Answers to qualitative questions

General instructions

Your responses should be coherent, clear and precise. Use of bullet points is acceptable.

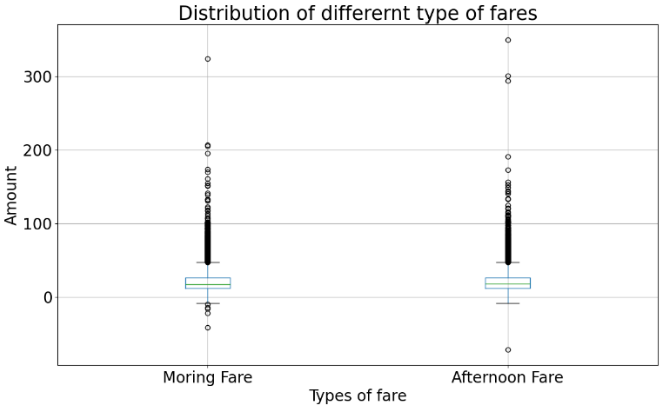
Task2

discussion (100-200 words in length) on the following - the definition of this task was analysing a particular type or abnormality. Explain two further types of properties that could be checked to look for highly abnormal records in this dataset. Be specific, make your properties as distinct as possible from each other and justify your reasoning.

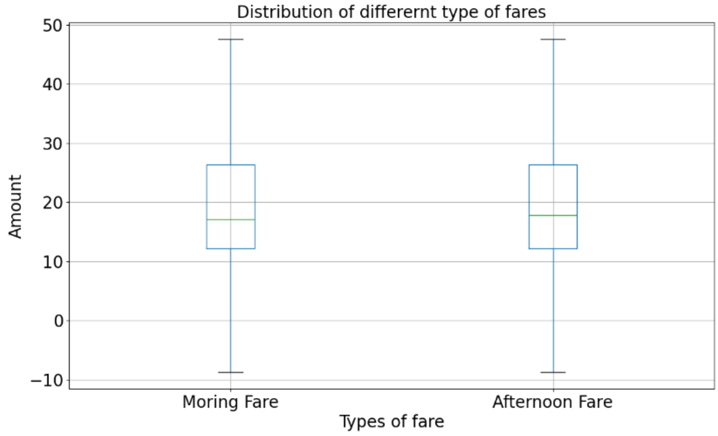
The other properties that could be checked is the total amount. The total amount should not be negative. We could exclude the entries that have total amount below 0. This is a range error and could be potentially caused by inconsistence data collection method that take total amount as amount taken. The other property we should check is the date. The data intend to entry by time. We could check if a adjacency date are same to check if there is a format error. Like putting month first instead of date first or there is wrong entry to the data.

Task3

discussion commenting on/comparing your two boxplots and discussing real world implications of the findings. should be 100-150 words in length.



The boxplot with fliers is hard to analysis because some abnormal data expand the range of the boxplot and make it hard to compare the box where most data lies.



This is the boxplot which doesn’t include the fliers. From this plot, the distribution of morning fare and afternoon fare are almost identical. That implies that different time period in a day doesn’t affect the amount of fare. Most of the fare amount lies between 10 to 30. Also, in the boxplot there is negative values. It is an range error and could be caused by data collection method which consider amount spent as negative values.

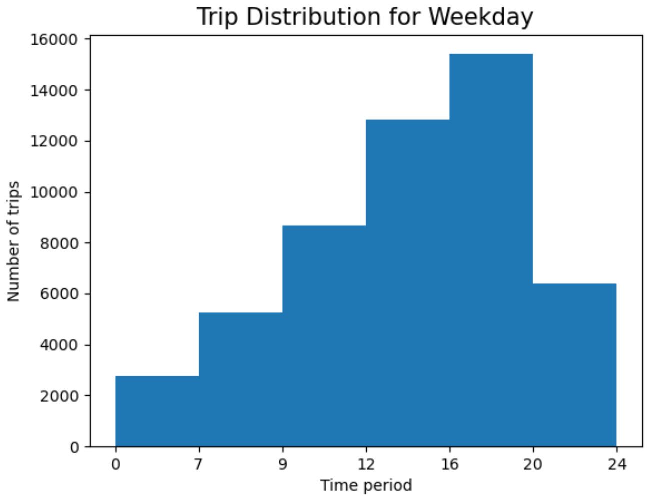
Task4

discussion analysing your calculated value and discussing real world implications of the finding. This should be 50-100 words in length.

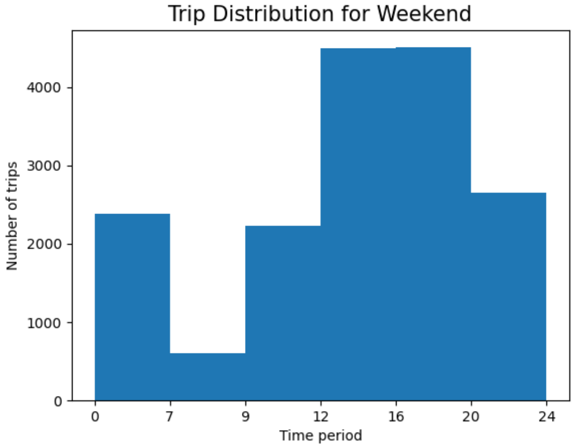
The percentage of weekend trip in January is 24.73%. In 2023 January there are 9 weekend days. 29.03% of days in January 2023 are weekend days. Which is higher that the percentage of weekend trip. This suggests that during weekends, people are less likely to take taxies than weekdays.

Task5

discussion analysing the two histograms individually and jointly and discussing real world implications of the findings. This should be 100-200 words in length.



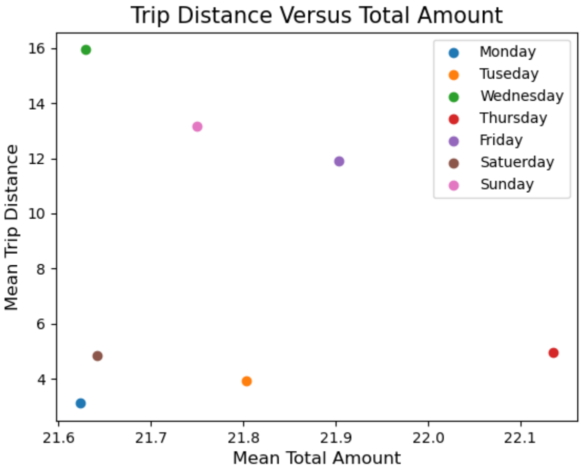
The distribution of trip for weekday shows that most trip happens between 12-20 o’clock. And are less likely to happened in mid-night.



The distribution for weekend shows that the during 7-9 o'clock it has the lowest number of trips. This is different from the weekday distribution. The decrease in the 7-9 time period is most likely to be affected by workdays. There may be some portion of people using taxies to go to work. The proportion of trips in midnights (20-7) for weekend is much higher compared to weekdays. The real-world reason could be people hang out late in weekends and use taxies to go home. Still, most of the trip happened in 12-20 o'clock.

Task6

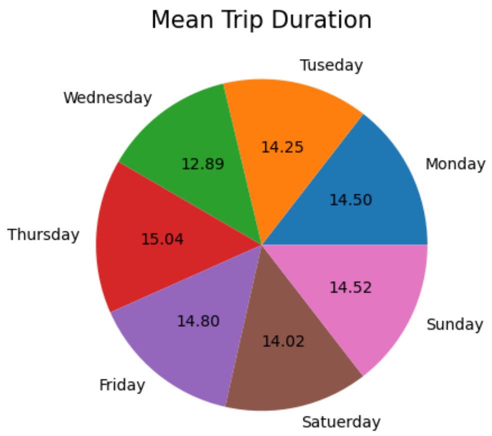
Discussion analysing your plot and the real world implications of the findings. This should be 100-200 words in length.



In this plot the x-axis is the mean amount and y-axis is the mean trip distance. The height in the plot suggest the average distance is high. And the x-coordinate suggest the amount is higher which is payment for taxi driver. For Wednesday, It have the highest trip distance and low total amount. Taxi driver should avoid working in Wednesday. And Thursday have the highest total amount and low trip distance, taxi driver should work more on Thursday based on the plot. However, the scale of mean trip distance varies too much. There is an increase of 14km average trip distance between Wednesday and Thursday. This is caused by abnormal data. There is an entry in Wednesday that have distance around 120000km which affects the mean.

Task7

a paragraph discussing your pie chart and discussing real world implications of the findings. This should be 50-100 words in length.



The graph shows the percentage of the mean trip duration for each day of the week. Thursday have the highest percentage and Wednesday have the lowest percentage. This suggest that in real world in Wednesday trip duration is generally lower than other day. And Thursday's trip duration is generally higher than other day in the week.