Extraction, Transformation, and Load Technical Report

Relationship Between Crime and Marijuana in Denver

**TABLE OF CONTENTS**

1. Introduction 3
   1. Summary 3
   2. Scope 3
   3. Technologies and resource contributions 3
   4. Definitions, Acronyms and Abbreviations 3
2. ETL Details 4
   1. Data Import/Extract Sources and Method 4
   2. Data Acquisition 4
   3. Data Transform 4
   4. Data Integrity 4
   5. Data Refresh Frequency 4
   6. Data Security 4
   7. Data Loading and Availability 5
3. Data Quality 6

|  |  |
| --- | --- |
| **1.** | **INTRODUCTION** |

*The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.*

# 1.1 Summary

The objective of the project is to compare marijuana related crimes to marijuana business sales in Denver, Colorado. The problem is our client, Snoop Dogg, does  not have a database to compare marijuana related crimes to business sales in Denver. The expected outcome of the ETL is to create a SQL database that will cohesively compare two data sets based on Denver crimes and its effect on business sales. By doing this, Snoop Dogg’s company can develop a marketing strategy to expand their business to other states, and work on changing legalizing marijuana all around the country.

Statement of Purpose: Our statement of purpose is to establish a relationship between marijuana business sales and crime rates in Denver. The datasets we are using is Marijuana\_gross\_sales.csv and crime\_marijuana.csv

# 1.2 Scope

This section explicitly outlines the disparate data sources that are to be integrated, which components of the overall data science project is in the scope for this initiative and also lists out the components of the data science project that are not in scope here.

* + Marijuana\_gross\_sales.csv and crime\_marijuana.csv are the data sets being used
  + We are joining the data by the date column for both data sets
  + Scope: analyzing the effect of marijuana on crimes in Denver and its relationship with business sales.
  + Components that are not in scope are the neighborhoods the crimes were reported in. We are only interested in the overall scope of marijuana and its effects on crime.

# 1.3 Technologies and resource contributions

This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.

* De-Anna Clarke did the extraction and part of transforming the data. She extracted the data from csv files and imported them into pandas and cleaned the data.
* Nicolas Colon was responsible for the transformation and helped with extraction. Nicolas changed the format of the dates listed in the tables to make the join.
* Chantay helped with transformation and loading the data into a SQL database. Chantay loaded the data from python into SQL and made a query view for Snook Dogg to see the joined table.

# 1.4 Definitions, Acronyms and Abbreviations

List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and

Load

|  |  |
| --- | --- |
| **2.** | **ETL DETAILS** |

*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

# 2.1 Data Import/Extract Sources and Method

This section provides information about the data and its source. For example, API names and URLs, key parameters available and its subset which will be preserved (loaded). Data extraction protocols (API, FTP, Web scraping etc.), any permissions required to access the said extraction dataset and any restriction placed on the usage and distribution of the acquired dataset.

Both datasets were imported from [www.data.world](http://www.data.world).

# 2.2 Data Acquisition

How to decide on the selection of data while re-obtaining or updating. Discuss, here the dimension of the obtained dataset and if updated what is the project growth rate of the data. Lastly, address any issues or pre-requisites that needs to be cleared prior to getting the data? (DE ANSWER THESE QUESTIONS)

* Data is dynamic and needs to be updated at the end of every year to be accurate, and expand the data
* The company will have to reach out to our team and pay for services to update and obtain the data again. We will then, find data reports of Denver crime in relation to marijuana over the past year and marijuana gross sales for that year, drop unnecessary rows of data, and load it to the ETL database. Once the load is complete, we will add a joint view of the data for the company to have for the next year.
* If the company attempts this process on their own, they have to transform:
  + REPORTDATE column in crime\_marijuana.csv to format (‘MONTH’- ‘YY’)
    - In order to make a new column called DATE
  + MONTH and YEAR columns have to be combined in marijuana\_gross\_sales.csv

# 2.3 Data Transform

In this section address any data transformation that needs to be performed to modify, clean, filter or create existing and new parameters. Address any technical analysis performed, include design specification or data models used (example linear interpolation etc.), and any calculations performed for any newly derived fields.

# 2.4 Data Integrity

In this section discuss the reliability of the extraction source data (e.g., missing data, dates stored as text, invalid code values, text fields with odd characters, etc.). Address the frequency with which the data sources are updated and if it is necessary to update the local data at the same frequency. Lastly, how if any notification can be received when the source data is updated; and what if any notification will be sent to the internal team when the local dataset is updated.

# 2.5 Data Refresh Frequency

This section explicitly lists the frequency with which this ETL process will refresh the local dataset (Daily, Weekly, Monthly, Quarterly, Semi-Annually, etc.).

# 2.6 Data Security

This section discusses any data anonymity and security requirements need to be satisfied. Address any federally mandated HIPAA considerations, any need to build in additional privacy, Encryption, Data masking, Auditing, Backups etc.

# 2.7 Data Loading and Availability

This section addresses the data schema and during of data retention. Discuss the interface that will allow your Client/Users to access the data.

|  |  |
| --- | --- |
| **3.** | **DATA QUALITY** |

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?