CITY~      0.35      1.06         38.1          27.9              5.2
## 10 econom~ CLAR~      0.47      0.7          24.7          37.7              5.17
## # ... with 402 more rows, and 9 more variables: PctMinorityFac <dbl>,
## #   PctFemaleFac <dbl>, PctFemaleStud <dbl>, PctIntlStud <dbl>,
## #   AvNumPhDs <dbl>, AvGREs <dbl>, TotFac <dbl>, PctAsstProf <dbl>,
## #   NumStud <dbl>
```

Is there much variation in the median completion time of a graduate program? What about by subject?

```
grad_programs %>%
  ggplot(aes(x = subject, y = MedianTimetoDegree)) +
  geom_boxplot()+
  ggtitle("Median completion time by subject")
```

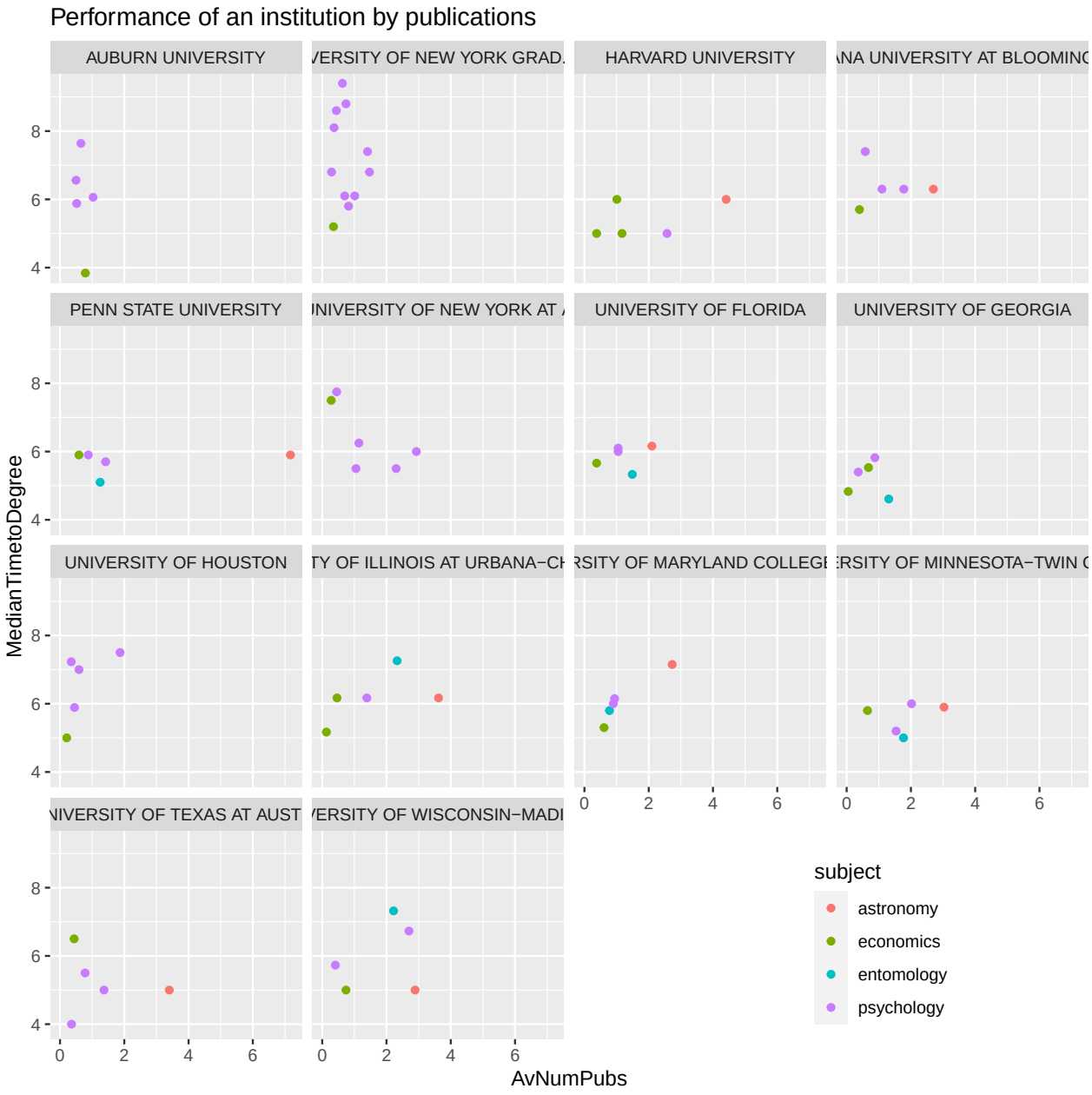
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



Does the performance of an institution (measured by the number of citations or publications) appear to have an influence on completion time? Is this different if considered by subject?

```
# first arrange the data by count and exclude any schools that have fewer than five programs
counts <- count(grad_programs, Inst, sort = TRUE)
counts <- filter(counts, n >= 5)
list_top_inst <- pull(counts, Inst)
top_inst = filter(grad_programs, Inst %in% list_top_inst)

# Publications with subjects as colours
ggplot(data = top_inst, aes(x = AvNumPubs,y = MedianTimetoDegree, colour = subject)) +
  geom_point() +
  facet_wrap(Inst~., nrow = 4) +
  theme(legend.position = c(0.8, 0.1))+
  ggtitle("Performance of an institution by publications")
```




```
# Citations with subjects as colours
ggplot(data = top_inst, aes(x = AvNumCits, y = MedianTimetoDegree, colour = subject)) +
  geom_point() +
  facet_grid(Inst ~ subject)+
  theme(legend.position = "bottom")+
  ggtitle("Performance of an institution by citations")
```


Performance of an institution by citations
astronomy
economics
entomology
psychology
MedianTimetoDegree
AvNumCits
subject
astronomy
economics
entomology
psychology