Capstone Project1: Identifying counties that are the most fire-prone and predicting the cause of a fire wildfire.

DataSet: 1.88 Million US Wildfires (Kaggle1)

This data publication contains a spatial database of wildfires that occurred in the United States from 1992 to 2015. It is the third update of a publication originally generated to support the national Fire Program Analysis (FPA) system. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. The following core data elements were required for records to be included in this data publication: discovery date, final fire size, and a point location at least as precise as the Public Land Survey System (PLSS) section (1-square mile grid). The data were transformed to conform, when possible, to the data standards of the National Wildfire Coordinating Group (NWCG). Basic error-checking was performed, and redundant records were identified and removed, to the degree possible. The resulting product, referred to as the Fire Program Analysis fire-occurrence database (FPA FOD), includes 1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned during the 24 years.

Approach: In this notebook, I will do the below steps:

1. Fetch Data and create a Pandas Dataframe for further analysis
2. Data Wrangling to clean and transform the data

Step1: Fetch Data from Database

As the dataset was in SQL file format, I did some cleaning activity in SQLite. There were many columns in SQL table "fires” which were related to Source of data and agency information which collected or prepared reports on Data. I omitted these columns and imported only relevant data to the "US_Wildfire_data.csv" file. Also while importing I filled in all the null/blank values with "Nan" value.

Step2:

By looking at the info of the dataframe I observed many columns with more than 50% of its total data with null values. As those columns were not helping the data set I omitted those columns.

Step3:

As County_Name is the column which I am interested in, I checked if there are any Nan values in that column.

Step4:

Filling the missing values of COUNTY_NAME column using 'LATITUDE' and 'LONGITUDE' columns. To do this I assigned rows with "Nan" values in FIPS_NAME(County name) to new dataframe and merged values of 'LATITUDE' and 'LONGITUDE' columns into a new column. Also reIndexed DataFrame. Retrieved CountyName by Passing Latitude and Longitude Value as tuple to "reverse_geocoder.search" method as parameter. As returned county value had "County" as a suffix, truncated it from column values.

At first look I observed few of the county names are not matching with expected states in "noCountyNameDF" DataFrame.(Note: When latitude and longitude lies in the US state border this discrepancy of value is observed). To get rows with the mismatching State value, a new column is mapped with State name and State Code as dictionary with former being key.

Merge DataFrames nullCountyNameDF(with null County names) and newCountyNamesDF(with county names derived from geocoder package) column wise forming new dataframe CountyNameDF. .Dropped rows with the mismatching State value after comparing US Wildfire DataSet value with County data values obtained from reverse_geocoder package.

Merge dataframes US_fires(1202317 entries) and CountyNameDF(636208 entries).

Step 5:

column "DISCOVERY_DATE" was in data type float64 and not in Datetime format as expected. Hence I converted it to a readable date format from julian date.Dropped DISCOVERY_DATE. Remaned "date" column and FOD_ID column names to more appropriate names.Rearranged columns in dataframe.

Step6: Created a column FIPS_CODE with FIPS code value for each County Name.

Step7: Checked if there are any outliers with respect to FIRE_YEAR column and found there were no outliers which needed to be discarded. Cleaned data is written into new csv file “**US_Widfires_cleanData.csv**”

Below is the link to Github repository of jupyter notebook file with Data wrangling code.
https://github.com/lasyabheemendra/Sprigboard-DatascienceProjects/blob/master/Capstone1_US-Wildfire-Prediction/Capstone1%20-%20Data%20Wrangling.ipynb