1.1) The last three columns with their integer positions are as follows:

59 awarding\_subtier\_agency\_abbreviation

60 funding\_subtier\_agency\_abbreviation

61 business\_categories

1.2) The longest line contains 7,031 characters.

1.3) bicycle appears in 32,122 rows.

1.4) There are 780 unique funding agencies.

1.5) The three largest transactions were all for Medicaid:

MEDICAID ENTITLEMENT FOR 7 – FY 2018 QUARTER 3 – T19: $65 billion  
 MEDICAID ENTITLEMENT FOR CALIFORNIA – FY 2015 QUARTER 3 – T19: $49 billion  
 MEDICAID ENTITLEMENT FOR 7 – FY 2018 QUARTER 2 – T19: $49 billion

2.1) First, I created my hw4.sh and submit.sh files in notepad. Once I finished working on them, I used sftp in my terminal pointing to my cluster address. Once inside the cluster, I could not simply use sbatch ./submit.sh to run the program, because of an error with DOS line breaks. To fix this, I used tr -d ‘\r’ < submit.sh | sbatch instead.

2.2) Because the first data set had 720 files and this data set has 780, the agency IDs are not identical. Hopefully I didn’t mess something up for this to happen!

2.3) The second bash command, cat file | grep “pattern” | sort is more efficient, because it only sorts as many entries as needed (that matches pattern), whereas the first option will require sorting of the entire file.

2.4) How many unique ZIP Codes are listed? What percentage is this of all US ZIP Codes? According to [www.zip-codes.com](http://www.zip-codes.com), in February 2019, there were 41,677 ZIP Codes. According to my code, 41,105 unique ZIP Codes were found in the dataset. This is an astounding 98.6% of all ZIP Codes!

**Appendix**

**hw4.sh**

DATA="/group/staclassgrp/transaction.zip"

# 1.1

unzip -p ${DATA} |

head --lines=1 |

tr "," "\n" |

nl |

tail -n 3 > last\_columns.txt

# 1.2

unzip -p ${DATA} |

wc --max-line-length |

cat > maxchars.txt

# 1.3

unzip -p ${DATA} |

grep --ignore-case 'bicycle' |

cat > bicycle.csv

wc --lines bicycle.csv

# 1.4

FUNDING\_AGENCY\_ID=18

unzip -p ${DATA} |

cut --delimiter=',' --fields=${FUNDING\_AGENCY\_ID} |

cat |

sort |

uniq |

cat > funding\_agency\_set.txt

wc --words funding\_agency\_set.txt # note: includes header

# 1.5

TOTAL\_OBLIGATION=8

TRANSACTION\_DESCRIPTION=25

unzip -p ${DATA} |

cut --delimiter=',' --fields=${TOTAL\_OBLIGATION},${TRANSACTION\_DESCRIPTION} |

sort --reverse --numeric-sort --key=1,1 --field\_seperator=',' |

uniq |

head --lines=30 > largest.csv

head -n 5 largest.csv

# 2.4

POP\_ZIP5=32

unzip -p ${DATA} |

cut --delimiter=',' --fields={POP\_ZIP5} |

cat |

sort |

uniq |

cat > zip\_set.txt

wc --words zip\_set.txt # note: includes header

**submit.sh**

#!/bin/bash -1

SBATCH --partition=staclass

SBATCH --cpus-per-task=2

SBATCH --job-name=Gaylor\_hw4

bash hw4.sh