Part_I_exploration

January 26, 2024

1 Part I - Student Questionaire and Standardized Exam Results

1.1 by Amanda Doty

1.2 Introduction

This dataset consists of demographic data and exam scores for students from around the world. Data includes information about immediate family members, socio-economic status, school attendance and performance, and exam scores.

1.3 Preliminary Wrangling

```
[1]: # import all packages and set plots to be embedded inline
     import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
```

```
[2]: sub_df = pd.read_csv('/Users/amand/WGU/Communicating Data Findings_A Doty_01.20.
      ↪24/student_qqq_sub.csv')

     sub_df.head(4)
```

```
[2]:            mother_edu          father_edu      qty_books    repeated  \
     0      ISCED level 3A      ISCED level 3A   26-100 books   No, never
     1      ISCED level 3A  ISCED level 3B, 3C   26-100 books   No, never
     2      ISCED level 3A      ISCED level 3A    11-25 books   No, never
     3  ISCED level 3B, 3C      ISCED level 3A  201-500 books   No, never

                    time_reading           outsider    belong  tmins  wealth  \
     0     30 minutes or less a day  Strongly disagree     Agree   1650       0
     1  I do not read for enjoyment           Disagree  Disagree   1620       0
     2     30 minutes or less a day           Disagree  Disagree   1350      -1
     3           1 to 2 hours a day  Strongly disagree     Agree   1650       0

        pv1math  pv1read  avg_score  tot_score
     0  697.233  650.848   674.0405   1348.081
     1  501.317  449.588   475.4525    950.905
     2  520.497  404.564   462.5305    925.061
```

```
3  563.845  522.237   543.0410   1086.082
```

```
[3]: sub_df.shape
```

```
[3]: (1211, 13)
```

```
[4]: sub_df.mother_edu.unique()
```

```
[4]: array(['ISCED level 3A', 'ISCED level 3B, 3C', 'ISCED level 2',
            'ISCED level 1', 'She did not complete  ISCED level 1'],
           dtype=object)
```

```
[5]: sub_df.father_edu.unique()
```

```
[5]: array(['ISCED level 3A', 'ISCED level 3B, 3C', 'ISCED level 2',
            'ISCED level 1', 'He did not complete  ISCED level 1'],
           dtype=object)
```

```
[6]: sub_df.qty_books.unique()
```

```
[6]: array(['26-100 books', '11-25 books', '201-500 books', '101-200 books',
            '0-10 books', 'More than 500 books'], dtype=object)
```

```
[7]: sub_df.outsider.unique()
```

```
[7]: array(['Strongly disagree', 'Disagree', 'Agree', 'Strongly agree'],
           dtype=object)
```

```
[8]: sub_df.tmins.unique()
```

```
[8]: array([1650, 1620, 1350, 1980, 2100, 1500, 1710, 1040, 2200, 1400, 1215,
            1800, 1540, 2500, 1530, 1395, 1575, 2400, 1200,  900, 1440, 1595,
            1485, 2115,  800, 1600, 1680, 2040, 2700,  540, 1750, 1625, 2145,
            1700, 1950, 1755, 1920, 2835,  550, 2720, 1305, 1640, 1960, 1925,
            2025, 1000, 2310, 1935, 1665, 2000, 2160,  720, 1280, 1160, 1260,
            2745, 2940, 1740, 1870, 1470, 2880, 1380, 2250, 2365, 3000, 1480,
            1050, 1300, 2450, 1560, 1850, 1100, 1760, 2430, 1520, 2275, 2900,
            2320, 1845, 1155, 2925, 1125, 1890, 1320, 1880, 2520, 2440,  855,
            1375, 1430, 2800,  425, 1690,  520, 2475, 1240, 2050, 2240, 2580,
            2170, 1900, 1720, 2750, 2385, 2640, 1250,  810,  675, 1875, 2460,
            2325,  400, 2960, 2790, 2565, 2035, 1815, 2150,  600, 1705, 2340,
            1035,  450, 1550,  495,  990, 1170, 1860, 2360, 2550, 2080, 1360,
            2070, 2205, 1020, 1435,  960, 1080, 2655,  480,  340,  360,  945,
             840,  225, 2280, 1615, 2650, 2760, 1210, 1645, 1770,  660, 2420,
             765, 2610, 2480, 2970, 1365, 1340, 2295, 2600,  825,  210,  650])
```

```
[9]: sub_df.repeated.unique()
```

[9]:
```
array(['No, never', 'Yes, once', 'Yes, twice or more'], dtype=object)
```

[10]:
```
sub_df['tmins']=sub_df['tmins'].astype('int')
sub_df['wealth']=sub_df['wealth'].astype('int')

sub_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1211 entries, 0 to 1210
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mother_edu    1211 non-null   object
 1   father_edu    1211 non-null   object
 2   qty_books     1211 non-null   object
 3   repeated      1211 non-null   object
 4   time_reading  1211 non-null   object
 5   outsider      1211 non-null   object
 6   belong        1211 non-null   object
 7   tmins         1211 non-null   int64
 8   wealth        1211 non-null   int64
 9   pv1math       1211 non-null   float64
 10  pv1read       1211 non-null   float64
 11  avg_score     1211 non-null   float64
 12  tot_score     1211 non-null   float64
dtypes: float64(4), int64(2), object(7)
memory usage: 123.1+ KB
```

[11]:
```
#creating summative columns to further assist in statistical visualizations.

sub_df['avg_score']=sub_df[['pv1read','pv1math']].mean(axis=1)
sub_df['tot_score']=sub_df[['pv1read','pv1math']].sum(axis=1)

sub_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1211 entries, 0 to 1210
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mother_edu    1211 non-null   object
 1   father_edu    1211 non-null   object
 2   qty_books     1211 non-null   object
 3   repeated      1211 non-null   object
 4   time_reading  1211 non-null   object
 5   outsider      1211 non-null   object
 6   belong        1211 non-null   object
 7   tmins         1211 non-null   int64
 8   wealth        1211 non-null   int64
 9   pv1math       1211 non-null   float64
 10  pv1read       1211 non-null   float64
 11  avg_score     1211 non-null   float64
 12  tot_score     1211 non-null   float64
dtypes: float64(4), int64(2), object(7)
memory usage: 123.1+ KB
```

[12]:
```
sub_df.describe()
```

[12]:
```
             tmins       wealth      pv1math      pv1read    avg_score  \
count  1211.000000  1211.000000  1211.000000  1211.000000  1211.000000
mean   1679.905037    -0.212221   492.234095   484.017983   488.126039
std     437.727380     0.862659    97.076571    98.277763    92.924435
min     210.000000    -4.000000   182.153000   207.258000   202.009500
25%    1440.000000    -1.000000   424.106000   413.819500   420.249250
50%    1620.000000     0.000000   494.838000   485.146000   491.400000
75%    1860.000000     0.000000   559.403000   555.586000   553.164750
max    3000.000000     4.000000   756.254000   813.600000   748.745000

         tot_score
count  1211.000000
mean    976.252078
std     185.848871
min     404.019000
25%     840.498500
50%     982.800000
75%    1106.329500
max    1497.490000
```

[13]:
```
# categorizing categorical data

books=['0-10 books','11-25 books','26-100 books','101-200 books','201-500␣
 ↪books','More than 500 books']
book_classes=pd.CategoricalDtype(ordered=True, categories=books)
sub_df['qty_books'] = sub_df['qty_books'].astype(book_classes)

mo_edu=['She did not complete  ISCED level 1','ISCED level 1','ISCED level␣
 ↪2','ISCED level 3A', 'ISCED level 3B, 3C']
mo_classes=pd.CategoricalDtype(ordered=True, categories=mo_edu)
sub_df['mother_edu'] = sub_df['mother_edu'].astype(mo_classes)

min_read=['I do not read for enjoyment','30 minutes or less a day','More than␣
 ↪30 minutes to less than 60 minutes a day','1 to 2 hours a day','More than 2␣
 ↪hours a day']
read_classes=pd.CategoricalDtype(ordered=True, categories=min_read)
sub_df['time_reading'] = sub_df['time_reading'].astype(read_classes)
```

```
outsider=['Strongly disagree','Disagree','Agree','Strongly agree']
out_classes=pd.CategoricalDtype(ordered=True, categories=outsider)
sub_df['outsider'] = sub_df['outsider'].astype(out_classes)

belong=['Strongly disagree','Disagree','Agree','Strongly agree']
belong_classes=pd.CategoricalDtype(ordered=True, categories=belong)
sub_df['belong'] = sub_df['belong'].astype(belong_classes)
```

[14]:

```
sub_df.to_csv('/Users/amand/WGU/Communicating Data Findings_A Doty_01.20.24/
↪student_qqq_sub.csv', index=False)
```

1.3.1 What is the structure of your dataset?

The base dataset is very large, with over 600,000 entries and 1100 columns. To make calculations and visualizations quicker and require less operating power, I took a subset of the dataset consisting of 1000 random entries and narrowed the columns down to those I am interested in studying (see below for details). I created two summary columns, avg_score (the average of the math and reading scores) and tot_score (the sum of the math and reading scores) to help with statistical analysis of the overall performance on the exams.

1.3.2 What is/are the main feature(s) of interest in your dataset?

I am interested in the role home life, socioeconomic status, and self-esteem have on student performance. I selected columns that deal with parent education, the number of books in the home, amount of time reading, emotions and belonging, and performance.

1.4 Univariate Exploration

[15]:

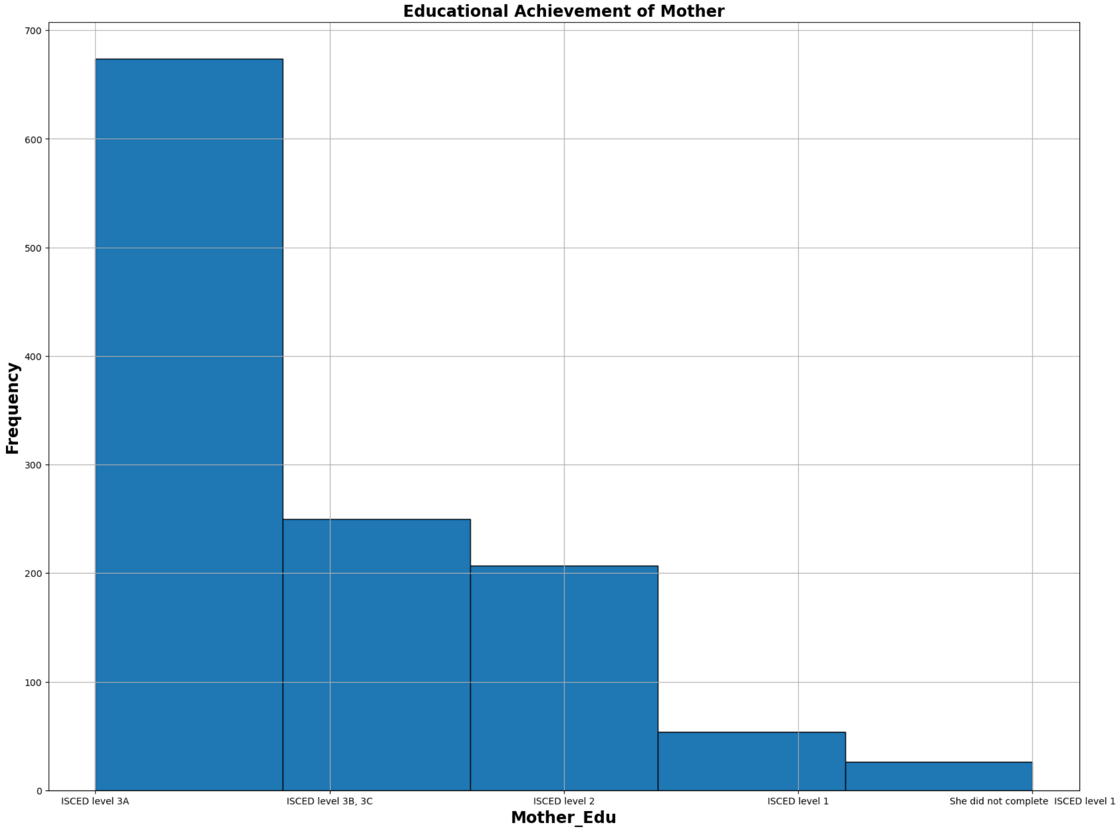
```
def plot_hist (df, X, bins=30, plt_title='none'):

    # this function plots histograms using the given parameters
    # it prevents repetitive code and saves time

    #inputs:
    ### df: dataframe to plot
    ### x: column name from which to gather data
    ### bins: number of bins in the histogram
    ### plt_title: the title of the figure

    #output: a plotted histogram with the given parameters

    df[X].hist(bins=bins,grid=True, edgecolor='black', color='tab:blue',␣
↪figsize = [20,15])
    plt.title(plt_title, weight='bold',size='xx-large')
    plt.xlabel(X.title(), weight = 'bold', size='xx-large')
    plt.ylabel('Frequency',weight='bold', size='xx-large')
```

1.5 Parental Education Distributions

1.5.1 Mother's Educational Background

Question What is the distribution of the highest level of education for the mothers in the dataset?

[16]:
```
plot_hist(sub_df, 'mother_edu',5,'Educational Achievement of Mother');
```



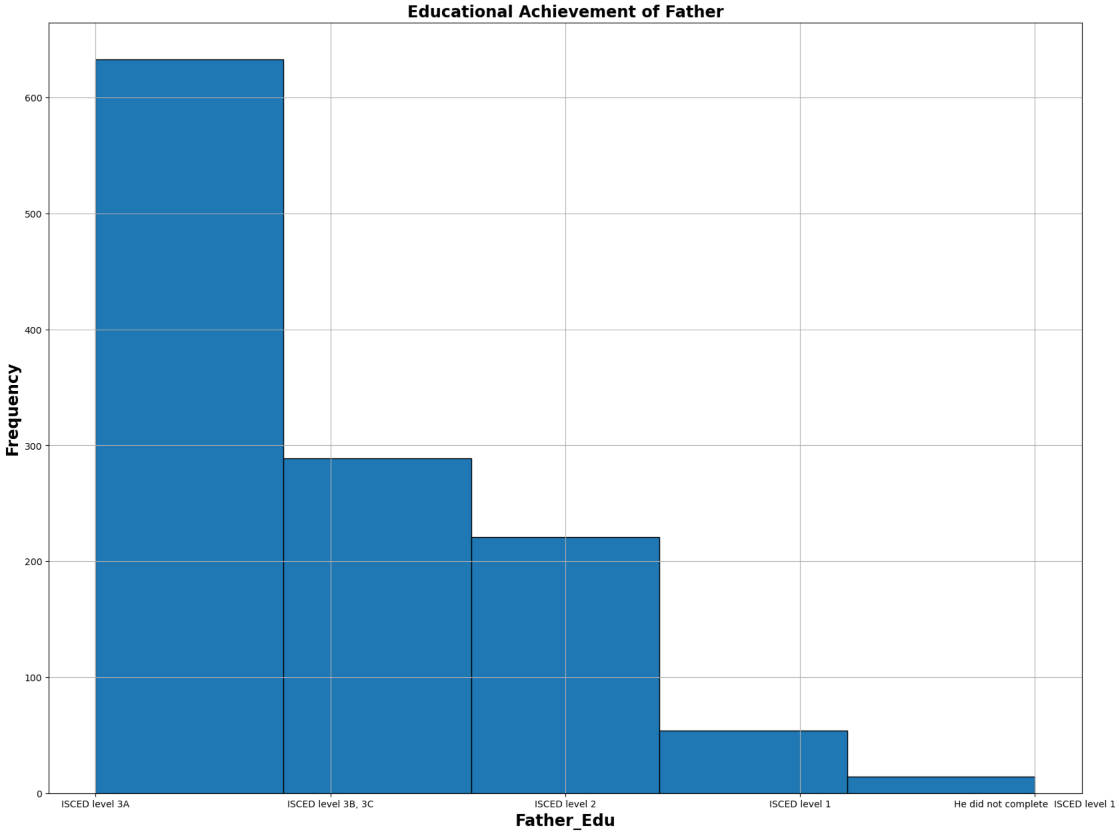
Answer Most mothers, over 50%, have graduated from secondary education (ISCED level 3A is roughly equivalent to a high school senior). Over 70% have a high school diploma or post-secondary education degree.

1.5.2 Father's educational background

Question

What is the distribution of fathers who have secondary or post-secondary education?

[17]:
```
plot_hist(sub_df, 'father_edu',5,'Educational Achievement of Father');
```



Answer

Once again, the majority of fathers, over 50%, have at least a high school diploma, and roughly 75% have a high school diploma or higher.

1.6 Parental education take aways

The vast majority of parents in this sample have at least a high-school level education. I am curious to see if students who have parents who completed high school or beyond score better overall on the exams.

1.7 Socio-Economic Factors

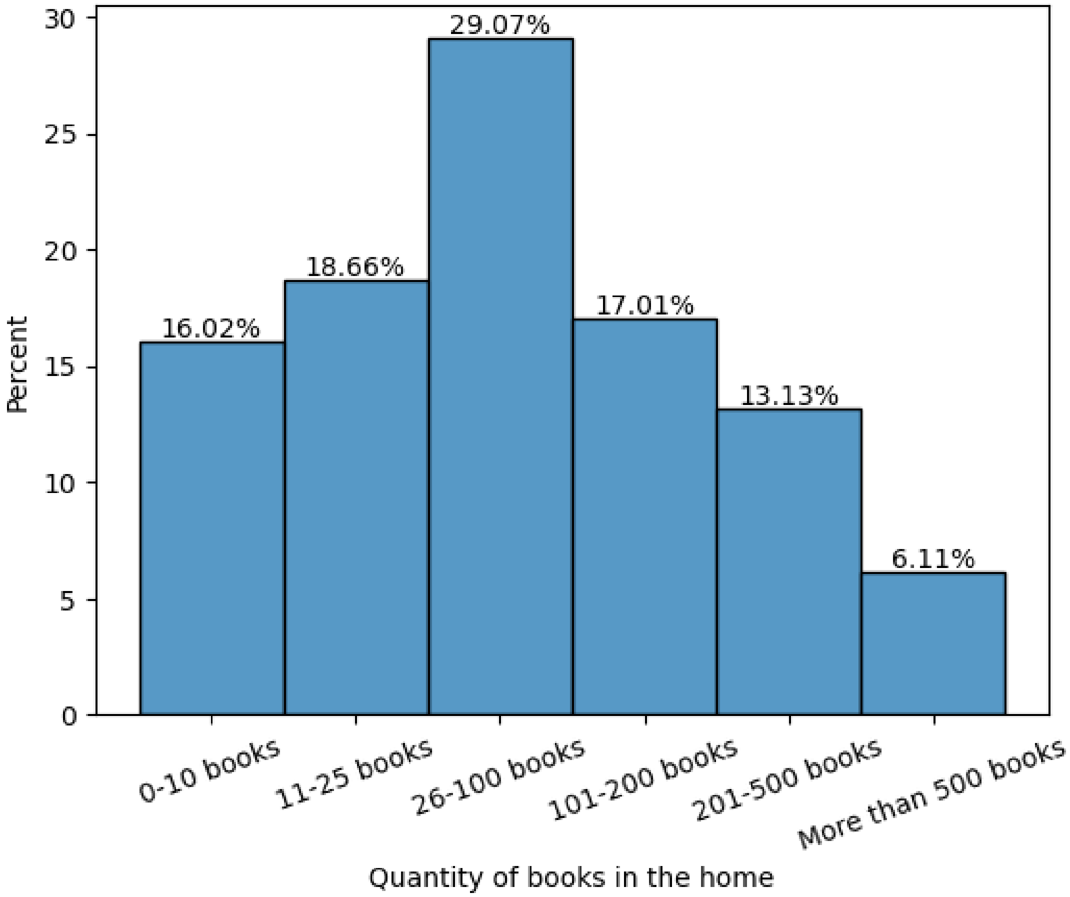
1.7.1 Availability of books in the home

Question

How many students have access to multiple books at home? Books are a luxury item that can point to socio-economic status, but access to books and reading material in the home is also a proven indicator of educational performance.

[18]:
```
ax = sns.histplot(data=sub_df, x='qty_books', stat='percent');
```

```
plt.xticks(rotation=20)
plt.xlabel('Quantity of books in the home')
ax.bar_label(ax.containers[0], fmt='%.2f%%');
```



Answer

Roughly 36% of students estimate that they have over 100 books at home. Most students estimate that they have 100 or fewer, with 16% stating they have ten or fewer. I would hypothesize that students who have fewer books at home read less and score lower on the exams, especially the reading exam.

1.8 School behavior

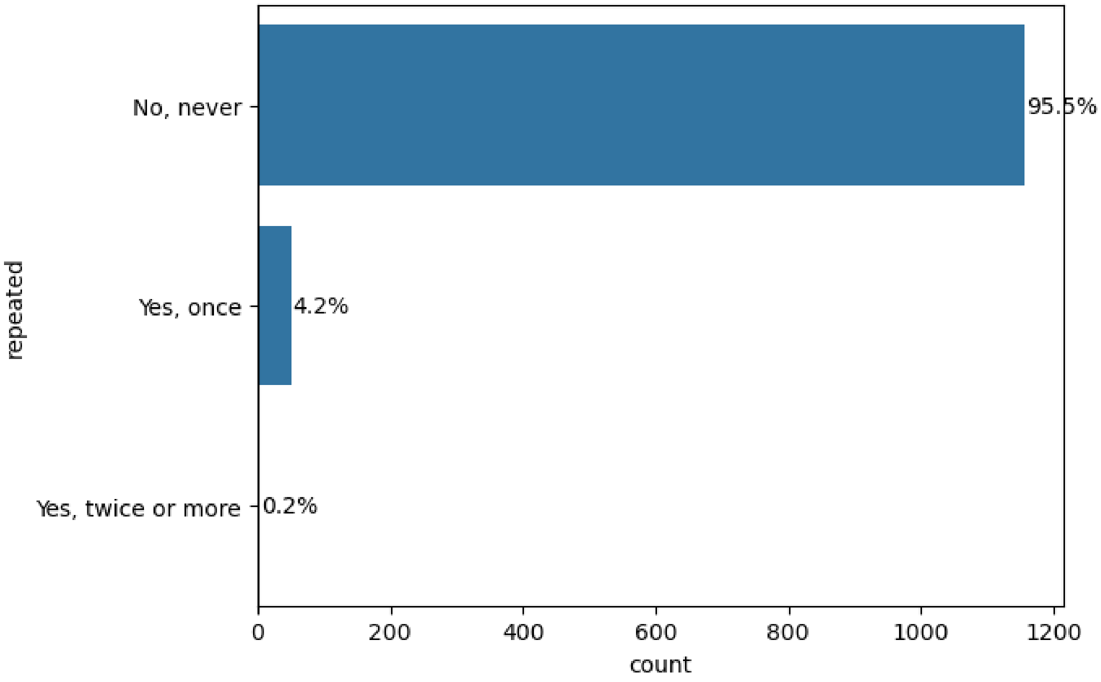
1.8.1 Repeating grades

Question

I would like to know how many students have repeated a grade.

[19]:

```
order = sub_df['repeated'].value_counts().index

repeat_counts = sub_df['repeated'].value_counts()
total_count = repeat_counts.sum()

sns.countplot(data=sub_df, y='repeated', color='tab:blue', order=order)
for i, count in enumerate(repeat_counts):
    # Convert count into a percentage, and then into string
    pct_string = f'{100*count/total_count:.1f}%'
    plt.text(count+1, i, pct_string, va='center')
```



Answer

The vast majority of students have not repeated a grade. Since there is such a large difference between the responses, it might be worth breaking this column down to research further if repeating a grade has an effect on student exam scores.

1.9 Exam Scores

Question

What is the breakdown of exam scores?

[20]:

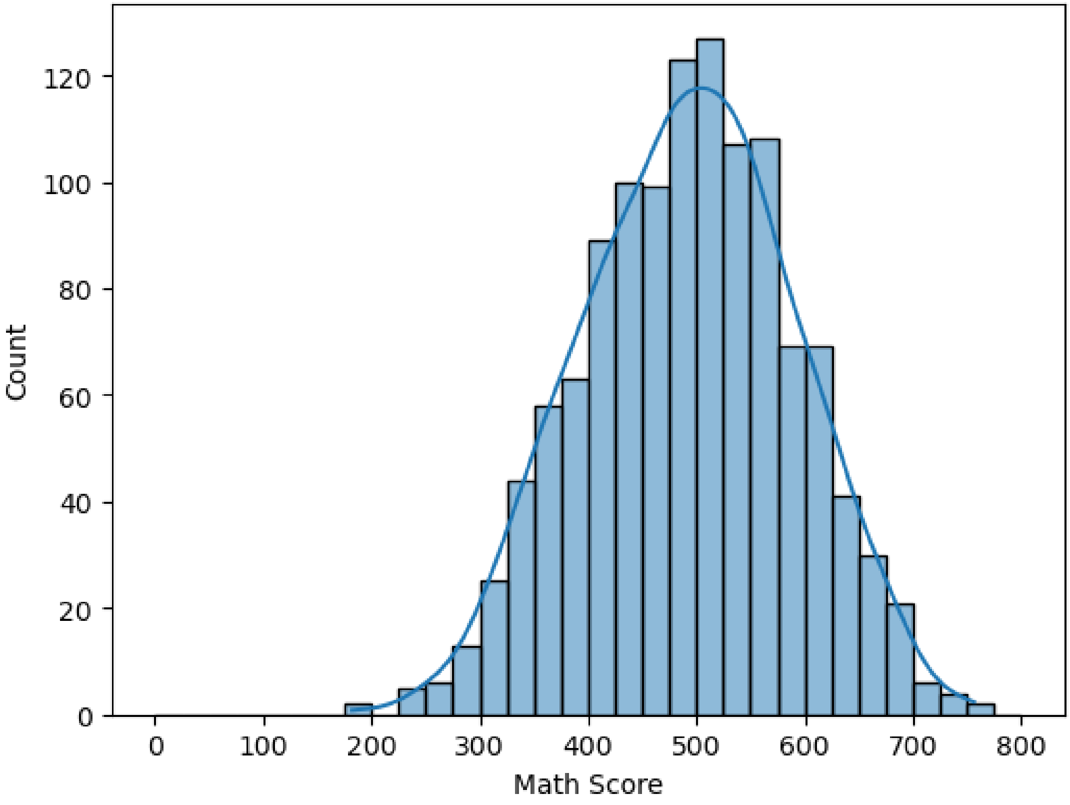
```
sub_df['pv1math'].value_counts()
```

[20]:

```
440.681    2
565.576    2
697.233    1
550.702    1
453.040    1
          ..
585.091    1
499.467    1
517.514    1
686.818    1
594.942    1
Name: pv1math, Length: 1209, dtype: int64
```

[21]:

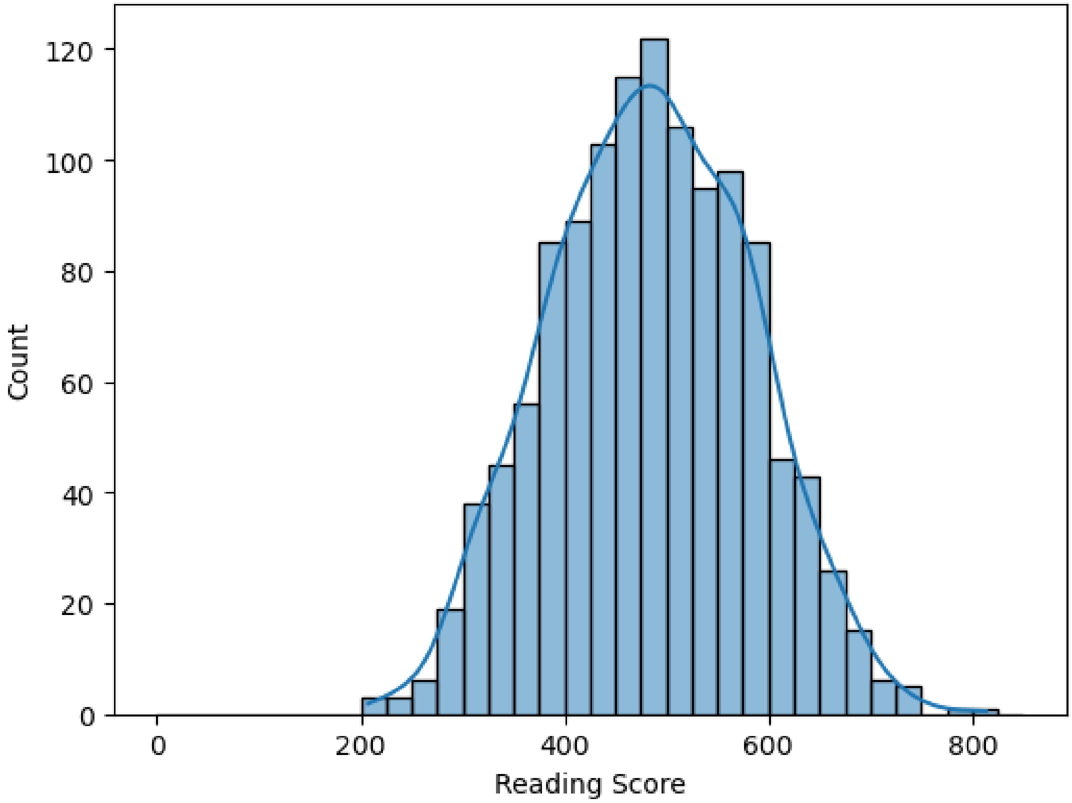
```
bins=np.arange(0, sub_df['pv1math'].max()+50, 25)

sns.histplot(data=sub_df, x='pv1math', bins=bins, stat='count', kde=True)

plt.xlabel('Math Score');
```



[22]:

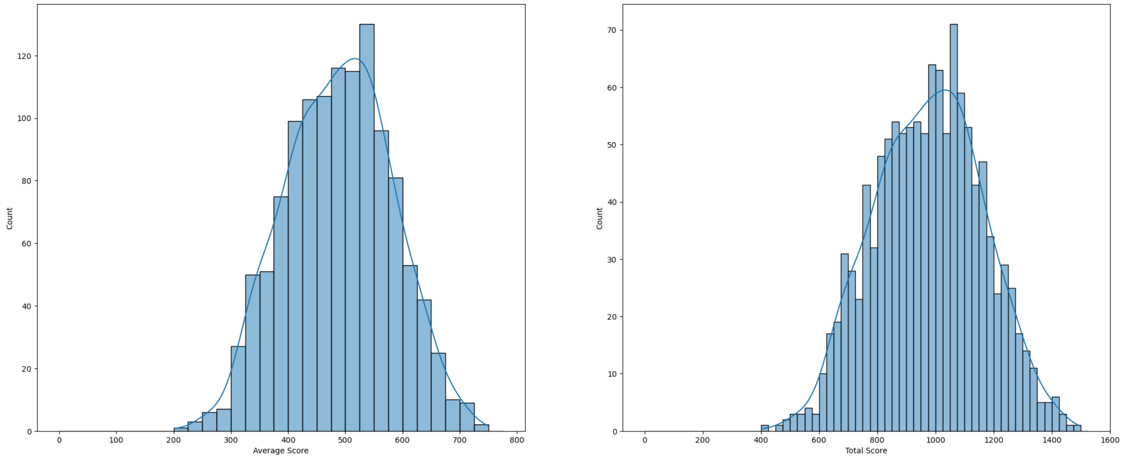
```
bins=np.arange(0, sub_df['pv1read'].max()+50, 25)
```

```
sns.histplot(data=sub_df, x='pv1read', bins=bins, stat='count', kde=True)

plt.xlabel('Reading Score');
```



[23]:

```
bins1=np.arange(0, sub_df['avg_score'].max()+50, 25)
bins2=np.arange(0, sub_df['tot_score'].max()+50, 25)

fig, ax = plt.subplots(1,2,figsize=(25,10))

sns.histplot(data=sub_df, x='avg_score', bins=bins1, kde=True, ax=ax[0])
ax[0].set_xlabel('Average Score');

sns.histplot(data=sub_df, x='tot_score', bins=bins2, kde=True, ax=ax[1])
ax[1].set_xlabel('Total Score');
```



Answer

The scores on the exam mostly follow a normal curve, which is unsurprising. It does appear that math scores are slightly higher on average than reading scores, but it is not likely statistically significant.

1.9.1 Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

I was slightly surprised at the level of education recorded for parents involved in the exams. I did not perform any transformations as there is not a ton of quantitative data, and the quantitative data that is available is relatively normal.

1.9.2 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I created two summary columns of exam scores, one for the average of the two exams and one for the total of the two exams. I ordered the values of several columns (mother_edu, father_edu, qty_books, time_reading) to ensure sequential visualizations.

1.10 Bivariate Exploration

1.11 School behavior

1.11.1 class time and exam score

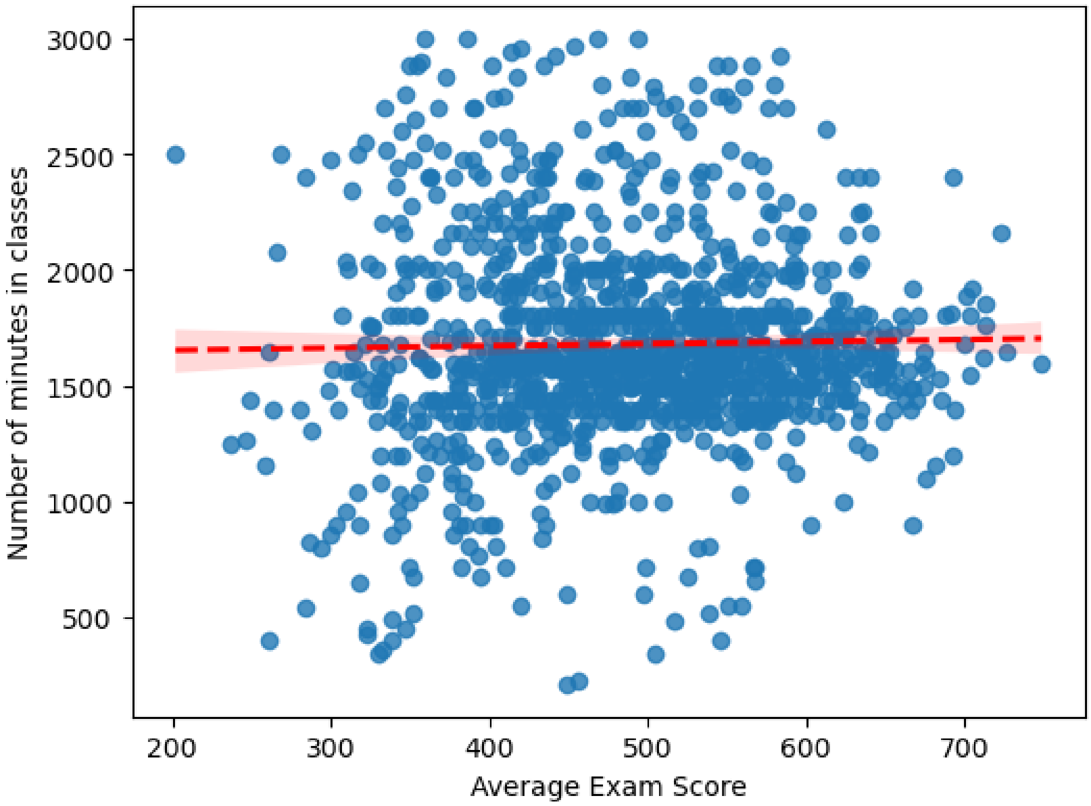
Question

Does more time in class mean a higher score on the exams?

[24]:
```
sns.regplot(data=sub_df, y='tmins',x='avg_score',line_kws=dict(color='r',␣
  ↪linestyle='--'));
```

```
plt.ylabel('Number of minutes in classes')
plt.xlabel('Average Exam Score')
```

```
[24]: Text(0.5, 0, 'Average Exam Score')
```



Answer

There does not appear to be. much of a correlation between time in class and exam scores.

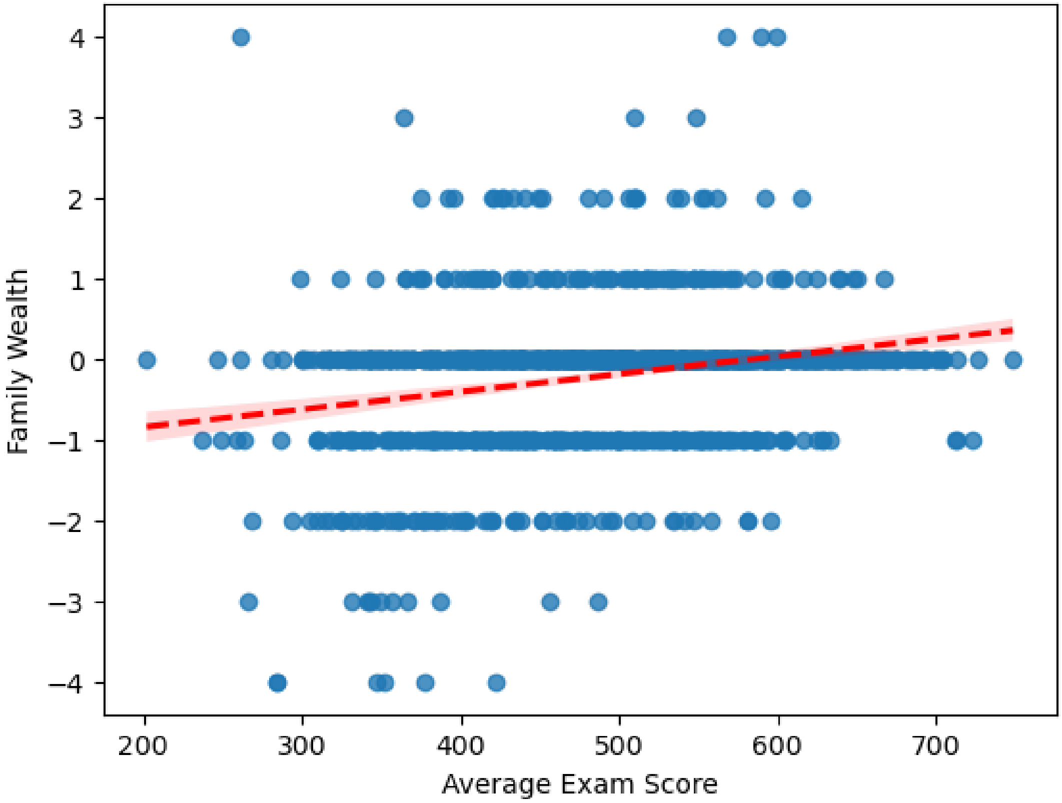
1.12 Socio-econic status

Question

Is there a correlation between family wealth and exam scores?

```
[25]: sns.regplot(data=sub_df, y='wealth',x='avg_score',line_kws=dict(color='r',␣
       ↪linestyle='--'));

      plt.ylabel('Family Wealth')
      plt.xlabel('Average Exam Score')
```

```
[25]: Text(0.5, 0, 'Average Exam Score')
```



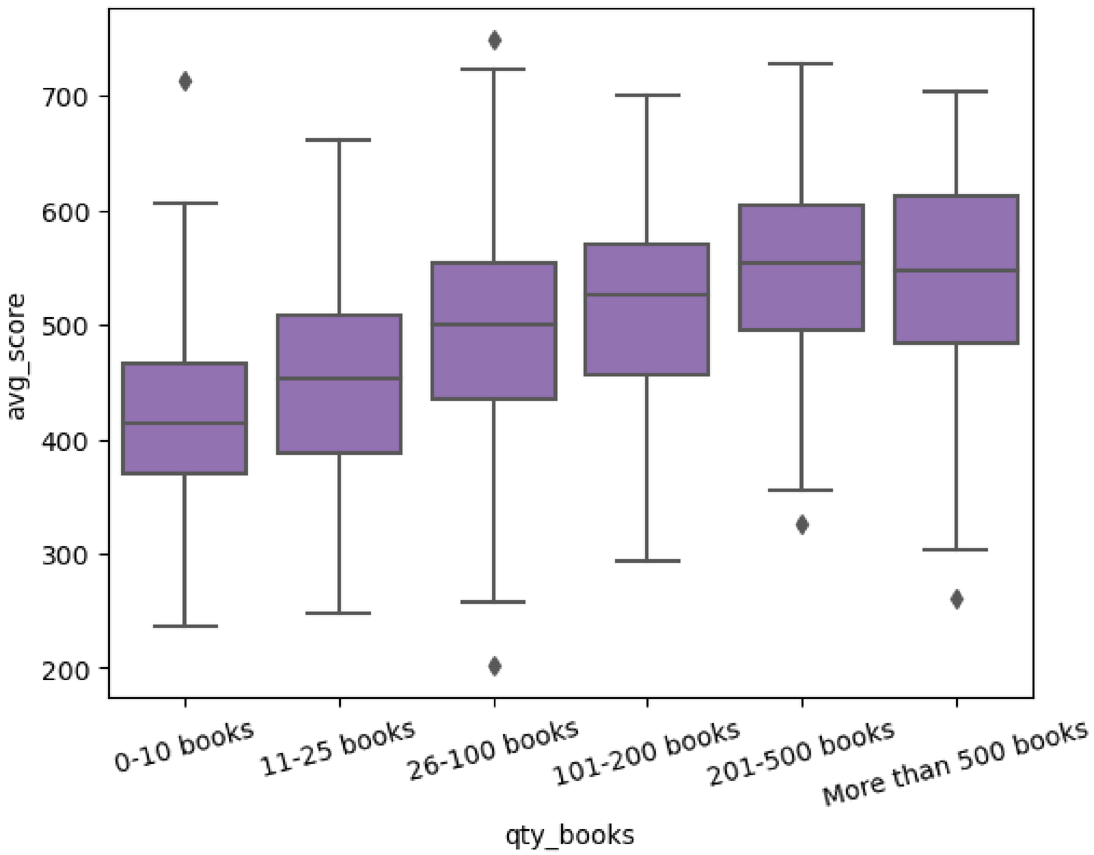
Answer

There is a slight positive correlation between wealth and exam scores. It would be valuable to see what other factors could contribute to the correlation.

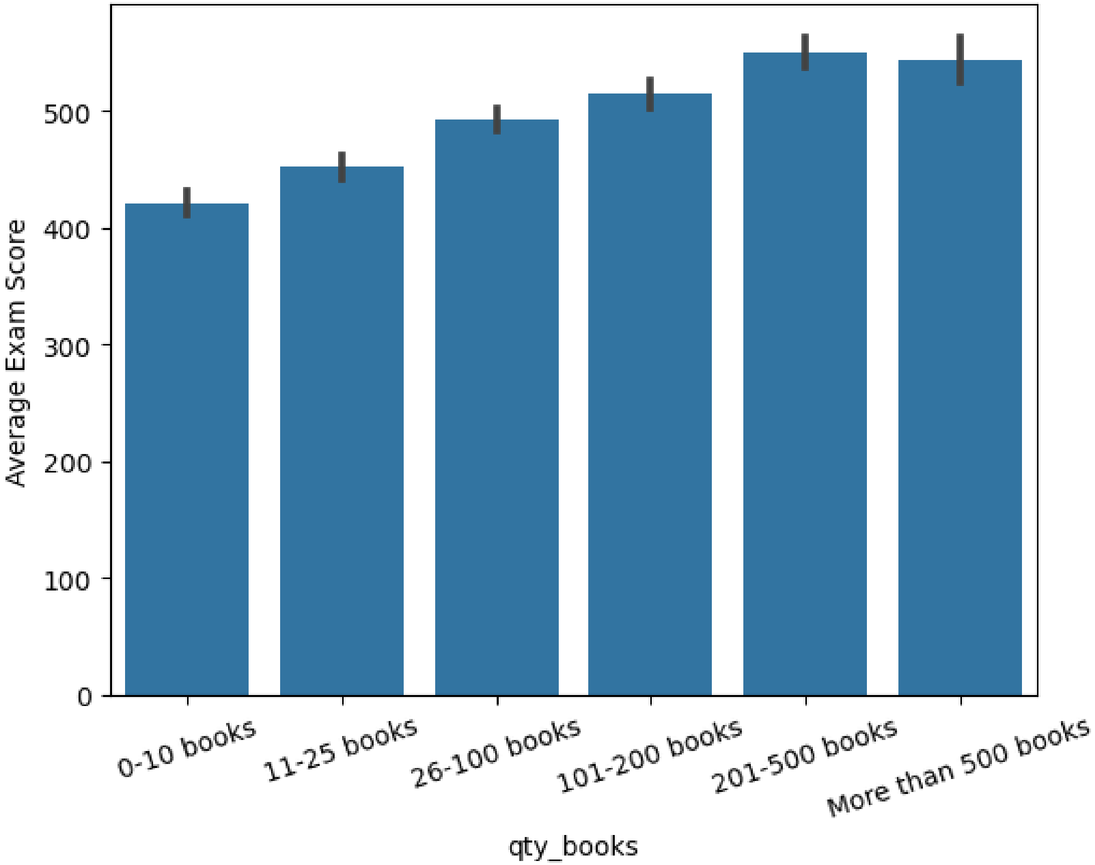
Question

Is there a correlation between the number of books in the home and exam score?

[26]:
```
ax1=sns.boxplot(data=sub_df, x='qty_books',y='avg_score', color='tab:purple')
plt.xticks(rotation=15)
plt.ylim(ax1.get_ylim());
```



[27]:
```
sns.barplot(data=sub_df, x='qty_books',y='avg_score',
            color='tab:blue');
plt.xticks(rotation=18)
plt.ylabel('Average Exam Score');
```



Answer

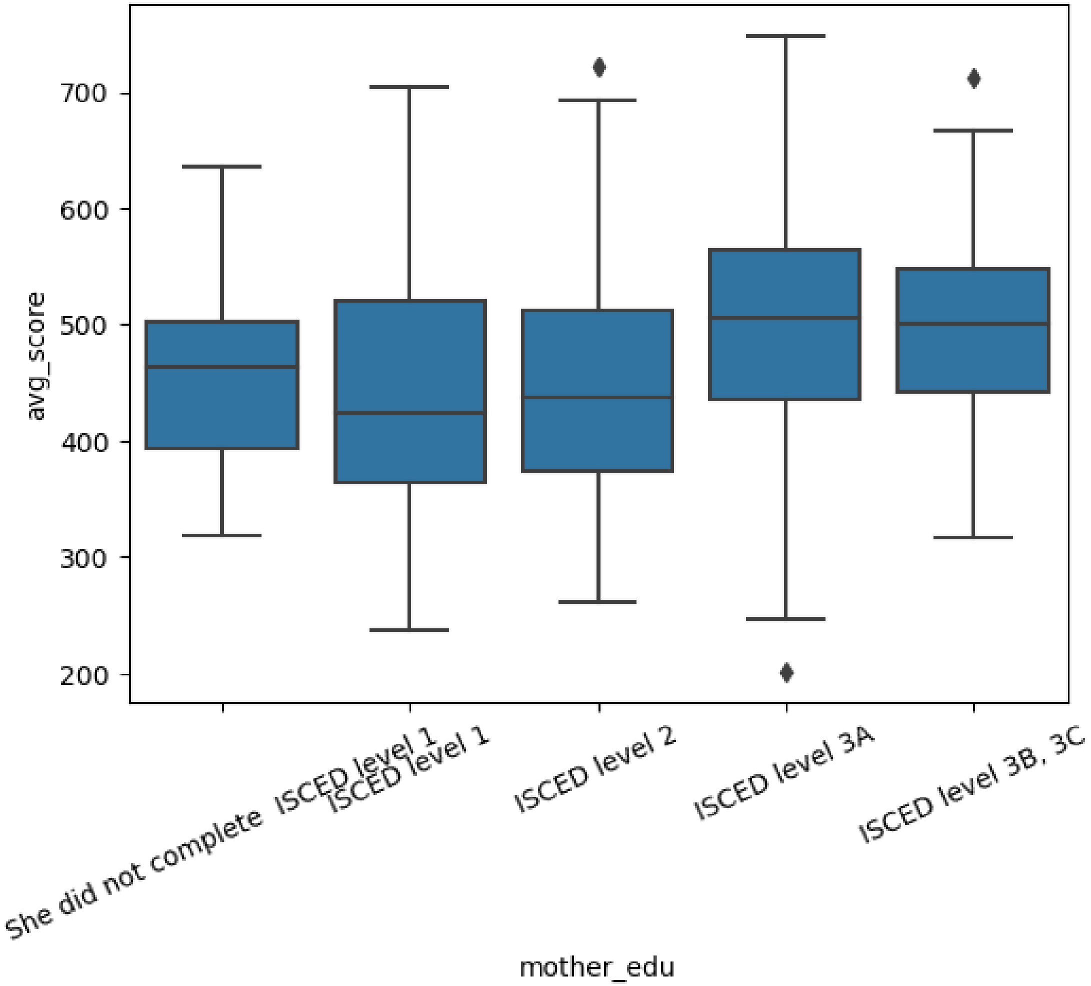
There is a positive correlation between the number of books in the home and the average exam score. It would be valuable to determine if there could be other causes of this correlation, like overall family wealth, or time spent reading.

1.13 Parents' education

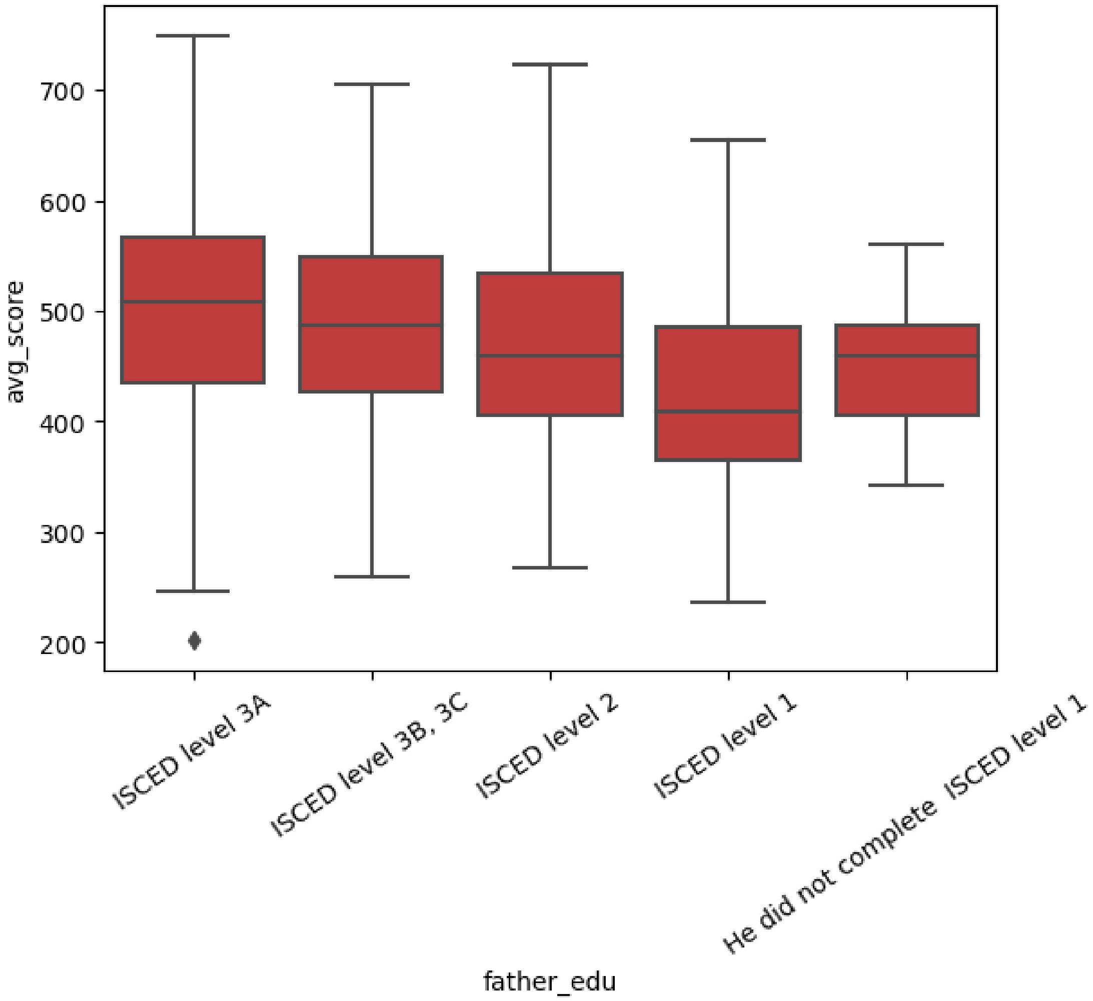
Question

Is there a correlation between parents' education and exam scores?

[28]:
```
ax1=sns.boxplot(data=sub_df, x='mother_edu',y='avg_score', color='tab:blue')
plt.xticks(rotation=25)
plt.ylim(ax1.get_ylim());
```



```
[29]: ax1=sns.boxplot(data=sub_df, x='father_edu',y='avg_score', color='tab:red')
      plt.xticks(rotation=35)
      plt.ylim(ax1.get_ylim());
```



Answer

There does appear to be a positive correlation between the level of education of household parents and exam scores of students. It would be interesting to see if parents with higher education also earn more, which has already shown a positive correlation to exam scores.

1.13.1 Deeper Dives

Question

Does the presence of more books in the home increase reading time?

[30]:

```
# Use group_by() and size() to get the number of books and each combination of␣
 ↪the two variable levels as a pandas Series
bk_counts = sub_df.groupby(['avg_score','qty_books']).size()
bk_counts
```

[30]:
```
avg_score  qty_books
202.0095   0-10 books           0
           11-25 books          0
           26-100 books         1
           101-200 books        0
           201-500 books        0
                               ..
748.7450   11-25 books          0
           26-100 books         1
           101-200 books        0
           201-500 books        0
           More than 500 books  0
Length: 7260, dtype: int64
```

[31]:
```
bk_counts = bk_counts.reset_index(name='count')
```

[32]:
```
bk_counts = bk_counts.pivot(index='qty_books', columns='avg_score',␣
 ↪values='count')
```

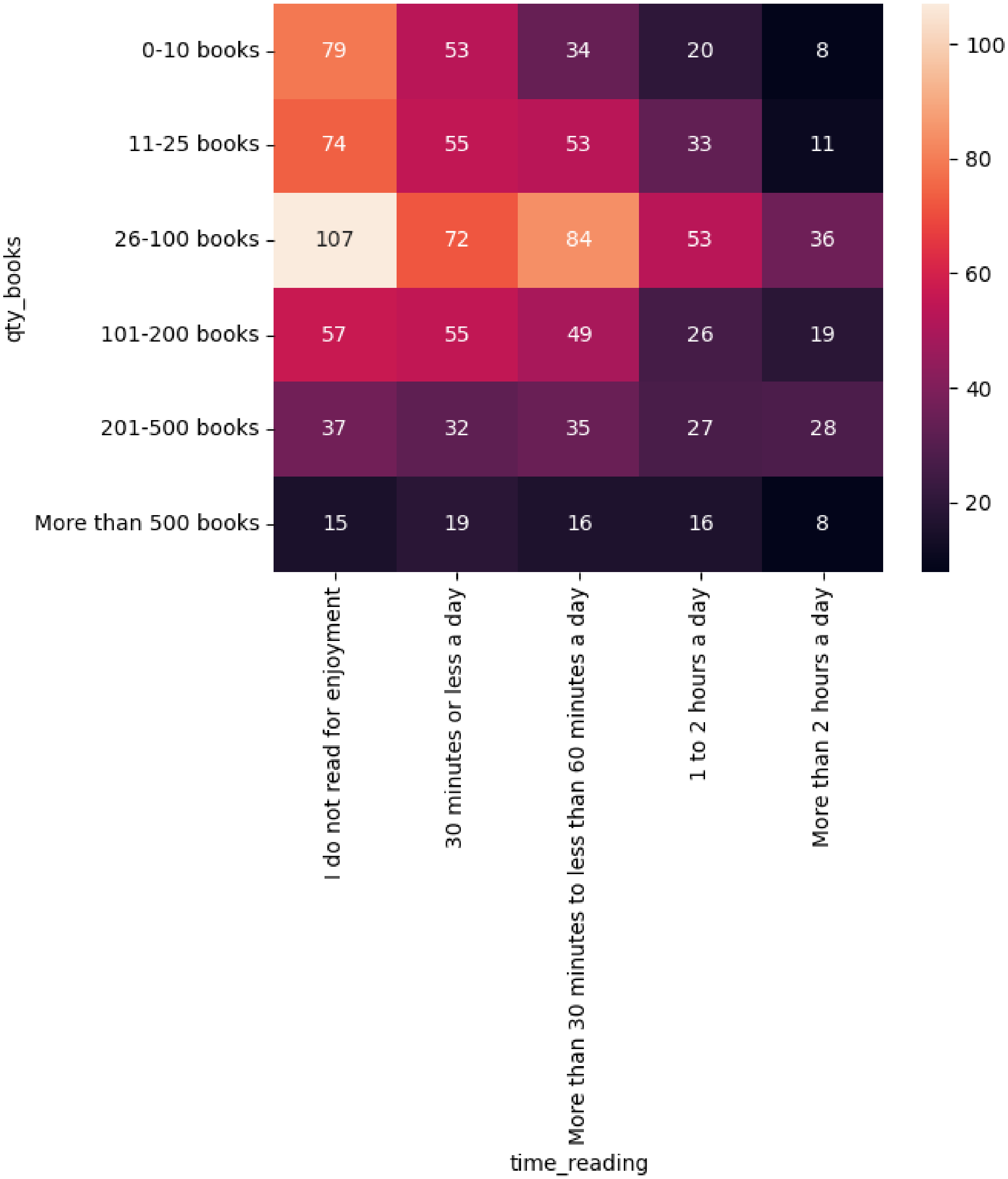
[33]:
```
bk_counts = sub_df.groupby(['qty_books', 'time_reading']).size()
bk_counts
```

[33]:
```
qty_books            time_reading
0-10 books           I do not read for enjoyment                              79
                     30 minutes or less a day                                 53
                     More than 30 minutes to less than 60 minutes a day       34
                     1 to 2 hours a day                                       20
                     More than 2 hours a day                                   8
11-25 books          I do not read for enjoyment                              74
                     30 minutes or less a day                                 55
                     More than 30 minutes to less than 60 minutes a day       53
                     1 to 2 hours a day                                       33
                     More than 2 hours a day                                  11
26-100 books         I do not read for enjoyment                             107
                     30 minutes or less a day                                 72
                     More than 30 minutes to less than 60 minutes a day       84
                     1 to 2 hours a day                                       53
                     More than 2 hours a day                                  36
101-200 books        I do not read for enjoyment                              57
                     30 minutes or less a day                                 55
                     More than 30 minutes to less than 60 minutes a day       49
                     1 to 2 hours a day                                       26
                     More than 2 hours a day                                  19
201-500 books        I do not read for enjoyment                              37
                     30 minutes or less a day                                 32
                     More than 30 minutes to less than 60 minutes a day       35
                     1 to 2 hours a day                                       27
                     More than 2 hours a day                             28
More than 500 books  I do not read for enjoyment                         15
                     30 minutes or less a day                            19
                     More than 30 minutes to less than 60 minutes a day  16
                     1 to 2 hours a day                                  16
                     More than 2 hours a day                              8
dtype: int64
```

```
[34]: bk_counts = bk_counts.reset_index(name='count')
      bk_counts = bk_counts.pivot(index='qty_books', columns='time_reading',␣
       ↪values='count')
```

```
[35]: sns.heatmap(bk_counts, annot=True, fmt='d')
```

```
[35]: <AxesSubplot:xlabel='time_reading', ylabel='qty_books'>
```



Answer

Most students do not read for enjoyment, regardless of the number of books they have. However, those who do read outside of school tend to have access to more books at home.

Question

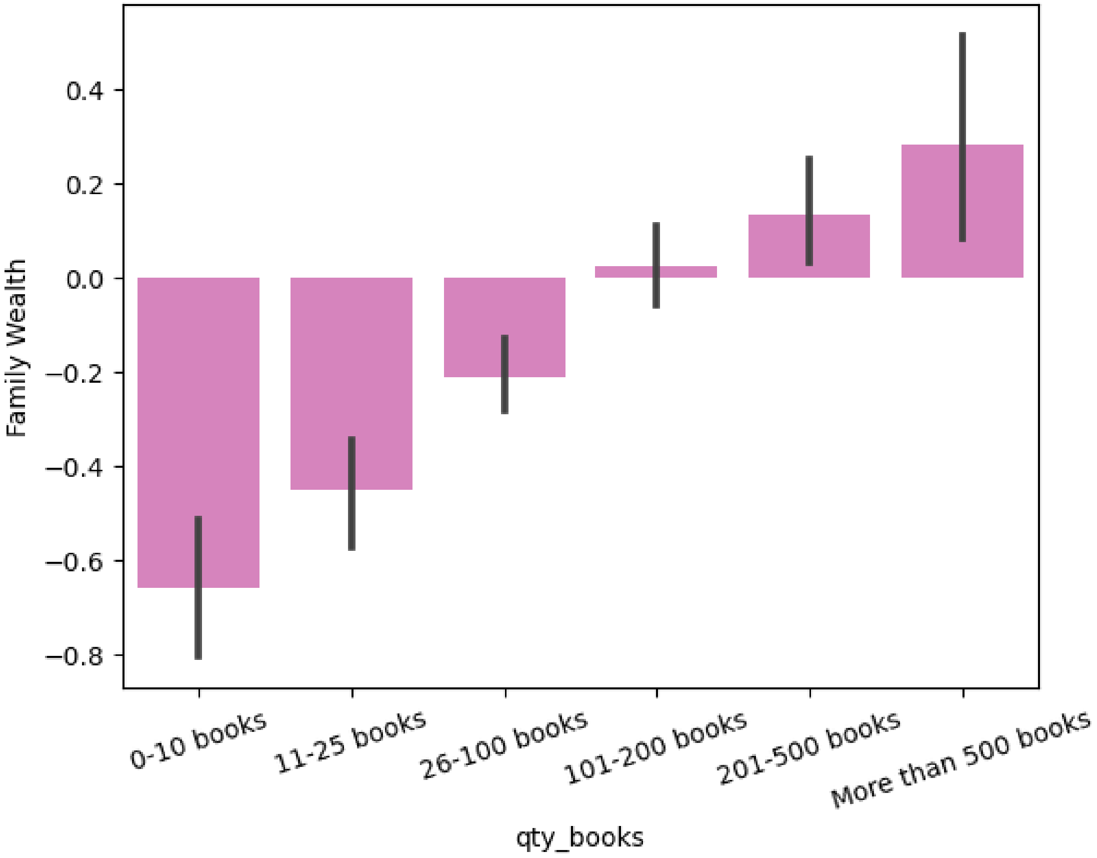
Is there a correlation between family wealth and the number of books in the home?

[36]:

```
colors=sns.color_palette()[6]

sns.barplot(data=sub_df, x='qty_books',y='wealth',
            color=colors);
plt.xticks(rotation=18)
plt.ylabel('Family Wealth')
```

[36]:

```
Text(0, 0.5, 'Family Wealth')
```



Answer

There is a definite positive correlation between family wealth and the number of books in the home. This is not wholely surprising as books are a luxury item.

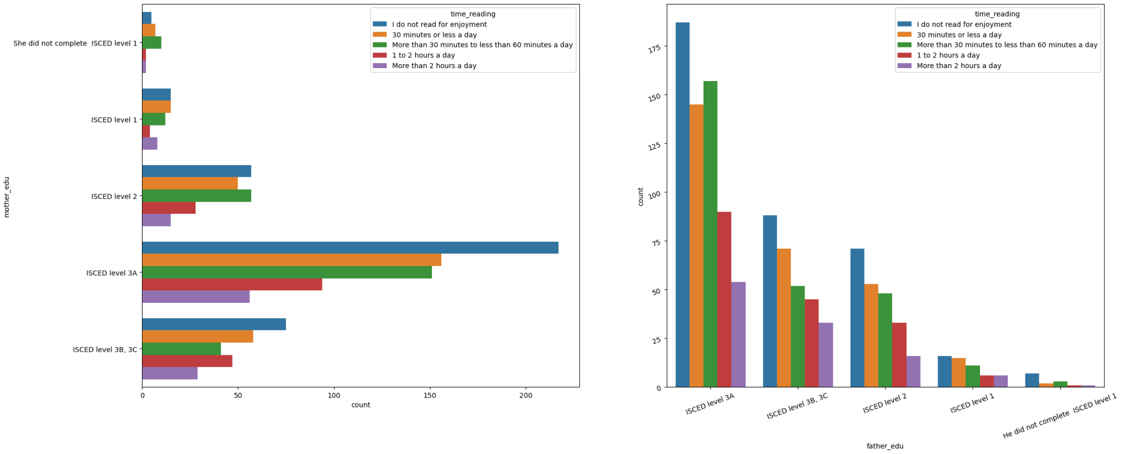
Question

Does the education level of parents correlate to reading outside of school?

[37]:

```
fig, ax = plt.subplots(1,2,figsize=(25,10))

#left plot - mother edu and time reading
```

```
sns.countplot(data=sub_df, y='mother_edu', hue='time_reading',ax=ax[0])
plt.yticks(rotation=20);

#right plot - father edu and time reading
sns.countplot(data=sub_df, x='father_edu',hue='time_reading',ax=ax[1])
plt.xticks(rotation=20);
```



Answer

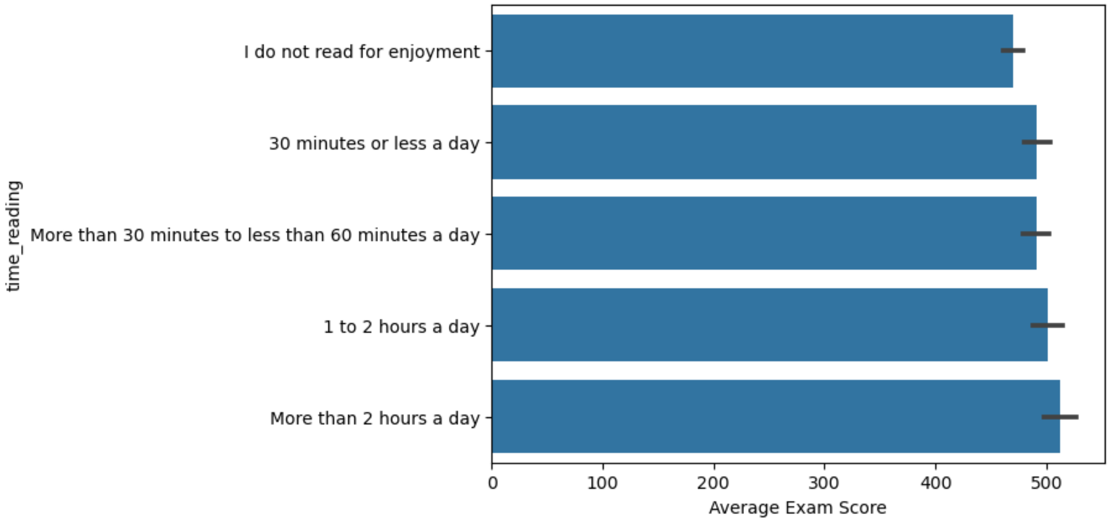
It is interesting that each category of education follow roughly the same shape (most do not read, then read for less than 30 minutes, and so on), except for students whose mother did not complete elementary school. Those students are more likely to read for somewhere between 30 minutes and an hour a day.

Question

Does time spent reading correlate with exam scores?

```
[38]: sns.barplot(data=sub_df, y='time_reading',x='avg_score',
                  color='tab:blue');

      plt.xlabel('Average Exam Score')
```

```
[38]: Text(0.5, 0, 'Average Exam Score')
```



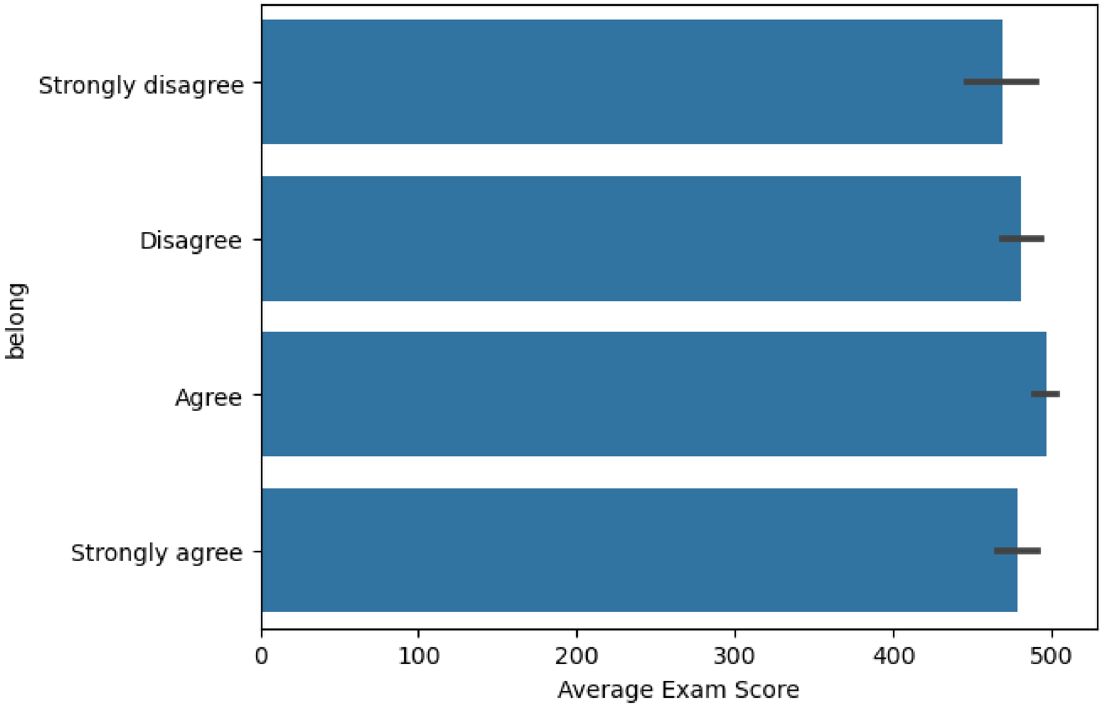
Answer

There is a slight positive correlation between time spent reading and average exam scores, but not likely enough of one to be statistically significant.

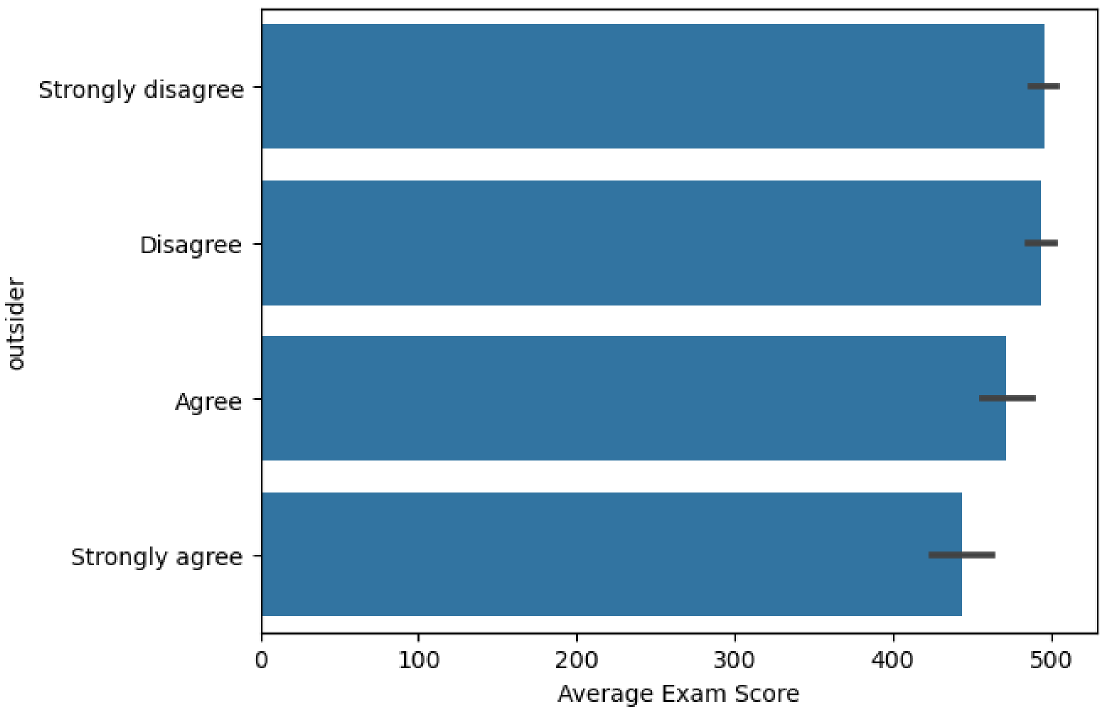
Question

How does a sense of belonging (or lack there of) at school affect exam scores?

```
[39]: sns.barplot(data=sub_df, y='belong',x='avg_score',
                  color='tab:blue');

      plt.xlabel('Average Exam Score');
```



[40]:
```
sns.barplot(data=sub_df, y='outsider',x='avg_score',
            color='tab:blue');

plt.xlabel('Average Exam Score');
```



Answer

There does appear to be a correlation between feeling like they belong at school and a better exam score, and the opposite is true of feeling like an outsider at school. Though the scores seem close, there is enough of a difference between the highest scoring bars and the lowest scoring bars to potentially be significant.

1.13.2 Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Wealth does interact with exam scores, both in general family wealth and factors of wealth, such as the number of books in the home. Another possible indicator of exam scores is a sense of belonging at school. For me, though, the most interesting finding is that students whose mother did not finish elementary school are more likely than other students to read for pleasure. That could be an interesting research project worth further study.

1.14 Multivariate Exploration

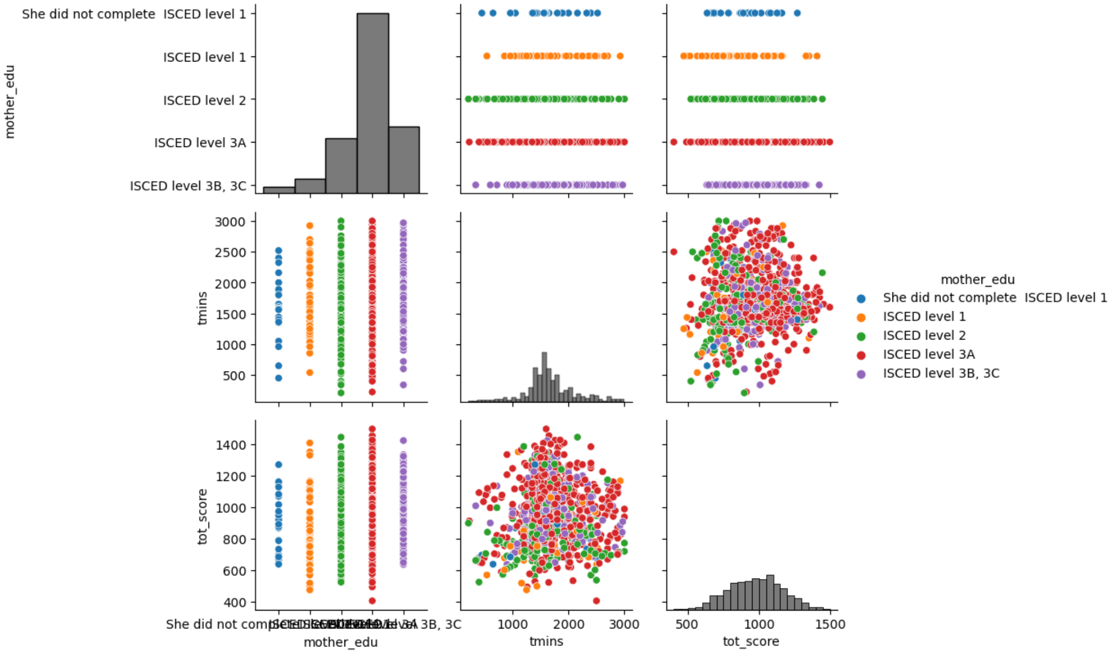
1.14.1 The effect of parental education level and classroom minutes on exam score

Question

Do these variables (parental education, classroom minutes) have a discernable affect on each other and/or overall exam score?

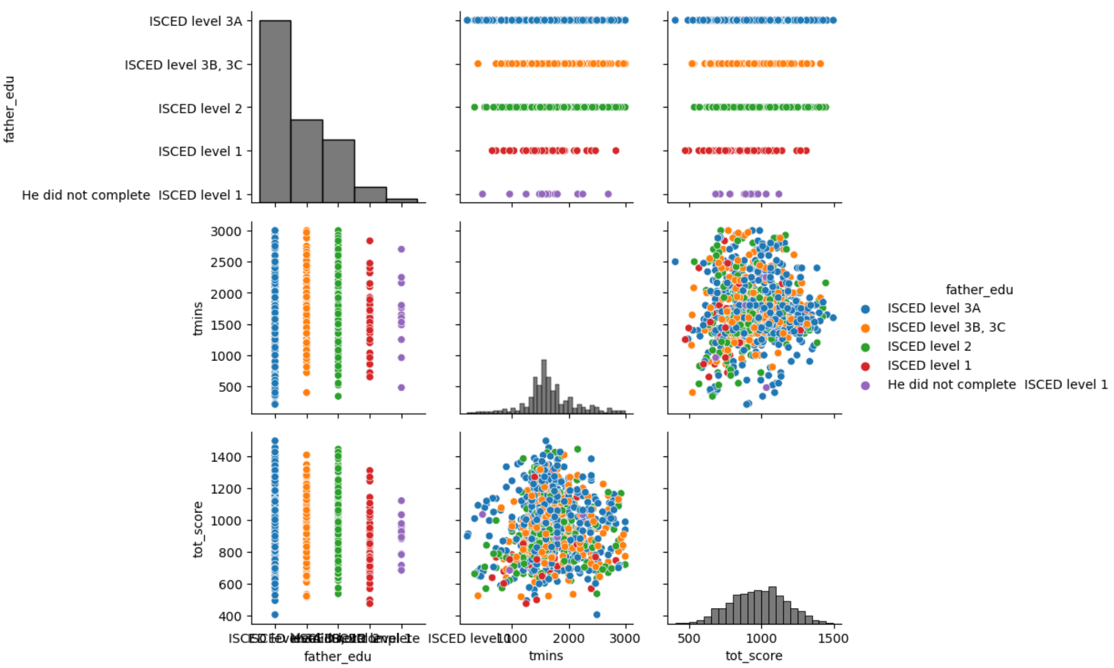
```
[41]: variables = ["mother_edu","tmins", "tot_score"]

      g = sns.PairGrid(sub_df, hue="mother_edu", vars=variables);
      g.map_diag(sns.histplot, hue=None, color=".3");
      g.map_offdiag(sns.scatterplot);
      g.add_legend()
```

```
[41]: <seaborn.axisgrid.PairGrid at 0x7f7fc87c2dc0>
```



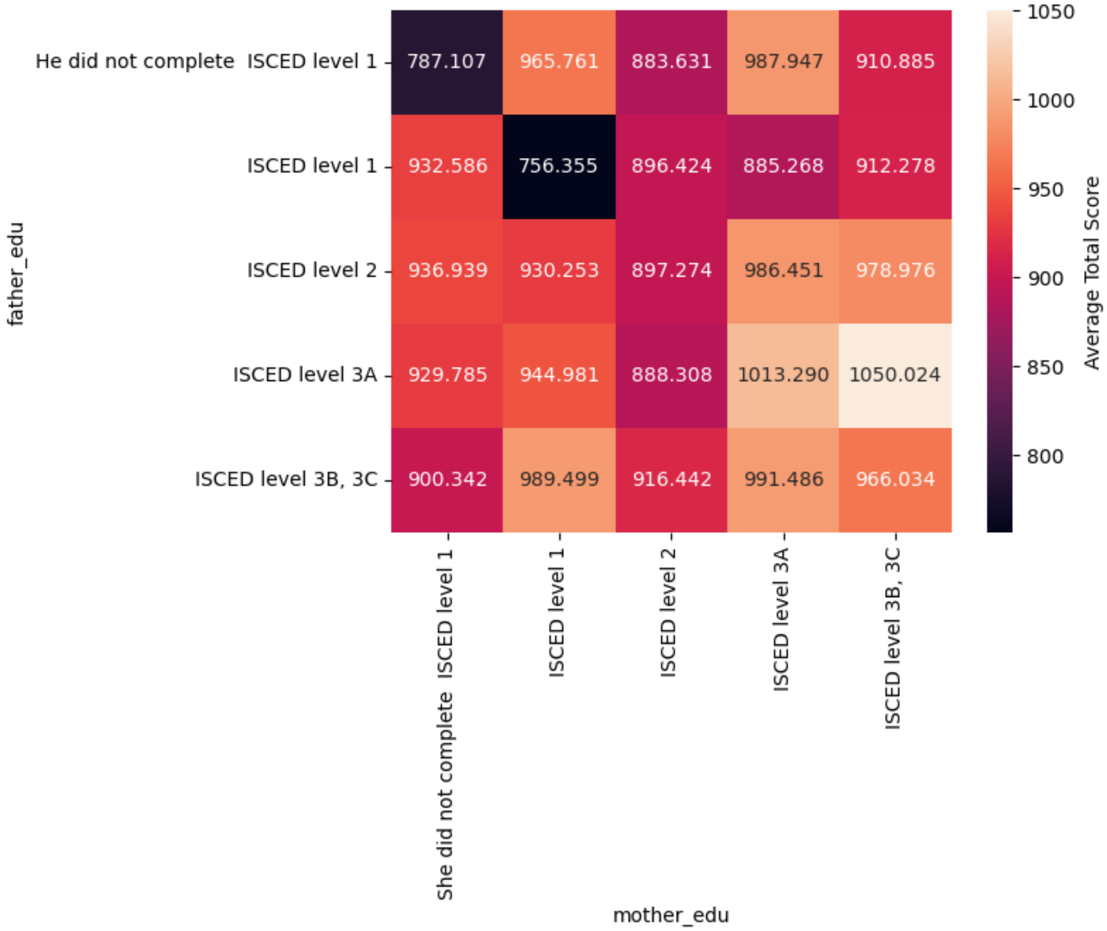
```
[43]: variables = ["father_edu","tmins", "tot_score"]

      g = sns.PairGrid(sub_df, hue="father_edu", vars=variables)
      g.map_diag(sns.histplot, hue=None, color=".3")
      g.map_offdiag(sns.scatterplot)
      g.add_legend()
```

```
[43]: <seaborn.axisgrid.PairGrid at 0x7f7fb8942ee0>
```



```
[44]: cat_means = sub_df.groupby(['mother_edu', 'father_edu']).mean()['tot_score']
      cat_means = cat_means.reset_index(name='tot_score_avg')
      cat_means = cat_means.pivot(index='father_edu', columns='mother_edu',
                                  values='tot_score_avg')
      sns.heatmap(cat_means, annot=True, fmt='.3f',
                  cbar_kws = {'label' : 'Average Total Score'})
```

```
[44]: <AxesSubplot:xlabel='mother_edu', ylabel='father_edu'>
```



Answer

There does appear to be a positive correlation between the academic achievement of parents and their student's exam scores.

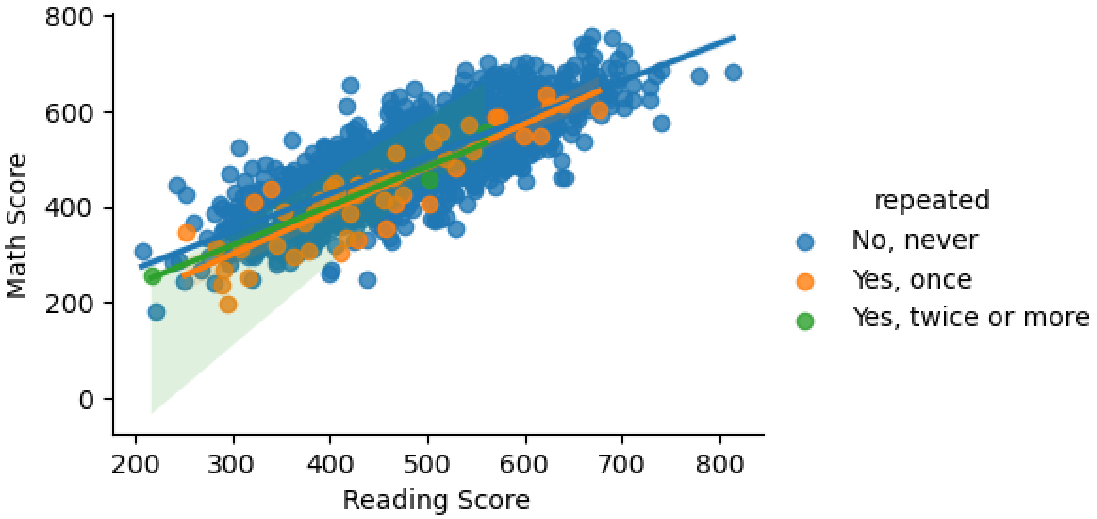
1.14.2 the affect of repeating on reading and math scores

Question

Do students who repeat grades struggle on reading and math as compared to those who do not?

[45]:
```
g = sns.FacetGrid(data=sub_df, hue='repeated',
                hue_order=['No, never','Yes, once', 'Yes, twice or more'],␣
 ↪aspect=1.5)

g.map(sns.regplot, 'pv1read','pv1math', fit_reg=True);

g.add_legend()
plt.xlabel('Reading Score')
plt.ylabel('Math Score')
```

[45]:
```
Text(56.84903549382716, 0.5, 'Math Score')
```



Answer

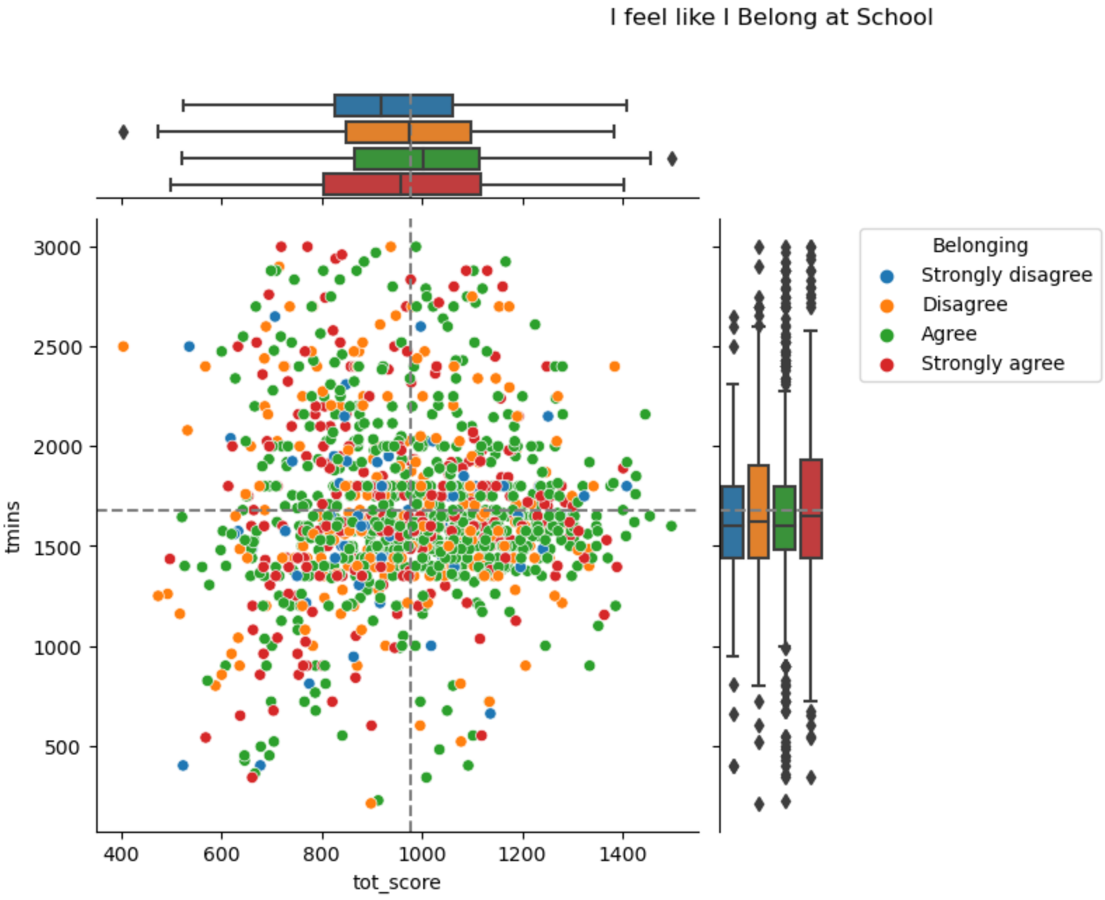
Though few students have repeated a grade, it does appear that those who have repeated score worse on the exams than students who have not.

1.14.3 Sense of belonging and exam performance

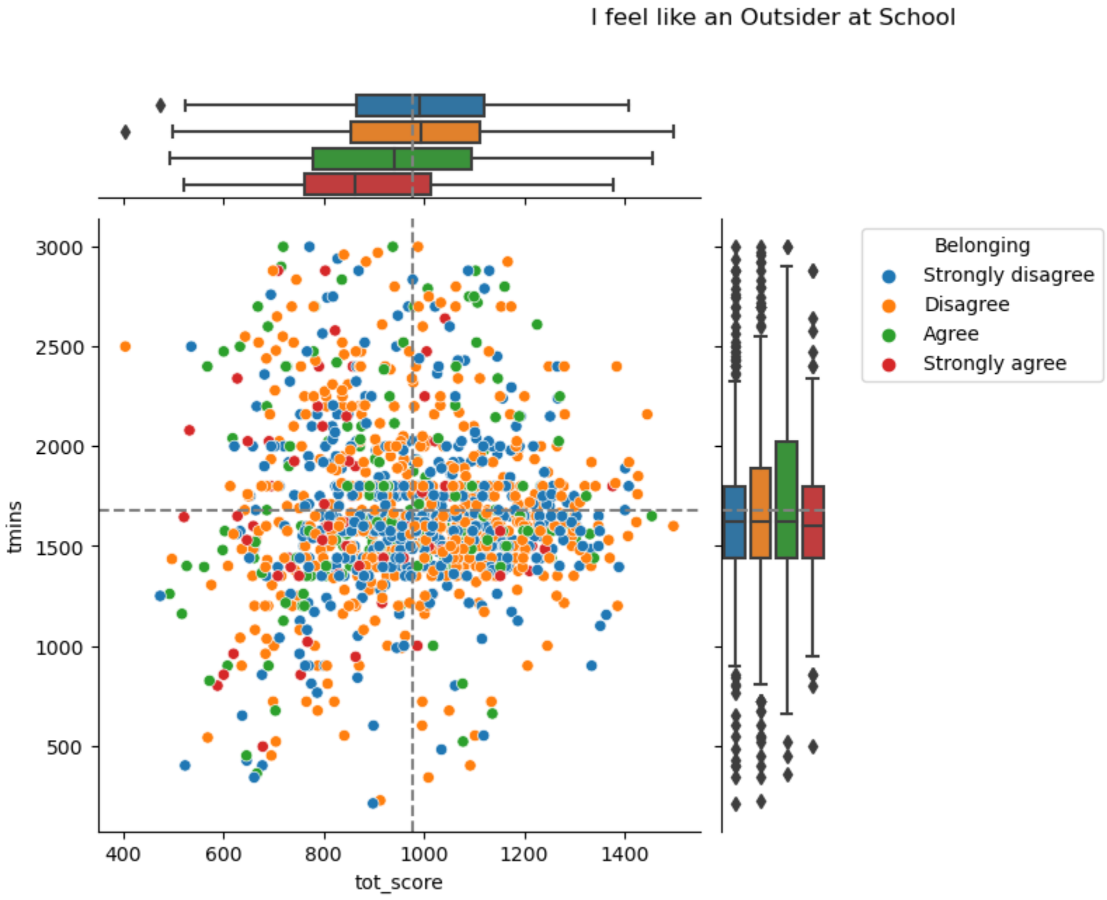
Question

Do students who feel like they belong at school attend classes more regularly and/or do better on exams?

[46]:
```
g = sns.JointGrid(data=sub_df, x="tot_score", y="tmins", hue="belong");
g.plot_joint(sns.scatterplot)
sns.boxplot(data=sub_df, x=g.hue, y=g.y, ax=g.ax_marg_y)
sns.boxplot(data=sub_df, y=g.hue, x=g.x, ax=g.ax_marg_x)
g.refline(x=sub_df['tot_score'].mean(), y=sub_df['tmins'].mean());

sns.move_legend(g.ax_joint, "upper left", title='Belonging',bbox_to_anchor=(1.
 ↪25, 1))

plt.title('I feel like I Belong at School',y=1.0,pad=100.0);
```



[47]:
```
g = sns.JointGrid(data=sub_df, x="tot_score", y="tmins", hue="outsider");
g.plot_joint(sns.scatterplot)
sns.boxplot(data=sub_df, x=g.hue, y=g.y, ax=g.ax_marg_y)
sns.boxplot(data=sub_df, y=g.hue, x=g.x, ax=g.ax_marg_x)
g.refline(x=sub_df['tot_score'].mean(), y=sub_df['tmins'].mean());

sns.move_legend(g.ax_joint, "upper left", title='Belonging',bbox_to_anchor=(1.
↪25, 1))

plt.title('I feel like an Outsider at School',y=1.0,pad=100.0);
```



Answer

Students who feel like they belong at school, overall, score better on the exams than students who do not feel like they belong. The students who feel like they belong also appear to spend more time in school.

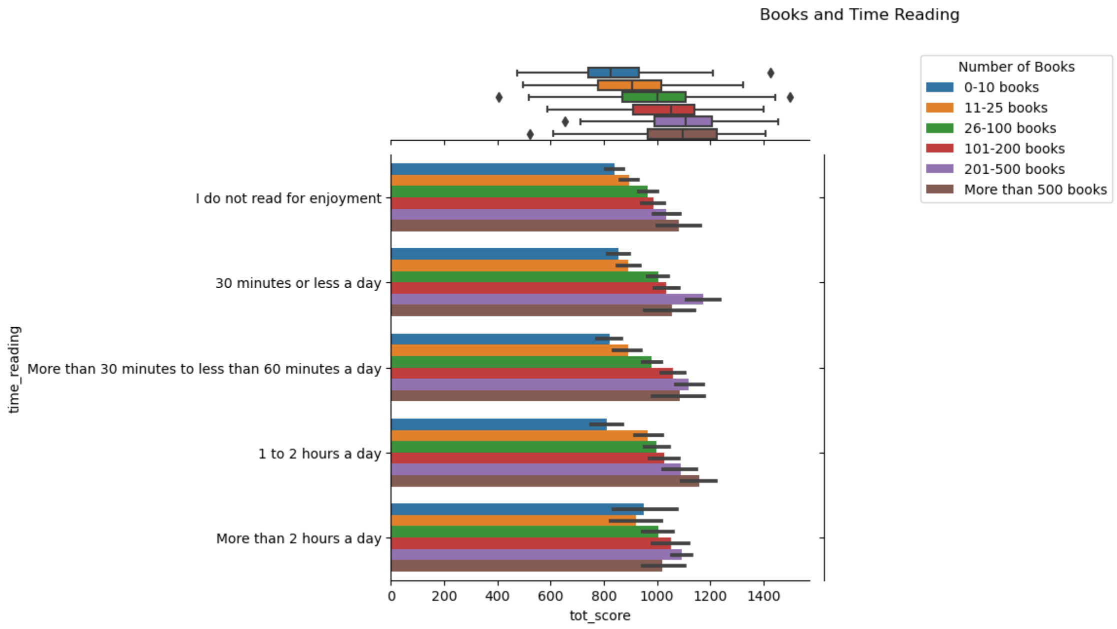
1.14.4 time reading, books at home, and exam scores

Question

Is there a relationship between the number of books in the home and the amount of time a student spends reading? Is there a relationship between those variables and exam scores?

```
[48]: g = sns.JointGrid(data=sub_df, x="tot_score", y="time_reading",␣
      ↪hue="qty_books");
      g.plot_joint(sns.barplot)
      sns.boxplot(data=sub_df, y=g.hue, x=g.x, ax=g.ax_marg_x);

      sns.move_legend(g.ax_joint, "upper left", title='Number of␣
      ↪Books',bbox_to_anchor=(1.25, 1.25))
```

```
plt.title('Books and Time Reading',y=1.0,pad=100.0);
```



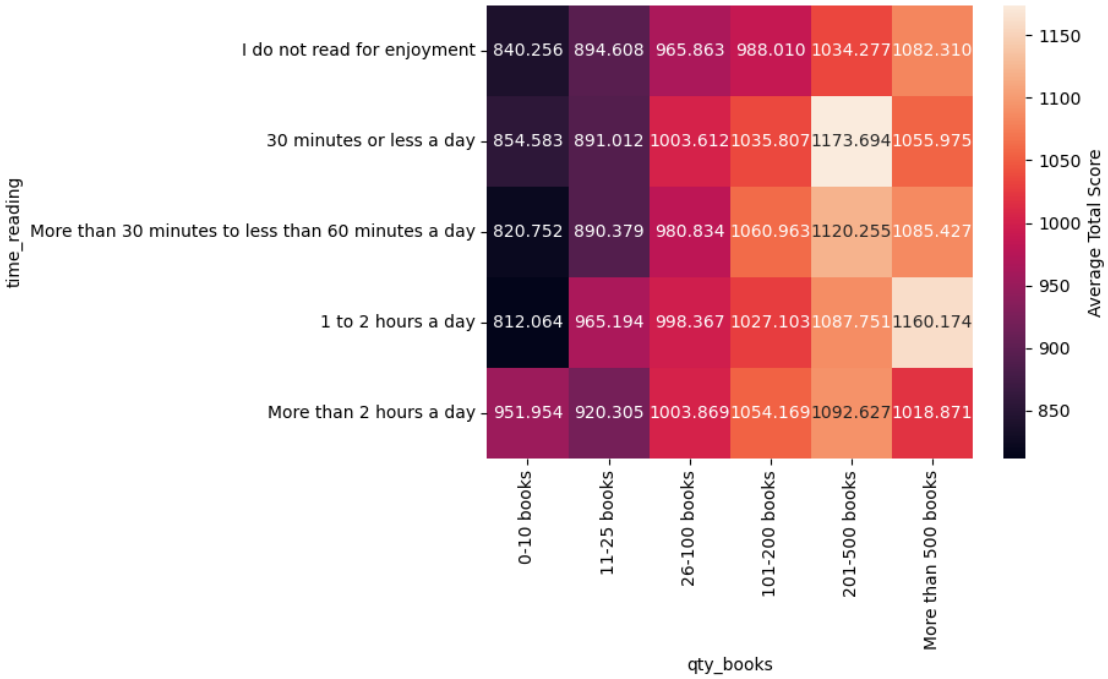
[49]:
```
bk_counts = sub_df.groupby(['qty_books', 'time_reading']).size()
bk_counts = bk_counts.reset_index(name='count')
bk_counts = bk_counts.pivot(index='qty_books', columns='time_reading',␣
  ↪values='count')
bk_counts
```

[49]:
```
time_reading          I do not read for enjoyment  30 minutes or less a day  \
qty_books
0-10 books                                     79                        53
11-25 books                                    74                        55
26-100 books                                  107                        72
101-200 books                                  57                        55
201-500 books                                  37                        32
More than 500 books                            15                        19

time_reading          More than 30 minutes to less than 60 minutes a day  \
qty_books
0-10 books                                                          34
11-25 books                                                         53
26-100 books                                                        84
101-200 books                                                       49
201-500 books                                                       35
More than 500 books                                                 16
```

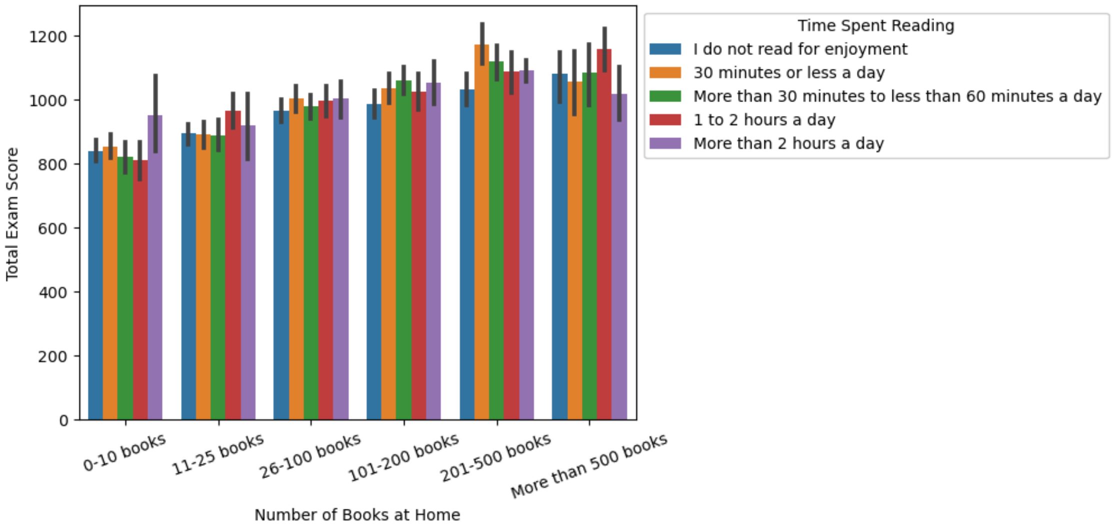
```
time_reading         1 to 2 hours a day  More than 2 hours a day
qty_books
0-10 books                           20                        8
11-25 books                          33                       11
26-100 books                         53                       36
101-200 books                        26                       19
201-500 books                        27                       28
More than 500 books                  16                        8
```

```
[50]: cat_means = sub_df.groupby(['qty_books', 'time_reading']).mean()['tot_score']
cat_means = cat_means.reset_index(name='tot_score_avg')
cat_means = cat_means.pivot(index='time_reading', columns='qty_books',
                            values='tot_score_avg')
sns.heatmap(cat_means, annot=True, fmt='.3f',
           cbar_kws = {'label' : 'Average Total Score'})
```

```
[50]: <AxesSubplot:xlabel='qty_books', ylabel='time_reading'>
```



```
[51]: ax = sns.barplot(data=sub_df, x='qty_books', y='tot_score', hue='time_reading')
ax.legend(loc='best', ncol=1, framealpha=1, title='Time Spent␣
 ↪Reading',bbox_to_anchor=(1, 1));
plt.xlabel('Number of Books at Home');
plt.xticks(rotation=20);
plt.ylabel('Total Exam Score');
```



Answer

It does appear that students who own more books are more likely to read outside of school, and students who read outside of school tend to score higher on the exams.

1.14.5 Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Overall, the higher the sense of belonging, the more time a student spends in school. There is also a positive correlation in the relationship between the education level of parents and total exam scores, as well as time spent reading and exam scores when wealth factors like the number of books in the home are factored in to the analysis.

1.14.6 Were there any interesting or surprising interactions between features?

It does not appear that students who feel like outsiders differ much in attendance, except for the min and max levels. The quartiles and means are about the same across the variable.

1.15 Conclusions

It appears the factors that have the largest impact on exam scores are parent education and a sense of belonging at school.

```
[ ]:
```

```
[ ]:
```