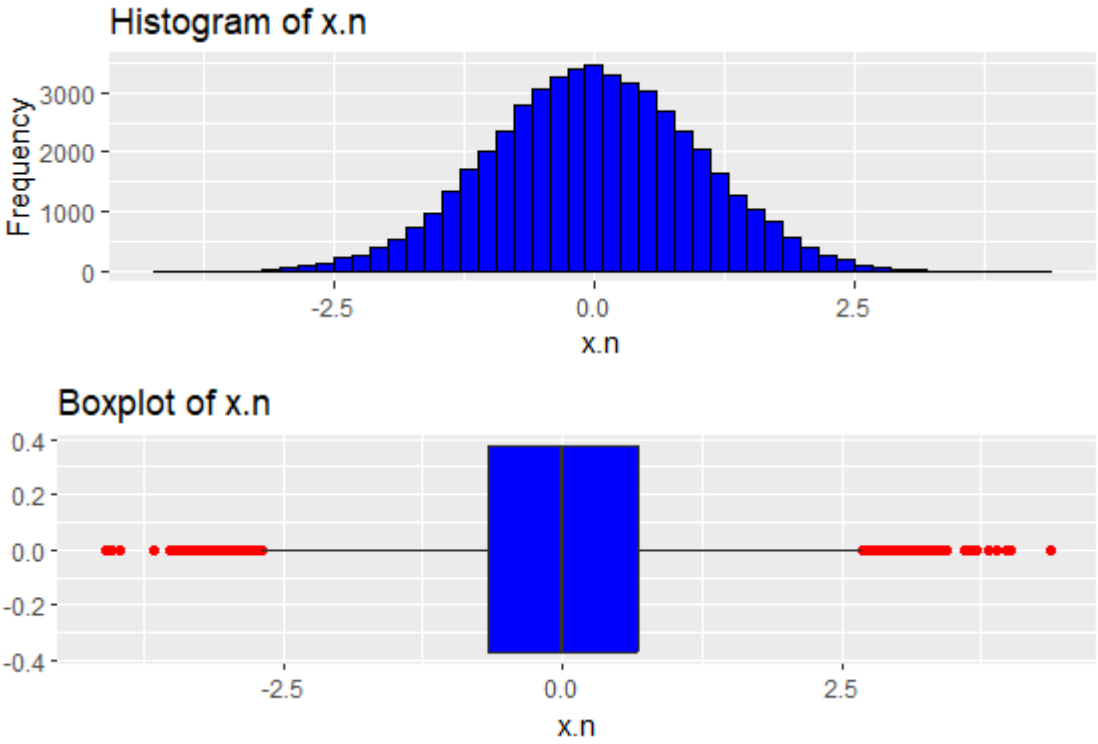
# Assignment

**Mykola Smyrnyi**

**21.06.2023**

**Task 1**

**1) RESULT:**




```
Data <- data.frame(x.n = rnorm(50000),
                   x.p = rPareto(50000, t = 1, alpha = 2))


histogram1 <- ggplot(data = Data, aes(x = x.n)) +
              geom_histogram(bins = 50, fill = "blue", color="black") +
              labs(x = "x.n", y = "Frequency", title = "Histogram of x.n")
boxplot1 <- ggplot(data=Data, aes(x = x.n)) +
            geom_boxplot(fill = "blue", outlier.colour = 'red') +
            labs(x = "x.n", title = "Boxplot of x.n")
grid.arrange(histogram1, boxplot1)
```

**2) RESULT:**

- Mean: -0.0002084956
- Standadrd deviation: 0.9989658

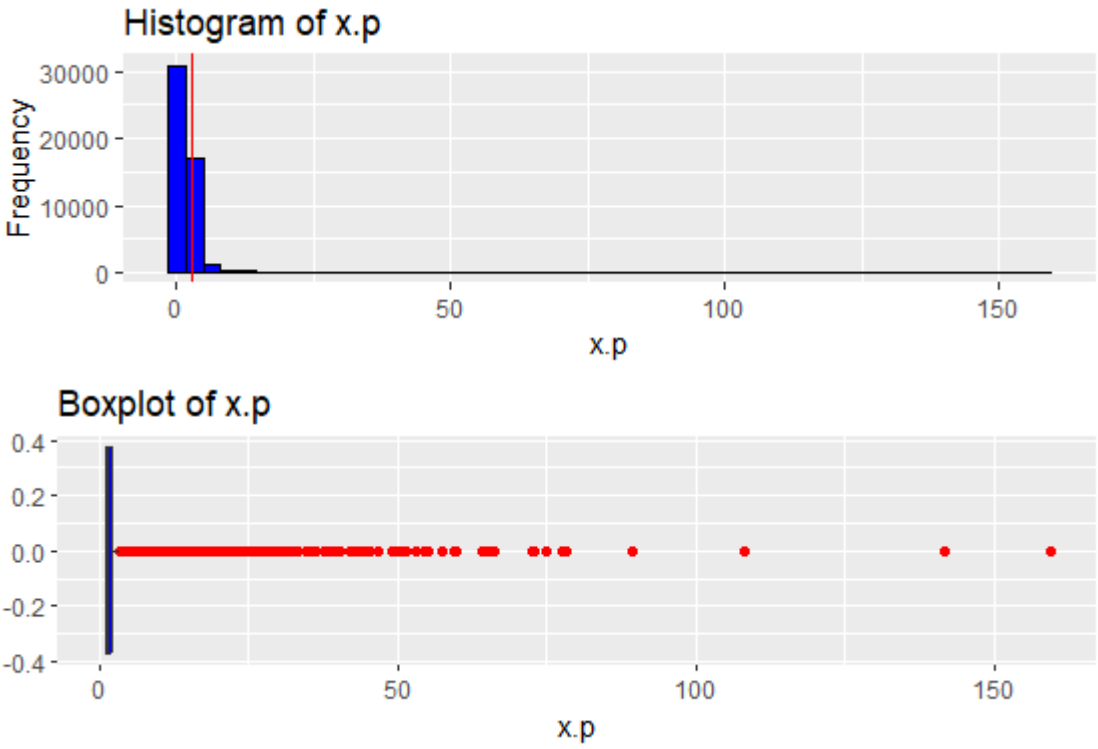
Yes, we would expect simillar values as we took sample from standardized normal distribution where mean equals 0 and sd equals 1. So, deviations that we got are not sufficient to consider them invalid for our expectations.

**3) RESULT:**

- We can say that 95% of future observations will lay in 2 standard deviations from the mean (-1.9980084956, 1.9975915044) because our data is normally distributed.
- Mean can be used to predict future observations because we have "bell" form data.

**4) RESULT:**

Lets plot histogramm and boxplot to observe x.p




We may notice that we are dealing with a lot of outliers, so lets get rid of them using **filter** method:

```
cutoff_up = mean_xp + 3*sd_xp
cutoff_down = mean_xp - 3*sd_xp
subset <- dplyr::filter(Data, (x.p >= cutoff_down) & (x.p <= cutoff_up))
```

Now we can compare mean and standard deviation of full sample and the subset to address the question.

- **Mean of x.p full**: 1.99390436133807
- **Standard Deviation of x.p full**: 2.6011734070772
- **Mean of x.p subset**: 1.80737030807044
- **Standard Deviation of x.p subset**: 1.13436458420453

Hence, we cannot sumarize the variable because outliers affect mean and standard deviation sufficiently. Also, mean is a bad predictor as we have right-skewed data.

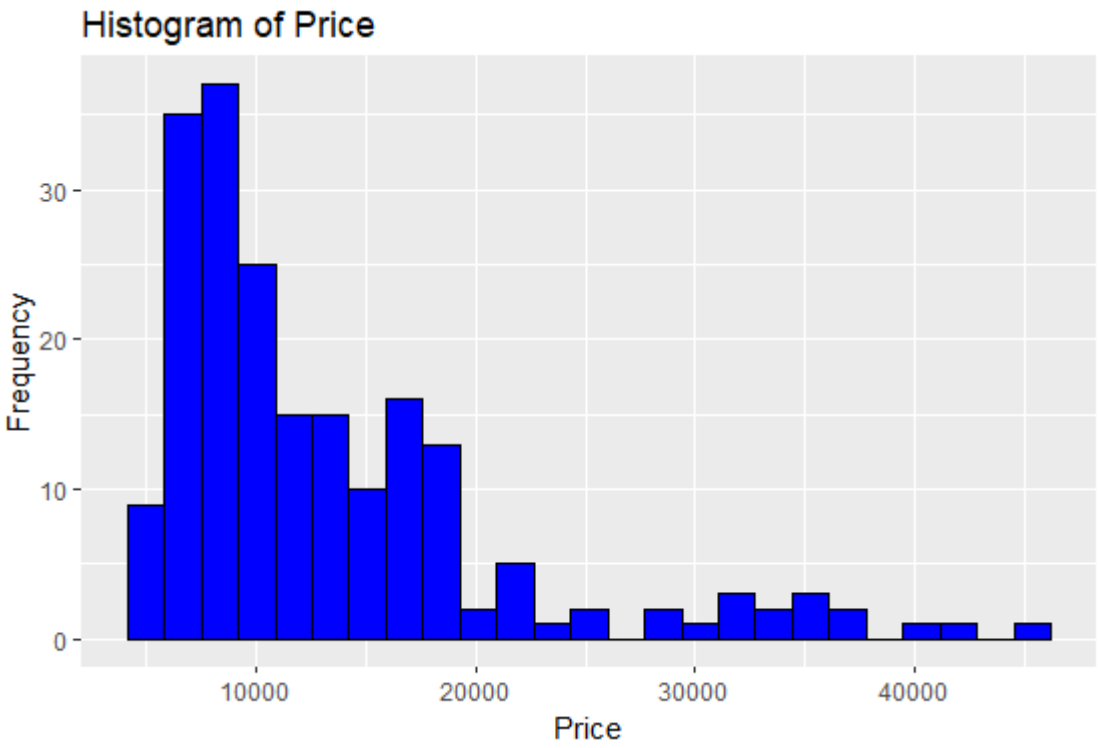
**Task 2**

**1) RESULT:**

```
carData=read.csv("Car_data.csv", na.strings = c("?"))
carDataCleaned <- carData[!is.na(carData$price), ]
```

**2) RESULT:**

```
ggplot(data = carDataCleaned , aes(x = price)) +
geom_histogram(bins = 25, fill = "blue", color = "black") +
labs(x = "Price", y = "Frequency", title = "Histogram of Price")
```




**3) RESULT:**


We can conclude:

- **Price and curb.weight** correlated positevly, we assume that they have linear dependency.
- **Price and engine.size** correlated positevly, we assume that they have linear dependency.
- **Price and horsepower** correlated positevly, we assume that they have exponential or linear dependency. Requires further investigation.
- **Price and highway.mpg** correlated negatively, we assume they have reciprocal dependency.

**4) RESULT:**

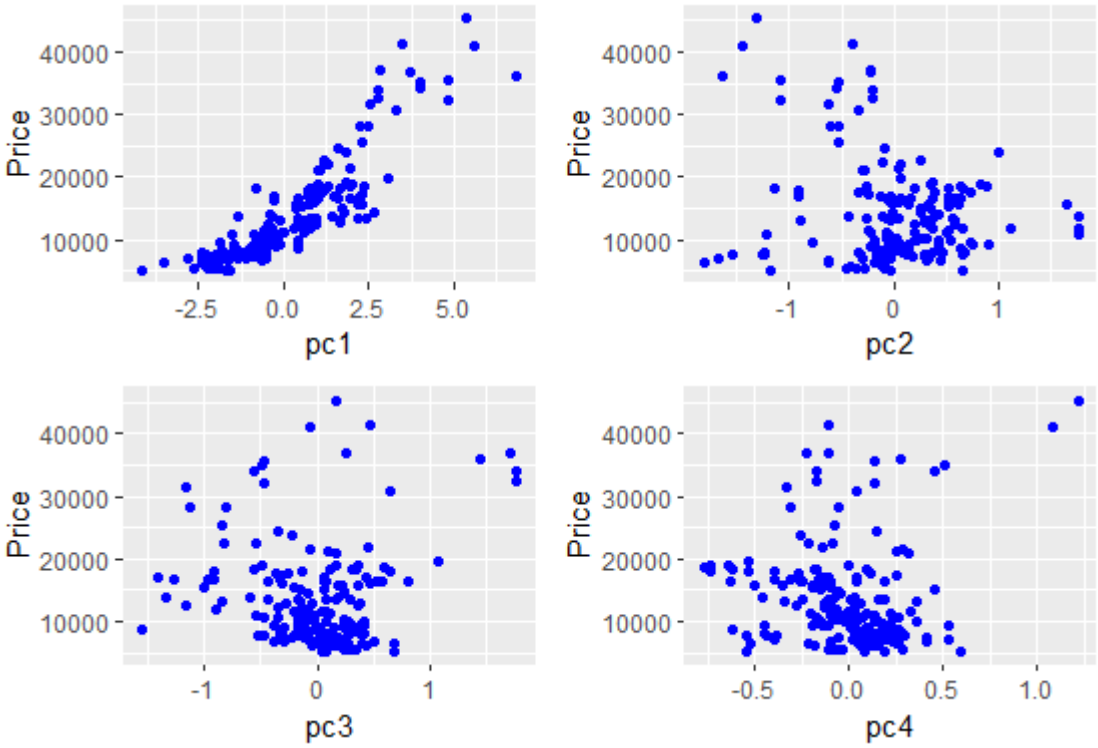
```
arr1 = c("curb.weight", "engine.size",
         "horsepower", "highway.mpg", "price")
arr2 = c("curb.weight", "engine.size", "horsepower", "highway.mpg")
pca_data <- na.omit(subset(carDataCleaned, select = arr1))
pca_dataCleaned <- subset(pca_data, select = arr2)
pca_result <- prcomp(pca_dataCleaned, scale. = TRUE)
pca_result$rotation
```

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| **curb.weight** | 0.5087483 | -0.3614817 | 0.4857817 | 0.5255770 |
| **engine.size** | 0.5009844 | -0.5764444 | -0.1685910 | -0.5784960 |
| **horsepower** | 0.5026465 | 0.2863146 | -0.7415238 | 0.4561544 |
| | | | | |

| | | | | |
|---|---|---|---|---|
| **highway.mpg** | -0.4873769 | -0.6745864 | -0.4309707 | 0.4254813 |

- PC1 becomes significantly larger with increase of **curb.weight**, **engine.size**, **horsepower** and with decrease of **highway.mpg**.
- PC2 becomes significantly larger with decrease of **highway.mpg** and **engine.size** and increase of **horsepower**
- PC3 becomes significantly larger with decrease of **horsepower** and increase of **curb.weight**

**5) Result:**




- PC1 correlates positevly and negatively with the same variables as price do, hence we can assume linear dependency
- PC2 & PC3 & PC4 do not have the same sign of correlation as price have with other variables, so there is no strong correlation between them and price.