Extraction, Transformation, and Load Technical Report

Relationship Between Crime rates and Marijuana sales in Denver, CO

Brought to you by:

**EasyData!**

Authors:

De-Anna Clarke

Chantay Drake

Nicolas Colón

**TABLE OF CONTENTS**

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

|  |  |
| --- | --- |
| **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.

* + 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

EasyData was able to come to confirm with the Data World and the City of Denver, that will not require any pre-requisites that will need to be cleared prior to getting the currently available data sets and any updates in the future. The sales data is being tracked monthly so we expect a constant rate of growth as a that data is updated. The legal sale of marijuana is a fairly new phenomenon so the dataset development team was not able to find previously established data on the rate at which crime is affected by the legalization of marijuana so we cannot present an accurate prediction on the rate of growth for Marijuana Related Crime data.

* Data is dynamic and needs to be updated at the end of every 4 years to be accurate, and build trends by establishing historical 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

The data any data transformation that needs to be performed to modify, clean, filter or create existing and new parameters. There was no technical analysis performed, including design specification nor data models used (example linear interpolation etc.), nor any calculations performed for any newly derived fields.

# 2.4 Data Integrity

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

While are information gathered within there datasets did not fall under any federally mandated HIPAA considerations, we did want to ensure that no individuals linked to this dataset would feel as though their activities were being unnecessarily tracked. The Marijuana Related Crime dataset, included addresses for where each incident occurred as well as dates of the first and last time an incident was reported at that address. EasyData chose to remove these columns in an effort to build in additional privacy for the human beings behind these incidents. As the analysis will be conducted on a city-wide level, our team does not believe the efforts of Greenway’s research team will be negatively impacted by the removal.

# 2.7 Data Loading and Availability

Greenway’s research group will have access to the data and any updates provided via the EasyData provided URL. The URL leads to a client named GitHub where files will be available 24/7 for the duration of the service agreement established between EasyData Inc. and Greenway Org.

|  |  |
| --- | --- |
| **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?