Analysis Report

The purpose of this assignment was to assist Maria, the chief data scientist for a city school district, to analyze data on student funding and students' standardized test scores and to aggregate the data and showcase trends in school performance.

In doing so, I started by importing the CSV file and creating a path.

import pandas as pd

import os

path = "Users\estel\Downloads"

mypath = os.path.join('c:\\',path,'new\_full\_student\_data.csv')

student\_df = pd.read\_csv(mypath)

print(student\_df.head()).

From there, I was able to get the first 5 set of the data. After I got my dataset, I prepared and cleaned the data by removing all the NaN values, removing all duplicated rows, removed the “th” suffix from the grade columns by using the “str” and replace function. Finally, I changed the grade column into “int” type as it shows in the figure below.

student\_df = student\_df.dropna()

student\_df = student\_df.drop\_duplicates()

student\_df["grade"] = student\_df["grade"].str.replace("th","”)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 103880842 | Travis Martin | 9 | Sullivan High School | 59.0 | 88.2 | Public | 961125 |
| **1** | 45069750 | Michael Brown | 9 | Dixon High School | 94.7 | 73.5 | Charter | 870334 |
| **2** | 45024902 | Gabriela Lucero | 9 | Wagner High School | 89.0 | 70.4 | Public | 846745 |
| **3** | 62582498 | Susan Richardson | 9 | Silva High School | 69.7 | 80.3 | Public | 991918 |
| **5** | 74579444 | Cynthia Johnson | 9 | Montgomery High School | 63.5 | 76.9 | Charter | 893368 |
| **..** |  |  |  |  |  |  |  |  |

student\_df["grade"] = student\_df["grade"].astype("int64")

student\_df.dtypes

student\_id int64

student\_name object

grade int64

school\_name object

reading\_score float64

math\_score float64

school\_type object

school\_budget int64

dtype: object

Following the steps above, I summarized the statistic of the dataFrame using the describe function, and displayed mean of the math\_ score and stored the minimum reading\_score in the “min\_reading\_score”. Also, I needed to display the grade column and the first three rows of Columns 3, 4, and 5 by using the “loc” and “iloc” respectively. These are the codes that I used for each function.

from statistics import mean

student\_df.describe()

mean(student\_df["math\_score"]) 64.6757332614119

min\_reading\_score = student\_df["reading\_score"].min() 10.5

Finally, I was able to compare the data. I compared the data from the average budget for each school type as well as their mean.

student\_df.groupby("school\_type").mean().loc[:,["school\_budget"]]. This is the result

school\_budget

school\_type

Charter 872625.656236

Public 911195.558251

The amazing thing that I discovered, is the fact that with jupyter notebook we can sort our dataset and choose and pick what row or column we to display. The perfect example is the figure below. I was able to find the average of the math score from the two types pf schools and grouped them by grades.

student\_df.groupby(by=["school\_type","grade"])["math\_score"].mean().round()

school\_type grade

Charter 9 70.0

10 66.0

11 68.0

12 60.0

Public 9 64.0

10 64.0

11 59.0

12 64.0

Name: math\_score, dtype: float64