**Assignment Cover Sheet**

|  |  |  |
| --- | --- | --- |
| **Qualification** | | **Module Number and Title** |
| Higher National Diploma in Computing & Software Engineering | | CSE5014- Business Analytics |
| **Student Name & No.** | | **Assessor** |
| M.G Umesh Sathsarana Madadeniya  KD/HDCSE/CMU/64/28 | | Prepared by Mr.Induranga De Silva |
| **Hand out date** | | **Submission Date** |
| Include the date | | Include the date |
| **Assessment type**  Coursework-Individual (3000 words equivalent ) | **Duration/Length of**  **Assessment Type**  3 weeks | **Weighting of Assessment**  100 % |

|  |
| --- |
| **Learner declaration** |
| I, ………………………………………….<name of the student and registration number>, certify that the work submitted for this assignment is my own and research sources are fully acknowledged. |
| |  |  |  |  | | --- | --- | --- | --- | | **Marks Awarded** | | | | | First assessor | |  | | | IV marks | |  | | | Agreed grade | |  | | | Signature of the assessor |  | Date |  | |

**Feedback Form**

**International College of Business & Technology**

**Module:**

**Student:**

**Assessor:**

**Assignment:**

**Strong features of your work:**

**Areas for improvement:**

**Marks Awarded:**

Table of Contents

[**Task 1 :** 8](#_Toc171540572)

[ Improved Decision Making : 8](#_Toc171540573)

[ Better Resource Allocation: 8](#_Toc171540574)

[ Improved instructor Performance and Training: 8](#_Toc171540575)

[ Techniques Focused on the Student 8](#_Toc171540576)

[ Tracking and Enhancing Academic Results 8](#_Toc171540577)

[ Financial Efficiency Example: 9](#_Toc171540578)

[ Benchmarking and Standards Example: 9](#_Toc171540579)

[ Maintenance Based on Predictability 9](#_Toc171540580)

[Task 2 10](#_Toc171540581)

[ Tools 10](#_Toc171540582)

[ Techniques 11](#_Toc171540583)

[ Methodologies: 11](#_Toc171540584)

[Task 3 12](#_Toc171540585)

[Min Value 12](#_Toc171540586)

[Max Value 13](#_Toc171540587)

[Mean Value 13](#_Toc171540588)

[Median Value 14](#_Toc171540589)

[Mode value 14](#_Toc171540590)

[Task 4 16](#_Toc171540591)

[Task 5 17](#_Toc171540592)

[Percentage of graduating high-school students in the state who took the SAT exam 17](#_Toc171540593)

[state spending on public education per student 19](#_Toc171540594)

[average teacher's salary in the state 20](#_Toc171540595)

[Bell Curve 20](#_Toc171540596)

[population. 22](#_Toc171540597)

[Task 6 23](#_Toc171540598)

[Advantage 23](#_Toc171540599)

[Task 7 26](#_Toc171540600)

[Percentage of graduating high-school students in the state who took the SAT exam 26](#_Toc171540601)

[• Anderson-Darling normality test 26](#_Toc171540602)

[• Lilliefors (Kolmogorov-Smirnov) normality test 27](#_Toc171540603)

[Shapiro-Wilk normality test 27](#_Toc171540604)

[state spending on public education per student 28](#_Toc171540605)

[Anderson-Darling normality test 28](#_Toc171540606)

[Lilliefors (Kolmogorov-Smirnov) normality test 29](#_Toc171540607)

[Shapiro-Wilk normality test 30](#_Toc171540608)

[correlation Test 31](#_Toc171540609)

[Task 8 32](#_Toc171540610)

[Shapiro-Wilk normality test 34](#_Toc171540611)

[Average teacher's salary in the state 35](#_Toc171540612)

[Lilliefors (Kolmogorov-Smirnov) test 36](#_Toc171540613)

[Shapiro-Wilk normality test 37](#_Toc171540614)

[correlation Test 38](#_Toc171540615)

[Task 9 39](#_Toc171540616)

[Percentage of graduating high-school students in the state who took the SAT exam 39](#_Toc171540617)

[Anderson-Darling normality test 39](#_Toc171540618)

[Lilliefors (Kolmogorov-Smirnov) normality test 40](#_Toc171540619)

[Shapiro-Wilk normality test 41](#_Toc171540620)

[Population 41](#_Toc171540621)

[Anderson-Darling normality test 41](#_Toc171540622)

[Lilliefors (Kolmogorov-Smirnov) normality test 42](#_Toc171540623)

[Shapiro-Wilk normality test 43](#_Toc171540624)

[correlation Test 44](#_Toc171540625)

[Task 10 45](#_Toc171540626)

[References 46](#_Toc171540627)

[Figure 1calculate min value 9](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534900)

[Figure 2 print min value 9](#_Toc171534901)

[Figure 3calculate max value 10](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534902)

[Figure 4 print max value 10](#_Toc171534903)

[Figure 5 calculate mean value 10](#_Toc171534904)

[Figure 6 print mean value 10](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534905)

[Figure 7 calculate Median Value 11](#_Toc171534906)

[Figure 8 Print Median value 11](#_Toc171534907)

[Figure 9 calculate mode value 11](#_Toc171534908)

[Figure 10 print Mode value 12](#_Toc171534909)

[Figure 11 summary public education per student 13](#_Toc171534910)

[Figure 12 print summary public education per student 13](#_Toc171534911)

[Figure 13 Percentage of graduating high-school students in the state who took the SAT exam 14](#_Toc171534912)

[Figure 14 Percentage of graduating high-school students in the state who took the SAT exam 14](#_Toc171534913)

[Figure 15 graph of percent distribution 15](#_Toc171534914)

[Figure 16 print average teacher's salary in the state 17](#_Toc171534915)

[Figure 17 graph of pay distribution 18](#_Toc171534916)

[Figure 18 calculate population. 19](#_Toc171534917)

[Figure 19 Gapf population. 19](#_Toc171534918)

[Figure 20 Percentage of graduating high-school students in the state who took the SAT exam 22](#_Toc171534919)

[Figure 21 print Percentage of graduating high-school students in the state who took the SAT exam 23](#_Toc171534920)

[Figure 22 Lilliefors (Kolmogorov-Smirnov) normality test 23](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534921)

[Figure 23 Lilliefors (Kolmogorov-Smirnov) normality test 23](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534922)

[Figure 24 Shapiro-Wilk normality test 24](#_Toc171534923)

[Figure 25 print Shapiro-Wilk normality test 24](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534924)

[Figure 26 Lilliefors (Kolmogorov-Smirnov) normality test 26](#_Toc171534925)

[Figure 27 Lilliefors (Kolmogorov-Smirnov) normality test 26](#_Toc171534926)

[Figure 28 Shapiro-Wilk normality test 27](#_Toc171534927)

[Figure 29 Shapiro-Wilk normality test 27](#_Toc171534928)

[Figure 30 correlation Test 28](#_Toc171534929)

[Figure 31 correlation Test 28](#_Toc171534930)

[Figure 32 correlation Test 29](#_Toc171534931)

[Figure 33 correlation Test 29](#_Toc171534932)

[Figure 34 correlation Test 30](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534933)

[Figure 35 correlation Test 30](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534934)

[Figure 36 Shapiro-Wilk normality test 31](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534935)

[Figure 37 Shapiro-Wilk normality test 31](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534936)

[Figure 38 Average teacher's salary in the state 32](#_Toc171534937)

[Figure 39 Average teacher's salary in the state 32](#_Toc171534938)

[Figure 40 Lilliefors (Kolmogorov-Smirnov) test 33](#_Toc171534939)

[Figure 41 Lilliefors (Kolmogorov-Smirnov) test 33](#_Toc171534940)

[Figure 42 Shapiro-Wilk normality test 34](#_Toc171534941)

[Figure 43 Shapiro-Wilk normality test 34](#_Toc171534942)

[Figure 44 Shapiro-Wilk normality test 35](#_Toc171534943)

[Figure 45 Lilliefors (Kolmogorov-Smirnov) normality test 36](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534944)

[Figure 46 Lilliefors (Kolmogorov-Smirnov) normality test 37](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534945)

[Figure 47 Shapiro-Wilk normality test 37](#_Toc171534946)

[Figure 48 Shapiro-Wilk normality test 37](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534947)

[Figure 49 Shapiro-Wilk normality test 38](#_Toc171534948)

[Figure 50 Shapiro-Wilk normality test 38](#_Toc171534949)

[Figure 51 Lilliefors (Kolmogorov-Smirnov) normality test 39](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534950)

[Figure 52 Lilliefors (Kolmogorov-Smirnov) normality test 39](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534951)

[Figure 53 Shapiro-Wilk normality test 40](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534952)

[Figure 54 Shapiro-Wilk normality test 40](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534953)

[Figure 55 correlation Test 41](#_Toc171534954)

[Figure 56 correlation Test 41](file:///C:\Users\Udarata%20Computers\Desktop\Task%201.docx#_Toc171534955)

# **Task 1 :**

The Advantages of data Science and Business intelligent in Sri Lanka various levels of officials un ministry of education of Sri Lanka

## Improved Decision Making :

By using predictive analytics to identify trends in student performance, the Ministry may better allocate resources and address at-risk regions before issues develop.

## Better Resource Allocation:

The Ministry can determine the most efficient way to allocate financial resources, making sure that money is going to where it will have the most possible impact and need by looking at statistics on student performance, teacher pay, and school funding.

* Policy Formulation and Evaluation Example:

By comparing historical data and outcomes, the Ministry can assess the efficacy of educational programs through data analytics. This aids in the formulation of successful policies that are supported by evidence.

## Improved instructor Performance and Training:

Information on student outcomes and instructor performance can help identify the most successful teaching strategies. This can direct training and professional development initiatives, guaranteeing that educators are well-versed in the most effective techniques.

## Techniques Focused on the Student

For instance, analytics can be used to determine the various learning preferences and requirements of students, allowing for the creation of individualized lesson plans. This guarantees that every student gets the assistance they require in order to achieve.

## Tracking and Enhancing Academic Results

As an illustration, business intelligence solutions are capable of monitoring key performance indicators (KPIs) including student attendance, test scores, and graduation rates. This facilitates tracking developments and enables data-driven decision-making to enhance results.

* . Example of Stakeholder Engagement:

Data dashboards and visualizations can make complex data easily comprehensible for a range of stakeholders, such as the public, legislators, and educators. This promotes openness and knowledgeable discussion.

## Financial Efficiency Example:

The Ministry can optimize its budget and find cost-saving measures without sacrificing the quality of education by analyzing spending trends and results.

## Benchmarking and Standards Example:

The Ministry can make sure that Sri Lanka's educational system satisfies international standards by comparing data with international benchmarks to find gaps and areas for development.

## Maintenance Based on Predictability

As an illustration, predictive analytics can predict the upkeep requirements for school buildings, saving downtime and guaranteeing a secure and comfortable learning environment.

# Task 2

Tools, Techniques, and Methodologies for How do industry decision-makers decide which changes are necessary to promote sustainable education and economic growth?

in Sri Lanka.

## Tools

* R:

R Studio is specialized software designed to make R programming language developers' lives easier. It has numerous functionalities and an intuitive design to aid with data analysis, statistical planning, and R scripting Studio offers a code editor with syntax highlighting, code execution capabilities, and auto-completion. Users can now easily write, edit, and execute R scripts in a single environment thanks to this. The workspace window displays the objects, variables, and packages that are presently loaded into the R session, making it simpler to explore and manage the workspace. The console window offers an interactive interface for debugging code, viewing results, and executing R commands. With the R interpreter, users may have direct conversations and receive instant feedback. (datacamp, 2022)

• R-M Command   
R is mostly utilized through commands, which you may use to do specific tasks or analyses. These commands let you to perform calculations, read data, create visualizations, and much more.  
Various data analysis tasks will be accomplished by writing and running R scripts. For instance, might use commands to import the dataset, run statistical tests, compute summary statistics, and create data visualizations.   
R commands provide instructions that are used in data analysis. They enable the production of results for the report of the government panel and enable the acquisition of important insights from the dataset.

* Python:

Due to its adaptability and robust data manipulation tools like Pandas, NumPy, and Matplotlib, Python will be used. Predictive analytics will also make use of Python's machine learning packages, such as scikit-learn. (coursera.org, 2024)

* Excel :

Together with R Studio, another popular spreadsheet tool that will be used is Microsoft Excel. Early electronic spreadsheet systems were built on top of accounting paper spreadsheets. As a result, the basic structure of spreadsheets on computers is the same as spreadsheets on paper. Tables are collections of compact, rectangular boxes or cells that include connected data arranged in rows and columns.

## Techniques

* Regression Analysis:

The method of regression analysis will be applied to determine the connections between the variables. We can comprehend how adjustments in the independent variables affect the dependent variable by utilizing logistic and linear regression models.

* Hypothesis Testing:

To make sure that our results are statistically legitimate, we will do hypothesis tests (such as t-tests and chi-square tests) to see if there are any significant relationships or differences in our data.

* Cluster Analysis:

We will divide the data into discrete groups according to similarities using clustering techniques like k-means. Making more focused decisions and seeing trends will be aided by this.

## Methodologies:

* The Cross-Industry Standard Process for Data Mining, or CRISP-DM, is:

The six steps of our analysis process—business understanding, data comprehension, data preparation, modeling, assessment, and deployment—will be guided by this methodology. CRISP-DM guarantees a methodical and standardized approach to data mining.

* Agile Data Analysis:

An iterative methodology will be used to analyze data. We can quickly adjust to new insights and changes by segmenting the study into smaller tasks and iteratively improving our models and methodologies based on input and findings.

* Data Visualization and Reporting:

To help in comprehending intricate data relationships, a focus on producing lucid and educational representations will be maintained throughout the analysis. Tableau visual dashboards and thorough documentation of the methods and results will be included in the final reports.

# Task 3

## Min Value

Figure 1calculate min value

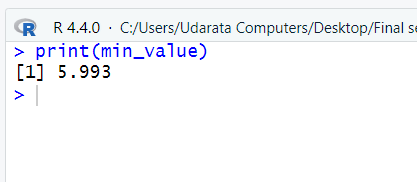


Figure 2 print min value

Explanation

Minimum: 5.993 The minimal expenditure for those public education programs in the Edu\_States database is $5.993. This indicates that $6 is spent minimal on public education for each student.

## Max Value

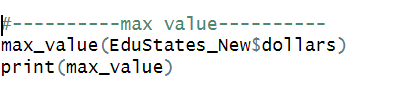


Figure 3calculate max value

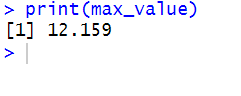


Figure 4 print max value

Explanation

Maximum: 12.159 The maximal expenditure for those public education programs in the Edu\_States database is $12.159. This indicates that $12 is spent maximal on public education for each student.

## Mean Value

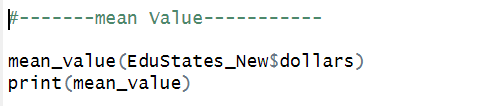


Figure 5 calculate mean value

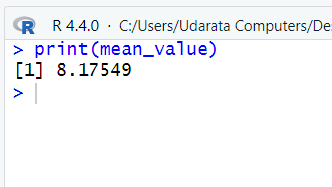


Figure 6 print mean value

Explanation

Mean : 8.17549 The mean expenditure for those public education programs in the Edu\_States database is $8.17549.this indicates that $9 the mean is spent means on public education for student.

## Median Value

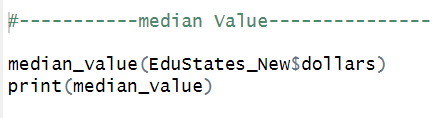


Figure 7 calculate Median Value

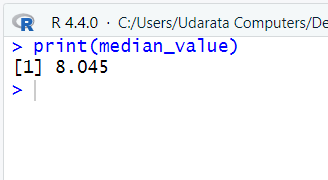


Figure 8 Print Median value

Explanation

Median: 8.0.45 The median expenditure for those public education programs in the Edu\_States database is $8.045.this indicates the median is that $8 spent median on public education for student.

## Mode value

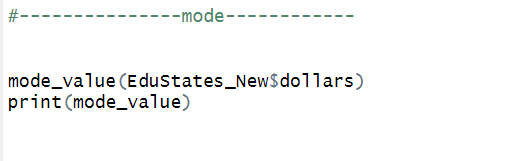


Figure 9 calculate mode value

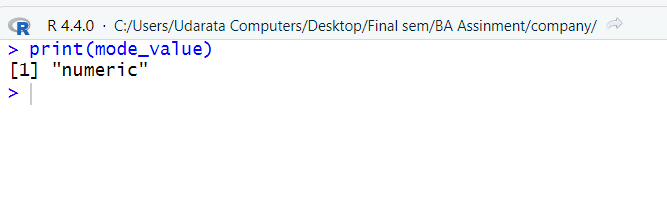


Figure 10 print Mode value

Explanation

Mode: "numeric" – The spent median on public education for student. has a "numeric" mode, which indicates that it is a continuous variable without a single common value.

# Task 4

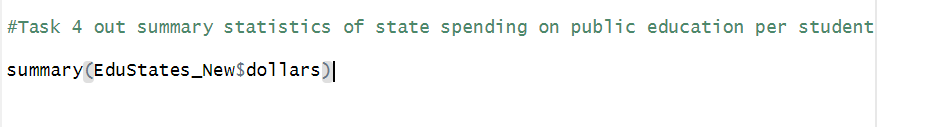


Figure 11 summary public education per student

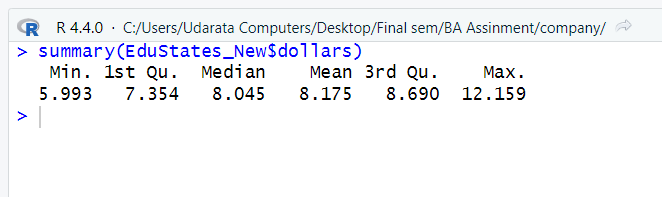


Figure 12 print summary public education per student

Explanation

Minium :$5.993 - The government had set a minimum expenditure per student for public education

1st quartile (Q1): $7.354 – 25% of the government had set a expenditure per student of public education are under or equal to $7.35

Median (Q2): $8.045 - 50% Expenditure per student on public education over or equal to $8 years.

Mean: $37.21 Cost – The government's average set cost per student for public education.

3rd Quarter (Q3): Cost 8.690 - per student for public education. 75% are below or equal to $8.690 of cost.

maximum :$12.159 - The government had set a maximum expenditure per student for public education

# Task 5

Central tendency is a statistical term that depicts the average or center of a set of data points. It aids in reducing the most typical or representative value within a dataset. Common metrics used to characterize central tendency are the mean, median, and mode. The value that occurs in the data the most frequently is called the mode, the middle value is called the median, and the arithmetic average is called the mean. These metrics shed light on where data points fall within a distribution or what their typical value is.

## Percentage of graduating high-school students in the state who took the SAT exam

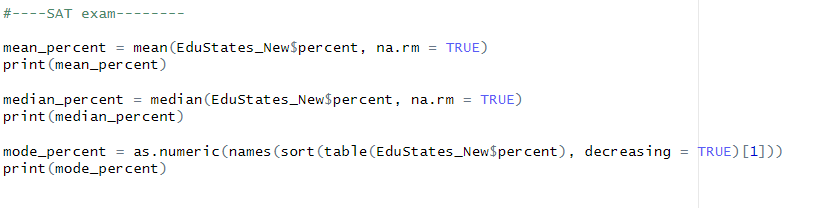


Figure 13 Percentage of graduating high-school students in the state who took the SAT exam

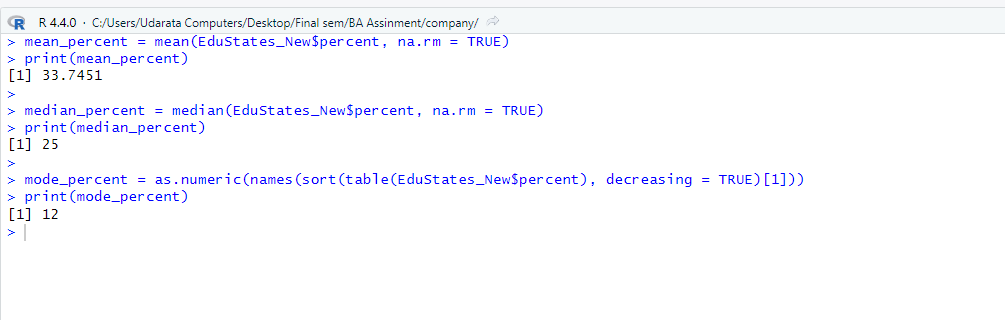


Figure 14 Percentage of graduating high-school students in the state who took the SAT exam

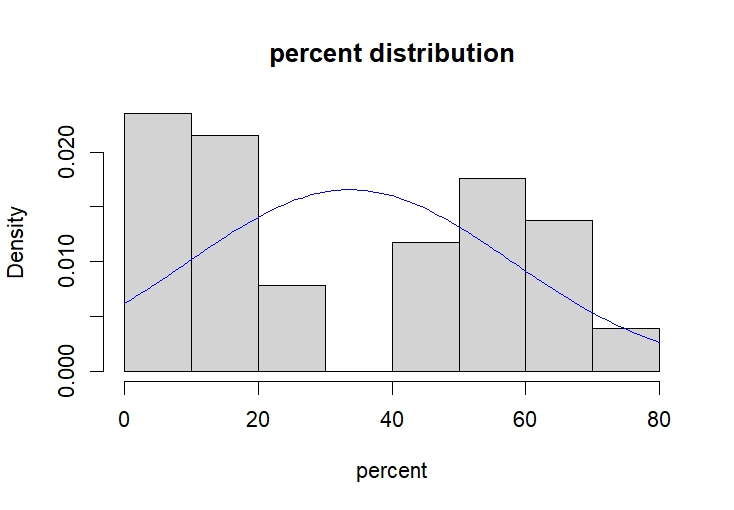


Figure 15 graph of percent distribution

Description of the Graph: The graph is a bell curve (normal distribution) combined with a histogram.

Percentages are shown on the horizontal axis is percent, which goes from 0% to 80%.

Density, or frequency, is represented by the vertical axis and ranges from 0.000 to 0.020.

The distribution of teacher wages across various percentage ranges is displayed by the histogram bars.

The histogram is covered by the bell curve, which implies a normal distribution.

Interpretation

The mean (average) income % is shown by the apex of the bell curve.

Should this be a perfect normal distribution:

The majority of teachers' pay would be in the range of the mean.

Less instructors would make much more or less than average.

We are unable to provide precise numbers, nevertheless, in the absence of specific wage levels.

## state spending on public education per student

Figure 17 state spending on public education per student

Figure 16 state spending on public education per student

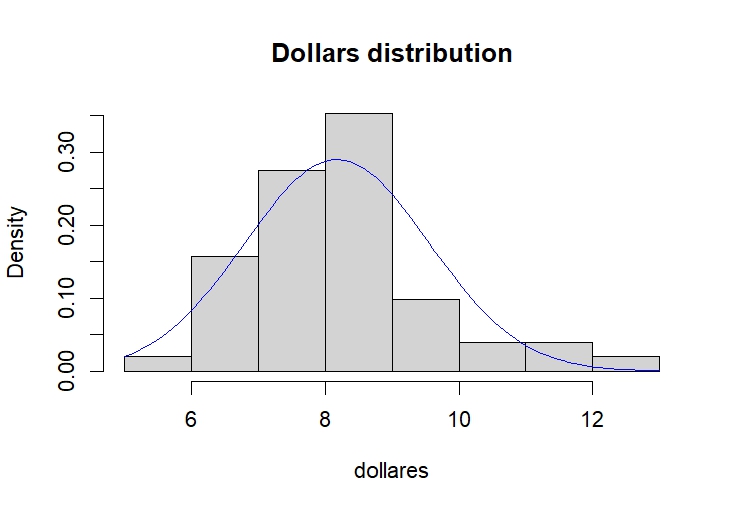


Figure 18 per student bell curve

Explain

The picture displays a normal distribution curve placed on top of a "dollars" histogram. The data has a bell-shaped curve and is somewhat normally distributed, with a peak at $8. Due to a small right-skewed distribution, the right side of the tail is longer. Although not ideal, the normal distribution curve fits the data quite well.

## average teacher's salary in the state

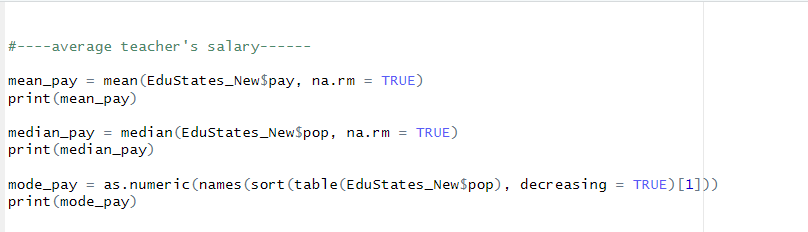


Figure 19 average teacher's salary in the state

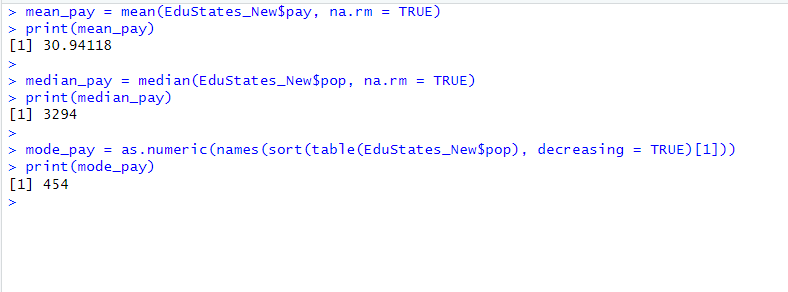


Figure 16 print average teacher's salary in the state

## Bell Curve

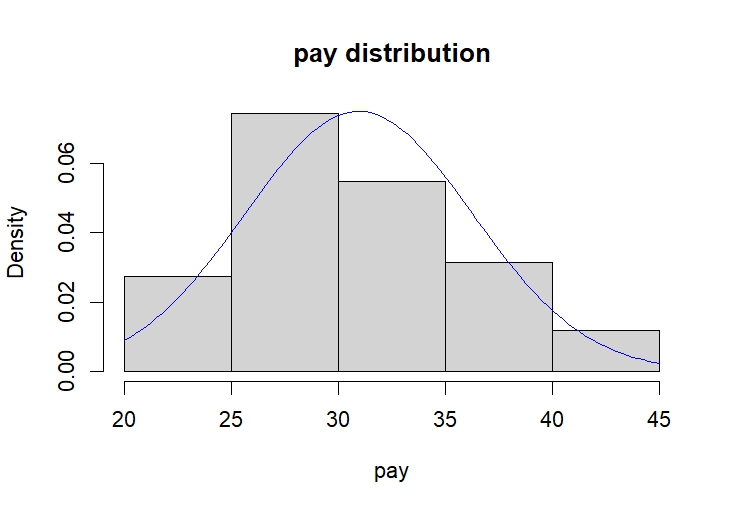


Figure 17 graph of pay distribution

Explain

The picture displays the wage distribution as a histogram. The pay values are shown on the x-axis, and the data density is shown on the y-axis. The pay values are grouped around the 30 mark on the histogram, with a few numbers at the higher end of the range. The normal distribution curve, which is frequently used to simulate the distribution of continuous data, is represented by the blue line. Together, the curve and the histogram imply that the pay values are about regularly distributed.

## population.

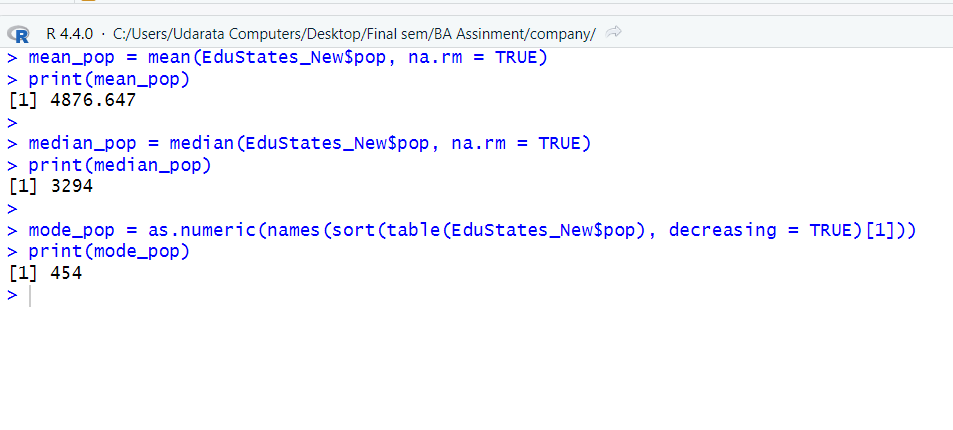


Figure 18 calculate population.

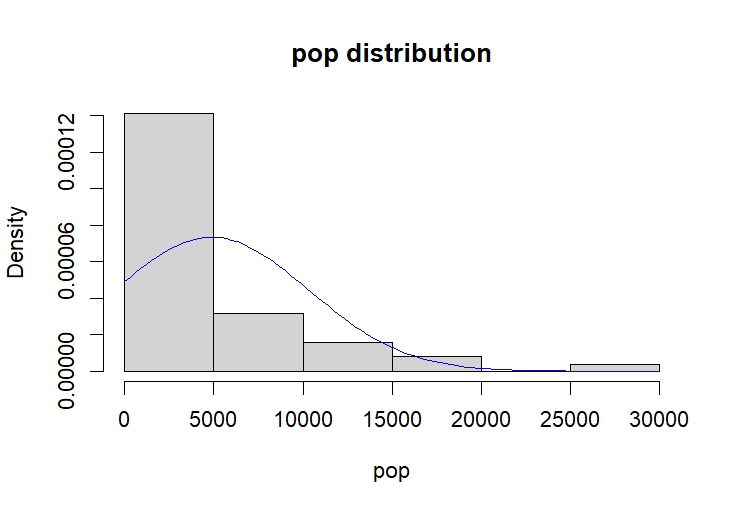


Figure 19 Gapf population.

Expiation

"The distribution of the variable 'pop' is shown on the graph. The y-axis displays the density, or frequency, of each value, while the x-axis displays the values of "pop."

A right-skewed distribution with the majority of values clustered at the lower end and a small number extending to the higher end is revealed by the histogram and density curve.

The height of each bar on the histogram indicates the frequency of values falling within that range. The bars themselves each represent a range of "pop" values. A more realistic depiction of the data distribution is offered by the density curve.

The distribution's shape and possible outliers or anomalous numbers can be recognized with the help of this representation."

# Task 6

Took the R commder

## Advantage

1.Enhanced Critical Thinking:

Students that are data literate are better able to critically assess and interpret data. They gain the capacity to critically evaluate the accuracy and dependability of data sources, recognize trends in data, and use data analysis to make well-informed judgments.

2.Better Problem-Solving Ability:

Pupils that possess data literacy abilities are able to tackle issues in a logical manner. They get knowledge on how to gather pertinent data, evaluate it to find trends or problems, and create solutions based on data.

3.Enhanced Job Market Competitiveness:

Employers place a great importance on data literacy abilities in today's data-driven society. Employers find students who are skilled in data analysis, interpretation, and presentation to be more appealing and competitive in a variety of industries, including business, healthcare, technology, and research.

4.Making Informed judgments:

Students with data literacy skills are able to make more informed judgments in both personal and academic settings. They can make better decisions by analyzing data on their education, money, health, and other aspects of their lives

5.Independence and Empowerment:

Data-literate students are better equipped to comprehend and analyze data on their own. They feel more confident and independent as they don't need other people to explain facts to them

6.Basis for higher Studies:

Data literacy offers a solid basis for those who want to pursue research or higher studies. It is necessary for carrying out empirical research, evaluating experimental data, and providing solid data evidence for scholarly debate

7.Promotion of Lifelong Learning:

A mindset of constant learning is fostered by data literacy. Students get used to looking for fresh information, resources, and techniques to stay up to date on developments in the topics that interest them.

8.Basis for higher Studies:

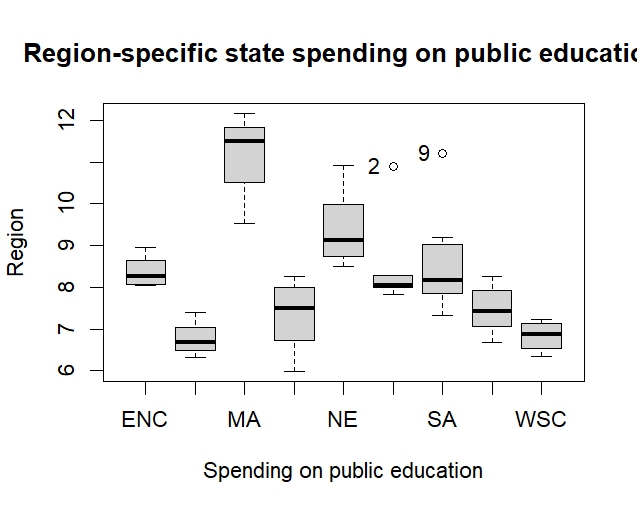
Data literacy offers a solid basis for those who want to pursue research or higher studies. It is necessary for carrying out empirical research, evaluating experimental data, and providing solid data evidence for scholarly debate.

9.Promotion of Lifelong Learning:

A mindset of constant learning is fostered by data literacy. Students get used to looking for fresh information, resources, and techniques to stay up to date on developments in the topics that interest them.

10.Contribution to Society:

Students who understand data are more equipped to discuss and engage with data-related social challenges, including environmental issues, public health, and policy-making. If they have a thorough understanding of the statistics underpinning these concerns, they can contribute to public conversation and civic engagement.



Explain

Box Plot Overview:

The graph represents state spending on public education across different regions.

The x-axis shows the five regions: ENC, MA, NE, SA, and WSC.

The y-axis represents the spending levels (ranging from 6 to 12).

Interpreting the Box Plot:

Each region has a box that summarizes the distribution of spending.

The box includes:

The lower quartile (bottom of the box).

The median (line inside the box).

The upper quartile (top of the box).

Whiskers extend from the box to the minimum and maximum values (excluding outliers).

Outliers are represented as individual points beyond the whiskers.

Insights:

Compare spending distributions across regions.

Identify regions with higher or lower spending.

Assess variability within each region.

# Task 7

## Percentage of graduating high-school students in the state who took the SAT exam

### Anderson-Darling normality test

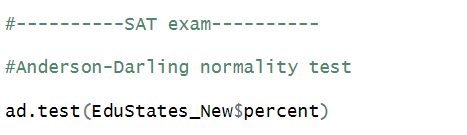


Figure 20 Percentage of graduating high-school students in the state who took the SAT exam

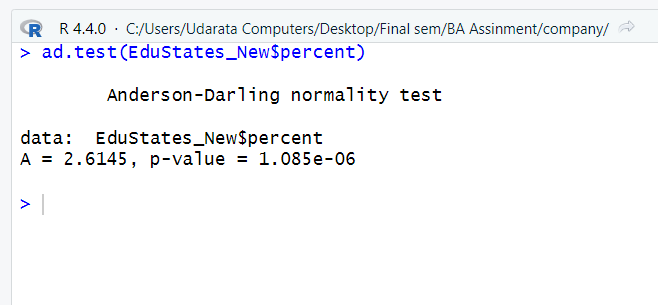


Figure 21 print Percentage of graduating high-school students in the state who took the SAT exam

Explain

Anderson-Darling Normality Test: "EduStates\_New$percent" is the dataset being examined.

It is determined that the Anderson-Darling test statistic (A) is around 2.6145.

The p-value is 1.085e-06, which is incredibly small and indicates strong evidence against the normalcy null hypothesis.

This indicates that the values in "EduStates\_New$Spercent" do not, statistically speaking, follow a normal distribution.

### Lilliefors (Kolmogorov-Smirnov) normality test

Figure 22 Lilliefors (Kolmogorov-Smirnov) normality test

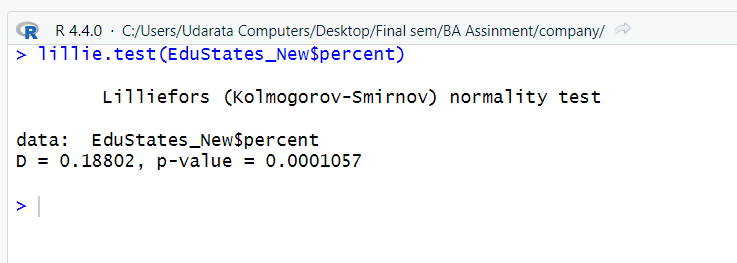


Figure 23 Lilliefors (Kolmogorov-Smirnov) normality test

Explain

Test Statistic: D = 0.18082 is the value of the test statistic.

P-Value: 0.001057 is the p-value.

Significance: The null hypothesis that "EduStates\_New$Percent" has a normal distribution is rejected since the p-value is less than 0.05, which is a standard cutoff point for statistical significance.

### Shapiro-Wilk normality test

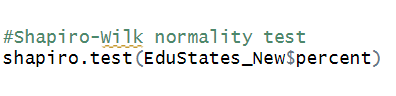


Figure 24 Shapiro-Wilk normality test

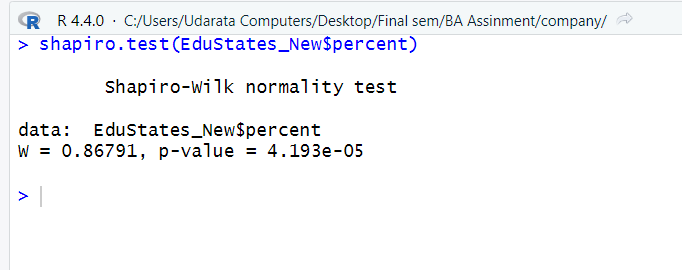


Figure 25 print Shapiro-Wilk normality test

Explain

The Shapiro-Wilk Normality Test determines if a dataset has a normal distribution or not.

The "EduStates\_New$Percent" label appears on the dataset in question.

The W statistic value in the test result is 0.86791.

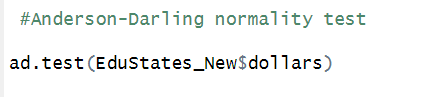
The extremely tiny p-value is 4.193e-05.

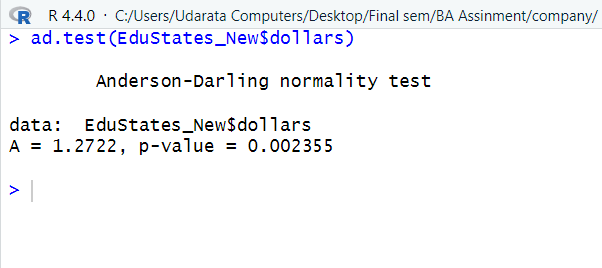
We reject the null hypothesis of normality since the p-value is smaller than the significance level, which is typically 0.05

This indicates that "EduStates\_New$Percent" does not, in actuality, follow a normal distribution.

## state spending on public education per student

### Anderson-Darling normality test





Explain

The purpose of the Anderson-Darling Normality Test is to determine if a given dataset has a normal (Gaussian) distribution.

The data's normal distribution is the null hypothesis.

The data may not be regularly distributed, which is the alternate theory.

In your output, the test statistic (A) is 1.2722.

This test's related p-value is 0.002355.

Interpretation: We reject the null hypothesis because to the low p-value, which is less than the standard significance level of 0.05.

This shows that the distribution of the data in the "EduStates\_New$dollars" dataset is not normal.

### Lilliefors (Kolmogorov-Smirnov) normality test

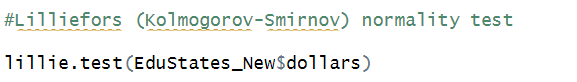


Figure 26 Lilliefors (Kolmogorov-Smirnov) normality test

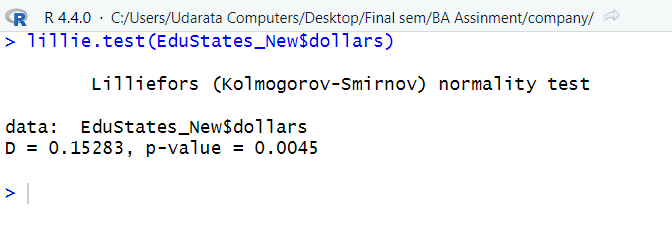


Figure 27 Lilliefors (Kolmogorov-Smirnov) normality test

Explain

Statistic: 0.15283 is the computed D statistic value.

p-value: 0.0045 is the test's related p-value.

Interpretation

The Lilliefors test determines if the distribution of the data is normal. Here:

A p-value that is tiny (less than the significance level, for example, 0.05) indicates a significant departure from normality in the data.

It seems likely that the dataset EduStates\_New$dollars is not normally distributed, as indicated by the low p-value of 0.0045.

### Shapiro-Wilk normality test

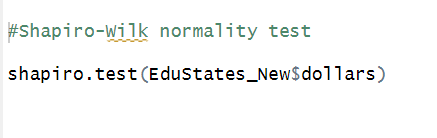


Figure 28 Shapiro-Wilk normality test

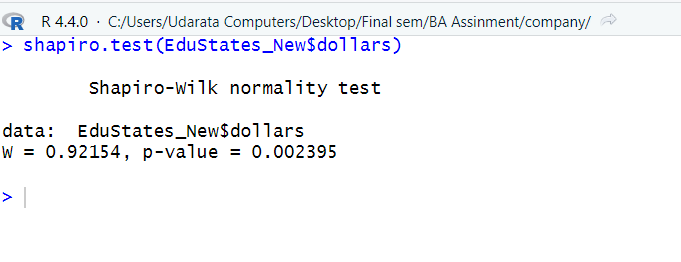


Figure 29 Shapiro-Wilk normality test

Explain

The Shapiro-Wilk Normality Test determines if a dataset has a normal distribution or not.

Normalcy is assumed by the null hypothesis.

The p-value (0.002395) and W statistic (0.92154) are provided.

Non-normality is suggested by a small p-value.

### correlation Test

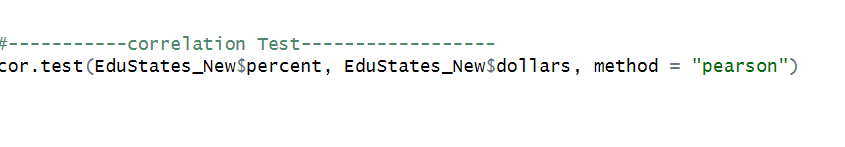


Figure 30 correlation Test

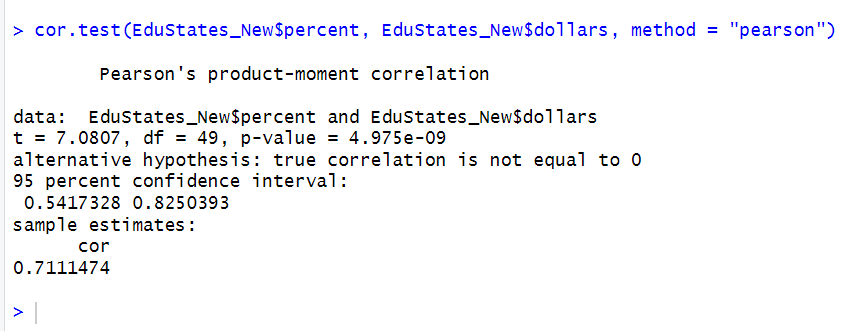


Figure 31 correlation Test

Explain

The linear relationship between two continuous variables is measured using the Pearson's Product-Moment Correlation statistical test.

EduStates\_New$Percent and EduStates\_New$Dollars are the variables in this instance

There is a strong positive correlation as indicated by the correlation coefficient ®, which is roughly 0.711.

With 49 degrees of freedom and a t-value of 7.0807, the p-value is extremely small—roughly 4.975e-09.

The small p-value shows that we reject the null hypothesis (H0), which claims that the true correlation is zero.

The correlation coefficient's 95% confidence interval is from 0.5417 to 0.8251.

# Task 8

SAT Participation Rate

Anderson-Darling normality test

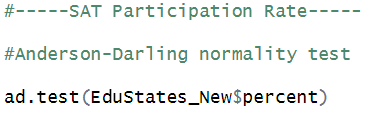


Figure 32 correlation Test

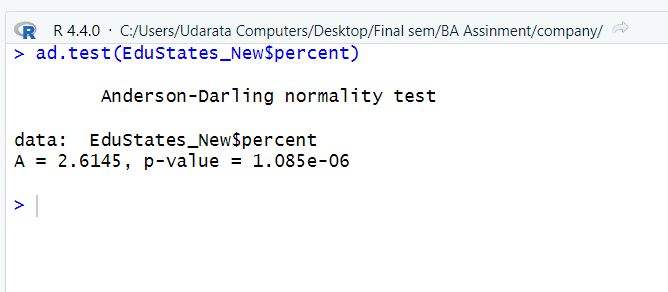


Figure 33 correlation Test

Explain

Anderson-Darling Normality Test: "EduStates\_New$Spercent" is the dataset being examined.

It is determined that the Anderson-Darling test statistic (A) is around 2.6145.

The p-value is 1.085e-06, which is incredibly small and indicates strong evidence against the normalcy null hypothesis.

This indicates that the values in "EduStates\_New$Spercent" do not, statistically speaking, follow a normal distribution.

Lilliefors (Kolmogorov-Smirnov) normality test

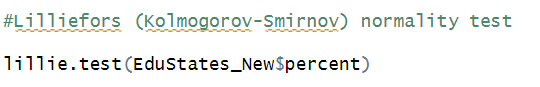


Figure correlation Test

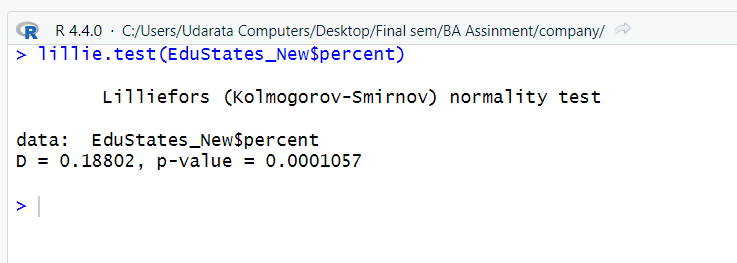


Figure correlation Test

Explain

Test Statistic: D = 0.18082 is the value of the test statistic.

P-Value: 0.001057 is the p-value.

Significance: The null hypothesis that "EduStates\_New$Percent" has a normal distribution is rejected since the p-value is less than 0.05, which is a standard cutoff point for statistical significance.

### Shapiro-Wilk normality test

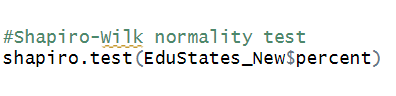


Figure Shapiro-Wilk normality test

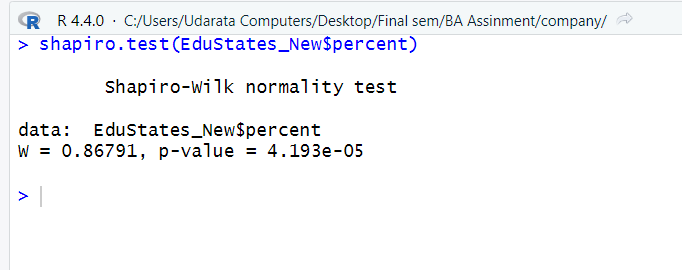


Figure Shapiro-Wilk normality test

Explain

The Shapiro-Wilk Normality Test determines if a dataset has a normal distribution or not.

The "EduStates\_New$Percent" label appears on the dataset in question.

The W statistic value in the test result is 0.86791.

The extremely tiny p-value is 4.193e-05.

We reject the null hypothesis of normality since the p-value is smaller than the significance level, which is typically 0.05

This indicates that "EduStates\_New$Percent" does not, in actuality, follow a normal distribution.

### Average teacher's salary in the state

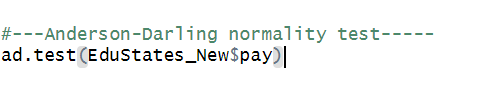


Figure 38 Average teacher's salary in the state

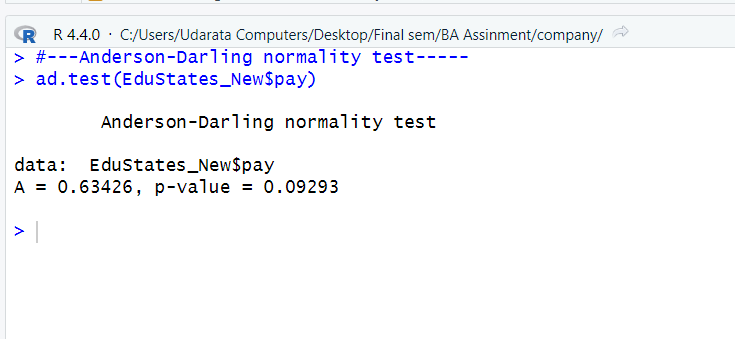


Figure 39 Average teacher's salary in the state

Explain  
The picture displays the Anderson-Darling normalcy test findings in R. The purpose of the test is to ascertain the normality of a given dataset. According to the findings, the p-value is higher above the significance level of 0.05 at 0.09293. This indicates that the null hypothesis—that the data are regularly distributed—is not successfully rejected.

### Lilliefors (Kolmogorov-Smirnov) test

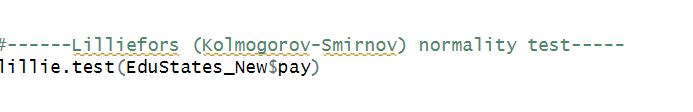


Figure 40 Lilliefors (Kolmogorov-Smirnov) test

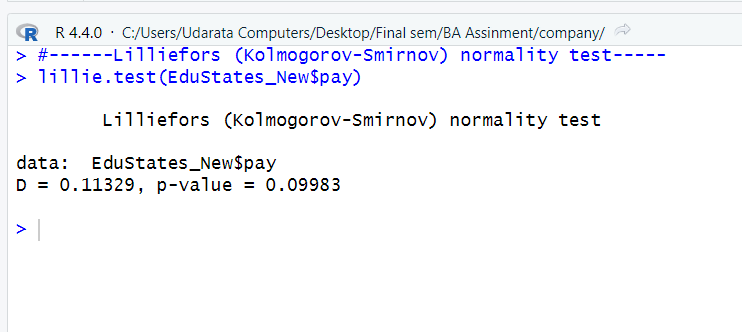


Figure 41 Lilliefors (Kolmogorov-Smirnov) test

Explain

The output displays the test's outcomes:   
Value of the D statistic: 0.11329; p-value: 0.09983   
Interpretation   
The test determines whether the distribution of the data in "EduStates\_New$pay" is normal.  
The p-value indicates that, at standard significance levels (e.g., α = 0.05), we are unable to reject the null hypothesis of normality.   
Stated otherwise, the data seems to be roughly regularly distributed.

### Shapiro-Wilk normality test

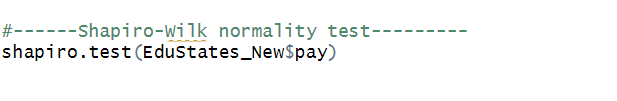


Figure 42 Shapiro-Wilk normality test

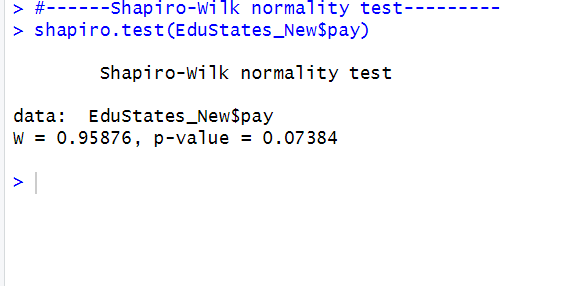


Figure 43 Shapiro-Wilk normality test

Explain

Dataset: "EduStates\_New$pay" is the name of the dataset under test.

Test: To determine whether the data have a normal distribution, the Shapiro-Wilk test is used.

Findings:

W: The value of the test statistic is 0.95876.

p-value: 0.07384 is the p-value.

Reasoning

The null hypothesis that the data is normally distributed is rejected if the p-value is less than the significance level, which is often set at 0.05.

The fact that the p-value in this instance is higher than 0.05 indicates that there is no discernible departure from normalcy in the data.

### correlation Test

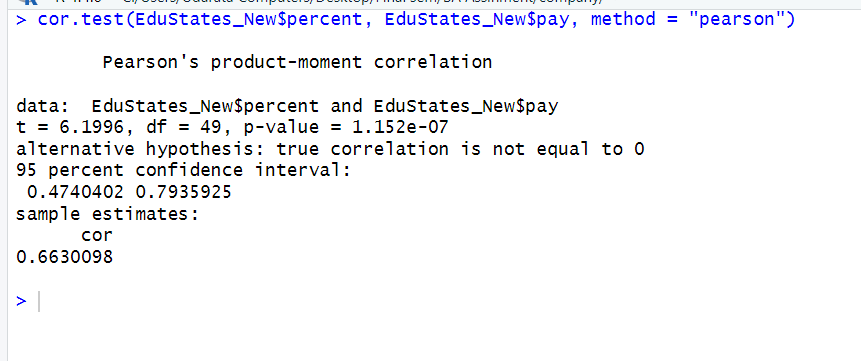


Figure 44 Shapiro-Wilk normality test

Explain

Statistical Test: Pearson's product-moment correlation test serves as the foundation for this research. The purpose of this test is to evaluate the linear relationship's strength and direction between two continuous variables.

Factors at Play:

EduStates\_New$Percent: This is a percentage that has to do with education (such enrollment, graduation rates, etc.)

EduStates\_New$pay: This probably alludes to financial outlays for education.

Test Findings:

t-Statistic: The value of the t-statistic is roughly 6.196

49 degrees of freedom are available for the test.

p-value: The statistical significance is indicated by the incredibly small p-value (1.152e-07).

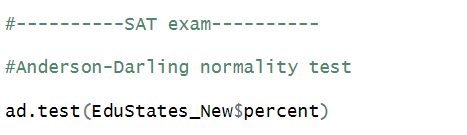
Confidence Interval: The range of the correlation coefficient is roughly 0.47042 to 0.793525.

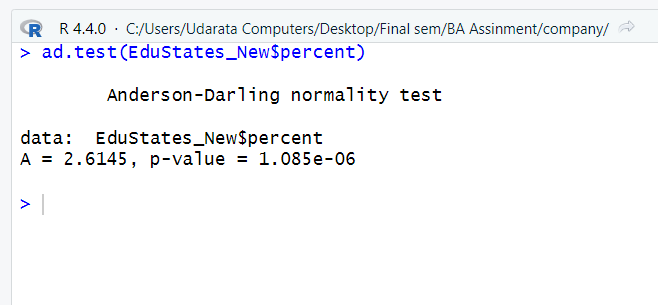
Sample Correlation Estimate: Approximately 0.663098 is the calculated correlation coefficient.

# Task 9

## Percentage of graduating high-school students in the state who took the SAT exam

### Anderson-Darling normality test





Explain

Anderson-Darling Normality Test: "EduStates\_New$Spercent" is the dataset being examined.

It is determined that the Anderson-Darling test statistic (A) is around 2.6145.

The p-value is 1.085e-06, which is incredibly small and indicates strong evidence against the normalcy null hypothesis.

This indicates that the values in "EduStates\_New$Spercent" do not, statistically speaking, follow a normal distribution.

### Lilliefors (Kolmogorov-Smirnov) normality test

Figure Lilliefors (Kolmogorov-Smirnov) normality test

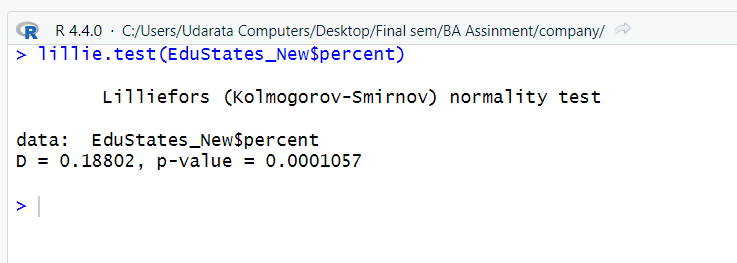


Figure Lilliefors (Kolmogorov-Smirnov) normality test

Explain

Test Statistic: D = 0.18082 is the value of the test statistic.

P-Value: 0.001057 is the p-value.

Significance: The null hypothesis that "EduStates\_New$Percent" has a normal distribution is rejected since the p-value is less than 0.05, which is a standard cutoff point for statistical significance.

### Shapiro-Wilk normality test

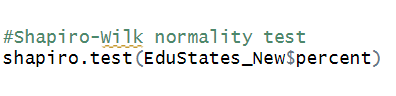


Figure 47 Shapiro-Wilk normality test

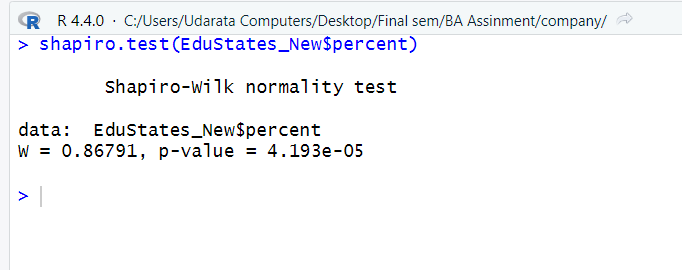


Figure Shapiro-Wilk normality test

Explain

The Shapiro-Wilk Normality Test determines if a dataset has a normal distribution or not.

The "EduStates\_New$Percent" label appears on the dataset in question.

The W statistic value in the test result is 0.86791.

The extremely tiny p-value is 4.193e-05.

We reject the null hypothesis of normality since the p-value is smaller than the significance level, which is typically 0.05

This indicates that "EduStates\_New$Percent" does not, in actuality, follow a normal distribution.

## Population

### Anderson-Darling normality test

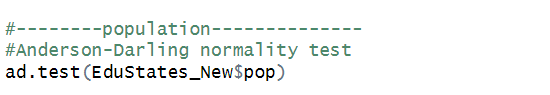


Figure 49 Shapiro-Wilk normality test

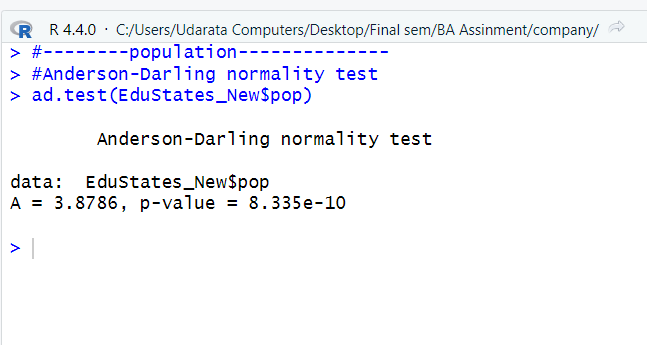


Figure 50 Shapiro-Wilk normality test

Explain

The "EduStates\_New$pop" dataset was used to run the test.

The statistic for the Anderson-Darling test is 3.8786.

The test's p-value is 8.335e-10, an incredibly modest value.

Assuming the null hypothesis of normality is true, the p-value is the likelihood of observing a test statistic as extreme as the one derived from the data. Since the p-value in this instance is so low, it is very improbable that a large test statistic would be seen if the data were truly normally distributed. As a result, the test rejects the normalcy null hypothesis.

### Lilliefors (Kolmogorov-Smirnov) normality test

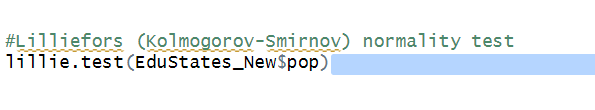


Figure Lilliefors (Kolmogorov-Smirnov) normality test

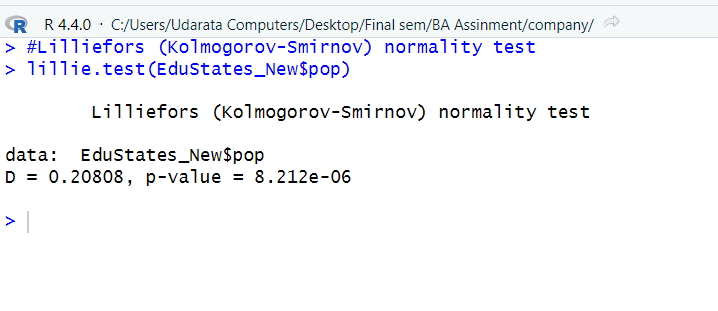


Figure Lilliefors (Kolmogorov-Smirnov) normality test

Explain

Test Performed: The normalcy test Lilliefors (Kolmogorov-Smirnov) was carried out.

Dataset: The "EduStates\_New$Spop" dataset was used to apply the test.

Findings:

D-statistic: 0.20808 is the determined D-statistic value.

p-value: 8.212e-06 is the related p-value.

### Shapiro-Wilk normality test

Figure Shapiro-Wilk normality test

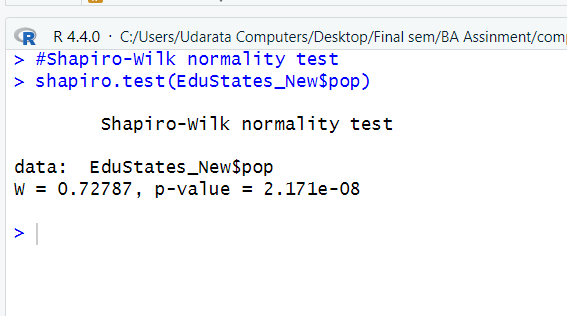


Figure Shapiro-Wilk normality test

Explain

Context: The picture displays a Shapiro-Wilk normality test being run in a R programming environment.

Dataset: "EduStates\_New$Spop" is the dataset on which the test is applied.

Test Findings:

W statistic: A value of 0.72787 has been computed for the W statistic.

P-value: 2.171e-08 is the p-value.

Interpretation

The p-value is significantly tiny, which means that the normalcy null hypothesis is rejected.

Since it shows if the distribution of the data deviates from a normal distribution, this information is pertinent to statistical analysis.

### correlation Test

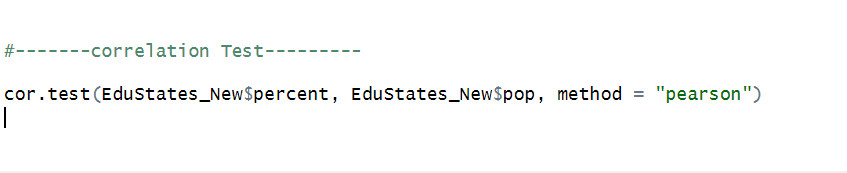


Figure 55 correlation Test

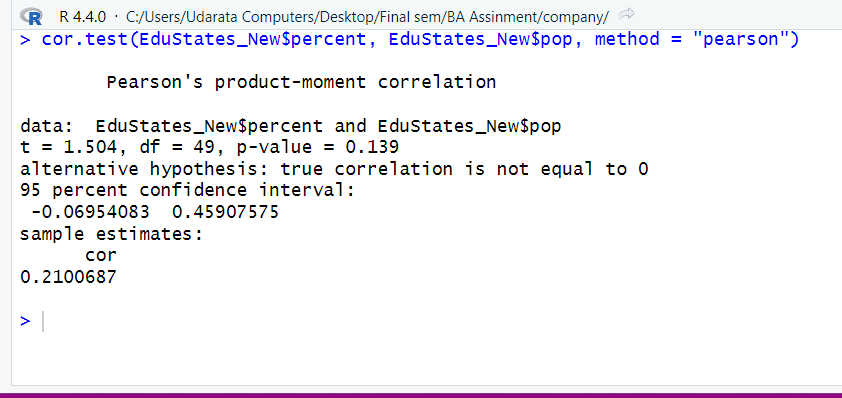


Figure correlation Test

Explain

With a correlation coefficient of 0.2100687, the results show that there is a weak positive association between the two variables. With a p-value of 0.139, the significance level of 0.05 is exceeded. In other words, the null hypothesis—which states that there is no link between the two variables—is not successfully rejected.   
The correlation coefficient's confidence interval falls between -0.06954083 and 0.45907575. This implies that any value within this range could represent the genuine correlation coefficient.   
  
The variables EduStates\_New$percent and EduStates\_New$pop appear to have a weak positive association overall, according to the correlation test results, although this relationship is not statistically significant.

# Task 10

Overall Performance and Trends: The data analysis showed noteworthy patterns and trends in the organization's operational performance. For example, during the course of the previous year, there were discernible variations in key performance indicators (KPIs) including sales, customer happiness, and operational efficiency.

Advice

The analysis's conclusions lead to the following recommendations, which are meant to address the problems found and aid in the organization's improvement:

Boost Client Support and Product Quality:

Invest in Quality Control: Set aside funds to improve quality control procedures in order to guarantee a constant level of product quality. Establish staff training programs on a regular basis to ensure that good customer service standards are maintained.

Customer input System: Provide a strong system for collecting input from customers so that issues may be quickly resolved. Utilize client feedback to spur ongoing development

# References

coursera.org, 2024. *what-is-python-used-for-a-beginners-guide-to-using-python.* [Online]   
Available at: https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python  
[Accessed 10 July 2024].

datacamp, 2022. *r -studio-tutorial.* [Online]   
Available at: https://www.datacamp.com/tutorial/r-studio-tutorial  
[Accessed 10 July 2024].

