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.

- Passenger count

According to Passenger rules highlighted in NYC311 Taxis page, "the maximum number of passengers allowed in a taxicab by law is 5", with an exception for 1 additional passenger, if that passenger is under age 7 and is held on the lap of an adult passenger seated in the rear seat." This implies that any taxi trip with more than 6 passengers should be considered an abnormality in the dataset.

- Drop-off datetime

By logic, the datetime of drop-off must be after the initial pick-up time. A passenger must be "picked up" before being "dropped off" . Hence, a trip record that has the drop-off occur before pick-up should be considered abnormal.

Reference:
"Taxis · NYC311." *Portal.311.Nyc.gov*, portal.311.nyc.gov/article/?kanumber=KA-01245.

Task3
discussion commenting on/comparing your two boxplots and discussing real world implications of the findings. should be 100-150 words in length.

The distributions of the recorded fares in both the morning and afternoon period have slight positive skews with the means closer to quartile 1. The centres(median) are similar, with morning trips median being $17.16 and afternoon $17.80. The spread(IQR) was $14.14 for both morning and afternoon periods, with their quartiles being approximately equal. Many outliers are present in this data, with the maximum being $324.40 for morning trips, $349.60 for afternoon trips, and the minimums being both negative(potentially caused by entry errors or taxi credits)

While there are many similarities, more morning trips were recorded(13068 trips) than afternoon trips(12308 trips), with the morning trips costing less on average, potentially suggesting that the morning period hosts slightly more and shorter taxi trips than the afternoon(which cost less). Additionally, the afternoon period having a larger overall range(421.10>366.18) suggests that the afternoon trips(costs) are more variable in nature.

Task4

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

With 24.73% of taxi trips being weekend trips, the data implies that a far larger portion of the trips occur on weekdays. Although weekends only occupy 2 out of 7 days of the week, the percentage or weekend trips is lower than its proportion of the week (24.73%<28.57%). The low percentage can potentially be explained by the lack of commuting for work, school or other activities that typically happen during the 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 distributions of trip hours on weekdays have a negative skew towards later hours, with the highest frequency being between 16:00–20:00 and lowest between 0:00-7:00. Trip hours on weekends is irregularly shaped, with the same highest frequency period as weekdays and the lowest between 7:00-9:00. Both distributions are centred between 12:00-16:00.

The relative frequency (in comparison with other time periods) of taxi trips at 20:00-7:00 saw a considerable increase on weekends, while 7:00-9:00 trips decreased dramatically. Although the total frequency was similar between weekends and weekdays at 0:00-7:00, given the small proportion of weekend trips highlighted in Task4, this can be seen as a significant relative increase. This trend is understandable since a large portion of NYC residents(with full-time positions) wouldn't need to commute to work or schools on weekends. Instead, more are likely to partake in the nightlife or leisure activities without a morning commute, which explains this relative increase in later hour taxi trips.

Similarly, hours close to the morning commute(9:00-12:00) had relatively less taxi trips on the weekends, while typical weekday working hours(9:00-5:00), where people would be occupied on weekdays had less taxi trips on weekdays.

Task6

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

The scatter plot shows an irregular distribution of mean trip fare and distance between the days of the week. To summarise the main findings:

- Trip fares are on average most expensive on Thursdays, followed by Fridays and Tuesdays
- Trip distance is on average the furthest on Wednesdays, followed by Sundays and Fridays

- Mondays on average have the lowest trip distance and cost, followed by Saturdays

According to “fare standards” highlighted in NYC311 Taxis, the taxi fare stays the same for any speed above 12mph, and is continuously charges by time (every 60 seconds) when under that speed.
The association between these two variables shouldn’t be assumed, however, the following speculations can be made in consideration with the fare standards:

- Thursdays and Tuesdays may have more traffic congestion, which causes a low trip distance and high costs.
- Wednesdays and Sundays may have less traffic congestion, which allows the taxi to drive faster and further, covering more distance with a lower fare.
- Destinations on Mondays and Saturdays may be closer in distance, which explains a lower mean cost and distance.
- Destinations on Fridays may typically be further, which also causes a higher cost.

Reference:
“Taxis · NYC311.” *Portal.311.Nyc.gov*, portal.311.nyc.gov/article/?kanumber=KA-01245.

Task7
a paragraph discussing your pie chart and discussing real world implications of the findings. This should be 50-100 words in length.

The pie chart shows taxi trips on Thursdays take up the largest proportion of “mean trip duration” in a week(15%), while Wednesdays take up the smallest proportion at 12.9%.

This suggests that taxi trips typically last the longest on Thursdays and is shortest on Wednesdays. The mean trip duration for Thursdays is approximately 19 minutes, and 16 minutes for Wednesdays.

However, trip duration alone without context isn’t enough for an accurate explanation. A longer trip duration can be caused by a variety of factors: heavy traffic, further destinations, route choice, weather conditions, passenger requests and more.

Additionally, from previous results we discovered Thursday trips costed the most on average, while Wednesday trips were cheaper and covered more distance. This could further suggest Thursdays having heavier traffic than Wednesdays.