McNulty MVP

My project will investigate US Visa applications from 2012-2016. The dataset was taken from Kaggle:

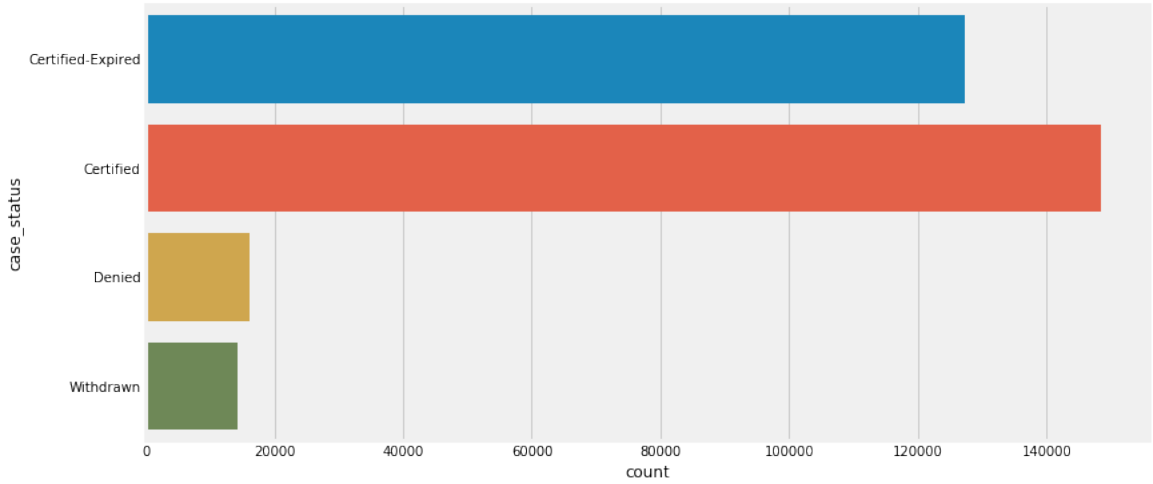
https://www.kaggle.com/jboysen/us-perm-visas

This dataset comprises 374,362 records with 154 columns. To narrow down the number of features, I took a count of the NaN values in each column and selected those columns with the fewest Nans. This resulted in 17 features:

na_counts.head(18)	# of NaN values
decision_date	0
case_status	0
employer_name	12
employer_city	14
employer_postal_code	37
employer_address_1	42
employer_state	42
job_info_work_city	102
job_info_work_state	103
pw_soc_code	397
pw_unit_of_pay_9089	1572
pw_source_name_9089	2099
pw_amount_9089	2216
pw_soc_title	2336
country_of_citizenship	20633
class_of_admission	22845
pw_level_9089	27627
---	
wage_offer_from_9089	114771

I did not select wage_offer_from_9089 or any of the features with more NaNs, because they comprised more than 30% values.

I then removed all rows with NaN values, reducing the dataset to 306,198 applications.



I have further cleaning to do (e.g. “NY” vs “New York” for employer_state), but I am close to beginning modeling. I also need to remove some of the features, such as employer address. When modeling, I will probably have to use one of the techniques for imbalanced outcomes, since very few applications have case_status = “Denied”.

A quick glance at the data leads me to believe that country of origin might not be the best predictor, and that wages and/or employer might be more likely to affect the case decision.

For visualization, I plan to use Tableau because of the map integration and will record a walkthrough of my dashboard.