Quantitative Analysis of Cash Withdrawals

Pre-selection assignment for ING Bank Śląski

Mateusz Dadej

28th of April 2019

Abstract

The Subject of herein analysis is modelling and risk analysis of cash withdrawals from one of the branches of ING Bank Śląski from the beggining of 2018 to end of march 2019. The analysis is mostly of quantitative character as many of the methods used, are present in statistical or mathematical textbooks. Although, every quantitative analysis later on is interpreted in a qualitative way. This analysis was conducted for pre-selection to Lion's Den Risk Modelling Challenge organized by ING Bank Śląski & ING Tech Poland.

Contents

Data set overview

The data set, separated originally by semicolon, consists of 455 observations of cash withdrawals from one of ING's branches. Each with 3 variables: date of the day `[DD.MM.YYYY]`, working day of the branch (categorical: YES or NO) and amount of cash withdrawals made by clients. These variables were given a following names **date, working, withdrawals**. Table below shows a table of first 6 observations from data set:

Table 1: First 6 observations

date	working	withdrawals
2018-01-01	NO	NA
2018-01-02	YES	241.88
2018-01-03	YES	168.55
2018-01-04	YES	116.79
2018-01-05	YES	105.72
2018-01-06	NO	NA

As the example above shows, the data set is not free from `NAs`. These would be a normal values had it not been for the branch being closed. Ergo, if we drop the observations with `NAs` the working column will consists only `YES` values.

Var1	Freq
NO	0
working	0
YES	315

```
df.na <- drop_na(df)

table(df.na$working)
```

For the sake of simplicity of code, later on we will analyse data sets with `NAs` which is `df`, as well as without `- df.na`.

Descriptive statistics

First task of the assignment is following:

> *Calculate: descriptive statistics: arithmetic mean, standard deviation, quartiles, skewness coefficient, kurtosis*

calculations of descriptive statistics below concerns values of withdrawals.

- Average: 233
- median: 211.69
- standard deviation: 152
- quartiles:
  - zeroth (minimum value) 105.72
  - first 162.71
  - second (median) 211.69
  - third 265.14
  - fourth (maximum value) 1855.42
- skewness coefficient: 7
- kurtosis: 74

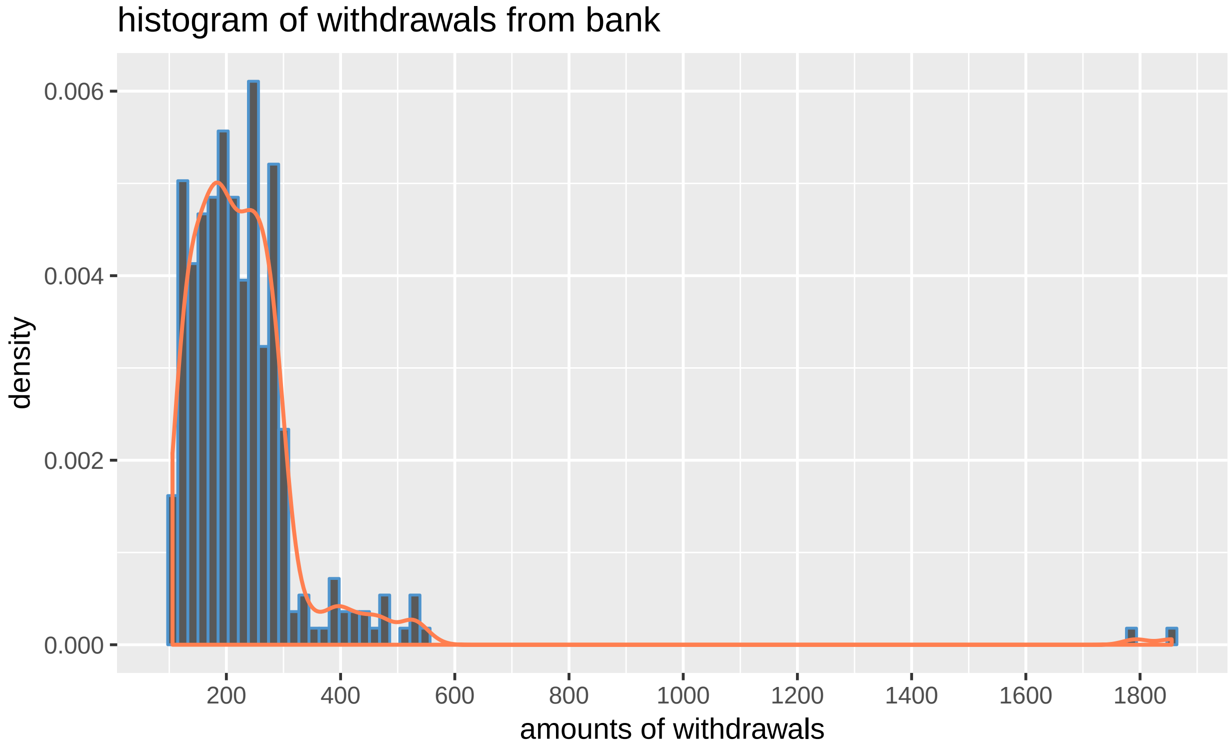
One might already notice potential outliers in data set. There is a substantial difference between maximum value and third quartile. These outliers could potentially influence standard deviation as it is also susceptible to outliers. We will leave it for now, as it is the subject of further task. We define every value in case they will be important later on.

histogram of withdrawals

Another task from assignment is to make a histogram of withdrawals.

> *Calculate histogram of withdrawals*

We will use `ggplot2` package here and later on for data visualization. Our histogram have 100 bins and additional density line.



It is now clearly visible that our data set consists of at least two outliers located around 1 800 000 zł. For the time being, we can state that non outliers are amounts of withdrawals under 600 000 zł. At the first sight, the distribution of withdrawals most likely does not follow the normal distribution. We will test normality hypothesis later.

outliers in the dataset

Next request is related to outliers which we spotted during previous task.

> *Calculate outliers. What could be the cause of increased demand for cash and subsequent withdrawals?*

Following our earlier remark about the range of non-outliers we shall check observations above our limit.

Table 3: Outliers

date	working	withdrawals
2018-12-21	YES	1855.42
2018-12-24	YES	1792.48

```r
filter(df.na, withdrawals > 600)
```

As one may suppose, the outliers are due to extreme seasonal events during **christmas times**. What is also likely is unique desynchronization between working days of bank branch and main Christmas events. Table below shows corresponding days of outliers.

```r
filter(df, date > "2018-12-17" & date < "2018-12-29")
```

Table 4: Outliers with their corresponding days of the month

date	working	withdrawals
2018-12-18	YES	265.00
2018-12-19	YES	263.25
2018-12-20	YES	257.52
2018-12-21	YES	1855.42
2018-12-22	NO	NA
2018-12-23	NO	NA
2018-12-24	YES	1792.48
2018-12-25	NO	NA
2018-12-26	NO	NA
2018-12-27	YES	254.92
2018-12-28	YES	264.63

December 21 is the last day with open branch of the bank (22th and 23th is closed) to withdraw cash for before-Christmas shopping during the weekend, when the Saturday is also exceptionally not restricted for shopping. It is commonly known that, people tend to postpone such a trivial activities like withdrawing cash from banks or ATM's until the very last moment.

Of course not every client want to go for a shopping during a weekend (or forget to withdraw cash), therefore there is also another day with working bank branch. December 24 is the last opportunity when it is possible to withdraw cash before Christmas (25th and 26th). These rationale, mostly explain existence of outliers in data set.

Does withdrawals follow Gaussian distribution?

> *Verify if the data on withdrawals follows normal distribution using statistical tests.*

As density plot shown earlier may point, it is more likely than not, that the distribution is not of normal distribution. Although, we should test it to be certain. We will use `shapiro.test()` to perform Shapiro Wilk test and make a following hypothesis

$$H_0 : X_w \sim \mathsf{N}(\mu, \sigma^2)$$

and first hypothesis that it does not follow normal distribution. $H_1 : H_0$ is wrong

The test produced following results:

```
##
##  Shapiro-Wilk normality test
##
## data:  df.na$withdrawals
## W = 0.47347, p-value < 2.2e-16
```

`p-value < 0.05` ergo, we can reject H_0 in favor of alternative hypothesis H_1. I.e Distribution of the withdrawals is not normal.

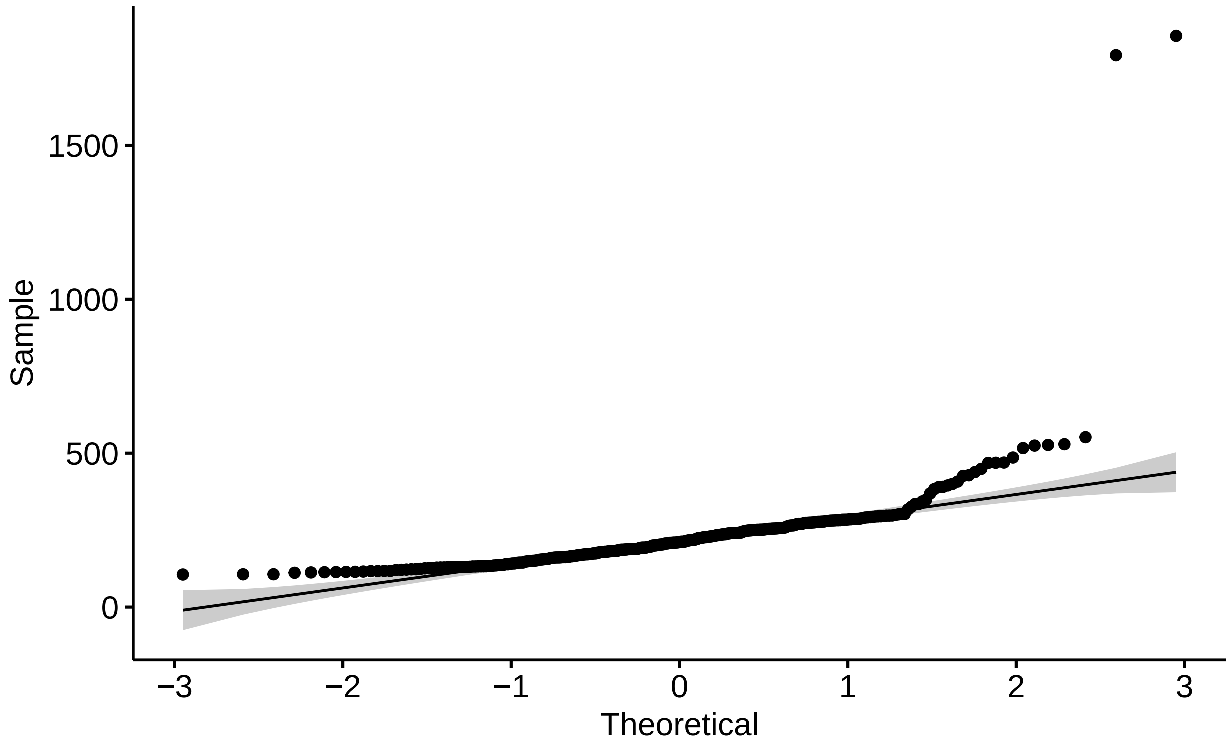
We should redo our statistical test on data set free from outliers.

```
##
##  Shapiro-Wilk normality test
##
## data:  df.na.out$withdrawals
## W = 0.89594, p-value = 8.093e-14
```

The inference on this data set is the same. Both are not of normal distribution.

For the sake of certainty and to analyse the way our data deviates from normal distribution, we can make a quantile - quantile plot to visualize the difference between actual distribution and Gaussian.

```
ggqqplot(df.na$withdrawals) # from ggpubr package
```



As we can see, although the outliers may significantly influence our inference, the withdrawals are still not close to follow normal distribution.

Day of the week impact on amount withdrawals

Following task was given to do with the use of previously done calculations :

> *Assess the impact of the* ***day of the week*** *on the size of withdrawals. Is there a relationship between these variables and how can these dependencies be justified?*

The easiest way to analyse the impact of the day of the week is to look at the averages of withdrawals for every day of the week. We will use popular `dplyr` package for this.

```r
mutate(df.na, week.day = weekdays(date))%>%
  group_by(week.day)%>%
  summarise(average = mean(withdrawals),
            median = median(withdrawals))%>%
  arrange(desc(average))
```

Table 5: Average and median withdrawal per day of the week.

week.day	average	median
piątek	252.0175	217.820
poniedziałek	251.7110	208.640
środa	222.7006	212.710
wtorek	221.0731	209.325
czwartek	215.2565	209.515

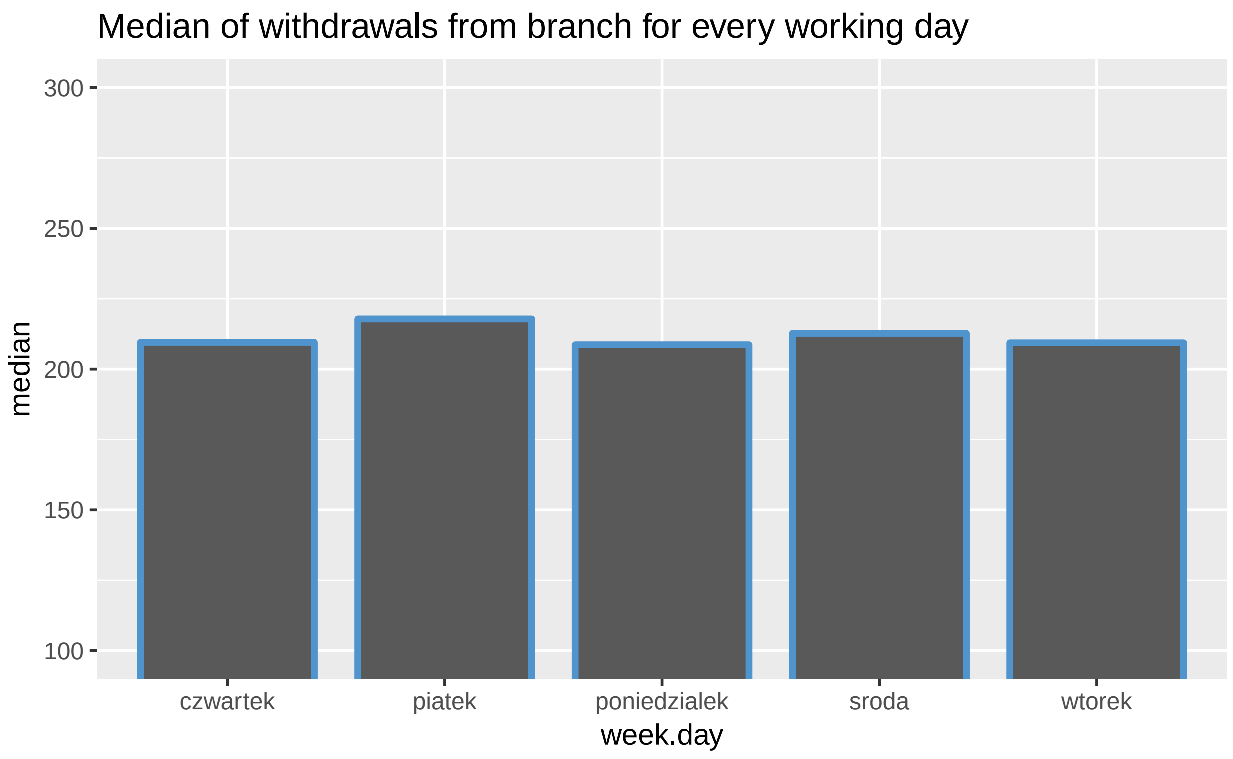
The average is indeed higher for Friday and Monday but its most likely due to the outliers if we look at the median, the difference virtually disappears. Friday is slightly higher than the general median. It might be explained by clients intentions to go for a shopping during upcoming weekend.

Again, we will try to analyse outliers free data set in an analogous way.

Table 6: Same table but data set without outliers

week.day	average	median
piątek	226.9644	215.065
poniedziałek	226.8598	204.385
środa	222.7006	212.710
wtorek	221.0731	209.325
czwartek	215.2565	209.515

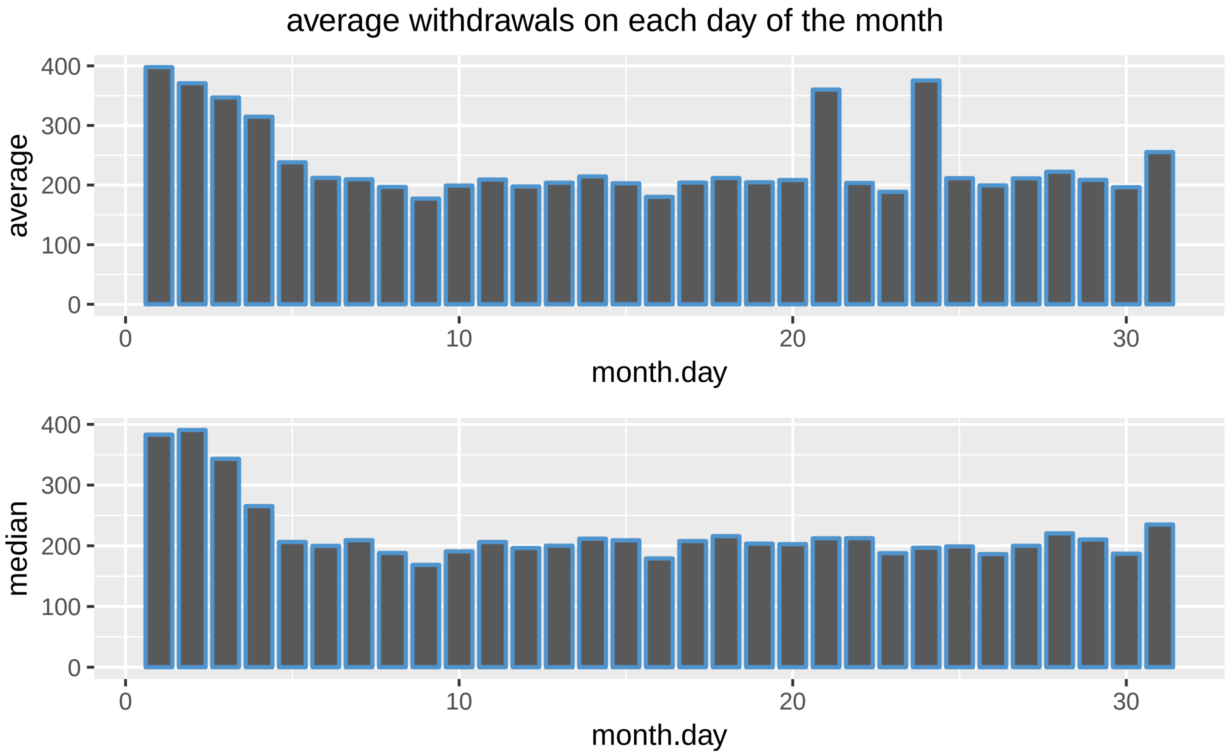
Given the fact that standard deviation of withdrawals from this data set is 83, the difference between withdrawals among different days and general mean is **not significant**.



day of the month impact

> *Assess the impact of the day of the month on the size of withdrawals. Is there a relationship between these variables and how can such dependencies be justified?*

The task above might be performed in an analogous way to the previous one. Although, for the sake of better embracement of the problem, we will limit analysis to graphical visualization.



First days of the month seems to be significantly higher than other days with next 3 days also higher but gradually closer to average. This might be due to wages being paid out to employees with various days at the beginning of each month depending on payday or weather it is a weekend or not. Supposedly, there are still many people living paycheck to paycheck. two bins are in solitude, for there are outliers within them.

Probability of withdrawal within given range

> *Assuming that withdrawals follow normal distribution, what is the probability that withdrawals from a given day are in the range from PLN 220,000 to PLN 250,000? Can this be confirmed by the historical data?*

Although, we already find out that the data of withdrawals does not follow normal distribution, this assumption will still be accepted. However, at the end of the task we will confront it with historical data. The probability can be calculated with a few methods but the author has chosen to compute definite integral of density function, applying the following equation:

$$\int_{r_1}^{r_2} f(x)dx = F(b) - F(a) = P(a < X < b) \tag{1}$$

And the function being integrated is a standard normal density function, such that:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, x \in R \tag{2}$$

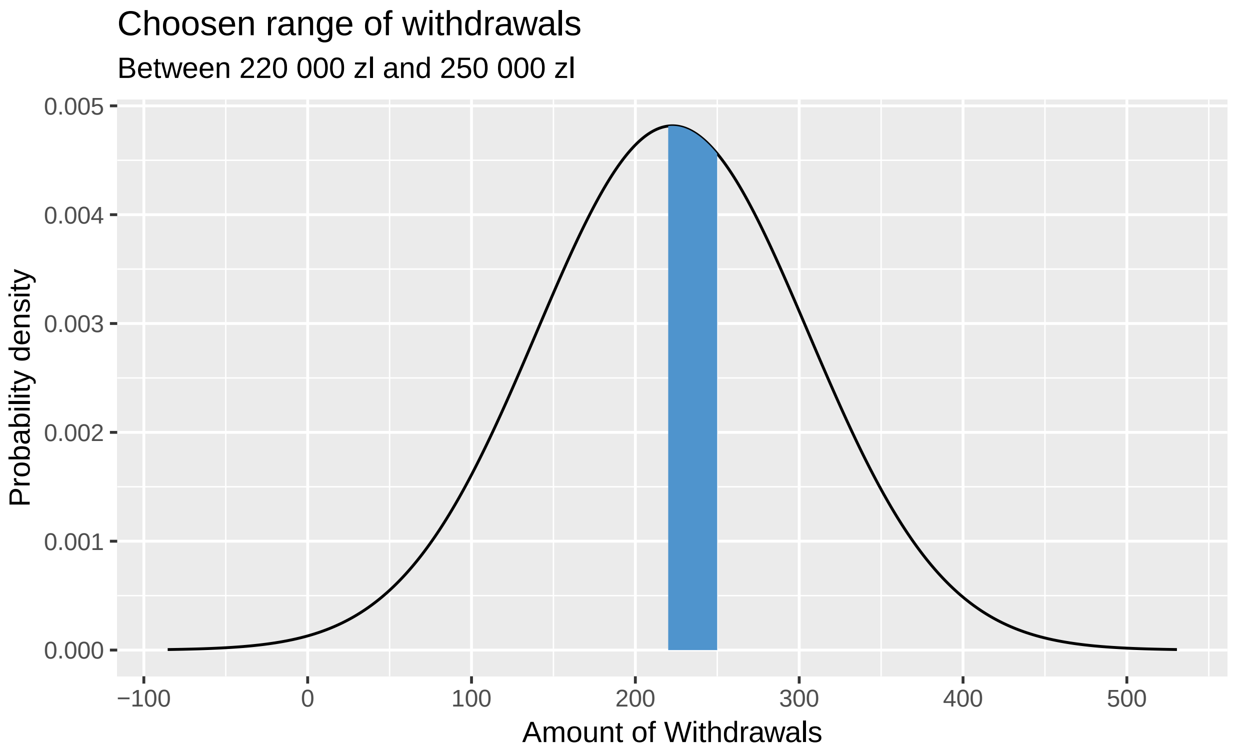
Where r_1 is lower limit of integration (lower range i.e. 220 k zł) and r_2 is upper limit of integration (upper range i.e. 250 k zł). μ is expected value of withdrawals and equals 223, σ is standard deviation of data, which is 83. The rest of letters are constants ($\pi = 3.1416, e = 2.718$).

```
avg.na.out <- mean(df.na.out$withdrawals) # average of withdrawals
sd.na.out <- sd(df.na.out$withdrawals)    # standard deviation of withdrawals

dens.funct <- function(x){dnorm(x = x,    # defined density function of withdrawals
                                mean = avg.na.out,
                                sd = sd.na.out)}

integrate(dens.funct,
          lower =  lower.bound,       # lower limit of integration (lower bound of given range)
          upper =  upper.bound)       # upper limit of integration (upper bound of a given range)
```

```
## 0.1421223 with absolute error < 1.6e-15
```

According to the result above, the likelihood that daily withdrawals will occur in a range of 220k and 250k zł is exactly 14.21%. Again, assuming the normal distribution of withdrawals. We can conclude that, given the range is close to the mean, the probability is rather low. It is due to the standard deviation, which is high.

Graph below shows the area of the probability of withdrawals on density function. Red area is also the product of integral performed earlier.



Now we will calculate how did the probability shape, based on historical data with up to 315 observations. For that, one shall simply divide number of observation in a given range by number of every observations in the data set.

```
nrow(filter(df.na, withdrawals >lower.bound & withdrawals < upper.bound))/nrow(df.na)
```

```
## [1] 0.1301587
```

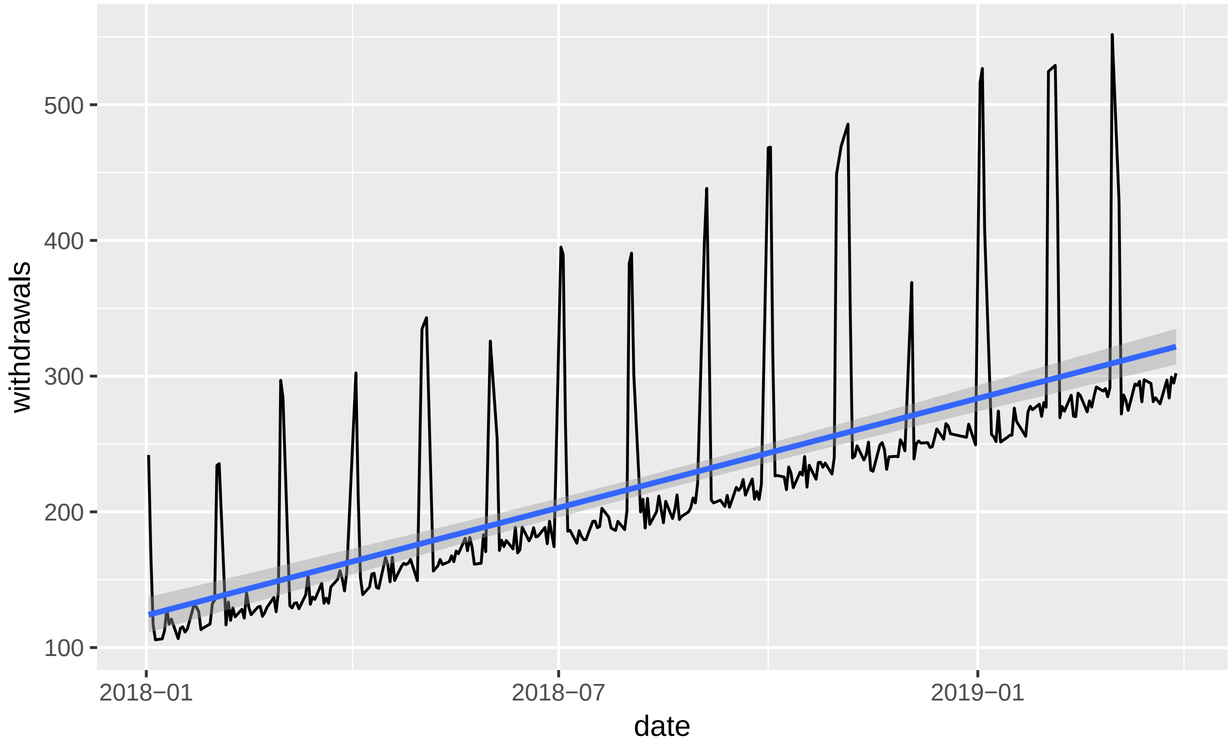
Historically, 13.02% of observed withdrawals were in a range of 220 000 to 250 000 zł. The difference is visible but not really big. It amounts to 1.2 percentage points.

Trend of the Withdrawals

The last task is to analyse time series related to withdrawals. With the emphasis on the occurrence of the trend.

> *Find the long-term trend of the withdrawals and justify the hypothesis that an increasing trend exists.*

A standard preliminary way to define whether the trend is indeed present, is to plot a variable of interest against time during which it had been occurring. Additionally, we will also plot linear function along with its confidence interval.



Increasing trend is clearly visible, as well as seasonality, somehow described earlier. To precisely asses and interpret the trend, we ought to estimate a parameters of linear model. For which, we will use ordinary least square method. The functional form of prespecified mode is following:

$$y_t = \alpha + \beta_1 x_t + \xi_t$$

Where, response variable y_t is amount of withdrawals in a day t, β_1 Coefficient, estimated with OLS, for variable x_t which represents date. And error term xi_t.

Summary below shows statistics of the fitted linear model, shown and described above.

```
##
## Call:
```

```
## lm(formula = df.na.out$withdrawals ~ df.na.out$date)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.11 -24.52 -18.99 -11.97 242.36
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -7.554e+03  4.543e+02  -16.63   <2e-16 ***
## df.na.out$date  4.380e-01  2.558e-02   17.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.52 on 311 degrees of freedom
## Multiple R-squared:  0.4852, Adjusted R-squared:  0.4836
## F-statistic: 293.1 on 1 and 311 DF,  p-value: < 2.2e-16
```

Variables are statistically significant. p-value is way below 0.05 threshold. Thus, the slope coefficient and upward trend is indeed significant. variable coefficient equals 0.438. Therefore, with every day (no matter if its not working) the amount of withdrawal theoretically increases by 437.95 zł.

References

- *Staystyka Matematyczna, M. Sobczyk, C. H. Beck, 2010*
- *Mathematics and Statistics for Financial Risk Management, M. B. Miller, Wiley, 2014*
- *Język R, H. Wickham, G. Grolemund, Helion, 2018*
- *Data set provided by ING Bank Śląski*

packages used

- `rlang`
- `readxl`
- `tidyverse`
- `lubridate`
- `e1071`
- `ggplot2`
- `ggpubr`
- `gridExtra`
- `knitr`