Student Perfomance Data Analysis

Import packages

In [2]:
```
import pandas as pd
import seaborn as sns
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
```

In [3]:
```
data = pd.read_csv("StudentsPerformance.csv")
```

Overview of the data

In [4]:
```
data.head()
```

Out[4]:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

In [5]:
```
data.shape
```

Out[5]:
```
(1000, 8)
```

In [71]:
```
race = data["race/ethnicity"]
math = data["math score"]
reading = data["reading score"]
writing = data["writing score"]
gender = data["gender"]
gscore = (math + reading + writing)/3
fdata = data[data["gender"] == "female"]
mdata = data[data["gender"] == "male"]
parent = data["parental level of education"]
lunch = data["lunch"]
```

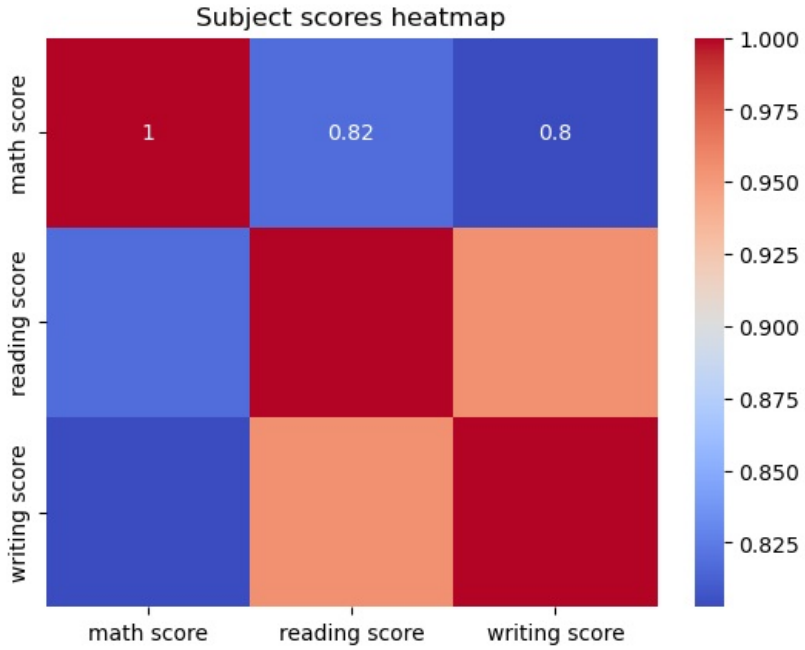
Correlations between different subject scores

In [72]:
```
a = ["math score", "reading score", "writing score"]
```

In [73]:
```
b = data[a]
```

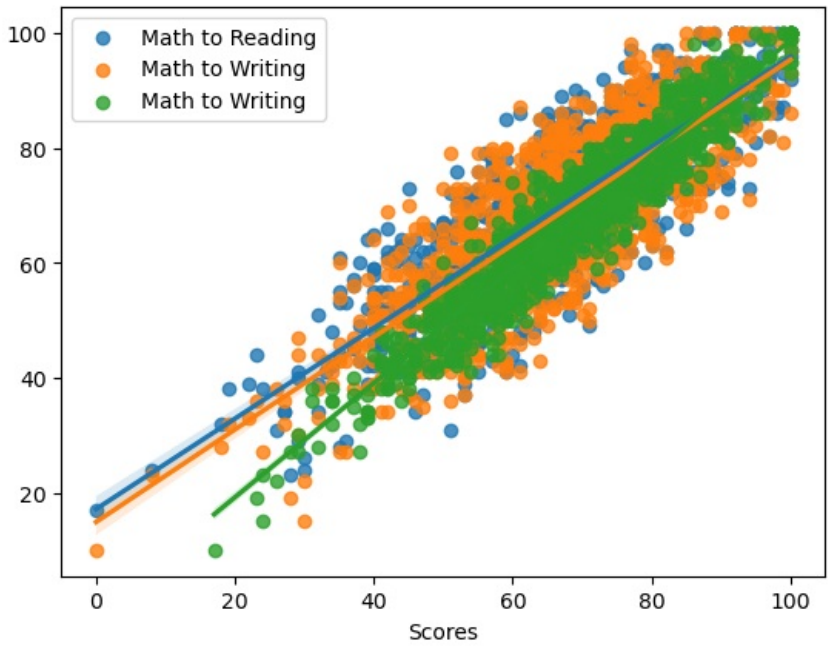
In [74]:
```
sns.heatmap(b.corr(), annot= True, cmap='coolwarm')
plt.title("Subject scores heatmap")
```

Out[74]:
```
Text(0.5, 1.0, 'Subject scores heatmap')
```



It seems that there is a stronger correlation between reading and writing compared to the correlation between math score

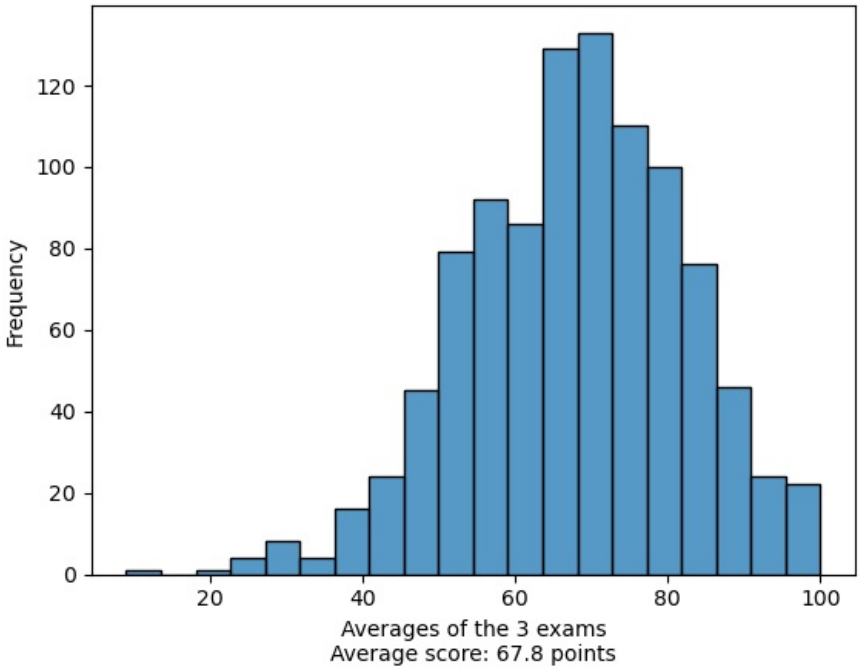
In [75]:
```
a1 = math.corr(reading)
b1 = math.corr(writing)
c1 = reading.corr(writing)
sns.regplot(x=math, y= reading, label="Math to Reading")
sns.regplot(x=math, y=writing, label="Math to Writing")
sns.regplot(x=reading, y=writing, label="Math to Writing")
plt.xlabel("Scores")
plt.ylabel("")
plt.legend()
plt.show()
print(f"Math to reading corr: {a1:0f}\nMath to writing: {b1:0f}\nReading to writing: {c1:0f}")
```




```
Math to reading corr: 0.817580
Math to writing: 0.802642
Reading to writing: 0.954598
```

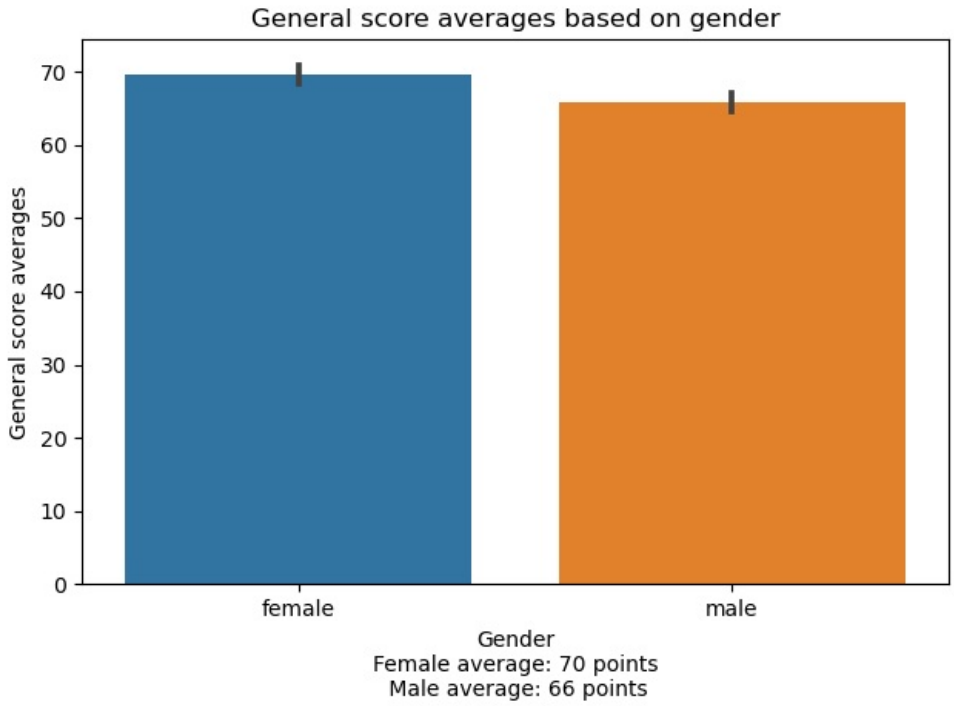
In [76]:
```
sns.histplot(gscore, bins=20)
gsm = gscore.mean()
plt.xlabel(f"Averages of the 3 exams\nAverage score: {gsm:.1f} points")
plt.ylabel("Frequency")

plt.show()
```

```
C:\Users\Lenovo\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

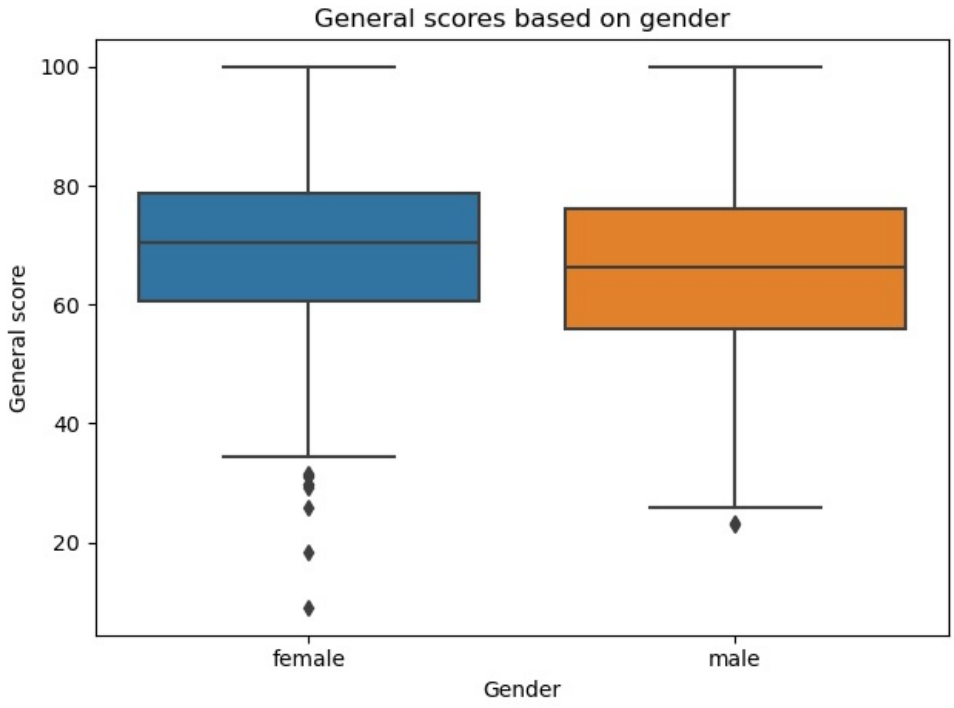


Gender stats

In [77]:
```
fg = ((fdata["math score"] + fdata["reading score"] + fdata["writing score"])/3).mean()
mg = ((mdata["math score"] + mdata["reading score"] + mdata["writing score"])/3).mean()
sns.barplot(x=gender, y=gscore)
plt.xlabel(f"Gender\nFemale average: {fg:.0f} points\n Male average: {mg:.0f} points")
plt.ylabel("General score averages")
plt.title("General score averages based on gender")
plt.tight_layout()
plt.show()
```

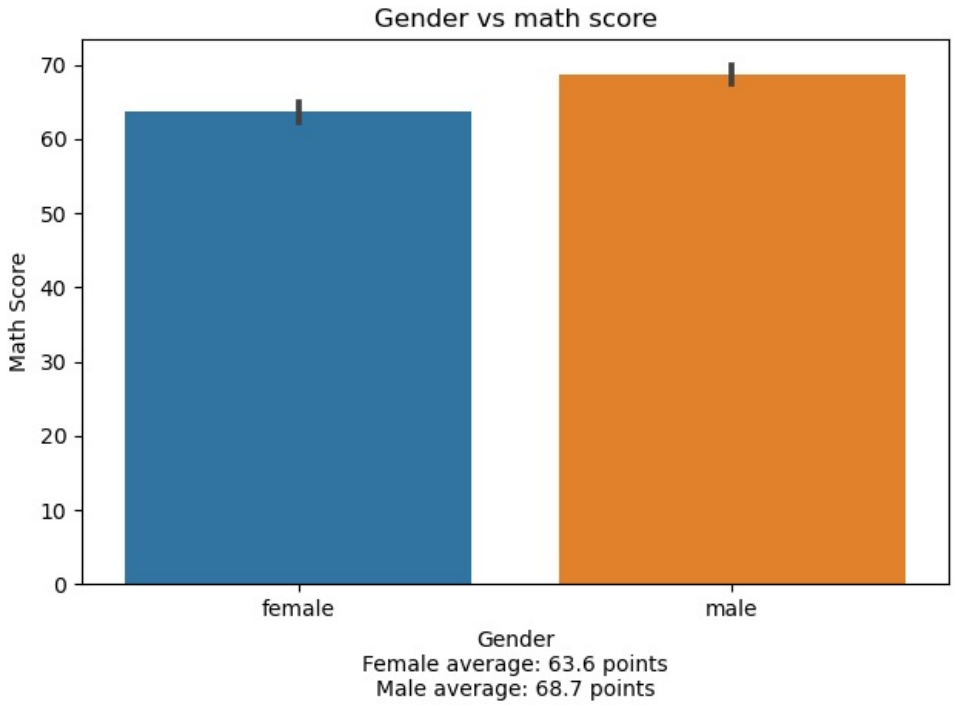


In [78]:
```
sns.boxplot(x=gender, y=gscore)
plt.xlabel("Gender")
plt.ylabel("General score")
plt.title("General scores based on gender")
plt.tight_layout()
plt.show()
```

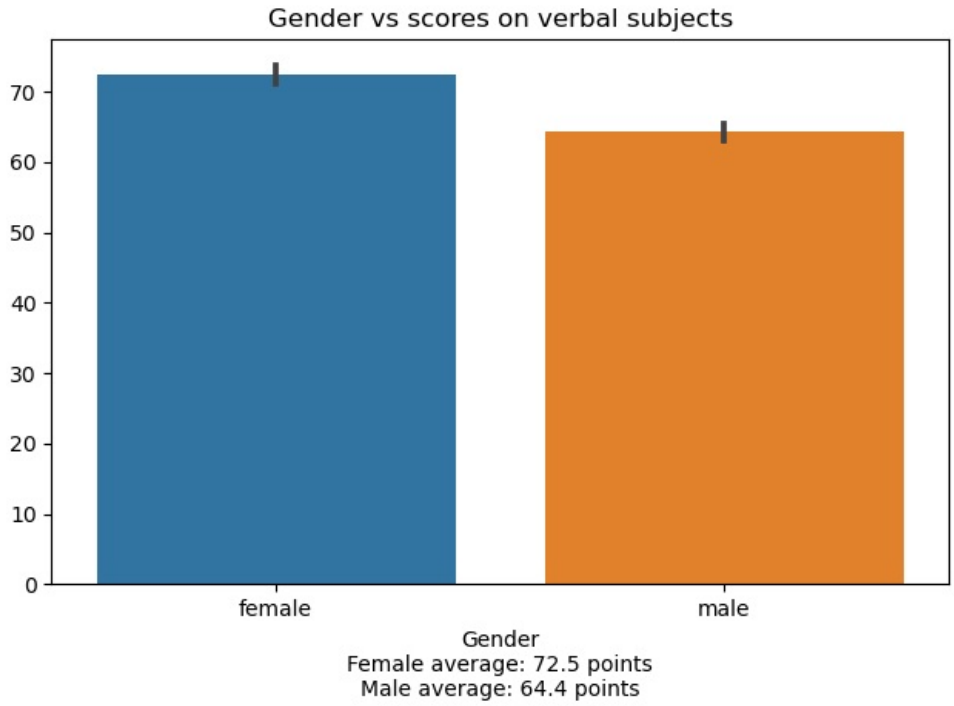


Gender specific differences in subjects

In [79]:
```
sns.barplot(x=gender, y=math)
plt.xlabel(f"Gender\nFemale average: {fdata['math score'].mean():.1f} points\nMale average: {mdata['math score'
plt.title("Gender vs math score")
plt.ylabel("Math Score")
plt.tight_layout()
plt.show()
```



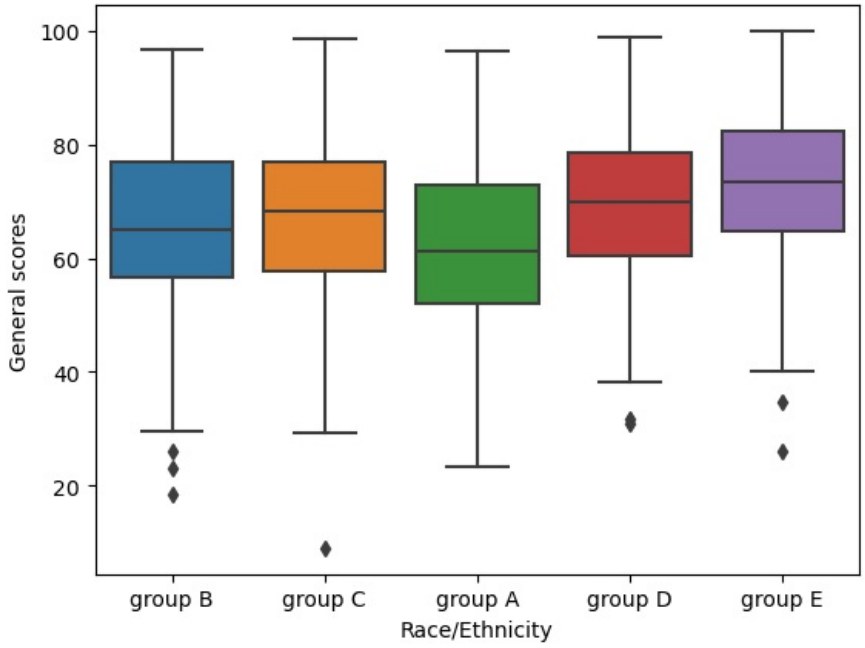
In [80]:
```
rw = ((data["reading score"] + data["writing score"])/2)
frw = ((fdata["reading score"] + fdata["writing score"])/2).mean()
mrw = ((mdata["reading score"] + mdata["writing score"])/2).mean()
sns.barplot(x=gender, y=rw)
plt.title("Gender vs scores on verbal subjects")
plt.xlabel(f"Gender\nFemale average: {frw:.1f} points\nMale average: {mrw:.1f} points")
plt.tight_layout()
plt.show()
```



We can counclude from the dataset that female students were slightly better in verbal subjects, whereas male students were better in maths

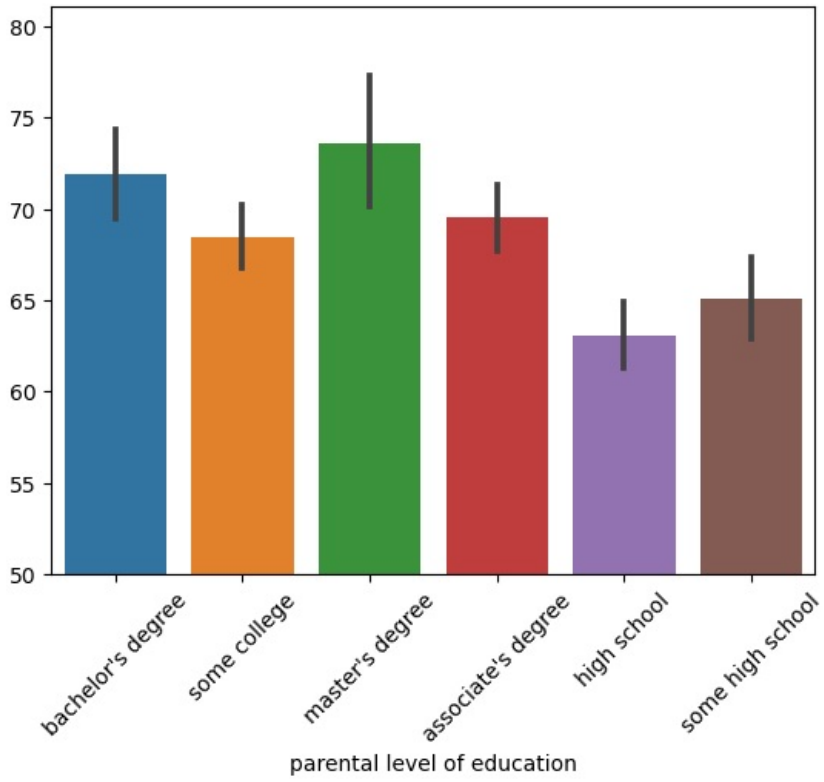
Score differences between different races/ethnicities

In [81]:
```
sns.boxplot(x=race, y=gscore)
plt.ylabel("General scores")
plt.xlabel("Race/Ethnicity")
plt.show()
```



parental level of education

In [82]:
```
sns.barplot(x=parent, y=gscore)
plt.xticks(rotation=45)
ax = plt.gca()
ax.set_ylim(50, None)
plt.show()
```



As expected, parents academic success is correlated with students scores