Modeling the Number of Bombs Dropped During WW2

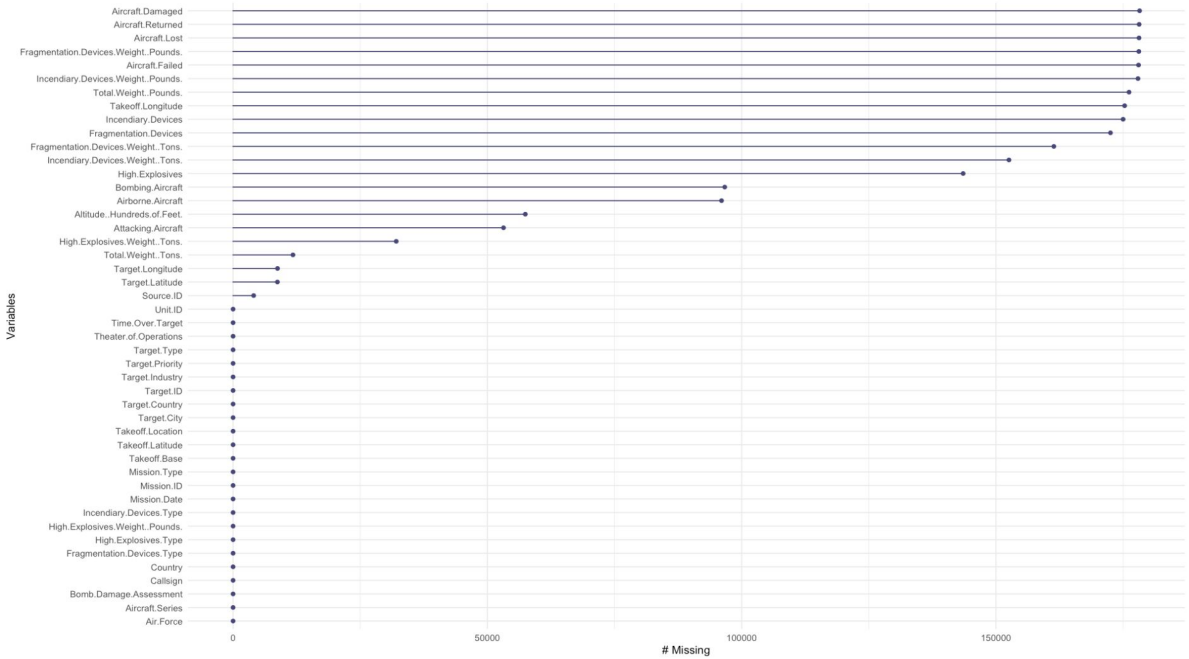
Using Poisson and
Negative Binomial Regression

Dataset Overview

- Was sourced from Kaggle.com, but was originally put together by Lt. Col Jenns Robertson of the US Air Force.
- Contained a total of approximately 175,000 observations were each observation is a single bombing mission between May 15, 1940 and May 2, 1945.
- There are 46 variables in the dataset, 22 of which are categorical and the remaining numerical.

Data Cleaning

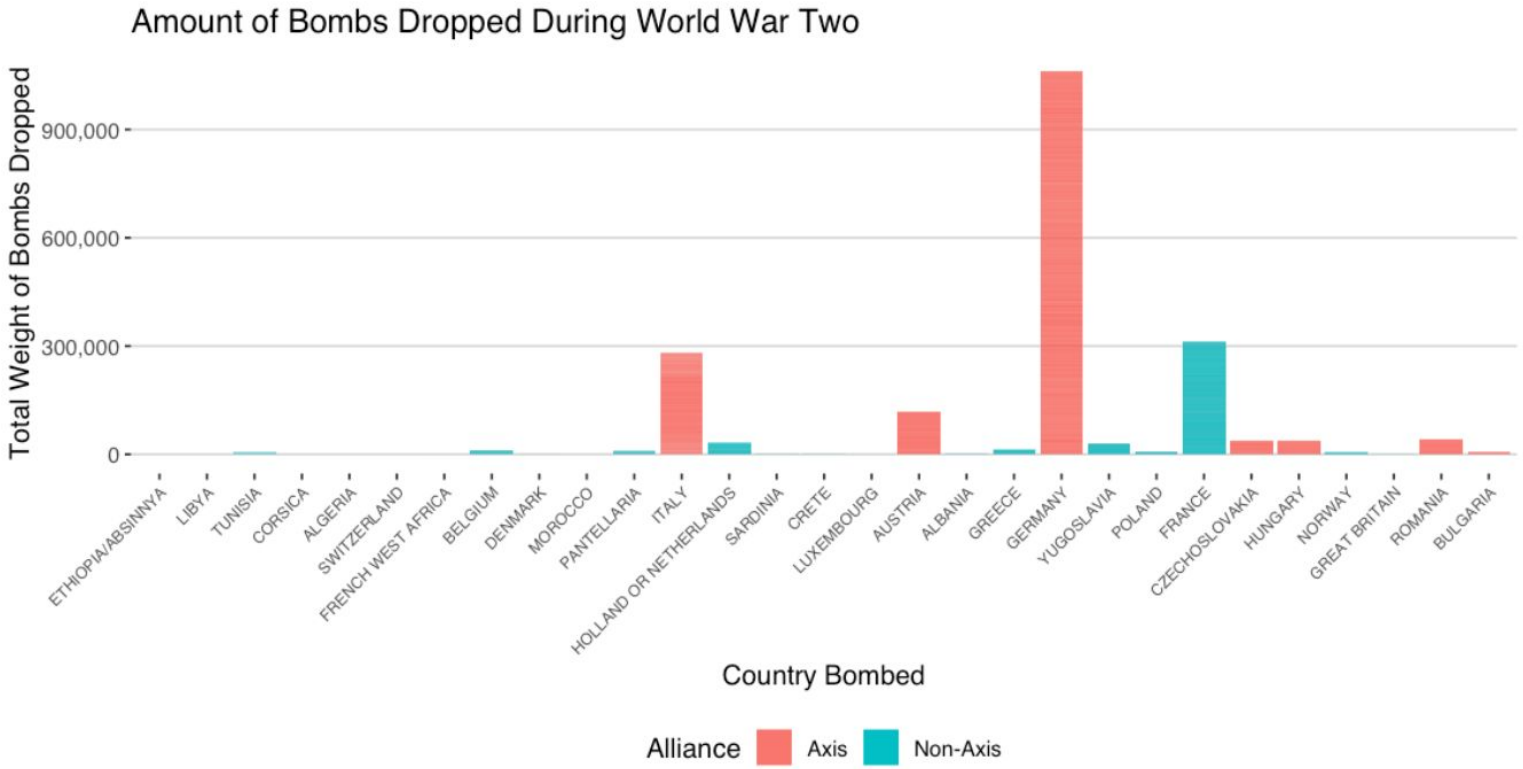
- Due to the nature of the data set there was a lot of wrangling needed.
    - Ultimately imputation was done where possible but a lot of the variables had too many missing values



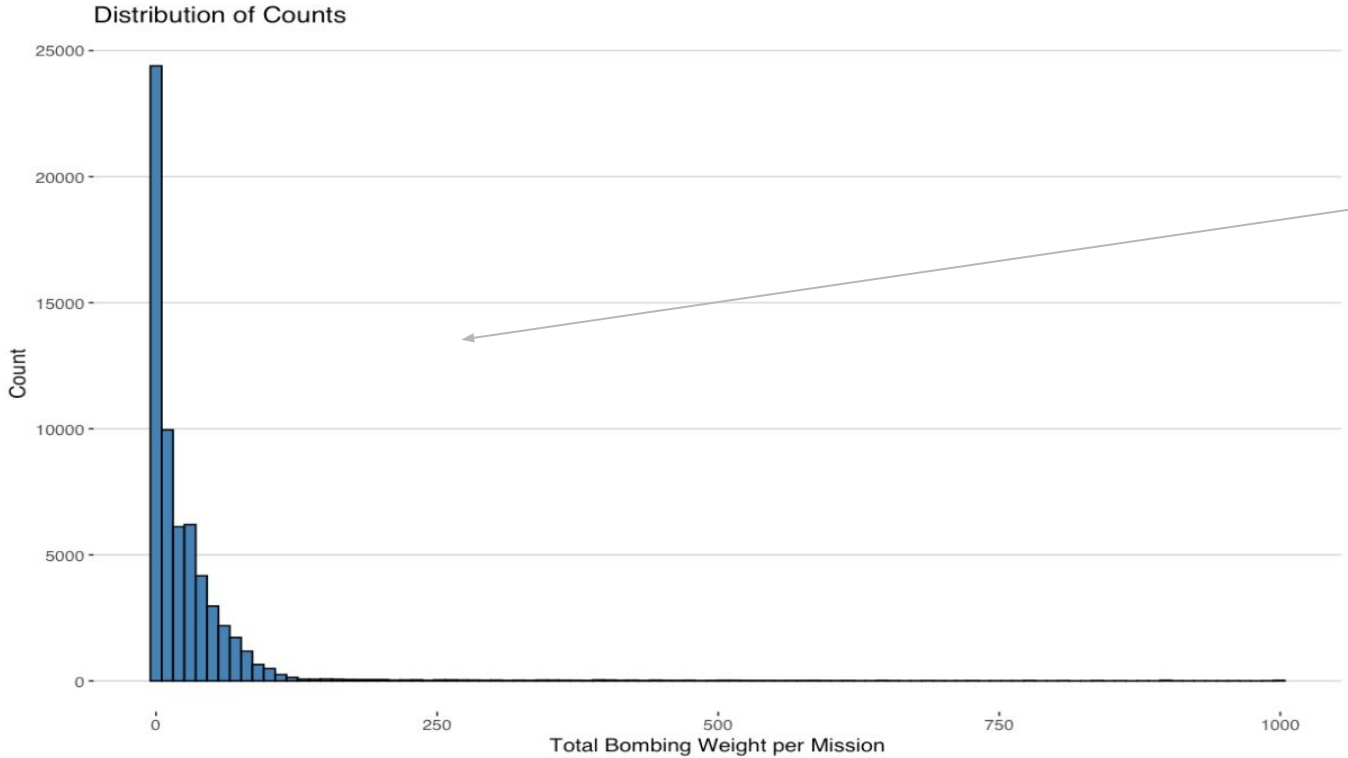
Data Cleaning and Exploration

Mission.ID	Date	Theater	Target_Country	Country	Total_Weight	Altitude	Attacking_Aircraft	Bombing_Aircraft	System_Time	Alliance	Industry_Type
12	1943-08-15	ETO	GERMANY	GREAT BRITAIN	1	250	11.462295	11.462295	-832550400	Axis	City
13	1943-08-15	ETO	GERMANY	GREAT BRITAIN	4	250	5.000000	5.000000	-832550400	Axis	City
58	1943-08-15	ETO	GERMANY	GREAT BRITAIN	87	135	11.462295	11.462295	-832550400	Axis	City
66	1943-08-15	MTO	ITALY	USA	2	95	16.857143	16.857143	-832550400	Axis	Non-City
67	1943-08-15	MTO	ITALY	USA	2	95	16.857143	16.857143	-832550400	Axis	Non-City
68	1943-08-15	MTO	ITALY	USA	17	95	6.733871	6.733871	-832550400	Axis	Non-City

Data Cleaning and Exploration



Data Cleaning and Exploration



Looks rather Poisson like...

Initial Model Fitting

```
Call:
glm(formula = Total_Weight ~ Attacking_Aircraft, family = poisson(link = "log"),
    data = ww2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-188.570    -6.048    -3.966     0.352    73.758

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.1734992  0.0007892    4021   <2e-16 ***
Attacking_Aircraft 0.0200166  0.0000166    1206   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4314893  on 62341  degrees of freedom
Residual deviance: 3803152  on 62340  degrees of freedom
AIC: 4062693

Number of Fisher Scoring iterations: 6
```

- Started out with a simple Poisson model but that went bad fast
  - Even with adding new variables and several types of transformations it was clear this was not going to work

Oh well onto the Negative Binomial

- After some transformations the over dispersion issue was drastically reduced
- Additionally it can now be seen what variables are significant or not.

```
Call:
glm.nb(formula = Total_Weight + 1 ~ log(Attacking_Aircraft) +
    log(Bombing_Aircraft) + Altitude + System_Time + Theater +
    Alliance + Industry_Type, data = ww2, init.theta = 1.217288066,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9157  -0.8016  -0.3203   0.0981  16.4306

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.195e+01  8.183e-01  14.598   <2e-16 ***
log(Attacking_Aircraft)  1.197e+00  1.499e+00   0.799    0.424
log(Bombing_Aircraft)   -3.538e-01  1.499e+00  -0.236    0.813
Altitude                 6.905e-04  7.121e-05   9.697   <2e-16 ***
System_Time              1.219e-08  1.005e-10 121.287   <2e-16 ***
TheaterETO              -2.023e-01  8.130e-01  -0.249    0.803
TheaterMTO              -7.031e-01  8.130e-01  -0.865    0.387
AllianceNon-Axis         3.561e-01  1.018e-02  34.993   <2e-16 ***
Industry_TypeNon-City   -6.557e-01  9.973e-03 -65.741   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.2173) family taken to be 1)

    Null deviance: 140932  on 62341  degrees of freedom
Residual deviance:  65859  on 62333  degrees of freedom
AIC: 499530
```

Diagnostic Plots

- QQ plot is not looking that great but the residual plot is not half bad considering its count data