DATA 602: Assignment 1 Summary Report

Sana Sharma

Baltimore County Liquor License Data Prediction

Background

The dataset is obtained from data.gov. It is provided by the Baltimore City Liquor License Board and describes the Baltimore County Liquor License Status from 2003 - 2017. It includes 20.8K observations and 19 columns.

Goal

The goal of the analysis is to predict zip code according to the rest of the data factors like fee, class etc.

Objective

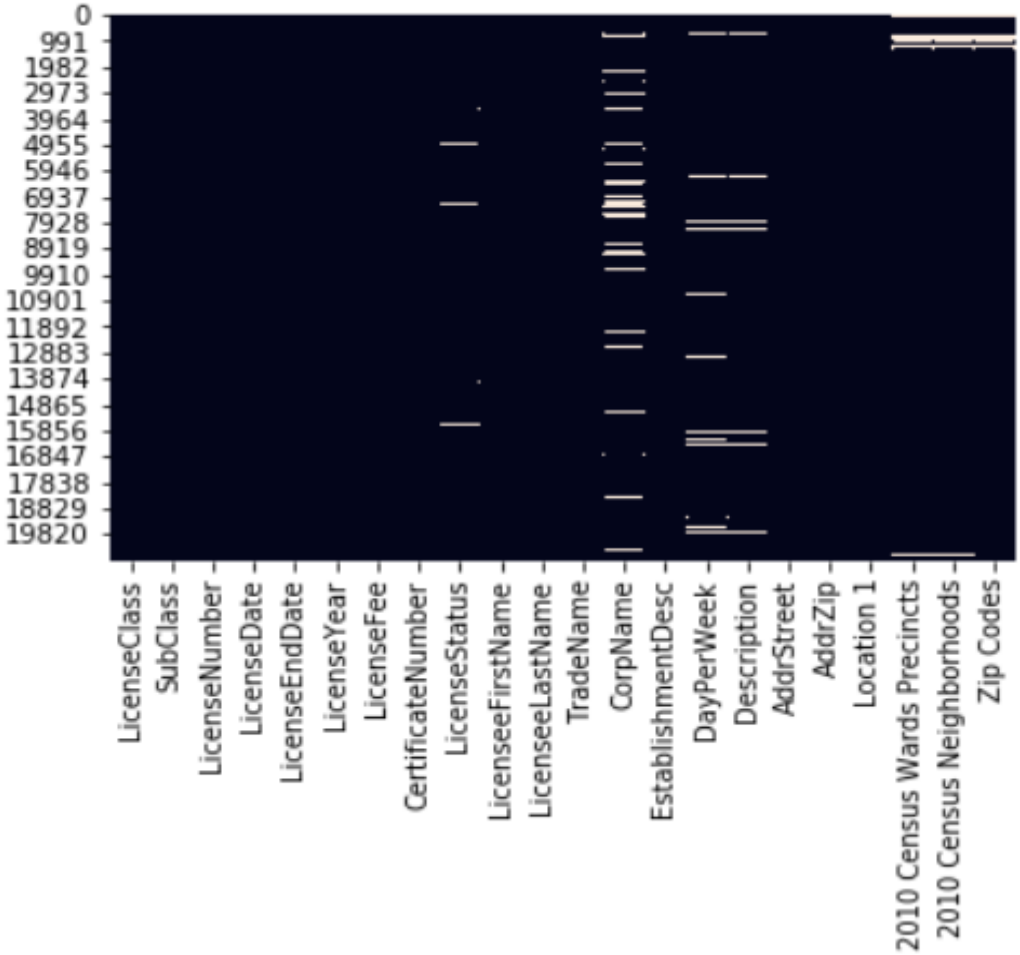
I performed by analysis in three steps –

1. Data Cleaning
2. Regression Modelling
3. Visual Analysis

1. Data Cleaning –

The dataset that I downloaded: Liquor_License.csv - had 20794 rows and 22 columns. It was a pretty-clean dataset and did not require much cleaning. Through heatmap I did observe some missing values in the dataset. After dropping the NaN values, the cleaned dataset was down to 18166 rows and 22 columns. Lastly, saved the clean dataset onto a new csv file: ll_cleaned.csv

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x1c1cb274888>
```



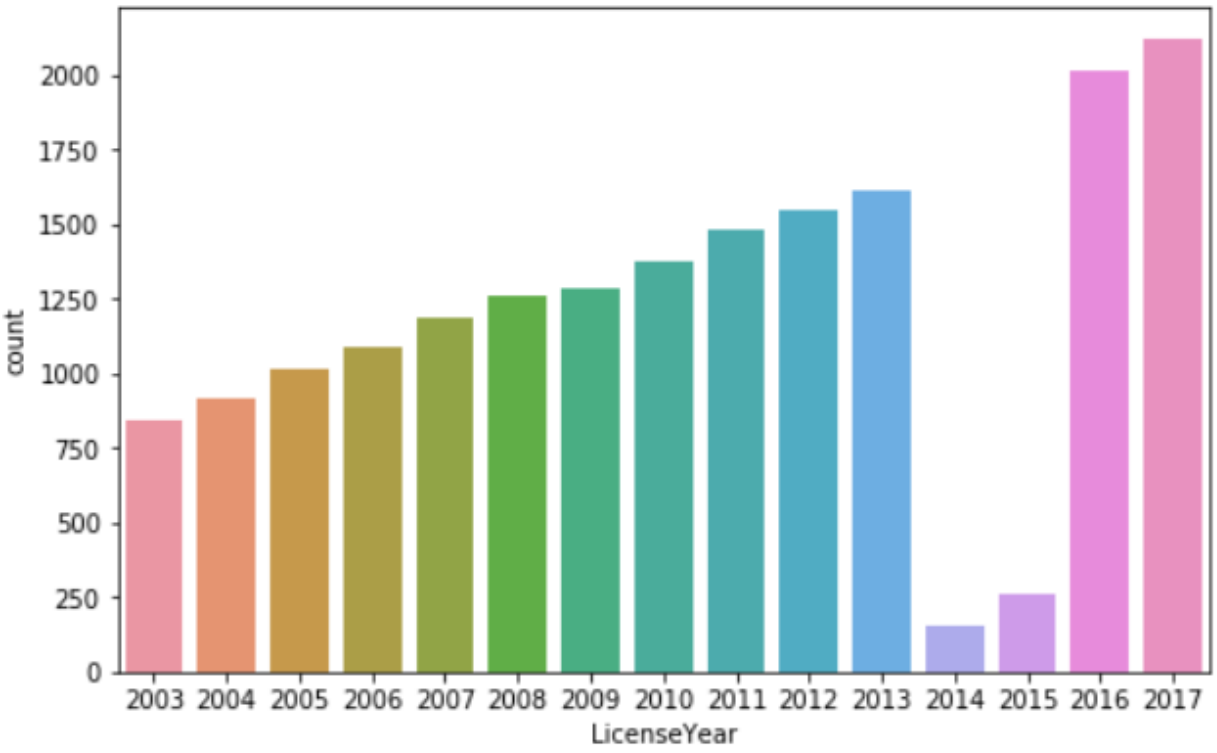
2. Regression Modelling –

Created a new data frame with only the columns required for the analysis and identified the unique value of Zip Code. I ran the regression using sklearn with a test-train split and reached at a prediction of 0.75

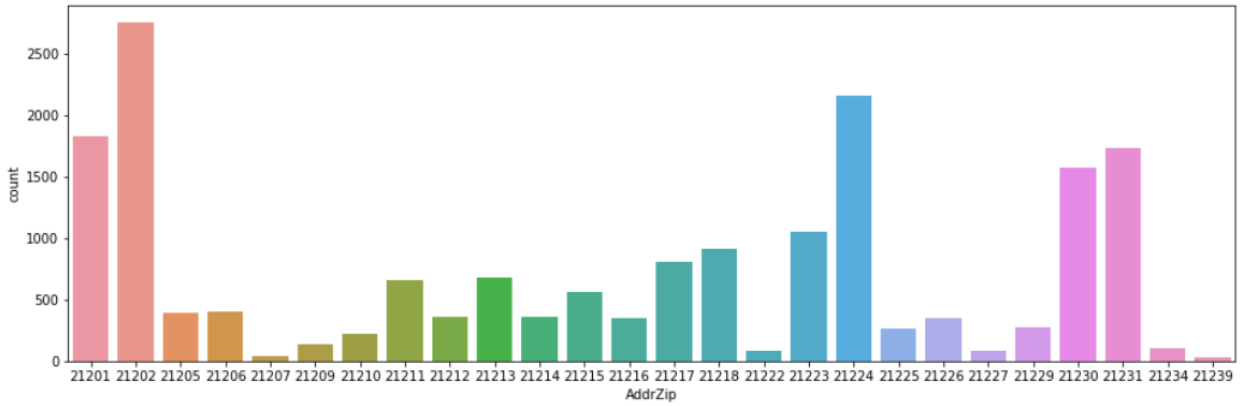
3. Visual Analysis –

For the visual analysis I used seaborn.

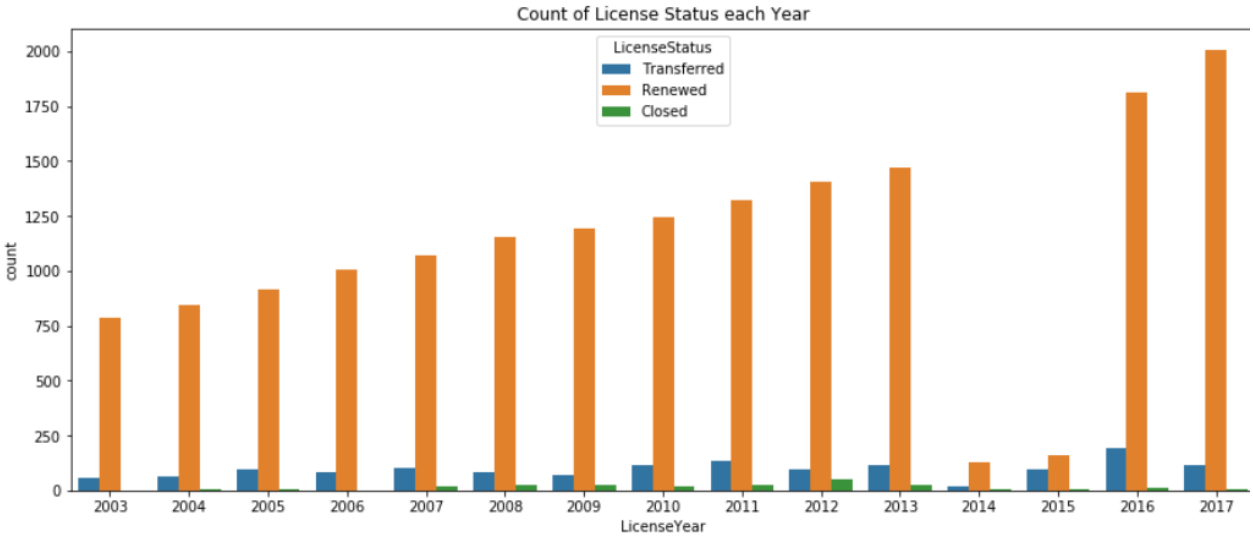
Through the count plot we can see that the count of license increased each year until 2013 and drastically dropped in 2014 and 2015. But sot back higher up in 2016 and 2017.



The below graph shows the number of licenses obtained in each Zip. 21202 has the highest license count and 21239 the least.



Every year the maximum number of licenses were renewed and only some transferred. Very few were closed.



Summary

The model can predict Zip Codes related to any other data with 75% accuracy. The zip code: 21202 has the highest license count.

Documents Submitted

Notebooks

1. Data Cleaning: LL_data_cleaning.ipynb
2. Regression Modelling: LL_modelling.ipynb
3. Visual Analysis: LL_analysis.ipnby

Data

1. Liquor_Licenses.csv
2. Ll_cleaned.csv

Summary Report

1. Data 602 Assignment 1 Summary Report.docx