Task - Solutions:

1.1 Project title

Interaction between motor vehicle Collisions and Weather in New York city borough's.

1.2 General description of the domain

The domain of the project focuses on the impact of weather conditions on road accidents in the boroughs of New York City over the specified time frame (2016 - 2019). The boroughs to be analyzed based on the collision data of New York City are Manhattan, Bronx, Queens, Brooklyn, and Staten Island. These boroughs experience a substantial volume of motor vehicle collisions. Motor vehicle collisions are a significant concern, impacting public safety and transportation infrastructure.

By examining collisions data factors such as.

- *Crash dates,*
- *Times,*
- *Severity,*
- *Contributing factors*

alongside weather conditions, such as.

- *visibility,*
- *wind speed,*
- *temperature,*
- *road condition,*

the project may influence the frequency, severity of accidents and mitigating risks. Moreover, the project seeks to contribute to the development of strategies and interventions aimed at reducing motor vehicle collisions and promoting safer driving practices, especially in adverse weather conditions.

1.3 Description of the analysis area with justification

I will be analyzing motor vehicle collision data in different brough's in conjunction with weather condition from 2016-2019 data. This analysis is going to be aimed to uncover correlations between weather patterns and the frequency and severity of collisions. How often and how bad those crashes are. By understanding these patterns, I aimed to identify risk factors contributing to accidents in NewYork city and devise strategies to improve road safety. However, there are multiple conditions that influences the collision of the vehicles.

1.4 identified Problems

Please state and comment the identified decision problem(s) in the assumed domain.

The identified decision problem in the domain is understanding the correlation between weather conditions and motor vehicle collisions in NewYork city. By delving into this correlation, we intent to see if certain weather patterns contribute to more frequent collisions.

- *Determining the most effective analytical approach to extract insights from the data as well as weather conditions.*
- *Addressing challenges related to data granularity and time aspects, such as analyzing data over specific time periods like seasons to understand seasonal patterns in collision occurs more often. This way we can understand if certain times of the year are riskier for accidents in NewYork city.*

1.5 Project Goal

Please identify basic users (types) – provide brief description of each type – and state expectations of each type. Do not select too many types – as it might be beneficial to focus on more details for 3-4 user types. Further present detailed needs by specify 10-15 OLAP user query types (do not write any SQL queries, these should be formulated in natural language – strictly in business terms). Note – these should not be very specific queries (using specific attribute values), but rather general query/need types. For each query/need please indicate how a user can utilise resultant information. Match each query to the identified user types. Provide results in section 1.5.1.

Perform basic analysis of these query types (1.5.1), try to infer some general user requirements. Focus on identifying event(s) and perspectives that the user is interested in. Further, perform basic analysis of user query types (1.5.1), try to infer some detailed user requirements. List all possible measures and all possible individual dimensions. Provide results in section 1.5.2

1.5.1 Expectations and detailed needs for decision support

- ***Users:***

NewYork Department of Transportation: *The primary user responsible for road safety initiatives and infrastructure development in NewYork city boroughs. They utilize data to implement measures that reduce motor vehicle collisions.*

Insurance Companies: *Insurance firms interested in understanding collision trends and risk factors to adjust premiums and develop risk management strategies.*

- ***Expectations:***

New York Department of Transportation: *Seeks insights to prioritize road maintenance and implement traffic safety measures based on collision data.*

Insurance Companies: *Expect to identify high-risk areas and factors contributing to collisions to determine insurance premiums and mitigate risks.*

- ***OLAP User Query Types:***

New York Department of Transportation:

- *Obtain the total number of collisions in New York city per month for the past year.*
- *Analyze the distribution of collision severity (injured and killed persons) by vehicle type on specific streets in New York city.*
- *Identify the most common contributing factors to collisions during nighttime hours in New York city.*

Insurance Companies:

- *Calculate the average number of collisions per day by vehicle type (e.g., sedan, truck, motorcycle) across all boroughs.*

- *Identify the most common contributing factors to collisions involving taxis in New York during rush hour.*
- *Determine the correlation between collision frequency and zip code demographics to assess risk factors for insurance premiums.*

Total Collisions Over Time

- *Total number of collisions per year, month, or day.*
- *Analyze trends in collision frequency over different time periods.*

Injuries and Fatalities

- *Calculate the sum of injured and killed persons over time.*
- *Compare the number of pedestrians and cyclists injured or killed in collisions.*

Location-Based Analysis

- *Analyze collision data by borough, ZIP code, or specific streets to identify high-risk areas.*
- *Compare collision rates between different locations to prioritize safety measures.*

Vehicle Type Analysis

- *Break down collision data by vehicle types to understand which types are involved in the most accidents.*
- *Identify trends in collisions involving specific vehicle types over time.*

Day and Time Analysis

- *Analyze collisions based on the time of day (day or night) to identify patterns in accident occurrence.*
- *Determine if there are specific hours or periods of the day with higher collision rates.*

Weather Impact Analysis

- *Correlate weather data with collision counts to analyze the impact of weather conditions on road safety.*
- *Determine if there's a relationship between maximum temperature, visibility, or wind speed and collision rates.*

1.5.2 Scope of analysis – aspects examined

There are multiple aspects of criteria we can examine the data by, depending on what we want to obtain. For example, in the case of insurance companies, they would expect get detailed data on specific crash factors, such as contributing factors (aggressive, slippery pavement), together with the 'edge cases' like low/high temperatures, high winds, snow, rain etc.), and time of the day could provide high risk scenarios and help in assessing claim patterns. Moreover, city transportation departments might be interested in general crash statistics such as monthly trends, locations, with high crash rates, and factors contributing to crashes. This data may inform targeted interventions and improvements based on the analyze to enhance road safety.

Detailed User Requirements (Measures and Dimensions):

Possible Measures ->

Count of Collisions

Number of Persons Injured

Number of Persons Killed

Number of Pedestrians/Cyclists/Motorists Injured or Killed

Possible Dimensions ->

Crash Time Dimension: *Surrogatekey, (crash time:2:39 than the id is: '0239'), Hour, minute, periodofday (Day-> 6.00 – 17.59,   Night -> 18.00 – 5.59)()*

Crash Date dimension: *Surrogatekey, fulldate(yyyy-mm-dd), dayofmonth, monthofyear, year, dayofweek*

Weather Dimensio*n: Surrogatekey(weather_id), MaximumTemperature, MinimumTemperature, WindSpeed, Humidity, Visibility, Cloud coverage, weather_description*

Location Dimension: *id, borough, zipcode, onstreetname, off street name, cross street name, zip code.*

Vehicle type dimension: *(vehicle id fk), vehicle type(sedan, Taxi, Box truck, Motorscooter, bike, bus, station wagon, garbage truck, motorcycle, etc.) contributing factor vehicle type.*

1.6 Data sources

1.6.1 Location, format, availability

Data sets are taken from:

Weather Data*: https://www.visualcrossing.com/weather/weather-data-services*

Motor Vehicle Collisions: *https://opendata.cityofnewyork.us/*

Motor Vehicle dataset which taken from the City of New York website, contains the information starting from 2016 to 2019 with the several locations like; Brooklyn, Queens, Bronx, Manhattan and Staten Island. However, in the visual crossing website I was only able to select one city and download weather history of only 2 years in once which pushed me to merge the data's by downloading them separately. All datasets are in csv format.

The Motor Vehicle Collisions data set contains information:

- *Crash date*
- *Crashtime*
- *Borough*
- *ZIPCODE*
- *LATITUDE*
- *LONGITUDE*

- *LOCATION*
- *ONSTREETNAME, CROSSSTREETNAME, OFFSTREETNAME*
- *NUMBER OF PERSONS INJURED, NUMBER OF PERSONS KILLED*
- *NUMBER OF PEDESTRIANS INJURED, NUMBER OF PEDESTRIANS KILLED*
- *NUMBER OF CYCLIST INJURED, NUMBER OF CYCLIST KILLED*
- *CONTRIBUTING FACTOR VEHICLE 1, CONTRIBUTING FACTOR VEHICLE 2,3,4,5*
- *COLLISION_ID*
- *VEHICLE TYPE CODE 1, VEHICLE TYPE CODE 2,3,4,5*

Weather dataset contains information: 2016-2019.

- *Datetime*
- *Tempmax*
- *Tempmin*
- *Temp*
- *Feelslikemax*
- *Feelslikemin*
- *Feelslike*
- *Dew*
- *Humidity*
- *Precip*
- *Precipprob*
- *Precipcover*
- *Preciptype*
- *Snow*
- *Snowdepth*
- *Windgust*
- *Windspeed*
- *Winddir*
- *Sealevelpressure*
- *Cloudcover*
- *Visibility*
- *Solarradiation*
- *solarenergy*
- *uvindex*
- *severerisk*
- *sunrise*
- *sunset*
- *moonphase*
- *conditions*
- *description*
- *icon*
- *stations*

1.6.2 Data source basic information

	Source	Number of rows	Number of attributes	Size	Update rate	Grain
1	Motor Vehicle Collisions	*1048576*	33	422MB	No update yet	Total,per year,per month,per day,per hour,min
2	weather 2016-2019	*7305*	14	2.2MB	Daily	year,per month,per day, time

1.6.3 Data source Initial assesment (Additional information. time span for facts, etc.)

Weather dataset – 7305 rows, 14 columns (There are some unnecessary columns which going to be removed before the start to process.

Motor Vehicle Collisions dataset contains borough about: Manhattan, Brooklyn, Staten Island, Queens, I have downloaded past years weather dataset of New York City.

Motor Vehicle Collisions – 1048576 rows, 33 columns (There are some unnecessary columns which going to be removed before the start to process.)

1.7-DIMENSIONAL model synopsis

1.7.1 facts and measures

	Fact	Measure(s)	Grain
	Fact_collisions		
1	*Collisions*	*Count (crashdate)*	*Total,per year,per month,per day,hour*
2	*Injured person number*	*Total number of injured person. Average number of people injured.*	*Total,per year,per month,per day,hour*
3	*Killed person number*	*Total number of killed person. Average number of killed person.*	*Total,per year,per month,per day,hour*
4	*Pedestrians injured number*	*Total number of pedestrians killed. Average number of pedestrians killed.*	*Total,per year,per month,per day,hour*
5	*Pedestrians killed number*	*Total number of pedestrians killed. Average number of pedestrians killed.*	*Total,per year,per month,per day,hour*
6	*Cyclist injured number*	*Total number of cyclists injured. Average number of cyclists injured.*	*Total,per year,per month,per day,hour*

7	*Cyclist killed number*	*Total number of cyclists killed. Average number of cyclists killed.*	*Total,per year,per month,per day,hour*
8	*Motorist injured number*	*Total number of motorists injured. Average number of motorists injured.*	*Total,per year,per month,per day,hour*
9	*Motorist killed number*	*Total number of motorists died. Average number of motorists killed.*	*Total,per year,per month,per day,hour*

1.7.2 Context for facts

	Dimension	Description	Grain
1	***Dimension_Weather***	*Weather conditions at the time of collision*	*Per year, month, day*
	Dimension_Location	*Location details of the collision*	*Per incident*
	Dimension_crash_time	*Time details of the collision*	*Per incident*
	Dimension_crash_date	*Date details of the collision*	*Per incident*
	Dimension_vehicle_type	*Type of vehicles involved in the collision and its contributing factor types,*	*Per vehicle involved in the collision*

1.7.3 Dimension overview

	Dimension	Hypothesized Attributes
1	***Dimension_Weather***	*Surrogatekey(weather_id), MaximumTemperature, MinimumTemperature, WindSpeed, Humidity, Visibility, Cloud coverage, description*
	Dimension_Location	*Id, borough, zip code, onstreet name, offstreetname, cross streetname,*
	Dimension_crash_time	*Surrogatekey, (crash time:2:39 than the id is: ‘0239’), Hour, minute, periodofday (Day-> 6.00 – 17.59, Night -> 18.00 – 5.59)()*
	Dimension_crash_date	*Surrogatekey, fulldate(yyyy-mm-dd), dayofmonth, monthofyear, year, dayofweek*
	Dimension_vehicle_type	*(vehicle id fk), vehicle type(sedan, Taxi, Box truck, Motorscooter, bike, bus, station wagon, garbage truck, motorcycle, etc.), contributing factor vehicle type.*

1.7 Data source details

1.7.1 Domain Data Dictionary

	Location	Attribute name	Attribute type	Description
1	*Motor_Vehicle_Collisions*	*CRASH DATE*	*DATE*	*Date of the collision*
2	*Motor_Vehicle_Collisions*	*CRASH TIME*	*Time (Interval)*	*Specific time of the collision*
3	*Motor_Vehicle_Collisions*	*BOROUGH*	*Text(Nominal)*	*Borough that where the collision was occurred*
4	*Motor_Vehicle_Collisions*	*ZIP CODE*	*Text(Nominal)*	*ZIP code of the collision location*
5	*Motor_Vehicle_Collisions*	*ON STREET NAME*	*Text(Nominal)*	*Name of the street where the collision occurred*
6	*Motor_Vehicle_Collisions*	*CROSS STREET NAME*	*Text(Nominal)*	*Name of the cross street*
7	*Motor_Vehicle_Collisions*	*OFF STREET NAME*	*Text(Nominal)*	*Name of the off-street location*
8	*Motor_Vehicle_Collisions*	*NUMBER OF PERSONS INJURED*	*Numerical(Ratio)*	*Number of persons injured in the collision*
10	*Motor_Vehicle_Collisions*	*NUMBER OF PERSON KILLED*	*Numerical(Ratio)*	*Number of persons killed in the collision*
11	*Motor_Vehicle_Collisions*	*NUMBER OF PEDESTRIANS INJURED*	*Numerical(Ratio)*	*Number of pedestrians injured in the collision*
12	*Motor_Vehicle_Collisions*	*NUMBER OF PEDESTRIANS KILLED*	*Numerical(Ratio)*	*Number of pedestrians killed in the collision*
13	*Motor_Vehicle_Collisions*	*NUMBER OF CYCLIST KILLED*	*Numerical(Ratio)*	*Number of cyclists injured in the collision*
14	*Motor_Vehicle_Collisions*	*NUMBER OF CYCLIST KILLED*	*Numerical(Ratio)*	*Number of cyclists killed in the collision*
15	*Motor_Vehicle_Collisions*	*NUMBER OF MOTORIST INJURED*	*Numerical(Ratio)*	*Number of motorists injured in the collision*
16	*Motor_Vehicle_Collisions*	*NUMBER OF MOTORIST KILLED*	*Numerical(Ratio)*	*Number of motorists killed in the collision*
17	*Motor_Vehicle_Collisions*	*VEHICLE TYPE*	*String*	*Vehicle type of the vehicle which belong to crash.*
18	*Motor_Vehicle_Collisions*	CONTRIBUTING FACTOR VEHICLE type	*String*	*Contributing factor for vehicle 1 in the collision*
19	*Weather_2016-2019*	*LocationName*	*Text(Nominal)*	*Borough where the collision occurred. This attribute represents the specific location or borough within*

				the city where the collision took place. Examples include Brooklyn, Queens, Bronx, Manhattan, and Staten Island.
20	*Weather_2016-2019*	*Tempmax*	*Numerical(Ratio)*	*The highest temperature recorded during the specified time period. This attribute indicates the maximum temperature measured in degrees Celsius or Fahrenheit, depending on the unit of measurement used.*
21	*Weather_2016-2019*	*Tempmin*	*Numerical(Ratio)*	*The lowest temperature recorded during the specified time period. This attribute indicates the minimum temperature measured in degrees Celsius.*
22	*Weather_2016-2019*	*WindSpeed*	*Numerical(Ratio)*	*The speed of the wind recorded at the observation time. This attribute represents the velocity of the wind such as kilometers per hour (km/h)*
23	*Weather_2016-2019*	*Humidity*	*Numerical(Ratio)*	*The level of humidity recorded at the observation time. This attribute represents the amount of moisture present in the air, typically expressed as a percentage.*
24	*Weather_2016-2019*	*Visibility*	*Numerical(Ratio)*	*The level of visibility recorded at the observation time. This attribute indicates the distance at which objects can be clearly seen, typically measured in kilometers (km).*
25	*Weather_2016-2019*	*Cloud Coverage*	*Numerical(Ratio)*	*The percentage of the sky covered by clouds at the observation time. This attribute represents the extent of cloud cover, where 0% indicates clear skies and*

				100% indicates completely overcast conditions.
26	*Weather_2016-2019*	*Weather Description*	*Text(Nominal)*	*A detailed description of the weather conditions observed during the specified time period. This attribute provides additional information about the weather, such as whether it was sunny, cloudy, rainy, foggy, or stormy, among other conditions.*

1.72.2 Quality assesment Sheet

In the data quality sheet for observing the number of unique values and number of null values, excel functions used -> =COUNTA(UNIQUE(AE1:AE10)), =COUNTBLANK(A1:A10).

	Location	Attribute name	Attribute type	Type of data	Number of Unique values	Number of Null values	*Quality assessment*
1	*Motor_Vehicle_Collisions*	*CRASH DATE*	*Date*	*Date/Time*	*1*	*0*	*Good - No missing values*
2	*Motor_Vehicle_Collisions*	*CRASH TIME*	*Time*	*Date/Time*	*1*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*BOROUGH*	*Categorical*	*Text (VarChar)*	*5*	*376686*	*Moderate - Significant missing values*
	Motor_Vehicle_Collisions	*ZIP CODE*	*Categorical*	*Text (VarChar) - Numerical*	*1*	*376870*	*Moderate - Significant missing values*
	Motor_Vehicle_Collisions	*ON STREET NAME*	*Categorical*	*Text (VarChar)*	*12511*	*257017*	*Moderate - Significant*

							missing values
	Motor_Vehicle_Collisions	*CROSS STREET NAME*	*Categorical*	*Text (VarChar)*	*9095*	*543024*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*OFF STREET NAME*	*Categorical*	*Text (VarChar)*	*166434*	*792886*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*NUMBER OF PERSONS INJURED*	*Quantitive*	*Numerical(Integer)*	*26*	*17*	*Good - Very few missing values*
	Motor_Vehicle_Collisions	*NUMBER OF PERSONS KILLED*	*Quantitive*	*Numerical(Integer)*	*7*	*30*	*Good - Very few missing values*
	Motor_Vehicle_Collisions	*NUMBER OF PEDESTRIANS INJURED*	*Quantitive*	*Numerical(Integer)*	*9*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*NUMBER OF PEDESTRIANS KILLED*	*Quantitive*	*Numerical(Integer)*	*7*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*NUMBER OF CYCLIST INJURED*	*Quantitive*	*Numerical(Integer)*	*4*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*NUMBER OF CYCLIST KILLED*	*Quantitive*	*Numerical(Integer)*	*3*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*NUMBER OF MOTORIST INJURED*	*Quantitive*	*Numerical(Integer)*	*24*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*NUMBER OF MOTORIST KILLED*	*Quantitive*	*Numerical(Integer)*	*5*	*0*	*Good - No missing values*
	Motor_Vehicle_Collisions	*VEHICLE TYPE CODE 1*	*Categorical*	*Text (VarChar)*	*937*	*8651*	*Good - Few*

							missing values
	Motor_Vehicle_Collisions	*VEHICLE TYPE CODE 2*	*Categorical*	*Text (VarChar)*	*1019*	*249012*	*Moderate - Significant missing values*
	Motor_Vehicle_Collisions	*VEHICLE TYPE CODE 3*	*Categorical*	*Text (VarChar)*	*160*	*975405*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*VEHICLE TYPE CODE 4*	*Categorical*	*Text (VarChar)*	*65*	*1031431*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*VEHICLE TYPE CODE 5*	*Categorical*	*Text (VarChar)*	*40*	*1043471*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*CONTRIBUTING FACTOR VEHICLE 1*	*Categorical*	*Text (VarChar)*	*59*	*3748*	*Good - Few missing values*
	Motor_Vehicle_Collisions	*CONTRIBUTING FACTOR VEHICLE 2*	*Categorical*	*Text (VarChar)*	*58*	*178532*	*Moderate-Significant missing values*
	Motor_Vehicle_Collisions	*CONTRIBUTING FACTOR VEHICLE 3*	*Categorical*	*Text (VarChar)*	*47*	*970992*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*CONTRIBUTING FACTOR VEHICLE 4*	*Categorical*	*Text (VarChar)*	*33*	*1030529*	*Poor - High level of missing values*
	Motor_Vehicle_Collisions	*CONTRIBUTING FACTOR VEHICLE 5*	*Categorical*	*Text (VarChar)*	*17*	*1043539*	*Poor - High level of missing values*

	Weather_2016-2019	*Location Name*	*Categorical*	*Text(varchar)*	1	0	*Good - No missing values*
	Weather_2016-2019	*Tempmax*	*Quantitative*	*Numerical(Decimal)*	719	0	*Good - No missing values*
	Weather_2016-2019	*Tempmin*	*Quantitative*	*Numerical(Decimal)*	645	0	*Good - No missing values*
	Weather_2016-2019	*WindSpeed*	*Quantitative*	*Numerical(Decimal)*	286	0	*Good - No missing values*
	Weather_2016-2019	*Humidity*	*Quantitative*	*Numerical(Decimal)*	531	0	*Good - No missing values*
	Weather_2016-2019	*Visibility*	*Quantitative*	*Numerical(Decimal)*	113	0	*Good - No missing values*
	Weather_2016-2019	*Cloud Coverage*	*Quantitative*	*Numerical(Decimal)*	751	0	*Good - No missing values*
	Weather_2016-2019	*Weather Description*	*Categorical*	*Text(varchar)*	79	0	*Good - No missing values*

General Conclusions:

The most challenging part for me was to select the suitable databases which I can easily work on. Complicated dataset selected :) From my research, I faced with the not good quality data's like: data with a lot of null values, incomplete or inconsistent entries or disparate formats of datasets. Additionally, most datasets were lacked proper documentation or context which makes it difficult to understand its variables and their relationships.

Based on the comprehensive analysis of motor vehicle collision data and weather conditions in New York City boroughs from 2016 to 2019, I'm intent to present how the weather conditions has a big effect on how often and how serious car crashes are. By this analyze, we reveal a potential correlation between weather conditions and collisions. it might be useful to potential users, such as New York Department of Transportation and Insurance companies may get benefit as a chance to *focus on making roads safer and improving infrastructure and understanding risk for getting insurance rates. However, since I have a missing data about accidents and although it’s a bit huge data, it’s still lack of information’s to get better idea of how the weather conditions affects to the accidents. To be aimed to better understanding by this analyze, we need more data from more places.*