Group 7: Project 1

Samuel Geddie + Bryce Emery

Dataset

Dataset Source Link

Description and Variables

The data describes student's exam scores along with various personal and social characteristics. Observed variables include: gender, race/ethnicity, parental education, lunch program, test preparation, along with math, reading, and writing exam scores.

`gender`

Describes the students sex designation.

- Categorical Binary, with two values 'female' and 'male'.

`race_ethinicity`

Describes the students racial or ethnic background. Groups are obfuscated within this data set.

- Categorical, with 5 values from "Group A" to "Group E".

`parental_level_of_education`

Describes the highest level of education held by the students parentage.

- Categorical, with 6 possible "bins" ranging from "Some Highschool" at the lowest to "Master's Degree" at the highest.

`lunch`

Describes the student's access or inclusion in a free or reduced price lunch program. - Categorical Binary, with values 'free/reduced' and 'standard'.

`test_preparation_course`

Describes the student's completion of a exam preparation course.

- Categorical Binary, with values 'completed' and 'none'.

`math_score`, `writing_score`, `reading_score`

Each describes the student's exam score in the given subject, using the conventional US grading scale. With a score 70 being the expected average.

- Numerical, with values 15-100

Variables Analyzed

- We intend to analyze every variable in our set.
- We will be looking at the Math, Reading, and Writing Scores as our response variables.
  - When looking at any one given score, the other's may be used as explanatory variables.

Data Cleaning

1. To start we load in the data, and run the `clean_names()` function from the janitor library to ensure consistency in variable names.
2. Next, we check the extant data types and examples values of our variables to see if anything needs conversion to a more appropriate type or could be further improved for sake of readability.

- It all looks good!

```
#1. - Loading Data
exams <- read.csv(here::here('data/exams.csv'))
exams <- exams %>% janitor::clean_names()
#2. - Checking data types of entries
str(exams)
```

```
'data.frame':   1000 obs. of  8 variables:
 $ gender                     : chr  "female" "male" "female" "male" ...
 $ race_ethnicity             : chr  "group D" "group D" "group D" "group B" ...
 $ parental_level_of_education: chr  "some college" "associate's degree" "some college" "some
 $ lunch                      : chr  "standard" "standard" "free/reduced" "free/reduced" ...
 $ test_preparation_course    : chr  "completed" "none" "none" "none" ...
 $ math_score                 : int  59 96 57 70 83 68 82 46 80 57 ...
 $ reading_score              : int  70 93 76 70 85 57 83 61 75 69 ...
 $ writing_score              : int  78 87 77 63 86 54 80 58 73 77 ...
```

3. We examine our entries more closely using the `table()` functions looking for any missing (`N/A`) or superfluous entries. entries (e.g. outliers).

- We find no missing or superfluous entries.

4. We look to note any key characteristics of our data and catch potentially problematic values (e.g. outliers).

- We find that our response variables of interest have no obvious problematic variables, and all have a similar expected shape for exam scares, being slightly left skewed.

```
#3.- Checking for missing data
table(exams$gender, useNA ='always')
```

```
female   male   <NA>
   492    508      0
```

```
table(exams$race_ethnicity, useNA ='always')
```

```
group A group B group C group D group E    <NA>
     79     198     323     257     143       0
```

```
table(exams$parental_level_of_education, useNA ='always')
```

```
associate's degree  bachelor's degree        high school    master's degree
               204                 105                215                 75
      some college   some high school              <NA>
               224                 177                  0
```

```
table(exams$lunch, useNA ='always')
```

```
free/reduced     standard         <NA>
         340          660            0
```

```
table(exams$test_preparation_course, useNA ='always')
```

```
completed      none      <NA>
      344       656         0
```

```
table(exams$math_score, useNA ='always')
```

```
 15  20  21  23  24  25  27  28  30  31  32  33  34  35  36  37
  1   1   1   1   1   1   1   2   2   2   1   2   4   4   3   4
 38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
  4   6   2   7   5   7  13  11   8  10   7  10  10  14  12   8
 54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69
 15  18  19  18  21  29  19  26  32  21  25  26  21  29  28  24
 70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
 23  21  23  22  30  25  22  18  11  19  23  23  27  25  20  13
 86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 <NA>
 11  10  15  11  11   9  11   4   6   3   6   4   5   4   9   0
```

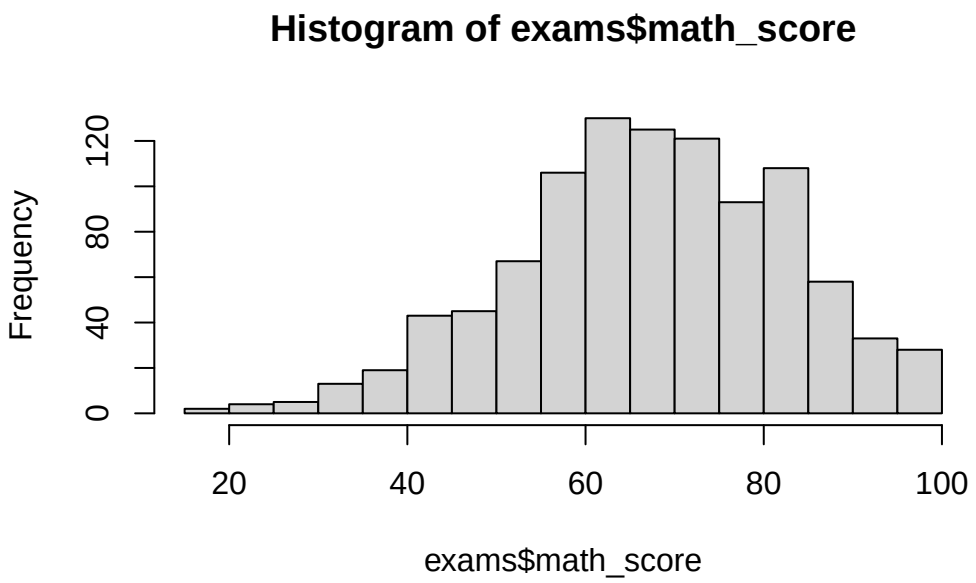
```
table(exams$reading_score, useNA ='always')
```

```
 25  27  30  32  33  34  35  37  38  39  40  41  42  43  44  45
  1   2   1   1   2   1   2   1   3   3   4   1   2   2   4  11
 46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61
  8   7  13  12  11   8   9  15   9   8  28  20  18  16  18  21
 62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77
 25  19  17  29  34  26  32  30  26  20  36  21  23  28  26  35
 78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93
 29  21  18  19  18  16  19  15  20  12  12  15  17  10  10  12
 94  95  96  97  98  99 100 <NA>
  6   8   9   8   5   2  10   0
```

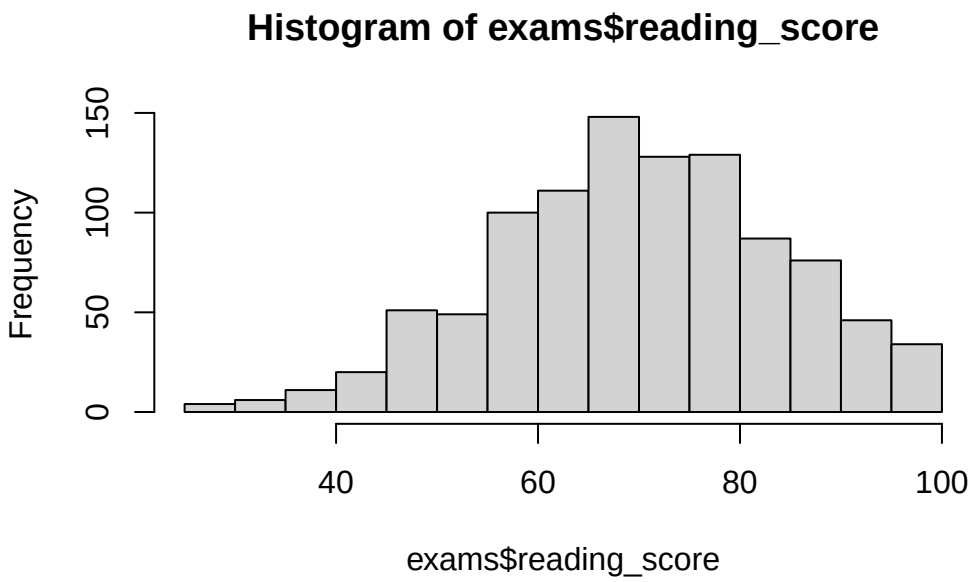
```
table(exams$writing_score, useNA ='always')
```

```
 15  19  23  26  27  28  30  32  33  35  36  37  38  40  41  42
  1   1   1   1   1   1   1   2   3   1   4   3   4   3   7   5
 43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  7  11   7  10  11   9   8   7  17  15  16  13  16  15  23  15
 59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
 23  22  20  22  29  24  16  23  27  24  18  31  22  32  22  31
 75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
 27  25  19  32  17  26  25  13  17  15  13  13  18  15  12  12
 91  92  93  94  95  96  97  98  99 100 <NA>
  7   4  13   8   6   7   6   4   8  13   0
```

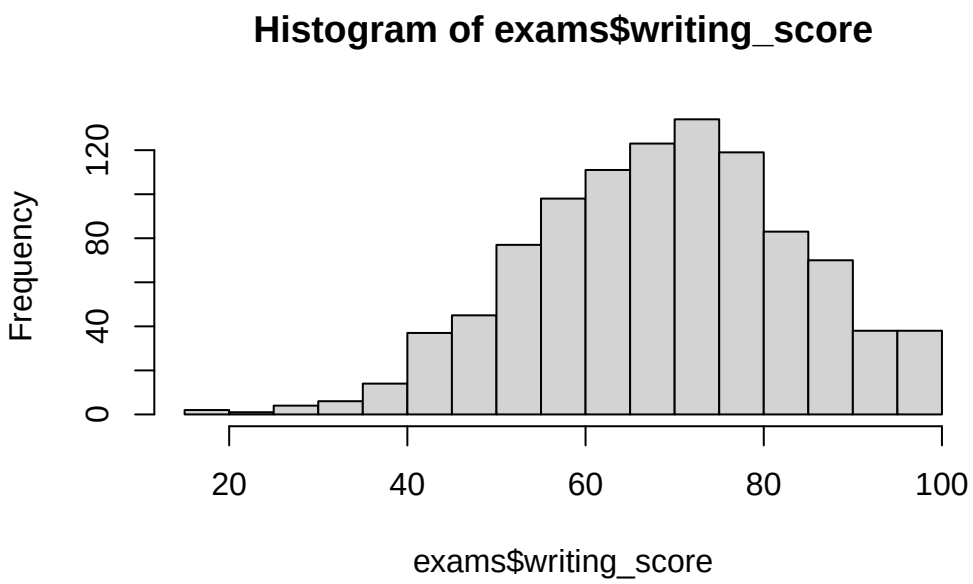
```
# Confirmed no missing entries found
#4. - Checking for shape of Data
hist(exams$math_score, breaks=20)
```




```
hist(exams$reading_score, breaks=20)
```




```
hist(exams$writing_score, breaks=20)
```




```
# Data shows a slightly left skewed distribution because of the way grades are averaged arou
summary(exams[c('math_score', 'reading_score', 'writing_score')])
```

```
   math_score      reading_score    writing_score
 Min.   : 15.00   Min.   : 25.00   Min.   : 15.00
 1st Qu.: 58.00   1st Qu.: 61.00   1st Qu.: 59.00
```

```
 Median : 68.00   Median : 70.50   Median : 70.00
 Mean   : 67.81   Mean   : 70.38   Mean   : 69.14
 3rd Qu.: 79.25   3rd Qu.: 80.00   3rd Qu.: 80.00
 Max.   :100.00   Max.   :100.00   Max.   :100.00
```

```
# No unexpected values.
```

Model Purposes

We wish to calculate which of the observed criteria is most correlated with marked higher or lower test scores in students.

Research Questions

1. "What parameters lead to the highest exam scores for Math, Reading, and Writing respectively?"
2. "If we combined our reading/writing scores, what parameters correlate to higher results? Do these differ from the individual score instances."
3. "Are higher scores in one domain predictive of higher scores in others?"
4. "What measurable effect do subsidized lunches have on student test scores?"
5. "Do subsidized lunches have stratified effects for students of different parental educational attainments? (Essentially, using educational attainment as a indicator for socioeconomic status."