Data Science Final Capstone Proposal

Ashley Steele

Project title:	*To Everything There is a Season*: Using Weather Data and Demographic Information in the Predictive Modeling of Crimes in Dallas, Texas
Project Summary:	In Texas we have a saying: “If you don’t like the weather, wait five minutes…. It will change.’ This is mainly based on the volatility of the weather in North Texas. As a resident of Dallas, Texas for over 20 years I have experienced first hand how the variable weather (including pollen production, sniff sniff) has impacted my mood and attitude. My goal for this project is to investigate the connection, if any, between Dallas’s demographic information, its crazy weather patterns, and incidence of crime (focusing on the type and severity of crime) to determine if a predictive model can be made to help improve public safety (and potentially police efficiency as well).
Hypothesis/Research Question:	What impact, if any, does the ever changing weather have on type and severity of reported crimes across different geographical areas of Dallas, Texas?
Project Problem:	With the assistance of both supervised and unsupervised learning methods I want to see if I can develop a predictive model that can foresee type and severity of future crimes based on weather and demographic information in Dallas, Texas.
Impact of Solution:	Currently Dallas is experiencing one of the highest cycles of violent crimes in the past decade. With the help of my predictive model I want to create an interactive, web-based dashboard that allows users to see projected high crime areas based on weather and demographic information in an attempt to help increase awareness, protect individuals from crimes, and, potentially, assist police departments in preemptive dispatching of officers.
Data Sources & Access:	● Dallas Police Incoming Calls from 2008- August 31st, 2019: The City of Dallas Open Data Website ● U.S. Census Demographic Information for Dallas, Texas 2008-present : U.S. Census Open Data Portal ● Historical Weather Information for Dallas, Texas : Various websites, accessed through web scraping
Techniques/Methods to be Utilized in Project:	● Time series analysis ● EDA

<table>
<tr><td></td><td>● Feature engineering
● Unsupervised and supervised learning
● Predictive modeling
● Data visualization
● <b>Extra</b>: Tableau Public interactive dashboards & GPS data tracking</td></tr>
<tr><td><b>Challenges/Potential Roadblocks:</b></td><td>● <b>Data cleaning and feature engineering</b>: the amount of data from the DPD is massive and it will be slightly difficult to clean up/manipulate
● <b>Time constraints</b>: this project has a very quick turn around (less than 3 weeks) and I want to make sure it is the best possible display of what I have learned throughout my course.</td></tr>
<tr><td><b>How is my Specialization Being Used in This Project:</b></td><td>Since my specialization is <b>time series analysis</b> I plan on utilizing the time-based nature of the crimes/reporting calls in the Dallas PD dataset. My base goal (analysis of crime related to date and the forecasting of future crimes using this, along with weather and demographic information) is at heart a completely time series based problem. Additionally, due to the seasonality and time-based nature of the weather, my supplemental data, historic weather for the areas in question, also fall into my specialization of <b>time series analysis</b>.</td></tr>
<tr><td><b>Goals/Milestones for This Project:</b></td><td>My project will be divided into several Jupyter notebooks, each with their own goals. They are as follows:
<table>
<tr><th>Notebook</th><th>Overarching Goal/Milestone</th></tr>
<tr><td>Data Extraction/Scraping</td><td>● Web scraping of weather information
● Collection and cleaning of the three main data sources
● Organization/merging of datasets in preparation for EDA</td></tr>
<tr><td>EDA and Basic Visualizations</td><td>● Analysis/breakdown of existing features
● Basic visualizations for distribution and content
● Feature engineering of relevant characteristics for modeling and analysis</td></tr>
<tr><td>Time Series Analysis</td><td>● Creation of sub-data frames for ARIMA modeling and trend observation
● Analysis of time related data and its</td></tr>
</table></td></tr>
</table>

	implications on predictive models ● Use of unsupervised learning models to analyze and cluster data to understand underlying patterns
Predicting Modeling	● Use of clean and organized time series data to forecast potential crime types, severities, and geographic locations based on EDA and past incidences ● Use of supervised learning, compared to crime calls after August 31st, 2019, to create and validate a predictive model of future criminal activity based on on-going weather reports
Reflection/Future Implications	● Formal write up of process, findings, and how this might impact future end-users
Interactive Dashboard/Visualization	● Tableau Public interactive dashboard published on the web for end-users to interact with past data and predictive models to help improve their safety relative to crime in Dallas, Texas

Variables to be Utilized in This Project:	**Since this project breaks down into several different aspects it is difficult to list all of the variables that will be used. Some of the current variables and features I plan on using, based on initial exploration of the data sets, are:** ● GPS coordinates (X and Y) of locations of criminal activity ● Crime occurrence dates ● Dummy variables of existing features for machine learning models (such as crime type, month, year, etc….) ● Sequential dates ● Temperature (in Fahrenheit) ● Pollen count for individual days ● Crime type/category ● Precipitation/humidity/other weather related information ● Geographical locations/city divisions ● There are many more but I will not know the full scope of what I will actually utilize until I am able to get into the data and work with it!