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.

A particular abnormality to look for in the dataset would be negative values in the 'total_amount' column, which would represent negative prices being charged to Green Taxi passengers. This is nonsensical and would be important to distinguish in our dataset. The actual range would be expected to have a lower bound in the single digit dollars, since taxis generally have a flat fee for passengers, or possibly $0 if a full refund could be made.

Another abnormality to look for would be 'passenger_count' being a non-positive integer. If this column were to contain negative numbers or zero, it would alter our results in data preprocessing/processing and is also nonsensical since it would represent no passengers or negative passengers, however all trips would require at least one passenger to be present for a trip to have occurred. The implication that a taxi drove 0 or a negative number of passengers does not make sense, and the expected range should be from 1 to 4 or possibly 6, depending on what sort of cars (5 or 7 seaters) Green Taxi supplied in January.

Task3

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

Both morning and afternoon boxplots present a large number of outliers, resulting in dense plots that obfuscate the data presentation and make it difficult for conclusions to be made. The mean prices for morning trips appear to be slightly more expensive than afternoon trips, although it cannot be confirmed if this is a statistically significant difference. The boxplots have also not accounted for errors in the data, as both plots display negative values for the prices, which is not possible in the real world. The large number of outliers is an implication that the time the trip started is not necessarily the central or only variable that affects trip fares, since both morning and afternoon boxplots present many outliers and variability, and other things such as traffic conditions and trip distance should also be considered and were likely more influential on the trip price than trip start time.

Task4

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

It is expected that the percentage of trips in the dataset that were during the weekend would account for a smaller portion of the whole (12.24%), since a weekend covers 2 of 7 days of the

week. This also indicates that a higher volume of commute and work-related travel occur during the weekdays, and a much lower volume occurs during the weekend, thereby having less trips occur on the weekends.

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.

Weekday: the histogram appears to be more left-skewed, with more trips beginning in the afternoon than at any other time, namely between 16:00-20:00 (~17,500 trips). This is in line with real world events where many jobs finish in the afternoon and people are travelling home or elsewhere. Additionally, many trips occurred between 12:00-16:00 (second largest bin), which could be indicative of people travelling to other sites or buildings during the day as an errand for their job.

Weekend: the histogram's frequency peaks much lower than the weekend histogram, with no more than 3000 trips occurring in a single bin (as opposed to the weekday's ~17,500 maximum). The weekend could be considered bimodal, with a large trip volume occurring between 00:00-7:00 and 12:00-20:00, with the former likely being due to people going to evening and late-night social events and needing to catch the taxi back home, while the former is likely attributed to general weekend travels. However, the weekday histogram had a higher frequency in this bin than the weekend, but this may be attributed to the overall greater volume of trips across the weekdays, since it was the lowest of the weekday bins.

Task6

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

The scatter plot appears to have no very little to no observable trend between the average trip cost and average trip distance for each day of the week. However, the scatter plot does show which days appear to have generally higher prices for a similar trip length; Wednesdays appear to be much cheaper (~$21.60/16mi) than those on Mondays (~$21.60/2mi), which is approximately 8 times the price per mile. Fridays and Sundays also appear to have cheaper prices per mile, while Mondays, Tuesdays, Thursdays and Saturdays on average have more expensive trip fares per mile. The more expensive weekdays could be attributed to high demand during peak times\ and so Green Taxi charges more per mile to increase profits, but Wednesday and Friday would be outliers to this claim.

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 that Thursday had the greatest mean trip duration compared to other days of the week. This indicates that passengers tend to have longer trips compared to any of other day of the week, however, this may not be a statistically significant difference between other days' mean trip durations, which vary from 12.89% (Wednesday) to 15.04% (Thursday). However, it could suggest that within the month of January, traffic may have been higher, or

passengers had to travel further on Thursdays compared to other days, resulting in longer trip durations on average.