COMP47350 Data Analytics Homework I

Name **Anil Vercruysse**
Student ID **22202474**

DATA QUALITY PLAN

Feature	Data Quality Issue	Handling Strategy	Justification and Alternatives
Missing Values	Inconsistent value entry for missing values (either ‘Missing’, ‘Unknown’, or a blank entry).	Replace every ‘Unknown’ or blank entry, where they correspond to missing values, to the consistent entry ‘NaN’.	Justification: Standardising missing values ensures consistency, facilitates analysis, and avoids potential misinterpretation. Alternative: leave as is, but this may lead to inconsistencies during analysis.
CPSI and COI (int64)	8,852 and 10,575 Missing values, respectively	Leave as is.	Justification: these features could provide insights into the relationships between reporting/testing and death. Alternative: drop the rows, but too many. Impute the missing values, but too many are missing.
CPSI (int64)	Negative values do not make logical sense	Replacement of negative values with ‘NaN’.	Justification: Replacing negatives with 'NaN' avoids introducing illogical data. Alternative: Change negative values to positive, but this assumes sign errors without confirmation.
COI (int64)	Negative values do not make logical sense	Replacement of negative values with ‘NaN’.	Justification: Replacing negatives with 'NaN' avoids introducing illogical data. Alternative: Change negative values to positive, but this assumes sign errors without confirmation.
CPSII (int64)	Outliers in the data	No measures taken as these might provide insights into extreme CPSI cases.	Justification: Outliers may represent genuine real-world scenarios and provide valuable information. Alternative: Remove or transform outliers, but this may result in loss of information.
COI (int64)	Outliers in the data	No measures taken as these might provide insights into extreme COI cases.	Justification: Outliers may represent genuine real-world scenarios and provide valuable information.

			Alternative: Remove or transform outliers, but this may result in loss of information.
State of residence (category)	1 Missing value	Leave as is.	Justification: Immaterial missing values have a negligible impact on analysis. Alternative: removal of the row, but the result is immaterial either way.
County of residence (category)	1,186 Missing values	Impute missing values based on the non-missing county feature distribution, but ensuring the state feature matches the new attribution.	Justification: Since about 6.3% of cases are missing, imputation is appropriate to improve data quality. Alternative: their removal was considered, but this would negatively impact sample size.
County of residence (category)	High cardinality of 868 compared to other categorical features	Leave as is.	Justification: the high cardinality is not due to a formatting issue (all county data in the same format). Alternative: Reformat the feature, but this is unnecessary.
Age Group (category)	29 Missing values	Removal of rows with missing age group information.	Justification: this feature is important for death prediction and the removal of the rows is immaterial. Alternative: impute the values but the low materiality makes this inefficient.
Sex (category)	398 Missing values	Impute missing values based on the non-missing sex feature distribution.	Justification: Since about 2.1% of cases are missing, imputation is appropriate to improve data quality. Alternative: their removal was considered, but this would negatively impact sample size.
Race (category)	4,534 Missing values	Impute missing values based on the non-missing race feature distribution.	Justification: Since about 24% of cases are missing, imputation is appropriate to improve data quality. Alternative: their removal was considered, but this would negatively impact sample size.
Ethnicity (category)	5,911 Missing values	Impute missing values based on the non-missing ethnicity feature distribution.	Justification: Since about 31.3% of cases are missing, imputation is appropriate to improve data quality. Alternative: their removal was considered, but this would negatively impact sample size.
Case identification process	17,251 Missing values and irrelevant for death prediction	Dropping this feature.	Justification: dropping is recommended as it is vastly incomplete and irrelevant for the death prediction analysis. Alternative: retain and impute the rows, but there is very

(category)			limited death prediction relation.
Exposure status (category)	16,996 Missing values and irrelevant for death prediction	Dropping this feature.	Justification: dropping is recommended as it is vastly incomplete and irrelevant for the death prediction analysis (exposure is implied with a positive case).. Alternative: retain and impute the rows, but there is no death prediction relation to explore.
Symptom status (category)	9,739 Missing values	Dropping this feature.	Justification: 50% of the data is missing so it is difficult to apply in a correlation analysis. Symptom status feature is changeable and therefore cannot be effectively captured reliably. Alternative: impute the values, but this is not recommended when +30% of data is missing. Keep as is, but the limited values are unreliable.
Hospitalisation outcome (category)	6,281 Missing values	Impute missing values based on the non-missing hospitalisation outcome feature distribution.	Justification: Since about 33.25% of cases are missing, imputation is appropriate, especially since the feature is very relevant. Alternative: their removal was considered, but this would negatively impact sample size and result in the loss of relevant data.
ICU outcome (category)	17,216 Missing values	Dropping this feature.	Justification: 91.15% of values are missing. Alternative: Drop the relevant rows or impute their value, but the missing rate is too high.
Underlying conditions (category)	17,237 Missing values	Dropping this feature.	Justification: 91.26% of values are missing. Alternative: Drop the relevant rows or impute their value, but the missing rate is too high.