Impact of Weather on Ridership of the Chicago Transit Authority

Dwight Brandt

Colorado State University Global

MIS581 – Capstone: Business Intelligence and Data Analytics

Dr. Kimberly A. Ford

November 7, 2021

Abstract

To better understand the impact that weather has on weekday transit ridership in the greater Chicago metropolitan area, research was conducted to assess whether a relationship exists between two types of weather, the amount of precipitation and snowfall, to that of transit ridership numbers on the Chicago Transit Authority and its two modes of transportation, bus and rail service. Daily summary counts of these modes were procured from the Chicago Data Portal. Daily weather data was acquired using the online data portal of the National Centers for Environmental Information, specifically the weather station for the Chicago O'Hare International Airport. As a whole, the data for the period of January 1, 2001 through December 31, 2019 was used in the analysis. Using SAS Studio, correlation analysis was conducted on a set of hypotheses that looked at each of the two weather variables and its relationship to each transit variable. The analysis found that overall, weather has a relationship to transit ridership on the CTA more specifically that as precipitation and snowfall increases, transit ridership on this system decreases. This overarching assessment aligns with past studies that have researched the impact of weather on other mass transit systems across the globe and have generally come to the same conclusion.

Impact of Weather on Ridership of the Chicago Transit Authority

Humanity has looked towards transportation as a means to get from one place to another throughout much of history. Over time, the modes have expanded to include such high-volume methods as bus and rail, capable of moving large numbers of people. Referred to as mass transit systems, they play an important role especially in urbanized settings (Farmer & Noonan, 2014) and to general economic growth of a region (Vajjarapu et al., 2020). As Kalkstein et al. (2009) explain, since 1980 many urban areas in the United States have invested heavily in modes like light rail systems in an effort to revitalize cities. In fact, bus ridership, long the most dominant in terms of popularity of the various types of mass transit, was surpassed by rail ridership in 2016. However, buses remain a dominant player in popularity (Merlin et al., 2021).

One component that is directly related to the success of any mass transit system is ridership, since often a large share of a budget that funds a mass transit system comes directly from those paying for the services in which they ride (Merlin et al., 2021). There are a multitude of factors that can impact ridership including the price of fuel, vehicle ownership, and accessibility to transit centers (Kalkstein et al., 2009). One major disruptor that has had a large negative impact on travel in general has been the COVID-19 pandemic. Parker et al. (2021) note that results of a survey indicated that 75% of participants stated a reduction in transit usage in the United States since the start of the pandemic. A reduction of this magnitude would likely have a substantial impact on any transit agency reliant on revenue from its riders.

A second more constant factor that can impact ridership is the weather, which has been long noted by those overseeing transit agencies (Kalkstein et al., 2009). This is apparent in examples throughout the United States where covered walkways are present in cities like Minneapolis to account for snow, or in cities like Phoenix where cooling stations are not uncommon at bus stops to help with the high summer temperatures (Kalkstein et al., 2009). Vajjarapu et al. (2020) outline that the scientific community at large acknowledges that climate change is impacting weather in many ways, causing more extremes in temperature swings as well as severity of storms. Especially vulnerable to climate change is transportation which includes mass transit systems.

Objectives

In order to build out the research for this project, there were several objectives that would become the guiding posts for the study. The first pertained to identifying a major transit system in the United States in which to study. Once one was selected, publicly available data for this transit agency would need to be acquired, representing a sufficient timespan and volume of data from which analysis could occur. The initial goal was a 10 year period.

As the study involved impact of weather, it was also necessary to acquire data that was publicly available and pertaining to the area where the transit system operates. The time period would need to match that of the transit data. Once that was acquired, the next objective would be to blend the two sets together to form a single dataset from which analysis could occur. Recalling the impact that COVID-19 has had on transit agencies (Parker et al., 2021), data after December 31, 2019 would then be removed to eliminate any influence on the results.

The final objective, which is the goal of this project, was to understand the relationship weather has on the ridership numbers of the transit system. Based on the results, recommendations could be provided to outline next steps, other areas of focus, and other items of note.

Overview of Study

In the 1850s, the city of Chicago was thought of as a young, growing city, uniquely located on Lake Michigan in the American Midwest. By 1892, it had an established electric trolley car network across its territory. Then, in 1897 an elevated structure nicknamed “the L” came to fruition in an effort to offset the severely congested downtown core. Bus service in the city became prevalent in the early decades of the twentieth century. Most of these systems, being privately owned and managed with little oversight or regulation, went bankrupt by the 1940s. As a result, the local government intervened in 1945 to form the Chicago Transit Authority (“the CTA”), first taking ownership of the rail systems followed by the street cars (since defunct), and finally the bus systems by the early 1950s. This created a publicly-owned system of mass transit in the city (Farmer & Noonan, 2014). Its first official date of operation began on October 1, 1947 (Chicago Transit Authority [CTA], 2021).

Today, the agency is among the largest mass transit agencies in the United States both in revenue and ridership, and it has grown over its existence to include not only the city of Chicago, but also the around three dozen bordering municipalities of the Greater Chicago Area (Tang & Thakuriah, 2012), amounting to a total service population of 3.5 million (CTA, 2021). On any given weekday, around 1.6 million riders take the L or a bus on its system, with ridership about evenly split between the two modes. The CTA provides roughly 80% of all the public transit trips in the six-county metro area, with Metra and Pace being the other two transit authorities in the region (CTA, 2021).

The L is made up of eight routes named after colors (Red Line; Blue Line; Brown Line; Green Line; Orange Line; Pink Line; Purple Line; Yellow Line) with a total of 145 rail stations. Chicago is unique in that is it one of the few cities in the world with rail that serves two international airports, with the Blue Line travelling to O’Hare International (since 1984) and the

Orange Line to Midway (since 1993). In addition to the L, the CTA manages 129 bus routes with close to 11,000 individual stops, amounting to around 160,000 miles travelled by buses daily (CTA, 2021).

From an operational viewpoint, the CTA is governed by the Chicago Transit Board made up of seven members, three of which are appointed by the Governor and the remaining members appointed by the Mayor of Chicago. As an organization, there are close to 10,000 employees on its payroll, positions that are funded through and included in its annual operating budget of over $1.5 billion. In addition, the CTA has an annual capital budget of close to $1.3 billion (CTA, 2021).

An Illinois state law requires the CTA to recover a minimum of one half of its operating budget from ridership revenue (CTA, 2021). Such an organization of this size, with its large budgets and number of employees, would likely be impacted by ridership swings, not only from substantial impacts of a pandemic, but perhaps weather-related impacts as well, adding to the constraints of the system.

As the CTA is quite reliant on ridership, it should seek to understand the various factors at play that could potentially negatively impact ridership. This includes attributes like precipitation. Understanding such impacts could better position the CTA with respect to transportation planning, route scheduling, or staffing, especially in times when budgets may be stressed or strapped. Data on daily ridership, when combined with daily precipitation, could provide the CTA this sort of valuable insight.

Research Hypotheses

An organization can often look for ways to improve its bottom line and to maximize profits through its endeavors. This may be focused at a high organization-wide angle or one narrowed and more focused to specific departments to help in the cause. With any of these, there is one valuable tool that, when used properly, can help an organization to optimize the success of its efforts, that of data. Leveraging data and the business intelligence that can be gleaned from it can help to revolutionize an organization in ways that it never even imagined could be possible. Technology has helped to make this possible (Marr, 2015).

Marr (2015) indicates that a critical step in all of this is for the organization to first understand what it is seeking to achieve, by defining its key objectives. It is no surprise, then, that Marr (2015) recommends that strategy itself is the first part of the journey that an organization needs to deeply understand as it sets the foundation for the remaining stepping stones in route to the goalpost. Pairing a strategic objective with the data that is used to validate it is possible through the formulation of business questions that are relevant and well-defined.

An opportunity for analysis of data that an organization has in its hands begins with looking to such business questions. This can be done in the form of a hypothesis which helps direct the focus of the analysis by putting forward a hunch about what the data may show. The nature of a hypothesis helps a researcher create a statement that is clear and concise about what will be tested as part of the analysis (O'Leary, 2017).

There are six research questions (RQs) and hypotheses developed as part of this project.

RQ 1: Does the amount of precipitation impact overall total weekday ridership?

H_0 1: There is no relationship between the amount of precipitation and the overall total weekday ridership

Ha 1: A relationship exists between the amount of precipitation and the overall total weekday ridership

RQ 2: Does the amount of precipitation negatively impact total weekday bus ridership?

H_0 2: There is no negative relationship between the amount of precipitation and the total weekday bus ridership

Ha 2: A negative relationship exists between the amount of precipitation and the total weekday bus ridership

RQ 3: Does the amount of precipitation positively impact total weekday rail ridership?

H_0 3: There is no positive relationship between the amount of precipitation and the total weekday rail ridership

Ha 3: A positive relationship exists between the amount of precipitation and the total weekday rail ridership

RQ 4: Does the amount of snowfall negatively impact overall total weekday ridership?

H_0 4: There is no negative relationship between the amount of snowfall and the overall total weekday ridership

Ha 4: A negative relationship exists between the amount of snowfall and the overall total weekday ridership

RQ 5: Does the amount of snowfall negatively impact total weekday bus ridership?

H_0 5: There is no negative relationship between the amount of snowfall and the total weekday bus ridership

Ha 5: A negative relationship exists between the amount of snowfall and the total weekday bus ridership

RQ 6: Does the amount of snowfall negatively impact total weekday rail ridership?

H_0 6: There is no negative relationship between the amount of snowfall and the total weekday rail ridership

Ha 6: A negative relationship exists between the amount of snowfall and the total weekday rail ridership

With RQs 2 and 3, the thought is that with more precipitation, roads could be more congested delaying buses, but not rail service, so there could potentially be more riders using that mode. With RQs 4 through 6, the thought is that with more snow there could be less ridership on both modes of transportation as it could be harder to get to the rail stations and that there could be fewer buses on the roadways..

Literature Review

A Decision of Behavior

Numerous studies have looked at the relationship between transit ridership and weather (Ngo, 2019). The decision on whether to ride on a mode of transit can be noted as a behavior. Najafabadi et al. (2019) indicate that transit ridership behavior with respect to weather has been studied in the past, noting the study done by Arana et al. (2014) as an example. Recent research has shown that adverse weather such as snow can have a negative impact on transit ridership (Miao et al., 2019). Riders are not wanting to walk or wait in snow or heavy rain for public transit, and as a result, this can reduce ridership (Miao et al., 2019). Li et al. (2018) indicate that past research points to weekday transit ridership as being more concrete and solid during times of increased weather like snow or rain, as opposed to the weekends where behaviors in riders are more determined by leisure activities that include shopping or attending sporting events.

Climate Considerations

Chicago has a climate marked with winters that are cold and summers that are warm. Its location at the midpoint between the Atlantic Ocean and the Continental Divide, paired with it sitting next to Lake Michigan contribute to these factors. Wind off the lake enhance what is referred as lake-effect snow, often leading to frequent snowfall during the winter months. The jet stream aids in the creation of precipitation during the fall, winter, and spring months especially (Angle, n.d.). Research by Tang and Thakuriah (2012) indicate that both precipitation and snow negatively impacts bus ridership in the city of Chicago (Ngo, 2019). Li et al. (2018) add to this point by noting the study by Guo et al. (2007) that looked at transit ridership of the Chicago Transit Authority specifically, which found that rain and snow can make ridership numbers vary.

Studies on transit ridership have been done on regions of different climates. The research by Ngo (2019) looked at impacts of precipitation on bus ridership in the mid-sized metropolitan area of Lane County, Oregon. The Pacific Northwest is noted as having a temperate climate, one that is generally cool and wet for the area west of the Cascade Mountain range, where snowfall is not a common occurrence (United States Fish and Wildlife Service, 2011). This includes Lane County. The overall results of the study (Ngo, 2019) indicate that heavy precipitation in particular has a negative impact on bus ridership.

Analysis conducted by Li et al. (2018) looked at the city of Nanjing, China, and whether precipitation has an impact on transit ridership. In this study, the mode of transportation used was not broken out, rather looking at overall ridership numbers. The results indicate that precipitation has a negative impact on transit ridership, with snow having a greater negative impact than rain. It is worth noting, however, that snow is not a common occurrence in Nanjing (Li et al., 2018), unlike Chicago where it is more prevalent.

Najafabadi et al. (2019) investigated the impact of weather on subway ridership in New York City, another large metropolitan area in the United States, focusing specifically on the island of Manhattan. Though not having the exact climate as that of Chicago, New York does experience rain in the spring months as well cold snowy winters. The results of this research align with past noted research in that increases in rainfall negatively impacts daily ridership.

This past research indicates that despite climates where weather such as rain or snow is common, and thus an assumption might be made that those living in such climates are accustomed to such weather and thus will not be impacted with respect to transit ridership, there is still a negative impact. Arana et al. (2014) appropriately calls out the distinction between climate and weather, whereas the former is often used to determine structural components of a transit system as opposed to the latter which is related to behavior of the riders in their decision to take a mode of transportation.

Selection of the Mode

Countless transit agencies across the globe provide multiple modes for the public to use as a means to get from one place to another. While there are various factors that determine which mode a rider chooses, including location (of the stop and final destination), schedule, and fare, all which are in control of the transit agency, weather in particular is widely considered one of the primary factors that influences which mode to take, whether it is the bus line, rail system, or subway, and yet, this is one factor that the transit agency cannot control (Miao et al., 2019).

Miao et al. (2019) points to the study done by Guo et al. (2007) that found that a subway line is the least negatively impacted from weather due to its structural nature of it being underground and not needing to use roadways, whereas buses were the most susceptible. The research by Singhal et al. (2014) concluded that for rail stations above ground, rain has a negative impact on transit ridership but that rail ridership increases during periods of heavy snowfall (Singhal et al., 2014), a factor that would likely negatively impact bus routes.

Research Design

Methodology

O'Leary (2017) recommends considering using existing data whenever possible, as it will most likely save time when doing research. Should the data itself not be in the desired format, refining it can be better than creating it in the first place. Han et al. (2012) suggest that there may be instances where data is sourced from more than one place, and as a result, they must then be blended together to create a single dataset for analysis. Such is the case with this project, as the weather data was obtained from the National Centers for Environmental Information ("NECI") and the CTA transit ridership data from the Chicago Data Portal. When combined, this creates an applicable dataset from which analysis on a topic like the impact precipitation may have on transit ridership can be conducted.

Source Data

Transit Ridership Data. Data for each ridership interaction is collected by the CTA, with ridership being defined and counted as the individual who boards one of its buses or rail cars. Transactional data captures each unique interaction or experience of a customer with an organization (Han et al., 2012), which in this instance is each boarding. A count occurs when an individual passes through a turnstile to catch the L or when one uses the fare box upon entry onto a bus.

Daily boarding totals, reflected as an aggregation of these transactions, are made publicly available as a dataset on the Chicago Data Portal, going back to 2001. Included in this dataset are the totals for each mode for each date, including weekends and holidays (City of Chicago, 2021). For the purposes of this project, while the entire dataset was downloaded due to a lack of date filters on the data portal, the author excluded dates on and after January 1, 2020 to offset any potential impacts in ridership as result of the COVID-19 pandemic. The remaining data reflects the full years 2001 through 2019 as a result. As the data values specifically reflect daily ridership counts for the CTA, this dataset was found to be aptly appropriate to be used as part of this research project.

Weather Data. The National Oceanic and Atmospheric Administration ("NOAA") is the agency for the United States responsible for understanding and providing critical information related to climate, weather, and the oceans (National Oceanic and Atmospheric Administration, 2021a). As part of its role, weather and climate data is collected at its various weather stations across the nation. According to NOAA (2021b), tens of terabytes of information is gathered from its data sources such as weather stations and satellites on a daily basis. Kimball and Ross (2013) note that this could be classified as what is referred to as Big Data. Marr (2015) adds that sensors are a type of mechanism that collects this Big Data which is apropos of what NOAA calls its Big Data Program (2021b).

Falling under the organizational structure of NOAA is NECI. One of the roles of the NECI is to provide the public access to its information, including weather station data (National Centers for Environmental Information, n.d.-a). NECI provides an online climate data portal for the public to search the various datasets that it manages and makes available. This includes totals for daily precipitation and snowfall.

For the purposes of this project, a dataset search specific to these types of weather attributes was conducted to query a dataset for weather station ID USW00094846 (Chicago O'Hare International Airport) for the time period between January 1, 2001 and ending on December 31, 2019 to match the dates of the CTA dataset. This particular weather station was selected due to its central geographic location in the greater Chicago area. In addition, the weather data is noted as having data going back to 1947 (National Centers for Environmental

Information (n.d.-b). While the dataset being used in the project will not go back that far, the author felt that the long tenure of the weather station is notable.

Creating the Single Dataset

There are countless tools available on the market for one to blend data together. One such tool that was used for this project is Power Query. This is a product by Microsoft that can transform and prepare data. As Microsoft (2020) states, Power Query can help with the extract, transform, and load process, commonly referred to as ETL. Kimball and Ross (2013) note that an ETL system is an important function of business intelligence environments. There are several products and services in which Power Query can be leveraged, including Power BI (Microsoft, 2020).

While Power BI is commonly known as a data visualization tool to build powerful dashboards and other important business intelligence functions (Microsoft, n.d.), Power Query is easily leveraged in this tool. As part of the ETL process, Power Query through the use of Power BI connected to the NECI and CTA datasets and pulled them into Power BI as tables as part of the extraction process.

Data Transformation Steps. For the data transformation, several steps were completed. First, the two tables were merged into a new table (named Combined_Data) through an inner join between the DATE field of the NECI dataset and the service_date field from the CTA dataset. Next, columns that were not needed including STATION (denoting the weather station ID), NAME (of the station), LATITUDE, and LONGITUDE were removed, as all of the rows contained the exact same values in each respective column. To add clarity to the dataset, all of the remaining fields but one were renamed as: Ridership_Date; Ridership_Total_Bus; Ridership_Total_Rail; Ridership_Total_All; Precipitation_Inches; Snowfall_Inches. The column day_type was left untouched (the reason explained in the next paragraph). Next, the two columns with inch measurement data were formatted as decimal numbers going out to two places, for consistency.

For the final step in the data transformation, records under the day_type column were filtered. Specifically, the rows with A to note Saturday and U to indicate either Sunday or holiday (City of Chicago, 2021) were filtered out, leaving the rows with W to indicate weekday (City of Chicago, 2021).

Exporting the Dataset and Understanding Its Components. Once the data transformation was complete, the new table with its dataset of 4,890 rows was exported out to Microsoft Excel for the next phase, that of analysis. Han et al. (2012) suggest that an important first step to any data analysis effort is to inspect the dataset itself, to understand which variables are included and what the data values comprise of overall, including meaning to an organization. This can be in the form of metadata, which is described as referring to data about data (Sharda et al., 2016). Table 1 outlines the metadata for the Combined_Data.xlsx dataset used in this project.

Table 1

Metadata for the Dataset

Variable	Data Type	Formatting	Definition	Source
Ridership_Date	Date	MM/DD/YYYY	Date from which data has been collected	CTA Ridership Dataset
Ridership_Total_Bus	Integer	#000,000	Total bus ridership for date	CTA Ridership Dataset
Ridership_Total_Rail	Integer	#000,000	Total rail ridership for date	CTA Ridership Dataset
Ridership_Total_All	Integer	#000,000	Ridership_Total_Bus + Ridership_Total_Rail	CTA Ridership Dataset
Precipitation_Inches	Decimal	#0.00	Total measured precipitation in inches for date	NECI Dataset
Snowfall_Inches	Decimal	#0.00	Total measured snowfall in inches for date	NECI Dataset

Methods

The collection and understanding of information is an important, almost expected, requirement of the process of data analysis. This is particularly true with respect to statistical analysis of quantitative datasets, where there are many software tools available including SAS, which O'Leary (2017) notes as being the institutional standard. Elliot and Woodward (2015) add that SAS is built around providing its users a full toolkit of packages from data wrangling to statistical analysis.

One such tool, available on the SAS OnDemand for Academics platform, is SAS Studio. As the dataset for this project will be used to better understand the relationship of the variables, SAS Studio is an appropriate tool to perform descriptive statistics and generate histograms to outline basic data understanding, as well as aid in the process of hypothesis testing through means of correlation analysis, the latter of which is the focus of this project, to assess the correlation of the ridership and weather variables.

As each hypothesis focuses on a relationship between two variables to assess whether a dependence exists, then the analysis approach used was that of correlation. The set of data specific to ridership in both modes, bus and rail, have been identified as integers, while the data specific to the precipitation measurements have been identified as decimals. To assess whether a correlation exists between the ridership and weather variables, a bivariate correlation analysis was performed per RQ. As Field (2013) notes, Pearson's correlation coefficient ("Pearson's r") must be used as the variables are continuous in nature. To conduct these statistical tests, SAS Studio, through the SAS OnDemand for Academics platform, was used. O'Leary (2017) notes SAS as being the institutional standard for analysis around statistical tests.

Limitations

The data provided by the two sources used to create the dataset used in the analysis are assumed to be accurate, and should any anomalies exist, these are not readily known to the author. As the data acquired from the Chicago Data Portal is limited to being an aggregation of daily totals at a system-wide level across its modes, no analysis will be done to look at specific rail lines, bus routes, or transit stops. Additionally, the analysis will not take into account any situations around system-related delays or downtime, road or rail construction impacts on routes, natural disasters, or any other event where such occurrence may have directly or indirectly impacted ridership on a specific date or span of dates.

Ethical Considerations

In understanding the challenges around security, privacy, and ethics with respect to the data used in this project, it does not appear that these would apply to the weather data that was acquired from NECI. This is because it does not pertain to any specific individual such as a customer or client, and as such there would not be any privacy concerns. As the data itself pertains to measurements of precipitation, one might take it at face value that the data thus is not sensitive in nature. Because the data was made publicly available from a government website, it is assumed that such acquisition would not violate any security concerns either.

The same security assumption can be made regarding the data pulled from the Chicago Data Portal for the same reason, being that it too was taken from a publicly available government website. From a privacy perspective, the ridership data was presented from the portal in a likely aggregated fashion to indicate overall daily totals of ridership, since no customer-level data was extracted. Thus, personally identifiable information like the name of a rider was not disclosed, and thus ridership privacy was preserved (Al-Hasnawi et al., 2019; Holtrop et al., 2017).

From the angle of addressing any ethical or bias concerns, the author intends to analyze the dataset without any conclusions being made going into the analysis effort. Additionally, the conclusions made on the results of the analysis will be focused from an objective statistically-driven explanation.

Findings

Review of the Dataset

Sharda et al. (2016) indicate that summary statistics are beneficial to analysis. This is because they can provide insight in noting patterns in the data at the variable level as well as for the dataset as a whole by describing important details about the data. As Figure 1 shows, there were a total of 4,890 records, which matches that of the Excel file. In addition, there were no missing values, and as such the dataset was deemed as complete. In reviewing the minimum values, there were none with negative values, and in reviewing the maximum values, there did not appear to be any there were very extreme in nature to indicate outliers that could skew the results. Therefore, the dataset was deemed accurate.

Figure 1

Summary Statistics of the Variables

Variable	Mean	Minimum	Maximum	N	N Miss
Ridership_Total_All	1601856.71	222071.00	2049519.00	4890	0
Ridership_Total_Bus	926264.91	124154.00	1211992.00	4890	0
Ridership_Total_Rail	675591.80	97917.00	1146516.00	4890	0
Precipitation_Inches	0.1091759	0	4.4500000	4890	0
Snowfall_Inches	0.1045194	0	13.6000000	4890	0

Messages User: dwightbrandt0

Tuesday, October 19, 2021

8:17 PM

Another way that can descriptive statistics can be used to gain insight into data is through the use of a visual like a graph or chart (Marr, 2015) Histograms were generated for the four main variables, shown Figure 2 through Figure 5. These were Ridership_Total_Bus; Ridership_Total_Rail; Precipitation_Inches; Snowfall_Inches. The two ridership variables showed a normal distribution of the data. The two weather variables had a large percentage of the data at the 0.0 mark along the X-axis. This is not surprising, since it indicated most days had either no precipitation or no snowfall. Additionally, both histograms showed the percentage value on the Y-axis trending down as the numbers increased along the X-axis. One might consider this a normal pattern for weather data.

Figure 2

Histogram for Ridership_Total_Bus Variable

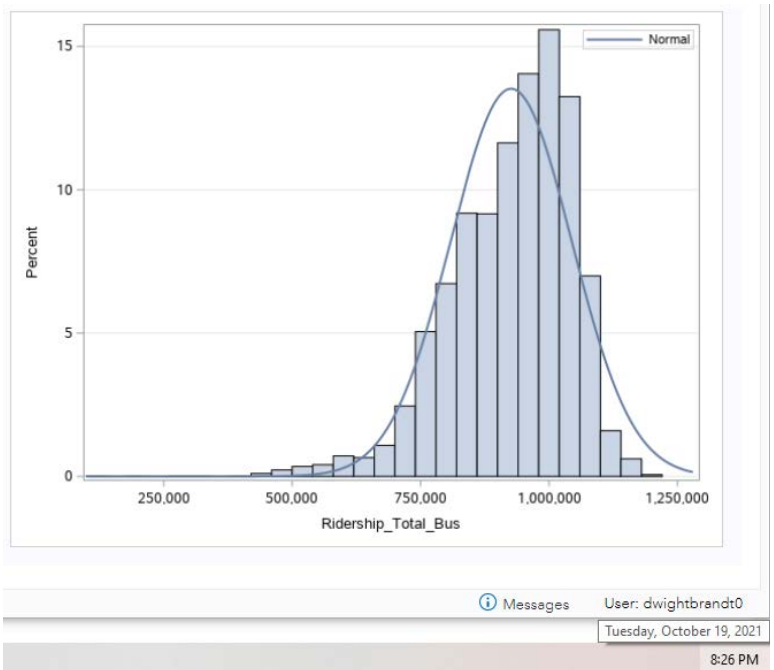


Figure 3

Histogram for Ridership_Total_Rail Variable

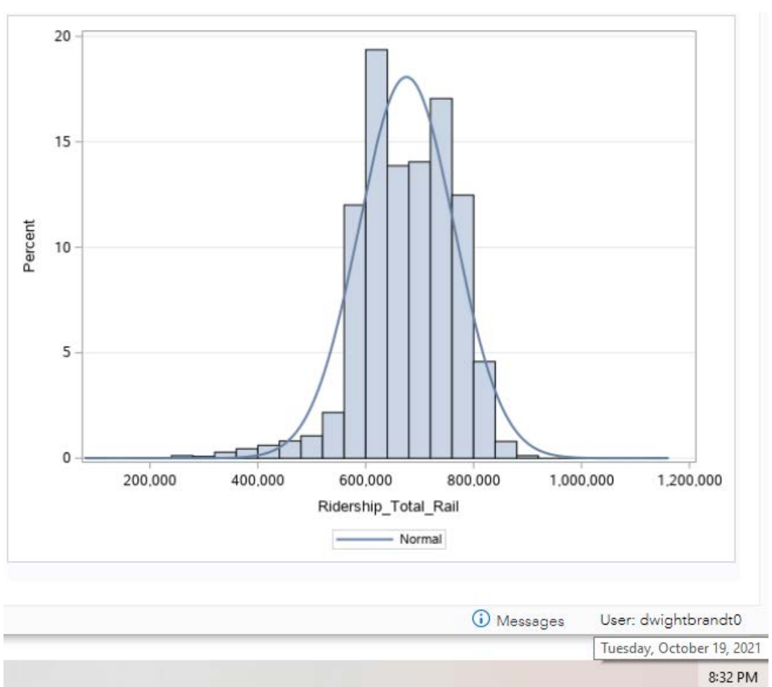


Figure 4

Histogram for Precipitation_Inches Variable

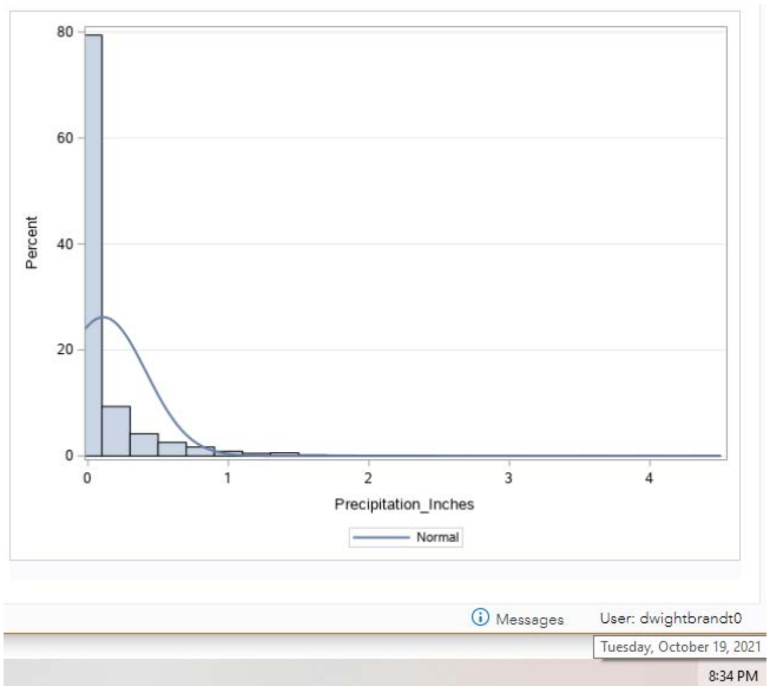
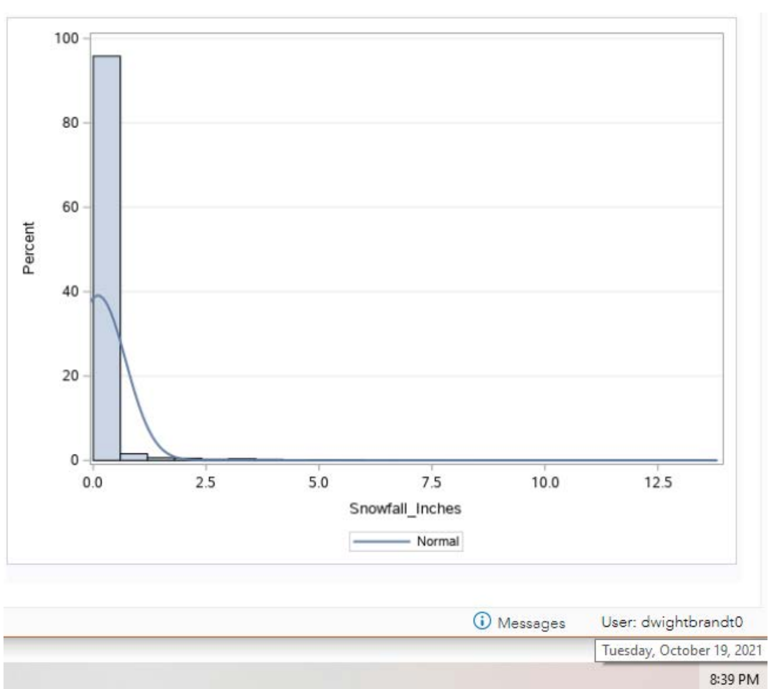


Figure 5

Histogram for Snowfall_Inches Variable



Correlation Analysis of the Dataset

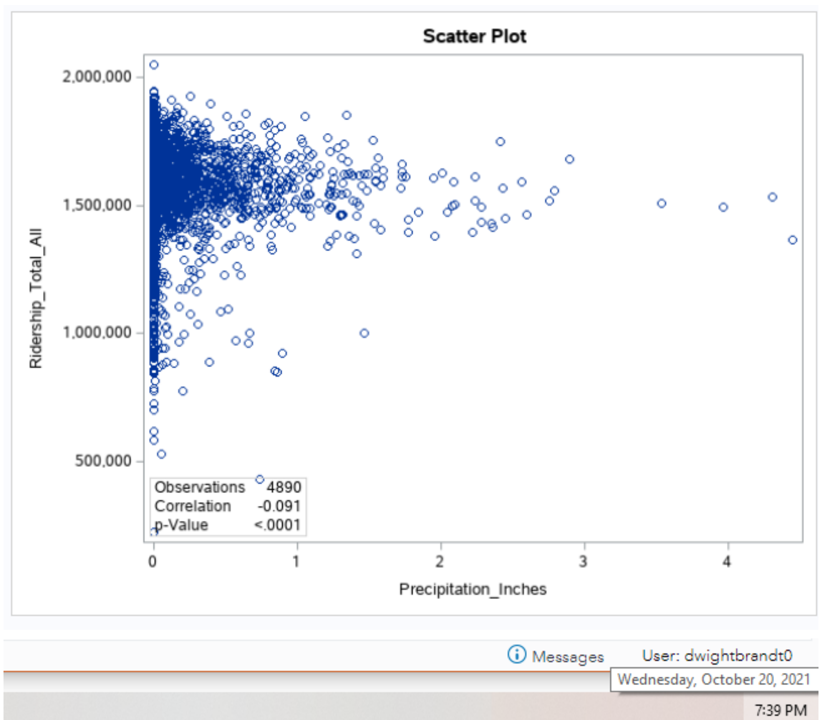
As earlier noted, in order to assess whether a correlation exists between the ridership and weather variables, a bivariate correlation analysis was performed per RQ, specifically a Pearson's r statistical test, as the variables are continuous in nature.

Results of RQ 1

RQ 1 asks whether the amount of precipitation impacts overall weekday ridership. In looking at the results as indicated in Figure 6, the amount of precipitation is significantly related to the overall total weekday ridership, $r = -.09$, $p < 0.001$. These two variables are weakly and negatively correlated. As the amount of precipitation increases, the ridership numbers decrease. Because the *p-value* is less than the significance level of 0.05, the null hypothesis is thus rejected and the alternative hypothesis is accepted.

Figure 6

Scatter Plot of Correlation between Precipitation_Inches and Ridership_Total_All



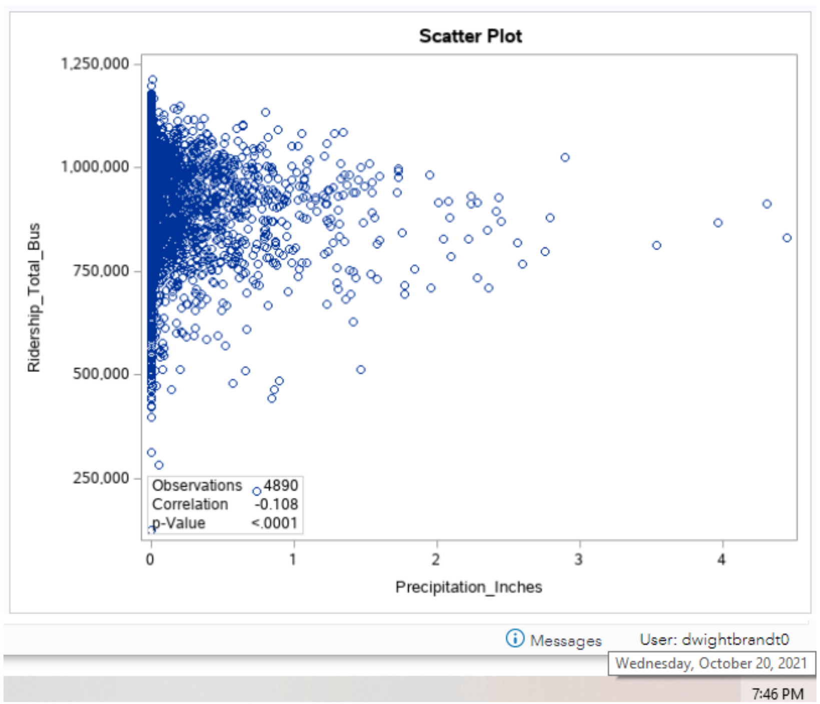
Note. The relationship between the variables is significant, though the correlation is weak.

Results of RQ 2

RQ 2 asks whether the amount of precipitation negatively impacts total weekday bus ridership. In looking at the results as indicated in Figure 7, the amount of precipitation is significantly related to the total weekday bus ridership, $r = -.11$, $p < .0001$. These two variables are weakly and negatively correlated. As the amount of precipitation increases, the bus ridership numbers decrease. Because the *p-value* is less than the significance level of 0.05, the null hypothesis is thus rejected and the alternative hypothesis is accepted.

Figure 7

Scatter Plot of Correlation between Precipitation_Inches and Ridership_Total_Bus



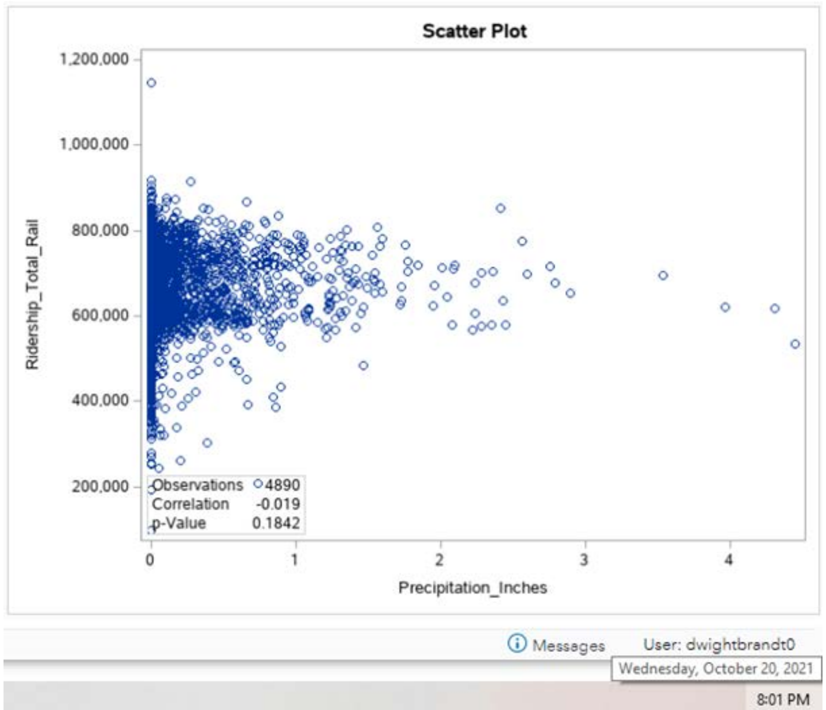
Note. The relationship between the variables is significant, though the correlation is weak.

Results of RQ 3

RQ 3 asks whether the amount of precipitation positively impacts total weekday rail ridership. In looking at the results as indicated in Figure 8, the amount of precipitation is not significant in relation to the total weekday rail ridership, $r = -.02$, $p = .1842$. Because the *p-value* is not less than the significance level of 0.05, the null hypothesis is not rejected and the alternative hypothesis is not accepted.

Figure 8

Scatter Plot of Correlation between Precipitation_Inches and Ridership_Total_Rail



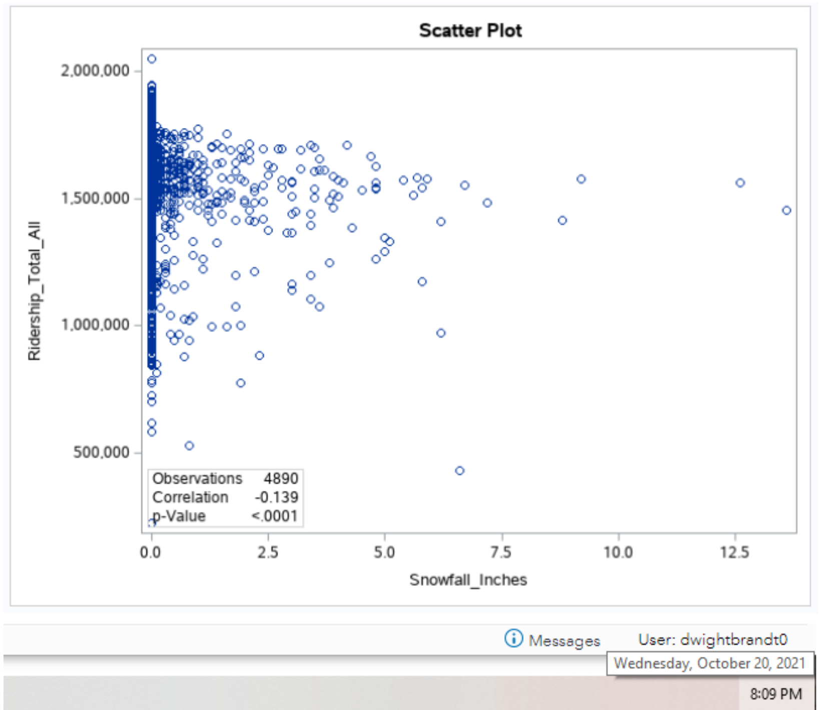
Note. The relationship between the variables is not significant.

Results of RQ 4

RQ 4 asks whether the amount of snowfall negatively impacts total overall weekday ridership. In looking at the results as indicated in Figure 9, the amount of snowfall is significantly related to the total overall weekday ridership, r = -.14, p < .0001. These two variables are weakly and negatively correlated. As the amount of snowfall increases, the overall ridership numbers decrease. Because the *p-value* is less than the significance level of 0.05, the null hypothesis is thus rejected and the alternative hypothesis is accepted.

Figure 9

Scatter Plot of Correlation between Snowfall_Inches and Ridership_Total_All



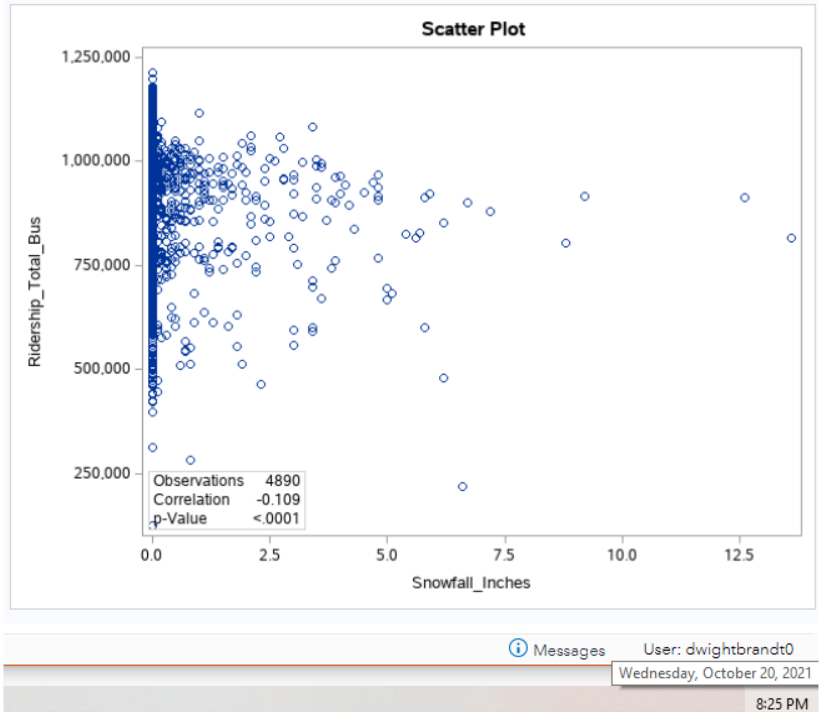
Note. The relationship between the variables is significant, though the correlation is weak.

Results of RQ 5

RQ 5 asks whether the amount of snowfall negatively impacts total weekday bus ridership. In looking at the results as indicated in Figure 10, the amount of snowfall is significantly related to the total weekday bus ridership, r = -0.11, p < .0001. These two variables are weakly and negatively correlated. As the amount of snowfall increases, the bus ridership numbers decrease. Because the *p-value* is less than the significance level of 0.05, the null hypothesis is thus rejected and the alternative hypothesis is accepted.

Figure 10

Scatter Plot of Correlation between Snowfall_Inches and Ridership_Total_Bus



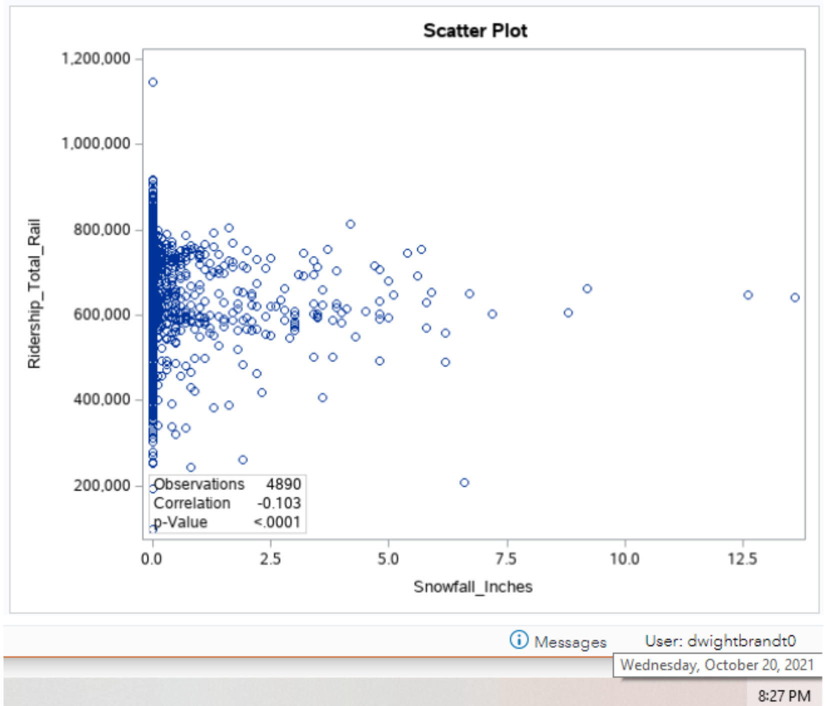
Note. The relationship between the variables is significant, though the correlation is weak.

Results of RQ 6

RQ 6 asks whether the amount of snowfall negatively impacts total weekday rail ridership. In looking at the results as indicated in Figure 11, the amount of snowfall is significantly related to the total weekday bus ridership, $r = -0.1$, $p < .0001$. These two variables are weakly and negatively correlated. As the amount of snowfall increases, the rail ridership numbers decrease. Because the *p-value* is less than the significance level of 0.05, the null hypothesis is thus rejected and the alternative hypothesis is accepted.

Figure 11

Scatter Plot of Correlation between Snowfall_Inches and Ridership_Total_Rail



Note. The relationship between the variables is significant, though the correlation is weak.

Results as a Whole

As Table 2 outlines, of all of the hypotheses tested, only RQ3 is where the null hypothesis was not rejected, as the *p-value* calculated as greater than the significance level of 0.05. On that note, all of the remaining RQs have negative correlation values, albeit weak. In addition, all of the correlation values are close numerically (aside from RQ3). The results suggest that as precipitation increases overall total ridership decreases (RQ1), and the same can be said for total bus ridership (RQ2). However, this cannot be said for rail ridership (RQ3). Regarding snowfall, as it increases, the results indicate that overall total ridership (RQ4), bus (RQ5), and rail (RQ6) ridership decreases. The RQ with the highest correlation, though minimally higher, is

RQ4, total overall ridership. Interestingly, for bus ridership specifically, the correlation was the same for both types of weather, precipitation and snowfall, at -.11.

Table 2

RQ Hypothesis Comparison

RQ	Variables	Relationship Noted in Ha	p-Value	Correlation	H_0 Result
1	Ridership_Total_All Precipitation_Inches	Neutral	<.0001	-.09	Rejected
2	Ridership_Total_Bus Precipitation_Inches	Negative	<.0001	-.11	Rejected
3	Ridership_Total_Rail Precipitation_Inches	Positive	0.1842	-.02	Not Rejected
4	Ridership_Total_All Snowfall_Inches	Negative	<.0001	-.14	Rejected
5	Ridership_Total_Bus Snowfall_Inches	Negative	<.0001	-.11	Rejected
6	Ridership_Total_Rail Snowfall_Inches	Negative	<.0001	-.10	Rejected

Conclusion

The results of this study indicate that overall, weather has a relationship to transit ridership on the CTA, more specifically that as precipitation and snowfall increases, transit ridership on this system decreases. This overarching assessment aligns with past studies that have researched the impact of weather on other mass transit systems across the globe and have generally come to the same conclusion.

Transportation has played an important role for humanity, and mass transit in particular has been a valuable asset to many communities including Chicago. The CTA is mandated by state law to recover a minimum of one half of its operating budget from ridership revenues, and thus, understanding impacts to ridership is a necessary endeavor especially as the world gets out of and recovers from the immeasurable impacts of the COVID-19 pandemic. Additionally as climate change continues to influence weather, knowledge and insight of how this force plays into its influence on transit ridership will only become a more valuable tool to have.

Recommendations

The weak correlation between weather and transit ridership suggests that other factors could be at play that may be influencing ridership numbers. As Miao et al. (2019) points out, there are internal factors such as the timeliness of service and fare pricing, as well other external factors in addition to weather like socioeconomic and demographic characteristics that can influence public transit ridership. While there can be absolute value in data to an organization for opportunities like innovation as well as social benefit from the analysis and subsequent reporting of insight, it remains nonetheless important though to value the privacy of information like that of a rider. Thus, the organization should allow for the mitigation of risk around collecting and protecting data (Davis, 2012), including personally identifiable information like that of a rider's name.

Another consideration may be to dig deeper by conducting analysis at specific stops or on specific rail lines, to better understand whether weather and these additional factors could have a more granular impact. While this study only analyzed weekday ridership data, there may be an opportunity to analyze the entire dataset to understand how weekends, holidays, or specific days of the week have any insight that can be provided.

References

Al-Hasnawi, A., Carr, S. M., & Gupta, A. (2019). Fog-based local and remote policy enforcement for preserving data privacy in the Internet of Things. *Internet of Things*, 7. https://doi.org/10.1016/j.iot.2019.100069

Angle, J. (n.d.). *Climate of Chicago – Description and normals.* Retrieved October 12, 2021, from https://www.isws.illinois.edu/statecli/general/chicago-climate-narrative.htm

Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. *Transportation Research Part A: Policy and Practice*, 59(1), 1–12. https://doi.org/10.1016/j.tra.2013.10.019

Chicago Transit Authority. (2021). *About us.* CTA. Retrieved September 15, 2021, from https://www.transitchicago.com/about/

City of Chicago. (2021). *CTA – Ridership – Daily Boarding Totals.* Chicago Data Portal. https://data.cityofchicago.org/Transportation/CTA-Ridership-Daily-Boarding-Totals/6iiy-9s97

Davis, K. (2012). *Ethics of big data*. O'Reilly Media, Inc.

Elliot, A.C., & Woodward, W.A. (2015) SAS essentials: Mastering SAS for analytics (2nd ed.). John Wiley and Sons.

Farmer, S., & Noonan, S. (2014). The contradictions of capital and mass transit: Chicago, USA. *Science & Society*, 78(1), 61-87. https://doi.org/10.1521/siso.2014.78.1.61

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage Publishing.

Guo, Z., Wilson, N. H. M., & Rahbee, A. (2007). Impact of weather on transit ridership in Chicago, Illinois. *Transportation Research Record*, 2034(1), 3–10. https://10.3141/2034-01

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques* (3rd ed.). Morgan Kaufmann.

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing*, 34(1), 154–172. https://doi.org/10.1016/j.ijresmar.2016.06.001

Kalkstein, A. J., Kuby, M., Gerrity, D., & Clancy, J. J. (2009). An analysis of air mass effects on rail ridership in three US cities. *Journal of Transport Geography*, 17(3), 198–207. https://doi.org/10.1016/j.jtrangeo.2008.07.003

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The complete guide to dimensional modeling* (3rd ed.). Wiley.

Li, J., Li, X., Chen, D., & Godding, L. (2018). Assessment of metro ridership fluctuation caused by weather conditions in Asian context: Using archived weather and ridership data in Nanjing. *Journal of Transport Geography*. 66. 356–368. https://doi.org/10.1016/j.jtrangeo.2017.10.023

Marr, B. (2015). *Big data: Using SMART big data, analytics, and metrics to make better decisions and improve performance*. John Wiley and Co.

Merlin, L. A., Singer, M., & Levine, J. (2021). Influences on transit ridership and transit accessibility in US urban areas. *Transportation Research Part A: Policy & Practice*, 150, 63–73. https://doi.org/10.1016/j.tra.2021.04.014

Miao, Q., Welch, E. W., & Sriraj, P. S. (2019). Extreme weather, public transport ridership and moderating effect of bus stop shelters. *Journal of Transport Geography*, 74, 125–133. https://doi.org/10.1016/j.jtrangeo.2018.11.007

Microsoft. (n.d.). *What is Power BI?* Retrieved September 22, 2021, from https://powerbi.microsoft.com/en-us/what-is-power-bi/

Microsoft. (2020, July 27). *What is Power Query?* https://docs.microsoft.com/en-us/power-query/power-query-what-is-power-query

Najafabadi, S., Hamidi, A., Allahviranloo, M., & Devineni, N. (2019). Does demand for subway ridership in Manhattan depend on the rainfall events? *Transport Policy*, 74, 201–213. https://doi.org/10.1016/j.tranpol.2018.11.019

National Centers for Environmental Information (n.d.-a). *About us*. NCEI. Retrieved September 16, 2021, from https://www.ngdc.noaa.gov/ngdcinfo/aboutngdc.html

National Centers for Environmental Information. (n.d.-b). *Climate Data Online Search.* NCEI. Retrieved September 16, 2021, from https://www.ncdc.noaa.gov/cdo-web/search

National Oceanic and Atmospheric Administration (2021-a, May 25). *About our agency*. NOAA. https://www.noaa.gov/about-our-agency

National Oceanic and Atmospheric Administration (2021-b, June 10). *Big Data program.* NOAA. https://www.noaa.gov/information-technology/big-data

Ngo, N. S. (2019). Urban bus ridership, income, and extreme weather events. *Transportation Research Part D*, 77, 464–475. https://doi.org/10.1016/j.trd.2019.03.009

O'Leary, Z. (2017). *The essential guide to doing your research project.* (3rd ed.). Sage Publishing.

Parker, M. E. G., Li, M., Bouzaghrane, M. A., Obeid, H., Hayes, D., Frick, K. T., Rodríguez, D. A., Sengupta, R., Walker, J., & Chatman, D. G. (2021). Public transit use in the United States in the era of COVID-19: Transit riders' travel behavior in the COVID-19 impact and recovery period. *Transport Policy*, 111, 53–62. https://doi.org/10.1016/j.tranpol.2021.07.005

Sharda, R., Delen, D., & Turban, E. (2016). *Business intelligence, analytics, and data science: A managerial approach*. Pearson.

Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice*, 69, 379–391. https://doi.org/10.1016/j.tra.2014.09.008

Tang, L., & Thakuriah, P. V. (2012). Ridership effects of real-time bus information system: A case study in the City of Chicago. *Transportation Research Part C*, 22, 146–161. https://doi.org/10.1016/j.trc.2012.01.001

United States Fish and Wildlife Service. (2011, October 19). *Climate change in the Pacific Northwest.* https://www.fws.gov/pacific/climatechange/changepnw.html

Vajjarapu, H., Verma, A., & Allirani, H. (2020). Evaluating climate change adaptation policies for urban transportation in India. *International Journal of Disaster Risk Reduction*, 47. https://doi.org/10.1016/j.ijdrr.2020.101528