THE UNIVERSITY OF TEXAS AT AUSTIN
CENTER FOR TRANSPORTATION RESEARCH

Technical Memorandum

To:	Bridget Barksdale
From:	CTR RS/Study Team: Michael Murphy, Zhe Han, Dr. Randy Machemehl, Carolina Baumanis, Meredith Brown, Darren Hazlett, Lisa Loftus-Otway, Sherri Greenberg, John Guttman, Michelle Surka, Kara Takasaki, Matt Kammer-Kerwick, Susanna Gallun, Taehoon Lim, Srijith Balakrishnan, Shidong Pan
Subject:	DPS-CTR IAC Contract – Technical Assistance to TxDPS Driver License Division, Technical Memorandum 4: Process and Prepare Information Obtained from Tasks 1 – 3 for Further Analysis
Date:	June 21, 2020

Table of Contents

List of Tables

List of Figures

Executive Summary

This technical memorandum contains five major chapters:

- Chapter 1 presents the data cleaning, filtering, and preparation processes used in this study
- Chapter 2 discusses data cleaning and filtering examples for surveys
- Chapter 3 outlines data cleaning, filtering, and preparation examples for agency databases
- Chapter 4 summarizes development of meeting transcripts or notes and preparation of meeting summary documents
- Chapter 5 provides conclusions and key findings

This technical memorandum discusses the processes and methods used to cleanse and filter data. Following are some notable findings presented in this technical memorandum:

- A process was developed by the Study Team for data cleaning and filtering to ensure high quality data and information was used in the various analyses.
- Extensive survey data and institutional datasets in various formats were obtained, reviewed, and cleansed/filtered.
- Examples are provided of the cleansing processes used and of variations found in the data.
- The effects of cleansing in terms of the amount of data remaining available for analysis are discussed when applicable.

Chapter 1. Data Cleaning, Filtering, and Preparation

This study involves obtaining large volumes of data through surveys; various state agency databases, including DPS, DMV, TxDOT, the Texas Demographic Center, the U.S. Census Bureau, and other sources. Each data source requires close examination of the data and data formats to develop methods for identifying inaccurate, out-of-range, corrupted, or otherwise unusable data. The methods that are applied vary significantly depending on whether the data has been entered by hand, or through automated methods, or by combining data from different sources.

Data entry by hand can result in human errors, such as misspelled words, misunderstanding the question and thus providing inappropriate response(s), 'careless' entries by individuals who did not take a survey seriously, and other factors. Different types of data errors might occur if the data is generated by automated processes, such as blank or missing entries; incorrect uploading of the raw data to the analysis database (columns out of alignment, etc.); and variations in the original data format (numeric), such as raw data whole numbers (1, 2, 3, 4, etc.) values that are incorrectly stored in the analysis database as rational numbers (1.0000, 2.0000, 3.0000, 4.0000, etc.). Other errors can occur due the incorrect use of lookup tables in the analysis database; these tables are used to convert raw numeric or abbreviated text entries to word or word strings in the analysis database.

The following sections discuss the process developed for reviewing, cleaning, filtering, and preparing data for further processing. The examples have been described in sufficient detail to explain the range of data errors that are possible, the various methods used to find and correct or extract the data from the analysis set, and to show the amount of effort necessary to create a quality set of data to ensure analysis result integrity.

1.1. The Data Cleaning and Filtering Process

Data cleansing (cleaning) and filtering is an extremely important process that helps ensure that the data used in the study analyses are of the highest quality possible. Figure 1.1 shows the data cleaning and filtering process used in this study.

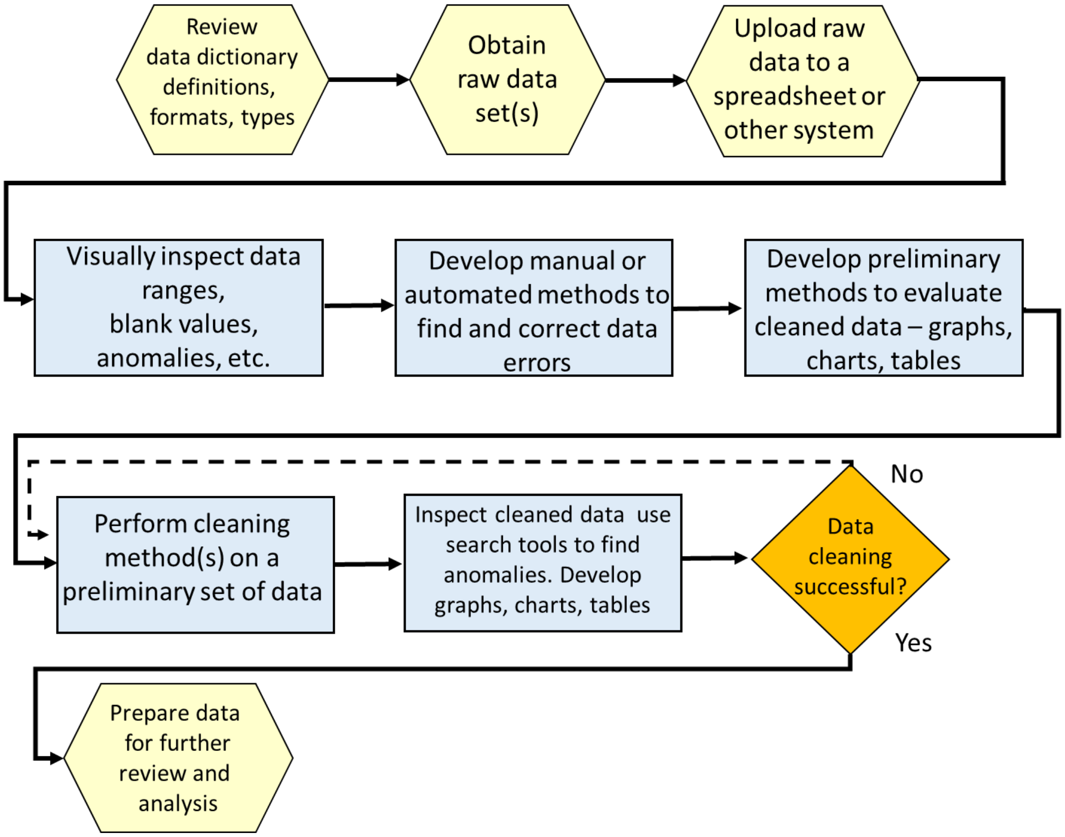

Review data dictionary definitions, formats, types
Obtain raw data set(s)
Upload raw data to a spreadsheet or other system
Visually inspect data ranges, blank values, anomalies, etc.
Develop manual or automated methods to find and correct data errors
Develop preliminary methods to evaluate cleaned data – graphs, charts, tables
Perform cleaning method(s) on a preliminary set of data
Inspect cleaned data use search tools to find anomalies. Develop graphs, charts, tables
Data cleaning successful?
No
Yes
Prepare data for further review and analysis

Figure 4.1 Data Cleaning and Filtering Flowchart Showing the Protocol Used in This Study

The data cleaning and filtering process seeks to ensure that:

a. Each set of data is examined to ensure problems are addressed as soon as practical. This is an interactive process in which very early stages of analysis might reveal data anomalies that require new methods for examining data or other actions to ensure data integrity;

b. Each set of data is examined in consideration of the data source and input methods (input by hand, automated, transcribed and so forth). This is because certain input methods (e.g., user input by hand) can result in many more variations in data entry and error types that require additional attention and check methods;

c. Each data field is of the proper format and type (text or numeric input, five-digit zip code, or other format) and is complete;

d. Data integration of values from different data sources are checked thoroughly to ensure the combined results are accurate;

e. Data examination, cleaning, and filtering continues throughout the analysis process to maintain high standards of data accuracy; and

f. Questionable data values are flagged or set aside for further evaluation or removed from the database to reduce potential errors.

1.2. Conclusions

Chapter 1 explained the importance of data cleaning and filtering and provided a diagram of the process. A series of more detailed steps in the data cleaning and filtering process were also given.

The following chapters summarize protocols that were applied to the various types of data used in this study. Chapter 2 outlines data cleaning and filtering for surveys and email databases. Chapter 3 covers data cleaning and filtering for databases obtained from public agencies. Chapter 4 discusses processing methods used for breakout session and focus group meeting documents. Chapter 5 summarizes the key findings of TM-4.

Chapter 2. Survey Data Cleaning and Filtering

Study Team members conducted online surveys with:

- Driver License Division – frontline staff (TM-5)
- Driver License Division – administrative staff (TM-5)
- Driver License Program customers
  - Driver License and ID Card Survey (TM-3)
  - Customer Renewal Choices Survey – designed to explore ways to incentivize online renewals (TM-8)
- County tax assessor-collectors (CTACs) – a survey of 31 CTACs to obtain general information (TM-3)
- CTACs – a survey of all 254 CTACs to obtain information about staffing, transactions, and related information (TM-3)

When a survey taker clicked on the survey link in the invitation email, the Qualtrics survey form was opened in the survey taker's internet browser. Additionally, Driver License and ID Card Survey takers might have accessed the survey using a URL or QR code on posters placed in every driver license office (DLO). The survey taker might have been using a desktop PC, laptop computer, tablet, or cell phone to take the survey—it is suggested that use of smaller keyboards could have contributed to some input mistakes.

Survey responses were automatically stored in the online Qualtrics database that is accessible by the Study Team survey creator and to those whom access is granted. The survey responses contain no personal information related to the survey taker and cannot be linked to the email address used to send the invitation. Under some instances a survey taker would provide contact information and offered further information.

In some cases, survey takers quickly reviewed the survey and closed it with no further input (reported by Qualtrics as 0% complete). Other survey takers completed the first few questions, then closed the survey (reported by Qualtrics as 19% complete). Yet other survey takers completed nearly all of the questions (86% complete), or completed all questions (100% complete). The analyst for the Driver License and ID Card Survey chose to use all surveys that were at least 86% complete for the next step in the data cleaning and filtering process. This is because the Driver License and ID Card Survey did not require a survey taker to answer all questions in order to provide useful information. In addition, certain questions might not have been applicable to the survey taker and were therefore skipped. The analyst for the Customer Renewal Choices Survey chose to use only 100% completed surveys.

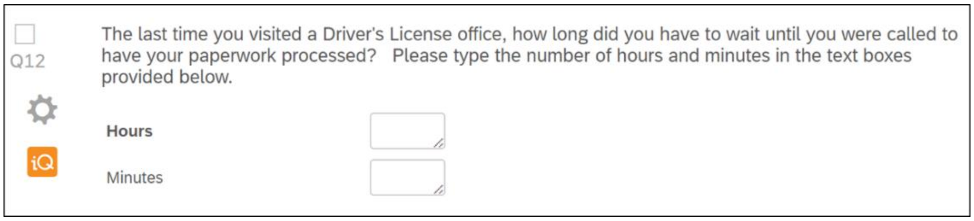
To ensure that the survey provided broad coverage of Texas counties and regions, survey results were tallied in relation to city or county populations. The survey asked for the respondent’s residence zip code and the city/county where they had last had a driver license or ID card transaction for this purpose. Thus, Harris County, which is the most populated Texas county, would be expected to have the largest number of survey responses. This information was maintained in the Driver License and ID Card Survey.

The following summary provides a comprehensive list of the types of data cleaning and filtering processes that were applied to surveys.

2.1. Survey Data Cleaning

Online surveys involve survey takers who input their responses by hand. Inputs for the study surveys included typing text or numeric responses, and/or selecting predetermined options from lists, and/or clicking radio buttons to choose ratings based on Likert Scale values (e.g., Very Good, Good, Fair, Poor, Very Poor), among other question response types. Questions that require the survey taker to type text or numeric entries can result in a variety of unexpected responses that must be evaluated and possibly adjusted to allow further analysis.

Figure 4.2 shows Q12 from the Driver License and ID Card Survey, which asked the survey taker to input their transaction wait time in two data entry fields labelled ‘Hours’ and ‘Minutes’. Figure 4.3 shows Q13, which asked the survey taker to rate their wait time by clicking on a radio button corresponding to a Likert Scale from Very Good to Very Poor, or indicate ‘No Opinion’.



□ Q12

The last time you visited a Driver's License office, how long did you have to wait until you were called to have your paperwork processed? Please type the number of hours and minutes in the text boxes provided below.

Hours

Minutes

Figure 4.2 Question 12 from the Driver License and ID Card Survey – Wait Time

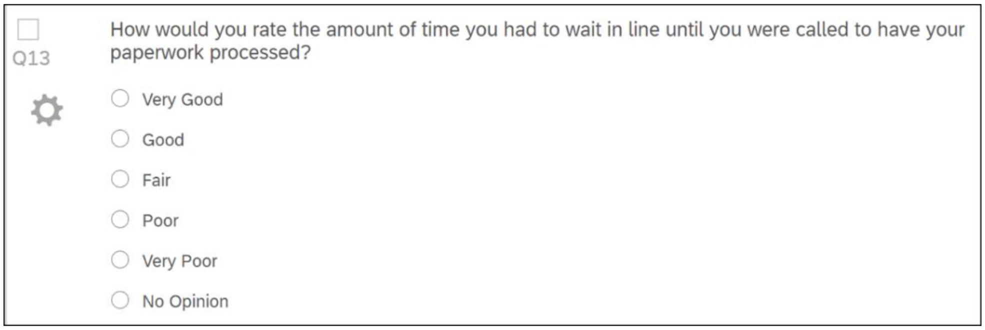

□
Q13
How would you rate the amount of time you had to wait in line until you were called to have your paperwork processed?
○ Very Good
○ Good
○ Fair
○ Poor
○ Very Poor
○ No Opinion

Figure 4.3 Question 13 from the Driver License and ID Card Survey – Wait Time Perception Rating

For example, if a customer's wait time was 1.5 hours, in Q12 they would input the number '1' in the box labeled 'Hours' and the number '30' in the box labelled 'Minutes'. For Q13, the survey taker would then click one of the radio buttons (such as 'Fair') to express their rating of their transaction wait time.

In order for wait time or processing time to be used in the analysis, the survey taker had to provide both the time waited and a rating value. Otherwise, the entry was recorded using one of the following four options and therefore was not used in the analysis. The number and types of incomplete responses was tracked and tallied with complete responses to help monitor data analysis progress. The four incomplete options were:

- **Blank Time**—applied to an entry in which the survey taker did not provide a wait time, but did provide a rating of their wait time.
- **No Opinion**—applied to an entry in which the survey taker provided a rating of 'No Opinion' regardless of whether they input a wait time.
- **All Blanks**—applied to an entry in which the survey taker did not provide a wait time or a rating value.
- **Time/Blk Qual**—applied to an entry in which a wait time was input, but no rating of any type was provided.

Some survey responses required editing, as shown in the following examples. These are a broad sampling only and do not imply that these specific corrected values were entered for every case regardless of the survey taker's input:

1. One, Thirty or One Hour, Thirty Minutes – Text responses had to be changed to numeric responses so that calculations could be performed in the analysis.
2. 1hr., 30min – The survey taker added abbreviations for hours and minutes that had to be deleted so that the input value would be read as a number.

3. Blank, 90 – The survey taker did not enter a value for hours, and typed 90 minutes in the Minutes field, which is equal to 1 hour 30 minutes. Responses of this type were converted to 1 hour 30 minutes.

4. One, 30 – The survey taker used a mix of text and numeric inputs that were converted to all numeric (e.g., 1 hour 30 minutes).

5. 1+/- 30+/- – The survey taker provided the wait time with the caveat that the times were approximate. In the case that specific numbers were used, these were converted to pure numeric values: 1 hour 30 minutes.

6. <1, Blank – The survey taker indicated the wait time was < 1 'less than' 1 hour, but was not specific—no entry was given for minutes. In these cases, 1 minute was subtracted from the hour value designated and entered in the minutes input cell as 59 minutes; otherwise the value would not have been usable.

7. ~1, or ~1hr., or ~1 hour – The diacritical mark tilde '~' means 'approximately' or 'approximately equal to'. Thus the survey taker who input ~1 waited approximately 1 hour. In these cases, the value was converted to 1 hour.

8. 0, 1 – Careful consideration was given to survey responses in which the wait time was shown to be a very small value, such as 1 minute or even zero minutes. However, it is possible that a customer was served immediately or nearly immediately upon arrival, as was discussed with Lubbock DLO personnel. In those cases, DLO employees may pull tickets for customers waiting in line outside the DLO early in the morning, and hand the ticket to each customer as they enter. The number is immediately called and thus the customer had essentially zero wait time or about 1 minute of wait time before being served.

9. 999, 0 – Entries for wait time that were not feasible, such as 999 hours resulted in close examination of the entire survey response. In some cases, data entered in other fields were also 'out of bounds' or reflected a 'careless' survey response. In these cases, the entire survey response was deleted. However, in cases that the survey taker left comments indicating that they had to return more than once to the DLO and waited long periods of time in each case, the value 999 was deleted as out of bounds, but the remainder of the survey response was retained in the database.

10. |, or |5 minutes – Some surveys had a vertical bar in place of the number 1 for wait or processing times. It is unknown if this was intentional and

intended by the survey taker to provide incorrect information, or a possible error introduced by the internet browser being used. In these cases that the obvious intent was for the bar to be the number 1, the vertical bar was replaced with the number 1.

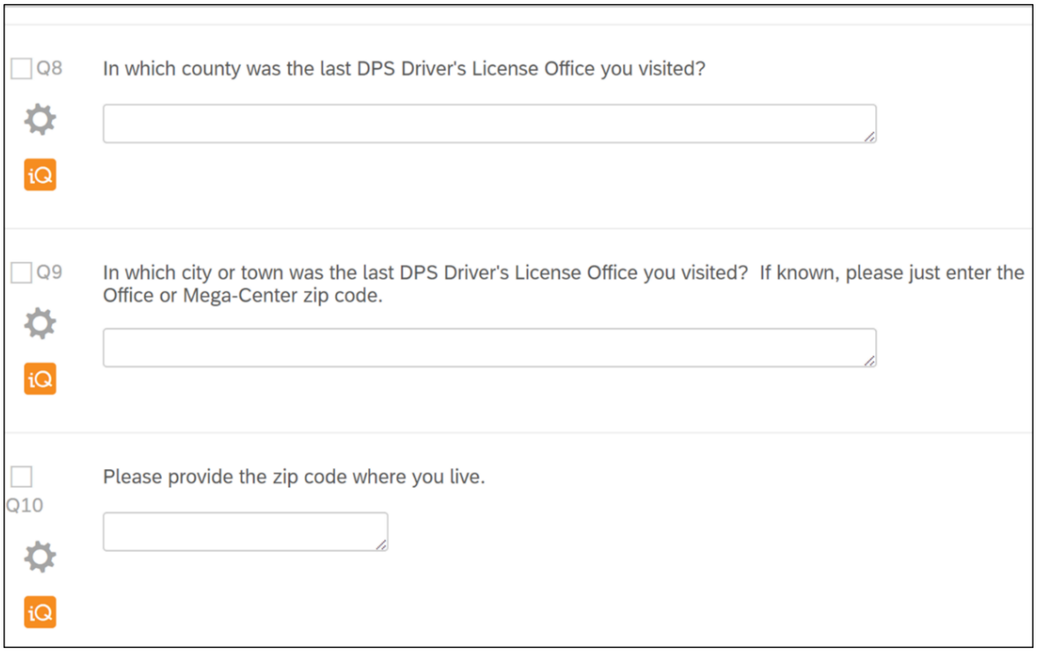
11. "I don't know", "I don't remember" entered for wait time, with a rating given for wait time – The text messages were deleted and the response recorded as "Blank Time".

12. Wait time entries that were recorded as a date (15-Oct), instead of a number or text entry – It was learned during this study that Excel has an internal error that can result in a CSV (comma spaced value) file, with numeric entries, being randomly converted to dates upon conversion to an Excel file. The random conversions of numeric entries to a date is rare for the amount of data contained in this survey; however, these date entries were discovered during data cleaning. It was further learned that once this conversion has occurred, there is no way to recover the original entry; thus the date was deleted and the remaining entry categorized accordingly. These dates were therefore deleted and the entry recorded as a 'Blank'.

13. 15–20, or 15/20 – In some cases, survey takers provided a range of time either in hours or minutes. For these cases, the average value was computed and used to replace the range. Thus, "15 – 20" minutes would be replaced with 17 minutes. An input of "1 – 2" hours would be replaced with 1 hour 30 minutes.

14. 1 1/2, mixed numeric values of whole numbers and fractions, or fractions entered to represent a fraction of an hour – These were converted to the nearest equivalent in hours and/or minutes; thus 1 hour 30 minutes was entered in the case of 1 1/2, or 0 hours 30 minutes for ½ hour.

Each survey taker was asked to provide their residence zip code and the name of the town/city and county where they last visited a DLO or mega center. Figure 4.4 shows Q8, Q9, and Q10, which requested this information.



□ Q8 In which county was the last DPS Driver's License Office you visited?

□ Q9 In which city or town was the last DPS Driver's License Office you visited? If known, please just enter the Office or Mega-Center zip code.

□ Q10 Please provide the zip code where you live.

Figure 4.4 County, City of the Last DLO Visited; Zip Code of Residence

Though Study Team members recommended creating a drop-down list from which the county name could be selected, it was not possible to use a drop-down list due to Americans with Disabilities Act (ADA) requirements. The University of Texas (UT) Office of Research Support and Compliance, which reviews survey instruments through an Institutional Review Board (IRB) evaluation, does not permit using certain Qualtrics question format types that cannot properly interpreted by a reading machine.

Thus, drop-down lists, rank order, and slider questions are not permitted for use in a survey that will be distributed to the public. This, however, can result in additional work in data cleaning due to:

1. Misspelled words. Any misspelled words found were corrected to ensure that the associated survey record would be found or correctly counted by the various search algorithms used.
2. Using the letter ‘O’ instead of the number ‘0’. Many errors were corrected when it was found that some survey takers had used the letter ‘O’ or ‘o’ instead of the number ‘0’. Corrections increased the number of usable survey records.
3. Variations in how certain city names were spelled. A single city may be spelled by survey respondents in multiple ways: Fort Worth, Ft. Worth, Ft Worth, ft Worth, FTW. A standard spelling was selected from references and used throughout the analysis database. Searches for spelling variations

were performed using the Excel 'Search and Replace' function to create consistency.

4. Variations in how a county name is provided. Examples include places that may or may not have a space between words: ElPaso, El Paso; LaSalle, La Salle.

5. Misinterpretation of the word 'county' for the word 'country'. A number of survey takers input U.S. or U.S.A (or Texas) in the county input field. It was apparent that these individuals had misinterpreted the word 'county' as 'country'. The city location was used in the majority of these cases to find the correct county using online resources.

6. Variations in how a zip code is provided. Survey respondents provided either the typical five-digit zip code (such as 78631) or the nine-digit zip code (such as 78631-1234). A lookup table of all Texas zip codes (five-digit) with the associated town and county was obtained and copied into the analysis spreadsheet. This allowed the user to determine the residence city/county by using the Excel 'countifs' command, which searches the lookup table for the survey entry zip code and returns the city and county in the appropriate location. However, since the nine-digit zip code was not applicable, all nine-digit zip codes had to be amended to a five-digit zip code using the search and replace function.

7. Providing a zip code for all three inputs (county name, city name, and residence zip code). Some survey takers typed in a zip code for county, city, and residence—in some cases different zip codes for residence and county/city where their last driver license transaction occurred. In these cases, the zip code location was searched on the internet and the city and county input in place of the zip code.

8. Military zip codes. In some cases, survey takers were active duty military personnel, as denoted in their comments; thus, if overseas, they provided their Army Post Office (APO) or Fleet Post Office (FPO) number. In these cases, since the APO or FPO number did not correspond to a Texas city/town, or county, these numbers were not used in tallying the number of county or city survey responses.

9. Zip codes from other states. In some cases, a survey taker would provide valid information for all fields, except their residence zip code. In almost every case, these residence zip codes were for cities outside Texas. The interpretation for this type of survey input was that the survey taker was a Texas resident who had conducted business at a DLO to obtain a driver license—but was either a new resident to Texas and had forgotten their Texas zip code or, out of habit, automatically provided their recent, out-of-state zip code. Out-of-state zip codes were not usable for residence

information and thus, though the survey results were valid, it was not possible to designate the city or county for these results.

These examples summarize some of the types of data cleaning processes that were applied to the surveys.

The Study Team obtained email addresses from DLD; customers provide these email addresses when applying for or renewing a driver license or ID card. In all, approximately 7.3 million email addresses were provided based on the following time frames for driver license or ID card transactions:

1.	January – December 2018	approximately 3.9 million email addresses
2.	January – September 2019	approximately 2.1 million email addresses
3.	October 2019 – February 2020	approximately 1.28 million email addresses

The DLD email addresses were provided in Excel spreadsheets, which are limited to approximately 1,050,000 entries per workbook page. During examination of the email addresses, it was found that duplicates existed—these were removed using the ‘Remove Duplicates’ feature in Excel. Duplicate email addresses occur due to customers who conduct more than one driver license or ID card transaction in the same year. For example, a customer might renew their license, then later have Lasik surgery, which eliminates the need to wear glasses. Thus, a second visit to the DLO is needed to have the eye glasses restriction removed and a new license prepared.

These email addresses were used for distribution of two customer surveys:

- TM-3 DPS Driver License and ID Card Survey (customer experiences and opinions)
    - Survey invitations were sent to all 7.3 million email addresses obtained from DLD.
- TM-8 Customer Renewal Choices Survey
    - Survey invitations were sent to approximately 2.1 million email addresses from DLD for the period January–September 2019.

The Study Team members conducting surveys requested email addresses with no other information appended. The Study Team did not want to have possession of any information that could associate the email addresses with a person’s name, location, or any other personal information. Thus, each spreadsheet prepared for mass emailing contained only the email addresses.

The email addresses were sub-divided into smaller subsets (numbers of emails) and stored in separate Excel or CSV spreadsheets. The number of emails depended on the distribution method used; several email distribution methods were used during the study as options to send a larger number of emails at one time became known.

1. Microsoft Word/Outlook. The Study Team used this software to send 500 emails at a time, distributing an invitation to take the survey and a link to a page on the CTR website with information about the study. The daily limit was 10,000 emails maximum. This method was considered too slow, but was used until more efficient, higher-volume distribution methods could be found. This process required the Study Team to create a series of Excel spreadsheets containing 500 email addresses each.

2. Qualtrics™ platform (free use). Each Qualtrics user at UT has a maximum allowable limit of 55,000 emailed survey distributions per week. This method was used for approximately 440,000 emailed survey invitations—but again this rate was considered too slow. This process required the Study Team to create a series of CSV files containing approximately 55,000 email addresses each.

3. A MailChimp™ account. This account allowed 500,000 email survey invitations to be sent at one time. The distribution occurred over approximately 2 hours. This process required the Study Team to create a series of Excel spreadsheets with approximately 500,000 email addresses each. However, concerns were raised about causing problems for the UT email system due this large volume of emails being distributed over a short time span. This method was discontinued after approximately 1.5 million email invitations had been sent.

4. Qualtrics™ platform (paid use). Discussions with UT's IT department and the Qualtrics technical support team resulted in a new option that had not been previously known. Qualtrics tech support indicated that the Study Team could engage Qualtrics to distribute the email invitations using the Qualtrics email server system. There was no limit on the number of emails that could be sent at one time using this process. Thus, the Study Team arranged a purchase order to distribute approximately 5.6 million email survey invitations in two tranches: one for the Driver License and ID Card Survey and the second for the Customer Renewal Choices Survey. CSV files were developed containing approximately 5.6 million email addresses and provided to Qualtrics for the email distribution.

5. The email invitation contained contact information in case the recipient had questions about why the survey was being conducted or requests for

additional information about specific questions. Hundreds of phone calls and emails were answered during these mass distributions. The Driver License and ID Card Survey was conducted in English/Spanish; thus, the email invitation was written in both English and Spanish.

2.2. Conclusions

The surveys performed for this study were thoroughly examined to correct errors, document missing information, and provide the best quality data sets for further analysis. The previous section summarized many, but not all, of the various methods used to perform data cleaning and filtering. The next section discusses cleaning and filtering that was performed for databases provided by state agencies.

Chapter 3. Data Cleaning and Filtering – Public Agency Databases

The Study Team obtained databases from DPS, DMV, and other public agency sources. Each set of data was examined to the extent required by the file size and contents of the database. The following summary discusses the processes applied to the DLD NEMO-Q queuing system data, the DPS-DLD High Value Datasets, and the DMV CTAC–Vehicle Title and Registration (VTR) transactions database from the Texas DMV.

3.1. NEMO-Q Data Cleaning and Preparation

The Study Team requested and obtained NEMO-Q queuing system raw data sets from DLD. At the beginning of this study, NEMO-Q queuing systems were installed in 77 DLOs and mega centers. The NEMO-Q queuing systems have now been replaced by Applus Appointment systems located in 226 offices.

The complete NEMO-Q data set contains seven CSV files, including two main tables containing customer transaction information (Event Table and Subservice Table) and five dictionary tables (User Category Table, Cashier Category Table, Office Category Table, Service Type Local Category Table, and Subservice Category Table). The features contained in Event Table are presented in Table 4.1.

Table 4.1 NEMO-Q Event Table Features

Data Field	Description
‘IINDEX’	A unique number assigned to each transaction.
‘OFFICE’	Station number of the driver license office where a transaction occurred
‘DATUM’	The date that a transaction occurred.
‘SERVICETYPE’	Service type provided for a transaction (e.g., 1-Shorts, 2-Longs, etc.)
‘CODE’	Subservice type provided for a transaction
‘START’	The time that a transaction placed in the queue for the service.
‘SERVICE’	The time that a transaction started being processed.
‘EOS’	The time that a transaction was completed.
‘WAITTIME’	The time interval for which a transaction had to wait after being placed in the queue and before the service actually occurred. (‘SERVICE’ – ‘START’)
‘SERVTIME’	The amount of time required to complete a transaction after the service started. (‘EOS’ – ‘SERVICE’)
‘BOOK_TIME’	Reserved time for a transaction. No value is provided if a transaction did not have a reservation.

The Subservice Table contains more detailed information regarding the subservice type and indication whether the transaction is completed or incomplete. Incomplete transactions refer to the transactions that could not be completed due to customers that failed to provide required documentation, fees, or information. The complete service and subservice type codes are provided in Table 4.2.

Table 4.2 Service and Subservice Type Codes

Field Name	Service Type Code	Description
Service Type Code	0	Unknown[1]
	1	Shorts[2]
	2	Longs[3]
	3	Americans with Disabilities Act (ADA)
	4	Not Listed
	5	Automated Driver License Testing System (ADLTS)
	6	Road Test
Subservice Type Code	1 - 7	Renewal – DL, ID, CLP, CDL, EIC, etc.
	8 - 14	Replacement –DL, ID, CLP, CDL, EIC, etc.
	15 - 21	Original – DL, ID, CLP, CDL, EIC, etc.
	22 - 26	Modification – DL, CLP, CDL, etc.
	27	Automated Driver License Testing System (ADLTS)
	28	Road Test
	29 - 31 34 - 45	Incomplete transactions
	32	No show
	33	None of the above

[1] This service type is not listed in the Service Type Local Category Table, but exists in the Event Table

[2] Transactions that would require short service time (e.g., driver license renewals)

[3] Transactions that would require long service time (e.g., original driver license)

The NEMO-Q raw data includes all transaction information from January 2017 to March 2020 at the 77 DLOs equipped with the kiosks, containing more than 21 million transactions. Due to the huge size of this dataset, the Study Team used the Python programming language to conduct the data processing and cleaning. Pandas, a free software library written for the Python programming language for data manipulation and analysis, was used to handle this process.

Step 1. Elimination of blank and missing values

The Study Team noticed that some data records in NEMO-Q data sets had blank or missing values in service type, subservice type, waiting time stamp, service time stamp, and/or total transaction time stamp. These data records cannot be used in future analyses due to missing data. The Study Team eliminated those incomplete records with missing values.

After this step, 50,336 records were eliminated and 20.98 million records remained.

Step 2. Integration of Event Table and Subservice Table

Since the Subservice Table contains information indicating whether the transaction was complete or incomplete, while the Event Table does not have such information, the Study Team integrated the Event Table with the Subservice in order to determine which transactions were completed and which were incomplete. The Study Team used "DATUM", "PRI_INDEX" and "OFFICE" to integrate the two tables. The attributes of data records with the same "DATUM", "PRI_INDEX" and "OFFICE" were integrated. During this process, the Study Team found that some records were mismatched. Therefore, the Study Team eliminated those records because it was unknown whether the transaction was complete or incomplete.

After this step, 1.18 million records were eliminated and about 19.85 million records remained.

Step 3. Elimination of DLOs that have been closed

During data examination, the Study Team noted that 4 of the 77 DLOs have been closed at least since May 2018. These include 116-Cedar Hill, closed on July 28, 2017; 299-Clear Lake, closed on May 2, 2018; 201-Houston-Winkler, closed on April 26, 2018; and 227-Pasadena, closed on April 30, 2018. Since one of the purposes of NEMO-Q data analysis was to evaluate the change of average wait time, service time, and transaction time before and after the DLD new hires began in September 2019, the Study Team eliminated transaction records from those four closed DLOs because they were closed before the new hiring began.

After this step, 268,759 records were eliminated and about 19.58 million records remained.

Step 4. Selection of Service Type

Although six service types were listed in Table 2, most of them were short transactions (e.g., DL or ID renewal) and long transactions (e.g., original DL or ID). The Study Team also noticed many transactions whose service type code was "0". However, there is no code "0" in the Service Type Local Category Table. The Study Team sent inquiring emails to NEMO-Q Inc. asking for clarification and explanation. No responses had been received as of June 20, 2020. Therefore, to ensure accuracy, the Study Team eliminated the transactions whose service type is "0" and extracted all transactions marked as shorts and longs for future analysis.

After this step, 4.02 million records were removed and 15.56 million records remained.

These 15.56 million transaction data records were used in later various analysis stages, including quantitative comparisons and generation of graphical results.

3.2. Driver License High Value Dataset

DPS maintains ten high value data sets on their website for public access. The Driver License High Value Dataset consists of monthly reports that can be downloaded in PDF format, which provide information on a range of Driver License Program activities that include:

- Numbers of transactions of different types
- Customer Service Center (Call Center) transactions
- Enforcement & Compliance Services
- Commercial Driver License (CDL) Program
- Impact Texas Driver (ITD) Program
- License and Record Services (LRS)

The Study Team downloaded all monthly reports dated from September 2017 to February 2020; monthly reports are stored based on fiscal years. The data from these reports were transcribed to an Excel database for further analysis and reference. The Driver License High Value Dataset is accompanied by a three-page document that provides definitions for the different data elements in the monthly reports.

The Study Team used this information to help evaluate the relationships between in-person and online or mail-in transactions, including trends over time. It is noted that the information contained in the dataset was for completed transactions only.

3.3. DMV VTR Transaction Dataset

Breakout sessions were conducted with DMV executive leadership, state agency subject matter experts, and CTACs. During these sessions, information was obtained about the VTR Program as well as basics about the interaction between DMV and the CTAC. Based on preliminary information, similarities between the size and complexity of the VTR Program were seen when compared to the Driver License Program in terms of:

- Number of transactions processed each year
- Transactions processed in-person, online, and through mail-in

- Operation of a call center
- Office located statewide to provide VTR services

However, there were also a number of differences between the Driver License Program, which is entirely operated by DLD employees, and the VTR Program. The VTR Program combines personnel and other resources from DMV and CTAC offices to provide customer service and process millions of transactions per year. Some basic statistics regarding the VTR program include:

- DMV VTR transaction processing involves 146 employees.
- DMV VTR transactions are focused on heavy truck and truck fleet transactions.
- DMV operates a centralized call center that handles nearly 700,000 calls per year with a high success rate of answered calls.
- DMV develops policies and guidelines regarding how VTR transactions area to be processed including the types of documents that are required to be collected and stored for each transaction. CTACs follow the DMV policies and guidelines when processing VTR transactions.
- DMV reviews county VTR transactions and may reject a transaction if incomplete or not documented according to policy.
- CTACs operate 514 offices statewide.
- CTACs are responsible for hiring, managing and evaluating performance of approximately 3,000 employees who perform VTR transactions in some capacity (customer-facing transaction, mail room, accounting, mail in transactions, answering VTR telephone calls or operating a call center.
- CTACs also manage a program of hundreds of ‘partner’ locations located in grocery stores and other types of business offices. Partner locations are deputized by the CTAC to perform vehicle registration sticker transactions.

The Study Team contacted DMV to request a database of CTAC VTR transactions for further evaluation. The following summary data was determined from the database regarding CTAC VTR transactions, based on monthly averages for a 24-month period (May 2018 through April 2020).

- CTACs process approximately 18,138.275 customer-facing, in-person VTR transactions annually;

- DMV/CTACs process approximately 4,103,623 online VTR transactions annually

- CTACs process approximately 728,655 mail-in transactions annually

- CTAC partner locations process approximately 1,539,040 registration sticker transactions annually.

- Thus, the vast majority of VTR transactions (more than 95 percent) are processed and managed by CTACs,

The CTAC VTR data helped provide the Study Team obtain a better understanding of the cooperative interaction between DMV and the 254 CTACs.

3.4. Conclusions

Chapter 3 summarizes examples of the data cleaning, filtering, and processing steps that were applied to datasets obtained from public agencies. Considering that state agency datasets receive close examination prior to use, there were fewer data cleaning and filtering processes required than for survey data.

Chapter 4. Information Filtering for Meeting Documents

The Study Team conducted ten WebEx breakout sessions with agency directors, key members of their management teams, and other subject matter experts. In addition, three in-person DLD customer focus group meetings were held in Austin (prior to COVID-19). These meetings are documented in TM-3. Additional DLD employee focus group meetings were held at the Boerne, Lubbock, and Houston offices. These meetings are documented in TM-5.

The three in-person DLD customer focus group meetings were conducted by IC^2. The WebEx breakout sessions were moderated by different members of the Study Team.

The ten WebEx breakout sessions are listed below:

1. Management, Operations, and Performance Standards (one breakout session)
2. REAL ID Compliance, Security, and Safety (one breakout session)
3. Customer Service (three breakout sessions)
4. IT (three breakout sessions)
5. Call Center Operations (two breakout sessions)

4.1. Breakout Session Documentation

Breakout sessions lasted from 1.5 to 2.5 hours each and comprised from three to nine subject matter experts and from three to five Study Team members. Due to COVID-19 sheltering, all breakout sessions were held using WebEx online meeting software. A series of questions was prepared in advance and discussed with the participants during the course of the meeting.

The breakout session Study Team leader used their preferred method to document the discussions for later evaluation and information extraction. These methods included:

1. WebEx session recording – WebEx provides a recording that can be played back to view the entire meeting and prepare transcripts. The breakout session attendee’s permission was requested to record the session.
2. Note taking – the meeting discussions were documented by a note-taker either as an observer or a participant in the meeting. The note-taker summarized the key comments as each question was asked. In certain cases, more than one set of notes was taken and later compiled to produce a comprehensive set of notes.

The WebEx recordings were later transcribed to provide a typed document of the entire meeting for later use. An approximately 1.5-hour session could produce a 25-page transcript. Meeting notes taken by hand were later transcribed by creating a Word document. Certain discussion leaders chose to maintain confidentiality of all comments; thus, each comment was denoted by a speaker number (Speaker 1, Speaker 2 etc.). In other cases, since it was thought important to know the person who made the comment, the speaker's name was recorded along with each comment. This allowed the Study Team to later contact a person to ask further questions or to identify knowledgeable individuals for inclusion in other breakout sessions. Handwritten notes converted to a Word document could exceed 10 or more type written pages.

4.2. Extracting Information to Inform Decision-making

Extracting information from transcripts or typed notes was guided by two primary considerations.

1. How does the information inform the benefits and drawbacks associated with the three options under consideration: DLD remains in DPS, DLD moves to DMV, or DLD becomes a stand-alone state agency?
2. How does the information inform the six criteria, listed below, which were used in the Option Ranking Scheme?
    1. Customer Service
    2. Compliance/Security
    3. Accountability/Trust
    4. Efficiency/Cost
    5. Culture/Staffing
    6. Disruption

The Study Team acknowledges that information and facts gleaned from these meetings often required subjective, qualitative assessments. These assessments helped determine how information applied, and the degree to which it applied, to one or more criteria and one or more options. These assessments were often not quantifiable and were often based on collective Study Team discussions that included individuals with different technical/professional backgrounds.

The Study Team sought to obtain information from many sources to inform all options and criteria. However, the amount and types of substantive information about customer service, performance, management, and operations was affected by the following considerations:

1. The Driver License Program is currently located in DPS. The program has never been in DMV, nor has it ever been a stand-alone state agency. Thus, information about DLD in DPS is extensive and partially quantifiable. However, considerations about the potential management, operation, and performance of the DLD within DMV or as a stand-alone agency are necessarily informed by expert opinions. These opinions were based on historical events involving merging or forming of other state agencies in Texas or other states. The formation of DMV from divisions within TxDOT provided information about potential improvements in customer service when a new state agency is created from portions of an existing agency. The state legislature has lauded DMV for its customer service, management, and operations.

2. The analysis of the three options applies to a program that is changing over time. For example, since the beginning of this study, DLD has implemented the following changes:

   I. Hired hundreds of new License and Permit Specialists (LPS) and given the majority of all LPS employees a substantial raise. This was funded by a $212 million funding provision to DLD in House Bill 1.

   II. Replaced the NEMO-Q™ queuing system located in 73 DLOs or mega centers with the Applus™ Appointment System, which is now located in 228 DLOs or mega centers;

   III. Replaced 1,600 PC workstations with new systems in DLOs and mega centers;

   IV. Upgraded the Windows Operating System on the 1,600 PCs to Windows 10;

   V. Awarded a new vendor contract, which has resulted in installation of new biometric data capture systems, including new cameras, thumbprint pads, and signature pads in all DLOs and mega centers.

   It is expected that the implementation of these new IT resources and the hiring of hundreds of new LPS employees will result in customer service improvements. However, it is too early to understand in detail exactly how, and to what extent, these changes will impact customer service. This is because COVID-19 ‘sheltering’ closure resulted in closing of DLOs and mega centers to driver license renewal customers. At the time of this writing, DLOs are primarily servicing CDL customers.

The breakout session and focus group meeting summaries were created to extract information that would help evaluate and rank the criteria for the three options by providing qualitative and at times quantitative information.

4.3. Conclusions

Preparing the documentation of the breakout sessions and focus groups provided an extremely important resource to the Study Team for subsequent evaluation of ranking criteria used to evaluate the three options. Different techniques were used to hold meetings, document the discussions, and prepare meeting summaries. The detailed information for the breakout sessions and focus groups are contained in TM-3 and TM-5.

Chapter 5. Conclusions and Key Findings

TM-4 provided information about data cleaning, filtering, and the preparation steps taken by the Study Team for survey data, agency databases, and documentation obtained during this study. A process was presented outlining the basic steps taken to ensure that data and information quality was maintained at a high standard. Examples were provided of the types of errors or data anomalies encountered and the steps taken by the Study Team to either remove or remedy the problem.

Key findings include:

1. Survey data cleaning, filtering and preparation is very labor intensive and time consuming. Human data entry results in many variations and/or data errors that must be found and either modified or deleted to produce the final analysis dataset.

2. Agency databases typically have already been cleaned and filtered by the respective agency's processes. Thus, the data cleaning and filtering processes used by Study Team members were still needed, but were less complex than for surveys.

3. Information obtained during breakout sessions or focus groups also requires a labor-intensive process to produce a transcript or comprehensive set of notes from which summaries can be written. The same transcript or set of notes can inform different analyses or questions under consideration by the Study Team.

   Information based on breakout sessions or focus group meeting is extremely valuable but often qualitative rather than quantitative in nature. Thus, evaluation of information, preparation of summaries, and extraction of key observations may require a consensus approach.