Data Task

As a **Data Engineer** at **Stacknexus**, you will be actively involved in integrating data from various sources (Kafka, Redshift, SQL, excel/csv files) to our datalake through different tools.

Here is one sample task that we will be working on at the data practices team. We would need you to work on the task mentioned below for the next 7 days (Starting from the date of receipt of this email) and present the results to us during your final round of interview.

Acknowledgement

The data was originally published by the City and County of San Francisco and made publicly available via San Fransisco Open Data (https://data.sfgov.org/).

Data Description

The data contains all fire service reports, fire incidents, fire safety checks and fire violations at the city and county of San Fransisco. You have 4 CSV files,

- Calls_for_Service_1.csv & Calls_for_Service_2.csv-> Contains all the fire department calls for service for the years 2021 & 2022. The entire table is split into two files as the data is large.
- Incidents.csv -> Contains all fire department related incidents during the years 2021 & 2022.
- Fire_Inspections.csv -> Contains information about fire safety inspections since the year 2004.
- Fire_Violations.csv -> Contains information about fire safety violations at various buildings across San Fransisco.

Task

Using the provided data, you must import the csv tables to any SQL server of your choice (MySQL, PostgreSQL, etc.) (added points if you can do the job on spark (pyspark, sparklyr, or basic spark-submit, etc.)), clean up the data and look for any anomalies and compute the following metrics as table views (more points for additional metrics) and visualise them for each metric.

The data is till March, 2022. You can also extract current data from the fire department opendata api (https://data.sfgov.org/Public-Safety/Fire-Incidents/wr8u-xric (https://data.sfgov.org/Public-Safety/Fire-Incidents/wr8u-xric)). Try to see if you can connect it with the live API.

Metrics

- Average number of incidents per day of the week
- Average number of incidents with the hour of the day
- Average time for despatch (Dispatch DtTm - Received DtTm) per neighbourhood.
- Average number of injuries and fatalities per neighbourhood for an incident

- Most common violation for each neighbourhood

Bonus

- On top of the metrics mentioned above, please try to come up with some more metrics which you feel would be useful to show a typical stakeholder from the San Fransisco Fire Department.
- Try to see if there are any correlations between violations and the incidents

Visualisations (Bonus, do it only if possible)

With the metrics that you have computed, please build visualisations of the above metrics using the tool of your choice (Tableu, PowerBI, plotly, etc.).

Bonus

See if you can use the GeoPoint columns to produce heat maps on top of regular San Fransisco map showcasing some of the above metrics on it.

Deliverables

After you have completed the task you can email us,

- A Microsoft word file, containing the SQL query you have used and the resulting table view that you got.
- A presentation containing the Visualisations (if you have built visualizations)