Analysis Stack Overflow survey 2022

Team: Peniaha Nazarii, Shamanskyi Kyrylo, Burak Vasyl.

The main aim of this project is to examine, analyze and then visualize open data set from stack Overflow developer survey 2022. In order to search the data we will be using a dataset provided by Stack Overflow for information about survey. Also, data set link: https://insights.stackoverflow.com/survey.

```
require(BSDA)
```

```
## Loading required package: BSDA
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
library(BSDA)
require(EnvStats)
```

```
## Loading required package: EnvStats
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
library(EnvStats)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library("fitdistrplus")
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
## The following object is masked from 'package:EnvStats':
##
##     boxcox
```

```
## Loading required package: survival
```

```
library("dplyr")
library("nortest")
```

Information technology. The main question is why people want to work in IT and why this area is so popular. The answer is simple - it is a high salary. So we decided why not to analyze pay as our data.

Reading the data from csv file.

```
# Below in path you should input your path to csv file
path <- "/Users/igor/Desktop/P&SR/Project/stack-overflow-developer-survey-2022/survey_results_public.csv"
data <- read.csv(path)
```

Hypothesis

Parsing the data for Hypothesis 1 and 2

```
year_salary <- data$ConvertedCompYearly
country <- data$Country
country_salary_na <- data.frame(country, year_salary)
country_salary <- na.omit(country_salary_na)
```

1. We will compare annual salary in Ukraine and India

\(H_0: \mu_0 = \mu_1\) vs \(H_1: \mu_0>\mu_1\)

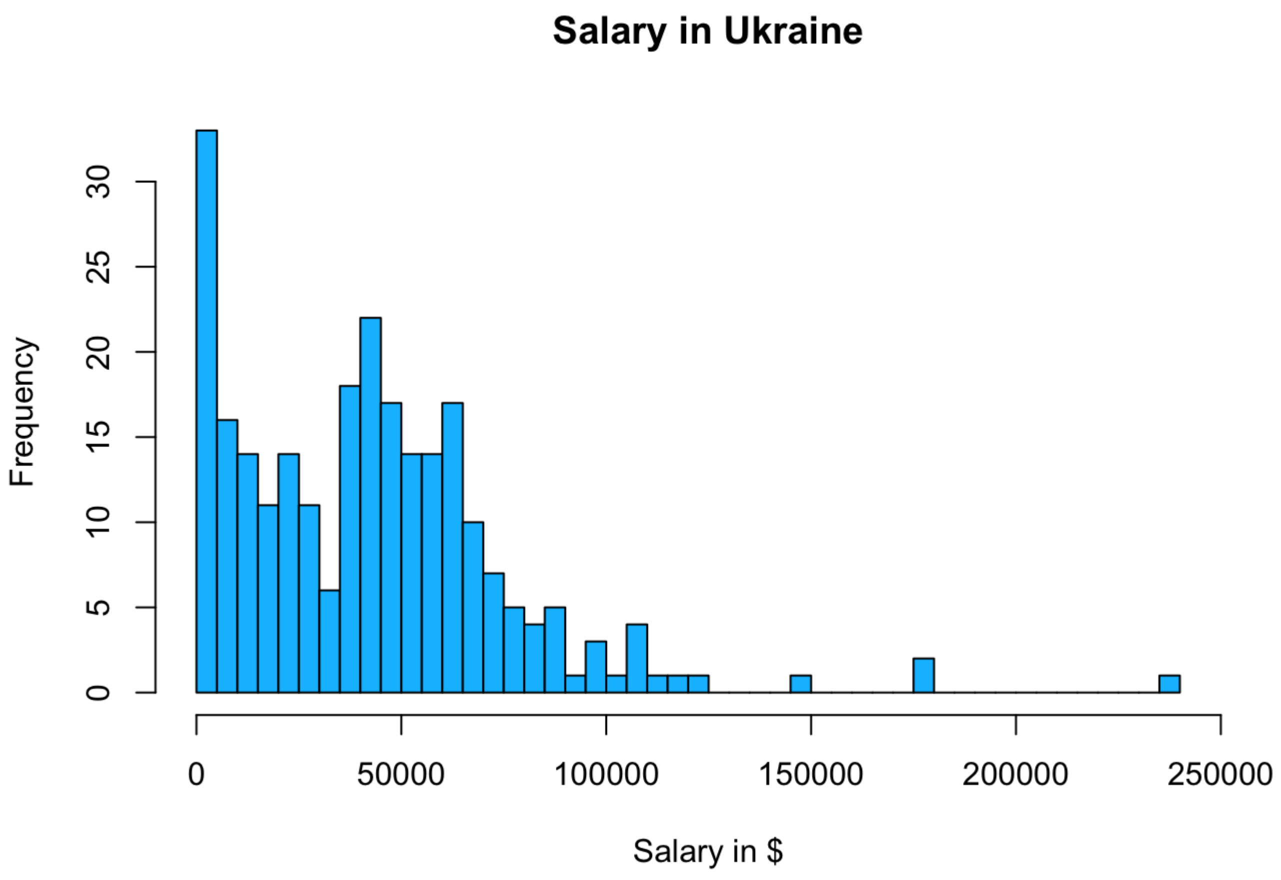
\(\mu_0\) - Ukraine; \(\mu_1\) - India.

```
ukraine_data <- subset(country_salary, country == "Ukraine" & year_salary < 10^6)
india_data <- subset(country_salary, country == "India" & year_salary < 10^6)
t.test(x = ukraine_data$year_salary, y = india_data$year_salary, alternative = "greater")
```

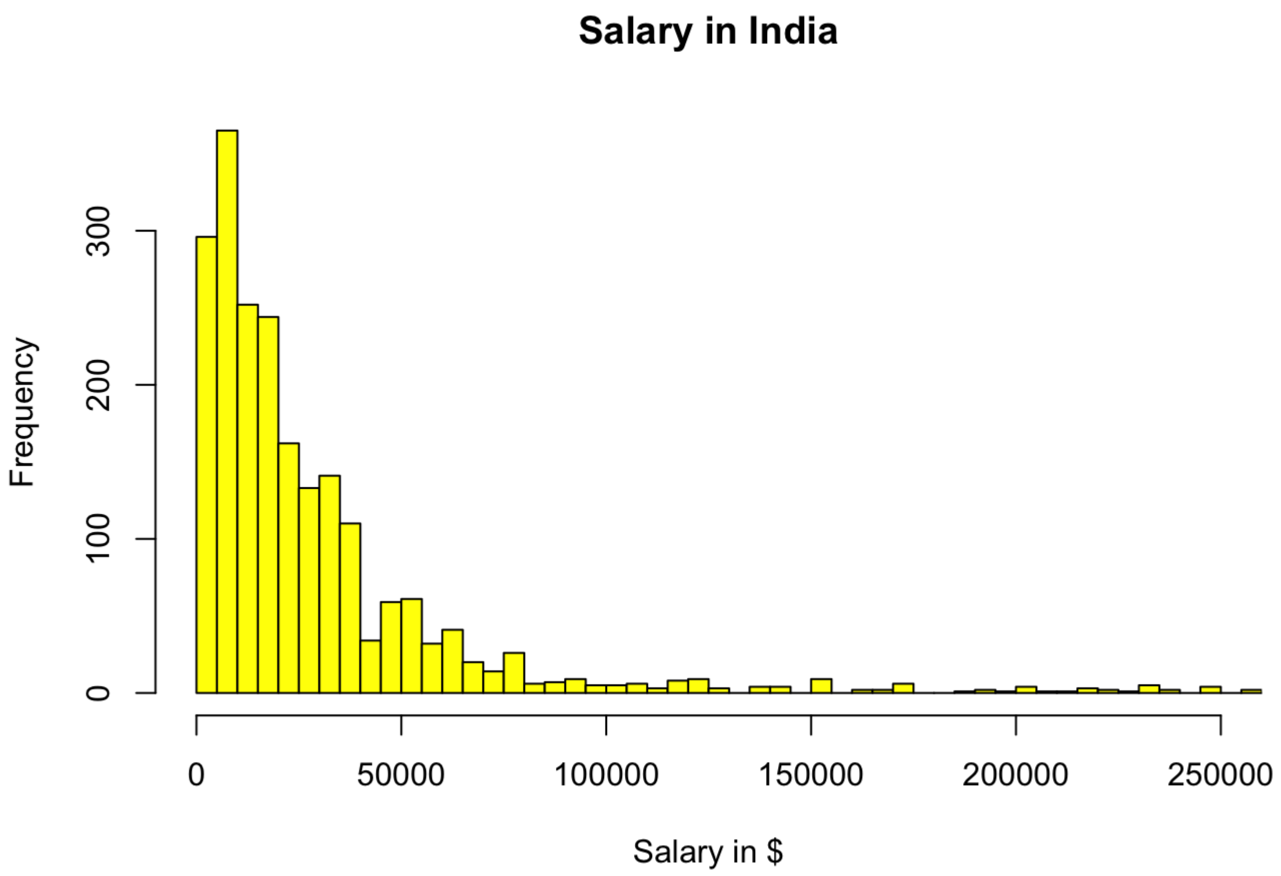
```
## 
##  Welch Two Sample t-test
## 
## data:  ukraine_data$year_salary and india_data$year_salary
## t = 1.016, df = 688.34, p-value = 0.155
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1740.824       Inf
## sample estimates:
## mean of x mean of y 
##  42251.80  39449.26
```

Here we can see the result of testing our hypothesis. From summary, we see that average salary in Ukraine is greater than in India but we should not reject null hypothesis because p-value = 0.155 and is greater than \(\alpha\)(which is equal to 0.05, in our case) Below, you can also see the histograms of two samples, Ukrainian and Indian salaries.

```
num <- 35
range <- c(0, 250000)
hist(ukraine_data$year_salary, xlim = range, breaks = num, col = "deepskyblue", main = "Salary in Ukraine", xlab = "Salary in $")
```



```
hist(india_data$year_salary, xlim = range, breaks = 4*num, col = "yellow", main = "Salary i
n India", xlab = "Salary in $")
```



2. Here we will compare annual salary in USA and Switzerland

\(H_0: \mu_0 = \mu_1\) vs \(H_1: \mu_0>\mu_1\)

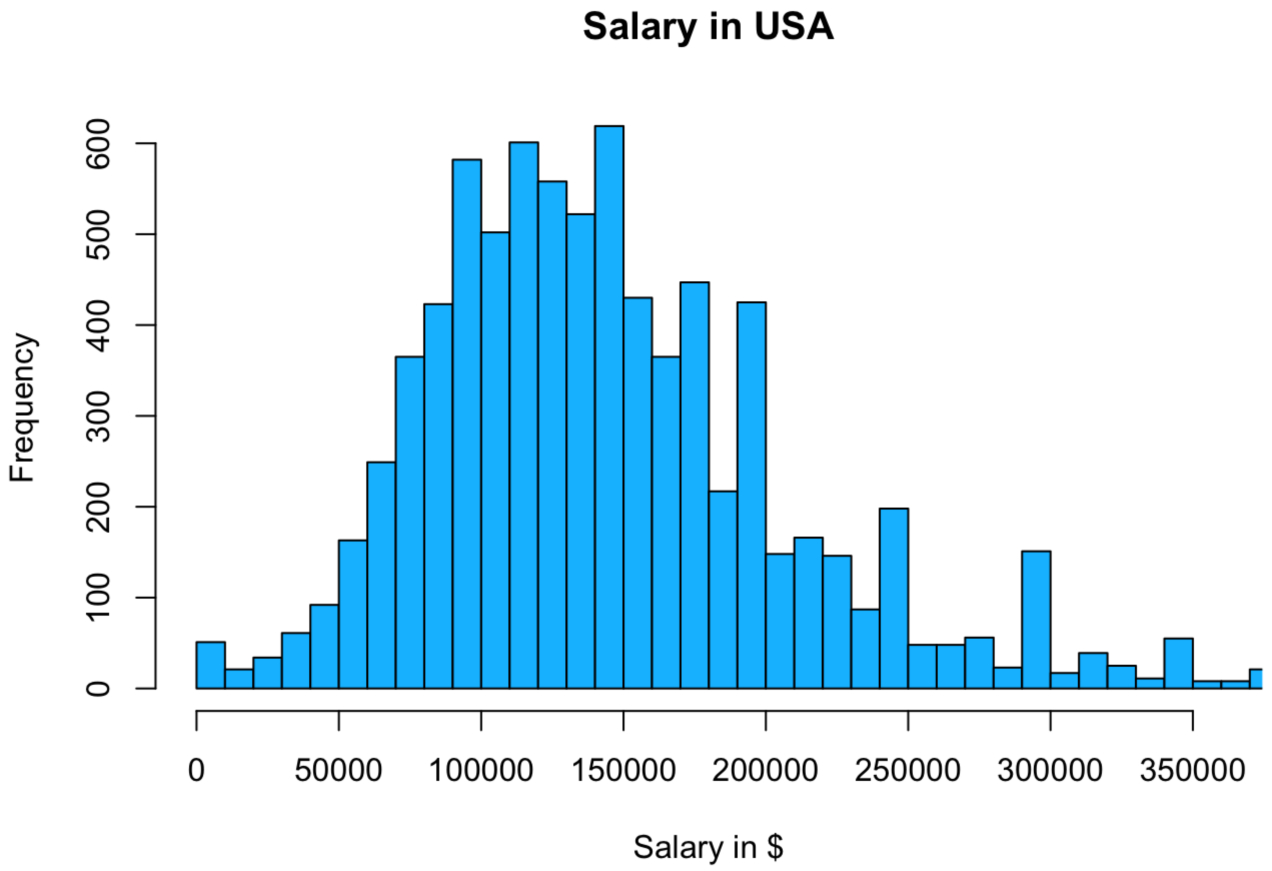
\(\mu_0\) - USA; \(\mu_1\) - Switzerland.

```
usa_data <- subset(country_salary, country == "United States of America" & year_salary <10^
6)
switzerland_data <- subset(country_salary, country == "Switzerland" & year_salary < 10^6)
t.test(x = usa_data$year_salary, y = switzerland_data$year_salary, alternative = "greater")
```

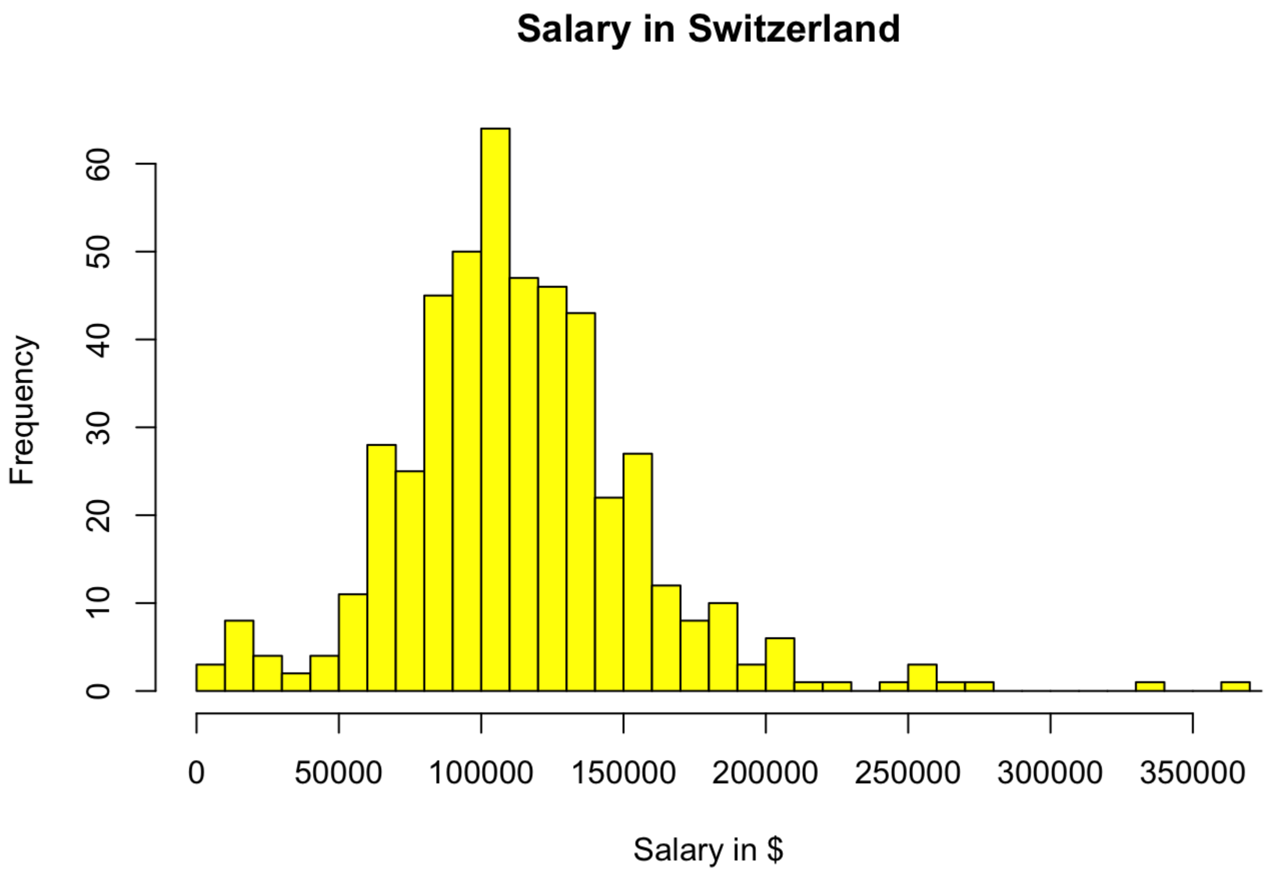
```
## 
##  Welch Two Sample t-test
## 
## data:  usa_data$year_salary and switzerland_data$year_salary
## t = 9.0882, df = 560.95, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  30948.35      Inf
## sample estimates:
## mean of x mean of y 
##  159909.6  122108.4
```

From the result of t.test, we see that difference in means of two samples is meaningfull. Moreover, p-value is significantly small and = 2.2 * 10^(-16). So we should reject our null hypothesis(because it is <\(\alpha\)). Below you can see the histograms of annual salary distribution of USA and Switzerland.

```
num <- 140
range <- c(0, 360000)
hist(usa_data$year_salary, xlim = range, breaks = num, col = "deepskyblue", main = "Salary
in USA", xlab = "Salary in $")
```



```
hist(switzerland_data$year_salary, xlim = range, breaks = num, col = "yellow", main = "Sala
ry in Switzerland", xlab = "Salary in $")
```



3. Average salary of people who have ever used Python vs C++.

\(\mu_0\) - Average salary of people who ever used Python.

\(\mu_1\) - Average salary of people who ever used C++.

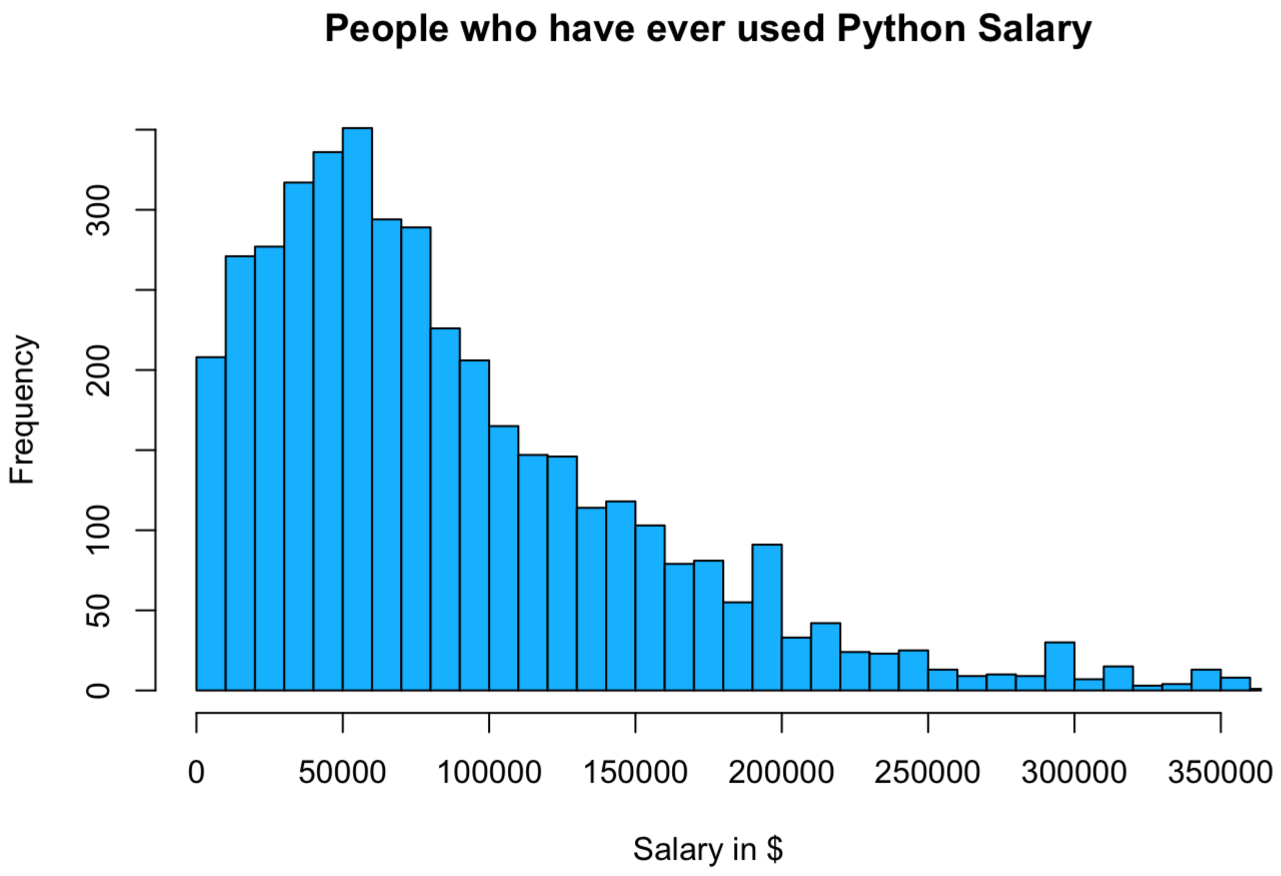
\(H_0: \mu_0 = \mu_1\).

\(H_1: \mu_0 > \mu_1\).

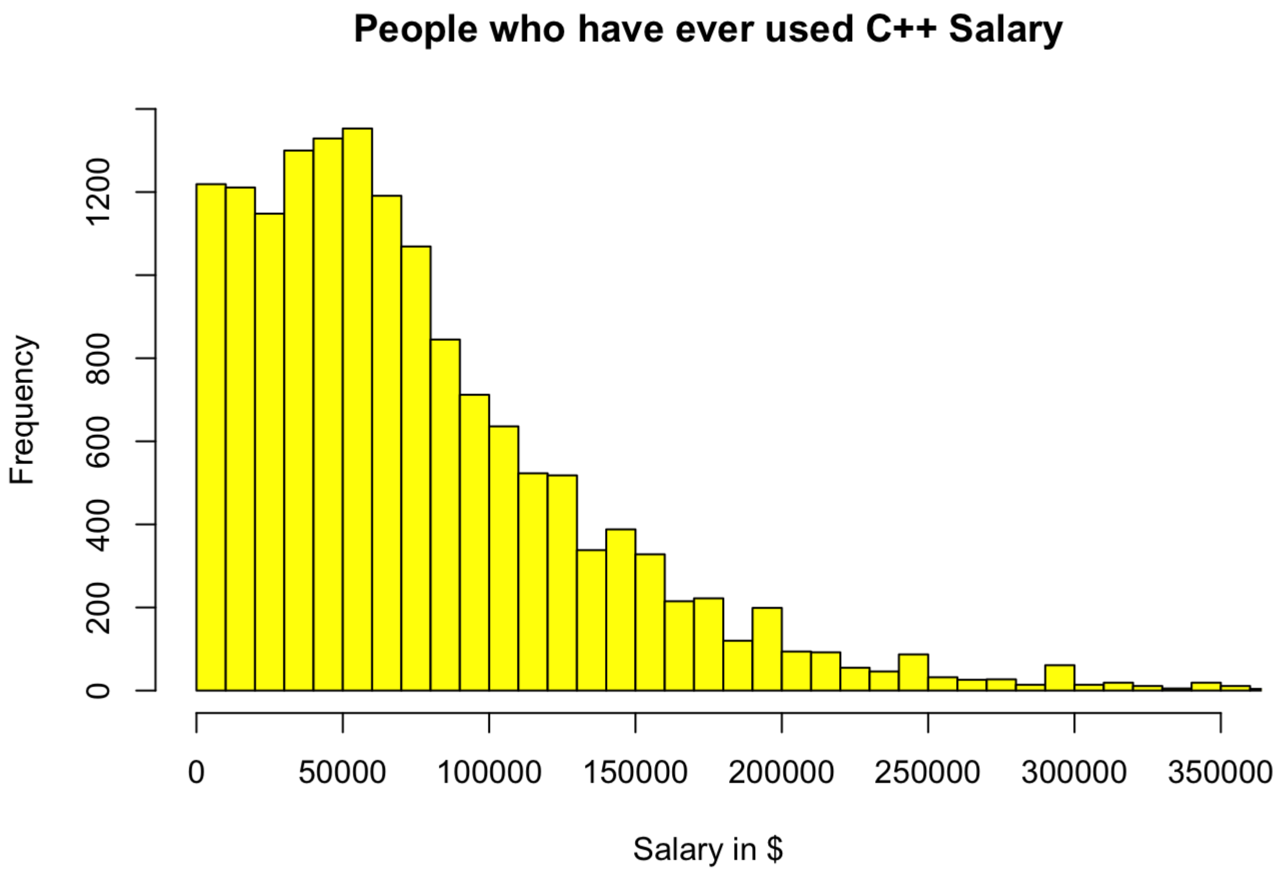
```
python_data <- subset(data, str_detect(LanguageHaveWorkedWith, "Python") & !(str_detect(LanguageHaveWorkedWith, "C++")) & ConvertedCompYearly > 0, select=c(LanguageHaveWorkedWith, ConvertedCompYearly))
c_data <- subset(data, str_detect(LanguageHaveWorkedWith, "C++") & !(str_detect(LanguageHaveWorkedWith, "Python")) & ConvertedCompYearly > 0, select=c(LanguageHaveWorkedWith, ConvertedCompYearly))
t.test(x = python_data$ConvertedCompYearly, y = c_data$ConvertedCompYearly, alternative = "greater")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  python_data$ConvertedCompYearly and c_data$ConvertedCompYearly
## t = 3.2805, df = 5334.8, p-value = 0.0005214
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  24416.65      Inf
## sample estimates:
## mean of x mean of y 
##  199750.6  150771.5
```

```
hist(python_data$ConvertedCompYearly, xlim = c(0, 350000), breaks = 9*320, main = "People w
ho have ever used Python Salary", xlab="Salary in $", col = "deepskyblue")
```



```
hist(c_data$ConvertedCompYearly, xlim = c(0, 350000), breaks = 9*320, main = "People who ha
ve ever used C++ Salary", xlab="Salary in $", col = "yellow")
```



As we can see from result of t-test, that compare salary Python and C++ developers, we should reject null hypothesis, because p-value = 0.0005214 is less than \(a\)(0.05 in our case). Also you can observe histograms of two samples.

4. Average salary Men vs Women.

\(\mu_0\) - Average salary of Men.

\(\mu_1\) - Average salary of Women.

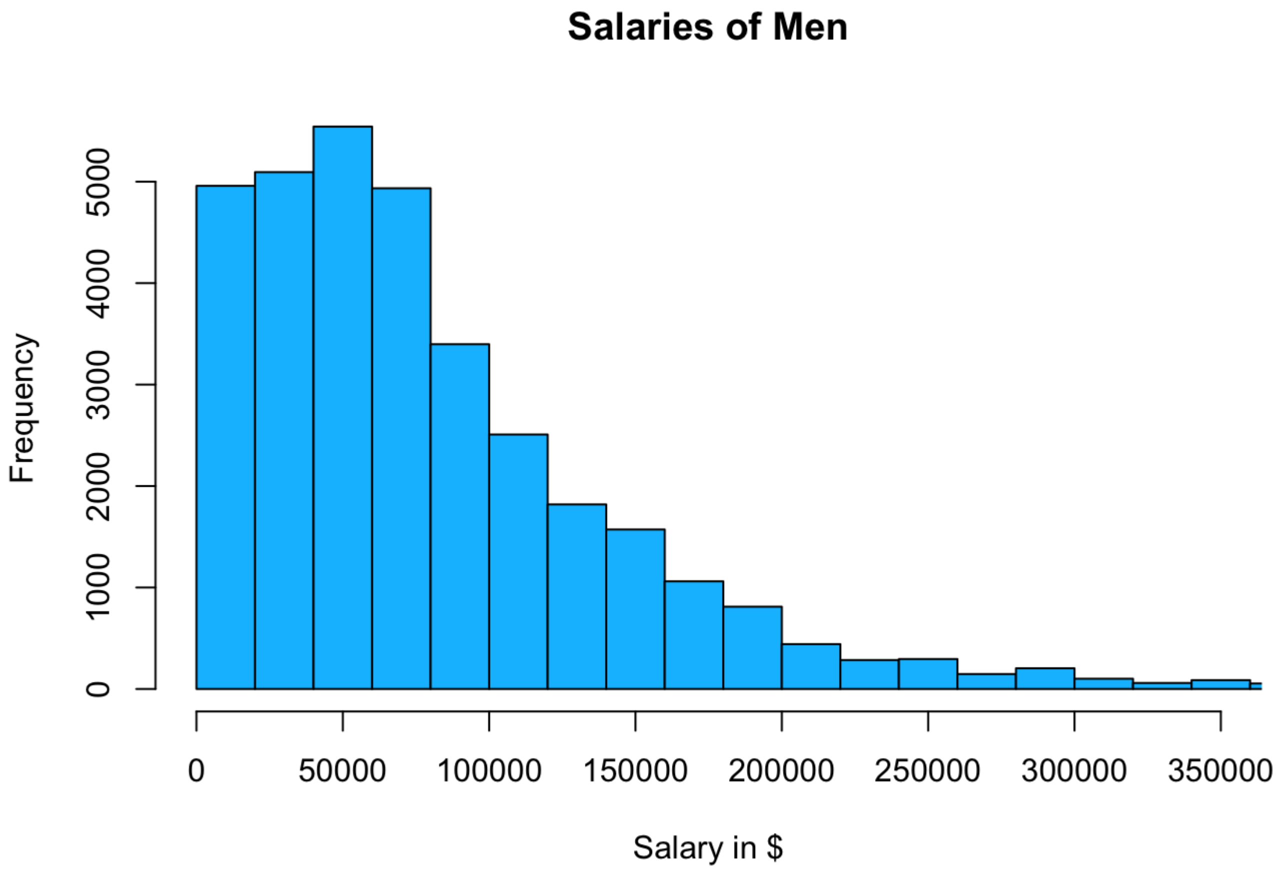
\(H_0: \mu_0 = \mu_1\).

\(H_1: \mu_0 > \mu_1\).

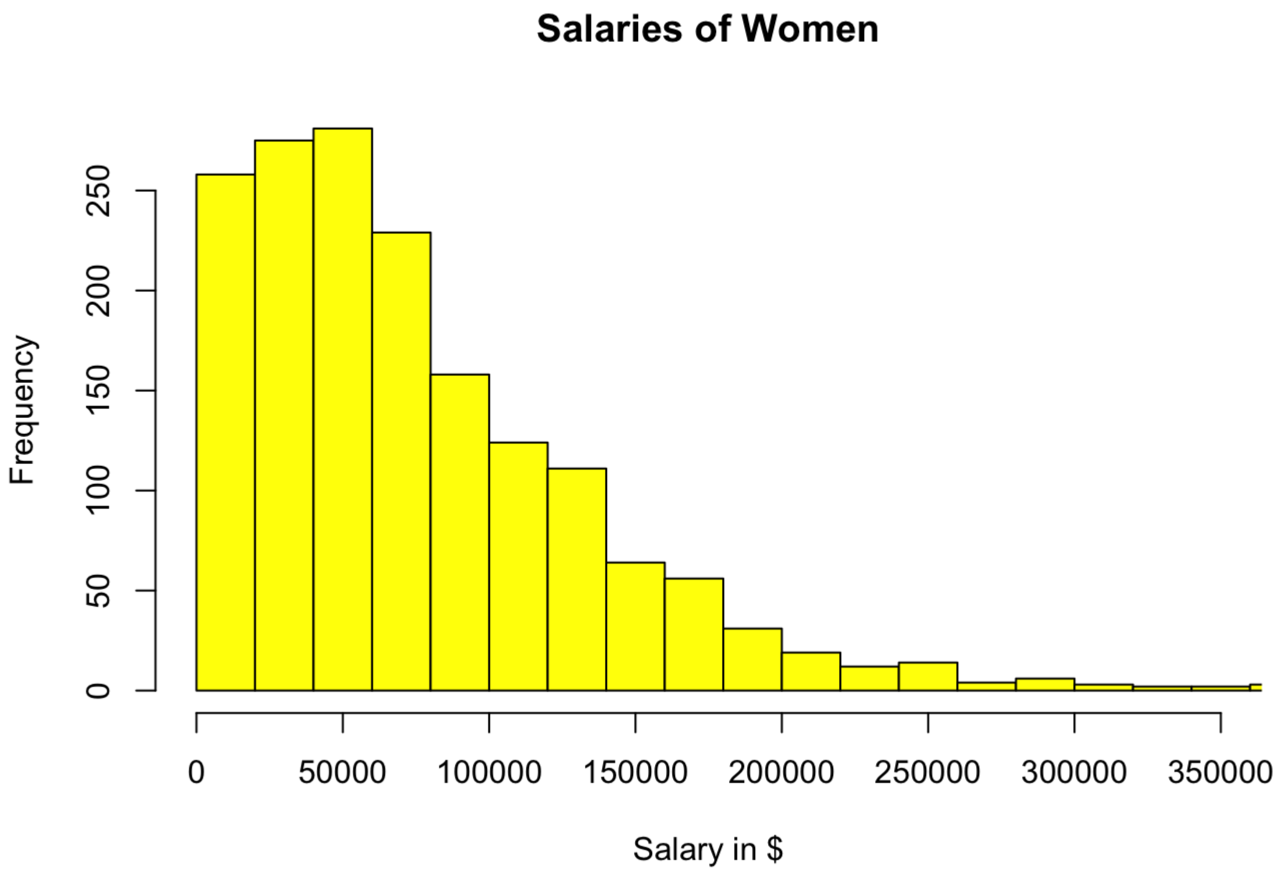
```
man_data <- subset(data, Gender == "Man" & ConvertedCompYearly > 0, select=c(Gender, ConvertedCompYearly))
woman_data <- subset(data, Gender == "Woman" & ConvertedCompYearly > 0, select=c(Gender, ConvertedCompYearly))
t.test(x = man_data$ConvertedCompYearly, y = woman_data$ConvertedCompYearly, alternative = "greater")
```

```
## 
##  Welch Two Sample t-test
## 
## data:  man_data$ConvertedCompYearly and woman_data$ConvertedCompYearly
## t = 0.53428, df = 1774.5, p-value = 0.2966
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -31240.91       Inf
## sample estimates:
## mean of x mean of y 
##  169166.2  154148.1
```

```
hist(man_data$ConvertedCompYearly, xlim = c(0, 350000), breaks = 5*320, main = "Salaries of
 Men", xlab="Salary in $", col = "deepskyblue")
```



```
hist(woman_data$ConvertedCompYearly, xlim = c(0, 350000), breaks = 9*320, main = "Salaries
of Women", xlab="Salary in $", col = "yellow")
```



As we can see from result of t-test, that compare salary of Men and Women developers, we shouldn`t reject null hypothesis, because p-value = 0.2966 is greater than \(a\)(0.05 in our case). Also you can observe histograms of two samples.

5. Fitting Distribution

Fitting distribution is the first thing that comes to mind when I think about data research.

Given that we don’t know anything at all at this point, we can only assume that our distribution is normal. After all, it is obvious that there are very few small and large salaries, and what is closer to the expected value is more.

For testing normality of our distribution I want to use Shapiro Wilk test, as it is the most powerful test when testing for a normal distribution. It has been developed specifically for the normal distribution and it cannot be used for testing against other distributions like for example the KS test. But in R this test have limit to sample size): So i decide to use Anderson-Darling normality test.

Our \(H_0:\) The data follows the normal distribution, \(H_1:\) The data do not follow the normal distribution

```
salary <- as.numeric(na.omit(data$ConvertedCompYearly))
```

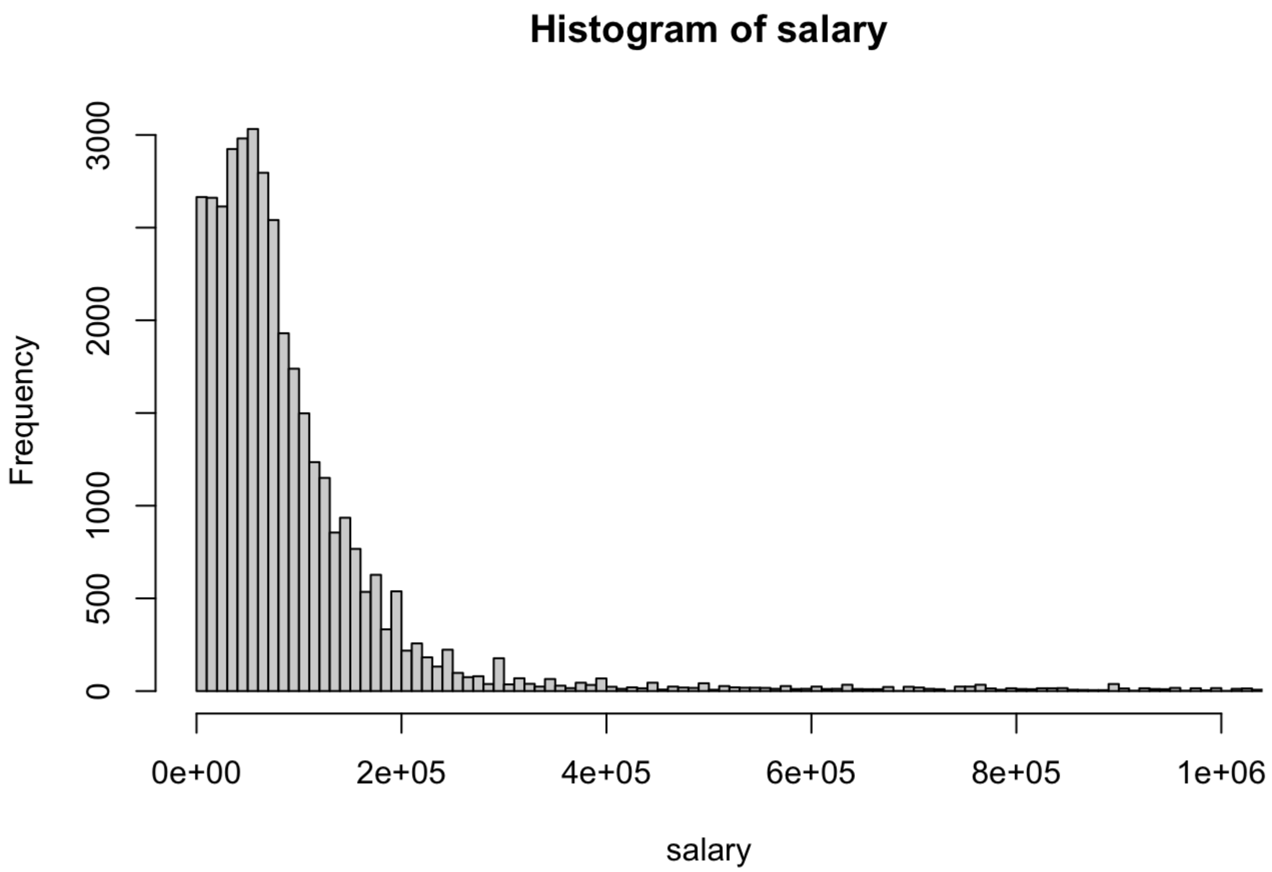
```
ad.test(salary)
```

```
##
##  Anderson-Darling normality test
##
## data:  salary
## A = 10764, p-value < 2.2e-16
```

The P-value of our test is small enough to reject the null hypothesis. We consider that data hasn’t normal distribution

The only way to make new hypothesis is visualize our data

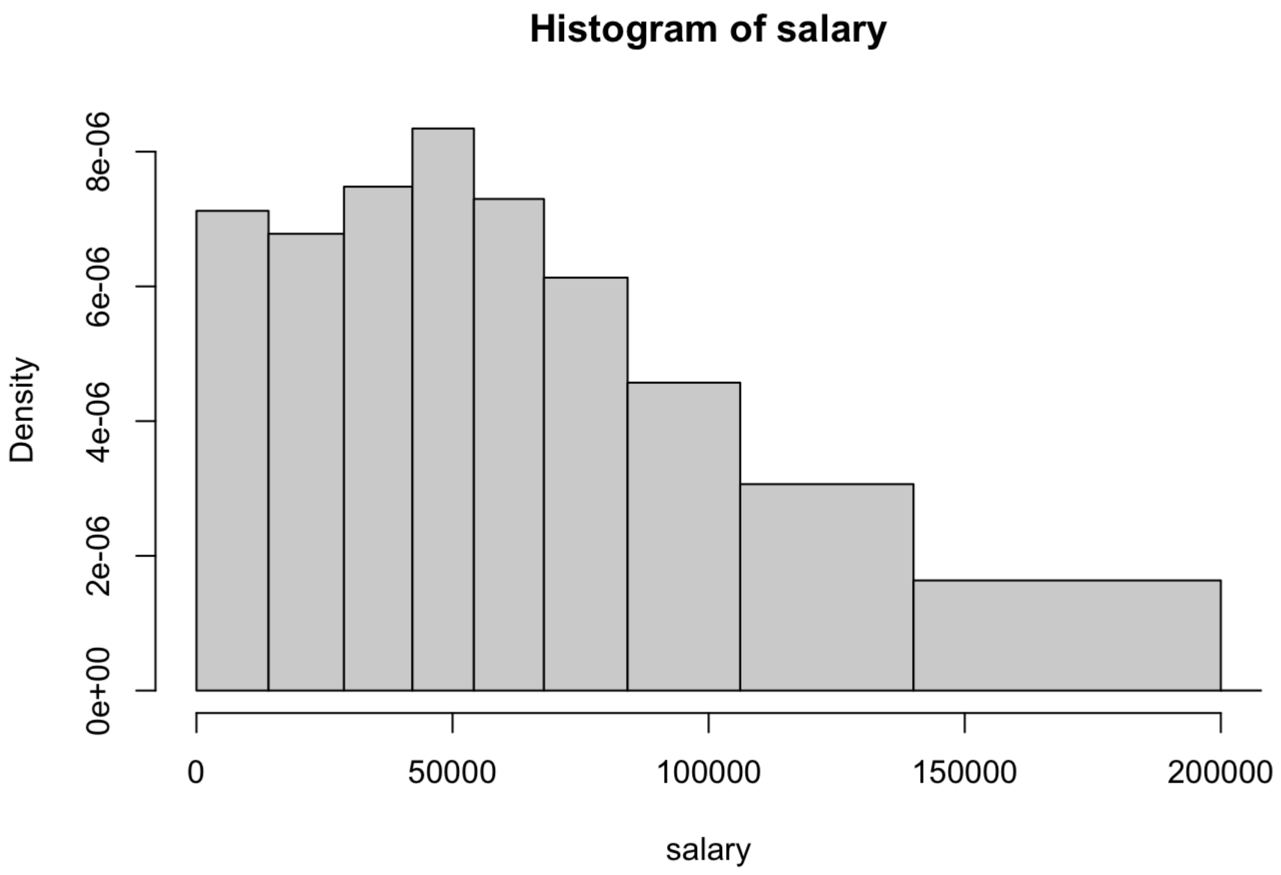
```
hist(salary, xlim = c(0, 10^6), breaks = 7*10^3)
```



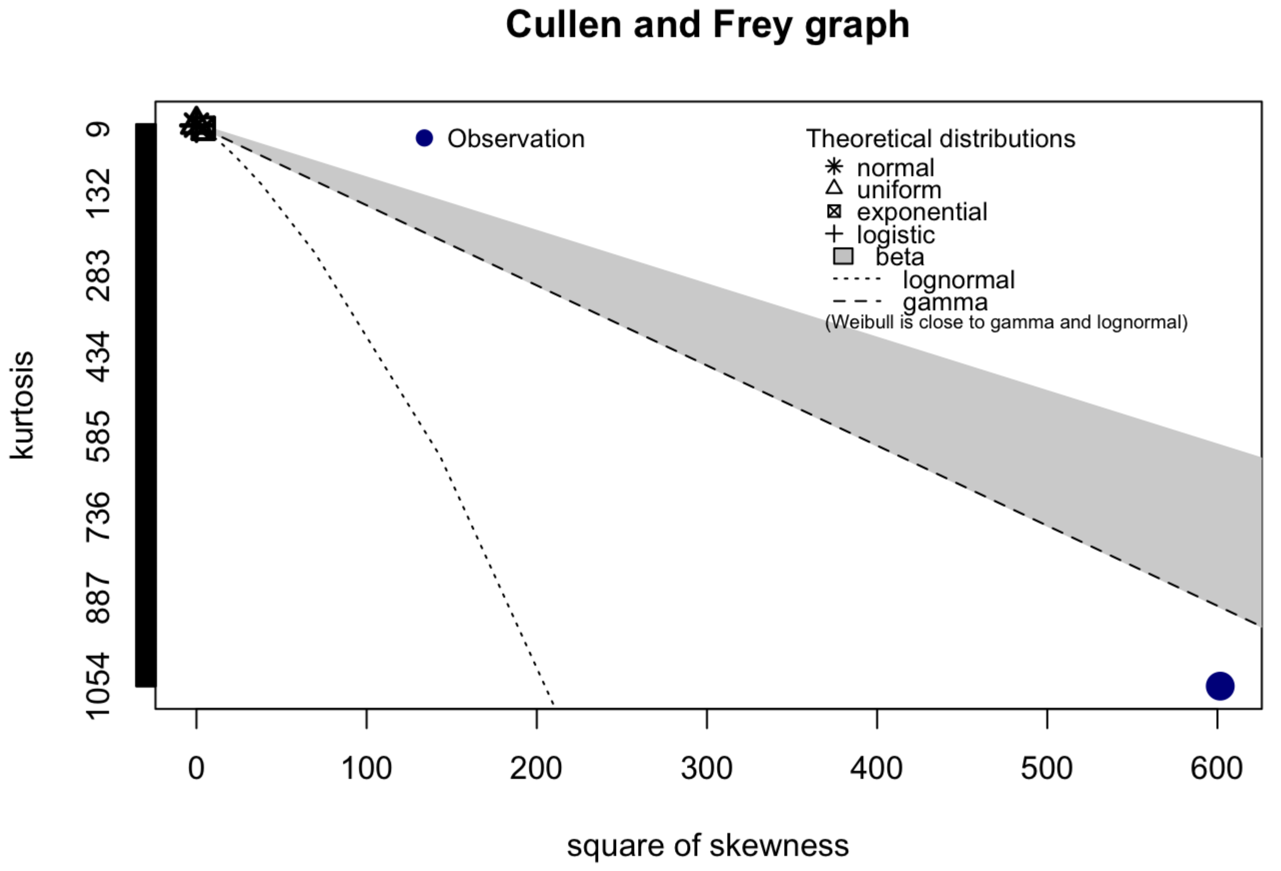
As we can see span of our data is huge. The only way to make better visualization it’s cut the data.To make live easier I use Cullen and Frey graph.

```
q <-seq(0, 1, .1)
quan <- as.numeric(quantile(salary, q, type = 1))

hist(salary,
     breaks = quan,
     xlim = c(quan[1], quan[10]))
```



```
descdist(salary)
```



```
## summary statistics
## ------
## min:  1   max:  5e+07 
## median:  67845 
## mean:  170761.3 
## estimated sd:  781413.2 
## estimated skewness:  24.52871 
## estimated kurtosis:  1053.408
```

Observation is far away from all distribution. I don’t know what distribution it can be. So I decide to use Kolmogorov - Smirnov test to find the best-fit distribution. As we can see, the closeset distribution to our observation is gamma, lognormal, weibull. Our hypothesis will look like: \(H_0\): the two distributions are identical \(H_1\):the two distributions are different

```
#mean and sd
mu <- mean(salary)
sd <- sd(salary)
#test for normal
ks.test(salary, 'pnorm')$p.value
```

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

```
#test for gamma
parameters <- egamma(salary)$parameters
ks.test(salary, 'pgamma', parameters[1], parameters[2])$p.value
```

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

```
#test for weibull
parameters <- eweibull(salary, method = "mle")$parameters
ks.test(salary, 'pweibull', parameters[1], parameters[2])$p.value
```

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

```
#test for lognormal
ks.test(salary, 'plnorm')$p.value
```

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

Our p-value in each test is 0, so we should reject all null hypothesis. Therefore, we need new distribution in P&S, which would fit to our data.

Thanks for watching our analysis.