TE360
Jack Chen
9/19/2022

Integrative Project Milestone 1

Part 1

I found it interesting to see a sudden increase of building permits between August 2005 to January 2006, increasing from 142 to 4181 building permits. It might be that the government began actively collecting the data.

I also noticed an annually recurring pattern that the number of building permits slowly increases during the year and peaks sometime in fall and decrease dramatically as time reaches New Year, before significantly increase again. I think it might be explained by the holidays and the cold weather.

Part 2

Q1

From the example subset, 'REVIEW_TYPE', 'APPLICATION_START_DATE','STREET DIRECTION','COMMUNITY_AREA','LATITUDE', 'LONGITUDE', 'LOCATION' all have missing values. And upon examination, the missing values are all “NaN” despite these fields having different dtypes, notably float64 and object.

Q2

I figured the best way to treat the missing values is to remove them. It keeps the originality of the data comparing to artifically manipulate the data using machine learning, and it is quick and simple. Since we have more than 700k rows, removing a portion of the data does not change our dataset significantly.

Q3

The subset now has 609559 rows and 10 columns.

Part 3

Hypotheses 1: Building fees tend to be cheaper when the location is far from the city.
Plan: I will gather the physical location of the building calculate the distance to the center of the city. I then will sort based on that distance and see whether there is a trend in price paid.

Hypothesis 2: There amount of out of state customers are increasing.
Plan: I will take the top 3 contacts information into account and accumulate with respect ot time. I will then analyze the percentage of customer from out of state.

Hypothesis 3: Some community areas are more favored.

Plan: I will find the number of buildings built in each community area and see whether there is a significant difference or a promising growth.