The number of units of Subway sold

Miza Syafiqah and Phichchaya Sutaporn

October 6th, 2022

1. Project Description

There are three Subway stores located Downtown, State College. They are at Burrows Street, Pugh Street, and East College Ave. The manager of the subways is interested in predicting how the sales are going to be after the pandemic, and life in State College is back to normal. Therefore, the data that will be assessed is from pre-pandemic as life in State College during that time is similar to nowadays. The data collected is over the course of 1 year and the number of units sold, along with some other variables is collected. The manager of Subway State College currently has completed the data collection portion of the study and has requested that the Statistics 470W students to assists her in formally answering the research questions using the data.

1.1. Research Question

Question 1: Are the differences between the store locations and days of the week associated with different average units sold for three subway stores in downtown state college?

Question 2: For Fridays and Saturdays (only), is there a difference in units sold for different stores and football versus non-football weekends, on average?".

1.2 Variables

The number of Units sold in the three Subway stores Downtown is collected over the span of 1 year; from January 2016 to March 2017, excluding Summer Break, Fall Break, and Winter break. Throughout the data collection, the Subway stores Downtown, Store 17 (Burrows st), Store 18 (Pugh St) & Store 26 (East College Ave), are collected to distinguish where the units are being made and sold. The date, days of the week, whether it is a weekend or weekday, and month are also collected along with whether it is a Football weekend or not. Note that the weekend is defined as Friday and Saturday.. A summary of all variables considered is included in table 1

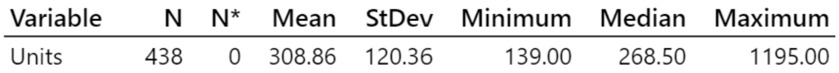
2. Exploratory Data Analysis (EDA)

It is always a good idea to define the variables used in the analysis before doing the analysis and answering the research question. In this analysis, the response variable is Units while the explanatory variables are Store, DayOfWk, and Football.

Variable	Type	Description	Level or range
Units	Numerical	Unit of food sold (not including drinks and cookies)	139 to 1195
Store	Categorical	Store locations in downtown state college	Str17xx: Burrows street Str18xx: Pugh street Str26xx: East College Ave.
DayOfWk	Categorical	What day of the week was that day	Mon/Tue/Wed/Thu/Fri/Sat/Sun
Football	Categorical	Whether that day had a football game or not	Football/ None football

Table 1: Summary of variables used in the analysis

Now let's take a closer look at the Unit variable. Table 2 shows that there is a total of 438 observation units with no missing value. The mean of Units is 308.86 while the median of MeanDiam is 268.50. This suggests that the data is right-skewed.



Variable	N	N*	Mean	StDev	Minimum	Median	Maximum
Units	438	0	308.86	120.36	139.00	268.50	1195.00

Table 2: Summary statistics for the Units variable.

From figure 1, the subway store located on Burrows street seems to have the highest average unit sold followed by East College Ave. and Pugh street in order. When comparing football and non-football game day, football game day seems to have a higher average of units sold than non-football game day. As we can see from the boxplots, there seem to exist some outliers.

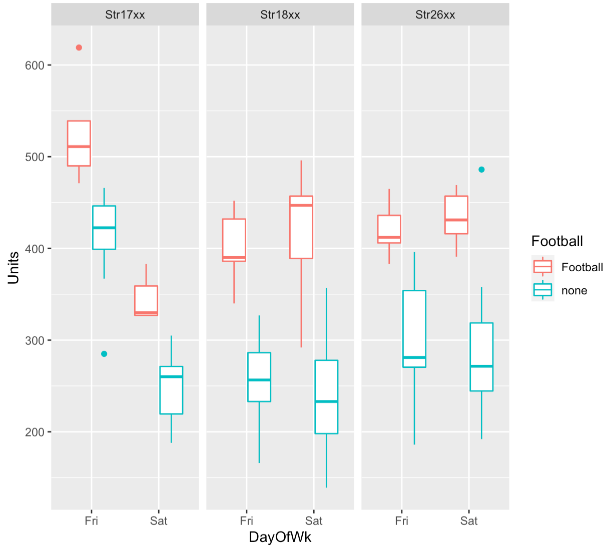
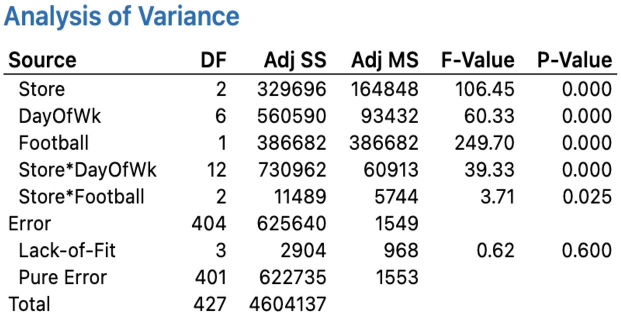


Figure 1: Box plots of units sold comparing three locations in downtown state college.

There are two research questions we want to explore, the first research question is are different store locations and days of the week associated with different average sales? The second research question is For Fridays and Saturdays (only), is there a difference in units sold for different stores and football versus non-football weekends, on average?

3. Statistical Analysis

To determine whether there are differences between the store locations and days of the week associated with different average units sold for three subway stores in downtown state college, we fit a model with Store, DayOfWk, Football, Month, and the two-way interaction terms. After using backward induction with alpha equal to 0.05 and getting rid of 10 outliers associated with the national sandwich day, St Patty's, and THON events because those days are special events that do not reflect the number of units sold on a regular basis, (see table x in appendix) and the standardized residuals are beyond ± 2. Per residual plots in the Appendix, the model assumptions are satisfied. Below is our final model:



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	329696	164848	106.45	0.000
DayOfWk	6	560590	93432	60.33	0.000
Football	1	386682	386682	249.70	0.000
Store*DayOfWk	12	730962	60913	39.33	0.000
Store*Football	2	11489	5744	3.71	0.025
Error	404	625640	1549		
Lack-of-Fit	3	2904	968	0.62	0.600
Pure Error	401	622735	1553		
Total	427	4604137			

Table 3: ANOVA table of the final model for the first research question

The variable Football is not indicated in research question one but, we decided to include it in our model because including it shows that the variable store and DayOfWk are statistically significant (p-value less than 0.05) meaning that Store and DayOfWk do associate with different average units sold.

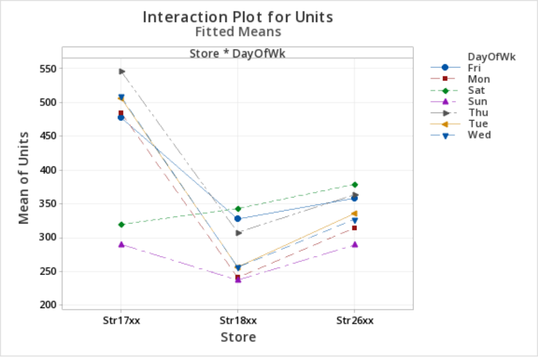
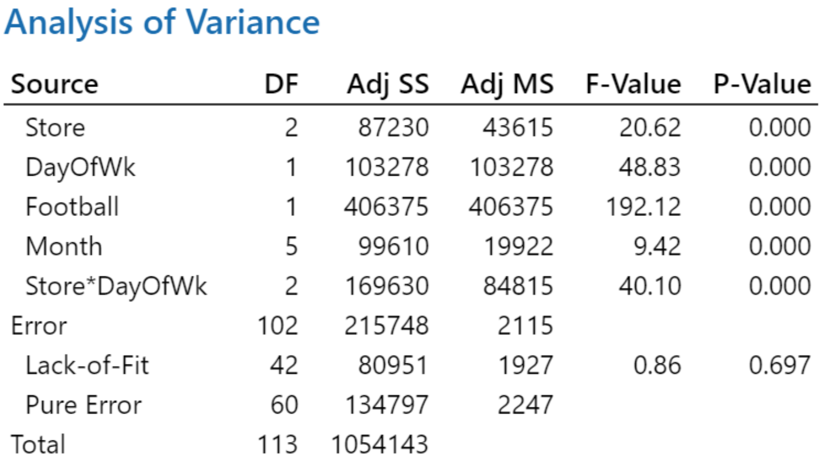


Figure 2: Interaction plot for units sold for the first research question

Other than ANOVA, we also perform the Tukey comparison (see appendix for the table) for the first research question and found that the average units sold by the subway store at Burrow street every day except for Saturday and Sunday are significantly higher than other stores every day even its own store on Saturday and Sunday. The reason might be because the subway store at Burrow street has the closest distance to campus, so it is possible that many students are more likely to get food during the day from the Burrow street location than other locations. Lastly, Subway at Pugh St has the lowest average units sold on weekdays, except for Thursday, compared to weekends where they have higher sales. This Tukey comparison is supported by figure 2 which shows that the Burrow street store has the highest average sales on Monday till Friday and Pugh St store has the lowest on weekdays, except for Thursday.

Similarly, for the second question, to determine whether there are differences between the store locations, football versus non-football on only Friday and Saturday associated with different average units sold for three subway stores in downtown state college, we fit a model with Store, DayOfWk, Football, Month, and the two-way interaction terms. After using backward induction with alpha equal to 0.05 we got rid of 6 outliers associated with the national sandwich day, St Patty's, and THON event (see table 6 in appendix) and the standardized residuals are beyond ± 2. Looking the figure 9 (in the appendix), all three assumptions of ANOVA are satisfied.

After fitting ANOVA models, we finally arrive at our final model below:



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	87230	43615	20.62	0.000
DayOfWk	1	103278	103278	48.83	0.000
Football	1	406375	406375	192.12	0.000
Month	5	99610	19922	9.42	0.000
Store*DayOfWk	2	169630	84815	40.10	0.000
Error	102	215748	2115		
Lack-of-Fit	42	80951	1927	0.86	0.697
Pure Error	60	134797	2247		
Total	113	1054143			

Table 4: ANOVA table of the final model for the second research question

Even though the variable Month is not indicated in the second research question, we still have to include it in our model because including it shows that the variable Store, DayOfWk, and Football are statistically significant (p-value less than 0.05) By looking at the final, we know that the variable Store, DayOfWk, Football, Month, and the interaction term of Store* DayOfWk do associate with different average units sold.

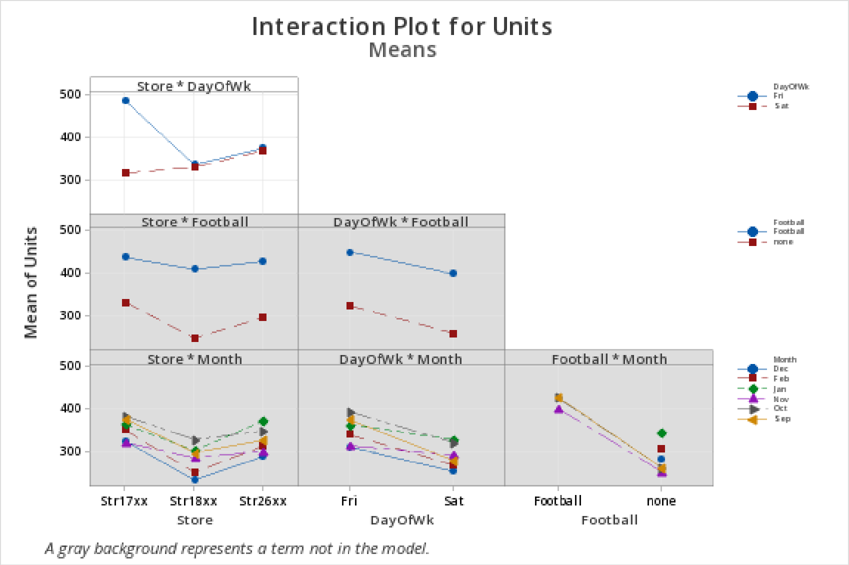


Figure 3: Interaction plot for units sold for the second research question

After performing the Tukey method for the second research question (see appendix to see the table), we found that the average units sold on football game day are significantly higher than none football day. Since there are a lot of people coming to town during the game day, this might be related to the higher units sold. In addition, the Burrow street store on Friday has the highest average units sold and it is significantly higher than the average units sold of other store locations on both Friday and Saturday. The reason might be because on Friday a lot of people go downtown, especially on Friday night and the subway store at Burrow street is the store that is the closest store to bars. This Tukey comparison is supported by figure 3. The first row of plots shows that the Pugh street store has the highest average units sold. Also, the second row indicates that football weekend has higher average units sold than non-football weekend.

4. Recommendations

The described analysis of your data results in the recommendations below:

Question 1: Are the differences between the store locations and days of the week associated with different average units sold for three subway stores in downtown state college?

Yes, store locations and days of the week are associated with different average units sold. Subway at Burrows St has the most average units sold on Monday till Friday compared to Saturday and Sunday. Subway at Pugh St has the lowest average units sold on weekdays, except for Thursday, compared to weekends where they have higher sales.

Question 2: For Fridays and Saturdays (only), is there a difference in units sold for different stores and football versus non-football weekends, on average?".

Yes, for Friday and Saturday (only) store locations and football versus non-football weekends are associated with different average units sold. Football weekend has significantly higher average units sold than none football weekend.

5. Resources

For resources related to the two-way ANOVA, see https://online.stat.psu.edu/statprogram/stat461/

6. Additional Considerations

We answered your research questions using a two-way main effects ANOVA model. In both cases, the conditions of the model were reasonably met and the results should be considered trustworthy.
For supporting figures, please see the Appendix. The Tukey comparison plot that is located in the Appendix shows the significance between interaction variables and if you want us to elaborate, please do not hesitate to contact us if you have any questions. There is also a caution below:
Association is not causation: This is an observational study, so we cannot make cause-and-effect conclusions. For example, we can only say that the difference in average units sold is related to store location but we cannot conclude that changing store from one location to another location causes the change in average units sold.

Technical Appendix

Further EDA

Date	Day of week	Location	Event
11/03/16	Thursday	Burrows street	National Sandwich day
02/24/17	Friday	Burrows street	St Patty's
02/25/17	Saturday	Burrows street	St Patty's
11/03/16	Thursday	Pugh street	National Sandwich day
02/18/17	Saturday	Pugh street	THON
02/24/17	Friday	Pugh street	St Patty's
02/25/17	Saturday	Pugh street	St Patty's
11/03/16	Thursday	East College Ave.	National Sandwich day
02/18/17	Saturday	East College Ave.	THON
02/24/17	Friday	East College Ave.	St Patty's

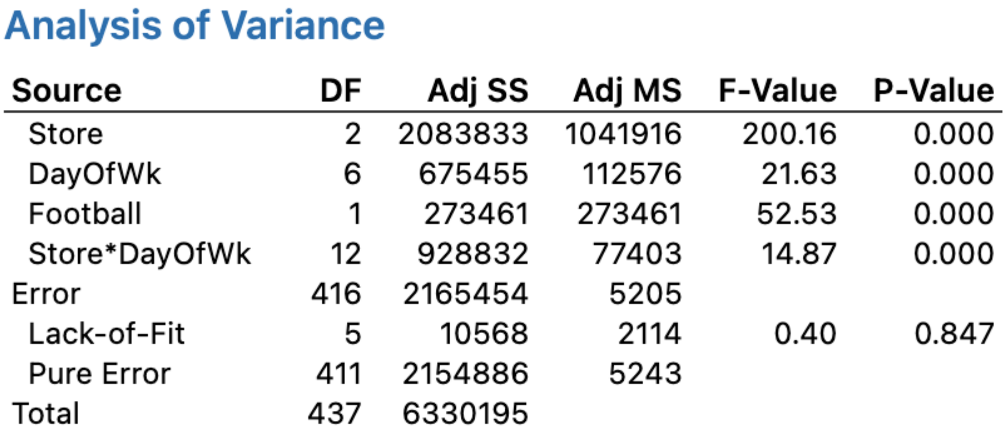
Table 5: Outliers table for the first research question showing dates, days of the week, and locations they correspond to.

Date	Day of week	Location	Event
2/25/17	Saturday	Burrows street	St Patty's
2/18/17	Saturday	Pugh street	THON
2/25/17	Saturday	Pugh street	St Patty's
2/18/17	Saturday	East College Ave.	THON
2/24/17	Friday	Burrows street	St Patty's
2/24/17	Friday	Pugh street	St Patty's

Table 6: Outliers table for the second research question showing dates, days of the week, and locations they correspond to.

Assessing ANOVA for the first research question:

1. **The first model before getting rid of outliers**



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	2083833	1041916	200.16	0.000
DayOfWk	6	675455	112576	21.63	0.000
Football	1	273461	273461	52.53	0.000
Store*DayOfWk	12	928832	77403	14.87	0.000
Error	416	2165454	5205		
Lack-of-Fit	5	10568	2114	0.40	0.847
Pure Error	411	2154886	5243		
Total	437	6330195			

Table 7: Analysis of Variance for the first model

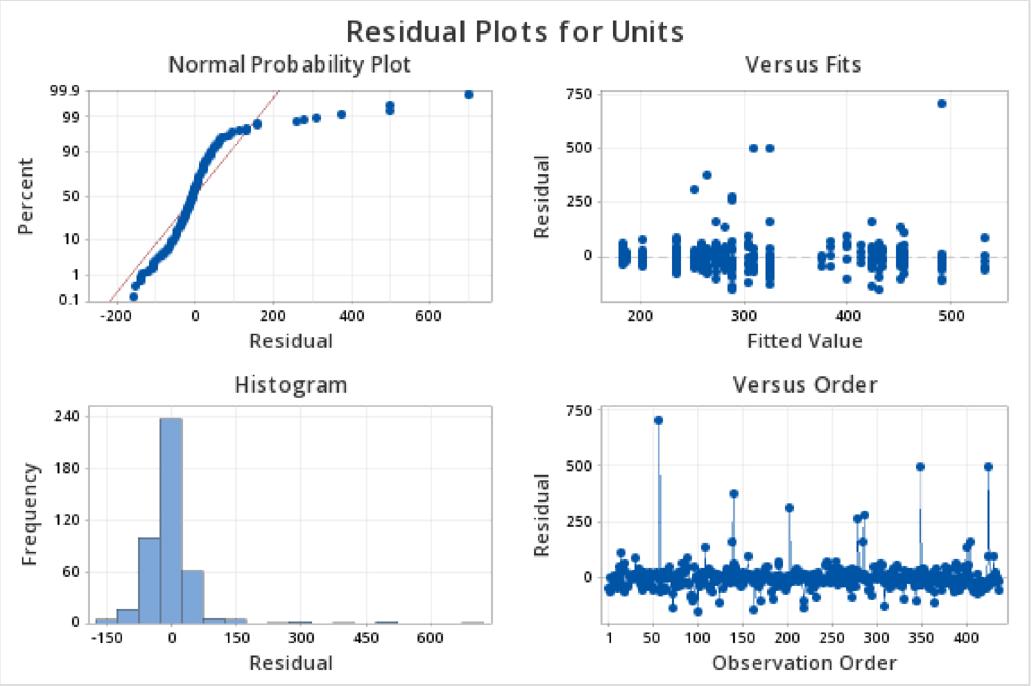
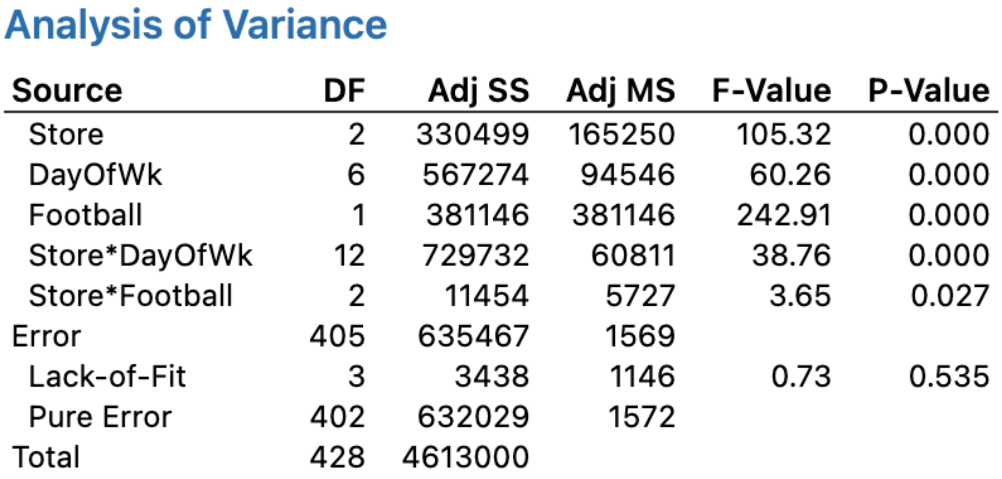


Figure 4: Residual plots of Units for the first model

2. The second model after getting rid of 9 outliers



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	330499	165250	105.32	0.000
DayOfWk	6	567274	94546	60.26	0.000
Football	1	381146	381146	242.91	0.000
Store*DayOfWk	12	729732	60811	38.76	0.000
Store*Football	2	11454	5727	3.65	0.027
Error	405	635467	1569		
Lack-of-Fit	3	3438	1146	0.73	0.535
Pure Error	402	632029	1572		
Total	428	4613000			

Table 8: ANOVA for the second model

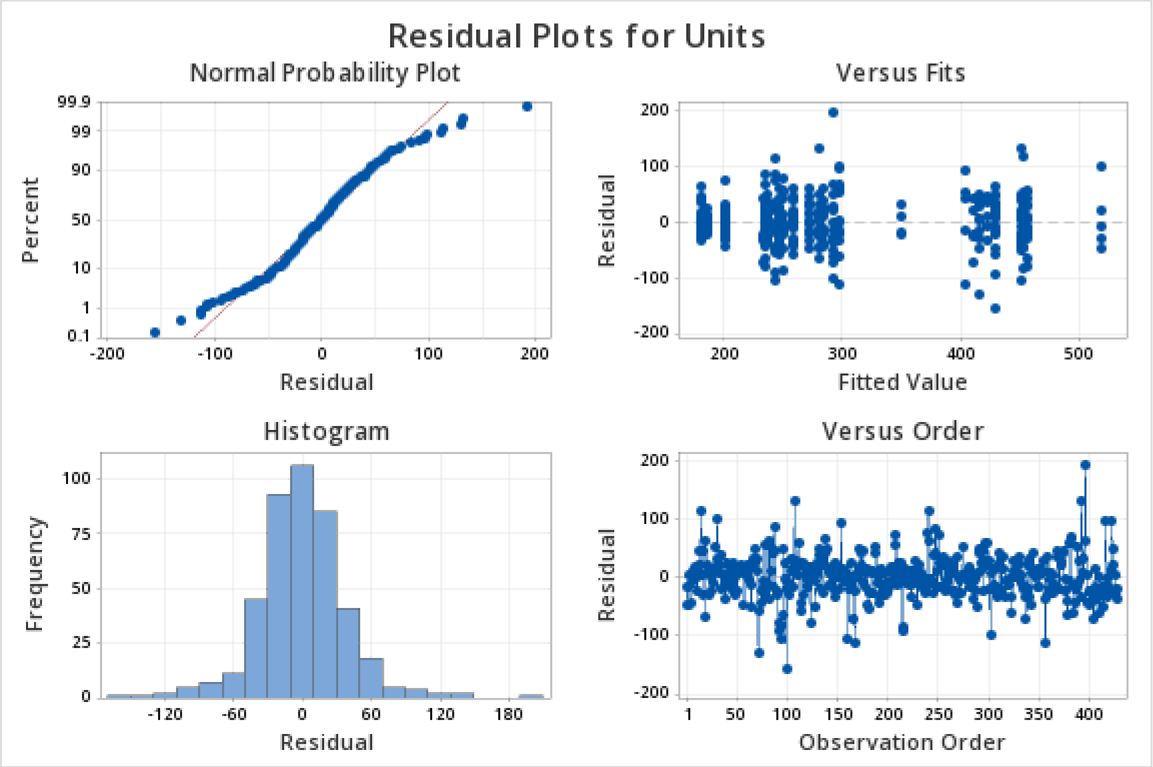


Figure 5: Residual plots for units for the second model

3. **The third and final model after getting rid of 1 more outlier can be found in the Statistical Analysis section and the residual plots are as below:**

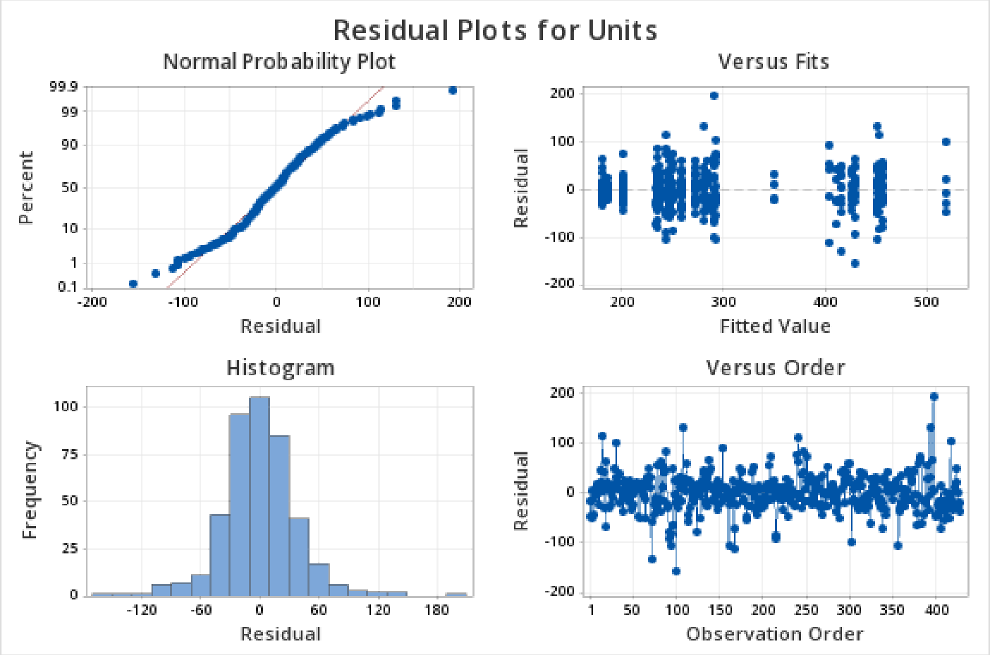
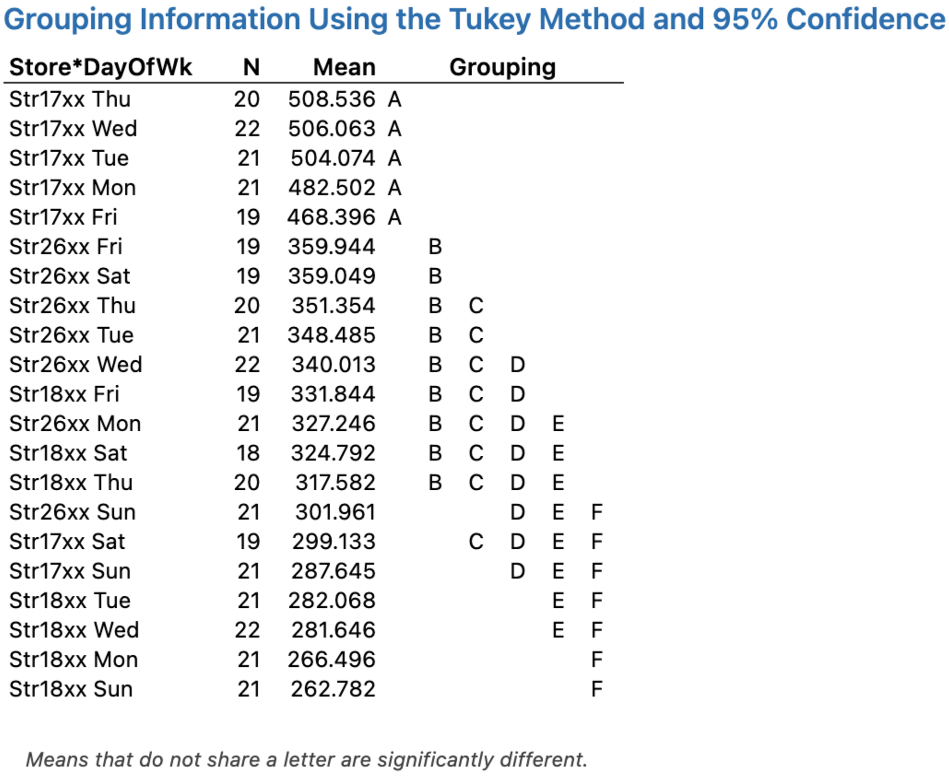


Figure 6: Residual plots for Units of ANOVA final model for the first research question

From the normal probability plot, the graph looks normally distributed as the majority of the points lie on the red line. However, it can be seen that there are a few outliers. After looking at these outliers, they are not from the national sandwich day, St Patty's, and THON so, we do not remove them because these outliers still reflect the sales on a regular basis. Looking at the residual vs fit plot, there is no clear cone pattern meaning that the assumption of equal variance is met. The samples were taken independently, so there is no indication that this assumption is violated. The units sold at different stores on different days do not depend on other stores and other days. So, it is independent. Therefore, the three assumptions are satisfied.

Tukey comparison for the first research question:

We assessed the relationship between the variables to see which are significant or not. Below is the Tukey table:



Grouping Information Using the Tukey Method and 95% Confidence

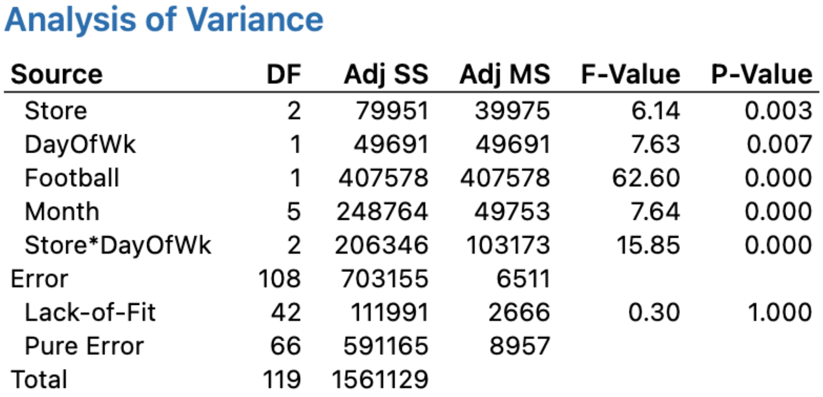
Store*DayOfWk	N	Mean	Grouping					
Str17xx Thu	20	508.536	A					
Str17xx Wed	22	506.063	A					
Str17xx Tue	21	504.074	A					
Str17xx Mon	21	482.502	A					
Str17xx Fri	19	468.396	A					
Str26xx Fri	19	359.944		B				
Str26xx Sat	19	359.049		B				
Str26xx Thu	20	351.354		B	C			
Str26xx Tue	21	348.485		B	C			
Str26xx Wed	22	340.013		B	C	D		
Str18xx Fri	19	331.844		B	C	D		
Str26xx Mon	21	327.246		B	C	D	E	
Str18xx Sat	18	324.792		B	C	D	E	
Str18xx Thu	20	317.582		B	C	D	E	
Str26xx Sun	21	301.961				D	E	F
Str17xx Sat	19	299.133			C	D	E	F
Str17xx Sun	21	287.645				D	E	F
Str18xx Tue	21	282.068					E	F
Str18xx Wed	22	281.646					E	F
Str18xx Mon	21	266.496						F
Str18xx Sun	21	262.782						F

Means that do not share a letter are significantly different.

Table 9: Tukey Method of comparison for the first research question

Assessing ANOVA for the second research question:

1. **First model before getting rid of outliers:**



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	79951	39975	6.14	0.003
DayOfWk	1	49691	49691	7.63	0.007
Football	1	407578	407578	62.60	0.000
Month	5	248764	49753	7.64	0.000
Store*DayOfWk	2	206346	103173	15.85	0.000
Error	108	703155	6511		
Lack-of-Fit	42	111991	2666	0.30	1.000
Pure Error	66	591165	8957		
Total	119	1561129			

Table 10: Analysis of Variance for the first model

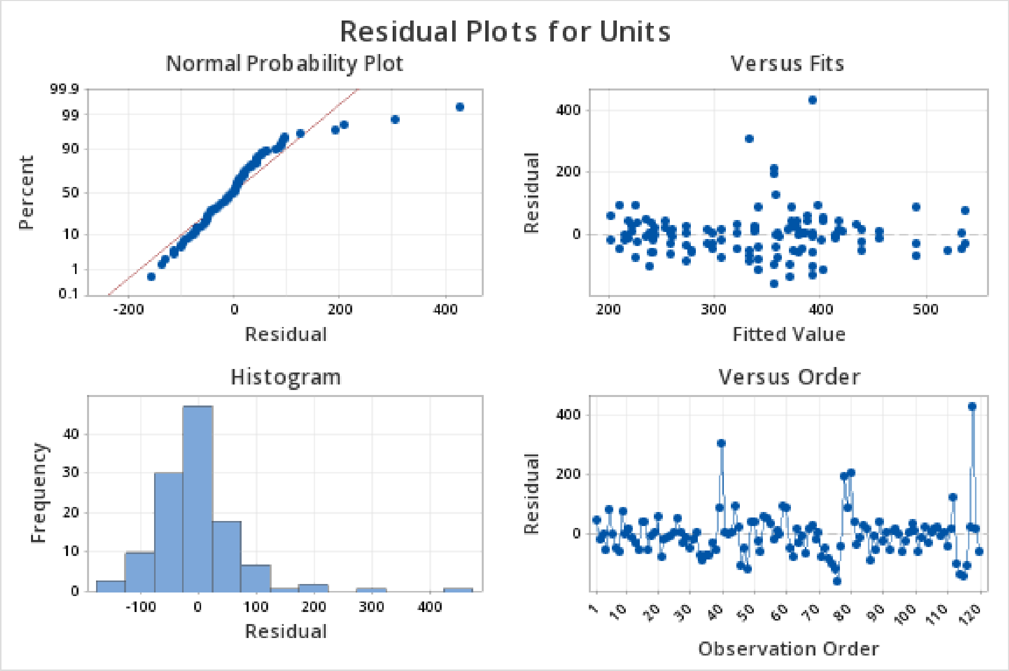
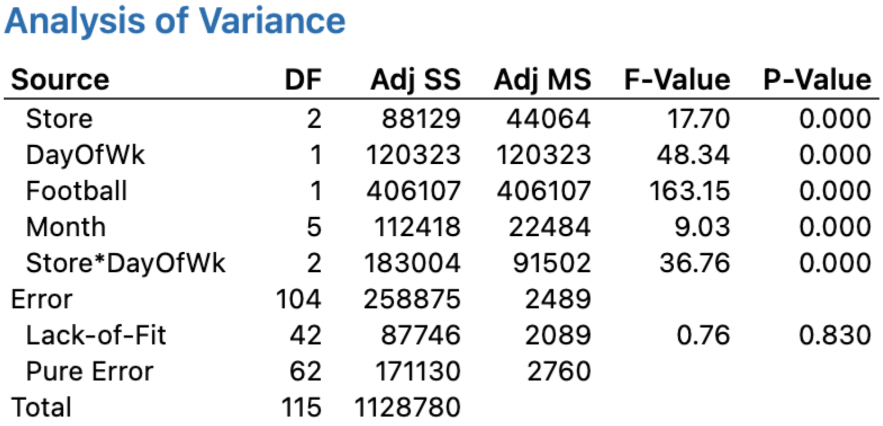


Figure 7: Residual plots for Units of ANOVA first model for the second research question

2. **Second model after getting rid of 4 outliers:**



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Store	2	88129	44064	17.70	0.000
DayOfWk	1	120323	120323	48.34	0.000
Football	1	406107	406107	163.15	0.000
Month	5	112418	22484	9.03	0.000
Store*DayOfWk	2	183004	91502	36.76	0.000
Error	104	258875	2489		
Lack-of-Fit	42	87746	2089	0.76	0.830
Pure Error	62	171130	2760		
Total	115	1128780			

Table 11: Analysis of Variance for the second model

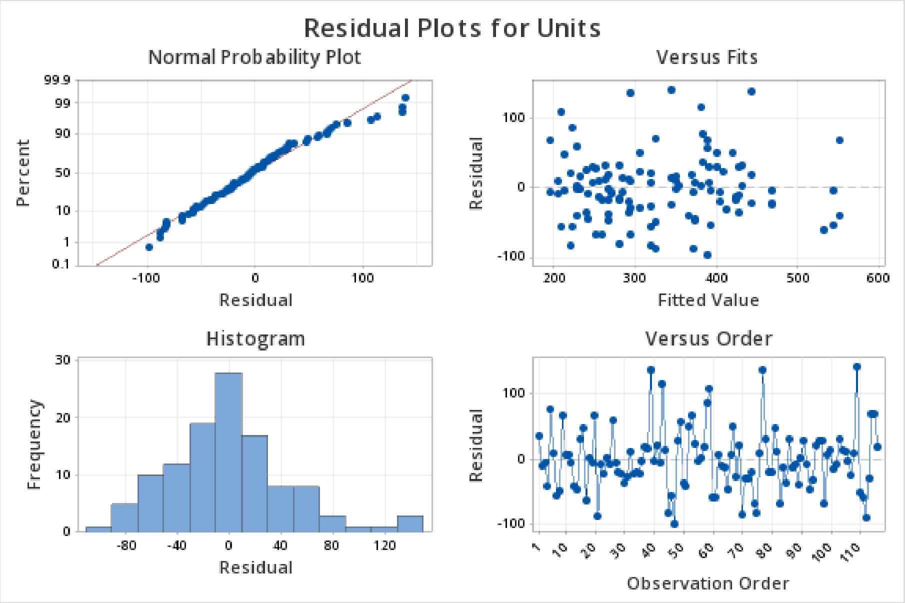


Figure 8: Residual plots for Units of ANOVA second model for the second research question

3. **The third and final model after getting rid of 4 more outliers can be found in the Statistical Analysis section and the residual plots are as below:**

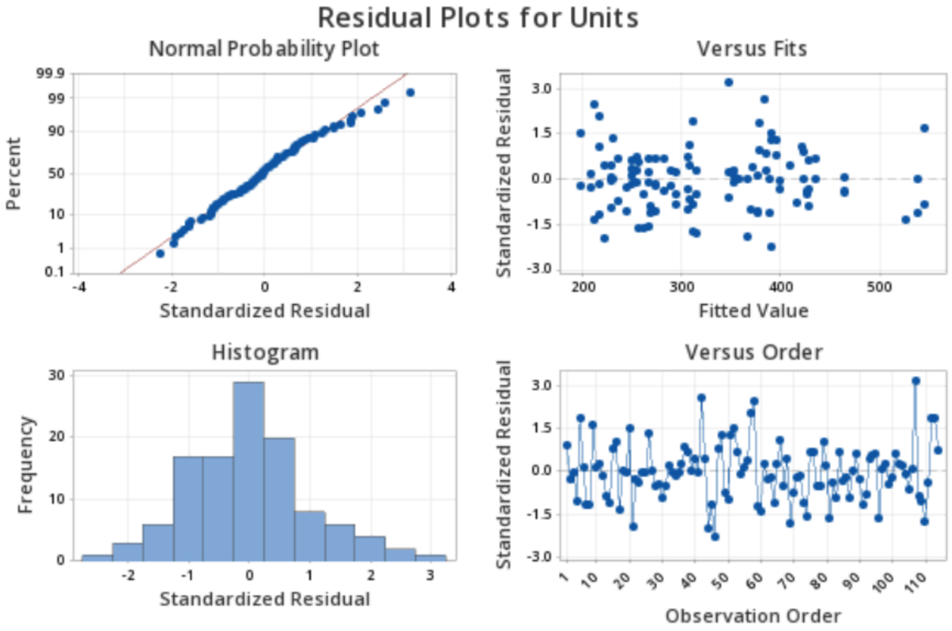
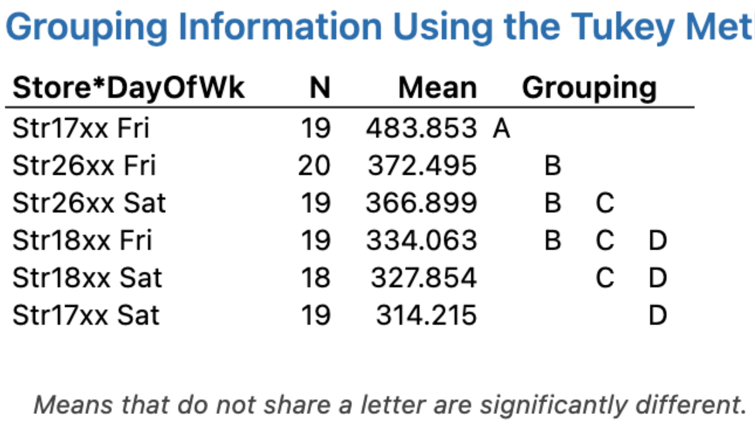


Figure 9: Residual plots for Units of ANOVA final model for the second research question

The normal probability plot suggests that the data is normally distributed because the points seem to follow the red line. Regardless of the fact that there are still some outliers in the normal probability plot, we still have to keep those observations since they are not special events. There is no clear cone pattern in the residual vs fit plot meaning that the assumption of equal variance is met. The samples were taken independently, so the independence assumption is satisfied. Therefore, all three assumptions are satisfied.



Grouping Information Using the Tukey Method and 95% Confidence

Store*DayOfWk	N	Mean	Grouping			
Str17xx Fri	19	483.853	A			
Str26xx Fri	20	372.495		B		
Str26xx Sat	19	366.899		B	C	
Str18xx Fri	19	334.063		B	C	D
Str18xx Sat	18	327.854			C	D
Str17xx Sat	19	314.215				D

Means that do not share a letter are significantly different.

Tukey comparison for the second research question:

We assessed the relationship between the variables to see which are significant or not. Below is the Tukey table:

Table 12: Tukey Method of comparison for the second research question