# Exploratory Data Analysis

For data analysis, we kept the visualizations simple to highlight the key features of every attribute. For categorical attributes, we mainly used bar charts since it allowed us to compare different variables, whereas histograms were used for continuous attributes to represent distributions of data. Our approach for handling longitude and latitude was to incorporate those values together into a global COVID-19 heatmap that is easy to visualize. For missing values, we simply checked the number of null entries within each attribute from both the cases\_train.csv and location.csv datasets and plotted them respectively using bar charts.

# Data Cleaning and Imputing Missing Values

To clean the age data, we took a list of every invalid entry in the datasets and manually replaced these values with the mean of that age range to reduce the amount of assumption that need to be made with the ages. After that, we replaced all empty values with the mean of the newly calculated dataset. Unfortunately, much of the training dataset has the age column empty so this creates an inflated number of values in the range of the mean values, as can be seen in the plot of age values.

With the other attributes, empty values were defaulted as either an “Unknown” or “None” value. The nature of these attributes did not allow for easy inferring of their values. Location-related information could not be inferred based on the values of the other rows, and neither could confirmation date and outcome. Sex being a binary attribute also made it impossible to impute.

# Dealing with Outliers

# Transformation

To transform the location data, we compiled the data from each individual county and aggregated them from the county level to the state level. Case numbers were aggregated by simply summing together the counties’ numbers. Incidence rate was aggregated by taking the mean of the individual rates, and case-fatality was manually recalculated by taking the quotient of the deaths and confirmed cases. Latitude and Longitude were aggregated by taking the mean values, and finally the last updated attribute was aggregated by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

# Joining the Cases and Location Datasets

# 1.6 Outcome Labels