Extraction, Transformation, and Load Technical Report

Relationship Between Crime rates and Marijuana sales in Denver, CO

**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 this project is to provide a dataset for our client to conduct analysis on the relationship between the legal sale of marijuana for medical and/or recreational purposes and rates of marijuana related crimes. Our client, Greenway Org, is a non-profit organization created by Snoop Dogg and other celebrity donors to advocate to the expansion of legal marijuana sales in the United States. Denver, Colorado has become a well-known for its marijuana friendly culture and economy and our client requested we provide as much historical data available on marijuana related crimes and marijuana related business sales for this city. 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 relationship to marijuana business sales. Greenway intend to use the analysis from our dataset to develop a marketing campaign for its social media sites as well as a proposal for its volunteers who meet with city council members to advocate for the legalization of marijuana all around the country.

Statement of Purpose: Our statement of purpose is to present our client with a database with which to explore the relationship between marijuana business sales and crime rates in Denver. The datasets we are using are 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
  + Both datasets include a date column with identical formatting for so an aggregate view of the data can be formed easily.
  + Scope: the rate of legal marijuana sales with crime rates in Denver.
  + 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

EasyData’s dataset development team consisted of Chantay Drake, De-Anna Clarke and Nicolas Colón

* De-Anna Clarke: extracted data from csv files into Jupyter lab Notebook then, using Python, began data cleaning by dropping any records that did not have gross sales numbers available.
* Nicolas Colón: created a column in both datasets to hold a new format of the dates listed in both tables so they can be joined for an aggregate view in SQL.
* Chantay Drake: assisted with data transformation and loaded data into SQL from Python and made a query for Greenway’s analysts to view the joined tables.

|  |  |
| --- | --- |
| **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

Both datasets were downloaded as csv files from [www.data.world](http://www.data.world). Marijuana Gross Sales and Marijuana Related Crime were both provided by the City of Denver <https://www.denvergov.org/> as the source from Data World. Since we had access to download the csv files from the Data World site, there was no need to perform further extraction protocols. No permissions were required to access the said extraction dataset nor any restrictions placed on the usage and distribution of the acquired datasets.

# 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?

* 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.

Obtaining the information directly from the City of Denver provides the most reliability for information sourced that can be expected. The only missing data was in the sales dataset in relation to dates prior to the when the sales of marijuana was deemed legal in the state of Colorado. We did not find any poor formatted data, invalid codes nor odd characters.

# 2.5 Data Refresh Frequency

The City of Denver updates both csv files provide to data world every 4 years. As outlined in the service agreement signed between EasyData Inc. and Greenway, our team will provide ETL services for the next update of datasets and upload our cleaned version of datasets to the GitHub provided as the final deliverable.

# 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?