Kimball Lifecycle Proposal Template

CIS 9440 - Data Warehousing for Analytics
Final Project Milestone 1
Group Number - 5
Student(s) - Hojin Lee, Mehdi Lahlou Charki, Ahmmed Hossain

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 at least 2 sources, dimensionally model the data 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**. Search for public datasets that interest you and your group. You may need to combine datasets to address your desired problem or opportunity. Below are just a few (of many) public data sources:

1. Kaggle
2. NYC Open Data
3. Opendata.gov
4. Gapminder
5. Zillow
6. NOAA Climate Data
7. Google's Public Datasets

To complete this Milestone, please fill in all bolded sections below:

Data Warehouse Project Title:
Overcoming Airbnb post-pandemic financial stress, by analyzing transit data.

Motivation for Project idea:
We are looking for Airbnb revenue factors generated by both CitiBike and taxi frequentation. Does a neighborhood containing one or more CitiBike stations make a rental more attractive and expensive? Where are the most served taxi locations, and do they contribute to higher attractiveness for airbnb rentals in that area? Analyzing these questions will help Airbnb promote certain locations more efficiently to generate more income.

Description of the issues or opportunities the project will address:
We are trying to understand the correlation between revenue of Airbnb listings and the transit data in regards to CitiBike and Yellow Taxi pick up and drop offs. The opportunity that may arise based on the findings of this project will allow Airbnb to understand where potentially lucrative short term rentals are. In turn, Airbnb can promote or incentivize new and/or existing listings in those respective hot zones.

Business Justification:

High-level Business Initiative:
Generate more revenue post-pandemic by analyzing current transit data from CitiBike and taxis.

BI Sponsors and Stakeholders (who will own this project?)
BI Sponsors: Project managers and Airbnb executives
Stakeholders: Guests, Hosts, communities, data architect, data managers, market researchers, data scientists

What's the Business Value?
Recover revenue lost from the global pandemic by improving customer satisfaction and promoting listings in regions highly served, and where revenue potential is heightened.

How long will this take? How much will this cost?
We are planning to complete the project within a defined scope and time period for 2 to 3 months. The project cost is based on datasets, labor, research, and technologies that we are using

Technical Justification:

Which data sources do we already have for this project?

Dataset 1:
http://insideairbnb.com/get-the-data.html
Data Complied 09/01/2021 New York City

Dataset 2:
https://www.kaggle.com/vineethakkinapalli/citibike-bike-sharingnewyork-cityjan-to-apr-2021

Dataset 3:
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
We will use historical data from 2019 as well as the 2021 for-hire vehicle trip records (the 2021 yellow taxi trip records haven't been provided yet) .

Dataset 4:
https://data.world/nyc-taxi-limo/taxi-zone-lookup/workspace/file?filename=taxi%2B_zone_lookup.csv

https://ride.citibikenyc.com/system-data

Is the data we have conformed, consistent, and current? (data quality)
The for-hire vehicle trip records for 2021 contains many missing values for the pickup locations but the other 3 are conformed to what we need for the analysis.

What technical skills will we need to complete this project?
SQL, Python and Tableau are the softwares we will use throughout the analysis.

Will we need any new types of technologies?
We will not need new types of technologies as our analysis will be based on the software cited above.

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

1. Average rental ratings/location
2. Number of CitiBike stations
3. Taxis available per rental
4. Bikes available per rental
5. Average price per rental in the area
6. Yearly revenue/dropoff [taxi]
7. Yearly revenue/pickup [taxi]
8. Yearly revenue/pickup [citibike]
9. Yearly revenue/dropoff [citibike]

10. Revenue per customer
11. Average customer review