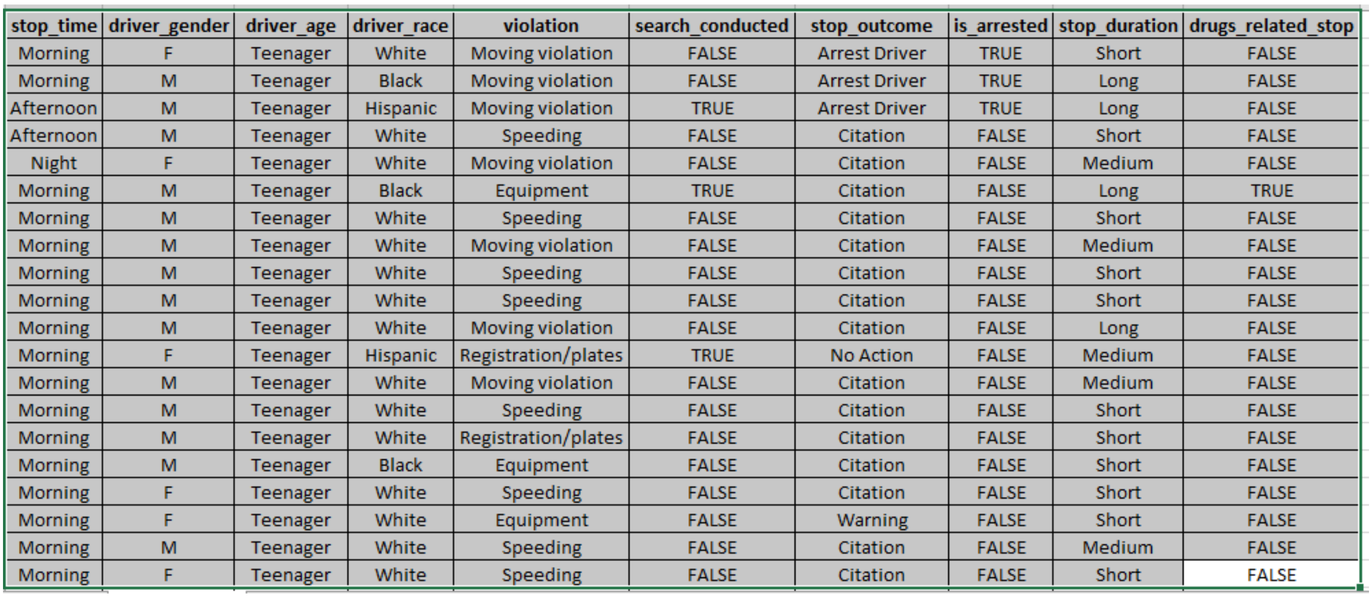
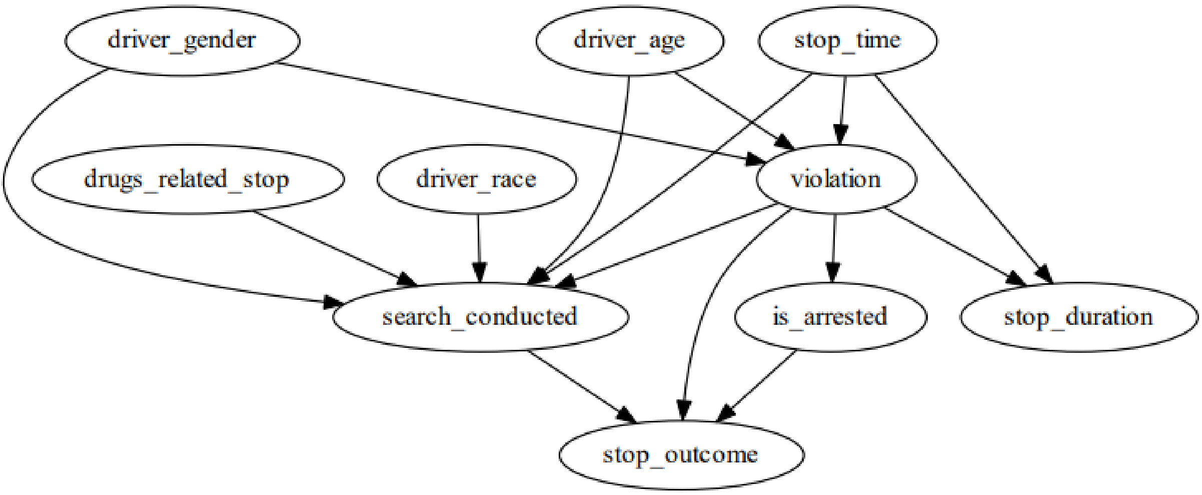
QUESTION 1:



stop_time	driver_gender	driver_age	driver_race	violation	search_conducted	stop_outcome	is_arrested	stop_duration	drugs_related_stop
Morning	F	Teenager	White	Moving violation	FALSE	Arrest Driver	TRUE	Short	FALSE
Morning	M	Teenager	Black	Moving violation	FALSE	Arrest Driver	TRUE	Long	FALSE
Afternoon	M	Teenager	Hispanic	Moving violation	TRUE	Arrest Driver	TRUE	Long	FALSE
Afternoon	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Night	F	Teenager	White	Moving violation	FALSE	Citation	FALSE	Medium	FALSE
Morning	M	Teenager	Black	Equipment	TRUE	Citation	FALSE	Long	TRUE
Morning	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Morning	M	Teenager	White	Moving violation	FALSE	Citation	FALSE	Medium	FALSE
Morning	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Morning	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Morning	M	Teenager	White	Moving violation	FALSE	Citation	FALSE	Long	FALSE
Morning	F	Teenager	Hispanic	Registration/plates	TRUE	No Action	FALSE	Medium	FALSE
Morning	M	Teenager	White	Moving violation	FALSE	Citation	FALSE	Medium	FALSE
Morning	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Morning	M	Teenager	White	Registration/plates	FALSE	Citation	FALSE	Short	FALSE
Morning	M	Teenager	Black	Equipment	FALSE	Citation	FALSE	Short	FALSE
Morning	F	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE
Morning	F	Teenager	White	Equipment	FALSE	Warning	FALSE	Short	FALSE
Morning	M	Teenager	White	Speeding	FALSE	Citation	FALSE	Medium	FALSE
Morning	F	Teenager	White	Speeding	FALSE	Citation	FALSE	Short	FALSE

The dataset is about the USA police officers work in stopping traffic for search and violations. It is sourced from Kaggle.com and available at https://www.kaggle.com/datasets/ibrahimelsayed182/stanford-open-policing. The reasons for choosing this dataset are that it has 10 relevant independent variables with 100000 samples suited for structural learning. Data pre-processing steps such as filling in missing blank values, categorising continuous variables such as Age, Time into discrete values are undertaken. This dataset with high dimensionality will enable the algorithms to learn accurate information about the causal relationships between variables. In USA, there are accusations of bias towards certain ethnic groups such as black community, gender, and misusing of policing power towards them. Inferring this causal structure through Structural learning will help in countering the assumed facts, analysing it, predicting certain outcomes with conclusive evidence for rectifying current system's risks and fixing it using new policy recommendations.

QUESTION 2:



The knowledge-based DAG is generated primarily with the help of information from the literature and a bit of personal knowledge. The literature [1] shows that Black drivers were stopped less after the evening hours when there is little visible sunlight for distant racial identification and time of stop during the day is taken as a parameter. We have accounted drug related stops as this study indicates that black and Hispanic drivers were stopped more for merely possessing drugs even after legalizing use of it in the United States [2].

Adding to it, the outcome of these stops such as duration of the stop/search and data about the arrested drivers with race/gender information are considered from the literature [3]. Furthermore, there are also studies [4] indicating women were stopped more likely during these searches showing necessity of gender taken into our learning. Drivers' race, age gender and possession of drugs are considered as one of few reasons for traffic stops and searching during policing. Relationships between these factors are identified for this structural learning and the graph is produced with information to determine the possibility of someone getting searched/arrested for actual violation or simply because of their race or gender.

1. https://journals.sagepub.com/doi/pdf/10.1177/09567976211051272
2. https://www.nature.com/articles/s41562-020-0858-1
3. https://onlinelibrary.wiley.com/doi/10.1111/j.1745-9125.2012.00285.x
4. https://www.cambridge.org/core/journals/journal-of-race-ethnicity-and-politics/article/at-the-intersection-race-gender-and-discretion-in-police-traffic-stop-outcomes/2C7329D70147351D843E50B4C1A47CDE

QUESTION 3:

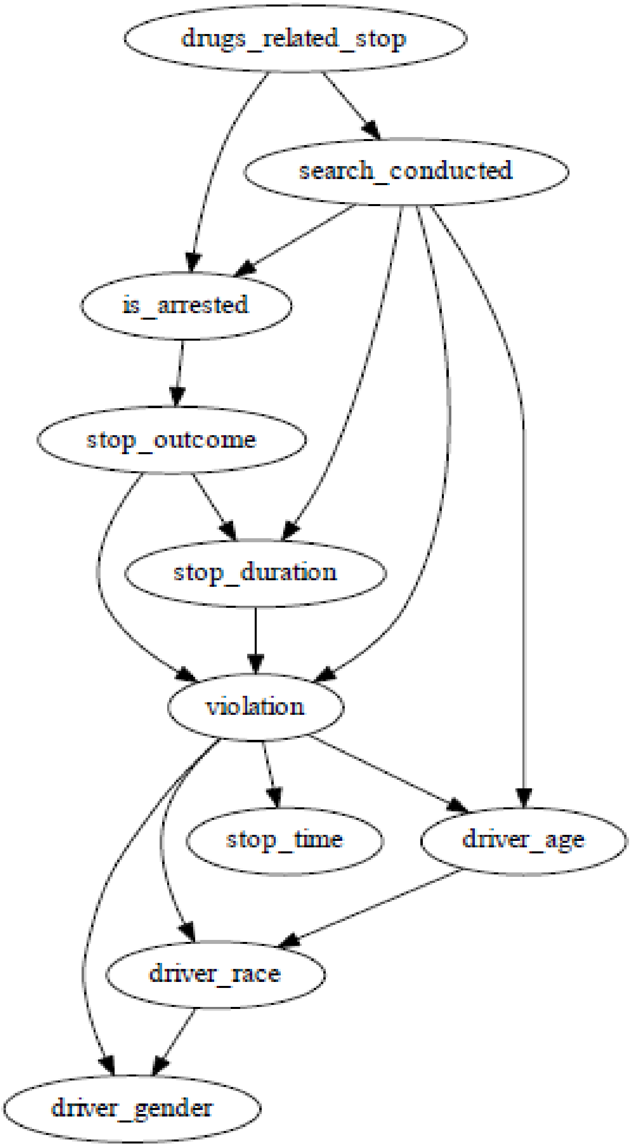
Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	#free parameters	Structure learning elapsed time
	BSF	SHD	F1				
HC	0.067	17.500	0.290	-752088.044	-755376.019	401	2 sec
TABU	0.067	17.500	0.290	-752088.044	-755376.019	401	3 sec
SaiyanH	0.133	15.000	0.308	-755945.889	-757249.599	159	11 sec
MAHC	0.067	17.500	0.300	-752371.869	-754954.692	315	9 sec
GES	0.067	17.500	0.290	-752088.044	-755376.019	401	3 sec

Table Q3. The scores of the five algorithms when applied to your data set.

For my dataset BSF, SHD and F1 scores of these algorithms are low compared to the average score estimated for different data in the manual. Structural learning algorithm performance depend on finding the causal relationships between variables and it is expected to be low when dealing with limited data sample size. F1 score depends on precision/ recall factors for determining the causal model accurately and it ranges between 0 and 1. The low range scores around 0.3 indicates a possibility of having missed several edges in the learned graph when compared to the true graph. There is a difference between predicted true positives and actual positives in these causal relationships.

As Balanced Score Function BSF indicates the accuracy of Bayesian network model based on similarities between learned graph and the true graph, and it wasn't quite similar in this case resulting in low scores. Structural hamming distance SHD score is influenced by the presence and direction of edges between the nodes in the model. When these relationships between the variables are non-linear or complex and it often tends to result in low scores while determining the causal structure. The noise and the inconsistency of some the data in the dataset also results in low scores. In addition, the synthetic data tends to have high scores as these algorithms perform better with it, than the real dataset that we have used in our case.

QUESTION 4:



All three causal classes are present in the CPDAG generated by the HC algorithm.

Common Cause Example:

search_conducted <- drugs_related_stop -> is_arrested

In this case, drugs_related_stop is the common cause for being search_conducted and getting is_arrested. Here, search_conducted and is_arrested are independent given there is a drugs_related_stop, but search_conducted gives information about is_arrested.

Common Effect Example:

search_conducted -> violation <-stop_outcome

In this scenario, getting search_conducted and charged during a stop_outcome results in a violation. Therefore, search_conducted and stop_outcome are independent, but they become dependent when there is a violation.

Common Chain Example:

drugs_related_stop-> search_conducted -> is_arrested

In this chain, drugs_related_stop has an influence on getting is_arrested through a search_conducted. However, drugs_related_stop and is_arrested are independent given a search being conducted.

QUESTION 5:

	Your rankings			**Rankings according to the Bayesys manual**		
Rank	**BSF[single score]**	**SHD[single score]**	**F1[single Score]**	**BSF [average score]**	**SHD [average score]**	**F1[average Score]**
1	SaiyanH[0.133]	SaiyanH[15.000]	SaiyanH[0.308]	SaiyanH[0.559]	MAHC[50.96]	SaiyanH[0.628]
2	MAHC[0.067]	MAHC [17.500]	MAHC[0.300]	GES [0.506]	SaiyanH[57.98]	MAHC [0.579]
3	GES[0.067]	HC [17.500]	GES[0.290]	MAHC [0.503]	HC [62.36]	GES [0.552]
4	TABU[0.067]	TABU [17.500]	TABU[0.290]	TABU [0.499]	TABU [62.63]	TABU [0.549]
5	HC[0.067]	GES [17.500]	HC[0.290]	HC [0.498]	GES [63.3]	HC [0.548]

Table Q5. Rankings of the algorithms based on your data set, versus ranking of the algorithms based.

Yes, the results are as expected and the rankings that we got are relatively consistent with the rankings according to the Bayesys manual. A parallel comparison of our network with four of the networks (of total 6) analysed in the manual have reasonably similar complexity with less than 40 nodes, small number of edges and real datasets with comparable sizes. The BSF, SHD, F1 scores of algorithms such as HC, TABU, GES are similar rendering a similar fully connected learnt graphs whereas MAHC performance were slightly better than these as it models continuously in search for a better graph structure.

Overall, a high Balanced Scoring Function (BSF) and F1 score of SaiyanH indicates a better accuracy match between the learned SaiyanH graph and true graph. A low Structural Hamming Distance SHD score (15.0) of SaiyanH proves that it has the lowest error rate and there was a smaller number of changes in the edges required during this modelling when compared of other models. Using real large sized datasets, an increase in the number of variables with more edges has a minimum effect on the performance of SaiyanH but affects HC, TABU, GES models considerably.

QUESTION 6:

Structural learning runtimes of all algorithms for my dataset is relatively consistent to the average runtime results shown in the Bayesys manual. Runtime comparison between my dataset and manual indicates that SaiyanH and MAHC had a longer runtime due to its phase level learning whereas HC, TABU and GES had shorter runtime, especially Hill climbing HC learnt the relationships quickly. Although we have taken a large sized dataset, the runtimes were ideal as we had only 10 nodes and 15 arcs in our graph. When the graph structure is more complex with large number of nodes and edges, the runtime increases considerably.

QUESTION 7:

Algorithm	**Your Task 3 results**			**Algorithm**	**Your Task 4 results**		
Your knowledge-based graph	**BIC score**	**Log-Likelihood**	**Free parameters**		**BIC score**	**Log-Likelihood**	**Free parameter**
	-782136.432	-763556.504	2266	**HC**	-755376.019	-752088.044	401
				TABU	-755376.019	-752088.044	401
				SaiyanH	-757249.599	-755945.889	159
				MAHC	-754954.692	-752371.869	315
				GES	-755376.019	-752088.044	401

Table Q7. The BIC scores, Log-Likelihood (LL) scores, and number of free parameters generated by each of the five algorithms during Task 3 and Task 4.

Yes, the results are as expected because all these algorithms from task 4 results have fit the data better compared to the Task 3 where Bayesys evaluated just based on knowledge-based graph. Log-Likelihood (LL) score shows if the model had fit the data well and all algorithms LL scores around -752088.044 have moved closer (less negative) to the positive value than the knowledge-based graph LL score (-763556.504).

Bayesian Information Criterion (BIC) scores of all five algorithms are lower indicating that they modelled better with the data dimensionality than knowledge-based graph.

Free parameters render the complexity of the model graph. The knowledge-based graph has FP score of 2266 indicating the knowledge-based graph is not accurate representation of data, needs more FP to fit the data and may result in overfitting. All five algorithms have comparatively less free parameters scores proving their graphs are better representation of data, especially SaiyanH with 159 FP score (and better LL/BIC scores) is less likely to overfit and can provide a good balance between right model fit of the data and ideal complexity. These results signify that structurally learnt model graphs are accurate representation than the knowledge-based graph.

<u>QUESTION 8:</u>

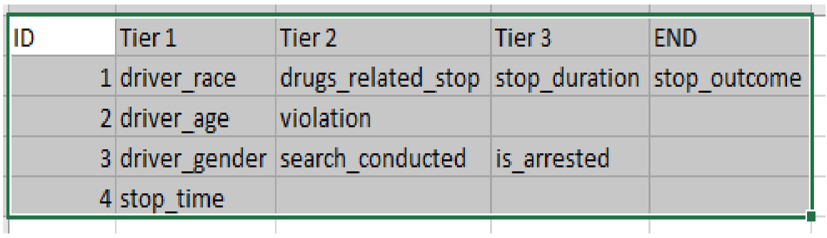
Knowledge approach	CPDAG scores			LL	BIC	Free parameters	Number of Edges	Runtime
	BSF	SHD	F1					
Without knowledge	0.067	17.500	0.290	-752088.044	-755376.019	401	16	2 sec
With knowledge – Target Node, r=8	0.200	14.000	0.394	-751408.615	-754040.635	321	19	3 sec
With knowledge –Temporal Order,3 tiers	0.333	13.000	0.552	-751296.909	-761945.659	445	14	3 sec

Table Q8. The scores of HC applied to your data, with and without knowledge.

Yes, the results are as expected after incorporating two knowledge constraint Target nodes and Temporal order approaches which has reduced the search-space of the graphs considerably. As noted from CPDAG scores above, there is improvement in the scores obtained from the models after incorporating knowledge as following. The BSF was near 0 without knowledge, but it is maximised about 5 times to 0.333. The F1 score of 0.552(About 100% increase) and SHD score has lowered to 13.000 resulting in more accurate causal graph since there is a decrease in error rate (about 25%).

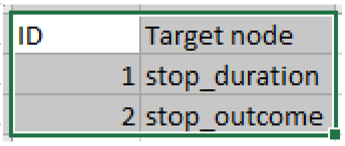
Log-Likelihood (LL) score has become more positive value than the earlier and there is a decrease in BIC scores after more edges are added in the graphs. These results reiterate that knowledge constraints have helped in increasing the performance and efficiency of the algorithm by reducing the search space as data has modelled closer to the true graph than it did without constraint incorporation.

Having said there is search-space reduction, the constraints have increased the runtime with more graphs processing, and it tends to work better with big data than limited data by restricting more number of edges resulting in better runtime.



ID	Tier 1	Tier 2	Tier 3	END
1	driver_race	drugs_related_stop	stop_duration	stop_outcome
2	driver_age	violation		
3	driver_gender	search_conducted	is_arrested	
4	stop_time			

constraintsTemporal



ID	Target node
1	stop_duration
2	stop_outcome

constraintsTarget

For Temporal order, the algorithm has considered variables given in different tiers by making sure final event (say stop_outcome or is_arrested) occurs after initial event (say search_conducted), thus final event cannot influence initial event and run the model by restricting edges of the nodes from same tier. For Target nodes, there is a decrease in the number of free parameters to 321 (using parameter r) and more causal relationships are produced for the target variables increasing the edges count to 19. The parameter r has been tuned to several ranges and finally optimized to r=8 which helped diminish the free parameters with a penalty.