Alcohol data consumption Analysis

Team 3: RuntimeTerror Anvi, Aruna Atreyi, Gauri, Riddhi, Vaibhavi

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


import warnings
warnings.filterwarnings('ignore')
```

```
df1 = pd.read_excel('/student-por.xlsx')
df2 = pd.read_excel('/student-por.xlsx')
```

```
# Add subject column
df1['subject'] = 'Math'
df2['subject'] = 'Portuguese'

# Combine datasets
df = pd.concat([df1, df2], ignore_index=True)

# Drop irrelevant columns if they exist
cols_to_drop = ['nursery', 'school']  # Add more if needed
df.drop(columns=[col for col in cols_to_drop if col in df.columns], inplace=True)

# Create Sum/60
df['Sum/60'] = df['G1'] + df['G2'] + df['G3']
```

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 34 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64 
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64 
 7   Fedu        649 non-null    int64 
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64 
 13  studytime   649 non-null    int64 
 14  failures    649 non-null    int64 
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64 
 24  freetime    649 non-null    int64 
 25  goout       649 non-null    int64 
 26  Dalc        649 non-null    int64 
 27  Walc        649 non-null    int64 
 28  health      649 non-null    int64 
 29  absences    649 non-null    int64 
 30  G1          649 non-null    int64 
 31  G2          649 non-null    int64 
 32  G3          649 non-null    int64 
 33  subject     649 non-null    object
dtypes: int64(16), object(18)
memory usage: 172.5+ KB
```

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 34 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64 
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64 
 7   Fedu        649 non-null    int64 
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64 
 13  studytime   649 non-null    int64 
 14  failures    649 non-null    int64 
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64 
 24  freetime    649 non-null    int64 
 25  goout       649 non-null    int64 
 26  Dalc        649 non-null    int64 
 27  Walc        649 non-null    int64 
 28  health      649 non-null    int64 
 29  absences    649 non-null    int64 
 30  G1          649 non-null    int64 
 31  G2          649 non-null    int64 
 32  G3          649 non-null    int64 
 33  subject     649 non-null    object
dtypes: int64(16), object(18)
memory usage: 172.5+ KB
```

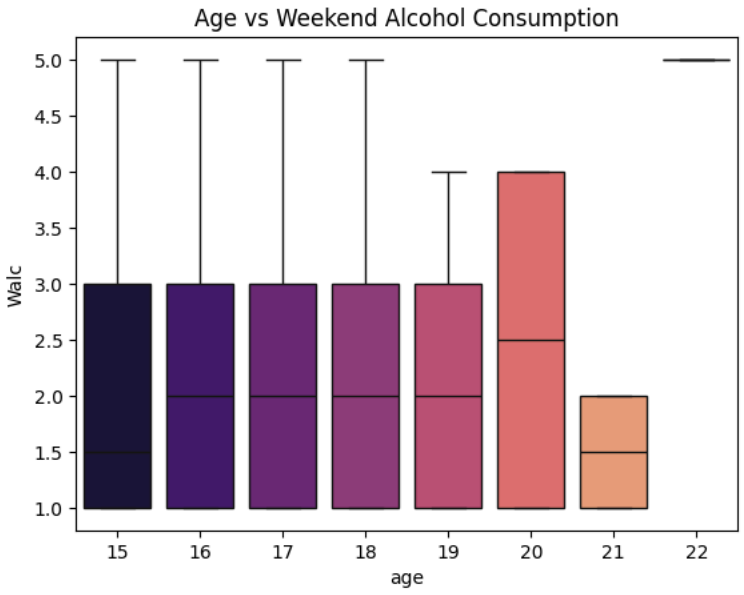
```
df1.describe()
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Wal
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.00000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880	3.930663	3.180277	3.184900	1.502311	2.28043
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235	0.955717	1.051093	1.175766	0.924834	1.28438
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.00000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.000000	1.00000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000	1.000000	2.00000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.000000	3.00000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.000000	5.00000

```
df2.describe()
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Wal
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.00000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880	3.930663	3.180277	3.184900	1.502311	2.28043
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235	0.955717	1.051093	1.175766	0.924834	1.28438
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.00000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.000000	1.00000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000	1.000000	2.00000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.000000	3.00000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.000000	5.00000

```
# Q1: What age is most prone to get into drinking habits?
sns.boxplot(data=df, x='age', y='Walc', palette='magma')
plt.title("Age vs Weekend Alcohol Consumption")
plt.show()
```

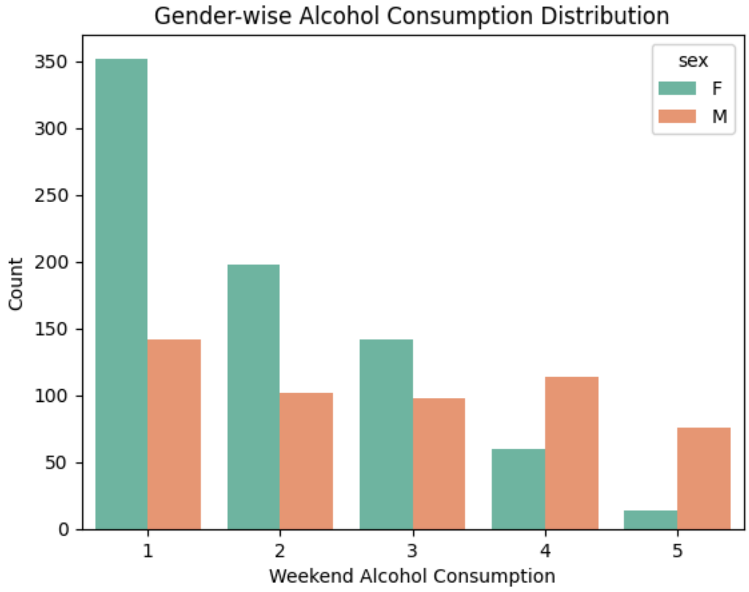


The above plot shows us how age impacts alcohol consumption in teenagers (per week) shown in the form of a box plot.

1. from the above representation we can tell that student of age group 20 have the most weekly alcohol consumption.
2. age groups 15-19 have the same alcohol consumption.

(basis of quantity of alc consumed)

```
# Q2: Which gender drinks more?
sns.countplot(data=df, x='Walc', hue='sex', palette='Set2')
plt.title("Gender-wise Alcohol Consumption Distribution")
plt.xlabel("Weekend Alcohol Consumption")
plt.ylabel("Count")
plt.show()
```



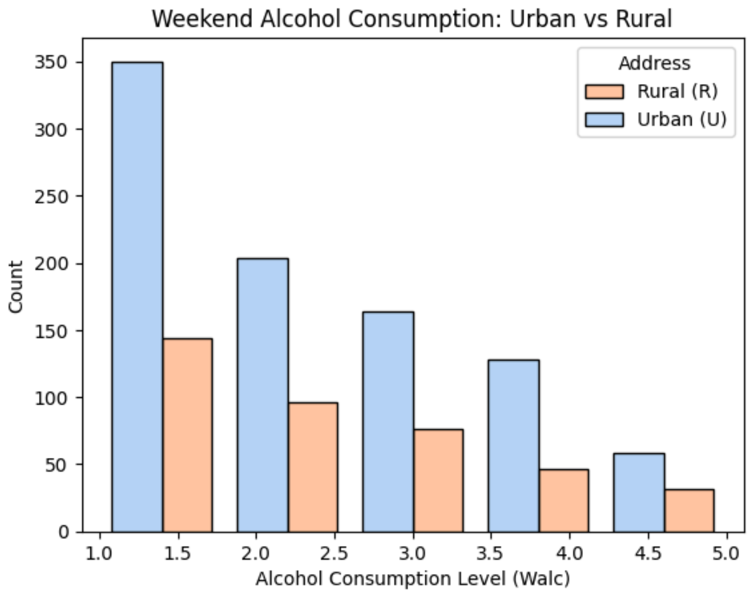
The above countplot shows the frequency of alcohol consumption betweeen male and female.

1. we can see that count of females descrease as quantity of alcohol consumed increases on a weekly basis.
2. on the other hand , count of men increases where there is more alcohol consumption

thus, we can say that among teens males drink more alcohol than females (if we consider quantity of alcohol consumed)

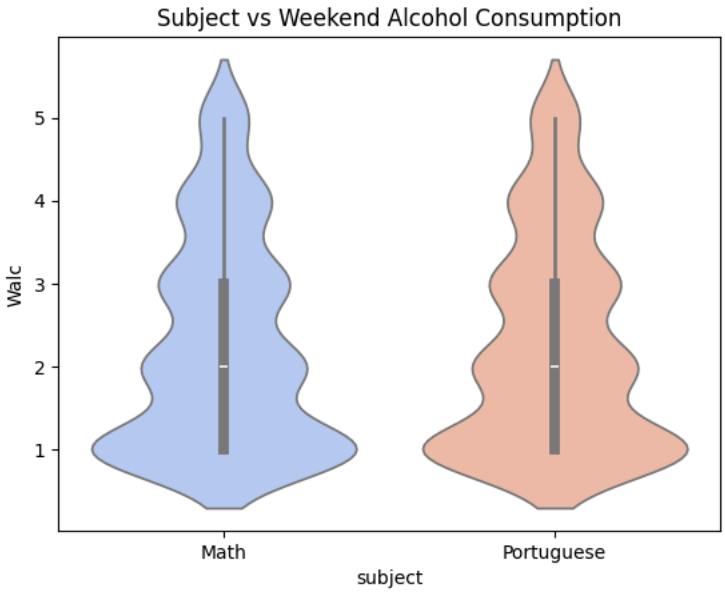
and more females drink than male (if the count of students is considered)

```
# Q3: Urban/rural area’s influence on alcoholism
sns.histplot(data=df, x='Walc', hue='address', multiple='dodge', bins=5, shrink=0.8, palette='pastel')
plt.title("Weekend Alcohol Consumption: Urban vs Rural")
plt.xlabel("Alcohol Consumption Level (Walc)")
plt.ylabel("Count")
plt.legend(title='Address', labels=['Rural (R)', 'Urban (U)'])
plt.show()
```



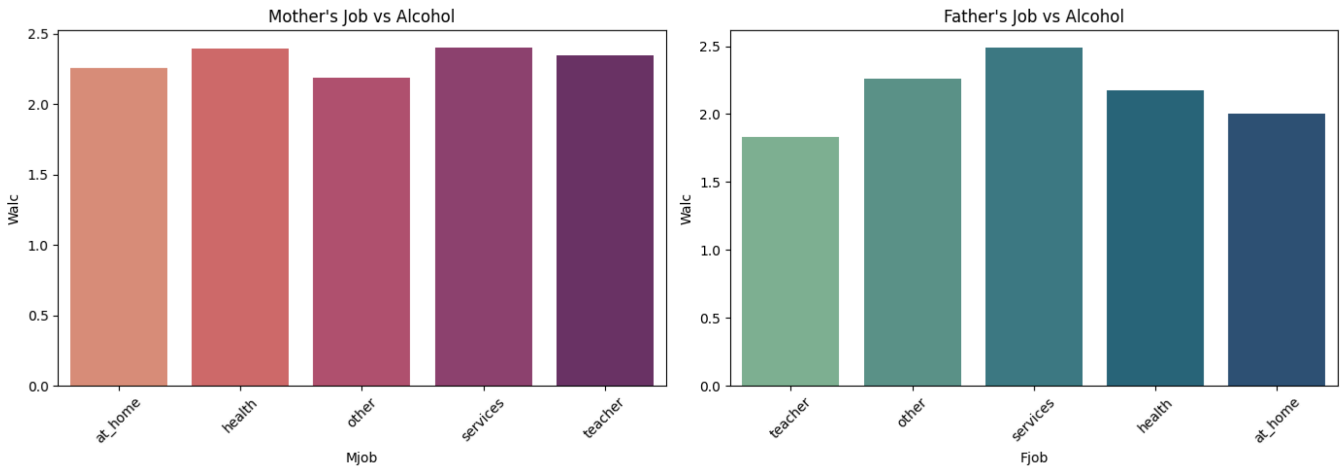
This histogram shows us the weekly alcohol consumption between urban students and rural students. From the plot we can see that urban students drink more alcohol weekly.

```
# Q4: Subject analysis – Which subject’s students drink more?
df['subject'] = ['Math'] * len(df1) + ['Portuguese'] * len(df2)
sns.violinplot(data=df, x='subject', y='Dalc', palette='coolwarm')
plt.title("Subject vs Weekend Alcohol Consumption")
plt.show()
```



The above violin plot shows amount of alcohol consumed per subject. We infer that there is not much disparity in levels of alcohol consumed. Hence we can say that subjects don't have much influence over alcohol consumption

```
# Q5: How does the type of parental job influence alcohol consumption?
fig, axs = plt.subplots(1, 2, figsize=(14, 5))
sns.barplot(data=df, x='Mjob', y='Walc', ci=None, ax=axs[0], palette='flare')
axs[0].set_title("Mother's Job vs Alcohol")
axs[0].tick_params(axis='x', rotation=45)
sns.barplot(data=df, x='Fjob', y='Walc', ci=None, ax=axs[1], palette='crest')
axs[1].set_title("Father's Job vs Alcohol")
axs[1].tick_params(axis='x', rotation=45)
plt.tight_layout()
plt.show()
```

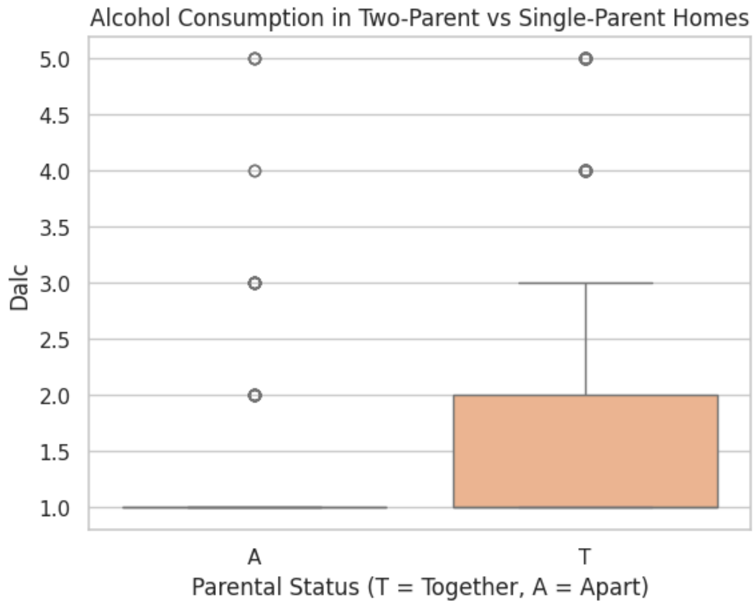


Overall Insights: Service sector jobs (for both parents) seem to correlate with higher alcohol consumption among students.

Parental involvement or presence at home (especially mothers) appears to relate to slightly lower alcohol consumption.

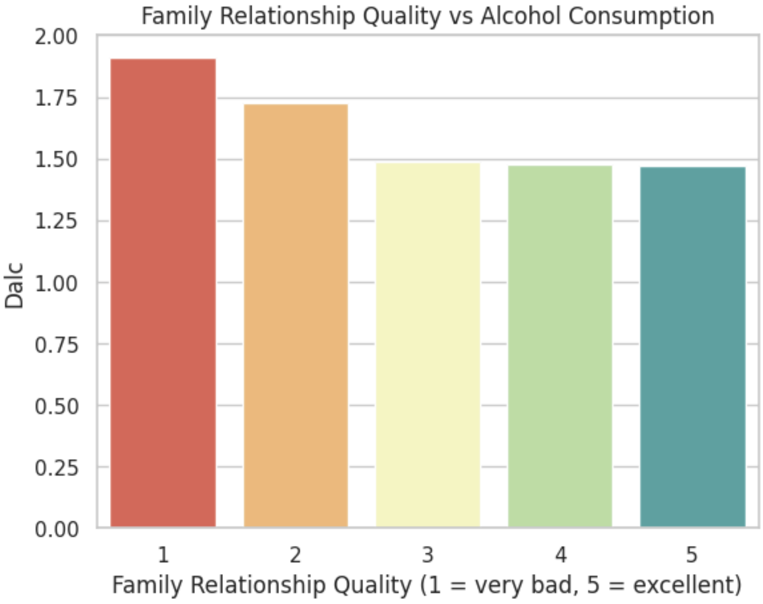
There is a more noticeable variation in alcohol consumption based on father’s job than mother’s, possibly indicating stronger correlation or influence.

```
# Q6: Alcohol difference between two-parent and single-parent families
sns.boxplot(data=df, x='Pstatus', y='Dalc', palette='pastel')
plt.title("Alcohol Consumption in Two-Parent vs Single-Parent Homes")
plt.xlabel("Parental Status (T = Together, A = Apart)")
plt.show()
```



Students from single parent home generally have a slightly higher median alcohol consumption than those from 2 parent homes.There are more outliers with high alcohol use in single parent home suggesting a possible impact of family structure on behaviour.

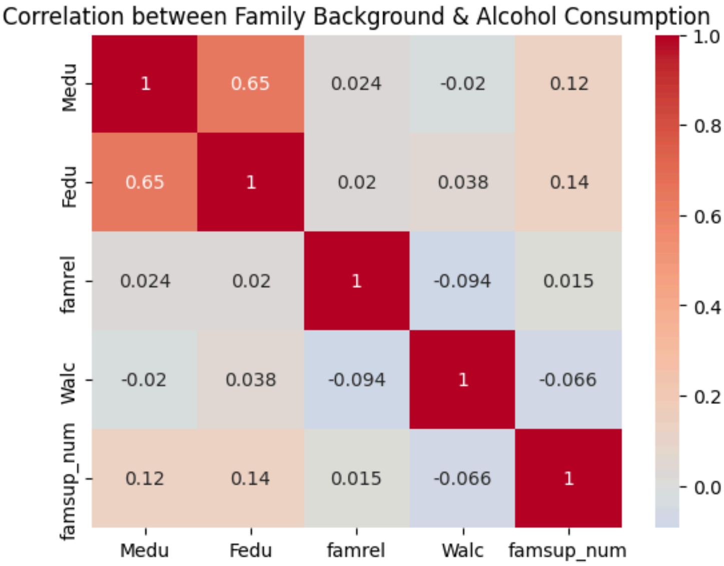
```
# Q7: What kind of relationship does the child have with family members?
sns.barplot(data=df, x='famrel', y='Dalc', ci=None, palette='Spectral')
plt.title("Family Relationship Quality vs Alcohol Consumption")
plt.xlabel("Family Relationship Quality (1 = very bad, 5 = excellent)")
plt.show()
```



Student with bad family relations tend to have a higher daily alcohol consumption level.

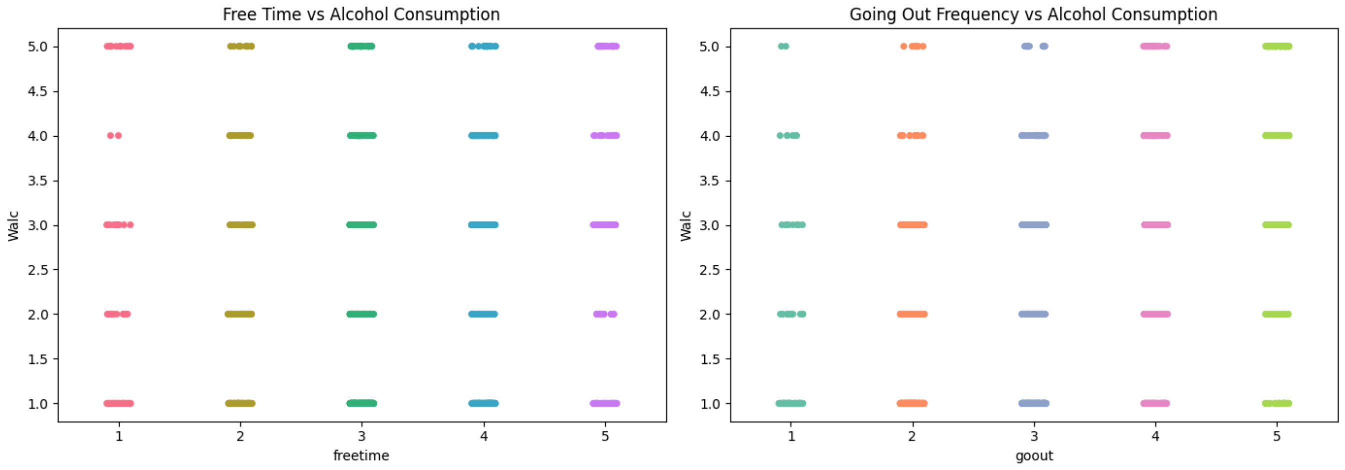
```
# Q8: Family background’s influence on alcoholism (correlation heatmap)
corr_data = df[['Medu', 'Fedu', 'famrel', 'Walc']].copy()
corr_data['famsup_num'] = df['famsup'].map({'yes': 1, 'no': 0})  # Convert to numeric

corr = corr_data.corr()
sns.heatmap(corr, annot=True, cmap='coolwarm', center=0)
plt.title("Correlation between Family Background & Alcohol Consumption")
plt.show()
```



The heat map shows the correlation between family background factors and alcohol consumption. A strong positive correlation exists between mothers and fathers education, while alcohol consumption on weekends has very weak or negative correlation with family background variables suggesting minimal influence.

```
# Q9: Does free time or going out relate to alcohol?
fig, axs = plt.subplots(1, 2, figsize=(14, 5))
sns.stripplot(data=df, x='freetime', y='Walc', ax=axs[0], palette='husl', jitter=True)
axs[0].set_title("Free Time vs Alcohol Consumption")
sns.stripplot(data=df, x='goout', y='Walc', ax=axs[1], palette='Set2', jitter=True)
axs[1].set_title("Going Out Frequency vs Alcohol Consumption")
plt.tight_layout()
plt.show()
```


Free Time vs Alcohol Consumption

Going Out Frequency vs Alcohol Consumption

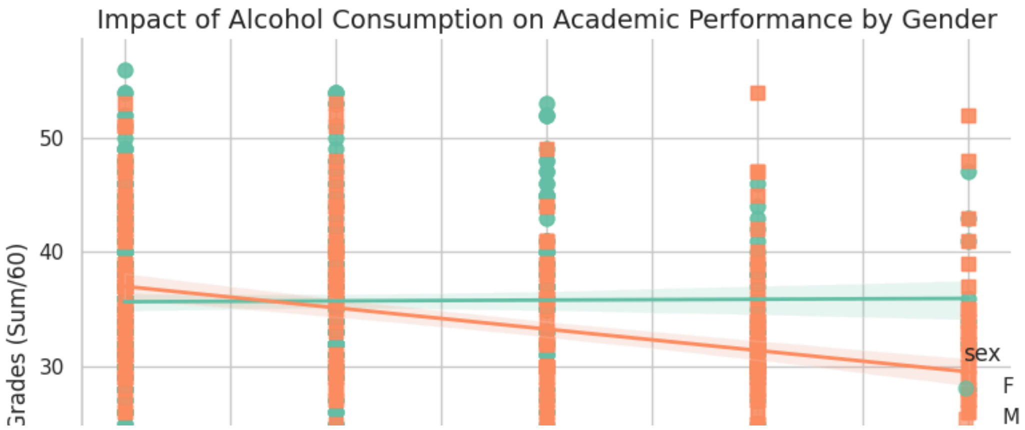
These plots explore how free time and going out frequency relate to alcohol consumption. The left plot shows no strong pattern between free time and alcohol use. However, the right plot indiciates a clear trend: Students who go out more frequenty tend to have high alcohol consumption levels.

```
# Q10: Alcohol consumption vs academic performance

sns.set_theme(style="whitegrid")
palette = sns.color_palette("coolwarm", as_cmap=True)

# Lmplot with improved aesthetics and gender hue
sns.lmplot(
    data=df,
    x='Walc',
    y='Sum/60',
    hue='sex',                      # highlight by gender
    palette='Set2',                 # bright and readable colors
    height=6,
    aspect=1.2,
    markers=["o", "s"],
    scatter_kws={'alpha': 0.7, 's': 60},  # transparency & point size
    line_kws={"linewidth": 2}
)

plt.title("Impact of Alcohol Consumption on Academic Performance by Gender", fontsize=14)
plt.xlabel("Weekend Alcohol Consumption (Walc)", fontsize=12)
plt.ylabel("Total Grades (Sum/60)", fontsize=12)
plt.tight_layout()
plt.show()
```


Impact of Alcohol Consumption on Academic Performance by Gender
50
40
30
Grades (Sum/60)
sex
F
M