Part 1

The problem being investigated is racial/ ethnic disparities in the data collected by the San Diego Police Department during traffic stops. This data for this study consisted of "259,569 records generated by SDPD officers following traffic stops occurring between January 1, 2014, and December 31, 2015". However, the data analyzed only included stops that occurred because of the suspect description, code enforcement effort, etc. This was done to try and limit the analysis to situations where the reason for the discrepancy in the decisions of post- stop outcomes could be attributed to an officer's use of race/ ethnicity bias.

The data collection was done using "vehicle stop cards". These recorded basic driver demographic information: race, gender, age, San Diego residency, date, time, location of stop, and reason for stop. Post stop outcomes are also recorded" issuance of a citation, field interview, search conducted, seizure of property, discovery of contraband, or arrest. There is also an optional space for qualitative descriptions. This study also only used data for stops that were recorded as an equipment or moving violation.

It is appropriate to address the problem because of the field of the stop cards and the additional qualitative data collected. The additional data, collected provides context around police- community relations. Focus groups extracted details on experiences and perceptions of police in different regions: Central, Mid- City, Southern, and Southeastern. There was also an officer surveys and interviews conducted department-wide that addressed police trust, police nationality, traffic stop data collection, and how they handle racial/ ethnical bias.

The majority of the studies have been focused on "Who gets pulled over", "Post- stop outcomes", "hit rate" (i.e. proportion of is contraband found), and "arrest rate". This analysis was done using "propensity score matching". The values range from 0 to 1 and represent "the probability that a driver will be of a certain race, given certain stop/demographic conditions".This technique "matched" drivers of different races to counterparts with similar demographic and stop- based characteristics in an effort to isolate the effect that race/ ethnicity had on post- stop outcomes.

Although this data is appropriate to answer the question at hand, there are several shortcomings. There is some important information not included that would improve the quality of analysis. This vehicle stop cards do not record race/ ethnicity of the officer, specific geo-location of the stop/ search, make, model, vehicle condition, and driver/ passenger demeanor. In other California- wide studies, the results have been scrutinized because the "report lacked details about whether stops occurred in high- or low-crime areas". The introduction of this specific geo-location could help with this complaint. The vehicle information [not included] has been used in other studies to control data for the confounding factor of economic status. Socio-economic status was one of the causes of racial disparity disparity in the North Carolina study. This information would help distinguish the cause of racial disparities in post stop outcomes; are they a result of differential criminality or something else ("Bad Apple" police, widely shared but individually subtle differences in driver treatment, implicit bias, or Institutional practices).

Several factors have also led SDPD to question the reliability of the data. There is a lot of missing data: 10.6% of citations, 7.9% of field interviews, 4.4% of searches, and 93% of contraband found data. Additionally, some of the search types were unable to be de-aggregated, and therefore making it difficult to assess racial disparities by search type.

Another issue with this data set is using total population as a benchmark to establish rates. The tends to inflate bias for several reasons: the driving population is not equal to the all ages of people and may be different across races, economic factors can influence who drives (and who takes public transportation), one person can be stopped multiple times,commuters may not be residents, etc.

Additionally, the data used was filtered based on the “Veil of Darkness” technique. This is based off the assumption that if offieres are engaged in racial profiling, they are less likely to be able to identify a drivers race at nigh than during the day. The issue with this is that vehicle make, year, and model often correlate with race and are still visible at night, which could lead to the test under-estimating the extent of racial profiling. This technique also doe not take into effect artificial lighting from street lighting.

One of the most popular studies on this topic was North Carolina 2002. In this study, a state- wide investigation of traffic stop data occurred, allowing for the analysis of different cities and different levels of disparity. This study spans a larger time period than the SDSU study and has more levels of disparity. One advantage the North Carolina study had was the presence of officer IDs for each traffic stop. This study also looked at “Who is searched “ and at stop rates.

Part 2

The data for this project originates on the public site: https://data.sandiego.gov/datasets/?department=police. This page has several hyperlinks that link to files that have specific descriptions of different codes and abbreviation used. I chose to use the data from 4 specific pages: police vehicle stops 2014- 2018 (https://data.sandiego.gov/datasets/police-vehicle-stops/), RIPA Police Stop (Racial and Identity Profiling) - basic details (https://data.sandiego.gov/datasets/police-ripa-stops/), RIPA reason for stop (https://data.sandiego.gov/datasets/police-ripa-stop-reason/), and RIPA result of stop (https://data.sandiego.gov/datasets/police-ripa-stop-result/).

Although this site is accessible publicly, and you can download the .csv files for free, there is a lack of data privacy. Since there are no names or driver's license IDs associated with the stop data, privacy for the driver is not a huge issue. Similarly, there is no specific ID for each officer in the data.

In designing the schema, I looked at the Stanford Open Policing cleaned data for an idea of a good organization. Stanford Open Policing uses attributes including:
Stop date - datetime
Stop time - datetime
Stop location(service ares) - int
Driver race - string
Driver sex - string
Driver age - float
SD_resident - bool (0/1)
Reason for stop - string
Search conducted - bool (0/1)
Reason for search - string
Contraband found - bool (0/1)

Property seized - bool (0/1
Citation issued - bool (0/1)
Warning issued - bool (0/1)
Arrest made - bool (0/1)
Outcome(autofilled from arrest, warning or citation) - string

I also used the Stanford Open Policing project (in addition to the SDSU study) to decide which data types would be best.

Old Format:
'stop_id', 'stop_cause', 'service_area', 'subject_race', 'subject_sex', 'subject_age', 'date_time', 'date_stop', time_stop', 'sd_resident', 'arrested', 'searched', 'obtained_consent', 'contraband_found', 'property_seized'

New Format:
'stop_id', 'ori', 'agency', 'exp_years', 'date_stop', 'time_stop', 'stopduration', 'stop_in_response_to_cfs', ‘officer_assignment_key', 'assignment', 'intersection', 'address_block', 'land_mark', 'address_street', 'highway_exit', 'isschool', 'school_name', 'address_city', 'beat', 'beat_name', 'pid', 'isstudent', 'perceived_limited_english', 'perceived_age', 'perceived_gender', 'gender_nonconforming', 'gend', 'gend_nc', 'perceived_lgbt'

From the newer format, I decided to keep the following attributes:
'stop_id', 'exp_years', 'date_stop', 'time_stop', 'stopduration', 'address_city', 'beat', 'pid', 'perceived_limited_english', 'perceived_age','gend'

I chose these because I felt that some of the location data was less useful, and actually had a large amount of missing date ('intersection', 'address_block', 'land_mark', 'address_street', 'highway_exit', 'isschool', 'school_name', 'address_city'). Additionally, I thought in the new format there was overlap in the gender/ sex that the old format didnt have. Also, since this data was only from San Diego, I chose not to include the agency.

Since I do not merge many csv files (and I drop columns I do not need as I go), I tried to minimize the storage needed. I also chose to keep columns that did not have a lot of missing date (and I believed were the most important).

The schema is set up to take in any year and output the traffic stop data for that year (pre- cleaned slightly). In the future, I would try to generalize the pipeline even more to be able to take in a general website and then navigate through it finding different data by year. For example, the Stanford open policing website has good data that is very well organized and nation- wide. Additionally, census data would be useful to incorporate but there are more security precautions for accessing that data.

Part 3

See github repo link: https://github.com/StephanieMoore14/DSC180A_Assignment1