

License Plate Detection

DS 4002 – Spring 2024

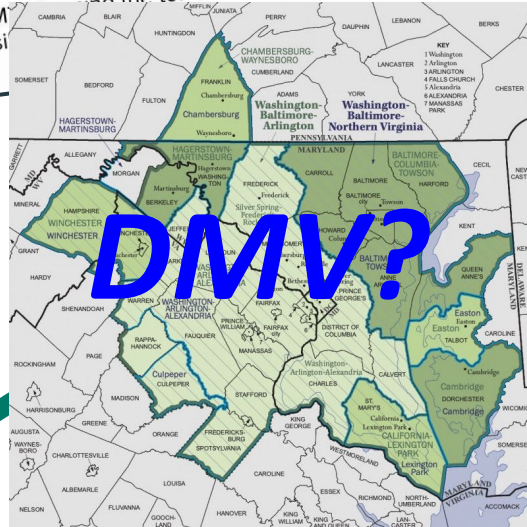
Group 15

Leader: Rishi Raghavan; Members: Ananya Goel, Aidan Wiktorowicz



Why New York will charge up to \$23/day to drive into Manhattan

Drivers crossing through Manhattan's central business district will face a new toll beginning in the Spring of 2024. The fee is expected to be between \$9 and \$23 per day for drivers entering or exiting during peak traffic hours. Regional planners believe "congestion pricing" will nudge more drivers onto transit. The MTA expects the toll revenues to reinvest in its aging infrastructure. The funds will go toward improvements to system reliability, accessibility, and safety.



Project Overview

- **Goal:** Analyze license plate images to determine which state a particular vehicle belongs to.
- **Motivation:** License plate data can be used to conduct research into travel patterns, specifically cross-state traffic. This data can be used to inform infrastructure and policy as well as population movements, economic trends, and regional disparities
- **Hypothesis:** We hypothesize that our model will be able to accurately predict the state of each license plate 90% of the time. We believe this is a good threshold to aim for in validating our model.
- **Research Question:** How accurately can our model detect the state from which a vehicle is from using license plate detection?
- **Modeling Approach:** We used PyTesseract to create our analysis model, after experimenting with alternative LPR programs.

Data Acquisition and Explanation

- **Source:**

- GitHub
- DMV websites of all US states





- **Data**

- 18,000+ distinct license plates from 50+ states of the US

- **Data Cleaning**

- Filtered data to include only DC, Virginia, Delaware, and Maryland plates
 - Xx license plates
 -

Column	Description
state	State postal code abbreviation.
plate_title	The title of the plate collected from the agency website.
plate_img	The filename of the local plate image file.
source_img	The plate image online at the time of collection.
source	The page that contains the plate listing.

State	Description ▲	Plate Image	Source
AL	Alabama - A Pink Breast Cancer Tag		Source
AL	Alabama - Active Reserve		Source
AL	Alabama - Ag Tag (Farming Feeds)		Source
AL	Alabama - Alabama A&M University		Source

Analysis Plan and Justification



Data Processing

```
attachEvent("onreadystatechange",H),e.attachE
boolean Number String Function Array Date RegE
=0; function F(e){var t=[e];return b.ea
[1])!==!&&e.stopOnFalse}{r=!1;real=n=!1,u&
?o=u.length:r&&(s=t,c(r))}return this},remove
ction(){return u=[],this},disable:function()
re:function(){return p.fireWith(this,argument
nding",r={state:function(){return n},always:
romise)?e.promise().done(n.resolve).fail(n.re
id(function(){n=s},t[1]?e[2].disable,t[2]?e
e,n,h.call(arguments),r=n.length,i=i+r||e&
r),i=Array(r);r=t;t+=n[i]&&b.isFunction(n[i]
<table></table>< href="/a">ac</input type
TagName("input")[0],r.style.cssText="top:top
est(r.getAttribute("style")),hrefNormalized:
```

Coding LPR Model

Statistical
Tests



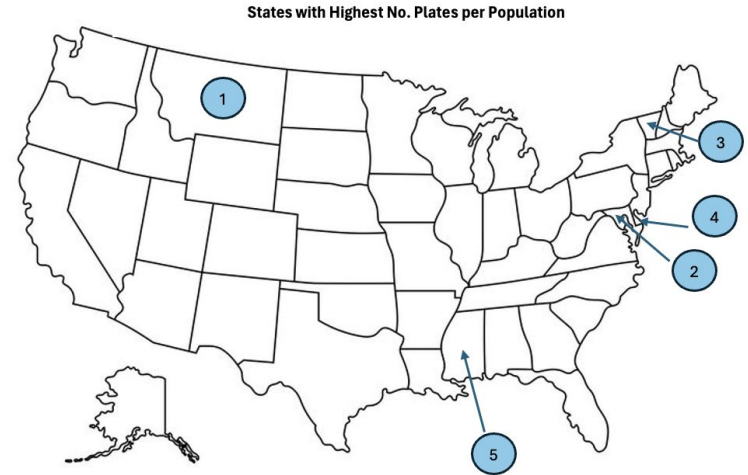
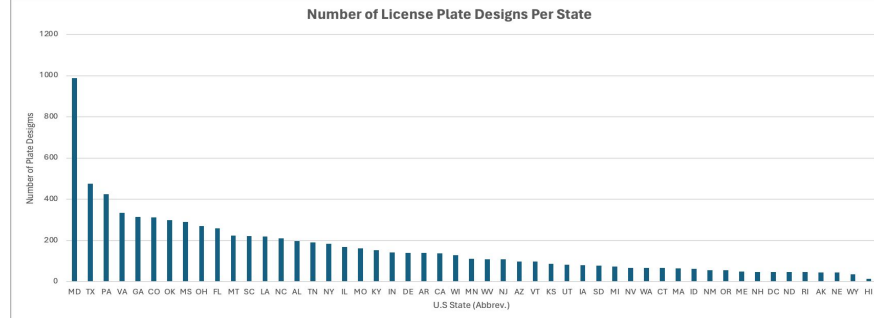
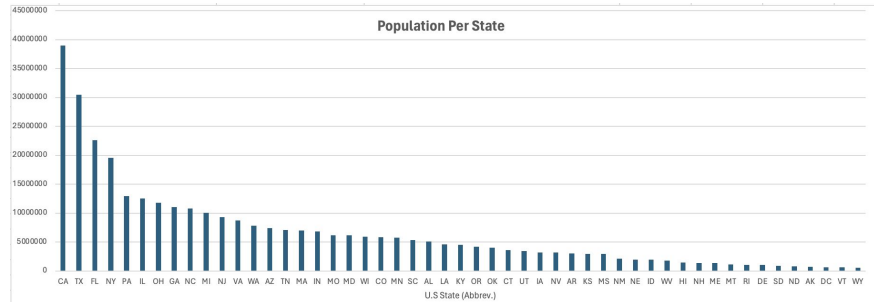
Execute Program on Data
Subset to Assess Accuracy

- 1) **Microsoft Excel** was used to explore data trends between number of license plate designs, population, and state location/geography
- 2) **Visual Studio Code (VSC)** was the main platform upon which we coded our LPR model using Python language. We experimented with using OpenALPR, or PyTesseract as our core engine for LPR.
- 3) **Analysis for Accuracy** was also performed on VSC, and we fed several sets of plates to assess efficacy and accuracy of identifying name of plate.

Data Exploration



Top 5 States by No. of Plates	Top 5 States by Population	States with Highest No. Plates per Population
Maryland	California	Montana
Texas	Texas	Maryland
Pennsylvania	Florida	Vermont
Virginia	New York	Delaware
Georgia	Pennsylvania	Mississippi



Tricky Analysis Decision

Source of Data:

- Originally wanted to create an LPR that can read the licence plate number → challenges with acquiring a large enough data set due to privacy concerns, lack of available data
- For Washington DC, we had 47 plates, much less than other states (350+)

Choosing an LPR Model:

- Open ALPR required extensive image preprocessing that would not be efficient to run on large batches of images given our time constraints
- With PyTesseract, we had large variability in detection of Maryland plates, and were unsure if we'd be able to refine our model with additional coding/processing



State	No. of Plates	Population	Ratio
AK	44	733,406	0.060
AL	197	5,108,468	0.039
AR	140	3,067,732	0.046
AZ	99	7431344	0.013
CA	138	38965193	0.004

Bias and Uncertainty

Sources of Bias:

- Our data source only has neatly oriented pictures of each plate → makes generalizability difficult
- Our character-recognition code may misclassify states, or in handling errors fail to classify plates that otherwise could be identified by the human eye
- Delaware was included as a possible state a plate could be registered to as a negative control (no plates should match to Delaware)



Results and Conclusions

Single-State Accuracy Tests

State	Accuracy	No. of Plates
Virginia	81.10%	334
Maryland	62.10%	990
Pennsylvania	92.50%	349
Washington D.C	91.50%	47

All States except Maryland had less than 7 plates match to Delaware (negative control)

Test Run #1

State	No. of Plates	No. of Detected Plates
Virginia	52	51
Maryland	50	34
Washington D.C	47	57
Delaware	0	7

Preliminary Conclusions:

- PA, VA, DC had relatively high accuracy in correct detection of plates
- PA and DC satisfied our initial hypothesis of > 90% accuracy
- Maryland has a poor accuracy (< 80%) in single-state test + test run

Discussion

- Detection of Virginia plates is the most accurate given both tests
- Maryland has huge variability in plate design, state font, location of state name → explains poor state detection

Next Steps

- More expansive dataset!
- Work on model refinement to increase accuracy for Maryland Detection
- Expand to more states along the East Coast
- Consider creating our own ML model from scratch to see if there's increased accuracy



References

1. <https://www.statsamerica.org/sip/Economy.aspx?page=pi&ct=S51>
2. <https://ggwash.org/view/79300/how-do-we-define-our-region-here-are-some-ways-to-look-at-it>
3. <https://www.nytimes.com/2024/03/21/nyregion/congestion-pricing-nyc.html>
4. <https://github.com/jonkeegan/us-license-plates?tab=readme-ov-file>
5. <https://survisiongroup.com/post-what-is-license-plate-recognition>
6. <https://www.nyclu.org/report/automatic-license-plate-readers>