Data Analytics Final: **Data-Driven Decision Making**

For your final, you will utilize data analytics to address a question or assist in making an informed decision. Questions you can answer can range from, “are there any anomalies in this cost data?” or “I was dealt a 20 in blackjack and the dealer shows a 9, can I predict the odds of winning this hand?” or “My boss asked me to identify the months a person may be more predisposed to buying x,y,z”.

The project involves several steps, but starts by asking a question that can be answered by data. Once you know what you are asking/trying to determine you will need to acquire the data. Data acquisition will aid in answering the question you pose. You will then clean the data if applicable. Once the data is ready, do some descriptive and predictive analysis, then visualize the data. Finally, you will creare a concise summary of the findings and their implications, then present to the class (5 min).

Introduction (10%)

* Define a clear question or decision-making scenario that can be addressed through data analytics.

How can we (L.A.P.D.) lower the crime rate within our city?

* Justify the significance of the chosen question or decision-making scenario in a real-world context.

This question is probably asked all the time in multiple cities. It's a question that can always be answered and improved. But, with data analytics, it can be answered even better and easier.

Data Acquisition (15%)

* Identify and acquire a relevant dataset from an online source.

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

* Import the dataset into a Snowflake database, detailing the steps taken.

I created a new schema called LACRIMEDATA under my database. I then created a table named CRIMEDATA and imported the csv file into that table, fixing the column names (replaced the spaces with "\_").

# Data Cleaning and Preparation (10%)

* Assess the dataset for quality and cleanliness, does any cleaning need to be done? Any nulls? Outliers?

First I checked for any NULLS to make sure the rest of the preparation and querying went smoothly. I found that there were no NULLS within the data. I then checked for any outliers within the Vict\_Age, Lon, and Lat columns. I proceeded to delete those rows to clean up the data.

* Perform necessary data cleaning and preprocessing steps, documenting the rationale and methods used.

I proceeded to delete the rows that came back for the outliers to clean up the data.

# Descriptive Analysis (15%)

* Conduct a descriptive analysis to summarize the main characteristics of the data.

First I did a descriptive analysis of the numerical columns to get the min, max, avg, and stddev of each one. I then did a descriptive analysis for the categorical columns to get the total counts of each of the data points they could be.

* Identify patterns, anomalies, or insights that could inform the subsequent predictive analysis.

I looked for any patterns in the date columns by getting their total counts. And I also looked at the location distribution of the lon and lat columns.

# Predictive Analysis using Snowflake Cortex ML Functions (20%)

* Select and justify a machine learning approach (Anomaly Detection, Prediction, or Classification) relevant to the project objective.

I chose the classification approach. These models can be used to predict crime locations, types, and victims. This information can be valuable for law enforcement to allocate resources effectively. These models can also be used to learn patterns which can help find the underlying causes. And overall these models can help identify high-risk areas and develop ways to reduce the crime rates.

* Train a model using Snowflake Cortex ML Functions, explaining the process and choice of parameters.
  + See <https://docs.snowflake.com/en/guides-overview-ml-functions>

First you need to give a role the CREATE SNOWFLAKE.ML.CLASSIFICATION privilege on the schema where you want the model to be created. You then use CREATE SNOWFLAKE.ML.CLASSSIFICATION to create and train a model. I created a model that used the Lon, Lat, and Vict\_Sex columns to get predictions of future crime locations. I then created a view containing binary data for training. Then used that to create and train a binary classification model. Then I used that to make a prediction on the data.

# Data Visualization with Power BI (10%)

* Connect your dataset(s) to Power BI to and visualize your findings.

Opened PowerBI selected “Get Data” then “Snowflake”. Inserted the corresponding information and imported my CRIMEDATA table.

* Ensure that the visualizations effectively communicate the insights and support the data-driven decision-making process. Below, paste a screenshot of your visualization and the question it answers/decision it drives.

A map of a country with different colored dots

Description automatically generated

Where are the crimes being reported? And what crimes are they?

A graph showing a blue line

Description automatically generated

What are the total crimes reported over the last few years? Are there any patterns or outliers?

A screenshot of a computer

Description automatically generated

What are the specifics behind the type of crime and victims?

A screenshot of a screenshot of a computer

Description automatically generated

Filters for the first report page.

A graph of blue bars

Description automatically generated

What are the total crimes against each sex? Who is targeted the most?

A graph of crime

Description automatically generated

What are the total crimes against each age? Who is targeted the most?

A screenshot of a computer

Description automatically generated

What is the most committed crime?

A screenshot of a computer

Description automatically generated

Filters for the second report page.

# Conclusion and Presentation (20%)

* Write a 250-word overview summarizing the key findings, their implications for the question or decision at hand, and potential future work. Include the conclusion below and in your powerbi report.

Here are some of the key findings I found when using the dashboard and other resources. Once I finished creating the dashboard I went and answered a few questions. The first question was which area has the most crimes reported. I found out that it was the north western side of L.A.. The next question was, how has crime changed over the years. It seems to be mostly steady with some large spikes between the years 2022 and 2023. There could be many reasons for this, one of those being inflation. The third question was, who gets targeted more? Here I looked at the bar graphs for the victim’s sex and age. It shows that males get targeted more, but not by much. We also must consider the X category. These are people who didn’t give their sex in the report or whoever filled out the report could have missed filling in that data. But since it could be either sex, we can assume it would stay about the same. Then when I was looking at the victim’s age graph, I saw that the targets were mostly between the ages of 20 and 40.

After the whole analysis I saw the data and trends that I was looking for. I did look at more specific data using the dashboard I created, and we could dig a lot deeper into this for many more questions we want answered like, what are the specifics for the X category or what are the most committed crimes in each area. But for the sake of keeping this conclusion a reasonable size I will stop here and can say I got what I was looking for.

* You will present your PowerBI report and findings to the class. Ensure you can navigate the report and explain the why and how behind it.

## Deliverables

* A completed final report
  + Submit to Canvas
* A link to your repo in GitHub. Use it to store your code, powerbi report, and findings
  + Submit to Canvas
* A Power BI dashboard showcasing the data visualization(s) and findings/conclusions. Upload the file to your github repo
  + Submit to Canvas
* A brief in-class presentation summarizing the project findings and implications
  + In class 4/25

## Assessment Criteria

* Clarity and relevance of the question or decision-making scenario.
* Technical proficiency in data handling, analysis, and visualization.
* Logical flow and coherence of the data analysis process.
* Depth of insight and analysis in the predictive modeling.
* Quality and effectiveness of the data visualizations.
* Conciseness and clarity of the final overview.