HW 4

Cory Costello

October 24, 2017

1. Tidy the data

```
library(rio)
library(janitor)
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(stringr)
library(knitr)
project_reads <- import("Project_Reads_Scores.csv") %>%
  # clean names because some of these variable names are not ideal
  clean_names()%>%
  # remove columns 5 to 9
  select(-(5:9))

# Going to rename the unit_5_6 because that will cause some problems later on
colnames(project_reads)[13:14] <- c("unit_56_score", "unit_56_percent")
# going to do the same thing for the total ones, since they also follow a different
# naming pattern
colnames(project_reads)[19:20] <- c("unit_total_score", "unit_total_percent")

project_reads_tidy <- project_reads %>%
  # tidy data such that each row is a score & percentage on a unit
  # use gather, call the new columns variable and score, and make sure
  #to tell gather not to gather the first 4 columns
  gather(variable, score, -(1:4)) %>%
  separate(variable, c("elim", "unit_num", "scale")) %>%
  select(-elim) %>%
  # changing unit 56 back to 5/6
  # Note sure that's ideal, because if you wanted to spread later, it'd be a a problem
  # But I can't think of a better name
  mutate(unit_num = recode(unit_num, "56" = "5/6")) %>%
  # Okay, so there are some rows that correspond to all
  # students at a site. Going to remove that in a potentially jenky way
  # separate student_id into two variables: site and id_num,
  # since student_id's are in the format "site #", except the all students
  # one that says "All_students"
  separate(student_id, c("site", "id_num")) %>%
  # Now remove the rows corresponding to all students,
```

```
  # by removing the rows where site contains all (which is the first part of the
  # student_id string for those total rows)
  filter(site !="All") %>%
  # Now re-unite the two parts of student id
  # this will put a '_' in between the parts, where it used to be a space
  # I actually like this better, so I'm going to leave it
  unite(student_id, c("site", "id_num"))
```

```
## Warning: Too many values at 48 locations: 46, 47, 48, 94, 95, 96, 142, 143,
## 144, 190, 191, 192, 238, 239, 240, 286, 287, 288, 334, 335, ...
```

```
project_reads_tidy$score <- as.numeric(parse_number(project_reads_tidy$score))
```

2. Summary Table

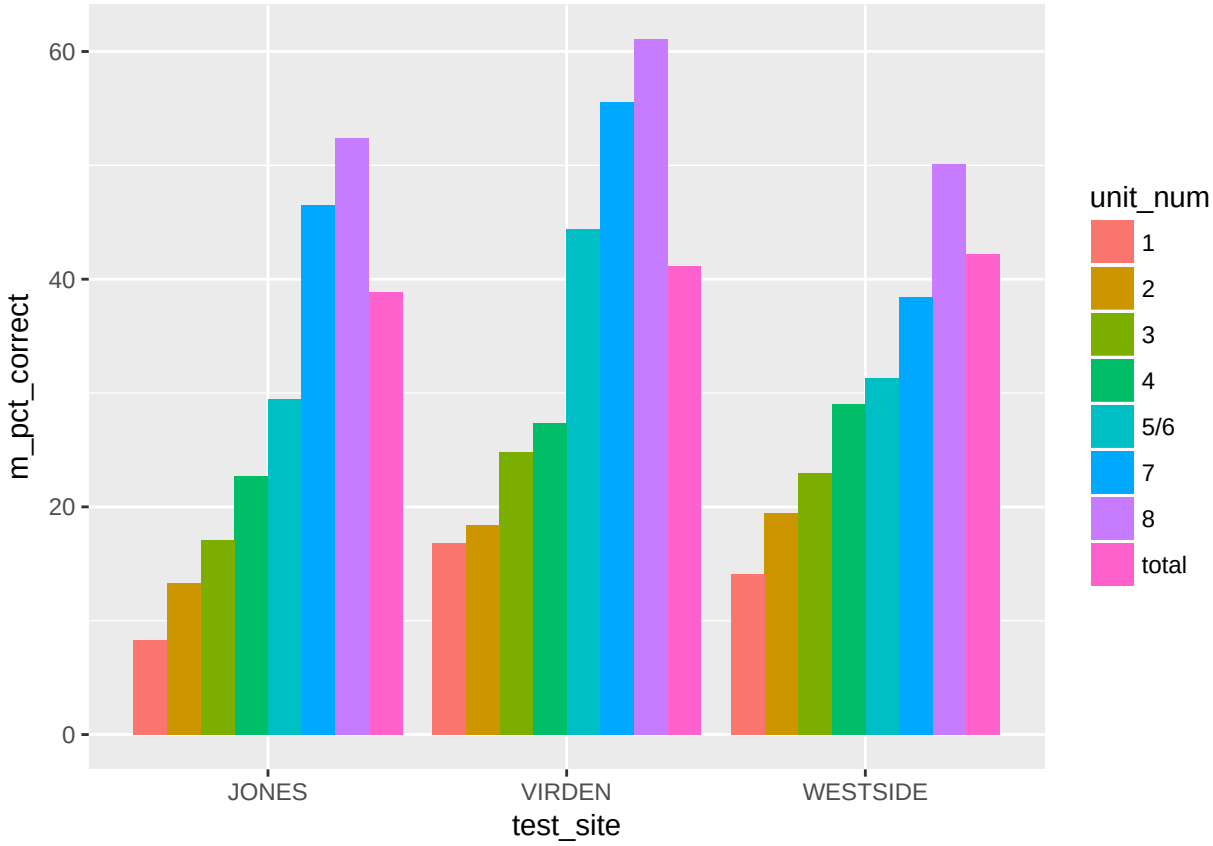
```
project_reads_summary<-project_reads_tidy %>%
  group_by(test_site, unit_num) %>%
  filter(scale == "percent") %>%
  summarize (m_pct_correct = mean(score, na.rm = TRUE))

kable(project_reads_summary, digits = 2)
```

test_site	unit_num	m_pct_correct
JONES	1	8.27
JONES	2	13.33
JONES	3	17.07
JONES	4	22.67
JONES	5/6	29.47
JONES	7	46.53
JONES	8	52.40
JONES	total	38.87
VIRDEN	1	16.80
VIRDEN	2	18.40
VIRDEN	3	24.80
VIRDEN	4	27.40
VIRDEN	5/6	44.40
VIRDEN	7	55.53
VIRDEN	8	61.07
VIRDEN	total	41.13
WESTSIDE	1	14.13
WESTSIDE	2	19.47
WESTSIDE	3	22.93
WESTSIDE	4	29.00
WESTSIDE	5/6	31.33
WESTSIDE	7	38.47
WESTSIDE	8	50.13
WESTSIDE	total	42.20

3. Plot of summary table

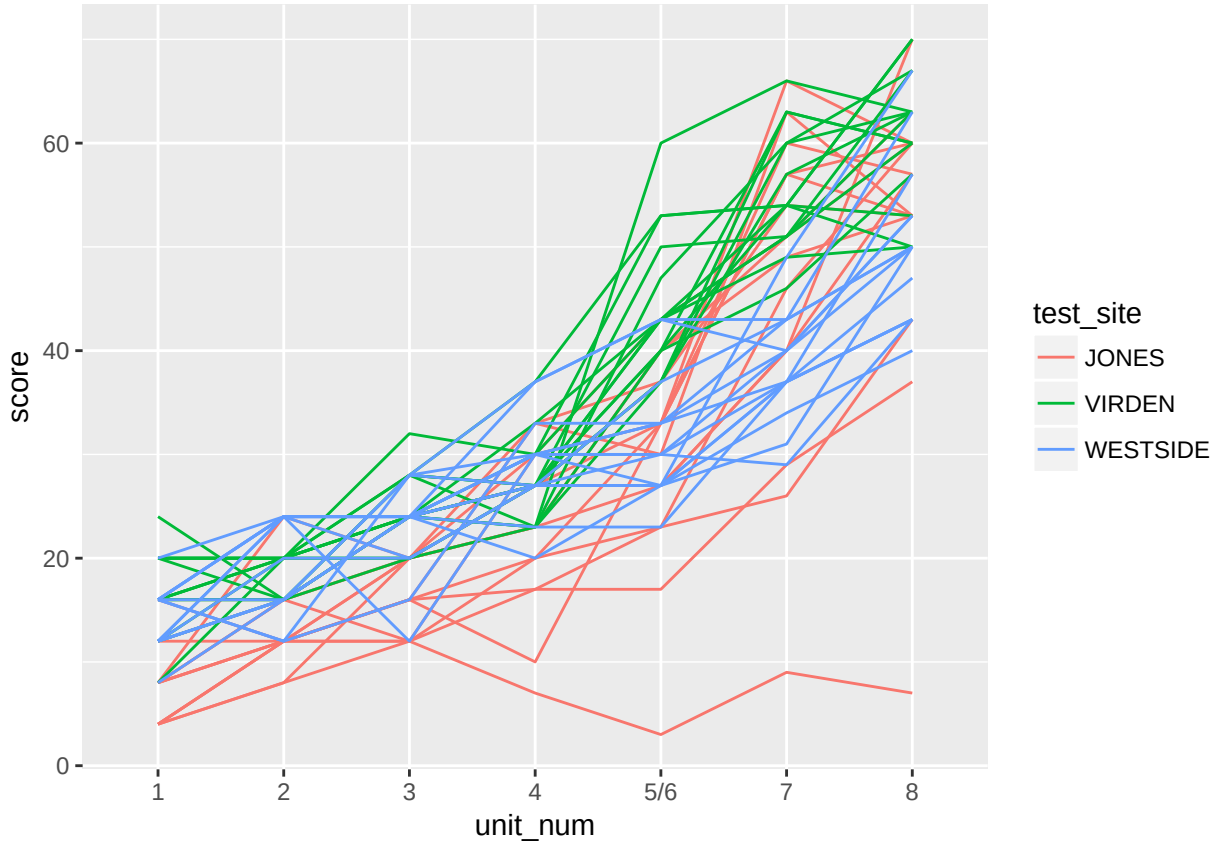
```
ggplot(project_reads_summary, aes(x = test_site, y = m_pct_correct, fill = unit_num)) +
  geom_bar(stat = "identity", position = "dodge")
```



Looks like each unit is better than the last (but the total is obviously not as high as the last unit, since it's being dragged down by the earlier ones).

4. Extra plot: score across units by student and site

```
project_reads_tidy %>%
  # filter for just the percent variable; also going to remove total
  # from this one, since I'm looking at trend across units
  filter(scale == "percent" & unit_num != "total") %>%
ggplot(aes(x = unit_num, y = score))+
  geom_line(aes(group = student_id, color = test_site))
```



This graph shows each student's progress across units, colored by school. I would have added a regression line too, with geom_smooth, but the unit 5/6 screws that up, and I'm not sure what the best solution is (could recode all of the factors maybe so that it's unit 0-7, could code 5/6 as 5.5 and leave the rest the same; the best solution would depend on what this 5/6 unit is, which I don't currently know).

BTW, looks like something isn't going so hot at Jones; they start off low, looks like a couple students don't improve much, and one student is getting worse.