W O M Fernando

Student No:  COMScDS241P-019

PRINCIPLES OF DATA SCIENCE

PORTFOLIO & GROUP ASSIGNMENT

Table of Contents

[Question 01 3](#_Toc192633730)

[Question 02 21](#_Toc192633731)

[References 22](#_Toc192633732)

[Question 2 22](#_Toc192633733)

# Question 01

1. **Data Collection:**

This dataset contains 5000 records from a private learning provider, and this includes key attributes to explore data, calculate correlations and insights related to academic performance.

A screenshot of a test

Description automatically generated

1. **Exploratory Data Analysis (EDA):**

Data cleaning and preprocessing

Detailed description of the dataset:

A table with numbers and a number of people

Description automatically generated with medium confidence

As per the above descriptive detail’s dataset contained a few missing values for the Attendance and Assignments Avg columns.

Following columns are identified as categorical data in the dataset:

* Gender
* Department
* Grade
* Extracurricular Activities
* Internet Access at Home
* Parent Education Level
* Family Income Level

Following columns are identified as numerical data in the dataset:

* Age
* Attendance (%)
* Midterm Score
* Final Score
* Assignments Avg
* Quizzes Avg
* Participation Score
* Projects Score
* Total Score
* Study Hours per Week
* Stress Level (1-10)
* Sleep Hours per Night

Following are the number of categories and their counts:

* Gender
  + Male – 2551
  + Female – 2449
* Department
  + CS – 2022
  + Engineering – 1469
  + Business – 1006
  + Mathematics – 503
* Grade
  + A – 1495
  + B – 978
  + C – 794
  + D – 889
  + F – 844
* Extracurricular Activities
  + No – 3493
  + Yes – 1507
* Internet Access at Home
  + Yes – 4485
  + No – 515
* Parent Education Level
  + PhD – 820
  + Bachelor’s – 810
  + High School – 796
  + Master’s – 780
* Family Income Level
  + Low – 1983
  + Medium – 1973
  + High - 1044

Identified Null values in the following columns:

Impute the missing numerical values with the mean. And impute the missing categorical values with the mode.

Missing numerical: Attendance (%), Assignments Average, Parent Education Level

A screenshot of a computer program

Description automatically generated

Descriptive statistics and visualizations

Descriptive statistics after cleaning the dataset

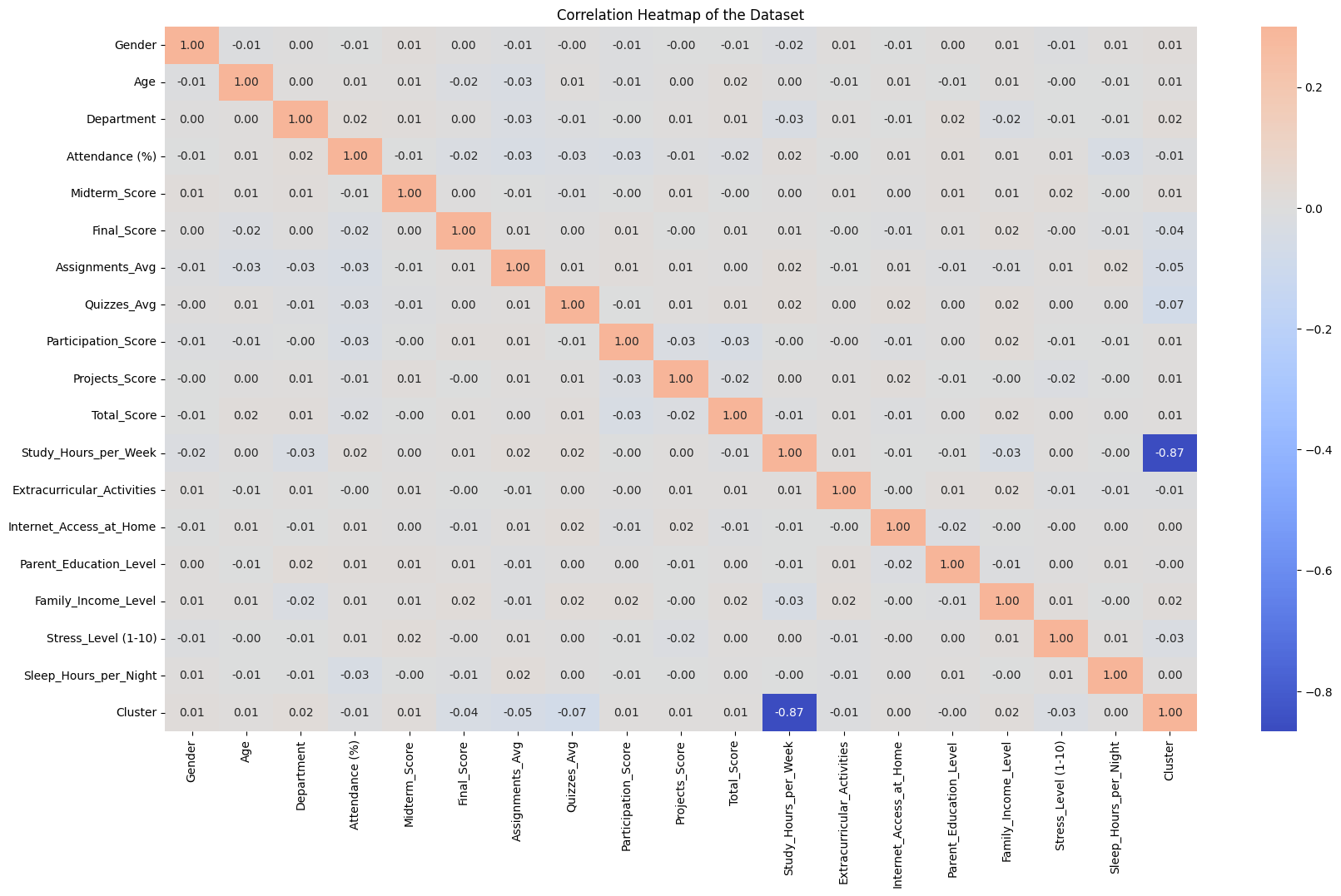
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| Gender | 5000 | 0.5 | 0.50 | 0 | 0.0 | 1.0 | 1.0 | 1 |
| Age | 5000 | 21.0 | 1.99 | 18 | 19.0 | 21.0 | 23.0 | 24 |
| Department | 5000 | 1.3 | 0.90 | 0 | 1.0 | 1.0 | 2.0 | 3 |
| Attendance (%) | 5000 | 75.4 | 13.61 | 50 | 64.7 | 75.4 | 86.2 | 100 |
| Midterm\_Score | 5000 | 70.3 | 17.21 | 40 | 55.5 | 70.5 | 85.0 | 100 |
| Final\_Score | 5000 | 69.6 | 17.24 | 40 | 54.7 | 69.7 | 84.5 | 100 |
| Assignments\_Avg | 5000 | 74.8 | 13.65 | 50 | 63.7 | 74.8 | 85.6 | 100 |
| Quizzes\_Avg | 5000 | 74.9 | 14.50 | 50 | 62.5 | 74.7 | 87.6 | 100 |
| Participation\_Score | 5000 | 5.0 | 2.89 | 0 | 2.4 | 5.0 | 7.5 | 10 |
| Projects\_Score | 5000 | 74.9 | 14.42 | 50 | 62.3 | 75.0 | 87.4 | 100 |
| Total\_Score | 5000 | 75.1 | 14.40 | 50 | 62.8 | 75.4 | 87.7 | 100 |
| Study\_Hours\_per\_Week | 5000 | 17.7 | 7.28 | 5 | 11.4 | 17.5 | 24.1 | 30 |
| Extracurricular\_Activities | 5000 | 0.3 | 0.46 | 0 | 0.0 | 0.0 | 1.0 | 1 |
| Internet\_Access\_at\_Home | 5000 | 0.9 | 0.30 | 0 | 1.0 | 1.0 | 1.0 | 1 |
| Parent\_Education\_Level | 5000 | 2.0 | 1.15 | 0 | 1.0 | 3.0 | 3.0 | 3 |
| Family\_Income\_Level | 5000 | 1.2 | 0.75 | 0 | 1.0 | 1.0 | 2.0 | 2 |
| Stress\_Level (1-10) | 5000 | 5.5 | 2.86 | 1 | 3.0 | 5.0 | 8.0 | 10 |
| Sleep\_Hours\_per\_Night | 5000 | 6.5 | 1.45 | 4 | 5.2 | 6.5 | 7.7 | 9 |

Histograms for numerical features:

A group of blue bars with numbers and lines

Description automatically generated with medium confidence

Correlation Heatmap:



Above heat map scale ranges from –0.2 (dark blue) to 1.0 (dark red). 1.0 (dark red) means a positive correlation (one variable increased the other increases too). 0 (light blue) shows no correlation between variables. Negative values (darker blue) show negative correlation. (one variable increases other one got decreased)

Key Observations:

Weak or no correlation (close to zero):

* Gender, Age, Department, internet access, parent education, family income, stress level, sleep hours per night show very weak correlation (close to zero). For example, Gender has a correlation of 0.01 with midterm score showing almost no relationship between the student gender and midterm score.
* Study hours per week also shows a weak correlation of –0.01 with total score showing that more study hours don’t translate into a better overall performance.

Weak negative correlation:

* Stress level shows a negative correlation with academic performance. This suggests higher stress might be associated with slightly lower academic performance, but the relation is weak.

Insights:

* External factors (gender, department, family income level, parent education level) have minimal or no impact on academic performance.
* Study habits and lifestyles like study hours, extracurricular activities, stress levels and sleep hours also have weak correlations with academic performance.
* We might expect higher attendance lead to higher grades but suggesting attendance might not be a strong predictor.

Scatter plots for key relationships:

A screenshot of a graph

Description automatically generated

Observations:

* Attendance: Can see a weak trend between the attendance and the scores. I suggest that attending more classes has a limited impact on the scores. This weak correlation can be seen in the above heatmap.
* Performance: There’s a positive trend between scores and student performance. Students who perform well in one subject tend to perform well overall.
* Data Distribution shape: This shows that the scores have a roughly a normal distribution. Attendance seems to have a skewed distribution with a peak of 100%.

Box plots to identify outliers:

A screenshot of a graph

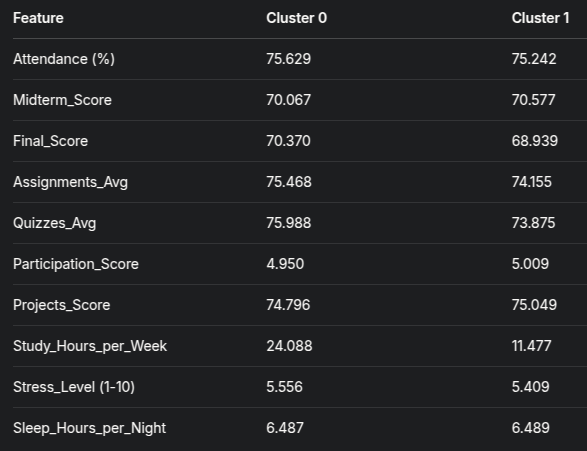
Description automatically generated

Observations:

* Attendance: This shows a outliers at 100%, we can suggest that a small number of students attend all classes.
* Categorical variables: Like gender, department and internet access have a slight variation showing that these variables are mostly uniform.
* Outliers: Outliers are present in most variables showing diversity in student population. For example, students with older age (Age outliers) and students with higher stress levels or students with exceptionally low academic ability stand out.

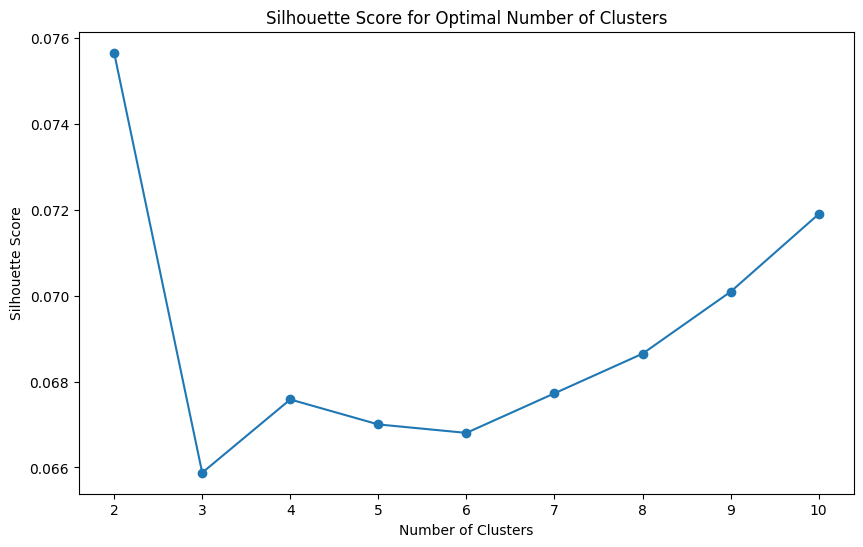
Clustering Analysis

* 1. We used K-mean clustering to identify the best number of clusters. As per the algorithm we got two as the best number of clusters. The following is the analyzed cluster data group by cluster and the feature.



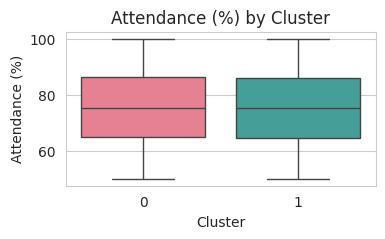
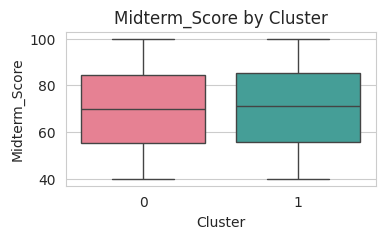
Insights from the above cluster analysis data:

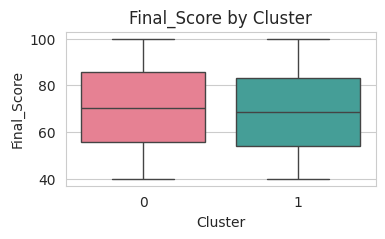
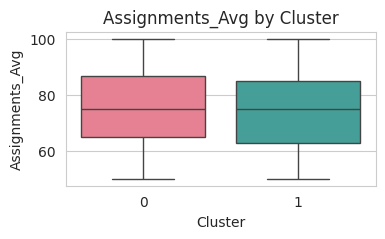
* Study Hours per Week: This shows a high difference between cluster 0 and cluster 1. Cluster 0 students study much more (24.088 hours/week) than cluster 1 students (11.477 hours/week).
* Academic scores are very similar between clusters. For example, midterm scores between clusters are (70.577 vs 70.067)
* Other features like sleep hours per night and attendance are very similar between clusters.

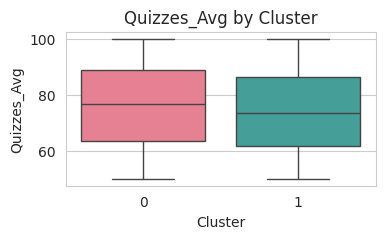
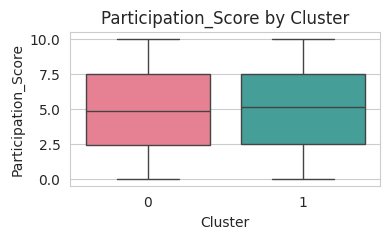


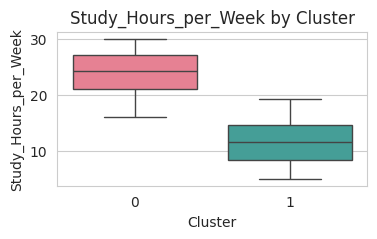
Optimal number of clusters: 2

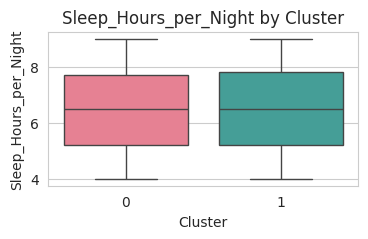
Visualize the cluster characteristics using box plots:

1. **Problem Solving**

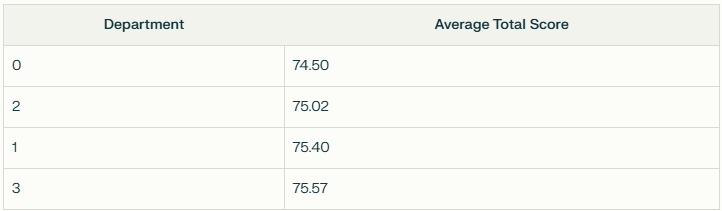
Identify the problem:

Students in some faculties tend to have low academic performance due to insufficient study time and high stress levels, which shows the need to improve students' academic performance and student well-being.

Proposed Solution:

We implement individualized intervention programs that focus on areas of underachievement and at-risk demographic groups and address academic support, stress management, and study habits.

Identify the underperforming departments:



Identify at-risk demographic groups based on stress level and study hours:

301 students identified with high stress level and low study hours.

1. **Model accuracy and business impact**

Model training and evaluation:

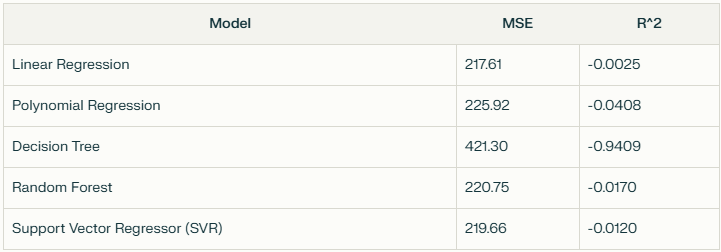
To find out the most accurately fit model, we trained a few selected models. Out of the trained models calculated MSC and R^2 values and plot the values to identify the best fit model.

Used models:

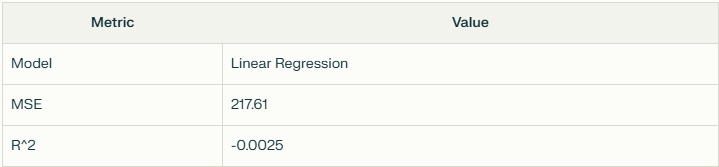
* Liner Regression
* Polynomial regression
* Decision tree
* Randon forest
* Support Vector Regression (SVR)

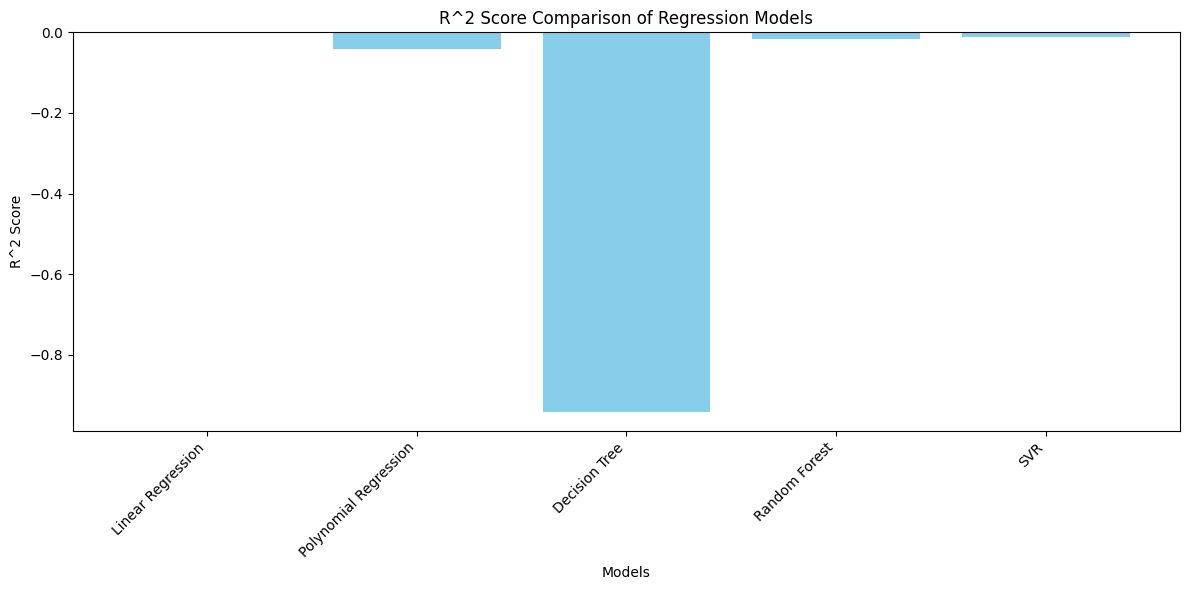
Comparison and selection:

Following are the resulted MSE and R^2 values from the above trained models.



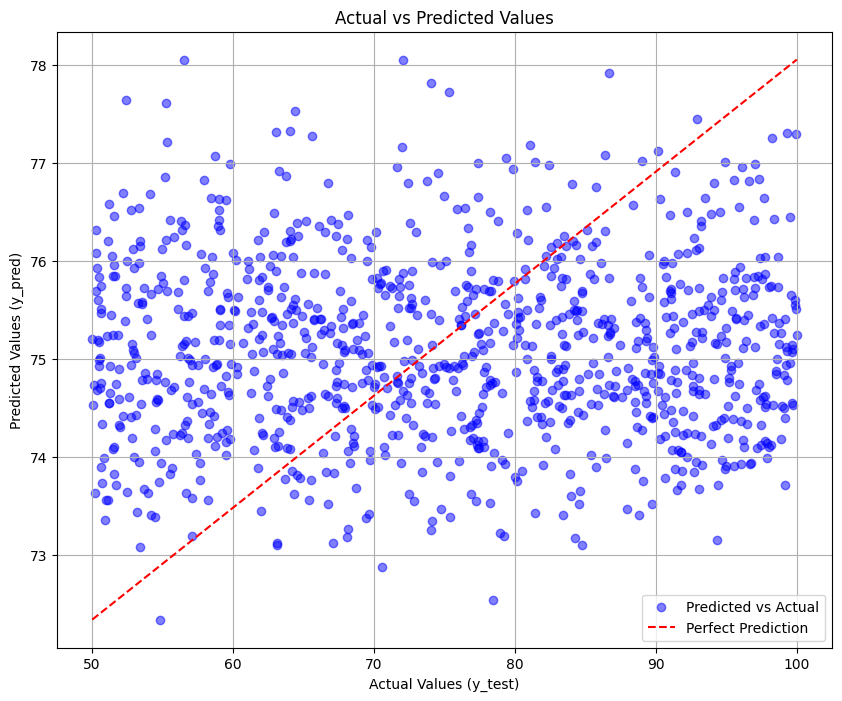
Best selected model:





Predicted data using Linear regression model:  
  
Actual vs Predicted Values  


Plotted actual vs predicted



Observations:

* Seems the data points have considerable spread which suggests the model has some predictions errors.
* Looking at the spread there can be some level of residuals in the model predictions.
* Fine tuning the features can improve the model prediction power.

# Question 02

Statistical Analysis of Yahoo’s Decline

**Abstract**: This report analyzes Yahoo's decline from its position as a dominant Internet company. Using publicly available financial data and relevant business statistics, the report identifies the key factors that contributed to Yahoo's collapse and gives a brief overview of the company's trajectory.

**Introduction**

Yahoo’s rise in the late 1990s represented the booming Internet era. In the beginning it’s a web directory, it quickly expanded into a fully-fledged portal offering search, email, news, and other content. However, Yahoo’s market position wear down dramatically in the decades that followed, culminating in its acquisition by Verizon in 2017. This analysis uses a business statistics perspective to understand the pivotal turns and decisions that led to this decline.

**Data and Methodology**

This analysis is based on openly accessible historical data, including:

* Revenue Data
* Market Capitalization
* User Statistics
* Strategic Missteps and Acquisition
* Competitor Data
* Comparative Analysis

**Analysis and Results**

Revenue Trends:

Yahoo's revenue peaked in 2008 and has been steadily declining since then. This decline has been attributed to several factors, including Google's dominance in search and advertising, as well as Yahoo's incapability to successfully adapt to the mobile Internet era.

Revenue Data:

A screenshot of a computer

Description automatically generated

Key Observation: With the revenue decline, from 2008 to 2016 Yahoo struggles to maintain the market position with its main competitor Google.1996 – 2005 period of robust growth with revenue peak at $5.26 billion in 2005. After 2005, decline in revenue and profitability significantly dropped with a net losses by 2015.

Market Capitalization Trends:

Yahoo's market capital went an all-time high during the dot-com bubble, with its stock price peaking at $118.75 in January 2000. However, after the bubble burst, Yahoo's market capital dropped drastically. By 2016, Yahoo's market capital was $48.08 billion, down from its peak in 2010.

Market Capital Data:

A screenshot of a screen

Description automatically generated

Revenue data comparison with its main competitor Google:

A screenshot of a graph

Description automatically generated

User Statistics:

Yahoo’s user engagement metrics, such as unique visitors and time spent on the platform, also declined. The table below compares Yahoo’s monthly unique visitors (in millions) to competitors like Google and Facebook.

A screenshot of a google search engine

Description automatically generated

Observations: While Yahoo has maintained a large user base, its growth rate has lagged significantly behind Facebook. Facebook’s social networking model has proven far more attractive, and provided superior data for targeted advertising. Yahoo’s attempts to create a similar social networking platform have largely failed.

Strategic Missteps and Acquisition:

Search Technology: Yahoo failed to develop or acquire superior search technology to compete with Google, and the decision to outsource search to Microsoft proved detrimental in the long run.

Mobile Strategy: Yahoo was slow to adapt to the mobile revolution and also Its mobile offerings were fragmented and lacked the cohesion of its competitors.

Innovation: Yahoo has struggled to foster a culture of innovation and failed to create innovative products that can drive and accelerate growth.

Overpriced Acquisitions: Yahoo purchased Tumblr for $1.1 billion in 2013 and later it was dropped to $712 million.

Missed Opportunities:

* Yahoo failed to acquire Google in 2002 and Facebook in 2006.
* Rejected Microsoft acquisition offer of $44.6 billion in 2008
* 2016 agreed to sell core internet operations to Verizon for $4.48 billion

From a business statistics perspective, the data clearly indicates that the company has failed to maintain sustainable growth in revenue, market value, and user engagement compared to its competitors. Strategic missteps in search, mobile, and acquisitions have worsened the situation. Yahoo's story serves as a warning about the importance of innovation, strategic decision-making, and adaptability in a rapidly evolving internet world. The data supports the conclusion that Yahoo's lack of ability to effectively address these challenges ultimately led to its decline and subsequent acquisition.

# References

## Question 2

* <https://www.investopedia.com/articles/markets/121015/how-yahoo-makes-money-yhoo.asp>
* <https://en.wikipedia.org/wiki/Yahoo!_Inc._(1995%E2%80%932017)>