S&DS 106 Final Project Report

Nathan Ahn, Erik Boesen, & Carmen Muniz-Almaguer

Due 12/22/2021

Introduction

In this analysis, we took advantage of a dataset on occupancy levels of Yale's 14 residential dining halls from spring 2019 through fall 2021 (though we were unfortunately forced to exclude data after March 2020 due to confounding variables related to COVID-19 service disruptions).

Our dining hall occupancy dataset contains over 5.6 million observations including timestamp, dining hall name, and a digit from 0-10 representing the fullness of that dining hall at the given point in time. This crowdedness is our outcome. Yale formerly provided this data through the Yale Dining website, from which it was collected constantly on the server of the Yale Menus app, which is run by one of our members.

We performed a great deal of data cleaning in order to make this observational data more straightforward to analyze. We started by deduplicating our data—removing sequential observations for the same dining hall with the same crowdedness rating. This allowed us to significantly slim down the size of our data file, enabling effective collaboration among our team via GitHub.

Then, we created an algorithm to compute summary data for crowdedness across entire meals, grouping observations and calculating weighted averages over time ranges according to historical dining hall opening times.

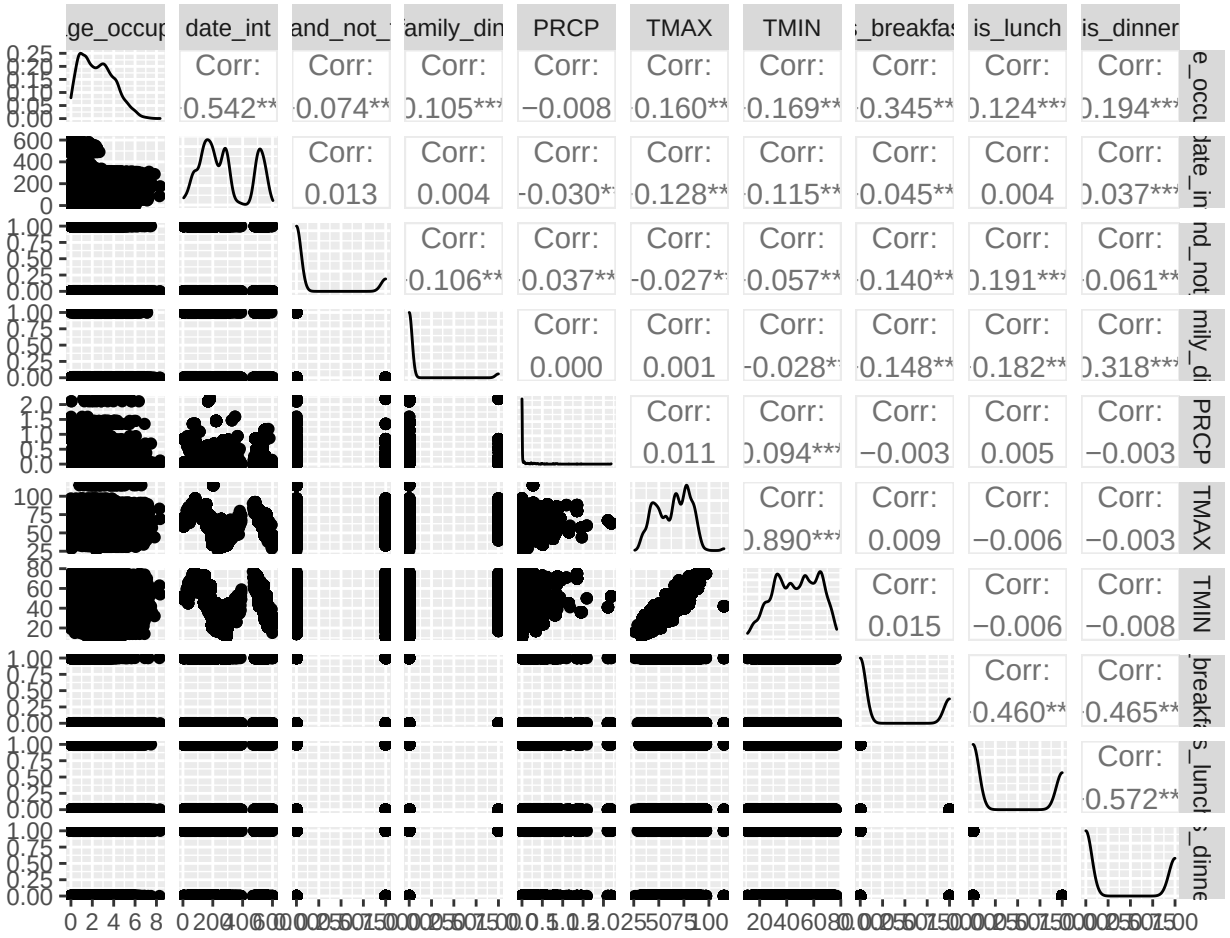
The second data set we used for our analysis is a New Haven daily weather data set from the National Oceanic and Atmospheric Administrations (NOAA). Through their website, we made a climate data online order. The weather data set begins on 01/02/2018 and ends on 12/05/2021. After filtering out some variables we did not find particularly useful for our analysis, our final data set contained information on: station name, geographic location (New Haven Tweed Airport), date, precipitation, max temperature, and min temperature.

Having this data, we set about to answer several questions:

- Is there a statistically significant difference between levels of dining hall attendance at different meals (breakfast, lunch, dinner)?
- Is there a statistically significant difference between attendance to dining halls on weekends vs weekdays?
- Is there a statistically significant decrease in dining hall attendance during "Family Dinner"? (i.e. Sunday nights)
- To what extent do weather conditions affect dining hall attendance?

We identified possible predictors out of the variables we had within our combined datasets (weather and occupancy). We then created a model using these variables to see which ones best predicted the average occupancy of the dining halls. Our initial predictors were the date, whether it was the weekend (and not family dinner), whether it was a family dinner night, the amount of precipitation, the maximum temperature for the day, and the minimum temperature for the day. Our final model removed the precipitation variable as we deemed it was statistically insignificant as a predictor. We determined that the most significant variable in the prediction of dining hall occupancy is the date.

Data exploration and visualization



The predictors we picked for our final model had a statistically significant correlation with average occupancy. In descending order of correlation, these are `date_int`, `name` (which meal: breakfast, lunch, or dinner), `TMIN` (minimum temperature for a given day), `TMAX` (maximum temperature for a given day), `is_family_dinner`, and `is_weekend_and_not_family_dinner`. `name` is encapsulates the Boolean variables of `is_breakfast`, `is_lunch`, and `is_dinner` for this correlation analysis. Later in our multiple regression, we can use the categorical variable of `name` for the same Boolean effect. `TMAX` was later removed in an effort to avoid multicollinearity within our model as we deemed the variable had too high of a correlation with `TMIN`. This correlation makes intuitive sense, as days with a high or low maximum temperature are likely to have a similar minimum temperature. `PRCP` (precipitation) was also not included in the final model as we deemed it too statistically insignificant to include in the model. Specifically, its correlation with `average_occupancy` is -0.008. All variables chosen for the final model have a significance code of three stars with `average_occupancy`.

Modeling/Analysis

We assume that our occupancy data is reasonably accurate to real-world occupancy of the dining halls. We were unable to acquire hourly weather data for New Haven, so we assume that our weather data per day has sufficient granularity to conduct an accurate analysis.

Our combined data is formatted with each row representing a unique meal block, dining hall, and day. Therefore each day has 3×14 data points, for 3 meals and 14 dining halls. Different columns represent different outcome variables (`average_occupancy` and `max_occupancy`) or potential predictor variables. In some cases, auxiliary dataframes were created for the purposes of data analysis; for example, conversion of

Boolean values to integers. The non-auxiliary variables are listed below:

```
 [1] "X"                                 "DATE"
 [3] "date_int"                          "hall_id"
 [5] "name"                              "is_weekend"
 [7] "is_family_dinner"                  "start"
 [9] "end"                               "average_occupancy"
[11] "max_occupancy"                     "PRCP"
[13] "TMAX"                              "TMIN"
[15] "is_weekend_and_not_family_dinner" "is_breakfast"
[17] "is_lunch"                          "is_dinner"
```

```
Call:
lm(formula = average_occupancy ~ name + date_int + is_weekend_and_not_family_dinner +
    is_family_dinner + TMIN, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1717 -0.6521 -0.0248  0.6124  6.0034

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.2180537  0.0438569  96.178  < 2e-16 ***
nameDinner                       1.3698780  0.0281851  48.603  < 2e-16 ***
nameLunch                        1.2755561  0.0276473  46.137  < 2e-16 ***
date_int                        -0.0055036  0.0000656 -83.896  < 2e-16 ***
is_weekend_and_not_family_dinner -0.5396927  0.0298944 -18.053  < 2e-16 ***
is_family_dinner                 0.1515140  0.0494215   3.066  0.00218 **
TMIN                            -0.0232663  0.0006842 -34.007  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.086 on 10327 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.5033,    Adjusted R-squared:  0.503
F-statistic:  1744 on 6 and 10327 DF,  p-value: < 2.2e-16
```

All of our predictors have a significance code of two or three stars, indicating that they are statistically significant in the determination of average occupancy for our model. The coefficients show that variable's effect on the average occupancy. We are assuming that for a given variable, the change is based on all other variables being held constant. This is an assumption, since our model inherently has some amount of collinearity.

We determine that there is a statistically significant difference between levels of dining hall attendance at different meals (breakfast, lunch, and dinner). For the above coefficients, breakfast happens to be our baseline of the Boolean variables. If it is lunch, then there is a predicted increase in average occupancy of about 1.28. If it is dinner, then there is a predicted increase in average occupancy of about 1.37. From these, we can determine that the occupancy for breakfast is predicted to be much lower than lunch and dinner.

The model predicts an increase in occupancy on family dinner nights by 0.152.

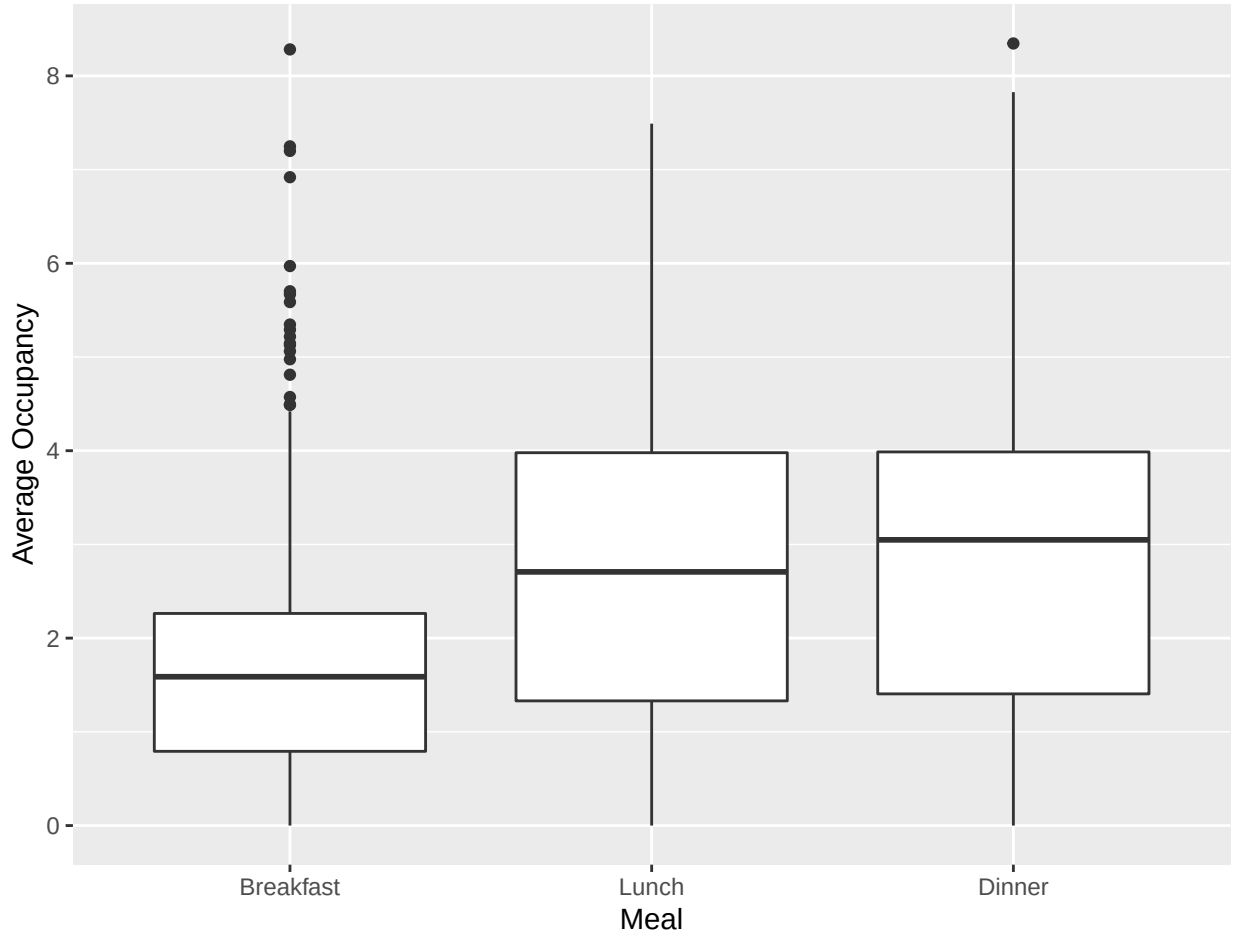
Some weather conditions are predicted to have an affect on dining hall attendance. Precipitation was removed earlier due to demonstrating a statistically insignificant effect in predicting occupancy. Therefore, the only weather condition we included in our model was `TMIN` (the minimum temperature for a given day). `TMIN` was chosen over `TMAX` due to having a higher correlation with `average_occupancy`. The negative coefficient for `TMIN` indicates that the model predicts the occupancy to decrease as the minimum temperature increases.

Lastly, if it is the weekend but not a family dinner night, then the occupancy is predicted to decrease by 0.540.

Our adjusted R^2 is 0.503, indicating that our model does well to predict the average occupancy of dining halls. This model is generally fairly appropriate for this kind of data, mixing Boolean and numerical variables to determine occupancy. The main issue with this model is `date` due to its cyclical nature, an issue we will further discuss in the improvements section of our conclusion.

Due to the real-life and applicable nature of the subject matter, it is rather easy to explain the data and the variables to a non-technical audience. For example, the high, positive coefficient for dinner may be explained by the general understanding that students are more free and awake during dinnertime resulting in the occupancy during dinner to be greater than breakfast (the baseline variable for the three meals).

Visualization and interpretation of the results




```
[1] "Breakfast"
```

```
 average_occupancy
 Min.   :0.0000
 1st Qu.:0.7922
 Median :1.5881
 Mean   :1.6145
 3rd Qu.:2.2634
 Max.   :8.2824
```

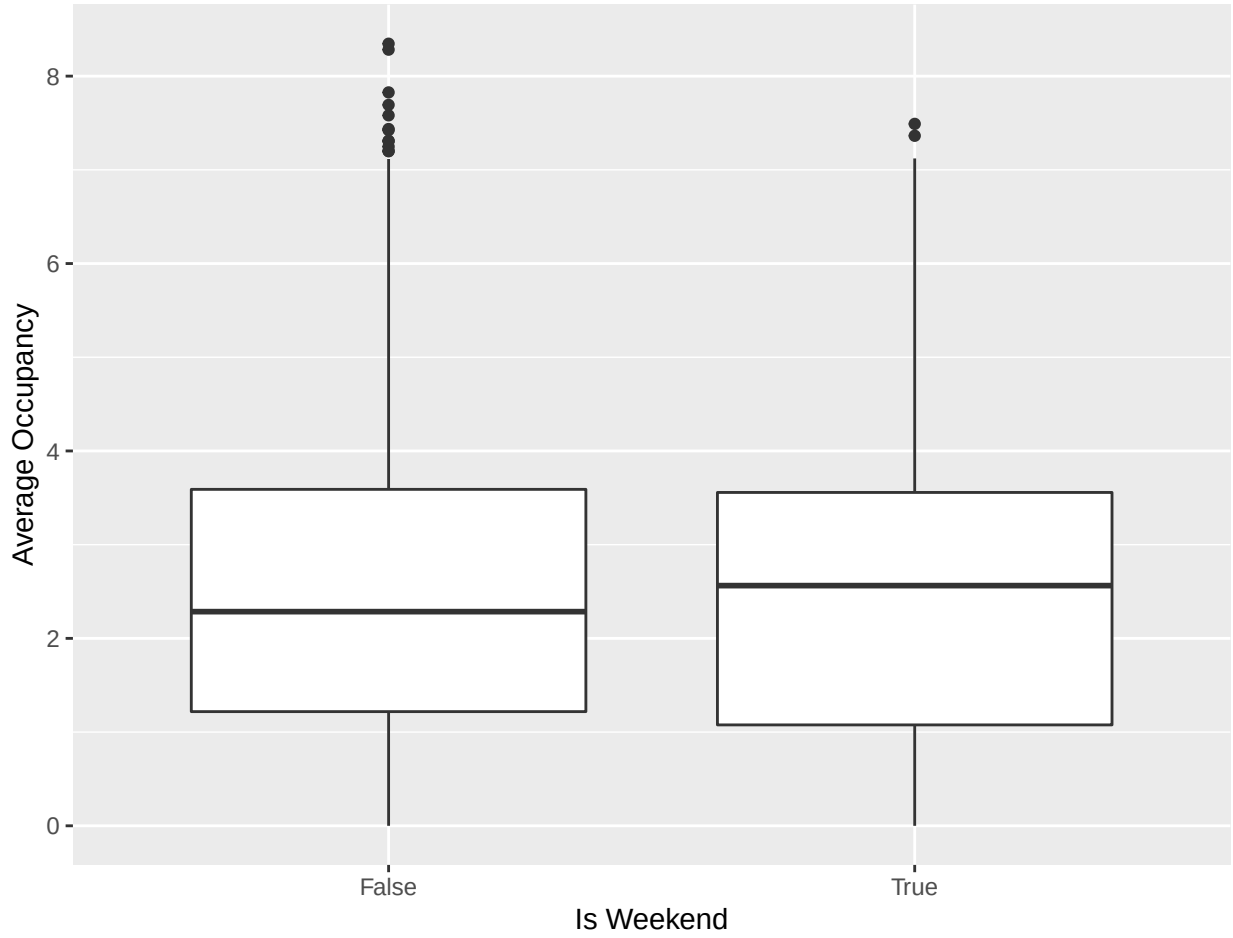
```
[1] "Lunch"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.331
 Median :2.707
 Mean   :2.737
 3rd Qu.:3.979
 Max.   :7.490
```

```
[1] "Dinner"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.405
 Median :3.049
 Mean   :2.876
 3rd Qu.:3.987
 Max.   :8.346
```

As previously determined through our use of multiple regression coefficients earlier, the average occupancy is generally higher for lunch and dinner. The above box plot supports this conclusion as the median average occupancy for lunch (2.707) and dinner (3.049) is higher than that of breakfast (1.588). Yale students seem to not attend breakfast as often as other meals. This could be explained by Yale students waking up late and missing breakfast, having to rush to class, or disliking the foods on offer. The outliers for breakfast could be explained by few students consistently waking up early for breakfast.



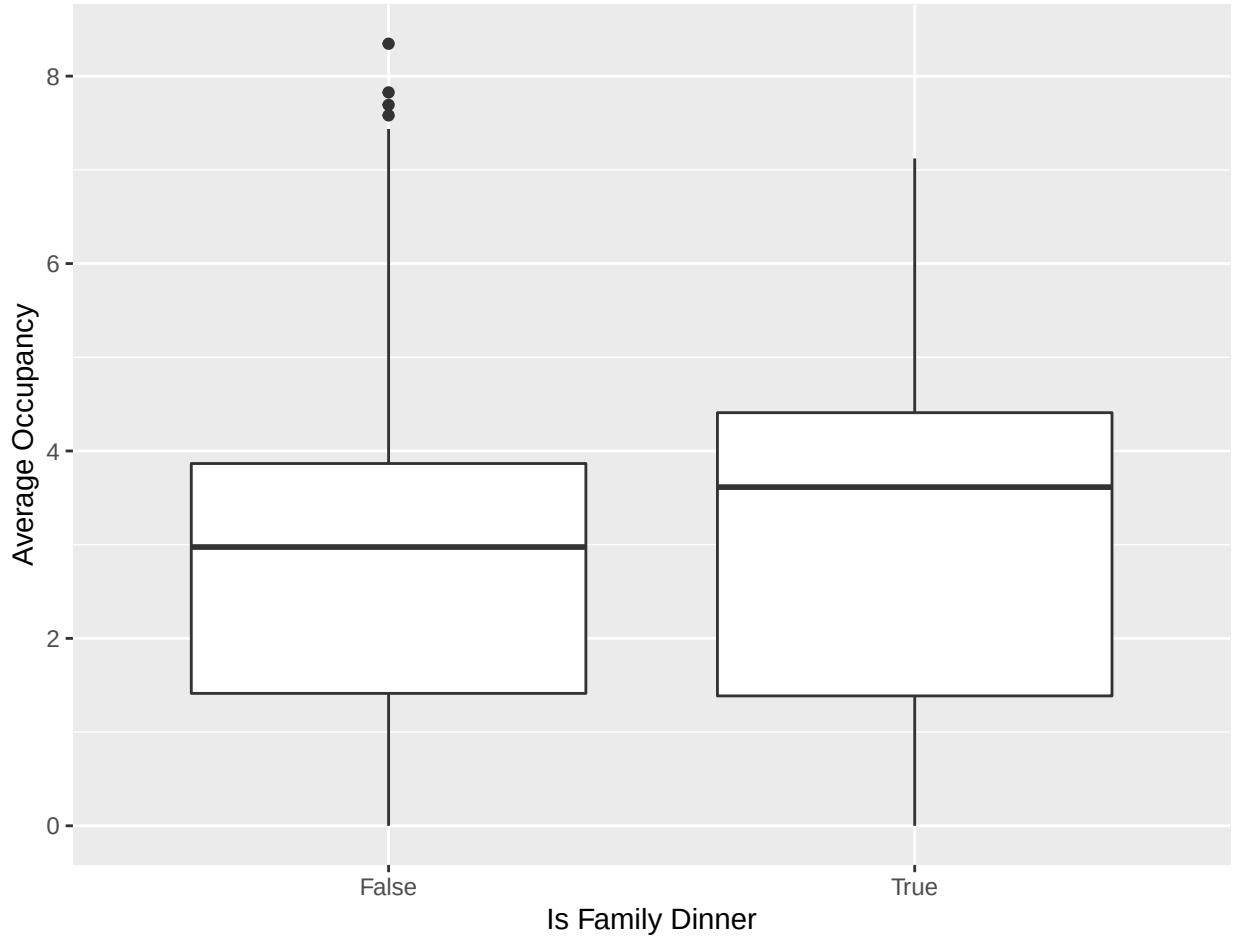

```
[1] "Is Not Weekend"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.218
 Median :2.285
 Mean   :2.489
 3rd Qu.:3.590
 Max.   :8.346
```

```
[1] "Is Weekend"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.077
 Median :2.563
 Mean   :2.459
 3rd Qu.:3.557
 Max.   :7.490
```

Due to multicollinearity issues with family dinner nights (Sundays), we did not include `is_weekend` within our multiple regression. Through this box plot we can see that the median average occupancy tends to be slightly higher on weekends (2.563) than that of weekdays (2.285). One explanation for this data could be that students are more free on weekends, allowing them to attend dining halls more frequently.




```
[1] "Is Not Family Dinner"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.413
 Median :2.975
 Mean   :2.827
 3rd Qu.:3.866
 Max.   :8.346
```

```
[1] "Is Family Dinner"
```

```
 average_occupancy
 Min.   :0.000
 1st Qu.:1.386
 Median :3.614
 Mean   :3.148
 3rd Qu.:4.408
 Max.   :7.121
```

As previously determined through our use of multiple regression coefficients earlier, the average occupancy is generally higher during family dinners. The above box plot supports this conclusion as the median average occupancy for family dinners (3.614) is higher than that of non-family dinners (2.975). A possible explanation for the higher occupancy on family dinner nights is that students enjoy eating with their fellow residential college residents, and that off-campus students who would not normally eat in the dining halls are able to

return to their college during this time.

Both boxplots and a multiple regression model allow us to gain insights into the dining habits of students based on certain variables. Boxplots are a good way of visualizing data to communicate information to the reader. They both convey similar trends, allowing us to arrive at the same conclusions for our questions regardless of which model is used. Multiple regressions allow us more exact prediction for a given variable, since it gives a numerical value effect on occupancy (assuming all other variables are held constant). Multiple regression analysis allows us to predict average occupancy based on multiple variables at once, whereas our boxplots focus on a single variable. Multiple regression is best for our model given the numerous variables we are considering at once. Therefore, the fact that it is based on more than one independent variable lends it to be a more realistic representation of average occupancy.

Conclusions and recommendations

We concluded that there is a statistically significant difference between levels of dining hall attendance at different meals (breakfast, lunch, and dinner). Specifically, average occupancy for breakfast tends to be lower than that of lunch and dinner.

We concluded that there is a difference in attendance to dining halls on weekends vs weekdays, though very small. Specifically, attendance on weekends in slightly higher than that of weekdays.

We concluded that there is a statistically significant increase in dining hall attendance during family dinners.

We concluded that precipitation does not have a statistically significant affect on dining hall attendance, whereas minimum temperature on a given day does have a statistically significant effect on average occupancy.

Some improvements that could be made to our analyses include finding hourly or sub-hourly weather data and determine whether date is better modeled periodically based on the year. Having a higher granularity in our weather data would allow us to better align the weather with a certain mealtime. According to our model, as time goes on, average occupancy of the dining halls decreases. While this could be accurate, it is worth considering the time since the beginning of the academic year as a variable instead, which would be periodic over the course of multiple years. This could be a more accurate predictor depending on student behavior. This variable could be visualized as a graph to determine yearly trends based on the season, breaks, or exam periods.

Future work could include analysis of the difference in breakfast attendance for dining halls that have continental breakfast vs dining halls with hot breakfast. We could also include new predictors based on the academic calendar, such as breaks or exam periods to see how they impact the occupancy of dining halls.