Milestone 5: Data Warehousing

Define at least 2 ETL workflows that combine and/or transform the external data sources. Provide screenshots of the ETL workflows, and describe the ETL workflows (components used and what they do). Include this in your final report.

Data Sources

The system uses three main data sources:

1. SPD Crime Data (Seattle Police Department Crime Data)
   - Source URL: [SPD Crime Data: 2008-Present | City of Seattle Open Data portal]()

Data Contents:
- Report Number — Unique identifier for each crime report
- Report DateTime — When the report was filed
- Offense ID — Unique offense identifier
- Offense Date — When the crime occurred
- NIBRS Group AB — Classification (Group A or B)
- NIBRS Crime Against Category — Category (Person, Property, Society)
- Offense Sub Category — Specific subcategory
- Offense Category — Parent group classification
- NIBRS Offense Code Description — Full offense name
- NIBRS_offense_code — Standardized offense code
- Block Address — Blurred/block-level address (privacy protection)
- Latitude/Longitude — Blurred coordinates (privacy protection)
- Precinct — Police precinct
- Sector — Police sector
- Beat — Police beat
- Neighborhood — MCPP neighborhood designation

How We Use This Data**:**
- Historical crime analysis: Load into crime_reports, report_offenses, and offense_types
- Risk scoring: Map incidents to street segments for ML-based risk calculation
- Crime visualization: Display on interactive map with filtering by type, time and location
- Route safety: Compute route risk scores by aggregating segment-level risks
- Temporal analysis: Support time-based filtering (24h, 7d, 30d, 90d, custom ranges)

2. Seattle Fire Real-Time 911 (Real-time 911 incidents)
   - Source URL: [Seattle Real Time Fire 911 Calls | City of Seattle Open Data portal]()

Data Contents:
- Incident Number — Unique identifier for each 911 call
- Type — Incident type (fire, medical, etc.)
- Datetime/DateTime — When the incident occurred
- Address — Location address
- Latitude/Longitude — Precise coordinates
- Report Location — Additional location information

How We Use This Data:

- Real-time alerts: Store in realtime_incidents for immediate safety notifications
- Live incident overlay: Display active incidents on the map
- Route adjustments: Consider active incidents when calculating route safety
- Temporal risk weighting: Weight recent incidents more heavily in risk calculations
- Emergency awareness: Provide up-to-date information about ongoing incidents

3. Seattle Streets Data (Geographic street data)
    - Source URL: [Seattle Streets]

Data Contents:

- UNITID — Unique street segment identifier
- ONSTREET — Street name
- INTKEYLO/INTKEYHI — Intersection keys (start/end points)
- INTRLO/INTRHI — Intersection names
- DIRLO/DIRHI — Direction indicators
- GIS_MID_X/GIS_MID_Y — Center point coordinates (longitude/latitude)
- SPEEDLIMIT — Speed limit
- ARTCLASS — Arterial classification
- STATUS — Street status
- SEGLENGTH — Segment length (meters)
- SURFACEWIDTH — Surface width
- SLOPE_PCT — Slope percentage
- OWNER — Ownership information
- ONEWAY — One-way indicator
- FLOW — Traffic flow direction

How We Use This Data:

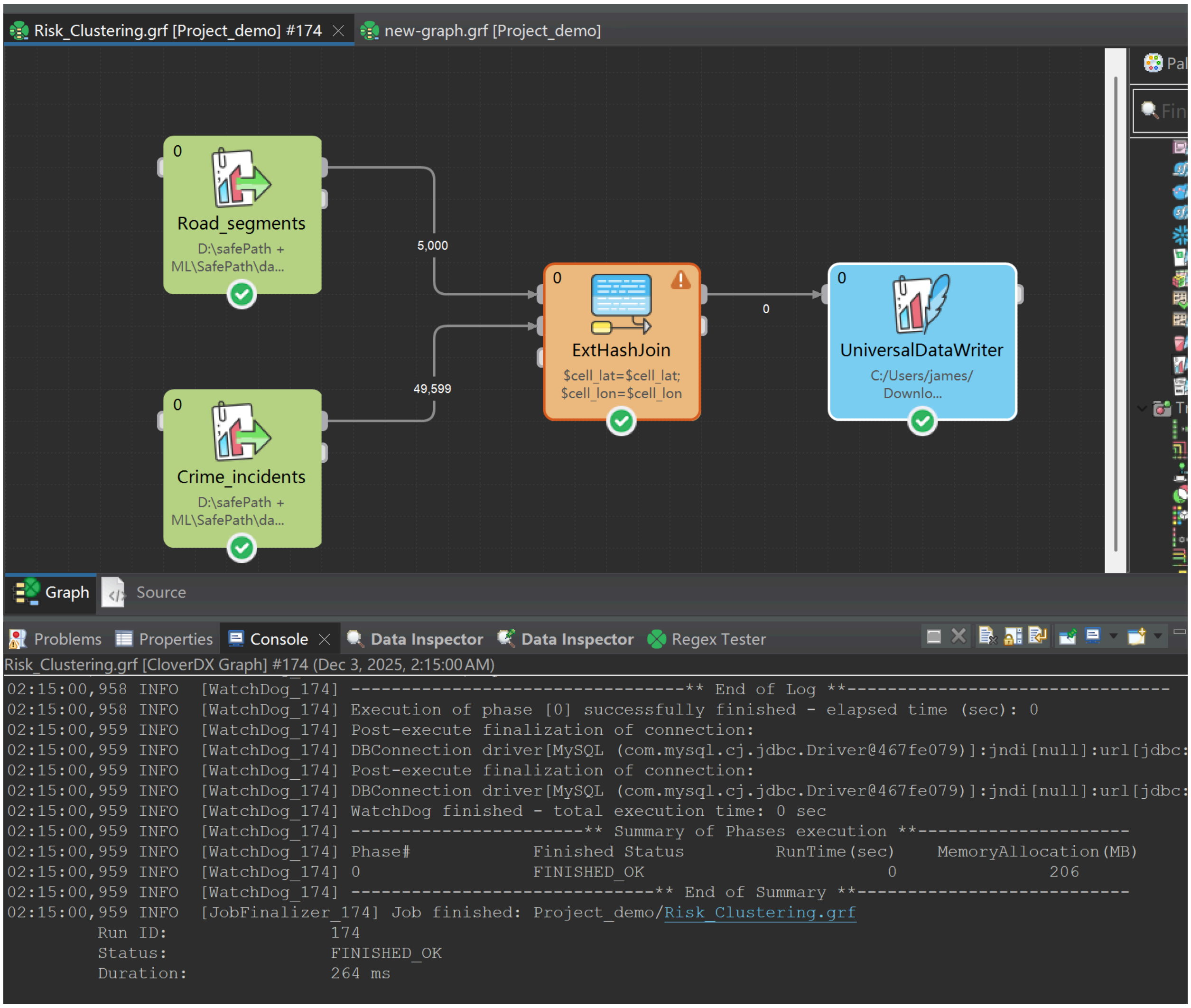
- Street network: Populate street_segments and intersections
- Spatial matching: Match route coordinates to street segments using Haversine distance
- Risk mapping: Associate crime incidents with specific street segments
- Route analysis: Map Google Directions polylines to our street network
- Geographic context: Provide street names and intersection information for route display

ETL workflows

Workflow 1: Risk Score Clustering

Components used and what they do:

- CSVReader – Road_segments: Reads the external road network CSV file and exposes road attributes (ID, name, location, …) as metadata.
- CSVReader – Crime_incidents: Reads the external SPD crime CSV file and exposes incident attributes (type, time, location, …).
- ExtHashJoin: Joins the two input streams on latitude/longitude, producing combined records that link roads with nearby crime incidents.
- FlatFileWriter / DatabaseWriter: Stores the joined dataset as a new table or CSV file in the data warehouse.

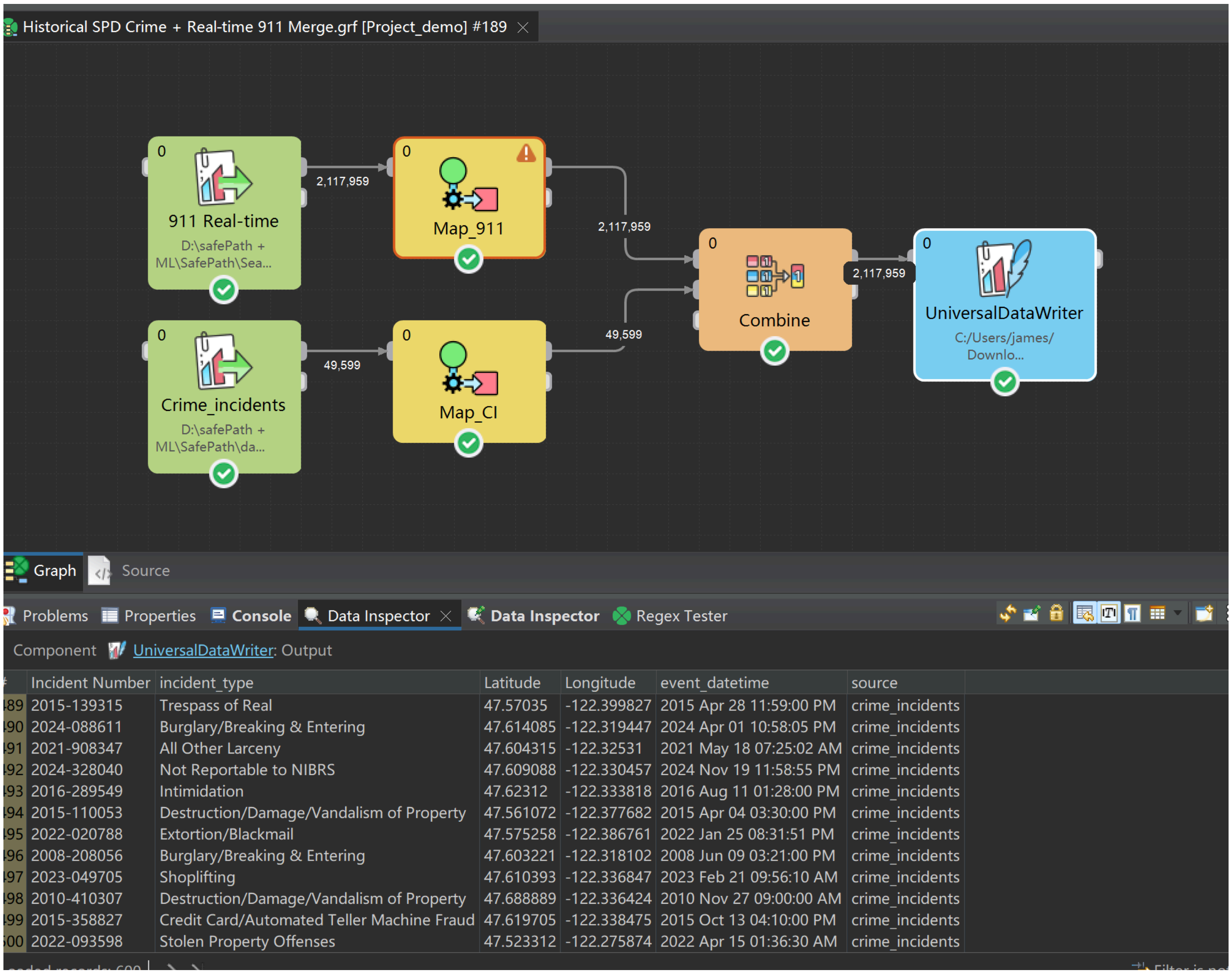


Return 0 because no matching latitude and longitude in the sample tables.

Workflow 2: Historical SPD Crime + Real-time 911 Merge

Components used and what they do:

- CSVReader – SPD_Crime: Reads the historical SPD crime CSV and exposes report-level fields.
- CSVReader – RT_911: Reads the Seattle Real-Time 911 calls CSV.
- Reformat (SPD): Maps SPD columns into the unified incident schema and tags records as SPD_HISTORY.
- Reformat (911): Maps 911 columns into the same schema and tags records as REALTIME_911.
- Union: Merges the two standardized incident streams into one.
- FlatFileWriter: Writes the merged incidents to a CSV file for downstream analysis.



Incident Number	incident_type	Latitude	Longitude	event_datetime	source
2015-139315	Trespass of Real	47.57035	-122.399827	2015 Apr 28 11:59:00 PM	crime_incidents
2024-088611	Burglary/Breaking & Entering	47.614085	-122.319447	2024 Apr 01 10:58:05 PM	crime_incidents
2021-908347	All Other Larceny	47.604315	-122.32531	2021 May 18 07:25:02 AM	crime_incidents
2024-328040	Not Reportable to NIBRS	47.609088	-122.330457	2024 Nov 19 11:58:55 PM	crime_incidents
2016-289549	Intimidation	47.62312	-122.333818	2016 Aug 11 01:28:00 PM	crime_incidents
2015-110053	Destruction/Damage/Vandalism of Property	47.561072	-122.377682	2015 Apr 04 03:30:00 PM	crime_incidents
2022-020788	Extortion/Blackmail	47.575258	-122.386761	2022 Jan 25 08:31:51 PM	crime_incidents
2008-208056	Burglary/Breaking & Entering	47.603221	-122.318102	2008 Jun 09 03:21:00 PM	crime_incidents
2023-049705	Shoplifting	47.610393	-122.336847	2023 Feb 21 09:56:10 AM	crime_incidents
2010-410307	Destruction/Damage/Vandalism of Property	47.688889	-122.336424	2010 Nov 27 09:00:00 AM	crime_incidents
2015-358827	Credit Card/Automated Teller Machine Fraud	47.619705	-122.338475	2015 Oct 13 04:10:00 PM	crime_incidents
2022-093598	Stolen Property Offenses	47.523312	-122.275874	2022 Apr 15 01:36:30 AM	crime_incidents

Using Excel and/or Google Sheets, create at least 5 charts from your data warehouse (so the charts should reflect the results of ETLs and should utilize the external data). Include this in your final report for each chart:

- *Your hypothesis for combining the data.*
- *The results of combining the data, and if it validates or invalidates your hypothesis.*
- *Briefly describe the chart's significance for your application and the action you could take (if any) given the new information.*

Chats

Chart 1 – Total Incidents Over Time (By Year) (Line Chart): Displays the trend of the total number of incidents over the years present in the data.

Hypothesis: Combining historical SPD crime data and real-time 911 incidents will reveal an overall upward trend in the number of incidents per year, reflecting population growth and increased reporting.

The **result** does not validate the hypothesis.

Conclusion: It is important to notice that the crime incidents has been decreasing since 2022. Therefore, we may need to consider other factors like economic trends or public health etc. in the analysis.

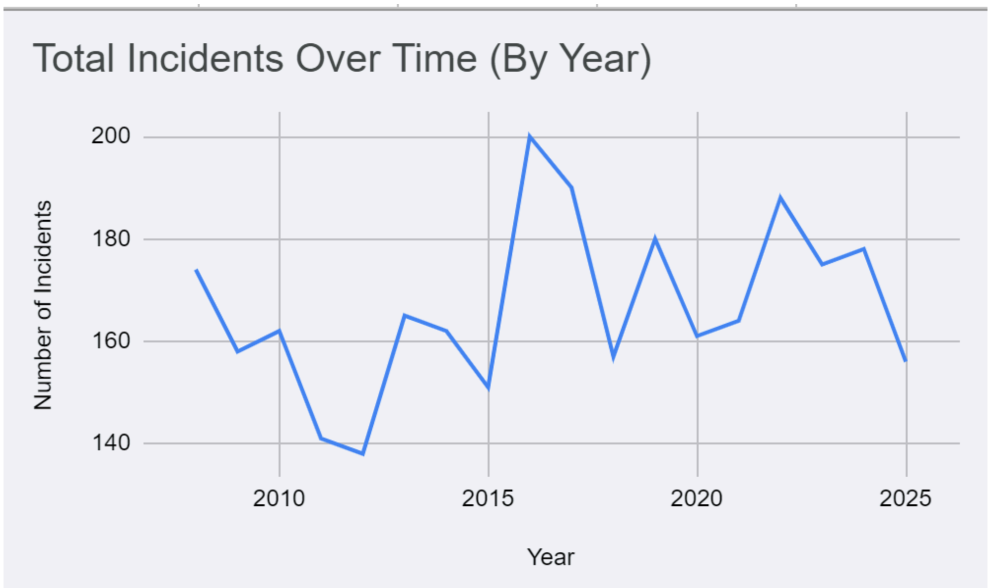


Chart 2 – Incidents by Day of the Week (Bar Chart): Reveals any patterns in incident frequency across the days of the week.

Hypothesis: When all incidents from both sources are combined, weekends (Friday–Sunday) will have higher incident counts than weekdays, due to increased nightlife and activity.

The **result** does not validate the hypothesis.
Conclusion: It’s surprising to see Monday has the most incidents, Day-of-week patterns are useful for time-aware safety guidance. SafePath can highlight elevated risk on specific days.

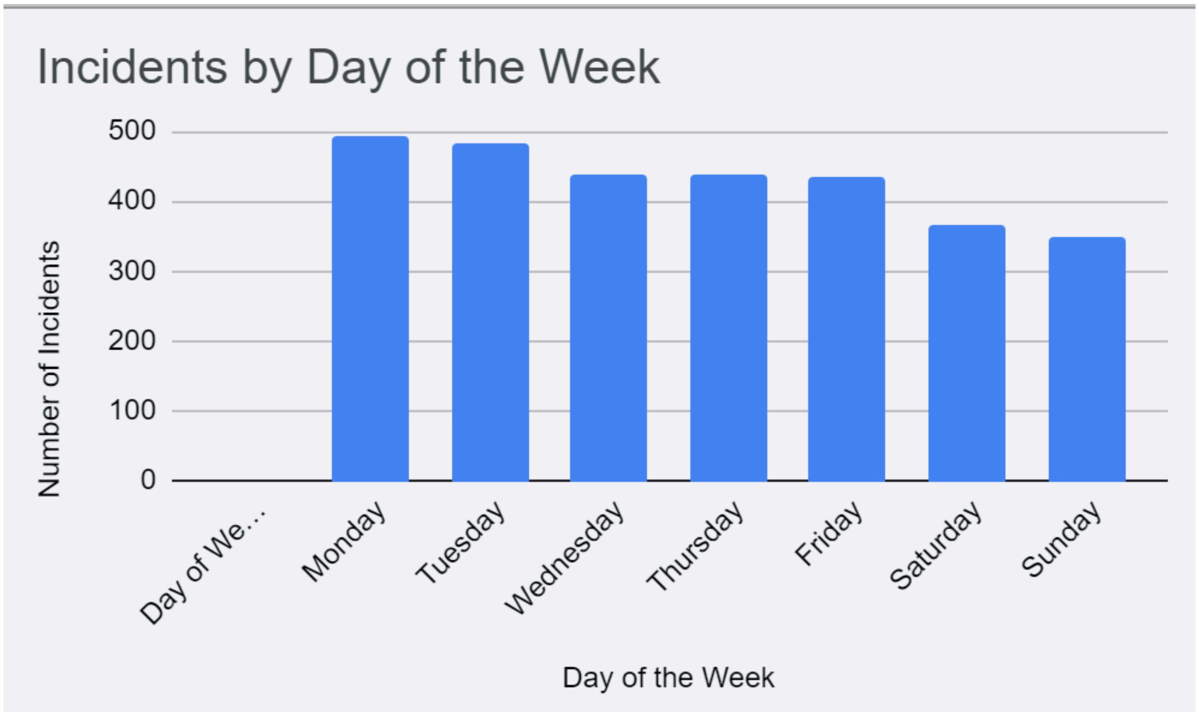


Chart 3 – Incidents by Hour of Day (Bar Chart): Shows the distribution of incidents across the 24 hours of the day, helping to identify peak times.
Hypothesis: Most incidents will occur in the evening and late night hours rather than in the early morning, when fewer people are outside.
The result **does not** validate the hypothesis.
Conclusion: For better recommendation, we can factor time of day into route risk scoring—penalizing segments more heavily at high-incident hours.

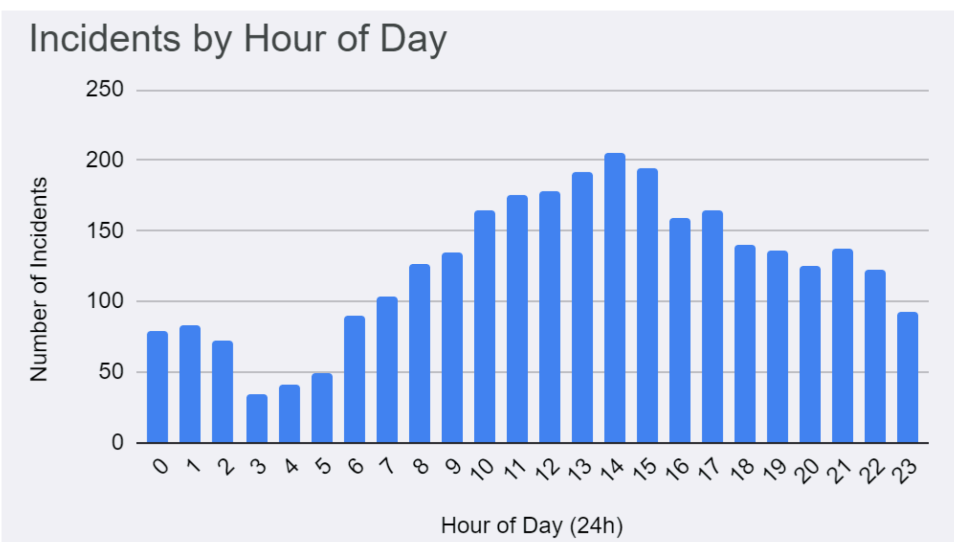


Chart 4 – Total Incident Counts by Type (Bar Chart): Shows the frequency of every unique incident type in the dataset.
Hypothesis: After combining both data sources, a small number of incident types will account for the majority of records, forming a clear "top risk categories" list.

Conclusion: The bar chart of incident counts by type shows that a few categories dominate the dataset, while many others appear much less frequently.
We can highlight these key categories in filters and legends, provide tailored explanations for them, and focus route-risk modeling and educational content on the types that actually drive most of the observed risk.

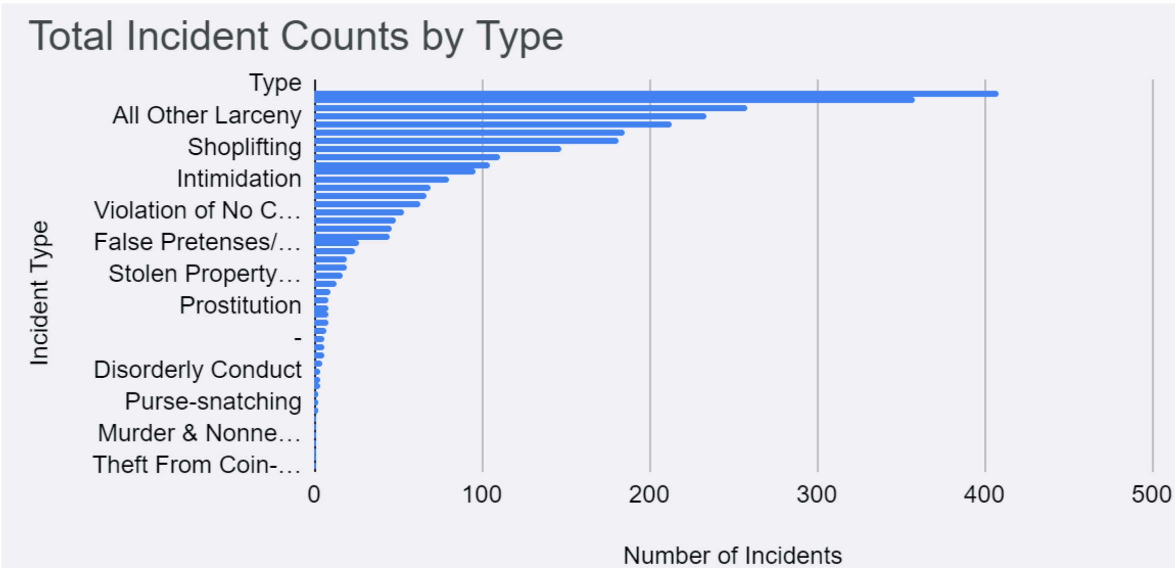


Chart 5 – Density Heatmap
Hypothesis: After combining both data sources, the graph will show a deeper color in the center/downtown areas.
The **result** conforms with the hypothesis.
Conclusion: The heatmap has more density in the center of the map. This is useful reference for us to avoid high risk areas when planning the route based on geological locations.