Subway Unit Sales
Garrett Peters

1. Introduction

Throughout this project, we will be looking at data collected by two subway stores located in downtown State College. Downtown State College is the hub for a large portion of the 40,000 Penn State students. With its numerous apartment buildings, restaurants, bars, and fast-food chains, it is always buzzing with activity. Among these, you will find two Subway locations, both under the ownership and operation of client Chris Paret. Chris reached out to me and my classmates to analyze units sold across the fall semester of 2016. He specifically tasked us with examining the difference in units sold throughout the week for each of his two stores. He seeks insights into which days require increased staffing levels and, conversely, when staffing can be scaled back. This data was pulled from each store's POS (Point of Sale) system which facilitates sales transactions, manages inventory, provides sales reporting, and tracks employee performance. The population of interest includes the units sold during upcoming fall semesters in the foreseeable future. We filtered the original 307 total days logged to focus on the fall semester, removing spring semester days and clear outliers such as Thanksgiving break and November 3rd, 2016, which was National Free Sandwich Day. Our analysis now centers on 178 fall semester days to accurately assess differences within each store.

1.1 - Research Questions

Question 1) For both stores, is there a difference in unit sales between days of the week during the fall semester?

Question 2) Can we predict the units sold for each store for Fridays and Saturdays when there is a home football game?

1.2 - Variables

Variable	Type	Description	Levels and Ranges
Units	Numeric	Number of units sold on a specific day	169 units to 606 units
Store	Categorical	Store Number	Str17, Str26
Date	Quantitative	Dates throughout the 2016 fall semester	9/07/2016 to 12/14/2016 (outliers deleted)
DayOfWk	Categorical	Day of the week	7 days of the week
Football	Categorical	Weekends where there are home football games	No = No game, Yes = Game

Table 1: Summary of variables for the Subway study

2. Exploratory Data Analysis

Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean	Standard Deviation	Sample Size
169	266	310	442.75	606	346.360	97.795	178

Table 2: Five number summary, mean, standard deviation and sample size for units sold

This table shows what is called a five-number summary of the units sold data plus a few other statistics. A five number summary displays the minimum value, 25th percentile where 25% of the data falls below and 75% above (Q1), the median where 50% of the data falls below and 50% above, the 75th percentile where 75% of the data falls below and 25% above (Q3), and the maximum value. The table also shows the mean, or average value, the standard deviation which is a measure of variation in the data (sd), and the number of observations in the data, or number of days being studied in this case (n).

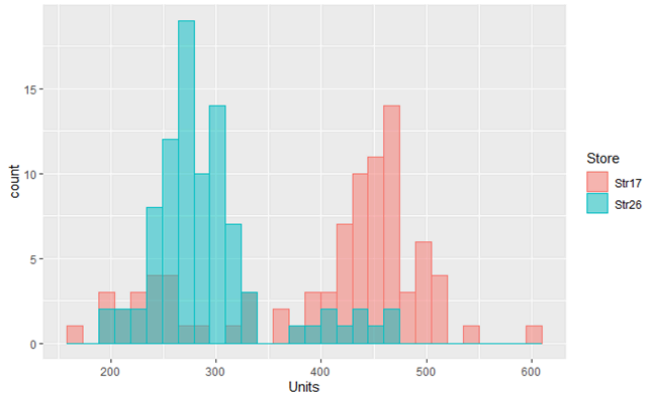

Figure 1: Histogram of Units Sold color coded by Store

This histogram shows the frequency of units sold for each day by store. The graph suggests that the majority of the days for Store 17 (Burrowes Street) the units sold are higher than the majority of the days for Store 26 (E College Ave). Each distribution seems to be somewhat normal, although there is a portion of days where the Burrowes St store sold a very low number of units, even lower than the average College Ave day. There is also a portion of days where College Ave had very high units sold, but this could potentially be explained by the football games. Overall though, this would suggest that Store 17 (Burrowes Street) is busier than Store 26 (E College Ave).

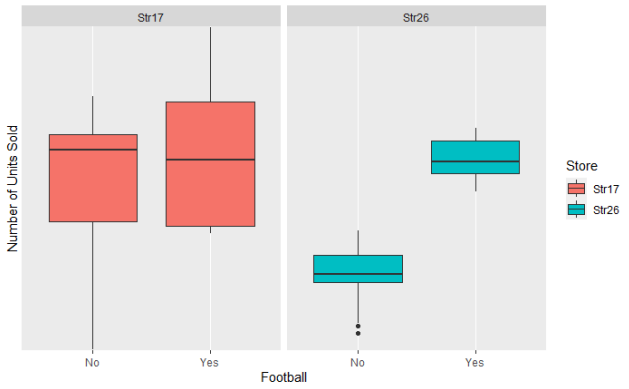

Figure 2: Boxplot of Units Sold per Store by Football Determination

This graph shows side-by-side boxplots for units sold for each store by whether or not it was a Friday or Saturday of a home football game. A boxplot is a visual representation of the five-number summary mentioned earlier where the black line in the middle is the median, the edges of the box are the first and third quartiles, and the top and bottom of the lines are the minimum and maximum. These plots would suggest something different for each store. For Store 17 (Burrowes Street), the median is essentially the same whether or not there is a football game. The minimum, maximum, and third quartile may increase during home games, but overall, there is nothing to suggest a significant change in units sold. On the other hand, Store 26 (E College Ave) experiences a clear increase in units sold when it is a home football Friday or Saturday, as evidenced by the boxplot for home game data being completely above the boxplot for non-home game data.

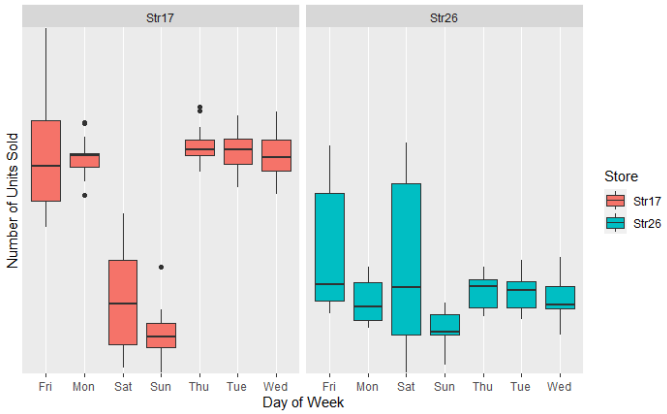

Figure 3: Boxplot of Units Sold per Store by Day of Week

This graph shows side-by-side boxplots for units sold for each store by which day of the week they were sold. Clearly, during the week, Store 17 (Burrowes Street) has more units sold on average. However, the Saturday and Sunday sales are much lower than the weekdays for that store. It also seems that the upper end of the Friday sales are far higher than any other day which makes sense considering it is both a workday and the day before a home game some weeks. The

median sales for Store 26 (E College Ave) seems to be fairly similar throughout the week outside of maybe a slight decrease on Sunday. Fridays and Saturdays have much larger ranges, though, which makes sense considering the last figure how much the unit sales for this store increases during home games.

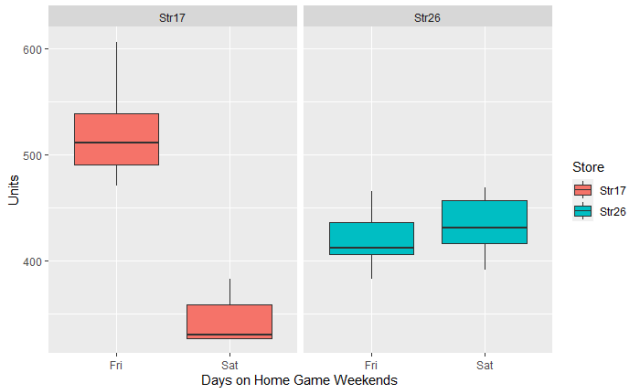

Figure 4: Boxplot of Units sold for each store per Day on Home Game Weekends

This graph shows side-by-side boxplots of units sold for each store for home football Fridays and Saturdays. From this visual representation, it is clear that for Store 17 (Burrowes Street) the number of units sold are significantly higher on Fridays than Saturdays for home football games. This makes sense as most people are at the football game and the store is far away from the stadium. On the other hand, there is no real difference in units sold between home football Fridays and Saturdays for Store 26 (E College Ave). This store is closer to the stadium, so it is possible that people stop at this store before going to the game.

3. Statistical Analysis

3.1 - Research Question 1

We chose to analyze this data using R Studio. To answer the first research question - “For both stores, is there a difference in unit sales between days of the week during the fall semester?” - conducting a Tukey Comparison Test is essential to identify significant differences among the days. Significance is the likelihood that the observed results, or more extreme ones, occurred by random chance alone, often represented by the p-value, indicating the strength of evidence against the null hypothesis in statistical analysis. Since we want to look at each store independently, we decided to split the data in two data sets, one for Store 17 (Burrowes Street) and one for Store 26 (College Avenue). In order to conduct a Tukey Comparison Test, we need to fit an Analysis of Variance (ANOVA) model for each store.

An ANOVA model will determine which variables have an effect on unit sales. The results included in Table 3 below show that both the Day of Week variable (DayOfWk) and Football variable are both significant for both stores. Including an interaction term between these two

variables is not necessary to include because it is known that football games will only be played on Saturdays. This also means that it was not necessary to do any backward elimination because

Store 17

ANOVA	Variable	P-value
Store 17	DayOfWk	0.000
	Football	0.000
Store 26	DayOfWk	0.000
	Football	0.000

Table 3: Final ANOVA model summary for both Store 17 and 26

all the variables in the data set were significant, and therefore no variables to eliminate.

It is important to note that after doing residual analysis, that all model assumptions were satisfied for both the final model for Store 17 and Store 26, these assumptions include normality, constant variance and independence. These plots will be listed in the Appendix below. In order to determine the significant difference in days, we ran a Tukey Multiple Comparison's Test using these models for each store. Significance is determined as a p-value less than .05. Tables describing the significant difference in days are located below:

Store 17

Variable - Days	Difference in Means	P-Value
Sat - Mon	-169.089744	0.000
Sat - Tue	-176.089744	0.000
Sat - Wed	-170.738095	0.000
Sat - Thur	-182.666667	0.000
Sat - Fri	-172.916667	0.000
Sun -Sat	-47.371795	0.0051
Sun - Mon	-216.461538	0.000
Sun - Tue	-223.461538	0.000
Sun - Wed	-218.109890	0.000
Sun - Thur	-230.038462	0.000
Sun - Fri	-220.288462	0.000

Table 4: Tukey Comparison - Significant DifferenceS in Days for Store 17

Store 26

Variable - Days	Difference	P-value
Fri - Mon	61.0705128	0.0000017
Fri - Tues	48.8397436	0.0001972
Fri - Wed	57.1309524	0.0000056
Fri - Thur	48.3333333	0.0003328
Fri - Sun	95.3782051	0.0000000
Sat - Mon	47.4038462	0.0003318
Sat - Tues	34.6666667	0.0177170
Sat - Wed	43.4642857	0.0010117
Sat - Thur	34.6666667	0.0246829
Sat - Sun	81.7115385	0.0000000
Sun - Mon	-34.3076923	0.0185328
Sun - Tues	-38.2472527	0.0003207
Sun - Wed	-46.5384615	0.0044454
Sun - Thu	-47.0448718	0.0003773

Table 5: Tukey Comparison - Significant Differences in Days for Store 26

Looking at Table 4, it tells us that both Saturday and Sunday are drastically different from days during the week for Store 17. Since the difference in means is negative, it tells us that the weekdays are more busy (more units sold) than on the weekends. This can be confirmed by

looking at the red box plots on Figure 3, the means for these days do seem significantly lower than the week days. It is also worth noting the difference between Saturday and Sunday is also significant. Since their difference is also negative, this tells us that Sunday is the slowest day for Units Sold, which can also be seen in Figure 3. Table 5 is also indirectly telling us that days in the week are all very similar and do not show a significant difference between each other.

Table 5 highlights significant differences between weekends and weekdays, with Friday being significantly busiest day for Store 26. Saturdays are also significantly busier than typical weekdays. While Friday is slightly busier than Saturday, the drop-off in sales from Friday to Sunday is insignificant. Sundays consistently have significantly slower sales compared to other days. It is also worth noting that the remaining days throughout the work week show no significant difference between each other. This can be visualized and confirmed when looking at Figure 3 in the EDA section above.

For both Stores, we also find significant differences in the determination of a football game, this can be seen in the Appendix below. Mean unit sales are significantly higher when there is a football game vs when there is not a football game. Figure 2 in the EDA section above, shows a good visual representation. This makes sense because these games attracted more people to the downtown State College Area, therefore these days will have greater mean unit sales vs when there are not football games.

3.2 - Research Question 2

To answer our second research question - Can we predict the units sold for each store for Fridays and Saturdays when there is a home football game? - we use the same ANOVA models for each store to find a 95% prediction interval. A prediction interval provides a range of values within which we expect a future observation to fall with a certain level of confidence. It accounts for both the variability in the data and the uncertainty in the prediction, offering a range rather than just a single point estimate. Below in Table 7 shows the prediction for each store for each day alone with its corresponding 95% Prediction Interval

Store	Home Football Game Day	Prediction	95% Prediction Interval
17	Friday	520.7583	(454.4438 - 587.0729)
17	Saturday	347.8417	(281.5271 - 414.1562)
26	Friday	433.4333	(378.3971 - 488.4696)
26	Saturday	419.7667	(364.7304 - 474.8029)

Table 7: Prediction Intervals for each Store

Table 7 shows that Fridays for both Store 17 and Store 26 will likely be busier than Saturdays when home football games occur. Looking at Figure 4 will confirm that for Store 17, there is a

pretty big drop off from Friday to Saturday, and that for Store 26, both Friday and Saturday show similar unit sales.

4. Recommendations

Question 1) For both stores, is there a difference in unit sales between days of the week during the fall semester?

For Store 17 (Burrowes Street), the days that there is significant difference in unit sales are Saturday and Sunday when compared to the work week days. These days have significantly lower than average unit sales. Sundays are also significantly lower than Saturdays. It is also important to note unit sales are higher when home football games occur on Saturday.

For Store 26 (East College Avenue) the days that there are significant differences in unit sales are Fridays, Saturdays and Sundays when compared to work week days. Fridays and Saturdays tend to have significantly higher than average unit sales. However, Sundays tend to have significantly lower than average unit sales.

Question 2) Can we predict the units sold for each store for Fridays and Saturdays when there is a home football game?

For Store 17 (Burrowes Street), we are 95% confident that the predicted unit sales falls between (454.4438 - 587.0729) on Fridays, and (281.5271 - 414.1562) on Saturdays.

For Store 26 (East College Avenue), we are 95% confident that the predicted unit sales falls between (378.3971 - 488.4696) on Fridays, and (364.7304 - 474.8029) on Saturdays.

5. Resources

For resources related to ANOVA and Tukey Multiple Comparison Tests please see:
https://online.stat.psu.edu/stat485/lesson/12/12.2

For resources related to Prediction Intervals please see:
https://online.stat.psu.edu/stat501/lesson/3/3.3

For resources related to R-Studio please see:
https://guides.libraries.psu.edu/c.php?g=894031&p=6430365

6. Additional Considerations

We answered the research question using a one way ANOVA model for each store. In both cases, the model assumptions were reasonably met suggesting that the model can be trusted in predicting units sold. We also created prediction intervals for Friday and Saturday when there's a

home football game for each store. This interval should suggest with 95% confidence what that sales on those days at those stores should be. There are a few additional considerations to remember when analyzing your results that are mentioned below.

Correlation and Causation: This data was studied observationally, in that it is not a true experiment, as the conditions for each store were not assigned randomly. Therefore, when discussing the relationship between the day of week and units sold as well as units sold on home football Fridays and Saturdays, it is important to be careful. There is no telling whether these results are specific to the year studied or not as football games at Penn State can be an unpredictable environment. The gap between the weekend and weekdays for Burrowes St, however, seems like it is significant enough to draw conclusions from along with the cap between home football Fridays and Saturdays at that store. It is also important to keep in mind the fickleness of studying home football games because not every home football game is created equal. Factors such as the team they are playing, the weather, what time the game is being played, and the weather all change from game to game and, more importantly, from year to year. Therefore, be careful not to extrapolate the football data too much as it is now eight years later.

Football and Day of Week Interaction: Left out of our ANOVA model was the interaction term that would have measured the effect of the football game and day of week variable on units sold. While this value was significant in our ANOVA model, it did not make sense to include in our final model because of the nature of the two variables. The football games generally happen every other Saturday in the fall, with everyone coming into town on Friday before the game, meaning that the football variable indicating whether or not there was a home football game will be highly correlated with those two days of the week. For that reason, it made sense to leave that term out of our model.

Outliers and Pieces Excluded From the Data: Again, because State College in the fall can be an unpredictable place, it is important to remember a few key pieces missing from the data. The exclusion of National Sandwich day was mentioned as the units were mostly free that day, but the fall break gap was also not considered. Those days, the sales were low enough to be counted as zero, but they were not actually zero. Specifically, the home football game against Michigan State on November 26th is curious for a few reasons. While this is still considered fall break, it is a Saturday before a Monday school day which means a lot of people were likely coming back up after Thanksgiving for the game and staying for class on Monday. Also, it is important to recognize the significance of that game. Penn State was playing for a trip to its first Big Ten Championship in eight years in a 3:30 game meaning that there were a lot of people at the game and a lot of people hanging around before and after. For this reason, we would have expected there at least to be some significant sales on and around that day that are not measured in the data.