# FINAL PROJECT REPORT

**NAME:** POOJAN VADALIYA (1281587)

**COURSE CODE :** DATA*6300: ANALYSIS OF BIG DATA

**TOPIC :** "DISASTER RELIEF FUND ALLOCATION ANALYSIS: IDENTIFYING EXCESSIVE GRANTS TO HOMEOWNERS"

## Table of Contents

# 1. Introduction And Background:

After hurricanes Katrina and Rita affected the state of Louisiana in 2005, the Road Home Program was established to aid Louisiana homeowners whose properties suffered extensive damage beyond what insurance and FEMA (Federal Emergency Management Association) assistance covered. This initiative provided various grants, such as compensation, elevation, and mitigation grants. The dataset acquired from the Louisiana Division of Administration contains detailed records of property owners who received these grants, including information on pre-storm property values, repair costs, insurance details, and total grant amounts disbursed.

Discrepancies in grant amounts have been observed due to factors like additional insurance payments or instances of fraud, affecting the actual funds owed by property owners to the state. Identifying properties that received more than the actual grant amount is essential for ensuring financial transparency, fair resource allocation, and enhancing program effectiveness.

# 2. Problem Statement:

Sometimes, it might be the case where homeowners might get extra money from their insurance or because of fraud, which meant they end up with more funds from the Road Home program than they should which became the need to identify to save the funds from improper allocations.

Here our task is to identify which homes/homeowners or might be the contractor in the state of Louisiana got more money than they were supposed to so we can make sure the money is used properly, shared fairly, and disaster recovery programs work better in a future scenario.

Also looking to the extensive-data it could by doing the exploratory data analysis of the acquired data we can try to identify the numbers which could have been saved at first if the frauds were detected before the disbursement of the grant leading to extensive help to other homeowners.

# 3. Methodology

## 3.1 Dataset Description:

Overall, the acquired Dataset consists of **130053 rows and 36 feature columns** describing the house demographics, bifurcation of accepted grant amounts along with the homes that have been identified as acquiring more funds than actually required.

Here, I was able to request the data from ProPublica DataStore [1] which acquired the data from Louisiana Division of Administration.

Features of dataset are as follows:

- **Structure Demographics:** It included the information and demographics of the structure of the house and its location which includes the city name state and the zip code along with the parish name in which its located. Also, NOLA Information was included which was looking forward for the redevelopment of the neighborhood based on the boundaries that were kept in place in the year of 2000.
  Here are some key features:
    - **Structure type**
    - **GIS state**
    - **GIS city**
    - **GIS zip**
    - **Parish**
    - **NOLA Planning Description:** It included District Number, District Name, Neighborhood Number, Neighborhood Name

- **Grant Statistics:** It includes the amount and the type of grant that has been disbursed to the home owners which includes the following grants.
    - **Total CG Amount :** Compensation grants – Max was $150,000.
    - **Total ACG Amount :** Additional Compensation grant – For the ones who made less than 80% of Area Median Income.
    - **Total Elevation Amount** : Grant up to $30,000 for raising their homes to prevent future flood damages.
    - **Total IMM Amount :** For mitigation measures other than compensation grants – Fixed amount of $7500
    - **Total Closing Amount :** Total grant allocated to the homeowner from the program.

- **Census Demographics:** It displays the block id's of the home as per the development formed during the year of 2000. It includes:
    - **Census Blocks**
    - **Block Groups**
    - **Census Tracks**

- **Closing Option:** Closing options states about how the houses and closing options were selected and whether or not to check if it would be better to reconstruct the complete house or to refurnish it. It included the following features:
    - **Closing Damage Assessment**
    - **PSV at Closing** – Value of the house before the storm.
    - **Closed File - Option 1 –** It stated to repair or rebuild the property.
    - **Closed File - Option 2/3 –** It stated to sell the property and purchase another one in Louisiana if Option 2 or in another state if Option 3.

  - **Closed with Approved Unmet Needs .**

- **Damage Assessment :** Damage assessment specifically defined about the total impact that the hurricane has made to the property to rebuild the house as it was before. It included :
  - **Current Damage Assessment :** Value after damage
  - Current Damage Assessment-Type1 – Estimated cost to repair the home.
  - **Current Damage Assessment-Type 2 –** Estimated cost to completely rebuild the home.
  - **Damage Type 1 or 2**
  - **Current PSV –** The most recent estimate of the property's pre storm value.

- **Target Feature :**
  - **ARS File (Yes/NO) –** It would say Yes if more grant was disbursed to the property owner and No if it was not.

## 3.2 Dataset Preprocessing:

- **Handling Missing Values:**
  - After checking the number of missing value it was found that the major part of the NOLA plan data was missing that accounted of almost 75% of NOLA data which lead to the decision of the removing the columns instantly.
  - As Current Total DOB Amount also had 6 values missing which was simply imputed by the mean of the all the other values of the data column.

- **Removing Unnecessary Feature Columns:**
  - Looking at the features in the dataset it was found out that feature columns like GIS zip, Census Demographics were not contributing well for the model learning purpose which led to the instant elimination of the columns as we have all the other data required to figure out the information about the property and the funds.

- **Minimizing the feature numbers:** It can be found that there are many features which can be removed and converted into single feature without diminishing the information like "Current Damage Assessment -Type 1" and "Current Damage Assessment Type –2", whose information can be extracted directly from the "Damage Type 1 or 2" feature column which is all we need.

- **Encoding Categorical Columns:**
  - After the minimization of the number of features the categorical column were identified from the dataset and were encoded with the following three types of encoding:
    - **One-Hot Encoding:** For the non-ordinal features having a smaller number of unique categories.

  - **Binary Encoding:** For the features having only two categories and the target column as well.
  - **Frequency Encoding:** Only for “GIS-city” feature as it was non ordinal in nature and also number of categories were more than 300 which would not be a good choice if one hot encoding is applied.

- **Handling Numerical Columns:**
  - Normalizing the data was one factor as the grant range was quite varied overall for different houses.
  - Also, the columns of having a fixed disbursement grants could be turned into a 0 or 1 value like IMM compensation where $7500 is granted which could be simply turned into 1 if it was given or 0 if not.

- **Handling the Class Imbalance:**
  - As of when started to model there was heavy class imbalance to be observed in the target column where the “No” had almost 121128 samples and “Yes” category had 8925 sample count.
  - For handling the sample count, sample was balanced by using “RandomOverSampler” for the under-sampled data and “RandomUnderSampler” for the oversampled data where after trial and error it was found that the 17850 was the optimum number of sample counts which helped in models to learn well.

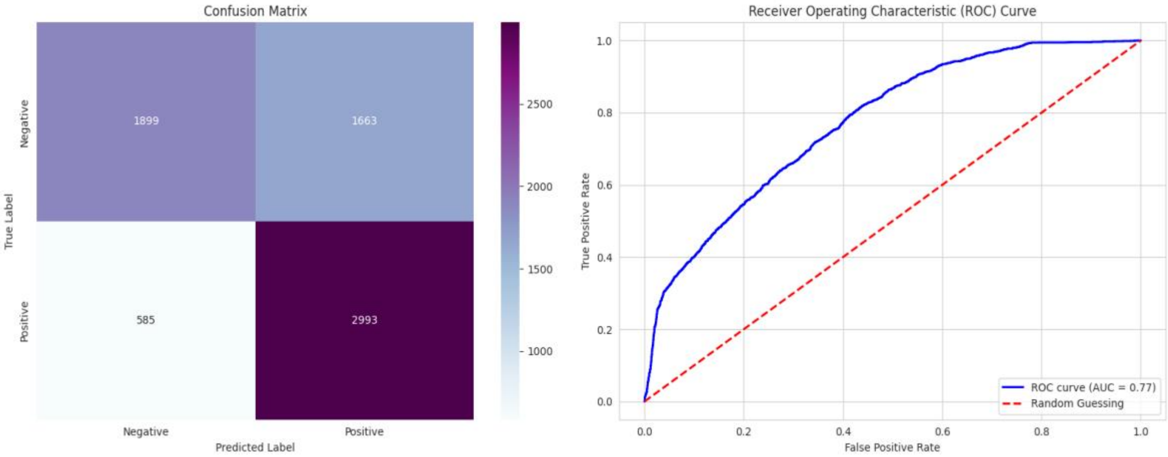
## 3.3 Modelling:

For modeling purpose three classification of models have been selected going from the simplest to the complex model to observe how the complexity of model can handle the preprocessed data. And also, after finding the best model Fine tuning was performed to enhance the model results even further.

1. **Logistic Regression:**

   Being a simple model and preferable for binary classification this model was preferred as it helps to provide significance of each feature and also when the dataset is not highly correlated or large.
   Here are the results after running the model:

   Logistic Regression gave **Test Accuracy of 68.52%** and here is the confusion matrix and the ROC Curve describing the model’s performance.

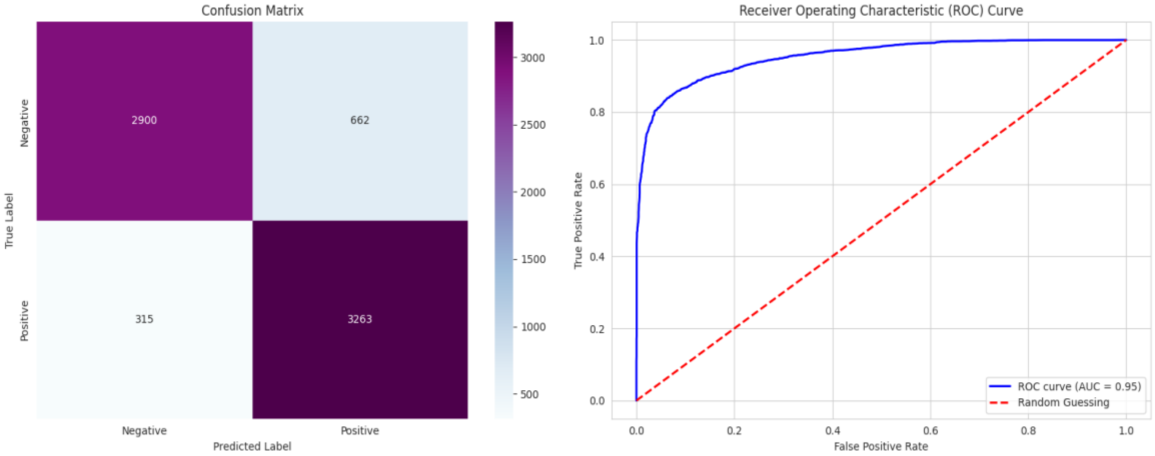


*1: Logistic Regression Classification Results*

2. **Random Forest:**

   Model which can help handle linear and non-linear classification problem led to the use of Random Forest where it is helping to compare multiple decision trees to find the best one. It did perform the best out of all three models showing its ability to classify the data and even handle the outliers well.

   Random Forest gave **Test Accuracy of 86.32%** and here is the confusion matrix and the ROC curve showing the model's performance.

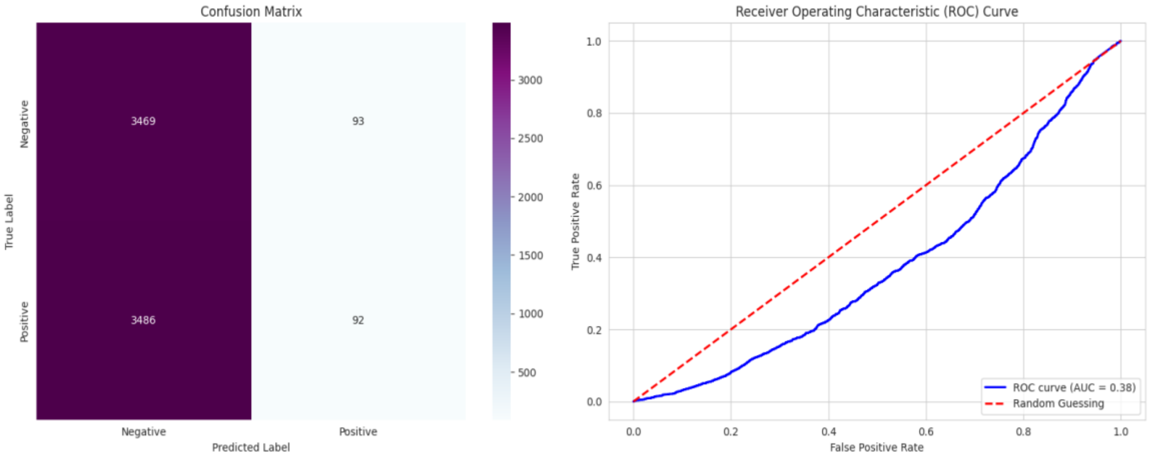


*2: Random Forest Classification Results*

3. **Support Vector Classifier:**

   Looking at the SVM's ability to classify the non-linear data well which has a clear margin for seperation classes led me to use this model but it seemed the preprocessed data did not work well in our case. I personally feel that it might be due to the highly correlated data or the margin of seperation can't be identified by the model.

SVM gave test accuracy of 49.87% and here is the confusion matrix and the ROC curve showing the model's performance.



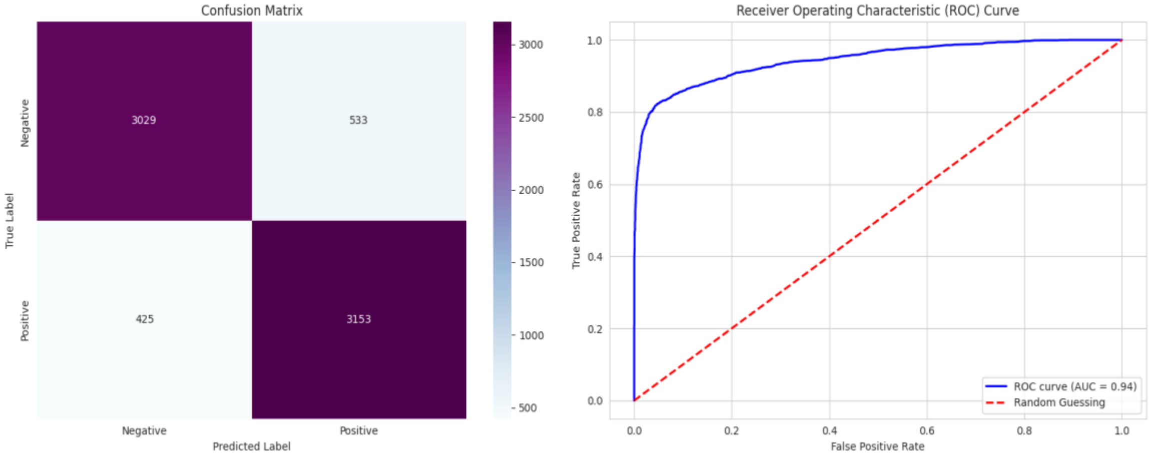
*3 : Support Vector Classifier Classificatioin Results*

## 4. After Fine Tuning:

Random forest performed the best overall so here I selected to use forward subset selection to identifying the best performing features and train the data on the subset of features and bagging to select the best subset of data.

### 4.2 Forward Feature subset selection:

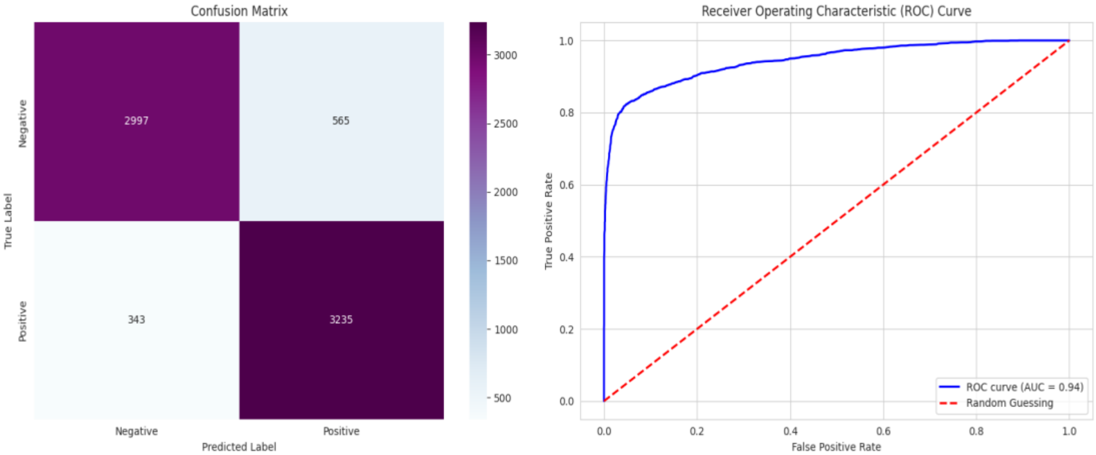
After selecting the features subset with 17 features performed the best with 86.58% of test accuracy. Here's the classification results of the Random Forest trained on the feature selected model.



*4. Feature Subset Selection Classification Result*

### 4.3 Bagging:

Used bagging to select the most optimal subset of data for training purpose which as a matter of fact showed the best results of all the other models after finetuning with a test accuracy of 87.28%. Here's the classification results of the Random Forest Model performed along with bagging.



*5: Random Forest with Bagging Classification results*

# 4. Results:

## 4.1 Model Results:

| Model Name | | Precision | Recall | F1-Score | Test Accuracy |
|---|---|---|---|---|---|
| **Logistic Regression** | 0 | 0.76 | 0.53 | 0.63 | 68.51% |
| | 1 | 0.64 | 0.84 | 0.73 | |
| **Random Forest** | 0 | 0.90 | 0.91 | 0.87 | 86.31% |
| | 1 | 0.83 | 0.91 | 0.87 | |
| **Support Vector Machines** | 0 | 0.50 | 0.97 | 0.66 | 49.92% |
| | 1 | 0.50 | 0.03 | 0.05 | |
| **After Fine Tuning** | | | | | |
| **Forward Subset Selection** | 0 | 0.88 | 0.85 | 0.86 | 86.58% |
| | 1 | 0.86 | 0.88 | 0.87 | |
| **Bagging** | 0 | 0.90 | 0.84 | 0.87 | 87.28% |
| | 1 | 0.85 | 0.90 | 0.88 | |

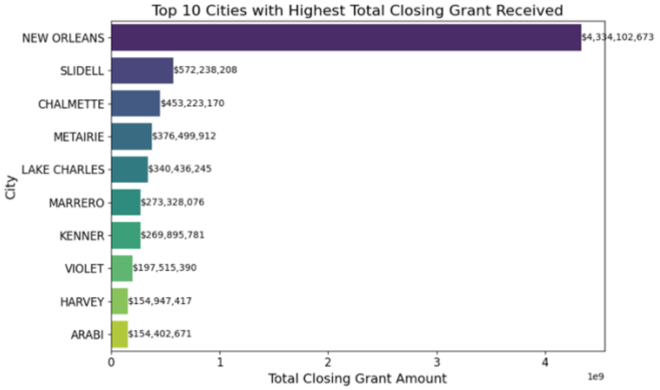
Looking at the score of each model in the above results we can see that without fine tuning random forest performed the best with an accuracy of 86.31% and adding more to it after fine-tuning bagging turned out to help to better the score to 87.28% to classify accurately out of all.

## 4.2 Exploratory Data Analysis:

Data was analyzed in terms of members who were found to have taken more grant than required as our primary focus is on those number.

### Top 10 cities with most Grant:

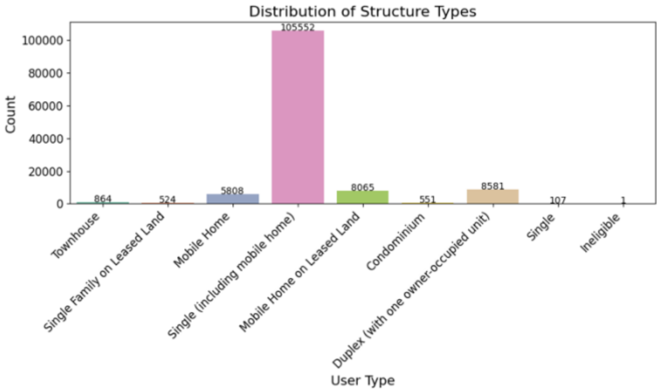
New Orleans turned out to be granted the maximum funds with over 4.3 million dollars which also shows that the hurricane impacted there the most.



*4 1: Most granted cities*

### Bifurcation based on Structure type:

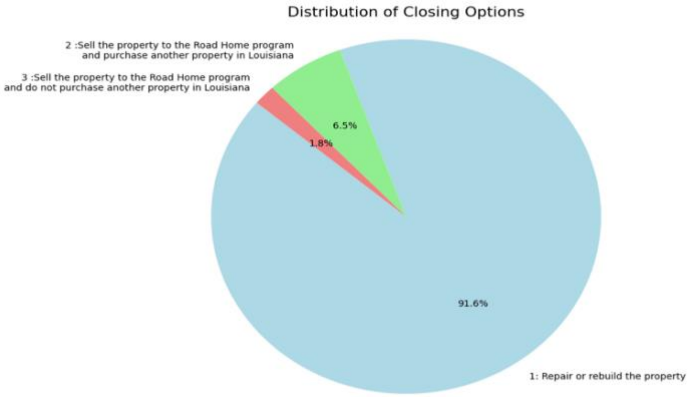
It was found that the maximum data included single mobile homes  with almost around 10500 house data.



*4 2: Structure based sample distribution*

### Options selected among grant Owners:

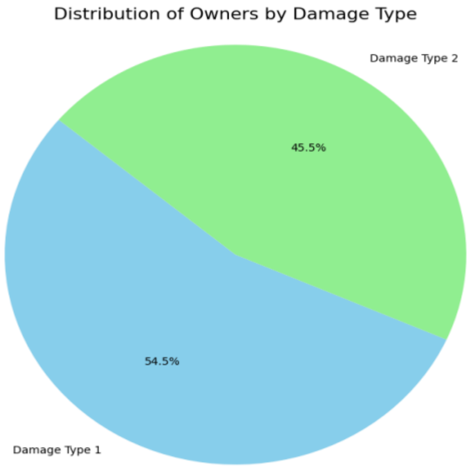
It seems that maximum people opted to repair and rebuild their property rather than selling it. Might be the reason to access the grant funds.



*4 3:  Selected Closing Option*

## Consideration of Damage type of houses:

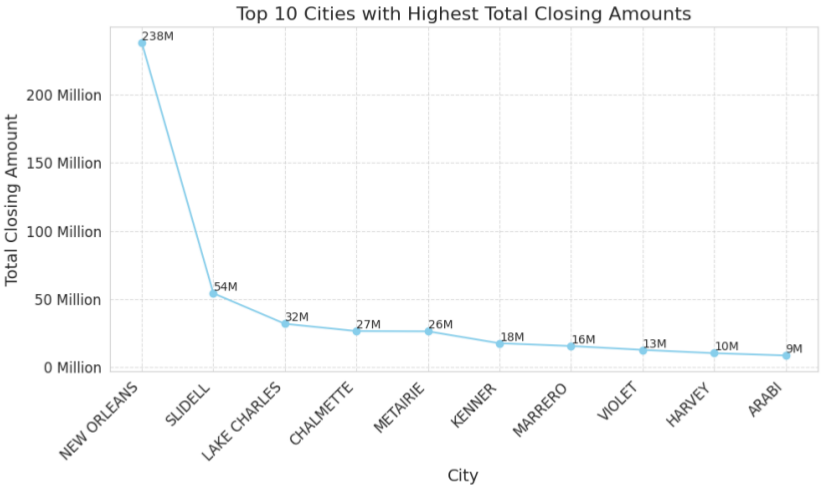
Almost 45.5% houses were classified to have been damaged above 50% which required a complete rebuild of the houses.



*4 4 Damage Type Consideration*

## Total Money that could have been saved:

It can be found that almost New Orleans had the highest amount of money transferred which were classified as fraudulent with about 238M followed by Slidell 54M and Lake Charles 32M.
As a total if these money had been found before then it might be the case that total of 542M dollars could have been saved from the program funds and could have reallocated elsewhere.



*4 5 Fraudulent money allocation*

## 5. Conclusion:

Looking at the nature of data there was a heavy class imbalance seen in the data which was hindering for the models to learn. As of modeling Random Forest performed the best overall and bagging helped to further the model accuracy with the final test accuracy of 87.2%.

Looking at the EDA New Orleans seemed to be the city with maximum grant allocation with around 4.3 Billion USD granted whereas. And a total of 542 Million USD was found to be allocated as fraudulent which could have been saved if detected before and could have been used for any other purpose.

## 6. References:

**[1] Source of the dataset:** https://www.propublica.org/datastore/dataset/road-home-rebuilding-grants

**[2] Source used for understanding the data and problem statement in depth:** https://www.google.com/url?client=internal-element-cse&cx=59278435ae7a1c194&q=https://www.doa.la.gov/media/qphoxrad/kr3_drgr_1stq2019.pdf&sa=U&ved=2ahUKEwiA0qGt09mEAxWNElkFHSuHBwsQFnoECAEQAg&usg=AOvVaw1gYzTucHUuaF-VXFkU7mfF