DATA DISCOVERY REPORT

This is an analysis of datasets recognized as the potential datasets useful in the field of data science for understanding various things about Chicago. We have identified attributes in datasets which can be integrated with another to extract information about Chicago. Each dataset individually can be used to analyze details of Chicago. In our search for interesting datasets, we have chosen the following as they could be the optimal datasets for better results.

Data Discovery Results:

Dataset-1: Chicago homicide data since 1957
Source: Chicago Tribune

Tabular representation of homicides in the city of Chicago, from 1957 to 2015 on monthly basis. It gives the average, highest, lowest recorded homicides in that year. This can be helpful in analyzing the crime trend and know the rate of murders occurring.

Integration:

<table>
<tr><th>Similar Attributes</th><th>Key Fields</th><th>Description</th></tr>
<tr>
<td>
<table>
<tr><th>Found-Data</th><th>Given-Data</th></tr>
<tr><td>year</td><td>Year</td></tr>
</table>
</td>
<td>Year, Month, Total, Population</td>
<td>We can draw patterns by evaluating the data from the consecutive years. For instance, If the number of homicides increases in that year, how it affects the rate of population growth in subsequent years.</td>
</tr>
</table>

Dataset-2: Apartment Listings, Rental Pricing
Source: Inside Airbnb, Zillow
This is a collection of current apartment listings in the Chicago. It is extracted from the public data repository of Airbnb. It can be analyzed to find the listings in a zip code, the number of apartments frequently being rented, income of the landlords, and to find leading property agencies in Chicago.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Zip Code / Zip Code	Zip code, Year	This data integrated with the business dataset to predict the success rate of the business in a neighbourhood based on the people. For instance, If migration rate to that location is large, the business can run with profits.

Dataset-3: Past Weather Data
Source: Wunderground

This data represents weather records containing weather attributes like temperature, precipitation, humidity of last 50 years. This can be used to predict the weather forecast in the city of Chicago. It can also be used to learn the trends about the changes in the atmospheric conditions due to various activities, impacts of crime in different seasons.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Date / Date, Latitude, Longitude	Date, Temperature	We can fabricate statistics by correlating the weather and the type of crime that has occured and determine the crime frequency with respect to the weather condition.

Dataset-4: Median Salaries of various jobs
Source: Glassdoor

A growing database of millions of company details around the globe. It provides free access to the data in major cities including Chicago. In data science, this can be cleaned and used for the analysis of living standards against their occupation.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Month / Year	Month, Median Salary, Population, Citizenship	Integrating with the census dataset can compute relative growth between the median salary of each profession with respect to the population growth.

Dataset-5: Clinical Data
Source: Chicago Health Atlas

The following data contains information about clinical care, mortality, etc classified based on attributes such as race, age, gender, etc. Information about the variations of disease effects can be studied to know the groups of people are affected by diseases. These trends can be understood so as to take precautions.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Zip code / Community area name	Percent aged 16+ unemployed, deaths of youth depression	The fact that does unemployment has any relationship with depression can be studied.

Dataset-6: United States environmental protection agency
Source: EPA

This is a data repository which contains everyday environmental data for a specific location and time period. It contains ambient air pollution data collected by EPA, state, local, and tribal air pollution control agencies. This data can be combined with weather data so as to get effects of pollution on the weather.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Longitude, Latitude / Zip code	Pollutant, License description.	By integrating this to Business Licenses, we can understand pollutants emitted in various industries, we can check how safe those business operations are. Licenses can also be verified based on this.

Dataset-7: School Ranking in Chicago
Source: CPS

The Chicago public school comprises of data related to performance of elementary and high schools in Chicago. It rates each school with School Quality Rating Policy(SQRP) index and also mentions the growth rate of each school by type of student(eg. Hispanic, African-American). It can rank schools and make recommendations to the user based on their preference .

Integration:

Similar Attributes	Key Fields	Description
Found-data / **Given-Data**: School Name / Name of School	SQRP Rating, School Growth Percentile Score	Using school name as common column, they can be combined to display schools rank-wise for any given zip code.

Dataset-8: Demographics of Chicago
Source: City-Data

It comprises a comprehensive data for each zipcode. It include population by race, median salary, media age, time to commute for work, business listings and many more interesting insights.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Zip code / Zip code	Business Activity	Based on business activity, a neighbourhood can be categorized as place popular for specific business.

Dataset-9: Project of Regional Transit System
Source: Regional Transportation Authority Mapping & Statistics

This lists the past, current and future projects for improving chicago transit system. It specifies each project's funding, status and year. Based on all the previous Transit improvement plan, data science can used to suggest improvement plans for various places in chicago.

Integration:

Similar Attributes	Key Fields	Description
Found-Data / **Given-Data**: Project Title / Region_id	Speed	Based on historical traffic data of the regions, we can predict when to start a project.

Dataset-10: Restaurant Data of Chicago
Source(s):
Foursquare
Google Places Reviews API
Yelp

The Data from these web sources is obtained in the form of HTML/JSON content. We can clean these to convert it into a tabular format by integrating the content from these websites. We can get the information about the ratings, reviews, address of restaurant so that we can produce interesting results, patterns based on the feedback.

Integration:

Similar Attributes	Key Fields	Description
<table><tr><th>Found-Data</th><th>Given-Data</th></tr><tr><td>ZipCode, Restaurant Name</td><td>Zip, DBA Name, AKA Name</td></tr></table>	Zip, Name, Ratings, Inspection Result, Inspection Name, Number of Inspections(Calculated using a different attribute “Name”)	Integrating these datasets with the food inspection dataset to study the impacts of reviews based on the food inspection results.