VSN Sai Krishna Mohan Kocherlakota

Title: Module 2 R Practice

Submission Date: 14th November 2023

INTRODUCTION:

In this project, a dataset of “Predict students' dropout and academic success” with almost 37 variables. The dataset was subjected to data cleaning, various data visualization and statistical analysis techniques using different R packages. Analysed student demographics, statistics, and grade comparison through R programming, showcasing marital status, age statistics, and visualizing grade distributions via plots and analysis.

TASKS:

1. Data loading and cleaning

```
> student <- read_csv("data.csv")
Rows: 4424 Columns: 37
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr  (1): Target
dbl (36): Marital status, Application mode, Application order, Course, Daytime/evening atte
ndance, Previous qua...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
>
> student <- clean_names(student)
>
> head(student)
# A tibble: 6 × 37
  marital_status application_mode application_order course daytime_evening_attendance previ
ous_qualification
           <dbl>            <dbl>             <dbl>  <dbl>                      <dbl>
<dbl>
1              1               17                 5    171                          1
1
2              1               15                 1   9254                          1
1
3              1                1                 5   9070                          1
1
4              1               17                 2   9773                          1
1
5              2               39                 1   8014                          0
1
6              2               39                 1   9991                          0
19
```

Loaded dataset into R, cleaned its column names, and displayed the first six rows of the data. The dataset appears to contain various columns such as `marital_status`, `application_mode`, `course`, and others, with corresponding numeric values for each column.

2. Computing Statistical Equations

```
> #Statistics Analysis of Admission Grade
> data_age
  stat_means_ag stat_var_ag
1      23.26514    57.57495
2      23.08000    54.66020
3      23.74000    69.70949
4      22.77000    45.33040

> # Create a bar plot with different colors for means and variances
> barplot(as.matrix(data_age),
+         beside = TRUE,
+         col = c("blue", "red"),  # Set different colors for means and variances
```

```
+         main = "Means and Variances of Admission grades",
+         xlab = "Comapring sample and original statitiscal values of Admission grade",
+         ylab = "Value"
+ )
```

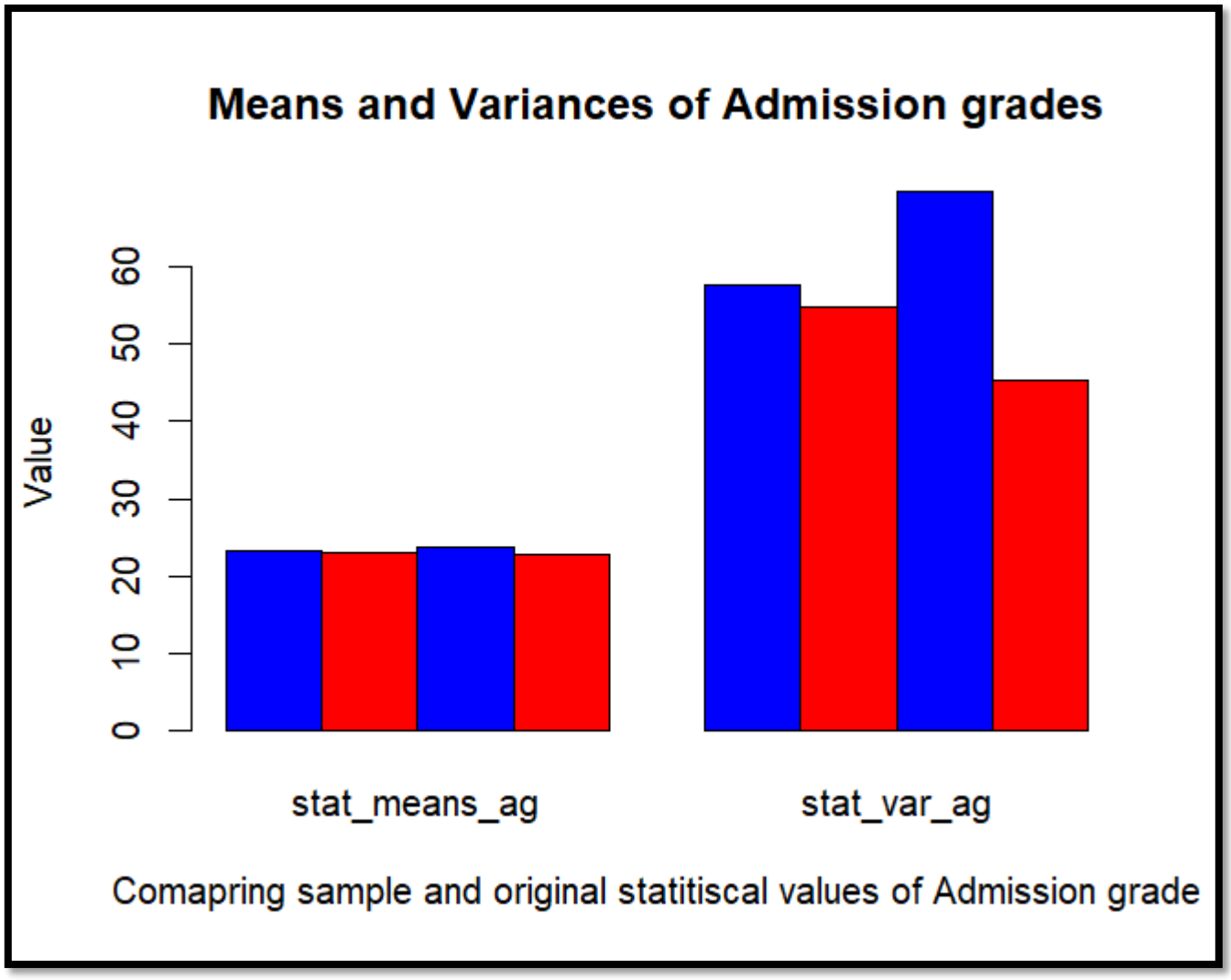

Figure 1: - Multiple Bar Graph of Admission Grade

```
> #Statistics Analysis of Age
> data_age
  stat_means_ag stat_var_ag
1      23.26514    57.57495
2      23.08000    54.66020
3      23.74000    69.70949
4      22.77000    45.33040


>
> # Create a bar plot with different colors for means and variances
> barplot(as.matrix(data_age),
+         beside = TRUE,
+         col = c("blue", "red"),  # Set different colors for means and variances
+         main = "Means and Variances of Age",
+         xlab = "Comapring sample and original statitiscal values of Age",
+         ylab = "Value"
+ )
```

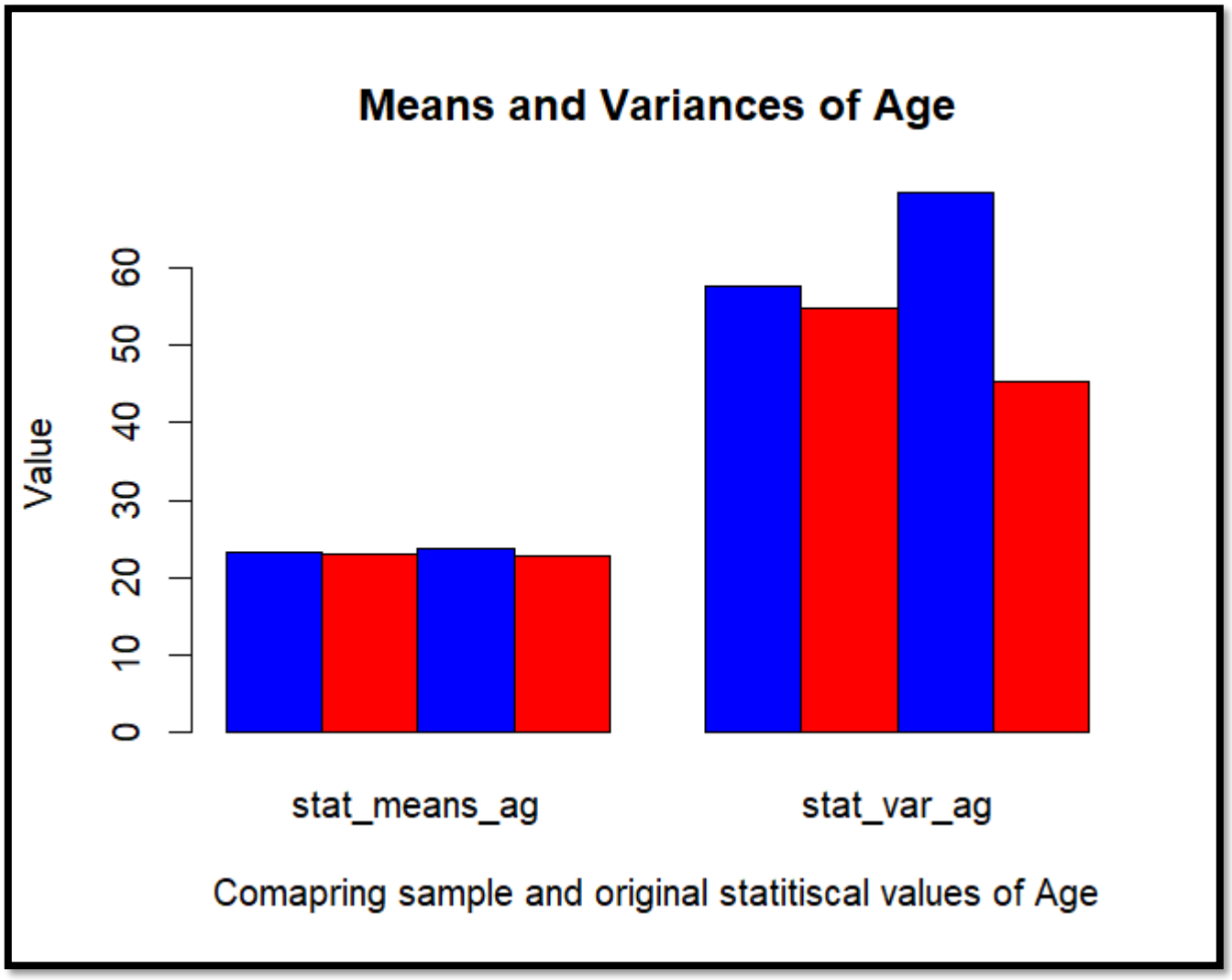

Figure 2: - Multiple Bar Graph of Age

The above R code conducts statistical analyses on the 'age_at_enrollment' data within the 'student' dataset. It generates three samples of size 100 each, computing mean and variance for each sample. Additionally, it calculates the mean and variance for the entire 'age_at_enrollment' data. Two separate data frames, 'data_age' for admission grades and 'data_age' for age, are created to summarize these statistics.

Each 'data_age' frame contains columns for sample means and variances alongside the overall mean and variance. The subsequent `barplot()` functions create side-by-side bar plots comparing these statistics, using different colors ('blue' and 'red') to differentiate between means and variances for both 'age_at_enrollment' and 'Admission grades'. However, there's a small typo ("sumarry") at the end, which seems to refer to the summarization performed in these bar plots. The analysis mainly showcases comparisons of sample means and variances against the overall statistics for these two variables.

3. Data Visualizations

a. Bar graph: -

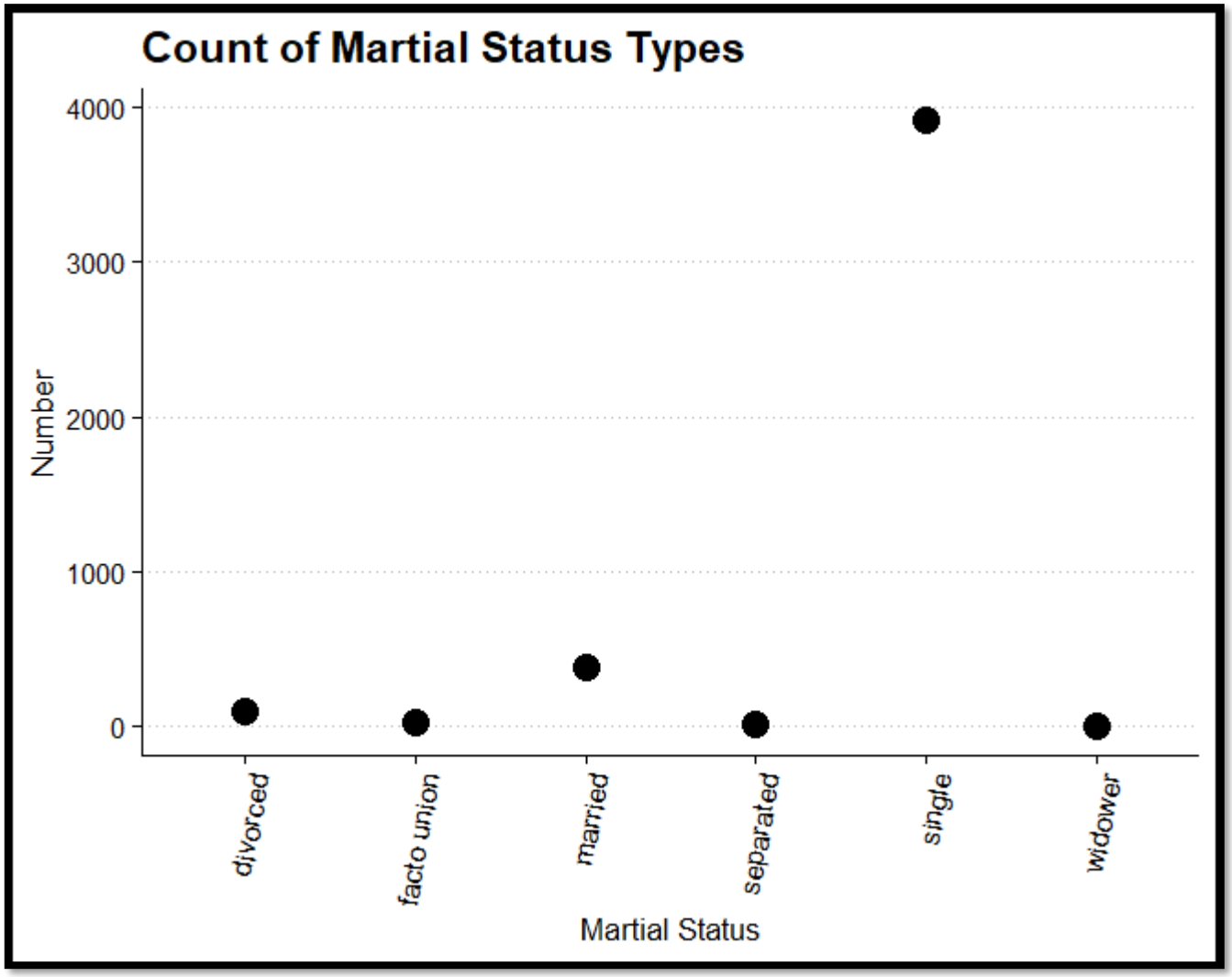

Figure 3: - Bar Graph

```
> bar_plot <- ggplot(ms_df, aes(x = martial_s , y = count)) +
+   geom_line(stat = "Identity") +
+   geom_point(size = 4)+
+   labs(x = "Martial Status", y = "Number") +
+   ggtitle("Count of Martial Status Types") + theme_clean() +
+   easy_rotate_x_labels(angle = 80, side = c("right"), teach = FALSE)
```

In this above code, `ms_df` is generated by grouping the `student` dataset by marital status (`marital_status`) and summarizing the count of occurrences for each status. A new column, `martial_s`, is created based on specific conditions using nested `ifelse()` statements to label each marital status category. However, the subsequent visualization with `ggplot2` attempts to use `geom_line()` and `geom_point()` intended for continuous data, which may not be suitable for categorical counts. A better approach would involve using `geom_bar()` to display the count of each category as bars in a bar plot, enhancing the visual representation of categorical data.

b. Scatter Plot: -

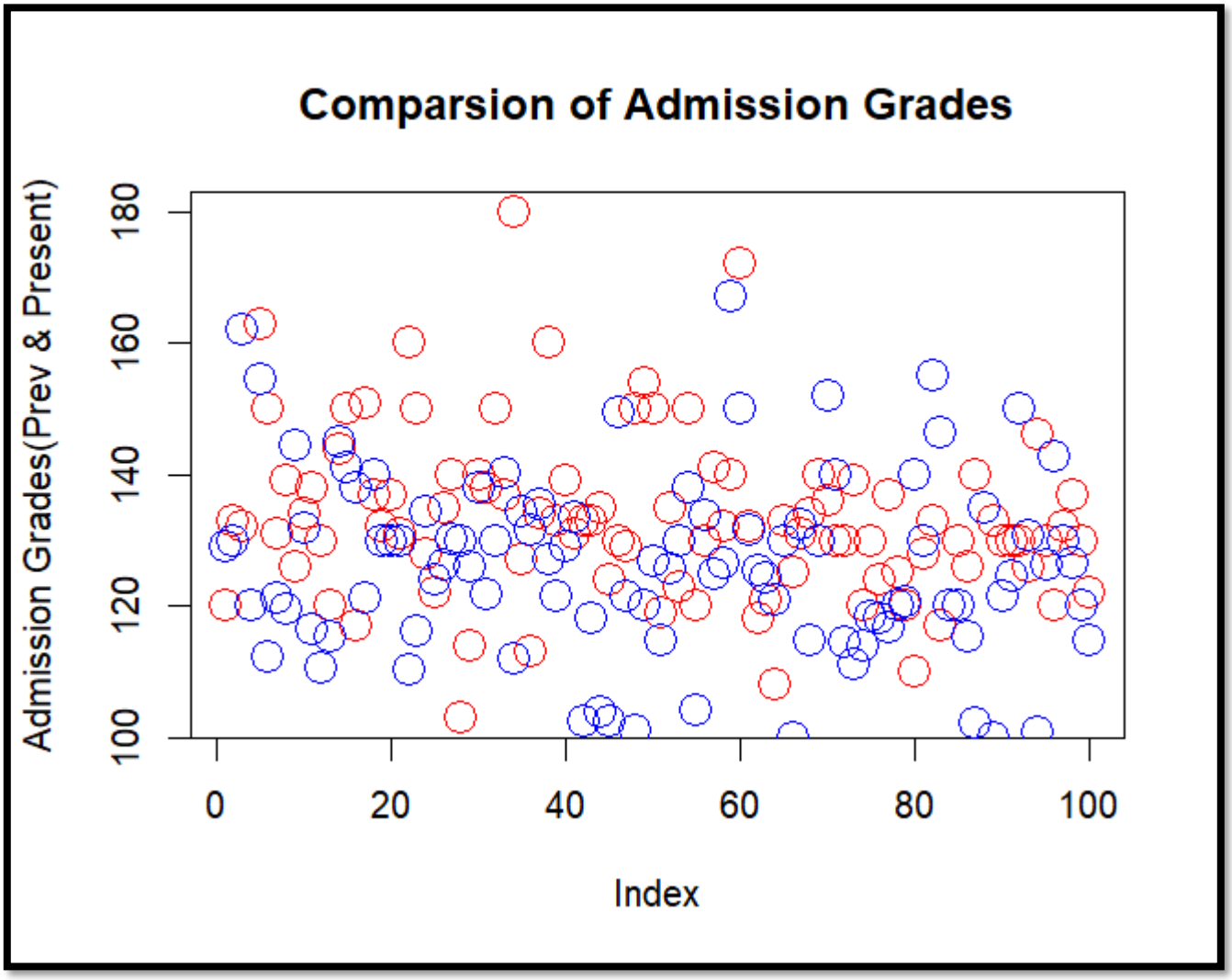

Figure 4: - Scatter Plot

```
> # Scatter Plot of Previous and Admission Grades
> # Set the seed
> set.seed(123)
>
> sample_size <- 100
>
> sample1_pg <- sample(student$previous_qualification_grade, size = sample_size)
> sample1_adg <- sample(student$admission_grade, size = sample_size)
> plot(sample1_pg, main="Comparsion of Admission Grades", ylab="Admission Grades(Prev & Pre
sent)", col="red", cex=2)
> points(sample1_adg, col="blue", cex=2)
```

The R code provides a scatter plot comparing 'previous_qualification_grade' against 'admission_grade' for a randomly sampled subset from the 'student' dataset. The seed is set to ensure reproducibility. Two samples of size 100 are drawn for 'previous_qualification_grade' (`sample1_pg`) and 'admission_grade' (`sample1_adg`). The `plot()` function initially creates a scatter plot for 'previous_qualification_grade' against its index, with red-colored points and a specific size (`cex=2`). The subsequent `points()` function overlays the 'admission_grade' scatter plot on the same graph with blue-colored points and the same size for better comparison between the two variables. This allows visual assessment of potential relationships or patterns between these grades in the subset of data.

b. Histogram: -

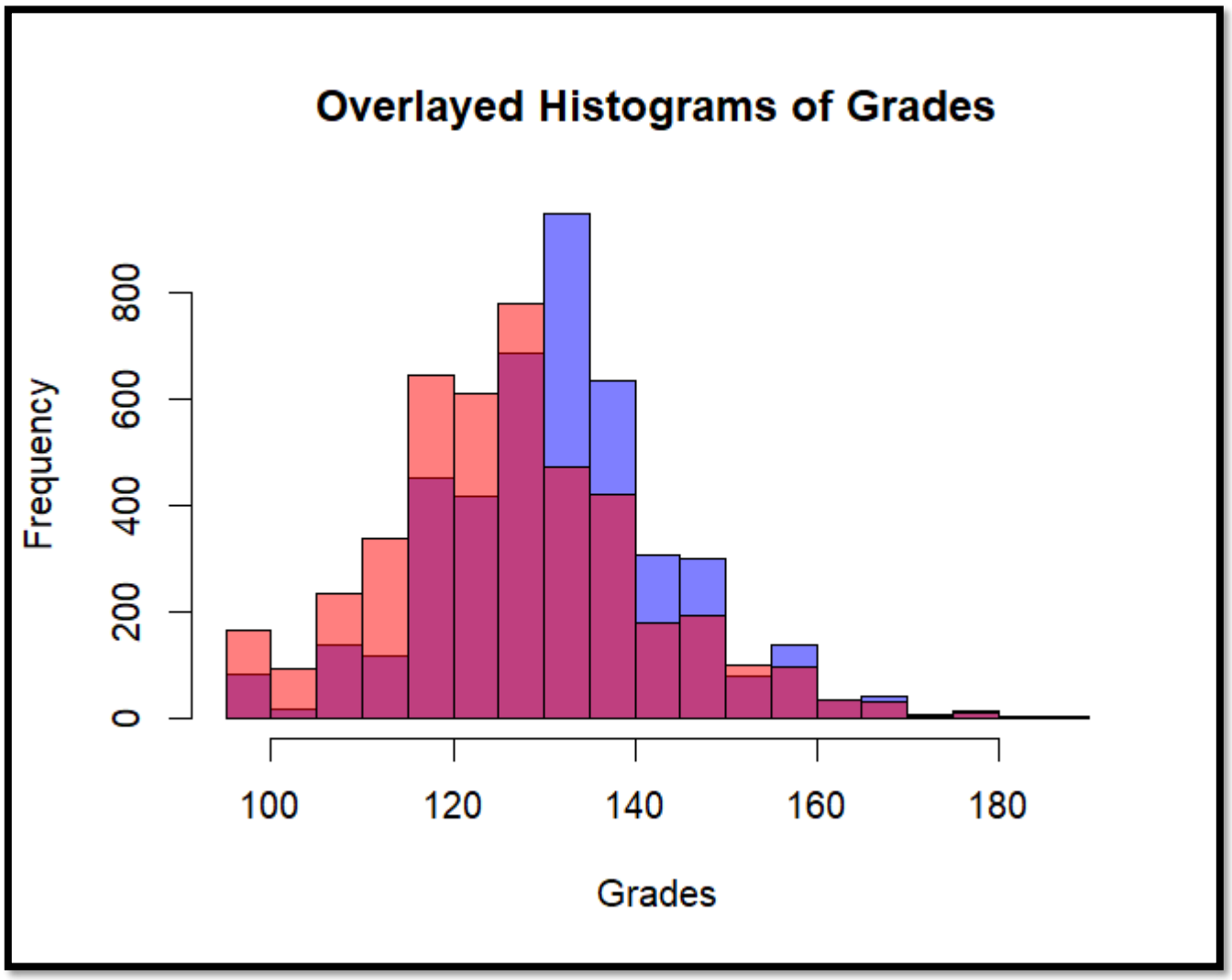

Figure 5: - Histogram

```
> # Generate the histogram data
> p_gd <- hist(student$previous_qualification_grade, plot = FALSE)
> ad_g <- hist(student$admission_grade, plot = FALSE)
>
> # Set colors for the histograms
> c1 <- "blue"
> c2 <- "red"
>
> # Plot the first histogram using a transparent color
> plot(p_gd, col = alpha(c1, 0.5), main = "Overlayed Histograms of Grades", xlab = "Grades"
)
```

The provided code generates histograms for 'previous_qualification_grade' and 'admission_grade' from the 'student' dataset, assigning colors 'blue' and 'red'. However, the attempt to plot the histograms using `plot()` seems to only apply transparency to the first histogram ('previous_qualification_grade'). A summary of the code in 50 words: It generates separate histograms for two grades, setting colors, but currently only displays the first histogram with transparency without overlaying the second.

CONCLUSION

The R code conducts analysis and visualization on student data, displaying marital status counts, statistical summaries (mean and variance) for 'Admission Grade' and 'Age,' a scatter plot comparing 'Previous' and 'Admission Grades,' and overlaid histograms for these grades. The conclusion might state: "The analysis reveals varying marital status counts, sample statistics for admission grades and age, showcasing some relationship between previous and admission grades in a scatter plot. Histograms overlay highlights differing distributions between admission and previous qualification grades."

CITATIONS

1. UCI Machine Learning Repository. (n.d.) https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
2. R Graphics - Scatter Plot. (n.d.). https://www.w3schools.com/r/r_graph_scatterplot.asp
3. Zach. (2021, January 25). How to Write a Nested If Else Statement in R (With Examples). Statology. https://www.statology.org/nested-ifelse-in-r/
4. Plot Two Histograms on one R chart: Tips and Tricks. Data Analytics (2019, November 15). https://www.dataanalytics.org.uk/plot-two-overlapping-histograms-on-one-chart-in-r/