TORONTO BUS DELAYS 2022: PROJECT INFORMATION

DATA: Toronto 2022 Bus Delays

- **Data Source:** This dataset was obtained from the Toronto Open Data Catalogue. It is data collected by the Toronto Transit Commission, a government organization in charge of public transit, and made available to the public. This data source is trustworthy because it is official data.

- **Data Collection:** Each bus is fitted with technology that tracks the location of the bus and the arrival times, it is not clear how the cause of delay is collected but I assume it is reported by the employees and recorded.

- **Contents:** This dataset contains information on TTC bus delays from Jan 2021 to June 2022. Each column and its description is included in the table below.

Column	**Description**
Date	The date (YYYY/MM/DD) when the delay-causing incident occurred
Route	The number of the bus route

Time	The time (hh:mm:ss AM/PM) when the delay-causing incident occurred
Day	The name of the day
Location	The location of the delay-causing incident
Incident	The description of the delay-causing incident
Min Delay	The delay, in minutes, to the schedule for the following bus
Min Gap	The total scheduled time, in minutes, from the bus ahead of the following bus
Direction	The direction of the bus route where B,b or BW indicates both ways. (On an east west route, it includes both east and
Vehicle	Vehicle Number

- **Limitations & Ethics:** Data is collected regularly with no time lag. There is little room for bias since most of the data is collected automatically, the only errors would occur if the technology malfunctions or a wrong cause of delay was somehow reported. There can be no unethical collection or use of this data since there is no private information and all the information is that of public transport vehicles.

- **Selection reasoning:** This dataset was selected because it worked well with my project goals, and I found it interesting to explore.

DATA PROFILE

Column Name	Data Type	Python Data Type	Wrangling procedures
Date	Structured, Qualitative (Ordinal)	object	-
Route	Structured, Qualitative (Ordinal), Time-invariant	Mixed to Object	Data types were mixed, changed to object
Time	Structured, Quantitative (continuous)	object	-
Day	Structured, Qualitative (ordinal)	object	-
Location-stop_name	Structured, Qualitative (Nominal), time-invariant	object	-Column name changed to stop_name
Incident	Unstructured, Qualitative (nominal), time-invariant	object	-
Min Delay	Structured, Quantitative (continuous)	int64	-
Min Gap	Structured, Quantitative (continuous)	int64	-

Direction	Structured, Qualitative (nominal), time-invariant	Mixed to object	-Data types were mixed, changed to object
Vehicle-Fleet	Structured, Qualitative (nominal)	int64 to object	-Column name changed to Fleet for clarity. -Data type changed from integer to object

Raw Dataset contains 10 columns and 75698 rows

	Min Delay	**Min Gap**
count	75698	75698
mean	19.337737	31.922389
std	45.411380	46.667887
min	0	0
25%	8	16
50%	11	22
75%	20	38
max	999	999

- **Duplicates:** Raw dataset contains 407 duplicate entries, a copy of the dataset without duplicates was created, and duplicate entries were exported to csv.

Column	Structural Errors	Missing values	Treatment
Date		-	
Route	119 routes do not exist and are put in wrongly	39	Missing values random, dropped Incorrect routes were deleted
Time		-	
Day		-	
Location	-Almost all locations entered wrongly, most were corrected -1936 rows were entered wrongly and could not be corrected		Created copy of dataset without incorrect rows
Incident		-	
Min Delay		-	
Min Gap		-	
Direction	Random non directional values inserted	13,877 entries missing	-column dropped
Vehicle		No bus fleet named fleet 0, rows assumed missing	-nothing was done to values -the fleet number is not very useful

			but other columns are

Wrangled and Cleaned delays dataset contains 72,831 rows and 9 columns