Marketing Analysis

Teodora Veljanoska

For this analysis I chose the data set:

https://www.kaggle.com/datasets/jacouchs/marketing-budget-and-actual-sales-dataset

from Kaggle.

The set contains numeric data for two features:

Company Marketing Budgets (in thousands of dollars) and Company Sales (in millions of dollars).

The columns have been renamed to marketing_budget_thousands and actual_sales_millions.

A. Part One

1. For the feature marketing budgets:

The sample size (n) is 222.

The number of intervals (k) in which the data should be grouped is approximately the square root of the sample size, i.e., √222 = 14.8996644257513, which rounds to 15.

The range of the data (difference between the largest and smallest values) is 396.88 – 101.91 = 294.97.

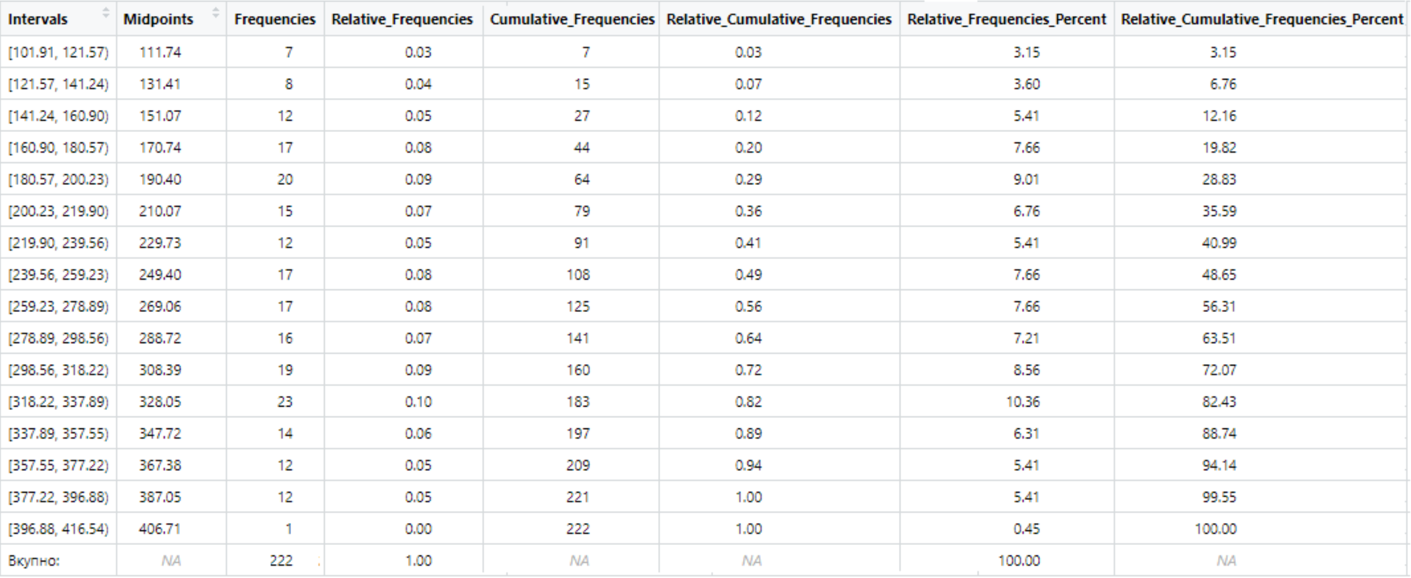
The width of the intervals (w) is: w ≥ R / k, i.e., 294.97 / 15 = 19.66

We extend the upper limit to ensure that values in the last interval are included.

We determine the intervals in R:

```
intervals <- seq(min(marketing_budgets), upper_limit, by = interval_width)
```

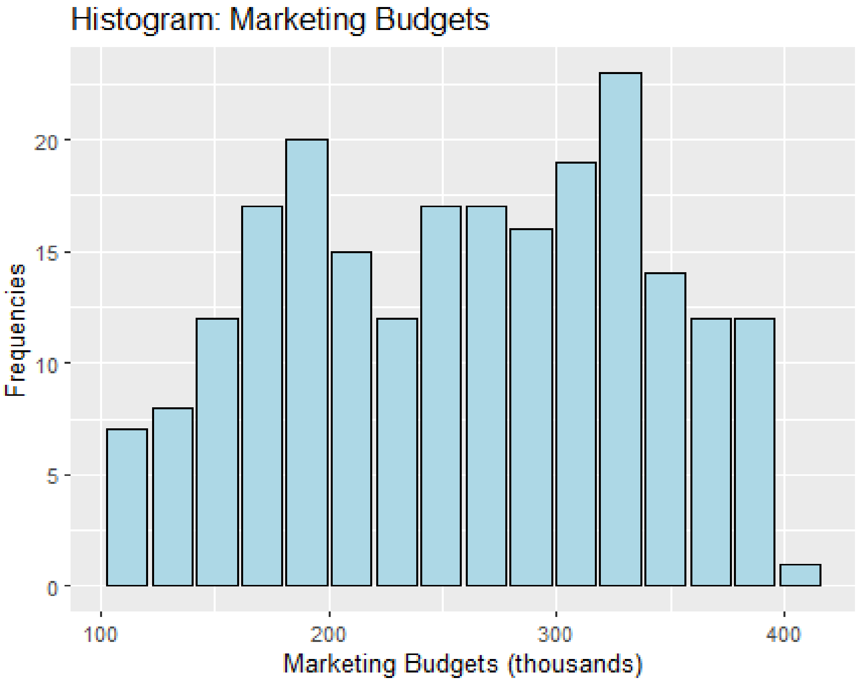
After determining the midpoints and frequencies, we create the final table in R using the data.frame command.



Intervals	Midpoints	Frequencies	Relative_Frequencies	Cumulative_Frequencies	Relative_Cumulative_Frequencies	Relative_Frequencies_Percent	Relative_Cumulative_Frequencies_Percent
[101.91, 121.57)	111.74	7	0.03	7	0.03	3.15	3.15
[121.57, 141.24)	131.41	8	0.04	15	0.07	3.60	6.76
[141.24, 160.90)	151.07	12	0.05	27	0.12	5.41	12.16
[160.90, 180.57)	170.74	17	0.08	44	0.20	7.66	19.82
[180.57, 200.23)	190.40	20	0.09	64	0.29	9.01	28.83
[200.23, 219.90)	210.07	15	0.07	79	0.36	6.76	35.59
[219.90, 239.56)	229.73	12	0.05	91	0.41	5.41	40.99
[239.56, 259.23)	249.40	17	0.08	108	0.49	7.66	48.65
[259.23, 278.89)	269.06	17	0.08	125	0.56	7.66	56.31
[278.89, 298.56)	288.72	16	0.07	141	0.64	7.21	63.51
[298.56, 318.22)	308.39	19	0.09	160	0.72	8.56	72.07
[318.22, 337.89)	328.05	23	0.10	183	0.82	10.36	82.43
[337.89, 357.55)	347.72	14	0.06	197	0.89	6.31	88.74
[357.55, 377.22)	367.38	12	0.05	209	0.94	5.41	94.14
[377.22, 396.88)	387.05	12	0.05	221	1.00	5.41	99.55
[396.88, 416.54)	406.71	1	0.00	222	1.00	0.45	100.00
Вкупно:	NA	222	1.00	NA	NA	100.00	NA

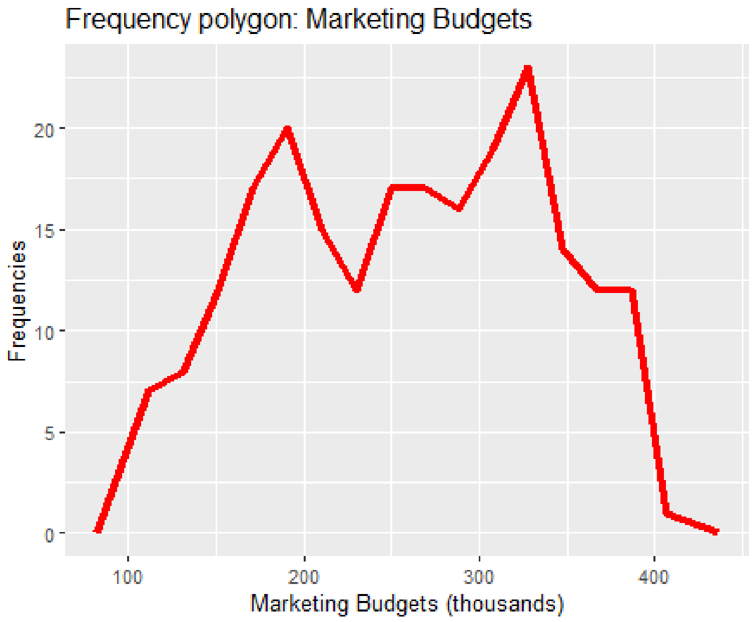
Histogram in R (Marketing Budgets):

```
# Histogram so biblioteka ggplot2
histogram <- ggplot(data.frame(x = hist_data$mids, y = hist_data$counts), aes(x = x, y = y)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +
  labs(title = "Хистограм: Маркетинг Буџети", x = "Маркетинг Буџети (илјадници)", y = "Честоти")
```



Polygon in R (Marketing Budgets):

```
# Presmetki za poligon
polygon_data <- data.frame(
  x = c(hist_data$breaks[1] - interval_width, hist_data$mids, hist_data$breaks[length(hist_data$breaks)] + interval_width),
  y = c(0, hist_data$counts, 0)
)
# Poligon so biblioteka ggplot2
polygon_plot <- ggplot(polygon_data, aes(x = x, y = y)) +
  geom_line(color = "red", size = 1.5) +
  labs(title = "Полигон на честоти: Маркетинг Буџети",
      x = "Маркетинг Буџети (илјадници)", y = "Честоти")
```



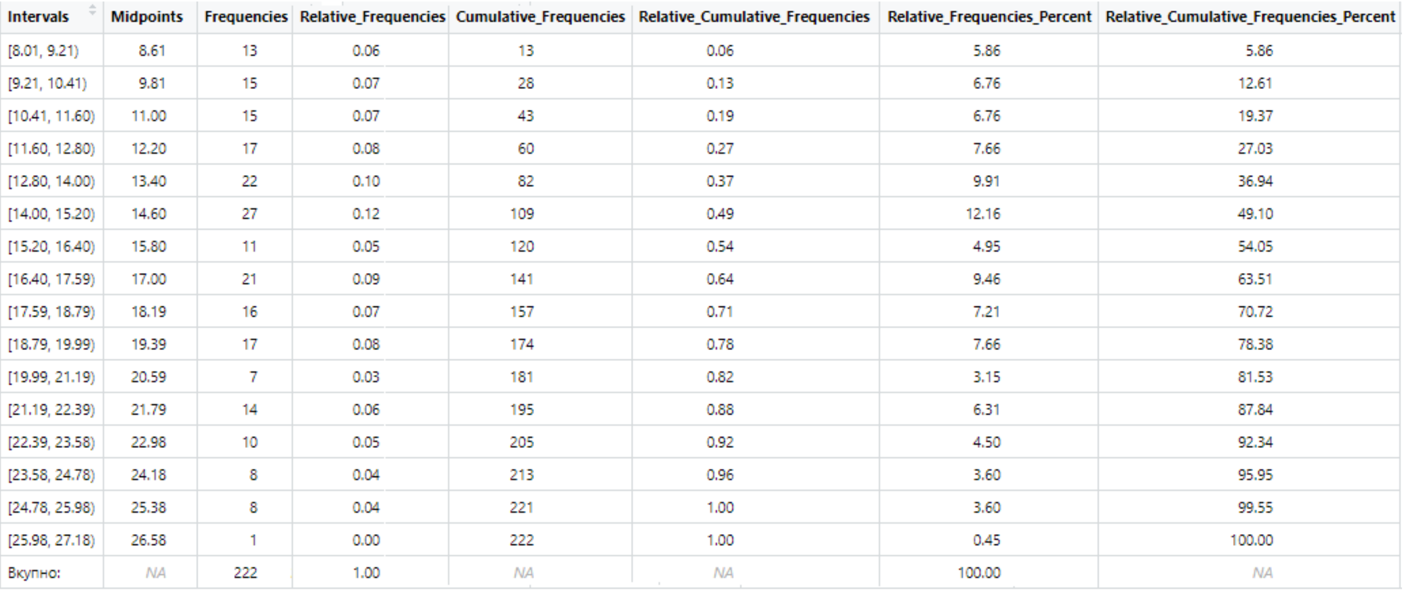
1. **For the feature actual sales:** The sample size (n) is 222.

The number of intervals (k) in which the data should be grouped is approximately the square root of the sample size, i.e., √222 = 14.8996644257513, which rounds to 15.

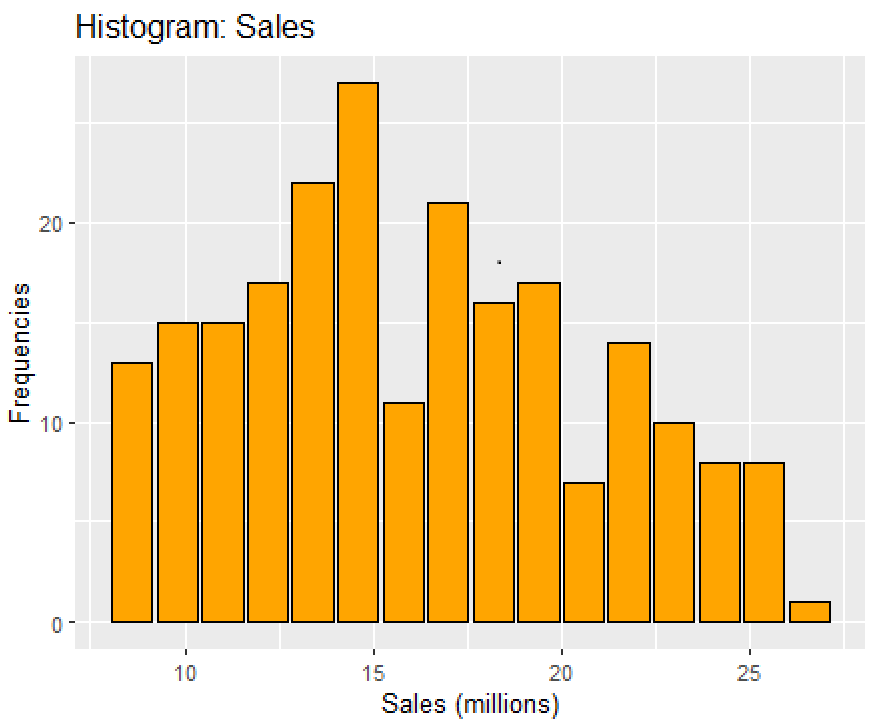
The range of the data is 25.98 – 8.01 = 17.97.

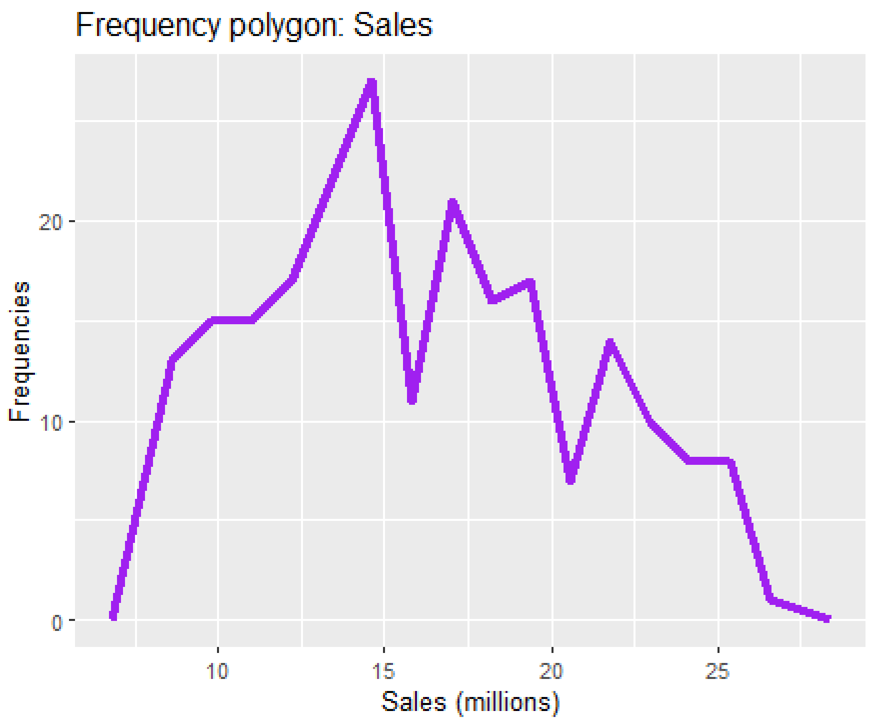
The width of the intervals (w) is: w ≥ R / k, i.e., 17.97 / 15 = 1.198.

We extend the upper limit to ensure that values in the last interval are included.



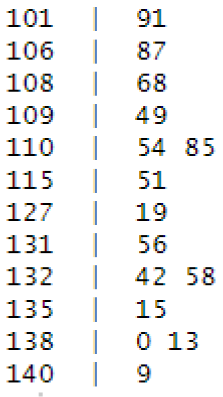
Intervals	Midpoints	Frequencies	Relative_Frequencies	Cumulative_Frequencies	Relative_Cumulative_Frequencies	Relative_Frequencies_Percent	Relative_Cumulative_Frequencies_Percent
[8.01, 9.21)	8.61	13	0.06	13	0.06	5.86	5.86
[9.21, 10.41)	9.81	15	0.07	28	0.13	6.76	12.61
[10.41, 11.60)	11.00	15	0.07	43	0.19	6.76	19.37
[11.60, 12.80)	12.20	17	0.08	60	0.27	7.66	27.03
[12.80, 14.00)	13.40	22	0.10	82	0.37	9.91	36.94
[14.00, 15.20)	14.60	27	0.12	109	0.49	12.16	49.10
[15.20, 16.40)	15.80	11	0.05	120	0.54	4.95	54.05
[16.40, 17.59)	17.00	21	0.09	141	0.64	9.46	63.51
[17.59, 18.79)	18.19	16	0.07	157	0.71	7.21	70.72
[18.79, 19.99)	19.39	17	0.08	174	0.78	7.66	78.38
[19.99, 21.19)	20.59	7	0.03	181	0.82	3.15	81.53
[21.19, 22.39)	21.79	14	0.06	195	0.88	6.31	87.84
[22.39, 23.58)	22.98	10	0.05	205	0.92	4.50	92.34
[23.58, 24.78)	24.18	8	0.04	213	0.96	3.60	95.95
[24.78, 25.98)	25.38	8	0.04	221	1.00	3.60	99.55
[25.98, 27.18)	26.58	1	0.00	222	1.00	0.45	100.00
Вкупно:	NA	222	1.00	NA	NA	100.00	NA


Histogram: Sales


Frequency polygon: Sales

2. For the feature marketing budgets: Steam-and-leaf plot

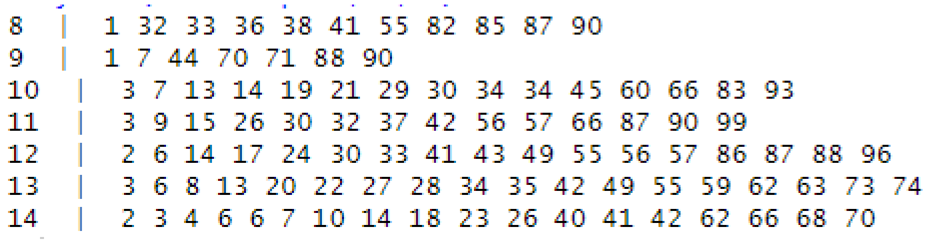
We sort the data using the sort command and create a stem-and-leaf plot for the first 15 elements using the myStem command. The stem represents the integer part of the number (before the decimal point), and the leaf consists of the decimal digits (after the decimal point).



```
101   |   91
106   |   87
108   |   68
109   |   49
110   |   54  85
115   |   51
127   |   19
131   |   56
132   |   42  58
135   |   15
138   |   0  13
140   |   9
```

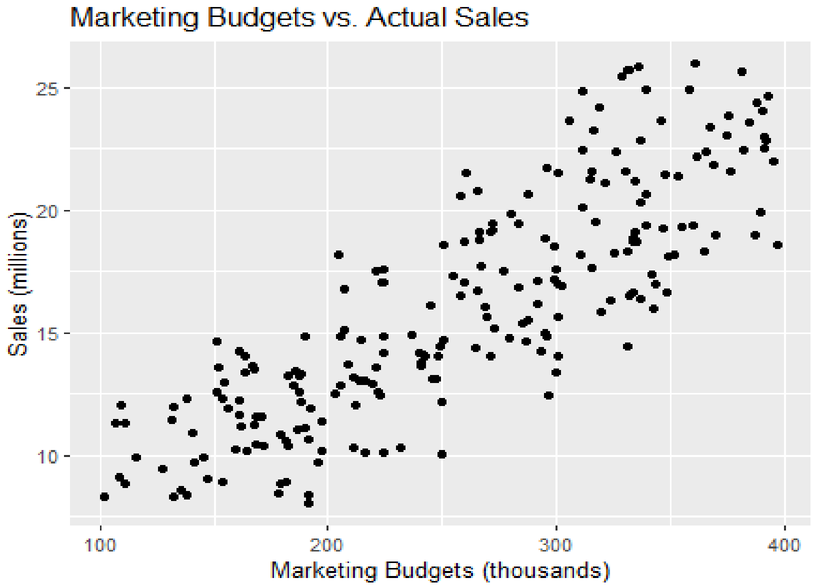
2. For the feature actual sales: Steam-and-leaf plot

We sort the data using the sort command and create a stem-and-leaf plot for the first 100 elements using the myStem command. The stem represents the integer part of the number (before the decimal point), and the leaf consists of the decimal digits (after the decimal point).



```
8   |   1  32  33  36  38  41  55  82  85  87  90
9   |   1  7  44  70  71  88  90
10    |   3  7  13  14  19  21  29  30  34  34  45  60  66  83  93
11    |   3  9  15  26  30  32  37  42  56  57  66  87  90  99
12    |   2  6  14  17  24  30  33  41  43  49  55  56  57  86  87  88  96
13    |   3  6  8  13  20  22  27  28  34  35  42  49  55  59  62  63  73  74
14    |   2  3  4  6  6  7  10  14  18  23  26  40  41  42  62  66  68  70
```

3. Scatter plot



```
#Grafik na rasejuvanje so ggplot2
scatter_plot <- ggplot(data, aes(x = marketing_budgets, y = actual_sales)) +
  geom_point() +
  labs(title = "Маркетинг буџети наспроти реални продажби",
       x = "Маркетинг Буџети (илјадници)",
       y = "Продажби (милиони)")
```

Based on the scatterplot, we can see that there is a strong positive relationship between marketing budgets and actual sales.

4. **For the feature marketing budgets: Mode, median and mean of the data**

```
median_value <- median(marketing_budgets)

mean_value <- mean(marketing_budgets)
```

The Median is: 262.485, The Mean is: 257.9929

To find the mode, we create a frequency table (showing how often each value occurs), sort the table in descending order, and extract the first value as the mode:

```
mode_value <- as.numeric(names(sort(table(marketing_budgets), decreasing = TRUE)[1]))
```

The Mode is: 287.93

4. **For the feature actual sales: Mode, median and mean of the data**

```
mode_value <- as.numeric(names(sort(table(actual_sales), decreasing = TRUE)[1]))

median_value <- median(actual_sales)

mean_value <- mean(actual_sales)
```

The Mode is: 18.18 , The Median is: 15.56 , The mean is: 16.0832

5. **For the feature marketing budgets: Quartiles, Range, and Interquartile Range**

```
q1 <- quantile(marketing_budgets, probs = 0.25)
q2 <- median(marketing_budgets)
q3 <- quantile(marketing_budgets, probs = 0.75)
iqr <- q3-q1
range <- max(marketing_budgets) - min(marketing_budgets)
```

Q1: 189.585 , Q2 (Median): 262.485， Q3: 327.7875

Interquartile Range (difference between the third and first quartile): 138.2025; Range: 294.97

5. **For the feature actual sales: Quartiles, Range, and Interquartile Range**

Q1: 12.445 , Q2 (Median): 15.56 , Q3: 19.31 , IQR: 6.865, Interquartile Range:17.97

6. For the feature marketing budgets: Variance and standard deviation

We use the functions var(x) and sd(x).

Variance: 6152.726, Standard Deviation: 78.43931

6. For the feature actual sales: Variance and standard deviation

We use the functions var(x) and sd(x).

Variance: 21.51985, Standard Deviation: 4.63895

7. Correlation Coefficient

```
correlation_coefficient <- cor(marketing_budgets, actual_sales)
```

The Correlation Coefficient is: 0.8367015

Since the correlation coefficient is close to 1, we can conclude that there is a strong positive linear relationship between marketing budgets and actual sales.

B. Part Two

1. Confidence interval – For the feature actual sales

To determine the confidence interval for the population mean (the parameter we are considering), we first calculate the sample size. It is known that n=222. Since the sample size is greater than 30, we use z-statistics.
The formula for calculating the confidence interval is: $\bar{X} \pm Z\alpha/2 * (\sigma / \sqrt{n})$. To find the critical value for a 95% confidence level, we use the qnorm() function in R: critical_value <- qnorm(0.975)

We calculated the standard error as $\sigma/\sqrt{n}$, and the margin of error by multiplying the standard error by the critical value (confidence factor, $Z\alpha/2$).

Standard Error: 0.3113459, Margin of Error: 0.6102268

The margin of error indicates how much the interval deviates from the sample mean.

We calculate the confidence interval with a sample mean of 16.0832:

```
 confidence_interval_lower <- mathematical_expectation - margin_of_error
confidence_interval_upper <- mathematical_expectation + margin_of_error
```

The confidence interval for the population mean of the sales feature is:

(15.47297 , 16.69342)

2. Hypothesis Testing – For the feature actual sales

Initial assumption: Sales are expected to be 16.0832 million, with a standard deviation of s=4.63895s based on the sample.
We test whether the expected sales (for the sample) deviate from the standard with a 95% confidence level.

Hypotheses:

- H0: μ = 16
- Ha (alternative hypothesis): μ > 16

Since we are dealing with a large sample and testing the population mean, we use z-statistics (according to the Central Limit Theorem).
Using the formula Z=(($\bar{X}-\mu$)/s)*$\sqrt{n}$ we find z= 0.2672211.

Using the qnorm() function, we determine the critical domain (for a one-tailed test):

(1.644854 , +∞)

Since z (test statistic value) does not fall within the critical domain, we do not reject the null hypothesis in a one-tailed test at a 95% confidence level.

Conclusion: The sales for the randomly selected sample do not significantly deviate from the expected sales of 16.0832 million.

3. Distribution Test – For the feature actual sales

The histogram does not clearly indicate the distribution of the feature, as its shape does not follow any studied distributions. For the actual sales feature, we choose to test for uniform distribution. The data (222 values) are divided into 8 intervals. The observed frequencies represent the number of values in each interval. The expected probability for each interval under the assumption of uniform distribution is 1/8.

Hypotheses:

- H0: The feature follows a uniform distribution.
- Ha: The feature does not follow a uniform distribution.

We use the chi-squared test:

```
chi_squared_test <- chisq.test(observed_counts, p = expected_probabilities)
```

df = 7 (Degrees of freedom) , X-squared = 24.032

This test yields a p-value of 0.001125.

Since the p-value is less than α(0.05), we reject the null hypothesis.

Conclusion: The feature does not follow a uniform distribution. The data are not uniformly distributed across intervals and show significant variations in frequencies.

4. Hypothesis Testing for Independence

It is not possible to test hypotheses for independence because we are working with numerical data. Independence tests are typically used when working with two categorical variables to determine if there is a statistically significant association or dependence between them. For purely numerical relationships, other statistical techniques, such as regression or correlation analysis, are more appropriate. These methods are designed to directly analyze relationships between numerical variables.

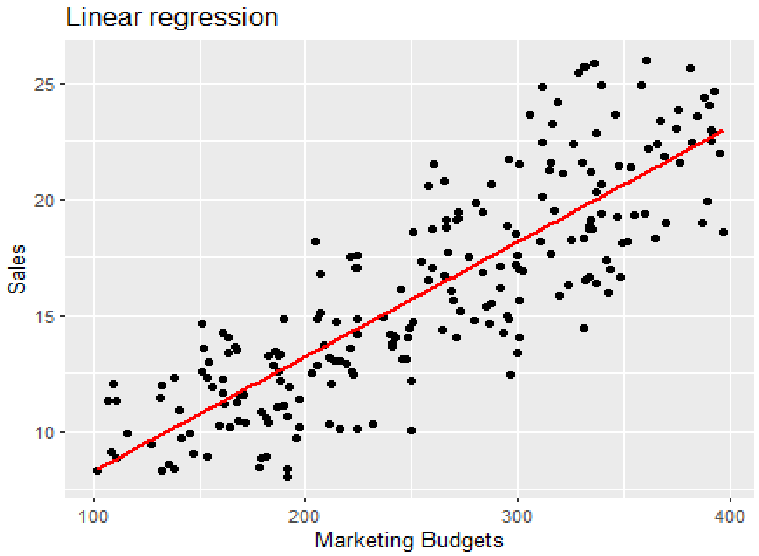
5. Regression Analysis

We determine the regression line for sales depending on marketing budgets. We calculate y based on x.

y (dependent variable) represents sales, x (independent variable) represents marketing budgets.

The regression line is: y = 3.32 + 0.05 * x . Using this equation, we can estimate sales for any given marketing budget.

The Coefficient of Determination is 0.7000694. Conclusion: Marketing budgets have a significant impact on sales. There is a strong linear relationship, which can also be observed graphically.



```
# Grafik so ggplot2
 scatter_plot <- ggplot(data, aes(x = marketing_budgets, y = actual_sales)) +
   geom_point() +
   geom_smooth(method = "lm", se = FALSE, color = "red") +  #dodavanje na pravata
   labs(
     title = "Линеарна регресија",
     x = "Маркетинг буџети",
     y = "Продажби"
   )
cat("Права на регресија: y = ", round(coefficients(regression)[1], 2), " + ", round(coefficients(regression)[2], 2), " * x\n")
```