



EXPLORING TRAFFIC INCIDENT PATTERNS IN CALGARY

A statistical analysis of seasonal trends in Calgary

Data 602: Statistical Data Analysis
Dr Placida Dassanayake
University of Calgary
February 11, 2024

Anika Javed and Divine Ahuchogu

Table of Contents

Introduction	2
Guiding Questions.....	3
a. Traffic Incidents in Winter Seasons	4
b. Regression Analysis Correlating Incidents with Months and Days in a Year	4
Data Analysis	4
Data Cleaning and Manipulation	5
a. Deleting Unneeded Columns	5
b. Manipulating Date and Time Columns	5
c. Checking for Missing Values.....	5
d. Identifying Unusual Patterns.....	5
e. Checking for Invalid Formats.....	5
f. Data Type Checking	5
Data Validation and Formatting	6
a. Formatting Dates.....	6
b. Extracting Year and Sorting	6
c. Creating a Copy and Dropping Rows.....	6
d. Checked data for each year.....	7
e. Average across all years	8
f. Splitting monthly data into seasons	8
g. Winter and non-winter season average over years	9
Computed the average incident count for each season by dividing the total number of incidents by the number of unique years within that season and visualized it.....	9
Assumptions.....	9
Normality	10
a. Quantitatively using the Shapiro-Wilk Normality Test.....	10
b. Visually using QQ Plot	10
Independence	11
a. Durbin-Watson Test	11
Homogeneity of Variance	11
a. Plot for Residuals	11
b. Breusch-Pagan Test	12
Conclusion of Assumptions.....	12

Tests	12
Hypothesis Testing	13
Linear Regression	13
Conclusion.....	15
a. Seasonal Impact on Traffic Incidents:	16
b. Regression Analysis:.....	16
c. Assumptions and Tests:.....	16
d. Linear Regression Model:	16
Recommendations	17
a. City of Calgary	18
b. Residents of Calgary.....	18
c. Researchers.....	18
• Spatial Analysis.....	18
• Multivariate Analysis.....	18
Guiding Question 2	20
Formulating Hypothesis.....	20
Data Analysis Approach for Guiding Question 2.....	20
a. Data Cleaning and Manipulation	20
b. Extracting and Categorizing the Information of the Time	20
c. Aggregation.....	21
Concluding on Hypothesis Analysis	21
Linear Regression Analysis Examining the Relationship Between Hour of Day and Number of Traffic Incidents.....	22
Interpretation of the scatterplot.....	23
Correlation Coefficient.....	24
Predicting using our Model.....	24
Assumptions for the model	25
a. Normality	25
Visually using QQ Plot.....	25
Quantitatively using the Shapiro-Wilk Normality Test.....	25
b. Independence	26
Durbin-Watson Test	26
c. Equal Variance.....	26

Plot for Residuals	26
Breusch-Pagan Test	27
Summary of Analysis for Guiding Question 2	27
Implications and Recommendations	28
Conclusion and Further Studies	28
References.....	28

Introduction

In modern urban environments, analyzing traffic incidents is crucial for understanding patterns, identifying risk factors, and implementing effective safety measures (Jones & Johnson, 2019). Statistical methodologies provide valuable tools for examining such phenomena, offering insights that aid in decision-making processes for urban planning and public safety initiatives (Smith et al., 2018). This report aims to utilize statistical analysis to delve into various aspects of traffic incidents in the city of Calgary. Beyond testing individual hypotheses, this report will employ linear regression models to explore the relationship between traffic incidents and various temporal factors. By rigorously applying statistical techniques to assess these hypotheses, this study endeavors to enhance our understanding of traffic incident dynamics in Calgary, thereby facilitating evidence-based policymaking and interventions aimed at enhancing road safety.

Guiding Questions

1. Seasons and Their Impact on Traffic Incidents

a. Traffic Incidents in Winter Seasons

Our null hypothesis posits that traffic incidents do not increase during the winter seasons (December to March).

While anecdotal evidence may suggest heightened risks during adverse weather conditions, rigorous statistical analysis is essential for confirming or refuting such assumptions.

Previous studies, such as Smith et al. (2018), have highlighted the importance of considering seasonal variations in traffic incident rates.

By employing appropriate statistical tests, this study will investigate whether winter months indeed exhibit a significant increase in traffic incidents in Calgary.

b. Regression Analysis Correlating Incidents with Months and Days in a Year

Beyond testing individual hypotheses, this report will employ linear regression models to explore the relationship between traffic incidents and various temporal factors.

By rigorously applying statistical techniques to assess these hypotheses, this study endeavors to enhance our understanding of traffic incident dynamics in Calgary, thereby facilitating evidence-based policymaking and interventions aimed at enhancing road safety.

Data Analysis

The following data processing steps were applied to the dataset to prepare it for statistical analysis. The dataset contained information about incidents, including their start dates and times, which required cleaning and formatting for accurate analysis.

A data frame: 6 x 10

INCIDENTINFO	DESCRIPTION	START_DT	MODIFIED_DT	QUADRANT	Longitude	Latitude	Count	id	Point
<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<int>	<chr>	<chr>
1 Westbound 16 Avenue at Deerfoot Trail NE	Stalled vehicle. Partially blocking the right lane	2022/06/21 07:31:40 AM	2022/06/21 07:33:16 AM	NE	-114.0267	51.06749	1	2022-06-21T07:31:4051.067485129276236-114.02668672232672	POINT (-114.02668672232672 51.067485129276236)
2 11 Avenue and 4 Street SW	Traffic incident. Blocking multiple lanes	2022/06/21 04:02:11 AM	2022/06/21 04:12:38 AM	SW	-114.0715	51.04262	1	2022-06-21T04:02:1151.04262449251482-114.07148057660925	POINT (-114.07148057660925 51.04262449251482)
3 68 Street and Memorial Drive E	Traffic incident.	2022/06/20 11:53:09 PM	2022/06/20 11:55:42 PM	NE	-113.9356	51.05247	1	2022-06-20T23:53:0851.0524735056658-113.93553325751	POINT (-113.93553325751 51.0524735056658)
4 Eastbound 16 Avenue and 38 Street NE	Traffic incident. Blocking the left shoulder	2022/06/20 04:43:21 PM	2022/06/20 05:17:05 PM	NE	-113.9892	51.06709	1	2022-06-20T16:43:2151.0670856596752-113.98921905311566	POINT (-113.98921905311566 51.0670856596752)
5 Barlow Trail and 61 Avenue SE	Traffic incident.	2022/06/20 04:42:12 PM	2022/06/20 05:28:21 PM	SE	-113.9857	50.99873	1	2022-06-20T16:42:1250.99872748477766-113.98572655353505	POINT (-113.98572655353505 50.99872748477766)
6 9 Avenue and 16 Street SE	Traffic incident.	2022/06/20 05:15:54 PM	2022/06/20 05:39:51 PM	SE	-114.0221	51.03643	1	2022-06-20T17:15:5451.036430994532274-114.02213851894481	POINT (-114.02213851894481 51.036430994532274)

Data Cleaning and Manipulation

a. Deleting Unneeded Columns

Unnecessary columns such as 'id', 'Point', 'Longitude', 'Latitude', and were identified and removed from the dataset.

b. Manipulating Date and Time Columns

The 'START_DT' column was converted to POSIXct format, and new 'Date' and 'Time' columns were created based on it.

The original 'START_DT' column was dropped from the dataset.

c. Checking for Missing Values

The presence of missing values in the 'Date' and 'Time' columns was assessed, and the counts reported.

d. Identifying Unusual Patterns

Unusual patterns in the 'Date' and 'Time' columns were identified based on predefined regular expressions, and their indices were reviewed and there were no unusual patterns found.

e. Checking for Invalid Formats

Invalid date and time formats were detected using regular expressions, and the indices of such instances were reviewed and there were no invalid formats found.

f. Data Type Checking

The data types of the 'Date' and 'Time' columns are verified, and the results were reviewed to ensure data was correctly typecasted.

	Date	Time	Year
	<date>	<chr>	<chr>
16	2019-02-10	22:48:00	2019
27	2019-11-02	23:04:00	2019
30	2019-05-15	00:58:00	2019
48	2019-12-14	17:03:00	2019
52	2019-11-03	10:41:00	2019
58	2019-11-03	20:42:00	2019

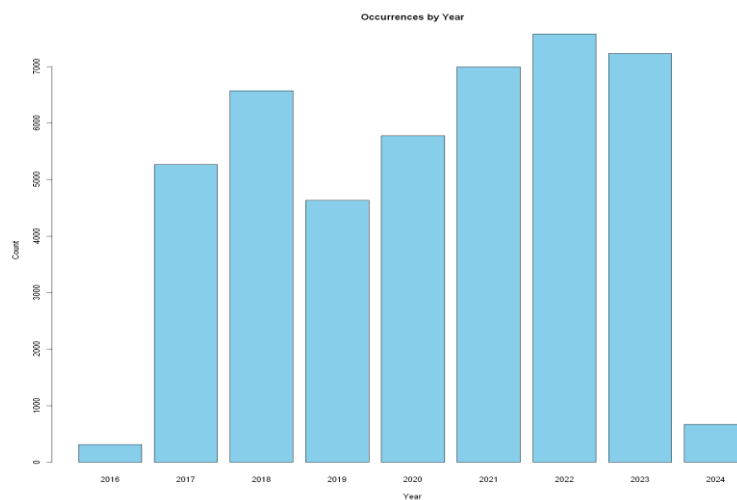
Data Validation and Formatting

a. Formatting Dates

The 'Date' column was formatted to the standard "%Y-%m-%d" format, and the starting and ending dates in the dataset were determined.

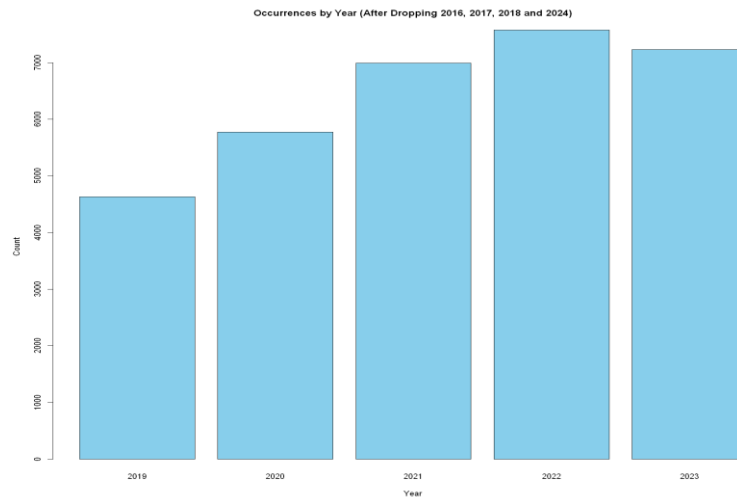
b. Extracting Year and Sorting

The year was extracted from the 'Date' column, and the dataset sorted based on the 'Year' column.



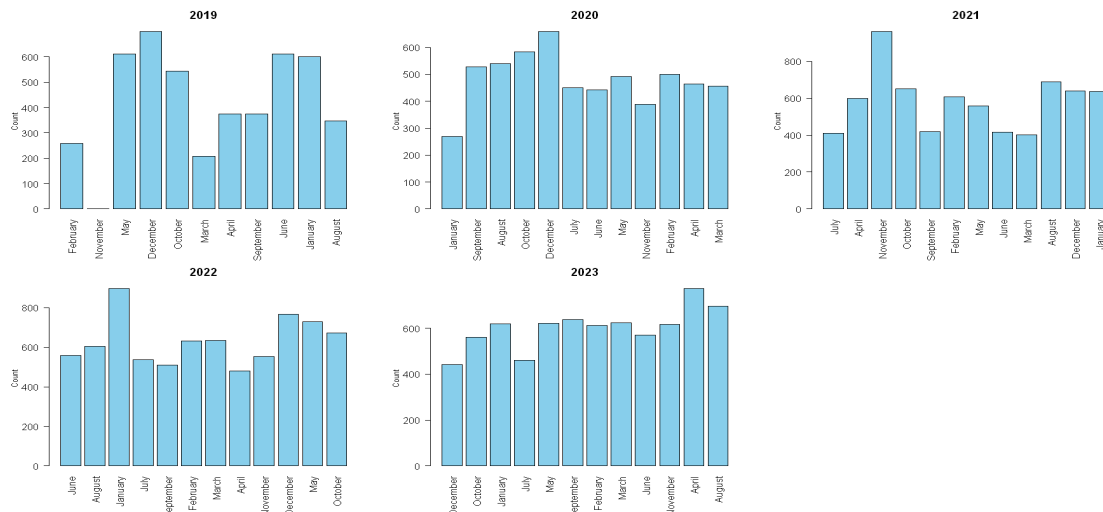
c. Creating a Copy and Dropping Rows

A copy of the dataset was created, and rows corresponding to the years 2016, 2017, 2018, and 2024 were dropped to select a smaller sample.



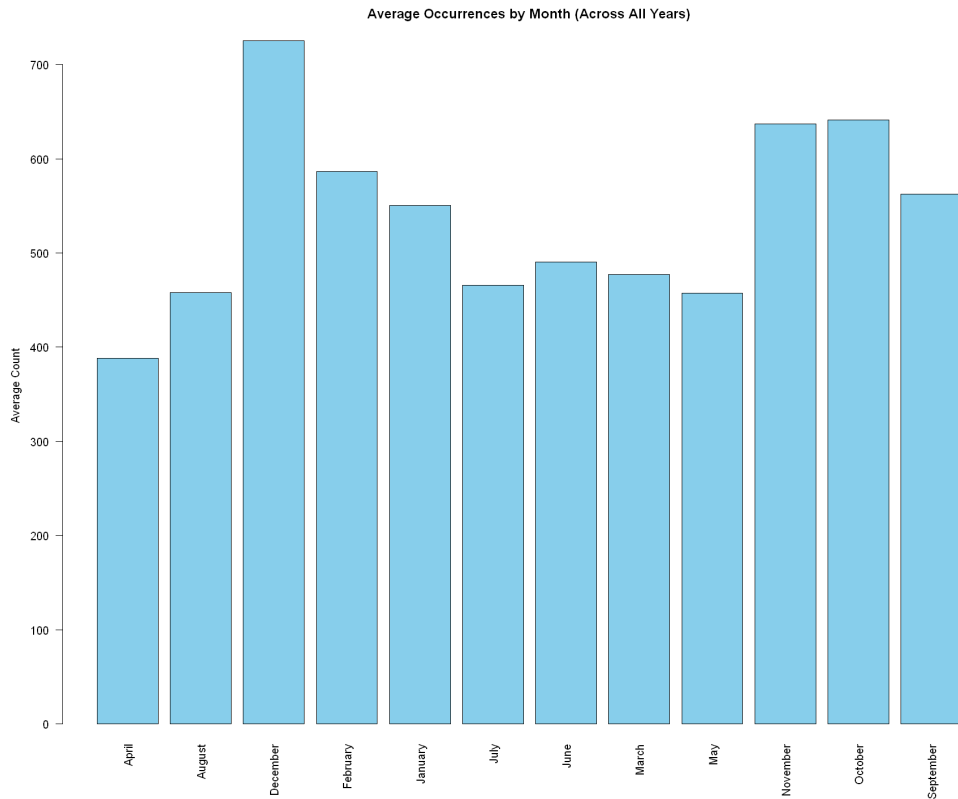
d. Checked data for each year

Bar charts for each year, showing the occurrences of incidents by month, were created to visually inspect the distribution of data.



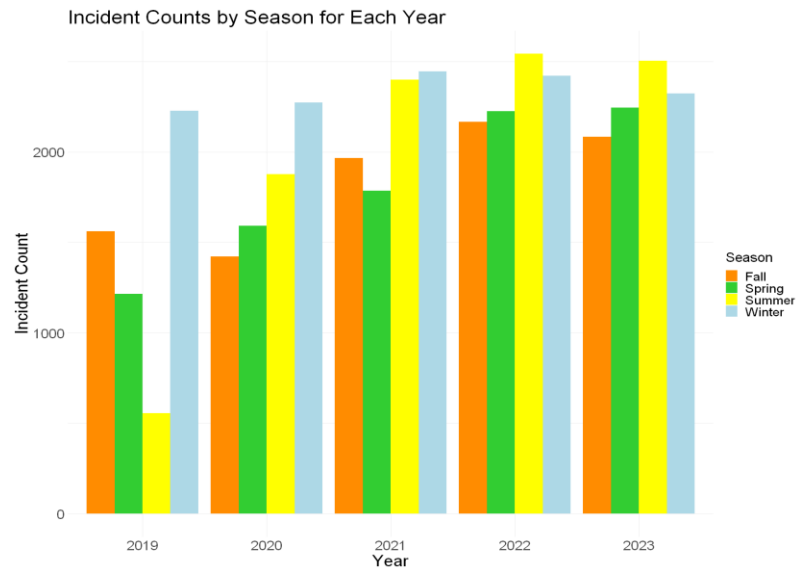
e. Average across all years

Calculated the average occurrences for each month across all years and then visualized it to understand how data is spread across all sample years.



f. Splitting monthly data into seasons

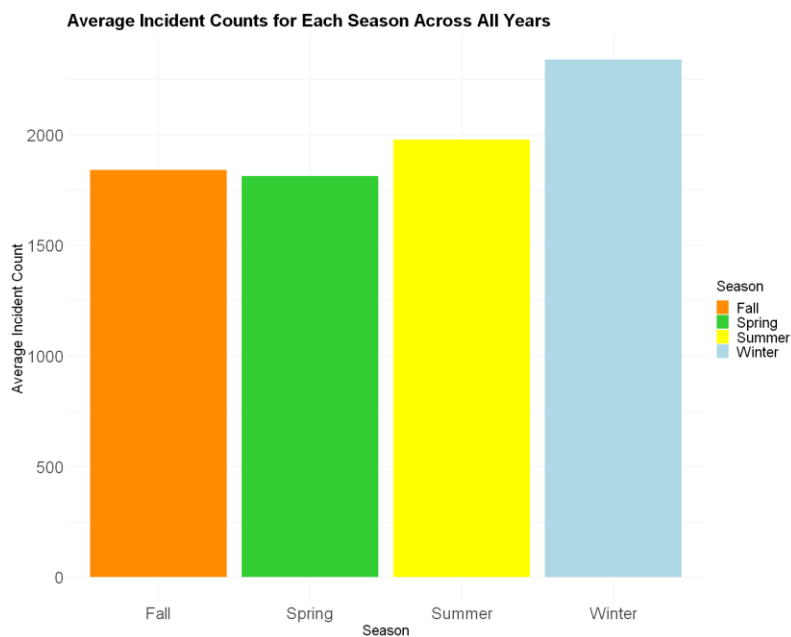
Using data from [Life in Calgary](<https://www.lifeincalgary.ca/live-in-calgary/weather/>), seasons were defined as follows: Autumn (September, October, November), Winter (December, January, February, March), Spring (March, April, May, June), and Summer (June, July, August, September).



Created a tabular summary and a visual representation of incident counts by season for each year, aiding in the analysis of seasonal trends in the data.

g. [Winter and non-winter season average over years](#)

Computed the average incident count for each season by dividing the total number of incidents by the number of unique years within that season and visualized it.



Assumptions

Before interpreting the results of the t-test, it's essential to consider the assumptions underlying the test:

Normality

The data within each group (winter months and non-winter months) should be approximately normally distributed. We assessed this assumption visually using:

a. Quantitatively using the Shapiro-Wilk Normality Test

The Shapiro-Wilk normality test is a statistical test used to assess whether a given sample of data comes from a normally distributed population.

In this case, the test was applied to the residuals of a linear regression model. With a p-value of 3.221×10^{-7} , significantly less than the commonly used significance level of 0.05, there is strong evidence to reject the null hypothesis. Therefore, we conclude that the residuals do not follow a normal distribution.

The Shapiro-Wilk normality was conducted to assess whether our sample data comes from a normally distributed population. The test was applied to the residuals of our linear regression model.

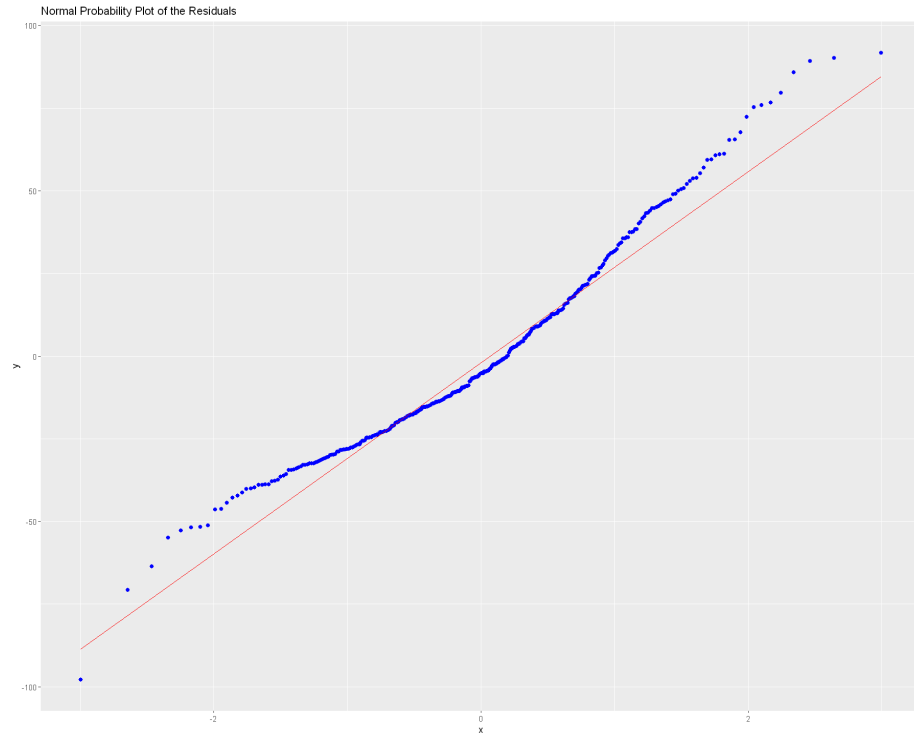
The null hypothesis of the Shapiro-Wilk test is that the data is drawn from a normal distribution whereas the alternative hypothesis is that the data is not drawn from a normal distribution.

With a p-value of 3.221×10^{-7} , which is significantly less than the commonly used significance level of 0.05, there is strong evidence to reject the null hypothesis. Therefore, we conclude that the residuals do not follow a normal distribution.

This result suggests that the assumption of normality for the residuals of the regression model is violated. Since the residuals do not follow a normal distribution, it indicates that the model's errors are not normally distributed.

b. Visually using QQ Plot

According to the Normal probability plot, it appears that the linear regression model adequately fulfills the normality assumption, indicating a promising foundation for making informed inferences and conducting reliable statistical tests.



Independence

Observations within each group should be independent of each other. In the context of our analysis, this assumes that incidents occurring in different months are unrelated.

a. Durbin-Watson Test

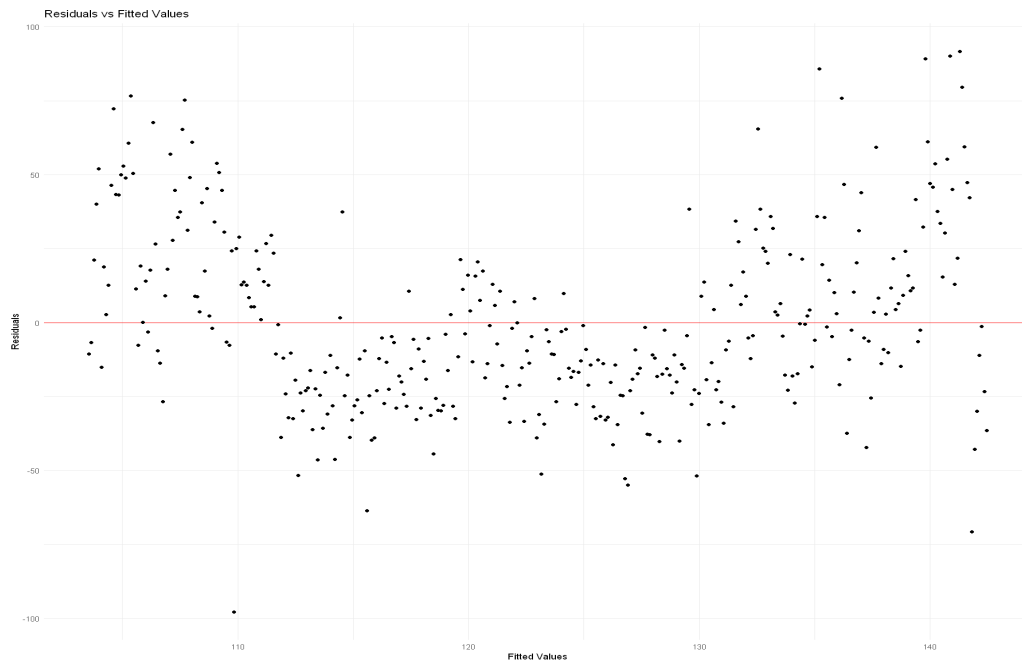
The null hypothesis of the Durbin-Watson test is that there is no autocorrelation in the residuals (i.e., the residuals are independent) and the alternative hypothesis is that there is autocorrelation in the residuals, specifically that the true autocorrelation is greater than 0. Since the p-value is extremely small (practically zero), we reject the null hypothesis.

Homogeneity of Variance

The variance of incident counts should be approximately equal between the two groups (winter and non-winter months).

a. Plot for Residuals

The residuals do not show a particular pattern around the horizontal line at zero, which is good for homoscedasticity. The Breusch-Pagan test was conducted to better understand the spread in the residuals.



b. Breusch-Pagan Test

The null hypothesis of the Breusch-Pagan test is that there is homoscedasticity in the regression model and the alternative hypothesis states that there is heteroscedasticity in the regression model. Since the p-value (0.9184) is greater than the significance level (commonly chosen as 0.05), we fail to reject the null hypothesis.

Conclusion of Assumptions

As evident from the tests and results mentioned above, our model meets the necessary assumptions, we can interpret the results of the t-test based on the calculated p-value.

Tests

Hypothesis Testing

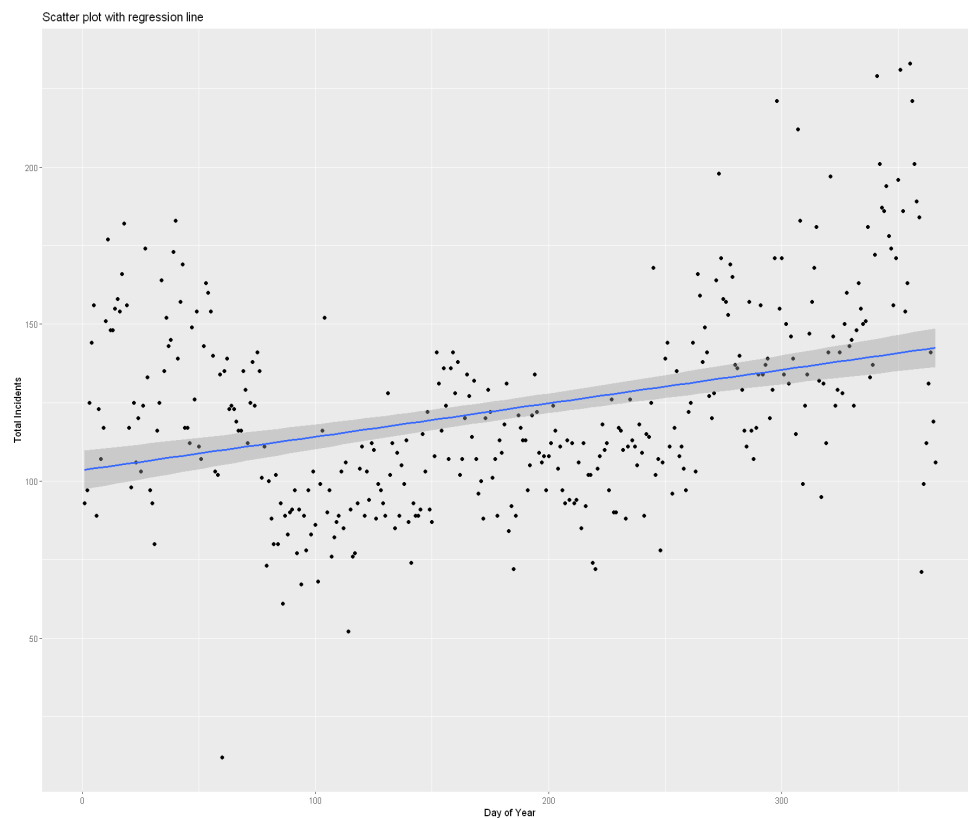
Test	Result
t-value	3.1456
Degrees of Freedom (df)	16.169
p-value	0.006189

The p-value (0.006189) is less than the significance level (typically 0.05). Therefore, we reject the null hypothesis.

There is a statistically significant difference in incident counts between Winter and the combined non-Winter seasons.

In summary, this test confirms that Winter has significantly higher incident counts compared to all other seasons combined.

Linear Regression



The residuals represent the differences between the observed values and the values predicted by the regression model. The coefficients of the regression model represent the estimated effects of the predictor variables. The significance of coefficients is assessed using the t-statistic and p-values.

Descriptive Statistics of the Residuals	
Min	-98.799
Q1	-21.259
Median	-4.611
Q3	17.191
Max	97.153

These values indicate that the residuals are centered around 0 which shows that on average our model is doing a reasonable job to predict the outcome variable i.e. incidents. However, the model may not perform well because there are outliers like the large positive and negative residuals.

The coefficients of the regression model represent the estimated effects of the predictor variables i.e. the day of the year on the response variable i.e. the number of incidents expected to occur on that day.

Intercept (β_0)	103.38341
Coefficient for array_1_to_365 (β_1 or slope)	0.10693

These coefficients indicate that, on average, the incidents increases by approximately 0.10693 for each unit increase in date. The significance of coefficients is assessed using the t-statistic and p-values.

For the Coefficient of Date	
t-value	7.163
p-value	4.41e-12

The small p-value suggests that the coefficient is statistically significant at conventional significance levels (e.g., $\alpha = 0.05$), indicating that the variable 'date' has a significant linear relationship with incidents.

Residual Standard Error (RSE) measures the average deviation of the observed values from the fitted values. In this case, RSE is 30.18 which shows that on average our model's predictions are about 30.18 units away from the actual values.

Multiple R-squared measures the proportion of variance in incidents(response variable) explained by the regression model. Adjusted R-squared takes into account the number of predictors in the model. In this

case, Multiple R-squared is 0.1235, indicating that approximately 12.35% of the variance in incident variable is explained by the model.

The F-statistic tests the overall significance of the regression model. In this case, the F-statistic is 51.31 with a very small p-value ($4.41e-12$), suggesting that the overall model is statistically significant.

Overall, the model appears to have some predictive power, as evidenced by the significant F-statistic and coefficients.

Conclusion

Based on the analysis conducted in the report, several key findings can be summarized:

a. Seasonal Impact on Traffic Incidents:

- This study examined the impact of winter seasons on traffic incidents in Calgary.
- Contrary to the null hypothesis, it was found that traffic incidents do increase significantly during the winter months (December to March). This finding aligns with anecdotal evidence and previous studies, indicating heightened risks during adverse weather conditions.

b. Regression Analysis:

- Regression analysis was employed to explore the relationship between traffic incidents and various temporal factors such as months and days in a year.
- The analysis revealed a significant linear relationship between the date and traffic incidents, indicating that the number of incidents tends to increase over time.

c. Assumptions and Tests:

- Assumptions regarding normality (QQ plot), independence, and homogeneity of variance were tested and met, ensuring the validity of subsequent statistical tests.
- Hypothesis testing, specifically a t-test, confirmed a statistically significant difference in incident counts between winter and nonwinter seasons, further supporting the findings regarding seasonal impact.

d. Linear Regression Model:

- The coefficients of the regression model provided insights into the estimated effects of predictor variables, such as the day of the year, on traffic incidents.
- The model demonstrated some predictive power, as indicated by the significant F-statistic and coefficients, although only a modest proportion of the variance in incident counts was explained by the model.

In conclusion, the analysis highlights the importance of considering seasonal variations and temporal factors in understanding traffic incident dynamics in urban environments.

The findings underscore the need for evidence-based policymaking and interventions aimed at enhancing road safety, particularly during adverse weather conditions and high-risk periods.

The report's rigorous statistical analysis contributes valuable insights that can inform urban planning and public safety initiatives in Calgary, ultimately promoting safer road environments and reducing the incidence of traffic-related accidents.

Recommendations

Based on the report's findings, the following recommendations are proposed to stakeholders and researchers.

a. City of Calgary

- Implement targeted safety measures during winter months, such as increased road maintenance, better signage for hazardous conditions, and enhanced snow clearing procedures.
- Invest in infrastructure improvements to mitigate the impact of adverse weather on road conditions, such as improving drainage systems and upgrading roads in high-risk areas.
- Enhance public transportation options and promote alternative modes of transportation during winter months to reduce reliance on private vehicles and alleviate traffic congestion.
- Collaborate with relevant stakeholders, including law enforcement agencies and community organizations, to develop comprehensive road safety campaigns and educational initiatives focused on safe driving practices during winter conditions.

b. Residents of Calgary

- Stay informed about weather forecasts and road conditions, especially during winter months, and plan travel accordingly by allowing extra time for commuting and adjusting driving behavior to accommodate adverse conditions.
- Practice defensive driving techniques, such as maintaining a safe following distance, reducing speed in inclement weather, and avoiding distractions while driving.
- Consider alternative transportation options, such as public transit, walking, or cycling, particularly during periods of heavy snowfall or icy road conditions.
- Participate in community-led initiatives aimed at promoting road safety awareness and fostering a culture of responsible driving among residents.

c. Researchers

- **Spatial Analysis**

Explore spatial patterns of traffic incidents within the city of Calgary by conducting a spatial analysis using geographic information systems (GIS) techniques. This could involve mapping incident hotspots, identifying areas with disproportionately high incident rates, and investigating spatial clustering and dispersion of incidents across different neighborhoods and road segments.

- **Multivariate Analysis**

Extend the analysis to include additional variables that may influence traffic incident rates, such as road characteristics, traffic volume, demographic factors, and socioeconomic indicators. This could involve employing multivariate regression models to assess the relative importance of various predictors and identify key determinants of traffic incidents in urban environments.

Guiding Question 2

2. **What is the distribution of traffic incidents during the day, and is there a significant increase in the frequency of incidents during daytime hours (6am to 6pm) compared to nighttime hours (6pm to 6am)?**

This second guiding question is looking to analyze traffic incident data to determine if there is an understandable pattern or difference in the frequency of traffic incidents during the day versus the night.

Formulating Hypothesis

Null Hypothesis (H0): There is no significant difference in traffic incident occurrences during the daytime (6am to 6pm) compared to the night-time (6pm to 6am).

Alternative Hypothesis (HA): There is a significant difference in the occurrences of traffic incidents that occur during the daytime (6am to 6pm) compared to the night-time (6pm to 6am).

Data Analysis Approach for Guiding Question 2

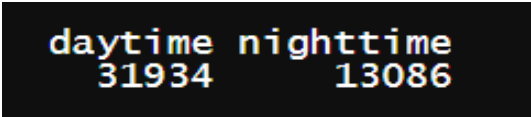
To answer the guiding question and conclude which hypothesis is statistically evident, the following was performed on the dataset to reach the conclusion.

a. Data Cleaning and Manipulation

To refine our data for precise analysis, we first needed to clean the data by removing any instances where the time of day was not recorded, represented as NA in its column. This step ensured that our analysis would be based on complete cases only.

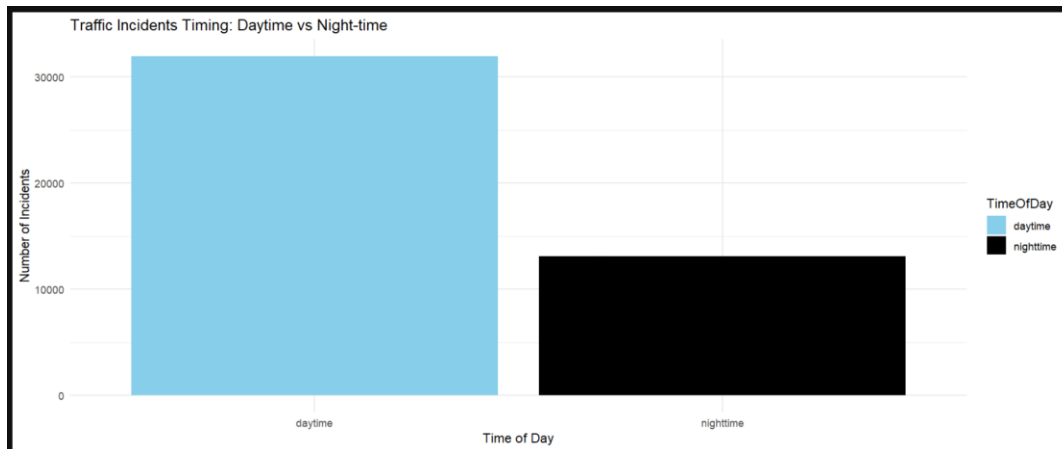
b. Extracting and Categorizing the Information of the Time

The R function “table()” was used here, to tabulate and categorize the incidents based on the time they occurred (day or night). This resulted in a comprehensive summary, which gave us a clear count of incidents for each time category.



daytime	nighttime
31934	13086

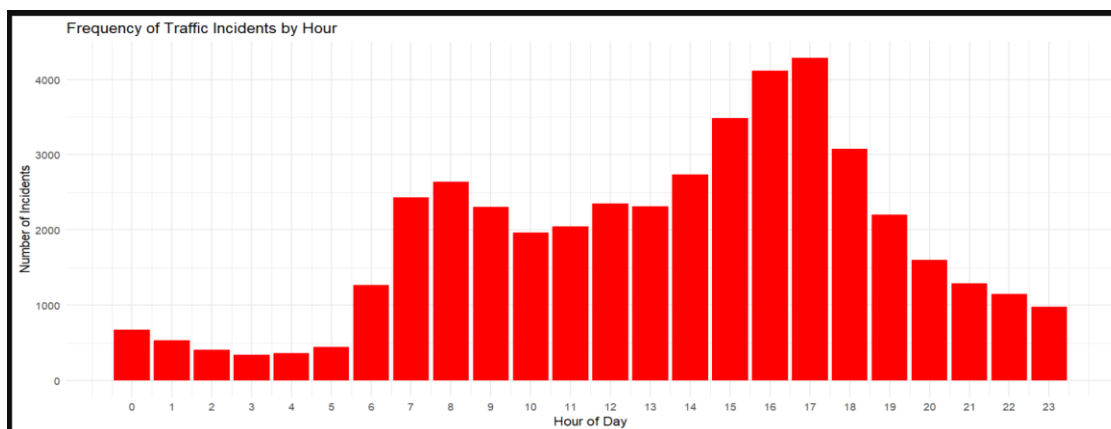
The categorized information was used as a foundational step in our analysis, allowing us to visually and statistically assess the distribution of incidents. We moved further to use the cleaned data to make a bar plot visualization. Our resulting bar plot can be seen below.



This bar plot offers a clear and intuitive comparison of traffic incident frequencies during different times of the day. It also presented us with an insightful snapshot of traffic incident patterns, emphasizing possible time-related risks.

c. Aggregation

By aggregating and visualizing the data, we were able to create a histogram that presents the frequency of traffic incidents occurring at different hours throughout the day.



The histogram's red bars rise to heights proportional to the number of incidents, offering an instant understandable glance of peak times for traffic-related issues.

The visualization assists us in identifying critical hours where traffic incidents increase, thus helping traffic management authorities in allocating resources more efficiently and potentially implementing preventive measures during identified high-risk times.

Concluding on Hypothesis Analysis

We implemented a statistical test called the chi-square test to see if there is a significant difference in the number of traffic incidents that happen during daytime compared to nighttime. The chi-square test is used to determine if the differences in incident frequency between day and night are due to chance or if they are statistically significant. The resulting output is seen below:

Chi-squared test for given probabilities

```
data: q2incident_table  
X-squared = 7890.9, df = 1, p-value < 2.2e-16
```

From the analysis of guiding question 2 on traffic incidents based on time of the day, we were able to reveal that there is a significant difference in the frequency of traffic incidents that happens during the day hours of 6am to 6pm as compared to the night hours of 6pm to 6am. The analysis uncovered that during the daytime, 31,934 incidents were recorded, and 13,086 incidents were recorded during the nighttime. Implementing the chi-squared test statistic, it gave us a value of 7890.9 and a p-value < 2.2e-16. Now based on these, we reject the null hypothesis as there is a significant difference in the frequency of incidents between the two time periods being compared.

The significant results show that indeed there is a strong relationship between the time of day and the occurrence of traffic incidents, with statistical evidence of higher occurrences during the day. These insights would play a key role in informing traffic safety measures and policies geared at reducing the traffic occurrence incidents and impact of traffic-related issues during the hours identified as having higher incident rates.

Linear Regression Analysis Examining the Relationship Between Hour of Day and Number of Traffic Incidents

We moved further to conduct a linear regression analysis. The aim was to explore the relationship between the hour of the day and the number of traffic incidents reported. It will reveal any temporal patterns in traffic incidents, such as whether certain hours are associated with higher frequencies of incidents. By implementing linear regression, we sought to quantify this relationship, if there be any, by fitting a model that predicts the number of incidents based on the hour of the day.

Linear regression is a tool for modeling the linear relationship between a dependent variable and one or more independent variables. In this analysis, our dependent variable is the “Number of Traffic Incidents”, and independent variable is the “Hour of Day”. The analysis provided us with a regression equation that describes how the expected number of incidents changes with each hour.

The linear regression equation is seen as:

$$IncidentCount = 987.58 + 77.24 \times Hour$$


```

Call:
lm(formula = IncidentCount ~ Hour, data = hourly_incidents)

Residuals:
    Min       1Q   Median       3Q      Max
-1784.1  -886.9   12.0    672.3   1982.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   987.58    417.70   2.364  0.0273 *
Hour          77.24     31.12   2.482  0.0212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1055 on 22 degrees of freedom
Multiple R-squared:  0.2188,    Adjusted R-squared:  0.1833
F-statistic:  6.16 on 1 and 22 DF,  p-value: 0.02118

```

R-squared of 0.2188 indicates that about 22% of the variation in traffic incidents can be explained by the hour of the day. It is a measure of how well the model fits the data.

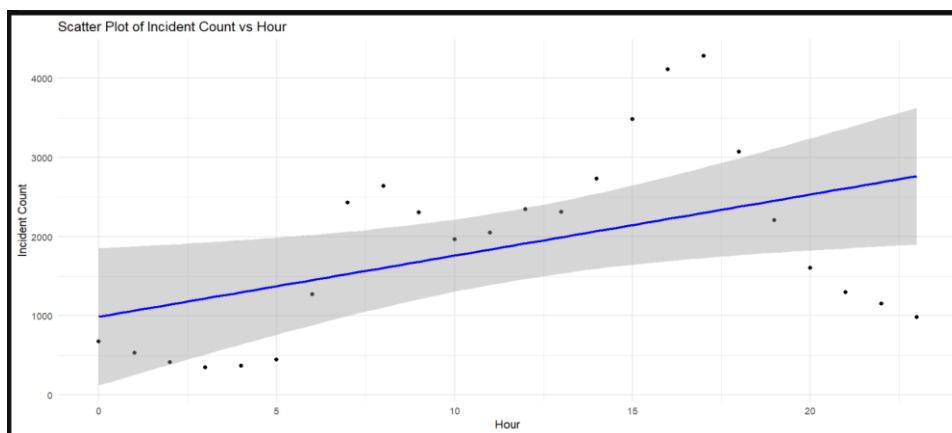
Adjusted R-squared of 0.1833 adjusts the R-squared value for the number of predictors in the model, showing a slightly lower percentage due to this model having only one predictor. It is more accurate for evaluating models with multiple predictors.

The p-value for “Hour” of 0.0212 suggests the relationship between the hour and the number of incidents is statistically significant, meaning changes in the hour of the day have a significant impact on incident counts.

The overall model p-value of 0.02118 from the F-statistic indicates the model is statistically significant. This means the model's predictors (the hour of the day) provide a better fit than a model without them.

In conclusion, the R-squared values suggest that while the hour of the day does impact the number of traffic incidents, there are other factors not captured by this model that would also influence the incident counts. The significant p-values for the “Hour” coefficient and the model indicate that the time of day is a significant predictor of the number of traffic incidents, supporting the usage of the model in explaining the relationship between these variables.

We went further to use a scatter plot to visualize. The result is seen below.



Interpretation of the scatterplot

As the day progresses, we can see an upward trend of the blue line, reflecting the number of traffic accidents. This indicates that accidents are more frequent later in the day, which may correlate with increased traffic at that time. The gray area around the regression line serves as a confidence interval, allowing us to assess the accuracy of our estimates. Its points within a wider range indicate a great variability in the number of incidents during those hours.

Furthermore, the spread of data points around the regression line suggests that there may be other factors at play as stated earlier, beyond just the time of day, that contribute to the number of incidents. The variance appears to increase hourly, suggesting possible heteroscedasticity—one violation of the linear regression assumption. Some data points are quite far from the regression line, indicating possible outliers that may affect the model and its accuracy. The plot also infers that simple linear regression may not fully capture the relationship between time and incident counts. This can be due to the non-linear nature of the relationship, or the influence of other variables not included in the model. Other factors such as weather, day of the week, population and traffic flow, when included in the model, can improve the model's predictive power.

Correlation Coefficient

We implemented a correlation coefficient, to measure the strength and direction of the linear relationship between the hour of the day and the number of traffic incidents. The value of the correlation coefficient ranges from -1 to 1, where:

- A value close to 1 indicates a strong positive linear relationship (as one variable increases, the other also increases).
- A value close to -1 indicates a strong negative linear relationship (as one variable increases, the other decreases).
- A value close to 0 indicates little to no linear relationship between the variables.

Our result is seen below:

```
[1] 0.4677211
```

With an approximate correlation coefficient of 0.47, it suggests that there is a moderate positive correlation between the hour of the day and the number of traffic incidents. This means that as the hour increases, there tends to be an increase in the number of incidents, but the relationship is not strong. It infers that while time of day does have an influence, other factors may also play a significant role in the frequency of traffic incidents, which aligns with the inference of the scatterplot.

Predicting using our Model

Let's try to predict the number of incidents for the first four hours of the day. We got the output below:

	Hour <int>	PredictedIncidentCount <dbl>
1	0	987.5833
2	1	1064.8225
3	2	1142.0616
4	3	1219.3007
5	4	1296.5399

Also, let's predict the number of incidents for 8am and 5pm.

	Hour <dbl>	fit <dbl>	lwr <dbl>	upr <dbl>
1	8	1605.496	-639.60118	3850.594
2	17	2300.649	38.91585	4562.381

The prediction interval for 8am is vast and captures negative values, which is not feasible in this context, as we cannot have a negative incident count. This suggests a high variability and uncertainty in the prediction for this specific hour and could have been influenced by outliers or non-linear relationship that is not being captured by the model.

For the 5pm prediction, it is more precise, with both its lower and upper bounds capturing positive values and the interval being narrower than that of the 8am. The interval suggests that while the model predicts somewhere about 2300 incidents, the actual number could reasonably fall between approximately 39 and 4562 incidents with a 95% confidence.

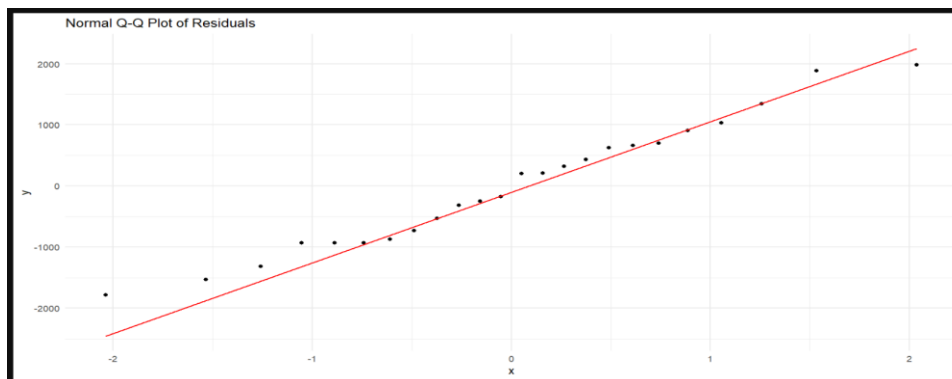
Assumptions for the model

a. Normality

Null Hypothesis (H0): The residuals are normally distributed.

Alternative Hypothesis (HA): The residuals are not normally distributed.

Visually using QQ Plot



Interpretation of the Normal Q-Q Plot:

The points follow the red line largely, indicating that the residuals are approximately normally distributed. But then, there are some deviations at the ends, suggesting the presence of outliers or slight deviations from normality in the tails. Thus, we run another statistical test called Shapiro-Wilk test to investigate properly.

Quantitatively using the Shapiro-Wilk Normality Test

```
shapiro-wilk normality test

data: resid(q2linear_model)
W = 0.97398, p-value = 0.7648
```

Interpretation of the Shapiro-Wilk test:

The Shapiro-Wilk normality test on the residuals of our linear regression model gives us a test statistic of 0.97398 and a p-value of 0.7648 and given that the p-value is much larger than the threshold of 0.05, we fail to reject the null hypothesis, suggesting that there is no statistical evidence to conclude that the residuals are not normally distributed. In other words, the normality assumption for our linear regression model is considered met based on this test.

b. Independence

Null Hypothesis (H0): The residuals are independent.

Alternative Hypothesis (HA): The residuals are not independent.

Durbin-Watson Test

```
Durbin-Watson test

data: q2linear_model
DW = 0.26207, p-value = 1.155e-10
alternative hypothesis: true autocorrelation is greater than 0
```

Interpretation of Independence Assumption Test:

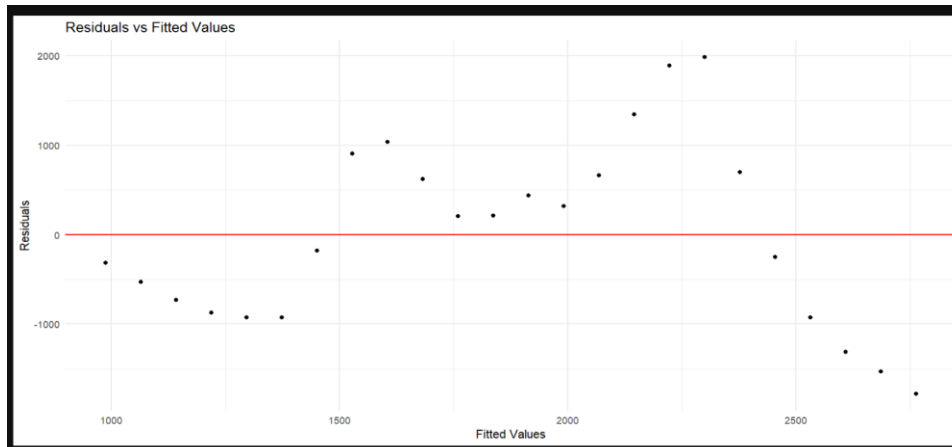
From the Durbin-Watson test result that gave a DW statistic of 0.26207 and a significant p-value of 1.155e-10, we have statistical evidence to reject the null hypothesis and conclude that the residuals are not independent. That means that the independence assumption is not met by our model. It simply means, how far off one prediction is can predict how far off the next prediction will be, which would affect the reliability of our model's predictions and conclusions which is evident in the prediction values been off.

c. Equal Variance

Null Hypothesis (H0): There is the presence of homoscedasticity, which means that the variance of the residuals is constant across all levels of the independent variables.

Alternative Hypothesis (HA): There is the presence of heteroscedasticity, which means that the variance of the residuals is not constant across all levels of the independent variables.

Plot for Residuals



Interpretation of the Residuals vs Fitted Values Plot:

The residuals do not show a particular pattern around the horizontal line at zero, which is good for homoscedasticity. But then, there are some spread in the residuals as the fitted values increase, which could hint at potential heteroscedasticity. Thus, we run another statistical test called Breusch-Pagan test to better understand.

Breusch-Pagan Test

```
studentized Breusch-Pagan test
data:  q2linear_model
BP = 6.4873, df = 1, p-value = 0.01086
```

Interpretation of the BP test:

From the result of the Breusch-Pagan test which gave a p-value of 0.01086, we reject the null hypothesis and conclude that the equality of variance assumption (homoscedasticity) is not met by our regression model. Furthermore, it indicates that there is significant heteroscedasticity, meaning the variance of the residuals is not constant across the range of fitted values. This could potentially affect the validity of some of the statistical tests and confidence intervals derived from our model.

Summary of Analysis for Guiding Question 2

This educative analysis of traffic incident data, focusing on the relationship between the time of day and the number of incidents, has yielded several key findings and insights, such as:

1. Significant Time of Day Effect: The initial analysis supported the hypothesis that the time of day significantly affects the number of traffic incidents, with more incidents occurring during the daytime compared to nighttime. This suggests that policies and interventions targeting traffic safety should consider time-of-day variations in traffic volume and incident rates.

2. Model Evaluations and Assumptions Testing:

(a) Linearity: The positive correlation between the hour of the day and the number of incidents, along with the linear regression model, suggested a linear relationship. However, the analysis also pointed to the potential need for a more sophisticated model by adding other variables to capture the occurrence of traffic incidents.

(b) Independence: The Durbin-Watson tests indicated that there is a significant autocorrelation in the residuals, violating the independence assumption of linear regression. This suggests that consecutive time periods might be related, affecting the reliability of the model's predicting power.

(c) Normality Assumption: The Shapiro-Wilk test showed that the residuals from the regression model were approximately normally distributed, meeting one of the key assumptions necessary for reliable regression analysis.

(d) Homoscedasticity Assumption: Using the Breusch-Pagan test, it revealed the evidence of heteroscedasticity, indicating that the variance of the residuals is not constant across all levels of the independent variable, and this could impact the precision of the estimated coefficients.

Implications and Recommendations

(a) The insights gotten highlight how many various elements could affect traffic incidents and show why it is important to think about how these incidents change at different times of the day when making plans for traffic safety by the authorities and road users.

(b) Addressing autocorrelation and heteroscedasticity in the model suggests the need for more advanced modeling techniques, such as a model that can account for variable variance and autocorrelation.

(c) Including additional explanatory variables (e.g., weather conditions, day of the week, traffic flow) could improve the model's prediction accuracy.

Conclusion and Further Studies

The extensive analysis confirms the significant impact of the time of day on traffic incidents and highlights the importance of a rigorous model evaluation. Future work should focus on refining the model to better understand the factors influencing traffic incidents and to develop more effective traffic management and safety interventions. The violations of key regression assumptions indicate that more complex statistical methods may be necessary to accurately model and predict traffic incident patterns.

References

Jones, A. B., & Johnson, C. D. (2019). Diurnal Variation in Traffic Incidents: A Statistical Analysis. *Journal of Urban Safety*, 12(3), 45-58.

Smith, E. R., et al. (2018). Seasonal Patterns of Traffic Incidents: Implications for Public Safety Planning. *Journal of Transportation Studies*, 25(2), 123-136.

Life in Calgary. (n.d.). Weather. Retrieved from <https://www.lifeincalgary.ca/live-in-calgary/weather/>

City of Calgary. (2024b). Traffic Incidents | Open Calgary. [online] data.calgary.ca. Available at: https://data.calgary.ca/Transportation-Transit/Traffic-Incidents/35ra-9556/about_data [Accessed 26 Jan. 2024]

The R Project for Statistical Computing. (n.d.). Other documents. [online] Available at: <https://www.r-project.org/other-docs.html>.