Task 2.5

Describe the insights that the five DDS measures can provide to a traffic   
data analyst in the context of this use case, without including the general   
definition of each measure. Instead, focus on what these measures reveal   
to a traffic analyst.

Range: As this measure indicates the spread of a given variable, in this case the speed on the 1083 site of Chester Road in Manchester, we can investigate the way drivers approach this segment. If we observe a large range, this may indicate that for some reason, users drive more dangerously or there are both high and low speed zones present. However, if a small range is observed, this may indicate a smoother traffic and general speed consistency between drivers. Overall, this measure in a way visualizes the “fluency” of the traffic. In the context of Traffic Volume, this indicates possible fluctuations in traffic density throughout a given period, which could be a result of inconsistent road usage or a specific event, such as road closure/obstruction. In general, we should strive to lower the range to keep the traffic flow consistent, which should result in less traffic jams and improved safety. This could be achieved through flow control means such as traffic lights.

1st Quartile: In the speed context, this is the speed below which 25% of all vehicles drive through the road segment. A low value would signify that a significant portion of drivers choose to drive at lower speeds. The reasons for that would include pedestrian-heavy or residential area, bad road surface maintenance, or any other possible obstructions. However, a high value would suggest that the road supports or encourages faster driving. In general, this suggests a trend in speeds less likely to be taken by a driver. In the context of Traffic Volume, this would further indicate consistency of either high or low road usage. Dependent on the where the road is placed (for example highway vs. school area), this can help an analyst pinpoint where and which measures should be used, or where traffic should be directed.

2nd Quartile: In other words, a value (median) that splits a given attribute into two halves. In the speed, this should be understood as the usual speed at which vehicles move through the segment. This value should be close to the actual speed limit present. If it is visibly lower, this may indicate frequent stopping or slowing, or traffic jams, and that the traffic should be made more relaxed. On the other hand, if it is much higher, then slowing means should be introduced, such as speed bumps. In the context of Traffic Volume, it is a good way of comparing different segments to visualize the “typical” car throughput, and analyze the direction trends which cars tend to follow.

3rd Quartile: In the speed context, this is the speed below which 75% of all vehicles drive at. This generally indicates the speed which a driver is very likely to be at or below. If this value is low, it indicates that drivers generally tend to drive slower at this segment. On the other hand, if this value is high, a significant portion of drivers move at higher speeds. In both cases, this could be concerning, and a point of either too high congestion or more reckless driving. In the context of Traffic Volume, it could signify the utilization of a road segment. If we observe a high value, this could indicate, for example, what the volume is during peak hours and how dense we expect the traffic to be. If we observe a low value, this might indicate less aggressive bursts, or low road utilization.

Interquartile Range (IQR): as the range between Q1 and Q3, this value visualizes the spread of 50% of the data. In the context of speed, this indicates the way drivers are “clustered” together, and how their speeds differ from one to another. A smaller IQR suggests that drivers tend to drive at similar speeds and share somewhat homogeneous driving behavior. On the other hand, a higher IQR suggests a larger spread in speeds, meaning some drivers tend to be noticeably faster than the others. In the context of Traffic Volume, it visualizes fluctuations within road utilization; smaller IQR suggests somewhat constant volume, and higher IQR suggests peaks and drops during a given period. This measure could be generally used as a way of optimizing flow control measures between and within lanes, as well as minimizing vehicle collisions.

In summary, these measures provide a set of useful information about both driver behavior and road utilization. Through thorough and careful analysis, necessary traffic interventions and other measures, such as infrastructure changes, can be utilized to introduce quality of life improvements for both drivers and pedestrians. It is also important to note that these measures should not be treated individually, but rather as a whole, to produce accurate and trustworthy observations about road traffic.

Task 2.6

Describe the relevant information or insights for this use case that the five   
DDS measures fail to provide to a traffic data analyst. Discuss the impact   
that the absence of this information has on the data analysis, specifically in   
the context of this use case, which involves analysing traffic volume and   
speed.

Although these measures provide great insight into the central tendencies of data, they inherently lack much information about outliers, which are crucial in understanding the complete picture of traffic behavior. In this context, these could represent anomalies such as roadworks or accidents. Such data can represent the likeliness of collisions, how road structure changes influence traffic, or the extremes of data density. If these are not identified and considered, we will in turn generate skewed interpretations which underfit or overfit given features. This will most likely lead to inadequate decisions.

By aggregating data, we generate a “general” or “typical” behavior of the road, but lack more detailed information, such as time-based patterns or unusual events. This might result in a “one solution fits all” approach, which will most likely fail at certain times of day. Hourly data is crucial for appropriate and adequate measures, in forms of for example unique traffic light patterns, to be applied at given times of the day. These measures also do not consider any mitigation strategies during events such as accidents or rush-hours, which results in inefficiencies when traffic situation is not close to its usual behavior.

Due to inherent static nature of these measures, we lack sequential information about the data; for example, how behavior of given hours/days affect the next, or how a speed trend changes with time, which is extremely crucial in traffic analysis, as its behavior is inherently sequential. This information helps us make more dynamic systems, which can adjust and reorganize its system according to similarly dynamic traffic behavior whenever needed. Understanding such sequences is of utmost importance to analysts, as omitting such measures when making decisions might lead to more aggressive traffic peaks, faster road degradation, under and over shooting volumes for different weekdays, or even increased accident rates due to inappropriate flow control systems.

As standalone values, these measures lack possible correlation between each other, such as how traffic volume might correlate to resultant vehicle speeds. Such correlations help analysts understand how given variables affect others. As traffic behavior might be much more complex than simple one-to-one relations between attributes, correctly identifying and utilizing appropriate correlation is crucial for adequate decision-making. If we disregard that information, we might negatively affect the problem we are trying to solve through accidental increase or decrease of an unexpected attribute. Through correlation usage, we might also introduce systems that are rather proactive then reactive.

Moreover, DDS measures lack consideration of external factors, which in the context of traffic analysis could be weather conditions, nearby accidents, special occasions (parades, etc.), or road maintenance works. Without this contextual data, we inherently lack actual understanding of what and why a certain phenomenon is happening. For example, if we do not account weather into our analysis, we might introduce inadequate flow control measures (for example, not increasing street light frequency after rain), which might lead to inefficiencies at best, and life-threatening situations at worst. However, if external factors are considered, the resultant systems become more adaptive and resilient to such events.

Without utilizing other measurements, we also do not capture cyclic patterns, such as seasonal, or even weekday, changes. Through understanding of cycles, we can much better optimize and understand traffic patterns, which results in better resource allocation and management. Such information also improves our actual representation of real world through data and helps better prepare for expected events, such as holidays.

In summary, DDS measures provide a great foundation-level understanding of what might be happening. However, lacking versatility and complexity, they should be utilized with other, more advanced measures and strategies. To be more representative, they should also utilize additional information outside their scope. Applying only insights from DDS might lead to a fragmented representation of real-world problems, which will almost always result in ill-suited solutions.

Task 2.7

Describe the conclusions/insights that can be drawn from the profile of the   
target road fragment. For this, consider commenting, for example, about   
speed and traffic volume for the North and South lanes

First, we focus on the north lanes. We can see that each lane has a visible number of outliers on both sides of the spectrum, but the low-end is noticeably more condensed, as seen in Figure 1. This indicates that an overwhelming number of drivers decide to drive slowly on the north segment of the road within given hours, independent of the lane. As specified in the coursework description, this is most likely due to the morning traffic jam, where the overall traffic flow is substantially affected by the sheer volume of drivers, which suggests business concentration in that direction. Surprisingly, some high-end outliers land well over the median, but these might either reading errors or data taken after the initial traffic peak, or even drivers speeding intentionally to leave the jam at last moments. Medians, minima and maxima are identical, which indicates that no matter the lane, the expected “general” driver and traffic behavior throughout this time remains the same.

When we focus on the south lanes, we see that, although still visible, the number of outliers is much smaller than of the north lanes, as seen in Figure 2. From that same figure we also notice that data is a bit more condensed, and the outliers a bit less extreme, which might indicate a less aggressive traffic condensation and peak, and a more generalized driver/traffic behavior. We also notice a comparatively higher median for all south lanes, which, when combined with higher condensation, might indicate a more relaxed and predictive behavior, with less frequent stopping and more fluent traffic flow, which might be due to drivers leaving the metropolitan area. Similarly to the north lanes, we do not see many discrepancies between each lane, which again indicates that the “typical” traffic flow remains similar across all south lanes.

When we look at Figure 3, we can see that north lanes experience a significantly more intense traffic volume than south lanes. Here, we can also see that although driver behavior might remain similar between lanes, the NB\_NS lane is prone to have gentler traffic volume, with a visibly lower median. Two other lanes, NB\_MID and NB\_OS, share a similar median, but due to much larger range, the NB\_OS lane is expected to have its traffic flow more chaotic and disruptive, which might in turn affect other lanes.

On the other hand, each south lane behaves very different to another, as they share no common features. As visible in Figure 3, SB\_MID has the middle median value, with the smallest range and maxima/minima, which indicates that although it is not the least utilized one, it experiences the smoothest traffic behavior of all three. SB\_NS lane seems to have one outlier, which might indicate that during data gathering period there was a week where this lane was somehow affected by outside circumstances which lowered the registered traffic volume significantly.

When we combine information retrieved from Figures 1, 2, and 3, we can see a strong correlation between traffic volume and resultant speeds. We see that with a higher traffic volume, the general (median) vehicle speed drops visibly. Additionally, higher traffic volumes seem to negatively affect the fluency of the traffic flow, and make the behavior much more chaotic and extreme, which might in turn lead to higher accident rates. This in turn calls for better flow control management during these hours to ease the intensity of the morning peak, through for example, better traffic light management, adding more lanes, or enhanced public transport.

Interestingly, we notice that although traffic volumes differ between both north and south lanes, the general driving behavior remains consistent. This might indicate that the volume alone does not play the only role in general behavior, and that other unseen features, such as road infrastructure or layout, play a crucial role for this segment. Additionally, the driving consistency between different traffic volume shows that drivers respond irrespective of the volume on given segments (north, south), which further showcases the importance of comprehensive road design.

Task 3.3

From the results you obtained in Task 3.I, what were you able to observe   
(feel free to consider any aspects associated with the obtained   
profile/results, profiling techniques, etc.)? To provide an answer to this   
question, please, proceed to the Blackboard Test “60711-Lab1-S-CW1”.   
Follow the instructions provided in the test to answer this task. Completing   
this will enable you to receive a mark for this task.

From Figure 4, we can see that the resultant box plots for Friday and Tuesday are very similar. Although Tuesday seems to have one outlier, all other data points have been placed within the maxima/minima limits. We see that for Tuesday, the boxplot has a smaller median and smaller IQR range, which might indicate that, when compared to Friday, it has on average less intense and more condensed traffic volume density throughout the day. For Friday, in contrary, one might expect a higher traffic volume for longer periods throughout the day.

When we look at Figure 5, we can see two distinct modes for both Tuesday and Friday, one around 7am, and the other around 4-5pm. These peaks indicate rush hours and are to be expected, as the data consists of average north and south lane traffic data combined. For both weekdays, we see a sharp decrease after 8am, a somewhat constant and more-relaxed period between 10am-2pm, with a sharp increase at 3-4pm, with a gradual decrease after 5pm. This is also expected, as these hours correspond to morning work commute and evening route back home, which results in a left-skewed histogram, which further showcases that comparably, only few people travel late in the evening and in the night.

Figure 5, combined with Figure 4, showcase a classic cyclical behavior for urban areas, as both days share similar characteristics across volume traffic data, which correspond to phenomena such as usual office hours (9am-5pm), school operating hours, or even public transport schedules. The corresponding evening and mid-day dips suggest an opportunity in terms of road maintenance and other related works, as the regular traffic would not be that disrupted.

Again, when we look at Figure 5, we see that Tuesday, being in the middle of work week, has higher traffic volume during morning and evening rushes, with more intense jumps. Whereas Friday, although similar in nature, has an overall gentler slope, which suggests a more homogenous traffic throughout the whole day, although only slightly. This gives us a deeper understanding of the mechanisms at hand and emphasizes the need for better traffic management during these peak rush hours. Moreover, Friday’s slightly more regular traffic and gradual slopes hint towards events such as early departures or weekend plans.

Furthermore, we should not forget that the way data was gathered could have also influenced the visible results and observations. As we do not know much about the way it was obtained, we can only speculate about its accuracy. For example, depending on the sensor reading frequency, or even their placement on specific road segments, we might miss crucial readings regarding all relevant attributes, including traffic volume. We also have no way of further analyzing the outliers, which could have provided more insight about traffic behavior. We could also add and analyze other data alongside traffic, such weather conditions, nearby events, or public transport schedules, as we are currently working “in a vacuum” of only traffic data with no additional context.

Moreover, it is worth remembering that these conclusions and insights should be validated against more data from other weekdays and months, or even years, especially when it comes to its cyclic nature. Such validation would also allow for better capturing of outliers (such as public holidays or events) and would further help in both removing possible bias in data and making the findings more robust. Furthermore, it is more than likely that these findings and spreads will not be applicable during the weekend, and even possibly Mondays, as these cases differ extremely from the analyzed weekdays due to their inherent more relaxed nature.

Task 4.1

You have been asked to develop the previous tasks using the Python   
programming language, which is one of the most popular languages for tabular   
data processing, adopted by programmers worldwide and in companies such as   
Netflix, Google and Amazon. However, not all data analysts possess programming   
skills. For those, alternative means are at disposal, including technology that have   
existed for decades, such as Database Management Systems and query   
languages, such as MySQL and SQL, as well as others that have more recently   
been developed, including OpenRefine and Knime, which offer Graphical User   
Interfaces to facilitate data profiling and preparation.   
   
Choosing a technology other than Python (from the suggested in Section 4),   
repeat Task\_3.I, calculating the average traffic volume per hour of the day for   
Tuesday and, separately, the same for Friday. Again, for simplification purposes,   
consider only the times of the day between 07:00:00 to 23:59:59, ignoring vehicles   
detected between 00:00:00 and 06:59:59, and make no distinctions between North   
and South lanes.   
   
I. Make sure you upload into your Gitlab COMP60711\_Part\_1 project the   
step-by-step code/recipe you developed using the technology of your   
choice. Please, name it ‘2ndSolutionForTask3\_I’.   
   
Also, describe in the “60711-Lab1-S-CW1” Blackboard Test text field the   
features of the technology you used, emphasising the ones that have made   
your work easier and the ones that made your work more difficult, in   
contrast to Python. In your description, you can make comments about   
functionality that is not offered within the technology, rendering you unable   
to satisfactorily complete Task 3.I.

As I have previous experience with DBMSs, I have decided to use MySQL from the offered choices. As both Python (Pandas, NumPy, etc.) and MySQL are very much associated with the field of data analysis, I believed transitioning from one environment to another would not require that much additional effort, but there are some differences that cannot be omitted.

1) Although I have not worked with MySQL previously, I have worked with Postgres and SQLite, where all of them share the same core and virtually the same functionalities. Because of that, I was able to easily write the necessary queries for desired data manipulation. Thanks to that, these queries can be moved into any other SQL-based management system with basically no effort, which makes it extremely versatile. Moreover, due to MySQL’s date-related filters, it was much easier to segregate data by days of the weeks and hours with little to no additional processing, whereas Python required, although not much, more work associated with those parameters.

2) MySQL supports structures such as indexes, which significantly enhance query performance, at a cost of additional memory of course. This is particularly useful for large datasets, where data retrieval optimization is crucial. Furthermore, as SQL-DBMS were strictly designed for data handling, they are much better optimized for data-related tasks than Python libraries. This made the debugging and testing much faster, as the queries were executed within less than half a second. If the dataset chosen for this task was larger, we would quickly see the obvious performance issues on the Python side, which would have made the task much more tedious and exhausting.

3) MySQL enforces much stricter constraints than Python from the initialization, which always ensures complete consistency and reliability of data during analysis. This is especially visible when comparing operations related to the datetime format in Python and MySQL. Where Python allowed casting to different data types, which could have led to major inconsistencies when not careful, the same would never be allowed in MySQL. This made the task easier in the terms of that when attributes were initially set up, I never had to worry about datatype integrity again.

4) In my opinion, SQL’s aggregate functions simplify the calculations of summary statistics, such as averages, much better than Pandas, which made the relevant work both easier to accomplish and then later understand the logic behind. It is also a simpler and more readable code, which makes it easier to come back after a while optimize.

5) Although MySQL can allow for quite a lot of operations, it was designed strictly for data storage and retrieval, making it much less flexible for any task beyond its core. For example, the task of visualizing relevant data becomes very awkward when more complex logic is involved, whereas Python does not struggle much. Additionally, if the task at hand was more convoluted, it would require a lot of previous SQL expertise to be carried out both successfully and optimally, whereas Python would handle most of it by itself. This made the work much stricter and careful.

6) Setting up MySQL database and server proved to be much more exhausting and irritating than initially expected. On my operating system (MacOS), I had to tackle several issues regarding permissions, server configuration, file access and storage. This took more than two hours, which is almost 3x the time I spent on writing the actual queries and testing. Whereas in Python, all that had to be done was importing the required libraries and loading data. Due to this, the initial setup and configuration made starting the work much more difficult.

7) Optimizing MySQL queries requires much more work, especially when handing large datasets. It may require advanced configurations, complex memory structures, and experience. Whereas in Python, most handling and optimization is done “under the hood”, even when loading in the initial file. Because of that, I did not spend much time optimizing the MySQL database and wrote very simple queries.

We can clearly see that although MySQL is much more adequate for the tasks of data storage and retrieval, it falls short when it comes to versatility and more complex tasks. It is also much faster due to its core design, as well as allows for superior control at a cost of greater expertise and work. However, Python is much easier to setup and use immediately, without having to worry too much about the data structure. It also allows for tasks outside of the data handling scope. Of course, this comes at a cost of possibly losing consistency and errors when not careful with operations.

In summary, MySQL being very much data-oriented worked well for the task at hand and had all the features needed but could very easily fall short if the specifications were more complex.