Chicago Crime Data Analysis 2016-2018

By Holly Michalak and Julian Gomez

1. Introduction

Any visualization of crime frequency will show a clear seasonal pattern – crime is low in the colder months, spikes in the summer, then declines as the temperature cools. The simplest intuitive explanation is the change in routine that comes with the change in season. People spend more time outside when the weather is warm and days are longer which presents more opportunities for crime. Other possible factors include more people going on vacations, leaving their homes susceptible to burglary, and children being off from school and unsupervised. Some studies have suggested that this relationship is not universal across all crimes and that violent crimes in particular will increase with heat, such as Anderson's 2001 paper.[1] Another study that analyzed crime data from Philadelphia found that crime increases on warmer days, regardless of the season. A possible explanation for increased crime on warmer days in cold months is that people are more willing to go outside if the weather is more pleasant. But this same idea does not hold up for warmer days in warmer months – crime still increases even when it is uncomfortably hot.[2]

Regardless of the cause, it is apparent that there is a relationship between crime rate and the season and this has been well-researched. Given this, we seek to

1 Anderson, Craig A. "Heat and Violence." *Current Directions in Psychological Science* 10, no. 1 (February 2001): 33–38. doi:10.1111/1467-8721.00109.

2 Schinasi, L.H. & Hamra, G.B. J Urban Health (2017) 94: 892. https://doi.org/10.1007/s11524-017-0181-y

find other relationships between aspects of crime and periods of time. Namely, we will test the following hypotheses:

1. The proportion of reported crimes that result in arrests will be higher in the warm months because police departments will be aware of the spike in crime. Police officers will also be more likely and willing to be in the community when the weather is comfortable.
2. There is a relationship between temperature (warm versus cool weather) and type of crime.
3. Examining the subgroup of crimes classified as domestic, we believe we will find that the average percentage of arrests per day will be higher in the warm months than in the cool months.
4. The proportion of crimes that are reported will not be evenly distributed across all days of the week. There will be a higher proportion of crimes on Fridays and Saturdays because weekend behavior presents more opportunities for crime.

Our crime data from January 2016 to July 2018 was obtained from the Chicago Data Portal. We also used total population estimates for the city of Chicago to calculate the crime rate and perform this analysis.

2. Analysis

2.1 Daily Crime Rates and Arrest Rates

2.1.1 Methods

In order to calculate the arrest percentages, we made a frequency table that counted the number of daily incidents from January 1, 2016 to July 31, 2018. In the table, we included the date and temperature. So that we would have the count for incidents by date, we outputted the table as a new dataset. To calculate the daily arrest percentage, we created another frequency table for number of arrests per day the same way we did for the frequency of reported crimes. Then, we merged the two datasets by date in order to divide the number of arrests by the number of reported crimes and multiply by 100. After that, we created a scatterplot of the percentages over time to see if there are any patterns. Afterwards, we used a t-test to check if the mean arrest percentage for colder weather is less than that of warmer weather.

2.1.2 Result

Once we got the daily arrest percentages in one dataset, we created a scatter plot (figure 1) for the arrest percentages by date. It shows seasonal patterns within each year, where it starts increasing around the first few months of each year, peaks around the early spring, decreases throughout the summer months, and becomes steady during the last months. This pattern suggests that daily arrests tend to change through different seasons. Following this, we proceeded to find out how the pattern behaves throughout each year using a t-test.

To analyze this further, we needed to test the mean differences in percentages by weather. The percentages are normally distributed; so we generated an t-test to check if the mean daily arrest percentage is lower in cool weather than in warmer weather (Table 1). By the 5% significance level, the difference in variances is significant, so we used the Satterthwaite method (t-value= -6.81, p-

value < 0.0001). In other words, the t-test shows significant evidence that the mean arrest percentage for cooler weather is less than the mean arrest percentage for warmer weather. The mean for warm weather is 20.5%, and the mean for cooler weather is 19%. The boxplot from this t-test (figure 2) shows that the distributions for both weathers are normal and have some outliers above their maximum values. However, the boxplot for cooler weather shows several outliers outside the minimum value. These represent some days in the first few months of 2017, see figure 1, where there have been very few arrests.

2.2 Primary Types of Crimes

2.2.1 Methods

After analyzing the mean daily arrest percentages by season, we proceeded to inspect mean crime percentages by primary type of crime. Since there were various different types of crimes committed during the past three years, we did a frequency table to find the most reported crimes since 2016. With this, we created a pie chart that would show the crimes with the highest frequency. Next, we created a contingency table (table 2.1) of the reported crimes from the pie chart by arrest status. We then used the row percentages of this table to create a bar graph of the percentages of reported crimes that resulted in an arrest (figure 4).

Next, we needed to figure out how the percent of arrests over reported crimes behaves through weather temperature. So, we created a bar graph based on two contingency tables on crimes and arrest status separated by weather.

Finally, we created another contingency table for primary types of crimes and temperature for all arrests made (table 4.1). In here, the percentages are

calculated by dividing the number of arrests in a crime in a weather category by the total number of arrests for that crime. We would use this table to create a bar graph to see how arrest percentages of crimes differ across weather temperatures.

2.2.2 Results

The pie chart (figure 3) features the top 9 most reported crimes in the city: theft, battery, criminal damage, assault, narcotics, deceptive practice, burglary, motor vehicle theft, and robbery. The most reported crime in the chart is theft, while the least reported is motor vehicle theft.

In the contingency table (table 2), the crime that deviates the most from its expected value is narcotics, since it is the crime with the highest count of arrests. The chi-squared test for independence (table 3) confirms that there is significant evidence that type of reported crime and arrest status are related (chi-square = 156904, p-value < 0.0001). The bar chart from this table shows that narcotics is indeed the crime where criminals are most likely to get arrested, while the others have a percentage less than a quarter from that of narcotics. These results are not surprising, since finding a certain illegal drug on someone's person is sufficient for that person to be arrested. The other crimes with the highest percentage of arrests are battery, assault, and theft.

The bar chart on arrest proportions in crimes by weather (figure 5) shows the percentages of number of arrests in a crime divided by the total number of incidents for that crime. In here, the arrest percentages are different between temperatures when it comes to primary type. However, some of the crimes have a

higher percentage of arrests in cold weather than warm weather. These crimes are battery, motor vehicle theft, and theft.

The chi-square test for the contingency table of crime and temperature for all arrests made (table 4.2) shows significant evidence that there is a relationship between crimes that resulted in arrests and weather temperature (chi-square = 149.46, p-value < 0.0001). The bar graph from this table (figure 6) shows that the differences between percentages across temperatures seem very distant for all crimes except motor vehicle theft. The difference is very close, though it is still lower in cool weather than warm weather.

2.3 Domestic Crimes

2.3.1 Methods

In this section we are examining a subgroup of our data more closely, namely crimes classified as domestic per the Illinois Domestic Violence Act. We seek to find out if this subgroup of domestic crimes follows the same temperature-dependent pattern as the arrest percentage overall. Using an upper-tailed t-test, we tested the hypothesis that the mean percentage of arrests for domestic crimes in the warm months is greater than the mean percentage of arrests for domestic crimes in the cool months.

This was accomplished by first filtering all the reported crimes to only include those where the domestic variable was equal to true. We then used frequency tables to get counts by date, which had an associated temperature, and counts of arrests per date. This allowed us to ultimately have a dataset that

contained each date, an associated temperature, and the number of arrests divided by the number of reported crimes fitting these criteria.

2.3.2 Results

The results of our t-test indicated that the variances of warm and cool months were not the same, so in examining the output we used the un-pooled variance. The test returned a high p-value of .4038 and a t-value of 0.24 (Table 5). The mean arrest ratio for warm months is 17.17 and for cool months it is 17.0730, a difference of only 0.0969. Based on the confidence interval, we are 95% confident that the difference in means is greater than -0.5280. Our results indicate that there is not sufficient evidence to suggest that the mean arrest percentage for warm months is larger than the mean arrest percentage for cool months.

In examining the distributions using figures 6 and 7, the arrest ratio looks approximately normal for warm months. For cool months, on the other hand, both the histogram and qq-plot indicate a deviation from normality. Cool months have a few days where the arrest percentage were 100% and 50%. There are also several observations where the percentage is zero. These observations are way off from the majority. Given that the sample size for cool months is 418, the violation of normality is not particularly concerning for the t-test.

2.4 Crime over the Week

2.4.1 Methods

Finally, we will introduce another measure of time to our analysis, the day of the week. We hypothesized that there will be more crimes reported on the weekend,

namely Friday and Saturday, because weekend activities could increase crime (e.g. alcohol consumption and more people out in public places). A simple bar graph (Figure 8) showed that Saturday and Sunday indeed had the highest frequency of reported crimes. We will use a two-way ANOVA on day of the week and season to test if the average frequency of reported crimes is statistically different for each day of the week and to see if there is an interaction between season and day of the week.

Essential for this test is to extract the day of the week from every day that there was a reported crime. As we employed several times before, we made a frequency table to obtain counts of reported crimes for each day and outputted this as a new dataset. We then used the built-in weekday function to create a new variable for the day of the week and added the season. The result was a dataset that consisted of the date, number of crimes, season, and day of the week.

2.4.2 Results

The two-way ANOVA indicated that number of crimes per day is associated with season and with day of the week. As shown in table 6, our F-value was 14.71 and we had a significant p-value of <.0001. Both season and day of week were labeled as individually significant for average frequency of reported crimes per day. The interaction term, however, was not with a p-value of 0.9466. This was supported by the interaction plot (Figure 9), were all seven lines followed the same pattern. The boxplot of count (Figure 10) demonstrates multiple days in the winter and fall where crime was exceptionally low. spring and summer do not have nearly as many outliers, which indicates that the frequency of crime per day is much more

consistent throughout those seasons. Perhaps the low crime days in winter and fall were days with severe weather.

The SNK test shows that the mean frequency of crime is different for summer and winter but the same for spring and fall, as shown in figure 11. Regarding the mean frequency for day of the week, the SNK test in figure 12 indicates that the average daily frequency of reported crimes is not significantly different for Saturday through Thursday, all of which are different from Friday. It also shows that Thursday, Friday, Saturday, and Monday are not significantly different. This supports the previously established fact that there is more crime in the summer (this had the highest average frequency per day of 719.19). It also supports our hypothesis that the mean frequency of crimes per day is not the same for every day of the week, though it does not confirm that Friday and Saturday have uniquely higher frequencies since both of these days overlapped with multiple other days of the week.

3.1 Conclusion

Though many of our results were statistically significant, we did not have many findings that were practically significant or important. Our first t-test test indicated that there is a statistically significant difference in the mean percentage of arrests for warm versus cool months, but the difference was only about 1.5%. It shows up as statistically significant because, with such a large number of observations, the confidence intervals become very narrow and less likely to

overlap. In the real world, a 1.5% difference in the mean arrest rate does not have much of an impact.

Our second analysis used a chi-squared test to determine if there is a relationship between arrest status and types of reported crime. We saw very clearly that arrests are much more common for crimes involving narcotics. This is reasonable given that most of the reports are probably police officers catching individuals with drugs, in which case there will be an arrest. It is less likely that a third party will report a drug dealer to the police, and then that person is never caught. Our test does not conclusively tell us which crimes have a statistically different rate of arrests, so we cannot critique it in the same fashion as the ANOVA test but it's possible that we are finding significance again due to the large sample size.

The examination of domestic crimes indicated that there is no difference in the mean daily arrest rate for crimes classified as domestic between the warm and cool months. This is a difference result than what we found in the first section, that the arrest rate is statistically different for each season. As shown in table 7, the overall percentage of reported crimes in our dataset that resulted in an arrest is about 19%. This is slightly higher than the average for domestic crimes, which is about 17% for both temperature groups. We do not know if this difference is statistically significant, but it could be explained by apprehension against reporting someone in one's household to the police. Perhaps part of the reason this did not turn out as significant is because we have a smaller sample size for this data.

Finally, our two-way ANOVA showed more practically significant results than some of our previous tests. We found that spring and fall have means that are not significantly different, but winter and summer do which coincides with the previous research discussed in our introduction. Day of the week, on the other hand, was not as convincing. Some were shown to be significantly different but all averages of reported crimes per day were within about 50 reported crimes. In a city of almost 3 million people, this is not practically important.

Our analysis certainly had significant limitations and there are several things that could have been done differently. Pairing down the data, for example by looking at only one crime, could have made our sample size more manageable and allowed us to reach more meaningful conclusions. In addition, having all categorical data put some limitations on the types of tests we were able to run. One potential solution could be to find supplemental quantitative data.

Appendix

Figure 1 – Time Plot on Arrest Percentage

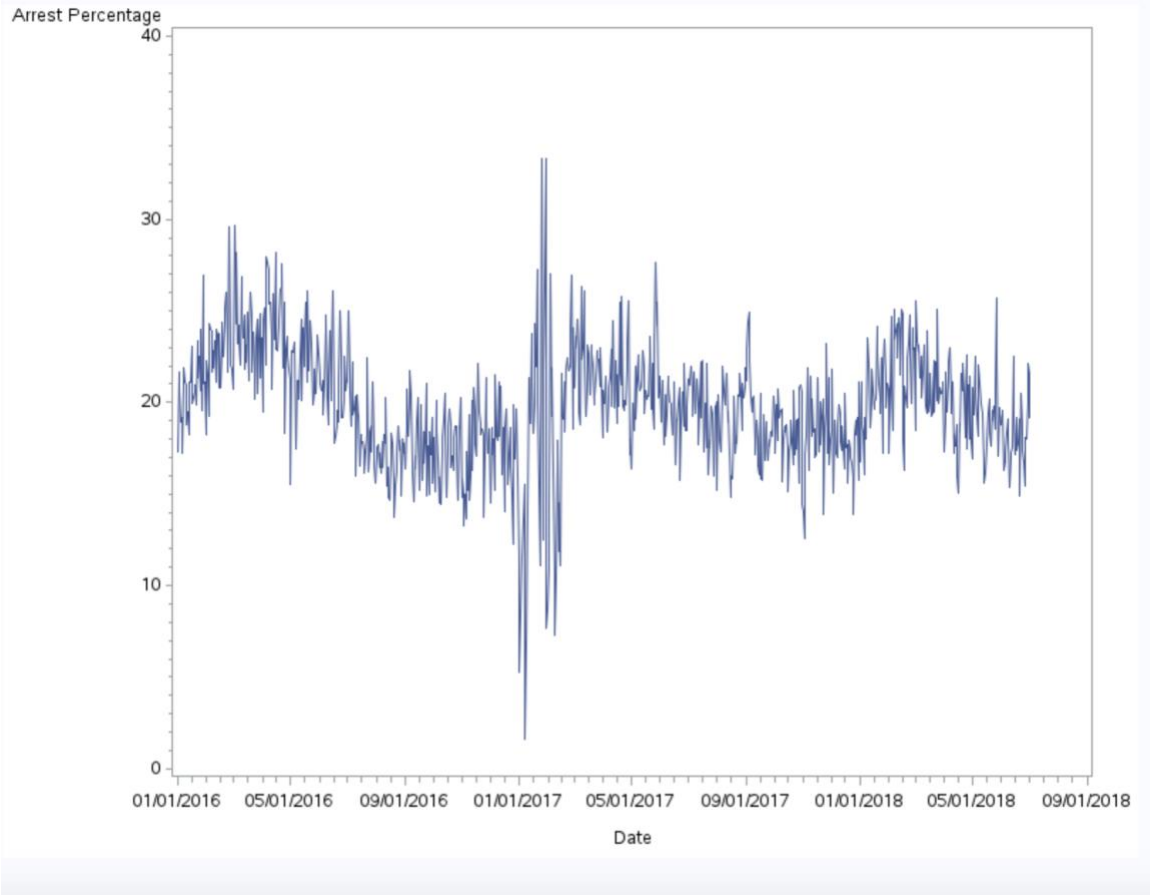

Table 1 – T- test on Arrest Percentages

The TTEST Procedure

Variable: Arrest_Percentage

Weather	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
cool		417	19.0224	3.5509	0.1739	1.6393	33.3333
warm		492	20.4703	2.7148	0.1224	13.7281	29.6724
Diff (1-2)	Pooled		-1.4479	3.1261	0.2081		
Diff (1-2)	Satterthwaite		-1.4479		0.2126		

Weather	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
cool		19.0224	18.6806	19.3642	3.5509	3.3251	3.8097
warm		20.4703	20.2299	20.7108	2.7148	2.5551	2.8959
Diff (1-2)	Pooled	-1.4479	-Infty	-1.1053	3.1261	2.9887	3.2770
Diff (1-2)	Satterthwaite	-1.4479	-Infty	-1.0977			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	907	-6.96	<.0001
Satterthwaite	Unequal	770.14	-6.81	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	416	491	1.71	<.0001

Figure 2 – Boxplot of Arrest Percentages by Weather

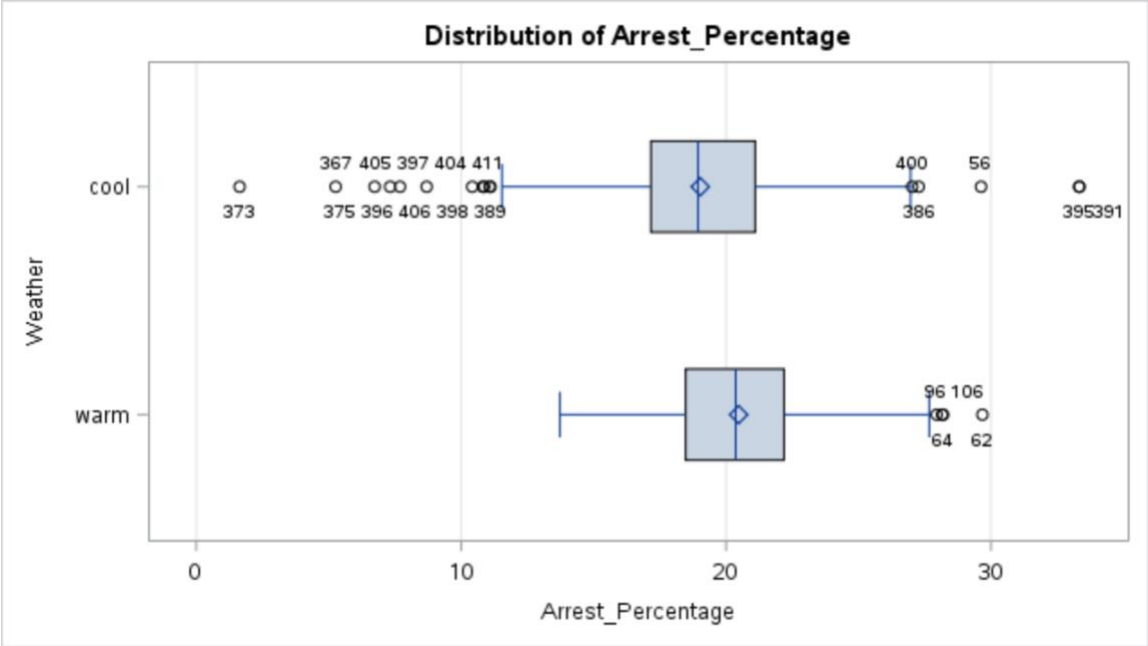

Figure 3 - Pie Chart of Most Reported Crimes in Chicago

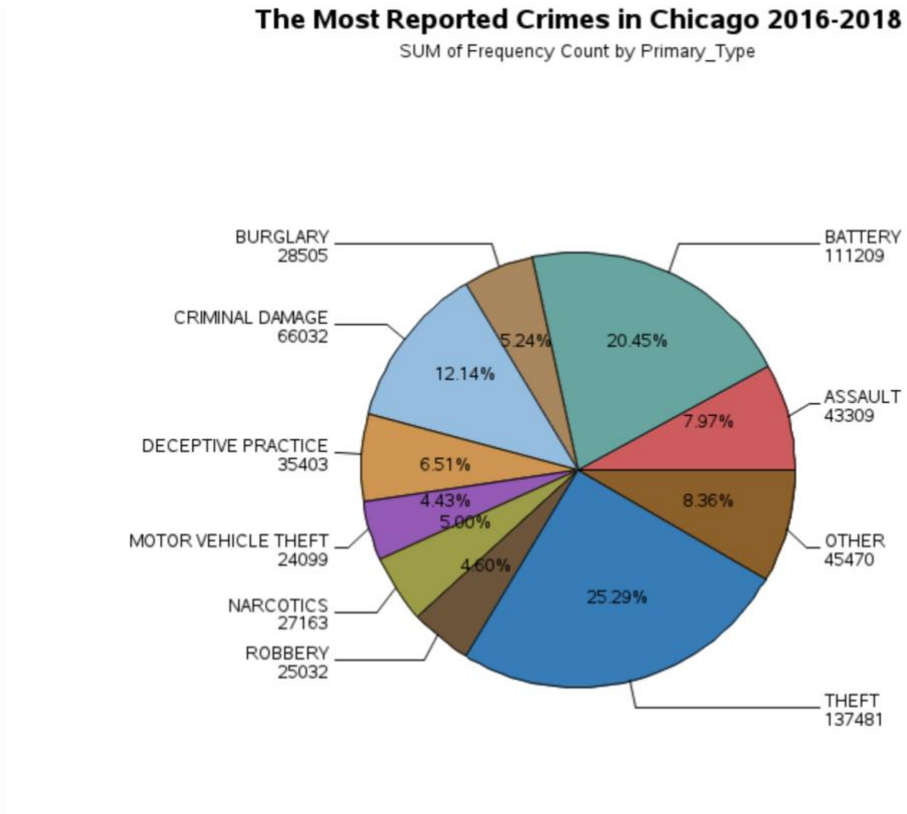

Table 2 – Contingency Table

The Most Reported Crimes in Chicago by Arrest

The FREQ Procedure

Frequency Expected Deviation Row Pct

Table of Primary_Type by Arrest			
	Arrest		
Primary_Type	FALSE	TRUE	Total
THEFT	123014 114710 8304.2 89.48	14467 22771 -8304 10.52	137481
BATTERY	88385 92789 -4404 79.48	22824 18420 4404.3 20.52	111209
CRIMINAL DAMAGE	62185 55095 7090 94.17	3847 10937 -7090 5.83	66032
ASSAULT	35565 36136 -570.7 82.12	7744 7173.3 570.67 17.88	43309
DECEPTIVE PRACTICE	33815 29539 4275.8 95.51	1588 5863.8 -4276 4.49	35403
BURGLARY	27284 23784 3500.3 95.72	1221 4721.3 -3500 4.28	28505
NARCOTICS	13 22664 -22651 0.05	27150 4499 22651 99.95	27163
ROBBERY	23202 20886 2316.1 92.69	1830 4146.1 -2316 7.31	25032
MOTOR VEHICLE THEFT	22247 20107 2139.5 92.32	1852 3991.5 -2140 7.68	24099
Total	415710	82523	498233

Table 3 - Chi-Square Test on Crimes and Arrest Status

Statistics for Table of Primary_Type by Arrest

Statistic	DF	Value	Prob
Chi-Square	8	156904	<.0001
Likelihood Ratio Chi-Square	8	122466	<.0001
Mantel-Haenszel Chi-Square	1	8783	<.0001
Phi Coefficient		0.56118	
Contingency Coefficient		0.48938	
Cramer's V		0.56118	

Sample Size = 498233

Figure 4

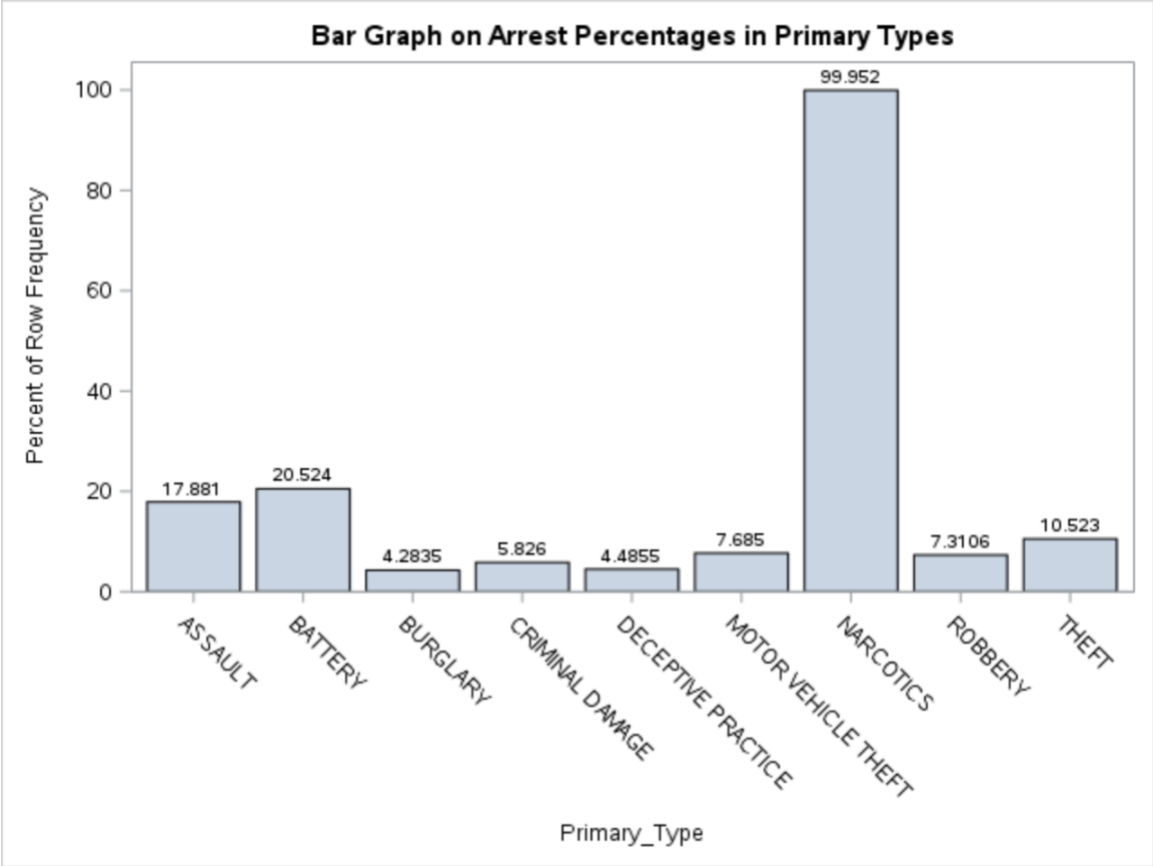

Figure 5

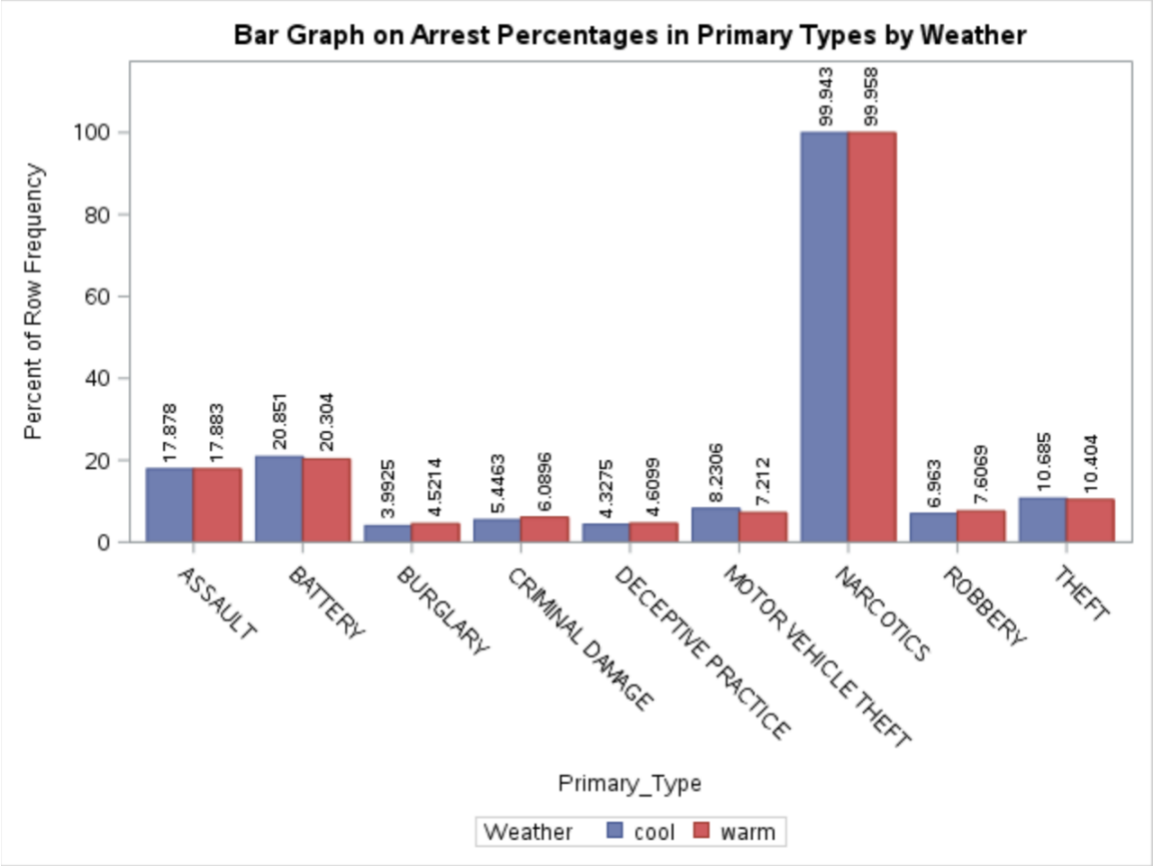

Table 4.1 - Contingency Table for Crimes and Weather Based on Arrests

The Most Reported Crimes in Chicago by temperature

The FREQ Procedure

Frequency
Expected
Deviation
Row Pct

Table of Primary_Type by Weather			
Primary_Type	Weather		
	warm	cool	Total
NARCOTICS	16565 16134 430.75 61.01	10585 11016 -430.7 38.99	27150
BATTERY	13531 13563 -32.47 59.28	9293 9260.5 32.467 40.72	22824
THEFT	8255 8597.2 -342.2 57.06	6212 5869.8 342.21 42.94	14467
OTHER OFFENSE	4692 4640.6 51.397 60.08	3117 3168.4 -51.4 39.92	7809
ASSAULT	4684 4602 82.024 60.49	3060 3142 -82.02 39.51	7744
CRIMINAL DAMAGE	2373 2286.1 86.869 61.68	1474 1560.9 -86.87 38.32	3847
MOTOR VEHICLE THEFT	931 1100.6 -169.6 50.27	921 751.42 169.58 49.73	1852
ROBBERY	1028 1087.5 -59.5 56.17	802 742.5 59.502 43.83	1830
DECEPTIVE PRACTICE	913 943.69 -30.69 57.49	675 644.31 30.69 42.51	1588
BURGLARY	709 725.6 -16.6 58.07	512 495.4 16.596 41.93	1221
Total	53681	36651	90332

Table 4.2 – Chi-Square

Statistics for Table of Primary_Type by Weather

Statistic	DF	Value	Prob
Chi-Square	8	149.4569	<.0001
Likelihood Ratio Chi-Square	8	148.3090	<.0001
Mantel-Haenszel Chi-Square	1	38.9302	<.0001
Phi Coefficient		0.0426	
Contingency Coefficient		0.0425	
Cramer's V		0.0426	

Sample Size = 82523

Figure 5

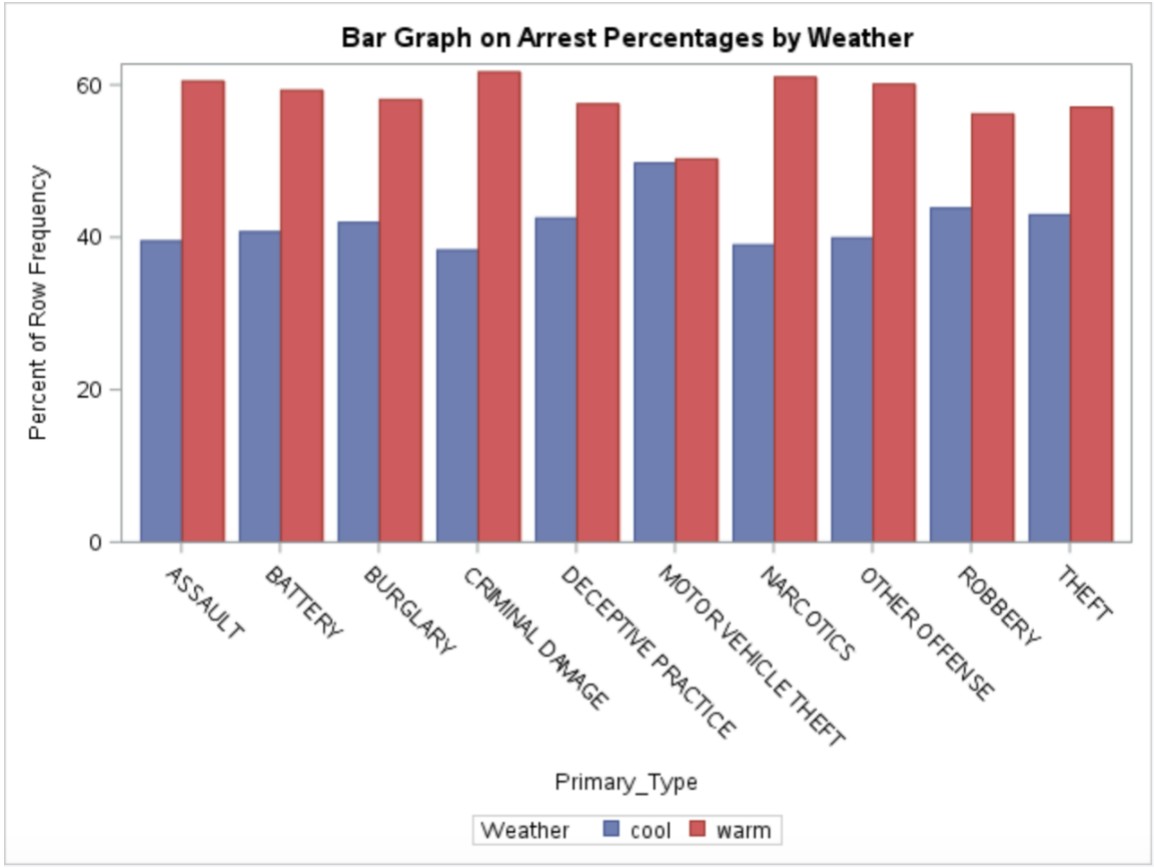

Table 5 – Upper-tailed T-test Percentage of Arrests for Domestic Crimes

The TTEST Procedure

Variable: arrest_ratio

temp	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Warm		492	17.1700	3.8145	0.1720	5.0633	28.5714
Cool		418	17.0730	7.3325	0.3586	0	100.0
Diff (1-2)	Pooled		0.0969	5.7061	0.3796		
Diff (1-2)	Satterthwaite		0.0969		0.3977		

temp	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Warm		17.1700	16.8321	17.5078	3.8145	3.5902	4.0691
Cool		17.0730	16.3680	17.7780	7.3325	6.8668	7.8664
Diff (1-2)	Pooled	0.0969	-0.5280	Infty	5.7061	5.4554	5.9813
Diff (1-2)	Satterthwaite	0.0969	-0.5583	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	908	0.26	0.3992
Satterthwaite	Unequal	603.7	0.24	0.4038

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	417	491	3.70	<.0001

Figure 6 – QQ Plot from T-test

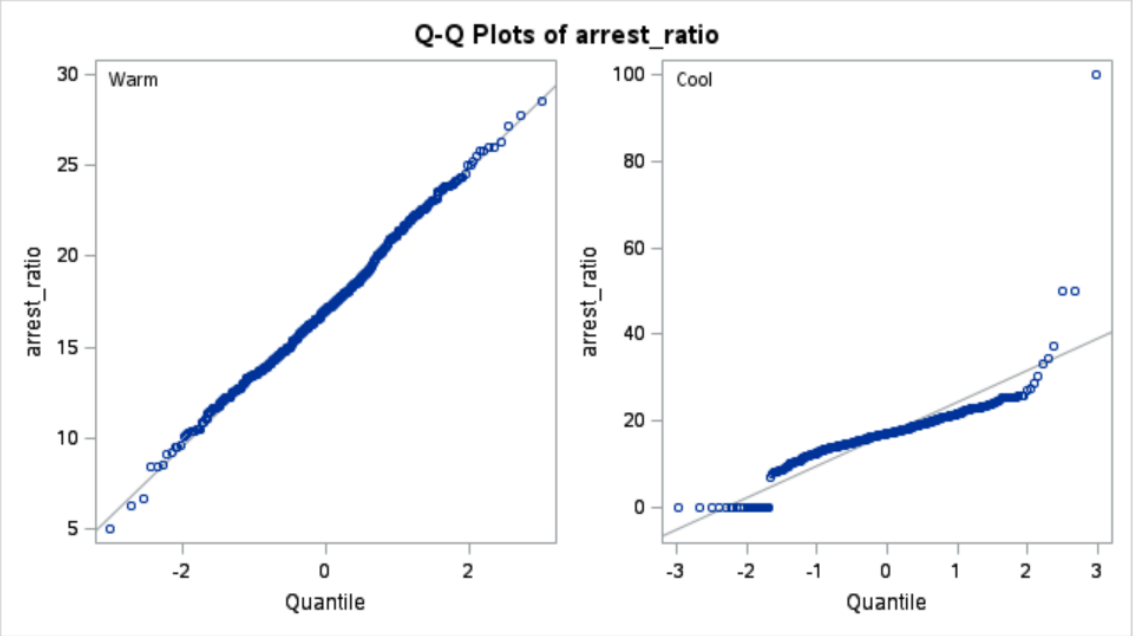

Figure 7 – Histograms from T-test

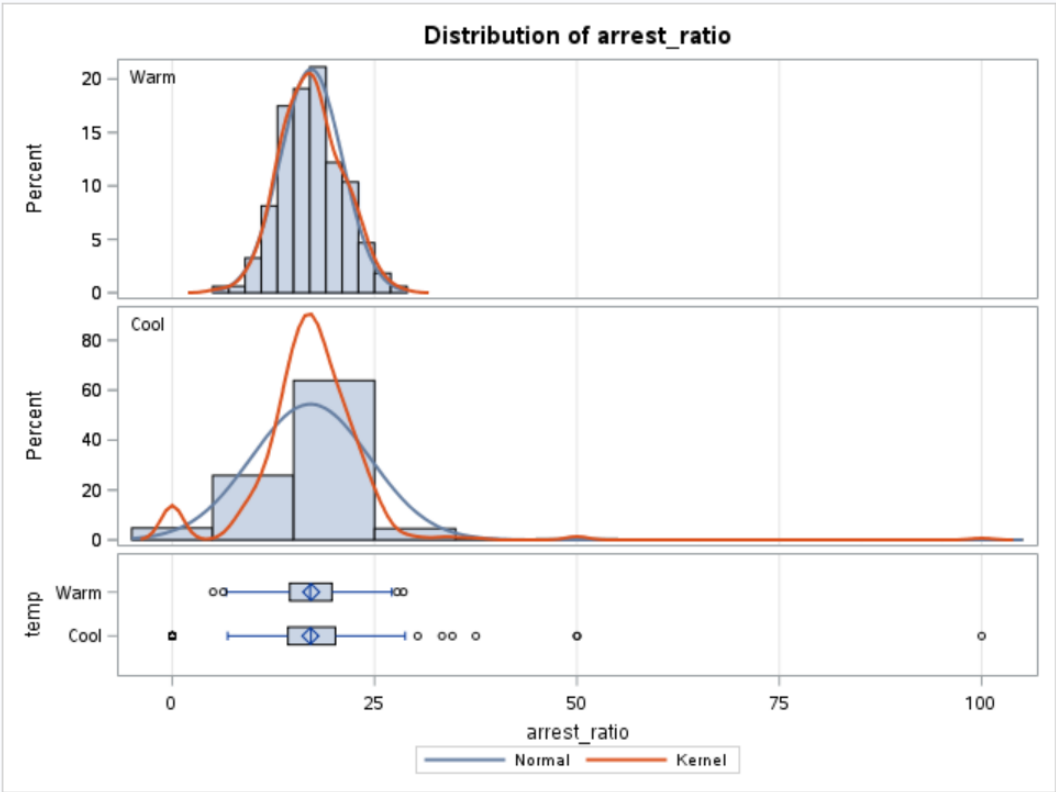

Figure 8

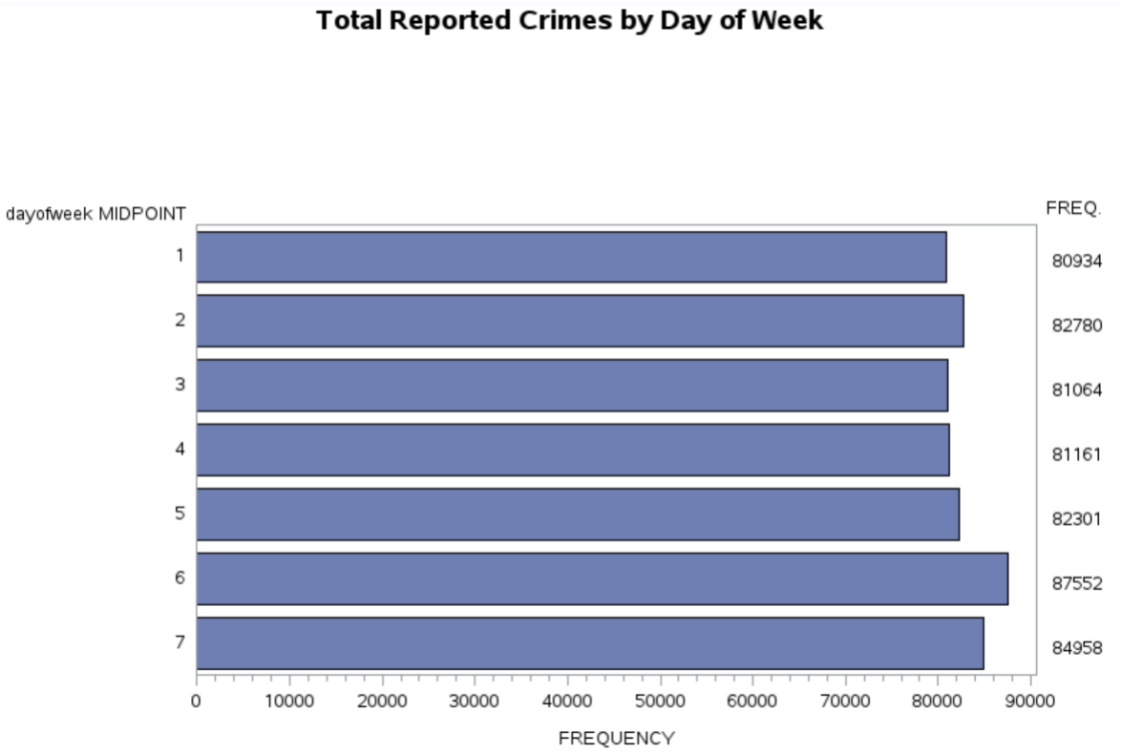

Table 6 – Two-Way ANOVA; Average Frequency of Crime per Day on Day of Week and Season

The GLM Procedure

Dependent Variable: COUNT Frequency Count

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	5841287.64	216343.99	14.71	<.0001
Error	886	13031074.56	14707.76		
Corrected Total	913	18872362.21			

R-Square	Coeff Var	Root MSE	COUNT Mean
0.309515	19.08667	121.2756	635.3939

Source	DF	Type I SS	Mean Square	F Value	Pr > F
season	3	5450387.320	1816795.773	123.53	<.0001
dayofweek	6	251453.439	41908.906	2.85	0.0094
season*dayofweek	18	139446.884	7747.049	0.53	0.9466

Source	DF	Type III SS	Mean Square	F Value	Pr > F
season	3	5450725.587	1816908.529	123.53	<.0001
dayofweek	6	255354.788	42559.131	2.89	0.0085
season*dayofweek	18	139446.884	7747.049	0.53	0.9466

Figure 9

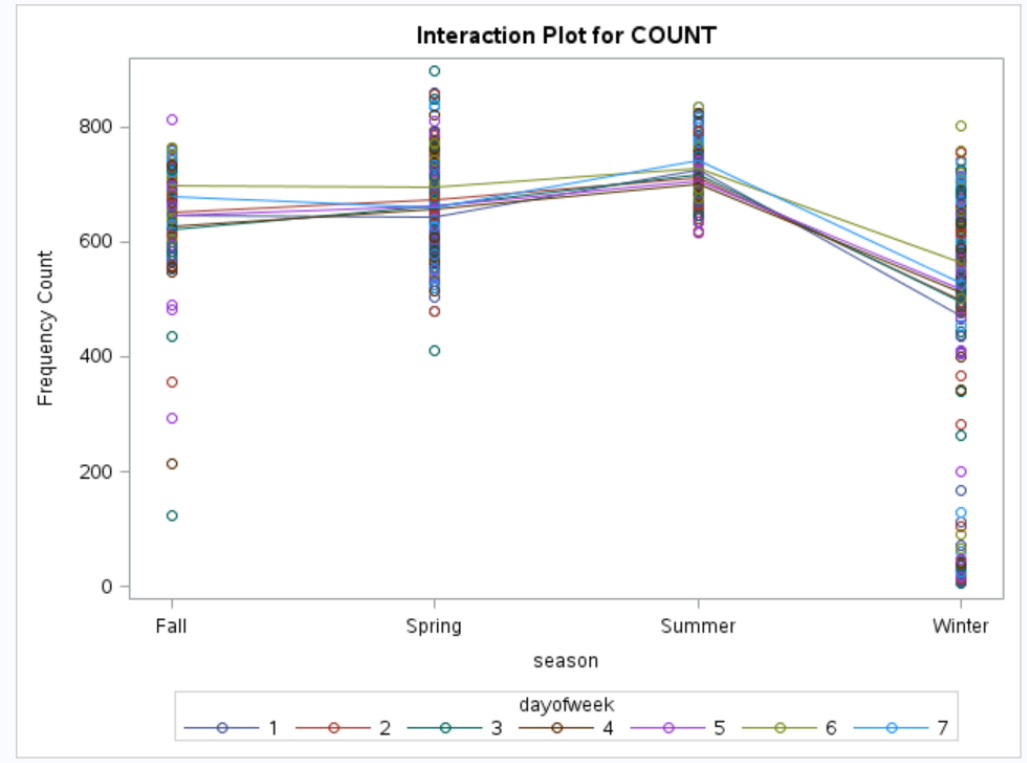

Figure 10

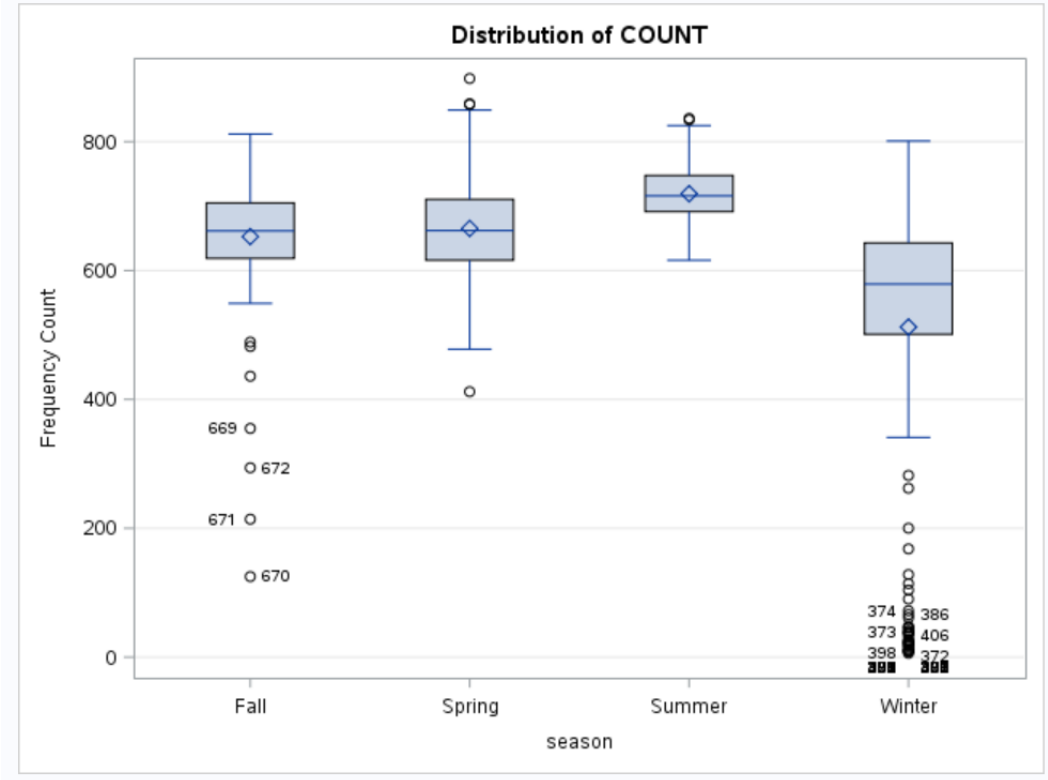

Figure 11 – ANOVA SNK Test

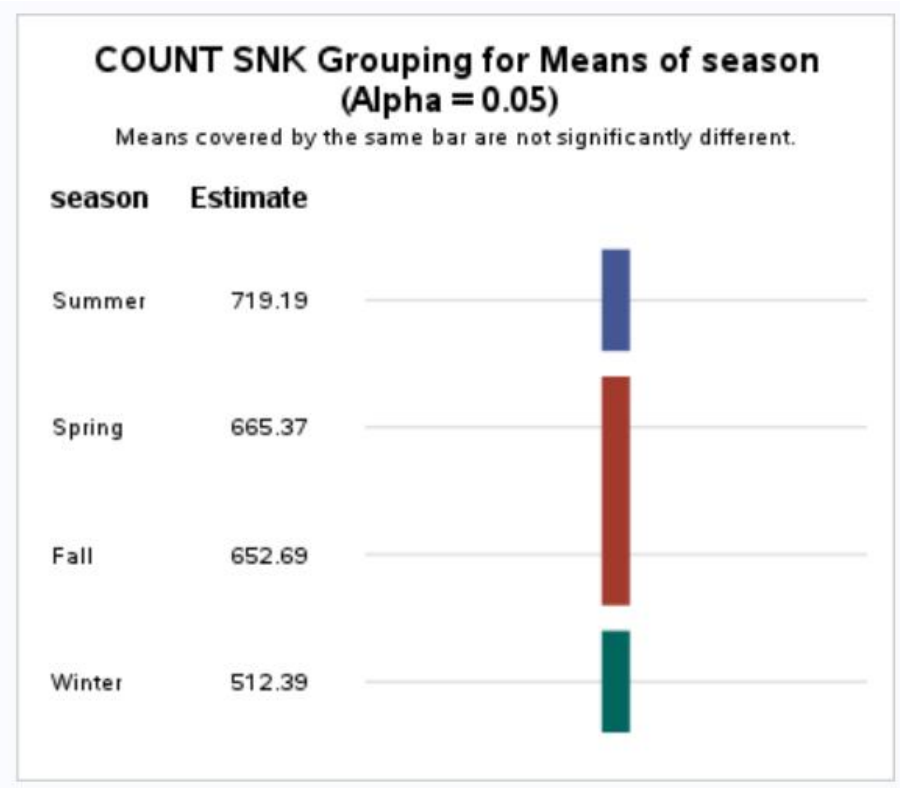

Figure 12 – ANOVA SNK Test

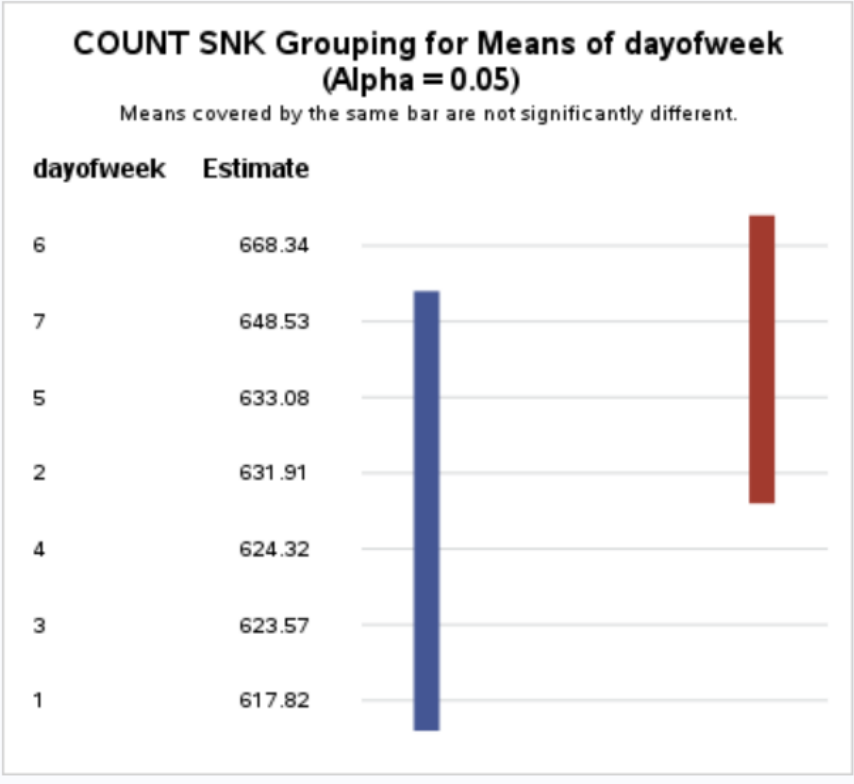

Table 7 – Overall Arrests Chicago January 2016 – July 2018

The FREQ Procedure

Arrest	Frequency	Percent	Cumulative Frequency	Cumulative Percent
FALSE	465237	80.11	465237	80.11
TRUE	115513	19.89	580750	100.00

References

Anderson, Craig A. "Heat and Violence." *Current Directions in Psychological Science* 10, no. 1 (February 2001): 33–38. doi:10.1111/1467-8721.00109.

Schinasi, L.H. & Hamra, G.B. J Urban Health (2017) 94: 892. https://doi.org/10.1007/s11524-017-0181-y

United States Census Bureau (2018). QuickFacts Chicago City, Illinois [Date File]. Retrieved from https://www.census.gov/quickfacts/fact/table/chicagocityillinois,US/PST045217.