University of Jeddah.

Collage of Computer Science and Engineering.

Department of Information System & Technology.

Course name: CCDS211 – Introduction to Data Science.

Project title: Linear Regression Analysis.

Rimas Almuntashiri – 2311631

Sadem awak – 2301922

Mariam Bourima - 2317479

5December2024.

-Linear regression is a statistical method used to model and analyze the relationship between two variables: an independent variable (input) and a dependent variable (output). It aims to find a straight line (regression line) that best fits the data, showing how changes in the independent variable influence the dependent variable.

-Our team worked on the database before saving it into an Excel file, which involved several preprocessing steps. As result, some differences in the code commands may appear due to these modifications. The steps included:

1-Cleaning missing values.

2-Normalizing data ranges.

3-Removing duplicate entries.

4-Formatting columns to ensure consistency.

5-Adding calculated fields for better analysis.

-The dataset consists of 153 entries with 6 columns related to air quality measurements. The columns include:

-Ozone: Ozone levels in parts per billion (116 non-null values).

-Solar.R: Solar radiation in Langley units (146 non-null).

-Wind: Wind speed in miles per hour (complete).

-Temp: Temperature in Fahrenheit(complete).

-Month: Month of the year (ranging from 5 to 9).

-Day: Day of the month (1 to 31).

The database contains some missing values in the Ozone and Solar.R columns. It captures daily air quality information, mainly between May and September.

1-Summatizing the database:

```
> summary(airquality)
     Ozone           Solar.R          Wind             Temp           Month      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000  
 NA's   :37       NA's   :7                                                       
      Day       
 Min.   : 1.0  
 1st Qu.: 8.0  
 Median :16.0  
 Mean   :15.8  
 3rd Qu.:23.0  
 Max.   :31.0  
```

-After summarizing the database using the summary(airquality) command, we observed key statistics for each variable, such as minimum, maximum, mean, and quartiles. To handle missing values, we executed the command is.na(airquality) to identify their locations. The command returns True for null values and FALSE for non-null values, helping us pinpoint where data is missing for further processing.

2-Determine null values:

```
> colSums(is.na(airquality))
  Ozone Solar.R    Wind    Temp   Month     Day 
     37       7       0       0       0       0 
```

-To determine the
total number of null values in each column, we used the command colSums(is.na(airquality)). This allowed us to identify that:

-The Ozone column has 37 missing values.

-The Solar.R columns has 7 missing values.

-The Wind, Temp, Month, and Day columns have no missing values.

Next, we cleaned the dataset by removing rows with null values and displayed the cleaned data for further analysis.

3-Cleaning the dataset:

```
> cleanedAIR = na.omit(airquality)
> cleanedAIR
```

-After removing rows with null values, we re-ran the command colSums(is.na(airquality)) to verify that all missing values had been successfully removed. The output showed that all columns now have 0 null values, confirming that the dataset is clean and ready for further analysis.

4-Compute correlation:

```
> cor( cleanedAIR$Ozone , cleanedAIR$Temp )
[1] 0.6985414
```

-Following the data cleaning process, we computed the correlation between the columns using the cor() function. The correlation results were close to 1,

indicating a strong positive relationship between the variables. Hence, it is suitable to proceed with Linear Regression for modeling and analysis.

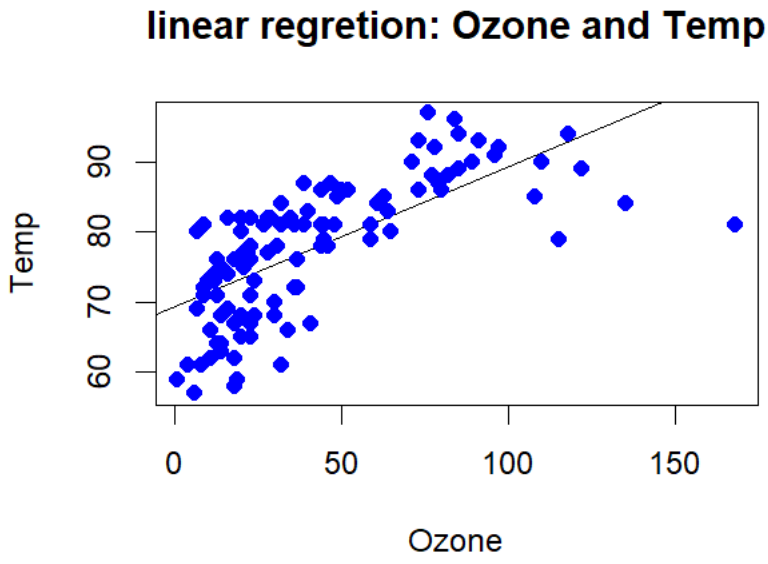
5-Create model & summary it:

```
> model1 = lm ( cleanedAIR$Temp ~ cleanedAIR$Ozone )
> model1

Call:
lm(formula = cleanedAIR$Temp ~ cleanedAIR$Ozone)

Coefficients:
      (Intercept)  cleanedAIR$Ozone
          69.3706            0.2001

> summary(model1)

Call:
lm(formula = cleanedAIR$Temp ~ cleanedAIR$Ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-21.980  -4.775   1.825   4.228  12.425

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      69.37059    1.05151   65.97   <2e-16 ***
cleanedAIR$Ozone  0.20006    0.01963   10.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.851 on 109 degrees of freedom
Multiple R-squared:  0.488,     Adjusted R-squared:  0.4833
F-statistic: 103.9 on 1 and 109 DF,  p-value: < 2.2e-16
```

-Following the data cleaning process and correlation analysis, we created a Linear Regression model and generated a summary of the model using the summary() function. The summary provided insights

into the model's performance, including coefficients, significance levels, and the overall fit of the model.

6-Create Linear Regression model:



-We created the plot for the linear regression model. The independent variable is the temperature, and the dependent variable is the ozone level. The plot demonstrates that temperature affects ozone levels, as an increase in temperature can trigger reactions that produce ozone.

7-Evaluating the model:

```
> summary(model1)$r.squared
[1] 0.4879601
> rmse <- sqrt(mean(residuals(model1)^2))
> print(rmse)
[1] 6.788569
```

-The code evaluates the linear regression model's performance:

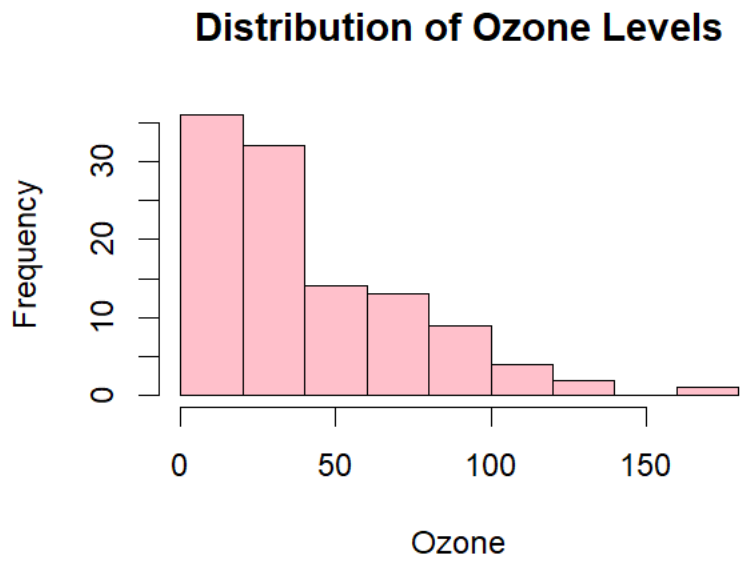
summary(model1)$r.squared: Returns the R-squared value (0.488), indicating that the model explains 48.8% of the variance in ozone levels.

rmse <- sqrt(mean(residuals(model1)^2)): Calculates the Root Mean Square Error (RMSE), which is 6.79. A lower RMSE suggests better model accuracy.

print(rmse): Prints the RMSE value.

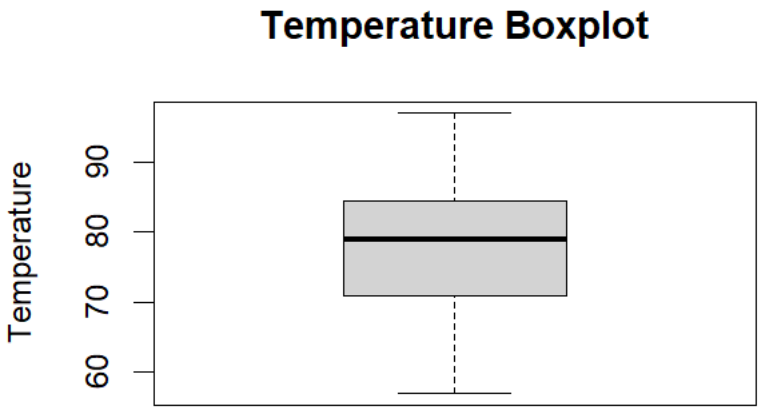
Together, R-squared shows model fit, and RMSE measures prediction error.

8-The histogram "Distribution of Ozone Levels":



-The histogram is skewed to the right, meaning that high Ozone levels are rare, and very high Ozone levels are almost nonexistent. This suggests that most of the data points are clustered at lower Ozone levels, with few instances of elevated Ozone.

9-The box plot:



-In the box plot, we analyze the temperature. The black line represents the median, which is close to 80, indicating that half of the values are below 80 and the other half are above it. From the plot, it's clear that there are no extreme outliers.

Thank you!