statmodata2

Ezo YEDİGÖL

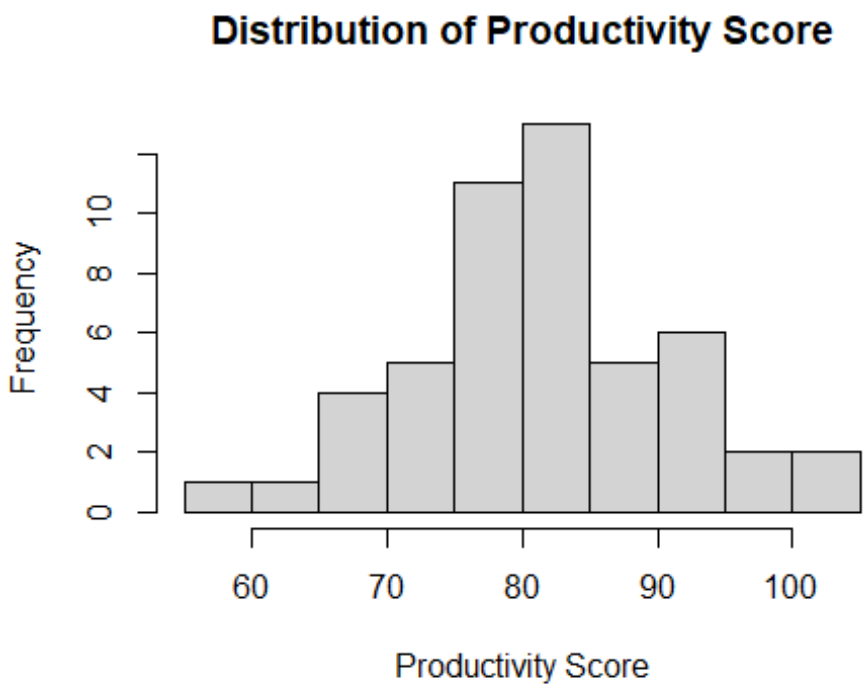
2024-04-18

#DATASET2 #The dataset comprises five variables representing employee performance metrics: Salary, Productivity Score, Experience Level, Training Hours, and Teamwork Rating. It consists of 50 observations (employees), and each variable may play a role in determining employee performance and salary.

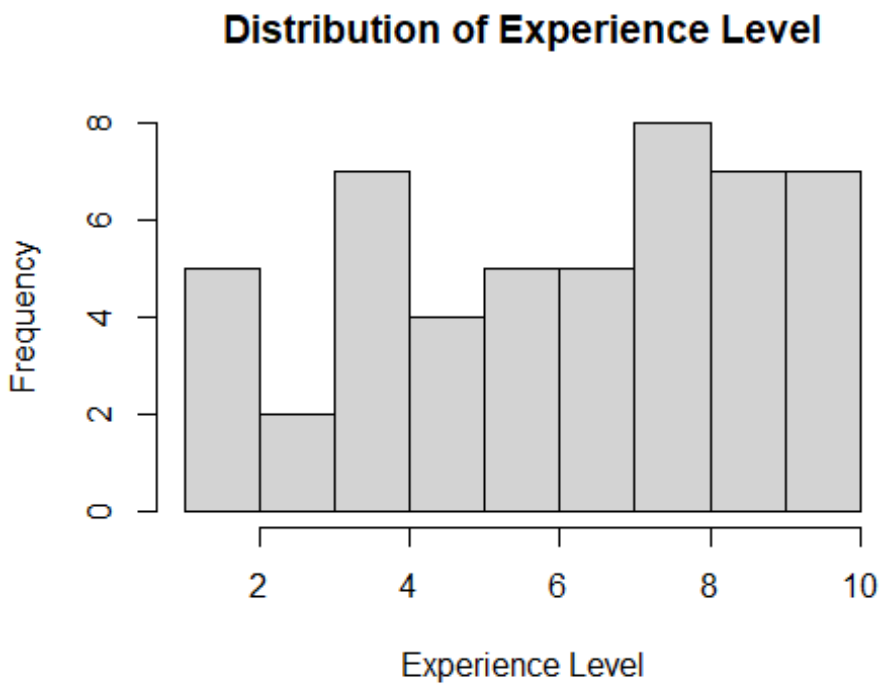
```
# Creating a synthetic data set for a different scenario
set.seed(123)  # Setting seed for reproducibility

# Sample data set size
n_samples <- 50

# Dependent variable: Salary
salary <- rnorm(n_samples, mean = 5000, sd = 1000)

# Independent variables: Productivity Score, Experience Level, Training
Hours, Teamwork Rating
productivity_score <- rnorm(n_samples, mean = 80, sd = 10)
experience_level <- sample(1:10, n_samples, replace = TRUE)
training_hours <- rnorm(n_samples, mean = 20, sd = 5)
teamwork_rating <- sample(1:5, n_samples, replace = TRUE)

# Creating a data frame
employee_data <- data.frame(Salary = salary, ProductivityScore =
productivity_score, ExperienceLevel = experience_level, TrainingHours =
training_hours, TeamworkRating = teamwork_rating)

# Showing the head of the created data set
head(employee_data)

##     Salary ProductivityScore ExperienceLevel TrainingHours TeamworkRating
## 1 4439.524          82.53319               4      14.53293              2
## 2 4769.823          79.71453               9      16.22268              2
## 3 6558.708          79.57130               8      18.68004              4
## 4 5070.508          93.68602               6      16.23769              3
## 5 5129.288          77.74229               4      22.20346              4
## 6 6715.065          95.16471               8      13.61275              3

# Exploring the structural characteristics of the data set
str(employee_data)

## 'data.frame':    50 obs. of  5 variables:
##  $ Salary           : num  4440 4770 6559 5071 5129 ...
```

```
##  $ ProductivityScore: num  82.5 79.7 79.6 93.7 77.7 ...
##  $ ExperienceLevel  : int  4 9 8 6 4 8 3 4 4 6 ...
##  $ TrainingHours    : num  14.5 16.2 18.7 16.2 22.2 ...
##  $ TeamworkRating   : int  2 2 4 3 4 3 3 3 5 3 ...

# Summary statistics for numerical variables
summary(employee_data)

##      Salary     ProductivityScore ExperienceLevel TrainingHours
##  Min.   :3033   Min.   : 56.91    Min.   : 1.00   Min.   : 8.004
##  1st Qu.:4441   1st Qu.: 76.39    1st Qu.: 4.00   1st Qu.:16.017
##  Median :4927   Median : 81.53    Median : 7.00   Median :19.206
##  Mean   :5034   Mean   : 81.46    Mean   : 6.46   Mean   :19.500
##  3rd Qu.:5698   3rd Qu.: 86.29    3rd Qu.: 9.00   3rd Qu.:24.122
##  Max.   :7169   Max.   :101.87    Max.   :10.00   Max.   :30.980
##  TeamworkRating
##  Min.   :1.00
##  1st Qu.:2.00
##  Median :3.00
##  Mean   :3.02
##  3rd Qu.:4.00
##  Max.   :5.00

# Graphical Exploratory Data Analysis
hist(employee_data$Salary, main = "Distribution of Salary", xlab = "Salary",
ylab = "Frequency")
```

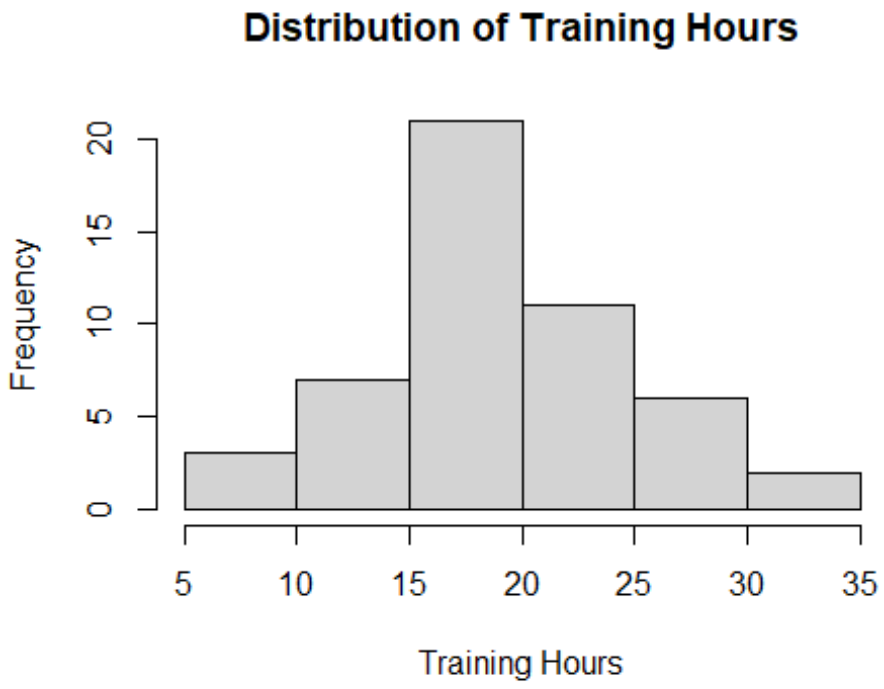
```r
hist(employee_data$ProductivityScore, main = "Distribution of Productivity
Score", xlab = "Productivity Score", ylab = "Frequency")
```



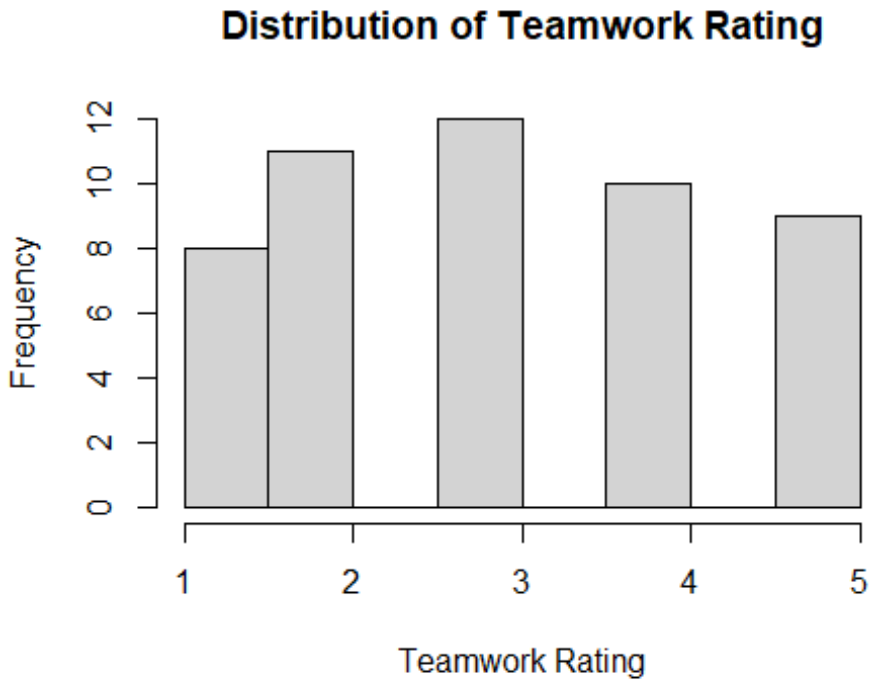

```r
hist(employee_data$ExperienceLevel, main = "Distribution of Experience
Level", xlab = "Experience Level", ylab = "Frequency")
```



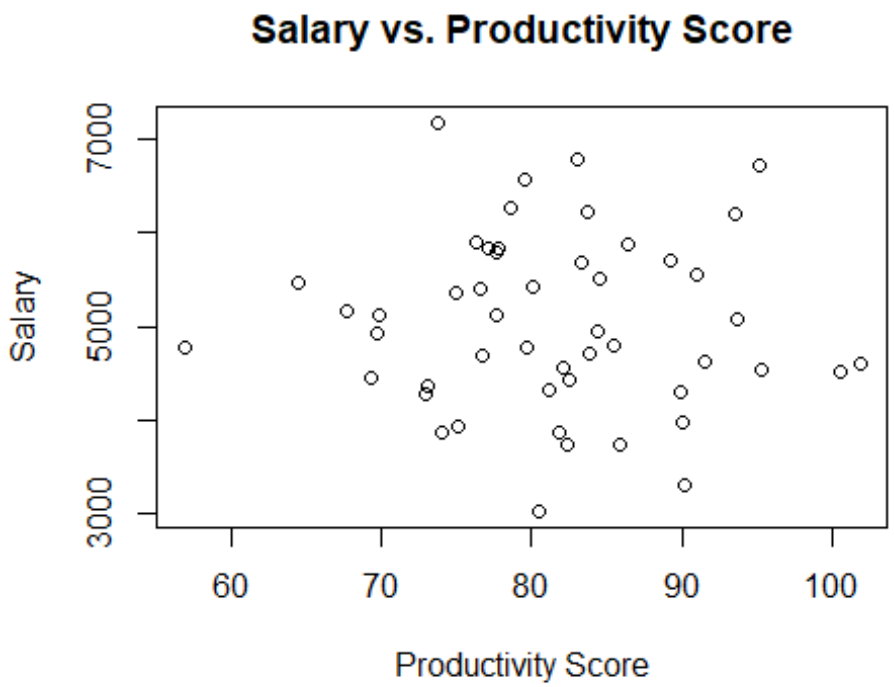

```r
hist(employee_data$TrainingHours, main = "Distribution of Training Hours",
xlab = "Training Hours", ylab = "Frequency")
```



```r
hist(employee_data$TeamworkRating, main = "Distribution of Teamwork Rating",
xlab = "Teamwork Rating", ylab = "Frequency")
```

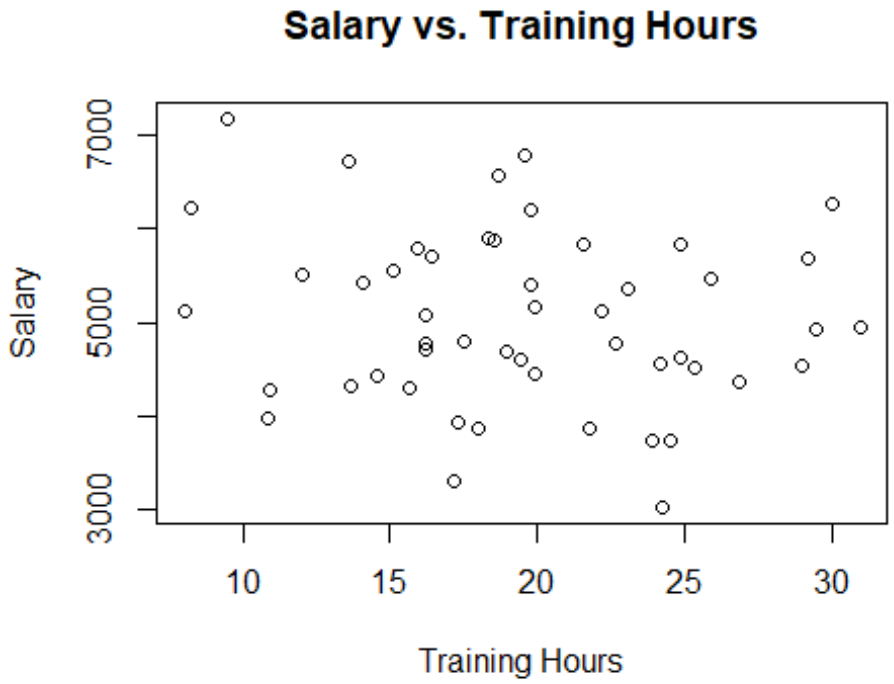



```r
plot(employee_data$ProductivityScore, employee_data$Salary, main = "Salary
vs. Productivity Score", xlab = "Productivity Score", ylab = "Salary")
```

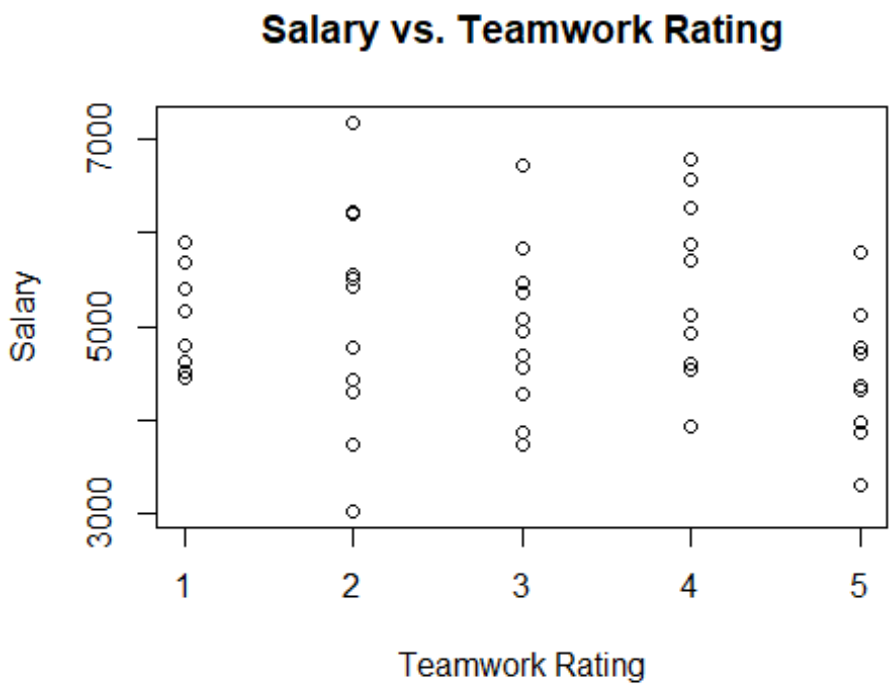



```r
plot(employee_data$ExperienceLevel, employee_data$Salary, main = "Salary vs. Experience Level", xlab = "Experience Level", ylab = "Salary")
```

```r
plot(employee_data$TrainingHours, employee_data$Salary, main = "Salary vs.
Training Hours", xlab = "Training Hours", ylab = "Salary")
```




```r
plot(employee_data$TeamworkRating, employee_data$Salary, main = "Salary vs.
Teamwork Rating", xlab = "Teamwork Rating", ylab = "Salary")
```




```
# Regression Analysis
# Modeling the relationship between Salary and independent variables

# Model creation
regression_model <- lm(Salary ~ ProductivityScore + ExperienceLevel +
TrainingHours + TeamworkRating, data = employee_data)

# Model summary
summary(regression_model)

##
## Call:
## lm(formula = Salary ~ ProductivityScore + ExperienceLevel + TrainingHours
+
##     TeamworkRating, data = employee_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1529.00  -729.52   -87.25   684.30  1701.33
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5393.860   1438.187   3.750 0.000502 ***
## ProductivityScore   -1.688     14.591  -0.116 0.908425
## ExperienceLevel     97.700     48.483   2.015 0.049889 *
## TrainingHours      -28.735     22.467  -1.279 0.207460
## TeamworkRating     -96.944     96.830  -1.001 0.322098
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 904.2 on 45 degrees of freedom
## Multiple R-squared:  0.1241, Adjusted R-squared:  0.04625
## F-statistic: 1.594 on 4 and 45 DF,  p-value: 0.1923

# ANOVA Analysis
# Testing the significance of the overall model

# ANOVA model creation
anova_model <- lm(Salary ~ ProductivityScore + ExperienceLevel +
TrainingHours + TeamworkRating, data = employee_data)

# ANOVA results
anova_result <- anova(anova_model)
print(anova_result)

## Analysis of Variance Table
##
## Response: Salary
##                   Df   Sum Sq Mean Sq F value  Pr(>F)
## ProductivityScore  1    54045   54045  0.0661 0.79827
## ExperienceLevel    1  3132862 3132862  3.8318 0.05651 .
## TrainingHours      1  1206721 1206721  1.4760 0.23074
## TeamworkRating     1   819502  819502  1.0023 0.32210
## Residuals         45 36791397  817587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# ANCOVA Analysis
# Testing the significance of the model with covariates

# ANCOVA model creation
ancova_model <- lm(Salary ~ ProductivityScore + ExperienceLevel +
TrainingHours + TeamworkRating, data = employee_data)

# ANCOVA results
summary(ancova_model)

##
## Call:
## lm(formula = Salary ~ ProductivityScore + ExperienceLevel + TrainingHours
+
##     TeamworkRating, data = employee_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1529.00  -729.52   -87.25   684.30  1701.33
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5393.860   1438.187   3.750 0.000502 ***
## ProductivityScore   -1.688     14.591  -0.116 0.908425
## ExperienceLevel     97.700     48.483   2.015 0.049889 *
## TrainingHours      -28.735     22.467  -1.279 0.207460
## TeamworkRating     -96.944     96.830  -1.001 0.322098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 904.2 on 45 degrees of freedom
## Multiple R-squared:  0.1241, Adjusted R-squared:  0.04625
## F-statistic: 1.594 on 4 and 45 DF,  p-value: 0.1923

#conclusions
#Regression Analysis (regression_model):
#Intercept (Constant): The intercept represents the expected mean value of
Salary when all independent variables are zero. In this model, it's
approximately $5393.86.

#Coefficients for Independent Variables:
#Productivity Score: For each unit increase in Productivity Score, the Salary
is expected to decrease by approximately $1.69.
#Experience Level: For each additional year of Experience Level, the Salary
is expected to increase by approximately $97.70.
#Training Hours: For each additional hour of Training, the Salary is expected
to decrease by approximately $28.74.
#Teamwork Rating: For each unit increase in Teamwork Rating, the Salary is
expected to decrease by approximately $96.94.

#ANOVA Analysis (anova_result):
#The ANOVA table tests the overall significance of the regression model by
comparing the variance explained by the model to the residual variance.
#The p-values associated with each independent variable (Productivity Score,
#Experience Level, Training Hours, Teamwork Rating) indicate whether these
variables are jointly significant in explaining the variation in Salary.
#In this case:
#Experience Level has a p-value of 0.05651, indicating it might be marginally
significant.
#Productivity Score, Training Hours, and Teamwork Rating have p-values above
0.05, suggesting they are not statistically significant in explaining Salary.

#ANCOVA Analysis (ancova_model):
#The results from ANCOVA are identical to the regression analysis because the
model formula and data used are the same.
#ANCOVA is essentially a regression analysis that includes quantitative and
categorical predictors (covariates).
#Overall, the Experience Level appears to be the most influential variable in
predicting Salary, while other variables such as Productivity Score, Training
Hours, and Teamwork Rating do not show significant associations with Salary
in this analysis.
```