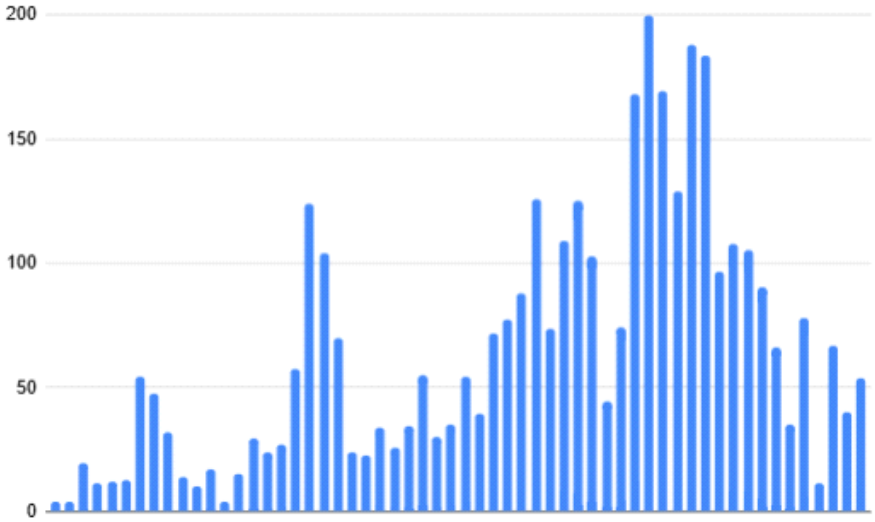
Project Land Occupation
by Shvets Ivan

For this project I have chosen the rate of occupation of Ukrainian land during russo-ukrainian War (2022 -). Dataset includes land occupied measured in km² beginning from January 2024 to February 2025. So in total it's going to be 59 weeks of measurements. I think this dataset became significantly important for Ukrainian defense matters, as we should understand how, where and how much land are we losing in order to stop this trend.



Graphical representation of my data by week/land_occupied

Data Analysis

I have done a Descriptive Data analysis in Excel and got these results from my dataset.

Descriptive Data	
Mean	64,25637931
Standard Error	6,677789425
Median	53,79
Mode	34,89
Standard Deviation	50,85652911
Sample Variance	2586,386553
Kurtosis	0,3442214293
Skewness	1,009853668
Range	196,21
Minimum	3,74
Maximum	199,95
Sum	3726,87
Count	58
Largest(1)	199,95
Smallest(1)	3,74
Confidence Level(95%)	13,3720423

Based on these variables, I can do my further analysis.

- For the **Center of my Data** (Measure of Location), I have found 3 variables: Mean (Arithmetic Average), Median (Middle Value of Data) and Mode (Most Frequent Value). Since: **Mean** (64,25) > **Median**(53,79) > **Mode**(34,89), it is safe to assume that my Data is **right-skewed** (long tail on the right).

- **Measure of Spread** (Variability) of the data helps understand how much the data varies and whether the values are close together or widely dispersed. Key variables for my measure of spread are: **Range** (Difference

between Maximum and Minimum Values), **Interquartile Range or IQR** (Spread of the Middle 50% of Data), **Variance & Standard Deviation** (How Far Data is From the Mean). I have calculated Range and SD using Excel, and IQR using R (IQR = 70). According to the table we can see that our variables are **extremely high**, and we are safe to assume that the data is **highly variable**, and predictions are less reliable.

Measure	Small (Low Variability)	Large (High Variability)
Range	< 50% of the mean	> 100% of the mean
IQR	< 50% of the mean	> 50% of the mean
Standard Deviation (SD)	< 15% of the mean	> 30% of the mean

Range = 196, IQR = 70, SD = 50, Mean = 64

- To determine whether data is **Symmetrically Distributed**, we can use two options: look at the **Skewness** of the data, and to look at 2 **Graphs** (QQ plot, Histogram and Box Plot). Since our Skewness > 0 (1,009), we determine that our data is right-skewed. As you can see below in the attached pictures on the Box Plot, Median is closer to Q1 (bottom), longer whisker above. That means that our data is not symmetrical and is right skewed. Also, looking at the QQ plot, we can see how the points in the upper tail of the plot deviate above the line. This shows us once again that the data is right skewed. And, finally, visually looking at histogram gives us an Idea of how the data is distributed with visual right skewness.

- Personally, I wouldn't want to apply **The Three Sigma Rule** on this dataset for a few reasons. **The Presence of Extreme Values**: The dataset contains some large values (e.g., 199.95, 169.24) that are far from the mean (64.26). These large values could be influencing the standard deviation significantly, which could make the three-sigma range very wide and potentially less meaningful. **Skewness**: The dataset is right-skewed, meaning there are more smaller values and a few larger values. In such cases, the three-sigma rule (which assumes a normal distribution) is not best. **Applicability**: The three-sigma rule works best when the data is approximately **normally distributed**. If the distribution is not normal, the

results will not accurately reflect the "true" spread of the data, and applying the rule could give misleading conclusions about what constitutes an "outlier." So no, I would use the three sigma rule.

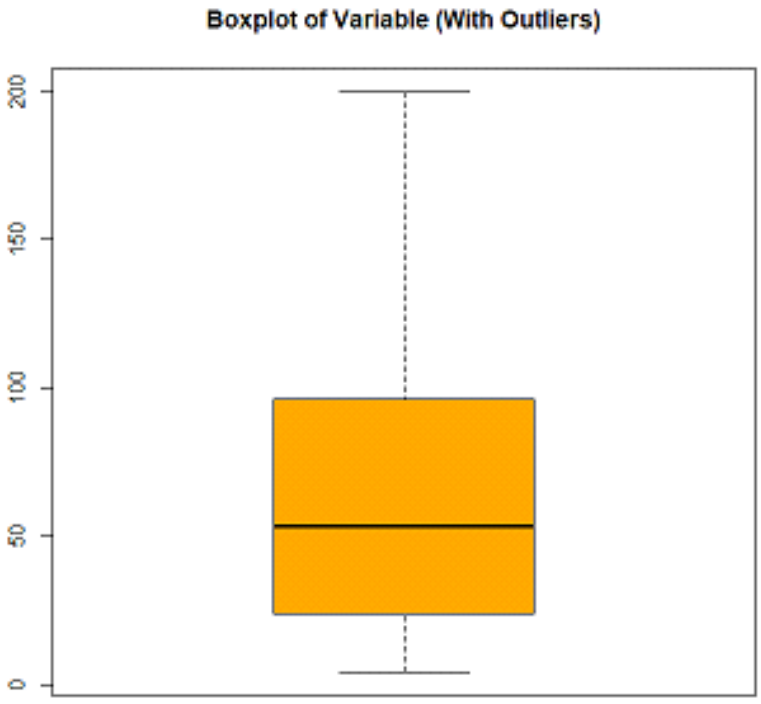
- There are a few ways to determine whether my data is **Normal** or not. First of all, **Visual Method**: We can see on the Histogram that it does not represent a classical bell-shaped curve, which is common to normally distributed data. QQ Plot shows us how point deviate in the upper tail above the line, which shows us that data is right-skewed and likely not normally distributed. **Descriptive Variables** show us skewness and kurtosis, both of which indicate that data is not normal (skewness = 1,009; kurtosis = 0,344). And finally, **Shapiro-Wilk Test.** It tests the null hypothesis that the data follows a normal distribution. If the p-value is low (< 0.05), you reject the null hypothesis, suggesting the data is not normally distributed. Our p value, according to R, is 0,00014, which shows that data is not normally distributed.

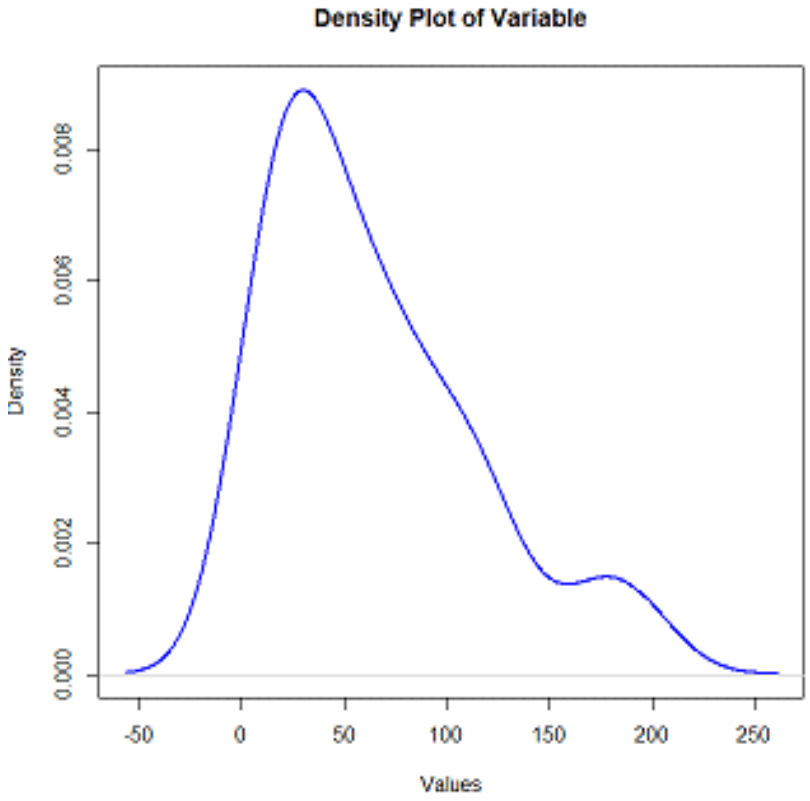
- In order to acknowledge if there are any **Outliers**, we need to do a step-by-step calculations. We need to calculate **Lower and Upper Bounds**, which can show us the outliers when the data is skewed. My Q1 = 24.48, and Q3 = 94.95. IQR = 70. Lower Bound=Q1−1.5×IQR, Upper Bound=Q3+1.5×IQR.

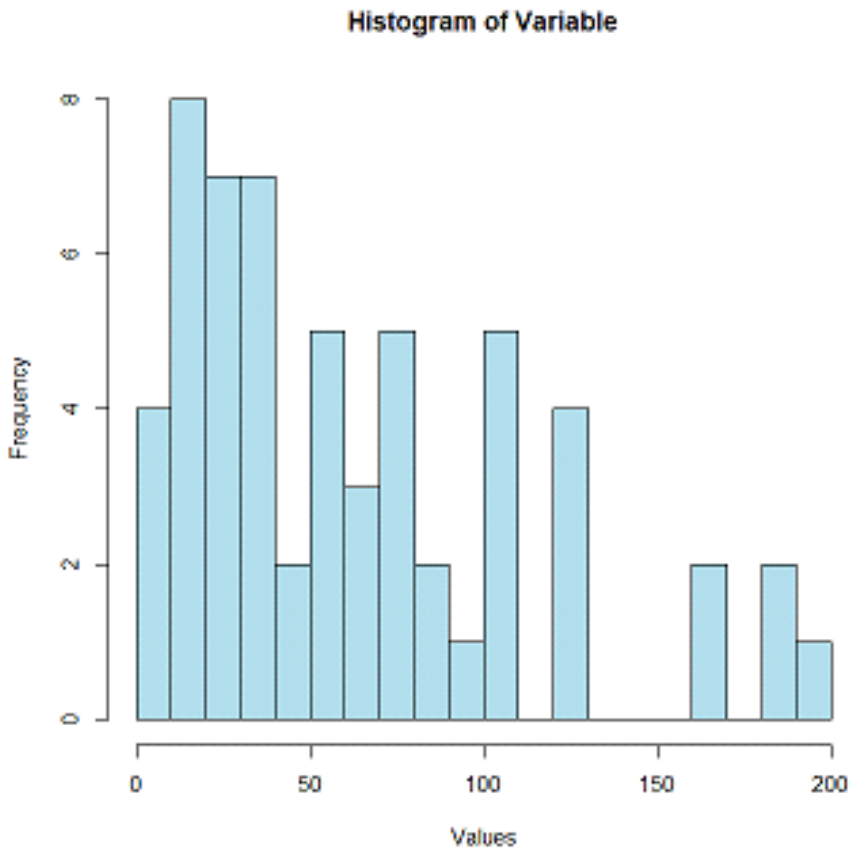
  Lower Bound = -81.22
  Upper Bound = 200.65

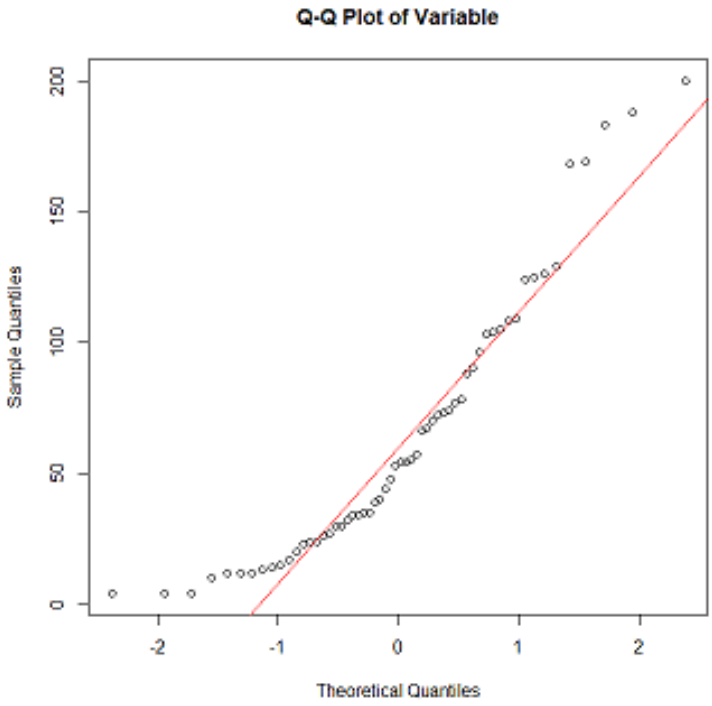
  Since all values in the dataset fall within this range, there were **no outliers**.

ATTACHMENTS




Density Plot of Variable
Density
0.000
0.002
0.004
0.006
0.008
-50
0
50
100
150
200
250
Values


Histogram of Variable
Frequency
0
2
4
6
8
0
50
100
150
200
Values



R Code:

```
R version 4.4.2 (2024-10-31 ucrt) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
```

```
> data <- read.csv("C:/Users/Ivan/Desktop/Jacyk.R/plotR.csv")
> head(data)   # View first few rows
   LO
1   4
2   4
3 20
4 12
5 12
6 13
> str(data)    # Check structure of dataset
'data.frame':    58 obs. of   1 variable:
 $ LO: int   4 4 20 12 12 13 54 48 32 14 ...
> summary(data)   # Get an overview of all numeric variables
       LO
 Min.    :   4.00
 1st Qu.: 24.50
 Median : 53.50
 Mean    : 64.26
 3rd Qu.: 94.50
 Max.    :200.00
> mean(var, na.rm = TRUE)
[1] NA
Warning message:
In mean.default(var, na.rm = TRUE) :
   argument is not numeric or logical: returning NA
> head(data)   # View first few rows
   LO
1   4
2   4
3 20
4 12
5 12
6 13
> str(data)    # Check structure of dataset
```

```
'data.frame':     58 obs. of   1 variable:
 $ LO: int   4 4 20 12 12 13 54 48 32 14 ...
> summary(data)   # Get an overview of all numeric variables
       LO
 Min.     :   4.00
 1st Qu.: 24.50
 Median : 53.50
 Mean     : 64.26
 3rd Qu.: 94.50
 Max.     :200.00
> data <- data$LO
> mean(data, na.rm = TRUE)
[1] 64.25862
> median(data, na.rm = TRUE)
[1] 53.5
> mode_func <- function(x) {
+       uniq_vals <- unique(x)
+       uniq_vals[which.max(tabulate(match(x, uniq_vals)))]
+ }
> mode_func(data)
[1] 4
> sd(data, na.rm = TRUE)
[1] 50.82704
> var(data, na.rm = TRUE)
[1] 2583.388
> range(data, na.rm = TRUE)
[1]     4 200
> IQR(data, na.rm = TRUE)
[1] 70
> hist(data, breaks = 20, col = "lightblue", main = "Histogram of
Variable",
+         xlab = "Values", ylab = "Frequency", border = "black")
> boxplot(data, main = "Boxplot of Variable", col = "lightgreen",
horizontal = TRUE)
> plot(density(data, na.rm = TRUE), main = "Density Plot of Variable",
```

```
xlab = "Values", col = "blue", lwd = 2)
> mean_val <- mean(data, na.rm = TRUE)
> sd_val <- sd(data, na.rm = TRUE)
>
> lower_bound <- mean_val - 3 * sd_val
> upper_bound <- mean_val + 3 * sd_val
>
> within_3sigma <- sum(data >= lower_bound & data <=
upper_bound) / length(data) * 100
> within_3sigma
[1] 100
> shapiro.test(data)

        Shapiro-Wilk normality test

data:  data
W = 0.89767, p-value = 0.00014

> qqnorm(data, main = "Q-Q Plot of Variable")
> qqline(data, col = "red")
> Q1 <- quantile(data, 0.25, na.rm = TRUE)
> Q3 <- quantile(data, 0.75, na.rm = TRUE)
> IQR_value <- Q3 - Q1
>
> lower_bound <- Q1 - 1.5 * IQR_value
> upper_bound <- Q3 + 1.5 * IQR_value
>
> outliers <- data[data < lower_bound | data > upper_bound]
> outliers
[1] 200
> boxplot(var, main = "Boxplot of Variable (With Outliers)", col =
"orange")
Error in sort.int(x, na.last = na.last, decreasing = decreasing, ...) :
  'x' must be atomic
In addition: Warning messages:
```

```
1: In is.na(x) : is.na() applied to non-(list or vector) of type 'closure'
2: In is.na(x) : is.na() applied to non-(list or vector) of type 'closure'
> text(x = 1, y = outliers, labels = outliers, col = "red", pos = 3)
Error in text.default(x = 1, y = outliers, labels = outliers, col = "red",  :
  plot.new has not been called yet
> boxplot(data, main = "Boxplot of Variable (With Outliers)", col =
"orange")
> text(x = 1, y = outliers, labels = outliers, col = "red", pos = 3)
```