Data Demo

Contents

```
df1 = read.csv('data/russia_losses_equipment.csv')
df2 = read.csv('data/russia_losses_personnel.csv')

df2_1 = subset(df2, select = c(personnel, personnel., POW))
df = cbind(df1, df2_1)

df %>% head()
```

```
##         date day aircraft helicopter tank APC field.artillery MRL military.auto
## 1 2022-02-25   2       10          7   80 516              49   4           100
## 2 2022-02-26   3       27         26  146 706              49   4           130
## 3 2022-02-27   4       27         26  150 706              50   4           130
## 4 2022-02-28   5       29         29  150 816              74  21           291
## 5 2022-03-01   6       29         29  198 846              77  24           305
## 6 2022-03-02   7       30         31  211 862              85  40           355
##   fuel.tank drone naval.ship anti.aircraft.warfare special.equipment
## 1        60     0          2                     0                NA
## 2        60     2          2                     0                NA
## 3        60     2          2                     0                NA
## 4        60     3          2                     5                NA
## 5        60     3          2                     7                NA
## 6        60     3          2                     9                NA
##   mobile.SRBM.system personnel personnel. POW
## 1                 NA      2800      about   0
## 2                 NA      4300      about   0
## 3                 NA      4500      about   0
## 4                 NA      5300      about   0
## 5                 NA      5710      about 200
## 6                 NA      5840      about 200
```

1. An overview of dataset

1.1 What does it include?

This is a dataset for 46 days of Equipment Losses & Death Toll & Military Wounded & Prisoner of War of russians in the Ukraine Russia War. The data is separated to equipment and personnel. df1 is equipment and df2 is personnel. df is combined dataset.

1.2 Where and how will you be obtaining it? Include the link and source.

https://www.kaggle.com/datasets/piterfm/2022-ukraine-russian-war

I got the dataset from Kaggle.

1.3 About how many observations? How many predictors?

It contains 45 observations and 18 variables. Data and Day variables are repeated. This dataset is time series. However, the observations could be updated later.

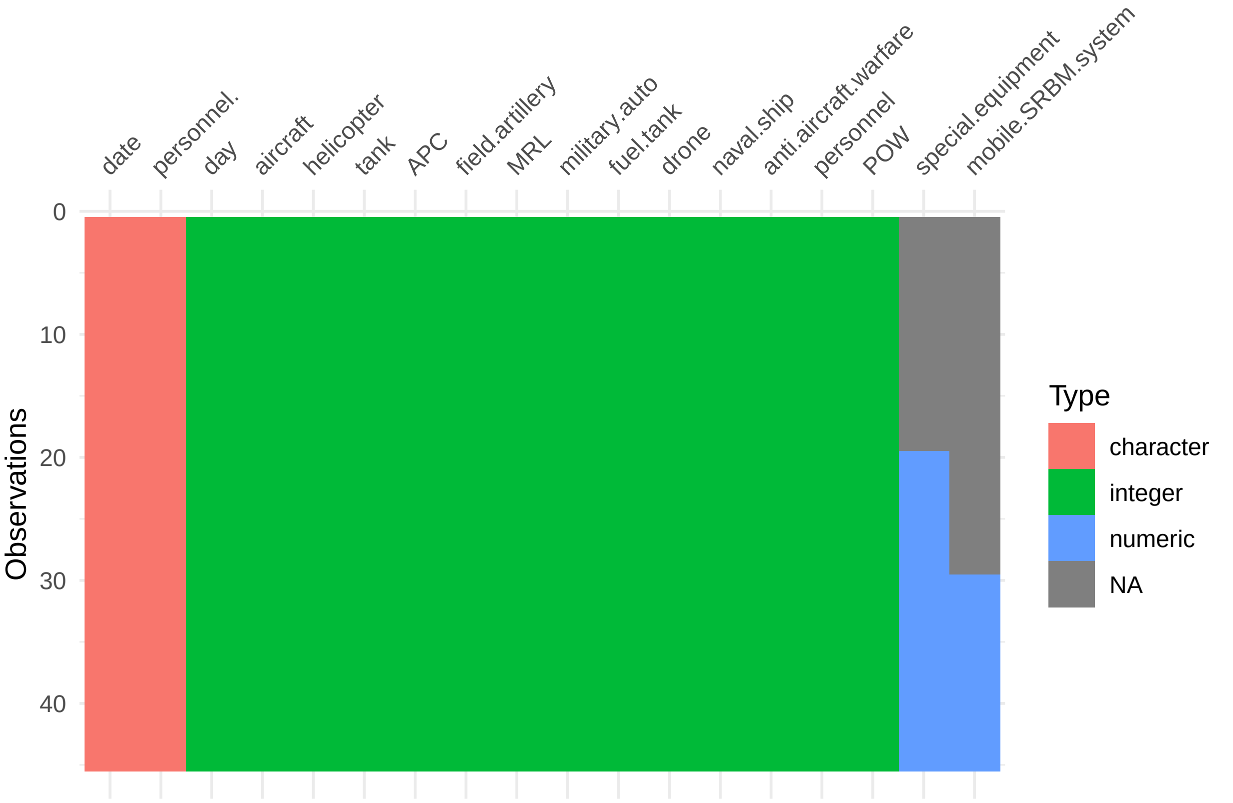
```
print(dim(df))
```

```
## [1] 45 18
```

1.4 What types of variables will you be working with?

```
vis_dat(df)
```

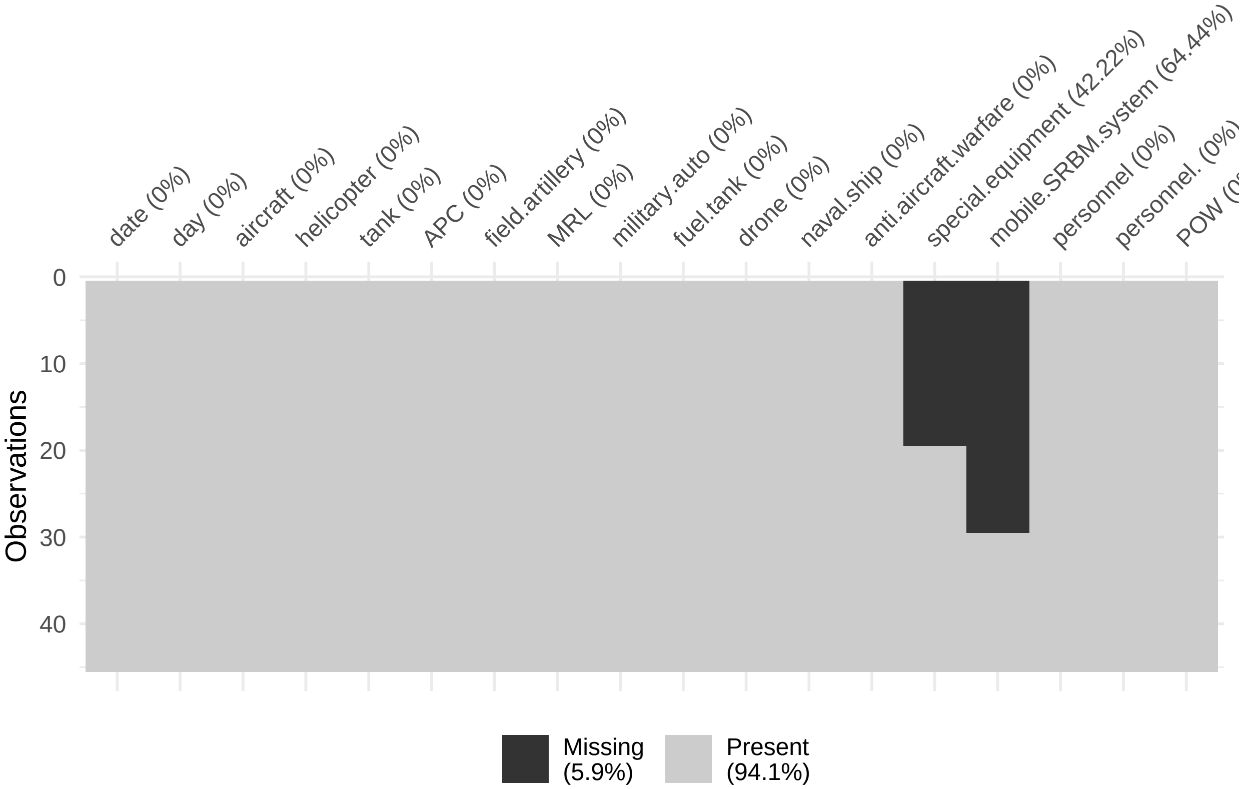
```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



There are character, integer and numeric.

1.5 Is there any missing data? About how much? Do you have an idea for how to handle it?

```
vis_miss(df)
```



The above two variables which have NA are both numeric. I can consider mean value to fill NA.

2. An overview of research question(s)

2.1 What variable(s) are you interested in predicting? What question(s) are you interested in answering?

I think human casualties is the most concerning issue, so I am interested in personnel column. Also I want to figure out prisoner of war(POW column).

2.2 Name your response/outcome variable(s) and briefly describe it/them.

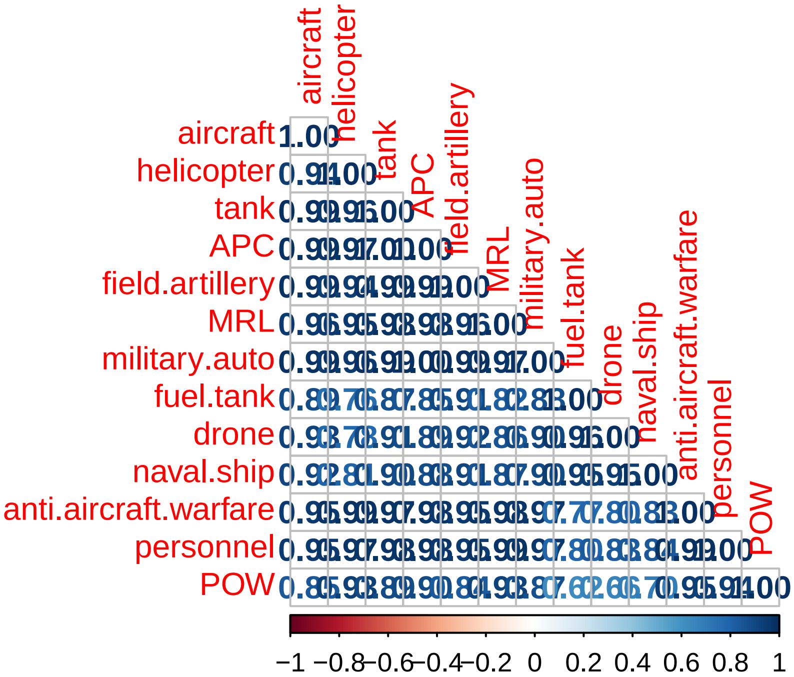
It will be personnel and POW column, but I am still not sure how to combine these two information. Personnel shows the number of Russian soldiers died and POW indicates the number of Russian prisoners of war due to war. It might be mostly soldiers.

2.3 Will these questions be best answered with a classification or regression approach?

This project needs regression approach. Those two concerning variables are both continuous and there is no categorical in whole dataset.

2.4 Which predictors do you think will be especially useful?

```
df_cor = subset(df, select = c(-day, -special.equipment, -mobile.SRBM.system))
M = cor(df_cor %>% dplyr::select(where(is.numeric)))
corrplot(M, method = "number", type = "lower")
```



All variables are strongly correlated. It needs more EDA.

2.5 Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

I will use descriptive and inferential model. I am going to focus on relationship and trend between variavles rather than prediction. KNN and SVM are examples for the model.

3. Project timeline

I will focus on EDA until this month. Dataset is quite small and there is no exact direction of this project - I am not sure to predict which variables, I plan to gain insight into data with EDA. Then, apply several models to predict and explain on next month.

4. Extra questions

The dataset sample size is very small so I don't have any idea to split train/test for prediction. It will hard to predict exact value such as the number of people died. Instead, I will focus to describe relationships between variables. This project will be unsupervised learning. My main goal is to gain interesting insight from the war data.