

Report – Project-Based Learning in Data Analytics

Academic Year: 2022 – 2023

Subject: CST322 Data Analytics

Faculty: Mrs. Asha George

Innovative Teaching Method: Project-Based learning and Hands-on Experience in R Studio

Topic/Question: Exploratory Data Analysis on HR Dataset

Objective

1. Gain practical experience in data analytics using R Studio.
2. Explore the HR dataset through descriptive statistics and visualization.
3. Understand the correlation between satisfaction levels and other variables.

Use of Appropriate Methods

1. Tool used: R Studio
2. Commands used: Various R functions for data loading, manipulation, and analysis.
3. Devices used: PC/Laptop for individual hands-on experience.

Tasks

1. Load the HR dataset into R Studio.
2. Explore the dataset structure and check for missing values.
3. Calculate mean and standard deviation of the satisfaction variable.
4. Create a histogram for visualization of satisfaction distribution.
5. Explore correlation between satisfaction and other variables.

Database Description

The HR dataset comprises ten variables that capture various aspects of employee information:

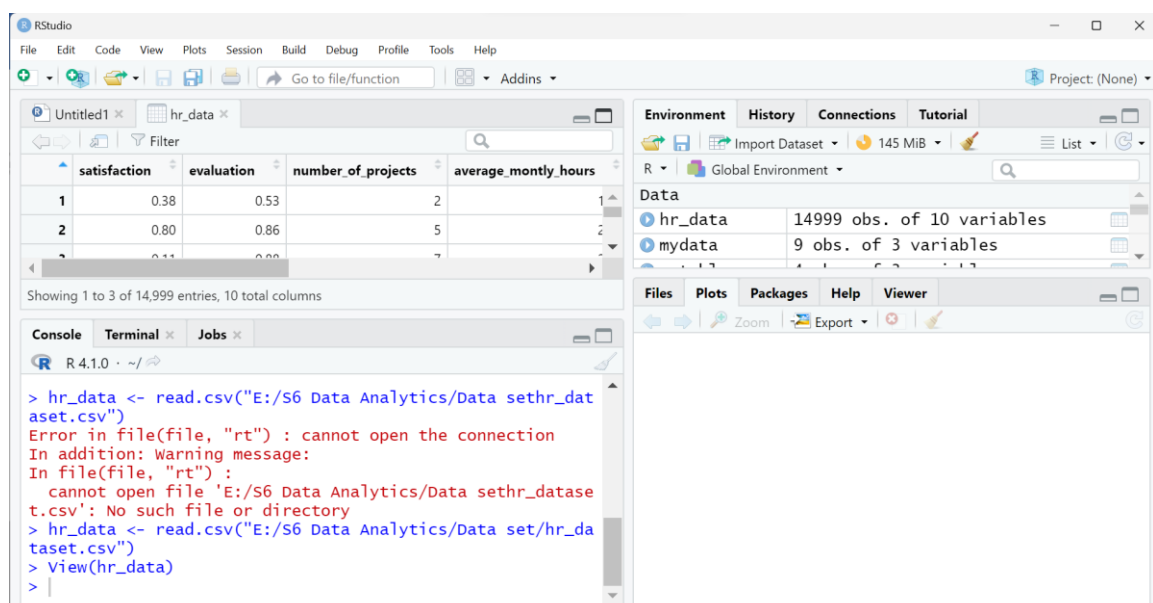
- **satisfaction:** Employee satisfaction level (numeric).
- **evaluation:** Performance evaluation score (numeric).
- **number_of_projects:** Number of projects assigned to the employee (numeric).
- **average_monthly_hours:** Average number of monthly working hours (numeric).
- **time_spent_company:** Number of years the employee has spent at the company (numeric).
- **work_accident:** Binary indicator of whether the employee had a work accident (0 for No, 1 for Yes).
- **churn:** Binary indicator of employee churn (0 for No, 1 for Yes).
- **promotion:** Binary indicator of employee promotion (0 for No, 1 for Yes).
- **department:** Department in which the employee works (categorical).
- **salary:** Employee salary level (categorical).

Project Implementation Steps

1. Load the HR dataset into R Studio:

```
hr_data <- read.csv("hr_dataset.csv")
```

In this step, we load the HR dataset into R Studio using the **read.csv** function.



2. Explore the dataset structure and check for missing values:

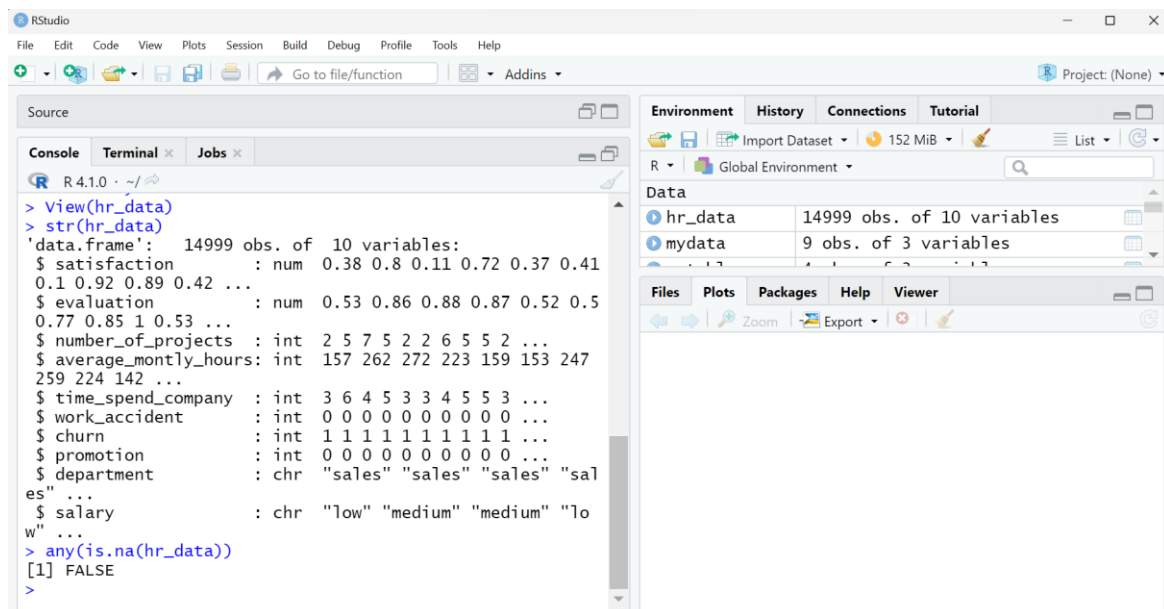
```
# Display the structure of the dataset

str(hr_data)

# Check for missing values in the dataset

any(is.na(hr_data))
```

Here, we examine the structure of the dataset using the **str** function to understand its columns and data types. Additionally, we check for any missing values in the dataset using the **any(is.na(...))** expression.



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins Project: (None)
Source
Console Terminal Jobs
R 4.1.0 ~ /
> View(hr_data)
> str(hr_data)
'data.frame': 14999 obs. of 10 variables:
 $ satisfaction : num 0.38 0.8 0.11 0.72 0.37 0.41
 0.1 0.92 0.89 0.42 ...
 $ evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5
 0.77 0.85 1 0.53 ...
 $ number_of_projects : int 2 5 7 5 2 2 6 5 5 2 ...
 $ average_monthly_hours: int 157 262 272 223 159 153 247
 259 224 142 ...
 $ time_spend_company : int 3 6 4 5 3 3 4 5 5 3 ...
 $ work_accident : int 0 0 0 0 0 0 0 0 0 ...
 $ churn : int 1 1 1 1 1 1 1 1 1 ...
 $ promotion : int 0 0 0 0 0 0 0 0 0 ...
 $ department : chr "sales" "sales" "sales" "sal
es" ...
 $ salary : chr "low" "medium" "medium" "lo
w" ...
> any(is.na(hr_data))
[1] FALSE
>
```

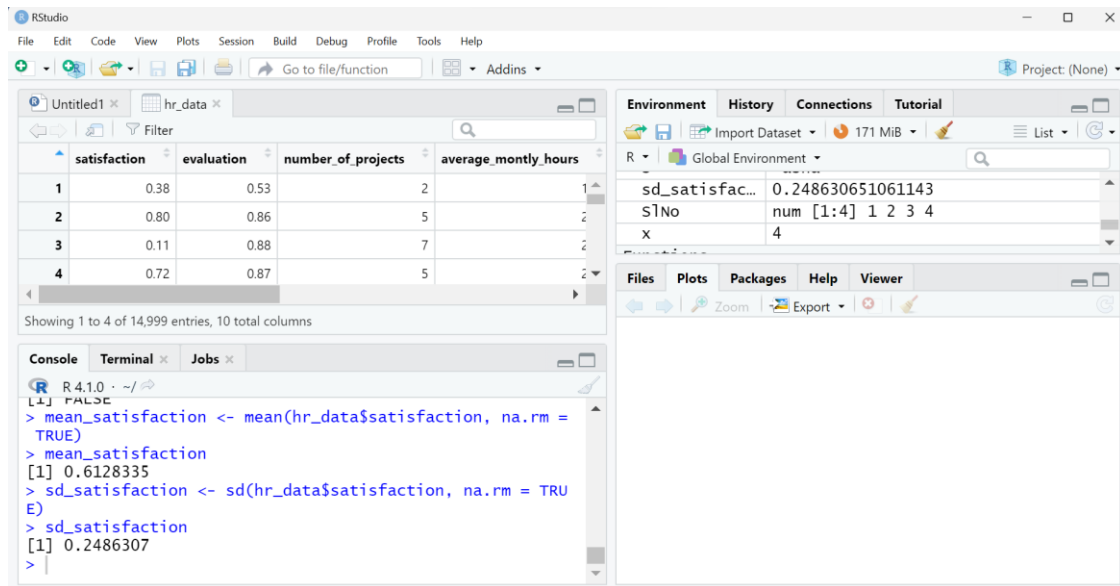
3. Calculate mean and standard deviation of the 'satisfaction' variable:

```
# Calculate mean and standard deviation of 'satisfaction'

mean_satisfaction <- mean(hr_data$satisfaction, na.rm = TRUE)

sd_satisfaction <- sd(hr_data$satisfaction, na.rm = TRUE)
```

These steps compute the mean and standard deviation of the 'satisfaction' variable, providing insights into the central tendency and variability of the satisfaction levels.



4. Create a histogram to visualize the distribution of satisfaction levels:

Create a histogram for 'satisfaction'

```
hist(hr_data$satisfaction, main = "Satisfaction Distribution", xlab = "Satisfaction Level",
col = "skyblue", border = "black")
```

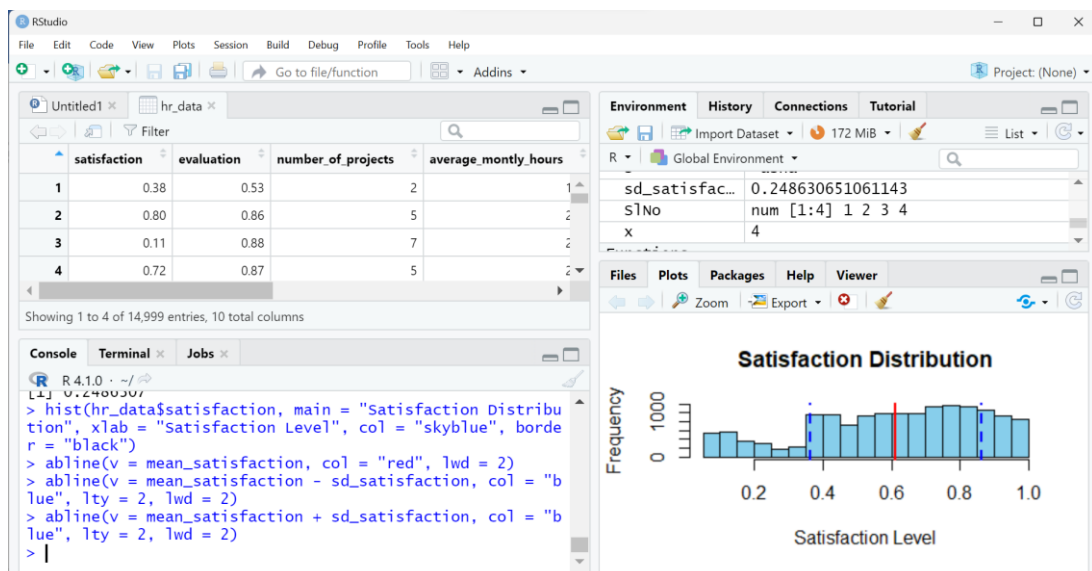
Add lines for mean and standard deviation to the histogram

```
abline(v = mean_satisfaction, col = "red", lwd = 2)
```

```
abline(v = mean_satisfaction - sd_satisfaction, col = "blue", lty = 2, lwd = 2)
```

```
abline(v = mean_satisfaction + sd_satisfaction, col = "blue", lty = 2, lwd = 2)
```

We generate a histogram to visually represent the distribution of satisfaction levels. The added lines indicate the mean and one standard deviation from the mean.



5. Explore the correlation between 'satisfaction' and other variables:

This following code calculates the correlation between 'satisfaction' and 'evaluation', providing insights into the relationship between these two variables.

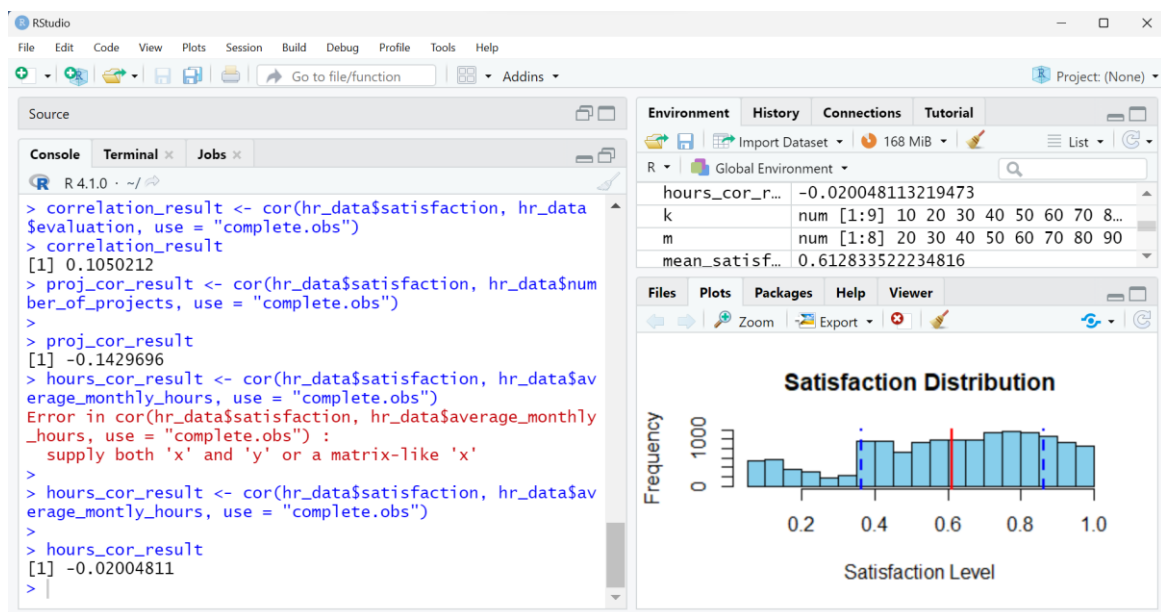
```
correlation_result <- cor(hr_data$satisfaction, hr_data$evaluation, use = "complete.obs")
```

To find the correlation between "satisfaction" and "number_of_projects", we run the following code.

```
proj_cor_result <- cor(hr_data$satisfaction, hr_data$number_of_projects, use =  
"complete.obs")
```

The relationship between employee satisfaction and average monthly hours worked can be found using the code:

```
hours_cor_result <- cor(hr_data$satisfaction, hr_data$average_monthly_hours, use =  
"complete.obs")
```



Insights from the HR Dataset

1. Overall Satisfaction Metrics:

- The organization maintains a moderate overall satisfaction level with a mean of 0.6128.
- The standard deviation of 0.248 indicates a notable variability in individual satisfaction.

2. Satisfaction Distribution:

- A histogram illustrates a generally normal distribution, centered around the moderate satisfaction range (approximately 0.4 to 0.8).
- A tail in the distribution suggests a subset of employees with lower satisfaction levels.

3. Correlation Analysis:

- The correlation between satisfaction and evaluation is 0.105. This indicates a weak positive correlation between satisfaction and evaluation. This suggests a slight tendency for employees with higher satisfaction levels to have higher evaluations. However, the weak strength implies that other factors contribute more significantly to the variation in these variables.
- The correlation between satisfaction and number_of_projects is -0.1429. This suggests a weak negative relationship between employee satisfaction and the number_of_projects. As the number of projects increases, there is a slight tendency for employee satisfaction to decrease.
- The correlation coefficient of -0.02 between satisfaction and average_monthly_hours indicate a very weak negative correlation. This suggests that there is little to no linear relationship between employee satisfaction and the number of hours they work on average per month.

Conclusion

This project provided valuable insights into the HR dataset, including descriptive statistics, visualization, and correlation analysis. The hands-on experience in R Studio enhances practical skills in data analytics.

1:22

Vg LTE 4G



CST322 DA CSE2...

Sooraj Sir Tki, +91 70346...



Your security code with Vimal Sir Tki changed.
Tap to learn more.

5 June 2023

Install R studio in laptop and bring tomorrow onwards

9:33 pm ✓✓

6 June 2023

You

Install R studio in laptop and bring tomorrow onwards



9:27 pm ✓✓

8 June 2023



hr_dataset.csv

567 kB • CSV



11:15 pm ✓✓

download this dataset. It should be available with you in tomorrow class

11:17 pm ✓✓

10 June 2023



Message

