**Los Angeles Arrest Data in 2010-2019 using SAP/SAC**

Authors: Gunzal Gaji, Raymond Tsang, Xiomara Iraheta, Bill Evashwick, Gia DiGiacobbe

BUS 5100-93 Data Analytics

[ggaji@calstatela.edu](mailto:ggaji@calstatela.edu), [rtsang@calstatela.edu](mailto:rtsang@calstatela.edu), [xirahet@calstatela.edu](mailto:xirahet@calstatela.edu), [bevashw@calstatela.edu](mailto:bevashw@calstatela.edu), [gdigiac@calstatela.edu](mailto:gdigiac@calstatela.edu)

**Abstract:** Our term project paper will demonstrate the relationships between arrest rates and demographic factors for arrests by the Los Angeles Police Department from 2010-2019. We considered demographic indicators such as sex, age, and arrest type in relation to the metro areas in Los Angeles. Utilizing SAP Analytics Cloud, we will present insights using visualizations such as tempo-spatial analysis and geo mapping and machine learning techniques like time series and regression.

**1. Introduction**

This project uses SAP-SAC to discover the correlation between police station location and arrest types. We will also explore other factors such as age, gender, and descent.

We have chosen this dataset because it reflects a major issue around safety in Los Angeles County. Since arrest and crime are huge issues in Los Angeles, we wanted to see whether there was a correlation between area and type of arrests made. We were also interested in confirming whether recent news reports by the Los Angeles Times that point to an overall decrease in arrests, starting in 2013, with rates dropping by 25% from 2013 to 2015 were accurate.[1]

**2. Background**

As one of the leading cities in the world, committed to transparency and accountability, Mayor Eric Garcetti launched the L.A. Open Data Portal in May of 2014. [2] This data portal makes available to the public raw data and easy to use dashboards for residents to analyze key and previously inaccessible data from governmental agencies such as the Los Angeles Police Department.

Our data is transcribed from original arrest reports that are typed on paper by police officers, which results in some inaccuracies and inconsistencies within the data. However, the dataset is monitored and updated on a weekly basis by LAPD with the last data update being on the 24th of February of 2021.

**3. Related Work**

On a national level, the Federal Bureau of Investigation’s Uniform Crime Reporting (UCR) program compiles official data on crime across the United States.[2] Eighteen thousand city, college, county, state, tribal, and federal law enforcement agencies participate voluntarily and submit their crime data either through a state UCR program or directly to the FBI's UCR Program.

The Vera Institute of Justice notes that the UCR data is large, complex, and contains only arrest data voluntarily reported by each of the country’s police agencies. These data complexities and the disparities between the way different agencies report information make it challenging to compare key indicators across different agencies. In order to address this issue, the Institute constructed *Arrest Trends*, an accessible platform where users can learn more about how arrest trends vary by offense type, time, and place.[2] Our LA dataset required clean up to account for some of the inconsistencies with self-reporting that the Institute mentioned; nonetheless, we were able to group the data to determine the relationships between police department location and key demographic variables.

Almanie, et al completed an analysis of crime types in Los Angeles using spatial and temporal criminal hotspots. They used this information to highlight factors that might affect the safety of neighborhoods and help agencies predict crimes by location and time. Their analysis consider the busiest months, days of the week, and times of day for crime, as well as the percentage of crimes by location/neighborhood. [8]

Another similar work in Los Angeles crime data analysis we reviewed was a Master’s thesis by Bianca Cung. Her work explores the relationships between demographic background, weather, and crime using LAPD crime data, weather data from the National Climate Data Center, and Census Data for demographics. The paper focused mainly on ethnicity and gender as demographic variables with the goal of identifying crime patterns in relation to weather patterns, day of the week, and location. The key variables were identified through contingency tables and Chi-squared test; clustering and regression were also included. This thesis analyzes over 300,000 crime incidents from 2009.[3] By comparison, our data did not layer on additional data sources like census or weather data, though this does give us ideas for opportunities for future research and analysis on this subject.

The correlation between age and arrests is frequently discussed in the literature. Crime peaks in early adulthood and then begins to taper off, with adults younger than 35 years old accounting for the greatest number of arrests, while those under 24 are responsible for a disproportionate amount of violent crime. [5] [6]

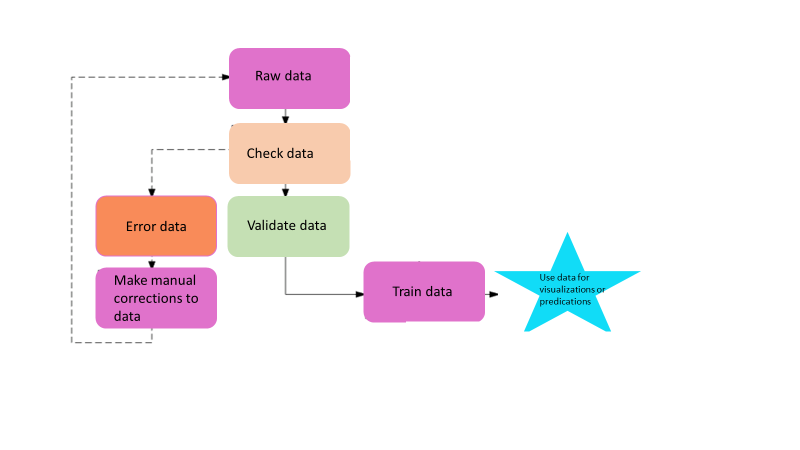
Meanwhile, the Public Policy Institute of California notes that prior to the pandemic, the crime rate had been decreasing and was at a lower level than 2010, before California undertook major criminal justice reforms, and is comparable to the crime rate in the 1960s. Crime further trended down in 2020, but that data may not be relevant for future predications due to the impact of the Covid-19 pandemic with more people staying home. [7]

**4. Specifications**

The dataset included location of police stations and we used this information to create a zip code column to enable us to complete the geomapping visualization. The dataset was sourced from data.lacity.org and was 245MB, spanning the years of 2010-2019. It originally included 1.32M rows with each row representing an arrest and 25 columns, which provided a variety of internal codes and details about the arrest. We explored age, ethnicity, arrest code type across Los Angeles. Since the dataset covers 10 years and is large, we decided to use it as a tool to predict future arrest numbers and the characteristics of those arrests.

**5. Implementation Flowchart**

The raw dataset was obtained from the city’s L.A. Open Data Portal on data.lacity.org. After retrieving the raw data set, we checked the data for quality, deleted erroneous data, and made changes within the excel file to prepare it for analysis. After all the additions and corrections were made, we were able to upload and use this edited dataset.



*Figure 1-implementation flowchart*

**6. Data Cleaning**

Raw files were downloaded to our computer and the following edits were made:

1. The original data set was obtained from the public site and named “Arrest\_Data\_from\_2010\_to\_2019”

<https://data.lacity.org/Public-Safety/Arrest-Data-from-2010-to-2019/yru6-6re4/data>

1. We created a file with the address of the 21 police stations obtained from the following site and named it “LAPD Address”

<http://www.laalmanac.com/crime/cr70a.php>

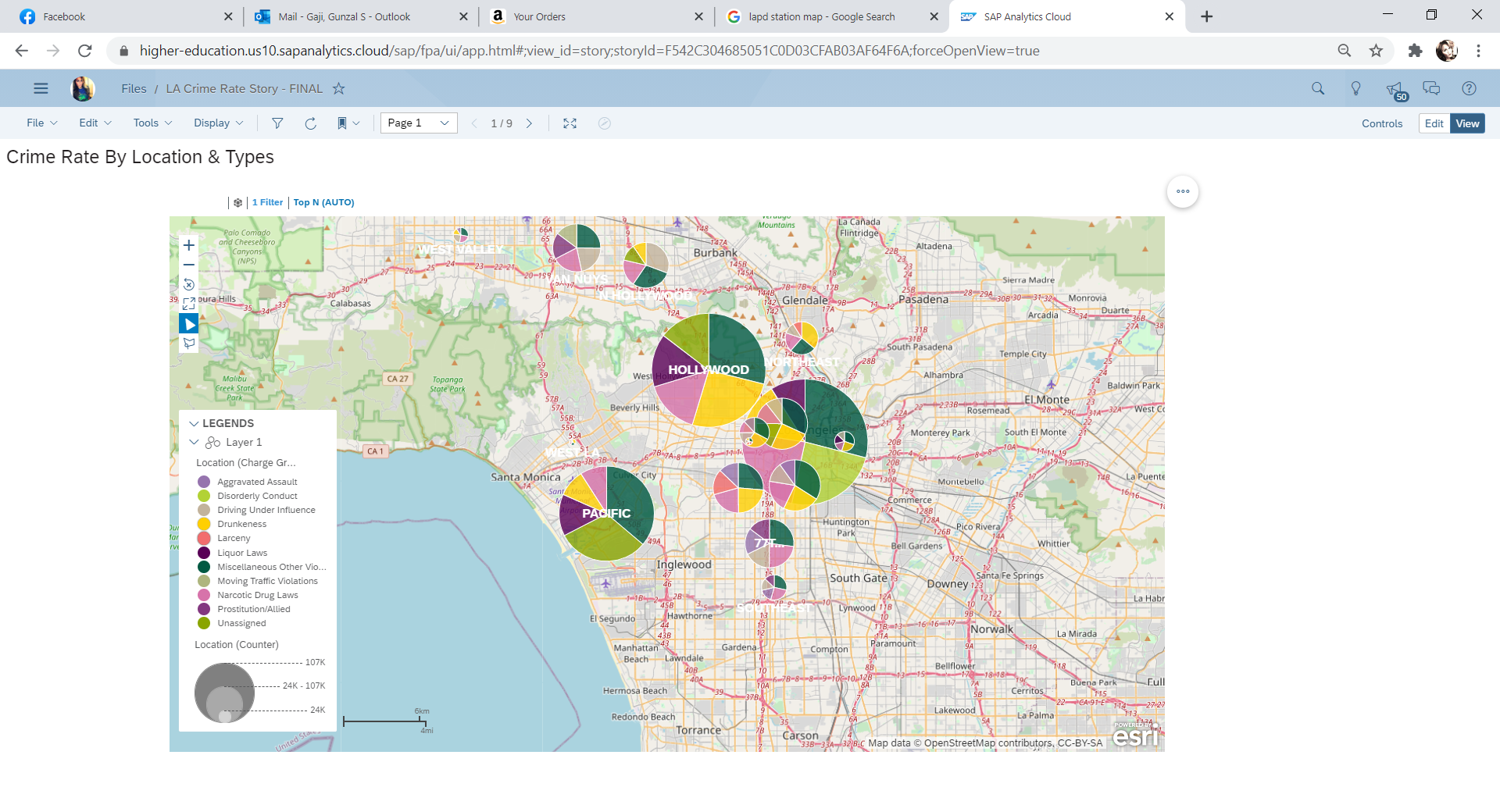
1. We located the latitude and longitude of the 21 police stations from the following site <https://www.findlatitudeandlongitude.com/>
2. We incorporated the above information onto the excel file “LAPD Address”
3. We then incorporated the additional data from the “LAPD Address” file to the original “Arrest\_Data\_from\_2010\_to\_2019”file via excel using the “VLOOKUP” function matching “Area Name” on “Arrest\_Data\_from\_2010\_to\_2019” file to “Area Name” of “LAPD Address” file, and added the columns “ZIP”, “LATITUDE”, “LONGITUDE” & “DEPLOYED” with the respective data from the “LAPD Address” file.
4. We changed the newly added columns - “LATITUDE”, “LONGITUDE”, “ZIP”, “DEPLOYED” format from formula to “VALUE”.
5. We then did the final clean up by eliminating the unnecessary columns, (J, N, Q, R, S, T, U, V, W, X, Y) -
6. “CHARGE GROUP”, “CHARGE”, “Address”, “Cross Street”, “LAT”, “LON”, “LOCATION”, “BOOKING DATE”, “BOOKING TIME”, “BOOKING LOCATION”, “BOOKING LOCATION CODE’.
7. We created a new column named “COUNTER” and populated the column with a value of “1” with a final total of 19 columns
8. The data set “Arrest\_Data\_from\_2010\_2019 (10) was ready for use.

**7. Analysis and Visualization**

After data cleaning and preparation for further analysis, files were extracted into Excel. We created a variety of different interactive visualizations in order to show arrest types per location, arrest types per charge group, arrest rate by gender, total arrests per location and top arrest types. We also developed a forecast for future arrest numbers.

**7.1 GeoMap**

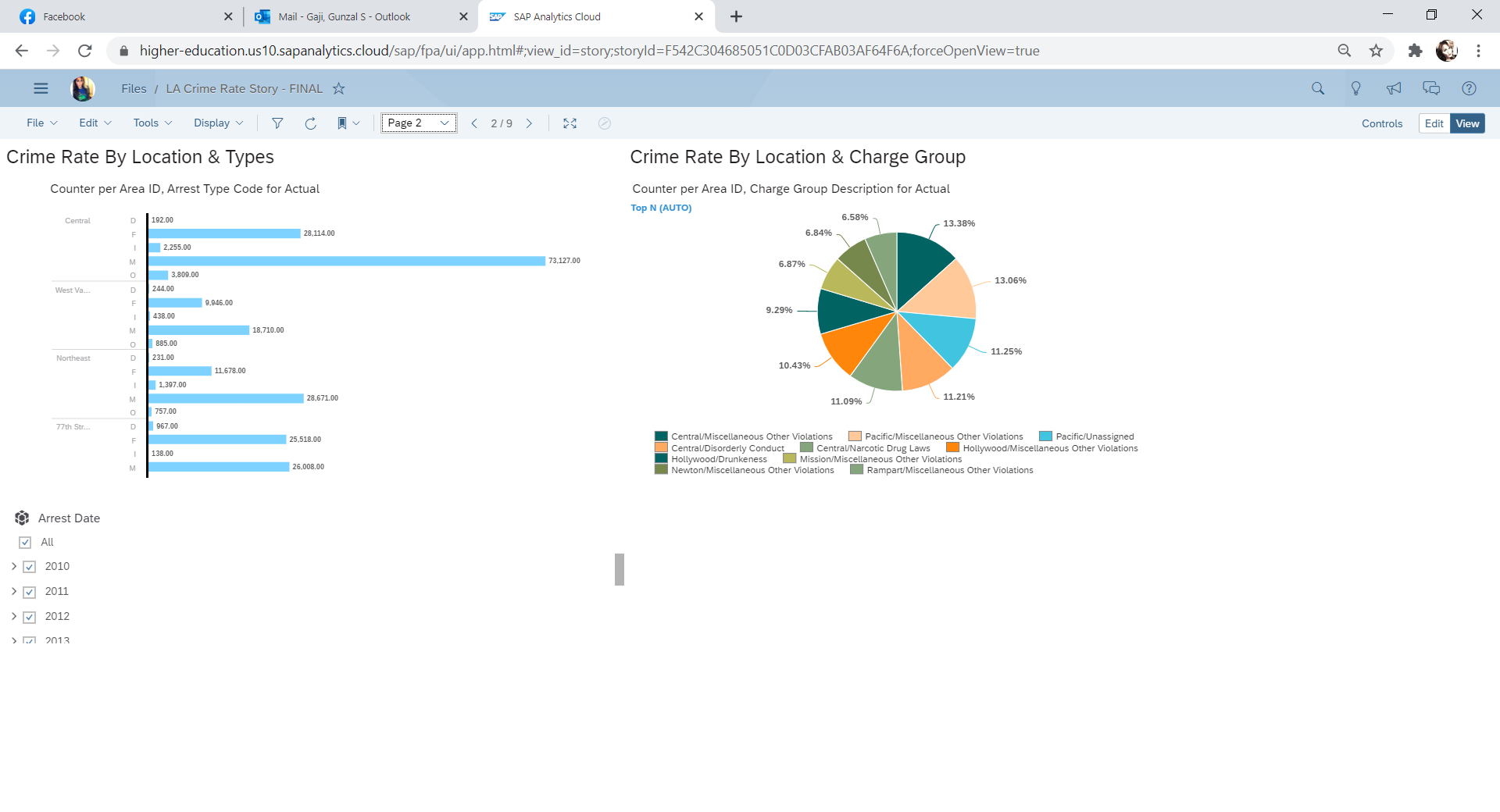
The first visualization (*Figure 2*) is a geomap that highlights the location where arrests were made and showing what type of charges were reported per area. As you can see, both Hollywood and Central have a comparatively large number of arrests.



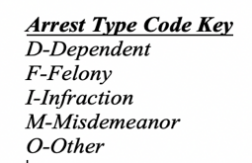
*Figure 2-Arrest Types per location in Los Angeles*

**7.2 Bar Graph**

The visualization *(Figure 3)* is a bar chart which show the highest types of arrest by code and area. You can also filter through the years to have a closer look at trends from year to year. The chart shows that miscellaneous arrests are the largest in every area followed by felonies. Miscellaneous arrests include infractions such as disturbing the peace, open container, curfew violations, and operating a taxi without a license. Arrest codes can be differentiated from one another by using *Figure 4* as a key.



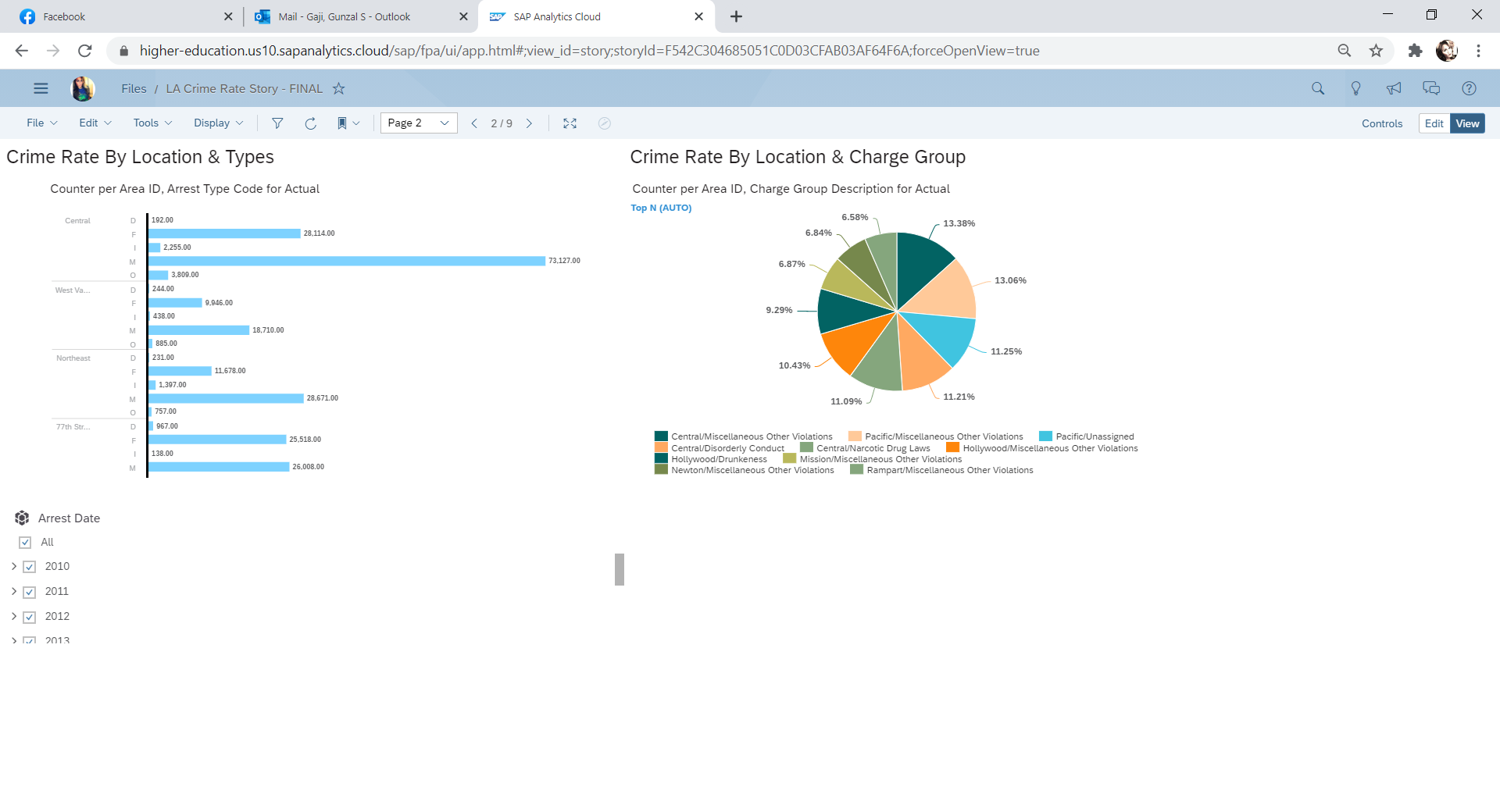
*Figure 3-Arrest Types per area and arrest code*



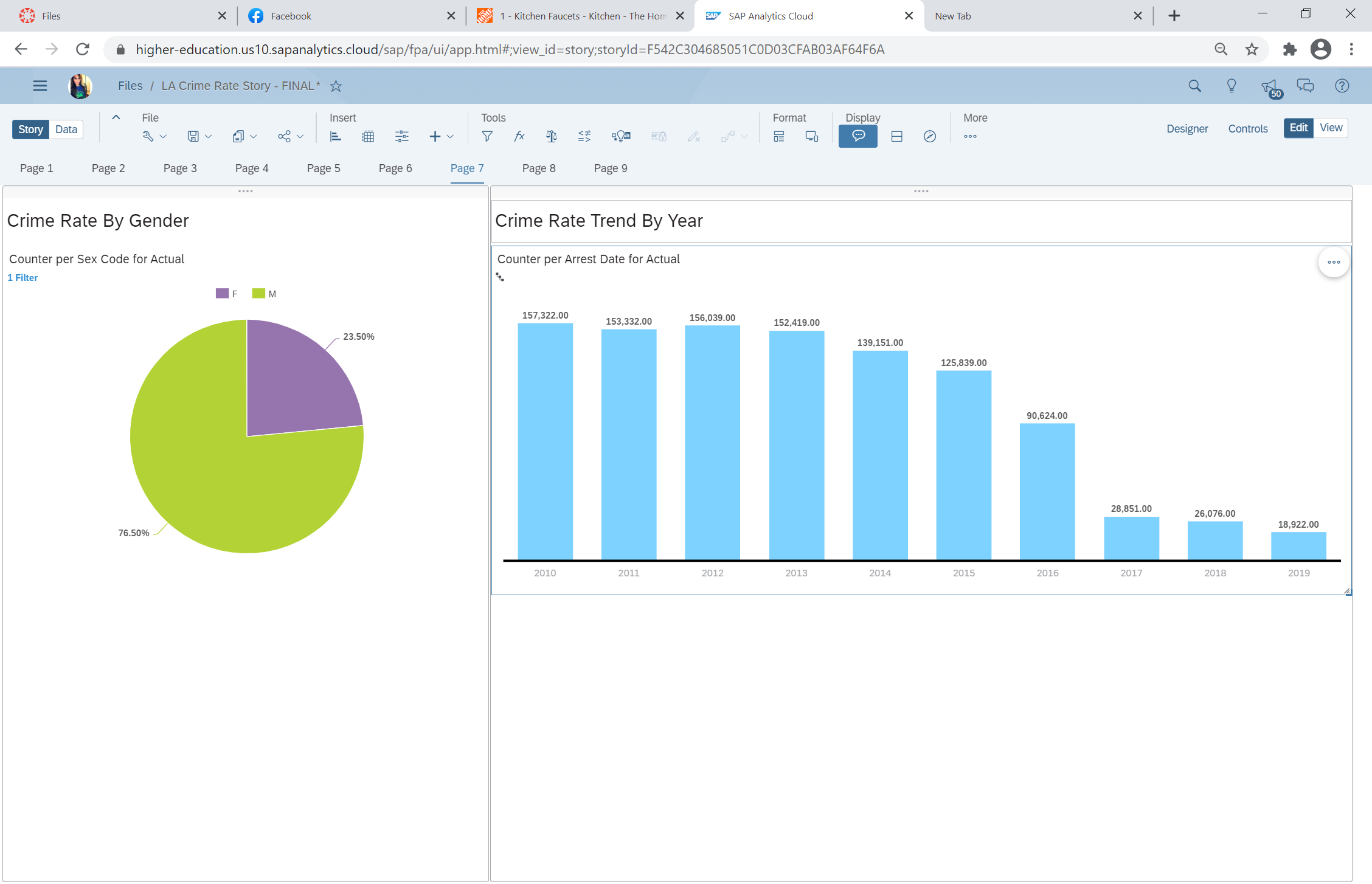
*Figure 4-Arrest Type Code Key*

**7.3 Pie Chart**

From *(Figure 5)* you can see arrests by charge group and Area ID in relation to one another. The chart notes that Central and Pacific have the highest “miscellaneous and other violations” as charge group descriptions reported.

*Figure 5-Arrest Types per charge group*

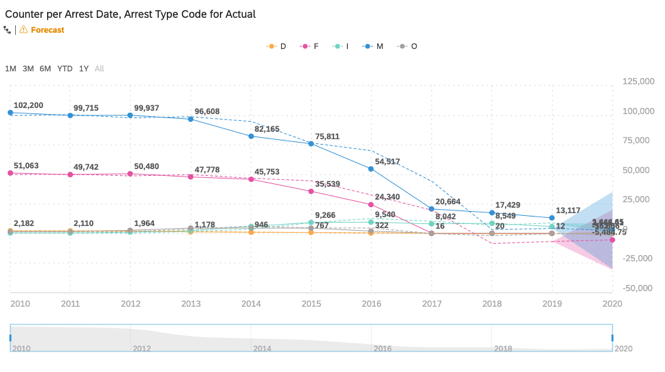
Our analysis *(Figure 5)* also showed that more men are arrested than women. You can see from the pie chart that men account for about 76.5%, while women only make up about 23.5% of total arrests.



*Figure 6-Arrest Types by gender*

**7.4 Time Series Forecast**

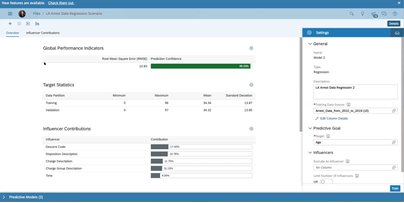
Based on the time series forecast *(Figure 6),* you can see that arrest rates have been on the decline since 2010. Based on the most reliable forecasting indicator in our analysis, we can predict that arrest rates will continue to decline through 2020.

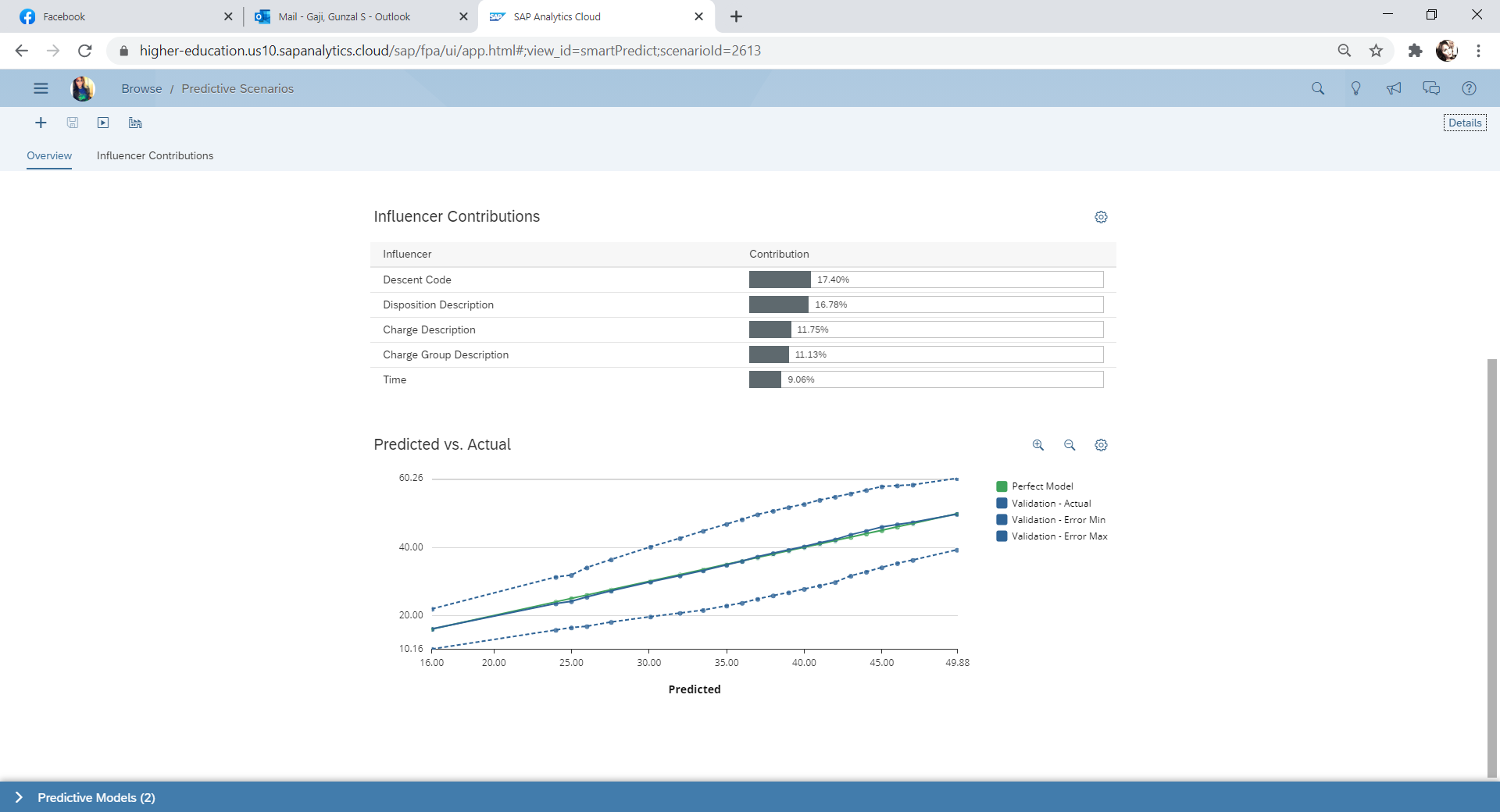


*Figure 7- Time Series Forecast*

**7.5 Regression**

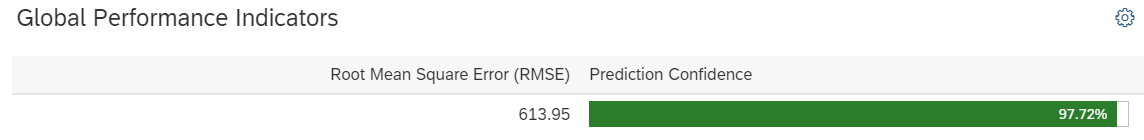
Our first regression (*Figure 7*) predicts the age of the perpetrator that was being arrested. The regression shows that the mean age for those arrested is 34 years old. The regression has a confidence level of 99.2% and a root mean square error of 11 years, which means that it predicts age within about 11 years of the actual age. If you compare the actual values to the predicted values in our forecast, the model was a good predictor, which you would expect from comparing the graph that shows predicted vs. actual.

****



*Figure 8 – Age Regression*

We then ran a second regression to predict the time arrests occur, which may have implications for agencies, as consider their staffing. We had to do additional data cleaning to perform this regression and it had a prediction confidence of over 97% and a root mean square error of 614, which is about four hours (as the data used military time).





*Figure 9 - Time Regression*

Our regression found that the type of charge was the greatest influencer on the time an arrest record, followed by disposition, reporting district, and age and descent of the perpetrator.

**8. Conclusion**

Based on our analysis, we have determined the following:

1. Misdemeanor and miscellaneous arrest charges were the most common/frequent made
2. Arrests have been declining since 2010
3. Men get arrested more than women.
4. The type of charge was most influential on the time of the arrest.

Based on the substantial data, we were able to predict that arrest rates will continue to decline through 2021. We were also able to see that miscellaneous and misdemeanor arrests were committed most frequently and that men were arrested at a much higher rate than women. More analysis can be done with additional datasets and more variables. For more information, dashboard, and codes visit our project’s GitHub link3.

**References**

**Github Link**: <https://github.com/BUS-5100-Group-2>

**SAP/SAC Story Location**: <https://higher-education.us10.sapanalytics.cloud/>

Cal State LA folder > Rtsang > Final Project Group 2

LA City Arrest Data Link: <https://data.lacity.org/Public-Safety/Arrest-Data-from-2010-to-2019/yru6-6re4?category=Public-Safety&view_name=Arrest-Data-from-2010-to-2019>

[1] Queally, J., Mather, K., & Chang, C.

(2017, April 1). Police arrests are plummeting

across California, fueling alarm and questions.

Retrieved March 21, 2021, from

[https://www.latimes.com/local/lanow/la-ln-police-](https://www.latimes.com/local/lanow/la-ln-police-slowdown-20170401-story.html)

[slowdown-20170401-story.html](https://www.latimes.com/local/lanow/la-ln-police-slowdown-20170401-story.html)

[2] Mayor garcetti launches open data portal for city

of los angeles at #techla conference. (2016,

February 11). Retrieved March 28, 2021, from

[https://www.lamayor.org/mayor-garcetti-launches-](https://www.lamayor.org/mayor-garcetti-launches-%20%20%20%20%20%20open-data-portal-city-los-angeles-techla-%20%20%20%20%20%20%20conference)

[open-data-portal-city-los-angeles-techla-](https://www.lamayor.org/mayor-garcetti-launches-%20%20%20%20%20%20open-data-portal-city-los-angeles-techla-%20%20%20%20%20%20%20conference)

[conference](https://www.lamayor.org/mayor-garcetti-launches-%20%20%20%20%20%20open-data-portal-city-los-angeles-techla-%20%20%20%20%20%20%20conference)

[3] O’Toole and Neusteter, R. (2019) Arrest Trends.

Retrieved From

<https://www.vera.org/downloads/publications/Arrest-Trends-Data-and-Methodology.pdf>

[4] Cung, B. (2013). Crime and Demographics: An Analysis of LAPD Crime Data. *UCLA*. ProQuest ID: Cung\_ucla\_0031N\_11420. Merritt ID: ark:/13030/m52b99n3. Retrieved from <https://escholarship.org/uc/item/2v76v571>

[5] Rocque, M., et al. (2016). Age and Crime. *The Encyclopedia of Crime and Punishment.* Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118519639.wbecpx275

[6] Steffensmeier, Darrell J., et al. “Age and the Distribution of Crime.” American Journal of Sociology, vol. 94, no. 4, 1989, pp. 803–831. JSTOR, www.jstor.org/stable/2780859. Accessed 26 Mar. 2021.

[7] Lofstrom, M., et al. (2021). “Crime Trends in California.” <https://www.ppic.org/publication/crime-trends-in-california/>. Accessed 29 Mar. 2021.

[8] Almanie, T., et al. (2015). “Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots.” *International Journal of Data Mining and Knowledge Management Process.* Vol.5,No.4.