# 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. That is, if the entry were “30-39”, for example, the entry would be replaced with “35”. After that, we replaced all empty entries with “Unknown”. We had considered using the mean age of the available data to fill in the missing values, but unfortunately too much of the training dataset had an empty age so this would have created an inflated number of values in the range of the mean 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

For data transformation, 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.

We decided to remove the last updated attribute because we felt it did not make sense to use any of the possible solutions. If we were to use the latest date, this would be inaccurate as not all the counties would have been last updated on that date, and this could lead to misconceptions on the data. Using either the earliest date or the average date also would not make much sense, so we decided to omit the column altogether.

# Joining the Cases and Location Datasets

To join the cases and location datasets, we decided to only add the incidence rate and case-fatality ratios to the cases datasets, with slight adjustments. We felt that active cases, deaths, etc. would not provide any useful additional information, so we instead look at the provincial and country level incidence rates and fatality rates. We also slightly adjusted incidence rate to be a percentage of 100,00 as opposed to the raw number per 100,000 as this would be easier to understand. Provincial level data could be heavily skewed due to small sample sizes, so we decided to also include the incidence rate and case-fatality ratio at the country level to give a better indicator of actual statistics of that area.

# Outcome Labels

The different ‘outcome’ labels: ‘hospitalized’, ‘nonhospitalized’, ‘deceased’, and ‘recovered’ represent the possible states of a person diagnosed with COVID-19. In the context of our dataset cases\_train.csv, they are the possible values of the class label ‘outcome’ that we will be training an algorithm to predict using all the other attributes in the dataset. The type of data mining task used for predicting the outcome labels is called classification.