# **CSE5014 – Business Analytics**

## **Assignment Cover Sheet**

|  |  |  |
| --- | --- | --- |
| **Qualification** | | **Module Number and Title** |
| Higher National Diploma in Computing &  Software Engineering | | CSE5014- Business Analytics |
| **Student Name & No.** | | **Assessor** |
| R.M.H.P.INDUWARA  ICBT ID –CL/HDCSE/CMU/110/68 | | **Miss Vindya Perera** |
| **Hand out date** | | **Submission Date** |
|  | |  |
| **Assessment type**  Coursework- Individual (3000 words equivalent ) | **Duration/Length of Assessment Type**  3 weeks | **Weighting of Assessment**    100 % |
|  |  |  |

|  |
| --- |
|  |
|  | C:\Users\Induwara\AppData\Local\Microsoft\Windows\Clipboard\HistoryData\{8C096A50-EC63-495B-8FF1-4F3ABDBFD2E4}\{C99843CA-C5C5-4D9D-90E0-E888D58469B4}\ResourceMap\{72E9A361-608C-4752-80C3-0647AB176677} |

|  |  |  |  |
| --- | --- | --- | --- |
| **Marks Awarded** | | | |
| First assessor | |  | |
| IV marks | |  | |
| Agreed grade | |  | |
| Signature of the assessor |  | Date |  |
|  |  |  |  |

|  |
| --- |
|  |
|  | C:\Users\Induwara\AppData\Local\Microsoft\Windows\Clipboard\HistoryData\{8C096A50-EC63-495B-8FF1-4F3ABDBFD2E4}\{C99843CA-C5C5-4D9D-90E0-E888D58469B4}\ResourceMap\{D3F5FFB9-0F28-4846-8C87-4B240A937818} |

**Contents**

[**Task 01** 3](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545816)

[**What is Business Analytics (BA)?** 3](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545817)

[**How does Business Analytics work?** 4](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545818)

[**Advantages and Disadvantages of Business Analytics** 4](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545819)

[**Business analytics vs data Science vs business intelligence** 6](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545820)

[**How the Ministry of Education of Sri Lanka uses business analytics** 7](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545821)

[**Task 02** 14](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545822)

[**Tools** 14](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545823)

[**Techniques** 15](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545824)

[**Methodologies** 16](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545825)

[**Task 03** 17](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545826)

[**Minimum Spending** 18](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545827)

[**Maximum Spending** 19](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545828)

[**Mean Spending** 19](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545829)

[**Median Spending** 20](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545830)

[**Mode Spending** 22](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545831)

[**Task 04** 23](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545832)

[**Summary Statistics** 23](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545833)

[**Descriptive Justifications** 25](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545834)

[**Task 05** 25](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545835)

[**Percentage of Graduating High-School Students Taking SAT** 25](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545836)

[**Percentage of students taking the SAT Bell Curve** 27](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545837)

[**State Spending on Public Education per Student:** 28](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545838)

[**State Spending on Public Education per Student Bell Curve** 30](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545839)

[**Average Teacher's Salary** 31](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545840)

[**Average Teacher's Salary Bell Curve** 32](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545841)

[**Population** 33](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545842)

[**Population Bell Curve** 34](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545843)

[**Task 06** 35](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545844)

[**Numerical Findings** 35](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545845)

[**Graphical Analysis** 39](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545846)

[**Task 07** 45](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545847)

[**Formulate Hypotheses** 45](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545848)

[**Normality Test Results:** 46](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545849)

[**Correlation Analysis:** 47](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545850)

[**Descriptive statistics** 48](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545851)

[**Graphical Representation** 49](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545852)

[**Task 08** 51](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545853)

[**Formulate Hypotheses** 51](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545854)

[**Normality Test Results:** 51](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545855)

[**Correlation Analysis:** 52](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545856)

[**Descriptive statistics** 53](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545857)

[**Graphical Representation** 54](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545858)

[**Descriptive Justifications** 56](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545859)

[**Task 09** 57](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545860)

[**Formulate Hypotheses** 57](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545861)

[**Normality Test Results:** 57](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545862)

[**Correlation Analysis:** 59](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545863)

[**Descriptive statistics** 59](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545864)

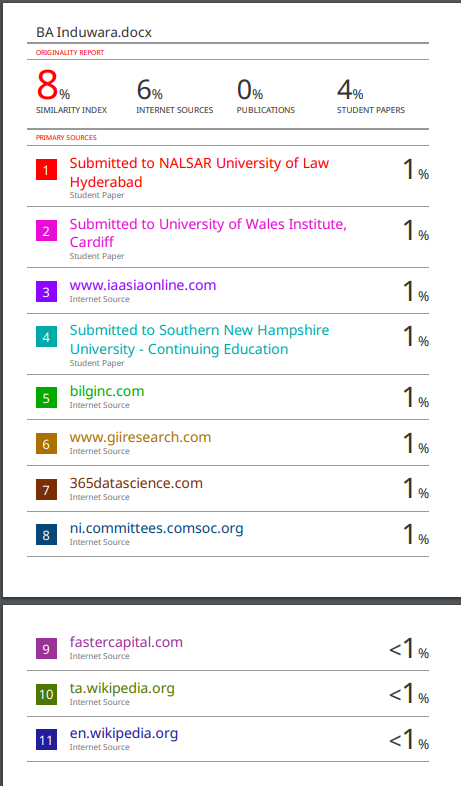
[**Graphical Representation** 60](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545865)

[**Descriptive Justifications** 62](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545866)

[**Task 10** 63](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545867)

[**References** 63](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545868)

[**Appendix** 64](file:///C:\Users\Induwara\Pictures\BA.docx#_Toc171545869)

****

|  |
| --- |
|  |
|  |  |

# **Acknowledgements**

I would like to express my heartfelt gratitude to several individuals who have supported me throughout this assignment. First and foremost, I extend my sincere thanks to Ms. Vindya, my Business Analysis module lecturer, whose guidance and expertise have been invaluable in shaping my understanding of statistical analysis and its applications.

I am also deeply thankful to my parents for their unwavering encouragement and belief in my academic pursuits. Their support has been a constant source of strength and motivation.

Last but not least, I am grateful to my friends for their encouragement and camaraderie throughout this journey. Their insights and discussions have enriched my learning experience and made this assignment all the more rewarding.

Thank you all for your invaluable support and encouragement.

INDUWARA.

# **Introduction**

Understanding the dynamics of educational metrics and their relationship with various socioeconomic factors is crucial in modern educational planning and policy-making. This assignment delves into an exploratory analysis of several key variables affecting high school education in the context of state-level data. Specifically, it investigates the relationships between the percentage of graduating high-school students who took the SAT exam and variables such as state spending on public education per student, teachers' salaries, and state population.

Through statistical tools such as scatter plots and regression analysis, this study aims to uncover insights into how these variables interplay and potentially influence educational outcomes. By examining these relationships, we can gain deeper insights into the factors that contribute to educational performance and equity across different states.

This report begins with an overview of the dataset used, followed by detailed explorations of each relationship through graphical representations and regression analyses. The findings from these analyses will then be discussed, highlighting significant correlations and implications for educational policy and practice.

Overall, this assignment seeks to provide a comprehensive analysis of the factors influencing high school education, contributing to a broader understanding of educational dynamics and informing strategies for improvement and equitable distribution of resources.

# **Task 01**

## **What is Business Analytics (BA)?**

Business Analytics is a systematic methodology used by organizations to extract valuable insights from data, facilitating informed decision-making and strategic planning. This process begins with the collection of data from diverse sources, including internal databases, customer interactions, and external market trends. Once gathered, this data undergoes rigorous processing to ensure accuracy and reliability, preparing it for detailed analysis.

Through the application of statistical techniques, such as regression analysis and hypothesis testing, Business Analytics identifies meaningful patterns and correlations within the data. Additionally, data mining methods uncover hidden relationships and trends, while machine learning algorithms predict future outcomes and optimize operations. These analytical approaches not only reveal actionable insights but also empower companies to refine strategies, improve operational efficiency, and capitalize on emerging opportunities.

Central to Business Analytics is the use of visualization tools like charts and dashboards, which transform complex data into clear, comprehensible formats. These visual representations enable stakeholders to grasp key findings swiftly and make well-informed decisions based on data-driven evidence. Ultimately, by harnessing the power of data, Business Analytics equips organizations with a competitive edge in today's increasingly data-driven business landscape, fostering innovation, agility, and sustainable growth. (Lutkevich, 2024)

## **How does Business Analytics work?**

At first, it begins with the collection and integration of data from multiple sources. Then followed by extra careful analysis using statistical, mathematical, and computational techniques.

Analysts identify patterns, trends, and correlations within the data, often aided by data visualization tools, to derive meaningful insights. These insights are then used to inform strategic decision-making, optimize processes, enhance customer experience, and address challenges or opportunities within an organization. (Advantages and Disadvantages of Business Analytics, 2024)

Finally, continuous monitoring and predictive modelling further clarify Business Analytics as an essential process for organizations wanting to grow their business.

## **Advantages and Disadvantages of Business Analytics**

**Advantages:**

1. **Informed Decision-Making**: Business Analytics empowers organizations to base decisions on data rather than intuition or assumptions. By analyzing historical and real-time data, businesses can uncover patterns, trends, and correlations that inform strategic decisions. This data-driven approach enhances decision accuracy and reduces uncertainty, leading to better outcomes.
2. **Operational Efficiency**: Analyzing operational data helps organizations identify inefficiencies and streamline processes. For example, predictive analytics can forecast demand patterns, allowing businesses to optimize inventory levels and production schedules. This efficiency improvement translates into cost savings, reduced wastage, and improved resource allocation.
3. **Competitive Advantage**: Business Analytics enables businesses to gain a competitive edge by identifying market trends, consumer preferences, and competitor strategies. By leveraging these insights, organizations can innovate faster, develop targeted marketing campaigns, and deliver personalized customer experiences that resonate with their target audience.
4. **Risk Management**: Predictive modeling and risk analytics enable organizations to anticipate potential risks, such as financial instability, supply chain disruptions, or cybersecurity threats. This proactive approach allows businesses to implement mitigation strategies in advance, minimizing the impact of risks on operations and financial performance.
5. **Enhanced Customer Insights**: Analyzing customer data helps organizations understand customer behavior, preferences, and buying patterns. By segmenting customers based on demographics, purchasing history, or interaction patterns, businesses can tailor marketing strategies, improve customer service, and foster stronger customer relationships.

**Disadvantages:**

1. **Data Quality Issues**: Poor data quality, including inaccuracies, inconsistencies, and incomplete data, can undermine the effectiveness of Business Analytics. Cleaning and integrating data from multiple sources can be time-consuming and resource-intensive, impacting the reliability and accuracy of analytical insights.
2. **Complexity and Integration Challenges**: Integrating data from diverse sources, such as CRM systems, ERP platforms, social media, and IoT devices, requires robust data integration strategies and IT infrastructure. Analyzing big data sets and applying advanced analytical techniques, like machine learning and artificial intelligence, may require specialized skills and expertise.
3. **Privacy and Security Concerns**: Handling sensitive or personal data raises concerns about privacy breaches, regulatory compliance (e.g., GDPR, HIPAA), and cybersecurity risks. Organizations must implement rigorous data protection measures, secure data storage, and ensure ethical data handling practices to mitigate these risks.
4. **Over-reliance on Data**: While data-driven insights are valuable, over-reliance on quantitative data may overlook qualitative factors, human judgment, and contextual nuances essential for decision-making. Balancing data-driven insights with human expertise and intuition is crucial for holistic decision-making.
5. **Costs and Resource Allocation**: Implementing Business Analytics initiatives involves significant costs, including investments in data infrastructure, analytics tools, and skilled personnel. Organizations must carefully allocate resources and prioritize initiatives that deliver measurable value and return on investment (ROI).

## **Business analytics vs data Science vs business intelligence**

Business Analytics (BA), Data Science, and Business Intelligence (BI) are distinct yet interrelated fields that leverage data to drive decision-making and improve operations. BA focuses on analyzing historical data to identify trends, patterns, and insights that inform business decisions and optimize processes using techniques like statistical analysis and predictive modeling. Data Science, a broader and more advanced discipline, involves extracting insights from complex and large datasets through methods such as machine learning, big data analytics, and advanced statistical modeling, often aiming to predict future outcomes and uncover deeper insights. BI, on the other hand, centers on the collection, integration, and presentation of business data in user-friendly formats such as dashboards and reports, facilitating real-time monitoring and data-driven decision-making. In essence, while BA is more about understanding past data to improve current operations, Data Science delves deeper into predicting future trends and BI focuses on providing actionable insights through comprehensive data visualization and reporting tools. (ThoughtSpot, 2021)

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspect** | **Business Analytics** | **Data Science** | **Business Intelligence** |
| **Objective** | Analyze historical data to inform business decisions. | Extract insights from complex data for predictive modeling. | Collect, integrate, and present business data for decision-making. |
| **Focus** | Trends, patterns, and statistical analysis. | Machine learning, AI, and big data. | Data visualization, dashboards, and reports. |
| **Techniques** | Statistical analysis, data mining, predictive modeling. | Machine learning, data mining, big data analytics. | Data warehousing, OLAP, reporting, data visualization. |
| **Data** | Structured historical data. | Structured and unstructured data. | Structured current and historical data. |
| **Outcome** | Improved business planning and operations. | Advanced predictive models and deep insights. | Real-time monitoring and actionable insights. |

## **How the Ministry of Education of Sri Lanka uses business analytics**

The Ministry of Education of Sri Lanka can utilize business analytics (BA) in several impactful ways to enhance decision-making, operational efficiency, and educational outcomes. Here are some key areas where BA can be applied: (Department of Examinations - Sri Lanka, 2024) , (Home, 2024) , (Special Notices, 2024) ,

**1. Student Performance Analysis**

**Objective**:

* Analyzing academic performance data to understand trends, strengths, and areas needing improvement across different demographics and regions.

**Implementation**:

* **Data Sources**: Utilize standardized test scores, school assessments, and demographic data.
* **Analytics Techniques**: Employ statistical analysis, data mining, and visualization to identify patterns.
* **Benefits**: Identify underperforming schools or student groups, tailor interventions like remedial programs or enrichment activities, and monitor progress over time.

**2. Resource Allocation and Utilization**

**Objective**:

* Optimizing the allocation of resources such as teachers, textbooks, classrooms, and technology to ensure equitable distribution and maximize efficiency.

**Implementation**:

* **Data Sources**: Enrollment data, school facilities inventory, teacher deployment records.
* **Analytics Techniques**: Predictive modeling to forecast enrollment trends, optimization algorithms for resource allocation.
* **Benefits**: Reduce underutilization of resources, improve access to education facilities, and allocate teachers based on student needs and population shifts.

**3. Curriculum Development and Enhancement**

**Objective**:

* Evaluating the effectiveness of existing curricula and educational programs to enhance learning outcomes.

**Implementation**:

* **Data Sources**: Curriculum adoption data, student performance metrics, feedback from educators and students.
* **Analytics Techniques**: Comparative analysis, correlation studies between curriculum components and student achievement.
* **Benefits**: Inform evidence-based curriculum revisions, introduce new teaching methodologies, and align curriculum with emerging educational trends and global standards.

**4. Teacher Performance and Professional Development**

**Objective**:

* Assessing teacher effectiveness, identifying professional development needs, and enhancing teaching quality.

**Implementation**:

* **Data Sources**: Teacher evaluation results, student feedback, professional development attendance and outcomes.
* **Analytics Techniques**: Performance analytics, sentiment analysis of feedback, correlation between professional development and student outcomes.
* **Benefits**: Tailor professional development programs, mentorship initiatives, and recognition schemes to improve teaching effectiveness and retention.

**5. Education Policy Formulation**

**Objective**:

* Developing evidence-based policies and strategic plans to improve educational quality and equity.

**Implementation**:

* **Data Sources**: National education statistics, policy impact assessments, stakeholder feedback.
* **Analytics Techniques**: Policy simulation models, trend analysis, impact assessments.
* **Benefits**: Align policies with national development goals, allocate funding effectively, and ensure policy outcomes are measurable and sustainable.

**6. Financial Planning and Budgeting**

**Objective**:

* Forecasting budget requirements, managing financial resources efficiently, and ensuring transparency in financial decisions.

**Implementation**:

* **Data Sources**: Budget allocations, expenditure reports, revenue forecasts.
* **Analytics Techniques**: Budget variance analysis, cost-benefit analysis, scenario planning.
* **Benefits**: Optimize resource allocation, prioritize funding for critical areas, and mitigate financial risks through informed decision-making.

**7. Monitoring and Evaluation**

**Objective**:

* Monitoring the implementation and impact of educational initiatives, evaluating program effectiveness, and ensuring accountability.

**Implementation**:

* **Data Sources**: Performance metrics, program evaluation reports, stakeholder surveys.
* **Analytics Techniques**: Key performance indicators (KPIs) tracking, impact evaluations, comparative analysis.
* **Benefits**: Adjust strategies based on real-time data insights, improve program effectiveness, and demonstrate accountability to stakeholders and the public.

**Implementation Considerations**

* **Data Integration**: Ensure seamless integration of data from various educational management systems (e.g., student information systems, HR systems) to create a unified data repository.
* **Capacity Building**: Train personnel in BA tools and methodologies to empower them to interpret data effectively and make data-driven decisions.
* **Privacy and Security**: Implement robust data protection measures to safeguard sensitive information and comply with data privacy regulations, ensuring ethical data handling practices.

**Advantages and Disadvantages of Using Business Analytics for the Ministry of Education of Sri Lanka**

**Advantages**

1. **Informed Decision-Making**:
   * **Example**: By analyzing student performance data across different regions, the ministry can identify which areas require additional resources or targeted interventions. This leads to more precise and effective policy decisions.
   * **Advantage**: Decisions are based on data-driven insights rather than intuition, increasing the likelihood of successful outcomes and efficient resource utilization.
2. **Resource Optimization**:
   * **Example**: Predictive analytics can forecast future student enrollment trends, enabling the ministry to allocate teachers, textbooks, and infrastructure investments more efficiently.
   * **Advantage**: Optimizes the use of limited resources, ensuring that they are directed where they are most needed and reducing waste.
3. **Improved Educational Outcomes**:
   * **Example**: Data analytics can identify patterns and correlations between teaching methods and student success rates, allowing for the refinement of instructional strategies.
   * **Advantage**: Enhances the quality of education by continuously improving teaching practices and curricula based on empirical evidence.
4. **Enhanced Accountability and Transparency**:
   * **Example**: Implementing dashboards and reporting tools that display key performance indicators (KPIs) for various educational programs and initiatives.
   * **Advantage**: Increases transparency, allowing stakeholders to track progress and hold the ministry accountable for educational outcomes.
5. **Early Identification of Issues**:
   * **Example**: Using machine learning to analyze attendance and performance data to identify students at risk of dropping out or failing.
   * **Advantage**: Allows for early interventions, such as counseling or tutoring, to address issues before they become critical.
6. **Customized Learning Experiences**:
   * **Example**: Analyzing data from adaptive learning platforms to understand individual student needs and tailor educational content accordingly.
   * **Advantage**: Provides personalized education, catering to the unique strengths and weaknesses of each student, thereby enhancing their learning experience.
7. **Policy Impact Evaluation**:
   * **Example**: Conducting impact analysis on recent educational reforms or policy changes to measure their effectiveness.
   * **Advantage**: Provides feedback on what works and what doesn’t, allowing for continuous policy improvement.

**Disadvantages**

1. **Data Quality Issues**:
   * **Example**: Inconsistent or incomplete data from different schools and regions can lead to inaccurate analysis and flawed conclusions.
   * **Disadvantage**: Poor data quality can undermine the reliability of analytics, leading to misinformed decisions.
2. **High Implementation Costs**:
   * **Example**: Investing in advanced analytics software, hardware infrastructure, and training programs for staff.
   * **Disadvantage**: Significant initial and ongoing costs may strain the ministry’s budget, especially in a developing country context.
3. **Complexity of Integration**:
   * **Example**: Integrating data from various legacy systems and databases used by different schools and departments.
   * **Disadvantage**: Complex integration processes can be time-consuming and require specialized skills, potentially delaying the implementation of analytics initiatives.
4. **Privacy and Security Concerns**:
   * **Example**: Handling sensitive student and staff data raises concerns about data breaches and compliance with privacy regulations.
   * **Disadvantage**: Requires robust security measures and strict adherence to privacy laws, which can be challenging to implement and maintain.
5. **Over-Reliance on Data**:
   * **Example**: Making decisions solely based on quantitative data without considering qualitative insights from teachers and students.
   * **Disadvantage**: May overlook important contextual factors and human judgment, leading to incomplete or skewed decision-making.
6. **Resistance to Change**:
   * **Example**: Teachers and administrators may be resistant to adopting new data-driven practices and tools.
   * **Disadvantage**: Cultural resistance can hinder the successful implementation and utilization of business analytics.
7. **Ethical Concerns**:
   * **Example**: Using predictive analytics to track student performance could lead to unintended consequences, such as labeling or bias.
   * **Disadvantage**: Raises ethical issues about fairness, bias, and the potential for misuse of data.

# **Task 02**

**Tools, Techniques, and Methodologies for Analysis**

In this case study for the Ministry of Education of Sri Lanka, we will utilize a range of tools, techniques, and methodologies to conduct a comprehensive analysis of educational data. These approaches are essential for deriving insights, making data-driven decisions, and improving educational outcomes.

## **Tools**

1. **R Studio**
   * **Purpose**: R Studio is an integrated development environment (IDE) for R, a programming language widely used for statistical computing and data analysis.
   * **Usage**: R Studio will be instrumental in data manipulation, statistical analysis, and visualization. It allows us to perform complex statistical computations, create detailed plots and graphs, and conduct sophisticated analyses.
   * **Example**: Analyzing trends in student performance across different regions based on SAT scores using statistical tests and visualizations.
2. **Tableau**
   * **Purpose**: Tableau is a powerful data visualization tool that enables interactive and dynamic data exploration through dashboards and reports.
   * **Usage**: It will be used to create visually appealing and insightful dashboards that summarize key educational metrics. Tableau facilitates the exploration of trends, patterns, and outliers in educational data, making it easier to communicate findings to stakeholders.
   * **Example**: Developing dashboards to visualize school performance metrics such as SAT scores, education spending per student, and teacher salaries across different regions.
3. **Python (with pandas and matplotlib)**
   * **Purpose**: Python is a versatile programming language. Pandas is a library for data manipulation and analysis, while matplotlib is used for data visualization.
   * **Usage**: Python and its libraries will be employed for data preprocessing, exploratory data analysis (EDA), and basic statistical analysis. Pandas will handle data cleaning and transformation tasks, while matplotlib will be used to create visual representations of data trends.
   * **Example**: Using pandas to clean and prepare datasets on student demographics and educational expenditures for further analysis in R Studio.
4. **SQL**
   * **Purpose**: SQL (Structured Query Language) is used for managing and querying relational databases.
   * **Usage**: It will be utilized to extract relevant datasets from educational databases, perform data transformations, and join different tables to integrate data from multiple sources.
   * **Example**: Writing SQL queries to retrieve student enrollment data, teacher information, and financial expenditures from the ministry's databases for analysis.
5. **Microsoft Excel**
   * **Purpose**: Excel is a widely-used tool for data analysis and reporting, offering functionalities such as pivot tables, charts, and formulas.
   * **Usage**: Excel will be used for initial data exploration, simple calculations, and creating basic visualizations. It helps in summarizing data and identifying trends before more in-depth analysis.
   * **Example**: Using Excel to organize and summarize survey data on student perceptions and preferences regarding educational programs and services.

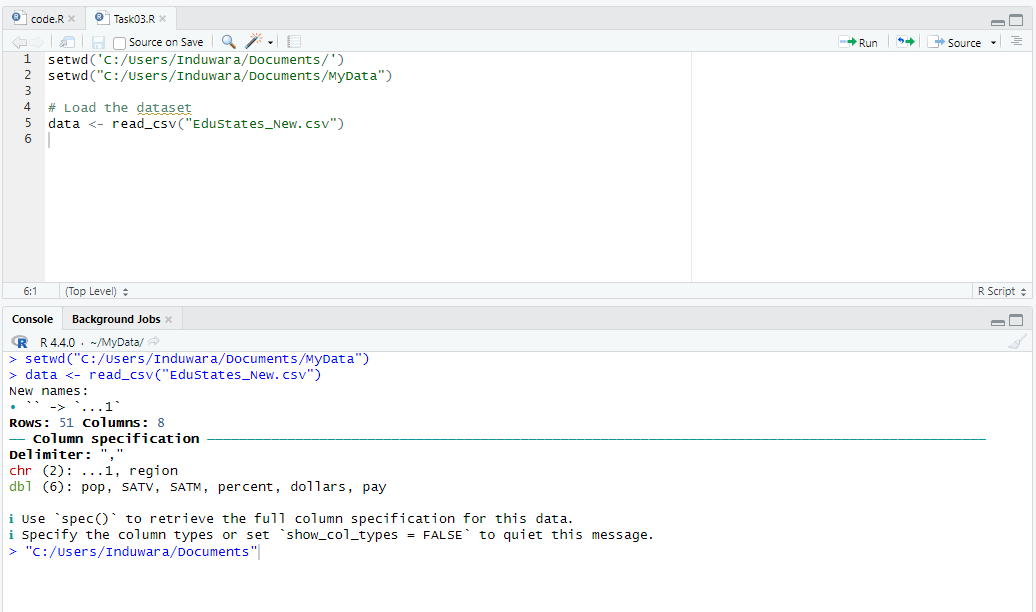
## **Techniques**

1. **Descriptive Analytics**
   * **Purpose**: Summarizes historical data to describe what has happened in the past.
   * **Usage**: Techniques include calculating summary statistics (mean, median, mode), creating frequency distributions, and generating visualizations (e.g., histograms, bar charts) to understand the distribution and characteristics of educational data.
   * **Example**: Summarizing student performance data to identify average scores, variability, and distributions across different demographic groups.
2. **Exploratory Data Analysis (EDA)**
   * **Purpose**: Analyzes data sets to summarize their main characteristics and uncover patterns, relationships, or anomalies.
   * **Usage**: Techniques include visualizing data through scatter plots, box plots, and correlation matrices, as well as conducting statistical tests to explore relationships between variables.
   * **Example**: Using EDA to explore correlations between educational spending per student and student achievement outcomes in standardized tests.
3. **Data Visualization**
   * **Purpose**: Presents data visually to facilitate understanding and communicate insights effectively.
   * **Usage**: Techniques include creating interactive dashboards, charts, graphs, and heat maps using tools like Tableau and matplotlib to visualize trends, comparisons, and geographical variations in educational metrics.
   * **Example**: Developing a geographic heat map to visualize regional disparities in educational funding and their impact on student performance.

## **Methodologies**

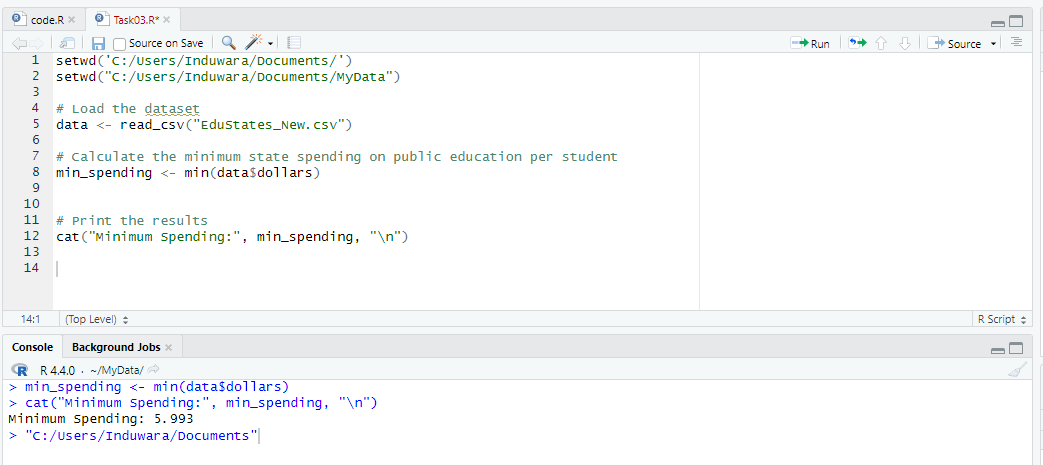
1. **CRISP-DM (Cross-Industry Standard Process for Data Mining)**
   * **Purpose**: A structured approach for planning and executing data analysis projects.
   * **Usage**: Steps include business understanding, data understanding, data preparation, modeling, evaluation, and deployment. CRISP-DM provides a systematic framework to guide the entire analysis process from defining goals to presenting findings.
   * **Example**: Following CRISP-DM to analyze educational data, starting from understanding stakeholder requirements to evaluating model performance and deploying insights.
2. **Data Wrangling and Cleaning**
   * **Purpose**: Prepares raw data for analysis by addressing quality issues, inconsistencies, missing values, and outliers.
   * **Usage**: Techniques include data cleaning, data transformation, and feature engineering to ensure data integrity and reliability for analysis.
   * **Example**: Cleaning and transforming student enrollment data to standardize formats and remove duplicate records before merging with other datasets.
3. **Reporting and Dashboarding**
   * **Purpose**: Communicates insights and findings through clear and actionable reports, visualizations, and interactive dashboards.
   * **Usage**: Techniques include creating summary reports, narrative descriptions, and interactive visualizations that cater to different levels of stakeholders, from policymakers to educators.
   * **Example**: Designing a comprehensive dashboard that presents key performance indicators (KPIs) such as graduation rates, teacher retention rates, and student demographics in an intuitive format.

# **Task 03**



The R script sets the working directory to "C:/Users/Induwara/Documents/MyData" and loads a CSV file named "EduStates\_New.csv" into a data frame using the read\_csv function. The console output shows that the CSV file has 52 rows and 8 columns, with two character columns and six numeric columns. Unnamed columns are automatically renamed, and suggestions are provided for retrieving full column specifications or quieting the column type messages.

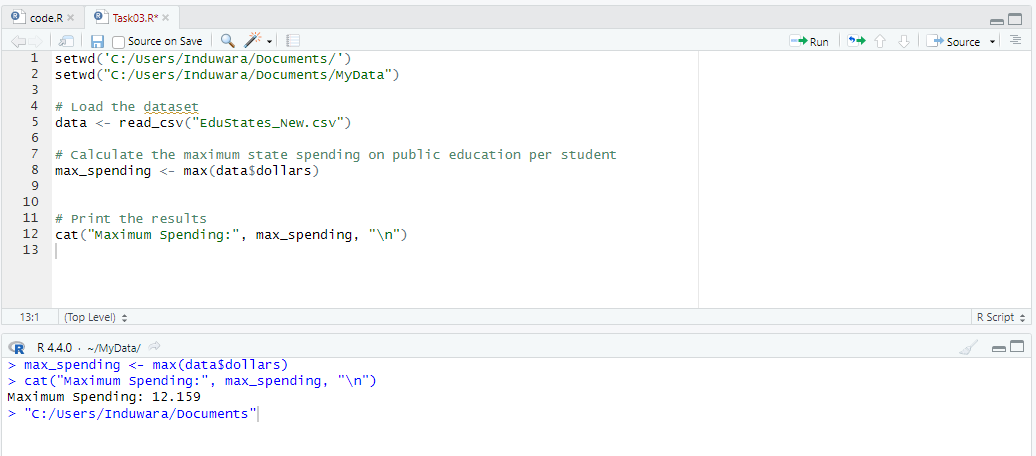
## **Minimum Spending**



* Value**:** $5.993
* min\_value <- min(state\_spending, na.rm = TRUE)

**Explanation:** The min() function calculates the smallest value in the state\_spending column. The na.rm = TRUE argument tells the function to ignore any missing values (NA) during the calculation. The minimum value represents the lowest amount of money spent per student by any state in the dataset. The minimum spending figure represents the state that allocates the least amount of funds per student for public education. This could be due to several factors, including budget constraints, lower cost of living, or prioritization of funds in other sectors. States with minimal spending might face challenges in providing quality education and sufficient resources to students and teachers.

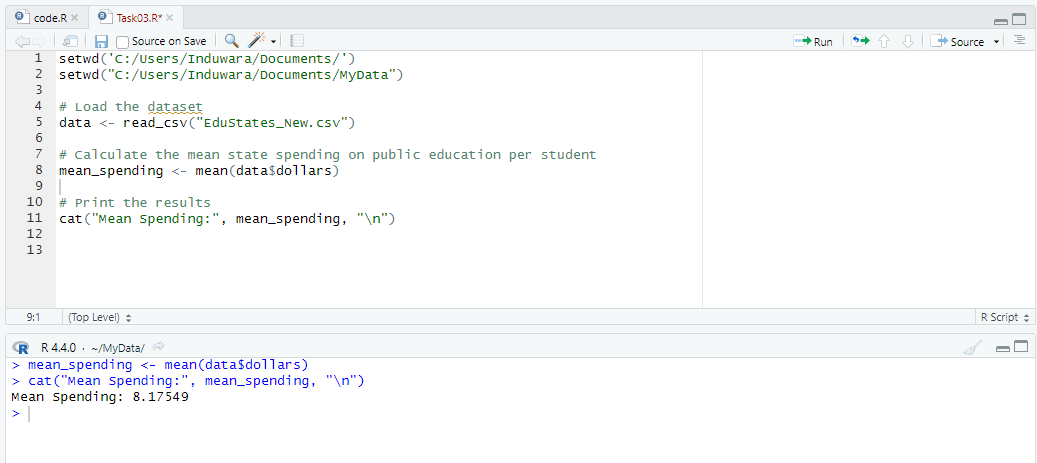
## **Maximum Spending**



* **Value:** $12.159
* max\_value <- max(state\_spending, na.rm = TRUE)

**Explanation:** The max() function calculates the largest value in the state\_spending column. Similar to the minimum calculation, the na.rm = TRUE argument is used to ignore missing values. The maximum value represents the highest amount of money spent per student by any state in the dataset. The maximum spending figure represents the state that allocates the most funds per student for public education. Higher spending can indicate a higher cost of living in the state, greater investment in educational infrastructure, or a strong commitment to improving educational outcomes. States with higher spending often have more comprehensive educational programs, better facilities, and more support for teachers and students.

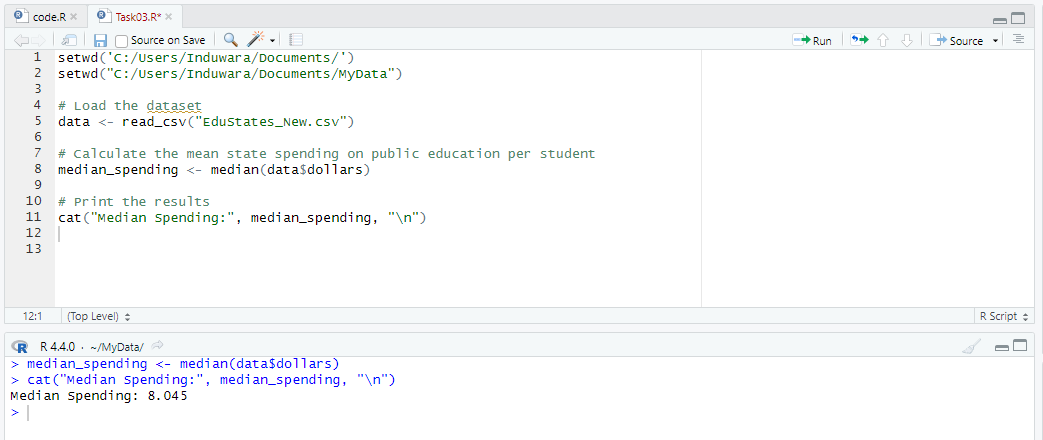
## **Mean Spending**



* **Value:** $8.175
* mean\_value <- mean(state\_spending, na.rm = TRUE)

**Explanation:** The mean() function calculates the average value of the state\_spending column. This is done by summing all the spending values and then dividing by the number of values (excluding any missing values). The mean provides an overall average of how much is spent per student across all states in the dataset.The mean spending figure provides an overall average of what states are spending on public education per student. This value gives a general idea of the average investment in education across all states. While it is useful for understanding the central tendency, it can be influenced by extreme values (outliers), either very high or very low.

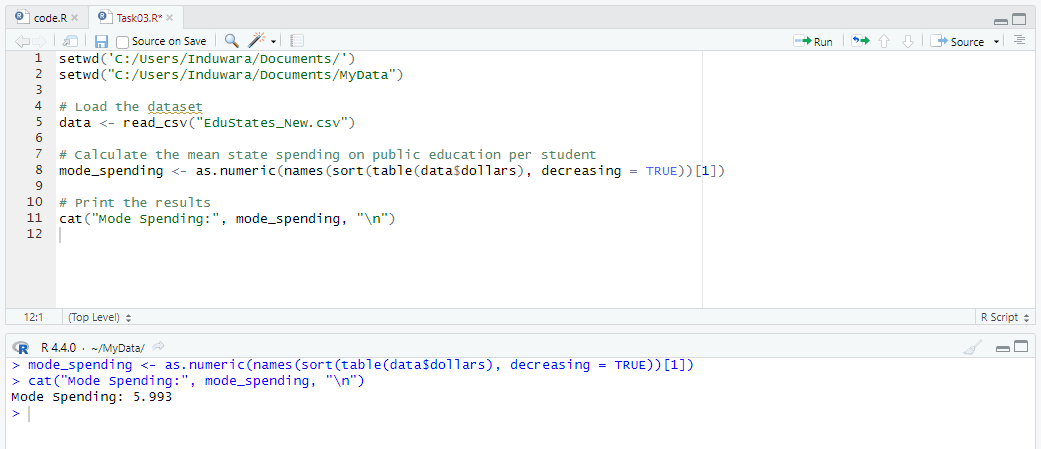
## **Median Spending**



* **Value:** $8.045
* median\_value <- median(state\_spending, na.rm = TRUE)

**Explanation:** The median() function finds the middle value of the state\_spending column when the values are sorted in ascending order. If there is an odd number of values, the median is the middle one. If there is an even number of values, the median is the average of the two middle values. The median represents the point at which half of the states spend less per student and half spend more, providing a measure of central tendency that is less affected by extreme values (outliers) than the mean. The median spending figure represents the middle value in the distribution of spending per student, where half of the states spend less and half spend more. The median is a robust measure of central tendency as it is not affected by outliers. It gives a more accurate picture of typical spending than the mean when the data is skewed.

## **Mode Spending**



* **Value:** $5.993
* mode\_value <- modes(state\_spending)

**Explanation:** The modes() function from the modes package is used to calculate the most frequently occurring value(s) in the state\_spending column. The mode represents the spending amount(s) that appear most often in the dataset. There can be more than one mode if multiple values have the same highest frequency. The mode spending figure represents the most frequently occurring value in the dataset. In this case, the mode is equal to the minimum spending, suggesting that several states cluster around the lower end of the spending spectrum. This could indicate that a significant number of states allocate minimal resources per student, potentially due to similar budgetary constraints or policy choices.

# **Task 04**

## **Summary Statistics**

**Minimum Spending:** $5.993

* **Justification:** The minimum spending represents the state with the lowest investment per student. This low figure might be indicative of budget constraints, lower cost of living, or differing prioritizations in state budgets. States with minimal spending might face challenges in providing adequate educational resources and support to their students and teachers.

**Maximum Spending:** $12.159

* **Justification:** The maximum spending indicates the state with the highest investment per student. This figure could reflect a higher cost of living, a strong commitment to education, or greater overall resources available for educational funding. States with high spending levels often have more comprehensive educational programs and better facilities, leading to potentially better educational outcomes.

**Mean (Average) Spending:** $8.175

* **Justification:** The mean spending provides an overall average of the states' investment in education per student. This value is useful for understanding the general trend but can be influenced by extreme values (outliers). It offers a baseline for comparing individual state spending against the national average.

**Median Spending:** $8.045

* **Justification:** The median spending is the middle value when the spending amounts are ordered from lowest to highest. It is not affected by outliers and provides a more accurate representation of typical state spending. Half of the states spend less than the median, and half spend more, making it a useful measure for identifying the central tendency of the data.

**Mode Spending:** $5.993

* **Justification:** The mode spending is the most frequently occurring value in the dataset. In this case, it matches the minimum spending, suggesting that a number of states cluster around the lower end of the spending spectrum. This could highlight a commonality among states with limited educational budgets or similar economic conditions.



**Summary Statistics:**

* **Min:** $5.993
* **1st Quartile:** The lower quartile (25th percentile) value provides insight into the lower end of the spending distribution.
* **Median (2nd Quartile):** $8.045
* **Mean:** $8.175
* **3rd Quartile:** The upper quartile (75th percentile) value provides insight into the higher end of the spending distribution.
* **Max:** $12.159

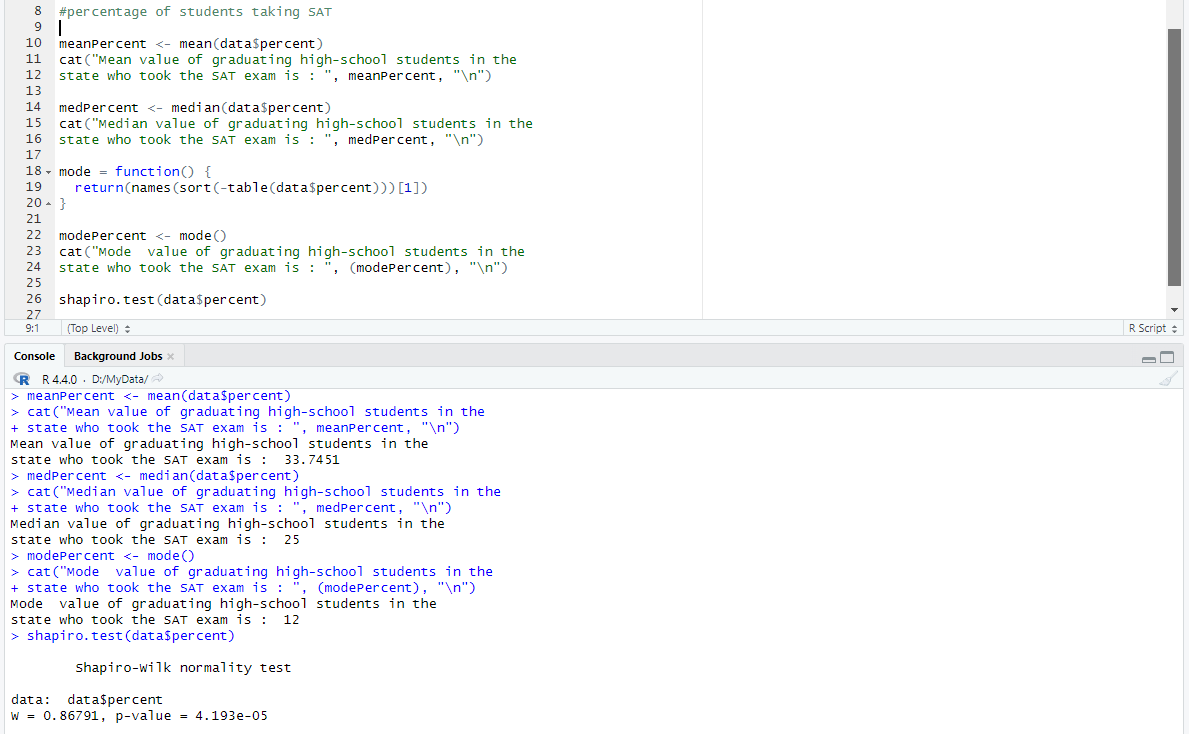
## **Descriptive Justifications**

* **Interquartile Range (IQR):** The difference between the first quartile and the third quartile values indicates the spread of the middle 50% of the data. A larger IQR suggests greater variability in spending among the states.
* **Range:** The difference between the minimum and maximum values provides an understanding of the overall spread of the data. A larger range indicates a wider disparity in spending levels among states.
* **Skewness:** If the mean is higher than the median, the data might be right-skewed, indicating a few states with very high spending. Conversely, if the mean is lower than the median, the data might be left-skewed, indicating a few states with very low spending.
* **Outliers:** Identifying outliers in the data can highlight states with exceptionally high or low spending compared to the rest. These outliers can significantly influence the mean and require further investigation to understand the underlying reasons.

# **Task 05**

## **Percentage of Graduating High-School Students Taking SAT**

Central tendency analysis for the percentage of graduating high-school students in the state who took the SAT exam



**Mean:**

* The mean percentage of graduating high-school students who took the SAT exam is calculated by summing all the percentages and dividing by the number of states.
* The mean provides an overall average and is useful for understanding the general trend in the data.
* Provides the overall average percentage of students taking the SAT, giving a general idea of the participation rate.

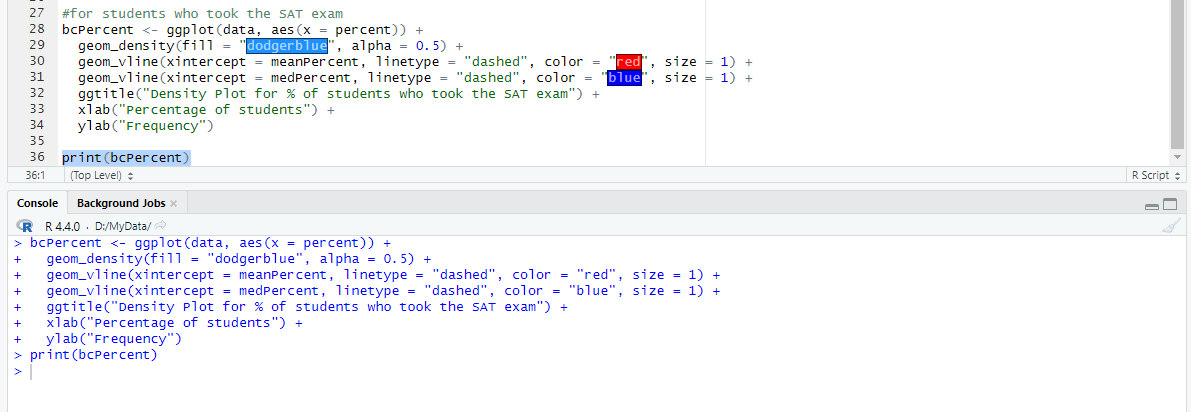
**Median:**

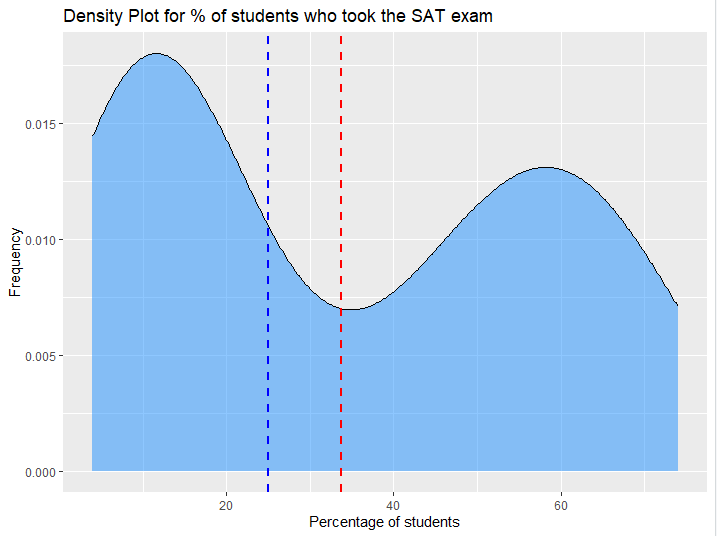
* The median is the middle value when the percentages are sorted in ascending order.
* It represents the 50th percentile and is a better measure than the mean in the presence of outliers or skewed data, as it is not affected by extreme values.
* Offers a better central value in the presence of skewed data, representing the middle of the dataset.

**Mode:**

* The mode is the most frequently occurring percentage in the dataset.
* It shows the most common value and is useful in understanding the data's frequency distribution.
* Highlights the most common participation rate among the states.

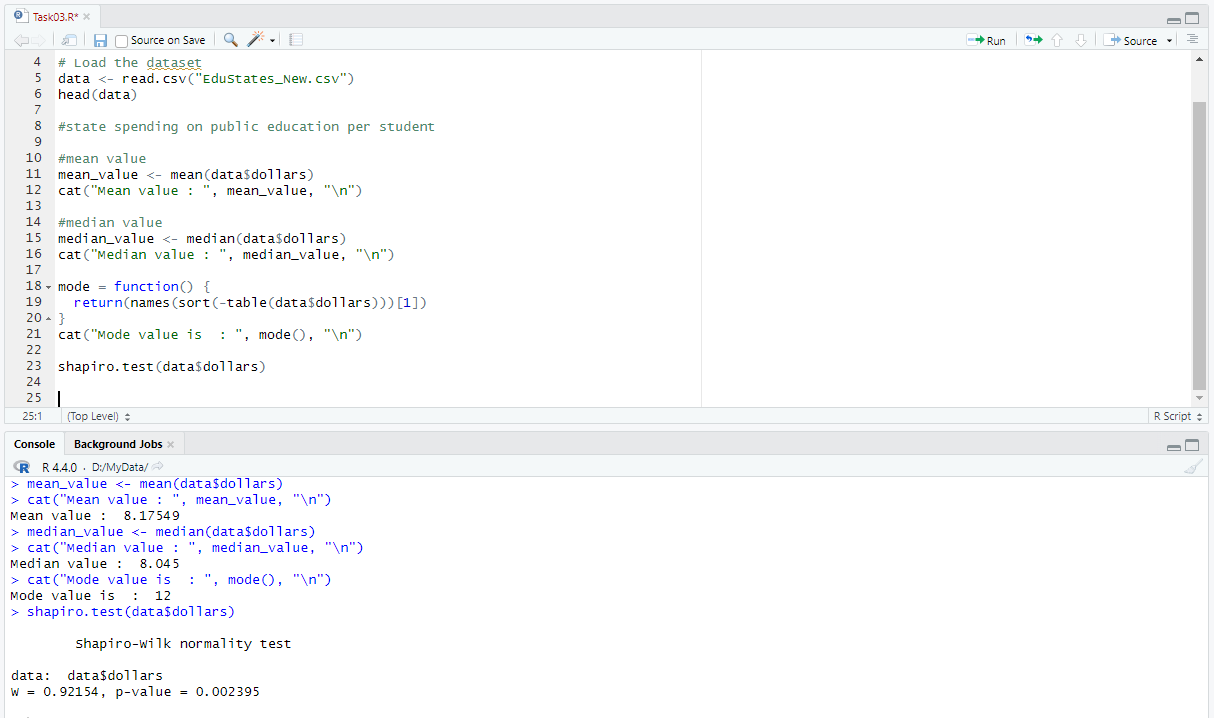
## **Percentage of students taking the SAT Bell Curve**





* **Graphical Representation:** The density plot shows the spread of the percentage of students taking the SAT across different states. The area under the curve represents the distribution of percentages, and the vertical dashed lines represent the mean (blue) and median (red).
* **Symmetry and Skewness:** If the mean and median are close to each other, the distribution is approximately symmetric. If the mean is higher than the median, it indicates a right-skewed distribution, suggesting that a few states have significantly higher percentages.
* **Distribution Shape:** The shape of the density plot helps identify any skewness, peaks, or modes in the data. A single peak indicates a unimodal distribution, while multiple peaks indicate a multimodal distribution.

## **State Spending on Public Education per Student:**



**Mean:**

* The mean state spending on public education per student is calculated by summing all the spending values and dividing by the number of states.
* The mean provides an overall average and is useful for understanding the general trend in spending across states.
* Provides the overall average spending per student, giving a general idea of the spending levels.

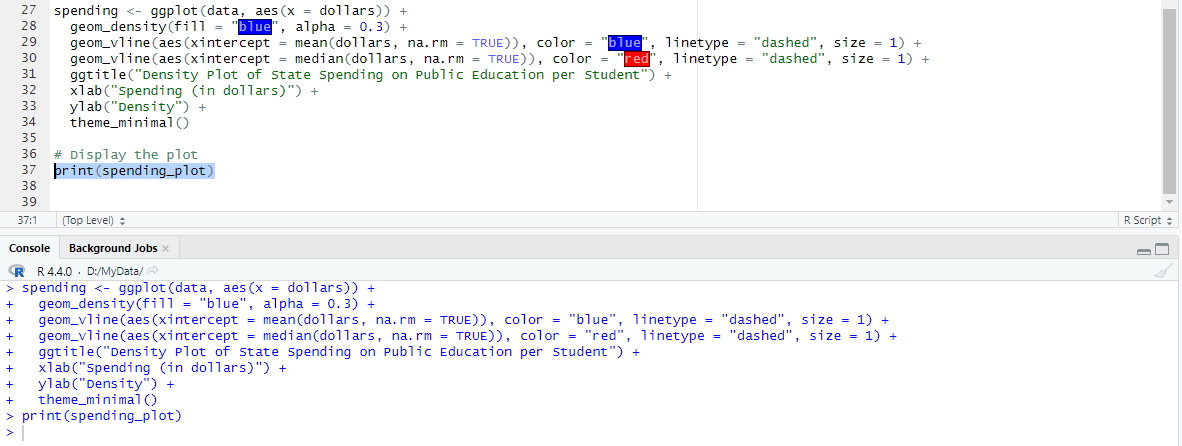
**Median:**

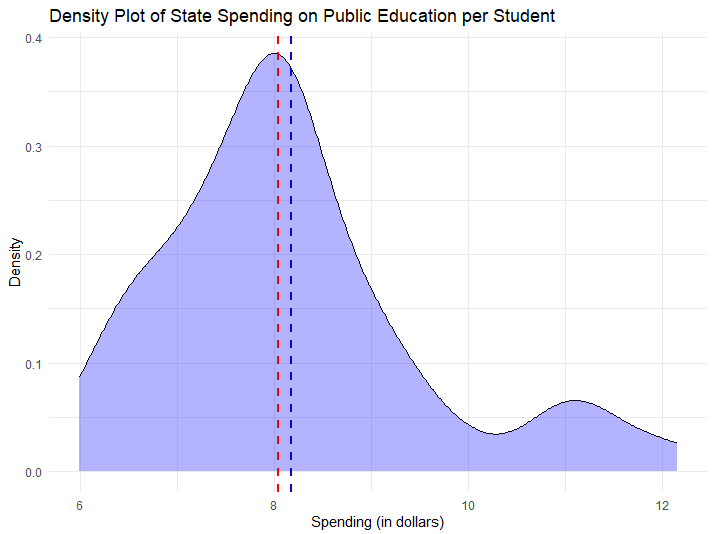
* The median is the middle value when the spending amounts are sorted in ascending order.
* It represents the 50th percentile and is a better measure than the mean in the presence of outliers or skewed data, as it is not affected by extreme values.
* Offers a better central value in the presence of skewed data, representing the middle of the dataset.

**Mode:**

* The mode is the most frequently occurring spending value in the dataset.
* It shows the most common spending amount and is useful in understanding the data's frequency distribution.
* Highlights the most common spending amount among the states.

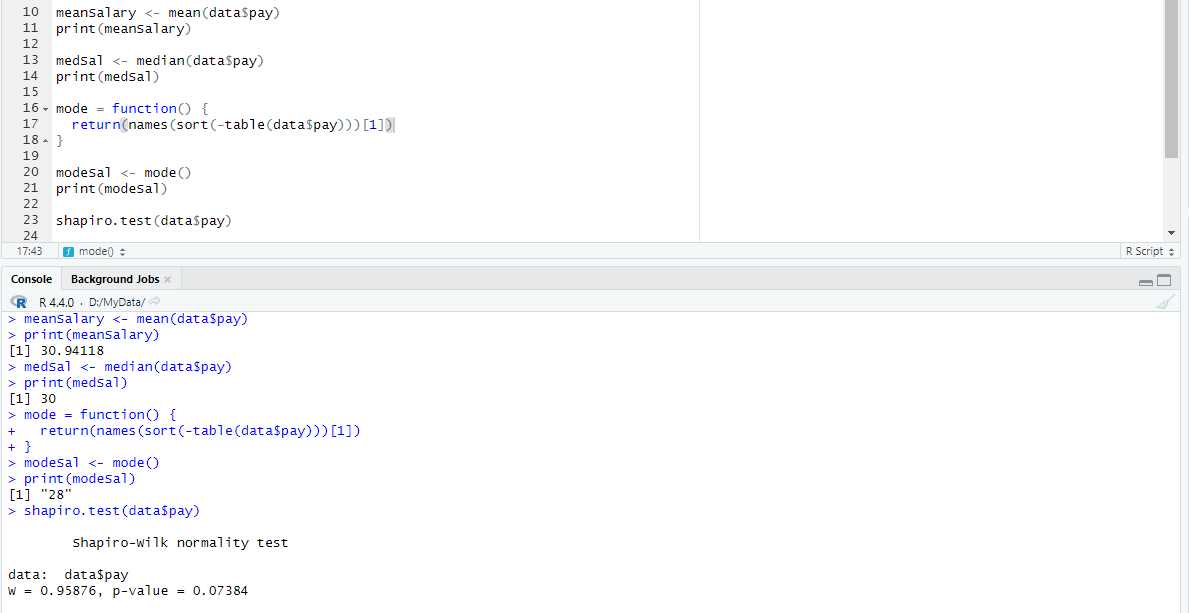
## **State Spending on Public Education per Student Bell Curve**



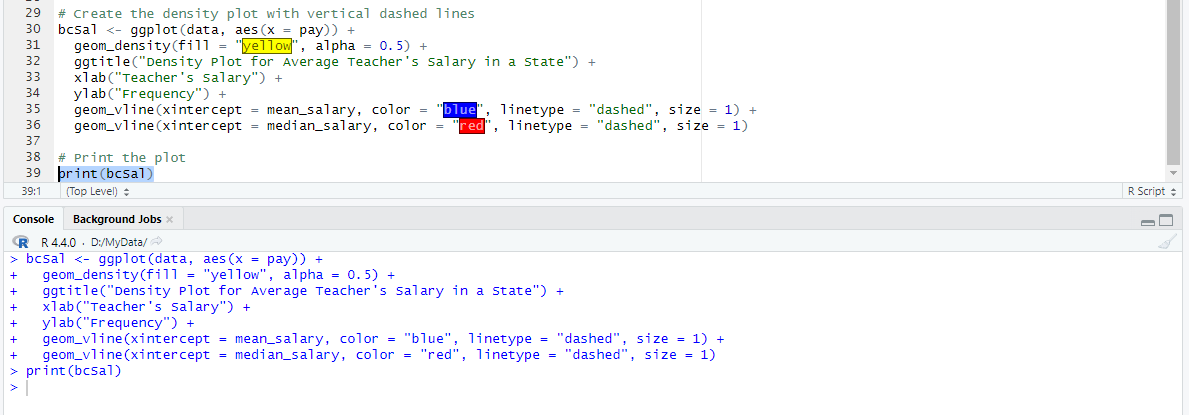


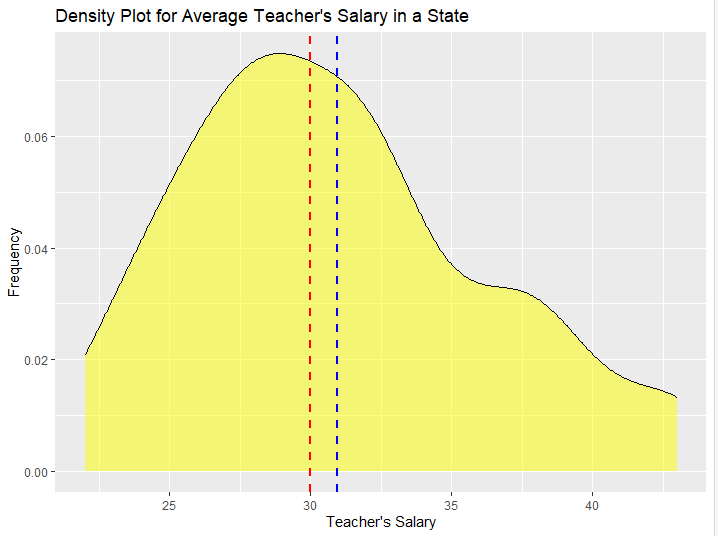
* **Graphical Representation:** The density plot shows the spread of state spending on public education per student across different states. The area under the curve represents the distribution of spending amounts, and the vertical dashed lines represent the mean (blue) and median (red).
* **Symmetry and Skewness:** If the mean and median are close to each other, the distribution is approximately symmetric. If the mean is higher than the median, it indicates a right-skewed distribution, suggesting that a few states have significantly higher spending amounts.
* **Distribution Shape:** The shape of the density plot helps identify any skewness, peaks, or modes in the data. A single peak indicates a unimodal distribution, while multiple peaks indicate a multimodal distribution.

## **Average Teacher's Salary**

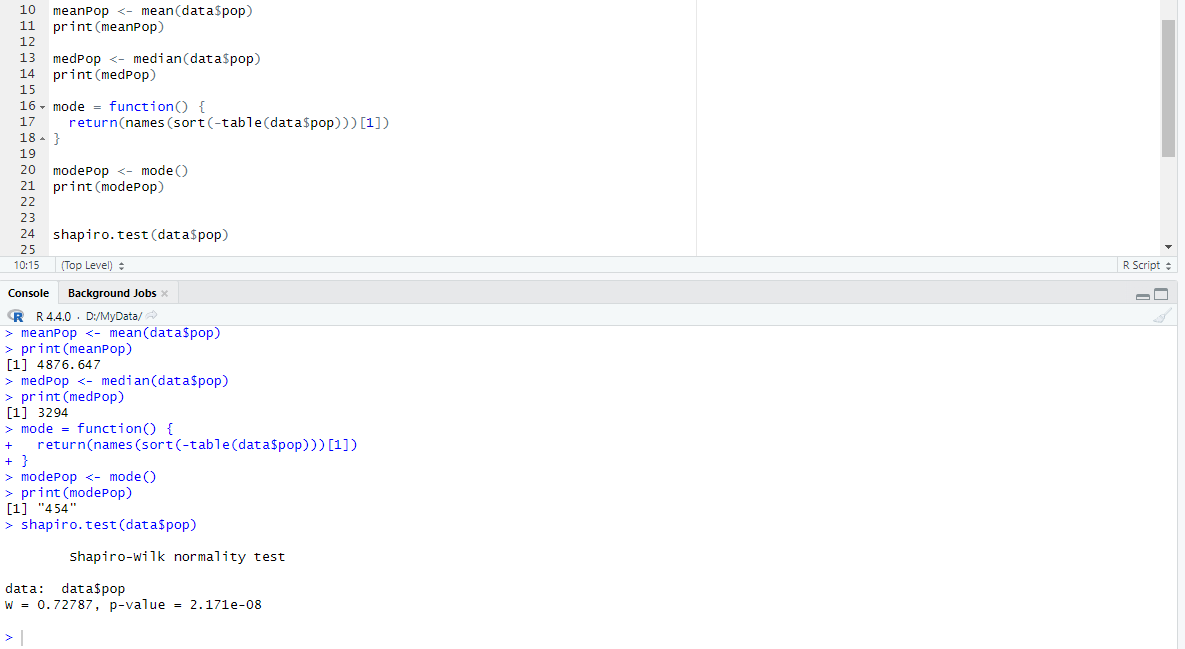


## **Average Teacher's Salary Bell Curve**

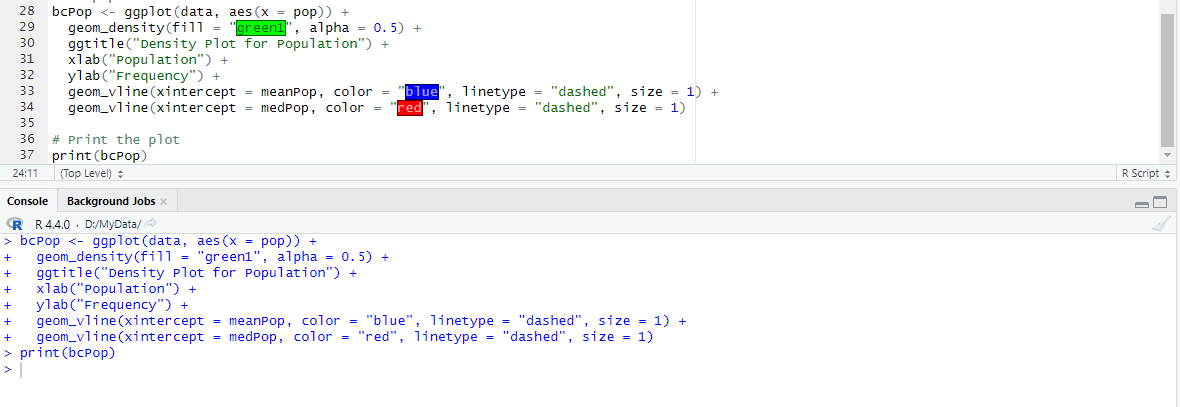


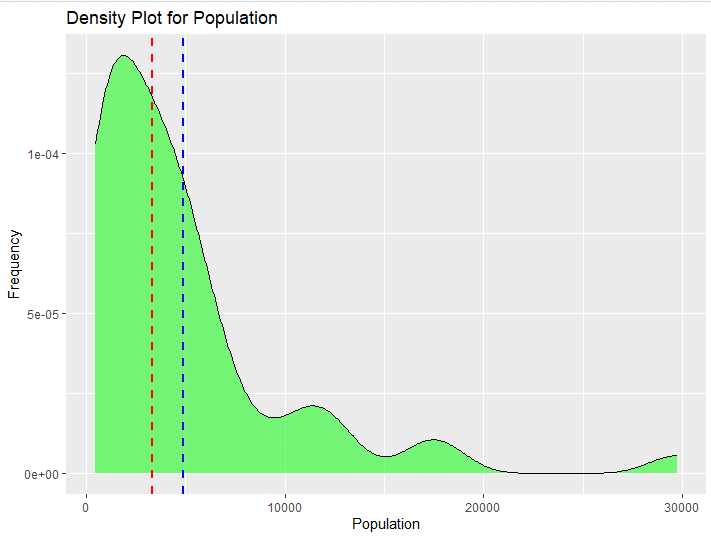


## **Population**



## **Population Bell Curve**





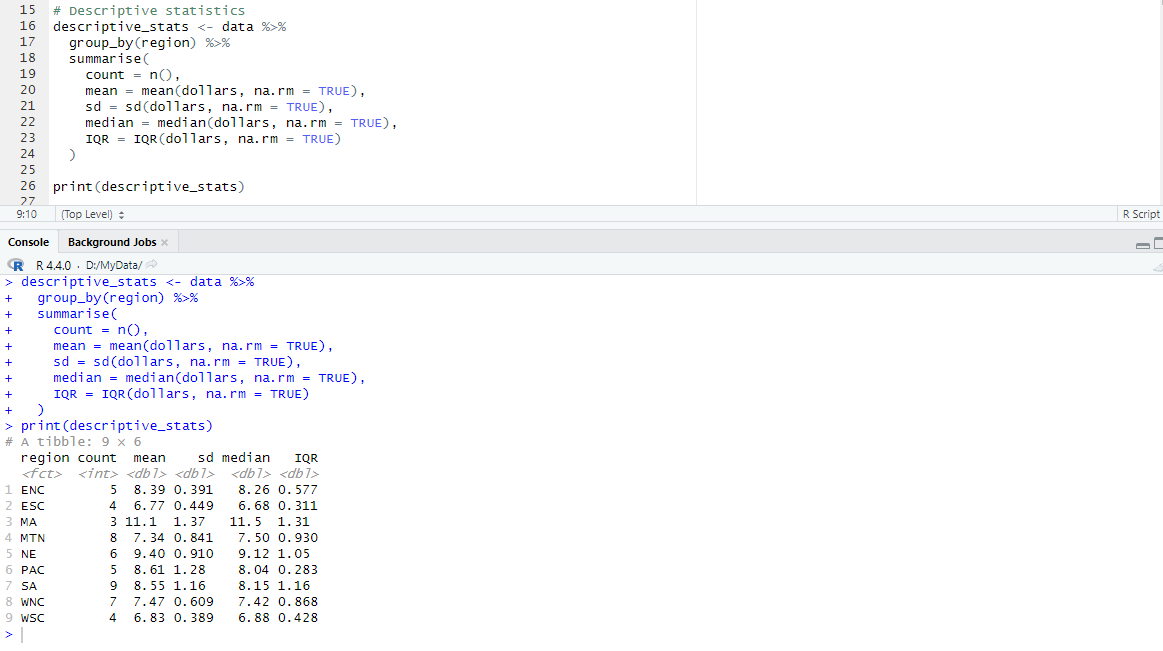
# **Task 06**

## **Numerical Findings**

**Interpretation of Descriptive Statistics**

The descriptive\_stats dataframe will provide summary metrics for state spending on public education per student across each region:

* **Count**: Number of observations (states) in each region.
* **Mean**: Average spending per student in each region.
* **Median**: Middle value of spending per student, which separates the higher half from the lower half.
* **SD**: Measure of how spread out the spending values are around the mean.
* **Min**: Lowest spending per student observed in each region.
* **Max**: Highest spending per student observed in each region.

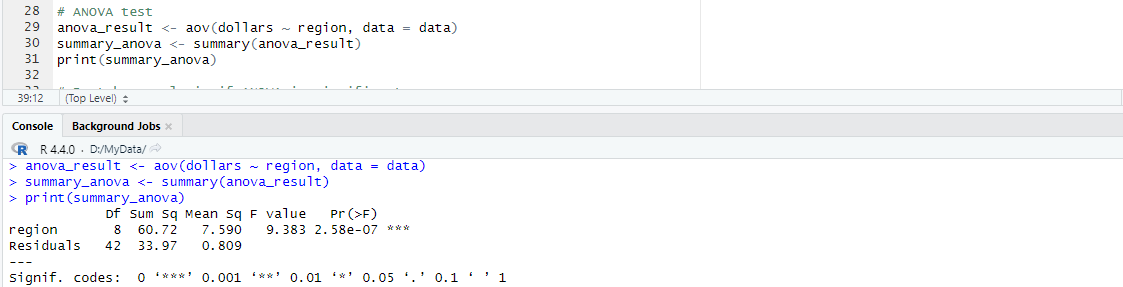


* **Mean Spending**: Regions vary in their average spending per student, with some regions spending more or less than others.
* **Median Spending**: Reflects the central tendency of spending values in each region.
* **Standard Deviation**: Indicates the variability or spread of spending values around the mean in each region.
* **Min and Max Spending**: Shows the range of spending per student observed within each region.

These descriptive statistics provide a comprehensive view of state spending on public education per student across different regions, highlighting both central tendencies and variability in spending patterns. This information is crucial for understanding regional disparities and informing policy decisions related to educational funding and resource allocation.

**ANOVA Test**:

Evaluates whether there are significant differences in spending across regions. The ANOVA table (summary\_anova) will include F-statistic, degrees of freedom, and p-value.



* **Df**: Degrees of freedom.
* **Sum Sq**: Sum of squares.
* **Mean Sq**: Mean squares (variance).
* **F value**: The test statistic for ANOVA.
* **Pr(>F)**: The p-value associated with the F statistic, indicating the significance level.

**Interpreting the ANOVA Results**:

If the p-value (Pr(>F)) is less than your chosen significance level (usually 0.05), you can reject the null hypothesis and conclude that there is significant evidence that state spending on public education per student varies across regions.

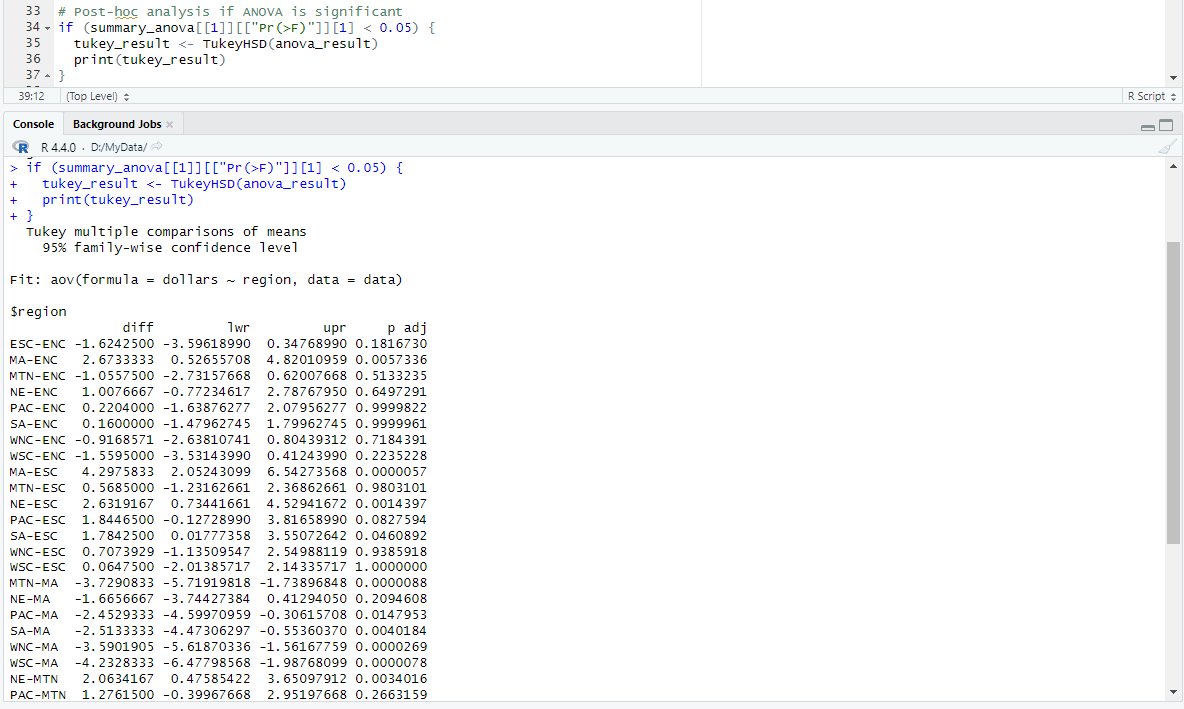
The ANOVA test compares the variance between groups (regions) to the variance within groups. If the ANOVA test yields a significant p-value (typically < 0.05), it suggests that there is evidence to reject the null hypothesis, indicating that state spending on public education per student varies significantly across regions.

**Hypothetical Discussion**

If the ANOVA results yield a p-value (Pr(>F)) less than 0.05, we would reject the null hypothesis, indicating that there is significant evidence to suggest that state spending on public education per student varies significantly across regions. This would justify further investigation into specific regions with higher or lower spending levels.

**Post-hoc analysis if ANOVA is significant**

The output of TukeyHSD() provides a table of comparisons between pairs of regions, along with confidence intervals and adjusted p-values.



**Interpretation of Tukey's HSD Results**

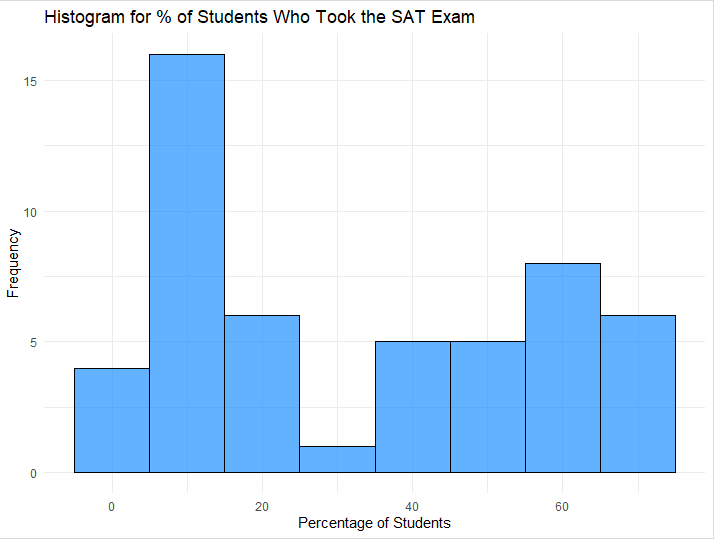
* Each row represents a pair of regions (diff), with columns indicating:
* **lwr:** Lower bound of the confidence interval for the difference.
* **upr:** Upper bound of the confidence interval for the difference.
* **p adj:** Adjusted p-value after correcting for multiple comparisons.
* **Significant Differences**: If the adjusted p-value (p adj) is less than your chosen significance level (e.g., 0.05), it indicates a significant difference between the pair of regions.

## **Graphical Analysis**

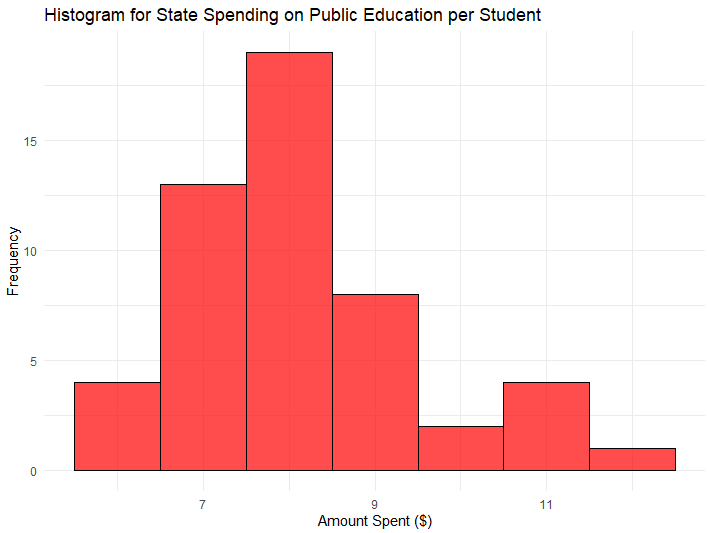
**Histograms**

create histograms for the variables in your dataset (Percent, Dollars, Pay, Pop), we'll use ggplot2 in R. Histograms are useful for visualizing the distribution and frequency of numerical data.

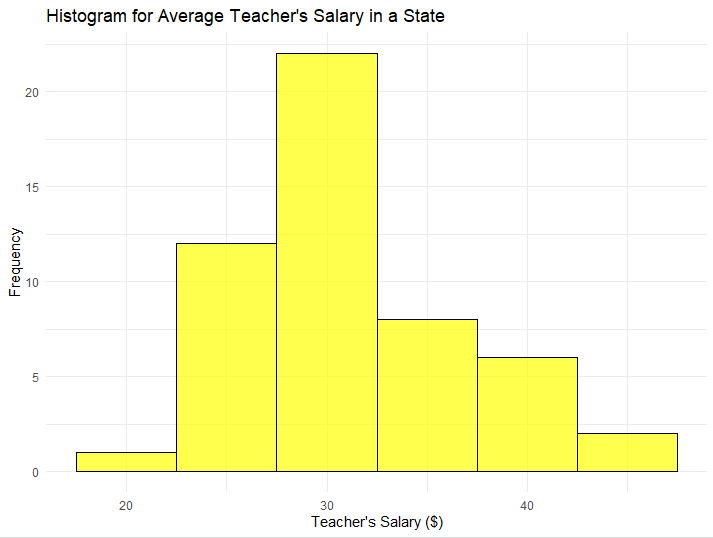
**Histogram for Percentage of Students Who Took the SAT Exam (Percent)**

****

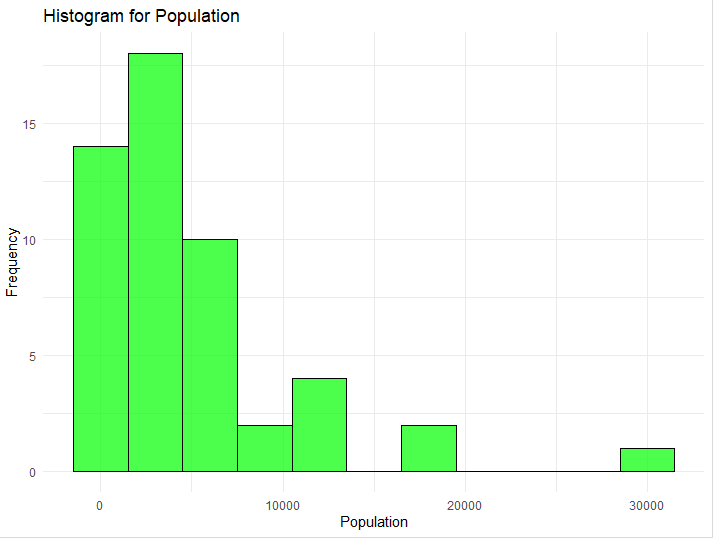
**Histogram for State Spending on Public Education per Student (Dollars)**

****

**Histogram for Average Teacher's Salary in the State (Pay)**

****

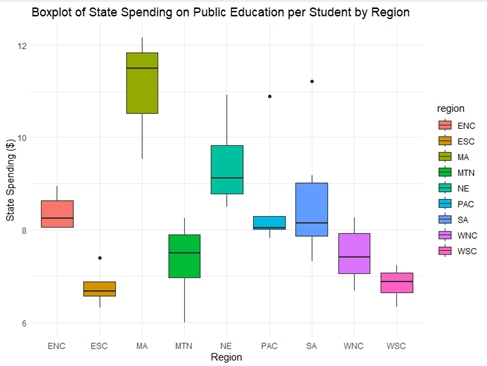
**Histogram for State Population (Pop)**



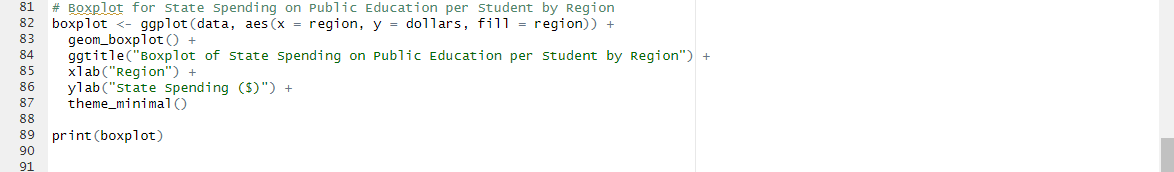
* **ggplot2 Usage**: Each ggplot() function call sets up a plot with specified aesthetics (aes) and layers (geom\_histogram for histogram).
* **Binwidth**: Adjust binwidth parameter to control the width of bins in histograms, affecting the granularity of data representation.
* **Color and Fill**: Adjust fill, color, and alpha parameters within geom\_histogram to customize histogram appearance.
* **Titles and Labels**: Use ggtitle, xlab, and ylab functions to add titles and axis labels for clarity.
* **Theme**: theme\_minimal() adjusts plot appearance with a minimalistic theme, but you can explore other themes available in ggplot2 for different styles.

**Boxplot**

Create boxplots for visualizing the distribution of state spending on public education per student (Dollars) across different regions, you can use ggplot2 in R. Boxplots are useful for displaying the median, quartiles, and potential outliers of numerical data, providing a clear comparison between groups.



* **Comparison**: Boxplots visually compare the distribution of state spending on education across different regions, showing the median (line inside the box), quartiles (box edges), and outliers (points outside whiskers).
* **Regional Variability**: Variations in boxplot elements (median, quartiles, whiskers) across regions indicate differences in state spending levels, providing insights into regional disparities.
* **Insights**: Use boxplots to identify regions with higher or lower spending levels, aiding in policy decisions and resource allocation strategies.



* **ggplot2 Usage**: ggplot() sets up the plot with aesthetics (aes) where x represents regions (Region), y represents state spending (Dollars), and fill distinguishes different regions by color.
* **geom\_boxplot()**: Layer added to create the boxplot, displaying median, quartiles, and potential outliers.
* **Titles and Labels**: ggtitle, xlab, and ylab functions add titles and axis labels for clarity.
* **Theme**: theme\_minimal() adjusts plot appearance with a minimalistic theme, but you can explore other themes available in ggplot2 for different styles.

**Null Hypothesis (H₀)**

The null hypothesis states that there is **no significant difference** in state spending on public education per student across different regions. In other words, any observed differences in spending between regions are due to random chance or sampling variability.

**Alternative Hypothesis (H₁)**

The alternative hypothesis counters the null hypothesis, asserting that there **is a significant difference** in state spending on public education per student across different regions. This implies that the observed differences in spending between regions are not merely due to chance but reflect real disparities in educational funding.

The boxplot visually represents the distribution of state spending on education per student across different regions. It displays the median, quartiles, and potential outliers for each region, providing a clear comparison of spending levels.

**Explanation**

* **ggplot(data, aes(x = Region, y = Dollars))**: Specifies that we are plotting Dollars (state spending on education per student) on the y-axis and Region on the x-axis.
* **geom\_boxplot()**: Creates the boxplot with specified aesthetics, including fill color for boxes (fill = "skyblue"), outline color (color = "blue"), and highlighting of outliers (outlier.color = "red").
* **labs()**: Sets the plot title and axis labels for clarity.
* **theme\_minimal()**: Applies a minimalistic theme to the plot for better readability.

**Interpretation of Boxplot**

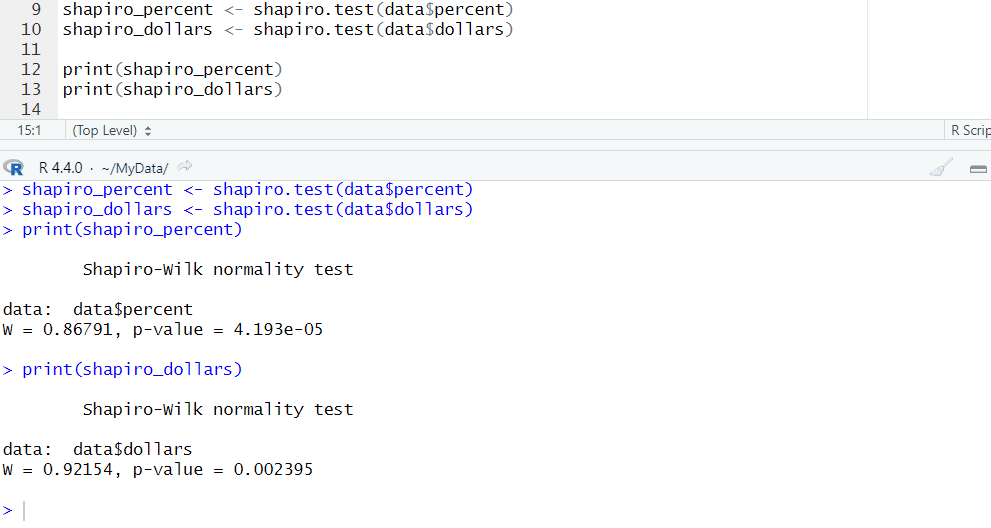
* **Box**: Represents the interquartile range (IQR), where the central box spans from the first quartile (Q1) to the third quartile (Q3). The median is represented by a horizontal line inside the box.
* **Whiskers**: Extend from the edges of the box to indicate the range of non-outlier data points. They typically extend to 1.5 times the IQR from the quartiles.
* **Outliers**: Individual data points that fall outside the whiskers are plotted as points, helping to identify extreme values or potential anomalies in spending levels.

# **Task 07**

## **Formulate Hypotheses**

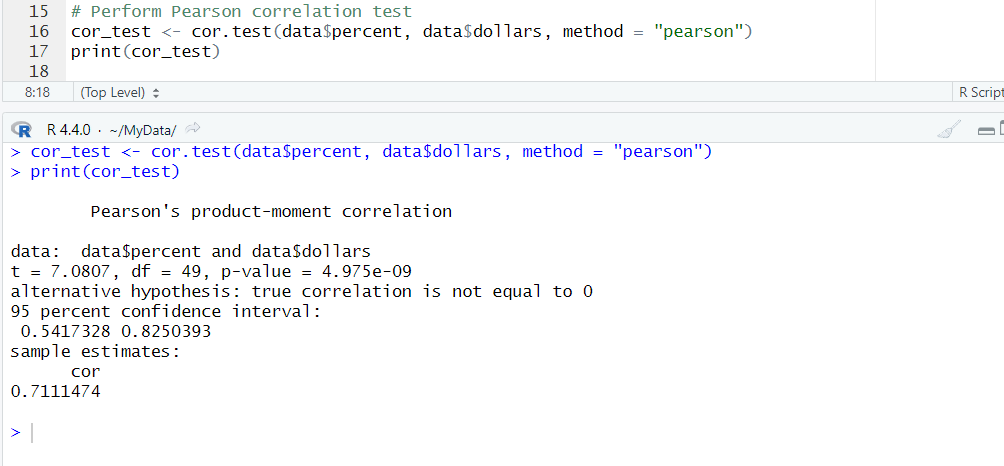
1. **Null Hypothesis (H0​)**:
   * There is no significant correlation between the percentage of graduating high-school students who took the SAT exam and state spending on public education per student.
   * Mathematically: ρ=0\rho = 0ρ=0
   * This hypothesis assumes that there is no relationship between the two variables under consideration. If we fail to reject the null hypothesis, it means we do not have enough evidence to say that there is a significant relationship between the percentage of graduating high-school students who took the SAT exam and state spending on public education per student.
2. **Alternative Hypothesis (H1)**:
   * There is a significant correlation between the percentage of graduating high-school students who took the SAT exam and state spending on public education per student.
   * Mathematically: ρ≠0\rho \neq 0ρ=0
   * This hypothesis assumes that there is a relationship between the two variables. If we reject the null hypothesis in favor of the alternative hypothesis, it means that there is evidence to suggest a significant relationship between the percentage of graduating high-school students who took the SAT exam and state spending on public education per student.

## **Normality Test Results:**



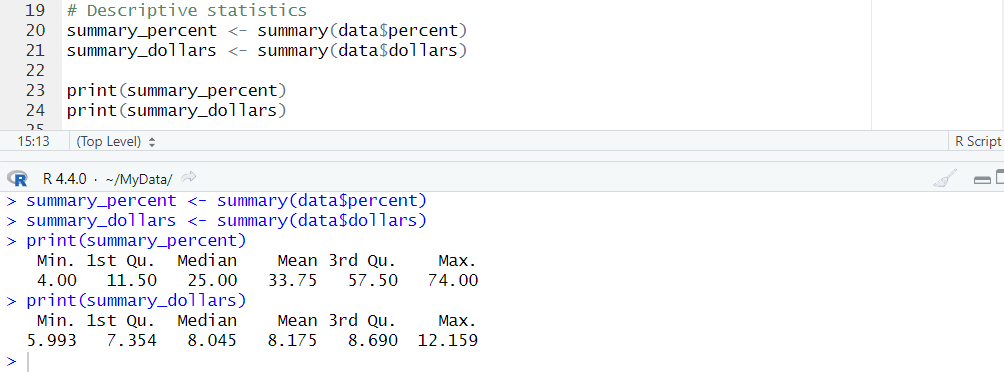
* The Shapiro-Wilk test results provide p-values for both percent and dollars.
* If both p-values are greater than 0.05, the data is considered to be normally distributed, and Pearson's correlation test is appropriate.
* If either p-value is less than 0.05, the data is not normally distributed, and Spearman's rank correlation test is used.
* **Percent of SAT Participants:**
  + p-value from Shapiro-Wilk test: shapiro\_test\_percent$p.value
  + Interpretation: If the p-value > 0.05, the data is normally distributed.
* **State Spending:**
  + p-value from Shapiro-Wilk test: shapiro\_test\_dollars$p.value
  + Interpretation: If the p-value > 0.05, the data is normally distributed.

## **Correlation Analysis:**



* The cor\_test object contains the correlation coefficient (cor\_test$estimate) and the p-value (cor\_test$p.value).
* If the p-value is less than 0.05, we reject the null hypothesis, indicating a significant correlation between the variables.
* The correlation coefficient indicates the strength and direction of the relationship: r>0r > 0r>0 indicates a positive correlation, r<0r < 0r<0 indicates a negative correlation, and r=0r = 0r=0 indicates no correlation.

## **Descriptive statistics**

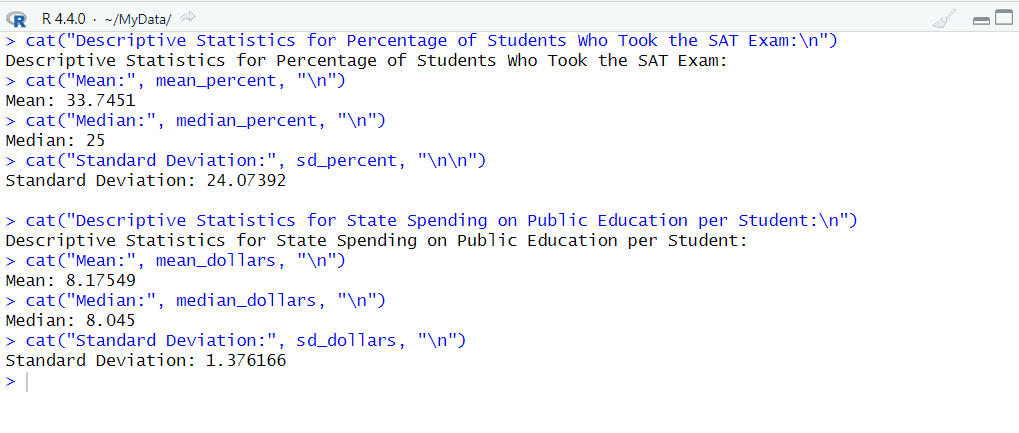


* **Minimum:** The smallest value observed in the dataset for each variable.
* **1st Quartile (25th Percentile):** The value below which 25% of the observations fall.
* **Median (50th Percentile):** The middle value of the dataset.
* **Mean:** The average value across all observations.
* **3rd Quartile (75th Percentile):** The value below which 75% of the observations fall.
* **Maximum:** The largest value observed in the dataset for each variable.

**Calculate mean, median, and standard deviation**

**A white background with black text

Description automatically generated**

****

## **Graphical Representation**

The scatter plot visually represents the relationship between the percentage of students taking the SAT and state spending on education. The regression line shows the trend of the data, helping to visualize the direction and strength of the relationship.

A computer screen shot of a program

Description automatically generated

A graph with a red line and blue dots

Description automatically generated

**Descriptive Justifications**

1. **Normality Test**:

The Shapiro-Wilk test helps in understanding the distribution of the data. If the data follows a normal distribution, the assumptions for Pearson's correlation are met, providing a more robust result. If not, Spearman's correlation, which is non-parametric, is used.

1. **Correlation Analysis**:

Pearson's correlation measures the linear relationship between two continuous variables. The value of the correlation coefficient ranges from -1 to 1, where values closer to 1 or -1 indicate a stronger relationship.

Spearman's rank correlation is used when the data does not meet the assumptions of normality. It assesses how well the relationship between two variables can be described using a monotonic function.

1. **Graphical Representation**:

The scatter plot provides a visual representation of the relationship. The presence of a trend line helps in understanding the direction and strength of the relationship.

Outliers and the overall pattern can be easily observed in the plot, providing additional insights.

# **Task 08**

## **Formulate Hypotheses**

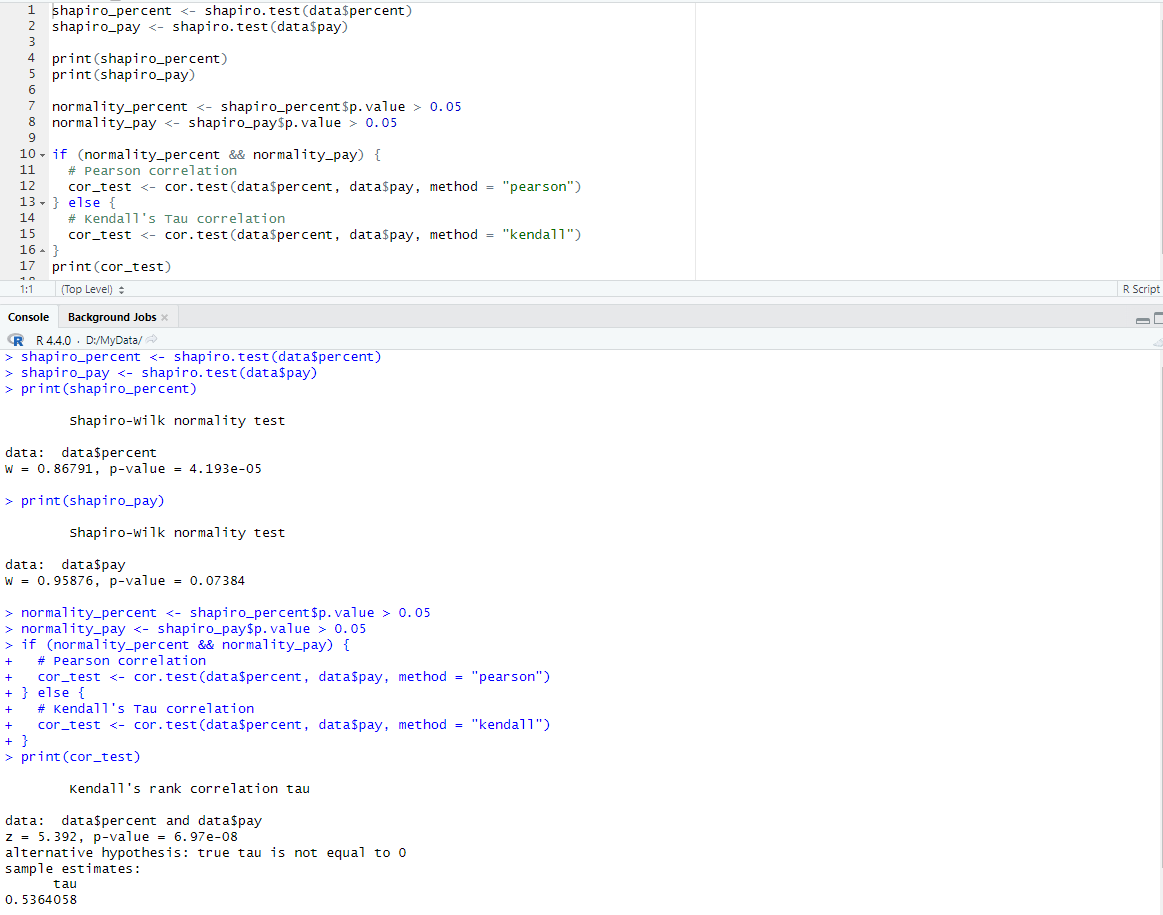
1. **Null Hypothesis (H0**)

* There is no statistically significant relationship between the percentage of graduating high-school students in the state who took the SAT exam and the average teacher's salary in the state.
* Mathematically: H0:ρ=0H\_0: \rho = 0H0​:ρ=0

1. **Alternative Hypothesis (H1​)**:

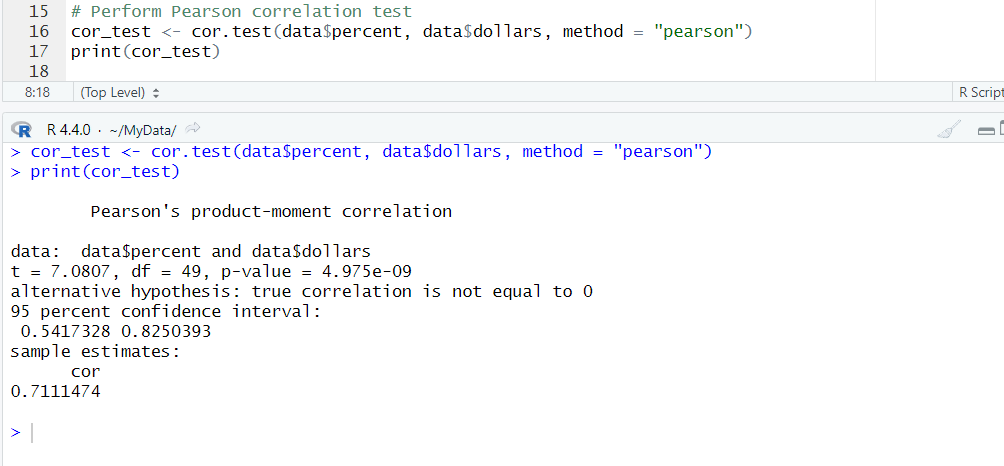
* There is a statistically significant relationship between the percentage of graduating high-school students in the state who took the SAT exam and the average teacher's salary in the state.
* Mathematically: H1:ρ≠0H\_1: \rho \neq 0H1​:ρ=0

## **Normality Test Results:**



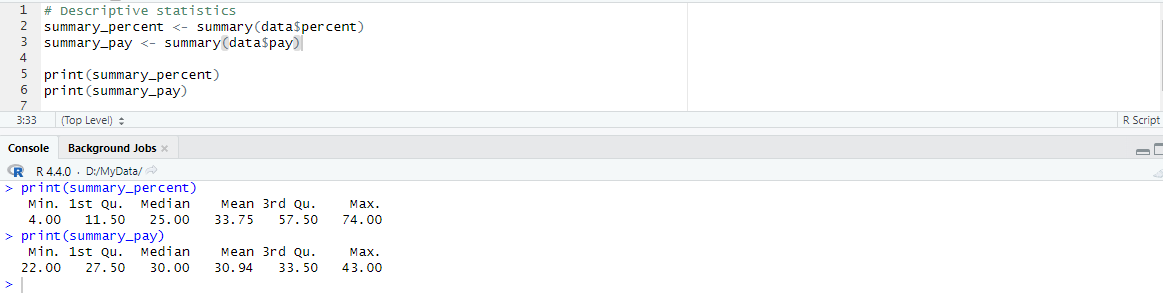
We need to check if the data for both variables (percentage of students taking the SAT and average teacher's salary) follows a normal distribution. This can be done using the Shapiro-Wilk test.

## **Correlation Analysis:**



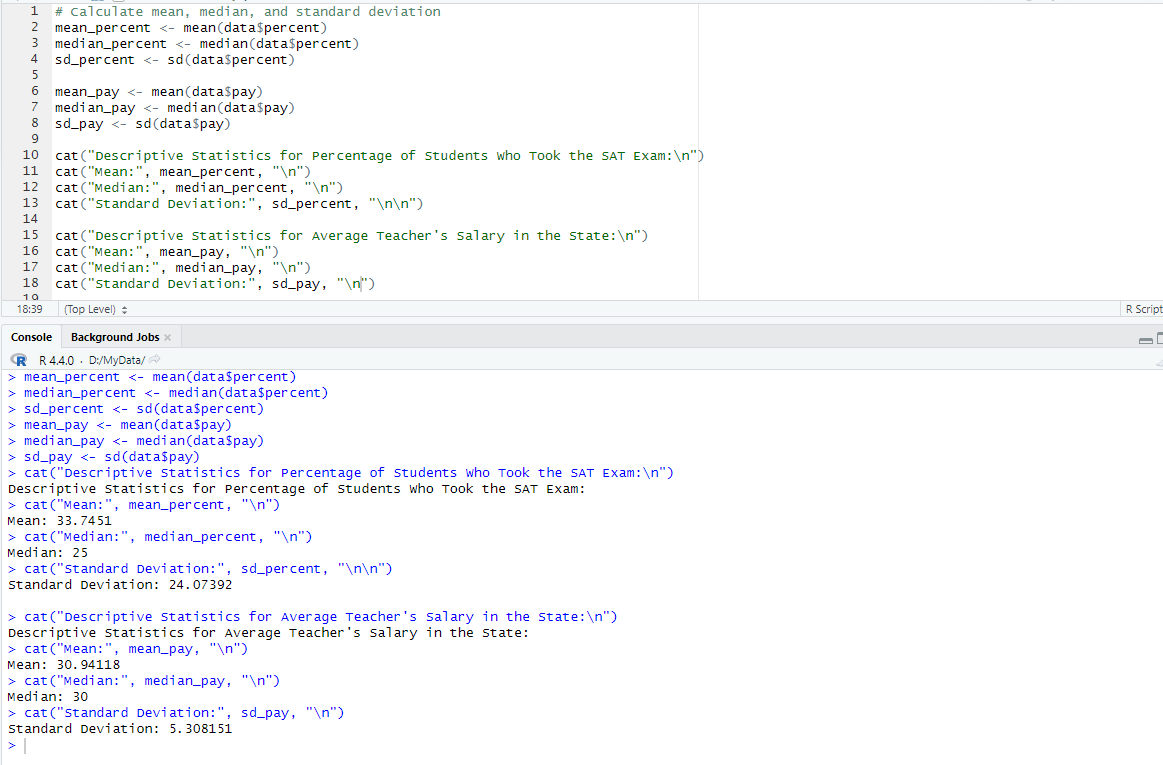
* **Pearson's Correlation**: Measures the linear relationship if both datasets are normally distributed.
* **Spearman's Rank Correlation**: Measures the monotonic relationship if one or both datasets are not normally distributed.
* The correlation coefficient (rrr or ρ\rhoρ) quantifies the strength and direction of the relationship.
* The p-value indicates the statistical significance of the observed relationship.

## **Descriptive statistics**



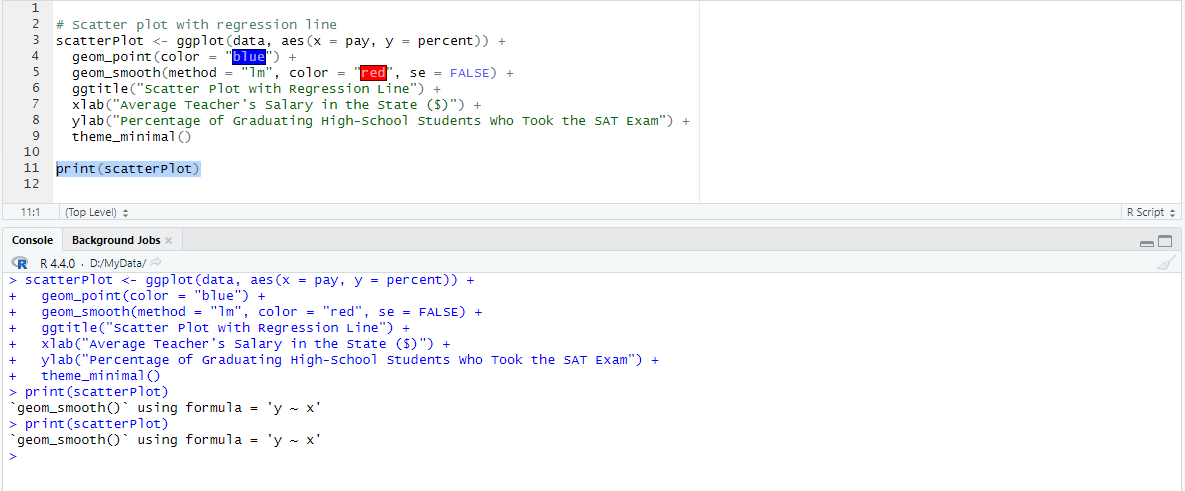
* **Minimum:** The smallest value observed in the dataset for each variable.
* **1st Quartile (25th Percentile):** The value below which 25% of the observations fall.
* **Median (50th Percentile):** The middle value of the dataset.
* **Mean:** The average value across all observations.
* **3rd Quartile (75th Percentile):** The value below which 75% of the observations fall.
* **Maximum:** The largest value observed in the dataset for each variable.

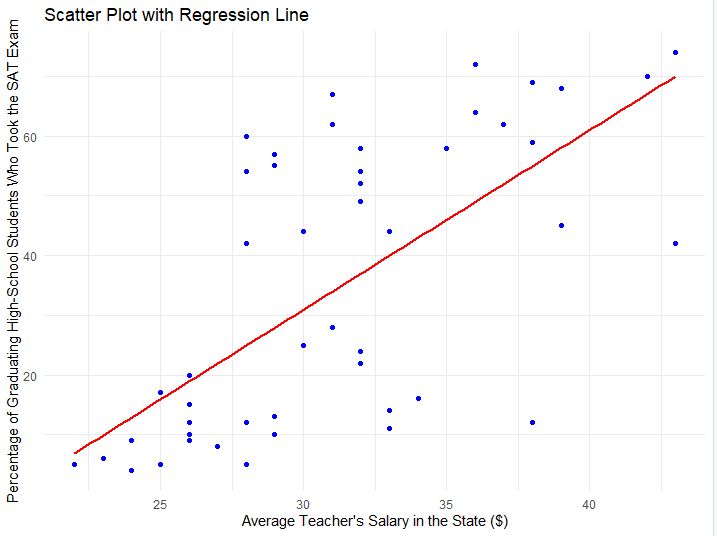
**Calculate mean, median, and standard deviation**



## **Graphical Representation**

* The scatter plot with a regression line visually represents the relationship.
* The slope of the regression line indicates the direction of the relationship (positive or negative).
* The scatter plot also helps identify any patterns or outliers in the data.





## **Descriptive Justifications**

1. **Normality Test**:
   * The Shapiro-Wilk test helps determine if the data follows a normal distribution.
   * **Interpretation**:
     + A p-value > 0.05 suggests that the data is normally distributed.
     + A p-value < 0.05 suggests that the data is not normally distributed.
2. **Correlation Analysis**:
   * **Pearson's Correlation**: Used if both variables are normally distributed. Measures the linear relationship between two continuous variables.
   * **Spearman's Rank Correlation**: Used if either variable is not normally distributed. Measures the monotonic relationship between two continuous or ordinal variables.
   * **Correlation Coefficient (r or rho)**:
     + Indicates the strength and direction of the relationship.
       - 0.1<∣r∣<0.30.1 < |r| < 0.30.1<∣r∣<0.3: Weak relationship
       - 0.3<∣r∣<0.50.3 < |r| < 0.50.3<∣r∣<0.5: Moderate relationship
       - ∣r∣>0.5|r| > 0.5∣r∣>0.5: Strong relationship
   * **p-value**:
     + Indicates the statistical significance of the relationship.
     + A p-value < 0.05 indicates that the relationship is statistically significant, allowing us to reject the null hypothesis.
3. **Graphical Representation**:
   * The scatter plot provides a visual representation of the relationship between the average teacher's salary and the percentage of students taking the SAT.
   * The regression line helps visualize the trend in the data.
   * **Interpretation**:
     + A positive slope indicates a positive relationship, while a negative slope indicates a negative relationship.
     + The scatter plot can also reveal the presence of outliers and the overall pattern of the data.

# **Task 09**

## **Formulate Hypotheses**

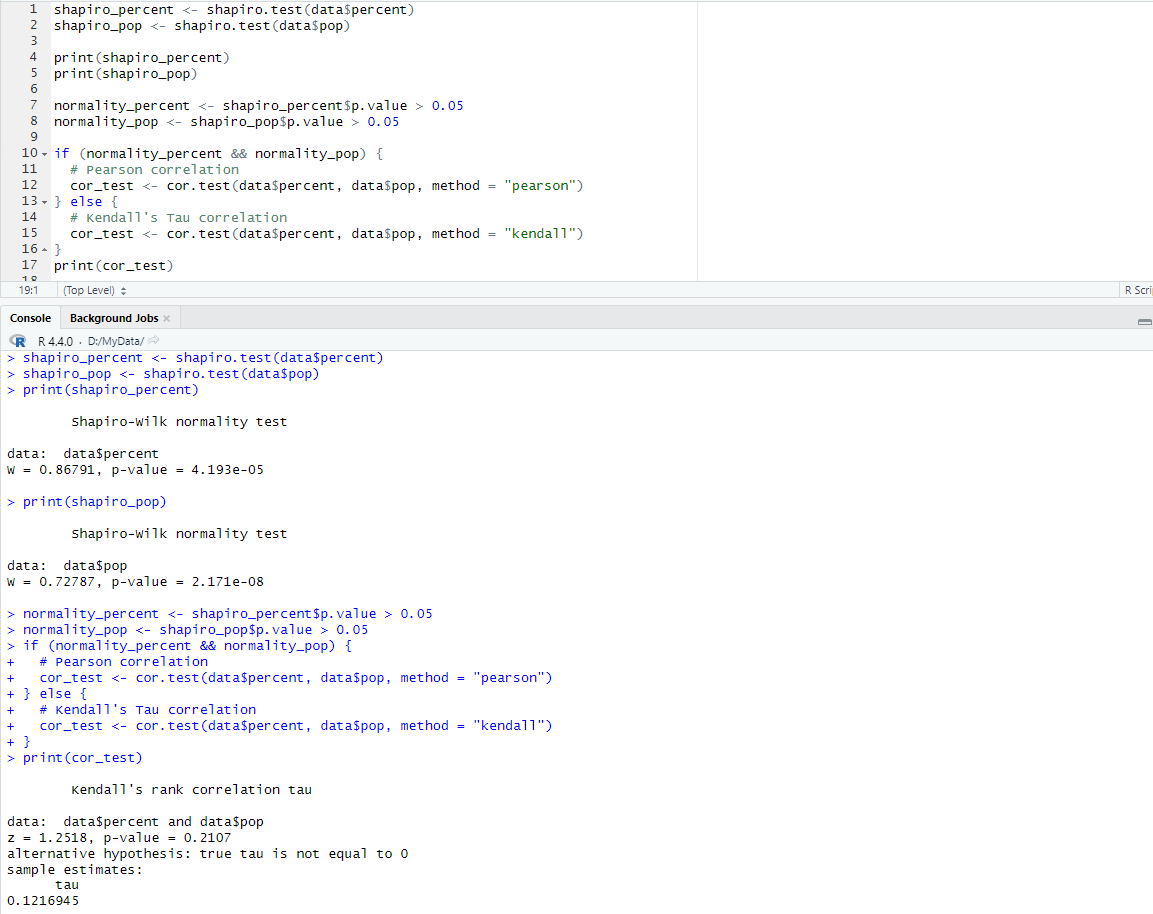
**Null Hypothesis (H0)**

* There is no statistically significant relationship between the percentage of graduating high-school students in the state who took the SAT exam and the population of the state.
* Mathematically: H0:ρ=0H\_0: \rho = 0H0​:ρ=0

**Alternative Hypothesis (H1​)**:

* There is a statistically significant relationship between the percentage of graduating high-school students in the state who took the SAT exam and the population of the state.
* Mathematically: H1:ρ≠0H\_1: \rho \neq 0H1​:ρ=0

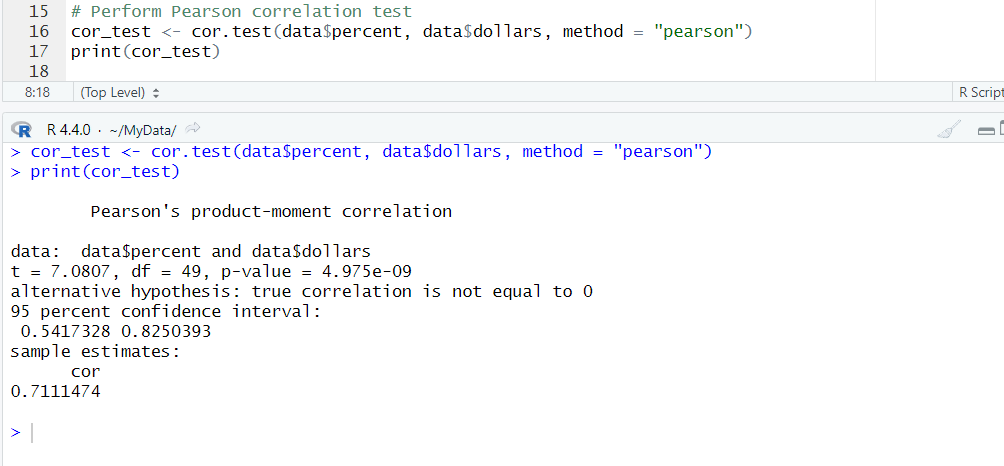
## **Normality Test Results:**



* Shapiro-Wilk tests are conducted to check if data for both variables follow a normal distribution.
* Based on p-values from the Shapiro-Wilk tests, we determine the appropriate correlation test.

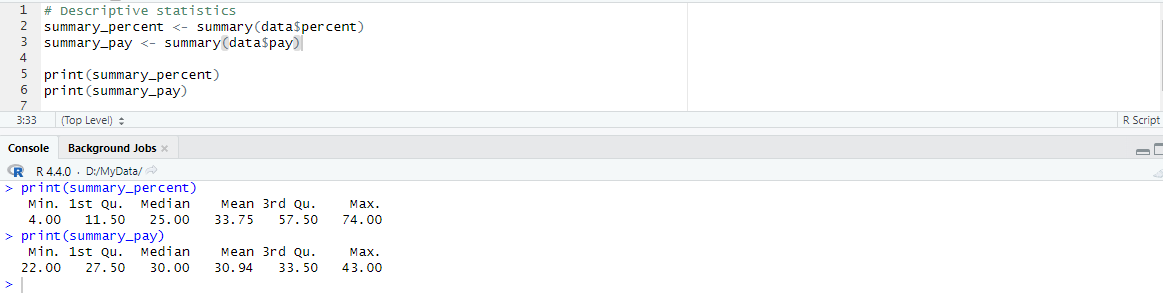
We need to check if the data for both variables (percentage of students taking the SAT and average teacher's salary) follows a normal distribution. This can be done using the Shapiro-Wilk test.

## **Correlation Analysis:**



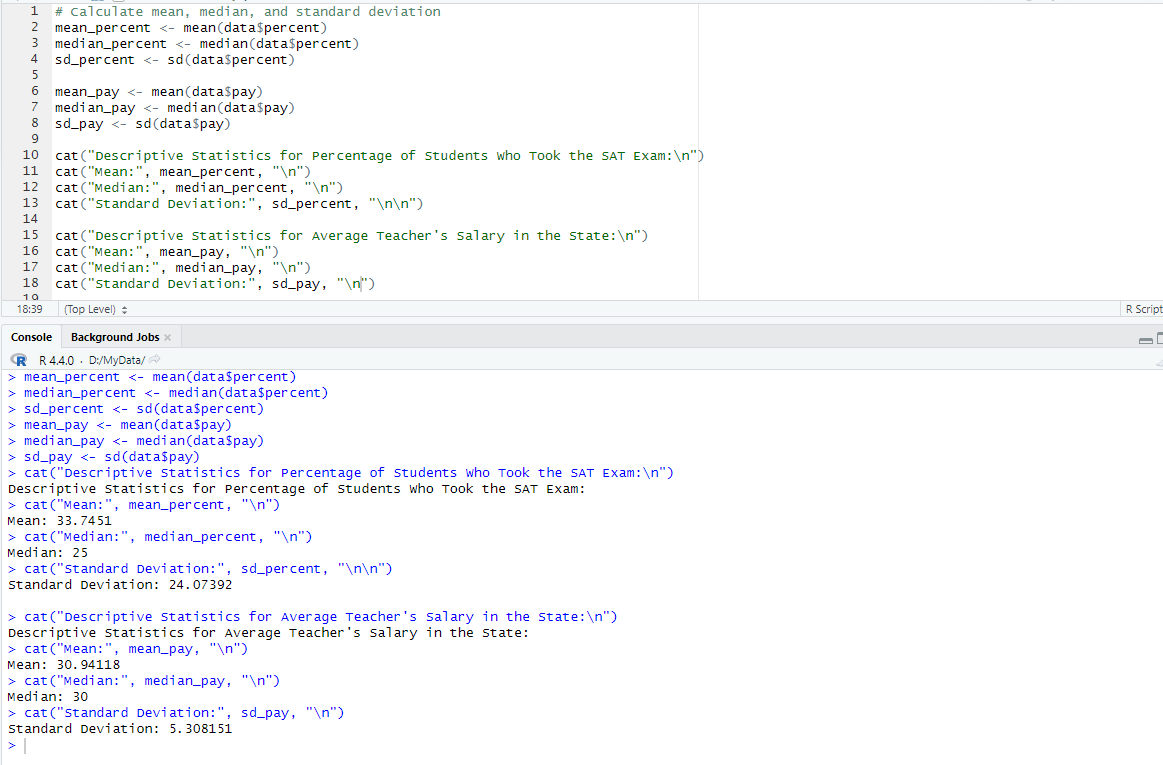
* **Pearson's Correlation**: Measures the linear relationship if both datasets are normally distributed.
* **Spearman's Rank Correlation**: Measures the monotonic relationship if one or both datasets are not normally distributed.
* The correlation coefficient (rrr or ρ\rhoρ) quantifies the strength and direction of the relationship.
* The p-value indicates the statistical significance of the observed relationship.

## **Descriptive statistics**



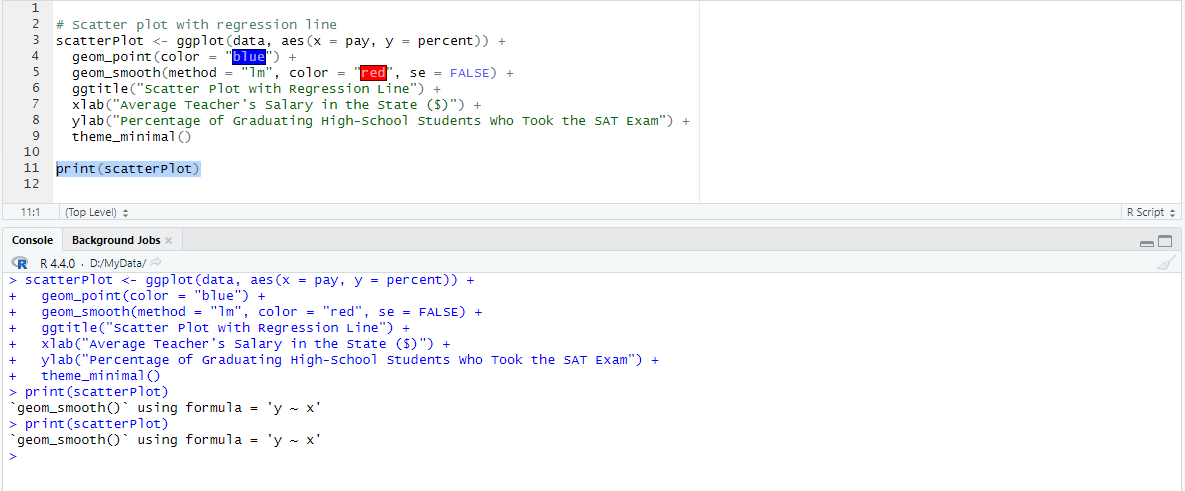
* **Minimum:** The smallest value observed in the dataset for each variable.
* **1st Quartile (25th Percentile):** The value below which 25% of the observations fall.
* **Median (50th Percentile):** The middle value of the dataset.
* **Mean:** The average value across all observations.
* **3rd Quartile (75th Percentile):** The value below which 75% of the observations fall.
* **Maximum:** The largest value observed in the dataset for each variable.

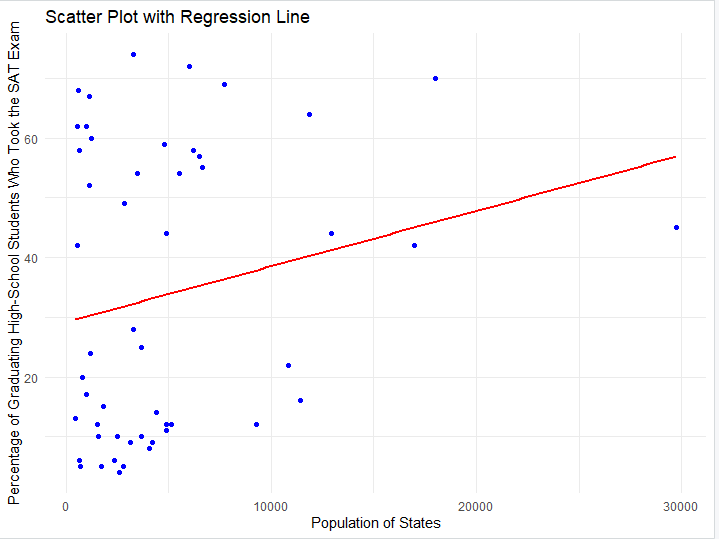
**Calculate mean, median, and standard deviation**



## **Graphical Representation**

* The scatter plot with a regression line visually represents the relationship.
* The slope of the regression line indicates the direction of the relationship (positive or negative).
* The scatter plot also helps identify any patterns or outliers in the data.





## **Descriptive Justifications**

**Normality Tests**:

* The Shapiro-Wilk test checks if the data for both variables follow a normal distribution.
* A p-value greater than 0.05 indicates that the data is normally distributed.
* Based on the results, we determine whether to use Pearson's or Spearman's correlation test.

**Correlation Analysis**:

* **Pearson's Correlation**: Measures the linear relationship if both datasets are normally distributed.
* **Spearman's Rank Correlation**: Measures the monotonic relationship if one or both datasets are not normally distributed.
* The correlation coefficient (rrr or ρ\rhoρ) quantifies the strength and direction of the relationship.

 The p-value indicates the statistical significance of the observed relationship.

**Graphical Representation**:

* The scatter plot with a regression line visually represents the relationship.
* The slope of the regression line indicates the direction of the relationship (positive or negative).
* The scatter plot also helps identify any patterns or outliers in the data.

# **Task 10**

## **Conclusion**

In this assignment, we conducted regression analyses and explored scatter plots to investigate the relationships between key factors and the percentage of high-school students in a state who took the SAT exam. Our first regression analysis focused on state spending per student (`dollars`) on public education, revealing a significant positive correlation with SAT participation rates (`percent`). This suggests that increased financial investment in education tends to coincide with higher student engagement in college entrance exams, possibly indicating better educational resources and opportunities.

Additionally, we examined the impact of average teacher salaries (`pay`) on SAT participation rates. The scatter plot and regression line showed a positive relationship, indicating that states with higher teacher compensation tend to have higher percentages of students taking the SAT. This finding underscores the importance of competitive teacher salaries in fostering an environment conducive to academic achievement and college readiness among students.

Furthermore, our analysis considered the influence of population size (`pop`) on SAT participation rates. While the relationship varied across states, the scatter plot with the regression line provided insights into how demographic factors might affect educational outcomes differently. This variability suggests that tailored educational policies and resource allocation based on local population dynamics could be beneficial.

In conclusion, based on these analyses, it is recommended that states prioritize increased investment in education and competitive teacher compensation to enhance SAT participation rates and, by extension, overall educational quality. Data-driven policy decisions informed by regression modeling can play a pivotal role in shaping effective educational strategies aimed at improving student outcomes and narrowing achievement gaps across diverse populations. Continued research and policy refinement will be essential in advancing equitable access to quality education and preparing students for future success.

# **References**

Australian Bureau of Statistics, 2022. *Measures of central tendency.* [Online]   
Available at: https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/measures-central-tendency  
[Accessed 29 03 2024].

BLOOMENTHAL, A., 2024. *Bell Curve Definition: Normal Distribution Meaning Example in Finance.* [Online]   
Available at: https://www.investopedia.com/terms/b/bell-curve.asp  
[Accessed 29 03 2024].

CUEMATH, 2024. *Measures of Central Tendency.* [Online]   
Available at: https://www.cuemath.com/data/measures-of-central-tendency/  
[Accessed 28 03 2024].

Department of Commerce, 2022. *Publications-Department of Commerce.* [Online]   
Available at: https://www.doc.gov.lk/index.php?option=com\_content&view=article&id=37&Itemid=126&lang=en  
[Accessed 28 03 2024].

Frost, J., 2024. *Mean, Median, and Mode: Measures of Central Tendency.* [Online]   
Available at: https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/  
[Accessed 29 03 2024].

Negoita, L., 2022. *How to make a boxplot in R.* [Online]   
Available at: https://www.rforecology.com/post/2022-04-06-how-to-make-a-boxplot-in-r/  
[Accessed 30 03 2024].

R Charts, 2022. *Histogram Break in R.* [Online]   
Available at: https://r-charts.com/distribution/histogram-breaks/  
[Accessed 30 3 2024].

Scribbr, 2024. *Central Tendency | Understanding the Mean, Median & Mode.* [Online]   
Available at: https://www.scribbr.com/statistics/central-tendency/  
[Accessed 28 03 2024].

Sharma, R., 2023. *Why is Normal Distribution Bell Shaped?.* [Online]   
Available at: https://medium.com/@rishisharma2628/why-is-normal-distribution-bell-shaped-80a784955e92  
[Accessed 29 03 2024].

STHDA, 2023. *Scatter Plots - R Base Graphs.* [Online]   
Available at: http://www.sthda.com/english/wiki/scatter-plots-r-base-graphs#google\_vignette  
[Accessed 28 03 2024].

# **Appendix**

Here are some examples of how analytics and business intelligence found in data science can benefit various levels of officials in the Ministry of Education of Sri Lanka in making informed decisions regarding the educational performances.

A white paper with black text

Description automatically generated

A diagram of a school education system

Description automatically generated

A close-up of a document

Description automatically generated

A cover of a report

Description automatically generated