Investigation of the Temperature and Traffic Volume Relationship in Chicago

Olivia M Rueschhoff

Advisor: Dr. Feridun Tasdan

Stats 471: Introduction to Mathematical Statistics

December 6, 2024

Abstract

A bivariate distribution is the shape of two data sets when they are put together, which can be used to help estimate where a new data point may land if the other variable is known. Several bivariate distributions were tested against the relationship between daily maximum temperatures and traffic volume in Chicago using data from 2006. While temperature and traffic are not commonly associated, extreme conditions can affect the traffic flow. Using data which are obtained from the National Centers for Environmental Information and Chicago.gov, this research explores whether a correlation exists and examines the underlying bivariate probability distributions of these two variables.

The data were analyzed using Excel and R programming, with graphical methods, correlation analysis, and goodness fit testing which are used as the core methodology. Explorative graphical analyses included scatterplots, histograms, and boxplots to visualize patterns and detect outliers. Advanced methods such as the Cullen and Frey graphs, Kolmogorov-Smirnov tests, and convex hull trimming were used to identify potential distributions and refine the analysis.

Results revealed a minimal correlation between maximum temperature and traffic volume, with an estimated correlation coefficient of r=0.12 (R 2 = 0.0144). After outlier removal, the correlation coefficient improved slightly to r=0.1412. Distribution analysis indicated that maximum temperature data resembled a uniform or beta distribution, while vehicle volume aligned more closely with gamma or log-normal distributions. Spearman's Rank Correlation Test confirmed a weak but statistically significant association between the variables, with a rho (ρ) value of 0.165 (p &lt; 0.05).

Despite identifying possible bivariate distributions, the weak correlation suggests that temperature and traffic volume are not strongly interrelated, making predictions based on these variables unreliable. However, the findings contribute to the understanding of bivariate distributions and the impact of outlier treatment on correlation results. Limitations of the study include potential data inaccuracies and missing

entries, notably for summer months, which may have influenced the results. Moreover, the study could be explored further if the data coverage is extended into several years instead of relying on 2006 only. While this knowledge may not help accurately predict anything, it is a good way to look at and analyze large data sets and try to determine their distributions.

1. Introduction

Temperature and traffic jams are not typically thought of as going hand in hand, but when truly considered extreme temperatures can affect traffic flow. But does the temperature tend to have any correlation to traffic on an average day? For this research, data from the City of Chicago was used to analyze the maximum temperature per day in 2006 as well as the number of cars involved in traffic jams on those same days. This data was then analyzed and tested for correlation, and the type of distribution both data sets followed was determined. With these sample sets and the analysis of the data sets, a conclusion on traffic and weather correlation in Chicago can be made.

2. Methods Used in Study

The data sets were put into an Excel spreadsheet to better compare and graph them. The temperature daily maximum in Chicago from 2006 is from the National Centers for Environmental Information, and the vehicle volume for traffic in 2006 was found on Chicago.gov. With the data now accessible to be prepared, testing began. Testing consisted of making a variety of graphs, finding the correlation variable, and checking for distributions using R-program.

2-a. Graphical Methods

To understand the data points and then examine them for correlation and distribution, six graphs were made. The first graph was date versus vehicle volume as a scatter plot to quickly see if there is any part of the year with more or less traffic. Similarly, the second graph was the date versus temperature maximum for that day to see the shape of the data plot. The next two graphs were temperature maximum versus vehicle volume and vehicle volume versus temperature maximum. These graphs were constructed to see if a correlation between the data sets can be visibly seen by graphing. They were both constructed because one might have possibly had a clearer shape to it than the other. With these graphs, Excel can calculate an R-squared value. The last two graphs constructed in Excel were histograms for vehicle volume and temperature maximum. By having these two histograms, the shape of the data sets is more apparent, and some assumptions can be made about their distributions, such as whether they are normally distributed.

2-b. Correlation Analysis

Excel is an Excellent tool for calculating the correlation between two data sets. Correlation tells us if two variables share a relationship or not and the intensity of that relationship. The first way correlation was found was once the maximum temperature and vehicle volume scatter plot was constructed, the built-in Excel function to show the linear R-squared value of the plot was used. The R-squared value shows the intensity of the relationship as it produces a number between zero and one. If the R-squared value is closer to zero, then there is little to no correlation and the closer the value is to one, the greater the correlation. While this value is for if the data sets have a linear correlation, it can still be a good starting point to estimate whether the data is going to have any correlation. The second way the correlation between the two data sets was calculated was using Excel's data analysis add-on. Within this tab, the data sets can be input, and Excel will produce a table to show the correlation between each variable input. Additionally, a boxplot was constructed to check both data sets for outliers. If there were outliers to be found, they would be removed, and a new R-squared value and correlation table were constructed and compared against those with the outliers.

2-c. R-program Implementation

Once both data sets were imported into R-program, they were easily running together and separately to see if they matched any bivariate and univariate distributions. The three main ways tested to try and tell the distribution from the data sets were Cullen and Frey Graphs, Kolmogorov-Smirnov Test, and bivariate box plotting and visualization. Spearman's Rank Correlation Analysis was lastly done to confirm findings of significant or insignificant association between the volume and temp variables. The Cullen and Frey graphs were constructed for each data set separately after researching that they are used to plot "the skewness-kurtosis [where] the skewness of the distribution indicates how symmetrical the distribution is [and] kurtosis stands for a measure of whether the distribution is peaked" (Quaresma, Pereira, & Fireman, 2021). Once the *fitdistrplus* package was installed into R-program the data sets were individually plugged into *descdist* to generate the graphs. Then both data sets were plotted to test for normal and uniform distributions as a double check. The Kolmogorov-Smirnov test is "a type of non-parametric test of the equality of discontinuous and continuous a 1D probability distribution that is used to compare the sample with the reference probability test or among two samples" (Kolmogorov-Smirnov Test in R Programming, 2023). For this data, we are interested in the two-sample test. This was done by importing the *dgof* package into R-program and then using it to plug both max temperature and vehicle volume into *ks.test* and plotted for visualization. To visualize the bivariate distribution between maximum temperature and vehicle volume in Chicago, the book An Introduction to Applied Multivariate Analysis with R was heavily consulted, specifically section 2.2 Looking at Multivariate Data: Visualisation: The Scatterplot. The first graph constructed was a visualization of maximum temperature versus vehicle volume and a bivariate boxplot. Using the book as guidance, a chart consisting of a scatterplot, histogram, and boxplot were made to give a side-by-side comparison. The second graph constructed was a scatterplot showing the convex hull of the data. A convex hull of bivariate data "allows robust estimation of the correlation" and "can eliminate isolated outliers without disturbing the general shape of the bivariate distribution" (Everitt & Hothorn, 2011). This can show whether the correlation between the data sets will further increase with the convex hull removed and can help determine the shape of the distribution. Lastly, to test whether or not the correlation between variables is truly significant, Spearman's Rank Correlation Analysis was conducted. This tests the correlation by comparing it with alpha=0.05 level of significance. If P value< alpha, reject the null hypothesis, Ho, being there no significant correlation between the variables, in favor of the alternative hypothesis, Ha, meaning a significant correlation exists between variables (How to Perform Spearman, 2024). If the P-value is greater than alpha=0.05, do not reject Ho.

3. Results of the Data Analysis

Vehicle volume in traffic and maximum daily temperature are not thought of as having an impact on each other. So, it is no surprise that there was a very small correlation found. However, that does not mean these data sets do not show anything interesting or noteworthy.

3-a. Graphical Analysis

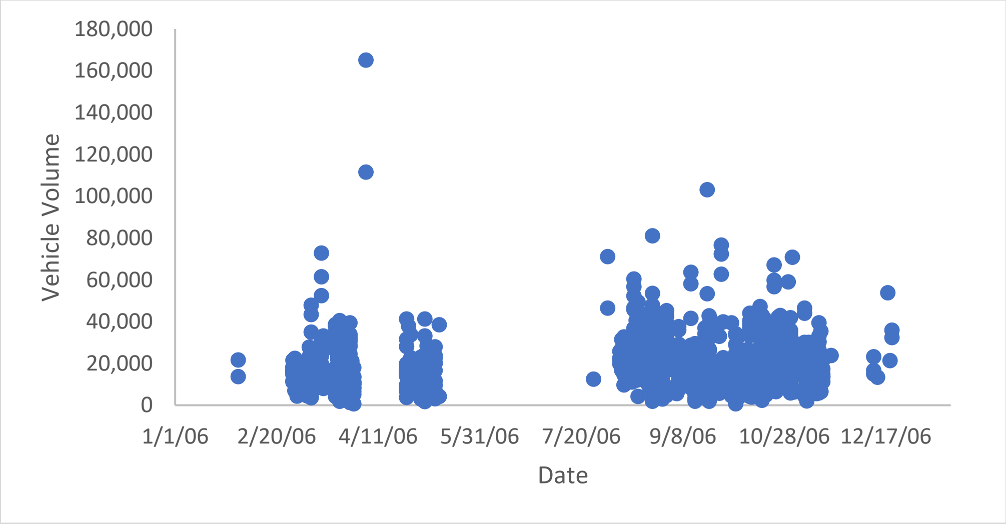


Figure 1: Date Versus Vehicle Volume

The first graph, Figure 1, is time versus vehicle volume to see if any time of the year had a visible data distribution or recognizable pattern. Figure 1 shows a massive gap in the data between May and August, suggesting there were no traffic jams during that period. However, that is an illogical assumption, so it is more likely that whoever

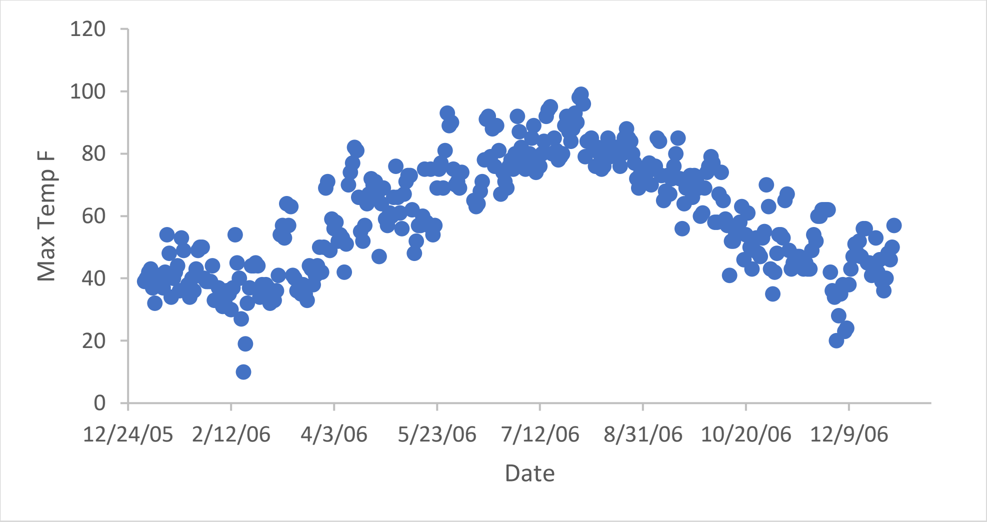


Figure 2: Date Versus Maximum Temperature

collected this data misreported or skipped the vehicle volume information for these months resulting in it missing from our data set. This is an error which will be considered in the final analysis.

The second graph, Figure 2, is time versus maximum temperature in °F. This graph shows an expected relationship where the maximum temperature fluctuates over time being hotter in the summer months and colder in the winter months. This scatterplot appears to have a sinusoidal curve to it, which makes sense as temperature fluctuates throughout the year.

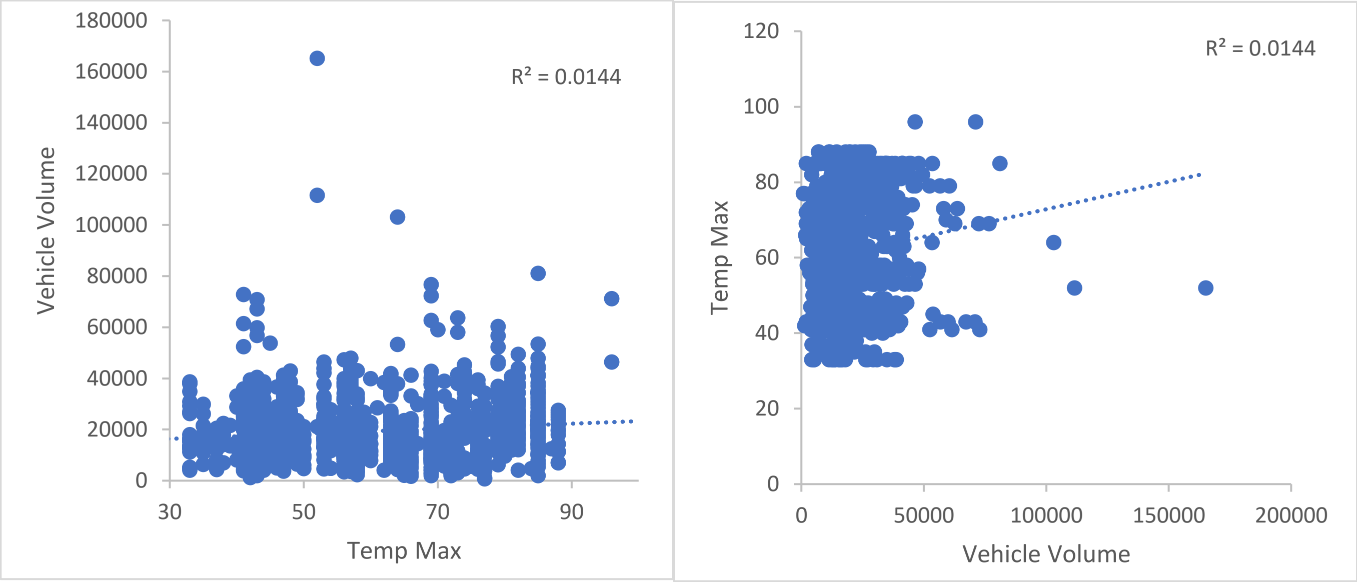


Figure 3: Temp Max Vs Vehicle Vol

Plotting the maximum daily temperature against the corresponding vehicle volume that day can help show whether there exists a correlation between the two variables. As seen in *Figure* 3, there is no obvious cure or model that follows the data regardless of which variable is the independent and which is the dependent variable. This

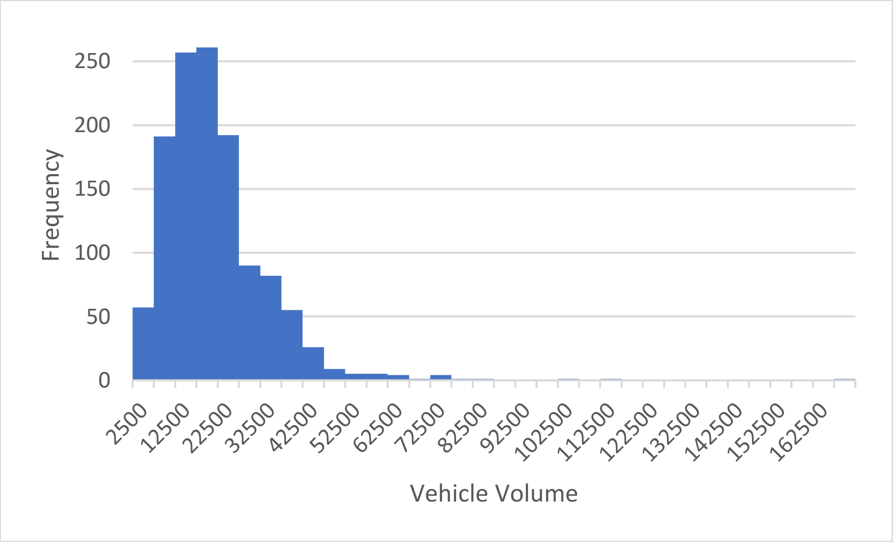


Figure 4: Histogram of Vehicle Vol

means there is no model to be generated that can be used to accurately predict one variable from the other. Note that the data appears to take on a slight appearance of a uniform distribution in the left plot of Figure 3, which was further investigated in R-program.

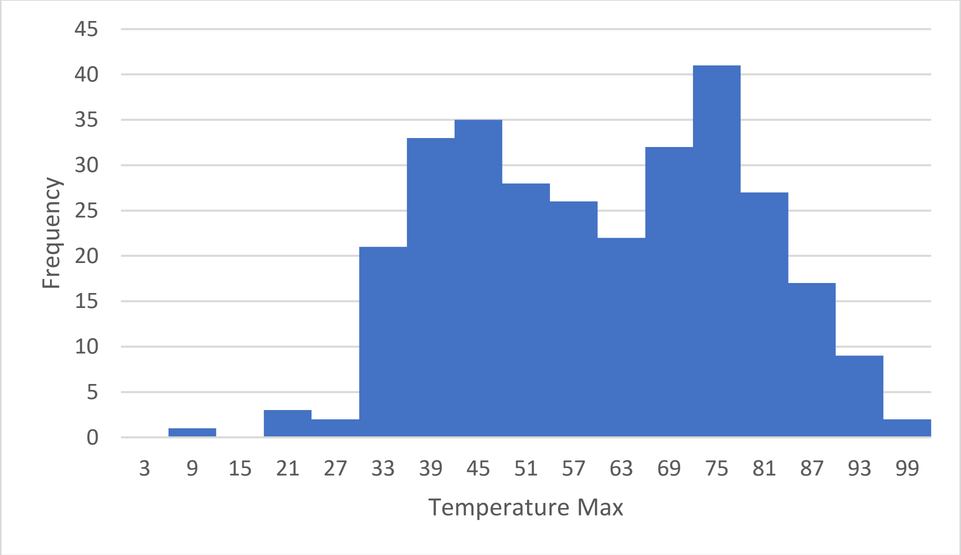


Figure 5: Histogram of Maximum Daily Temperature

The last graphs constructed were histograms for both data sets, Figures 4 and 5. Histograms can generally be a good start to see the shape of a distribution. For Figure 4, the shape of this histogram is right-skewed with possible outliers. These outliers will be further investigated. Figure 5 looks to have a bimodal shape with no outliers. This shape will be helpful when further exploring the distributions. Despite not being able to accurately predict one variable from the other, the distributions of each set and their importance to the time of year can be investigated further.

3-b. Correlation Testing

The R-squared value can be a starting point in seeing whether there exists a correlation between two variables. Figure 3 shows the R-squared value of the maximum temperature versus vehicle volume being 0.0144. This value is very close to 0 which suggests little to no correlation between the data sets. This value is especially low considering our data set has well over 1000 data points being analyzed.

Correlation	Temperature Maximum °F	Total Passing Vehicle Volume
Temperature Maximum °F	1	0.120039383
Total Passing Vehicle Volume	0.120039383	1

Table 1: Correlation Table between Maximum Temperature and Vehicle Volume

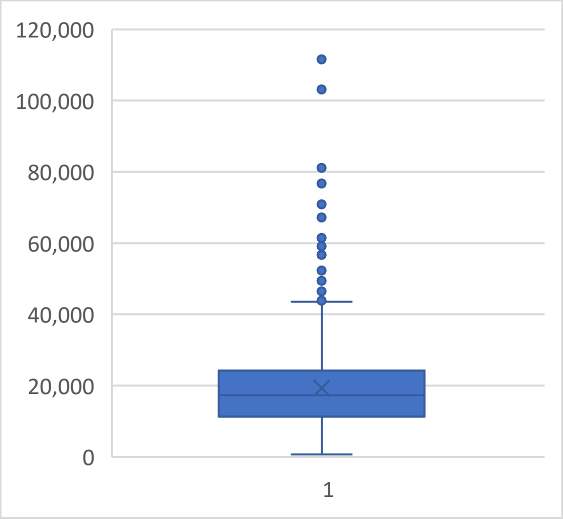


Figure 6: Vehicle Volume Boxplot to see outliers

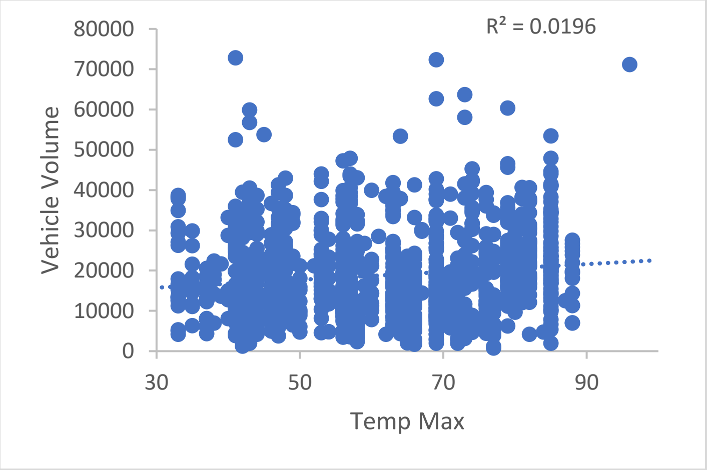


Figure 7: Temp Max Vs Vehicle Vol, outliers removed

Then the correlation table, table 1, shows that the correlation coefficient is 0.1200 for temperature and vehicle volume. This suggests the two data sets have a very small relationship as the ideal relationship between variables is as close to one as possible. While this is larger than the R-squared value that was first examined as a starting point to gauge correlation, this is still not a very significant value to suggest a relationship. However, it was noticed that the vehicle volume data contained several potential outliers, so a boxplot, Figure 6, was constructed to quickly gather those exact values. The outliers were then removed from the data set and the scatterplot and correlation table were calculated. The new R-squared value to estimate correlation is 0.0196, seen in Figure 7. Removing the outliers caused this value to increase by 0.005 for linear correlation. This value is not very significant, but it does show that the outliers were affecting the correlation of the variables. The new correlation table where the outliers have been removed shows an increase in the correlation coefficient of almost 0.02. This amount is somewhat significant which supports the idea that the outliers were negatively affecting the correlation and could also make it difficult to truly capture the distribution of the data. The correlation was also calculated in R program which was 0.1411972. This result is not

Correlation	Temperature Maximum °F	Total Passing Vehicle Volume
Temperature Maximum °F	1	0.139927138
Total Passing Vehicle Volume	0.139927138	1

Table 2: Correlation Table between Maximum Temperature and Vehicle Volume with outliers removed

very different from the other results, suggesting that while the outliers were affecting the correlation, it still is rather small.

3-c. Investigation of the Possible Distributions

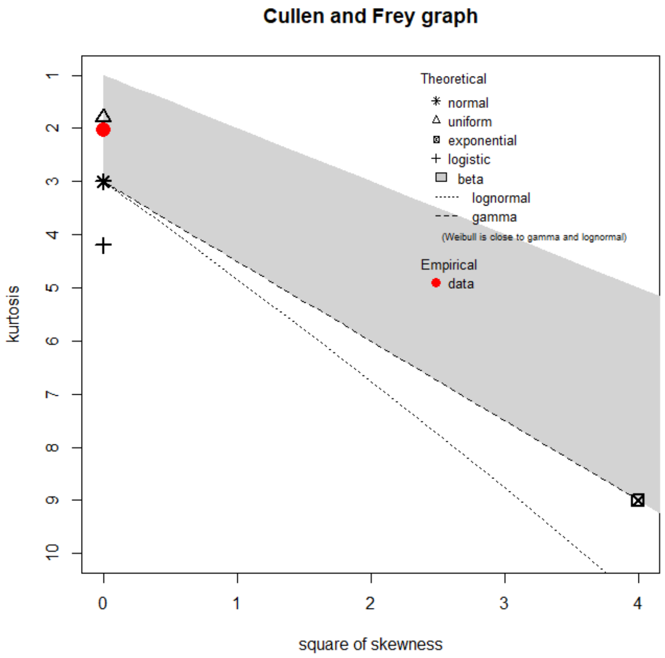


Figure 8: Cullen and Frey Graph of Maximum Temperature

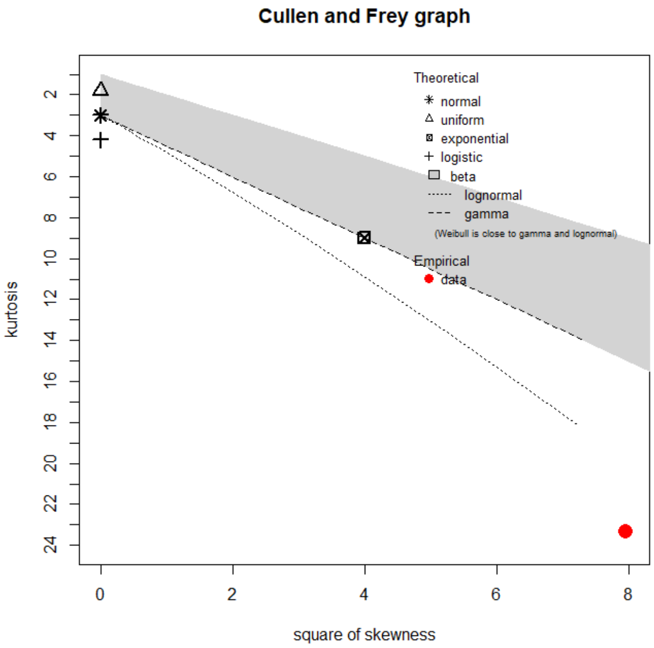


Figure 9: Cullen and Frey Graph of Vehicle Volume

The two Cullen and Frey graphs, Figures 8 and 9, were generated using R-program. Figure 8 shows that the maximum temperature data set is within the area to be considered a beta distribution, but even more interestingly, it is very close to the triangle which suggests that the maximum temperature has a distribution close to uniform. The Cullen and Frey graph in Figure 9, however, suggests that the vehicle volume data set follows none of the distributions it was tested for. This means it is not normal, uniform, exponential, logistic, beta, lognormal, or a gamma distribution. While there are many different statistical distributions, the vehicle volume does not follow any of the most commonly occurring.

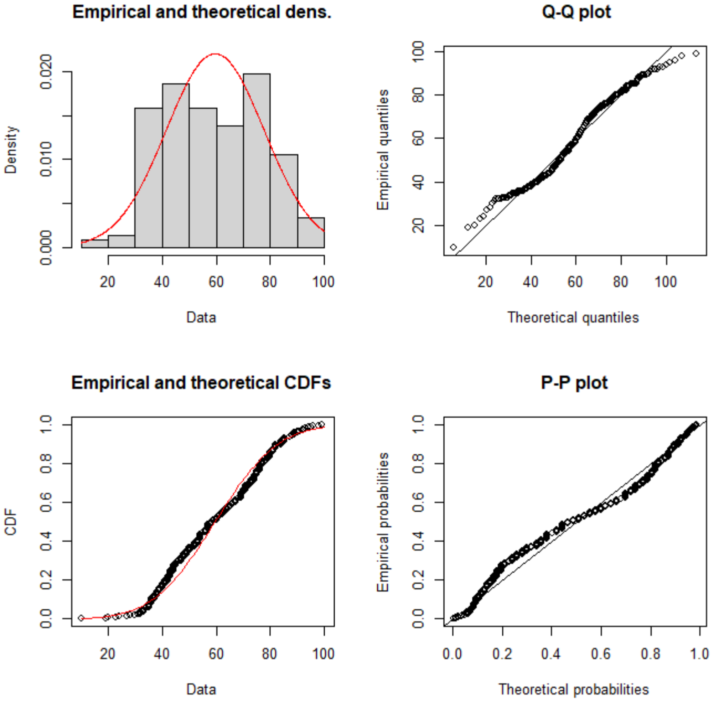


Figure 10: Temp Normal Distribution

Next, the data sets were analyzed against the most common distribution, normal distribution, as well as uniform distribution since the maximum temperature's Cullen and Frey graph showed it was very close to uniform. The first one plotted was maximum temperature against the normal distribution in Figure 10. The data does fit well for things such as the cumulative distribution function (CDF), the Q-Q plot, which helps asses "if a set of data plausibly came from some theoretical distribution" (Ford, 2015), specifically comparing "the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions" (Comparison of P-P plots and Q-Q plots, 1999) and the P-P plot which "compares the empirical cumulative distribution function of a data set with a

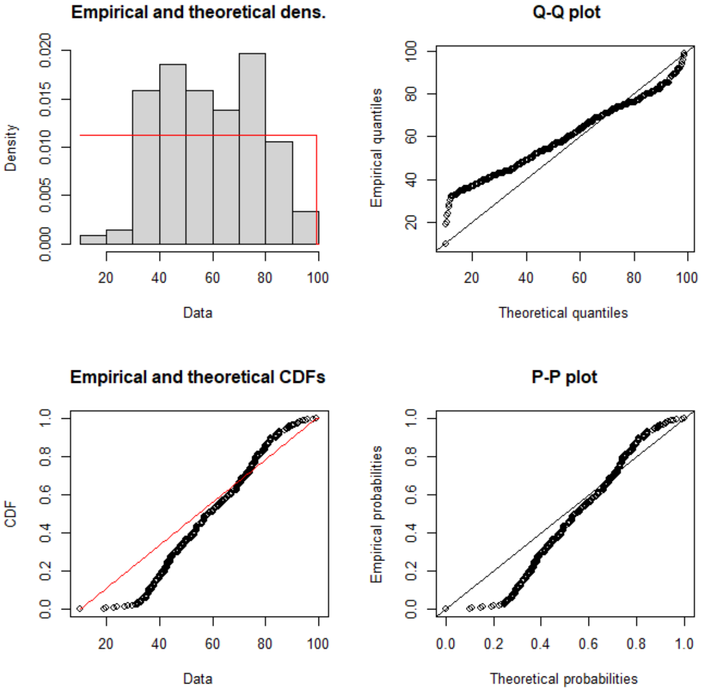


Figure 11: Temp Uniform Distribution

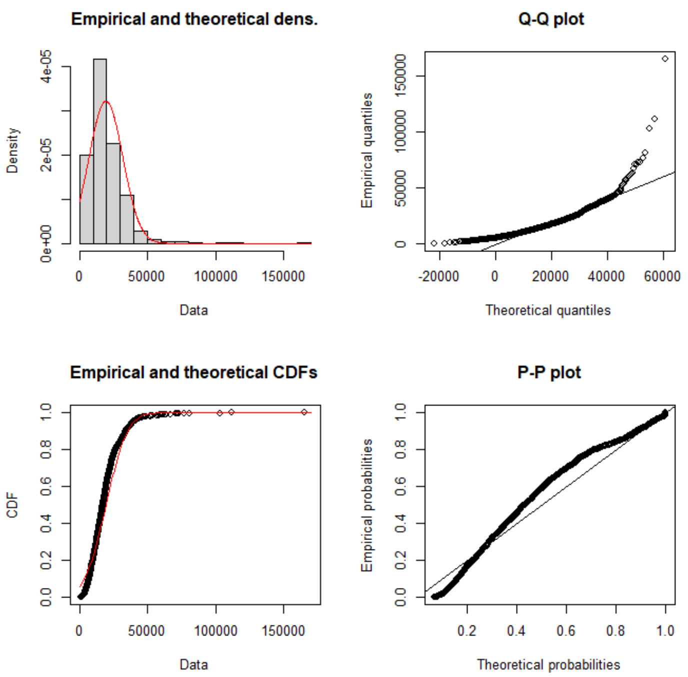


Figure 12: Vehicle Volume Normal Distribution

specified theoretical cumulative distribution function" (Comparison of P-P plots and Q-Q plots, 1999). However, it clearly does not match the density of the function which greatly suggests the maximum temperature data is not normally distributed. Then the temperature was plotted against a uniform distribution, shown in Figure 11. While the density of this plot seems to somewhat follow the uniform distribution, it still is not great and the cumulative distribution function, P-P plot, and Q-Q plot don't resemble that of a uniform distribution, as they have large areas that don't follow the line. The vehicle volume was then plotted against normal and uniform distributions. The normal distribution plot for vehicle volume, Figure 12, shows a somewhat normally distributed density and cumulative distribution function, however, the P-P and Q-Q plots have shapes that don't follow the normal distribution's probability. Next, vehicle volume was checked against uniform distribution in Figure 13. The data set does not follow a uniform distribution at all, evident by how it does not align with the density, CDF, Q-Q plot, or P-P plot. While maximum temperature distribution

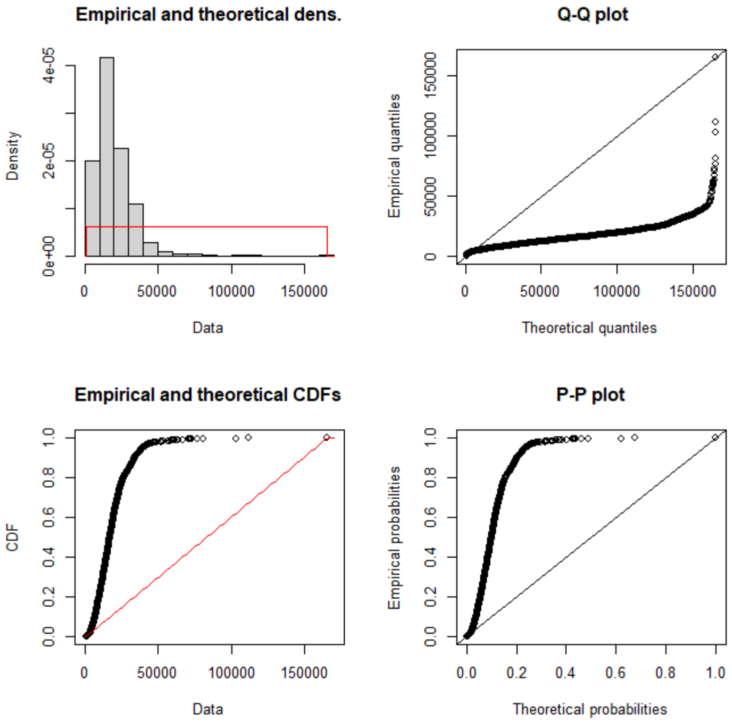


Figure 13: Vehicle Volume Uniform Distribution

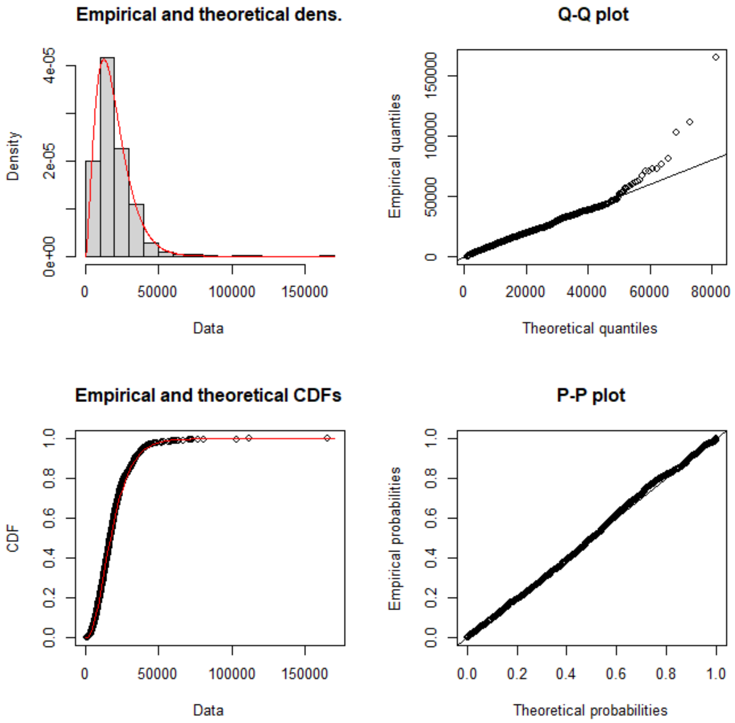


Figure 14:Vehicle Volume Gamma Distribution

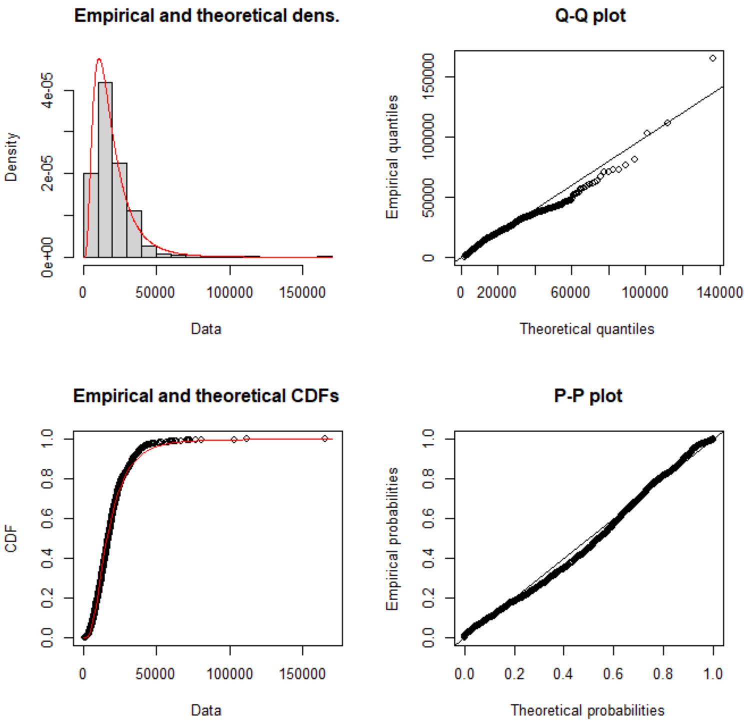
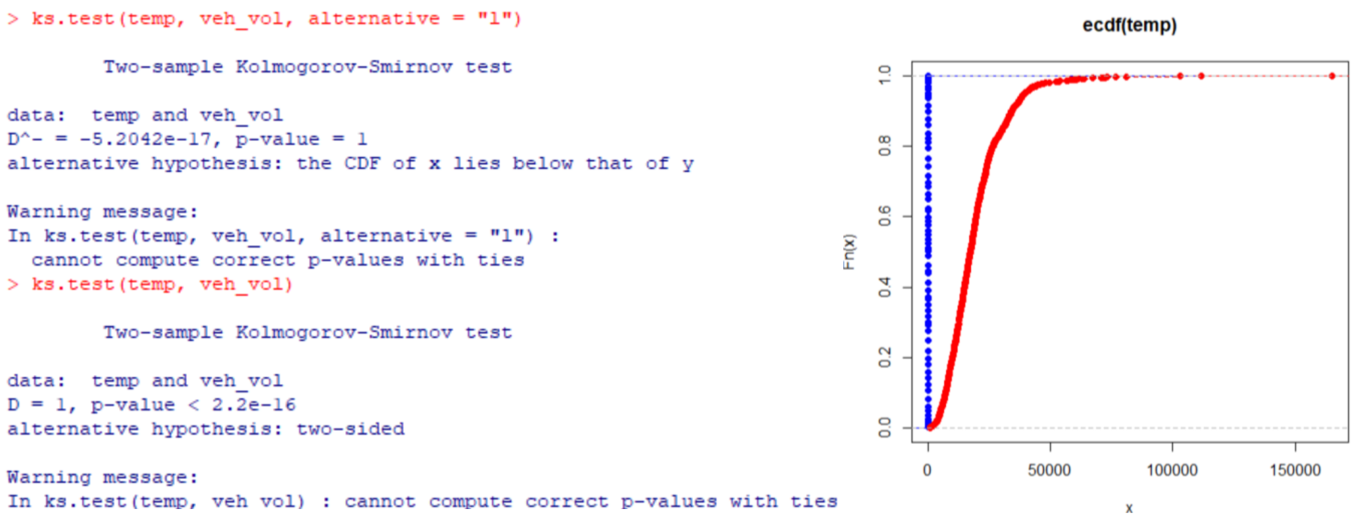


Figure 15: Vehicle Volume Log-normal Distribution

appears to be similar to normal and uniform distributions, a similar distribution for vehicle volume was not discovered. So, the data set was run against several other distributions to see if any were even close to similar. It was found that the vehicle volume closely resembled both a gamma distribution and a log-normal distribution, as seen in Figure 14 and 15. While the vehicle volume still does not perfectly match either of the distributions, it is a lot better for all four plots for both distributions than it was for the normal and uniform distributions.



```
> ks.test(temp, veh_vol, alternative = "l")

        Two-sample Kolmogorov-Smirnov test

data:  temp and veh_vol
D^- = -5.2042e-17, p-value = 1
alternative hypothesis: the CDF of x lies below that of y

Warning message:
In ks.test(temp, veh_vol, alternative = "l") :
  cannot compute correct p-values with ties
> ks.test(temp, veh_vol)

        Two-sample Kolmogorov-Smirnov test

data:  temp and veh_vol
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Warning message:
In ks.test(temp, veh vol) : cannot compute correct p-values with ties
```

Figure 16: Code results and Graph for Kolmogorov-Smirnov test between both data sets

Once the distribution for each data set was estimated, then the bivariate distribution between both of the data sets could be examined. The Kolmogorov-Smirnov test was conducted, with the outlier values removed. The output code and the graph, in Figure 16, show that the empirical distribution functions for each data set are seemingly far apart from each other. This suggests that both data sets do not share the same distribution type. While this was already concluded through previous tests, the idea is strengthened so some bivariate distributions can be eliminated such as the normal bivariate distribution where it requires that both data sets be normally distributed.

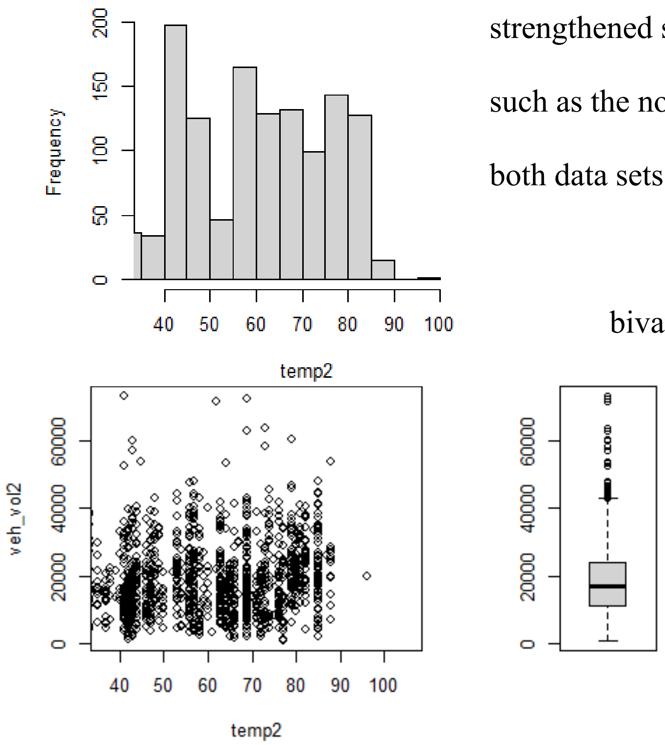


Figure 17: Side-by-side plots for bivariate distribution

In order to truly visualize whether or not there is a bivariate distribution between the data sets, a side-by-side boxplot, scatterplot, and histogram were constructed in Figure 17. These plots were constructed with the outlier values removed. While it does show that once the outliers were removed, more were discovered because all of the quartiles and maximum shifted, it does also take on a very blocky shape. The histogram matches the scatterplot very closely and the scatterplot makes sense for the boxplot. This suggests that there is a bivariate distribution between the data sets as all the plots make sense and show a similar picture of the shape. To finish assessing the bivariate distribution between the data, convex hull trimming was used to get a better assessment of the correlation between the data sets. Once the convex hull was found in R-program. It was then plotted against the data sets in a scatter plot in Figure 18. Now that the convex hull can be visualized against the data, the correlation coefficient can be recalculated. The correlation coefficient in R-program after removing the outliers but before removing the convex hull was 0.141197. The correlation coefficient after the convex hull was removed is 0.148166. This is only about a 0.007 difference in correlation coefficient, which is not very significant. This suggests that while there might be a bivariate distribution between traffic vehicle volume and maximum daily temperature, it does not have a strong correlation and it would not be advised to try to model either traffic or temperature from the other variable following the distribution.

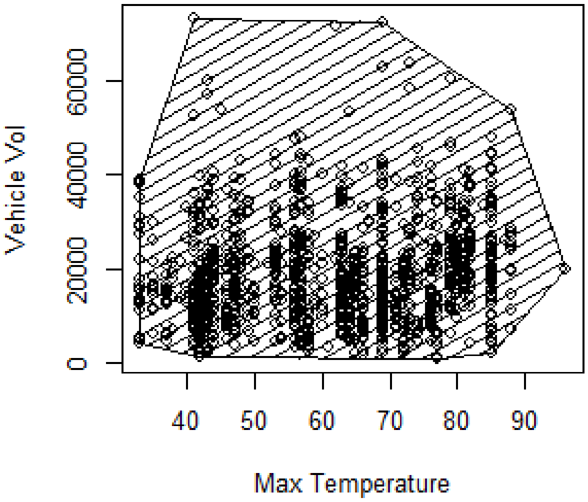
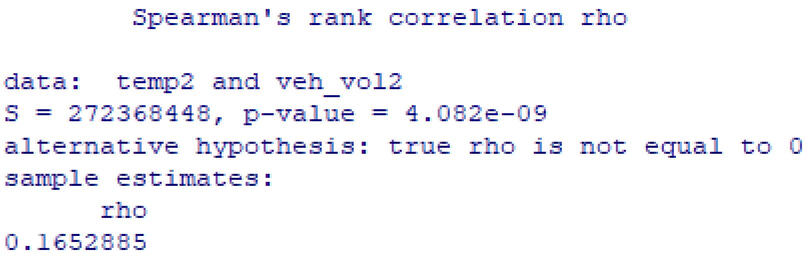


Figure 18: Convex Hull Scatter Plot

Finally, in order to see if the correlation is significant or not Spearman's Rank Correlation test was used in R program. Figure 19 shows the results of the correlation test. The rho found is 0.165, which even though it is a very small or weak correlation, this rho is



```
                    Spearman's rank correlation rho

data:  temp2 and veh_vol2
S = 272368448, p-value = 4.082e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1652885
```

Figure 19: Spearman's Rank Correlation Test Results

statistically significant at alpha=0.05, meaning the null hypothesis should be rejected and that the correlation is significant as proven by Spearman’s rank correlation test.

4. Possible Errors

While the correlation tested very low and not much is to be concluded with high certainty, these data sets could be compared again in the future and have better results if several changes were made. The first error is the validity of the data sets. While they were both taken from internet databases for the same location and year, the accuracy of the data reporting is questionable, especially for the vehicle volume data set. This data set is allegedly daily traffic reports, yet the dates tend to only be at the end of the month, and the whole months of June and July are missing from this data set. It is hard to believe that Chicago had no traffic jams during these months, especially since it is summer. Another possible error is in the reporting, in order to use this data frame in both Excel and R-program, they had to go from a pdf to Excel manually or be copied and pasted over. While mistakes were avoided as best they could, some of the data could’ve been entered inaccurately, or outliers were removed inaccurately due to human error. Overall, if these things were addressed, the chances of getting a stronger correlation and more apparent distribution could improve.

5. Conclusion

There is a possibility of having a uniform-gamma, uniform-lognormal, normal-gamma, or normal-lognormal bivariate distribution between the data. However, because the correlation continuously tested so low, but was proven somewhat significant through Spearman’s test, regardless of the distribution, it will not provide any accurate information whether it be percentile, probability, vehicle volume, or maximum temperature. The data sets were carefully examined in Excel and R-program with 17 graphs and 2 tables generated to show that there is a low correlation between data sets regardless of outliers and each data set resembles some distribution being either normal or uniform, or gamma or lognormal. While this knowledge does not help accurately predict anything, it is a good way to look at and analyze large data sets and try to determine distributions.

References

Chicago Department of Transportation. (n.d.). *Average daily traffic counts.* City of Chicago. https://www.chicago.gov/city/en/depts/cdot/dataset/average_daily_trafficcounts.html

Comparison of P-P plots and Q-Q plots. (1999, September). Retrieved from Simon Fraser University: https://www.sfu.ca/sasdoc/sashtml/qc/chap8/sect9.htm

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied.* (R. Gentleman, K. Hornick, & G. Parmigiani, Eds.) London, Munchen: Springer.

Ford, C. (2015, August 26). *Understanding QQ Plots.* Retrieved from University of Virgina Library: https://library.virginia.edu/data/articles/understanding-q-q-plots

Hogg, R. V., Tanis, E. A., & Zimmerman, D. L. (2021). In *Probability and Statistical Inference* (10th ed.). Pearson.

How to Perform Spearman. (2024). Retrieved from OnlineSPSS: https://www.onlinespss.com/spearmen-correlation-in-r/

Kolmogorov-Smirnov Test in R Programming. (2023, March 10). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/kolmogorov-smirnov-test-in-r-programming/

National Centers for Environmental Information. (2024). *Record of climatological observations: Chicago Northerly Island, IL US (USC00111550)*. U.S. Department of Commerce, National Oceanic and Atmospheric Administration.

Quaresma, D. F., Pereira, T. E., & Fireman, D. (2021). *Validation of a simulation model for FaaS performancebenchmarking using predictive validation.* Campina Grande: Federal University of Campina Grande. Retrieved from https://www.researchgate.net/publication/353060353_Validation_of_a_simulation_model_for_FaaS_performance_benchmarking_using_predictive_validation