University of San Diego
SHILEY-MARCOS
SCHOOL OF ENGINEERING

# ADS-506 – Final Team Project Start Form

Fill out this form and submit it by the end of Module 2 in Blackboard.

Team Number: **1**

Team Leader/Representative: **Andrew Kim**

Full Names of Team Members:

1. **Andrew Kim**
2. **Jonathan Yoon**
3. **Grigor Tashchyan**

Title of Your Time Series Final Project: **Addressing the Issue of Potholes in San Diego**

Motivation for choosing this project:

**This topic is one that affects us on a daily basis. Potholes can cause traffic to slow down and in severe cases, even cause interior damage to vehicles.**

Problem Statement: Short Description of Your Time Series Project and Objective(s):
**By analyzing the pothole data request dataset, we can determine just how effectively the city is repairing the potholes and thus giving them insights on where they are doing well versus where they might need improvement. Classification zones can be helpful for predicting if a pothole would be repaired, in order to address certain countries that have repair rates. For example, zones with higher population the more potholes there are leading to a higher number of pothole requests being closed; while rural districts have less reports to have potholes repaired. By using our Data Science techniques, we can help identify what causes these potholes and eventually help prevent them.**

Name of Your Selected Dataset: **Pothole Repair Requests.**

Description of your selected dataset:
**The dataset contains 26,620 cases of pothole repair service requests. Each service requests contains its unique service request identification number, requested and closed dates, age of the request in days, case record type, status, location of the service in latitude, longitude, address, and zip code, district number, community identification in code and name, how the request was submitted (web, phone, mobile, etc.), whether the case was referred to any outside vendors, and public description. Some of these variables contain missing values, and investigations on these missing values and potential outliers will be performed, along with analysis.**

Data source, number of variables, size of dataset, etc: **The dataset was retrieved from the City of San Diego datasets tab. It contains 26,620 entries (rows) and 23 variables (columns).**

Notable findings from your initial EDA:

**The first notable discovery that was made towards the dataset was that nearly half of the variables contained missing data. The variable “public_description” also contained duplicated message data entries. For the purpose of our project, we have decided to drop these factors from the dataset to avoid generating biased estimates or invalid conclusions.**

**Another discovery made towards the dataset was that some variables duplicated data values of other variables, which included “iamfloc” and “floc”. These variables along with “service_request_id” were imputed because the values or data within the variables were not deemed important nor could be used as predictors.**

**A box plot was used to evaluate any outliers within the dataset. It was discovered from the visualization that the data based on the feature “case_age_days” was skewed to the right, with a majority of ticket requests addressing pothole cases**

**ranging to 200 days before the cases were closed. Another boxplot that was computed generated a report that the TSW department was assigned a multitude of ticket cases involving pothole incidents compared to other departments with a value of 75,223. With the remaining departments each being assigned ticket cases that did not exceed double digits, this indicates that the dataset is highly unbalanced.**

Github link: **https://github.com/hjyoon16/ADS506-Team1/tree/main**