**Prepare the Data**

* The datasets are merged into a single DataFrame
  + Left merge on “Mouse ID”
* The number of mice are shown from the merged DataFrame
  + Length of “Mouse ID” unique values
* Each duplicate mice is found based on the Mouse ID and Timepoint
  + .duplicated function on a .loc to find the duplicate row
  + All data retrieved on .loc where “Mouse ID” column = duplicated mouse
* A clean DataFrame is created with the dropped duplicate mice
  + .loc function to return only rows not equaling (!=) duplicated mouse
* The number of mice are shown from the clean DataFrame
  + Length of “Mouse ID” unique values

**Generate Summary Statistics**

* The mean of the tumor volume for each regimen is calculated using groupby.
  + Groupby drug regimen -> tumor volume -> .mean
* The media of the tumor volume for each regimen is calculated using groupby.
  + Groupby drug regimen -> tumor volume -> .median
* The variance of the tumor volume for each regimen is calculated using groupby.
  + Groupby drug regimen -> tumor volume -> .var
* The standard deviation of the tumor volume for each regimen is calculated using groupby.
  + Groupby drug regimen -> tumor volume -> .std
* The SEM of the tumor volume for each regimen is calculated using groupby.
  + Groupby drug regimen -> tumor volume -> sem
* A new DataFrame is created with using the summary statistics.
  + pd.DataFrame with columns for each of above calculations

**Create Bar Charts and Pie Charts**

* A bar plot showing the total number of timepoints for all mice tested for each drug regimen using Pandas is generated.
  + Created DataFrame off of “Drug Regimen” value counts
  + .plot bar chart
* A bar plot showing the total number of timepoints for all mice tested for each drug regimen using pyplot is generated.
  + plt.bar function
* A pie plot showing the distribution of female versus male mice using Pandas is generated.
  + Groupby sex -> mouse ID counts -> to new DataFrame
  + .plot pie chart
* A pie plot showing the distribution of female versus male mice using pyplot is generated.
  + plt.pie function

**Calculate Quartiles, Find Outliers, and Create a Box Plot**

* A DataFrame that has the last timepoint for each mouse ID is created using groupby.
  + First, used .isin function to filter for the four regimens
  + Groupby mouse ID -> returned max ‘timepoint’
  + .reset\_index to reset index
  + Merged the grouped DataFrame with the filtered DataFrame on ‘mouse ID’ and ‘timepoint’, left type
* The index of the DataFrame is reset.
* Retrieve the maximum timepoint for each mouse.
* The four treatment groups, Capomulin, Ramicane, Infubinol, and Ceftamin, are put in a list.
  + regimens = ["Capomulin", "Ramicane", "Infubinol", "Ceftamin"]
* An empty list is created to fill with tumor volume data.
  + tumor\_vol\_list = []
* A for loop is used to display the interquartile range (IQR) and the outliers for each treatment group
  + for x in regimens: tumor\_volume = .loc where ‘Drug Regimen’ equals ‘regimens’, also return ‘tumor volume’ column
  + append the tumor\_volume to the empty list
  + calculate quartiles, iqr, then upper/lower bounds based off iqr
  + outliers = .loc to find any value where tumor\_volume is less than lower bound or more than upper bound
* A box plot is generated that shows the distribution of the final tumor volume for all the mice in each treatment group.
  + plt.boxplot

**Create a Line Plot and a Scatter Plot**

* A line plot is generated that shows the tumor volume vs. time point for one mouse treated with Capomulin.
  + X axis is .loc for ‘mouse ID’ of l509, and the timepoint
  + Y axis is .loc for ‘mouse ID’ of l509, and the tumor volume
  + plt.plot
* A scatter plot is generated that shows average tumor volume vs. mouse weight for the Capomulin regimen.
  + Filter for only Capomulin rows using .loc of ‘drug regimen’ = capomulin
  + Groupby mouse ID and return mean
  + X values are weight column of groupby DataFrame above
  + Y values are tumor volume column of groupby DataFrame above

**Calculate Correlation and Regression**

* The correlation coefficient and linear regression model are calculated for mouse weight and average tumor volume for the Capomulin regimen.
  + correlation = st.pearsonr(x\_values, y\_values)
  + (slope, intercept, rvalue, pvalue, stderr) = linregress(x\_values, y\_values)
  + regress\_values = x\_values \* slope + intercept
  + plt.plot(x\_values,regress\_values,"r-")
  + overlayed onto scatter plot from above