Motivation:

The Vietnam War is a pivotal and controversial period of American history. Political and policy issues aside, the Vietnam War brought upon a major change in military tactics, especially in terms of air doctrine. The lessons the United States Air Force learned during the Vietnam War spurred the development of precision-guided munitions, and saw the final use of strategic bombing as a large-scale tactic. Finally, given the prevailing statistics/analytics-based approach Secretary of Defense Robert McNamara, it only seems poetic that further analysis of the air war be conducted.

The four main questions that I sought to answer were:

1. Which country flew the most missions, and which branch flew the most missions for that country?
2. Which year saw the heaviest bombing, by tonnage dropped?
3. Given some sortie information, can we predict the country that flew the sortie?
4. Using mapping software, how did bombing locations and the countries that flew them change over the years?

Data Source:

The data set was taken from Kaggle; it documents the over 4 million sorties flown by the nations involved in the Vietnam War. The link is here. The data is in a CSV format, with three distinct tables. One table consists of all the aircraft used, with some accompanying information. The next is a table of all the ordnance types used, with accompanying information. Finally, the main table contains all the sorties flown. This table contains 47 distinct columns. The main variables are the latitude and longitude in float format, mission date as a string, country of origin, weapon weight in integer, and weapons delivered in integer form. In total, there are over 5 million total rows; the sortie table itself contains around 4.8 million rows.

The time covered starts in 1965, and goes to 1975, over a span of 10 years.

Methods:

1. Which country flew the most missions, and which branch flew the most missions for that country?
    a. Data manipulation
        i. There was no initial cleaning of the data.
    b. Missing/incomplete/noisy data
        i. Any data issues would automatically get dealt with by the data analysis process.
    c. Data analysis
        i. To figure out which country flew the most combat missions, I first extract only the COUNTRYFLYINGMISSION column, and then did value counts. I separated the country and the count into two variables, then plotted using a seaborn bar plot. To annotate graph, which seaborn has no intuitive in-built method, I used a method taken from Stack Overflow.

ii. To figure out which branch flew the most missions for the country that flew the most missions, I selected for all rows that had the United States as the country flying mission, and then took only the MILSERVICE column. Using the above method for plotting, I did the same for the branch that flew the most missions.

d. Challenge/solutions
    i. One of the biggest challenges was figuring out how to annotate the bar plot. However, a cursory google turned up an easy solution.

2. Which year saw the heaviest bombing, by tonnage dropped?
    a. Data manipulation
        i. I first had to filter down the dataset, and selected only sorties that actually delivered ordnance (NUMWEAPONSDELIVERED > 0).
        ii. Next, I had to extract the year from the MSNDATE column. There were two primary formats for date: YYYY-MM-DD, or YYYYMMDD. A simple regex found the first format, and the second was done with string casting and indexing. Any other formats that may have slipped in were ignored, and no year was assigned to those sorties. This was assigned to a new column, year.
        iii. Finally, I calculated the sortie tonnage for each sortie I had. This was done by applying a function to each row, where the number of weapons delivered was multiplied by the weapon weight, and then divided by 2000 (2000 lbs in a US ton). This was assigned to a new column, sortie_tonnage.
    b. Missing/incomplete/noisy data
        i. Any missing data was dropped through the course of analysis. My initial selection of sorties that delivered ordnance filtered through many of the recon/support flights that would be considered noise.
    c. Data analysis
        i. I grouped by year, then did a sum, and only took the sortie_tonnage column. The index needed to be reset, so that graphing could occur. Then, like above, I plotted using a seaborn bar plot, and annotated the graph.
    d. Challenge/solutions
        i. The hardest part was figuring out that even the US government, which provided this dataset, has issues with standardizing date format. This was easily dealt with; try/except blocks, with if/else statements nested inside. Anything format that was not common enough to show up was duly ignored. While this may have caused issues with 100% accurate sortie_tonnage calculations, the number of rows that ended up being dropped as a result is negligible.

3. Given some sortie information, can we predict the country that flew the sortie?
    a. Data manipulation
        i. First, I converted the country flying mission into numerical values, using a numpy select. There was a total of 5 countries.
        ii. Next, I dropped a significant number of columns. If the column had a significant amount of unique categorical values (in the order of 1 or 2 million), then I deemed it would be irrelevant for my classifiers. As well, any rows where year was null were dropped as well. Finally, any row that had less

than 23 populated columns got dropped, to decrease the size of the dataset and facilitate the generation of dummy variables.
iii. A random selection of 50,000 rows was selected as a subset, using code taken from here.
iv. Any columns that had null values were manually filled with placeholder values.
v. Dummies were taken, and the dataset was split into a 70/30 train/test split.

b. Missing/incomplete/noisy data
i. See above.

c. Data analysis
i. I created two models: a random forest classifier and a logistic regression model. They followed the standard pipeline of all machine learning models. First, the data was trained on all columns of the training set, except for the type column. Predictions were made on the test set, with the same type restriction. This prediction was added to the scoring set (a copy of the test set), and accuracy was assessed by using sklearn. If the model allowed for feature importance to be extracted, it was done so, using code we had previously covered in class.

d. Challenge/solutions
i. The main challenge was first getting the dataset down to a size so that dummies could be made, without exceeding machine memory limits. It quickly became obvious using all 4.8 million rows would quickly exceed even the 20 GB of memory I had locally. Thus, I cut down as many columns as I could, and cutting down the number of rows used. The best method for this was to use a random selection, and seed the random generator to make sure I got the same results each time. This way, there was no user bias in selection the 50,000 rows.

4. Using mapping software, how did bombing locations and the countries that flew them change over the years?
a. Data manipulation
i. The first step was to select on the country flying mission, latitude, longitude, and year columns. Then, any rows with latitude, country, or year being null were dropped. Latitude and longitude were rounded to 2 decimals, and any duplicated values were dropped, to make the mapping process run quicker. Each country was associated with a custom color, and that color was assigned to a new column.

b. Missing/incomplete/noisy data
i. See above for missing value removal.

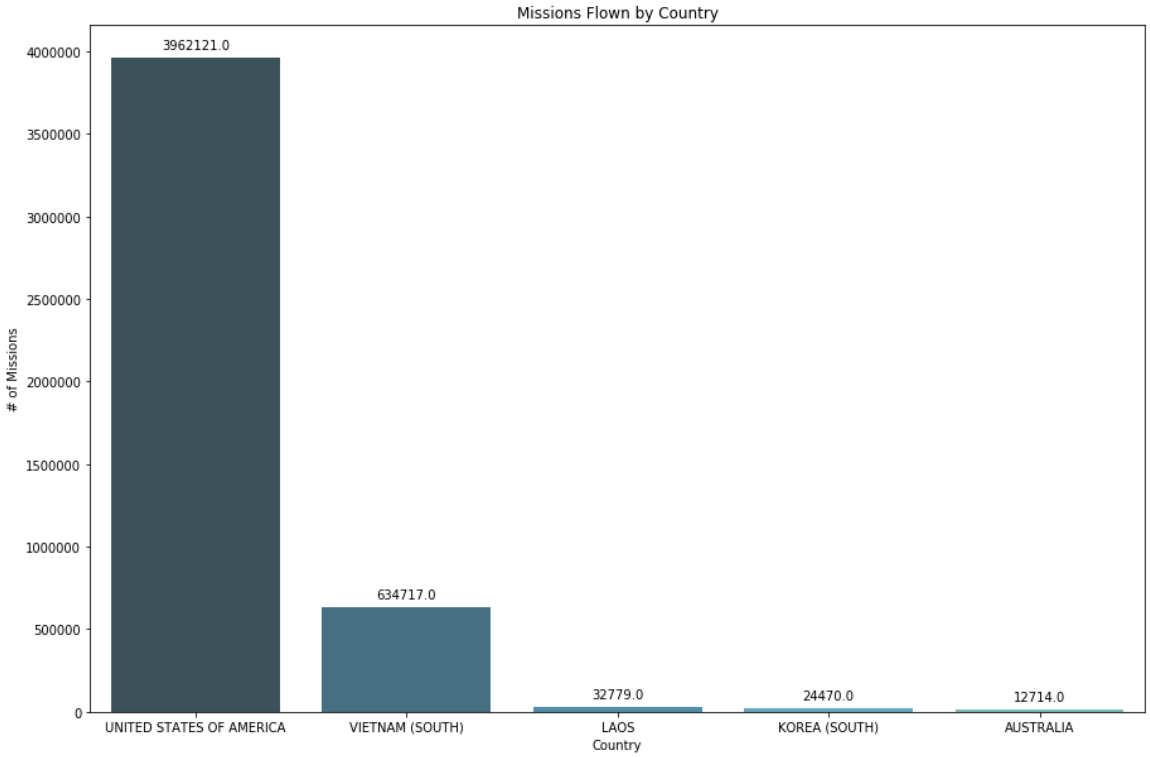
c. Data analysis
i. A custom function was defined, to plot for a specific year. Latitude, longitude, and country color was extracted for all rows that matched the year. A Basemap figure was created, and appropriate map-related functions were called, to make it look prettier. Finally, the latitude and longitudes were plotted on the map.

d. Challenge/solutions

i. Getting the map to center correctly, and work properly was the most challenging. Originally, I had planned to use Cartopy, since it was the currently supported library for Python/Matplotlib mapping. However, I had significant issues with getting a map with passable resolution. Thus, I had to resort back to Basemap, which is currently being phased out in favor of Cartopy.

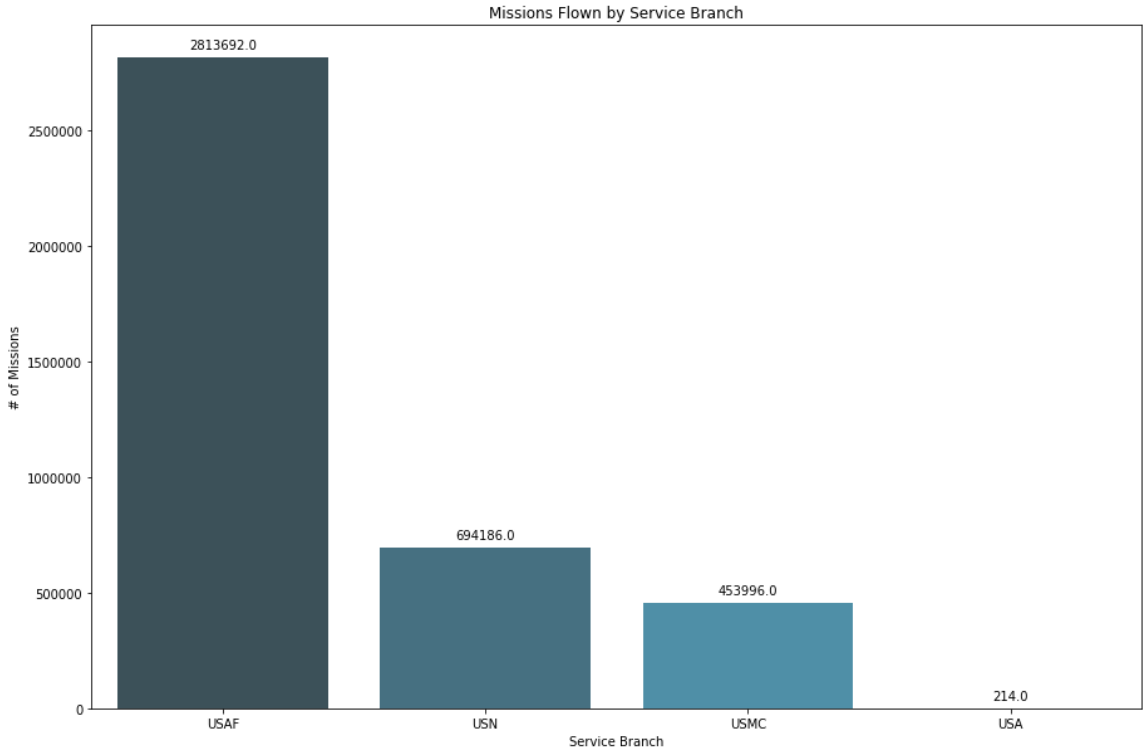
Analysis and Results:

1. Which country flew the most missions, and which branch flew the most missions for that country?
    a. As to be expected, the United States flew the most combat missions during the Vietnam War. The forces of South Vietnam also conducted a not-insignificant portion of the sorties as well. Given the amount of resources and training the United States invested in South Vietnam, this is not suprising. The involvement of South Korea, however, is quite interesting, as the United States had been involved with training ROK forces during the 50s. Australia flying sorties is not surprising, given the country's proximity to the Vietnam region. As well, Laos was a nominal ally of the United States, and did fly sorties in support of the conflict.



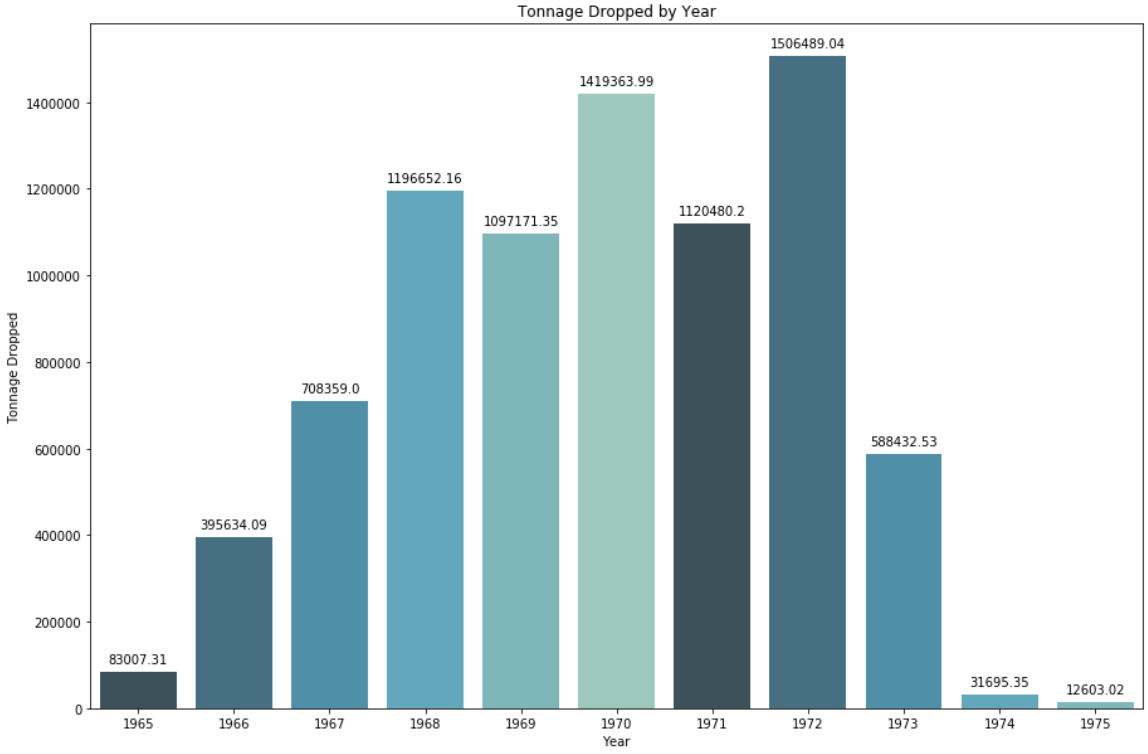
    b. It is obvious that the Air Force dropped the most ordnance during the Vietnam War, given the intensity of the campaigns carried out, such as Rolling Thunder or Arc Light. As well, the Air Force flew many interdiction missions in North Vietnam and Cambodia/Laos, the cut the infamous Ho Chi Minh Trail. The Navy and the Marine Corps also dropped a significant tonnage, but nothing comparable to the Air Force. These missions were primarily for close-air support. Interestingly enough, the Army had some tonnage dropped. This, however, cannot be accounted by rotary-wing

sorties. After further research, it appears this tonnage is accounted for by liaison and observation aircraft, which in some cases also had ordnance.



2. Which year saw the heaviest bombing, by tonnage dropped?
    a. The year 1972 saw the heaviest bombing (1,506,489.04 tons). In fact, the 82 missions flown in the first three months of 1972 exceeded the tonnage of all missions dropped in 1971. There was a noticeable drop-off after 1972 and 1973, because of the US administration beginning to negotiate a cease-fire. This eventually became the 1973 Paris Peace Accords, which ceased military activities in Vietnam, and US troops

began to withdraw.



3. Given some sortie information, can we predict the country that flew the sortie?
    a. Using the random forest classifier that performed the best when we classified malicious/benign websites, it was able to achieve an accuracy score of 99.91% when predicting the country of origin in the test set. The most important features seemed to stem from what branch the mission flew in. The Vietnamese National Air Force and the US Air Force seemed important, when predicting the country of origin. Given

their close relation to the country of origin, this makes logical sense.

Random Forest Classifier ¶

```
df_rf = skens.RandomForestClassifier(n_estimators = 10, oob_score = True, criterion = 'entropy', random_state = 42)
df_rf.fit(bombing_train.iloc[:,bombing_train.columns != 'type'], bombing_train.type)
```

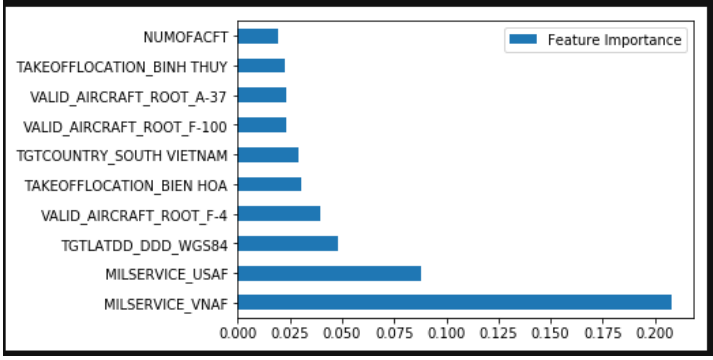
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
            oob_score=True, random_state=42, verbose=0, warm_start=False)
```

```
pred_rf = df_rf.predict(bombing_test.iloc[:,bombing_test.columns != 'type'])
bombing_score['pred_rf'] = pred_rf
skmetric.accuracy_score(bombing_score.type, bombing_score.pred_rf)
```

```
0.9990666666666667
```

```
feat_importance = df_rf.feature_importances_
feat = pd.DataFrame({'Feature Importance': feat_importance},
            index = bombing_train.iloc[:,bombing_train.columns != 'type'].columns)
feat.sort_values(by = 'Feature Importance', ascending = False).head(10).plot(kind = 'barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xb358066630>
```



b. Using a baseline logistic regression, it was able to achieve an accuracy of 91.29%.

Logistic Regression

```
df_lr = sklm.LogisticRegression()
df_lr.fit(bombing_train.iloc[:,bombing_train.columns != 'type'], bombing_train.type)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

```
pred_lr = df_lr.predict(bombing_test.iloc[:,bombing_test.columns != 'type'])
bombing_score['pred_lr'] = pred_lr
skmetric.accuracy_score(bombing_score.type, bombing_score.pred_lr)
```

```
0.9128666666666667
```

4. Using mapping software, how did bombing locations and the countries that flew them change over the years?
    a. 1965 – To see the full gif, refer to vietnam.gif in the submission file. By looking at the gif, it is clear that the US flew a majority of the sorties during the early years of the war. However, starting 1970, South Vietnam began to fly more sorties in country,

in support of the ground operations. From 1971 to 1973, South Korea, Australia, and Laos flew a small amount of sorties, probably to cut off the Ho Chi Minh Trail, and deny the Vietcong on the North Vietnamese Army their logistics.