Observations:

<div class="section-header text-left">

<p>Research Question</p>

<h2>

Can we predict the type of crime that will take place based

on a variety of spatial, temporal and categorical factors?

<p>

Independent Variables:

<div class="service-text">

<h3>Visualizing Spatial Trends</h3>

<p> Larger population sizes do not always yield more reported arrests. </p>

<br>

<p>The proportion of arrests by crime type (can we mention what crime type is Category 1, 2 and 3?), vary slightly based on location;

however, Crime Against Category 3 appears to represent the largest proportion

for most counties.</p>

<br>

<p>The NIBRS database does not contain data for several Texas counties;

the lack of coverage is currently unknown.</p>

</div>

<div class="service-text">

<h3>Visualizing Seasonal and Temporal Trends </h3>

<p>There appears to be little variability in seasonal trends, likely due to mild seasonal

changes in the weather throughout most of Texas.</p>

<br>

<p>Temporally, the Crime Against categories of person and society follow similar trends, peaking in the late night/early morning hours. The property category varies in that it peaks in the late afternoon, evening hours.

</p>

</div>

<div class="service-text">

<h3>Probable Multicollinearity Between Some Independent Variables </h3>

<p>The p value for the suburban area flag 1 and 2 (Can we mention what do we mean we area flag 1 and 2?) is significantly low (<0.0001)</p>

<br>

<p>The r squared value for the suburban area flag 1 and 2 is within .02 of each other;

.99 and .97 respectively.</p>

<br>

<p>Given the statistically significant correlation, we can remove two of the three independent

variables for the purposes of our multinomial logistic regression—using the binary

variable for the suburban area flag will make our machine learning more efficient because

it will not need to evaluate many discrete, integer values (population and total employees).

</p>

</div>

<div class="service-text">

<h3>Measuring Correlation Between Variables</h3>

<p>

Offense Type has a strong, negative correlation with Crime Against Type;

however, this variable must be removed because it has a hierarchical relationship with

our dependent variable.</p>

<br>

<p>The high correlation between Population, Total Employees (police department) and Suburban Area Flag is visible in our heat map.

</p>

</div>

<div class="section-header text-center">

<h2>AWS's relational database (RDS) service and Storage S3

</h2>

</div>

<div class="container banner-text">

<p>Our team hosted the raw and processed data in Amazon RDS. This allows to encrypt the database using keys that are manage through AWS Key Management Service (KMS).

The data connects through a cloud database that runs in our local computers.

Also, the team also created a S3 "datacrime-bucket" that stores the unstructured and reference content. This allows scalability, data availability, security, and performance.

The selected data was applied for initial data analysis and machine learning modeling. The data sets were published to a

postgres database and queried using SQLAlchemy to run our models. The SQL

queries and joins were conducted to building relationships within the database between

multiple tables.

</p>

</div>

<div class="blog-text">

<h2>SQL Schema</h2>

<div class="blog-meta">

</div>

<p>

This is the logic structure of the database where it is grouped the objects of the tables used in the exploratory analysis and the Supervised ML model.

</p>

</div>

</div>

</div>

<div class="col-lg-6">

<div class="blog-item wow fadeInUp" data-wow-delay="0.3s">

<div class="blog-img">

<img src="img/S3.JPG" alt="Blog" />

</div>

<div class="blog-text">

<h2>PySpark AWS S3 Read CSV files</h2>

<div class="blog-meta">

</div>

<p>

The Colab script performs reading operations on AWS S3 using Apache Spark Python API PySpark that allows access to the bucket and read the data into DataFrames. Then, the data is pushed to the postgres environment.

</p>

</div>

<div class="blog-img">

<img src="join/sql.jpg" alt="Blog" />

</div>

<div class="blog-text">

<h2>Postgres database and SQL queries</h2>

<div class="blog-meta">

</div>

<p>

Using the database-crime with defined public access, the endpoint was aggregated as server in pgAdmin to execute SQL commands and review data outcomes.

</p>

</div>

</div>

</div>

<div class="col-lg-6">

<div class="blog-item wow fadeInUp" data-wow-delay="0.3s">

<div class="blog-img">

<img src="join/query.png" alt="Blog" />

</div>

<div class="blog-text">

<h2>Create SQL tables</h2>

<div class="blog-meta">

</div>

<p>

The first function of CRUD is creating data. The team created tables and inserted data into them. Inserting is the main part of creating data to be stored within the database.

</p>

</div>

<div class="img/aly.jpg">

<h2>Alejandra Rivera</h2>

<h4>Financial Analyst</h4>

<p> Interests include research in financial and green economics as well as the application of financial modeling techniques to improve efficiencies and optimize processes in dynamic environments.

</p>

<div class="team-social">

<a class="btn" href="https://twitter.com/cocolirio"><i class="fab fa-twitter"></i></a>

<a class="btn" href="https://www.linkedin.com/in/alejandra-rivera/"><i class="fab fa-linkedin-in"></i></a>

</div>

</div>

<div class="team-text">

<h2>Becky Jones</h2>

<h4>Geospatial Intelligence Analyst</h4>

<p> Interests include developing advanced proficiencies in spatiotemporal data analysis.

</p>

<div class="team-text">

<h2>Matt Lane</h2>

<h4>FP&A Financial Analyst</h4>

<p>Interests include creating efficiencies of repetitive deliverables.

</p>