DATA WRANGLING

Springboard Chapter 4.3 Submission
Prepared by Mohini Singh

CAPSTONE PROJECT DATASET

My Capstone project is to predict the enrollment count of California School Districts. The dataset is the publicly available enrollment data from the California Department of education at https://www.cde.ca.gov/ds/ .

THE DATA SET

1. The data set is available for the years 1981 to 2018. There is an individual file for each year going back to 1994. 1981 to 1993 is in one single file. The file format and column names vary from file to file.
2. I am interested in enrollment count by school district by year.
3. Each row in the data files has an enrollment count grouped by school, ethnicity and gender.

DATA WRANGLING

1. *Grouping data by school district* - Some files have the school district name as a column but some don't. All files have a code CDS_CODE which is set of numbers. The first part of the code is the code for the school district. I extracted the school district code from the CDS code to determine the school district for each row of data. I also created a data dictionary of all school district codes and names.
2. *Enrollment count by school district*- I created a summary data file of enrollment count by school district for all years of available data.
3. *Enrollment count grouped by school district and school*- I similarly created a summary data file of school district, school, enrollment count for each year of available data. I also created a data dictionary of all school codes and their names.
4. *Ethnicity missing values* - Some of the data files had missing values in the ethnicity column. If a particular year had ethnicity data missing, I decided to ignore that

whole year of data. I summarized the data to get enrollment count by school district and ethnicity. I also created a data dictionary of all ethnicity codes and names.

METHODOLOGY AND TOOLS

1. I used Jupyter to read the files and mainly used the pandas package for reading and writing the summary data files.
2. I created a BigQuery database on Google Cloud and linked it to Mode Analytics to examine the data in SQL. Exploring the data in SQL gave me some quick insights into distinct values, missing values and group counts.