GroupDReport

Rana Barghout, Alex Liu, Daniel Njoo

12/15/2017

Introduction

We decided to investigate wh

No more than 1 page. Briefly describe the dataset you are studying and the research question(s) you aim to answer. Provide ample motivations to convince your audience of the relevance of your research question(s), i.e. why the question(s) you proposed is important, meaningful, or interesting.

Data

Dimensions

Our cleaned and wrangled dataset contains 1923 rows and 14 columns. Each observational unit corresponds to a single meal at the Valentine Dining Hall in the 3 academic years starting 2014, 2015, and 2016.

```
cleaned %>% dim()
```

```
## [1] 1922   14
```

```
cleaned$year %>% table()
```

```
## .
## 2014 2015 2016 2017
##  317  638  645  322
```

Variables

Our 14 available variables were:

```
cleaned %>% names()
```

```
##  [1] "date"            "count"           "event"
##  [4] "semester"        "type"            "tag75"
##  [7] "tag69"           "tag66"           "substitute_time"
## [10] "datetime"        "posix"           "weekday"
## [13] "year"            "semester_week"
```

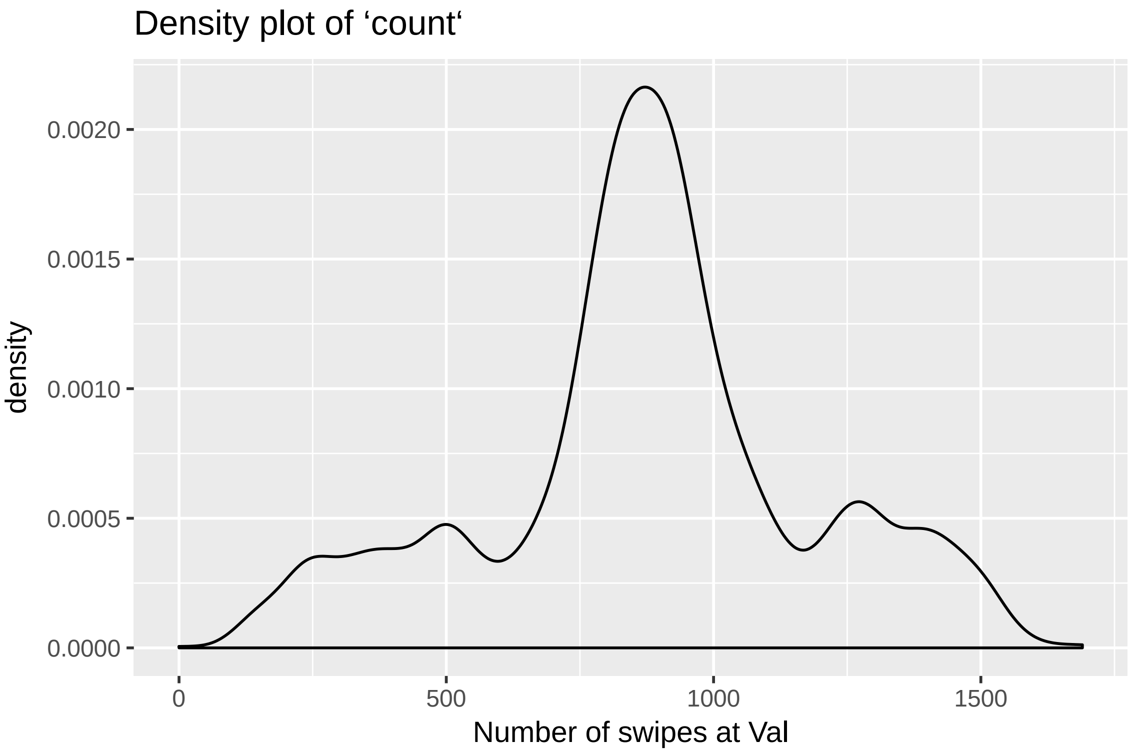
Response Variable

Our response variable was `count`, which is the number of swipes in any given meal period. A summary description of the stats, and a histogram of count is shown below. We see that the distribution is unimodal with a mean of 875 (roughly half the student body), and a standard deviation of 306, which is surprisingly high: a 1 standard deviation range runs from 569 to 1181 which covers 68% of all meals (68-95-99.7 rule). The mean is very close to the median which implies a lack of skew.

```r
favstats(cleaned$count)
```

```
##  min  Q1 median      Q3  max     mean       sd    n missing
##    0 748    876 1023.75 1690 875.5838 305.8551 1922       0
```

```r
cleaned %>%
  ggplot(aes(count)) +
  geom_density() +
  xlab("Number of swipes at Val") +
  ggtitle("Density plot of `count`")
```



Explanatory Variable

Our explanatory variables were:

- `event`, which refers to whether there was a specific event on the day such as Finals, or Family weekend.
- `semester`, either Fall or Spring
- `type`, either Breakfast, Lunch or dinner
- `tag75`, `tag69`, `tag66`, what menu was being offered for the meal, the number represents the string matching threshold used to cluster these labels
- `weekday`, Monday through Sunday
- `semester_week`, refers to how many weeks into a given semester a meal was, ranges from 1-17

Other variables `substitute_time`, `datetime`, and `posix` were variables we used in wrangling and did not feature in our model.

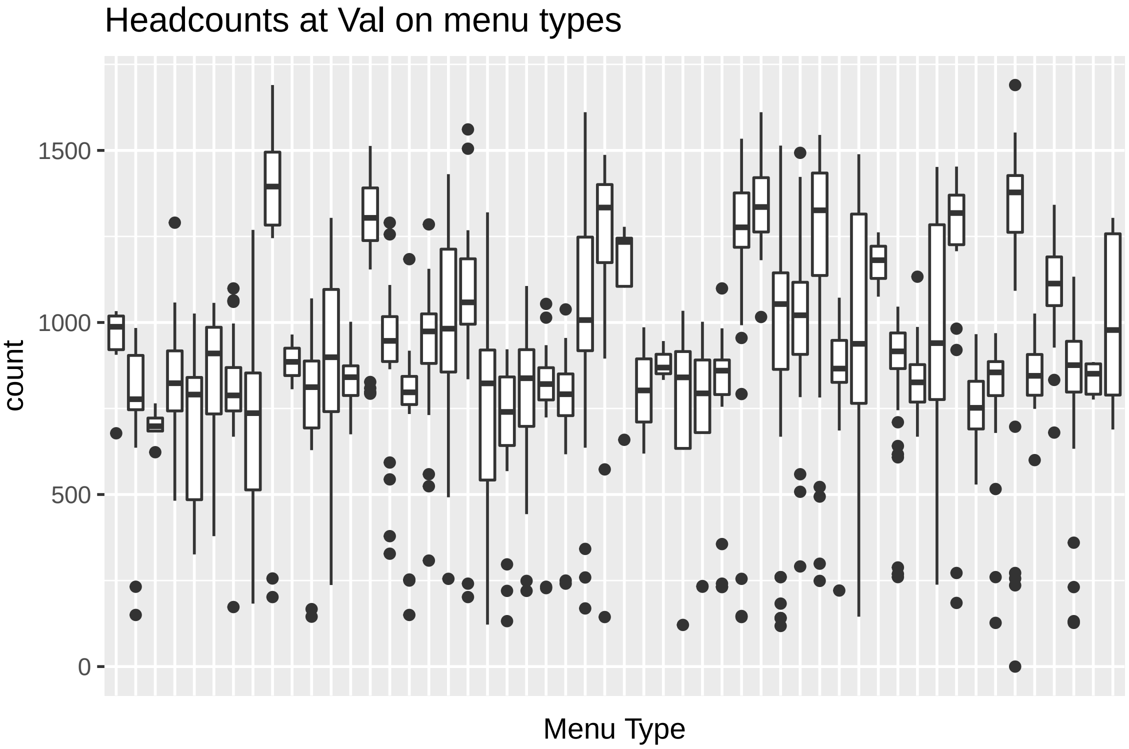
tag66, what was on the menu

tag66 represents what was on the menu that day, scraped from the school's website, the details of this will be explained in the next section. As there are 55 categories, and each is a long string, the labels of these are omitted in the following plot which is displayed to show the variance in count between menu types. In light of this variance, we used tag66 in our model.

- It's worth noting that the count of each these varied significantly, and the most common meal type was 'free cage' because this represented every breakfast and a few brunches (790 vs 612), and that fortunately burger was the second most common mean type because as we later found out, this meal had quite a strong positive effect on count.

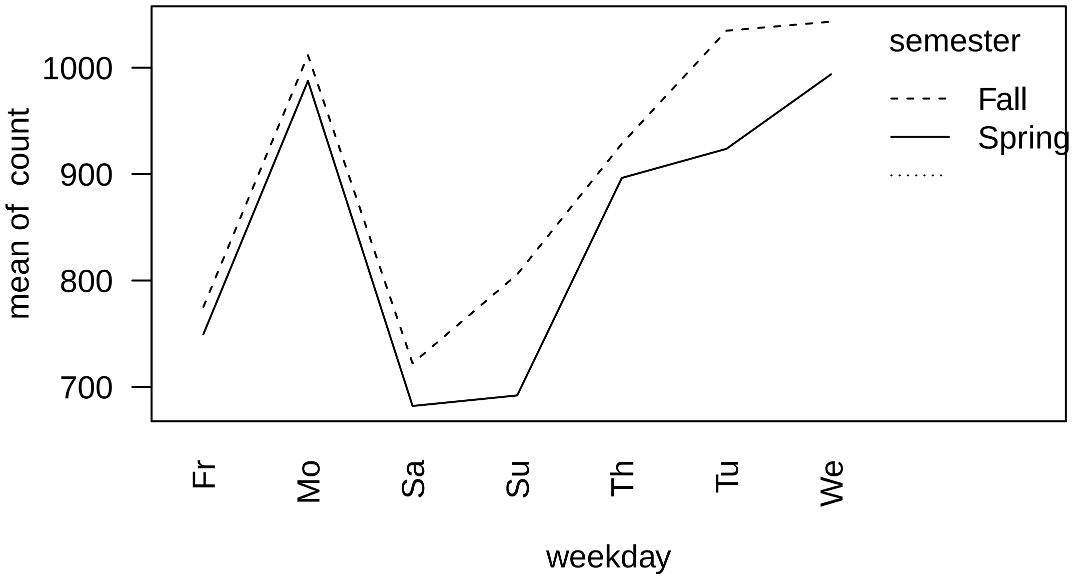
```
##             . Freq
## 1  free cage  789
## 2  hot burge   69
```

```
## .
## Breakfast    Dinner     Lunch
##       611       653       658
```



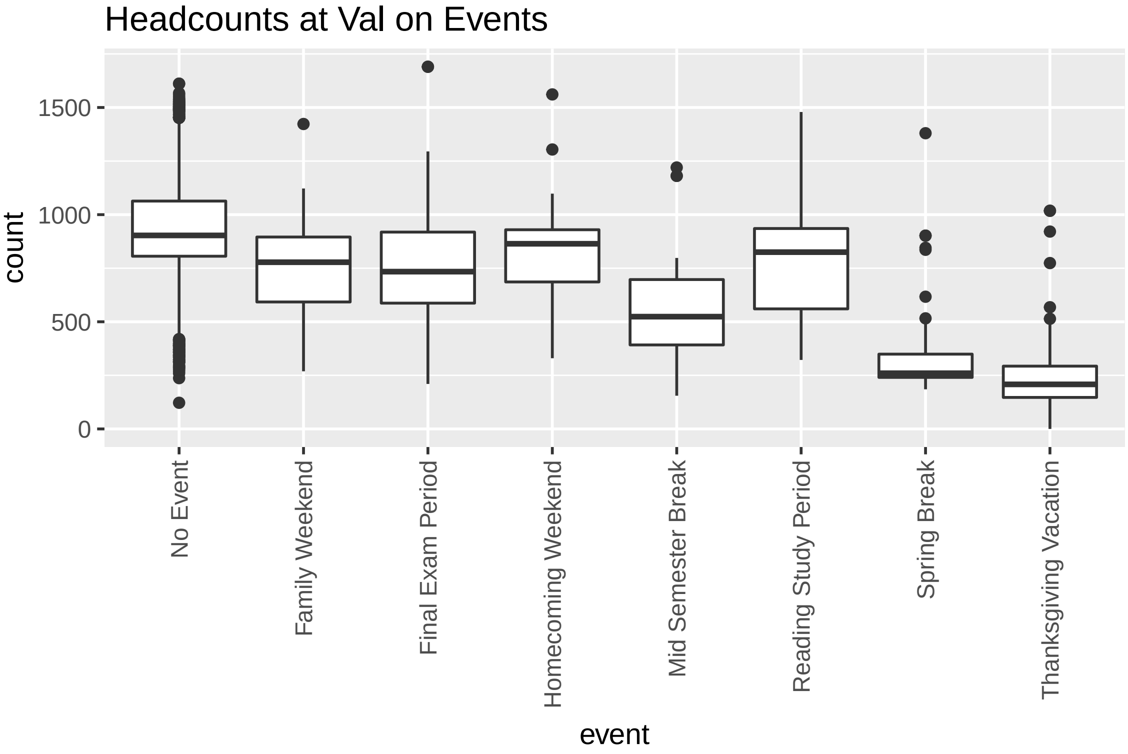
Semester*Weekday, the interaction between them

During our analysis we found a relatively significant interaction between semester and weekday, this is indicated by the relative absence of parallel slopes in this interaction plot where we notice Spring turnout to be consistently lower than Fall (perhaps due to study abroad?):



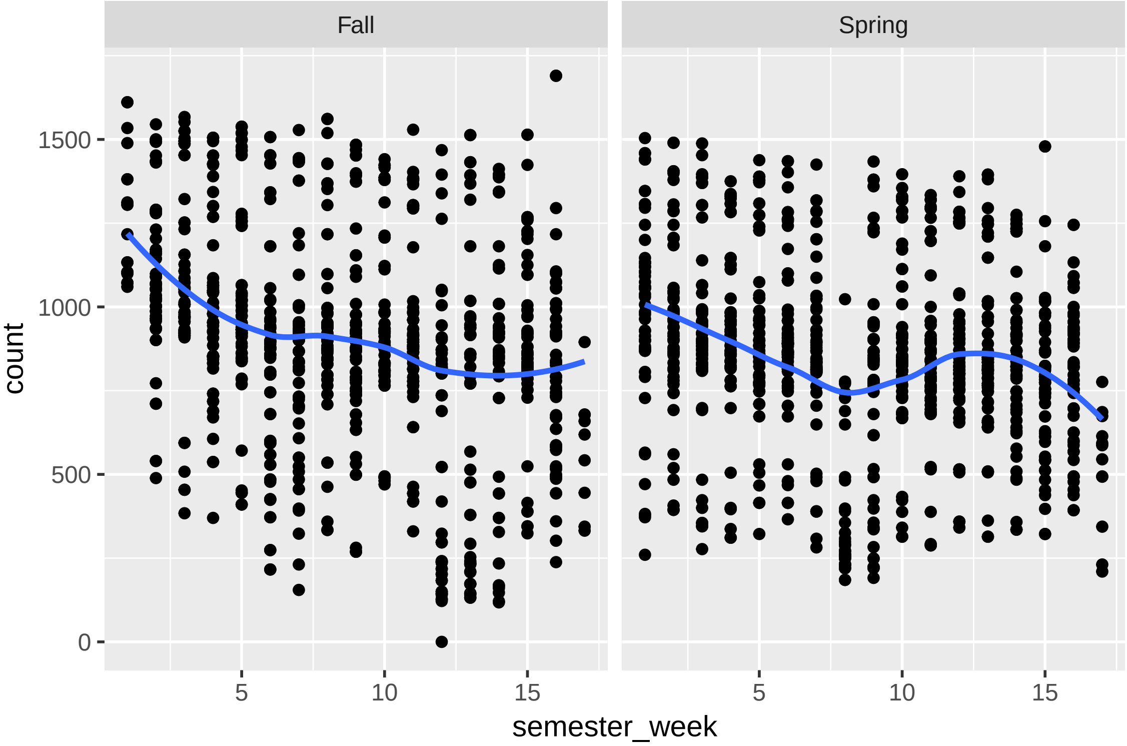
events, what was going on at Amherst

The next plot shows the variation in `count` when explored through `event`, what was going on on campus that day. `Event` encodes things such as Family Weekend or Thanksgiving Break. We noticed significant variation as evidenced by the vastly different medians and IQRs in boxplots.



semester_week, the week into the semester

`semester_week` also turned out to be a useful predictor, and this is evidenced by the variation in `count` that it is correlated with. Here we faceted it with `semester` because we found it intuitive that secondary modes occurred in difference places due to Spring Break (`semester_week`~=7) and Thanksgiving Break (`semester_week`~=12), but we did not end up pursuing a three-way interaction between `semester_week`, `semester`, and `weekday`, because this would have exponentiated the number of coefficients in our final and model and would have made interpretation more difficult.



Analysis

Present the analysis corresponding to your research question(s). For any statistical tech- niques that you used, check to see if the assumptions/conditions are appropriate and fully explain the results in the context of your research question(s). If you used methods that were not taught in our class (which is allowed!), please include a brief tutorial.

Conclusions

Summarize your findings. Do you have any concerns about the data? Could your analysis have benefited from other data sources? What, if anything, would you do differently if you could collect your own data or design your own experiment to address the same research question?

Appendix

While you may include some short bits of R code (less than 5 lines) in the main body of the report, it's often better to put your longer R code chunks here. Similarly, you might want to include only the most important plots and R output in the main body of the report, and leave all other relevant ones here; you can always refer to those in the Appendix using some sentence like "Please see Plot/Output XX in the Appendix". How to organize R output and plots to support your statements in the report is certainly one key learning point of this project.