GRM Final Project

Connor Putnam

11/14/2020

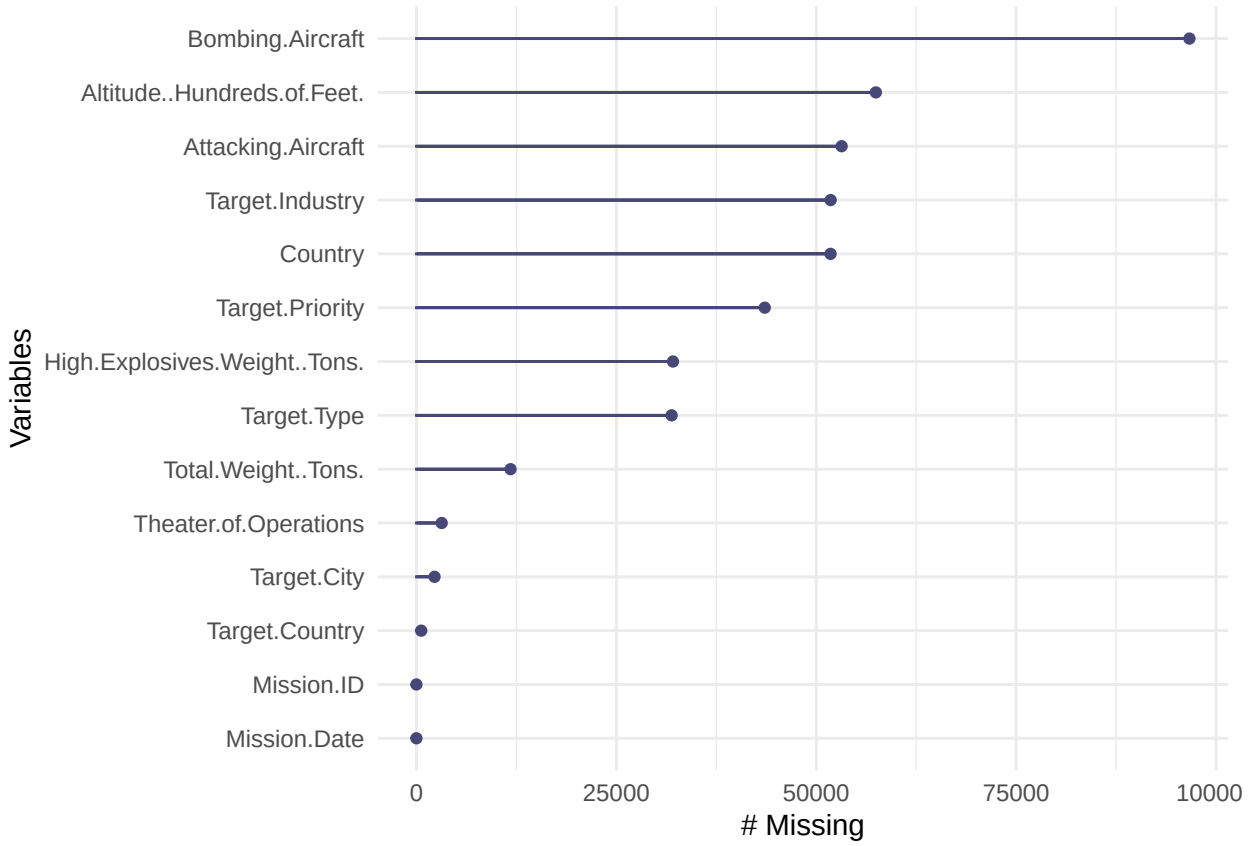
Question: What factros lead to the amount of bombs dropped being mission in WW2?

```
ww2 <- read.csv("operations.csv")

#variables of interest
ww2 <- ww2 %>%
  dplyr::select(Mission.ID, Mission.Date, Theater.of.Operations, Country, Target.Country, Target.City,
         Target.Industry, Target.Type, Altitude..Hundreds.of.Feet., Target.Priority,
         High.Explosives.Weight..Tons., Total.Weight..Tons., Attacking.Aircraft, Bombing.Aircraft) %>%
  mutate_all(na_if, "")

#Then going to see what variables have alot of missing values
ww2 %>%
  summarise_all(funs(sum(is.na(.))))
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
##   Mission.ID Mission.Date Theater.of.Operations Country Target.Country
## 1          0            0                  3158   51787            599
##   Target.City Target.Industry Target.Type Altitude..Hundreds.of.Feet.
## 1        2263           51802       31911                       57455
##   Target.Priority High.Explosives.Weight..Tons. Total.Weight..Tons.
## 1           43561                         32080               11773
##   Attacking.Aircraft Bombing.Aircraft
## 1              53173            96669
```

```
naniar::gg_miss_var(ww2)
```




```
ww2 <- ww2 %>%
  dplyr::select(Mission.ID, Mission.Date, Theater.of.Operations, Target.Country, Country, Target.Type, T
ww2 <- ww2 %>%
  drop_na(Total.Weight..Tons.) %>% #This will be my explainitory variable, so no missing
  drop_na(Theater.of.Operations) %>% # no way to impute this
  drop_na(Country) %>%
  drop_na(Target.Type) %>%
  drop_na(Target.Industry) %>%
  group_by(Target.Industry) %>%
  mutate(Altitude..Hundreds.of.Feet. =
           ifelse(is.na(Altitude..Hundreds.of.Feet.), mean(Altitude..Hundreds.of.Feet., na.rm = TRUE),
                  Altitude..Hundreds.of.Feet.)) %>%
  mutate(Attacking.Aircraft =
           ifelse(is.na(Attacking.Aircraft), mean(Attacking.Aircraft, na.rm = TRUE),
                  Attacking.Aircraft)) %>%
  mutate(Bombing.Aircraft =
           ifelse(is.na(Bombing.Aircraft), mean(Bombing.Aircraft, na.rm = TRUE),
                  Bombing.Aircraft)) %>%
  drop_na(Altitude..Hundreds.of.Feet.) %>%
  drop_na(Attacking.Aircraft) %>%
  drop_na(Target.Country)
ww2 <- ww2 %>%
  rename(Date = Mission.Date,
         Theater = Theater.of.Operations,
         Target_Country = Target.Country,
         Target_Type = Target.Type,
         Total_Weight = Total.Weight..Tons.,
```

```
         Industry = Target.Industry,
         Altitude = Altitude..Hundreds.of.Feet.,
         Attacking_Aircraft = Attacking.Aircraft,
         Bombing_Aircraft = Bombing.Aircraft)
ww2 <- ww2 %>%
  mutate(Total_Weight = round(Total_Weight)) %>%
  mutate_at(vars(Total_Weight), as.integer) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%Y")) %>%
  mutate(System_Time = as.numeric(as.POSIXct((Date)))) %>%
  mutate(Alliance = if_else(grepl("AUSTRIA", Target_Country), "Axis",
                            if_else(grepl("GERMANY", Target_Country), "Axis",
                            if_else(grepl("BULGARIA", Target_Country), "Axis",
                            if_else(grepl("SLOVAKIA", Target_Country), "Axis",
                            if_else(grepl("HUNGARY", Target_Country), "Axis",
                            if_else(grepl("ROMANIA", Target_Country), "Axis",
                            if_else(grepl("BULGARIA", Target_Country), "Axis",
                            if_else(grepl("CROATIA", Target_Country), "Axis",
                            if_else(grepl("IRAQ", Target_Country), "Axis",
                            if_else(grepl("ITALY", Target_Country), "Axis",
                            if_else(grepl("FINLAND", Target_Country), "Axis",
                            if_else(grepl("THAILAND", Target_Country), "Axis",
                            if_else(grepl("CROATIA", Target_Country), "Axis", "Non-Axis")))))))))))))
ww2
```

```
## # A tibble: 62,342 x 13
## # Groups:   Industry [50]
##    Mission.ID Date       Theater Target_Country Country Target_Type Total_Weight
##         <int> <date>     <fct>   <fct>          <fct>   <fct>              <int>
##  1         12 1943-08-15 ETO     GERMANY        GREAT ~ CITY AREA              1
##  2         13 1943-08-15 ETO     GERMANY        GREAT ~ CITY AREA              4
##  3         58 1943-08-15 ETO     GERMANY        GREAT ~ CITY AREA             87
##  4         66 1943-08-15 MTO     ITALY          USA     SHIPPING               2
##  5         67 1943-08-15 MTO     ITALY          USA     SHIPPING               2
##  6         68 1943-08-15 MTO     ITALY          USA     ROAD                  17
##  7         69 1943-08-15 MTO     ITALY          USA     ROAD                  17
##  8         70 1943-08-15 MTO     ITALY          USA     MARSHALL Y~           42
##  9         71 1943-08-15 MTO     ITALY          USA     MARSHALL Y~           42
## 10         72 1943-08-15 MTO     ITALY          USA     SUPPLIES               7
## # ... with 62,332 more rows, and 6 more variables: Industry <fct>,
## #   Altitude <dbl>, Attacking_Aircraft <dbl>, Bombing_Aircraft <dbl>,
## #   System_Time <dbl>, Alliance <chr>
```

```
ww2 <- ww2 %>%
  mutate(Industry_Type = if_else(grepl("CITIES TOWNS AND URBAN AREAS",
                                       Industry), "City", "Non-City")) %>%
  dplyr::select(-Industry, -Target_Type)
```

```
## Adding missing grouping variables: `Industry`
```

```
formattable::formattable(ww2 %>% head())
```

Industry

Mission.ID

Date

Theater

Target_Country

Country

Total_Weight

Altitude

Attacking_Aircraft

Bombing_Aircraft

System_Time

Alliance

Industry_Type

CITIES TOWNS AND URBAN AREAS

12

1943-08-15

ETO

GERMANY

GREAT BRITAIN

1

250

11.462295

11.462295

-832550400

Axis

City

CITIES TOWNS AND URBAN AREAS

13

1943-08-15

ETO

GERMANY

GREAT BRITAIN

4

250

5.000000

5.000000

-832550400

Axis

City

CITIES TOWNS AND URBAN AREAS

58

1943-08-15

ETO

GERMANY

GREAT BRITAIN

87

135

11.462295

11.462295

-832550400

Axis

City

SHIPS

66

1943-08-15

MTO

ITALY

USA

2

95

16.857143

16.857143

-832550400

Axis

Non-City

SHIPS

67

1943-08-15

MTO

ITALY

USA

2

95

16.857143

16.857143

-832550400

Axis

Non-City

HIGHWAYS AND VEHICLES

68

1943-08-15

MTO

ITALY

USA

17

95

6.733871

6.733871

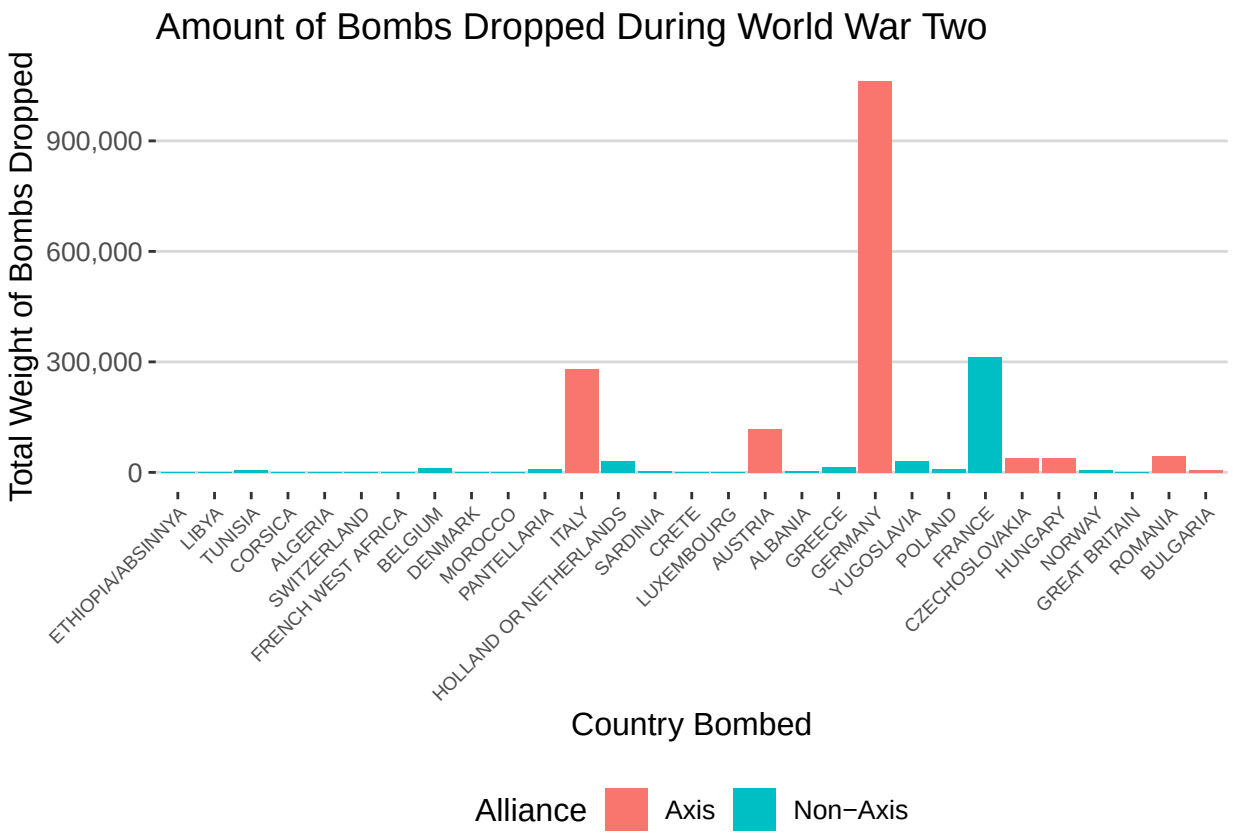
-832550400

Axis

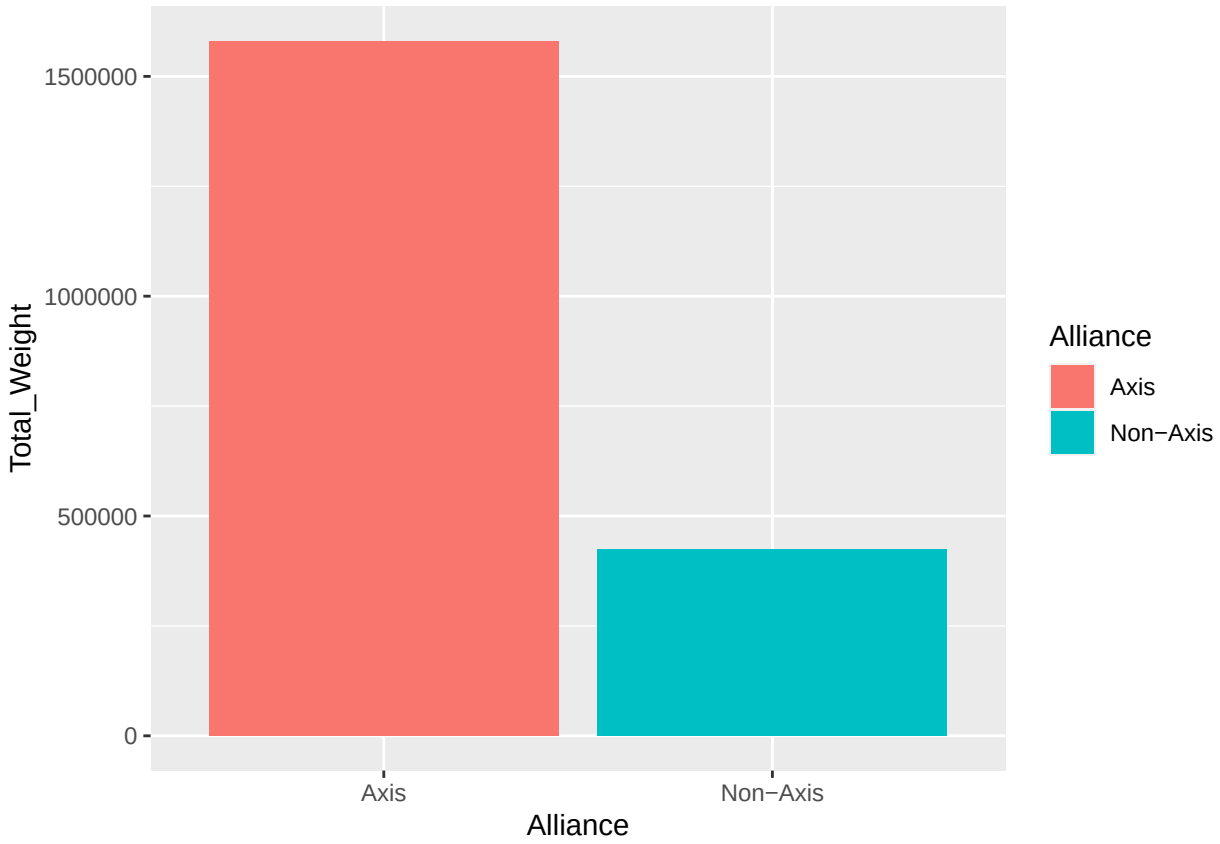
Non-City

EDA

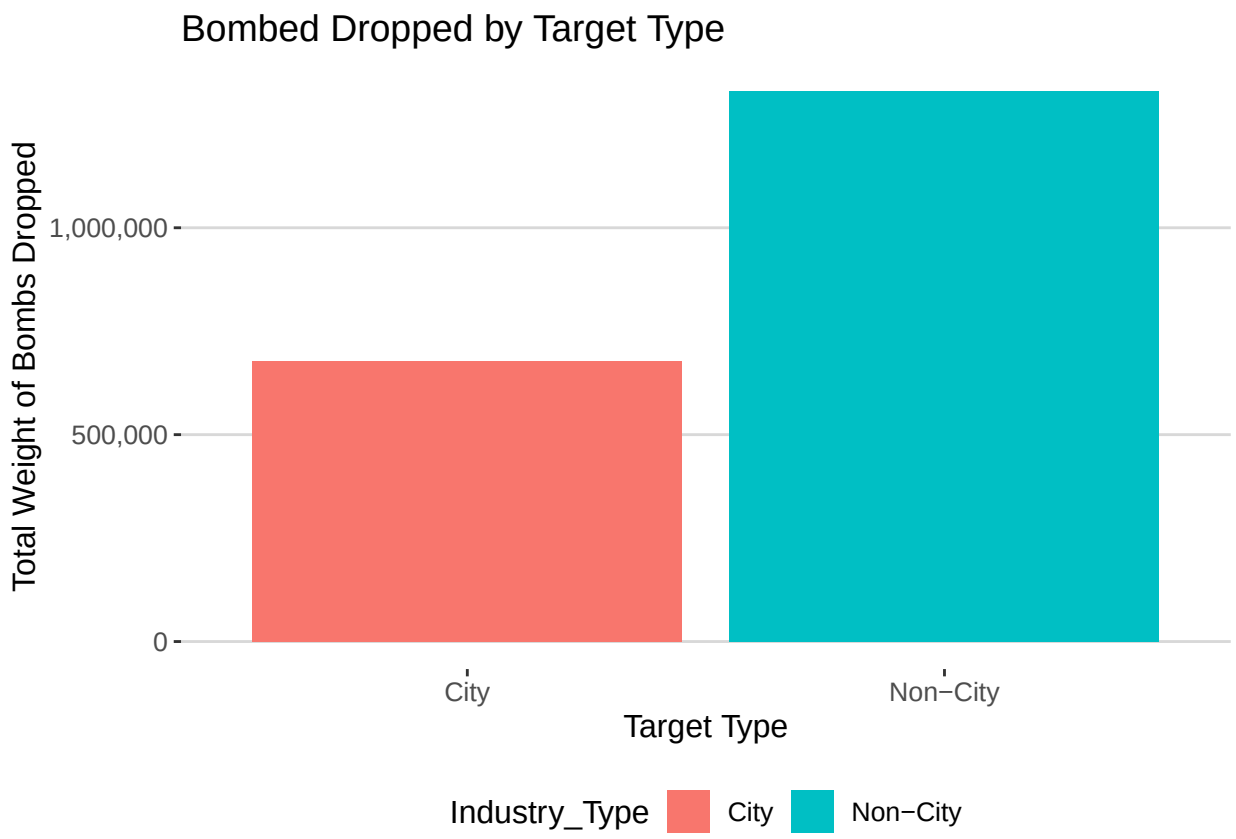
```
ggplot(ww2, aes(reorder(Target_Country, Total_Weight),
                Total_Weight, fill = Alliance)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(labels = comma) +
  theme_hc() +
  theme(axis.text.x = element_text(angle = 45, hjust=1, size = 7)) +
  labs(title = "Amount of Bombs Dropped During World War Two") +
  xlab("Country Bombed") +
  ylab("Total Weight of Bombs Dropped")
```




```
  #ggtitle("Bombs", size = .5)

ggplot(ww2, aes(Alliance, Total_Weight, fill = Alliance))  +
  geom_bar(stat = "identity")
```




```
ggplot(ww2, aes(Industry_Type, Total_Weight, fill = Industry_Type))  +
  geom_bar(stat = "identity") +
  theme_hc() +
  scale_y_continuous(labels = comma) +
  labs(title = "Bombed Dropped by Target Type") +
  xlab("Target Type") +
  ylab("Total Weight of Bombs Dropped")
```



library(MASS) fitdistr(ww2$Total_Weight, “negative binomial”)

negbin_sim <- reshape2::melt(ww2$Total_Weight) negbin_sim <- as.data.frame(table(negbin_sim))

ggplot(ww2, aes(Total_Weight)) + geom_histogram(bins = 100, color = “black”, fill = “steelblue”) + theme_hc() + xlab(“Total Bombing Weight per Mission”) + ylab(“Count”) + ggtitle(“Distribution of Counts”)

first_model <- glm(Total_Weight ~ Attacking_Aircraft, data = ww2, family = poisson(link = “log”)) summary(first_model)

Sense there is a tone of over dispersion I will try the negative binomial

neg_model <- glm.nb(Total_Weight + 1 ~ log(Attacking_Aircraft) + log(Bombing_Aircraft) + Altitude + System_Time + Theater + Alliance + Industry_Type, data = ww2) summary(neg_model) AIC(neg_model)

https://www.youtube.com/watch?v=vN5cNN2-HWE

QQ_nb_war <- qplot(sample = .stdresid, data = neg_model, stat = “qq”) + geom_abline() + ggtitle(“QQ Plot”)

Res_nb_war <- qplot(.fitted, .resid, data = neg_model) + geom_hline(yintercept = 0) + geom_smooth(se = FALSE) + ggtitle(“Residual Plot”)

gridExtra::grid.arrange(QQ_nb_war, Res_nb_war, ncol =2)

```
#boxcox(neg_model, lambda = seq(-1, 1))
```

```
min(ww2$Total_Weight)
```

```
## [1] 0
```

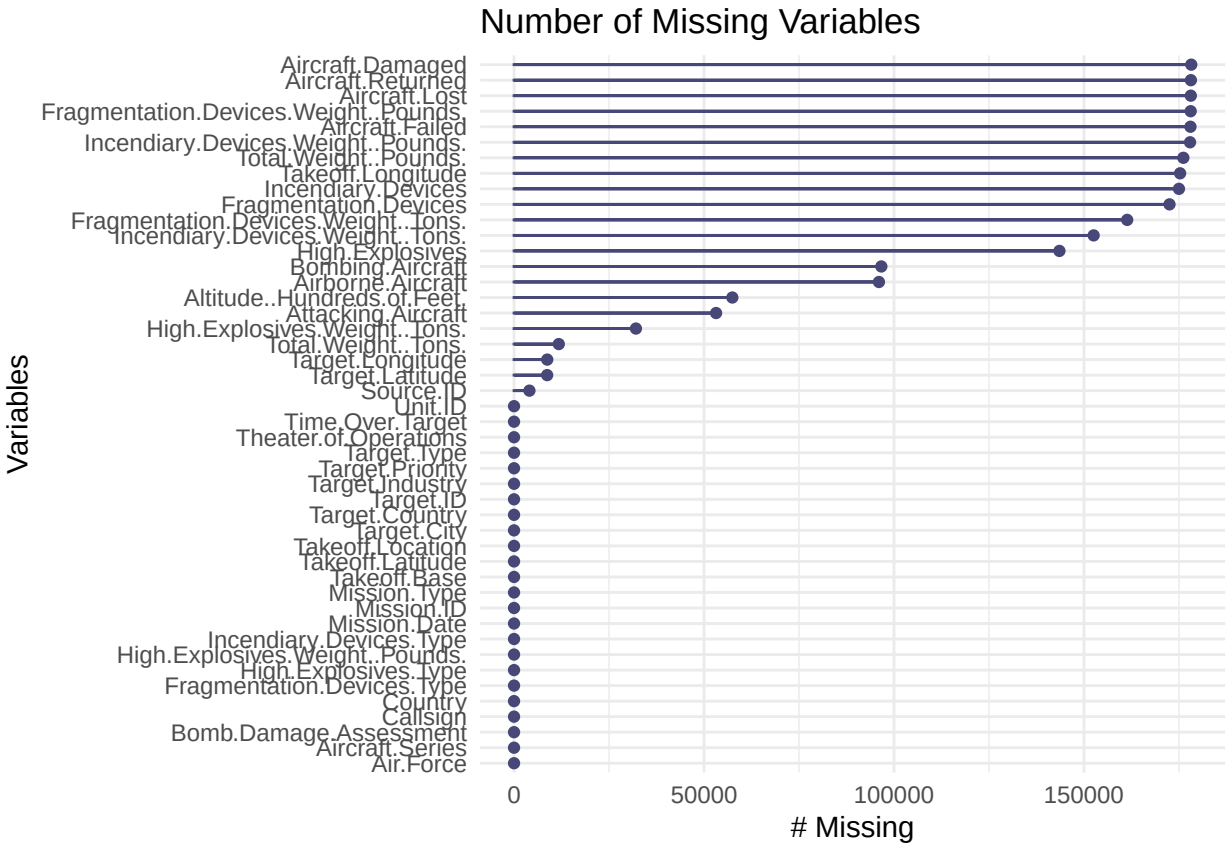
Overview

The goal of this project was to model the amount of bombs drops on a given mission during World War 2. When setting out to do this project I wanted to implimentent generalized linear models as per the assignemnt requirements, but I also what to explore between generalized linear models and big data. In order to get my toes wet with this I decided abone this this Arieal Bombing Dataset that can be found at https://www.kaggle.com/usaf/world-war-ii. The dataset contains 178,281 observations, each indicating a single bombing mission between the May 15, 1940 and May 2, 1945. This data only covers Allied operations an thus this is why the records begin approximately eight months after the German advancement into Poland. With final observations coming a few days before the German surrender. The data set also contains 46 variables.

Question of Interest:

What factors lead to the amount of bombs dropped on a mission in WW2?

I will set the response variable to be the Total Weight of bombs dropped during a single mission. Some of the variables are not relevant to answering this question so I will discard those. Additionally, many of the variables are categorial variables and have many different factor levels. In order for my analysis more interpretable I might have to perform some data cleaning and categorization. As with many large data sets there was alot of missing variable, this can also be explained in part becasue this is historical data during a time of war. The plot below shows that many of the variables have too many missing values and thus will have to be dropped from potential analysis.



After addressing this the next step was to impute and as well decide what variable to keep for analaysis. Based on the ability to impute as well as the relevence to myt quesiton of interest I decided on the following varibales for analysis. The table below list out variables being used after data wrangling/cleaning as well the

first six obervations for those values.

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

Poisson Regression

After looking at the counts of bomb droped I devided to process with the following Poisson regression,

```
glm(Total_Weight ~ Attacking_Aircraft + Bombing_Aircraft + Altitude + System_Time +
Theater + Alliance, data = ww2, family = poisson(link = "log"))
```

This resulted in some interesting findings regarding signifigance but ulitmately cannot be used becasue the deviance is clocking in at 3.3044099×10^6 on 62334 degrees of freedom. So there is clearly an issue with overdispersion at play. The next time to is to take the quasi poisson but that also resulted in a similiar overdisprerion issue. Addtioanlly even with vearious transformations should as inverse and logarithmic is there was no meaningful improvement in this model.

Negative Binomial

Sense both the Poisson and the Quasi-Poisson reuslted in over disperion I decided to try my luck on a negative binomial model instead. The first negative binomial model I choose was the following:

```
glm.nb(Total_Weight ~ Attacking_Aircraft + Bombing_Aircraft + Altitude + System_Time +
Theater + Alliance + Industry_Type, data = ww2)
```

This resulted in a much better deviance value of 6.8687553×10^4 on 62333 degrees of freedom. This level off overdispersion is much more managebale, esspeically considering the size of the data. The overdispersion ration sits at approximaley 1.1 Now the next step was to to see if I could improve upon this in terms of AIC values and even less overdisperison.