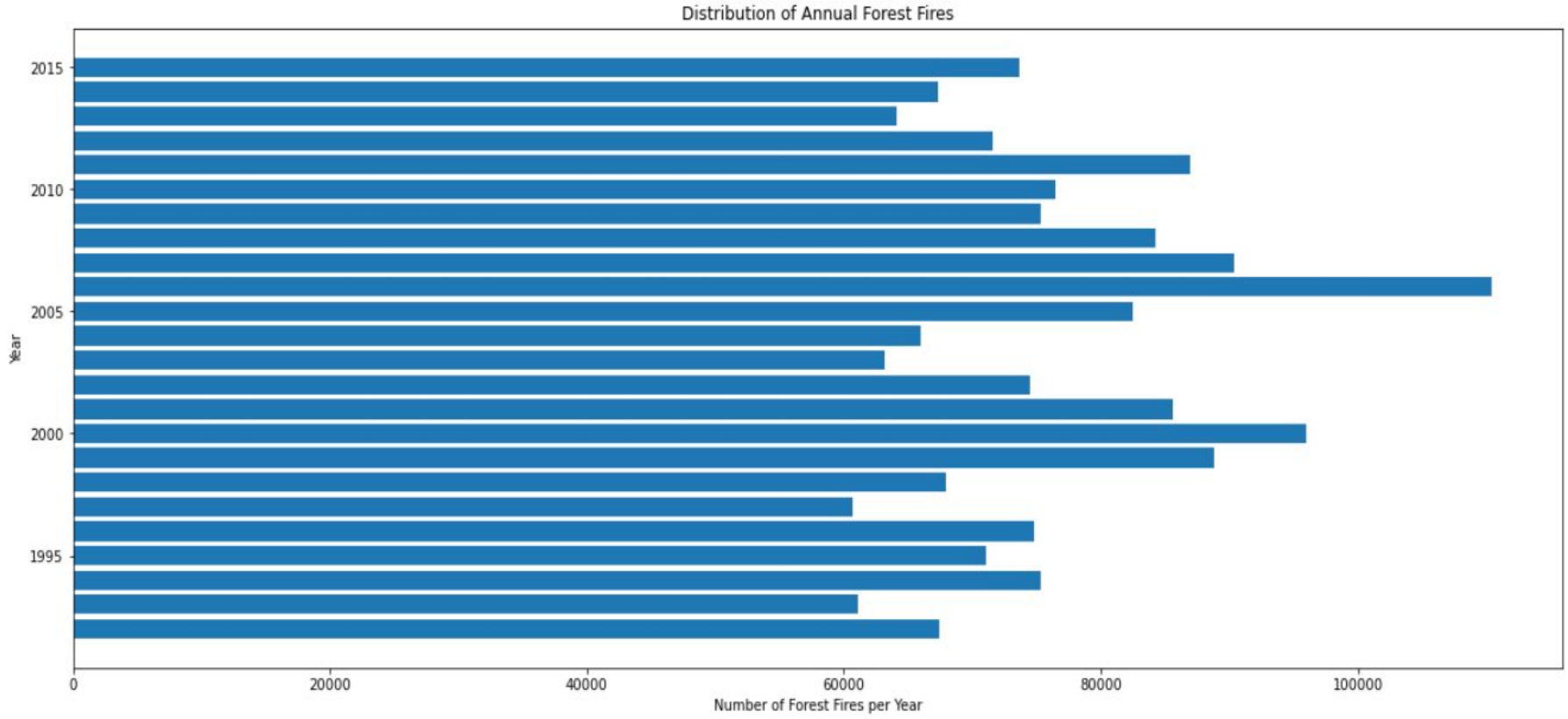
# Project 5: Forest Fires DS

By Evgeny Didenko

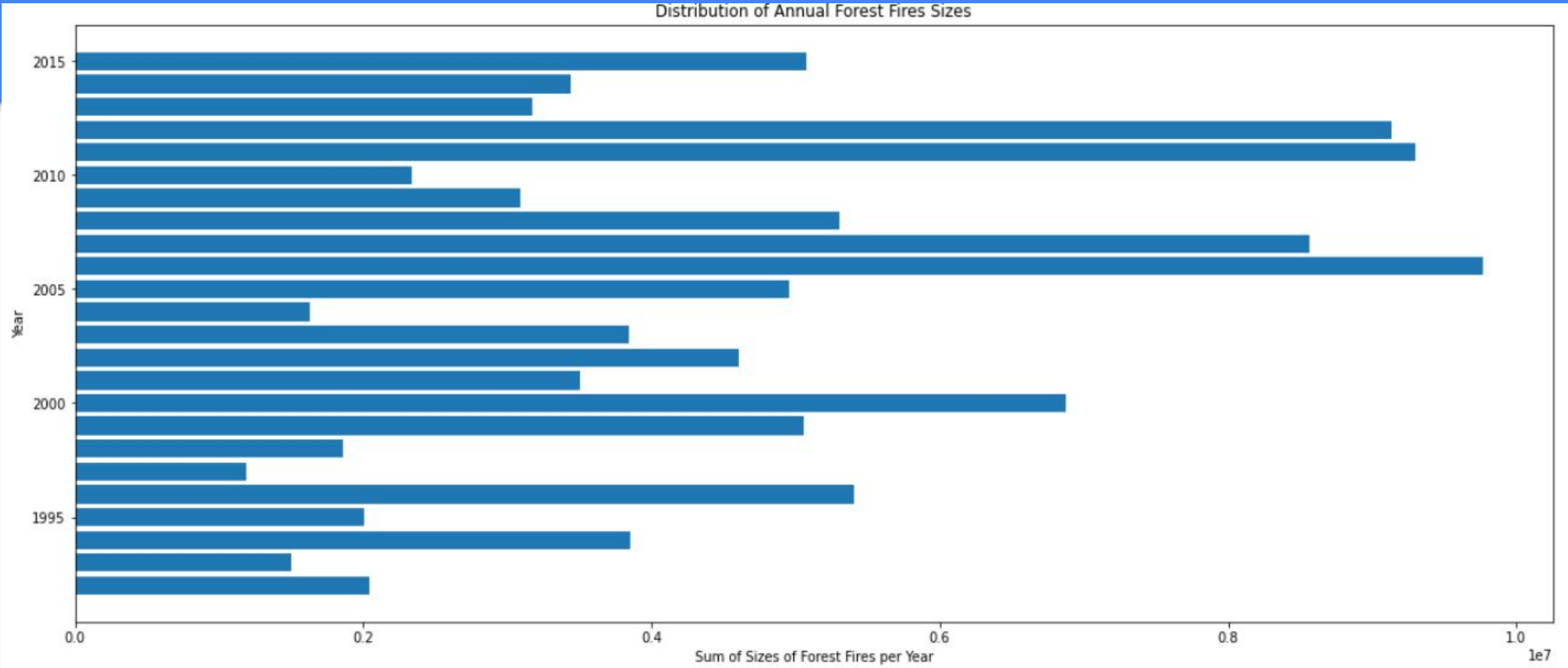
# 1.88 Million US Wildfires

- Public Kaggle Dataset
- Information about forest fires in US between 1993 and 2015
- SQLite format and over 700MB of data
- Data about Alaska, Hawaii and Puerto Rico was dropped in an attempt to make maps prettier and easier to make

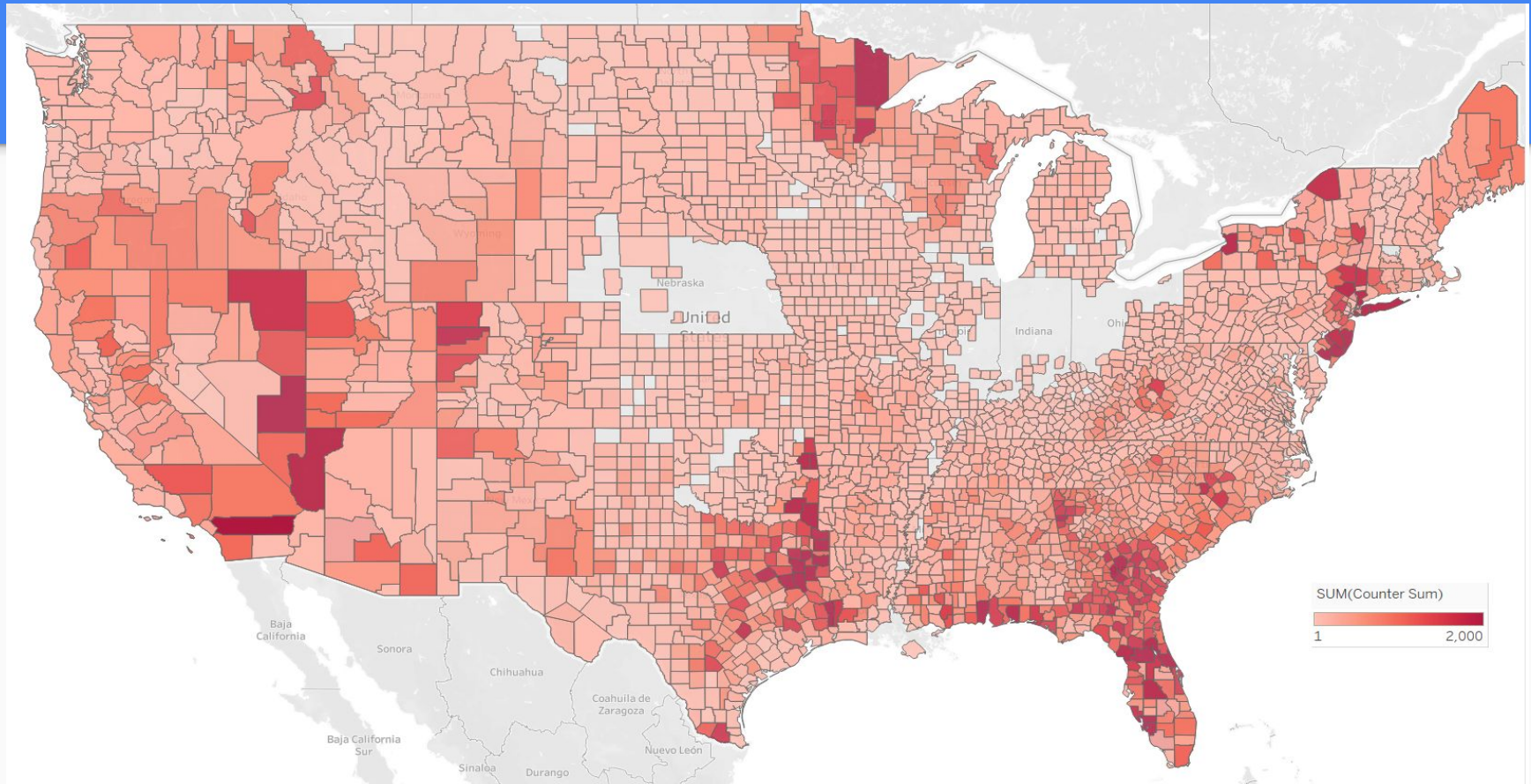
# Annual Forest Fires Numbers



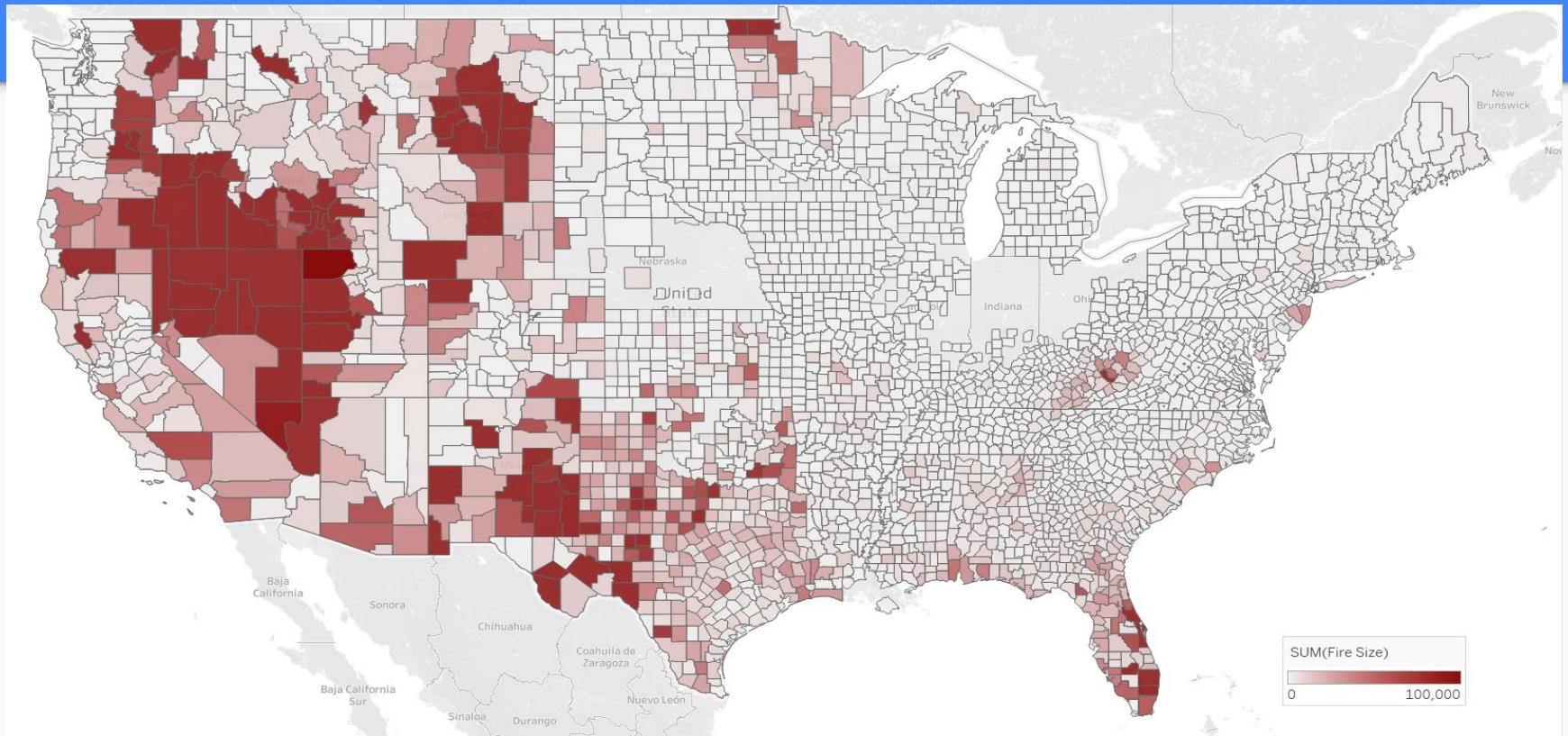
# Sum of Forest Fire Sizes per Year



# Counties Most Affected by Forest Fires (Number)



# Counties Most Affected by Forest Fires (Size)



# Modeling

- One thing that machine learning can predict is the cause of a forest fire based on available data.
- Due to extreme size of the dataset, complex models or gridsearch weren't completed.
- Baseline theory has 23% accuracy.
- Extra Trees multiple class classifier has 60% accuracy on testing data.
- Model is extremely overfit with 99% accuracy on training data.

# ~~Regrets~~ Conclusions

- Climate zone dataset was available on a private website with a worldwide set of coordinates.
- Function that takes the coordinates, finds closes match and adds climate zone to a record is simple, but extremely poorly optimized.
- Original dataset needs to be split to be processed in portions, but best way to split it is using climate zones which requires data to be processed first.
- Disregarding small fires is also a viable approach that I didn't think of at the time.
- Despite my original intention to use unsupervised learning I wasn't able to wrap my head around a way to actually apply this method to this data.

# Questions and Complains