ANALYZING MOTOR VEHICLE COLLISIONS
PROJECT REPORT

ABSTRACT:

- The goal of the project is to leverage integrated crash data from the Department of Transportation portals in three major cities: Austin, Chicago, and New York. This effort aims to investigate the complexities of traffic crash incidents, their patterns, contributing variables, and subsequent effects on public safety, with the primary goal of improving urban traffic safety measures. This collaborative initiative highlights the need for a group effort to address pressing issues related to public safety and urban mobility.

STEPS INVOLVED:

- **Data Analysis and Profiling**: Using the profile tools Alteryx and Ydata, a thorough data profiling exercise is conducted at the beginning of the project. The team's goal is to obtain a deep understanding of the qualities, traits, and possible problems present in the crash data sets through painstaking study.
- **Structured Data Staging and ETL:** The project moves on to data staging using Talend ETL operations after data profiling. By following industry best practices, the team makes sure that the data is consistent and of high quality during the staging phase, which provides a solid basis for further research.

  Dimensional Modeling: Creating a dimensional model that incorporates dimensions and facts is essential to the project's architecture. The group outlines the complex connections between source data columns and the associated destination entities through painstaking documentation and mapping operations, guaranteeing traceability and clarity all the way through the data transformation process.
- **Real-time Problem Solving:** The initiative addresses issues related to accident incidence, severity, contributing variables, and temporal patterns that are common in urban traffic safety in real-time. The team utilizes SQL queries on the dimensional data model to derive meaningful insights that can guide policy decisions and reduce the likelihood of traffic-related incidents.

  Visualization and Insights: Utilizing cutting-edge technologies like Tableau and Power BI, the project moves into the visualization phase in its last phase. Stakeholders can get profound insights into the crash data analysis process through visually appealing representations. This allows for proactive interventions to support urban traffic safety measures and informed decision-making processes.

TEAM MEMBERS- GROUP 17:

- Anisha Gandhi
- Anushka Paradkar
- Dhir Thacker
- Lakshmi Kumar

DELIVERABLES:

Part 1:

1. Data profiling Alteryx / Y-data profile

› Analysis document
› Data staging (Staging tables). Use Talend for ETL jobs
› For database connections, use Azure SQL server / MySQL / SQL Server

2. Dimensional model (Target tables)

› Facts and Dimensions to be created
› Create mapping document
› Clearly explain the source column name and where it finally maps to target column
› Stage to Target
› Document all transformations if any

Note:

• Must configure at least one dimension as SCD2
• Address null values appropriately
• Maintain Source DIM table and audit columns wherever applicable

Part 2:

1. Staging to Integration

› Using Talend ETL jobs
› Query dimensional to validate data
› If any rows rejected clearly explain the reason for rejection
› Query dimensional data model using SQL for the provided business questions

Part 3:

1. Visualizations

› Tableau and Power BI
› Upload all screen shots
› Upload source workbooks

PART-1

DATA PROFILING AND GENERAL DOCUMENTATION OF DATASETS

➔ **AUSTIN:**

There are 54 columns in this dataset.

Column Name	**Description**	**Type**
crash_id	TxDOT C.R.I.S. system-generated unique identifying number for a crash	Number
crash_fatal_fl	Fatal Crash Identifier - Indicates that the crash involved one or more fatalities	Plain Text
crash_date	Crash Date	Date & Time
crash_time	Crash Time - Time crash occurred	Plain Text
case_id	Case ID	Plain Text
rpt_latitude	Reported Latitude	Number
rpt_longitude	Reported Longitude	Number
rpt_block_num	Reported Block Number (road on which crash occurred)	Plain Text
rpt_street_pfx	Reported Street Prefix (road on which crash occurred)	Plain Text
rpt_street_name	Reported Street Name (road on which crash occurred)	Plain Text
rpt_street_sfx	Reported Street Suffix (road on which crash occurred)	Plain Text
crash_speed_limit	Speed Limit	Number
road_constr_zone_fl	Construction Zone - Indicates whether the crash occurred in or was related to a construction, maintenance, or utility work zone, regardless of whether workers were present at the time of the crash	Plain Text
latitude	Derived Latitude map coordinate of the crash	Number
longitude	Derived Longitude map coordinate of the crash	Number
street_name	Derived Street Name - Name of the road crash occurred on, as determined by the Locator application.	Plain Text
street_nbr	Derived Street Number - Block number of primary street where crash occurred as determined by the Locator application	Plain Text
street_name_2	Derived Street Name 2 - The road name for the secondary road related to the crash location (If applicable)	Plain Text
street_nbr_2	Derived Street Number 2 - Block number of secondary street related to the crash location as determined by the Locator application (If applicable)	Plain Text
crash_sev_id	Crash Severity - Most severe injury suffered by any one person involved in the crash (0=UNKNOWN, 1=INCAPACITATING INJURY, 2=NON-INCAPACITATING INJURY, 3=POSSIBLE INJURY, 4=KILLED, 5=NOT INJURED)	Number
sus_serious_injry_cnt	Total Suspected Serious Injury Count	Number
nonincap_injry_cnt	Total Non-incapacitating Injury Count	Number
poss_injry_cnt	Total Possible Injury Count	Number
non_injry_cnt	Total Not Injured Count	Number
unkn_injry_cnt	Total Unknown Injury Count	Number
tot_injry_cnt	Total Injury Count	Number
death_cnt	Total Death Count	Number
contrib_factr_p1_id	The first factor for a given vehicle which the officer felt possibly contributed to the crash	Plain Text
contrib_factr_p2_id	The second factor for a given vehicle which the officer felt possibly contributed to the crash	Plain Text
units_involved	Mode of units involved in crash	Plain Text
atd_mode_category_metadata	Description of units involved in crash	Plain Text
pedestrian_fl	Pedestrian involved crash flag	Plain Text
motor_vehicle_fl	Motor vehicle involved crash flag	Plain Text
motorcycle_fl	Motorcycle involved crash flag	Plain Text
bicycle_fl	Bicyclist involved crash flag	Plain Text
other_fl	Other involved crash flag	Plain Text
point	Point datatype created with crash latitude and longitude to enable request of GeoJSON.	Point
apd_confirmed_fatality	APD Fatality flag	Plain Text
apd_confirmed_death_count	APD Fatality Count	Number

motor_vehicle_death_ count		Number
motor_vehicle_serious_ injury_count		Number
bicycle_death_count		Number
bicycle_serious_injury_cou nt		Number
pedestrian_death_count		Number
pedestrian_serious_ injurycount		Number
motorcycle_death_ count		Number
motorcycle_serious_ injury_count		Number
other_death_count		Number
other_serious_injury_ count		Number
onsys_fl	Flag indicates whether primary road of crash was on the TxDOT highway system.	Plain Text
private_dr_fl	Flag indicating whether crash occurred on a private drive or road/private property/parking lot.	Plain Text
micromobility_serious_ injury_count		Number
micromobility_death_ count		Number
micromobility_fl		Plain Text

Dataset Statistics:
Number of Variables: **54**
Number of Observations: **147,750**
Missing Cells: **1,725,084**
Missing Cells (%): **21.6%**
Duplicate Rows: **0**
Duplicate Rows (%): **0.0%**
Total Size in Memory: **60.9 MiB**
Average Record Size in Memory: **432.0 B**

Variable Types:
Numeric: **17 |** Boolean: **11 |** Date Time: **2 |** Text: **8 |** Categorical: **15 |** Unsupported: **1**

Missing Values:
The following columns have missing values which have to be handled using appropriate handling methods:

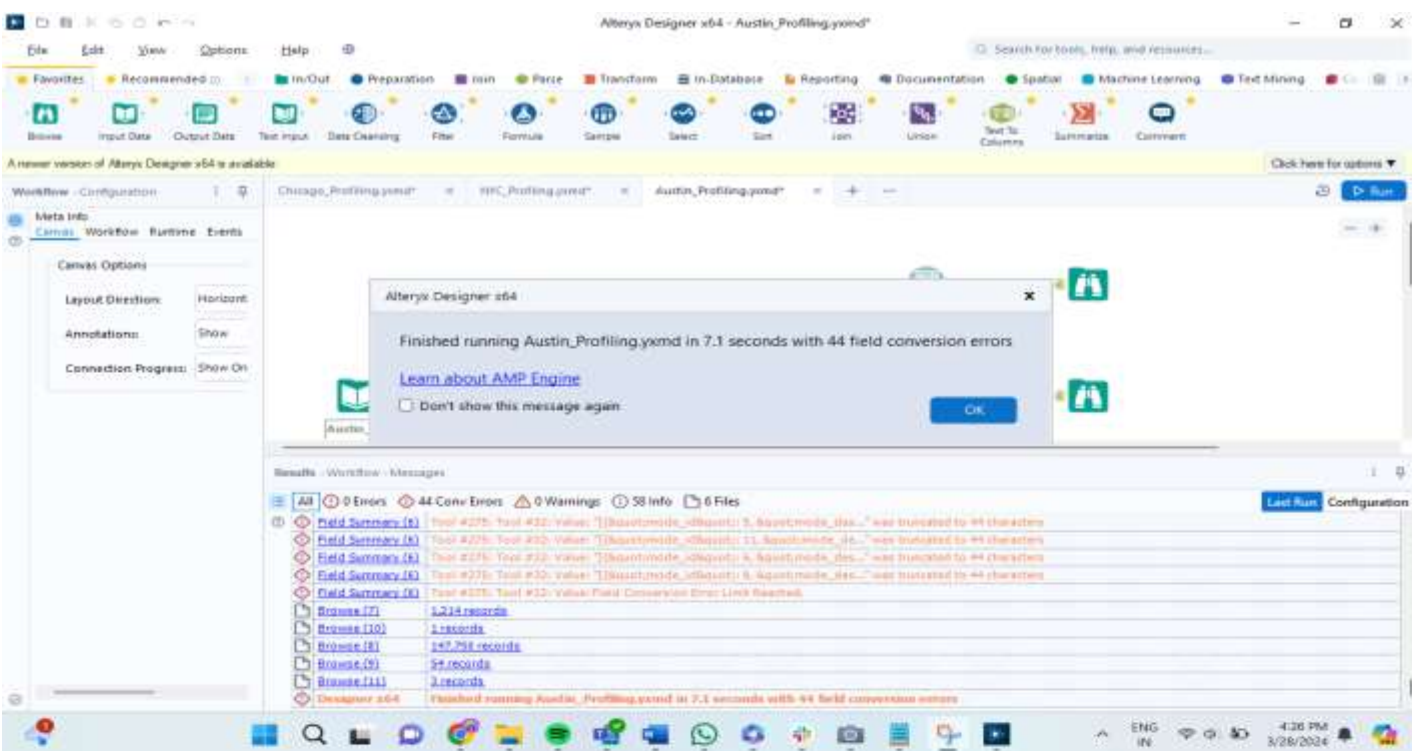
- ➔ case_id has 1858(1.3%)
- ➔ point has 2243 (1.5%)
- ➔ latitude has 2243 (1.5%)
- ➔ longitude has 2243 (1.5%)
- ➔ rpt_block_num has 19611 (13.3%)
- ➔ rpt_street_sfx has 50340 (34.1%)
- ➔ rpt_street_pfx has 67805 (45.9%)
- ➔ street_name_2 has 81474 (55.1%)
- ➔ street_nbr has 87038 (58.9%)
- ➔ contrib_factr_p1_id has 119143 (80.6%)

- ➔ rpt_latitude has 137456 (93.0%)
- ➔ rpt_longitude has 137456 (93.0%)
- ➔ other_fl has 142905 (96.7%)
- ➔ contrib_factr_p2_id has 143235 (96.9%)
- ➔ pedestrian_fl has 144245 (97.6%)
- ➔ motorcycle_fl has 144148 (97.6%)
- ➔ bicycle_fl has 145306 (98.3%)
- ➔ micromobility_fl has 147439 (99.8%)
- ➔ street_nbr_2 has 147750 (100.0%)

Observations:

- ➔ **case_id** has the least number of missing values and **street_nbr_2** has the greatest number of missing values.
- ➔ **street_nbr_2** has no values in any of the rows, it is in an unsupported datatype format. Since it has no values, the column can also be completely dropped.
- ➔ **crash_id** has unique values and is a good fit for primary key.
- ➔ **crash_date** contains both the crash date and time. But there is also a separate **crash_time** column that contains the time of accident. crash_time must be deleted or crash_date must be handled to separate the crash_time from it.
- ➔ The **crash_date** column ranges from 03/26/2014 to 03/11/2024, indicating no discrepancies in the dates being reported after the date of analysis.
- ➔ Most columns have generic datatypes which must be handled. For example, all the flag values are stored as plain text, but the best practice is to convert them as Boolean. The same has to be cross verified and appropriate steps to handle data types must be taken.

Alteryx Workflow with the profiling run statistics:



REPORT:

Numeric Fields:

Name	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
apd_confirmed_death_count	0.0%	5	0.000	0.006	0.000	4.000	0.082	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
micromobility_death_count	0.0%	2	0.000	0.000	0.000	1.000	0.006	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
rpt_latitude	93.0%	7,977	25.837	30.297	30.295	36.500	0.377	This field has over 10% missing values. Consider imputing these values.
sus_serious_ injry_cnt	0.0%	7	0.000	0.034	0.000	10.000	0.207	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
motor_vehicle_serious_injury_count	0.0%	6	0.000	0.023	0.000	5.000	0.176	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
contrib_factr_ p2_id	96.9%	66	1.000	36.347	22.000	79.000	20.583	This field has over 10% missing values. Consider imputing these values.
motorcycle_ death_count	0.0%	3	0.000	0.001	0.000	2.000	0.030	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
motorcycle_ serious_injury_count	0.0%	3	0.000	0.005	0.000	2.000	0.071	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
other_death_ count	0.0%	1	0.000	0.000	0.000	0.000	0.000	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
contrib_factr_ p1_id	80.6%	71	1.000	33.358	20.000	80.000	19.899	This field has over 10% missing values. Consider imputing these values.
pedestrian_ death_count	0.0%	3	0.000	0.002	0.000	2.000	0.047	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
rpt_longitude	93.0%	7,265	-106.646	-97.747	-97.731	-93.508	0.537	This field has over 10% missing values. Consider imputing these values.
pedestrian_ serious_injury_count	0.0%	5	0.000	0.004	0.000	9.000	0.071	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
bicycle_serious_injury_count	0.0%	4	0.000	0.002	0.000	3.000	0.043	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
micromobility_serious_injury_count	0.0%	3	0.000	0.000	0.000	2.000	0.018	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
motor_vehicle_death_count	0.0%	5	0.000	0.003	0.000	4.000	0.059	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
bicycle_death_count	0.0%	2	0.000	0.000	0.000	1.000	0.014	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
other_serious_injury_count	0.0%	3	0.000	0.000	0.000	3.000	0.009	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".

death_cnt	0.0%	5	0.000	0.006	0.000	4.000	0.082	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
crash_sev_id	0.0%	8	0.000	3.707	5.000	99.000	1.754	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
unkn_injry_cnt	0.0%	17	0.000	0.116	0.000	41.000	0.418	
poss_injry_cnt	0.0%	17	0.000	0.339	0.000	20.000	0.728	
crash_speed_limit	0.0%	29	-1.000	34.758	40.000	85.000	23.233	
crash_id	0.0%	147,750	1,001.000	16,810,778.317	16,758,228.500	180,290,542.000	1,832,197.355	
tot_injry_cnt	0.0%	19	0.000	0.642	0.000	21.000	0.935	
non_injry_cnt	0.0%	47	0.000	1.850	2.000	56.000	1.637	
nonincap_injry_cnt	0.0%	15	0.000	0.269	0.000	14.000	0.625	

Date Fields:

Name	% Missing	Unique Values	Latest Date	Earliest Date	Interval	Remarks
crash_time	0.0%	1,440	01/01/1400 23:00	01/01/1400 00:00	Unknown	

String/Character Fields:

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
pedestrian_fl	97.6%	2	Y	Y	3,505	144,245	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
bicycle_fl	98.3%	2	Y	Y	2,444	145,306	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
motorcycle_fl	97.6%	2	Y	Y	3,602	144,148	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash_fatal_fl	0.0%	2	N	N	866	146,884	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
latitude	1.5%	96,362	30.4	30.368659130525828	1	2,243	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
motor_vehicle_fl	0.8%	2	Y	Y	1,116	146,634	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
case_id	1.3%	145,679	.	10000 BLK US HIGHWAY	1	1,858	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
longitude	1.5%	96,265	-97.767	-97.69380787510062	1	2,243	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
micromobility_fl	99.8%	2	Y	Y	311	147,439	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

road_constr_ zone_fl	0.0%	3	N	N	2	139,901	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
other_fl	96.7%	2	Y	Y	4,845	142,905	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
point	1.5%	97,740	POINT (-97.65954 30.4)	POINT (-97.69380787510062 30.368659130525828)	1	2,243	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
units_involved	0.0%	1,113	Motorcycle	Other/Unknown & Passenger car & Passenger car & Large passenger vehicle & Large passenger vehicle & Large passenger vehicle & Large passenger vehicle & Passenger car & Motor vehicle â€“other & Large passenger vehicle & Passenger car & Other/Unknown & Pa	1	34,141	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
street_name_2	0.0%	3,398	N/A	E ANDERSON EB TO N 35 SB RAMPN IH35 SB	1	81,472	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
street_name	0.0%	4,631	441	PRIVATE DRIVE TO THE CATHERINE APARTMENTS	1	24,841	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
rpt_street_pfx	45.9%	9	W	SW	30	67,805	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash_date	0.0%	144,667	03/30/2014 10:58:00 AM	03/30/2014 10:58:00 AM	1	4	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
street_nbr	58.9%	9,828	0	11099	1	86,964	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
rpt_street_name	0.0%	9,796	1	ED BLUESTEIN BLVD SB TO ED BLUESTEIN BLVD SVRD SB	1	10,178	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

rpt_block_num	13.3%	4,789	E	11100 BLK	1	19,611	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
apd_confirmed_fatality	0.0%	2	N	N	842	146,908	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
rpt_street_sfx	34.1%	19	RD	BLVD	40	50,340	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
atd_mode_category_metadata	0.0%	147,744	[{"mode_id": 3, "mode_desc": "Motorcycle", "unit_id": 2262910, "death_cnt": 0, "sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 0, "poss_injry_cnt": 0, "non_injry_cnt": 1, "unkn_injry_cnt": 0, "tot_injry_cnt": 0}]	[{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2262335, "death_cnt": 0, "sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 0, "poss_injry_cnt": 0, "non_injry_cnt": 1, "unkn_injry_cnt": 0, "tot_injry_cnt": 0}, {"mode_id": 1, "mode_desc": "Passeng	1	7	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
private_dr_fl	0.0%	1	N	N	147,750	147,750	
onsys_fl	0.0%	2	N	N	73,041	74,709	

➔ **CHICAGO:**

There are 48 columns in this dataset.

Column name	Description	Type	Data Type
crash_record_id	This number can be used to link to the same crash in the Vehicles and People datasets. This number also serves as a unique ID in this dataset.	Plain Text	String
crash_date_est_i	Crash date estimated by desk officer or reporting party (only used in cases where crash is reported at police station days after the crash)	Plain Text	String
crash_date	Date and time of crash as entered by the reporting officer	Date & Time	Date
posted_speed_ limit	Posted speed limit, as determined by reporting officer	Number	Int
traffic_control_ device	Traffic control device present at crash location, as determined by reporting officer	Plain Text	String
device_condition	Condition of traffic control device, as determined by reporting officer	Plain Text	String
weather_ condition	Weather condition at time of crash, as determined by reporting officer	Plain Text	String
lighting_ condition	Light condition at time of crash, as determined by reporting officer	Plain Text	String
first_crash_ type	Type of first collision in crash	Plain Text	String
trafficway_type	Trafficway type, as determined by reporting officer	Plain Text	String
lane_cnt	Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection)	Number	Int
alignment	Street alignment at crash location, as determined by reporting officer	Plain Text	String
roadway_ surface_cond	Road surface condition, as determined by reporting officer	Plain Text	String
road_defect	Road defects, as determined by reporting officer	Plain Text	String
report_type	Administrative report type (at scene, at desk, amended)	Plain Text	String
crash_type	A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away	Plain Text	String
intersection_ related_i	A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection.	Plain Text	String
not_right_of_way_i	Whether the crash begun, or first contact was made outside of the public right-of-way.	Plain Text	String
hit_and_run_i	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid	Plain Text	String
damage	A field observation of estimated damage.	Plain Text	String
date_police_ notified	Calendar date on which police were notified of the crash	Date & Time	String
prim_contributory_cause	The factor, which was most significant in causing the crash, as determined by officer judgment	Plain Text	String
sec_contributory_cause	The factor, which was second most significant in causing the crash, as determined by officer judgment	Plain Text	String
street_no	Street address number of crash location, as determined by reporting officer	Number	Int
street_direction	Street address direction (N, E,S,W) of crash location, as determined by reporting officer	Plain Text	Int

street_name	Street address name of crash location, as determined by reporting officer	Plain Text	String
beat_of_ occurrence	Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74	Number	String
photos_taken_i	Whether the Chicago Police Department took photos at the location of the crash	Plain Text	String
statements_taken_i	Whether statements were taken from unit(s) involved in crash	Plain Text	String
dooring_i	Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash	Plain Text	String
work_zone_i	Whether the crash occurred in an active work zone	Plain Text	String
work_zone_type	The type of work zone if any	Plain Text	String
workers_present_i	Whether construction workers were present in an active work zone at crash location	Plain Text	String
num_units	Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.	Number	Int
most_severe_ injury	Most severe injury sustained by any person involved in the crash	Plain Text	String
injuries_total	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer	Number	Int
injuries_fatal	Total persons sustaining fatal injuries in the crash	Number	Int
injuries_ incapacitating	Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they could perform before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries.	Number	Int
injuries_non_ incapacitating	Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations.	Number	Int
injuries_reported_not_e vident	Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness, claims of injuries not evident, limping, complaint of pain, nausea, and hysteria.	Number	Int
injuries_no_ indication	Total persons sustaining no injuries in the crash as determined by the reporting officer	Number	Int
injuries_unknown	Total persons for whom injuries sustained, if any, are unknown	Number	Int
crash_hour	The hour of the day component of CRASH_DATE.	Number	Int
crash_day_of_ week	The day of the week component of CRASH_DATE. Sunday=1	Number	Int
crash_month	The month component of CRASH_DATE.	Number	Int
latitude	The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	Number	Int
longitude	The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash	Number	Int
location	The crash location, as determined by reporting officer, as derived from the reported address of crash, in a column type that allows for mapping and other geographic analysis in the data portal software	Point	Float/Double

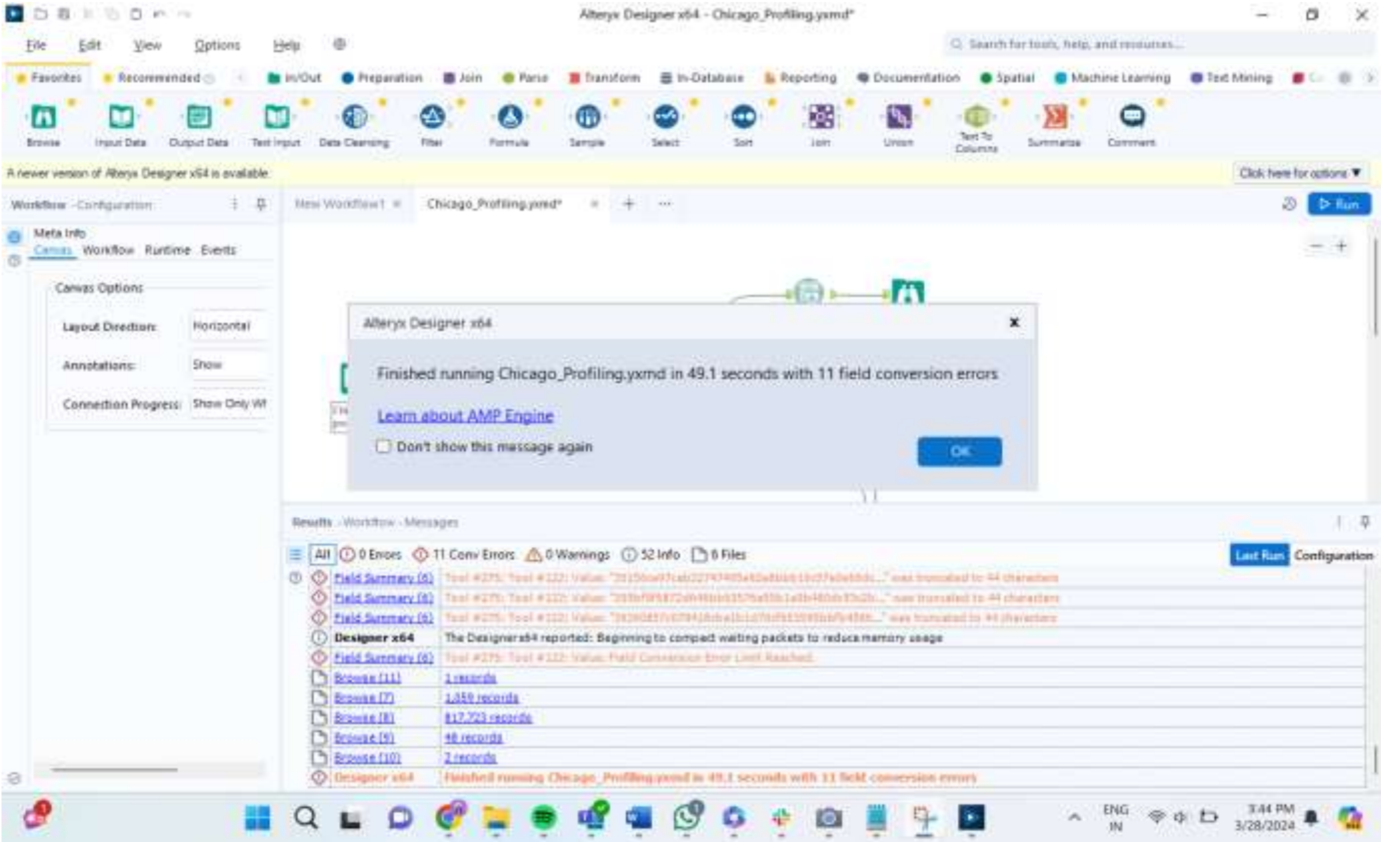
Dataset statistics:
Number of variables: **48**
Number of observations: **817723**
Missing cells: **8268003**
Missing cells (%): **21.1%**
Duplicate rows: **0**
Duplicate rows (%): **0.0%**
Total size in memory: **299.5 MiB**
Average record size in memory: **384.0 B**

Variable types:
Text: **3 |** Boolean: **9 |** Date Time: **2 |** Numeric: **15 |** Categorical: **19**

Missing values:
The following columns have missing values:
report_type 24314 (3.0%)
hit_and_run_i 561774 (68.7%)
lane_cnt 618714 (75.7%)
intersection_related_i 630174 (77.1%)
crash_date_est_i 756594 (92.5%)
not_right_of_way_i 780015 (95.4%)
statements_taken_i 799465 (97.8%)
photos_taken_I 806948 (98.7%)
work_zone_i 813053 (99.4%)
work_zone_type 814105 (99.6%)
dooring_i 815211 (99.7%)
workers_present_i 816529 (99.9%)

Observations:
- ➔ **report_type** has the least number of missing values and **workers_present_i** has the greatest number of missing values.
- ➔ **crash_record_id** has unique values and is a good fit for primary key.
- ➔ **crash_date** contains both the crash date and time. crash_date must be handled to separate the crash_time from it.
- ➔ The **crash_date** column ranges from 03/03/2013 to 03/26/2024, indicating no discrepancies in the dates being reported after the date of analysis.

Alteryx Workflow with the profiling run statistics:



REPORT:
Numeric Fields:

Name	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
lane_cnt	75.7%	42	0.000	13.330	2.000	1,191,625.000	2,961.601	This field has over 10% missing values. Consider imputing these values.
injuries_fatal	0.2%	6	0.000	0.001	0.000	4.000	0.037	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
injuries_unknown	0.2%	2	0.000	0.000	0.000	0.000	0.000	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
crash_day_of_week	0.0%	7	1.000	4.123	4.000	7.000	1.981	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
injuries_no_indication	0.2%	49	0.000	2.003	2.000	61.000	1.157	
longitude	0.7%	300,055	-87.936	-87.674	-87.674	0.000	0.684	
num_units	0.0%	17	1.000	2.035	2.000	18.000	0.453	

street_no	0.0%	11,728	0.000	3,689.634	3,201.000	451,100.000	2,886.884	
latitude	0.7%	300,092	0.000	41.855	41.875	42.023	0.336	
beat_of_occurrence	0.0%	277	111.000	1,243.735	1,212.000	6,100.000	705.271	
injuries_reported_not_evident	0.2%	14	0.000	0.062	0.000	15.000	0.319	
crash_month	0.0%	12	1.000	6.656	7.000	12.000	3.453	
injuries_non_incapacitating	0.2%	20	0.000	0.107	0.000	21.000	0.422	
crash_hour	0.0%	24	0.000	13.198	14.000	23.000	5.570	
injuries_incapacitating	0.2%	11	0.000	0.020	0.000	10.000	0.165	
injuries_total	0.2%	21	0.000	0.190	1.000	21.000	0.566	
posted_speed_limit	0.0%	46	0.000	28.406	30.000	99.000	6.166	

String/Character Fields:

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
weather_condition	0.0%	12	SNOW	BLOWING SAND, SOIL, DIRT	7	641,373	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
date_police_notified	0.0%	620,545	09/05/2023 07:05:00 PM	09/05/2023 07:05:00 PM	1	12	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash_date_est_i	92.5%	3	Y	Y	7,862	756,594	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
lighting_condition	0.0%	6	DUSK	DARKNESS, LIGHTED ROAD	13,689	522,973	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
dooring_i	99.7%	3	Y	Y	824	815,211	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
road_defect	0.0%	7	OTHER	DEBRIS ON ROADWAY	616	657,425	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
street_name	0.0%	1,642	82ND	MICHIGAN AVE 175 E CHESTNUT AVE	1	22,319	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash_record_id	0.0%	817,723	23a79931ef555d54118f64dc9be2cf2dbf59636ce253f7a1179c4a1c091442a6eeab8352220c7c56ca1ff7c4b4b0fc345c74e3e85ecb9d43deeb66b5f803d4a0	23a79931ef555d54118f64dc9be2cf2dbf59636ce253f7a1179c4a1c091442a6eeab8352220c7c56ca1ff7c4b4b0fc345c74e3e85ecb9d43deeb66b5f803d4a0	1	1	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

first_crash_type	0.0%	18	ANGLE	SIDESWIPE OPPOSITE DIRECTION	45	190,062	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
report_type	3.0%	4	AMENDED	NOT ON SCENE (DESK REPORT)	240	447,696	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
trafficway_type	0.0%	20	RAMP	DIVIDED - W/MEDIAN (NOT RAISED)	168	355,079	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
statements_takn_i	97.8%	3	Y	Y	3,384	799,465	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash_date	0.0%	536,888	09/05/2023 07:05:00 PM	09/05/2023 07:05:00 PM	1	30	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
street_direction	0.0%	5	S	S	4	292,260	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
alignment	0.0%	6	CURVE, LEVEL	STRAIGHT ON HILLCREST	364	797,888	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
traffic_control_device	0.0%	19	OTHER	PEDESTRIAN CROSSING SIGN	25	464,816	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
location	0.7%	300,266	POINT (0 0)	POINT (-87.665902342962 41.854120262952)	1	5,615	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
hit_and_run_i	68.7%	3	Y	Y	10,994	561,774	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
work_zone_i	99.4%	3	Y	Y	1,052	813,053	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
photos_taken_i	98.7%	3	Y	Y	2,665	806,948	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
device_condition	0.0%	8	OTHER	WORN REFLECTIVE MATERIAL	95	470,240	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
most_severe_injury	0.2%	6	FATAL	NONINCAPACITATING INJURY	899	703,419	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
not_right_of_way_i	95.4%	3	Y	Y	3,452	780,015	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

sec_contributory_cause	0.0%	40	ANIMAL	OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER	56	335,814	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
prim_contributory_cause	0.0%	40	ANIMAL	OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER	23	318,311	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
intersection_related_i	77.1%	3	Y	Y	8,907	630,174	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
workers_present_i	99.9%	3	Y	Y	138	816,529	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
work_zone_type	99.6%	5	UTILITY	CONSTRUCTION	228	814,105	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
roadway_surface_cond	0.0%	7	DRY	SAND, MUD, DIRT	303	603,295	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
damage	0.0%	3	OVER $1,500	$501 - $1,500	93,373	507,718	
crash_type	0.0%	2	NO INJURY / DRIVE AWAY	INJURY AND / OR TOW DUE TO CRASH	218,340	599,383	

➔ **NYC:**

There are 28 columns in this dataset.

Column name	**Description**	**Type**	**Data Type**
crash date	Occurrence date of collision	Date & Time	Date
crash time	Occurrence time of collision	Plain Text	String
borough	Borough where collision occurred	Plain Text	String
zip code	Postal code of incident occurrence	Plain Text	String
latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	Number	Int
longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)	Number	Int
location	Latitude, Longitude pair	Location	Float/Double
on street name	Street on which the collision occurred	Plain Text	String
cross street name	Nearest cross street to the collision	Plain Text	String
off street name	Street address if known	Plain Text	String
number of persons injured	Number of persons injured	Number	Int
number of persons killed	Number of persons killed	Number	Int
number of pedestrians injured	Number of pedestrians injured	Number	Int
number of pedestrians killed	Number of pedestrians killed	Number	Int
number of cyclists injured	Number of cyclists injured	Number	Int
number of cyclists killed	Number of cyclists killed	Number	Int
number of motorists injured	Number of vehicle occupants injured	Number	Int
number of motorists killed	Number of vehicle occupants killed	Number	Int
contributing factor vehicle 1	Factors contributing to the collision for designated vehicle	Plain Text	String
contributing factor vehicle 2	Factors contributing to the collision for designated vehicle	Plain Text	String
contributing factor vehicle 3	Factors contributing to the collision for designated vehicle	Plain Text	String
contributing factor vehicle 4	Factors contributing to the collision for designated vehicle	Plain Text	String
contributing factor vehicle 5	Factors contributing to the collision for designated vehicle	Plain Text	String
collision_id	Unique record code generated by system. Primary Key for Crash table.	Number	Int
vehicle type code 1	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	Plain Text	String
vehicle type code 2	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	Plain Text	String
vehicle type code 3	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	Plain Text	String
vehicle type code 4	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	Plain Text	String
vehicle type code 5	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)	Plain Text	String

a) **Profiling using Y-data profile.**

Dataset statistics:
Number of variables: **29**
Number of observations: **2075427**
Missing cells: **17761579**
Missing cells (%): **29.5%**
Duplicate rows: **0**
Duplicate rows (%): **0.0%**
Total size in memory: **459.2 MiB**
Average record size in memory: **232.0 B**

Variable types:
Date Time: **2 |** Categorical: **6** | Text: **13** | Numeric: **8**

Missing values:

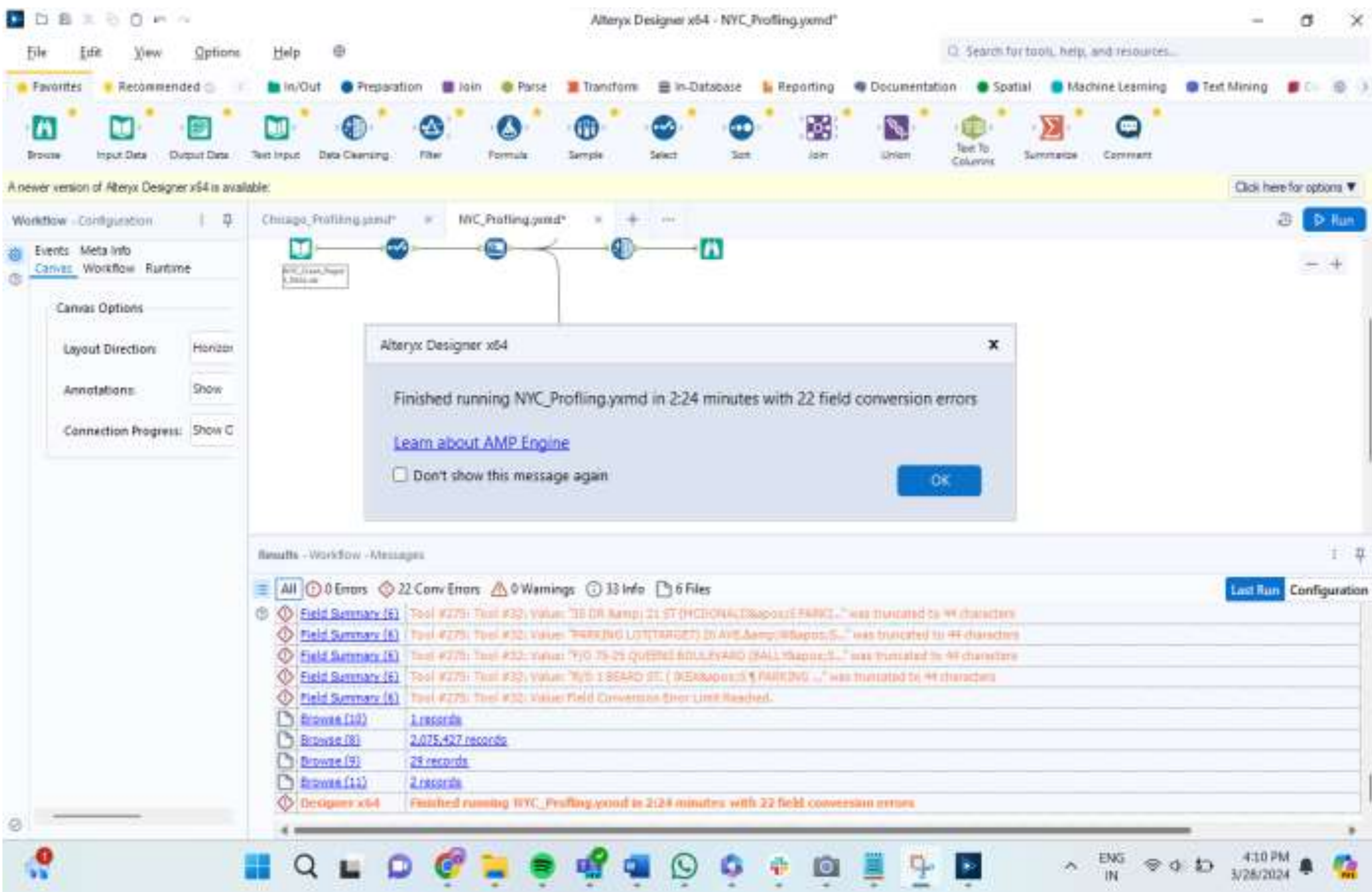
➔ latitude has 233626 (11.3%)
➔ longitude has 233626 (11.3%)
➔ location has 233626 (11.3%)
➔ contributing factor vehicle 2 has 321736 (15.5%)
➔ vehicle type code 2 has 396691 (19.1%)
➔ on street name has 440569 (21.2%)
➔ borough has 645746 (31.1%)
➔ zip code has 645996 (31.1%)
➔ cross street name has 784436 (37.8%)
➔ off street name has 1727231 (83.2%)
➔ contributing factor vehicle 3 has 1927163 (92.9%)
➔ vehicle type code 3 has 1932530 (93.1%)
➔ contributing factor vehicle 4 has 2041953 (98.4%)
➔ vehicle type code 4 has 2043115 (98.4%)
➔ contributing factor vehicle 5 has 2066358 (99.6%)
➔ vehicle type code 5 has 2066635 (99.6%)

Observations:

➔ **Latitude, Longitude, and location** have the least number of missing values and **contributing factor vehicle 5 and vehicle type code 5** has the greatest number of missing values.
➔ **collision_id** has unique values and is a good fit for primary key.
➔ The **crash_date** column ranges from 7/27/2012 to 03/07/2024, indicating no discrepancies in the dates being reported after the date of analysis.

b) Using Alteryx

Alteryx Workflow with the profiling run statistics:



REPORT:
Numeric Fields:

Name	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
longitude	11.3%	98,352	-201.360	-73.752	-73.926	0.000	3.723	This field has over 10% missing values. Consider imputing these values.
number of motorists killed	0.0%	6	0.000	0.001	0.000	5.000	0.027	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
number of persons killed	0.0%	8	0.000	0.001	0.000	8.000	0.041	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
latitude	11.3%	126,595	0.000	40.628	40.722	43.344	1.981	This field has over 10% missing values. Consider imputing these values.

number of pedestrians killed	0.0%	4	0.000	0.001	0.000	6.000	0.028	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
number of cyclists injured	0.0%	5	0.000	0.027	0.000	4.000	0.163	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
number of cyclists killed	0.0%	3	0.000	0.000	0.000	2.000	0.011	This field has a small number of unique values and appears to be a categorical field. Consider changing the field data type to "string".
number of persons injured	0.0%	33	0.000	0.310	1.000	43.000	0.700	
number of motorists injured	0.0%	31	0.000	0.223	0.000	43.000	0.661	
number of pedestrians injured	0.0%	14	0.000	0.057	0.000	27.000	0.244	
collision_id	0.0%	2,075,427	22.000	3,159,626.962	3,673,954.000	4,712,252.000	1,505,149.883	

String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
crash time	0.0%	1,440	2:39	11:45	94	28,391	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
vehicle type code 1	0.7%	1,635	.	Enclosed Body – Non-removable Enclosure	1	576,659	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
on street name	21.2%	13,081		WILLIAMSBURG BRIDGE OUTER ROADWAY	1	440,569	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
contributing factor vehicle 4	98.4%	42	Glare	Traffic Control Device Improper/Non-Working	1	2,041,953	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
contributing factor vehicle 5	99.6%	31	Glare	Traffic Control Device Improper/Non-Working	1	2,066,358	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

location	11.3%	283,007	(0.0, 0.0)	(40.8359271, -73.9029039)	1	233,626	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
contributing factor vehicle 2	15.5%	62	1	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	3	1,476,469	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
vehicle type code 5	99.6%	71	C3	Station Wagon/Sport Utility Vehicle	1	2,066,635	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
off street name	83.2%	200,266		26-45 Brooklyn queen’s expressway wes	1	1,727,231	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
vehicle type code 4	98.4%	102	PK	Station Wagon/Sport Utility Vehicle	1	2,043,115	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
cross street name	37.8%	13,699		CROSS BRONX EXPRESSWAY EXTENSION	1	784,436	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
crash date	0.0%	4,283	09/11/2021	09/11/2021	94	1,161	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
contributing factor vehicle 3	92.9%	52	1	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	1	1,927,163	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
vehicle type code 3	93.1%	263	PK	Station Wagon/Sport Utility Vehicle	1	1,932,528	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
borough	31.1%	6	BRONX	STATEN ISLAND	60,012	645,746	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
contributing factor vehicle 1	0.3%	62	1	Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	10	706,732	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
vehicle type code 2	19.1%	1,824	0	Enclosed Body - Nonremovable Enclosure	1	403,529	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
zip code	31.1%	236		11208	1	645,996	This field has over 10% missing values. Consider imputing these values. Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

DIMENSIONAL MODEL:

HOW DOES OUR DIMENSIONAL MODEL ANSWER THE BUSINESS QUESTIONS:

~ How many accidents occurred in NYC, Austin, and Chicago?

In the dimensional model, the Fct_Accident fact table holds records of individual accidents, including details about each incident's outcome, such as injuries and fatalities. The Dim_Location table contains attributes such as City, StreetName, and Latitude/Longitude, which allow us to filter and count accidents based on the city. By joining Fct_Accident with Dim_Location on the LocationKey and grouping the results by the city attribute, we can determine the number of accidents for each specified city.

~ Best way to present these values on the dashboard:

For visual representation on a dashboard, a bar chart is effective for comparing the number of accidents across cities. Alternatively, a map visualization with pins or heatmaps can provide a geographical context, making it immediately apparent which cities have higher accident rates.

~ Which areas in the 3 cities had the greatest number of accidents?

Using the Dim_Location table, we can identify specific areas with the highest accidents by grouping accident data by City and AreaName (if this attribute is captured in the location dimension) or by other geographic identifiers available. Sorting these groups by the count of related accidents in descending order and then selecting the top three will give us the areas with the greatest number of accidents in each city.

~How many accidents resulted in just injuries?

To find accidents that resulted only in injuries, we would use the Fct_Accident fact table and filter records where InjuriesCount is greater than zero and FatalitiesCount is zero. We can then aggregate this data to get a total count for the overall report and group by City for the city-level report.

~ How often are pedestrians involved in accidents?

To determine the frequency of pedestrian involvement, we can sum the PedestriansInvolvedCount from the Fct_Accident table. For an overall count, sum this value across all records. For city-level data, perform the same sum but group the results by the City field from the Dim_Location table.

~ When do most accidents happen? (Seasonality report)

For the seasonality report, you will need to join Fct_Accident with the Dim_Date table on the relevant date key. Then, you can aggregate accident counts based on the Month or Season fields. This will reveal patterns or trends indicating when accidents are most frequent, such as particular seasons or months of the year.

~ How many motorists are injured or killed in accidents?

By referencing the MotoristsInjuredCount and MotoristsKilledCount in the Fct_Accident table, we can sum these figures for the overall statistics. To break down the numbers by city, you would perform the sum within groups determined by joining with Dim_Location on the LocationKey and grouping by City.

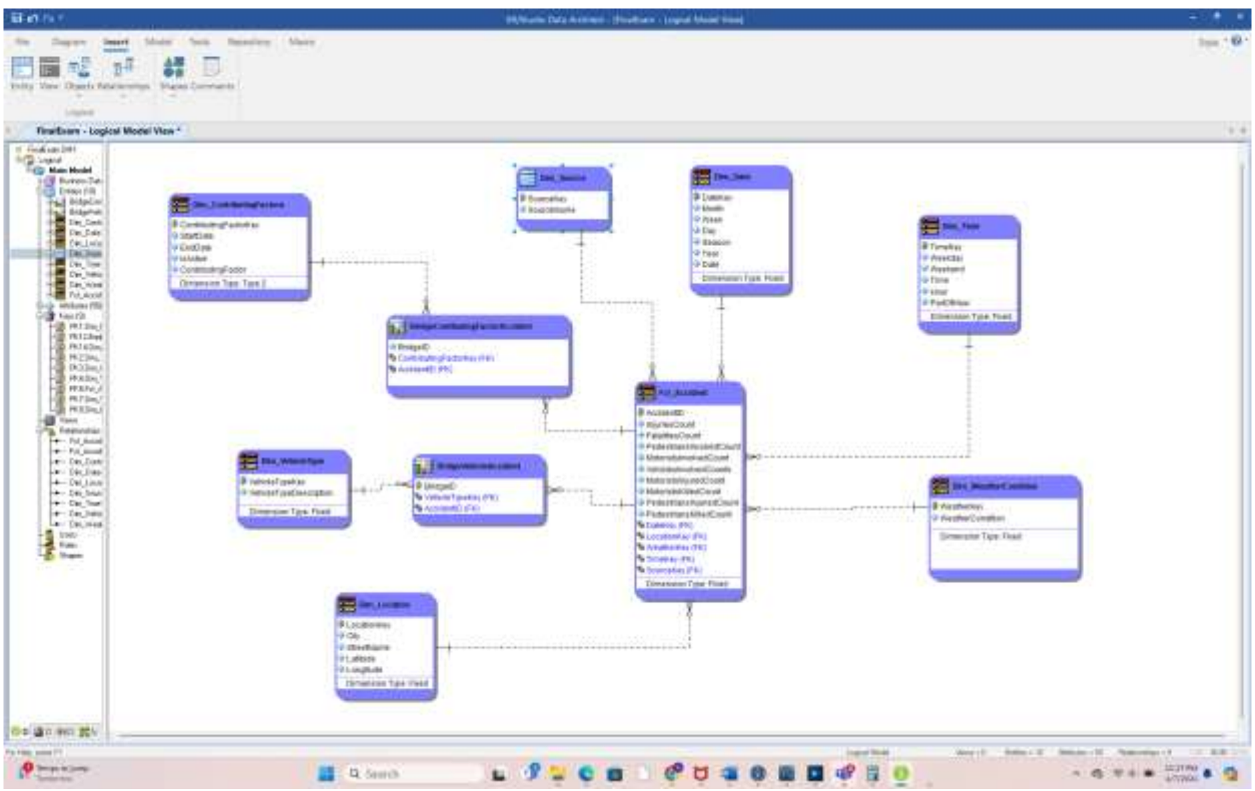
~ Which top 5 areas in 3 cities have the most fatal number of accidents?

This requires analyzing the FatalitiesCount from the Fct_Accident table. By joining with the Dim_Location table and grouping by location-related attributes, you can sort the results by the number of fatalities in descending order. Limiting the result set to the top five will provide the areas with the most fatal accidents.

~ Time-based analysis of accidents:
Join Fct_Accident with Dim_Time on TimeKey and then group and count the accidents by different time dimensions such as Hour, Weekday vs. Weekend, and Time (morning, afternoon, evening). This will allow you to analyze and visualize the data to understand accident patterns relative to the time of day or week.

~ Fatality analysis:
Comparing pedestrian fatalities to those of motorists involves aggregating PedestriansKilledCount and comparing it against MotoristsKilledCount from the Fct_Accident table. Additionally, comparing these numbers to the total FatalitiesCount will indicate which group is more at risk.

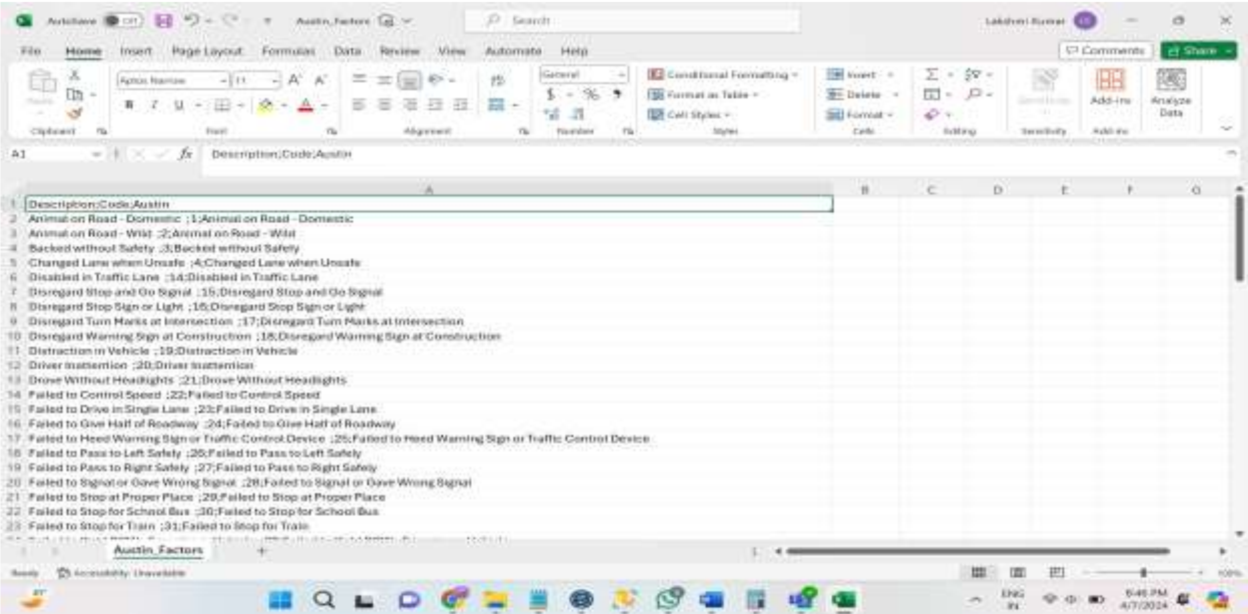
~ What are the most common factors involved in accidents?
This can be ascertained by analyzing the Dim_ContributingFactors dimension table. By joining this table with the BridgeContributingFactorAccident table, which relates contributing factors to specific accidents, and then counting the frequency of each contributing factor, you can identify which are most common in accidents.



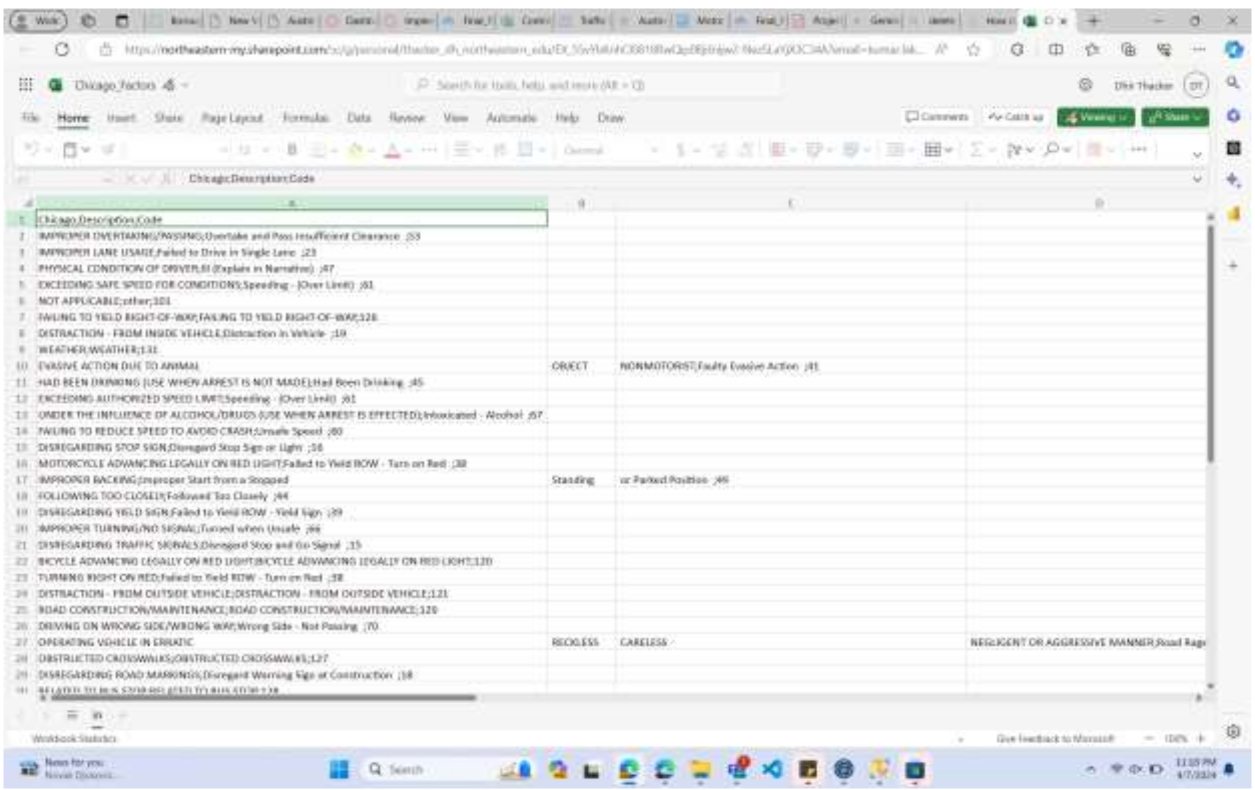
CONTRIBUTING FACTORS MAPPING DOCUMENT:

Took the codes and descriptions from the Contributing Factors Mapping document provided in One drive and performed a VLOOKUP with the normalized contributory cause columns and created separate csv sheets for each dataset. Used these csv sheets as input with the final normalized table to map code with the contributory cause.

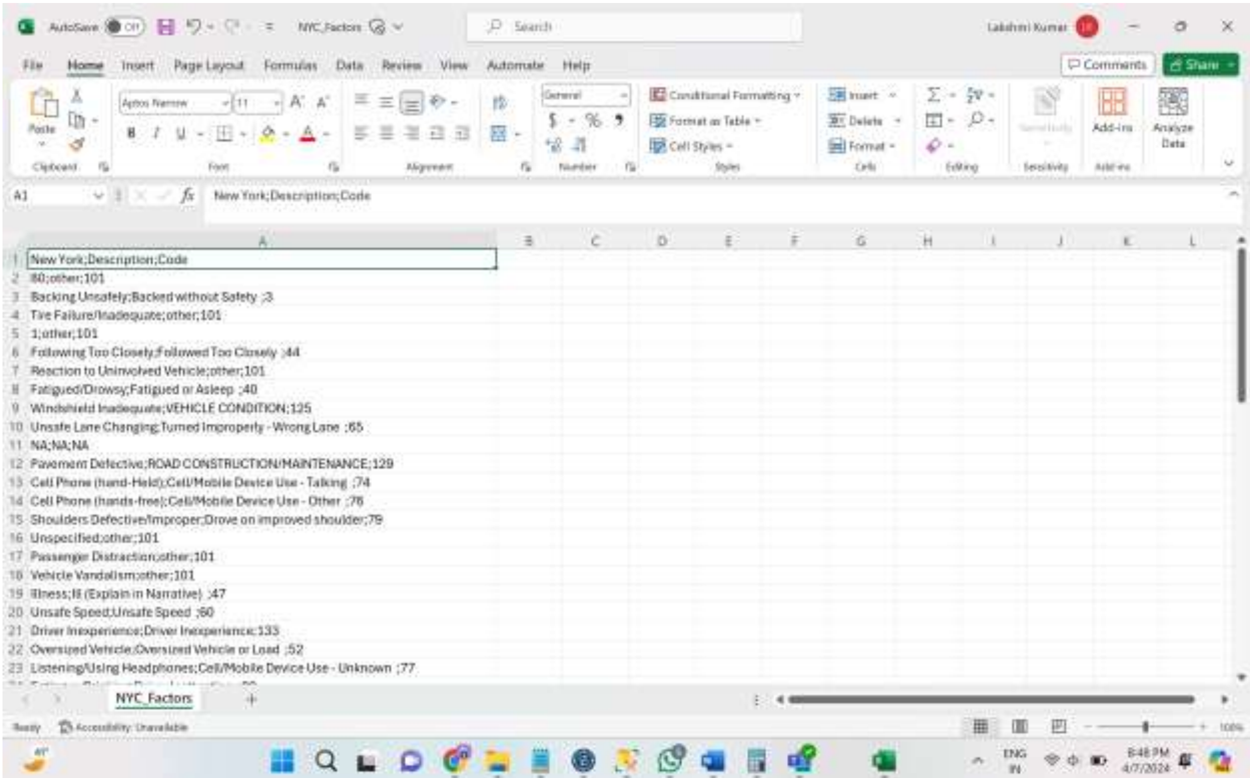
- ➔ AUSTIN:



- ➔ CHICAGO:

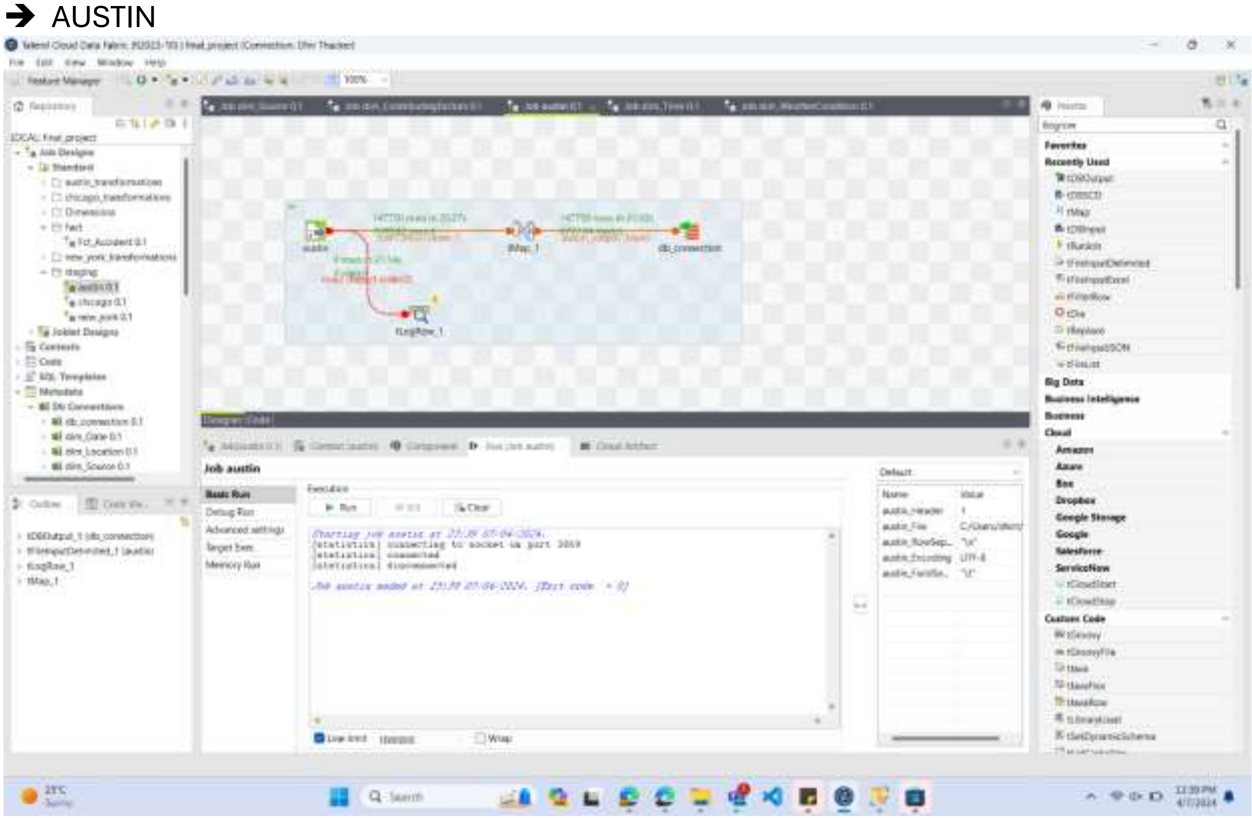


➔ NYC:

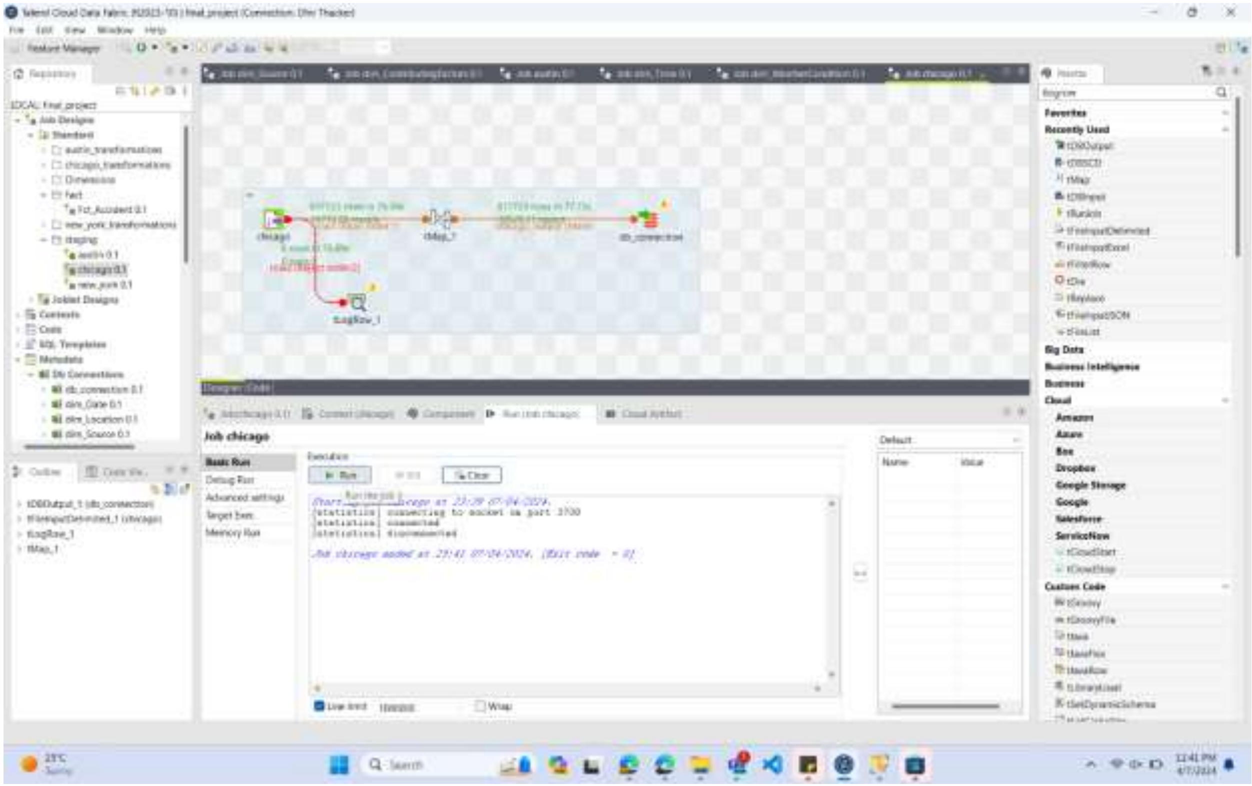


BASIC STAGING OF ALL THREE DATASET USING TALEND:

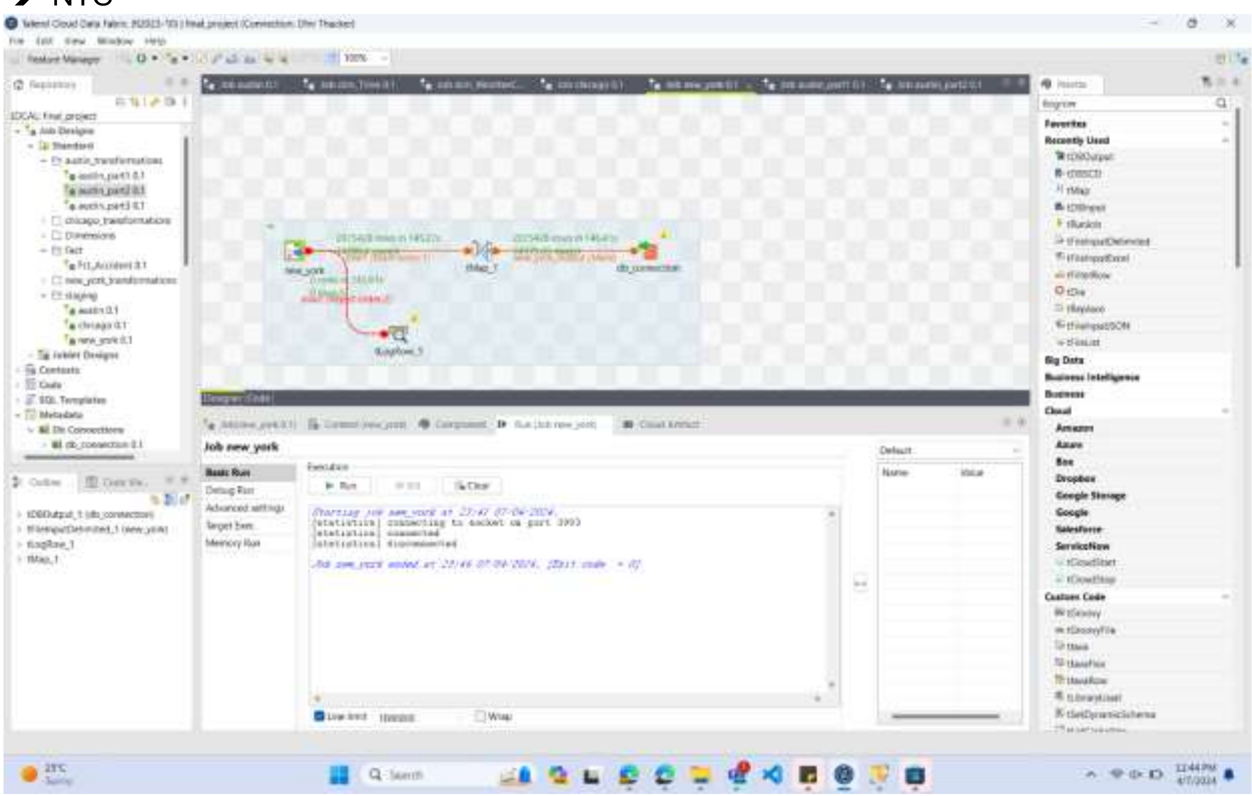
➔ AUSTIN



➔ CHICAGO

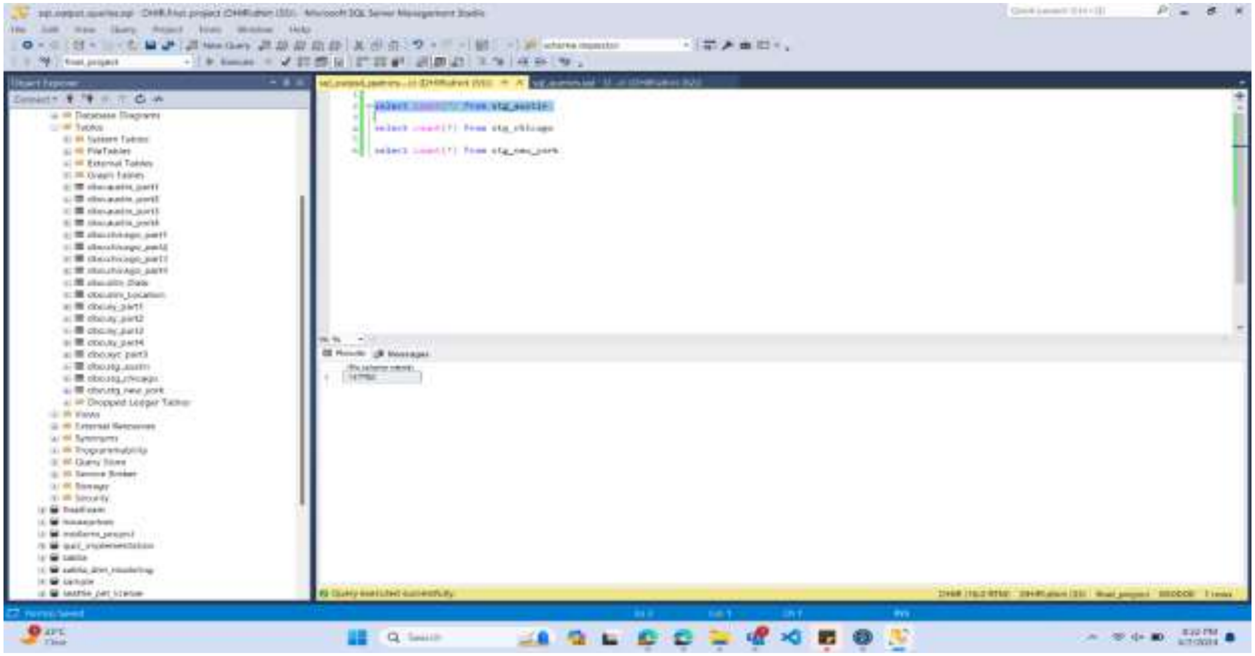


➔ NYC

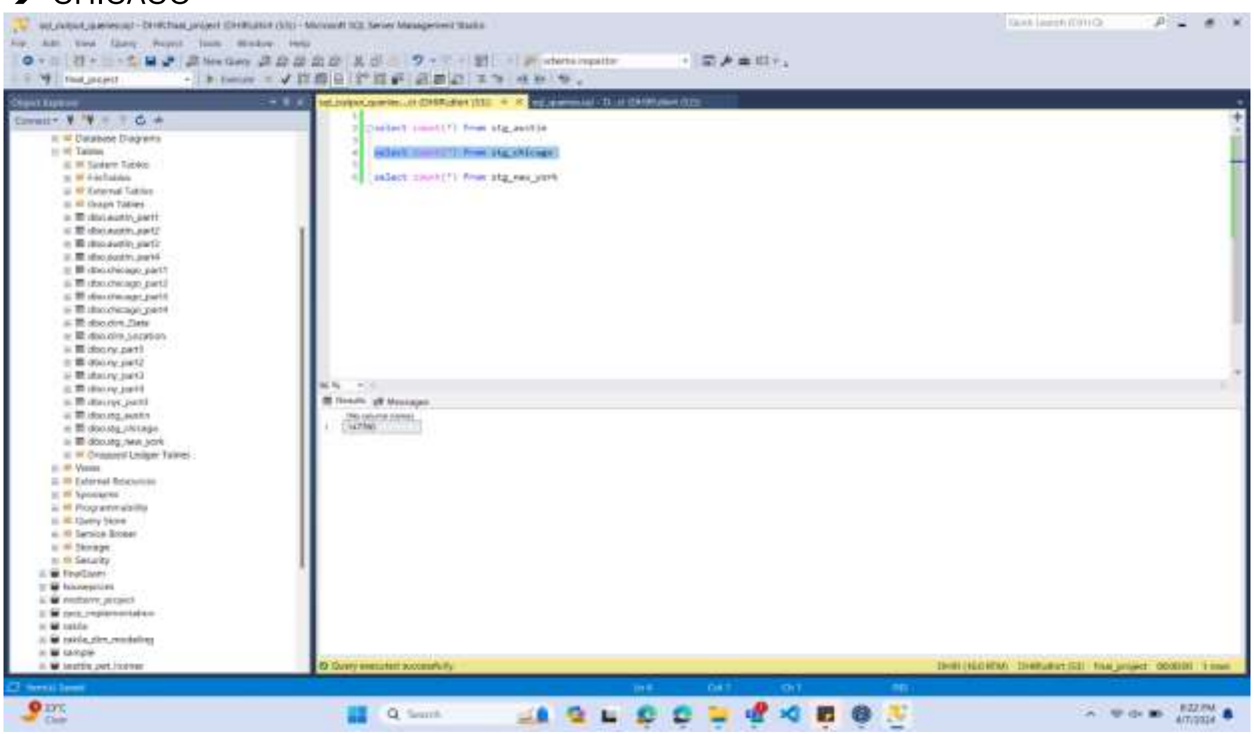


SQL OUTPUTS FOR COUNT OF EACH DATASET:

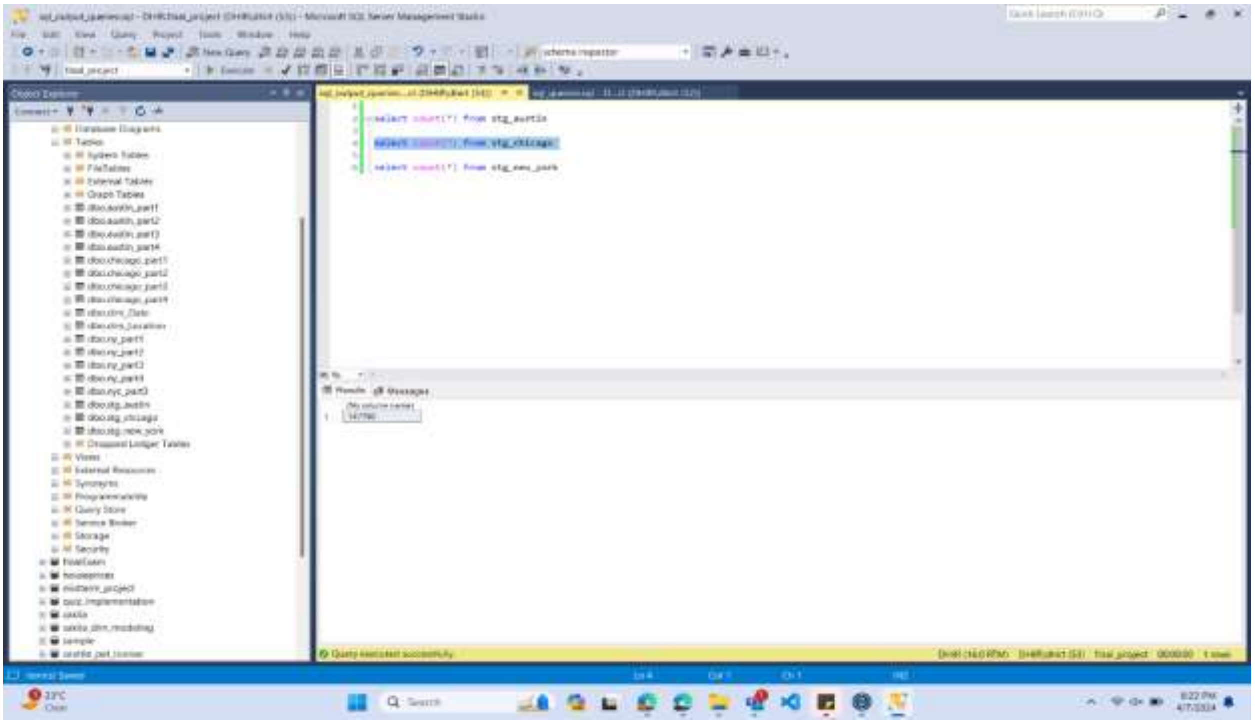
➔ AUSTIN



➔ CHICAGO

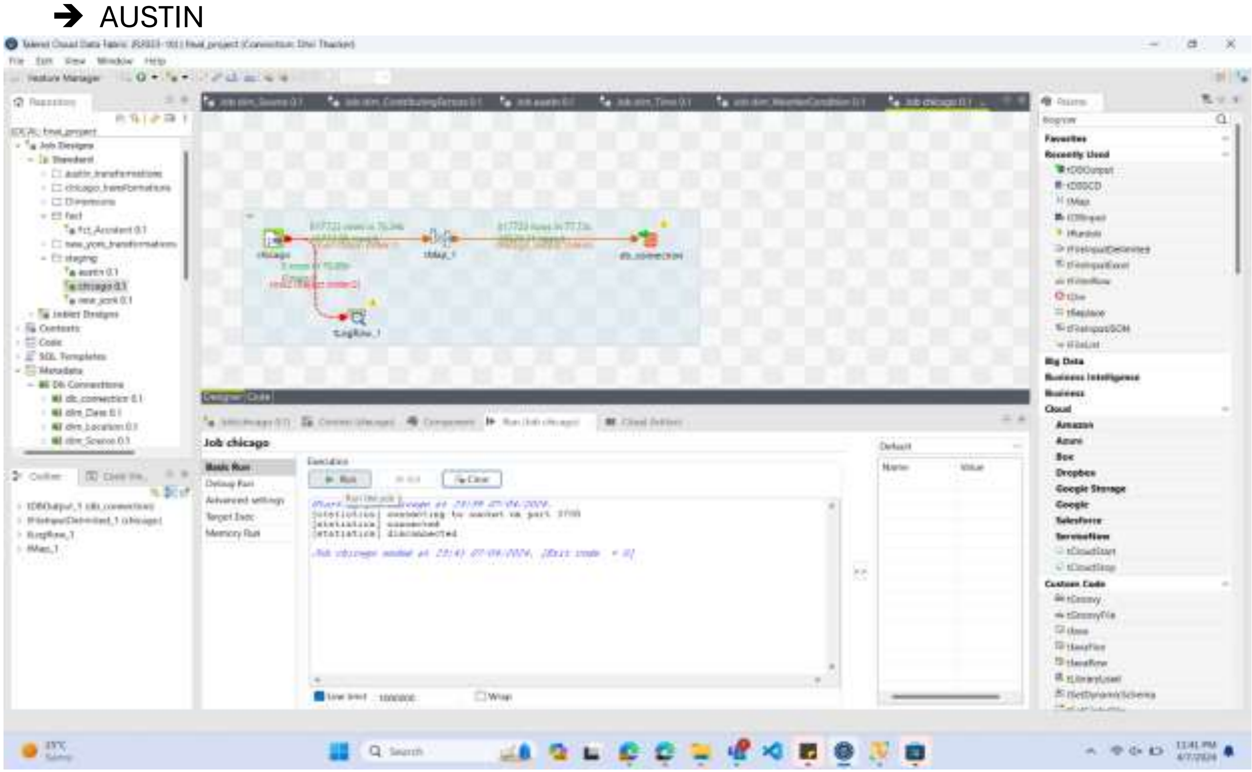


➔ NYC

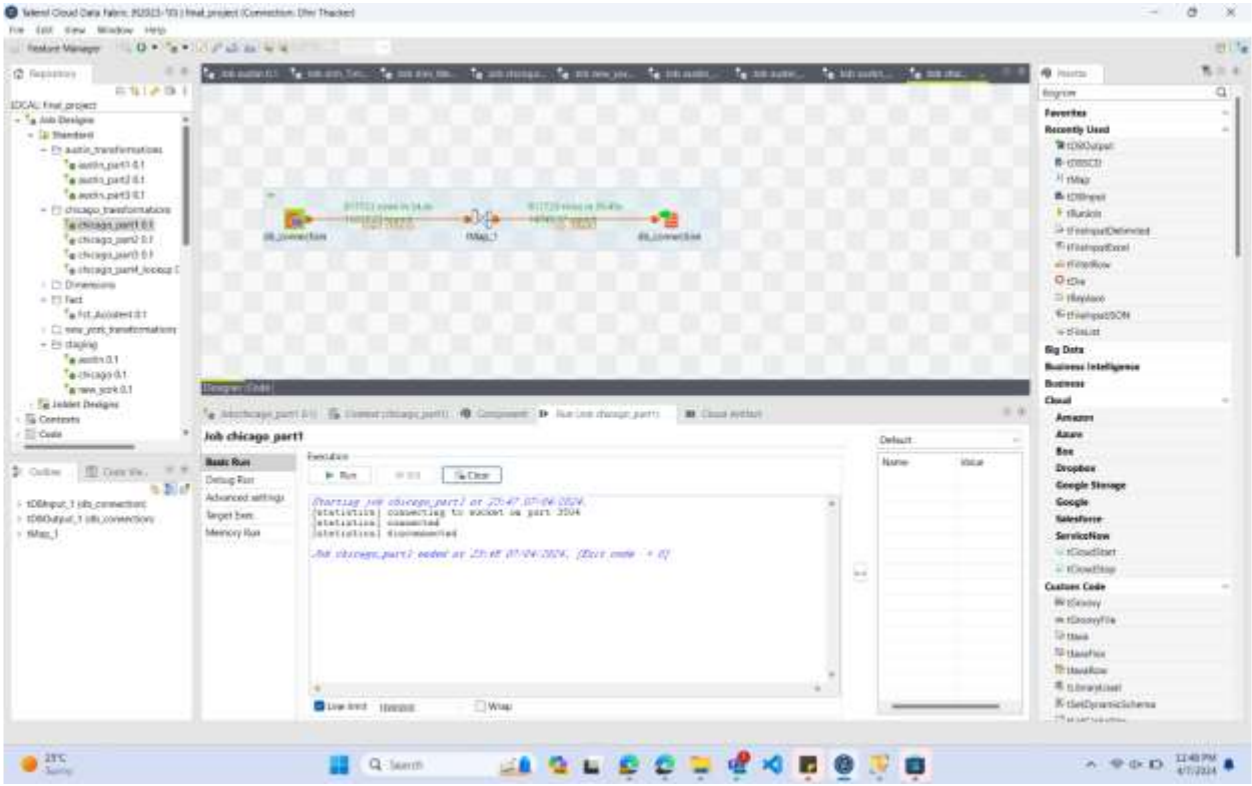


TRANSFORMATIONS PART 1:
Removing columns that are not required for the business questions and handling null values till contributing_factors column.

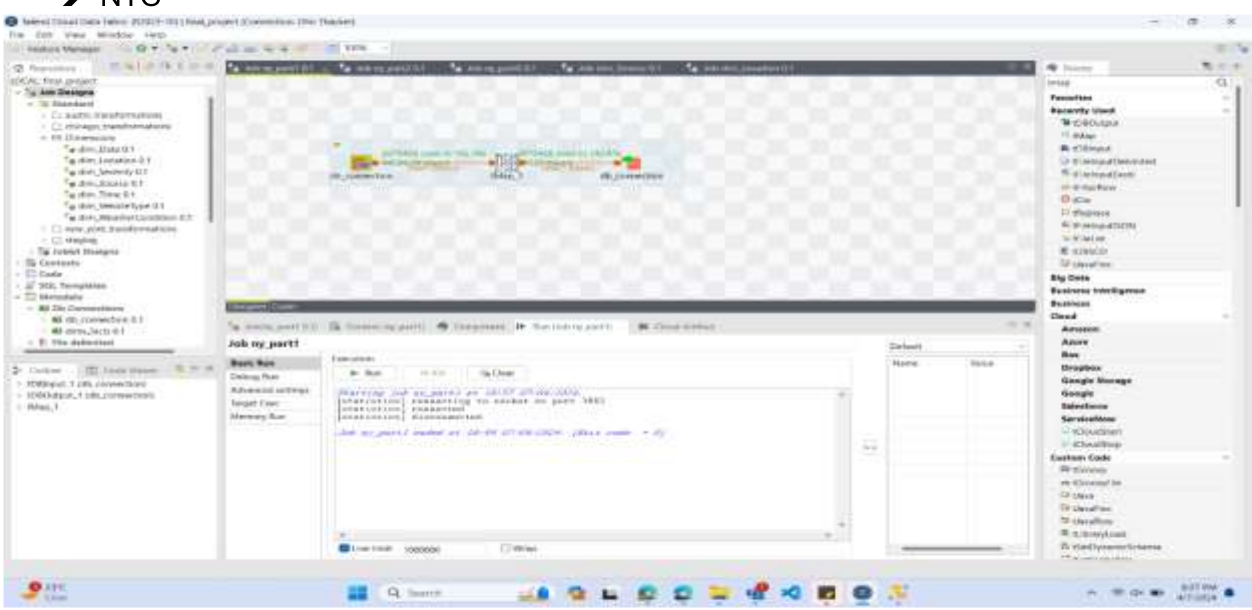
➔ AUSTIN



→ CHICAGO



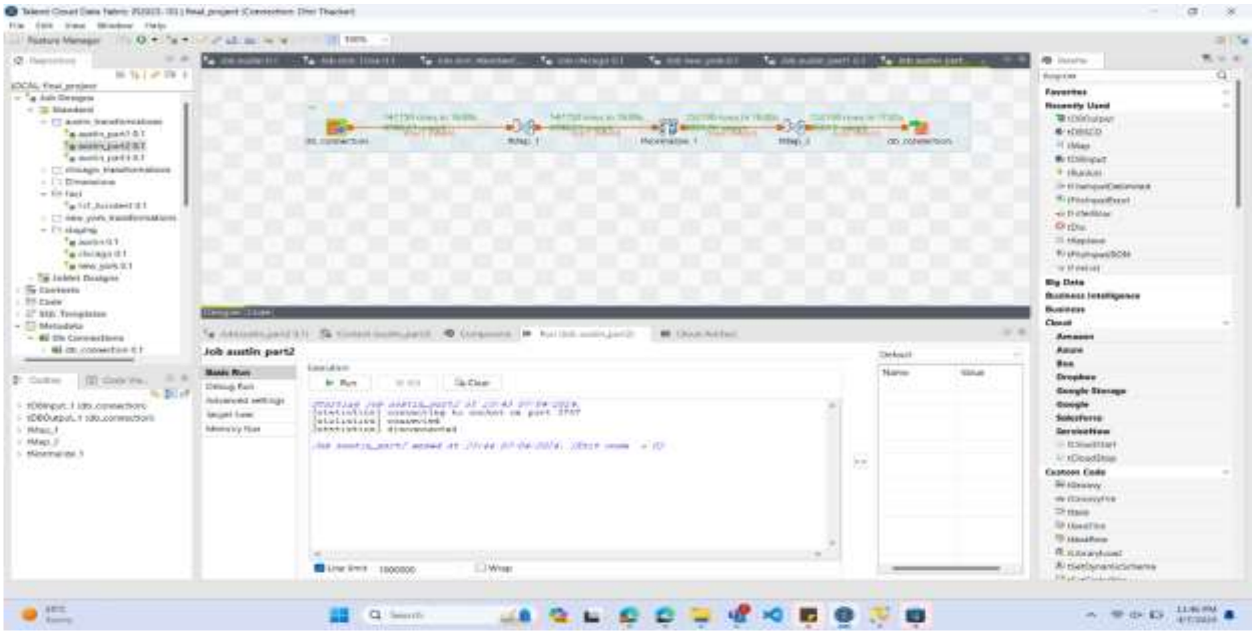
→ NYC



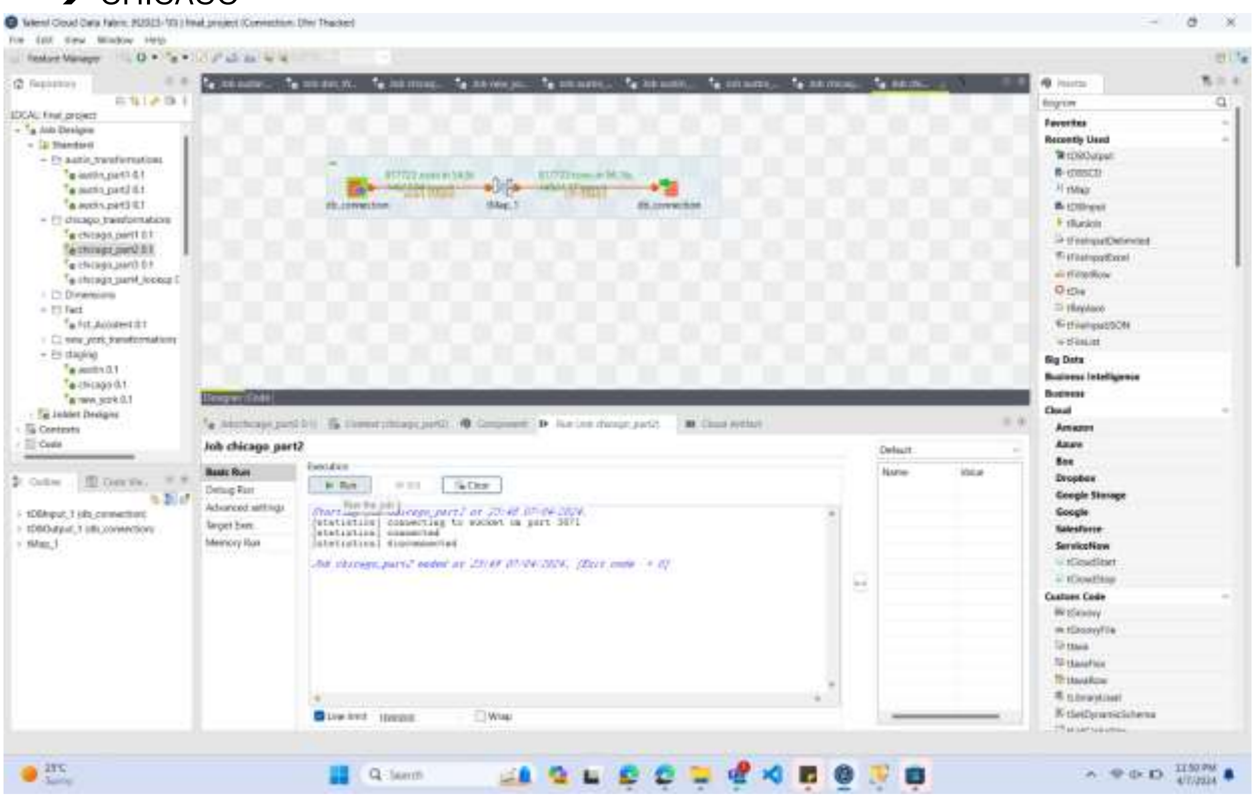
TRANSFORMATIONS PART 2 CONTRIBUTING FACTORS

Handling null values and merging contributing factors column and eventually normalizing it.

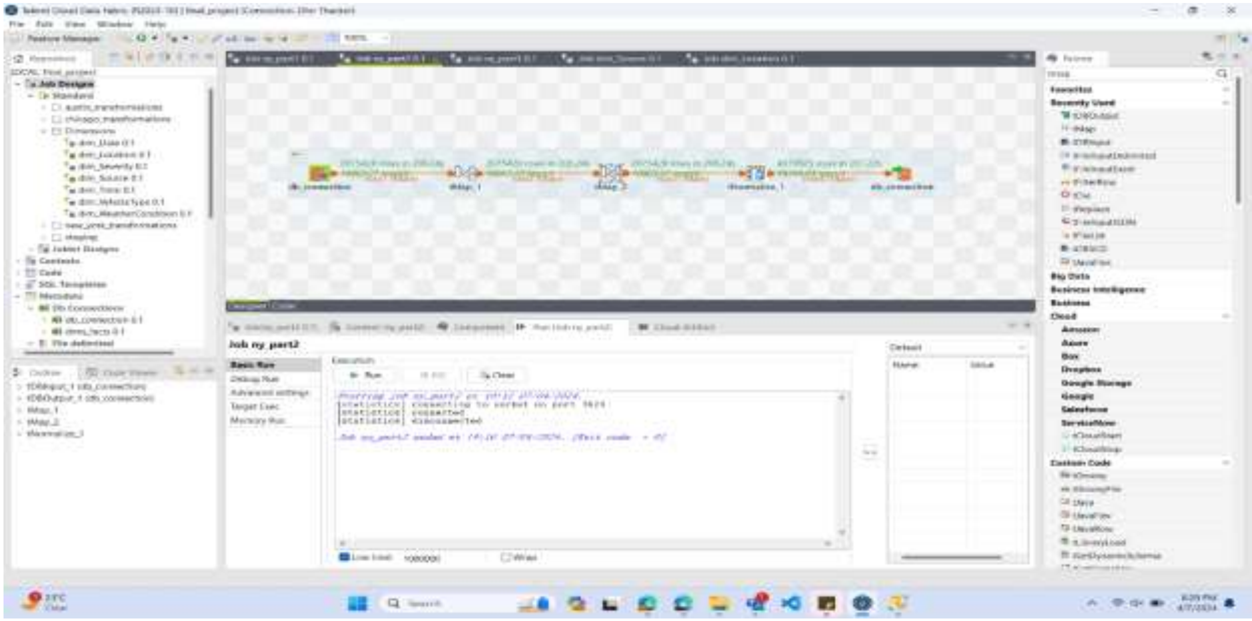
➔ AUSTIN



➔ CHICAGO



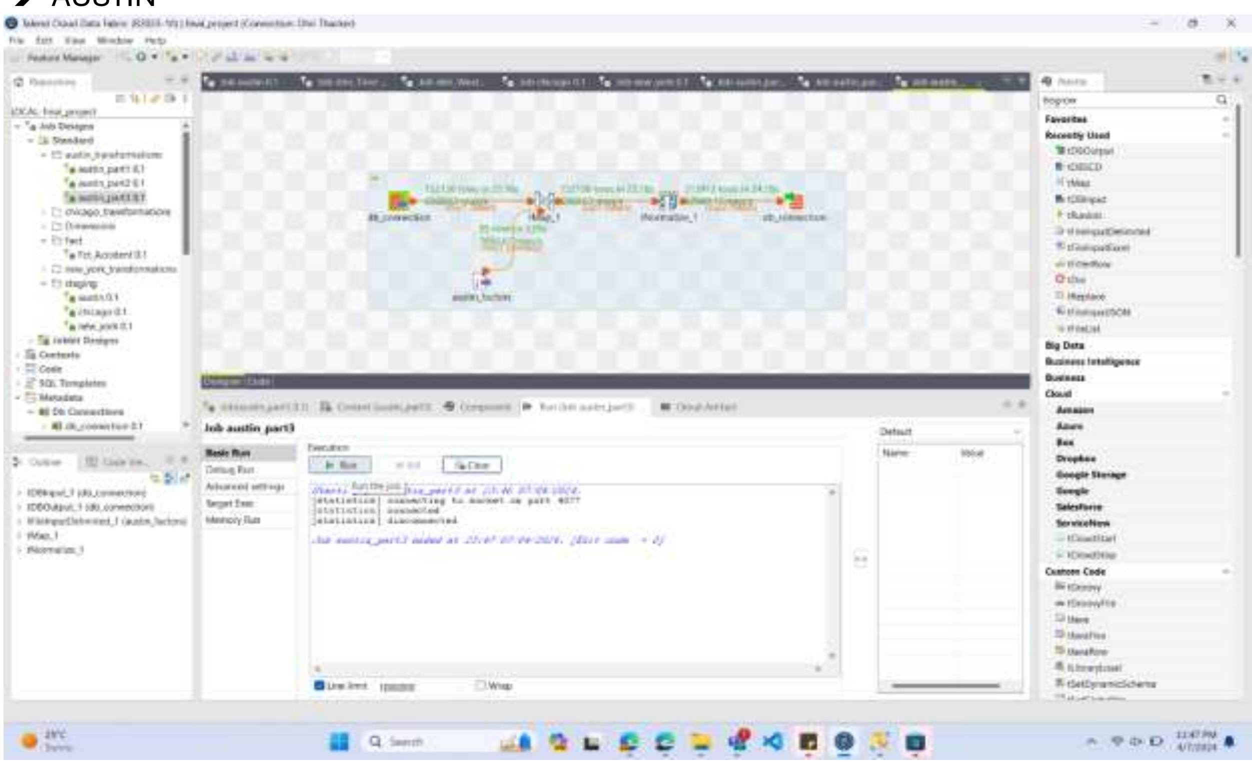
➔ NYC



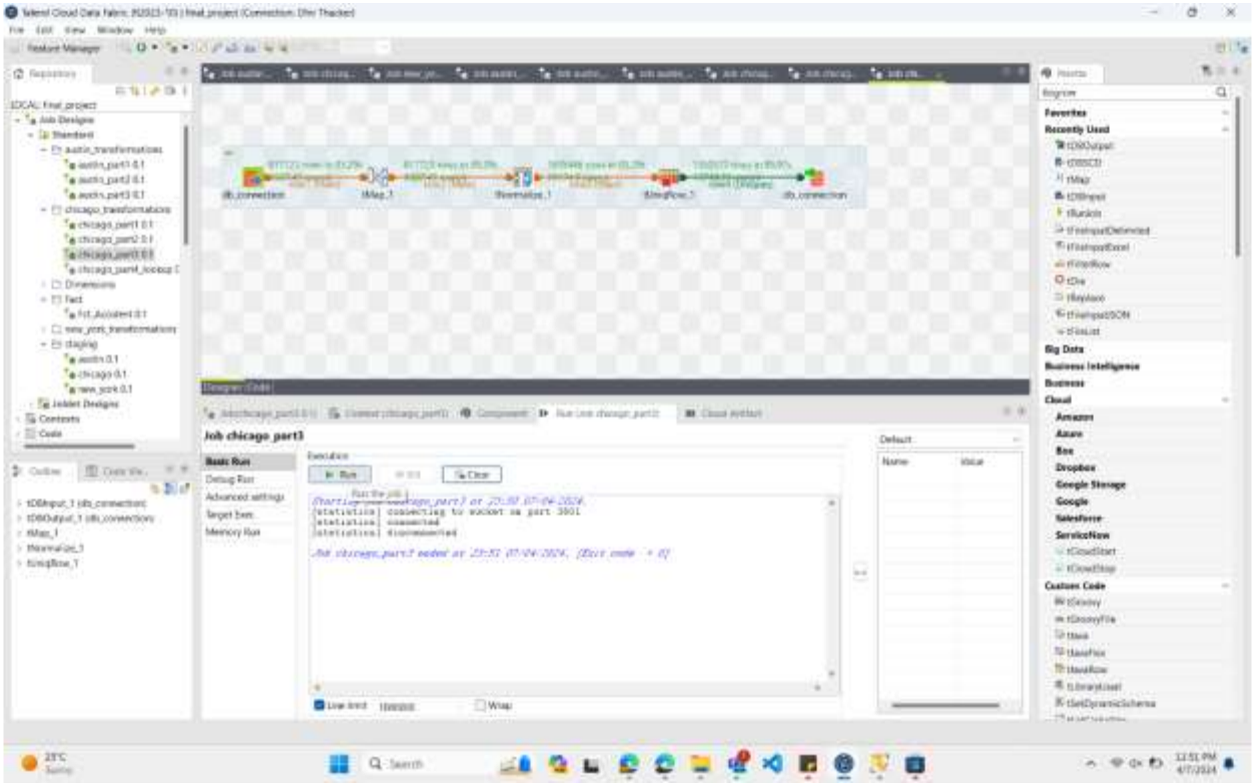
TRANSFORMATIONS PART 3

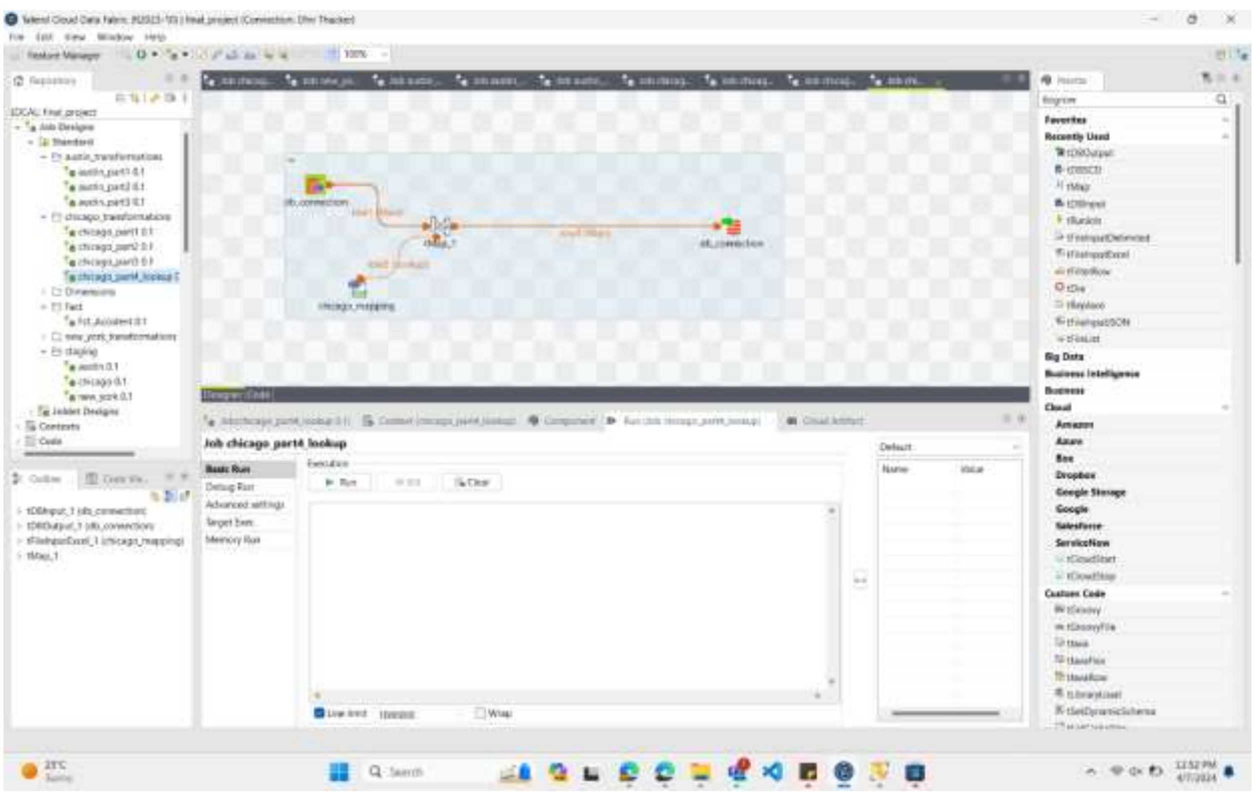
Handling null values, combining columns and normalizing

➔ AUSTIN

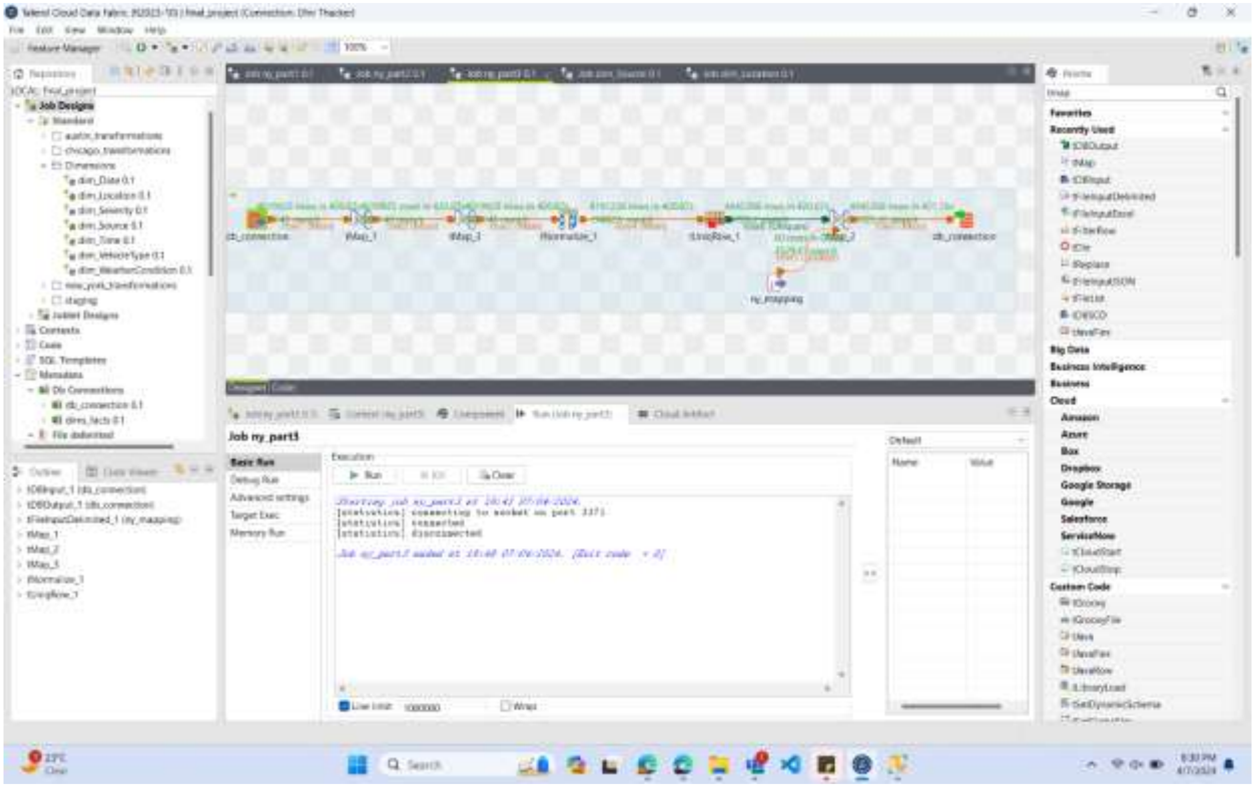


➔ CHICAGO



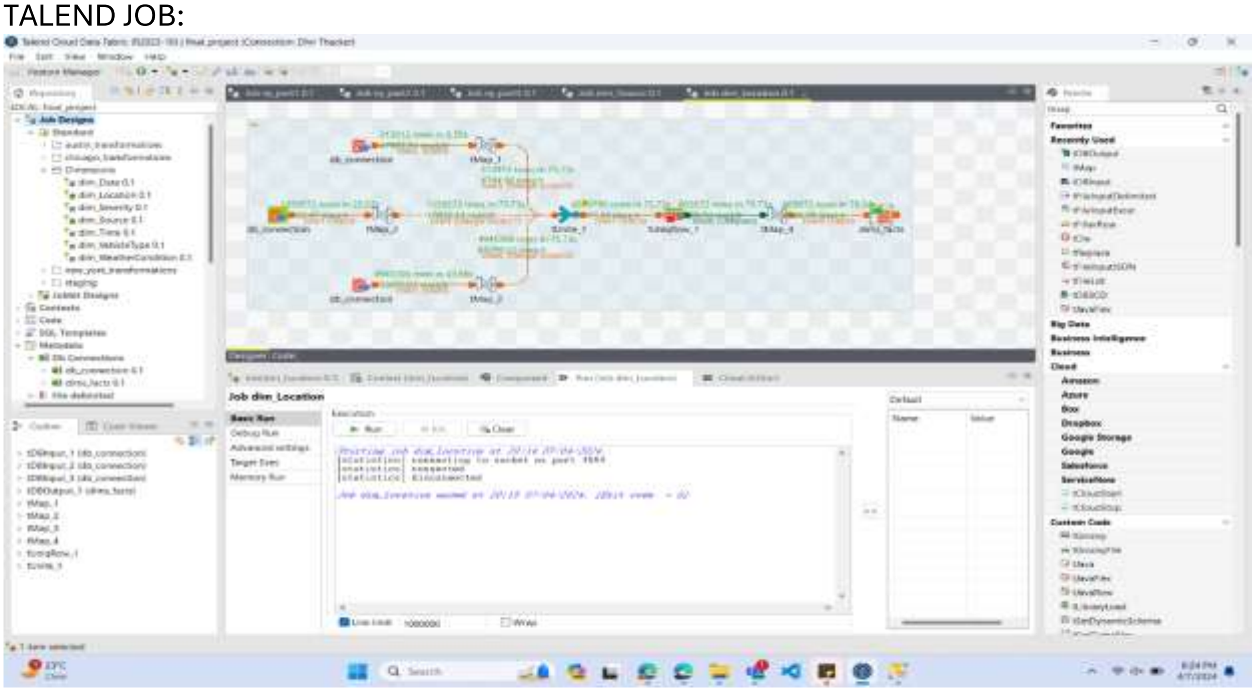


➔ NYC

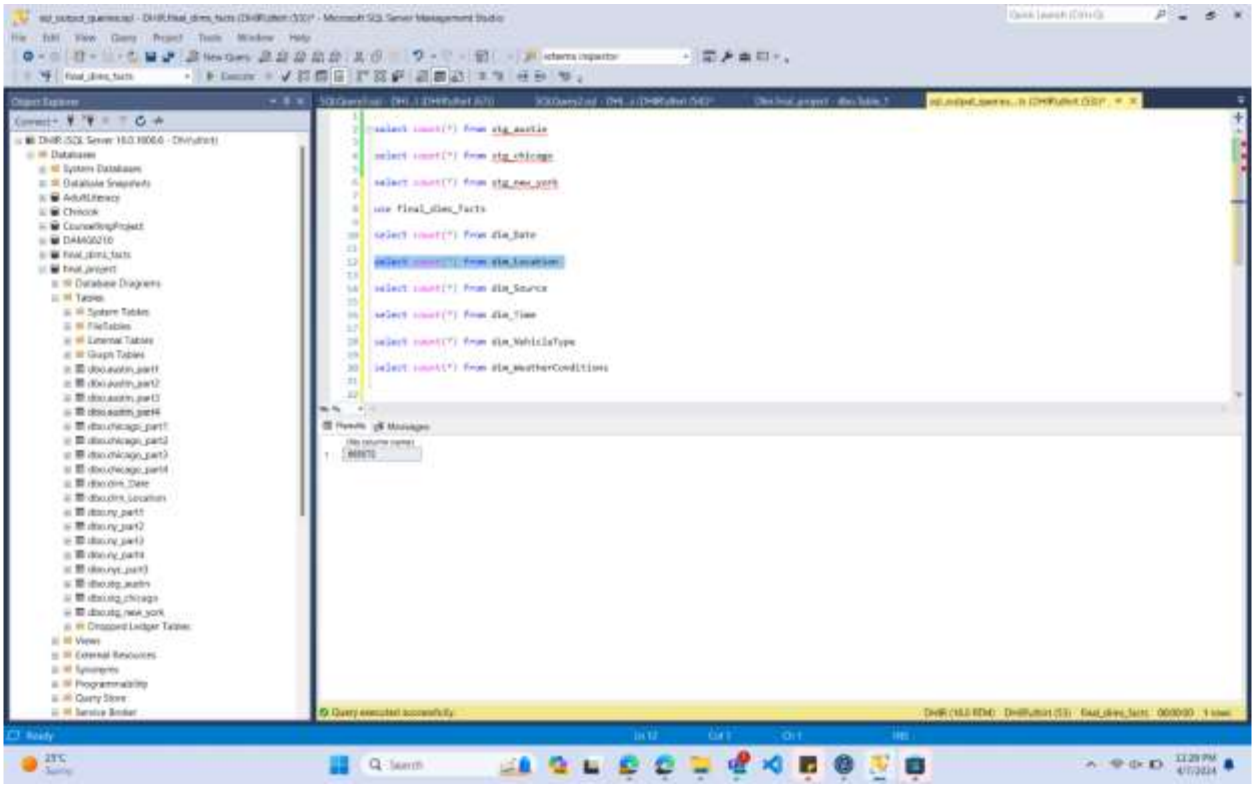


LOADING INTO FACT TABLE AND DIMS

1) Dim_Location
TALEND JOB:

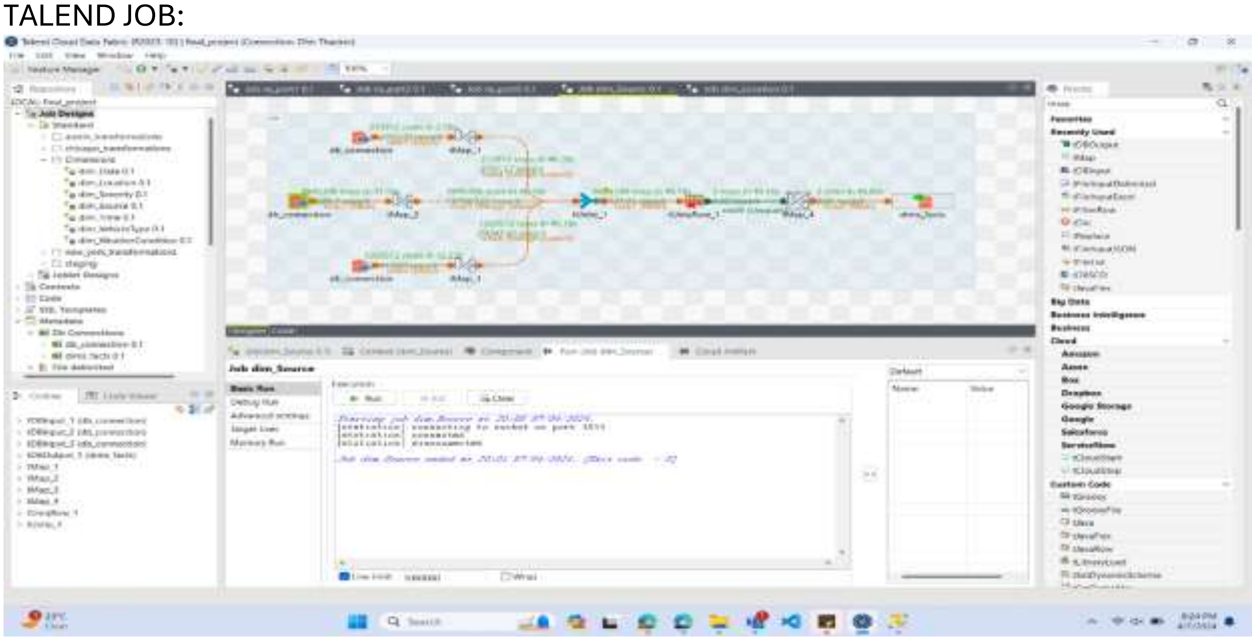


SQL OUTPUT:

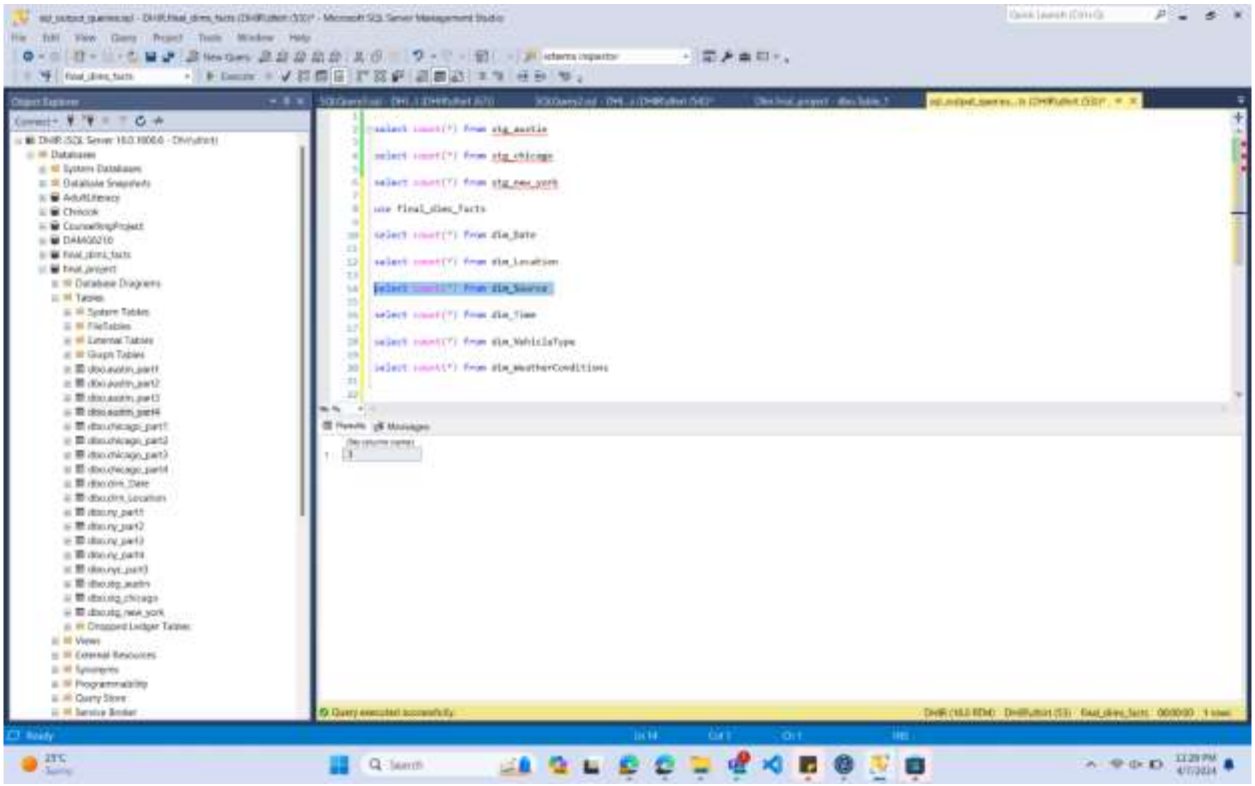


2) Dim_Source

TALEND JOB:

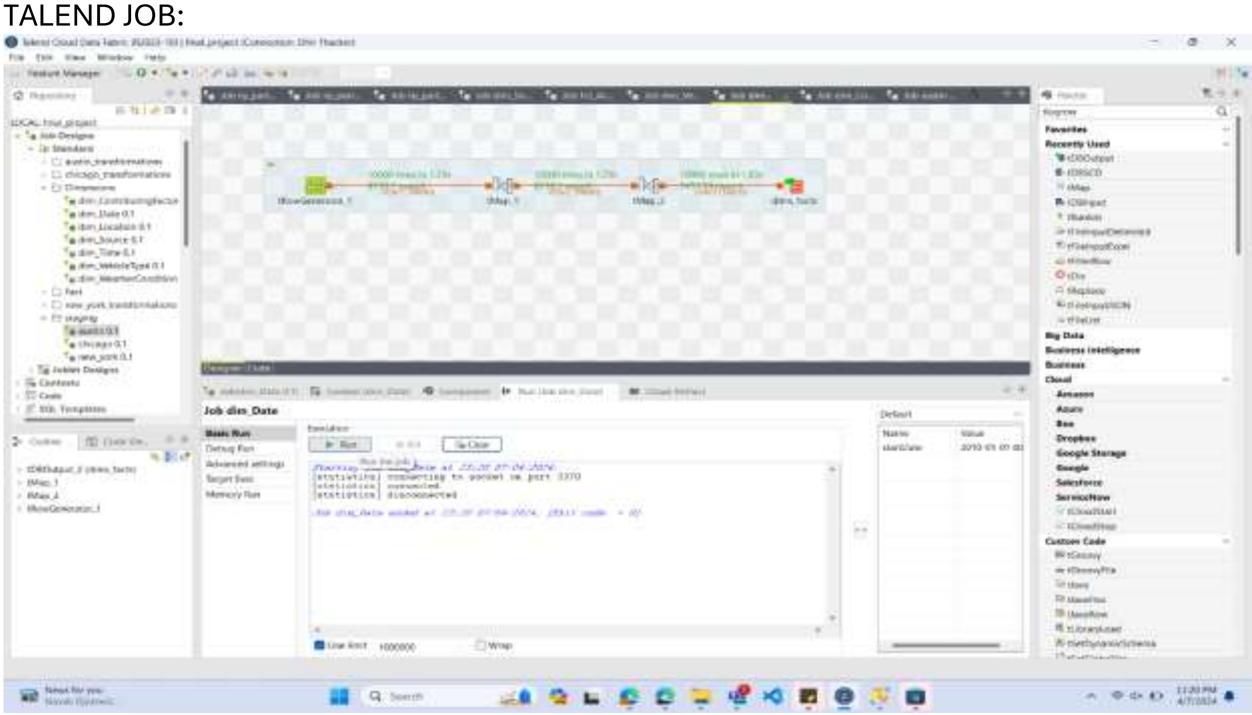


SQL OUTPUT:

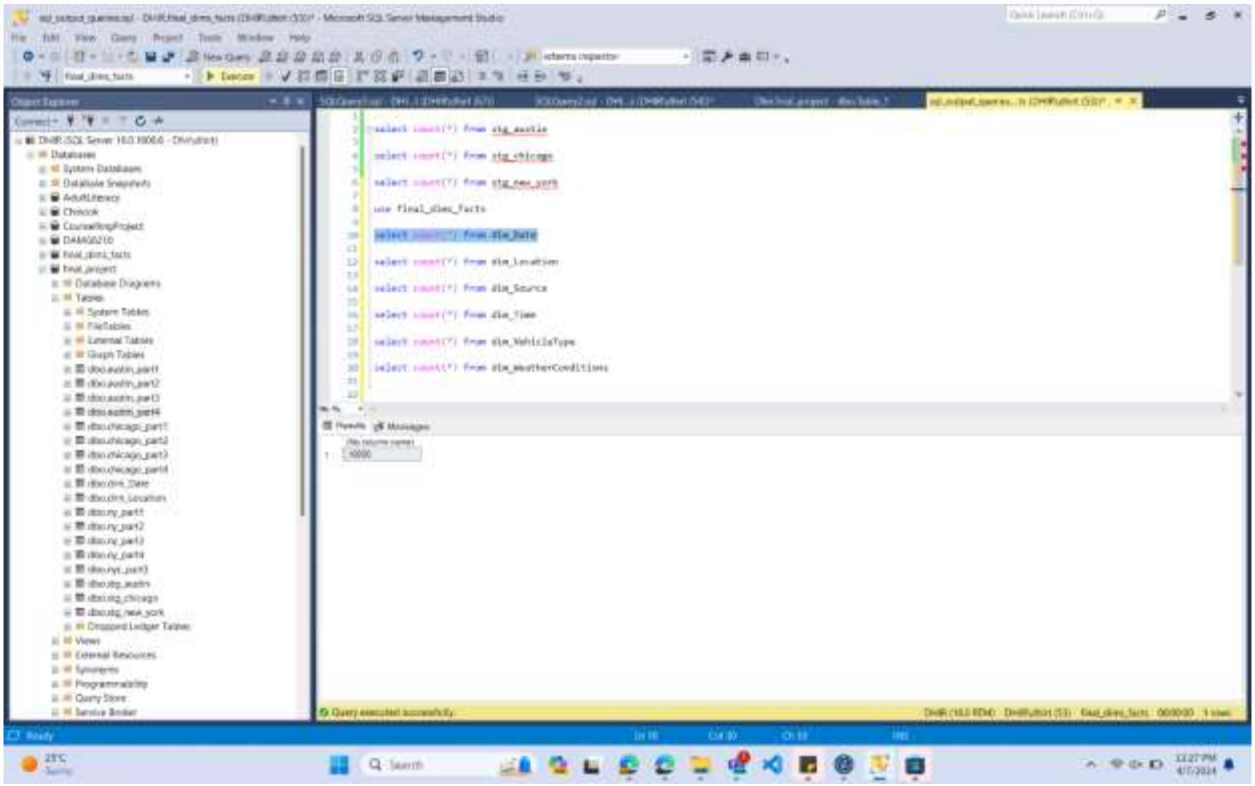


3)Dim_Date

TALEND JOB:

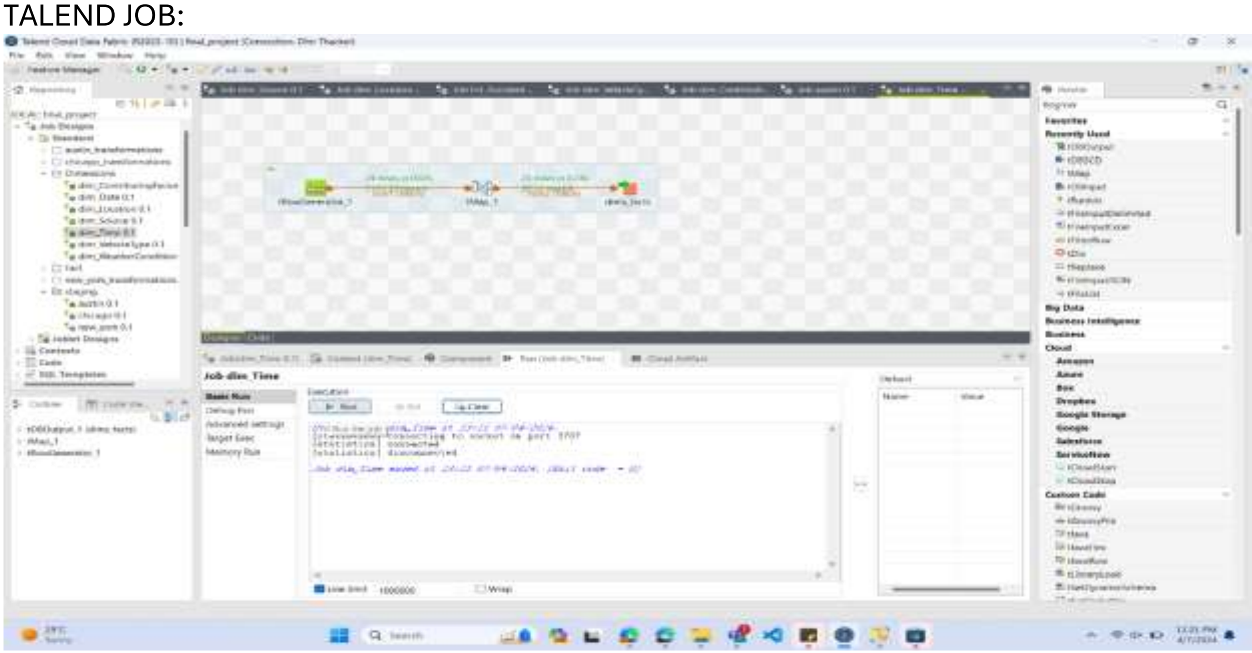


SQL OUTPUT:

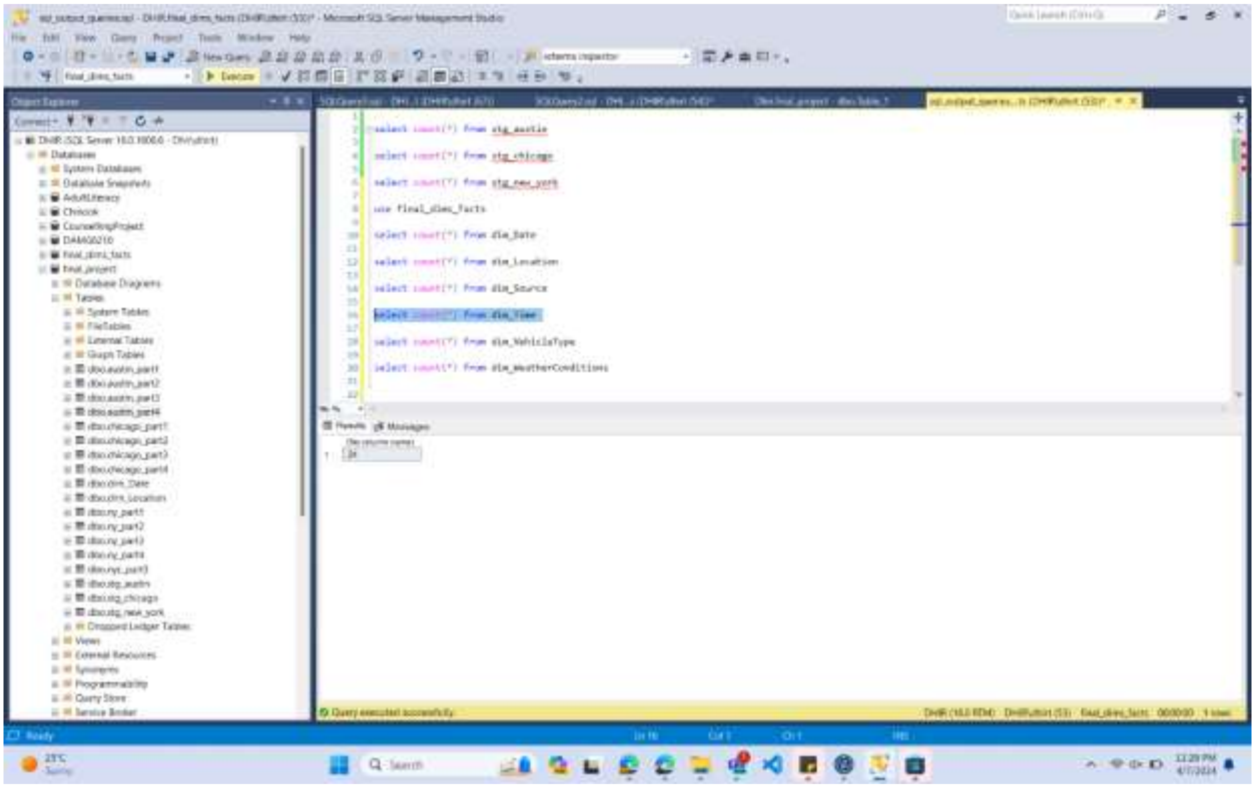


4)Dim_Time

TALEND JOB:

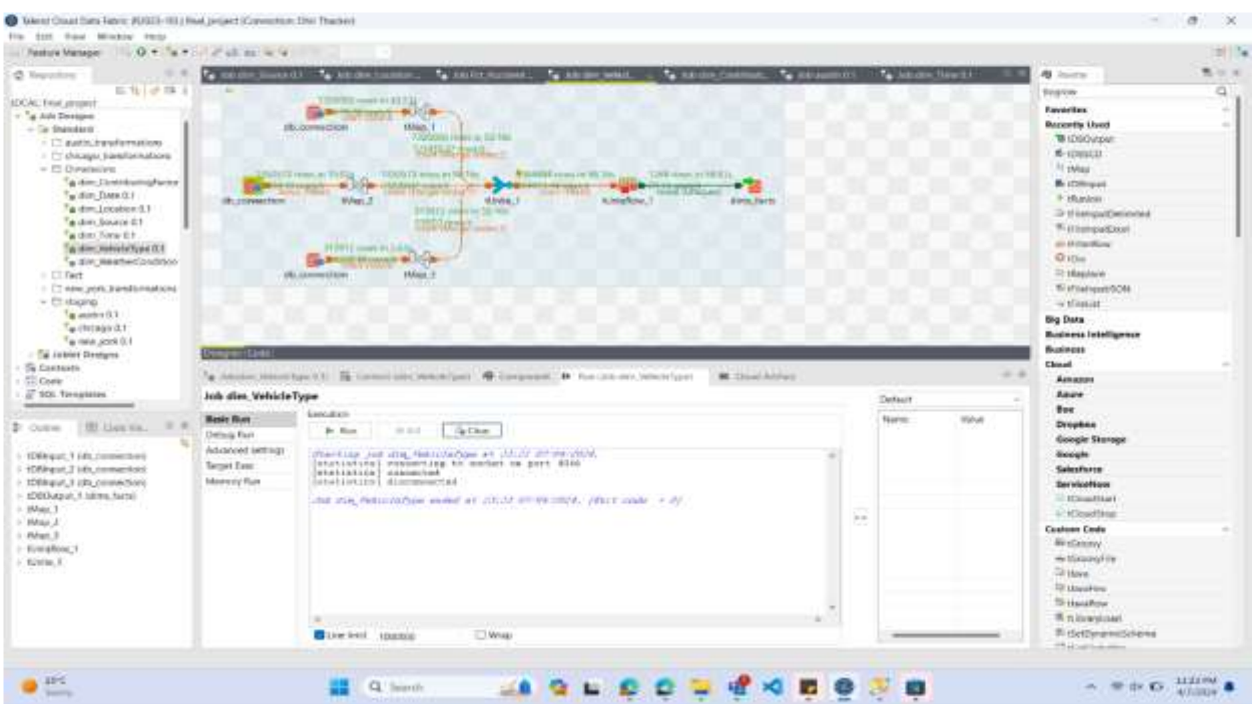


SQL OUTPUT:

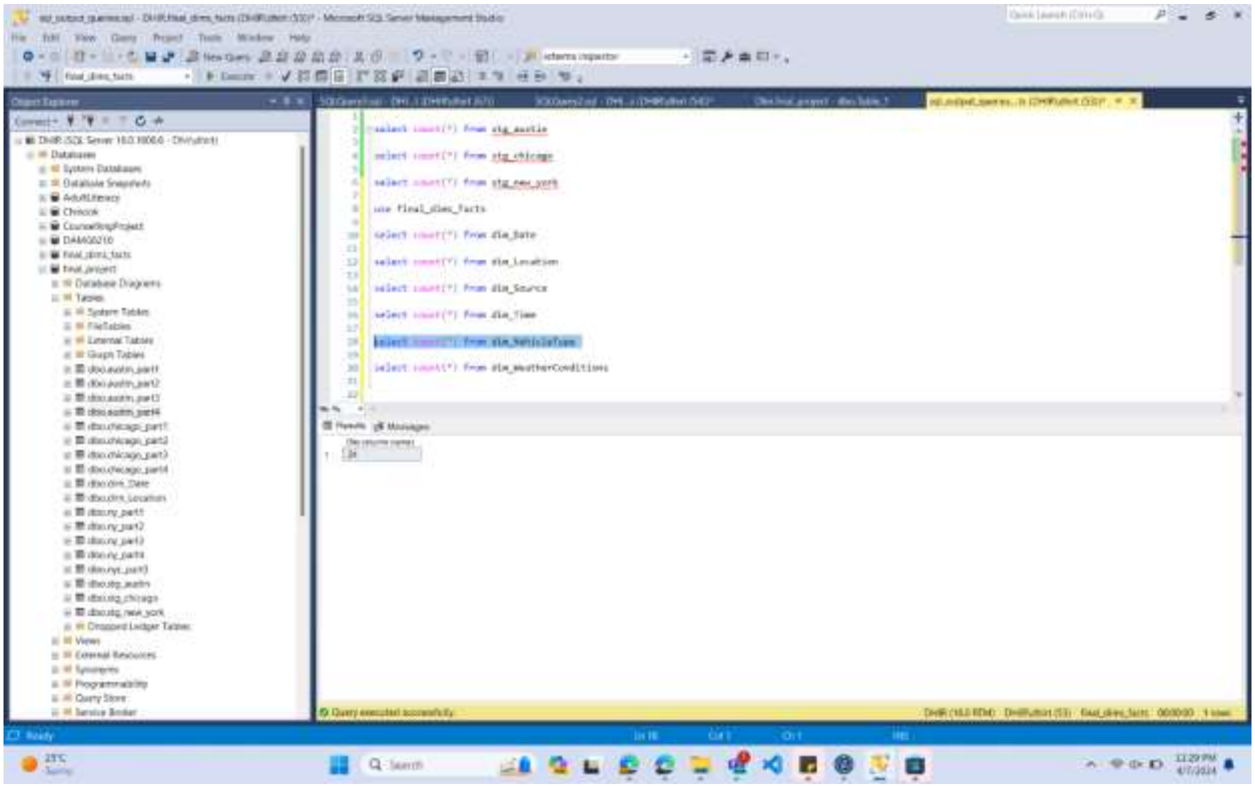


5)Dim_VehicleType

TALEND JOB:

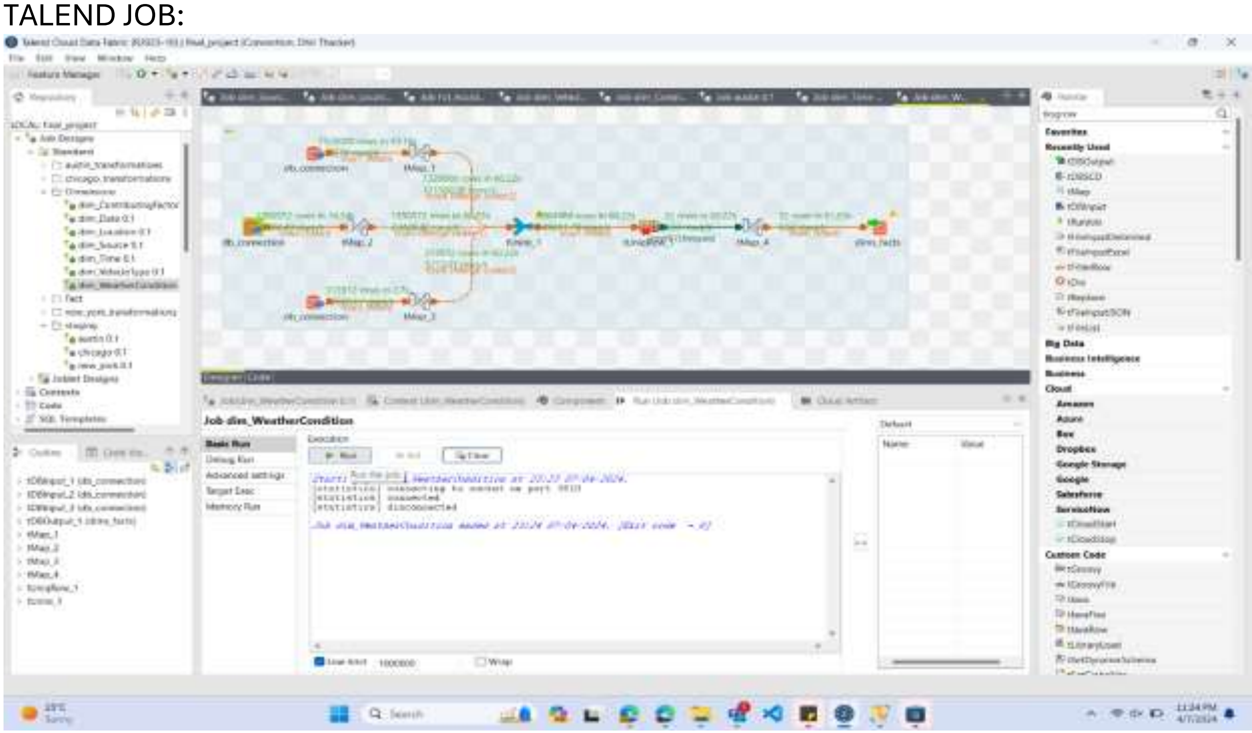


SQL OUTPUT:

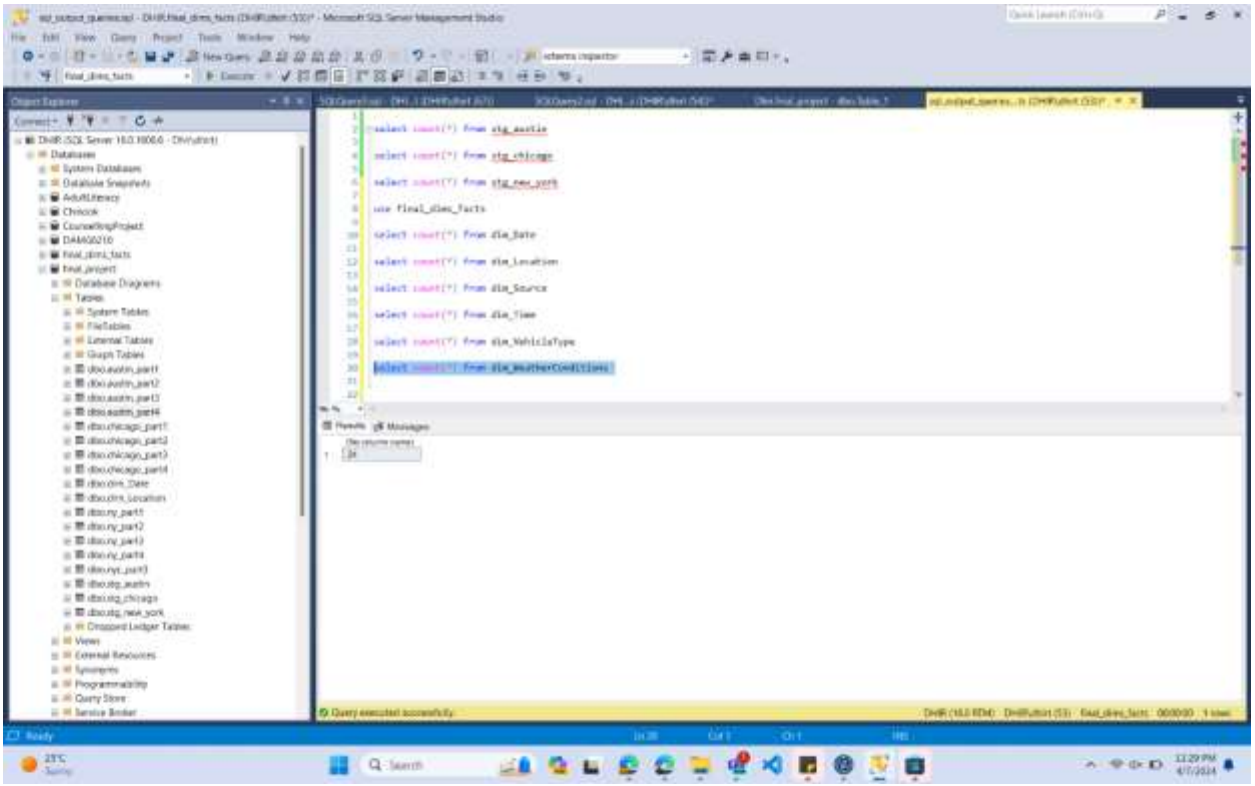


6)Dim_WeatherCondition

TALEND JOB:

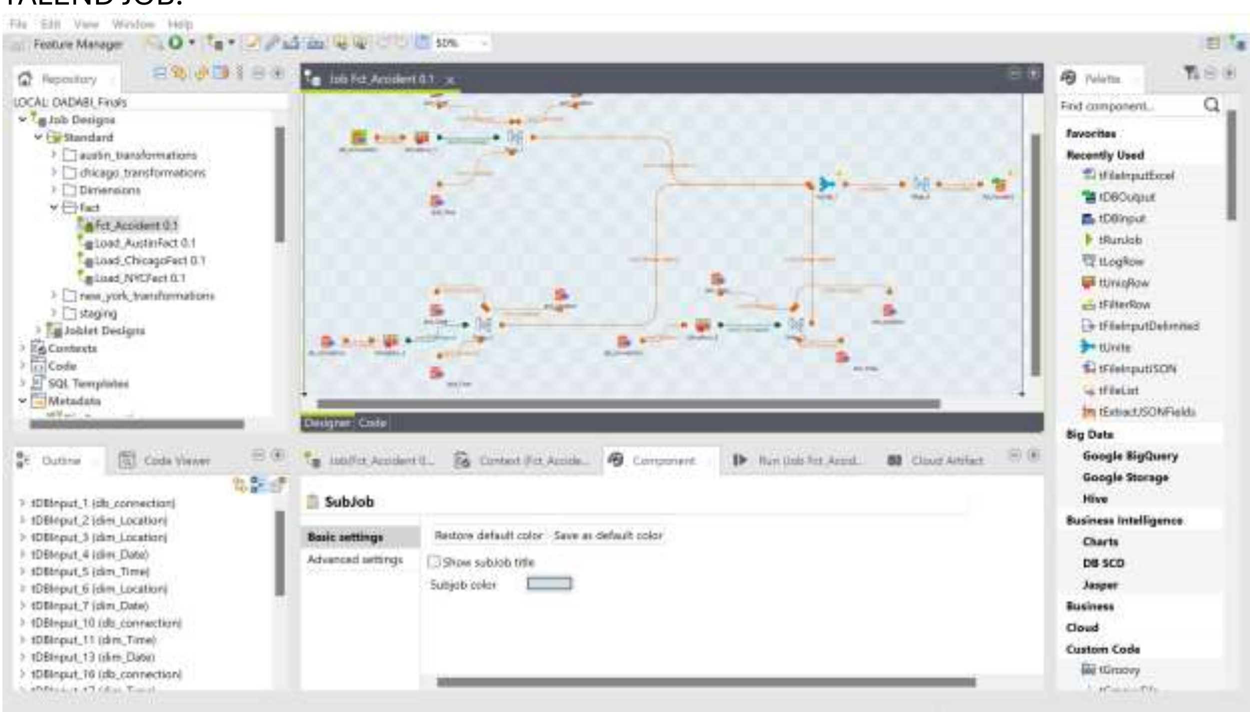


SQL OUTPUT:

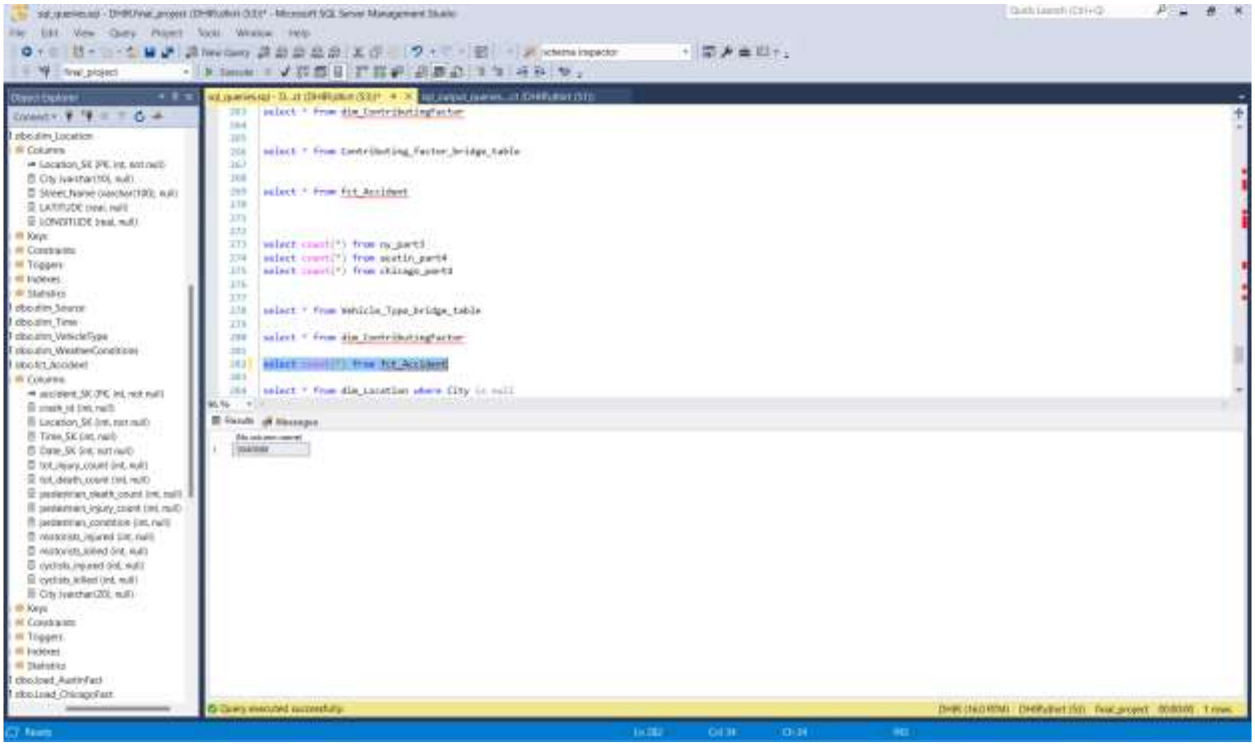


7)Fct_Accident

TALEND JOB:

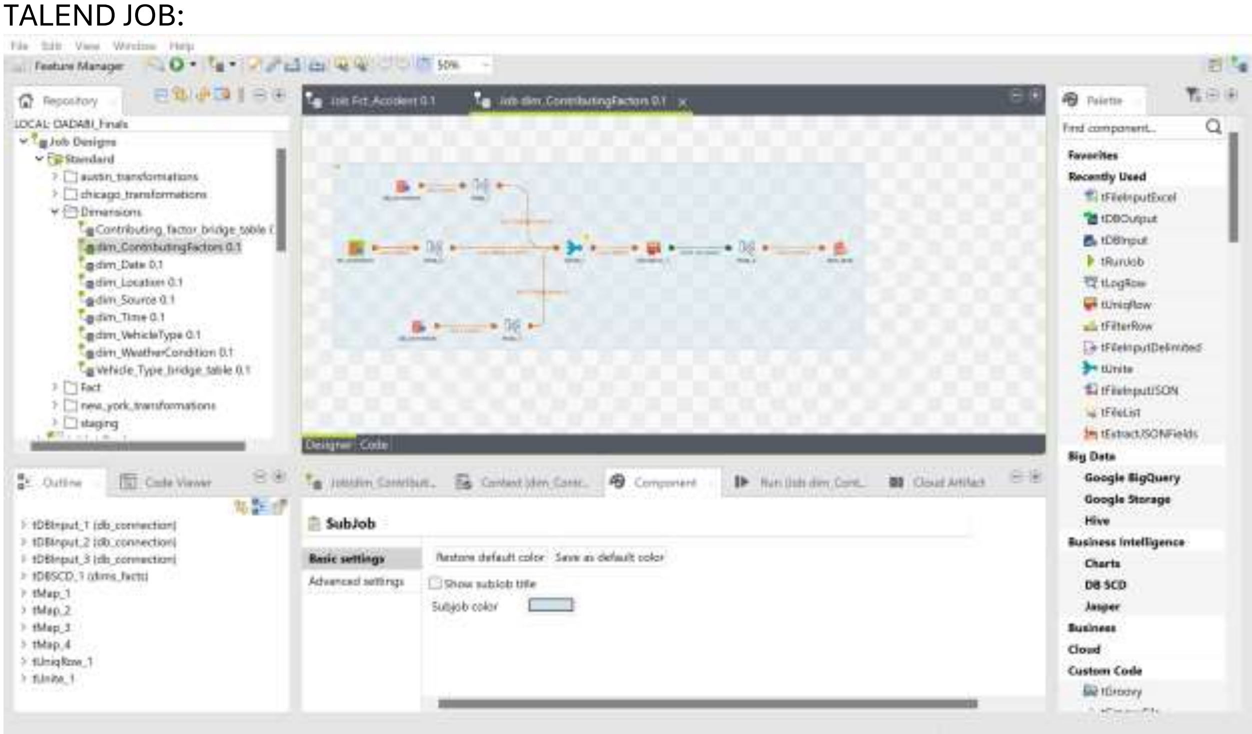


SQL:

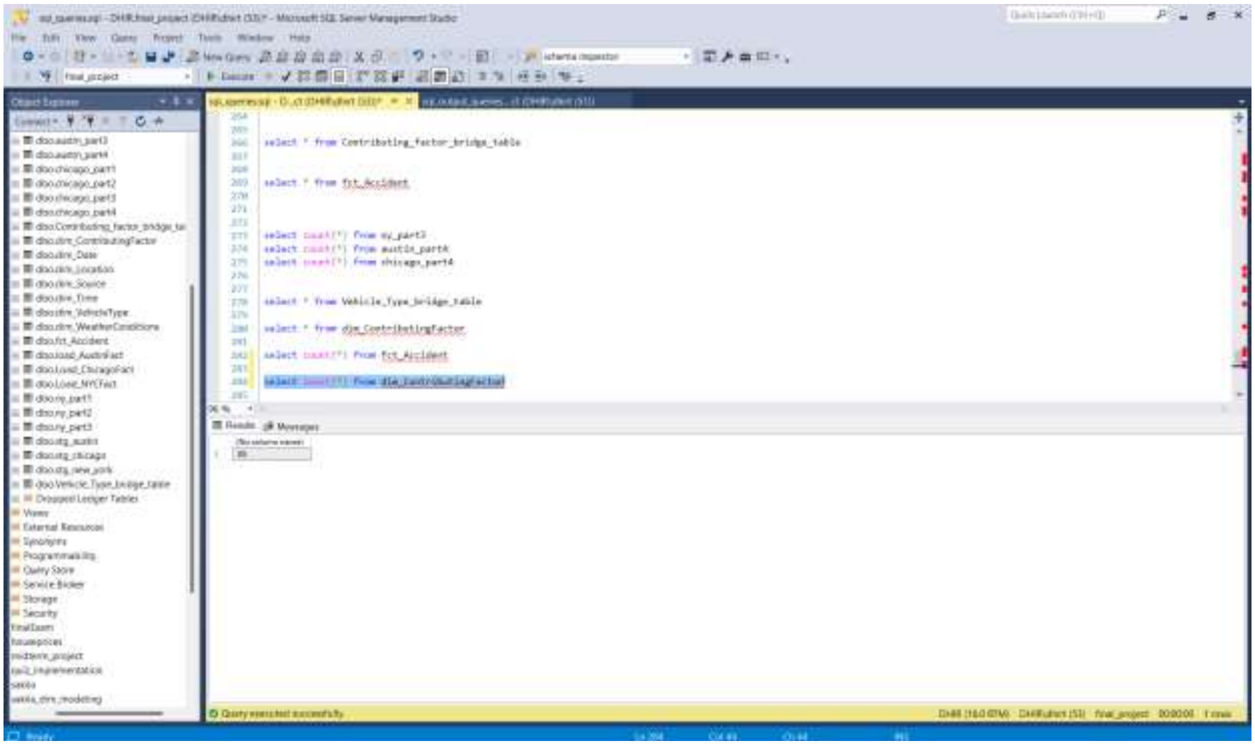


8) Dim_ContributingFactors

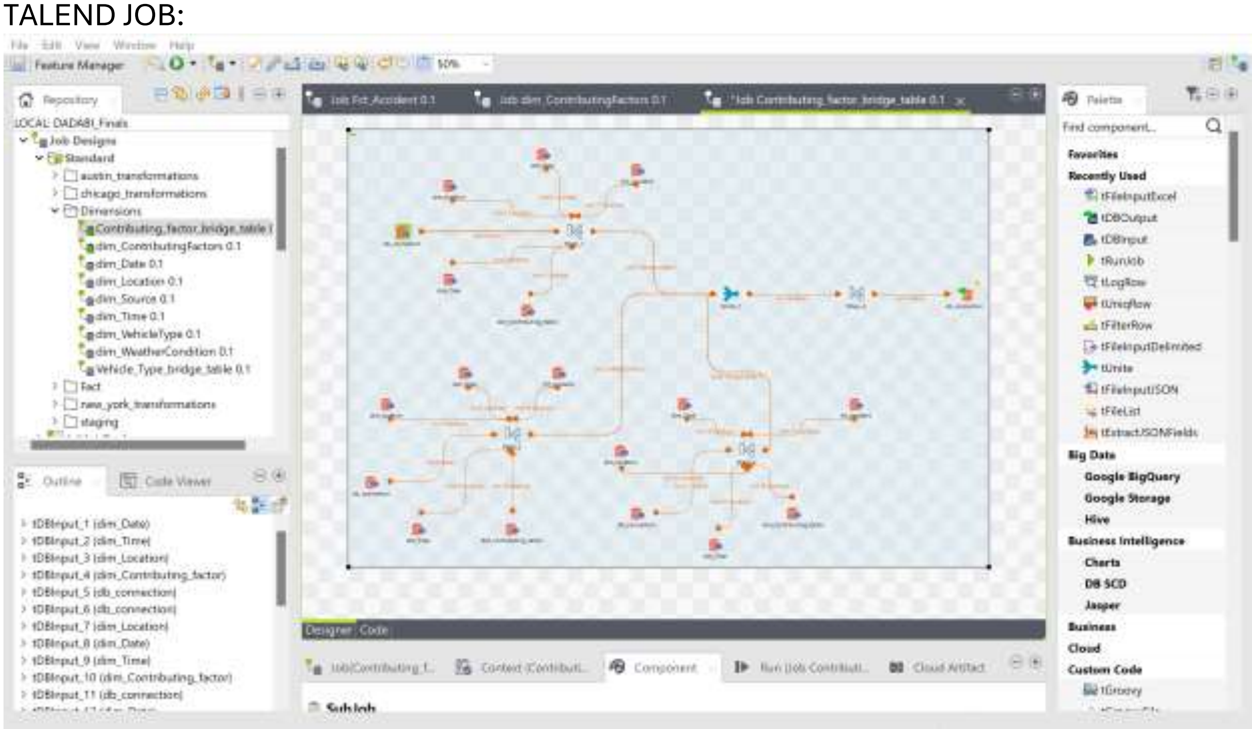
TALEND JOB:



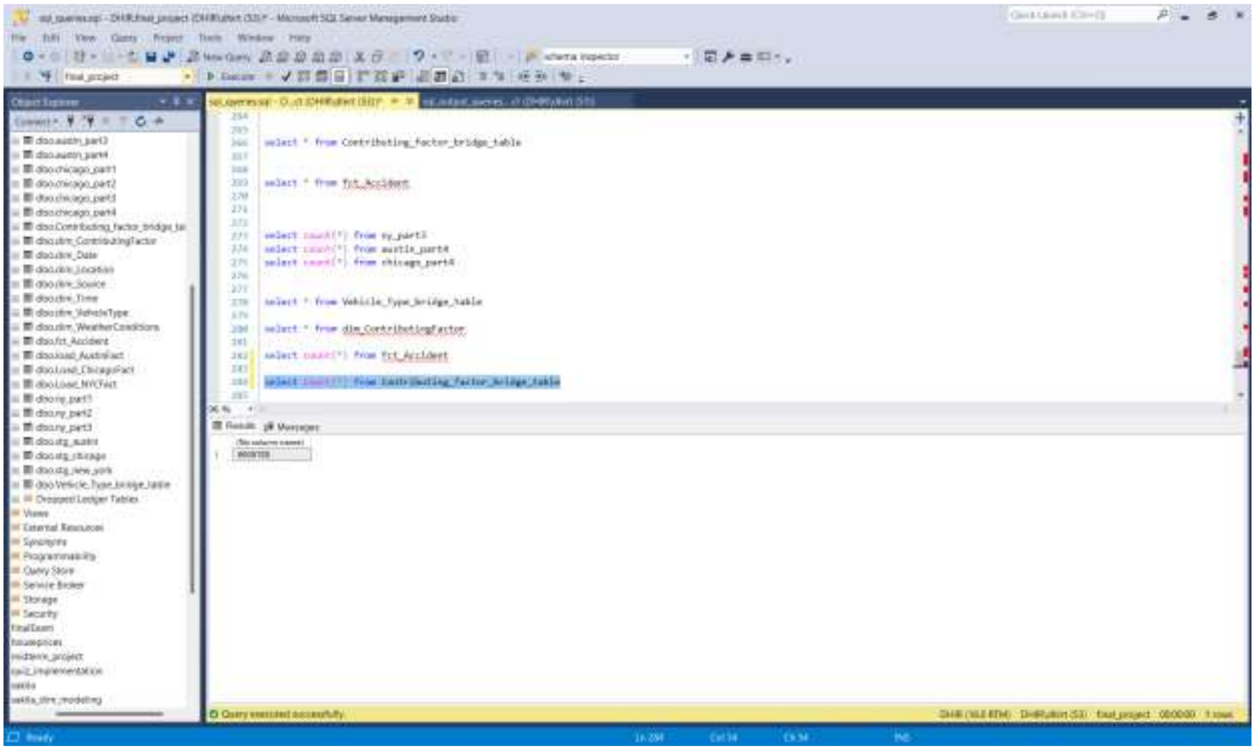
SQL:



9) Contributing Factors Bridge Table
TALEND JOB:

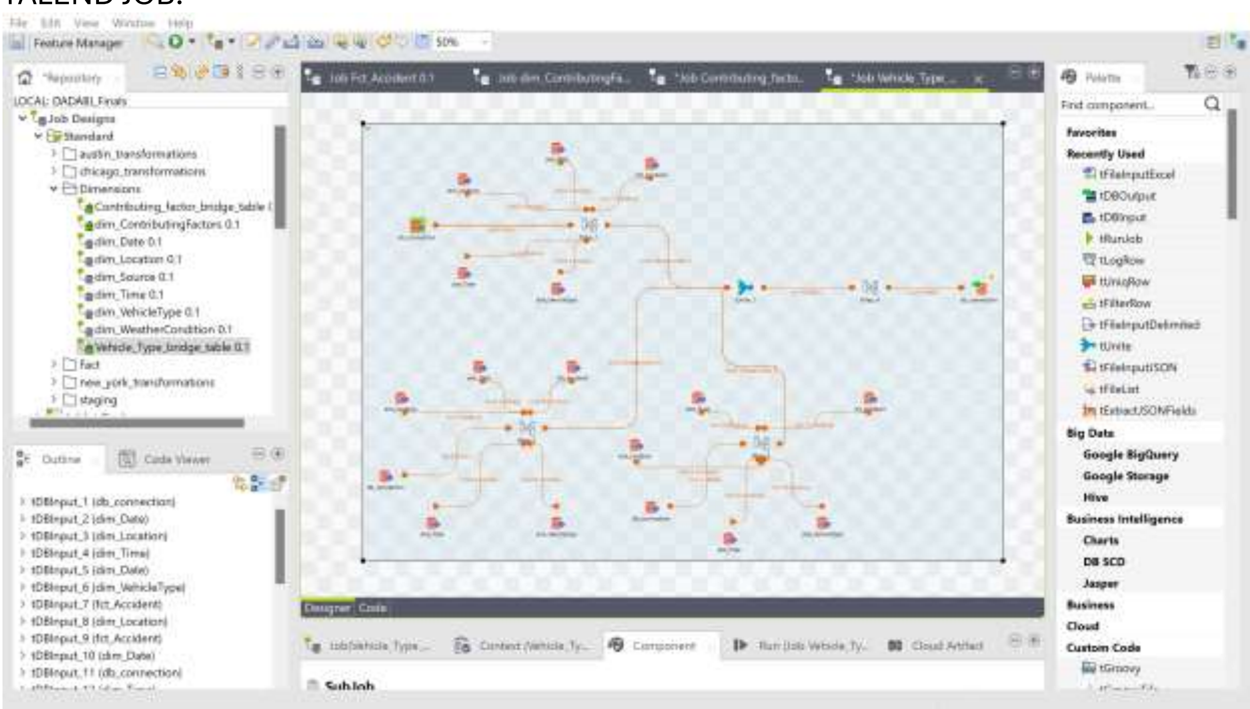


SQL:

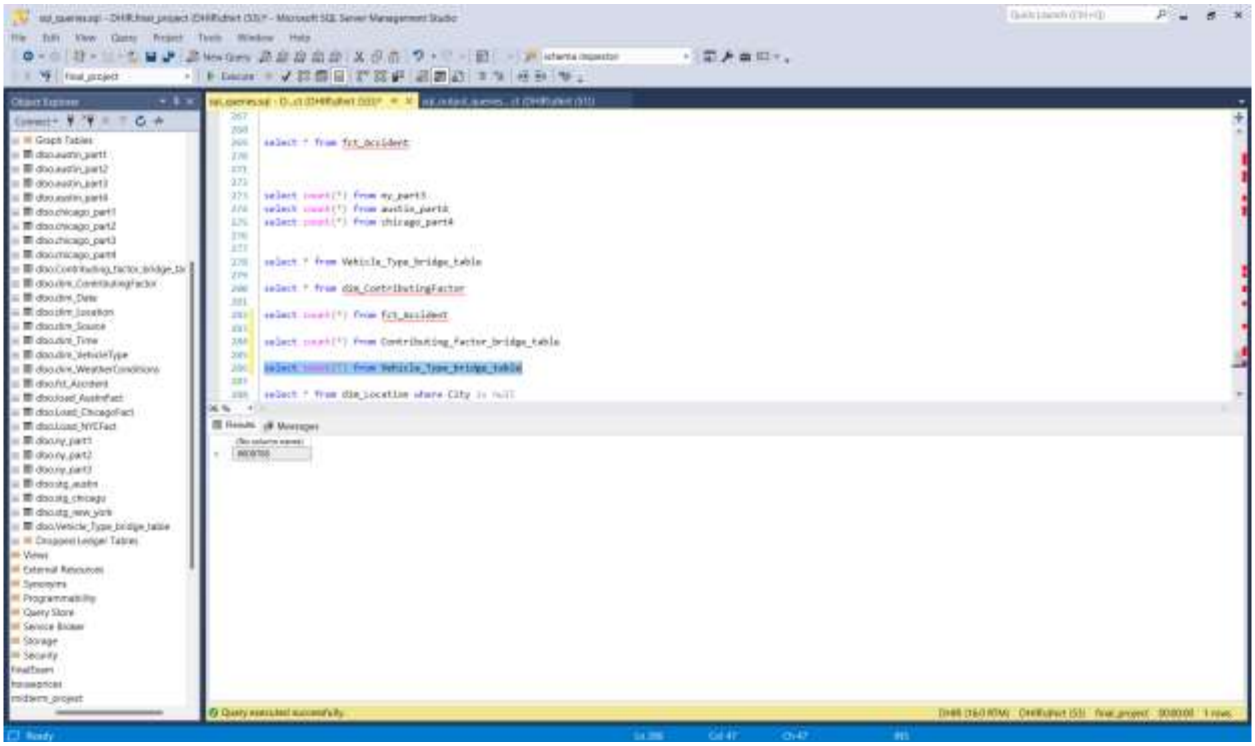


10) Vehicle Type Bridge Table

TALEND JOB:



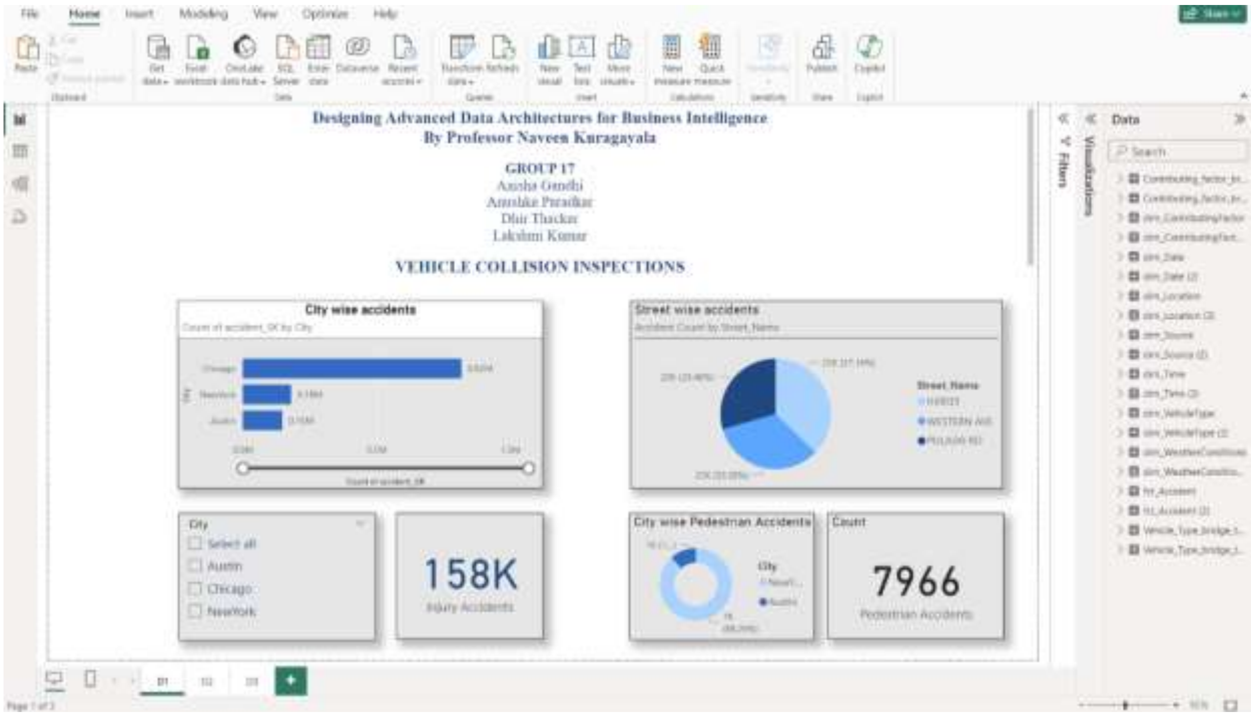
SQL:


Microsoft SQL Server Management Studio
select * from fct_accident
select count(*) from ny_part3
select count(*) from austin_part4
select count(*) from chicago_part4
select * from Vehicle_Type_bridge_table
select * from dim_ContributingFactor
select count(*) from fct_accident
select count(*) from Contributing_Factor_bridge_table
select count(*) from Vehicle_Type_bridge_table
select * from dim_Location where City is null
Results
Messages
Query executed successfully.
Ready

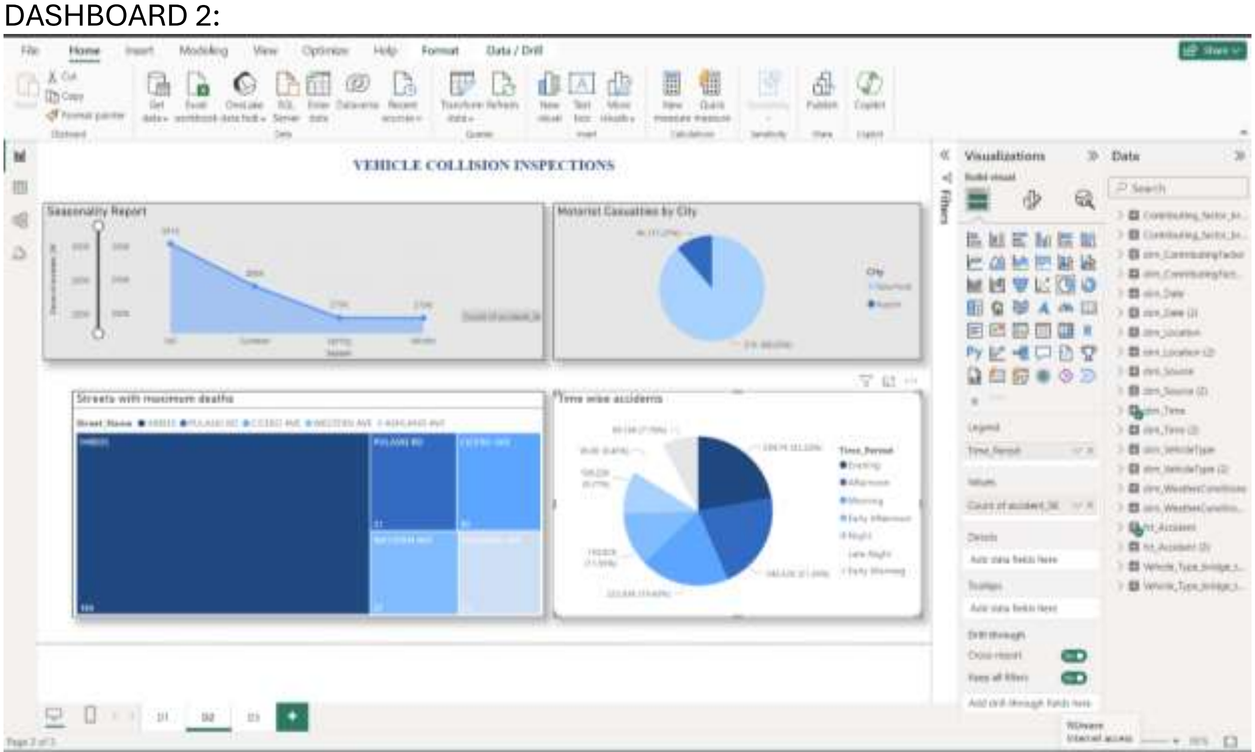
VISUZALIZATIONS:

POWERBI:

DASHBOARD 1:



DASHBOARD 2:



DASHBOARD 3:

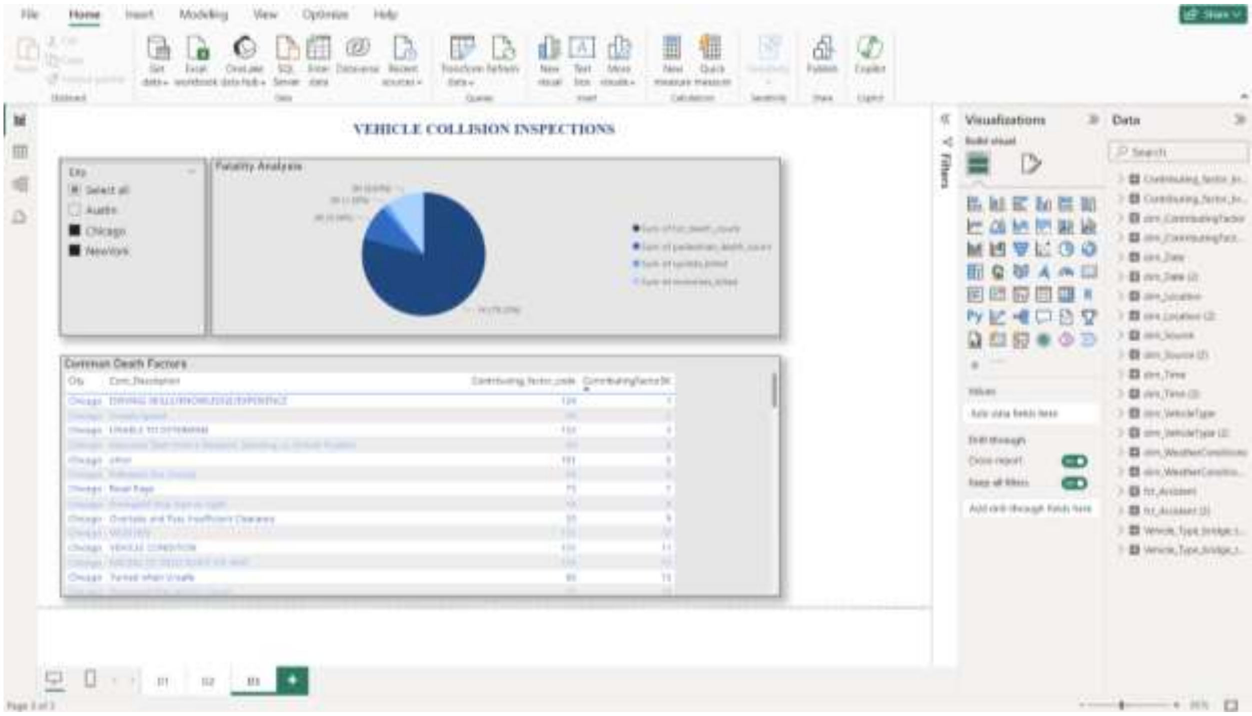
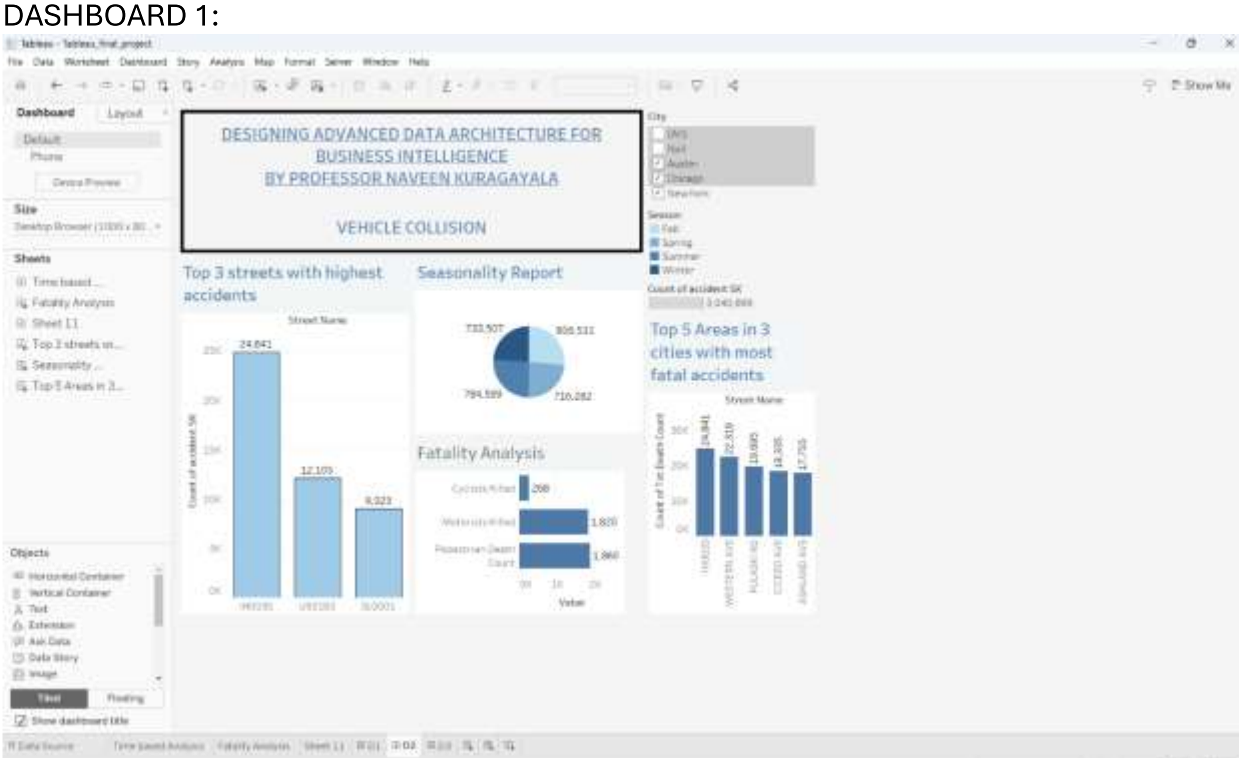
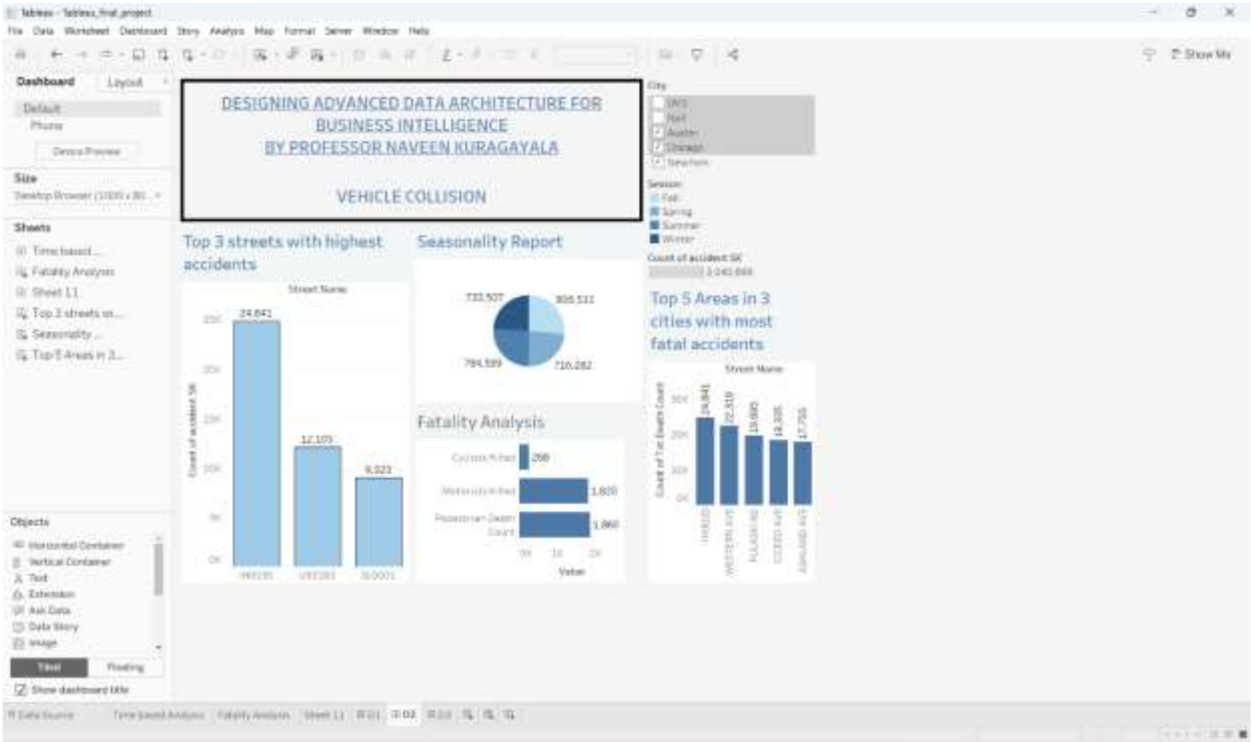


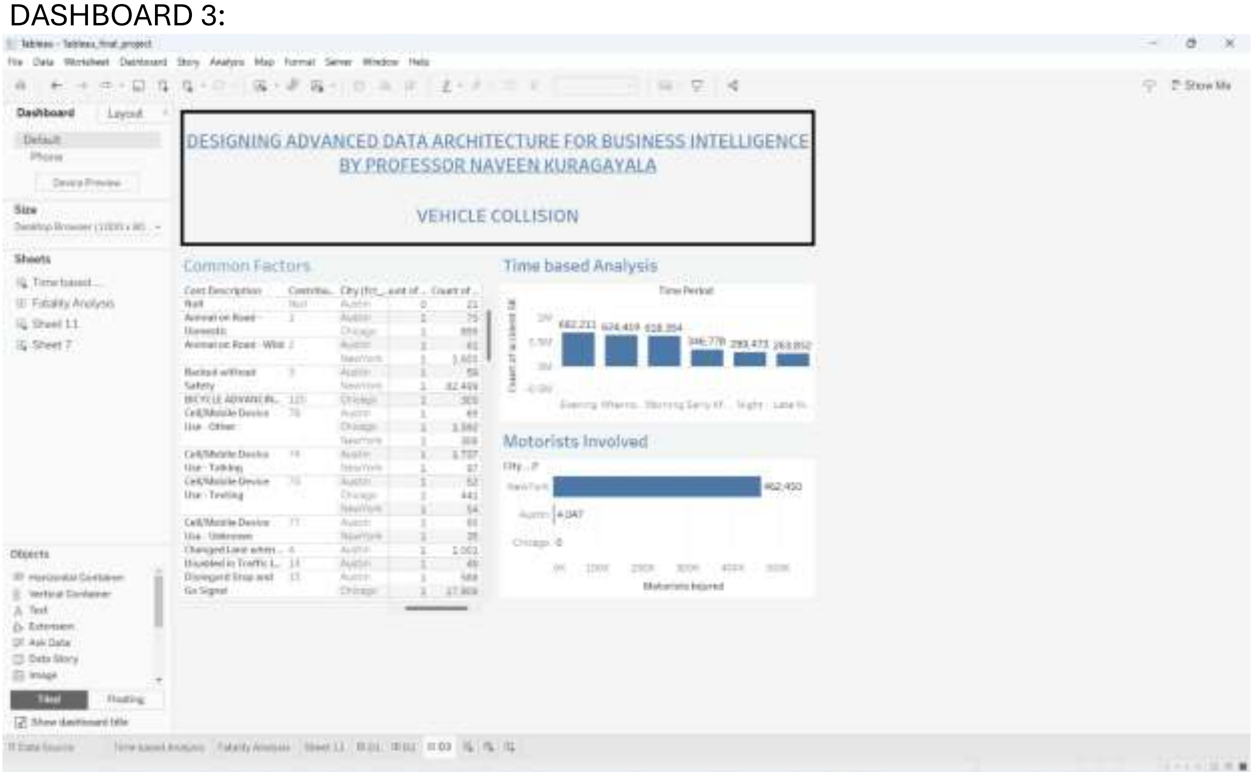
TABLEAU
DASHBOARD 1:



DASHBOARD 2:



DASHBOARD 3:



CHANGE REQUEST IMPLEMENTATION: VISUALIZATION