Task 1:

Question 1

A. We trimmed whitespace with quotes. We did this to ultimately standardize the categorical strings in order to transform them into factors. Subsequently, this allows for one hot encoding those columns for input into the models. Without the factor columns and we could not use those columns either in the t-SNE or the actual classification algorithms.
B. We filled null values with means for numeric columns and modes for factor columns. Using mode and median for the respective columns allows us to capture the central tendency of the data. This avoids unnecessary complexity and artificial values. More importantly this preprocess is necessary in order for the models to run in r code.

Question 2

The t-SNE plot shows substantial overlap between good and bad credit classes. Therefor there are limitations on the accuracy of the models before running them. It wont be easy for this models to achieve a high accuracy score. Running the t-SNE on multiple perplexity values, at 50, there is several more clearly discernable cluster of bad and good credit.

Task 2:

Part 2

Decision Tree — Rule Summary Table

Rule	Outcome	Key Conditions
1	0.14	No checking; duration ≥ 48; no known savings
2	0.25	No checking; duration < 23; no known property; non-critical credit history
3	0.29	No checking; duration 23–48; no known savings; credit ≥ 8,098; used car
4	0.30	No checking; duration 12–23; no known property; credit < 1,286; no other parties; installment ≥ 3.5
5	0.39	No checking; duration 23–48; no known savings; not used car; checking < 200
6	0.42	No checking; duration ≥ 23; known savings; checking not in [0,200)
7	0.59	No checking; duration 12–23; no known property; credit < 1,286; no other parties; installment < 3.5
8	0.72	No checking; duration 12–23; no known property; credit ≥ 1,286
9	0.73	No checking; duration < 23; no known property; critical credit history
10	0.80	No checking; duration 23–48; no known savings; not used car; checking ≥ 200

Rule	Outcome	Key Conditions
11	0.84	No checking; duration < 12; no known property
12	0.88	Has checking account
13	0.90	No checking; duration 12–23; no known property; credit < 1,286; other parties present
14	0.94	No checking; duration 23–48; no known savings; credit < 8,098; used car
15	0.95	No checking; duration ≥ 23; known savings; checking in [0,200)

PART — Rule Summary Table

Rule	Prediction	Key Conditions	Coverage
1	Good	No checking; no other payment plans; not used car; no delayed credit; no other parties; ≤1 dependent; credit ≤ 4,530; age > 31	103
2	Good	No checking; no other payment plans; not used car; critical credit history; owns home	51
3	Good	No checking; used car; no other payment plans; some prior credit history	45
4	Bad	No prior credit fully paid; not divorced/separated male; does not own home	15
5	Good	No checking; employment 4–7 yrs; owns telephone	19
6	Good	Duration ≤ 15; guarantor present	26
7	Good	Savings ≥ 1,000; not education/repairs; credit ≤ 4,530	24
8	Good	No checking; radio/TV purpose; ≤1 existing credit; not married/widowed male	32
9	Good	Checking ≥ 200; paid credit history; installment ≤ 3	11
10	Bad	Purpose = education; ≤2 credits; owns home; employment < 4 yrs	11

JRip (RIPPER) Classifier — Rule Summary Table

Rule	Prediction	Key Conditions	Coverage
1	Bad	Has checking account; duration ≥ 18; credit amount ≤ 2,319	98

Rule	Prediction	Key Conditions	Coverage
2	Bad	Has checking account; credit amount ≥ 7,763	50
3	Bad	Has checking account; duration ≥ 39	42
4	Bad	Has checking account; credit ≤ 959; duration ≥ 9; no real estate property	26
5	Good (default)	All remaining cases not covered by above rules	784

Task 3:

Summary

The one-way ANOVA test, which measures the difference between means of the models, shows significance between our classifier results.

```
            Df  Sum Sq  Mean Sq F value Pr(>F)
model        2 0.00794 0.003968    3.65   0.03 *
Residuals   87 0.09456 0.001087
```

Digging further into the data using Tukey HSD, we can see that the Decision Tree model outperformed, especially compared to PART. Part is roughly 2 percentage points worse than Decision Tree. Their respective values were approximately:

```
Significant difference detected (p < 0.05). Running Tukey HSD...
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = accuracy ~ model, data = acc_df)

$model
                              diff         lwr          upr     p adj
PART-DecisionTree        -0.02300000 -0.043297836 -0.002702164 0.0223506
Ripper_JRip-DecisionTree -0.01166667 -0.031964503  0.008631169 0.3607576
Ripper_JRip-PART          0.01133333 -0.008964503  0.031631169 0.3817646

Saved Tukey HSD output to: task3_tukeyHSD.txt

Highest mean accuracy model: DecisionTree
Means:
DecisionTree         PART  Ripper_JRip
   0.7236667    0.7006667    0.7120000
```