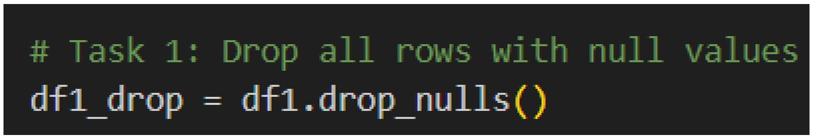
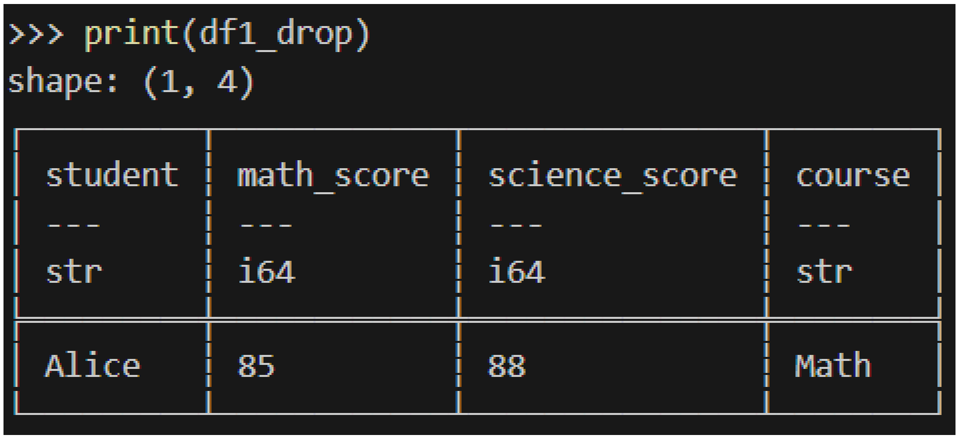
Name: Maclang, Waken Cean C.
Course: BSCS 3A

CSDS 312: Lab Activity 4
Date: September 29, 2025

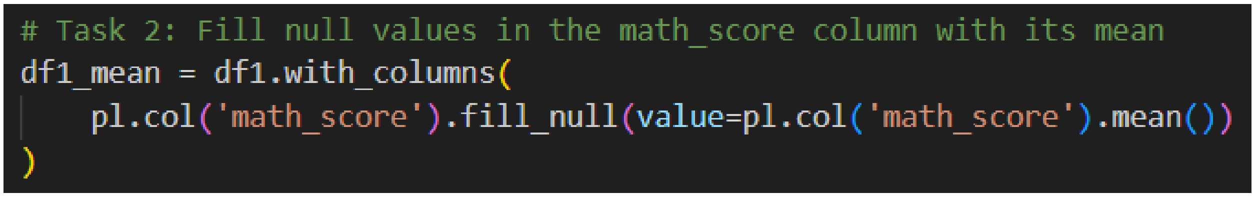
Exercise 1: Handling Missing Data



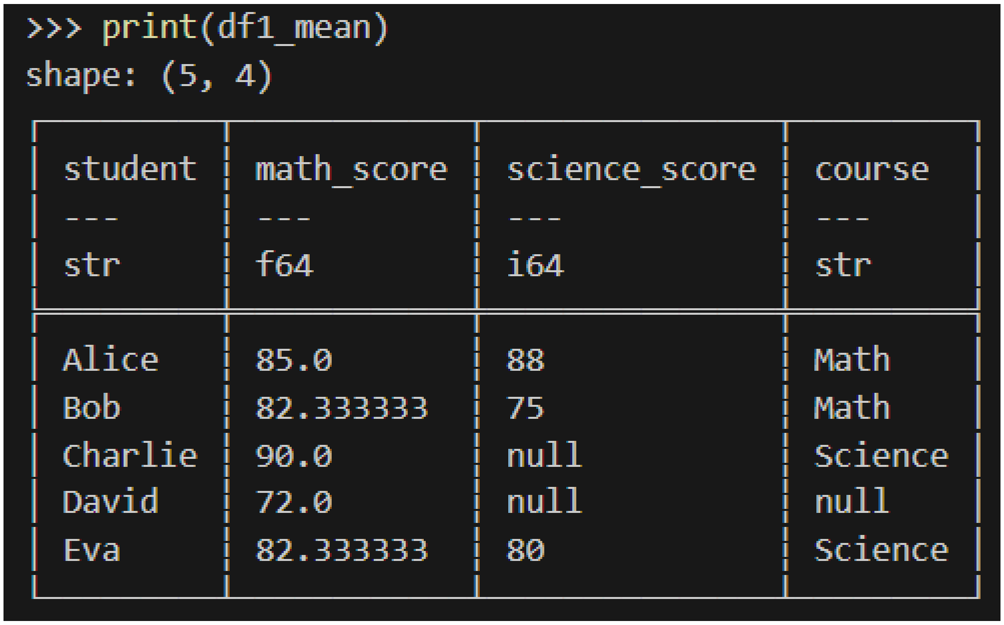
```python
# Task 1: Drop all rows with null values
df1_drop = df1.drop_nulls()
```



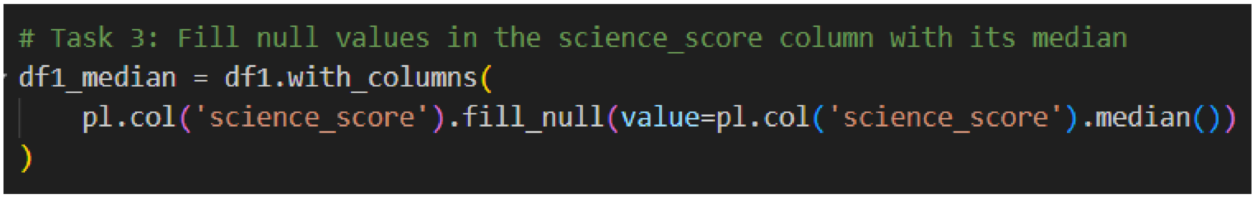
```
>>> print(df1_drop)
shape: (1, 4)
┌─────────┬────────────┬───────────────┬────────┐
│ student ┆ math_score ┆ science_score ┆ course │
│ ---     ┆ ---        ┆ ---           ┆ ---    │
│ str     ┆ i64        ┆ i64           ┆ str    │
╞═════════╪════════════╪═══════════════╪════════╡
│ Alice   ┆ 85         ┆ 88            ┆ Math   │
└─────────┴────────────┴───────────────┴────────┘
```



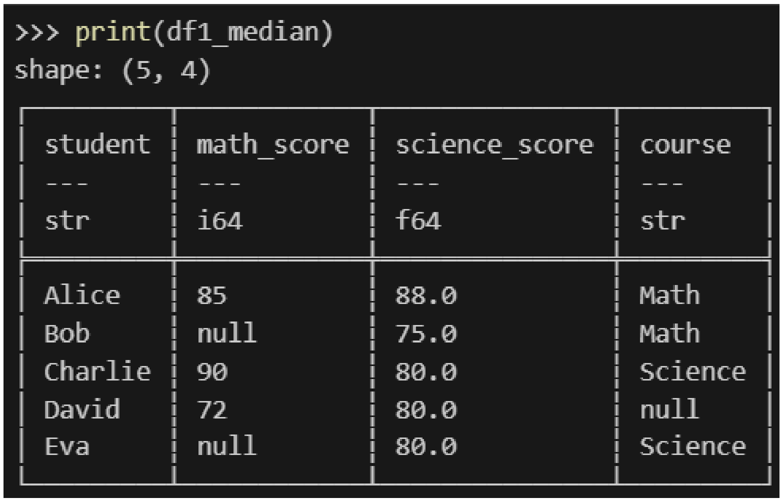
```python
# Task 2: Fill null values in the math_score column with its mean
df1_mean = df1.with_columns(
    pl.col('math_score').fill_null(value=pl.col('math_score').mean())
)
```



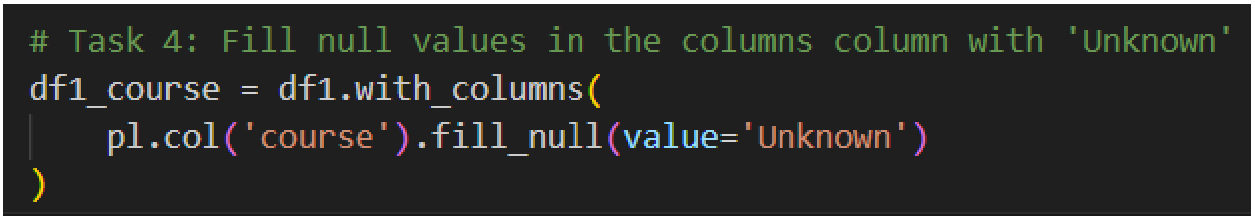
```
>>> print(df1_mean)
shape: (5, 4)
┌─────────┬────────────┬───────────────┬─────────┐
│ student ┆ math_score ┆ science_score ┆ course  │
│ ---     ┆ ---        ┆ ---           ┆ ---     │
│ str     ┆ f64        ┆ i64           ┆ str     │
╞═════════╪════════════╪═══════════════╪═════════╡
│ Alice   ┆ 85.0       ┆ 88            ┆ Math    │
│ Bob     ┆ 82.333333  ┆ 75            ┆ Math    │
│ Charlie ┆ 90.0       ┆ null          ┆ Science │
│ David   ┆ 72.0       ┆ null          ┆ null    │
│ Eva     ┆ 82.333333  ┆ 80            ┆ Science │
└─────────┴────────────┴───────────────┴─────────┘
```



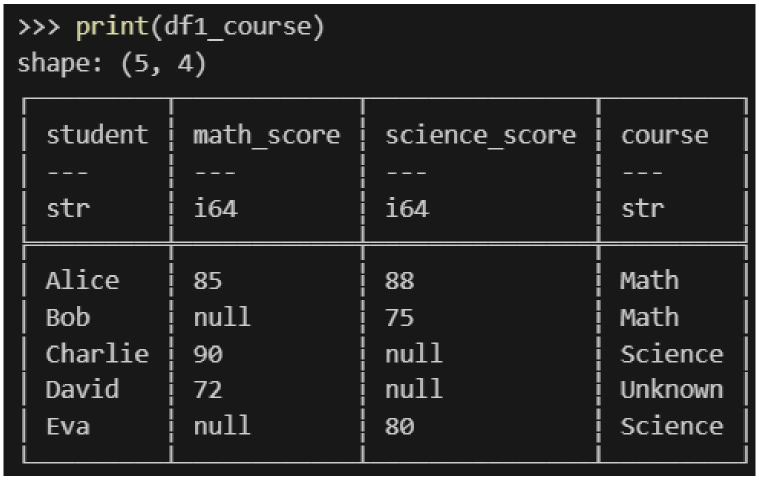
```python
# Task 3: Fill null values in the science_score column with its median
df1_median = df1.with_columns(
    pl.col('science_score').fill_null(value=pl.col('science_score').median())
)
```



```
>>> print(df1_median)
shape: (5, 4)
┌─────────┬────────────┬───────────────┬─────────┐
│ student ┆ math_score ┆ science_score ┆ course  │
│ ---     ┆ ---        ┆ ---           ┆ ---     │
│ str     ┆ i64        ┆ f64           ┆ str     │
╞═════════╪════════════╪═══════════════╪═════════╡
│ Alice   ┆ 85         ┆ 88.0          ┆ Math    │
│ Bob     ┆ null       ┆ 75.0          ┆ Math    │
│ Charlie ┆ 90         ┆ 80.0          ┆ Science │
│ David   ┆ 72         ┆ 80.0          ┆ null    │
│ Eva     ┆ null       ┆ 80.0          ┆ Science │
└─────────┴────────────┴───────────────┴─────────┘
```



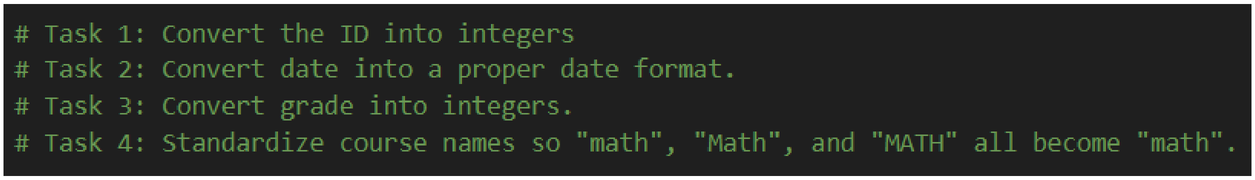
```python
# Task 4: Fill null values in the columns column with 'Unknown'
df1_course = df1.with_columns(
    pl.col('course').fill_null(value='Unknown')
)
```



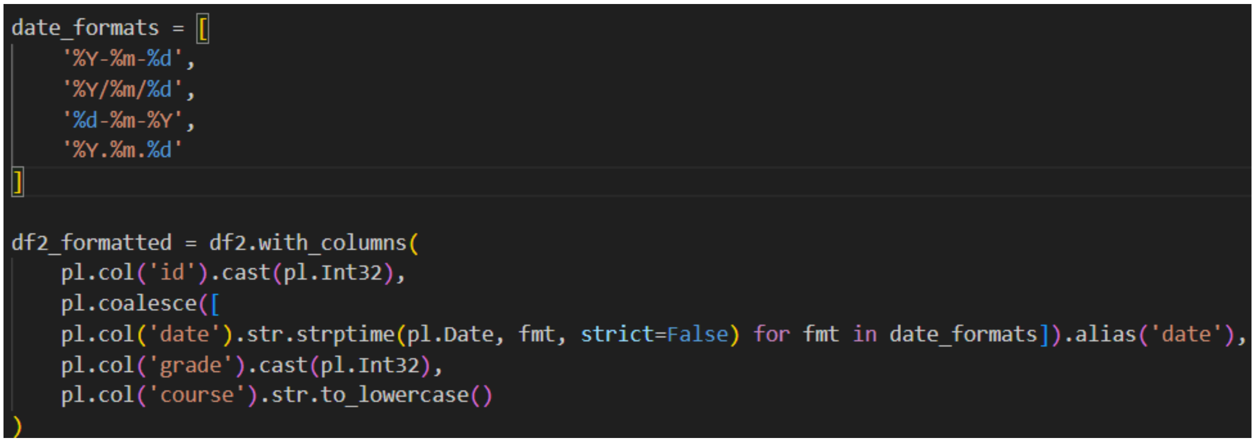
```
>>> print(df1_course)
shape: (5, 4)
┌─────────┬────────────┬───────────────┬─────────┐
│ student ┆ math_score ┆ science_score ┆ course  │
│ ---     ┆ ---        ┆ ---           ┆ ---     │
│ str     ┆ i64        ┆ i64           ┆ str     │
╞═════════╪════════════╪═══════════════╪═════════╡
│ Alice   ┆ 85         ┆ 88            ┆ Math    │
│ Bob     ┆ null       ┆ 75            ┆ Math    │
│ Charlie ┆ 90         ┆ null          ┆ Science │
│ David   ┆ 72         ┆ null          ┆ Unknown │
│ Eva     ┆ null       ┆ 80            ┆ Science │
└─────────┴────────────┴───────────────┴─────────┘
```

Name: Maclang, Waken Cean C. CSDS 312: Lab Activity 4
Course: BSCS 3A **Date:** September 29, 2025

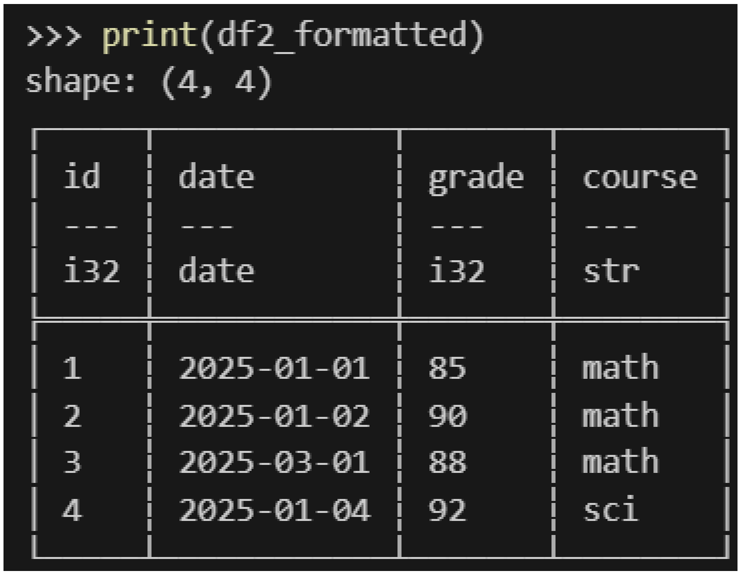
Exercise 2: Formatting Data



```
# Task 1: Convert the ID into integers
# Task 2: Convert date into a proper date format.
# Task 3: Convert grade into integers.
# Task 4: Standardize course names so "math", "Math", and "MATH" all become "math".
```



```
date_formats = [
    '%Y-%m-%d',
    '%Y/%m/%d',
    '%d-%m-%Y',
    '%Y.%m.%d'
]

df2_formatted = df2.with_columns(
    pl.col('id').cast(pl.Int32),
    pl.coalesce([
    pl.col('date').str.strptime(pl.Date, fmt, strict=False) for fmt in date_formats]).alias('date'),
    pl.col('grade').cast(pl.Int32),
    pl.col('course').str.to_lowercase()
)
```

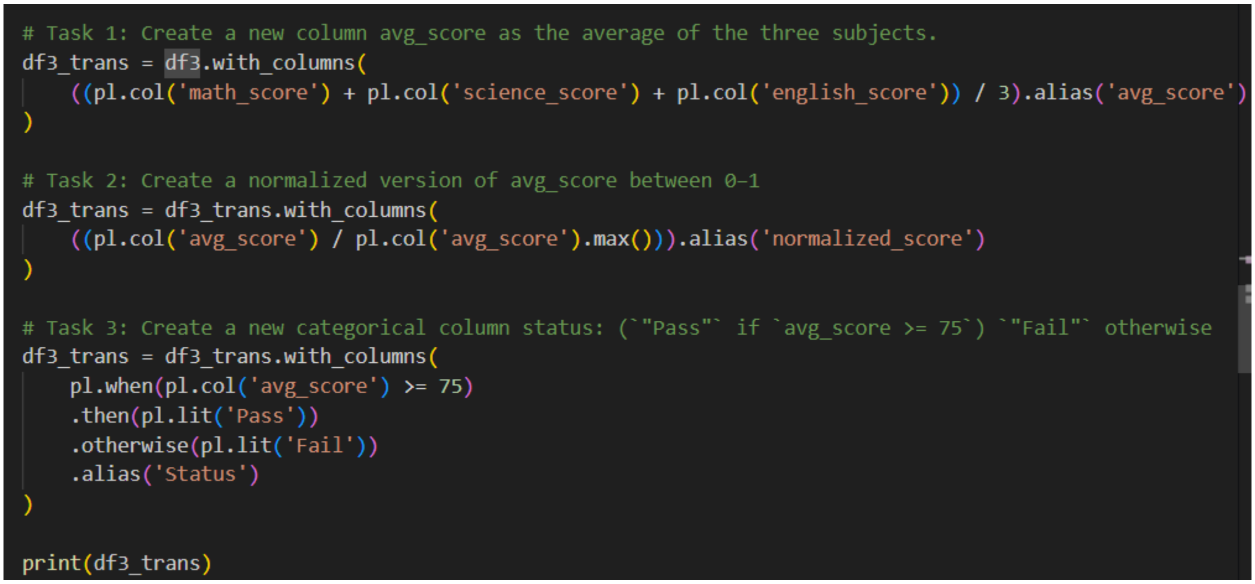


```
>>> print(df2_formatted)
shape: (4, 4)
┌─────┬────────────┬───────┬────────┐
│ id  ┆ date       ┆ grade ┆ course │
│ --- ┆ ---        ┆ ---   ┆ ---    │
│ i32 ┆ date       ┆ i32   ┆ str    │
╞═════╪════════════╪═══════╪════════╡
│ 1   ┆ 2025-01-01 ┆ 85    ┆ math   │
│ 2   ┆ 2025-01-02 ┆ 90    ┆ math   │
│ 3   ┆ 2025-03-01 ┆ 88    ┆ math   │
│ 4   ┆ 2025-01-04 ┆ 92    ┆ sci    │
└─────┴────────────┴───────┴────────┘
```

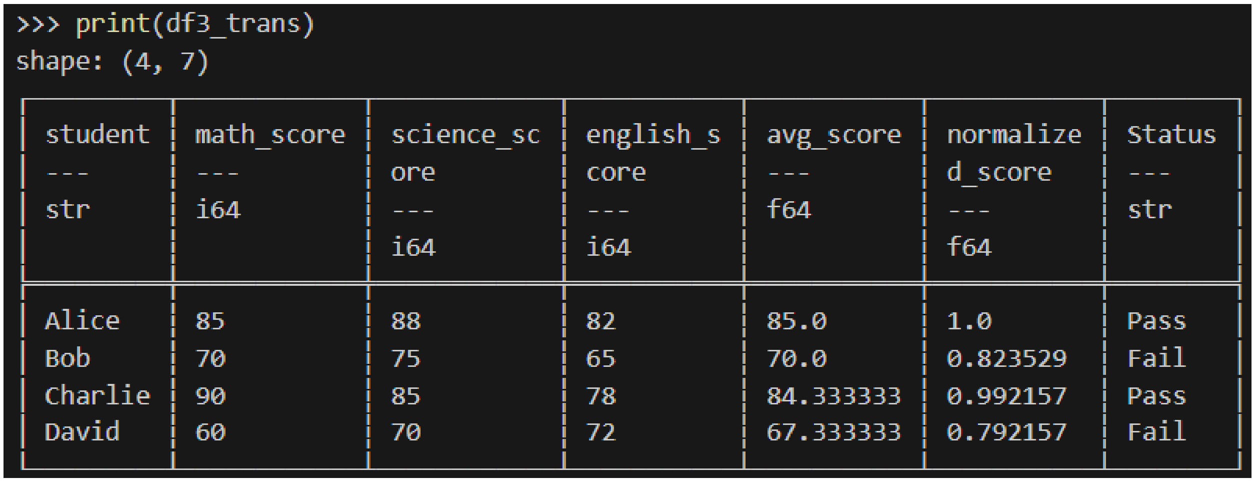
Name: Maclang, Waken Cean C.
Course: BSCS 3A

CSDS 312: Lab Activity 4
Date: September 29, 2025

Exercise 3: Transforming Data

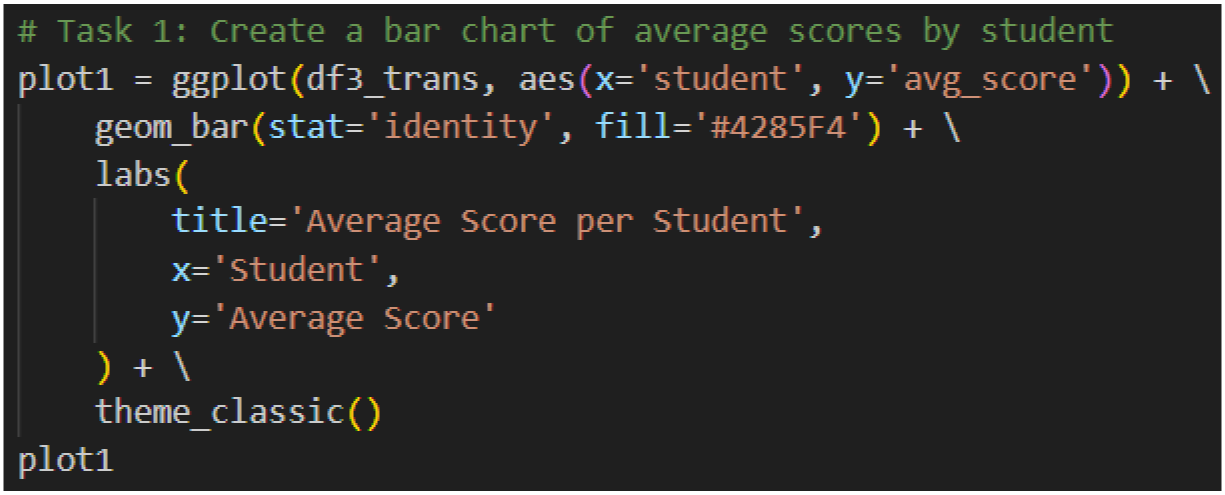


```
# Task 1: Create a new column avg_score as the average of the three subjects.
df3_trans = df3.with_columns(
    ((pl.col('math_score') + pl.col('science_score') + pl.col('english_score')) / 3).alias('avg_score')
)

# Task 2: Create a normalized version of avg_score between 0-1
df3_trans = df3_trans.with_columns(
    ((pl.col('avg_score') / pl.col('avg_score').max())).alias('normalized_score')
)

# Task 3: Create a new categorical column status: (`"Pass"` if `avg_score >= 75`) `"Fail"` otherwise
df3_trans = df3_trans.with_columns(
    pl.when(pl.col('avg_score') >= 75)
    .then(pl.lit('Pass'))
    .otherwise(pl.lit('Fail'))
    .alias('Status')
)

print(df3_trans)
```

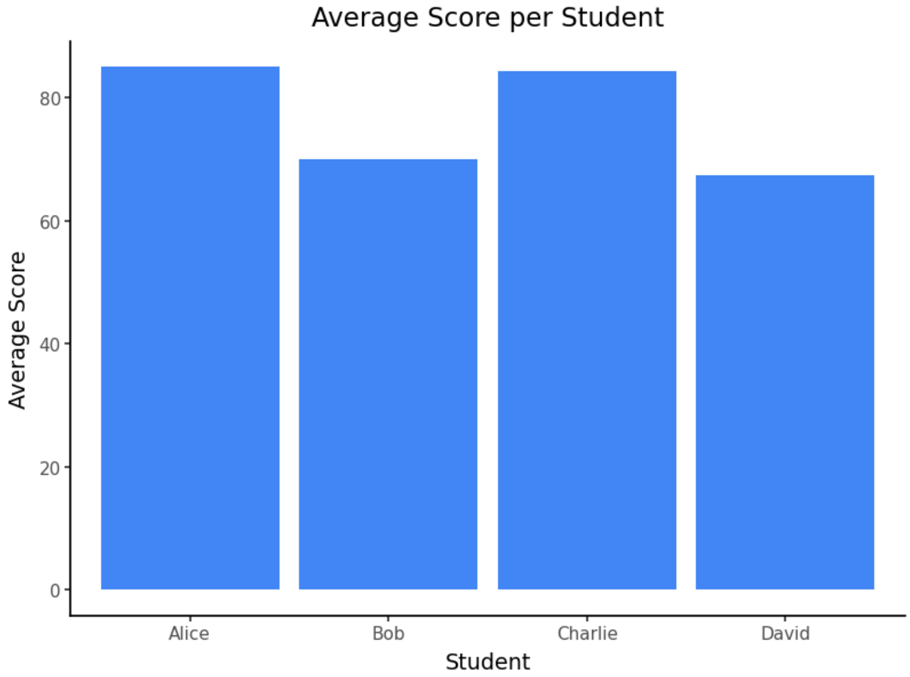


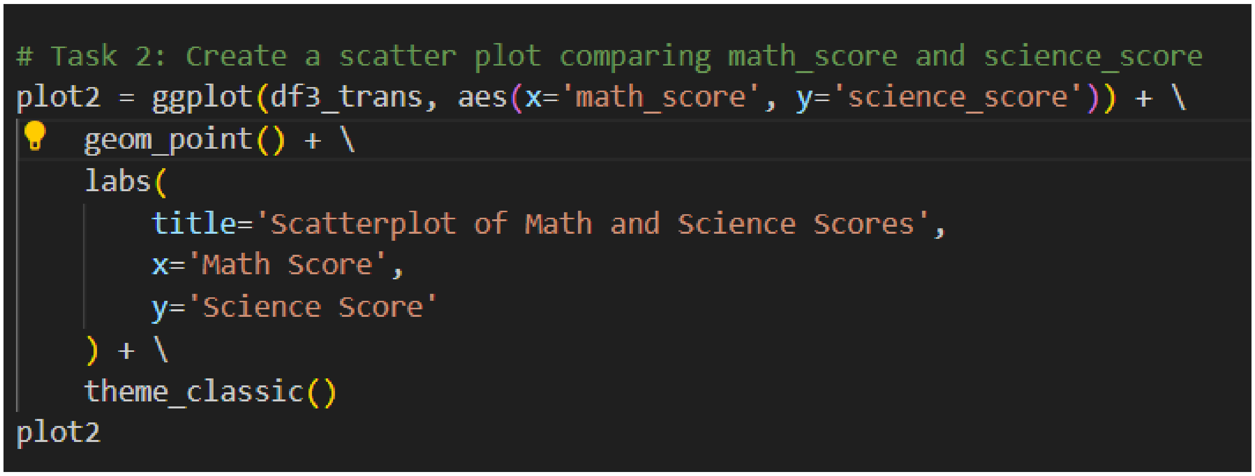
```
>>> print(df3_trans)
shape: (4, 7)
┌─────────┬────────────┬────────────┬────────────┬───────────┬────────────┬────────┐
│ student ┆ math_score ┆ science_sc ┆ english_s  ┆ avg_score ┆ normalize  ┆ Status │
│ ---     ┆ ---        ┆ ore        ┆ core       ┆ ---       ┆ d_score    ┆ ---    │
│ str     ┆ i64        ┆ ---        ┆ ---        ┆ f64       ┆ ---        ┆ str    │
│         ┆            ┆ i64        ┆ i64        ┆           ┆ f64        ┆        │
╞═════════╪════════════╪════════════╪════════════╪═══════════╪════════════╪════════╡
│ Alice   ┆ 85         ┆ 88         ┆ 82         ┆ 85.0      ┆ 1.0        ┆ Pass   │
│ Bob     ┆ 70         ┆ 75         ┆ 65         ┆ 70.0      ┆ 0.823529   ┆ Fail   │
│ Charlie ┆ 90         ┆ 85         ┆ 78         ┆ 84.333333 ┆ 0.992157   ┆ Pass   │
│ David   ┆ 60         ┆ 70         ┆ 72         ┆ 67.333333 ┆ 0.792157   ┆ Fail   │
└─────────┴────────────┴────────────┴────────────┴───────────┴────────────┴────────┘
```

Exercise 4: Visualization of Cleaned Data

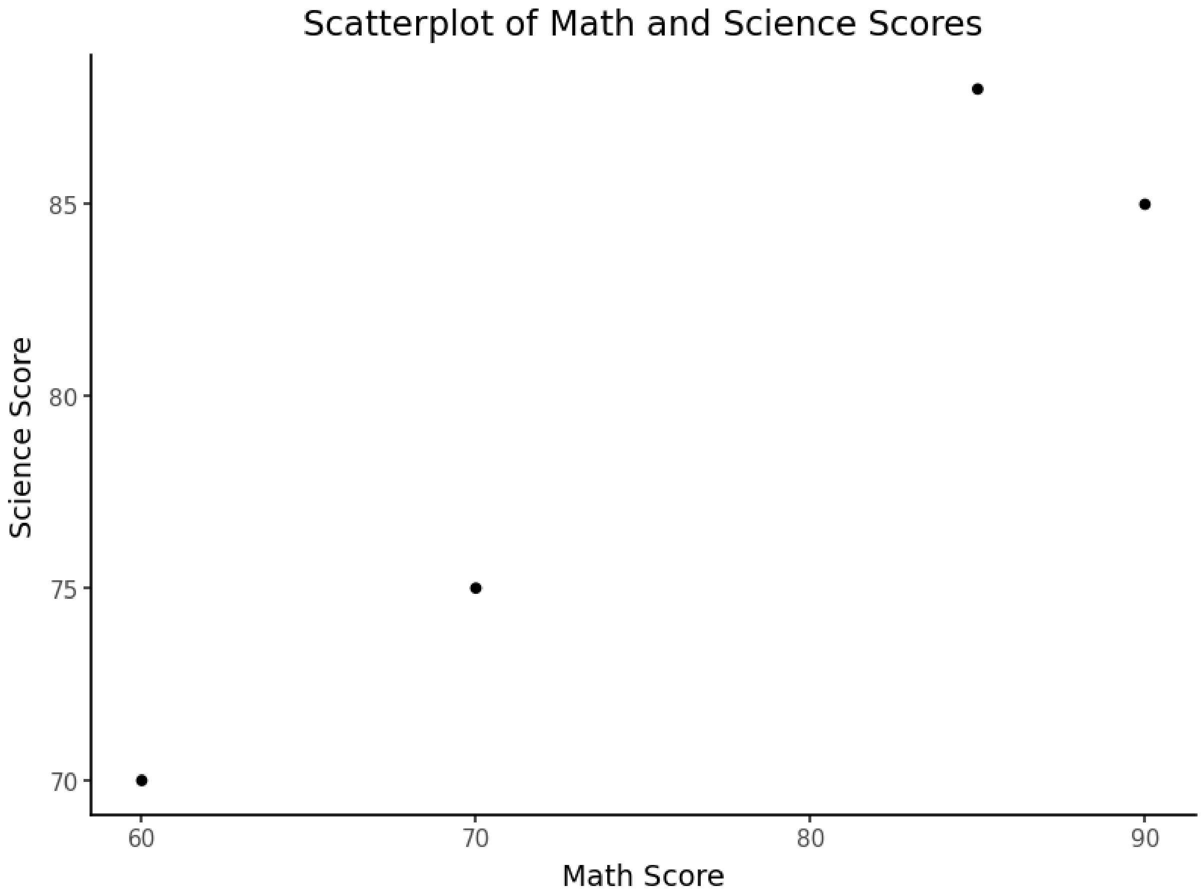


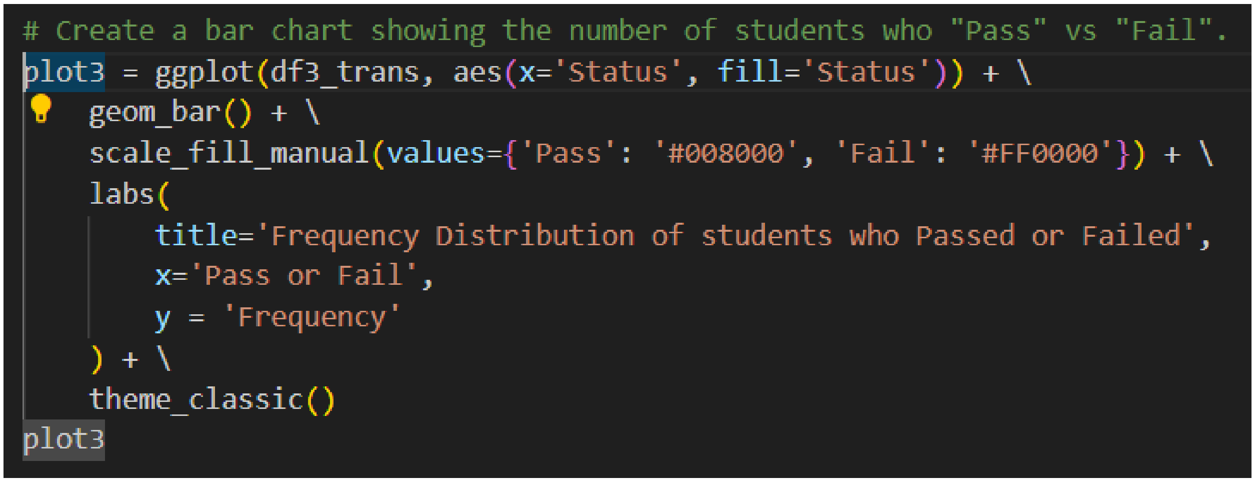
```
# Task 1: Create a bar chart of average scores by student
plot1 = ggplot(df3_trans, aes(x='student', y='avg_score')) + \
    geom_bar(stat='identity', fill='#4285F4') + \
    labs(
        title='Average Score per Student',
        x='Student',
        y='Average Score'
    ) + \
    theme_classic()
plot1
```



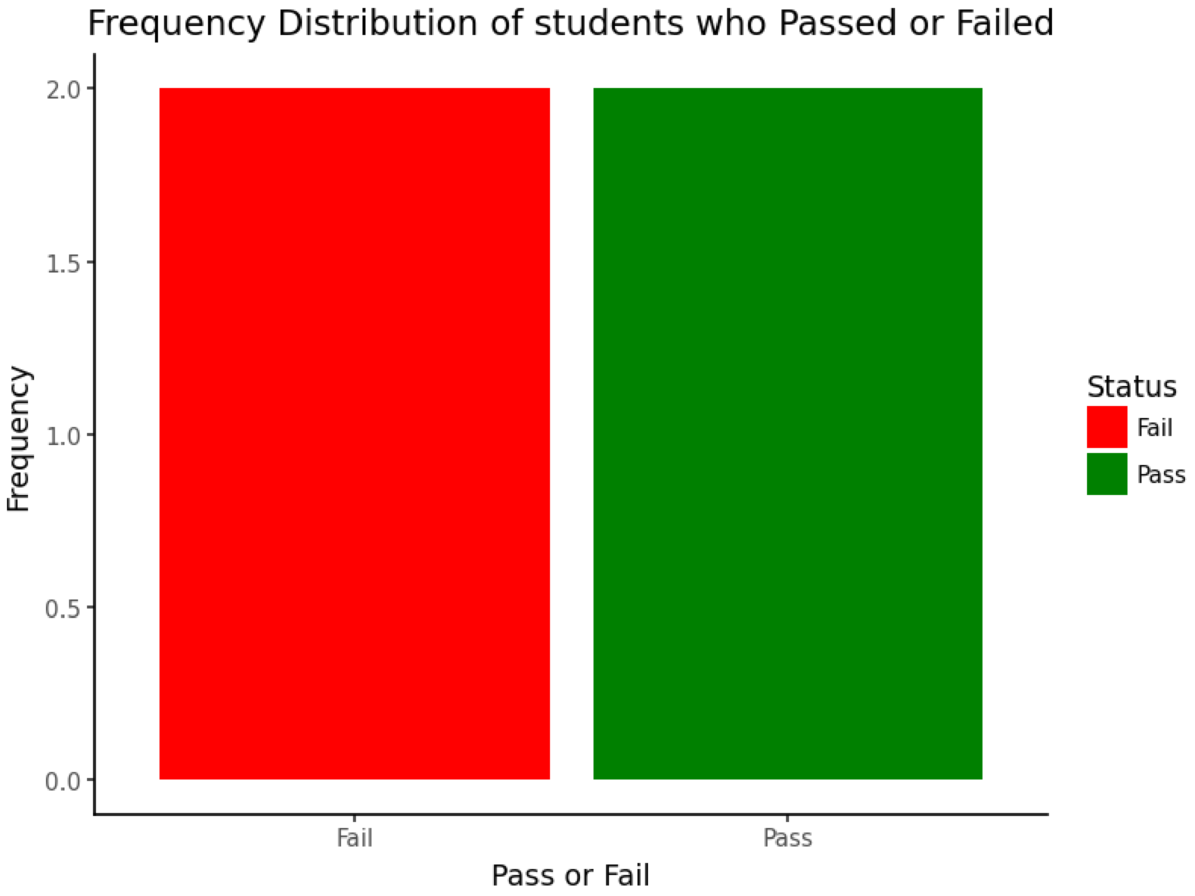


```
# Task 2: Create a scatter plot comparing math_score and science_score
plot2 = ggplot(df3_trans, aes(x='math_score', y='science_score')) + \
    geom_point() + \
    labs(
        title='Scatterplot of Math and Science Scores',
        x='Math Score',
        y='Science Score'
    ) + \
    theme_classic()
plot2
```





```
# Create a bar chart showing the number of students who "Pass" vs "Fail".
plot3 = ggplot(df3_trans, aes(x='Status', fill='Status')) + \
    geom_bar() + \
    scale_fill_manual(values={'Pass': '#008000', 'Fail': '#FF0000'}) + \
    labs(
        title='Frequency Distribution of students who Passed or Failed',
        x='Pass or Fail',
        y = 'Frequency'
    ) + \
    theme_classic()
plot3
```


Frequency Distribution of students who Passed or Failed
Frequency
Fail
Pass
Pass or Fail
Status
🟥 Fail
🟩 Pass