CIS 9440 - Data Warehousing for Analytics
Final Project Milestone 1
Group Number: 5
Student(s): Gabriel Fernandez, Jason Jiang

This Proposal is the beginning of your semester-long Final Project. The goal of the project is to develop a working Data Warehouse using a commercial database management system. Your project will use data from a public source(s), transform the data into a dimensional model inside your Data Warehouse, and connect to a Business Intelligence application to produce valuable, actionable insights.

For motivation on project ideas, **think about interesting problems, opportunities, or insights that could be shown, solved, or highlighted with data about New York City**. Search for datasets on NYC Open Data (https://opendata.cityofnewyork.us/) that interest you and your group.

Data Warehouse Project Title:
NYC Motor Vehicle Collision Transparency Data Warehouse Project

Motivation for Project idea:
A recent congestion pricing plan will charge drivers $23 to enter Manhattan if implemented. It was part of NYC's strategy to tackle congestion and other problems relating to the city's health (source). This sparked our group's interest in taking a look into the traffic congestion statistics in NYC - specifically relating to traffic accidents. This is especially relevant now that the city is implementing more bike lanes and reducing carbon emissions. For example, in 2013, Citi Bike was launched as an alternative transportation method to help the environment and keep the city's citizens healthy.

We want to see how traffic collisions in NYC changed over the years and if additions similar to Citi Bike made traffic conditions better or worse for the city. Our group hopes to bring more transparency to traffic collisions in NYC.

Description of the issues or opportunities the project will address:

- What are collision statistics broken down by vehicle type? (Cars, bikes, motorcycles, buses, etc.)
    - Is there seasonality to this?
    - Driver's license status
- What are the areas in the city most prone to collisions? (Boroughs, Zip Codes, etc.)
- What time are collisions more likely to occur? (Rush hour, mornings, etc.)
- Which are contributing factors of collisions in NYC?
    - Person emotional status
    - Pre-crash action (right turns, merges, etc.)
- What are the demographics of those involved in collisions?
    - Gender
    - Age
    - Borough

Business Justification:

High-level Business Initiative:

- We want to discover where and when most motor vehicle accidents happen in NYC and which demographics are more affected. Our group ultimately wants to bring transparency to the traffic collisions in NYC.

BI Sponsors and Stakeholders (who will own this project?)

- This project is by Baruch students working for the NYC Department of Transportation (NYC DOT).

What's the Business Value?

- By identifying the context of motor vehicle accidents, we can come up with suggestions to guide city leaders in preventing these types of accidents.

How long will this take? How much will this cost?

- We used the help of this article from Cooladata.com (written in 2017, so we **adjusted prices for 2022** using a CPI Inflation Calculator) to help determine the average cost this project can be expected to run.

- Ideally, we will need a Data Analyst, a Database Architect, a Backend Developer, and an I.S. Project Manager. The total headcount is four, and the cost breakdown can be summarized in the table below given that we have ~2.5 months for this project to be finished:

Position	Approximate Monthly Cost
1x Data Analyst	$7,000.00
1x Database Architect	$14,000.00
1x Backend Developer	$11,000.00
1x I.S. Project Manager	$11,000.00
Total Monthly Cost:	$43,000.00

- $43,000.00 * (2.5) = $107,500.00 for the entire duration of this project.

Technical Justification:

Which data sources do we already have for this project?

Dataset 1: Motor Vehicle Collisions - Crashes

Dataset 2: Motor Vehicle Collisions - Person

What new data sources do we need (if any)?

- Dataset 3: Motor Vehicle Collisions - Vehicles

Is the data we have conformed, consistent, and current? (data quality)

- Conformed
    - Yes, the datasets we have are conformed. The units within the datasets are the same and follow the same convention.
- Consistent
    - Yes, the datasets are consistent. They do share a primary key, but that would be used to obtain unique information from the other table.
- Current
    - No, two of our datasets are up-to-date, but the third is not:
        - Motor Vehicle Collisions - Crashes (last updated September 22, 2022)
        - Motor Vehicle Collisions - Person (last updated September 22, 2022)
        - Motor Vehicle Collisions - Vehicles (last updated December 8, 2021)

What technical skills will we need to complete this project?

- Requirements Gathering
- Data Modeling
- Dimensional Modeling
- ETL Creation
- BI Application Design and Implementation
- Data Warehouse Engineering
- Standardized Report Development

Will we need any new types of technologies?

- Lucidchart
    - For design, drawings, and models
- Google BigQuery
    - Data Warehousing
- Tableau
    - Business Intelligence and Dashboarding

Key Performance Indicators (KPI's) your Data Warehouse will display:

1. Number and average collisions per Borough
2. Number and average collisions per hour
3. Number and average collisions per Zip Code
4. Number and average collisions per Vehicle
5. Number and average collisions per street
6. Injury per safety equipment (from the person dataset)
7. Number of collisions per sex (from Vehicles)
8. Number of persons killed by collisions
9. Number of cyclists killed by collision
10. Number of cyclists injured by collisions
11. Average number of deaths
12. Percentage of collisions that resulted in deaths