Potential crime influencing factors analysis

1. Project idea presentation

Problem domain:

We aim to explore how various factors influence criminal tendencies and behaviour. This problem is multidimensional, as crime rates can depend on numerous factors such as premise, crime type, date, victim descent, victim sex, and time of the crime. In this report, we will focus on visualizing and analysing crime data through various graphs to uncover potential patterns and correlations. Collectively, these graphs will help us understand the complex relationships between these factors and crime, offering valuable insights for law enforcement, policy-makers, and researchers.

Assumptions:

To get good data we need to focus on specific area rather than global parameters. We've chosen the city of Los Angeles, Southern California.

Purpose:

To look whether there's correlation between time of the week / year, where number of crimes is changing, (or type of specific crime).

Scope:

The city of Los Angeles.

Requirements:

Provide results and insights about how particular conditions is or is not correlated with specific types of crimes.

In particular we want to answer those questions:

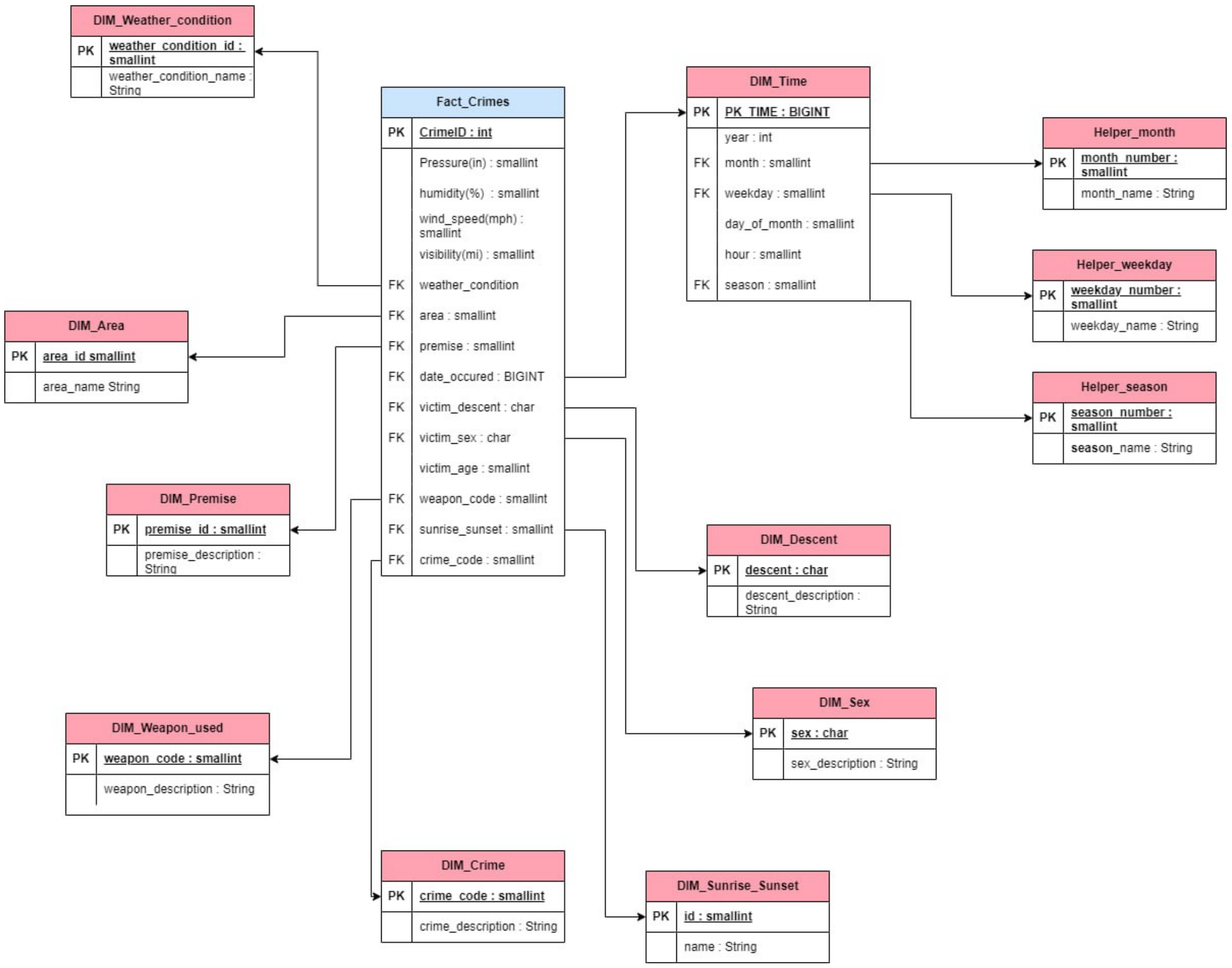
- Does the humidity have an influence on the number of crimes?
- Does the wind speed have an influence on the number of crimes?
- Does the atmospheric pressure have an influence on the number of crimes?
- Does the visibility have an influence on the number of crimes?

Data Domain Dictionary:

- pressure – atmospheric pressure, in inches of water (in.)
- humidity – air humidity, in percentage
- wind speed – average wind speed, measured in miles per hour
- visibility – visiblity measured by miles of clear air
- weather condition – word description of the general weather condition, e.g.:(Fair, Cloudy, Haze, Mostly Cloudy, Fog, Light Rain, Duststorm, Thunder, etc.)

- area – area of the city where the crime occurred, e.g.:(Hollywood, Southwest, 77th Street, Central, Harbour, etc.)
- premise – specific type of location of the crime e.g.: (Street, Sidewalk, Bank, Parking, Bus Stop, Beach, etc.)
- date occurred – date when the crime occurred (year, season, month, day of the week)
- victim descent – descent of the victim, e.g.: (Hispanic, White, Black, Chinese, Other, etc.)
- victim age – age of the victim (years)
- weapon code – type of weapon, e.g.: (Hand gun, Stick, Strong Arm, Axe, Glass, etc.)
- sunrise_sunset – Day or night, specifies time of day when the crime happened
- crime code – type of crime, e.g.: (Simple assault, Burglary, Vandalism, Child abuse, etc.)

Data schema:



2. Data source preparation and assessment

Below are described original data columns from both data sources, the yellow ones are the ones we used in our study.

We used 2 data sources (both .csv files):

1. Crimes - from Los Angeles crimes data: https://www.kaggle.com/datasets/venkatsairo4899/los-angeles-crime-data-2020-2023

Column	Description	Data Type
DR_NO	Case ID	calendar_today
Date Rptd	Reported Date	calendar_today
DATE OCC	Date Occurred	text_format
TIME OCC	Time occurred (format in hhmm)	text_format
AREA	Area of Crime	text_format
AREA NAME	Name of the Area	text_format
Rpt Dist No	District Number of Incident	text_format
Part 1-2	Part 1 or Part 2 Crime Indicator	text_format
Crm Cd	Police Code for Crime	text_format
Crm Cd Desc	Crime Description	text_format
Mocodes	Modus Operandi (MO) Codes	text_format
Vict Age	Victim Age	text_format
Vict Sex	Victim Sex	text_format
Vict Descent	Victim Descent/Ethnicity	text_format
Premis Cd	Premise Code	text_format
Premis Desc	Premise Description	text_format
Weapon Used Cd	Weapon Used Code	text_format
Weapon Desc	Weapon Description	text_format
Status	Status of the Case	text_format
Status Desc	Status Description	text_format
Crm Cd 1	Additional Crime Code 1	text_format
Crm Cd 2	Additional Crime Code 2	text_format
Crm Cd 3	Additional Crime Code 3	text_format
Crm Cd 4	Additional Crime Code 4	text_format
LOCATION	Location of the Incident	text_format
Cross Street	Nearest Cross Street to the Incident	navigation
LAT	Latitude of the Incident Location	navigation
LON	Longitude of the Incident Location	navigation

2. Weather – from US countrywide car accidents https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Column	Description	Data Type
ID	This is a unique identifier of the accident record.	text_format
Source	Source of raw accident data.	text_format
Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	calendar_today
Start_Time	Shows start time of the accident in local time zone.	calendar_today
End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	text_format
Start_Lat	Shows latitude in GPS coordinate of the start point.	text_format
Start_Lng	Shows longitude in GPS coordinate of the start point.	text_format
End_Lat	Shows latitude in GPS coordinate of the end point.	text_format

End_Lng	Shows longitude in GPS coordinate of the end point.	text_format
Distance(mi)	The length of the road extent affected by the accident in miles.	text_format
Description	Shows a human-provided description of the accident.	text_format
Street	Shows the street name in the address field.	text_format
City	Shows the city in the address field.	text_format
County	Shows the county in the address field.	text_format
State	Shows the state in the address field.	text_format
Zipcode	Shows the zipcode in the address field.	Flag
Country	Shows the country in the address field.	text_format
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	text_format
Airport_Code	Denotes an airport-based weather station which is the closest one to the location of the accident.	calendar_today
Weather_Timestamp	Shows the timestamp of the weather observation record (in local time).	text_format
Temperature(F)	Shows the temperature (in Fahrenheit).	text_format
Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	text_format

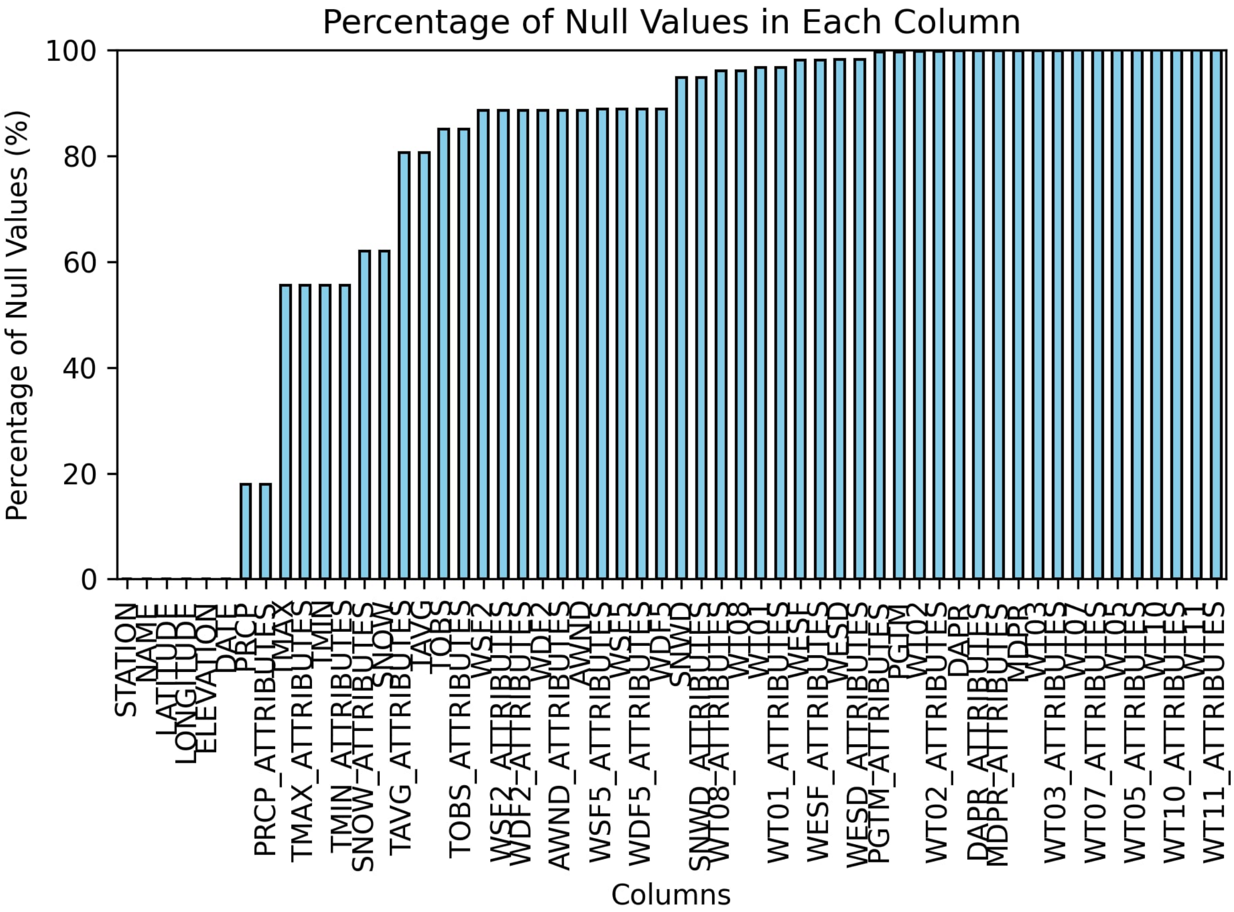
Humidity(%)	Shows the humidity (in percentage).	text_format
Pressure(in)	Shows the air pressure (in inches).	text_format
Visibility(mi)	Shows visibility (in miles).	text_format
Wind_Direction	Shows wind direction.	text_format
Wind_Speed(mph)	Shows wind speed (in miles per hour).	text_format
Precipitation(in)	Shows precipitation amount in inches, if there is any.	text_format
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.).	check
Amenity	A POI annotation which indicates presence of an amenity in a nearby location.	check
Bump	A POI annotation which indicates presence of a speed bump or hump in a nearby location.	check
Crossing	A POI annotation which indicates presence of a crossing in a nearby location.	check
Give_Way	A POI annotation which indicates presence of a give way in a nearby location.	check

Junction	A POI annotation which indicates presence of a junction in a nearby location.	check
No_Exit	A POI annotation which indicates presence of a no exit in a nearby location.	check
Railway	A POI annotation which indicates presence of a railway in a nearby location.	check
Roundabout	A POI annotation which indicates presence of a roundabout in a nearby location.	check
Station	A POI annotation which indicates presence of a station in a nearby location.	check
Stop	A POI annotation which indicates presence of a stop in a nearby location.	check
Traffic_Calming	A POI annotation which indicates presence of traffic calming in a nearby location.	check

Traffic_Signal	A POI annotation which indicates presence of a traffic signal in a nearby location.	check
Turning_Loop	A POI annotation which indicates presence of a turning loop in a nearby location.	text_format
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	text_format
Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	text_format
Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	text_format
Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	text_format

Analysis and problems:

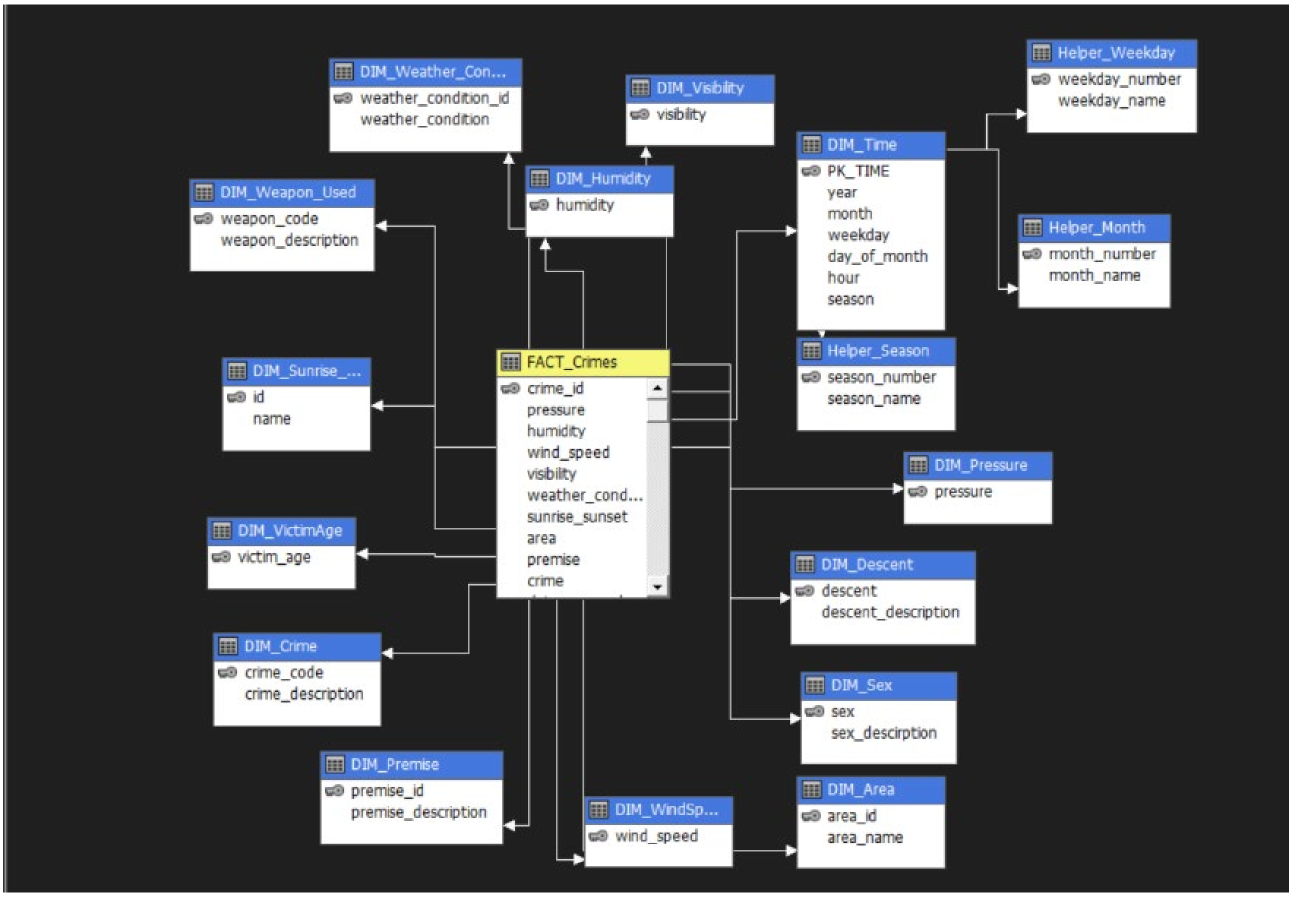
It took us a lot of time to find proper data source for daily weather data in that period of time 2020-2023, because a lot of good meteorological datasets were not free. On the other hand in datasets we found, almost all interesting values were empty, below is example of one of datasets which covered Los Angeles on our target period of time. But when check closer it contained a lot of missing values:



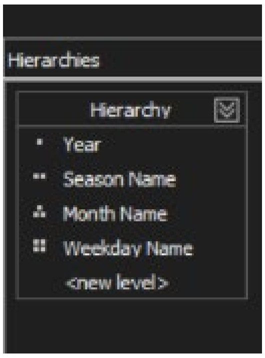
In the end we decided to use Cat accidents data, as it also contains good and precise weather descriptions for each car accident. Then we filtered the car accidents data to the city of Los Angeles, and surprisingly we got really got coverage (day and hour of the crime matching car accident). From the original ~800k columns with crimes, we were able to get the weather data from closest car accident (time and space wise, because sometimes we had multiple car accidents at the same time of crime but in different locations in the city), and in the end having 687k records were we could find the weather data matching the crime.

Due to such process we lost some data, but also due to our interpolation of weather for each crime, the data is not as precise as it could be, which could result in some small additional margin of error which will be taken into account in the analysis part.

3. Multidimensional model



Hierarchies:

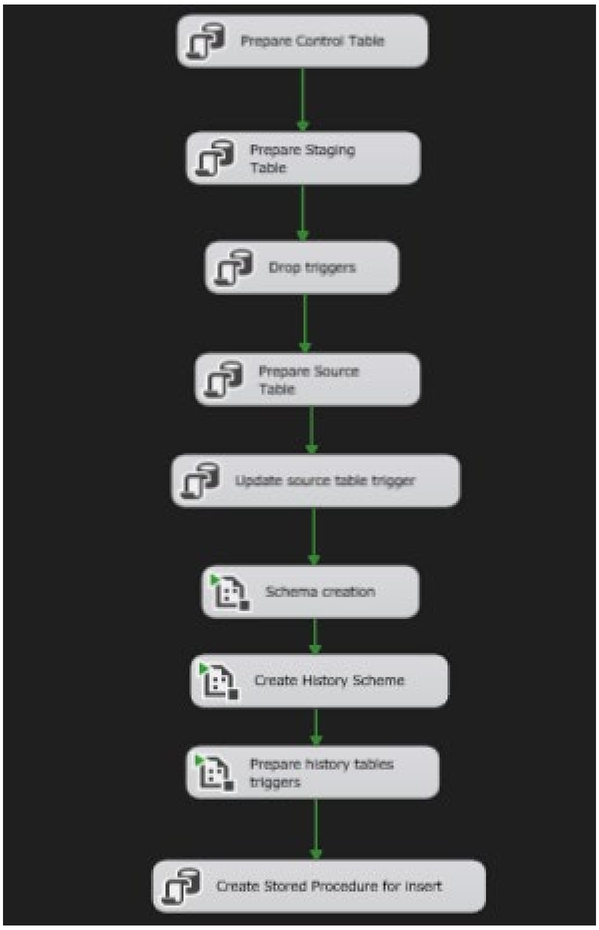


- Year
- Season Name
- Month Name
- Weekday Name
- <new level>

4. The ETL process

1. Preparation of Database Schema

The ETL process begins with the preparation of the database schema. This step ensures that the structure of the data warehouse is defined and optimized to accommodate the incoming datasets. Schemas, Source Table, Processing Table, database schema, and History Tables are created along with the necessary triggers.



(DB setup.dtsx)

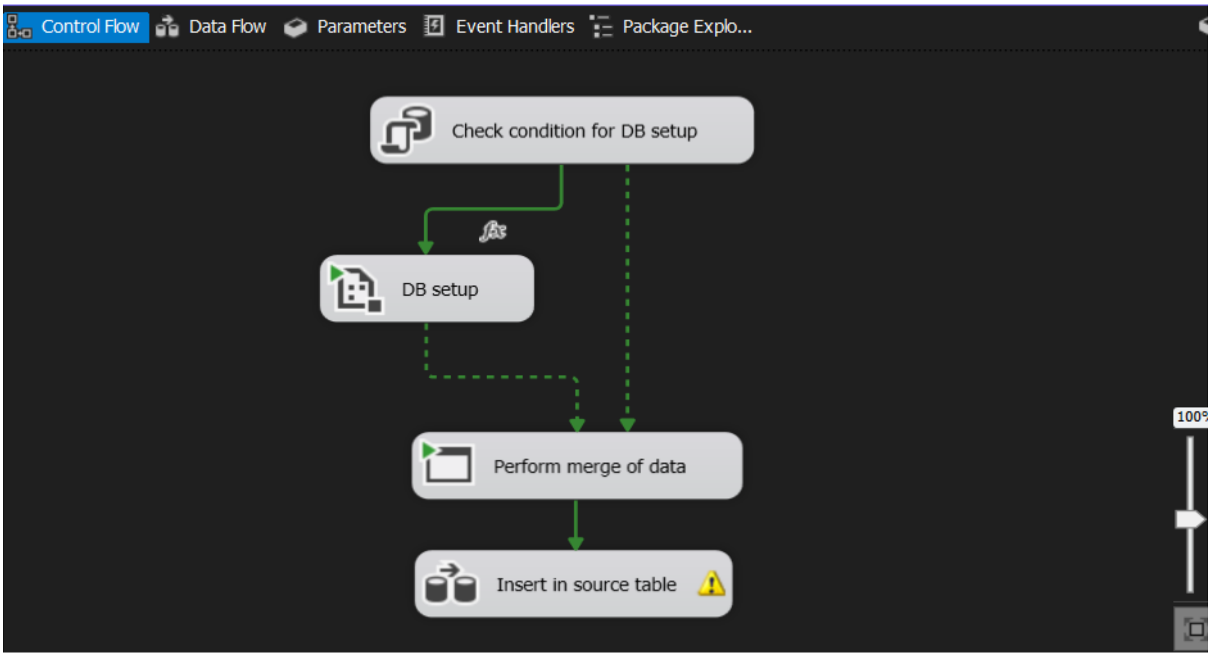
2. Data Merging

A Python script is executed in this step, which is responsible for merging two datasets into a single cohesive dataset. This merged dataset is then prepared for loading into the source table. **(main.py file)**

3. Data Loading into Source Table

Initially, data is loaded into the source table. This step supports incremental loading to handle new and updated records efficiently. A dedicated package manages this process, ensuring that only the new data is ingested during each run.

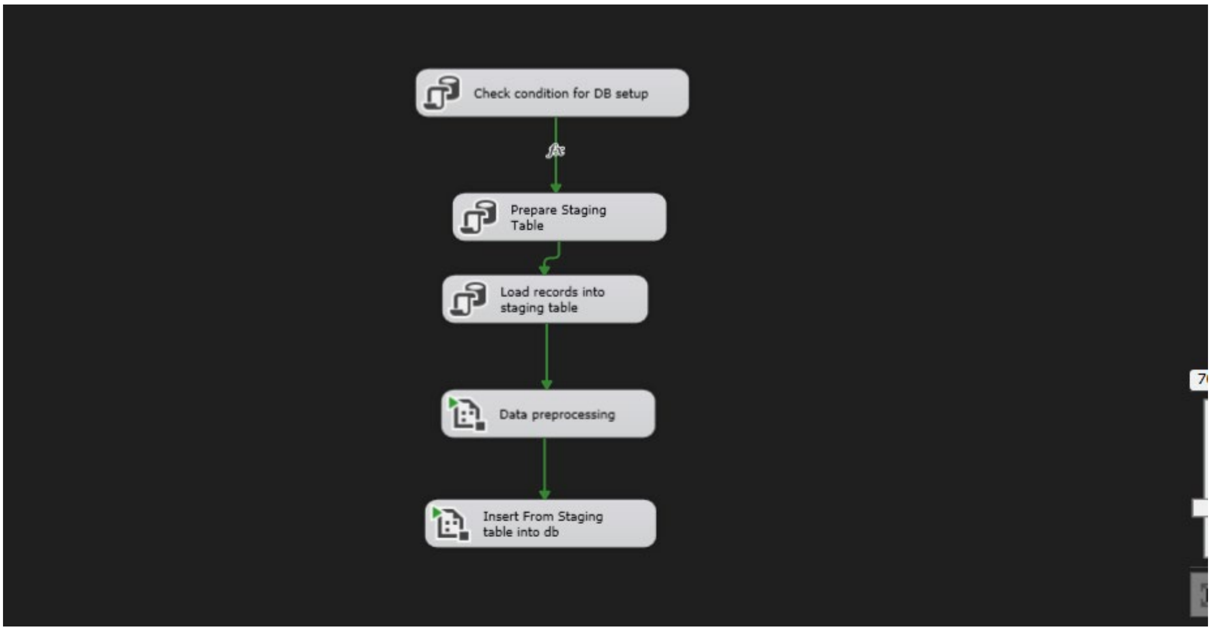
(Insert new records in source table.dtsx)



4. Staging Process

When the data is ready to be loaded into the data warehouse schema, another package is executed. This package identifies and processes the newly added and modified records since the last execution. The process involves the following steps:

- **Control Table and Triggers:** A control table and triggers on the source table track the last update time of each record. This mechanism ensures that only records with recent changes are considered.
- **Staging Table:** The identified records are then loaded into the StagingTable. This staging area serves as temporary storage for data that will undergo further processing before being moved to the final schema.


Check condition for DB setup
Prepare Staging Table
Load records into staging table
Data preprocessing
Insert From Staging table into db

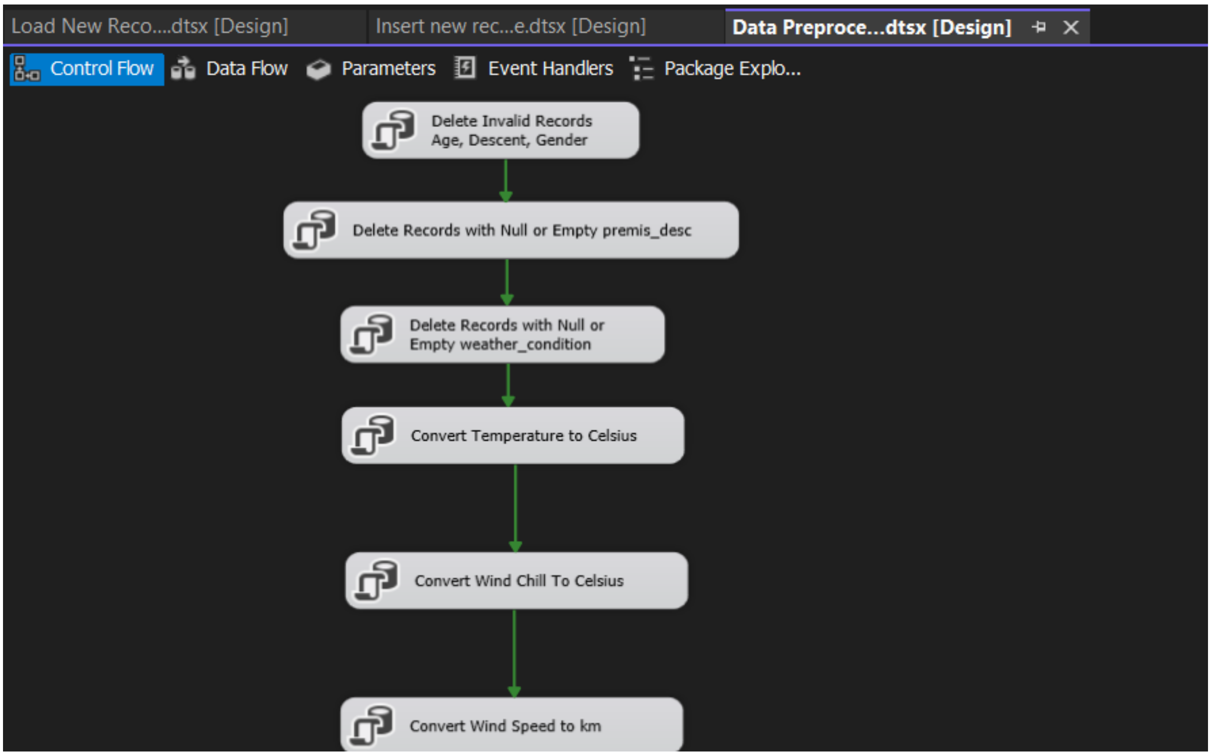
(Load New Records from source.dtsx)

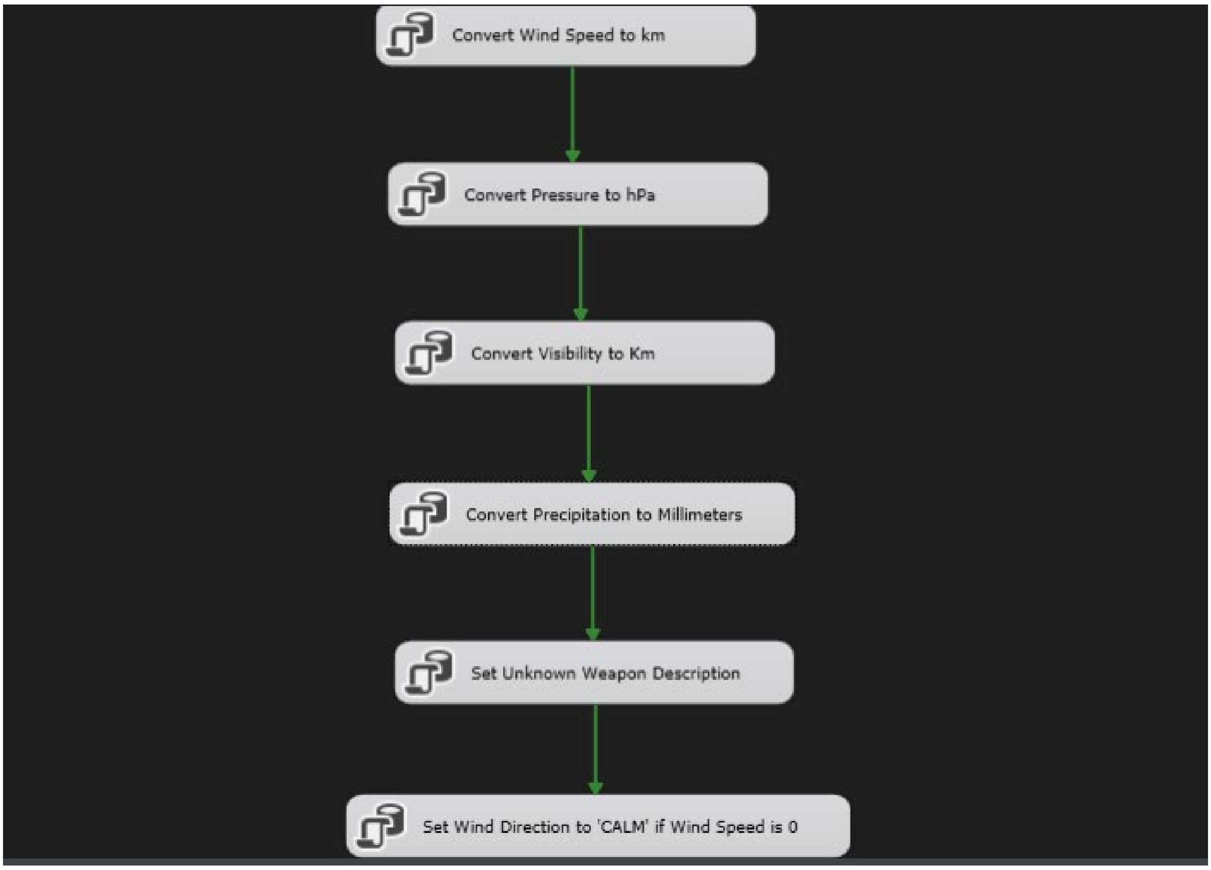
5. Data Cleaning and Transformation

Once the data is in the StagingTable, it undergoes a series of cleaning and transformation steps:

- **Handling Missing and Null Values:** Missing and null values are addressed to ensure data completeness.
- **Dropping Incorrect Records:** Records that do not meet predefined criteria are removed.
- **Units Transformation:** Data units are standardized to ensure consistency.

(Data Preprocessing.dtsx)


Load New Reco....dtsx [Design]
Insert new rec...e.dtsx [Design]
Data Preproce...dtsx [Design]
Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Delete Invalid Records Age, Descent, Gender
Delete Records with Null or Empty premis_desc
Delete Records with Null or Empty weather_condition
Convert Temperature to Celsius
Convert Wind Chill To Celsius
Convert Wind Speed to km



Convert Wind Speed to km

Convert Pressure to hPa

Convert Visibility to Km

Convert Precipitation to Millimeters

Set Unknown Weapon Description

Set Wind Direction to 'CALM' if Wind Speed is 0

6. Loading Data into the Schema

After the data is cleaned and transformed, it is loaded from the StagingTable into the final schema of the data warehouse. This step integrates the processed data into the main analytical framework.

7. Control Table Update

Post-loading, the control table is updated to reflect the latest data load timestamp. This update prevents the reprocessing of the same records in subsequent ETL runs, ensuring efficiency and data integrity.

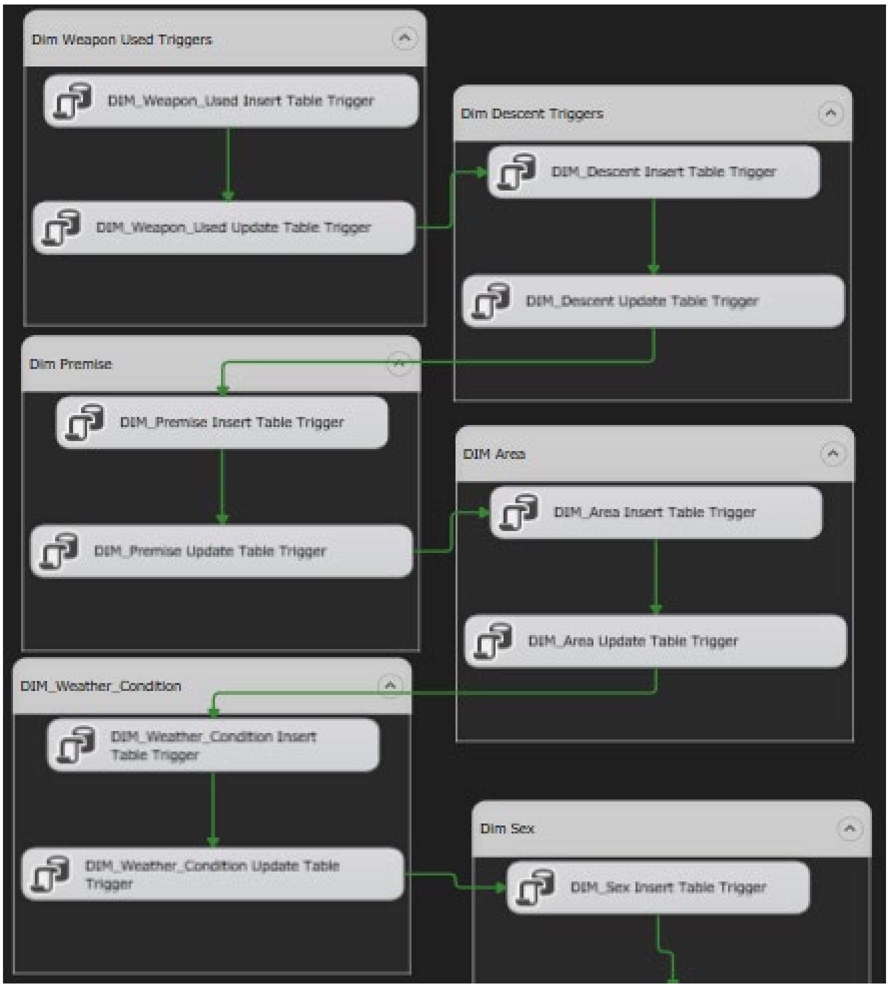
8. Implementation of Slowly Changing Dimensions (SCD)

The data warehouse schema incorporates Slowly Changing Dimensions (SCD) Type 4 to maintain a comprehensive history of changes:

- **SCD Type 4:** This type was chosen to preserve the historical data without influencing the current data analysis results, a challenge often associated with SCD Type 2.

(Prepare History Tables Triggers.dtsx)

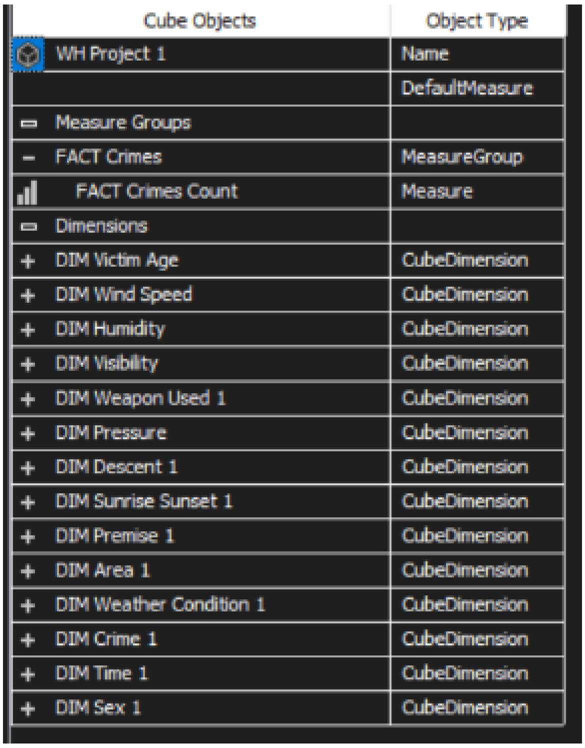
(History Tables Creation.dtsx)


Dim Weapon Used Triggers
DIM_Weapon_Used Insert Table Trigger
DIM_Weapon_Used Update Table Trigger
Dim Descent Triggers
DIM_Descent Insert Table Trigger
DIM_Descent Update Table Trigger
Dim Premise
DIM_Premise Insert Table Trigger
DIM_Premise Update Table Trigger
DIM Area
DIM_Area Insert Table Trigger
DIM_Area Update Table Trigger
DIM_Weather_Condition
DIM_Weather_Condition Insert Table Trigger
DIM_Weather_Condition Update Table Trigger
Dim Sex
DIM_Sex Insert Table Trigger

P.S The entire ETL process structure is encapsulated within a Visual Studio project,

5. Preparation of the cube

Thanks to the way the database schema was designed we didn’t need to change or add any more relations in the cube itself, which would be the case if the database was to be used for production and more uses than only analytical. But as in our case we only needed the analytical use, we could create the schema accordingly and simply the process later on.
Thanks to that we also didn’t need to create any calculated measures, or aggregations, as already had the data in the form that we wanted.



Cube Objects	Object Type
WH Project 1	Name
	DefaultMeasure
− Measure Groups	
− FACT Crimes	MeasureGroup
FACT Crimes Count	Measure
− Dimensions	
+ DIM Victim Age	CubeDimension
+ DIM Wind Speed	CubeDimension
+ DIM Humidity	CubeDimension
+ DIM Visibility	CubeDimension
+ DIM Weapon Used 1	CubeDimension
+ DIM Pressure	CubeDimension
+ DIM Descent 1	CubeDimension
+ DIM Sunrise Sunset 1	CubeDimension
+ DIM Premise 1	CubeDimension
+ DIM Area 1	CubeDimension
+ DIM Weather Condition 1	CubeDimension
+ DIM Crime 1	CubeDimension
+ DIM Time 1	CubeDimension
+ DIM Sex 1	CubeDimension

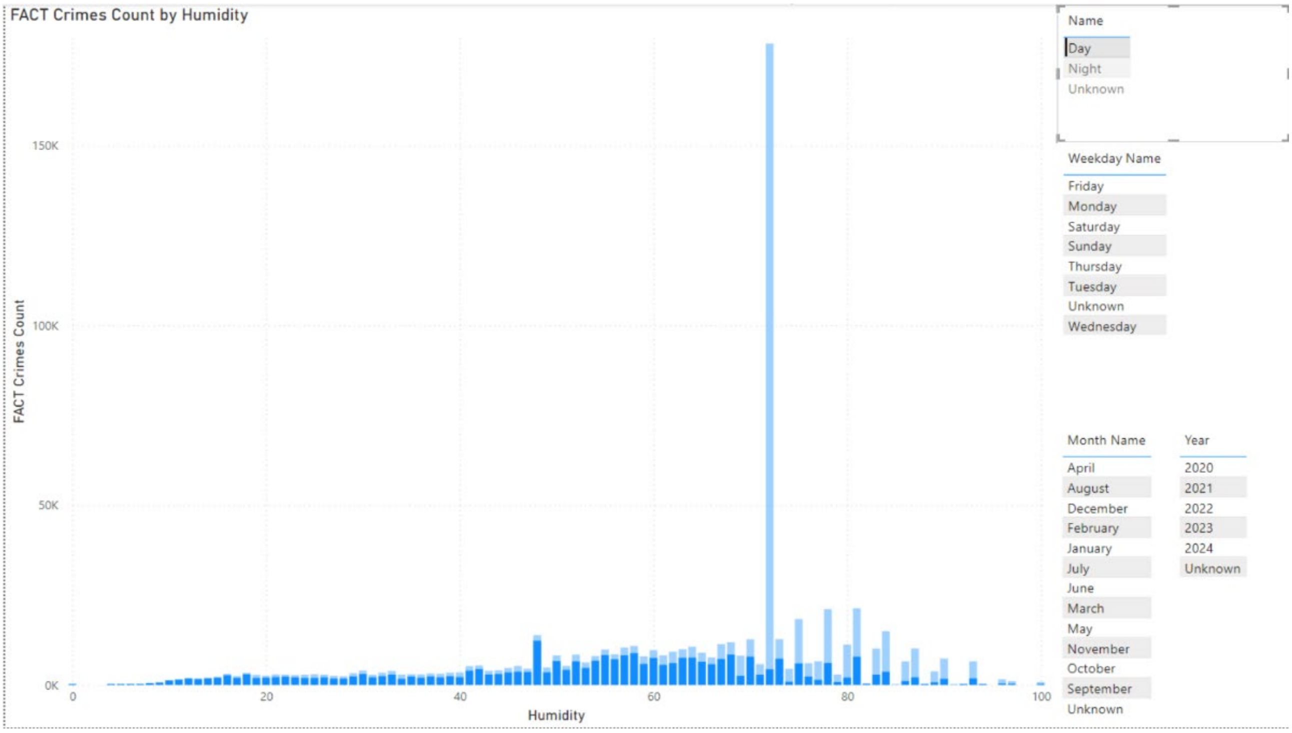
6. Consumption

Humidity on crime:

- Number of crimes by the humidity

Can be filtered by year, month, day of the week

Here we checked if there's some correlation with pressure and number of crimes. While we didn't find it, we discovered potential discrepancy in the dataset (or the way the officers fill in the reports), where humidity was not measured properly during the night, and given default value 72, during some humid period of time, or there was really in fact very specific and constant level of humidity for a long time, which isn't impossible as the average humidity in Los Angeles is 65%, but it still points to bad reporting due to the fact that it only happened for night-time reports.

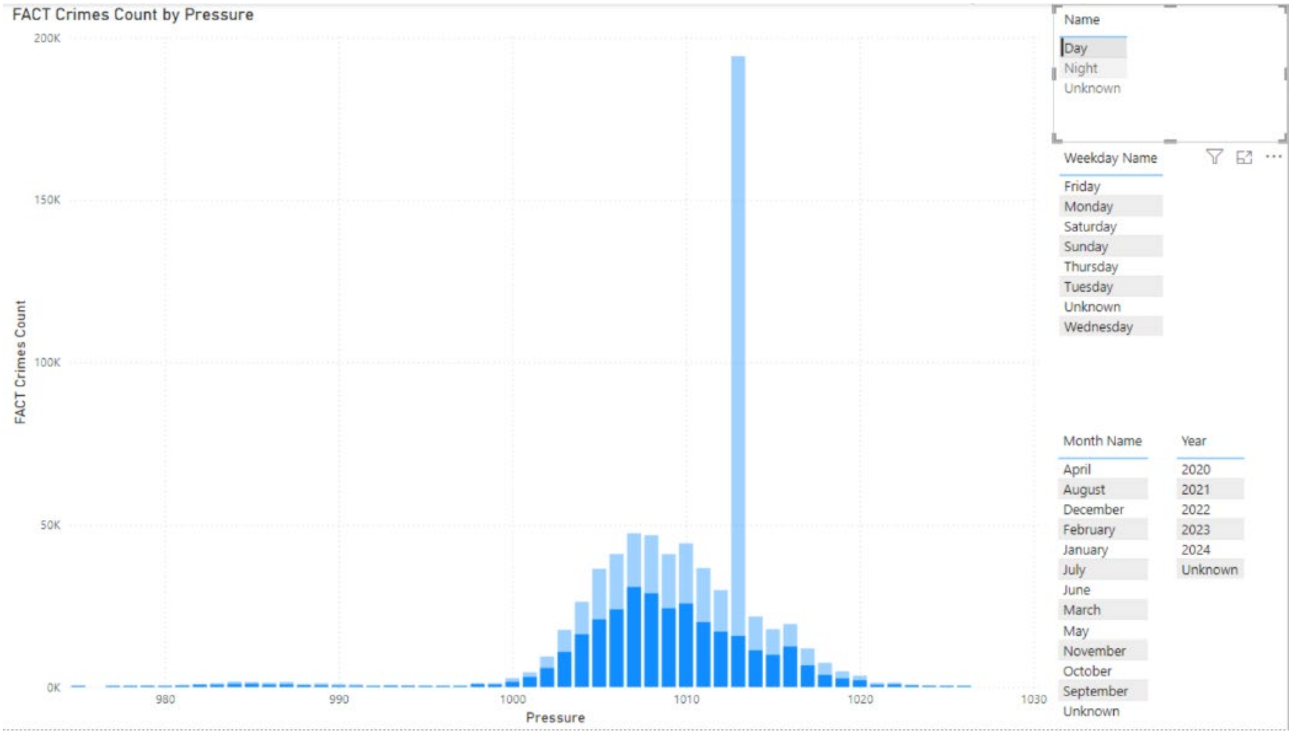


Crimes by atmospheric pressure:

- Number of crimes

With option to filter by day/night, year, month, and day of the week.

Here we checked if there's some correlation with pressure and number of crimes. While we didn't find it, we discovered potential discrepancy in the dataset (or the way the officers fill in the reports), where pressure during night was possibly not check properly and taken default 1013 hPa.

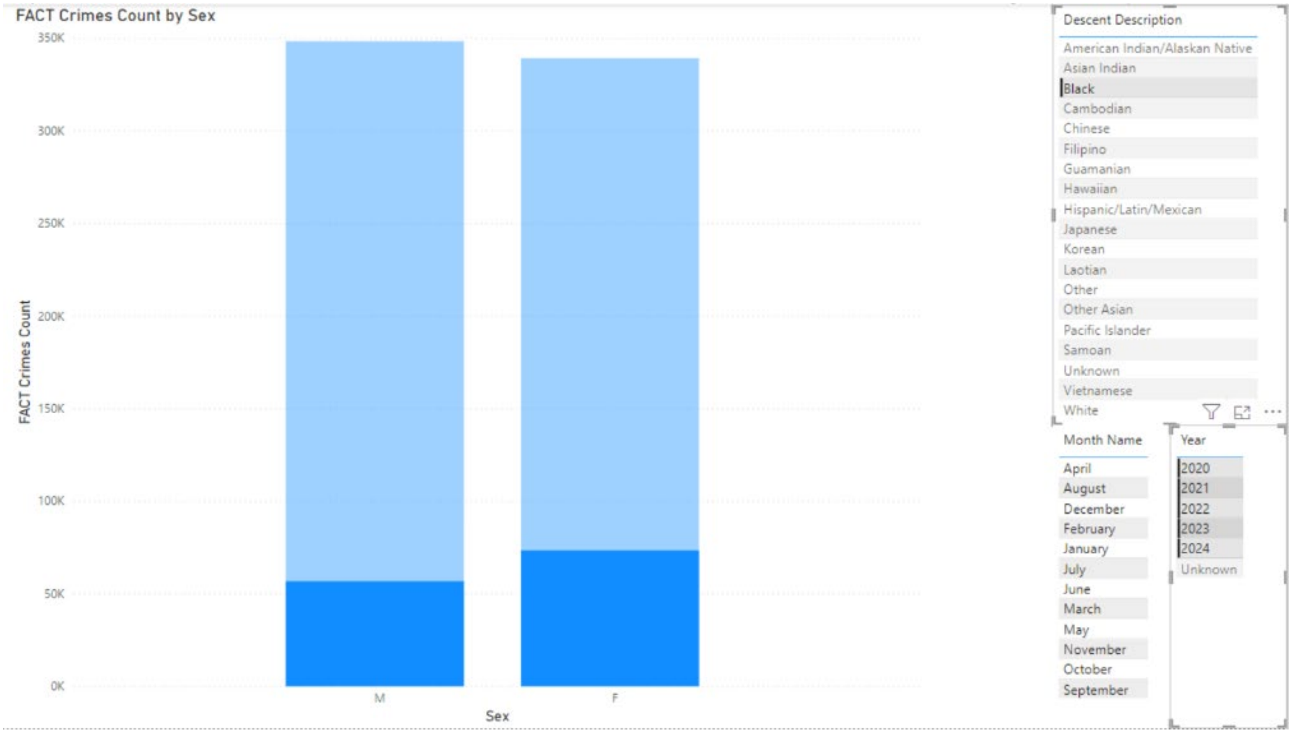


Victim sex by ethnicity:

- victim sex by the ethnicity and time

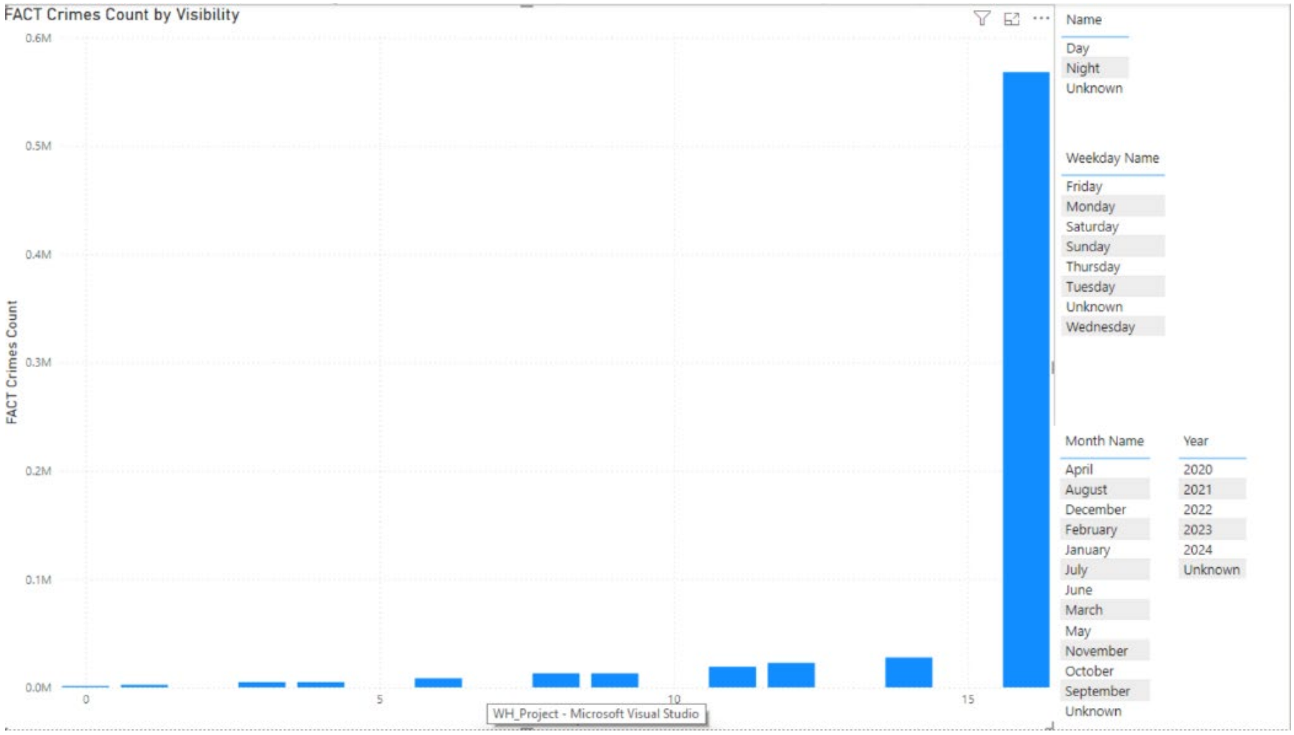
With option to filter by ethnicity, year and month.

Here we can analyse and keep track of any disparities on racial/gender prejudice victims.



Crimes by visibility

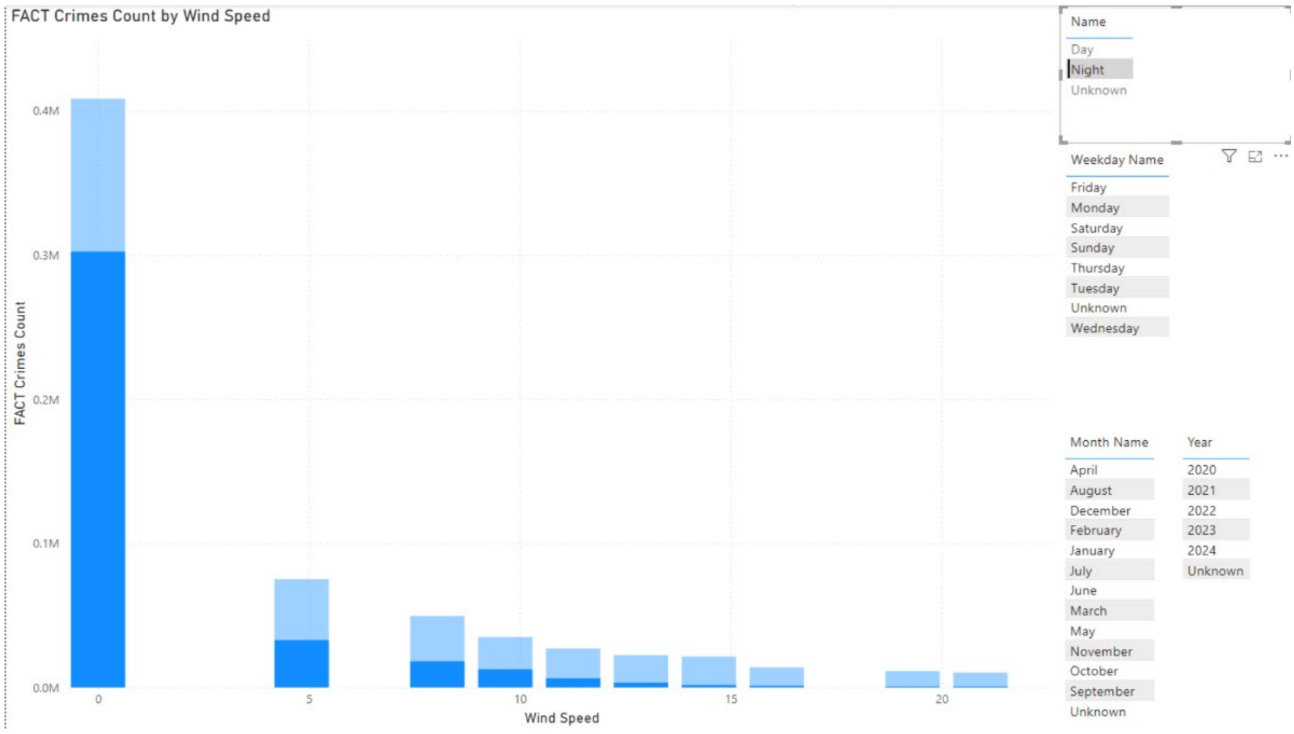
- Number of crimes by visibilitHere we checked if there’s some correlation with visibility and number of crimes. And we could assume that most of the crimes happened when there’s high visibility. Which could actually be true, as general visibility in Los Angeles seems to be about 10miles.



Crimes by wind speed

- Number of crimes by wind speed

Here we check another weather measure if the wind speed affects the number of crimes. We can see that crimes at night happen when the wind is generally weaker. But at the same time we need to remember that the wind is stronger during the day than at night.

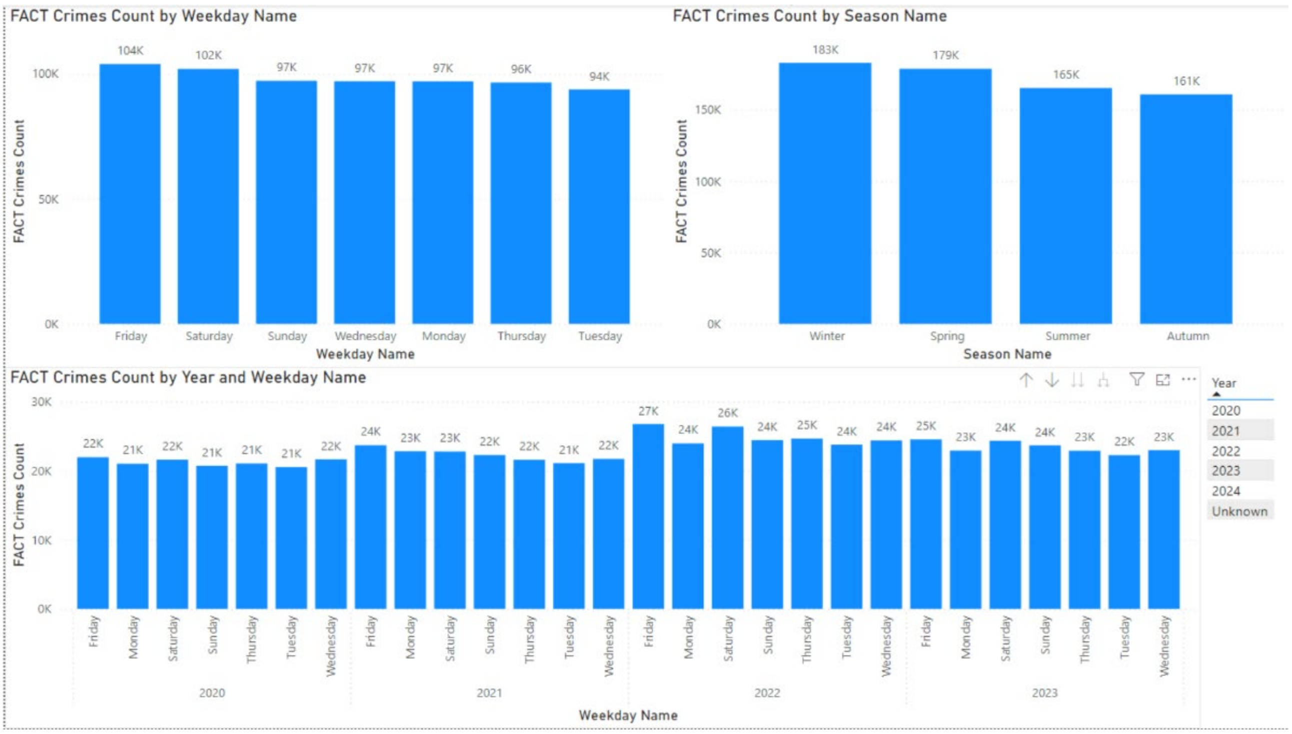


Number of crimes by time:

- based on day of the week
- based on season
- general overview

With option to filter by specific year.

Here we can analyse and keep track if any anomaly may happened regarding general crime count by time of the year, or day of the week.

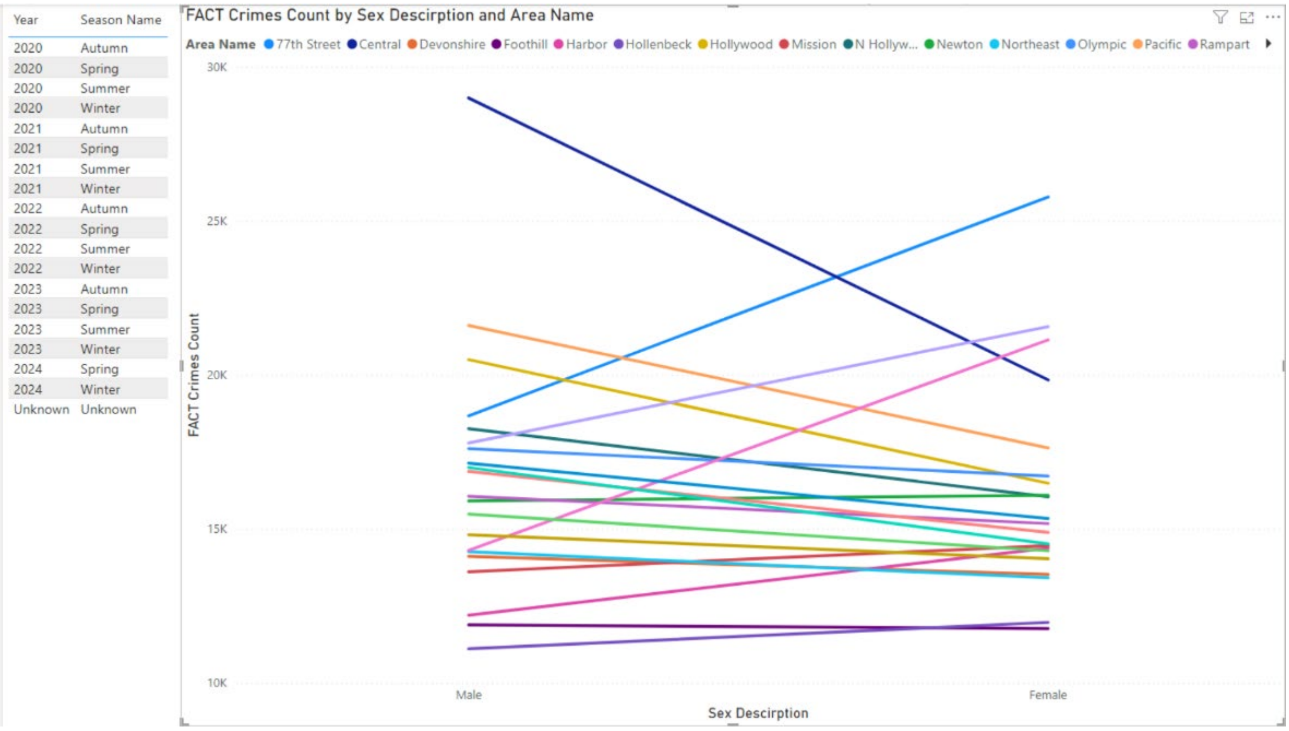


Crimes by victim sex and the crime location:

- Number of crimes committed on males/females on different regions of the city

With option to filter by the year and month.

Here we can see in which parts of the city we can expect specified gender victims more than the others. For example we can look for specific area and check the ratio of male to female victims, which could give us more information about the area itself.



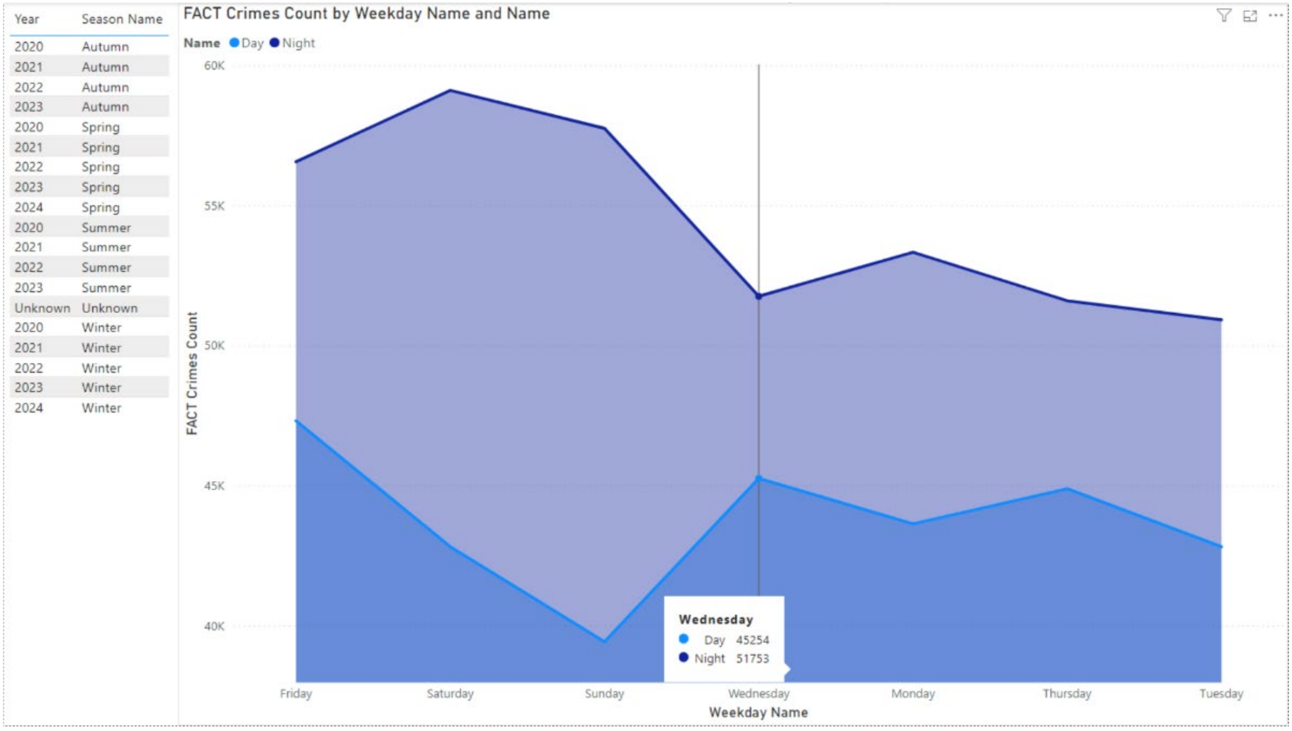
Year	Season Name
2020	Autumn
2020	Spring
2020	Summer
2020	Winter
2021	Autumn
2021	Spring
2021	Summer
2021	Winter
2022	Autumn
2022	Spring
2022	Summer
2022	Winter
2023	Autumn
2023	Spring
2023	Summer
2023	Winter
2024	Spring
2024	Winter
Unknown	Unknown

Crimes by time of day:

- number of crimes by the time of day

With option to filter by year and season

Here we check if there’s any correlation of day of the week and number of crimes. And discovered that Tuesday is in general the most ‘peaceful’ of days, while the weekends seem to be more active, as people have more time.



Year	Season Name
2020	Autumn
2021	Autumn
2022	Autumn
2023	Autumn
2020	Spring
2021	Spring
2022	Spring
2023	Spring
2024	Spring
2020	Summer
2021	Summer
2022	Summer
2023	Summer
Unknown	Unknown
2020	Winter
2021	Winter
2022	Winter
2023	Winter
2024	Winter

7.Conclusion, final remarks

We did found a little bit higher number of crimes during weekends than the rest of the days in general.
As well as that in even though in general most of the crimes were committed by night, in some years number of crimes happening during the day was much larger than anticipated.

Additionally we found that even though men make up for more crime victims, which also stays true for almost all ethnicities, except for black: where women are more often crime victims than men.

The main problem with analysing weather correlation to number of crimes is that, it can be coincidence, as in Los Angeles we have some general weather distribution of humidity/pressure etc.. So even when we see that most crimes happen in visibility is really high (15miles), but that’s the common case for Los Angeles, so there may be not much coincidence.
To solve this uncertainties and ensure conclusive results, we would need to additionally gather weather data on each day for analysed years, to check if the crimes distribution by weather follows weather distribution or there are some differences, which would indicate that for example: with high winds there’s less crimes happening.

Some insights gained during the work:

- During designing the schema we need to anticipate and define what is the MEASURE and what is the DIMENSION we will want to filter the measure by. (we had some issue with this, and needed to redo it, as not all desired filters were defined as dimensions)
- Adding labels to graphs is really useful, to add more specific numeric information (e.g. in column chart)
- It's good to ask another person about the clarity of the graphs, as not everybody has good sense of design, and can make very unreadable graphs without knowing it.

Conclusion:

In the end we couldn't prove that there's any correlation between the weather and the number of crimes in the city of Los Angeles during 2020-2023 years. From our analysis we deduced that the weather didn't play any significant role in crime rates in our specified study domain.

Mateusz Molenda 259905
Volodymyr Shepel 266617