MidPoint Report

Project Objective

For the project objective there are basically three parts we are going to approach which is Statistic, visualization and direction finding.

For statistic:

The package we will use for this part of the project is **Pandas** which contains many statistical functions that we can use. For Pandas, there are two types of data structure which is dataframe and series.

- A five-number summary (min, max, mean, lower quartile, upper quartile)
    - For this question, we will be using a series struct and “describe” function which will give us many useful information such as count, mean, std, min, 25%, 50%, 75% and max.
    - Calculate the maximum, minimum, average, lower quartile and upper quartile of each route by pandas and make those data be a new table. Then analyze the mean value of all routes, and draw a box plot for those data, so that you can intuitively visualize the position and size of the largest, smallest, median, upper quartile, and lower quartile. We can also observe the skewness of the box plot, and the comparison of data intervals is also easy.
    - We can analyze the maximum data and minimum data to get the largest deviation of all the route and smallest deviation of all the route and so on.
- A measure of the central tendency of deviation (mean, median, IQM)
    - For this section, we will first need the function mean() for getting mean value
    - For median, we will use median() for result
    - For IQM(Interquartile mean), in wikipedia they state the way to calculate IQM:”In calculation of the IQM, only the data between the first and third quartiles is used, and the lowest 25% and the highest 25% of the data are discarded.” This means that we need to find the lowest 25% by quantile(0.25) and highest 25% by quantile(0.75). And then select the data

value between this two number and find their media by median()

$$x_{\mathrm{IQM}} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

  - We will use countplot to show differences(mean, median, IQM) between trips.
- A measure of the dispersion of the deviation (e.g., standard deviation of the breadcrumb reading deviations, variance, range, interquartile range,)
  - For standard deviation, we use std() function
  - For variance, we use var() function
  - We can get range from min value and max value
  - Find quantile(0.25) and quantile(0.75) and IQR = quantile(0.75)-quantile(0.25)
- A measure of skewness (e.g., Pearson's skewness coefficient)
  - $$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$
  - print(route.skew())
  - Calculate the mean value of each route and place it in the table. Draw a histogram, the x-axis represents the magnitude of the deviation value, and the y-axis represents the number of the route which is in the range of x-axis. In this way, an intuitive skew diagram is obtained. We know that when the image becomes more symmetrical, its skewness value will be close to zero. The left skew of the image shows that the route number of low deviation is more, and the skew of the right image shows that the route number of high deviation is more. We can use pandas to calculate the skew of the data and observe its positive and negative.
- 90th percentile of deviation (that is, 90% of the deviations are less than what amount of deviation?)
  - quantile(0.9)

`

For visualization:

For this part, we will use the **Seaborn** package for visualization, and use the **matplotlib** package to visualize data.

- Histogram
    - Use matplotlib to draw histogram plots -- plt.hist()
    - A measure of the central tendency of deviation (mean, median, IQM)
    - A measure of skewness
    - https://seaborn.pydata.org/generated/seaborn.distplot.html
- Box plot
    - Use matplotlib to draw box plots -- plt.boxplot()
    - Routes mean data,min, max, median, lower quartile and upper quartile
    - A measure of the dispersion of the deviation
    - https://seaborn.pydata.org/generated/seaborn.boxplot.html
- Countplot
    - Use countplot to show the difference of mean, median and IQM for different trips
    - http://seaborn.pydata.org/generated/seaborn.countplot.html?highlight=countplot#seaborn.countplot
- Kernel density estimaton
    - Use *kdeplot()* to show a general deviation for each trip so that the data would be more intuitive
    - http://seaborn.pydata.org/generated/seaborn.kdeplot.html?highlight=kdeplot#seaborn.kdeplot

For direction of deviation:

For this section, we will use **math** package and by using atan2(x2-x1,y2-y1), we can get a number associated with pi, and from the result of pi we got, we will know which quadrant the point is in. Then we will have a basic idea of which direction and a specific angle for which direction the deviation is.

Project Approach

We have abandoned the machine learning part, and come up with a more detailed plan for data statistics and analysis. Also, we have been playing around with data and have already done some exercise with it. For project planning, we have been doing a lot of research for packages and libraries.

Team Structure

We divided these large problems into smaller problems. After getting the data from Bruce which will contains record longitude, record latitude, corrected latitude, corrected longitude and a time stamp, and a route ID and a distance of deviation, Tianhui will do basic data processing such as mean, maximum and minimum, and analyze the basic data and skewness. Yixuan will analyze the central tendency of deviation, the dispersion of the deviation and 90th percentile of deviation.

Project Milestones

Starting with looking into data

Have do some implement with data

Doing research for information while waiting for final data

Date / time of scheduled midpoint meeting

7/23/2020 2:20PM