PROJECT DOCUMENTATION

STUDENT\_PERFOMANCE\_ANALYSIS

|  |  |
| --- | --- |
| TITLE: | STUDENT\_PERFOMANCE\_ANALYSIS |
| NAME: | GOWTHAM.K |
| COURSE: | DA/DS, Offline |
| BATCH: | JUNE 2025 |

TABLE OF CONTENT

|  |  |
| --- | --- |
| **1.** | **Introduction** |
| **2.** | **Aim** |
| **3.** | **Business Problem / Problem Statement** |
| **4.** | **Project Workflow** |
| **5.** | **Data Understanding** |
| **6.** | **Data Cleaning**  Missing Values Imputation  Outlier Treatment |
| **7.** | **Filtering Data for Analysis** |
| **8.** | **Statistical Analysis**  Descriptive analysis  Test statistics and hypothesis testing |
| **9.** | **Exploratory Data Analysis (EDA) - Univariate Analysis** |
| **10.** | **Bivariate Analysis** |
| **11.** | **Multivariate Analysis** |
| **12.** | **Overall Insights from Analysis** |
| **13.** | **Conclusion** |

1. **INTRODUCTION**

This analysis project aims to explore a dataset containing student performance data. The primary objective is to investigate the relationship between various factors, such as gender, race/ethnicity, parental education, lunch type, and test preparation course completion, and student scores in math, reading, and writing. The project's key focus is to use data cleaning, exploration, and statistical testing to uncover insights into these relationships. The dataset, which includes information on 1,000 students, provides a context for understanding potential influences on academic achievement.

1. **AIM OF THE PROJECT**

The project's aim as an analysis of a student performance dataset. The key goal is to investigate the relationships between various student attributes and their academic scores in math, reading, and writing.

* Examining how factors like **gender, parental education level, and test preparation** correlate with student performance.
* Utilizing data cleaning, exploration, and statistical testing (such as the **Z-test**) to identify significant patterns and insights.
* Drawing conclusions about which factors may have the most notable impact on student achievement.

1. **PROBLEM STATEMENT**

The business problem or problem statement is centered on understanding the factors that influence student performance.

The real-world implications of this problem are significant, especially in the field of education. By identifying the key determinants of academic success, educators, policymakers, and parents can develop more effective strategies and interventions.

The analysis is valuable because it can:

* Help schools tailor their teaching methods and resources to better support students from different backgrounds.
* Inform the development of policies aimed at reducing disparities in educational outcomes.
* Provide a data-driven basis for implementing programs that address specific challenges, such as a lack of test preparation or socioeconomic factors.

1. **PROJECT WORKFLOW**

**1. Data Preparation:**

* Import Libraries: The project begins by importing standard libraries for data analysis and visualization, including NumPy, pandas, matplotlib, and seaborn.
* Data Loading: A dataset named Students\_Performance\_knn.csv is loaded into a pandas DataFrame. The columns are then renamed for clarity.

**2. Data Cleaning:**

* Handling Null Values: The workflow checks for missing values in the dataset. It identifies 25 null values in the Lunch column and fills them with the most frequent value (mode).
* Checking for Duplicates: The code also checks for duplicate rows, but the analysis confirms that there are no duplicates in the dataset.

**3. Statistical Analysis:**

* Z-Test: A Z-test is performed to compare the writing scores of male and female students.
* Key Findings: The Z-test results indicate a statistically significant difference between the writing scores of male and female students. The negative Z-statistic suggests that female students scored higher on average than male students in writing.

**4. Data Visualization & Exploration**

The project includes several steps to explore and visualize the data to understand the relationships between different variables.

* Descriptive Statistics:

The notebook uses the .describe() method to get a statistical summary of the dataset's numerical columns.

* **Unique Values:**

The code also checks for the number of unique values in each column, which helps in identifying categorical versus continuous data.

* **Box Plots:**

Box plots are used to visualize the distribution of math, reading, and writing scores across different genders and race/ethnicity groups. This visualization helps in identifying potential outliers and comparing score distributions.

* **Kernel Density Estimate (KDE) Plots:**

The notebook generates KDE plots to show the distribution of scores in math, reading, and writing. This provides a smooth representation of the data distribution.

* **Correlation Heatmap:**

A heatmap is created to visualize the correlation between the numerical columns (math, reading, and writing scores). This helps to quickly identify if any scores are strongly correlated with each other. The analysis found a strong positive correlation between all three scores.

1. **Key Insights from Analysis:**

In addition to the Z-test on writing scores, the project highlights other key findings:

* **Parental Level of Education:**

The analysis indicates that students with parents who have an associate's degree tend to perform better than those with parents who have a high school diploma or some college education.

* **Preparation Course:**

Students who completed the test preparation course performed better on all three tests (math, reading, and writing) compared to those who did not.

1. **DATA UNDERSTANDING**

**1. Dataset Overview:**

The dataset contains information on **1,000 students** with **8 key features**.

**2. Key Variables:**

* **Gender:** Male/Female
* **Race/Ethnicity:** Five groups (A, B, C, D, E)
* **Parental Level of Education:** Level of education attained by a student's parents.
* **Lunch:** Type of lunch (standard vs. free/reduced)
* **Test Preparation Course:** Whether a student completed a test preparation course (completed vs. none)
* **Math Score:** The numerical score on the math test.
* **Reading Score:** The numerical score on the reading test.
* **Writing Score:** The numerical score on the writing test.

**3. Data Cleaning & Pre-processing**

The notebook details the following steps taken to ensure the data is clean and ready for analysis:

* **Handling Null Values**: The dataset was checked for missing values, and it was found that the Lunch column had 25 null values. These were filled using the **mode** (most frequent value) of that column.
* **Checking for Duplicates**: The analysis also included a check for duplicate rows. The notebook confirms that **no duplicate entries** were found in the dataset.

**4. Statistical Analysis**

The project used statistical methods to derive insights from the data:

* **Z-Test**: A Z-test was performed to compare the **writing scores** between male and female students.
* **Finding**: The Z-test showed a statistically significant difference between the writing scores of male and female students.
* **Interpretation**: The negative Z-statistic value indicated that **female students scored higher on average** than male students in writing.

1. **DATA CLEANING**

To ensure accurate analysis and meaningful insights, the dataset underwent a thorough cleaning process. This step corrected inconsistencies, handled missing values, and prepared the data for analysis.

1. **Missing Values Imputation**

Filling in missing data to maintain dataset completeness.

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

* Detected **null values** in the columns: Lunch
* For **categorical variables** (L), null values were filled using the **mode** (most frequent value).
* For the **numerical variable** (Education), missing values were filled using the **mean**.
* Used the fillna() function in Python for replacing missing values with appropriate statistics.
* Ensured no missing values remain in the dataset after imputation, making it ready for further analysis.

1. **Outlier Treatment**

Outliers are unusual values that deviate significantly from the overall distribution of data. If not handled properly, they can distort statistical analysis and negatively impact machine learning models. Therefore, identifying and treating outliers was an important step in the Student Performance Analysis

* Used Interquartile Range (IQR) method:

A screenshot of a computer screen

AI-generated content may be incorrect.

1. **FILTERING DATA FOR ANALYSIS**

**1. Identifying Null Values:**

The checks for missing values in the dataset using the .isnull().sum() function.

* The output shows that the 'Lunch' column has 25 null values, while all other columns have no missing data.

**2. Handling Missing Values:**

To handle the missing values in the 'Lunch' column, the notebook uses the .fillna() function with the mode of the 'Lunch' column.

* This approach is appropriate because 'Lunch' is a categorical data type.
* After this step, a check confirms that there are no remaining null values in the dataset.

**3. Checking for Duplicate Rows**

The notebook also checks for duplicate rows in the dataset using the duplicated() function.

* The check reveals that there are no duplicate values in the dataset.
* The notebook proceeds with a .drop\_duplicates() call, which results in the same number of rows as before (1000).

1. **STATISTICAL ANALYSIS**
   1. **Descriptive Analysis**

Used to summarize and understand the central tendency, spread, and distribution of data.

**A screenshot of a calculator

AI-generated content may be incorrect.**

**2. Hypothesis testing:**

Hypothesis testing was carried out to statistically verify the relationship between employee attributes and attrition. Both **categorical** and **numerical** features were tested against the target variable (Attrition) to identify significant factors.

**1. The Hypotheses**

The file implicitly defines the following hypotheses to be tested

* Null Hypothesis (H0​): There is no significant difference between the average writing scores of male and female students.
* Alternative Hypothesis (Ha​): There is a significant difference between the average writing scores of male and female students.

**2. The Test and Results**

* The notebook uses the ztest function to compare the writing scores of the two groups.
* The Z-statistic is calculated as -9.98. The negative value indicates that on average, female students scored higher than male students.
* The p-value is calculated as approximately 0.0.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

1. **EXPLORATORY DATA ANALYSIS (EDA)**

**UNIVARIATE ANALYSIS**

Univariate analysis focuses on exploring individual variables to understand their distribution, patterns, and anomalies.

| **Variable** | **Chart Type** | **Key Insights** |
| --- | --- | --- |
| Math Score | Histogram Plot | Most students scored between **60–80**, followed by **40–60**. Very few scored below 20 or above 90. |
| Parental Level of Education | Pie Chart | Highest share from **some college (22.6%)** and **associate’s degree (22.2%)**; lowest from **master’s degree (5.9%)**. |
| Race & Ethnicity | Count Plot | **Group C** has the highest representation, followed by **Group D**; **Group A** has the lowest. |

**A pie chart with text on it

AI-generated content may be incorrect.**

**A graph of a number of blue bars

AI-generated content may be incorrect.**

**A graph of a student math score

AI-generated content may be incorrect.**

**10.BIVARIATE ANALYSIS**

Bivariate analysis helps explore relationships between two variables, uncovering trends, correlations, and dependencies that impact housing prices.

**A graph of scatter plot

AI-generated content may be incorrect.**

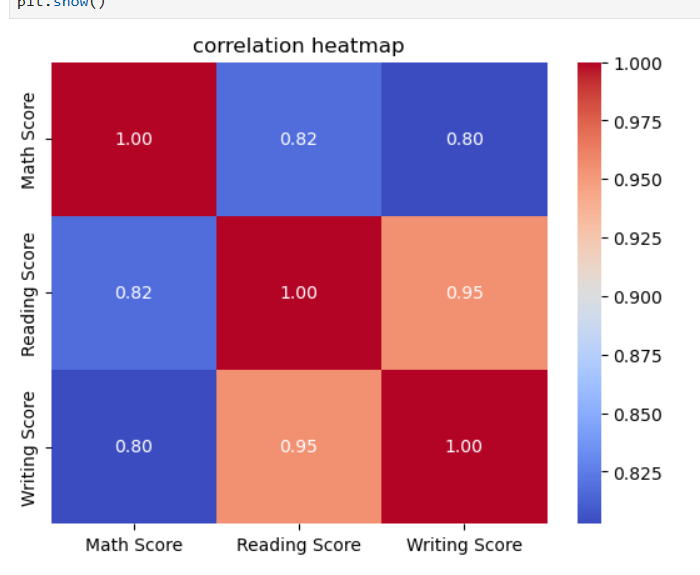
A graph of scatter plot

AI-generated content may be incorrect.

A graph showing a number of scatters

AI-generated content may be incorrect.

**11.MULTIVARAIATE ANALYSIS**



INSIGHTS:

* THIS CORRELATION HEATMAP TO REPRESENT TO THREE SCORES ARE **HIGHLY POSITIVELY CORRELATED (ABOVE 0.80)**
* **READING AND WRITING** SCORES HAVE **THE STRONGEST RELATIONSHIP (0.95)**
* **MATH SCORE** HAS SLIGHTLY LOWER CORRELATION WITH **READING (0.82) AND WRITING (0.80) COMPARED TO INTERRELATION**

A graph of different colored bars

AI-generated content may be incorrect.

INSIGHTS:

* **ASSOCAIATE'S DEGREE** AND **SOME COLLEGE** ARE THE TOP EDUCATION LEVELS FOR BOTH MALE AND FEMALE STUDENT
* MOST PARENTAL EDUCATION LEVELS FALL BEWTWEEN SOME HIGH SCHOOL AND ASSOCIATE DEGREE, WITH FEWER AT THE EXTREMES (LOW OR HIGH).

A graph with blue dots

AI-generated content may be incorrect.

INSIGHTS:

* MATH,READING,AND WRITING SCORES ARE ALL CLOSELY RELATED TO HIGH DCORES IN ONE SUBJECT OFTEN MEAN HIGH SCORES IN THE OTHERS
* HIGH PERFORMERS MANY STUDETN ARE CONCENTRATE IN THE UPPER RANGE (70-100) ACROSS ALL THREE SUBJECTS

**12.OVERALL INSIGHTS FROM ANALYSIS**

**1. General Observations:**

* The dataset contains 1000 rows and 8 columns of student performance data.
* The analysis identified and handled 25 missing values in the 'Lunch' column.
* No duplicate rows were found in the dataset.
  1. **Factors Influencing Performance:**
* **Gender:**

Female students, on average, perform better than male students in all three subjects: math, reading, and writing**.**

* **Parental Education:**

There's a positive relationship between a parent's education level and a student's test scores. Students with parents who have a master's or bachelor's degree tend to achieve higher scores compared to those whose parents have a high school education or less.

* **Test Preparation:**

Students who completed the test preparation course consistently scored higher than those who did not.

* **Lunch Type:**

Students with a standard lunch generally outperform those with a free or reduced lunch**.**

* 1. **Recommendations and Future Work**

Encourage all students to take the test preparation course to improve outcomes.

* Provide additional academic support to students receiving free or reduced lunch to help close the performance gap.
* Implement targeted interventions to support boys in reading and writing through literacy initiatives.
* Provide extra practice for girls in math to balance performance**.**

**13.CONCLUSION**

**Key Findings**

* **Socioeconomic Factors**: Students with a **standard lunch** and those with parents who have a **higher level of education** (e.g., Master's or Bachelor's degree) generally perform better. This suggests that both economic background and parental support play a role in academic success.
* **Demographic Factors**: **Female students** consistently outperform male students across all subjects, with a statistically significant difference in writing scores. This highlights a gender-based performance gap.
* **Proactive Measures**: Students who **completed a test preparation course** scored higher than those who did not. This indicates that proactive steps taken by students can directly improve their performance.

**Recommendations**

To improve overall student performance, the analysis suggests the following:

* Encourage all students to participate in **test preparation courses**.
* Implement **targeted support programs** for students from less-privileged backgrounds, such as those on free or reduced lunch.
* Develop **subject-specific interventions** to address gender-based performance gaps, such as providing additional literacy support for boys and extra math practice for girls.