Highway Hustle - Unraveling Metro Interstate Traffic Trends

Group Members:

Abraar Patel (2069545)

Talha Sohail (2092858)

Ujwal Joshi (2018577)

Kevin Zheng (2022183)

Alton Phan (2020230)

INTRODUCTION

In the ever-evolving world of urban landscapes, we find inspiration in the vibrant heartbeat of cities and the endless potential for progress. As population growth and urbanization continue their inexorable march, our motivation to decode the intricate patterns of urban life intensifies. Urban transportation networks face increasing challenges due to population growth and urbanization. Understanding traffic patterns and congestion on metro interstate highways is crucial for efficient urban planning and management. Therefore, we chose a comprehensive **metro interstate traffic** dataset sourced from Kaggle to analyze and uncover valuable insights into traffic flow, congestion and accurately forecast traffic volume. The dataset consists of **48,205 observations** and following **9 variables**:

- ***holiday***: a categorical variable representing a national or regional holiday.
- ***temp***: a numeric variable that represents the average temperature in kelvin.
- ***rain_1h***: a numeric variable that shows the amount of rain occurring in an hour in mm.
- ***snow_1h***: a numeric variable that shows the amount of snow occurring in an hour in mm.
- ***clouds_all***: numeric variable representing the percentage of cloud cover.
- ***weather_main***: categorical variable containing a short textual description of the current weather.
- ***weather_description***: categorical variable providing longer textual description and detail for the current weather.
- ***date_time***: datetime variable that shows the hour of the data collected in local CST time.
- ***traffic_volume***: numeric variable that shows the hourly I-94 reported westbound traffic volume.

The **main question** we want to answer is: **Which variables or factors have the most significant impact on traffic volume?**

```
# Importing the dataset and displaying the summary
traffic = read.csv("C:/Users/alton/OneDrive/Documents/MATH4322_GroupProject/Metro_Intersta
summary(traffic)
```

```
traffic_volume   holiday               temp          rain_1h
Min.   :   0   Length:48204       Min.   :  0.0   Min.   :   0.000
1st Qu.:1193   Class :character   1st Qu.:272.2   1st Qu.:   0.000
```

```
Median :3380   Mode  :character   Median :282.4   Median :   0.000
Mean   :3260                      Mean   :281.2   Mean   :   0.334
3rd Qu.:4933                      3rd Qu.:291.8   3rd Qu.:   0.000
Max.   :7280                      Max.   :310.1   Max.   :9831.300
   snow_1h            clouds_all     weather_main       weather_description
Min.   :0.0000000   Min.   :  0.00   Length:48204       Length:48204
1st Qu.:0.0000000   1st Qu.:  1.00   Class :character   Class :character
Median :0.0000000   Median : 64.00   Mode  :character   Mode  :character
Mean   :0.0002224   Mean   : 49.36
3rd Qu.:0.0000000   3rd Qu.: 90.00
Max.   :0.5100000   Max.   :100.00
 date_time
Length:48204
Class :character
Mode  :character
```

```r
# checking if there are any missing values in the entire dataset
# (returns TRUE if there are missing values , otherwise FALSE)
any(is.na(traffic))
```

```
[1] FALSE
```

There are no missing values in the dataset.

```r
# Converting the temperatures from Kelvin to Farenheit
traffic$temp <- (traffic$temp - 273.15) * 9/5 + 32
```

After conducting some data visualizations, it became evident that data cleaning needed. Implementing data cleaning is an important step in the data analytics process.

Data Cleaning (Abraar and Talha): During data cleaning, we formatted the date_time column, extracted time-related features, and categorized hours into distinct periods. We converted the holiday variable to a binary format, where 0 represents no holiday, and 1 represents a holiday. The outliers in temp and rain_1h column were removed. We also simplified weather conditions into 'thunderstorm,' 'mist,' 'fog,' and 'haze,' since they are the most distinct weather conditions and then created dummy variables for weather_description column. Additionally, we separated the snow_1h variable into two categories: "snow" and "no_snow" and then created a new binary variable, snow_present, which indicates the presence of snow (1) or the absence of snow (0) in the dataset. The unnecessary columns

were then dropped. The cleaned dataset has 15 variables: ***traffic_volume***, ***holiday***, ***temp***, ***rain_1h***, ***clouds_all***, ***date***, ***weekday***, ***hour***, ***month***, ***year***, ***fog***, ***haze***, ***mist***, ***thunderstorm***, ***snow_present.***

```
# Removing the outlier in temp variable and rain_1h variable
traffic <- traffic[traffic$temp > -400, ]
traffic <- traffic[traffic$rain_1h <2500, ]

traffic$date_time <- strptime(traffic$date_time, format = "%d-%m-%Y %H:%M")
# Formatting the date_time column in the desired format (%Y-%m-%d %H:%M)
traffic$date_time <- strftime(traffic$date_time, format = "%Y-%m-%d %H:%M")

# Extracting additional features from date_time variable
#(For weekdays, Monday is 0 and Sunday is 6)
traffic$date_time = as.POSIXct(traffic$date_time)
traffic$date = as.Date(traffic$date_time)
traffic$weekday = as.numeric(format(traffic$date_time, "%w"))
traffic$hour = as.numeric(format(traffic$date_time, "%H"))
traffic$month = as.numeric(format(traffic$date_time, "%m"))
traffic$year = as.numeric(format(traffic$date_time, "%Y"))
```

The full data cleaning code can be found in the R source code file.

```
# Separating snow_1h into categories such as "snow" and "no_snow"
traffic$snow_1h <- ifelse(traffic$snow_1h > 0, "snow", "no_snow")

# creating new column snow_present (binary variable) which specifies
# if there is a snow or not
traffic$snow_present <- ifelse(traffic$snow_1h == "snow", 1, 0)

# Dropping the unnecessary or not required columns
# (date_time, weather_descriptionother column, weather_description
# and weather_main column)
# Dropped the date_time column because we already extracted the
# features from the date_time column.
# Dropped snow_1h since we already one-hot encoded these columns.
traffic <- traffic[, !colnames(traffic) %in%
                     c("date_time", "snow_1h", "weather_description",
                       "weather_descriptionother",
                       "weather_main")]
```

METHODS

Linear Regression Model (Abraar and Talha)

In this section, we are focusing on utilizing a linear regression model to see which predictors are significant in impacting and predicting the interstate highway traffic volume. Linear Regression model tries to find the direct correlation between the response variable, **traffic_volume**, against the other potential predictors. We are using linear regression model because the response variable is a continuous or quantitative variable. The biggest advantage of this model is the ability to clearly understand how the response variable is changing with a one unit increase/decrease in each of the predictors. Linear Regression model is also easy to understand and interpret and is also quick to train and make predictions.The disadvantage of linear regression model is that it does not capture complex and non-linear relationships in the data. Furthermore, linear regression models exhibit sensitivity to outliers, leading to potential distortions in regression coefficients and predictive outcomes.

The linear regression formula would be as follows:

$$\begin{aligned}\text{traffic\_volume} =& \beta_0 + \beta_1 * \text{holiday} + \beta_2 * \text{temp} + \beta_3 * \text{rain\_1h} \\ &+ \beta_4 * \text{clouds\_all} + \beta_5 * \text{date} + \beta_6 * \text{weekday} + \beta_7 * \text{hour} \\ &+ \beta_8 * \text{month} + \beta_9 * \text{year} + \beta_{10} * \text{fog} + \beta_{11} * \text{haze} \\ &+ \beta_{12} * \text{mist} + \beta_{13} * \text{thunderstorm} + \beta_{14} * \text{snow\_present} + \epsilon\end{aligned}$$

```
# creating the linear regression model
traffic.lm = lm(traffic_volume ~ ., data=traffic)
summary(traffic.lm)
```

```
Call:
lm(formula = traffic_volume ~ ., data = traffic)

Residuals:
    Min      1Q  Median      3Q     Max
-5227.8 -1650.2   -63.2  1533.5  4873.3

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.778e+06  6.893e+05  -6.931 4.22e-12 ***
holiday      -1.190e+03  2.354e+02  -5.055 4.32e-07 ***
temp          1.012e+01  3.929e-01  25.756  < 2e-16 ***
rain_1h      -5.210e+01  8.498e+00  -6.131 8.78e-10 ***
clouds_all    3.774e+00  2.219e-01  17.009  < 2e-16 ***
```

```
date         -6.660e+00  9.577e-01  -6.954 3.59e-12 ***
weekday       7.554e+01  4.183e+00  18.058  < 2e-16 ***
hour          9.438e+01  1.229e+00  76.788  < 2e-16 ***
month         1.872e+02  2.930e+01   6.387 1.71e-10 ***
year          2.426e+03  3.499e+02   6.934 4.14e-12 ***
fog          -2.421e+02  6.168e+01  -3.925 8.67e-05 ***
haze          2.470e+02  5.084e+01   4.858 1.19e-06 ***
mist         -1.944e+02  2.616e+01  -7.434 1.07e-13 ***
thunderstorm -4.755e+02  5.925e+01  -8.025 1.03e-15 ***
snow_present -1.614e+02  2.316e+02  -0.697    0.486
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1834 on 48178 degrees of freedom
Multiple R-squared:  0.148, Adjusted R-squared:  0.1477
F-statistic: 597.6 on 14 and 48178 DF,  p-value: < 2.2e-16
```

From the above summary of our linear regression model, it is evident that almost all predictors are statistically significant in predicting the traffic volume except snow_present variable. The p-value for snow_present variable is greater than 0.05 (0.486). Hence, snow_present variable is not statistically significant in predicting the traffic volume and we would exclude it from consideration. Rest of the predictors have a p-value less than 0.05 and are considered statistically significant in predicting the traffic volume. Therefore, we would retain these predictors in our model.